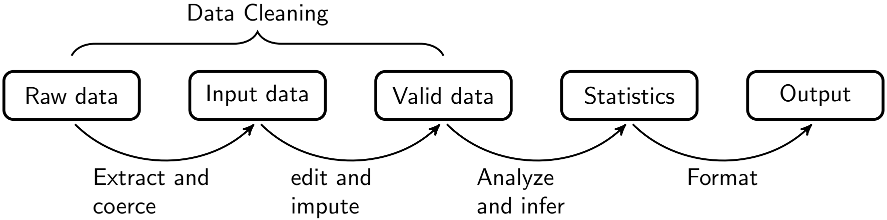
# Theory of data validation

Mark van der Loo and Edwin de Jonge

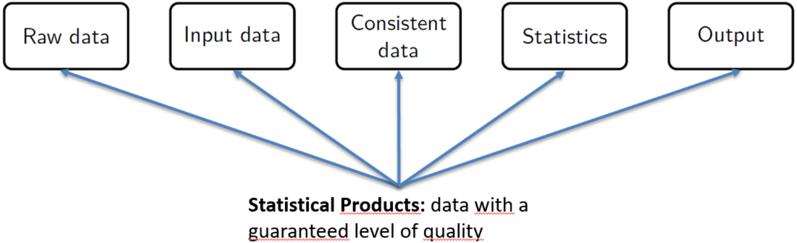
Statistics Netherlands Research & Development
@markvdloo @edwindjonge

useR!2021

# Statistical Value Chain



# Statistical Value Chain


Raw data

Input data

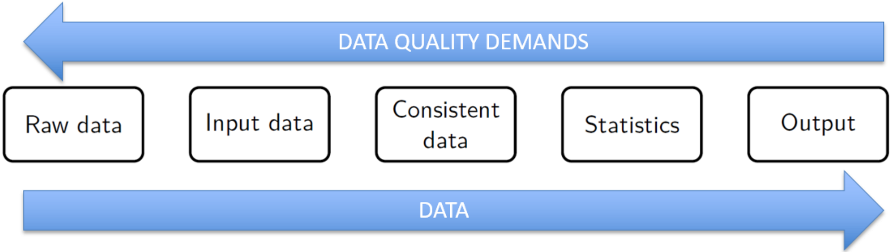
Consistent data

Statistics

Output

**Statistical Products:** data with a guaranteed level of quality

# Statistical Value Chain



# Data validation

*Data validation is an activity in which one verifies whether a combination of values is acceptable.*

## Examples

- Is the *Age* nonnegative?
- Does *Turnover* − *Cost* equal *Profit*?
- Is the average *Profit* positive?
- Does the mean *Profitratio* differ less than 10% from last year's?

# Why data validation rules?

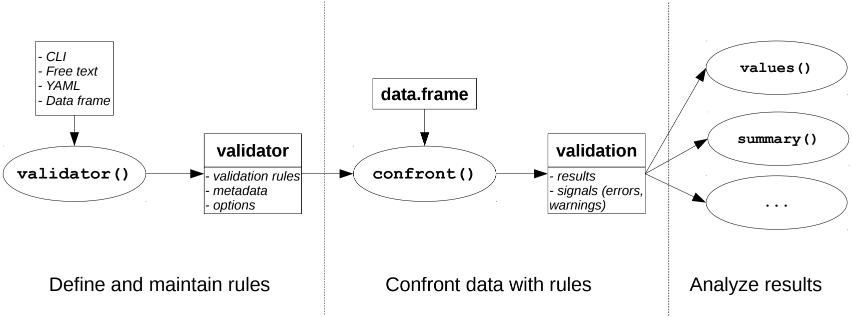
## Because

- you want to clearly **communicate** your data quality
- validation rules have a **life cycle**
  - treat like data (CRUD, analyze)
  - treat like code (version control, review, test)
- they are **Input** for algorithms that improve data quality.

## `validate`

Define, use, analyze, manipulate data validation rules and validation results.

# The `validate` package: basic workflow



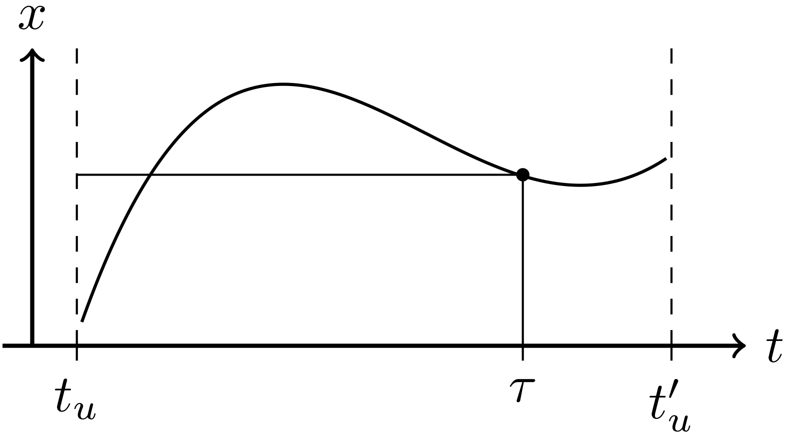
# Rule complexity

# How complex is a validation rule?

**Intuition**

A rule is ‘complex’ if I need different a lot of different information to evaluate it.

# To label a data point



### Intuition

A *data point* is a key-value pair, where the key determines what the value means.

### From the previous picture, a key should at least label

- What population (entity type) we are measuring: $U$
- When did we make the measurement: $\tau$
- Which element of the population (entity) was measured: $u$
- Which variable was measured: $X$

$\rightarrow$ mnemonic: $U\tau uX$

# A measure for rule complexity

**To evaluate my rule, do I need values from *one* or *more***

1. populations (entity types) $U$?
2. measurements $\tau$?
3. population units $u$?
4. variables $X$ ?

- → For each 'yes' denote a $m$ (multiple)
- → For each 'no', denote a $s$ (single)
- The number of $m$'s is the complexity level of your rule.

# Examples

| Rule | labels | level |
|---|---|---|
| $Age >= 0$ | *ssss* | 0 |
| $Turnover - Cost = Profit$ | *sssm* | 1 |
| $Mean(Profit) >= 10$ | *ssms* | 1 |
| $\lvert Mean(Profit/Turnover)_t - Mean(Profit/Turnover)_{t-1}\rvert < 5$ | *smmm* | 3 |

# Not all 4-sequences of *m*'s and *s*'s are possible

| Validation level | | | | |
|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 |
| ssss | sssm | ssmm | smmm | mmmm |
| | ssms | smsm | msmm | |
| | smss | smms | | |

More information: arxiv.org/abs/2012.12028

# Assignment 2

`pdf/assignment2.pdf`