# VISUALIZANDO LA INFORMACIÓN EN CIENCIA DE DATOS

Mag. María Alejandra Malberti, Mag. Graciela Elida Beguerí, Mag. Raúl Oscar Klenzi, Lic. Manuel Ortega, Prog. Luis Olguín, Lic. Fabrizio Amaya, Joaquín Cortez

Instituto de Informática / Departamento de Informática /Facultad de Ciencias Exactas Físicas y Naturales / Universidad Nacional de San Juan

Av. Ignacio de la Roza 590 (O), Complejo Universitario "Islas Malvinas", Rivadavia, San Juan, Teléfonos: 4260353, 4260355 Fax 0264-4234980, Sitio Web: http://www.exactas.unsj.edu.ar
e-mail: {amalberti, grabeda, rauloscarklenzi, manuel.ortega, lolguinunsj, fabrizio.amaya88, joaquincortez19}@gmail.com

RESUMEN

El presente trabajo plantea los avances del proyecto "Evaluación de visualizaciones eficientes en Ciencia de Datos" que tiene, entre otros, los siguientes objetivos: Análisis de distintos aspectos que atañen a una representación visual, búsqueda de datos abiertos y otras fuentes de datos provenientes de actividades de cooperación, análisis de herramientas de software libre en cuanto a sus potencialidades de visualización y análisis comparativo de los lenguajes Python y JavaScript como soporte de visualizaciones. Para tal fin, se están considerando diversas métricas tales como escala, longitud, área, color, entre otros. Así como visualizaciones de distintos tipos de datos e información con softwares libres y lenguajes de código abierto.

**Palabras clave:** Visualización, Ciencia de Datos, Lenguajes de código abierto, Software Libre.

CONTEXTO

El Proyecto articula líneas de investigación de un grupo de investigadores y docentes de las carreras Licenciatura en Sistemas de Información y Licenciatura en Ciencias de la Computación del Departamento de Informática de la Facultad de Ciencias Exactas Físicas y Naturales de la Universidad Nacional de San Juan (FCEFN-UNSJ), y se encuentra contenido en el Laboratorio de Sistemas Inteligentes para Extracción de Conocimiento en Datos Masivos del Instituto de Informática de la misma facultad. Asimismo, el proyecto se encuentra aprobado y subsidiado por Consejo de Investigaciones Científicas y Técnicas y de Creación Artística CICITCA- UNSJ.

La propuesta cuenta con antecedentes logrados en el tema conforme a sucesivos proyectos aprobados y subsidiados por el ente mencionado en los que el grupo viene trabajando desde el año 2003, siendo los desarrollados desde el año 2014:

- "Extracción de Conocimiento en Datos Masivos" 21/E-951, período 2014-2015
- "La Ciencia de Datos en grandes colecciones de datos" 21/E1014, período 2016-2017
- "Visualización y Deep Learning en Ciencia de Datos" 21/E1071, período 2018-2019

## 1. INTRODUCCIÓN

El constante aumento de datos y la complejidad de los mismos, ha traído como consecuencia problemas y más desafíos a la hora de su visualización.

Entre los problemas se pueden citar espacios de alta dimensión y relaciones complejas y como gran desafío la cantidad de datos que se generan permanentemente.

Es por ello que el proyecto citado "Evaluación de visualizaciones eficientes en

ciencia de datos" tiene como objetivo general proponer criterios para la evaluación de visualizaciones eficientes para Ciencia de Datos.

Behrisch, en el paper "Quality Metrics for Information Visualization" expone el pipeline de visualización –figura 1– con el cual da el caso en que el encargado de diseñar las visualización se enfrenta con el dilema de elegir una entre una infinidad de posibilidades de procesamiento de datos y una cantidad, aún mayor, de opciones de visualizaciones potenciales.

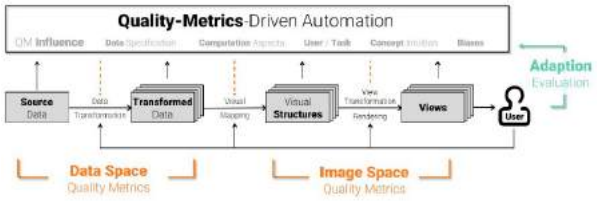


Figura 1. Analítica Visual basada en métricas de calidad con la capa adicional de automatización (Behrisch, et al., 2018)
https://bib.dbvis.de/uploadedFiles/QMSTAR_QualityMetricsForInformationVisualization_FINAL.pdf

En la publicación, ¿Cuántos datos se producen en un minuto?, del Grupo Bit se afirma que "*Un buen uso y análisis de los datos le puede dar a una marca o a una empresa la posibilidad de conocer características o Insights de sus consumidores que antes eran simplemente una hipótesis o eran desconocidos*" (Grupo Bit, 2020)

La Ciencia de Datos es el campo interdisciplinario tendiente a extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sean estructurados o no estructurados.

Dada la gran cantidad de herramientas posibles de aplicar en cada proceso de visualización, el determinar cuáles son los objetivos perseguidos al visualizar es de relevancia. Una caracterización inicial propone dos grandes grupos: Visualizaciones para el análisis vs. Visualizaciones para la comunicación. En el primero de los grupos, lo que se busca es la exploración visual de los datos para que *ellos mismos hablen de su estructura y patrones (NIST/SEMATECH, 2012)*. Herramientas como Rstudio y Phyton poseen numerosas opciones para visualización.

El visualizar para comunicar implica conocer el usuario hacia quien va dirigido el mensaje de la visualización, por lo que se hace necesario *traducir la complejidad* hacia una forma más empática con el observador. Herramientas como Flourish, Data Studio, Tableau, RawGraphics aportan soluciones en este sentido. (Vega, R., 2019)

Jankun-Kelly et al. (2006) propone que para comprender la Visualización, es recomendable analizarla a partir de responder a tres preguntas claves: cómo se crea la visualización, qué sucede durante la visualización, y qué beneficio recibió el usuario o qué lo motivó para trabajar con la visualización. Estas preguntas sugieren tres tipos de modelos básicos:

- Modelo de transformación para describir el método de la visualización
- Modelo de exploración para describir el uso de la visualización, y
- Modelo de diseño para predecir o medir el éxito de un método de visualización. (Ponjuán, D., 2010) (Kelly, et al., 2006)

## 2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Las tareas que se están llevando a cabo comprenden varias líneas de investigación relacionadas con:

- Ciencia de Datos, principalmente lo relativo a Visualización de Información.
- Deep Learning a través del análisis y diferencias entre Boosted Decision Tree Regression y Redes Neuronales del tipo LSTM (Long Short Term Memory) en entorno de visualización open source Knowage.
- Herramientas de software libre para arquitecturas secuenciales, paralelas y

distribuidas particularmente KNIME Analytics cuyo enorme potencial didáctico permite visualizar mediante workflows las diferentes etapas que constituyen el pipeline del Data Science.
- Lenguajes de programación de código abierto tales como Python, JavaScript y R.

## 3. RESULTADOS OBTENIDOS Y ESPERADOS

Se está realizando, sobre diversos conjuntos de datos abiertos, una comparativa entre las bibliotecas de visualización disponibles en el lenguaje Python. Hasta ahora las bibliotecas exploradas son: Mathplotlib, Seaborn, Plotly, Bokeh, Altair, y Folium. Se encuentra en desarrollo un documento con los siguientes criterios de evaluación de las bibliotecas citadas:

- Código e interfaz: qué características o particularidades tiene el código necesario para producir las visualizaciones en estas bibliotecas. Cuál es el nivel de su interfaz.
- Entorno de ejecución: sobre qué plataformas se puede ejecutar, otros programas que sean necesarios para el funcionamiento de la biblioteca, donde se visualizan los gráficos.
- Tipo de visualizaciones: qué tipo de visualizaciones va a estar en condiciones de generar la biblioteca en cuestión.
- Interactiva: Si la biblioteca genera gráficos con los que se pueda interactuar, o, por el contrario, son estáticos.
- Aportes y documentación: en este criterio se analiza qué tan completa es la documentación y cuántos aportes existen para esta biblioteca.
- Dificultad de uso: qué tan sencillo o complejo se puede llegar a tornar trabajar con estas bibliotecas.

También, se han creado varios Jupyter notebooks, con diversos conjuntos de datos libres, con los que se ejemplifican el uso de las bibliotecas.
En el caso particular del espacio vectorial de las palabras, área del procesamiento del lenguaje natural, se aplicaron técnicas de aprendizaje automático basadas en redes neuronales y se empleó el método Word2vec mediante la librería Gensim de Python, para su visualización.

Para el caso de datos temporales, se han revisado varias publicaciones y vale la pena mencionar, en particular, uno de los artículos científicos: “Integration of temporal data visualization techniques in a KDD-based DSS Application in the medical field” (Mohamed, et al. 2014) dado que en éste, se desarrollan temas tales como análisis de datos temporales y la visualización de estos; además de proporcionar una integración de técnicas de visualización (para representaciones interactivas). Ver figura 2

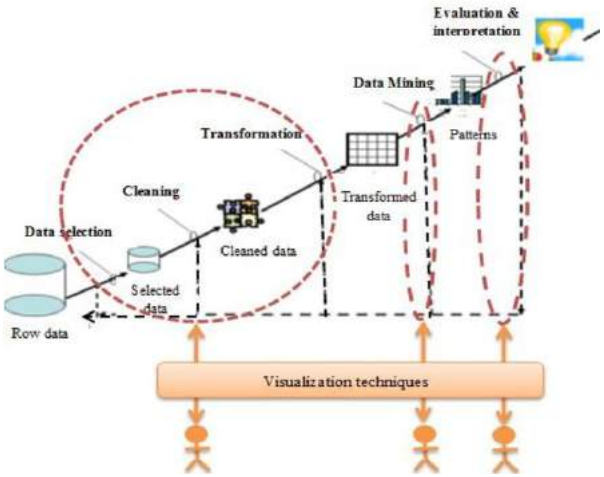


Figura 2. Integración de técnicas de visualización. Recuperado de “Integration of temporal data visualization techniques in a KDD-based DSS Application in the medical field”, de Mohamed, E, (2014).
https://www.semanticscholar.org/paper/Integration-of-temporal-data-visualization-in-a-DSS-Mohamed-Ltifi/0acdc5c6758c34769ae5898a57781e4bd0580be4

## 4. FORMACIÓN DE RECURSOS HUMANOS

El proyecto de investigación se encuentra conformado por ocho docentes-investigadores de distintas áreas de conocimientos que atañen a la Ciencia de Datos, tales como Estadística, Inteligencia Artificial, Lenguajes de Programación y Estructuras de Datos, y cuatro alumnos.
Los investigadores trabajan desde hace varios años en forma conjunta lo que ha permitido generar y asesorar, de modo integral, varios trabajos finales de grado, tesis de posgrado y Beca CIN - Consejo Interuniversitario Nacional- periodo 2020-2021.
Pertenecen a las líneas de investigación los siguientes trabajos finales de la carrera Licenciatura en Ciencias de la Computación.

- "Herramienta de apoyo al aprendizaje de Metaheurísticas" (Autor: Olivares Juan Ignacio. Defendido en 2020)
- "Descripción de los procesos de recolección de datos y extracción de información en Redes Sociales en Ambientes Paralelos y Distribuidos" (Autor: Gouric Guillermo. Defendido 2020)
- Herramienta tecnológica de apoyo al aprendizaje: Problema del Viajante de Comercio, caso asimétrico (Autor: Cocinero, Pablo. En ejecución)

Con docentes investigadores del Instituto de Automática de la Facultad de Ingeniería INAUT-FI y personal del Instituto Nacional de Tecnología Agropecuaria INTA-San Juan, se lleva adelante la tesis de Maestría en Informática "Análisis de Fenómenos en Estaciones Agrometeorológicas mediante Ciencia de Datos" a cargo de un integrante del equipo de trabajo.
Es de destacar que en todos los trabajos citados, está planteado un apartado específico sobre visualización.
La Beca CIN "Abordaje de la Analítica Visual desde un lenguaje de programación-Python como caso de estudio" ha sido otorgada al estudiante Joaquín Cortez, alumno de quinto año de la carrera Licenciatura en Ciencias de la Computación.

Como aporte a la sociedad se ha planificado un curso destinado a las pequeñas y medianas empresas de la provincia de San Juan, a partir de una iniciativa del gobierno y con el aval éste. Dicho curso tiene como finalidad introducir los conceptos básicos de Ciencia de Datos y mostrar las virtudes o ventajas de extraer conocimiento de los datos.

## 5. BIBLIOGRAFÍA


- Abela, A. (2008). *Advanced presentations by design: Creating communication that drives action*. John Wiley & Sons.
- Barcellos, R., Viterbo, J., Bernardini, F., & Trevisan, D. (2018, July). An Instrument for Evaluating the Quality of Data Visualizations. In 2018 22nd International Conference Information Visualisation (IV) (pp. 169-174). IEEE.
- Behrisch, M., Blumenschein, M., Kim, N. W., Shao, L., El-Assady, M., Fuchs, J., ... & Keim, D. A. (2018, June). Quality metrics for information visualization. In *Computer Graphics Forum* (Vol. 37, No. 3, pp. 625-662). https://bib.dbvis.de/uploadedFiles/QMSTAR_QualityMetricsForInformationVisualization_FINAL.pdf
- Benoit, G. (2019). Introduction to Information Visualization: Transforming Data Into Meaningful Information. Rowman & Littlefield.
- Cady, F. (2017). The data science handbook. John Wiley & Sons.
- Chen, M., Feixas, M., Viola, I., Bardera, A., Shen, H. W., & Sbert, M. (2017). Information theory tools for visualization. AK Peters/CRC Press
- Cielen, D., Meysman, A., & Ali, M. (2016). Introducing data science: big data, machine learning, and more,

using Python tools. Manning Publications Co..
- Erraissi, A., Mouad, B., & Belangour, A. (2019, April). A Big Data visualization layer meta-model proposition. In 2019 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO) (pp. 1-5). IEEE.
- Grupo Bit (2020) ¿Cuántos datos se producen en un minuto? https://business-intelligence.grupobit.net/blog/cuantos-datos-se-producen-en-un-minuto
- Grus, J. (2019). Data science from scratch: first principles with python. O'Reilly Media.
- How to choose a visualization (2019). https://www.kdnuggets.com/2019/06/how-choose-visualization.html
- Klenzi, R. O., Malberti, A., & Beguerí, G. (2020). Evaluación de visualizaciones eficientes en ciencia de datos. In *XXII Workshop de Investigadores en Ciencias de la Computación (WICC 2020, El Calafate, Santa Cruz)*.
- Knaflic, C. N. (2015). Storytelling with data: A data visualization guide for business professionals. John Wiley & Sons.
- Laura Igual Muñoz, & Santi Seguí Mesquida. (2017). Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications. Springer.
- Lo, L. Y. H., Ming, Y., & Qu, H. (2019). Learning Vis Tools: Teaching Data Visualization Tutorials. arXiv preprint arXiv:1907.08796.
- Mohamed, E.B., Ltifi, H., & Ayed, M.B. (2014). Integration of temporal data visualization techniques in a KDD-based DSS Application in the medical field. https://www.semanticscholar.org/paper/Integration-of-temporal-data-visualization-in-a-DSS-Mohamed-Ltifi/0acdc5c6758c34769ae5898a57781e4bd0580be4
- Murray, S. (2017). Interactive data visualization for the web: an introduction to designing with D3" O'Reilly Media, Inc.".
- Sosulski, K. (2018). Data Visualization Made Simple: Insights Into Becoming Visual. Routledge.
- VanderPlas, J. (2016). Python data science handbook: essential tools for working with data. " O'Reilly Media, Inc.".
- Wang, J., Hazarika, S., Li, C., & Shen, H. W. (2018). Visualization and visual analysis of ensemble data: A survey. IEEE transactions on visualization and computer graphics.
- Wang, C., & Shen, H. (2011). Information Theory in Scientific Visualization. Entropy, 13, 254-273.
- Ware, C. (2012). Information visualization: perception for design. Elsevier
- Wenqiang Cui (2019) Visual Analytics: A Comprehensive Overview. EEE Xplore Digital Library. Volume 7: 2019 https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8740868
- Why Data Visualization Is The Most Important Skill in a Data Analyst Arsenal (2019). https://www.kdnuggets.com/2019/08/simpliv-data-visualization-data-analyst.html