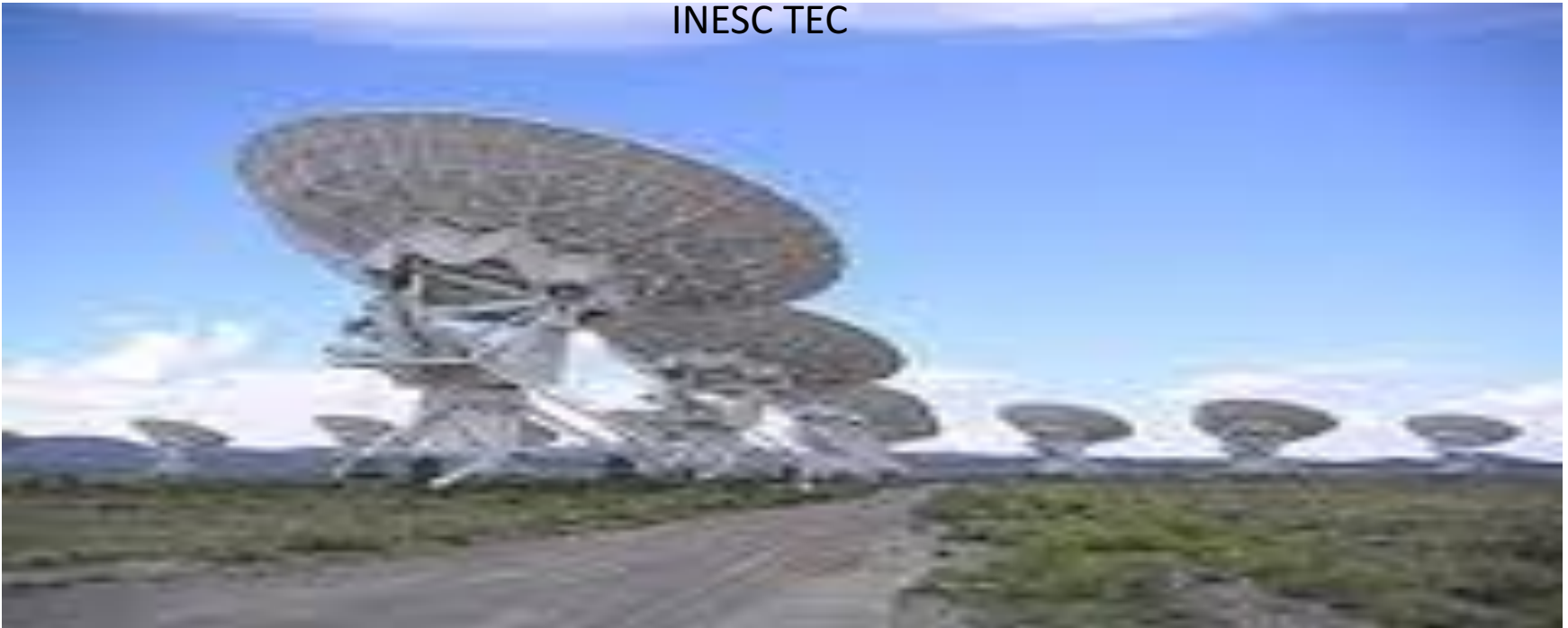


Data Sciences for the XXI Century

João Gama
jgama@fep.up.pt
INESC TEC



Tribute to Sir Ronald Fisher



Nowadays



Tsunami de dados

Tribuna Redes de Investigação
Pedro Velga

A observação do bóson de Higgs foi considerada a descoberta do ano de 2012 para a revista *Science*. Sabia que, para a descoberta do bóson de Higgs, o acelerador de partículas do CERN, quando em funcionamento, produz um enorme volume de dados que daria para encher cem mil CD a cada segundo? E que esses dados têm que ser distribuídos por investigadores localizados em todo o mundo, para poderem ser tratados em sofisticados computadores?

Este é um dos vários exemplos daquilo que, no mundo da ciência, se designa por *tsunami* de dados.

Com efeito, nos últimos anos e para todas as áreas científicas, começaram a ser produzidas enormes quantidades de dados. E espera-se que nos próximos anos esta tendência se venha a acentuar. Um exemplo do que virá a acontecer, dentro de uma década, é o que se relaciona com o Projecto SKA (Square Kilometer Array). O SKA vai ser o maior e mais sensível radiotelescópio do mundo, ficando instalado na região Sul de África e na Austrália. Quando entrar em funcionamento, irá produzir volumes de dados extremamente elevados, muito maiores do que os actualmente produzidos no CERN, e obrigará a que as redes de investigação e ensino, que foram sendo criadas em meados da década de 80 do séc

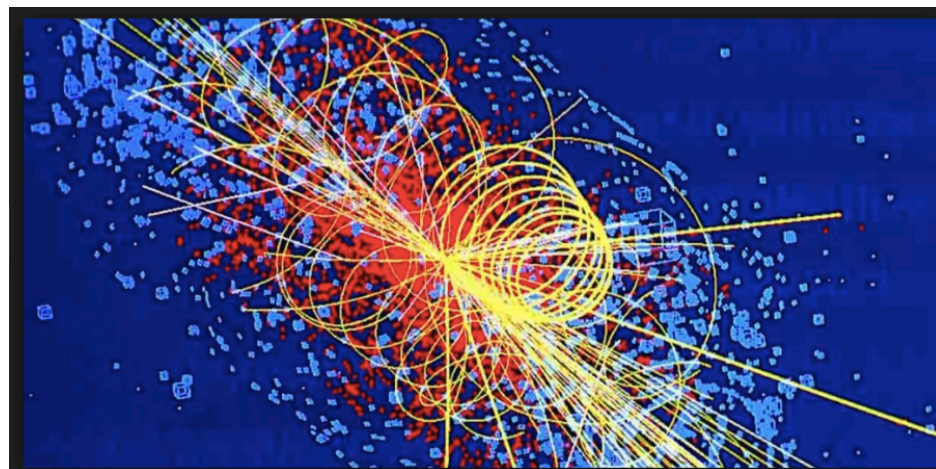
Em Portugal, a FCCN – Fundação para a Computação Científica Nacional – gere a rede RCTS, a rede que liga as nossas instituições de investigação e ensino superior à rede pan-europeia GÉANT e às outras redes mundiais congéneres. A maioria das nossas universidades e politécnicos já está ligada com capacidade de múltiplos acessos a 10 Gbit/seg. Um destes acessos é usado para o tráfego Internet normal. Os outros acessos são usados para projectos específicos do universo académico, como seja a ligação de centros de computação GRID para o tratamento dos dados recolhidos no acelerador do CERN.

A rede criada pela FCCN é uma poderosa rede de comunicações para a comunidade científica nacional, baseada em 1000km de fibra óptica e que chega a cerca de 83% dos utilizadores. Para os restantes são alugados circuitos aos operadores de telecomunicações. Esta fibra óptica está ligada à NREN espanhola nas fronteiras de



**Apesar da
nossa posição
geográfica,
estamos
perto de todo
o mundo
científico**

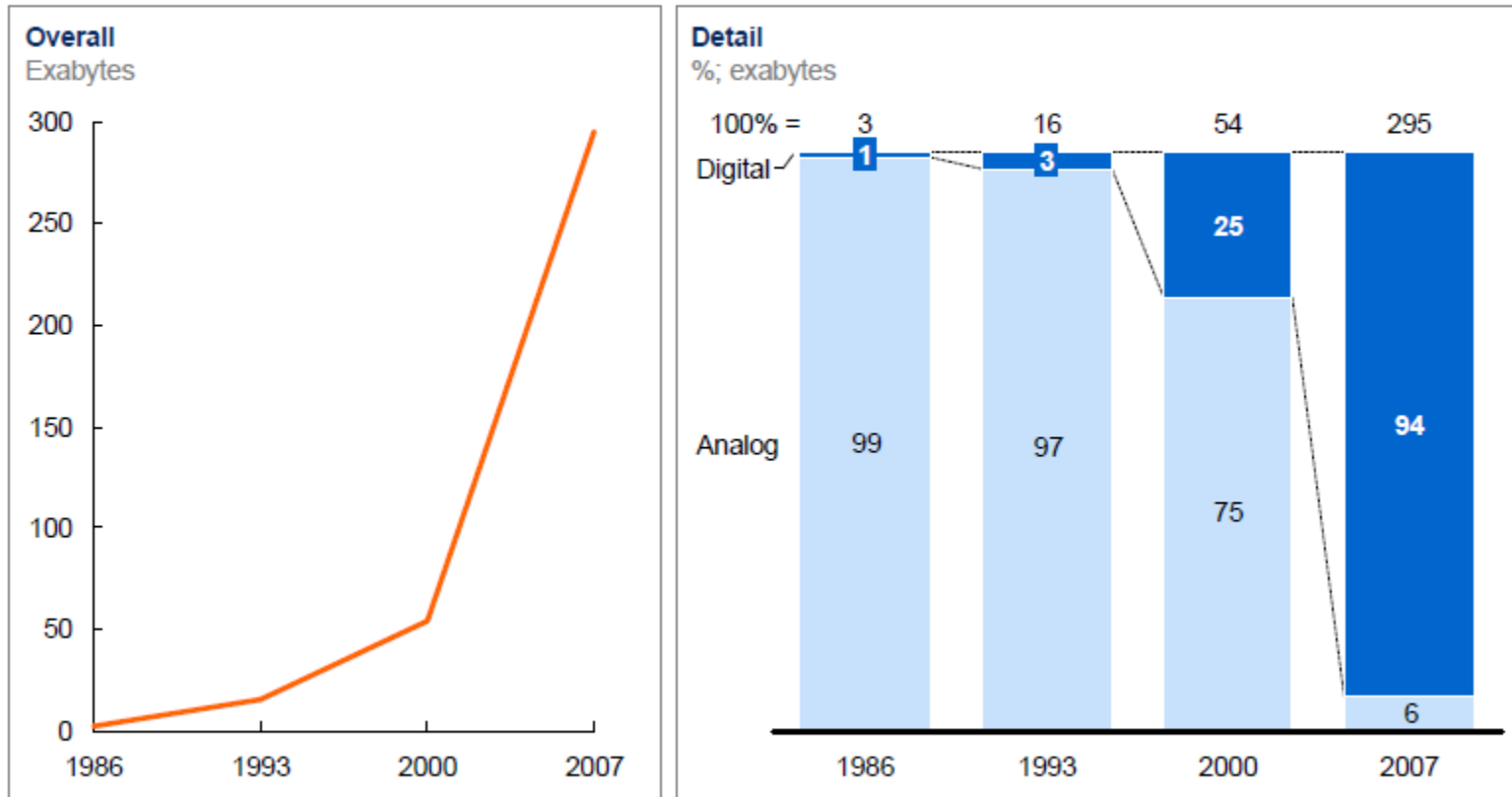
Valença e Elvas e, através desta, liga-nos à infraestrutura europeia e mundial de redes de investigação. Trata-se de uma infraestrutura científica electrónica - hoje em dia designada por *e-Infrastructure*, em terminologia inglesa - que é reconhecida como um dos sustentáculos



Growth of Digital Data

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

Global installed, optimally compressed, storage

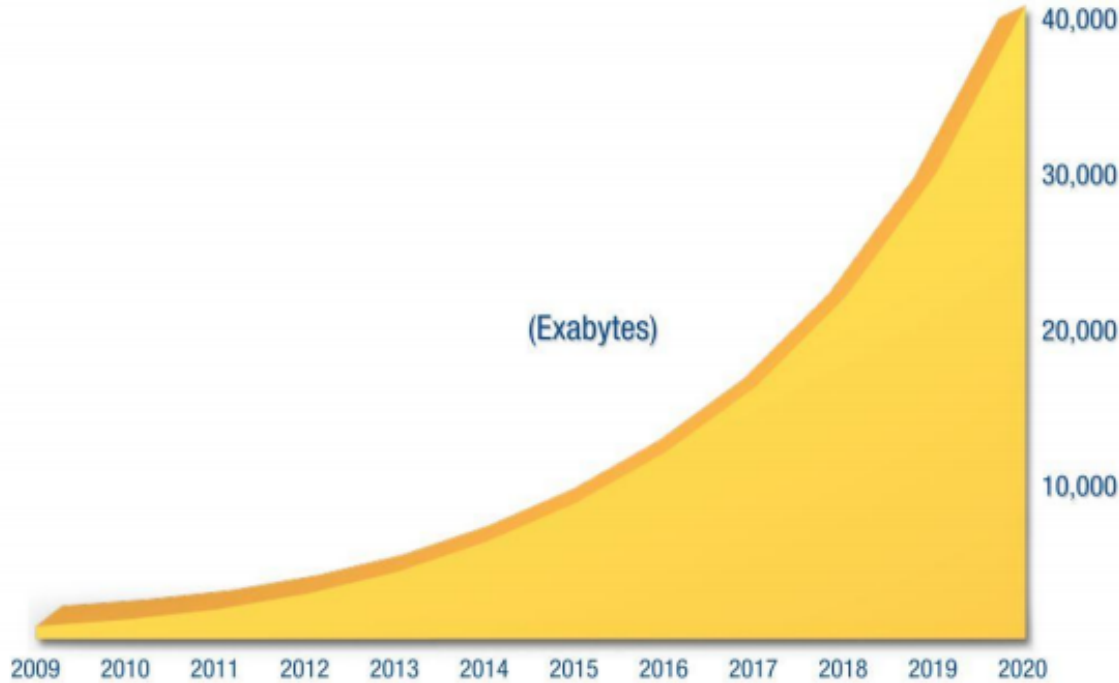


NOTE: Numbers may not sum due to rounding.

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

Growth of Digital Data

The Digital Universe: 50-fold Growth from the Beginning of 2010 to the End of 2020



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

Last few years

TOOLS

Problems



Understanding Data

A brief history of big data, the Noam Chomsky way



Text Size - +

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

By: Eric Rosenbaum | CNBC.com

Recommend 47

Twitter 104

g+1 6

LinkedIn 15

Share



ChinaFotoPress | Getty Images

Noam Chomsky

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

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.

A report from 2001



• 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

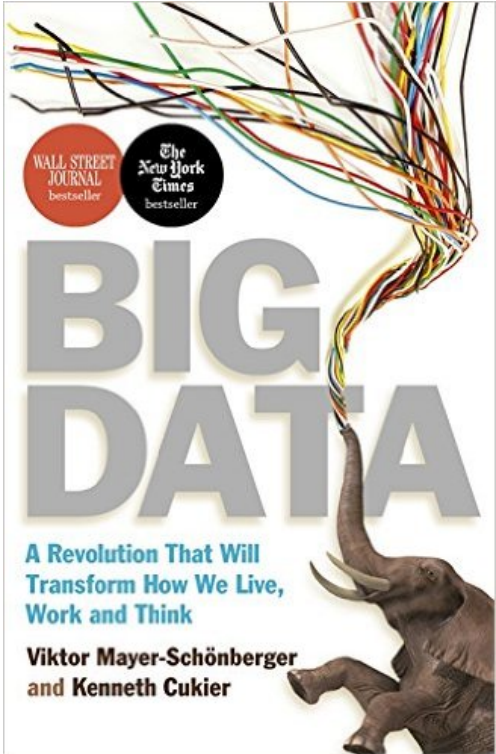
A report from 2011 and a book from 2014

McKinsey Global Institute

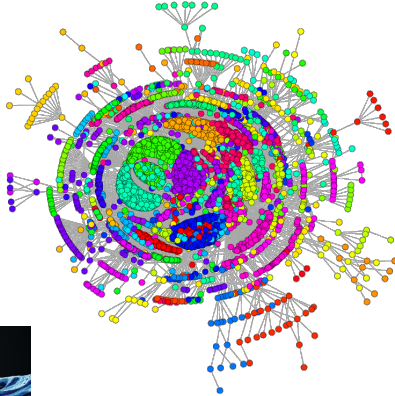


June 2011

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



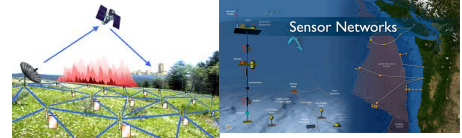
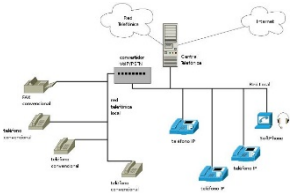
Who is Generating Data?



Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)



Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

The Model Has Changed...

- **The Model of Generating/Consuming Data has Changed**

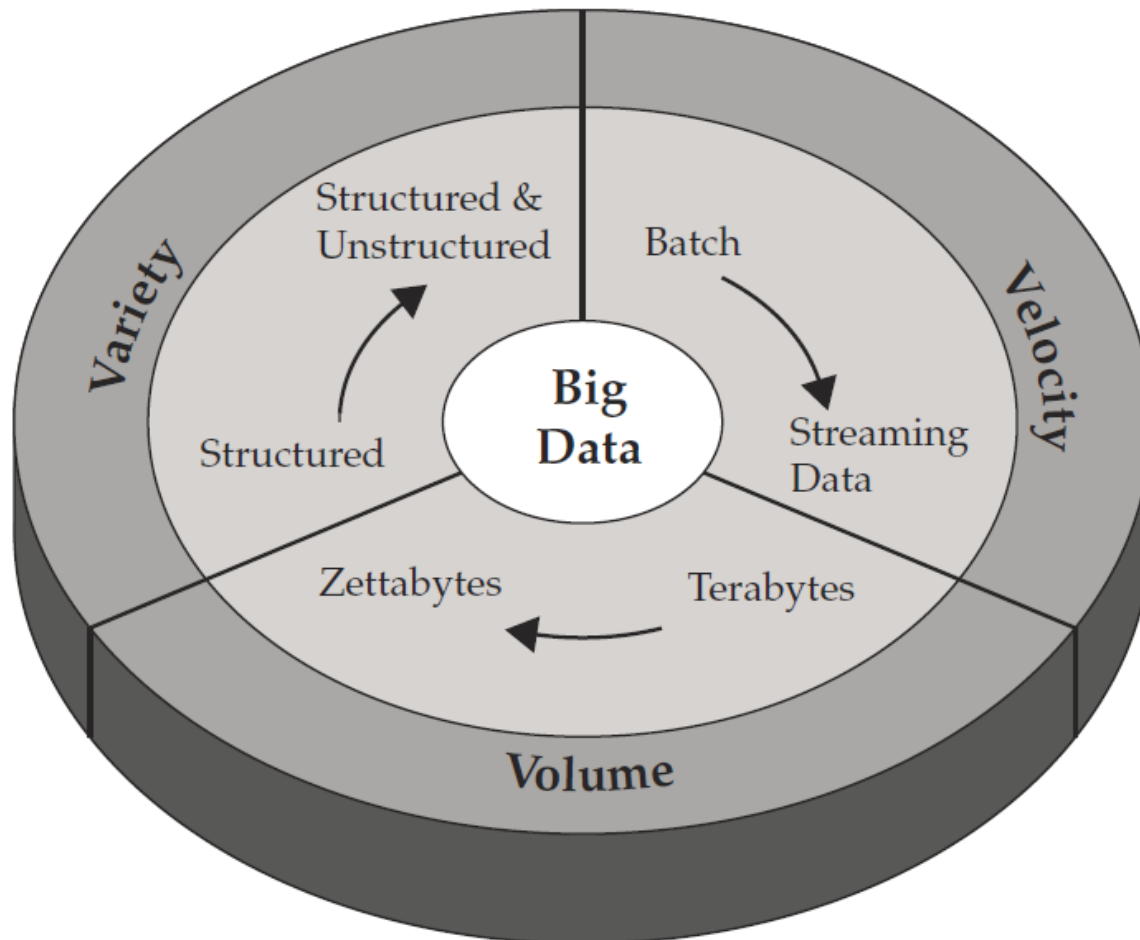
Old Model: Few companies are generating data, all others are consuming data



New Model: all of us are generating data, and all of us are consuming data



The 3V's

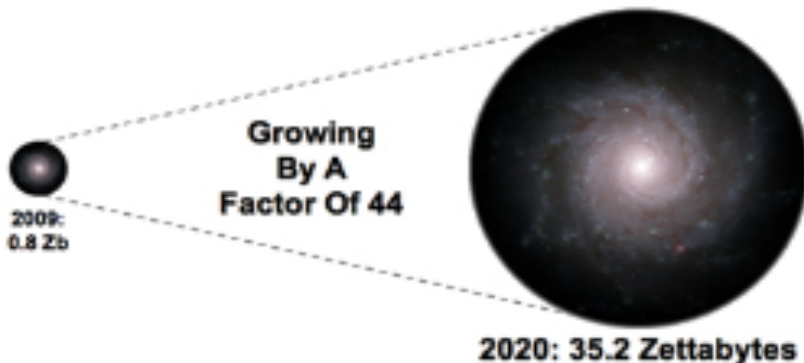


1-Scale (Volume)

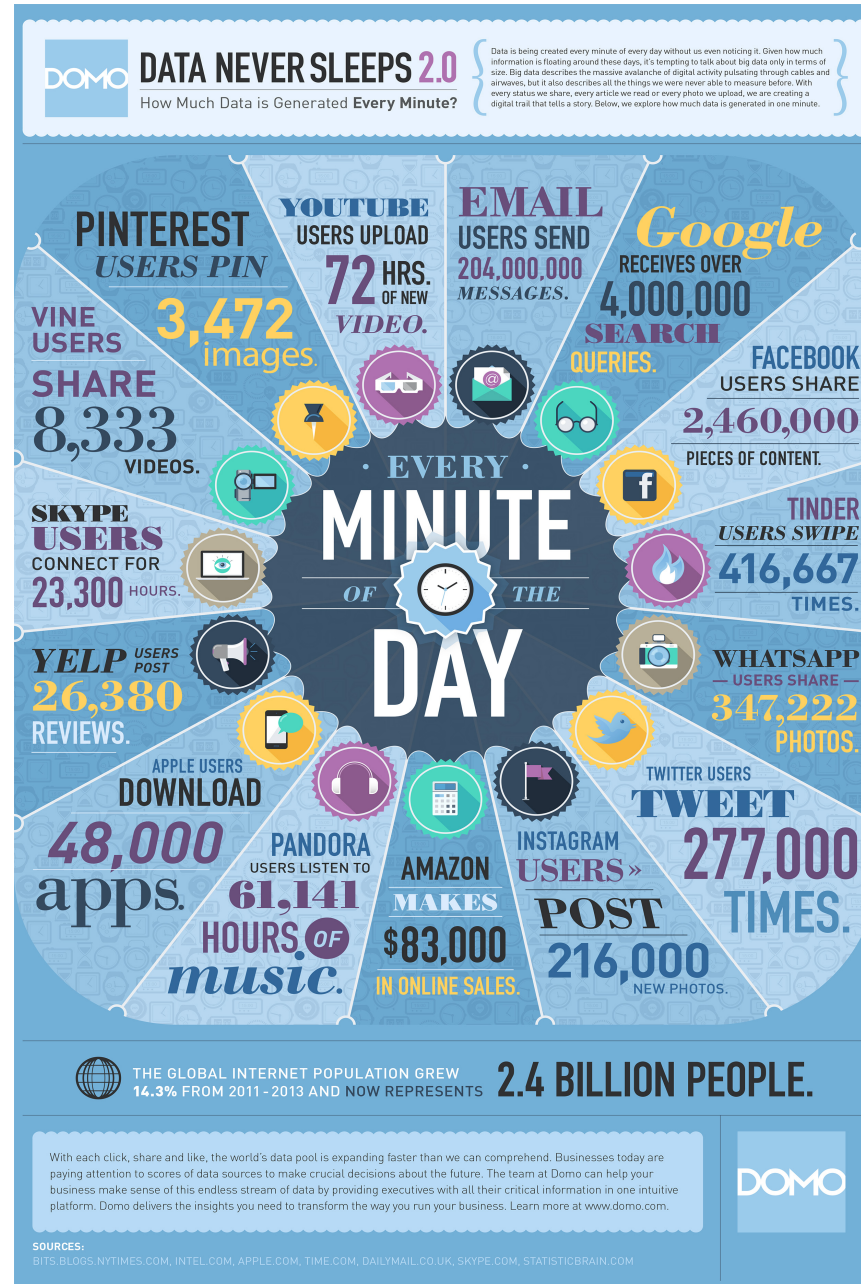
• Data Volume

- 44x increase from 2009-2020
- From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially

The Digital Universe 2009-2020

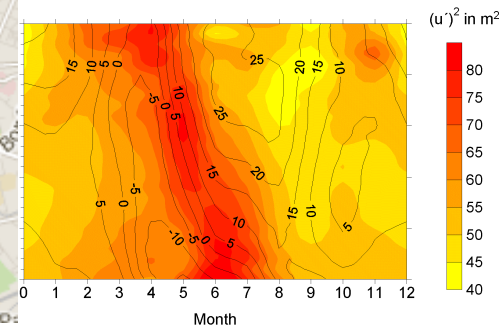
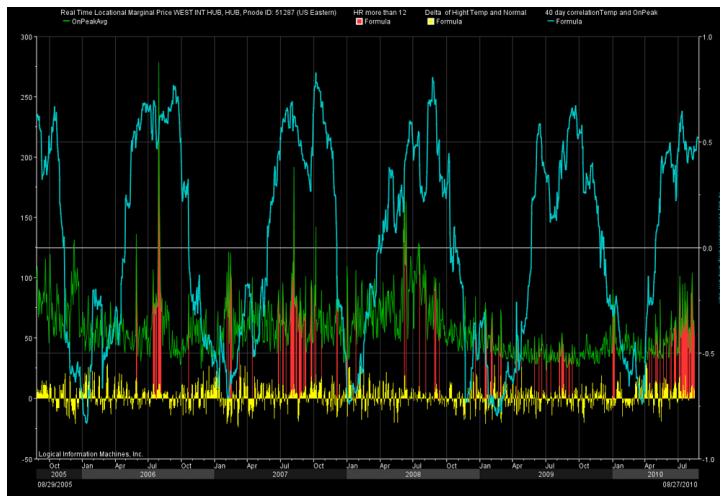
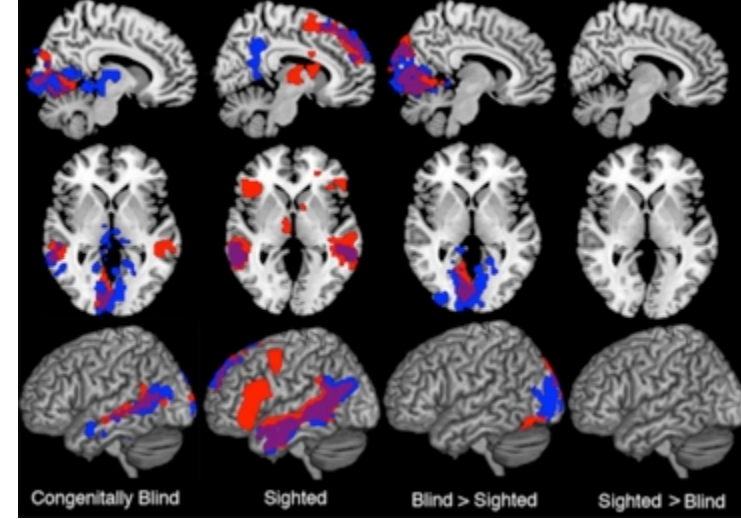


EMC



2-Complexity (Varity)

- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dimensional arrays, location, etc...
- Static data vs. streaming data
- A single application can be generating/collecting many types of data



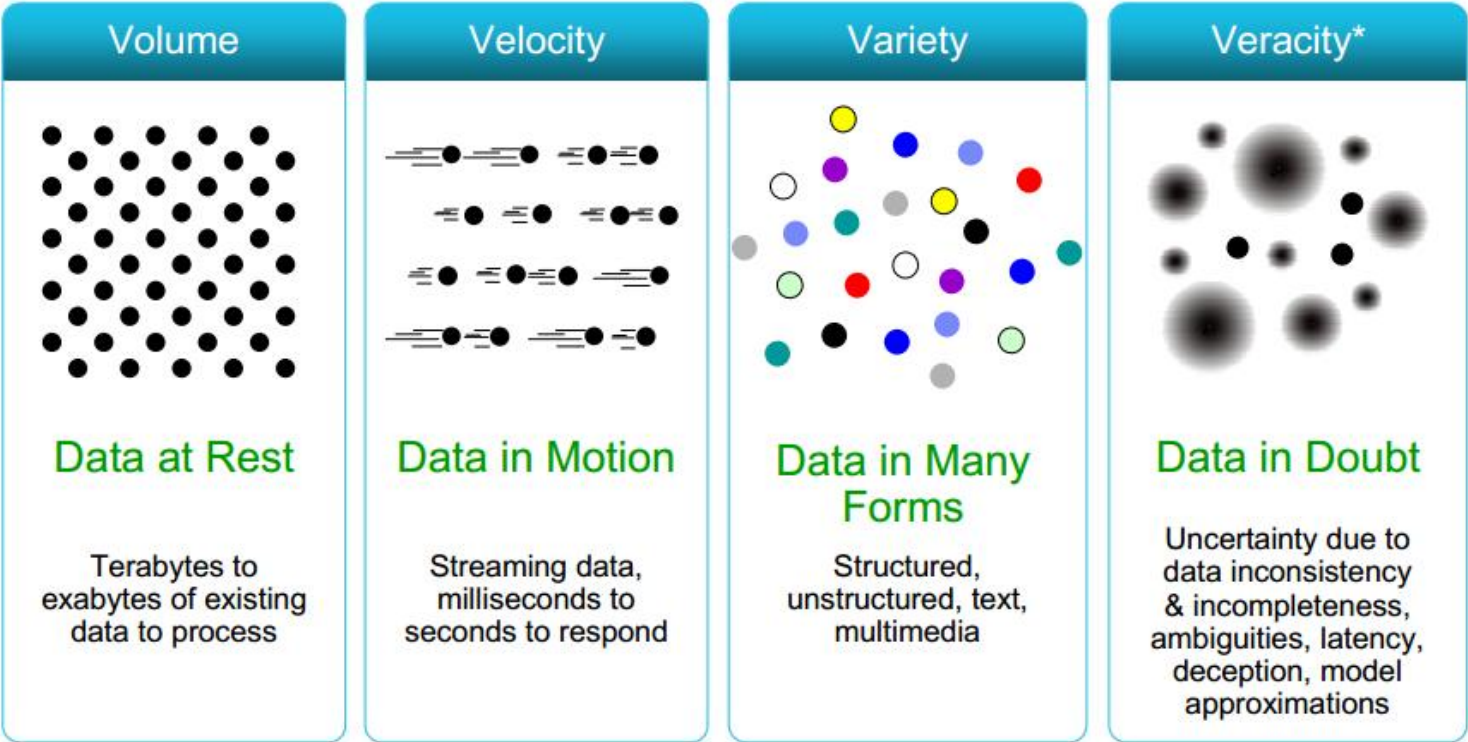
To extract knowledge → all these types of data need to be linked together

3-Speed (Velocity)

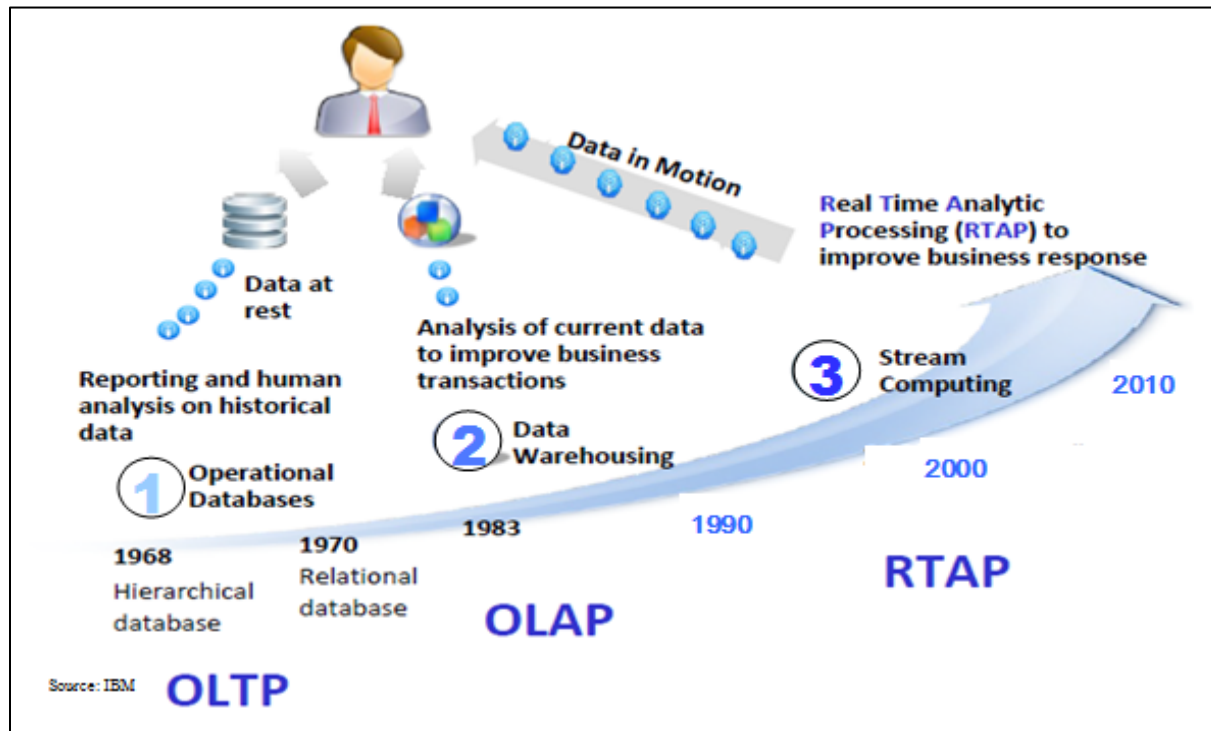
- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions → missing opportunities
- **Examples**
 - **E-Promotions:** Based on your current location, your purchase history, what you like → send promotions right now for a store next to you
 - **Healthcare monitoring:** sensors monitoring your activities and body → any abnormal measurements require immediate reaction
 - **Event Detection** from telecommunication networks -> bursts in calls activity



Some Make it 4V's

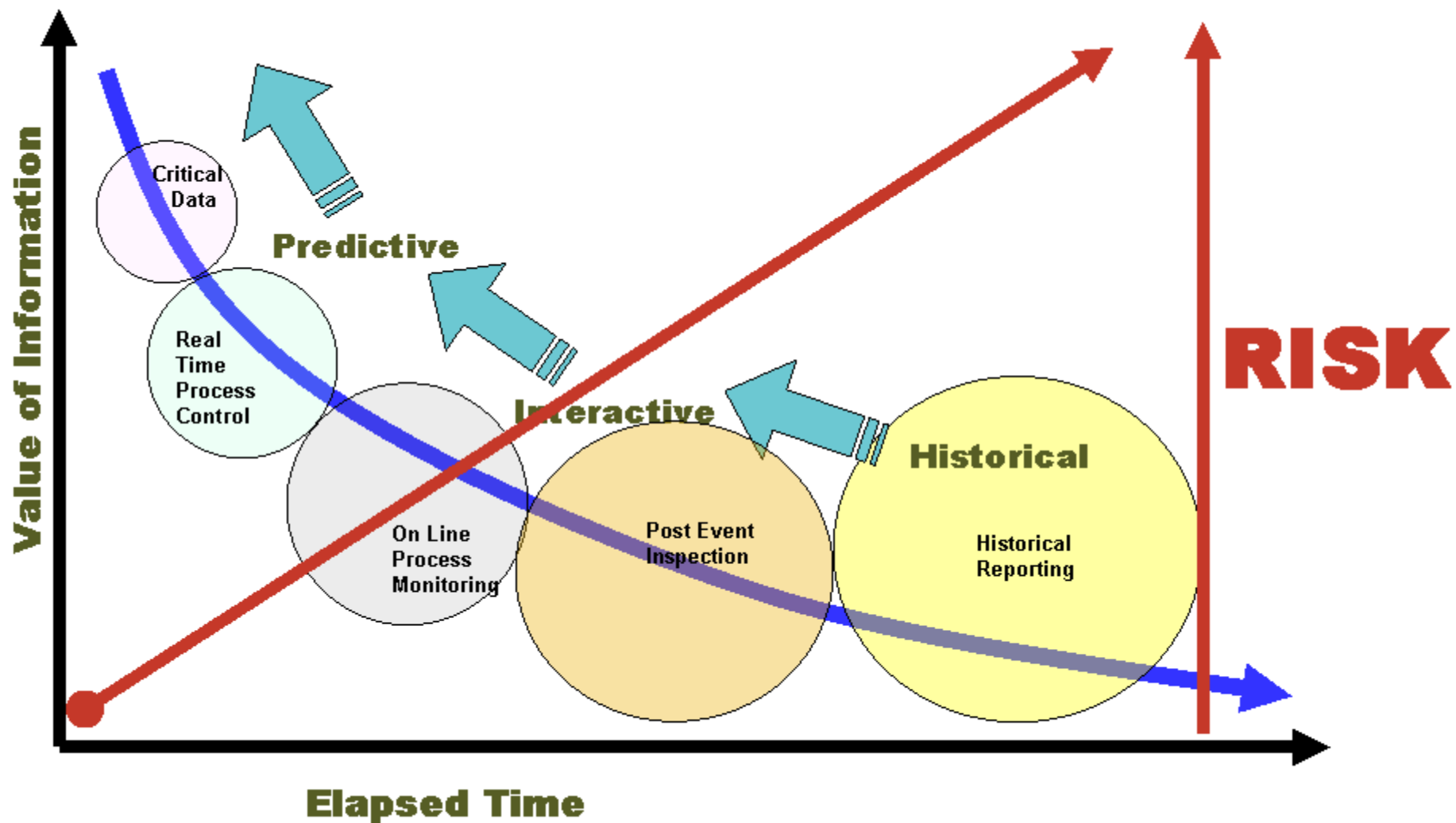


Data in Motion



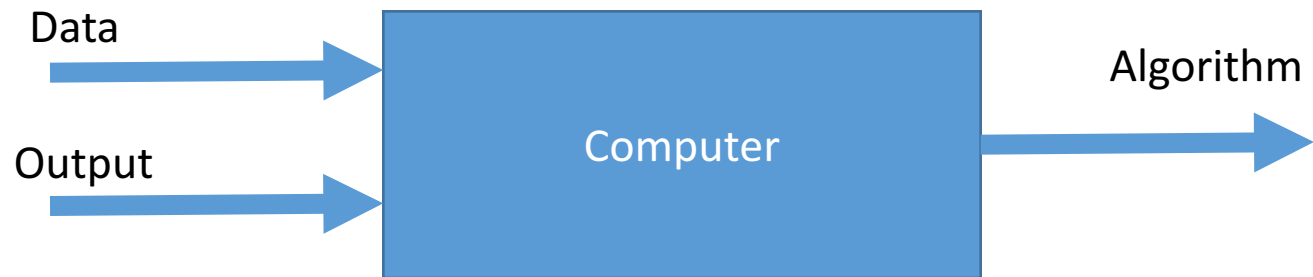
- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

Value of Information



Machine Learning: Understanding data

gives computers the ability to learn without being explicitly programmed



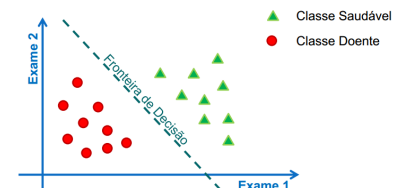
Supervised: Classification

Classifying people or things into groups by recognizing patterns

Person	A28202_ac	AB00014_at	AB00015_at	...	Class
Person1	1144.0	321.0	2567.2	...	normal
Person2	105.2	586.1	759.2	...	cancer
Person3	586.3	559.0	3210.2	...	normal
Person4	42.8	692.0	812.2	...	cancer

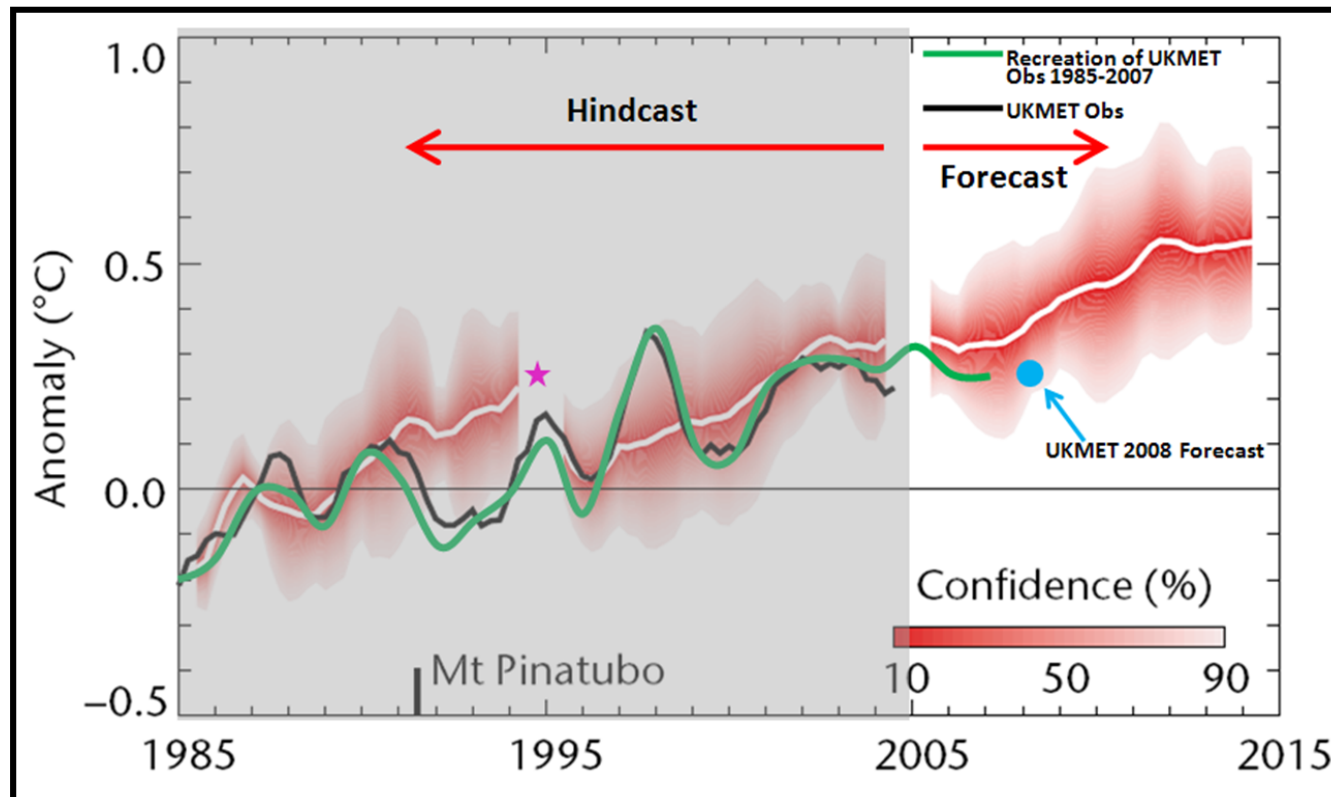
Learning Problems:

- Find a function:
$$\text{Class} = f(\text{A28202_ac}, \text{AB00014_at}, \text{AB00015_at}, \dots)$$
- Given the expression level of genes of a Person, predict if he has cancer or not.



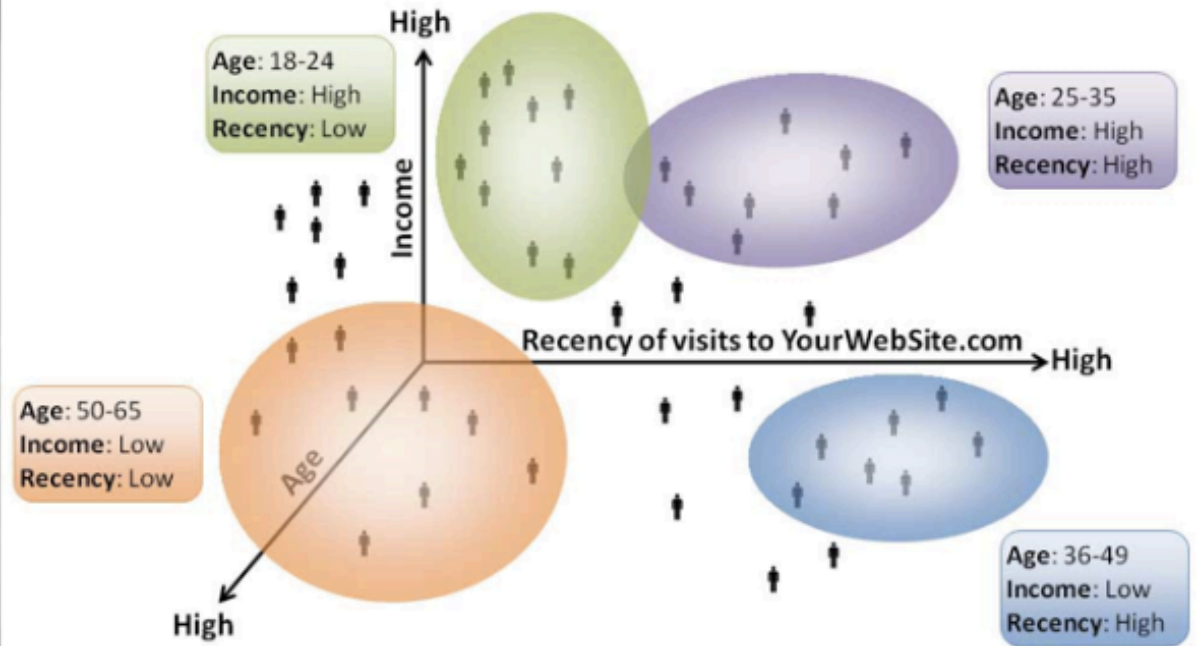
Regression: Function Approximation

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



Cluster Analysis

- Clustering people or things into groups based on their attributes



Association Analysis

What does it goes with?



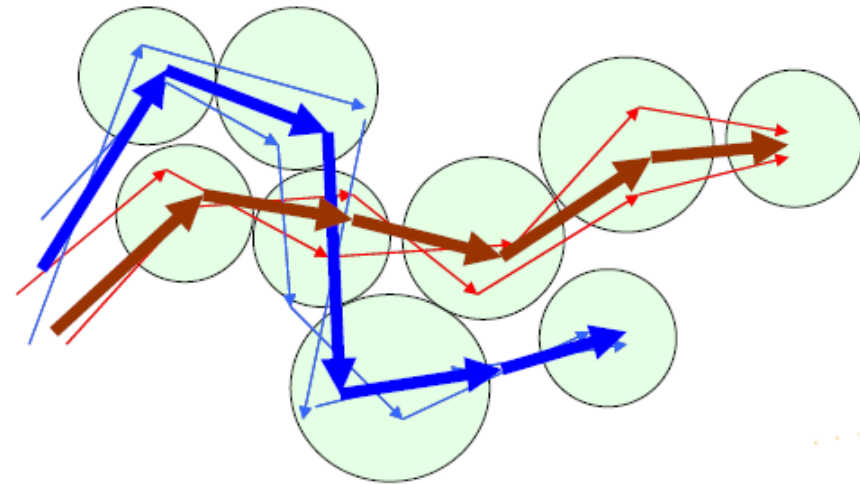
[3, 75%]



[3, 100%]

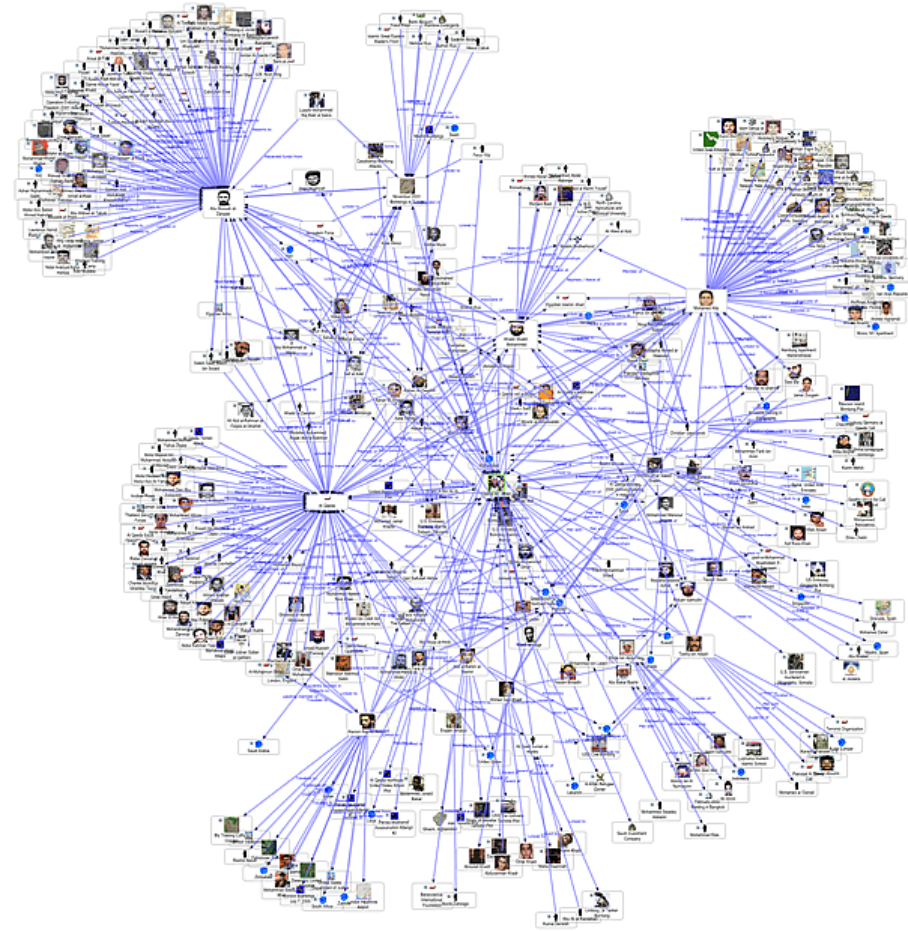
Sequence Analysis

- Trajectories from GPS devices. Where are we going?



Social Network Analysis

- People communicate with people. Who are the pivots?



Applications

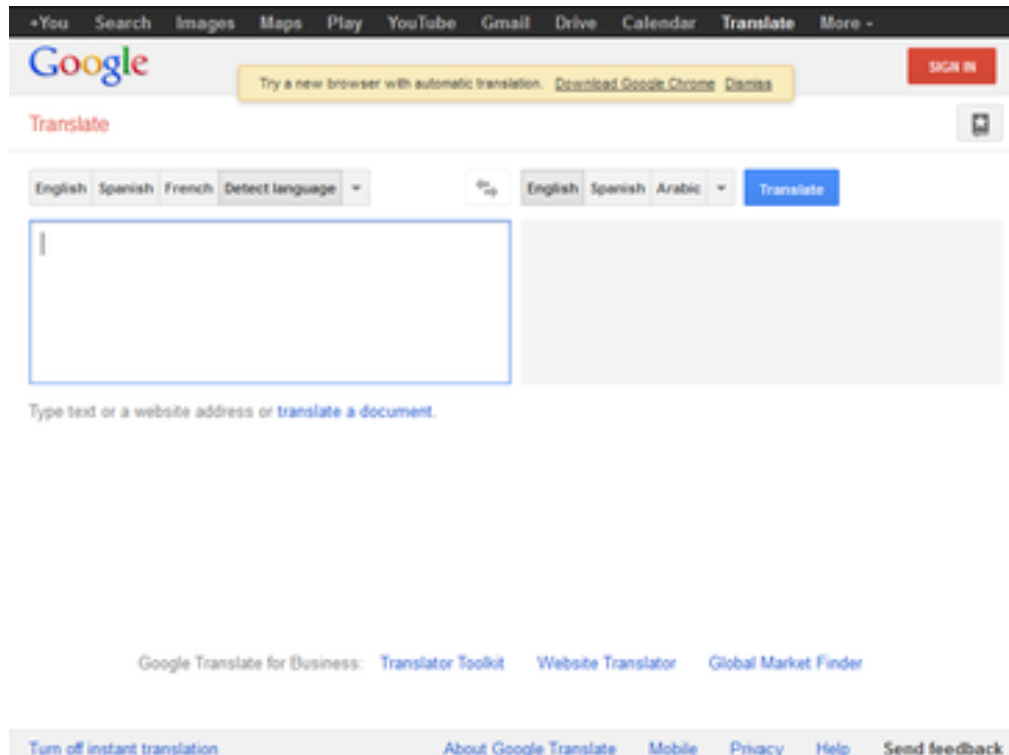
See the differences



- Machine learning used to take place behind the scenes:
 - Amazon mined your clicks and purchases for recommendations,
 - Google mined your searches for ad placement,
 - Facebook mined your social network to choose which posts to show you.
- Nowadays, machine learning is on the front pages of newspapers, and the subject of heated debate:
 - Learning algorithms drive cars,
 - Translate most popular human languages, speech to text
 - Won Kasparov, won Lee Sedol, won at Jeopardy!


Automatic Translation

- Google's approach:
 - Looking **at masses of data** in parallel produced far better translations than the old algorithm-driven method.
 - the English and French translations of various public-domain texts
 - EC legislation translated to several European languages
 - **The bigger the corpus, or body of parallel texts, the better the results.**



Text Mining

Sentiment140

 Tweet 374

 Like 167

 +1 83

English ▾

Search

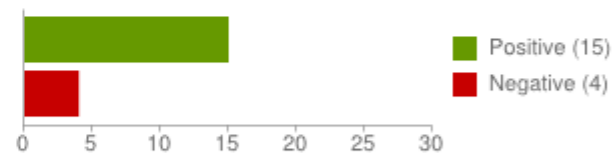
[Save this search](#)

Sentiment analysis for Toyota

Sentiment by Percent



Sentiment by Count



Tweets about: Toyota

[Roshan_Gunner](#): RT @FAC7S: In Malaysia the ad of **Toyota** Altis featuring Brad Pitt inferior.

Google Flu Trend

- *Five years ago, a team of researchers from [Google](#) announced a remarkable achievement in one of the world's top scientific journals, [Nature](#).*
- *Without needing the results of a single medical check-up, they were nevertheless able to track the spread of influenza across the US. What's more, they could do it [more quickly](#) than the Centers for Disease Control and Prevention (CDC).*
- *Google's tracking [had only a day's delay](#), compared with the week or more it took for the CDC to assemble a picture based on reports from doctors' surgeries. Google was faster because it was tracking the outbreak by finding [a correlation between what people searched for online and whether they had flu symptoms](#).*
 - Big data: are we making a big mistake?
 - Tim Harford, FT Magazine, March 2014

Telecommunications

Mobile phones provide maps of the Portuguese population density near real-time

Publico 24/11/2014

Onde estão os portugueses quando trabalham e vão de férias?

Distribuição da população
(pessoas por km²)

Férias em Julho e Agosto



Período de trabalho

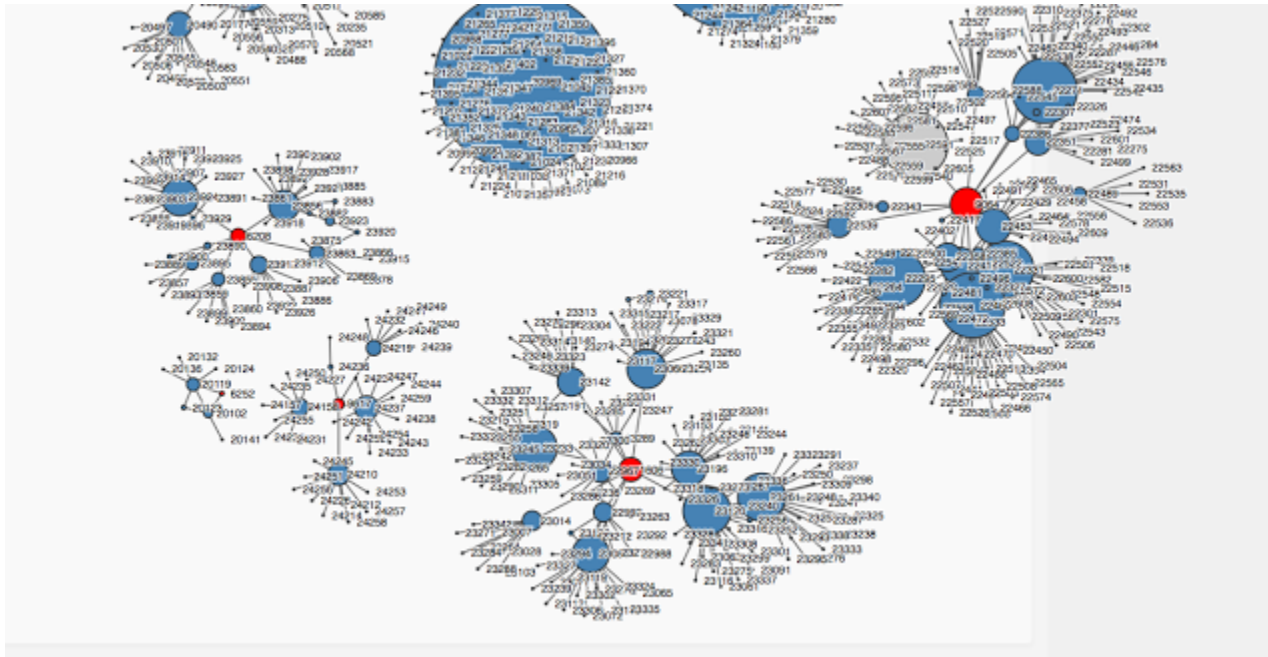


Large Scale Social-Network Analysis

- The streams of Call Detail Records (CDR's) generating from these devices provide a powerful abstraction of social interactions between individuals, representing social structures.
- A case study
 - Call Detail Records (CDR) log files:
 - 6 million of users.
 - 10 million calls per day (on average).
 - CDRs implicitly defines a network,
 - nodes are clients.
 - edges corresponds to a call between two clients.
 - The stream of phone calls defines a network stream.
 - Goals:
 - Identify communities
 - Track the evolution of communities
 - Business Goals:
 - Fraud detection
 - Prevent churn
 - Tarifs

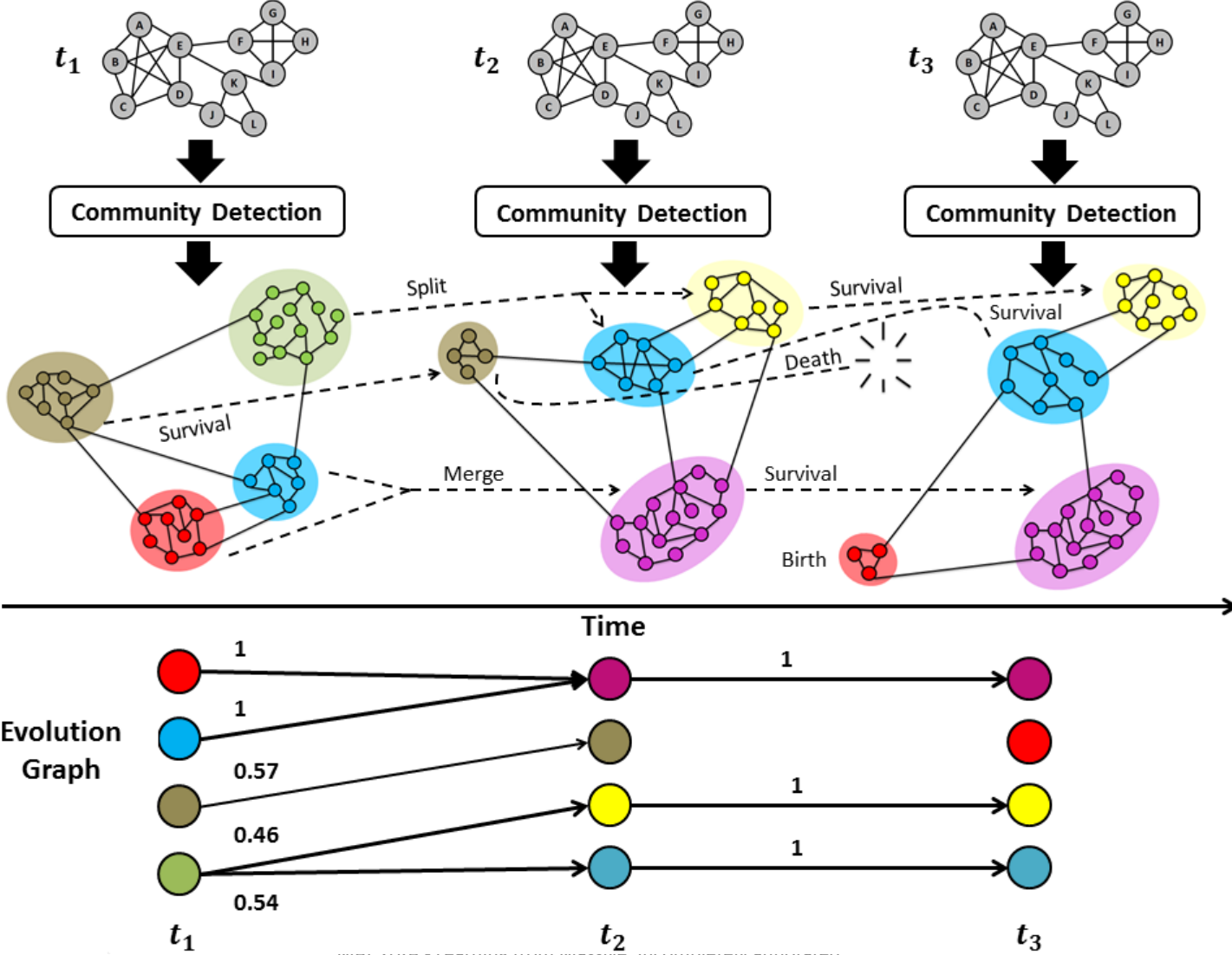
Large Scale Social-Network Analysis

- Mobile phones are powerful tools to **connect people**.
- The streams of Call Detail Records (CDR's) generating from these devices provide a powerful abstraction of **social interactions** between individuals, representing **social structures**.



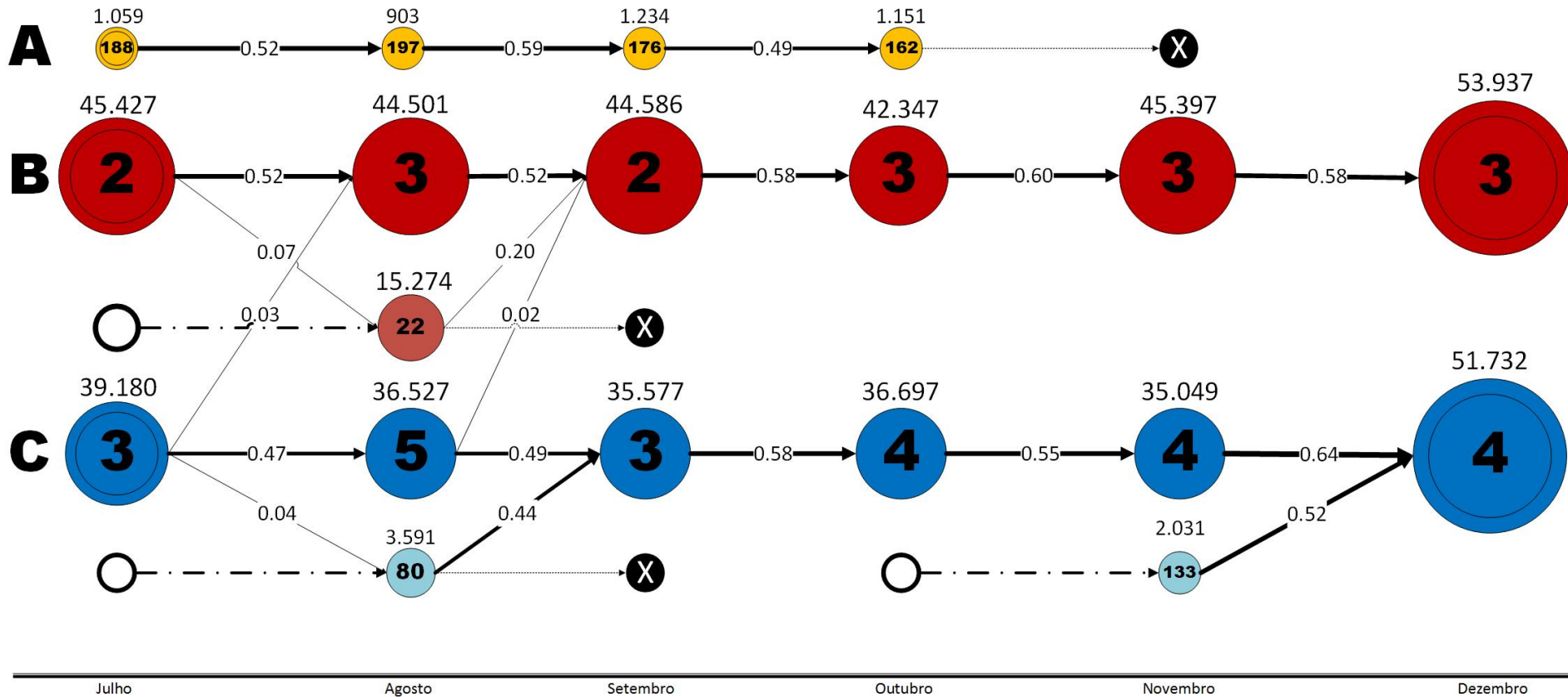
- 1) **unsocial** - rarely makes phone calls..
- 2) **small network** - few calls to neighbors
- 3) **nagging** - often calls to call centers
- 4) **social** - connected to a lot of friends which are interconnected together

Dynamic Community Mining

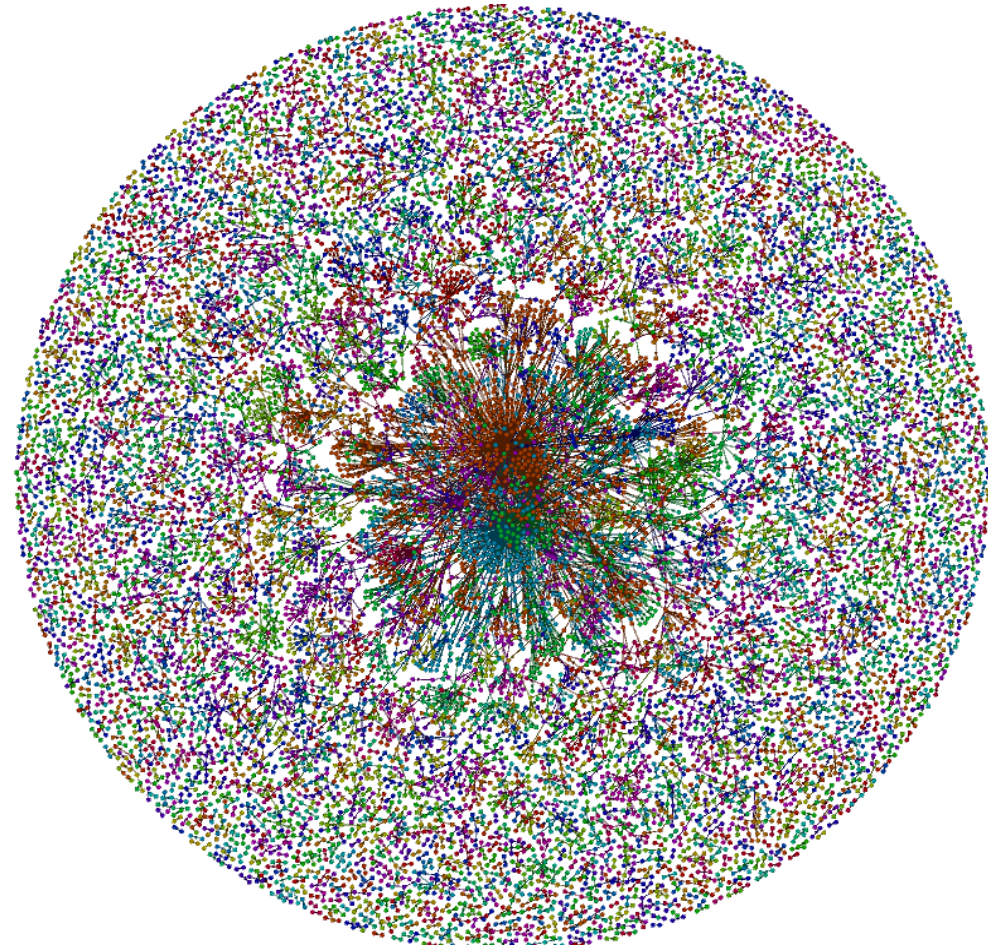
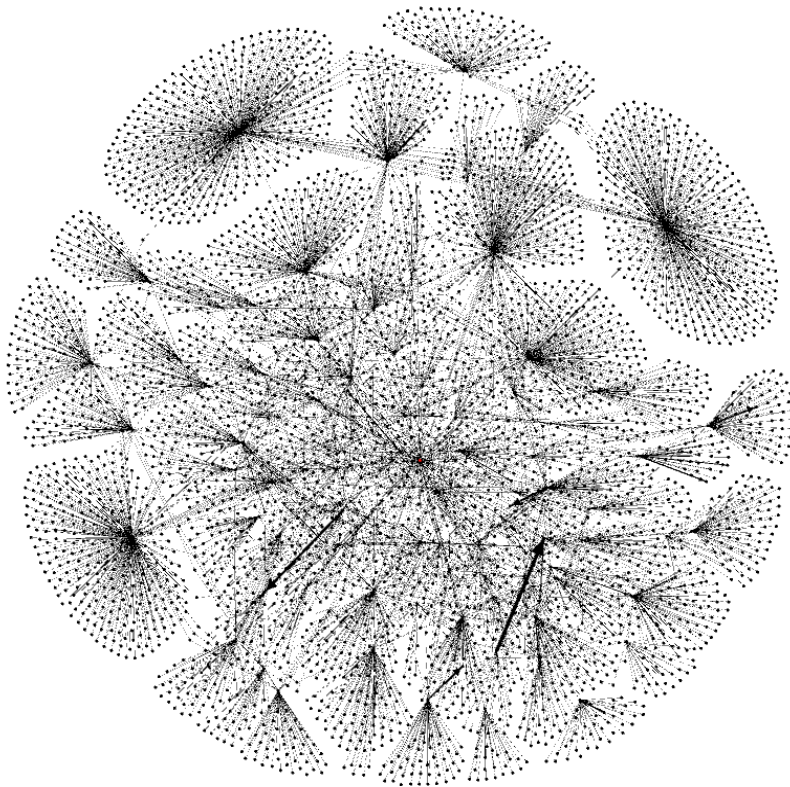


Dynamic Community Mining and Tracking

Ciclo de vida das comunidades detetadas



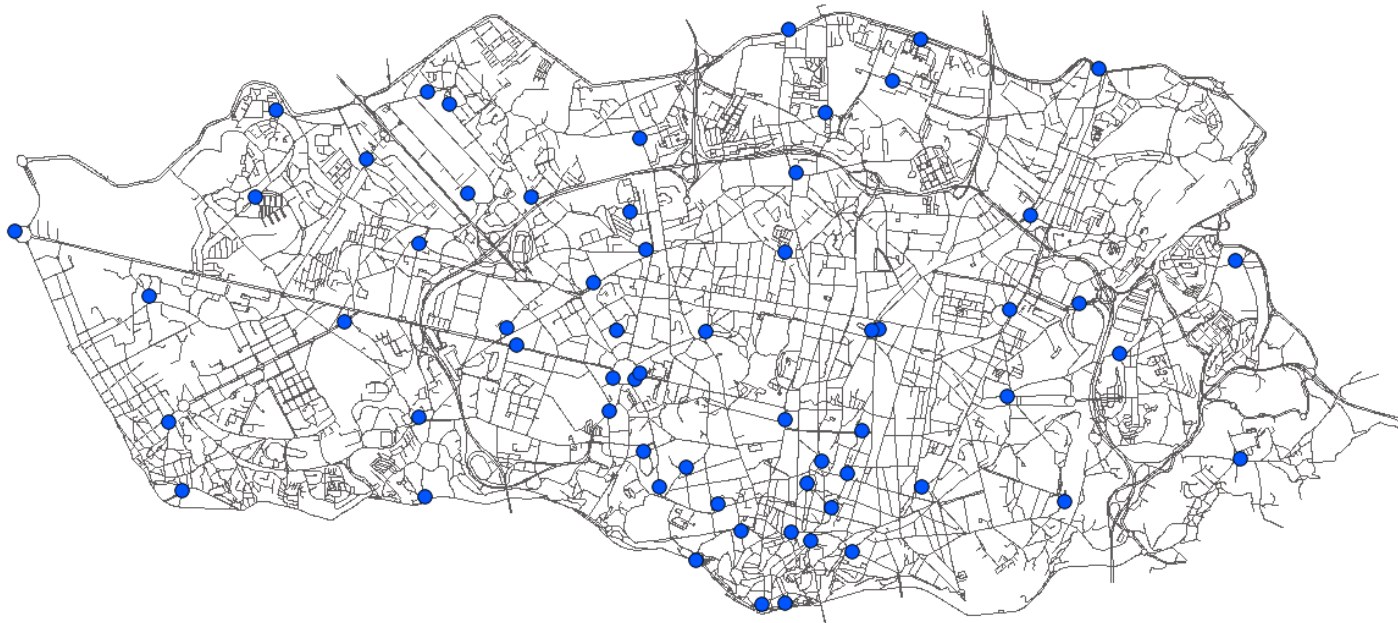
High-speed streaming Networks



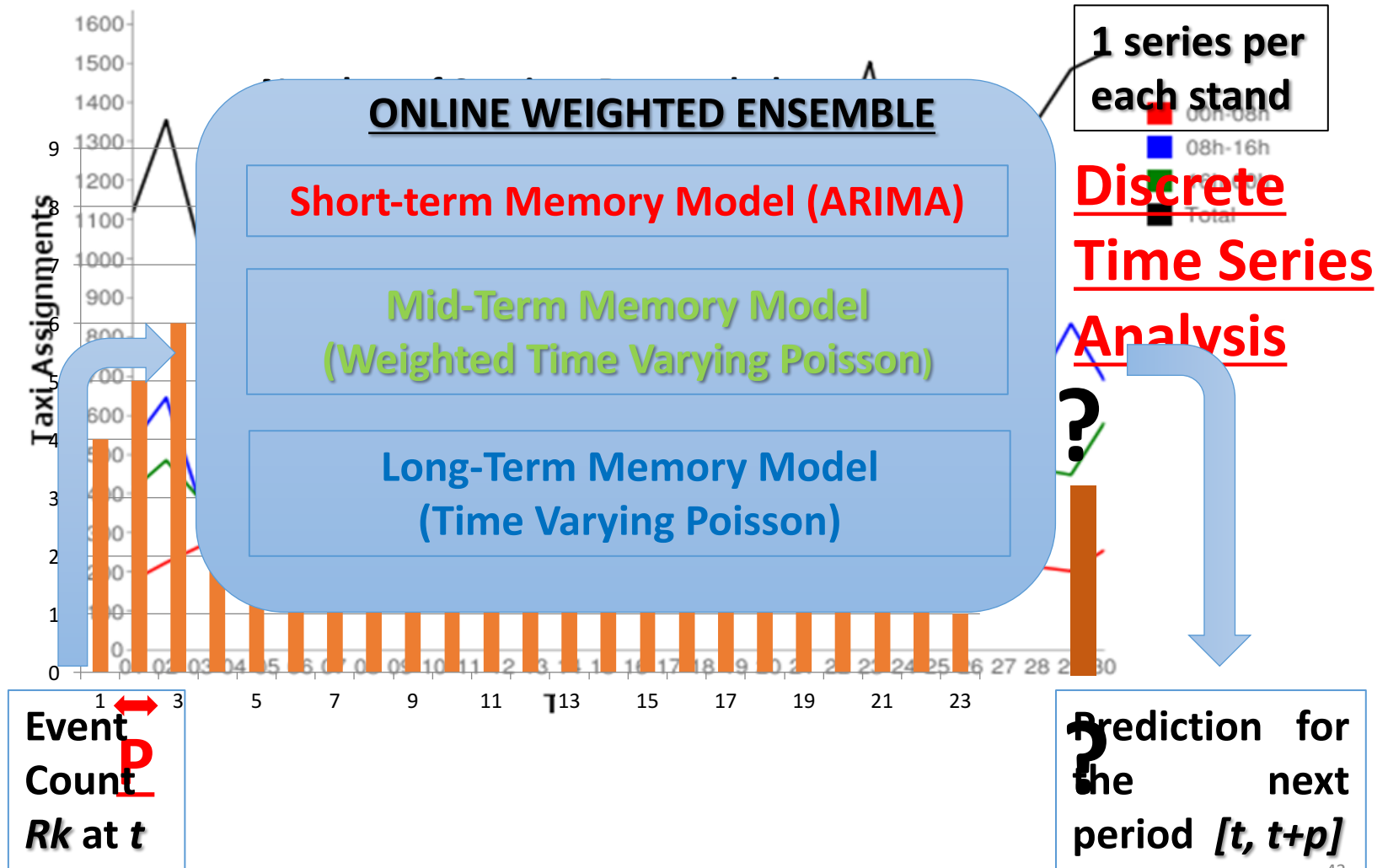
Transportation Systems

Real-time Taxi-Passenger Demand Prediction

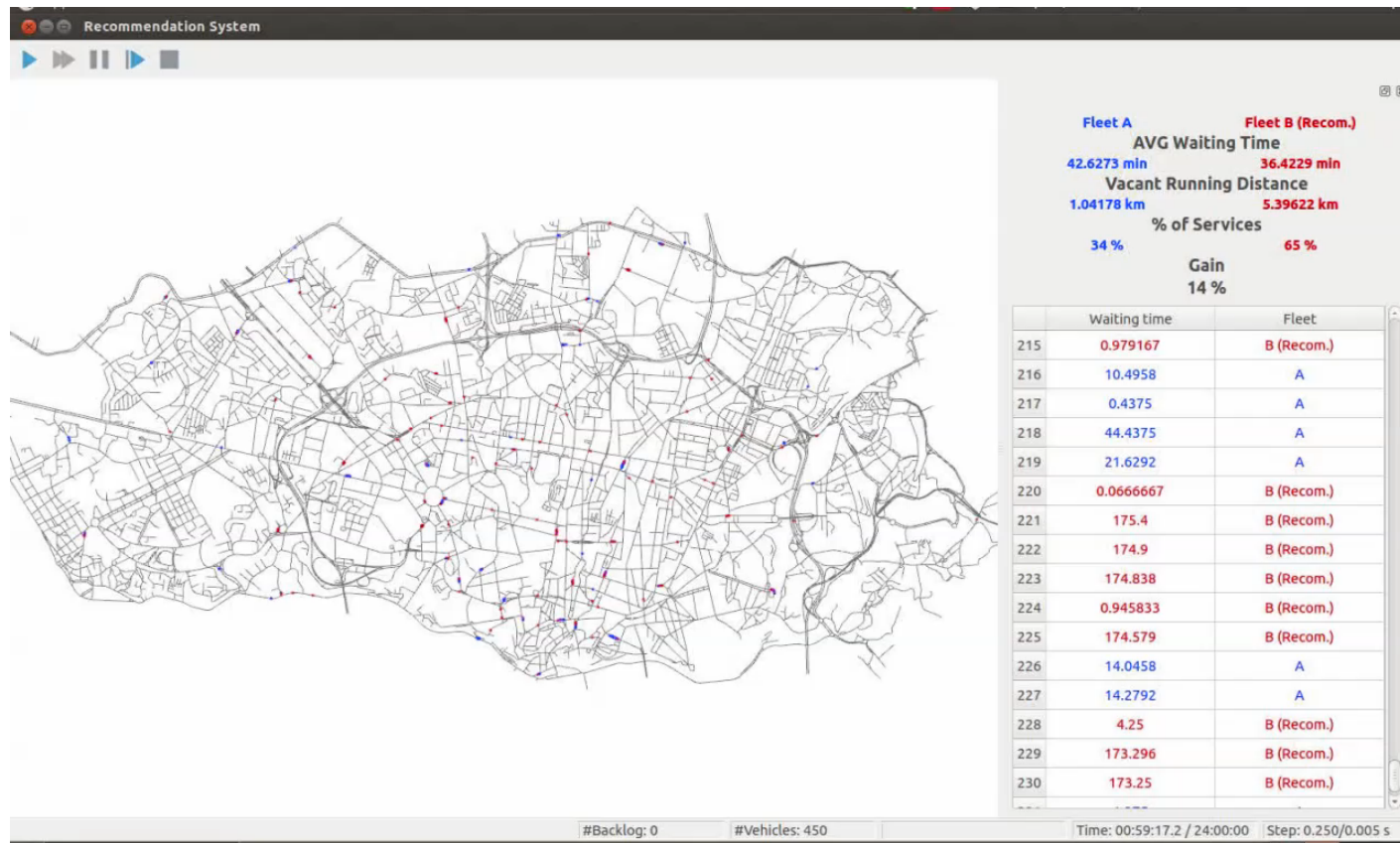
- Predict for each taxi-stand in Porto, the passenger demand for a forecasting **horizon of 30 minutes**;
- **Real World Deployment on Porto, Portugal;**



Taxi-Passenger Demand Prediction



A Prototype on the Recommendation Model



[Moreira-Matias *et al.*, 2012,2013,2013a]

Health

The future of clinical decision

Hospital S. João, Porto

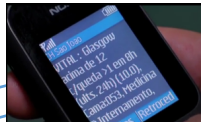
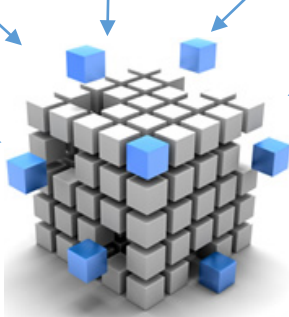
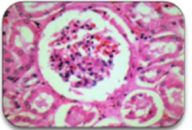
JN 13 May 2014

O Hospital de S. João desenvolveu um sistema inovador que deteta o risco de morte em 50% dos óbitos registados nos doentes internados, através da monitorização constante de todos os parâmetros clínicos.

A ferramenta informática - denominada Vital (Vigilância, Monitorização e Alerta) - foi distinguida pela Microsoft com o Prémio Mundial de Inovação em Saúde, ontem entregue na Florida, EUA. **O carácter inovador e pioneiro reside no facto de agregar todo o tipo de informação (clínica e laboratorial) e calcular índices de risco.**

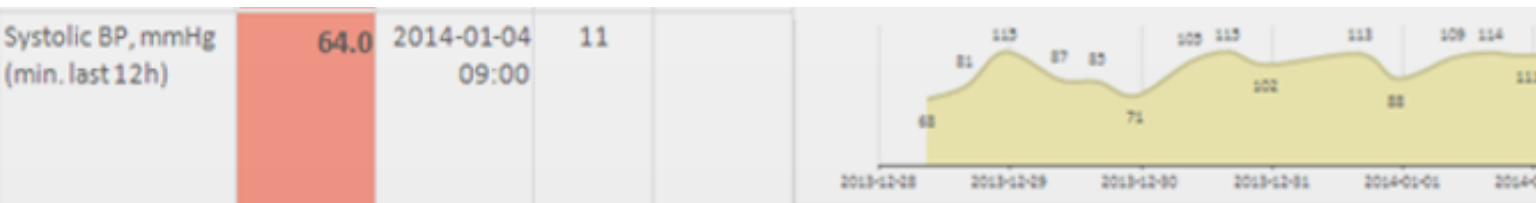
Na prática, isto quer dizer que o Vital indica, por vezes com dias de antecedência, quais os doentes cujo estado vai provavelmente agravar--se, com base não só no estado atual, mas também na evolução dos múltiplos registos.

Arquitetura



HITAL
HEALTH INTELLIGENCE





São João Hospital
HEALTH INTELLIGENCE 1.0.0.0

Date: 1/14/2015 Service: All

All	32
Critical Alerts	4
Worsening	8
Stable	21
Improving	1

Monitoriza em tempo real, avalia o perfil de cada paciente e o seu risco relativo

Alerta as equipas médicas para eventos críticos, tendências e relações problemáticas entre factos aparentemente não relacionados.

VITAL: risk indices



Sofia
57 anos, Feminino
Processo

Emilia
64 anos, Feminino
Processo 0

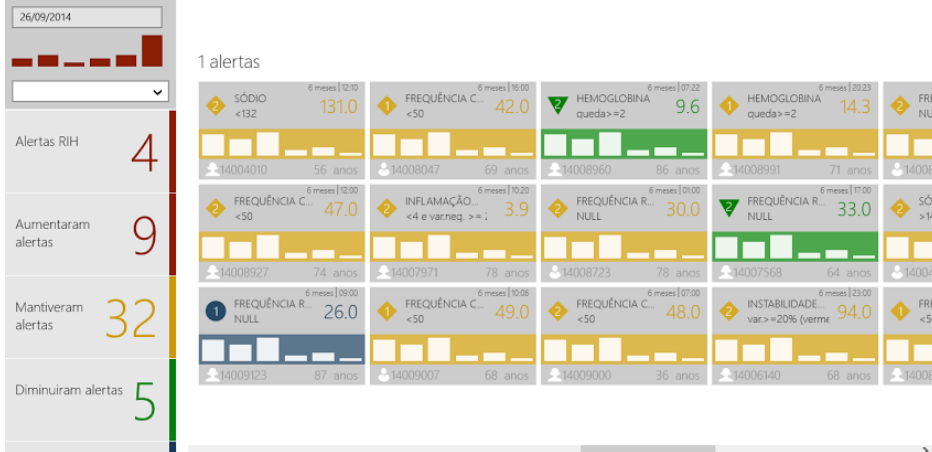
Alertas

- Enterobacteriaceae ESBL + (URINA) Ver mais
- Enterobacteriaceae XDR (URINA) Ver mais
- FR <10 cpm (min. ults. 12h) Ver mais
- TAS <70 mmHg (min. ults. 12h) Ver mais

Episódio Internamento Cama

HOSPITAL SÃO JOÃO
VITAL

Dr. João Lourenço
Cardiologista



Reconhecimentos Internacionais



Microsoft
Health Users Group
Innovation Awards 2014
WINNER

**Microsoft Innovation Award 2014, Florida,
E.U.A, Fevereiro 2014**



**Big Data & Analytics Solution of the
Year, Londres, Março 2014**



**Reconhecimento como mais inovadora
solução no apoio à Decisão Clínica**

Recommender Systems



amazon.com



VIEW CART | WISH LIST | YOUR ACCOUNT | HELP



João's Gold Box

WELCOME JOÃO'S STORE BOOKS APPAREL & ACCESSORIES ELECTRONICS TOYS & GAMES MUSIC COMPUTER & VIDEO GAMES SEE MORE STORES

Recommended for João Gama (If you're not João Gama, click here.)

BROWSE RECOMMENDED

Recommendations

All Stores

- Baby Books DVD Electronics Outdoor Living Tools & Hardware Kitchen & Housewares Magazine Subscriptions Music Computers Camera & Photo Software Toys & Games Video Computer & Video Games

Your recommendations are based on 1 items you own and more.

More results

view: All | New Releases | Coming Soon | Bargains

1.



Machine Learning

by Tom M. Mitchell Average Customer Review: 5 stars Publication Date: March 1, 1997 Our Price: \$143.45 Used & new from \$49.00

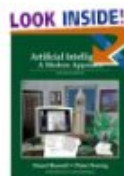
Add to cart Add to Wish List

See related items

Why was I recommended this?

Rate this item x | 5 stars I own it Not interested

2.



Artificial Intelligence: A Modern Approach (2nd Edition)

by Stuart J. Russell, Peter Norvig Average Customer Review: 4.5 stars Publication Date: December 20, 2002 Our Price: \$78.32 Used & new from \$39.00

Add to cart Add to Wish List

See related items

Why was I recommended this?

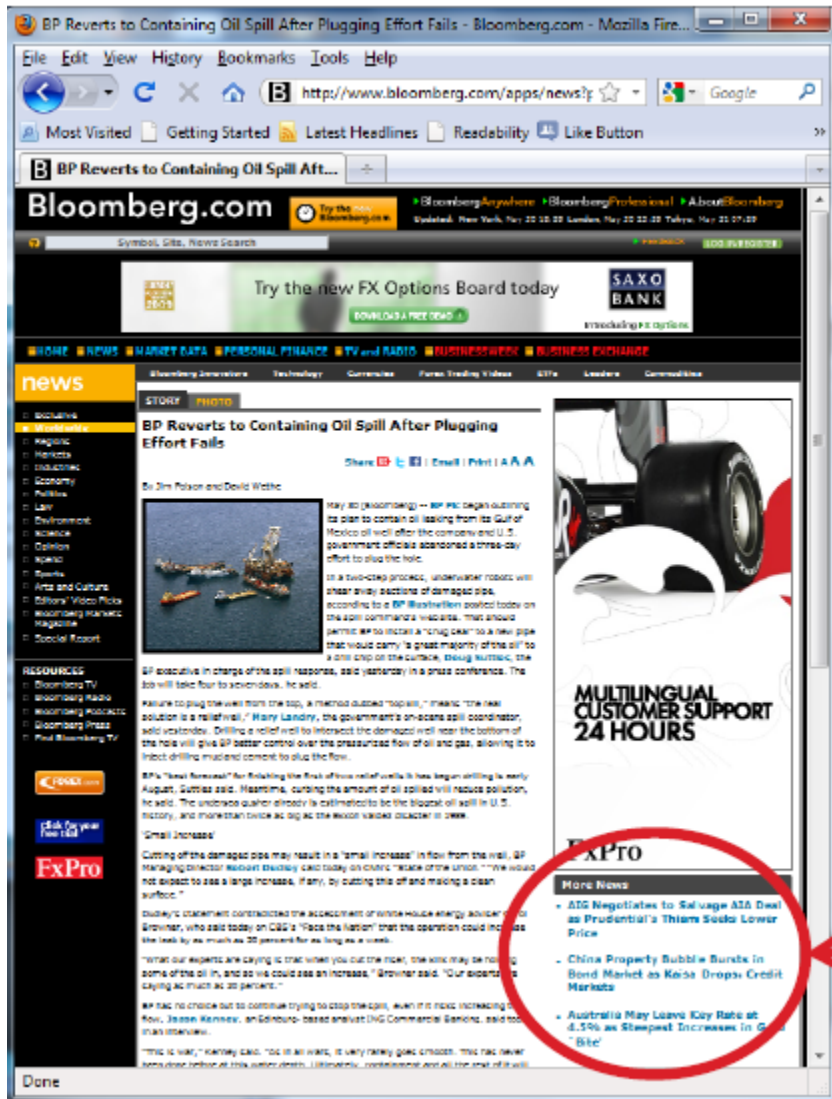
Rate this item x | 4.5 stars I own it Not interested

3.



Neural Networks for Pattern Recognition

Recommender Systems



- ▶ Good recommendations can make a big difference when keeping a user on a web site
 - ...the key is how rich the context model a system is using to select information for a user
 - Bad recommendations <1% users, good ones >5% users click
 - 200clicks/sec

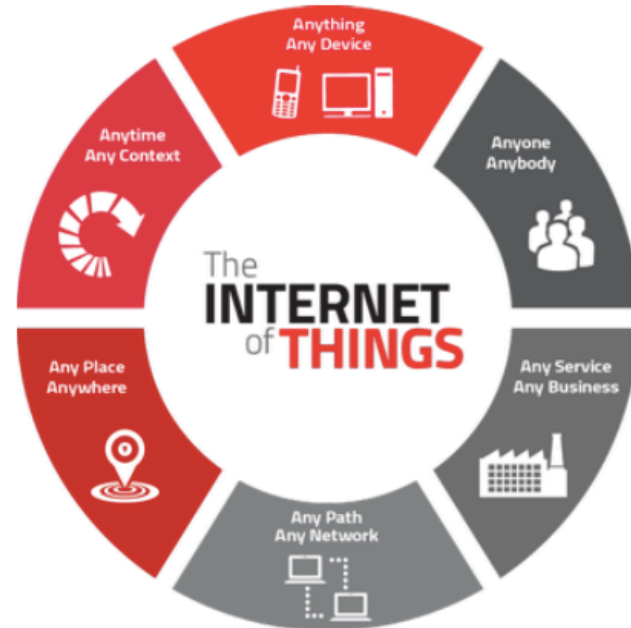
Contextual personalized recommendations generated in ~20ms

To Conclude

Virtualization of the World

- We are living in a connected virtualized world
 - Large luxury fashion (FARFETCH) has no shops
 - Largest taxi company owns no taxis (UBER)
 - Largest phone company own no telco infrastructure (Skype, WhatsApp)
 - Most popular media creates no content (Facebook)
 - Largest movie house owns no cinema (Netflix)

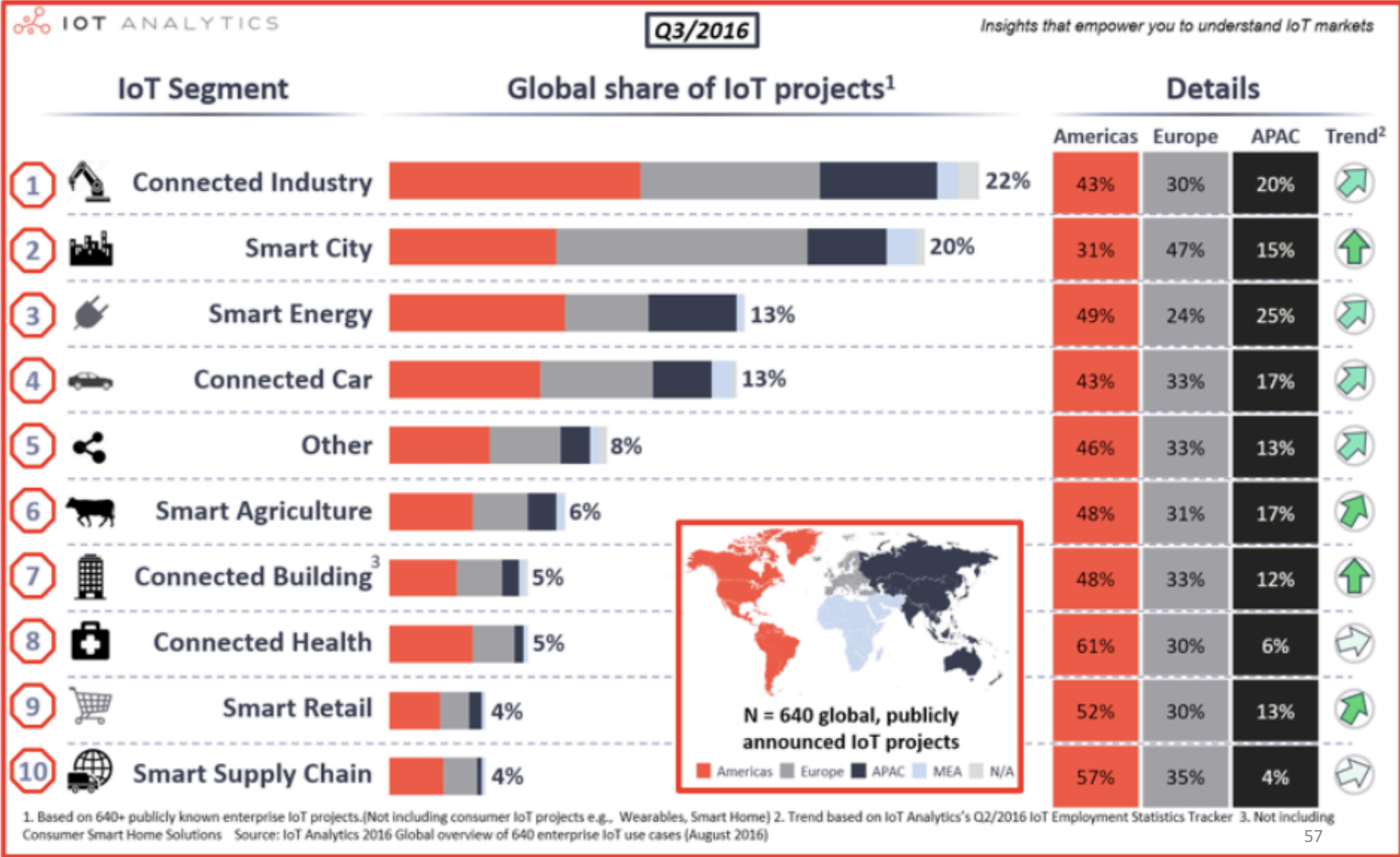
INTERNET OF THINGS



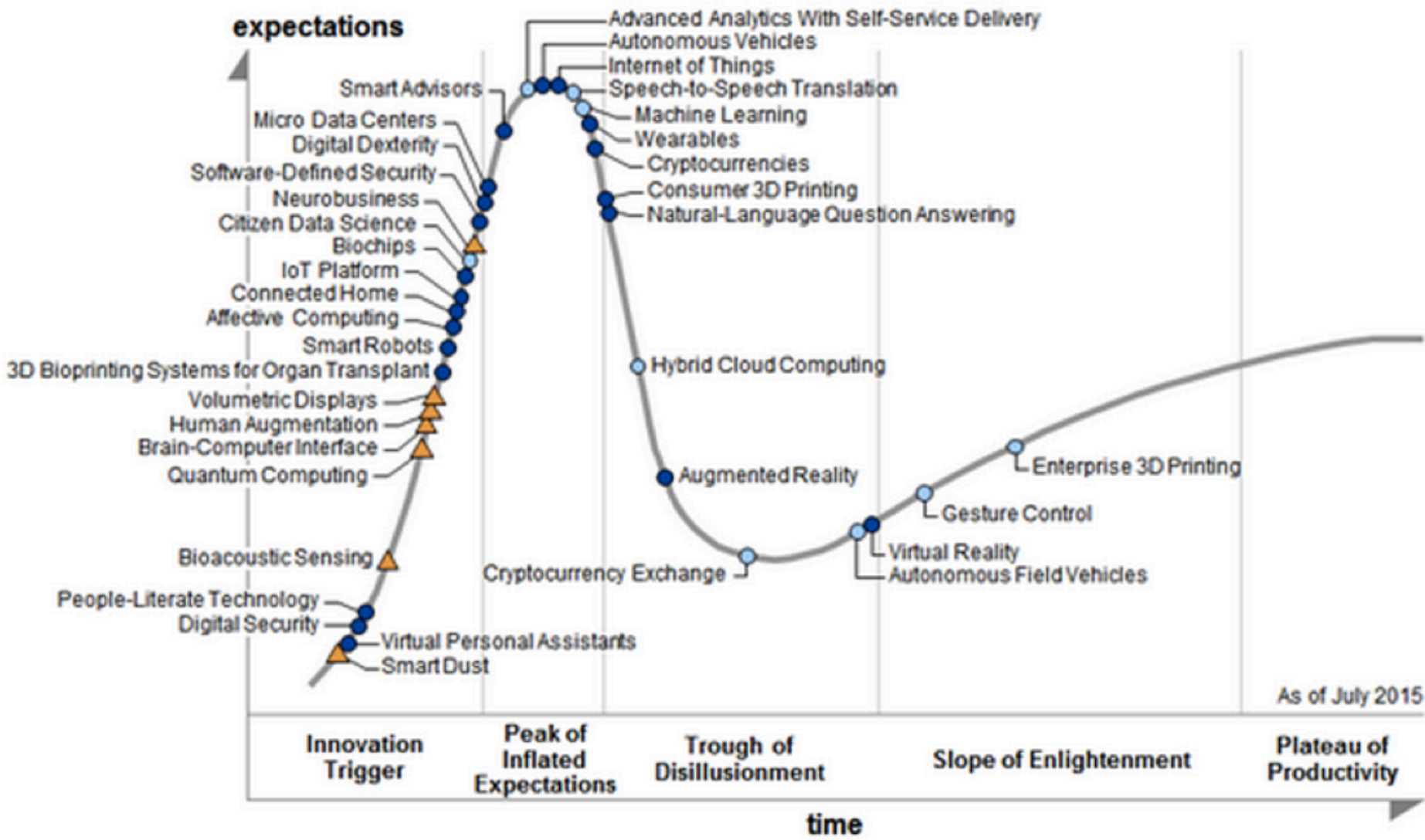
IoT: sensors and actuators connected by networks to computing systems.

- Gartner predicts 20.8 billion IoT devices by 2020.
- IDC projects 32 billion IoT devices by 2020

Applications IoT Analytics



That's all folks !



Plateau will be reached in:

- less than 2 years
- 2 to 5 years
- 5 to 10 years
- ▲ more than 10 years
- ⊗ obsolete before plateau

The Future is here

- The world is faster and smaller
- Small devices are becoming intelligent and reactive
- Able of predictive self-diagnosis;

Strengths

- Any time any where
- Information 24/7
- Personalization

Weaknesses

- Much more cahotic, and risky
- Laws and Ethics not yet defined

Opportunities

- Full of Innovation opportunities
- Is the time of small companies

Threats

- Privacy
- Trustabilty

Quotes

- Data is the new oil. ~ Clive Humby
- Information is the oil of the 21st century, and analytics is the combustion engine. ~ Peter Sondergaard, SVP, Gartner Research
- In God we trust. All others must bring data. ~ W. Edwards Deming,
- Data Scientist: The Sexiest Job of the 21st Century

The image shows a website banner for the EURO IOTA'17 European IOT Analytics Summit. The banner features a dark green header with the event name and navigation links. The main content area has a background image of a cityscape with a river and a bridge. The event title is prominently displayed in white text, along with the date and location. Two buttons, 'INTRODUCTION' and 'PROGRAM', are visible at the bottom.

EUROIOTA'17

INTRODUCTION PROGRAM REGISTRATION ORGANIZATION VENUE

**EUROPEAN IOT ANALYTICS
SUMMIT**

Porto, 24 November 2017

INTRODUCTION PROGRAM



EPIA 2017

18th EPIA Conference on Artificial Intelligence

Porto - 5th-8th September

Thank you!!