KPMG

cutting through complexity

We are Big Data

The future of the information society

prof. dr. Sander Klous

Big Data Ecosystems in Business and Society
University of Amsterdam

Managing Director Big Data Analytics
KPMG Advisory

klous.sander@kpmg.nl
@sanderklous
http://nl.linkedin.com/in/sanderklous

Extreme expectations

https://www.youtube.com/watch?v=2vXyx_qG6mQ

Big, Bigger, Biggest

Key:

q quark

g gluon

e electron

m muon t tau

n neutrino

W, Z bosons

qq̄ meson

qqq baryon

ion

atom

photon

star

galaxy

black hole

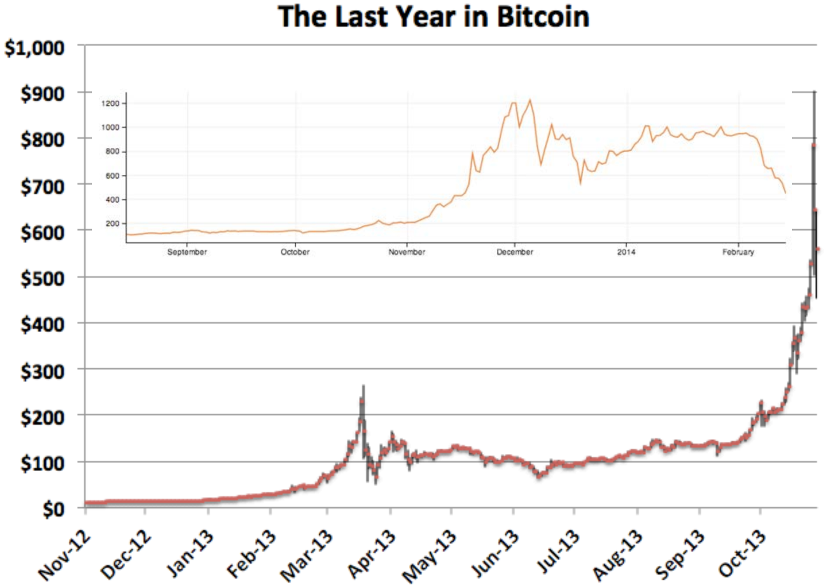
Why do particles have mass?

ter Reid, 2009 (peter.reid@ed.ac.uk)

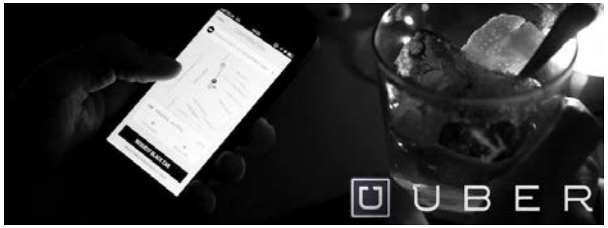
The new information society

Torrents in the music industry, Bitcoins in the financial sector

Download on iTunes



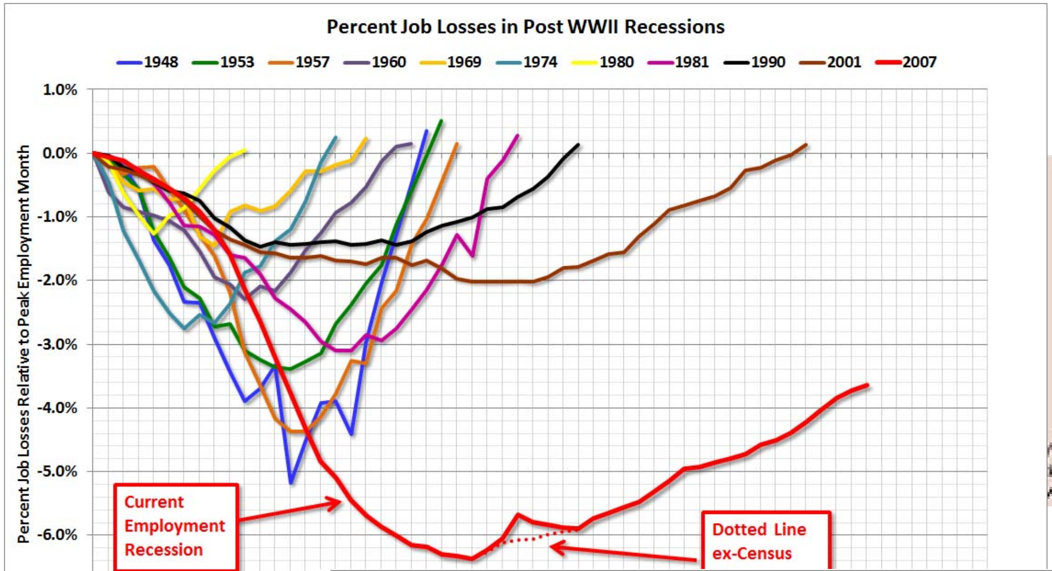
Spotify®

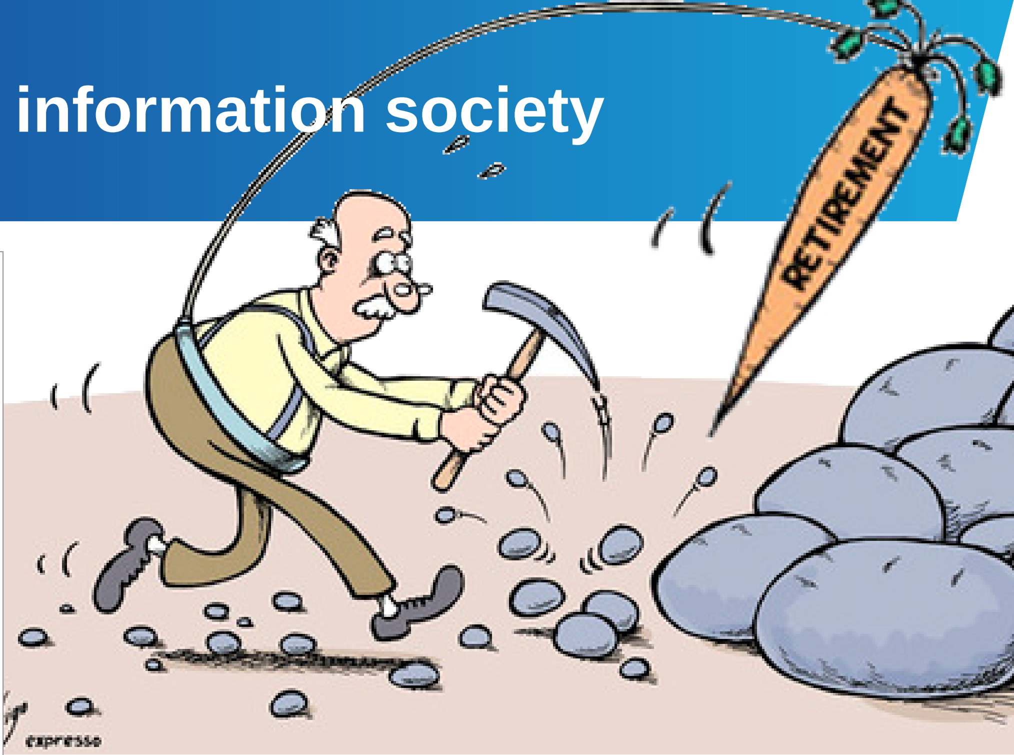


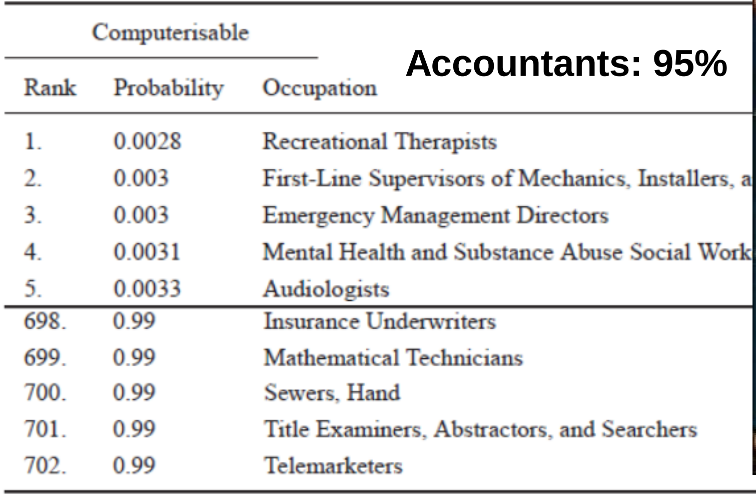
BOOKING.COM
online hotel reservations

When is a trend going to disrupt your business?

Jobless recovery in a developing information society

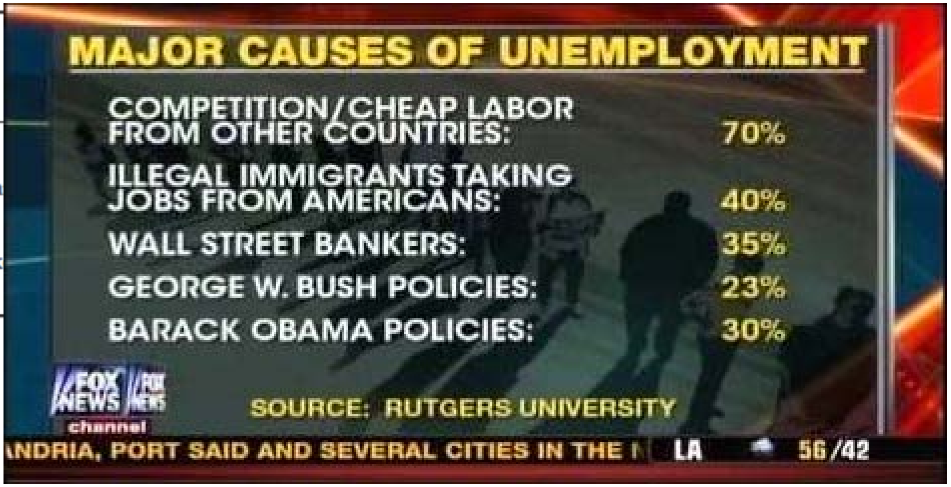






Rank	Computerisable Probability	Occupation
1.	0.0028	Recreational Therapists
2.	0.003	First-Line Supervisors of Mechanics, Installers, a
3.	0.003	Emergency Management Directors
4.	0.0031	Mental Health and Substance Abuse Social Work
5.	0.0033	Audiologists
698.	0.99	Insurance Underwriters
699.	0.99	Mathematical Technicians
700.	0.99	Sewers, Hand
701.	0.99	Title Examiners, Abstractors, and Searchers
702.	0.99	Telemarketers

Accountants: 95%



MAJOR CAUSES OF UNEMPLOYMENT

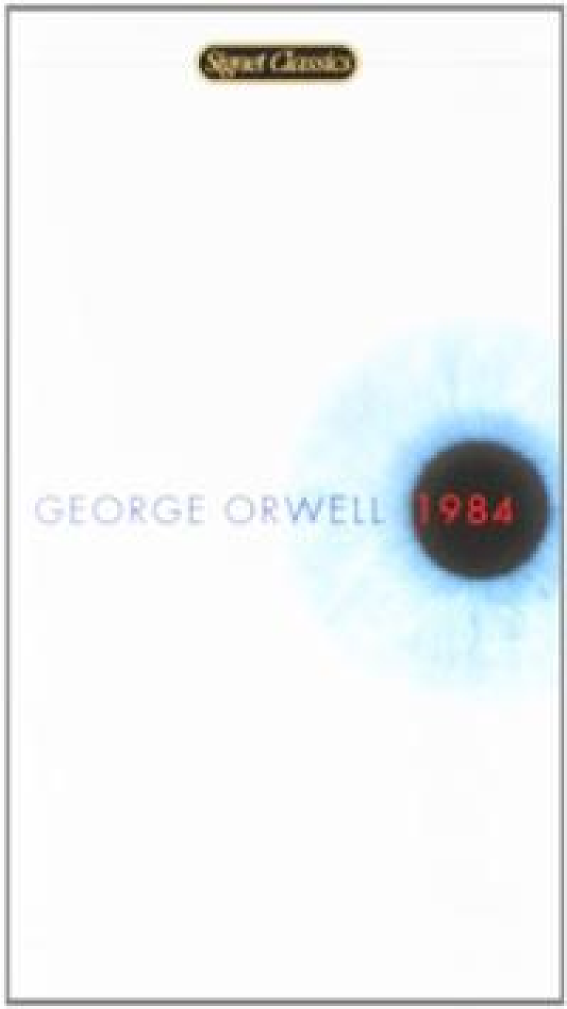
COMPETITION/CHEAP LABOR FROM OTHER COUNTRIES:	70%
ILLEGAL IMMIGRANTS TAKING JOBS FROM AMERICANS:	40%
WALL STREET BANKERS:	35%
GEORGE W. BUSH POLICIES:	23%
BARACK OBAMA POLICIES:	30%

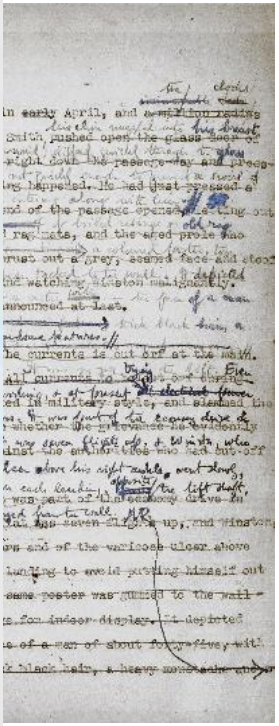
SOURCE: RUTGERS UNIVERSITY

http://www.futuretech.ox.ac.uk/sites/futuretech.ox.ac.uk/files/The_Future_of_Employment_OMS_Working_Paper_1.pdf

The dark side of Big Data

Big Brother is watching you



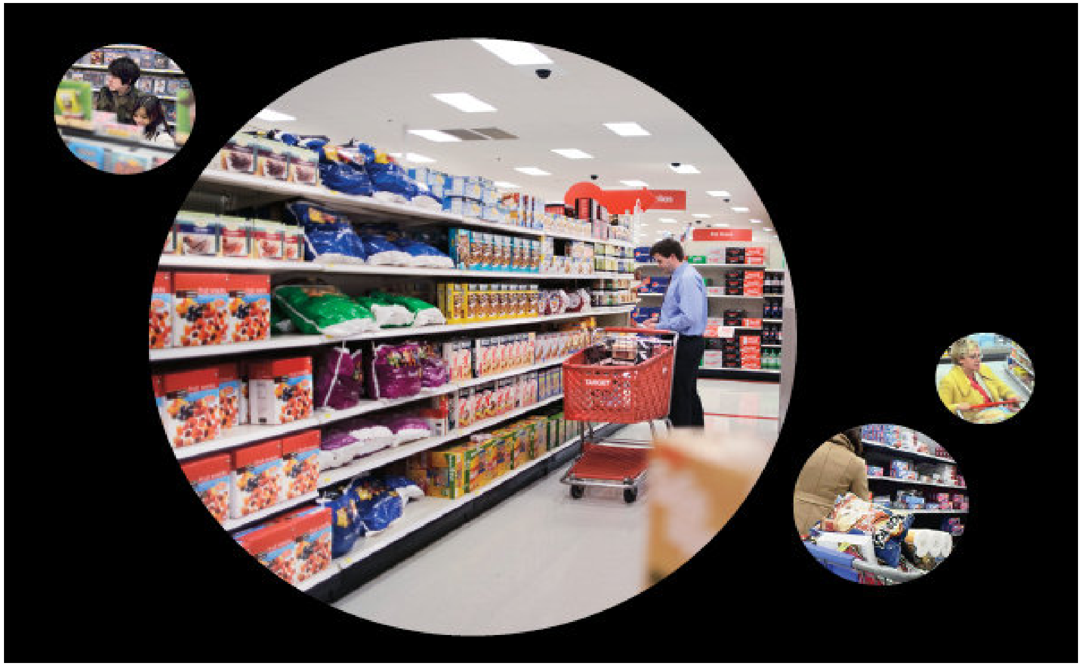

the / clocks
in ~~early~~ April, and ~~a million radios~~
his chin nuzzled into his breast
Smith ~~pushed open the glass door of~~
vile wind, slipped quickly through the glass
~~right down the passage-way and press-~~
not quickly enough to prevent a swirl of
~~ing happened. He had just pressed a~~
entering along with him.
~~end of the passage opened~~ ~~ting out~~
of boiled cabbage and old rag
~~d rag mats, and the aged prole who~~
a coloured poster, too
~~rust out a grey, seamed face and stood~~
had been tacked to the wall. It depicted
~~nd watching Winston malignantly.~~
the face of a man
~~announced at last.~~
thick black hair, a
handsome features.//
~~he currents is cut off at the main.~~
trying Even
~~All currents to~~
working, & at present ~~the electrical power~~
~~ed in military style, and slammed the~~
It was part of the economy drive in
~~whether the grievance he evidently~~
was seven flights up, & Winston, who
~~inst the authorities who had cut off~~
cer above his right ankle, went slowly,
each landing, opposite the lift shaft,
~~was part of the economy drive in~~
ed from the wall. //R
~~was seven flights up, and Winston,~~
~~es and of the varicose ulcer above~~
~~landing to avoid putting himself out~~
~~same poster was gummed to the wall~~
~~s for indoor display. It depicted~~
~~s of a man of about forty-five, with~~
~~k black hair, a heavy moustache and~~

The Target case: life changing events

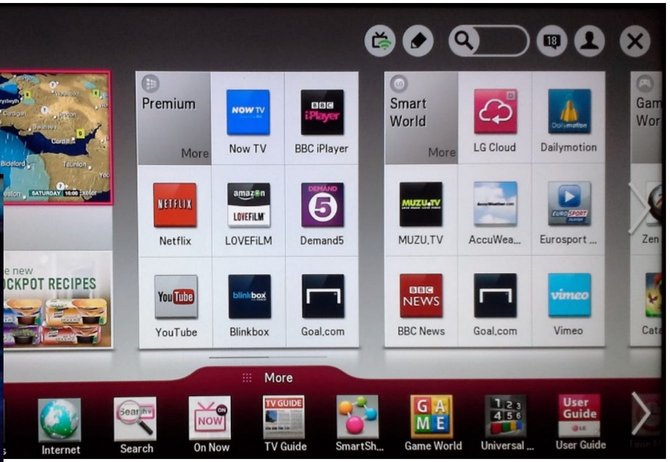
The New York Times

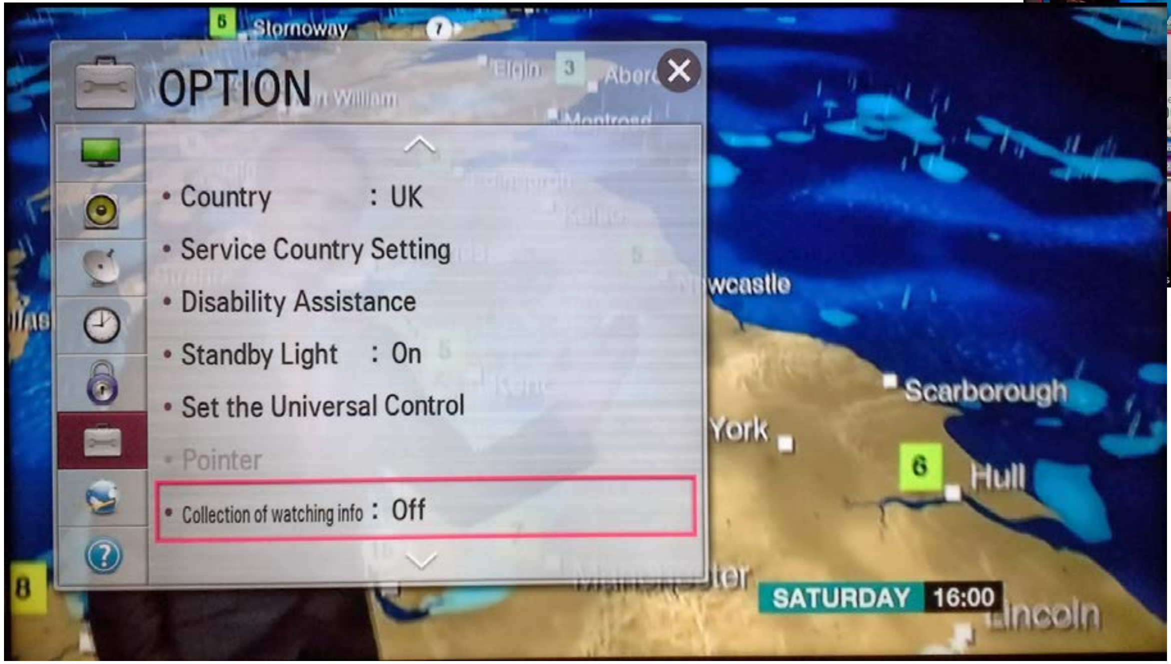
Andrew Pole had just started working as a statistician for Target in 2002…

http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html



Lack of transparency: LG and TP-Link (Philips) example





OPTION

- Country : UK
- Service Country Setting
- Disability Assistance
- Standby Light : On
- Set the Universal Control
- Pointer
- Collection of watching info : Off

http://doctorbeet.blogspot.nl/2013/11/lg-smart-tvs-logging-usb-filenames-and.html
http://www.cbpweb.nl/downloads_pb/pb_20130822-persoonsgegevens-smart-tv.pdf

KPMG
cutting through complexity

Big Brother?

That’s us!

If something is for free, you’re not the client, you’re the product!

Fundamental questions about Big Data

The profile of the data scientist

Apples and Pears

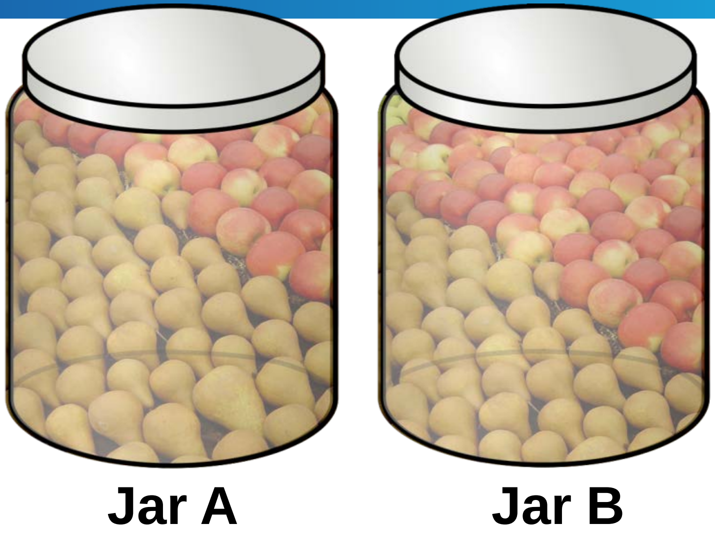
Suppose we have two jars with apples and pears.

Our prior knowledge is that:

- Jar A contains 10 apples and 30 pears
- Jar B contains 20 of each

Fred picks a jar, without further evidence there is a 50% chance this is jar A (or B).

Fred pulls out a pear. The new probability that Fred picked bowl A is 0.75 x 0.5 / (0.75 x 0.5 + 0.5 x 0.5) = 0.6



$$P(H_n|E) = \frac{P(E|H_n)P(H_n)}{\mathrm{Sum}_1^N\,(P(E|H_n))}$$

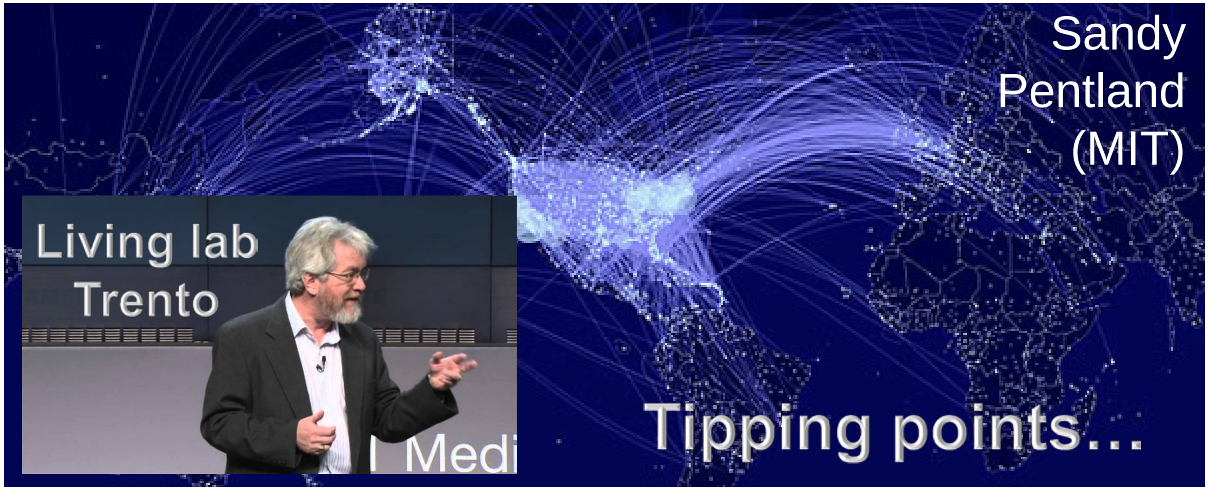
Simpson’s paradox

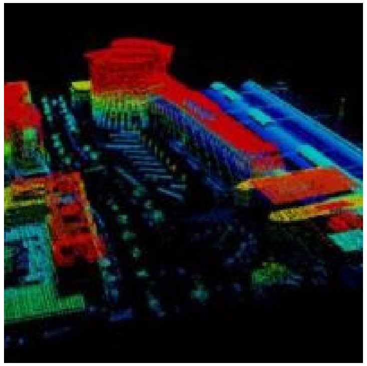
Although in this simple example Bayesian statistics appears to be straight forward, many subtleties arise in analysis. One of the more famous pitfalls is called Simpson’s paradox: http://en.wikipedia.org/wiki/Simpson's_paradox

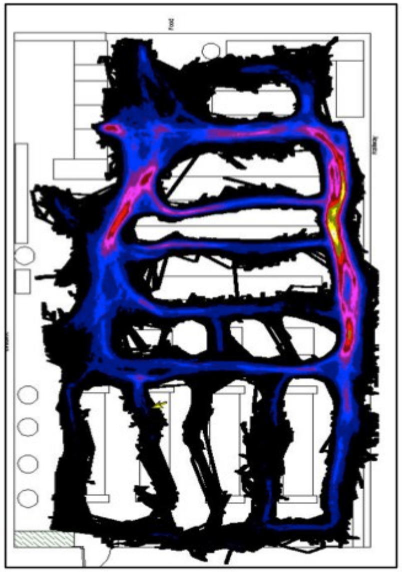
Question:
Do you have a higher chance of survival when falling overboard with or without a life jacket?

Correlation or Causality?

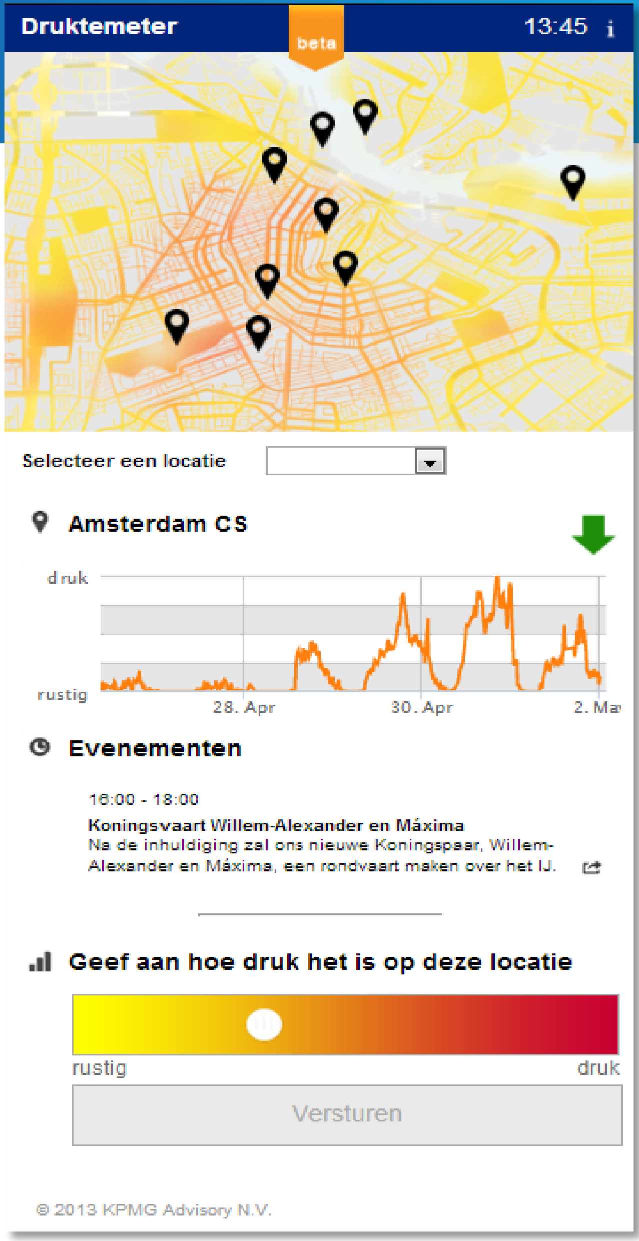
Mechanism Design (Nobel prize 2007)







Understanding large scale human mobility (UvA project)



Druktemeter beta 13:45

Selecteer een locatie

Amsterdam CS

Evenementen

16:00 - 18:00

Koningsvaart Willem-Alexander en Máxima

Na de inhuldiging zal ons nieuwe Koningspaar, Willem-Alexander en Máxima, een rondvaart maken over het IJ.

Geef aan hoe druk het is op deze locatie

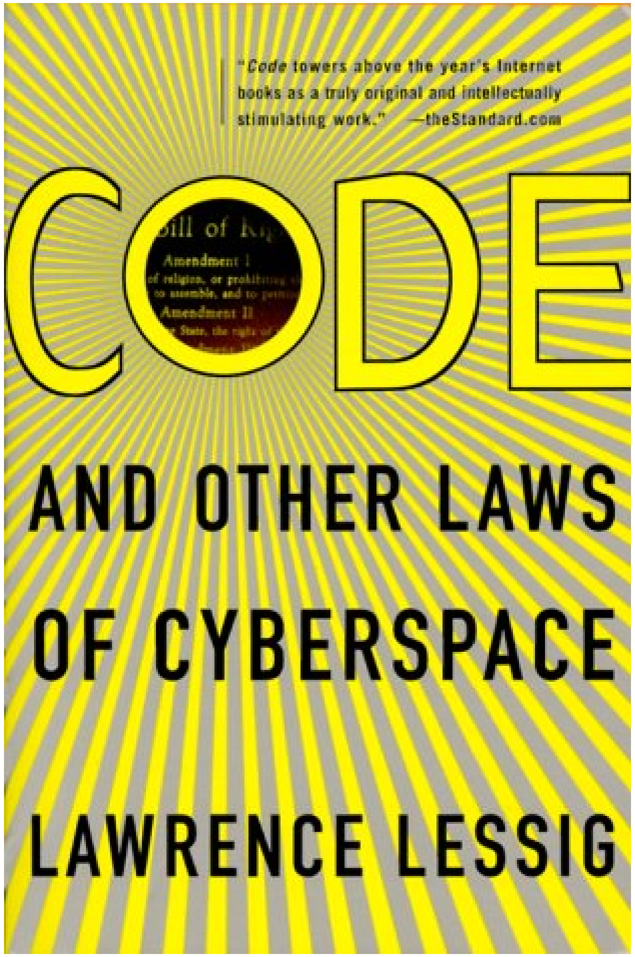
rustig druk

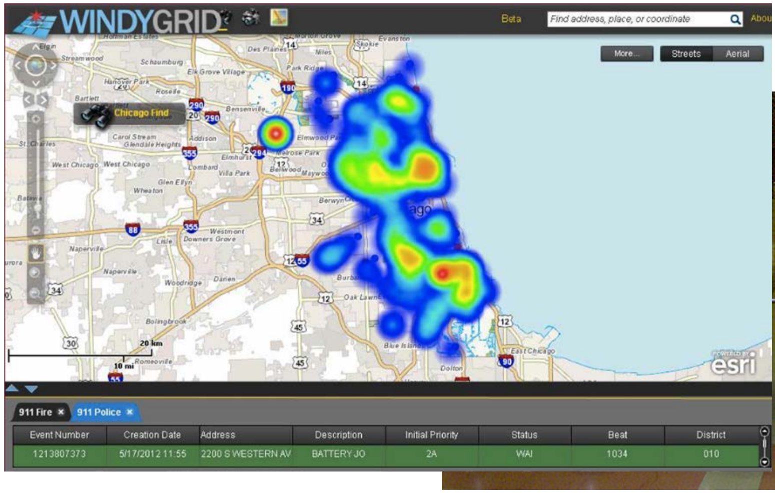
Versturen

© 2013 KPMG Advisory N.V.

Rules and regulations in a do it yourself society

Our behavior is determined by the systems we use

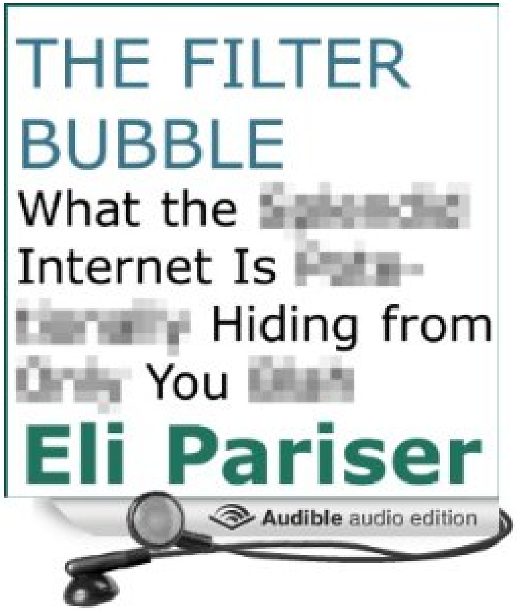

"Code towers above the year's Internet books as a truly original and intellectually stimulating work." —theStandard.com
CODE
AND OTHER LAWS OF CYBERSPACE
LAWRENCE LESSIG

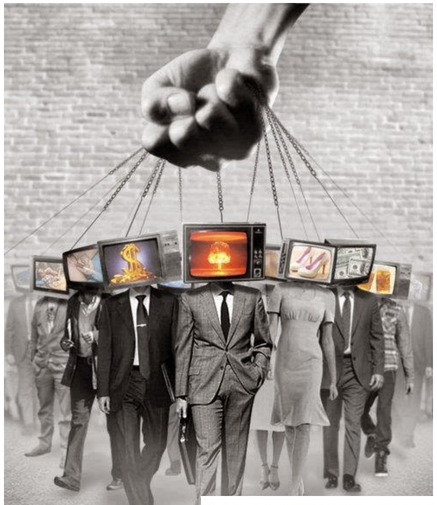


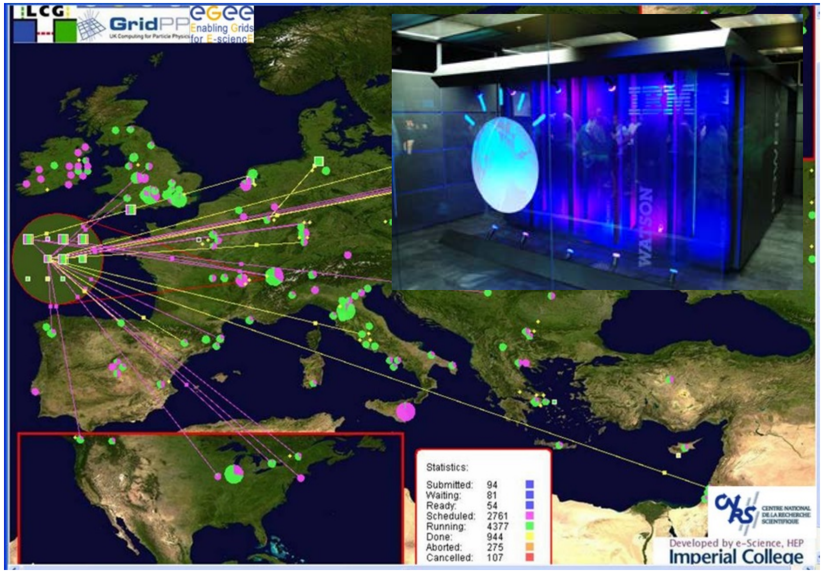
911 Fire ✕ 911 Police ✕

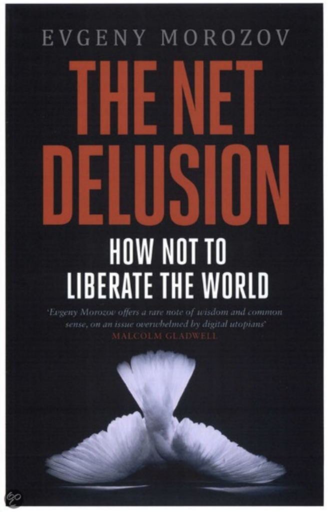
Event Number	Creation Date	Address	Description	Initial Priority	Status	Beat	District
1213807373	5/17/2012 11:55	2200 S WESTERN AV	BATTERY JO	2A	WAI	1034	010

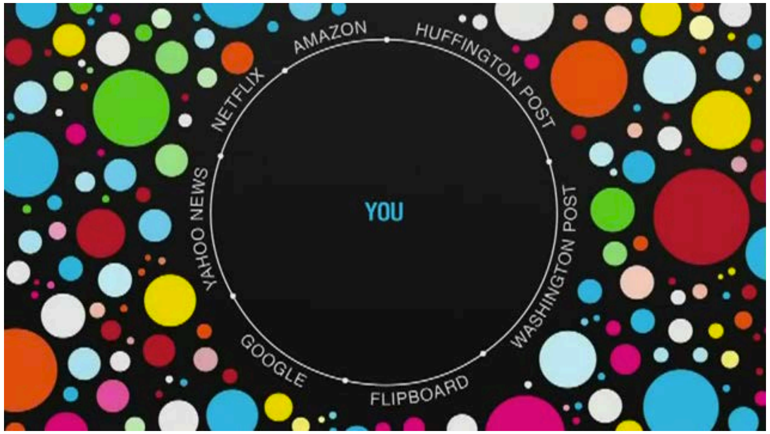
Big Data and Society, an ethical perspective





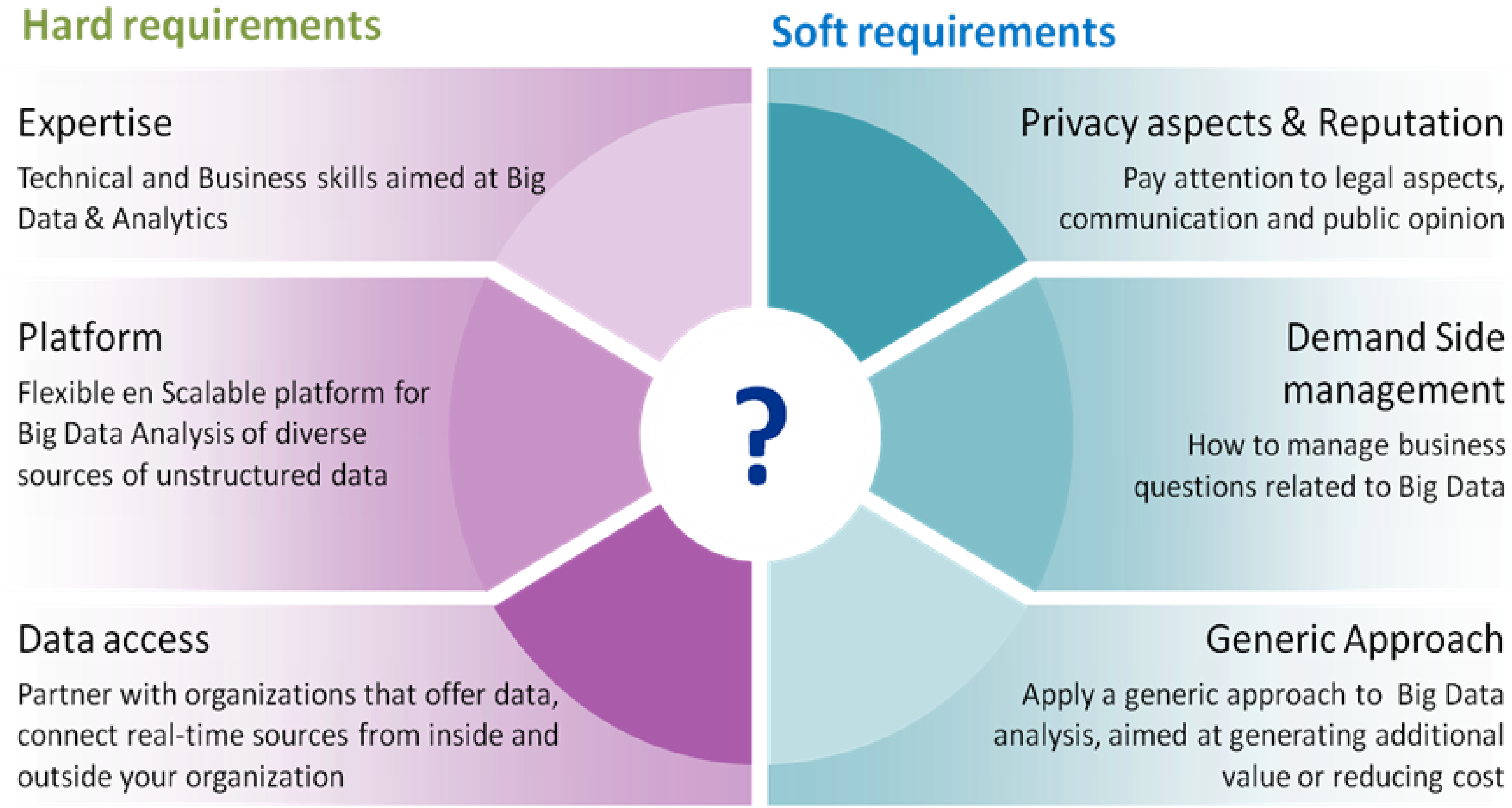



EVGENY MOROZOV
THE NET DELUSION
HOW NOT TO LIBERATE THE WORLD
'Evgeny Morozov offers a rare note of wisdom and common sense, on an issue overwhelmed by digital utopians'
MALCOLM GLADWELL



A Practical Guide to Big Data

Boundary conditions



Hard requirements

Expertise

Technical and Business skills aimed at Big Data & Analytics

Platform

Flexible en Scalable platform for Big Data Analysis of diverse sources of unstructured data

Data access

Partner with organizations that offer data, connect real-time sources from inside and outside your organization

Soft requirements

Privacy aspects & Reputation

Pay attention to legal aspects, communication and public opinion

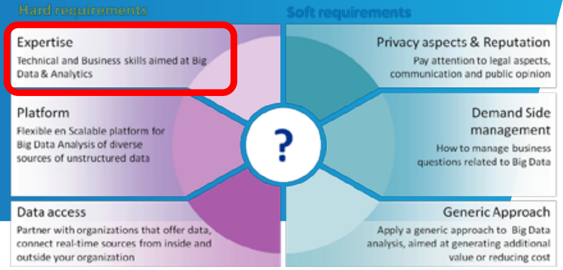
Demand Side management

How to manage business questions related to Big Data

Generic Approach

Apply a generic approach to Big Data analysis, aimed at generating additional value or reducing cost

Expertise



Expertise
Technical and Business skills aimed at Big Data & Analytics

Platform
Flexible en Scalable platform for Big Data Analysis of diverse sources of unstructured data

Data access
Partner with organizations that offer data, connect real-time sources from inside and outside your organization

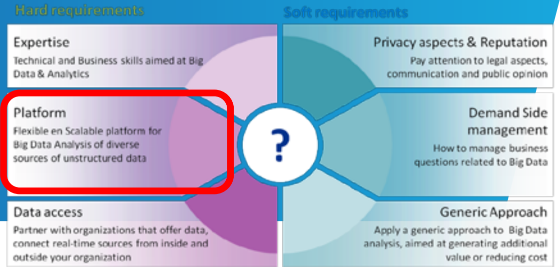
Privacy aspects & Reputation
Pay attention to legal aspects, communication and public opinion

Demand Side management
How to manage business questions related to Big Data

Generic Approach
Apply a generic approach to Big Data analysis, aimed at generating additional value or reducing cost

			Learn Phase I	**Scale Up** Phase II	**Accelerate** Phase III
Client cases	#	0	→ 3	→ 30	→ 90
Data to analyze	TB	0	→ 16	→ 160	→ 500
Resources	# FTE	1	→ 7	→ 23	→ 40
Partners	#	1*	→ 1*	→ 3	→ TBD
Clients	#	1*	→ 2	→ 20	→ TBD
Characteristics per phase			- Light weight project based organisation - Small investments, no fixed costs - Start building expertise on Big Data environments - Small technology investments - Use mostly internal data sets - One or few partners required	- Build data driven organization according to blue print - Limited investment & fixed costs - Start partnerships with external parties - Extend expertise on Big Data environments - Ready for executing bigger and more complex projects - Add data sets from (snowball effect)	- Data driven organization ready to address all market opportunities - Organisational structure and technical infrastructure supports partnering model - Create pull by adding value in Big Data domain

Platform



Hard requirements

Soft requirements

Expertise
Technical and Business skills aimed at Big Data & Analytics

Platform
Flexible en Scalable platform for Big Data Analysis of diverse sources of unstructured data

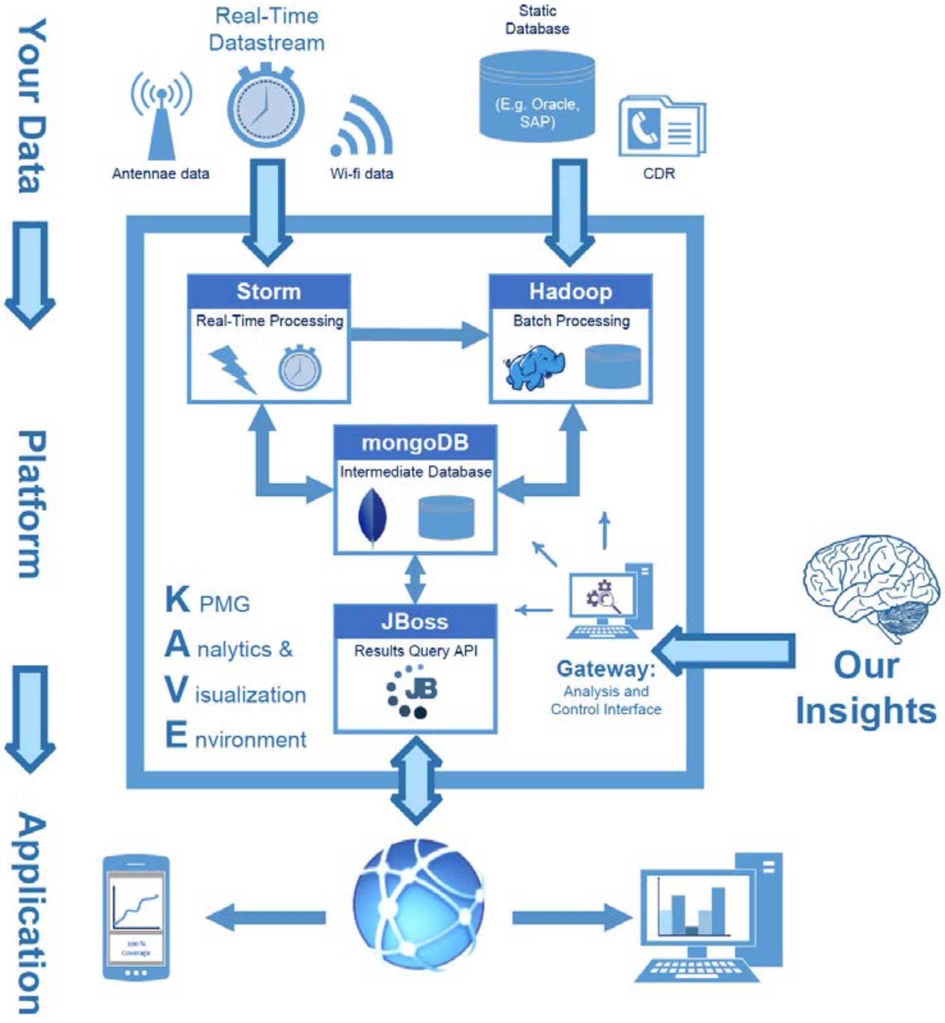
Data access
Partner with organizations that offer data, connect real-time sources from inside and outside your organization

?

Privacy aspects & Reputation
Pay attention to legal aspects, communication and public opinion

Demand Side management
How to manage business questions related to Big Data

Generic Approach
Apply a generic approach to Big Data analysis, aimed at generating additional value or reducing cost



Your Data

Platform

Application

Real-Time Datastream

Antennae data

Wi-fi data

Static Database

(E.g. Oracle, SAP)

CDR

Storm
Real-Time Processing

Hadoop
Batch Processing

mongoDB
Intermediate Database

JBoss
Results Query API

Gateway:
Analysis and Control Interface

K PMG
A nalytics &
V isualization
E nvironment

Our Insights

KAVE
The KPMG Analytics & Visualization Environment (KAVE) has proven to provide an excellent tool stack for processing a wide variety of data like telecommunications and financial services data, containing Data processing, Visualization, Analytics and Management tools.

The Overview

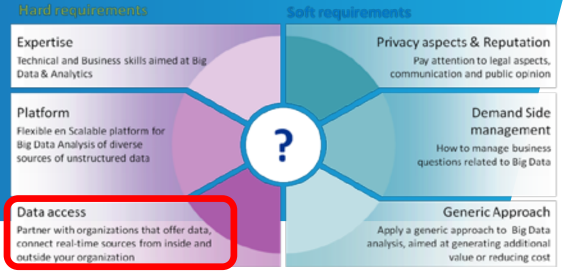
- Horizontally Scalable
- Open Source
- Configurable
- Modular
- Secure

The Implementation

- Remote (or local) hosting
- Dedicated hardware
- Virtualized system
- Secure internal network
- Modular and extendable with anonymization by TTP

http://beta.kave.io

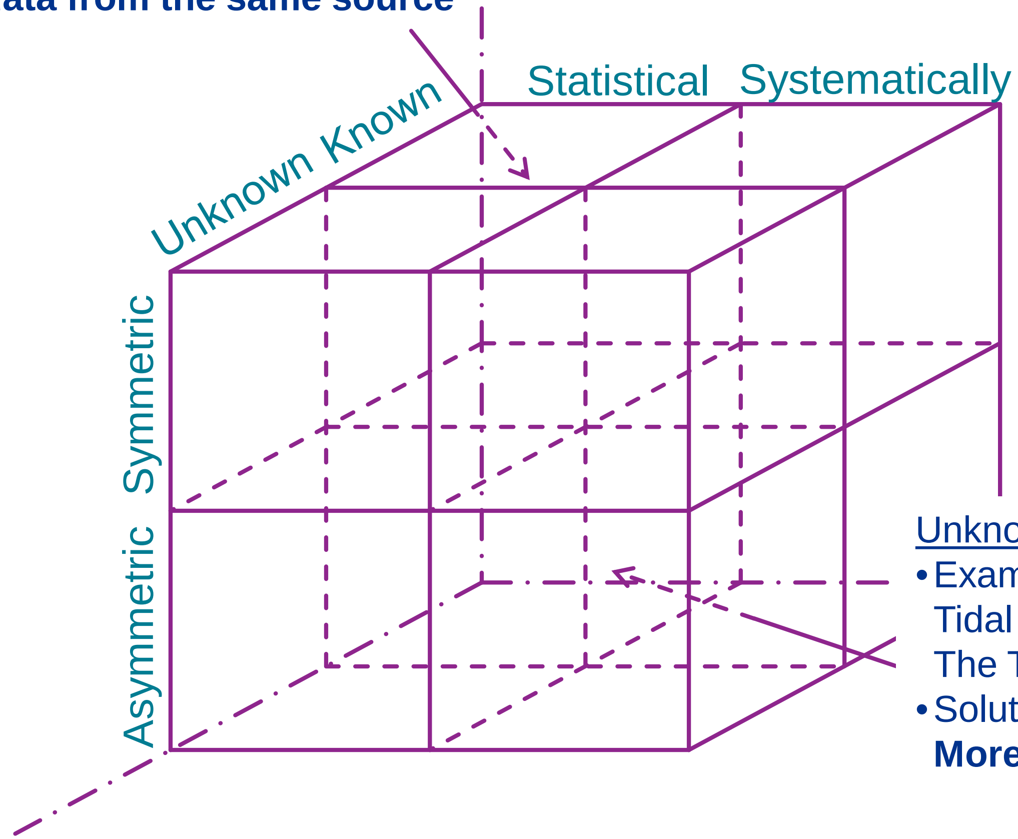
Quantity over Quality



Hard requirements
Soft requirements
Expertise
Technical and Business skills aimed at Big Data & Analytics
Privacy aspects & Reputation
Pay attention to legal aspects, communication and public opinion
Platform
Flexible en Scalable platform for Big Data Analysis of diverse sources of unstructured data
?
Demand Side management
How to manage business questions related to Big Data
Data access
Partner with organizations that offer data, connect real-time sources from inside and outside your organization
Generic Approach
Apply a generic approach to Big Data analysis, aimed at generating additional value or reducing cost

Known symmetric statistical error

- Example:
  Typical Gaussian distributed measurement errors
- Solution to get a more accurate mean value:
  More data from the same source



Unknown asymmetric systematically error

- Example:
  Tidal effects in the lake of Geneva
  The TGV on the train track near CERN
- Solution to get a more accurate results:
  More data from different sources

Governance in a data driven organization

Hard requirements

Soft requirements

Expertise

Technical and Business skills aimed at Big Data & Analytics

Privacy aspects & Reputation

Pay attention to legal aspects, communication and public opinion

Platform

Flexible en Scalable platform for Big Data Analysis of diverse sources of unstructured data

?

Demand Side management

How to manage business questions related to Big Data

Data access

Partner with organizations that offer data, connect real-time sources from inside and outside your organization

Generic Approach

Apply a generic approach to Big Data analysis, aimed at generating additional value or reducing cost

D&A Board

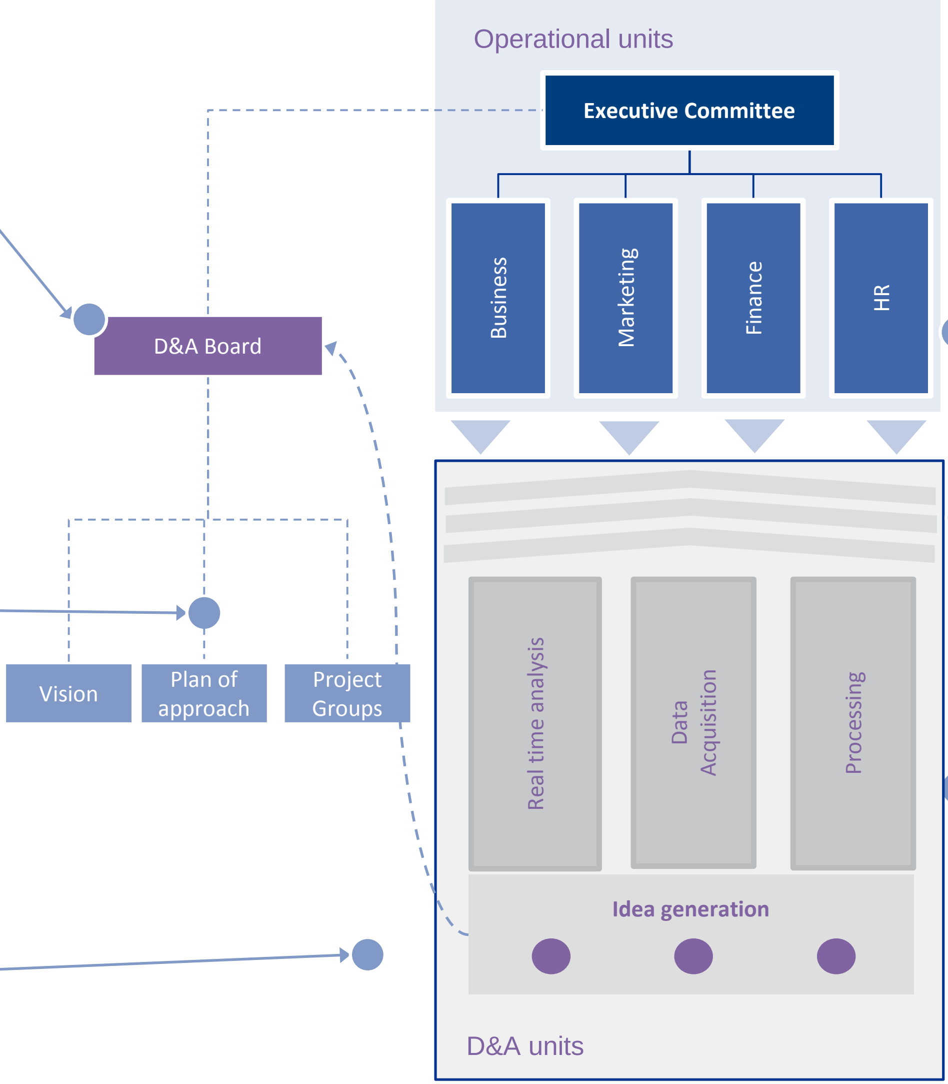
The D&A board is accountable for the success of D&A projects. The board consists of multiple delegates from the operational and D&A units. The board is headed by an executive that reports directly to the CEO.

From ideas to projects

The board is responsible for the vision, plan of approach and formation of project groups. Project groups consist of talent from within the organization to work out their own initiatives and ideas. Participation in these projects is by invitation only and considered a career accelerator. Project development occurs in short cycles.

From data to ideas

The D&A services are managed by the board. Ideas are collected from all sources within the organization (not only from within the D&A Services groups).



Aggregating data

The D&A units facilitate the operational units to capitalize on data through data driven business models.

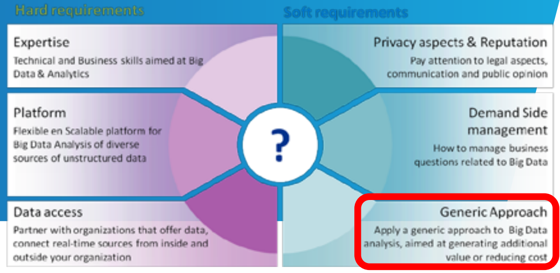
Becoming ‘data driven’

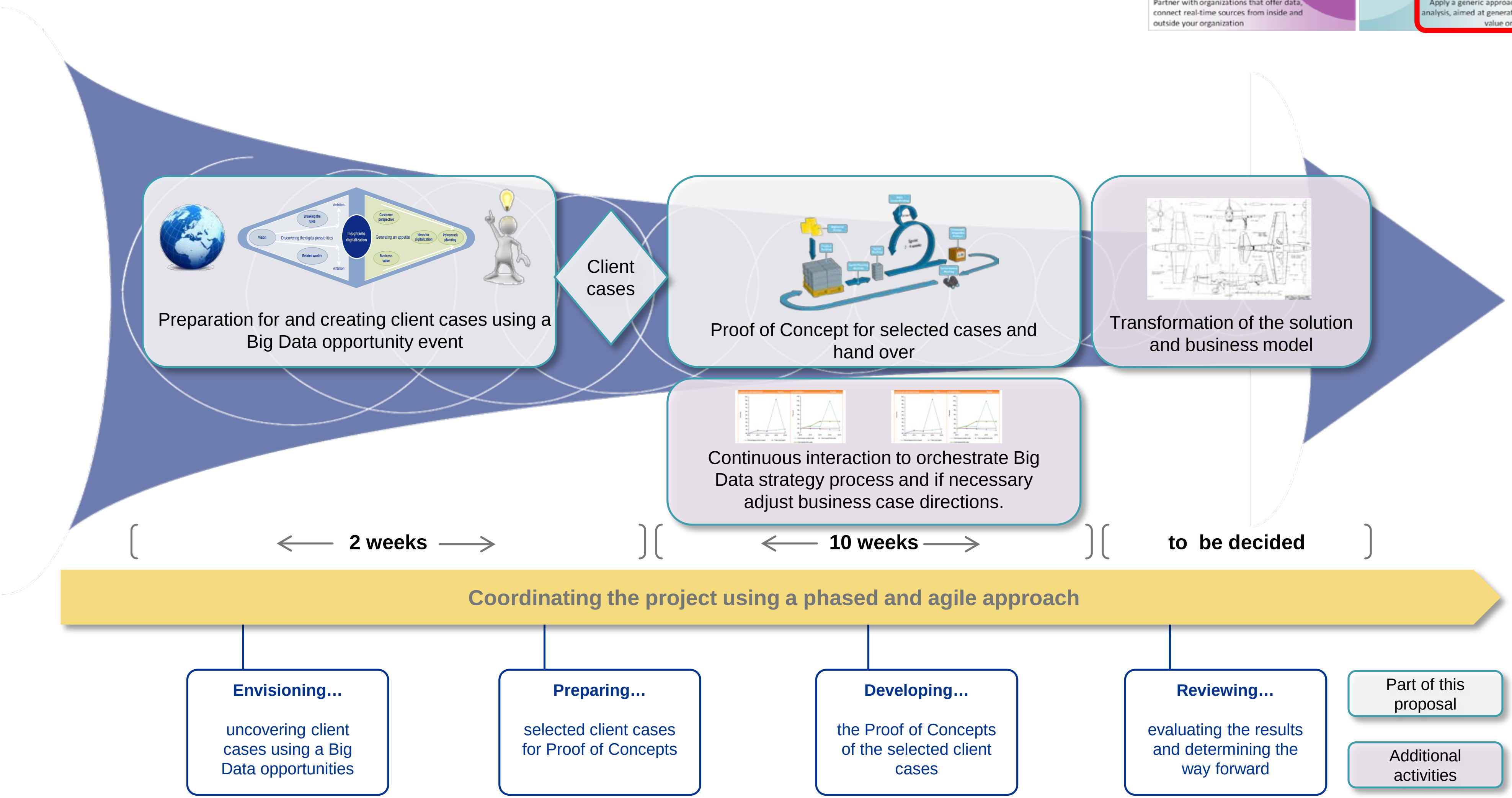
As the D&A units mature, the division of data-oriented functionalities into operational silos decreases.

Understanding D&A services

The D&A units provide a variety of solutions to the operational units. Operational units will continue to run and be responsible for their own activities and associated data.

Agile project management


Hard requirements
Soft requirements
Expertise
Technical and Business skills aimed at Big Data & Analytics
Privacy aspects & Reputation
Pay attention to legal aspects, communication and public opinion
Platform
Flexible en Scalable platform for Big Data Analysis of diverse sources of unstructured data
?
Demand Side management
How to manage business questions related to Big Data
Data access
Partner with organizations that offer data, connect real-time sources from inside and outside your organization
Generic Approach
Apply a generic approach to Big Data analysis, aimed at generating additional value or reducing cost



Preparation for and creating client cases using a Big Data opportunity event

Client cases

Proof of Concept for selected cases and hand over

Transformation of the solution and business model

Continuous interaction to orchestrate Big Data strategy process and if necessary adjust business case directions.

2 weeks

10 weeks

to be decided

Coordinating the project using a phased and agile approach

Envisioning…

uncovering client cases using a Big Data opportunities

Preparing…

selected client cases for Proof of Concepts

Developing…

the Proof of Concepts of the selected client cases

Reviewing…

evaluating the results and determining the way forward

Part of this proposal

Additional activities

Applying Big Data

Start Small

Mortgage Risk Portfolio Management at a large Dutch bank

Background

A Dutch retail bank started a project combining internal and external data sources. The goal is to model risk for a collection of mortgages based on life-changing events in order to reduce costs on financially unhealthy clients. By combining public information on an online house-selling website with internal bank data, a model has been built to describe and predict the length of mortgages and to analyze the difference in house-selling price and the mortgage.

Requirements / Challenges

- More than 200,000 mortgages & over 250,000 homes for sale on Funda.nl, while limited or no understanding of customers selling houses
- Privacy challenge: sensitive personal information needs to be anonymized

Resources

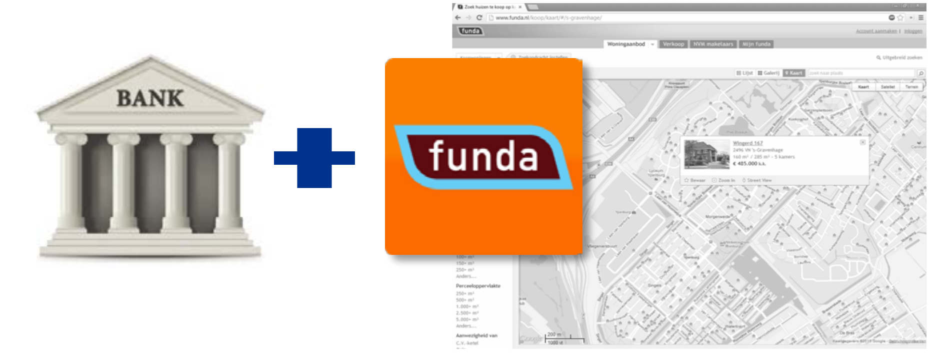
- Data sources: internal bank data, open data crawled from Funda.nl
- Duration (develop phase): 6 –10 weeks

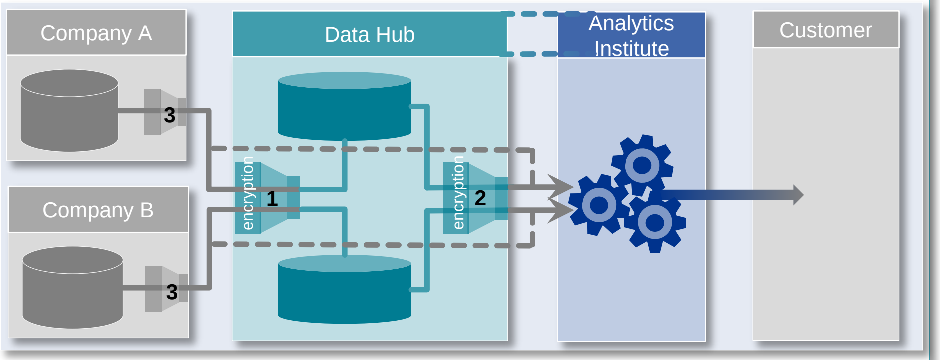
Solution

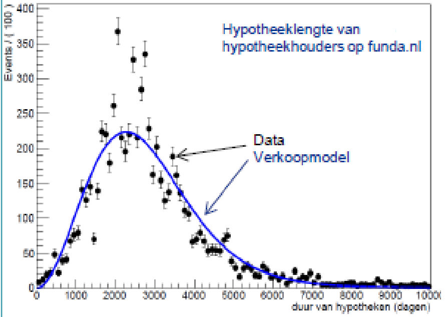
- An application crawls the Funda.nl pages on a daily basis
- The results are loaded into a Teradata Aster Cluster and combined with bank information on customers and products
- Mortgages and Funda.nl results are linked based on postal code and house number

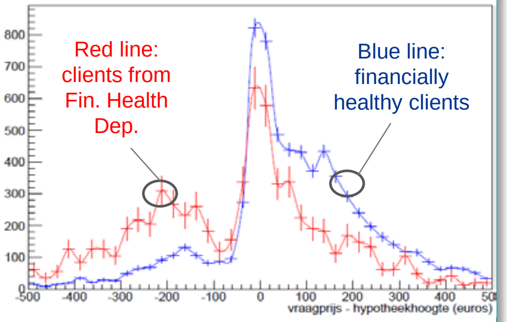
Benefits

Daily insight into the housing market: prices, locations and time and understanding customers that offer a home or apartment for sale









Service

Predicting Credit Card Remediation: targeting valuable customers the right way

Background

Usage of multiple credit cards for everyday payments is very common in Taiwan, and competition between banks with attractive credit card promotion campaigns is high. To avoid customer churn, a bank needs to come up with smart and interesting offers at the right moment. Internal data from the bank, combined with data from the country's credit card information center (JCIC), is used to predict which inactive credit card holders are susceptible to a specific remediation campaign and start using the bank's credit card again.

Requirements / Challenges

- Combine transaction data, campaigning data, and external JCIC data
- Find powerful, discriminating variables in credit card data (300 GB)

Resources

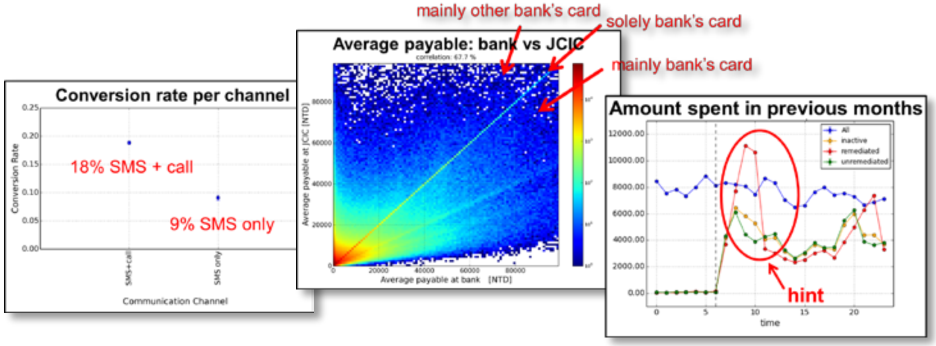
- KPMG Analytics & Visualization Environment (KAVE) for data processing
- Duration (develop phase): 3 weeks with a team of data scientists

Solution

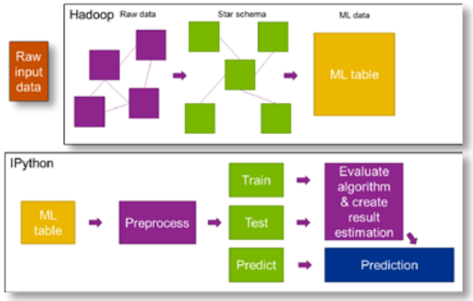
- Combine machine learning with data-driven discovery approach: machine learning to identify correlations and data-driven discovery to study the identified correlations in detail, i.e. *moving from correlation to causality*.
- Identified groups of valuable customer with highest remediation chance.

Benefits

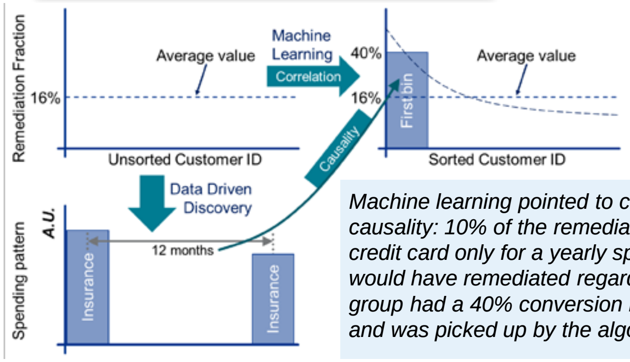
Raise in effectiveness of a remediation campaign: budget only spent on inactive customers that are potentially valuable and are considered susceptible. An estimated increase in revenue regain of € 10-50 million per month. Enhanced loyalty due to offerings that are relevant for the customer.



Top three plots:
data-discovery plots with conversion rates, loyalty, and spending patterns for various types of customers.



Left two diagrams:
A star schema was developed to combine internal and external data as input for a machine learning process



Machine learning pointed to correlations, data-discovery to causality: 10% of the remediated card holders use their credit card only for a yearly spending on insurance and would have remediated regardless any campaign. This group had a 40% conversion rate compared 16% in total and was picked up by the algorithms to be contacted first.

Asset management and planning at a University Medical Center

Background

Monte Carlo simulation to analyze operational efficiency of Operation Room planning at a large academic hospital. A new blue print for OR planning was created and needed to be evaluated on performance, expected production numbers and where issues might arise.

Requirements / Challenges

- Handling poor data quality
- Translate highly flexible human decision-making to rules in model
- Client involvement (expert knowledge) was needed throughout the process in order to create an expert system for planning

Resources

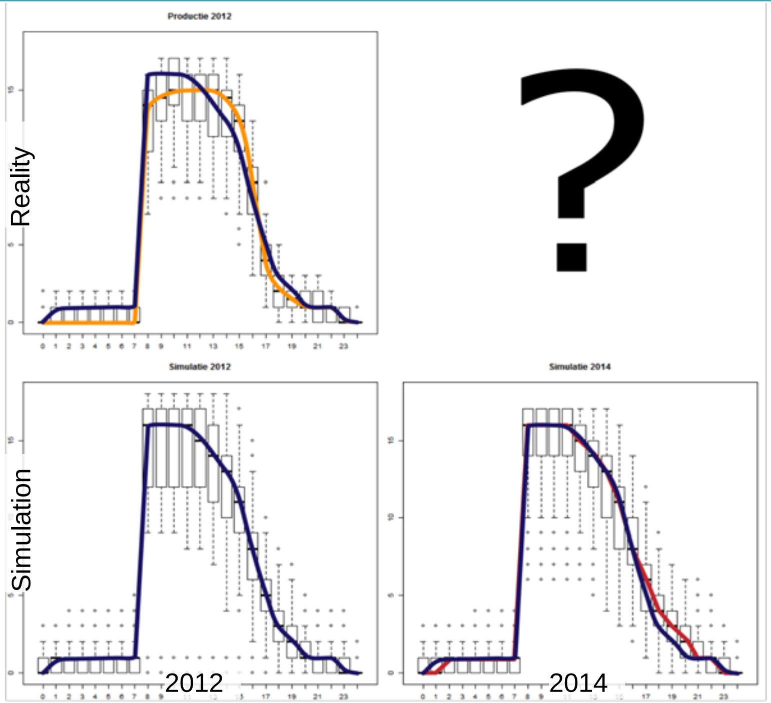
- Data sources: blueprint, historical OR usage data
- Duration (develop phase): 10 –16 weeks

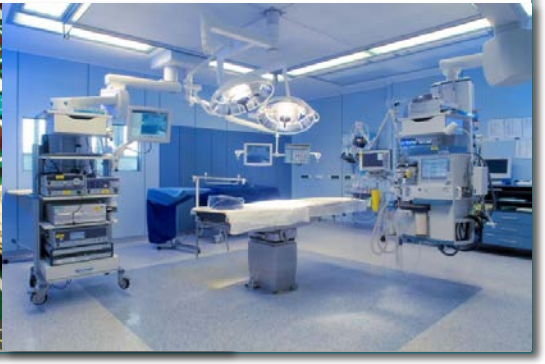
Solution

Built a Monte Carlo simulation model to evaluate the expected performance of the blueprint in various scenario’s and to quantify these results.

Benefits

With our results, the client gained insight into the expected performance of the new blueprint, including the expected performance for several alternative scenario’s.





Service

Enhancing traffic management through combining multiple large data sources

Background

The Dutch governmental body NDW involves various authorities working closely together to develop a database providing information on the current traffic situation on the motorways, secondary roads and urban thoroughfares of the participating authorities. NDW distributes this data to road authorities and traffic information providers, who then inform road users of the traffic situation. This means the road user can be better informed before and during a journey, and road authorities can optimize traffic flow over the road network.

Requirements / Challenges

How can NDW data be leveraged with external data sources to further enhance traffic management? Two sources were identified and investigated:

- Weather radar data (KNMI, www.nationaleregenradar.nl).
- Tweets on motorways in the Netherlands.

Resources

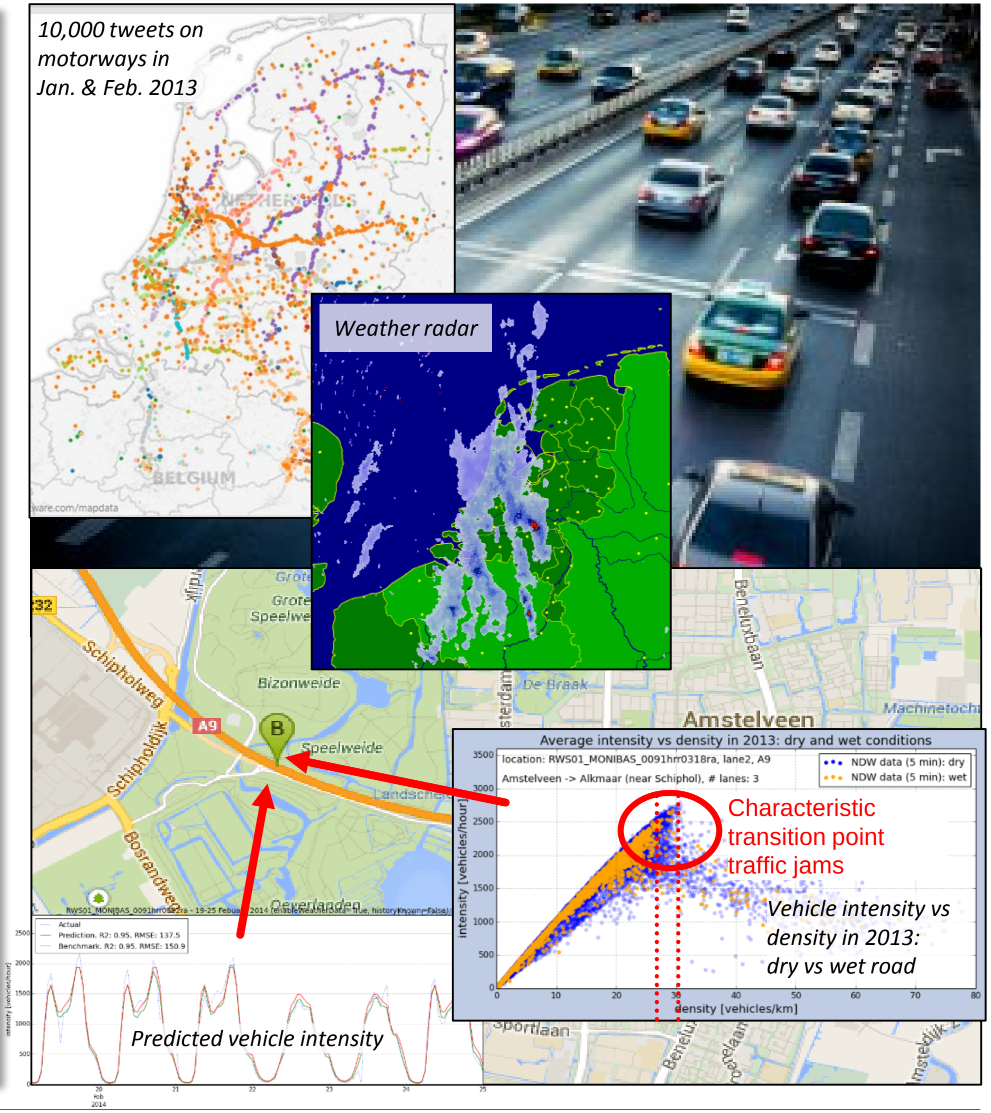
- Data: 200 TB from 24,000 measuring points along 6,000 km of roads; 2.5 GB of tweets with GPS in The Netherlands between Jan-Feb 2013; 10 GB of weather radar data on precipitation per 5 minute interval in 2013.
- Duration (pilot) : 2-4 weeks, with 2 data scientists.

Solution

The data sources were combined on the KAVE platform (KPMG Analytics and Visualization Environment) to find correlations. The ‘Random Forest’ machine-learning technique was applied on historic data to predict traffic situations in advance based on current and historic road and weather conditions.

Benefits

This pilot demonstrated the possibilities with the data at hand and the technical feasibility of two use cases relying on weather data and social media through the usage of Big Data enabling technologies and algorithms. To be continued.



10,000 tweets on motorways in Jan. & Feb. 2013

Weather radar

Characteristic transition point traffic jams

Vehicle intensity vs density in 2013: dry vs wet road

Predicted vehicle intensity

© 2014 KPMG Advisory N.V., registered with the trade register in the Netherlands under number 33263682, is a subsidiary of KPMG Europe LLP and a member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative (‘KPMG International’), a Swiss entity. All rights reserved. Printed in the Netherlands. The KPMG name, logo and ‘cutting through complexity’ are registered trademarks or trademarks of KPMG International.

28

Behavioral patterns:
Crowd & Mobility management at the Dutch coronation event

Background

Our work for a large Dutch telecom provider allowed us to gain experience with various cases regarding crowd control and location based services. Especially for the coronation of the King in 2013, a website and widget were developed in close collaboration with several parties.

Requirements / Challenges

- Intensity and movement of crowd on street level based on antenna data
- Sentiment analysis of social media
- Real-time feedback

Resources

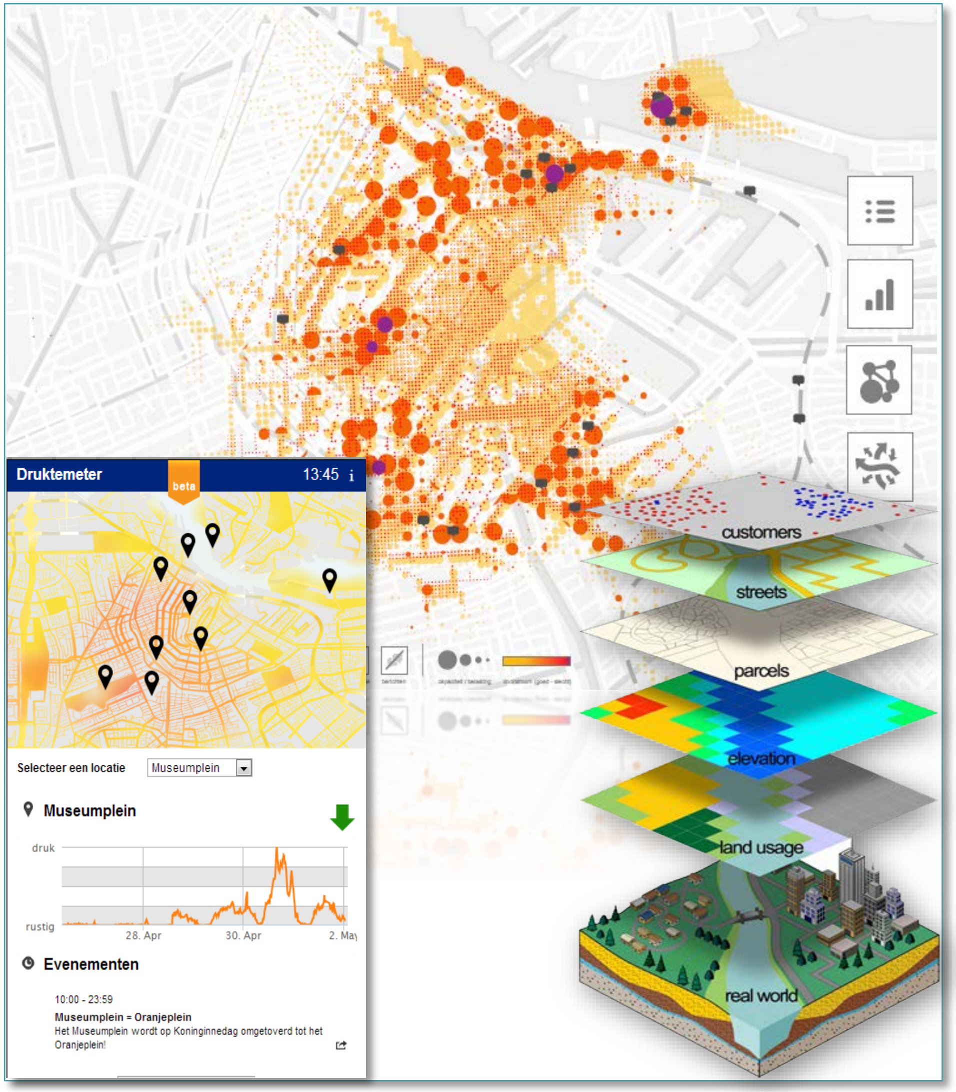
- Data sources: aggregated transaction data of mobile phones, Tweets including GPS coordinates
- Duration (develop phase): 5 – 10 weeks

Solution

Using open standards for Big Data Analytics, we have developed a base solution that is scalable to millions of mobile phone users, tweets and other data points. An intuitive user interface is developed, which is important to represent analysis results to the public.

Benefits

By combining crowd and sentiment monitoring, valuable real-time information is available for crowd control and safety purposes. This information can be used to prevent hazardous situations and instantly respond to incidents. The strength of this tool is its flexibility to adapt to different types of use; e.g. emergency services or taxi companies.



Druktemeter beta 13:45

Selecteer een locatie: Museumplein

Museumplein

Evenementen

10:00 - 23:59

Museumplein = Oranjeplein

Het Museumplein wordt op Koninginnedag omgetoverd tot het Oranjeplein!

Conclusions and Reflections

Conclusions

The Golden Case:

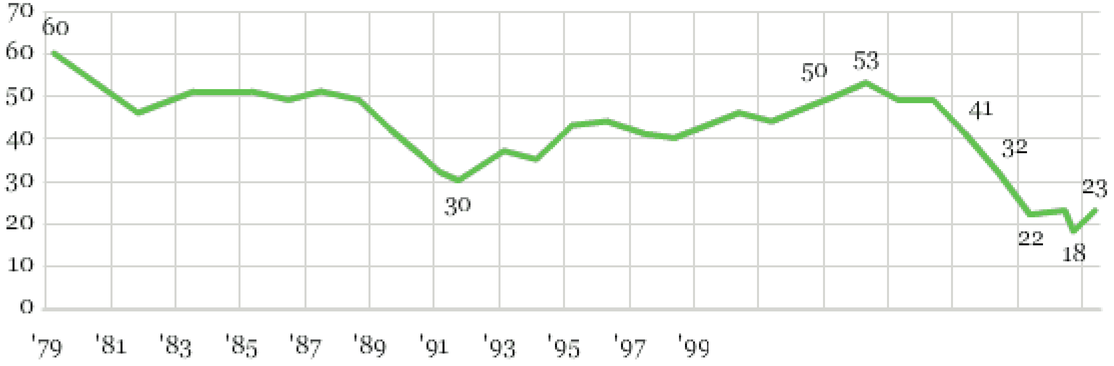
1. Adds value for society and clients.
2. Shows the potential of Big Data
3. Starts small but is meaningful.
4. Has the potential to scale fast.

Maybe trust is overrated

Confidence in Banks, 1979-2011 Trend

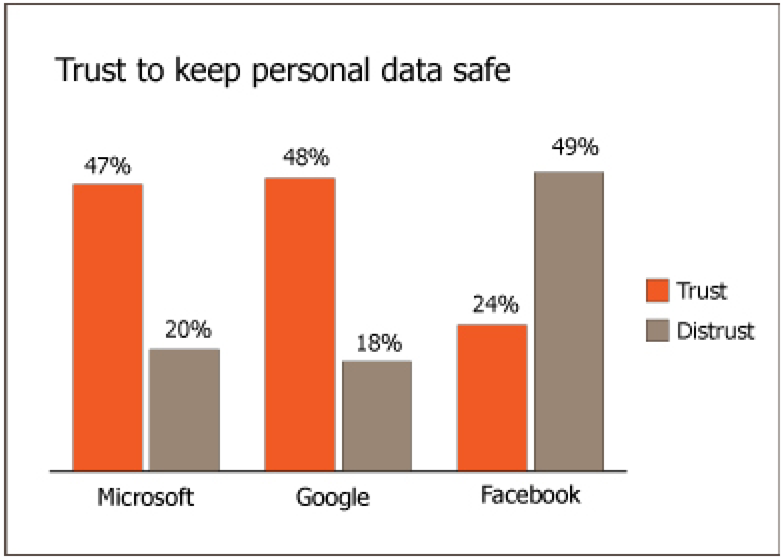
Please tell me how much confidence you, yourself, have in banks -- a great deal, quite a lot, some, or very little?

🟩 % Great deal/Quite a lot



Source: Gallup

GALLUP



Trust to keep personal data safe

Overview and insight in your own data.

That's Qiy!

Trust them to not loose our money / data

versus

Trust in them doing the right thing

cutting through complexity

Sander Klous
KPMG Management Consulting
Laan van Langerhuize 1
1186 DS, Amstelveen

Tel: +31 20 656 7186
klous.sander@kpmg.nl