Production Processes of Official Statistics & Analytics Processes Augmented by Analytics of Things: Friends or Foes?

Prof. Dr. Diego Kuonen, CStat PStat CSci

Statoo Consulting, Berne, Switzerland

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

BigData UN Global Working Group

'Seminar on Data Science Campus activities in Kigali, Rwanda' — April 29, 2019

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).

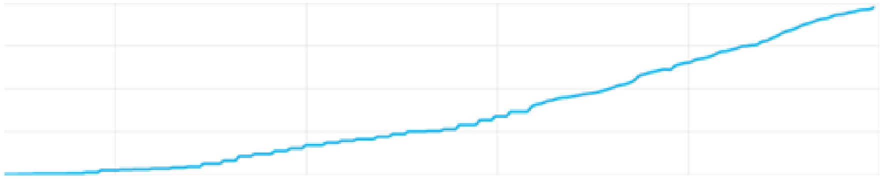
▷ Founder, CEO & CAO, Statoo Consulting, Switzerland (since 2001).

▷ 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
- 12.04.2019: 20’366

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING
TOP 100 INFLUENCERS AND BRANDS

About Statoo Consulting (www.statoo.info)

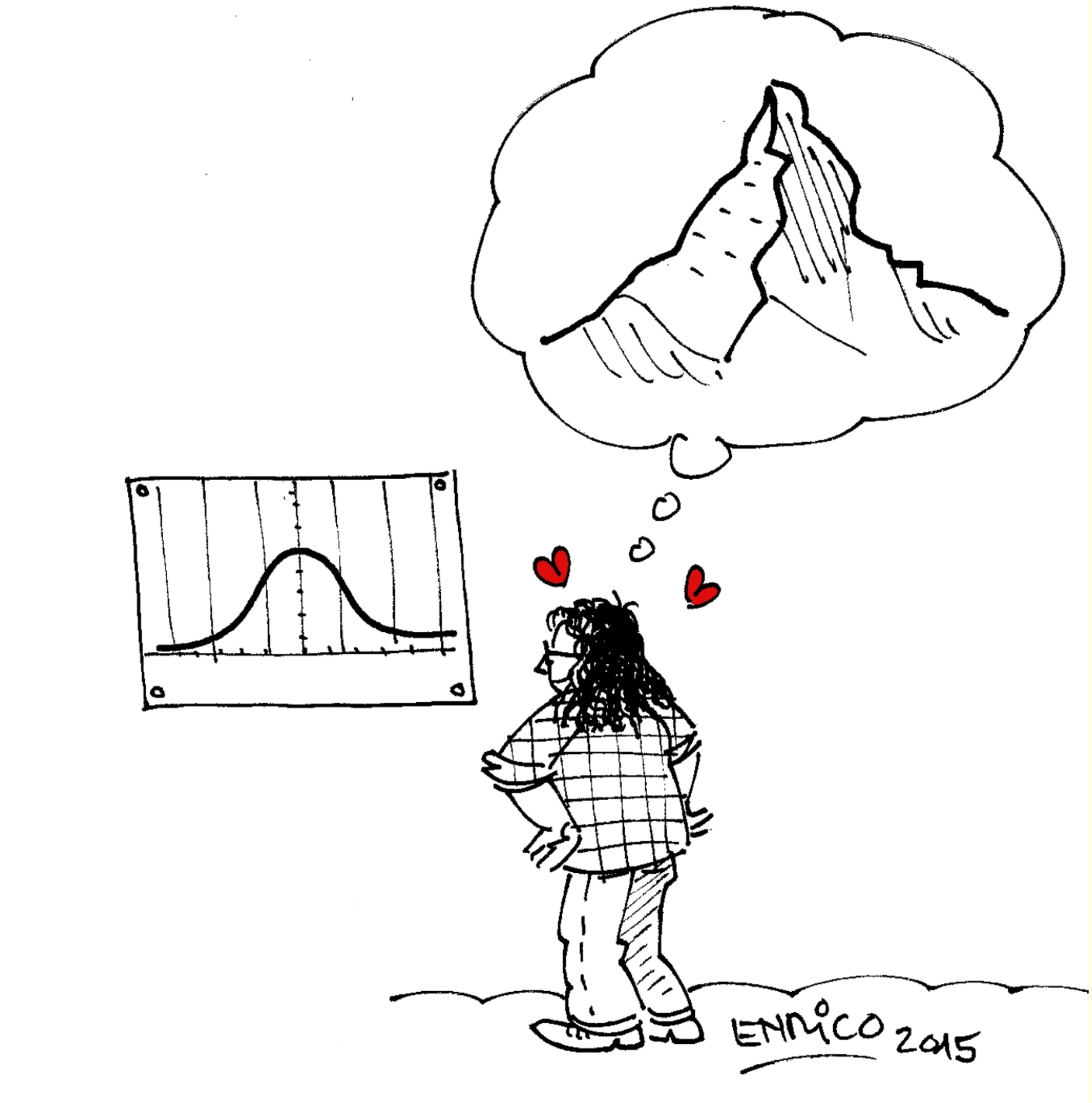
- Founded Statoo Consulting in **2001**.

$$\boxed{\rightsquigarrow 2019 - 2001 = 18 + \epsilon.}$$

- 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.

⇝ **Are you drowning in uncertainty and starving for knowledge?**

⇝ **Have you ever been Statooed?**


ENRICO 2015

Contents

‘Just as haute cuisine must incessantly reinvent itself in order to stay at the forefront of gastronomy, official statistics is also confronted with a rapidly changing context and needs. They are currently facing an impressive number of challenges: the ‘data revolution’ and the emergence of ‘big data’, ...’

Walter J. Radermacher, 2018

Source: Radermacher, W. J. (2018). Official statistics in the era of big data opportunities and threats. *International Journal of Data Science and Analytics* (`doi.org/10.1007/s41060-018-0124-z33`).

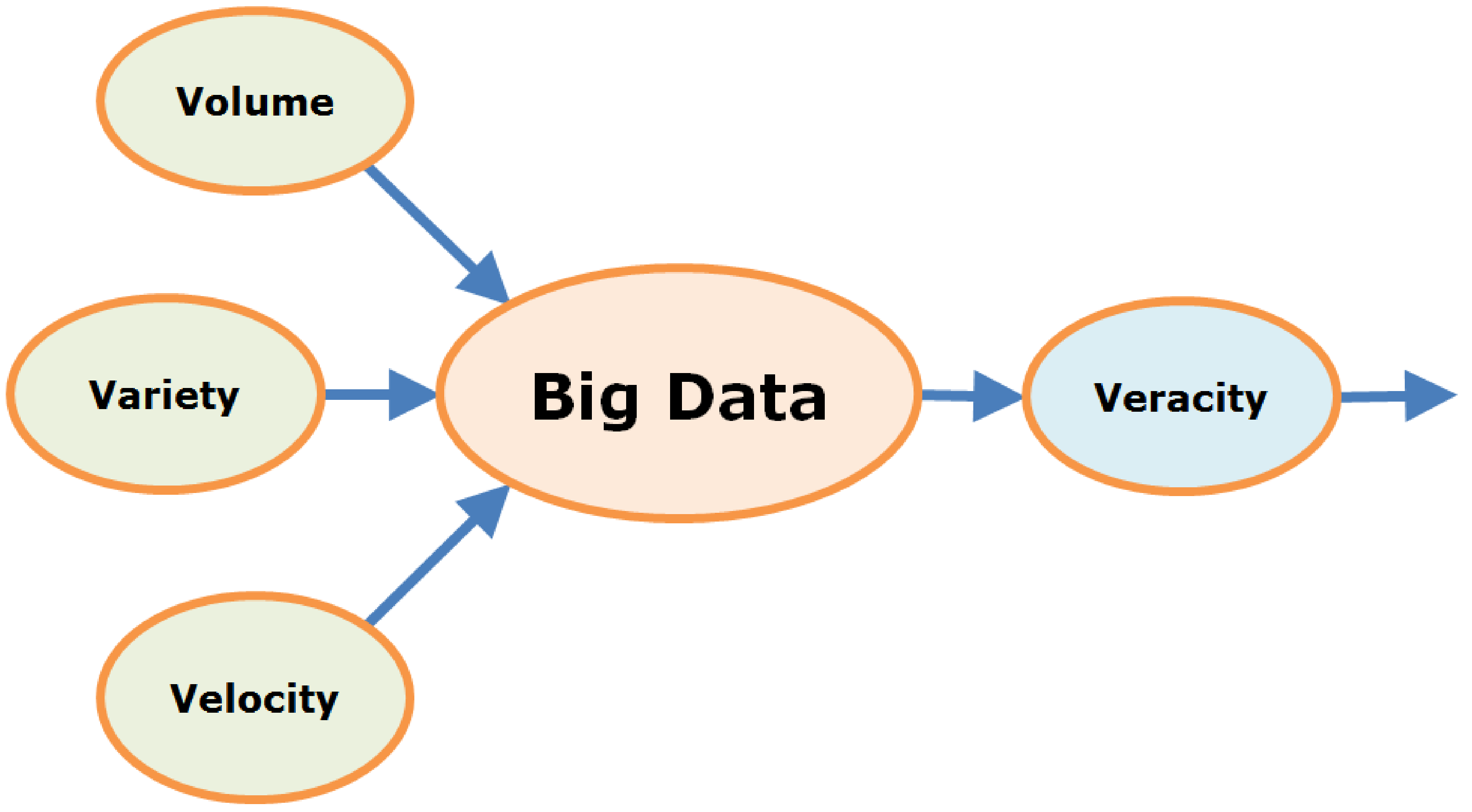
1. Demystifying the ‘big data’ hype

- 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’).



⇝ The $\boxed{\text{‘Veracity’}}$ (*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!

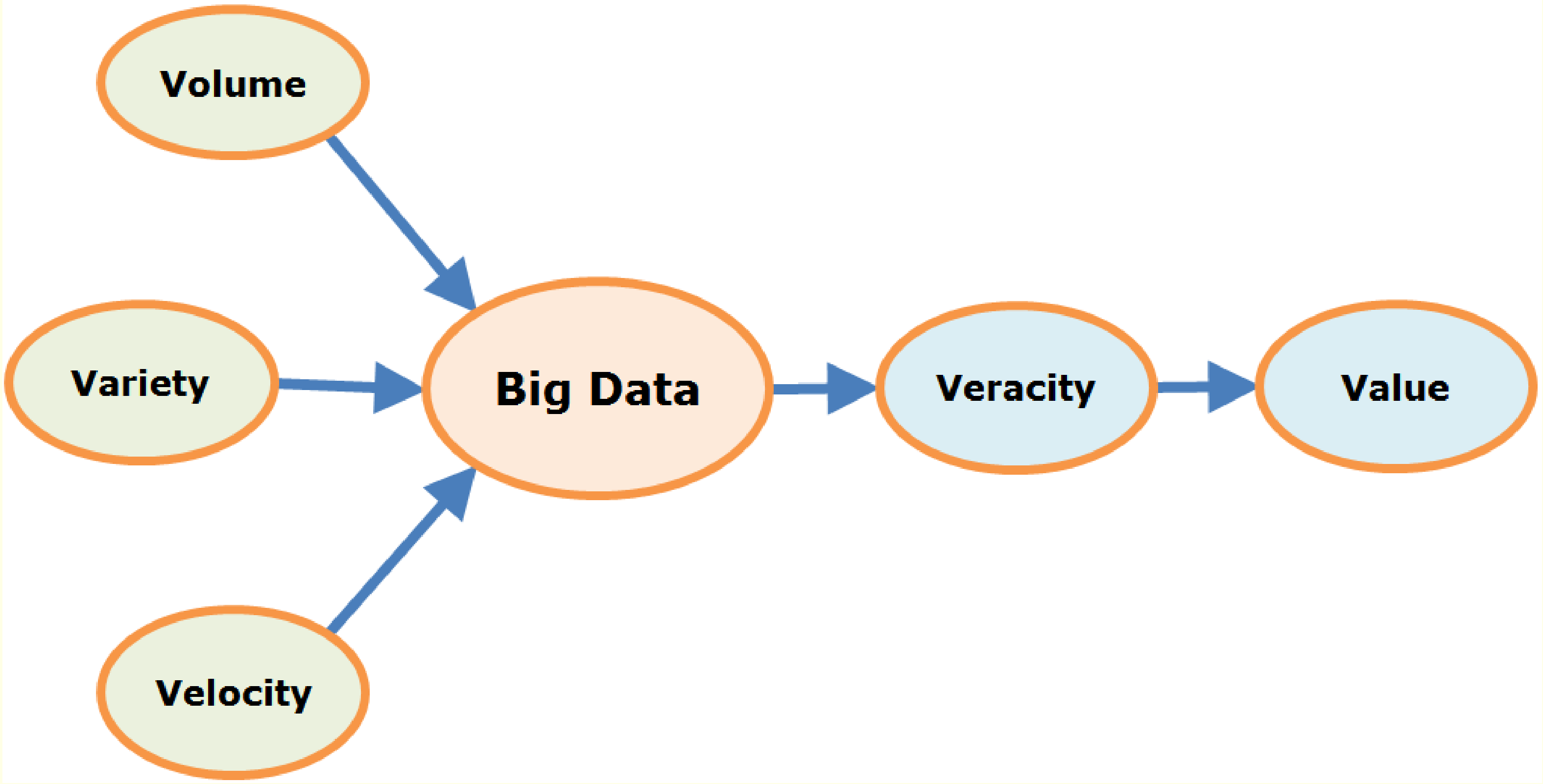
‘Data is part of Switzerland’s infrastructure, such as road, railways and power networks, and is of great value. The government and the economy are obliged to generate added value from these data.’

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



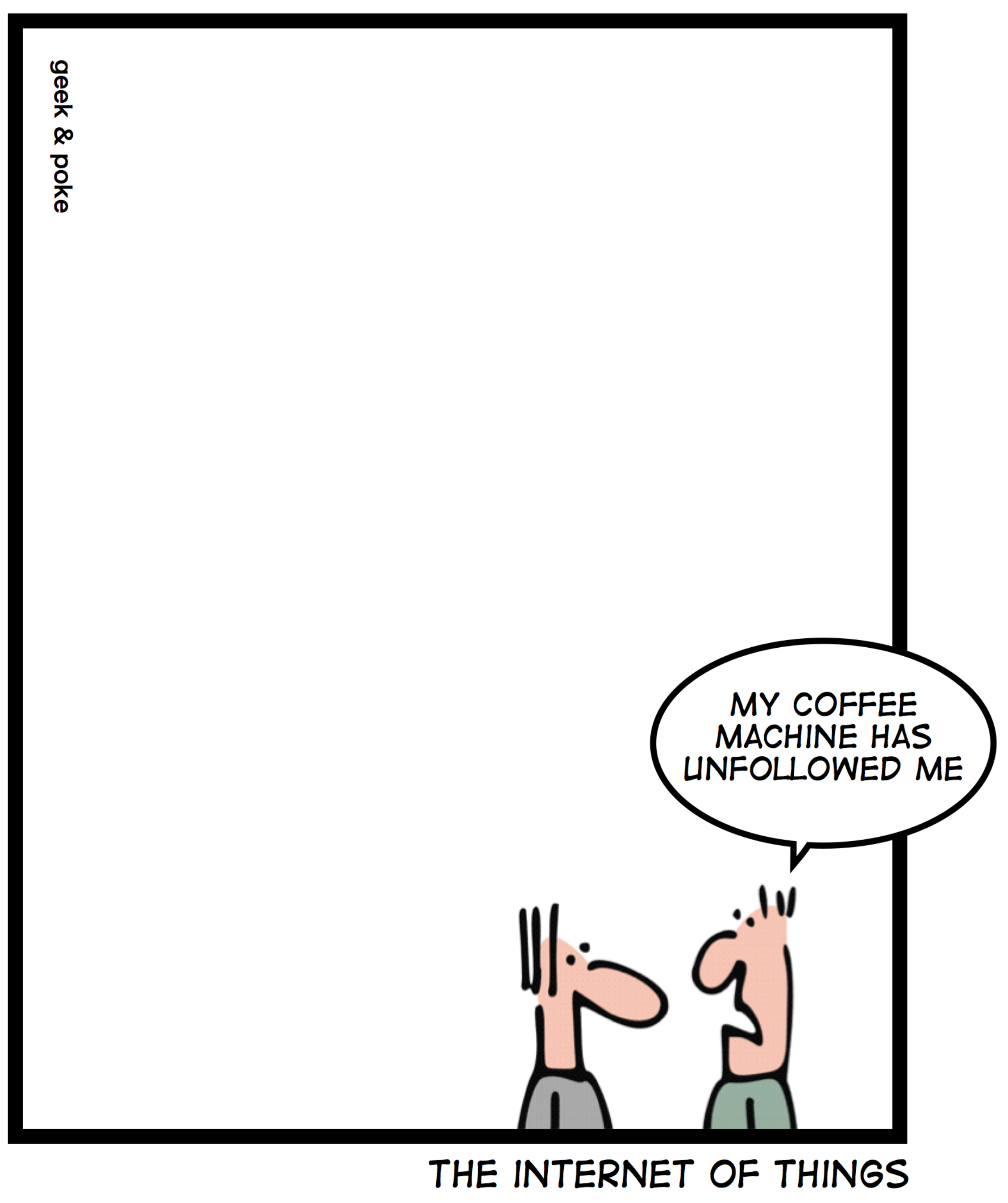
◇ 'Volume', 'Variety' and 'Velocity' are the 'essential' characteristics of (big) data;

◇ 'Veracity' and 'Value' are the 'qualification for use' characteristics of (big) data.

‘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.’

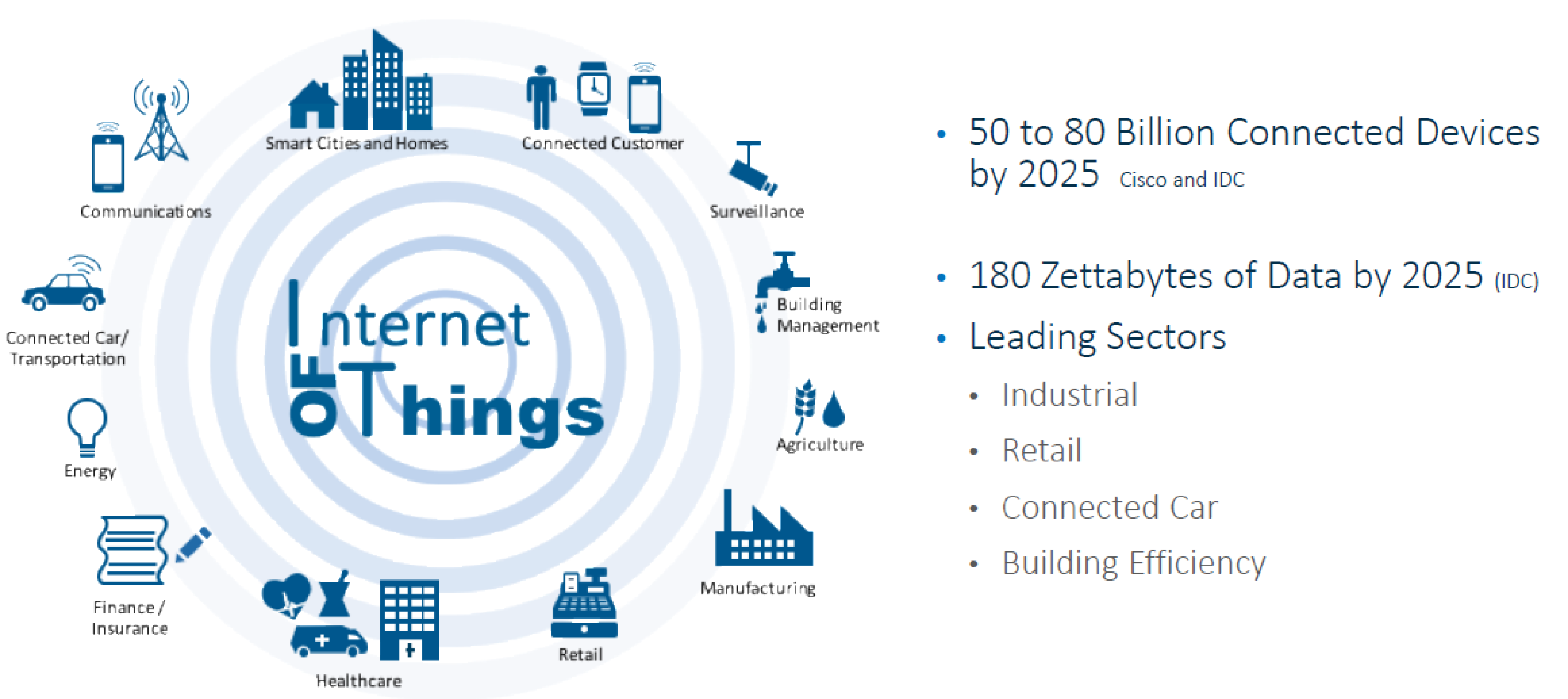
Swiss Federal Council, September 5, 2018

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


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

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’!

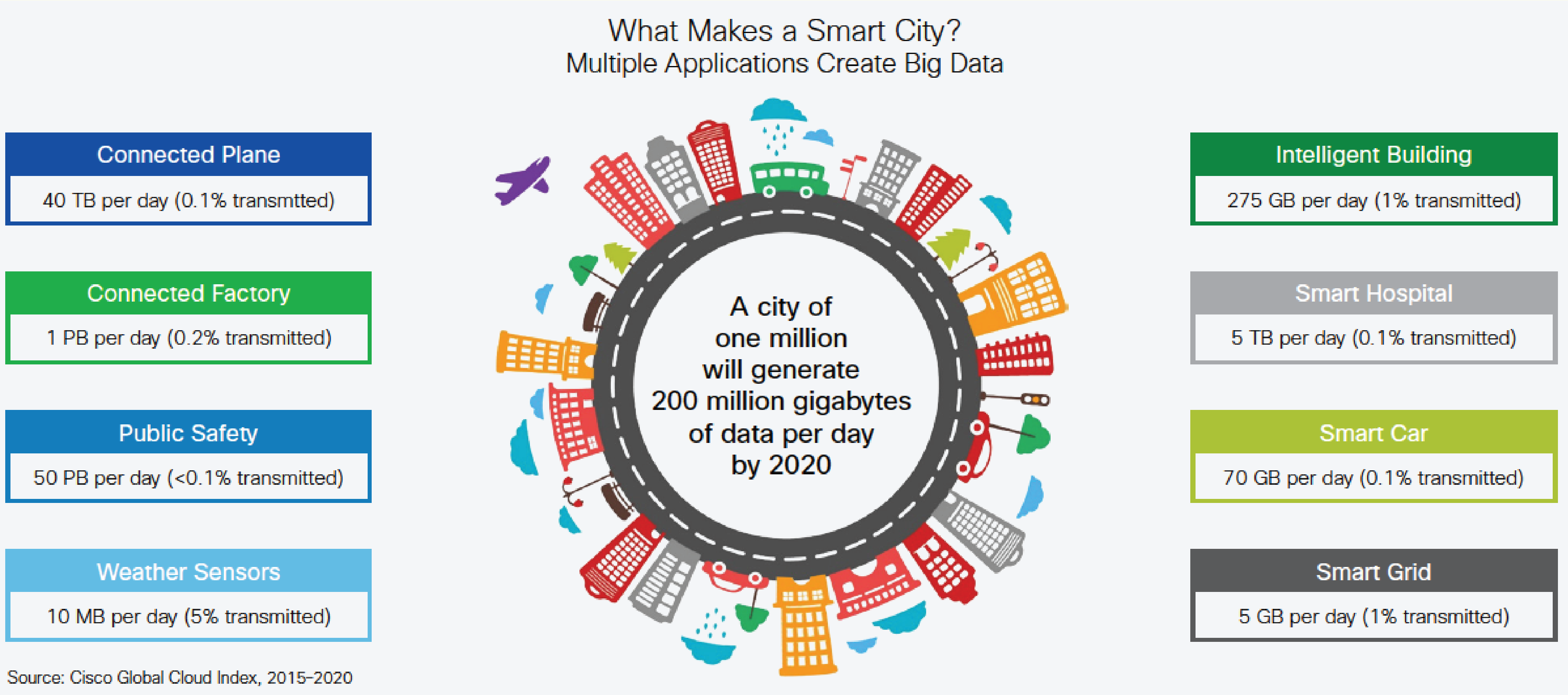


- 50 to 80 Billion Connected Devices by 2025 Cisco and IDC
- 180 Zettabytes of Data by 2025 (IDC)
- Leading Sectors
  - Industrial
  - Retail
  - Connected Car
  - Building Efficiency

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

⇝ However, IoT is about data, not things!

s+a+oo



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)

A city of
one million
will generate
200 million gigabytes
of data per day
by 2020

Intelligent Building
275 GB per day (1% transmitted)

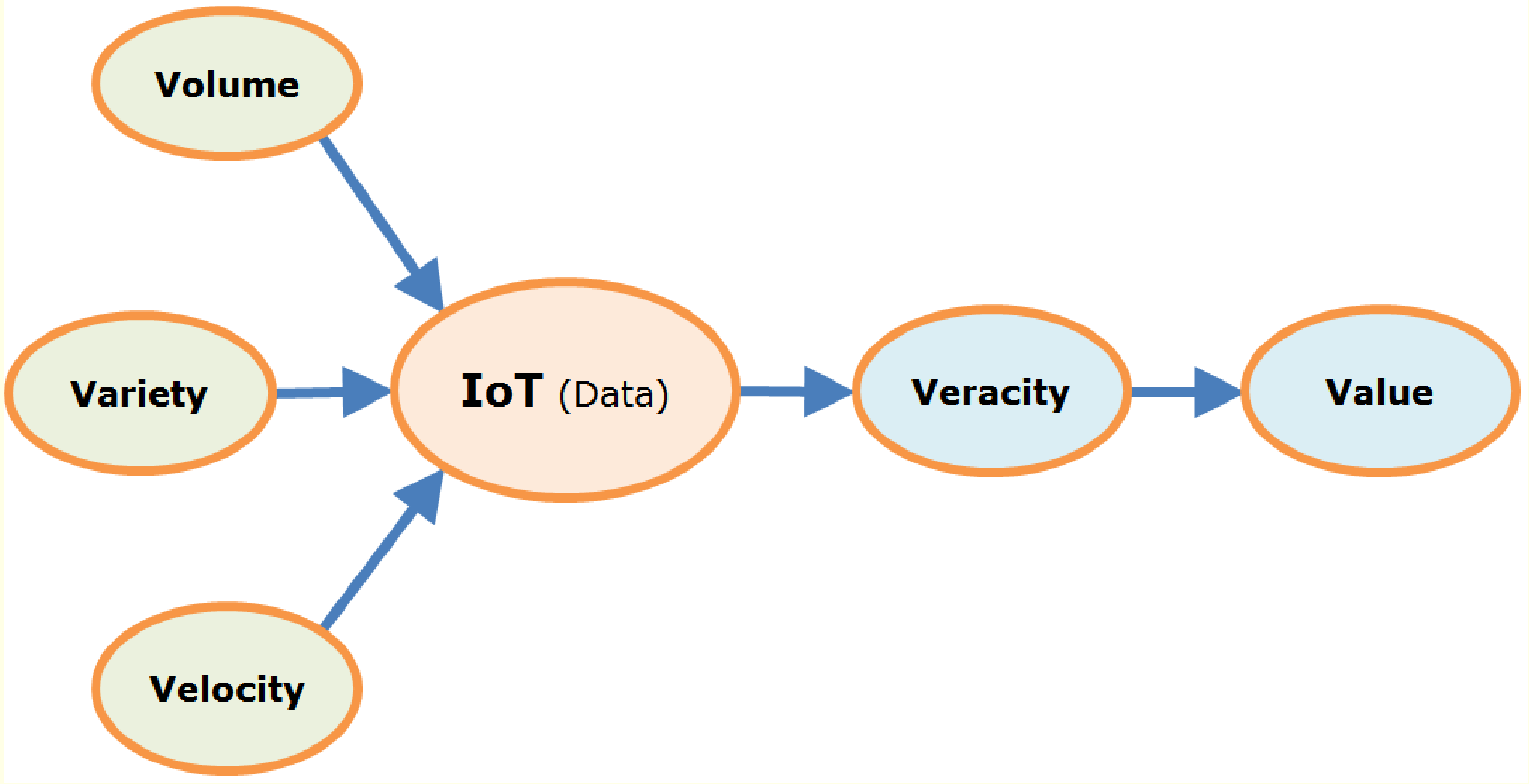
Smart Hospital
5 TB per day (0.1% transmitted)

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).

> ‘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’**!

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) 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’.

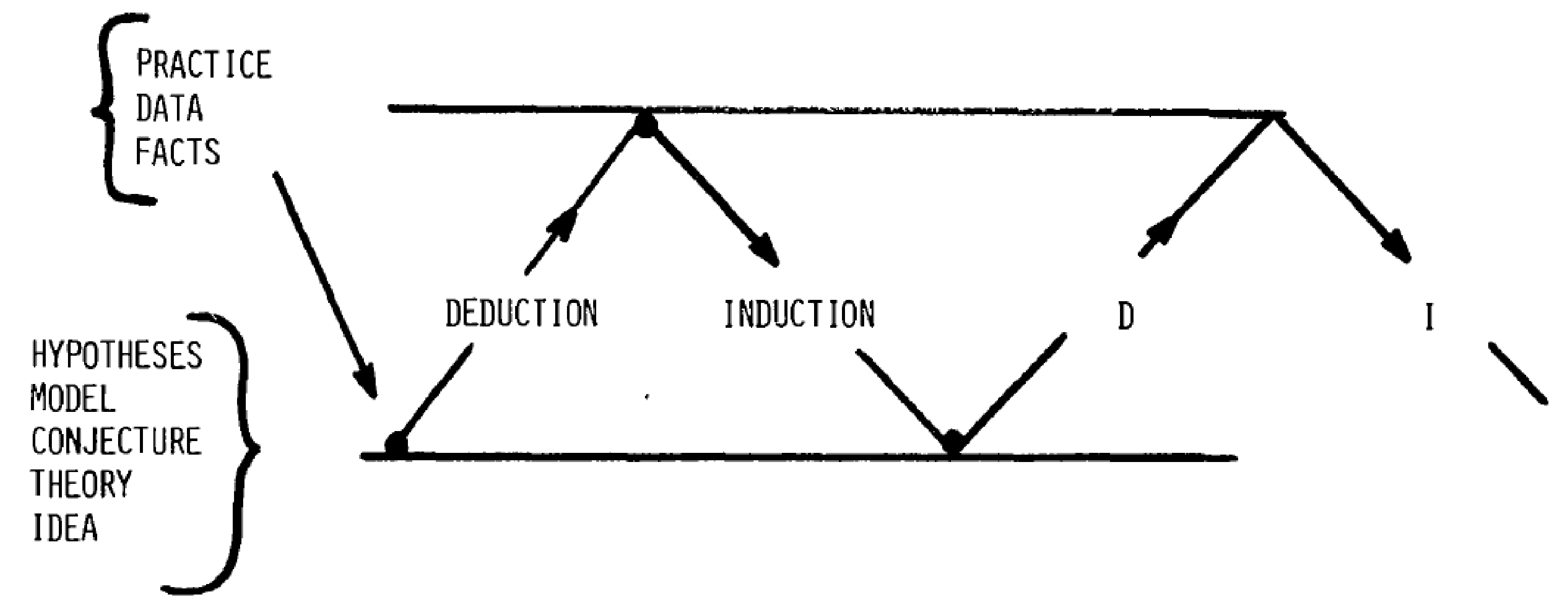
‘AI *[(‘Artificial Intelligence’)]* algorithms are not natively ‘intelligent’. They learn inductively by analyzing data.’

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`).

- 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, data-informed policy making and proper continuous improvement.

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

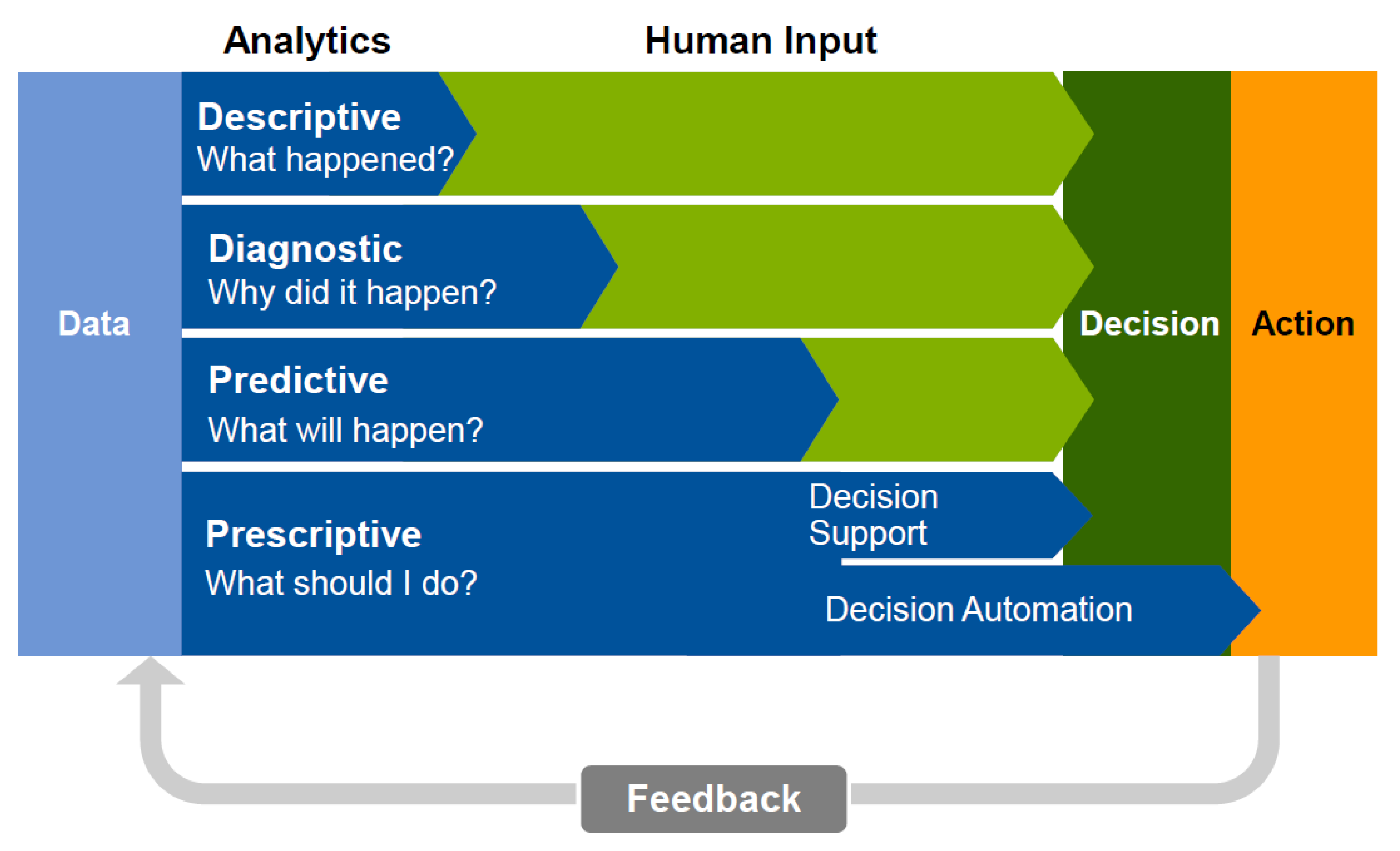


Source: Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71, 791–799.

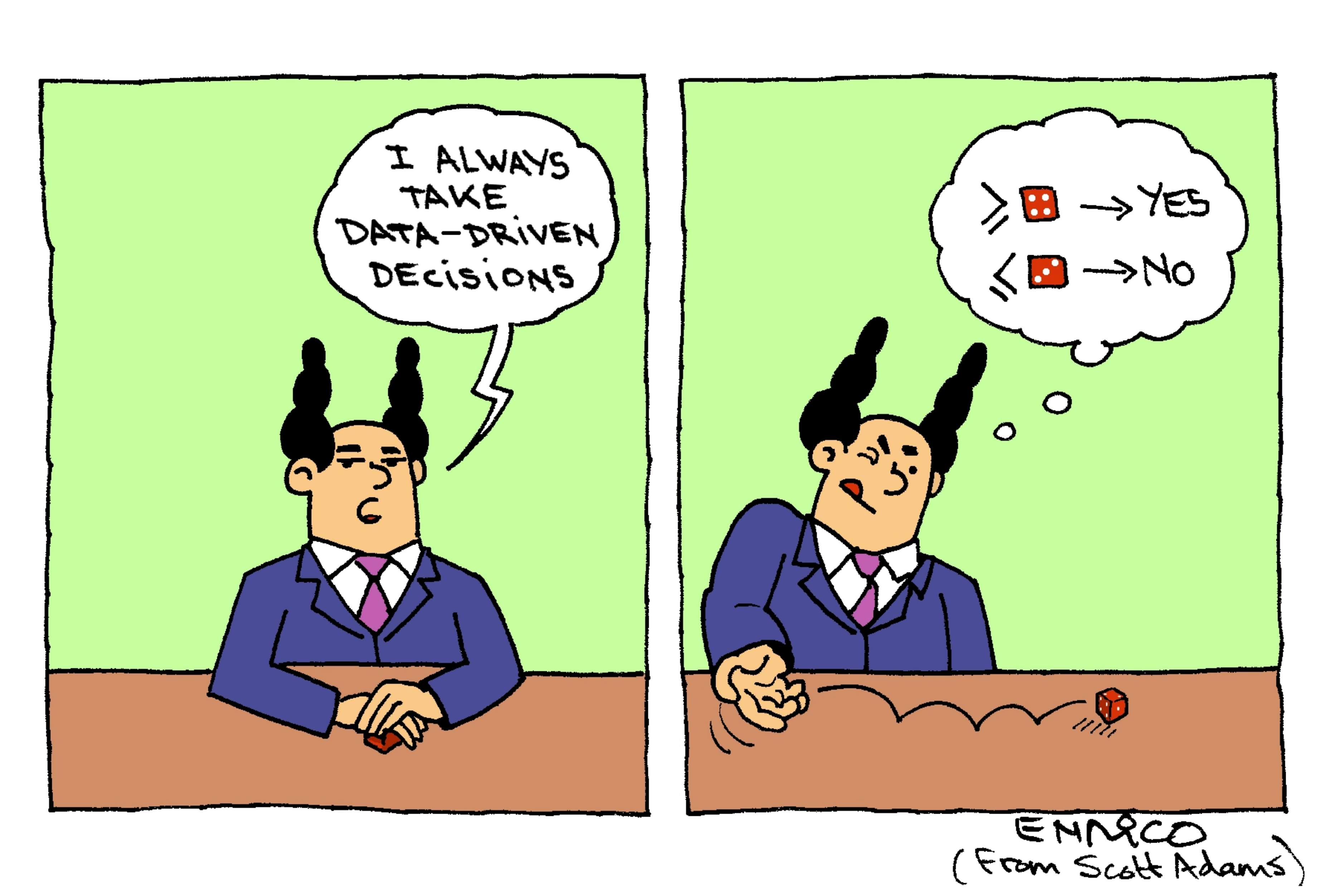
‘Neither exploratory nor confirmatory is sufficient alone. To try to replace either by the other is madness. We need them both.’

John W. Tukey, 1980

Questions analytics tries to answer & the 'analytics continuum'



Source: João Tapadinhas, VP Business Analytics and Data Science, Gartner, June 2014 (`goo.gl/YmjFPB`).


I ALWAYS TAKE DATA-DRIVEN DECISIONS
≥ ⚃ → YES
≤ ⚂ → NO
ENRICO
(From Scott Adams)

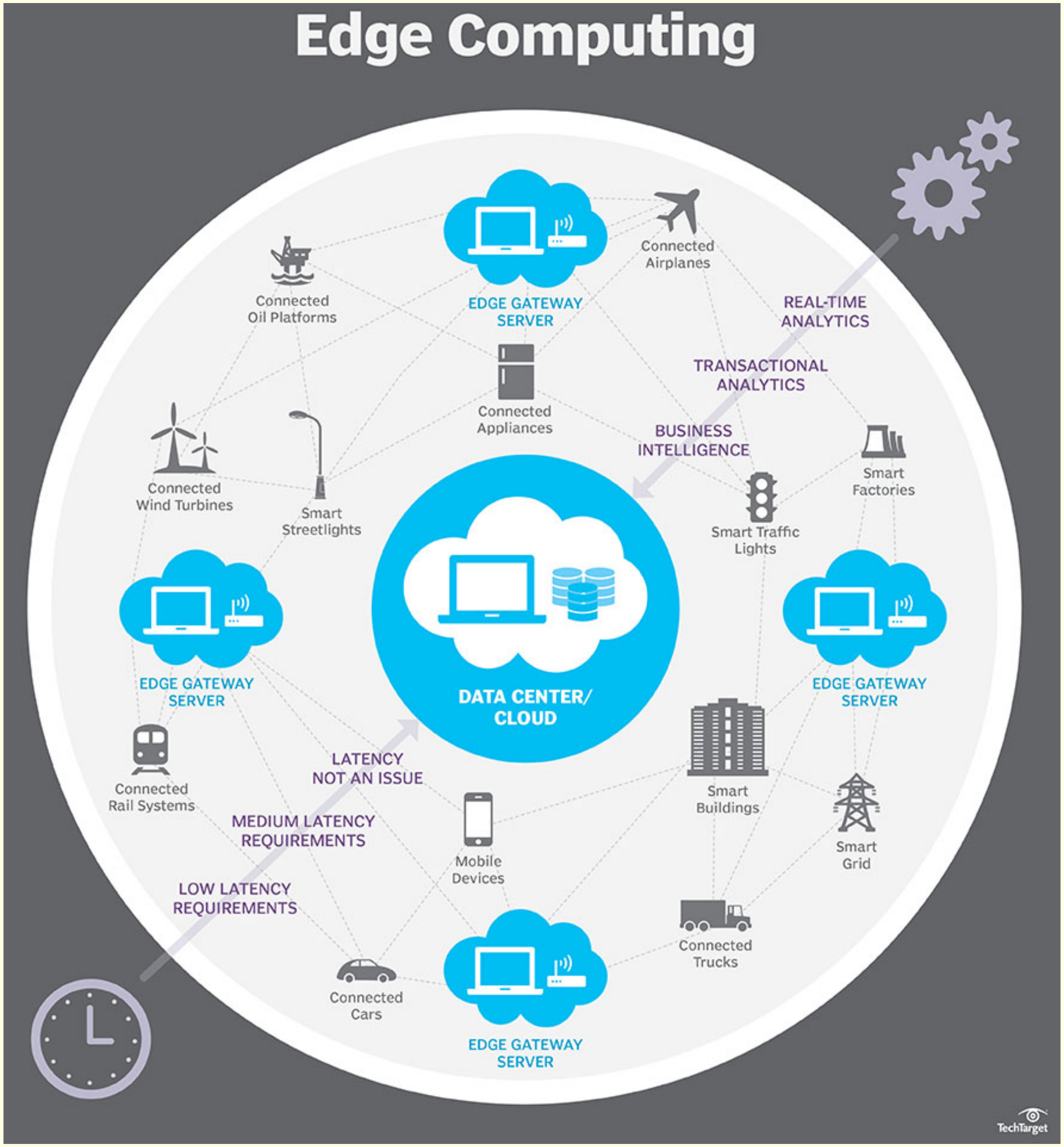
4. Demystifying ‘analytics of things’

• 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.


Edge Computing
EDGE GATEWAY SERVER
Connected Airplanes
Connected Oil Platforms
REAL-TIME ANALYTICS
TRANSACTIONAL ANALYTICS
Connected Appliances
BUSINESS INTELLIGENCE
Smart Factories
Connected Wind Turbines
Smart Streetlights
Smart Traffic Lights
EDGE GATEWAY SERVER
DATA CENTER/ CLOUD
EDGE GATEWAY SERVER
Connected Rail Systems
LATENCY NOT AN ISSUE
Smart Buildings
MEDIUM LATENCY REQUIREMENTS
Mobile Devices
Smart Grid
LOW LATENCY REQUIREMENTS
Connected Trucks
Connected Cars
EDGE GATEWAY SERVER
TechTarget

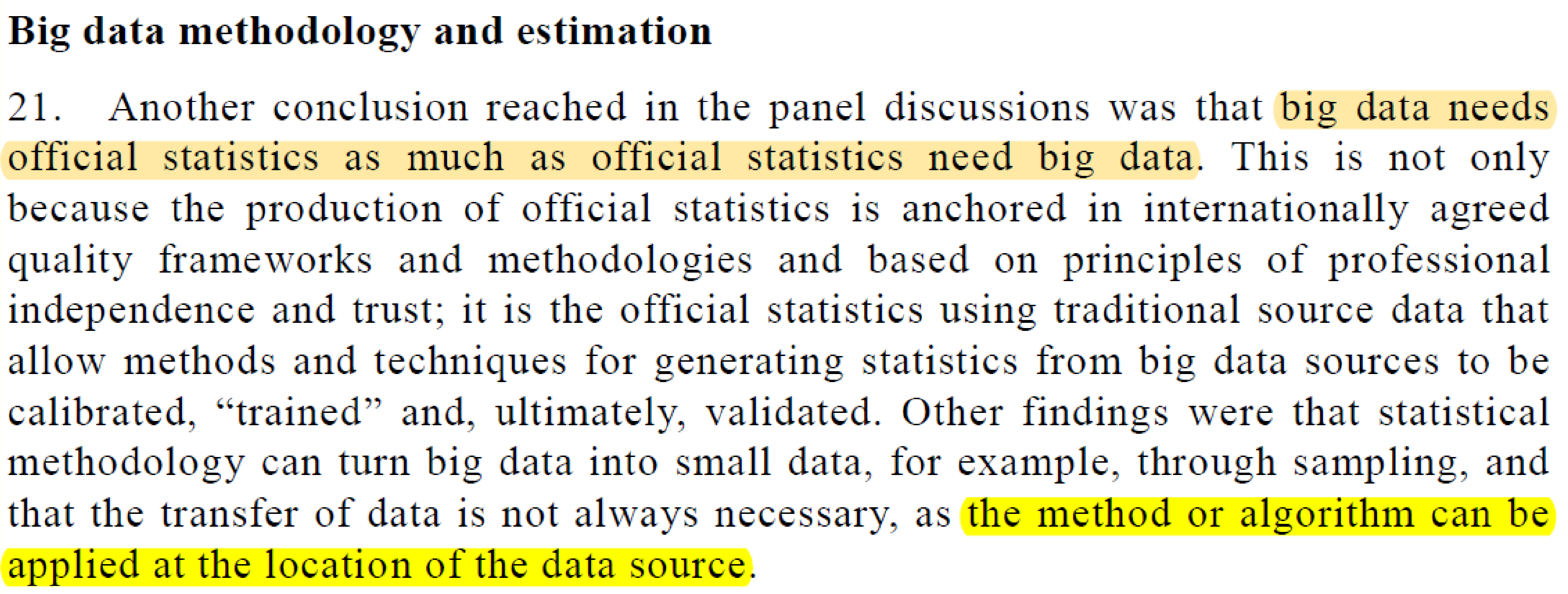
⇝ It is about moving the analytics and the ‘data quality frameworks’ to the data and not the data to the (centralised) analytics and (centralised) ‘data quality frameworks’.

⇝ 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.

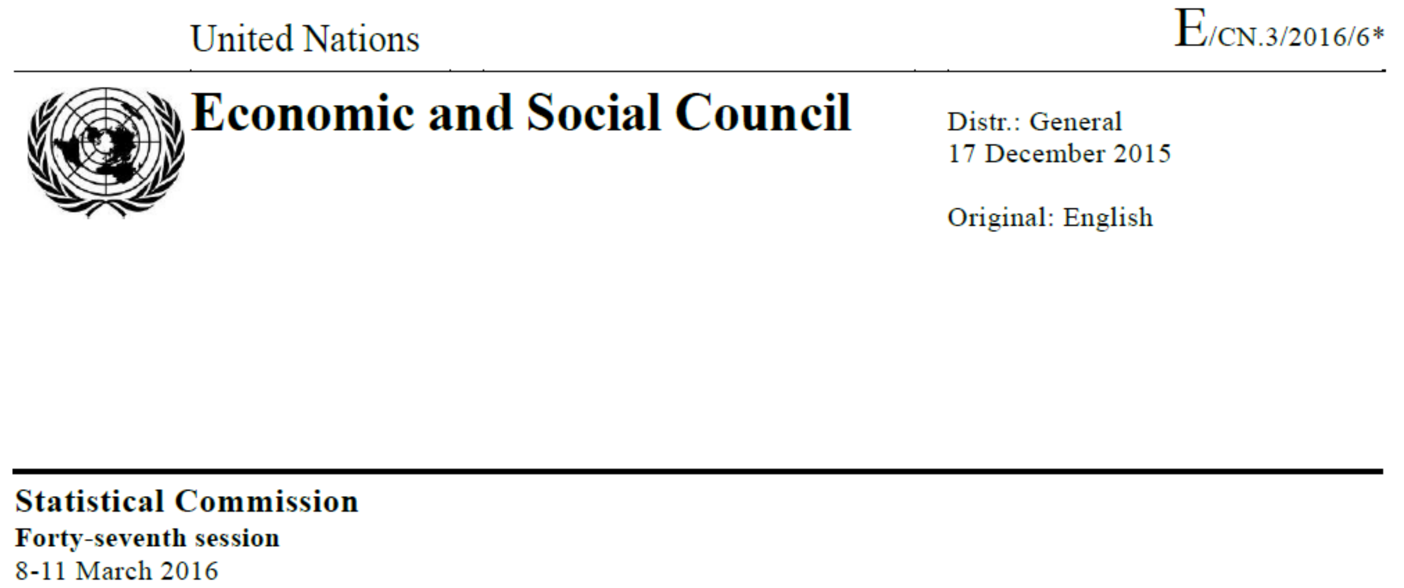
- Current key challenges: lack of (glocalised) standards of both IoT data and analytics, and of AoT approaches, *e.g.* ‘edge analytics’.

⇝ Standardisation efforts needed (by official statistics by augmenting existing ones?)!



Big data methodology and estimation

21. Another conclusion reached in the panel discussions was that big data needs official statistics as much as official statistics need big data. This is not only because the production of official statistics is anchored in internationally agreed quality frameworks and methodologies and based on principles of professional independence and trust; it is the official statistics using traditional source data that allow methods and techniques for generating statistics from big data sources to be calibrated, "trained" and, ultimately, validated. Other findings were that statistical methodology can turn big data into small data, for example, through sampling, and that the transfer of data is not always necessary, as the method or algorithm can be applied at the location of the data source.



United Nations E/CN.3/2016/6*

Economic and Social Council

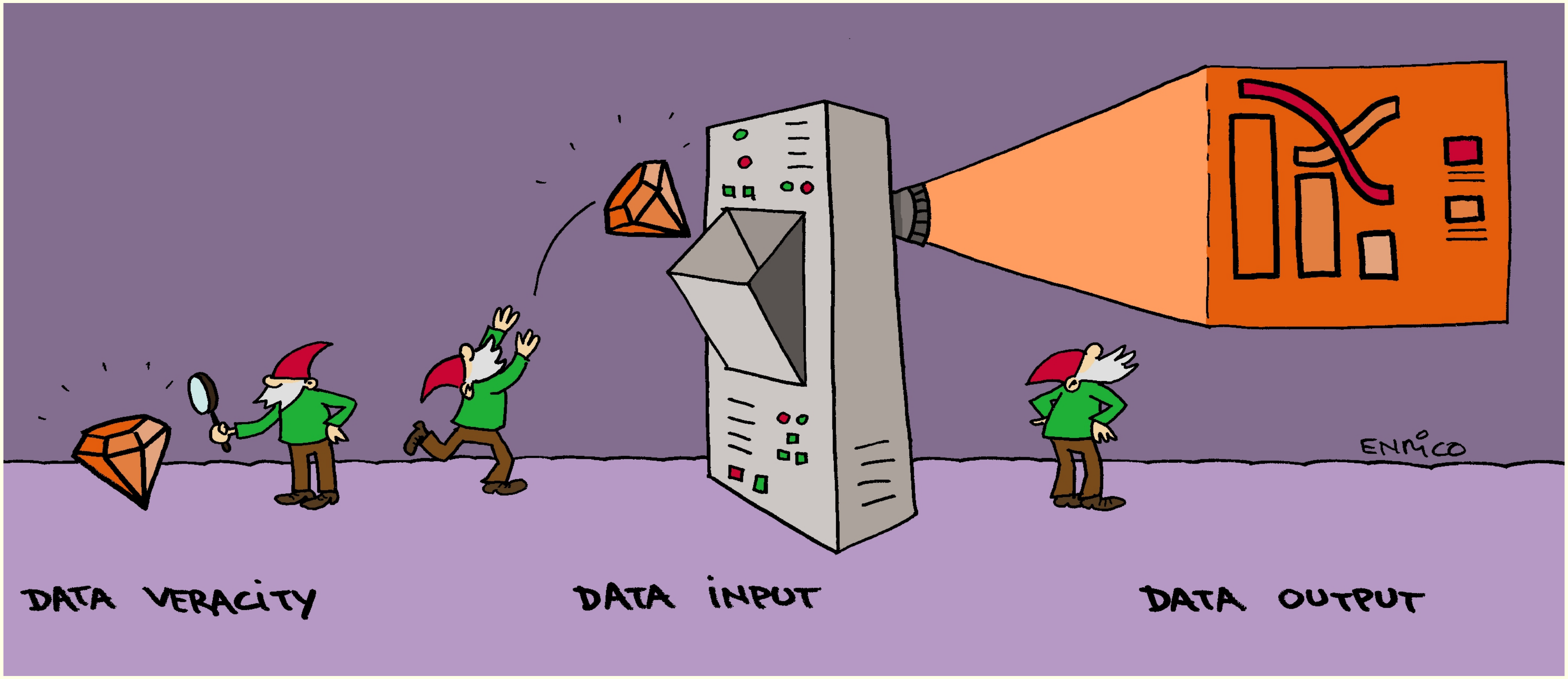
Distr.: General
17 December 2015

Original: English

Statistical Commission
Forty-seventh session
8-11 March 2016

Intermediate summary

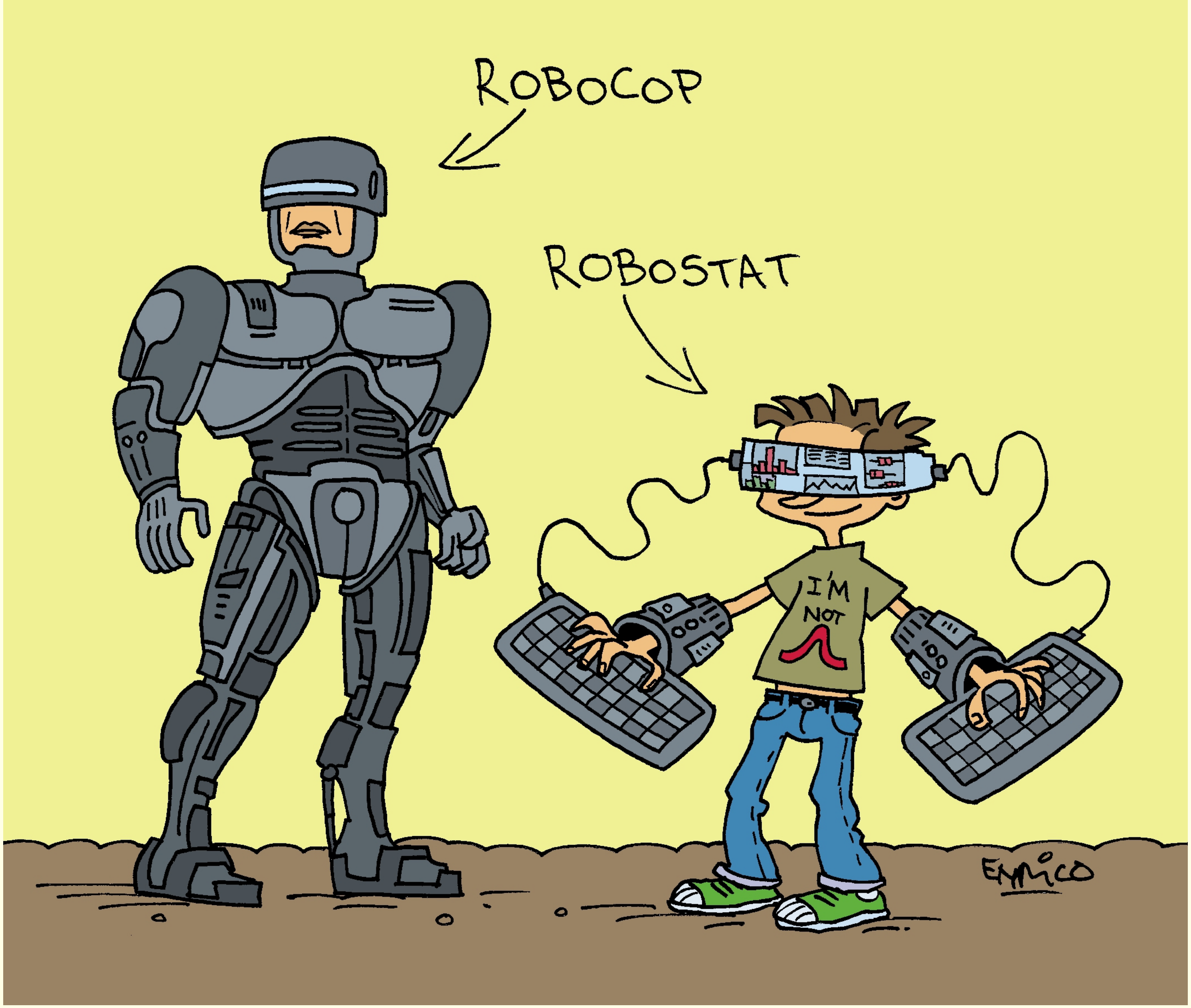
- In a world of (big) data, IoT (data) and also post-truth politics, the veracity of data, *i.e.* the trustworthiness of data (including the related data quality), is more important than ever!



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

- Data and analytics should be envisaged to complement and augment (official) statistics, not replacements for it!

⇝ 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.


ROBOCOP
ROBOSTAT
I'M NOT
ENRICO

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

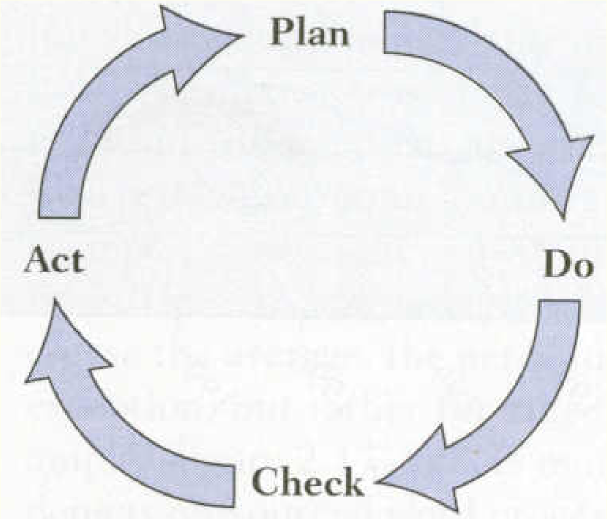
W. Edwards Deming

⇝ **Analytics is a** whole iterative problem solving and continuous improvement **cycle/process** that must be mastered through interdisciplinary and transdisciplinary team effort!

5. Process models for continuous improvement

- The $\boxed{\text{‘Plan–Do–Check–Act’}}$ (**PDCA**) cycle is often referred to as the Deming cycle, Deming wheel or the Shewhart cycle.

⇝ Walter A. Shewhart proposed this approach in the field of ‘quality control’ in the 1920s, and W. Edwards Deming later popularised PDCA as a general management approach based on the scientific method.

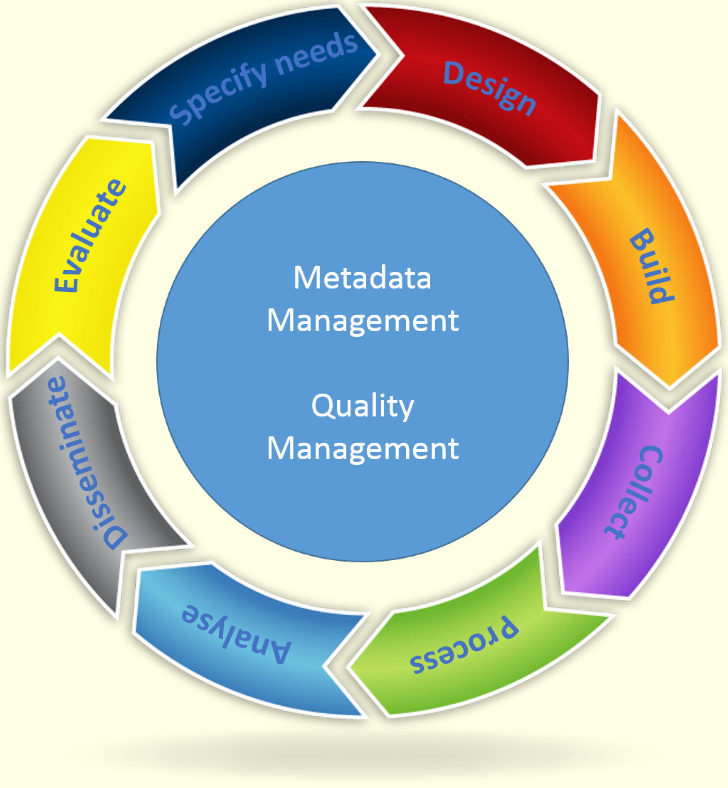


‘All improvement takes place project by project and in no other way.’

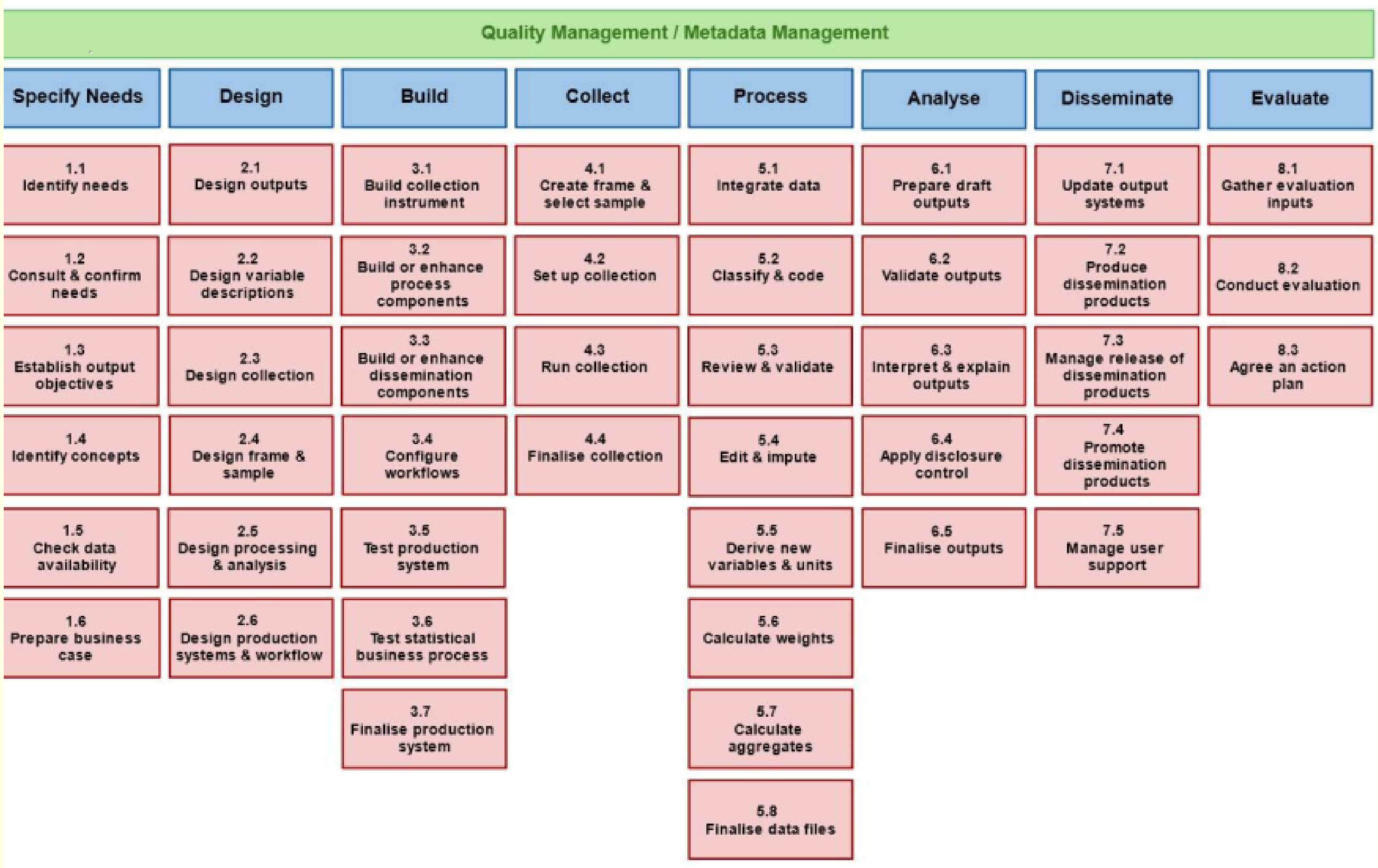
Joseph M. Juran, 1989

A process model for the production of official statistics

- The '**Generic Statistical Business Process Model**' (GSBPM) — coordinated through the 'United Nations Economic Commission for Europe' (Version 5.0 as of December 2013) — is consistent with the PDCA cycle:



- GSBPM is a key conceptual framework for the modernisation (and standardisation of the production) of official statistics.



Quality Management / Metadata Management

Specify Needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Build collection instrument	4.1 Create frame & select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult & confirm needs	2.2 Design variable descriptions	3.2 Build or enhance process components	4.2 Set up collection	5.2 Classify & code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Build or enhance dissemination components	4.3 Run collection	5.3 Review & validate	6.3 Interpret & explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame & sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit & impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing & analysis	3.5 Test production system		5.5 Derive new variables & units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare business case	2.6 Design production systems & workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production system		5.7 Calculate aggregates			
				5.8 Finalise data files			

⇝ But, where is the continuous (quality) improvement cycle?

‘The author believes the reason *[operational]* cost *[of different parts of the statistical business process]* has not been a central focus is a difference between NSIs focus on *measuring* quality of their products and services, rather than continuous *improvement* of quality.’

David A. Marker, 2017

Source: Marker, D. A. (2017). How have national statistical institutes improved quality in the last 25 years? *Statistical Journal of the IAOS*, 33, 951–961.

⇝ Emphasis needs to **move from measuring quality to improving quality**!

⇝ Moreover, the GSBPM is a deductive reasoning and a sequential approach.

⇝ For example, the first GSBPM steps are entirely focused on deductive reasoning for primary data collection and are not suited for inductive reasoning applied to (already existing) secondary data.

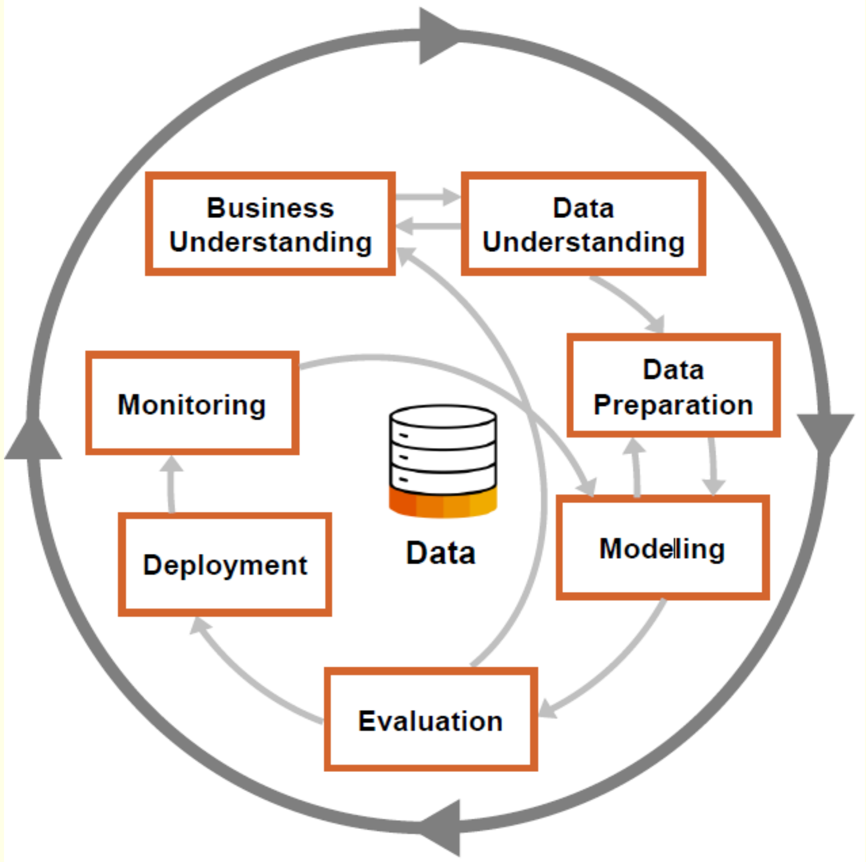
⇝ Moreover, the evaluation ('Evaluate' step) is only performed at the end.

⇝ This process model needs to be adapted to incorporate analytics by taking into account both approaches of analytics (*i.e.* inductive and deductive reasoning) and through the usage of, for example, data-informed continuous evaluation at any GSBPM step.

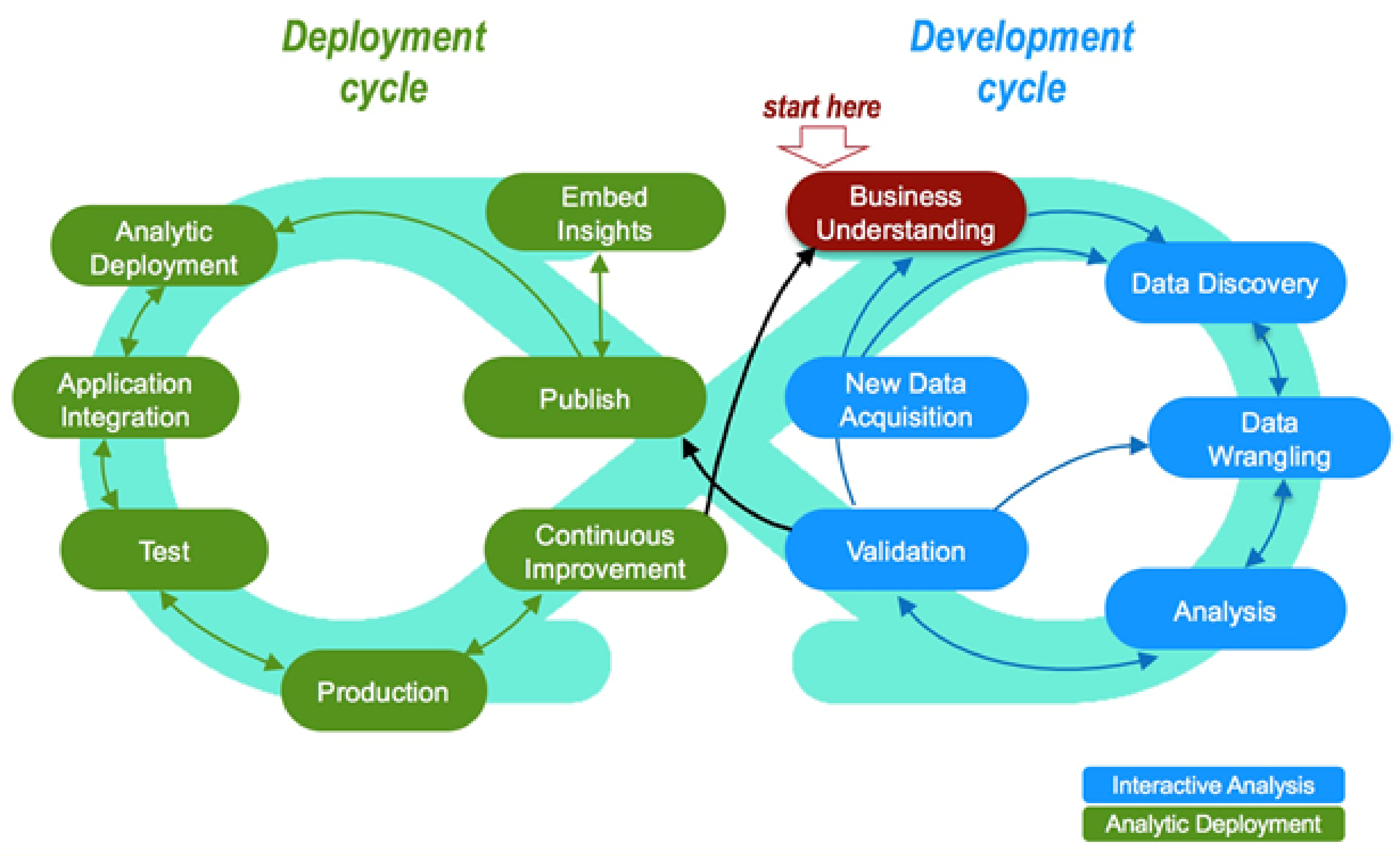
⇝ Current **production processes of official statistics need to be augmented and empowered by analytics (of things)!**

A process model for analytics

- The CRISP-DM (‘**CRoss Industry Standard Process for Data Mining**’) process — initially conceived in 1996 — is also consistent with the PDCA cycle:



The complementary cycles of developing & deploying ‘analytical assets’

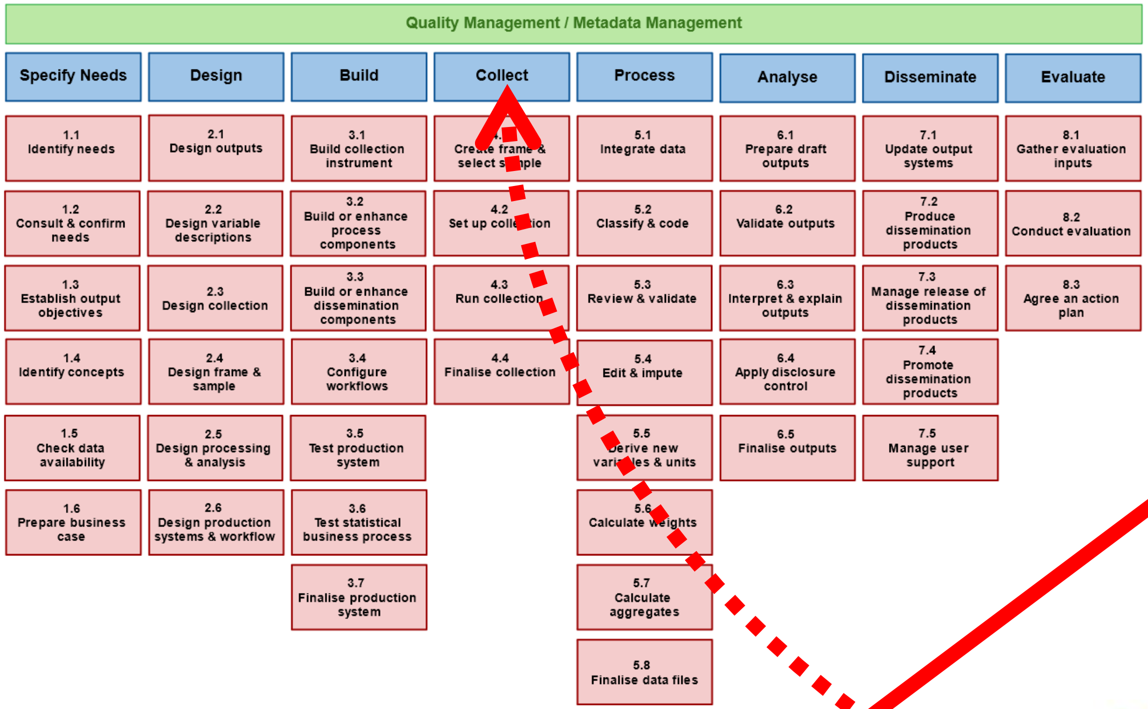


Source: Erick Brethenoux, Director, IBM Analytics Strategy & Initiatives, August 18, 2016 (`goo.gl/AhsG1n`).

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

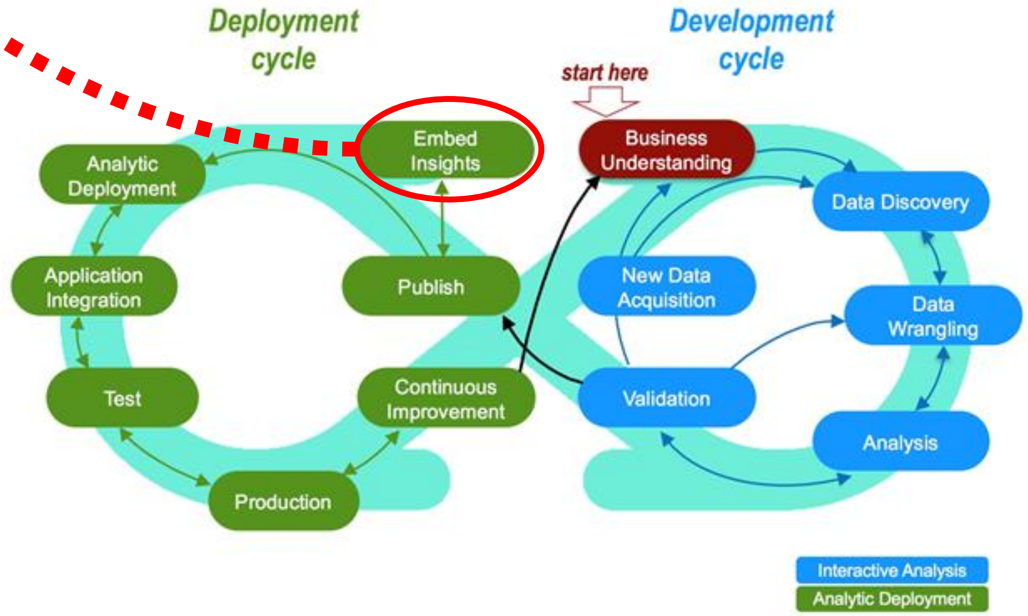
Thomas H. Davenport and Jinho Kim, 2013

GSBPM («current statistical production»)



Quality Management / Metadata Management

Specify Needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Build collection instrument	4.1 Create frame & select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult & confirm needs	2.2 Design variable descriptions	3.2 Build or enhance process components	4.2 Set up collection	5.2 Classify & code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Build or enhance dissemination components	4.3 Run collection	5.3 Review & validate	6.3 Interpret & explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame & sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit & impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing & analysis	3.5 Test production system		5.5 Derive new variables & units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare business case	2.6 Design production systems & workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production system		5.7 Calculate aggregates			
				5.8 Finalise data files			

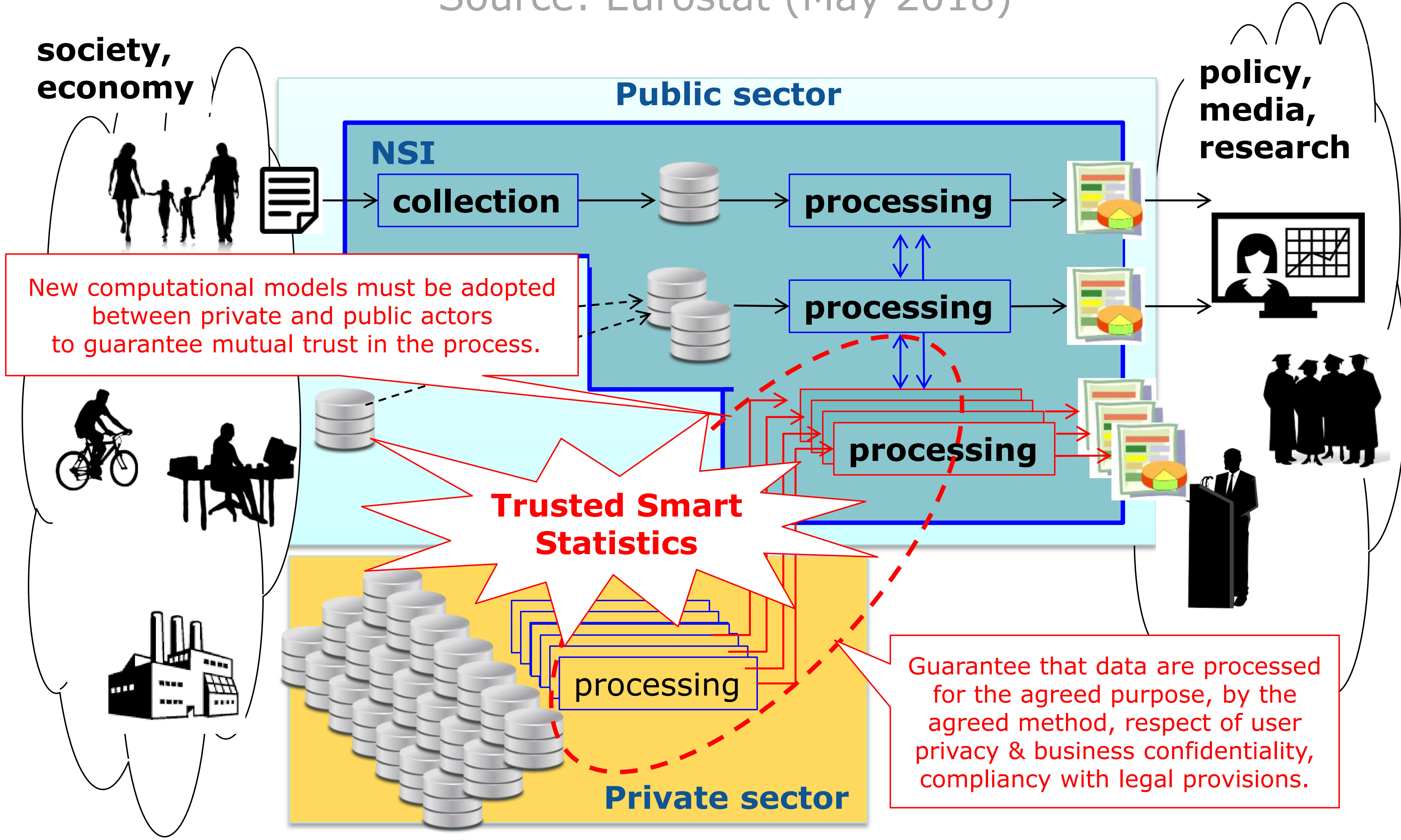


«Analytics process model»

«Handle the new in new ways»

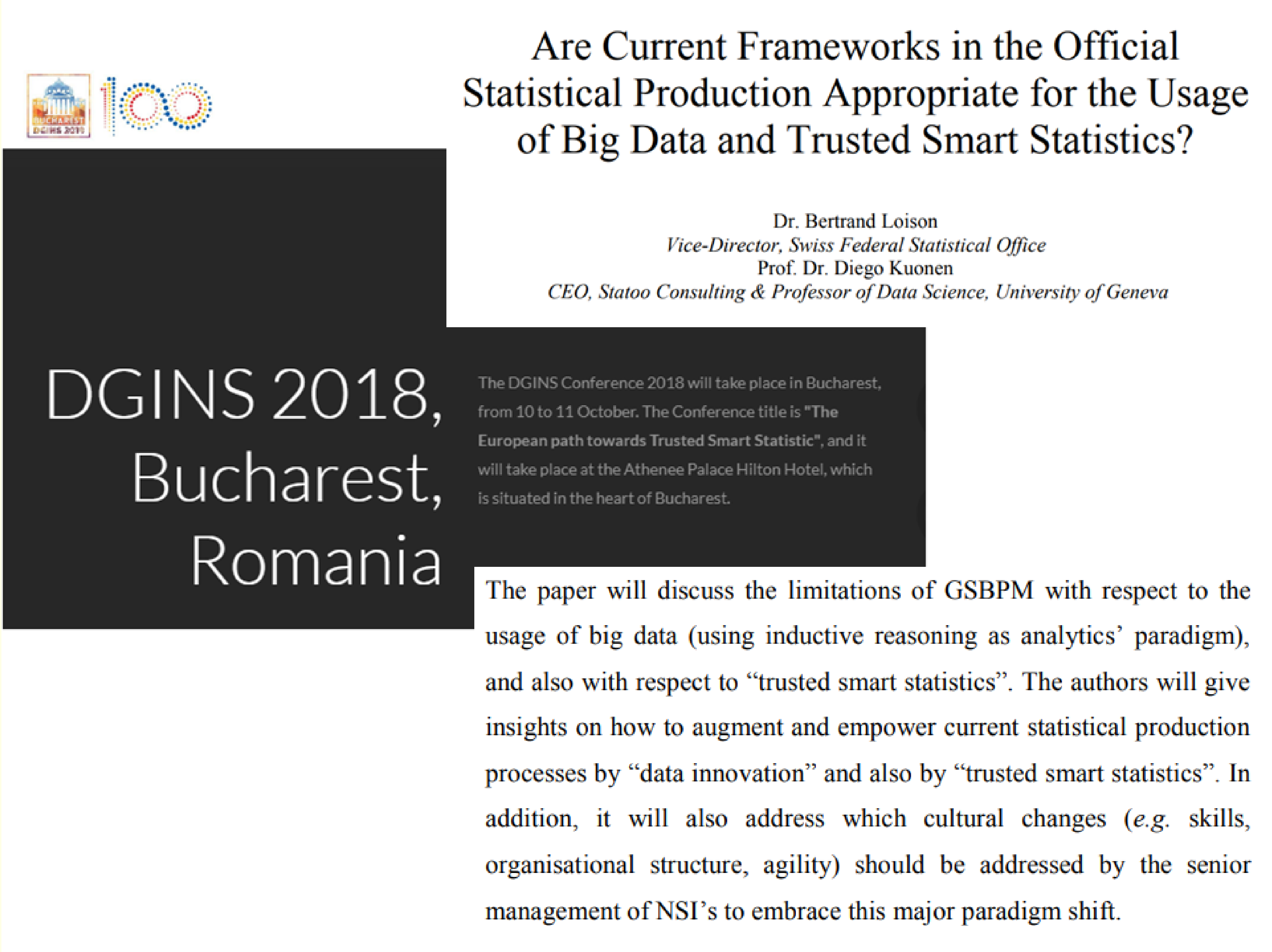
«Push computation out (partially)»

Source: Eurostat (May 2018)



New computational models must be adopted between private and public actors to guarantee mutual trust in the process.

Guarantee that data are processed for the agreed purpose, by the agreed method, respect of user privacy & business confidentiality, compliancy with legal provisions.



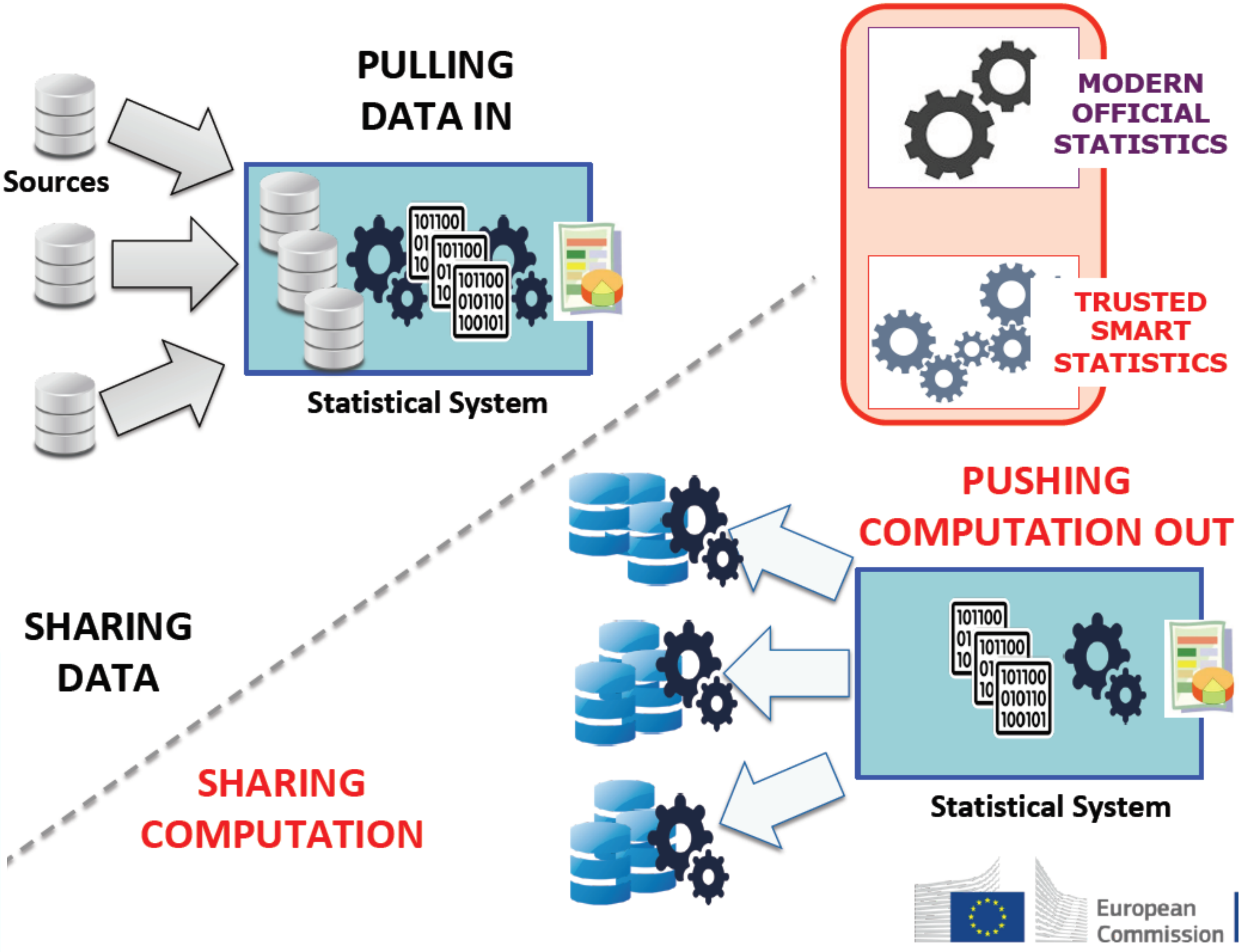
Are Current Frameworks in the Official Statistical Production Appropriate for the Usage of Big Data and Trusted Smart Statistics?

Dr. Bertrand Loison
Vice-Director, Swiss Federal Statistical Office
Prof. Dr. Diego Kuonen
CEO, Statoo Consulting & Professor of Data Science, University of Geneva

DGINS 2018, Bucharest, Romania

The DGINS Conference 2018 will take place in Bucharest, from 10 to 11 October. The Conference title is **"The European path towards Trusted Smart Statistic"**, and it will take place at the Athenee Palace Hilton Hotel, which is situated in the heart of Bucharest.

The paper will discuss the limitations of GSBPM with respect to the usage of big data (using inductive reasoning as analytics' paradigm), and also with respect to "trusted smart statistics". The authors will give insights on how to augment and empower current statistical production processes by "data innovation" and also by "trusted smart statistics". In addition, it will also address which cultural changes (*e.g.* skills, organisational structure, agility) should be addressed by the senior management of NSI's to embrace this major paradigm shift.



PULLING DATA IN

Sources

Statistical System

MODERN OFFICIAL STATISTICS

TRUSTED SMART STATISTICS

PUSHING COMPUTATION OUT

Statistical System

SHARING DATA

SHARING COMPUTATION

Deep data and shared computation: Shaping the future Trusted Smart Statistics

Fabio Ricciato
Albrecht Wirthmann
Michail Skaliotis
Fernando Reis
Konstantinos Giannakouris

EUROSTAT

NTTS 2019, Brussels 12.3.2019

⇝ As soon as it works, no one calls it ‘production process of official statistics empowered by analytics and augmented by analytics of things’ any more!

‘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?

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: 'UN.BigData.2019'.

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

Compilation date: 12.04.2019.