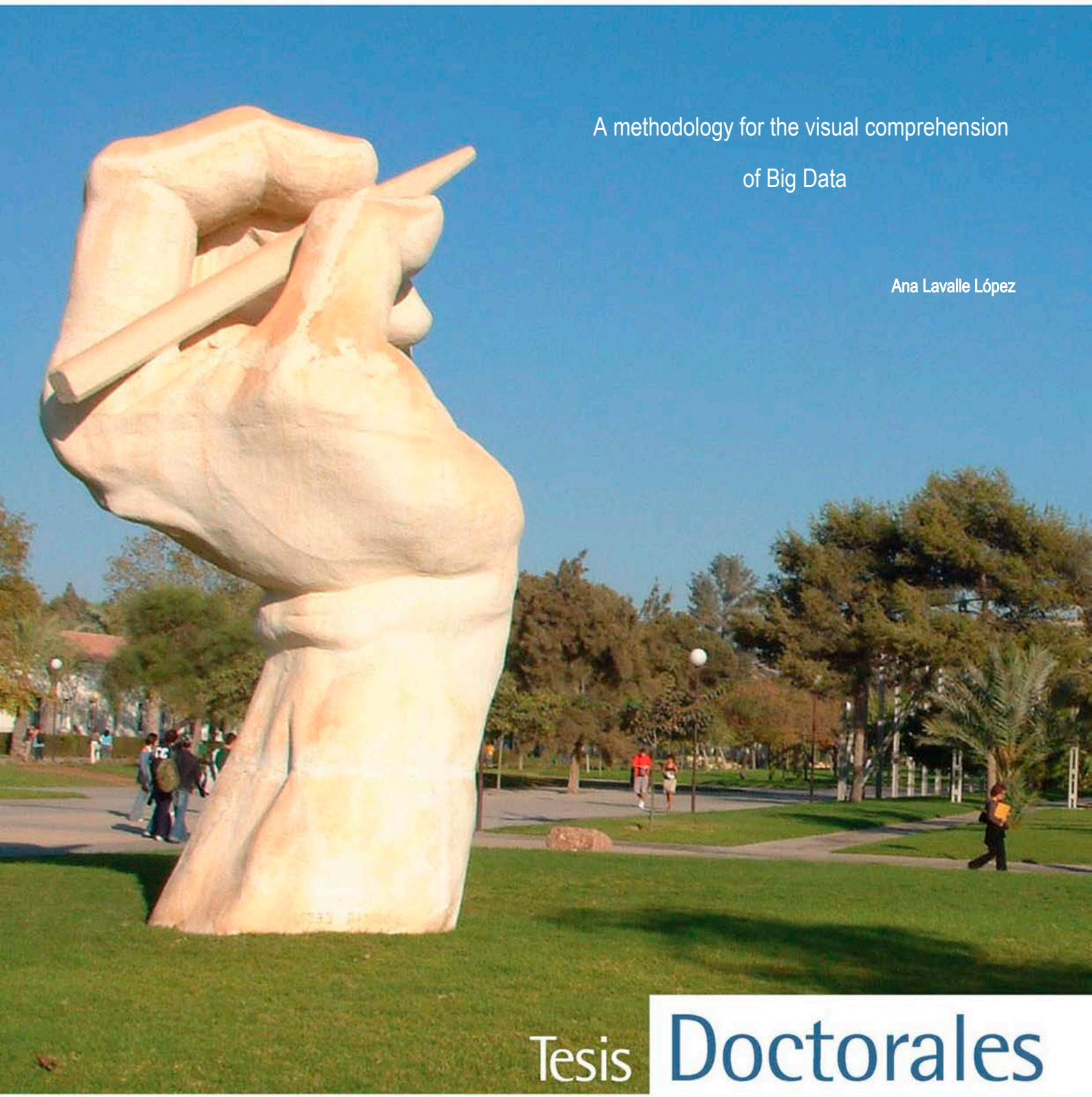
Universitat d'Alacant
Universidad de Alicante

A methodology for the visual comprehension of Big Data

Ana Lavalle López



Tesis Doctorales
UNIVERSIDAD de ALICANTE

Unitat de Digitalització UA
Unidad de Digitalización UA

DEPARTAMENTO DE LENGUAJES Y SISTEMAS INFORMÁTICOS

ESCUELA POLITÉCNICA SUPERIOR

A methodology for the visual comprehension of Big Data

ANA LAVALLE LÓPEZ

Tesis presentada para aspirar al grado de
DOCTORA POR LA UNIVERSIDAD DE ALICANTE
MENCIÓN DE DOCTORA INTERNACIONAL
DOCTORADO EN INFORMÁTICA

Dirigida por:
Dr. Juan Carlos Trujillo Mondéjar
Dr. Alejandro Maté Morga

Financiada por:
Universidad de Alicante y la empresa Lucentia Lab S.L.

Julio 2021

TESIS DOCTORAL POR COMPENDIO DE PUBLICACIONES

El presente documento contiene una síntesis del trabajo realizado por Ana Lavalle López bajo la dirección del Dr. Juan Carlos Trujillo Mondéjar y el Dr. Alejandro Maté Morga, para optar por el grado de Doctora en Informática. Se presenta en la Universidad de Alicante y se estructura según la normativa establecida para la presentación de tesis doctorales en forma de compendio de publicaciones.

Julio 2021

A mis padres, Felicidad y Bruno,
mi corazón y razón, mi soporte y libertad.
A mi Tita.
A Cristina, Pablo y Pedro.
A mi dream team Eva, Ponsoda, Bravo, Aurea y Susana.
A los amaneceres y atardeceres que me han acompañado estos tres años.

Agradecimientos

En primer lugar, mi sincero agradecimiento a mis directores de tesis, Juan Carlos Trujillo Mondéjar y Alejandro Maté Morga. A Juan Carlos, por ser mi guía y ayudarme a tomar las mejores decisiones desde el cariño y la experiencia como catedrático de universidad y director del grupo de investigación Lucentia. A Alejandro, por su gran apoyo incondicional, por guiarme en el día a día y seguir mis pasos de cerca. A ambos, por transmitirme toda su experiencia en investigación y transferencia de tecnología a lo largo de todo el desarrollo de la tesis doctoral.

Agradecer también a los miembros del grupo de investigación Lucentia, por su apoyo personal y profesional. Deseo hacer extensivo este agradecimiento al Departamento de Lenguajes y Sistemas Informáticos (DLSI) y al Instituto Universitario de Investigación en Informática (IUII) de la Universidad de Alicante, por los apoyos institucionales y las facilidades concedidas para realizar la presente tesis doctoral.

Un agradecimiento especial al catedrático Stefano Rizzi y a la Universidad de Bolonia, por acogerme en esa *bella* ciudad durante los 4 meses de mi estancia de investigación. Ha sido un placer trabajar con ellos y poder colaborar en artículos de investigación que forman parte del núcleo de la tesis.

Mención especial para la Universidad de Alicante y la empresa Lucentia Lab S.L., spin-off de la Universidad de Alicante, por haber cofinanciado mi beca predoctoral I-PI 03-18, enmarcada en la convocatoria de la Universidad de Alicante de ayudas destinadas a la formación predoctoral en colaboración con empresas.

Mi experiencia en Lucentia Lab ha sido crucial para el desarrollo de la tesis, dado que me ha permitido poner en práctica, a través de proyectos industriales y de innovación tecnológica, todo el marco teórico desarrollado,

así como testear e implantar en proyectos con clientes finales algunas de las soluciones desarrolladas en el seno de la presente tesis.

De mi paso por Lucentia Lab, me llevo no solo la inmensa experiencia profesional, sino también el apoyo y cariño recibido de todos sus empleados y directivos. A todos ellos, les estoy muy agradecida, en especial a Miguel Ángel Teruel, con quien trabajé de forma continua en varios proyectos y de quien obtuve ayuda en todo momento.

Por último y, no menos importante, me gustaría apuntar que la presente tesis doctoral ha sido cofinanciada también por varios proyectos de investigación en los cuales se ha enmarcado la metodología de visualización de *Big Data* desarrollada en esta tesis doctoral, entre ellos los proyecto del Plan Nacional de I+D+i ECLIPSE-UA (RTI2018-094283-B-C32) *(Enhancing data quality and security for improving business processes and strategic decisions in cyber physical systems)* y Aether-UA *(A smart data holistic approach for context-aware data analytics: smarter machine learning for business modelling and analytics)*, ambos subvencionados por el Ministerio de Ciencia, Innovación y Universidades. A este se viene a sumar la cofinanciación del proyecto de investigación de excelencia *Big data e inteligencia artificial para mejorar el diagnóstico de los afectados por la Covid-19*, concedido por la Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital de la Generalitat Valenciana.

«*La ciencia más útil es aquella cuyo fruto es el más comunicable.*»

Leonardo da Vinci (1452-1519)

Resumen

En una era donde el análisis de *Big Data* está a la orden del día, la analítica visual se convierte en un componente clave. Sin embargo, establecer unos objetivos analíticos y encontrar las visualizaciones que mejor se adapten a un contexto determinado es una tarea desafiante, especialmente cuando se trata con usuarios no expertos en visualización de datos. El uso de un tipo de visualización inadecuado puede llevar a malinterpretar los datos y a tomar decisiones equivocadas, provocando pérdidas significativas.

Por ello, el objetivo principal de la presente tesis doctoral es definir una metodología que agrupe una serie de técnicas y aproximaciones para mejorar la comprensión visual de *Big Data*. En concreto, se han analizado las necesidades actuales en la toma de requisitos para la generación de visualizaciones y se ha propuesto una metodología completa, desde la definición de requisitos hasta la implementación de visualizaciones, que guía al usuario en la definición de sus objetivos analíticos y genera automáticamente la mejor visualización para cada uno, agrupando dichas visualizaciones en cuadros de mandos.

La metodología está compuesta por (i) un modelo de requisitos de usuario, (ii) un modelo de perfilado de datos que extrae de forma semiautomática información sobre las características de las fuentes de datos y (iii) un modelo de visualización de datos. Nuestra propuesta ha sido evaluada y aplicada en distintos ámbitos, tales como ciudades inteligentes, procesos de producción industrial y entornos sanitarios. Además, con los resultados obtenidos y que se presentan en el trabajo, podemos concluir que se logra el objetivo principal del estudio, ya que, en línea con los experimentos realizados en el núcleo de la presente tesis doctoral, nuestra propuesta: (i) permite a los usuarios cubrir más cuestiones analíticas; (ii) mejora el conjunto de visualizaciones generadas; (iii) produce una mayor satisfacción

general en los usuarios.

La investigación realizada en la presente tesis doctoral ha dado como resultado diferentes artículos científicos que han sido presentados en congresos internacionales y revistas científicas de alto impacto, es por ello por lo que se elige presentar la tesis doctoral por compendio de publicaciones.

Palabras clave: *big data;* analítica de datos; requisitos de usuario; visualización de datos; cuadros de mando; inteligencia artificial; arquitectura dirigida por modelos.

Abstract

Big Data Analytics is continuously growing, especially since the last decade, and Visual Analytics have become a key component in order to analyze it. However, defining analytical goals and using the most suitable visualizations is a complex task, especially for non-expert users in data visualization. Consequently, it is possible that graphics are misinterpreted, contributing to making wrong decisions that lead to missed opportunities.

Therefore, the main goal of this doctoral thesis is to define a methodology that groups a series of techniques and approaches to improve the visual understanding of Big Data. Specifically, the current needs in the taking of requirements for the generation of visualizations have been analyzed and a complete methodology has been proposed, from the definition of the user requirements to the implementation of the visualizations. This methodology guides the user in the definition of their analytical goals and automatically generates the best suited visualization for each goal by grouping them into dashboards.

The methodology is composed by; (i) a User Requirements Model, (ii) a Data Profiling Model that semi-automatically extracts information about the characteristics of the data sources, and (iii) a Data Visualization Model. Our proposal has been evaluated and applied in different areas such as: Smart Cities, industrial production processes, and health environments. Furthermore, with the results obtained and presented in this work, we can conclude that the main goal of this doctoral thesis is achieved since, in line with the experiments carried out in the core of this doctoral thesis, our proposal; (i) allows users to cover more analytical questions, (ii) improves the set of generated visualizations, and improves (iii) the overall satisfaction of the users.

As a result of the research carried out in this doctoral thesis, nume-

rous scientific articles have been obtained and presented in international congresses and high impact journals. That is why this doctoral thesis is presented by a compendium of articles.

Keywords: big data; data analytics; requirements engineering; data visualization; dashboards; artificial intelligence; model-driven architecture.

Índice general

Parte I

Síntesis

Capítulo 1

Introducción

1.1. Motivación

Los datos están creciendo continuamente, especialmente desde la última década. En esta nueva era de *Big Data Analytics*, se ha generado un interés creciente tanto desde mundo académico como desde la industria [45].

El *Big Data*, definido principalmente como datos con un gran volumen, variedad, velocidad, veracidad y valor, se ha convertido en un área de investigación fundamental [44]. El interés por ellos ha aumentado en las diferentes fases del ciclo de vida de los datos: desde el almacenamiento hasta el análisis, limpieza o integración y, por supuesto, la visualización.

Por parte de la industria, los avances en las tecnologías de la información llevan a las organizaciones a aumentar su valor comercial y a buscar ventajas competitivas a través de la recopilación, almacenamiento, procesamiento y análisis de grandes cantidades de datos, en muchas ocasiones heterogéneos y generados a un ritmo cada vez mayor [22].

El concepto ha ganado una notoriedad significativa durante los últimos años, ya que, con los avances tecnológicos, muchas áreas de negocio pueden beneficiarse de este fenómeno. El análisis de datos desempeña un papel esencial en diversos sectores, tales como la inteligencia de negocio, sistemas de recomendación, sanidad, ciudades inteligentes, transporte, maquinaria industrial, detección de fraude, publicidad web y *marketing* entre otros [37].

El *Big Data* como tema de investigación enfrenta innumerables desafíos y la visualización de datos se está convirtiendo en un elemento estratégico

para la exploración de grandes conjuntos de datos debido al gran impacto que estos tienen [2]. El uso de unas visualizaciones adecuadas es crucial para extraer información precisa de los datos y guiar a los usuarios de estas a tomar decisiones mejor informadas.

De hecho, según [4], el tamaño del mercado global de visualización de datos se situó en 8,85 mil millones de dólares en 2019 y se prevé que alcance los 19,20 mil millones de dólares en 2027. La evolución de las técnicas de análisis y visualización se encuentra en el centro de las estrategias comerciales, y cada vez más líneas de investigación se centran en la visualización de datos.

Según una encuesta de Salesforce [43], el 76 % de grandes empresas coinciden totalmente en otorgarle un gran valor a las herramientas analíticas para obtener información estratégica a partir de los datos.

Una representación eficaz, eficiente e intuitiva de los datos a analizar puede resultar tan importante como el propio proceso analítico [9]. Sin embargo, elegir e implementar las visualizaciones más adecuadas para cada conjunto de datos es una tarea realmente complicada, especialmente cuando se trabaja con *Big Data*. En estos escenarios, es común encontrar fuentes de datos heterogéneas que requieren un amplio conocimiento de los datos subyacentes para crear una visualización adecuada [8]. Además, el uso de un tipo de visualización inadecuado puede llevar a malinterpretar los datos y tomar decisiones equivocadas. Algunos tipos de gráficos no son recomendables para comunicar cierto tipo de información, y pueden surgir problemas a la hora de interpretar los datos. Por ejemplo, utilizar un gráfico circular para comparar demasiadas variables puede dificultar el visionado de los valores de dichas variables.

Por lo tanto, encontrar la visualización que mejor se adapte a un contexto determinado es una tarea desafiante, y el problema se agrava cuando los usuarios no son expertos en visualización de datos, cosa que es habitual, incluso es común que no tengan una idea clara de los objetivos para los que están construyendo las visualizaciones.

1.2. Definición del problema

A pesar del interés mostrado, elegir e implementar las visualizaciones más adecuadas para cada contexto sigue siendo todo un reto, especialmente

cuando se trabaja con *Big Data* y con usuarios no expertos en visualización de datos.

En estos escenarios, es habitual encontrar fuentes de datos heterogéneas que exigen un amplio conocimiento de los datos subyacentes para crear una visualización adecuada [8]. Esto requiere un gran esfuerzo por parte de los usuarios de las visualizaciones, que muy probablemente no sepan exactamente qué tipo de información quieren extraer de los datos o cuál sería el mejor tipo de visualización a utilizar.

En consecuencia, cabe la posibilidad de que los gráficos sean malinterpretados. El uso de un tipo de visualización inadecuado puede llevar a malinterpretar los datos y tomar decisiones equivocadas que provoquen pérdidas significativas. [20] defiende que una de las razones de la falta de visualizaciones avanzadas son los usuarios, ya que a menudo no saben cómo representar sus datos. [7] afirma que los motivos de los problemas con las visualizaciones pueden ser dobles: por la codificación (causada por el diseñador/desarrollador de la visualización), o por la decodificación (causada por el lector/usuario), es decir, los problemas que puede tener el usuario de la visualización en la interpretación de esta.

Es fundamental considerar los posibles riesgos y errores que se pueden cometer durante el diseño y generación de visualizaciones. [39] señala que el proceso de renderizado puede introducir incertidumbre en tres áreas: en la recopilación de datos, errores algorítmicos, y exactitud y precisión computacional. Además, en [21] los autores identifican como posibles fuentes de incertidumbre en la representación visual, la adquisición, el modelo, la transformación y la visualización.

Otro aspecto crítico es que, aparentemente, un gran conjunto de tipos de visualizaciones puede ser igualmente válido para cualquier conjunto de datos dado, lo cual se ha demostrado que es incorrecto [47]. Cada conjunto de datos y cada análisis tiene sus características particulares y no siempre todos los tipos de visualización son válidos.

Para abordar este problema, algunos trabajos han propuesto diferentes formas de encontrar la mejor visualización para cada análisis. [5] examina las principales clasificaciones propuestas en la literatura y las integra en un marco basado en seis requisitos de visualización. [10] propone un marco para elegir la mejor visualización donde los principales tipos de gráficos están relacionados con los objetivos de los usuarios y con la dimensionalidad, cardinalidad y el tipo de datos que soportan. Finalmente, [40] propone

una clasificación más detallada de los tipos de datos y relaciona cada tipo común de gráfico con los objetivos de los usuarios con los que cumple mejor.

Otro propuesta es SkyViz [19], un enfoque donde los usuarios definen un contexto de visualización estructurado para automatizar la traducción del contexto en una visualización adecuada. Sin embargo, tal y como reconocen los autores, definir un contexto de visualización desde cero puede ser un desafío para los usuarios que habitualmente no son expertos en visualización de datos.

A pesar del trabajo realizado en este campo, ninguno de los enfoques mencionados proporciona metodologías o herramientas que orienten a los usuarios no expertos desde la especificación de sus objetivos hasta la generación de visualizaciones y su agrupación en cuadros de mandos *(dashboards)* adecuados que faciliten la extracción de conocimiento.

Por otro lado, el uso de algoritmos de inteligencia artificial (IA) también se ha convertido en un componente clave de muchos procesos. En estos casos es crucial tener en cuenta el sesgo de los datos. Cuando este no se advierte, puede afectar significativamente a la interpretación de los datos y, en consecuencia, tener un impacto devastador en los resultados de la IA [6].

Un área donde los sesgos pueden conducir a consecuencias mortales es en la atención médica. Identificar como sano a un paciente con una enfermedad grave puede retrasar su tratamiento. En este sentido, [3] destaca la existencia de sesgos como una de las amenazas más comunes para la validez de la investigación en el ámbito de la salud.

Sin embargo, no solo la atención médica se ve afectada significativamente por el sesgo de datos. Ha habido múltiples casos de aprendizaje sesgado que han dado lugar a una controversia significativa. Por ejemplo, en la plataforma LinkedIn, cuando se busca un contacto femenino, en ciertos casos se recomienda un nombre masculino similar [14]. Otro caso controvertido que ha aparecido en los titulares de prensa ha sido la aplicación de fotos de Google, debido a que identificaba, erróneamente, a personas de color como gorilas [34].

Como tal, el sesgo de datos se ha convertido en una preocupación importante no solo en la comunidad científica, también en grandes empresas como Amazon, Facebook, Microsoft, Google e IBM, las cuales invierten

recursos y esfuerzos para abordar el problema [46].

Se acepta ampliamente que el sesgo se puede clasificar en dos tipos. *Class Imbalance*, cuando las clases no están representadas uniformemente en los datos, es decir, algunas categorías del conjunto de datos tienen una representación más alta que el resto de categorías; y *Dataset Shift*, cuando la distribución de datos en el conjunto de datos de entrenamiento y de prueba es diferente.

Autores de esta área han sugerido diferentes técnicas para abordar estos problemas. Por lo general, proponen enfoques para tratar los problemas de datos desequilibrados en las tres siguientes perspectivas [33]:

- Perspectiva de datos: Se utilizan técnicas para reequilibrar artificialmente la distribución de datos, ya sea mediante un sobremuestreo creando de más datos de las clases menos representadas [11] o bien, submuestreando las clases mayoritarias eliminando datos de estas [18].

- Perspectiva algorítmica: En este caso se intentan ajustar los algoritmos durante el proceso de entrenamiento para mejorar su rendimiento sobre conjuntos de datos sesgados [35].

- Perspectiva mixta: Se utilizan perspectivas de datos y algorítmicas para determinar la predicción final [16].

Desafortunadamente, la mayoría de los enfoques desarrollados hasta el momento se centran principalmente en el aprendizaje automático y en el reequilibrio de los conjuntos de datos sesgados. Sin embargo, ese enfoque no siempre es válido, ya que dichas propuestas modifican la distribución de los datos y se podrían descartar u ocultar aspectos importantes de los mismos. Como argumenta [12], la equidad de las predicciones debe evaluarse en el contexto de los datos.

Por consiguiente, se percibe la falta de un enfoque general que advierta a los usuarios de la existencia de sesgos y permita analizarlos visualmente desde diferentes perspectivas sin alterar el conjunto de datos.

Además, cuando se trabaja con inteligencia artificial es habitual añadir aprendizaje automático predictivo para conocer si el proceso se está ejecutando tan bien como se esperaba [42]. Estas técnicas a menudo se basan en el uso de redes neuronales. La entrada de las redes neuronales suele ser

el estado general del sistema (tuplas de datos generados por todo el sistema) [48], mientras que la salida proporciona una clasificación binaria sobre "el sistema está funcionando correctamente" o "habrá un problema". Es decir, las redes neuronales actúan como una caja negra, no proporcionan información sobre la parte del sistema que va a producir el problema [15].

Por tanto, también es crucial completar la información proporcionada por redes neuronales mediante detalles visuales sobre la evolución de los datos del proceso. Estas visualizaciones permitirán identificar anomalías en ciertas partes del sistema. Aun así, la creación de tales visualizaciones es algo trivial, debido al gran volumen de datos producidos en múltiples magnitudes, lo que provoca que sea todo un desafío facilitar a los usuarios solo la información necesaria sin sobrecargarla.

1.3. Hipótesis de partida y objetivos

Por lo tanto, según todo lo expuesto con anterioridad, la presente tesis doctoral parte de la hipótesis inicial de que es factible mejorar y sistematizar las aproximaciones actuales para la visualización de *Big Data*.

En consecuencia, el objetivo de investigación de la presente tesis doctoral es definir una metodología que aglutine una serie de técnicas y aproximaciones para mejorar la comprensión visual de *Big Data* mediante los siguientes objetivos específicos:

O1 - Definir una metodología sistemática que permita derivar visualizaciones orientadas a usuarios no expertos.

O2 - Definir un marco de requisitos que permita especificar de manera clara los objetivos analíticos que se persiguen.

O3 - Automatizar la obtención de las visualizaciones a partir del marco de requisitos definido.

O4 - Aplicar y evaluar el impacto de las técnicas desarrolladas en distintas áreas.

1.4. Método de trabajo

Dado que la presente tesis doctoral se publica mediante compendio de publicaciones, siguiendo la normativa establecida por la Escuela de Doctorado de la Universidad de Alicante (Consejo de Gobierno de la Universidad de Alicante de 29 de septiembre de 2020, BOUA de 16 de octubre de 2020), se desarrollaron y publicaron diversos artículos científicos para abordar los objetivos previamente nombrados.

Durante la investigación se siguió una metodología de trabajo basada en la aproximación investigación-acción [32], llevando a cabo reuniones de seguimiento periódicas entre la doctoranda y los directores, a fin de evaluar los avances obtenidos hasta dicho momento. Posteriormente, una vez que se intensificó la fase de prueba y testeo de las soluciones en entornos industriales, en estas reuniones periódicas se incluyó al personal de la empresa involucrado en los proyectos donde se testeaban las soluciones.

A continuación, la figura 1.1 resume las actividades principales desarrolladas a lo largo de los cuatro cursos que han compuesto la presente tesis doctoral para alcanzar los objetivos planteados. Donde se encuentra: coloreado en gris el trabajo desarrollado en los cursos; coloreado en lila, las actividades comunes de la escuela de doctorado; coloreados en naranja, congresos nacionales e internaciones; coloreado en amarillo, la estancia de investigación; coloreado en verde, la publicación de artículos en revistas; finalmente, en color rojo, el registro de una herramienta *CASE*.

- **Curso 17/18:** En primer lugar, se llevaron a cabo las actividades comunes de la escuela de doctorado de la Universidad de Alicante. Esta tarea esta formada por las cuatro actividades siguientes: herramientas para la gestión y recuperación de la información; fines y objetivos de la investigación; modelos de comunicación científica; modelos de transferencia del conocimiento. Simultáneamente, se realizó una revisión del estado del arte de las aproximaciones para la visualización de datos, permitiendo detectar problemas en el estado actual del campo de estudio. Al finalizar el curso, se presentó un póster sobre el trabajo desarrollado hasta el momento en las Jornadas del doctorado en Informática de 2018, organizada por la Universidad de Alicante [23]. En el mes de junio, comenzó la estancia de investigación en la Universidad de Bolonia, que se extendió hasta octubre del siguiente curso.

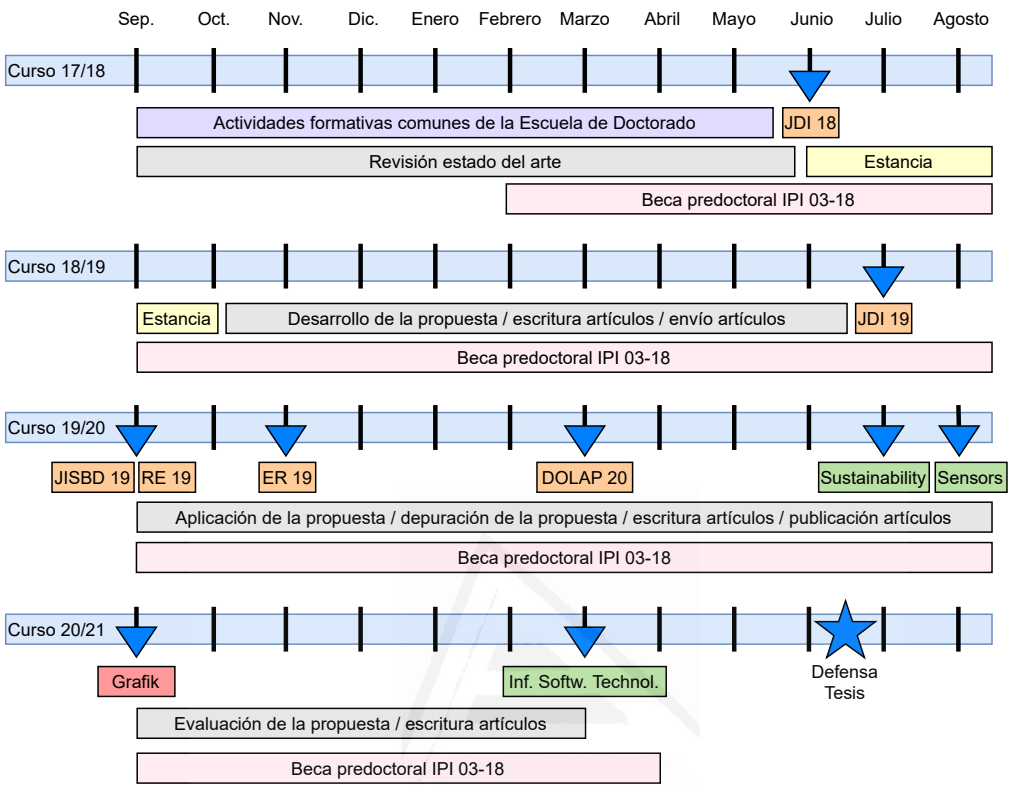


Figura 1.1: Linea temporal del las actividades desarrolladas

- **Curso 18/19:** Los conocimientos adquiridos en la estancia sirvieron para desarrollar una primera aproximación donde se definió el marco conceptual de la metodología de generación de visualizaciones. Se desarrolló la propuesta, se escribieron artículos científicos y se enviaron a distintos congresos de alto impacto. Finalmente, coincidiendo con la clausura del curso, se realizó una presentación del trabajo desarrollado hasta ese momento en las Jornadas del doctorado en Informática de 2019 organizadas por la Universidad de Alicante [24].

- **Curso 19/20:** Se presentaron distintos artículos en congresos científicos de gran relevancia, tanto nacionales como internacionales. El primer artículo se presentó en las *Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2019)* [25], celebradas en Cáceres (España)

del 2-4 de septiembre. A finales del mismo mes, del 23-27 de septiembre, se presentó el artículo [28] en el *27th IEEE International Requirements Engineering Conference (RE 2019)* en Jeju (Corea del Sur). Del 4-7 de noviembre se presentó el artículo [26] en el *38th International Conference on Conceptual Modeling (ER 2019)* en Salvador de Bahía (Brasil). El 30 de marzo de 2020 se presentó el artículo [27] en el *22nd International Workshop On Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP 2020)*, celebrado en Copenhague (Dinamarca), que debido a la pandemia COVID-19 finalmente se realizó *online*.

Paralelamente, durante ese curso, se continuó trabajando en la ampliación de la propuesta, aplicándose en distintos escenarios y entornos industriales, lo que nos permitió descubrir nuevas necesidades y depurarla. Al final del curso, se publicó en el mes de julio el artículo [31] en la revista *Sustainability*, y en agosto, el artículo [30] en la revista *Sensors*.

- **Curso 20/21:** En el último curso de la tesis, se registró la propiedad intelectual de la herramienta Grafik en el Servicio de Transferencia de Resultados de Investigación de la Universidad de Alicante. Se desarrolló el artículo científico [29] que contenía una evaluación de la propuesta y fue publicado en la revista *Information and Software Technology*. Finalmente, se preparó la presente tesis doctoral y su defensa.

1.5. Resultados

Las tesis doctorales por compendio de publicaciones reproducen de forma íntegra y literal los resultados de investigación que han sido publicados en revistas indexadas, capítulos de libros o congresos de alto impacto. Por tanto, en las siguientes secciones sintetizamos las publicaciones resultantes del trabajo desarrollado que compone la presente tesis doctoral.

Para resumir de forma visual el trabajo desarrollado en cada capítulo, el cuadro 1.1 recoge los objetivos abordados en cada uno de los diferentes artículos y el capítulo que lo recoge de forma íntegra y literal. Por su parte, la figura 1.2 resume las principales aportaciones de cada una de las publicaciones.

Cuadro 1.1: Alcance en cada capítulo de los objetivos definidos

	O1	O2	O3	O4
Capítulo 4 (RE 2019)	++	++	++	+
Capítulo 5 (ER 2019)	++	++	++	+
Capítulo 6 (Sustainability)	+	+	+	++
Capítulo 7 (Sensors)	+	+	+	++
Capítulo 8 (Inf. Softw. Technol.)				++
Capítulo 9 (DOLAP 2020)	+	+	+	++

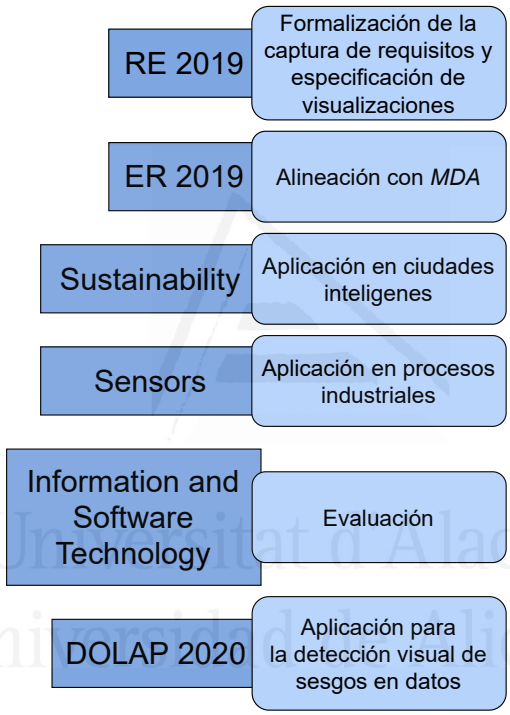


Figura 1.2: Principales aportaciones de cada publicación

1.5.1. Síntesis de *Visualization Requirements for Business Intelligence Analytics: A Goal-Based, Iterative Framework* (RE 2019)

El problema fundamental al que nos enfrentamos consiste en ayudar a los usuarios no expertos a obtener sus visualizaciones. Para ello, el primer paso

consiste en disponer de una aproximación que permita obtener visualizaciones. En este sentido, como hemos visto anteriormente, existen distintas técnicas, si bien ninguna de ellas cubre por completo las necesidades de los usuarios. Como base de este trabajo, partimos de la aproximación presentada en [19] en la que se propone SkyViz, un enfoque donde los usuarios definen un contexto de visualización estructurado basado en siete coordenadas, con el fin de automatizar la traducción del contexto en una visualización adecuada. Sin embargo, como reconocen los autores, definir un contexto de visualización desde cero puede ser un desafío para los usuarios que habitualmente no son expertos en visualización de datos.

Es por ello por lo que en el comienzo de nuestro trabajo, tal y como se refleja en el capítulo 4, se complementa SkyViz con la definición de un enfoque de modelado iterativo basado en objetivos y haciendo uso del lenguaje i* [13]. A su vez, se han definido un conjunto de pautas y guías para capturar las necesidades de los usuarios y poder así derivar las visualizaciones de datos más adecuadas.

Nuestra propuesta proporciona: (i) una secuencia de pautas y guías para ayudar a usuarios no expertos a definir sus objetivos y alcanzarlos, haciendo uso de las fuentes de datos disponibles; (ii) traduce los objetivos del usuario en un contexto de visualización; (iii) extrae de forma semiautomática información de las fuentes de datos a visualizar; (iv) proporciona un diseño racional para cuadros de mando *(dashboards)*. De esta forma, los usuarios no expertos en visualizaciones pueden descubrir sus objetivos de análisis, comunicarlos y obtener, con un esfuerzo limitado, las visualizaciones que mejor se adapten a sus necesidades. Además, estas visualizaciones se agruparán en cuadros de mando para permitir a los usuarios monitorear y medir sus objetivos de manera efectiva.

Para ello en este artículo se propone: (i) una metodología que abarca todo el proceso; (ii) un metamodelo para la formalización de la especificación de los requisitos de las visualizaciones basado en i* en su versión 2.0 [13] y en la extensión de i* para almacenes de datos [36]; (iii) una serie de guías y pautas para asistir a los usuarios en la metodología; finalmente, (iv) un analizador de fuentes de datos creado para extraer información de forma semiautomática de las fuentes de datos.

Finalmente, la propuesta es llevada a la práctica con un ejemplo ilustrativo centrado en una compañía de recaudación de impuestos, lo que permite evaluar la validez de nuestra propuesta.

1.5.2. Síntesis de *Requirements-Driven Visualizations for Big Data Analytics: A Model-Driven Approach* (ER 2019)

Continuando con nuestro objetivo de brindar autonomía al usuario y, al mismo tiempo, automatizar la obtención de visualizaciones a partir de un marco de requisitos definido, en esta propuesta presentada en el capítulo 5, proseguimos con nuestro trabajo, proporcionándole a la metodología una alineación con arquitectura dirigida por modelos (*MDA*, acrónimo en inglés de *model-driven architecture*) [38].

Se propone un enfoque basado en *MDA* para facilitar la analítica visual adaptada para usuarios no expertos en visualizaciones. El enfoque se basa en tres modelos: (i) modelo de requisitos de usuario; (ii) modelo de perfilado de datos; (iii) modelo de visualización de datos donde se recogen detalles que deberán representar las visualizaciones, independientemente de la tecnología seleccionada para su implementación. Junto a estos modelos, se proponen un conjunto de transformaciones modelo-modelo y modelo-texto que permiten obtener la implementación correspondiente de forma semiautomática, evitando así la intervención de los usuarios no expertos en el proceso. De esta forma, los usuarios no necesitan centrarse en las características de las visualizaciones, solo lo harán en sus requisitos de información y obtendrán las visualizaciones que mejor se adapten a sus necesidades.

Siguiendo los principios básicos de *MDA*, se definen las siguientes capas y transformaciones:

- Capa CIM: Contiene el modelo de requisitos de usuario cuyo principal objetivo es capturar las necesidades analíticas de los usuarios y determinar qué tipos de visualizaciones necesitan para alcanzarlos.
- Capa PSM: Contiene el modelo de perfilado de datos el cual extrae información, de forma semiautomática, sobre las características de las fuentes de datos seleccionadas en el modelo de requisitos de usuario.
- Transformación modelo-modelo: La información proveniente del modelo de requisitos de usuario y del modelo de perfilado de datos se transforma generando el modelo de visualización de datos.
- Capa PIM: Contiene el modelo de visualización de datos. Este modelo permite a los usuarios especificar detalles de visualización indepen-

dientemente de la tecnología utilizada para la implementación. También permite determinar si las visualizaciones propuestas cumplen con los requisitos esenciales para los que fueron creadas y si contribuyen a alcanzar los objetivos de los usuarios.

- Transformación modelo-texto: Tiene como entrada el modelo de visualización de datos y transforma cada elemento de la especificación de visualización en texto a nivel de código para la biblioteca gráfica seleccionada.

Gracias a la alineación con *MDA*, los tomadores de decisiones obtienen las visualizaciones que mejor se adaptan a sus necesidades y a los datos disponibles de forma semiautomática. Esta propuesta permite refinar gradualmente las visualizaciones hasta obtener su implementación ideal. Asimismo, otorga independencia a los usuarios de los aspectos técnicos subyacentes a las visualizaciones y y les deja centrarse en sus objetivos analíticos.

Finalmente, se presenta un caso de estudio basado en una compañía de recaudación de impuestos.

1.5.3. Síntesis de *Improving Sustainability of Smart Cities through Visualization Techniques for Big Data from IoT Devices* (Sustainability)

Una vez definimos la propuesta, pasamos a aplicarla a distintos escenarios del mundo real. Esto nos permite comprobar la efectividad de la metodología e incorporar elementos que no se habían tenido en cuenta.

En el capítulo 6 nos centramos en el ámbito de las ciudades inteligentes *(Smart Cities)* con el objetivo de resaltar el valor que se puede extraer de los datos generados en ellas.

Como sucede en cualquier proyecto de *Big Data*, es necesario capturar, almacenar, procesar y analizar grandes cantidades de datos de diferentes fuentes para transformarlos en conocimiento útil para los usuarios. El objetivo principal de este trabajo es proporcionar una metodología que ayude a los responsables de las ciudades a tomar decisiones que estén respaldadas por los datos, conduciendo así a la ciudad hacia un crecimiento más sostenible.

En nuestros trabajos anteriores [28] [26] definimos una metodología que

ayuda y guía a los usuarios a definir y alcanzar sus objetivos. En este artículo se muestra cómo nuestra propuesta es aplicada al contexto de las *Smart Cities*, explotando conjuntos de datos reales y combinándolos con algoritmos de inteligencia artificial. Además, se mejoran las propuestas anteriores, incorporando elementos requeridos para procesar datos en tiempo real. Y finalmente, se representa la información a los usuarios de manera que sirva de ayuda para comprender mejor el resultado de algoritmos de inteligencia artificial.

Así pues, las ventajas de esta propuesta se resumen en: (i) ayuda a los usuarios a definir sus objetivos y alcanzarlos a través de la toma de decisiones respaldadas por visualizaciones históricas y en tiempo real; (ii) ayuda a comprender visualmente el resultado de los algoritmos de inteligencia artificial; (iii) permite a los usuarios recopilar evidencias para tomar decisiones estratégicas y tácticas en el contexto de las *Smart Cities*. Sin los beneficios introducidos por esta propuesta, sería difícil para los usuarios comprender el estado en el que se encuentran sus procesos, y así poder tomar las mejores decisiones en relación a ellos.

Para evaluar el impacto de la propuesta, se presenta un caso de estudio basado en la gestión de llamadas del departamento de bomberos de la ciudad de San Francisco [1]. También se ha realizado un experimento con 12 usuarios no expertos en visualización de datos.

1.5.4. Síntesis de *Fostering Sustainability through Visualization Techniques for Real-Time IoT Data: A Case Study based on Gas Turbines for Electricity Production* (Sensors)

La realización de la tesis doctoral en colaboración con la empresa Lucentia Lab nos ha brindado la posibilidad de testear e implantar propuestas teóricas en proyectos con clientes finales. En este caso, el artículo presentado en el capítulo 7, muestra la aplicabilidad de nuestro enfoque en un caso de estudio real de una empresa de producción de electricidad mediante turbinas de gas.

En él se propone un enfoque metodológico para monitorear maquinaria industrial a través de técnicas de visualización basadas en datos en tiempo real provenientes del internet de las cosas (*IoT*, acrónimo en inglés de *Internet of Things*). El objetivo principal de este trabajo es ayudar a

usuarios no expertos en visualizaciones de datos a localizar y comprender visualmente fallos de los sistemas que podrían surgir en un proceso de producción, permitiéndoles así tomar las decisiones más sostenibles en cada situación a través de técnicas de visualización para datos en tiempo real.

Además, en este escenario se trabaja con redes neuronales, las cuales actúan como una caja negra, sin proporcionar información sobre la parte del sistema donde se podría producir un fallo. Por tanto, en este enfoque también se completa la información proporcionada por las redes neuronales mediante visualizaciones sobre la evolución de los datos del proceso, lo que permitirá identificar anomalías en las distintas partes del sistema.

En este artículo, (i) se muestra nuestra propuesta aplicada al contexto de la maquinaria industrial; (ii) se amplía nuestro metamodelo de especificación de visualizaciones, agregando nuevos elementos para adecuarlo a escenarios en tiempo real; (iii) proporcionamos una metodología novedosa para monitorear maquinaria industrial dividida en dos fases (la primera realizada antes del inicio proceso de producción y la segunda ejecutada durante el proceso de producción). Esta metodología (a) permite a los usuarios definir los objetivos y requisitos del proceso de producción; (b) deriva automáticamente el tipo de visualización más adecuado para cada contexto; (c) ayuda a los usuarios a comprender visualmente el resultado de los modelos de inteligencia artificial; (d) proporciona visualizaciones para ayudar a los usuarios a tomar la decisión más sostenible en cada situación.

Con el fin de probar y demostrar la aplicabilidad de nuestra propuesta, se presenta un caso de estudio basado en turbinas de gas para la generación de electricidad. Estas turbinas recopilan datos procedentes de 80 sensores en tiempo de ejecución.

La complejidad de los datos, la velocidad a la que se generan y la importancia de detectar fallos brindan un escenario perfecto para probar que el enfoque presentado mejora la sostenibilidad del proceso, es decir, mejora el rendimiento del proceso al evitar la ruptura de las máquinas. Gracias a las visualizaciones generadas, los operadores de la maquinaria podrán monitorear la calidad de los sistemas, ayudarán a prevenir averías en las máquinas, a identificar si el proceso de producción está funcionando tan bien como se esperaba y a comprender y correlacionar los resultados de algoritmos de IA.

1.5.5. Síntesis de *A Methodology to Automatically Translate User Requirements Into Visualizations: Experimental Validation* (Inf. Softw. Technol.)

Habiendo aplicado la propuesta en distintos escenario del mundo real, con el objetivo de evaluar el impacto real percibido por los usuarios, el capítulo 8 se centra en la evaluación de la propuesta.

En este artículo se presenta la metodología completa de definición de objetivos y derivación de visualizaciones, desde la definición de los requisitos del usuario hasta la implementación de las visualizaciones. En él se proponen los modelos presentados en trabajos anteriores: el modelo de requisitos de usuario [28], para capturar las necesidades analíticas de los usuarios; el modelo de perfilado de datos [26], para extraer de forma semiautomática las características de las fuentes de datos y el modelo de visualización de datos [26].

Para evaluar el impacto de nuestra propuesta, se ha presentado un ejemplo ilustrativo y se han realizado una serie de experimentos con usuarios no expertos en visualización de datos. Los experimentos fueron levados a cabo por 97 participantes, entre ellos 84 estudiantes de ingeniería informática y 13 empleados de una empresa tecnológica, todos ellos no expertos en visualización de datos. Estos experimentos confirmaron la validez de nuestra propuesta, ya que se demostró que nuestra metodología, (i) permite a los usuarios cubrir más cuestiones analíticas; (ii) mejora el conjunto de visualizaciones generadas; (iii) produce una mayor satisfacción general en los usuarios.

Por lo tanto, siguiendo nuestra propuesta, los usuarios no expertos en visualizaciones de datos podrán expresar con mayor eficacia sus necesidades analíticas y obtener el conjunto de visualizaciones que mejor se adapte a sus objetivos.

1.5.6. Síntesis de *An Approach to Automatically Detect and Visualize Bias in Data Analytics* (DOLAP 2020)

Finalmente, otro de los problemas presentados en la presente tesis doctoral y de gran actualidad se trata del manejo del sesgo. La mayoría de los enfoques desarrollados hasta el momento se centran principalmente en el reequilibrio de los conjuntos de datos, pero esto no siempre es recomendable, es más, cuando este no se advierte, puede afectar notablemente a la

interpretación de los datos, lo que provoca consecuencias nefastas.

En el capítulo 9 se presenta un enfoque que complementa los trabajos anteriores [26, 28] al incluir una aproximación para la detección y visualización de sesgos en la analítica de datos.

Concretamente se propone un enfoque que incluye un algoritmo novedoso para detectar sesgos en conjuntos de datos. Dicho algoritmo permite detectar automáticamente sesgos en conjuntos de datos mediante un análisis de estos, teniendo en cuenta el alcance del análisis. Además, se representa el resultado obtenido de forma visual, para que sea comprensible para usuarios no expertos.

De esta manera, los usuarios pueden percibir no solo la existencia de sesgos en sus conjuntos de datos, sino también cómo pueden estos afectar a sus análisis y a los algoritmos de IA, evitando así resultados no deseados.

La gran ventaja de esta propuesta es que ayudamos a los usuarios a comprender sus datos y tomar decisiones considerando el sesgos desde diferentes perspectivas, sin alterar nunca el conjunto de datos. Además, los usuarios pueden beneficiarse de la reducción del tiempo necesario para inspeccionar y comprender los sesgos existentes dentro de sus conjuntos de datos y, al mismo tiempo, evitan que los sesgos pasen desapercibidos, con los problemas que esto conlleva.

Capítulo 2

Publicaciones y visibilidad

2.1. Publicaciones

Como se ha comentado anteriormente, la presente tesis doctoral se publica mediante compendio de publicaciones, siguiendo la normativa establecida en el reglamento de régimen interno de la Escuela de Doctorado de la Universidad de Alicante. Una tesis por compendio de publicaciones reproduce de forma íntegra y literal los resultados de investigación de la tesis doctoral que han sido publicados mediante artículos científicos en revistas indexadas, capítulos de libros o congresos.

Como resultado de la investigación realizada en la presente tesis doctoral se han elaborado diferentes artículos que han sido presentados en congresos internacionales y revistas científicas de alto impacto, es por ello por lo que se elige defender la presente tesis doctoral por compendio de publicaciones.

Durante el periodo del doctorado, han sido publicados un total de 7 artículos, 6 de los cuales forman parte del núcleo de la presente tesis doctoral. Además, una herramienta CASE *(Computer Aided Software Engineering)* ha sido registrada y está en proceso de depuración con el objetivo de realizar pruebas con usuarios antes de su licitación formal.

Las secciones 2.1.1 y 2.1.2 recogen información sobre la lista de congresos y revistas en los que se han publicado nuestros artículos.

2.1.1. Congresos

Esta sección detalla los artículos recogidos en las actas de los congresos internacionales presentados en el cuadro 2.1, incluyendo información sobre el nombre del congreso, la clasificación según *GII-GRIN-SCIE (GGS)* [17] y según *The Computing Research and Education Association of Australasia (CORE)* [41], así como la localización y las fechas en las que se celebró el congreso.

Todos los congresos en los que se han publicado artículos cuentan con procesos de revisión por pares y han sido un punto clave en el desarrollo de la investigación, ya que compartir el desarrollo de los trabajos con expertos en el dominio ha permitido descubrir nuevos puntos de vista. Estas nuevas aportaciones han servido para explorar distintas ramas que se han transformado en nuevos objetivos de investigación.

Cabe destacar que la clasificación proporcionada por *GII-GRIN-SCIE (GGS)* [17] y la cual se detalla en la tercera columna del cuadro 2.1 *(Conf. Rating)* fue actualizada por última vez el 30 de mayo de 2018, mientras que la clasificación proporcionada por *The Computing Research and Education Association of Australasia (CORE)* [41] fue actualizada en 2020.

Cuadro 2.1: Publicaciones presentadas en congresos

	Congreso	**Conf. Rating**	**Ciudad/País**	**Fecha**
C1	27th IEEE International Requirements Engineering Conference (RE 2019)	Clase 2 Core A	Jeju Corea del Sur	23-27 Sep. 2019
C2	38th International Conference on Conceptual Modeling (ER 2019)	Clase 3 Core A	Salvador de Bahía Brasil	4-7 Nov. 2019
C3	22nd International Workshop On Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP 2020)	Clase 2 Core B	Copenhague Dinamarca	30 Marzo 2020

Los artículos presentados en los distintos congresos se detallan a continuación así como el capítulo que los recoge de forma íntegra:

C1 - Visualization Requirements for Business Intelligence Analytics: A

Goal-Based, Iterative Framework *(Disponible en el capítulo 4)*

C2 - Requirements-Driven Visualizations for Big Data Analytics: A Model-Driven Approach *(Disponible en el capítulo 5)*

C3 - An Approach to Automatically Detect and Visualize Bias in Data Analytics *(Disponible en el capítulo 9)*

2.1.2. Revistas

Los artículos publicados en revistas científicas encuentran su referencia en el cuadro 2.2. La primera columna representa la identificación de la revista; la segunda columna, el nombre de la revista junto con su ISSN; la tercera columna muestra el factor de impacto por el *Journal Citations Report (JCR)*; la cuarta columna, el cuartil en el que se encuentra posicionada; la quinta columna, la posición de la revista; la sexta, el área de conocimiento en el que se encuentra posicionada; finalmente, el año en el que se ha publicado.

Cabe destacar que dicha clasificación hace referencia al *Journal Citations Report (JCR)* publicado en el año 2019. Dicho reporte se publica anualmente, aproximadamente en el mes de junio.

Cuadro 2.2: Publicaciones presentadas en revistas

	Revista	**F.I.**	**Cuartil**	**Posición**	**Área**	**Año**
J1	Sustainability ISSN: 2071-1050	2,576	Q2	120/265	Environmental sciences	2020
J2	Sensors ISSN: 1424-8220	3,275	Q1	15/64	Instruments & instrumentation	2020
J3	Information and Software Technology ISSN: 0950-5849	2,726	Q2	28/108	Comp. science, soft. engineering	2021

A continuación se detallan los artículos que han sido publicados en las revistas previamente mencionadas y el capítulo donde se reproduce de forma íntegra:

J1 - Improving Sustainability of Smart Cities through Visualization Techniques for Big Data from IoT Devices *(Disponible en el capítulo 6)*

J2 - Fostering Sustainability through Visualization Techniques for Real-Time IoT Data: A Case Study based on Gas Turbines for Electricity Production *(Disponible en el capítulo 7)*

J3 - A Methodology to Automatically Translate User Requirements Into Visualizations: Experimental Validation *(Disponible en el capítulo 8)*

2.2. Visibilidad

Con el objetivo de recoger las publicaciones y aumentar la repercusión de la investigación, se crearon perfiles académicos en diversas plataformas. El cuadro 2.3 recoge las direcciones de cada uno de los perfiles de la autora de la presente tesis doctoral.

La estrategia para difundir y aumentar la visibilidad del trabajo publicado ha aumentado el impacto de la investigación. Se ha obtenido un total de 29 citas en Google Scholar y 15 citas en Scopus, a 15 de abril de 2021.

Cuadro 2.3: Perfiles de Ana Lavalle en bases de datos internacionales

Base de Datos	**URL**
Scopus	`https://www.scopus.com/authid/detail.uri?authorId=57212223210`
ORCID	`https://orcid.org/0000-0002-8399-4666`
dblp	`https://dblp.org/pid/251/6230`
Google Scholar	`https://scholar.google.es/citations?user=stUrU8cAAAAJ&hl=es&oi=ao`

2.3. Proyectos de investigación relacionados

Asimismo, las publicaciones desarrolladas en la presente tesis doctoral se han elaborado en el marco de varios proyectos de investigación que se detallan en el cuadro 2.4.

El proyecto principal que proporciona el marco donde se ha desarrollado la presente tesis doctoral es el proyecto ECLIPSE-UA (RTI2018-094283-B-C32), financiado por el Ministerio de Ciencia, Innovación y Universidades. Este proyecto tiene como objetivo principal mejorar la calidad y la

Cuadro 2.4: Proyectos que sustentan esta tesis doctoral

Acrónimo	Título	Referencia	Entidad
ECLIPSE-UA	Enhancing data quality and security for improving business processes and strategic decisions in cyber physical systems	RTI2018-094283-B-C32	Ministerio de Ciencia, Innovación y Universidades
DQIoT	Development of the framework of data quality management for vitalization of IoT products - a case of gas turbine	INNO-20171060	Centro para el Desarrollo Tecnológico Industrial (CDTI), Ministerio de Economía, Industria y Competitividad
Aether-UA	A smart data holistic approach for context-aware data analytics: smarter machine learning for business modelling and analytics	-	Ministerio de Ciencia, Innovación y Universidades
Covid-IA	Big data e inteligencia artificial para mejorar el diagnóstico de los afectados por la Covid-19	DECRETO 180/2020 de la GVA	Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital de la Generalitat Valenciana

seguridad de los datos para optimizar procesos comerciales y decisiones estratégicas en sistemas ciberfísicos.

El proyecto DQIoT (INNO-20171060), financiado por el Centro para el Desarrollo Tecnológico Industrial (Ministerio de Economía, Industria y Competitividad), tiene como objetivo el desarrollo de un marco de gestión de la calidad de datos para la mejora de productos de IoT, en el que colaboró el fabricante Siemens. Esta convocatoria Innoglobal financia proyectos de acciones bilaterales entre entidades de Corea del Sur y España que hayan superado superado la convocatoria de proyectos Europeos EUREKA. En este caso concreto, el proyecto EUREKA es E!11737DQIOT.

El proyecto Aether-UA, financiado por el Ministerio de Ciencia, Inno-

vación y Universidades tiene como objetivo el desarrollo de un enfoque de analítica de datos para modelos de análisis de negocio mediante aprendizaje automático. Un aspecto crucial de este proyecto es la visualización de fuentes de *Big Data* y, la evolución de las investigaciones y tecnología desarrollada en el seno de la presente tesis doctoral son el punto de partida de este proyecto.

Por último, el proyecto Covid-IA, concedido por la Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital de la Generalitat Valenciana (GVA) (DECRETO 180/2020 de la GVA) tiene como objetivo el desarollo de una propuesta que aplique *Big Data* e inteligencia artificial para mejorar el diagnóstico de los afectados por la Covid-19. De nuevo, una parte crucial en este proyecto es desarrollar cuadros de mando y visualizaciones para que el personal sanitario pueda interpretar los datos de forma correcta y mejorar su toma de decisiones.

A modo de resumen, esta tesis doctoral se alinea con varios de los objetivos de estos proyectos, ya que se ha creado una metodología para el análisis visual de *Big Data* que mejora procesos comerciales y sistemas ciberfísicos. Se ha creado una aproximación para la mejora de sistemas de producción mediante el análisis de datos de *IoT*, y la metodología propuesta sirve para abordar problemas como el diagnóstico de los afectados por la Covid-19.

Parte II

Trabajos publicados

Capítulo 3

Compendio

El cuadro 3.1 recoge los artículos seleccionados para la solicitud de presentación de la tesis doctoral por compendio de publicaciones, aprobada por la directora de la Escuela de Doctorado de la Universidad de Alicante a 4 de marzo de 2021.

La primera columna detalla el identificador de la publicación; la segunda columna, el título de la publicación; la tercera columna, el impacto del congreso/revista donde fue publicado; finalmente, la última columna detalla el capítulo donde se reproduce una copia íntegra y literal del artículo.

Cuadro 3.1: Publicaciones que conforman el compendio

Id	Título	Impacto	Ubicación
C1	Visualization Requirements for Business Intelligence Analytics: A Goal-Based, Iterative Framework	Clase 2 Core A	Capítulo 4
J1	Improving Sustainability of Smart Cities through Visualization Techniques for Big Data from IoT Devices	Q2	Capítulo 6
J2	Fostering Sustainability through Visualization Techniques for Real-Time IoT Data: A Case Study Based on Gas Turbines for Electricity Production	Q1	Capítulo 7
C3	An Approach to Automatically Detect and Visualize Bias in Data Analytics	Clase 2 Core B	Capítulo 9

Capítulo 4

Visualization Requirements for Business Intelligence Analytics: A Goal-Based, Iterative Framework

Lavalle, A., Maté, A., Trujillo, J., & Rizzi, S. (2019, September). Visualization Requirements for Business Intelligence Analytics: A Goal-Based, Iterative Framework. In *2019 IEEE 27th International Requirements Engineering Conference (RE 2019)* (pp. 109-119). IEEE.

Conference Rating por GII-GRIN-SCIE: **Clase 2**
Conference Rating por CORE: **A**

Disponible en:
DOI: https://doi.org/10.1109/RE.2019.00022

Visualization Requirements for Business Intelligence Analytics: A Goal-Based, Iterative Framework

Ana Lavalle, Alejandro Maté, Juan Trujillo
Lucentia Research, DLSI, University of Alicante
Carretera San Vicente del Raspeig s/n, 03690
San Vicente del Raspeig, Alicante, Spain
alavalle@dlsi.ua.es, amate@dlsi.ua.es, jtrujillo@dlsi.ua.es

Stefano Rizzi
DISI, University of Bologna
V.le Risorgimento 2, 40136, Bologna, Italy
stefano.rizzi@unibo.it

***Abstract*—Information visualization plays a key role in business intelligence analytics. With ever larger amounts of data that need to be interpreted, using the right visualizations is crucial in order to understand the underlying patterns and results obtained by analysis algorithms. Despite its importance, defining the right visualization is still a challenging task. Business users are rarely experts in information visualization, and they may not exactly know the most adequate visualization tools or patterns for their goals. Consequently, misinterpreted graphs and wrong results can be obtained, leading to missed opportunities and significant losses for companies. The main problem underneath is a lack of tools and methodologies that allow non-expert users to define their visualization and data analysis goals in business terms. In order to tackle this problem, we present an iterative goal-oriented approach based on the i* language for the automatic derivation of data visualizations. Our approach links non-expert user requirements to the data to be analyzed, choosing the most suited visualization techniques in a semi-automatic way. The great advantage of our proposal is that we provide non-expert users with the best suited visualizations according to their information needs and their data with little effort and without requiring expertise in information visualization.**

***Index Terms*—Data Visualization, Data Analysis, Model-driven development, Requirements engineering**

I. Introduction

Data visualization plays a key role in business intelligence analytics. With ever larger amounts of data that need be interpreted, finding effective visualizations is key to understanding the underlying patterns and the results obtained by analysis algorithms. Without this understanding, users are more likely to distrust the results, following their gut feeling instead of making well-informed decisions. Indeed, according to a survey by Salesforce [20], 73% of high performers strongly agree that analytic tools are valuable for gaining strategic insights from the data. A large number of companies and researchers are very interested in its application.

Despite this interest, finding the right visualization is still a challenging task. Business users are rarely expert in data visualization, and they may not exactly know what type of information they want to extract from data or which would be the best visualization type. Consequently, misinterpreted graphs and wrong results can be obtained, leading to missed opportunities and significant losses for companies. Another relevant point to be considered is related to dashboard design. A dashboard is a visualization tool that groups multiple tables and charts, ideally aiming to provide a 360° view of the phenomenon being analyzed. Dashboards play a key role in the analysis and visualization of data because they enable users –even those with limited ICT skills– to get their insights and make informed decisions. Although predefined dashboards have been designed for specific sectors, each business and each user may have particular needs different from those already included in predefined dashboards. To design a dashboard, users should state their goals and precisely delimit the information to be represented. However, in most cases, users do not have a clear idea of the most effective visualization techniques for each piece of data.

Although some studies have proposed models to automatically generate dashboards (e.g., [23], [21], [13]) they do not consider the best visualization types and tools for each situation. In this direction, some approaches have been proposed to automate data visualization from user requirements (e.g., [4], [14], [17], [10], [9]). However, these approaches do not *guide* users in the discovery of their objectives nor in the definition of the necessary requirements to generate the most appropriate visualizations for each situation; indeed, they still require users to explicitly state what they wish to visualize and, most importantly, how exactly they want to visualize it.

A very recent approach for the derivation of visualization requirements in analytics is SkyViz [10]. In SkyViz, first the user specifies her visualization objectives and describes the dataset to be visualized by defining a *visualization context* based on seven prioritizable visualization requirements. Then, this visualization context is automatically translated into a set of most-suitable visualization types (e.g., pie chart and bar chart) via a skyline-based technique.

As recognized in [10], defining a visualization context from scratch may indeed be a challenge for non-expert users. So, in this paper, we complement SkyViz by defining a goal-based [6], [8], [15] modeling approach and a set of guidelines to capture user goals and derive the corresponding visualization contexts. Our proposal is meant to work on top of SkyViz making it better usable by non-expert users. Specifically, it improves SkyViz from several points of view: (i) it provides a sequence of steps and guidelines to help users define their goals and achieve them by using the available data sources;

2332-6441/19/$31.00 ©2019 IEEE
DOI 10.1109/RE.2019.00022

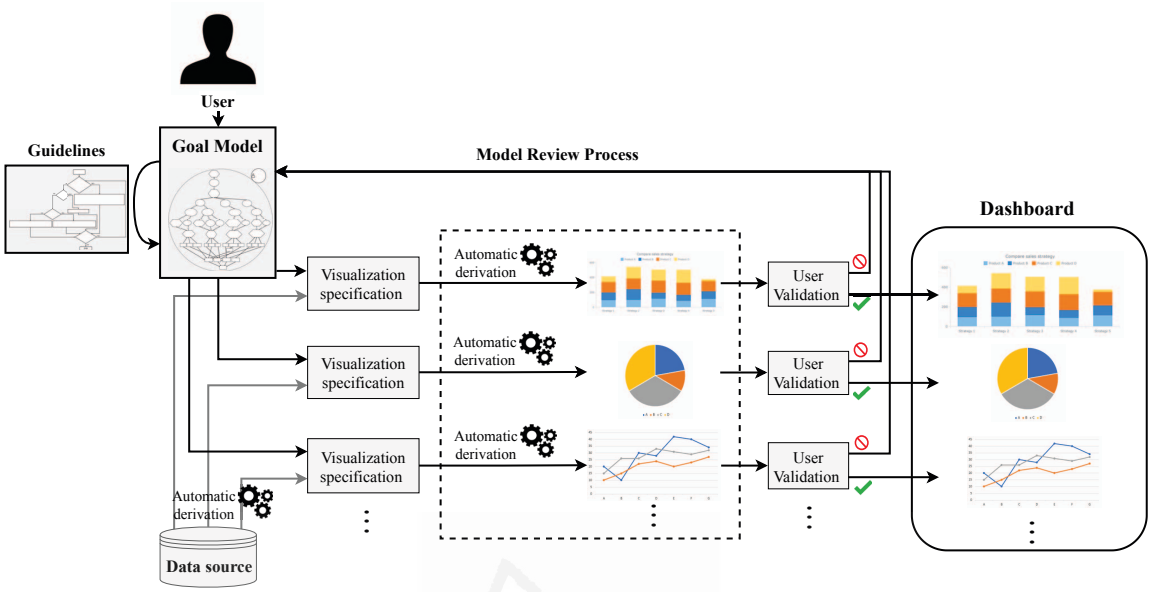


Fig. 1. Overall view of the process proposed

(ii) it translates the user's goals into a visualization context; (iii) it semi-automatically extracts visualization requirements from the data sources to be analyzed; and (iv) it provides a rationale for dashboard design. In this way, business users can stay focused on their analysis goals and they can eventually obtain, with a limited effort, the visualizations that best suit their needs. Besides, these visualizations will be grouped into dashboards to allow users to effectively monitor and measure their goals.

Fig. 1 summarizes the process followed in our proposal. In current practice the user is accompanied by a data analyst to help her to follow the process. We are continuing the work on [2] where we propose a visualization model for representing visualization details regardless of their implementation technology with the aim to develop a web tool and let users follow the process on its own. In this case, firstly, a sequence of questions guides the user in creating a Goal-Based model that captures her needs. This model encompasses all the visualizations required to tackle the user's objectives. Then, this Goal-Based model is completed by analyzing the features of the data sources to be visualized. At this stage, the model is translated into a a set of visualization contexts, which are then handed to SkyViz to find the best visualization types (process represented within the dashed lines). Finally, each generated visualization is validated by the user to verify if it fulfills the essential requirements for which it was created. The validation process is performed through a questionnaire that is automatically generated from the Goal-Based model, asking the user if the visualization obtained does contribute to answer her goals. Each visualization validated is added to the dashboard. An unsuccessful validation points out to the existence of missing or wrongly-defined requirements that must be reviewed; in this case, a new cycle is started by reviewing the existing model to identify which aspects were not taken into account, generating in turn an updated model. This process is repeated until all user requirements are fulfilled.

The rest of the paper is structured as follows. Section 2 presents the related work in this area. Section 3 presents our Goal-Based modeling approach for data visualization. Section 4 describes the implementation of our approach. Section 5 discusses an illustrative example in the fiscal domain. Section 6 presents limitations and validity threats of our approach. Finally, Section 7 summarizes the conclusions and our future work.

II. Related Work

Several works are focused on finding ways to automatically generate visualizations or dashboards. In [23], the authors propose an automatic dashboard generator with the capacity to alter dashboard design and functionality without requiring significant development time. In [21], a technique is proposed that allows users to modify or add new visualizations as desired, including filters in real time. In [13], a users-and-roles model is introduced, enabling the automatic generation of user-specific monitoring dashboards, properly displaying the information needed by each user in an organization. All these approaches require that the final user chooses the type of visualization for the representation of the data, without trying to determine which is the most adequate one for the current

context. Clearly, this requires the user to be an expert or at least knowledgeable in data visualization techniques.

In order to tackle this problem, some works have proposed different ways to find the best visualization for each analysis. [4] surveys the main classifications proposed in the literature and integrates them into a single framework based on six visualization requirements. In [14], authors propose a framework for choosing the best visualization where the main types of charts are related to users goals and to the data dimensionality, cardinality, and the type they support. Finally, [17] proposes a more detailed classification of data types and relates each common type of chart to the users goals it is most compliant with.

In [10] the authors propose SkyViz, an approach to automate the translation of a structured visualization context specified by the user into a suitable visualization. A visualization context consists of seven coordinates, namely goal, interaction, user skills, dimensionality, cardinality, type of the independent variables, and type of the dependent variables. Furthermore, in [9] a novel utility function and a suite of search schemes for recommending top-k aggregate data visualizations is presented. The utility function recognizes the impact of numerical dimensions on visualization, which is captured by means of multiple objectives, namely, deviation, accuracy, and usability.

Other works are focused on additional issues related to visualization. In [11] it is argued that one of the reasons for the lack of advanced visualizations are users, who do not often know how they may represent their data. In [7] the authors propose a classification of causes of pitfalls, the designer or the user, and they list three types of (negative) effects: *cognitive*, *emotional*, and *social*. More specifically, they state that the cause of a visualization problem can be twofold: the *encoding* (that is, caused by the designer/developer) or the *decoding* (that is, caused by the reader/user). In the latter case, the user who reads the visualization makes a mistake in the interpretation.

Other works [16] also point out that the rendering process introduces uncertainty in all three areas: from the *data collection process*, *algorithmic errors*, and *computational accuracy and precision*. In addition, others like [12] have started thinking about visual representations of errors and uncertainties; possible sources of uncertainty are *acquisition* (instrument measurement error, numerical analysis error, statistical variation), *model* (both mathematical and geometric), *transformation* (errors introduced from resampling, filtering, quantization, and rescaling), and *visualization*.

While certainly adding value to visualizations, these researches focus on the potential pitfalls of blindly using visualization methods without fully understanding the limitations and assumptions of each method and the rationale behind visualizations. In this sense, visualizations should consider the evolving needs of users, taking into account high-level semantics, reasoning about unstructured and structured data, and providing a simplified access and better understanding of data [3]. As such, although often overlooked when designing visualizations, requirement modelling is an important activity [19], that compensates the little or no attention often paid to (explicitly) representing the reasons, i.e., the *why*, in terms of motivations, rationale, goals, and requirements. This is specially true for goal-based modeling approaches, where the motivations become first-class citizens in the models.

It is very important that users understand what they are visualizing and why this visualization contributes to reach a goal. Visualizations must be precise and understandable to users to minimize the interpretation mistakes made by both users and designers. In this sense, [1] shows how IBM Watson Analytics can be used to visualize and analyze data derived from goal-based conceptual models of regulations and regulatory initiatives.

To sum up, none of the approaches summarized above provides a methodology that guides non-expert users in specifying the most adequate set of visualizations and facilitates their implementation into dashboards to be used for data analysis. To bridge the gap between user needs and visualization, goal-based modeling approaches —which we apply this paper— emerge as a natural solution.

III. A Goal-Based Modeling Approach for Visualization Requirements

Eliciting from users visualization requirements is considered to be a challenging task, which we aim at supporting in this paper. To this end, we use a combination of modeling and automatic derivation. Initially, we create a Goal-Based model that guides non-expert users towards the specific visualizations they need according to their data analysis objectives. Then, starting from this model we semi-automatically derive a visualization context to be fed into SkyViz and it will recommend us the most suitable visualization type.

The approach we take to formally define our model is through a metamodel that follows the specification given in [10] in terms of the coordinates required to build a visualization context (*Goal*, *Interaction*, *User*, *Dimensionality*, *Cardinality*, *Independent Type*, and *Dependent Type*) and the values these coordinates can take. Our metamodel is shown in Fig. 2 and is an extension of the one used for social and business intelligence modeling, namely i* in its 2.0 version [8] and its *i* for Data Warehouses* extension [15]. Existing elements in the i* core are represented in cyan , whereas those included in i* for Data Warehouses are represented in red. The new concepts added by our proposal are represented in light green and yellow. In the following, we describe the concepts included in the metamodel by following the process required for its application.

The aim of the proposed metamodel is to support users in better understanding their objectives and in determining which visualization type they need. To this end, the first element is the *VisualizationActor*, which models the user of the system. There are two types of Visualization Actors: Lay, if she has no knowledge of complex visualizations, and Tech, if she has previous experience and is accustomed to business intelligence analytics.

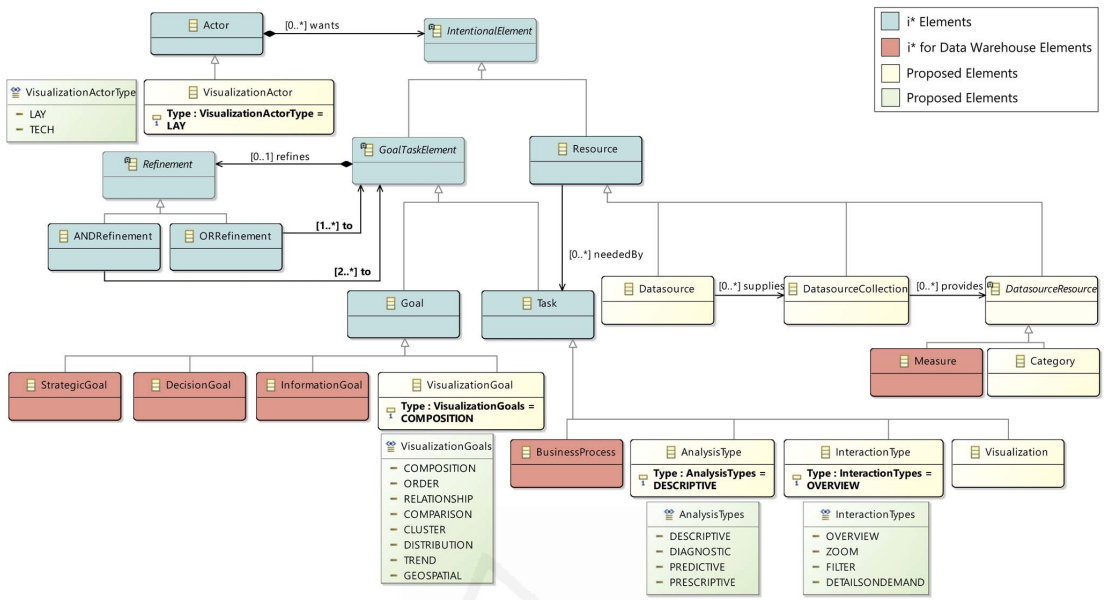


Fig. 2. Visualization Specification Metamodel

Once the actor has been defined, the next elements to be defined are the *BusinessProcesses* on which users will focus their analysis. The business process will serve as the guideline for the definition of *Goals*. A goal represents a desired state of affairs with reference to the business process at hand. Goals can be divided into *Strategic*, *Decision*, *Information*, and *Visualization*.

The top-level goals are *StrategicGoals*. They are the main objectives of the business process and are meant as changes from a current situation into a better one. Strategic goals are achieved by means of analyses that support the decision-making process.

The *AnalysisType* allows users to express which kind of analysis they wish to perform. The definition of a type of analysis will also give the advantage of determining the visualizations to be grouped in the same dashboard. The type of analysis can be determined by selecting which question from the following ones is to be answered [22]:

- **Prescriptive**: How to act?
- **Diagnostic**: Why has this happened?
- **Predictive**: What is going to happen?
- **Descriptive**: What to do to make it happen?

Once the types of analysis to be performed over the strategic goals have been defined, the next elements are *DecisionGoals* and *InformationGoals*. A *DecisionGoal* aims to take appropriate actions to fulfill a strategic goal and explains how it can be achieved. *DecisionGoals* communicate the rationale followed by the decision-making process; however, by themselves they do not provide the necessary details about the data to be visualized. Therefore, for each decision goal there are one or more *InformationGoals*, i.e., lower-level abstraction goals representing the information to be analyzed.

For each *InformationGoals* there will be one *Visualization*. *Visualization* is defined as a task because we understand it as the visualization process, not as the visualization representation. A *Visualization* is characterized by one or more *VisualizationGoals* which describe which aspects of the data the visualization is trying to reflect, and one or more kinds of *InteractionType* that users will need to have with the visualization. *VisualizationGoal* can be defined as Composition, Order, Relationship, Comparison, Cluster, Distribution, Trend, or Geospatial, while *InteractionType* can be Overview, Zoom, Filter, or Details-on-demand [10]. Moreover, a *Visualization* will make use of one or more *DatasourceResource* elements to get the relevant data from the data source.

In the following subsection, we describe in detail the visualization specification process we propose.

A. Visualization Specification

As argued in [10], inexperienced users may find it difficult to properly give values to the seven coordinates included in a visualization context. To facilitate their task, we observe that the coordinates can be split into two families (Fig. 3): *user-related* (namely, *Goal*, *Interaction*, *User*) and *data-related* (namely, *Dimensionality*, *Cardinality*, *Independent Type*, and *Dependent Type*). In current practice the user is accompanied by a data analyst to help her to follow the process. We are continuing the work on [2] where we propose a visualization model for representing visualization details regardless of their

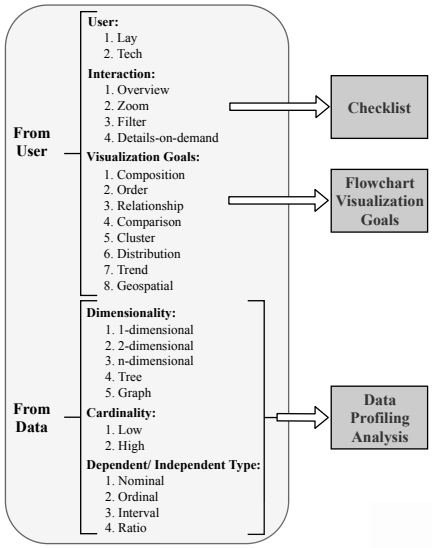


Fig. 3. Specification of Visualization context

implementation technology with the aim to develop a web tool and let users follow the process on its own.

Therefore, the first step we take concerns user-related coordinates, and consists in guiding users to specify what *Visualization Goals* they aim to achieve and which kind of *Interaction* they would like to have.

- **Interaction:**
  The possible interactions are *Overview* (gain an overview of the entire data collection), *Zoom* (focus on items of interest), *Filter* (quickly focus on interesting items by eliminating unwanted items), and *Details-on-demand* (select an item and get its details). We show a checklist to the user (Fig. 4) from which she can choose one or more types of interaction.

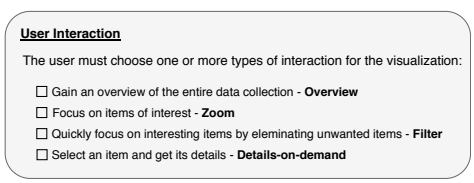


Fig. 4. Interaction with the visualization

- **Visualization Goals:**
  A visualization goal can be *Composition*, *Order*, *Relationship*, *Comparison*, *Cluster*, *Distribution*, *Trend*, or *Geospatial*. Since choosing the right goal can be difficult depending on the context, to aid users in finding which visualization goal they are pursuing we use the flowchart in Fig. 5, which contains a series of Yes/No questions to be answered by users. The flowchart provides an easy way to discern which visualization goals should be included for each visualization, thus simplifying the task for non-expert users.

As to data-related coordinates, we semi-automatically extract their values by analyzing the features of the data sources. In this way, users do not need to manually inspect the data or have a deep understanding of their characteristics to obtain the most adequate visualizations, and we avoid the introduction of errors in the process.

- **Data Profiling Analysis:**
  In addition to the requirements provided by users, we extract the values of the remaining coordinates by analyzing the features of the data sources. Users need only to provide the source dataset; then, **Dimensionality**, **Cardinality**, and **Dependent/INdependent Type** will be extracted as explained below.
  First, users specify a connection to the source dataset they wish to visualize. A menu is provided where users can choose if they want to know the Data type, Cardinality or Dimensionality of the selected column. Finally, this software returns the information requested by users. This development has been created to collect information about the data in a simple way for users. To know how to delimit the values for each coordinate we have followed the proposed in [10]. In this way we classify the **Dimensionality**, **Cardinality**, and **Dependent/Independent Type** as follows:
  - **Dependent/Independent Type** is used to declare the type of each variable. It can be *Nominal* when it is qualitative and each variable is assigned to one category, *Ordinal* when it is qualitative and categories can be sorted, *Interval* when it is quantitative and equality of intervals can be determined, or *Ratio* when it is quantitative with a unique and non-arbitrary zero point.
    We delimited each category as follows: If the value is a number, we determine *Ratio* if is a numeric with a unique and non-arbitrary zero point or *Interval* if is a numeric with under 0 values. In the cases where the value is a string of characters, the program shows a grouped list of the values. Then the user is available to determine if in the list there is an order, then it would be *Ordinal*, and if the user can not determine an order it would be *Nominal*.
  - **Cardinality** represents the cardinality of the data, and it can be defined as *Low* or *High* depending of the numbers of items to represent. It will be *Low* cardinality from a few items to a few dozens items and *High* cardinality if there are some dozens items or more. Some visualization types support a larger number of items than others (for example, a pie chart can only visualize low-cardinality data, while a heat map is also fit for high-cardinality data).
  - **Dimensionality** is used to declare the number of variables to be visualized. Specifically, it can be *1-*

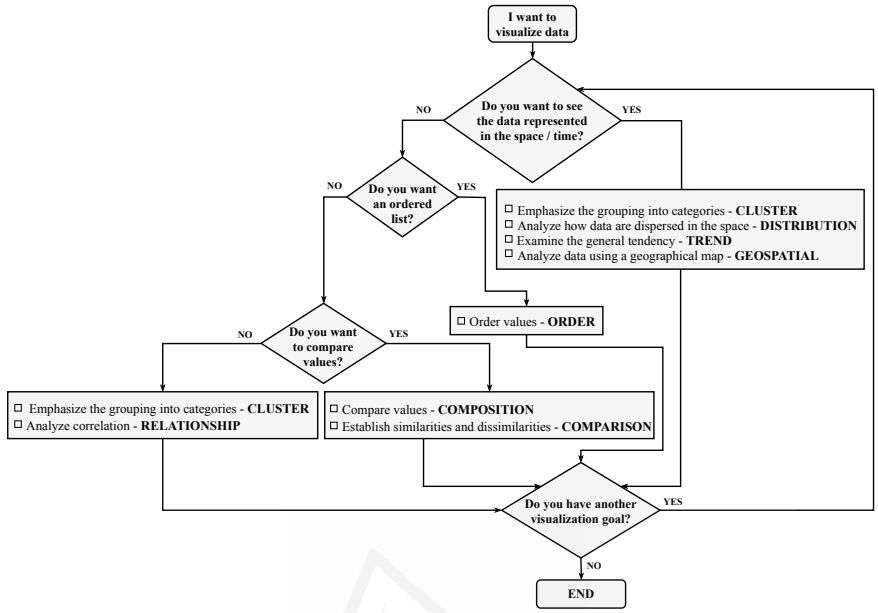


Fig. 5. Guidelines expressed as a flowchart to help non-expert users in defining visualization goals

dimensional when the data to represent is a single numerical value or string, *2-dimensional* when one variable depends on other, *n-dimensional* when a data object is a point in an n-dimensional space, *Tree* when a collection of items have a link to one other parent item, or *Graph* when a collection of items are linked to arbitrary number of other items.

Once all the requirements have been gathered, we can use SkyViz to get the best type of visualization suited for each particular case while taking into account the preferences of users. However, to check if the visualization generated really fulfills the essential requirements for which it was created, a questionnaire is submitted to the user. The questionnaire will be generated automatically from the information specified by users in the model. Specifically, users will be asked if the visualization contributes to answering the *InformationGoal* defined in the model. If the visualization passes the validation, it will be added to the dashboard. Conversely, if it does not pass the validation, a review of the model will be done to know what aspects were not taken into account and thus generate an updated model. This review gives users an assisted path to improve the obtained visualizations and helps them to achieve their goals.

IV. Implementation

The implementation of our approach relies on four integrated components as Fig. 6 shows: (i) the CASE tool aimed at creating the model through the definition of a metamodel, represented as "Visualization Requirements Modeler"; (ii) "Data Analyzer" component that semi-automatically extracts the dataset features, through queries using the (iii) "Data Source Connector"; (iv) the "Visualization Generator", component that selects and renders the best visualization following the process described in [10]. These four components are integrated into our system. The system extracts information from users and data sources and it realizes a communication between the components to generate a visualization.

The CASE tool is implemented in Eclipse by using the Ecore metamodel as a baseline. Defining our metamodel in Ecore enables the automatic generation of the diagram editors

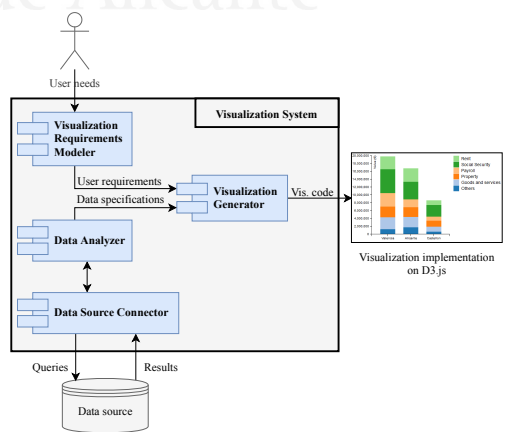


Fig. 6. System architecture

for models. Using the Ecore framework we are able to generate the java class objects that support the creation of requirements models.

The Data Analyzer software created to extract the data profiling has been implemented in Java. It allows users to specify the data source where they need to extract information and performs in an automated and guided way the extraction of information. The MySQL relational database has been used to make the connection, but other types of data sources can be connected as well. In order to use another type of data source we just need to replace the Data Source Connector.

Fig. 7 shows an example of the interactive version of the code that is executed to extract information from the data source. This code will be connected to the user-defined model allowing users to automatically obtains the requested data information.

```
What do you want to know about amount?
1 - Data type
2 - Cardinality
3 - Dimensionality
1
amount --> RATIO
```

Fig. 7. Data profiling analysis example

Part of the visualization requirements represented in the case tool are elicited from users by following our model, the rest comes from the analysis of the data sources. Once we have all these requirements defined, using SkyViz the visualization context is automatically translated into the most-suitable visualization type. Then, the visual requirements are translated into a call to the D3 JavaScript library [5] which renders the visualization. In the cases where a map has to be rendered, the Plotly library [18] is used, it can be developed on JavaScript, Python or R. Fig. 8 shows the final result of the process using D3.js. In the next section, we apply our approach to an illustrative example in the field of tax collection.

V. Illustrative Example

In order to evaluate the validity of our approach we have applied it to an illustrative example. In this case, a tax collection organization has been selected to evaluate the validity of our approach. A tax collection organization requires a set of visualizations to analyze their data in order to help them detect underlying patterns in their unpaid bills and tax collection

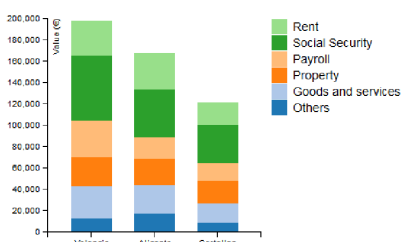


Fig. 8. Visualization rendered in D3.js

distribution. Due to the sensibility of their data, we are not allowed to show the real values; besides, data had to be anonymized.

Our approach has been applied to the tax collection organization, producing the model shown in Fig. 9. In this model, the company wants to analyze the unpaid debts. Therefore, the analysis will focus on the *"Tax collection"* business process. Defining a business process helps determining which concepts are involved in the analysis and what kind of goals are pursued. Here the user is a tax collector who is not a specialist in analytics but rather an expert in tax management, thus she is defined as *"Lay user"*.

A. Specifying Goals

The main objectives of the business process are defined as shown in Fig. 9. Specifically, the user defined her strategic goal as *"Reduce the unpaid bills"*. Strategic goals are achieved by means of analyses that support the decision-making process. The analysis type allows users to express what kind of analysis they wish to perform. In this case, the user wishes to know why bills are unpaid. Thus, she decides to perform a *"Diagnostic analysis"*. Having defined a specific type of analysis, we are aware that all context information for *"Diagnostic analysis"* should be gathered in the same dashboard in order to provide a complete answer to the user.

The diagnostic analysis is decomposed into decision goals. A decision goal aims to take appropriate actions to fulfill a strategic goal and explains how it can be achieved. The user defined her decisions goals as: *"Identify unpaid bills"* *"Identify the quantities unpaid"*, and *"Evolution of unpaid bills"*. Decisions goals communicate the rationale followed by the decision-making process; however, by themselves they do not provide the necessary details about the data to be visualized. Therefore, for each decision goal we specify one or more information goals, i.e., lower-level abstraction goals representing the information to be analyzed.

From each of the decision goals she listed, the user refined the following information goals: *"Analyze where are the places with more unpaid bills"*, *"Analyze the type of unpaid bills"*, *"Analyze who has unpaid bills"*, and *"Analyze the evolution of unpaid bills by province"*. Information goals represent the lowest level of goal abstraction.

At this point, the user has the necessary information about her goals to start defining the visualization context. For each information goal, we will have one visualization to achieve it. A visualization is characterized by one or more visualization goals which describe what aspects of the data the visualization is trying to reflect, and one or more kinds of interaction that they will like to have with the visualization. Moreover, a visualization will make use of one or more data source elements to get the relevant data from the database.

In this case, the user defines the interactions she wants to have with each visualization following the checkbox shown in Fig. 4. *"Overview"*, *"Zoom"* and *"Details-on-demand"* have been defined. Additionally, following the flowchart shown in

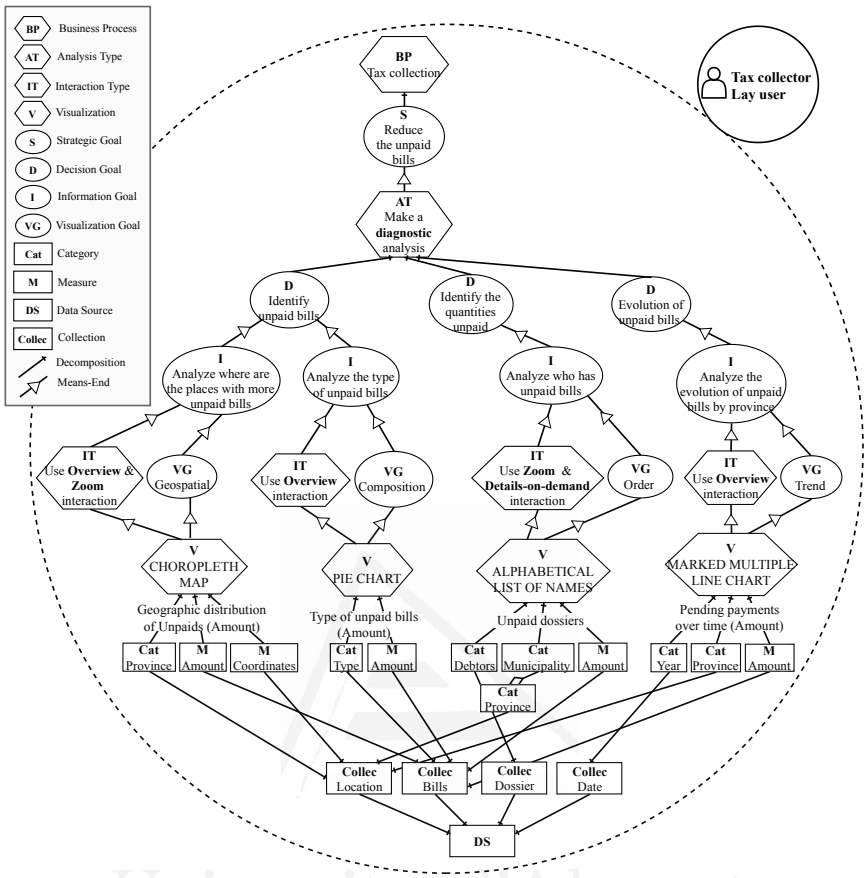


Fig. 9. Application of our metamodel to the illustrative example

Fig. 5, the user specified her visualization goals: *"Geospatial"*, *"Composition"*, *"Order"*, and *"Trend"*.

Finally, the visualizations are decomposed into Categories and Measures that will populate them. In this case, the visualization of *"Geographic distribution of unpaids"* includes *"Province"* as category, and *"Amount"* and *"Coordinates"* as measures. These attributes come from the data source collections *"Location"* and *"Bills"*, respectively. For the visualization of *"Type of unpaid bills"* the user picked *"Type"* as relevant category, and *"Amount"* as measure. These are obtained from the data source collection *"Bills"*. Next, in the case of *"Unpaid dossiers"*, categories *"Debtors"*, *"Municipality"*, and *"Province"* as well as measure *"Amount"* are selected from the data source collections *"Dossier"*, *"Location"* and *"Bills"*. The last visualization is *"Pending payments over time"*, that makes use of categories *"Year"* and *"Province"* and of measure *"Amount"*. These data are obtained from the collections *"Date"*, *"Location"*, and *"Bills"*.

Once user have defined the data sources and collections from where the data will be extracted, it is possible to profile data sources to determine Dimensionality, Cardinality and Dependent/Independent Type.

B. Profiling Data Sources

Tax data are divided into different collections as follows: **Location** collects information about where tax was unpaid; **Date** holds data about when it was unpaid; **Dossier** represents who is the debtor, whether a person or an entity; and **Bill** joins the set of previously mentioned data by means of bills. Each collection is futter decomposed into measures and categories.

Next step is to analyze the data sources in order to extract information about information about Dimensionality, Cardinality, Dependent/Independent Type from the data sources, using our Data Analyzer tool as shown in Section 3.

We focus on the **"Type of unpaid bills"** visualization from our Goal-Based model, which requires information about the category *"Type"* and measure *"Amount"*. Firstly, using the data profiling tool, the independent variable *"Type"* are classified as **Nominal** and the dependent variable *"Amount"*

as **Ratio**. Dimensionality is set to **2-dimensional**, because the user has defined 2 variables to visualize. Finally, the tool computes the Cardinality of data through a query. The tool defines Cardinality as **Low** because the data contains a few items to represent, there are 6 types of bills. Overall, the values obtained through data profiling are:

- **Dimensionality**: 2-dimensional
- **Cardinality**: Low
- **Independent Type**: Nominal
- **Dependent Type**: Ratio

C. Validation and Results

Once the visualization has been specified, the visualization context is used as input to be fed into SkyViz and it will recommend us the most suitable visualization type by a table with suitability scores.

The suitability scores from Table 1 are proposed by [10] and shows the visualization context we derived from the information goal *"Analyze the type of unpaid bills"*, together with the suitability scores for tree visualization types, namely, *"Stacked Column Chart"*, *"Bubble Chart"*, and *"Pie Chart"*. The semantics of the suitability scores in this context is as follows:

- **Fit**: Means that the visualization type is fully compatible.
- **Acceptable**: Means that the visualization type is compatible with the coordinate value, though it may fail to emphasize some of the required features.

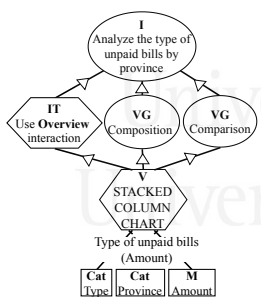


Fig. 10. Model update due the review

- **Discouraged**: Means that the visualization type can be used in principle for the coordinate value, but it may distort the very nature of the required features.
- **Unfit**: Means that the visualization type should not be used for the coordinate value.

Accordingly the suitability scores in Table 1, the most suitable visualization for our analysis is *"Pie Chart"*. A mockup of the pie chart visualization is shown to the user and the user detects that this visualization does not reach exactly her goals. Consequently, a model review is done by the user and she detects that the information goal *"Analyze the type of unpaid bills"* is not correctly defined, she adds information modifying it to *"Analyze the type of unpaid bills"*. She continues reviewing the model and she extends the visualization goals by adding *"Comparison"*. Finally, she select from the collection *"Location"* the category *"Province"* to be represented in the visualization. This review modifies the model updating it as shown in Fig. 10. The goals have change to *"Composition"* and *"Comparison"* and dimensionality now is *"n-dimensional"*.

Now, with the update of the context, the most suitable visualization for our visualization context has changed to *"Stacked Column Chart"*. Again a mockup of the visualization is shown to the user and now the visualization can effectively answer the user information goal *"Anayze the type of unpaid bills by province"*. Therefore, following the derivation process we make a call to the D3 JavaScript library, obtaining the visualization. Consequently, the visualization is added to the dashboard, as shown in the lower-left corner of Fig. 11.

This visualization, combined with those generated from the informational goals *"Analyze where are the places with more unpaid bills"*, *"Analyze who has unpaid bills"* and *"Analyze the evolution of unpaid bills by provinces"*, are grouped into the dashboard layout proposed in Fig. 11, aimed at satisfying the analytic requirements of our tax collector user with the most adequate visualizations.

VI. Limitations and Validity Threats

In this section, we summarize the main limitations we envision for our approach.

- Up to now, we have had satisfactory results in our applications to real cases and testing the proposal with a focus group. However, since we have not yet tested the

TABLE I
SUITABILITY SCORES NEEDED FOR THE TWO VISUALIZATION CONTEXT OF THE ILLUSTRATIVE EXAMPLE.

	VISUALIZATION CONTEXT	**Stacked Column Chart**	**Bubble Chart**	**Pie Chart**
Goal:	Composition	fit	unfit	fit
	Comparison	fit	fit	unfit
Interaction:	Overview	acceptable	acceptable	fit
User:	Lay	fit	acceptable	fit
Dimensionality:	2-dimensional	unfit	unfit	fit
	n-dimensional	fit	fit	unfit
Cardinality:	Low	fit	acceptable	fit
Independent Type:	Nominal	fit	unfit	fit
Dependent Type:	Ratio	fit	fit	fit

proposal in a comprehensive set of contexts, it may be the case that some specific user profiles have not yet been identified.

- In principle, our proposal is context-independent; up to now, it has been applied in the economic, educational, and gas turbine contexts. However, some other context may raise specific issues that we have not contemplated yet. For instance, some contexts may require visualizations to be produced in real time, which is currently out of the scope of our approach.
- In our experiments, we have worked with a data analyst supporting each non-expert user in defining her visualization requirements. We are currently working to conclude the development of the tool proposed in [2] to verify that users are actually qualified to define visualization requirements completely on their own.
- At the time of defining the visualization specification, the user has to know the features of the dataset to be visualized. Besides, she is required to be expert in the application domain for which visualization is required.
- We rely on SkyViz to derive the best suited visualizations type, however SkyViz itself has some limitations [10]. First of all, it currently includes a limited number of visualization types. On the other hand, if a significantly larger number of visualization types were included, the seven coordinates might no longer be sufficient to distinguish them, in which case the user would be provided with a large number of (probably similar) visualization types. To cope with this situation, other coordinates should be added, but the research questions to be addressed would be (i) how to select them in order to actually improve the discriminatory power of SkyViz, and (ii) how to deal with these new coordinates in the goal-based model.
- While our proposal would be able to represent collaborative visualizations through *Strategy Dependency* diagrams, this aspect has not been yet fully explored and is considered out of the scope of this paper.

VII. Conclusions and Future Work

In this paper, we have presented an iterative goal-based modeling approach in order to help non-expert users define their data analysis goals and derive the most adequate visualizations to facilitate the analysis of data. Compared to other approaches, our proposal covers the whole process from the definition and modeling of user requirements to the implementation of the visualizations. The great advantage of our proposal is that non-technical users can effectively communicate their visual analytic needs without needing deep knowledge of visualization technologies or data sources descriptions. Furthermore, visualizations are easily modified by altering requirements such as the type of interaction or the visualization goal pursued.

As part of our future work, we are working on improving the data analysis step to better support the detection of independent and dependent variables when multiple measures and categories are present. Furthermore, we are working on the implementation of a user-friendly diagram editor by using Graphiti. In this way, we will be able to provide better support for users even when there is no analyst available to aid them when building the requirements model. We will also consider capturing non-functional requirements or quality goals to help decide between visualizations.

VIII. Acknowledgment

This work has been co-funded by the ECLIPSE-UA (RTI2018-094283-B-C32) project funded by Spanish Ministry of Science, Innovation, and Universities. Ana Lavalle holds an Industrial PhD Grant (I-PI 03-18) co-funded by the University of Alicante and the Lucentia Lab Spin-off Company.

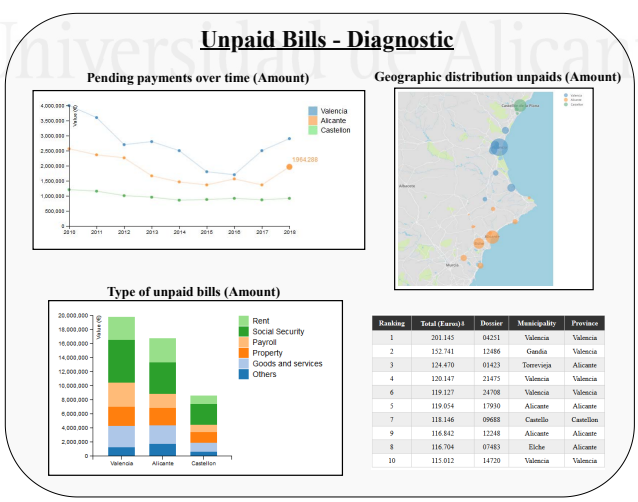


Fig. 11. Dashboard for tax collection analysis

REFERENCES

[1] Akhigbe, O., Heap, S., Amyot, D., Richards, G.: Exploiting IBM watson analytics to visualize and analyze data from goal-based conceptual models. In: Proceedings of the ER Forum 2017 and the ER 2017 Demo Track co-located with the 36th International Conference on Conceptual Modelling. pp. 338–342 (2017)

[2] Ana Lavalle, A.M., Trujillo, J.: Requirements-driven visualizations for big data analytics: a model-driven approach. In: International Conference on Conceptual Modeling ER 2019, to appear. Springer (2019)

[3] Aufaure, M.: What's up in business intelligence? A contextual and knowledge-based perspective. In: Conceptual Modeling - 32th International Conference, ER 2013. pp. 9–18. Springer (2013)

[4] Börner, K.: Atlas of knowledge (2014)

[5] Bostock, M.: Data-driven documents (2019), https://d3js.org/

[6] Bresciani, P., Perini, A., Giorgini, P., Giunchiglia, F., Mylopoulos, J.: Tropos: An agent-oriented software development methodology. Autonomous Agents and Multi-Agent Systems **8**(3), 203–236 (2004)

[7] Bresciani, S., Eppler, M.J.: The pitfalls of visual representations: A review and classification of common errors made while designing and interpreting visualizations. Sage Open **5**(4), 2158244015611451 (2015)

[8] Dalpiaz, F., Franch, X., Horkoff, J.: istar 2.0 language guide. CoRR **abs/1605.07767** (2016)

[9] Ehsan, H., Sharaf, M.A., Chrysanthis, P.K.: Efficient recommendation of aggregate data visualizations. IEEE Trans. Knowl. Data Eng. **30**(2), 263–277 (2018)

[10] Golfarelli, M., Rizzi, S.: A model-driven approach to automate data visualization in big data analytics. Information Visualization, to appear (2019)

[11] Gray, C.C., Teahan, W.J., Perkins, D.: Understanding our analytics: A visualization survey. Journal of Learning Analytics, to appear (2017)

[12] Johnson, C.R., Sanderson, A.R.: A next step: Visualizing errors and uncertainty. IEEE Computer Graphics and Applications **23**(5), 6–10 (2003)

[13] Kintz, M., Kochanowski, M., Koetter, F.: Creating user-specific business process monitoring dashboards with a model-driven approach. In: Proceedings of the 5th International Conference on Model-Driven Engineering and Software Development, MODELSWARD. pp. 353–361 (2017)

[14] Madhu Sudhan, S., Chandra, J.: Iba graph selector algorithm for big data visualization using defence dataset. International Journal of Scientific & Engineering Research **4**(3) (2013)

[15] Maté, A., Trujillo, J., Franch, X.: Adding semantic modules to improve goal-oriented analysis of data warehouses using i-star. Journal of systems and software **88**, 102–111 (2014)

[16] Pang, A., Wittenbrink, C.M., Lodha, S.K.: Approaches to uncertainty visualization. The Visual Computer **13**(8), 370–390 (1997)

[17] Peña, O., Aguilera, U., López-de-Ipiña, D.: Exploring LOD through metadata extraction and data-driven visualizations. Program **50**(3), 270–287 (2016)

[18] Plotly: Dash (2019), https://plot.ly/

[19] Quartel, D.A.C., Engelsman, W., Jonkers, H., van Sinderen, M.: A goal-oriented requirements modelling language for enterprise architecture. In: Proceedings of the 13th IEEE International Enterprise Distributed Object Computing Conference, EDOC 2009. pp. 3–13. IEEE (2009)

[20] Salesforce: State of analytics (2015), www.lubbersdejong.nl/wp-content/uploads/2015/10/Salesforce-2015-State-of-Analytics-report.pdf

[21] Santos, H., Dantas, V., Furtado, V., Pinheiro, P., McGuinness, D.L.: From data to city indicators: A knowledge graph for supporting automatic generation of dashboards. In: The Semantic Web - 14th International Conference, ESWC. pp. 94–108. Springer (2017)

[22] Shi-Nash, A., Hardoon, D.R.: Data analytics and predictive analytics in the era of big data. Internet of Things and Data Analytics Handbook pp. 329–345 (2017)

[23] Vázquez-Ingelmo, A., García-Peñalvo, F.J., Therón, R.: Application of domain engineering to generate customized information dashboards. In: Learning and Collaboration Technologies. Learning and Teaching - 5th International Conference, LCT. pp. 518–529. Springer (2018)

Capítulo 5

Requirements-Driven Visualizations for Big Data Analytics: A Model-Driven Approach

Lavalle, A., Maté, A., & Trujillo, J. (2019, November). Requirements-Driven Visualizations for Big Data Analytics: A Model-Driven Approach. In *International Conference on Conceptual Modeling (ER 2019)* (pp. 78-92). Springer, Cham.

Conference Rating por GII-GRIN-SCIE: **Clase 3**
Conference Rating por CORE: **A**

Disponible en:
DOI: https://doi.org/10.1007/978-3-030-33223-5_8

Requirements-Driven Visualizations for Big Data Analytics: A Model-Driven Approach

Ana Lavalle[1,2( )], Alejandro Maté[1,2], and Juan Trujillo[1,2]

[1] Lucentia (DLSI), University of Alicante, Carretera San Vicente del Raspeig s/n, San Vicente del Raspeig, 03690 Alicante, Spain
{alavalle,amate,jtrujillo}@dlsi.ua.es
[2] Lucentia Lab, C/Pintor Pérez Gil, N-16, Alicante, Spain

Abstract. Choosing the right Visualization techniques is critical in Big Data Analytics. However, decision makers are not experts on visualization and they face up with enormous difficulties in doing so. There are currently many different (i) Big Data sources and also (ii) many different visual analytics to be chosen. Every visualization technique is not valid for every Big Data source and is not adequate for every context. In order to tackle this problem, we propose an approach, based on the Model Driven Architecture (MDA) to facilitate the selection of the right visual analytics to non-expert users. The approach is based on three different models: (i) a requirements model based on goal-oriented modeling for representing information requirements, (ii) a data representation model for representing data which will be connected to visualizations and, (iii) a visualization model for representing visualization details regardless of their implementation technology. Together with these models, a set of transformations allow us to semi-automatically obtain the corresponding implementation avoiding the intervention of the non-expert users. In this way, the great advantage of our proposal is that users no longer need to focus on the characteristics of the visualization, but rather, they focus on their information requirements and obtain the visualization that is better suited for their needs. We show the applicability of our proposal through a case study focused on a tax collection organization from a real project developed by the Spin-off company Lucentia Lab.

Keywords: Data visualization · Big Data Analytics · Model Driven Architecture · User requirements

1 Introduction

Data is continuously growing, specially since the last decade. With ever larger amounts of data that need to be interpreted and analyzed, using the right visualizations is crucial to help decision makers to properly analyze the data and guide them to take better informed decisions.

A. H. F. Laender et al. (Eds.): ER 2019, LNCS 11788, pp. 78–92, 2019.
https://doi.org/10.1007/978-3-030-33223-5_8

In this new era of Big Data Analytics, there has been an increasing interest from both the academic and industry worlds in different phases of the data life cycle: from the storage to the analysis, cleaning or integration and, of course, the visualization. Data and Information Visualization are becoming strategic for the exploration and explanation of large data sets due to the great impact that data have from a human perspective. An effective, efficient and intuitive representation of the analyzed data may result as important as the analytic process itself [6]. However, larger data sets and their complexity in terms of heterogeneity contribute to make the representation of data more complex [5].

In this context, defining and implementing the right visualization for a given data set is a complex task for companies, specially in the age of Big Data where heterogeneous and external data sources require knowledge of the underlying data to create an adequate visualization. As such, choosing the wrong visualizations and misunderstanding the data leads to wrong decisions and considerable losses. One of the key difficulties for defining the right visualization technique is the lack of expertise in information visualization of decision makers. Another critical aspect is that, apparently, a large set of visualizations may be equally valid for any given data sets, which has been proven to be absolutely wrong [21], each data set and each analysis has its particular characteristics and not always all the types of visualization are valid to represent them.

In order to tackle the above-presented problems, we propose an approach, based on the Model Driven Architecture (MDA) [16] proposed by the Object Management Group (OMG).

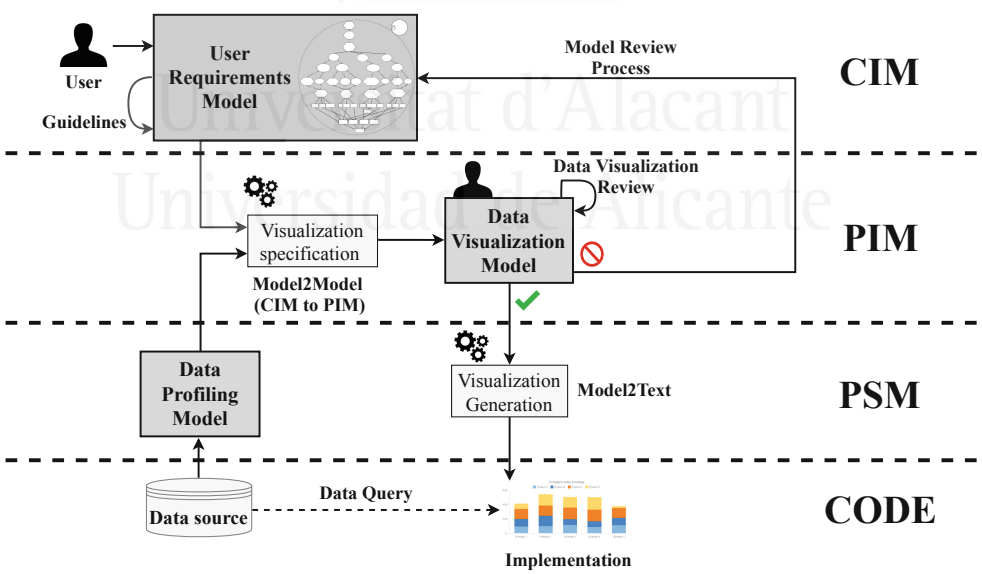


Fig. 1. Overall view of the process proposed.

Figure 1 summarizes the process followed in our proposal, aligned with MDA. Firstly, a sequence of questions guides users in creating a Goal-Oriented [12]

model that captures their needs. This model (CIM layer) enables them to capture all the visualizations that are needed to tackle their information needs. The user requirements together with the data profiling information coming through the Data Profiling Model (at PSM layer) are used as a visualization specification that is input to a model to model transformation. This transformation generates the Data Visualization Model (PIM layer). This model allows users to specify exactly how they need to visualize the data. It also allows them to determine if the proposed visualization is adequate to satisfy the essential requirements for which it was created. The validation process is performed through a questionnaire according to user goals model. If the proposed visualization passes the validation. Otherwise, an unsuccessful validation points out to the existence of missing or wrongly defined requirements that must be reviewed. This process is repeated until all user requirements are fulfilled. Finally, a model to text transformation generates the implementation of the visualization using the data visualization model as input.

The great advantage of our proposal is that users no longer focus on the underlying technical aspects or finding the most adequate visualization technique to be used in every different data analytic process. By following our approach, decision makers obtain the visualization technique that is better suited to their information needs and the characteristics of the data at hand in a semi-automatic way. This is achieved thanks to our alignment with MDA, enabling us to incrementally refine the visualization until its implementation is obtained.

The rest of the paper is structured as follows. Section 2 presents the related work in this area. Section 3 presents the different proposed models of the approach based on the MDA. Section 4 discusses a real case study in the fiscal domain. Finally, Sect. 5 summarizes the conclusions and our future work.

2 Related Work

Several works have focused on proposing different ways to find the best visualization. [2] surveys the main classifications proposed in the literature and integrates them into a single framework based on six visualization requirements. In [11], authors propose a framework for choosing the best visualization where the main types of charts are related to users goals and to data dimensionality, cardinality, and data type they support. Finally, [9] proposes a model to automate the translation of visualization objectives specified by the user into a suitable visualization type based on seven visualization requirements.

Additionally, several approaches are focused on the analysis of visualization representations. [15] describes an information visualization taxonomy. [18] make a revision of visualization techniques for Big Data to determine which are the most optimistic when analyzing Big Data. [4] propose a metamodel to represent tree and graph views by modeling nodes and edges. Similarly, [7] uses nodes and edges to draw basic shapes like lines and circles.

Other works are focus on visual analytics recommendation systems. [20] detail the key requirements and design considerations for a visualization recommendation system and identify a number of challenges in realizing this vision

and describe some approaches to address them. [8] propose EventAction, a prescriptive analytics interface designed to present and explain recommendations of temporal event sequences. Additionally, [21] propose SEEDB, a visualization recommendation engine to facilitate fast visual analysis, SEEDB explores the space of visualizations, evaluates promising visualizations for trends, and recommends those it deems most "useful" or "interesting". In [14] authors propose a new language VizDSL for creating interactive visualizations that facilitate the understanding of complex data and information structures for enterprise systems interoperability.

To the best of our knowledge, the only approaches that follow the MDA philosophy in the Big Data Context are presented within the TOREADOR project. In [1], the authors propose a Model-driven approach that aims to lower the amount of competences needed in the management of a Big Data pipeline. [10] illustrates a use case exploiting the Model-driven capabilities of the TOREADOR platform as a way to fast track the uptake of business-driven Big Data models. [13] provides a layered model that represents tools and applications following the Dataflow paradigm.

Despite all the work presented so far, none of the approaches provide a way to easily translate user requirements into visual analytics implementations. Furthermore, there is an absence of a methodology that guides users in obtaining the most adequate visualization, allowing them to focus on their own needs rather than on the characteristics of the visualization.

3 A MDA Approach for Visual Analytics

As previously introduced in the paper, specifying the right visualization for a user is a challenging task. User has not only to take into account her needs, which are on a completely different abstraction level, but also consider characteristics of the data that make inadequate the use of certain visualizations. In order to let the user focus on her information needs, we aim to bridge the gap between the user requirements and their visualization implementation.

To this aim, we propose a development approach Fig. 1 in the context of the Model Driven Architecture [16]. Our main goal is to help users to generate the visualizations that are better suited to meet their information needs. Following the basic principles of MDA, our proposal builds on three types of models:

- **User Requirements Model (CIM layer):** Allows users to capture their information needs and certain visualization aspects that are needed to tackle them.
- **Data Visualization Model (PIM layer):** Enables users to specify the characteristics of their visualizations before obtaining their implementation.
- **Data Profiling Model (PSM layer):** Abstracts the required information from the data sources to (i) aid in determining the most adequate visualization and (ii) take certain aspects of data into account for their representation (such as whether they are numeric or categorical).

The process starts by capturing information needs at the CIM level. Then, a data profiling process is run to generate a data profiling model at the PSM level that contains the relevant data characteristics for the process. Once both models have been obtained, they are processed through a model to model transformation that generates a data visualization model at the PIM level. This model provides the user with the better suited visualization for her needs and the data available, and allows her to modify different aspects of the visualization such as the axis where each attribute should be positioned, the orientation, or the color range among others. Once the model refinement process is finished, a model to text transformation generates the implementation using a visualization library, such as D3.js in our case.

3.1 User Requirements Model

Our approach starts from a goal-oriented requirements model that allows us to capture information needs. To describe the coordinates required to build a visualization context (*Goal, Interaction, User, Dimensionality, Cardinality, Independent Type, and Dependent Type*) we follows the specification to automate data visualization in Big Data Analytics given in [9], in this way we make sure that the visualization specification is addressed in terms of Big Data. Due to paper constraints, we cover only the main aspects of our requirements model.

Our metamodel shown in Fig. 2 is an extension of i* and the *i* for Data Warehouses* extension [12]. Existing elements in the i* core are represented in blue (light grey), whereas those in i* for Data Warehouses are represented in red (dark grey). The new concepts added by our proposal are represented in white.

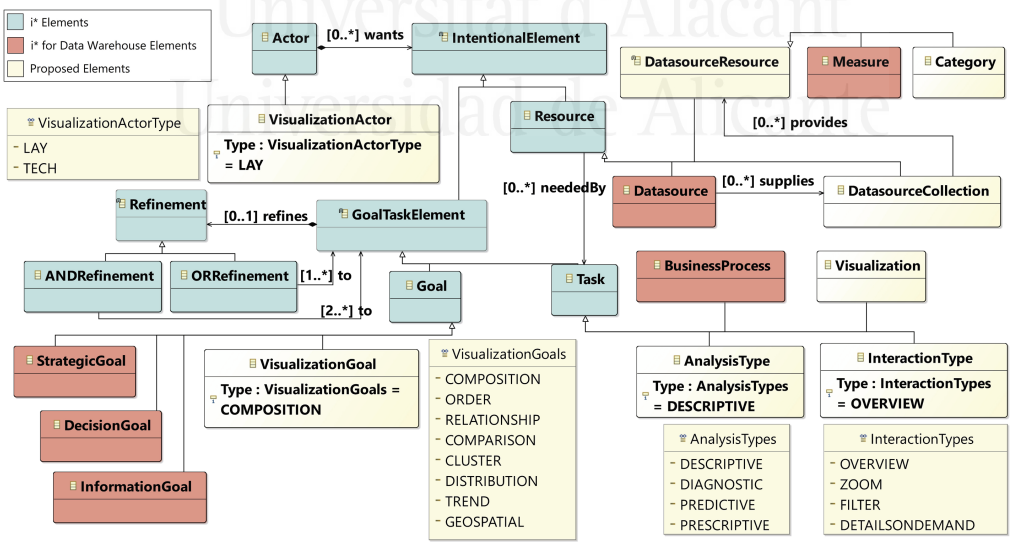


Fig. 2. User requirements metamodel. (Color figure online)

The first element is the *VisualizationActor*, which models the user of the system. There are two types of Visualization Actors: Lay, if she has no knowledge of complex visualizations, and Tech, if she has previous experience and is accustomed to Big Data Analytics. Next is the *BusinessProcess* on which users will focus their analysis. The business process will serve as the guideline for the definition of *Goals*.

The *AnalysisType* allows users to express which kind of analysis they wish to perform. The type of analysis can be determined by selecting which question from the following ones [19] is to be answered: How to act? (Prescriptive), Why has it happened? (Diagnostic), What is going to happen? (Predictive) or What to do to make it happen? (Descriptive).

Next, a Visualization represents a specific visualization that will be implemented to satisfy one or more *VisualizationGoals*. Each *VisualizationGoal* describes an aspect of the data that the visualization should reflect. These goals can be Composition, Order, Relationship, Comparison, Cluster, Distribution, Trend, or Geospatial, as considered in [9].

Along with *VisualizationGoals*, Visualizations have one or more *InteractionTypes*, that capture how the user will interact with the visualization. The different kinds of interaction are Overview, Zoom, Filter, or Details on Demand as [9] consider to data visualization in Big Data Analytics. Finally, a Visualization makes use of one or more *DatasourceResource* elements which feed the data to the visualization.

Using these concepts we allow users to define their needs instead of focusing on technical details that are not relevant at this level.

3.2 Data Profiling Model

Our second model is the Data Profiling Model. This model captures the data characteristics that are relevant to the visualization and is generated through a data profiling process. Firstly, users will select the data sources that they want to be represented in the visualization. Consecutively, the data analyst will analyze the data sources extract the values of the coordinates by analyzing the features of the data sources. In this way, users do not need to manually inspect the data or have a deep understanding.

To know how to delimit the values for each coordinate we have use the values proposed in [9] to Big Data Analytics. In this way we classify the Dimensionality, Cardinality, and Dependent/Independent Type as follows:

Cardinality represents the cardinality of the data. It can either be *Low* or *High*, depending of the numbers of items to represent. *Low* cardinality considers a few items to a few dozens of items while *High* cardinality is set if there are some dozens of items or more.

Dimensionality is used to declare the number of variables to be visualized. Specifically, it can be *1-dimensional* when the data to represent is a single numerical value or string, *2-dimensional* when one variable depends on other, *n-dimensional* when a data object is a point in an n-dimensional space, *Tree*

when a collection of items have a link to one other parent item, or *Graph* when a collection of items are linked to arbitrary number of other items.

Type of Data: is used to declare the type of each variable. It can be *Nominal* when each variable is assigned to one category, *Ordinal* when it is qualitative and categories can be sorted, *Interval* when it is quantitative and equality of intervals can be determined, or *Ratio* when it is quantitative with a unique and non-arbitrary zero point.

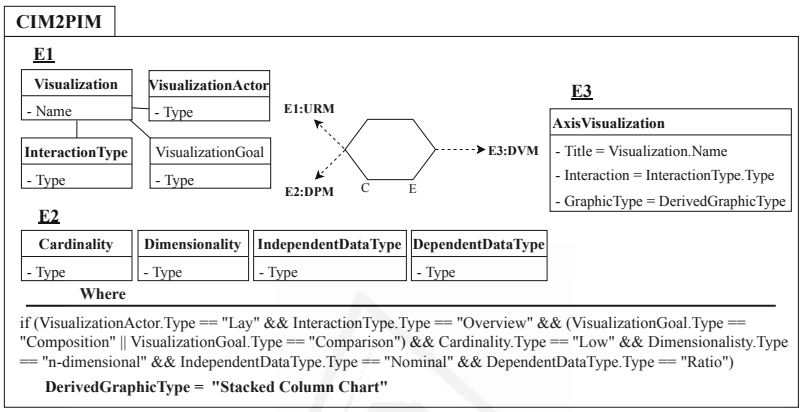


Fig. 3. Generation of axis based visualizations from user requirements.

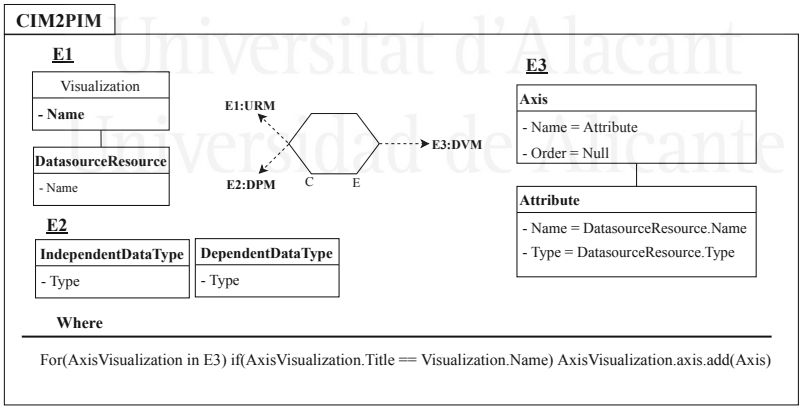


Fig. 4. Generation of axes for axis based visualizations from user requirements.

3.3 Visualization Specification Transformation - (Model to Model)

Information coming from User Requirements Model and the Data Profiling Model form the Visualization Specification. This specification is transformed

into a data visualization model using a set of model to model transformations, presented in Figs. 3 and 4 by the OMG standard language QVT [17]. According to the nature of the visualization to be derived, there are two types of transformations. On the one hand, we can have axis-based visualizations, such as column chart, line chart, bubble chart, etc. On the other hand, some visualizations such as dendrogram, chord or graphs require graph-based visualizations, which make use of nodes and edges instead of axis.

Due to space constraints, we will focus on how axis-based visualizations are derived. Our first transformation (Fig. 3), generates the visualization element, an AxisVisualization in this case. An AxisVisualization is derived according to the graphic type established by the transformation. This value is derived using the imperative part of the transformation (Where clause) according to the specific criteria established by [9] for the each graphic type. The values Cardinality, Dimensionality, IndependentDataType and DependentDataType are obtained from the data profiling. Finally, the visualization name and interaction type defined in the User Requirements Model are used to establish the title and interaction of the Axisvisualization.

Next, as Fig. 4 shows, each of the axes is generated individually. An axis is generated for each measure or category (abstracted by the DatasourceResource element) in the User Requirements Model. Afterwards, each axis is assigned their corresponding visualization by iterating over the data visualization model, completing the derivation of the visualization.

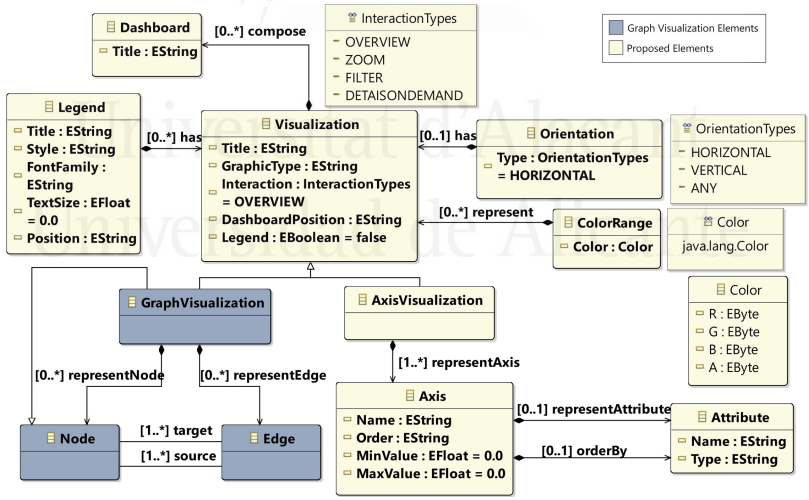


Fig. 5. Data visualization metamodel. (Color figure online)

3.4 Data Visualization Model

In order to verify if the recommended visualization is adequate to satisfy the information needs of the user and allow her to customize each visualization, we

require an abstraction of the visualization to be generated. Despite our best efforts, there is no metamodel proposed so far to model visual analytics. Thus, to support our process, we have defined a novel visualization metamodel.

Our metamodel shown in Fig. 5 is composed of elements extracted from [4] to define tree and graph visualizations, represented in blue (grey) color, while new concepts added by our proposal to detail the specification of visualizations represented in white. In the following, we describe the concepts included in the proposed metamodel.

The main element is *Visualization*, this element collects all the visualization requirements that should be met. It contains a visualization Title; a Legend, that may be shown or not; a Graphic Type that determines the type of visualization; a set of interactions that contain the type of interaction that must be supported (Overview, Zoom, Filter or Details-on-demand); and a Dashboard Position, in the event that the visualization will be part of a *Dashboard*.

In order to define the representation of a visualization, other elements are necessary. A visualization has and *Orientation*, either Horizontal, Vertical, or Any (when the graphic type does not have orientation). Moreover, a visualization has a *ColorRange*, that represents the range of colours that will be used by the visualization, an aspect of special importance for color-blind users.

A visualization will be instanced as either a *GraphVisualization* or an *AxisVisualization* depending on the type of visualization. A *GraphVisualization* may contain several *Nodes* and *Edges* [4]. Meanwhile, an *AxisVisualization* constaints a series of axes that represent the data. An *Axis* is may have a Name, Order, Minimum Value and Maximum Value. Each Axis represents an Attribute at most. An *Attribute* has a Name and a Type. Attributes can be used to be represented or to set the order of the data in the visualization.

3.5 Visualization Generation Transformation - (Model to Text)

The Visualization Generation Transformation has as input the data visualization model from the previous step. This transformation transforms each element within the visualization specification into a code level specification for a graphic library. In our case, we use the D3 JavaScript library [3] for generating the visualization. The GraphicType and the Orientation determine the type of visualization to implement. Categories and measures and their respective axes determine how the data is assigned to each axis. Meanwhile, the Color Range is translated into custom color scales. Moreover, if a Legend has been defined, the type of the legend, title, position, font family and text size are be translated attributes in the corresponding d3.legend function call. Finally, the title is used to provide a name to the visualization created, and the dashboard position is used to assign a position to the visualization.

4 Case Study

In order to evaluate the validity of our approach we have applied it to a real case study, based on a tax collection organization. Due to space constraints, we

provide a reduced example including enough data in order to allow readers to completely understand the approach. Therefore, the example is constrained to a Tax Region Area covering only three provinces. The organization requires a set of visualizations to analyze their data in order to help them detect underlying patterns in their unpaid bills and tax collection distribution. Due to the sensitivity of their data, we are not allowed to show the real values.

4.1 Specifying User Requirements

Through the application of our User Requirements Model to a tax collection organization, the Fig. 6 has been generated. A tax collector user wants to analyze the unpaid debts. Therefore, the analysis will focus on the *"Tax collection"* business process. Defining a business process helps determining the scope of the analysis and the goals pursued. The user is not a specialist in Big Data Analytics but rather an expert in tax management, thus she is defined as *"Lay user"*.

Next, the main objectives of the business process are defined as shown in Fig. 6. Specifically, the user defined her strategic goal as *"Reduce the unpaid bills"*. Strategic goals are achieved by means of analyses that support the decision-making process. The analysis type allows users to express what kind of analysis they wish to perform. In this case, the user wishes to know why bills are unpaid. Thus, the user decides to perform a *"Diagnostic analysis"*.

The diagnostic analysis is decomposed into decision goals. The user defined her decisions goals as: *"Identify unpaid bills"*, *"Identify the quantities unpaid"*, and *"Analyze the evolution"*. Decisions goals communicate the rationale followed by the decision-making process; however, by themselves they do not provide the necessary details about the data to be visualized. Therefore, for each decision goal we specify one or more information goals.

From each of the decision goals the user refined the following information goals: *"Identify places with more unpaid bills"*, *"Identify the type of unpaid bills"*, *"Identify who has unpaid bills"*, and *"Evolution of unpaid bills"*. Information goals represent the lowest level of goal abstraction. And for each information goal, we will have one visualization to achieve it. A visualization is characterized by one or more visualization goals which describe what aspects of the data the visualization is trying to reflect, and one or more kinds of interaction that they will like to have with the visualization. Moreover, a visualization will make use of one or more data source elements to get the relevant data from the database. In this case, the user defines the interactions she want to have with each visualization and her visualization goals following user guidelines. *"Overview"*, *"Zoom"* and *"Details-on-demand"* have been defined as interactions and *"Geospatial"*, *"Composition"*, *"Comparison"*, *"Order"*, and *"Trend"* as visualization goals. Finally, the user specifies the data source where the analysis will be performed and selects the Categories and Measures that will populate the visualizations.

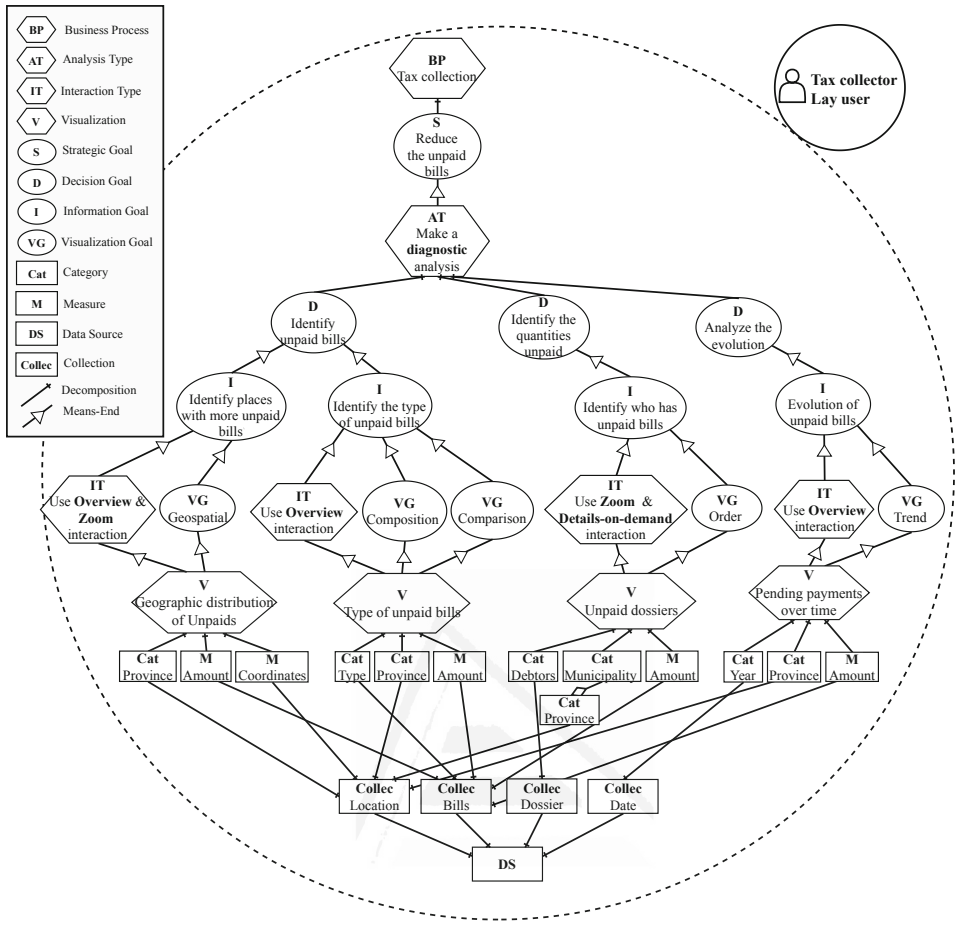


Fig. 6. Application of our user requirements metamodel to the case study.

4.2 Profiling Data Sources

Once user have defined the data sources and collections from where the data will be extracted, it is possible to profile data sources to determine Dimensionality, Cardinality and Dependent/Independent Type.

We focus on the **"Identify the type of unpaid bills"** Information Goal from our Goal-Oriented model, which requires information about categories *"Type"* and *"Province"* and measures *"Amount"*. Firstly, by the Data Profiling Model, are classified the independent variables *"Type"* and *"Province"* as **Nominal** and the dependent variable *"Amount"* as **Ratio**. Dimensionality is set to **n-dimensional**, because the user has defined 3 variables to visualize. Finally, the Cardinality is defined as **Low** Cardinality because the data contains a few items to represent 3 provinces to represent and there are 6 types of bills.

Overall, the visualization specification obtained through User Requirements Model and Data Profiling Model are:

- **Visualization Goal:** Composition & Comparison
- **Interaction:** Overview
- **User:** Lay
- **Dimensionality:** n-dimensional
- **Cardinality:** Low
- **Independent Type:** Nominal
- **Dependent Type:** Ratio

With the definition of this visualization specification, by applying our visualization specification transformation, the visualization type generated is "Stacked Column Chart".

4.3 Specifying Data Visualizations Requirements

The visualization specification is used as input of the Data Visualization Model. A visualization tool will be generated as Fig. 7 shows using the information collected in the process.

The tool shows the most suitable visualization type, the integration type defined by the user and a representation of the visualization. It also shows the selected elements to be represented in the visualization. The user will have to choose in which **axes** she want to see each element represented. In this case, we have *"Province"* in X axis, *"Amount"* in Y axis and *"Type"* as Color. The user also has to select the element that determines the **order** in the visualization. Other element to specify is the **orientation** of the visualization, this can be defined as horizontal, vertical or any if the visualizations have no orientation. In this case the user has decide to user a horizontal orientation. Next element is the **legend**, which can be shown or not. A legend may have a title, a type (in this case the user has decide to represent it like a list), a position on the visualization, a font family, and a text size. The **range of colours** used to represent the visualization also has to be choose, the user can choose one of the color ranges proposed or personalize a range. Finally, the user can give a dashboard position to the visualization and a title.

The user will review the data visualization model until she achieves her visualization requirements. Once all the elements have been customized, the user has to validate if the visualization obtained does contribute to answer her informational goal, in this case *"Identify the type of unpaid bills"*. If the visualization is validated, it will be generated making a call to the D3 JavaScript library [3], obtaining the visualization shown in Fig. 8. Otherwise, an unsuccessful validation would generated a review of the existing user requirements model, to start a new iteration and generating in turn an updated model.

This visualization, combined with those generated for the others information goals, will be grouped into a dashboard, aimed at satisfying the analytic requirements of our tax collector user with the most adequate visualizations and covering all the data required by the analysis.

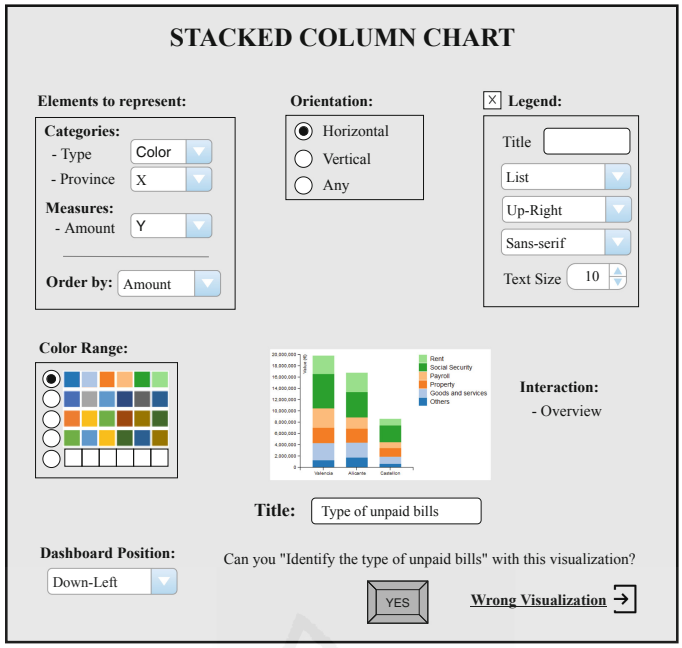


Fig. 7. Application of our data visualization metamodel to the case study.

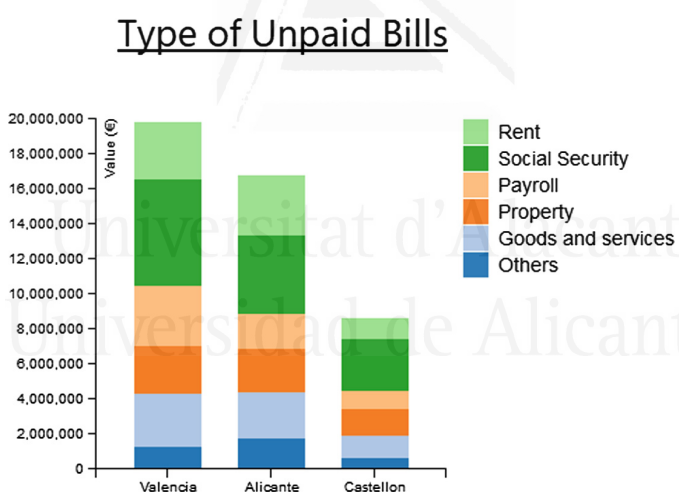


Fig. 8. Visualization rendered in D3.js.

5 Conclusions and Future Work

In this paper, we have presented an approach in the context of the Model Driven Architecture (MDA) standard in order to help users derive the most adequate visualizations. Our approach envisages three different models, (i) a requirements model based on goal-oriented modeling for representing information requirements; (ii) a data profiling model that abstracts the required information from

the data sources; and, (iii) a visualization model for capturing visualization details regardless of their implementation technology. Together with these models, we have proposed a series of transformations that allow us to bridge the gap between information requirements and the actual implementation. The great advantage of our proposal is that users can focus on their information needs and obtain the visualization that is better suited for their particular case, without requiring visualization expertise. In order to check the validity of our approach, we have applied our approach to a real use case focused on a tax collection organization. The results obtained, as well as a currently ongoing family of experiments, support the approach presented.

As part of our future work, we are working on the definition and generation of dashboards as a whole. In this way, we will simplify and reduce the resources required to obtain visual analytics, which is of special interest for small and medium companies who cannot afford hiring several analysts in order to cover data, visualization, and business expertise required for Big Data analytics.

Acknowledgments. This work has been co-funded by the ECLIPSE-UA (RTI2018-094283-B-C32) project funded by Spanish Ministry of Science, Innovation, and Universities. Ana Lavalle holds an Industrial PhD Grant (I-PI 03-18) co-funded by the University of Alicante and the Lucentia Lab Spin-off Company.

References

1. Ardagna, C.A., Bellandi, V., Ceravolo, P., Damiani, E., Bezzi, M., Hébert, C.: A model-driven methodology for big data analytics-as-a-service. In: International Conference on Big Data, pp. 105–112. IEEE (2017)
2. Börner, K.: Atlas of Knowledge. MIT Press, Cambridge (2014)
3. Bostock, M.: Data-driven documents (2019). https://d3js.org/
4. Bull, R.I., Favre, J.: Visualization in the context of model driven engineering. In: MDDAUI, vol. 159 (2005)
5. Caldarola, E.G., Rinaldi, A.M.: Improving the visualization of wordnet large lexical database through semantic tag clouds. In: International Congress on Big Data, pp. 34–41. IEEE (2016)
6. Caldarola, E.G., Rinaldi, A.M.: Big data visualization tools: a survey - the new paradigms, methodologies and tools for large data sets visualization. In: Proceedings of the 6th International Conference on Data Science, Technology and Applications, DATA. INSTICC, SciTePress (2017)
7. Domokos, P., Varró, D.: An open visualization framework for metamodel-based modeling languages. Electr. Notes Theor. Comput. Sci. **72**(2), 69–78 (2002)
8. Du, F., Plaisant, C., Spring, N., Shneiderman, B.: Eventaction: visual analytics for temporal event sequence recommendation. In: 2016 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 61–70. IEEE (2016)
9. Golfarelli, M., Rizzi, S.: A model-driven approach to automate data visualization in big data analytics. Inf. Vis. (2019, to appear)
10. Leida, M., Ruiz, C., Ceravolo, P.: Facing big data variety in a model driven approach. In: RTSI, pp. 1–6. IEEE (2016)
11. Madhu Sudhan, S., Chandra, J.: IBA graph selector algorithm for big data visualization using defense data set. Int. J. Sci. Eng. Res. (IJSER) **4**(3), 1–7 (2013). ISSN: 2229-5518

12. Maté, A., Trujillo, J., Franch, X.: Adding semantic modules to improve goal-oriented analysis of data warehouses using I-star. J. Syst. Softw. **88**, 102–111 (2014)
13. Misale, C., Drocco, M., Aldinucci, M., Tremblay, G.: A comparison of big data frameworks on a layered dataflow model. Parallel Process. Lett. **27**, 1740003 (2017)
14. Morgan, R., Grossmann, G., Stumptner, M.: VizDSL: towards a graphical visualisation language for enterprise systems interoperability. In: BDVA. IEEE (2017)
15. de Oliveira, E.C., de Oliveira, L.C., Cardoso, A., Mattioli, L., Junior, E.A.L.: Metamodel of information visualization based on treemap. Univ. Access Inf. Soc. **16**(4), 903–912 (2017)
16. (OMG), O.M.G.: Model driven architecture guide rev. 2.0 (2014). https://www.omg.org/cgi-bin/doc?ormsc/14-06-01
17. (OMG), O.M.G.: MOF 2.0 query/view/transformation specification (2016). https://www.omg.org/spec/QVT/1.3/PDF
18. Peña, L.E.V., Mazahua, L.R., Hernández, G.A., Zepahua, B.A.O., Camarena, S.G.P., Cano, I.M.: Big data visualization: review of techniques and datasets. In: International Conference on Software Process Improvement, pp. 1–9. IEEE (2017)
19. Shi-Nash, A., Hardoon, D.R.: Data analytics and predictive analytics in the era of big data. In: Internet of Things and Data Analytics Handbook, pp. 329–345 (2017)
20. Vartak, M., Huang, S., Siddiqui, T., Madden, S., Parameswaran, A.: Towards visualization recommendation systems. ACM SIGMOD Record **45**(4), 34–39 (2017)
21. Vartak, M., Rahman, S., Madden, S., Parameswaran, A., Polyzotis, N.: SEEDB: efficient data-driven visualization recommendations to support visual analytics. Proc. VLDB Endowment **8**(13), 2182–2193 (2015)

Capítulo 6

Improving Sustainability of Smart Cities through Visualization Techniques for Big Data from IoT Devices

Lavalle, A., Teruel, M. A., Maté, A., & Trujillo, J. (2020). Improving Sustainability of Smart Cities through Visualization Techniques for Big Data from IoT Devices. *Sustainability, 12*(14), 5595.

Factor de Impacto: **2,576**
Clasificación JCR: **Q2 (120/265 Environmental sciences)**

Disponible en:
DOI: https://doi.org/10.3390/su12145595

sustainability

MDPI

Article

Improving Sustainability of Smart Cities through Visualization Techniques for Big Data from IoT Devices

Ana Lavalle [1,2,*], Miguel A. Teruel [1,2], Alejandro Maté [1,2] and Juan Trujillo [1,2]

[1] Lucentia Research, DLSI, University of Alicante, Carretera San Vicente del Raspeig s/n, San Vicente del Raspeig, 03690 Alicante, Spain; materuel@dlsi.ua.es (M.A.T.); amate@dlsi.ua.es (A.M.); jtrujillo@dlsi.ua.es (J.T.)

[2] Lucentia Lab, C/Pintor Pérez Gil, N-16, 03540 Alicante, Spain

* Correspondence: alavalle@dsi.ua.es

Received: 15 May 2020; Accepted: 9 July 2020; Published: 11 July 2020

Abstract: Fostering sustainability is paramount for Smart Cities development. Lately, Smart Cities are benefiting from the rising of Big Data coming from IoT devices, leading to improvements on monitoring and prevention. However, monitoring and prevention processes require visualization techniques as a key component. Indeed, in order to prevent possible hazards (such as fires, leaks, etc.) and optimize their resources, Smart Cities require adequate visualizations that provide insights to decision makers. Nevertheless, visualization of Big Data has always been a challenging issue, especially when such data are originated in real-time. This problem becomes even bigger in Smart City environments since we have to deal with many different groups of users and multiple heterogeneous data sources. Without a proper visualization methodology, complex dashboards including data from different nature are difficult to understand. In order to tackle this issue, we propose a methodology based on visualization techniques for Big Data, aimed at improving the evidence-gathering process by assisting users in the decision making in the context of Smart Cities. Moreover, in order to assess the impact of our proposal, a case study based on service calls for a fire department is presented. In this sense, our findings will be applied to data coming from citizen calls. Thus, the results of this work will contribute to the optimization of resources, namely fire extinguishing battalions, helping to improve their effectiveness and, as a result, the sustainability of a Smart City, operating better with less resources. Finally, in order to evaluate the impact of our proposal, we have performed an experiment, with non-expert users in data visualization.

Keywords: internet of things; data visualization; big data analytics; smart city; methodology; artificial intelligence; dashboards

1. Introduction

Needless to say that the cities we live on are increasingly becoming Smart Cities [1]. Smart city refers to a type of urban development based on sustainability [2] that is capable of adequately responding to the basic needs of institutions, companies, and the inhabitants themselves, both economically and in operational, social and environmental aspects. Thus, in order to achieve such sustainability, Smart Cities need what has been called the 21st century's oil [3], namely data.

Data can be collected through different methods [4], for instance, cameras and sensors distributed throughout the city, communication between devices, or the interaction of human with machines. It can be generated by meteorological observatories, financial markets or social networks. Even Internet traffic of a Smart City can be analyzed. A large volume of data is generated continuously and it

increases sharply as time passes. Therefore, a Smart City packed with inhabitants and devices can be considered as a very vast source of data from which we can extract valuable knowledge [5,6].

The extraction of knowledge from data can improve many sectors. Indeed, economists in [7] estimate that the global Gross domestic product (GDP) will be up to 14% higher in 2030 as a result of the accelerating development and take-up of Artificial Intelligence (AI) [8], the equivalent of an additional $15.7 trillion. This scenario represents a significant boost for the world economy with potential to raise a new industry and generate numerous business opportunities. Despite the fact that cities occupy only 2% of the planet's territory, they represent between 60% and 80% of world energy consumption and generate 70% of greenhouse gas emissions [9]. Therefore, in order to foster sustainability and reduce energy consumption and greenhouse gas emissions, we require to improve the efficiency of cities and transform them into Smart Cities through data analysis. Hence, Big Data will play a fundamental role for the management of resources and the provision of important services.

Analyzing all dataflows coming from Smart Cities represents a big challenge, partially addressed by the Big Data paradigm [10]. Their characteristics match Big Data V's, providing large amounts of data (Volume), the data comes from different sources, some of them are unstructured (Variety), coming from reliable and unreliable sources (Veracity), some of them are generated and have to be processed in real-time (Velocity), there is knowledge that the data provides value (Value), it is possible to process and analyze the captured data (Viability) and represent it in an understandable way to the final users (Visualization).

The Big Data paradigm enables us to gather all this data and process it in order to extract value from the data sources and represent the information in a visual and attractive manner for the users. It also helps to obtain behavior patterns that will enable us to design solutions and modify different processes of cities to make them more sustainable.

Ref. [11,12] list the different areas and propose examples in which initiatives could be applied to convert a city into a Smart City. Figure 1 represents this areas graphically. Such areas are explained in the following:

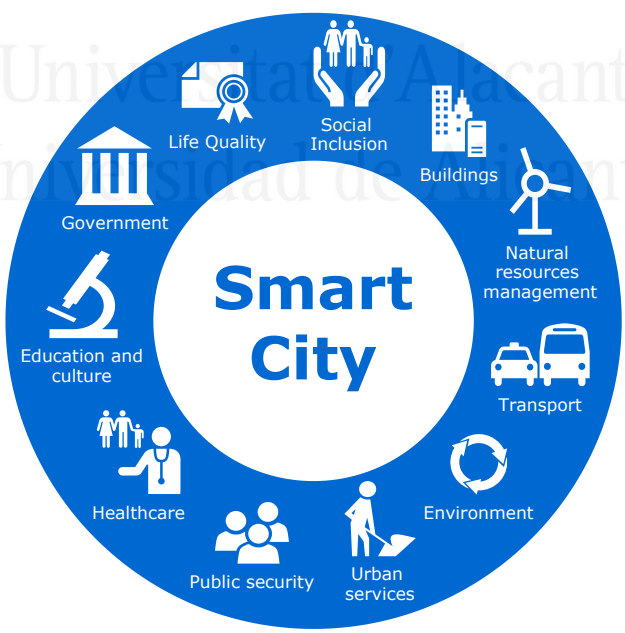


Figure 1. Areas of a Smart City.

- Social Inclusion: Underlining the importance of creativity, social integration, education and tolerance, with the aim of bring the people and society inside to create a participatory and innovative environment.
- Buildings: Intelligent systems in buildings could control the electrical and mechanical equipment. It will reduce the maintenance costs and will improve their safety.
- Natural resources management: Sensors could automatically control the quality of resources, as well as locate possible leaks (water, gas, oil and so on) or incidents.
- Transport: By analyzing data from cameras and sensors distributed throughout the city, it is possible to reduce traffic jams or act on them efficiently. It will be possible to interact with the network of traffic lights to make traffic more fluid. Also, citizens could also be informed of the traffic situation and indicate the optimal alternative routes.
- Environment: Better energy management will save money and improve the environment. It will also be possible to use different types of devices to analyze the emissions of greenhouse gases and control them.
- Urban services: Through social networks, call centers and other systems, it will be possible to know the opinion of citizens and tourists about the city's services in real-time. Therefore, it will be possible to know what services are needed and in which locations.
- Public security: Data could be collected and monitored to prevent crime through the correlation of the information collected by the different systems installed in the city.
- Healthcare: Studying and analyzing clinical data will improve the decision-making in medical treatments. Also, the digitization of hospital management, will make it more efficient.
- Education and culture: A better planning for each student can be offered by personalizing their curriculum and analyzing their progression in order to detect situations of risk and reduce the dropout rates.
- Government: Digitization of the public administration in order to optimize the services and to be able to offer them through Internet.
- Life Quality: Solutions to provide information about the places of interest of a city and cultural activities in order to facilitate the dissemination of information and motivate people to get involved in them.

As described above, there are many areas in which initiatives could be applied in order to aid in the conversion of a city into a Smart City. Unfortunately, this conversion is not automatic, it must be done guided by conscious decisions city representatives. As such, data and AI outputs themselves can be worthless if they are not represented in a proper manner. Indeed, to enable Smart Cities representatives to take actions toward the correct development on such cities, it is necessary to provide such data as comprehensive and straightforwardly possible, ideally in the form of key visualizations that drive change. This can be cumbersome due to the complexity and volume of the data generated by the Smart City's devices. Furthermore, as [10] discuss, one of the main issues in Smart City development is to transform the great amount of data streams into knowledge and finally, into strategic and tactical decisions. Hence, using the correct visualizations is very important.

Therefore, in this paper we focus on the area of Smart Cities, with the aim of highlighting the value that can be extracted from the data generated in them. Like any Big Data project, it is necessary to capture, store, process and analyze large amounts of data from different sources in order to transform them into useful knowledge. The main goal of our work is to provide a methodology that helps Smart City representatives to make smart decisions aided by visualizations, leading the city towards a more sustainable growth.

In previous works [13,14] we have defined a methodology that helps users to define and achieve their goals. It extracts in a semi-automatically manner characteristics of the data sources and derives automatically the best type of visualizations according to the defined context.

In this paper, we have significantly improved our previous works as follows. First of all, (i) we show how our proposal can be applied to the context of Smart Cities, exploiting existing datasets in combination with Artificial Intelligence to provide city representatives with evidences to make decisions that improve sustainability through city development and resource management. The Artificial Intelligence approach is a novel issue introduced in this paper for the first time in our overall and wider methodology. In addition, we (ii) improve our previous works by incorporating the required steps and solutions to process data in real-time, and (iii) facilitate context information to users, helping them better understand the output of (iv) an Artificial Intelligence algorithm we trained for this particular case study.

Therefore, to the best of our knowledge, this is the first work presenting a complete methodology, based on visualization techniques for non-experts, to cover all the data value chain from its generation, through its gathering and processing and finally, offering an easy visualization technique (including AI) to facilitate the decision making in the context of Smart Cities.

In order to put into practice the proposed methodology, we have used a case study based on the service calls management from the fire department of San Francisco. The objective of this city is to improve the sustainability of their processes by analyzing the responses of the emergency services. A complete dashboard will help them to reorganize the services as needed and to be ready for the action. It will lead to the reduction of the number and severity of serious fires in the city.

Furthermore, in order to evaluate the impact of our proposal, we have performed an experiment, with 12 non-expert users in data visualization. Each user was tasked with filling a questionnaire with two exercises. In the first exercise, a dataset was provided to users with the aim that they to made an analysis on them. Besides, in the second exercise, users were tasked with do different analysis by following our proposed methodology. The results obtained from the experiment have been analyzed.

The advantages of our proposal are that (i) it helps users to define their goals and achieve them through decision making supported by the most adequate historical and real-time visualizations in the context of Smart Cities, (ii) it provides a rationale for dashboard design, (iii) it helps to visually understand the output of Artificial Intelligence algorithms (iv) it enables users to gather evidence for making strategic and tactical decisions. Without the benefits introduced by our proposal, it would be hard for the users to understand the state in which their processes are, and to be able to make the best decisions in relation to them.

The rest of the paper is structured as follows. Section 2 presents the related work in this area. Section 3 describes our proposed methodology to fostering sustainability through visualizations. Section 4 shows our approach applied in a Smart City case study. Section 5 presents a evaluation of our approach. Finally, Section 6 summarizes the conclusions and our future work.

2. Related Work

During the last few decades, sustainability and sustainable development have become popular topics not only for scholars in the fields of environmental economics, technology and science, urban planning, development and management. It has also become popular for urban policy makers and professional practitioners [15].

In a city context, a city may be called smart when investments in human, social capital and ICT, foster sustainable economic growth and a high quality of life, as well as wise management of natural resources, through participatory government [16].

Regarding Internet of Things (IoT) for Smart Cities, significant research effort and technological development have been devoted. The main reason is the exponential growth of devices/smart objects that can participate in an IoT infrastructure [17].

According to [18], the typical challenges raised by the application of the IoT on Smart Cities are: Security and Privacy (the system can be subjected to attacks), Heterogeneity (each system component is knitted to the particular application context), Reliability (the communication between smart devices may not reliable enough), Large Scale (the large scale of information requires suitable storage and

computational capability), Legal and Social Aspects (when the service is based on user-provided information), Big Data (it is certainly necessary to pay attention to transferring, storing and recalling and also analyzing such a huge amount of data produced by smart devices) and Sensor Networks (process the large-scale data of the sensors in terms of energy and network limits and various uncertainties).

Some researchers propose approaches to provide solutions to interconnect IoT elements. In [19] is proposed an Artificial Intelligence-based semantic IoT hybrid service architecture which enables flexible connections among heterogeneous IoT devices. In [20] is proposed an IoT-based platform for the development of cyber–physical systems suitable for Smart Cities services and applications. The authors provide a set of abstractions suitable to hide the heterogeneity of the physical sensor/actuator devices embedded in the system. Besides, in [21] authors are focused in locating and optimizing the traffic in cities through a swarm-based architecture that interconnect their elements.

In [22] is introduced a Machine Learning approach to automate and help crime analysts to identify the connected entities and events. They collect, integrate and analyze diverse data sources to generate alerts and predictions.

Besides, several works have focused in how to visualize IoT data coming from Smart Cities. Ref. [23] highlights the challenge of real-time data stream visualization in the fields of Smart Cities as traffic, pollution, social media activity, citizens dynamics, etc. They apply different types of glyphs for showing real-time stream evolution of data gathered in the city. In [24], they propose a Service Oriented Architecture software platform aimed to providing Smart City services on top of 3D urban city models. In [25] a solution is proposed for enhancing the visualization of IoT information. This study proposes a framework to integrate IoT data into an environment based on Augmented Reality (AR). Finally, Ref. [26] shows a use case scenario, where live air quality data that is visualized and monitored via sensors installed on top of mobile post vans, driving around in the City of Antwerp.

However, none of these approaches take into account the users' goals. As [10] argue, one of the main issues in Smart City development is to transform the great amount of data streams into knowledge and finally, into strategic and tactical decisions. Hence, using the correct visualization is very important. The manner people perceive and interact with a visualization tool can strongly influence their data understanding as well as the system's usefulness [27].

Therefore, we propose a process that helps users to define their goals and derive automatically the best type of visualization in the context of Smart Cities. Moreover, we provide visual techniques to easily understand the output of Artificial Intelligence algorithms. It will enable users of Smart Cities to make strategic and tactical decisions in order to improve the sustainability of their processes.

3. Process to Fostering Sustainability through Historical Visualizations and Instant Analysis

In this section we will describe our proposed process. Figure 2 summarizes the process followed in our proposal. By following this process, users will be able to communicate their analytical needs and to obtain automatically the set of visualizations most suitable to achieve their goals. These visualizations will be grouped in a powerful dashboard that will help them to make strategic and tactical decisions in order to help to improve the evidence-gathering process by assisting users in visualising data.

Firstly, users define through a series of guidelines the User Requirements Model presented in [13]. They will also define a set of Key Performance Indicators (KPIs) in order to measure the degree of achievement of the goals.

Once the requirements have been defined, we differentiate 2 types of analysis. On the one hand, a study to analyze the historical data that will enable users to have an image of the current situation of the process. On the other hand, an instant analysis, where the incoming data will be processed at the moment it arrives with the aim of anticipating to the events, as it will be explained later (Section 3.3).

Therefore, our process provides a set of visualization techniques to understand the current situation of a process, thus enabling users to visually represent the output of Artificial Intelligence Algorithms. In the following, we describe in detail the elements that compose our process (Figure 2).

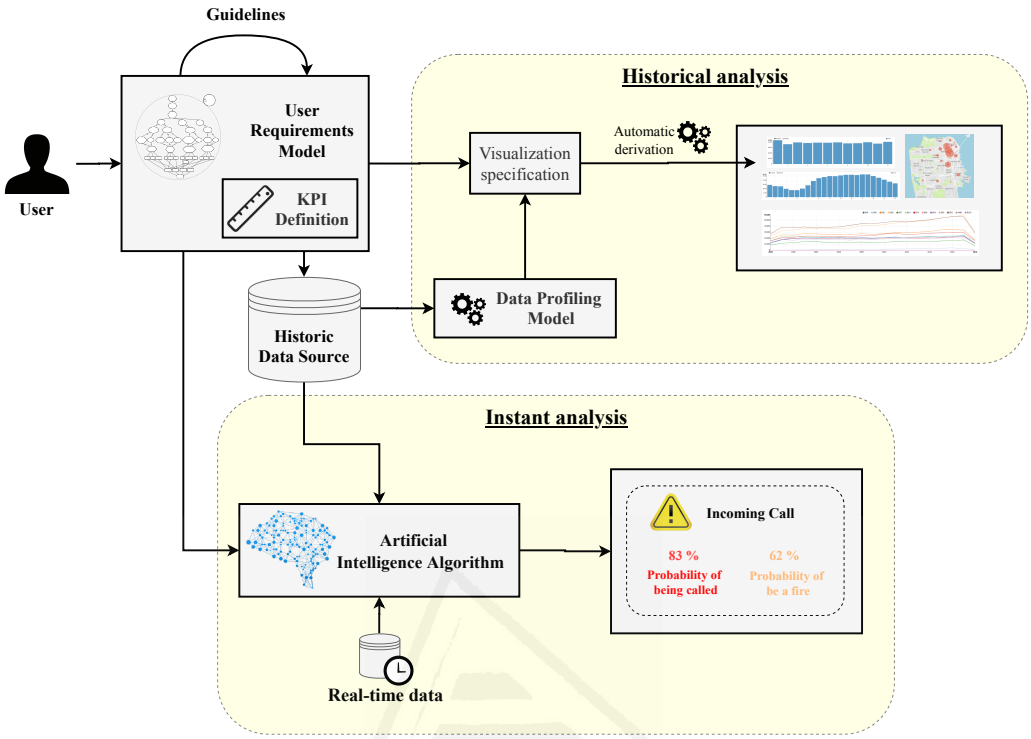


Figure 2. Proposed Process.

3.1. Definition of Requirements and KPIs

The first element that in our process is the User Requirements Model. This element will help users to define their data analysis objectives and to achieve it through the visualizations that best suit them. We can find an example of the application of this model into a real case of study in Section 4. In order to formally define our model, in [13] we proposed this metamodel that we can see in Figure 3).

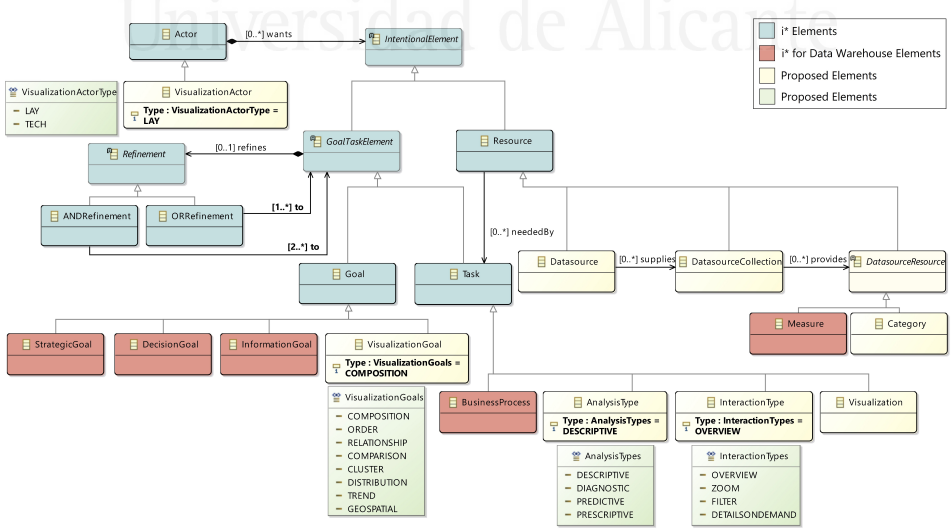


Figure 3. User Requirements Metamodel.

The proposed metamodel is an extension of the one used for social and business intelligence modeling [28], namely i* [29] and the *i* for Data Warehouses* extension [30]. i* elements are represented in blue, and i* for Data Warehouses elements are represented in red. The elements added by us are represented in light yellow and green. In the following, we will describe the elements that compose the metamodel.

The *Visualization Actor* refers to the user of the system. There are two types of Visualization Actors: Lay, if the user has no knowledge of complex visualizations, or Tech, if she has experience. The next element is the *Business Process* on which users will focus their analysis. The business process will serve as the guideline for the definition of *Goals*. A goal represents a desired state of affairs with reference to the business process at hand. Goals can be divided into *Strategic*, *Decision*, *Information*, and *Visualization*.

The *Analysis Type* enables users to express which kind of analysis they wish to perform. The type of analysis can be determined by selecting which question from the following ones [31] is to be answered: How to act? (Prescriptive), Why has it happened? (Diagnostic), What is going to happen? (Predictive) or What to do to make it happen? (Descriptive).

A *Visualization* represents a specific visualization type that will be implemented to satisfy one or more *Visualization Goals*. Each *Visualization Goal* describes an aspect of the data that the visualization should reflect. These goals can be Composition, Order, Relationship, Comparison, Cluster, Distribution, Trend, or Geospatial, as considered in [32]. Along with *Visualization Goals*, *Visualizations* have one or more *Interaction Types*, that capture how the user will interact with the visualization. The different kinds of interaction are Overview, Zoom, Filter, or Details on Demand as [32] consider. Finally, a *Visualization* will make use of one or more *Data Source Resource* elements which will feed the data to the visualization.

As argued in [32], inexperienced users may find it difficult to properly give values to these elements. Consequently, in [13], we proposed a series of guidelines to assist non-expert users in definition of the model elements.

In order to improve the definition and measure the degree of achievement of the goals defined in the User Requirements Model, users may define a set of Key Performance Indicators (KPIs). To do this, users will use a table like the one shown in Table 1. This table represents an example of the goals defined in the User Requirements Model that require extra information. For each goal, users may define a KPI to measure it and a threshold to identify its possible states of the KPI.

Table 1. Example of table to the definition of KPIs.

Goal	KPI	Threshold
Increase profit	Revenue - Cost	>3 M$ Good, 3 M$–1 M$ Acceptable, <1 M$ Bad
Increase the number of visits	Number of visits per week	>1000 Good, 1000–700 Acceptable , <700 Bad
...	...	...

Once all the requirements have been defined, we differentiate 2 types of analysis as we see in the following.

3.2. Historical Analysis

The Historical Analysis is aimed to make a summary of the current situation of the process. The visualizations derived in this analysis will help users to understand which are the critical areas of their processes and to be able to make right decisions about them.

3.2.1. Data Profiling Model

In order to derive the types of visualizations that will fit best in the Historical Analysis, we use the Data Profiling Model to capture the characteristics of the data that are relevant for the visualizations.

The **Dimensionality**, **Cardinality**, and **Dependent/Independent Data Type** will be extracted in a semi-automatic manner as explained below.

- **Cardinality** represents the cardinality of the data, it depends of he numbers of items to represent. It can be:
  - *Low* from a few items to a few dozens of items.
  - *High* if there are some dozens of items or more.
- **Dimensionality** represent the number of variables to be visualized. It can be:
  - *1-dimensional* when the data to represent is a single numerical value or string.
  - *2-dimensional* if one variable depends on other.
  - *n-dimensional* when a data object is a point in an n-dimensional space.
  - *Tree* when a collection of items have a link to one other parent item.
  - *Graph* provided a collection of items are linked to arbitrary number of other items.
- **Type of Data:** is used to declare the type of each variable. It can be:
  - *Nominal* when each variable is assigned to one category.
  - *Ordinal* when it is qualitative and categories can be sorted.
  - *Interval* if it is quantitative and equality of intervals can be determined.
  - *Ratio* provided it is quantitative with a unique and non-arbitrary zero point.

3.2.2. Derivation of Visualizations

Once the Data Profiling Model is completed, this information will be combined with the information coming from the User Requirements Model and it will result in the Visualization specification. This Visualization specification will enable us to derive the best type of visualization for each specification following the guidelines proposed in [32]. Furthermore, [14] explains how to transform the Visualization specification into a visualization following a Model Driven Architecture (MDA) standard. This approach also introduces a Data Visualization Model in order to facilitate the selection of the right visual analytics to non-expert users.

Therefore, at the end of this analysis, users will get a set of visualizations that will form a dashboard. This dashboard will enable non-expert users to understand the current status of their processes. Furthermore, by following our approach, the visualizations that make up this dashboard will be the most appropriate for each case and users will able to extract knowledge from them properly.

3.3. Instant Analysis

On the other hand, compared to historic analysis we have the type of analysis focused on a specific moment, usually when an event has occurred. The Instant analysis will process the data as soon it arrives with the aim of be anticipated to the incidents. Hence, we propose to use an Artificial Intelligence (AI) algorithm to enable users to make predictions in real-time.

Firstly, the AI algorithm is trained with the Historic Data Source, this action will enable the algorithm to learn about the correlation between the data source variables. Then, once the algorithm has been trained, when new data come into the process, the algorithm will be able to automatically and instantly predict what is going to happen.

Depending on the user's goals, one type of algorithm or another will be chosen. In the next section, we apply our approach into an illustrative case of study. In this case, we have applied a Deep Neural Network, which is able to predict attributes of data that are unknown at that time and generates additional information that can be interpreted in combination with the information facilitated through the historical analysis. This way, users have more information and better understanding of the context, allowing them to better to interpret the data for making decisions.

4. Case Study: Fire Department Calls for Service

In this section, we will apply our approach into a real case of study. The aim of this case of study is to demonstrate how a city that collects and offers its data can improve the sustainability of their processes thanks to the evidence gathered by means of Big Data techniques and visualizations.

The data provisioning process can be done from many different sources. In this case we have used Open Data Portals. The way to operate with Open Data is the same as with any other data source, with the advantage that there are no limits of use or publication. Open Data is defined as "that content that can be freely used, modified, and shared by anyone for any purpose" [33]. For this reason, we have taken as input the open dataset of San Francisco city [34]. More specifically, we took the Fire Department Calls for Service dataset [35]. This dataset contained 5.27 millions of rows and 34 columns by April 2020. Each row corresponds to a call to fire units. In order to improve sustainability, this city would require a set of visualizations to analyze their data in order to improve the response of their emergency services. We assume the role of a user that is trying to reduce the severity and number of serious fires in the city.

As is shown below, we will apply our approach to this case study by following our proposed process (Figure 2).

4.1. User Requirements Model and KPI Definition

Following our proposed process (Figure 2), the first element is the **User Requirements Model**. Figure 4 shows the result of its application. In this case, the user is the Fire Department Supervisor of the city of San Francisco. However, such user is not a specialist in visualization of Data Analytics. Therefore, the user is defined as *"Lay user"*. In this case, the user has decided to perform the analysis about the *"Service Calls Managment"* process. It helps to determine the scope of the analysis and what kind of goals will be pursued. Following the process, the strategic goal is defined as *"Reduce serious fires"*. Strategic goals are achieved by means of analyses that support the decision-making process.

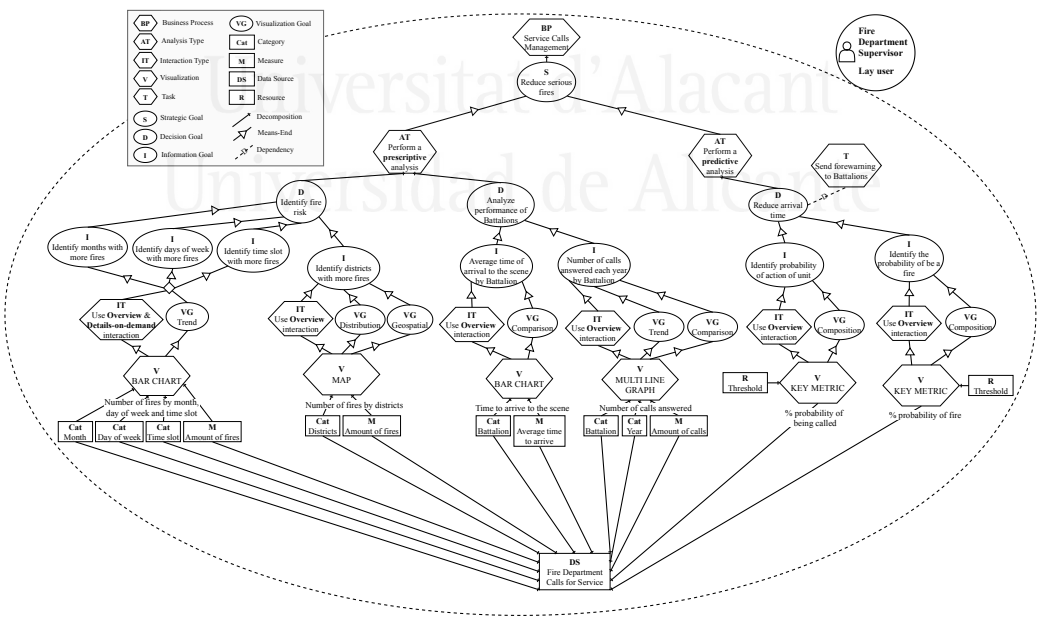


Figure 4. Application of our User Requirements Model to the Case Study.

Next, in this case study, the user has decided to create two types of analysis. One *"Prescriptive analysis"* in order to make an historical analysis and discover how to turn the city into a more sustainable

Smart City. Furthermore, the user also decides to create a *"Predictive analysis"* with the aim of predicting what is going to happen in order to be able to act more efficiently.

Each type of analysis is decomposed into decision goals. A decision goal aims to take appropriate actions to fulfill a strategic goal and it explains how it can be achieved. In the case of the **Prescriptive analysis**, the user defines *"Identify fire risk"* and *"Analyze the performance of Battalions"* as a decision goals. For each decision goal the user also specify one or more information goals to provide more detailed information, the information goals represent the lowest level of goal abstraction. To achieve the decision goal *"Identify fire risk"*, the user defines the information goals: *"Identify month with more fires"*, *"Identify days of week with more fires"*, *"Identify time slot with more fires"* and *"Identify districts with more fires"*. Besides, in order to achieve the decision goal *"Analyze the performance of Battalions"* the information goals defined by the user are *"Average time of arrival to the scene by Battalion"* and *"Number of calls answered each year by Battalion"*.

Otherwise, in order to make the **Predictive analysis**, the user defines *"Reduce arrival time'* as a decision goal, which is specified in detail by the information goals *" Identify probability of action of unit"* and *"Identify the probability of be a fire"*.

For each information goal a visualization will be automatically derived in order to achieve it (the visualizations used to achieve the information goals *"Identify month with more fires"*, *"Identify days of week with more fires"* and *"Identify time slot with more fires"* follow the same needs, for that reason they have been grouped in the model). A visualization is characterized by one or more visualization goals and kinds of interaction. The visualization goals describe which aspects of the data the visualization is trying to reflect and the interaction type describes how users would like to interact with the visualization. Users may use the guidelines that are published in [13] in order to define these elements. In this case the user has selected *"Trend"* ,*"Distribution"*, *"Geospatial"*, *"Comparison"* and *"Composition"* as visualization goals. And *"Overview"* and *"Details-on-demand"* as interaction types.

Finally, the user specifies the data source where the analysis will be performed over and selects the Categories and Measures that will populate the visualizations.

Once the User Requirements Model is done, users will define the KPIs through a table like Table 2. In this table the KPIs to measure the degree of achievement of the goals (for the goals which needs KPIs) are defined. Furthermore, it is necessary to identify the thresholds of each KPI in order to define its possible states and be able to decide whether the KPI is succeeding or failing.

Table 2. KPI Definition.

Goal	KPI	Threshold
Reduce serious fires	Number of serious fires	
Identify action probability of unit	% probability of being called	>70% high, 70% −40% medium, <40% low
Identify the probability of be a fire	% probability of fire	>70% high, 70% −40% medium, <40% low

In this case, the user has defined the number of serious fires as KPI in order to measure the strategic goal *"Reduce serious fires"*. For the information goal *"Identify action probability of unit"* a percentage to measure the probability that the unit will be called is proposed. The thresholds to define the states that this variable can take have been defined as *High probability* when there is a probability of action higher than 70%, *Medium probability* when the probability of action is between 70% and 40%, and finally, *Low probability*, when the probability of action is lower than 40%. For the case of the information goal *"Identify the probability of be a fire"* a percentage to measure it is also proposed. In this case, the threshold is defined as *High* when there is a probability of fire higher than 70%, *Medium* when the probability is between 70% and 40%, and *Low* when the probability is lower than the 40%.

4.2. Historical Analysis (Prescriptive)

Following our approach, in order to perform the historical analysis, the next step is to apply our **Data Profiling Model** (Section 3.2.1) to determine the Dimensionality, Cardinality and Data Type of the selected data source.

For example, the visualization **"Identify districts with more fires"** (from the model shown in Figure 4) requires information about the category *"Districts"* and the measurement *"Amount of fires"*.

Firstly, using the data profiling tool (Section 3.2.1), the independent variable *"Districts"* is classified as *Nominal* and the dependent variable *"Amount of fires"* as *Ratio*. Dimensionality is set to 2-*dimensional*, because the user has defined 2 variables to visualize. Finally, the Cardinality is defined as *High* Cardinality because the data contains many items to represent.

Overall, the visualization specification obtained through **User Requirements Model** and **Data Profiling Model** is:

- **Visualization Goal**: Distribution
- **Visualization Goal**: Geospatial
- **Interaction**: Overview
- **User**: Lay
- **Dimensionality**: 2-dimensional
- **Cardinality**: High
- **Independent Type**: Nominal
- **Dependent Type**: Ratio

With this visualization specification we are able to derive it into the most suitable visualization type following the approach proposed in [32]. As we specify in Section 3.2.2, the explanation of how to do this process in an automatic manner is covered in [14]. In this case, the visualization type that better fits this specification is a **"MAP"**. This whole process is repeated with the rest of the visualizations that compose the model shown Figure 4 in order to derive the most suitable visualization type for each specification.

4.3. Predictive Analysis

However, not all the necessary information can be extracted from historical analysis. In order to perform the **Predictive analysis**, we will face a real-time scenario, where the user wants to predict in real-time the *"Probability of action of unit"* and *"Probability of be a fire"*, as it is specified in Figure 4.

In order to predict events, an AI algorithm will be integrated into the system. This algorithm will analyze each incoming call and will make a prediction before be answer with the aim of send a forewarning to the Battalions when they have a high probability of action.

The AI algorithm will be trained with the Historic Data Source. This algorithm will learn about which are the variables in the data source that determine the type of call and the probability of action of the Battalions. The reason why we use the complete data source is because, in the User Requirements Model (Figure 4), the user has selected the entire data source to feed the visualizations of the Predictive Analysis.

In this case, as is shown in Figure 5, we have used a dense Deep Neural Network (DNN) with 6 layers and (input, output and 4 hidden ones). In order to avoid the problems that data bias brings, as we explain in [36], first step is to shuffle the whole dataset and separate it into Train, Validation and Test data. In this case 70% of the data is used to train the algorithm, 20% to validate it and the 10% to test it. The train and validation data will form the input of the algorithm, these data will be used by the algorithm to learn how to classify. Then, test data will be use to check the learning result.

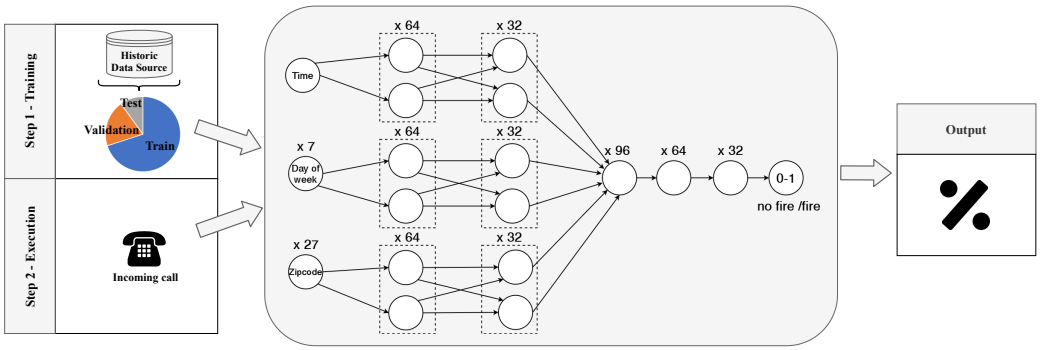


Figure 5. Application of the AI algorithm to predict the probability that a call was a fire.

In this case, as Figure 5 shows, the algorithm input will be formed by the variables *"Time"*, *"Day of week"* and *"Zipcode"*. Each variable will have a set neurons to represent each of its types of attributes. More specifically, the Zipcode will be represented by using a one-hot [37] encoding into 27 input neurons. In the same sense, the Day of Week will codified into 7 neuron. Besides, since the Time is not categorical, it will be represented by using one neuron.

Then, a layer with 64 neurons by each variable will be launched. Following, these neurons will be mixed with a layer of 32 neurons by variable. Finally, all these neurons will be grouped into a 96 neurons, them will be reduced to 64 neurons, to 32 neurons, and the output will be a binary neuron. This is the most complex task and the one that will need more computational capacity.

Once the algorithm has been trained and tested, when new data come into the process, as a new call, the algorithm will be able to automatically and instantly predict information about the process. The output of this algorithm will be a number between 0 and 1, where a number close to 0 will means that there is no fire risk and a number close to 1 will mean that there is high fire risk. This information will be represented in a Key Metric visualization as Figure 4) indicates. These visualizations are composed by resources that specify the threshold values. These thresholds has been previously defined in Table 2. Therefore, these visualizations will be created by taking into account these specific thresholds and will enable users to visually understand the algorithms output.

Moreover, beside to visually represent the probability of action, in order to achieve the decision goal **"Reduce arrival time"**, the task *"Send forewarning to Battalions"* will be executed. Therefore, those Battalions with a higher probability to be called will be forewarned. As a result, they will be able to be ready for the action and in this way, the reduction of the arrival time will consecutively reduce the impact of the fire.

Therefore, by applying this AI algorithm, when a call comes in, users will know automatically extra information that has not yet been received and the probability of this information to be true. Furthermore, our approach provides visual techniques to easily understand the algorithm output and allows users to define thresholds in order to distinguish when the result of the algorithm is relevant to them. These visualizations will be grouped into a dashboard that will be updated with each call, thus always offering the most up-to-date information.

4.4. Final Dashboard

Once all visualizations have been developed following the recommended visualization types proposed in Figure 4, a dashboard like the one shown in Figure 6 will be generated. The dashboard will combine all the generated visualizations aimed at satisfying the analytic requirements of our fire department supervisor user. The visualizations that make up the dashboard are grouped following the types of analysis defined in the User Requirements Model (Figure 4).

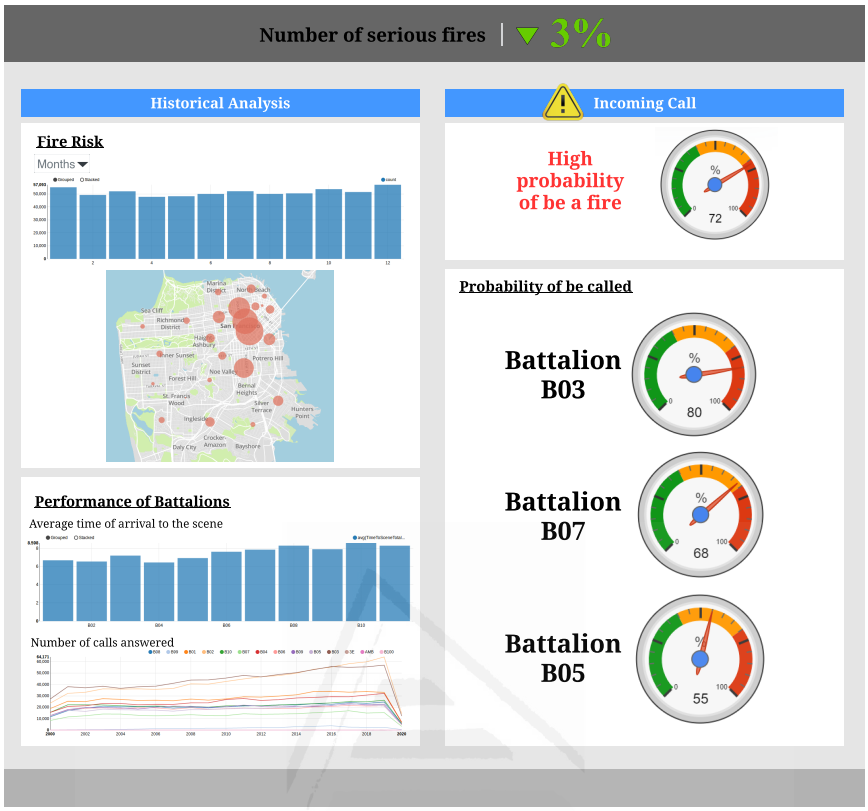


Figure 6. Dashboard for Service Calls Management analysis.

This dashboard represents the overall analysis of the Service Calls Management process. On the left side, the historical analysis is represented. Here the user is able to identify the fire risk at a certain moment as well as the fires distribution through the different districts. Furthermore, the user may also analyze the performance of the Battalions thought the average time to arrival to a scene and the number of calls answered by each Battalion.

On the other hand, on the right side, it will be represented the key metrics that will made possible to analyze each incoming call before answering it. These visualizations will be updated for each incoming call and they will represent the probability for the call to be a fire, as well as the probability for the Battalions to be called for action. Moreover, as is defined in Figure 4, an automatic forewarning will be send to the Battalion with the highest probability of action, in order to enable them to be ready for action, thus reducing the arrival time to the scenarios.

Finally, at the top of the dashboard there is a key metric that measures the strategic goal "*Reduce serious fires*". This KPI will enable users to follow the impact of their decisions and check whether the measures in place are really improving the sustainability of the city.

5. Evaluation

In order to evaluate the impact of our proposal, we have performed an experiment, with non-expert users in data visualization. The experiment consisted in filling a questionnaire with two exercises. On each exercise, a dataset was presented to users and they were asked to define a series of visualizations to perform one type of analysis.

In the first exercise, users were asked to define a set of visualizations to perform an Historical Analysis without following any methodology. While, in the second exercise, users were tasked to perform an Instant Analysis by following our proposed methodology. The tasks posed were

interchanged between two questionnaire models, and distributed equally among the participants as is shown in Table 3.

Table 3. Experiment design.

	Historical Analysis	**Instant Analysis**
With Methodology	Group A	Group B
Without Methodology	Group B	Group A

Once both exercises were finished, they were asked more concrete questions that had to be answered by using the created visualizations.

We took as non-expert users in data visualization software developers and non-IT users from a company with experience in joint projects with public institutions from Alicante (Spain) who performed the experimental task from home. A total of 12 non-expert participants filled the questionnaires. Since we were unable to perform a fully controlled experiment due the COVID-19 measures, we used a video conference tool to have direct contact with the participants. During the session, there were no dropouts. Table 4 shows some statistics about the experimental sessions.

Table 4. Participant's information.

Number of Participants	12
Female participants	25%
Average age	31.5
Regular use of visualization tools	33.3%
Unusual use of visualization tools	66.7%

According to the results obtained, the set of visualizations proposed without following any methodology were able to answer the 23% of the specific questions posed, whereas this number rose to 39% coverage when following our proposed methodology.

A T-test of the results showed statistical significance for the results obtained, confirming the impact of the method proposed.

Threats to the Validity

In this section, we summarize the main limitations we envision for our approach.

- The subjects were randomly assigned within the groups to cancel out fatigue effects. However, the experiment could not be fully controlled due to the COVID-19 measures. Nevertheless, we tried to overcome this shortcoming by using a real-time remote meeting tool.
- The lack of a CASE tool forces users to be accompanied by a data analyst in order to follow the proposed method, making it necessary to implement a user-friendly CASE tool to overcome this limitation.
- The methodology increases the capability to answer questions, however, domain expertise can still be considered a significant factor to define a more complete set of visualizations.
- In principle, our proposal is context-independent. However, since we have not yet tested the proposal in a comprehensive enough set of contexts, it may be the case that some specific user profiles have not yet been identified.

6. Conclusions and Future Work

Nowadays, the cities we live on are transforming into Smart Cities, incorporating sensors and other devices that generate large volumes of IoT data. The different nature of these data makes it complex to process, combining a variety of sources at different speeds that require the application of Big Data techniques. With the aim of solving these difficulties, a technological revolution emerges, with potential to boost a new industry and to generate numerous business opportunities and helping to improve the sustainability of our cities by improving their efficiency and resource management.

However, resource optimization cannot be achieved unless decision makers -city representatives- obtain a clear view of the information processed, in turn having the necessary evidence to make the correct decisions.

In this paper, we have proposed an approach that improves the evidence-gathering process in the decision making through visualization techniques for Big Data in the context of Smart Cities. We have combined our methodology to help users define their goals and derive the best type of visualization with the possibility of including real-time data in the context of Smart Cities. These visualizations represent incoming real-time data from IoT sensors in order to enable users to gather the necessary evidence to make strategic and tactical decisions.In addition, they help understand the output of AI algorithms, providing the necessary context to generate trust regarding the output. The set of visualizations created takes into account the expertise of users, facilitating their understanding and their translation into actions through decision making. Without the holistic view about the process and its outputs provided by our proposal, it would be hard for the users to understand the state in which their processes are, and therefore there would not be enough evidence for making the correct decision.

In order to assess the suitability of our proposal, we presented a case study based on fire department's call service, where data coming from IoT devices and incoming calls are analyzed. In this particular case, our findings are applied to data coming from citizen calls. Thus, the results of this work contribute to the optimization of resources by facilitating evidence for the decision making process. Namely, fire extinguishing battalions will be better suited to respond when a call is received.

Our proposal has been evaluated through an experiment with 12 users without expertise in data visualization. The results show that the visualizations obtained by our proposal are considered better by users than the ones they created on-demand and, in fact, support better their information needs.

As part of our future work, we are working on extending the case study by applying it in other Smart City processes with the aim of helping Smart Cities to improve the sustainability in all their processes. Moreover, we plan to perform an evaluation of the understandability of the models created with our methodology, similar to the one presented in [38]. Finally, we intend to develop a CASE tool in order to ease the usage of our methodology. This this aim, we will follow the guidelines established in [39].

Author Contributions: Conceptualization, A.L., M.A.T., A.M. and J.T.; methodology, A.L., M.A.T. and A.M.; investigation, A.L., M.A.T. and A.M.; writing–original draft preparation, A.L. and M.A.T.; writing–review and editing, A.L., M.A.T., A.M. and J.T.; visualization, A.L., M.A.T. and A.M.; supervision, M.A.T., A.M. and J.T.; funding acquisition, J.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been co-funded by the ECLIPSE-UA (RTI2018-094283-B-C32) project funded by Spanish Ministry of Science, Innovation, and Universities and the DQIoT (INNO-20171060) project funded by the Spanish Center for Industrial Technological Development, approved with an EUREKA quality seal (E!11737DQIOT). Ana Lavalle holds an Industrial PhD Grant (I-PI 03-18) co-funded by the University of Alicante and the Lucentia Lab Spin-off Company.

Acknowledgments: We would like to thank the Lucentia Lab Spin-off Company for providing us with the algorithms and the data necessary for the application of our proposal.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dameri, R.P. Searching for smart city definition: A comprehensive proposal. *Int. J. Comput. Technol.* **2013**, *11*, 2544–2551. [CrossRef]
2. Ahvenniemi, H.; Huovila, A.; Pinto-Seppä, I.; Airaksinen, M. What are the differences between sustainable and smart cities? *Cities* **2017**, *60*, 234–245. [CrossRef]
3. ET Bureau. Data is the 21st Century's Oil, says Siemens CEO Joe Kaeser. 2018. Avalaible online: https://economictimes.indiatimes.com/magazines/panache/data-is-the-21st-centurys-oil-says-siemens-ceo-joe-kaeser/articleshow/64298125.cms?from=mdr (accessed on 15 April 2020).
4. Jin, J.; Gubbi, J.; Marusic, S.; Palaniswami, M. An information framework for creating a smart city through internet of things. *IEEE Internet Things J.* **2014**, *1*, 112–121. [CrossRef]
5. Li, C.; Dai, Z.; Liu, X.; Sun, W. Evaluation System: Evaluation of Smart City Shareable Framework and Its Applications in China. *Sustainability* **2020**, *12*, 2957. [CrossRef]
6. Wu, Y.C.; Sun, R.; Wu, Y.J. Smart City Development in Taiwan: From the Perspective of the Information Security Policy. *Sustainability* **2020**, *12*, 2916. [CrossRef]
7. Rao, A.S.; Verweij, G. Sizing the Prize What's the Real Value of AI for Your Business and How Can You Capitalise? Avalaible online: https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf (accessed on 15 April 2020).
8. Nikitas, A.; Michalakopoulou, K.; Njoya, E.T.; Karampatzakis, D. Artificial Intelligence, Transport and the Smart City: Definitions and Dimensions of a New Mobility Era. *Sustainability* **2020**, *12*, 2789. [CrossRef]
9. Nations, U. *World Population Prospects: The 2017 Revision, Key Findings and Advance Tables*; United Nations: New York, NY, USA, 2017.
10. Villanueva, F.J.; Aguirre, C.; Rubio, A.; Villa, D.; Santofimia, M.J.; López, J.C. Data stream visualization framework for smart cities. *Soft Comput.* **2016**, *20*, 1671–1681. [CrossRef]
11. Neirotti, P.; De Marco, A.; Cagliano, A.C.; Mangano, G.; Scorrano, F. Current trends in Smart City initiatives: Some stylised facts. *Cities* **2014**, *38*, 25–36. [CrossRef]
12. Capdevila, I.; Zarlenga, M. Smart city or smart citizens? The Barcelona case. *Barc. Case* **2015**. [CrossRef]
13. Lavalle, A.; Maté, A.; Trujillo, J.; Rizzi, S. Requirements-Driven Visualizations for Big Data Analytics: A Model-Driven. In Proceedings of the 27th IEEE International Requirements Engineering Conference, RE 2019, Jeju Island, Korea, 23–27 September 2019; pp. 109–119. [CrossRef]
14. Lavalle, A.; Maté, A.; Trujillo, J. Requirements-Driven Visualizations for Big Data Analytics: A Model-Driven Approach. In Proceedings of the Conceptual Modeling—38th International Conference, ER 2019, Salvador, Brazil, 4–7 November, 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 78–92. [CrossRef]
15. Yigitcanlar, T.; Kamruzzaman, M. Planning, development and management of sustainable cities: A commentary from the guest editors. *Sustainability* **2015**, *7*, 14677–14688. [CrossRef]
16. Caragliu, A.; Del Bo, C.; Nijkamp, P. Smart cities in Europe. *J. Urban Technol.* **2011**, *18*, 65–82. [CrossRef]
17. Theodoridis, E.; Mylonas, G.; Chatzigiannakis, I. *Developing an Iot Smart City Framework*; IEEE: New York, NY, USA, 2013; pp. 1–6. [CrossRef]
18. Arasteh, H.; Hosseinnezhad, V.; Loia, V.; Tommasetti, A.; Troisi, O.; Shafie-Khah, M.; Siano, P. Iot-based smart cities: A survey. In Proceedings of the 2016 IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC), Florence, Italy, 7–10 June 2016; pp. 1–6. [CrossRef]
19. Guo, K.; Lu, Y.; Gao, H.; Cao, R. Artificial intelligence-based semantic internet of things in a user-centric smart city. *Sensors* **2018**, *18*, 1341. [CrossRef] [PubMed]
20. Cicirelli, F.; Guerrieri, A.; Spezzano, G.; Vinci, A. An edge-based platform for dynamic smart city applications. *Future Gener. Comput. Syst.* **2017**, *76*, 106–118. [CrossRef]
21. Chamoso, P.; De La Prieta, F. Swarm-based smart city platform: A traffic application. *ADCAIJ Adv. Distrib. Comput. Artif. Intell. J.* **2015**, *4*, 89–98. [CrossRef]
22. Ghosh, D.; Chun, S.A.; Shafiq, B.; Adam, N.R. Big data-based smart city platform: Real-time crime analysis. In Proceedings of the 17th International Digital Government Research Conference on Digital Government Research, New York, NY, USA, 8 June 2016; pp. 58–66.
23. Villanueva, F.J.; Aguirre, C.; Villa, D.; Santofimia, M.J.; López, J.C. Smart City data stream visualization using Glyphs. In Proceedings of the 2014 Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, Birmingham, UK, 2–4 July 2014; pp. 399–403. [CrossRef]

24. Prandi, F.; Soave, M.; Devigili, F.; Andreolli, M.; De Amicis, R. Services oriented smart city platform based on 3D city model visualization. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2014**, *2*, 59. [CrossRef]
25. Phupattanasilp, P.; Tong, S.R. Augmented Reality in the Integrative Internet of Things (AR-IoT): Application for Precision Farming. *Sustainability* **2019**, *11*, 2658. [CrossRef]
26. Braem, B.; Latré, S.; Leroux, P.; Demeester, P.; Coenen, T.; Ballon, P. Designing a smart city playground: Real-time air quality measurements and visualization in the City of Things testbed. In Proceedings of the 2016 IEEE International Smart Cities Conference (ISC2), Trento, Italy, 12–15 September 2016; pp. 1–2. [CrossRef]
27. Tory, M.; Moller, T. Human factors in visualization research. *IEEE Trans. Vis. Comput. Graph.* **2004**, *10*, 72–84. [CrossRef]
28. Teruel, M.A.; Maté, A.; Navarro, E.; González, P.; Trujillo, J.C. The New Era of Business Intelligence Applications: Building from a Collaborative Point of View. *Bus. Inf. Syst. Eng.* **2019**, *61*, 615–634. [CrossRef]
29. Dalpiaz, F.; Franch, X.; Horkoff, J. iStar 2.0 Language Guide. *arXiv* **2016**, arXiv:1605.07767.
30. Maté, A.; Trujillo, J.; Franch, X. Adding semantic modules to improve goal-oriented analysis of data warehouses using I-star. *J. Syst. Softw.* **2014**, *88*, 102–111. [CrossRef]
31. Shi-Nash, A.; Hardoon, D.R. Data analytics and predictive analytics in the era of big data. *Internet Things Data Anal. Handb.* **2017**, 329–345. [CrossRef]
32. Golfarelli, M.; Rizzi, S. A model-driven approach to automate data visualization in big data analytics. *Inf. Vis.* **2020**, *19*. [CrossRef]
33. Open Knowledge Foundation. The Open Definition. Available online: https://opendefinition.org/ (accessed on 15 April 2020).
34. San Francisco Open Data Portal. 2019. Available online: https://datasf.org/opendata/ (accessed on 15 April 2020).
35. Fire Department Calls for Service Dataset. 2019. Available online: https://data.sfgov.org/Public-Safety/Fire-Department-Calls-for-Service/nuek-vuh3 (accessed on 15 April 2020).
36. Lavalle, A.; Maté, A.; Trujillo, J. An Approach to Automatically Detect and Visualize Bias in Data Analytics. In Proceedings of the 22nd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data co-located with EDBT/ICDT 2020 Joint Conference, Copenhagen, Denmark, 30 March–2 April 2020; pp. 84–88.
37. Harris, S.; Harris, D. *Digital Design and Computer Architecture: Arm Edition*; Morgan Kaufmann: Burlington, MA, USA, 2015.
38. Teruel, M.A.; Navarro, E.; López-Jaquero, V.; Montero, F.; Jaen, J.; González, P. Analyzing the understandability of Requirements Engineering languages for CSCW systems: A family of experiments. *Inf. Softw. Technol.* **2012**, *54*, 1215–1228. [CrossRef]
39. Teruel, M.A.; Navarro, E.; López-Jaquero, V.; Montero, F.; González, P. A CSCW Requirements Engineering CASE Tool: Development and usability evaluation. *Inf. Softw. Technol.* **2014**, *56*, 922–949. [CrossRef]

Capítulo 7

Fostering Sustainability through Visualization Techniques for Real-Time IoT Data: A Case Study Based on Gas Turbines for Electricity Production

Lavalle, A., Teruel, M. A., Maté, A., & Trujillo, J. (2020). Fostering Sustainability through Visualization Techniques for Real-Time IoT Data: A Case Study Based on Gas Turbines for Electricity Production. *Sensors, 20*(16), 4556.

Factor de Impacto: **3,275**
Clasificación JCR: **Q1 (15/64 Instruments & instrumentation)**

Disponible en:
DOI: `https://doi.org/10.3390/s20164556`

sensors

MDPI

Article

Fostering Sustainability through Visualization Techniques for Real-Time IoT Data: A Case Study based on Gas Turbines for Electricity Production

Ana Lavalle [1,2,*], Miguel A. Teruel [1,2], Alejandro Maté [1,2] and Juan Trujillo [1,2]

[1] Lucentia Research, DLSI, University of Alicante, Carretera San Vicente del Raspeig s/n, 03690 Alicante, Spain; materuel@dlsi.ua.es (M.A.T.); amate@dlsi.ua.es (A.M.); jtrujillo@dlsi.ua.es (J.T.)

[2] Lucentia Lab, Avda. Pintor Pérez Gil, N-16, 03540 Alicante, Spain

* Correspondence: alavalle@dlsi.ua.es

Received: 9 July 2020; Accepted: 13 August 2020; Published: 14 August 2020

Abstract: Improving sustainability is a key concern for industrial development. Industry has recently been benefiting from the rise of IoT technologies, leading to improvements in the monitoring and breakdown prevention of industrial equipment. In order to properly achieve this monitoring and prevention, visualization techniques are of paramount importance. However, the visualization of real-time IoT sensor data has always been challenging, especially when such data are originated by sensors of different natures. In order to tackle this issue, we propose a methodology that aims to help users to visually locate and understand the failures that could arise in a production process.This methodology collects, in a guided manner, user goals and the requirements of the production process, analyzes the incoming data from IoT sensors and automatically derives the most suitable visualization type for each context. This approach will help users to identify if the production process is running as well as expected; thus, it will enable them to make the most sustainable decision in each situation. Finally, in order to assess the suitability of our proposal, a case study based on gas turbines for electricity generation is presented.

Keywords: Internet of Things; data visualization; Big Data analytics; sustainable production; gas turbines; Artificial Intelligence

1. Introduction

Global energy consumption is increasing on a daily basis [1,2]. New lifestyle trends are increasing the need for electricity generation. In order to cope with this ever-growing need, a sustainable energy production process is required [3]. In this sense, one approach to aiding the sustainability of energy production is to exploit the potential of the Internet of Things (IoT). The adoption of IoT by industry has led to highly sensorized machinery [4]. Thus, thanks to the data provided by these sensors, it is possible to better understand how an electricity production process is performing, and thus it is possible to take actions aimed at improving the throughput and sustainability of the whole process [5].

The introduction of Artificial Intelligence (AI) processing data provided by sensors has enabled the determination of whether a generation process is running as well as expected [6]. Indeed, Predictive Machine Learning can be applied in order to assess whether or not machinery may fail in the near future [7]. Nevertheless, such techniques are often based on the usage of neural networks, whose input is usually the general status (or a subset) of the whole system (i.e., tuples of the data generated from all the system's sensors) [8]. Thus, since neural networks act as a black box, it is unlikely that they can provide information regarding the part of the system which is going to cause the predicted failure [9].

However, even if the output of the neural network can only determine whether the process is going to fail or not, the information of the production process can be complemented with visual

details regarding the evolution of the machinery sensors. Thanks to these visualizations, machinery operators can identify abnormalities in certain parts of the system, enabling them to identify certain problems which could not be detected otherwise. Still, the creation of these visualizations is not trivial; the large volume of sensor data produced across multiple magnitudes makes it challenging to present the necessary information to users without making it overbearing.

Therefore, in order to make such visualizations possible, we propose a new methodological approach to monitor industrial machinery using an IoT-based visualization technique. The main goal of this work is to help non-expert users in data visualization to visually locate and understand the failures that could arise in a production process, thus enabling them to make the most sustainable decision in each situation.

In previous works [10,11], we defined a model that helped users to specify and achieve their goals. It extracted the characteristics of the data sources and automatically derived the best type of visualizations according to the defined context. Moreover, in [12], we have published an approach that is focused on the context of Smart Cities; in this work, we proposed a methodology, based on visualization techniques, with the aim of improving the evidence-gathering process by assisting users in their decision making in the context of Smart Cities.

In this paper, we have significantly improved and complemented our previous works as follows: (i) we show how our proposal can be applied to the context of industrial machinery, (ii) we broaden our proposed metamodel by adding new elements to make it suitable for real-time scenarios, and (iii) we provide a novel methodology to monitor industrial machinery, which is divided into two phases (the first phase is performed before runtime, and the second phase is executed at runtime). This methodology (a) to define the goals and requirements of the production process, (b) automatically derives the most suitable visualization type for each context, (c) helps users to visually understand the output of Artificial Intelligence models and (d) provides visualizations to help users to make the most sustainable decision in each situation.

Furthermore, in order to test and show the applicability of our proposal, we have presented a case study based on gas-based electricity generation turbines. Gas turbines are large machines that can be heavily sensorized. A picture of a gas turbine can be seen in Figure 1. The gas turbine in our case study includes 80 sensors from which data are gathered at runtime. The complexity of the data, the speed at which data are generated and the importance of detecting failures make this a perfect scenario to test how our approach improves the sustainability of the process; i.e., how it improves the performance of the process by preventing the breakdown of the machines. The results of this case study confirm that our proposal helps to improve the sustainability of the process.

The advantages of our proposal are that (i) it enables users to monitor the quality of the systems, (ii) it aids in preventing the breakdown of the machines, (iii) it helps to identify if the production process is running as well as it was expected, and (iv) it helps users to understand and co-relate the outputs of an AI engine. Without the benefits introduced by our proposal, users would find it more difficult to determine the optimality of the execution of the production process. Moreover, it would be difficult for them to identify whether stopping production is the most sustainable decision.

The rest of the paper is structured as follows. Section 2 presents the related work in this area. Section 3 describes our proposed methodology for fostering sustainability through visualizations. Section 4 shows our approach, which is applied in a real case study for electricity generation based on gas turbines. Section 5 summarizes the limitations of our work. Finally, Section 6 summarizes the conclusions and sketches our future work.

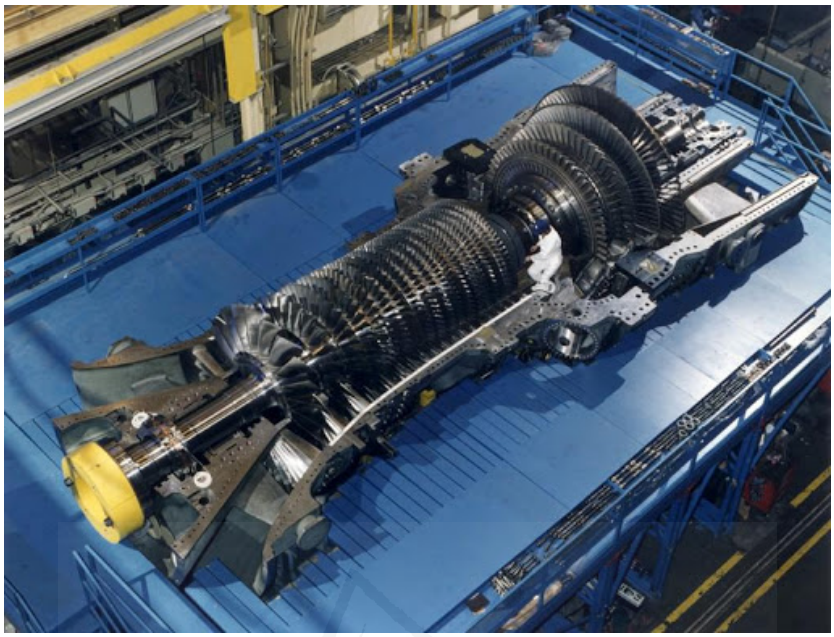


Figure 1. Gas turbine.

2. Related Work

The importance of the sustainable development in industry is increasing. In 1992, the concept of sustainable production emerged at the United Nations Conference on Environment and Development [13]. There, it was determined that the main cause of the deterioration of the global environment is the unsustainable pattern of consumption and production, especially in industrialized countries.

The sustainability strategy includes indicators giving a measurable overview of trends and involves action by all sectors, especially industrial systems. This sector should play an important part in the attainment of sustainability goals [14]. The set of strategic metrics for assessing sustainability includes [15] (i) reflecting the status of a system, (ii) providing early warning information, (iii) anticipating future conditions and trends, (iv) comparing across places and situations and (v) highlighting what is happening in a large system.

In [16], a new methodology was presented to promote and measure sustainable production in business. The authors proposed 22 indicators and provided guidance to select additional, production-specific indicators.

As [17] argues, visualizations may help in making energy-saving management decisions. A visualization of the incoming data can provide insights. However, visualizing big data in real-time is a challenge itself. The growth of the Internet of Things (IoT) means that the amount of available real-time data is increasing rapidly; therefore, the development of analysis programs for IoT platforms is a complex task [18].

Cyber-physical systems are successful in various scientific communities, specifically regarding production issues [19]. The industry represents a rich data environment, and increasingly large volumes of data are constantly being generated by its processes. However, only a relatively small portion of the data is actually exploited by manufacturers [20].

Several works have focused on IoT visualization. For example, in [17], a platform is proposed to transform sensor data to context-based visualized data. One sector in which the visualization of IoT sensors is used is in the Smart Cities domain. These systems generate massive amounts of data that can be analyzed and visualized to better understand people's dynamics [21]. Another sector is healthcare: the visualization of data, metadata and sensor networks is becoming one of the most

important aspects of the health monitoring process [22]. In [23], an intelligent healthcare framework based on IoT technology is proposed, providing ubiquitous healthcare to users during their workout sessions. In [24], the authors propose an ambient intelligence environment for cognitive rehabilitation at home, combining physical and cognitive activities. They implement a Fuzzy Inference System in which smart sensors and actuators attempt to compensate for the absence of the therapist.

The visualization of a large data set is a demanding task. The traditional manners of presenting data face a few limitations as the amount of data grows constantly. In [25], the authors identified challenges in big data visualizations, such as perceptual scalability, real-time scalability and interactive scalability. They argue that visualization tools and techniques are able to help users in the identification of missing, erroneous or duplicate values.

The authors in [26] contribute methods for the visualization of big data in real-time. They present techniques to address perceptual and interactive scalability, following the principle that scalability should be limited by the chosen resolution of the visualized data, rather than the number of records. In [20], an Intelligent Data Analysis and Real-Time Supervision (IDARTS) framework is proposed that combines distributed data acquisition, machine learning and run-time reasoning to assist in fields such as predictive maintenance and quality control. The goal of their framework is to allow manufacturers to translate their data into a business advantage.

In [18], the authors present I^2—an interactive development environment that coordinates running cluster applications and corresponding visualizations, where only the currently depicted data points are processed and transferred. They present a model for the real-time visualization of time series and show how cluster programs can adapt to changed visualization properties at runtime to enable interactive data exploration on data streams. Additionally, [27] presented Hashedcubes—a data structure for answering queries from interactive visualization tools that explores and analyzes large, multidimensional datasets. This enables the real-time visual exploration of large datasets with low memory requirements and low query latencies.

The aforementioned works highlight the importance of the use of visualizations in IoT scenarios. On the other hand, other works such as [28] highlight the importance of fault detection and isolation in safety-critical systems, such as gas turbine engines. They discuss the necessity of a decision-support system to prescribe corrective actions so that the system can continue to function without jeopardizing the safety of the personnel and equipment involved. The authors [28] propose the use of Self-Organizing Maps (SOM) in order to visually explore the data in a two-dimensional space, understand the nature of the input signal and gain insights into the difficulty of the fault classification task. SOM transforms complex, nonlinear relationships between high-dimensional data into topological relationships in a low-dimensional space.

Other works, such as [29], visualize turbulent flow behavior between turbines in a physical space and allow the viewer to see intricate vortex-blade intersection configurations in a static-blade view. In [30], examples of the implementation of optical techniques employed to visualize flow structure, fuel spray patternation, liquid fuel penetration and combustion species are presented.

In [31], an OSRDP architecture framework for sustainable manufacturing is proposed. The authors propose a system that is capable of processing massive sensor data efficiently when the amount of sensors, data and devices increases. The system uses data mining based on Random Forest to predict the quality of products. However, the proposed system classifies sensors as normal/abnormal on an individual basis; it does not take into account problems that are only reflected by the readings of the system as a whole. Moreover, it does not analyze which visualizations would be most adequate to troubleshoot the underlying problems, making it more difficult to make adequate decisions for their correction.

One of the core benefits of visualizations is that it enables people to discover visual patterns that might otherwise be hidden [32]. However, it is very important to be mindful of which types of visualizations are used in each context. Not all types of visualizations are suitable for visually detecting anomalies; as [32] discusses, it is possible to create visualizations that seem "plausible"

(design parameters are within normal bounds and pass the visual sanity check) but hide crucial data features.

As we have shown, different approaches highlight the importance and challenges of visualizing real-time data from IoT systems. Other approaches highlight the importance of systems that detect and predict failures in order to achieve sustainable production. However, none of the approaches listed above provide a complete methodology that captures information from an IoT system in order to predict when the system may potentially fail and enables users to make the most sustainable decision with the aid of real-time visualizations.

Therefore, we propose a methodology that chooses the best type of visualization based on users' analytical needs. Moreover, visual techniques are provided so that users can understand the output of Artificial Intelligence models. This will enable users to monitor the quality of the systems and to make the most sustainable decision in each situation.

3. Methodology to Foster Sustainability through Visualizations

Once the related work has been presented, this section will describe our methodology. The main aim of our proposal is to help users to visually locate and understand the failures that could arise in a production process. Our methodology includes two phases. Phase 1 is the setup phase, performed before production (runtime). In this phase, users define the goals and requirements of the production process; this information is used to generate the best suited visualizations. Phase 2 is executed during the production process (at runtime). In this phase, the production process is monitored with the objective of aiding users in making the most sustainable decisions. In the following, we describe these two proposed phases in detail.

3.1. Phase 1—Definition of Goals and Visualizations

As mentioned above, Phase 1 is executed prior to the production process. The objective of this phase is for users to define the goals that they are aiming to achieve during the production process. Therefore, the most proper type of visualization to achieve these goals will be automatically derived. These visualizations, defined in the pre-production process, will be used to detect and monitor failures in the production process. In this sense, we ensure that the visualizations shown to users are the most suitable to meet their goals and help them to make decisions about the production process.

Figure 2 summarizes the process followed in Phase 1, which defines visualizations. Firstly, users create a User Requirements Model aided by a sequence of guidelines published in [10]. This model guides non-expert users to capture their analytical needs. Furthermore, through this User Requirements Model, users define, among others, which elements of the data source they wish to represent in the visualizations. Complementary to this model, a Data Profiling Model [10] is obtained by analyzing the features of the data sources to be visualized in a semi-automatic manner.

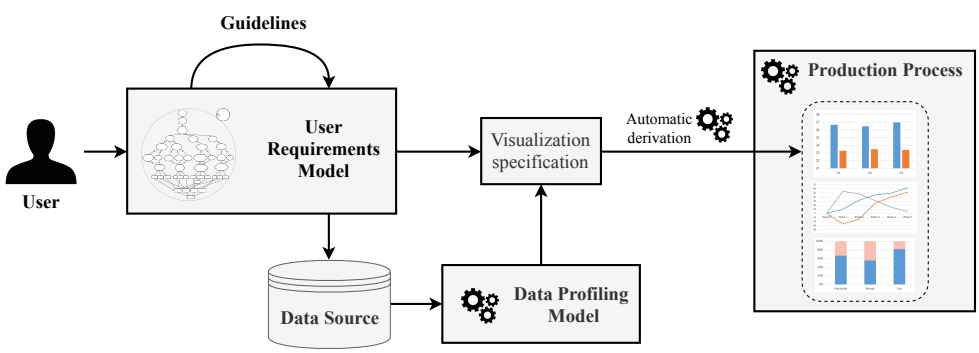


Figure 2. Phase 1—visualization definition process.

Once both models have been obtained, they are translated into a visualization specification. Following [33], we are able to derive the visualization specification into the most suitable visualization to achieve each specified goal in an automated manner.

These generated visualizations are introduced in the production process at each defined moment. Therefore, users will be able to monitor the production and make decisions more accurately based on the visualizations. In the following, we describe the elements included in the visualization definition orocess.

3.1.1. User Requirements Model

Our approach starts from a User Requirements Model that guides non-expert users towards the definition of specific visualizations that they would need to achieve their data analysis objectives. It is possible to find an example of the User Requirements Model applied to a real case in Section 4.

In order to formally define our novel model, we propose a metamodel (see Figure 3). This metamodel is an extension of the model used for social and business intelligence modeling [34], namely i* [35] and the i* for data warehouses extension [36]. It is worth noting that i* has already been extended and used to model other real-time IoT-enabled domains [37].

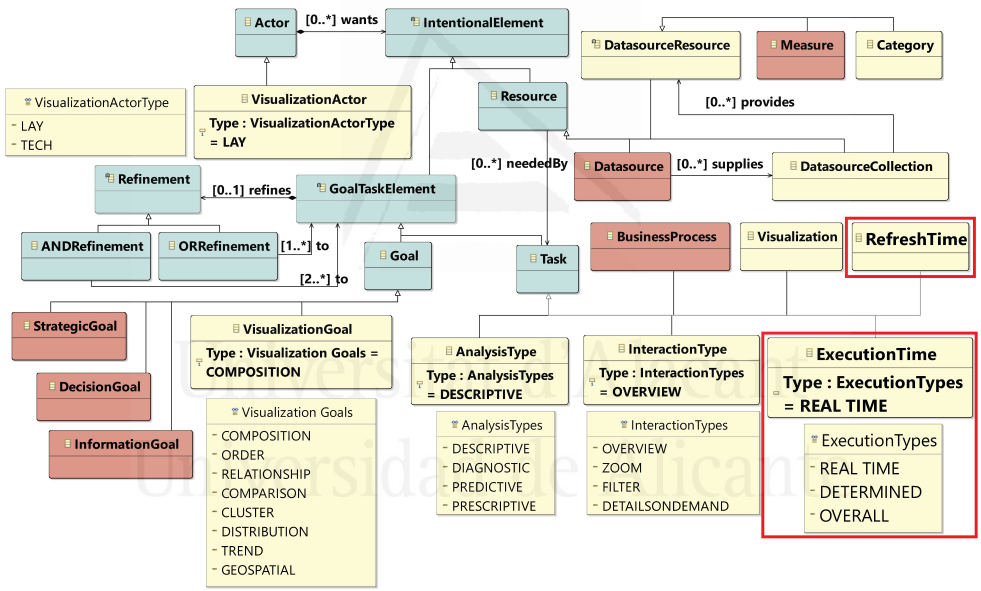


Figure 3. User Requirements Metamodel.

In Figure 3, elements from i* are represented in blue, elements from i* for data warehouses are represented in red and the elements added in our proposal are in yellow, including the new elements introduced to work with real-time scenarios (represented within a red square). In the following, we will describe the elements of the metamodel.

The user of the system is represented with the visualization actor element. We can find two types of visualization actors: lay, when the user is not expert in complex data visualizations, or tech, when the user has experience in data visualization. The next element is the business process on which users will focus their analysis. This process will serve as a guideline for the definition of different goals.

Then, the analysis type enables users to define which kind of analysis they want to perform. In order to determine the type of analysis, the user may select which of the following questions [38] needs to be answered: (prescriptive) How to act? (diagnostic) Why has it happened? (predictive) What is going to happen? or (descriptive) What should be done to make it happen?

The visualization element represents a visualization type that will be created to satisfy the visualization goals. The aspect of the data that the visualization should describe is represented with the visualization goal. These goals can be defined as comparison, trend, relationship, composition, cluster, geospatial, distribution, order or cluster, as considered in [33]. Furthermore, the visualizations have one or more interaction type; this element represents the interaction that the user aims to have with the visualization. As considered in [33], the different kinds of interaction are the following: details on demand, zoom, overview or filter. Finally, a visualization will make use of a datasource resource, which will feed the data to the visualization.

Furthermore, in order to cope with real-time scenarios, we have added new elements that capture the execution time and the refresh time. The execution time element defines whether the visualization will be executed in real-time, at a specific moment, or if it shows an image of the overall process, while the refresh time element defines the interval of time in which the visualization will be updated.

As argued in [33], it can be difficult for non-expert users to give proper values to these elements. For example, choosing the correct visualization goal can be difficult. Therefore, our proposal includes some guidelines as shown in the flowchart in Figure 4. This element helps users to choose which visualization goal best suits their needs. In [10], we propose other alternatives to make the definition of model elements easier for non-expert users.

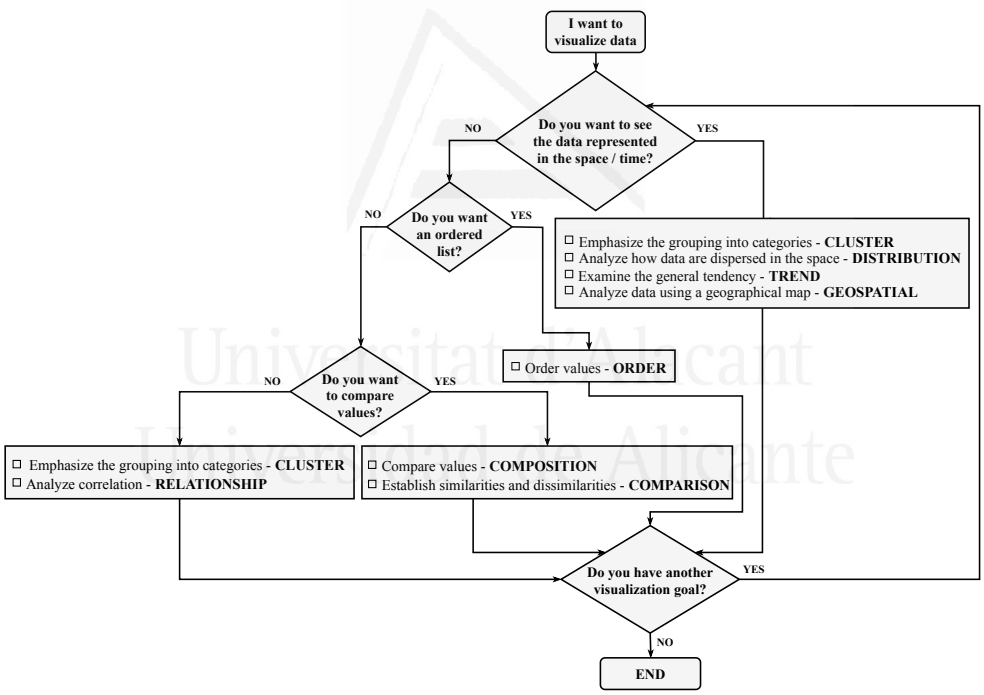


Figure 4. Guidelines expressed as a flowchart to help non-expert users to define visualization goals.

3.1.2. Data Profiling Model

The next model involved in the process is the Data Profiling Model; this model captures characteristics of the data that are relevant for visualization. Firstly, through the User Requirements Model, users select the data elements that they want to represent in the visualizations. Then, through the Data Profiling Model, the data characteristics of dimensionality, cardinality and dependent/independent type are extracted in a semi-automatic manner, as explained below.

- Cardinality can be defined as low or high, depending on the number of items it is necessary to represent. Low cardinality is defined as when there are few dozens of items to represent, while high cardinality is when there are several dozens of items or more.
- Dimensionality represents the number of variables to be visualized. It can be defined as one-dimensional when the data to represent are a single numerical value or string, two-dimensional when one dependent variable depends on one independent variable, n-dimensional if each data object is a point in an n-dimensional space, Tree if a collection of items have a link to one parent item, or graph when a collection of items is provided and each item is linked to an arbitrary number of other items.
- The type of data defines the data type of each variable. It can be defined as nominal if each variable is assigned to one category, ordinal when each variable is assigned to one category and the categories can be sorted, interval when it is possible to determine the equality of intervals or ratio when there is a unique and non-arbitrary zero point.

3.1.3. Derivation of Visualizations

Once the User Requirements Model and the Data Profiling Model are completed and all the requirements have been gathered, a visualization specification can be built. This process is covered in [11], where the transformation from a visualization specification into a visualization implementation is performed following a Model-Driven Architecture (MDA) standard.

3.2. Phase 2—Monitoring of Production Process

Once Phase 1 is completed, users will have defined their goals. Furthermore, the best types of visualization to achieve and measure their goals will have been proposed. Then, the production process starts. Figure 5 summarizes the approach to the production process in our proposal. In the figure, we can see how visualizations generated through the visualization definition process (Figure 2) are integrated and how users intervene during the severity and sustainability check in order to decide whether the production should be stopped or not. In the following, we describe the different components depicted in Figure 5 in more detail.

3.2.1. Cloud Computing Architecture

In order to integrate the real-time data from the sensors with the final dashboard, we have designed the Cloud computing architecture shown in Figure 6. Firstly, the data from the sensors in the production process are collected through a Pub/Sub queue. After that, a streaming analysis pipeline will read the data from the queue and send the data to the AI Engine. Then, the data from the sensors, along with the output data from the Artificial Intelligence model, are stored in a data warehouse. From this data warehouse, the visualizations are fed by the data to be represented in the dashboards that will be presented to the final user.

3.2.2. Artificial Intelligence Model

The first element of the process (Figure 5) is an Artificial Intelligence model. This element is used to detect if there is any potential failure in the production process. A detailed explanation regarding how predictive neural networks work is beyond the scope of this paper. Our proposal is focused on providing techniques to visually understand the output of the models; however, we will briefly explain how these models work together in order to make our proposal more comprehensible.

As Figure 7 shows, the first Artificial Intelligence model is fed with data from the different sensors of the process and divided into two steps. Firstly, as Step 1 in Figure 7 shows, a clustering algorithm is used [39]. This kind of algorithm analyzes the incoming data from the sensors in order to differentiate the phases that compose the production process by analyzing the different values that the sensors have in the whole process. Therefore, the output of this algorithm will be a model for the definition of the phases that compose the process.

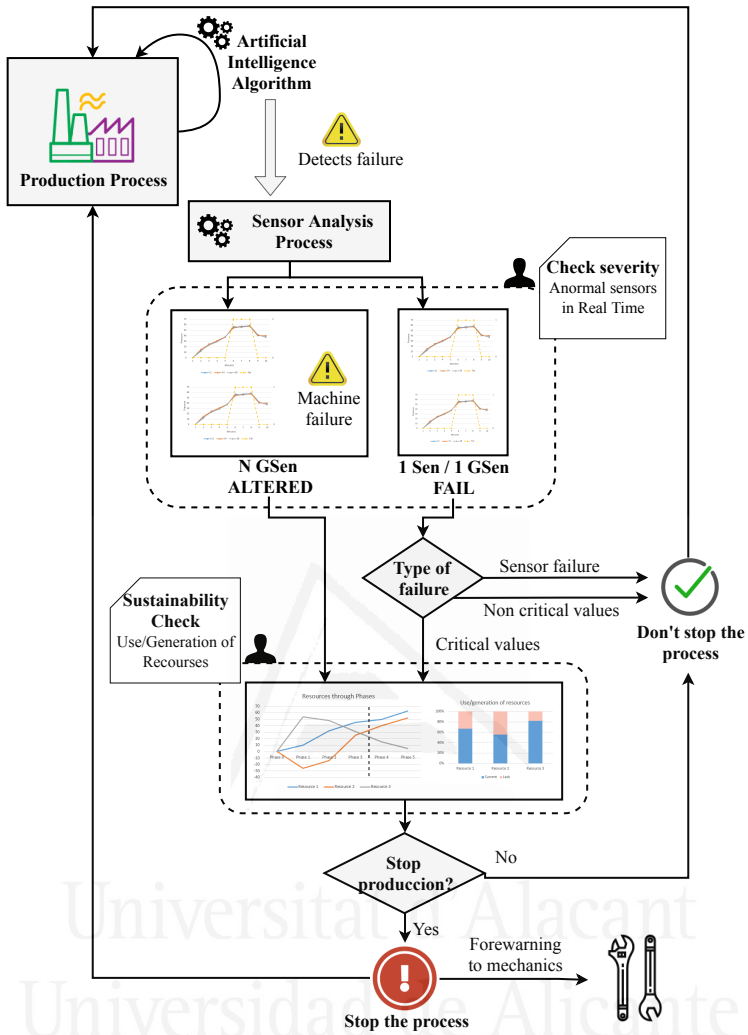


Figure 5. Phase 2—production process.

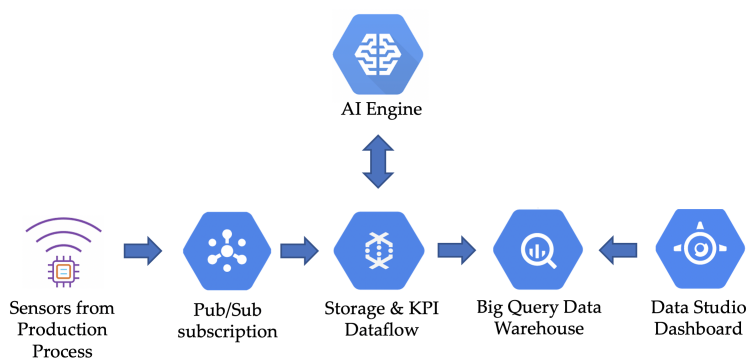


Figure 6. Cloud computing architecture of the system.

Once the phases have been identified, a Deep Neural Network [40] based on Variational Autoencoders (VAEs) for anomaly detection [41] is trained in each cluster (phase). Once the neural

network is trained, as Step 2 in Figure 7 shows, the data from the values of the sensors are analyzed in real-time. First, the incoming data are analyzed by the clustering model in order to discover the phase in which the data have been generated. Once the phase is identified, the neural network corresponding to that phase is called for prediction. This neural network identifies whether there are potential failures present in the production process. Therefore, the output will be a data tuple encoded by the corresponding VAE. The Euclidean distance between input and output tuples will be used to assess whether or not the input of the model corresponded to an anomalous situation of the machinery.

With the information provided from the neural network and the clustering model, users are able to determine if a potential failure has been detected, as well as the phase of the process in which it was detected. However, due to the black-box nature of neural networks, this information is insufficient to understand the root cause of the problem. Therefore, our approach introduces the next element: the sensor analysis process.

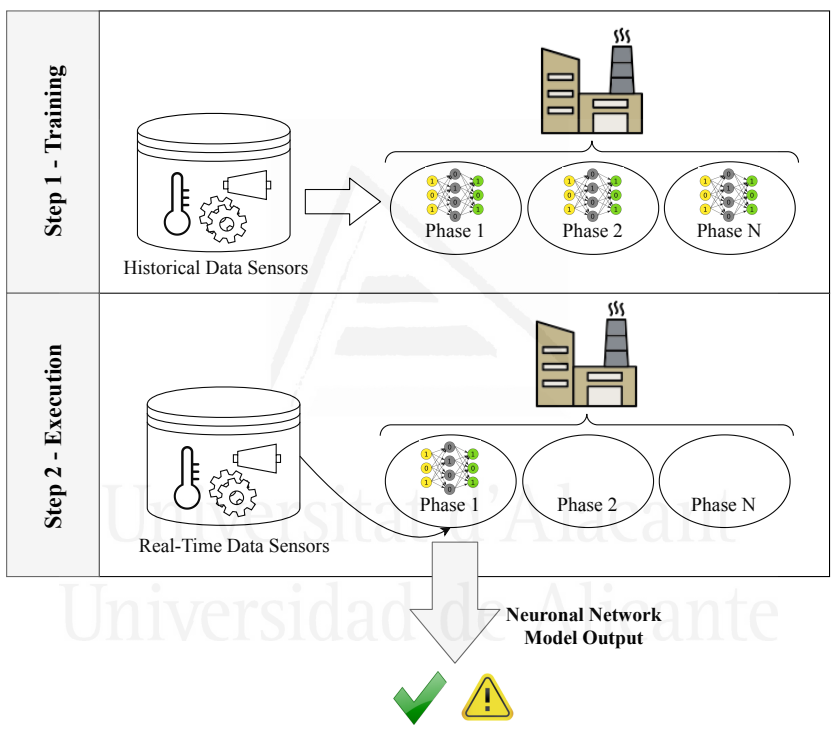


Figure 7. Artificial Intelligence model.

3.2.3. Sensor Analysis Process

Once the Artificial Intelligence model has detected that there is a potential failure in the process, the sensor analysis process (see Figure 5) enables users to detect what type of fault has occurred in real-time and make decisions according to the severity of the problem.

The sensor analysis process compares the values of the sensors in order to detect which are out of range. There are two situations in which our system detects a failure: on the one hand, our proposal defines that a sensor is out of range when the current value exceeds the limits defined in its hardware specification; on the other hand, from time to time, a system failure is not produced by the failure of an individual sensor—in these cases, the fault is identified by the anomalous values of a set of sensors. These sensors may have individual values within adequate operation ranges; however, their combined status can be abnormal with regards to the production process. As an example of this situation, an energy-generation engine's throughput sensor could send a value of 1% while a related

temperature sensor could be measuring 300 ºC. Despite both measurements being correct according to their hardware specifications, it is illogical that an engine could work at that capacity while having such a high temperature. Thus, taking into account both of the explained scenarios, our approach covers them as follows.

- N GSen ALTERED (Machine failure): N groups of sensors are altered. An alteration means that there is a small alteration in the values of the sensors but that no sensor is out of its acceptable ranges. Therefore, groups of visualizations are generated. These visualizations represent all sensors of the machine, grouped by the unit of measure and the localization in the machine. Furthermore, warnings will be considered, thus warning users that the machine is presenting an abnormal status and that it is possible that the production optimal.

  When this scenario arises, additional information will be necessary in order to make decisions. This new information will help users to decide if, at that moment, it is sustainable to stop the production or not.
- 1 Sen/1 GSen FAIL: There is one sensor or a group of sensors which is out of range. In these cases, a group of visualizations are generated in which the anomalous sensor/sensors with their real-time values are represented, split by the unit of measurement. Furthermore, in order to display a reference, the historical average value of these sensors is also represented. Moreover, these visualizations include the values of sensors located physically close to the relevant sensor which do not present anomalies.

  When this case arises, users should make their first decision. As Figure 5 shows, users should decide, relying on the visualizations, if the failure is a device failure or is not critical. Otherwise, they must decide if it is a critical moment and therefore necessary to consider the possibility of stopping the production process.

  - Sensor failure or non-critical values: If users decide that the failure is caused due to a broken sensor or if the values that the sensor is showing are acceptable or are located in non-critical areas, the production process will continue. However, if users deem it necessary, it is possible to use the visualizations to continuously monitor the values of these abnormal sensors, thus allowing users to visualize the values of these sensors in real-time and take measures if at any time the sensors reach critical values.
  - Critical values: If users decide that the values of the sensors are critical fir the production process, it will be necessary to present additional information in order to help to users to decide if, at that moment, it would be sustainable to stop the production or not.

After this process, users are able to check the severity of failures during the production process and locate the problem by analyzing the sensors through visualizations. Furthermore, if users detect a severe problem, more information will be shown so that they will be able to decide whether it is sustainable to stop the production at that moment or not. In the following section, we describe this sustainability check in more detail.

3.2.4. Sustainability Check

The sustainability check (see Figure 5) is performed when users have detected that there is a potential critical failure in the production process. Therefore, at that moment, users need more information in order to decide whether it would be sustainable to stop production or not. They can decide if it is more optimal that production continues with some risk of failure, even if sensors or some machinery pieces may be damaged. In order to make these decisions, a set of visualizations is needed that measure the used/generated resources at each phase of the production process, enabling users to analyze the situation and make decisions according to the expected consequences.

Thanks to the application of the AI models, we are aware of the exact phase which the production process has reached. With this knowledge, a set of visualizations are generated by following the

design defined in phase 1 of the process (Figure 2), in which the visualizations required to achieve user goals were derived. This visualizations present the expected evolution of the system in terms of costs, risks and resources. For example, users may decide that during the initial phase, many resources have been spent and the production has been low. Therefore, it would not be sustainable to stop the production at this moment since the cost would be too high. However, during a more advanced phase, the resources spent have already been amortized, and stopping the production at this moment will lead to an acceptable reduction of the profit without significant resource losses.

Once users have analyzed the visualizations, if they decide to avoid stopping the production, the affected parts of the machinery will be monitored. The system will create a very detailed visualization of the sensors of each part of the machinery in order to enable users to stop the process at the moment at which the sensors reach critical values. This could potentially avoid risky situations for the machinery as well as for the corresponding operators.

In the case that users decide to stop the production process, a forewarning will be sent to the mechanics with all the information of the affected parts and the values of the sensors. Therefore, the mechanics will be able to study the cause of the failure and will be able to intervene as soon as production is completely stopped in order to make the necessary repairs.

4. Case Study: Gas Turbines for Electricity Generation

This scenario has been developed in the context of an international project under a non-disclosure agreement (NDA). Since the data are industrial property, we provide real data in an anonymized manner and thus do not provide details of the turbine or specific sensors. Moreover, the data shown in this work have been altered to avoid presenting real data protected by the NDA.

In the following, we show how our approach is applied to a real case study of a company that produces electricity using gas turbines. The main goal of the company is to improve the sustainability of the process. In order to achieve this goal, the company requires a set of visualizations to analyze their data in real-time in order to foster the decision-making process regarding when it is optimal and sustainable to stop the production process at a given point in time. The gas turbines for electricity generation used in this case of study consist of 80 sensors, from which data are gathered at runtime. These sensors are located along the machine and measure all relevant magnitudes, including the temperature, pressure, frequency, speed, humidity, etc. of different parts of the gas turbine. Some of them are replicated to ensure correct measurements.

Following the Cloud computing architecture shown in Figure 6, in this specific case study, we used Google Cloud Dataflow to collect and process the data from sensors in real-time. These data, as well as the information from the output of the Artificial Intelligence models, have been stored into a BigQuery data warehouse. Finally, we have chosen Google Data Studio to perform the visualizations.

4.1. Phase 1—Definition of Goals and Visualizations

Following the application of our approach (Figure 2), the first step in phase 1 is to create a User Requirements Model. In Figure 8, we can see the result of its application. In this case, the user is a production supervisor; however, this user is not an expert in data visualization. Therefore, the user is defined as a "lay user", and the analysis will therefore be focused on the "Electricity Generation" business process.

Next, the strategic goal is defined as "improve sustainability", and the type of analysis to perform is "prescriptive analysis", meaning that the user wants to know how to act in the process; specifically, whether the process should be stopped or not.

The prescriptive analysis is decomposed into decision goals. These goals are defined by the user as "prevent breakage", "identify when production should be stopped", and "optimize resources". By themselves, the decisions goals do not provide the necessary details about the data to be visualized. Therefore, for each decision goal, the user has to specify information goals.

From each of the decision goals, the user decided upon the following information goals: "analyze damaged pieces", "analyzed used/generated resources at a certain moment" and "analyze the production though phases". For each information goal, one visualization will be created to achieve it.

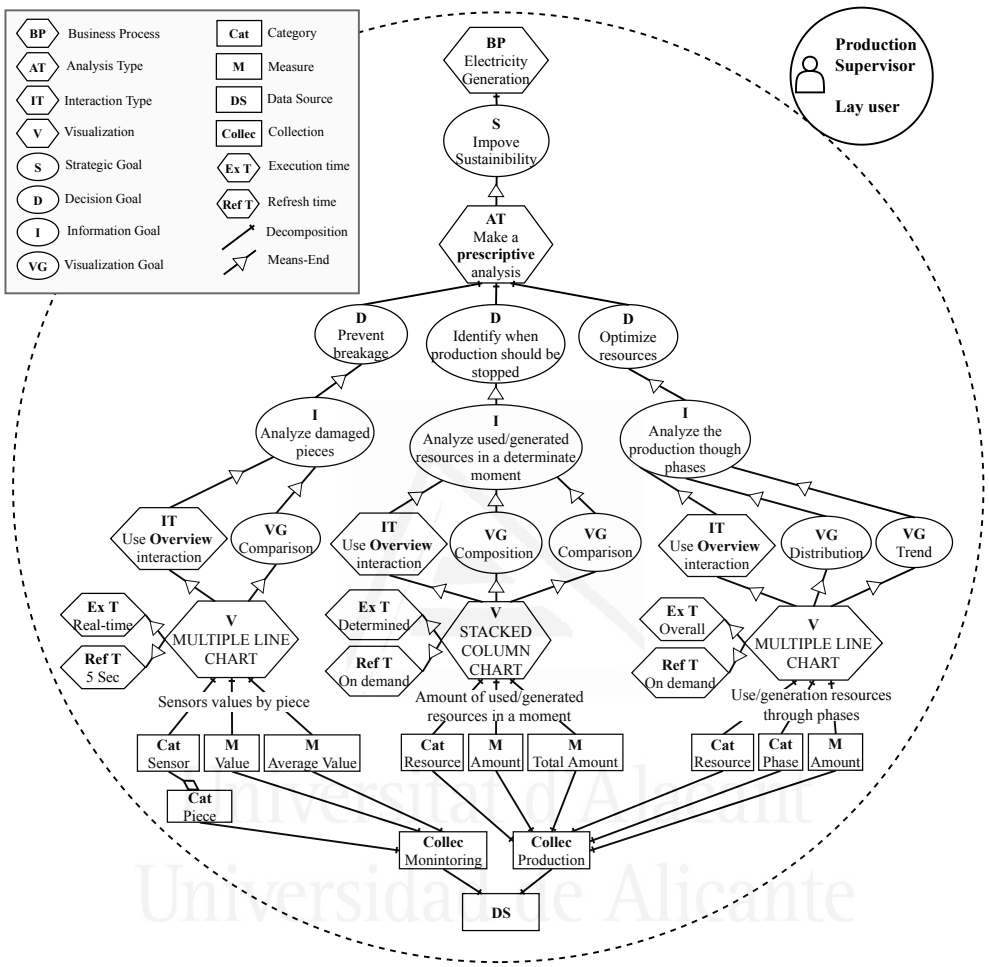


Figure 8. Application of our User Requirements Model to the case study.

The user defines the visualization goals following the guidelines shown in Figure 4. In this case, the user defines "comparison", "composition", "distribution" and "trend" as visualization goals. The user also defines the kind of interaction that they would like to have in the visualization as "overview". Furthermore, since we are using a real-time scenario, the user must define the execution and refresh time of the visualizations. In this case, the user defines as execution times "real-time", "determined" and "overall". As refresh time, the user defines "5 sec" or "on demand". Finally, the user specifies the data source that will feed the information for the analysis and selects the categories and measures that will populate the visualizations.

Once the data sources and collections are defined by the users, it is possible to apply our Data Profiling Model. This model will determine, in a semi-automatic manner, the dimensionality, cardinality and type of the data.

We focus on the "sensor values by piece" visualization from the goal-based model (Figure 8). This visualization will require information about the category "sensor" and the measures "value" and "average Value". Fist, the data profiling tool classifies the independent variable "sensor" as

nominal and the dependent variables "value" and "average value" as interval. The dimensionality is set to n-dimensional, due the fact that the user has selected only three variables to visualize. Finally, the cardinality is defined as high because the data contain a large number of items to represent. Overall, the visualization specification obtained through the User Requirements Model and the Data Profiling Model is as follows:

- **Visualization goal**: Comparison
- **Interaction**: Overview
- **User**: Lay
- **Dimensionality**: n-dimensional
- **Cardinality**: High
- **Independent Type**: Nominal
- **Dependent Type**: Interval

Following [33], we are able to automatically translate this visualization specification into the most suitable visualization type. As specified in Section 3.1.3, this process is covered in [11]. In this case, the visualization type that best fits this specification is "multiple line chart". This whole process is repeated with the rest of the visualizations that compose the model (Figure 8) in order to derive the most suitable visualization type for each specification.

4.2. Phase 2—Monitoring of Production Process

Once users have defined the goals of the process and the system has derived the best visualization types (phase 1), it is possible to start monitoring the production process. First, as Figure 5 shows, the Artificial Intelligence model (previously trained) is launched. When the model detects a failure, the sensor analysis process is executed in order to detect whether the fault has been caused by an alteration of the whole machine, or otherwise if the failure has been caused by a specific sensor or group of sensors.

In the event that the sensor analysis process detects that the fault has been caused by an alteration of the whole machine (N GSen ALTERED), a dashboard like the one shown in Figure 9 is generated. Following the recommendation of the model shown in Figure 8, a multiple-line chart visualization has been generated to achieve the goal of analyzing damaged parts.

Figure 9. Dashboard of N GSen ALTERED.

This dashboard represents the overall status of the machine and warns users about the fact that the machine is failing. Thus, all machine sensors are represented, split by the unit of measurement and the localization in the machine. Each visualization represents the evolution of sensor values during the time of the process execution as well as the historical average value of these sensors, which serves as reference to the users. In each visualization, we can see the names of the sensors. The X-axis represents the the date and time when the data were read from the sensors and the Y-axis represents the values of the readings. Additionally, the right side of Y-axis represents whether the process is failing or not.

As we can see in Figure 9, there is no sensor that is out of range, although the machine is failing; therefore, it is possible that the production is not optimal. In order to make decisions and decide whether it is sustainable to stop the production or not, additional visualizations will be necessary.

Figure 10 represents the additional visualizations needed to check the sustainability of the process. On the left side, the visualization represents the use/generation of resources through the phases and marks the stage that the process has reached. This visualization achieves the information goal of the model (Figure 8): "analyze the production through phases". In this case, the process has almost reached phase 4, and the spent resources are already amortized. Therefore, stopping the production at this moment will only lead to a reduction of the profit; however, if the process were in phase 1, the process would have just begun, and therefore many resources would have been spent and the production would be very low.

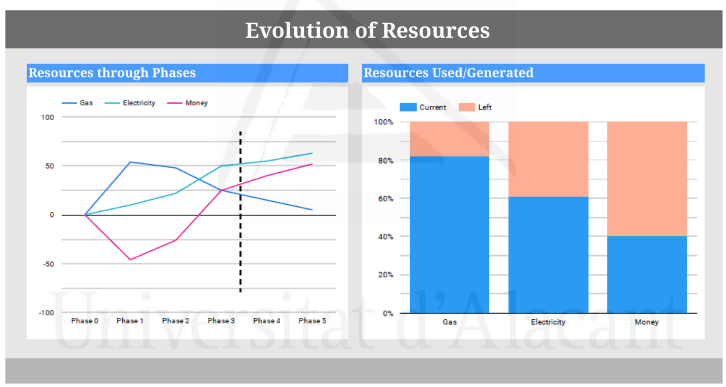


Figure 10. Dashboard showing the evolution of resources.

On the other hand, the right-side visualization represents the resources that have been used/generated at a specific moment in order to achieve the information goal "analyze the used/generated resources at a certain moment". This visualization enables users to be more precise in their decisions.

In the case that the sensor analysis process detects that there is a sensor or group of sensors that is out of range (1 Sen/1 GSen FAIL), one of the dashboards shown in Figure 11 is generated. These dashboards represent the sensors detected as out of range and, for reference, the historical average value of these sensors. Furthermore, the values of sensors located physically close and which do not present anomalies can be included.

Figure 11a,b shows two possible cases that users may face. On the one hand, if the visualization looks like Figure 11a, this means that there is a defective sensor and that it has made an incorrect reading (in this case, sensor S22 is defective). Therefore, the machinery is not affected, and it will not be necessary to stop the production process. The production can continue, and information about the damaged sensor is sent to the AI model to ignore the values of this sensor.

On the other hand, if the system generates a visualization like Figure 11b, this means that there is a problem in this area. It is possible to see in this figure that sensors S17 and S18 show values that are out of the average range. In this case, users should decide if it is a critical moment or if the sensor values

are in a critical range; otherwise, if these sensors are not critical to the production, production can continue regardless of this damaged piece.

(**a**) Dashboard sensor failure (**b**) Dashboard sensors out of range

Figure 11. 1 Sen/1 GSen FAIL Dashboards.

In the case that it is not a critical moment or if the sensor values are not in a critical range, the production process can continue. However, in this case, we cannot ignore the values of these sensors; a visualization is created in order to allow users to monitor the damaged area. This visualization will enable users to take measures if, at any time, the sensors reach a critical point that may affect the normal operations of the machine or at which the machine would become potentially unsafe for operators, allowing users to stop the machine and proceed to perform maintenance work.

Otherwise, in the case that the values of the sensors are classified as critical to the production process, a dashboard like the one shown in Figure 10 will be presented to users in order for them to decide if it is sustainable or not to stop the production process.

5. Limitations

In this section, we summarize the limitations of our work.

- Our proposal has been applied to a specific case study of gas turbines for electricity generation. In principle, the proposal is context-independent, but it should be tested in other production contexts to verify that the results are accurate.
- Our methodology has been developed for non-expert users; however, the user's domain expertise can be a crucial factor in the definition of more complex dashboards.
- In order to allow users to follow the methodology by themselves, the creation of a CASE tool is necessary.
- Further evaluation of our proposal is required; to this end, we are conducting an empirical evaluation, analyzing the obtained results through the application of our methodology in other production contexts.

6. Conclusions and Future Work

Global energy consumption is growing daily, and new lifestyle trends are increasing the need for electricity generation. Industry is benefiting from the rise of technologies such as IoT that enable us to better understand and monitor how production processes are performing. Effective use of these technologies will enable users to take actions aimed at improving both the throughput as well as the sustainability of the process. However, this requires data to exploited from real-time IoT sensors,

which is a challenging task due to the size, speed and variety of the data. This is especially cumbersome in industrial IoT devices featuring hundreds of sensors producing measurements which are prone to fail due to several conditions (degradation of sensors, inconsistency among replicated sensors, incomplete data, etc.).

In order to tackle this issue, we have proposed a new methodological approach to monitor industrial machinery through an IoT-based visualization technique. Our approach collects users' goals and the requirements of the production process, analyzes the incoming data from IoT sensors and automatically derives the most suitable visualization type for each context. It presents a set of visualizations that are intended for non-expert users in data visualization and created by taking into account the level of knowledge of the users. In this sense, our approach makes it easier to visually locate and understand the failures that could arise in a production process and enables users to make the most sustainable decision in each situation.

When this kind of industrial system features AI prediction engines, its complexity is even greater. This is because a neural-network-based AI will commonly not work as a block box and usually provides binary classification results such as "the system is working correctly" or "there will be a problem". Because of this, it is cumbersome to relate the output of this model with the status of the system's sensors measurements. However, our approach takes this issue into account by offering visualizations that help users to co-relate AI outputs and sensor's data, thus enabling them to identify where and when the problem was caused. Otherwise, it is difficult to identify the problematic part within a systems consisting of hundreds of sensors.

Moreover, in order to assess the suitability of our proposal, we have presented a case study based on gas turbines for electricity generation. Our proposal will contribute to the avoidance of unexpected maintenance stops, thus improving the sustainability of the energy-production industry.

As part of our future work, we are working on a further evaluation of our proposal; we are conducting an empirical evaluation, analyzing the results obtained through the application of our methodology. Furthermore, we are working on the creation of a CASE tool in order to facilitate the use of our process, which will be evaluated as in our previous experiments [42].

Author Contributions: Conceptualization, A.L., A.M. and J.T.; methodology, A.L., A.M. and J.T.; investigation, A.L., A.M. and J.T.; writing—original draft preparation, A.L. and M.A.T.; writing—review and editing, A.L., M.A.T., A.M. and J.T.; visualization, A.L., A.M. and J.T.; supervision, A.M. and J.T.; funding acquisition, J.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been co-funded by the ECLIPSE-UA (RTI2018-094283-B-C32) project funded by Spanish Ministry of Science, Innovation, and Universities and the DQIoT (INNO-20171060) project funded by the Spanish Center for Industrial Technological Development, approved with an EUREKA quality seal (E!11737DQIOT). Ana Lavalle holds an Industrial PhD Grant (I-PI 03-18) co-funded by the University of Alicante and the Lucentia Lab Spin-off Company.

Acknowledgments: We would like to thank the Lucentia Lab Spin-off Company for providing us with the models and the data necessary to develop our proposal. Figures 3 and 4 reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer, Lecture Notes in Computer Science, Requirements-Driven Visualizations for Big Data Analytics: A Model-Driven Approach, Ana Lavalle, Alejandro Maté & Juan Trujillo, COPYRIGHT 2019.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pérez-Lombard, L.; Ortiz, J.; Pout, C. A review on buildings energy consumption information. *Energy Build.* **2008**, *40*, 394–398. [CrossRef]
2. Sun, J. Changes in energy consumption and energy intensity: a complete decomposition model. *Energy Econ.* **1998**, *20*, 85–100. [CrossRef]
3. Pant, D.; Van Bogaert, G.; Diels, L.; Vanbroekhoven, K. A review of the substrates used in microbial fuel cells (MFCs) for sustainable energy production. *Bioresour. Technol.* **2010**, *101*, 1533–1543. [CrossRef] [PubMed]
4. Mourtzis, D.; Vlachou, E.; Milas, N. Industrial Big Data as a result of IoT adoption in manufacturing. *Procedia Cirp* **2016**, *55*, 290–295. [CrossRef]

5. Bedi, G.; Venayagamoorthy, G.K.; Singh, R.; Brooks, R.R.; Wang, K. Review of Internet of Things (IoT) in Electric Power and Energy Systems. *IEEE Internet Things J.* **2018**, *5*, 847–870. [CrossRef]
6. Ribeiro, R.P.; Pereira, P.M.; Gama, J. Sequential anomalies: A study in the Railway Industry. *Mach. Learn.* **2016**, *105*, 127–153. [CrossRef]
7. Ge, Z.; Song, Z.; Ding, S.X.; Huang, B. Data Mining and Analytics in the Process Industry: The Role of Machine Learning. *IEEE Access* **2017**, *5*, 20590–20616. [CrossRef]
8. Widrow, B.; Rumelhart, D.E.; Lehr, M.A. Neural Networks: Applications in Industry, Business and Science. *Commun. ACM* **1994**, *37*, 93–105. [CrossRef]
9. Dayhoff, J.E.; DeLeo, J.M. Artificial neural networks: opening the black box. *Cancer Interdiscip. Int. J. Am. Cancer Soc.* **2001**, *91*, 1615–1635. [CrossRef]
10. Lavalle, A.; Maté, A.; Trujillo, J.; Rizzi, S. Visualization Requirements for Business Intelligence Analytics: A Goal-Based, Iterative Framework. In Proceedings of the 27th IEEE International Requirements Engineering Conference (RE 2019), Jeju Island, Korea, 23–27 September 2019; pp. 109–119.
11. Lavalle, A.; Maté, A.; Trujillo, J. Requirements-Driven Visualizations for Big Data Analytics: A Model-Driven Approach. In Proceedings of the Conceptual Modeling—38th International Conference (ER 2019), Salvador, Brazil, 4–7 November 2019; pp. 78–92.
12. Lavalle, A.; Teruel, M.A.; Maté, A.; Trujillo, J. Improving Sustainability of Smart Cities through Visualization Techniques for Big Data from IoT Devices. *Sustainability* **2020**, *12*, 5595. [CrossRef]
13. Weiss, E.B. United Nations conference on environment and development. *Int. Leg. Mater.* **1992**, *31*, 814–817. [CrossRef]
14. Krajnc, D.; Glavič, P. Indicators of sustainable production. *Clean Technol. Environ. Policy* **2003**, *5*, 279–288. [CrossRef]
15. Gallopin, G.C. Indicators and their use: information for decision-making. *Scope Sci. Comm. Probl. Environ. Int. Counc. Sci. Unions* **1997**, *58*, 13–27.
16. Veleva, V.; Ellenbecker, M. Indicators of sustainable production: framework and methodology. *J. Clean. Prod.* **2001**, *9*, 519–549. [CrossRef]
17. Chang, K.M.; Dzeng, R.J.; Wu, Y.J. An automated IoT visualization BIM platform for decision support in facilities management. *Appl. Sci.* **2018**, *8*, 1086. [CrossRef]
18. Traub, J.; Steenbergen, N.; Grulich, P.M.; Rabl, T.; Markl, V. I^2: Interactive Real-Time Visualization for Streaming Data. In Proceedings of the 20th International Conference on Extending Database Technology (EDBT 2017), Venice, Italy, 21–24 March 2017; pp. 526–529.
19. Cardin, O. Classification of cyber-physical production systems applications: Proposition of an analysis framework. *Comput. Ind.* **2019**, *104*, 11–21. [CrossRef]
20. Peres, R.S.; Rocha, A.D.; Leitão, P.; Barata, J. IDARTS—Towards intelligent data analysis and real-time supervision for industry 4.0. *Comput. Ind.* **2018**, *101*, 138–146. [CrossRef]
21. Sobral, T.; Galvão, T.; Borges, J. Visualization of urban mobility data from intelligent transportation systems. *Sensors* **2019**, *19*, 332. [CrossRef]
22. Napolitano, R.; Blyth, A.; Glisic, B. Virtual environments for visualizing structural health monitoring sensor networks, data, and metadata. *Sensors* **2018**, *18*, 243. [CrossRef]
23. Bhatia, M.; Sood, S.K. A comprehensive health assessment framework to facilitate IoT-assisted smart workouts: A predictive healthcare perspective. *Comput. Ind.* **2017**, *92–93*, 50–66. [CrossRef]
24. Oliver, M.; Teruel, M.A.; Molina, J.P.; Romero-Ayuso, D.; González, P. Ambient intelligence environment for home cognitive telerehabilitation. *Sensors* **2018**, *18*, 3671. [CrossRef] [PubMed]
25. Agrawal, R.; Kadadi, A.; Dai, X.; Andrès, F. Challenges and opportunities with big data visualization. In Proceedings of the 7th International Conference on Management of Computational and Collective intElligence in Digital EcoSystems, Caraguatatuba, Brazil, 25–29 October 2015; pp. 169–173.
26. Liu, Z.; Jiang, B.; Heer, J. *imMens*: Real-time Visual Querying of Big Data. *Comput. Graph. Forum* **2013**, *32*, 421–430. [CrossRef]
27. de Lara Pahins, C.A.; Stephens, S.A.; Scheidegger, C.; Comba, J.L.D. Hashedcubes: Simple, Low Memory, Real-Time Visual Exploration of Big Data. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 671–680. [CrossRef] [PubMed]
28. Donat, W.; Choi, K.; An, W.; Singh, S.; Pattipati, K. Data visualization, data reduction and classifier fusion for intelligent fault diagnosis in gas turbine engines. *J. Eng. Gas Turbines Power* **2008**, *130*. [CrossRef]

29. Shafii, S.; Obermaier, H.; Linn, R.; Koo, E.; Hlawitschka, M.; Garth, C.; Hamann, B.; Joy, K.I. Visualization and Analysis of Vortex-Turbine Intersections in Wind Farms. *IEEE Trans. Vis. Comput. Graph.* **2013**, *19*, 1579–1591. [CrossRef]
30. Hicks, Y.R.; Locke, R.J.; Anderson, R.C. Optical measurement and visualization in high-pressure high-temperature aviation gas turbine combustors. In Proceedings of the SPIE Symposium on Applied Photonics, Glasgow, UK, 21 May 2000; pp. 66–77.
31. Syafrudin, M.; Fitriyani, N.L.; Li, D.; Alfian, G.; Rhee, J.; Kang, Y.S. An open source-based real-time data processing architecture framework for manufacturing sustainability. *Sustainability* **2017**, *9*, 2139. [CrossRef]
32. Correll, M.; Li, M.; Kindlmann, G.L.; Scheidegger, C. Looks Good To Me: Visualizations As Sanity Checks. *IEEE Trans. Vis. Comput. Graph.* **2019**, *25*, 830–839. [CrossRef]
33. Golfarelli, M.; Rizzi, S. A model-driven approach to automate data visualization in big data analytics. *Inf. Vis.* **2020**, *19*, 24–47. [CrossRef]
34. Teruel, M.A.; Maté, A.; Navarro, E.; González, P.; Trujillo, J.C. The New Era of Business Intelligence Applications: Building from a Collaborative Point of View. *Bus. Inf. Syst. Eng.* **2019**, *61*, 615–634. [CrossRef]
35. iStar 2.0 Language Guide. Availiable Online: https://arxiv.org/abs/1605.07767 (accessed on 01 July 2020).
36. Maté, A.; Trujillo, J.; Franch, X. Adding semantic modules to improve goal-oriented analysis of data warehouses using I-star. *J. Syst. Softw.* **2014**, *88*, 102–111. [CrossRef]
37. López-Jaquero, V.; Rodríguez, A.C.; Teruel, M.A.; Montero, F.; Navarro, E.; Gonzalez, P. A bio-inspired model-based approach for context-aware post-WIMP tele-rehabilitation. *Sensors* **2016**, *16*, 1689. [CrossRef] [PubMed]
38. Shi-Nash, A.; Hardoon, D.R. Data analytics and predictive analytics in the era of big data. In *Internet of Things and Data Analytics Handbook*; Wiley: Hoboken, NJ, USA, 2017; pp. 329–345.
39. He, X.; Cai, D.; Shao, Y.; Bao, H.; Han, J. Laplacian Regularized Gaussian Mixture Model for Data Clustering. *IEEE Trans. Knowl. Data Eng.* **2011**, *23*, 1406–1418. [CrossRef]
40. Ng, A. Sparse autoencoder. *CS294A Lect. Notes* **2011**, *72*, 1–19.
41. An, J.; Cho, S. Variational autoencoder based anomaly detection using reconstruction probability. *Spec. Lect. IE* **2015**, *2*, 1–18.
42. Teruel, M.A.; Navarro, E.; López-Jaquero, V.; Simarro, F.M.; González, P. A CSCW Requirements Engineering CASE Tool: Development and usability evaluation. *Inf. Softw. Technol.* **2014**, *56*, 922–949. [CrossRef]

Capítulo 8

A Methodology to Automatically Translate User Requirements Into Visualizations: Experimental Validation

Lavalle, A., Maté, A., Trujillo, J., Teruel, M. A., & Rizzi, S. (2021). A Methodology to Automatically Translate User Requirements Into Visualizations: Experimental Validation. *Information and Software Technology, 136*, 106592.

Factor de Impacto: **2,726**
Clasificación JCR: **Q2 (28/108 Computer science, software engineering)**

Disponible en:
DOI: https://doi.org/10.1016/j.infsof.2021.106592

Information and Software Technology 136 (2021) 106592

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

A methodology to automatically translate user requirements into visualizations: Experimental validation

Ana Lavalle [a,b,*], Alejandro Maté [a,b], Juan Trujillo [a,b], Miguel A. Teruel [a,b], Stefano Rizzi [c]

[a] Lucentia Research, Department of Software and Computing Systems, University of Alicante, Carretera San Vicente del Raspeig s/n, 03690, San Vicente del Raspeig, Alicante, Spain
[b] Lucentia Lab, C/Pintor Pérez Gil, N-16, 03540, Alicante, Spain
[c] DISI, University of Bologna, V.le Risorgimento 2, 40136, Bologna, Italy

ARTICLE INFO

Keywords:
Data visualization
Big data analytics
Model-driven development
Requirements engineering
Experimental validation

ABSTRACT

Context: Information visualization is paramount for the analysis of Big Data. The volume of data requiring interpretation is continuously growing. However, users are usually not experts in information visualization. Thus, defining the visualization that best suits a determined context is a very challenging task for them. Moreover, it is often the case that users do not have a clear idea of what objectives they are building the visualizations for. Consequently, it is possible that graphics are misinterpreted, making wrong decisions that lead to missed opportunities. One of the underlying problems in this process is the lack of methodologies and tools that non-expert users in visualizations can use to define their objectives and visualizations.
Objective: The main objectives of this paper are to (i) enable non-expert users in data visualization to communicate their analytical needs with little effort, (ii) generate the visualizations that best fit their requirements, and (iii) evaluate the impact of our proposal with reference to a case study, describing an experiment with 97 non-expert users in data visualization.
Methods: We propose a methodology that collects user requirements and semi-automatically creates suitable visualizations. Our proposal covers the whole process, from the definition of requirements to the implementation of visualizations. The methodology has been tested with several groups to measure its effectiveness and perceived usefulness.
Results: The experiments increase our confidence about the utility of our methodology. It significantly improves over the case when users face the same problem manually. Specifically: (i) users are allowed to cover more analytical questions, (ii) the visualizations produced are more effective, and (iii) the overall satisfaction of the users is larger.
Conclusion: By following our proposal, non-expert users will be able to more effectively express their analytical needs and obtain the set of visualizations that best suits their goals.

1. Introduction

Information visualization is paramount for the analysis of Big Data. The volume of data requiring interpretation is continuously growing. Visual analytics in software engineering is also gaining importance [1]. In fact, according to [2], the global data visualization market size stood at USD 8.85 billions in 2019 and is projected to reach USD 19.20 billions by 2027. The evolution of analytics and visualization techniques lies at the core of business strategies, and more and more research lines are focusing on the visualization of data.

However, users are typically unskilled in information visualization. Thus, finding the visualization that best suit a determined context is a very challenging task for them. Moreover, it is often the case that users do not have a clear idea of what objectives they are building the visualizations for. Consequently, it is possible that graphics are misinterpreted, making wrong decisions that lead to missed opportunities. One of the underlying problems in this process is the lack of methodologies and tools that users who are not experts in visualizations can use to define their objectives and the corresponding visualizations.

Choosing and implementing the most suitable visualizations for each dataset is a really complicated task, particularly when working with Big Data. In these scenarios, it is common to find heterogeneous data sources that require extensive knowledge of the underlying data

* Corresponding author at: Lucentia Research, Department of Software and Computing Systems, University of Alicante, Carretera San Vicente del Raspeig s/n, 03690, San Vicente del Raspeig, Alicante, Spain.
E-mail address: alavalle@dlsi.ua.es (A. Lavalle).

https://doi.org/10.1016/j.infsof.2021.106592
Received 30 November 2020; Received in revised form 26 March 2021; Accepted 27 March 2021
Available online 5 April 2021
0950-5849/

to create a suitable visualization [3]. Moreover, using an unsuitable type of visualization can lead to misunderstanding the data and making wrong decisions. In this sense, an approach such as SkyViz [4] can support users in creating visualizations. In SkyViz, the suitable visualization types for a given dataset are selected and created based on a *visualization context* defined by users; however, as the authors recognize, defining a visualization context from scratch can be a challenge for users who are not expert in data visualization.

To fill this gap, in this paper we present a process that helps non-expert users define their analytical goals and derive automatically the suitable visualizations according to the defined context. Our proposal covers the whole process, from the definition of the user requirements to the implementation of the visualizations. In our previous work we proposed (i) a User Requirements Model [5] to capture the users' analytical needs, (ii) a Data Profiling model [6] to extract semi-automatically the characteristics of the data sources, and (iii) a Data Visualization Model [6] that enables users to specify the visualization details regardless of the technology used for the implementation. Therefore, by following our proposal non-expert users will be able to communicate their analytical needs and obtain the visualizations that best suit them to achieve their goals. Besides, a dashboard will also be generated to group the visualizations and help users to carry out strategic decisions as such as the monitoring and measuring of their goals. Moreover, in this paper we put into practice the proposed methodology, by applying it to a case study focused on the Incidents Management from the Police Department of San Francisco.

To assess the validity of our proposal we have performed an experiment with 97 non-expert users in data visualization. In the experiment each user was tasked with two exercises. In the first exercise, participants were tasked with carrying out an analysis over a dataset without following any particular methodology. In the second exercise, each participant carried out a different analysis than the one they had seen before, this time following our proposed methodology. The results obtained from the experiment have been analyzed and represented graphically in order to show the improvements achieved by our methodology.

Therefore, the main contributions of this paper are to show the overall steps of the process, the application of the approach to a new scenario (illustrative example) to show its generalizability, and the validation of the proposal through the analysis of the results obtained by several groups of participants.

The rest of the paper is structured as follows. Section 2 presents the related work in the area of visualizations and analytics. Section 3 describes our process to automatically create visualizations. Section 4 shows our approach applied to an illustrative example. Section 5 presents an evaluation of the proposal by means of an experiment with non-expert users. Section 6 describes the validity threats to our proposal. Finally, Section 7 summarizes the conclusions and sketches future works.

2. Related work

Several approaches highlight the importance of visual analytics. For instance, [7] and [8] show the potential of visual analytics in software engineering. In [7] a visualization framework is presented that utilizes heat-maps to explore the evolution of a source code repository. Meanwhile, [8] presents visualization approach that captures significant aspects of the development process, and then tightly integrates and synchronizes them with product artifacts created by it.

Due to the relevance of this field, numerous authors are working in this area. In [9,10] and [11], techniques are proposed to automatically generate visualizations or dashboards. However, all of them rely on the user to choose the type of visualization to be used. That is why some other approaches propose ways to find the best type of visualization. For instance, authors in [12] review the main classifications proposed in the literature and integrate them into a single framework. In [13] a framework is proposed that chooses the best type of visualization. Similarly, in [14] some visualization types are related to those types of users objectives that could be more compliant with. Finally, the SkyViz approach asks users to specify a structured visualization context and determines the suitable types of visualization [4].

Other works are focused on the possible limitations of graphic representations. [15] argues that one of the reasons for the lack of advanced visualizations are users, who often do not know how they may represent their data. Similarly, in [16], the authors point out that users are often seen as the "weakest link" in the security chain. For this reason, the authors propose an approach that improves systems by ensuring that problems are mitigated even when the users deviate from their expected behavior. In [17] a classification of causes of pitfalls is proposed, where pitfalls are responsibility of either the designer or the user. They list three types of (negative) effects: *cognitive*, *emotional*, and *social*. The distinction between designer and user-induced mistakes is particularly valuable in pragmatic terms, as it can give immediate insights to the producers or to the evaluators of visualizations respectively. In this sense, visualization designers should look at the encoding of the visualization, while users should pay attention to pitfalls in the decoding.

It is crucial to consider the possible risks and errors that can be made during the design and generation of visualizations. [18] points out that the rendering process introduces uncertainty in three areas: *data collection process*, *algorithmic errors*, and *computational accuracy and precision*. Moreover, in [19] the authors presented an initial study about the representation of errors and uncertainties visually. The possible sources of uncertainty are acquisition, model, transformation and visualization.

It is also relevant for users to understand the visualization that they are seeing and what is the goal that this visualization pursues. Visualizations are required to be precise and easy to comprehend by users in order to minimize the interpretation errors made by users as well as designers. Visualizations must also contemplate the changing needs of users, considering high-level semantics, and reasoning about unstructured and structured data, providing easier access and better understanding of the data [20]. Moreover, although often overlooked in visualization design, requirement modeling is a paramount activity [21] that compensates for the little attention usually paid to (explicitly) representing the reasons, i.e., the *why*, in terms of motivations, rationale, objectives, and requirements.

Despite all the work done in this field, none of the approaches previously mentioned provides a methodology that guides non-expert users from the start in the specification of the most adequate set of visualizations and facilitates their generation and grouping into suitable dashboards used for the extraction of knowledge. In this sense, our proposal aims to better bridge the gap between user requirements and visual analytics.

3. Process to create visualizations automatically

In this section we describe our methodology. Fig. 1 represents the proposed process. The first model is the **User Requirements Model** presented in [5]. The main aim of this model is to capture the users' analytical needs. Since we are dealing with non-expert users, the model is completed by following a sequence of guidelines. Then, a **Data Profiling Model** [5] is obtained. This model is created by semi-automatically analyzing the features of the data sources selected in the previous model. Once both models are completed, a **Visualization Specification** is derived according to their information. This specification contains enough information to automatically derive the suitable visualization types by following [4]. This transformation generates the **Data Visualization Model** [6], which allows users to specify visualization details regardless of the underlying technology used for the implementation. Using this model, users are also able to confirm whether the proposed visualizations fulfill the essential requirements for which they were created and whether they contribute to reach the users goals by

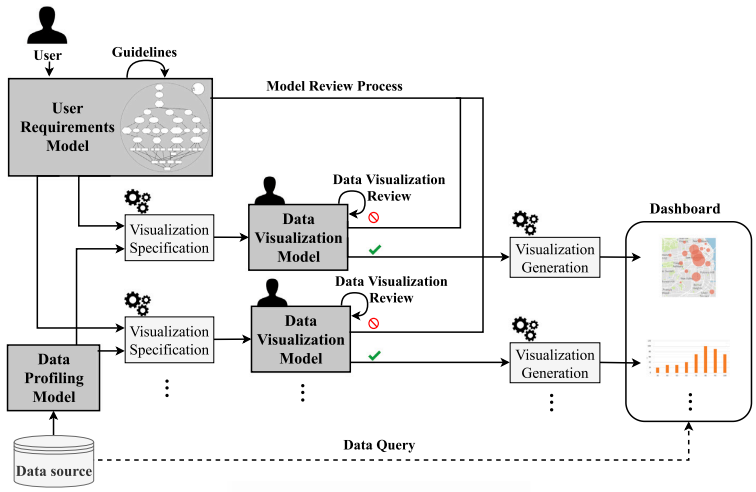


Fig. 1. Process proposed to create visualizations.

providing the necessary answers or not. If a visualization does not pass the validation, it means there are missing or ill-defined requirements. In this case, the models will be reviewed to identify which aspects were not taken into account. Otherwise, if a visualization passes the validation, it will be implemented in the selected technology and added to a dashboard.

In the following sections we explain in more detail the different components of the process.

3.1. User Requirements Model

The User Requirements Model supports the users in the definition of their data analysis objectives and helps to determine which visualization types they need to achieve these objectives. This model collects the **Interaction** and **Visualization Goals** that compose the Visualization Specification. Section 4.1 shows this model applied in an illustrative example.

In order to formally define our model, in [5] we proposed the metamodel shown in Fig. 2. This metamodel is an extension of [22], used for social and business intelligence modeling, and derived from the widely known i*, in its 2.0 version [23] and its specialized *i* for Data Warehouses* extension [24]. I* is one of the most widespread frameworks and has been successfully applied to a large number of fields, such as [25–27]. Moreover, it facilitates the communication with the user, structures the information (objectives and mechanisms to achieve them) in an intuitive way, and provides a structure to the requirements.

Elements from i* are represented in blue, elements from i* for Data Warehouses in red, and the new concepts added in yellow. In the following we describe in detail the main elements of the metamodel.

- **Visualization Actor**: the user who will interact with the system. It can been classified as either *Tech* or *Lay* depending on whether she is expert or not in complex data visualizations.
- **Business Processes**: the process at the core of users' analysis. It serves as a guideline for the definition of their *Goals*.
- **Strategic Goals**: the main objectives of the business process; achieving them translates into an improvement from a current situation into a better one.
- **Analysis Type**: it allows users to express which kind of analysis they wish to perform, as classified by [28]:
  - **Prescriptive**: How to act?
  - **Diagnostic**: Why has this happened?
  - **Predictive**: What is going to happen?
  - **Descriptive**: What to do to make it happen?
- **Decision Goals**: decisions aimed at taking appropriate actions to fulfill a strategic goal. They also explain how the associated strategic goal can be achieved.
- **Information Goals**: the lower-level abstraction goals that represent the analysis to be carried out over the available information.
- **Visualization**: a specific visualization type that will be implemented to satisfy one or more information goals.
- **Visualization Goals**: they describe the data aspects that the visualization tries to reflect. Work in [5] proposed a flowchart to aid users in finding which visualization goal they are pursuing. The flowchart contains a series of Yes/No questions to be answered by users, and provides an easy way to discern which visualization goals should be included for each visualization. The possible goals that users can choose from are [4]:
  - **Composition**: Highlight how the parts of data are composed to form a total.
  - **Order**: Order values.
  - **Relationship**: Analyze correlation.
  - **Comparison**: Establish similarities and dissimilarities.
  - **Cluster**: Emphasize the grouping into categories.
  - **Distribution**: Analyze how data are dispersed in the space.
  - **Trend**: Examine the general tendency.
  - **Geospatial**: Analyze data using a geographical map.
- **Interaction Type**: Type of interaction that the visualization must support. In [5] a series of guidelines was proposed to help users choose one or more types of interaction they want to be supported by the visualization. The possible interactions that users can choose from are [4]:
  - **Overview**: Gain an overview of the entire data collection.
  - **Zoom**: Focus on items of interest.
  - **Filter**: Quickly focus on interesting items by eliminating unwanted items.
  - **Details-on-demand**: Select an item and get its details.

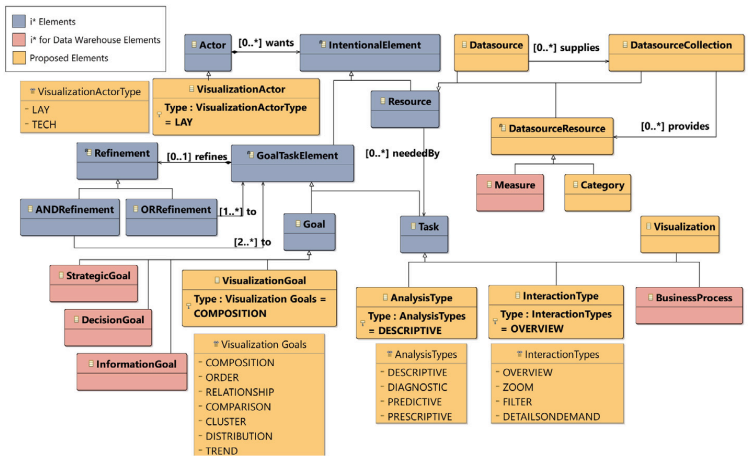


Fig. 2. Visualization specification metamodel.

- **Datasource Resource**: elements that provide relevant data from the data source.

3.2. Data Profiling Model

Following the proposed process, the next model is the Data Profiling Model. At first, in the User Requirements Model, the users have captured the data elements to be represented in each visualization. Then, the Data Profiling Model captures the data characteristics that are relevant for that visualization, such as **Dimensionality**, **Cardinality**, and **Dependent/Independent Type**. In [5], a Java implementation of a Data Analyzer to carry out data profiling was described. This software allows users to specify the data source from which they need to extract information and performs the extraction in an automated and guided way. Section 4.2 shows an example.

These characteristics are extracted in a semi-automatic manner, as explained below. First, the users specify a connection to the source dataset they wish to visualize. Then, a menu is provided where users can choose if they wish to retrieve the Data type, Cardinality, or Dimensionality of the selected column. Finally, the software returns the information requested by users. This tool has been created to collect information about the data in a simple way for users. In order to know how to delimit the values for each coordinate we have followed the approach proposed in [4], which classifies the Dimensionality, Cardinality, and Dependent/Independent Type as follows:

- **Dimensionality** is used to declare the number of variables to be visualized. Specifically, it can be *1-dimensional* when the data to represent is a single numerical value or string, *2-dimensional* when one variable depends on another, *n-dimensional* when a data object is a point in an n-dimensional space, *Tree* when each item in a collection is linked to one parent item, or *Graph* when each item in a collection is linked to an arbitrary number of items.
- **Cardinality** represents the number of data items. It is set to *Low* if this number is below a few dozens, to *High* otherwise.
- **Type of Data** is used to declare the type of each variable v. We identify each category as follows. If v is numerical, it is labeled as *Interval* if it supports the determination of equality of intervals or differences, as *Ratio* if it also has a unique and non-arbitrary zero point. If v is alphanumeric, the program shows a list of values; the user can then specify if in the list there is an order (in which case v is *Ordinal*) or not (*Nominal*).

3.3. Visualization Specification

Once the User Requirements Model and the Data Profiling Model are completed, the information coming from the models composes the Visualization Specification. We follow the SkyViz approach to discover which type of visualization suits bests each particular case, taking into account users preferences. Section 4.3 shows an example.

As described in [4], SkyViz operates by (i) asking the user to define a visualization context based on seven prioritizable coordinates for assessing her objectives and describing the dataset to be visualized; (ii) translating the visualization context into a set of suitable visualization types; (iii) asking the user to select one preferred visualization type among those proposed at the previous step; (iv) finding the best bindings between the columns of the dataset and the graphic coordinates used by the visualization type chosen by the user, and (v) asking the user to select one preferred binding among those proposed at the previous step. Specifically, as to (i), the seven coordinates composing the visualization context are filled starting the User Requirements Model and the Data Profiling Model. Step (ii) is performed based on a *suitability function* that assesses to which extent (fit, acceptable, discouraged, unfit) each visualization type is suitable for each possible value of each visualization coordinate; for instance, pie charts are discouraged for high-cardinality data, and bubble graphs are fit for n-dimensional data. The scores in the suitability function were mainly derived from the best practices found in the literature [29–31]. The set of suitable visualization types is then defined as those that are Pareto-optimal; a visualization type is Pareto-optimal when no other visualization type dominates it, being better along one coordinate and not worse along all the other coordinates. Given one preferred visualization type among the Pareto-optimal ones, step (iv) requires to decide how each variable in the dataset will be visualized, i.e., to establish a binding between each variable and each graphic coordinate of that visualization type. This is done by relying on a set of scores that indicates to which extent each graphic coordinate of each visualization type is suitable for each data type; for instance, the 'X' graphic coordinate of a single line chart is fit for variables of interval and ratio type, and the 'size' graphic coordinate of a bubble graph is unfit for variables of nominal type. Like for step (ii), the bindings proposed to the user are all the Pareto-optimal ones.

In [6] we explain in detail how to transform the Visualization Specification into a visualization following a Model-Driven Architecture (MDA) standard. As Fig. 3 shows, we transform the Visualization Specification by means of a set of model-to-model transformations using the QVT language [32], a standard from the OMG. For example, to

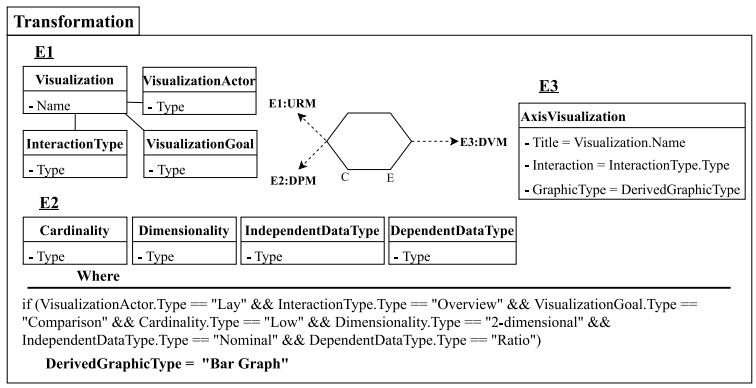


Fig. 3. Transformation of Visualization Specification into a visualization type.

derive an axis-based visualizations, our transformation generates an AxisVisualization element according to the graphic type established by the transformation. To derive this value we use the imperative part of the transformation (Where clause) according to the specific criteria established by [4] for each graphic type. The *Cardinality*, *Dimensionality*, *IndependentDataType*, and *DependentDataType* values are obtained from the **Data Profiling Model**, while the *VisualizationActor*, *InteractionType*, and *VisualizationGoal* are obtained from the **User Requirements Model**.

3.4. Data Visualization Model

Once a suitable visualization is obtained, we allow users to customize it using the Data Visualization Model [6]. This model shows a representation of the visualization and facilitates non-expert users in selecting the right visual analytics. For instance, users may define the element that determines the order in the visualization, the orientation of the visualization, or other elements as the legend, font family, and range of colors. Section 4.4 shows an example.

Users will modify the visualization through the Data Visualization Model until it meets their requirements. Once all the elements have been customized, users have to confirm whether the visualization obtained achieve their goals. If the visualization passes the validation, it will be generated as described in the next subsection. Otherwise, an unsuccessful validation generates a review of the existing User Requirements Model. The review process consists in reviewing the User Requirements Model to add or modify missing goals. If the visualization does not meet the goal for which it was created, the users will have the opportunity to redefine that goal or create a new one.

3.5. Visualization Generation

If the visualization achieves the requirements for which it was created, it will pass the validation and will be generated. This step transforms each element specified in the Data Visualization Model into a code-level specification for a graphic library. The transformation is done by means of code-to-text transformations that generate the code according to the target library. In our case, we use either the D3 JavaScript [33] or the Plotly [34] libraries to render the visualizations. The visualization, combined with the other ones generated in the process, will be grouped into a dashboard so that the user has access to all the information simultaneously. Section 4.5 shows an example of the generation of a visualization.

4. Illustrative example

This section shows the approach applied to an illustrative example based on the Police Department Incident Reports dataset [35] from the open dataset of San Francisco city [36]. In this case, the Police Department of the city requires a set of visualizations to analyze their data in order to help them improve the responsiveness of their services and reduce the incidents. We have applied our proposal to this case study by following the process in Fig. 1.

4.1. User Requirements Model

The first element involved in our process is the User Requirements Model (Section 3.1); Fig. 4 shows the result of its application. In this case, the final user is the Police Department Supervisor of the city of San Francisco, represented as a *"Lay user"* because she is not a specialist in visualization of Data Analytics. The Business Process which the user wants to analyze is *"Incidents Managment"*, and the strategic goal that she wishes to achieve is to *"Reduce incidents"*.

In order to achieve this strategic goal, the user decides to perform a *"Prescriptive analysis"* and decomposes it into two decision goals, *"Identify risk of the incident"* and *"Identify workload of police districts"*, that aim to fulfill the strategic goal.

Afterwards, the user specifies information goals for each decision goal. These goals represent the lowest level of goal abstraction. In the case of decision goal *"Identify risk of the incident"*, the user refines it into two information goals, *"Analyze neighborhoods with more incidents"* and *"Analyze the categories of incidents"*. Decision goal *"Identify workload of police districts"* is refined it into the information goal *"Analyze the number of incidents attended by police district"*.

At this moment, the user has the essential information about her goals, and she can start to define the visualization context. For each information goal, a visualization will be automatically derived in order to achieve it. Each visualization represents one or more visualization goals (aspects of the data the visualization is trying to reflect) and one or more kinds of interaction (how users would like to interact with the visualization). A set of guidelines that may be used by users to aid in the definition of these elements can be found in [5]. In this case the user has selected for the different visualizations *" Distribution"*, *"Geospatial"*, *"Comparison"*, and *"Trend"* as visualization goals and *"Overview"* as interaction type.

Finally, visualizations are decomposed into Categories and Measures that will populate them. In this case, the visualization of *"Number of incidents by neighborhood"* includes *"Neighborhoods"* as category, and *"Amount incidents"* as measure. For the visualization of *"Number of incidents by category"* the user picked *"Incident category"* as category,

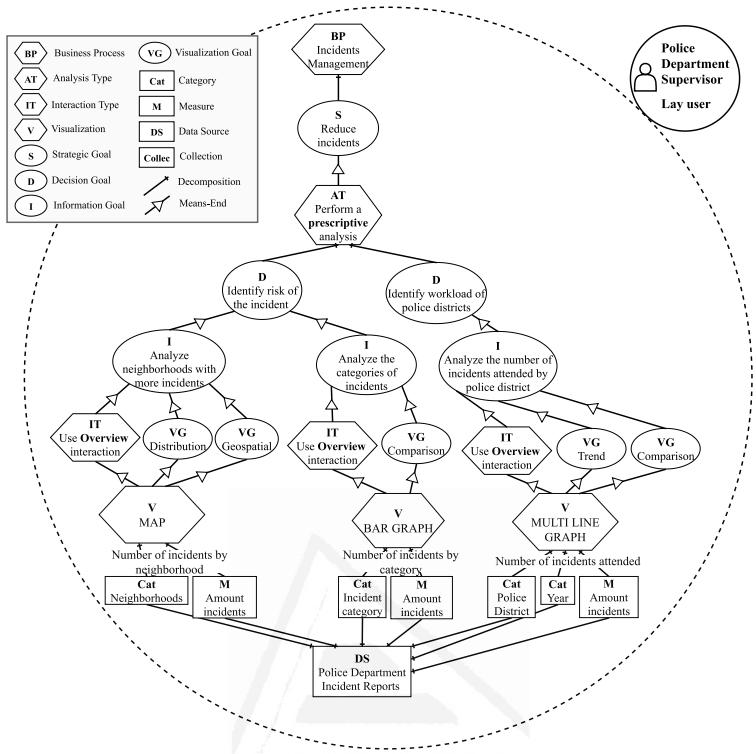


Fig. 4. Application of our User Requirements Model to the case study.

and *"Amount incidents"* as measure. Finally, in the case of *"Number of incidents attended"*, it makes use of the categories *"Police District"* and *"Year"* and the measure *"Amount incidents"*.

4.2. Data Profiling Model

Once the data sources and collections that will feed the visualizations have been defined by the user, we apply the Data Profiling Model (Section 3.2) this model determines, in a semi-automatic way, the Dimensionality, Cardinality, and Dependent/Independent Type of the data. We focus on the "Number of incidents by category" visualization from the User Requirements Model, which requires information about category *"Incident category"* and measure *"Amount incidents"*.

First, through the Data Profiling Model, the independent variable *"Incident category"* is classified as *Nominal* and the dependent variable *"Amount incidents"* as *Ratio*. Dimensionality is set as *2-dimensional*, because the user has selected 2 variables to visualize. Finally, the Cardinality is defined as *Low* because the independent variable contain 19 items to represent.

4.3. Visualization Specification

Once the Visualization Specification has all the necessary information from the previous models, it is used as input of the approach presented in [4]. This approach performs a suitability function that assesses to which extent (fit, acceptable, discouraged, unfit) each visualization type is suitable for the information stored in the Visualization Specification (Section 3.3).

Table 1 shows the Visualization Specification with its suitability scores (though for brevity we only include three visualization types, all the available visualization types were actually compared). According to the suitability scores, the most suitable visualization for the case at hand is *"Bar Graph"*. Fig. 3 shows how we use transformations to automate this process.

Table 1
Suitability scores for different visualizations types.

Vis. specification	Bar graph	Bubble graph	Single line graph
Goal: Comparison	*fit*	*fit*	*unfit*
Interaction: Overview	*fit*	*fit*	*fit*
User: Lay	*fit*	*acceptable*	*fit*
Dimensionality: 2-dim.	*fit*	*unfit*	*fit*
Cardinality: Low	*fit*	*acceptable*	*acceptable*
Independent type: Nominal	*fit*	*unfit*	*unfit*
Dependent type: Ratio	*fit*	*fit*	*fit*

4.4. Data Visualization Model

Once the visualization type has been established as *"Bar Graph"*, a Data Visualization Model (Section 3.4) is built as Fig. 5 shows to verify that the visualization satisfies the users' needs and allow them to customize it.

This model shows a mockup of the visualization with a series of characteristics that the user can customize. For example, the user has selected *"Amount incidents"* for the X axis, *"Incident category"* for the Y axis, and the orientation has been determined as *Horizontal*. When the user has finished customizing the visualization, she will have to test if the visualization makes it possible to satisfy the information goal *"Analyze the categories of incidents"* (i.e., if all the necessary information

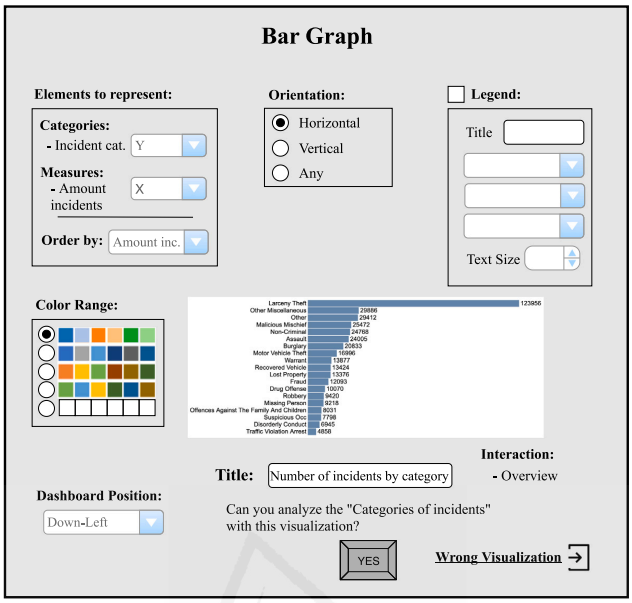


Fig. 5. Data Visualization Model.

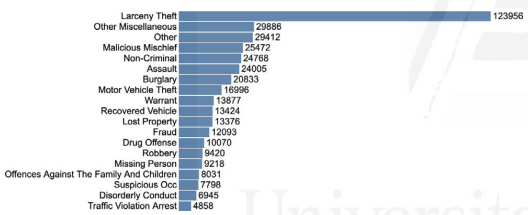


Fig. 6. Generated visualization.

can be analyzed). If the visualization passes the validation, it will be generated.

4.5. Visualization Generation

After the validation is passed, visual requirements are translated into an implementation by means of calls to the D3 JavaScript library [33] (Section 3.5), obtaining the visualization shown in Fig. 6. Consequently, this visualization, combined with those generated from the other informational goals *"Analyze neighborhoods with more incidents"* and *"Analyze the number of incidents attended by police district"* will be added to the dashboard that will enable non-expert users in data visualizations – such as the Police Department Supervisor – to monitor their processes.

5. Evaluation

In this section, we present the performance of our proposal in a controlled experiment. This experiment is part of a set of experiments for assessing the validity and impact of the proposal. In [37] it is possible to find a copy of the experimental materials in order to reproduce the experiments. We have followed the guidelines for experimentation in software engineering proposed in [38]. We have carried out our experiments with non-expert users in data visualization coming from the University of Castilla la Mancha (UCLM) Campus of Albacete (Spain) and from a small IT company located in Alicante (Spain).

5.1. Experiment context

The main goal of these experiments is to analyze the proposed methodology and evaluate its understandability and effectiveness from the viewpoint of non-expert users in data visualization. In the experiments, a total of 97 non-expert participants filled in the questionnaires. The set of participants included 2nd-year computer engineering students and employees of a technological company. In both cases none of the participants had knowledgeable skill in data visualization.

The students were recruited through an email from their teachers, and participated voluntarily. They were rewarded with 0.25 out of 10 in the final mark of the subject, however, their performance had no impact on the mark. The participants of the company participated on a voluntary basis without any benefit.

Due to the COVID measures, not all the participants could meet in the same room and this is why they had to be divided into the groups shown in Table 2. The group of instructors was composed by two developers of the method and two professors from the University of Castilla la Mancha. The professors were instructed to know the experiment and what kind of assistance they could provide. In the case of the experiment in the company, the instructors were two developers of the method. During the experiment it was not explicitly explained who were the developers of the method.

As to the assistance provided during the experiments, it was focused on the development of the exercises, not on the content. The different elements of the model were explained so that users were able to generate the goal tree by themselves. Some additional help was provided to derive the visualizations, since the experiment was made on paper and the prototype CASE tool [39] was not ready at that time. Moreover, using the prototype would have introduced additional risks and noise, since it would have been difficult to understand whether an

Table 2
Participants to the experiment.

	No. of participants	Background	Assistance
Group 1	9	UCLM	Yes
Group 2	15	UCLM	Yes
Group 3	21	UCLM	Yes
Group 4	39	UCLM	No
Group 5	13	IT company	Yes

Table 3
Main features of the experiment.

Null hypothesis	- H_{0A}: The use of the proposed methodology does not allow users to cover more analytical questions - H_{0B}: The use of the proposed methodology does not improve the set of generated visualizations - H_{0C}: Users do not find any improvement between performing the exercises with or without the proposed
Dependent variable	- Number of questions answered - Perceived value of the visualizations - Perceived improvement
Independent variable	- Whether the methodology was used or not - Whether there has been assistance in the performance of the experiment
Location	- Albacete - Alicante
Date	- October 2020
Subjects	- 84 Computer Engineering Students - 13 Employees of an IT company

improvement in the results was derived from the methodology itself or from the usage of the tool. In the case of the participants who had no assistance, there was no interaction at all between them and the instructors.

Importantly, only the help that the tool would have provided was indeed given to the participants. There was no help in applying the methodology, as this would have posed a threat to the validity of the experiment.

As Table 3 shows, the experiment seeks to discover whether (i) the proposed methodology really helps in answering more analytical questions, (ii) it increases the perceived value of the set of visualizations created, and (iii) whether users perceive an improvement when doing an exercise with or without the methodology. Then, the independent variables were defined as (i) whether the methodology was used or not and (ii) whether there has been assistance to carry out the experiment or not. And finally, the null hypothesis that the experiment tried to accept/refuse.

5.2. Experiment design

The experiment consisted of performing two exercises related to a tax collection topic and an evaluation, the first exercise without following any methodology and the second by following our methodology. Usually different cases are used, however, in this experiment we decided to use the same case to avoid fatigue effect risk, since the experiment was very long.

Firstly, before starting the exercises, we requested the participants to fill a short anonymous survey where they were asked about their age, gender, studies, and level of experience with data visualization tools. In this way we could then identify non-expert users and evaluate how our proposal improves their results. Both the survey and the experiment exercises were always filled in an anonymous manner, making us unable to identify the author behind the survey and the corresponding exercise.

Then, users performed the requested exercises. On each exercise, participants were assigned with a different strategic goal to achieve related to the tax collection topic. In the first exercise, participants were asked to define visualizations by knowing the strategic goal and having all the data available. In this first case, the participants did not follow any method. In the second exercise, participants were assigned with a different strategic goal and, in this case, they were asked to follow our methodology. Once both exercises were finished, the participants completed the evaluation by answering concrete questions that required the usage of the visualizations they had created. They also had to rate the visualizations they had defined as well as the improvement perceived when doing an exercise with or without the methodology.

Table 4
2 × 2 factorial design.

	Exercise 1 Without methodology	Exercise 2 With methodology
Experiment mode A	Strategic goal 1	Strategic goal 2
Experiment mode B	Strategic goal 2	Strategic goal 1

Everyone did the experiment first without the method and then with it. To avoid the learning effect, a 2 × 2 factorial design with confounded interaction [40] was used, as shown in Table 4. In this sense, the strategic goal to achieve and the analytical questions were swapped, i.e., the participants with Experiment Mode A received strategic goal 1 to do exercise 1 (without using our methodology) and strategic goal 2 to do exercise 2 (using our methodology). Conversely, in Experiment Mode B the participants received strategic goal 2 to do exercise 1 (without using our methodology) and strategic goal 1 to do exercise 2 (using our methodology). The experiment modes were distributed equally among the participants.

The analytical questions that participants had to answer by using the created visualizations are listed below. These questions were established by the authors during a brainstorming process. Moreover, the questions were tested in a pilot experiment and some of them were removed. For each question, participants must state whether they can or cannot answer it using their previously defined visualizations. It was not possible to answer all the questions with a single visualization since the questions were designed to force participants to use more than one visualization.

Reduce unpaid bills (Strategic Goal 1)

1. Identify the areas with most unpaid bills
2. Identify the types of taxes with most unpaid bills
3. Identify the tax records with most unpaid bills
4. Analyze the evolution over time of unpaid bills

Reduce the bill collection time (Strategic Goal 2)

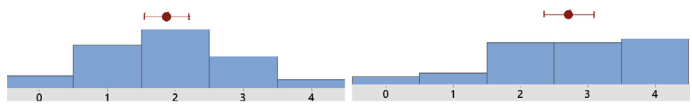
1. Identify the amount of bills paid on and after the deadline
2. Identify the types of bills that are mostly paid after the deadline
3. Indicate in which areas there are payment delays
4. Identify the most delayed tax records

Finally, in order to rate their confidence on the visualizations created, they were tasked to fill the rubric shown in Table 5. This table allows participants to communicate the perceived value of the set of visualizations created in Exercise 1 and Exercise 2 and the improvement perceived between performing the exercises with or without our methodology. This is a subjective aspect that allows us to know if users can really feel that there is an improvement in the performance of the exercise by following our methodology.

Therefore, the information collected was: (i) information regarding participants demographics, (ii) number of analytical questions answered, (iii) score of the rubric (Table 5), and (iv) time required by the participants to complete the experiment, which was only collected for statistical purposes.

Table 5
Rubric to evaluate the set of visualizations.

		Score			
		1	**2**	**3**	**4**
Ex.1	Are the visualizations useful? Is the information well represented? Are the visualizations suitable for the information?	Strongly disagree	Disagree	Agree	Totally agree
Ex.2	Are the visualizations useful? Is the information well represented? Are the visualizations suitable for the information?	Strongly disagree	Disagree	Agree	Totally agree
Ex.1 vs. Ex.2	Did you perceive any improvement in Ex. 2 over Ex. 1?	No improvement	Little improvement	Reasonable improvement	Remarkable improvement



(a) Without using our methodology (b) Using our Methodology

Fig. 7. Histogram for number of questions answered for Group 123.

Table 6
Comparison of the analytical questions and rubric results based on whether or not the methodology was used.

		Average	
		Without methodology	With methodology
Group 123	Answered questions	1.87	2.72
	Perceived value	1.82	2.74
	Perceived improvement	2.95	
Group 4	Answered questions	2.13	2.34
	Perceived value	1.89	2.58
	Perceived improvement	2.42	
Group 5	Answered questions	1.08	2.15
	Perceived value	1.85	2.62
	Perceived improvement	2.92	

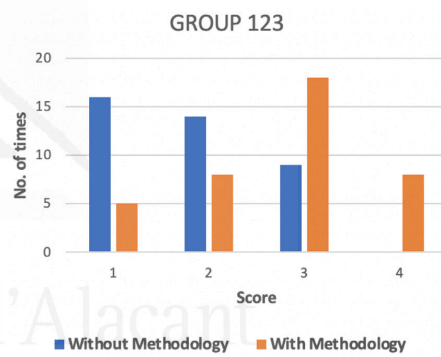


Fig. 8. Perceived value of visualizations for Group 123.

5.3. Experiment results

After manually transcribing the survey, the analytical questions, and the rubric for a subsequent analysis, we obtained the results shown in Table 6. We have grouped the results in:

- **Group 123:** 45 Computer Engineering students from UCLM who were assisted while carrying out the experiment.
- **Group 4**: 39 Computer Engineering students from UCLM who were not given assistance in carrying out the experiment.
- **Group 5**: 13 Employees of a small IT company who were assisted while carrying out the experiment.

In the following, the results from each group will be analyzed.

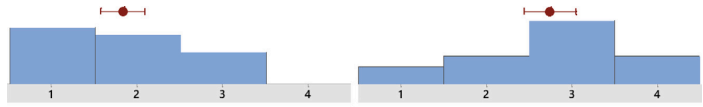
5.3.1. Group 123 - Students with assistance

Group 123 included 45 Computer Engineering Students from the University of Castilla la Mancha. In this case, we gave assistance to the participants through a detailed explanation of the methodology and solved all their doubts.

According to the results obtained, shown in Table 6, the set of visualizations generated without following any methodology can answer 1.87/4 (47%) of the specific questions proposed, while this number grows until 2.72/4 (68%) coverage when following the proposed method. Furthermore, a 2-Sample T-Test was performed with an alpha of 0.05. Thanks to this test, we could conclude that the mean number of questions answered differs at the 0.05 level of significance, with a p-value < 0.001. Therefore, for Group 123, with a 95% confidence level, we can reject the null hypothesis *"H_{0A} - The use of the proposed methodology does not allow users to cover more analytical questions"*. Thus, the number of analytical questions answered by using the methodology is significantly different (higher) from the number of analytical questions answered without using the methodology. Moreover, Fig. 7(a) reflects the normality of the data since it corresponds to the structure of a Gaussian distribution [41].

In order to accept or reject the null hypothesis *"H_{0B}: The use of the proposed methodology does not improve the set of generated visualizations"*, the answers of the rubric of Table 5 have been analyzed. As Table 6 shows, the participants scored the visualizations created without methodology with an average 1.82/4. Comparatively, the visualizations generated using the methodology presented were scored with an average of 2.74/4. Fig. 8 represents the perceived value of the resulting visualizations.

The 2-Sample T-Test was performed with an alpha of 0.05. Thanks to this test, we could conclude that in the case of the perceived value of the visualizations the means differ at the 0.05 level of significance, with



(a) Without using our methodology (b) Using our Methodology

Fig. 9. Histogram for the confidence on the visualizations score for Group 123.

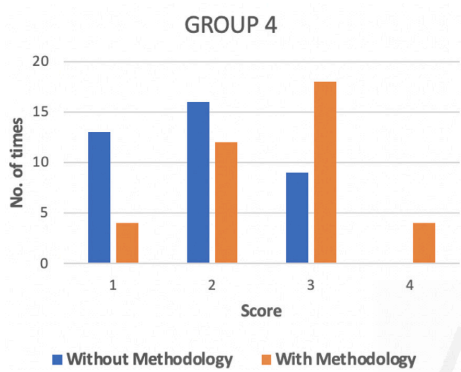


Fig. 10. Perceived value of visualizations for Group 4.

a p-value < 0.001. Therefore, for Group 123, with a 95% confidence level, we can reject the null hypothesis *"H_{0B}: The use of the proposed methodology does not improve the set of generated visualizations"*, meaning that, for this group, the perceived value of the visualizations is indeed higher when using our methodology. As in the previous case, Fig. 9 reflects the normality of the data, as well as the difference of the averages. Therefore, the normality of our data is confirmed. We conclude that the results show a statistical significance that confirms the impact of the methodology proposed.

5.3.2. Group 4 - Students without assistance

Group 4 is composed of 39 Computer Engineering Students from the University of Castilla la Mancha. In this case we let them carry out the experiment without offering them any assistance.

In accordance with the results obtained (Table 6), the set of visualizations generated without following any methodology can answer 2.13/4 (53%) of the specific questions proposed, while this number grows until 2.34/4 (59%) when following the proposed method. However, in this case, the 2-Sample T-Test concludes that with a p-value of 0.430, the number of questions answered is not significantly different. Therefore, for "Group 4", we cannot reject the null hypothesis *"H_{0A} - The use of the proposed methodology does not allow users to cover more analytical questions"*.

In order to accept or reject the null hypothesis *"H_{0B}: The use of the proposed methodology does not improve the set of generated visualizations"*, the answers of the rubric of Table 5 have been analyzed. As Table 6 shows, the participants scored the visualizations created without methodology with an average 1.89/4. Comparatively, the visualizations generated using the methodology presented were scored with an average of 2.58/4. Fig. 10 represents the perceived value of the resulting visualizations.

The 2-Sample T-Test was performed with an alpha of 0.05. Thanks to this test, we can conclude that in the case of the perceived value of the visualizations the means differ at the 0.05 level of significance, with a p-value < 0.001. Therefore, for Group 4, with a 95% confidence level, we can reject the null hypothesis *"H_{0B}: The use of the proposed methodology does not improve the set of generated visualizations"*, meaning that in this group the perceived value of the visualizations is not the same using or not using our proposed methodology.

Fig. 11 reflects the normality of the data, as well as the difference of the averages. Therefore, the normality of our data is confirmed.

5.3.3. Group 5 - Company employees with assistance

The last group, number 5, was composed of 13 employees from the small technological company. In this case, we gave assistance to the participants through a detailed explanation of the methodology and solved all their doubts.

According to the results obtained, shown in Table 6, the set of visualizations generated without following any methodology can answer the 1.08/4 (27%) of the specific questions proposed, while this number grows until 2.15/4 (54%) coverage when following the proposed method. Furthermore, a 2-Sample T-Test was performed with an alpha of 0.05. Thanks to this test, we could conclude that in the case of the number of questions answered means differ at the < 0.05 level of significance, with a p-value of 0.019. Therefore, for "Group 5", with a 95% confidence level, we can reject the null hypothesis *"H_{0A} - The use of the proposed methodology does not allow users to cover more analytical questions"*, meaning that the number of analytical questions answered by using the methodology is significantly different (again, higher) from the number of analytical questions answered without using the methodology.

Fig. 12 reflects the normality of the data, as well as the difference of the averages. Therefore, the normality of our data is confirmed.

In order to accept or reject the null hypothesis *"H_{0B}: The use of the proposed methodology does not improve the set of generated visualizations"*, the answers of the rubric of Table 5 have been analyzed. As Table 6 shows, participants scored the visualizations created without methodology with an average 1.85/4. Comparatively, the visualizations generated using the methodology presented were scored with an average of 2.62/4. Fig. 13 represents the perceived value of the resulting visualizations.

The 2-Sample T-Test was performed with an alpha of 0.05. Thanks to this test, we could conclude that in the case of the perceived value of the visualizations the means differ at the 0.05 level of significance, with a p-value of 0.047. Therefore, for "Group 5", with a 95% confidence level, we can reject the null hypothesis *"H_{0B}: The use of the proposed methodology does not improve the set of generated visualizations"*, meaning that in Group 5 the perceived value of the visualizations is higher when following our proposed methodology.

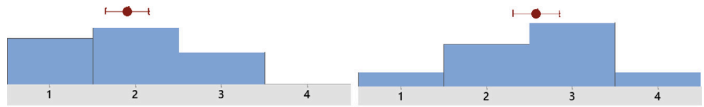
Finally, as in the previous cases, Fig. 14 reflects the normality of the data, as well as the difference of the averages. Therefore, the normality of our data is confirmed.

In conclusion, the T-test results showed statistical significance for the results obtained, confirming the impact of the methodology proposed. Fig. 15 summarizes the score given to the improvement of one method over the other through the third question of rubric shown in Table 5. In most cases the participants have detected an improvement when using our methodology.

5.3.4. Analysis of visualizations

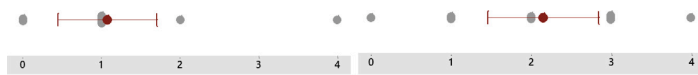
Finally, we analyzed the visualizations generated freely and those generated using our methodology.

The first outcome is that, by following our methodology, a larger number of visualizations were created than by creating them freely.



(a) Without using our methodology (b) Using our Methodology

Fig. 11. Histogram for the confidence on the visualization score for Group 4.



(a) Without using our methodology (b) Using our Methodology

Fig. 12. Distribution of data of number of questions answered for Group 5.

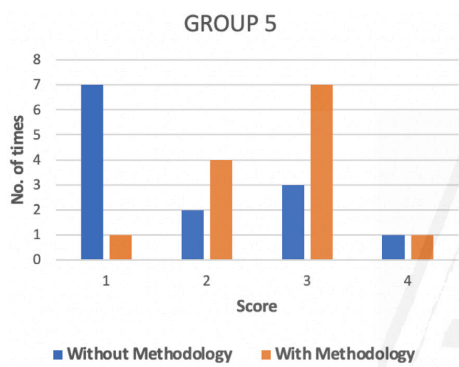


Fig. 13. Perceived value of visualizations for Group 5.

A total of 205 visualizations were created freely (an average of 2.11 per participant), while 257 visualizations were created by following our methodology (an average of 2.65 per participant).

Moreover, we have analyzed the types of visualizations selected in each case. When the participants did the exercise freely, the most used visualization types were Column Graph, Pie Chart, and Map as Fig. 16(a) shows. Nevertheless, when the participants did the exercise following our methodology, the most used visualizations were Column Graph, Map, and Bubble Graph as Fig. 16(b) shows.

Therefore, we can conclude that: (i) following the methodology, participants tend to use more visualizations; (ii) the visualization type most used by the participants is also the one most recommended by our methodology; and (iii) when participants use visualizations that are not suitable for non-expert users, such as histograms, it is common to create erroneous visualizations that do not really represent what they expected.

5.4. Meta-analysis

In this section the results from the different groups are discussed. Table 7 summarizes the results obtained.

The results increase our confidence about the utility of our methodology, because (i) it allows users to cover more analytical questions, (ii) it improves the set of generated visualizations, and (iii) users find improvements when they use it to execute the exercises. For the group that carried out the experiment without any assistance (Group 4), it was not possible to verify statistically that the use of the proposed methodology allows users to cover more analytical questions. Therefore, given the positive results obtained in the remaining groups, we can infer that some assistance or prior training is required for the effective application of the methodology. In addition, these results point in the direction of emphasizing the development and usage of a user-friendly tool to apply our proposal more effectively, reducing the users' knowledge burden and improving the results obtained.

Table 7
Summarized results.

	Group 123	Group 4	Group 5
H_{0A}: The use of the proposed methodology does not allow users to cover more analytical questions	Rejected	Not rejected	Rejected
H_{0B}: The use of the proposed methodology does not improve the set of generated visualizations	Rejected	Rejected	Rejected
H_{0C}: Users do not find any improvement between performing the exercises with or without the proposed methodology	Rejected	Rejected	Rejected

6. Validity threats

In this section, we summarize the main limitations and validity threats for the performed experiments. Although we did our best to avoid that the outcome is affected by undesired factors, there are some aspects that must be taken into account when reproducing these experiments:

- When performing the experiments, we had a data analyst supporting non-expert users in order to aid in following the methodology. Such actor may not be always available, which may alter the results (i.e., Group 4). We are working to a user-friendly CASE tool to verify that users are able to define visualization requirements completely on their own.
- Our proposal is meant to be context-independent. We have applied it in educational, economic, smart cities, and gas turbine contexts. However, we have not applied our proposal yet in a full set of contexts, so there may be some specific user profiles we have not considered yet.
- Our methodology increases the capability to answer analytical questions. However, it is still recommended that the user who defines the visualizations is an expert in the application domain for which the visualizations are required.
- We rely on [4] to derive suitable visualization types. This means that our proposal inherits the associated limitations when deriving the visualizations. One of such limitations is that not all visualization types are supported. Furthermore, if a significantly

(a) Without using our methodology (b) Using our Methodology

Fig. 14. Distribution of data of confidence on the visualization score for Group 5.

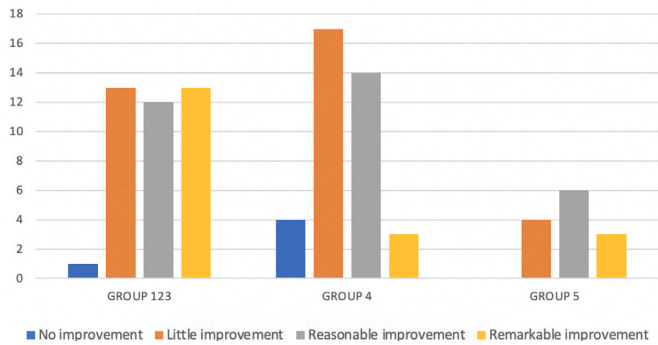
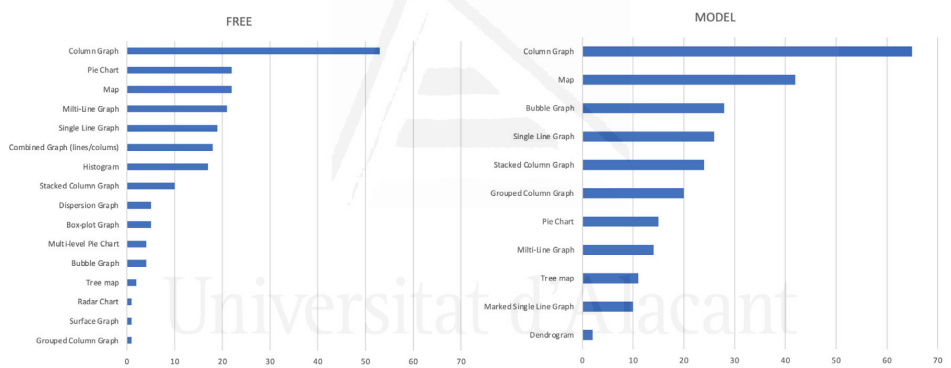


Fig. 15. Perceived improvement results.



(a) Without using our methodology (b) Using our methodology

Fig. 16. Most used visualizations.

larger number of visualization types were to be included, the seven coordinates we rely on might no longer be sufficient to distinguish them.

- The participants in the experiment received a predefined template (essentially, a tree with empty nodes) as a guide to facilitate the creation of the User Requirements Model. Then they completed the model independently by filling the nodes and adding or eliminating branches as necessary.
- Although the objective of the experiment was to test our approach on non-expert users only, the experience of the users can be considered a validity threat.

7. Conclusions and future work

The volume of data that needs be analyzed and interpreted is continuously growing. Data visualization plays a key role in this analysis. However, finding the most effective visualizations is a difficult task. Normally, users are not experts in data visualization, and they rarely know which is the visualization type that will best suit them, nor they know exactly what information they are trying to extract from them. Unfortunately, there is a lack of methodologies that guide non-expert users, taking into account their analysis goals to define the visualizations they need. For this reason, in this paper we have presented a process that helps non-expert users define their analytical goals and achieve them by automatically deriving the visualizations that best suit a certain context.

Compared to other approaches, our proposal covers the whole process, from defining user requirements to implementing visualizations. Therefore, the great advantage of our proposal is that non-expert users will be guided to reflect their analytical needs and automatically obtain a set of visualizations that will help them to achieve their goals.

To evaluate the impact of our proposal, we have presented a case study and performed a set of experiments with non-expert users in data visualization. The experiments have been carried out by 97 participants, including 84 Computer Engineering Students and 13 employees of a technological company, all of them non-expert in data visualization. These experiments confirmed the validity of our proposal since it has been shown that our methodology (i) allows users to cover more analytical questions, (ii) improves the set of generated visualizations and, (iii) users themselves perceive improvements when adopting our

Table 8
Experiment data 1.

Group	Id	Mode	No. of questions answered without methodology	No. of questions answered with methodology	Value of visualizations without methodology	Value of visualizations with methodology	Comparative	Time
1	1	A	0	4	1	4	4	49
1	2	A	2	3	2	4	4	49
1	3	A	2	4	3	3	4	56
1	4	A	1	4	2	3	3	54
1	5	B	1	3	1	3	3	53
1	6	B	2	2	1	2	2	
1	7	B	2	2	1	2	2	
1	8	B	3	4	2	4	4	49
1	9	B	1	4	1	3	4	47
2	1	A	4	4	2	3	4	54
2	2	A	1	4	3	4	3	58
2	3	A	2	2	1	1	2	61
2	4	A	1	4	2	4	4	60
2	5	A	1	3	1	3	3	60
2	6	A	1	4	2	3	3	68
2	7	A	3	3	3	2	2	63
2	8	A						62
2	9	B	2	4	1	4	4	63
2	10	B	2	2	3	4	2	62
2	11	B	4	3	3	3	2	62
2	12	B	1	1	3	3	3	59
2	13	B	2	1	1	3	3	59

Table 9
Experiment data 2.

Group	Id	Mode	No. of questions answered without methodology	No. of questions answered with methodology	Value of visualizations without methodology	Value of visualizations with methodology	Comparative	Time
2	14	B						57
2	15	B	3	1	2	3	4	54
3	1	A	2	4	1	2	2	44
3	2	A	1	2	2	3	2	45
3	3	A	3	3	2	1	2	47
3	4	A	1	0	3	3	2	62
3	5	A	3	2	2	1	2	60
3	6	A	1	3	2	3	4	50
3	7	A	3	2	3	3	3	50
3	8	A						50
3	9	A	2	0	2	1	1	51
3	10	A	2	3	3	4	2	51
3	11	A						55
3	12	B	3	3	1	2	4	62
3	13	B	2	3	2	3	3	57
3	14	B						57
3	15	B	0	2	1	2	4	55
3	16	B	3	3	1	3	3	53
3	17	B	2	2	2	2	3	50
3	18	B	0	2	1	1	2	49
3	19	B						47
3	20	B	2	4	1	2	4	47

methodology. Although the majority of user groups in the experiments have shown a statistical significance in favoring the methodology, for the group that carried out the experiment without any assistance it has not been possible to verify statistically that the use of the proposed methodology allows users to cover more analytical questions. The other improvements have also been confirmed with this group.

Therefore, considering that the assistance in following the method has a positive impact on its application, we are implementing a user-friendly tool [39]. As part of our future work we are going to test the usability of the tool through new controlled experiments. This will allow us to adjust the tool to users' needs. Moreover, we will explore the possibility of taking into account changing needs to our methodology.

CRediT authorship contribution statement

Ana Lavalle: Conceptualization, Methodology, Investigation, Data curation, Writing - original draft, Visualization. **Alejandro Maté:** Conceptualization, Methodology, Writing - review & editing, Supervision, Project administration. **Juan Trujillo:** Conceptualization, Methodology, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition. **Miguel A. Teruel:** Formal analysis, Investigation, Data curation, Writing - review & editing. **Stefano Rizzi:** Conceptualization, Methodology, Writing - review & editing.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.infsof.2021.106592.

Acknowledgments

This work has been co-funded by the ECLIPSE-UA (RTI2018-0942-83-B-C32) project funded by Spanish Ministry of Science, Innovation. Ana Lavalle holds an Industrial PhD Grant (I-PI 03-18) co-funded by the University of Alicante, Spain and the Lucentia Lab Spin-off Company.

Table 10
Experiment data 3.

Group	Id	Mode	No. of questions answered without methodology	No. of questions answered with methodology	Value of visualizations without methodology	Value of visualizations with methodology	Comparative	Time
3	21	B	2	2	1	3	3	45
4	1	A	2	4	2	4	4	46
4	2	A	2	2	2	2	2	46
4	3	A	2	2	2	3	3	46
4	4	A	3	2	2	1	1	47
4	5	A	3	3	2	3	3	48
4	6	A	3	0	2	3	3	48
4	7	A	1	1	2	2	3	47
4	8	A	3	2	2	3	3	50
4	9	A	1	2	1	1	2	50
4	10	A	3	4	3	3	2	53
4	11	A	2	3	3	4	2	52
4	12	A	0	3	1	2	2	54
4	13	A	2	3	3	2	2	56
4	14	A	3	4	3	3	2	55
4	15	A	2	4	2	3	4	58
4	16	A	2	2	2	2	2	57
4	17	A	2	3	1	3	3	66
4	18	A	1	4	2	3	3	60
4	19	B	3	2	2	3	3	65
4	20	B	3	4	2	2	3	64
4	21	B	1	1	1	2	2	64

Table 11
Experiment data 4.

Group	Id	Mode	No. of questions answered without methodology	No. of questions answered with methodology	Value of visualizations without methodology	Value of visualizations with methodology	Comparative	Time
4	22	B	1	3	3	3	3	62
4	23	B						62
4	24	B	0	2	1	2	2	61
4	25	B	4	3	1	3	4	58
4	26	B	1	1	2	2	2	56
4	27	B	4	3	3	3	3	56
4	28	B	3	2	1	3	3	55
4	29	B	2	2	3	2	1	55
4	30	B	0	0	1	1	1	55
4	31	B	0	2	1	3	2	54
4	32	B	1	2	1	2	2	51
4	33	B	3	3	3	4	3	50
4	34	B	3	2	1	3	2	50
4	35	B	3	1	3	3	2	50
4	36	B	4	4	2	4	3	49
4	37	B	1	1	1	1	1	49
4	38	B	3	1	2	3	2	47
4	39	B	4	2	1	2	2	46
5	1	B	0	3	1	3	4	
5	2	B	1	3	1	3	3	
5	3	B	2	1	3	3	2	
5	4	B	1	2	2	3	2	

Table 12
Experiment data 5.

Group	Id	Mode	No. of questions answered without methodology	No. of questions answered with methodology	Value of visualizations without methodology	Value of visualizations with methodology	Comparative	Time
5	5	B	0	3	3	3	3	
5	6	B	1	2	1	2	3	
5	7	A	4	2	4	1	2	
5	8	A	1	3	2	2	3	
5	9	A	1	1	1	2	2	
5	10	A	1	0	1	2	4	
5	11	A	0	1	1	3	3	
5	12	A	1	3	3	3	3	
5	13	A	1	4	1	4	4	60

We would like to thank Elena Navarro, Pascual González and Victor López from the University of Castilla-La Mancha (Spain) for their collaboration in the experiment.

Appendix. Experiment data

See Tables 8–12.

References

[1] M. Staron, H.A. Sahraoui, A. Telea, Special section on visual analytics in software engineering, Inf. Softw. Technol. 98 (2018) 117, http://dx.doi.org/10.1016/j.infsof.2018.03.001, URL https://doi.org/10.1016/j.infsof.2018.03.001.

[2] R. Arterburn, Data visualization market growth, future prospects and competitive analysis (2020-2027) — fortune business insights, 2020, URL https://fresnobserver.com/data-visualization-market-growth-future-prospects-and-competitive-analysis-2020-2027-fortune-business-insights/.

[3] E.G. Caldarola, A.M. Rinaldi, Improving the visualization of wordnet large lexical database through semantic tag clouds, in: International Congress on Big Data, IEEE, 2016, pp. 34–41.

[4] M. Golfarelli, S. Rizzi, A model-driven approach to automate data visualization in big data analytics, Inform. Vis. 19 (1) (2020) http://dx.doi.org/10.1177/1473871619858933, URL https://doi.org/10.1177/1473871619858933.

[5] A. Lavalle, A. Maté, J. Trujillo, S. Rizzi, Visualization requirements for business intelligence analytics: A goal-based, iterative framework, in: 27th IEEE International Requirements Engineering Conference, RE 2019, Jeju Island, Korea (South), September 23-27, 2019, 2019, pp. 109–119, http://dx.doi.org/10.1109/RE.2019.00022, URL https://doi.org/10.1109/RE.2019.00022.

[6] A. Lavalle, A. Maté, J. Trujillo, Requirements-driven visualizations for big data analytics: A model-driven approach, in: Conceptual Modeling - 38th International Conference, ER 2019, Salvador, Brazil, November 4-7, 2019, Proceedings, Springer, 2019, pp. 78–92, http://dx.doi.org/10.1007/978-3-030-33223-5_8, URL https://doi.org/10.1007/978-3-030-33223-5_8.

[7] A. Fernandez, A. Bergel, A domain-specific language to visualize software evolution, Inf. Softw. Technol. 98 (2018) 118–130, http://dx.doi.org/10.1016/j.infsof.2018.01.005, URL https://doi.org/10.1016/j.infsof.2018.01.005.

[8] M. Alshakhouri, J. Buchan, S.G. MacDonell, Synchronised visualisation of software process and product artefacts: Concept, design and prototype implementation, Inf. Softw. Technol. 98 (2018) 131–145, http://dx.doi.org/10.1016/j.infsof.2018.01.008, URL https://doi.org/10.1016/j.infsof.2018.01.008.

[9] H. Santos, V. Dantas, V. Furtado, P. Pinheiro, D.L. McGuinness, From data to city indicators: A knowledge graph for supporting automatic generation of dashboards, in: The Semantic Web - 14th International Conference, ESWC, Springer, 2017, pp. 94–108, http://dx.doi.org/10.1007/978-3-319-58451-5_7, URL https://doi.org/10.1007/978-3-319-58451-5_7.

[10] M. Kintz, M. Kochanowski, F. Koetter, Creating user-specific business process monitoring dashboards with a model-driven approach, in: Proceedings of the 5th International Conference on Model-Driven Engineering and Software Development, MODELSWARD, 2017, pp. 353–361, http://dx.doi.org/10.5220/0006135203530361, URL https://doi.org/10.5220/0006135203530361.

[11] A. Vázquez-Ingelmo, F.J. García-Peñalvo, R. Therón, Application of domain engineering to generate customized information dashboards, in: Learning and Collaboration Technologies. Learning and Teaching - 5th International Conference, LCT, Springer, 2018, pp. 518–529, http://dx.doi.org/10.1007/978-3-319-91152-6_40, URL https://doi.org/10.1007/978-3-319-91152-6_40.

[12] K. Börner, Atlas of Knowledge, MIT Press Cambridge, MA, 2014.

[13] S. Madhu Sudhan, J. Chandra, IBA Graph selector algorithm for big data visualization using defence dataset, Int. J. Sci. Eng. Res. 4 (3) (2013).

[14] O. Peña, U. Aguilera, D. López-de-Ipiña, Exploring LOD through metadata extraction and data-driven visualizations, Program 50 (3) (2016) 270–287, http://dx.doi.org/10.1108/PROG-12-2015-0079, URL https://doi.org/10.1108/PROG-12-2015-0079.

[15] C.C. Gray, W.J. Teahan, D. Perkins, Understanding our analytics: A visualization survey, J. Learn. Anal. (2017) in press.

[16] T.T. Tun, A. Bennaceur, B. Nuseibeh, OASIS: weakening user obligations for security-critical systems, in: 28th IEEE International Requirements Engineering Conference, RE 2020, IEEE, 2020, pp. 113–124, http://dx.doi.org/10.1109/RE48521.2020.00023, URL https://doi.org/10.1109/RE48521.2020.00023.

[17] S. Bresciani, M.J. Eppler, The pitfalls of visual representations: A review and classification of common errors made while designing and interpreting visualizations, Sage Open 5 (4) (2015) http://dx.doi.org/10.1177/2158244015611451, URL https://doi.org/10.1177/2158244015611451.

[18] A. Pang, C.M. Wittenbrink, S.K. Lodha, Approaches to uncertainty visualization, Vis. Comput. 13 (8) (1997) 370–390, http://dx.doi.org/10.1007/s003710050111, URL https://doi.org/10.1007/s003710050111.

[19] C.R. Johnson, A.R. Sanderson, A next step: Visualizing errors and uncertainty, IEEE Comput. Graph. Appl. 23 (5) (2003) 6–10, http://dx.doi.org/10.1109/MCG.2003.1231171, URL https://doi.org/10.1109/MCG.2003.1231171.

[20] M. Aufaure, What's up in business intelligence? A contextual and knowledge-based perspective, in: Conceptual Modeling - 32th International Conference, ER 2013, Springer, 2013, pp. 9–18.

[21] D.A.C. Quartel, W. Engelsman, H. Jonkers, M. van Sinderen, A goal-oriented requirements modelling language for enterprise architecture, in: Proceedings of the 13th IEEE International Enterprise Distributed Object Computing Conference, EDOC, IEEE, 2009, pp. 3–13, http://dx.doi.org/10.1109/EDOC.2009.22, URL https://doi.org/10.1109/EDOC.2009.22.

[22] M.A. Teruel, A. Maté, E. Navarro, P. González, J.C. Trujillo, The new era of business intelligence applications: Building from a collaborative point of view, Bus. Inform. Syst. Eng. 61 (5) (2019) 615–634, http://dx.doi.org/10.1007/s12599-019-00578-3, URL https://doi.org/10.1007/s12599-019-00578-3.

[23] F. Dalpiaz, X. Franch, J. Horkoff, Istar 2.0 language guide, CoRR, abs/1605.07767, 2016, [arXiv:1605.07767] URL http://arxiv.org/abs/1605.07767.

[24] A. Maté, J. Trujillo, X. Franch, Adding semantic modules to improve goal-oriented analysis of data warehouses using I-star, J. Syst. Softw. 88 (2014) 102–111, http://dx.doi.org/10.1016/j.jss.2013.10.011, URL https://doi.org/10.1016/j.jss.2013.10.011.

[25] J. Horkoff, E. Yu, Interactive goal model analysis for early requirements engineering, Requir. Eng. 21 (1) (2016) 29–61, http://dx.doi.org/10.1007/s00766-014-0209-8, URL https://doi.org/10.1007/s00766-014-0209-8.

[26] C. Cares, X. Franch, A metamodelling approach for i* model translations, in: H. Mouratidis, C. Rolland (Eds.), Advanced Information Systems Engineering - 23rd International Conference, CAiSE 2011, London, UK, June 20-24, 2011. Proceedings, in: Lecture Notes in Computer Science, 6741, Springer, 2011, pp. 337–351, http://dx.doi.org/10.1007/978-3-642-21640-4_26, URL https://doi.org/10.1007/978-3-642-21640-4_26.

[27] M. Asadi, G. Gröner, B. Mohabbati, D. Gasevic, Goal-oriented modeling and verification of feature-oriented product lines, Softw. Syst. Model. 15 (1) (2016) 257–279, http://dx.doi.org/10.1007/s10270-014-0402-8, URL https://doi.org/10.1007/s10270-014-0402-8.

[28] A. Shi-Nash, D.R. Hardoon, Data analytics and predictive analytics in the era of big data, in: Internet of Things and Data Analytics Handbook, Wiley Online Library, 2017, pp. 329–345, http://dx.doi.org/10.1002/9781119173601.ch19.

[29] A. Abela, Advanced Presentations by Design, Pfeiffer, 2008.

[30] R. Marty, Applied Security Visualization, Addison-Wesley, 2009.

[31] K. Börner, Atlas of Knowledge: Anyone Can Map, MIT Press, 2015.

[32] Object Management Group (OMG), MOF 2.0 query/view/transformation specification, 2016, https://www.omg.org/spec/QVT/1.3/PDF.

[33] M. Bostock, Data-driven documents, 2019, URL https://d3js.org/.

[34] Plotly, Dash, 2019, URL https://plot.ly/.

[35] Police Department Incident Reports, 2020, https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783. (Accessed 15 November 2020).

[36] San Francisco Open Data portal, 2020, https://datasf.org/opendata/. (Accessed 15 November 2020).

[37] A. Lavalle, A. Maté, J. Trujillo, M.A. Teruel, S. Rizzi, Experimental materials, Zenodo, 2021, http://dx.doi.org/10.5281/zenodo.4637123, URL https://doi.org/10.5281/zenodo.4637123.

[38] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell, A. Wesslén, Experimentation in Software Engineering, Springer Science & Business Media, 2012.

[39] L. Márquez Carpintero, A. Lavalle, A. Maté, J. Trujillo, Grafik – Herramienta para la generación automática de visualizaciones de dashboards en Data Analytics, 2020, URL http://hdl.handle.net/10045/109458.

[40] B.J. Winer, Statistical Principles in Experimental Design, McGraw-Hill Book Company, 1962.

[41] N.R. Goodman, Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction), The Ann. Math. Statist. 34 (1) (1963) 152–177.

Capítulo 9

An Approach to Automatically Detect and Visualize Bias in Data Analytics

Lavalle, A., Maté, A., & Trujillo, J. (2020, March). An Approach to Automatically Detect and Visualize Bias in Data Analytics. In *Proceedings of the 22nd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data co-located with EDBT/ICDT (DOLAP 2020)* (vol. 2572 of CEUR Workshop Proceedings, pp. 84–88). CEUR-WS.org.

Conference Rating por GII-GRIN-SCIE: **Clase 2**
Conference Rating por CORE: **B**

Disponible en:
URL: `http://ceur-ws.org/Vol-2572/short11.pdf`

An Approach to Automatically Detect and Visualize Bias in Data Analytics

Ana Lavalle
Lucentia Research (DLSI)
University of Alicante
San Vicente del Raspeig, Spain
Lucentia Lab, Alicante, Spain
alavalle@dlsi.ua.es

Alejandro Maté
Lucentia Research (DLSI)
University of Alicante
San Vicente del Raspeig, Spain
Lucentia Lab, Alicante, Spain
amate@dlsi.ua.es

Juan Trujillo
Lucentia Research (DLSI)
University of Alicante
San Vicente del Raspeig, Spain
Lucentia Lab, Alicante, Spain
jtrujillo@dlsi.ua.es

ABSTRACT

Data Analytics and Artificial Intelligence (AI) are increasingly driving key business decisions and business processes. Any flaws in the interpretation of analytic results or AI outputs can lead to significant economic loses and reputation damage. Among existing flaws, one of the most often overlooked is the use biased data and imbalanced datasets. When unadverted, data bias warps the meaning of data and has a devastating effect on AI results. Existing approaches deal with data bias by constraining the data model, altering its composition until the data is no longer biased. Unfortunately, studies have shown that crucial information about the nature of data may be lost during this process. Therefore, in this paper we propose an alternative process, one that detects data biases and presents biased data in a visual way so that the user can comprehend how data is structured and decide whether or not constraining approaches are applicable in his context. Our approach detects the existence of biases in datasets through our proposed algorithm and generates a series of visualizations in a way that is understandable for users, including non-expert ones. In this way, users become aware not only of the existence of biases in the data, but also how they may impact their analytics and AI algorithms, thus avoiding undesired results.

1 INTRODUCTION

Nowadays, Data Analytics have become a key component of many business processes. Whether driving business decisions or offering new services through Artificial Intelligence (AI) algorithms, data serves as the main resource for improving business performance. Therefore, any flaws within the data or its use will be translated into significant performance and economic loses.

One of such flaws is data bias and the use of imbalanced datasets. When unadverted, data bias can significantly affect the interpretation of data, and has devastating impact on AI results as recently reported by the Gartner Group [6]. One area where biases lead to life-threatening consequences is Healthcare, where identifying as healthy a patient that is incubating a severe illness may delay its treatment [2].

As such, data bias has become an important concern in the community, with Big companies like Amazon, Facebook, Microsoft, Google, etc. investing resources and effort to tackle the problem. Amazon Web Services [23] has published information about fairness in their machine-learning services in terms of accuracy, false positive and false negative rates. Facebook [19] has shown one of its internal anti-bias software tools, "Fairness Flow" which measures how a model interacts with specific groups.

Unfortunately, most approaches developed until now are mainly focused on machine-learning and rebalancing the biased datasets. As [7] argues, the fairness of predictions should be evaluated in context of the data, and unfairness induced by inadequate samples sizes or unmeasured predictive variables should be addressed through data collection rather than by constraining the model. As such, a general approach that automatically warns the user of the existence of biases and lets her analyze the data from different perspectives without altering the dataset is missing.

Therefore, in this paper we focus our work on detecting and presenting in a humanly understandable way the existence of data bias and imbalanced datasets, with a special focus on enabling the analysis through data analytics without altering the dataset.

Our approach complements our previous work [15] [14] where we presented an iterative Goal-Based modeling approach based on the i* language for the automatic derivation of data visualizations and we aligned it with the Model Driven Architecture (MDA) in order to facilitate the creation of the right visual analytics for non-expert users. Now, we include a Biases Detection Process that automatically detects the existence of biases in the datasets and enables users to measure them and select those ones which are relevant to them. Our process includes a novel algorithm that takes into account the scope of the analysis, detects biases, and presents them in a way that is understandable for users, including non-expert ones. In this way, users become aware not only of the existence of biases in their datasets, but also how they may impact their analytics and AI algorithms, thus avoiding unwanted results.

The rest of the paper is structured as follows. Section 2 presents a classification of types of biases. Section 3 summarizes the related work in this area. Section 4 describes our proposed process. Section 5 presents our Biases Detection Approach. Section 6 describes results of the experiments applying our approach. Finally, Section 7 summarizes the conclusions and our future work.

2 BIASES IN DATA

In order to illustrate the negative impact of data bias, in this section, we provide a classification of types of biases. There are different types of biases in datasets, the most common being Class Imbalance and Dataset Shift.

Class Imbalance is the case where classes are not equally represented in the data, this means that one or more categories on the dataset have a higher representation than the rest of the categories. It is usual to find this kind of bias in real word datasets [12]. This bias causes several problems, specially when people are trying to analyze this data and/or applying AI algorithms.

Dataset Shift refers to the case where the distribution of the data within the training dataset does not match the distribution in the test and real datasets. In real word datasets often train and test datasets have not been generated by the same distribution.

Artificial Intelligence Algorithms trained on biased training sets tend not to generalize well on test data that is from the true underlying distribution of the population, which has an negative effect on the quality of a machine learning model. As [18] argue, there are three potential types of dataset shift:

Covariate Shift: It happens when the input attributes have different distributions between the training and test datasets.

Prior Probability Shift: In this case, it happens when the class distribution is different between the training and test datasets.

Concept Shift: It happens when the relationship between the input and class variables changes. Usually occurs when training data is collected at a different point in time than testing data.

Biased datasets are very common and they can cause severe problems if bias are not taken into account and treated properly depending on the type of bias we are facing, the context, and the objective that the dataset is being used for. Therefore, it is paramount to show users how biased their data are, in order to enable them to take into account those biases which are determinant to them. Otherwise, their decisions will likely have unexpected and negative consequences.

3 RELATED WORK

The class imbalance problem has been encountered in multiple areas, some of them with a serious impact, such as in the interpretation of medical data [5]. This problem has been also considered one of the top 10 problems in data mining and pattern recognition [24]. The issue with imbalance in class distribution becomes more pronounced with the applications of the AI algorithms. Mining and learning classifiers from imbalanced datasets are indeed a very important problem from both the algorithmic and performance perspective [13]. Not choosing the right distribution can introduce bias towards the most represented class. Since most AI algorithms expect a balanced class distribution [11], an algorithm trained with imbalanced datasets will tend to unadvertedly return results of the most populated classes.

Different authors have proposed several techniques to handle with these problems. Generally, the approaches to deal with **Imbalanced Data** issues involve three categories [16]:

Data perspective: uses techniques to artificially re-balance the class distribution by sampling the data space to diminish the effect caused by class imbalance. As [10] argues, one intuitive method is undersampling the majority classes by dropping training examples. This approach leads to smaller data sets, but important examples could be dropped during the process. Another method is oversampling the minority classes.

Algorithmic perspective: these solutions try to adapt or modify cost adjustment within the learning algorithm to make it perform better on imbalanced data sets during the training process. For example, [17] proposes an algorithm that is able to deal with the uncertainty that is introduced in large volumes of data without disregarding the learning in the underrepresented class.

Ensemble approach: this type of solutions uses aspects from both perspectives to determine the final prediction. [9] proposes an integrated method for learning large imbalanced dataset. Their approach examines a combination of metrics across different learning algorithms and balancing techniques. The most accurate method is then selected to be applied on real large, imbalanced, and heterogeneous datasets.

In the case of **Dataset Shift** (when the training data and test data are distributed different), a common approach is to reweight data such that the reweighted distribution matches the target distribution [20]. In [22] authors analyze, the relationship between the class distribution of training data to determine the best class distribution for learning. [10] have recently proposed decision tree learning for finding a model that is able to distinguish between training and test distributions.

On the other hand, some works have focused on the impact of data flaws on the visual features of visualization. M. Correll et al. in [8] show how it is possible to create visualizations that seem "plausible" (design parameters are within normal bounds and pass the visual sanity check) but hide crucial data flaws. The biases can be considered as data flaws if the context determines so. It is possible to detect biases in datasets when the classification categories are not approximately equally represented.

As we have shown, most approaches developed until now are mainly focused on machine-learning and rebalancing the biased datasets. However, our goal is not to balance the biased datasets. As [7] argues, the fairness of predictions should be evaluated in context of the data, and unfairness induced by inadequate samples sizes or unmeasured predictive variables should be addressed through data collection rather than by constraining the model. For this reason, we propose an approach that automatically warns the user of the existence of biases and lets her analyze the data from different perspectives without altering the dataset. Since one of the core benefits of visualizations is enabling people to discover visual patterns that might otherwise be hidden [8].

4 PROPOSED PROCESS

In this section, we will describe our proposed process. Fig. 1 summarizes the process followed in our proposal, representing in a red cloud the new elements introduced in this paper. Rest of the elements were introduced in our previous work [15] [14].

In our process, firstly, a sequence of questions guides users in creating a **User Requirements Model** [15] that captures their needs and analysis context. Then, this Model is complemented by the **Data Profiling Model** [15] that analyzes of the features of the data sources selected to be visualized. The user requirements, together with the data profiling information, are translated into a **Visualization Specification** that enables users to derive the best visualization types [15] in each context automatically. This transformation generates a **Data Visualization Model** [14].

The **Data Visualization Model** enables users to specify visualization details regardless of their implementation technology. This model also enables users to determine if the proposed visualization is adequate to satisfy the essential requirements for which it was created or not. If the proposed visualization does not pass the user validation, it will point out the existence of missing or wrongly defined requirements. In this case, a new cycle is started by reviewing the existing model to identify which aspects were not taken into account, generating in turn an updated model. Otherwise, a successful validation will start the **Biases Detection Process**. Once users have validated the visualization, the attributes of the collections that have been selected in the process to be represented in the visualization are analyzed. Our novel algorithm examines the data to automatically detect biases and presents this information to the users. Users may define thresholds to adapt the **Biases Detection Process** to their specific needs. The definition of thresholds is performed in an easy way, adapted for non-expert users by defining two variables through the interface. This new functionality will make users aware of biases that could significantly alter the interpretation of their data, as well as the techniques to be used for the analysis.

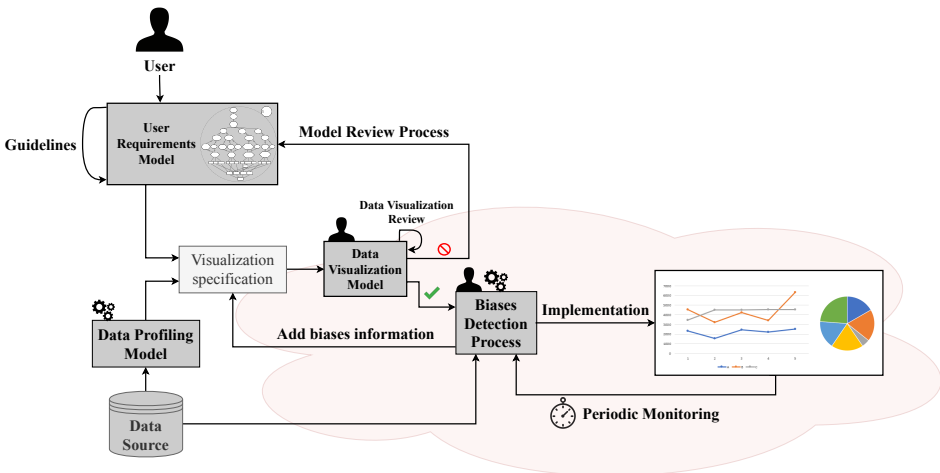


Figure 1: Overall view of the proposed process

As a result of the process, users will obtain a **visual representation** of the bias, being offered the option to include information in their analytics about each of the attributes detected as biased by the algorithm. If they decide to add information about a biased attribute, they can integrate this information within the visualization that they had created for the initial analysis, or, alternatively, in a new visualization that is dynamically connected with the visualization of the process, so that when one of the visualizations is interacted with, the other one is updated.

If users decide to add new information about some biased attribute, a new visualization specification will be generated. Therefore, in the **Data Visualization Model**, users will be able to customize the visualization/visualizations and select how to represent the biases information. Once users validate the new visualizations and do not wish to add further information, the corresponding implementation will be generated.

Finally, when the visualization has been implemented and users are working with it, it is possible to program a Periodic Monitoring. The aim of this continuous monitoring is to ensure that, as new data populates the data sources, no new biases are introduced unadvertedly. The **Periodic Monitoring** event will trigger an execution of our Biases Detection Algorithm with the aim of automatically detect if the data has exceeded the defined thresholds. If a new threshold has been exceed, an alert will be shown to users. This will enable them to return to the Biases Detection Process and choose if they want to edit or add information about this new bias in the visualizations.

By following this process, we facilitate the data analysis and bias awareness for non-expert users in data visualization. Furthermore, all users may benefit from the reduction in time involved in using this approach, since skipping the existing biases will lead to problems, requiring users to manually identify the biases that originated them and requiring to rebuild all the visualizations or re-train their AI algorithms. Therefore, the process enables users to retain control of how data biases affect their data and makes them aware of the impact on their analytics and AI algorithms.

5 BIASES DETECTION

Our proposal starts from the result of our process for automatic derivation of visualizations, shown in Fig. 1. In this sense, we assume the user has defined her requirements, the information that she wants to analyze and that the visualization that best suits her needs has been automatically derived. Once the user has validated the visualization, it is possible that certain elements are changing the interpretation of the data and the user is unaware of them. Therefore, at this point we introduce our novel Biases Detection Process to detect biases in the data, based on the algorithm proposed in this paper that will facilitate this task. It is important to note that, although we assume that the user has followed our previous approach, the process proposed can be applied to visualizations obtained through other tools, as long as the necessary information is facilitated as input to the algorithm.

The first step in our Biases Detection Process proposed is to automatically analyze the attributes of the collections used for the visualization defined in the process through **Algorithm 1**. This algorithm enables us to automatically detect biases in the data by an analysis of the datasets, giving us information as to how biased data are. Users can alter the limits for bias detection in order to tailor the algorithm to their particular case.

It is important to note that, although we exemplify the implementation of our algorithm assuming an existing relational database, our proposal can be applied to any context where structured or semi-structured data is being analyzed.

Algorithm 1 starts with the input of the data tables (**tables_vis**) that are used for the visualization. These tables will come automatically to the algorithm from the previous step of our process. On the other side, the variables **thdCategorical** and **thdBiases** define the thresholds to delimit the biases and attributes, these thresholds do not need to be defined, as they are already assigned default values according to our experience analyzing datasets.

To define the thresholds, we have analyzed different studies. Academic research [11] suggest that there is a situation of class imbalance when the majority-to-minority class ratio is within the

Algorithm 1: Biases Detection Algorithm

```
/* tables_vis comes automatically from the process
   thdCategorical and thdBiases are defined by default, but
   users may personalize it                                  */
Input  : tables_vis[] = list of tables used in the
         visualization, thdCategorical = 0,05; number
         that represents the maximum percentage of the
         total elements of the table to be considered a
         categorical attribute, thdBiases = 8; number
         between 0 and 10 that establishes the admissible
         bias ratio of the attributes (being 0 equally
         distributed and 10 very biased)
Output : biasedAtt = list of attributes and their bias
1  foreach table in tables_vis do
2      Statement stmt = con.createStatement();
       /* Query 1                                           */
3      String rowsQuery = "SELECT COUNT(*) FROM " +
         table;
       /* Query 2                                           */
4      String attributesQuery = "SELECT COLUMN_NAME
         FROM INFORMATION_SCHEMA.COLUMNS
         WHERE TABLE_NAME = " + table;
       /* number of rows from the table                     */
5      ResultSet rsRN = stmt.executeQuery(rowsQuery);
6      int RN = rsRN.getInt(1);
       /* list of attributes from the table                 */
7      ResultSet attributes =
         stmt.executeQuery(attributesQuery);
       /* for each attribute from the table                 */
8      foreach attribute in attributes do
           /* Query 3                                       */
9          String groupAttrQuery = "SELECT COUNT(" +
             attribute + " ) FROM " + table + " GROUP BY " +
             attribute ;
           /* number of times that each different value of the
              attribute appears                             */
10         ResultSet rsGroupAttr =
             stmt.executeQuery(groupAttrQuery);
           /* number of distinct values of the attribute    */
11         rsGroupAttr.last();
12         int RND = rsGroupAttr.getRow();
           /* if it is a categorical attribute              */
13         if RND < RN*thdCategorical then
               /* is extracted the value that is repeated more
                  and less times                            */
14             int max = max(rsGroupAttr);
15             int min = min(rsGroupAttr);
               /* is calculated and normalized the bias in the
                  attribute                                 */
16             float biasAttribute = ((max - min)/max)*10;
               /* if the bias is bigger than the threshold
                  defined by the user                       */
17             if biasAttribute > thdBiases then
18                 biasedAtt.append(attribute, biasAttribute);
19             end
20         end
21     end
22     return (biasedAtt);
23 end
```

range of 100:1 to 10000:1. However, from the viewpoint of effective problem solving, lower class imbalances that make modeling and prediction of the minority class a complex and challenging task (i.e. in the range of 50:1 and lower) are considered high class imbalance by domain experts [21].

In our case, the variable **thdCategorical** is a number that represents the maximum percentage of the total elements of the table to be considered a categorical attribute. An attribute is categorical when it can only take a limited number of possible values. The default threshold for this variable has been defined heuristically, setting the value of this variable to 5% (0,05). This threshold enables us to discover categorical attributes within the data, even when a schema is not available, such as with NoSQL databases or file-based systems.

Moreover, the variable **thdBiases** is a number between 0 and 10 that establishes the admissible bias ratio of the attributes (being 0 equally distributed and 10 very biased). The bias ratio represents the relationship between the values that appear the least and most in an attribute. Therefore, adjusting this variable, users may limit when an attribute is considered biased, i.e. when the difference between the most and least common value is decisive for them. We propose 8 as default value. Therefore, if the most common value has 8 times or more the representation of the least common value, then it will be considered as highly biased.

Finally, the output of this algorithm will be **biasedAtt**, a list with the information about each attribute and its bias ratio.

The algorithm will be executed for each table used for the visualization *(line 1)*. For each table, it stores the number of rows from the table in the variable **RN** *(lines 5-6)*. Then, the attributes of the table are included in the variable **attributes** *(line 7)*. For each attribute in the list *(line 8)*, a ResultSet **rsGroupAttr** with the number of repetitions of each different value is stored *(line 10)*. In *(lines 11-12)*, the number of distinct values of this attribute is calculated and stored in **RND**. Afterwards, the algorithm evaluates whether this attribute is categorical or not *(line 13)*. An attribute is considered categorical when the number of distinct values of this attribute (**RND**) is be lower than the number of rows from the table (**RN**) multiplied by the categorical threshold defined earlier (5%) **thdCategorical**. If this comparison is fulfilled, the values that have the highest (**max**) *(line 14)* and lowest (**min**) *(line 15)* representation are extracted from the ResultSet **rsGroupAttr** that contains the number of times that each different value of the attribute appears. Then, the bias of each attribute is calculated and normalized in **biasAttribute** *(line 16)* using the following formula:

$$\frac{max - min}{max} * 10 \tag{1}$$

We have used Min-Max normalization because it guarantees that all attributes will have the exact same scale and highlights outliers. This is a desirable characteristic in our case, since detecting the existence of these outlier biases and warning the user is one of our main goals. With this normalization, we will have a ratio for each attribute in **biasAttribute** that will provide an indication in the 0 to 10 range how biased is the attribute, 0 being equally distributed and 10 very biased.

If the **biasAttribute** is bigger than the threshold **thdBiases** *(line 17)*, it means that the attribute has a considerable bias that should be analyzed. Then, the name of the attribute and the bias ratio of the attribute previously calculed in **biasAttribute** will be stored in **biasedAtt** *(line 18)*. Therefore, when the algorithm

concludes, the variable **biasedAtt** will contains a list of attributes with their bias ratio.

6 PERFORMANCE ANALYSIS

In order to do an implementation of the experiment, we have downloaded the Fire Department Calls for Service dataset from [1] where we have get an 1,75 GB file.

We have chosen Apache Spark [3] to process this file because its speed, ease of use, advanced and in-memory analytical capabilities. Specifically we have used as a development environment Apache Zeppelin [4] 0.8. The configuration is as default.

We have run the experiment on a single laptop with the following characteristics: Intel Core i5 CPU M 460 @ 2.53GHz × 4, HDD at 7200 rpm, 6GB of RAM and OS: Ubuntu 16.04 LTS.

Although in the definition of **Algorithm 1** we establish connections with the database, since we are running the algorithm on Spark this is not necessary, loading the dataset into the framework using a load instruction instead. We have loaded the *Fire_Department_Calls_for_Service.csv* into the variable **dfCalls** and we run the following queries as part of the algorithm:

(1) **Number of rows from the table:** dfCalls.count()
(2) **List of attributes from the table:** dfCalls.columns
(3) **Number of distinct values from each attribute:**
dfCallsG = dfCalls.groupBy(attribute).count()
dfCallsG.count()

The execution times of our approach over a 5.1 millions of rows and including all the passes to process the 34 columns of the dataset (1,75 GB) are: 46 seconds to be load the data table. Query 1 takes 27 seconds. Query 2 is executed in under 1 second and finally, Query 3 takes 993 seconds. Therefore, the time required to run Algorithm 1 in this experiment is a total of 1066 seconds, i.e. 17 minutes and 46 seconds.

7 CONCLUSIONS AND FUTURE WORK

Data bias is becoming a prominent problem due to its impact in data analytics and AI. Current solutions focus on the problem from an AI outputs perspective, centering their efforts in constraining the model to re-balance the data at hand. The side effect is that the datasets are altered without understanding whether there is a problem at the data gathering step or the data is representing the actual distribution of the sample. In turn, potentially important information about the nature of the data is lost, which can have implications for interpreting the data and finding the root causes of the original imbalance.

Compared to these solutions, in this paper we have presented a Bias Detection Approach. Our proposal complements our previous works [14, 15] by including a novel algorithm that takes into account the scope of the analysis, detects biases, and presents them in a way that is understandable for users, including non-expert ones. The great advantage of our proposal is that we enable users to understand their data and make decisions considering biases from different perspectives without altering the dataset. Furthermore, all users may benefit from the reduction in time required to inspect and understand existing biases within their datasets, while at the same time they avoid biases going unadverted with the problems that it entails.

As a part of our future work, we are continuing our work on new techniques to present biased attributes with a high number of categories. We are also applying our approach to unstructured data and including analytic requirements as an input to estimate the impact of data biases for each particular user.

ACKNOWLEDGMENTS

This work has been co-funded by the ECLIPSE-UA (RTI2018-094283-B-C32) project funded by Spanish Ministry of Science, Innovation, and Universities. Ana Lavalle holds an Industrial PhD Grant (I-PI 03-18) co-funded by the University of Alicante and the Lucentia Lab Spin-off Company.

REFERENCES

[1] 2019. Fire Department Calls for Service dataset. https://data.sfgov.org/Public-Safety/Fire-Department-Calls-for-Service/nuek-vuh3. Accessed: 23/10/2019.
[2] Alaa Althubaiti. 2016. Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of multidisciplinary healthcare* 9 (2016), 211.
[3] Apache. 2019. Apache Spark. https://spark.apache.org/. Accessed: 23/10/2019.
[4] Apache. 2019. Apache Zeppelin. https://zeppelin.apache.org/. Accessed: 23/10/2019.
[5] Colin B Begg and Jesse A Berlin. 1988. Publication bias: a problem in interpreting medical data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 151, 3 (1988), 419–445.
[6] Kenneth Brant, Moutusi Sau, Anthony Mullen, Magnus Revang, Chirag Dekate, Daryl Plummer, and Whit Andrews. 2017. Predicts 2018: Artificial Intelligence. https://www.gartner.com/en/documents/3827163/predicts-2018-artificial-intelligence. Accessed: 23/10/2019.
[7] Irene Y. Chen, Fredrik D. Johansson, and David Sontag. 2018. Why Is My Classifier Discriminatory?. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*. 3543–3554.
[8] Michael Correll, Mingwei Li, Gordon Kindlmann, and Carlos Scheidegger. 2018. Looks Good To Me: Visualizations As Sanity Checks. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 830–839.
[9] Mojgan Ghanavati, Raymond K Wong, Fang Chen, Yang Wang, and Chang-Shing Perng. 2014. An effective integrated method for learning big imbalanced data. In *2014 IEEE International Congress on Big Data*. IEEE, 691–698.
[10] Patrick O. Glauner, Petko Valtchev, and Radu State. 2018. Impact of Biases in Big Data. *CoRR* (2018).
[11] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21, 9 (2009), 1263–1284.
[12] Richard A. Bauder Joffrey L. Leevy, Taghi M. Khoshgoftaar and Naeem Seliya. 2018. A survey on addressing high-class imbalance in big data. *J. Big Data* 5 (2018), 42.
[13] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, et al. 2006. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering* 30, 1 (2006), 25–36.
[14] Ana Lavalle, Alejandro Maté, and Juan Trujillo. 2019. Requirements-Driven Visualizations for Big Data Analytics: a Model-Driven approach. In *International Conference on Conceptual Modeling ER 2019, to appear*. Springer.
[15] Ana Lavalle, Alejandro Maté, Juan Trujillo, and Stefano Rizzi. 2019. Visualization Requirements for Business Intelligence Analytics: A Goal-Based, Iterative Framework. In *27th IEEE International Requirements Engineering Conference RE 2019, to appear*.
[16] Chaoliang Li and Shigang Liu. 2018. A comparative study of the class imbalance problem in Twitter spam detection. *Concurrency and Computation: Practice and Experience* 30, 5 (2018).
[17] Victoria López, Sara Del Río, José Manuel Benítez, and Francisco Herrera. 2015. Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data. *Fuzzy Sets and Systems* 258 (2015), 5–38.
[18] Victoria López, Alberto Fernández, and Francisco Herrera. 2014. On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed. *Information Sciences* 257 (2014), 1–13.
[19] Jerome Pesenti. 2018. TAI at F8 2018: Open frameworks and responsible development. https://engineering.fb.com/ml-applications/ai-at-f8-2018-open-frameworks-and-responsible-development/. Accessed: 23/10/2019.
[20] Sashank Jakkam Reddi, Barnabás Póczos, and Alexander J. Smola. 2015. Doubly Robust Covariate Shift Correction. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
[21] Isaac Triguero, Sara del Río, Victoria López, Jaume Bacardit, José Manuel Benítez, and Francisco Herrera. 2015. ROSEFW-RF: The winner algorithm for the ECBDL'14 big data competition: An extremely imbalanced big data bioinformatics problem. *Knowledge-Based Systems* 87 (2015), 69–79.
[22] Gary M Weiss and Foster Provost. 2003. Learning when training data are costly: The effect of class distribution on tree induction. *Journal of artificial intelligence research* 19 (2003), 315–354.
[23] Matt Wood. 2018. Thoughts On Machine Learning Accuracy. https://aws.amazon.com/es/blogs/aws/thoughts-on-machine-learning-accuracy/. Accessed: 23/10/2019.
[24] Qiang Yang and Xindong Wu. 2006. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making* 5, 04 (2006), 597–604.

Parte III

Transferencia de tecnología

Desarrollos tecnológicos

La realización de la tesis doctoral en colaboración con la empresa Lucentia Lab ha brindado numerosas oportunidades al permitir poner en práctica, a través de proyectos industriales y de innovación tecnológica, el marco teórico desarrollado. Por un lado, testeando e implantando algunas de las soluciones desarrolladas en proyectos con clientes finales, por otro, posibilitando el desarrollo tecnológico de herramientas basadas en los artículos científicos desarrollados en el núcleo teórico la tesis doctoral.

En esta sección se presenta la herramienta Grafik, cuya propiedad intelectual ha sido registrada por el Servicio de Transferencia de Resultados de Investigación de la Universidad de Alicante y en el momento de la redacción del presente documento se está elaborando el contrato de explotación de licencia por parte de Lucentia Lab.

Capítulo 10

Grafik – Herramienta para la generación automática de visualizaciones y *dashboards* en *Data Analytics*

10.1. Descripción de la herramienta

Grafik consiste en una aplicación web que permite visualizar de forma automática y sencilla conjuntos de datos. El *software* interroga al usuario acerca de los objetivos de su visualización (si quiere obtener una visión global de sus datos, si está comparando, etc.) y analiza el conjunto de datos elegido para sugerir las visualizaciones más adecuadas a la hora de analizarlos. Comparado con otros programas como, por ejemplo Excel, el *software* se basa en un proyecto de investigación [28, 26] que tiene como uno de los puntos la definición de criterios para evaluar las bondades de cada visualización. Además, tiene en cuenta los objetivos del usuario a la hora de evaluar las visualizaciones posibles, lo cual no existe en otros programas que se encuentran en el mercado actualmente.

El objetivo principal de esta herramienta *CASE* es que cualquier usuario sin conocimientos de *Bussines Intelligence* o *Big Data* pueda crear y modificar cuadros de mando *(dashboards)* profesionales que aporten valor a sus organizaciones, pudiéndose así beneficiar de las oportunidades que ofrece el *Bussines Intelligence* antes de realizar grandes inversiones. Ofrece a los usuarios un entorno fácil de usar, permitiendo delegar en el *software* determinadas decisiones que de forma habitual realizan los analistas de datos. Permite descubrir las bondades del análisis de datos antes de externalizar el servicio, adquirir un *software* nuevo o contratar más personal.

10.2. Funcionalidades

Entre las funcionalidades que ofrece la herramienta podemos encontrar, como muestra la figura 10.1, la posibilidad de conectar el *software* a distintos tipos de fuentes de datos.

Al seleccionar la fuente de datos donde se quiere realizar el análisis, el sistema analiza dicha fuente y extrae información sobre las variables que contiene, tal y como muestra la figura 10.2. El usuario solo tendría que arrastrar las variables que seleccione como variables dependientes y no dependientes hacia su correspondiente cuadro. Además, seleccionaría mediante una serie de preguntas realizadas por el sistema el/los objetivo/s

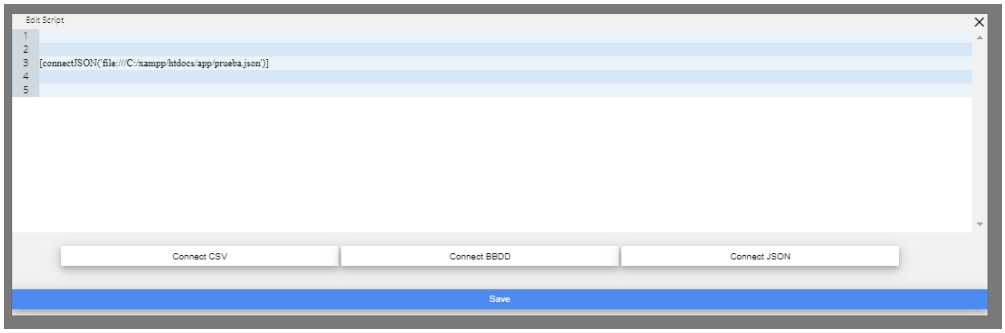


Figura 10.1: Configuración de la conexión a una fuente de datos

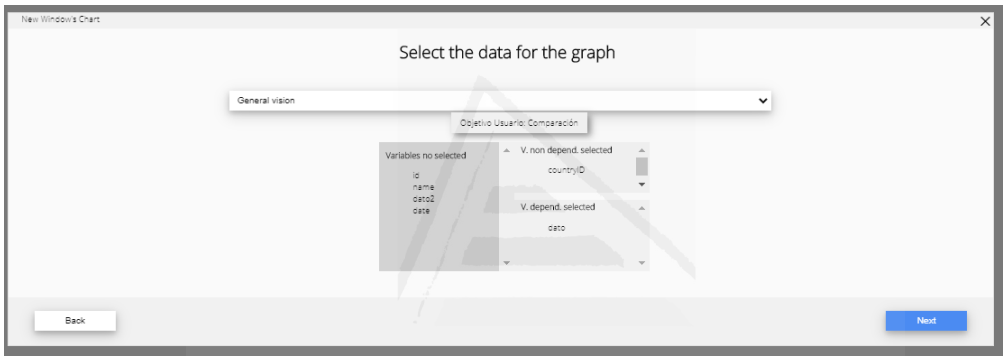


Figura 10.2: Selección características de la nueva visualización

de la visualización y el tipo de interacción que desea tener con ella.

De esta manera, la información que necesita el sistema para derivar automáticamente el mejor tipo de visualización se encuentra ya definida con muy poco esfuerzo para el usuario.

Como muestra la figura 10.3, el siguiente paso es darle un estilo a la visualización. El usuario tiene la posibilidad de elegir el formato del texto, el fondo y la gama de colores que va a utilizar la visualización.

Una vez el estilo ha sido seleccionado y el sistema ya conoce la fuente de datos y las necesidades del usuario, le propone los tipos de visualizaciones que mejor se adaptan a sus necesidades, tal y como muestra la figura 10.4. El sistema recomienda los tipos de visualización que presentan mayor idoneidad con los requisitos especificados, sin embargo también ofrece al

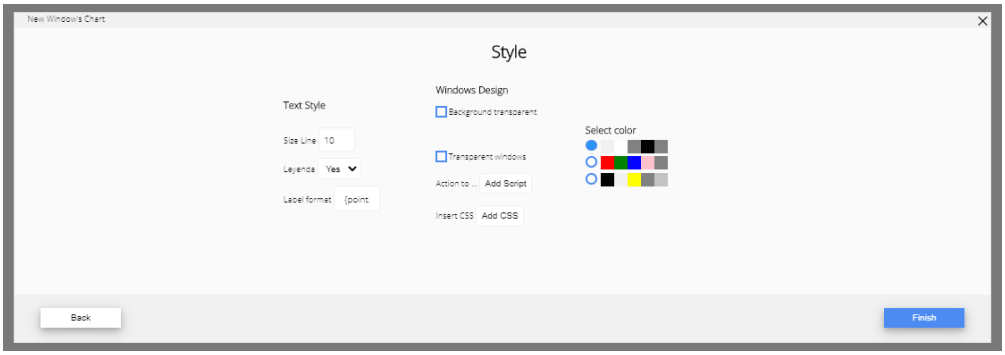


Figura 10.3: Selección de estilo de la nueva visualización

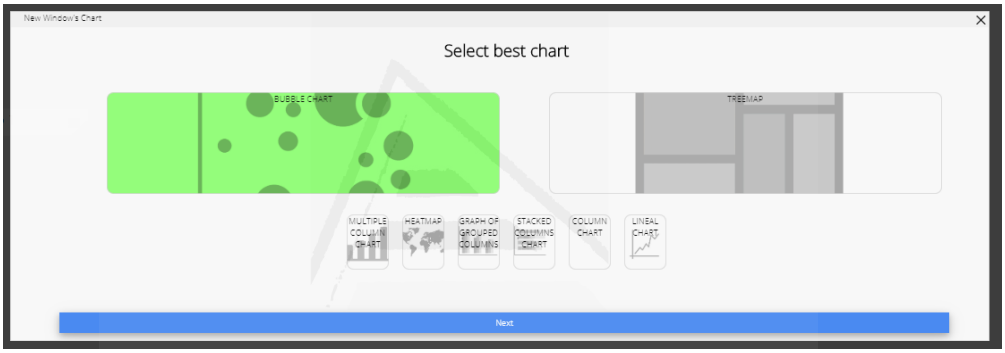


Figura 10.4: Selección del mejor tipo de visualización

usuario la posibilidad de seleccionar otro tipo de gráfico aunque este no sea el mas idóneo.

Finalmente, cuando el usuario selecciona el tipo de visualización que desea utilizar, la visualización se crea de forma totalmente automática y se incorpora al cuadro de mandos como muestra la figura 10.5, el cual recogerá todas las visualizaciones que componen el análisis.

De esta forma, el usuario seguirá estos sencillos pasos para generar el resto de visualizaciones que componen su cuadro de mandos, como muestra la figura 10.6. Así pues, el usuario no experto en visualización de datos será capaz de completar un *dashboard* analítico que le permita obtener información de sus fuentes de datos de una forma guiada y muy sencilla.

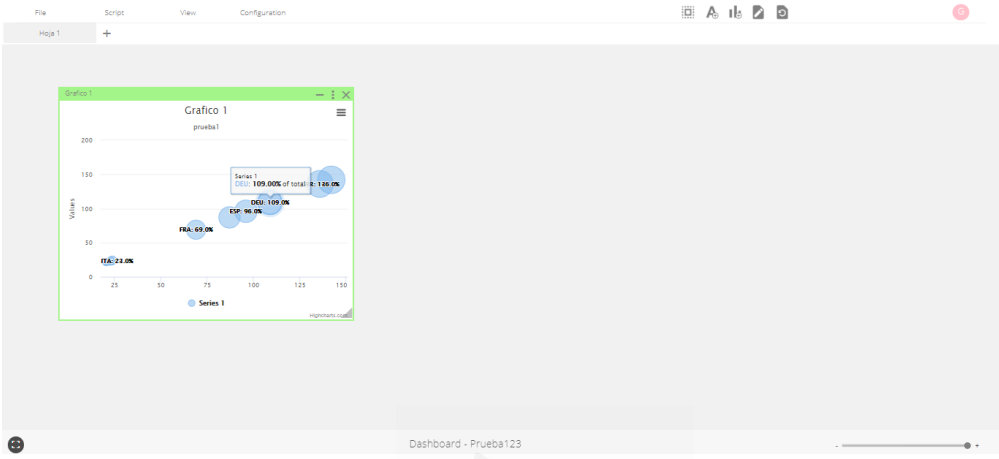


Figura 10.5: Visualización generada automáticamente

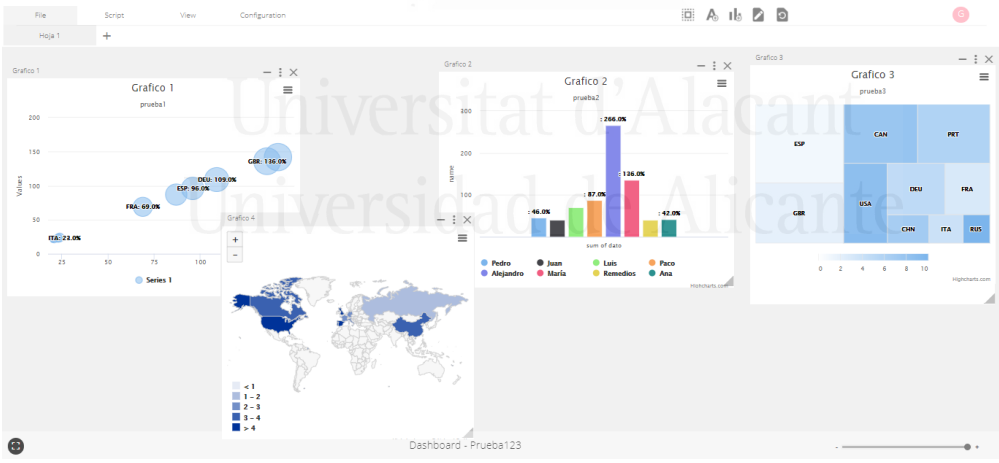


Figura 10.6: Cuadro de mandos con visualizaciones

Parte IV

Conclusiones

Capítulo 11

Conclusiones

En los últimos años, el volumen de datos a analizar e interpretar está aumentando exponencialmente, y el procesamiento y análisis de *Big Data* se trata de uno de los principales temas de interés, tanto para investigadores como para desarrolladores en entornos industriales.

Dentro de este amplio tema de investigación, la visualización de *Big Data* ha generado un interés creciente, ya que el uso de unas visualizaciones adecuadas es determinante para extraer información precisa de los datos y así guiar a los usuarios de estas a tomar decisiones mejor informadas. Es crucial poder ofrecer de una forma automática la visualización más adecuada en función de las necesidades particulares de los usuarios finales y de la naturaleza y características de las fuentes de *Big Data*.

No obstante, encontrar las visualizaciones que mejor se adapten a un contexto determinado es una tarea desafiante, especialmente cuando se trabaja con *Big Data*. Y el problema se agrava cuando los usuarios no son expertos en visualización de datos y no tienen una idea clara de los objetivos para los que están construyendo las visualizaciones.

Por esta razón, en la presente tesis doctoral se han analizado las necesidades actuales en la toma de requisitos para la generación de visualizaciones y se ha presentado una metodología para la automatización de la toma de requisitos y la derivación automática de las visualizaciones más adecuadas.

A continuación, detallamos los cuatro objetivos específicos planteados en la tesis doctoral y los resultados que se han obtenido para alcanzar cada uno de ellos.

O1 - Definir una metodología sistemática que permita derivar visualizaciones orientadas a usuarios no expertos.

- Metodología completa, desde la definición de requisitos hasta la implementación de visualizaciones.

O2 - Definir un marco de requisitos que permita especificar de manera clara los objetivos analíticos que se persiguen.

- Formalización del modelo de requisitos de usuario y metamodelo para la especificación de objetivos analíticos y la toma de requisitos sobre la generación de visualizaciones.
- Formalización del modelo de visualización de datos y metamodelo para la especificación de características detalladas de las visualizaciones.
- Adaptación de los modelos para escenarios en tiempo real.

O3 - Automatizar la obtención de las visualizaciones a partir del marco de requisitos definido.

- Transformación automática de los requisitos tomados en los modelos a conjuntos de visualizaciones reales agrupadas en cuadros de mando.
- Formalización del modelo de perfilado de datos para la extracción automática de características de las fuentes de datos utilizadas en las visualizaciones.
- Desarrollo de una herramienta *CASE* que implementa el mejor tipo de visualización para unas características especificadas por el usuario.

O4 - Aplicar y evaluar el impacto de las técnicas desarrollados en distintas áreas.

- Aplicación del enfoque en distintos ámbitos, tales como ciudades inteligentes y procesos de producción industrial.
- Evaluación del enfoque mediante experimentos realizados sobre 97 participantes que confirman la validez de nuestra propuesta demostrando que: (i) permite a los usuarios cubrir

más cuestiones analíticas; (ii) mejora el conjunto de visualizaciones generadas; (iii) produce una mayor satisfacción general en los usuarios.

En conclusión, considerando que se han cumplido todos los objetivos específicos, es posible afirmar que el objetivo principal de la presente tesis doctoral, especificado como: "definir una metodología que agrupe una serie de técnicas y aproximaciones para mejorar la comprensión visual de *Big Data*" ha sido alcanzado. Por consiguiente, también se puede afirmar que la hipótesis inicial de la presente tesis doctoral, definida como: "es factible mejorar y sistematizar las aproximaciones actuales para la visualización de *Big Data*" ha sido confirmada.

Capítulo 12

Trabajos futuros

A pesar de los resultados obtenidos en la presente tesis doctoral todavía es posible profundizar en distintos campos de esta línea de investigación, así como descubrir otras nuevas.

En primer lugar, se podría profundizar la propuesta recomendando no solo los tipos de visualización más adecuados, sino también proponiendo unos valores en los ejes específicos para cada tipo de objetivo analítico, de forma que una misma visualización pueda tener distintos valores en los ejes en función del objetivo por el cual ha sido creada o qué tipo de usuario la vaya a consultar.

Por otra parte, aunque nuestra propuesta es independiente del contexto de aplicación, se podría profundizar teniendo en cuenta otros sectores con diversas necesidades, para comprobar si se cubren todas las necesidades analíticas o falta algún elemento por incorporar. En este sentido, todavía existe margen para ampliar la investigación en modelado de requisitos de visualización, incluyendo aspectos no considerados en el marco de la presente tesis, tales como las visualizaciones colaborativas o profundizando en otros como las visualizaciones compuestas.

En referencia a la herramienta para la generación automática de visualizaciones y *dashboards* Grafik, se continúa depurando y se están incorporando los pasos previos en los que se defina el modelo de requisitos de usuario, para que estos puedan no solo definir las visualizaciones necesarias para alcanzar sus objetivos, sino medir el grado de su alcance.

Una vez la herramienta se encuentre completamente finalizada y testeada, se realizará una evaluación con nuevos participantes donde se medirá

el grado de mejora aportado.

Finalmente, cabe destacar que el campo de la visualización de datos es tan amplio y transferible que podría aplicarse a otros sectores y abrir prácticamente cualquier línea de investigación. Una de esas nuevas líneas de investigación que estamos explorando actualmente, y en la cual nos hemos querido centrar, es la detección de sesgos y cómo estos influyen en los algoritmos de inteligencia artificial.

El aprendizaje automático se está incorporando cada vez más en más sectores, y tanto un aprendizaje sesgado como un aprendizaje con datos rebalanceados manualmente puede traer consecuencias nefastas. Esto ocurre, por ejemplo, con los sesgos de género. Si en un cierto escenario entrenamos el algoritmo con datos históricos, el algoritmo seguirá reproduciendo los clásicos problemas de discriminación. Estudiar cómo incorporar ética en la inteligencia artificial y cómo evitar sesgos discriminatorios en los algoritmos es un tema crucial.

Con este fin, estamos concluyendo la redacción de un artículo científico que complementa el presentado en el capítulo 9, en el que se propuso un proceso para analizar visualmente cuánto de sesgada está cada clase de un conjunto de datos. En esta nueva propuesta analizamos el impacto del sesgo en distintos sectores y cómo afecta ello en la ejecución de diferentes algoritmos de inteligencia artificial. Así, podemos recomendar a usuarios no expertos en analítica y visualización de datos, cuándo es apropiado y cuándo no alterar los conjuntos de datos para suavizar el sesgo en función de los objetivos del análisis. Es de prever que este análisis se convierta en un paso fundamental de cara al futuro conforme el aprendizaje automático pase a ser una pieza básica del funcionamiento de nuestra sociedad.

Bibliografía

[1] Fire department calls for service dataset. https://data.sfgov.org/Public-Safety/Fire-Department-Calls-for-Service/nuek-vuh3, 2019. Accessed: 10/04/2021.

[2] Agrawal, R., Kadadi, A., Dai, X., and Andrès, F. Challenges and opportunities with big data visualization. In *Proceedings of the 7th International Conference on Management of computational and collective intElligence in Digital EcoSystems* (2015), ACM, pp. 169–173.

[3] Althubaiti, A. Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of multidisciplinary healthcare 9* (2016), 211.

[4] Arterburn, R. Data visualization market growth, future prospects and competitive analysis (2020-2027) fortune business insights, 2020. Accessed: 10/04/2021.

[5] Börner, K. Atlas of knowledge, 2014.

[6] Brant, K., Sau, M., Mullen, A., Revang, M., Dekate, C., Plummer, D., and Andrews, W. Predicts 2018: Artificial intelligence. https://www.gartner.com/en/documents/3827163/predicts-2018-artificial-intelligence, 2017. Accessed: 10/04/2021.

[7] Bresciani, S., and Eppler, M. J. The pitfalls of visual representations: A review and classification of common errors made while designing and interpreting visualizations. *Sage Open 5*, 4 (2015).

[8] Caldarola, E. G., and Rinaldi, A. M. Improving the visualization of wordnet large lexical database through semantic tag clouds. In *International Congress on Big Data* (2016), IEEE, IEEE Computer Society, pp. 34–41.

[9] Caldarola, E. G., and Rinaldi, A. M. Big data visualization tools: A survey - the new paradigms, methodologies and tools for large data sets visualization. In *Proceedings of the 6th International Conference on Data Science, Technology and Applications, DATA* (2017), INSTICC, SciTePress, pp. 296–305.

[10] Chandra, J., and Shudan, M. Iba graph selector algorithm for big data visualization using defence dataset. *International Journal of Scientific & Engineering Research 4*, 3 (2013), 1–7.

[11] Chawla, N. V. Data mining for imbalanced datasets: An overview. In *Data Mining and Knowledge Discovery Handbook, 2nd ed*, O. Maimon and L. Rokach, Eds. Springer, 2010, pp. 875–886.

[12] Chen, I. Y., Johansson, F. D., and Sontag, D. A. Why is my classifier discriminatory? *CoRR abs/1805.12002* (2018).

[13] Dalpiaz, F., Franch, X., and Horkoff, J. istar 2.0 language guide. *CoRR abs/1605.07767* (2016).

[14] Day, M. How linkedin's search engine may reflect a gender bias. `https://www.seattletimes.com/business/microsoft/how-linkedins-search-engine-may-reflect-a-bias/`, 2016. Accessed: 10/04/2021.

[15] Dayhoff, J. E., and DeLeo, J. M. Artificial neural networks: opening the black box. *Cancer: Interdisciplinary International Journal of the American Cancer Society 91*, S8 (2001), 1615–1635.

[16] Ghanavati, M., Wong, R. K., Chen, F., Wang, Y., and Perng, C.-S. An effective integrated method for learning big imbalanced data. In *2014 IEEE International Congress on Big Data* (2014), IEEE, pp. 691–698.

[17] GII, GRIN, and SCIE. The gii-grin-scie (ggs) conference rating. `http://scie.lcc.uma.es/gii-grin-scie-rating/`, 2018. Accessed: 10/04/2021.

[18] Glauner, P. O., Valtchev, P., and State, R. Impact of biases in big data. *CoRR abs/1803.00897* (2018).

[19] Golfarelli, M., and Rizzi, S. A model-driven approach to automate data visualization in big data analytics. *Information Visualization 19*, 1 (2020), 24–47.

[20] Gray, C. C., Teahan, W. J., and Perkins, D. Understanding our analytics: A visualization survey. *Journal of Learning Analytics, to appear* (2017).

[21] Johnson, C. R., and Sanderson, A. R. A next step: Visualizing errors and uncertainty. *IEEE Computer Graphics and Applications 23*, 5 (2003), 6–10.

[22] Labrinidis, A., and Jagadish, H. V. Challenges and opportunities with big data. *Proceedings of the VLDB Endowment 5*, 12 (2012), 2032–2033.

[23] Lavalle, A. Poster - una metodología para la comprensión visual de big data. In *Jornada del doctorado en Informática (JDI-2018)* (2018).

[24] Lavalle, A. Presentación - una metodología para la comprensión visual de big data. In *Jornada del doctorado en Informática (JDI-2019)* (2019).

[25] Lavalle, A., Maté, A., and Trujillo, J. Modelado conceptual basado en objetivos para la definición de visualizaciones. In *Jornadas de Ingeniería del Software y Bases de Datos, JISBD 2019, Caceres, Spain, September 2-4, 2019* (2019).

[26] Lavalle, A., Maté, A., and Trujillo, J. Requirements-driven visualizations for big data analytics: A model-driven approach. In *Conceptual Modeling - 38th International Conference, ER 2019, Salvador, Brazil, November 4-7, 2019, Proceedings* (2019), Springer, pp. 78–92.

[27] Lavalle, A., Maté, A., and Trujillo, J. An approach to automatically detect and visualize bias in data analytics. In *Proceedings of the 22nd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data co-located with EDBT/ICDT 2020 Joint Conference, DOLAP@EDBT/ICDT 2020, Copenhagen,*

Denmark, March 30, 2020 (2020), I. Song, K. Hose, and O. Romero, Eds., vol. 2572 of *CEUR Workshop Proceedings*, CEUR-WS.org, pp. 84–88.

[28] Lavalle, A., Maté, A., Trujillo, J., and Rizzi, S. Visualization requirements for business intelligence analytics: A goal-based, iterative framework. In *27th IEEE International Requirements Engineering Conference, RE 2019, Jeju Island, Korea (South), September 23-27, 2019* (2019), pp. 109–119.

[29] Lavalle, A., Maté, A., Trujillo, J., Teruel, M. A., and Rizzi, S. A methodology to automatically translate user requirements into visualizations: Experimental validation. *Information and Software Technology 136* (2021), 106592.

[30] Lavalle, A., Teruel, M. A., Maté, A., and Trujillo, J. Fostering sustainability through visualization techniques for real-time iot data: A case study based on gas turbines for electricity production. *Sensors 20*, 16 (2020), 4556.

[31] Lavalle, A., Teruel, M. A., Maté, A., and Trujillo, J. Improving sustainability of smart cities through visualization techniques for big data from iot devices. *Sustainability 12*, 14 (2020), 5595.

[32] Lewin, K., Tax, S., Stavenhagen, R., and Fals, O. La investigación acción participativa. *La investigación-acción y los problemas de las minorías* (1946).

[33] Li, C., and Liu, S. A comparative study of the class imbalance problem in twitter spam detection. *Concurrency and Computation: Practice and Experience 30*, 5 (2018).

[34] Liedtke, M. Google's new app blunders by calling black people 'gorillas'. `https://www.seattletimes.com/nation-world/google-apologizes-after-app-tagged-black-people-gorillas/`, 2015. Accessed: 10/04/2021.

[35] López, V., Del Río, S., Benítez, J. M., and Herrera, F. Cost-sensitive linguistic fuzzy rule based classification systems under the mapreduce framework for imbalanced big data. *Fuzzy Sets and Systems 258* (2015), 5–38.

[36] Maté, A., Trujillo, J., and Franch, X. Adding semantic modules to improve goal-oriented analysis of data warehouses using i-star. *Journal of systems and software 88* (2014), 102–111.

[37] Mohamed, A., Najafabadi, M. K., Yap, B. W., Kamaru-Zaman, E. A., and Maskat, R. The state of the art and taxonomy of big data analytics: view from new big data framework. *Artificial Intelligence Review 53*, 2 (2020), 989–1037.

[38] (OMG), O. M. G. Model driven architecture guide rev. 2.0, 2014.

[39] Pang, A., Wittenbrink, C. M., and Lodha, S. K. Approaches to uncertainty visualization. *The Visual Computer 13*, 8 (1997), 370–390.

[40] Peña, O., Aguilera, U., and López-de-Ipiña, D. Exploring LOD through metadata extraction and data-driven visualizations. *Program 50*, 3 (2016), 270–287.

[41] Research, T. C., and of Australasia, E. A. Core rankings portal. `https://www.core.edu.au/`, 2020. Accessed: 10/04/2021.

[42] Ribeiro, R. P., Pereira, P. M., and Gama, J. Sequential anomalies: a study in the railway industry. *Mach. Learn. 105*, 1 (2016), 127–153.

[43] Salesforce. State of analytics, 2015. Accessed: 10/04/2021.

[44] Storey, V. C., and Song, I. Big data technologies and management: What conceptual modeling can do. *Data & Knowledge Engineering 108* (2017), 50–67.

[45] Storey, V. C., Trujillo, J., and Liddle, S. W. Research on conceptual modeling: Themes, topics, and introduction to the special issue, 2015.

[46] Teich, P. Artificial intelligence can reinforce bias, cloud giants announce tools for ai fairness. `https://www.forbes.com/sites/paulteich/2018/09/24/artificial-intelligence-can-reinforce-bias-cloud-giants-announce-tools-for-ai-fairness/#27fffe5d9d21`, 2018. Accessed: 10/04/2021.

[47] Vartak, M., Rahman, S., Madden, S., Parameswaran, A. G., and Polyzotis, N. SEEDB: efficient data-driven visualization recommendations to support visual analytics. *Proceedings of the VLDB Endowment 8*, 13 (2015), 2182–2193.

[48] Widrow, B., Rumelhart, D. E., and Lehr, M. A. Neural networks: Applications in industry, business and science. *Commun. ACM 37*, 3 (1994), 93–105.