

What is Data Science?

(a personal view)

Jordi Vitrià, PhD
Universitat de Barcelona

Data Science
Big Data

Taking (big)data-based decisions is not new but now it is easier.

Sir William Davenant
@SirWilliamD

The world before computers - staff sorting 4M used tickets from #London Underground to analyse line use in 1939.

Respon Retuitar Marca com a preferit Pocket ... Més



RETUITS 105 PREFERITS 49

8:50 - 8 ag. 2014 Marca contingut

Old Pics Archive
@oldpicsarchive

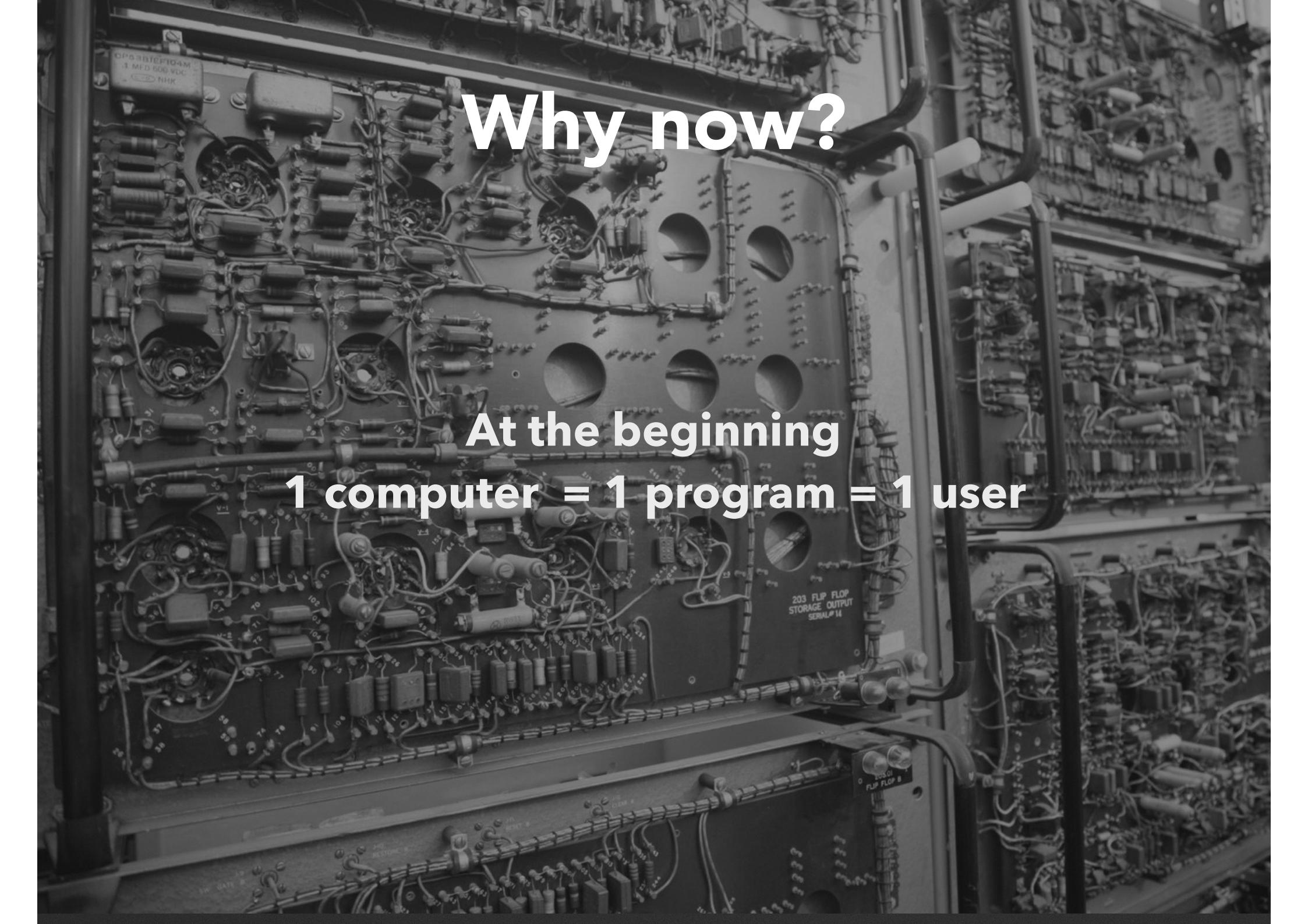
Computing Division at the Department of the Treasury, mid 1920s

...
COMPUTING DIVISION COMPUTING SECTION



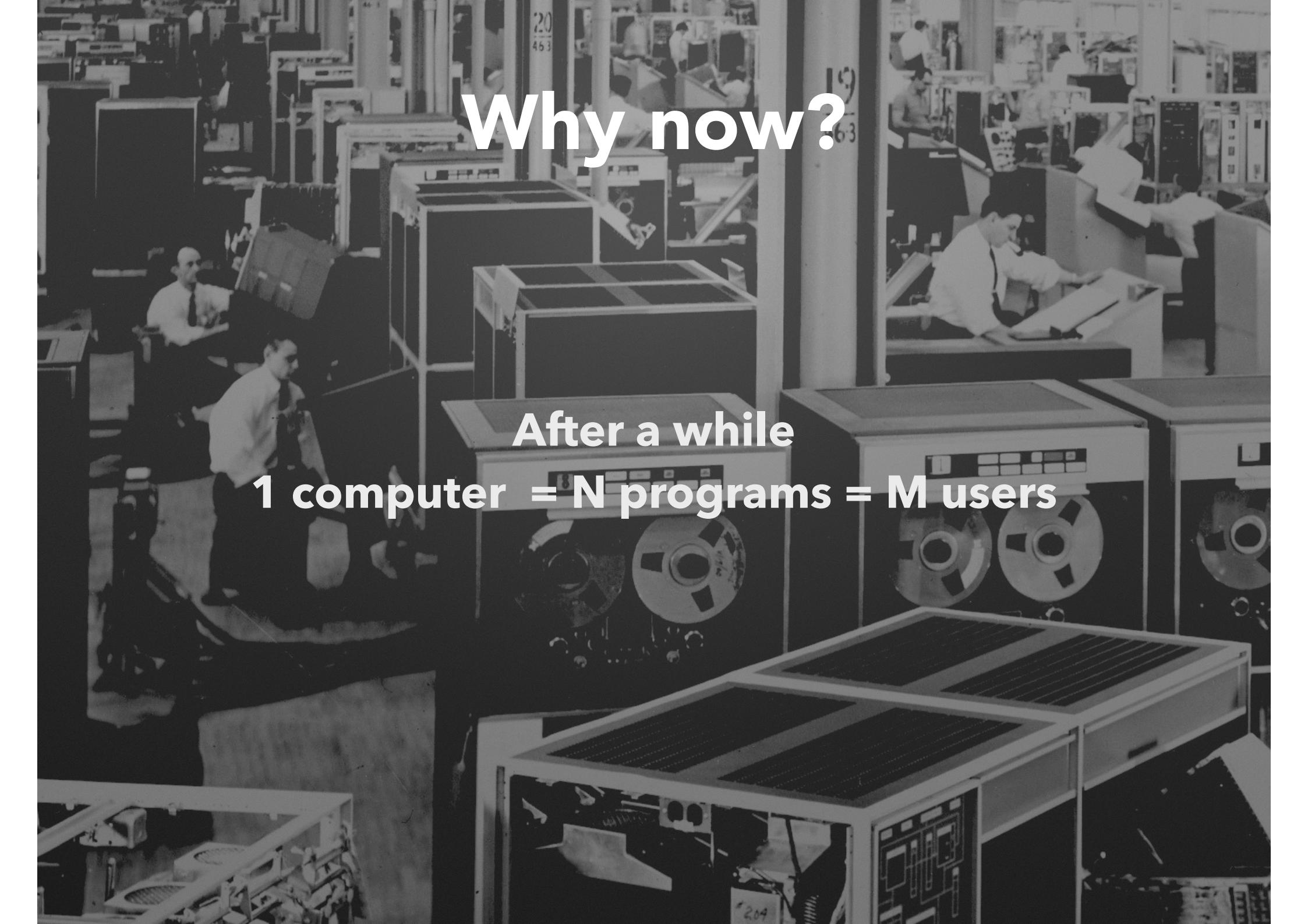
RETUITS 264 PREFERITS 152

21:49 - 20 set. 2014



Why now?

At the beginning
1 computer = 1 program = 1 user

A collage of black and white photographs from the 1960s. It includes images of large mainframe computers with multiple tape drives, people working at terminals in control rooms, and close-up views of computer components like circuit boards and power supplies.

Why now?

After a while
1 computer = N programs = M users



Why now?

Then

1 computer = N programs = 1 user

Why now?

**Meanwhile
Internet & the Web**

Why now?

**A few years ago we reach the present situation.
From a user perspective:**

M computers = N programs = 1 user

Why now?

From a “dev-ops” perspective we are implementing “the network is the computer” idea:

$$2^N \text{ computers} = 2^M \text{ programs} = 2^P \text{ users}$$

Why now?

The “cloud” is a necessary condition to process big data, but not the main cause of the Big Data fever.

Big Data

What is Big Data?

- **For some people, they have big data when its size > 65536 x 256.**
- **In general we have big data when its size does not allow its storage and analysis in a big computer.**

More common

Fat Data

Big Data

Less common

Big Data

Wal-Mart handles over one million customer transaction per hour, the information is stored on a database sized in excess of 2.5 Petabytes (2.0×10^{16} bits).

By 2016 it is likely that a typical hospital will create 665 terabytes (5.32×10^{15} bits) of data a year.

Big Data

With a personal computer:

- You can find an element in a 1 MB file in less than a second.
- You can find an element in a 1 GB file in less than a minute.
- You can find an element in a 1 TB file in less than sixteen hours.
- You can find an element in a 1 PB file in less than two years.
- You can find an element in a 1 EB file in less than two thousand years.

Big Data

Big data is more than size.

It is commonly characterized with four

V:

Volume

Velocity

Variety

Veracity

Big Data

The cloud is key to deal with the three V, but the main phenomenon behind Big Data is **datification**.

Key enabler

The three V are a consequence of it.

Big Data

We are rendering into data many aspects
of the world that have never been
quantified before:

business networks

books I'm reading

location

physical activity

consumed food

purchases

physiological signals

straight thoughts

friendship

gaze

driving behavior

Big Data

Information comes from:

- Corporate Data Bases (structured information).
- Unstructured information in documents, Wikipedia, textbooks, journals, blogs, tweets, etc.
- Images in the web, public cameras, phones, TV, YouTube, etc.
- Public APIs: smart cities, government, search engines, etc.
- Sensor Data: GPS, accelerometer, physico-chemical sensors, sociometric sensors, super-colliders, telescopes, etc.

Big Data

There are several Big Data flavors:

- Big multidimensional arrays (homogeneous data).
- Big tables (structured data).
- Big text.
- Big image.
- Big sound.
- Big sequential data (sensors, tweets, etc.)

Big Data

There are several problems:

- ETL (Extract, Transform, Load)
- BI/Analytics (Think you can do in SQL)
- **Advanced Analytics.**
- **Machine Learning.**
- Visualization.

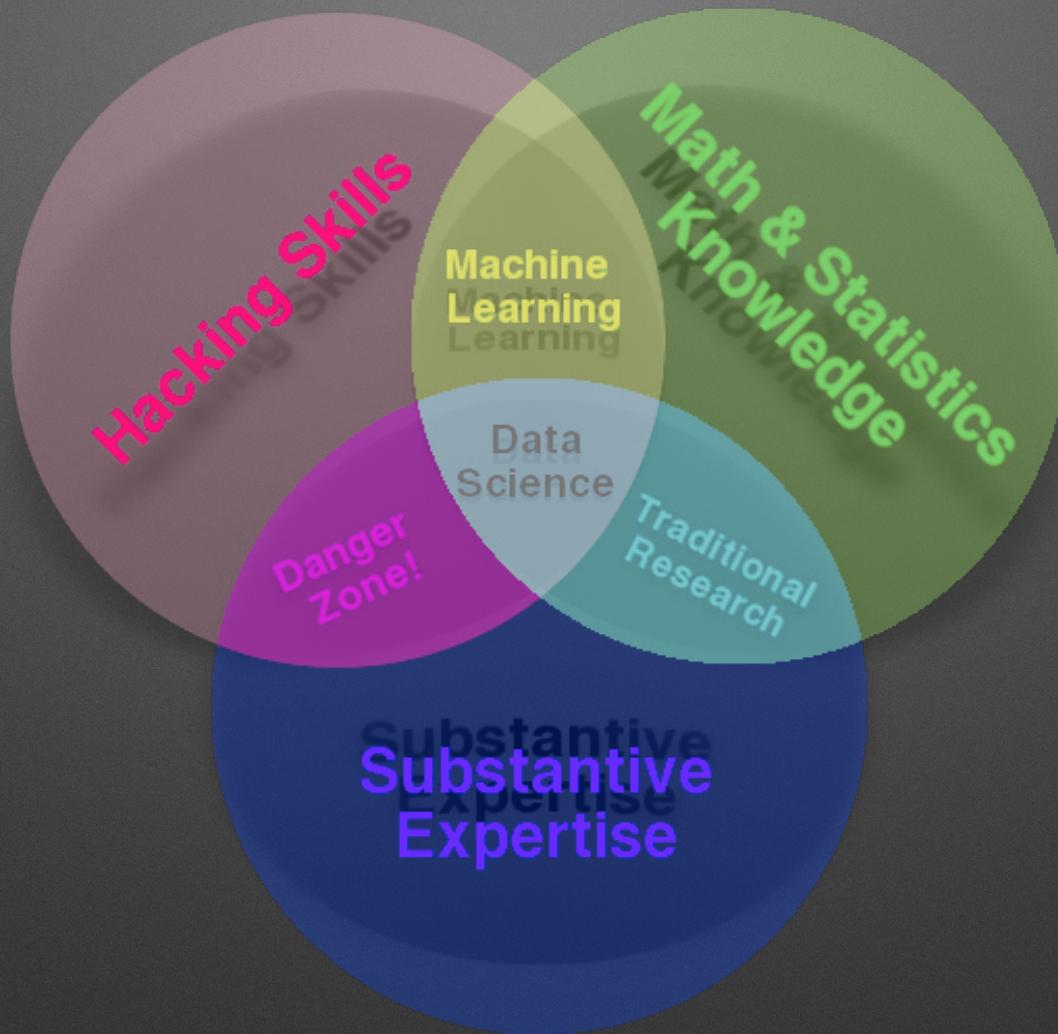
Data Science

Technology is the collection of tools, including machinery, modifications, arrangements and procedures used by humans.

Big Data is a key **technology** to process massive amounts of data (f.e. to count items).

Methodology is the systematic, theoretical analysis of the methods applied to a field of study.

Data Science is a **methodology** to define what we want to do with data, how do we evaluate our actions, what decisions can be grounded on data, how do we combine evidences from several sources, etc.



Drew Conway's Data Science Venn Diagram

Data Science

Data Science is **not** a **science** but a methodology based on multidisciplinar knowledge.

Currently, most company decisions are based on intuition and best practices. The alternative is to integrate data-based knowledge in the decision process.

Data Science is a new data processing model focused on turning data into actions.

Data Science

Steps:

- Ask a question.
- Get the data. They can be heterogeneous and non structured.
- Data Processing (cleaning, munging, etc.).
- Data Analysis (computer science, linguistics, economy, sociology, etc.).
- Take a decision and act.

Data Science

Data Science is a new job!



Data Science

What are the limits of data science?

- Data science is a tool to inform, not to explain.
- Data science cannot substitute intuition or creativity.

If I had asked people what they wanted,
they would have said faster horses.
Henry Ford.

Data Science

	 COMPANY Mastercard	 INDUSTRY Finance
	 EMPLOYEES 67,000	 TYPE Behavioral Analytics

PURPOSE:

With 1.8 billion customers, MasterCard is in the unique position of being able to analyze the behavior of customers in not only their own stores, but also thousands of other retailers. The company teamed up with Mu Sigma to collect and analyze data on shoppers' behavior, and provide the insights it finds to other retailers in benchmarking reports.

Data Science



COMPANY

Starbucks Coffee



INDUSTRY

Food & Beverage



EMPLOYEES

160,000



TYPE

Behavioral
Analytics

PURPOSE:

Starbucks collects data on its customers' purchasing habits in order to send personalized ads and coupon offers to the consumers' mobile phones. The company also identifies trends indicating whether customers are losing interest in their product and directs offers specifically to those customers in order to regenerate interest.

Data Science

COMPANY		INDUSTRY	TYPE
 Spotify		Entertainment	Customer Segmentation & Behavioral Analytics
EMPLOYEES			
5,000			
PURPOSE: Spotify uses data from user profiles and users' playlists, and historical data on music played to provide recommendations for each user. By combining data from millions of users, Spotify is able to make recommendations even if a particular user doesn't have an extensive history with the site.			

Data Science



COMPANY

Union Pacific
Railroad



INDUSTRY

Transportation



EMPLOYEES

44,000



TYPE

Predictive Support

PURPOSE:

With predictive analytics and tools such as visual sensors and thermometers, Union Pacific can detect imminent problems with railway tracks in order to predict potential derailments days before they would likely occur. So far the sensors have reduced derailments by 75 percent.

Data Science



COMPANY

Coca-Cola Co.



INDUSTRY

Food



EMPLOYEES

146,200



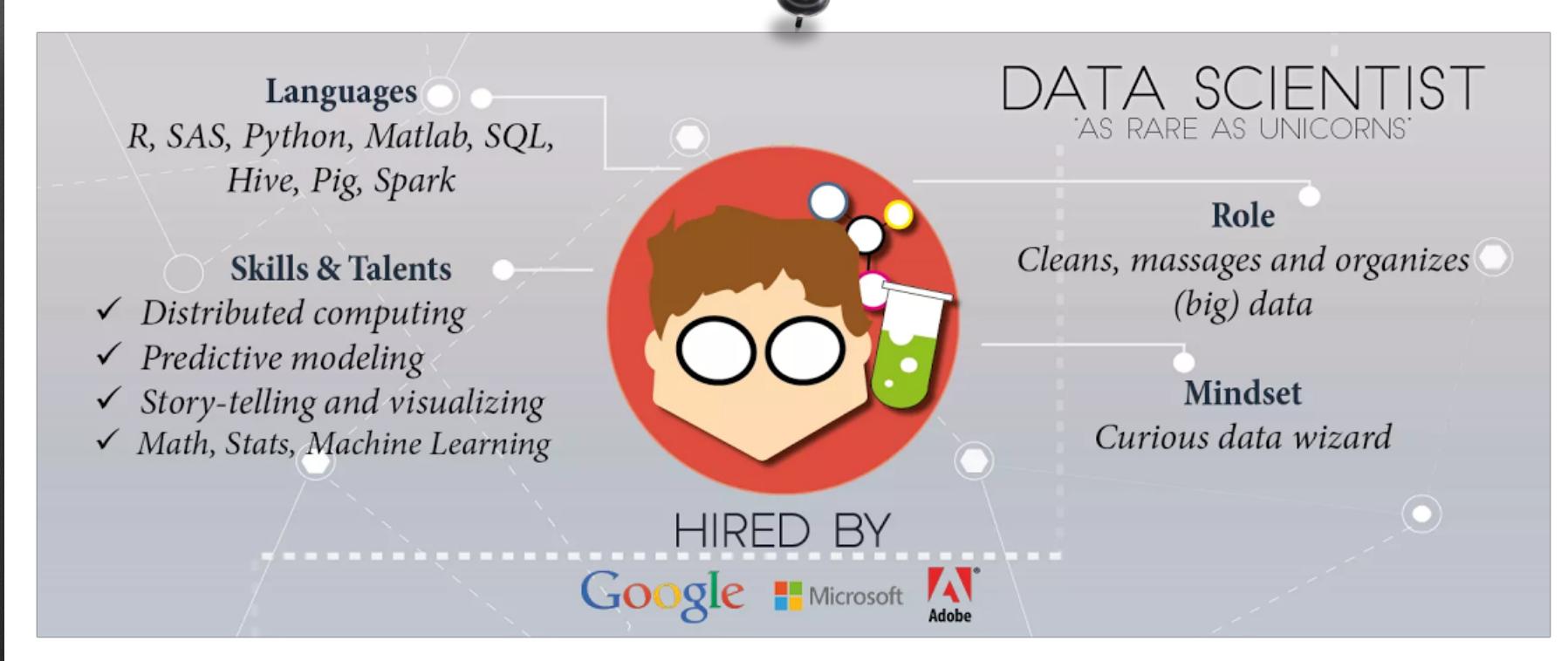
TYPE

Market Basket Analysis

PURPOSE:

Coca-Cola uses an algorithm to ensure that its orange juice has a consistent taste throughout the year. The algorithm incorporates satellite imagery, crop yields, consumer preferences and details about the flavours that make up a particular fruit in order to determine how the juice should be blended.

Data Science Jobs



Data Science Jobs

A graphic pinned to a dark background, resembling a resume or job listing. It features a central circular icon of a person's head with a circuit board pattern inside, set against a red gradient background.

DATA ANALYST
‘DATA DETECTIVE’

Role
Collects, processes and performs statistical data analyses

Mindset
Intuitive data junkie with high “figure-it-out” quotient

Languages
R, Python, HTML, Javascript, C/C++, SQL

Skills & Talents

- ✓ *Spreadsheet tools (e.g. Excel)*
- ✓ *Database systems (SQL and NO SQL based)*
- ✓ *Communication & visualization*
- ✓ *Math, Stats, Machine Learning*

HIRED BY

IBM **hp** **DHL**

Data Science Jobs

A graphic pinned to a dark background, resembling a post-it note. It features a central circular icon of a person wearing a hard hat and safety glasses, with the text "HIRED BY" and logos for Spotify, Facebook, and Amazon below it. The graphic is divided into sections by white dashed lines.

DATA ENGINEER
"SOFTWARE ENGINEERS BY TRADE"

Role
Develops, constructs, tests and maintains architectures (such as databases and large-scale processing systems)

Mindset
All-purpose everyman

Languages
SQL, Hive, Pig, R, Matlab, SAS, SPSS, Python, Java, Ruby, C++, Perl

Skills & Talents

- ✓ Database systems (SQL & NO SQL based)
- ✓ Data modeling & ETL tools
- ✓ Data APIs
- ✓ Data warehousing solutions

Spotify

Data Science Jobs

A graphic of a resume pinned to a dark background with a pushpin at the top center. The resume is white with black text and features a central cartoon illustration of a person with large eyes and a red circle around their head.

DATA ARCHITECT
THE CONTEMPORARY DATA MODELLER[®]

Languages
SQL, XML, Hive, Pig, Spark

Skills & Talents

- ✓ Data warehousing solutions
- ✓ In-depth knowledge of database architecture
- ✓ Extraction Transformation and Load (ETL), spreadsheet and BI tools
- ✓ Data modeling
- ✓ Systems development

HIRED BY
VISA **Coca-Cola** **logitech**

Role:
Creates blueprints for data management systems to integrate, centralize, protect and maintain data sources

Mindset:
Inquiring ninja with a love for data architecture design patterns

Conclusions

- Big Data will be soon a commodity that will be used mainly for data munging and counting at scale.
- The most difficult part of Big Data is getting insight.
- Data Science is a new job with a bright future.