

Ciencia de datos: Aprendiendo lo básico

Roberto Muñoz, PhD
Astronomer and Data Scientist
Pontificia Universidad Católica de Chile



github.com/rpmunoz

Organiza:



Colabora:



Patrocinan:



Auspicia:



Evolución procesamiento de datos

- 1890: Se usa la máquina tabuladora de Hollerith para procesar los datos del censo de EE.UU.

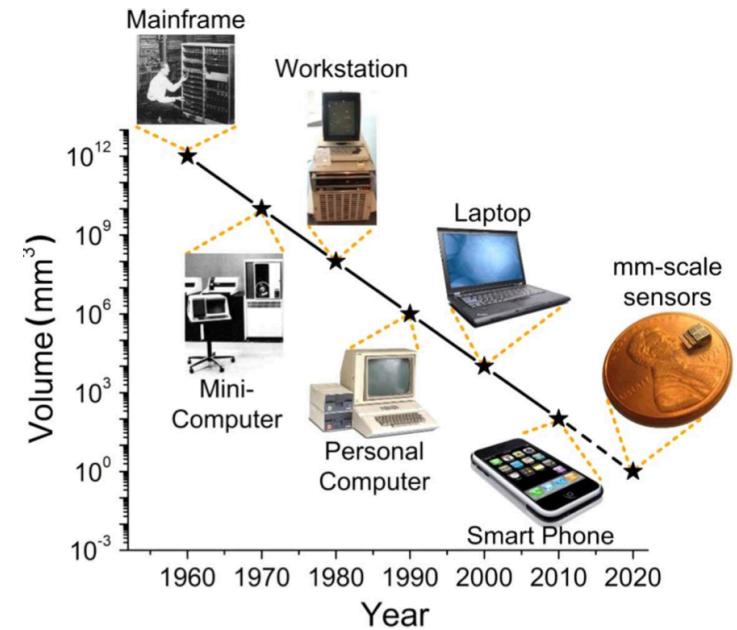
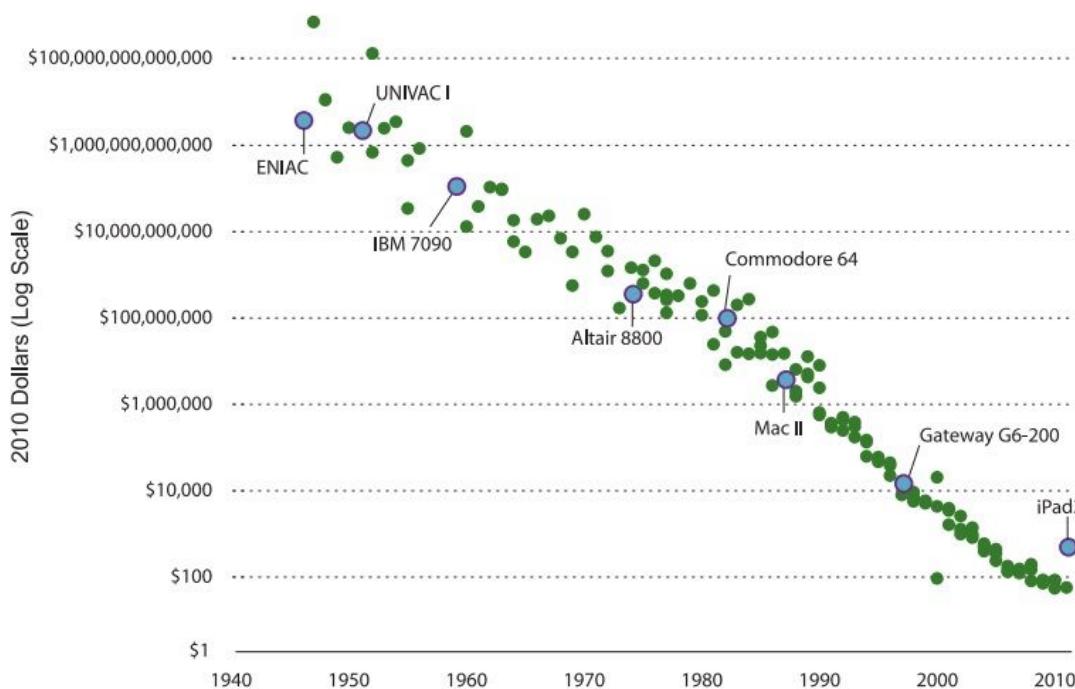


- 1951: Se diseña el primer computador electrónico con fines comerciales, UNIVAC I.



Costo del cómputo

- Desde la invención de los computadores electrónicos, tanto el precio como el tamaño han disminuido sostenidamente.

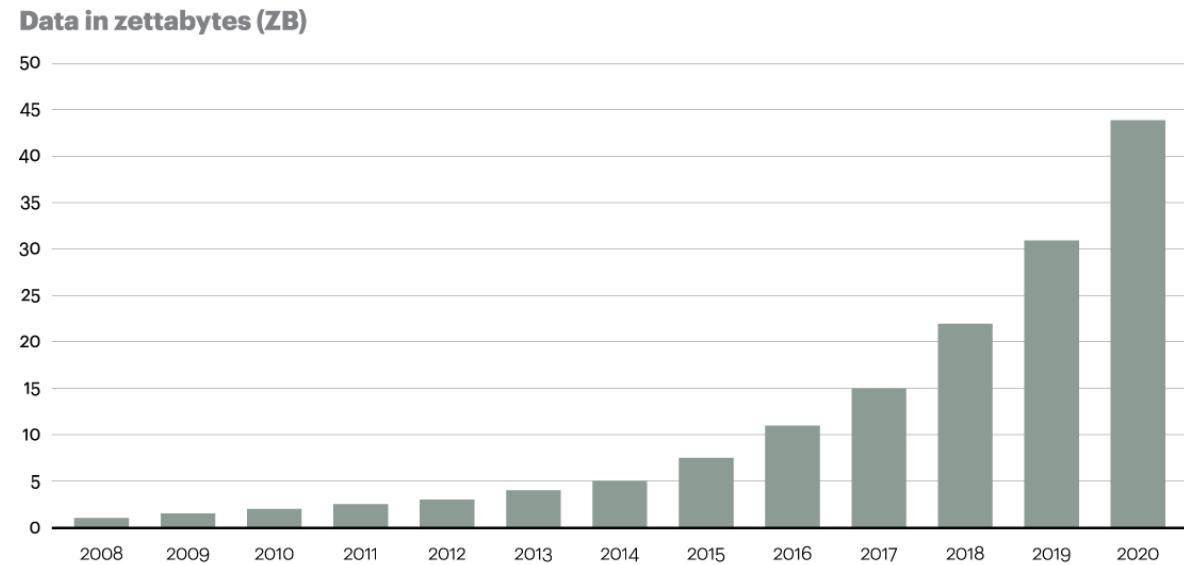


Tsunami de datos

- Durante las últimas décadas la sociedad en su conjunto se ha digitalizado.
- Mayor capacidad de cómputo y tecnología más asequible han permitido un crecimiento explosivo de los datos.

Los datos crecen a una tasa anual del 40%.
Se estima una producción de 45 ZB para el 2020.

Fuente: Oracle, 2012



Comunidad Open Source

- Una mayor variedad y cantidad de datos trae consigo nuevos desafíos.
- Desarrollo continuo de herramientas y métodos para analizar los datos.
- Transición de software empaquetado y comercial a uno desarrollado por comunidad open source.

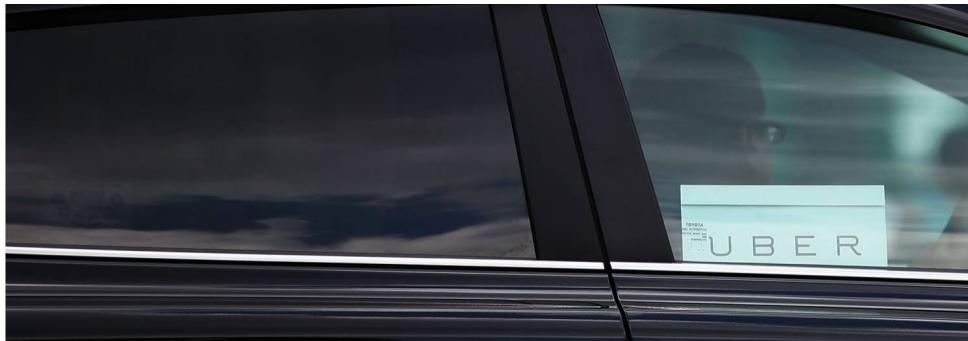


Casos notables

Análisis de uso de Taxis y Uber en NYC Open Data+Open Source



nyc-taxi-data
uber-tlc-foil-response



An Uber car. SPENCER PLATT / GETTY IMAGES

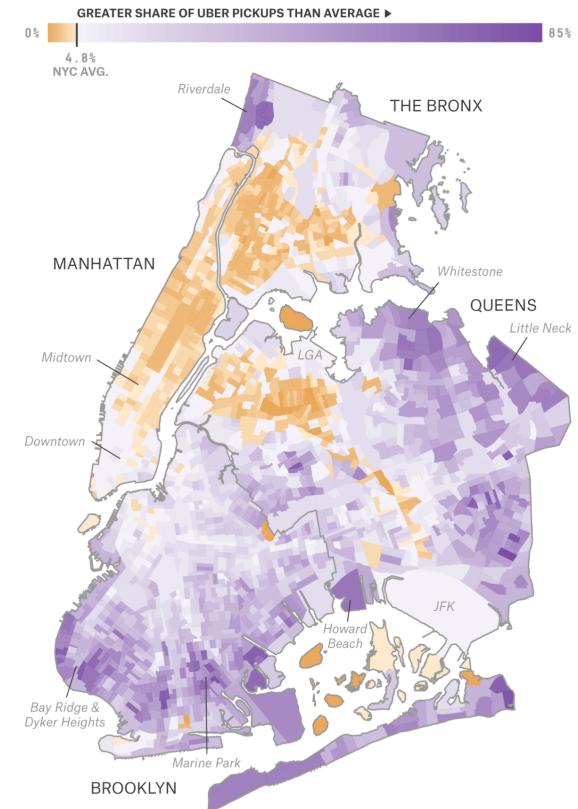
AUG 10, 2015 AT 2:06 PM

Uber Is Serving New York's Outer Boroughs More Than Taxis Are

But most of its rides, like those of taxis, still start in Manhattan.

Fuente: FiveThirtyEight

New York City's Edges Are Uber-Heavy
Share of all Uber, yellow cab and green cab pickups that were by
Ubers from April through September 2014, by census tract



REUBEN FISCHER-BAUM

SOURCE: TAXI & LIMOUSINE COMMISSION



Casos notables

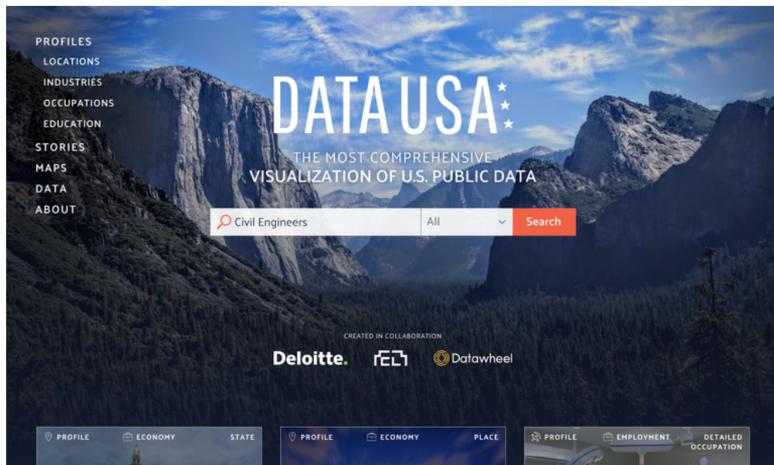
Análisis de datos públicos de USA Open Data+Open Source



datausa

Website Seeks to Make Government Data Easier to Sift Through

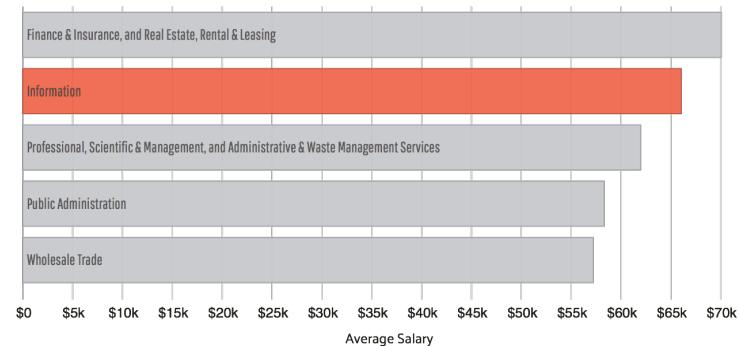
By STEVE LOHR APRIL 4, 2016



Data USA, a project by the M.I.T. Media Lab, seeks to better organize and visualize government data.

Fuente: The New York Times

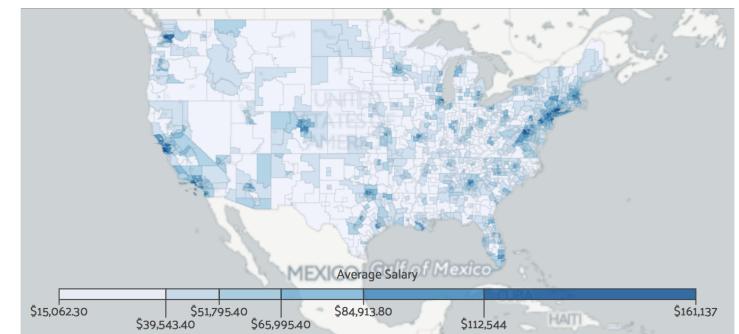
Average Salary for Information



Dataset: 2014 ACS PUMS 1-year Estimate
Source: Census Bureau

DATAUSA:

Wage by Location for Information



Dataset: 2014 ACS PUMS 5-year Estimate
Source: Census Bureau

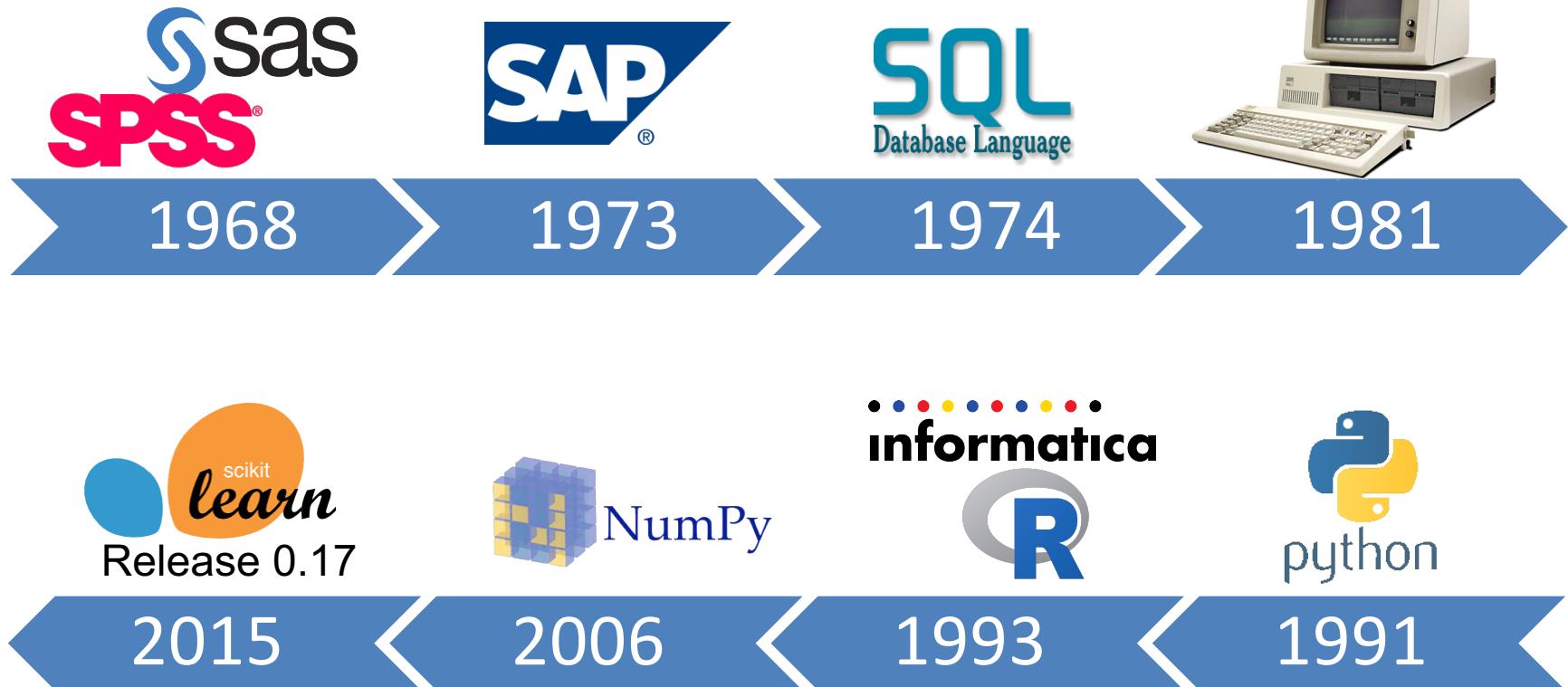
DATAUSA:

Roberto Muñoz



github.com/rpmunoz

Evolución del Analytics



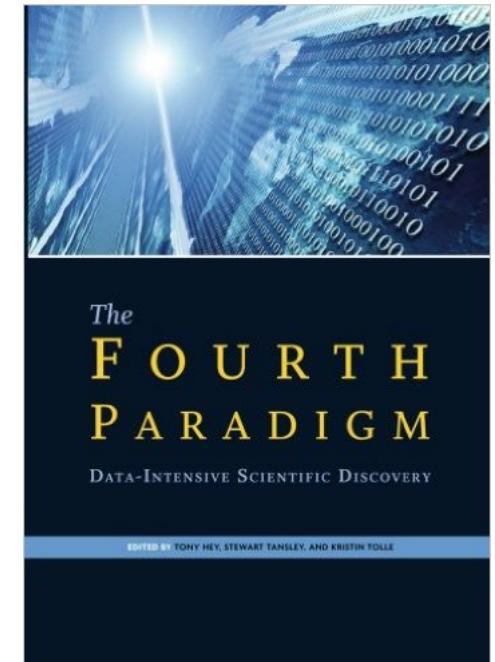
¿Qué es la Ciencia de datos?

- La Ciencia de datos o **Data Science** es un campo interdisciplinario que se ocupa de los procesos y sistemas usados en la extracción de conocimiento a partir del análisis de datos.
- Se dice interdisciplinario pues requiere conocimientos de los campos de la computación, matemáticas y estadística.



¿Cambio de paradigma?

- Los datos digitales y las tecnologías han cambiado la manera en cómo vivimos y cómo entendemos el mundo.
- Jim Gray, investigador de Microsoft y pionero en bases de datos introdujo el concepto del cuarto paradigma.
- Era experimental, teórica, computacional y últimamente la Era del dato.



Carácter interdisciplinario

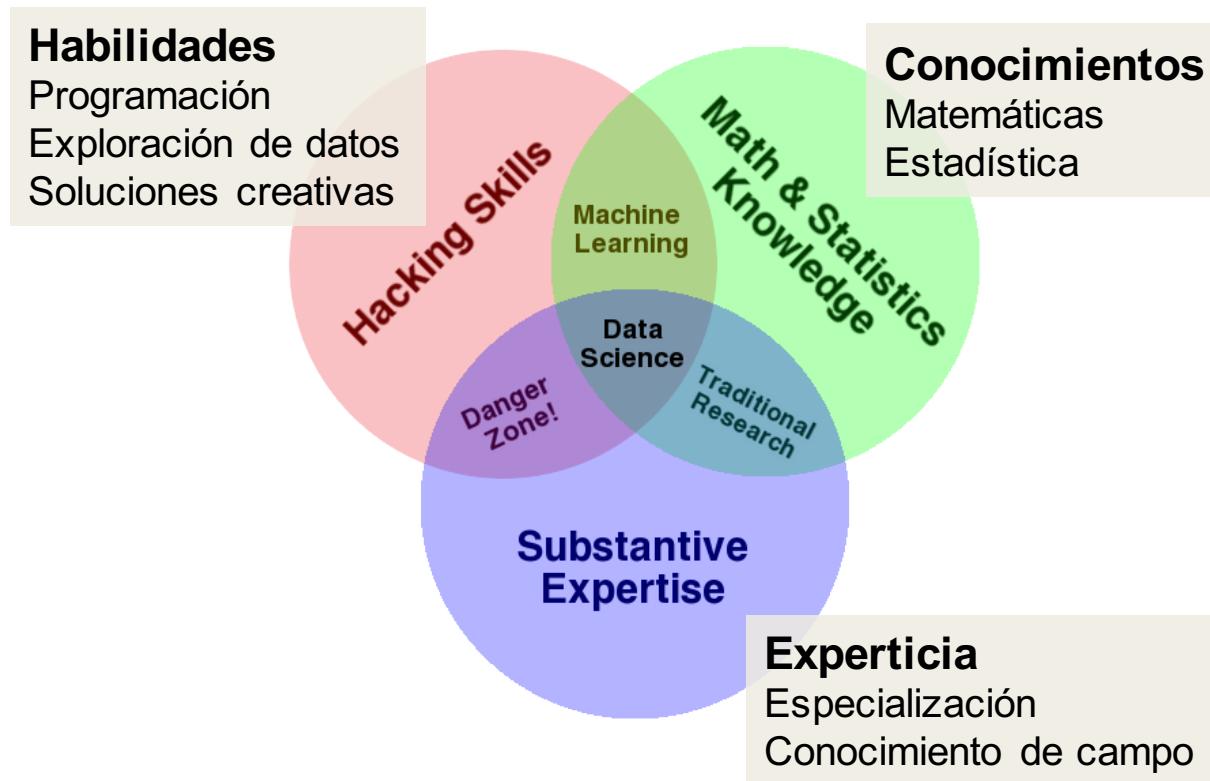


Diagrama de Venn para Data Science
Drew Conway (2010)

¿Qué hace un Data Scientist?

- Profesional que posee las herramientas y los conocimientos necesarios para:
 - **Recolectar y filtrar** datos de diversas fuentes
 - **Explorar** de manera efectiva un set de datos
 - **Obtener** información valiosa oculta en los datos
 - **Construir** modelos que permitan tomar decisiones informadas.

Data Scientist: Persona que es mejor en estadística que cualquier ingeniero de software y que es mejor en ingeniería de software que cualquier estadístico.



Conocimientos y Habilidades

- Formación universitaria en las áreas de Ingeniería y Ciencias Naturales. Idealmente tienen Magister y PhD.
- Poseen conocimientos de Matemáticas, Estadística y Programación computacional.
- Se caracterizan por su **curiosidad intelectual**, son capaces de **diseñar experimentos** y **comunicar de manera efectiva** los resultados.



Roles en la Organización

Usuario de negocio



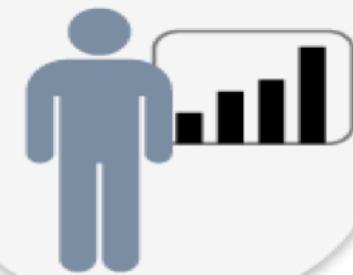
Sponsor del Proyecto



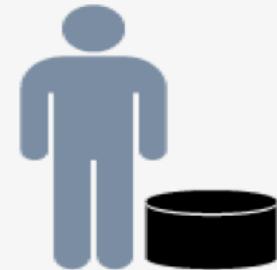
Administrador del Proyecto



Analista de BI



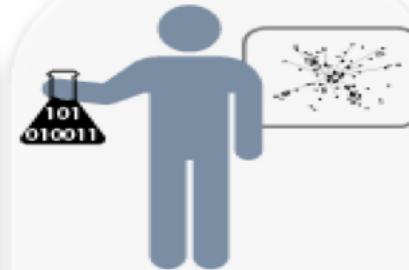
Administrador de BD



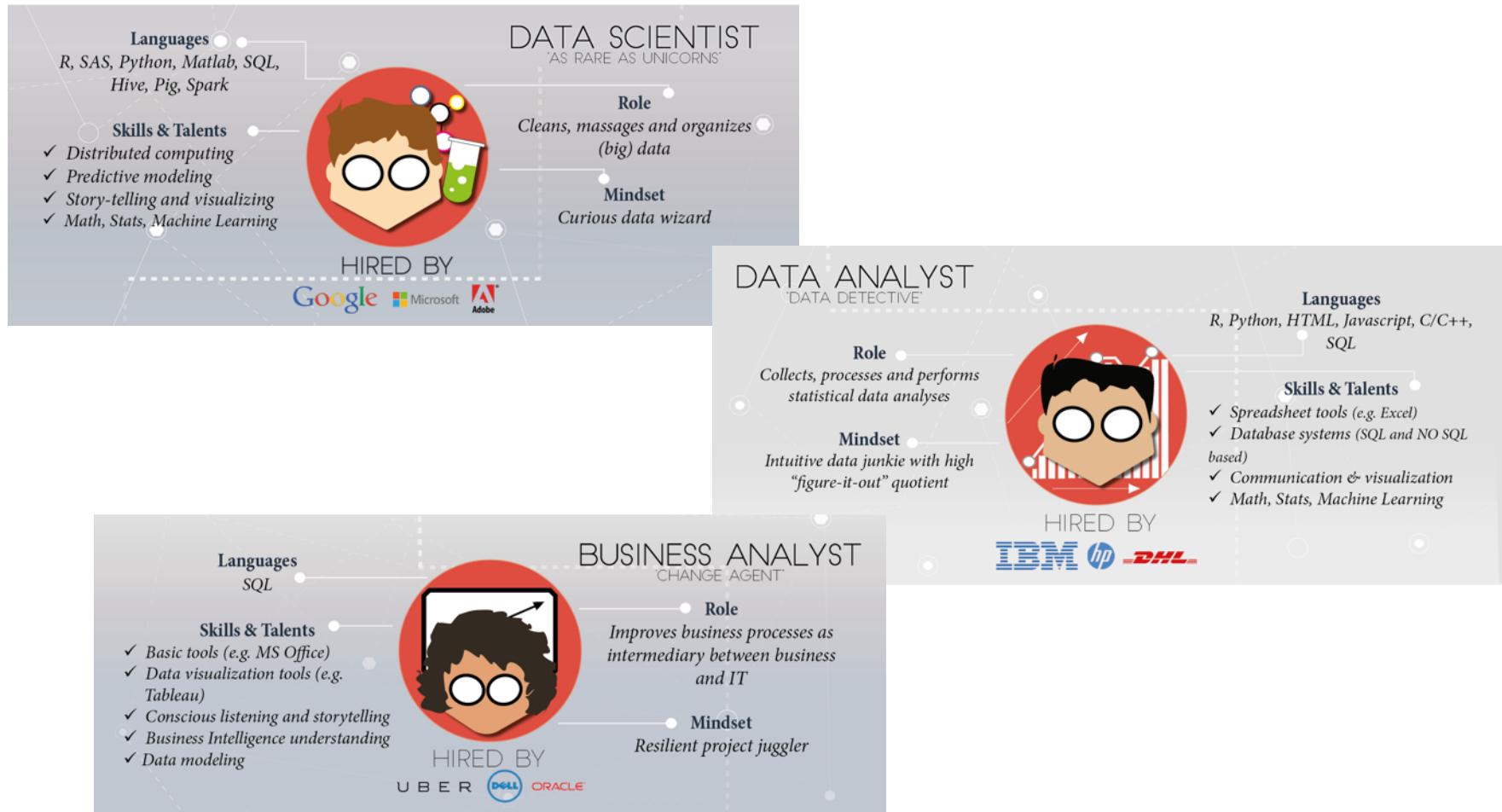
Ingeniero de Datos



Data Scientist



Trabajo en equipo

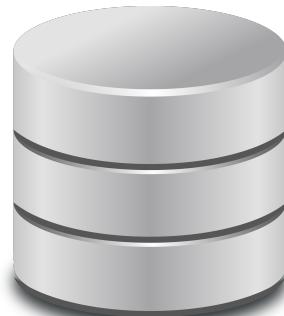


Fuente: DataCamp



Tipos de datos

- Los datos son el punto de partida para todo análisis.
- Tipos de datos de acuerdo a organización
 - **Estructurados:** Están altamente organizados. Se almacenan en una base de datos relacional.

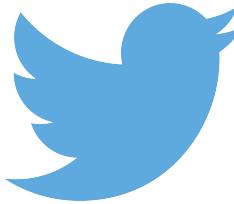


BD

Año	PIB (\$ Millones)	Consumo Eléctrico (GWh)
1993	32.559.292	21.011,3
1994	34.416.724	22.730,7
1995	38.028.591	24.910,2
1996	40.831.596	27.969,0
1997	43.526.546	30.351,5
1998	44.944.340	33.015,8
1999	44.616.349	35.921,3
2000	46.605.199	38.867,4

Tipos de datos

- **No estructurados:** Son datos crudos y no están organizados. Deben ser procesados y transformados para luego ser almacenados en una base de datos.



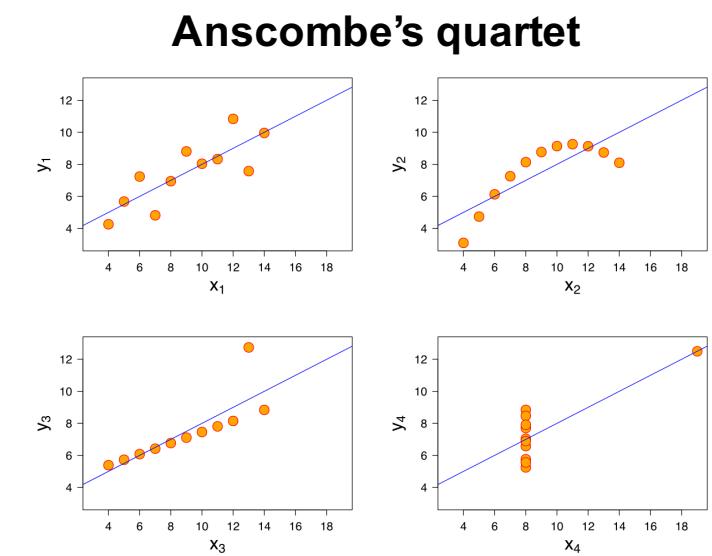
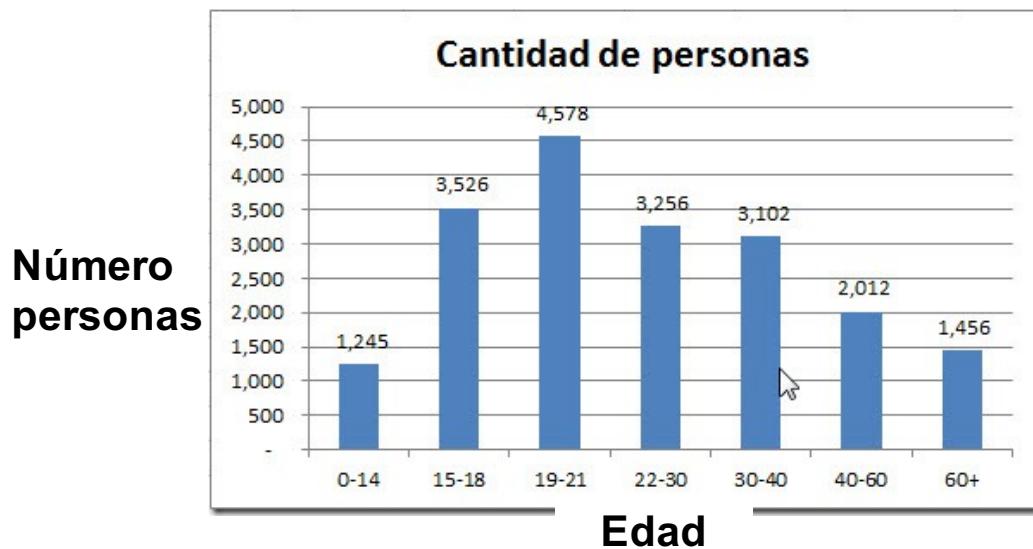
Lenguajes de programación

- Los lenguajes más usados por la comunidad de Data Science son Python y R. Se estima que Python tiene más de 30M de usuarios y R más de 16M.
 - R es más funcional y los módulos de análisis estadístico vienen incorporados.
 - Python es más orientado a objetos y deben cargarse módulos para hacer análisis.



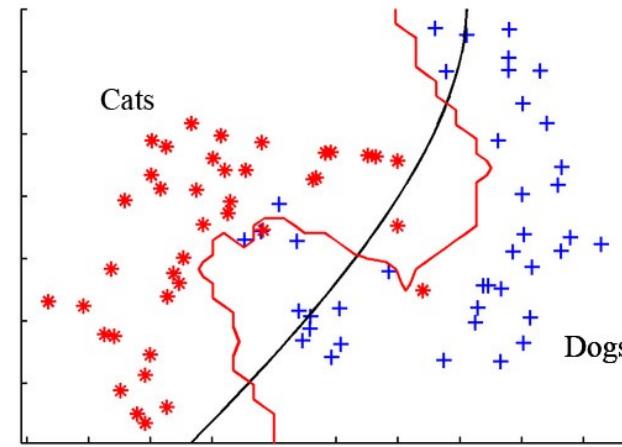
Visualizaciones

- Las visualizaciones juegan un rol importante en todo el proceso de análisis de datos. Permiten explorar los datos, examinar resultados y comparar cualitativamente los modelos.

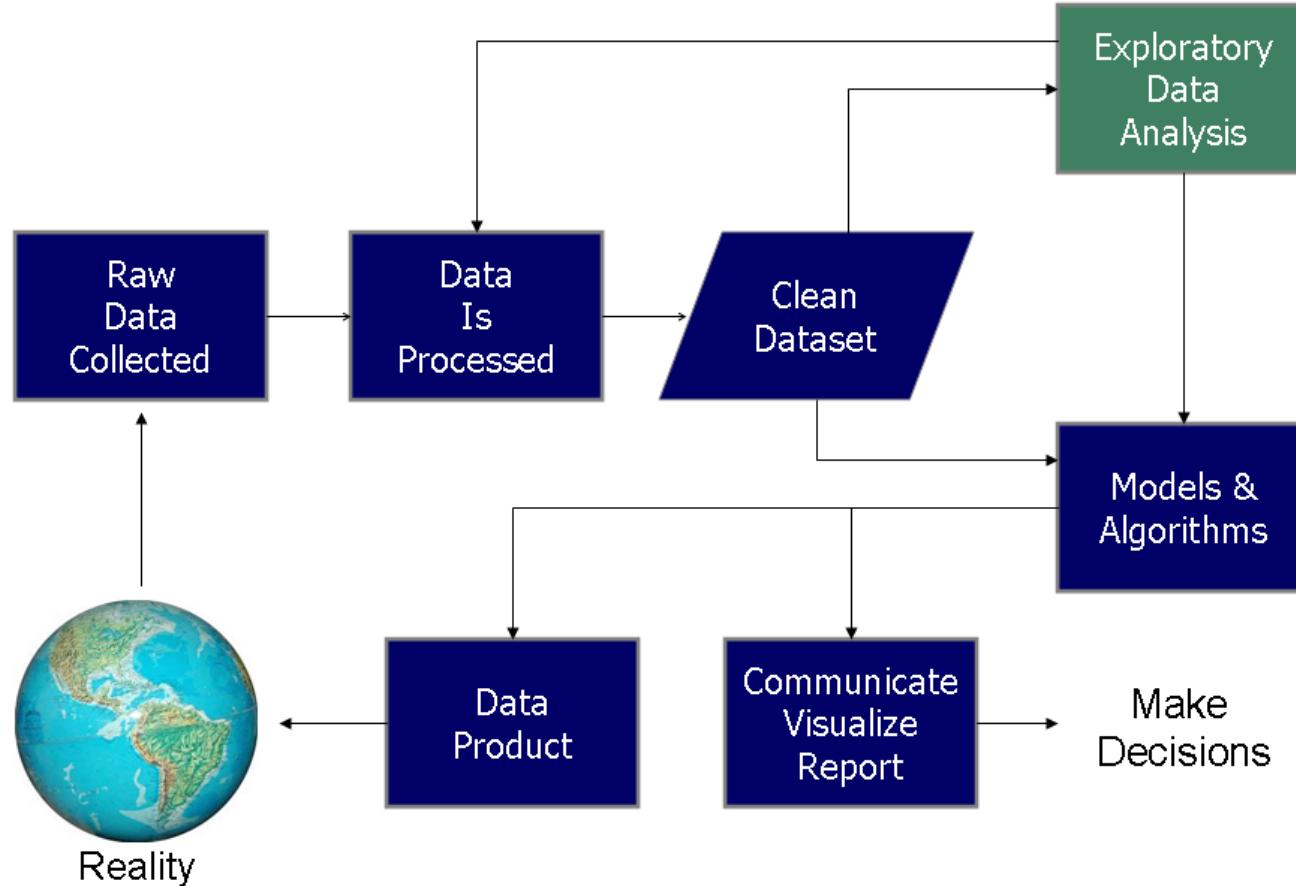


Construcción de modelos

- La construcción y validación de modelos son es clave para los objetivos.
- Permiten entender el comportamiento del sistema, definir cantidades de interés, buscar outliers en los datos y últimamente hacer análisis predictivo.



Esquema simple Data Science



Recursos en Internet



Data Science Central

<http://www.datasciencecentral.com>

kaggle™

<http://blog.kaggle.com>



<http://www.kdnuggets.com>

Quora

<https://www.quora.com/topic/Data-Science>



Tutorial análisis de datos

- Usaremos un notebook de Jupyter y Python 3 para mostrar parte del proceso que se hace en Data Science.
- El notebook de Python se puede descargar desde la carpeta **CIO2016** en github.com/rpmunoz/datascience

IP[y]:
IPython



**“AN APPROXIMATE ANSWER TO THE RIGHT
PROBLEM IS WORTH A GOOD DEAL MORE THAN
AN EXACT ANSWER TO AN APPROXIMATE
PROBLEM.”**

JOHN TUKEY

Gracias

Organiza:



Colabora:

GERENCIA

Patrocinan:

ACTI
Instituto Chileno de Estudios de Tecnología de Información S.A.

CETIUC

Chiletec
gacita - Software y Servicios Chile A.G.

Auspicia:

HDI
Grupo HDI

soluciones
BI Consulting

Bibliografía

- A very short history of Data Science
- From Data Mining to Knowledge Discovery in Databases
- A Project-Based Case Study of Data Science Education

