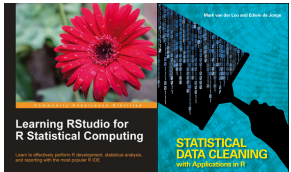
Data Cleaning with applications in R

Mark van der Loo, Statistics Netherlands

Complutense University of Madrid, Spring 2019

$> whoami

- PhD Molecular Physics (2008)
- In Official Statistics (methodology) since 2007
- Research interests:
  - Statistical Computing (Mostly R and C)
  - Data cleaning
  - Network analyses (just started last year)
- Author, with Edwin de Jonge



R Packages

`dcmodify`
`deductive`
`extremevalues`
`errorlocate`
`gower`
`hashr`
`lintools`
`lumberjack`
`rspa`
`simputation`
`stringdist`
`tinytest`
`validate`
`validatetools`

Contents

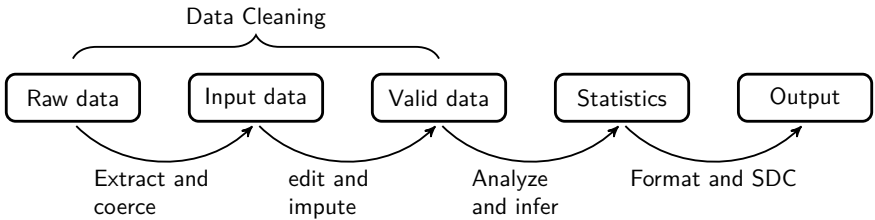
Statistical Value Chain

Value Chains

Porter's value chain (1985)

The idea of the value chain is based on the process view of organizations, the idea of seeing a manufacturing (or service) organization as a system, made up of subsystems each with inputs, transformation processes and outputs.

Statistical Value Chain



Notes

- This part only pertains to the data processing stage. Collection, design, dissemination is not included.
- The fixed points are well-defined statistical products.

Stages in the SVC

1. **Raw Data** is data as it arrives
   - Can differ in quality/source: survey/admin/big data
2. **Input data** satisfies technical demands:
   - File type is known and can be read
   - Variables are of correct type (number/date/text/categorical...)
   - Records identified with statistical objects
   - Variables identified with statistical properties

Stages in the SVC

3. **Valid data** satisfies domain knowledge constraints
   - Age cannot be negative
   - Someone under 15 yrs old cannot have income from work
   - mean economic growth/decline does not exceed 5% in a certain sector
   - ...

Justification

Invalid data leads to invalid statistical results.

Stages in the SVC

4. **Statistics** are the target output values (aggregates) describing the population characteristic of interest.
   - Economic growth
   - Unemployment
   - Income distribution
   - GDP
   - ...

Note

Statistics also need to satisfy domain knowledge constraints.

Stages in the SVC

5. **Output** are statistics, formatted and annotated for publication
   - Figures, tables
   - Definitions
   - ..

The SVC: Remarks

- Actual data processing is not necessarily linear across the chain
- In production architectures a more flexible model is often used where the definition of interfaces between processing steps play a crucial role. The chain shown here is a general example covering most steps in some way.
- The general idea scales really well.