The Process of Getting Actionable Insights from Data

(Big) Data as the Fuel and Analytics as the Engine of the Digital Transformation

Prof. Dr. Diego Kuonen, CStat PStat CSci

Statoo Consulting, Berne, Switzerland

@DiegoKuonen + kuonen@statoo.com + www.statoo.info

Berne, Switzerland — August 23, 2018

About myself (about.me/DiegoKuonen)

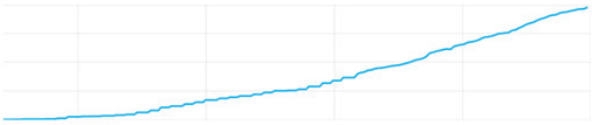
- ◇ PhD in Statistics, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.
- ◇ MSc in Mathematics, EPFL, Lausanne, Switzerland.

- CStat ('Chartered Statistician'), Royal Statistical Society, UK.
- PStat ('Accredited Professional Statistician'), American Statistical Association, USA.
- CSci ('Chartered Scientist'), Science Council, UK.

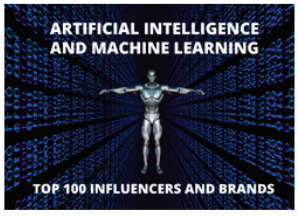
- Elected Member, International Statistical Institute, NL.
- Senior Member, American Society for Quality, USA.
- President of the Swiss Statistical Society (2009-2015).

- ▷ Founder, CEO & CAO, Statoo Consulting, Switzerland (since 2001).
- ▷ Professor of Data Science, Research Center for Statistics (RCS), Geneva School of Economics and Management (GSEM), University of Geneva, Switzerland (since 2016).
- ▷ Founding Director of GSEM's new MSc in Business Analytics program (started fall 2017).
- ▷ Principal Scientific and Strategic Big Data Analytics Advisor for the Directorate and Board of Management, Swiss Federal Statistical Office (FSO), Neuchâtel, Switzerland (since 2016).

@DiegoKuonen



- 30.11.2013: 3 followers
- 18.11.2014: 1'404
- 20.08.2018: 17'552



About Statoo Consulting (www.statoo.info)

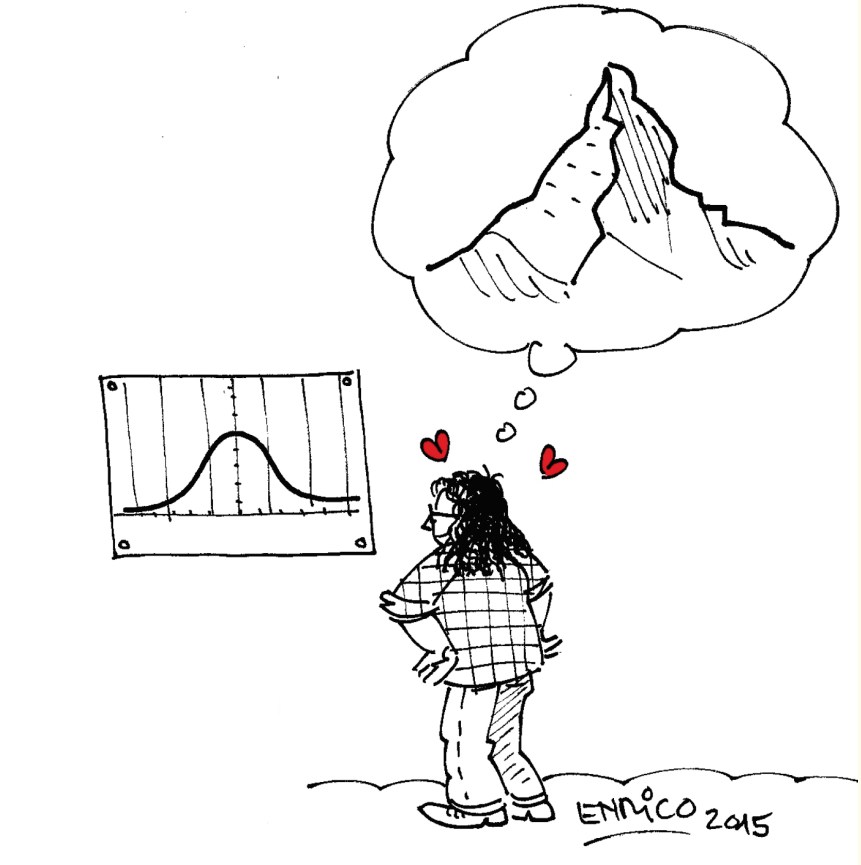
- Founded Statoo Consulting in **2001**.

$$\rightsquigarrow 2018 - 2001 = 17 + \epsilon.$$

- Statoo Consulting is a software-vendor independent Swiss consulting firm specialised in statistical consulting and training, data analysis, data mining (data science) and big data analytics services.
- Statoo Consulting offers consulting and training in statistical thinking, statistics, data mining and big data analytics in English, French and German.

$\rightsquigarrow$ Are you drowning in uncertainty and starving for knowledge?

$\rightsquigarrow$ Have you ever been Statooed?



Contents

‘Data is arguably the most important natural resource of this century. ... Big data is big news just about everywhere you go these days. Here in Texas, everything is big, so we just call it data.’

Michael Dell, 2014

1. Demystifying the ‘big data’ hype

• ‘Big data’ have hit the business, government and scientific sectors.

⇝ The term ‘big data’ — coined in 1997 by two researchers at the NASA — has acquired the trappings of religion.

• But, what exactly are ‘big data’ ?

◇ The term ‘big data’ applies to an accumulation of data that can not be processed or handled using traditional data management processes or tools.

⇝ Big data are a data management infrastructure which should ensure that the underlying hardware, software and architecture have the ability to enable ‘learning from data’ or ‘making sense out of data’, *i.e.* ‘analytics’.

- The following characteristics — 'the four Vs' — provide a definition:
  - 'Volume': 'data at rest', *i.e.* the amount of data ($\rightsquigarrow$ 'data explosion problem'), with respect to the number of observations ($\rightsquigarrow$ 'size' of the data), but also with respect to the number of variables ($\rightsquigarrow$ 'dimensionality' of the data);
  - 'Variety': 'data in many forms', 'mixed data' or 'broad data', *i.e.* different types of data (*e.g.* structured, semi-structured and unstructured, *e.g.* log files, text, web or multimedia data such as images, videos, audio), data sources (*e.g.* internal, external, open, public), data resolutions (*e.g.* measurement scales and aggregation levels) and data granularities;
  - 'Velocity': 'data in motion' or 'fast data', *i.e.* the speed by which data are generated and need to be handled (*e.g.* streaming data from devices, machines, sensors, drones and social data);

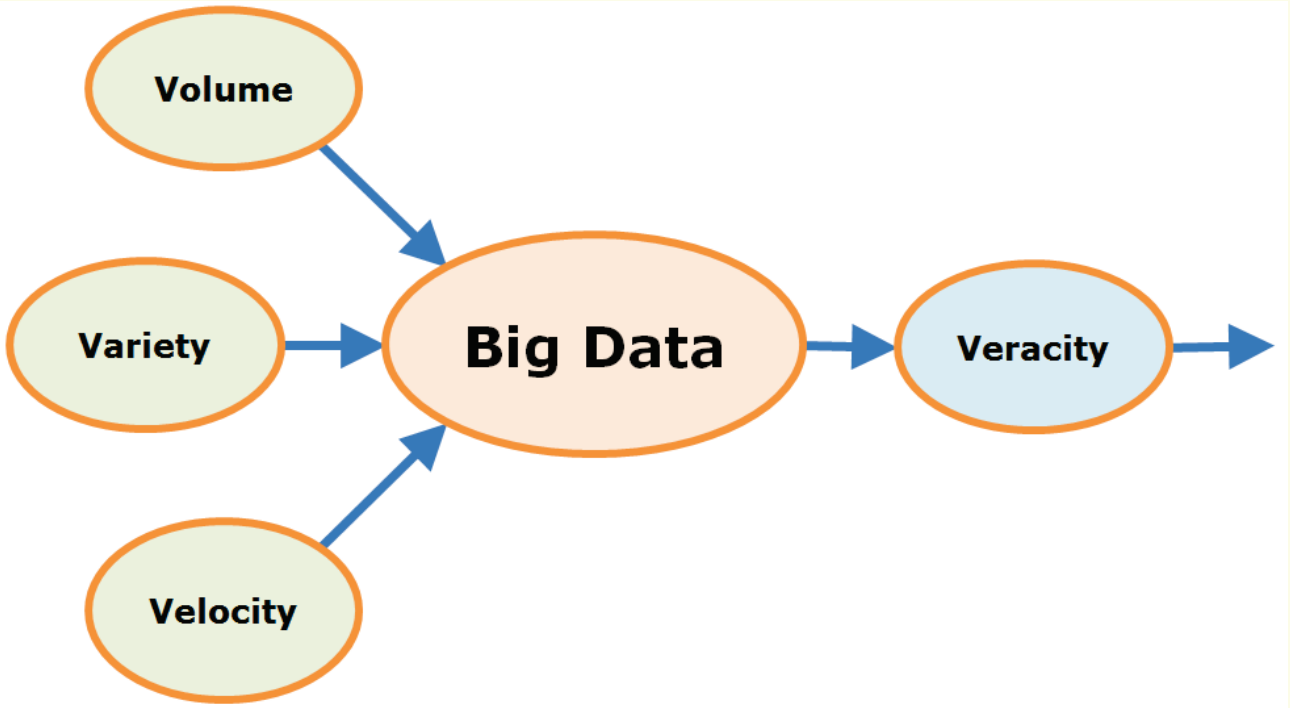
  - 'Veracity': 'data in doubt' or 'trust in data', *i.e.* the varying levels of noise and processing errors, including the reliability ('quality over time'), capability and validity of the data.

- 'Volume' is often the least important issue: it is definitely not a requirement to have a minimum of a petabyte of data, say.

$\rightsquigarrow$ Bigger challenges are 'variety' (*e.g.* combining different data sources such as internal data with social networking data and public data) and 'velocity', but most important is 'veracity' and the related quality of the data.

$\rightsquigarrow$ Indeed, big data come with the data quality and data governance challenges of 'small' data along with new challenges of its own!

$\rightsquigarrow$ Existing 'small' data quality frameworks need to be extended, *i.e.* augmented!



Big data in 1939

8 March 1939: Some of the four million tickets collected from London Underground passengers are examined in a survey by London Transport to discover the most and least used routes to help future infrastructure development

Source: 'The history of the Tube in pictures' (`goo.gl/dmJymR`).

⇝ Criticism of sceptics: these four Vs have always been there! But, what is new?

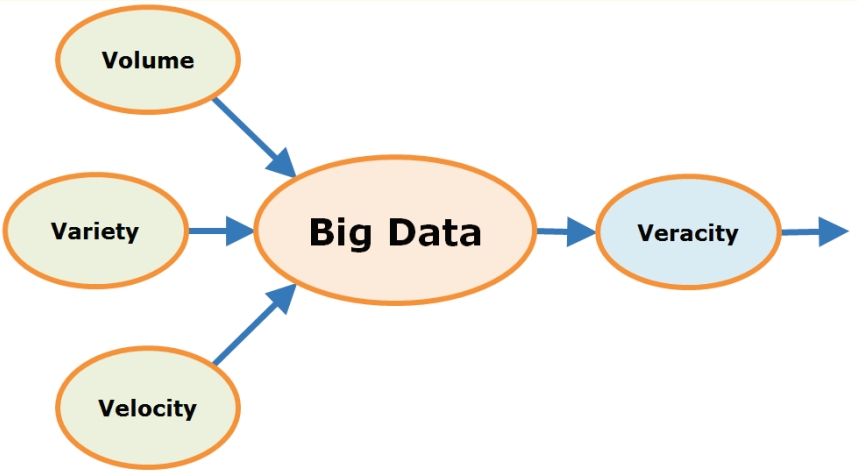
‘Data is part of Switzerland’s infrastructure, such as road, railways and power networks, and is of great value. The government and the economy are obliged to generate added value from these data.’

digitalswitzerland, November 22, 2016

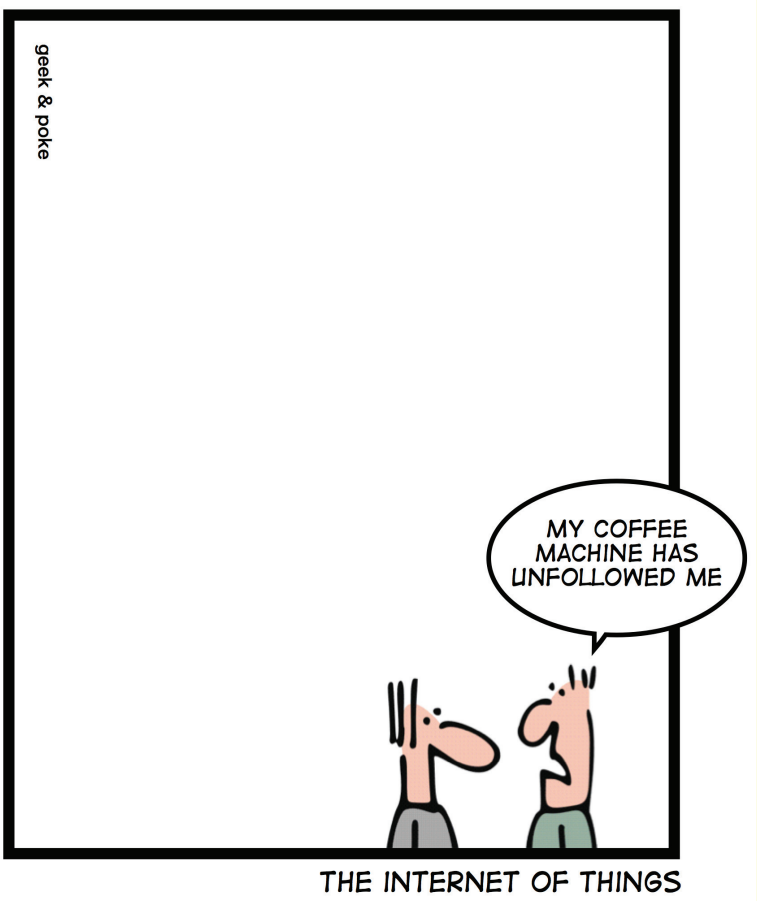
Source: digitalswitzerland’s ‘Digital Manifesto for Switzerland’ (`digitalswitzerland.com`).

$\rightsquigarrow$ The 5th V of big data: ‘Value’, *i.e.* the ‘usefulness of data’.

Intermediate summary: the ‘five Vs’ of (big) data



- ‘Volume’, ‘Variety’ and ‘Velocity’ are the ‘essential’ characteristics of (big) data;
- ‘Veracity’ and ‘Value’ are the ‘qualification for use’ characteristics of (big) data.



2. Demystifying the 'Internet of things' hype

• The term 'Internet of Things' (IoT) — coined in 1999 by the technologist Kevin Ashton — starts acquiring the trappings of a 'new religion'!

Source: Christer Bodell, 'SAS Institute and IoT', May 30, 2017 (goo.gl/cVYCKJ).

⇝ However, IoT is about data, not things!

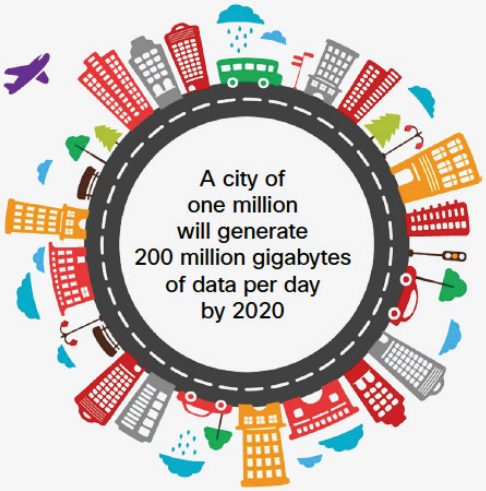
What Makes a Smart City?
Multiple Applications Create Big Data

Connected Plane
40 TB per day (0.1% transmtted)

Connected Factory
1 PB per day (0.2% transmitted)

Public Safety
50 PB per day (<0.1% transmitted)

Weather Sensors
10 MB per day (5% transmitted)



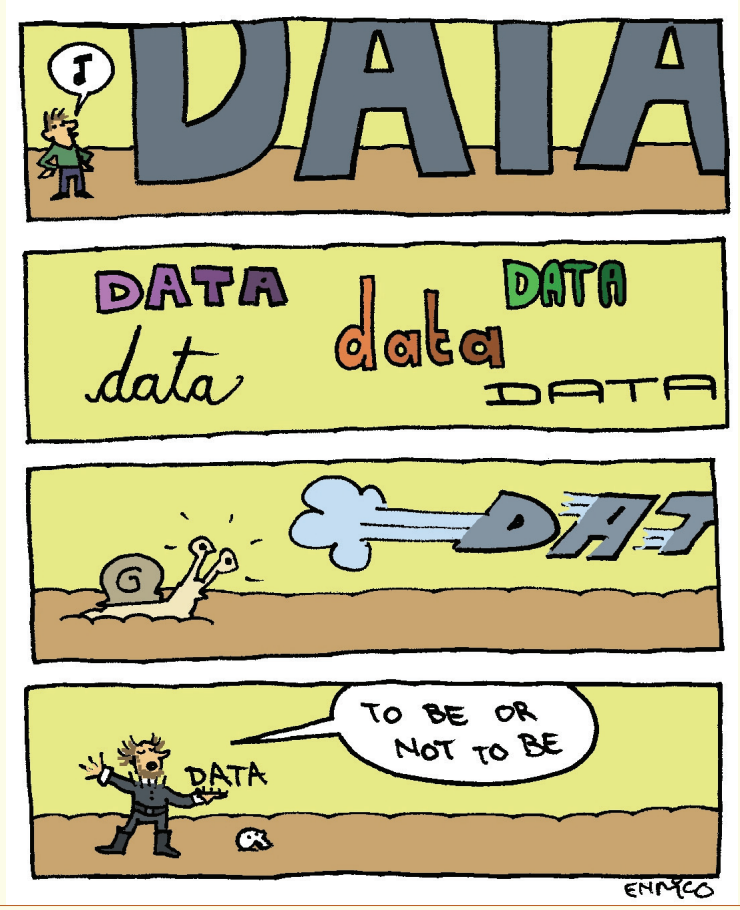
Intelligent Building
275 GB per day (1% transmitted)

Smart Hospital
5 TB per day (0.1% transmitted)

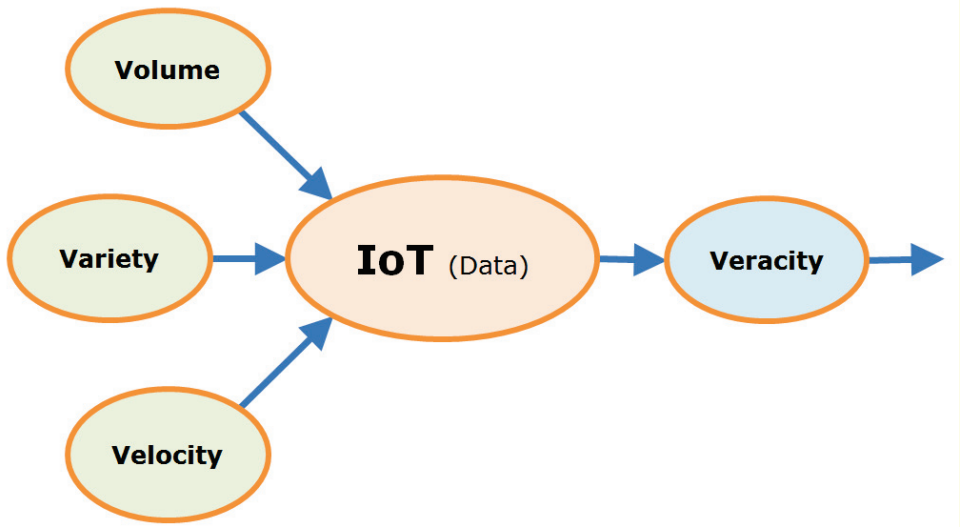
Smart Car
70 GB per day (0.1% transmitted)

Smart Grid
5 GB per day (1% transmitted)

Source: Cisco Global Cloud Index, 2015–2020



The ‘five Vs’ of IoT (data)



◇ ‘Volume’, ‘Variety’ and ‘Velocity’ are the ‘essential’ characteristics of IoT (data);

◇ ‘Veracity’ and ‘Value’ are the ‘qualification for use’ characteristics of IoT (data).

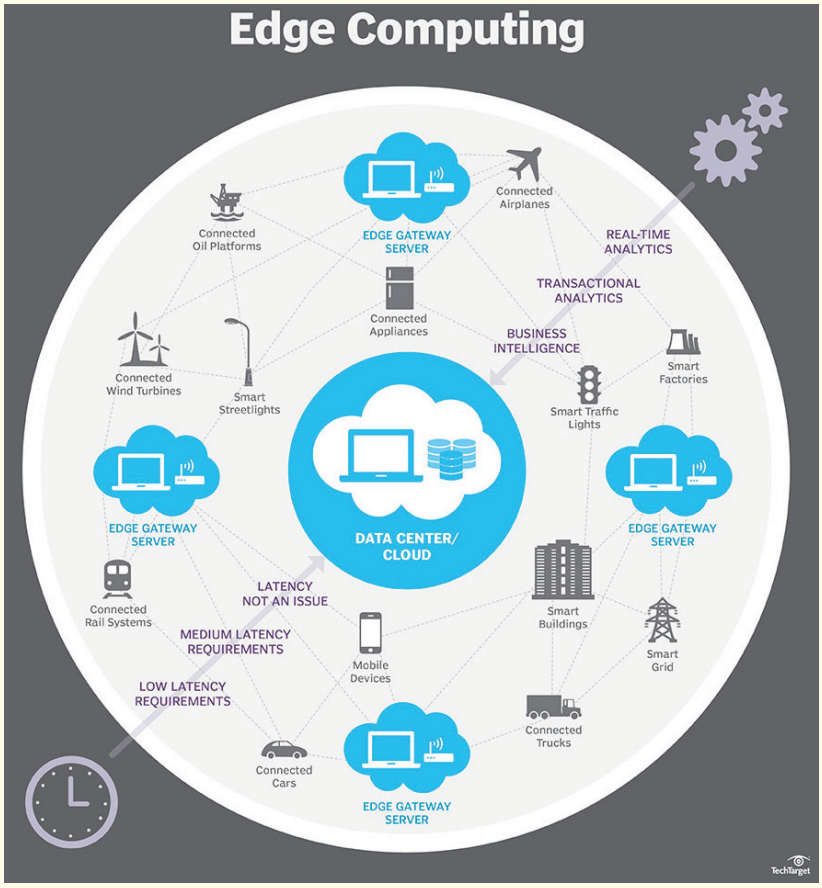
‘Analytics of things’

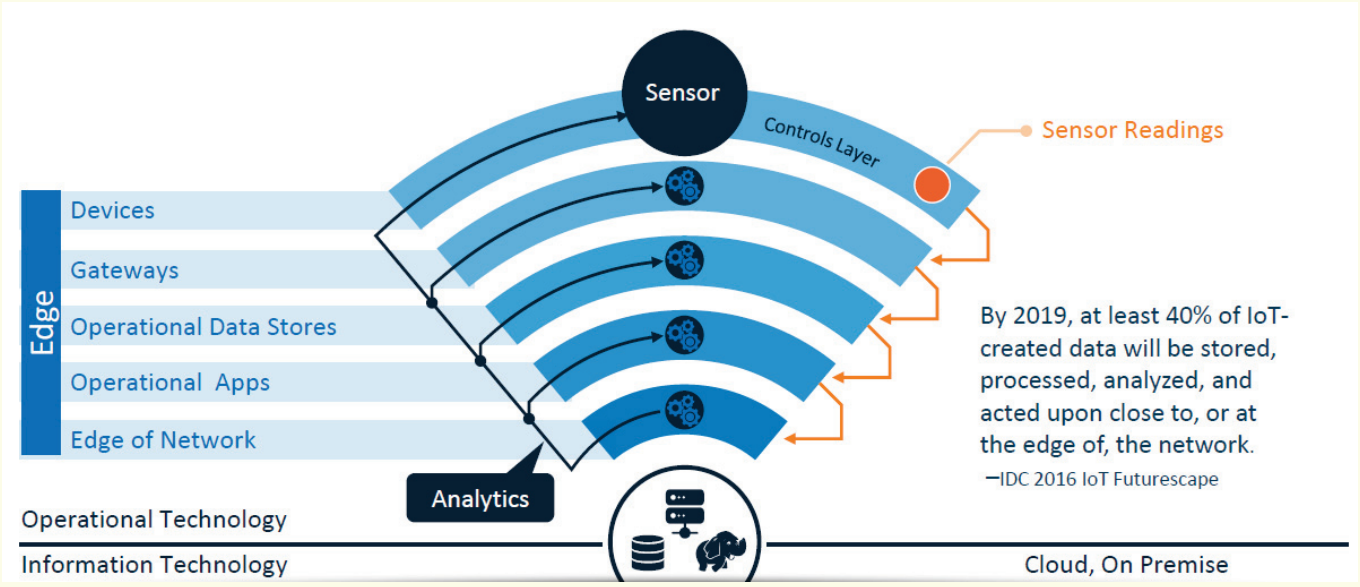
• The ‘Analytics of Things’ (AoT) corresponds to the ‘analytics layer’ that occurs with the IoT devices and their generated data.

⇝ It becomes key to execute analytics and related ‘data quality processes’ on the data-gathering devices themselves, *i.e.* at the edge (or at the ‘endpoint’), or as close to the originating data source as possible.

⇝ ‘Analytics at the edge’ or ‘**edge analytics**’ (based on a distributed ‘IT architecture layer’ called ‘edge computing’).

⇝ For example, in practice, the most efficient way to control data quality is to do it at the point where the data are created, as cleaning up data downstream (and hence centralised) is expensive and not scalable.





Source: Christer Bodell, 'SAS Institute and IoT', May 30, 2017 (goo.gl/cVYCKJ).

⇝ It is about moving the analytics and the 'data quality frameworks' to the data and not the data to the (centralised) analytics and (centralised) 'data quality frameworks'.

⇝ To do so, a centralised management of analytics will be needed; consisting, for example, of transparent central analytics model and rule development and maintenance, a common repository for all analytics models, *i.e.* 'algorithms', and a related analytics model version management.

⇝ Additional concerns are security (*e.g.* will be improved by reducing complexity), privacy (*e.g.* sensitive data will be retained at the edge), analytics governance (*e.g.* no strong governance needed as the algorithms are decentralised and publicly available), reliability and scalability of the edge devices, and (public) trust.

PROF. DR. ÈS SC. DIEGO KUONEN
CEO / CAO, Statoo Consulting
Adjunct Professor of Data Science, University of Geneva
@DiegoKuonen

My favourite new Internet of Things (IoT) product is Analytics of Things (AoT). IoT devices generate a lot of data and statistical principles and rigour are necessary to correctly collect the "right" data and to make sense out of these big data. Therefore, at the heart of AoT lies statistics.

Source: '12 incredible IoT products — Why are these experts excited about the future?', Manthan, India, April 29, 2016 (goo.gl/ZymF7y).

- Technology is **not** the real challenge of the digital transformation!

⇝ Digital is not about the technologies (which change too quickly)!

'Data are not taken for museum purposes; they are taken as a basis for doing something. If nothing is to be done with the data, then there is no use in collecting any. The ultimate purpose of taking data is to provide a basis for action or a recommendation for action.'

W. Edwards Deming, 1942

$\rightsquigarrow$ **Data are the fuel and analytics**, *i.e.* 'learning from data' or 'making sense out of data', **is the engine of the digital transformation** and the related data revolution!

3. Demystifying the two approaches of analytics

Statistics, data science and their connection

◇ Statistics traditionally is concerned with analysing **primary** (*e.g.* experimental or 'made' or 'designed') **data** that have been collected (and designed) to explain and check the validity of specific existing 'ideas' ('hypotheses'), *i.e.* through the operationalisation of theoretical concepts.

$\rightsquigarrow$ Primary analytics or **top-down** (*i.e.* explanatory and confirmatory) analytics.

$\rightsquigarrow$ 'Idea (hypothesis) evaluation or testing'.

$\rightsquigarrow$ Analytics' paradigm: '**deductive reasoning**' as 'idea (theory) first'.

◇ Data science — a rebranding of 'data mining' and as a term coined in 1997 by a statistician — on the other hand, typically is concerned with analysing **secondary** (*e.g.* observational or 'found' or 'organic' or 'convenience') **data** that have been collected (and designed) for other reasons (and often not 'under control' or without supervision of the investigator) to create new ideas (hypotheses or theories).

⇝ Secondary analytics or **bottom-up** (*i.e.* exploratory and predictive) analytics.

⇝ 'Idea (hypothesis) generation'.

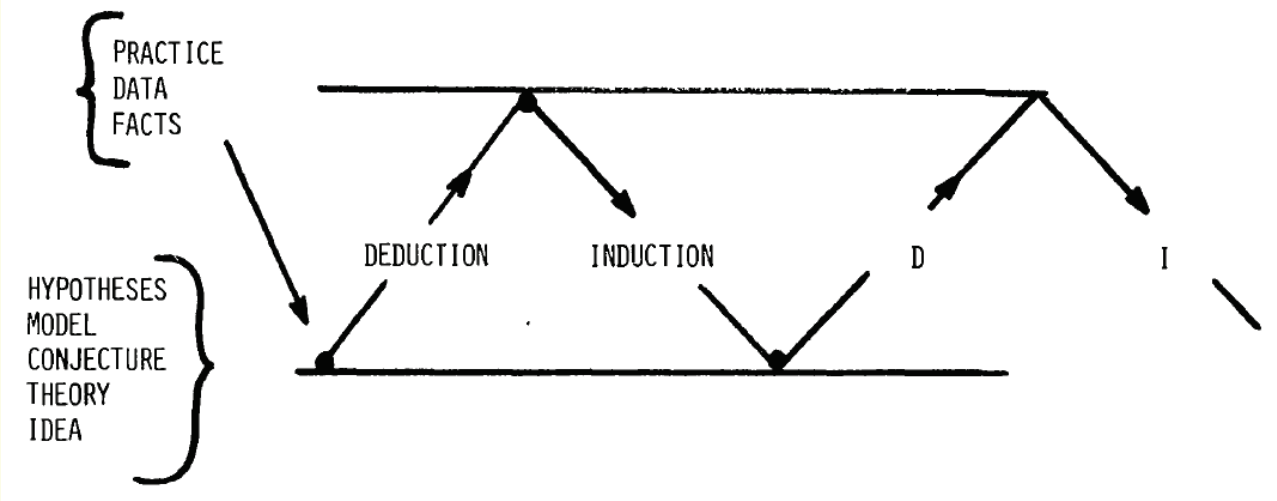
⇝ Analytics' paradigm: '**inductive reasoning**' as 'data first'.

'Neither exploratory nor confirmatory is sufficient alone. To try to replace either by the other is madness. We need them both.'

John W. Tukey, 1980

• The two approaches of analytics, *i.e.* deductive and inductive reasoning, are complementary and should proceed iteratively and side by side in order to enable 'data-driven decision making' and proper continuous improvement.

⇝ The **inductive–deductive reasoning cycle**:



Source: Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71, 791–799.

Example. When historical data are available the idea to be generated from a bottom-up analysis (*e.g.* 'predictive analytics' using a mixture of so-called 'ensemble techniques') could be

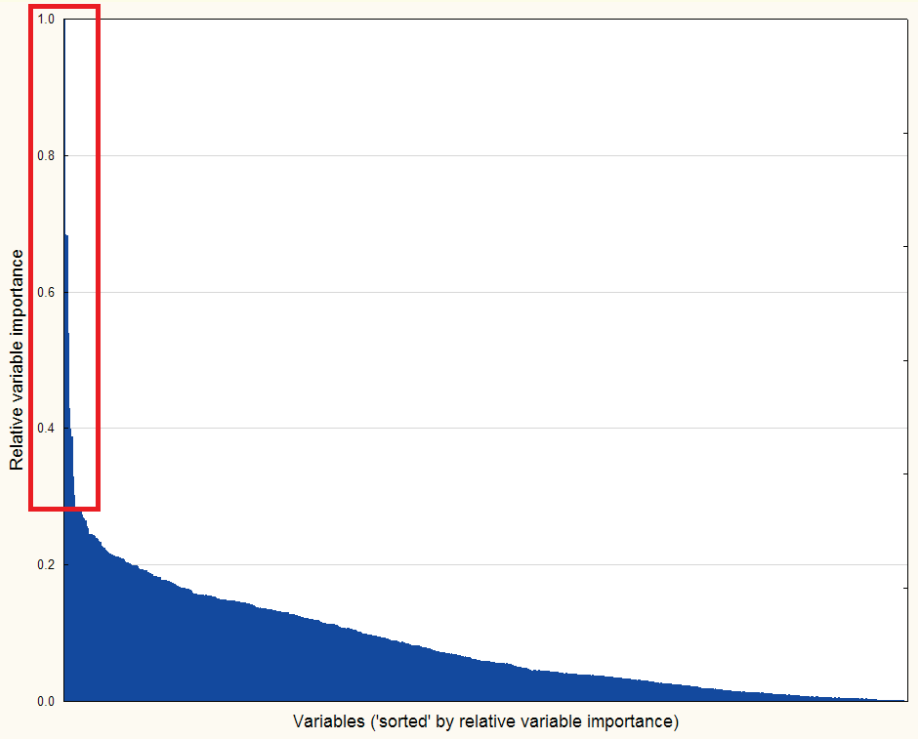
> 'which are the most important (from a predictive point of view) factors (among a 'large' list of candidate factors) that impact a given process output (or a given 'Key Performance Indicator', KPI)?'.

⇝ Mixed with subject-matter knowledge this idea could result in a list of a 'small' number of factors (*i.e.* 'the critical ones').

⇝ The confirmatory tools of top-down analysis (statistical 'Design Of Experiments', DOE, in most of the cases) could then be used to confirm and evaluate this idea.

⇝ By doing this, new data will be collected (about 'all' factors) and a bottom-up analysis could be applied again — letting the data suggest new ideas to test.

Example. 'Relative variable, *i.e.* factor, importance' measures (resulting from so-called 'stochastic gradient tree boosting' using real-world data on 679 variables):



Do **not** forget the term 'science' in 'data science'!

'Experiments may be conducted sequentially so that each set may be designed using the knowledge gained from the previous sets.'

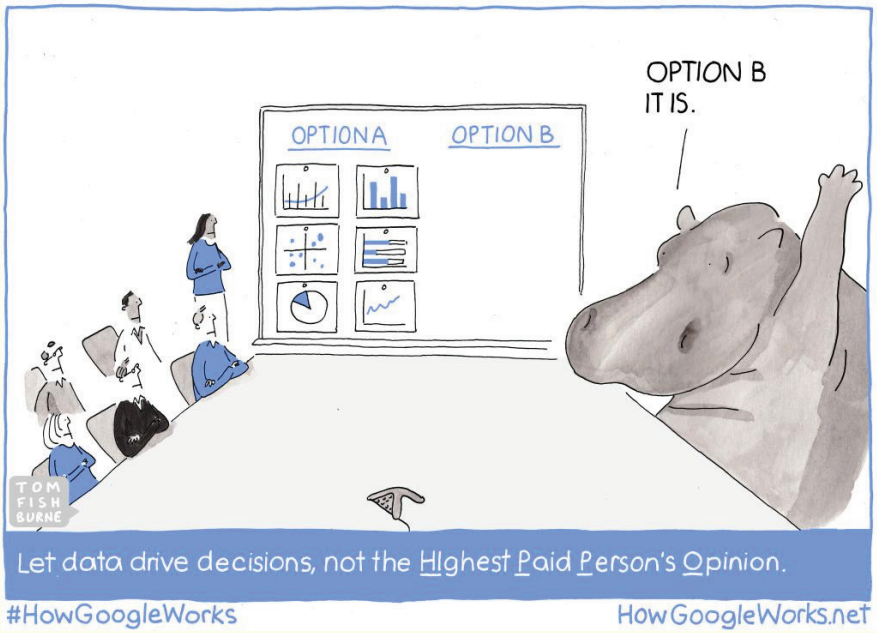
George E. P. Box and K. B. Wilson, 1951

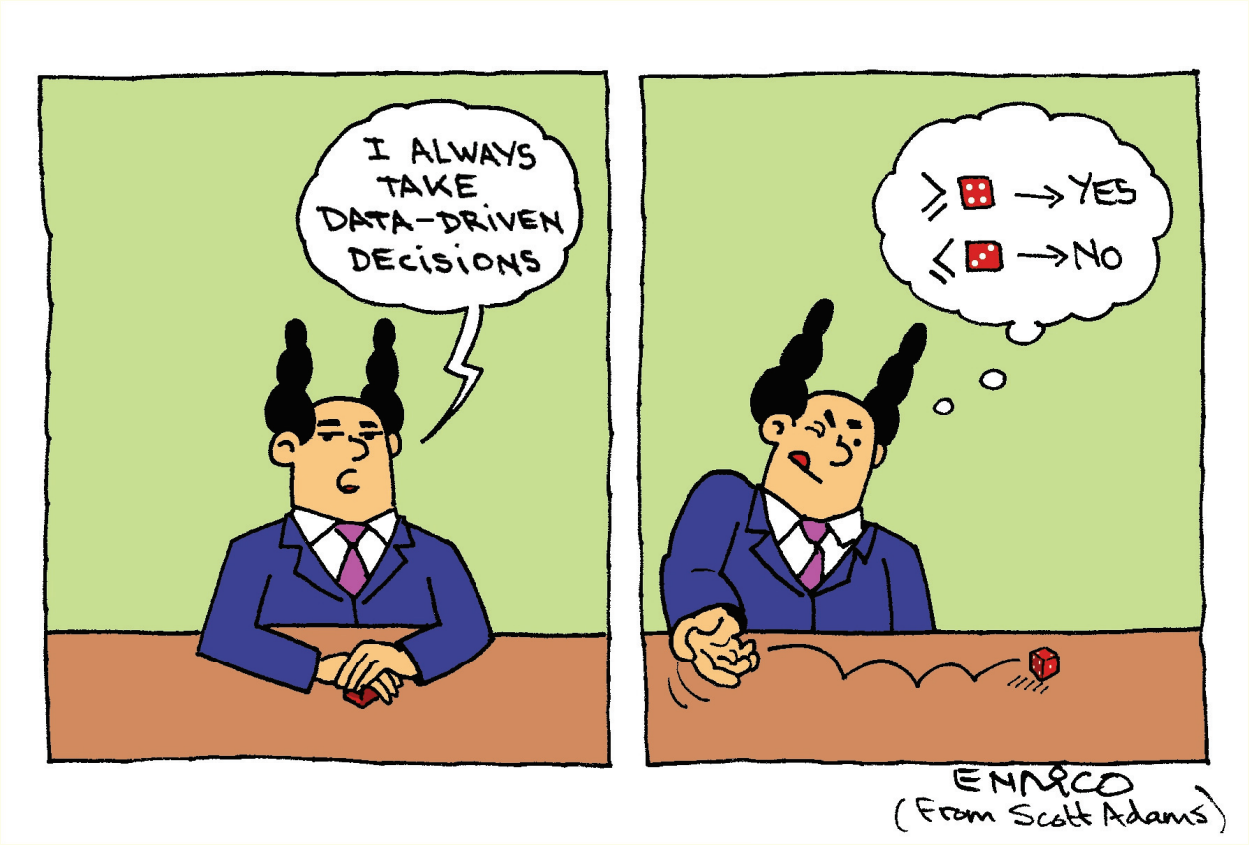
$\rightsquigarrow$ Scientific investigation is a sequential learning process!

$\rightsquigarrow$ Statistical methods allow investigators to accumulate knowledge!

Data-driven decision making

• Data-driven decision making: refers to the practice of basing decisions on the analysis of data, rather than purely on gut feeling and intuition:



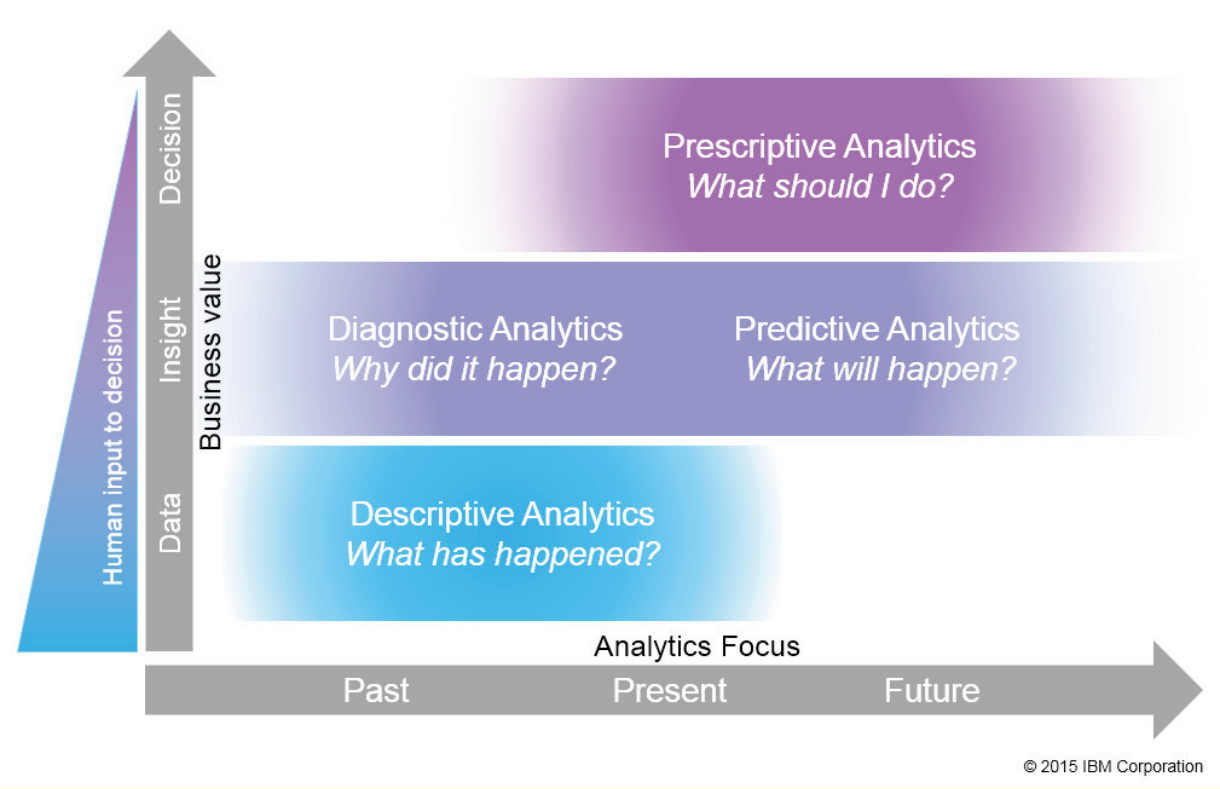


'Big data and analytics based on it promise to change virtually every industry and business function over the next decade.'

Thomas H. Davenport and Jinho Kim, 2013

4. Questions analytics tries to answer

- Analytics can answer questions like
  - 'what happened?' or 'what is happening now?' (*i.e.* describe, 'inform', or 'sense' $\rightsquigarrow$ 'hindsight') $\rightsquigarrow$ 'descriptive analytics' (*e.g.* 'Business Intelligence', BI);
  - 'why did it happen?' or 'why is it happening?' or 'what is happening?' or 'what are the trends?' or 'what patterns are there?' (*i.e.* explain, understand or 'respond' $\rightsquigarrow$ 'oversight') $\rightsquigarrow$ 'explanatory (or **diagnostic**) analytics';
  - 'what will happen?' (*i.e.* predict $\rightsquigarrow$ 'foresight') $\rightsquigarrow$ '**predictive** analytics';
  - 'what to do?' or 'how to make it happen?' or 'what is the best that could happen?' or 'how to optimise what happens?' (*i.e.* optimise, 'advise', 'recommend' or 'act' $\rightsquigarrow$ 'insight') $\rightsquigarrow$ '**prescriptive** analytics'.



Source: Jean-Francois Puget, Chief Architect, IBM Analytics Solutions, September 21, 2015 (goo.gl/V1412d).

Analytics Will Dramatically Improve Your Ability to ...

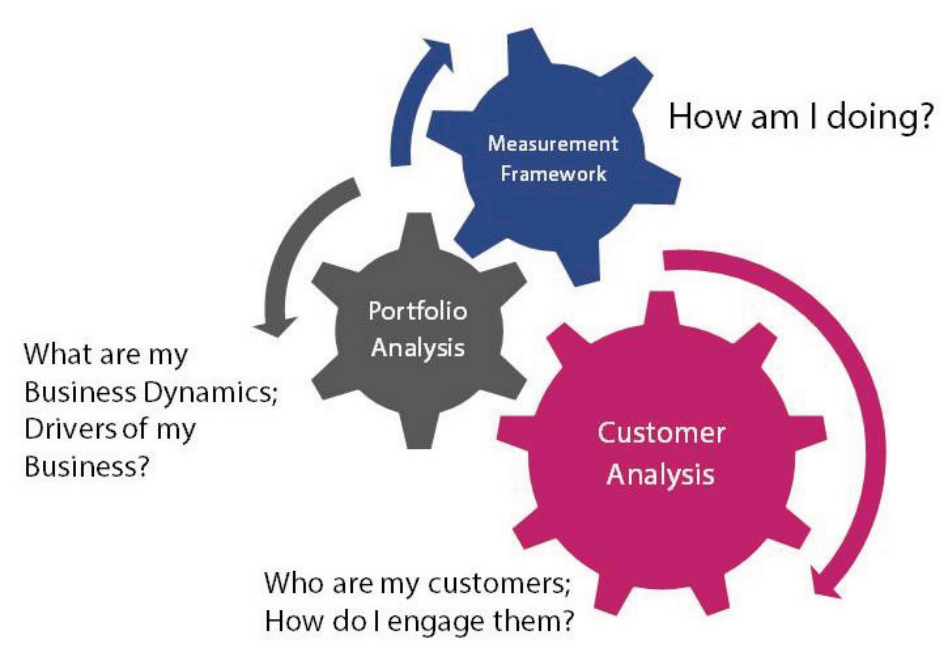
Measure	Understand	Decide
Identify Key Performance Indicators	Identify the Most Important Attributes	Identify Creative Choices
Set Goal/Target Values for Performance	Find Associations	Build Transparency Into Who and How Decisions Are Made
Assiduously Monitor Actual Values	Create Custom Groups	Provision: Simulation, Optimization, Experimental Design and Driver-Based Planning
Determine Leading Indicators	Blend Disparate Data Sources	Extreme Devil's Advocacy
Forecast Performance Measures	Drill Into Details	
	Build Taxonomy and Ontology	

Warning: Improving These Skills Has a Profoundly Positive Impact on Your Business

#GartnerSYM

Gartner

Three key 'business' analytics questions



Source: Piyanka Jain, 'Key analytics questions to ask your big data', *Forbes*, August 2012 (goo.gl/UrQo4a).

1. How am I doing?

   ⇝ Understand and agree on KPIs.

2. What drives my 'business'?

   ⇝ Once you have KPIs identified, you need to understand what drives these KPIs.

   ⇝ Which are the most important (from a predictive point of view) drivers (factors) that influence a given KPI?

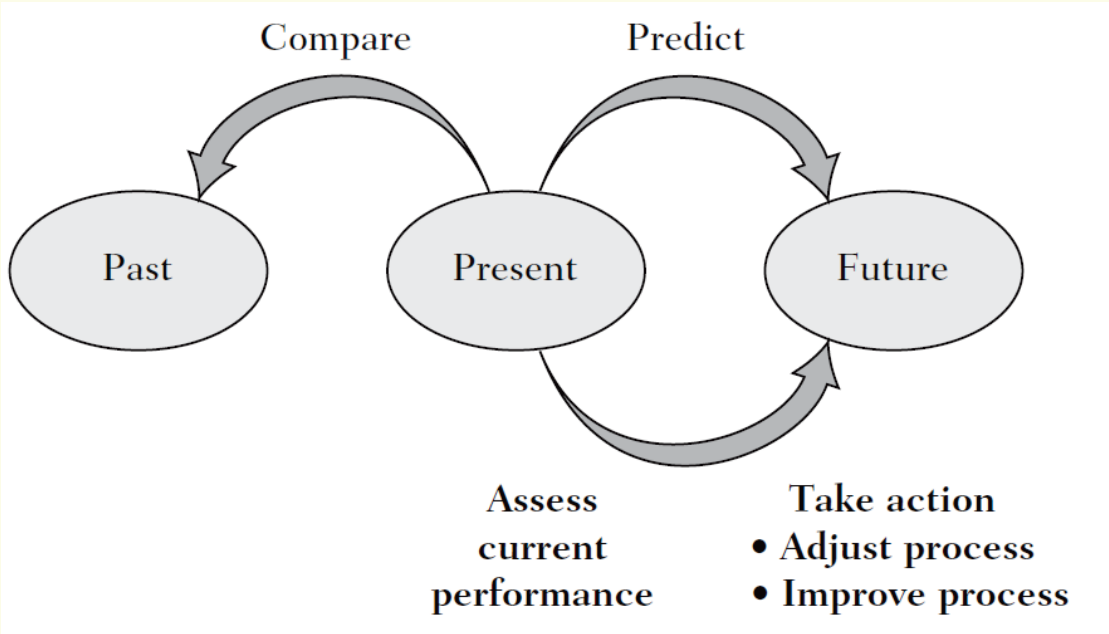
3. Who are my customers? What are their needs?

   ⇝ Understand customers and customise their offering, messaging, marketing channel accordingly, delight the customers, securing their future revenue (or KPI).

   ⇝ Drive a given KPI in the 'right' direction.

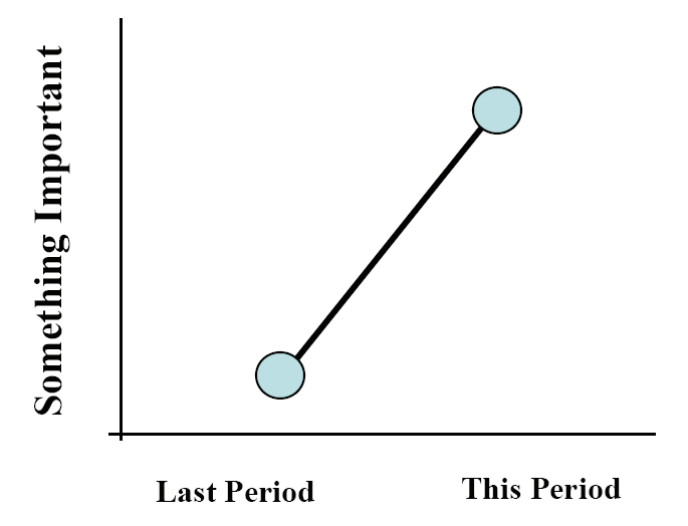
KPIs tracked over time (⇝ 'process tracking')

- KPIs tracked over time enable the analysis of a 'business' process as follows:



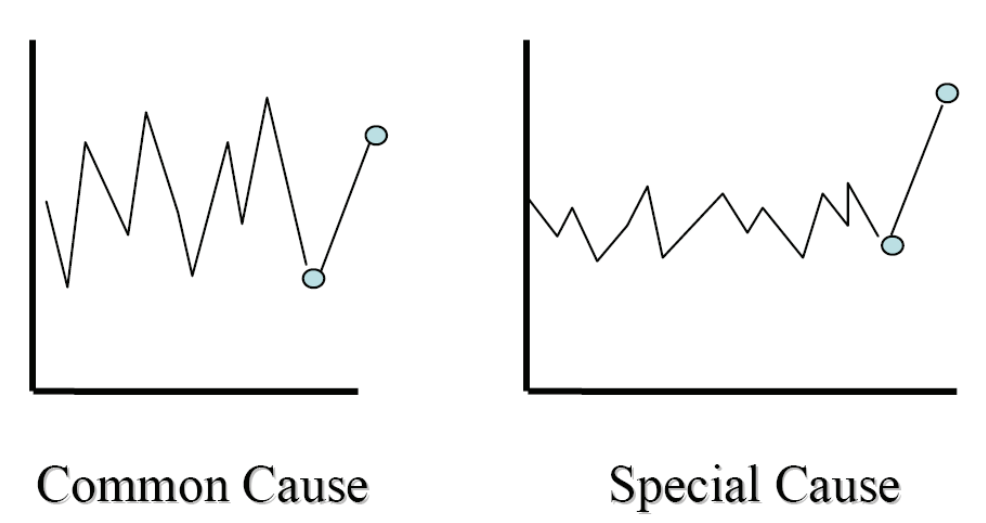
‘Do not treat common cause variation as special cause variation!’

Example. Giving two numbers, one will always be bigger:



⇝ What action is appropriate?

⇝ It depends:



⇝ Do not treat common cause variation as special cause variation (⇝ ‘tampering’)!

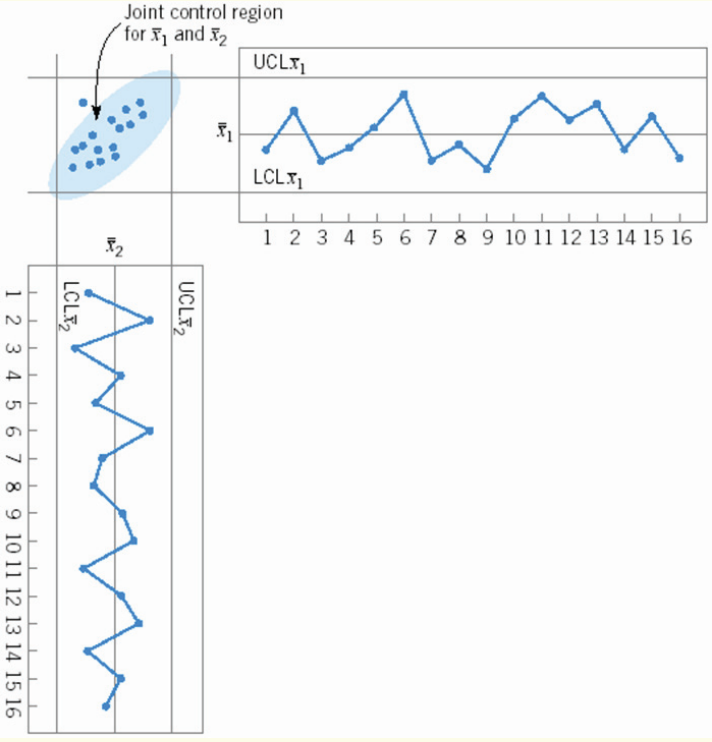
- The differences between special cause and common cause variation:

Special cause variation	Common cause variation
Called ‘assignable cause of variation’.	Called ‘chance cause of variation’.
Temporary and unpredictable.	Always present and predictable.
Few sources but each has large effect.	Numerous sources but each has small effect.
Often related to a specific event.	Part of the ‘normal’ behaviour of the process.
Process is unstable.	Process is stable.

⇝ The benefits of stable processes include:

- process performance is predictable; therefore there is a rational basis for planning;
- the effect of changes in the process can be measured faster and more reliably.

- Monitoring KPIs independently can be very misleading:



‘Confusing common causes with special causes will only make things worse.’

W. Edwards Deming

‘Machine learning has tremendous potential to transform companies, but in practice it is mostly far more mundane than robot drivers and chefs. Think of it simply as a branch of statistics, designed for a world of big data.’

Mike Yeomans, July 7, 2015

Source: Yeomans. M. (2015). What every manager should know about machine learning. *Harvard Business Review* (goo.gl/BfAWVN).

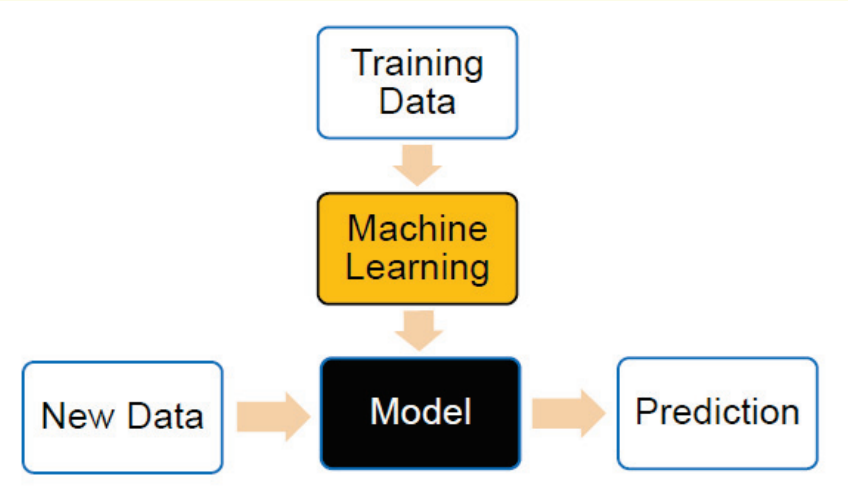
5. Demystifying the 'machine intelligence and learning' hype

◇ John McCarthy, one of the founders of 'Artificial Intelligence' (AI) (now sometimes referred to as 'Machine Intelligence', MI) research, defined in 1956 the field of AI as 'getting a computer to do things which, when done by people, are said to involve intelligence', *e.g.* visual perception, speech recognition, language translation and visual translation.

⇝ AI is about (smart) machines capable of performing tasks normally performed by humans (⇝ 'learning machines'), *i.e.* 'making machines smart'.

◇ In 1959, Arthur Samuel defined 'Machine Learning' (ML) as one part of a larger AI framework 'that gives computers the ability to learn'.

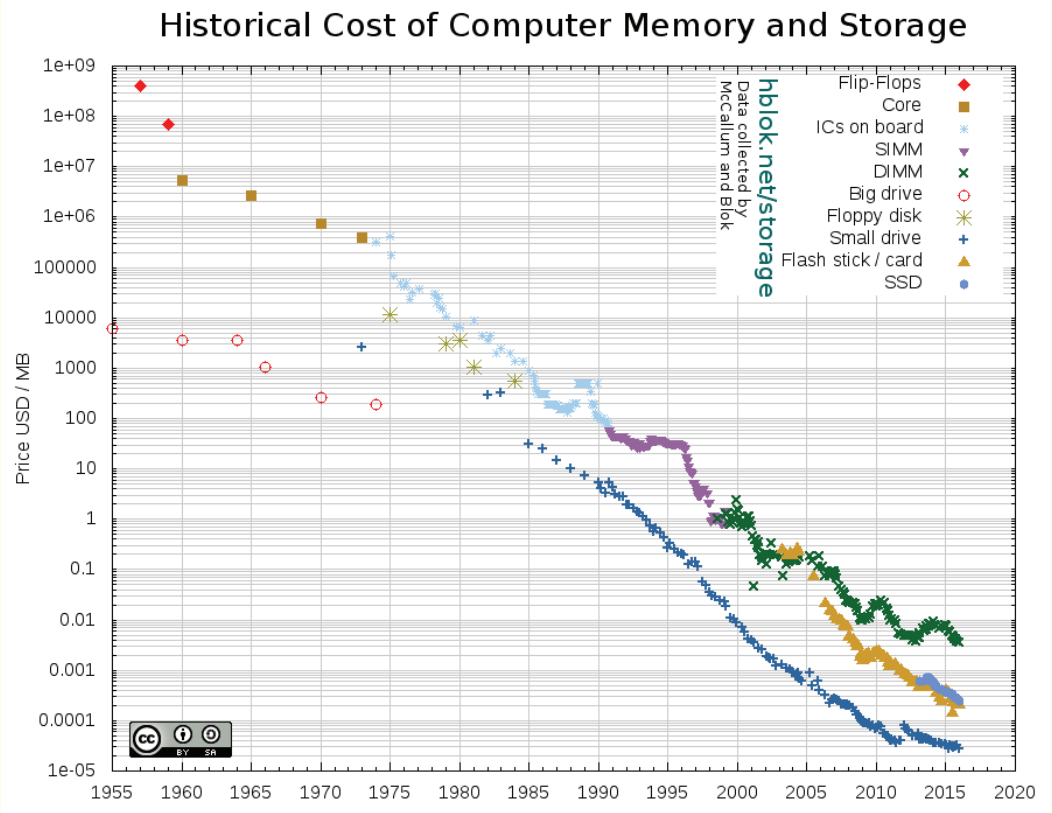
⇝ ML explores the study and construction of algorithms that can learn from and make predictions on (yet-to-be-seen) data, *i.e.* 'prediction making' through the use of computers.



‘In short, the biggest difference between AI then and now is that the necessary computational capacity, raw volumes of data, and processing speed are readily available so the technology can really shine.’

Kris Hammond, September 14, 2015

Source: Kris Hammond, ‘Why artificial intelligence is succeeding: then and now’, *Computerworld*, September 14, 2015 (goo.gl/Q3giSn).



Source: ‘Historical cost of computer memory and storage’ (hblok.net/blog/storage/).

• Most of the recent progress in machine learning involves mapping from a set of inputs to a set of outputs.

⇝ Some examples of such 'supervised (machine) learning systems':

Input X	Output Y	Application
Voice recording	Transcript	Speech recognition
Historical market data	Future market data	Trading bots
Photograph	Caption	Image tagging
Drug chemical properties	Treatment efficacy	Pharma R&D
Store transaction details	Is the transaction fraudulent?	Fraud detection
Recipe ingredients	Customer reviews	Food recommendations
Purchase histories	Future purchase behavior	Customer retention
Car locations and speed	Traffic flow	Traffic lights
Faces	Names	Face recognition

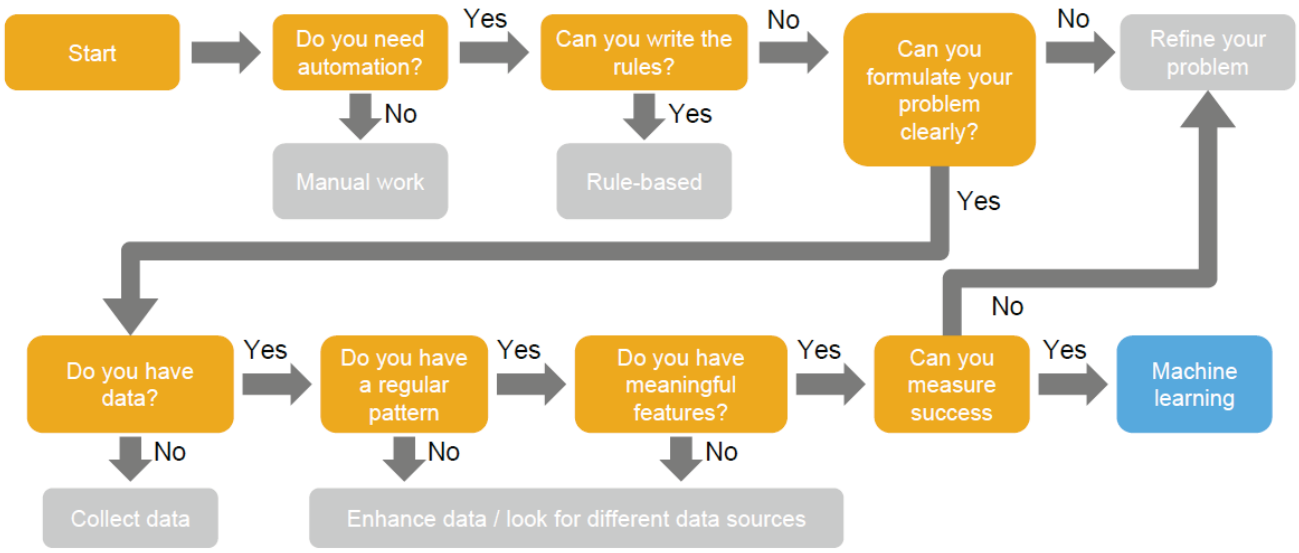
Source: Brynjolfsson, E. & McAfee, A. (2017). The business of artificial intelligence: what it can — and can not — do for your organization. *Harvard Business Review*, Digital Article, July 2017.

'Old theories never die, just the people who believe in them.'

Albert Einstein

From Business Problem to Machine Learning Problem: A Recipe

The "cheat sheet"



TASKS WELL SUITED TO MI

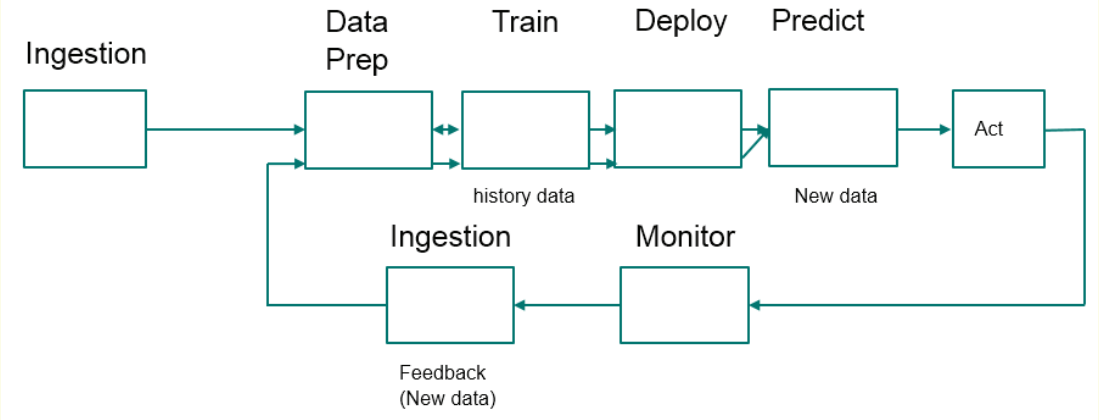
- Simple, requires little or no context
- Involves finding patterns in data
- Characterized by large amounts of data—preferably labeled
- Occurs in a static environment or one with little uncertainty

TASKS POORLY SUITED TO MI

- Complex, requires contextual understanding
- Requires explaining patterns in data
- Little or no data to characterize the problem
- Occurs in a dynamic environment with lots of uncertainty

Source: Booz Allen Hamilton (2017). *The Machine Intelligence Primer* (goo.gl/oi1AC5).

An example of a machine learning workflow



⇝ **Monitoring** and using the resulting **feedback** are at the core of machine learning.

⇝ Implementation requires the automation of the monitoring step and the feedback ingestion step. Assuming this is done, we have a 'learning machine'.

Source: Jean-Francois Puget, Chief Architect, IBM Analytics Solutions, 'Machine learning algorithm ≠ learning machine', April 27, 2016 (goo.gl/7nK4pR).

'AI algorithms are not natively 'intelligent'. They learn inductively by analyzing data. ... Sophisticated algorithms can sometimes overcome limited data if its quality is high, but bad data is simply paralyzing. Data collection and preparation are typically the most time-consuming activities in developing an AI-based application, much more so than selecting and tuning a model.'

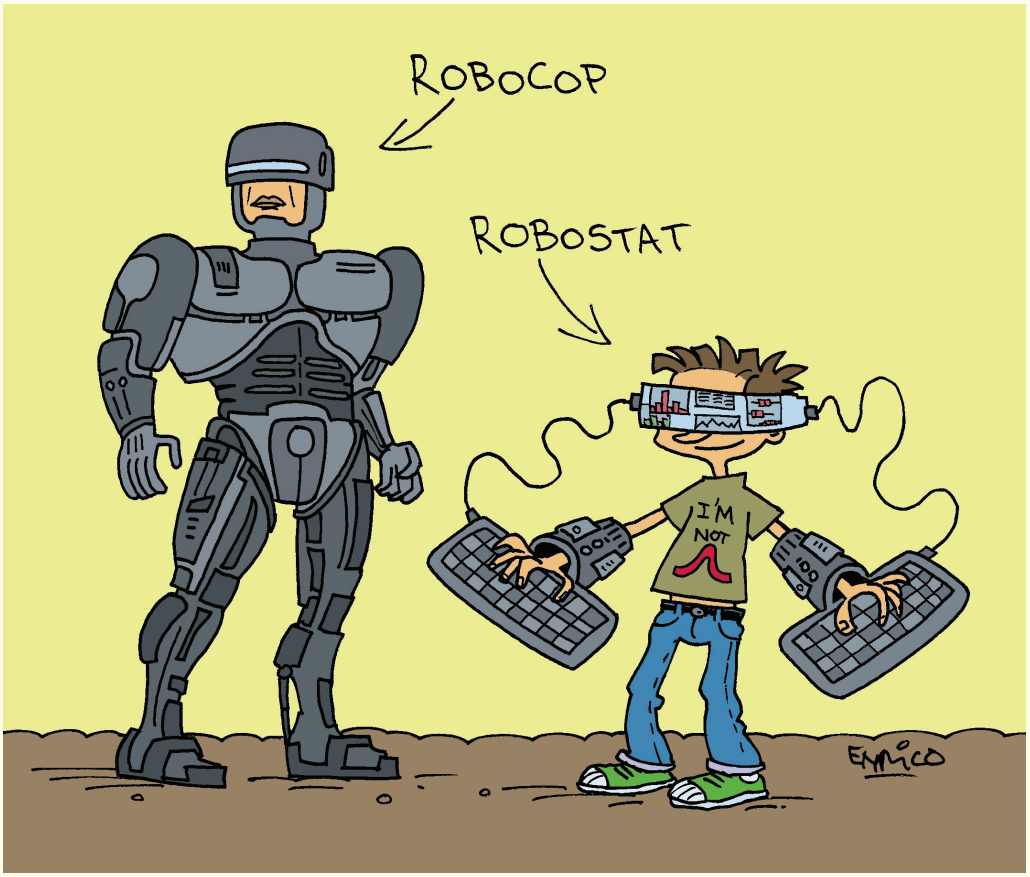
Sam Ransbotham, David Kiron, Philipp Gerbert and Martin Reeves, 2017

Source: Ransbotham, S., Kiron, D., Gerbert, P. & Reeves M. (2017). *Reshaping Business With Artificial Intelligence*. MIT Sloan Management Review & The Boston Consulting Group (goo.gl/wnGqr3).

• However, without humans as a guide, current AI is no more capable than a computer without software!

'Business is not chess; smart machines alone can not win the game for you. The best that they can do for you is to augment the strengths of your people.'

Thomas H. Davenport, August 12, 2015



> ‘In the anticipated symbiotic *[man–computer]* partnership, men will set the goals, formulate the hypotheses, determine the criteria, and perform the evaluations. Computing machines will do the routinizable work that must be done to prepare the way for insights and decisions in technical and scientific thinking. ... In one sense of course, any man-made system is intended to help man, to help a man or men outside the system.’
>
> Joseph C. R. Licklider, 1960

Source: Licklider, J. C. R. (1960). Man–computer symbiosis. *IRE Transactions on Human Factors in Electronics*, 1, 4–11.

Intermediate summary, principles for success and skills

• Decision making that was once based on hunches and intuition should be driven by data ($\rightsquigarrow$ data-driven decision making, *i.e.* muting the HIPPOs after having listened carefully to their opinions!).

• Despite an awful lot of marketing hype, big data are here to stay — as well as the IoT — and will impact every single ‘business’!

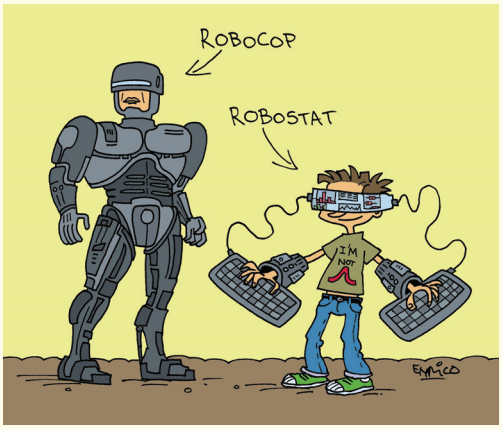
• Extracting useful knowledge from data to solve ‘business’ problems must be treated systematically by following a process with reasonably well-defined stages (see later).

• The key elements for a successful analytics future are statistical principles and rigour of humans!

- Analytics is an aid to thinking and not a replacement for it!

- Data and analytics should be envisaged to complement and augment humans, and not replacements for them!

$\rightsquigarrow$ Nowadays, with the digital transformation and the related data revolution, **humans need to augment their strengths to become more 'powerful'**: by automating any routinisable work and by focusing on their core competences.



'By 'augmenting human intellect' we mean increasing the capability of a man to approach a complex problem situation, to gain comprehension to suit his particular needs, and to derive solutions to problems.'

Douglas C. Engelbart, 1962

Source: Engelbart, D. C. (1962). 'Augmenting human intellect: a conceptual framework' (1962paper.org).

‘Digital strategies ... go beyond the technologies themselves. ... They target improvements in innovation, decision making and, ultimately, transforming how the business works.’

Gerald C. Kane, Doug Palmer, Anh N. Phillips, David Kiron and Natasha Buckley, 2015

Source: Kane, G. C., Palmer, D., Phillips, A. N., Kiron, D. & Buckley, N. (2015). Strategy, not technology, drives digital transformation. *MIT Sloan Management Review* (goo.gl/Dkb96o).

⇝ **Digital is not about the technologies** (which change too quickly)!

My key principles for analytics’ success

- **Do not neglect** the following four principles that ensure successful outcomes:
  - use of sequential approaches to problem solving and improvement, as studies are rarely completed with a single data set but typically require the sequential analysis of several data sets over time;
  - carefully considering data quality and assessing the ‘data pedigree’ before, during and after the data analysis; and
  - having a strategy for the project and for the conduct of the data analysis; including thought about the ‘business’ objectives (⇝ ‘strategic thinking’);
  - applying sound subject matter knowledge (‘domain knowledge’ or ‘business knowledge’, *i.e.* knowing the ‘business’ context, process and problem to which analytics will be applied), which should be used to help define the problem, to assess the data pedigree, to guide data analysis and to interpret the results.

‘It is getting better. . . A little better all the time.’

The Beatles, 1967

‘The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.’

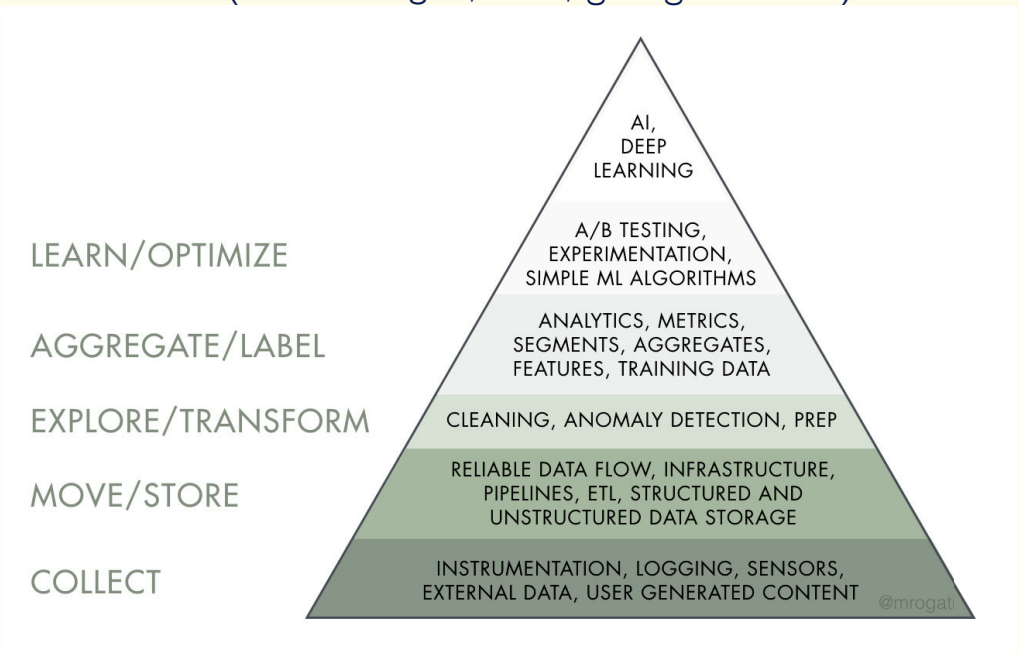
John W. Tukey, 1986

'To properly analyze data, we must first understand the process that produced the data. While many ... take the view that data are innocent until proven guilty, ... it is more prudent to take the opposite approach, that data are guilty until proven innocent.'

Roger W. Hoerl, Ronald D. Snee and Richard D. De Veaux, 2014

Source: Hoerl, R. W., Snee, R. D. & De Veaux, R. D. (2014). Applying statistical thinking to 'big data' problems. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6, 222–232.

• The largest and most basic 'need' in the analytics hierarchy is the need for a 'strong' data collection (Monica Rogati, 2017; goo.gl/F7hKH7):



⇝ **Data should be treated as a key strategic asset, so ensuring their veracity and the related data quality become imperative!**

‘Data is rapidly becoming the lifeblood of the global economy. It represents a key new type of economic asset. Those that know how to use it have a decisive competitive advantage in this interconnected world, through raising performance, offering more user-centric products and services, fostering innovation — often leaving decades-old competitors behind.’

European Political Strategy Centre (EPSC), January 2017

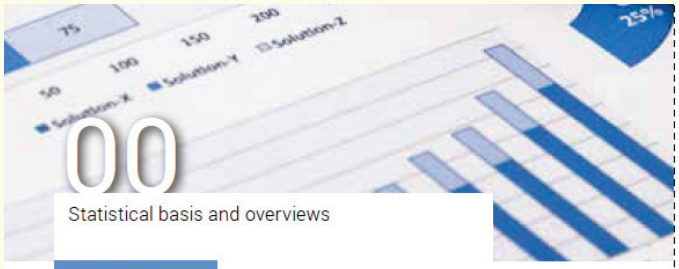
Source: EPSC, *Enter the Data Economy: EU Policies for a Thriving Data Ecosystem*, EPSC Strategic Notes, Issue 21, January 11, 2017 (goo.gl/RAeota).

‘You do not need a digital strategy. You need a better *(‘business’)* strategy, enabled by digital. ’

George Westerman, 2018

Source: Westerman, G. (2018). Your company doesn’t need a digital strategy. *MIT Sloan Management Review*, 59(3), 14–15 (goo.gl/mSb5yd).

⇝ **Focus on transformation instead of technology!**



1790-1700

Swiss Federal Statistical Office Data Innovation Strategy

Purpose, strategic objectives and implementation steps

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Statistical Office FSO

Neuchâtel 2017

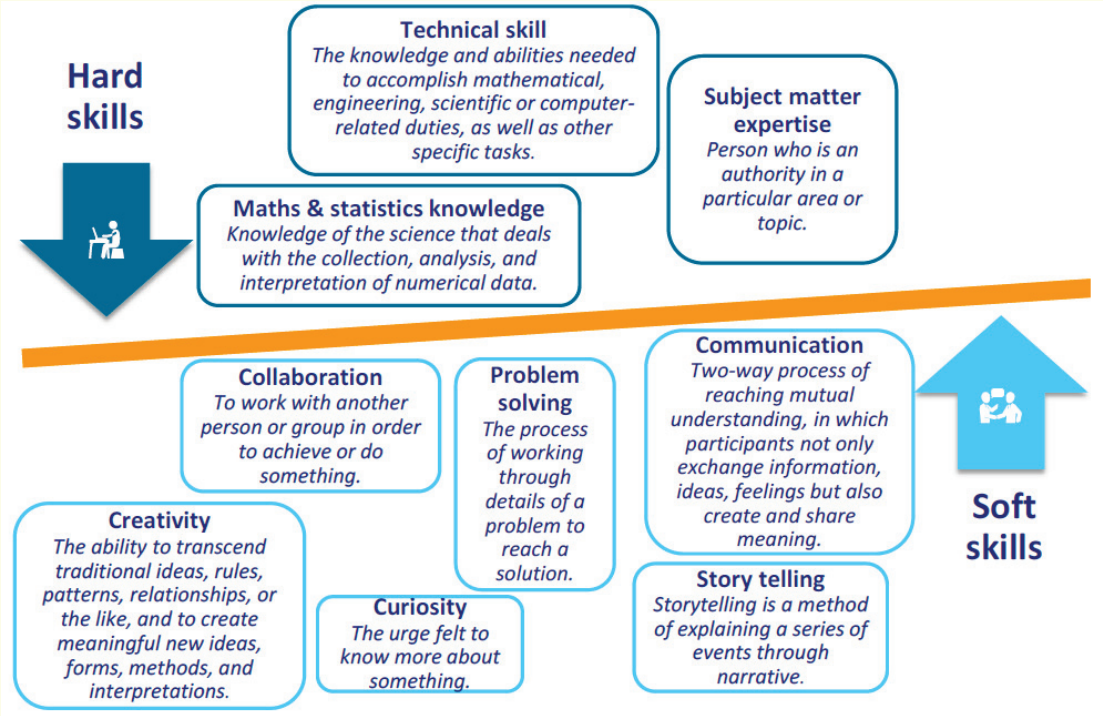
Published by:	Federal Statistical Office (FSO)
Information:	Bertrand Loison, FSO, tel. +41 58 463 67 70, bertrand.loison@bfs.admin.ch
Editor:	Bertrand Loison, FSO Diego Kuonen, Statoo Consulting
Series:	Swiss Statistics
Topic:	00 Statistical Basis and Overviews
Original text:	English
Translation:	FSO language services
Layout:	DIAM Section, Prepress/Print
Front page:	FSO; Concept: Netthoevel & Gaberthüel, Biel; Photograph: © vinnstock – Fotolia.com
Copyright:	FSO, Neuchâtel 2017 Reproduction with mention of source authorised (except for commercial purposes).
Print format orders:	Federal Statistical Office, CH-2010 Neuchâtel, tel. +41 58 463 60 60, fax +41 58 463 60 61, order@bfs.admin.ch
Price:	Free of charge
Downloads:	www.statistics.admin.ch (free of charge)
FSO number:	1790-1700

⇝ Available at `goo.gl/tW85FP` in English, German, French and Italian.

The focus of the strategy is to augment and/or complement existing basic official statistical production at the Swiss Federal Statistical Office (FSO) in the areas where data innovation (as defined below) makes sense.

> By understanding **analytics** as the science of learning from data (or of making sense of data), the FSO defines
>
> > **data innovation** as the application of *complementary analytics methods* (*e.g.* predictive analytics using approaches from advanced statistics, data science and/or machine learning) to existing (or traditional) and/or new (or non-traditional) data sources
>
> to sustain the role of official statistics in the democratic process in Switzerland by ensuring that the information we provide remains reliable, transparent and trustworthy.

‘Digital skills’ — ‘Hard and **soft skills** to work with data’



Source: European Commission, *Open Data Maturity in Europe 2016*, Figure 32, October 2016 (goo.gl/WPHR54).

‘We can not solve problems by using the same kind of thinking we used when we created them.’

Albert Einstein

⇝ Do not let culture eat strategy — have them feed each other!

⇝ Culture change is key in the digital transformation!

'The transformation can only be accomplished by man, not by hardware (computers, gadgets, automation, new machinery). A company can not buy its way into quality.'

W. Edwards Deming, 1982

'The only person who likes change is a wet baby.'

Mark Twain

‘It is not necessary to change. Survival is not mandatory.’

W. Edwards Deming

Another version by W. Edwards Deming: ‘Survival is not compulsory. Improvement is not compulsory. But improvement is necessary for survival.’

‘All improvement takes place project by project and in no other way.’

Joseph M. Juran, 1989

‘If you can not describe what you are doing as a process, you do not know what you are doing.’

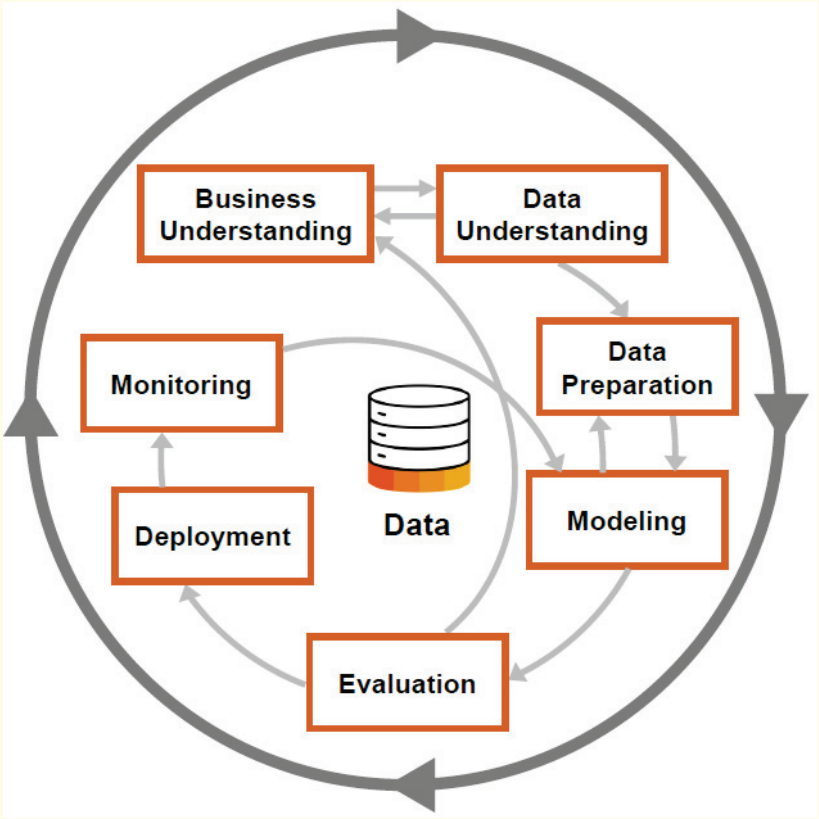
W. Edwards Deming

6. A process model for data-driven decision making

• Analytics is not only modelling and prediction, nor a product that can be bought, but a whole iterative problem solving and improvement cycle/process that must be mastered through interdisciplinary and transdisciplinary team effort.

$\rightsquigarrow$ The methodological framework called CRISP-DM (‘CRoss Industry Standard Process for Data Mining’ (now rebranded as data science) provides a comprehensive standard process for fitting analytics into the general problem solving and improvement strategy of a ‘business’.

$\rightsquigarrow$ CRISP-DM places a structure on the problem, allowing reasonable consistency, repeatability and objectiveness:



‘Coming together is a beginning. Keeping together is progress. Working together is success.’

Henry Ford

‘Analytics seen as a process of making soup in a soup bar’

◇ Phase 0: choose your ‘battles’ carefully, *i.e.* what do you want to do?

◇ Phase 1: project definition, *i.e.* understand the needs, priorities, desires and resources. ⇝ Take the order:

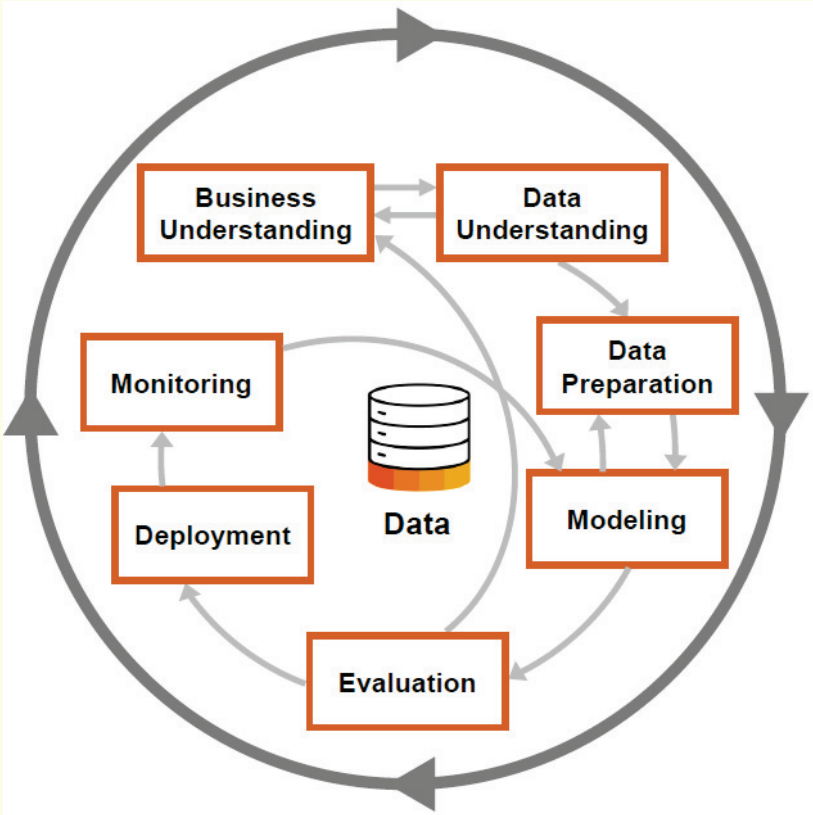
◇ Phase 2: data preparation, *i.e.* '*mise en place*':

◇ Phase 3: model building, *i.e.* cook the soup by choosing exactly those ingredients that blend into a 'great' result:

◇ Phase 4: model validation, *i.e.* taste the soup before it is served:

◇ Phase 5: model usage, *i.e.* present and professionally serve the soup:

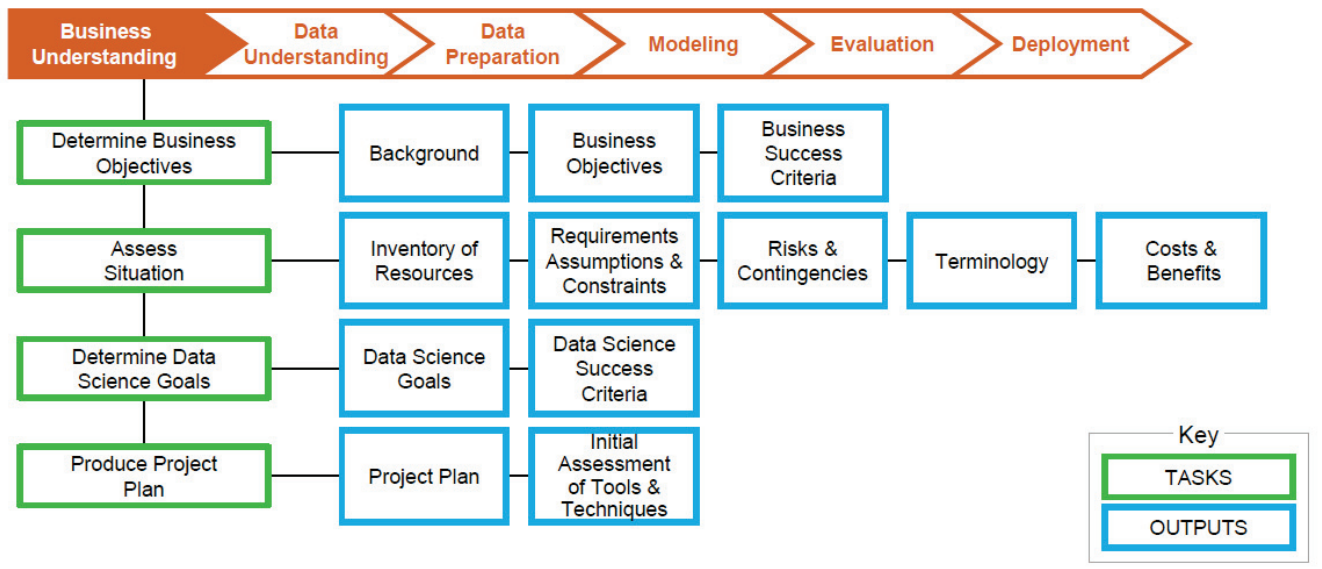
Source: Geert Verstraeten, 'Predictive analytics — A soup story', *LinkedIn*, August 27, 2015 (goo.gl/90HU42).



‘If I had only one hour to save the world, I would spend fifty-five minutes defining the problem, and only five minutes finding the solution.’

Albert Einstein

The ‘Business Understanding’ phase



The 'Determine Business Objectives' task

◇ Start gathering background information about the current 'business' situation:

- determine organisational structure, *e.g.* identify key individuals, identify an internal sponsor, identify 'business' units that will be affected;
- describe problem area, *e.g.* identify the problem area, describe the problem in general terms, clarify the prerequisites, status of the analytics project;
- describe current solution (if any).

◇ Define specific 'business' objectives, questions and any other 'business' requirements as precisely as possible, and specify expected benefits in 'business' terms.

◇ Agree upon objective and/or subjective criteria used to determine success from a 'business' perspective as precisely as possible ($\rightsquigarrow$ 'business' success criteria).

The 'Assess Situation' task

◇ Take an accurate inventory of resources available to the project, *e.g.* personnel, data sources, computing resources (hardware), software.

◇ Make an honest assessment of liabilities to the project:

- determine requirements, *e.g.* project scheduling requirements, requirements on results deployment;
- clarify assumptions, *e.g.* economic factors that might affect the project, data quality assumptions, expectations on how to 'view' the results;
- verify constraints, *e.g.* legal, financial, timescales, resources.

◇ Document each possible risk, *e.g.* 'business', scheduling, financial, data, results, and document a contingency plan for each risk.

◇ Ensure that 'business' and analytics teams are 'speaking the same language', *i.e.* keep a list of terms or jargon confusing to team members by including both 'business' and analytics terminology.

◇ Perform a costs and benefits analysis:

- estimate costs for data collection;
- estimate costs of developing and implementing a solution;
- identify benefits;
- estimate operation costs.

The 'Determine Data Science Goals' task

◇ Translate the 'business' goals into 'data science' goals which state project objectives in technical terms, and specify analytics tasks.

◇ Define the objective and/or subjective criteria for a successful outcome to the project in technical terms ($\rightsquigarrow$ 'data science' success criteria).

$\rightsquigarrow$ Document technical goals using specific units of, for example, time.

• Moreover, specify the requirements for the analytics models, *i.e.* algorithms, with respect to, for example, interpretability, reproducibility and stability, model flexibility and adequacy, run-time, interestingness and use of expert knowledge.

The 'Produce Project Plan' task

◇ Design preliminary project plan (including deployment) to achieve the objectives.

⇝ The project plan is the master document for all of the analytics work.

⇝ It can inform everyone associated with the project of the goals, resources, risks, and schedules for all phases of analytics.

◇ Initial assessment of tools and techniques for analytics most appropriate for the 'business' needs.

Ready for the next CRISP-DM phase?

- Before continuing, be sure you have answered also the following key questions.
  - What is the unit of analysis, *i.e.* the major entity that you are analysing and modelling — that is, the 'who' and 'what'?
  - What is the population of interest, *i.e.* the collection of 'rows' that will form the data set you will analyse and model?
  - Is the data you consider representative of, for example, your 'customers' and 'prospects'?
  - What are the time frames or periods you will consider, *i.e.* the time horizon for prediction (⇝ How far into the future?), the time window of relevant behaviour (⇝ 'How far to 'look' back?) and the time base of the population?

- Three key issues to consider when exploring observational data include:
  - the **relevance of the sample** to the current study;

    ⇝ Are there differences in time, product, service or process that may change patterns or relationships?
  - the **quality of the data**;

    ⇝ How were the 'units of the analysis' selected?

    ⇝ Have all potentially relevant process inputs, *i.e.* the drivers (factors) influencing a given KPI, been gathered?

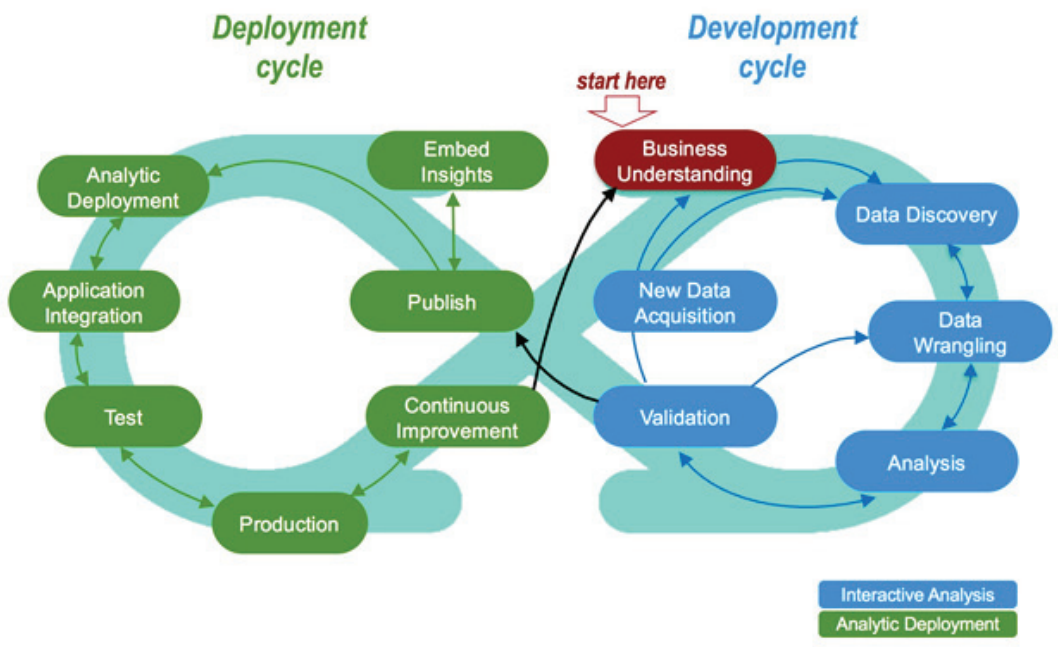
    ⇝ Do you understand the processes and system that generated the data?

    ⇝ Were operational definitions (for process inputs and outputs, *i.e.* the KPIs) used to collect the data?
  - **how can the data be used** to advance the study without drawing conclusions that are not justifiable, given the lack of established causality?

'The error of the third kind is the error committed by giving the right answer to the wrong problem.'

Allyn W. Kimball, 1957

The complementary cycles of developing & deploying 'analytical assets'



Source: Erick Brethenoux, Director, IBM Analytics Strategy & Initiatives, August 18, 2016 (goo.gl/AhsG1n).

'The key to success is to make sure that the beginning and ending steps of the analysis are well thought out.'

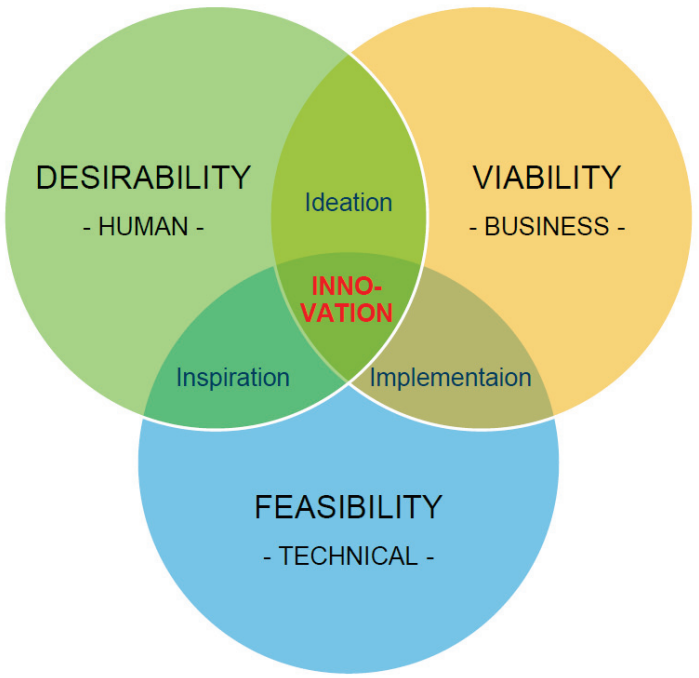
Thomas H. Davenport and Jinho Kim, 2013

Reflections on choosing 'pilot projects' ('use cases')

◇ Develop the following information:

- problem definition: clear statement of the problem, projected impact, ...;
- create structure: objectives, constraints, metrics for success (*i.e.* the KPIs), potential drivers for these metrics, related operational definitions, ...;
- context: identify important stakeholders (*e.g.* customers, organisations, individuals, management), issue relevant background, sources of relevant data, practical restrictions, ...;
- strategy: critical elements of an overall, high level approach to attacking the problem, based on the problem's structure and context;
- establish tactics on how to implement the strategy.

From an application point of view...



‘What we have to learn to do, we learn by doing.’

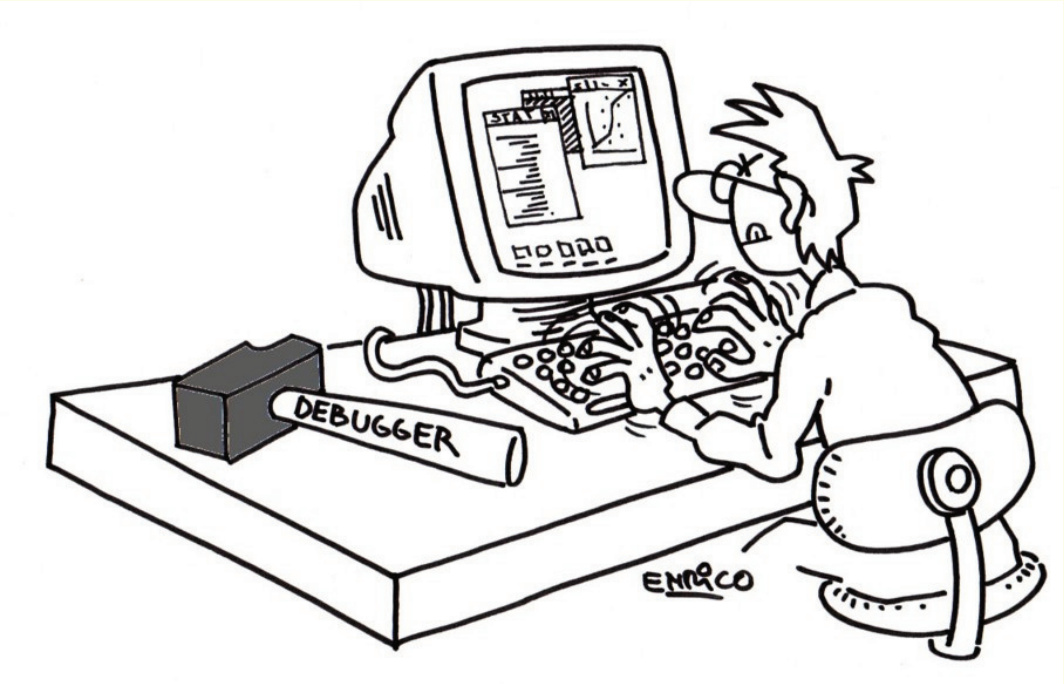
Aristotle

‘One must learn by doing the thing; for though you think you know it, you have no certainty until you try.’

Sophocles

‘It is getting better. . . A little better all the time.’

The Beatles, 1967



Have you been Statooed?

Prof. Dr. Diego Kuonen, CStat PStat CSci

Statoo Consulting

Morgenstrasse 129

3018 Berne

Switzerland

email kuonen@statoo.com

@DiegoKuonen

web www.statoo.info

Presentation code: 'AdNovum/Bern.Aug23.2018'.

Typesetting: LaTeX, version 2ε. PDF producer: pdfTeX, version 3.141592-1.40.3-2.2 (Web2C 7.5.6).

Compilation date: 20.08.2018.