DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DAS

DATA SCIENCE – UNIT 1

33965-MÁSTER UNIVERSITARIO EN INGENIERÍA Y TECNOLOGÍA DE SISTEMAS SOFTWARE

Mª José Ramírez

(based on material from José Hernández-Orallo)

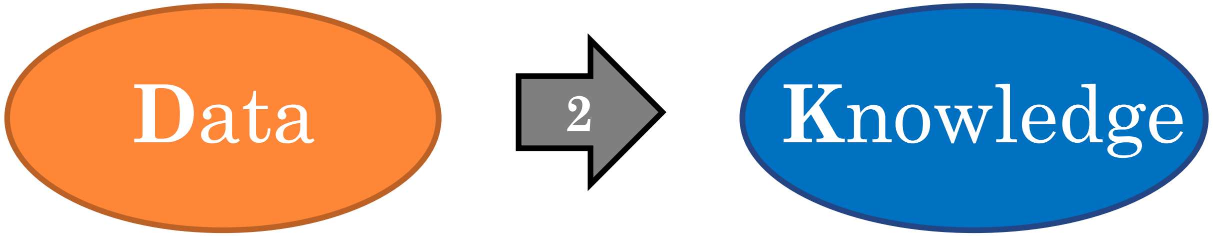
DSIC, UPV, mramirez@dsic.upv.es

DESCRIPTION

- Unit 1: Introduction
  - Data science: the data scientist role.
  - The value of data: examples
  - The D2K process (Data to Knowledge).
  - Introduction to R
- Unit 2: Data integration and manipulation
  - Types of data and repositories.
  - Data integration and cleansing.
  - Data ownership, privacy and security
  - Visualisation and data understanding.
  - Data wrangling in R. Reasoning about privacy and discrimination using R.
- Unit 3: Data analysis
  - Types of predictive and descriptive tasks
  - Supervised learning
  - Unsupervised learning
  - Model evaluation.
  - Model construction and evaluation in R.

Data Science

- Data Science:
  - "Data science is the study of the generalizable **extraction of knowledge from data**"*
  - "Data science is a set of principles that guide the **extraction of knowledge from data**"**
- A.k.a., **Data to Knowledge (D2K)**:

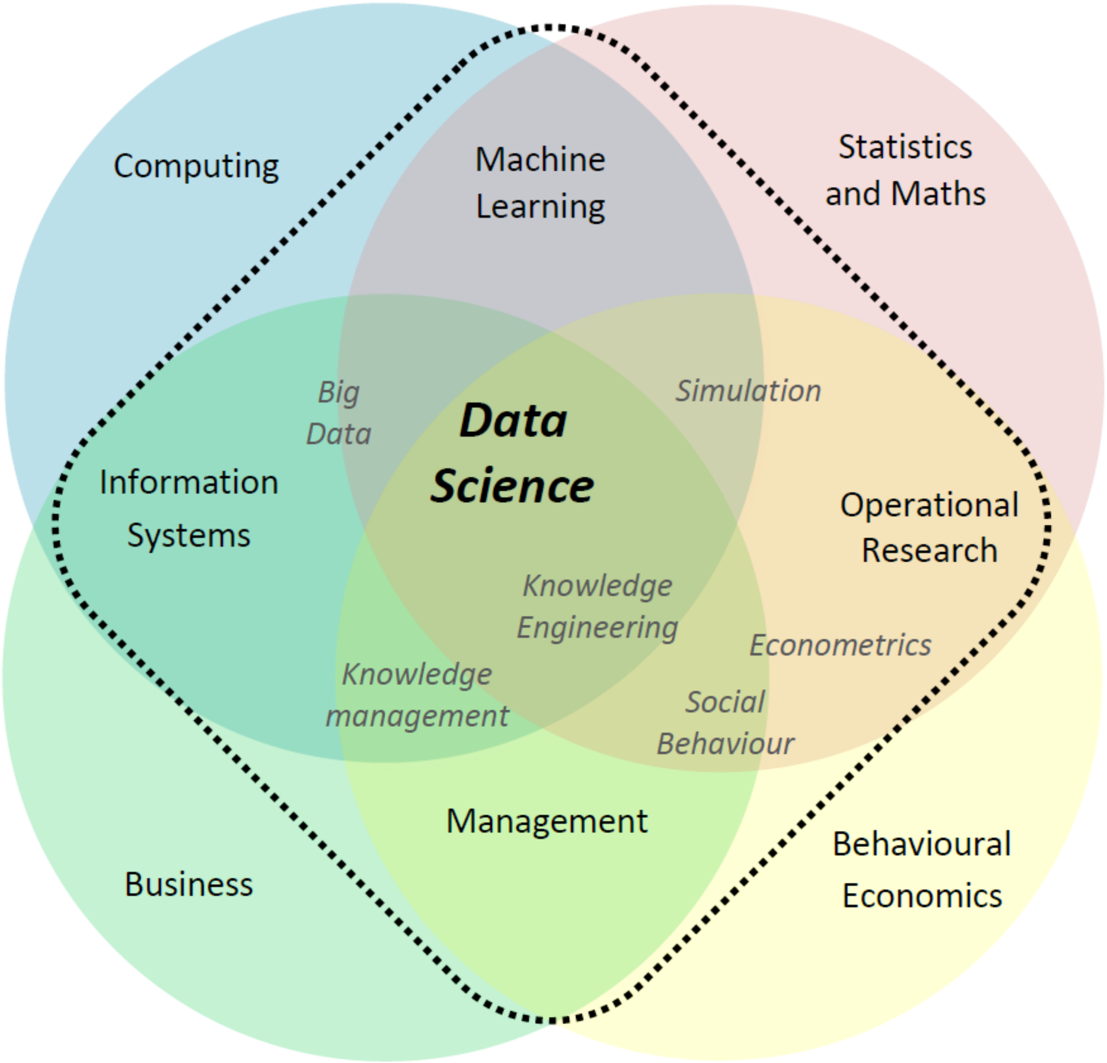


* Foster Provost and Tom Fawcett Data Science for Business: Fundamental principles of data mining and data analytic thinking, O'Reilly Media, 2013

** Communications of the ACM, Dhar, V. "Data Science and Prediction", December 2013, http://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/fulltext

DATA SCIENCE

- Data Science: a crossroads



DATA SCIENCE

- Data Science: related terms
  - Data Mining:
    - A classical term. Now seen as less general than data science.
    - Data mining is more associated to tools.
    - Data science is associated to an inquisitive profession.
  - (Intelligent) Data Analysis
    - The act of drawing an inference from some data.
    - Similar term to Data Mining, used mostly in statistics.
  - Data Analytics
    - A more fashionable term than Data Analysis
  - Big Data
    - Not all big data projects do analytics.
    - Not all data science projects require big data infrastructure.
  - Knowledge Discovery (from Databases), KDD
    - A classical term emphasising the whole process.

THE DATA SCIENTIST ROLE

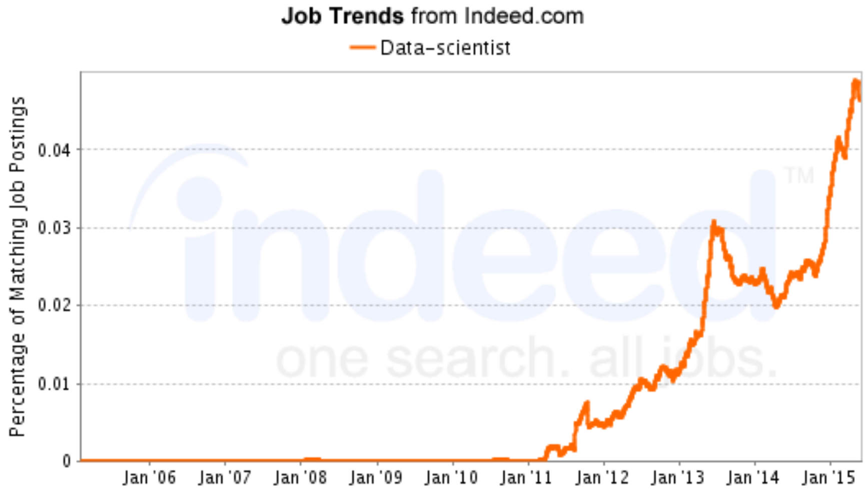
- Professionals
  - Chief Information Officer (CIO):
    - Traditional term for the most senior executive for IS & IT.
  - Data Manager
    - Traditional term for the responsible person for DB management.
  - Chief Data Officer (CDO)
    - "responsible for enterprise-wide governance and utilization of information as an asset, via data processing, analysis, data mining, information trading and other means"*
  - **Data Scientist**
    - "an integrated skill set spanning mathematics, machine learning, artificial intelligence, statistics, databases, and optimization, along with a deep understanding of the craft of problem formulation to engineer effective solutions"**.

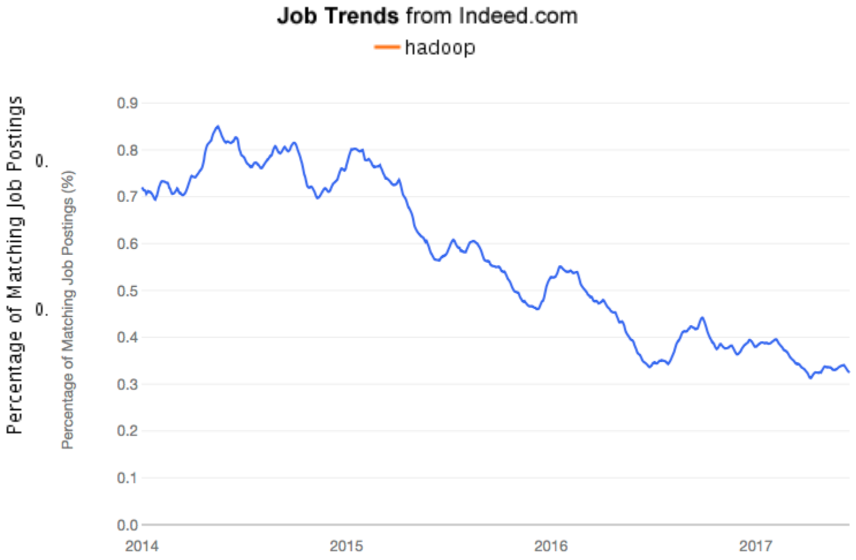
* http://en.wikipedia.org/wiki/Chief_data_officer, November 2014

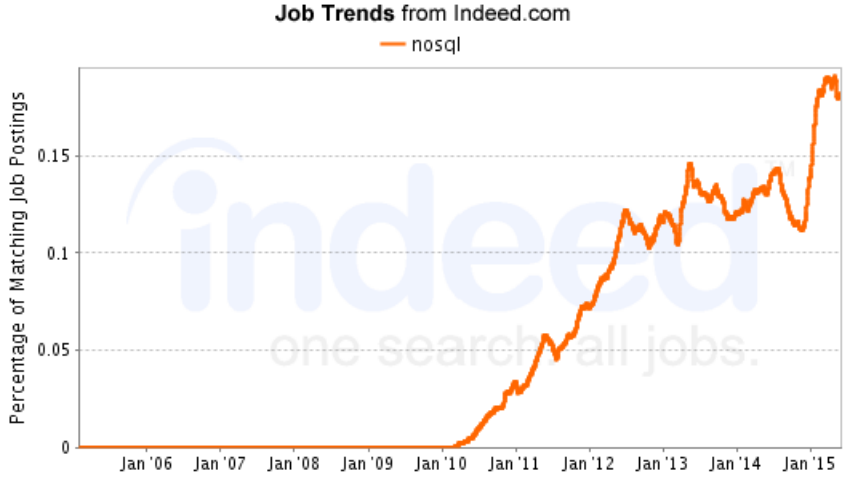
** Communications of the ACM, Dhar, V. "Data Science and Prediction", December 2013, http://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/fulltext

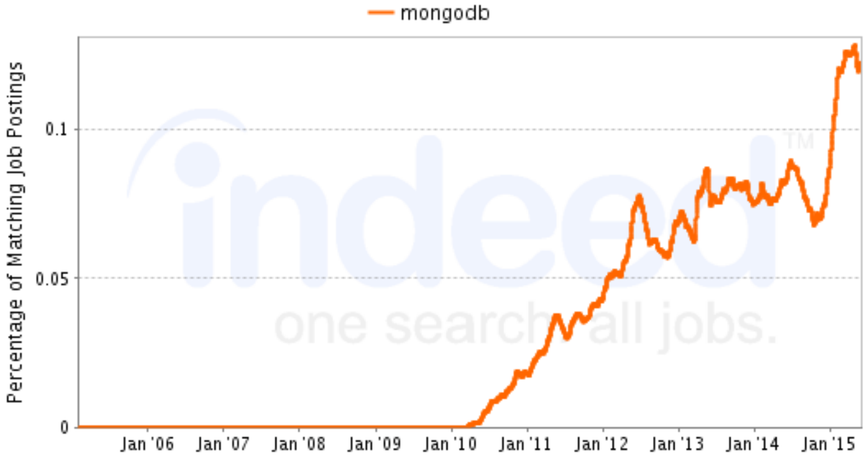
THE DATA SCIENTIST ROLE

- Professionals
  - Do facts corroborate this?


Job Trends from Indeed.com

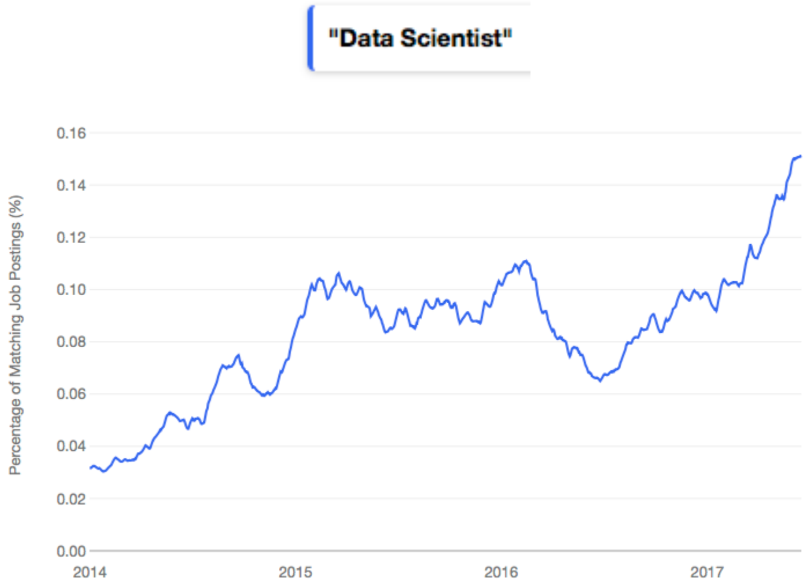

Job Trends from Indeed.com

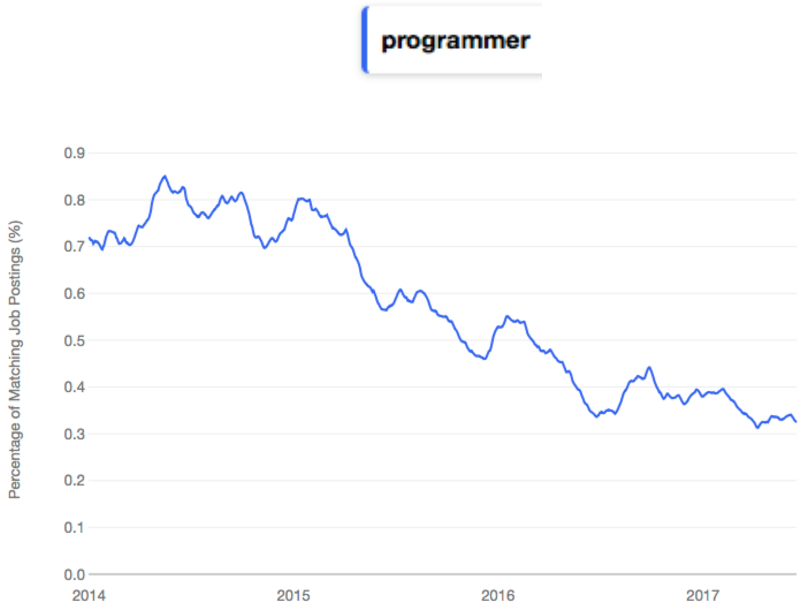

Job Trends from Indeed.com



The Data Scientist Role

- Professionals
  - Do facts corroborate this?



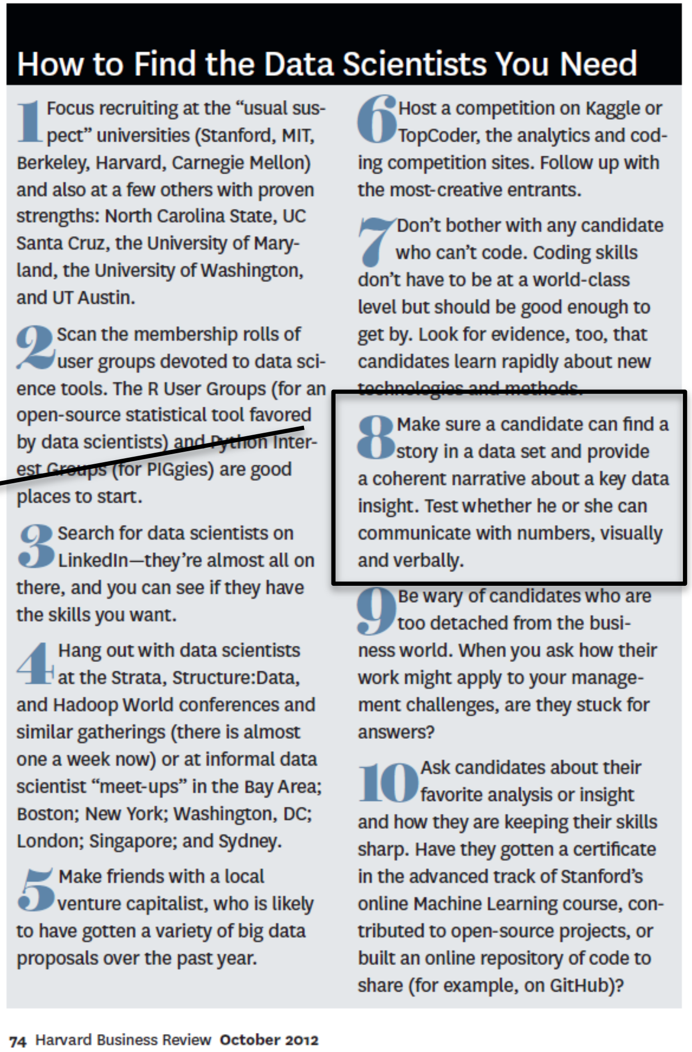


The Data Scientist Role

- Professionals
  - Know who hires and how:

"Perhaps **the most important skill a data scientist possesses**, however, **is the ability to explain the significance of data** in a way that can be easily understood by others"

from Margaret Rouse (WhatIs.com)



How to Find the Data Scientists You Need

1 Focus recruiting at the "usual suspect" universities (Stanford, MIT, Berkeley, Harvard, Carnegie Mellon) and also at a few others with proven strengths: North Carolina State, UC Santa Cruz, the University of Maryland, the University of Washington, and UT Austin.

2 Scan the membership rolls of user groups devoted to data science tools. The R User Groups (for an open-source statistical tool favored by data scientists) and Python Interest Groups (for PIGgies) are good places to start.

3 Search for data scientists on LinkedIn—they're almost all on there, and you can see if they have the skills you want.

4 Hang out with data scientists at the Strata, Structure:Data, and Hadoop World conferences and similar gatherings (there is almost one a week now) or at informal data scientist "meet-ups" in the Bay Area; Boston; New York; Washington, DC; London; Singapore; and Sydney.

5 Make friends with a local venture capitalist, who is likely to have gotten a variety of big data proposals over the past year.

6 Host a competition on Kaggle or TopCoder, the analytics and coding competition sites. Follow up with the most-creative entrants.

7 Don't bother with any candidate who can't code. Coding skills don't have to be at a world-class level but should be good enough to get by. Look for evidence, too, that candidates learn rapidly about new technologies and methods.

8 Make sure a candidate can find a story in a data set and provide a coherent narrative about a key data insight. Test whether he or she can communicate with numbers, visually and verbally.

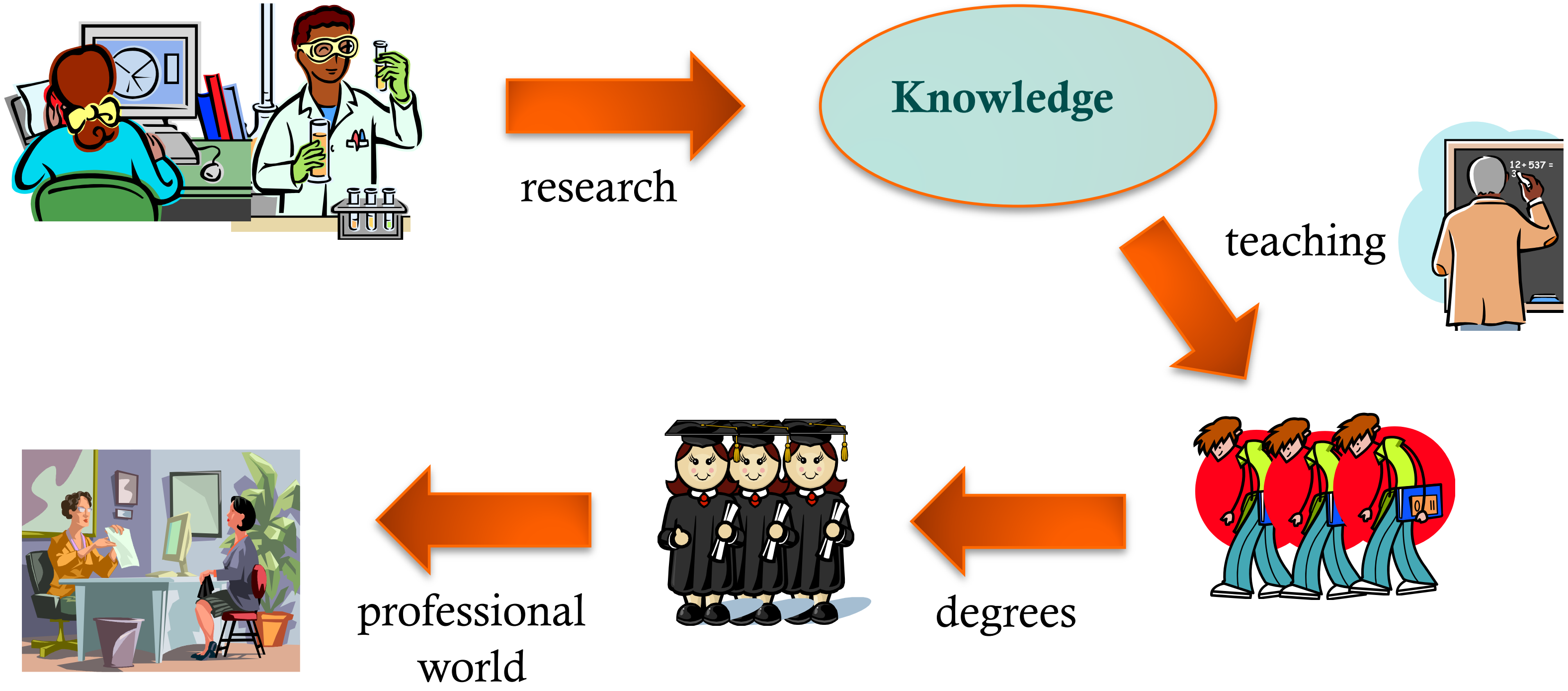
9 Be wary of candidates who are too detached from the business world. When you ask how their work might apply to your management challenges, are they stuck for answers?

10 Ask candidates about their favorite analysis or insight and how they are keeping their skills sharp. Have they gotten a certificate in the advanced track of Stanford's online Machine Learning course, contributed to open-source projects, or built an online repository of code to share (for example, on GitHub)?

74 Harvard Business Review October 2012

The Data Scientist Role

- Universities and companies
  - Overhauling how knowledge is created (and who).
    - Traditional schema:
      - Universities generate knowledge through research. Students acquire this knowledge at university. They apply this knowledge as professionals.



Knowledge

research

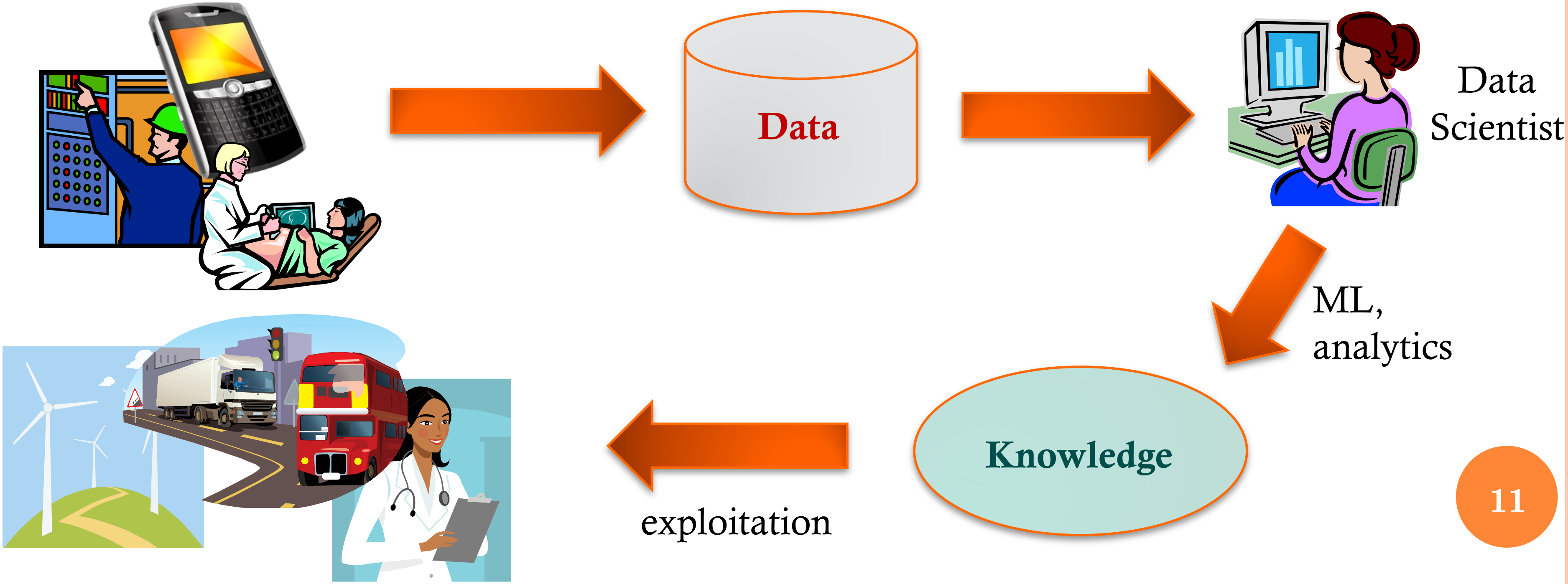
teaching

degrees

professional world

The Data Scientist Role

- Universities and companies
  - Overhauling how knowledge is created (and who).
    - New schema:
      - People, companies and organisations deal with changing phenomena. Lots of *data* are stored. *New*, *domain-specific actionable knowledge* has to be extracted and deployed



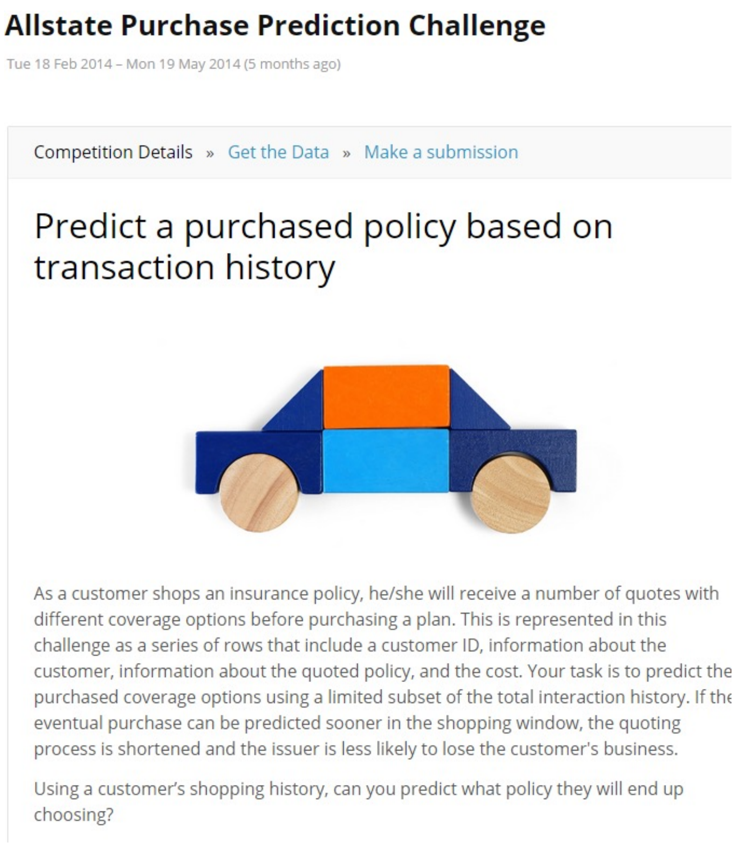
THE VALUE OF DATA

- My data is valuable for me (in → in).
  - Internal data for the organisation.
  - Classical business intelligence… Still many opportunities.
- That data is valuable for me (out → in).
  - External data for the organisation.
  - Social media, Internet, open data, … Many new opportunities.
- My data is valuable for other (in → out).
  - Internal data for other organisations.
  - My data has a value for others, … Many new opportunities.
- That data is valuable for others (out → out).
  - External data for other organisations.
  - That data has a value for others, … Freelance data scientist!
- Creating the data ($\varnothing \rightarrow$ out).
  - Collect data that may have a value. Data entrepreneur!

The Value of Data

- Examples of data-driven products (in → in):
  - A car insurance company, *Allstate* wants to predict the policy that will be purchased given the transaction history.

https://www.kaggle.com/c/allstate-purchase-prediction-challenge



Allstate Purchase Prediction Challenge

Tue 18 Feb 2014 – Mon 19 May 2014 (5 months ago)

Competition Details » Get the Data » Make a submission

Predict a purchased policy based on transaction history

As a customer shops an insurance policy, he/she will receive a number of quotes with different coverage options before purchasing a plan. This is represented in this challenge as a series of rows that include a customer ID, information about the customer, information about the quoted policy, and the cost. Your task is to predict the purchased coverage options using a limited subset of the total interaction history. If the eventual purchase can be predicted sooner in the shopping window, the quoting process is shortened and the issuer is less likely to lose the customer's business.

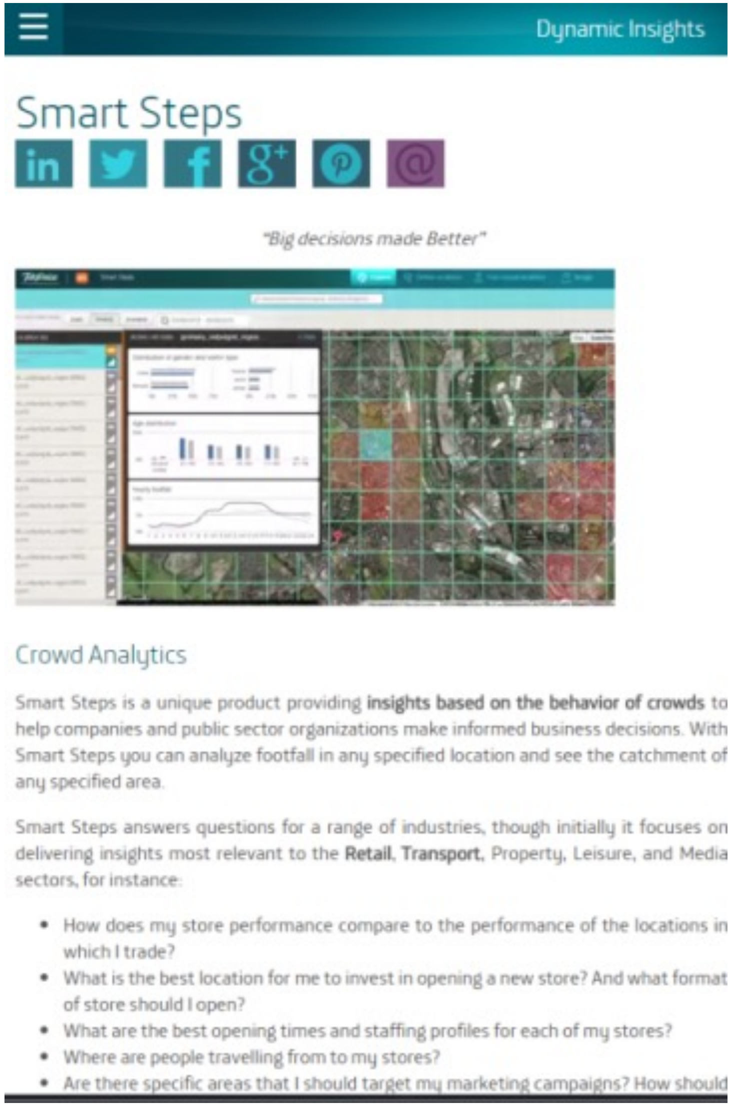
Using a customer's shopping history, can you predict what policy they will end up choosing?

The Value of Data

- Examples of data-driven products (in → out):

https://www.business-solutions.telefonica.com/en/products/big-data/business-insights/smart-steps/

- Smart Steps is real-time data gathered by Telefonica branches (Movistar, O2, …).
- They sell this data, the tools and the expertise to analyse and represent it to other companies.



Dynamic Insights

Smart Steps

"Big decisions made Better"

Crowd Analytics

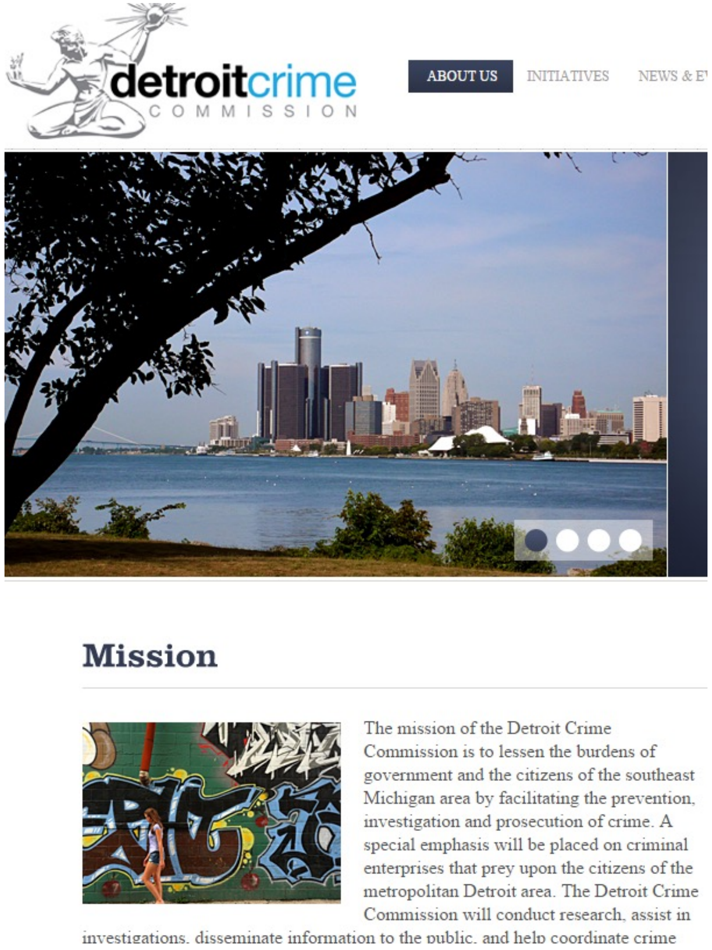
Smart Steps is a unique product providing **insights based on the behavior of crowds** to help companies and public sector organizations make informed business decisions. With Smart Steps you can analyze footfall in any specified location and see the catchment of any specified area.

Smart Steps answers questions for a range of industries, though initially it focuses on delivering insights most relevant to the **Retail**, **Transport**, Property, Leisure, and Media sectors, for instance:

- How does my store performance compare to the performance of the locations in which I trade?
- What is the best location for me to invest in opening a new store? And what format of store should I open?
- What are the best opening times and staffing profiles for each of my stores?
- Where are people travelling from to my stores?
- Are there specific areas that I should target my marketing campaigns? How should

The Value of Data

- Examples of data-driven products (out → in):
  - The Detroit Crime Commission (DCC) recognized that many criminals were posting about their crimes across various social media platforms, announcing potential plans, flaunting drugs and weapons on Facebook, Twitter, and Instagram, and organizing their next move. However, by making such information transparent to the public, the DCC decided to **take advantage of this open data** by partnering with *Semantria* to introduce text analytics that would allow the team to track criminal elements, activities, and consequences.
  - http://www.1to1media.com/weblog/2014/04/using_social_cues_to_combat_cr.html#sthash.BAlMAaOs.dpuf

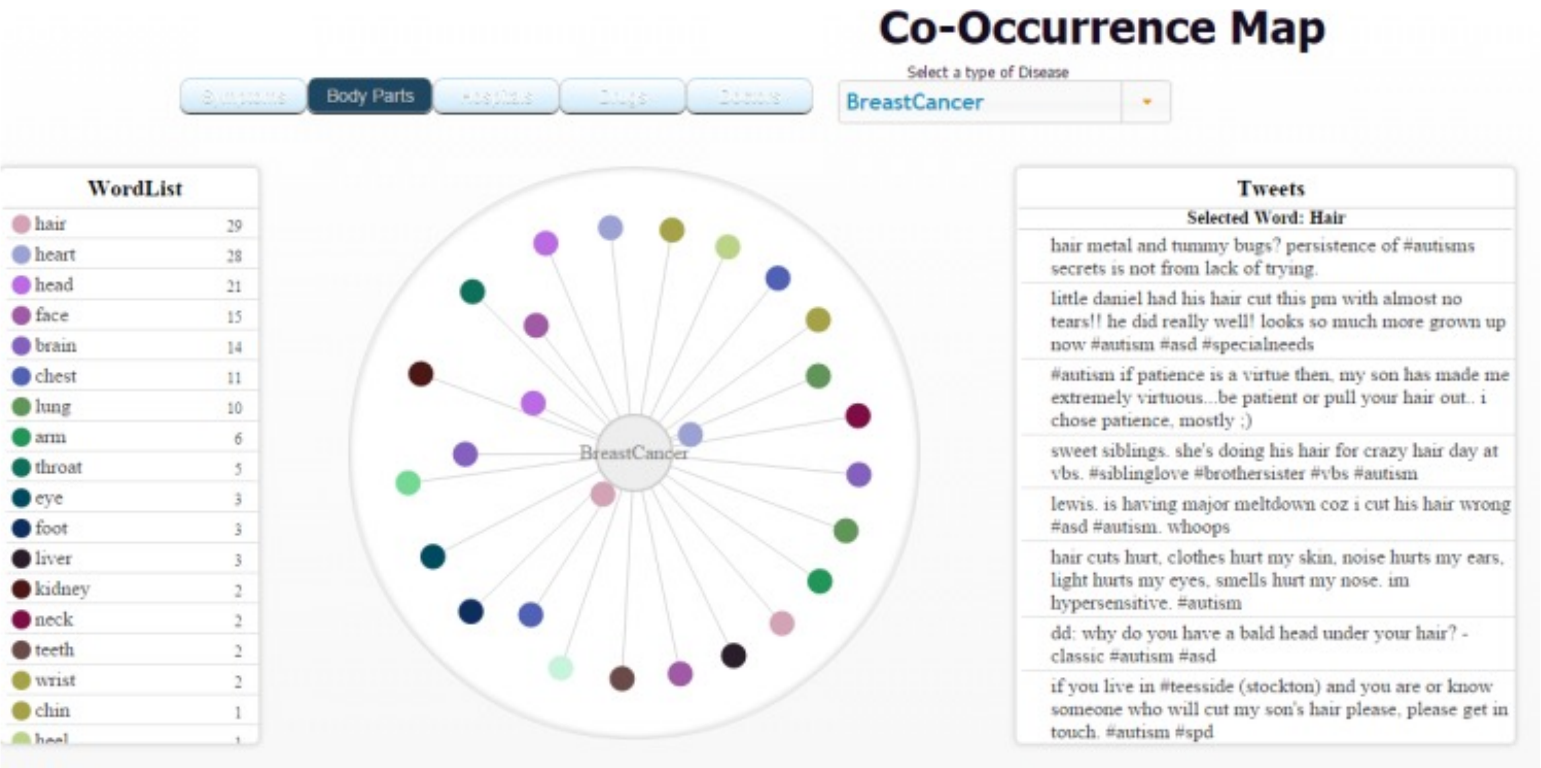


Mission

The mission of the Detroit Crime Commission is to lessen the burdens of government and the citizens of the southeast Michigan area by facilitating the prevention, investigation and prosecution of crime. A special emphasis will be placed on criminal enterprises that prey upon the citizens of the metropolitan Detroit area. The Detroit Crime Commission will conduct research, assist in investigations, disseminate information to the public, and help coordinate crime

THE VALUE OF DATA

- Examples of data-driven products (out → out):
  - *www.healthcaredataanalysis.org* was an experiment to show that tweets could give valuable information about the effect of diseases and the relation with symptoms and drugs.



Co-Occurrence Map

WordList

●	hair	29
●	heart	28
●	head	21
●	face	15
●	brain	14
●	chest	11
●	lung	10
●	arm	6
●	throat	5
●	eye	3
●	foot	3
●	liver	3
●	kidney	2
●	neck	2
●	teeth	2
●	wrist	2
●	chin	1
●	heel	1

Tweets

Selected Word: Hair

hair metal and tummy bugs? persistence of #autisms secrets is not from lack of trying.

little daniel had his hair cut this pm with almost no tears!! he did really well! looks so much more grown up now #autism #asd #specialneeds

#autism if patience is a virtue then, my son has made me extremely virtuous...be patient or pull your hair out.. i chose patience, mostly ;)

sweet siblings. she's doing his hair for crazy hair day at vbs. #siblinglove #brothersister #vbs #autism

lewis. is having major meltdown coz i cut his hair wrong #asd #autism. whoops

hair cuts hurt, clothes hurt my skin, noise hurts my ears, light hurts my eyes, smells hurt my nose. im hypersensitive. #autism

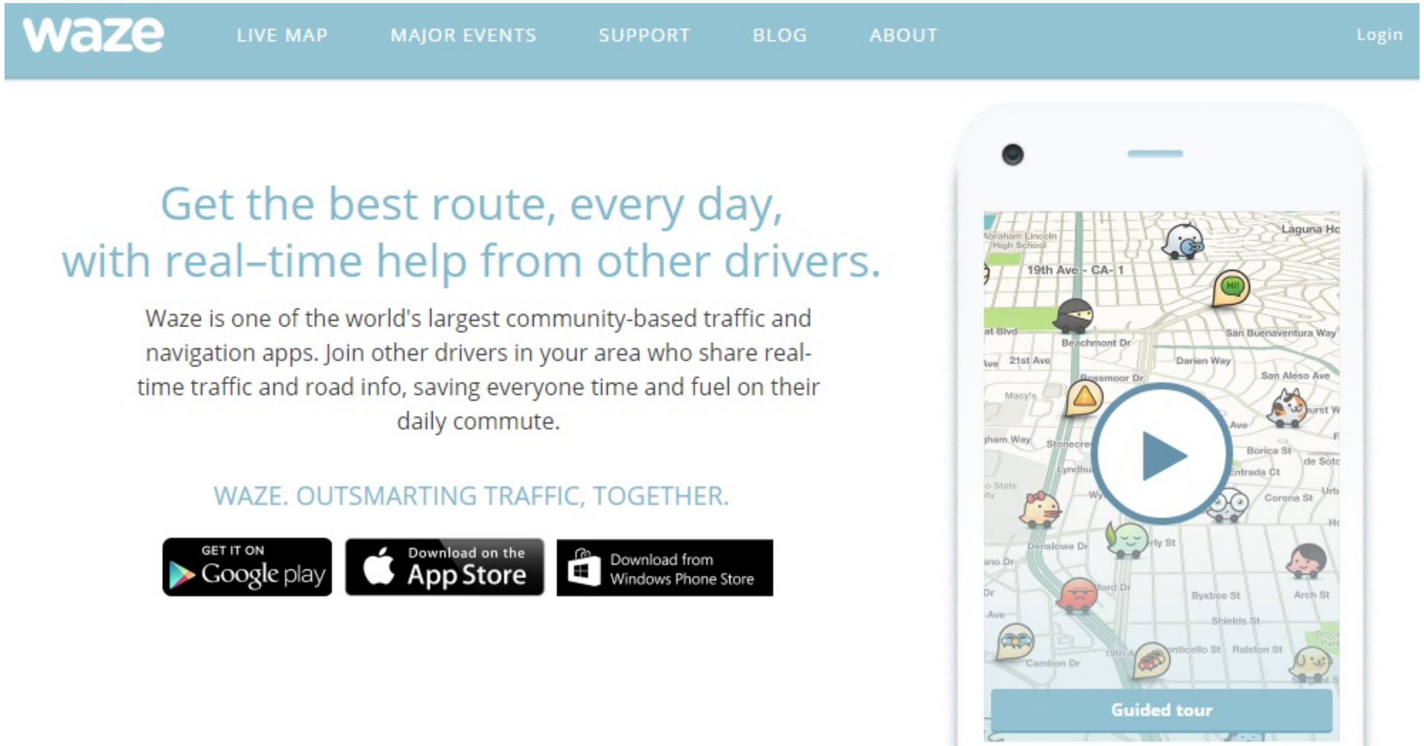
dd: why do you have a bald head under your hair? - classic #autism #asd

if you live in #teesside (stockton) and you are or know someone who will cut my son's hair please, please get in touch. #autism #spd

http://www.healthcaredataanalysis.org/demos/index.html

The Value of Data

- Examples of data-driven products (∅ → out):
  - By collecting and sharing information from drivers, an app was created to give real-time information, knowledge and advice about routes.



Get the best route, every day, with real–time help from other drivers.

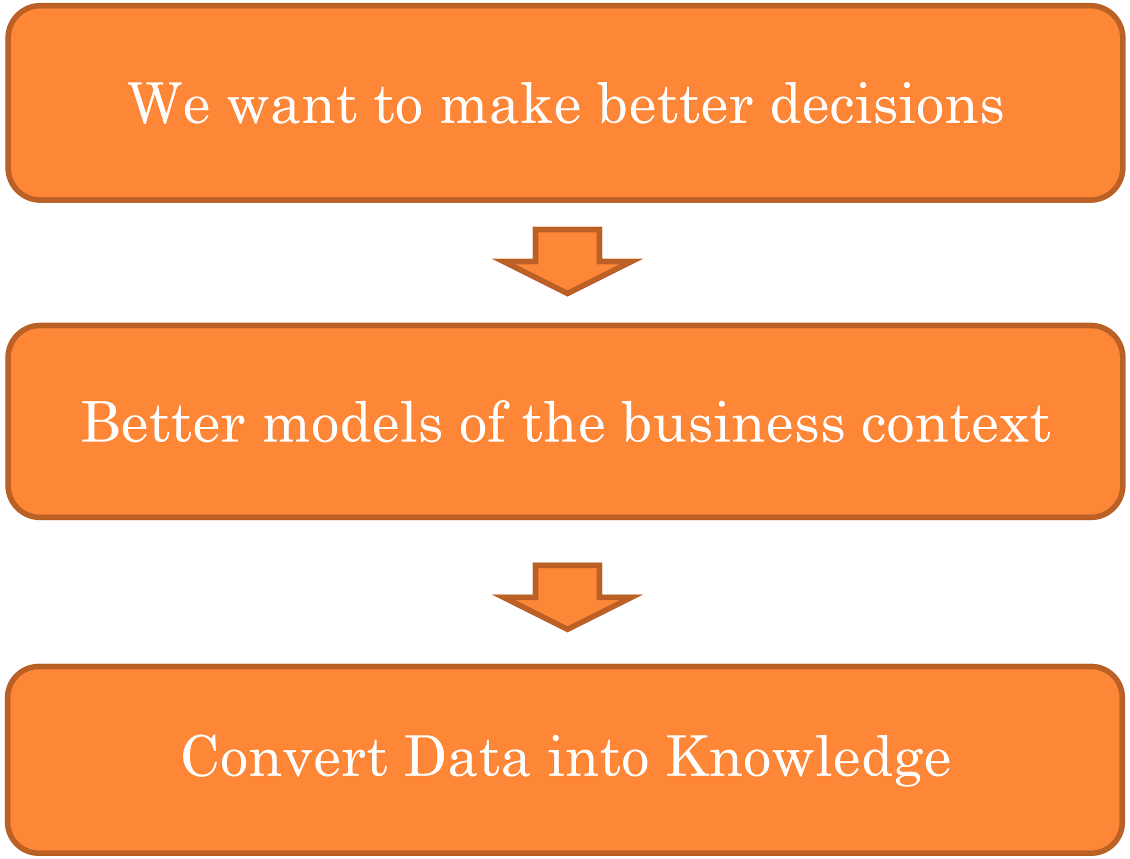
Waze is one of the world's largest community-based traffic and navigation apps. Join other drivers in your area who share real-time traffic and road info, saving everyone time and fuel on their daily commute.

WAZE. OUTSMARTING TRAFFIC, TOGETHER.

http://waze.com

The D2K Process

- What problems do we want to solve?



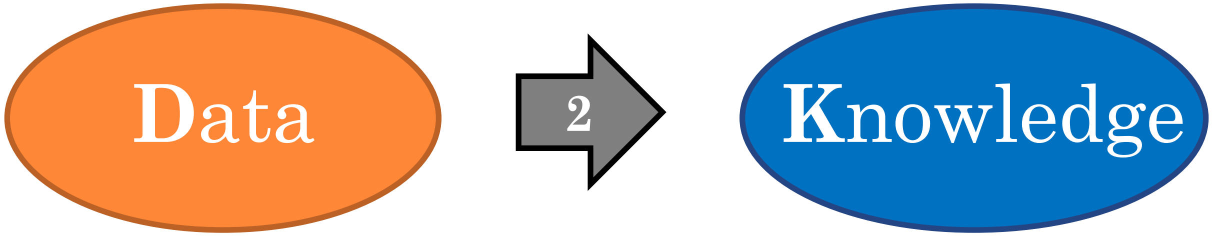
We want to make better decisions

Better models of the business context

Convert Data into Knowledge

The D2K Process

- Focus on the goal, **knowledge**, and not on the source, **data**:



- *"The extraction of actionable knowledge from the vast amounts of available digital information seems to be the natural next step in the ongoing evolution from the **Information Age** to the **Knowledge Age**"**

* United Nations ECLAC "Big Data for Development" http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2205145, January 2013.

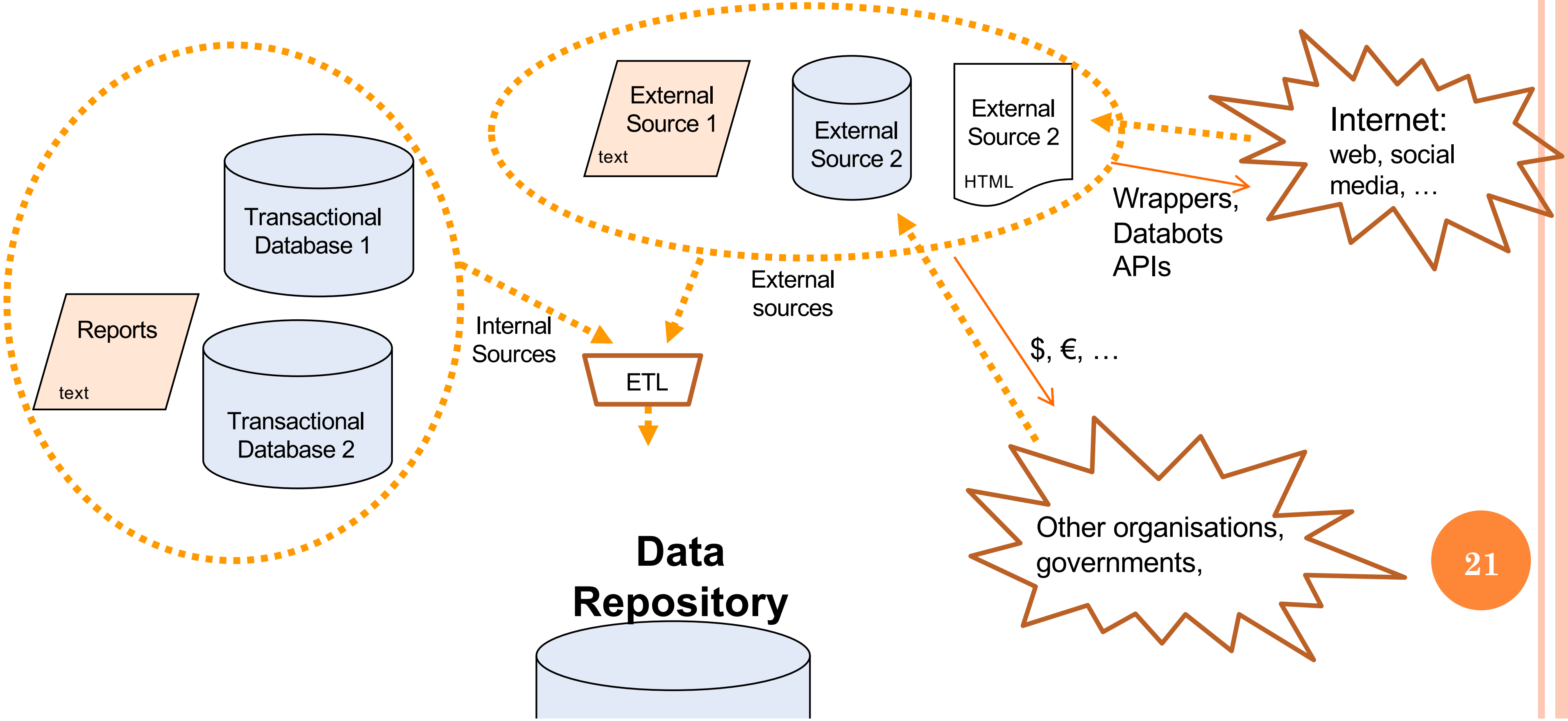
THE D2K PROCESS

- D2K
  - D: What Kind of Data?
  - K: What Kind of Knowledge?
  - 2: What Kind of Processes and Resources?

Only when these three things are examined, we can determine the viability and the technologies for a D2K problem.

The D2K Process

- D: Kinds of Data
  - Do I own the data?
    - Internal: easier, cheaper, no privacy issues.
    - External: more difficult, more expensive, privacy issues.



THE D2K PROCESS

- D: Kinds of Data
  - What does it look like?
    - Structured
      - Scalar (numerical, nominal, date, ...)
      - Non-scalar: trees, lists, graphs, ...
    - Semi-structured
      - XML, other markup languages, ..
      - Source code: programs, protocols, law, experiments, ...
      - Social media.
    - Non-structured
      - Web pages.
      - Natural language.
    - Hypermedia
    - Multimedia
    - Semantic

The D2K Process

- D: Kinds of Data
  - Who generates the data?
    - Human-generated
      - Transactions (through applications).
      - Mobile devices.
      - Social media.
      - Documents.
      - Photos, music, videos, …
    - Machine-generated
      - Sensors.
      - Logs.

THE D2K PROCESS

D: Kinds of Data

- Is it good?
  - Biased
    - Only part of the data.
  - Unbiased
    - The data is representative of the population of interest.
  - Accurate
    - Controlled data acquisition, quality control.
  - Non-accurate, missing, inconsistent, …
    - As it comes…

A real purchase from the internal database is much more reliable than a “like” in a social network.

THE D2K PROCESS

- D: Kinds of Data
  - Is the data changing?
    - Mostly static
      - Data is historical, data is assumed to be stable for a time (e.g., days, weeks or months)
        - Just refreshed from the sources periodically.
    - Stream, real-time
      - Data comes and changes very quickly (e.g., every second).

Not only the data may change, but also the structure of the data.

THE D2K PROCESS

- D: Kinds of Data
  - Is it cumulative?
    - Incremental:
      - Past data is (almost) never modified.
    - Modifiable:
      - Past data can be modified or corrected with new information.

THE D2K PROCESS

- D: Kinds of Data
  - Is it free?
    - Open
      - Anyone can have access and produce value from it
        - For their interests or those who produce the data.
    - Restricted
      - Many reasons: economic, technical, lack of transparency (governments), etc.
    - Personal
      - Privacy protection issues.

THE D2K PROCESS

- D: Kinds of Data
  - Size and complexity?
    - Small
      - Few examples and few features.
    - Big
      - Many examples and/or many features
    - Data understanding effort can change dramatically:
      - how much regular the data is.
      - whether sampling is possible.

Value		Metric
1000	kB	kilobyte
1000^2	MB	megabyte
1000^3	GB	gigabyte
1000^4	TB	terabyte
1000^5	PB	petabyte
1000^6	EB	exabyte
1000^7	ZB	zettabyte
1000^8	YB	yottabyte

- Measuring the size of data in GB, TB, … is misleading
  - Storage space can change dramatically depending on the organisation
    - Redundancy
    - Partial compression

THE D2K PROCESS

- K: Kinds of Knowledge
  - How elaborate knowledge is?
    - Simple statistical indicators
      - Means, correlations, etc.
    - Rules
      - Simple rules: e.g., propositional rules (if A then B).
    - Probabilistic
      - Knowledge with degrees of uncertainty.
    - Complex models
      - Regions are non-linear.
    - Relational, deep
      - Models relate several features and examples.
      - New features are created.
      - Models create new constructs and concepts.
      - Models are recursive.

The D2K Process

- K: Kinds of Knowledge
  - Representation?
    - Graphical
      - Visualisation
    - Declarative
      - Rules
    - Mathematical
      - Kernels, distances, weights, …

The D2K Process

- K: Kinds of Knowledge
  - Does it produce an output?
    - Descriptive
      - Helps to describe and understand the data.
    - Predictive
      - Also makes it possible to predict or estimate unknown data.

THE D2K PROCESS

- K: Kinds of Knowledge
  - Is it valid?
    - Accurate
      - Validation must be central to the use of knowledge.
    - Non-accurate
      - Models are never perfect, but they can lead to better decisions than before.
    - Reliable
      - The error of the model is bounded.
    - Unreliable
      - The error of the model is unpredictable or the model has not been well evaluated.

The D2K Process

- K: Kinds of Knowledge
  - Is it intelligible?
    - Comprehensible
      - Experts and users can understand knowledge and better revise, validate and integrate it.
    - Non-comprehensible
      - Black-box, complex models may be very accurate but less useful and inspectable.

The D2K Process

- 2: Kinds of Process and Resources
  - How are the process and the data arranged?
    - ❑ Centralised
      - ❑ Data and/or analysis
    - ❑ Distributed (data and/or process)
      - ❑ Data and/or analysis

Distribution principle: if the mountain won't come to Muhammad then Muhammad must go to the mountain.

The D2K Process

- 2: Kinds of Process and Resources
  - Where is the process and the data?
    - In-house
      - Leads to infra-utilisation (idle processors, empty storage)
      - Leads to saturation.
    - External
      - Cloud: easy to dimension depending on process and data.
      - Issues about privacy.

THE D2K PROCESS

2: Kinds of Process and Resources

- How is the analysis performed?
  - Through specific tools
    - E.g, OLAP tools, front-ends
  - Through the web
    - E.g., BigML
  - Through querying languages
    - English-like (e.g., pig)
    - SQL-like (e.g., Hive)
  - Through programming languages
    - E.g., R, Python, …
  - Through graphical suites
    - E.g., data mining suites: RapidMiner, IBM SPSS Modeler.
  - Through “cogs”
    - E.g., IBM Watson

The D2K Process

- 2: Kinds of Process and Resources
  - Who performs the analysis?
    - A single person inside the organisation
      - Small projects, no need for coordination.
    - A team in the organisation
      - Tools and process sharing, documentation.
    - Partially outsourced.
      - Coordination, cost, privacy issues, responsibilities, …
    - Completely outsourced.
      - Flexibility, control…
    - Through crowdsourcing: e.g., kaggle.
      - Cost-effective, no control, social impact, publicity, recruiting.

THE D2K PROCESS

- 2: Kinds of Process and Resources
  - How many analytic requirements do we have?
    - Is it an occasional analysis?
      - Effort in data design and organisation does not compensate.
    - Or is it a regular analysis?
      - We put more effort on repositories and tools.
  - How is the analysis driven?
    - Goal-driven?
      - We have a clear business goal to start with.
    - Data-driven?
      - We need more explorative and curious professionals.
  - Number and kinds of users:
    - Only decision makers?
    - All the organisation?
    - External users?

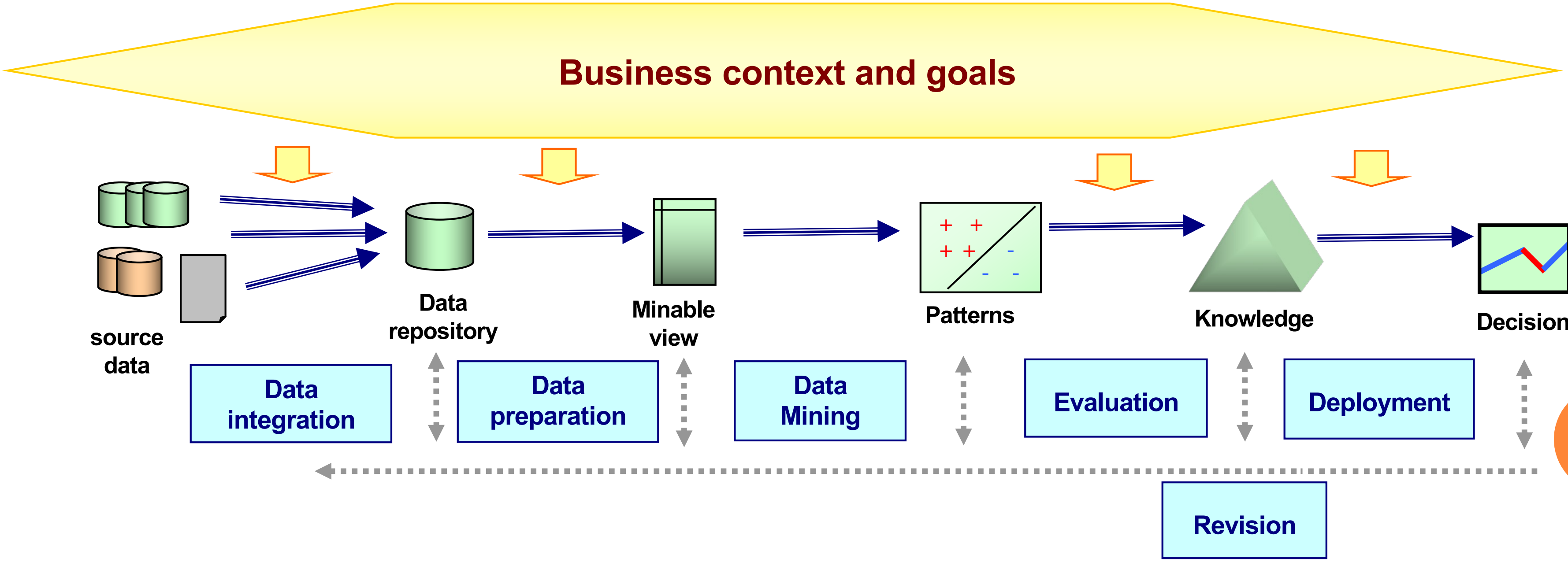
The D2K Process

- 2: Kinds of Process and Resources
  - Who is affected by the process?
    - Is it too intrusive?
      - We ask information that people are not happy to provide (forms).
      - We collect information people are not aware we are collecting.
      - Models can be used at an inconvenient moment or situation (e.g., browsing in front of your students)

Privacy, ethics, security issues (unit 2)

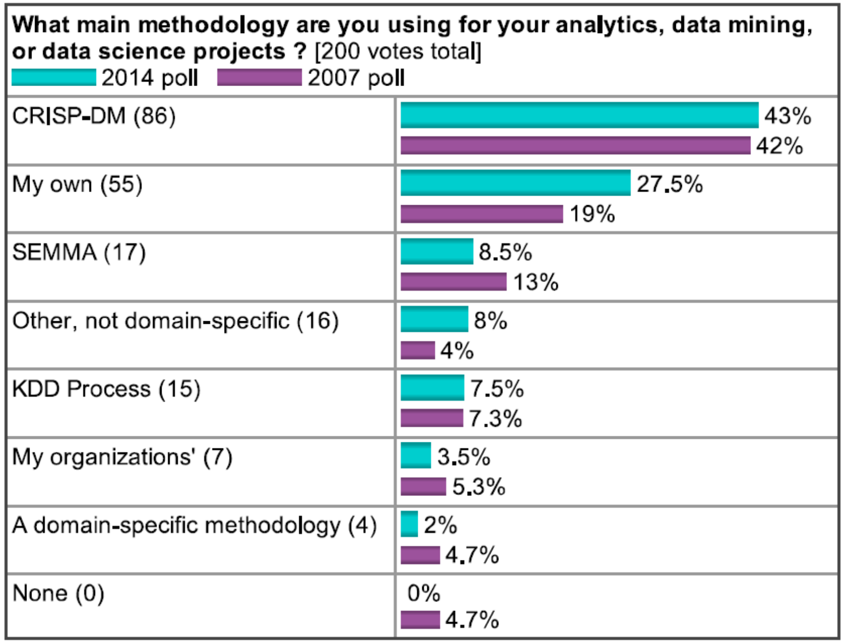
THE D2K PROCESS

- The methodology
  - Classical business intelligence process:
    - ❑ Does the business requirement require inference and patterns?
      - ❑ No, aggregate data (SQL or OLAP queries + visualisation)
        - Use your human insight to see trends and patterns.
      - ❑ Yes, use the Knowledge Discovery process:
        - Use analytical tools to get patterns and models.



The D2K Process

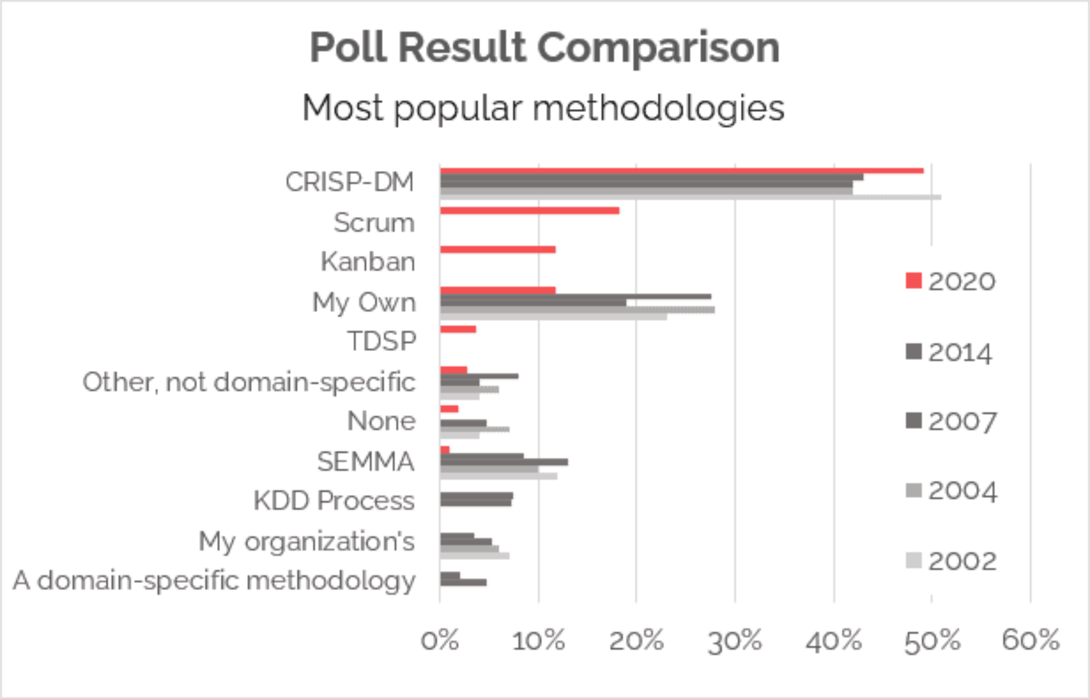
- The methodology
  - CRISP-DM (CRoss-Industry Standard Process for Data Mining)
    - ❑ An old company consortium (with funding from the European Commission), which includes SPSS, NCR and DaimlerChrysler.
  - CRISP-DM is still the most common methodology:



* From kdnuggets.com

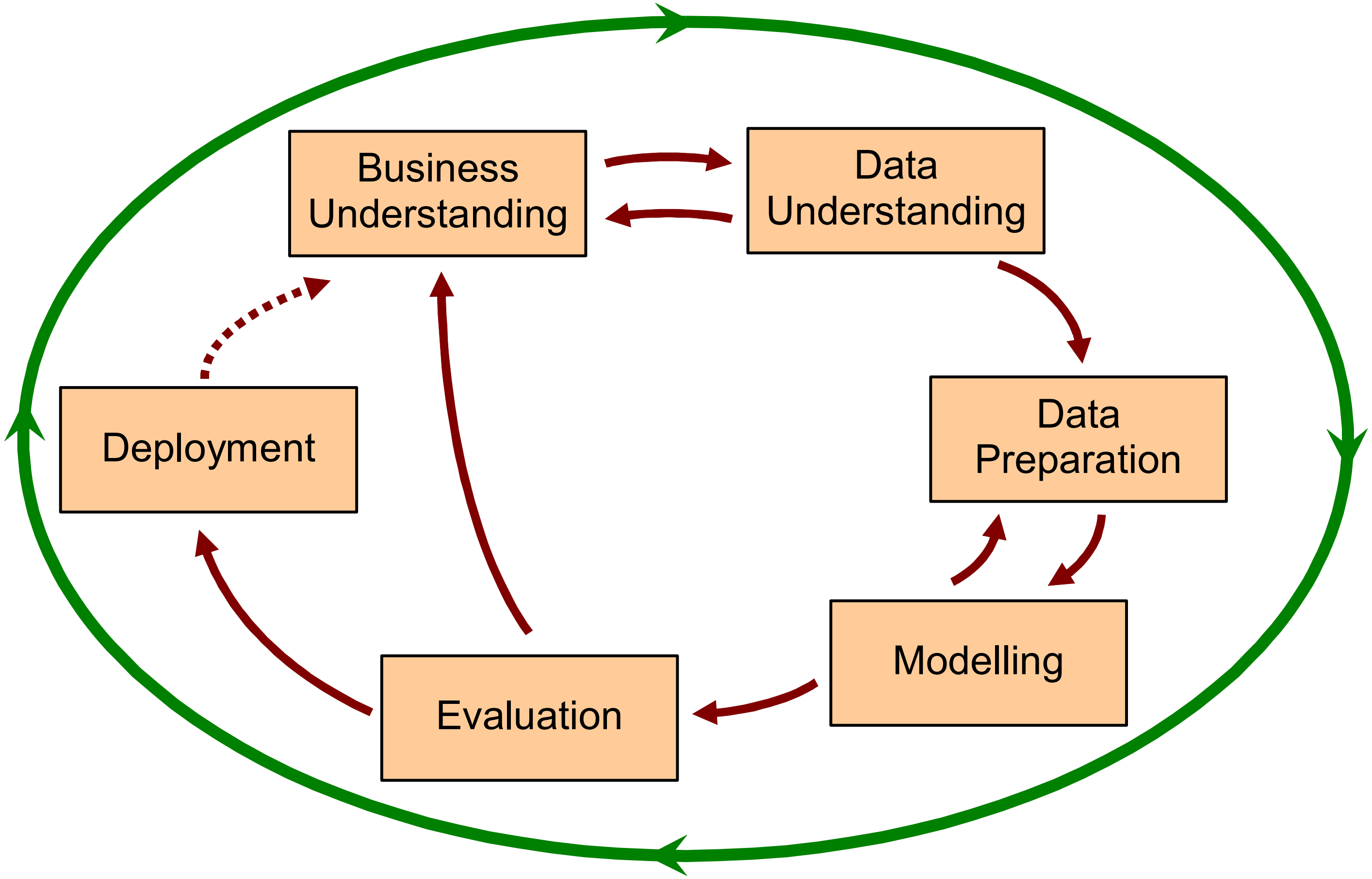
THE D2K PROCESS

- A new pull performed by Data Science Process Alliance (https://www.datascience-pm.com/) confirm that CRISP-DM is still the most commonly use methodology used by project management teams to execute their data science projects



The D2K Process

- The methodology
  - CRISP-DM



THE D2K PROCESS

- **Business Understanding**:
  - Understand the project goals and requirements from a business perspective. Substages:
    - **establishment of business objectives** (initial context, objectives and success criteria),
    - **evaluation of the situation** (resource inventory, requirements, assumptions and constraints, risks and contingences, terminology and costs and benefits),
    - **establishment of the data mining objectives** (data mining objectives and success criteria) and,
    - **generation of the project plan** (project plan and initial evaluation of tools and techniques).

The D2K Process

- **Data understanding**:
  - Collect and familiarise with data, identify the data quality problems and see the first potentialities or data subsets which might be interesting to analyse (according to the business objectives from the previous stage). Substages:
    - **initial data gathering** (gathering report),
    - **data description** (description report),
    - **data exploration** (exploration report) and
    - **data quality verification** (quality information).

THE D2K PROCESS

- **Data preparation**:
  - The goal of this stage is to obtain the “minable view”. Here we find: integration, selection, cleansing and transformation. Substages:
    - **data selection** (inclusion/exclusion reasons),
    - **data cleansing** (data cleansing report),
    - **data construction** (derived attributes, generated records),
    - **data integration** (mixed data) and
    - **data formatting** (reformatted data).

THE D2K PROCESS

- **Data modelling**:
  - It is the application of modelling techniques or data mining to the previous minable views. Substages:
    - **selection of the modelling technique** (modelling technique, modelling assumptions),
    - **evaluation design** (test design),
    - **model construction** (chosen parameters, models, model description) and
    - **model evaluation** (model measures, revision of the chosen parameters).

THE D2K PROCESS

- **Evaluation**:
  - It is necessary to evaluate (from the view point of the goal) the models of the previous stage. In other words, if the model is useful to answer some of the business requirements. Substages:
    - **result evaluation** (evaluation of the data mining results, approved models),
    - **revise the process** (process revision) and,
    - **establishment of the following steps** (list of possible actions, decisions).

THE D2K PROCESS

- **Deployment**:
  - The idea is to exploit the potential of the extracted models, integrate them in the decision-making processes of the organisation, spread reports about the extracted knowledge, etc. Substages:
    - **deployment planning** (deployment plan),
    - **monitoring and maintenance planning** (monitoring and maintenance plan),
    - **generation of the final report** (final report, final presentation) and,
    - **project revision** (documentation of the experience).

A PRACTICAL CASE: THE PROBLEM

- ***Identity Theft***

A bank with 1 million customers has all the historical data of each of them. The identity theft fraud is not a frequent problem (it annually only affects 0.1% of customers), but generates considerable losses (an average of 5000€/customer) with a total of €5 million per year. The bank wants to detect these cases as soon as possible, in order to block fraudulent transactions.

How to detect those cases?
What data do we need?

A PRACTICAL CASE: THE MODEL

- ***Option 1: to build a fraud pattern***

To analyse historical fraudulent cases (using any data science technique) in order to learn a fraud pattern or model for predicting future fraud cases.

- ***Option 2: to detect anomalous cases***

To analyse historical normal (non-fraudulent) cases in order to learn a pattern or model for "normality" and then to consider as fraudulent those behaviors that are considerably deviated from the normal behavior.

Frauds are detected in 3 days with a 70% of accuracy →

loss is reduced: $5*10^6*0,7=3,50*10^6$

A PRACTICAL CASE: THE DECISION

- Let us suppose that our models also have an accuracy of 70%.
- What is the real benefit for the bank?
  - Undetected fraud (30% of losses): 1,5*10^6 €
  - Detected fraud
    - Cost for the 3 days before the fraud is detected: 1000€/cliente => 700*10^3€
    - Cost derived from the infrastructure needed to carry out the project as well as the compensations in case of error: 1*10^6 €

3,5*10^6 € - 1,7*10^6

1,8 millions of cost saving for the bank