

# 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

Segueix

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



REUTSIS 105 PREFERITS 49

8.50 - 8 ag. 2014 Marca contingut

PIOS 868 - 02.8

REUTSIS 102 PREFERITS 64

PIOS JEE DS - 04.15

REUTSIS 462 PREFERITS 251

Old Pics Archive  
@oldpicsarchive

Segueix

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

Retuitar Marca com a preferit Pocket Més

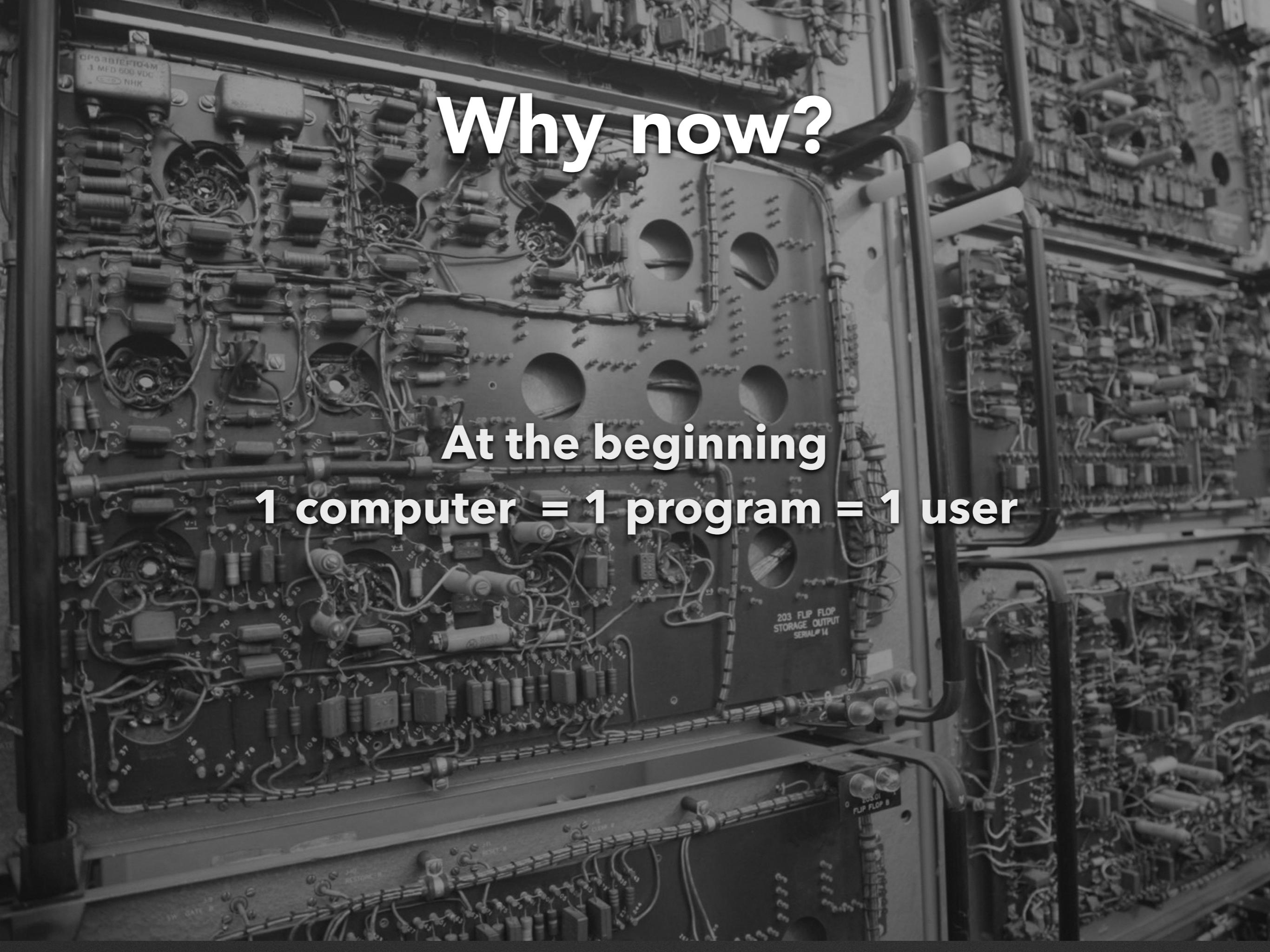


REUTSIS 264 PREFERITS 152

21:49 - 20 set. 2014

PIOS JEE DS - 04.15

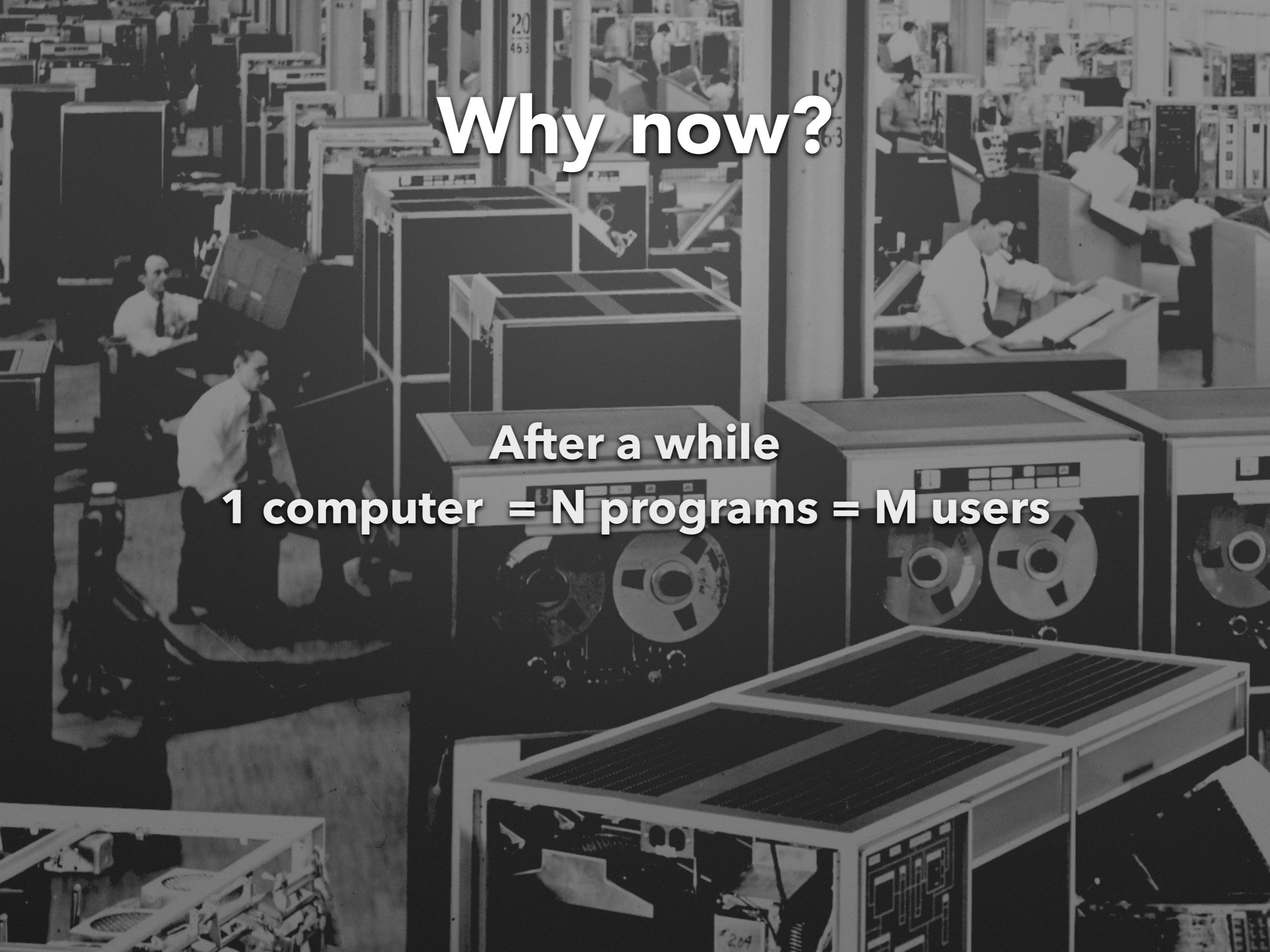
REUTSIS 462 PREFERITS 251



# Why now?

At the beginning

**1 computer = 1 program = 1 user**



# 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:

**$2^N$  computers =  $2^M$  programs =  $2^P$  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  $> 65536 \times 256$ .
- 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 ( $2,0 \times 10^{16}$  bits).**

**By 2016 it is likely that a typical hospital will create 665 terabytes ( $5.32 \times 10^{15}$  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.

# A PETABYTE IS A LOT OF DATA

1	PETABYTE	20 MILLION FOUR-DRAWER FILING CABINETS FILLED WITH TEXT
1	PETABYTE	13.3 YEARS OF HD-TV VIDEO
1.5	PETABYTES	SIZE OF THE 10 BILLION PHOTOS ON → FACEBOOK
20	PETABYTES	THE AMOUNT OF DATA   PER PROCESSED BY GOOGLE   DAY
20	PETABYTES	TOTAL HARD DRIVE SPACE MANUFACTURED IN 1995
50	PETABYTES	THE ENTIRE WRITTEN WORKS OF MANKIND, FROM THE BEGIN- NING OF RECORDED HISTORY, IN ALL LANGUAGES

(all approximate)

# 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.

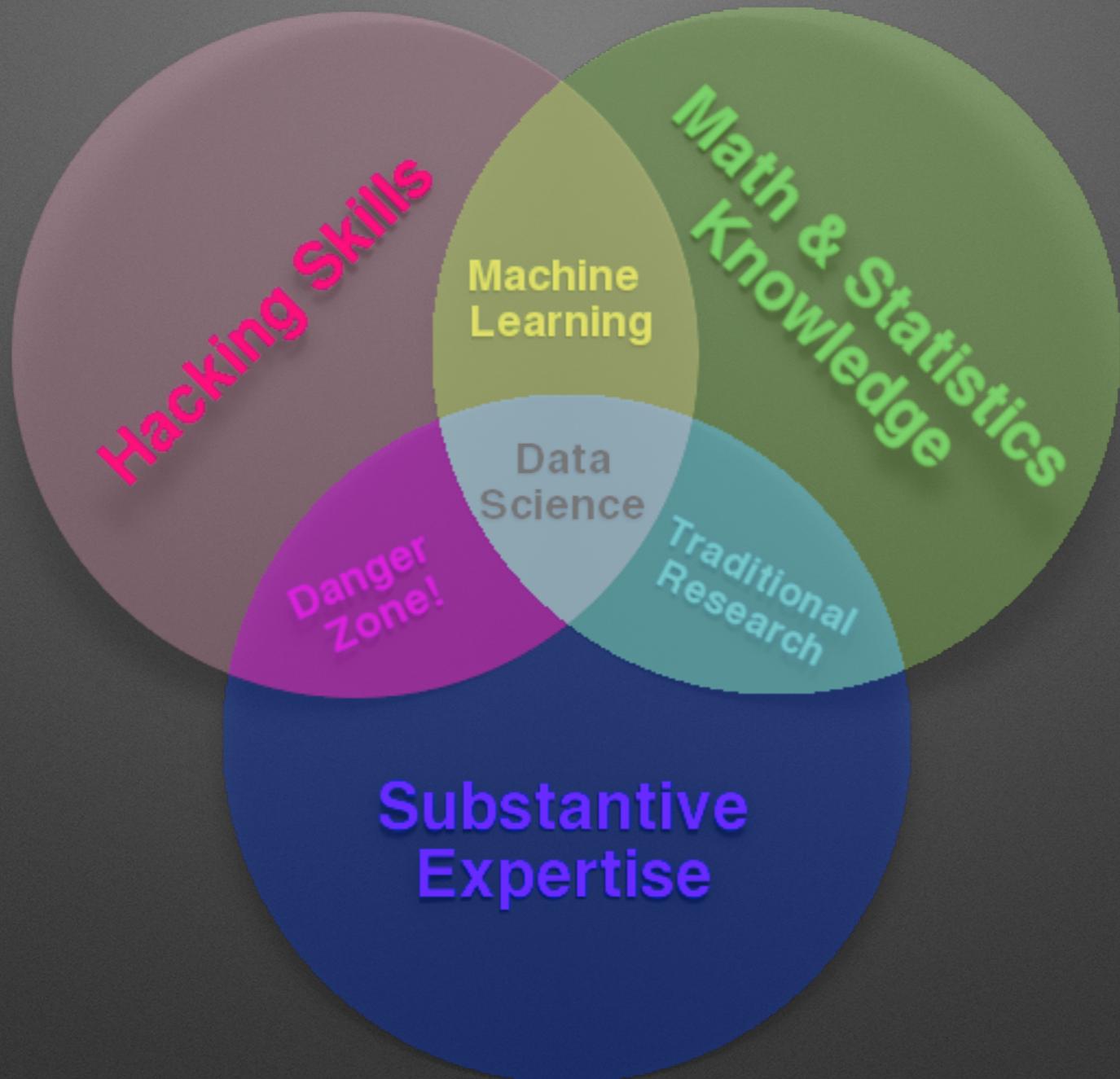
# Data Science

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

<b>Classification</b>	To which category does this data point belong?	Medical diagnosis: does this tissue show signs of disease? Banking: is this transaction fraudulent? Computer vision: what type of object is in this picture? Is it a person? Is it a building?
<b>Regression</b>	Given this input from a dataset, what is the likely value of a particular quantity?	Finance: what is the value of this stock going to be tomorrow? Housing: what would the price of this house be if it were sold today? Food quality: how many days before this strawberry is ripe? Image processing: how old is the person in this photo?
<b>Clustering</b>	Which data points are similar to each other?	E-commerce: which customers are exhibiting similar behaviour to each other, how do they group together? Video Streaming: what are the different types of video genres in our catalogue, and which videos are in the same genre?

# Data Science

Data Science is a new job!

**THE SEXIEST JOB OF  
THE 21TH CENTURY.**  
HARVARD BUSINESS REVIEW,  
OCT. 2012

# 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.

# What are the limits of Data Science

- Data science models reproduce what we do and how we do it (including bad things and wrong strategies).

Rich Caruana gives the example of a **pneumonia risk prediction** model on which he had worked. The purpose of the model was to evaluate whether a patient with **pneumonia** was at high or low risk, to help decide whether or not the patient should be admitted to the hospital. "On the basis of the patient data," says Caruana, "the model had found that patients with a history of **asthma** have a lower risk of dying from pneumonia. In reality, everybody knows that asthma is a very high risk factor for pneumonia. What the model found is the result of the fact that asthma patients get healthcare faster, which lowers their chance of dying compared to the general population."

<https://cacm.acm.org/news/214618-in-black-box-algorithms-we-trust-or-do-we/fulltext>

# Data Science

	<b>COMPANY</b> Mastercard	<b>INDUSTRY</b> Finance
<b>EMPLOYEES</b> 67,000	<b>TYPE</b> 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

 Spotify®	 COMPANY Spotify	 INDUSTRY Entertainment
	 EMPLOYEES 5,000	 TYPE Customer Segmentation & Behavioral Analytics

## 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



<b>COMPANY</b> Union Pacific Railroad	<b>INDUSTRY</b> Transportation
<b>EMPLOYEES</b> 44,000	<b>TYPE</b> 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

	<b>COMPANY</b> Coca-Cola Co.	<b>INDUSTRY</b> Food
<b>EMPLOYEES</b> 146,200	<b>TYPE</b> 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

The screenshot shows the homepage of the bodas.net website. At the top, there is a navigation bar with links for 'Área Empresas', 'ACCEDE', and 'REGISTRATE'. Below the navigation is a large banner featuring a close-up of a person's hands holding a saxophone. The banner has text overlay: 'Encuentra todo lo que necesitas para tu boda' and '¡Tienes más de 45.000 proveedores para elegir!'. Below the banner is a search bar with fields for '¿Qué buscas?' and '¿Dónde?', and a red 'Buscar' button. A list of categories follows: Banquetes, Fotógrafos, Música, Coches de boda, Invitaciones de boda, Tiendas de novia, Floristerías, Trajes novio. To the right of the banner, there is a photo credit: 'Antoni & Rebeca (Burriana)' and 'Toni Vida Fotógrafo's'. Below the banner, there are five icons with corresponding text: 'Encuentra tus proveedores', 'Gestiona tu lista de invitados', 'Comunidad de novi@s', 'Crea tu web de boda gratis', and 'Comparte tu lista de boda'. To the right of these icons is a call-to-action: '...y mucho más con tu organizador de boda' and a red 'Empezar' button. At the bottom, there is a section titled 'Bodas reales' with a sub-instruction: 'Inspírate en las bodas de otros novios y si te gustan contacta con los proveedores que las organizaron'. Below this are four small wedding photos. A promotional overlay on the left says 'Gana 5.000€ PARA TU BODA'. Another overlay on the right says 'Participa en el sorteo mensual de un cheque para los preparativos de tu boda' and 'Y si te casaste en 2018 ¡también puedes!' with a red 'Participa' button.

# Data Science



# Data Science



**Data Science is for all,  
small and big, old and new, etc.**



Swimming companies

Walking companies

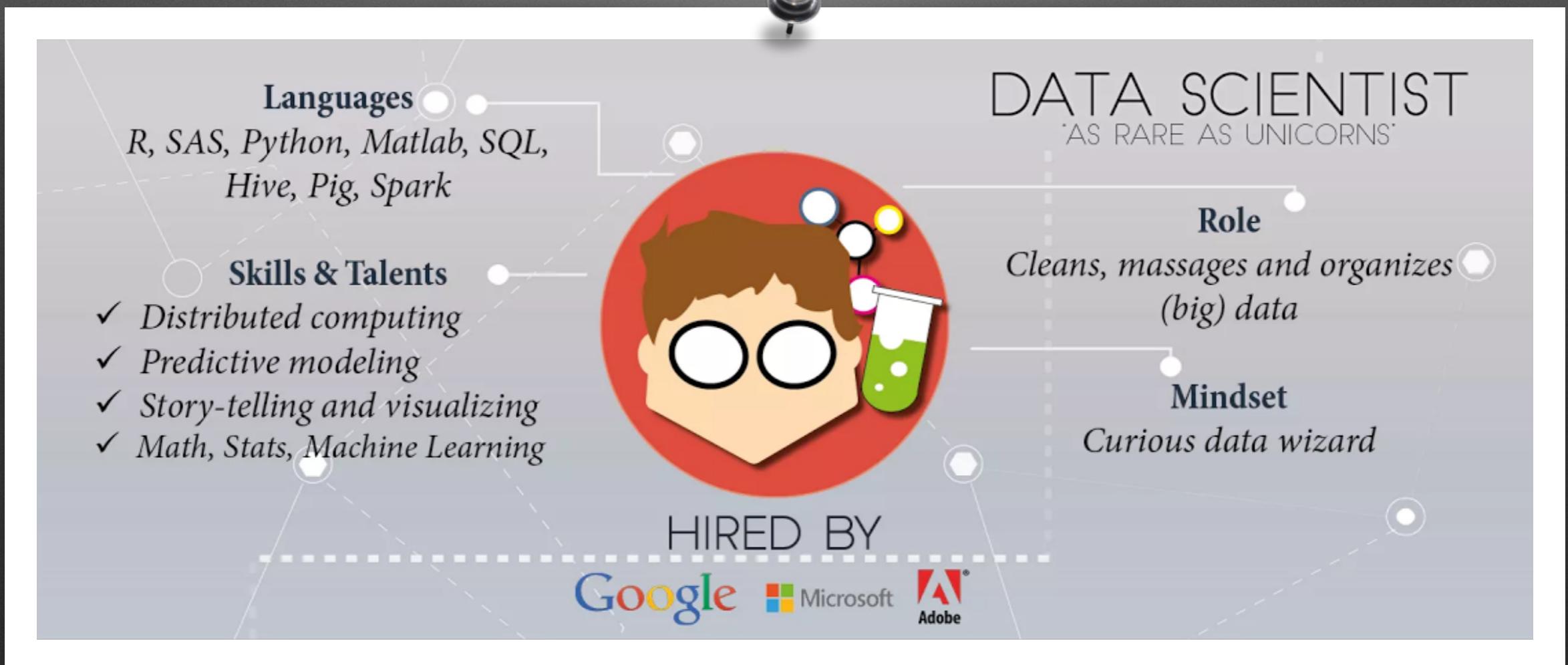


Running companies

**Data Science is for all,  
small and big, old and new, etc.**

All these companies can be better by  
knowing better their **customers**,  
improving by their operational **processes**  
and even by creating new **business**  
**models** with data products.

# Data Science Jobs



# Data Science Jobs

A graphic pinned to a dark background, featuring a red circular icon of a person with glasses and a brain-like pattern, surrounded by network connections. The text is organized into sections:

- 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, featuring a red circular icon of a person wearing a hard hat and glasses. The text around the icon describes the role, mindset, languages, and skills required for a Data Engineer.

**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

HIRED BY

Spotify

# Data Science Jobs

A pinned-up infographic on a dark background. The title 'DATA ARCHITECT' is at the top right, with the subtitle 'THE CONTEMPORARY DATA MODELLER' below it. In the center is a circular cartoon character with large white eyes and a small mouth, wearing a red hoodie. Below the character is the text 'HIRED BY' followed by logos for VISA, Coca-Cola, and logitech.

**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*

**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.