



What is Data Science?

(a personal view: connecting data to reality)

Jordi Vitrià, PhD

Machine Learning

Fat Data

Data Science

Dirty Data

Big Data

Data

Mining

Artificial Intelligence

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

Data Science Path

What do I want?
Does it have sense?

What are my data
sources? How reliable
are they?

How do I develop an
understanding of the
content of my data?

What are the key
relationships in my
data?

How do I develop an
understanding of the
content of my data?

What are the likely
future outcomes?

Are my expectations
fulfilled?

Question

Acquire

Describe

Discover

Analyze

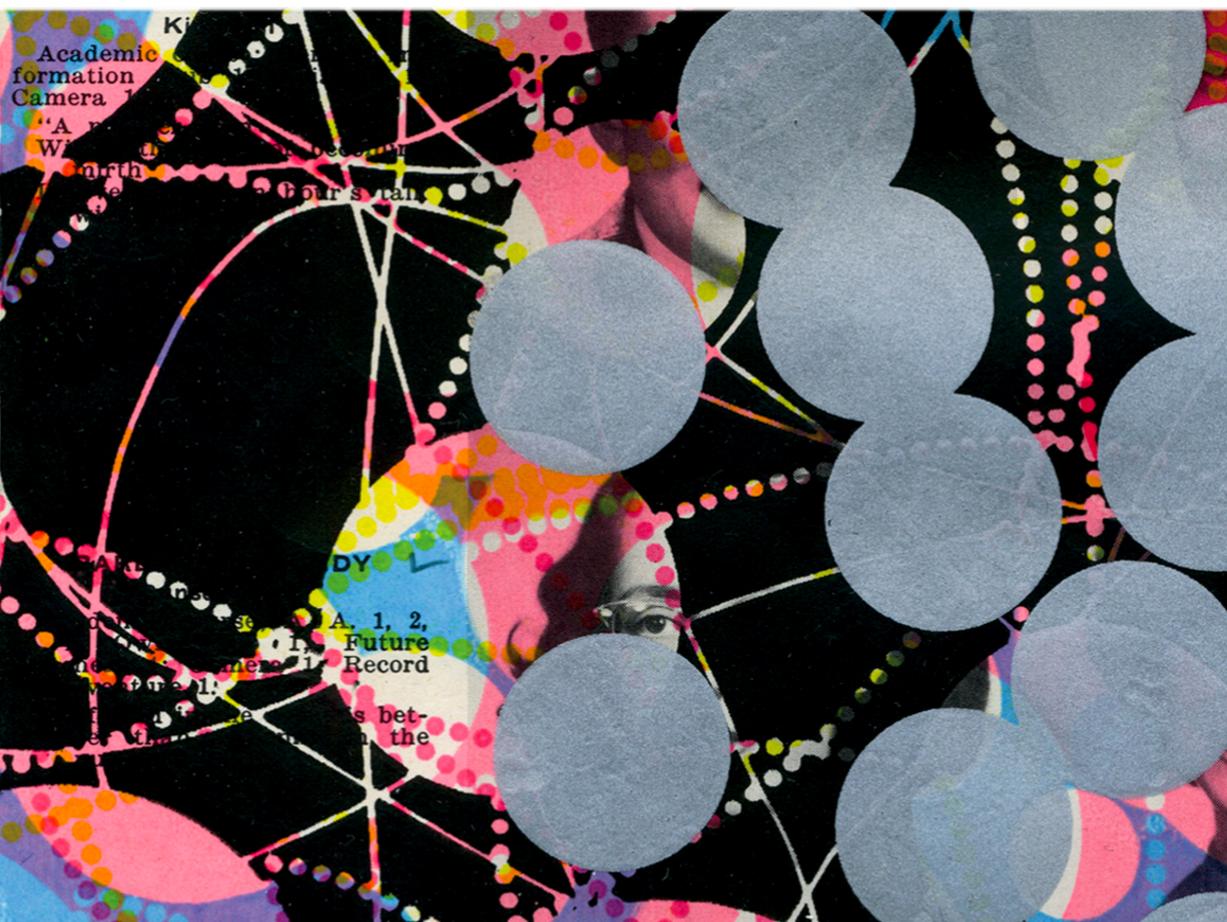
Predict

Evaluate

In this era, where a **huge amount** of information from different fields is gathered and stored, its analysis and the **extraction of value** have become one of the most attractive tasks for companies and society in general. The design of solutions for the new questions emerged from data has required multidisciplinary teams. Computer scientists, statisticians, mathematicians, physicists, journalists and sociologists, as well as many others are now working together in order to provide **knowledge from data**. This new interdisciplinary field is called data science.

Data is only as **valuable** as the questions that it can help answer.

The answers to these questions may result in operational efficiencies, better market sensing, higher quality service to the customer, or nothing at all...



DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

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



REUTS 105 PREFERITS 49

8.50 - 8 ag. 2014 Marca contingut

PIOS 868 - 02.8

REUTS 102 PREFERITS 64

PIOS JES 05 - 04.5

REUTS 462 PREFERITS 25

Old Pics Archive
@oldpicsarchive

Segueix

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

RETUTS 264 PREFERITS 152

21:49 - 20 set. 2014



PIOS JES 05 - 04.5

REUTS 462 PREFERITS 25

Big Data

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.**

10 Megabyte Hard Disk \$3,495*



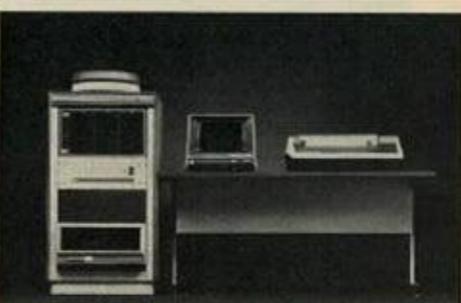
5440-12 Top Load Drive

* Factory rebuilt 10MB cartridge disk drive only.
A new Cameo Data Systems controller is available for \$1,495
\$4,495 for a brand new Ampex 10MB drive only



We are the CP/M** and MP/M** specialist of Southern California. We can supply you with the latest CP/M (\$150) or MP/M (\$300) and with Standard BIOS (\$150) or Custom BIOS (\$300). Immediate delivery worldwide. Domestic and foreign inquiries invited... dealers too.

**CP/M and MP/M are Trademarks of Digital Research



We are a full service computer retailer. We totally integrate hardware and software into high quality, high reliability systems. Systems for use in development, process control and general business. Word processing naturally, multi tasking and multi processing too.

COMPUTER COMPONENTS

Circle 279 on inquiry card.

5848 Sepulveda Boulevard Van Nuys, California 91411 213•786-7411

BYTE July 1980 291

July 1980.

More common

Fat Data

Big Data

Less common



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

With over 20,000 stores in 28 countries, Walmart is the largest retailer in the world. So it's fitting then that the company is in the process of building the world's largest private cloud, big enough to cope with 2.5 petabytes of data every hour. (2.5×10^{16} bits = one million gigabytes).

Big Data

- On average, people send about 500 million tweets per day.
- The average U.S. customer uses 1.8 gigabytes of data per month on his or her cell phone plan.
- Amazon sells 600 items per second.
- On average, each person who uses email receives 88 emails per day and send 34. That adds up to more than 200 billion emails each day.
- MasterCard processes 74 billion transactions per year.

Big Data

Big data is more than size.

It is commonly characterized with several V:

Volume

Velocity

Variety

Big Data

The main phenomenon behind Big Data
is **datification**.

The V's are a consequence of it.

Big Data

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

A grid of colored boxes containing various data points, arranged in four rows. The colors of the boxes are green, red, blue, orange, yellow, and dark green. The text inside the boxes includes:
Row 1: business networks, books I'm reading, location
Row 2: physical activity, consumed food, purchases
Row 3: physiological signals, straight thoughts, friendship
Row 4: gaze, driving behavior

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 problems:

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

Analyzing the past

Predicting the future (predictive analytics)
Evaluating alternative worlds (prescriptive analytics)

Artificial Intelligence and Machine Learning

Artificial intelligence is an academic discipline devoted to the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, language recognition, decision-making, planning, reasoning, etc.

Artificial intelligence is classified into two parts, General AI and Narrow AI. General AI refers to making machines intelligent in a wide array of activities that involve thinking and reasoning. Narrow AI, on the other hand, involves the use of artificial intelligence for a very specific task.

Machine learning is a subset of artificial intelligence that uses algorithms to learn from data (inductive behavior).

Data Science

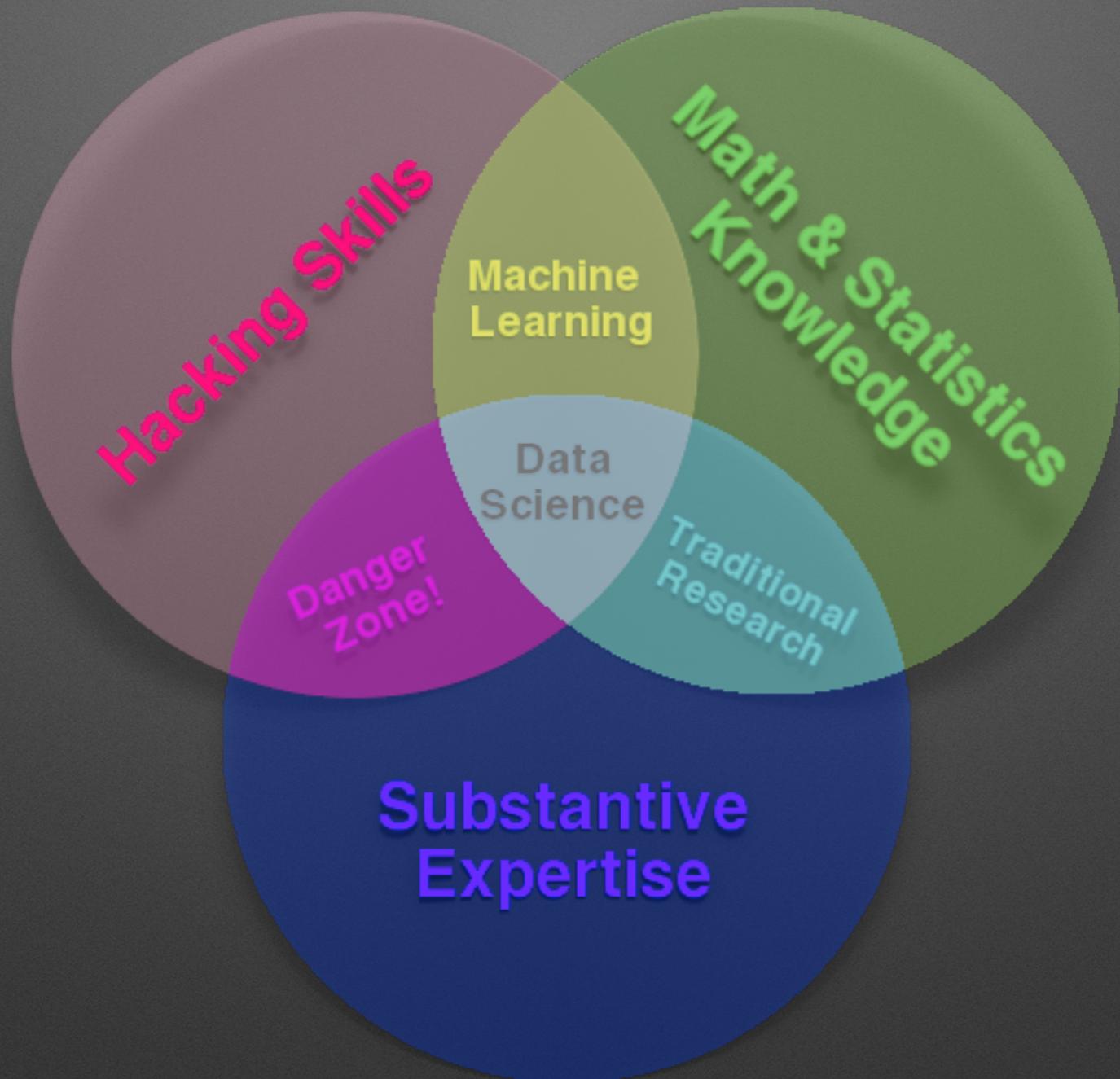
Data Science

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.

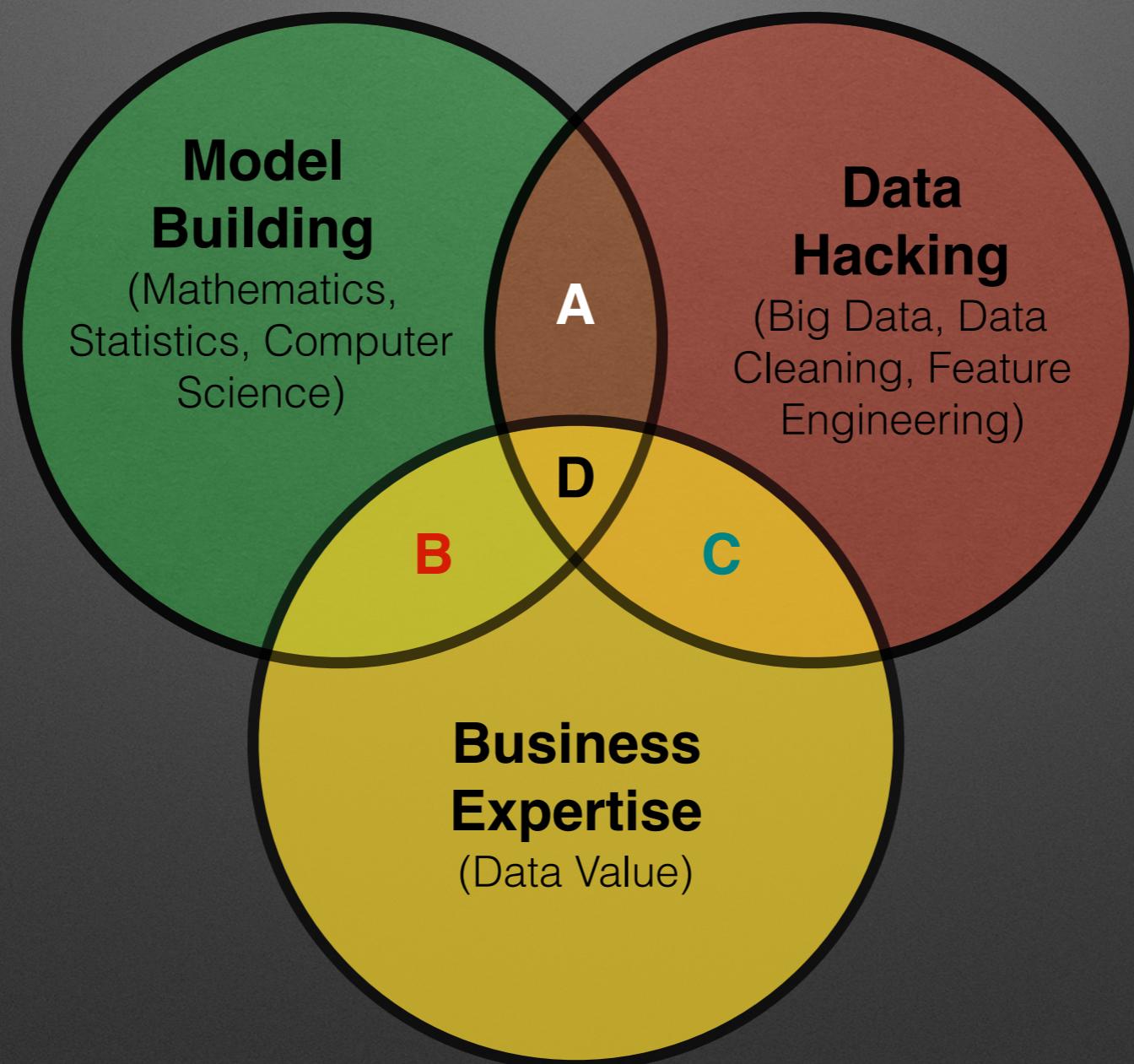
Data Science is a **methodology** to define what we want to do with data, how do we evaluate our actions, what decisions can be grounded on data, how do we combine evidences from several sources, etc.



Drew Conway's Data Science Venn Diagram

D is an empty set!

$$A + B + C = D$$



Data Science Tasks

Background

Domain Knowledge, Causality, Decision Making, Human Behavior

Domain Knowledge, Statistics, Machine Learning, Complex Systems, etc.

Data Processing,
Visualization

Data Processing

Data Engineering

Data Engineering

Output

Prescriptive Decisions:
Why? What is best?

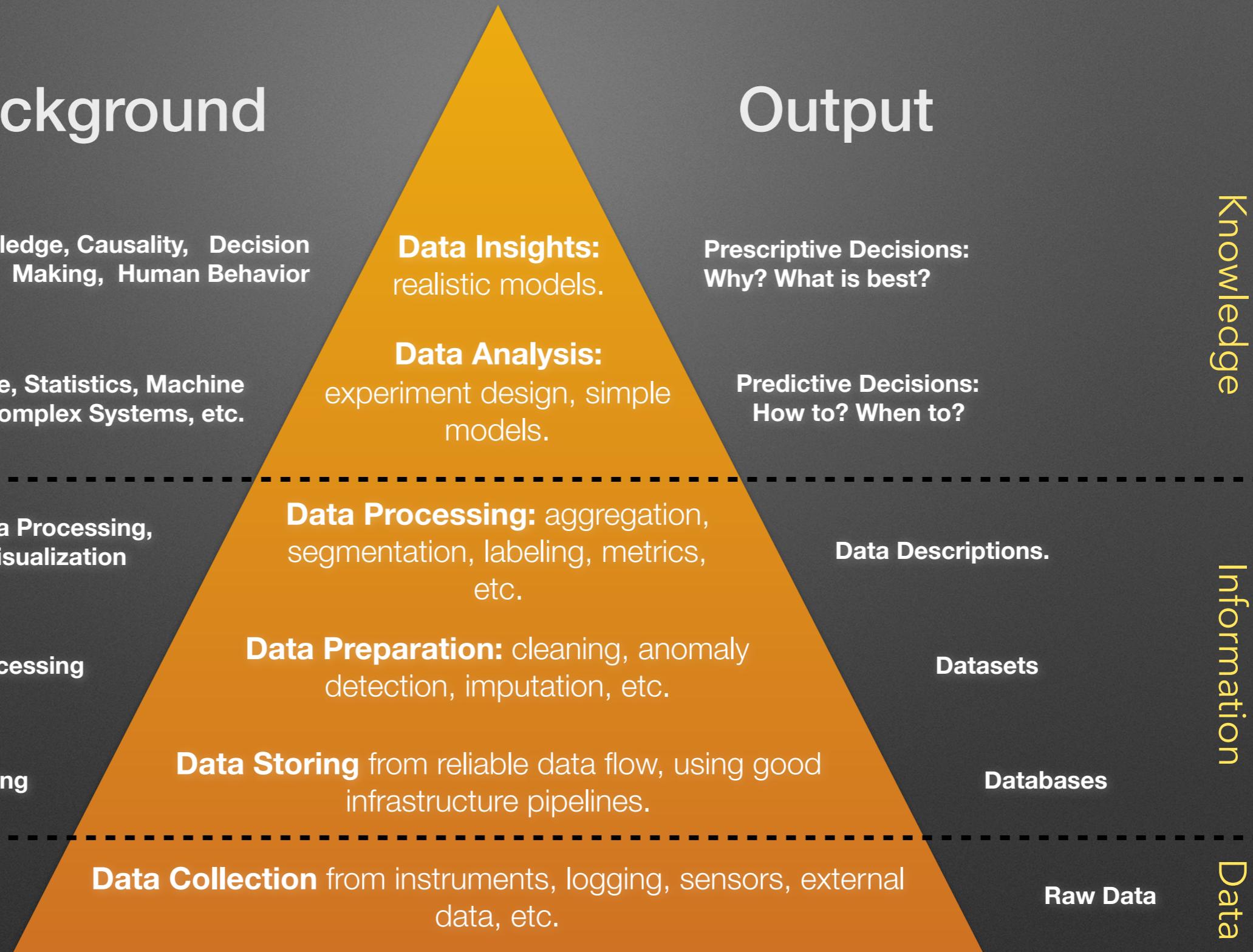
Predictive Decisions:
How to? When to?

Data Descriptions.

Datasets

Databases

Raw Data



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!

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

What are the limits of Data Science

- Data science must be bounded by ethical limits.
- 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). Prediction is a dangerous game!

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."

Ethical Data Science

If a DS system is making automatic decisions, someone has the **responsibility** of those decisions.

Problems:

- Choosing a wrong model.
- Building a model with inadvertently discriminatory rules.
- Not providing explanations about decisions.
- Not respecting privacy.
- Etc.

Ethical Data Science

Responsible data science challenges:

- Data science **without prejudice** - How to avoid unfair conclusions even if they are true?
- Data science **without guesswork** - How to answer questions with a guaranteed level of accuracy?
- Data science that **ensures confidentiality** - How to answer questions without revealing secrets?
- Data science that **provides transparency** - How to clarify answers such that they become indisputable?

Canonical Problems and Tools

Classification	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?
Regression	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?
Clustering	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?
Dimensionality reduction	What are the most significant features of this data and how can these be summarised?	E-commerce: what combinations of features allow us to summarise the behaviour of our customers? Molecular biology: how can scientists summarise the behaviour of all 20,000 human genes in a particular diseased tissue?
Semi-supervised learning	How can labelled and unlabelled data be combined?	Computer vision: how can object detection be developed, with only a small training data set? Drug discovery: which of the millions of possible drugs could be effective against a disease, given we have so far only tested a few?
Reinforcement learning	What actions will most effectively achieve a desired endpoint?	Robots: how can a robot move through its environment? Games: which moves were important in helping the computer win a particular game?

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:

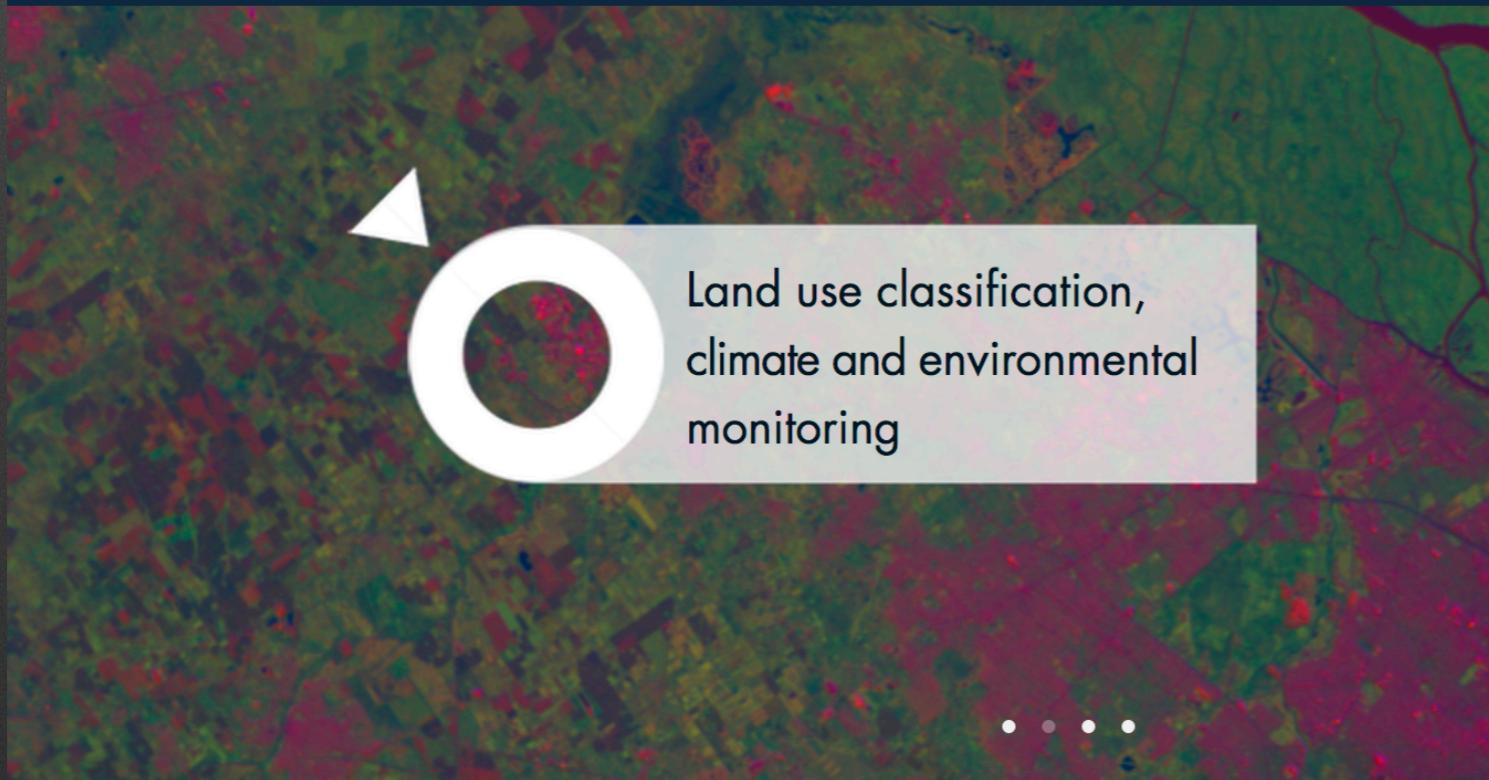
Starbucks collects data on its customers' purchasing habits in order to send personalized ads and coupon offers to the consumers' mobile phones. The company also identifies trends indicating whether customers are losing interest in their product and directs offers specifically to those customers in order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

Data Science



Home Smart Data Industry Solutions Hyperspectral Company Jobs

ENABLING LIVE
GEO-INFORMATION ANALYTICS



Sole supplier of
high resolution
hyperspectral
data

Data Science

[HOME](#)[TEAM](#)[CAREERS](#)

Your Personal Doctor Online

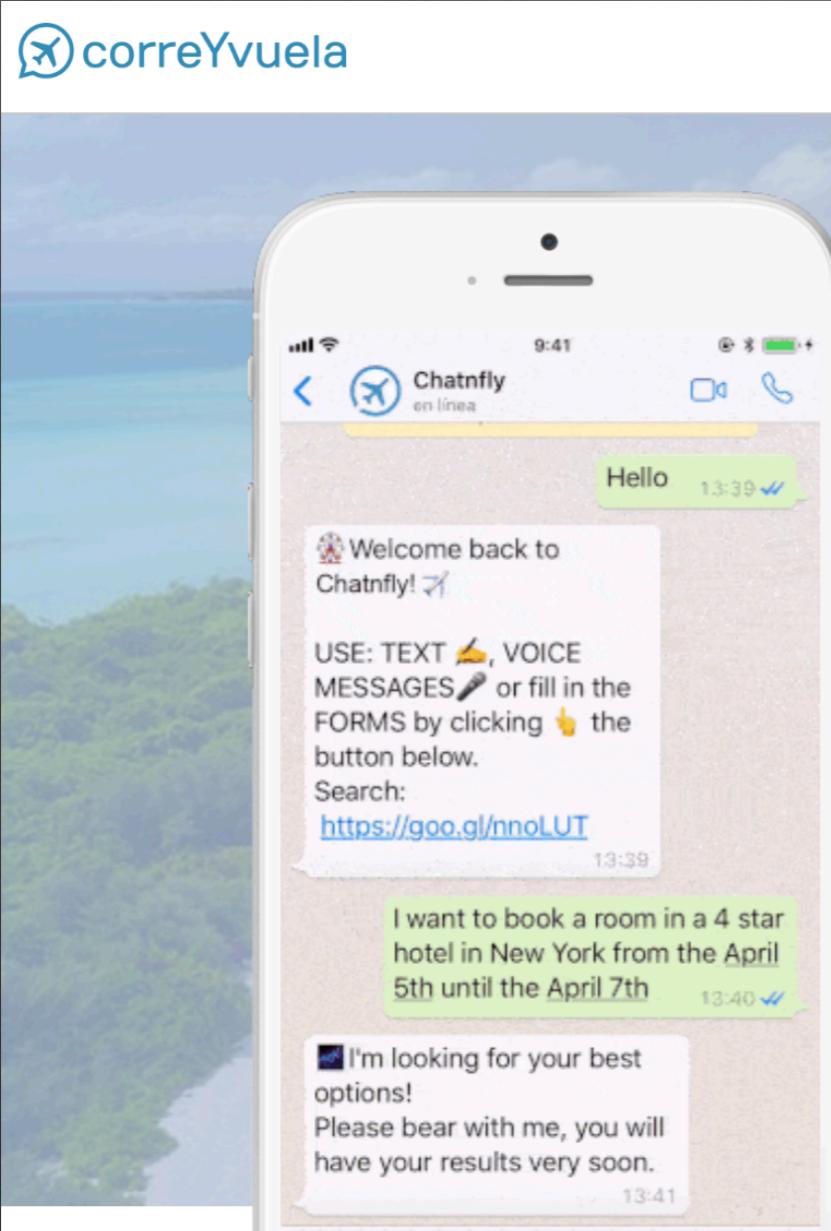
OUR MISSION

Scaling the world's best healthcare to every human being

OUR APPROACH

We are using artificial intelligence / machine learning with a user-centric focus to provide instant medical expertise that is accurate, trustworthy, relevant, and actionable.

Data Science



The screenshot shows a mobile application interface for booking flights and hotels. At the top left is the logo "correYvuela". At the top right are links for "How it works", "FAQs", "SAAS", "Contact", and "Language: ". Below the header is a large image of a tropical beach.

Chatnfly en linea

Hello 13:39

Welcome back to Chatnfly!

USE: TEXT , VOICE MESSAGES or fill in the FORMS by clicking the button below.

Search: <https://goo.gl/nnoLUT>

I want to book a room in a 4 star hotel in New York from the April 5th until the April 7th 13:40

I'm looking for your best options!
Please bear with me, you will have your results very soon. 13:41

Book your flight and hotel through our app
Download it!

Play Store App Store

iPuedes probarnos en nuestro chat web!

Data Science

The image shows the homepage of the Social Point website. At the top, there's a navigation bar with the "socialpoint" logo, a search icon, and links for HOME, GAMES, JOBS, ABOUT, BLOG, PRESS, and COMMUNITY.

The main banner features the game "Monster Legends" with various cartoonish monsters like a blue dragon, a green lizard-like creature, and a yellow bird-like creature. It includes download links for Facebook, App Store, and Google Play.

Below the banner, on the left, is a section titled "DISCOVER WHO WE ARE AT SOCIAL POINT!" which includes a video thumbnail showing office life and a quote from Alba Rodriguez.

On the right, there's a "WE'RE HIRING" section with a photo of a woman holding a small toy and a "CHECK OUT ALL JOBS" button.

At the bottom left, there's a footer with social media icons for Google+ and Facebook, and a comment count of "0 comments".

Discover who we are at Social Point!

Discover who we are at Social Point!

"We share what we learn and we learn from each other."

Alba Rodriguez
Head of Influencer Marketing

OUR OFFICES ARE BECOMING MORE AND MORE HEALTHY AND ECO-FRIENDLY EVERY DAY

G+ 0 comments

CHECK OUT ALL JOBS

Data Science

Kernel
analytics



Analytics at the core

Data helps businesses make better decisions.
We help businesses make the most of their data.

Want to know more about us?

[CONTACT US](#)

Want to work at Kernel Analytics?

[SEE JOB OFFERS](#)

Some of our clients

Data Science

glassdoor Jobs Company Reviews Salaries Interviews Know Your Worth Sign In Write Review For Employers Post Jobs Free

data scientist Jobs Barcelona Search

Job Type Date Posted Salary Range Distance More Create Job Alert

Data Scientist Jobs in Barcelona 95 Jobs

Data Scientist Social Point - Barcelona 21 days ago  models that can be embedded in our ongoing data processes, working closely with Data Engineers and Tooling Developers - Debugging... people to join our teams. About the role: As part of our Data Science team you will be working to apply scientific models to...

Data Scientist Chartboost - Spain  EASY APPLY We're Hiring  We're seeking a superb datascientist to build and evolve our monetization models for both the demand (advertisers) and supply... failures and successes backed by data. Define, advocate and maintain high standards for data quality. Who you are: 3-5...

Data Analyst TravelPerk - Barcelona  New  relevant experience interpreting data in a business intelligence, analytics, or data science/datascientist role, including relevant tools. Management with visual data

Data Scientist Social Point- Barcelona, ES  Apply Now Save

Job Company Rating Reviews

About us:
We are a rapidly growing social game developer, with top ranked games on Mobile and Facebook.
We have over 50 million fans worldwide playing our games all over the world.
There are about 350 of us creating super fun games in our offices, located 10 minutes from the beach, in the beautiful and sunny Barcelona.
Last year we were acquired by Take Two Interactive (GTA, Bioshock, XCOM) and have been expanding our activities, creating exciting new opportunities for top people to join our teams.

About the role:

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



Swimming companies

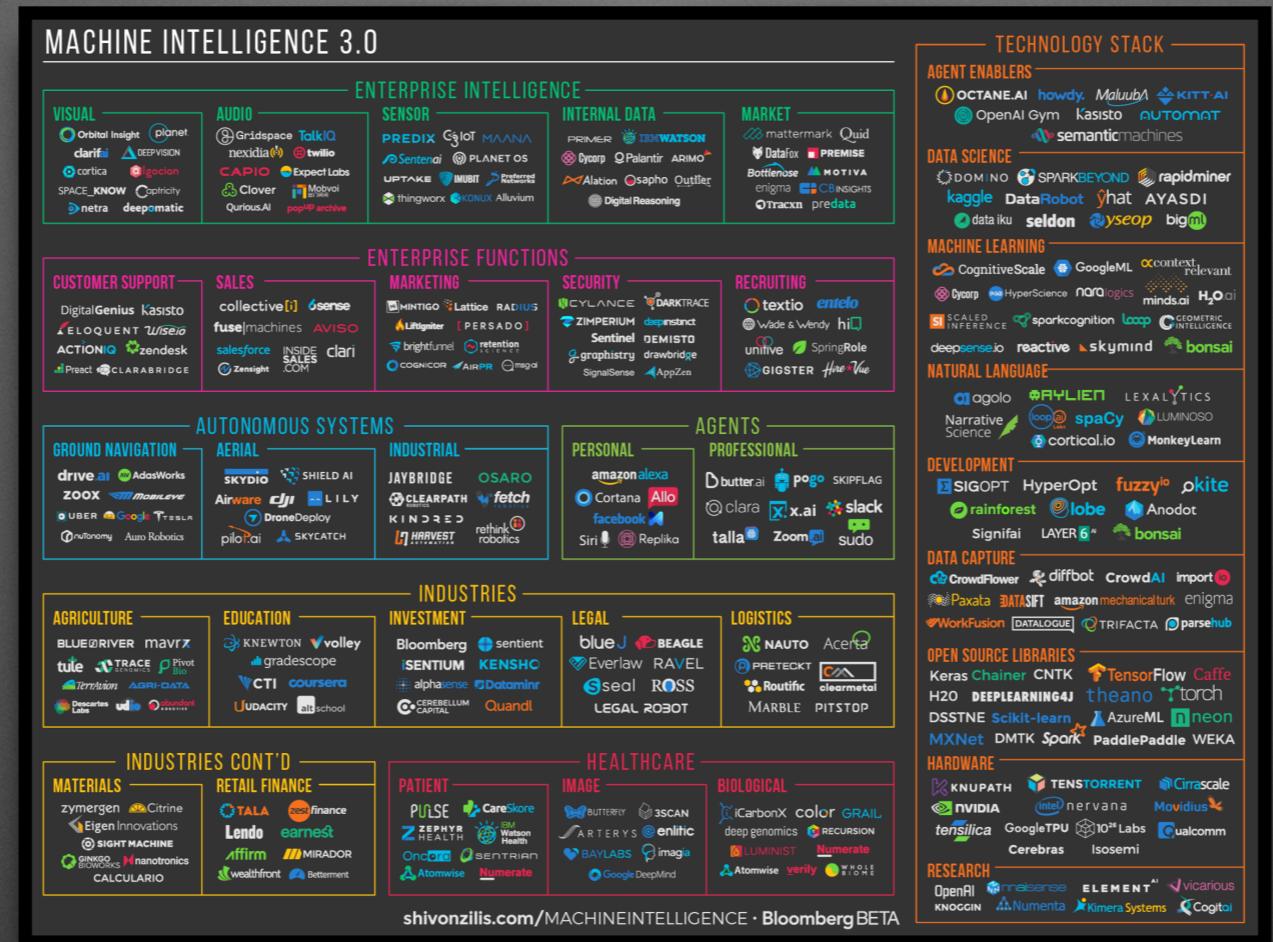
Walking companies



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.



Datification is not the only ingredient of the data science revolution. The other ingredient is the **democratization** of data analysis.

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 **Data**.
- **Data Science** is a new job that is here to stay.