Data Analytics y Data Science

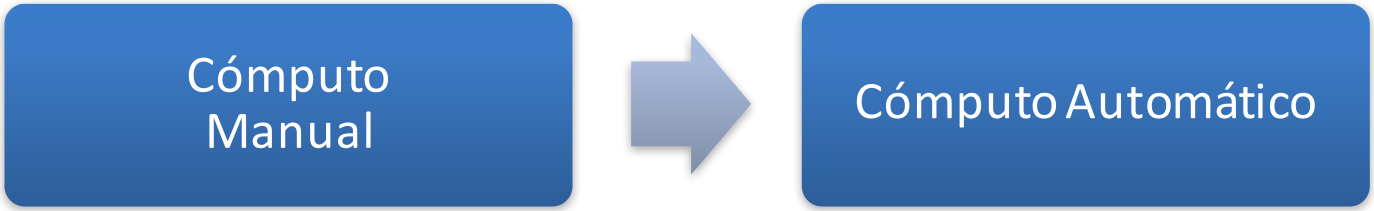
Roberto Muñoz, PhD
Astrónomo y Data Scientist
MetricArts

github.com/rpmunoz

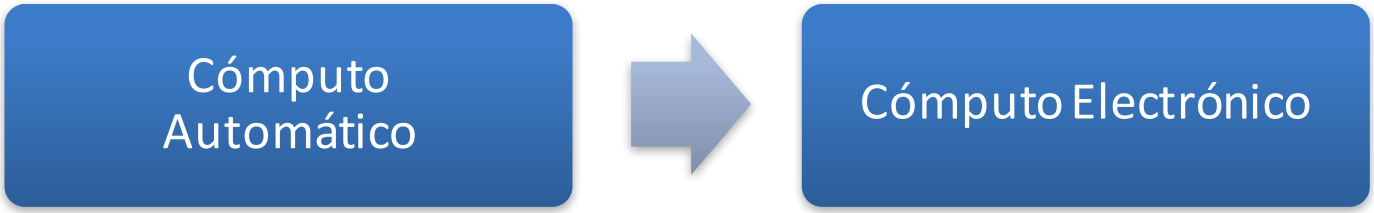
@RobertoKPax

Evolución procesamiento de datos

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

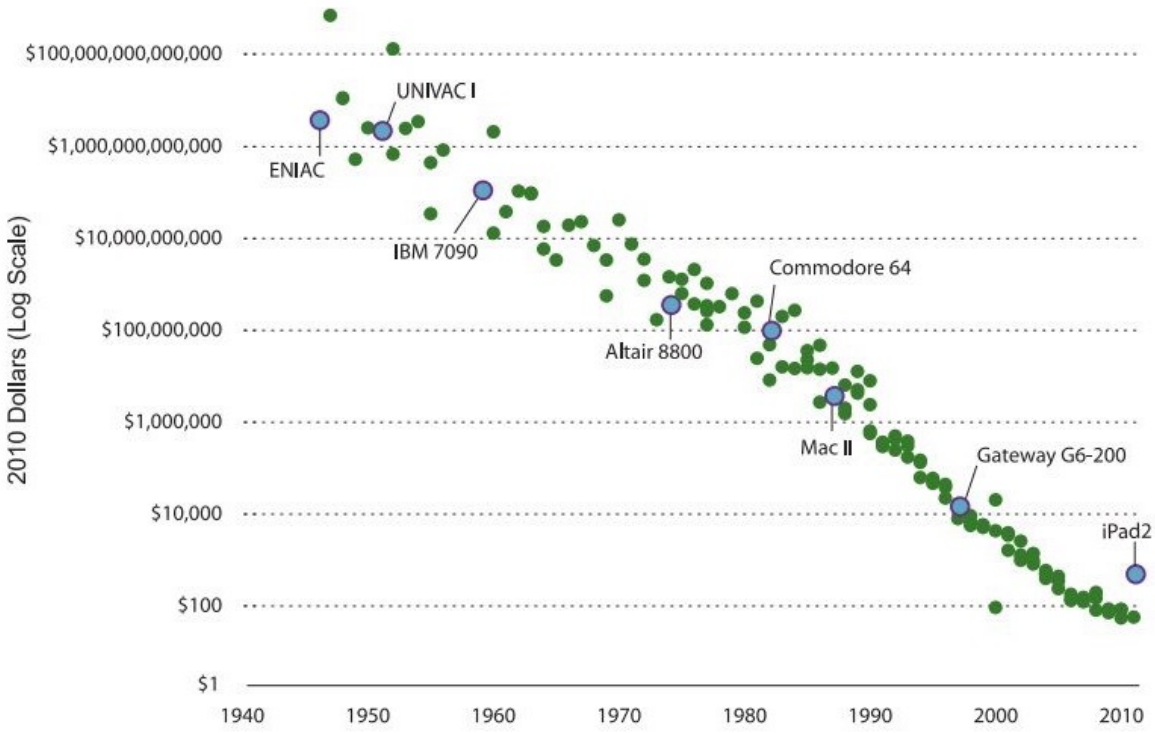


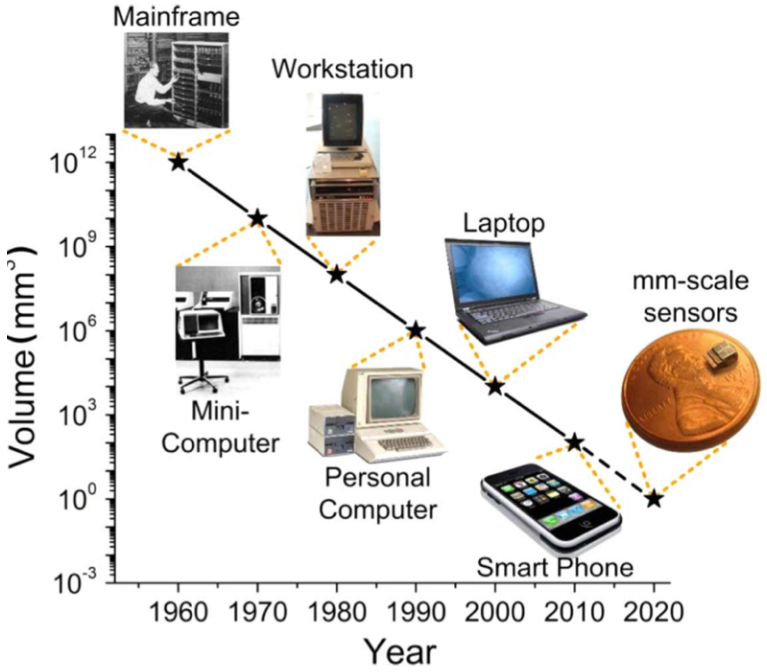
- 1951: Se diseña el primer computador electrónico con fines comerciales, UNIVAC I.



Costo del cómputo

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





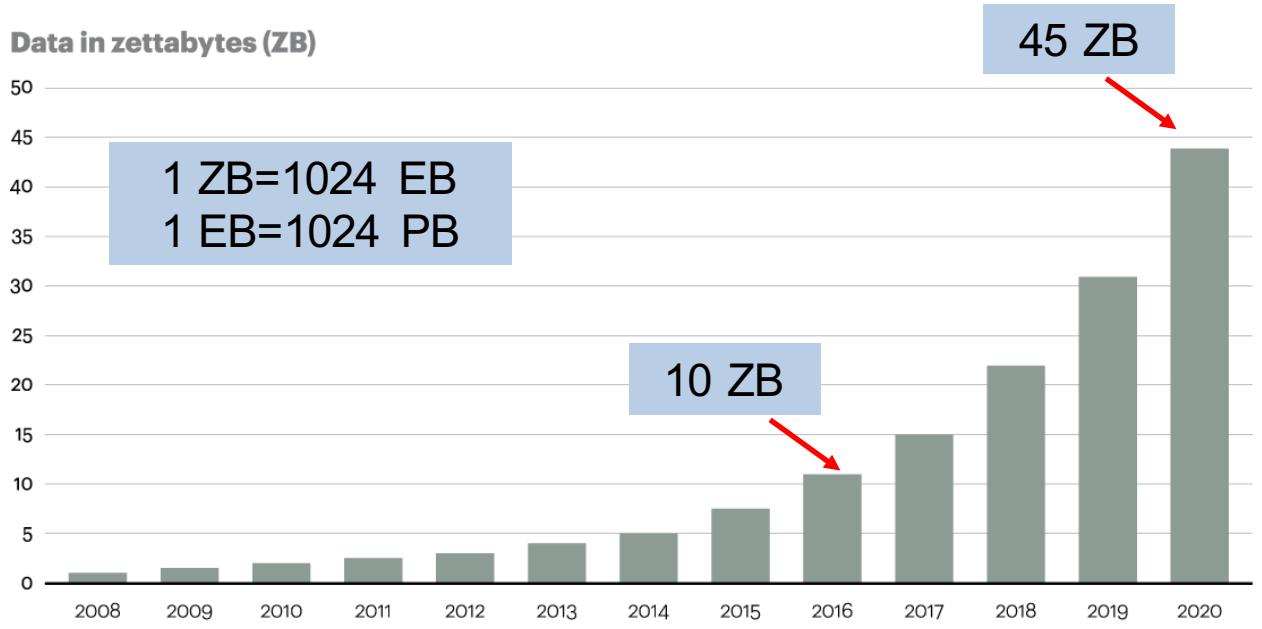
Tsunami de datos

- 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.

GitHub

kaggle

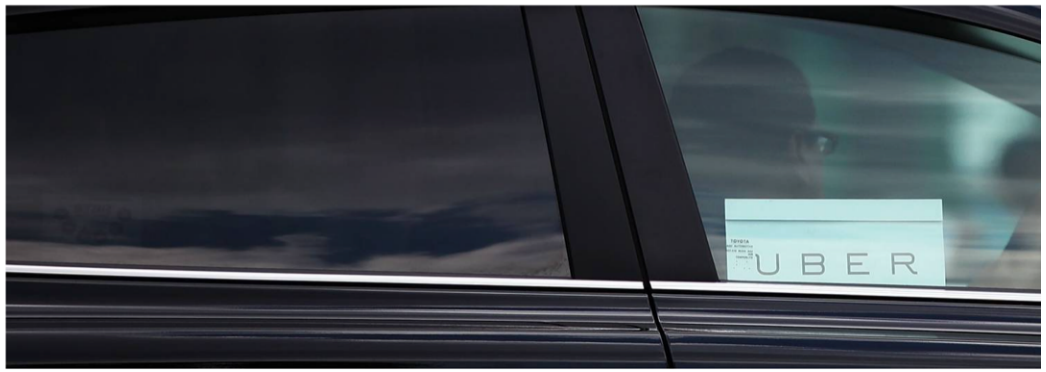
Casos notables

Análisis de uso de Taxis y Uber en NYC

Open Data+Open Source

nyc-taxi-data

uber-tlc-foil-response

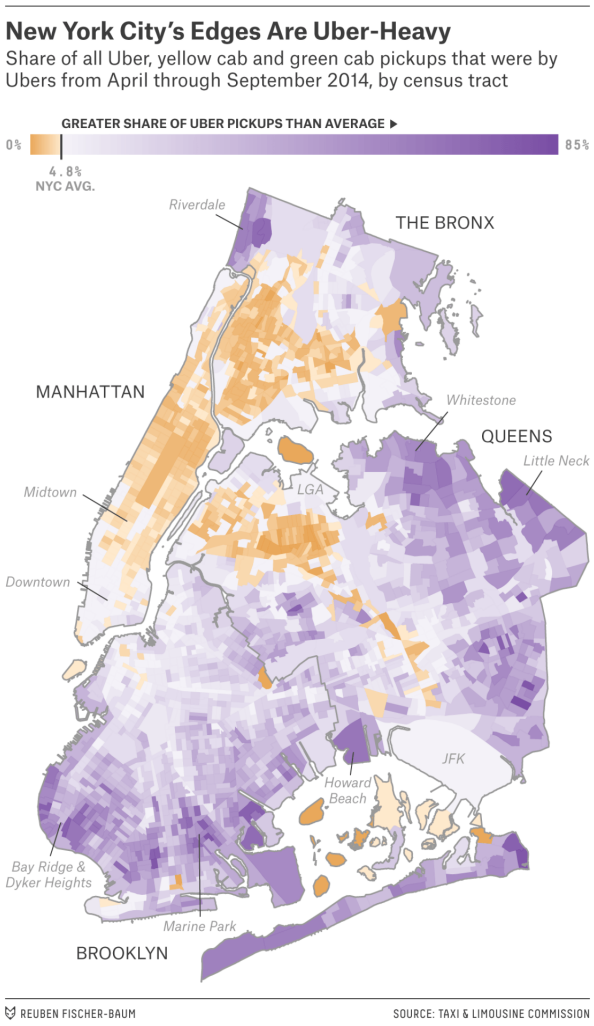


An Uber car. SPENCER PLATT / GETTY IMAGES

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

¿Qué es el Analytics?

- Analytics es entendido como el uso intensivo de datos, estadística y análisis cuantitativo, modelos predictivos y explicativos y gestión basada en hechos para dar soporte al proceso de toma de decisiones, la creación de ventajas competitivas y la generación de valor en las organizaciones.

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.

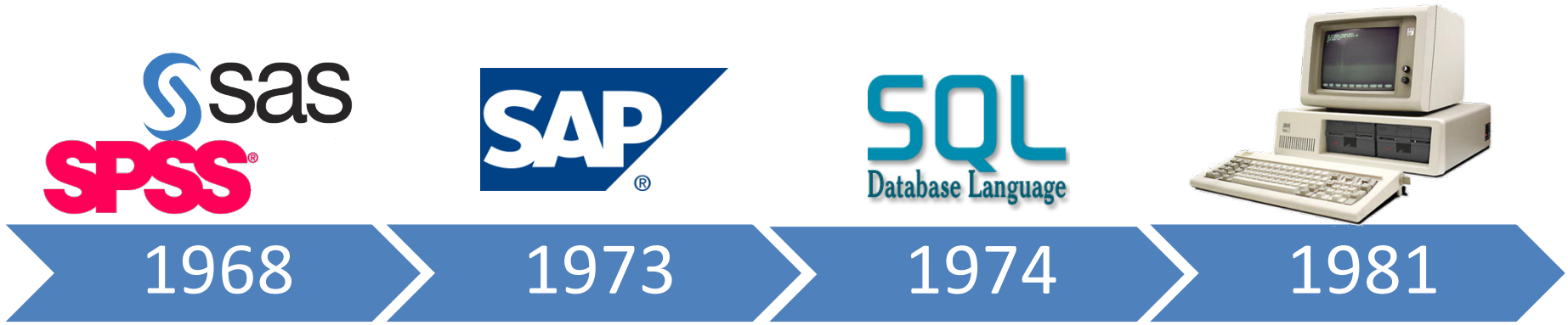
BD

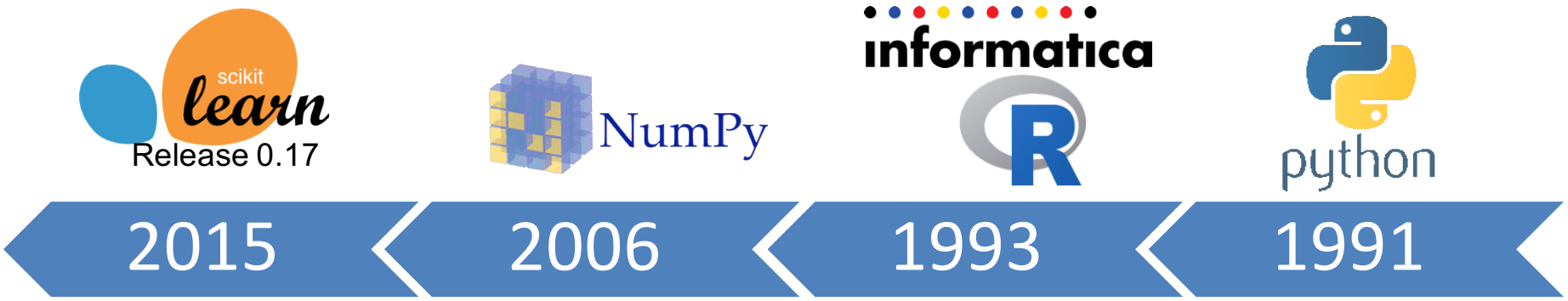
Año	PIB ($ Millones)	Consumo Eléctrico (GWh)
1993	32.559.292	21.011,3
1994	34.416.724	22.730,7
1995	38.028.591	24.910,2
1996	40.831.596	27.969,0
1997	43.526.546	30.351,5
1998	44.944.340	33.015,8
1999	44.616.349	35.921,3
2000	46.605.199	38.867,4

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.

Evolución del Analytics





Software comercial

- Los dos software más usados por las empresas en Chile y el mundo son SAS y SPSS
  - SAS (Statistical Analysis System) fue desarrollado en la Universidad de North Carolina (EE.UU.) y fue planteado originalmente para analizar grandes cantidades de datos agrícolas.
  - SPSS (Statistical Package for the Social Sciences) fue desarrollado en la Universidad de Stanford (EE.UU.) y fue planteado para analizar datos en las Ciencias sociales.

Lenguajes abiertos

- Los dos lenguajes más usados por las empresas y comunidad open source son R y Python

  - R fue desarrollado por investigadores en la Universidad de Auckland (Nueva Zelanda) y es de código abierto bajo licencia GNU GPL v2

  - Python fue desarrollado por el programador holandés Guido van Rossum y es de código abierto bajo licencia de la Python Software Foundation

Comparación

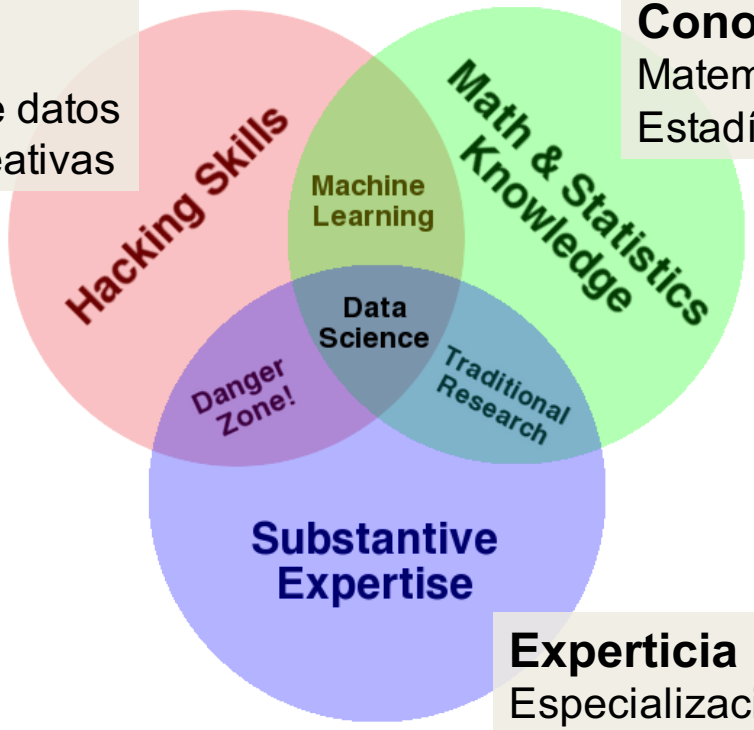
	SAS	SPSS	R	Python
Advantages	1. High adoption rate in major industries 2. Flow based interface with drag and drop 3. Official support 4. Handling large datasets 5. 'PROC SQL'	1. Used a lot in universities 2. Good user interface with extensive documentation 3. Click & Play functionality 4. Writing code made easy using the ´paste´ button. 5. Official support	1. Big community who creates libraries 2. Free 3. Early adopter in explanatory and predictive modeling. 4. Easy to connect to data sources, including NoSQL and webscraping.	1. Scalability 2. General purpose language 3. Easy to learn 4. Good in machine learning 5. Big community 6. Free
Disadvantages	1. Relatively high cost 2. For not-standard options not in interface, you'll need to write the code 3. Slow adapting to new techniques 4. Different programs for visualization or Data Mining	1. Relatively high cost 2. different licenses for different functionalities. 3. Syntax limited 4. Slow adapting to new techniques 5. Slow in handling large datasets	1. Can be slow with big datasets 2. Steep learning curve 3. No official support 4. No user interface	1. Not as strong in explanatory modeling 2. Choice of version: 2.7 or 3.5? 3. No user interface 4. No official support

¿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.

Carácter interdisciplinario

Habilidades
Programación
Exploración de datos
Soluciones creativas



Conocimientos
Matemáticas
Estadística

Experticia
Especialización
Conocimiento de campo

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.

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 estadísitico vienen incorporados.

- Python es más orientado a objetos y deben cargarse módulos para hacer análisis.