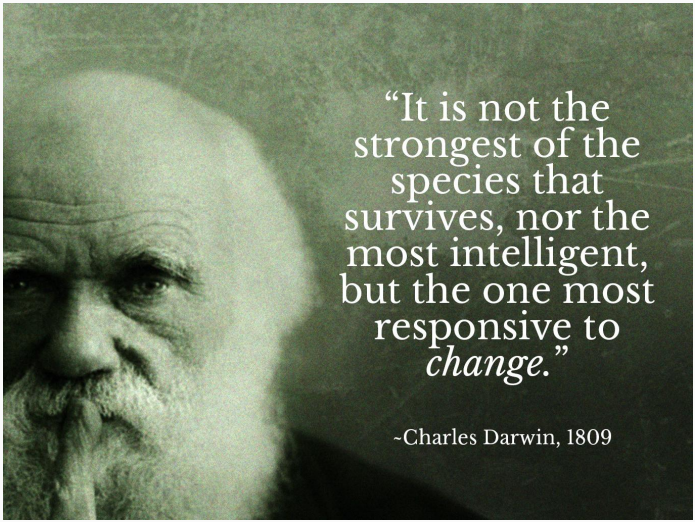
Data, Big Data, and Artificial Intelligence... The Data Science

Seminario TAP - Marzo 2022

Giovanni Giuffrida

Overview

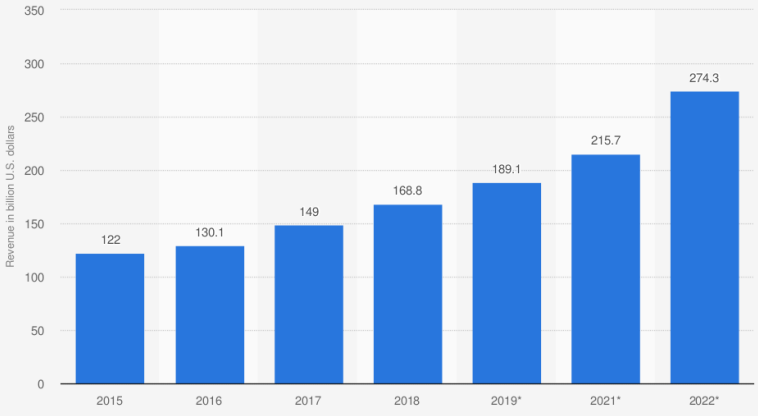
- How important (big) data is today
- Big Data & Artificial Intelligence Introduction
- Some considerations on our Big Data based society
- Cultural shifts imposed by AI
- Application examples


“It is not the strongest of the species that survives, nor the most intelligent, but the one most responsive to *change*.”
~Charles Darwin, 1809

Introduction to Big Data

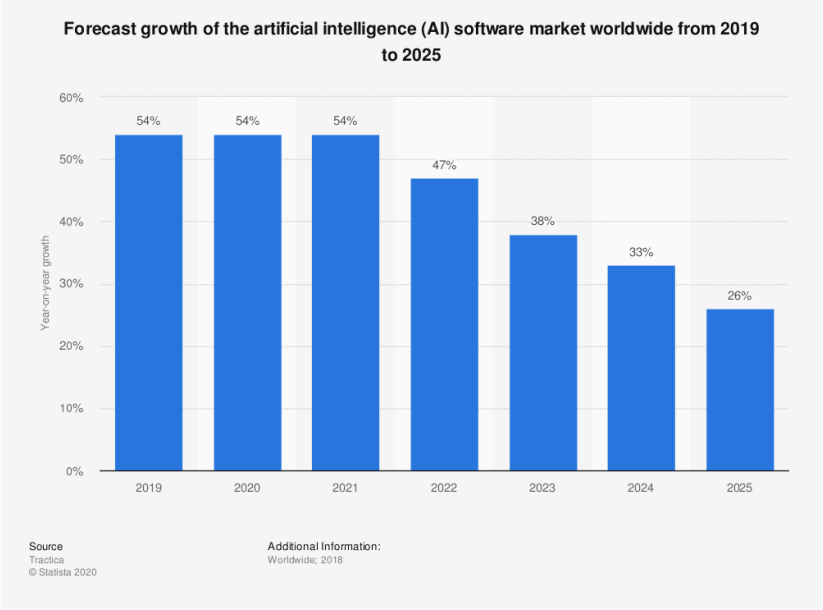

BIG
DATA
INFORMATION
TECHNOLOGIES
MANAGEMENT
ANALYSIS
RESEARCH
SEARCH
INTERNET
PETABYTES
EXABYTES
STORAGE
LARGE
SETS
MILLION
SOFTWARE
SAN
NAS
SOCIAL
GOVERNMENT
PROCESS
VOLUME
SIZE
DATABASES
PROCESSING
FRAMEWORK
BUSINESS
VISUALIZATION
ANALYTICS
SIMULATIONS
AMOUNT
APPLICATIONS
INFRASTRUCTURE
LEARNING
WORLD
EVERY
BASED
DEVELOPMENT
TECHNOLOGY
SCIENCE
FUTURE
BILLION
NEW
PARALLEL
INITIATIVE
ECONOMIC
TERABYTES
CRITIQUES
NETWORKS
INCLUDE
HYPOTHESIS
PAST
CHALLENGES
HANDLE
SOURCES
DRIVE
REPLICATION
RESULTS
YEARS
FACTOR
SYSTEMS
MANAGE
ONLINE
DAY
WORLDWIDE
MASSIVE
DEFINITION
BETWEEN
STATISTICS
USING
DETERMINE
WELL
SET
DEPARTMENT
BUREAUS
HIGH
USERS
REAL
STRUCTURE
WORK
RATE
HIGHER
ALGORITHMS
COMPANIES
MARKING
INTEGRATION
ANNOUNCED
LIMITS
COMPLEX
CAPACITY
HAND
ONE
PRIVACY
INTERNATIONAL
IMPLEMENTATION
BECOMES
DISCUSS
PRIVATE
LOSS
HUGE
TERMS
YEAR
CAPABILITIES
SUCCESSFUL
NEEDED
USES
COMPANIES
VARIOUS
TIMES
SHARED
APPROACHES
PROCESSES
INTELLIGENCE
NATIONAL
FLOW
ARCHITECTURE
PARADIGM
FUNDING
INSIGHT
TOOLS
CENTER
RECORDS
COST
PER
DISTRIBUTED
RELEVANT
USED
PROJECT
ANNUAL
NOW
VARIETY
MARKET
SENSOR
TIME
RELATED
CURRENTLY
USE
PEOPLE
TRAFFIC
STATE
CRITIQUE
MAPREDUCE
NEARLY
MOST
END
APPROACH
MANY
OUT
PAST

Revenue from big data and business analytics worldwide from 2015 to 2022 (in billion U.S. dollars)



Source
IDC
© Statista 2021

Additional Information:
Worldwide; 2015 to 2021


Forecast growth of the artificial intelligence (AI) software market worldwide from 2019 to 2025
Year-on-year growth
60%
50%
40%
30%
20%
10%
0%
54%
54%
54%
47%
38%
33%
26%
2019
2020
2021
2022
2023
2024
2025
Source
Tractica
© Statista 2020
Additional Information:
Worldwide; 2018

Roots

- **Datification** is the revolution behind Big Data
- People have always tried to quantify the world around them

- **Datification** is the revolution behind Big Data
- People have always tried to quantify the world around them

Always-connected digital technologies made this a reality

Big Data releases

- **1.0**: Big Data become available and get collected

Big Data releases

- **1.0**: Big Data become available and get collected
- **2.0**: Technology to process Big Data develops

Big Data releases

- **1.0**: Big Data become available and get collected
- **2.0**: Technology to process Big Data develops
- **3.0**: Getting value out of Big Data (The most difficult!)

The top 6 most capitalized companies in the world

2008

Company	USbn
Exxon Mobil	453
PetroChina China	424
General Electric	369
Gazprom Russia	300
China Mobile	298
Bank of China	278

The top 6 most capitalized companies in the world

2008

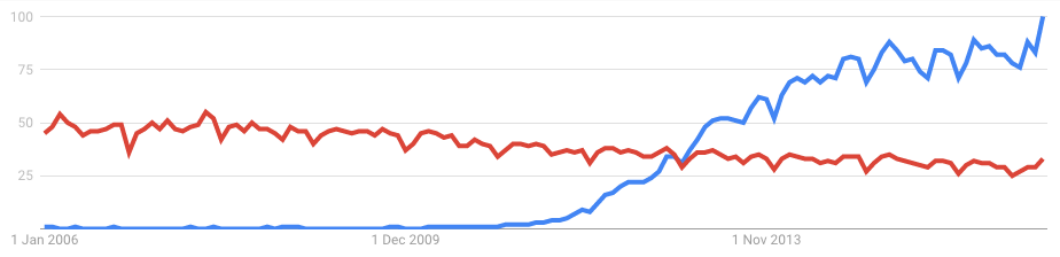
Company	USbn
Exxon Mobil	453
PetroChina China	424
General Electric	369
Gazprom Russia	300
China Mobile	298
Bank of China	278

2018

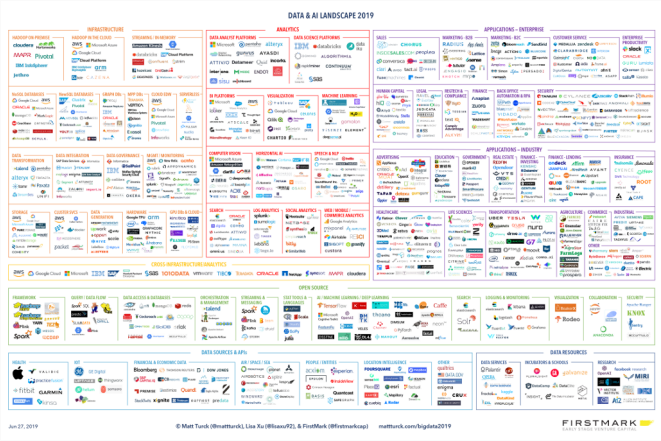
Company	USbn
Apple	927
Amazon	778
Google	767
Microsoft	750
Facebook	542
Alibaba	500

When it became trendy?

Big Data vs Business Intelligence



A complicate landscape



Big Data according to Oxford Dictionary

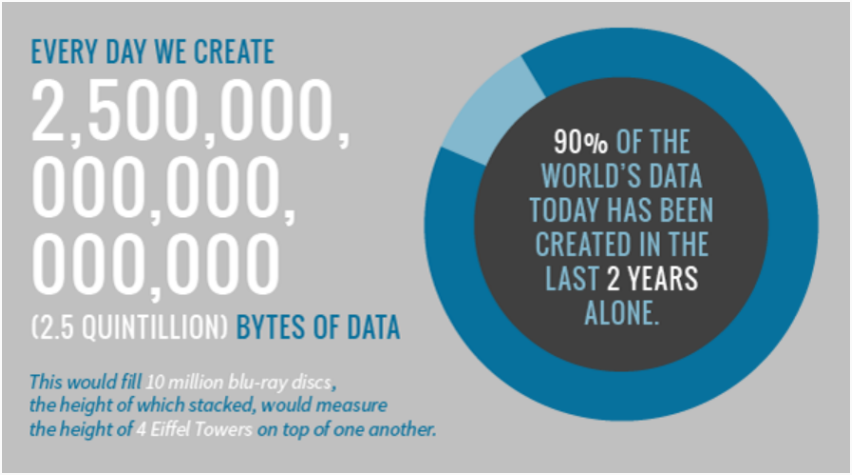
big data n. Computing (also with capital initials): *data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges; (also) the branch of computing involving such data.*

A simpler view

Anything that does not fit in Excel

How big is big?

How big is big?



Big Data definition...

- According to Gartner:

  "Big data is high-**Volume**, high-**Velocity** and high-**Variety** information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

Big Data definition...

- According to Gartner:

  "Big data is high-**Volume**, high-**Velocity** and high-**Variety** information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

- **Value** and **Veracity** added later

Big Data definition...

- According to Gartner:

  "Big data is high-**Volume**, high-**Velocity** and high-**Variety** information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

- **Value** and **Veracity** added later
- **Volatility** and **Validity** added later

Big Data definition...

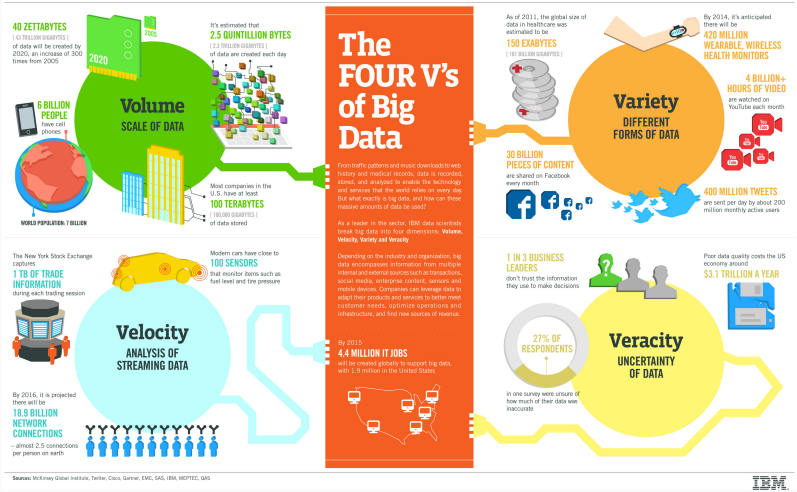
- According to Gartner:

  "Big data is high-**Volume**, high-**Velocity** and high-**Variety** information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

- **Value** and **Veracity** added later
- **Volatility** and **Validity** added later
- **Virality** added even later

Frankly... No formal definition yet!

- These numbers outstrip our machines and our imagination
- Technology to process all this is behind
- ”Big” depends on the context for many



http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg

Volume: Data at rest

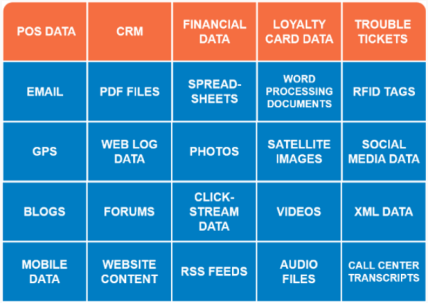
- About: Amount of data
- Unit: bytes
- Information about the general population, education, health, medicine, travel, geographic locations, shopping, financial transactions, jobs, scientific experiments, emails, sensors, texts, photos, videos, activity on social networks, etc.
- How **big** is ”Big Data”?

Velocity: Data in motion

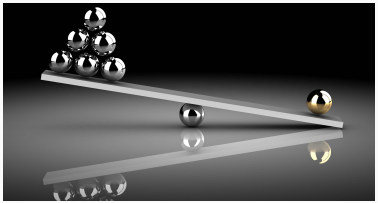
- About: Moving data
- Unit: Bytes per second
- Two possible interpretations
  - Data Generation Rate
  - Data Processing Rate

Variety: Data in many forms

- About: Form of data
- Three basic types of data
  - Structured = Data in a fixed field within a record (spreadsheets, Relational Database)
  - Semi-Structured = XML, JSON, CSV
  - Unstructured = Data stored without any model, or that does not have any organisation
- Any of those types can be big
- Today, only 20% of data is ”structured”



Veracity: Data in doubt



- Uncertainty due to many factors
  - Incompleteness
  - Inconsistency
  - Ambiguity
  - Model approximation
  - Technical constraints
- Often overlooked
- But... it could as important as the other Vs

People do not need data, they need insights!!

- Hidden in the data
- Value is a *concentrated data-juice*
- Gaining correct but irrelevant or un-actionable information is a (big) waste of time

The Artificial Intelligence evolution

In 1956 the term **Artificial Intelligence** was coined by John McCarthy

Artificial Intelligence: a (very) old friend in town

In 1956 the term **Artificial Intelligence** was coined by John McCarthy

"The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."

The human brain... the most powerful pattern recognition machine?

What’s a “chair”?

What’s a “chair”?

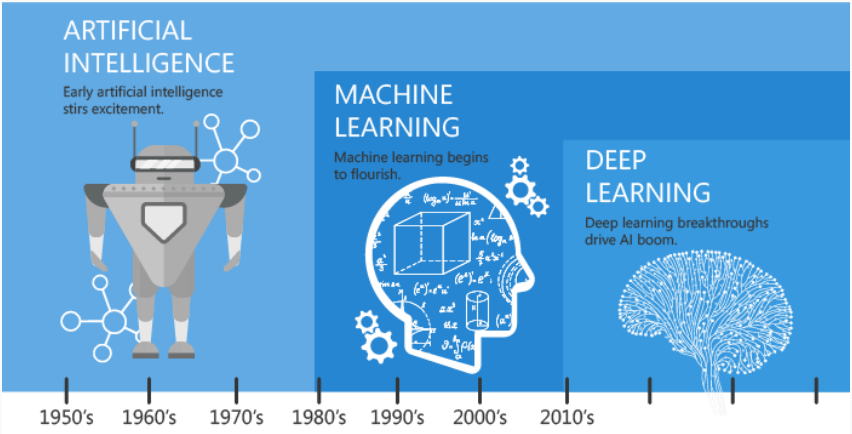
What's a "chair"?

What's a "chair"?

What's a "chair"?

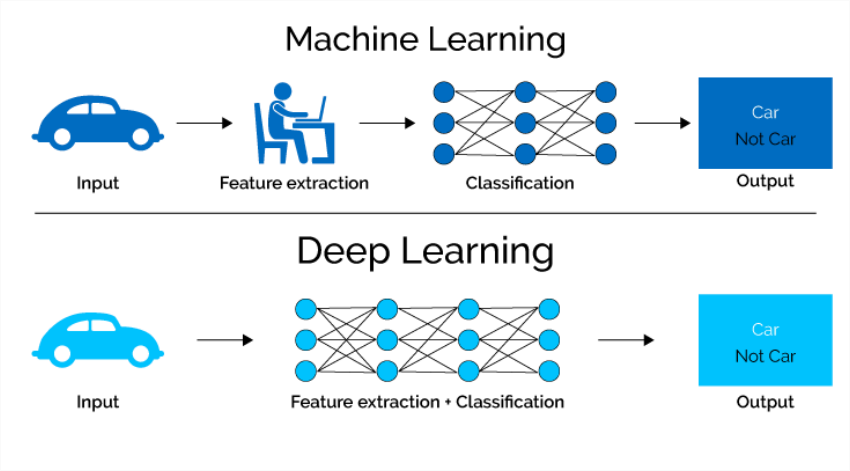



The AI evolution



Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.

Fully automated learning



A new vision of Artificial Intelligence today

- Before: Intelligence was **hardcoded** into machines
- Today: Machines learn by **observing** Big Data

A new vision of Artificial Intelligence today

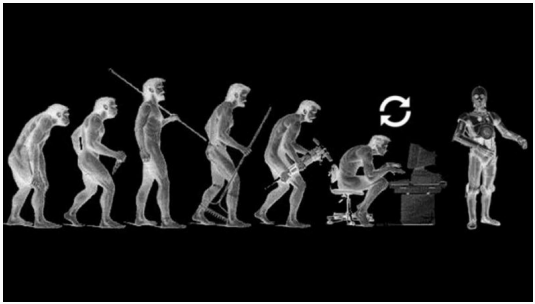
- Before: Intelligence was **hardcoded** into machines
- Today: Machines learn by **observing** Big Data

Big Data $\Longrightarrow$ A.I.

The old vs the new school

- In the past, many attempts to make machines ”Intelligent”: Expert systems, Artificial Intelligence, etc.
- Today, Big data/Artificial Intelligence is about deriving math models (insights) from huge data bases
- Being able to observe and learn models leads to *intelligent behavior*
  - IBM Watson
  - AlphaGo

The old vs the new school



- CYC vs Watson
- Two (very) different approaches
- CYC was “embedding” knowledge
- Watson is able to “learn” from huge amount of data

Cyc

From Wikipedia, the free encyclopedia

For other uses, see CYC (disambiguation).

Cyc (/saɪk/) is the world's longest-lived artificial intelligence project,[*citation needed*] attempting to assemble a comprehensive ontology and knowledge base that spans the basic concepts and "rules of thumb" about how the world works (think common sense knowledge but focusing more on things that rarely get written down or said, in contrast with facts one might find somewhere on the internet or retrieve via a search engine or Wikipedia), with the goal of enabling AI applications to perform human-like reasoning and be less "brittle" when confronted with novel situations that were not preconceived.

Douglas Lenat began the project in July 1984 at MCC, where he was Principal Scientist 1984–1994, and then, since January 1995, has been under active development by the Cycorp company, where he is the CEO.

Contents [hide]

Cyc

Original author(s)	Douglas Lenat
Developer(s)	**Cycorp, Inc.**
Initial release	1984; 35 years ago
Stable release	6.1 / 27 November 2017; 15 months ago
Written in	Lisp, CycL
Type	Ontology and Knowledge Base and Knowledge Representation Language and Inference engine
Website	www.cyc.com

Watson


PIENSE
THINK
DENKE
PENSER
$24,000
Who is Stoker?
(I FOR ONE WELCOME OUR NEW COMPUTER OVERLORDS)
$ 1,000
$77,147
Who is Bram Stoker?
$ 17,973
$21,600
WHO IS BRAM STOKER?
$5600

Jeopardy

Oscar Wilde said of this title place "The warder is despair"	At the beginning of "A Tale of Two Cities", these 2 kings sit on the thrones of England & France	Around 1912, while recovering in a sanatorium, this former seaman decided to become a playwright
The accompanying text to this book was published separately as "Ornithological Biography" in the 1830s	In May 1973 Sports Illustrated ran one of his short stories under the title "A Day of Wine and Roses"	This author & biochemist who died in 1992 has at least one book in all 10 main Dewey Decimal categories
The Prague tombstone of this German-language writer who died in 1924 is inscribed in Hebrew	D.H. Lawrence called him "an adventurer into the vaults and... horrible underground passages of the human soul"	In 1935 she sent a telegram to a Macmillan editor: "Please send manuscript back I've changed my mind"

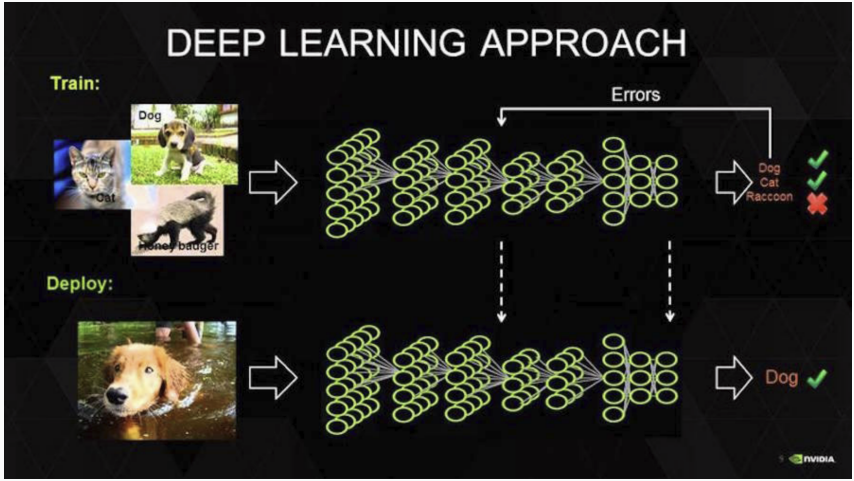
WIRED Technology | Science | Culture | Gear | Business

Credit **IBM**

IBM's Watson -- the language-fluent computer that beat the best human champions at a game of the US TV show *Jeopardy!* -- is being turned into a tool for medical diagnosis. Its ability to absorb and analyse vast quantities of data is, IBM claims, better than that of human doctors, and its deployment through the cloud could also reduce healthcare costs.

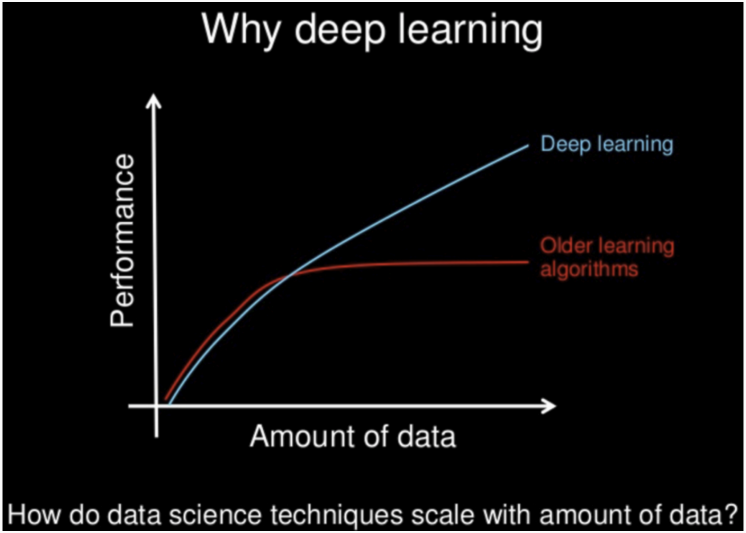
Two years ago, IBM announced that Watson had "learned" the same amount of knowledge as the average second-year medical student. For the last year, IBM, Sloan-Kettering and Wellpoint have been working to teach Watson how to understand and accumulate complicated peer-reviewed medical knowledge relating to oncology. That's just lung, prostate and breast cancers to begin with, but with others to come in the next few years). Watson's ingestion of more than 600,000 pieces of medical evidence, more than two million pages from medical journals and the further ability to search through up to 1.5 million patient records for further information gives it a breadth of knowledge no human doctor can match.

According to Sloan-Kettering, only around 20 percent of the knowledge that human doctors use when diagnosing patients and deciding on treatments relies on trial-based evidence. It would take at least 160 hours of reading a week just to keep up with new medical knowledge as it's published, let alone consider its relevance or apply it practically. Watson's ability to absorb this information faster than any human should, in theory, fix a flaw in the current healthcare model. Wellpoint's Samuel Nessbaum has claimed that, in tests, Watson's successful diagnosis rate for lung cancer is 90 percent, compared to 50 percent for human doctors.


DEEP LEARNING APPROACH
Train:
Dog
Cat
Errors
Dog
Cat
Raccoon
Deploy:
Dog
NVIDIA

Self learning example

(Learning to walk)


Why deep learning
Performance
Deep learning
Older learning algorithms
Amount of data
How do data science techniques scale with amount of data?

In 1997

Alpha GO

- 3000 years old game
- Simple board
- Before 2016 it was considered to be **impossible** to model
- Many (many) more combinations compared to chess
- It was said:
  - *"the most elegant game that humans have ever invented"*;
  - *"simple rules that give rise to endless complexity"*;
  - *"more possible Go positions than there are atoms in the universe"*
- Mostly based on **intuition**

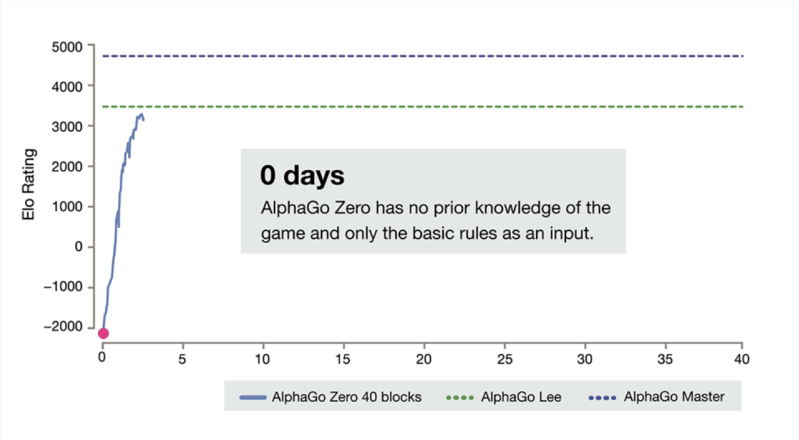
In 2016

It gets better

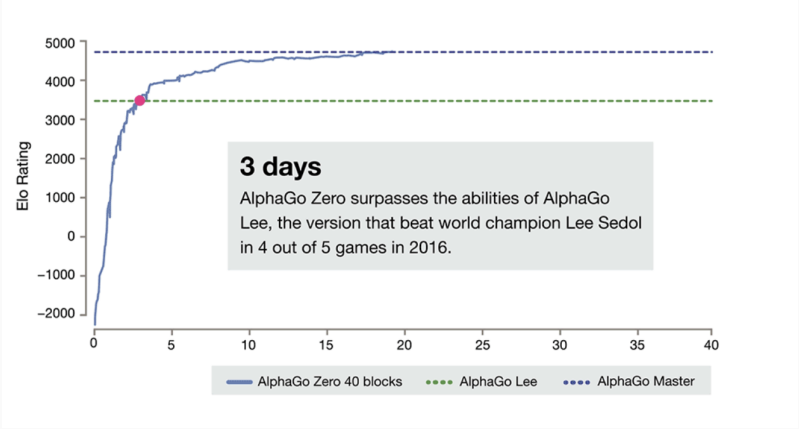
- In 2018 AlphaGo-Zero
- A new version based on Deep Learning techniques

> Previous versions of AlphaGo initially trained on thousands of human amateur and professional games to learn how to play Go. AlphaGo Zero skips this step and learns to play simply by playing games against itself, starting from completely random play. In doing so, it quickly surpassed human level of play and defeated the previously published champion-defeating version of AlphaGo by 100 games to 0.

At the beginning

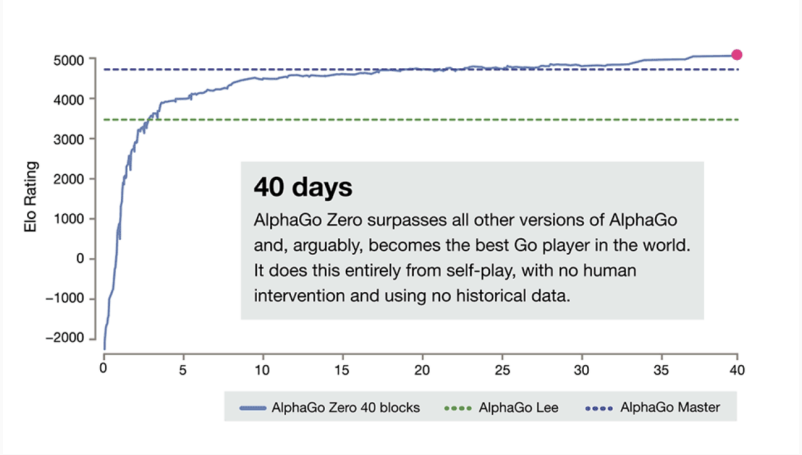


After 3 days


Elo Rating
5000
4000
3000
2000
1000
0
−1000
−2000
0
5
10
15
20
25
30
35
40
3 days
AlphaGo Zero surpasses the abilities of AlphaGo Lee, the version that beat world champion Lee Sedol in 4 out of 5 games in 2016.
AlphaGo Zero 40 blocks
AlphaGo Lee
AlphaGo Master

After 21 days

After 40 days



The Turing test

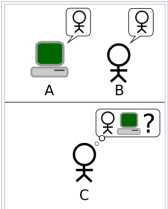
The Turing test

Test di Turing

Da Wikipedia, l'enciclopedia libera.

Il **test di Turing** è un criterio per determinare se una macchina sia in grado di pensare. Tale criterio è stato suggerito da Alan Turing nell'articolo *Computing machinery and intelligence*, apparso nel 1950 sulla rivista *Mind*.[1]

Indice [nascondi]



Rappresentazione grafica del test di Turing.

Descrizione [modifica | modifica wikitesto]

Nell'articolo Turing prende spunto da un gioco, chiamato "gioco dell'imitazione", a tre partecipanti: un uomo A, una donna B, e una terza persona C. Quest'ultimo è tenuto separato dagli altri due e tramite una serie di domande deve stabilire qual è l'uomo e quale la donna. Dal canto loro anche A e B hanno dei compiti: A deve ingannare C e portarlo a fare un'identificazione errata, mentre B deve aiutarlo. Affinché C non possa disporre di alcun indizio (come l'analisi della grafia o della voce), le risposte alle domande di C devono essere dattiloscritte o similarmente trasmesse.

Il test di Turing si basa sul presupposto che una macchina si sostituisca ad A. Se la percentuale di volte in cui C indovina chi sia l'uomo e chi la donna è simile prima e dopo la sostituzione di A con la macchina, allora la macchina stessa dovrebbe essere considerata intelligente, dal momento che - in questa situazione - sarebbe indistinguibile da un essere umano.

Per *macchina intelligente* Turing ne intende una in grado di pensare, ossia capace di concatenare idee e di esprimerle. Per Turing, quindi, tutto si limita alla produzione di espressioni non prive di significato. Nell'articolo, riprendendo il *Cogito* cartesiano, si legge:

> « Secondo la forma più estrema di questa opinione, il solo modo per cui si potrebbe essere sicuri che una macchina pensa è quello di essere la macchina stessa e sentire se si stesse pensando. [...] Allo stesso modo, la sola via per sapere che un uomo pensa è quello di essere quell'uomo in particolare. [...] Probabilmente A crederà *"A pensa, mentre B no"*, mentre per B è l'esatto opposto *"B pensa, ma A no"*. Invece di discutere in continuazione su questo punto, è normale attenersi alla educata convenzione che ognuno pensi. »

Big Data is surely helping computers towards passing the Turing test!

The innovation cycle

Innovation follows three steps...

First: Innovation

Second: Democratization

Third: Responsability

Innovation cycle

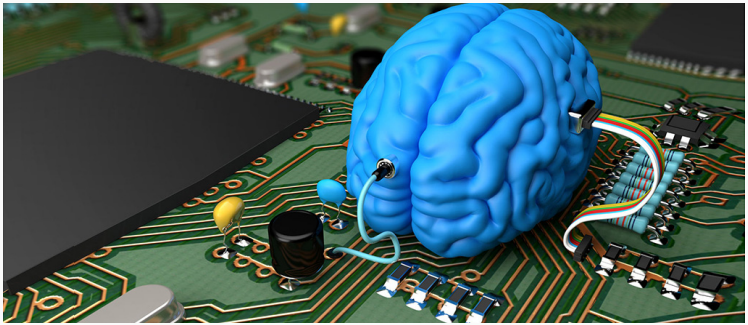
Innovation

Democratization

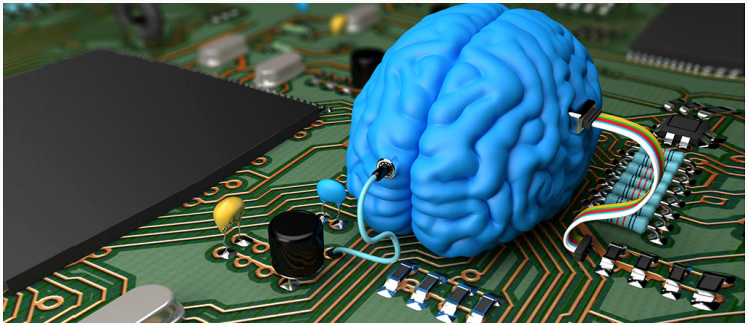
Responsability

Loop back and forth between the various phases

What about Artificial Intelligence?



What about Artificial Intelligence?



- We just passed the “democratization” phase
- Starting with the “responsability”
- … while continuing with the “innovation”

The new Big Data society

The Big Data based new society

First time in history

Human and machines to interact on a **cognitive** level

More integrated

⇒

More integrated

⇒

Huge social implications

More integrated

$\Rightarrow$

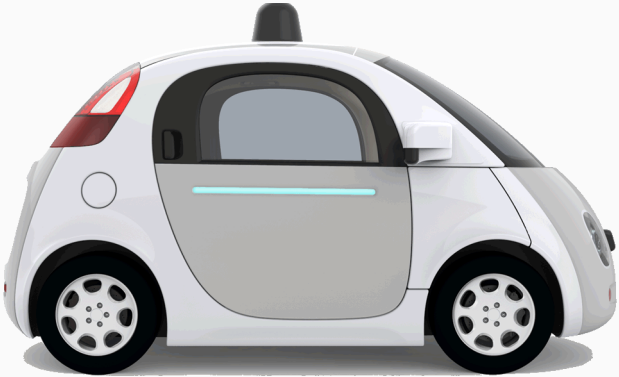
Huge social implications

“Democratization” and “Responsability” to be eavily reconsidered!

Can we really hyperconnect in real-time?

(how about this)

New society: Self driving cars



It’s happening now!

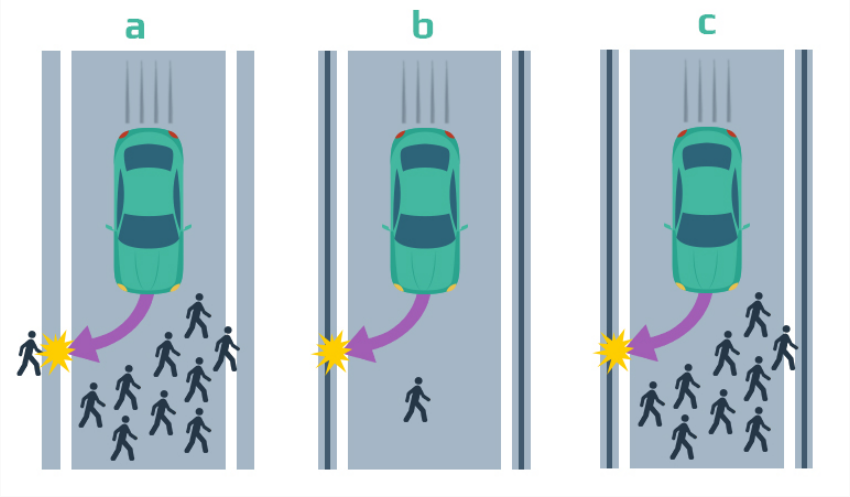
New society: Self driving cars

Great impacts

to this


ROAD TRAFFIC
INJURIES: THE FACTS
ROAD TRAFFIC INJURIES ARE THE
8th
LEADING CAUSE OF DEATH
1.35 MILLION
ROAD TRAFFIC DEATHS EVERY YEAR
1ST
CAUSE OF DEATH AMONG CHILDREN AGED 5-14 AND AMONG YOUNG ADULTS AGED 15-29
3%
28%
23%
54%
OF DEATHS ARE PEDESTRIANS, CYCLISTS AND MOTORCYCLISTS
LOW-INCOME COUNTRIES
1% OF THE WORLD'S VEHICLES 13% OF ALL DEATHS
HIGH-INCOME COUNTRIES
40% OF THE WORLD'S VEHICLES 7% OF ALL DEATHS
EVERY 24 SECONDS
SOMEONE DIES ON THE ROAD
#RoadSafety
World Health Organization
source: WHO Global status report on road safety 2018
www.who.int/violence_injury_prevention/road_safety_status/2018/en/

The ethical dilemma



Big Data all over the place

- Not only business: Big Data have implications far beyond marketing and consumer goods
- It will profoundly change how governments work and alter the nature of politics and our daily life too (e.g., smart cities).
- *"When it comes to generating economic growth, providing public services, or fighting wars, those who can harness big data effectively will have a significant edge over others."* [1]

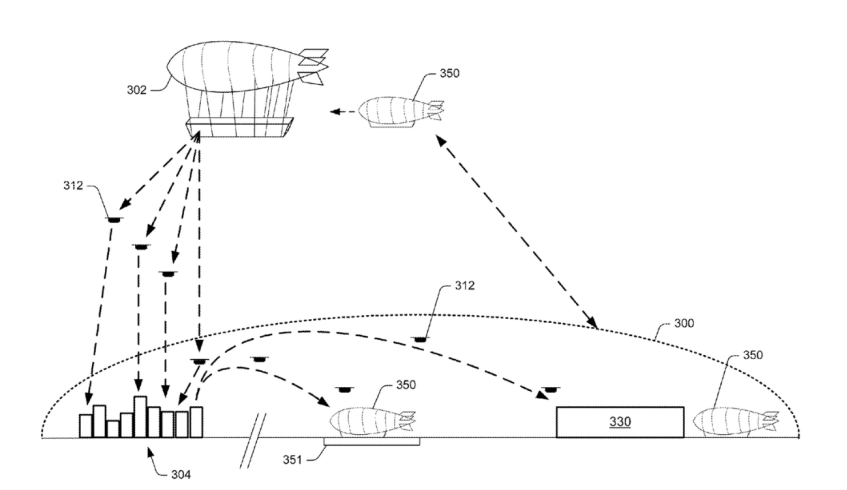
[1] The Rise of Big Data: *How It's Changing the Way We Think About the World*, Kenneth Cukier and Viktor Mayer-Schoenberger, Foreign Affairs

Big Data has consequences

Forbes thinks that Big Data will greatly influence us in 5 ways:

- how we spend
- how we vote
- how we study
- how we stay healthy
- how we keep/lose privacy

New society: Flying objects



Is it a “fair” barter?

“Data to the people”

- People are now *appreciating better* the (economical) value of personal data
- GDPR is a booster for this process
- Many new business initiatives are building on this
- Plan your company strategy to *share* that value with your customers

The risk of Big Data

(Cathy O'Neal Ted Talk)

Big data, big shifts

Big Data revolution implies big shifts

- From exact to approximate
- From sampling to all (n = All)
- From causality to correlation

From exact to approximate

- Increasing data size and speed leads to *inexactitude*
- Data in database are never *clean*: With small data we can afford to clean those
- Good enough is “good enough” with big data
- Willing to sacrifice a little accuracy in favor of general trends
- Big data transforms exact figures into probabilities... And this is ok in many cases!!

From exact to approximate

- *"Inexactitude"* difficult to digest for old school statisticians
  - They were fighting against mess
  - Data were scarce and every data point was useful
  - Wrong data could "deviate" the model
- Messiness is not inhereht to Big Data per se...
  - due to *imperfect* collection tools
    - perfect tools will lead to perfect Big Data
  - *For the time being this is what we have to deal with!*

From exact to approximate

- Try to have **precise** Big Data may not make economical sense
- We do not to live in a clean world of data
- Relational model still required
  - Banks
  - Legal
  - Landing on the moon
  - Etc.

From exact to approximate: technological implications

- Relational Model **cracks** under the Big Data weight
- Technology is adapting to messiness... Example: “NoSQL”
  - Designed to handle messy data
  - A variable can accept data of various type
  - **“Eventually consistent”** concept

From exact to approximate: technological implications

- Relational Model **cracks** under the Big Data weight
- Technology is adapting to messiness... Example: “NoSQL”
  - Designed to handle messy data
  - A variable can accept data of various type
  - **“Eventually consistent”** concept

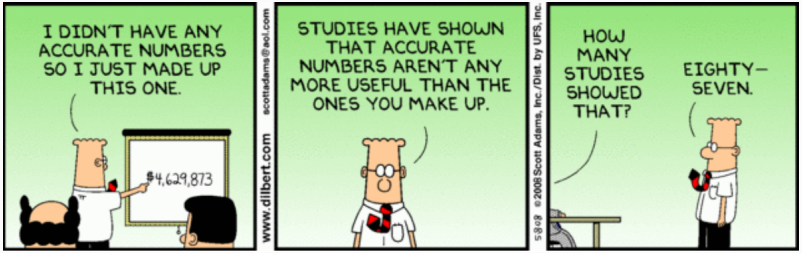
Back to the blackboard!

- Relational Model **cracks** under the Big Data weight
- Technology is adapting to messiness... Example: “NoSQL”
  - Designed to handle messy data
  - A variable can accept data of various type
  - **“Eventually consistent”** concept

Back to the blackboard!

- Hectic competition
- No standards yet
- It may take some time
- **Hadoop/MapReduce** model of distributing computing
  - Large number of cheap computers
  - Resistent to failure

n = All... No more sampling



- Statistics in the past hundred years based on *sampling*
  - find the **best**, **smallest**, and **most representative** sample
- This was accepted as a matter of life
- Reality: poor technology to process ALL data
- Some industries developed around this concept, e.g.: Surveys

n = All... some issues

- Sampling works well at macro-level
- Like a picture: good from the right distance, blurry when close
- In a sense, the sample is chosen depending on the ”distance”
- You may need to reprocess data with new samples in order to change ”distance”
- Sampling may not work well for *outliers detection*

From causality to correlation

- Big data is about *what* not *why*
- Stop searching for ”causality”
- Correlation doesn’t tell us why something is happening
- *If Millions of cancer patients who drinks orange juice and get one aspirin a day get better... do we really care why?*
- Prediction based on correlations is central in Big Data
- It leads to big (BIG) social implications

CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE

"Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all."

Harvard Business Review

MANAGING ORGANIZATIONS

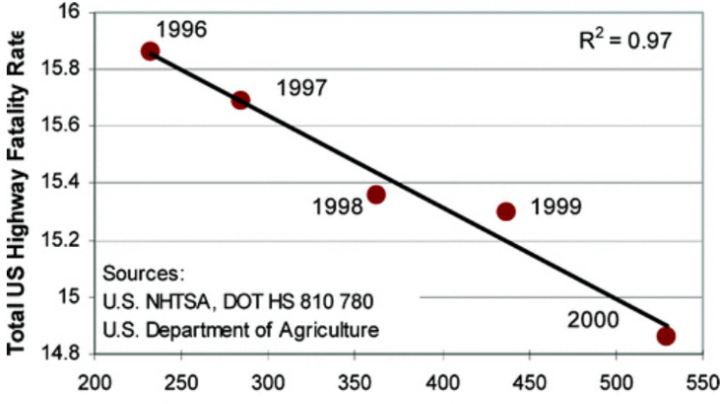
Beware Spurious Correlations

FROM THE JUNE 2015 ISSUE

We all know the truism "Correlation doesn't imply causation," but when we see lines sloping together, bars rising together, or points on a scatterplot clustering, the data practically begs us to assign a reason. We want to believe one exists.

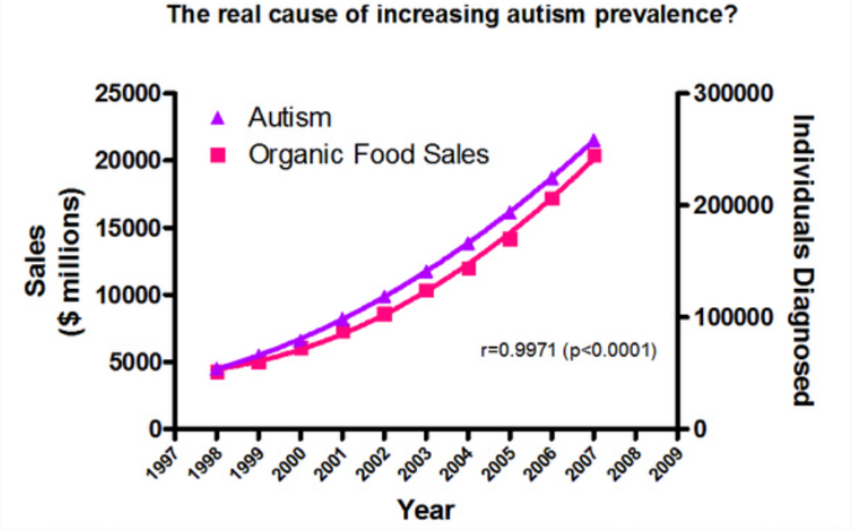
Statistically we can't make that leap, however. Charts that show a close correlation are often relying on a visual parlor trick to imply a relationship. Tyler Vigen, a JD student at Harvard Law School and the author of *Spurious Correlations,* has made sport of this on his website, which charts farcical

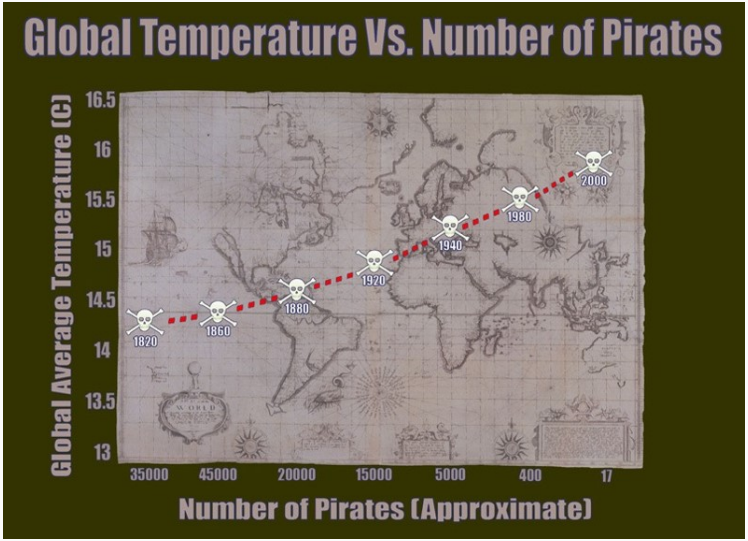
7. Mexican lemon imports prevent highway deaths.



Using Internet Explorer leads to murder.

. Eating organic food causes autism.




Global Temperature Vs. Number of Pirates
Global Average Temperature (C)
16.5
16
15.5
15
14.5
14
13.5
13
1820
1860
1880
1920
1940
1980
2000
35000
45000
20000
15000
5000
400
17
Number of Pirates (Approximate)

Some old-school Big Data based applications

Historical Google Flu Trend model

Historical Google Flu Trend model

2009: H1N1 new flu virus was discovered

Center for Disease Control and Prevention (CDC) asked doctors to inform CDC of any new case

Historical Google Flu Trend model

2009: H1N1 new flu virus was discovered

Center for Disease Control and Prevention (CDC) asked doctors to inform CDC of any new case

- Couple of weeks delay
  - People may wait before consulting a doctor
  - Doctors may take some time to inform CDC
  - CDC may take time to process data

Historical Google Flu Trend model

2009: H1N1 new flu virus was discovered

Center for Disease Control and Prevention (CDC) asked doctors to inform CDC of any new case

- Couple of weeks delay
  - People may wait before consulting a doctor
  - Doctors may take some time to inform CDC
  - CDC may take time to process data
- Google was able to predict flu spread in almost real-time
  - Based on geotargeting and search terms
  - Model based on past data from CDC about flu spread
  - Looking for correlation with search terms

Historical Google Flu Trend model

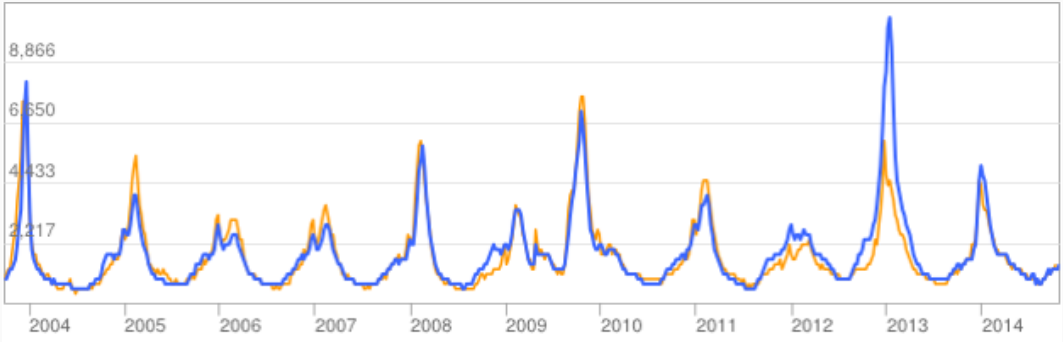
This is how Google describes it:

> *We have found a close relationship between how many people search for flu-related topics and how many people actually have flu symptoms. Of course, not every person who searches for "flu" is actually sick, but a pattern emerges when all the flu-related search queries are added together. We compared our query counts with traditional flu surveillance systems and found that many search queries tend to be popular exactly when flu season is happening. By counting how often we see these search queries, we can estimate how much flu is circulating in different countries and regions around the world.*

Big Data available to support research

- 50 Million most common search terms in the 2003-2008 time range
- Geolocalized
- Find a time-space correlation between search terms and infected areas provided by CDC
- Tested 450 Millions different mathematical models
- Outcome: 45 search terms model strongly correlated with flu spread
- As CDC but in real-time
- Others tried before but not much data was available

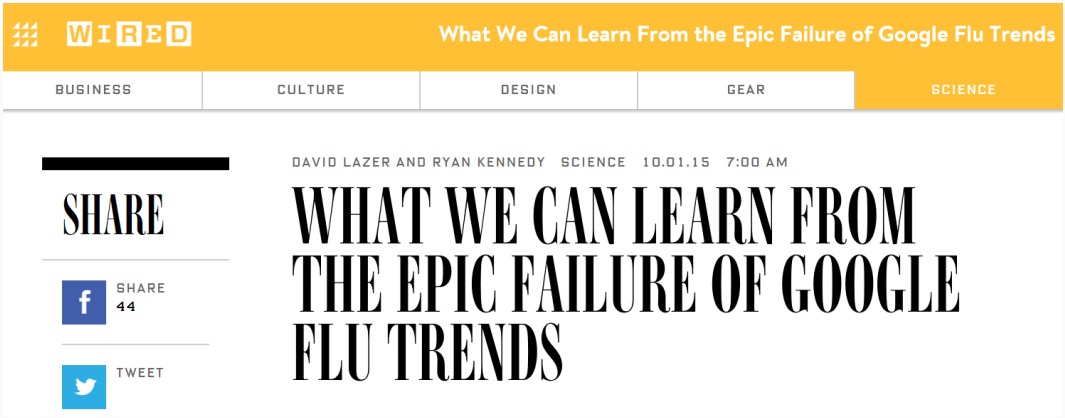
Historical Google Flu Trend model



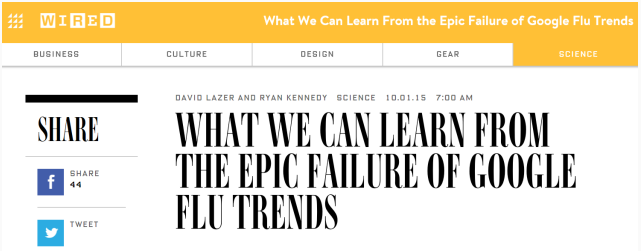
- Orange: US real data
- Blue: Google predictions based on keywords
- Published in Nature in 2009

... and then??

Historical failure

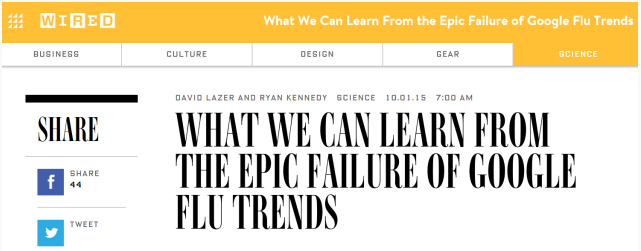


Historical failure



- “GFT failed, and failed spectacularly, missing at the peak of the 2013 flu season by 140 percent”
- GFT model started to deteriorate over time
- Google dismissed the project

Historical failure



- “GFT failed, and failed spectacularly, missing at the peak of the 2013 flu season by 140 percent”
- GFT model started to deteriorate over time
- Google dismissed the project

… simply a wrong model

Historical failure

GFT overestimated the prevalence of flu in the 2012-2013 and 2011-2012 seasons by more than 50%. From August 2011 to September 2013, GFT over-predicted the prevalence of the flu in 100 out 108 weeks. During the peak flu season last winter, GFT would have had us believe that 11% of the U.S. had influenza, nearly double the CDC numbers of 6%. If you wanted to project current flu prevalence, you would have done much better basing your models off of 3-week-old data on cases from the CDC than you would have been using GFT's sophisticated big data methods. "It's a Dewey beats Truman moment for big data," says David Lazer, a professor of computer science and politics at Northeastern University and one of the authors of the *Science* article.

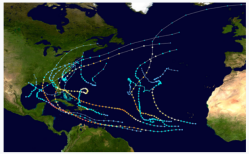
2004 Atlantic hurricane season

From Wikipedia, the free encyclopedia

The **2004 Atlantic hurricane season** was a very deadly, destructive, and hyperactive Atlantic hurricane season, with over 3,200 deaths and more than $61 billion (2004 USD) in damage. More than half of the 16 tropical cyclones brushed or struck the United States. Due to the development of a Modoki El Niño – a rare type of El Niño in which unfavorable conditions are produced over the eastern Pacific instead of the Atlantic basin due to warmer sea surface temperatures farther west along the equatorial Pacific – activity was above average. The season officially began on June 1 and ended on November 30, though the season's last storm, Otto, dissipated on December 3, extending the season beyond its traditional boundaries. The first storm, Alex, developed offshore of the Southeastern United States on July 31, one of the latest dates on record to see the formation of the first system in an Atlantic hurricane season. It brushed the Carolinas and the Mid-Atlantic, causing one death and $7.5 million (2004 USD) in damage.[nb 1] Several storms caused only minor damage, including tropical storms Bonnie, Earl, Hermine, and Matthew. In addition, hurricanes Danielle, Karl, and Lisa, Tropical Depression Ten, Subtropical Storm Nicole and Tropical Storm Otto had no effect on land while tropical cyclones.

There are four notable storms: Hurricane Charley, that made landfall in Florida as a Category 4 hurricane on the Saffir–Simpson hurricane wind scale (SSHWS), causing $16 billion in damage in the United States alone. Later in August, Hurricane Frances struck the Bahamas and Florida, causing at least 49 deaths and $10.1 billion in damage. The costliest and most intense storm was Hurricane Ivan. It was a Category 5 hurricane that devastated multiple countries adjacent to the Caribbean Sea, before entering the Gulf of Mexico and causing catastrophic damage on the Gulf Coast of the United States, especially in the states of Alabama and Florida. Throughout the countries it passed through, Ivan caused 129 fatalities and over $26.1 billion in damage. The deadliest storm was Hurricane Jeanne. In Haiti, torrential rainfall in the mountainous areas resulted in mudslides and severe flooding, causing at least 3,006 fatalities. Jeanne also struck Florida, inflicting extensive destruction. Overall, the storm caused at least $7.94 billion in damage and 3,042 deaths, ranking it as one of the deadliest Atlantic hurricanes in history.

2004 Atlantic hurricane season



Season summary map

Seasonal boundaries	
First system formed	July 31, 2004
Last system dissipated	December 3, 2004
Strongest storm	
Name	Ivan
• Maximum winds	165 mph (270 km/h) (1-minute sustained)
• Lowest pressure	910 mbar (hPa; 26.87 inHg)
Seasonal statistics	
Total depressions	16
Total storms	15
Hurricanes	9

What Wal-Mart Knows About Customers' Habits

By CONSTANCE L. HAYS NOV. 14, 2004

Correction Appended

HURRICANE FRANCES was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons, something that the company calls predictive technology.

A week ahead of the storm's landfall, Linda M. Dillman, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier. Backed by the trillions of bytes' worth of shopper history that is stored in Wal-Mart's computer network, she felt that the company could "start predicting what's going to happen, instead of waiting for it to happen," as she put it.

The experts mined the data and found that the stores would indeed need certain products – and not just the usual flashlights. “We didn’t know in the past that strawberry Pop-Tarts increase in sales, like seven times their normal sales rate, ahead of a hurricane,” Ms. Dillman said in a recent interview. “And the pre-hurricane top-selling item was beer.”

Learning sensitive data

THE HOME FRONT

How Target Found Out a Teen Was Pregnant Before Her Father Did

By NANCY FRENCH | February 21, 2012 7:51 PM

Forbes has an interesting piece that will make you think twice before sliding your debit card:

> Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

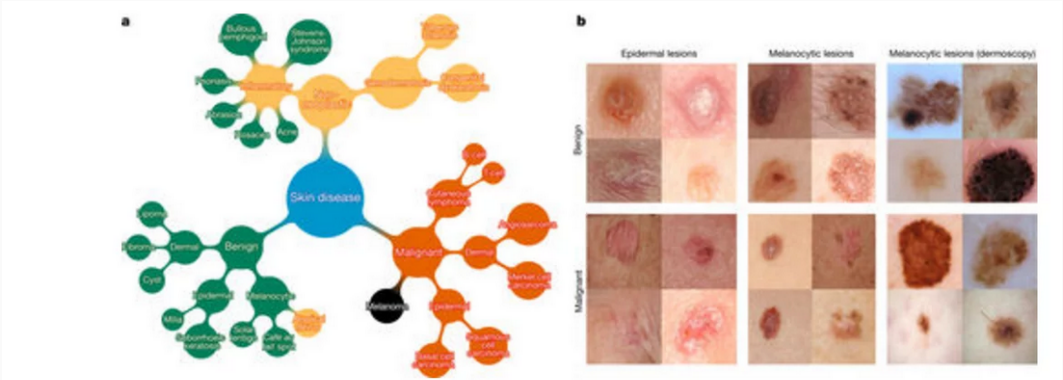
TOP STORIES

1. **Meghan McCain Is Right about AR-15 Confiscation, and You Know It**
   CHARLES C. W. COOKE

2. **Buttigieg Defends Abortion by Suggesting the Bible**

TARGET

AI & Heathcare

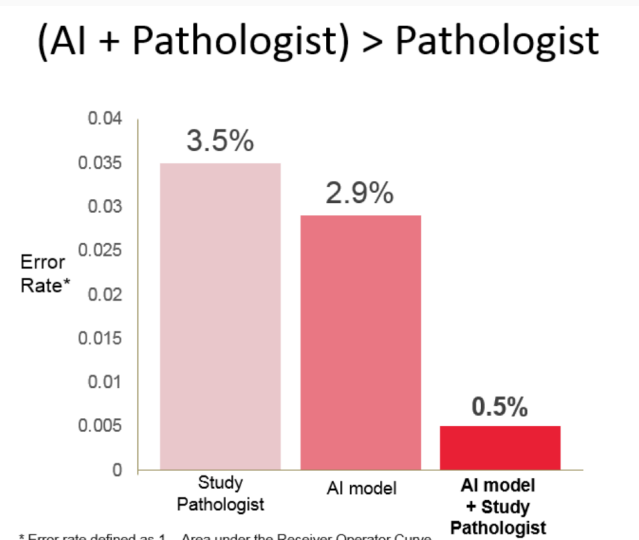


From the journal Nature: Dermatologist-level classification of skin cancer with deep neural networks

Stanford's deep learning algorithm was tested against 21 board-certified dermatologists who reviewed a reported 370 images and were asked if "they would proceed with biopsy or treatment, or reassure the patient" based on each image. Results showed that the algorithm had the same ability as the 21 dermatologists in determining the best course of action across all images.

AI & Healthcare - Breast cancer

The application of Artificial Intelligence (AI) technology with deep learning algorithms to whole-slide pathology images can potentially improve diagnostic accuracy of breast cancer and metastases (Lamy et al., 2019). Some scholars have assessed the performance of automated deep learning algorithms at detecting metastases in tissue sections of lymph nodes of women with breast cancer and compared results with pathologists' diagnoses. In the setting of a challenge competition, some deep learning algorithms achieved better diagnostic performance than a panel of 11 pathologists participating in a simulation exercise designed to mimic routine pathology workflow; in short, algorithm performance was comparable with an expert pathologist interpreting whole-slide images without time constraints (Ehteshami Bejnordi et al., 2017). Hence, this experiment has shown that deep learning algorithms could identify metastases in sentinel axillary lymph nodes slides with 100% sensitivity, whereas 40% of the slides without metastases could be identified as such. This approach can significantly reduce the workload of pathologists and improve the management decisions on whether or not to administer a therapy, perform a surgical intervention, etc. Overall, then, this interesting result shows the potential of

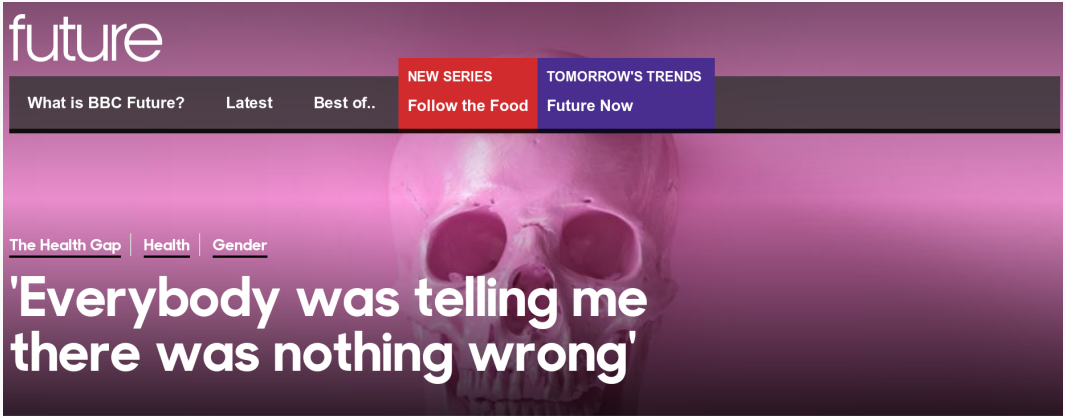

(AI + Pathologist) > Pathologist
Error Rate*
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
3.5%
2.9%
0.5%
Study Pathologist
AI model
AI model + Study Pathologist
* Error rate defined as 1 – Area under the Receiver Operator Curve
** A study pathologist, blinded to the ground truth diagnoses, independently scored all evaluation slides.
© 2016 PathAI

The real problem is “speed”

The real problem is “speed”

... early detection is crucial to save lives!!

In the U.S., there are approximately 5.4 million new skin cancer diagnoses each year and early detection is critical for a greater rate of survival. For example, early detection correlates with a 97 percent five-year survival rate but quickly decreases with later stages, hitting the 15-20 percent margin at stage IV. In 2017, an estimated 9,730 people will die of melanoma and one person dies of melanoma every 54 minutes.



Women are more likely to wait longer for a health diagnosis and to be told it's 'all in their heads'. That can be lethal: diagnostic errors cause 40,000-80,000 deaths in the US alone.

Healthcare - Improve diagnostic

Compared to many other diseases, diagnosing a brain tumour is fairly straightforward. Promptly detecting it comes down to being concerned enough about the early symptoms – which range from fatigue to seizures to personality change – to get an image of the brain. Either the tumour is there, or it isn't.

But in 2016, the Brain Tumour Charity released a report on **the treatment of brain tumour patients** in the United Kingdom. It found that almost one in three of them had visited a doctor more than five times before receiving their diagnosis. Nearly a quarter weren't diagnosed for more than a year.

Women, as well as low-income patients, experienced longer delays. They were more likely than men to see 10 or more months pass between their first visit to a doctor and diagnosis –and to have made more than five visits to a doctor prior to diagnosis.

Machine Learning for Medical Diagnostics: Insights Up Front

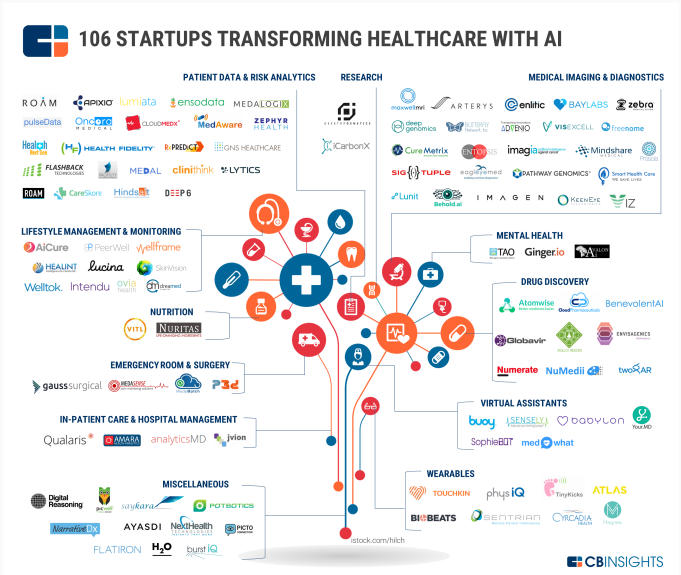
The Institute of Medicine at the National Academies of Science, Engineering and Medicine reports that "diagnostic errors contribute to approximately 10 percent of patient deaths," and also account for 6 to 17 percent of hospital complications. It is important to note that physician performance is typically not the direct cause of diagnostic errors. In fact, researchers attribute the cause of diagnostics errors to a variety of factors including:

- Inefficient collaboration and integration of health information technologies (Health IT)
- Gaps in communication among clinicians, patients and their families
- A healthcare work system which, by design, does not adequately support the diagnostic process

AI can surely help improving healthcare... two main directions:

- Improve diagnostic skills
- Patient relationship management & data collection

Healthcare & AI - a rich echosystem today



Unexpected applications of AI

AI Creativity

AI Creativity

The New York Times

AI Art at Christie's Sells for $432,500

"Edmond de Belamy, from La Famille de Belamy," by the French art collective Obvious, was sold on Thursday at Christie's New York. Christie's

Digital news

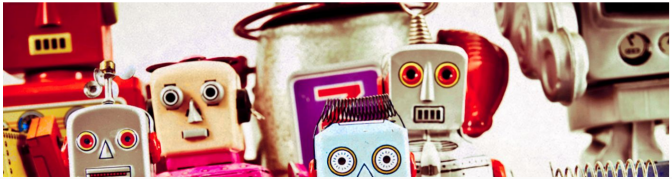
(click here)

Future Finance

Who to Sue When a Robot Loses Your Fortune

The first known case of humans going to court over investment losses triggered by autonomous machines will test the limits of liability.

By Thomas Beardsworth and Nishant Kumar
May 6, 2019, 2:00 AM GMT+2



Partners Inc. It's the first-known instance of humans going to court over investment losses triggered by autonomous machines and throws the spotlight on the "black box" problem: If people don't know how the computer is making decisions, who's responsible when things go wrong?

"People tend to assume that algorithms are faster and better decision-makers than human traders," said Mark Lemley, a law professor at Stanford University who directs the university's Law, Science and Technology program. "That may often be true, but when it's not, or when they quickly go astray, investors want someone to blame."

Raffaele Costa *Photographer: Andreas Rentz/Getty Images*

The new Big Data Science

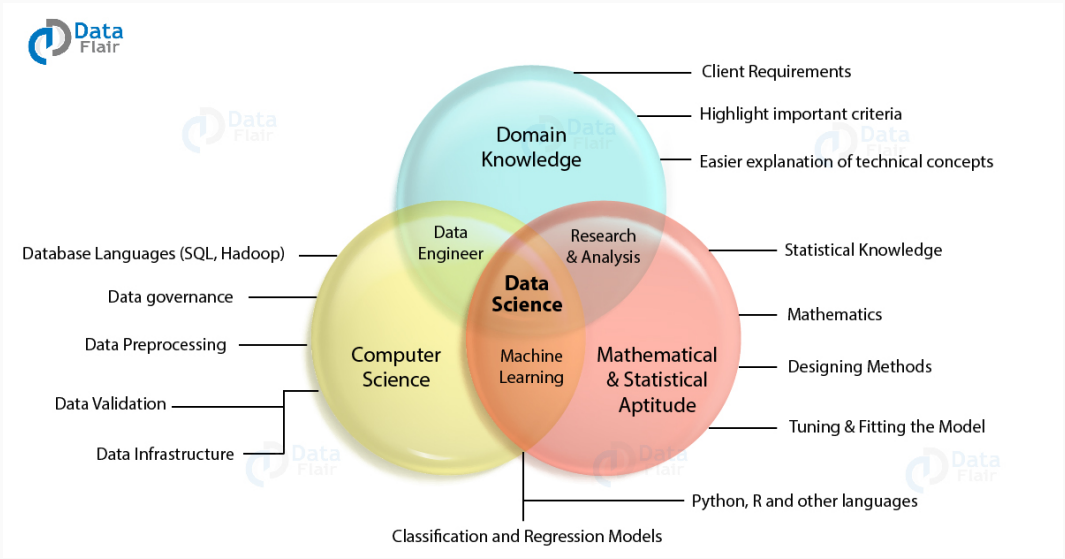

INTERNET
THINK
OUTSOURCE
ONLINE
50% SALE
60% SALE
INVEST
DEBIT
PAYMENT
MERCHANT
LIKE
LOADING
USER1
USER2
CLOUD
CONCEPT
SALE
SUCCESS
MARKET
ra2studio/ShutterStock.com

Big Data Science is a special mix

- **Science**: You can learn it
- **Crafts**: You can learn it
- **Creativity**: in part natural, in part it may came through experience
- **Common Sense**: You need to have/develop it

Big Data Science is multidisciplinary

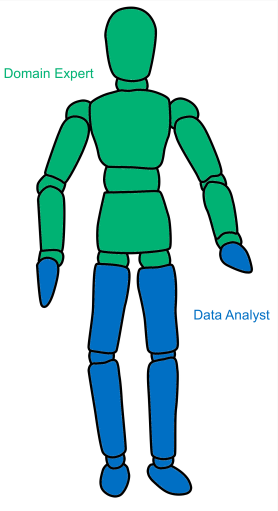
{width=50%}


Data Flair
Client Requirements
Highlight important criteria
Domain Knowledge
Easier explanation of technical concepts
Database Languages (SQL, Hadoop)
Data governance
Data Preprocessing
Data Validation
Data Infrastructure
Data Engineer
Research & Analysis
Statistical Knowledge
Data Science
Mathematics
Computer Science
Machine Learning
Mathematical & Statistical Aptitude
Designing Methods
Tuning & Fitting the Model
Python, R and other languages
Classification and Regression Models

Data Science is becoming ubiquitous

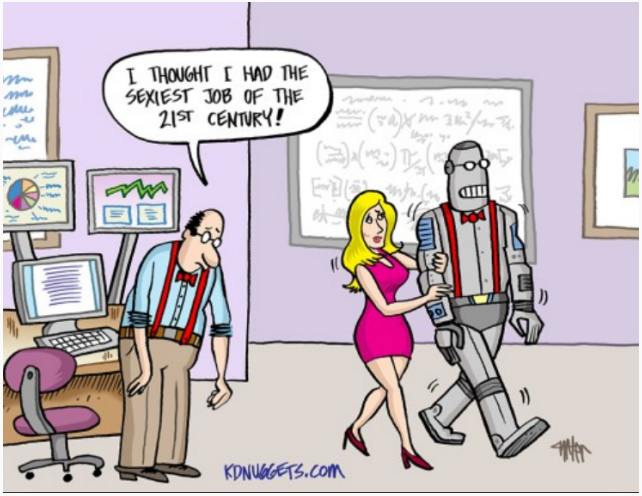
- Analytical thinking will be *pervasive* in companies
- Managers should have a *basic understanding of fundamental principles*
- Managers will lead data-analytics teams and data-driven projects
- Company's culture should support data-driven processes
- All aspects: from production to marketing to sales

The ideal data scientist profile



- *Mostly a domain expert (70%)*
- A strong background on data (30%)
- Ideally, *all managers should be like this*
- See A.I. like ”Excel”

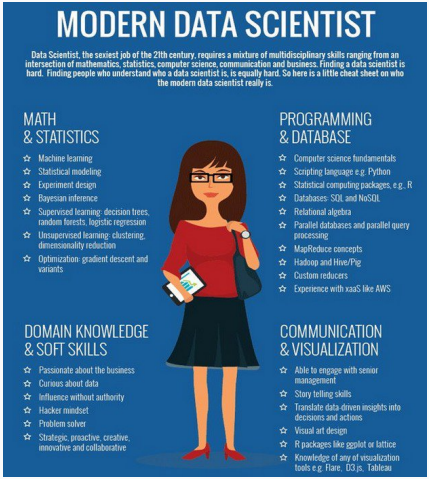
The new “Data Scientist” job profile

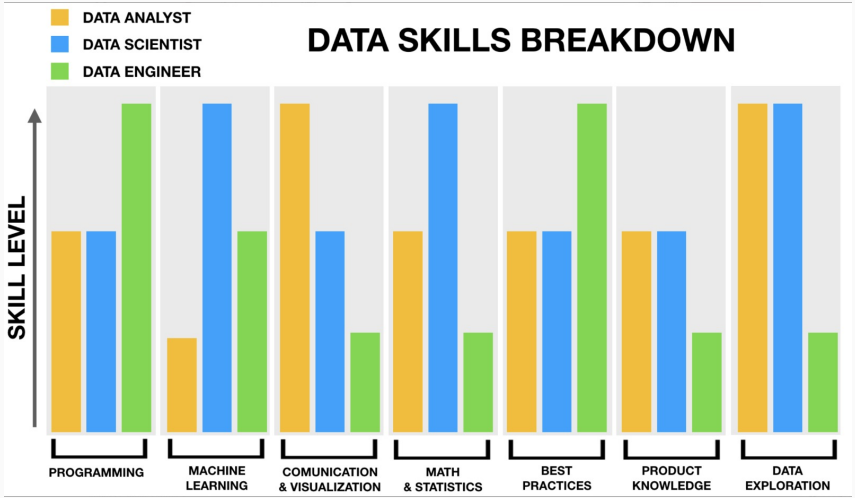


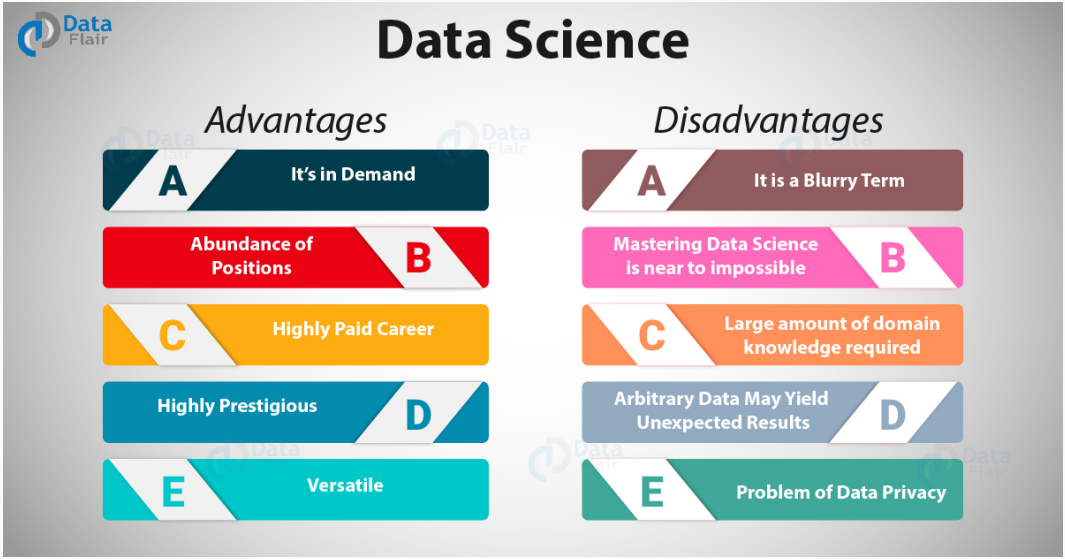
The “Data Scientist” must-have skills

- Business understanding
- Basic statistics
- Basic statistical programming: R and Python, SQL
- Machine learning concepts
- Data Visualization: Many tools available
- Thinking like a “Data Scientist”: see data everywhere
- Thinking like a problem solver
- Exceptional oral and written communication abilities
- Customer oriented approach

The data scientist skills




DATA ANALYST
DATA SCIENTIST
DATA ENGINEER
DATA SKILLS BREAKDOWN
SKILL LEVEL
PROGRAMMING
MACHINE LEARNING
COMUNICATION & VISUALIZATION
MATH & STATISTICS
BEST PRACTICES
PRODUCT KNOWLEDGE
DATA EXPLORATION


Data Flair
Data Science
Advantages
A
It's in Demand
Abundance of Positions
B
C
Highly Paid Career
Highly Prestigious
D
E
Versatile
Disadvantages
A
It is a Blurry Term
Mastering Data Science is near to impossible
B
C
Large amount of domain knowledge required
Arbitrary Data May Yield Unexpected Results
D
E
Problem of Data Privacy

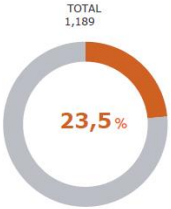
The new company and the Big Data

How to take full advantage of Big Data and A.I. today?

How to take full advantage of Big Data and A.I. today?

TO WHAT EXTENT DOES YOUR COMPANY HAVE A CLEAR STRATEGY ON BIG DATA?

Figure 6: World companies adopting a strategy on big data

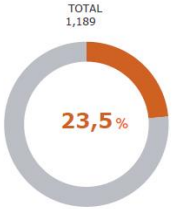


Europe	Asia	North America	Central South America	Large Companies >1000 employees	Leaders
n=368	n=615	n=82	n=65	n=207	n=82
23,9%	24,9%	19,5%	16,9%	39,6%	100,0%

How to take full advantage of Big Data and A.I. today?

TO WHAT EXTENT DOES YOUR COMPANY HAVE A CLEAR STRATEGY ON BIG DATA?

Figure 6: World companies adopting a strategy on big data



Europe	Asia	North America	Central South America	Large Companies >1000 employees	Leaders
n=368	n=615	n=82	n=65	n=207	n=82
23,9%	24,9%	19,5%	16,9%	39,6%	100,0%

It is MUCH MORE than just technology!

“If you went to bed last night as an industrial company, you’re going to wake up today as a software and analytics company”

Jeff Immelt (chairman and CEO of GE)

Three main reasons

Three main reasons

- Cultural resistance

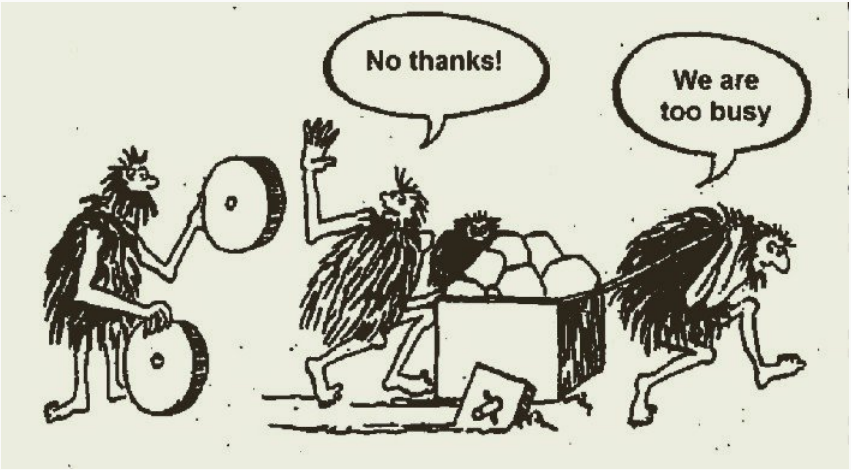
Three main reasons

- Cultural resistance
- A.I. is not a tool! It’s a mindset

Three main reasons

- Cultural resistance
- A.I. is not a tool! It’s a mindset
- Lack of necessary skills on the market

#1 Cultural resistance



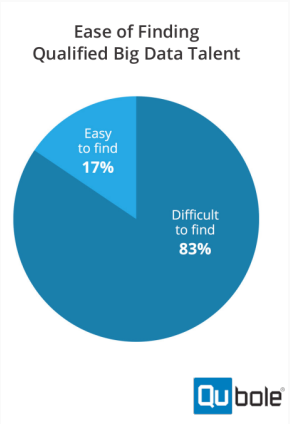
Imprese alla ricerca di un milione di scienziati dei dati. Luca Tremolada. Il Sole 24 ore, Marzo 2019:

> In ogni settore servirà sempre di più un esperto in grado di formulare le giuste domande a grandi moli di dati. In questo senso la data science non è più una disciplina o un insieme di discipline che vivono dentro le divisioni IT di una azienda. Ma diventa una forma di sapere diffuso. In un futuro sempre più vicino non esisterà l'idea di un professionista che risponde a domande come fosse un elaboratore elettronico. Ogni settore del sapere imparerà a interrogare i dati del proprio universo di competenza per prevedere e comprenderne meglio fenomeni e dinamiche.

#2 A.I. is not a tool! It's a mindset

- Change has to come from inside
- You cannot buy an A.I. tool to solve your problems... yet
- A.I. is very vast! No single supplier can offer everything

#3 Lack of necessary skills on the market



- Data science is not a common background… yet
- Ability to see business through the data glasses
- Not necessarily *data miners* are the right answer
- Universities are starting now

A modest advice

- Start building a unified company-wide infrastructure to collect and process data: *Data Lake*

A modest advice

- Start building a unified company-wide infrastructure to collect and process data: *Data Lake*
- Make it *easily sharable* with outside A.I. suppliers

A modest advice

- Start building a unified company-wide infrastructure to collect and process data: *Data Lake*
- Make it *easily sharable* with outside A.I. suppliers
    - Be careful: Data transfer process issues and privacy concerns arise

A modest advice

- Start building a unified company-wide infrastructure to collect and process data: *Data Lake*
- Make it *easily sharable* with outside A.I. suppliers
  - Be careful: Data transfer process issues and privacy concerns arise
  - You may need more than one supplier of A.I. solutions

A modest advice

- Start building a unified company-wide infrastructure to collect and process data: *Data Lake*
- Make it *easily sharable* with outside A.I. suppliers
  - Be careful: Data transfer process issues and privacy concerns arise
  - You may need more than one supplier of A.I. solutions
- Start diffusing the *data culture* in your company... please remember that:

A modest advice

- Start building a unified company-wide infrastructure to collect and process data: *Data Lake*
- Make it *easily sharable* with outside A.I. suppliers
  - Be careful: Data transfer process issues and privacy concerns arise
  - You may need more than one supplier of A.I. solutions
- Start diffusing the *data culture* in your company... please remember that:
  - A.I is a mindset not a tool

A modest advice

- Start building a unified company-wide infrastructure to collect and process data: *Data Lake*
- Make it *easily sharable* with outside A.I. suppliers
  - Be careful: Data transfer process issues and privacy concerns arise
  - You may need more than one supplier of A.I. solutions
- Start diffusing the *data culture* in your company... please remember that:
  - A.I is a mindset not a tool
  - It will take a *significant* amount of time

A modest advice

- Start building a unified company-wide infrastructure to collect and process data: *Data Lake*
- Make it *easily sharable* with outside A.I. suppliers
  - Be careful: Data transfer process issues and privacy concerns arise
  - You may need more than one supplier of A.I. solutions
- Start diffusing the *data culture* in your company... please remember that:
  - A.I is a mindset not a tool
  - It will take a *significant* amount of time
- Consider **sharing the value of data** with your customers as part of your data driven strategy

Learning Models

Type of learning

Mostly, two different learning paradigms:

- Supervised
- Unsupervised

Supervised learning

- Data are *labelled*
- Labels are the targets (or output, or class): what we want to *learn*
- So, for each observation we have:
  - Input values
  - Label

Supervised learning

- Data are *labelled*
- Labels are the targets (or output, or class): what we want to *learn*
- So, for each observation we have:
  - Input values
  - Label

The machine learning algorithm learns such associations over time

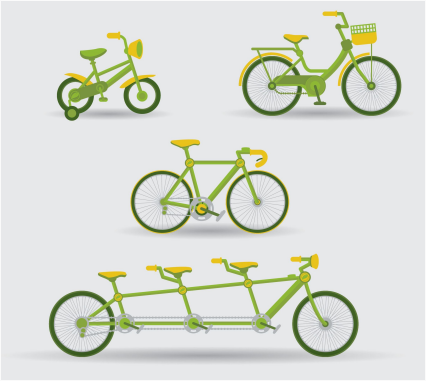
Supervised learning - example

We want to teach a small kid how to distinguish a *bike* from a *car*

He has not ever seen those before

Input = *a set of labelled images*

Supervised learning - example

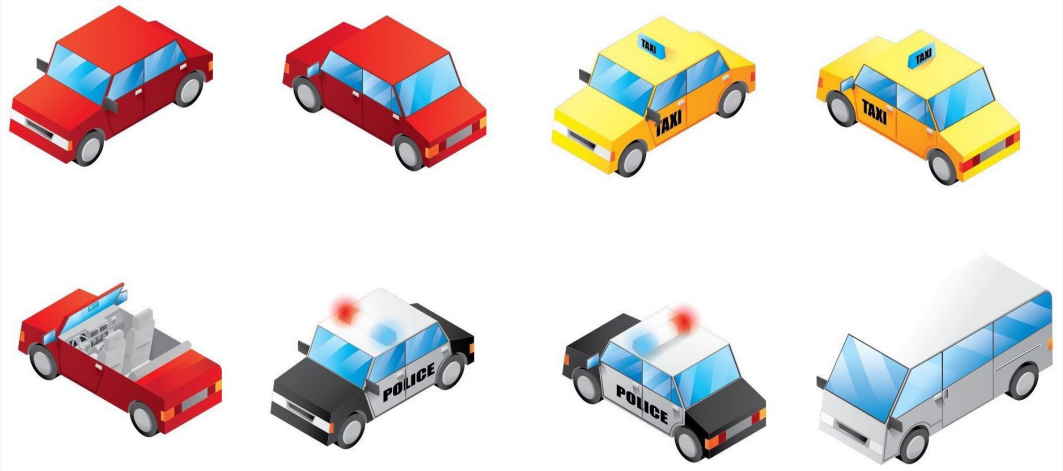


Lets proceed as follows:

1. Let’s show the images of the bikes
2. We tell him those are ”bikes”
3. We do not teach him about any specific characteristic

So we let the kid analyse those images to understand what makes those objects a “bike”

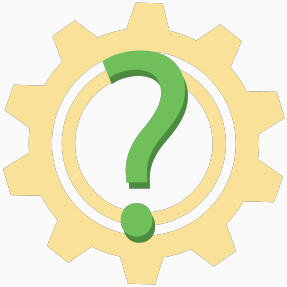
We do the same with the **cars**



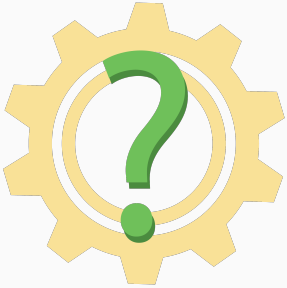
We let him “think and learn”

Supervised learning - example

Eventually, we show him a picture and ask him to identify it



Eventually, we show him a picture and ask him to identify it



Notice: It’s a new picture, he has not seen it before

Unsupervised learning

- Here the algorithm learns without any label
- The input to the algorithm is just a set of observations

- Here the algorithm learns without any label
- The input to the algorithm is just a set of observations

In general, this is a more challenging class of problems

Unsupervised learning - Example

- Let's repeat the previous example with no *supervision*

Unsupervised learning - Example

- Let's repeat the previous example with no *supervision*
- This time we show the kid the images at once, *bikes and cars together*

Unsupervised learning - Example

- Let's repeat the previous example with no *supervision*
- This time we show the kid the images at once, *bikes and cars together*
- **We don't tell him anythig about the two type of objects!**

Unsupervised learning - Example

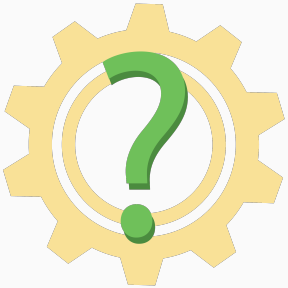
The kid has to learn by himself the two categories and what makes those different from each other

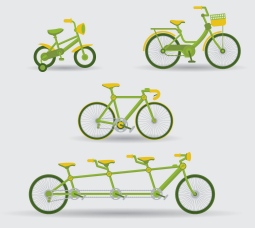
Unsupervised learning - Example

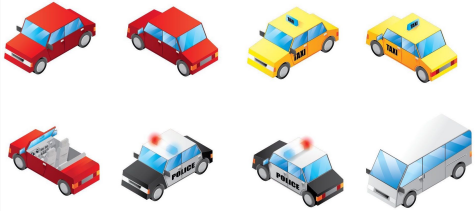
The kid has to learn by himself the two categories and what makes those different from each other

He will use a different logical path to *cluster* the input images

Then, like before, we show him a new **unseen** image







The kid in his learning may use

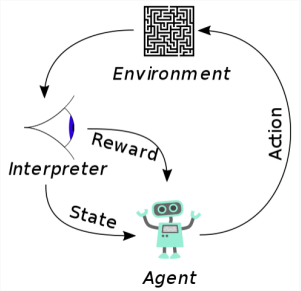
- more than two categories
- a very different set of categories of what we expect

For instance, he may decide to put together objects based on color, size, or number of wheels (that he sees!)

Unsupervised learning - Considerations

- The results is greatly dependent upon the quality of the input images
- As usual, *the more data in input the more accurate the learning*, at least, until a certain point

Reinforcement learning - The basic principle



- Learning is based on a *gain function*
- Each time the *machine* reachs a positive state it gains something
- The obectivive is to *maximize gain*
- Used by DeepMind to develop AlphaGo

The basic of “Modeling”

We do it *naturally* in life, maybe without knowing it

- Basic task for ”data miners”
- Statisticians have been doing it for many years
- It takes many different forms
- Today, *all managers should have at least a basic understanding*

The art of modeling: Classification

Modeling: A simple “supervised” example

Age	Income	Out
30	65k	Y
68	83k	Y
43	61k	N
30	25k	Y
51	82k	N
78	67k	Y

Modeling: A simple “supervised” example

Age	Income	**Out**
30	65k	Y
68	83k	Y
43	61k	N
30	25k	Y
51	82k	N
78	67k	Y

Goal: **Build a model to predict the *dependent* variable** Out

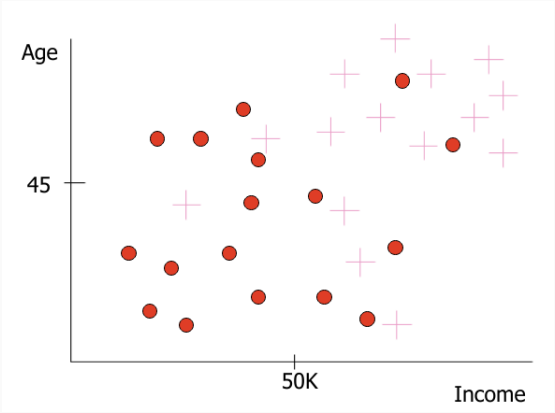
Modeling: A simple “supervised” example

Age	Income	Out
30	65k	Y
68	83k	Y
43	61k	N
30	25k	Y
51	82k	N
78	67k	Y

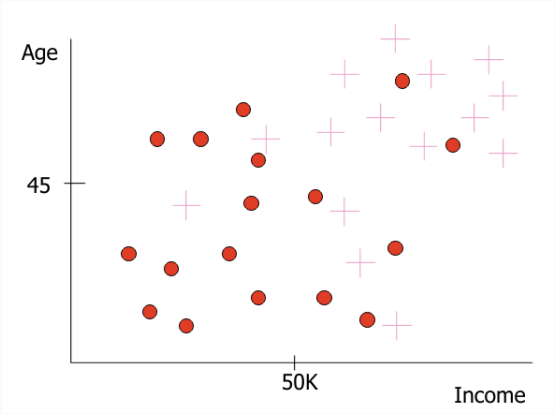
Goal: **Build a model to predict the *dependent* variable Out**

- Out is a categorical variable
- Age and $Income$ are the *independent* variables

Let's model

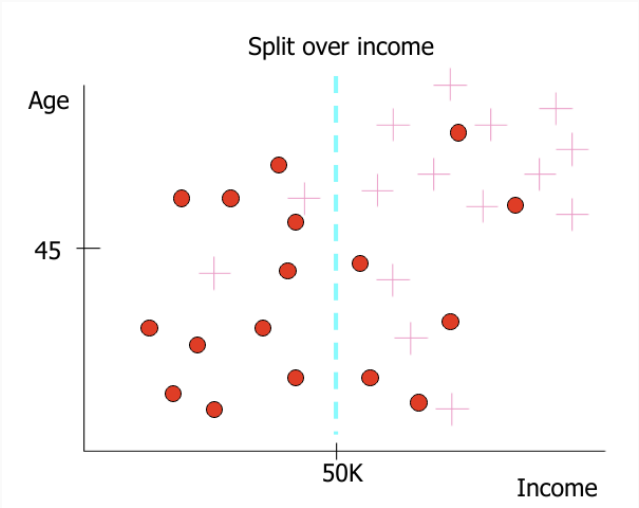


- We project the *dependent* variable Out, as $+$ and $\circ$, on a two-dimensional space

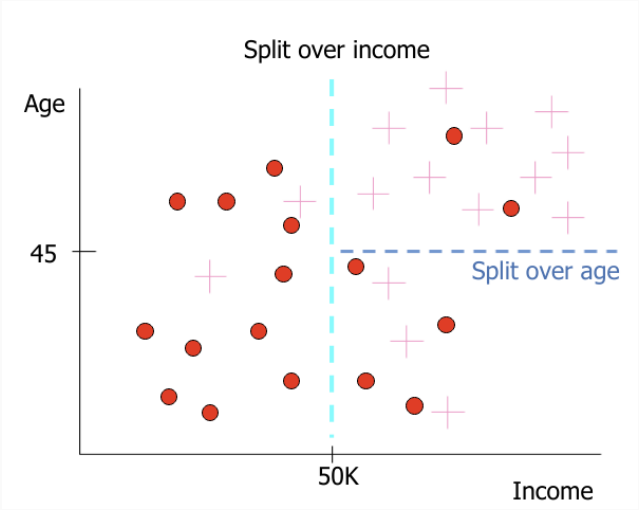


Do we notice anything?

Let’s model

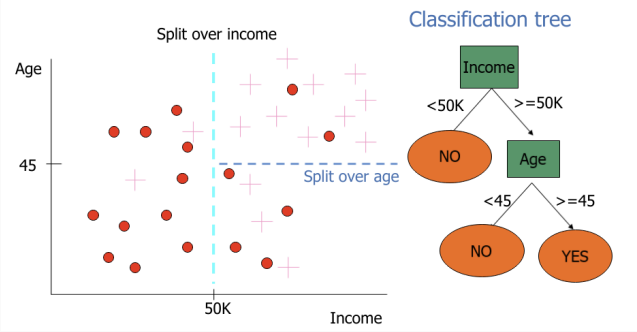


Let’s model



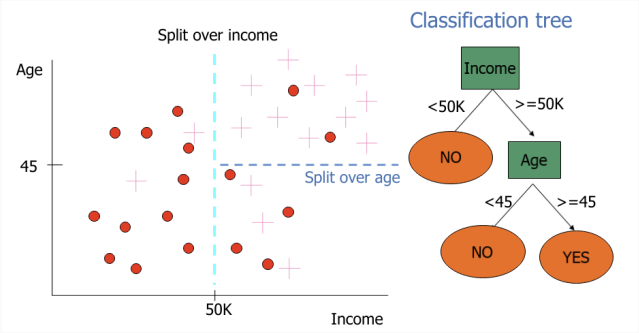
Let's model

- We can model it through a **Classification tree**



Let's model

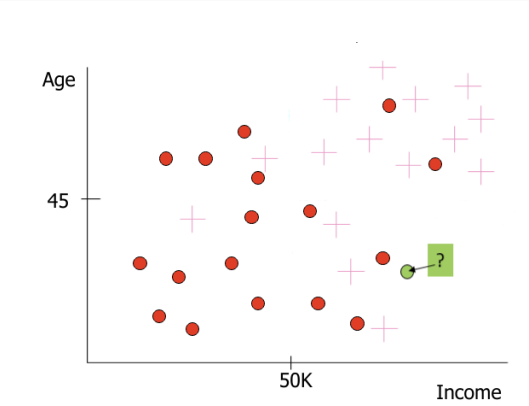
- We can model it through a **Classification tree**



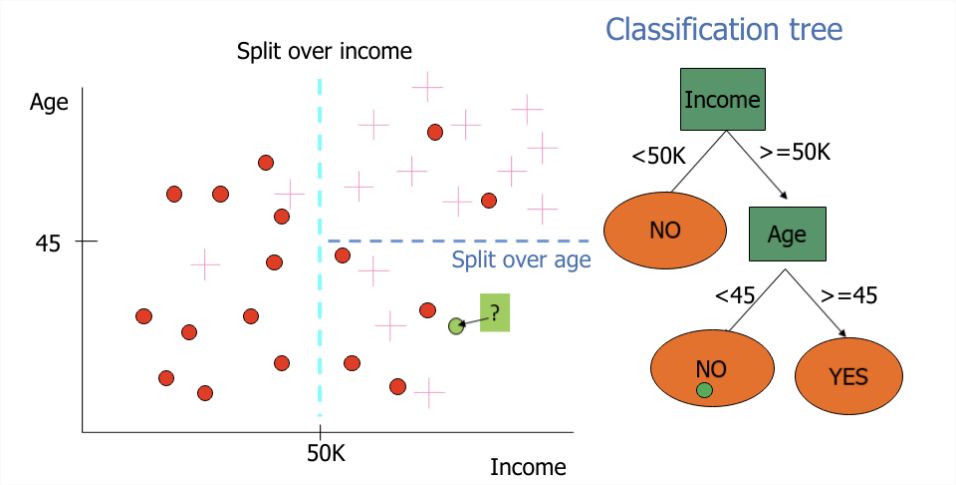
- The algorithm **finds** the optimal splits
- It maximizes **prediction confidence**

Let's apply the model now

- Let's now **generalize** the model
- Say, we receive new data and we want to predict the Out variable
  - For instance, a new person 25 years old and with an income of 70k

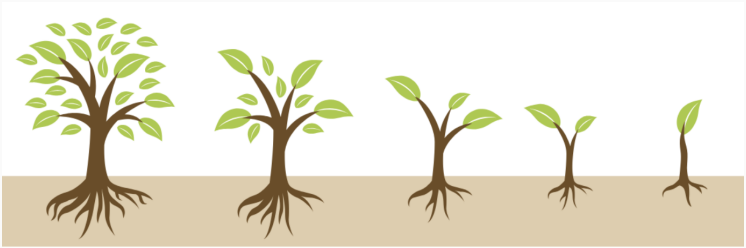


We can now predict



Classification tree considerations

- **C4.5** was the original idea
- Old technique very well established
- Works for *categorical* dependent variable
- It works with both *continuos* and *categorical* independent variables
- Fast to execute
- Easy to find free code on the web



The art of modeling: Clustering

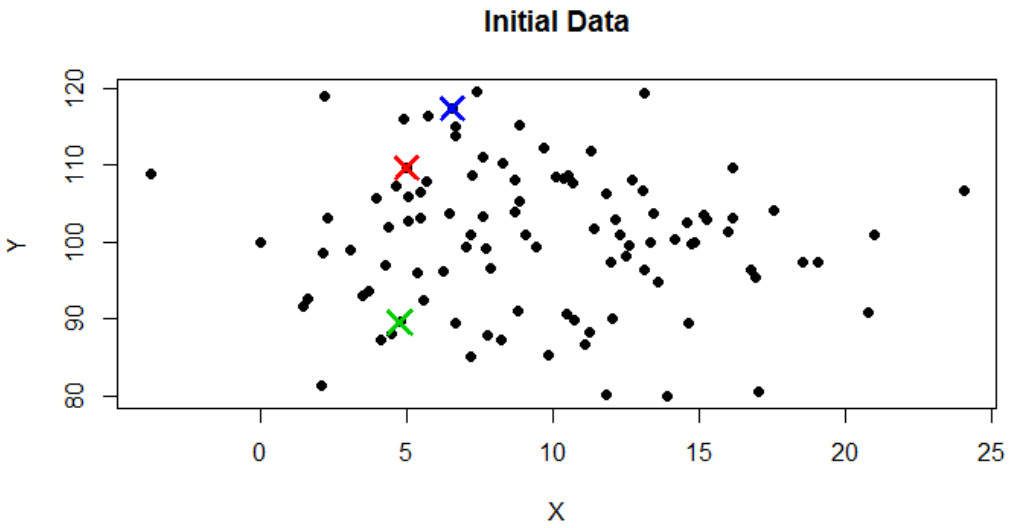
Clustering

- **Unsupervised** learning model
- Widely used in business/marketing applications
- Gives a structure to unsorted data points
- In simpler words: Aggregate *similar* items

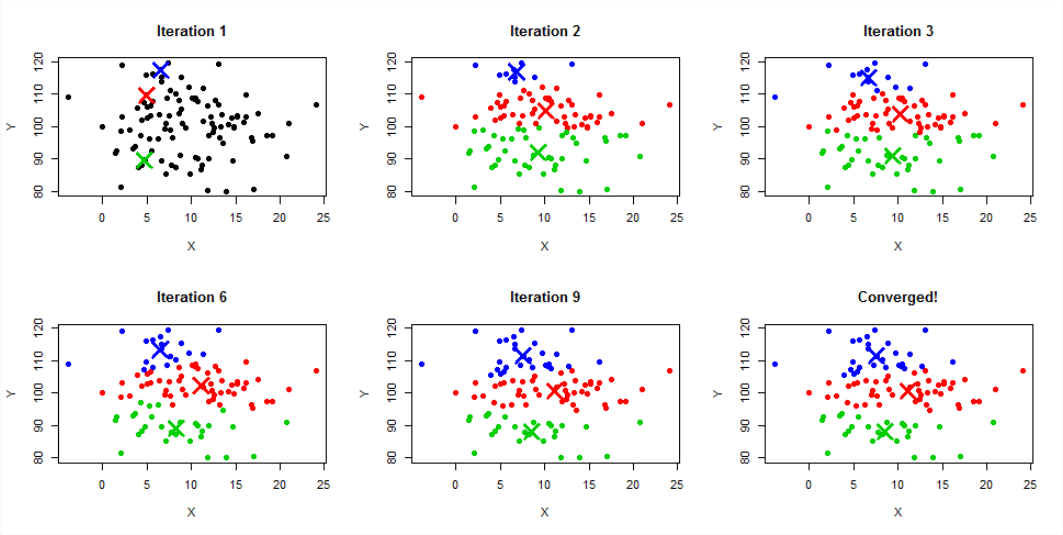
k-means Clustering algorithm

- **Step 0**: Initialize K random centroids (*just pick randomly K data points)*
- **Step 1** For every data point:
  - *assign* it to the closest centroid (*any distance metric works*);
- **Step 2** For every centroid
  - move the centroid to the *average* among all its points;
- Repeat **Step 1** and **Step 2** until all centroids do not change anymore;

The algorithm **converged**


Initial Data
Y
120
110
100
90
80
0
5
10
15
20
25
X

k-means example... Iterations



k-means, some considerations

- Easy to apply
- Various algorithms available
  - You can find free, ready to use, code on the web
- Not suitable for categorical variables
- Need to *normalize* variable for scale uniformity
  - Otherwise, distance calculation may be affected
- It scales on Big Data, that is, it can be *parallelized*
- How to pick initial K?

The art of modeling: Associations

What can we infer just by observing?

Market-basket analysis: Understanding meaningful *patterns* by analysing baskets

A *basket* is a generic set of items

- **Pattern**: A set of items
- **Frequent pattern**: A pattern that appears *frequently*
- We infer rules such as: $A, B, ..., C \implies X$
- *Beer and diapers on friday evening?!*
- It may depend on the context: "Have kids?", "Travelling for work?", etc.

Metrics: Support and Confidence

- We are only interested in rules with *high support*
  - Statistically meaningful
  - *support* $s(X \implies Y)$ = "# of transactions containing both X and Y"

Metrics: Support and Confidence

- We are only interested in rules with *high support*
  - Statistically meaningful
  - *support* $s(X \implies Y)$ = "# of transactions containing both X and Y"

- We are only interested in rules with *high confidence*
- *Confidence* $c(X \implies Y)$ = "conditional probability that a transaction containing X also contains Y"

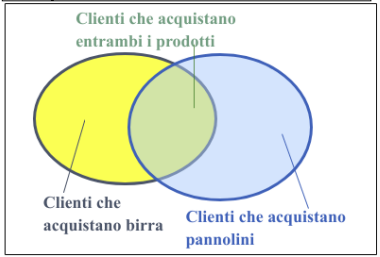
Metrics: Support and Confidence

- We are only interested in rules with *high support*
  - Statistically meaningful
  - *support* $s(X \implies Y)$ = "# of transactions containing both X and Y"

- We are only interested in rules with *high confidence*
- *Confidence* $c(X \implies Y)$ = "conditional probability that a transaction containing X also contains Y"

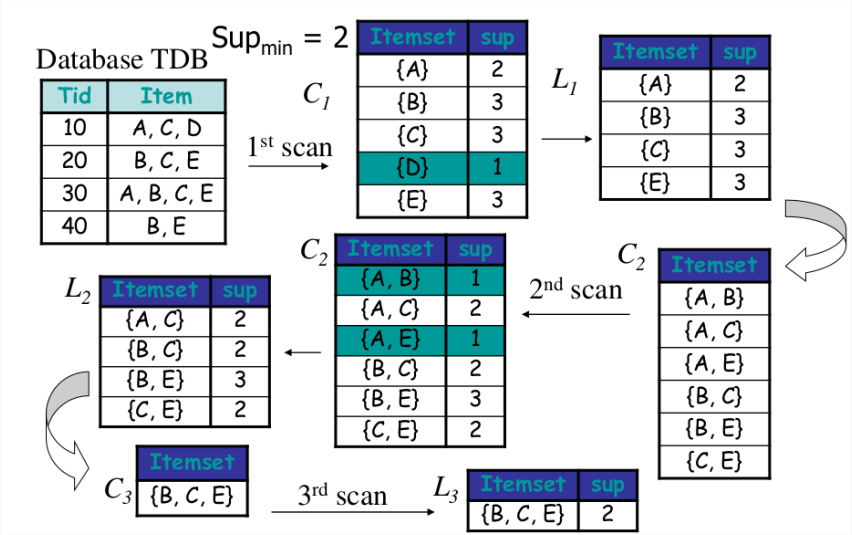
- Be careful: The rule $milk, butter \implies bread$ may derive from the fact that many baskets contain $bread$
- Basically, we need to select rules such as:
  - $c(milk, butter \implies bread) \gg c(bread)$

Tid	Prodotti acquistati
10	Birra, Noccioline, Pannolini
20	Birra, Caffè, Pannolini
30	Birra, Pannolini, Uova
40	Noccioline, Uova, Latte
50	Noccioline, Caffè, Pannolini, Uova, Latte



- Find all $X \implies Y$
- Supports in the given example:
  - $s(Birra) = 3$
  - $s(Noccioline) = 3$
  - $s(Pannolini) = 4$
  - $s(Uova) = 3$
  - $s(Birra, Pannolini) = 3$

- $Birra \implies Pannolini(3, 100\%)$
- $Pannolini \implies Birra(3, 75\%)$
- $Latte \implies Uova(?, ?)$
- $Noccioline \implies Latte(?, ?)$


Database TDB
Tid | Item
10 | A, C, D
20 | B, C, E
30 | A, B, C, E
40 | B, E
Sup_min = 2
C1
Itemset | sup
{A} | 2
{B} | 3
{C} | 3
{D} | 1
{E} | 3
1st scan
L1
Itemset | sup
{A} | 2
{B} | 3
{C} | 3
{E} | 3
C2
Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}
2nd scan
C2
Itemset | sup
{A, B} | 1
{A, C} | 2
{A, E} | 1
{B, C} | 2
{B, E} | 3
{C, E} | 2
L2
Itemset | sup
{A, C} | 2
{B, C} | 2
{B, E} | 3
{C, E} | 2
C3
Itemset
{B, C, E}
3rd scan
L3
Itemset | sup
{B, C, E} | 2

Association rules applications

Association rules are used in many different contexts:

- Combined promotions
- Optimized shelf allocation
- Text documents based on shared concepts
- Targeted promotions of movies/books/articles/etc
- CTR banner optimization
- Mechanical fault prevention
- Medical diagnosis
- etc.

Thanks