# Management of Scientific Data

ChemBioSys

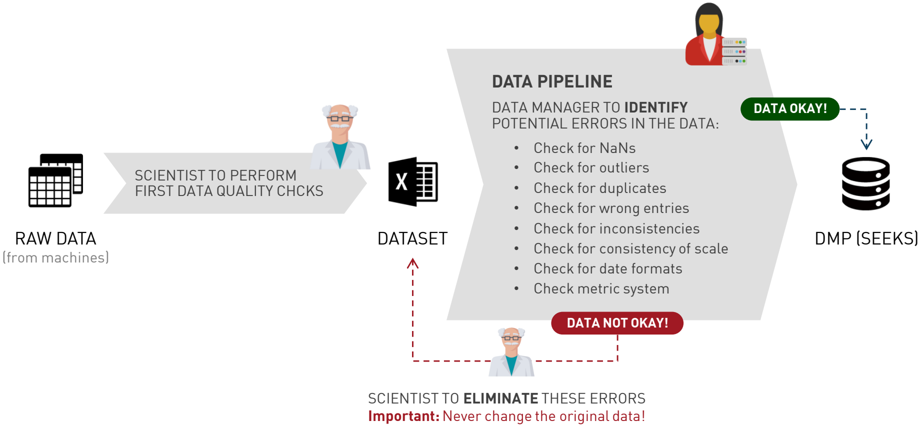
Matthias Bruhns, Tobias Giesemann, Lucas Schneider, Tabitha Uphoff

13.05.2019

Friedrich-Schiller-Universität Jena

# General principles

# General principles



RAW DATA

(from machines)

SCIENTIST TO PERFORM
FIRST DATA QUALITY CHCKS

DATASET

**DATA PIPELINE**

DATA MANAGER TO **IDENTIFY**
POTENTIAL ERRORS IN THE DATA:

- Check for NaNs
- Check for outliers
- Check for duplicates
- Check for wrong entries
- Check for inconsistencies
- Check for consistency of scale
- Check for date formats
- Check metric system

**DATA OKAY!**

DMP (SEEKS)

**DATA NOT OKAY!**

SCIENTIST TO **ELIMINATE** THESE ERRORS

**Important:** Never change the original data!

# Inconsistency, wrong entries and duplicates

# Inconsistency, wrong entries and duplicates

## Data Quality

Inconsistencies

Formats: Time Stamps, Date, Numericals, …

Scales: (kg, g) (m, mm)

System: (metric, imperial)

Wrong Entries:

Impossible Values -> set to NULL for analysis

Outliers -> exclude from analysis

NEVER overwrite original set

Duplicates:

In DBS: Primary Keys

Consistent IDs should be used

# Inconsistency, wrong entries and duplicates

## Data Quality @ ChemBioSys

Inconsistencies

Formats: Machine-created Data

Scales: Machine-created Data e.g. NMR-Machine, Mass-Spectrometry, …

System: metric

Wrong Entries:

Impossible values: Machine Errors

Outlier Detection:

Checkup Routine done by individual Scientists

Duplicates:

Data Management Platform (SEEK)

ID-consistency in each spreadsheet

Overlooked by each individual scientist

# NaNs and Outliers

# NaNs and Outliers

## NaNs

- NaNs can appear when the machine could not measure a value
- Add comments in Meta Data file
    - text based: not computer-readable
    - might lead to misunderstandings
- Replace NaNs with zero
    - might lead to wrong analysis
- Delete NaNs from data file
    - if more than 20% of the data are NaNs, data should not be used or pulished
- Indicate “% of missed values” in meta data file

# NaNs and Outliers

## Outliers

- Outliers can appear when irregularities during the experiment happen
- Use statistical tests to identify “real” outliers → expensive
- Exclude entries above or below a certain threshold

# Usage of machines/computers

# Usage of machines/computers

## Pro

- Most machines provide meta data file for each experiment
- Machines can perform first data quality check internally
- Errors can only be caused by external factors (assuming good software)

## Con

- Proprietary software might become a problem
  - File formats: Updates can render old formats useless
  - Bugs: No way to check the quality of the internal software

→ Some formats might be convertable using scripts

Questions?