Machine Learning

João Gama
jgama@fep.up.pt

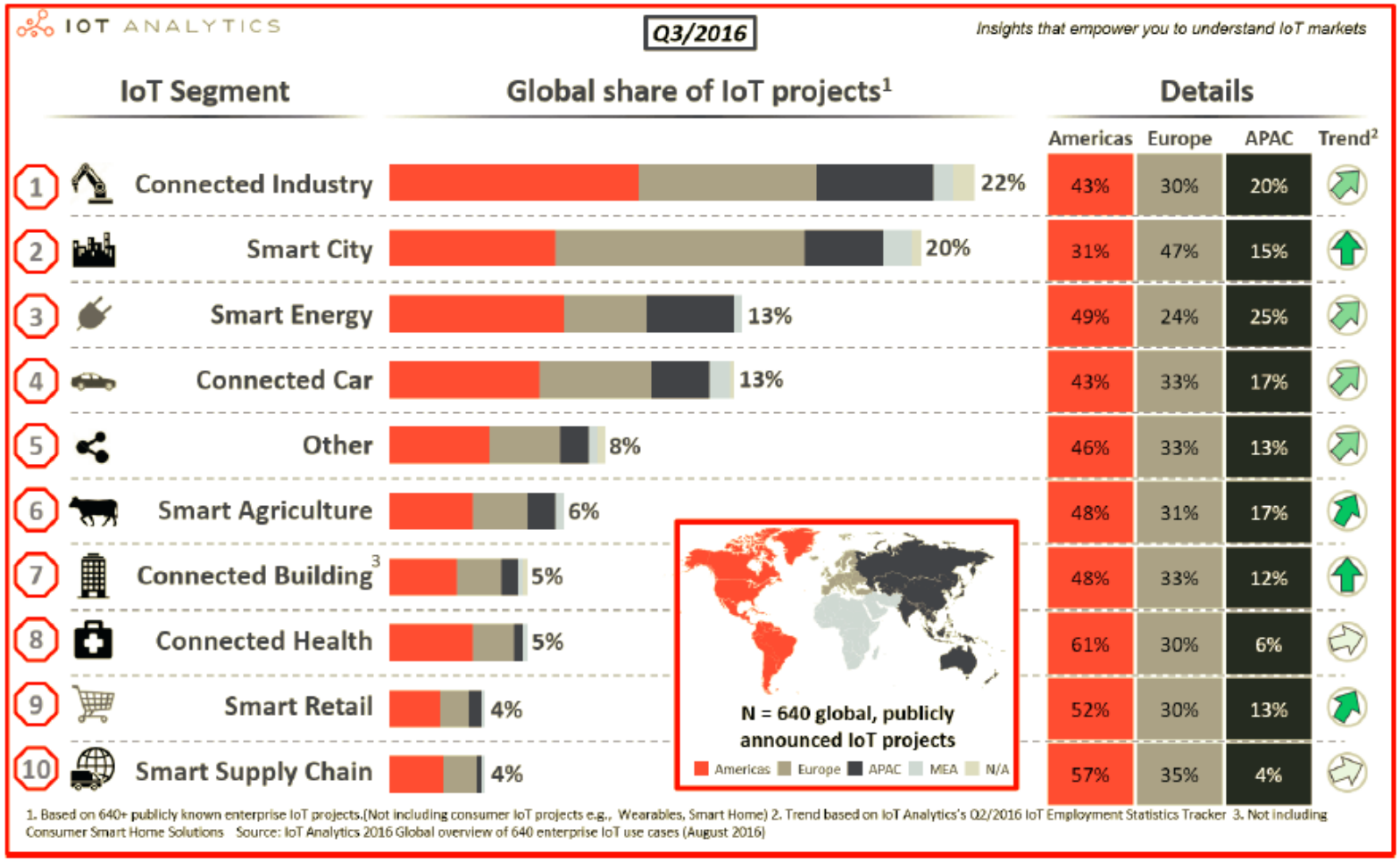
LIAAD-INESC Porto, University of Porto, Portugal

December 2018

Industry 4.0

We have machines that collect, process, and send information to other machines

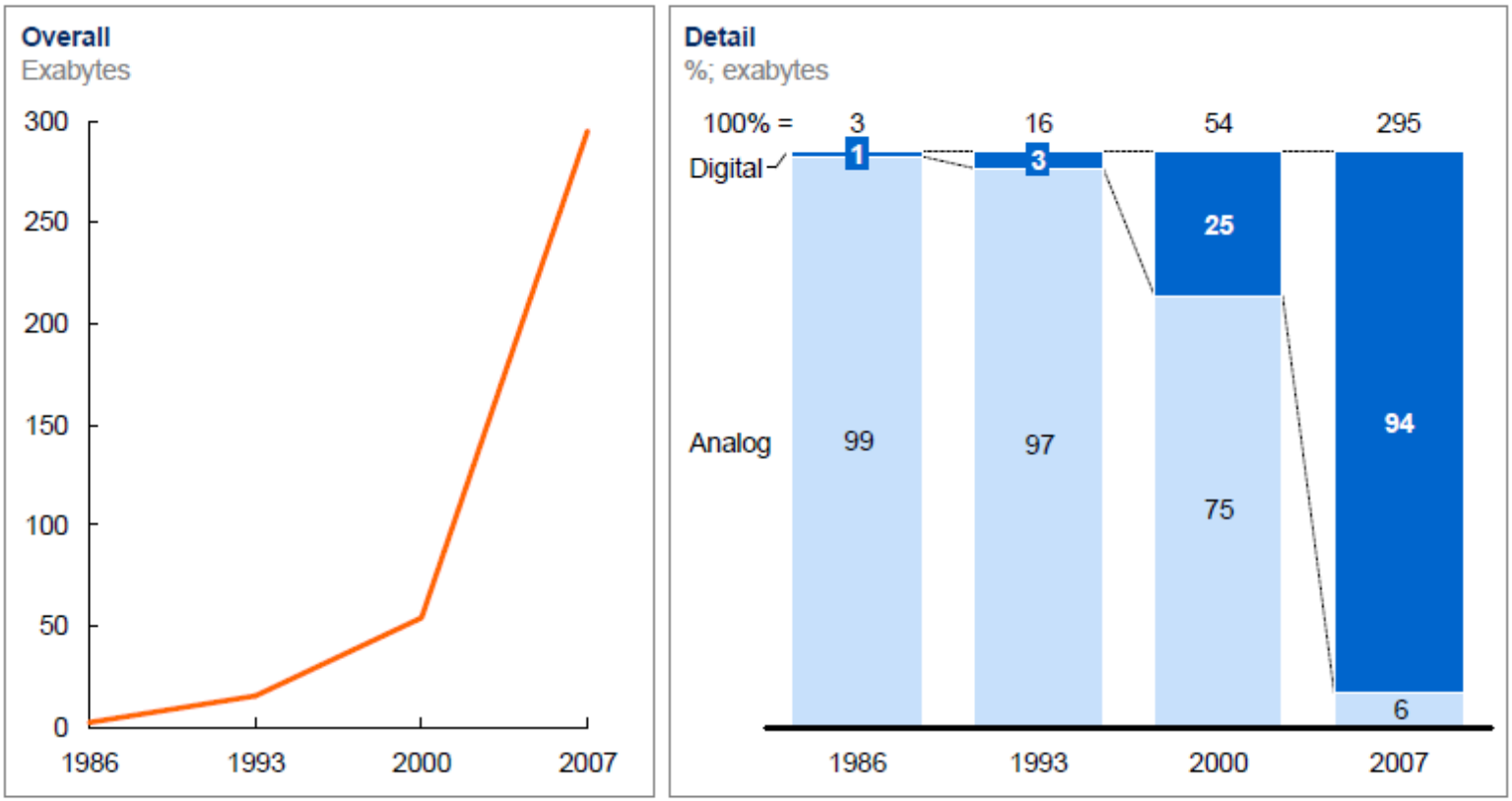
Internet of Things



The Big Bang of digital data ...

Data storage has grown significantly, shifting markedly from analog to digital after 2000

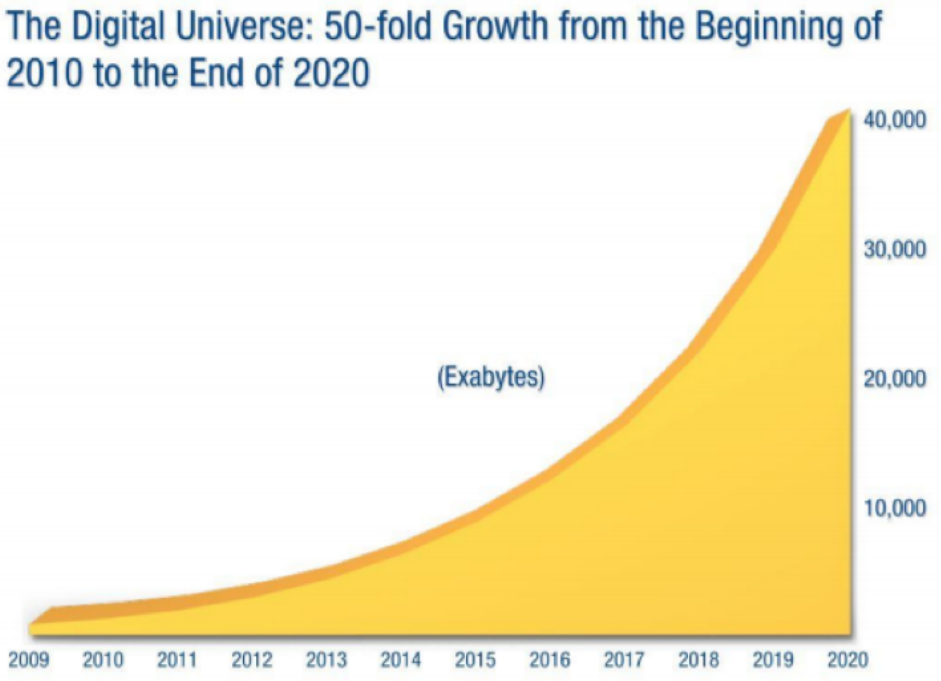
Global installed, optimally compressed, storage



NOTE: Numbers may not sum due to rounding.

SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

The Growth of Digital Data...



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Memory unit	Size	Binary size
kilobyte (kB/KB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}

Tools: time ago ...

Tools seemed quite powerful

Tools

Problems

Tools: nowadays ...

Last few years

Main Goal: Understanding Data

A brief history of big data, the Noam Chomsky way

CNBC

Text Size

Published: Saturday, 23 Nov 2013 | 7:00 AM ET

By: Eric Rosenbaum | CNBC.com

Recommend 47 Twitter 104 +1 6 LinkedIn 15 Share

*The latest news from the fast-evolving world of the **Data Economy**:*

For those familiar with Noam Chomsky, the pioneering linguist whose theory of recursion seeks to find the universal in all human languages, you probably also know that Chomsky often has not-so-nice things to say about the U.S. government, and has also made a career of finding the universal

ChinaFotoPress | Getty Images

Noam Chomsky

Big data is a step forward, but our problems are not lack of access to data, but understanding them. Big data is very useful if I want to find out something without going to the library, but I have to understand it, and that's the problem.

Motivation

- Consultant
- Speaker
- Trends
- Scenarios
- Planning

Global Future

Strategies for a Global Age

Home | What's New | Global Future Reports™ | Book Reviews | Bibliographies | Contact Us

Global Future Report™ January 15th, 2001

10 Emerging Technologies That Will Change the World

© Dr. Terry J. van der Werff, CMC

MIT's Technology Review has identified 10 emerging areas of technology that will soon have a profound impact on the economy and how we live and work.

Regular readers of *Global Future Report*™ know I am a sucker for lists of things that matter. I even write lists of my own, e.g. my "Ten Tips for Harnessing the Future" or the four forces converging to alter global telecommunications in "Calling the Future."

To launch the New Millennium the January/February issue of *Technology Review*, MIT's magazine of innovation, focuses on "The Technology Review Ten" - "10 emerging areas of technology that will soon have a profound impact on the economy and how we live and work." For each, one innovator's work is highlighted.

Drum roll, please! The ten emerging technologies that will change the world are:

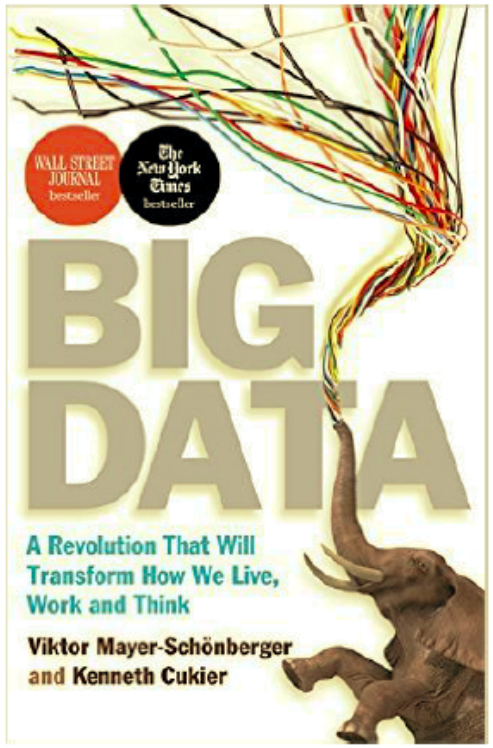
- **Brain-Machine Interfaces** - In essence, researchers try both to understand how the brain works and to use this knowledge to implant electrodes in specific parts of the brain to permit control of computers, robotic arms, or other artificial devices designed to restore lost sensory and motor functions.
- **Flexible Transistors** - Silicon does not bend readily, so a new class of hybrid materials are being developed that marry the speed of inorganic compounds with the flexibility of organic polymers. They have the advantage of being able to be dissolved and printed onto paper or plastic as if they were ink particles.
- **Data Mining** - Ever get an e-mail from amazon.com suggesting a book that relates to an earlier one you ordered from them? You have been the subject of data mining, which is nothing more than the extraction of meaningful information and patterns from huge data sets.
- **Digital Rights Management** - Think Napster! The Internet permits the sharing of digital content far and wide at little cost. But originators of the content - articles, data, graphics, songs - may lose control of their intellectual property. Digital rights management combines encryption, with payment software to

Motivation

McKinsey Global Institute

June 2011

Big data: The next frontier for innovation, competition, and productivity



Definitions

Extração de Conhecimento de Dados;
Gama, Carvalho, Faceli, Lorena, Oliveira; 3 edition,
Silabo, 2017

Definitions

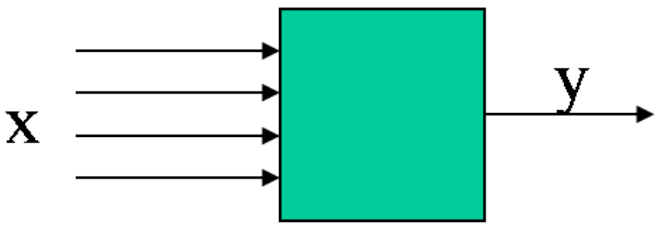
- *Self-constructing or self-modifying representations of what is being experienced for possible future use* Michalski, 1990
- *Analysis of observational data to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful for the data owner* Hand, Mannila, Smyth, 2001
- *Obter representações em compreensão a partir de representações em extensão.*

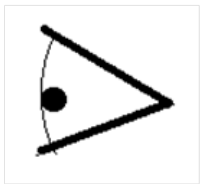
Scientific Areas

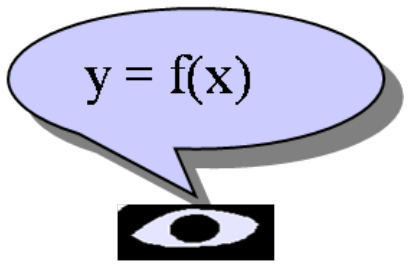
Scientific Areas

- Statistics
  - Inference
- Computer Science
  - Artificial Intelligence
  - Machine Learning
- Data Bases
  - Multidimensional Data Bases

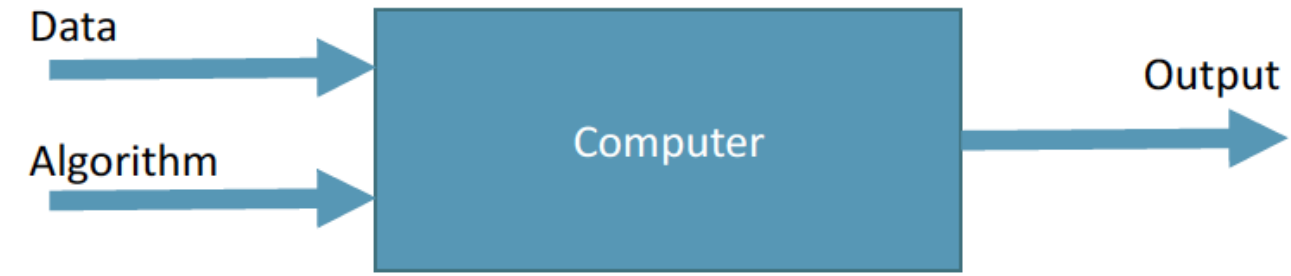
Outline

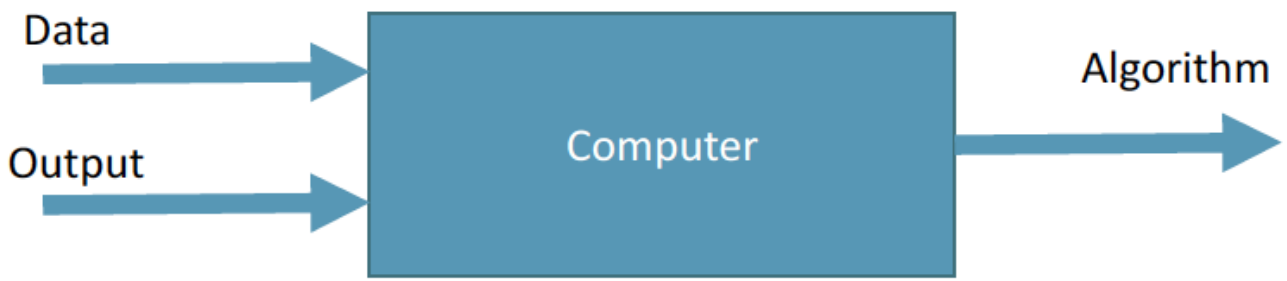

x
y



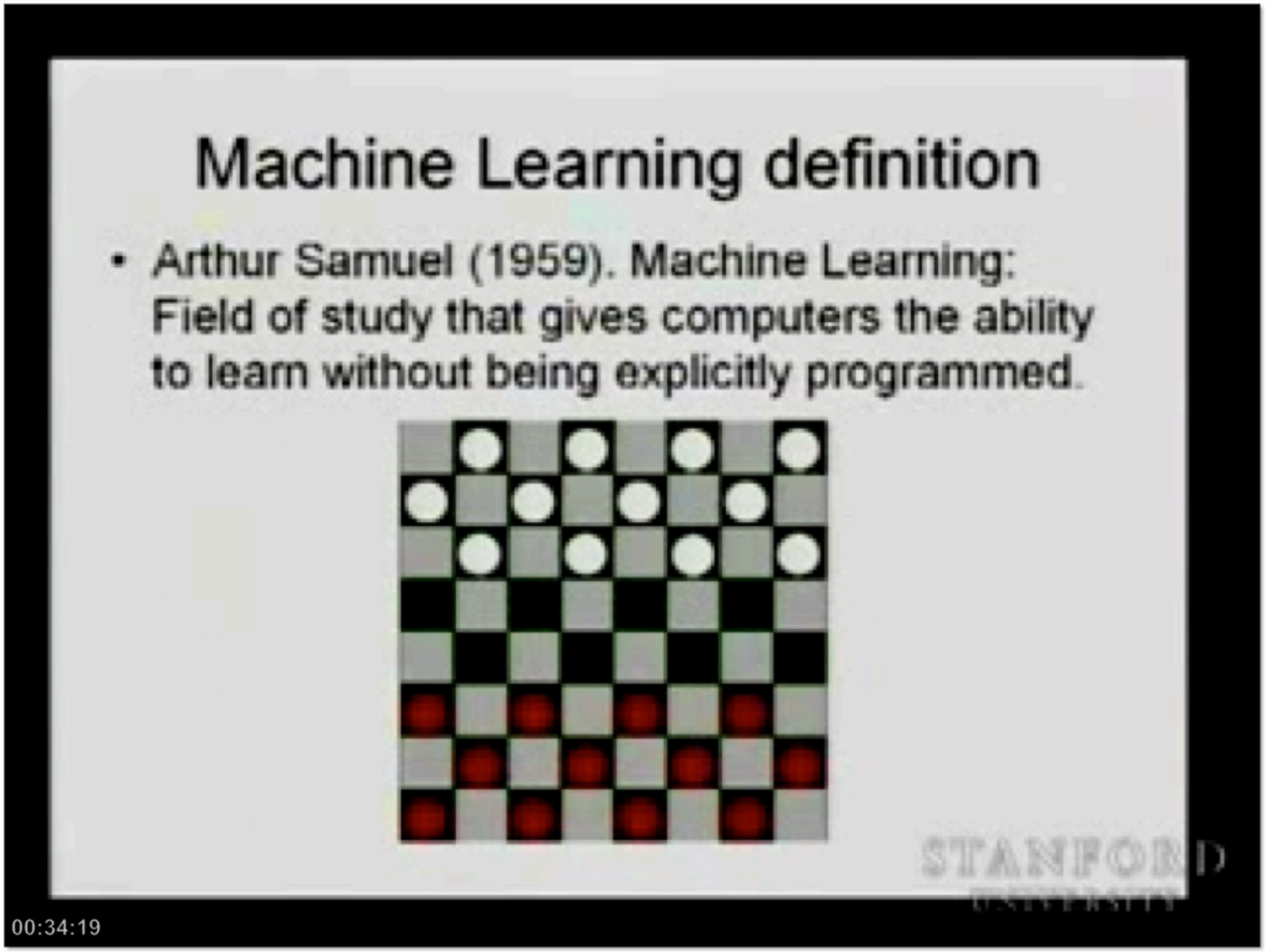

y = f(x)

Self-Programming machines

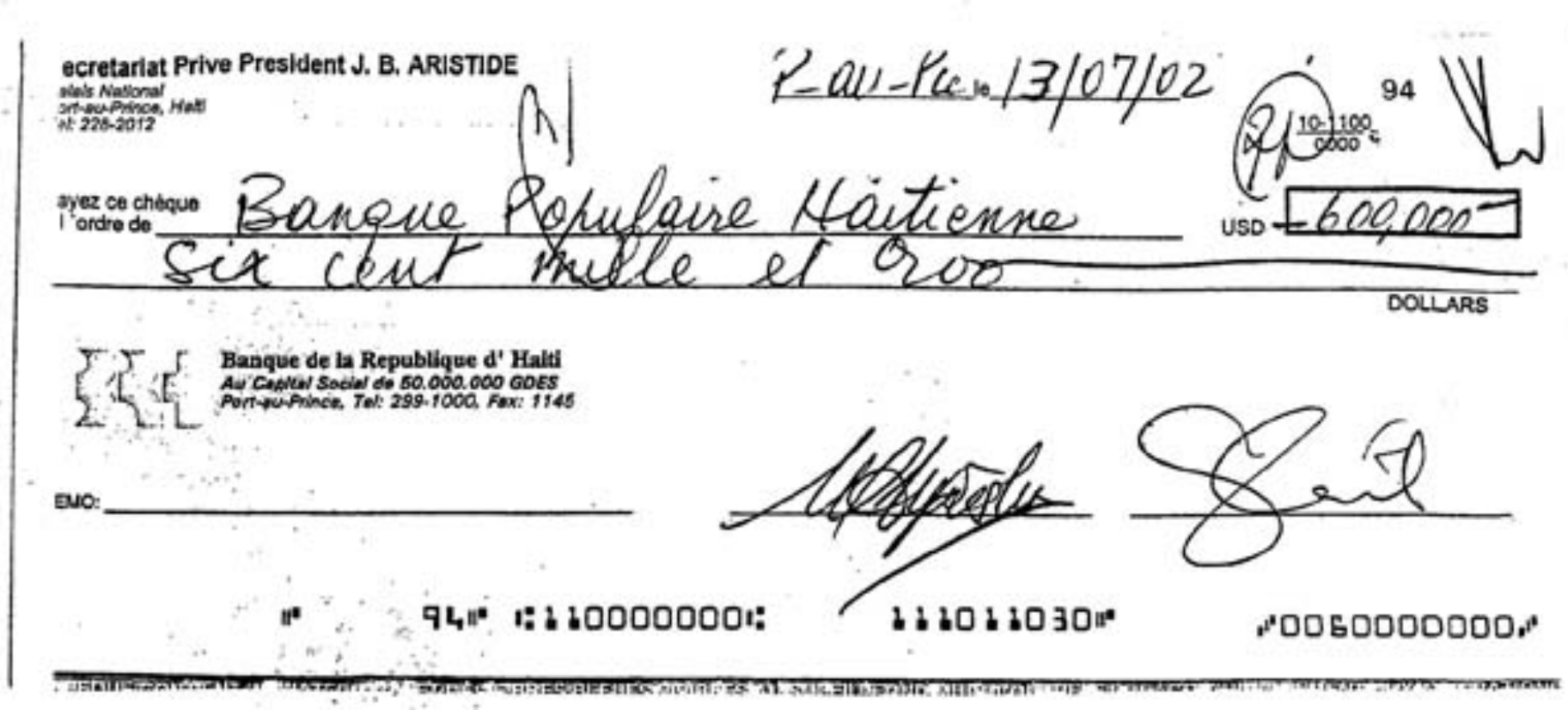




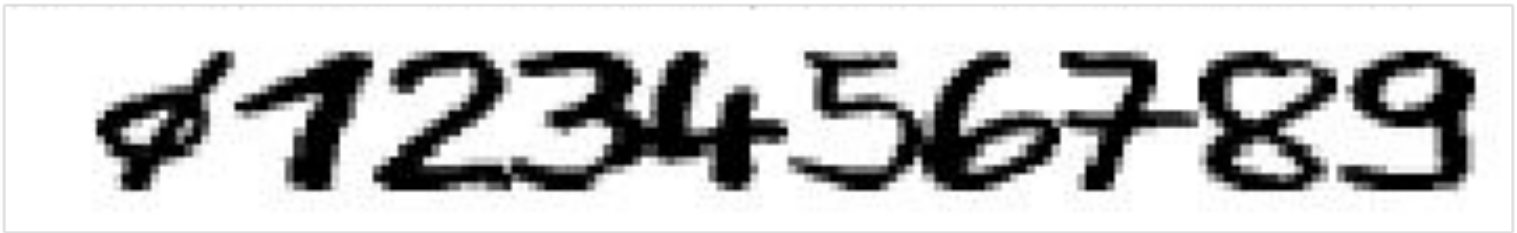
Lecture 1 - The Motivation & Applications of Machine Learning.flv



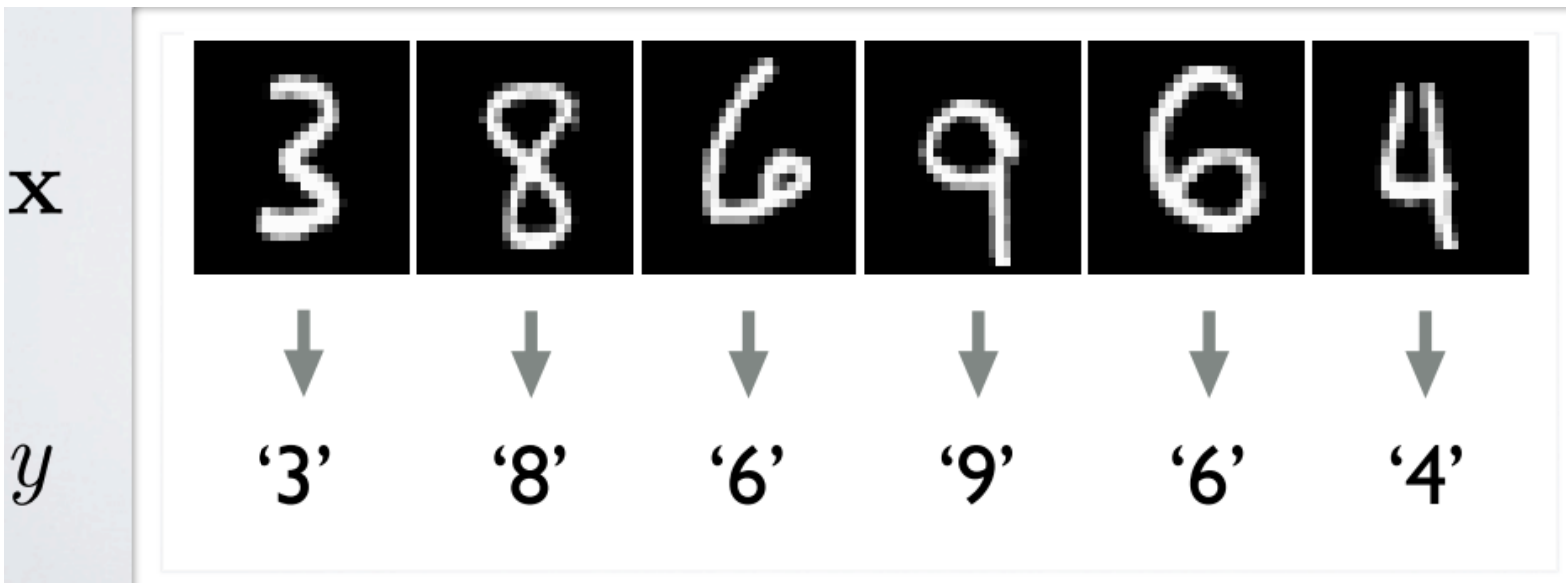
Bank Check


ecretariat Prive President J. B. ARISTIDE
Port-au-Prince, Haiti
13/07/02
94
10-1100
0000
ayez ce chèque
l'ordre de
Banque Populaire Haïtienne
USD
600,000
Six cent mille et 00
DOLLARS
Banque de la Republique d' Haiti
Au Capital Social de 50.000.000 GDES
Port-au-Prince, Tel: 299-1000, Fax: 1145
EMO:
94
0000000
1030
0060000000

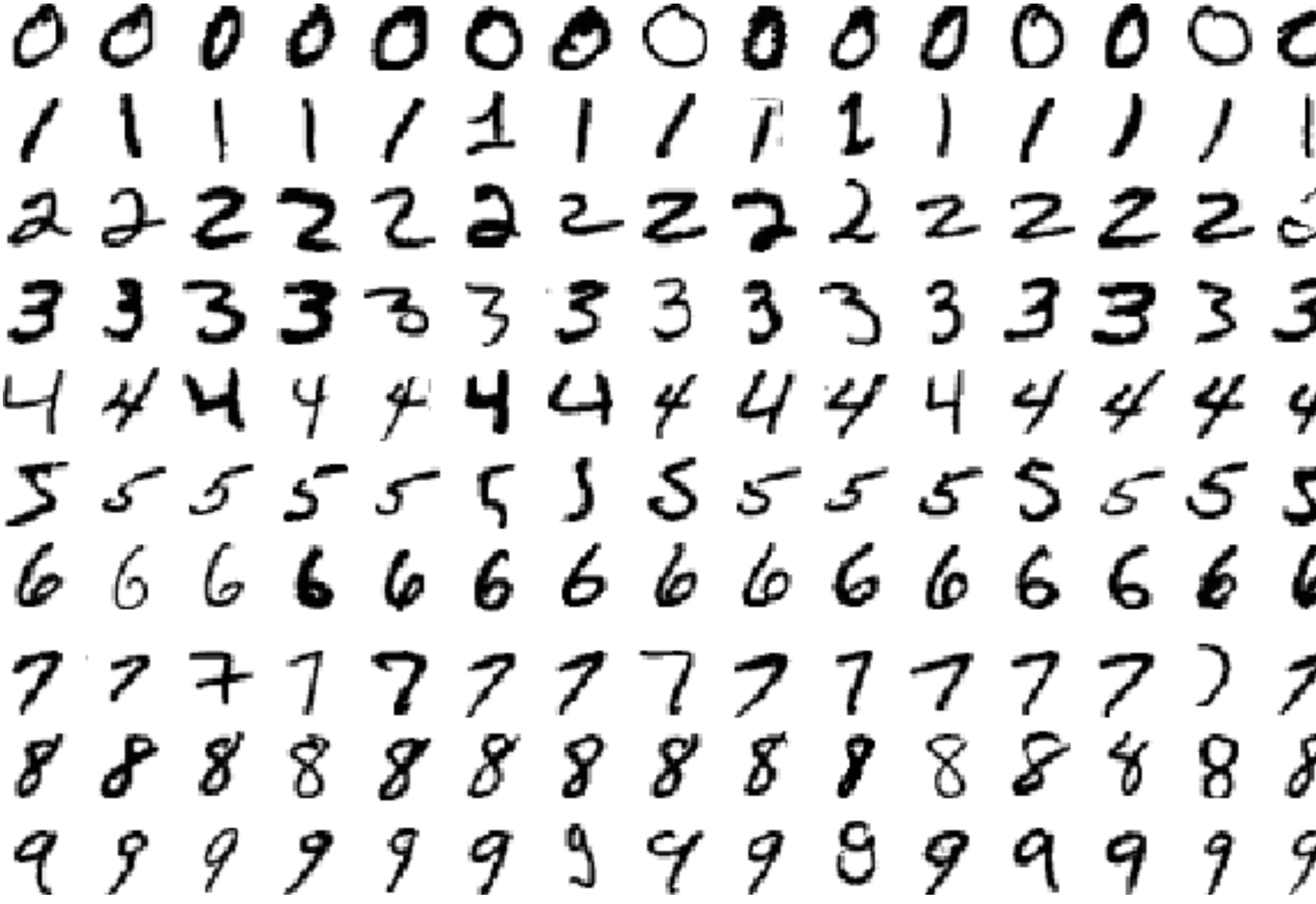
Postal Codes



Postal Codes



Postal Codes



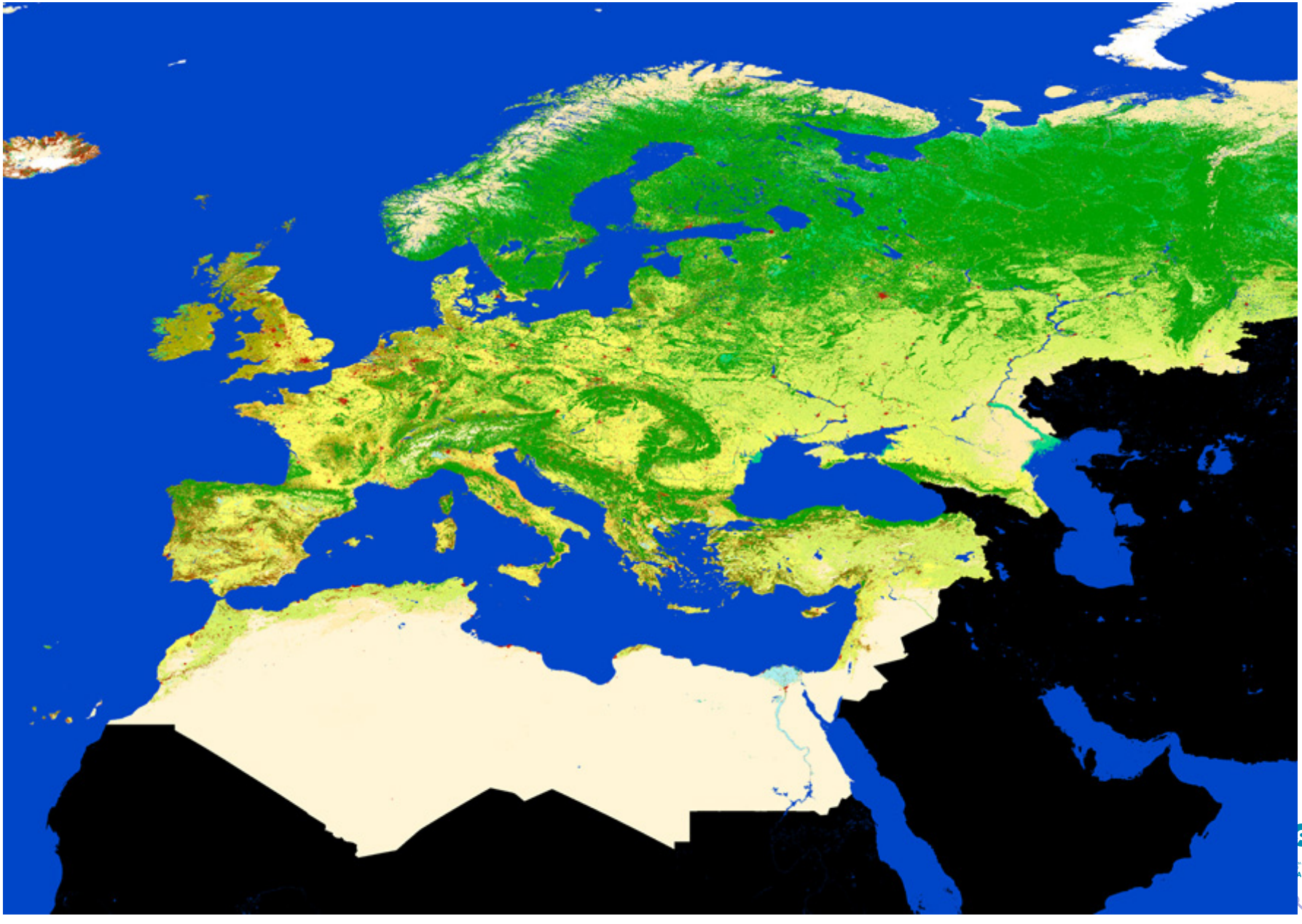
Satélite traça mapa ambiental do continente europeu em tempo recorde - Publico 25.11.2010

Segundo a Agência Espacial Europeia (ESA), o mapa foi realizado em nove meses, a partir dos dados recolhidos pelo satélite Envisat, graças ao seu espectógrafo MERIS (Medium Resolution Imaging Spectrometer). Este mapa, resultado do projecto GlobCorine, *é o primeiro do género a ser produzido em tão pouco tempo, nove meses em comparação com vários anos*, salienta.

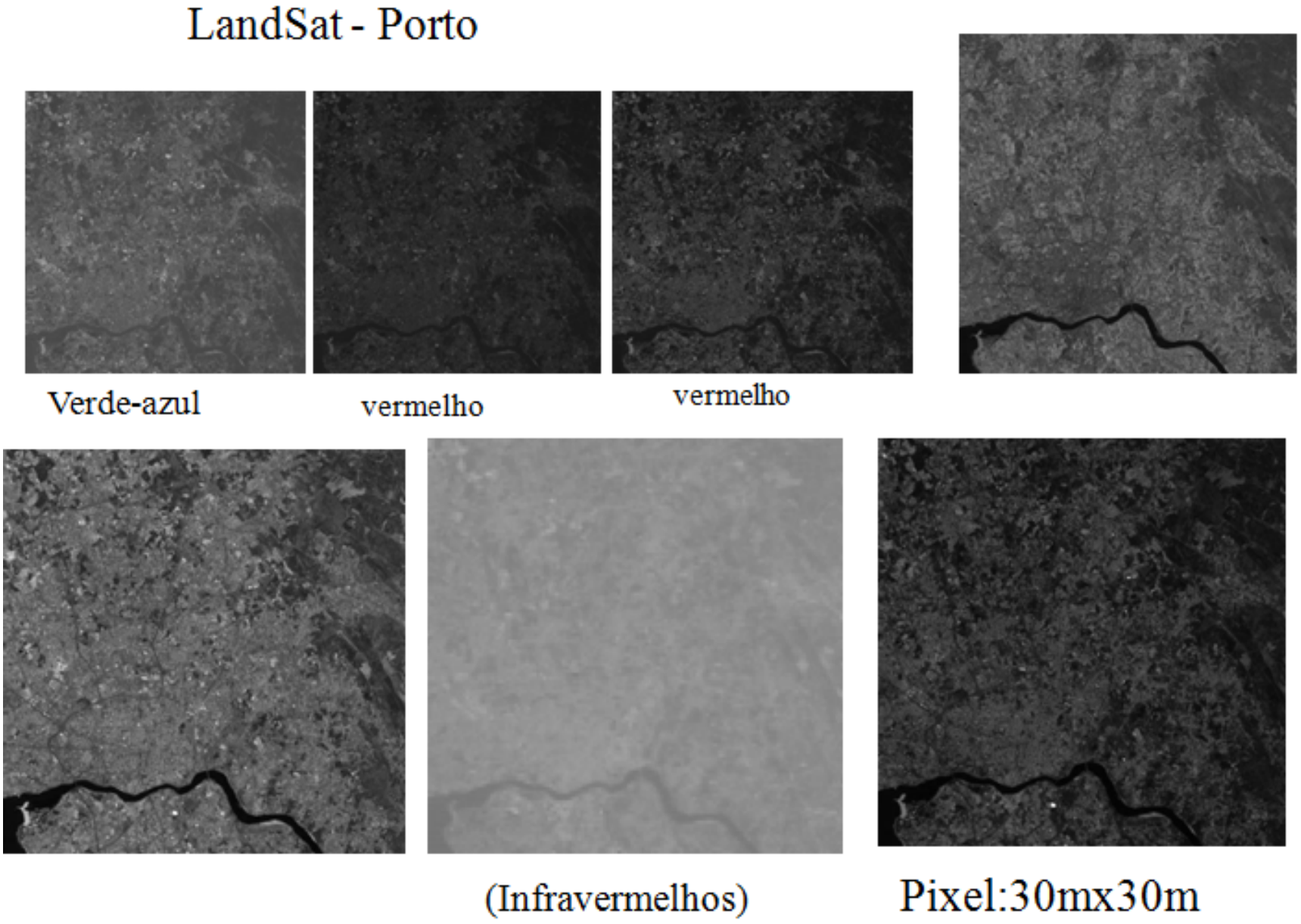
O objectivo foi desenvolver um serviço automático de produção de cartografia de ocupação e utilização do solo à escala europeia, incluindo os países da bacia mediterrânica e a Rússia. Agora, o mapa relativo a 2009, deverá ser actualizado todos os anos, algo considerado essencial para as agências de Ambiente dos vários países. *Satisfazer as necessidades ambientais da Europa exige informação sobre o uso do solo consistente e actualizada*, explica a ESA.

Finalmente podemos ter informação relevante e actual que suporte os processos de tomada de decisão, comentou Chris Steenmans, responsável pelo projecto na ESA.

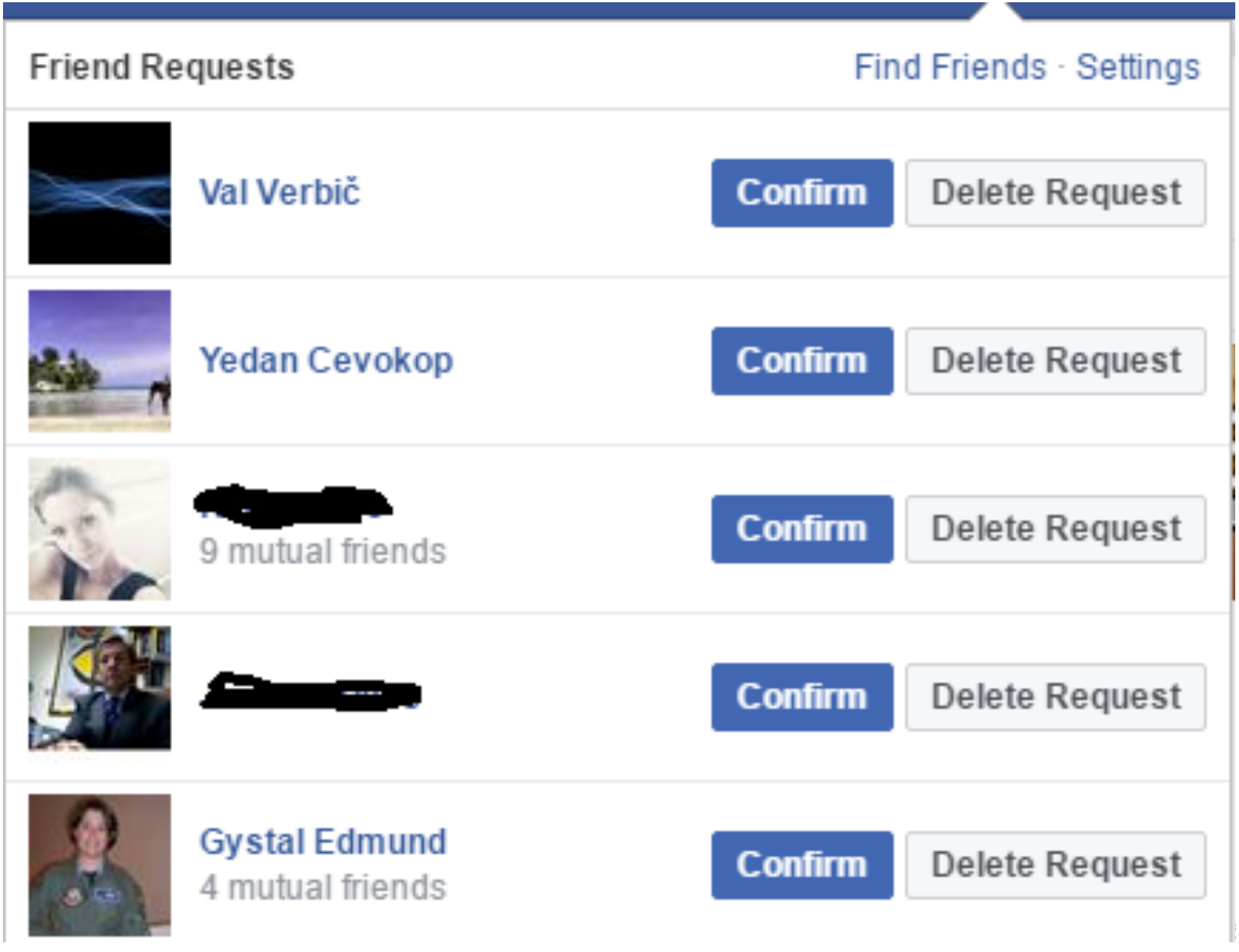
Predicting Land Usage



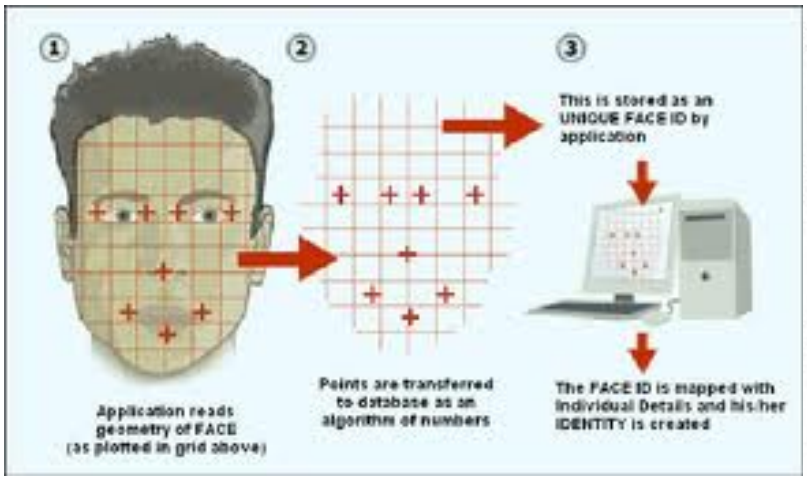
Porto


LandSat - Porto
Verde-azul
vermelho
vermelho
(Infravermelhos)
Pixel:30mx30m

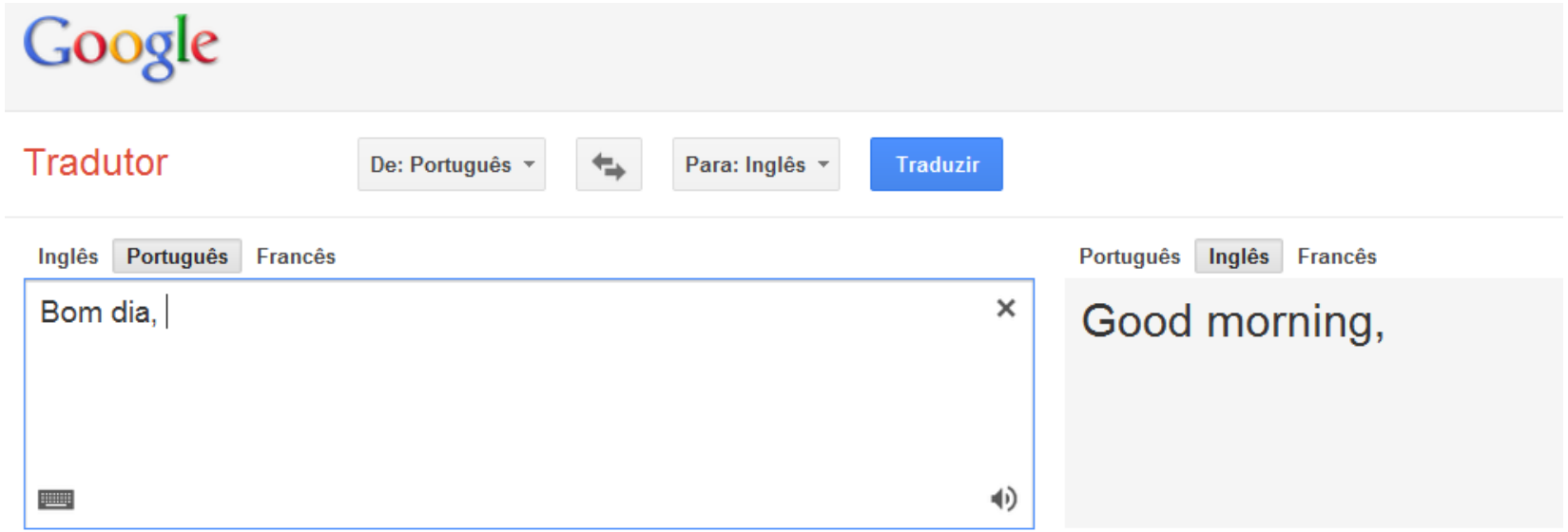
FACEBOOK: The Missing Link



Face Recognition



Google (2)


Google
Tradutor
De: Português
Para: Inglês
Traduzir
Inglês Português Francês
Bom dia,
Português Inglês Francês
Good morning,

Google (3)

Google Photos

From Wikipedia, the free encyclopedia

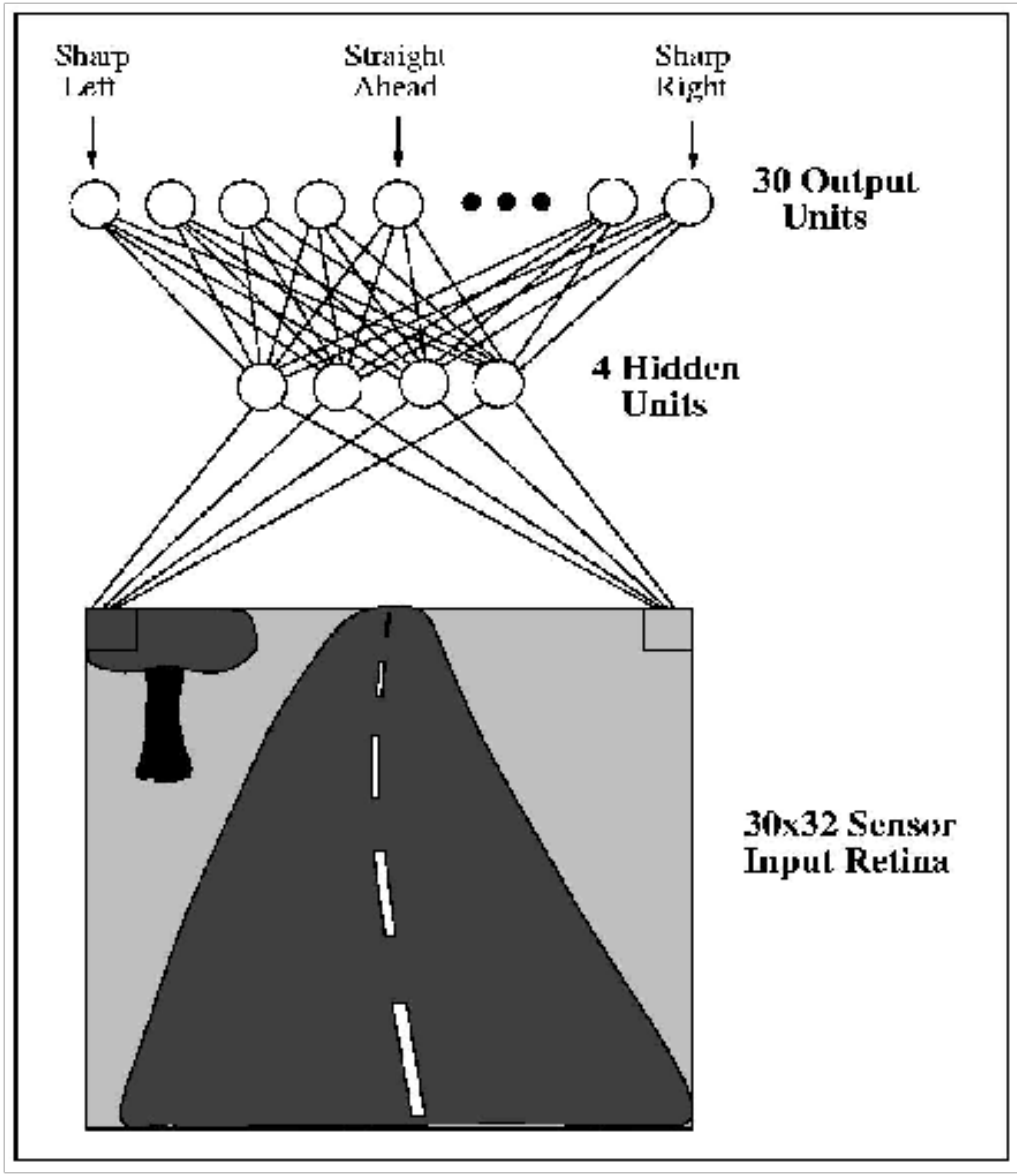
Google Photos is a photo sharing and storage service developed by Google. It was announced in May 2015 and spun out from Google+, the company's social network.

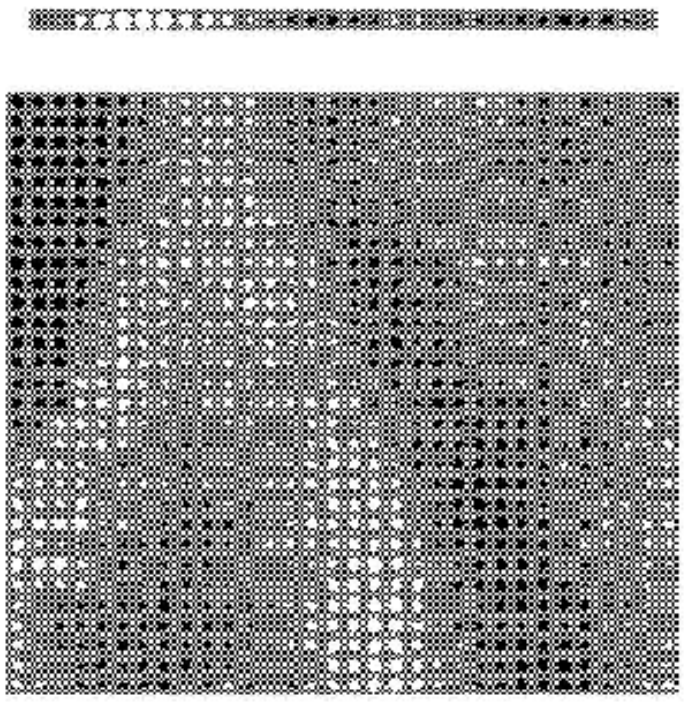
Google Photos gives users free, unlimited storage for photos up to 16 megapixels and videos up to 1080p resolution. The service automatically analyzes photos, identifying various visual features and subjects. Users can search for anything in photos, with the service returning results from three major categories: People, Places, and Things. Google Photos recognizes faces, grouping similar ones together; geographic landmarks (such as the Eiffel Tower); and subject matter, including birthdays, buildings, animals, food, and more. Google implements different forms of machine learning into the Photos service, particularly its recognition of photo contents, as well as enabling features that can automatically generate albums, animate similar photos into quick videos, surface past memories at significant times, and improve the quality of photos and videos. In May 2017, Google announced several updates to Google Photos, including reminders for and suggested sharing of photos, shared photo libraries between two users, and physical albums, with Photos automatically suggesting collections based on face, location, trip, or other distinction.

Google Photos received critical acclaim after its decoupling from Google+ in 2015. Reviewers liked the updated Photos service for its recognition technology, search, apps, and loading times. Nevertheless, privacy concerns were raised, including Google's motivation for building the service, as well as its relationship to governments and possible laws requiring Google to hand over a user's entire photo history. Google Photos has seen strong user adoption. It reached 100 million users after five months, 200 million after one year, and 500 million as of May 2017, with Google announcing that over 1.2 billion photos are uploaded to the service every day, with the grand total of all uploaded content measuring over 13.7 petabytes of storage.

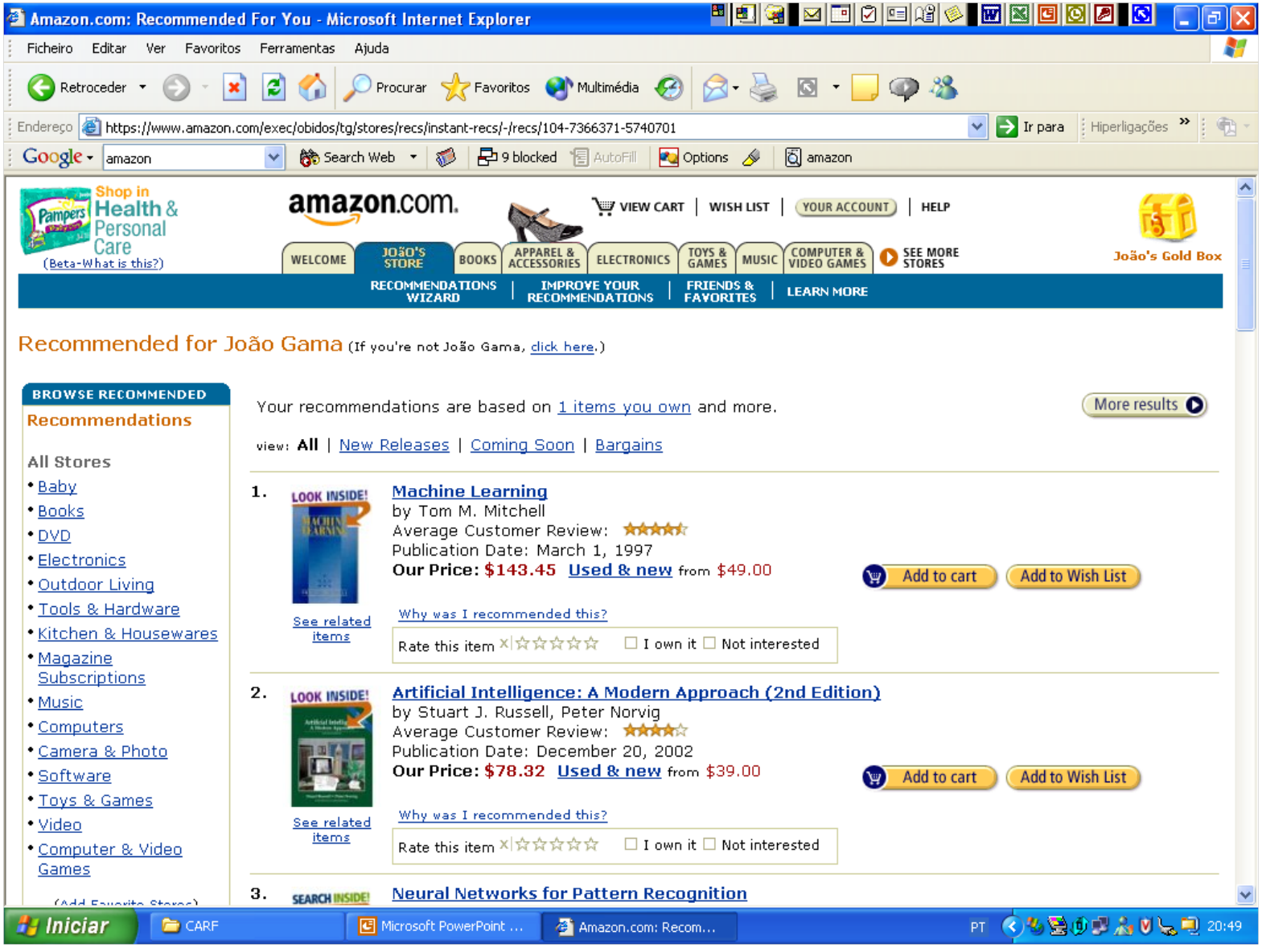
Autonomous Vehicles

Autonomous Vehicles





Recommender Web Sites



Jeopardy!


PIENSE
THINK
ΣΚΕΨΟΥ
DENKE
सोचिए
PENSER
$300,000
Who is Stoker?
$1,000
$1,000,000
Who is Bram Stoker?
$ 17,973
$200,000
$5600

Diagnosis

- Sensors:
  - gsr_low_average
  - heat_flux_high_average
  - near_body_temp_average
  - pedometer
  - skin_temp_average
  - longitudinal_accelerometer_SAD
  - longitudinal_accelerometer_average
  - transverse_accelerometer_SAD
  - transverse_accelerometer_average

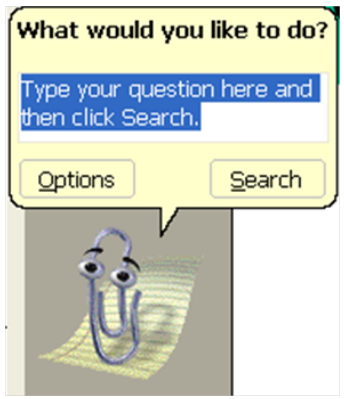
The SenseWear armband, shown in the figure below, is a sleek, wireless and accurate wearable body monitor that enables continuous physiological monitoring outside the laboratory.

Help Systems

ERIC HORVITZ, a researcher at Microsoft and a guru in the field of Bayesian statistics, feels bad about the paperclip, but he hopes his latest creation will make up for it. The paperclip in question, as even casual users of Microsoft's Office software will be aware, is a cheery character who pops up on the screen to offer advice on writing a letter or formatting a spreadsheet. That was the idea, anyway. But many people regard the paperclip as annoyingly over-enthusiastic, since it appears without warning and gets in the way.

Mobile Manager evaluates incoming e-mails on a user's PC and decides which are important enough to forward to a pager, mobile phone or other e-mail address. Its Bayesian innards give it an almost telepathic ability to distinguish junk mail from genuinely important messages.

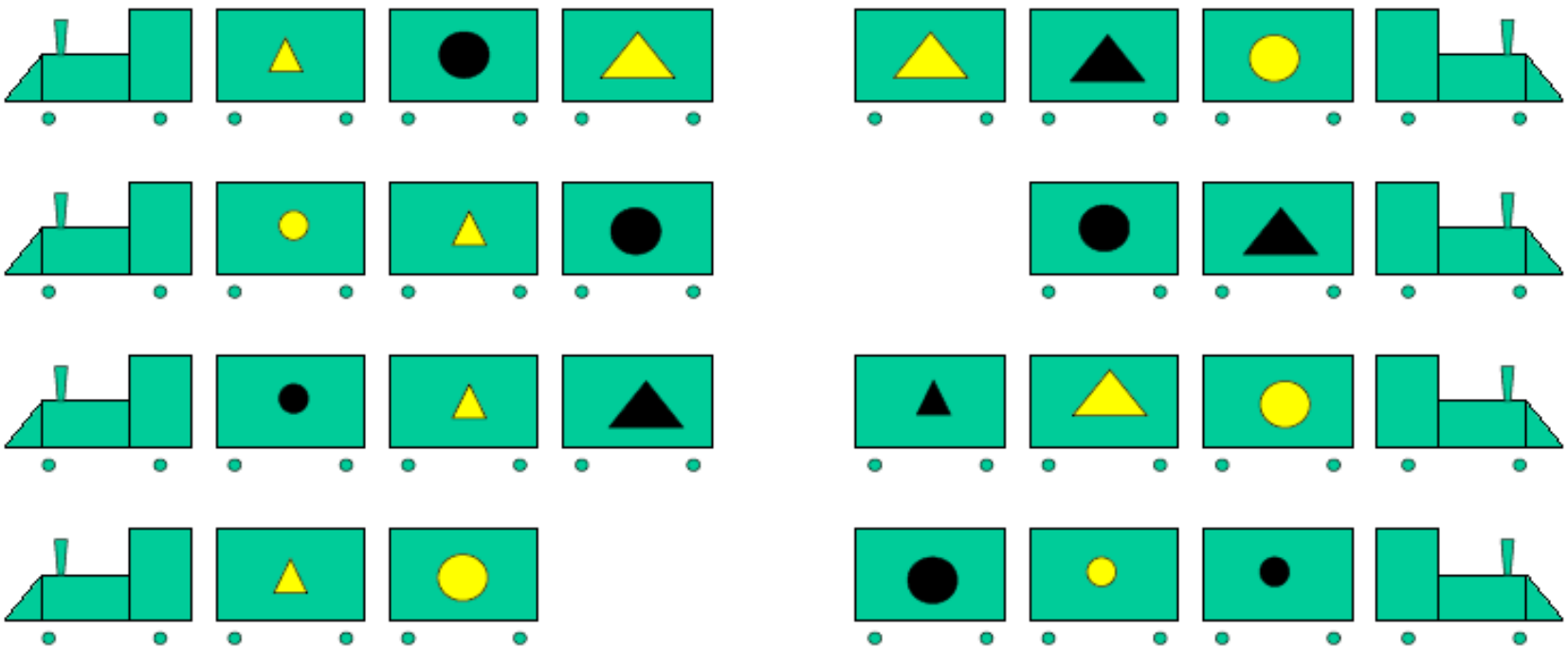
The *Clip* is an interface for a Bayesian Network:



Outline

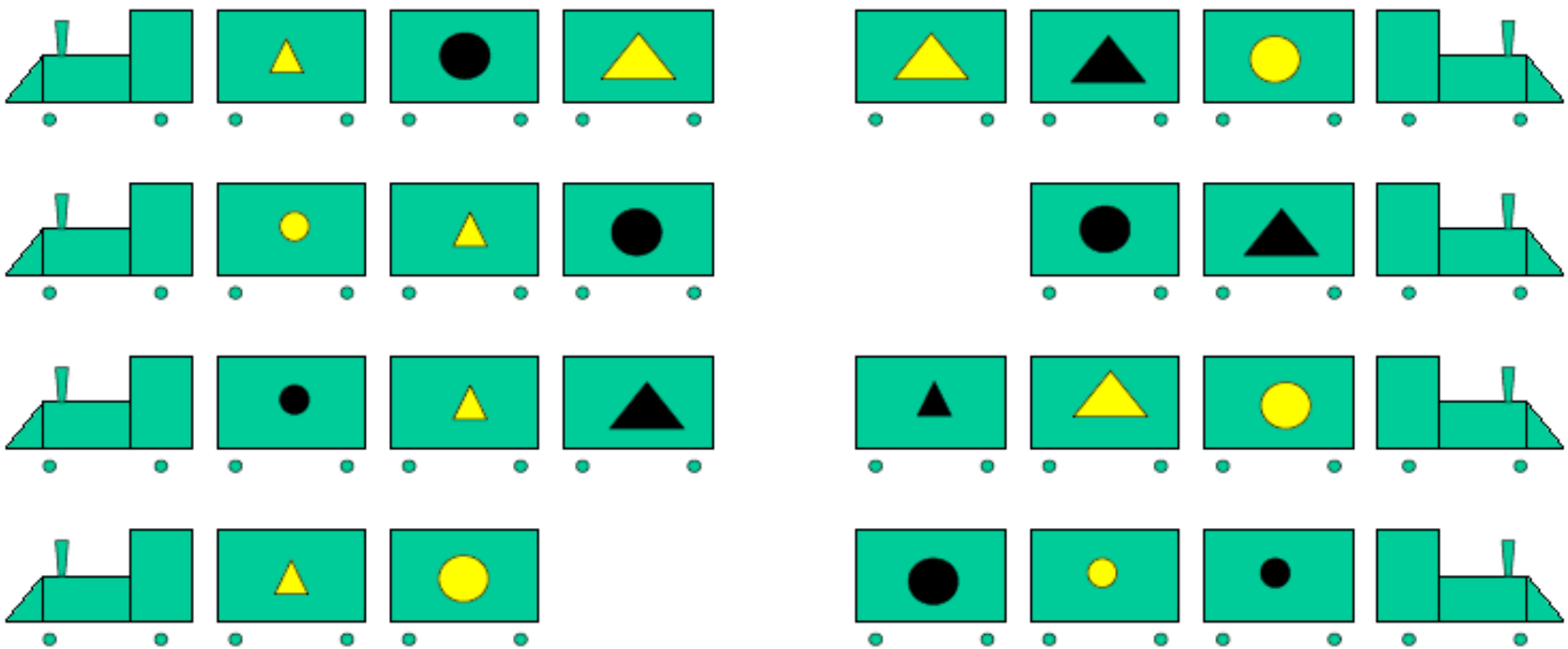
The Michaslky trains: data

- **Problema**
  - O que caracteriza os comboios que vão para o Ocidente (←) relativamente aos que vão para o Oriente (→)?
- **Dados**



The Michaslky trains: patterns

- **Padrões observados (entre muitos)**
  - Os que vão para Ocidente têm um vagão com um pequeno triângulo amarelo, os que vão para Oriente não têm.
  - Todos os comboios que têm um triângulo amarelo, têm também um círculo



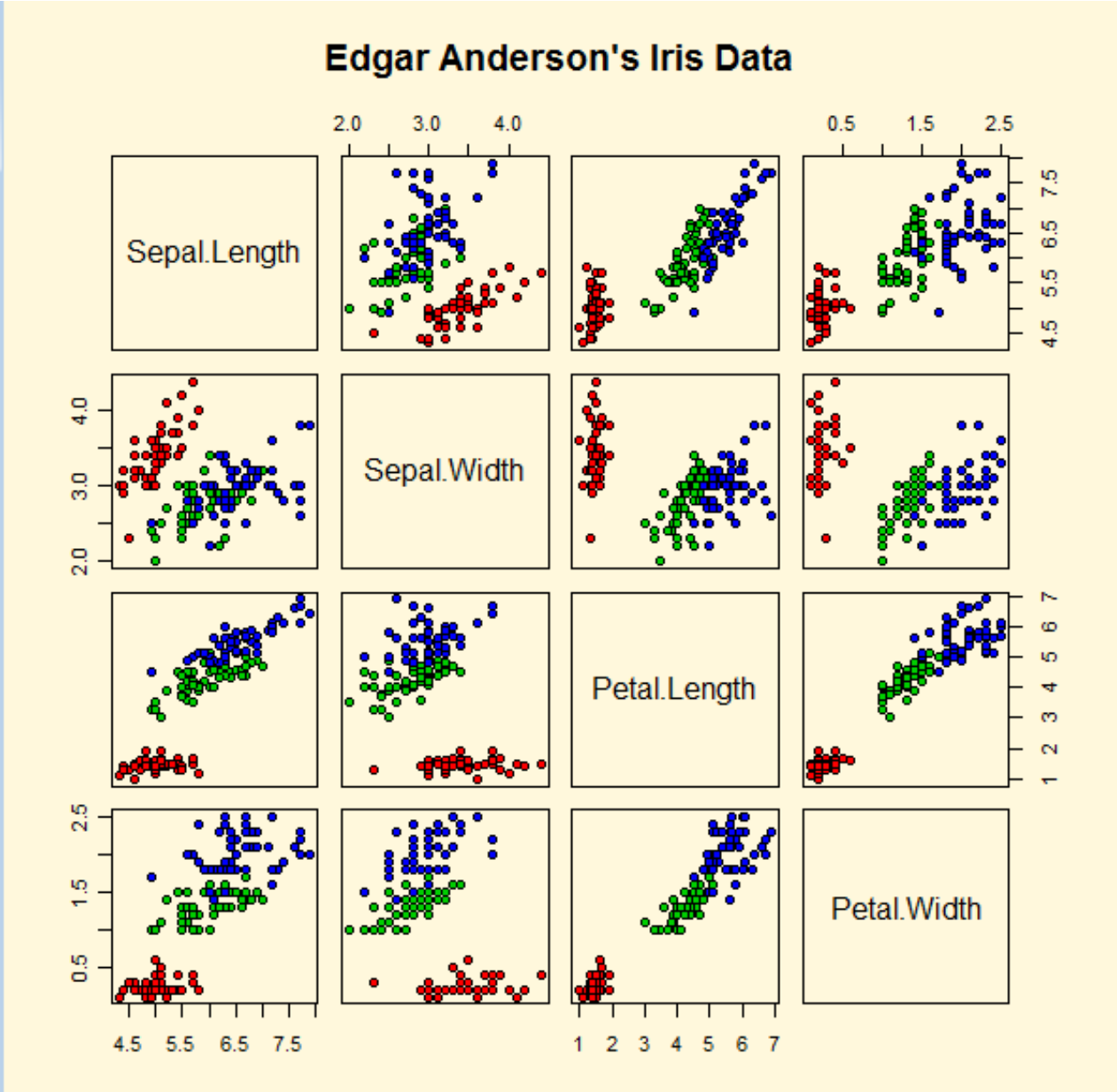
The Iris Problem

The Iris problem: data

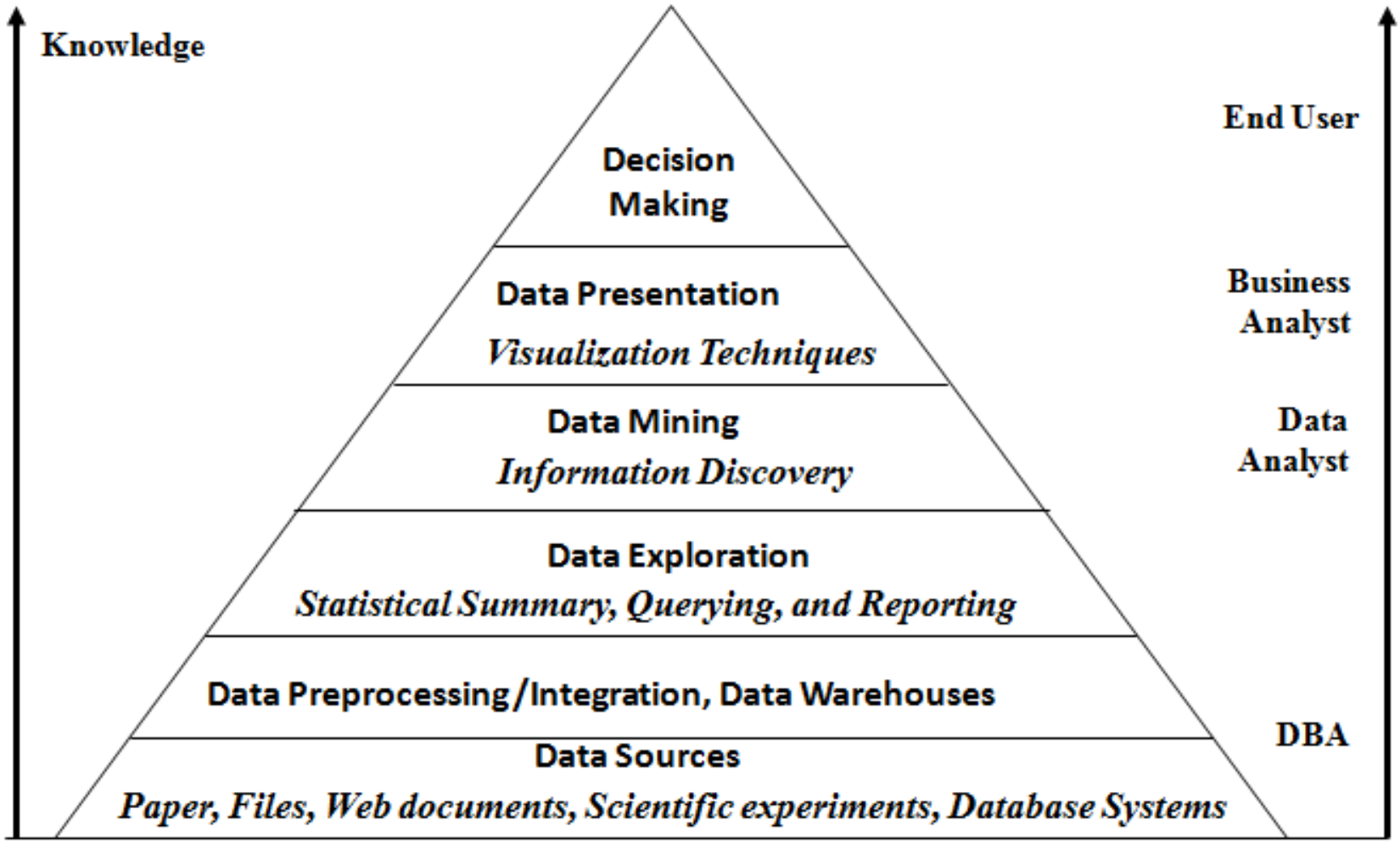
```
> data(iris)
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1
> 
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
5.8	2.7	5.1	1.9	virginica
7.1	3.0	5.9	2.1	virginica
6.3	2.9	5.6	1.8	virginica

Where are the patterns?



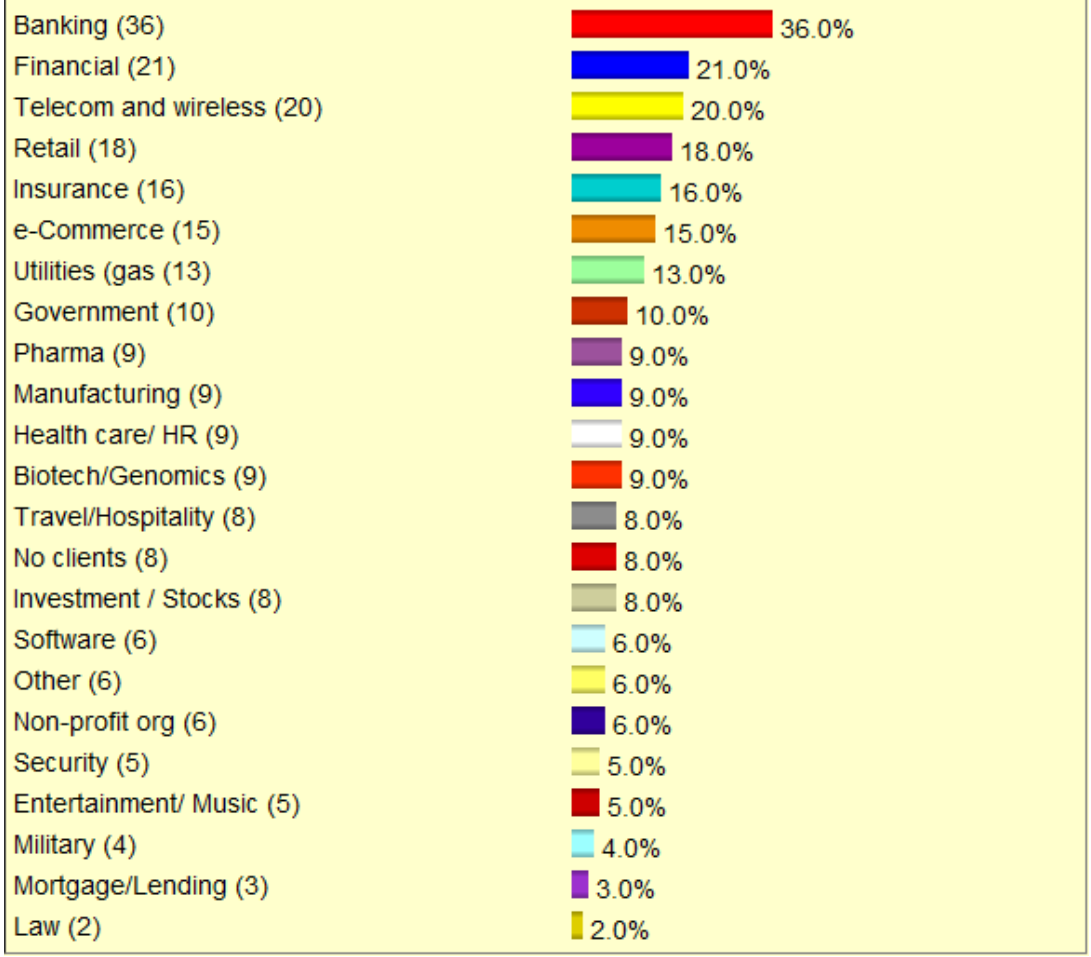
Data and Knowledge



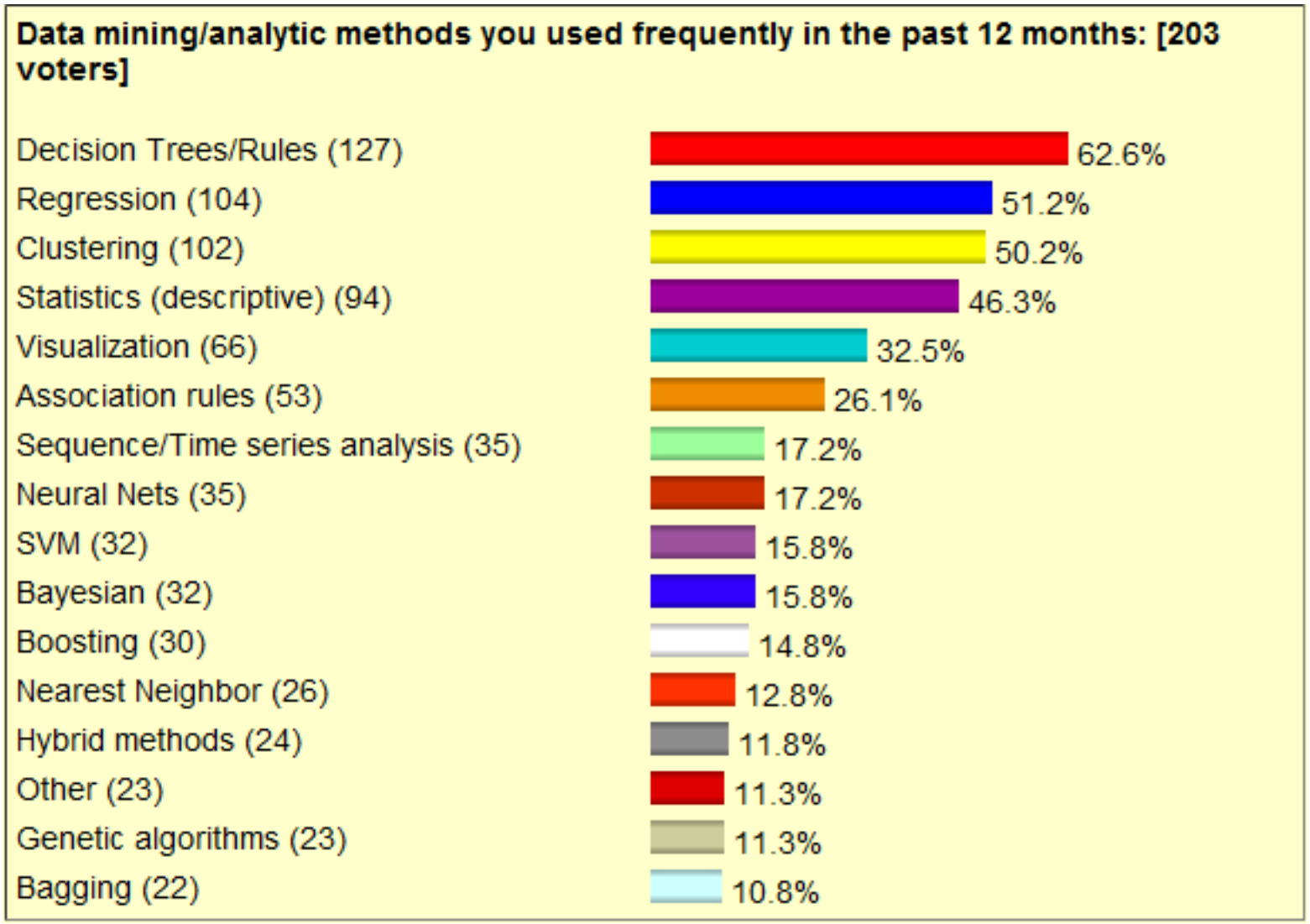
Where is DM used?

Poll

In what industries/sectors were your data mining clients in 2007-2008? [100 voters]

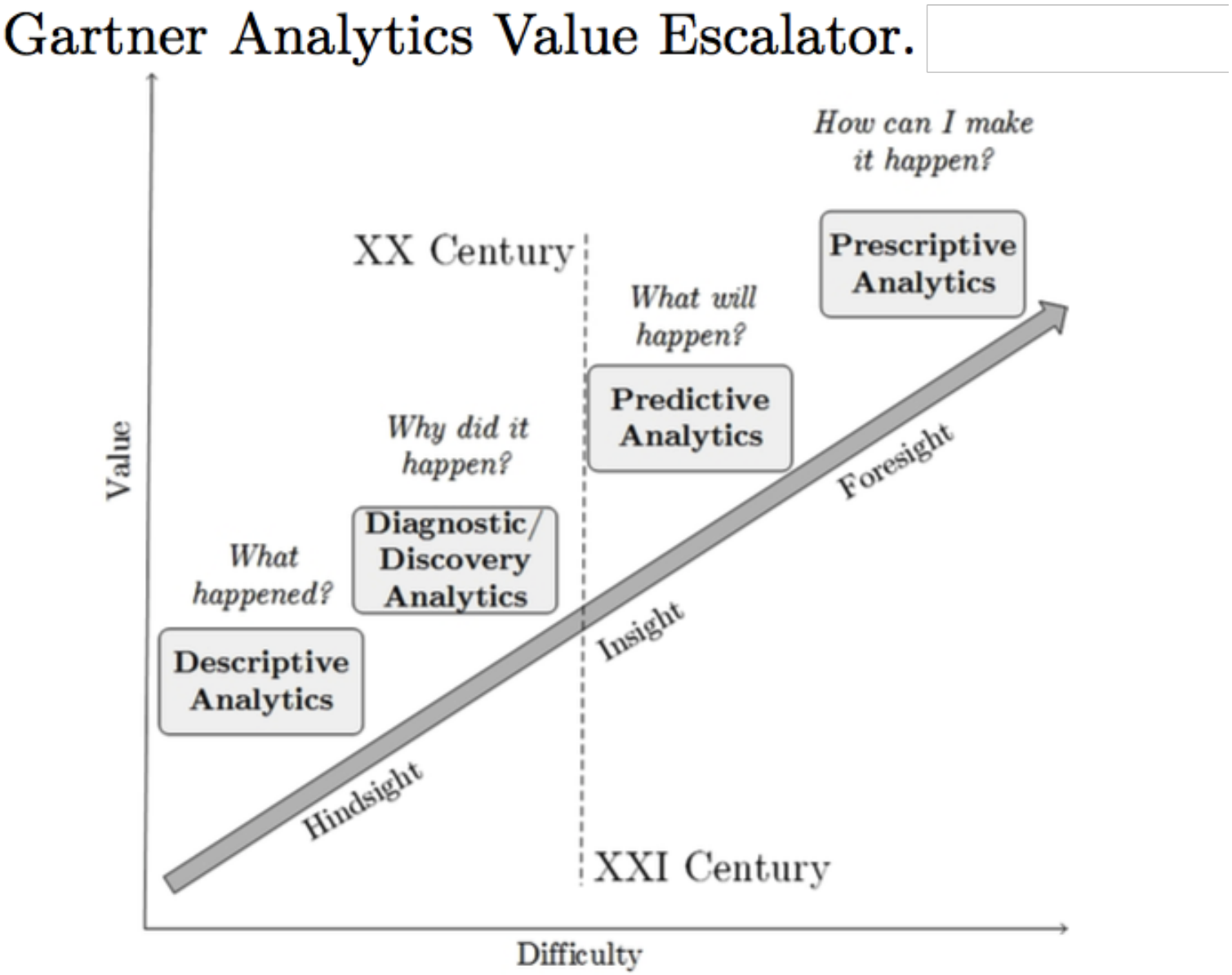


The most used techniques



Outline

The Evolution of Computational Data Analysis - II



Examples of What People are Doing with Data Mining

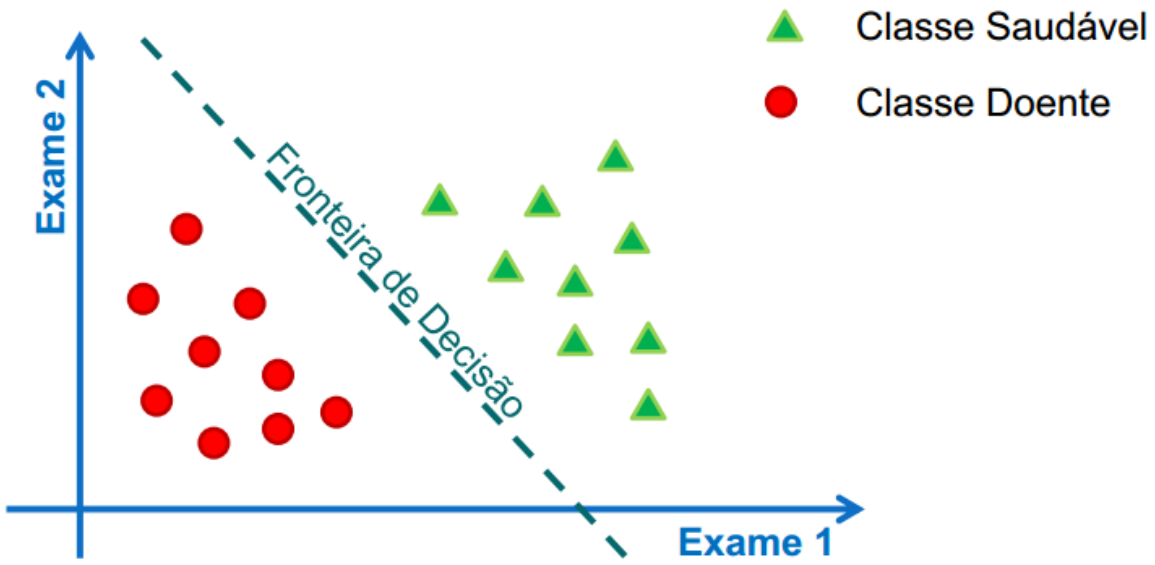
- Fraud/Non-Compliance Anomaly detection
  - Isolate the factors that lead to fraud, waste and abuse
  - Target auditing and investigative efforts more effectively
- Credit/Risk Scoring
- Intrusion detection
- Parts failure prediction
- Recruiting/Attracting customers
- Maximizing profitability (cross selling, identifying profitable customers)
- Service Delivery and Customer Retention
- Build profiles of customers likely to use which services
- Web Mining
- Social Networks analysis

Types of Problems

- Supervised
  - **Classifying** people or things into groups by recognizing patterns
  - **Regression** relationship between a dependent continuous variable and one or more independent variables.
    - **Forecasting** what may happen in the future
- Unsupervised
  - **Clustering** people or things into groups based on their attributes
  - **Associating** what events are likely to occur together
  - **Sequencing** what events are likely to lead to later events
  - **Link Analysis** who is connected with

Classification

Classifying people or things into groups by recognizing patterns

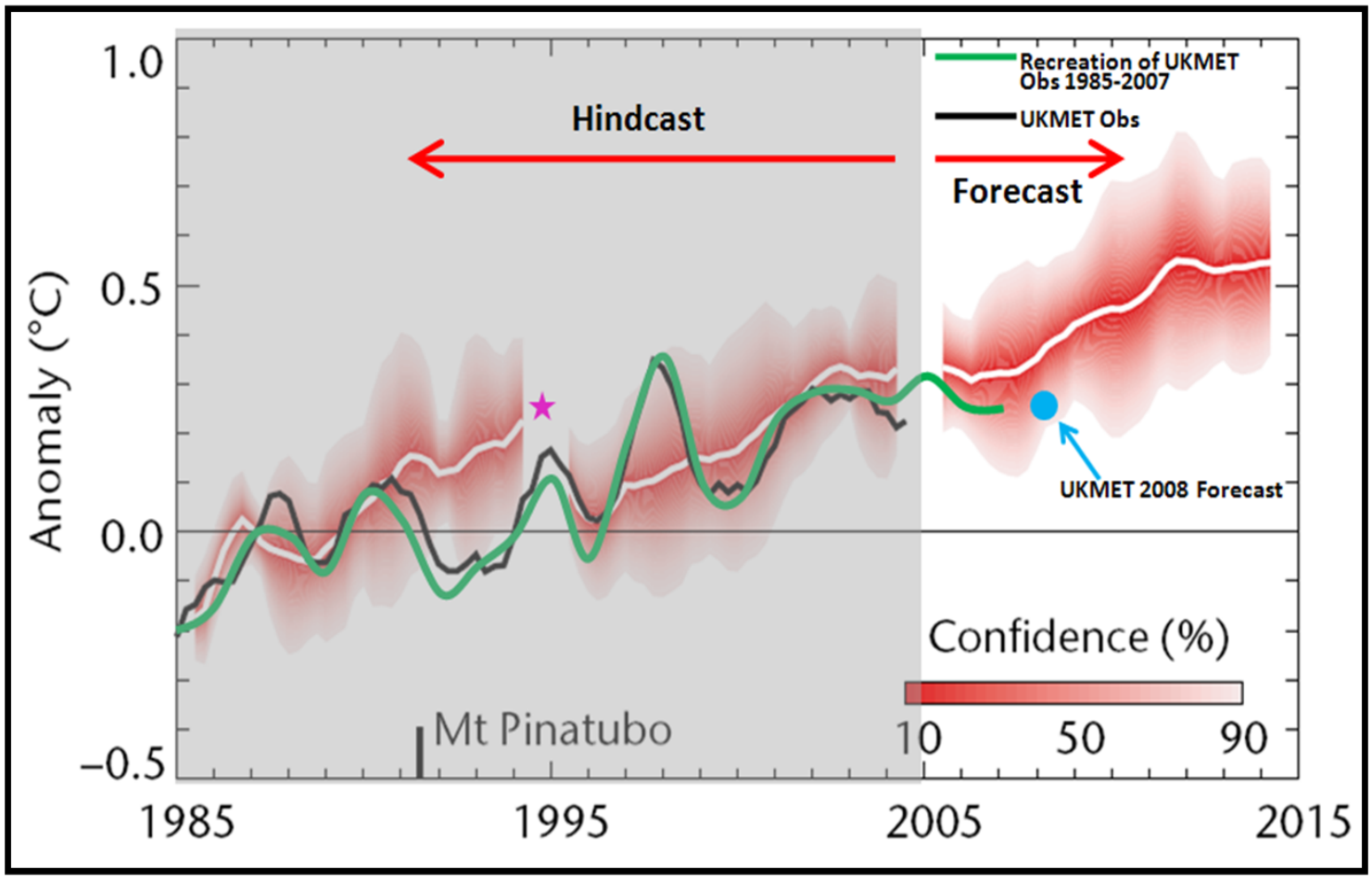


Regression: Function Approximation

relationship between a dependent continuous variable and one or more independent variables.

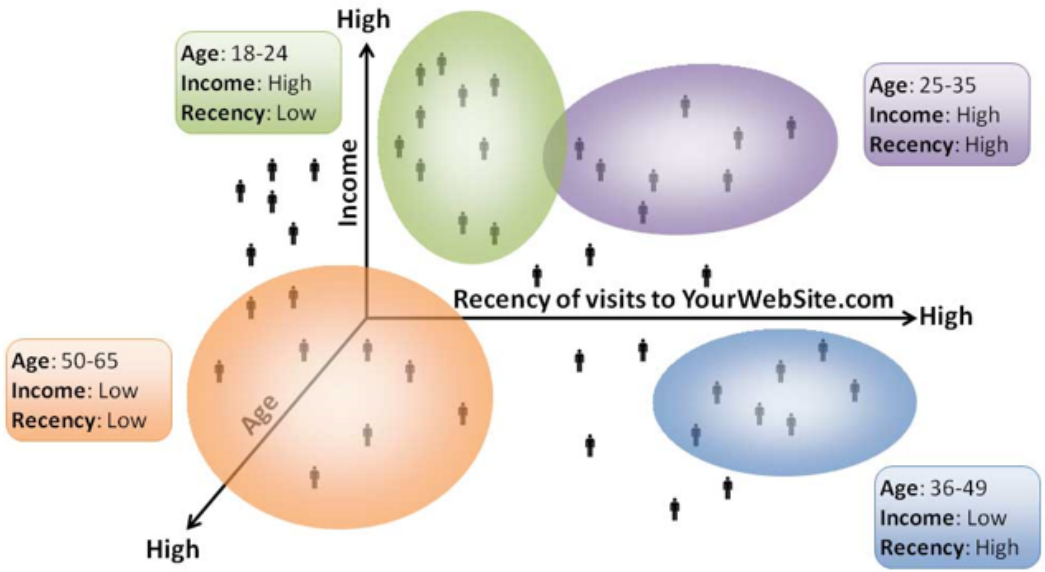
Forecast

Predicting the Future:



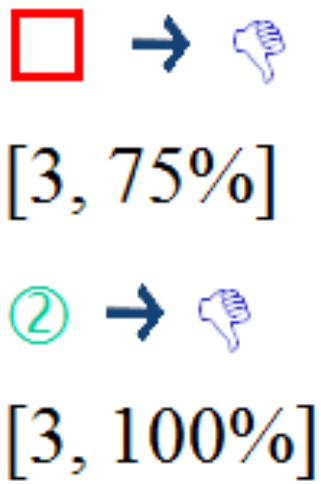
Cluster Analysis

Clustering people or things into groups based on their attributes



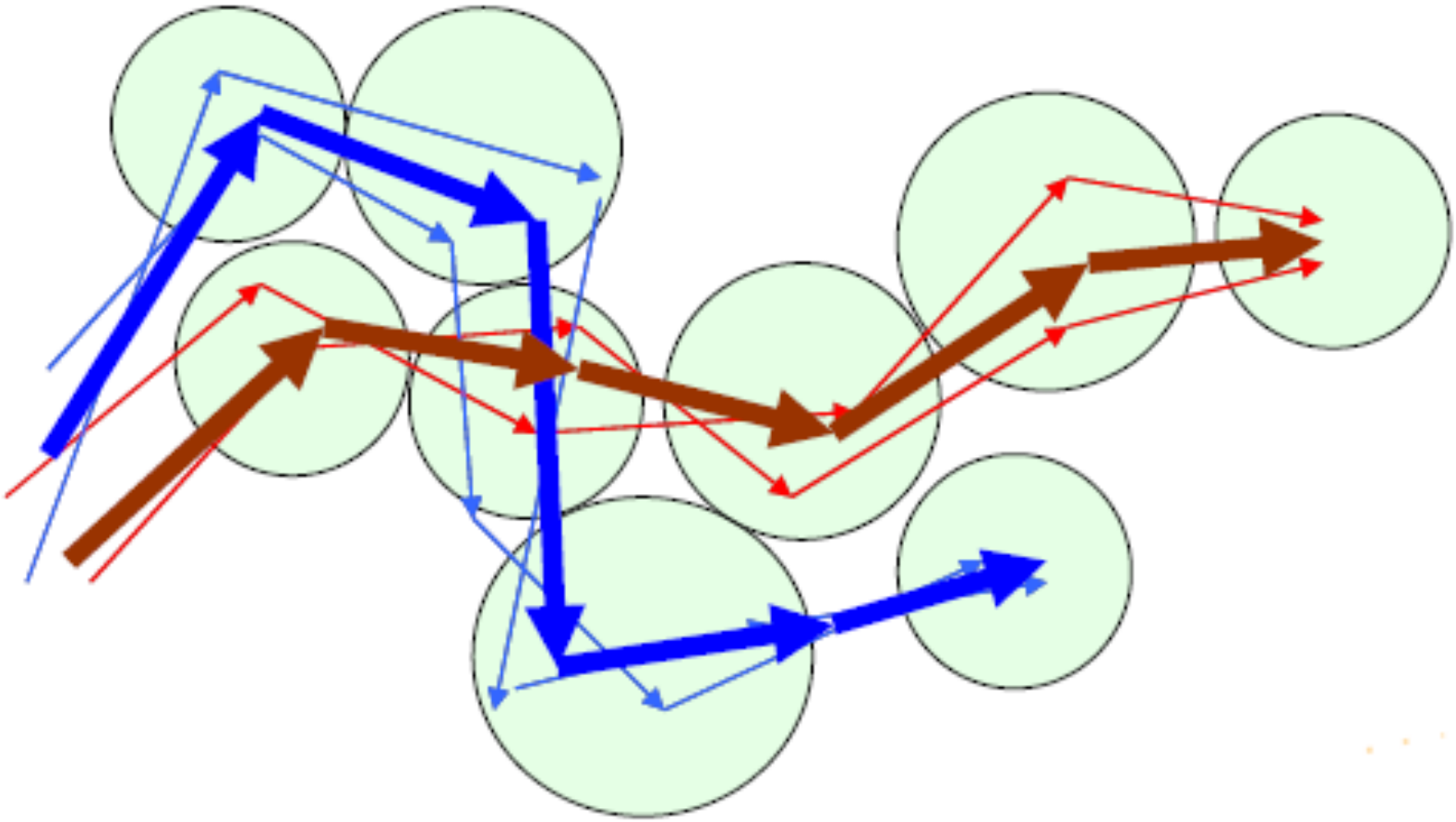
Association Analysis

What does it goes with?



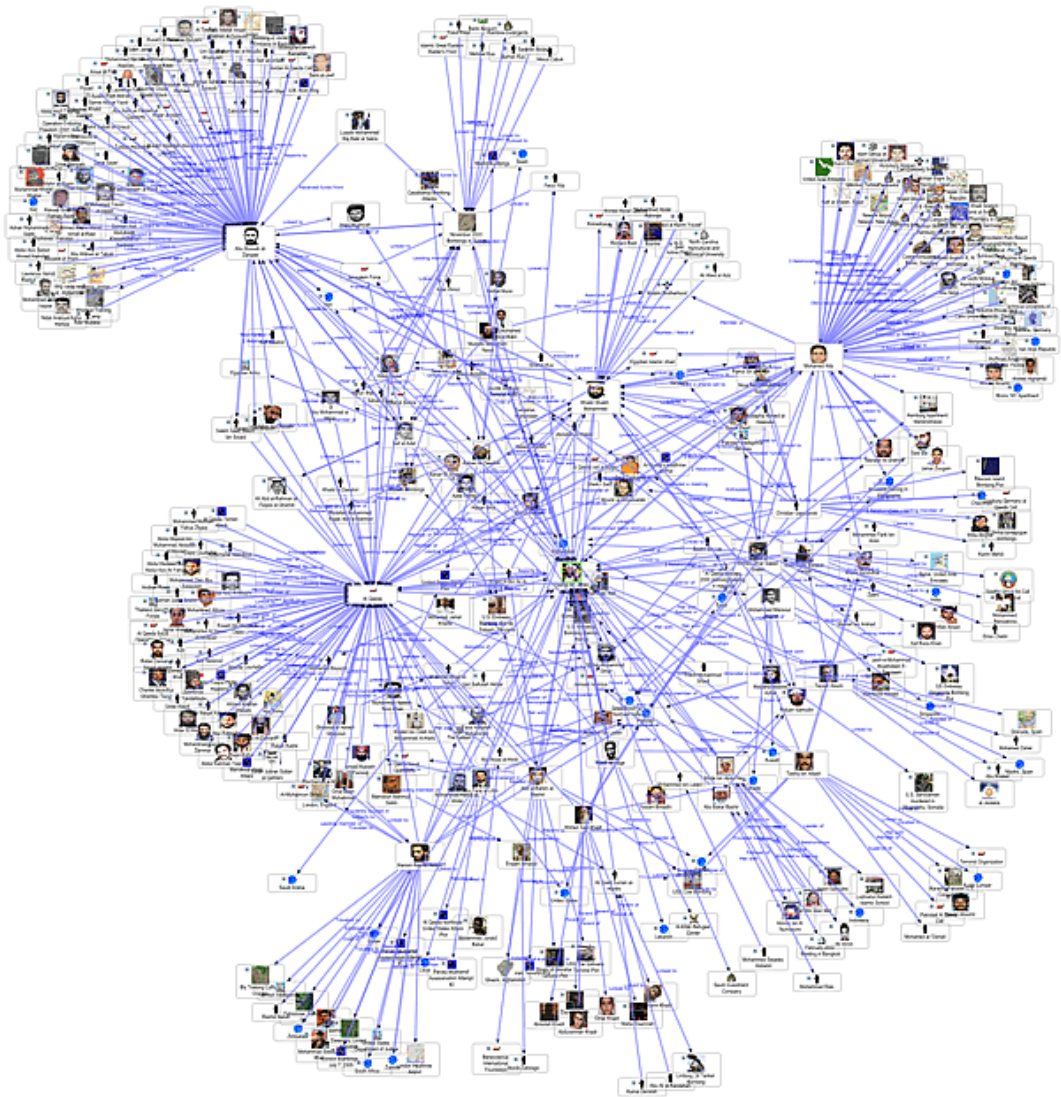
Sequence Analysis

Record trajectories from GPS devices. Which ones are most frequent?



Social Network Analysis

People communicate with people. Who are the pivots?



Outline

Recursos

- Video Lectures: http://videolectures.net/Top/Computer_Science/Machine_Learning
- KDD nuggets: http://www.kdnuggets.com
- Data Sets at UCI: http://archive.ics.uci.edu/ml/
- http://www.sciencemag.org/feature/data/compsci/machine_learning.dtl
- http://www.microstrategy.pt/data-mining/
- http://www.kdubiq.org/
- http://www.acm.org/sigs/sigkdd/explorations/
- http://home.earthlink.net/ dwaha/research/machine-learning.html