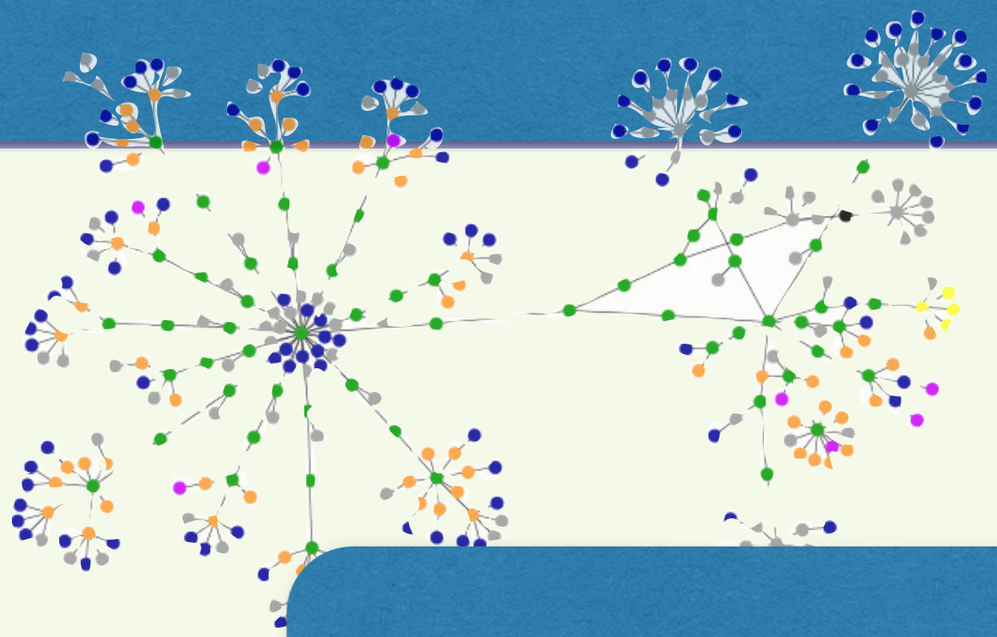
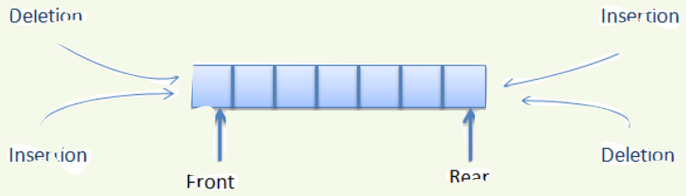
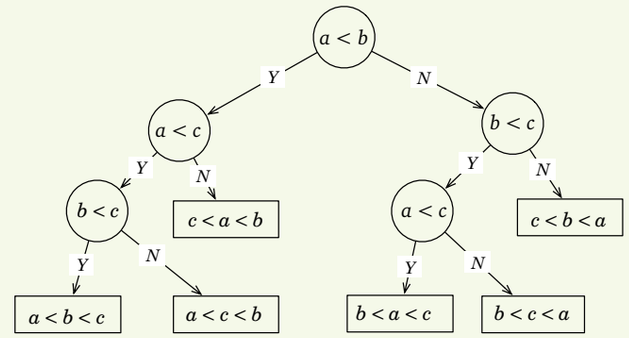






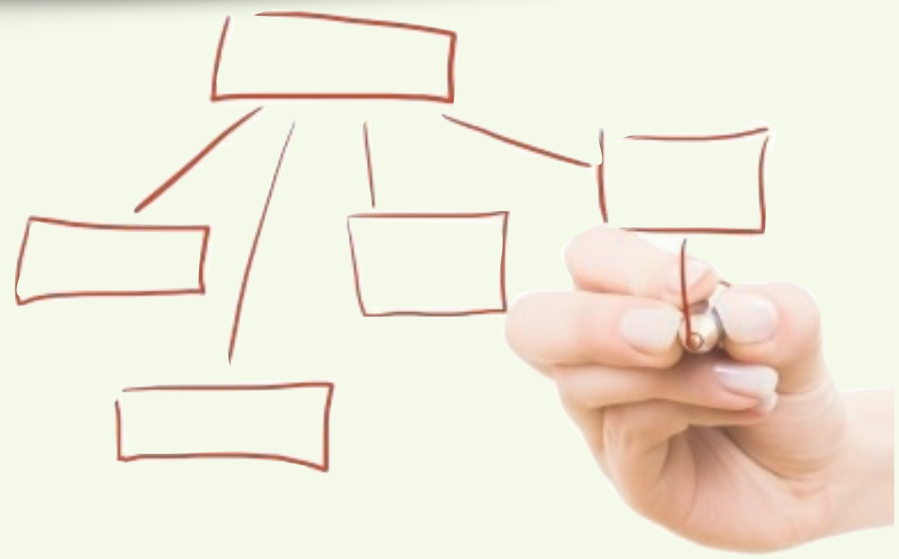
COURSE
"Técnicas Matemáticas para Big Data"
University of Aveiro
2025/2026

Algorithm 2.1: INSERTION-SORT(A)

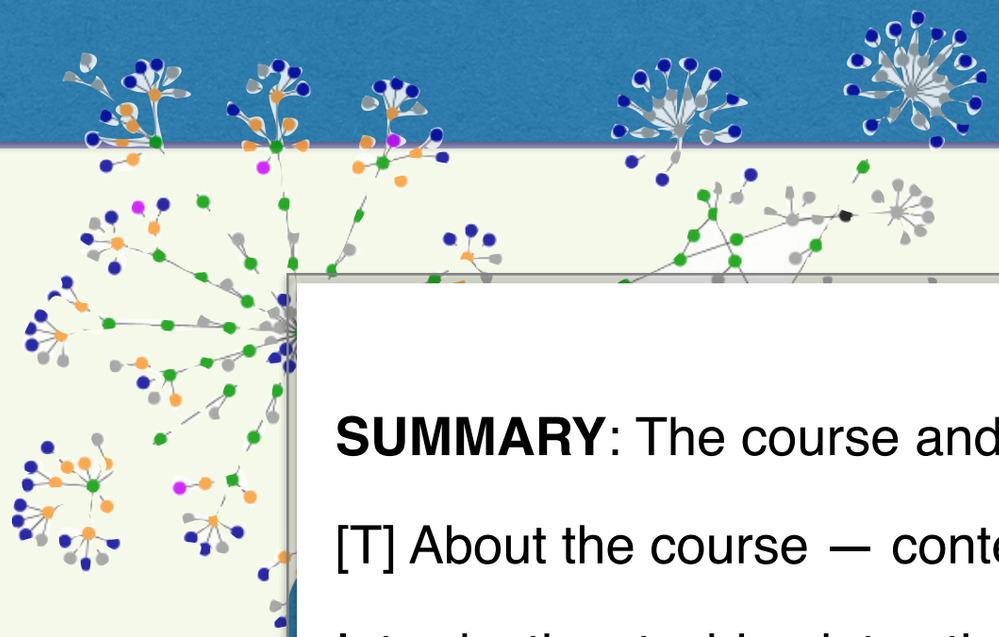
```
1  for j ← 2 to A.size do
2      key ← A[j]
       // Insert A[j] into the sorted sequence A[1 . . j − 1]
3      i ← j − 1
4      while i > 0 and A[i] > key do
5          A[i + 1] ← A[i]
6          i−−
7      A[i + 1] ← key
```

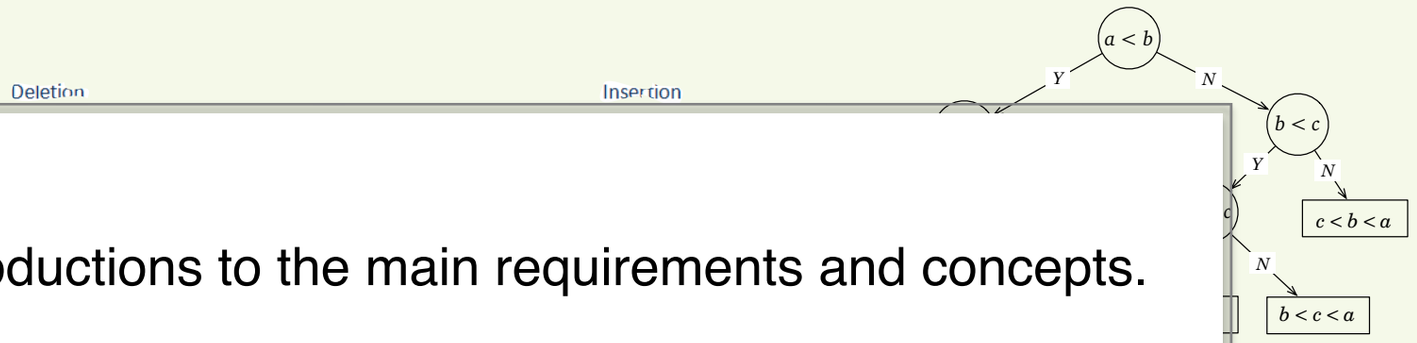
Master in Data Science

Master in Mathematics and Applications



EXTRA INFO - this label will appear in slides with complementary information





SUMMARY: The course and Introductions to the main requirements and concepts.

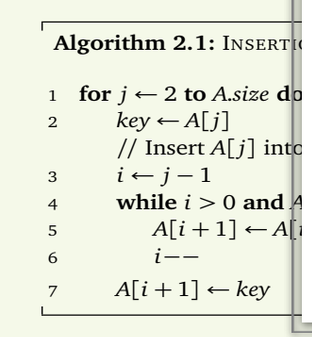
[T] About the course — content, practical works, seminars, evaluations.

Introduction to big data: the four V's. Opportunities and challenges. Big data vs machine learning.

Mathematical preliminaries (e.g. Bachmann–Landau notation).

[P] General overview about tools for big data and machine learning: micro-services, IoT (digital twin), ETL, databases (SQL vs noSQL), dashboarding.

Micro-services deploy pipeline for TMBD.



EXTRA INFO - this label will appear in slides with complementary information

Big Data vs Machine Learning ?
Any difference ?

Big Data vs Machine Learning ?
Any difference ?

Goal: The goal is to turn data into information

Challenges: Capture, curation, time-limitations, storage, search, sharing, transfer, analysis, and visualization of the data.

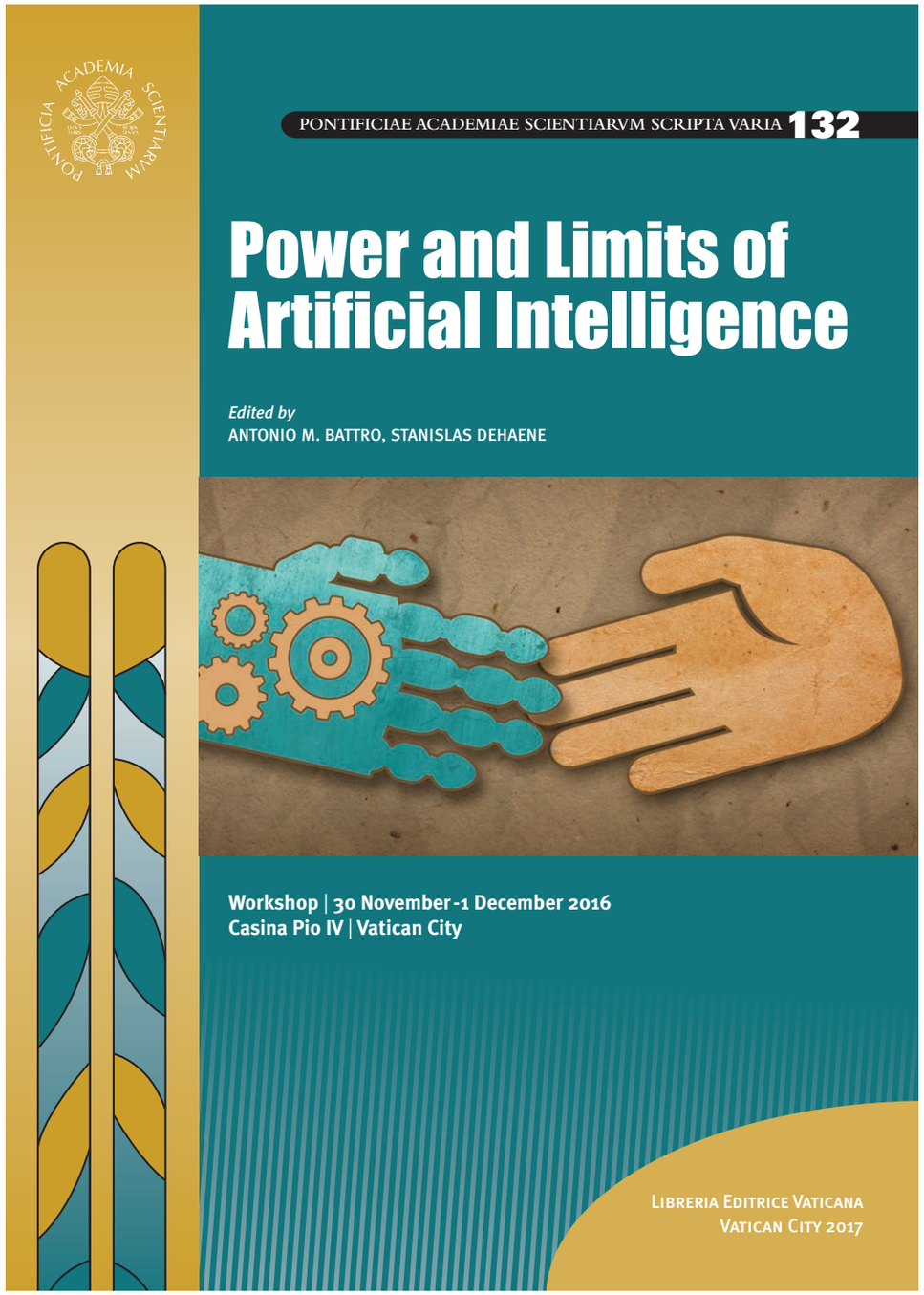
Data can be massive, non-static, multi-modal, incomplete, noisy, non-random, unstructured, dynamic, streaming, ...

Big Data vs Machine Learning ?
Any difference ?

Mathematics ?

OR

Computer Science ?



* see dropbox folder

Contents

Preface 11
Programme 12
List of Participants 14

State of the Art in Artificial Intelligence, Robotics, Brain Modeling, Brain-Computer Interfaces

Artificial Intelligence – Big Achievements and Huge Questions Viewed from Mathematics
Cédric Villani 19

The Cerebral Cortex: An Evolutionary Breakthrough
Wolf Singer 37

Comments: The Ethics of Artificial Intelligence
Stephen Hawking 50

Optimal Strategies for Decision-Making and Their Neural Basis
Alexandre Pouget 51

Motivations and Drives Are Computationally Messy
Patricia Smith Churchland 55

Children and Robots
Antonio M. Battro and Magela Fuzatti 60

Putative Prerogatives of the Human Brain: Education, Reasoning, Creativity, Consciousness, Sense of Self, Ethics...Could They be Captured in Machines?

Ghost In the Machine
Olaf Blanke 69

What Is Consciousness, and Could Machines Have It?
Stanislas Dehaene 75

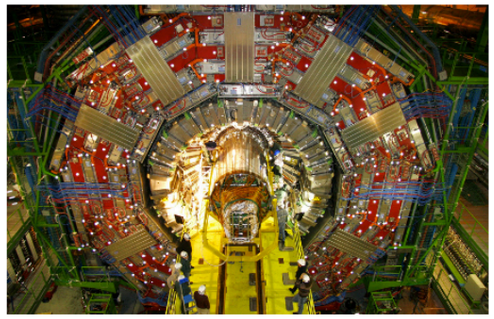
Big Data everywhere:

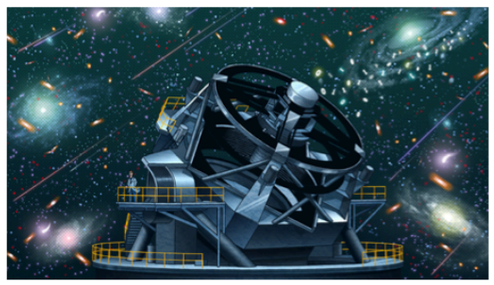
Lots of data is being collected and warehoused

- Web data (often user-provided)
- e-commerce, purchases at stores
- Medical data, health care
- Bank/Credit Card transactions
- Social Network
- Traffic, GPS, ...
- Scientific experiments
- ...

- YouTube contains 120 million videos and 72 hours of video uploaded every minute.
- Google processes 3.5 billion requests per day
- There is currently an estimate of 3.8 trillion photographs, 10% of them taken in the last year.
- Facebook has about 140 billion images with about 300 million new images a day.
- 2.5PB are flowing through Walmart's databases
- NYSE collects 1 TB each day.

- CERN's Large Hydron Collider generates 15 PB a year
- The BRAIN initiatives produce terabytes of data a day
- The Large Synoptic Survey Telescope in Chile will collect 30TB per night. Headed by Tony Tyson from UC Davis





Big Data does not just mean massive amounts of data
Big Data also means complex data

- Heterogeneous data
- Incomplete data
- Unstructured/semi-structured Data
- Graph Data
- Social Network, Semantic Web
- Streaming Data

Usual tasks for BDt (?!)

- Discovery of useful, possibly unexpected, patterns in data
- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Finding outliers (security threat, credit card theft, ...)
- Clustering
- Classification
- Object recognition
- Visualization, dimension reduction
- “Data cleaning”: denoising, smoothing, grouping, ...
- Association Rule Mining (Costumers who buy X often buy Y, Costumer 123 likes product p10)
- Collaborative filtering: users collaborate in filtering information to find information of interest (Amazon, Netflix)

Thomas Strohmer
Department of Mathematics
University of California, Davis

Usual tasks for BDt (?!)

The idea is 100 years old (see Karl Pearson), but its full potential will be unleashed only now.

Example:
In a recent analysis researchers developed a framework for comparing classiffers common in Machine Learning (Boosted decision trees, Random Forests, SVM, KNN, PAM and DLDA) based on a standard series of datasets.

Result: A simple (but mathematically rigorous) method gave better classification results across the data sets than the “glamorous” methods.

The dawning Age of Big Data will make it not just possible but very common (and perhaps necessary?) to validate methods via such meta data analyses.

Understanding deeply the problem
leads to better results
than
knowing a lot of "fancy" algorithms

"understand" = have a rigorous mathematical description

WHAT IS BIG DATA ?

WHAT IS MACHINE LEARNING ?

WHAT IS ... ?

Q5.1: What is the meaning of each concept and differences?

Your ideas:
- ...

EXTRA INFO

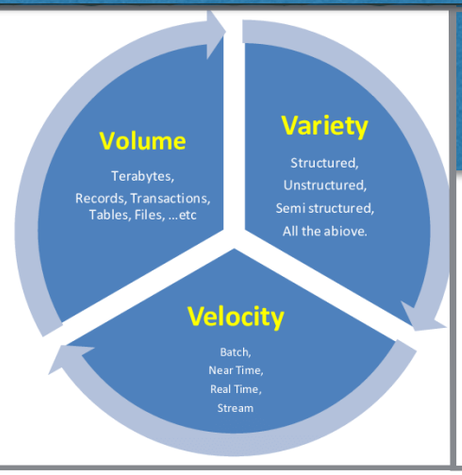
KI - Künstliche Intelligenz (2018) 32:3–8
https://doi.org/10.1007/s13218-017-0523-7

EDITORIAL

From Big Data to Big Artificial Intelligence?

Algorithmic Challenges and Opportunities of Big Data

Kristian Kersting[1] · Ulrich Meyer[2]



What is Big Data?

Some "definitions"

Definition of big data

The ability to collect huge amounts of data (i.e., database scale) about many different phenomenon (Anonymous)

Large volumes of interrelated data with multiple modalities (Sriraam Natarajan, UT Dallas, USA)

"Big Data" is the body of technology—algorithms, programming systems, and hardware, that stress our abilities to handle the data. A related, and apparently more modern term is "Data Science", which really means the same thing, but includes the application areas to which "big-data" technology is applied. What it DOESN'T mean is "AI" or "Machine Learning" or "Statistics done right", as some have claimed (Jeffrey D. Ullman, Stanford, USA)

I define it by size but also by the fact that it is typically collected as a side-effect of some other process (Anonymous)

One or more of: petabytes, high velocity, high complexity—high velocity (more data coming at you faster) predominant (Sofus Macskassy, Branch Metrics, USA)

Large volumes of data from diverse sources that can be used to learn new facts, or predict future events (Anonymous)

When it is faster to transport it on tape than to transfer it via satellite (Anonymous)

Big Data is a collection of data instances which is significantly larger than the data that has existed/been used in a given area before (Anonymous)

Computational approaches to solving computational problems lie on a spectrum of model complexity. Big data is at one end of the spectrum, effectively attempting to solve problems without explicit models. I would therefore define Big Data as the set of methods and efforts that attempt to solve problems without the necessity of explanation or understanding through explicit models (Oliver Brock, TU Berlin, Germany)

Enough data so that only very basic/generic priors (such as in NNs or CNNs) are sufficient ensure generalization to test data (Marc Toussaint, MIT, USA)

Mainly high dimensions. In fact rich SVD spectrum (Nikolaos Vasiloglou, MLtrain, USA)

"Big" is a relative term. A couple of decades ago, datasets with a million elements were considered big. But now, one routinely comes across datasets with trillions of elements (Anonymous)

Opportunities of using Big Data	Risks of using Big Data
High robustness in statistical learning (Christian Bauckhage, U. Bonn, Germany)	Privacy concerns when dealing with personal data, particularly for healthcare. Inability to verify the correctness of learned models or discovered patterns when analyzing large datasets, particularly when using "black-box" models. Security issues, dealing with adversaries, information overload, etc. (Anonymous)
Almost every field of human endeavor. I see lots of activity and promise in biology and biomedicine. Lots of commercial applications too, as companies like Google or Facebook use their data gathered from billions of people to understand much about human activity, from detecting spam, to knowing what I really mean when I misspell something. I don't see how this question can be answered in a few lines. My favorite example is providing driving directions based on real-time traffic estimates obtained from cell phone location data (Jeffrey D. Ullman, Stanford, USA)	Hype and expectation (Sriraam Natarajan, UT Dallas, USA)
Deeper holistic understanding of complex data (Sofus Macskassy, Branch Metrics, USA)	too high reliance on what is learned as truth (bad models, bad assumptions, not understanding the data). biased models in ways we do not understand (e.g., gender biases) (Sofus Macskassy, Branch Metrics, USA)
Quick solutions for problems that we do not understand but for which abundant data exists (Oliver Brock, TU Berlin, Germany)	If the learning systems leveraging these Big Data sets are not designed carefully, they may end up codifying our biases and stereotypes (e.g., possible racial bias in automated airport profiling) and this may result in these biases getting even more deeply ingrained (Deepak Ajwani, Nokia Bell Labs, Ireland)
Understanding and getting the data is the key to solving many problems. It may not sound very complex or challenging on the surface, however is the key to be successful in AI / ML initiatives (Anonymous)	Frankly, I think the biggest risk is that governments, especially in Europe, will worry more about privacy than about the advantages that can come from exploiting "big data". Privacy is a modern invention. 200 years ago, before the anonymity of cities was available, people didn't imagine that they could keep their lives secret from those around them (Jeffrey D. Ullman, Stanford, USA)

Opportunities of using Big Data	Risks of using Big Data
To solve low-level language and image perception tasks. Discover correlations for rare medical conditions (Guy Van den Broeck, UCLA, USA)	That science in its current form ceases to exist because people just believe in the power of data but not in the power of understanding (Oliver Brock, TU Berlin, Germany)
There are two very different instances of Big Data: the one makes modelling very easy, because the large data show underlying regularities (all models are wrong and we we don't need them any more). This is a huge opportunity for analysis. The other are of the needle in the haystack type and make analysts' work harder (Anonymous)	How to verify analysis methods and results? How to reproduce results? (Anonymous)
The opportunity to analyse data streams in real-time moves machine learning into the direction of real-time control. This opens up tremendous applications.(Anonymous)	Algorithms that require large amounts of data will be less useful compared to algorithms that can work on smaller amounts of rare occurrence data (Anonymous)
With the design of more scalable learning algorithms and better parallel hardware, vast amount of collected data can be leveraged for myriad applications, particularly in augmented intelligence (Deepak Ajwani, Nokia Bell Labs, Ireland)	To get 'stupid' machines that don't reason to yield an answer, that cannot learn from few data (e.g., in personal conversations) (Marc Toussaint, MIT, USA)

- Challenges and opportunities with big data
  (Proceedings of the VLDB Endowment 5(12):2032-2033)

- For a Career Path in Big Data
  (Edureka 2019 Tech Career Guide)

- Supply Chain Talent of the Future - Findings from the third annual supply chain
  (Deloitte)

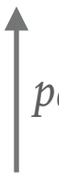
WHAT IS BIG DATA ?

WHAT IS MACHINE LEARNING ?

WHAT IS ... ?

E.Rocha's understanding

***BIG DATA**: set of technologies and methodologies to (efficiently) deal with the transport, processing and storage of a (nowadays considered) big amount of information.*



***MACHINE LEARNING**: a group of techniques and algorithms to model phenomena based on the information contained in datasets, rooted on mathematical models and optimization techniques, and implemented on well-known computer science frameworks.*

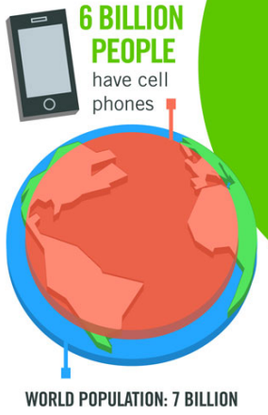
The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States



Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

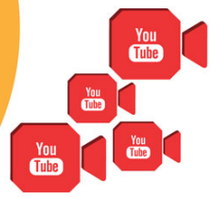
Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

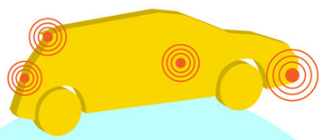
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month



400 MILLION TWEETS
are sent per day by about 200 million monthly active users

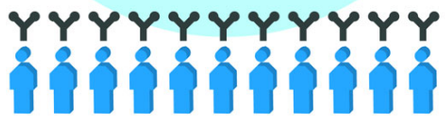
Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session



Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth



Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

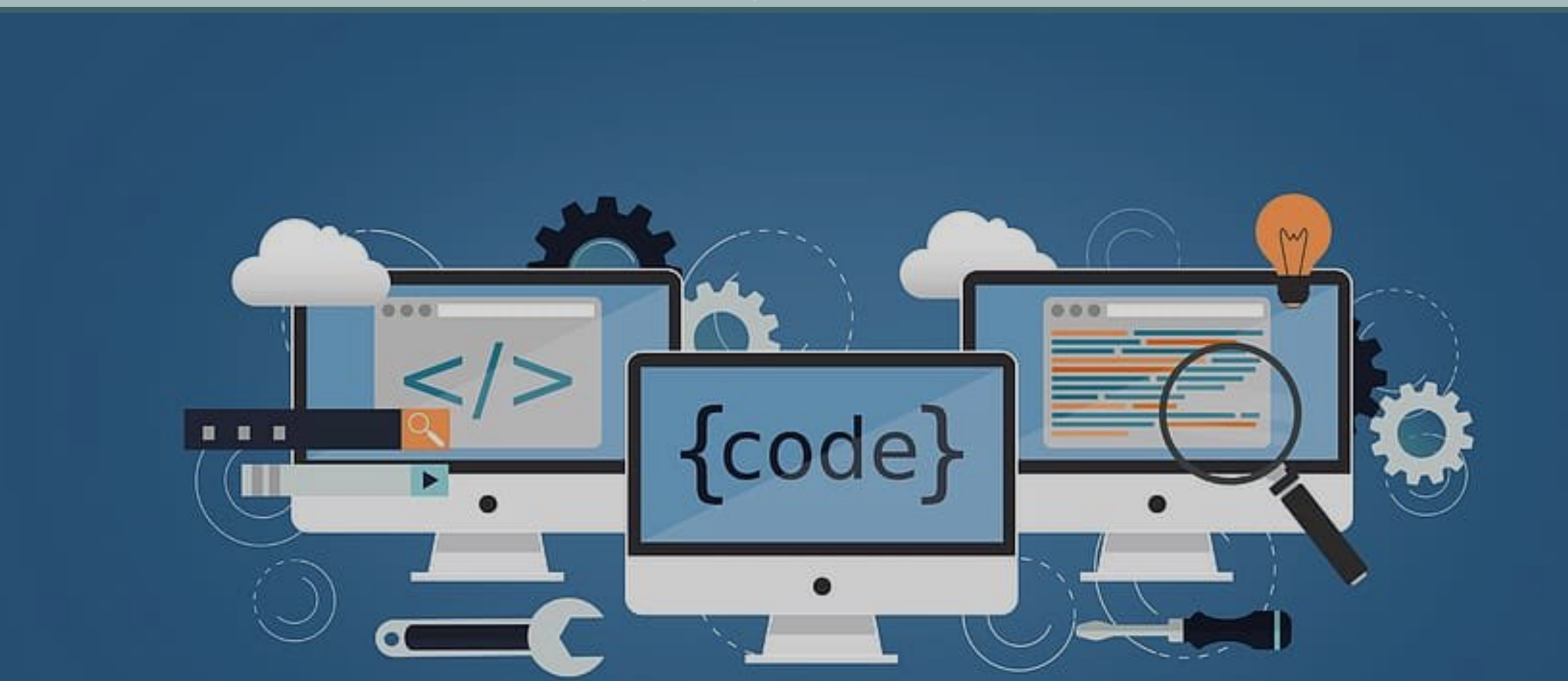
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Now it is common to find an additional V of Big Data: the Value.

QUESTION

Is AI, ML and BigData “good” or “bad” to humanity?



What is the oldest tool in the human history?

TOOLS FOR CUTTING

Oldest stone tools pre-date earliest humans

By Rebecca Morelle
Science Correspondent, BBC News

They were unearthed from the shores of Lake Turkana in Kenya, and date to 3.3 million years ago.

They are 700,000 years older than any tools found before, even pre-dating the earliest humans in the *Homo* genus.

The find, **reported in Nature**, suggests that more ancient species, such as *Australopithecus afarensis* or *Kenyanthropus platyops,* may have been more sophisticated than was thought.

2017 London Bridge Attack

London Bridge Attack: 7 Killed, 48 Wounded; 3 Suspects Shot Dead by Police

Convicted murderer hailed as 'hero' in London Bridge attack

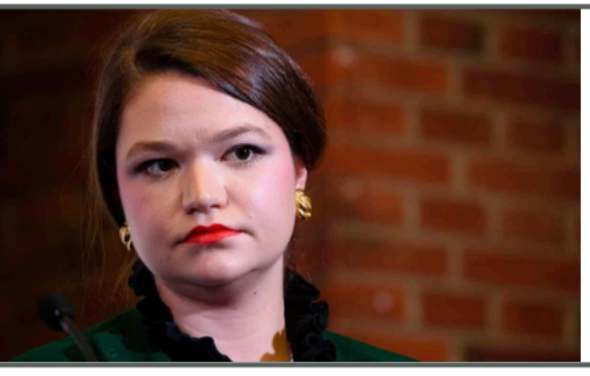
Man convicted of killing woman, 21, joined others to stop terrorist who went on knife rampage

Dec 02, 2019 06:00 am

LONDON: One of the men who helped to stop the London Bridge terrorist from killing more people was once a killer himself.

EXTRA INFO

A RECENT (NON)STANDARD EXAMPLE



Fresh Cambridge Analytica leak 'shows global manipulation is out of control'

Company's work in 68 countries laid bare with release of more than 100,000 documents

Carole Cadwalladr

@carolecadwalla

Sat 4 Jan 2020 16.55 GMT

42,334

Q5.2: What is the biggest intrusion to your privacy?

Q5.3: Can AI be predict if a couple breaks after 4 years?

Q5.4: Can AI model the couple interaction dynamics?

Q5.5: Can AI be used for choosing innovation successfully ?

QUESTION

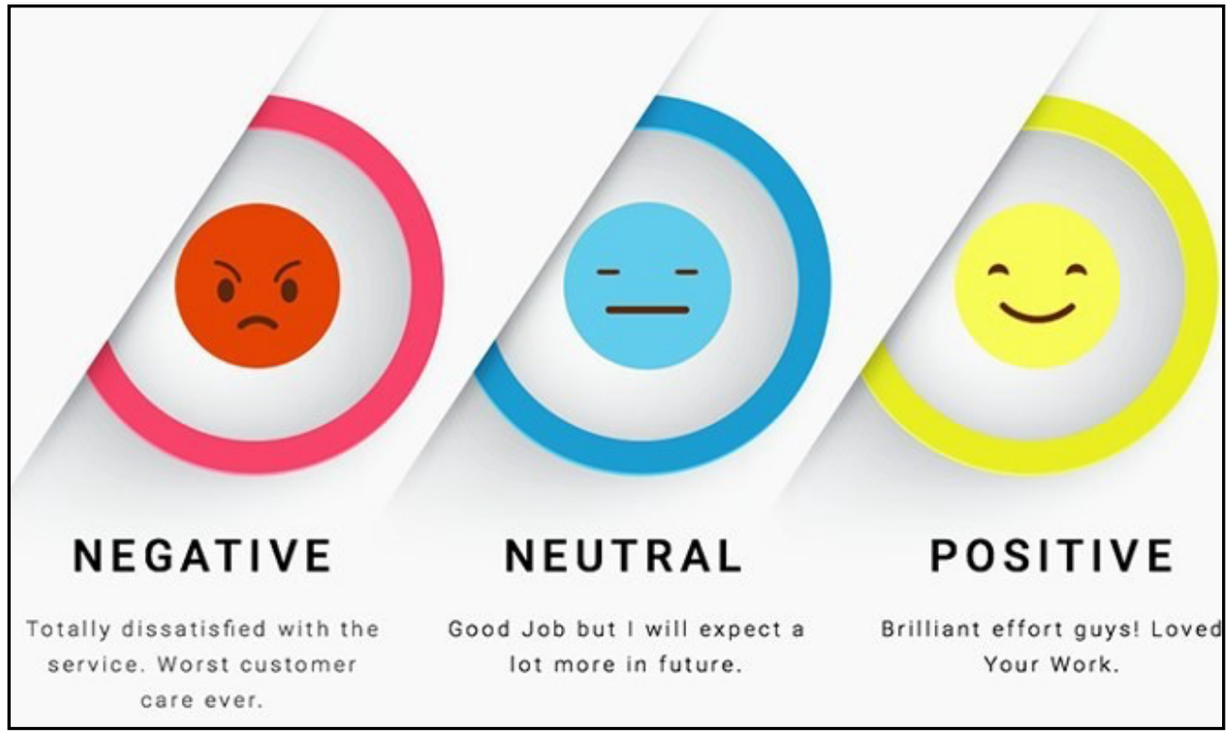
EXTRA INFO

How human sentiments may improve the economy but destroy creativity ? Predicting crowdfunding campaigns...

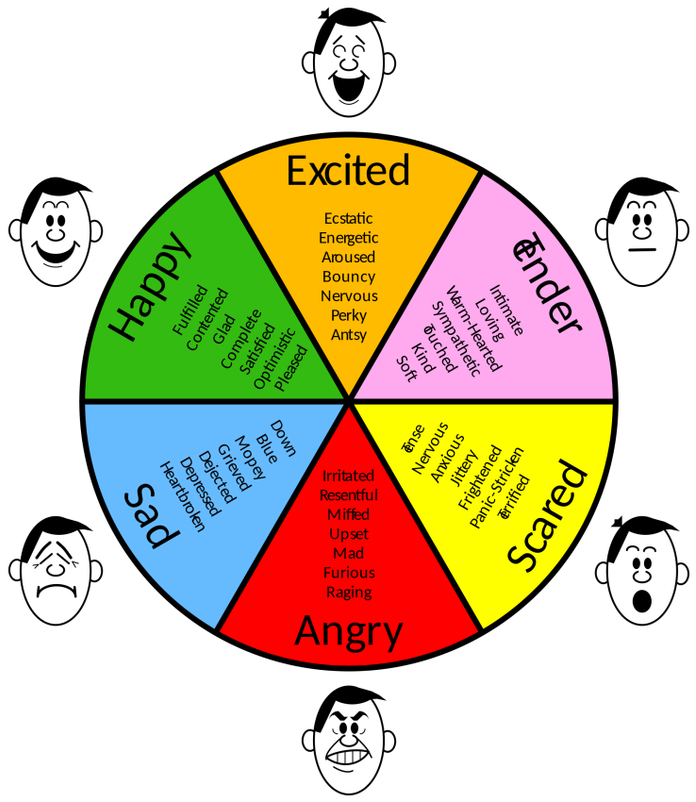
Definição (António Damásio):

- **EMOÇÃO**: conjunto de reacções corporais, automáticas e inconscientes, face a determinados estímulos provenientes do meio onde estamos inseridos;
- **SENTIMENTO**: surge quando tomamos consciência das nossas emoções, isto é, o sentimento dá-se quando as nossas emoções são transferidas para determinadas zonas do nosso cérebro.

Análise de Sentimentos



Análise de Emoções



Análise de Sentimentos

(opinion mining)

— Positivo, neutro ou negativo?

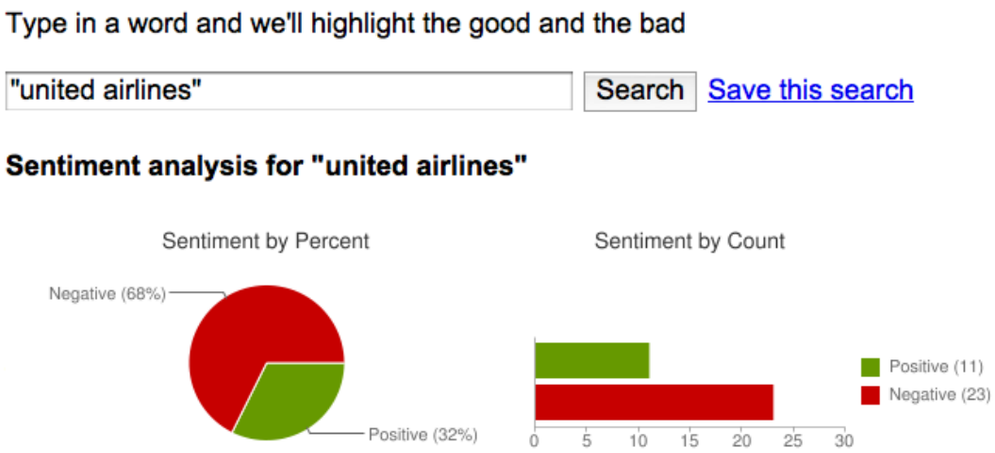
- unbelievably disappointing
- Full of zany characters and richly applied satire, and some great plot twists
- this is the greatest screwball comedy ever filmed
- It was pathetic. The worst part about it was the boxing scenes.

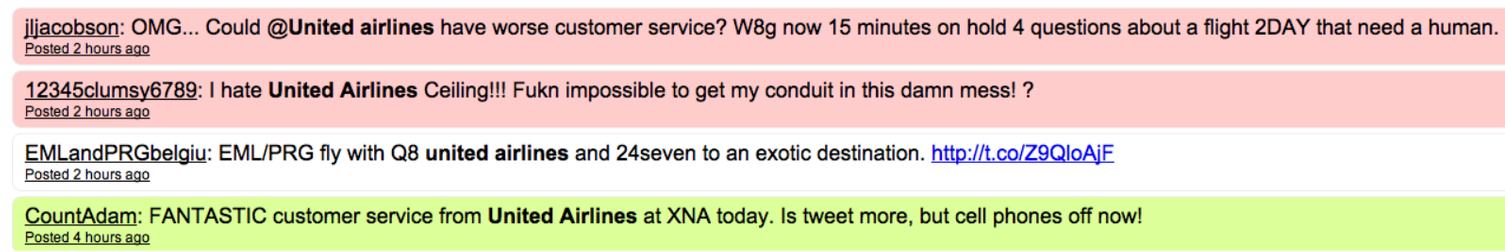
— Várias fontes:

- Google product search
- Bing shopping
- ★ **Twitter**
- Etc.

Sentiment Analysis (NLP) on Twitter

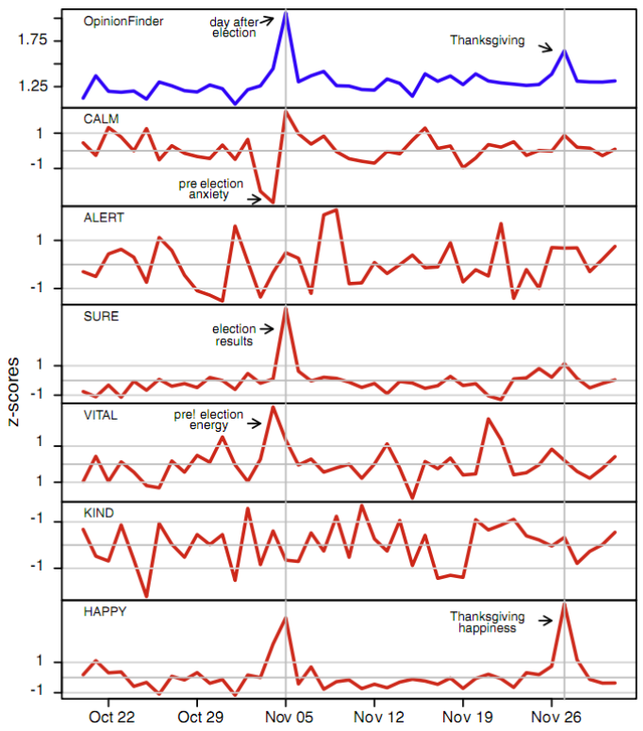
Alec Go, Richa Bhayani, Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision



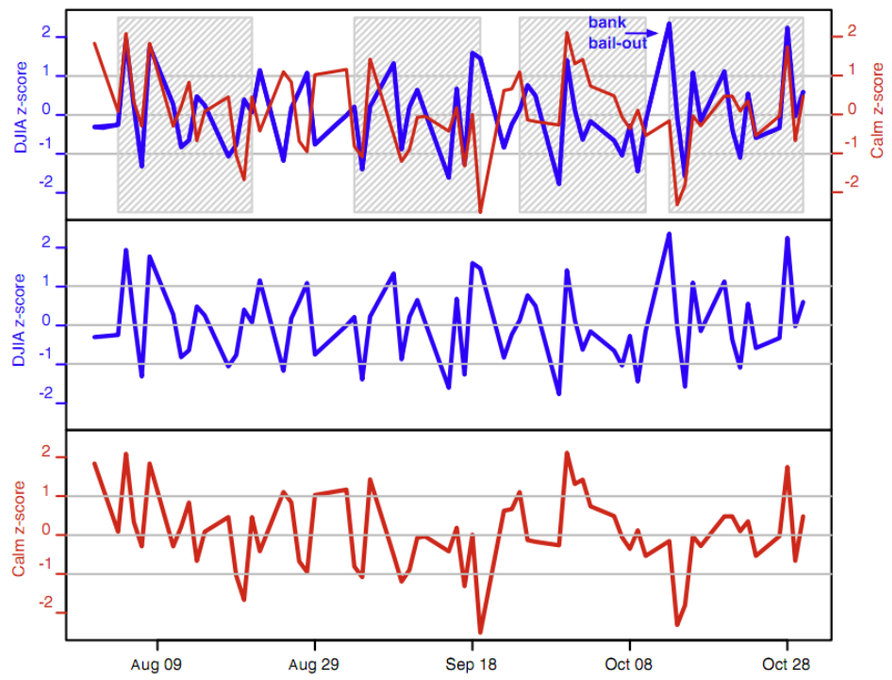


Can we predict Stock Market from Twitter information?

Johan Bollen, Huina Mao, Xiaojun Zeng. 2011. Twitter mood predicts the stock market, Journal of Computational Science 2:1, 1-8. 10.1016/j.jocs.2010.12.007.



FINDING: CALM predicts DJIA 3 days later



An Interpretable Prediction Model based on Campaigns Aspects Emotions

work with a GERMAN PRIVATE COMPANY

Commercial Web App

Campaign Goal ($): 20000

Last Seasons: 3

Show all clusters

EXTRA INFO

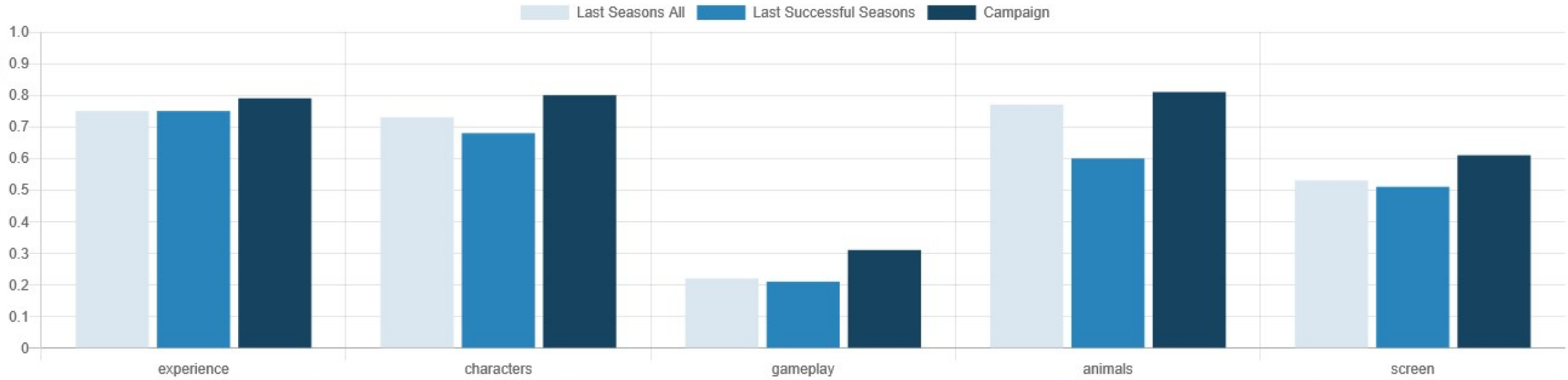
Campaign Description:

* For those new to Kickstarter, scroll down to read the FAQ's below. Basically, you pledge a $ amount to this campaign and only get charged if the project is fully funded. You will receive a Kickstarter survey at the end of the campaign to select reward choices.Watch the Fungisaurs animated origin story below!Hope you enjoyed that as much as we enjoyed making it. This is our first step towards

Start

Campaign Prediction:

Scoring emotional intensity on Clusters.
Benchmarked on campaigns of the last *3* seasons.



Cluster	experience (0,79)	characters (0,8)	gameplay (0,31)	animals (0,81)	screen (0,61)
Feature Cluster Need Index Benchmarking	Optimization Opportunity	Optimization Opportunity	Major revisions required	Optimization Opportunity	Minor revisions required
Trend Need Index (All)	0,75 ↓	0,73 ↓	0,22 ↓	0,77 ↓	0,53 ↓
Trend Need Index (Successful)	0,75 ↓	0,68 ↓	0,21 ↓	0,6 ↓	0,51 ↓
Lead Mood Season Trend (All) / Lead Mood Season Trend (Susscessfull) / Campaign Lead Mood	Joy /Joy / Joy	Joy /Joy / Joy	Joy /Admiration / Joy	Joy /Disgust / Admiration	Joy /Admiration / Admiration
Domain Aspect hits - Needindex	experience: 0,76	character: 0,73	adventure: 0,72	cat: 0,6	screen: 0,74

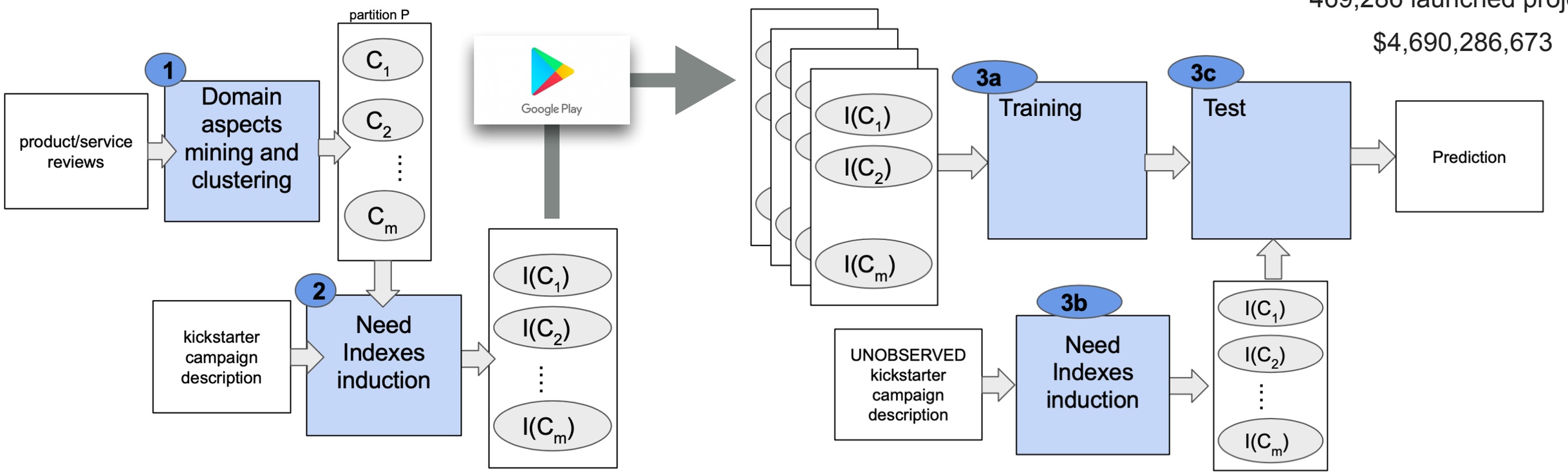
An Interpretable Prediction Model based on Campaigns Aspects Emotions

work with S.Rittinghaus + N.Samsami (GERMANY) and S.Faralli + D.Distante (ITALY)

Kickstarter

469,286 launched projects

$4,690,286,673



State-of-Art
(on 16/01/2020)

To the best of our knowledge, by the art prediction accuracy on the basis of static Kickstarter data is **76,4%**

NLP + ML (python)

NeedIndex

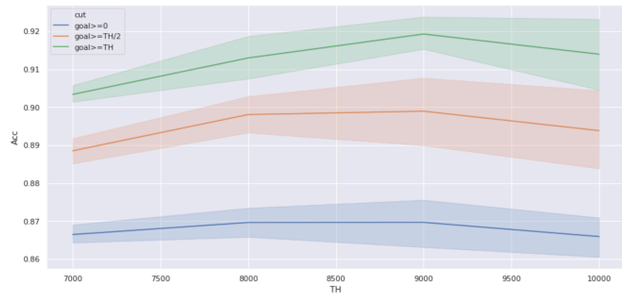
Unsupervised: *P_MG_2009-2019*				
system	P	R	F1	A
MultiLayerPerceptron	0.75	0.87	0.81	0.87
SupportVectorMachines	0.75	0.87	0.81	0.87
RandomForestClassifier	0.78	0.85	0.80	0.85
GradientBoostingClassifier	0.75	0.87	0.81	0.87
Supervised: *P*_MG_2009-2019*				
system	P	R	F1	A
MultiLayerPerceptron	0.81	0.90	0.85	0.90
SupportVectorMachines	0.81	0.90	0.85	0.90
RandomForestClassifier	0.81	0.89	0.85	0.89
GradientBoostingClassifier	0.81	0.90	0.85	0.90
baseline				
system	P	R	F1	A
Random	0.77	0.50	0.58	0.51

recreate all the pipeline in **Spark**

Ideas from the theoretical part about PDEs

Reproducing Kernel Banach Spaces

Convexity



Our final accuracy **94%**

Best cut: 9000$

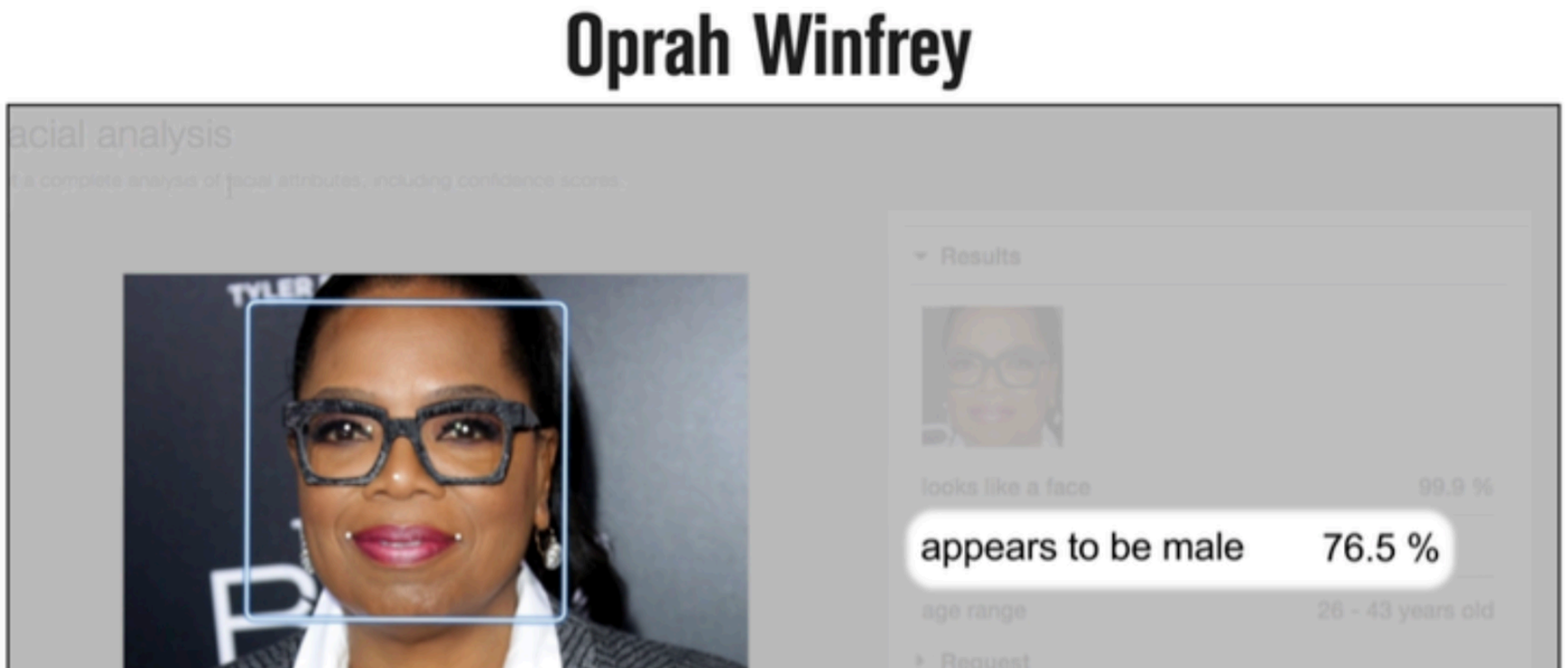
QUESTION

ML/AI algorithm is (in)parcial? Avoid social discrimination?!

Joy Buolamwini /The Algorithmic Justice League at MIT Media Lab, Joy Buolamwini /The Algorithmic Justice League, 2018/2018, From the collection of: Barbican Centre

Joy Buolamwini

machines are neutral, but they aren't. My research uncovered racial bias in AI systems sold by tech giants like IBM, Amazon. Given the task of guessing the gender of a face, all med substantially better on male faces than female faces. The ated had error rates of no more than 1% for lighter-skinned kinned women, the errors soared to 35%. AI systems from s have failed to correctly classify the faces of Oprah Winfrey, and Serena Williams. When technology denigrates even these iconic women, it is time to re-examine how these systems are built and who they truly serve.

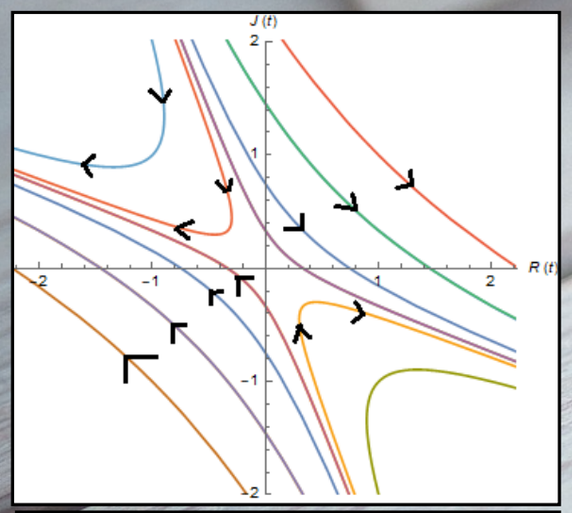


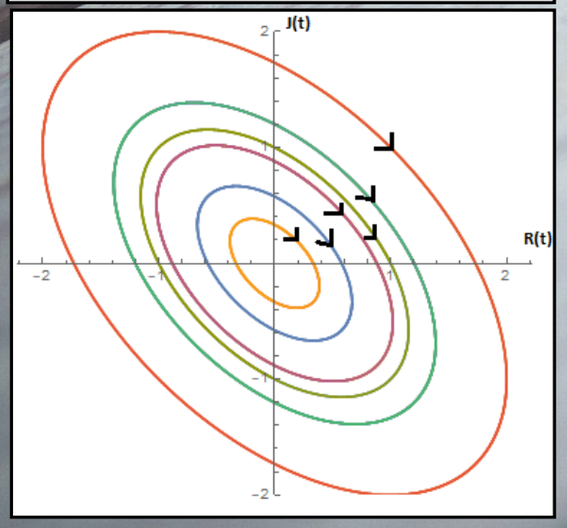
QUESTION

Can we model “love” or “friendship” ?

QUESTION

Can we model “love” or “friendship” ?





There are several mathematical models in the literature, e.g.

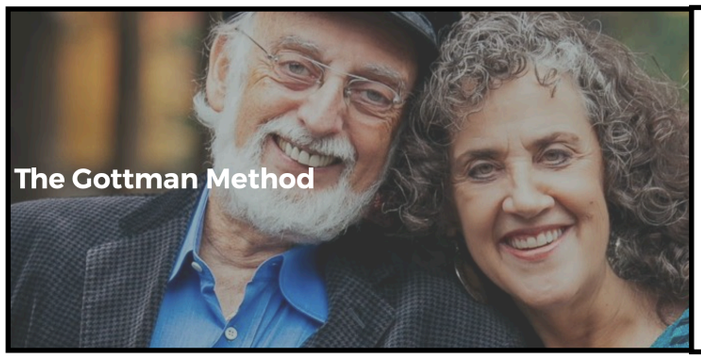
$$\dot{R}(t) = \frac{dR(t)}{dt} = aR(t) + bJ(t)$$

$$\dot{J}(t) = \frac{dJ(t)}{dt} = cR(t) + dJ(t)$$

BG/ML techniques can be used to find parameters a,b,c,d

Can we predict if a couple is going to break after 4 years of relationship?

Trabalho do Gottman Institute (Washington)



Goals and Principles of the Gottman Method

The goals of Gottman Method Couples Therapy are to disarm conflicting verbal communication, increase intimacy, respect, and affection, remove barriers that create a feeling of stagnancy in conflicting situations, and create a heightened sense of empathy and understanding within the context of the relationship.

> In our laboratory's multi-method research on marital interaction we have shown, in four longitudinal studies, that we can predict with over 90% accuracy whether a couple will divorce or stay married, and their marital satisfaction if they do stay married (Gottman, 1994; 1999).

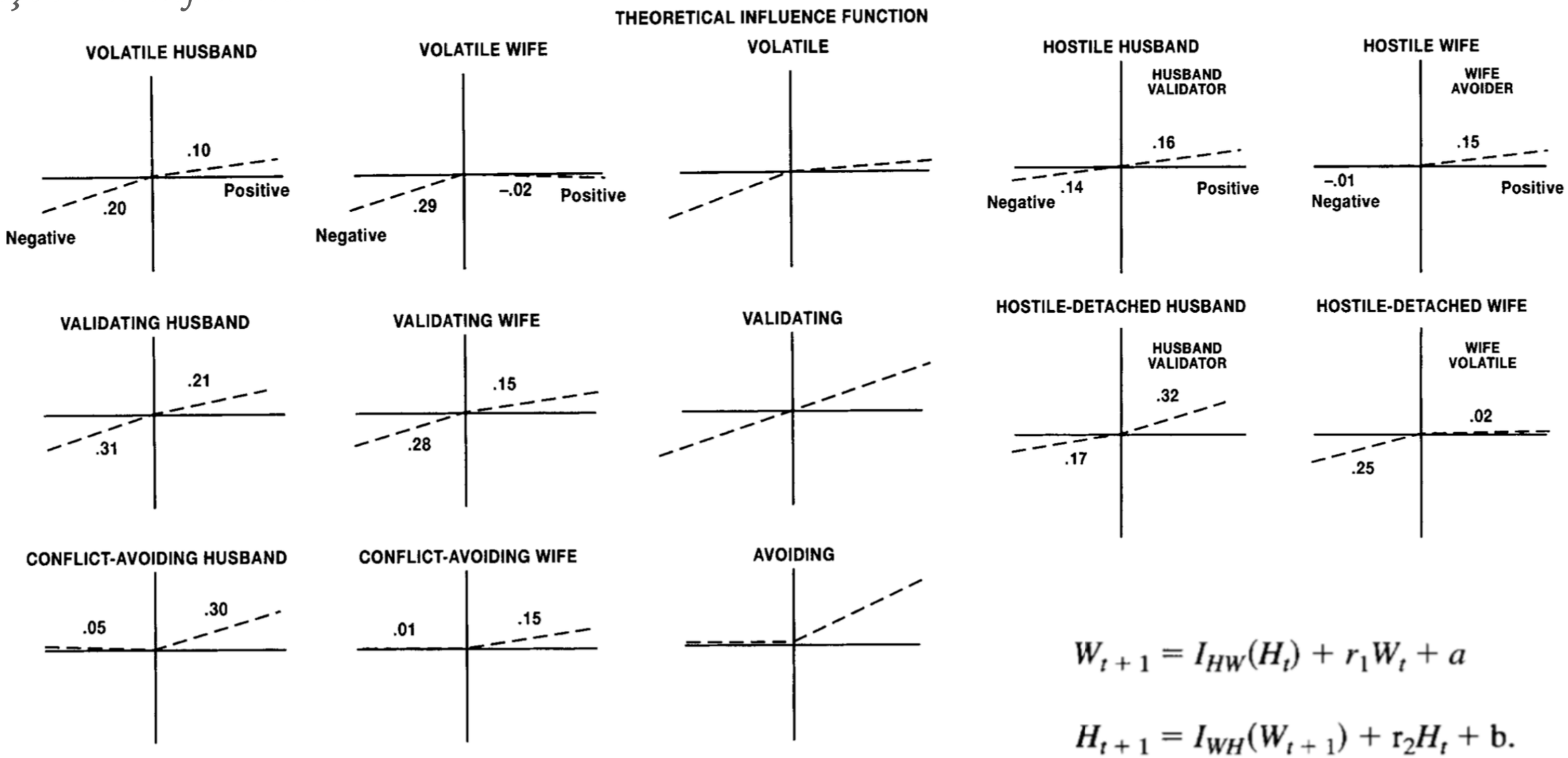
> 90% !!!

Trabalho do Gottman Institute (Washington)

> In our laboratory's multi-method research on marital interaction we have shown, in four longitudinal studies, that we can predict with over 90% accuracy whether a couple will divorce or stay married, and their marital satisfaction if they do stay married (Gottman, 1994; 1999).

> 90% !!!

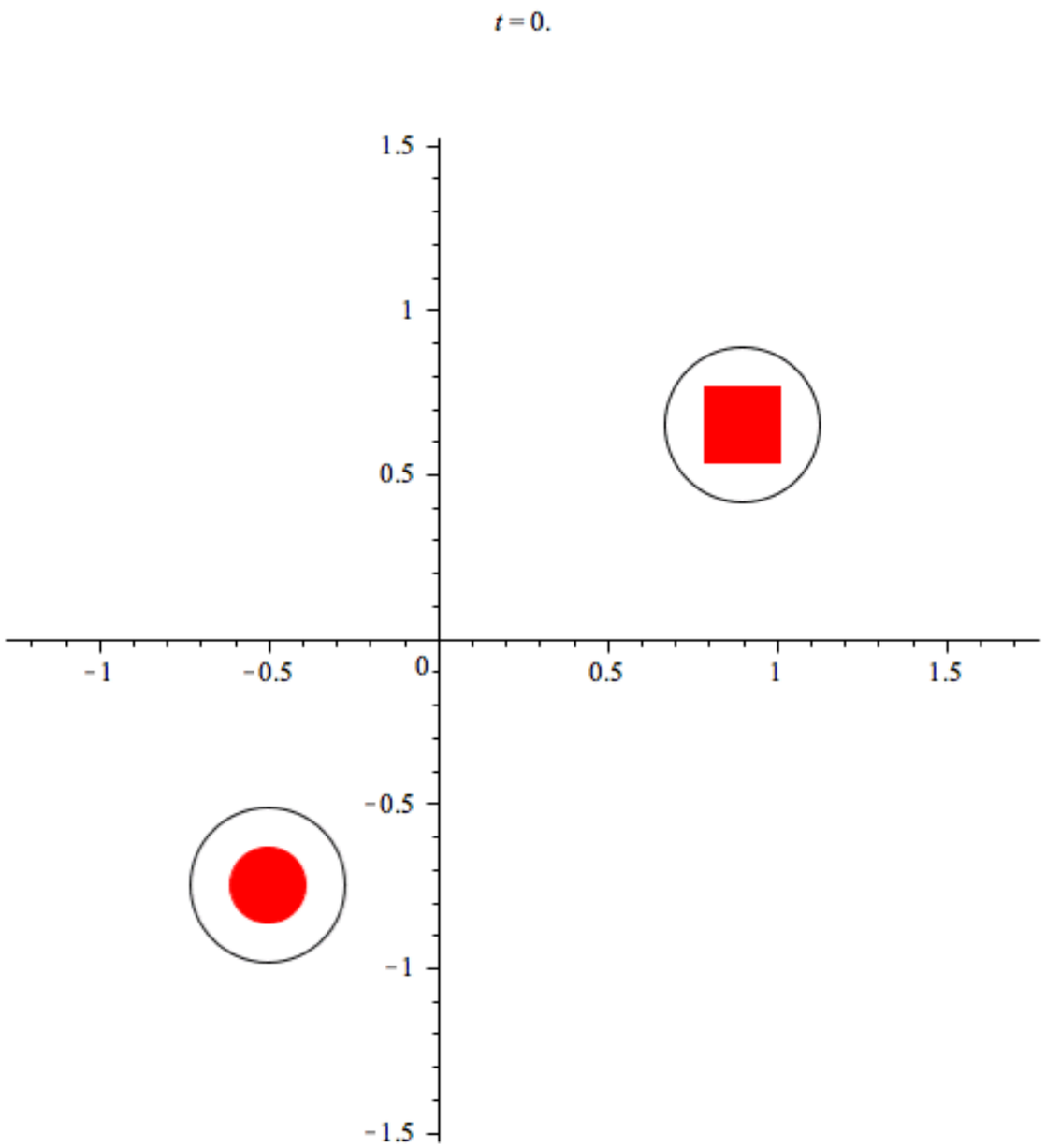
Funções de influência

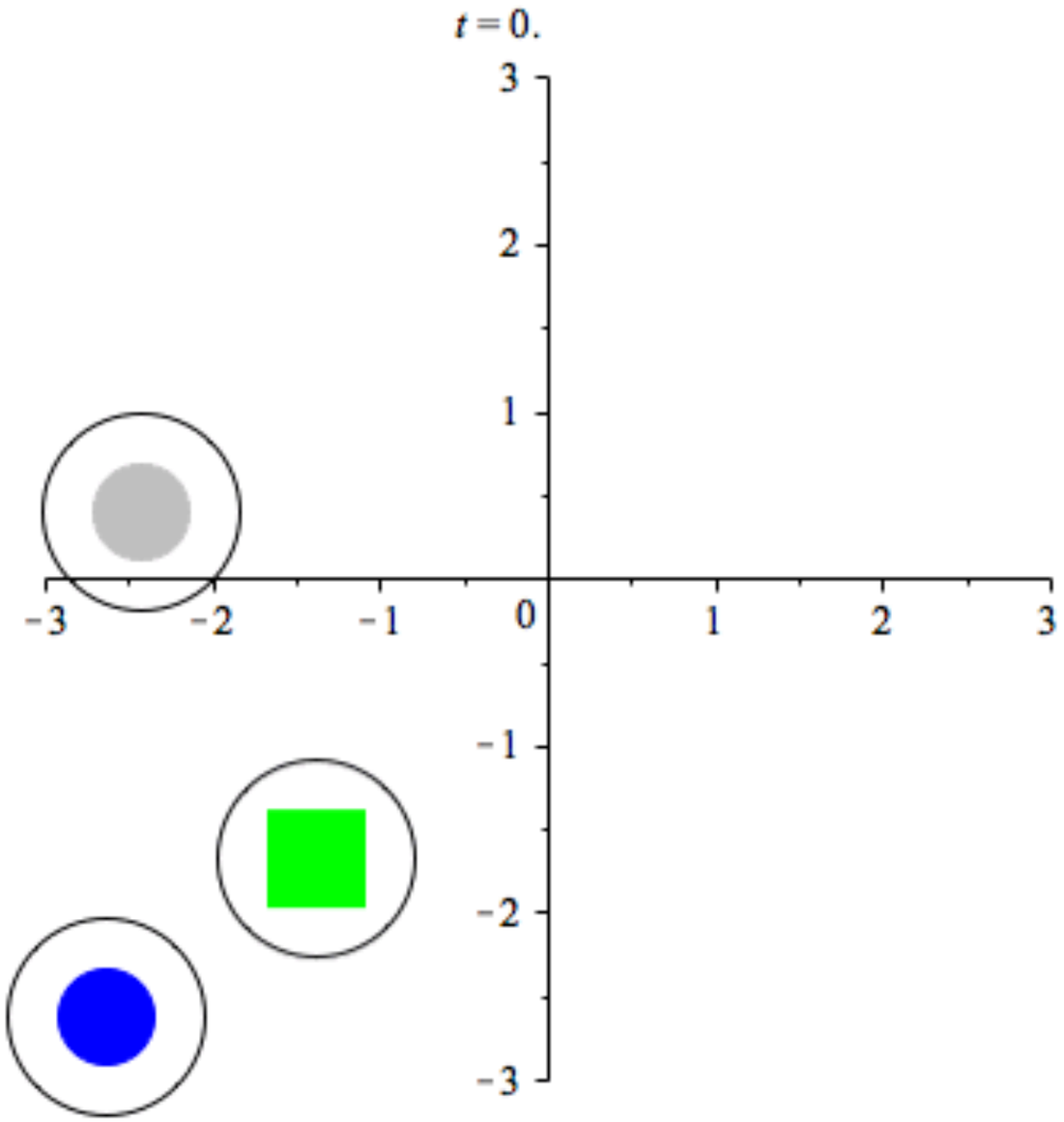


$$W_{t+1} = I_{HW}(H_t) + r_1 W_t + a$$

$$H_{t+1} = I_{WH}(W_{t+1}) + r_2 H_t + b.$$

Example 1

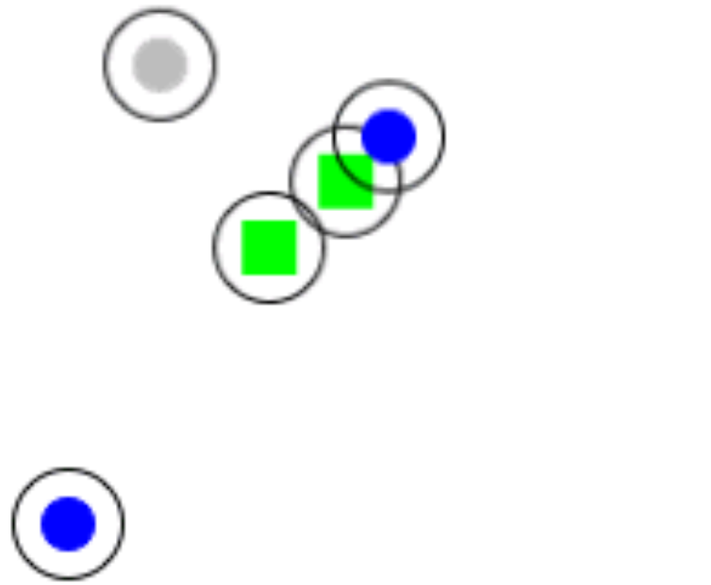




LEGENDA:
verde (volatile), **azul** (validating), **cinzento** (conflit-avoiding), **vermelho** (hostile), **laranja** (hostile-deatached)
bola (woman), quadrado (man)

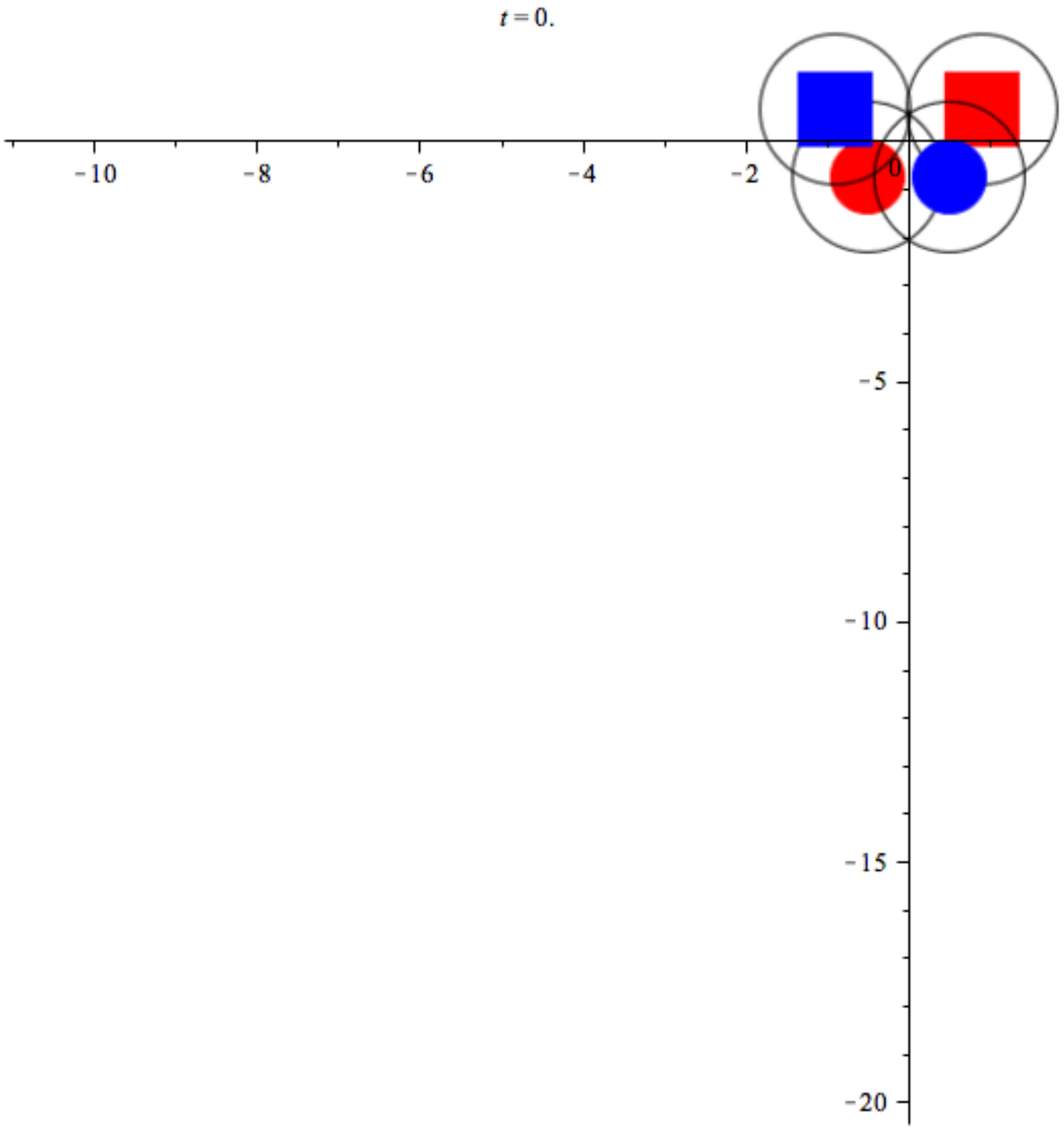
Example 2

$t = 0$.



LEGENDA:
verde (volatile), **azul** (validating), **cinzento** (conflit-avoiding), **vermelho** (hostile), **laranja** (hostile-deatached)
bola (woman), quadrado (man)

Example 3



LEGENDA:
verde (volatile), **azul** (validating), **cinzento** (conflit-avoiding), **vermelho** (hostile), **laranja** (hostile-deatached)
bola (woman), quadrado (man)

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

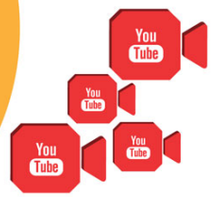
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

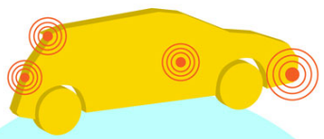


30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session



Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

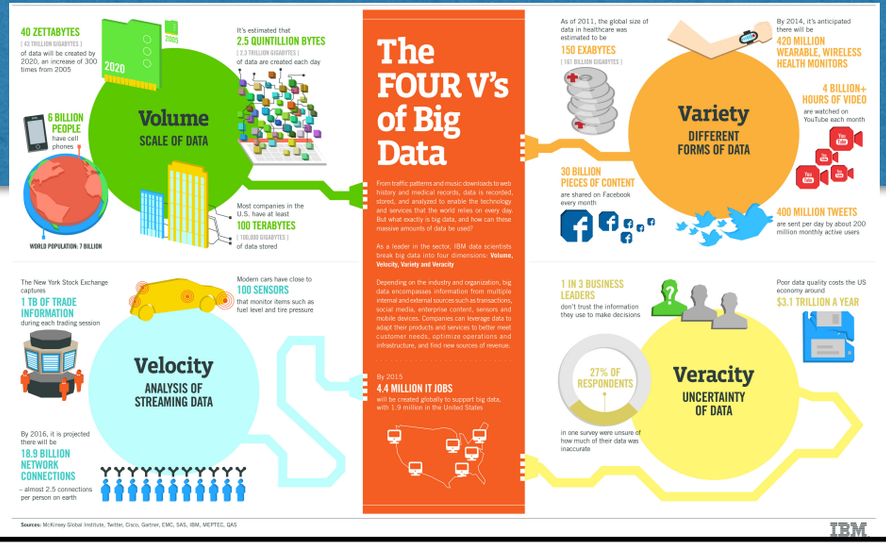
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

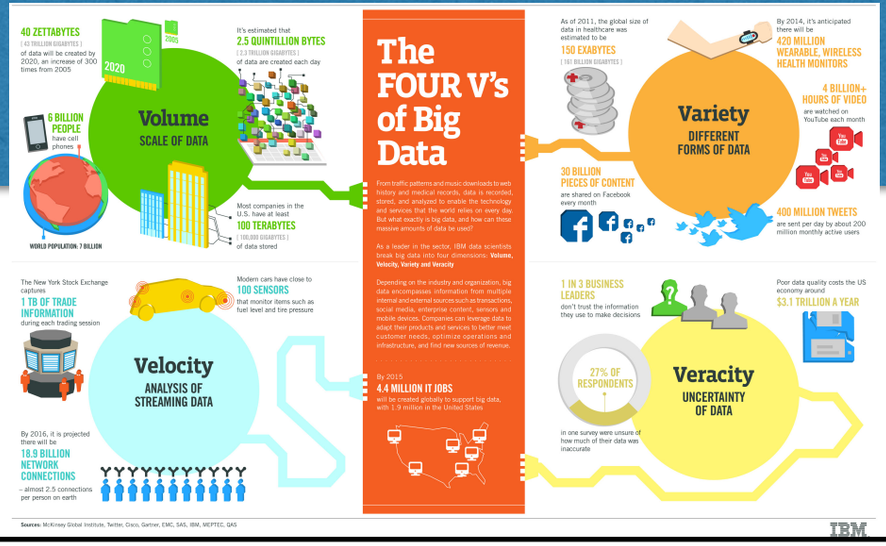
IBM



"Técnicas Matemáticas para Big Data"

Q: Difficulties?

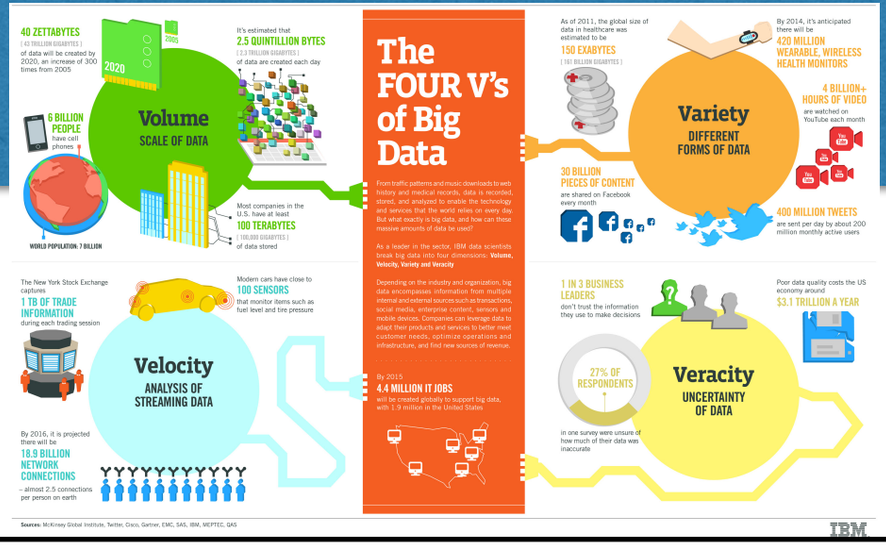
- **Low Veracity**
  - Veracity, validity, …, falsification, ethics
  - Discrete/Continuous Markov Chains
  - Bayesian Networks
  - Fuzzy Logics



"Técnicas Matemáticas para Big Data"

- Low Veracity
  - Veracity, validity, …, falsification, ethics
  - Discrete/Continuous Markov Chains
  - Bayesian Networks
  - Fuzzy Logics
- **High Velocity**
  - About data streams (DS)
  - Statistical algorithms for DS

Q: Difficulties?



"Técnicas Matemáticas para Big Data"

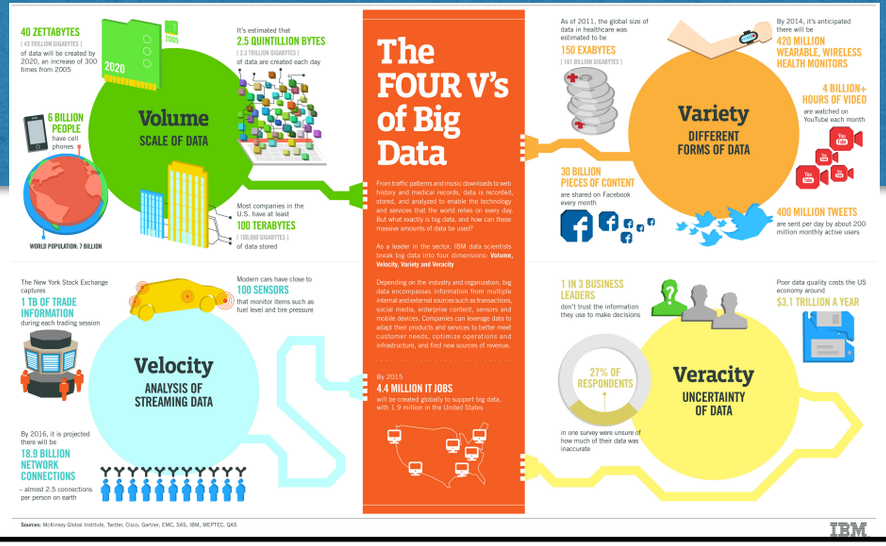
- Low Veracity
  - Veracity, validity, …, falsification, ethics
  - Discrete/Continuous Markov Chains
  - Bayesian Networks
  - Fuzzy Logics
- **High Velocity**
  - About data streams (DS)
  - Statistical algorithms for DS

Q: Difficulties?

Problem 1: Calculate the mean and standard deviation without keeping the values of the previous temperatures.

$$W = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

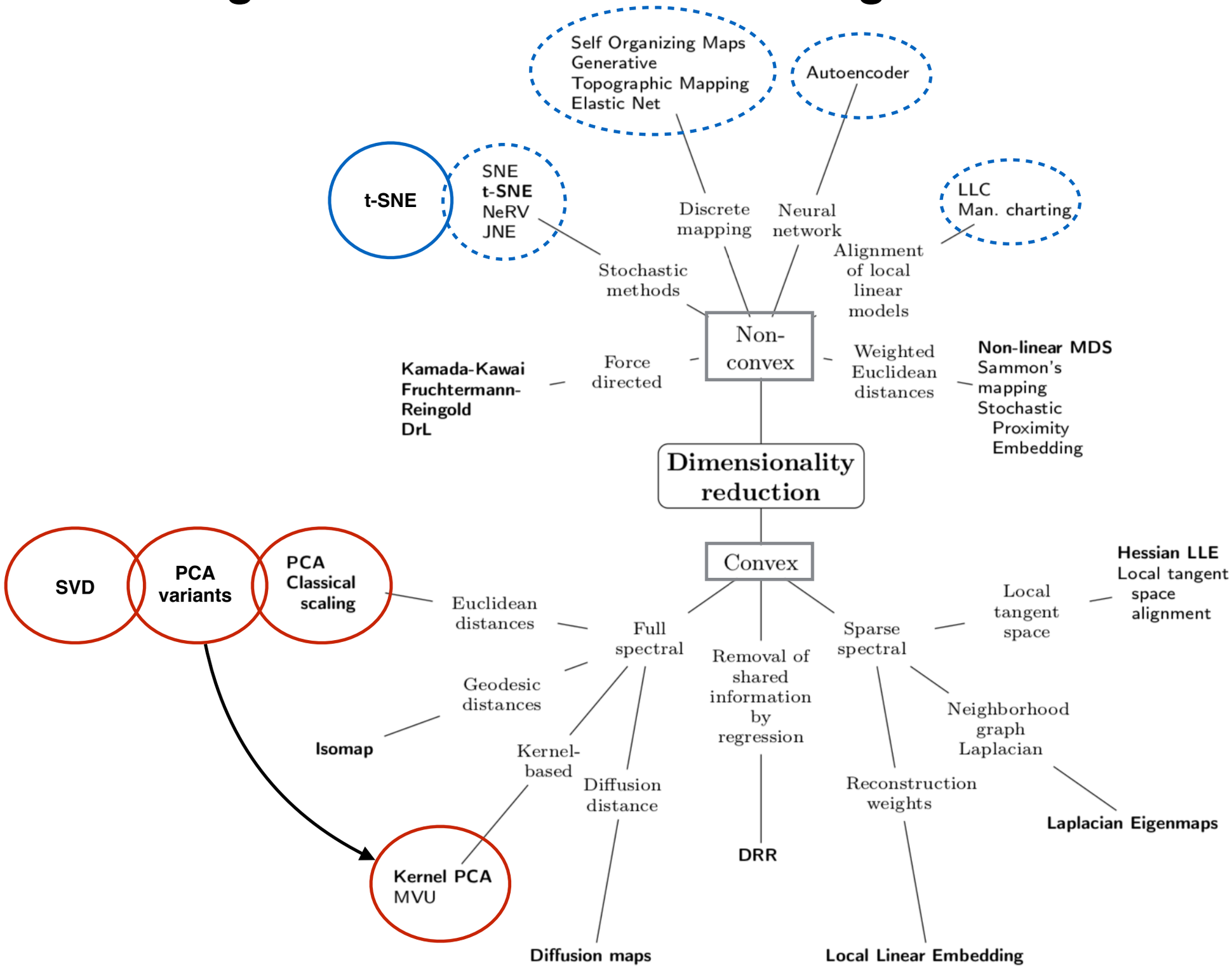


"Técnicas Matemáticas para Big Data"

- Low Veracity
  - Veracity, validity, …, falsification, ethics
  - Discrete/Continuous Markov Chains
  - Bayesian Networks
  - Fuzzy Logics
- High Velocity
  - About data streams (DS)
  - Statistical algorithms for DS
- **Large Volume**
  - Dimension reduction methods

Q: Difficulties?

Brief Catalog of Dimension Reduction Algorithms



see notes/N01-Higher_Dimensions.pdf

1 Surprises in high dimensions

Our intuition about space is based on two and three dimensions and can often be misleading in high dimensions. It is instructive to analyze the shape and properties of some basic geometric forms, which we understand very well in dimensions two and three, in high dimensions. To that end, we will look at the sphere and the cube as their dimension increases.

1.1 Geometry of the d-dimensional Sphere

Consider the unit sphere is d dimensions. Its volume is given by

$$V(d) = \frac{\pi^{\frac{d}{2}}}{\frac{d}{2}\,\Gamma\left(\frac{d}{2}\right)}$$

where Γ is the Gamma function. Recall that for positive integers n, $\Gamma(n) = (n-1)!$. Using Stirling's Formula,

$$\Gamma(n) \sim \sqrt{\frac{2\pi}{n}}\left(\frac{n}{e}\right)^n$$

we can see $\Gamma\left(\frac{d}{2}\right)$ grows much faster than $\pi^{\frac{d}{2}}$, and hence

$$V(d) \to 0 \qquad \text{as} \qquad d \to \infty.$$

Volume of a d-dimension unitary sphere

Consider the unit sphere is d dimensions. Its volume is given by

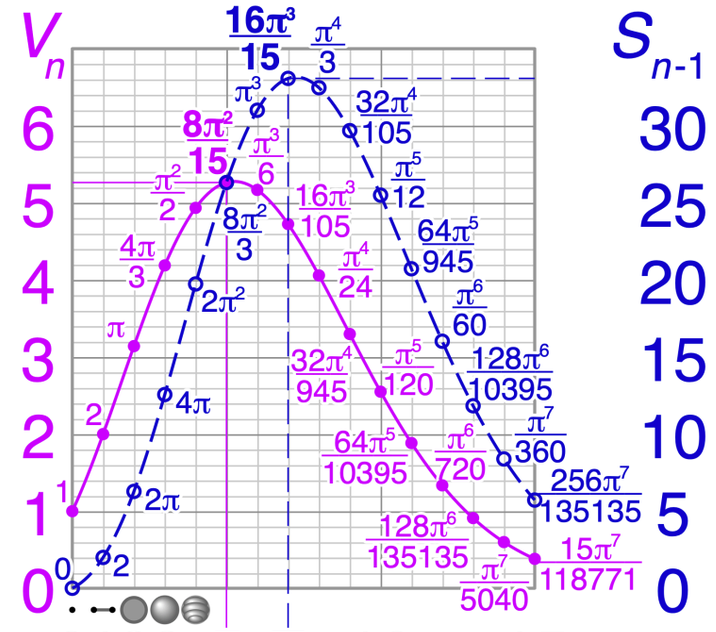
$$V(d) = \frac{\pi^{\frac{d}{2}}}{\frac{d}{2}\,\Gamma\left(\frac{d}{2}\right)}$$

where Γ is the Gamma function. Recall that for positive integers n, $\Gamma(n) = (n-1)!$. Using Stirling's Formula,

$$\Gamma(n) \sim \sqrt{\frac{2\pi}{n}}\left(\frac{n}{e}\right)^n$$

we can see $\Gamma\left(\frac{d}{2}\right)$ grows much faster than $\pi^{\frac{d}{2}}$, and hence

$$V(d) \to 0 \qquad \text{as} \qquad d \to \infty.$$



Volume of a d-dimension unitary cube

Claim: Most of the volume of the high-dimensional cube is located in its corners.

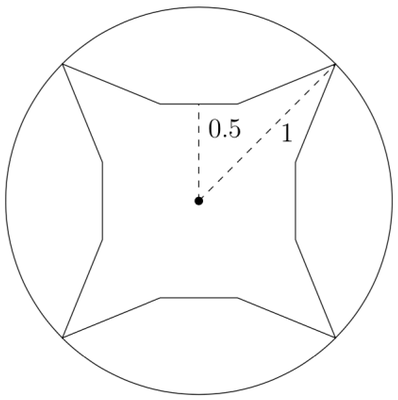


Figure 4: Projections of the 4-dimensional unit sphere and unit cube, centered at the origin (4 of the 16 vertices of the hypercube are shown).

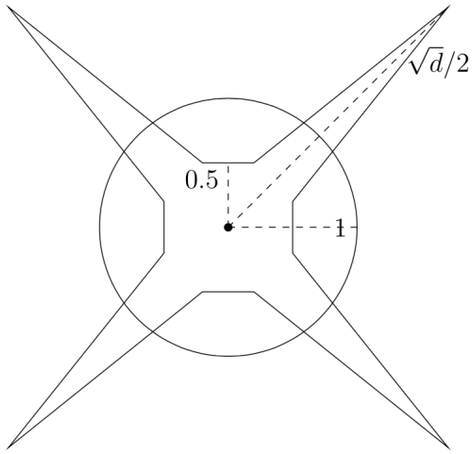
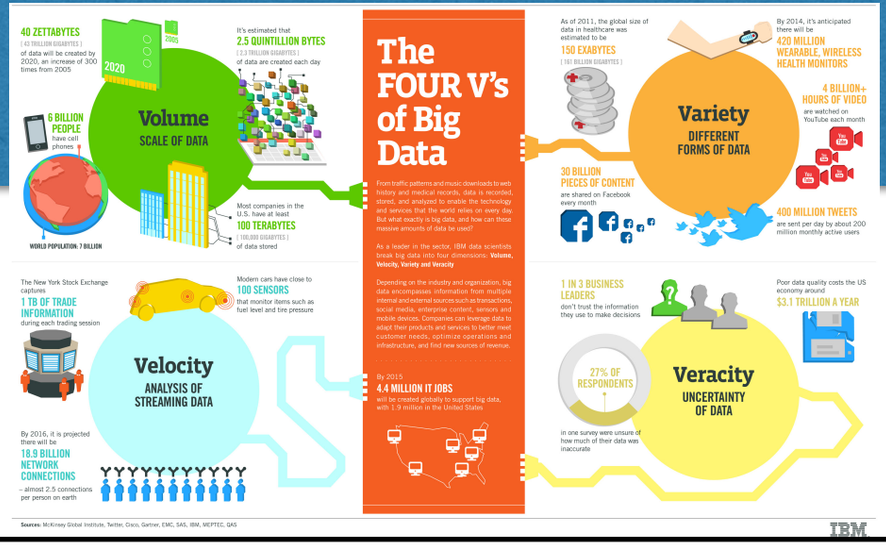


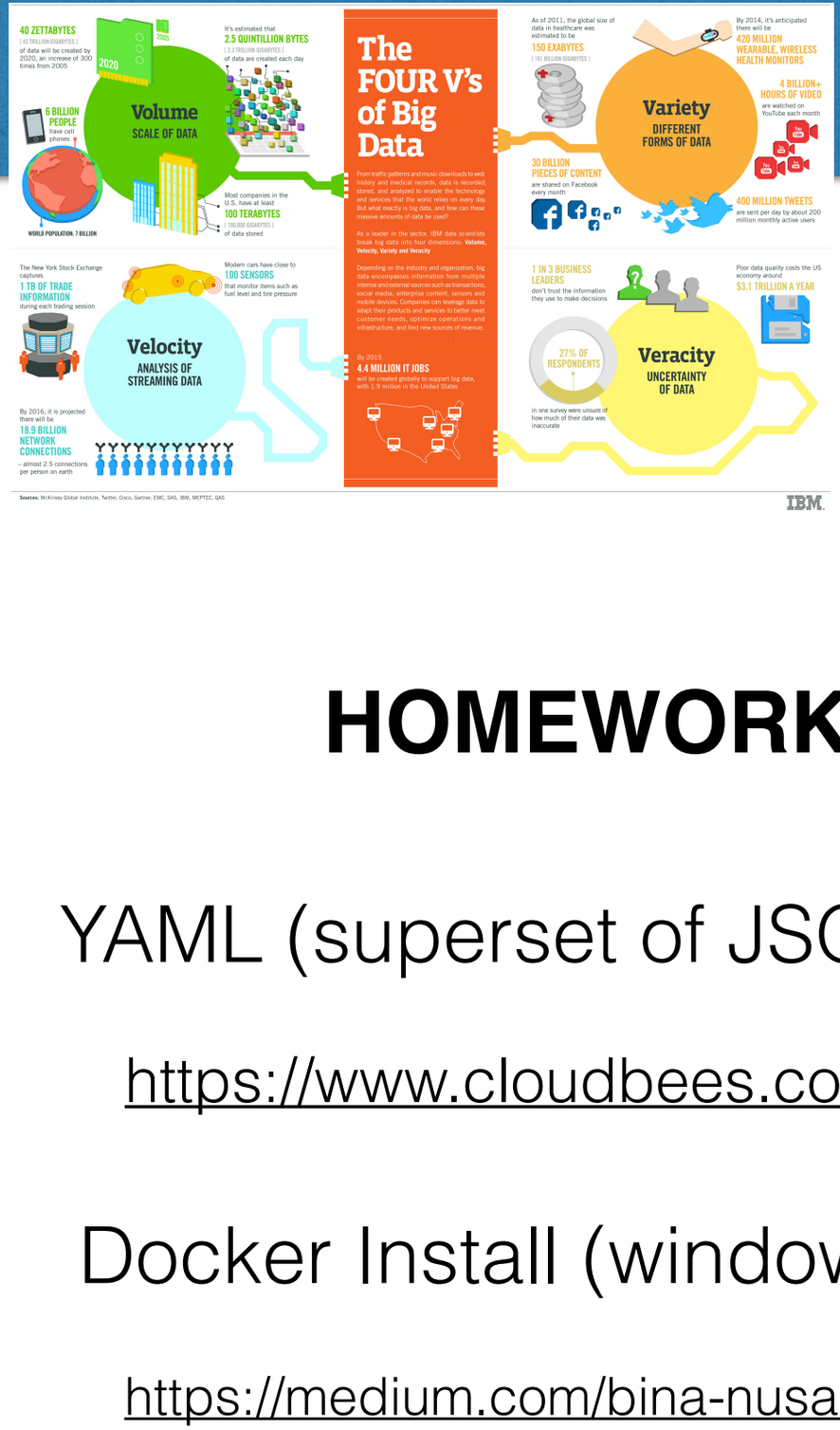
Figure 5: Projections of the d-dimensional unit sphere and unit cube, centered at the origin (4 of the 2^d vertices of the hypercube are shown).



"Técnicas Matemáticas para Big Data"

- Low Veracity
  - Veracity, validity, …, falsification, ethics
  - Discrete/Continuous Markov Chains
  - Bayesian Networks
  - Fuzzy Logics
- High Velocity
  - About data streams (DS)
  - Probabilistic algorithms for DS
- Large Volume
  - Dimension reduction methods
- **High Variety**
  - PEGs for documents classification
  - Introduction to NLP

Q: Difficulties?



HOMEWORK (Basics for Tools Deployment)

YAML (superset of JSON):

https://www.cloudbees.com/blog/yaml-tutorial-everything-you-need-get-started

Docker Install (windows):

https://medium.com/bina-nusantara-it-division/tutorial-for-docker-on-windows-c9af1162f0f2

Docker Tutorial:

https://docs.docker.com/get-started/