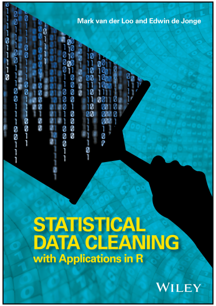

Mark van der Loo and Edwin de Jonge
STATISTICAL DATA CLEANING
with Applications in R
WILEY

Data Validation

Mark van der Loo and Edwin de Jonge

Statistics Netherlands Research & Development
@markvdloo @edwindjonge

useR!2019

Try the code

`03valid/check_validity.R`

Data validation

Verify that data satisfy technical restrictions and does not contradict expert knowledge.

Examples of technical demands

- Number of records must equal 60
- Financial variables are numeric
- Records have a unique `id`
- Zipcode consists of 4 numbers followed by 2 letters

Examples of domain knowledge demands

- turnover is nonnegative
- turnover − costs = profit
- profit not larger then 60% of turnover
- average profit is larger than 0
- average profit differs less than 10% from last year's average

Data validation rules

A domain specific language to express demands.

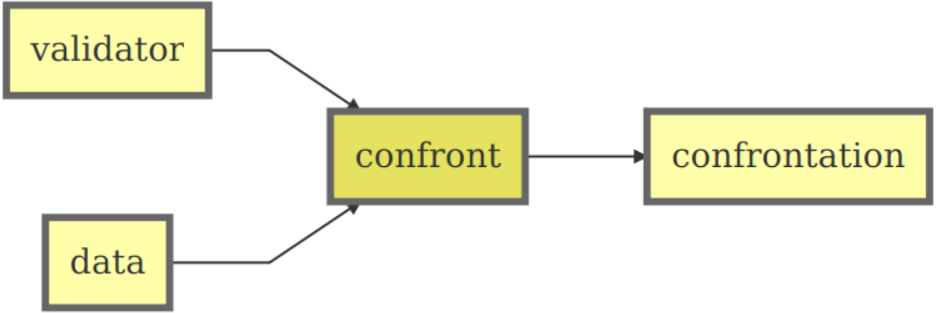
Why?

- Communicate data quality without ambiguities
- Make knowledge explicit and organize it
- Create custom data quality reports
- Reuse ruleset for data cleaning purposes

How?

```r
library(validate)
companies <- read.csv("02input/input.csv",stringsAsFactors = FALSE)
rules     <- validator(.file="02input/rules.R")
result    <- confront(companies, rules)
```

Core concepts of the `validate` package



Comparing numbers

```
rules <- validator(turnover + costs == profit)
summary( confront(companies, rules) )
```

```
##   name items passes fails nNA error warning
## 1   V1     0      0     0   0  TRUE   FALSE
##                                expression
## 1 abs(turnover + costs - profit) < 1e-08
```

Data validation: informal definitions

Data validation

Check if a value, or combination of values is in a certain set of valid values or valid value combinations.

Data validation language in `validate`

Any R expression that results in a `logical`.

Expressions that are validation rules

Basic syntax

- Any type check: `is.numeric`, `is.character`,...
- Any comparison: `<`, `<=`, `==`, `identical` `!=`, `%in%`, `>=`, `>`
- Logical operators `|`, `&`, `if`, `!`, `all`, `any`
- Pattern match: `grepl`

Sugar

Dot "`.`" stands for the whole data set:

```
nrow(.) >= 50        # at least 40 rows
"id" %in% names(.)   # 'id' must be present
```

More, see `?syntax` or `vignette("introduction", package="validate")`

Challenges

1. Express the following restrictions on `companies`. Then `confront` and `summary`
   - Profit does not exceed 60% of turnover
   - turnover less costs equals profit
   - Average profit is larger than zero
   - correlation (`corr`) between total cost and staff exceeds 0.5
   - zipcode is 4 numbers followed by two upper case letters (you need to know regex)
2. Read the rules in `rules.R`. Then, `confront`, and `summary`.