Cybersecurity in the Era of Big Data, Machine Learning and Artificial Intelligence

Prof. Dr. Diego Kuonen, CStat PStat CSci

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

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

About myself (about.me/DiegoKuonen)

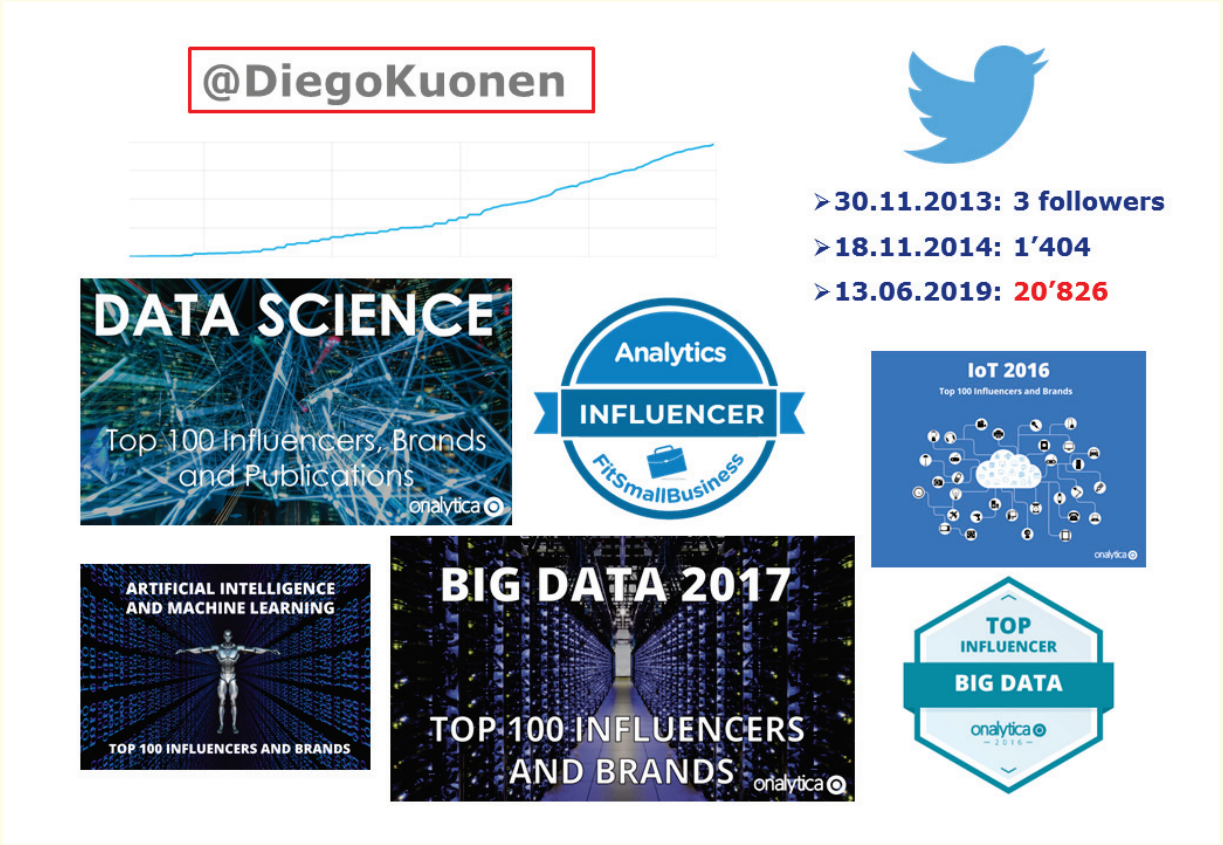
- ◇ PhD in Statistics, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.
- ◇ MSc in Mathematics, EPFL, Lausanne, Switzerland.

- 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).

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.



About Statoo Consulting (www.statoo.info)

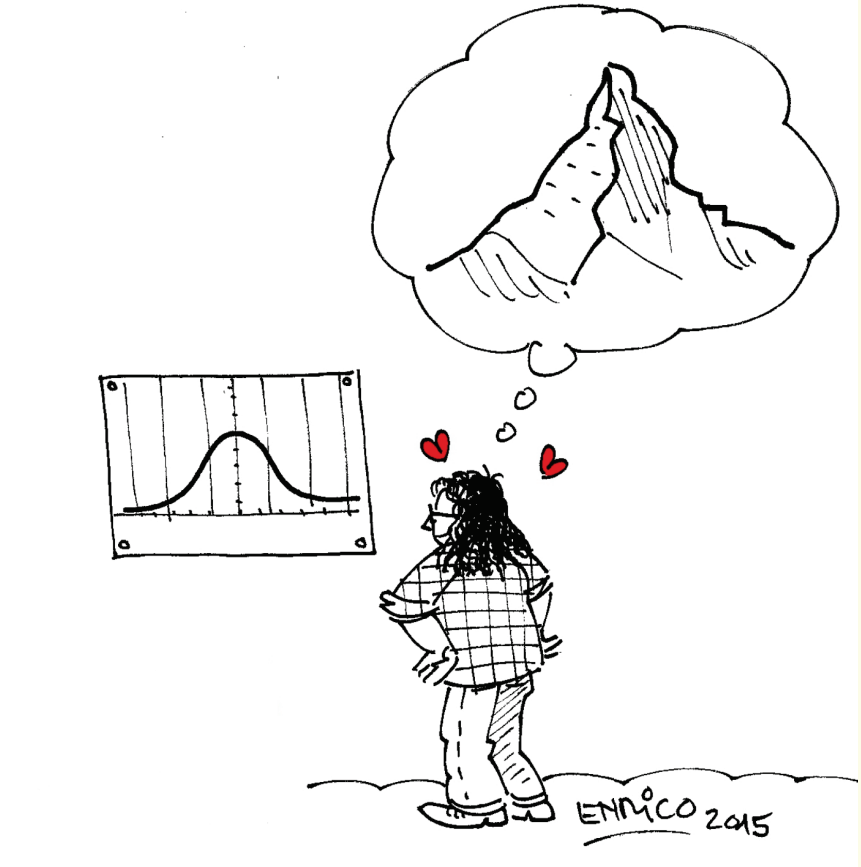
- Founded Statoo Consulting in **2001**.

$$\rightsquigarrow 2019 - 2001 = 18 + \epsilon.$$

- 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.

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

$\rightsquigarrow$ Have you ever been Statooed?



'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

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

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'?

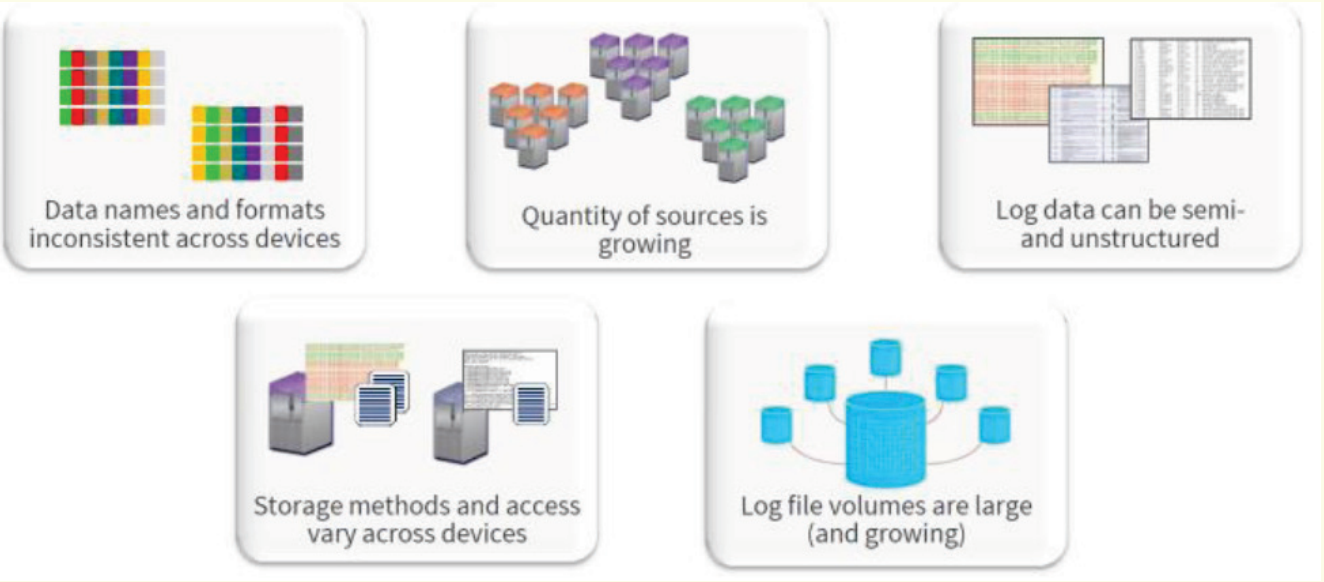
> ◇ The term 'big data' applies to an accumulation of data that can not be processed or handled using traditional data management processes or tools.

⇝ 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-driven cybersecurity').

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

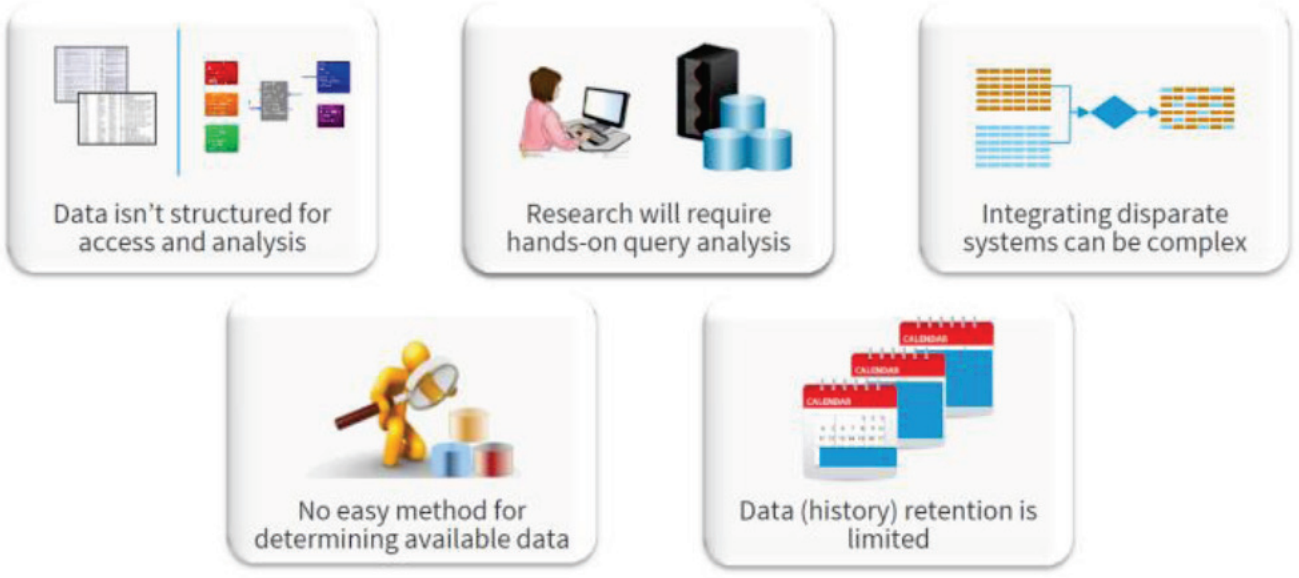
Data-driven cybersecurity starts with data management!

⇝ Examples of cybersecurity data sources:



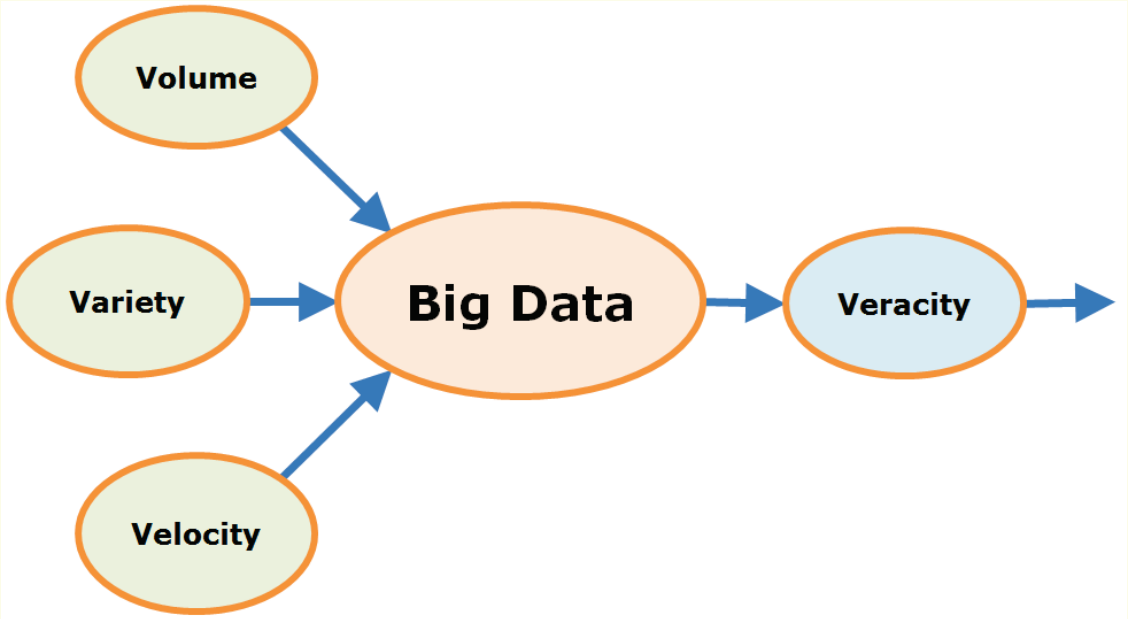
Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

⇝ Examples of cybersecurity data access and usage:



Source: Christopher Smith & Evan Levy, 'Stronger Cybersecurity Starts with Data Management', *International Institute for Analytics*, 2016 (goo.gl/3oogK8).

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.



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

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

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

‘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.’

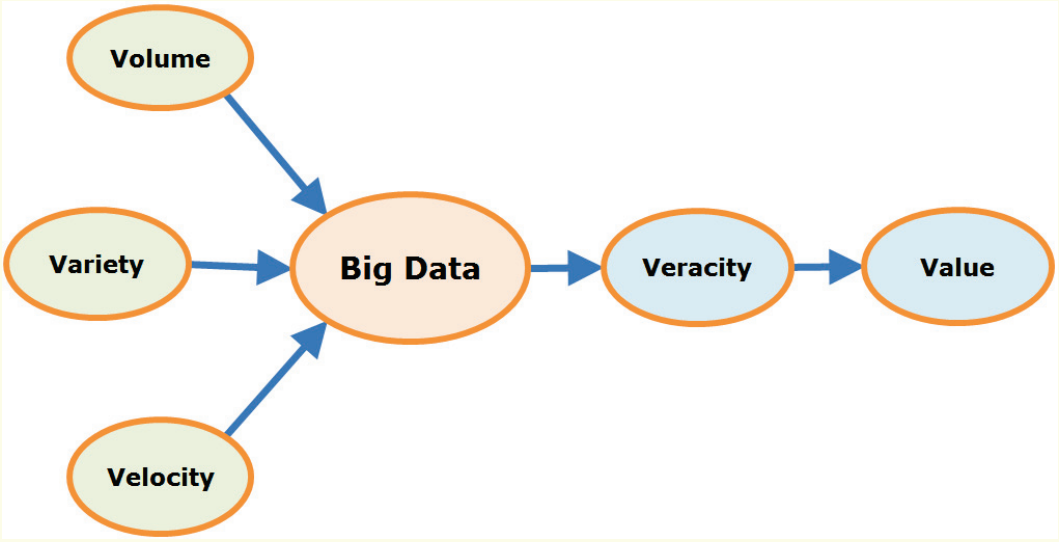
Swiss Federal Council, September 5, 2018

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

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

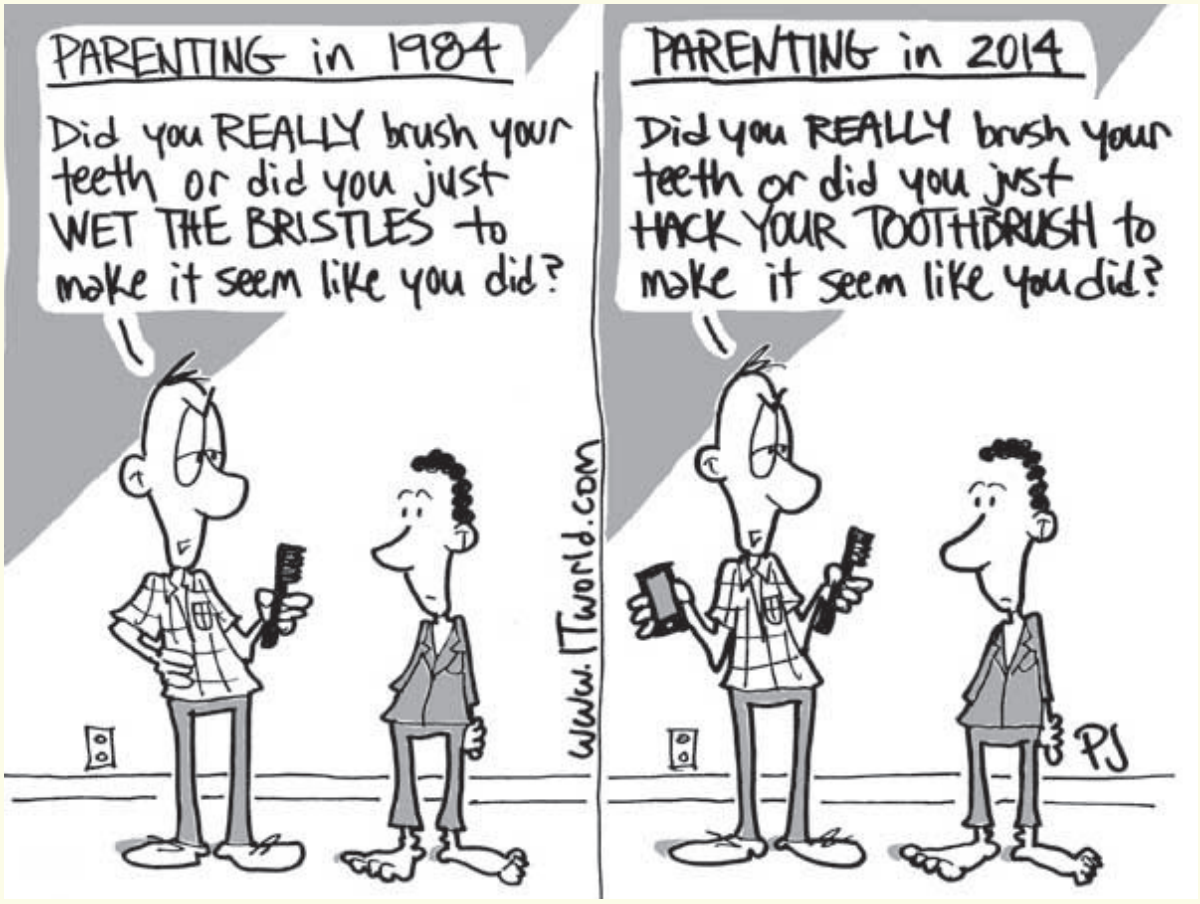
Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

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.

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.



Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

2. Demystifying the 'Internet of things' hype

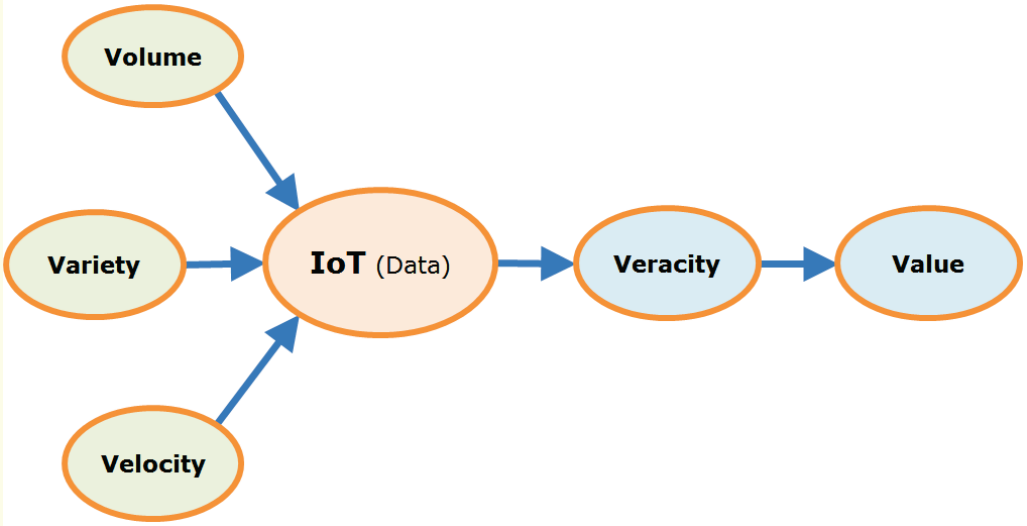
• The term 'Internet of Things' (IoT) — coined in 1999 by the technologist Kevin Ashton — starts acquiring the trappings of a 'new religion'!

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

⇝ However, IoT is about data, not things!

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

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).

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

‘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 **data-driven cybersecurity**!

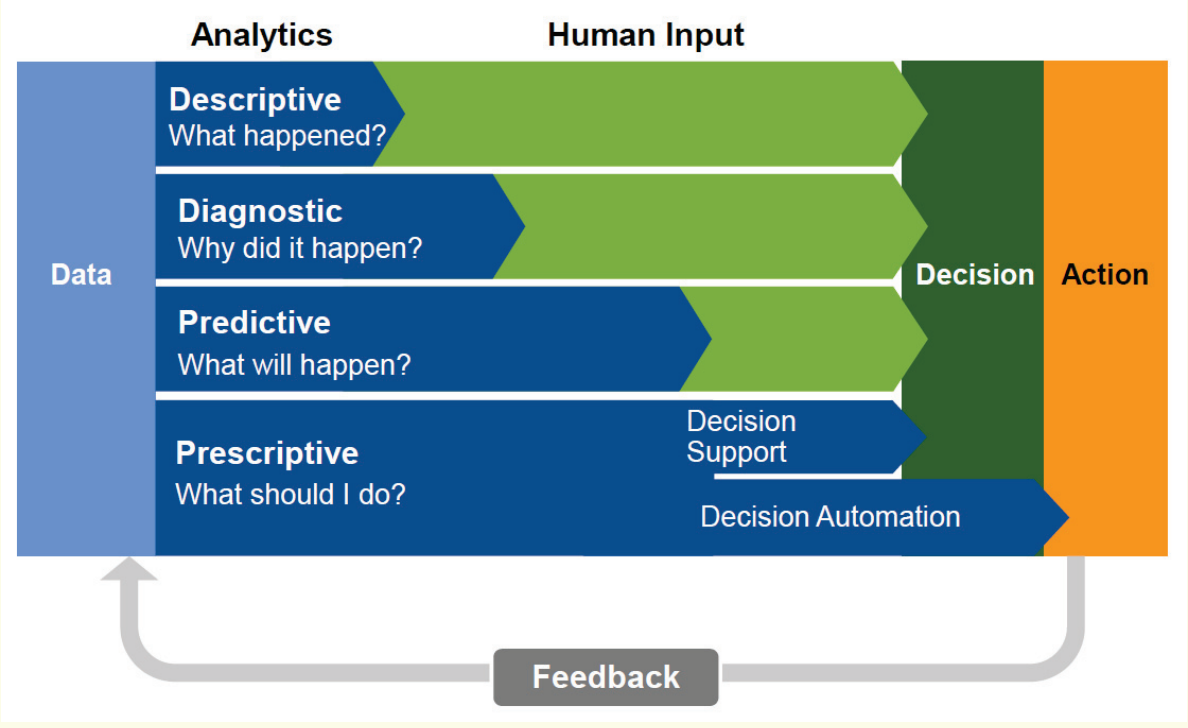
Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

‘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

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

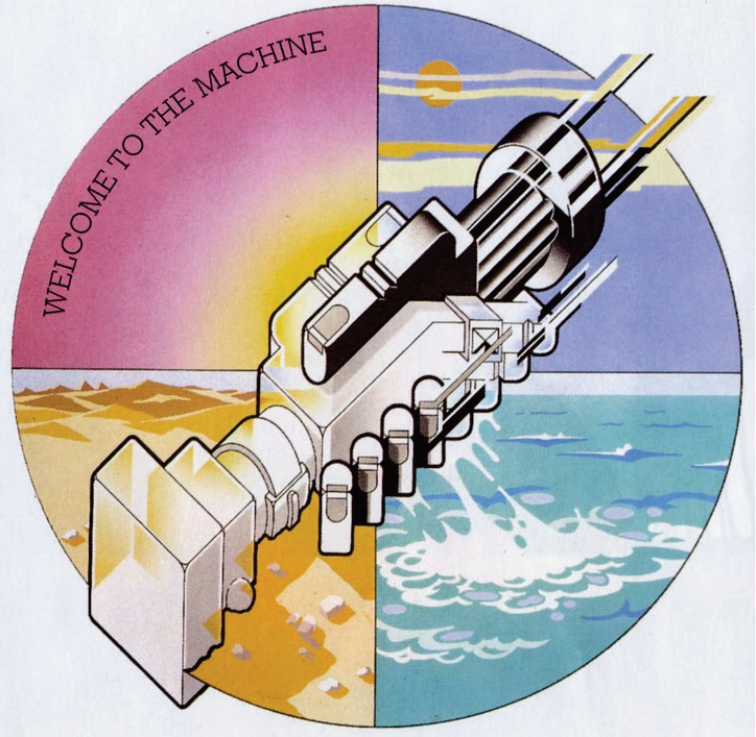
Questions analytics tries to answer & the ‘analytics continuum’



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

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

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



Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

3. Demystifying the ‘machine intelligence and learning’ hype

◇ John McCarthy, one of the founders of ‘Artificial Intelligence’ (AI) (now sometimes referred to as ‘machine intelligence’) 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’,

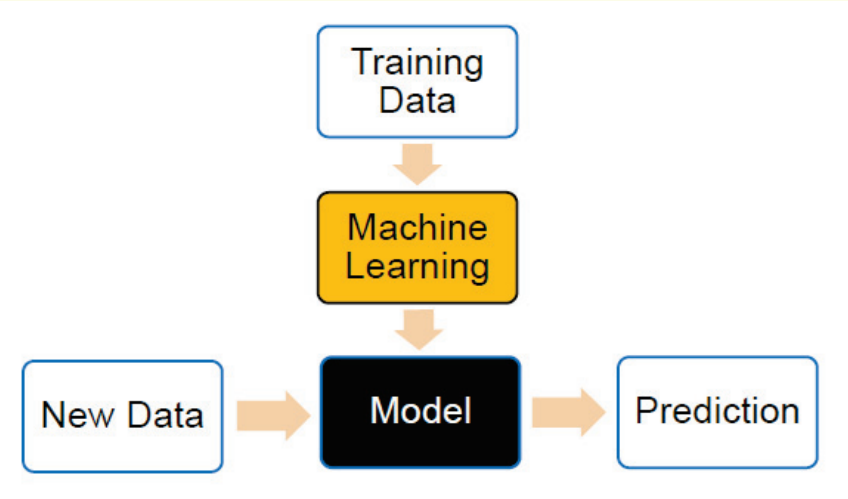
e.g. visual perception, speech recognition, language translation, visual translation and playing games (with concrete rules).

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

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

◇ 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, and help make decisions.



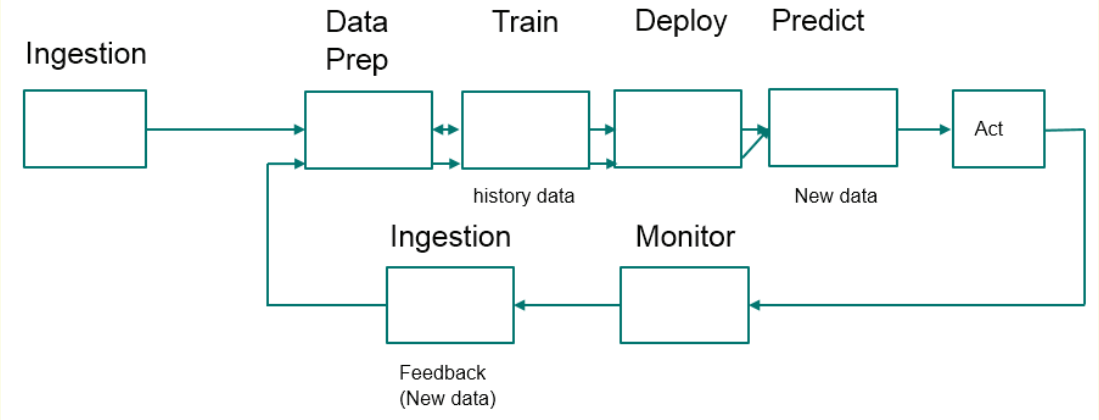
Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

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

Albert Einstein

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

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).

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

⇝ Some examples of common (data-driven) cybersecurity use cases of such '(machine) learning systems':

Malware infection	Detecting malware, especially unseen variants, is a daunting challenge in organizations with a large number of endpoints, many users, and a wide public Internet presence. Machine intelligence approaches can learn characteristics of malware previously observed in order to predict potential malware infections that signature-based approaches would miss.
Network anomalies	High volumes of traffic traverse typical networks (internally and externally) each day and it is difficult to distinguish benign traffic from malicious or risky activities. By employing machine intelligence, deviations from normal network traffic can be extracted in real time and evaluated by algorithms, saving massive amounts of time manually sifting through logs.

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

Intrusion detection	Sophisticated threat actors are capable of intruding into networks and covering their tracks to look like a typical user, which makes detection and remediation very difficult. By modeling patterns seen in malicious traffic, machine intelligence can learn over time, so intrusion detection can get ahead of the threat, rather than requiring frequent rule updates.
Rank aggregation	Many organizations are successfully able to implement first-order analytics (queries, statistics, patterns), but in isolation these datasets miss the big picture threat landscape. Advanced machine intelligence analytical techniques can learn how to integrate multiple analytic data products to tell a more cohesive story regarding the aggregate threat.
Deep packet inspection	Network threats are continually evolving, and organizations must move past signature matching to uncover malicious content contained within network packets at network speed. Modern computing architectures, such as GPUs, are being designed specifically for machine intelligence workloads at an attainable price-point using open source software.

Source: Guerra, P. & Tamburello. P. (2018). *Modernizing Cybersecurity Operations with Machine Intelligence*. O'Reilly Media, Sebastopol, CA.

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

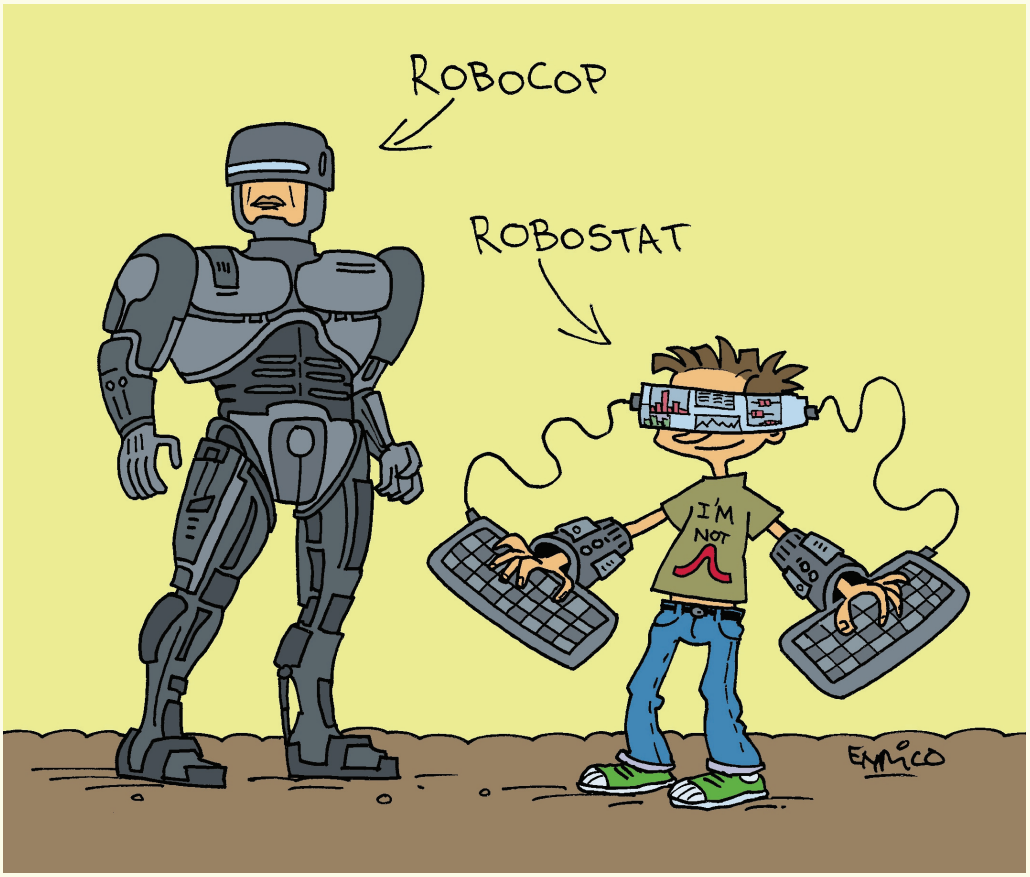
- However, without humans as a guide, current AI is no more capable than a computer without software!
- AI without trustworthy data is like a swimming pool without trustworthy water!
- There is nothing artificial about AI: it is inspired by humans, it is created by humans and impacts humans!
- Data-driven cybersecurity starts with trust, simple because data collected for analytics must be trusted!

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

‘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

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.



Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

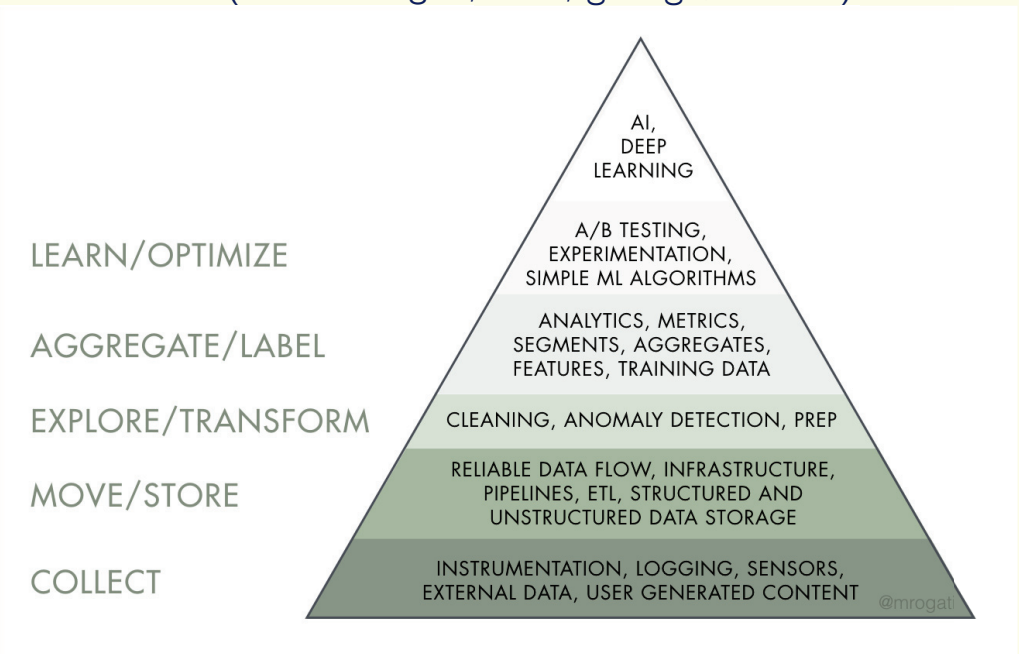
‘AI algorithms are not natively ‘intelligent’. They learn inductively by analyzing data. ... Sophisticated algorithms can sometimes overcome limited data if its quality is high, but bad data is simply paralyzing.’

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).

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

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

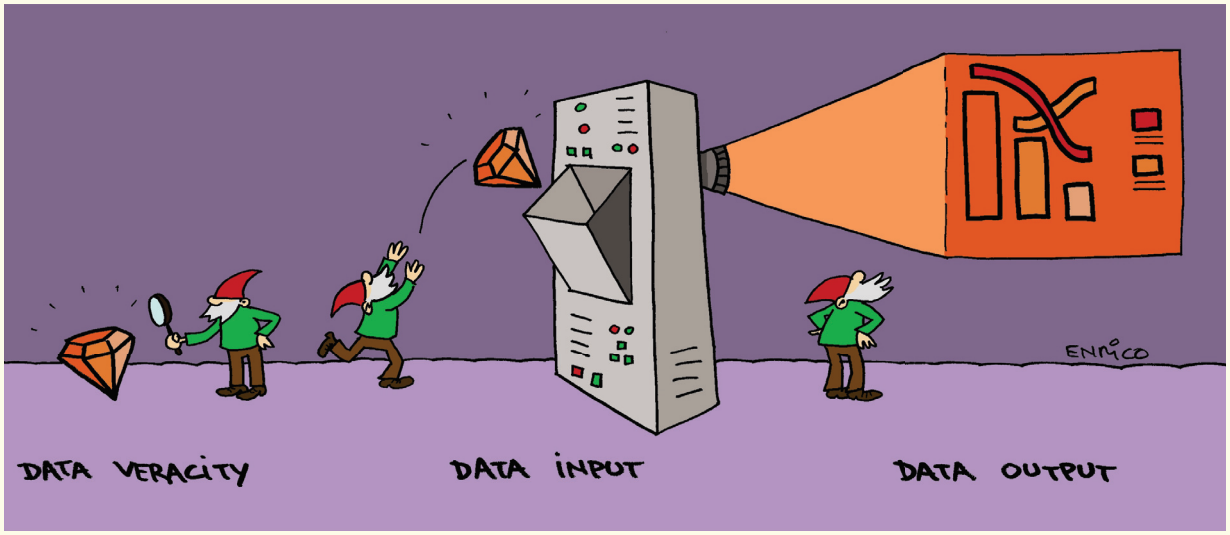


⇝ **Data should be treated as a key strategic asset, so ensuring their veracity and the related data quality become imperative**!

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

Conclusion, challenges and opportunities

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

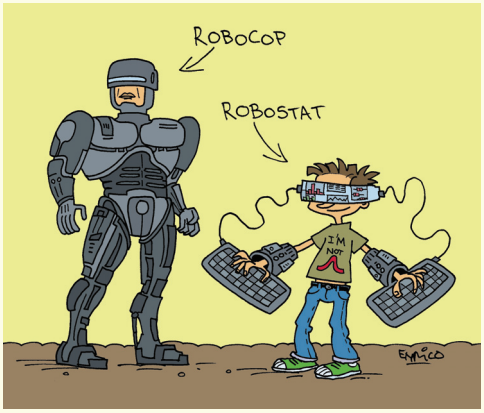


Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

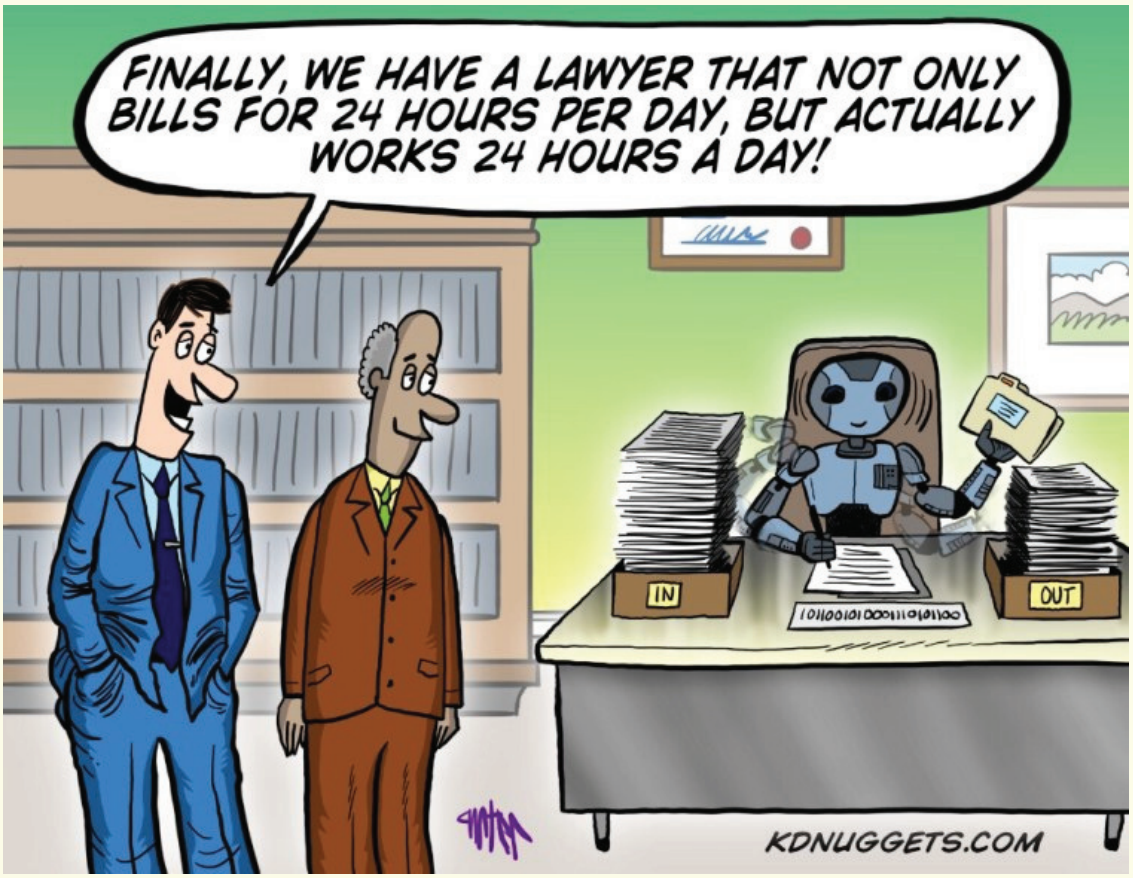
• Analytics is an aid to thinking and not a replacement for it!

• Data and analytics should be envisaged to complement and augment humans to strengthen existing cybersecurity setups, *e.g.* by extending existing security rules, and not replacements for them!

⇝ **Humans (and cybersecurity teams as well!) need to augment their strengths to become more 'powerful'**: by automating any routinisable work and by focusing on their core competences.



Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.



Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

'It is getting better... A little better all the time.'

The Beatles, 1967

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

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

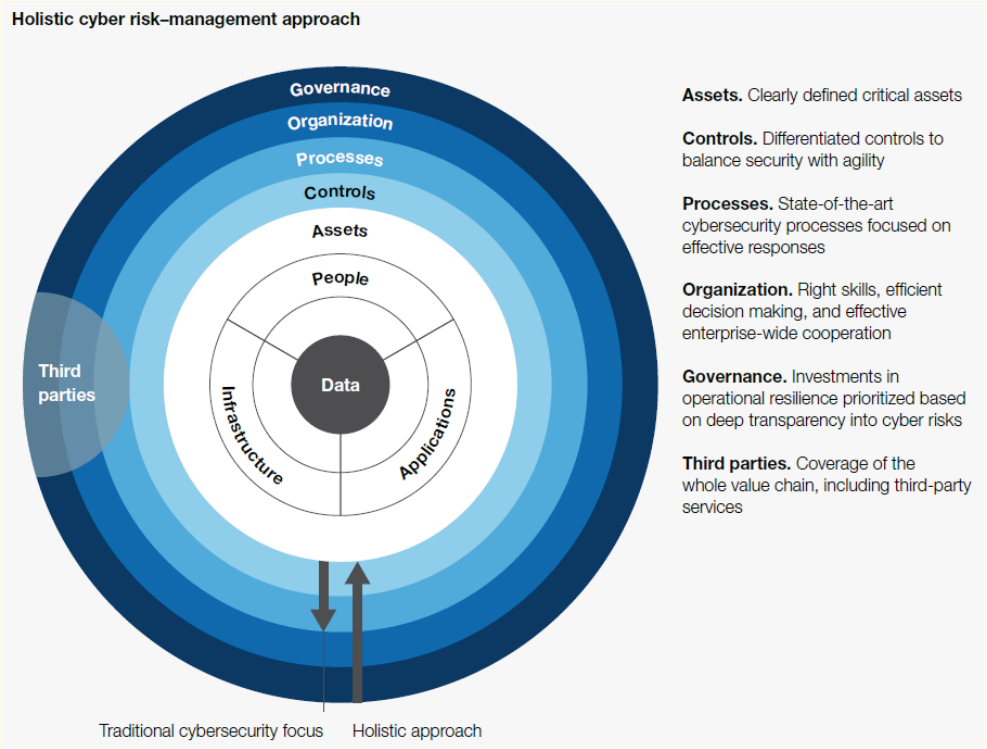
George Westerman, 2018

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

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

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

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.



Source: Boehm, T., Merrath, P., Poppensieker, T., Riemenschnitter, R. & Stähle, T. (2018). *Cyber Risk Measurement and the Holistic Cybersecurity Approach*. McKinsey & Company (goo.gl/2Dj2M7).

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

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

Albert Einstein

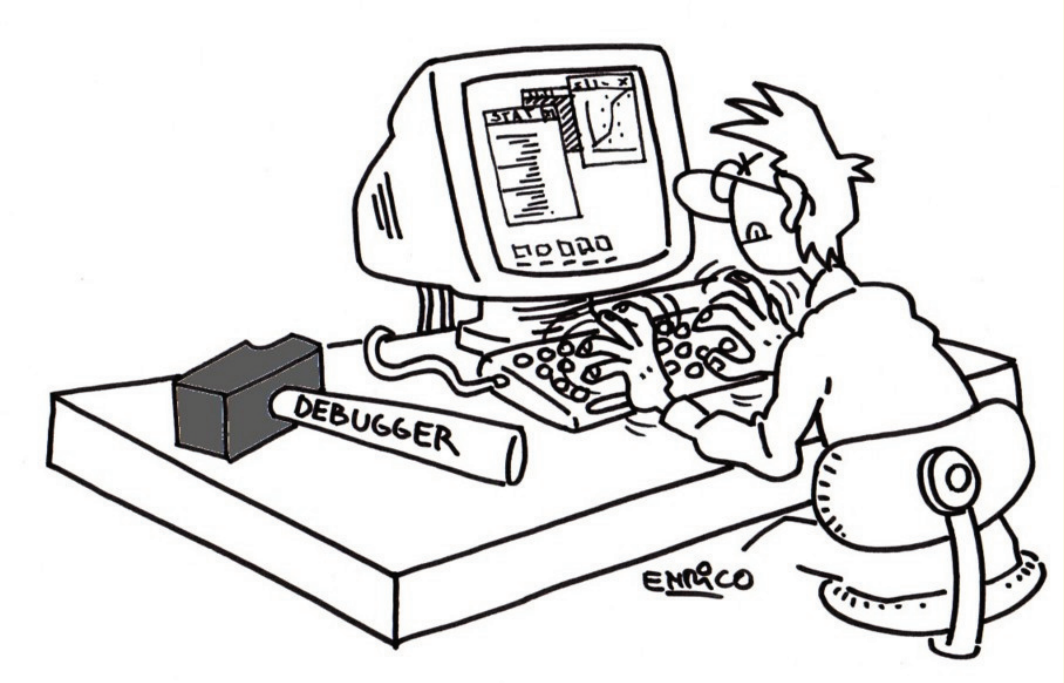
Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

'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

Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.



Copyright © 2001–2019, Statoo Consulting, Switzerland. All rights reserved.

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: 'Geneva.CybersecurityLiability.Jun20.2019/MyKeynote'. Typesetting: LaTeX, version 2ε. Compilation date: 13.06.2019.