What is Data Science?

(a personal view: connecting data to reality)

Jordi Vitrià, PhD

UNIVERSITAT DE BARCELONA

jordi.vitria@ub.edu

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?

Question

What are my data sources? How reliable are they?

Acquire

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

Describe

What are the key relationships in my data?

Discover

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

Analyze

What are the likely future outcomes?

Predict

Are my expectations fulfilled?

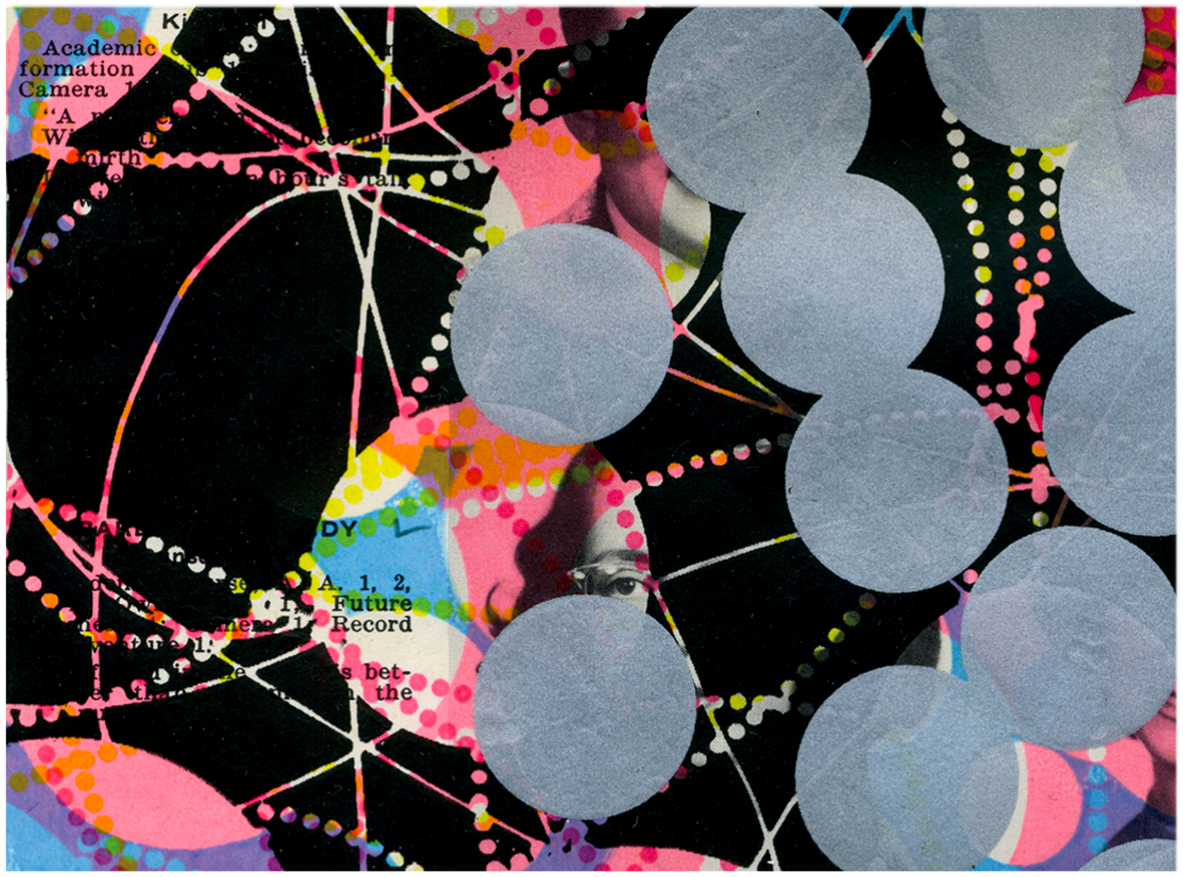
Evaluate

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…

Harvard
Business
Review



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.

Big Data

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.

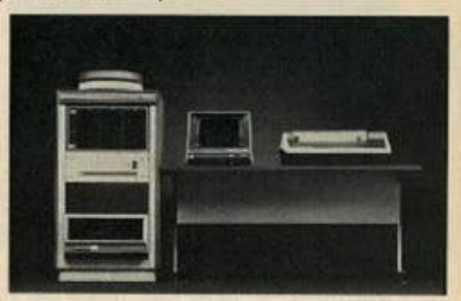
10 Megabyte Hard Disk

$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

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:

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

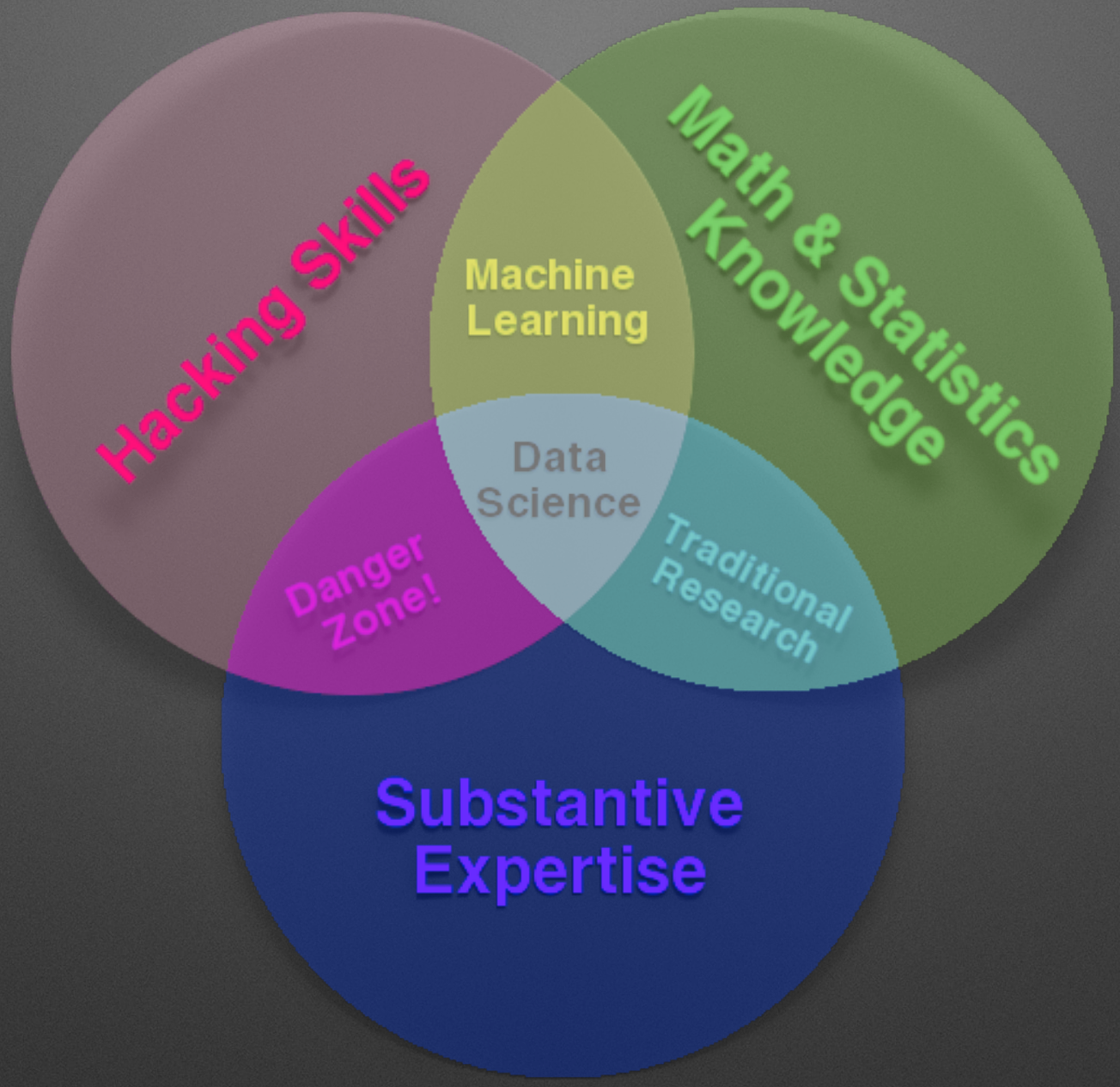
Data Science

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.

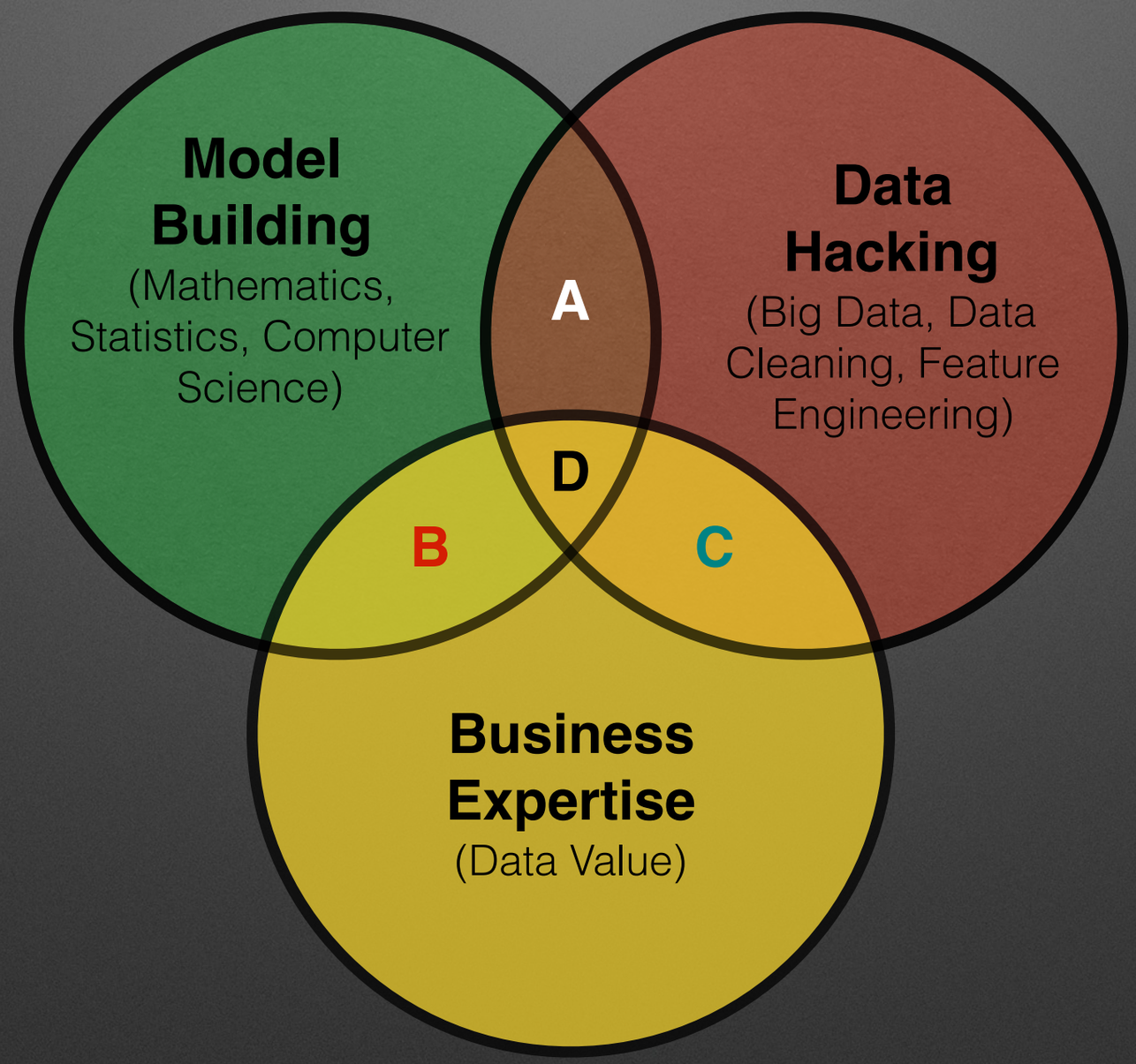
Data Science is a **methodology** to define what we want to do with data, how do we evaluate our actions, what decisions can be grounded on data, how do we combine evidences from several sources, etc.



Drew Conway's Data Science Venn Diagram

D is an empty set!

A+B+C = D



Data Science Tasks

Background

Output

Domain Knowledge, Causality, Decision Making, Human Behavior

Data Insights: realistic models.

Prescriptive Decisions: Why? What is best?

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

Data Analysis: experiment design, simple models.

Predictive Decisions: How to? When to?

Knowledge

Data Processing, Visualization

Data Processing: aggregation, segmentation, labeling, metrics, etc.

Data Descriptions.

Data Processing

Data Preparation: cleaning, anomaly detection, imputation, etc.

Datasets

Data Engineering

Data Storing from reliable data flow, using good infrastructure pipelines.

Databases

Information

Data Engineering

Data Collection from instruments, logging, sensors, external data, etc.

Raw Data

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!

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."

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

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

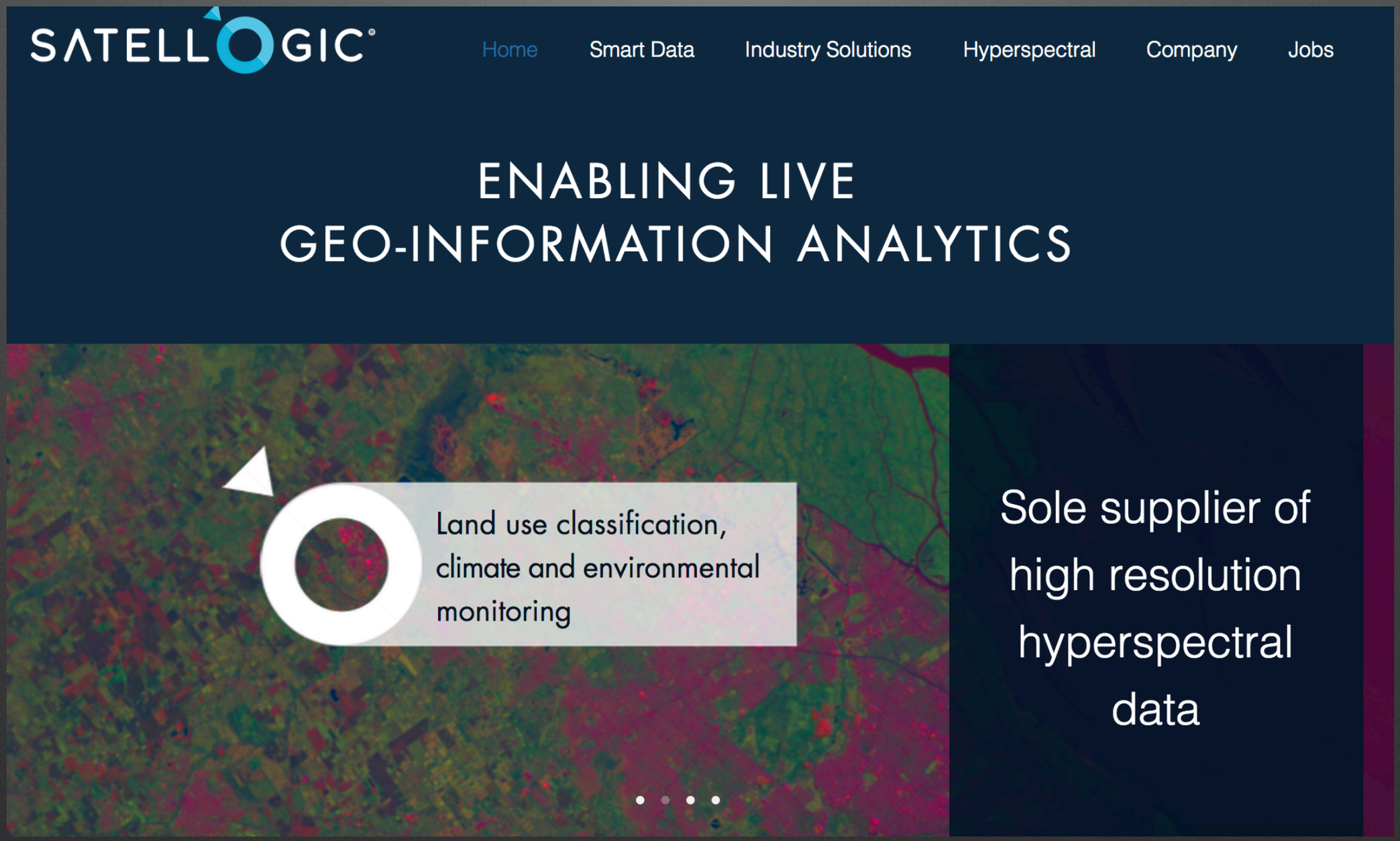
EMPLOYEES
160,000

TYPE
Behavioral Analytics

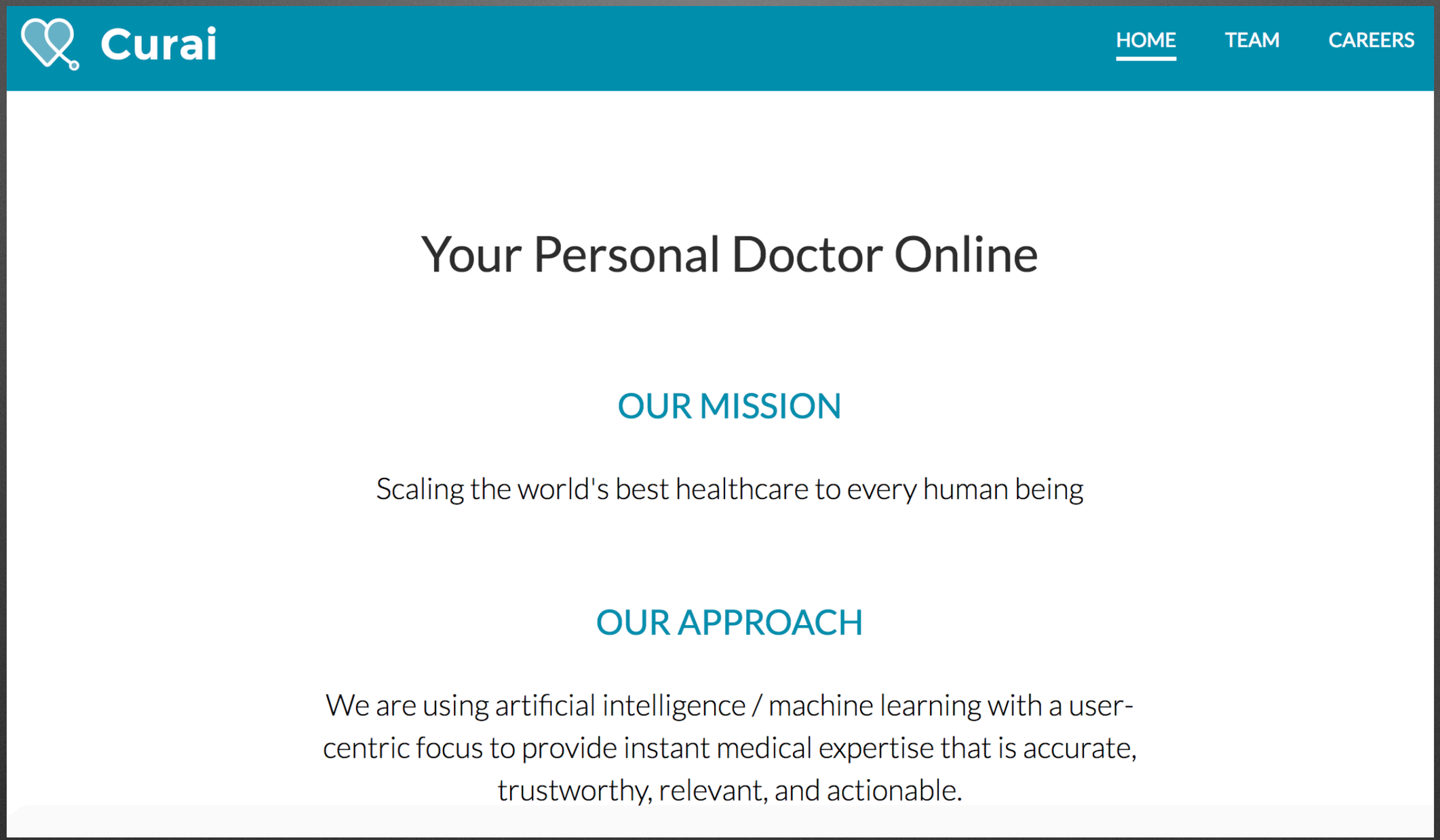
PURPOSE:

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.

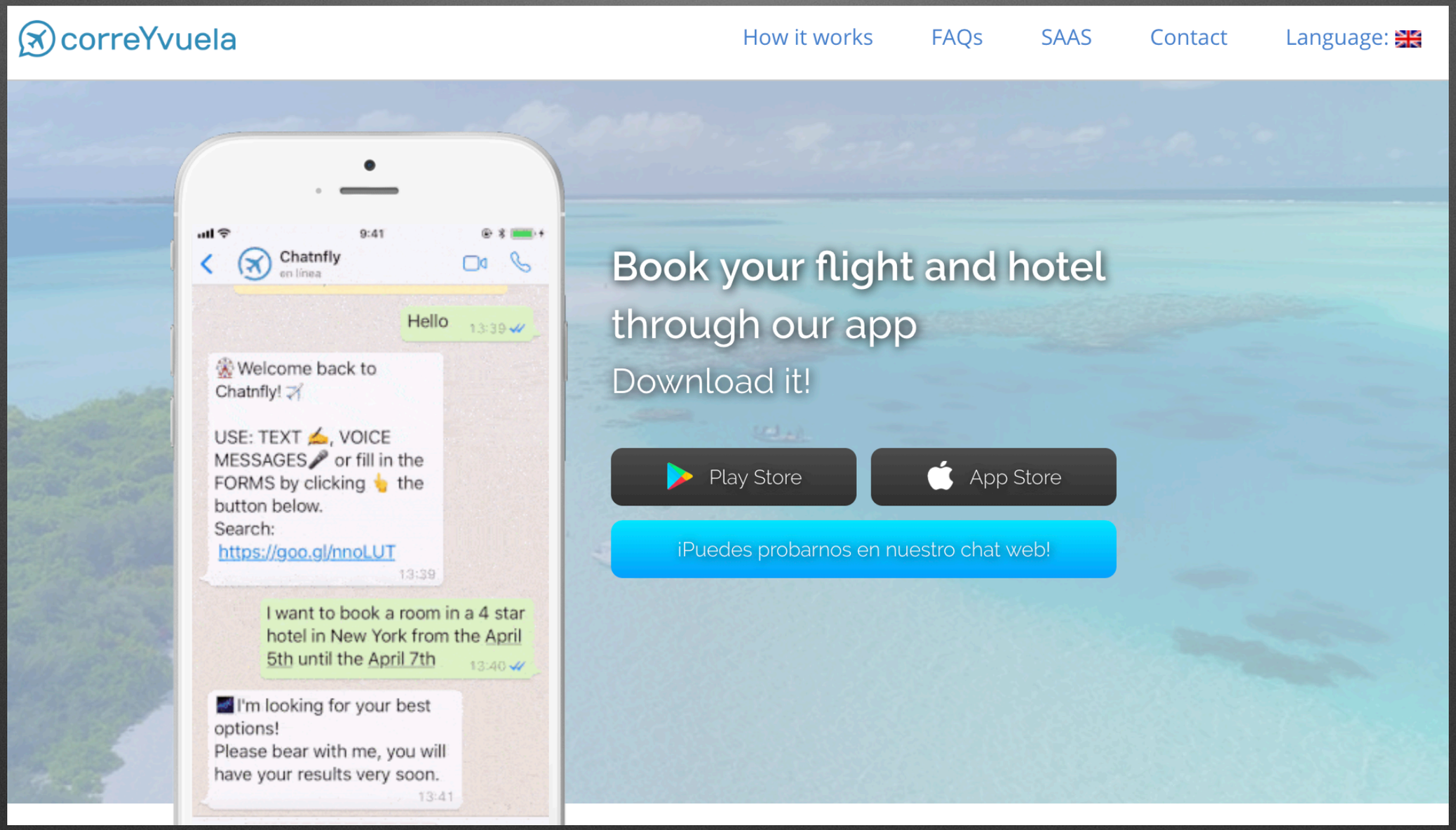
Data Science



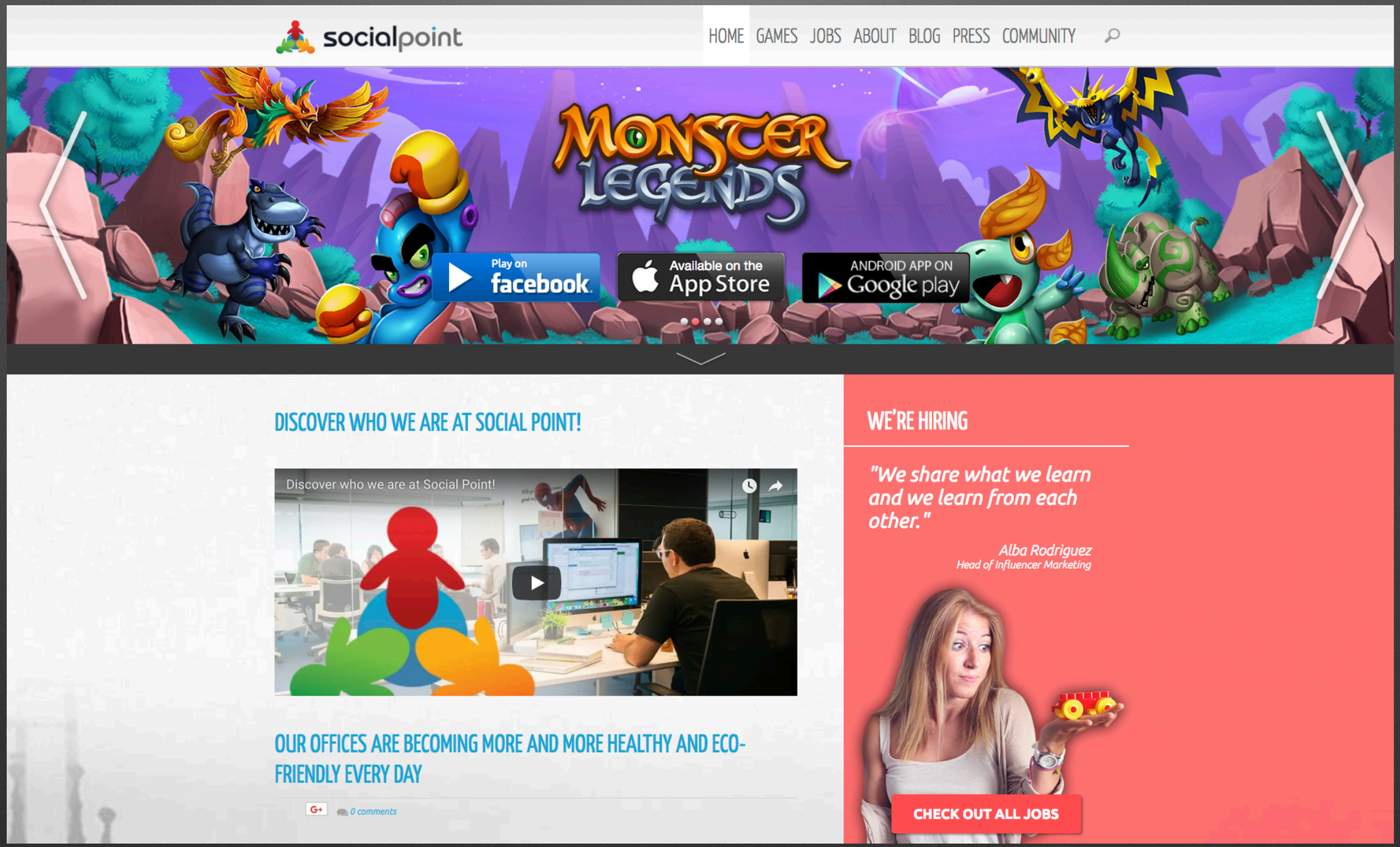
Data Science



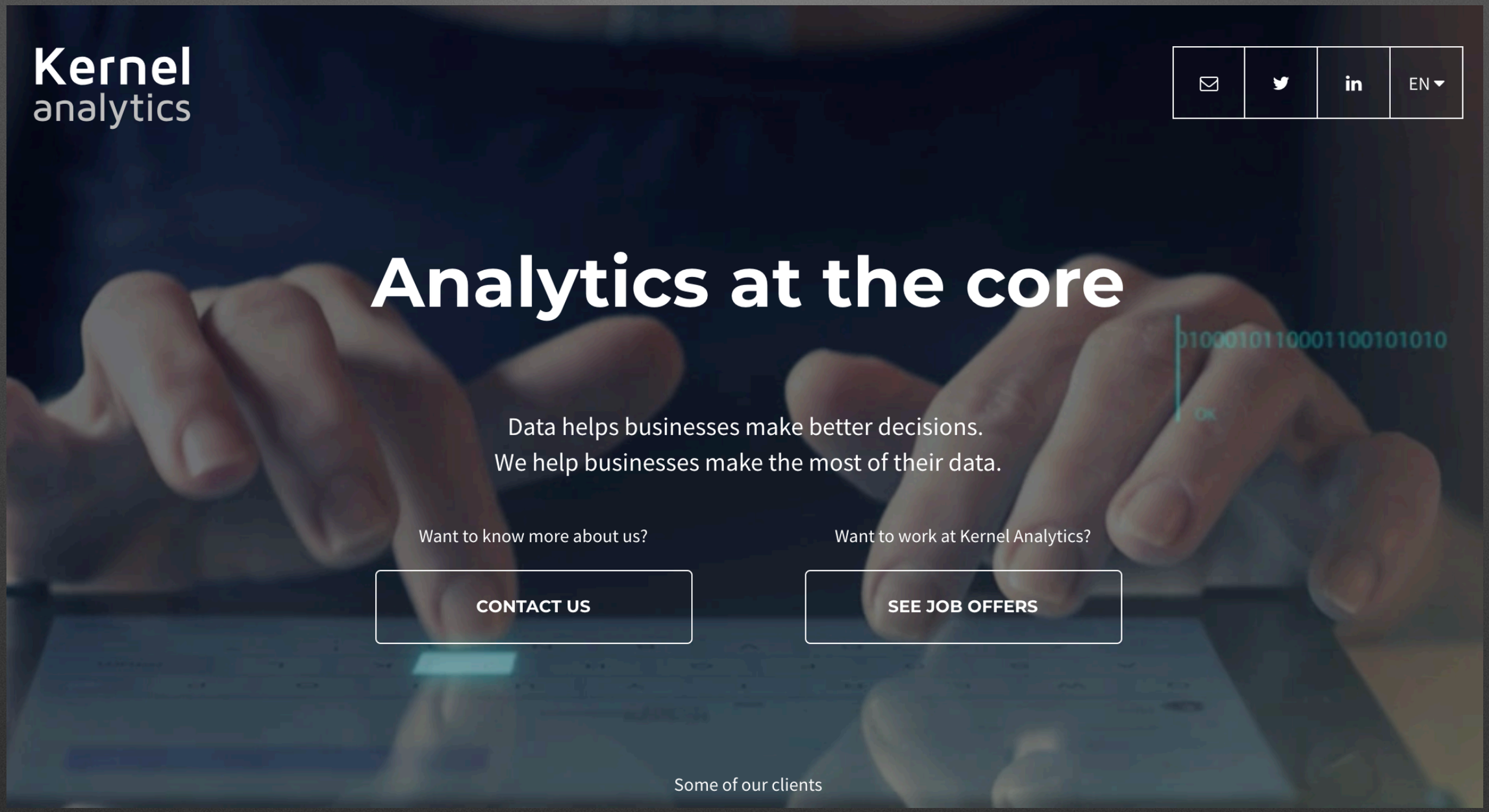
Data Science



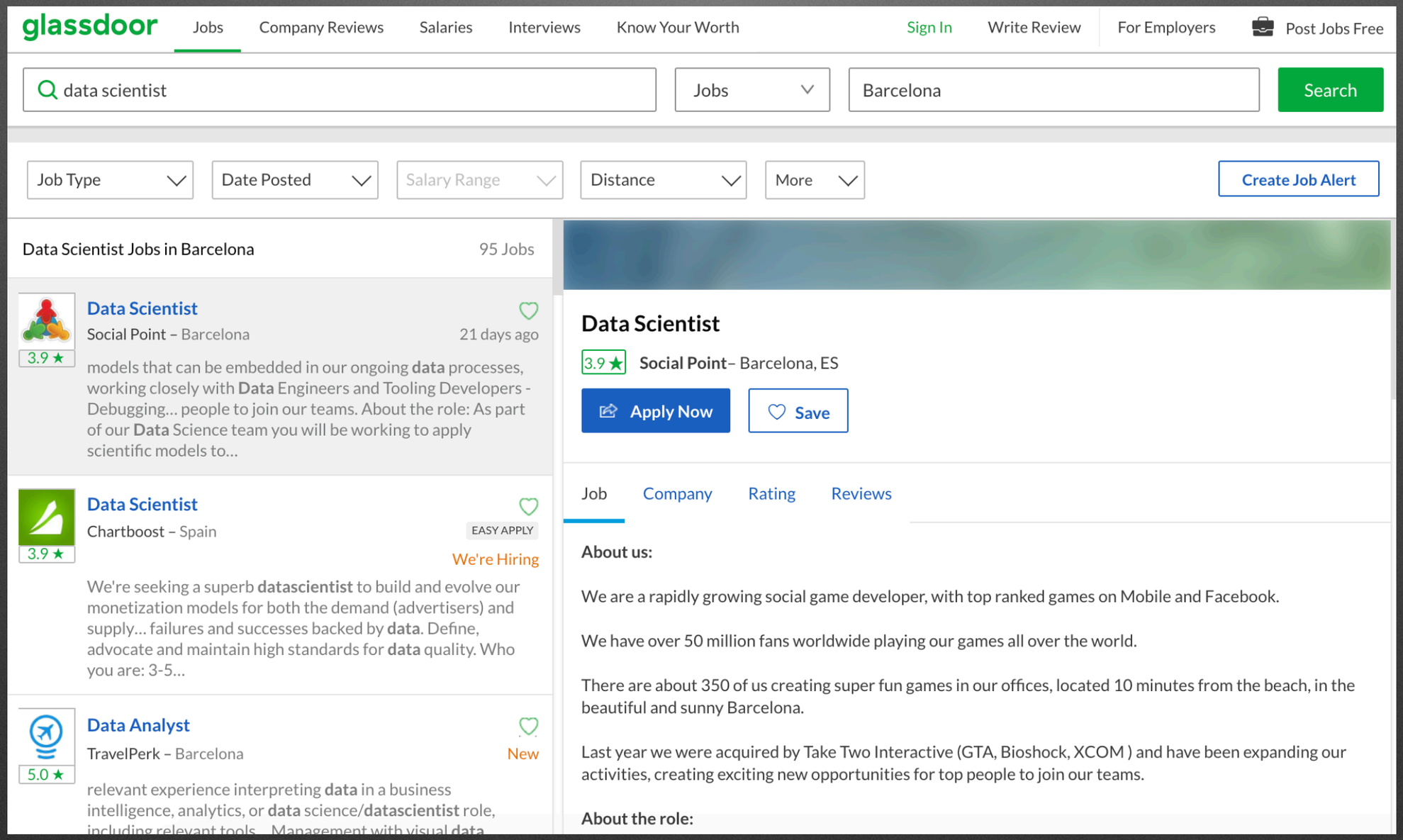
Data Science



Data Science



Data Science



Data Science is for all,

small and big, old and new, etc.

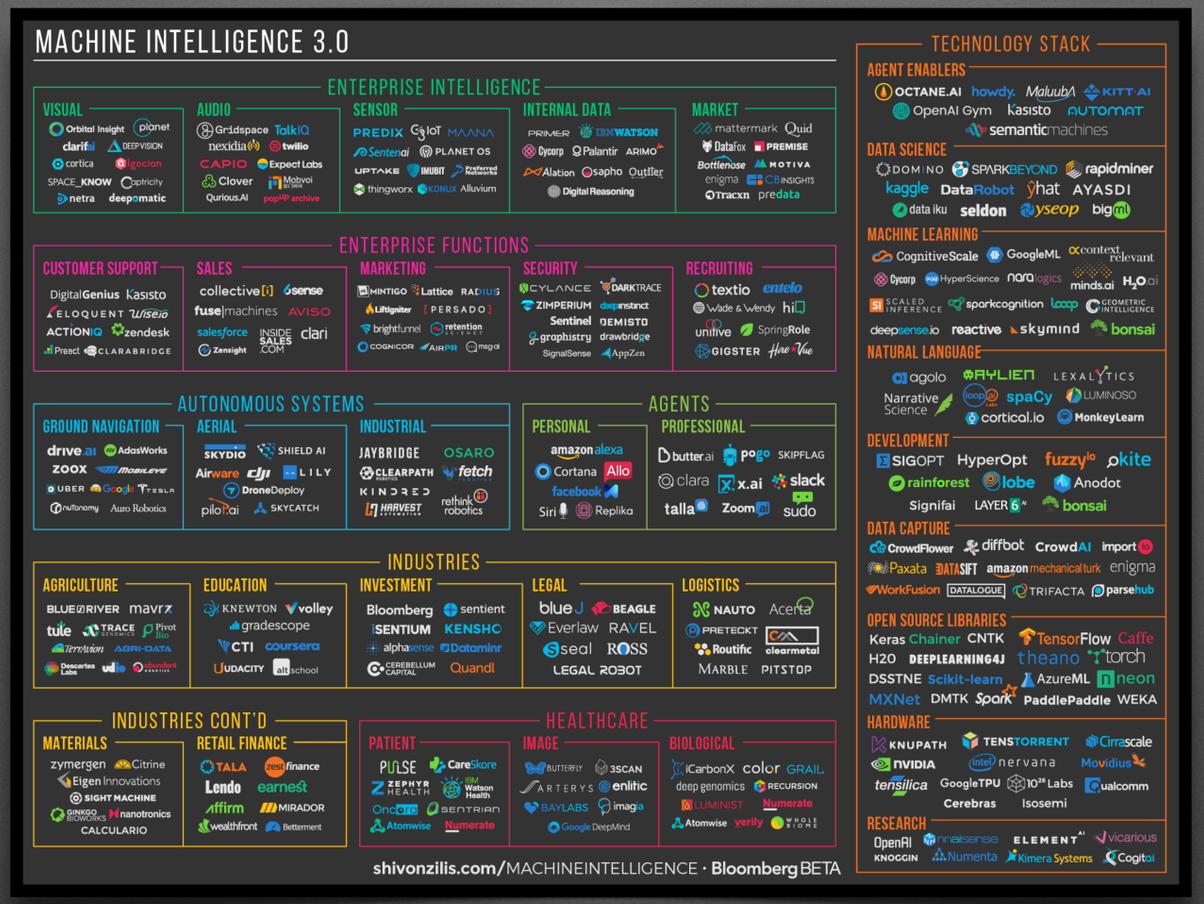
Swimming companies

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.



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.