



# Data Analytics y Data Science

**Roberto Muñoz**

Doctor en Astrofísica

Director Centro I+D MetricArts

 @RobertoKPax

**METRICARTS**

# Evolución procesamiento de datos

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

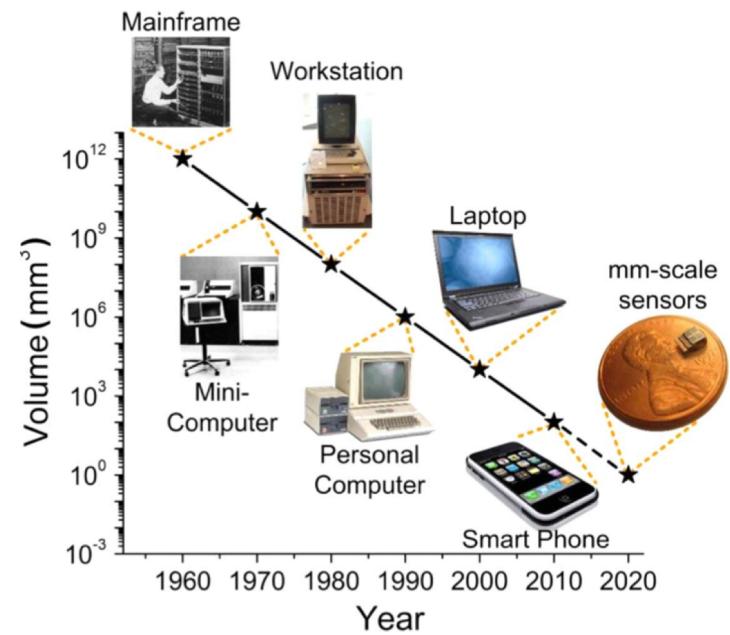
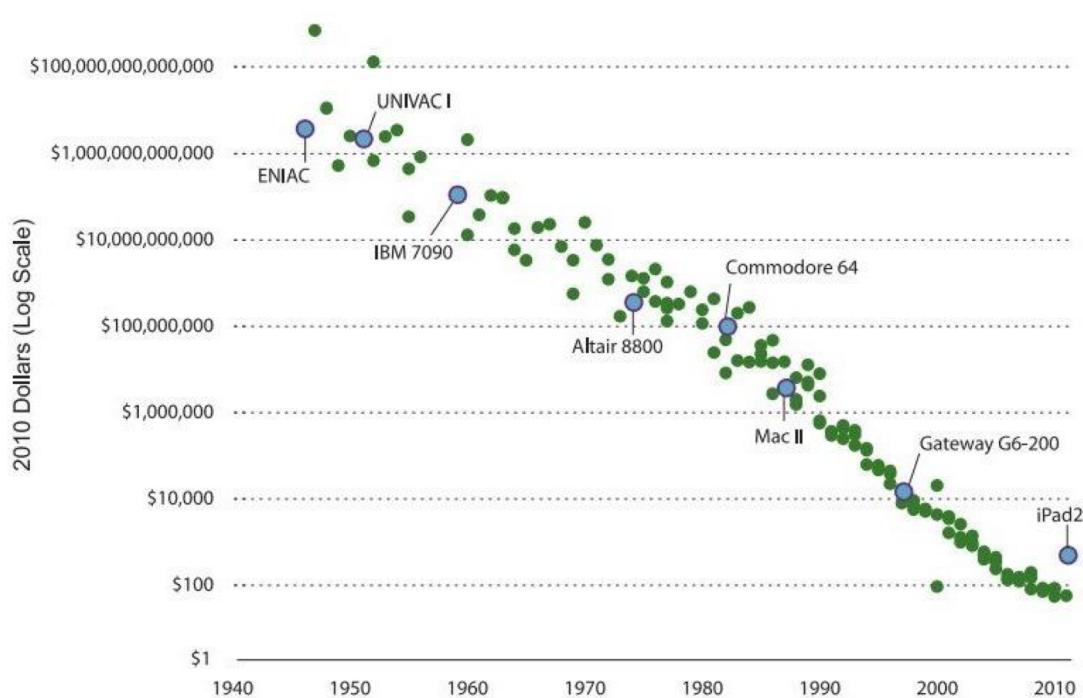


- 1951: Se diseña el primer computador electrónico con fines comerciales, UNIVAC I.



# Costo del cómputo

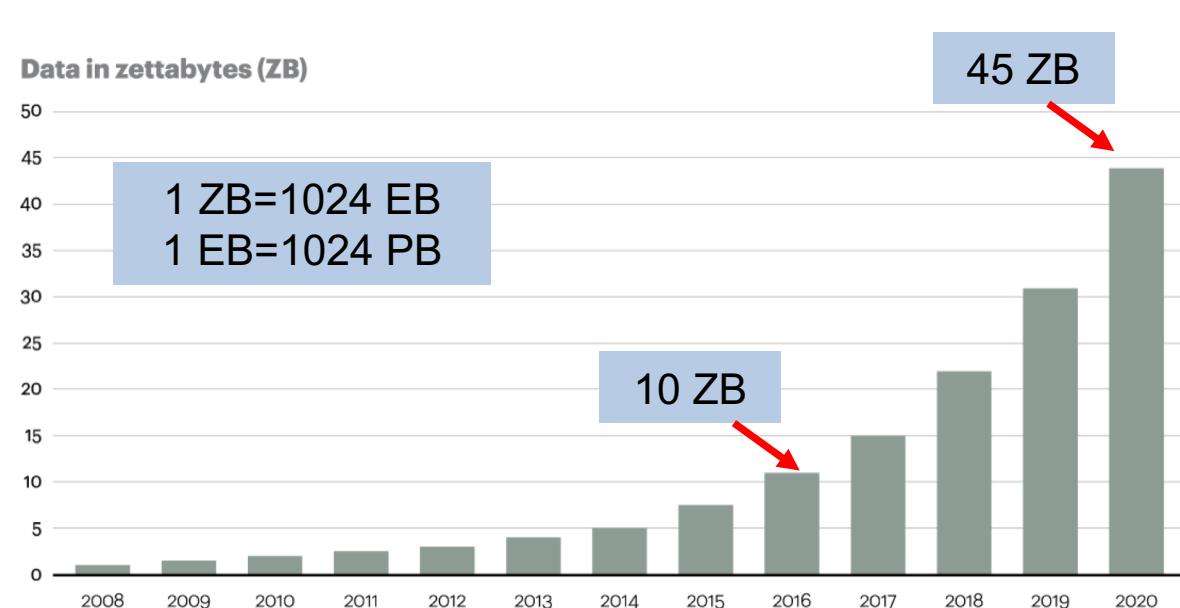
- Desde la invención de los computadores electrónicos, tanto el precio como el tamaño han disminuido sostenidamente.



# 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.



# 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.

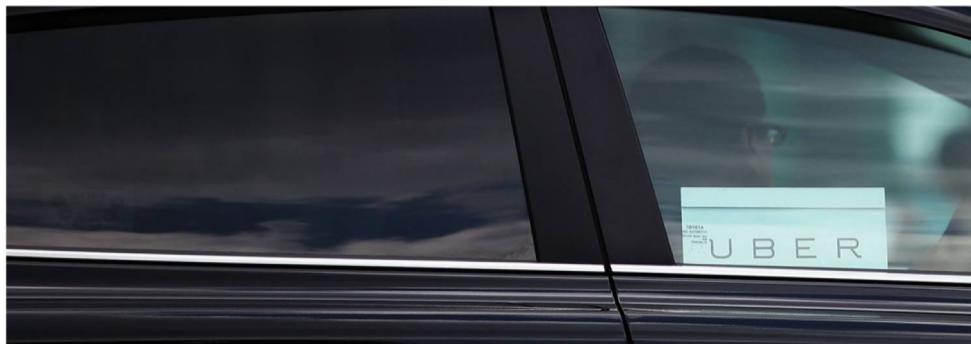


# Casos notables

## 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

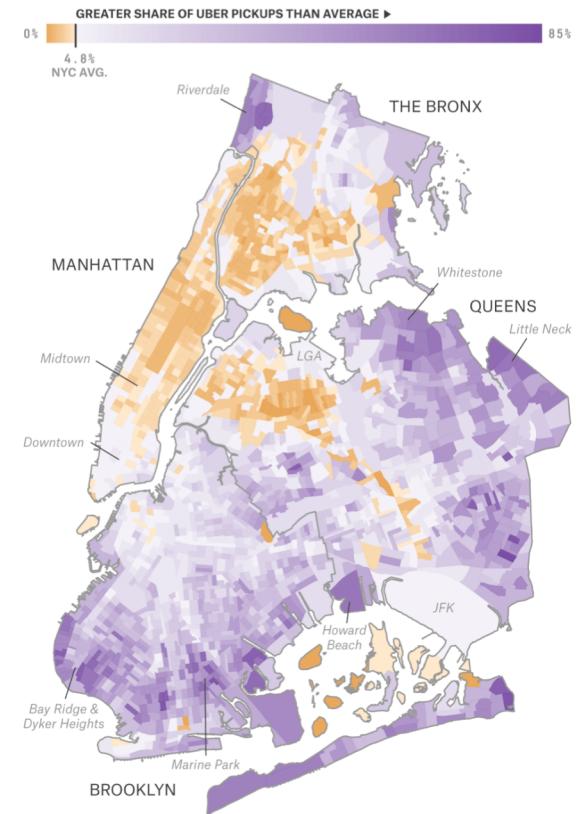
### Uber Is Serving New York's Outer Boroughs More Than Taxis Are

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

# ¿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

## ¿Qué es el Data Science?

Visitar siguiente link

<https://pollev.com/robertomunoz211>

# ¿Qué es el Data Science?

Es una disciplina que se ocupa de los sistemas necesarios para analizar grandes volúmenes de datos

Es una disciplina que se ocupa de los procesos necesarios para explorar y visualizar datos

Es un campo multidisciplinario que se ocupa de los procesos y sistemas usados en la extracción de conocimiento a partir de datos

Ninguna de las anteriores

## ¿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.



# Data Analytics vs Data Science

The world is generating data at a higher rate, and so the need of "Data Science" & "Data Analytics" tools increases to analyze and manage this "Big Data".

**Data  
Science**

**Vs**

**Data  
Analytics**

**Vs**

**Big  
Data**

## WHAT IS DATA SCIENCE?

**Data Science** is a field that refers to the collective processes, theories, concepts, tools and technologies that enable the review, analysis and extraction of valuable knowledge and information from raw data.

## WHAT IS DATA ANALYTICS?

**Data Analytics (DA)** is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems & software.

## WHAT IS BIG DATA?

**Big Data** refers to voluminous amounts of structured or unstructured data that organizations can potentially mine & analyze for business gains.

## APPLICATION AREAS

1. Digital advertisements
2. Internet Research
3. Recommender System
4. Image/Speech Recognition

1. Gaming
2. Travel
3. Energy Management
4. Healthcare

1. Communication
2. Retail
3. Financial services
4. Education

## TOOLS & LANGUAGES

1. Python
2. SAS
3. SQL

1. R
2. Tableau Public
3. Apache Spark

1. Hadoop
2. NoSQL
3. Hive

# Carácter interdisciplinario

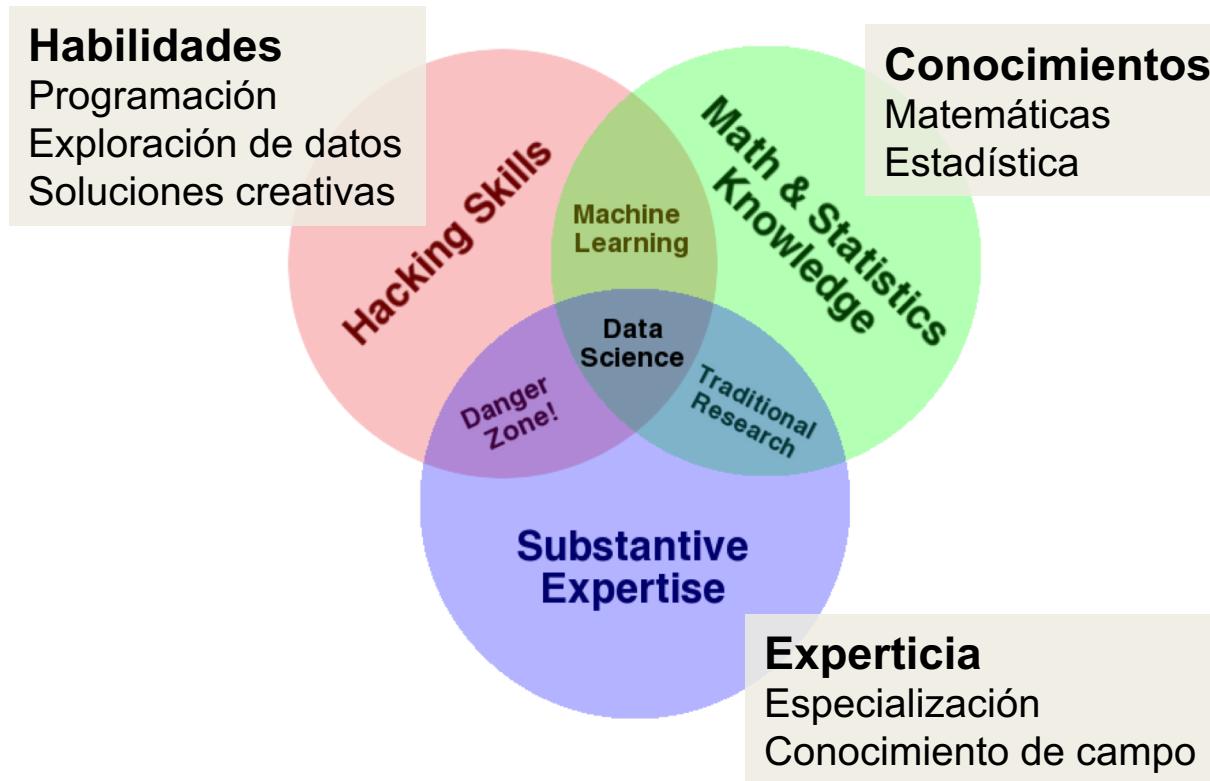


Diagrama de Venn para Data Science  
Drew Conway (2010)

## ¿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.

# Infraestructura y Analítica



Google  
Cloud Platform



Microsoft  
Azure

