Sharing data pipelines

Why sharing data may not be enough, and what to do about it

David Käthner
Deutsches Zentrum für Luft- und Raumfahrt e.V.

07.12.2020

Three claims about data pipelines

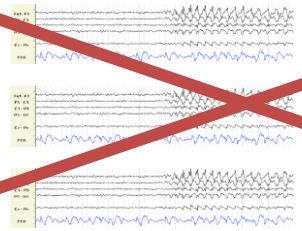
1. Providing only primary data is often insufficient to make results of an analysis sufficiently transparent.
2. Data pipelines and data analysis are inseparably linked.
3. The idea of data pipelines is easy to understand, but can be difficult to implement.

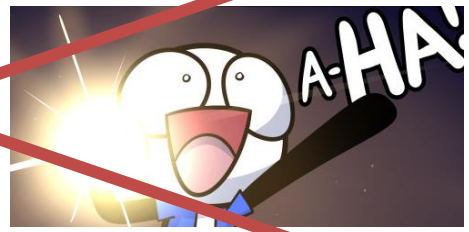
Data Pipelines have a direct effect on reproducibility

Research question (actual study):

Is a certain pattern in EEG-profiles indicative of a diagnosis from the autism spectrum disorder (ASD)?

design study record data analyse data

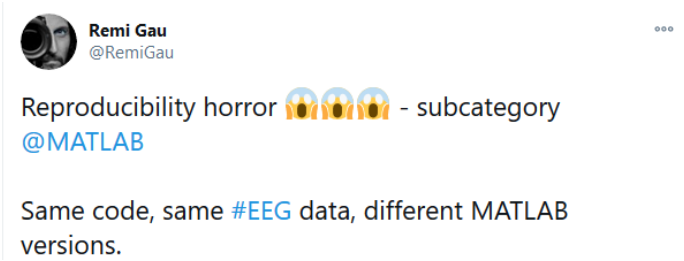




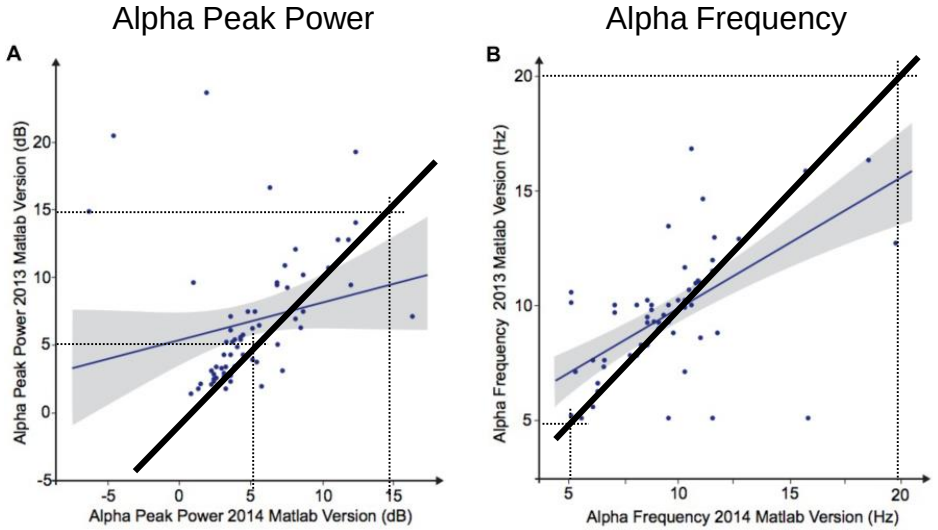
loadingartist.com

Answer:

Not sure, because very different outcomes using the same data. Reason for that: Different data pipelines involving different software.



https://twitter.com/RemiGau/status/1334936565638979592



Lefebvre et al. (2018). Alpha Waves as a Neuromarker of Autism Spectrum Disorder: The Challenge of Reproducibility and Heterogeneity. *Frontiers in Neuroscience, 12,* 33.

Why worry about pipelines?

Bugs

Pandas reads values incorrectly from csv

link

Features

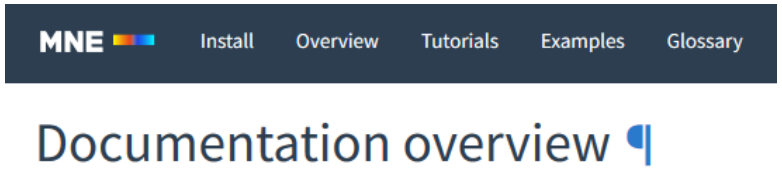
link

Parameter settings

```
# Filtering
data.filter(phase='zero',
            fir_window='hamming',
            fir_design='firwin',
            l_freq=None, h_freq=48,
            l_trans_bandwidth='auto',
            h_trans_bandwidth='auto',
            filter_length='auto')
```

```
19    import mne

62            data = mne.io.read_raw_edf(op.join(inputFolder, f_name),
```



link

You have always been creating data pipelines

Research question:

Is a certain pattern in EEG-profiles indicative of a diagnosis from the autism spectrum disorder (ASD)?

source 1: socio-demographic

ID	values
02AN	...
95BC	...
20NQ	...
...	...

source 2: diagnosis

ID	values
02AN	...
95BC	...
20NQ	...
...	...

source 3: EEG

ID	time	values (raw)
02AN	1	...
	2	...
	3	...
	...	...

↓

ID	*time*	values (aggregated)
02AN		...

join

analysis and visualization format

ID	socdem	diagnosis	EEG
02AN	...	...	...
95BC	...	...	...
20NQ	...	...	...
...	...	...	...

1NF-normalization, also known as Tidy Data

observation	diagnosis	EEG	sociodemographic	ID
1	...	...	...	02AN
2	...	...	...	95BC
3	...	...	...	20NQ
...	...	...	...	...

regression y = x0 + x1 + x2

each row = one observation

- uniquely identifiable through one or more keys
- observation ≠ point in time

each column = one variable

- no further decomposition possible
- is of one of these types:
  - identification (key)
  - factor
  - measurement
  - factor from measurement

normalization: link

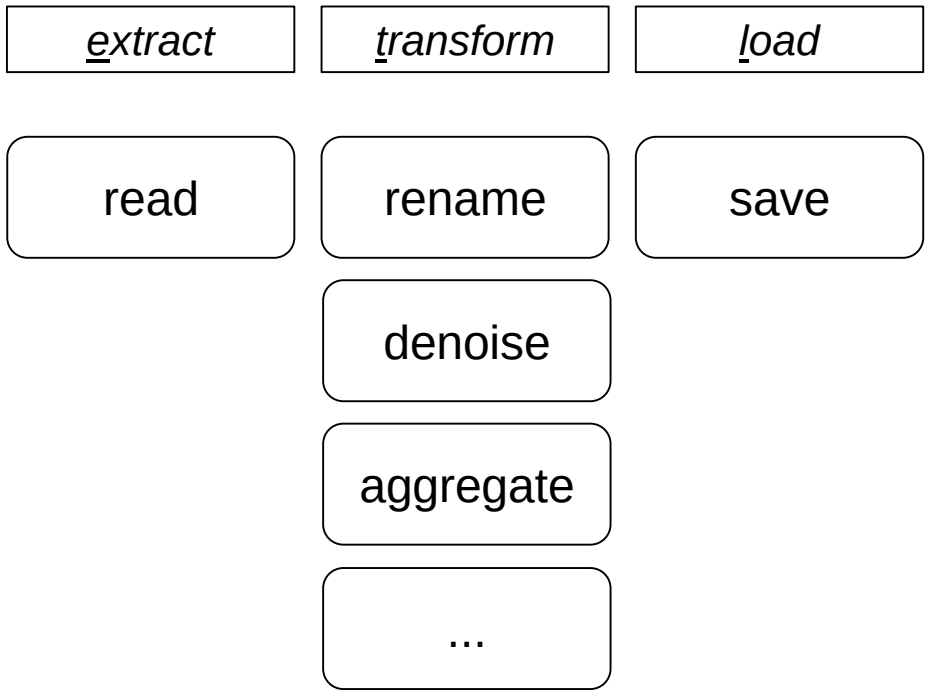
Tidy Data: Wickham, H. (2014)

Functions are the core concept of a working pipeline

ID	source
1	source 1
2	source 1
3	source 1
...	...

ID	source
1	source 2
2	source 2
3	source 2
...	...

ID	source
1	source 3
2	source 3
3	source 3
...	...



application of functions to data

```
for data in source:
  for participant in ID :
      read_data()  |>
      function_1()  |>
      function_2()  |>
      save_data()
```

pipe operator:
output function 1 = input function 2

R and Tidyverse: Guiding you in the right direction

- „opinionated collection of R packages“
- addresses every important aspect of data preparation and analysis specifically
- accessible tutorials and documentation
- create literal data pipelines

link

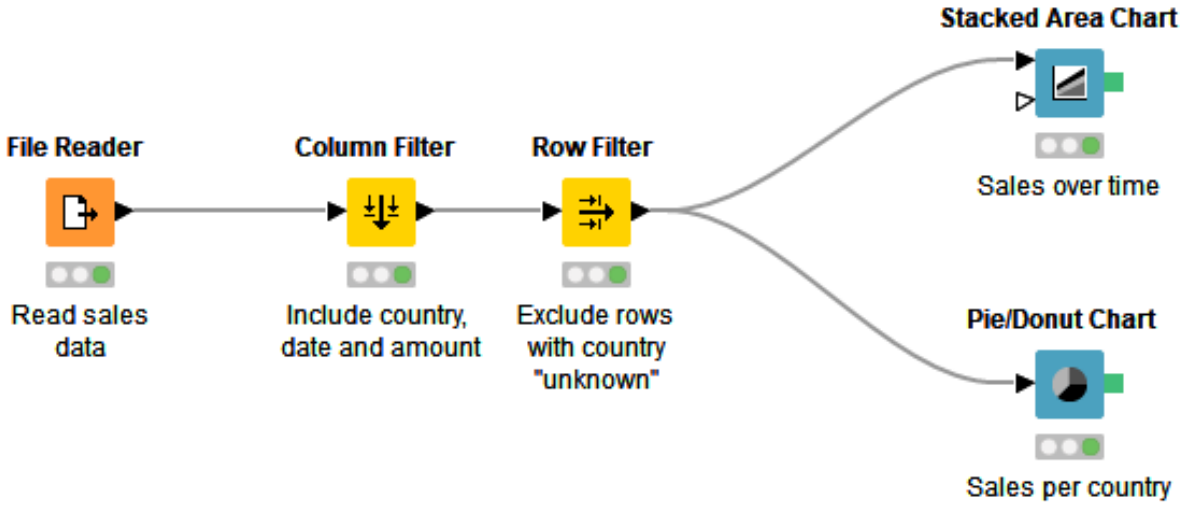
	country	beer_servings	spirit_servings	wine_servings	total_litres_of_pure_a…
1	Afghanist…	0	0	0	0
2	Albania	89	132	54	4.9
3	Algeria	25	0	14	0.7

```
drinks_smaller <- drinks %>%
  filter(country %in% c("USA", "China", "Italy", "Saudi Arabia")) %>%
  select(-total_litres_of_pure_alcohol) %>%
  rename(beer = beer_servings, spirit = spirit_servings, wine = wine_servings)
```

	country	beer	spirit	wine
1	China	79	192	8
2	Italy	85	42	237
3	Saudi Arabia	0	5	0
4	USA	249	158	84

link

KNIME (Konstanz Information Miner): Taking pipeline literally



read | transform | visualise

- each node is a processing step
- nodes can be configured via GUI or call a script (R, Python, Matlab)

personal opinion:

- usage requires programming skills
- really good with pipelines that are much re-used
- visual approach works for simpler workflows, complex tasks still need to be scripted
- lots of machine learning for finance and life sciences

Sharing *reproducible* pipelines

Pipeline must be deterministic

- same input = same output
  - most likely problem in EEG-parameter extraction: dependency issue (e.g. changed defaults)
- 2 conditions
  - ensure actually same input is used
  - solve dependencies

How to

1. make sure the right data is used
   - provide a hash value for your data like so:

```
kaet_da@TS-010435 MINGW64 ~/projects/
$ tar -cf - raw | md5sum
9d8a7cdc804237438252af02374065d4 *-
```

2. reserve the dependencies
   - R and Packrat: provide project containing all dependencies as local copies

```r
# Initializing Packrat
packrat::init()

# tidyverse
library(dplyr)
library(ggplot2)

demo <- mtcars %>%
  mutate(type = row.names(.)) %>%
  select(type, everything()) %>%
  filter(mpg > 16) %>%
  glimpse

# Export the whole thing into a single tar archive
packrat::bundle(file = 'export.tar.gz', overwrite = TRUE)
```

link

more options

- R: checkpoint
- Python: virtual environments
- everything: Docker container

link

Tips for a shareable data pipeline

practicalities

- isolate functions in separate file
- load functions into cleaning file, notebook, ...
- ship files to others via Github, servers of your organization, Email, ...

improve usability for others

- can be executed without (external) explanation
  - README is a good starting point
- group steps of cleaning/analysis in coherent units that communicate intent
- document settings in an obvious way (e.g. config-files)

make your own life easier

- design the pipeline before implementation
  - better even before data recording
  - best to already know planned analysis
- adopt a functional programming style (40-line rule-of-thumb)
- be very careful with interactive programming (best to avoid it for cleaning)

SPSS

- try to do as much as possible in SPSS, and export the syntax file

Three claims about data pipelines revisited

1. Providing only primary data is often insufficient to make results of an analysis sufficiently transparent.
   - bugs, features, parameter settings, user error, missing or changing dependencies
2. Data pipelines and data analysis are inseparably linked.
   - basis for many types of analysis, visualisations, ... are tidy data
3. The idea of data pipelines is easy to understand, but can be difficult to implement.
   - Good: Effective and accessible concepts, tools, and tutorials exist.
   - Bad: Complex pipelines still require substantial programming and experience.
   - Solution: Teach basic concepts and their application as part of regular curriculum.

Thank you for your attention!

References

Lefebvre, A., Delorme, R., Delanoë, C., Amsellem, F., Beggiato, A., Germanaud, D., et al. (2018). Alpha Waves as a Neuromarker of Autism Spectrum Disorder: The Challenge of Reproducibility and Heterogeneity. *Frontiers in Neuroscience, 12,* 33.

Wickham, H. (2014). Tidy Data. *Journal Of Statistical Software, 59*(10).

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., et al. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686.

Websources

Database normalization
https://de.wikipedia.org/wiki/Normalisierung_(Datenbank)

Docker
https://www.docker.com/resources/what-container
https://www.docker.com/blog/containerized-python-development-part-1/

Packrat
https://rstudio.github.io/packrat/
https://github.com/c06n/Packrat_HowTo

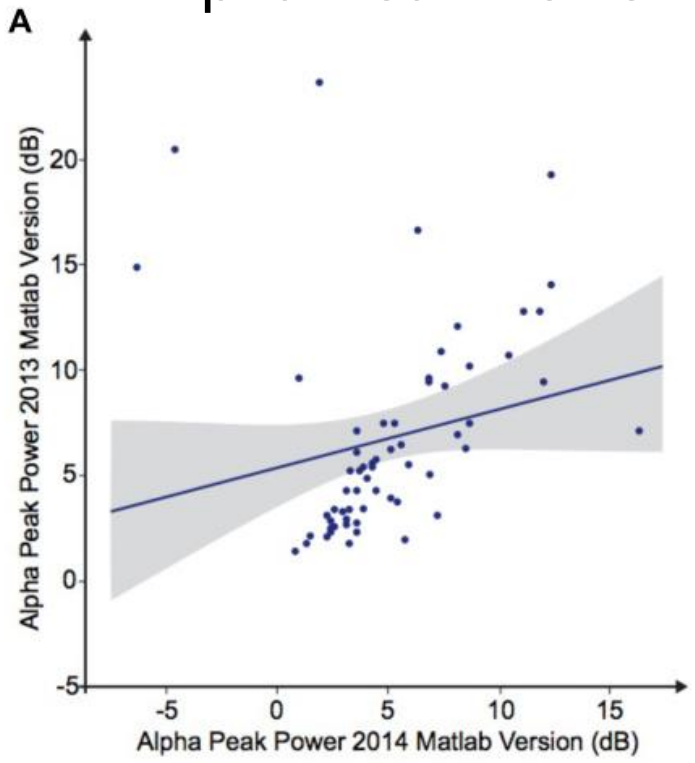
Pandas issue
https://github.com/pandas-dev/pandas/issues/34120
https://github.com/c06n/Pandas_readcsv_issue/blob/master/demonstration.ipynb

Tidyverse
https://moderndive.com/index.html
https://r4ds.had.co.nz/

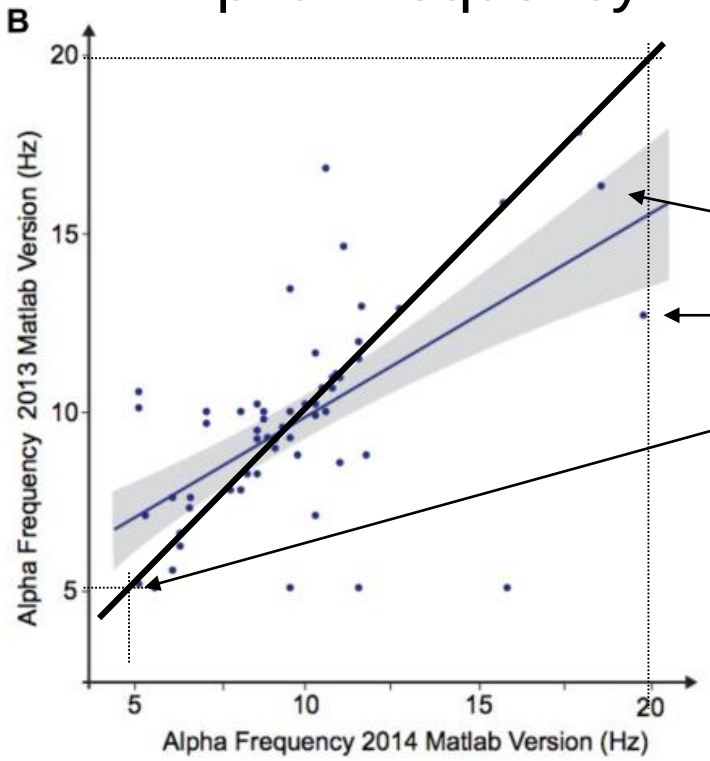
KNIME
https://www.knime.com/learning

Different pipelines, different Alpha peak and Alpha Frequencies

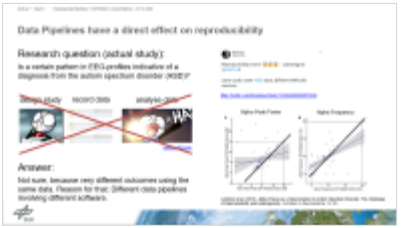
Alpha Peak Power



Alpha Frequency



Matlab 2013	Matlab 2014
16	18
13	20
5	5



Lefebvre, A., Delorme, R., Delanoë, C., Amsellem, F., Beggiato, A., Germanaud, D., Bourgeron, T., Toro, R, and Dumas, G. (2018). Alpha Waves as a Neuromarker of Autism Spectrum Disorder: The Challenge of Reproducibility and Heterogeneity. *Frontiers in Neuroscience, 12,* 33.

Data Pipeline of Lefebvre et al. (2018)

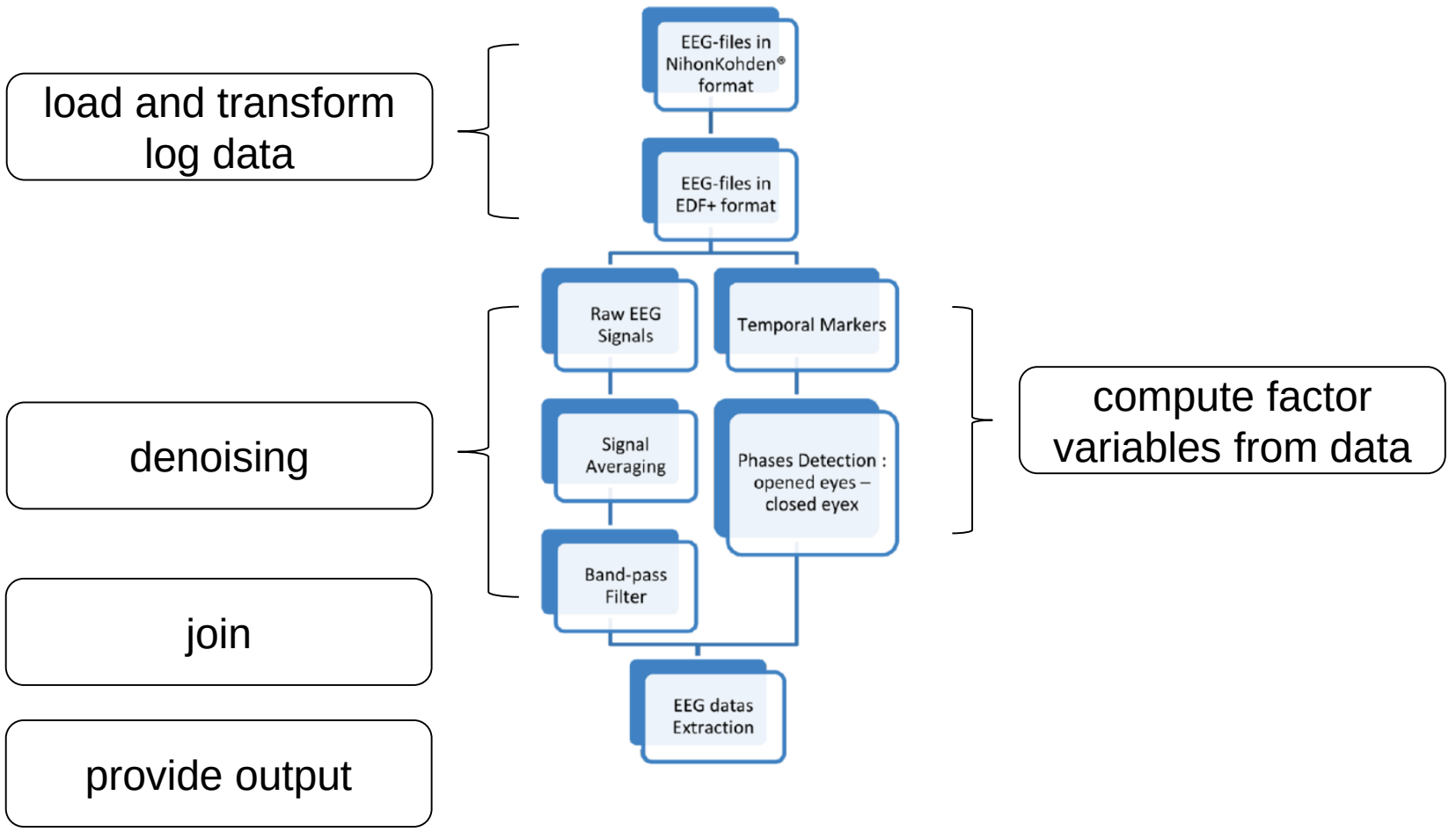


FIGURE S2 | Preprocessing and processing chain of the data.

Big data and data models

target	source	typical frequencies
human	eye tracking	30 / 120 / 500 / 1000 Hz
	ECG	500 / 1000 Hz
context	vehicle data	20 / 25 Hz
	positional data	1 Hz

back-of-the-envelope calculation for logging at 120 Hz:

- 100 variables, float64, 30 min, 50 participants = 8.64 GB

- Big data problem
  - per participant easily 2-5 GB
  - *all* recorded data does not fit in standard working memory anymore
- Ontological problem
  - tidy model is intuitive, but forcing every data set into this model leads to inefficient data structures
  - error prone for ontologically different dimensions

Complex KNIME pipeline