



Aalto University
School of Science



Department of
Computer Science

Kernel Methods, Pattern Analysis and Computational Metabolomics (KEPACO)

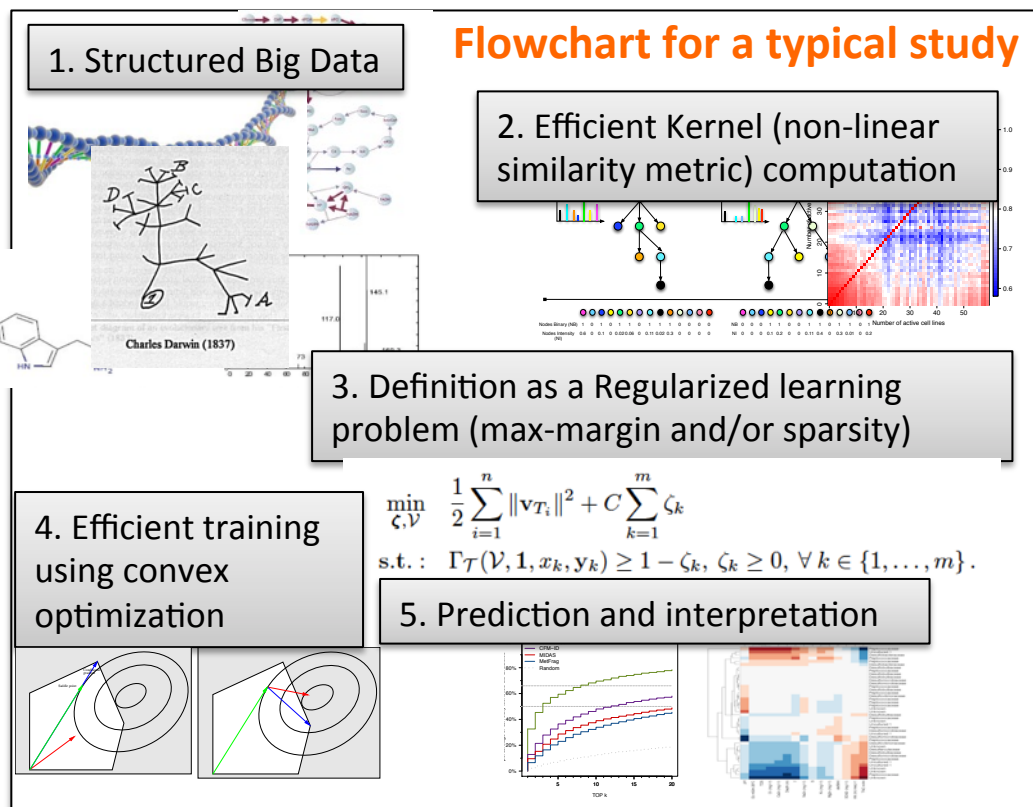
Associate professor Juho Rousu

KEPACO in a Nutshell

We develop kernel-based machine learning methods for **predicting of structured, non-tabular data** arising in biomedicine and digital health, especially applications on small molecules such as drugs and metabolites.

Examples of research questions:

- How to identify a metabolite from its mass spectrum?
- How to find multivariate associations between SNPs and metabolites?
- How to classify an enzyme according to its metabolomic role?



KEPACO Mission: Predicting Structured data

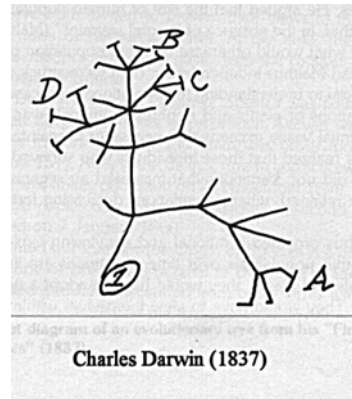
Many data have **logical structure**, how can we make use of that in making models **faster** and **more accurate**?

In particular, we are interested of predicting structured **output**

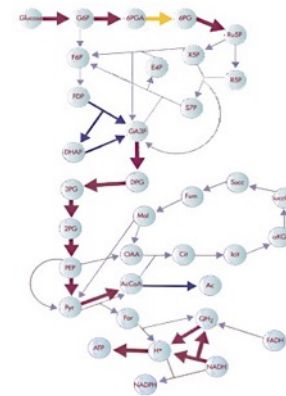
Sequences



Hierarchies



Networks



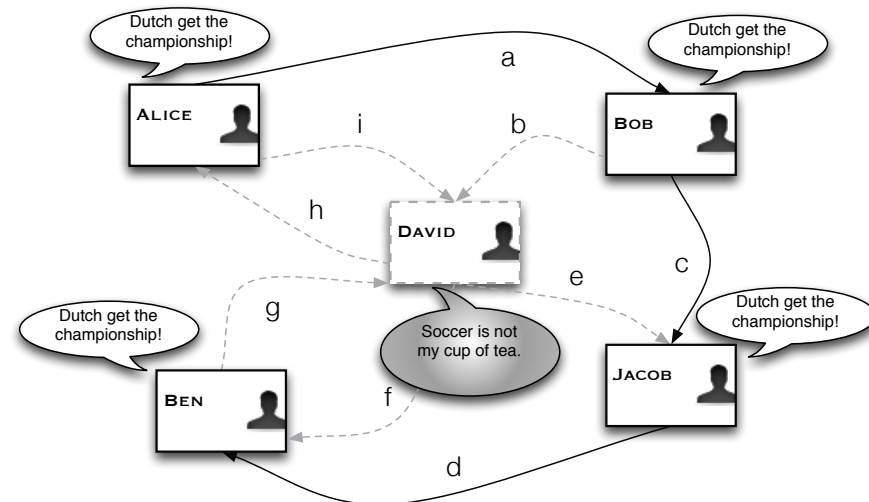
Example: Predicting Network Response

Problem:

- To predict the spread of activation in a complex network, given an action/stimulus

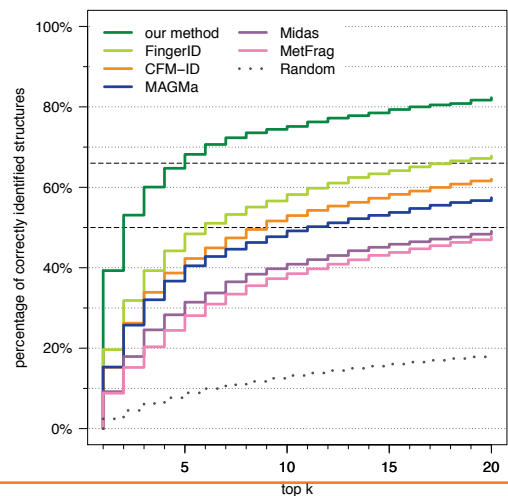
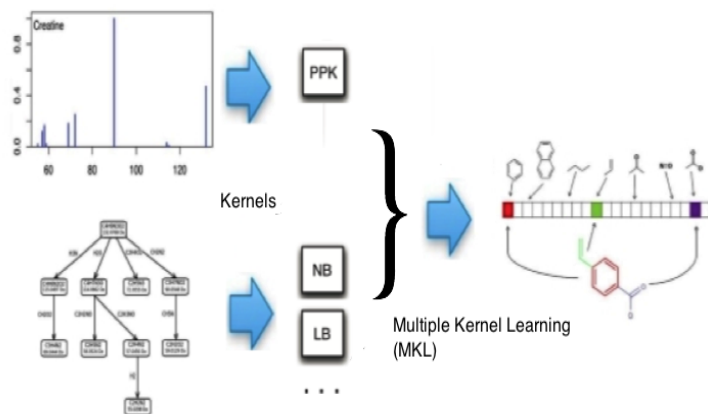
Example:

- Input: post in a social network
- Output: subnetwork of friends who share the post



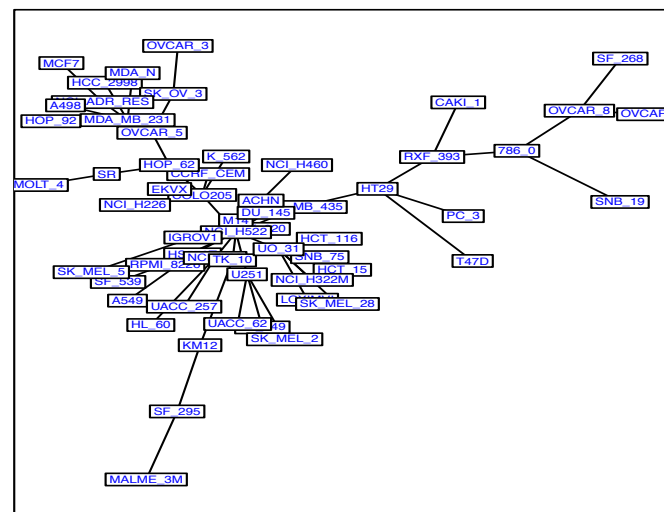
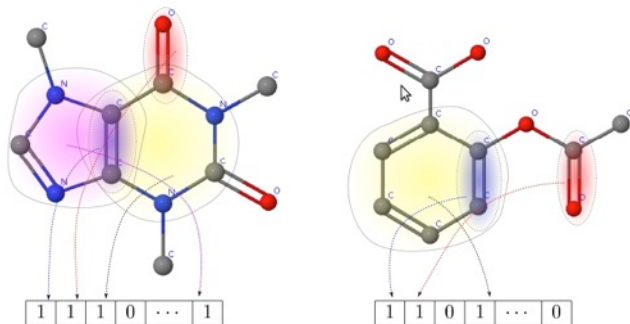
Example: Metabolite identification through machine learning

- From a set of MS/MS spectra (x = structured input), learn a model to predict molecular fingerprints (y = multilabel output)
- With predicted fingerprints retrieve candidate molecules from a large molecular database
- Our CSI:Fingerid (<http://www.csi-fingerid.org>) tool is the most accurate metabolite identification tool to date



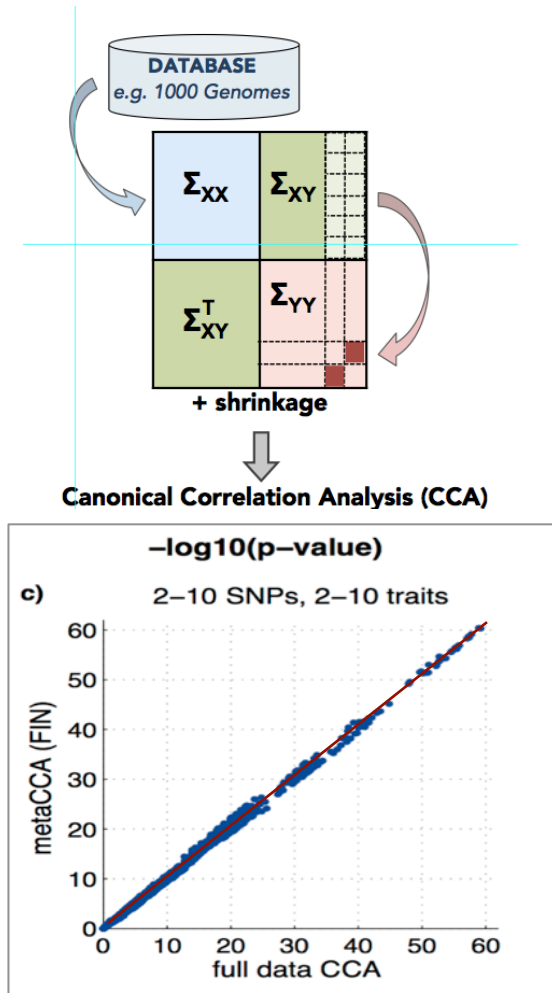
Example: Drug Bioactivity Classification

- Data: x = Molecule, y = labeling of a graph of cell lines
- Compatibility score $F(x,y)$: how well subset of cell lines go together with the drug
- Co-occurring features: molecular fingerprints vs. subsets of cell lines
- Prediction: finding the best scoring labeling of the graph



Example: Multivariate association meta-analysis

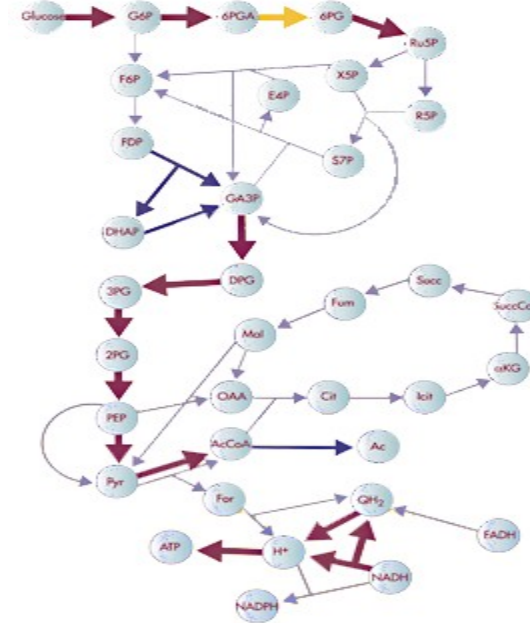
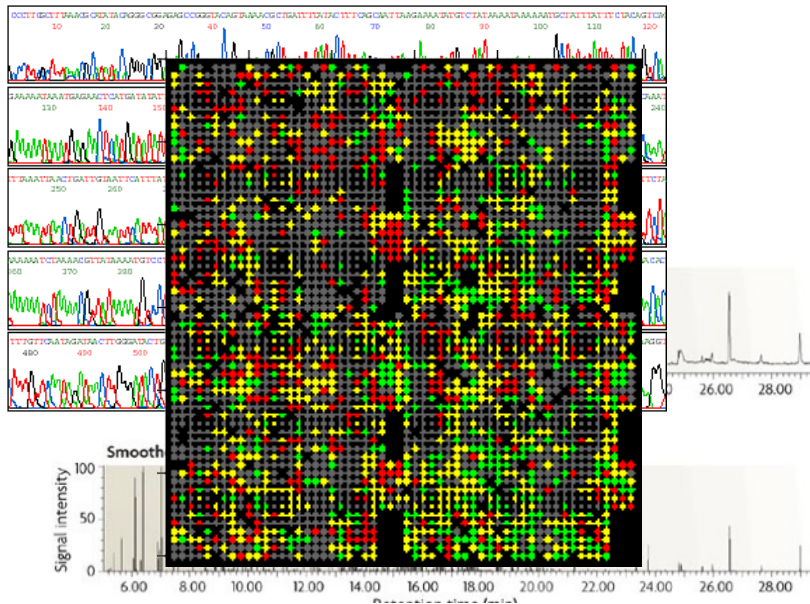
- **Problem: find associations btw. groups of SNPs (=genotype) and groups of metabolites (=phenotype)**
- **Meta-analysis setting**
 - Several studies analyzed together
 - No original data available, only summary statistics of individual studies
- **Our metaCCA tool can recover multivariate associations nearly as accurately as if original data was available**



Example: Biological network reconstruction

Input: genomic measurements

Output: biological network



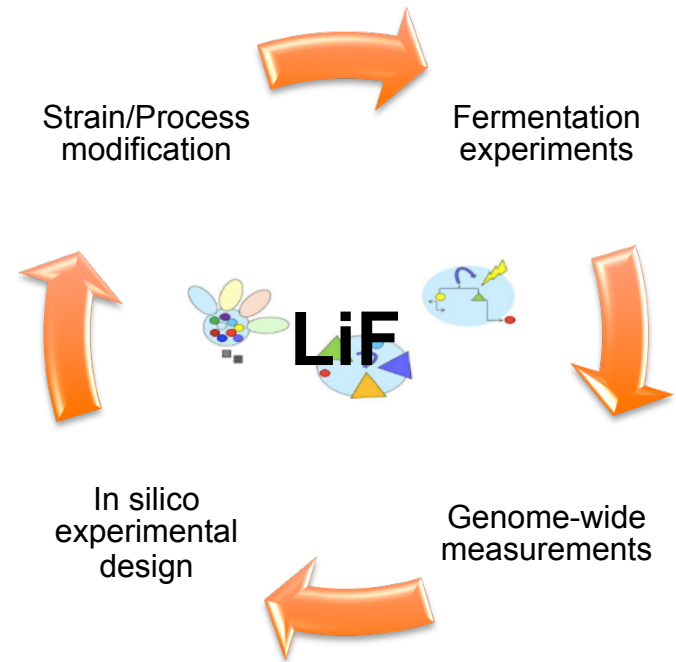
Example: Synthetic biology

Tekes Strategic Opening “LiF – Living Factories”, 2014-

- **Goal: energy and carbon efficient industrial processes through synthetic biology**
- **Consortium: VTT, Aalto, U. Turku**

Our focus

- **In silico design, modelling and optimization of synthetic biosystems**
- **Systems Modelling + Machine Learning**



Kernel Methods, Pattern Analysis & Computational Metabolomics (KEPACO)

KEPACO research group

Dr. Sandor Szedmak

Dr. Sahely Bhadra (w. S.Kaski)

Dr. Markus Heinonen (w. H. Lähdesmäki)

Dr. Celine Brouard

Dr. Elena Czeizler

Dr. Hongyu Su

MSc Huibin Shen

MSc Anna Cichonska (HIIT&FIMM)

MSc Viivi Uurtio

+ students



KEPACO group summer 2015

Funding

- Academy of Finland
- EU FP7
- Tekes



ACADEMY
OF FINLAND



Tekes