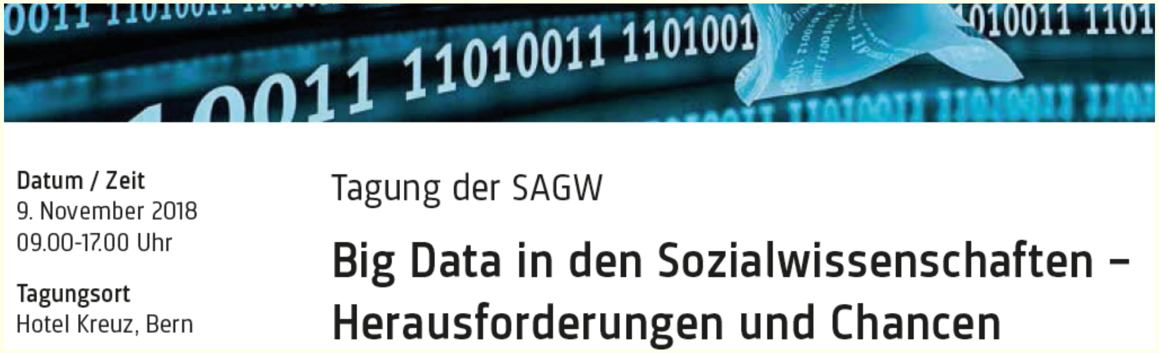
A Statistician's 'Big Tent' View on the Veracity and the Trustworthiness of Data

Demystification, Challenges, Opportunities and Principles for Success

Prof. Dr. Diego Kuonen, CStat PStat CSci

Statoo Consulting, Berne & GSEM, University of Geneva, Switzerland

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



Datum / Zeit
9. November 2018
09.00-17.00 Uhr

Tagungsort
Hotel Kreuz, Bern

Tagung der SAGW

Big Data in den Sozialwissenschaften – Herausforderungen und Chancen

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).

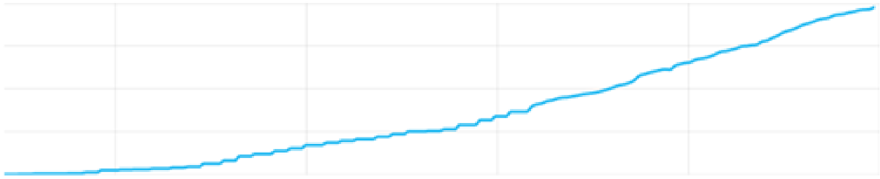
▷ Founder, CEO & CAO, Statoo Consulting, Switzerland (since 2001).

▷ 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).

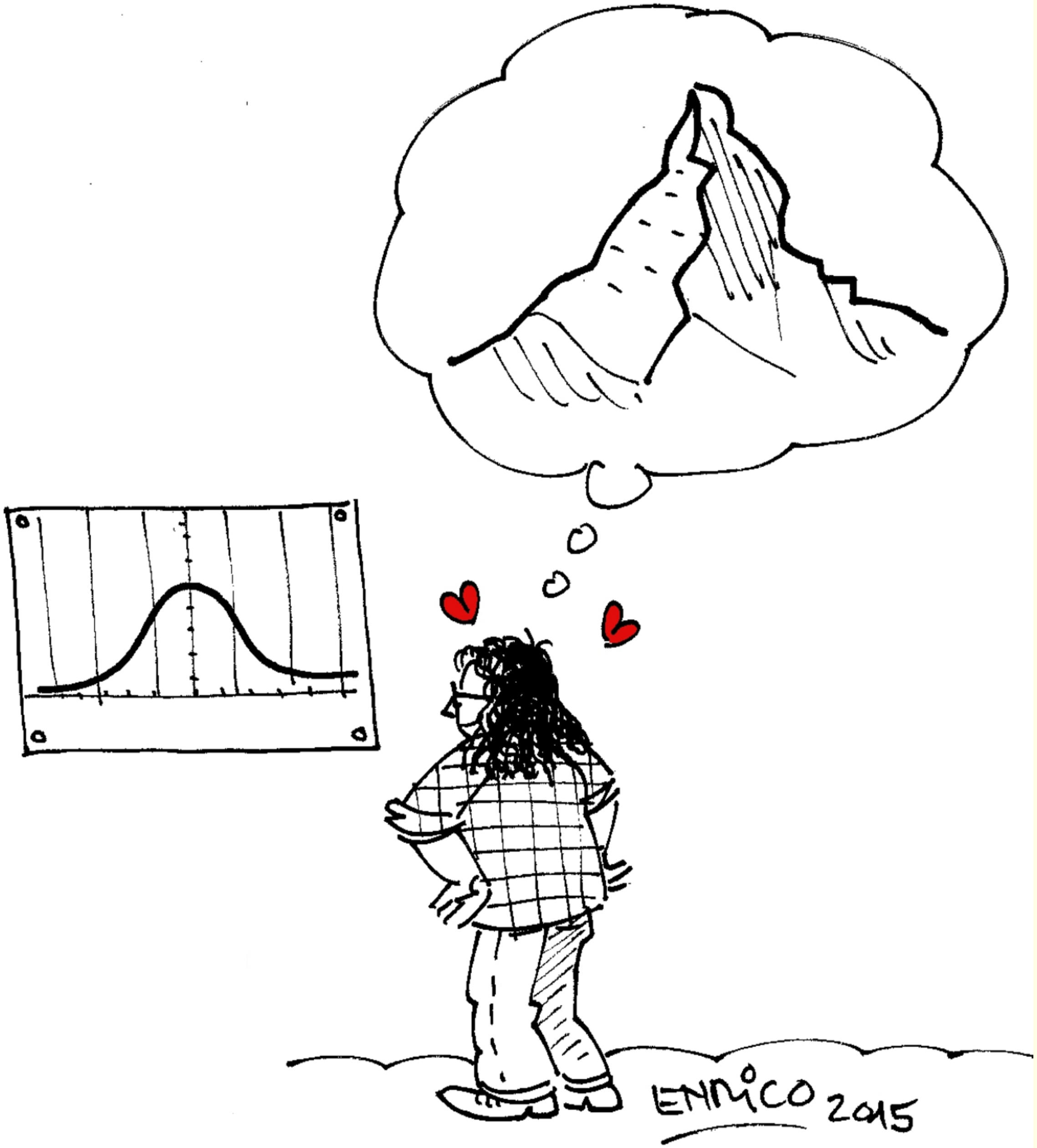
▷ Principal Scientific and Strategic Big Data Analytics Advisor for the Directorate and Board of Management, Swiss Federal Statistical Office (FSO), Neuchâtel, Switzerland (since 2016).

@DiegoKuonen



- 30.11.2013: 3 followers
- 08.11.2018: 18’555

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING
TOP 100 INFLUENCERS AND BRANDS


ENRICO 2015

‘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

Demystifying the ‘big data’ hype

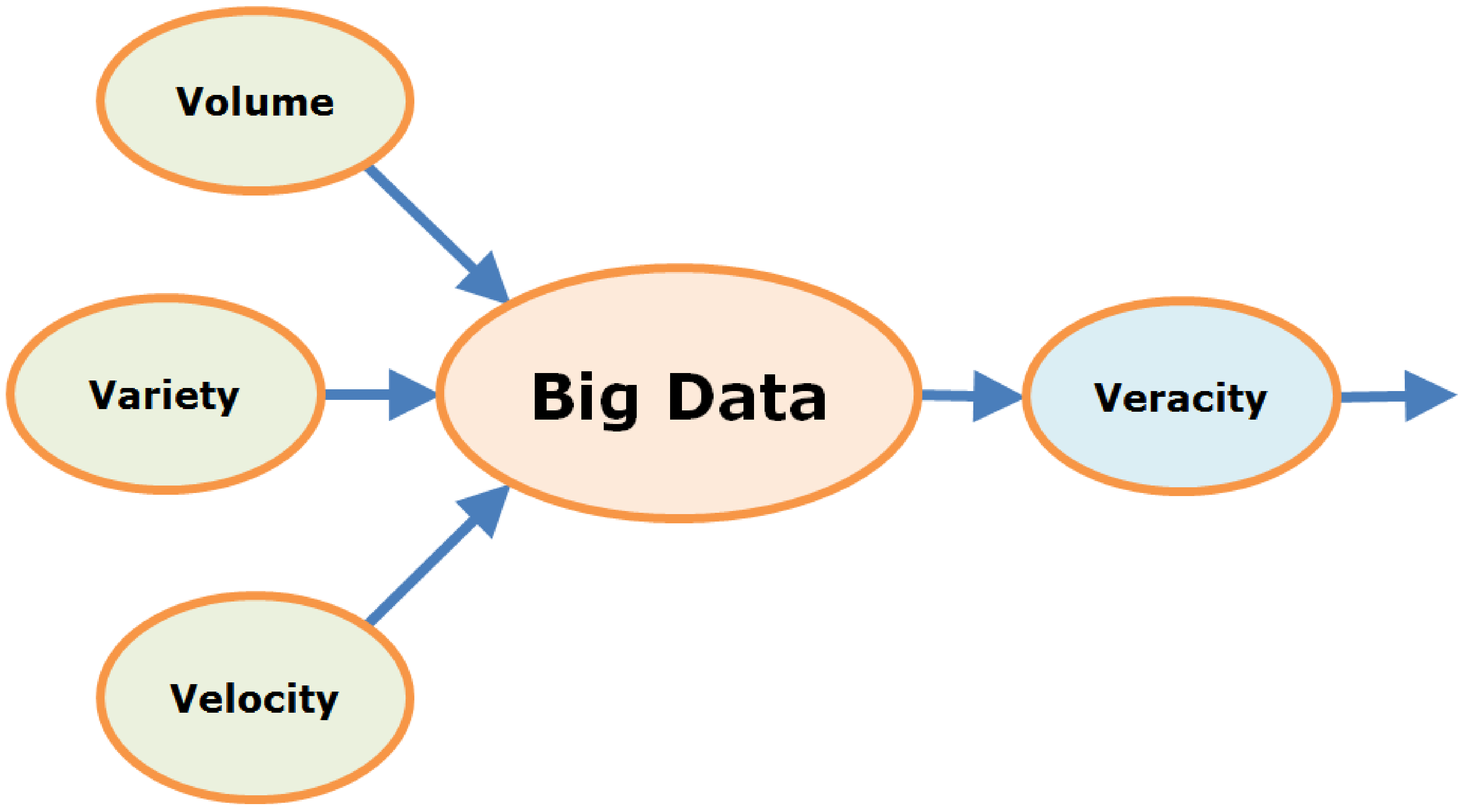
- ‘Big data’ have hit the business, government and scientific sectors.

⇝ The term ‘big data’ — coined in 1997 by two researchers at the NASA — has acquired the trappings of a ‘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 IT 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’ (⇝ ‘data-driven decision making’ and ‘data-informed policy making’).



⇝ The ‘Veracity’ (*i.e.* ‘trust in data’), including the reliability (‘quality over time’), capability and validity of the data, and the related quality of the data are key!

⇝ Existing ‘small’ data quality frameworks need to be extended, *i.e.* augmented!

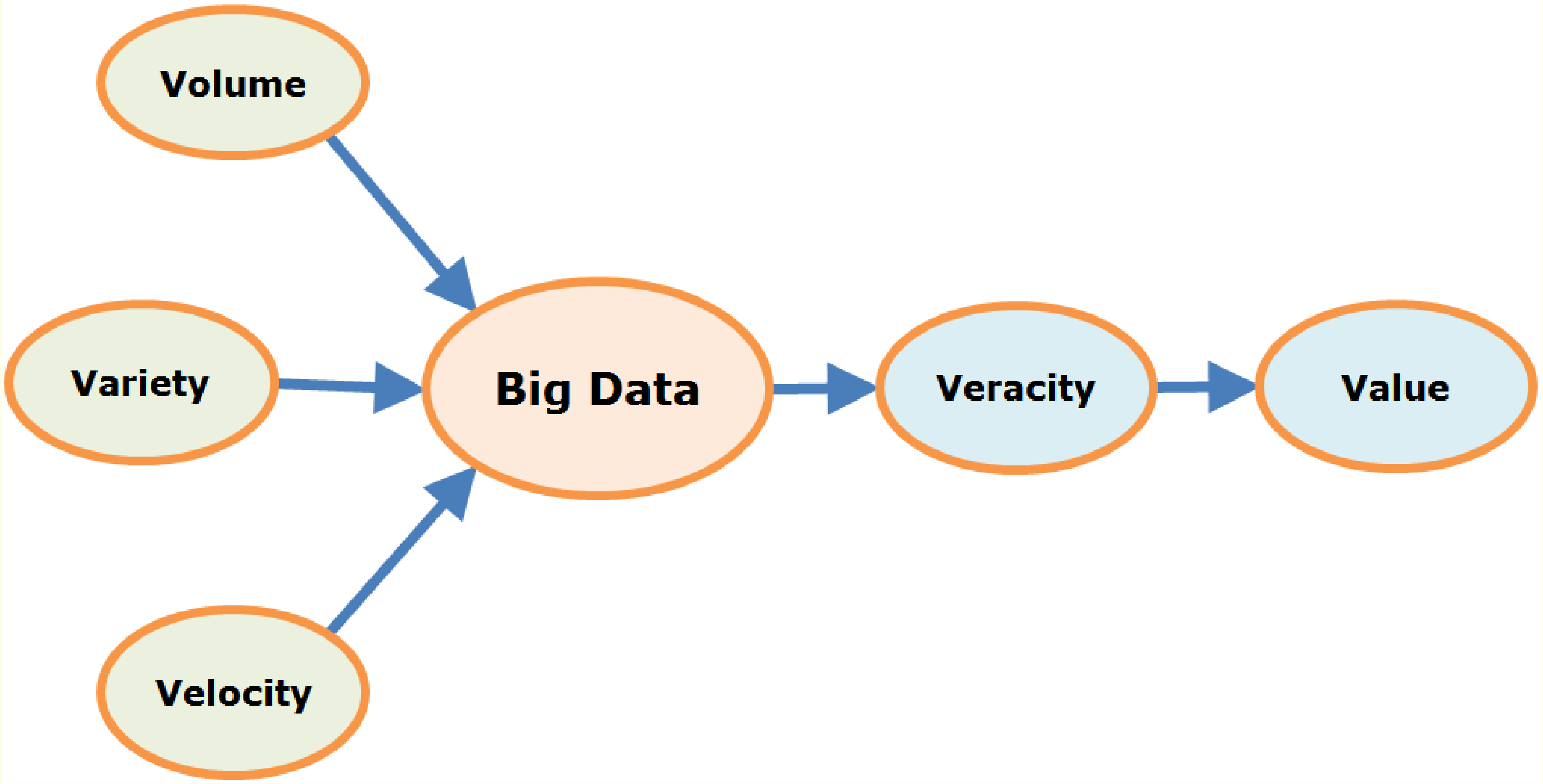
‘Data is part of Switzerland’s infrastructure, such as road, railways and power networks, and is of great value. The government and the economy are obliged to generate added value from these data.’

digitalswitzerland, November 22, 2016

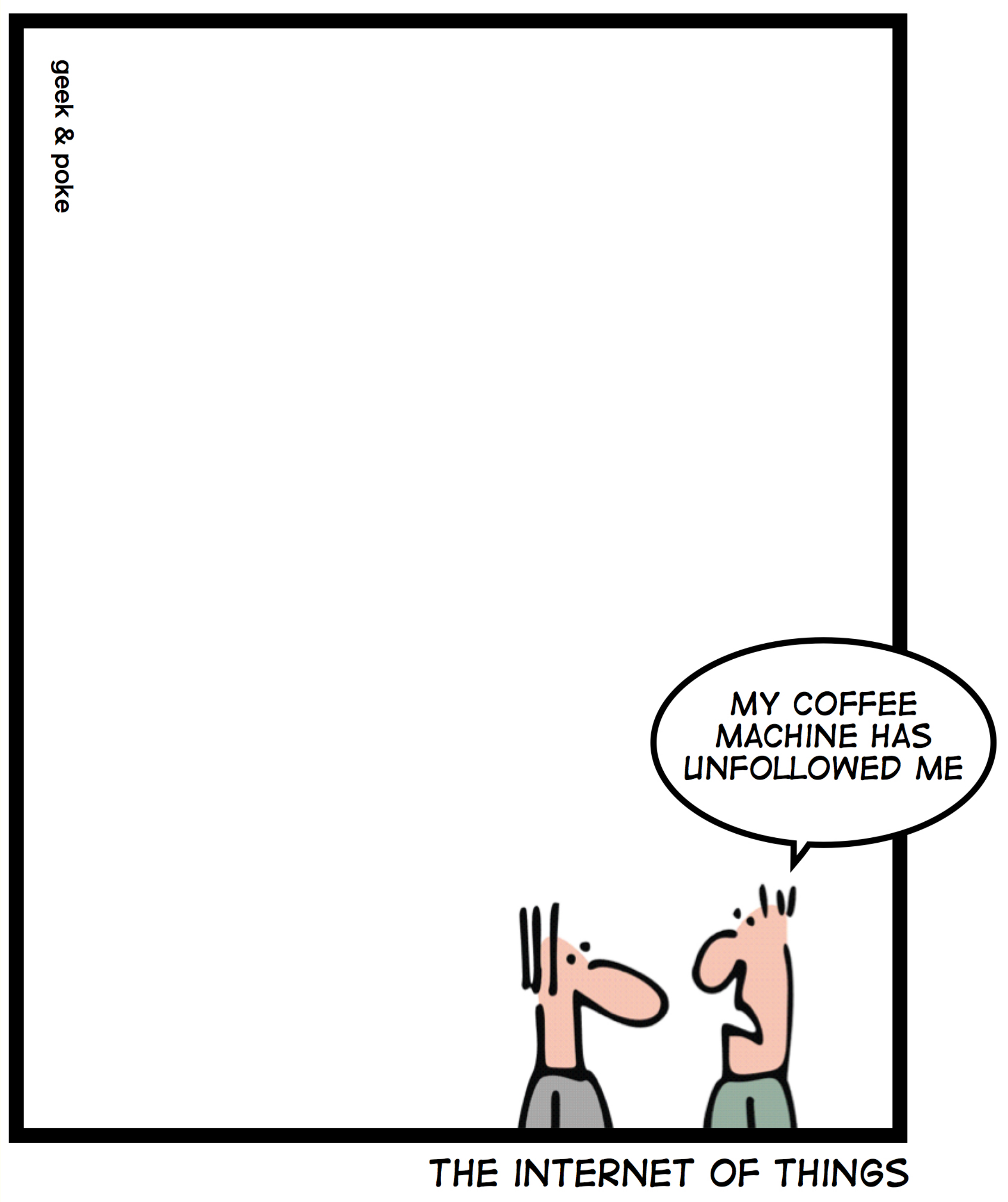
Source: digitalswitzerland’s ‘Digital Manifesto for Switzerland’ (`digitalswitzerland.com`).

⇝ The 5th V of big data: $\boxed{\text{‘Value’}}$, *i.e.* the ‘usefulness of data’.

Intermediate summary: the 'five Vs' of (big) data

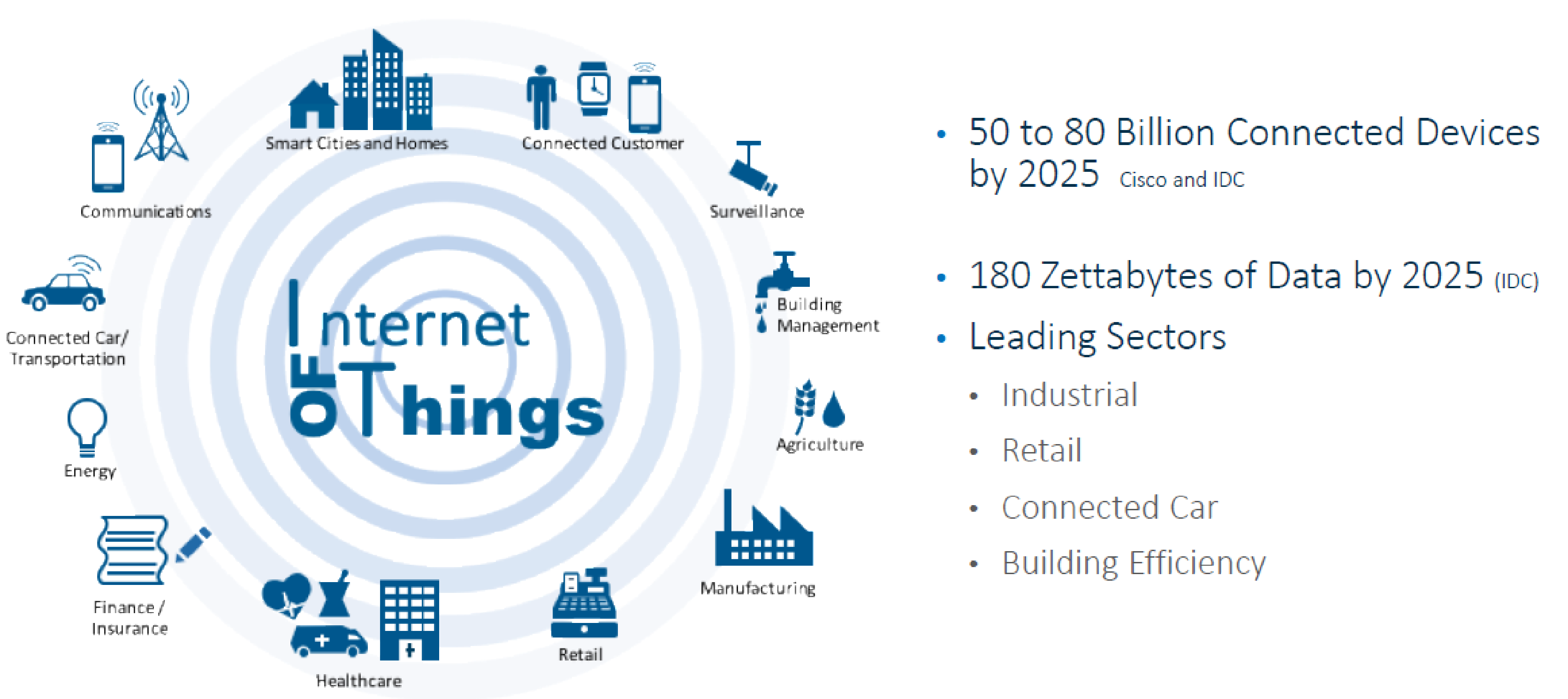


- ◇ 'Volume', 'Variety' and 'Velocity' are the '<u>essential</u>' characteristics of (big) data;
- ◇ 'Veracity' and 'Value' are the '<u>qualification for use</u>' characteristics of (big) data.


geek & poke
MY COFFEE MACHINE HAS UNFOLLOWED ME
THE INTERNET OF THINGS

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’!

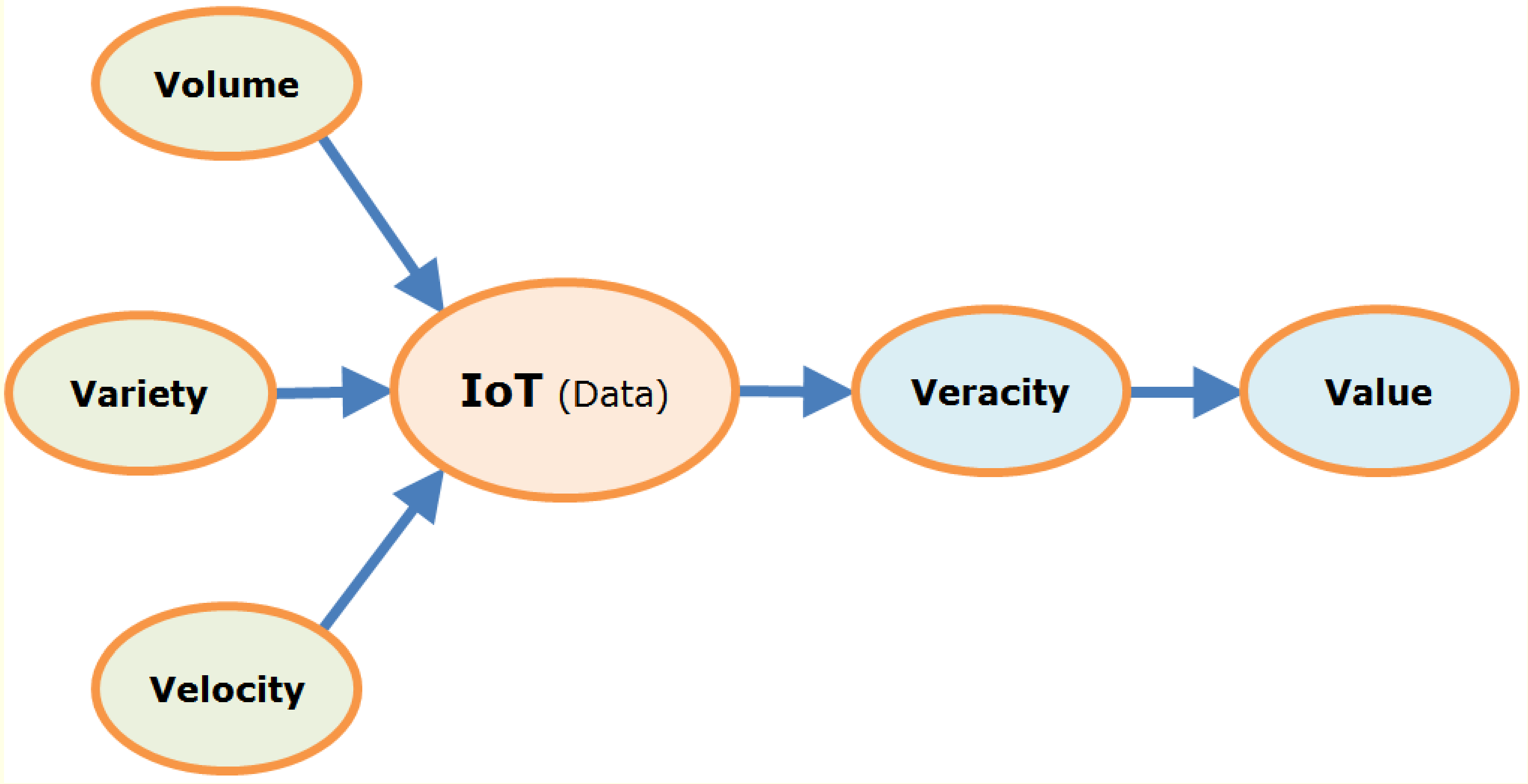


- 50 to 80 Billion Connected Devices by 2025 Cisco and IDC
- 180 Zettabytes of Data by 2025 (IDC)
- Leading Sectors
  - Industrial
  - Retail
  - Connected Car
  - Building Efficiency

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

⇝ However, IoT is about data, not things!

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).

> ‘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

⇝ **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**!

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) for statistical purposes to explain and check the validity of specific existing 'ideas' ('hypotheses'), *i.e.* through the operationalisation of theoretical concepts.

⇝ Primary analytics or **top-down** (*i.e.* explanatory and confirmatory) analytics.

⇝ 'Idea (hypothesis) evaluation or testing'.

⇝ Analytics' paradigm: '**deductive reasoning**' as 'idea (theory) first'.

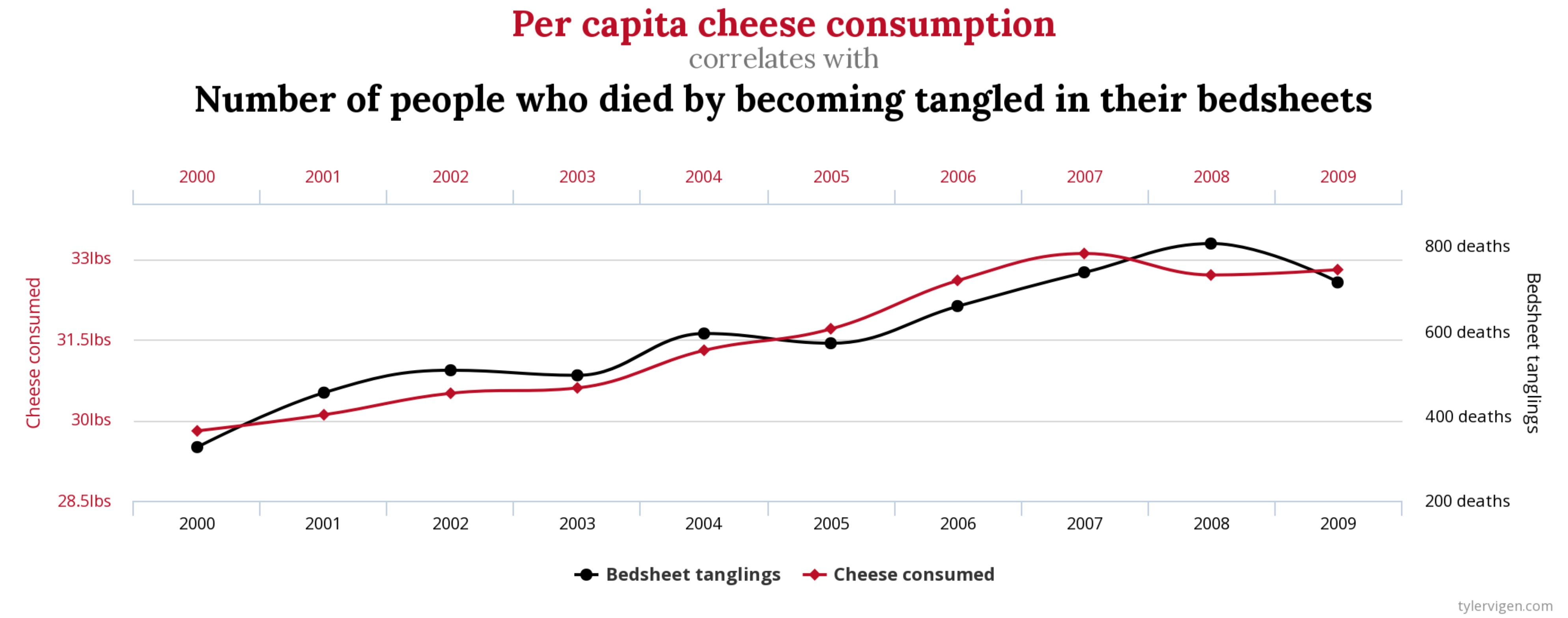
◇ Data science — a rebranding of 'data mining' and as a term coined in 1997 by a statistician — on the other hand, typically is concerned with analysing **secondary** (*e.g.* observational or 'found' or 'organic' or 'convenience') **data** that have been collected (and designed) for other reasons (and 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'.

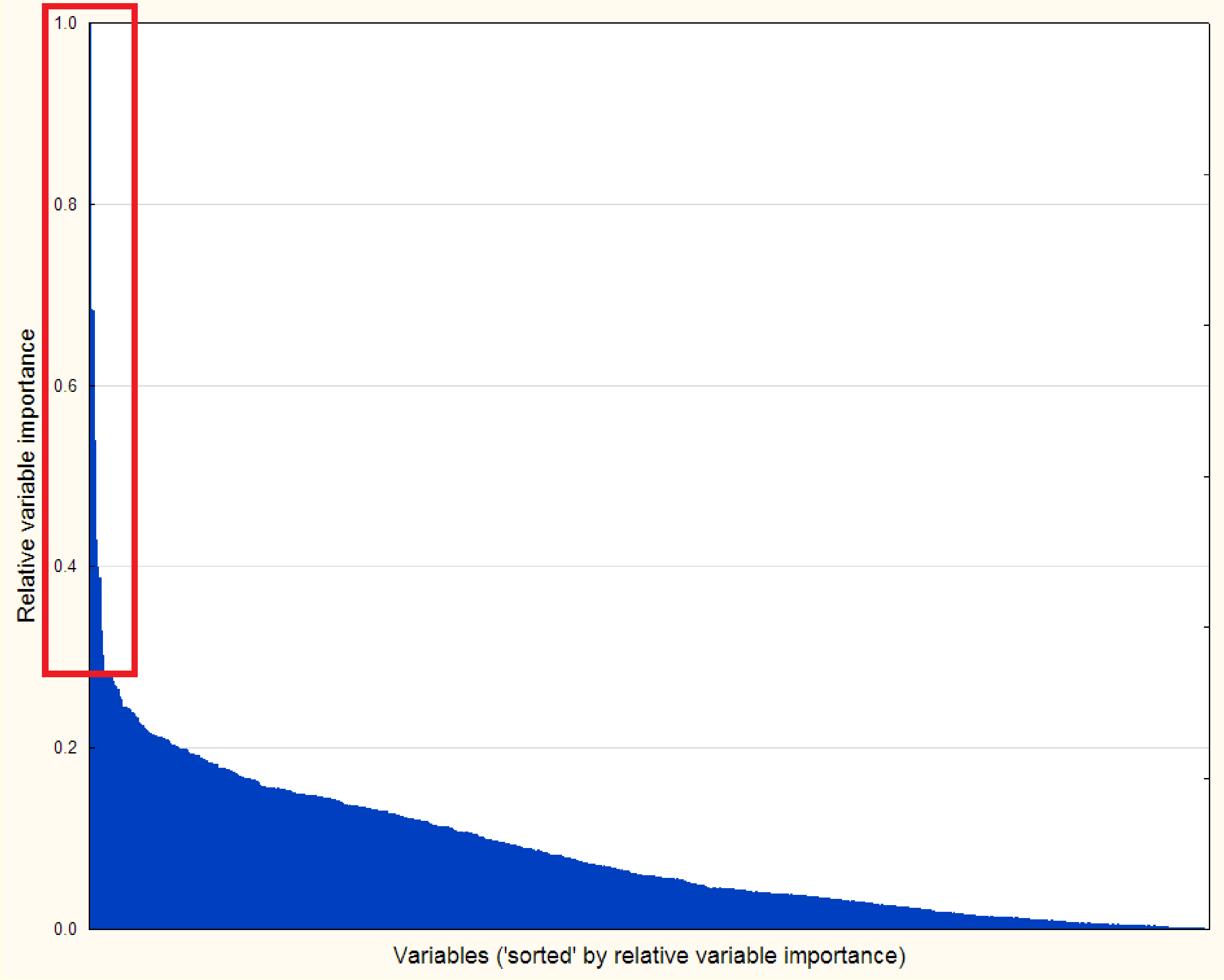
‘Spurious correlation is not causation!’



‘Any claim coming from an observational study is most likely to be wrong.’

S. Stanley Young and Alan Karr, 2011

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

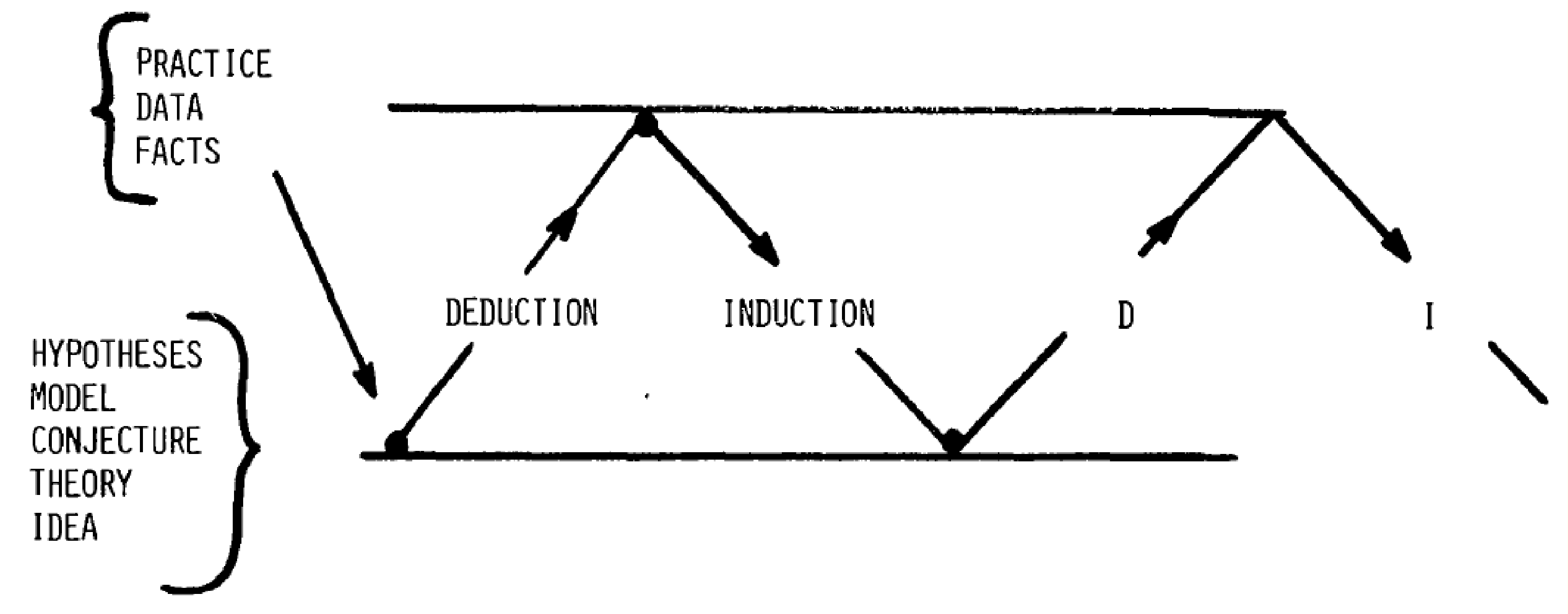


‘Neither exploratory nor confirmatory is sufficient alone. To try to replace either by the other is madness. We need them both.’

John W. Tukey, 1980

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

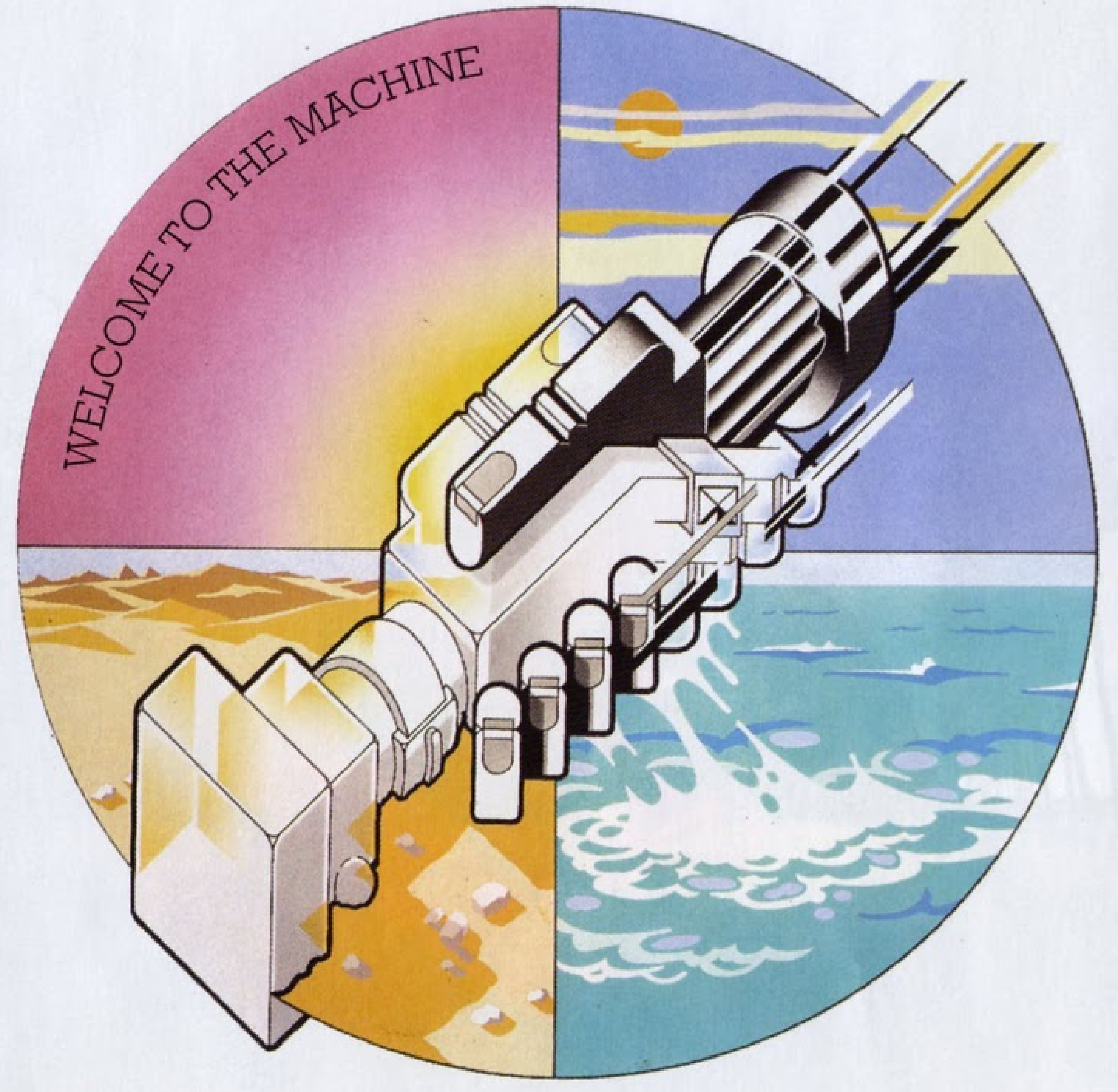
‘Experiments may be conducted sequentially so that each set may be designed using the knowledge gained from the previous sets.’

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

⇝ Scientific investigation is a sequential learning process!

⇝ Statistical methods allow investigators to accumulate knowledge!

‘Welcome to the Machine’ (Pink Floyd, 1975)

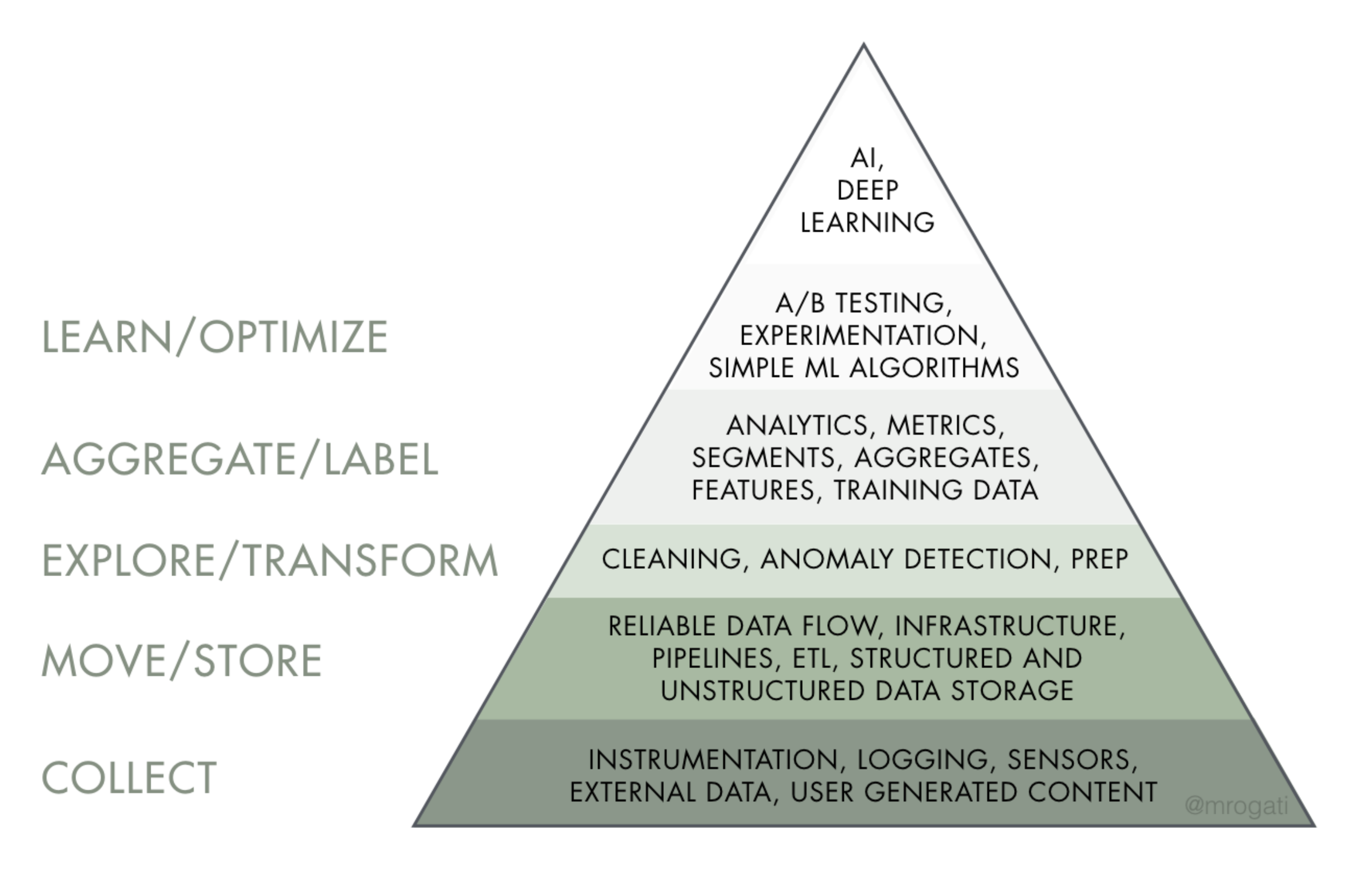


‘Artificial intelligence algorithms are not natively ‘intelligent’. They learn inductively by analyzing data. … Sophisticated algorithms can sometimes overcome limited data if its quality is high, but bad data is simply paralyzing.’

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).

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


LEARN/OPTIMIZE
AGGREGATE/LABEL
EXPLORE/TRANSFORM
MOVE/STORE
COLLECT
AI, DEEP LEARNING
A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS
ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA
CLEANING, ANOMALY DETECTION, PREP
RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE
INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT
@mrogati

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

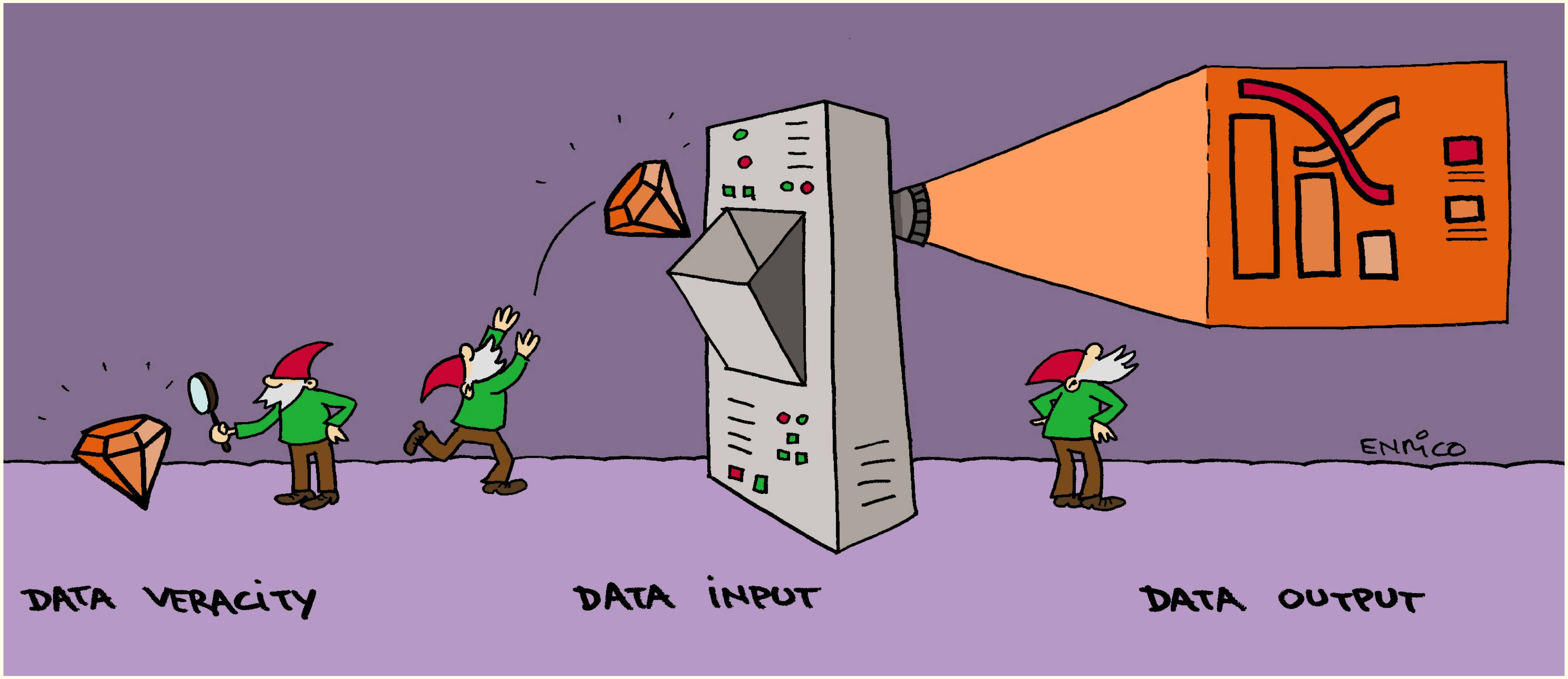
‘Data themselves are a central raw material of the knowledge society. However, this means that the data must be of high quality, accessible and trustworthy.’

Swiss Federal Council, September 5, 2018

Source: ‘*Digitale Schweiz*’ strategy, adopted by the Federal Council on September 5, 2018 (`goo.gl/b7K8aE`).

Challenges, opportunities and principles for success

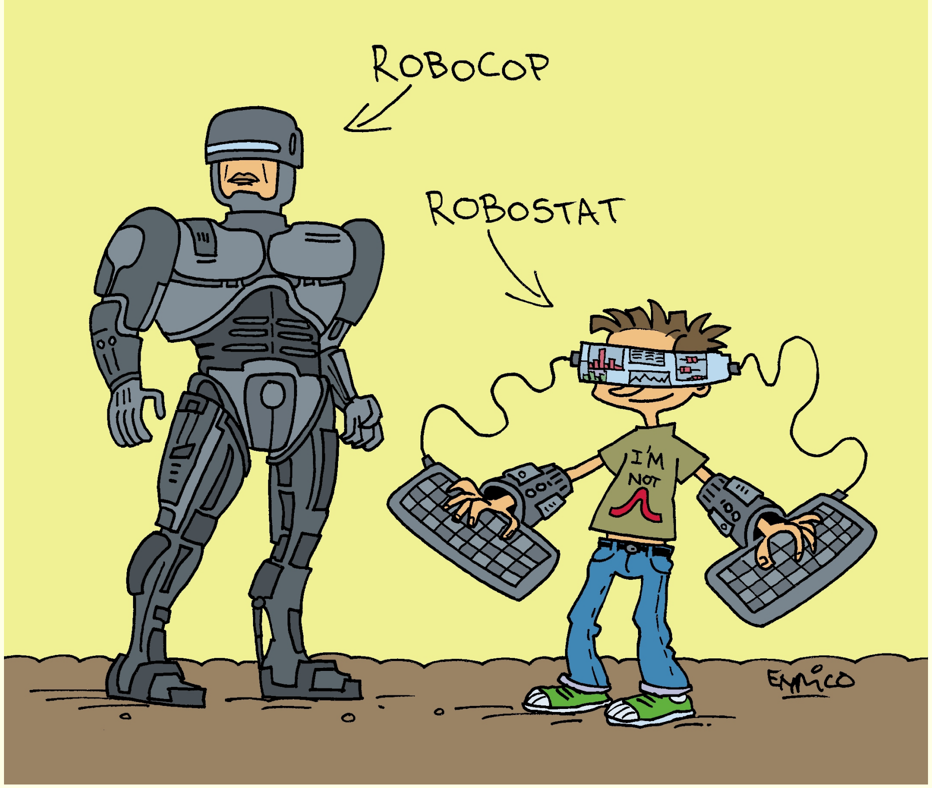
- In a world of (big) data, IoT (data) and also post-truth politics, the veracity of data, *i.e.* the trustworthiness of data (including the related data quality), is more important than ever!



- 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, not replacements for it!



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 (⇝ 'continuous improvement');
  - having a strategy for the conduct of the data analysis; including thought about the 'business' objectives (⇝ 'strategic thinking');
  - carefully considering data quality and assessing the 'data pedigree' before, during and after the data analysis; and
  - applying sound subject matter knowledge ('domain knowledge' or 'business knowledge', *i.e.* knowing the 'business' context and process 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

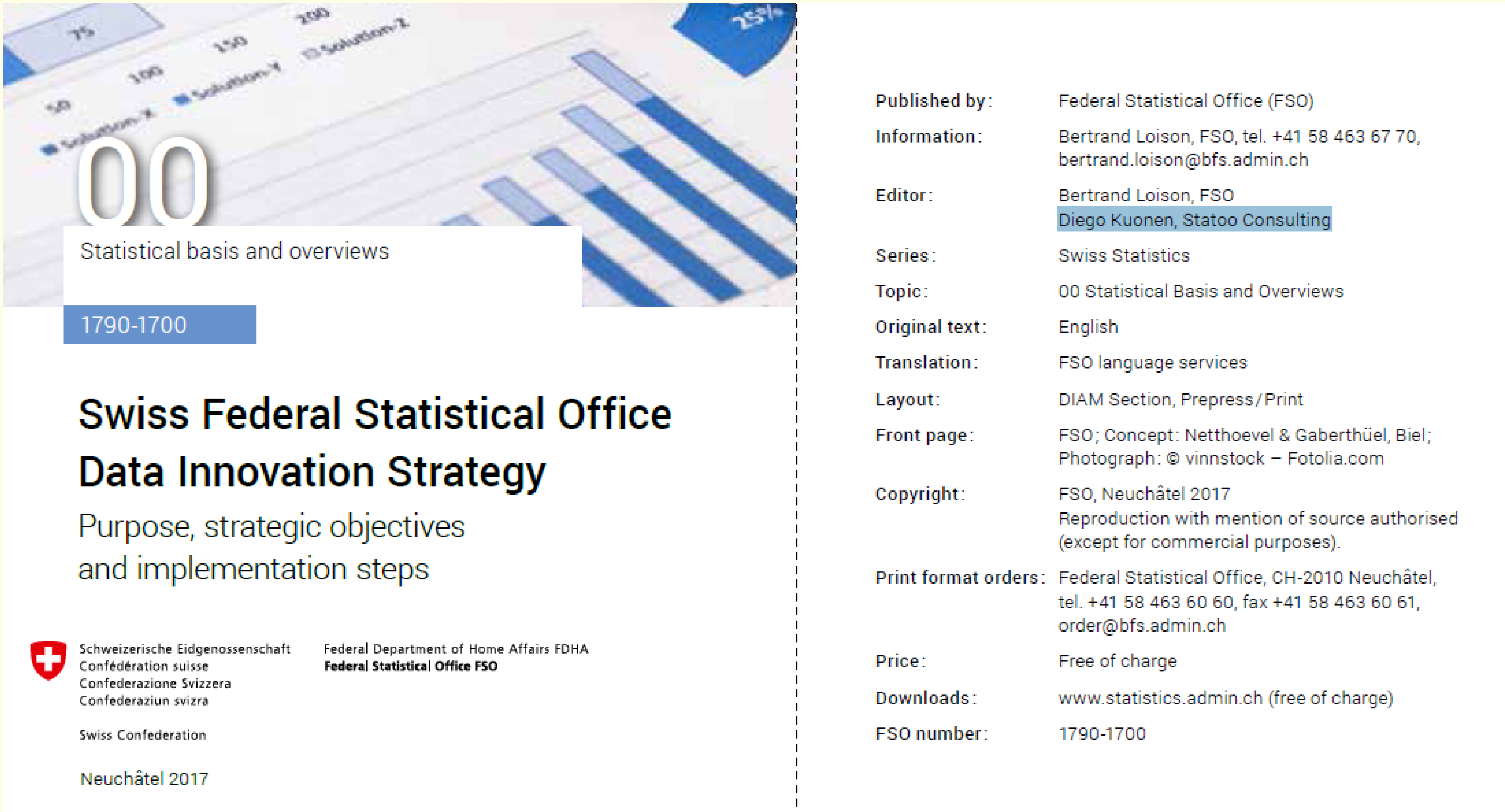
‘You do not need a digital strategy. You need a better *(‘business’)* strategy, enabled by digital. ’

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`).

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

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



00

Statistical basis and overviews

1790-1700

Swiss Federal Statistical Office Data Innovation Strategy

Purpose, strategic objectives
and implementation steps

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

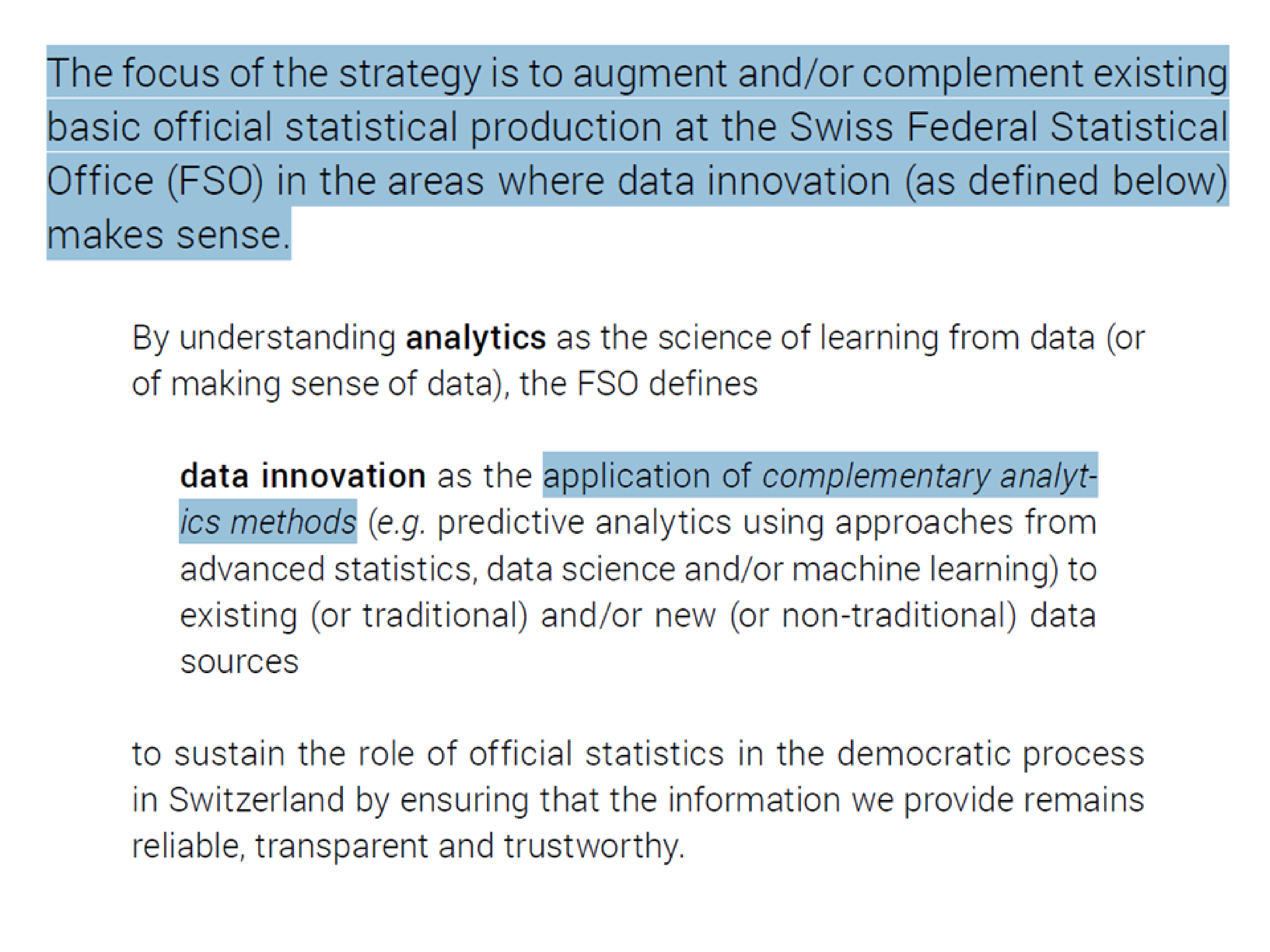
Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Statistical Office FSO

Neuchâtel 2017

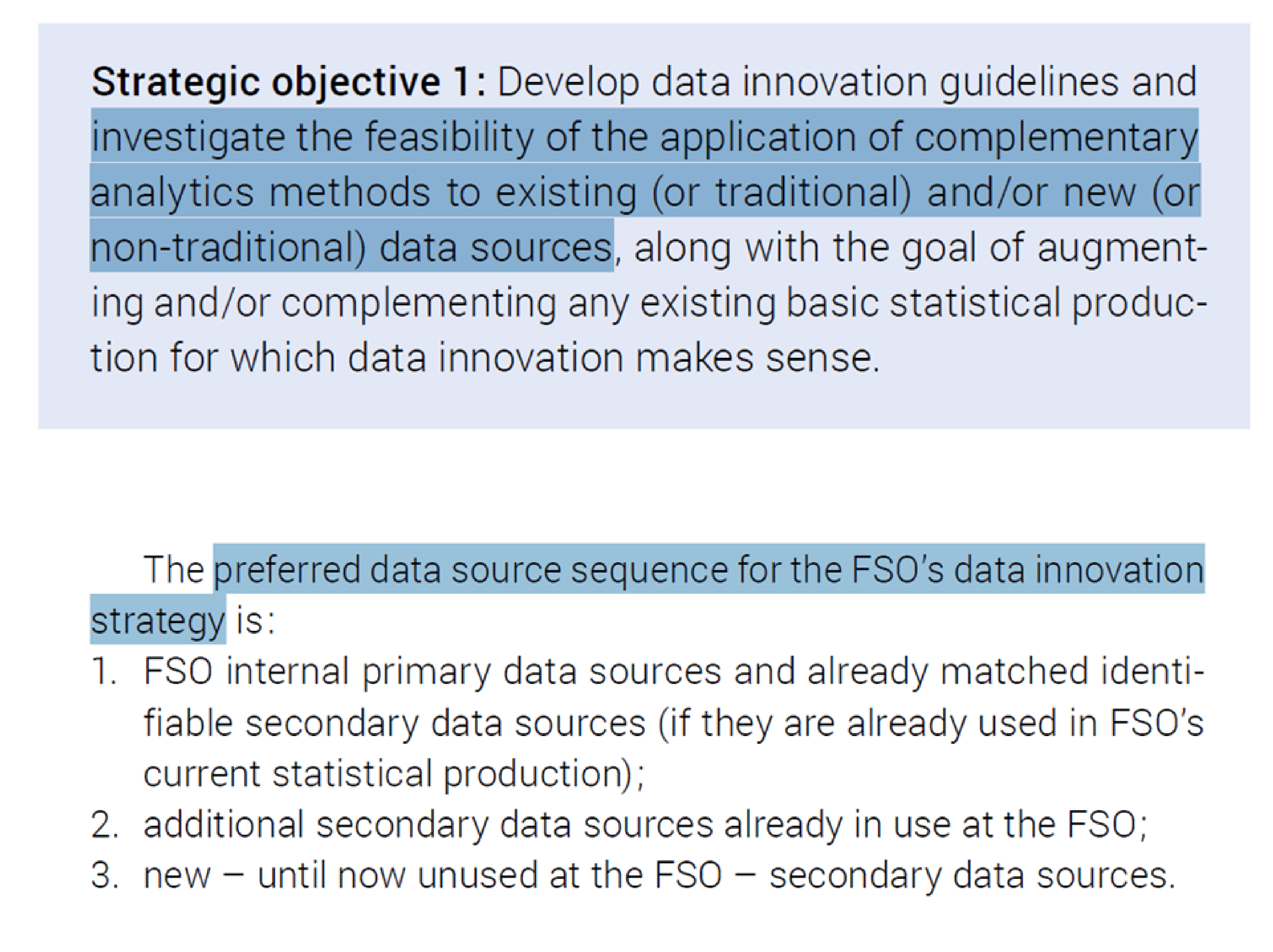
Published by:	Federal Statistical Office (FSO)
Information:	Bertrand Loison, FSO, tel. +41 58 463 67 70, bertrand.loison@bfs.admin.ch
Editor:	Bertrand Loison, FSO Diego Kuonen, Statoo Consulting
Series:	Swiss Statistics
Topic:	00 Statistical Basis and Overviews
Original text:	English
Translation:	FSO 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.



Strategic objective 1: Develop data innovation guidelines and investigate the feasibility of the application of complementary analytics methods to existing (or traditional) and/or new (or non-traditional) data sources, along with the goal of augmenting and/or complementing any existing basic statistical production for which data innovation makes sense.

The preferred data source sequence for the FSO's data innovation strategy is:

1. FSO internal primary data sources and already matched identifiable secondary data sources (if they are already used in FSO's current statistical production);
2. additional secondary data sources already in use at the FSO;
3. new – until now unused at the FSO – secondary data sources.

DR
ENRICO

‘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

Have you been Statooed & GSEMed?

Prof. Dr. Diego Kuonen, CStat PStat CSci

Statoo Consulting
Morgenstrasse 129
3018 Berne
Switzerland

GSEM, University of Geneva
Bd du Pont-d'Arve 40
1211 Geneva 4

email	`kuonen@statoo.com`	`Diego.Kuonen@unige.ch`
web	`www.statoo.info`	`gsem.unige.ch/rcs/kuonen`
🐦	`@DiegoKuonen`	

Presentation code: 'SAGW.BigData.Nov9.2018'. Typesetting: LaTeX, version 2ϵ. PDF producer: pdfTeX, version 3.141592-1.40.3-2.2 (Web2C 7.5.6).
Compilation date: 08.11.2018.