Homework guide

Peter Antal
antal@mit.bme.hu

Academic calendar

- Scientific student contest (no classes) 14 November (Wednesday)
- Open day for secondary schools (no classes) 30 November (Friday)
- Homework presentation (Last day of classes): 7 December (Friday)
- Repeat week (resits and late submission of home assignments): 10-14 December (Monday-Friday)
- Exams begin: 17 December (Monday)
- Duration of examination period 21 working days
- Last day of examination period: 22 January (Tuesday)

Requirements

- Grading:
  - Homework, obligatory, min.40%, weight: 50%
  - Final exam, min.40%, weight: 50%
  - Overall
    - 40<: satisfactory
    - 50<: fair
    - 65<: good
    - 80<: excellent
- Final exam is a closed-book exam.

Topics

- Intelligence in the data analysis process.
- Intelligent (complex) models in data analysis.
- Optimization, Bayesian model averaging and sensitivity analysis using resampling methods in data analysis.
- Semantic data repositories, data visualization, dimensionality reduction, data engineering/transformations using ontologies, data cleaning and imputation.
- Unsupervised learning.
- Supervised learning: decision trees, regression, kernel methods, multilayer perceptron, deep neural networks.
- Probabilistic graphical models: Markov Random Fields, Bayesian networks.
- Reinforcement, active, budgeted and online learning.

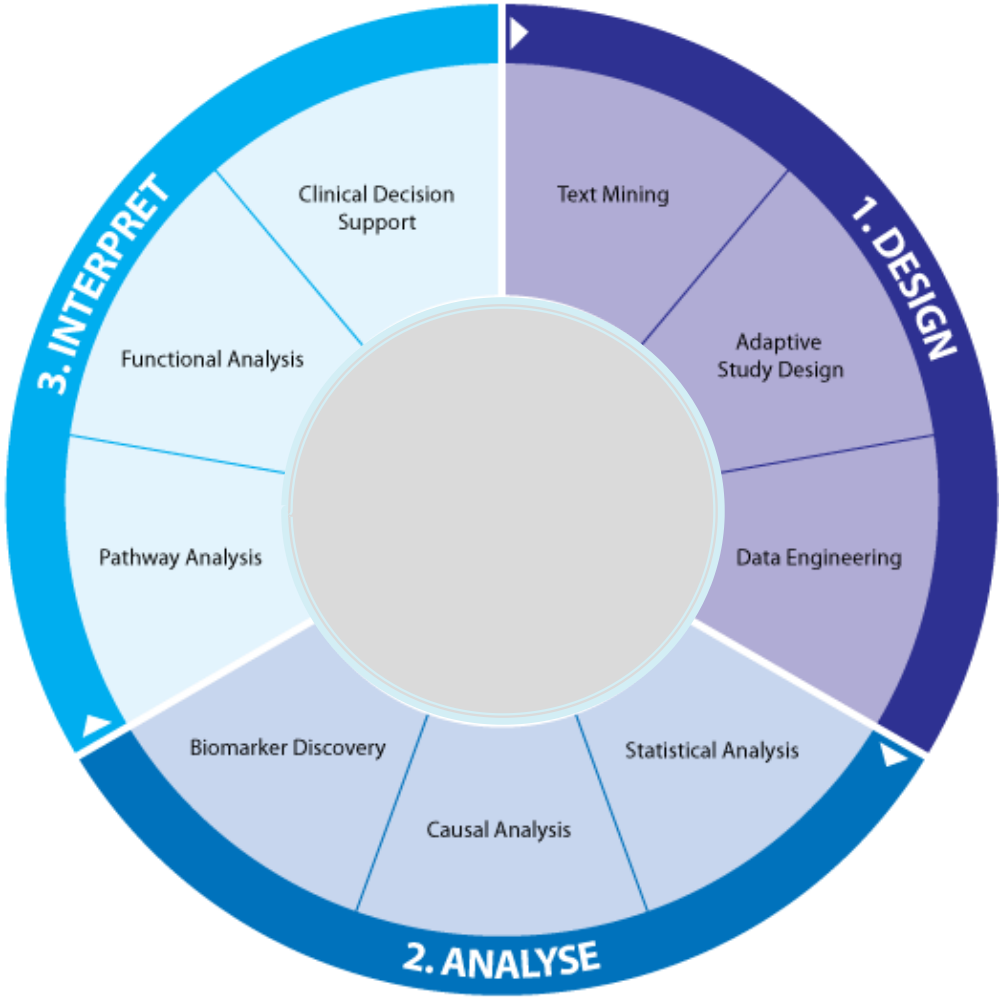
Goals of inductive inference

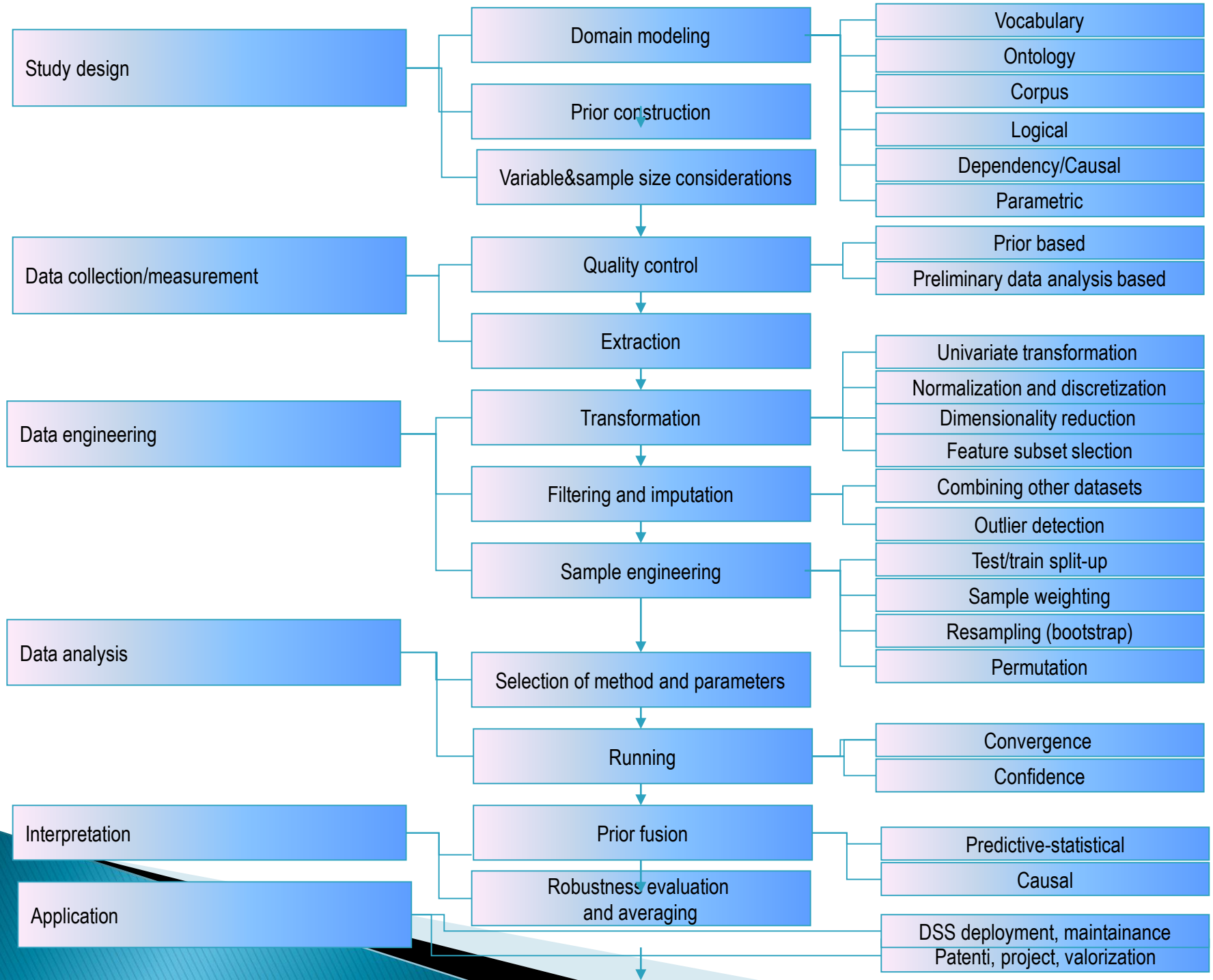
- Predictive: P(Future|Past)
  - (Passive, observational) inference
    - P(Query|Observations, Observational data)
  - Interventionist inference
    - P(Query|Observations, Interventions)
  - Counterfactual inference
    - P(Query| Observations, Interventions, Counterfactual conditionals)

- „Parametric”: ~about model(s): P(Model|Data)
  - Model structure
  - Parameters
  - Model properties
  - **Ultimate goal of inference about models?**

The complex process of IDA

- Text mining
- Study design
- Data engineering
- Analysis
- Interpretation
- Application




Study design
Domain modeling
Prior construction
Variable&sample size considerations
Vocabulary
Ontology
Corpus
Logical
Dependency/Causal
Parametric
Data collection/measurement
Quality control
Extraction
Prior based
Preliminary data analysis based
Data engineering
Transformation
Filtering and imputation
Sample engineering
Univariate transformation
Normalization and discretization
Dimensionality reduction
Feature subset slection
Combining other datasets
Outlier detection
Test/train split-up
Sample weighting
Resampling (bootstrap)
Permutation
Data analysis
Selection of method and parameters
Running
Convergence
Confidence
Interpretation
Prior fusion
Robustness evaluation and averaging
Predictive-statistical
Causal
Application
DSS deployment, maintainance
Patenti, project, valorization

Tools for IDA

- Tools
  - Visual data analytics: Mondrian,..
  - Data analysis workflows: Knime, Taverna, Kepler, PipelineP
  - Data mining - machine learnig(ML): WEKA, RapidMiner
  - General statistical: SPSS, SAS
  - Languages for stat/ML: STAN, EDWARD
  - General languages with stat/ML/DM packages: R, Python

Homework

- Study design: selection of a domain
- Data engineering: selection of dataset from UCI
  - http://archive.ics.uci.edu/ml/index.php
- Selection of language or tool
  - Python, scikit: https://scikit-learn.org/stable/index.html#
  - RapidMiner: https://rapidminer.com/get-started/
- Learn
- Prepare report and presentation
  - 5-10 pages
- Send in email
- Present your work!