DATA SCIENCE

Javier Lezama

Incluit

Mayo - Junio de 2019

Presentación

- Doctor en Matemática.

Presentación

- Doctor en Matemática.
- Post doc en Data Science.

Presentación

- Doctor en Matemática.
- Post doc en Data Science.
- Profesor universitario e investigador.

Presentación

- Doctor en Matemática.
- Post doc en Data Science.
- Profesor universitario e investigador.
- Me gusta enseñar lo que he aprendido :-)

Presentación

- Doctor en Matemática.
- Post doc en Data Science.
- Profesor universitario e investigador.
- Me gusta enseñar lo que he aprendido :-)

Presentación

- Doctor en Matemática.
- Post doc en Data Science.
- Profesor universitario e investigador.
- Me gusta enseñar lo que he aprendido :-)

Presentación

- Doctor en Matemática.
- Post doc en Data Science.
- Profesor universitario e investigador.
- Me gusta enseñar lo que he aprendido :-)

https://www.linkedin.com/in/javier-lezama-806b5579/

Introducción

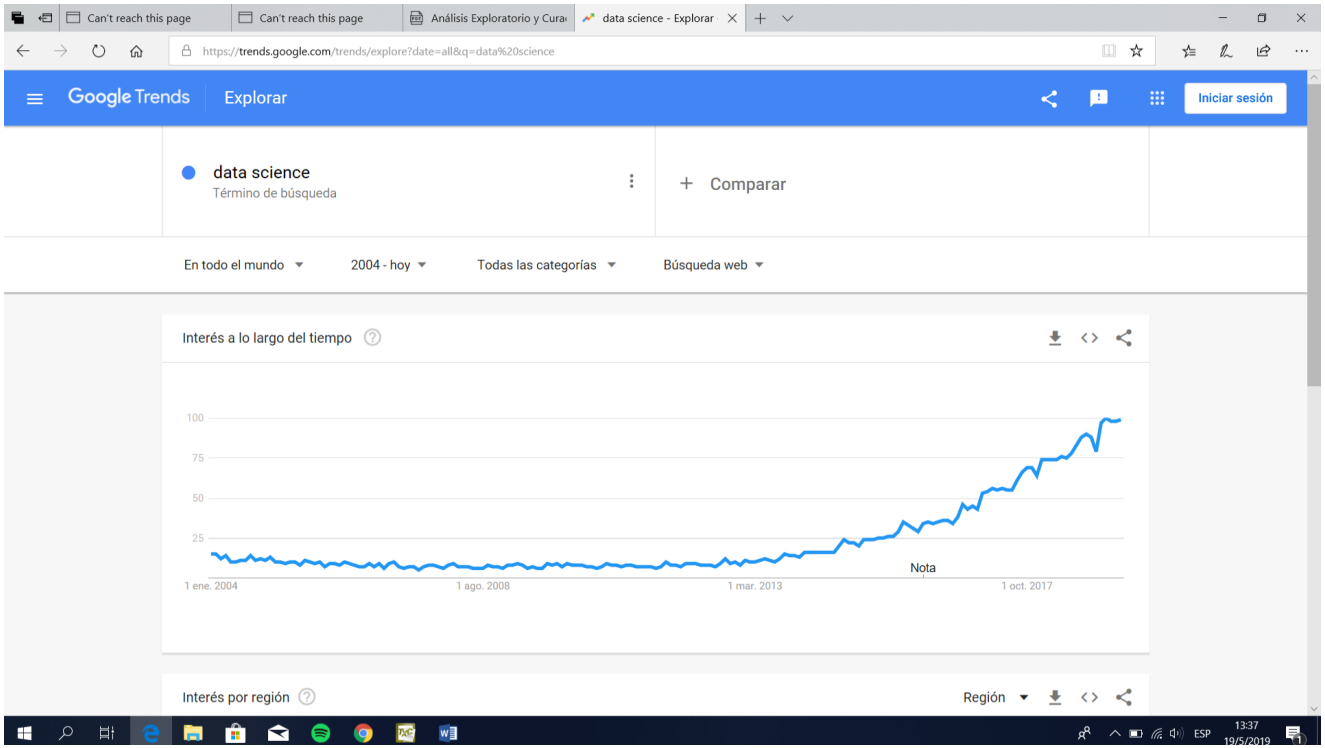
- ¿Que es la ciencia de datos?

Introducción

- ¿Que es la ciencia de datos?

Introducción

- ¿Que es la ciencia de datos?



Que es la ciencia de datos

- ¿El trabajo mas sexi?

Que es la ciencia de datos

- ¿El trabajo mas sexi?
- Harvard Business Review en 2012.

Que es la ciencia de datos

- ¿El trabajo mas sexi?
- Harvard Business Review en 2012.
- https://hbr.org/2012/10/data-scientistthe-sexiest-job-of-the-21st-century

Que es la ciencia de datos

- ¿El trabajo mas sexi?
- Harvard Business Review en 2012.
- https://hbr.org/2012/10/data-scientistthe-sexiest-job-of-the-21st-century

Que es la ciencia de datos

- ¿El trabajo mas sexi?
- Harvard Business Review en 2012.
- https://hbr.org/2012/10/data-scientistthe-sexiest-job-of-the-21st-century

Que es la ciencia de datos

La ciencia de datos es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados.
https://es.wikipedia.org/wiki/Ciencia_de_datos

Que es la ciencia de datos

https://www.oreilly.com/ideas/what-is-data-science (2010).
La ciencia de datos es la práctica de crear productos de datos.
Un producto de datos es una aplicación que no solo manipula datos sino que obtiene su valor creando información a partir de esos datos.

Productos de Datos

- Google Search

Productos de Datos

- Google Search
- Sistemas de recomendación de: Amazon, Netflix, Spotify

Productos de Datos

- Google Search
- Sistemas de recomendación de: Amazon, Netflix, Spotify
- Sistemas de publicidad online

Productos de Datos

- Google Search
- Sistemas de recomendación de: Amazon, Netflix, Spotify
- Sistemas de publicidad online
- Grammarly

Productos de Datos

- Google Search
- Sistemas de recomendación de: Amazon, Netflix, Spotify
- Sistemas de publicidad online
- Grammarly
- etc.

¿Cómo hacer productos es Ciencia y no Ingeniería?

El estado del arte de ingeniería está más cerca que nunca de la ciencia.
Particularmente por el cambio de paradigma del diseño inteligente al descubrimiento de conocimiento, empujado por las metodologías ágiles/lean.
Ciencia de datos es para hacer énfasis en la generación de conocimiento a partir de los datos, proceso propio de las ciencias.
Ingeniería de datos existe y se refiere a las técnicas y prácticas de manipulación de datos, más propias de la ingeniería.

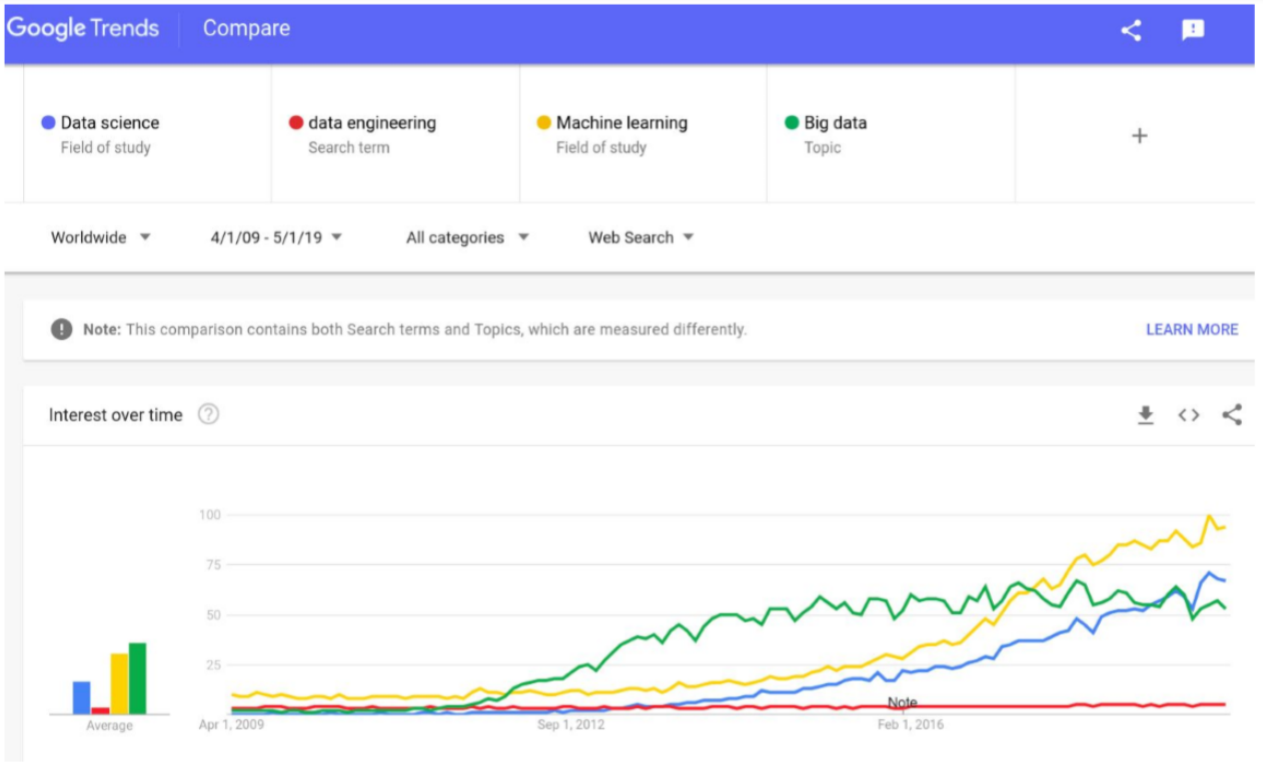
Puestos de datos

Data scientist: diseño, análisis, evaluación, KPI
Machine learning engineer: implementación de modelos, entrenamientos
Data engineer: ETL, data pipelines
Otros perfiles:

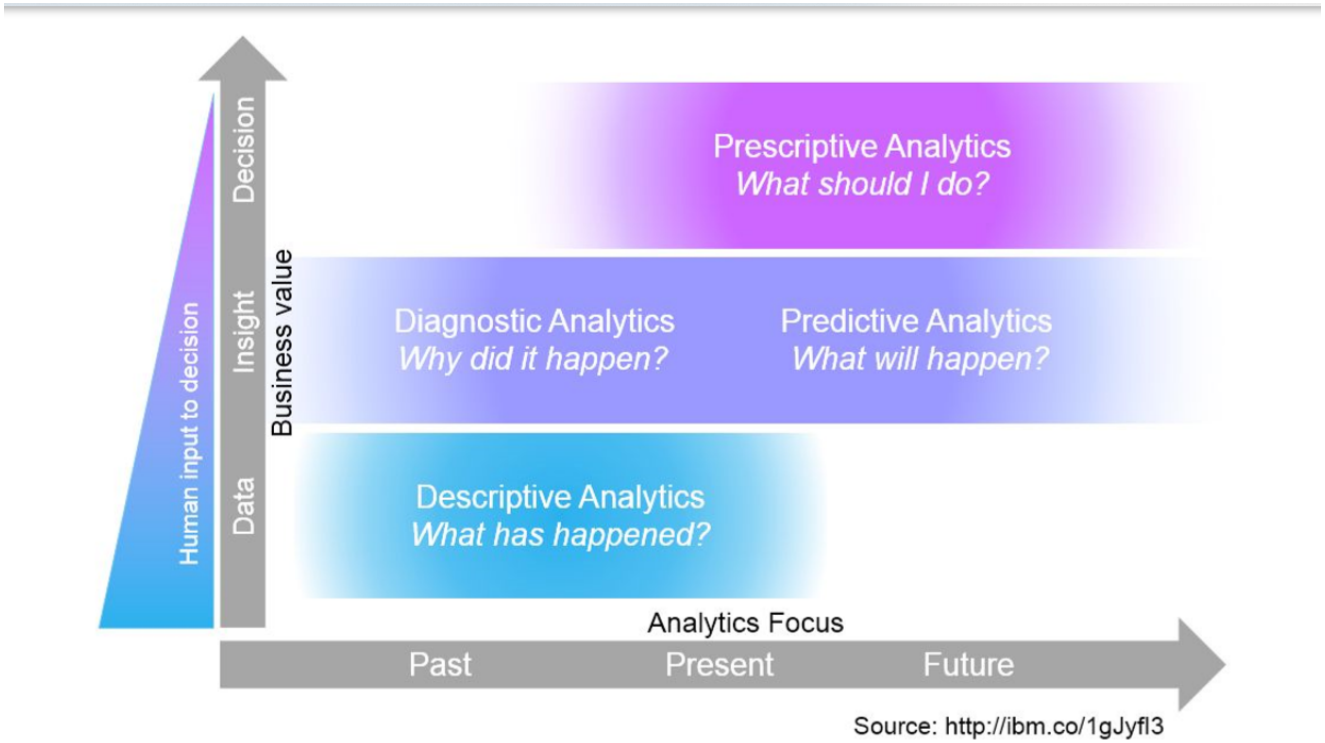
* Data analyst
* Business analyst
* Business Intelligence developer

https://www.oreilly.com/ideas/why-a-data-scientist-is-not-a-data-engineer
https://www.zarantech.com/blog/top-10-data-science-career-options-shaping-our-future/

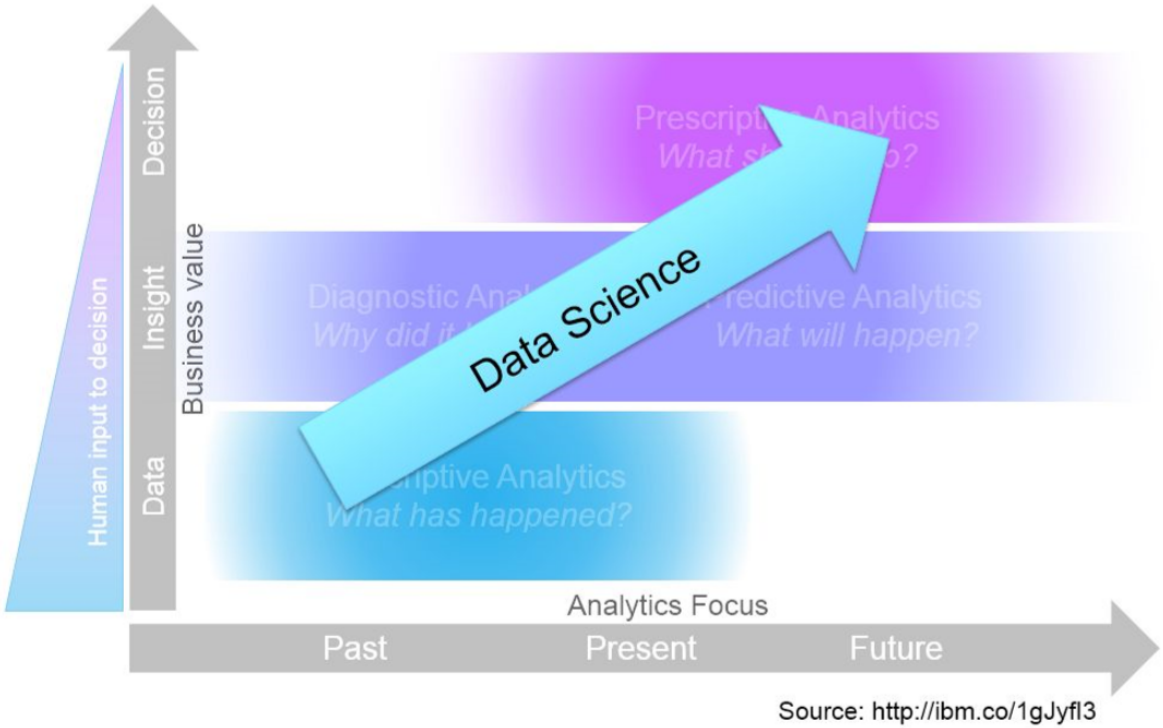
Popularidad de disciplinas



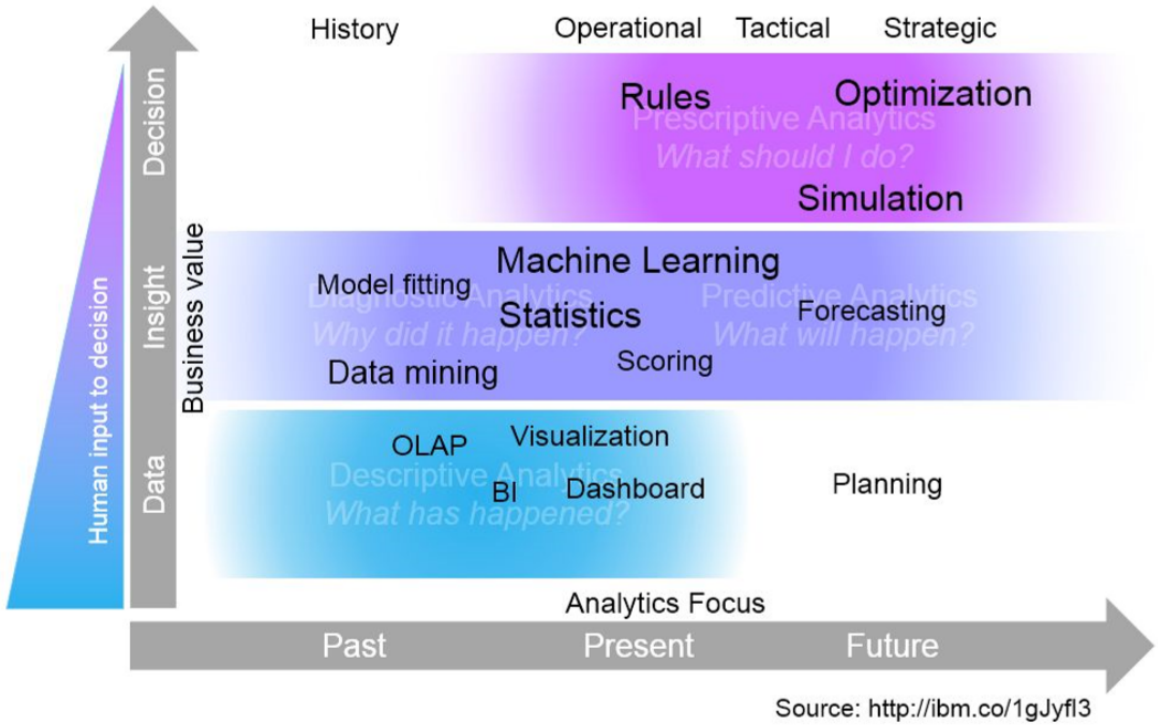
Ciencia de datos y organizaciones



Ciencia de datos y organizaciones



Ciencia de datos y organizaciones



Data Science Process

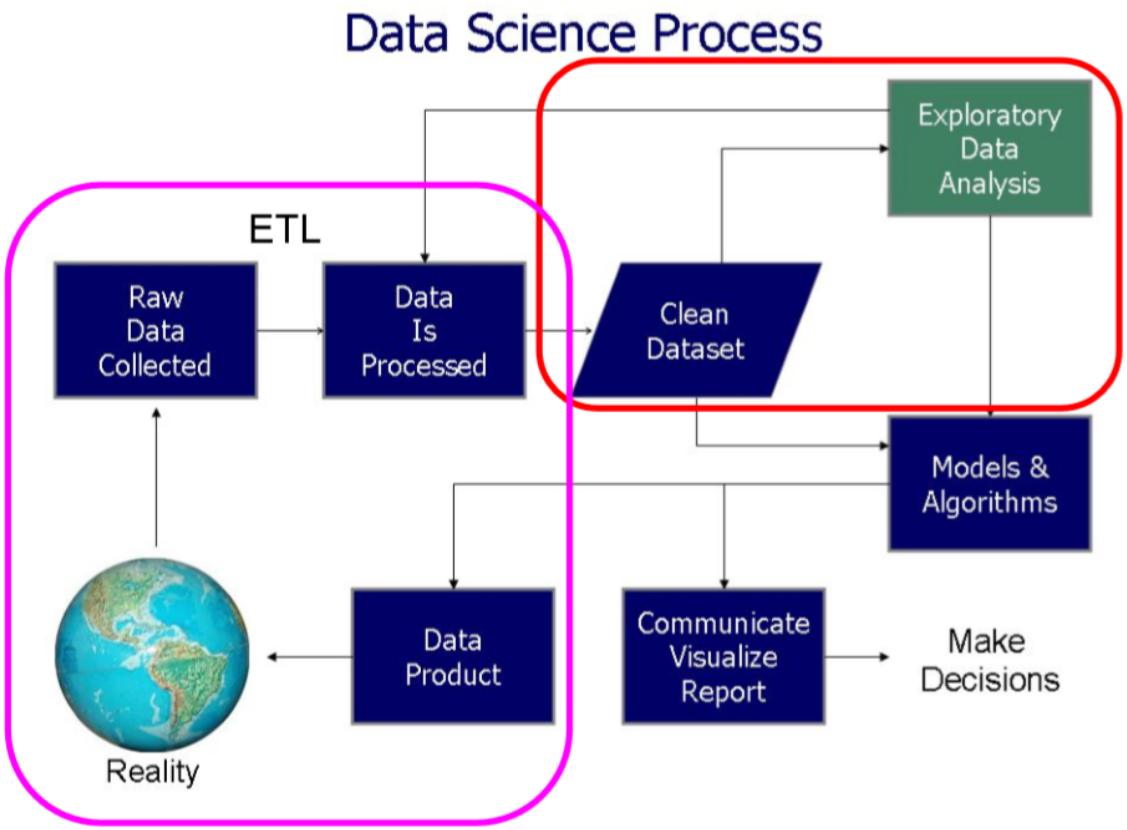
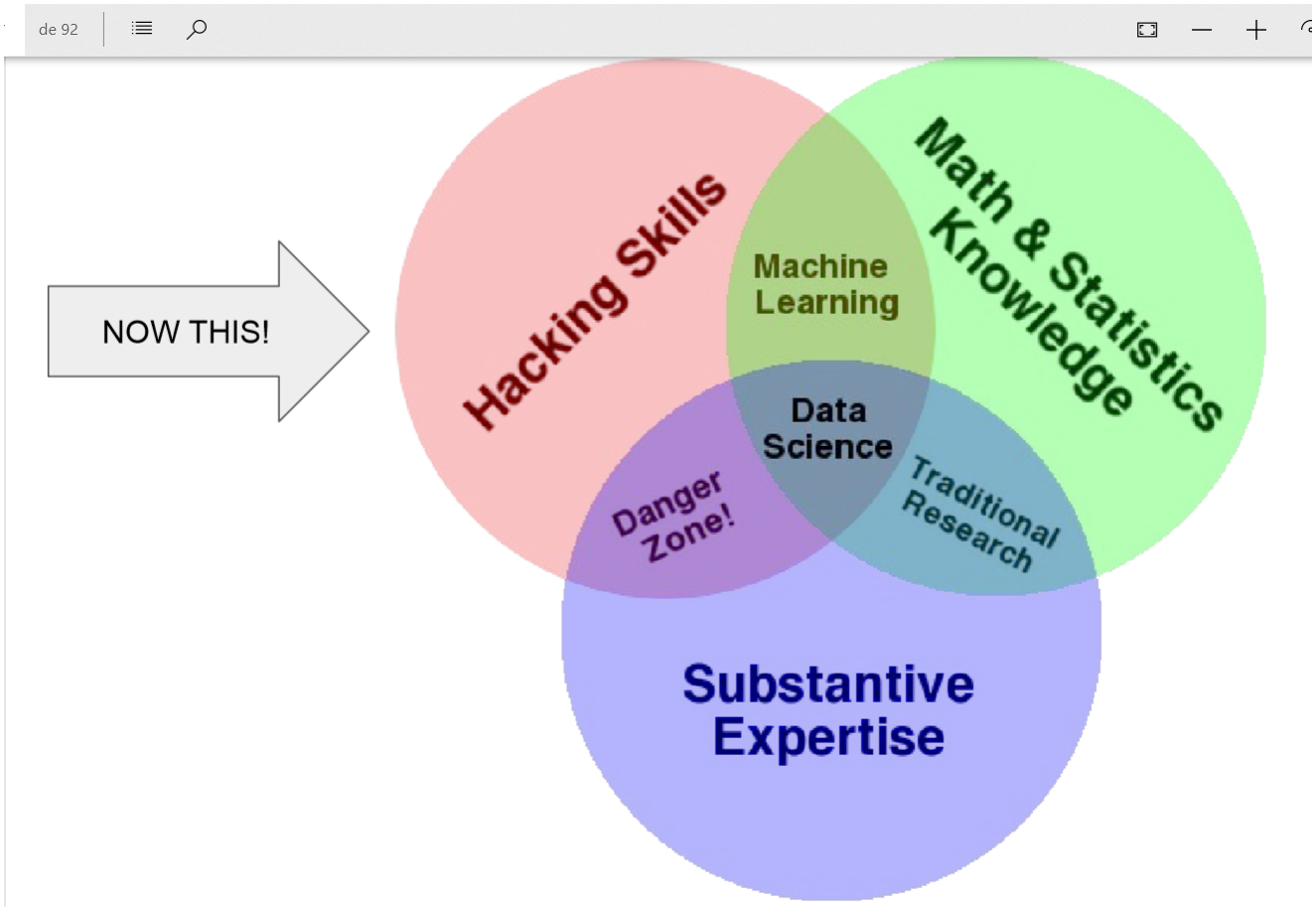
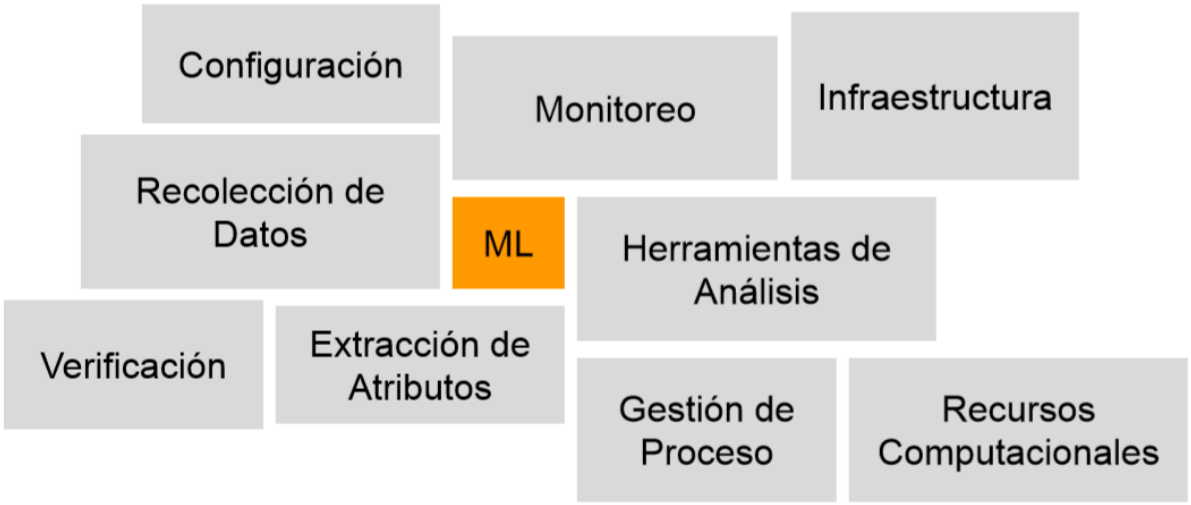


Diagrama de Venn

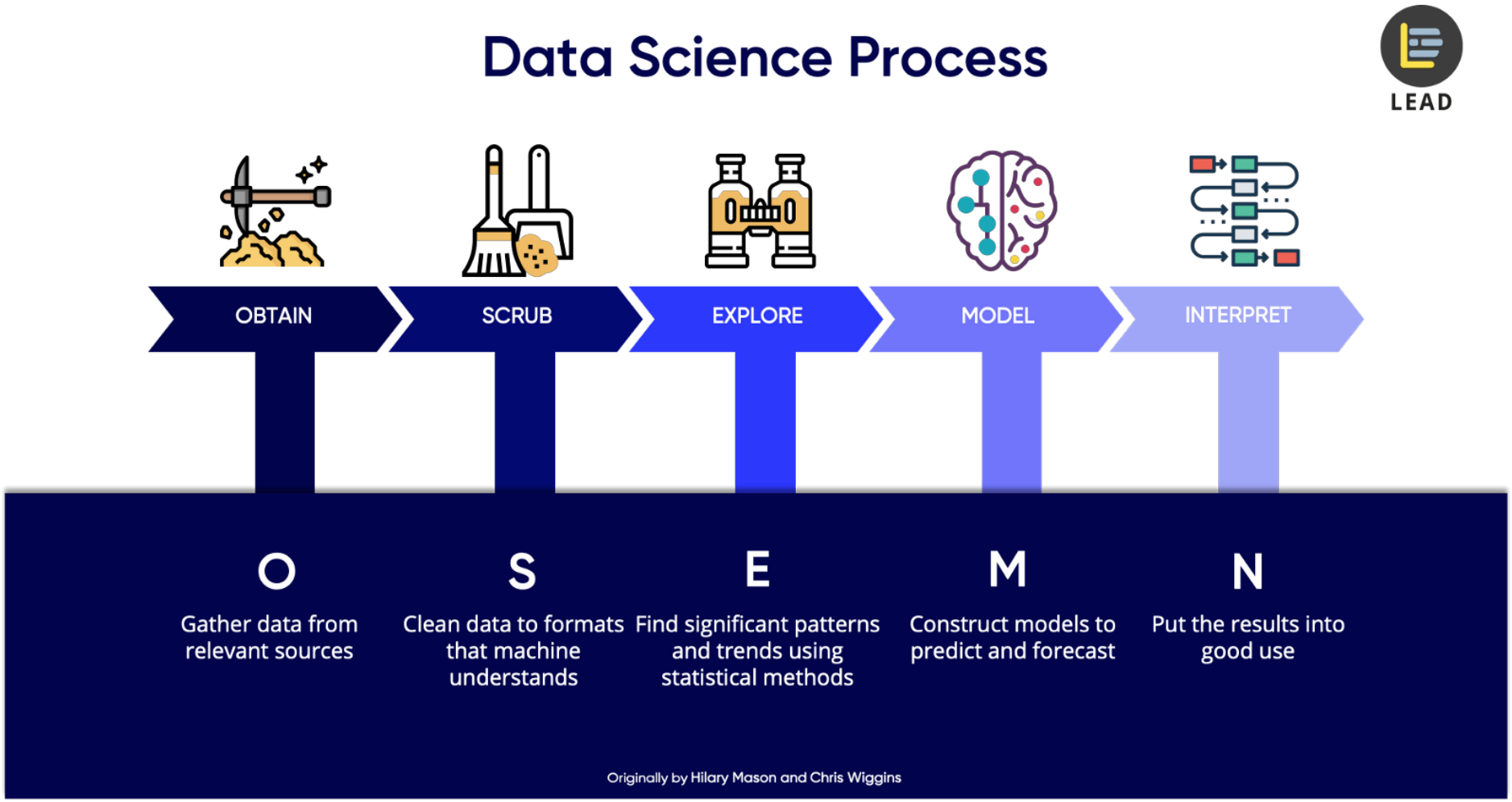


Producto de datos



https://youtu.be/vdG7uKQ2eKk?t=107

Data science process



Workflow de un proyecto en Data Science

- Objetivo.

Workflow de un proyecto en Data Science

- Objetivo.
- Obtención de los datos.

Workflow de un proyecto en Data Science

- Objetivo.
- Obtención de los datos.
- Exploración de los datos y limpieza de los mismos.

Workflow de un proyecto en Data Science

- Objetivo.
- Obtención de los datos.
- Exploración de los datos y limpieza de los mismos.
- Modelado: entrenamiento.

Workflow de un proyecto en Data Science

- Objetivo.
- Obtención de los datos.
- Exploración de los datos y limpieza de los mismos.
- Modelado: entrenamiento.
- Prueba.

Workflow de un proyecto en Data Science

- Objetivo.
- Obtención de los datos.
- Exploración de los datos y limpieza de los mismos.
- Modelado: entrenamiento.
- Prueba.
- Visualización e interpretación.

Workflow de un proyecto en Data Science

- Objetivo.
- Obtención de los datos.
- Exploración de los datos y limpieza de los mismos.
- Modelado: entrenamiento.
- Prueba.
- Visualización e interpretación.

Workflow de un proyecto en Data Science

- Objetivo.
- Obtención de los datos.
- Exploración de los datos y limpieza de los mismos.
- Modelado: entrenamiento.
- Prueba.
- Visualización e interpretación.

Herramientas con las que trabajaremos

- Conda.

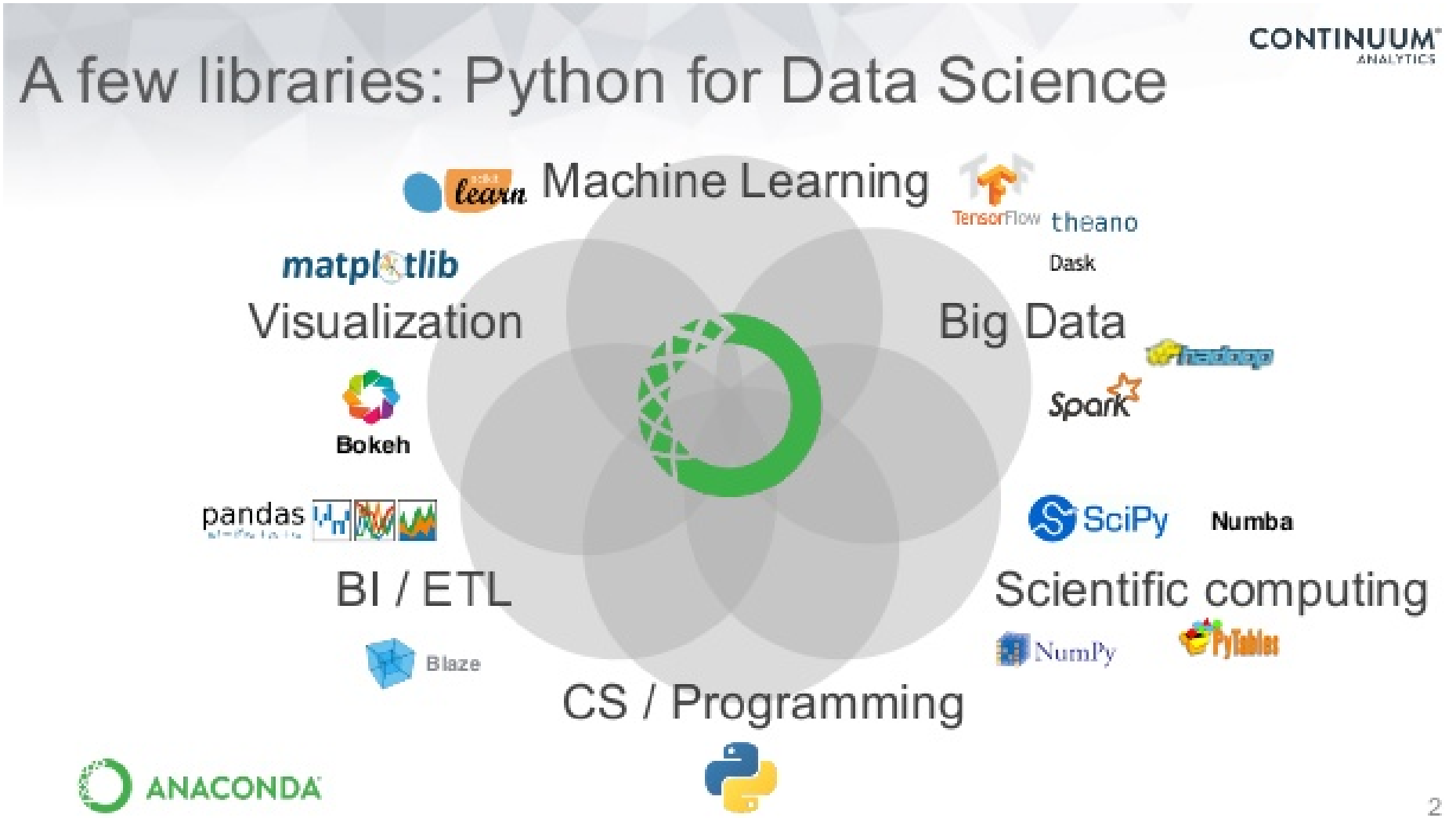
Herramientas con las que trabajaremos

- Conda.
- Jupyter notebook.

Herramientas con las que trabajaremos

- Conda.
- Jupyter notebook.
- Librerias: Numpy, Matplotlib, Pandas, Scikit-learn, etc.

Librerias de python



Numpy

- A powerful N-dimensional array object.

Numpy

- A powerful N-dimensional array object.
- Vectorización.

Numpy

- A powerful N-dimensional array object.
- Vectorización.
- Sophisticated (broadcasting) functions.

Numpy

- A powerful N-dimensional array object.
- Vectorización.
- Sophisticated (broadcasting) functions.
- Tool for integrating c/c++ and Fortan code.

Numpy

- A powerful N-dimensional array object.
- Vectorización.
- Sophisticated (broadcasting) functions.
- Tool for integrating c/c++ and Fortan code.
- Useful linear algebra, Fourier transform, and random number capabilities.

Matplotlib

- Es una libreria de Python de dibujos 2D.

Matplotlib

- Es una libreria de Python de dibujos 2D.
- Funciones.

Matplotlib

- Es una libreria de Python de dibujos 2D.
- Funciones.
- Histogramas.

Matplotlib

- Es una libreria de Python de dibujos 2D.
- Funciones.
- Histogramas.
- Gráficas de barras.

Matplotlib

- Es una libreria de Python de dibujos 2D.
- Funciones.
- Histogramas.
- Gráficas de barras.
- Grficos de dispersión, etc.

Pandas

- Python Data Analisis Library.

Pandas

- Python Data Analisis Library.
- Herramienta de manipulación de datos de alto nivel.

Pandas

- Python Data Analisis Library.
- Herramienta de manipulación de datos de alto nivel.
- Estructura de datos clave: DataFrame.

Pandas

- Python Data Analisis Library.
- Herramienta de manipulación de datos de alto nivel.
- Estructura de datos clave: DataFrame.
- DataFrame: permite almacenar y manipular datos tabulados en filas de observaciones y columnas de variables.

Pandas

- Python Data Analisis Library.
- Herramienta de manipulación de datos de alto nivel.
- Estructura de datos clave: DataFrame.
- DataFrame: permite almacenar y manipular datos tabulados en filas de observaciones y columnas de variables.

Pandas

- Python Data Analisis Library.
- Herramienta de manipulación de datos de alto nivel.
- Estructura de datos clave: DataFrame.
- DataFrame: permite almacenar y manipular datos tabulados en filas de observaciones y columnas de variables.

https://pandas.pydata.org/pandasdocs/stable/getting_started/10min.html.

¿Dudas? ¿Comentarios?

Bibliografia

https://www.anaconda.com/distribution/
https://www.hackerrank.com/dashboard
https://www.numpy.org/
https://pandas.pydata.org/
https://matplotlib.org/
https://seaborn.pydata.org/