SMART DATA

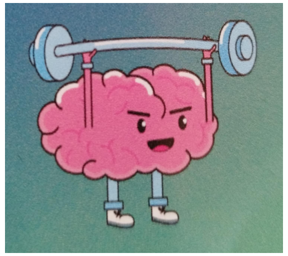
A Meinerzhagen Original

Our background: The Team

- **Moritz Hannemann:** IT Developer, JCNS@MLZ, Garchingen - Cloud Computing, DAPHNE Computing Infrastructure
- **Kai Zhou:** Frankfurt Institute for Advanced Studies, Focus in Theoretical Physics, Heavy Ion Physicist
- **Cyrus Walther:** Machine Learning Application in Astroparticle Physics, International relations
- **Sebastian Wozniewski:** Grid-Computing for CMS & ATLAS, Data Analyst with Machine Learning Focus

Smart data are data sets that have been extracted from larger data volumes (cf. Big Data) by means of algorithms according to certain structures and obtain meaningful information. This data has already been collected, sorted and analyzed beforehand and prepared for the end user.

~

A Definition: What is  - Data?

- Is Smart Data just data reduced down to suit a certain purpose?
- Can Smart Data display a path to approach information preservation?
- Does Smart Data imply a new data format or rather smart handling of the data?

Multiple Aspects - Primary Usage

Purpose of Smart Data:

- Efficient storage in case of Big Data
- Preparation for further efficient usage, i.e. “transformation into smart data” (Examples: Filtering, accessible & understandable format)

Methods:

- For reduction:
  - Define rules based on expert knowledge (what would possibly be needed, what is irrelevant)
  - Implement automated data transformation (collaborations handling big data provide already good examples)
- For transformation:
  - Satisfy needs of current analysts (“easily” identifiable)

Multiple Aspects - Data Preservation

Purpose:

- Reproduce analysis, check for mistakes:
  - Is the full data set needed?
    - Reduce not only attributes but also the number of samples since only some will be necessary for that (“Life cycle Approach”)
    - After some time keep only a zero-bias subset?
- Usage of data for new research goals:
  - Complex to identify needs of scientists in multiple decades
  - How to find and retrieve data efficiently after a long time?
    - Even technical language may change. One idea in progress to overcome this: Ontology of keywords describing data sets.

Content:

- Extended meta-data (or links between related data, e.g detector information)
  - guaranteeing usability even after decades (“self-explanatory data”)