A Statistician's View on Data Quality

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

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

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

'1st Data Science Seminar: Exploring Data Quality' — November 11, 2020

‘In God we trust; all others must bring data.’

W. Edwards Deming

‘Data is a key raw material in the knowledge society and the digital economy. ... The trustworthiness, security, serviceability, accessibility, verifiability and availability of data therefore become key concerns in a digitalised society.’

Swiss Federal Council, September 11, 2020

Source: ‘*Digital Switzerland Strategy*’, adopted by the Federal Council on September 11, 2020 (`bit.ly/2U95yuG`).

Provenance of the word *data*

- The word *data* is derived from the Latin meaning 'given'.

$\rightsquigarrow$ However, instead of considering data as given — as data are never simply given — it would be more appropriate to **think of data as taken**, for which the Latin would be *capta*.

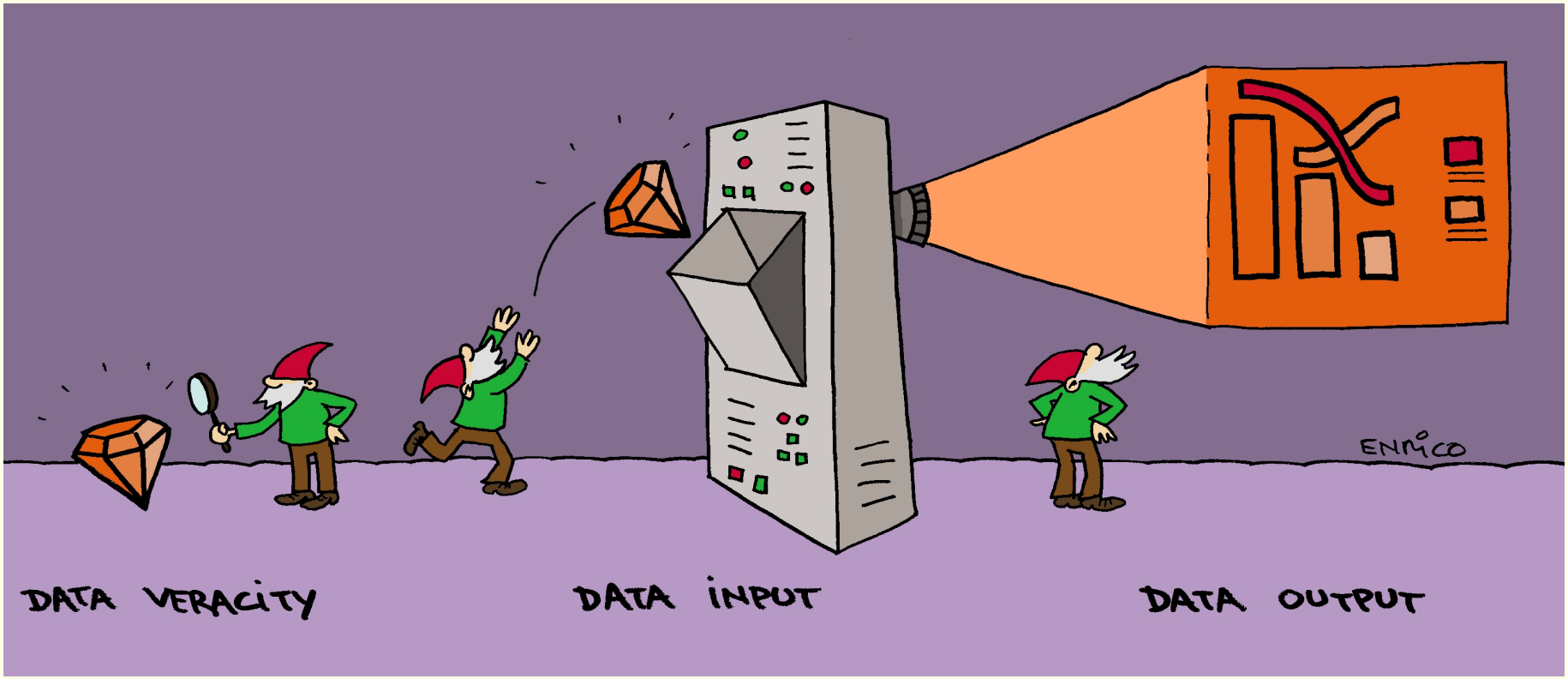
$\rightsquigarrow$ How data are construed, recorded and collected is the result of human decisions, *i.e.* decisions about what exactly to measure, when and where to do so and by what methods.

$\rightsquigarrow$ Inevitably, what gets measured and recorded has an impact on the conclusions that are drawn.

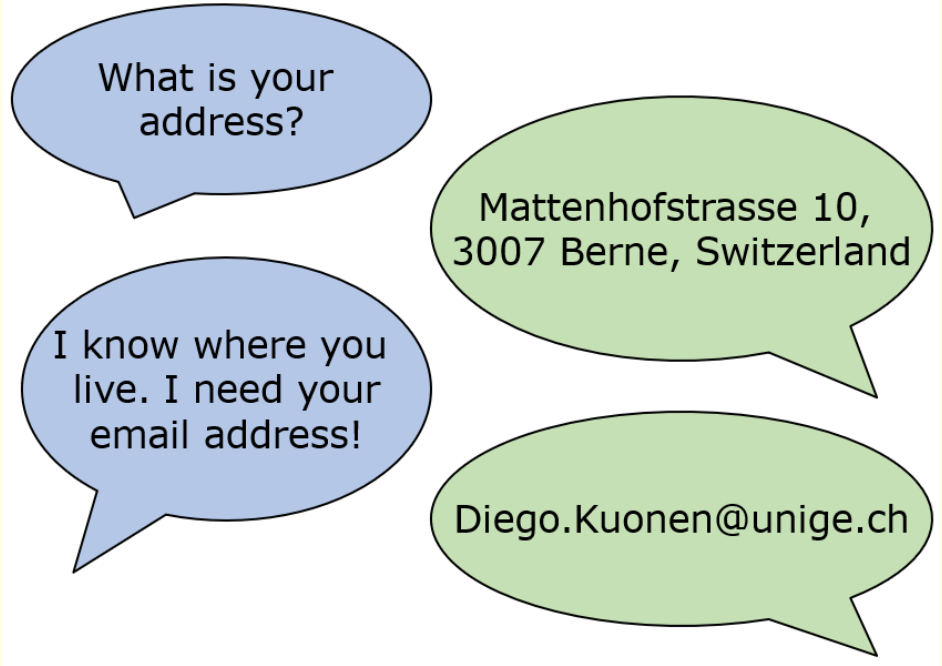
Source: Barrowman, N. (2018). Why data is never raw. *The New Atlantis*, 56, 129–135 (see `bit.ly/2s2I7Fd`).

Data veracity and data quality more important than ever!

• The **data veracity**, *i.e.* the trustworthiness of data, including the reliability ('quality over time'), capability, validity and security of the data, and the related data quality are more important than ever!



What is data quality?



$\rightsquigarrow$ 'Data quality is in the eye of the beholder', *i.e.* the customer.

$\rightsquigarrow$ There are two different aspects of data quality: 'the data being right' and 'the right data' for the task at hand.

Source: Thomas C. Redman, *What in the World is Data Quality?* Medium, November 2020 (`bit.ly/36RsHaL`).

The definition of data quality

• Data are of high quality if they are 'fit for use' by **customers** (*i.e.* anyone who uses the data) in their intended operational, decision making, planning and other roles ($\rightsquigarrow$ 'fitness for use' or 'fit for purpose').

> 'Data is of high quality if it is fit for its intended use (by customers) in operations, analytics, decision making, and planning. To be fit for use, data must be 'free from defects' (*i.e.* 'right') and 'possess desired features' (*i.e.* be the 'right data').'
>
> Thomas C. Redman, 2016

For Data, Only Two Moments Really Matter

Critical roles: Data creator and data customer.
When people step into these roles, quality improves quickly!

The Moment of Creation

The Moment of Use

PATH FROM CREATOR TO CUSTOMER

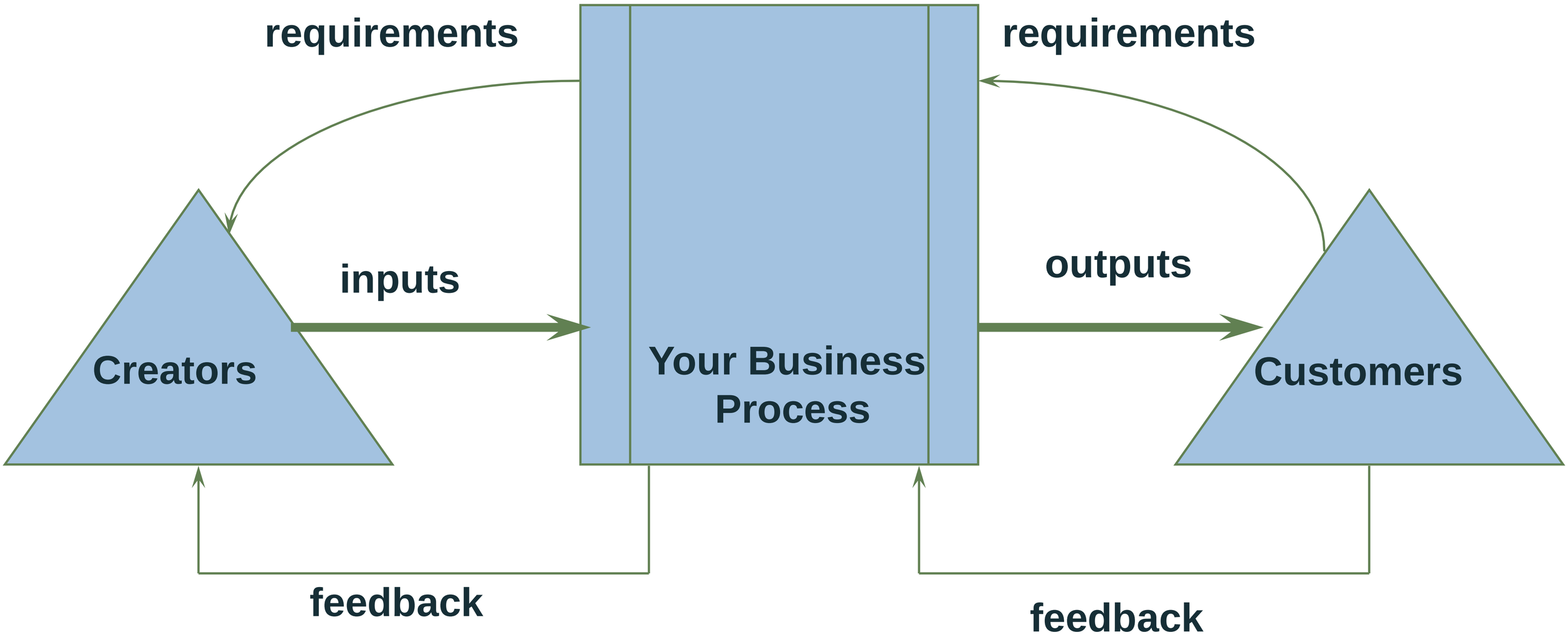
The whole point of data quality management is to connect the two!

DATA CREATOR

DATA CUSTOMER

The most important tool in all of data quality management (and maybe all management)

CUSTOMER-SUPPLIER MODEL



Everyone Bears Four Basic Accountabilities

1. **Be a good customer**: Ensure the data you need from the outside is of high quality.
2. **Be a good creator**: Ensure that the data you/your process creates meets the needs of the next person/process in line.
3. **Build the communications channels** necessary to be a good customer and creator.
4. **Understand the impact of bad data** on your work and grow increasingly intolerant of its impact.

‘You can not understand the data unless you first understand the process that produced it.’

Roger W. Hoerl and Ronald D. Snee, 2019

'Always ask yourself, 'Do I really understand how the data were collected? Can I trace back and identify the origin of each data point?' A good principle to remember is that data are guilty until proven innocent, not the other way around.'

Ronald D. Snee and Roger W. Hoerl, 2012

• **Data quality is not an one time project but a continuous improvement process** (which itself is a process that must be managed!).

⇝ Data quality management is an ongoing process that requires continuous data monitoring and reporting.

⇝ 'Data quality management is **a marathon, not a sprint**!'

‘Blame the process not the person. We need to ask, ‘how did the process allow this to happen?”

Brian L. Joiner, 1994

‘Failure is only the opportunity to begin again more intelligently.’

Henry Ford

‘Quality comes not from inspection but from improvement of the process.’

W. Edwards Deming

‘In god we trust; all others bring *fit for use* data.’

Modified version of the original quote by W. Edwards Deming: ‘In god we trust; all others bring data.’

Have you been Statooed & GSEMed?

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

Statoo Consulting
Morgenstrasse 129
3018 Berne
Switzerland

GSEM, RCS, 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: '1st.DS.Seminar.Nov11.2020'. Typesetting: LaTeX, version 2ϵ. Compilation date: 20.11.2020.