Veracity and Trustworthiness of Data in the Context of Evaluation

Demystification, Challenges and Opportunities

Prof. Dr. Diego Kuonen, CStat PStat

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

kuonen@statoo.com + @DiegoKuonen + Diego.Kuonen@unige.ch

SEVAL

Schweizerische Evaluationsgesellschaft
Société suisse d'évaluation
Società svizzera di valutazione

Keynote @ 'Congrès 2022 de la SEVAL', Fribourg, Switzerland — September 2, 2022

About myself (about.me/DiegoKuonen)

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

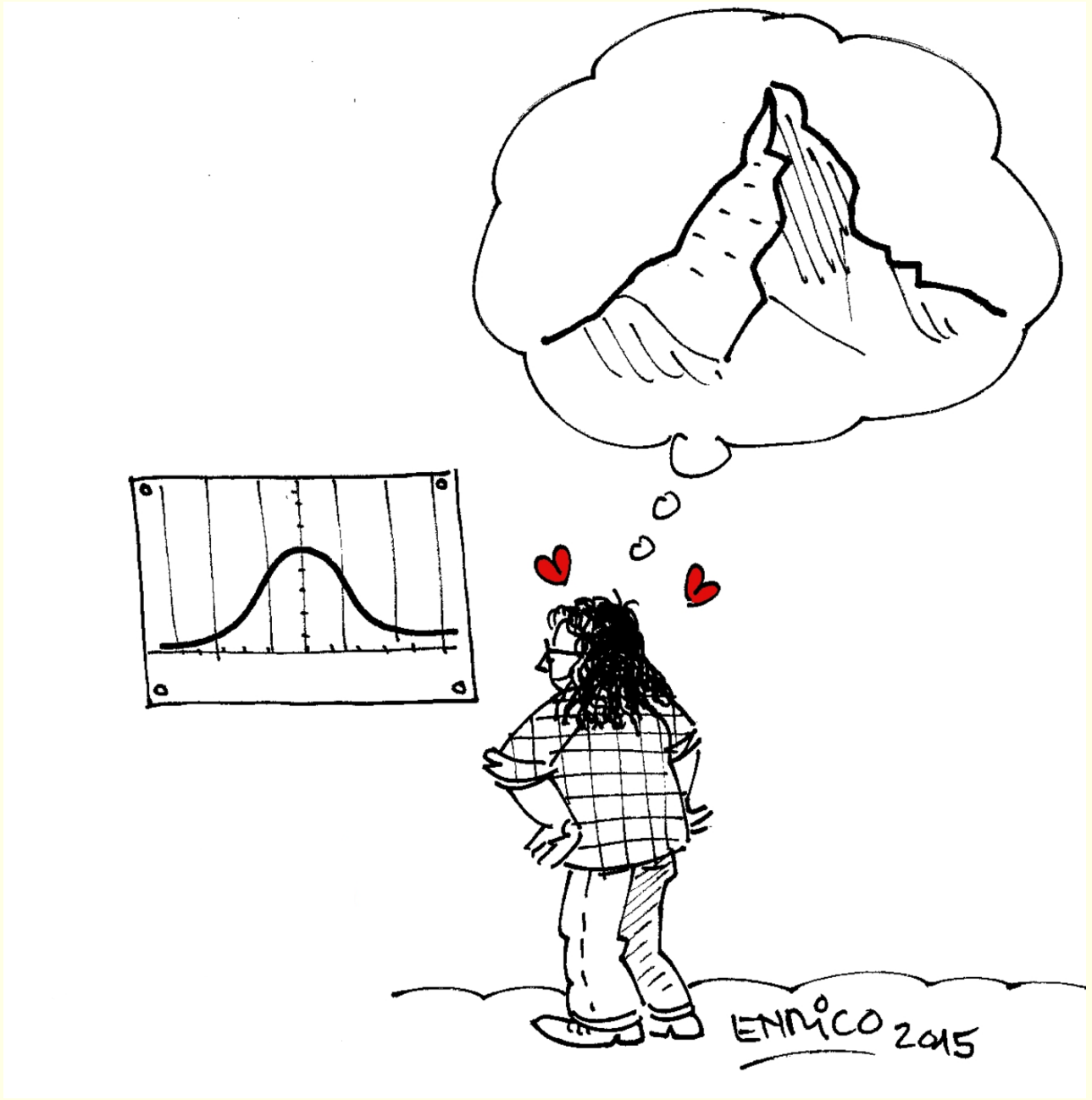
- CStat ('Chartered Statistician'), Royal Statistical Society, UK.
- PStat ('Accredited Professional Statistician'), American Statistical Association, USA.

- 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 & Data Science Advisor and Consultant for the Directorate and the Board of Management, Swiss Federal Statistical Office (FSO), Neuchâtel, Switzerland (since 2016).



‘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

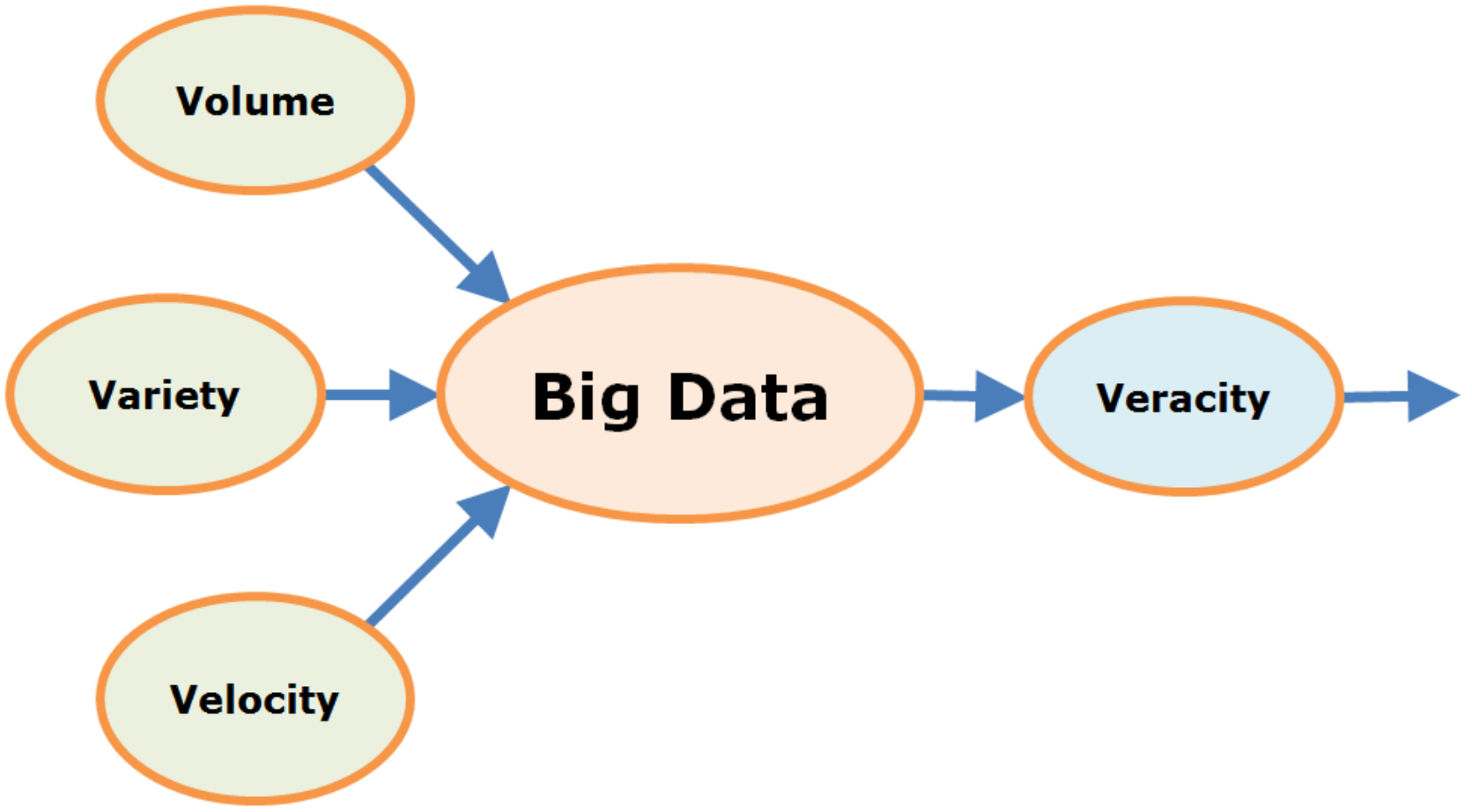
1. Demystifying the 'big data' hype

• 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').



$\rightsquigarrow$ 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!

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

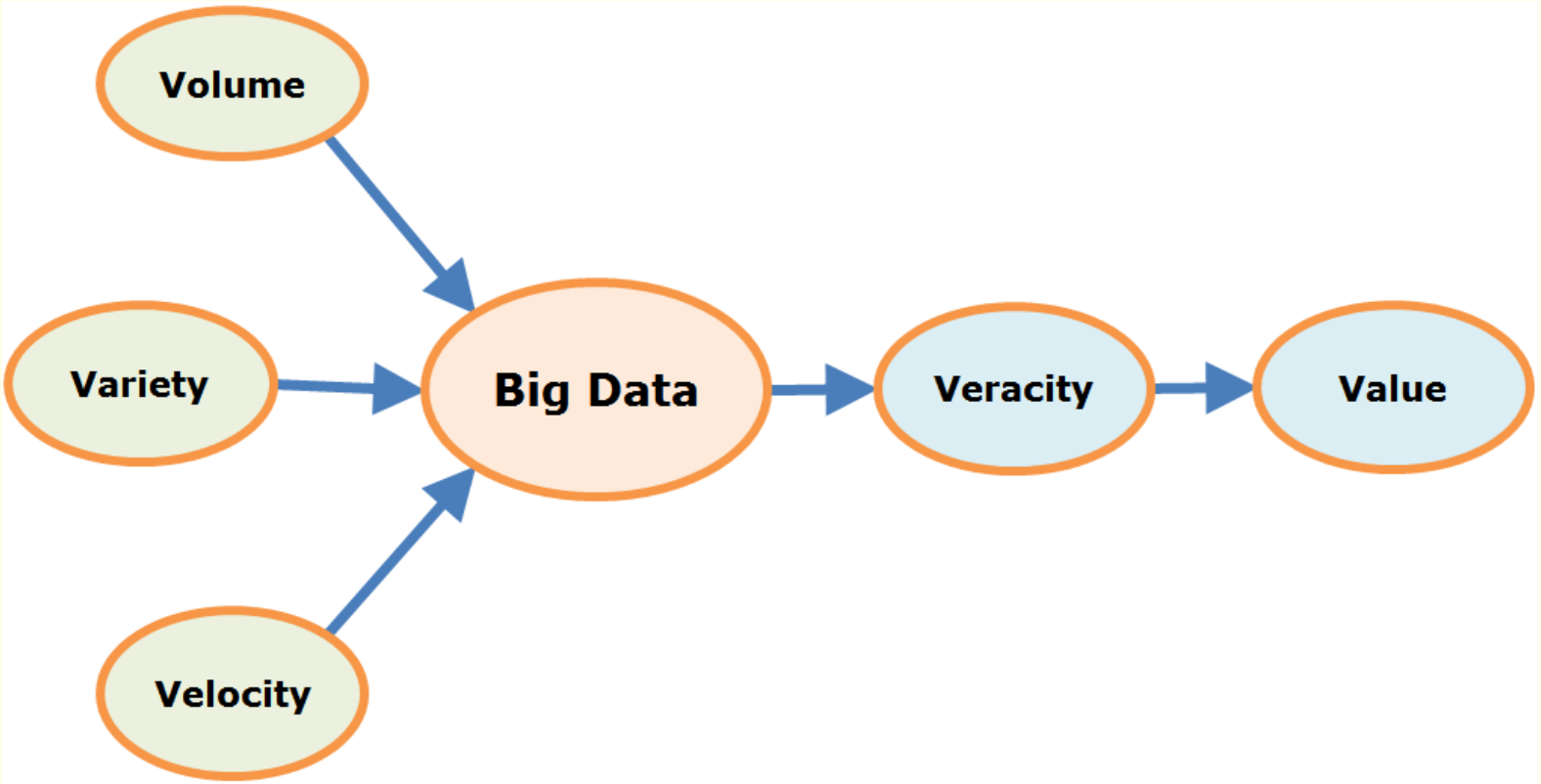
‘Data is part of Switzerland’s infrastructure, such as road, railways and power networks, and is of great value. The government and the economy are obliged to generate added value from these data.’

digitalswitzerland, November 22, 2016

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

$\rightsquigarrow$ 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.

'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**!

2. 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'.

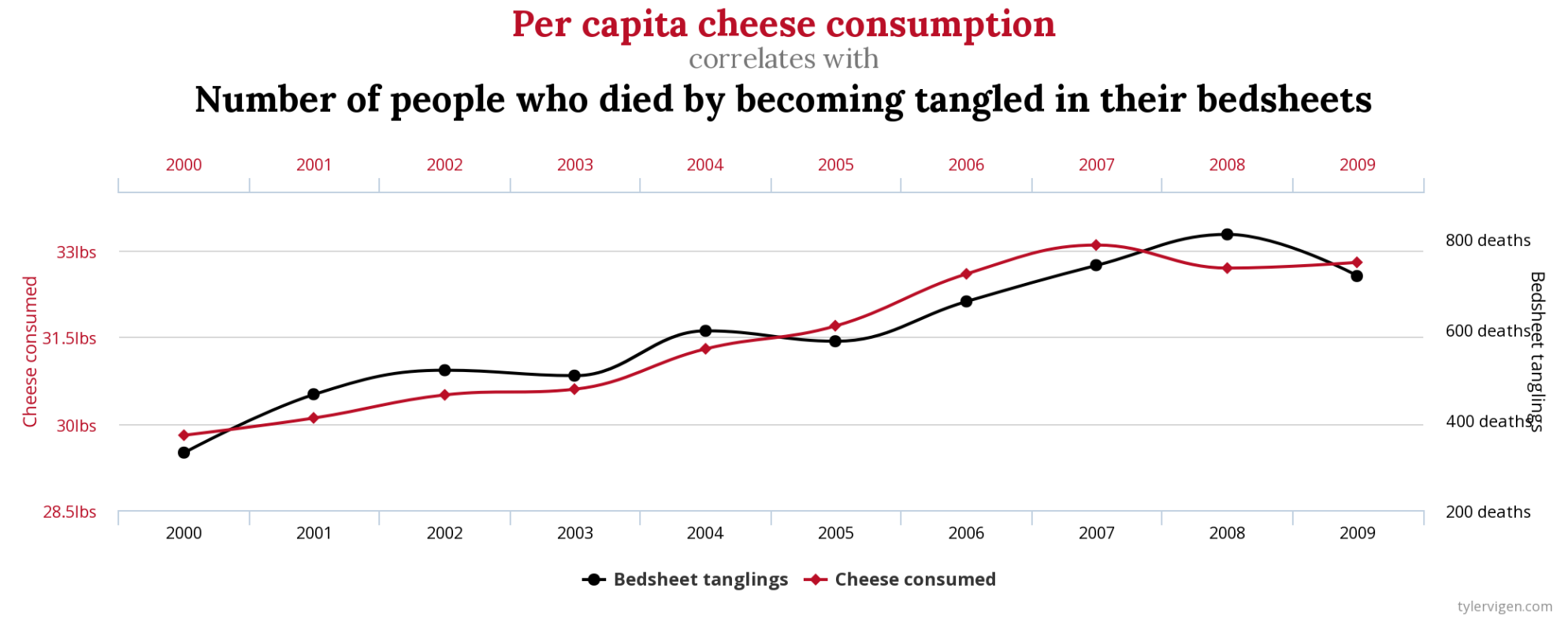
⇝ Analytics' paradigm: '**inductive reasoning**' as 'data first'.

'AI *[('Artificial Intelligence')]* algorithms are not natively 'intelligent'. They learn inductively by analyzing data.'

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).

‘Spurious correlation is not causation!’

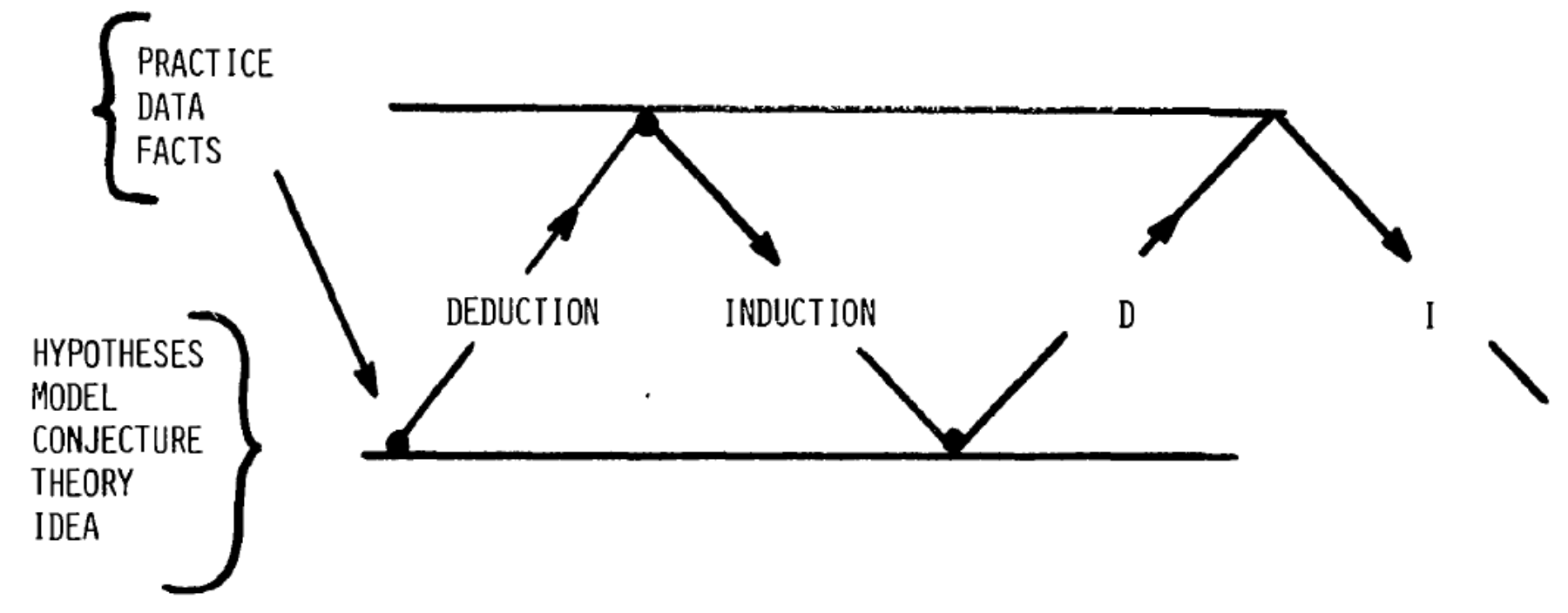


‘Any claim coming from an observational study is most likely to be wrong.’

S. Stanley Young and Alan Karr, 2011

• 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**:

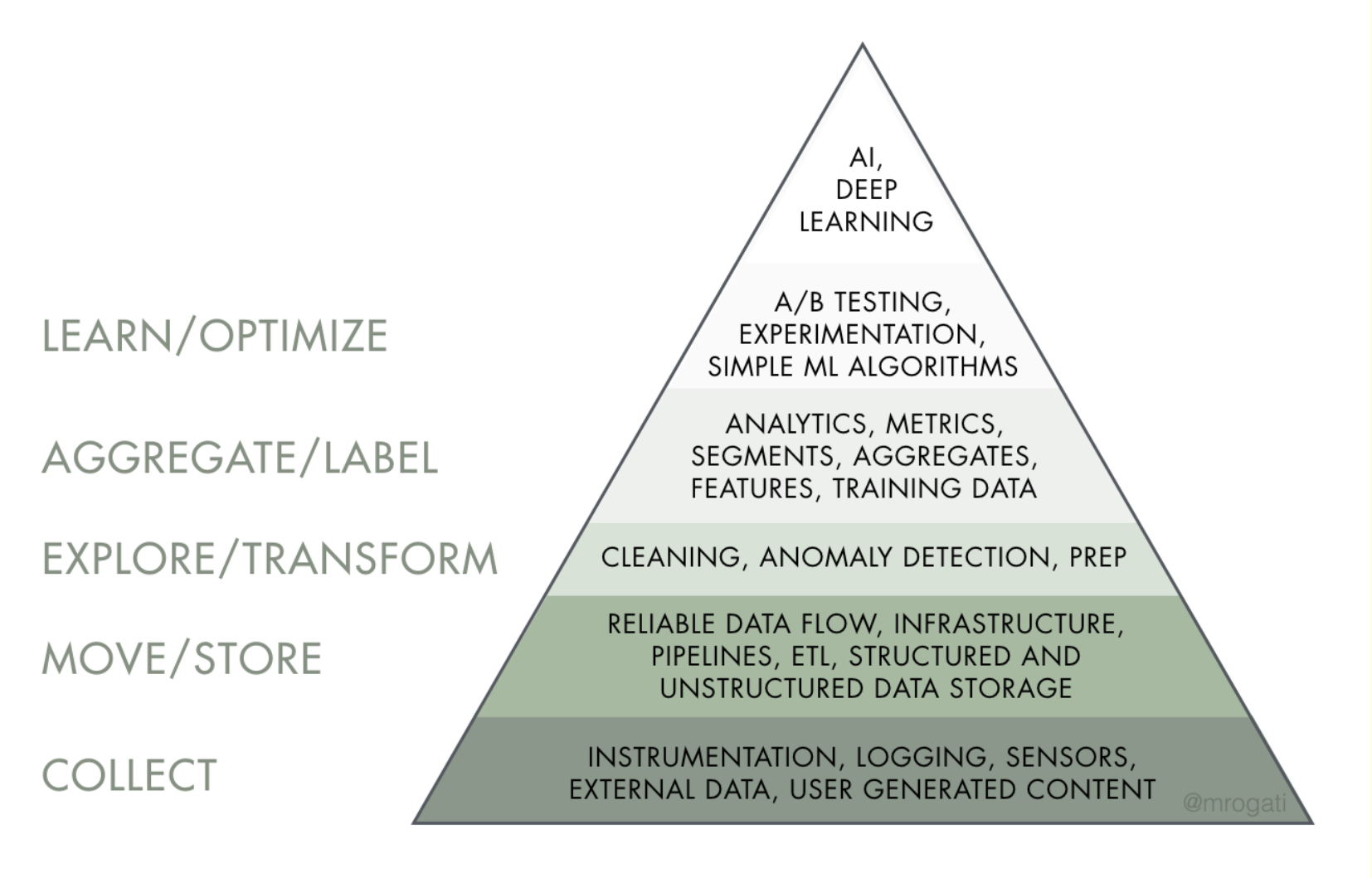


Source: Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71, 791–799.

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



‘Les données elles-mêmes sont une matière première essentielle de la société de la connaissance. Il faut toutefois pour cela que des données d’une qualité et d’une fiabilité élevées soient disponibles et accessibles.’

Conseil fédéral, 5 septembre 2018

Source: stratégie ‘*Suisse numérique*’, adoptée par le Conseil fédéral le 5 septembre 2018 (goo.gl/T8eJUS).

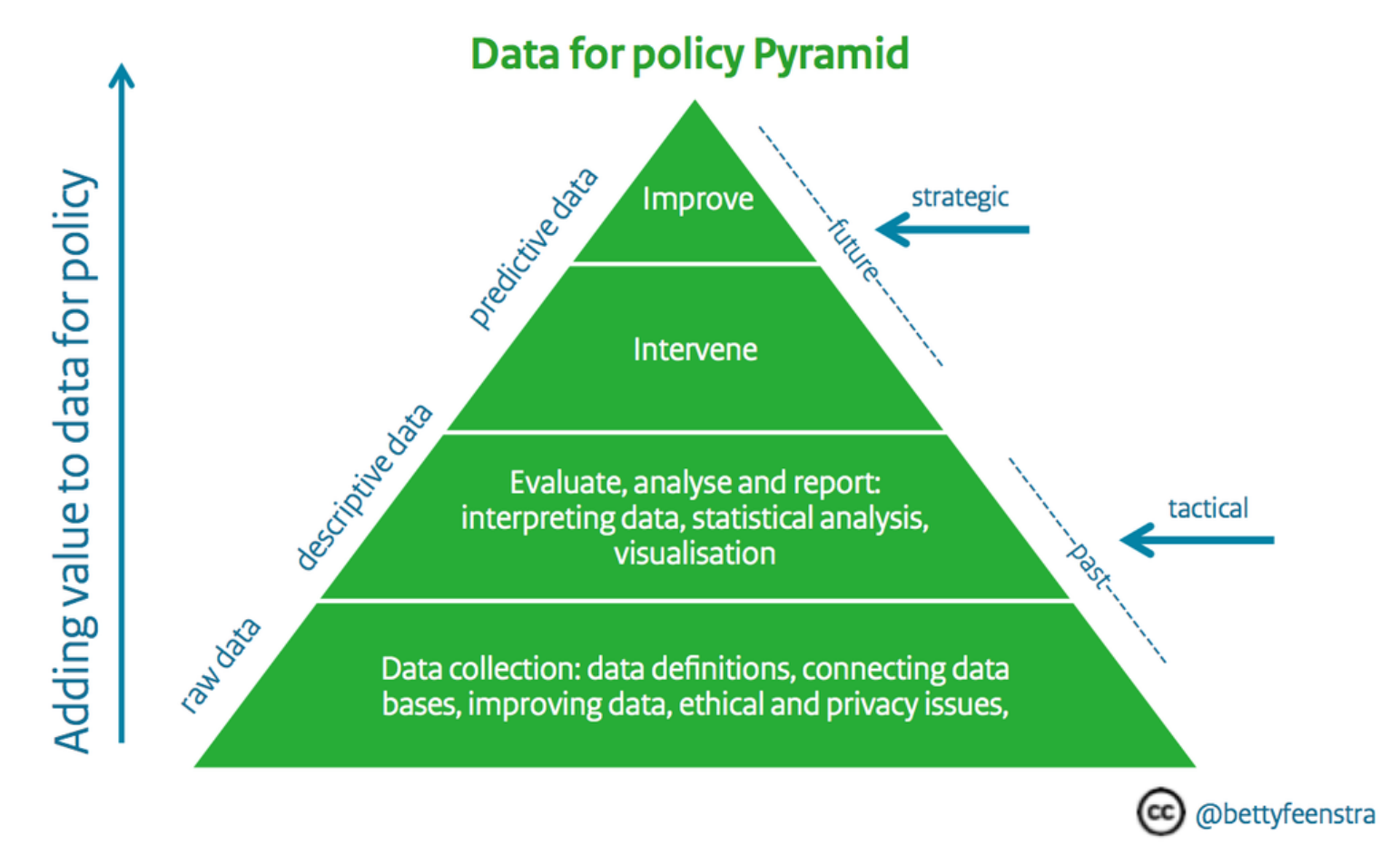
3. Data-informed policy making

'Without data we are flying blind, and we can not do evidence-based policy decisions — or any decision at all.'

Johannes P. Jütting, 2015

Source: Johannes Jütting, Manager of the Partnership in Statistics for Development in the 21st Century (PARIS21), quoted in Sarah Shearman's article 'Data 'crucial' to eradicating poverty' in the *Guardian*, September 28, 2015 (goo.gl/DBTwza).

• Policy makers want the top of the iceberg, but they need to remember the stuff beneath sea (adapted from `@HetanShah`):



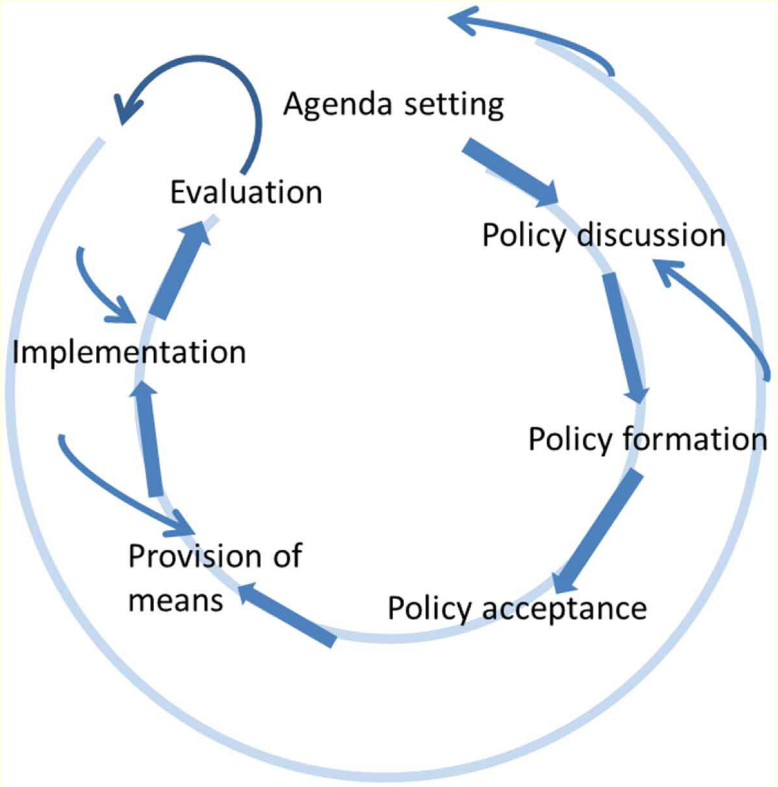
⇝ In a world of post-truth politics, the veracity of data is more important than ever!

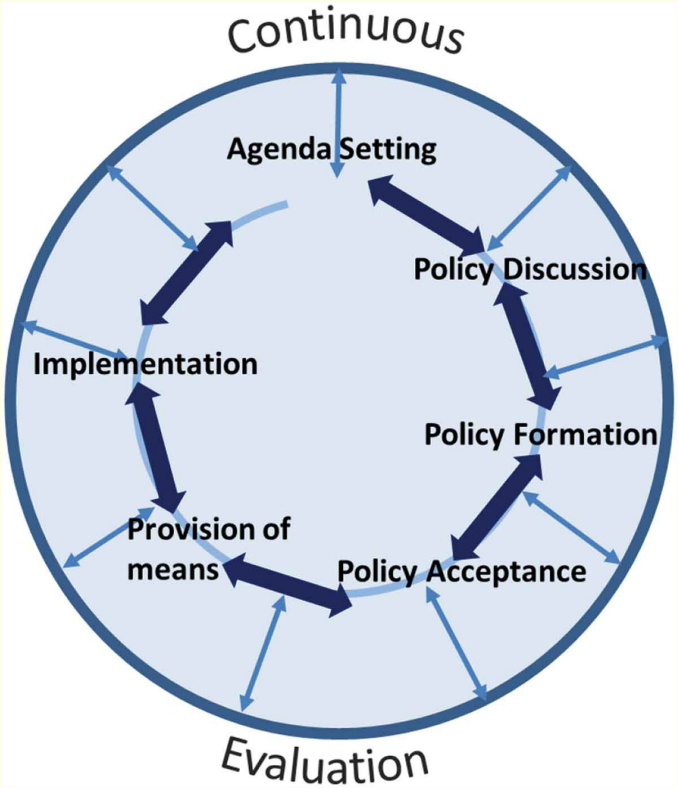
‘Data are the lifeblood of decision-making and the raw material for accountability. Without high-quality data providing the right information on the right things at the right time; designing, monitoring and evaluating effective policies becomes almost impossible.’

IEAG, 2014

Source: United Nations Secretary-General’s ‘Independent Expert Advisory Group on a Data Revolution for Sustainable Development’ (IEAG), *A Word That Counts: Mobilising The Data Revolution for Sustainable Development*, November 6, 2014 (`www.undatarevolution.org/report/`).

The policy cycle and the (big) data-revised policy cycle



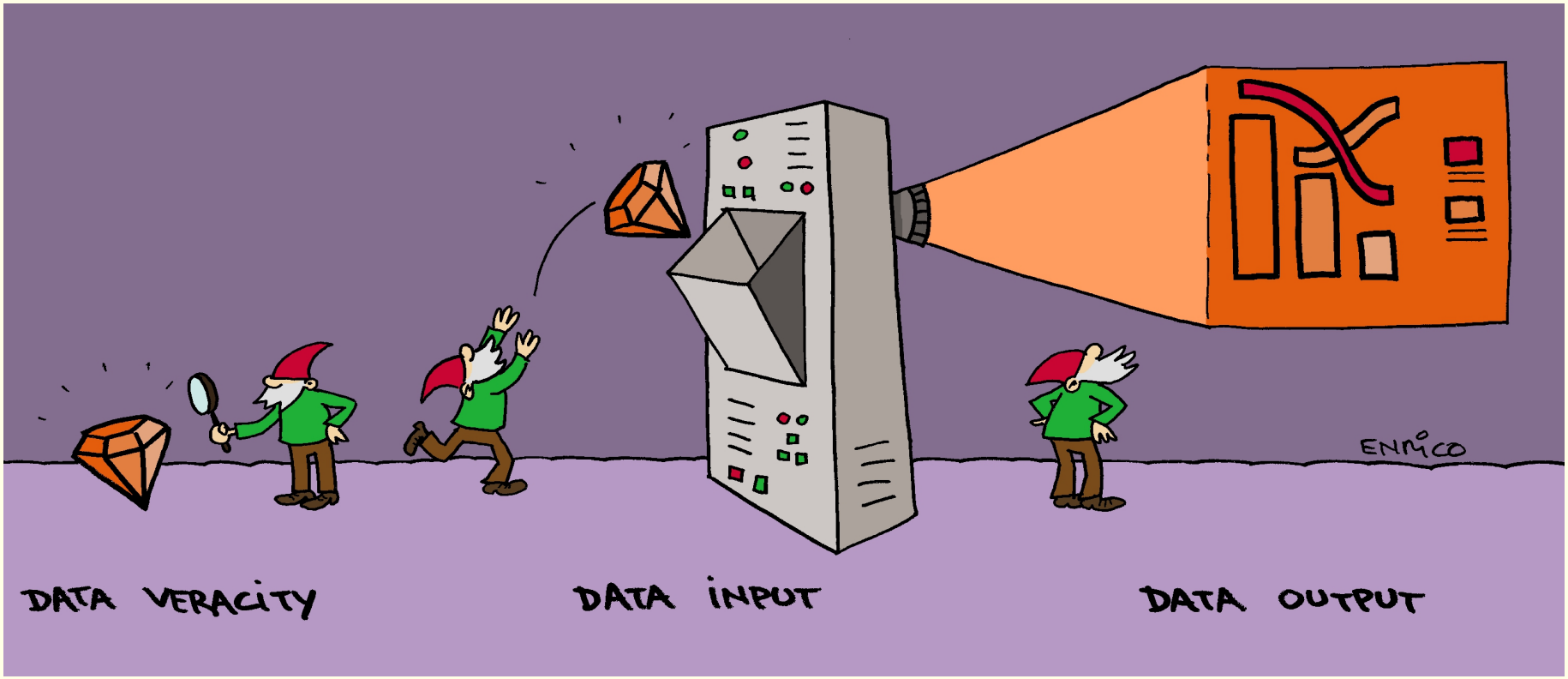


$\rightsquigarrow$ The revised policy cycle (right) takes into account (big) data analytics using data-informed continuous evaluation at any stage ($\rightsquigarrow$ 'e-policy cycle').

Source: Höchtl, J., Parycek, P. & Schöllhammer, R. (2016). Big data in the policy cycle: policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce*, 26, 147–169.

4. Conclusion and opportunities

• In a world of (big) 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!

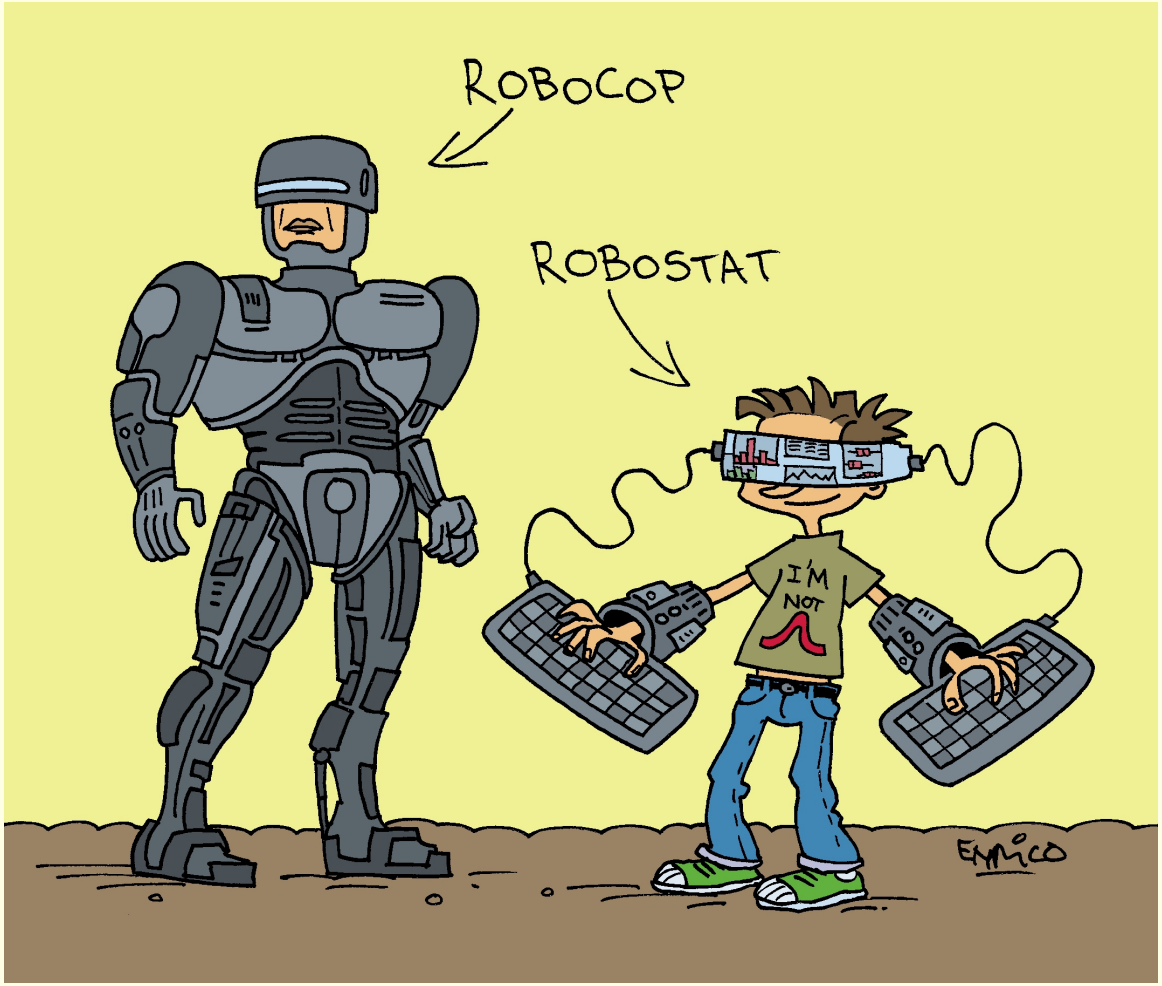
$\rightsquigarrow$ Data and analytics should be envisaged to complement and augment humans, not replacements for them!

‘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).

• 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.



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

The Beatles, 1967

Have you been Statooed & GSEMed?

Prof. Dr. ès sc. Diego Kuonen, CStat PStat

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: 'SEVAL.Sep.2022'. Typesetting: LaTeX, version 2ϵ. Compilation date: 31.08.2022.