

Data Science (CDA)

UNIT 1: Introduction

- José Hernández Orallo, DSIC, UPV, jorallo@dsic.upv.es

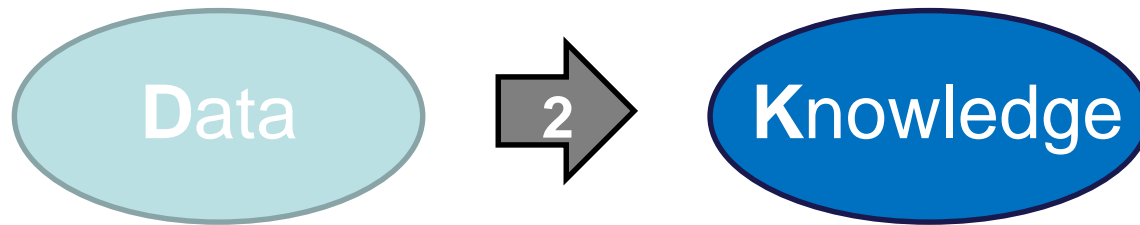


■ Unit 1: Introduction

- Data science: the role of the data scientist.
- The value of the data: examples.
- The D2K (Data to Knowledge) process.
- Big Data: challenges and solutions.



- 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)**:

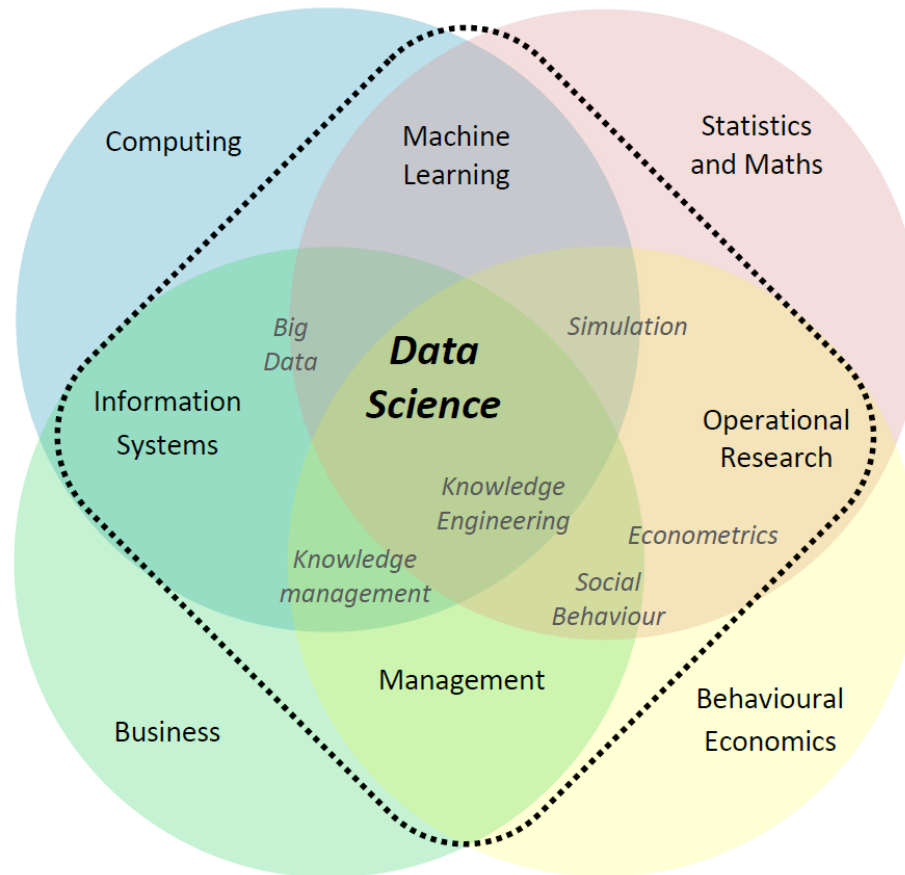


* Foster Provost and Tom Fawcett Data Science for Business: Fundamental principles of data mining and data analytic thinking, O'Reilly Media, 2013

** 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: a crossroads:



- Data Science: related terms

- Data Mining:

- A classical term. Now seen as less general than data science.
 - Data mining is more associated with tools.
 - Data science is associated with an inquisitive profession.

- (Intelligent) Data Analysis

- Similar term to Data Mining, used mostly in statistics

- Data Analytics

- A fancier term for 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.



■ 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”**.

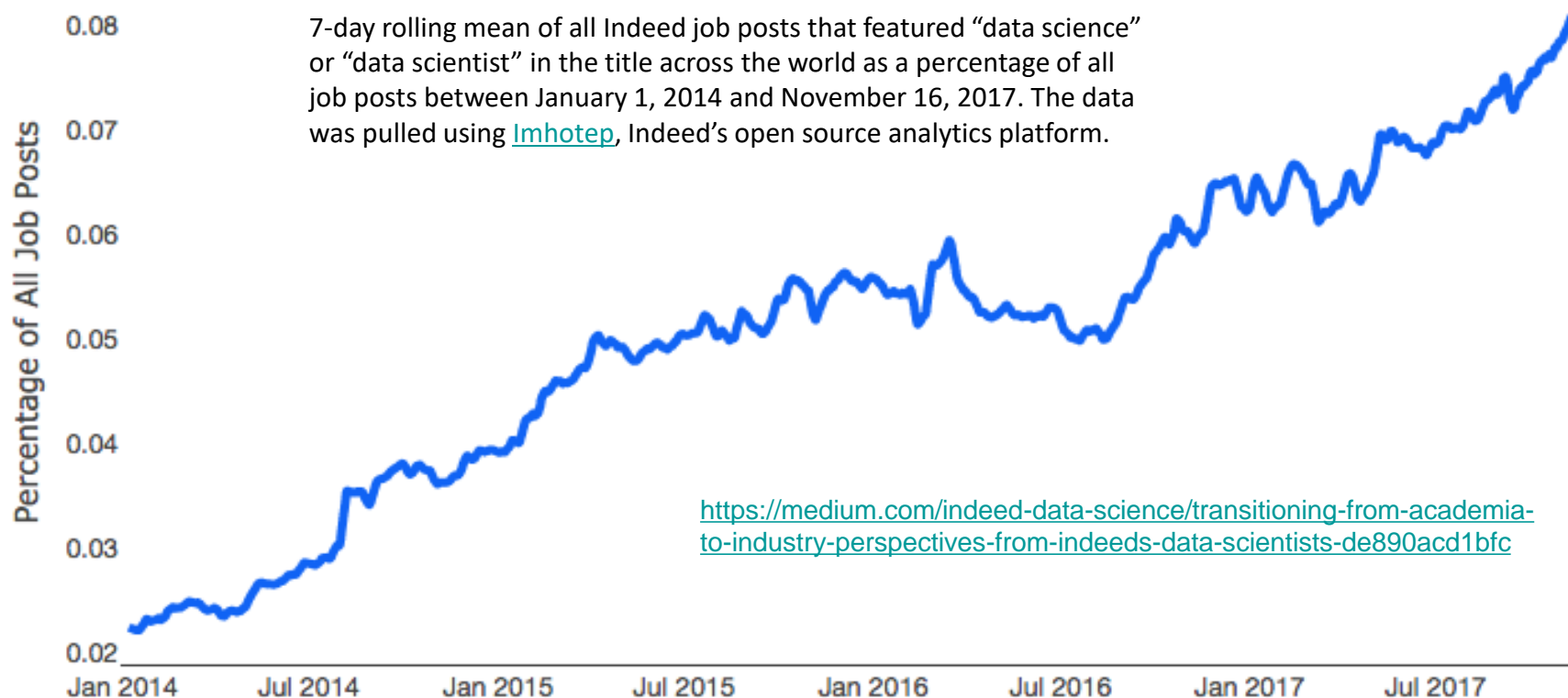
* http://en.wikipedia.org/wiki/Chief_data_officer, November 2014

** Communications of the ACM, Dhar, V. “Data Science and Prediction”, December 2013.



■ Professionals: Do facts corroborate this?

Data Science Jobs on Indeed Have Quadrupled over the last 4 years



■ Professionals (US)

31 Aug 2018 | 13:21 GMT

Desperate for Data Scientists

LinkedIn reports dramatically increasing shortage of data scientists across U.S.

By Tekla S. Perry



<https://spectrum.ieee.org/view-from-the-valley/at-work/tech-careers/desperate-for-data-scientists>

<https://economicgraph.linkedin.com/resources/linkedin-workforce-report-august-2018>

Table 2: The intensification of local shortages for data science skills, July 2015 to July 2018

	Metro Area	July 2015	July 2018	3Y Delta
1	New York City, NY	+4,132	+34,032	+29,900
2	San Francisco Bay Area, CA	+10,995	+31,798	+20,803
3	Los Angeles, CA	+425	+12,251	+11,826
4	Boston, MA	+1,667	+11,276	+9,609
5	Seattle, WA	+1,182	+9,688	+8,506
6	Chicago, IL	-1,826	+5,925	+7,751
7	Washington, D.C.	+735	+7,686	+6,951
8	Dallas-Ft. Worth, TX	-2,496	+3,641	+6,137
9	Atlanta, GA	-2,301	+3,350	+5,651
10	Austin, TX	+26	+4,949	+4,923



- Professionals (Spain):
 - Job postings “empleo.trovit.es” (2014-today).
 - Other synonyms are used as well.

Profile	06/11/2014	SQL-ratio	06/09/2016	SQL-ratio	04/09/2018	SQL-ratio	04/09/2020	SQL-ratio
"Analytics"	904	22%	1159	29%	700	34%	2468	63%
"Data Scientist" / "Data Science"	63	2%	440	11%	263	13%	1784	45%
"Análisis de datos"	195	5%	1808	46%	1364	66%	1529	39%
"Business intelligence"	1065	26%	1055	27%	562	27%	1091	28%
"NOSQL"	144	3%	264	7%	137	7%	656	17%
"Data mining"	144	3%	104	3%	58	3%	196	5%
"Hadoop"	131	3%	230	6%	123	6%	316	8%
"Big Data"	328	8%	783	20%	786	38%	1762	45%
"OLAP"	68	2%	46	1%	12	1%	32	1%
"Data warehouse"	101	2%	247	6%	72	3%	188	5%
"SQL"	4164	100%	3949	100%	2071	100%	3935	100%
"Machine learning"	64	2%	125	3%	204	10%	924	23%
"Modelización"	96	2%	116	3%	74	4%	80	2%



- Professionals
 - Know who hires and how:

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

10

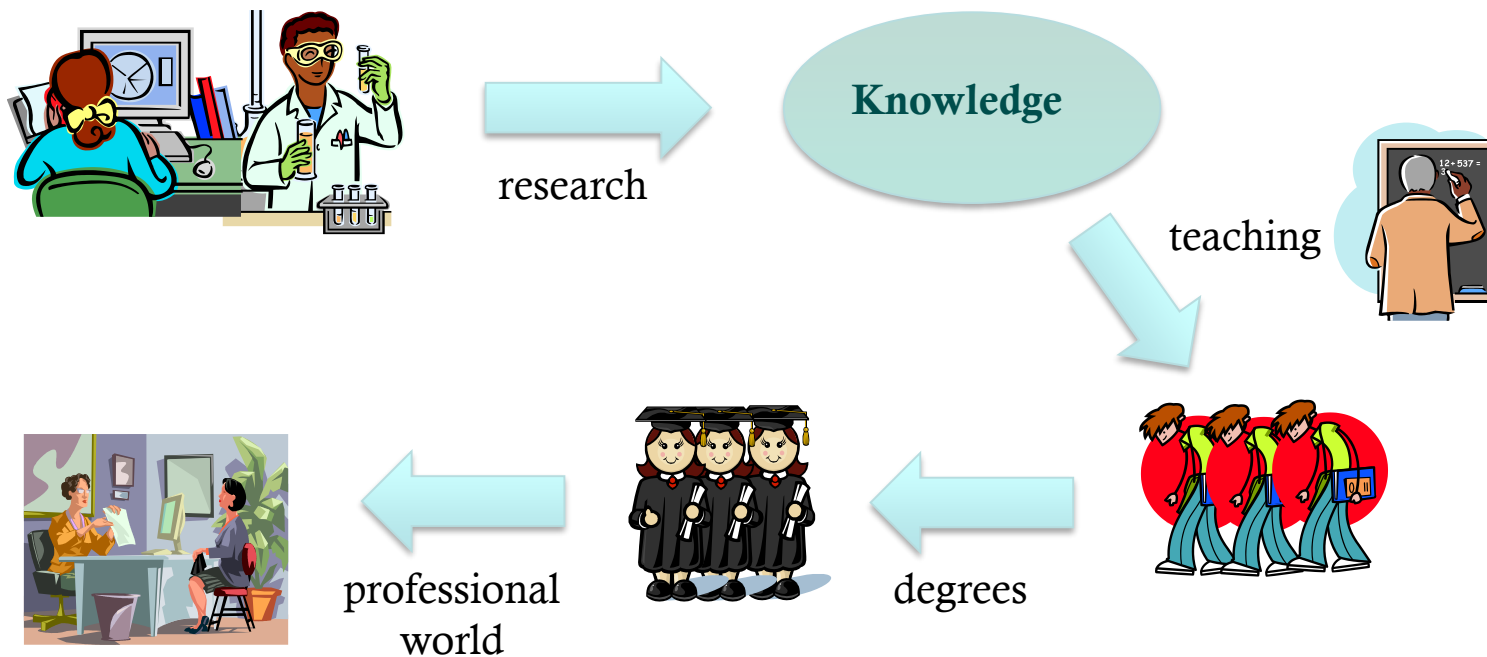


- Universities and companies

- 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.

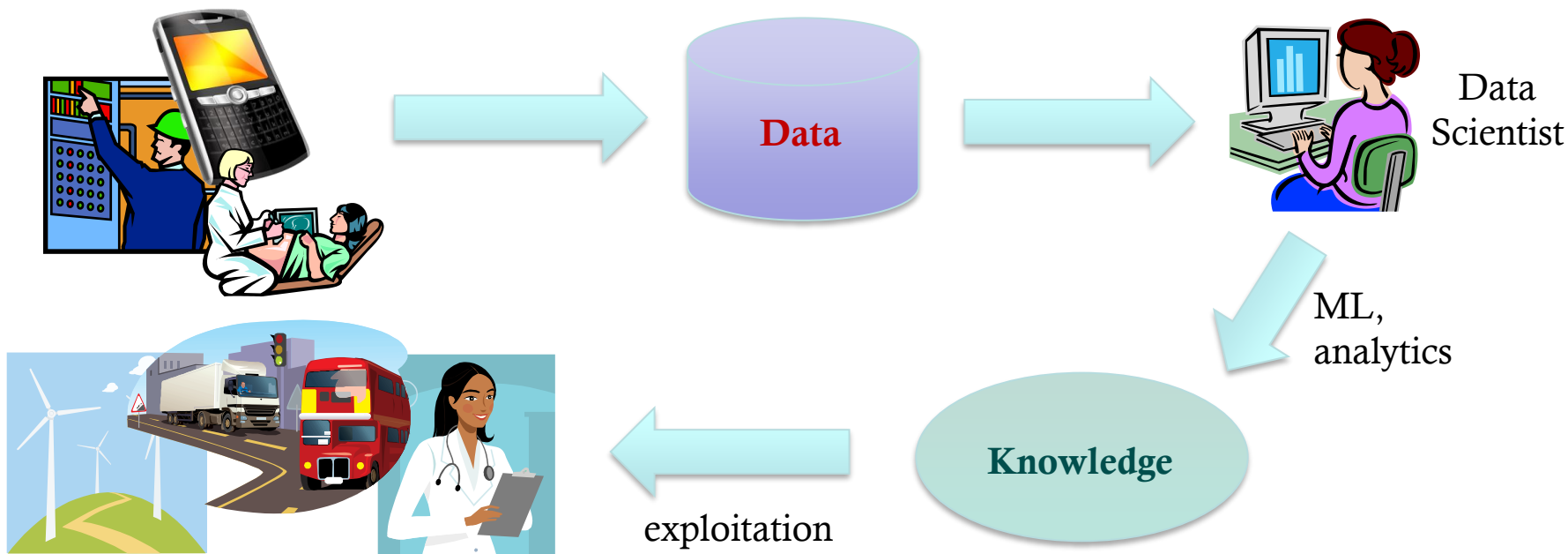


- Universities and companies

- 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



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



- 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?



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

- 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.

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



Smart Steps



"Big decisions made Better"

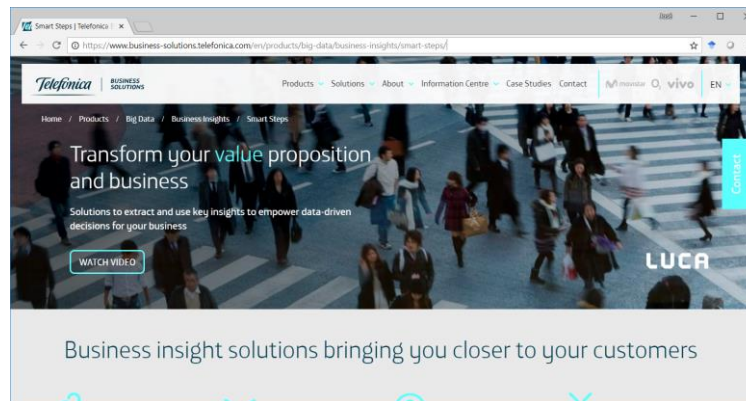


Crowd Analytics

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



■ 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.BA1MAaOs.dpuf>



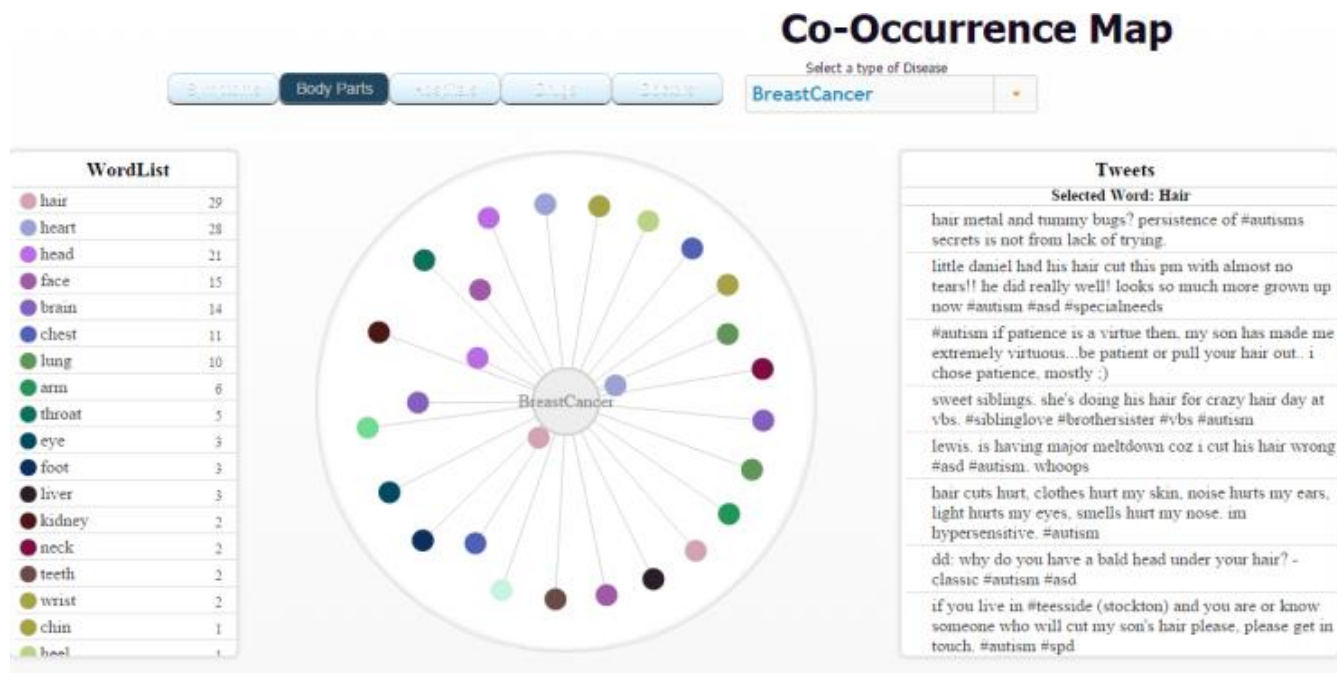
Mission



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



- 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.

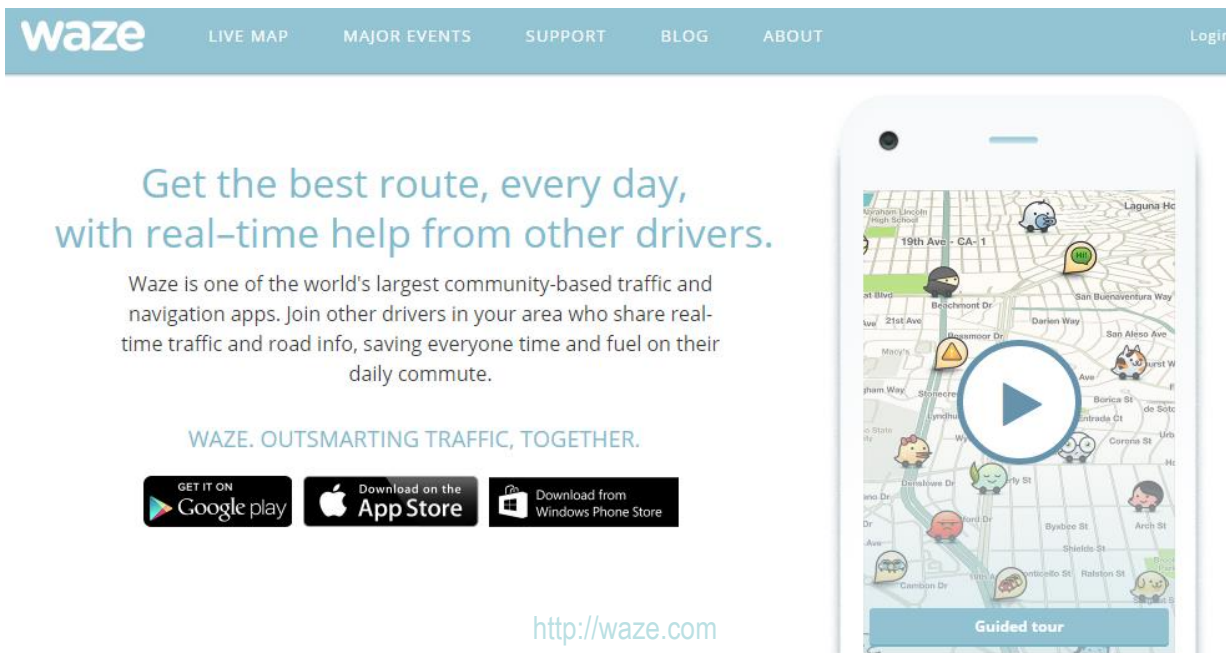


<http://www.healthcaredataanalysis.org/demos/index.html>

17



- Examples of data-driven products ($\emptyset \rightarrow \text{out}$):
 - By collecting and sharing information from drivers, an app was created to give real-time information, knowledge and advice about routes.



The image shows the Waze website and its mobile app interface. The website header includes the Waze logo and navigation links: LIVE MAP, MAJOR EVENTS, SUPPORT, BLOG, ABOUT, and a Login button. The main content area features the text: "Get the best route, every day, with real-time help from other drivers." Below this, it states: "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." Further down, it says "WAZE. OUTSMARTING TRAFFIC, TOGETHER." and provides download links for Google Play, the App Store, and the Windows Phone Store. On the right, a smartphone displays the Waze app interface, showing a map with various traffic icons and a large play button in the center. A "Guided tour" button is visible at the bottom of the app screen.

<http://waze.com>

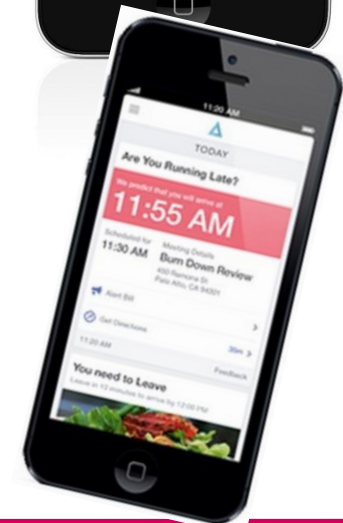
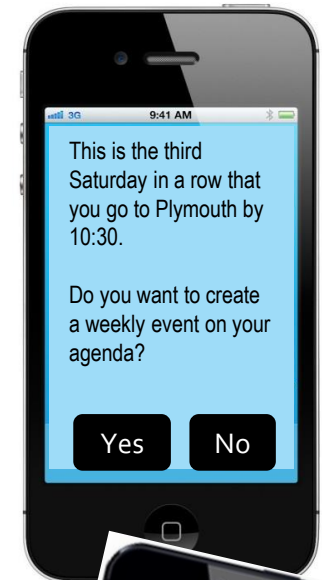


Examples of data-driven products:

Smartphone apps that mine personal data in order to anticipate a person's needs

NAME	Cue	Google Now	Osito	Tempo AI	Dark Sky
<p><i>Now dominated by the big four: Google Assistant, Siri, Cortana, and Alexa, but still not very proactive/predictive.</i></p>					
PREDICTIONS	Summarizes a person's day based on information scavenged from calendar, e-mail, and documents	Directions, traffic, and weather based on a person's location and calendar	Handles transactions like checking in for a flight or calling a cab after you land at the airport	Directions to appointments. Also sends messages if you're running late	Provides minute-by-minute weather forecasts for user's exact location

* From: Tom Simonite "With Personal Data, Predictive Apps Stay a Step Ahead", MIT Technology Review, <http://www.technologyreview.com/news/514366/with-personal-data-predictive-apps-stay-a-step-ahead/>



■ Areas and sectors

- Telecommunication data
 - Valuable for retailers, traffic, city council, police, ...
- Other geolocated data (Flickr, Instagram, Wikiloc, ...)
 - Valuable for tourism agencies, ...
- Energy consumption data
 - Valuable for TV advertising, ...
- Public transport data (bus, subway, train, taxi, traffic, ...)
 - Valuable for tourism, energy consumption, retailing...
- Social media data.
 - Valuable for almost everything, ...
- Credit card usage data
 - Valuable for retailers, city council, ...
- Police data
 - Valuable for insurance companies, estate agents, ...
- Retailing data (amazon, ebay, segundamano.es, ...)
 - Valuable for healthcare, demographics, sociology, ...
- Weather data
 - Valuable for retailers
- Web search data.
 - Valuable for almost everything



- Products and companies based on data
 - Selling data
 - Selling knowledge
- Users and customers as data producers
 - The goal of many companies (telecoms, social networks, games, etc.) is to have as many users as possible.
 - How could Minecraft be worth \$2.5 billion in 2014???
 - 18 million players.



“Facebook and Twitter along with many other ‘Digital 100’ companies, have high valuations due primarily to data assets they are committed to capturing or creating”

Foster Provost and Tom Fawcett, “Data Science and its relationship to Big Data and data-driven decision making”, *Big Data*, Volume: 1 Issue 1: February 13, 2013



- Estimating the value of data is known as “monetisation”.
 - Example, *Flutura** distinguishes five monetisation models for the analysis of the IoT (and IoS):
 - Remote Monitoring as a Service
 - E.g., Industrial device monitoring.
 - Predictive Action as a Service
 - E.g., Prediction when sensors need re-calibration.
 - Value Added Services
 - E.g., Recommendations from consumption patterns of similar customers.
 - Extreme Pricing Personalisation
 - E.g., Car measuring device (braking, acceleration, etc.) to infer driving habits to personalise the insurance price.
 - Machine Data as a Service
 - E.g., Information exchange or syndication (selling information).

* <http://blog.fluturasolutions.com/2014/09/show-me-money-5-industrial-iot.html>



■ Not everyone was so convinced...

Customer trust is put on top of everything

CincoDías

miércoles, 12 de noviembre de 2014

Inicio ▾ Mercados ▾ Empresas ▾ Economía ▾ Tecnología ▾

ESTÁ PASANDO > Test estrés banca | Operación Púnica | Caso tarjetas Caja Madrid | Reforma fiscal | F

Atraemos el Talento que necesita su organización. Talento Objetivo y Global
Descubra las soluciones digitales para su identificación + Info RAY

Supercomputador para Hacendado y Bosqueverde

Un nuevo cerebro gestionará la Mercadona del futuro

FERNANDO SANZ SÁNCHEZ DE ROJAS | 29-09-2014 09:25

f 499 t 263 in 689 18

Temas relacionados: Mercadona Juan Roig Gestión empresarial Supermercados Mercados Empresas
Establecimientos comerciales Tecnología Economía Ciencia Comercio



La caja de hormigón que alberga el supercomputador.

Son solo poco más de 60 metros cuadrados, pero sin duda es la superficie más protegida, reservada, importante y estratégica de Mercadona. En ellos están montados los servidores, modems, procesadores y discos duros que conforman el nuevo centro de mando del líder de la distribución en España.

CincoDías logró luz verde de los altos ejecutivos de la cadena para estar a su lado. Ser testigo durante unos minutos de su funcionamiento 'on line', de lo que es capaz de hacer ahora y de lo que será posible hacer con él cuando su instalación culmine dentro de unos meses.

De su importancia y de la consideración estratégica que tiene para la empresa da cuenta como se reflejó a él, de pasada, el presidente Juan Roig, con ocasión de la presentación de los resultados de 2013. Todo el mundo allí pudo escuchar estas palabras: "nosotros innovamos mucho, somos una empresa de innovación. Estamos construyendo un centro informático, y se va a hacer un cambio informático en 2015 y 2016, que será revolucionario y que inauguraremos el año que viene. Es un salto en el que nosotros podemos jugar la empresa pero yo estoy muy convencido de que va a triunfar".

El salto lo ha dado en un solar rodeado de naranjos de una localidad valenciana, que los responsables de la empresa prefieren que se obvie su nombre. Allí ha construido un caja gris, un edificio de 2.000 metros cuadrados, de aspecto muy parecido a los que albergan los reactores nucleares, diseñado por R Studio Arquitectura. El objetivo de las toneladas de hormigón armado, que forman sus paredes y de las medidas de seguridad desplegadas, es proteger su cerebro.

El salto lo ha dado en un solar rodeado de naranjos de una localidad valenciana, que los responsables de la empresa prefieren que se obvie su nombre. Allí ha construido un caja gris, un edificio de 2.000 metros cuadrados, de aspecto muy parecido a los que albergan los reactores nucleares, diseñado por R Studio Arquitectura. El objetivo de las toneladas de hormigón armado, que forman sus paredes y de las medidas de seguridad desplegadas, es proteger su cerebro.

Un nuevo cerebro artificial, instalado en una sala de 300 metros cuadrados, que controla en tiempo real, las 24 horas del día, todas las variables de las 1.500 tiendas, los bloques logísticos y sus almacenes. Los ejecutivos de Mercadona saben ya, gracias a su nuevo CPD (así han bautizado al superordenador), cuanto, cómo y a que precio esta vendiendo en cada una de sus cajas, segundo a segundo.

Los datos no se usarán

Tener a tu disposición un caudal de información tan importante. Disponer de un termómetro diario de como evoluciona el consumo doméstico en España, da acceso a conocer mucho del gasto de cada tipo de cliente. Algo así podría no tener precio en un momento en el que las grandes cadenas de distribución luchan por cada céntimo de consumo. Que un ordenador de estas características pueda almacenar todas las acciones de compra de sus clientes, hace factible que la compañía se incline a hacer descuentos ad-hoc incentivando la fidelidad. Sin embargo la empresa asegura que no utilizará lo que puede llegar a conocer. "En Mercadona los clientes no son productos. Para conocer sus preferencias y hábitos preferimos la relación directa y para ello invertimos más de 15 millones al año a través de nuestros monitores en tienda y centros de innovación. Preferimos esta forma de interactuar con nuestros clientes que estar monitorizando todas sus acciones".

Su complejidad técnica, no apta para la comprensión de los neófitos, se adivina cuando su potencialidad se traduce en cifras. La empresa ha presupuestado para su desarrollo, desde 2012, año en el que se puso en marcha su diseño, hasta 2015, ejercicio en el que está prevista que acabe su instalación definitiva, 126 millones.

En su puesta en marcha inicial han trabajado 1.600 técnicos, tanto de la propia cadena como de las multinacionales informáticas, que se han encargado de su diseño y montaje. Y para hacerlo posible Mercadona se ha tenido que sentar, asociar y pactar con multinacionales de la talla de IBM, HP, Oracle Cisco, Redhat, Capgemini, Sopra, Atos, Indra, Sothis y Telefónica.

Y juntos han diseñado una revolución tecnológica en la cadena comparable a la que llevó a cabo en 1982, cuando decidió ser la primera empresa de distribución en España que incorporaba algo tan habitual ahora como el código de barras y los escáneres.

2019: <https://es.linkedin.com/jobs/view/business-analyst-mercadona-online-at-mercadona-1072026281>

23



- What problems do we want to solve?

We want to make better decisions



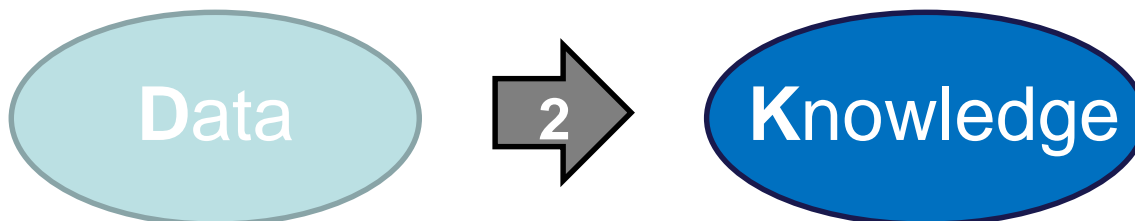
Better models of the business context



Convert Data into Knowledge



- 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.



■ D2K

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

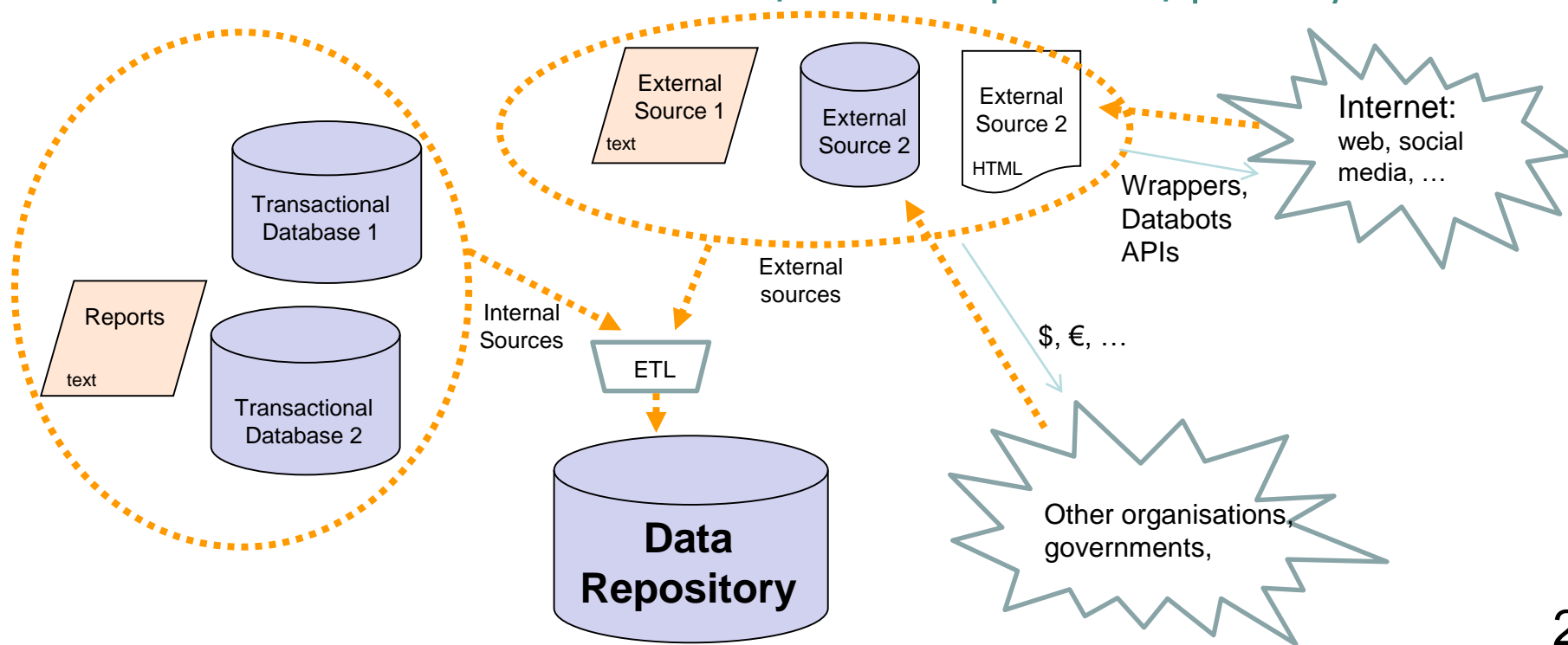
Only when these three things are examined, can we determine the viability and the technologies for a D2K problem.



■ D: Kinds of Data

○ Do I own the data?

- Internal: easier, cheaper, fewer privacy issues.
- External: more difficult, more expensive, privacy issues.



■ 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



■ 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
- Cognitive generators (e.g., GANs).



■ 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.



■ 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 may the data change, but also the structure of the data.



■ 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.



■ D: Kinds of Data

○ Is it free?

- Open

- Anyone can have access and produce value from it
 - » For their interests or for those who produce the data.

- Restricted

- Many reasons: economic, technical, lack of transparency (governments), etc.

- Personal

- Privacy protection issues.



■ 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.

○ Measuring the size of data in GB, TB, ... is misleading

- Storage space can change dramatically depending on the organisation
 - Redundancy
 - Partial compression

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



■ 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.



■ K: Kinds of Knowledge

○ Representation?

- Graphical
 - Visualisation
- Declarative
 - Rules
- Mathematical
 - Kernels, distances, weights, ...



■ 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.



■ K: Kinds of Knowledge

○ Is it valid?

- Accurate / Non-accurate

- Validation must be central to the use of knowledge.
- Models are never perfect, but they can lead to better decisions than before.

- Reliable / Unreliable

- The error of the model should be bounded and well estimated instead of unpredictable.

- Fair / Biased

- The model behaves equally beneficially for all (protected) groups of the population.



■ K: Kinds of Knowledge

○ Is it intelligible?

- Comprehensible / non-comprehensible

- Experts and users can understand knowledge and better revise, validate and integrate it.
- Black-box, complex models may be very accurate but less useful and inspectable.

- Explainable?

- Even if we can't open the model, can we still explain how it behaves?

○ Is it traceable/actionable/operational?

- Transparent?

- We should determine the provenance from predictions / descriptions to the data, and goals.

○ Is it vulnerable?

- Can we get unwanted results by changing inputs adversarially?

39



- 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.



■ 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.



- 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 numerical computation platforms
 - E.g., Keras over TensorFlow, ...
 - 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



■ 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.



- 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 pay off.
 - 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?
 - The whole organisation?
 - External users?



■ 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)

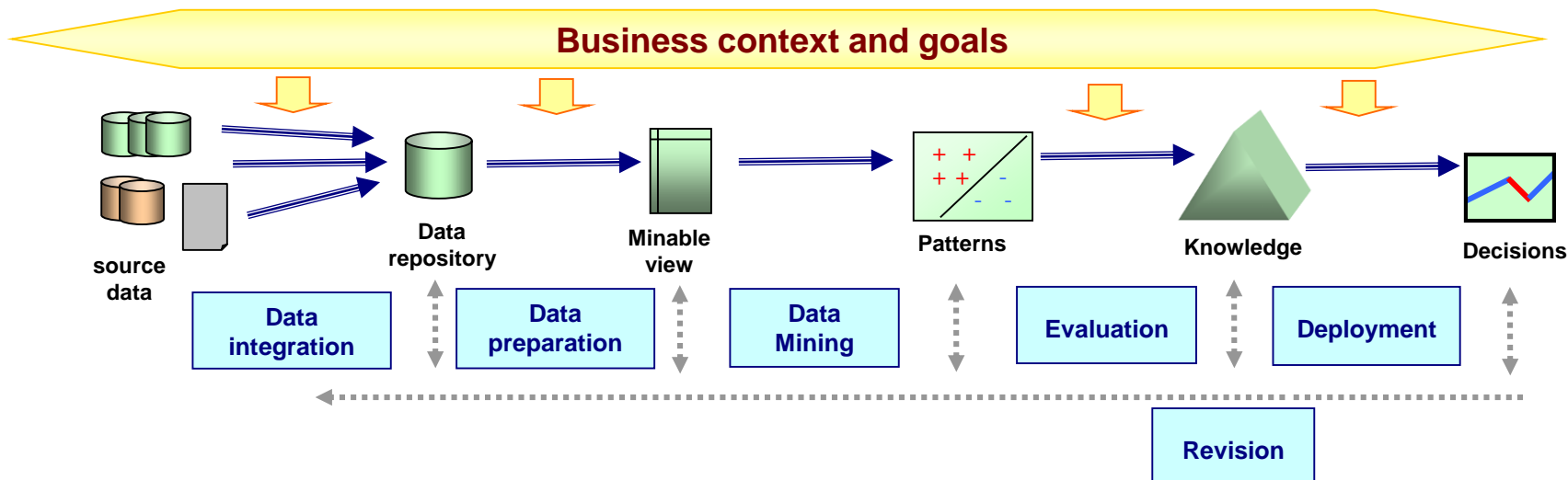
Privacy, ethics, security issues (unit 2)



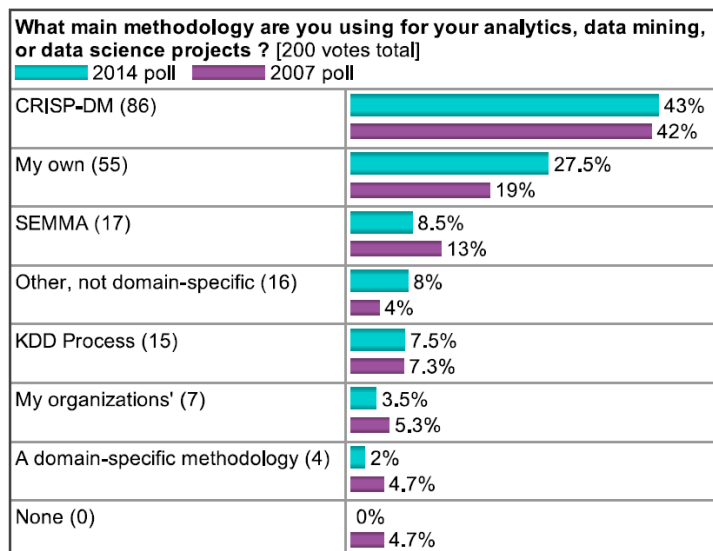
■ The methodology

○ 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.



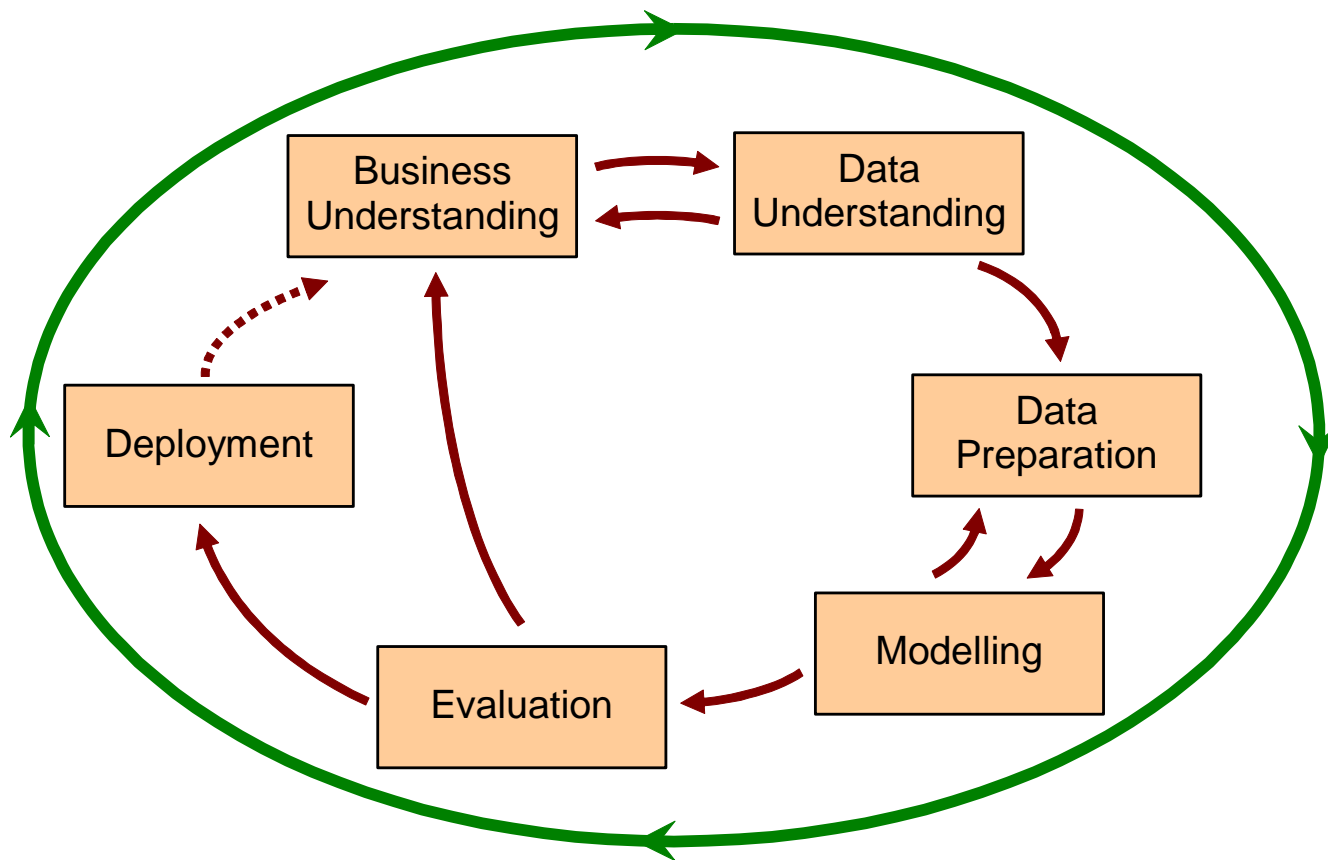
- The methodology
 - CRISP-DM (CRoss-Industry Standard Process for Data Mining)
 - An old company consortium (funded by the European Commission), including SPSS, NCR and DaimlerChrysler.
 - CRISP-DM is still the most common methodology:



* From kdnuggets.com



- The methodology
 - CRISP-DM



- **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).



- **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).



- **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).



- **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).



- **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).

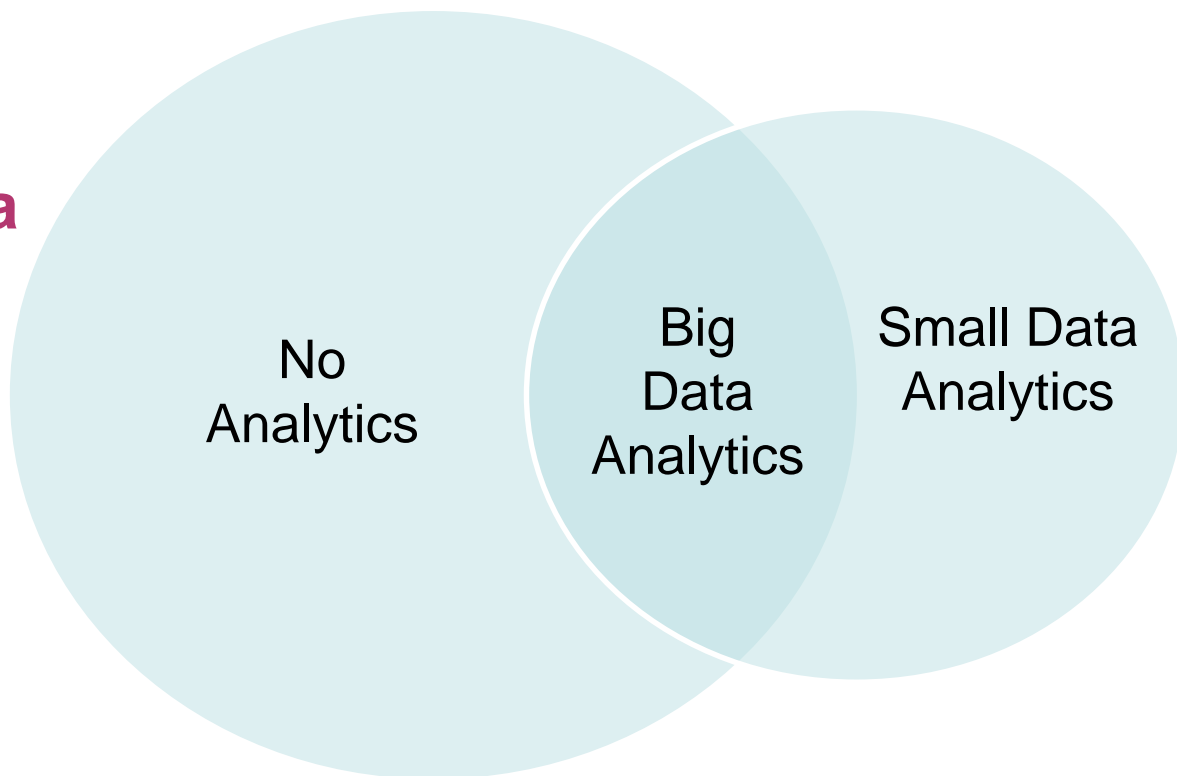


- **Deployment:**
 - The idea is to exploit the potential of the extracted models, integrate them in the decision-making processes of the organisation, spread reports about the extracted knowledge, etc. Substages:
 - **deployment planning** (deployment plan),
 - **monitoring and maintenance planning**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



- Big Data: not all data science is big data. Not all big data is data science.

**Big
Data**



***Data
Science***



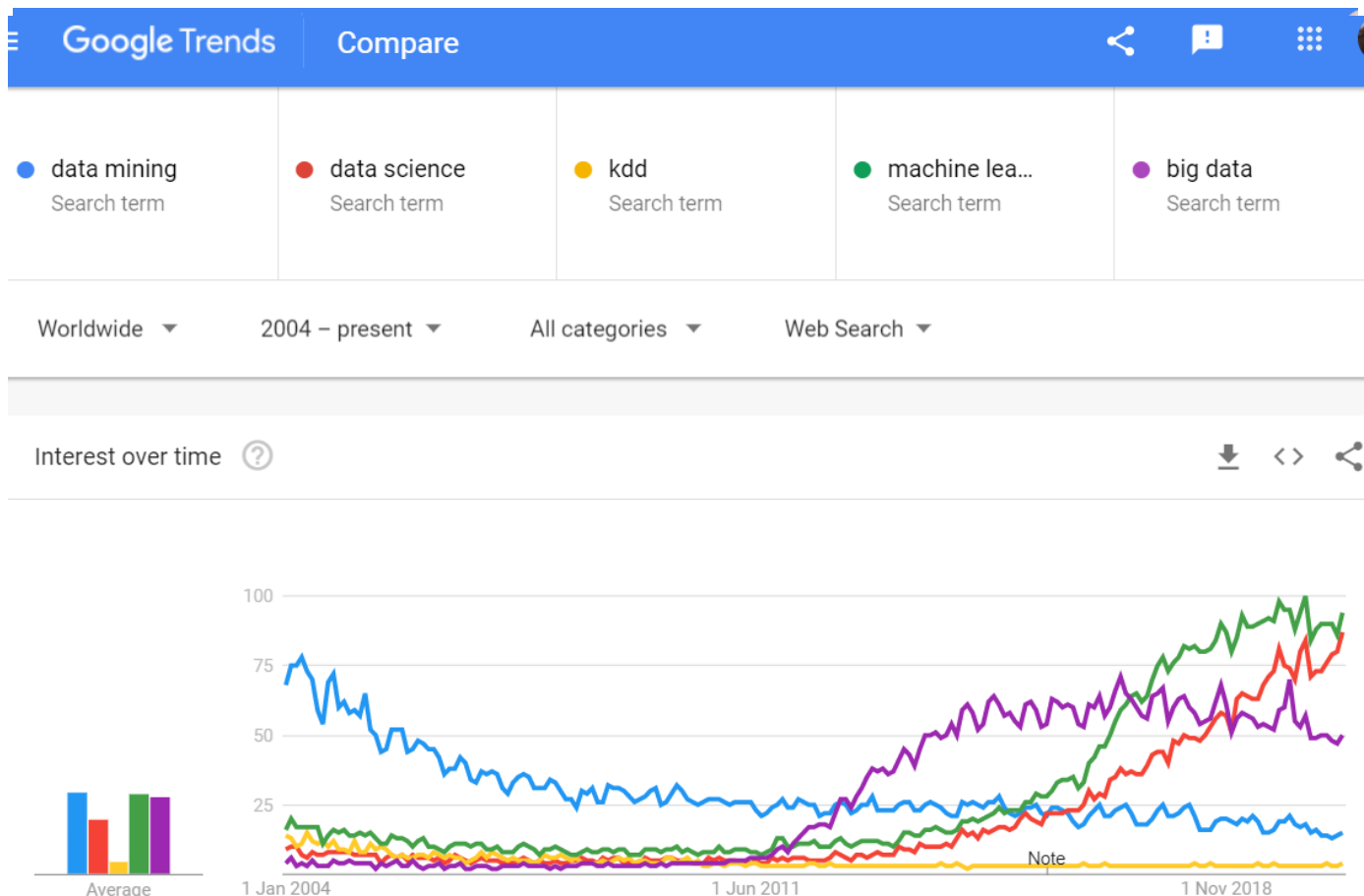
- data “**that exceeds the processing capacity of conventional database systems**” (O’Reilly “Planning for Big Data”, 2012)
- data “**beyond current database technology**” (Akerkar (ed.) “Big Data Computing”, 2014).
- data for which “**our jungle-surplus wetware can’t keep up**” (O’Reilly “Planning for Big Data”, 2012)
- data processing that is “**not scalable**” (Akerkar (ed.) “Big Data Computing”, 2014).
- information “**that can’t be processed or analyzed using traditional processes and tools**” (Eaton et al. “Understanding Big Data”, 2012)
- data that “**defies traditional storage**” (Kerzner and Maniyam, “Hadoop illuminated”, 2014).
- ...



- Which are the old and new technologies?
 - Relational databases
 - No-SQL
 - Data Mining
 - Data Science
 - Data Warehouses
 - VLDB: Very Large Databases
 - Cloud computing
 - Dashboards
 - Distributed databases
 - Data Analysis / Analytics
 - OLAP
 - KDD: Knowledge Discovery from Databases
 - Statistical Modelling
 - Machine learning
 - Cloud computing
 - ...



- The answer is blowing in ... Google Trends!



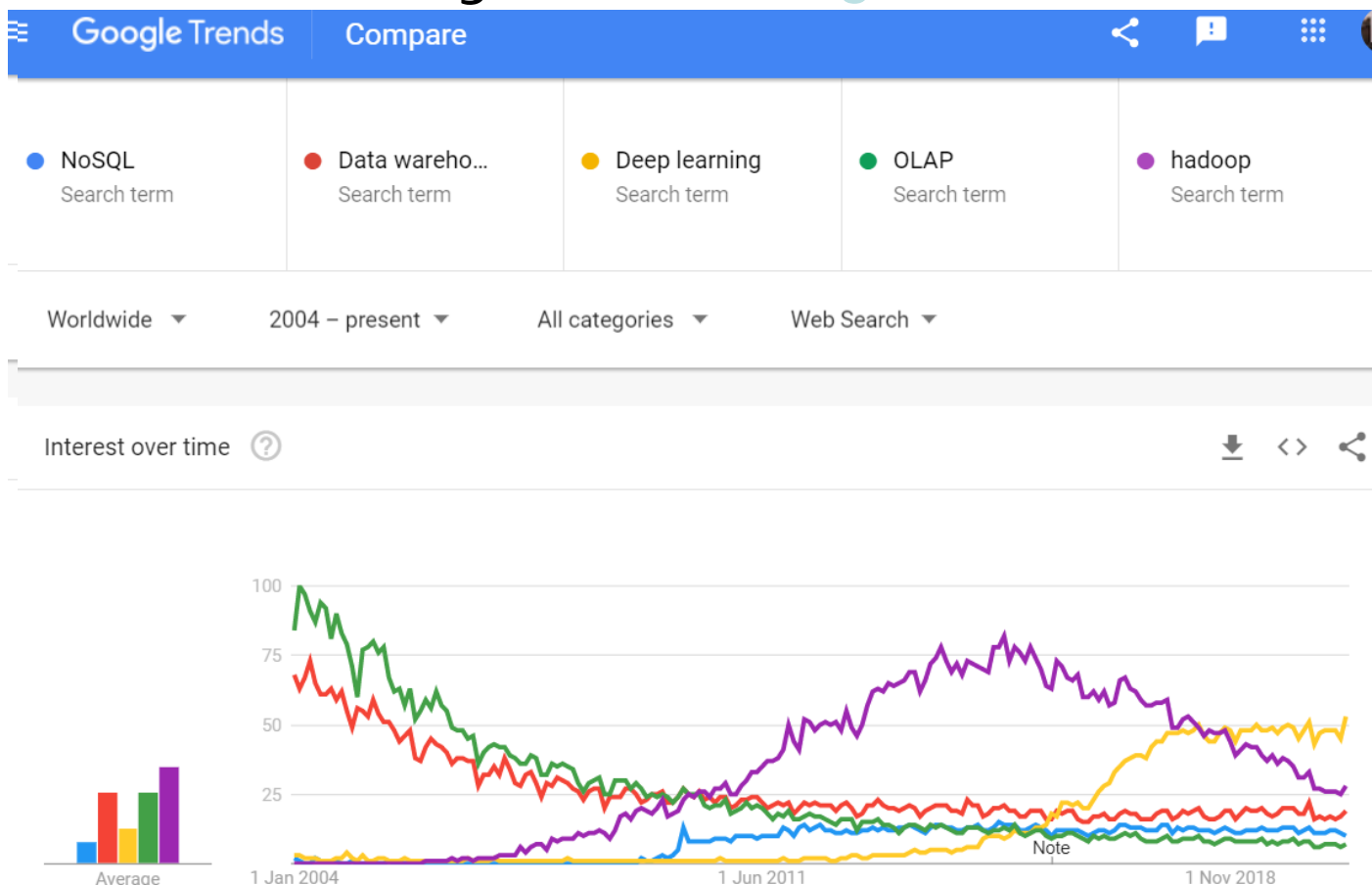
disciplines

58



- The answer is blowing in ...

Google Trends!



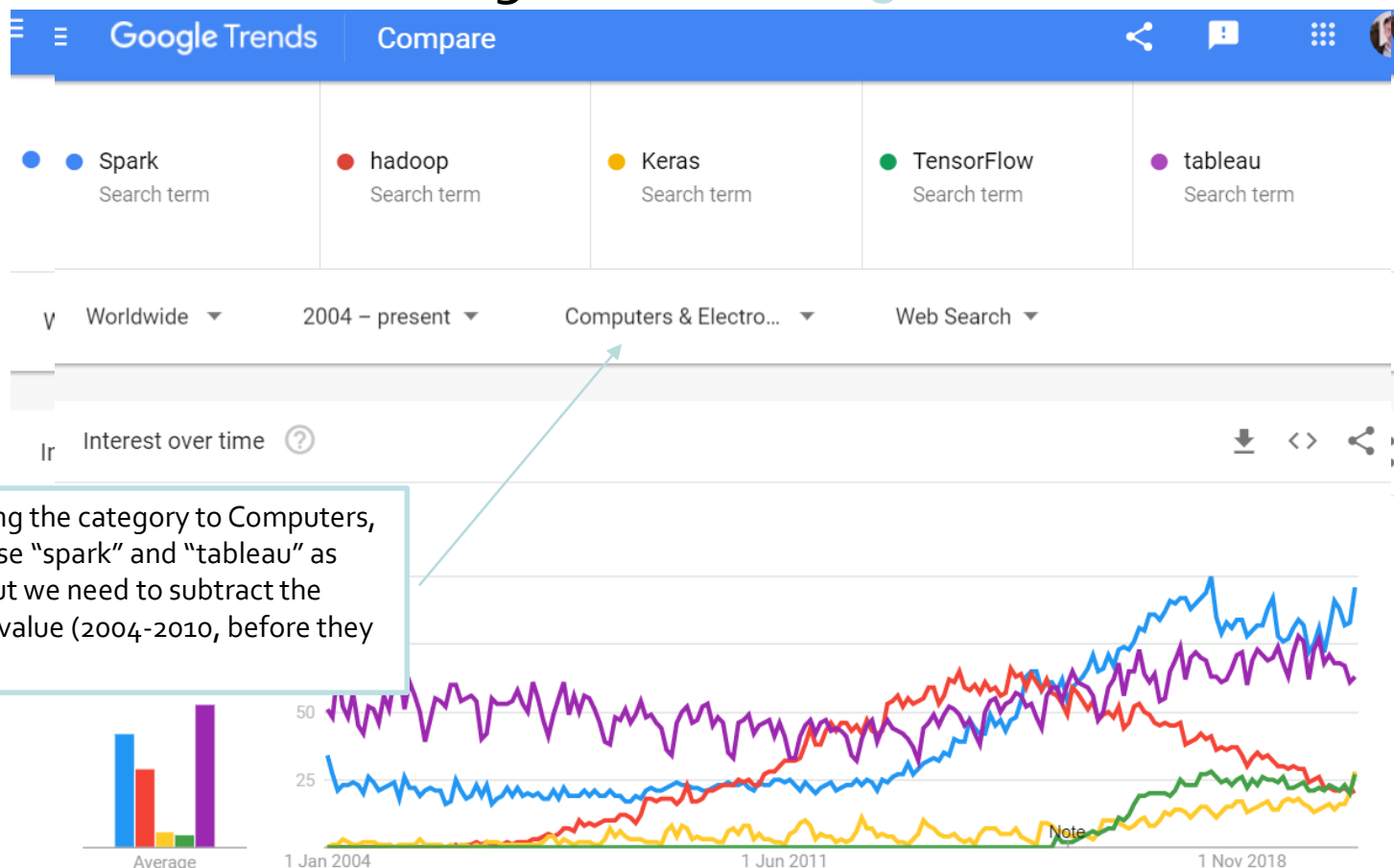
» technologies

59



- The answer is blowing in ...

Google Trends!

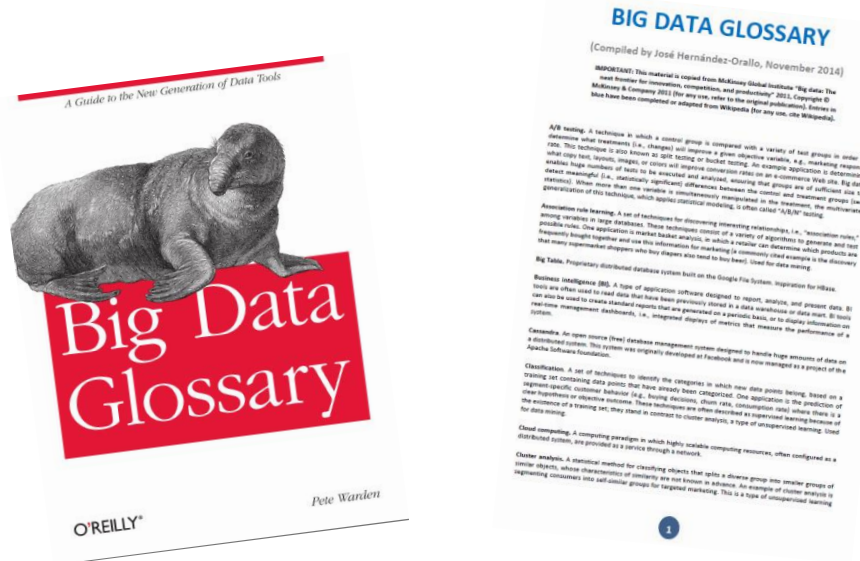


» tools

60



- Trends are biased by novelty.
- Many new terms and words appear everyday.
 - Check out for glossaries!



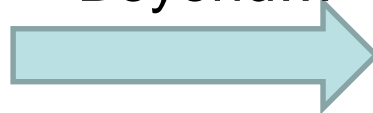
- We cannot tell whether things are new, or even valuable just by their names
 - Some are, some others are not.



- It's not easy to define an area *just* as a challenge.
- Or in terms of current technology or terms...

Current DB technology
Current Analytic Tools

Beyond...



What will *big data* be like
in 5 years time?



- Is *Big Data* just characterised by “*Big*”?
 - What's big? What's small?
 - 48KB was big!



- Eureka!

“What’s driving the need for *Big Data* solutions [...] isn’t some pre-defined ‘bigness’ of the data, but a combination of all three Vs”
(Fujitsu “White Book of Big Data”, 2012)

“Vs” ?

“Vig Data”!



- The V series!



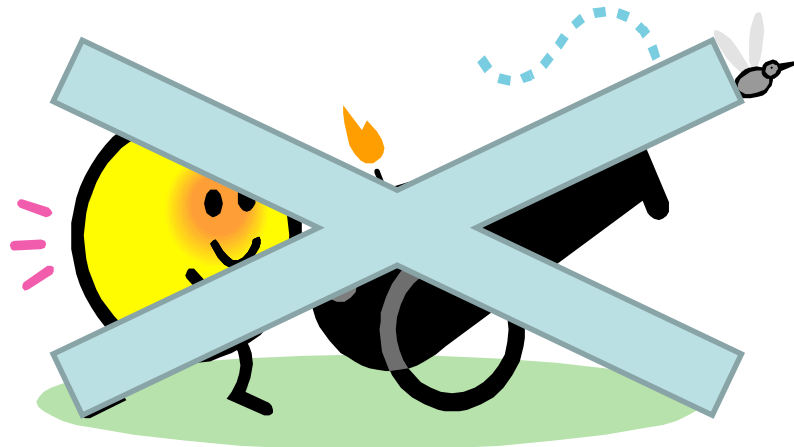
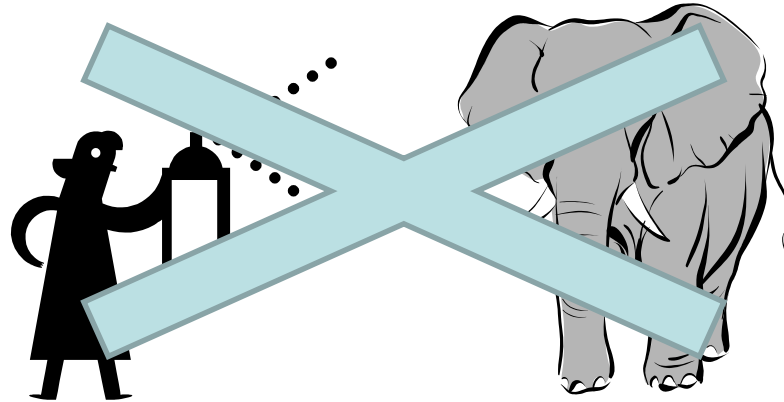
- Big Data 1.0: The 3 Vs (focus on capacity)
 - Volume
 - Velocity
 - Variety
- Big Data 2.0: The 4-5 Vs (focus on quality)
 - Veracity
 - Variability
- Big Data 3.0: More Vs. (focus on applicability)
 - Value
 - Volatility
 - Versatility
 - Visualisation
 - ...

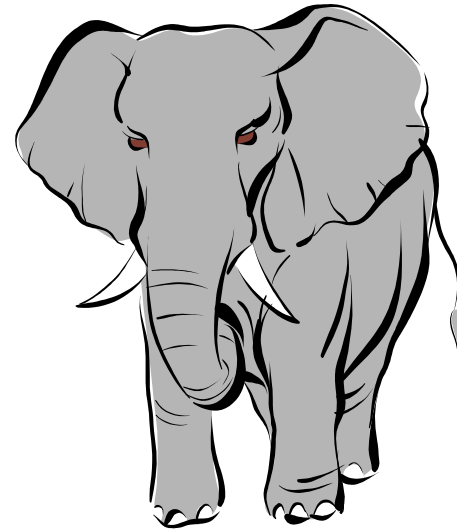
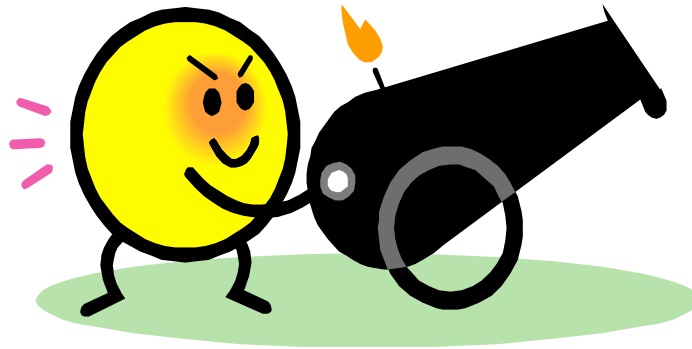


A giant with feet of clay?

- Let's build upon **goals** and not upon challenges.
- Remember the D2K goal and process.







■ Choosing the solutions

Only when we know the kind of data we can get, the kind knowledge we are interested in, the kind of processes and resources we want to commit with, can we choose the technology!

- Even so, the set of technologies is never-ending
 - “data capture and storage; search, sharing, and analytics; big data technologies; data visualization; architectures for massively parallel processing; data mining tools and techniques; machine learning algorithms for big data; cloud computing platforms; distributed file systems and databases; and scalable storage systems” *.
- Technologies are usually categorised into three groups:
 - Data Infrastructure
 - Data Analytics
 - Data Curation

* From the journal of big data www.journalofbigdata.com/, November 2014.

70



■ Data Infrastructure

- Data warehouses
- NoSQL databases
- Distributed file systems
- Scalable data storage
- Distributed databases
- Cloud computing
- ...

In some cases, a .CSV file could be enough!

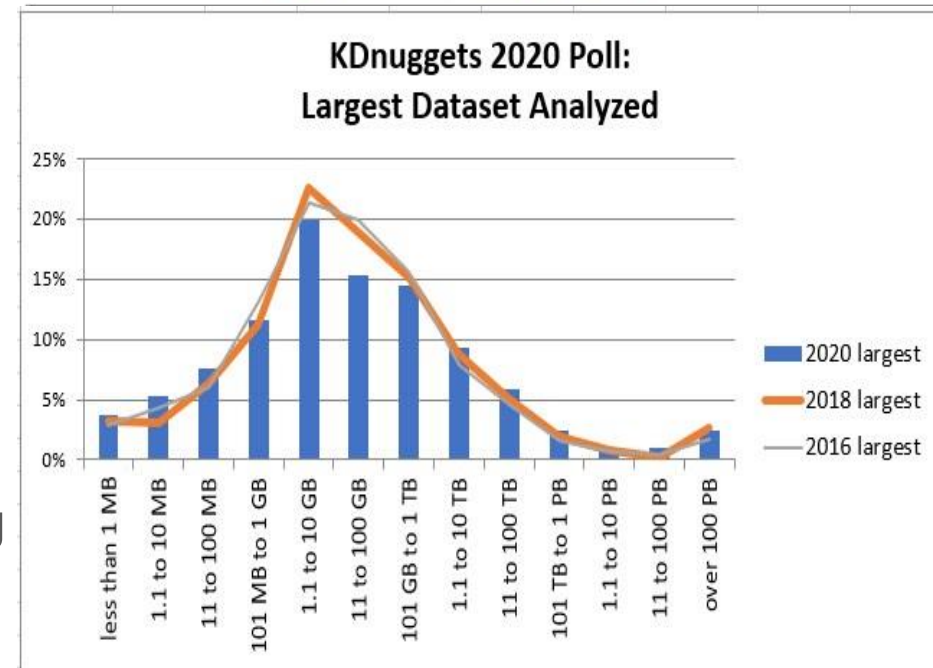
In some cases, “Hadoop projects distract businesses away from using Big Data to solve their business problems faster and instead tempt them onto the rocky road of developing their ‘ideal Big Data solution’ – which often ends up delivering nothing!” (White Book of Big Data, Fujitsu 2011)



■ Data Infrastructure

○ Don't "supersize".

- "Most data isn't "big," and businesses are wasting money pretending it is" (Quartz, qz.com, May 6, 2013)
 - "Even web giants like Facebook and Yahoo generally aren't dealing with big data, and the application of Google-style tools is inappropriate."
 - "Supersizing your data is going to cost you and may yield very little."



Surprising stability. The sizes are not growing!

* From kdnuggets.com



■ Data Infrastructure

○ When what you have is *really* insufficient:

- Do not just choose one solution because everybody is using it.
- Analyse all the alternatives, follow experts' advice.
 - Example: choosing a database for big data

DB Pros/Cons	HBase	Cassandra	Vertica	CloudTran	HyperTable
Pros	Key-based NoSQL, active user community, Cloudera support	Key-based NoSQL, active user community, Amazon's Dynamo on EC2	Closed-source, SQL-standard, easy to use, visualization tools, complex queries	Closed-source optimized on line transaction processing	Drop-in replacement for HBase, open-source, arguably much faster
Cons	Steeper learning curve, less tools, simpler queries	Steeper learning curve, less tools, simpler queries	Vendor lock-in, price, RDMS/BI - may not fit every application	Vendor lock-in, price, transaction-optimized, may not fit every application, needs wider adoption	New, needs user adoption and more testing
Notes	Good for new, long-term development	Easy to set up, no dependence on HDFS, fully distributed architecture	Good for existing SQL-based applications that needs fast scaling	Arguably the best OLTP	To be kept in mind as a possible alternative

Table 7.1, Kerzner and Maniyam "Hadoop Illuminated", 2014



■ Data Analytics

- Reporting with SQL
- OLAP tools
- Data visualisation
- Data mining tools
- Machine learning libraries
- Statistical packages
- ...

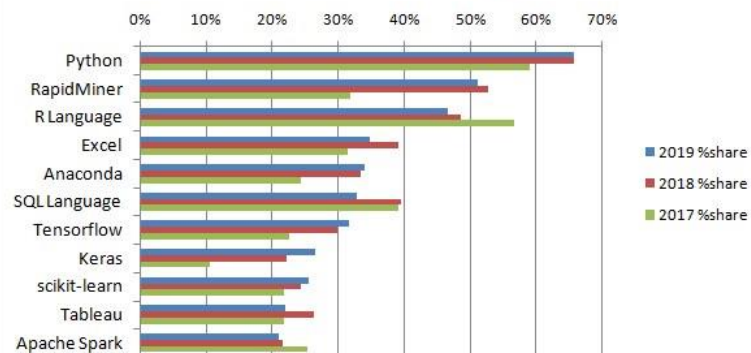
In some cases, a good OLAP query or a visualisation chart may be enough!

In some cases, an ensemble of support vector machines with a very complex kernel may give you just a very little push in your accuracy with respect to an elegant, comprehensible and manageable linear model.



■ Data Analytics: choose wisely...

Top Analytics, Data Science, Machine Learning Software 2017-2019, KDnuggets Poll



KDnuggets 2018 Data Science, Machine Learning Software Poll: Top Tools Associations



* From kdnuggets.com

75



■ Data Curation

- Data integration and manipulation tools.
- ETL tools.
- Data cleansing, wrangling and munging.
- Data quality: uncertainty and inconsistency handling.
- Privacy and ownership evaluation.
- Security.
- Metadata platforms and languages.
- ...

Not as fashionable as the other two legs.
But *curation* takes most of the time!

