ALGODAN Algorithmic Data Analysis

Advances in Neuroinformatics

Aapo Hyvärinen

Aalto University

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Neuroinformatics Team

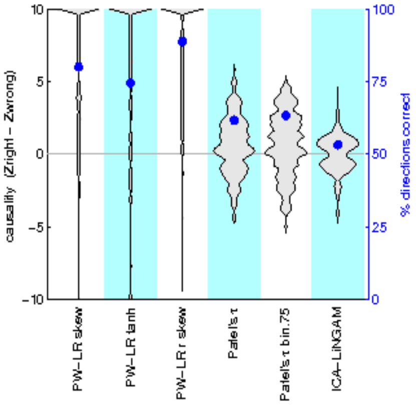
- Mission:
  - Develop statistical data analysis methods, with focus on
    - Unsupervised machine learning methods
    - Neuroscience applications
  - Non-Gaussianity a central theoretical framework

- Members:
  - Aapo Hyvärinen, leader
  - Patrik Hoyer, co-leader (until 8/2013, started own company)
  - 2-4 postdocs, 2-4 PhD students
  - From 2012, partly in CoE of Inverse Problems Research

Highlight 1:
Causal analysis

- Passive observation vs. interventions
  - Completely passively observed data (our LiNGAM from 2006)
  - Experiment with (optimal?) interventions

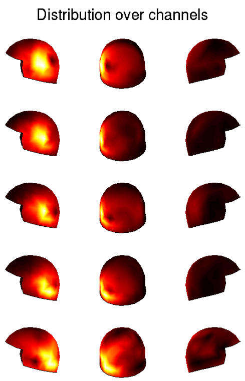
    (Hyttinen, Eberhardt, Hoyer, *JMLR*, 2012, 2013a)
- Causality in fMRI, jointly with Stephen Smith
  - Oxford Centre for Functional Imaging of the Human Brain
  - Developer of simulated data for comparing algorithms
  - Our tailor-made methods (*JMLR*, 2013b)
    - Have best performance on simulated data
    - Are particularly simple variants of LiNGAM

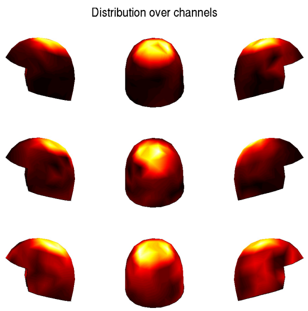


Simulation 1 (5 nodes, 10 minute sessions, TR=3.00s, noise=1.0%, HRFstd=0.5s)

Highlight 2: Testing independent components

- In independent component analysis, testing almost inexistent
  - Components could be local minima, or random effects
- We developed a method which uses a proper null hypothesis and the theory of classical hypothesis testing
  - Do ICA on multiple datasets (e.g. subjects), and see if you get the same component in more than one data set
  - Applications in MEG (*NeuroImage*, 2011):





- Application on fMRI needed further theory (*Frontiers in Human Neuroscience*, 2013)

Highlight 3: Practical brain imaging data analysis

- Decoding brain state from MEG (*NeuroImage*, 2013)
  - Optimal combination of ICA with classification methods
  - Must use nonlinear classification
- Two-person neuroscience: measuring interacting subjects
  - Riitta Hari's ERC AdG for constructing a system of two MEG scanners with video connection
  - Extremely challenging, still ongoing
- Analysing nonstationary dynamics
  - Result of sabbatical at ATR, Japan, in 2013,
    a leading centre in brain imaging

Future

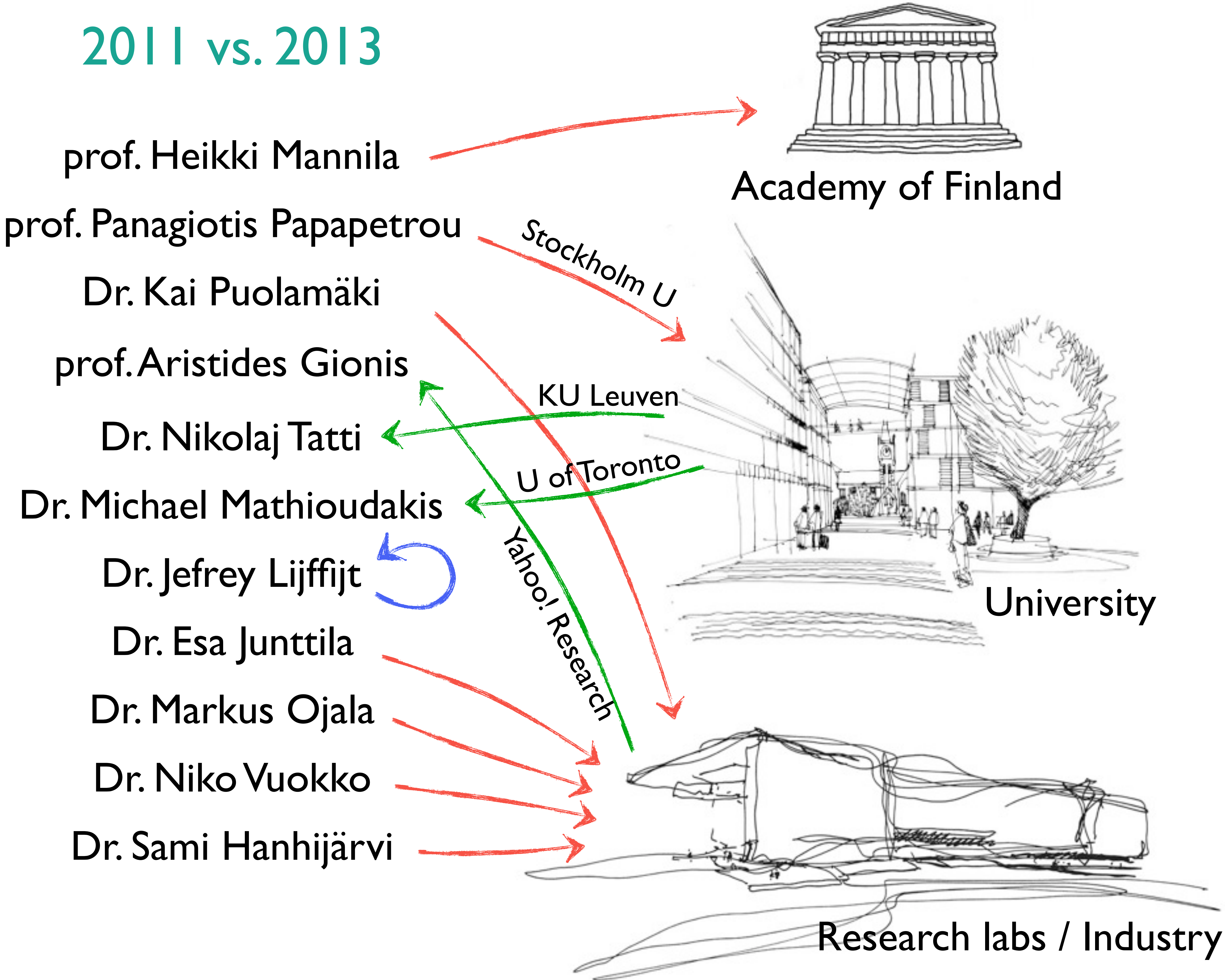
- Co-leader Patrik Hoyer left academia
  - Group size reduced
  - Causal analysis given less emphasis
- New planned project: Modelling spontaneous brain activity
  - Very popular topic in brain imaging
  - But: our approach is to model the computations happening in the brain
    - Theoretical neuroscience instead of brain imaging

Data mining : theory and applications

Aristides Gionis

18 March, 2014

2011 vs. 2013



prof. Heikki Mannila

prof. Panagiotis Papapetrou

Dr. Kai Puolamäki

prof. Aristides Gionis

Dr. Nikolaj Tatti

Dr. Michael Mathioudakis

Dr. Jefrey Lijffijt

Dr. Esa Junttila

Dr. Markus Ojala

Dr. Niko Vuokko

Dr. Sami Hanhijärvi

Academy of Finland

Stockholm U

KU Leuven

U of Toronto

Yahoo! Research

University

Research labs / Industry

research activities

- foundations in pattern discovery
  - statistical significance of patterns
- sequence analysis
  - episodes, segmentation, surprising events
- applications
  - biology, paleontology, linguistics, ...

selected publication venues (2012-2014)

- TODS 2014
- 2 x DMKD 2014
- 3 x DMKD 2013
- 2 x ECML PKDD 2013
- ACM Transactions on Applied Perception 2013
- Proceedings of the Royal Society B 2012
- International Journal of Data Mining and Bioinformatics 2012
- VLDB 2012

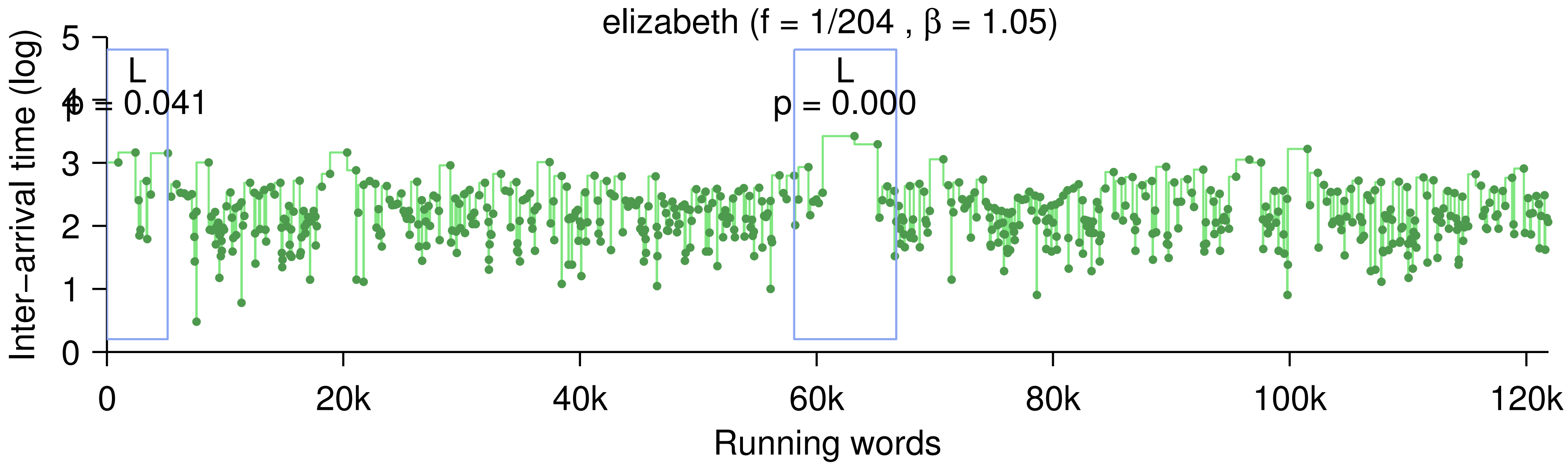
research highlights

comparison and exploration of event sequences

- Jefrey Lijffijt, PhD dissertation, Dec 2013
  - best doctoral dissertation in the Aalto school of Science in 2013
- data: event sequences
  - DNA, texts, sensor readings
- problems:
  - are two data sets equivalent with respect to pattern X?
  - are there parts of the data different from the whole?
  - which set of granularities to use when looking for patterns?

are there parts of the data that are different?

- multiple testing



- challenge: provide accurate correction without randomization/ simulation

- computational question: given a Bernoulli process that runs for n steps, what is the probability that in any subsequence of length m, there are k or more events?

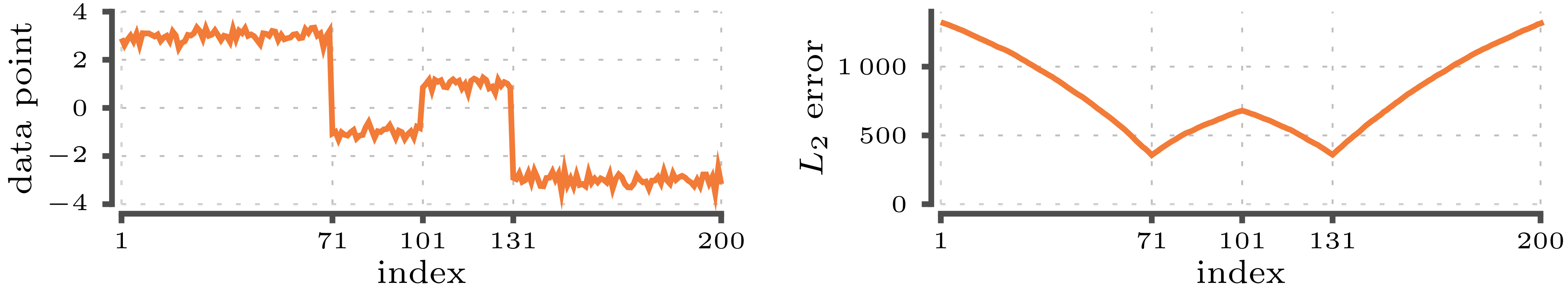
- thesis introduces upper-bound that works well in practice

finding informative window lengths

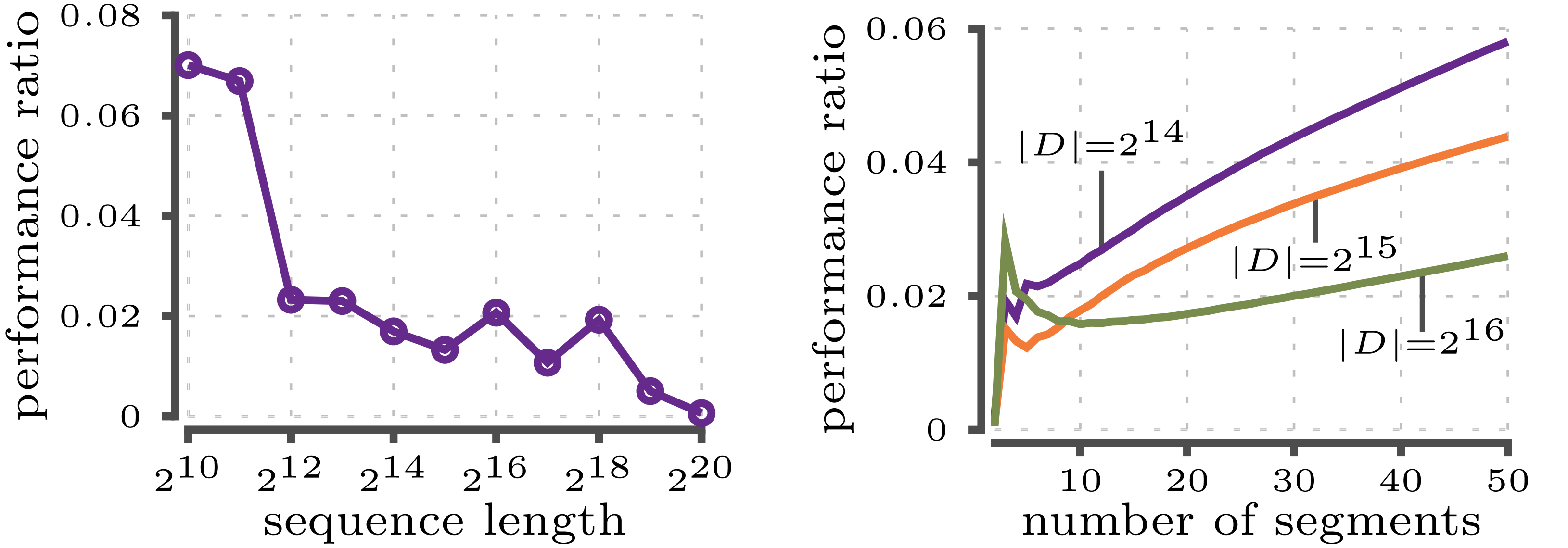
- [Lijffijt, Papapetrou, Puolamäki, PKDD 2012]
- many sequence algorithms use sliding windows
- how to choose window lengths?
- treat as an optimization problem
- pick a set of window lengths that explains most of the variability in statistics over all possible window lengths

fast sequence segmentation using log-linear models

- [Tatti, DMKD 2013]



(a) Sequence (b) Cost of segmentation



(a) speedup vs. sequence length (b) speedup vs. # of segments

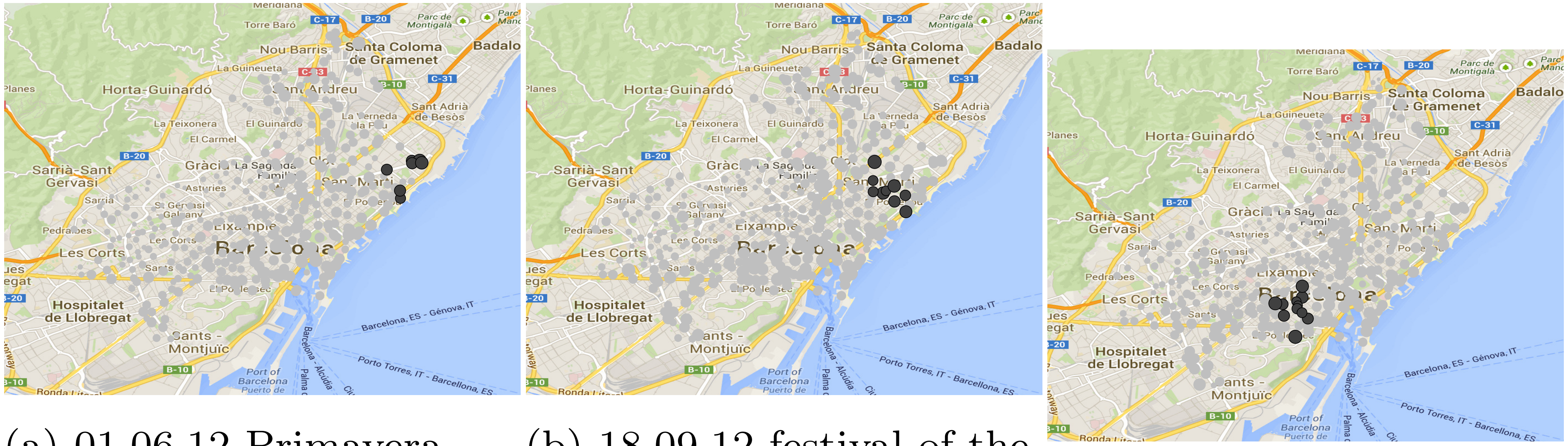
future directions

new research directions

- graph mining and social network analysis
- analysis of information networks
- analysis of evolving networks
- smart cities

recent paper

- given a graph with weights on the nodes, find dense and heavy subgraph
- solutions using submodular function maximization and semidefinite programming
- applications in finding events in cities



(a) 01.06.12 Primavera sound music festival

(b) 18.09.12 festival of the Poblenou neighborhood

(c) 31.10.12 Halloween

ALGODAN Algorithmic Data Analysis

Regression models for data streams with missing values

Indrė Žliobaitė
Postdoctoral Researcher

Aalto University

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Problem setting

- Predictive modelling for streaming data
  - data arrives and needs to be mined in real time
  - real valued inputs, real valued target variable
  - linear regression models

Examples of streaming data

Sensor data (monitoring)

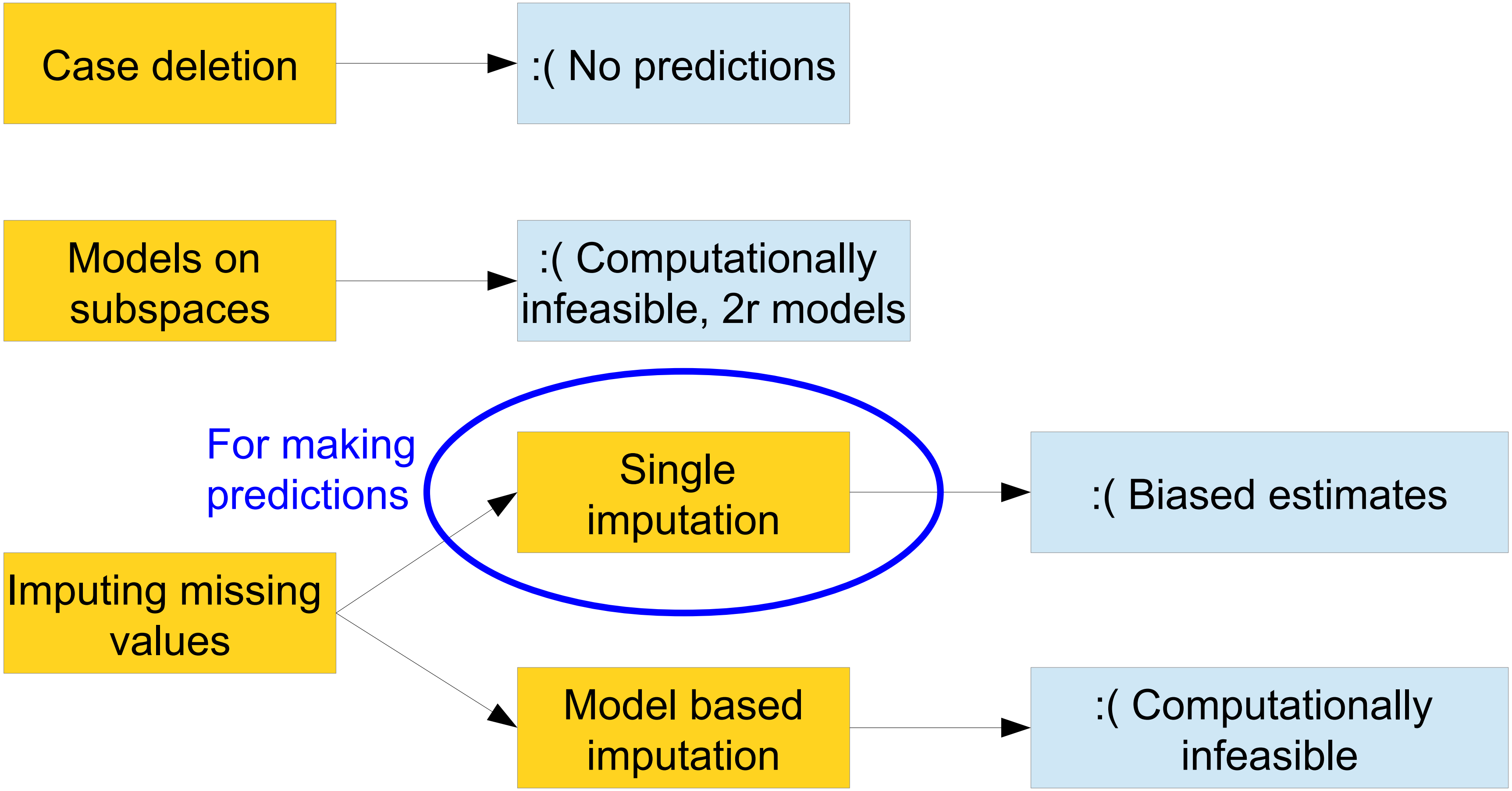
Transactional data (events)

Web data (user generated content)

Problem setting

- Predictive modelling for streaming data
    - data arrives and needs to be mined in real time
    - real valued inputs, real valued target variable
    - linear regression models
- During operation predictive models can be regularly updated with recent data
- **Problem**: *massively* missing input data, while predictions are needed continuously
- **Our approach**: make predictive models robust to missing data, use simple mean imputation

Possible solutions



Case deletion

:(No predictions

Models on subspaces

:(Computationally infeasible, 2r models

For making predictions

Single imputation

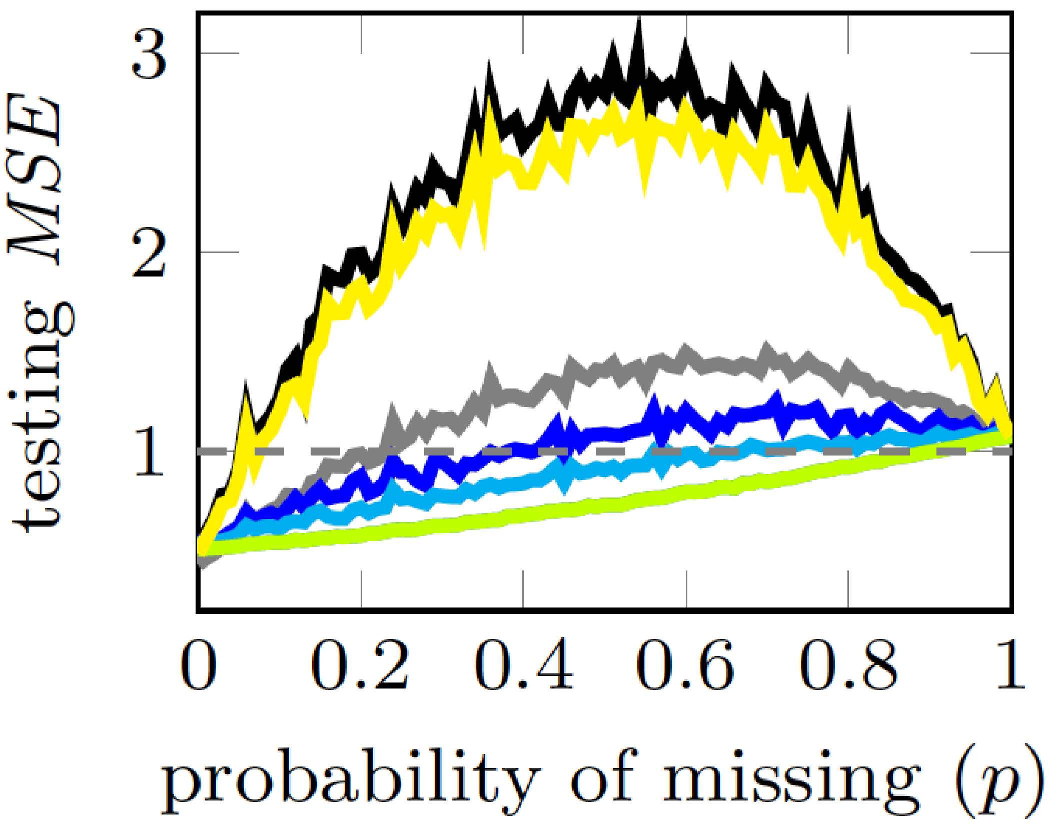
:(Biased estimates

Imputing missing values

Model based imputation

:(Computationally infeasible

Predictions by different linear models



What makes
a predictive model
robust to missing data?

Analysis of the expected error

- Expected MSE of a linear model

p – prior probability of missing, ß – regression coefficients,

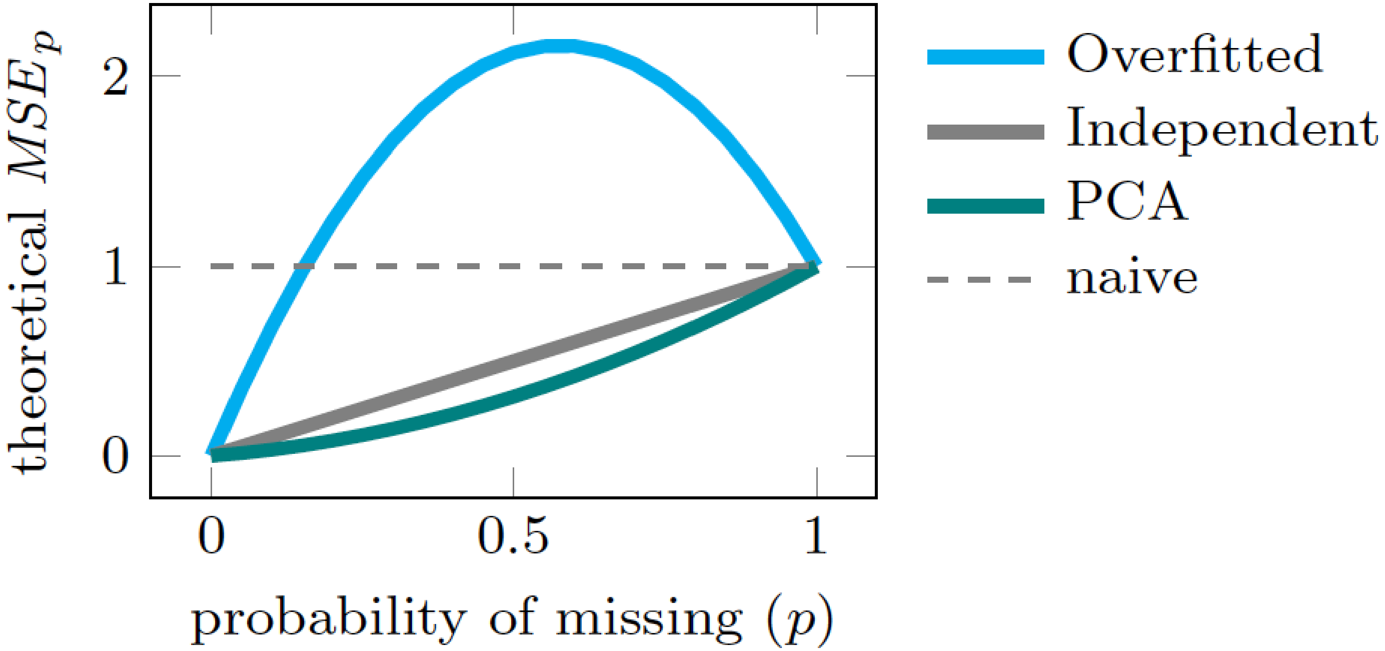
C – covariance matrix of inputs, I – identity matrix

$$E[MSE_p^\star] = \underbrace{(1-p)E[MSE_0] + p}_{\text{MSE grows linearly with number of missing inputs}} - \underbrace{p(1-p)}_{\text{Quadratically}}\underbrace{\boxed{\beta^T(\mathbf{C}-\mathbf{I})\beta}}_{\textbf{Deterioration Index D}}$$

- If D = 0 inputs are treated as independent
- We can make use of dependency in inputs to ensure sub-linear MSE growth

Illustrative example

- Data: x1 = x2 = x3 = x4 = y ~ N(0,1)



Independent: $y = x_1$
PCA: $y = 0.25x_1 + 0.25x_2 + 0.25x_3 + 0.25x_4$
Overfitted: $y = 2x_1 - 1.5x_2 + x_3 - 0.5x_4$

Theoretically optimal model

$$\hat{\beta}_{ROB} = \left((1-p)\mathbf{X}^T\mathbf{X} + pn\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

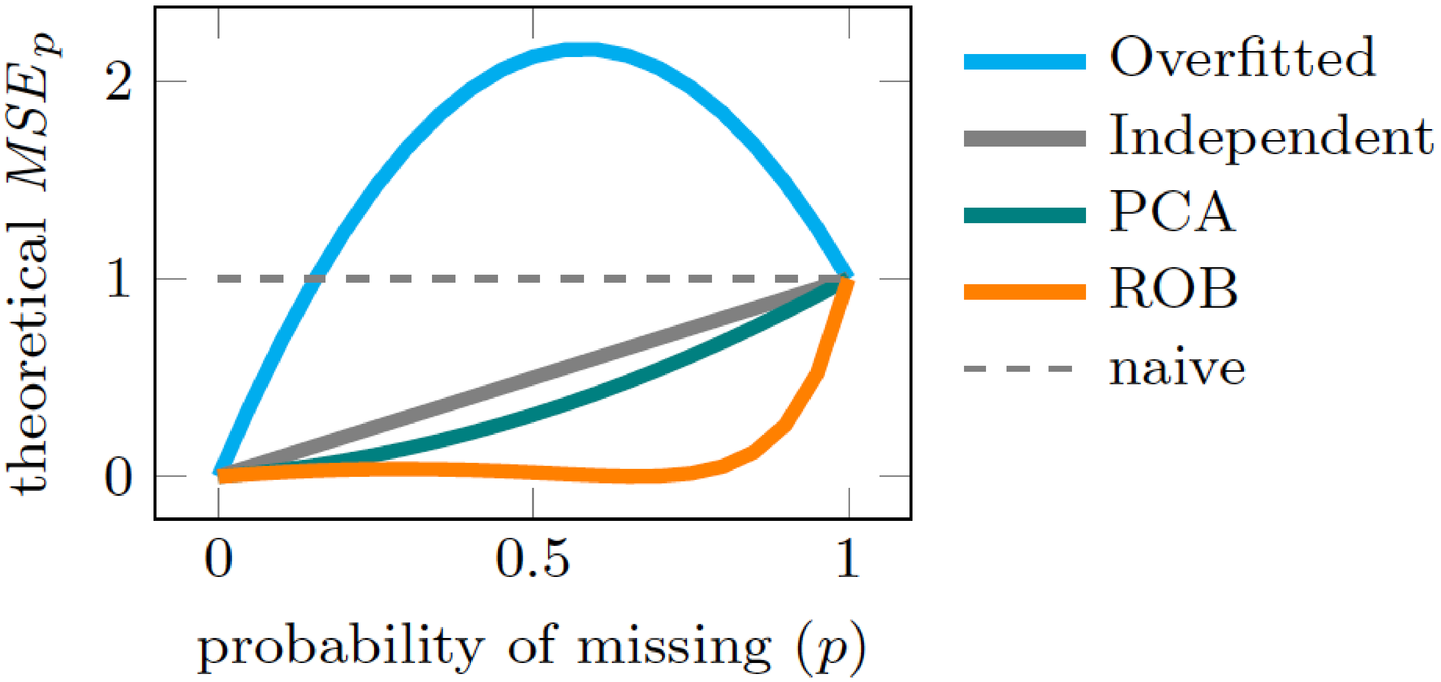
prior probability
of missing values

minimizes MSE given
prior probability of missing values

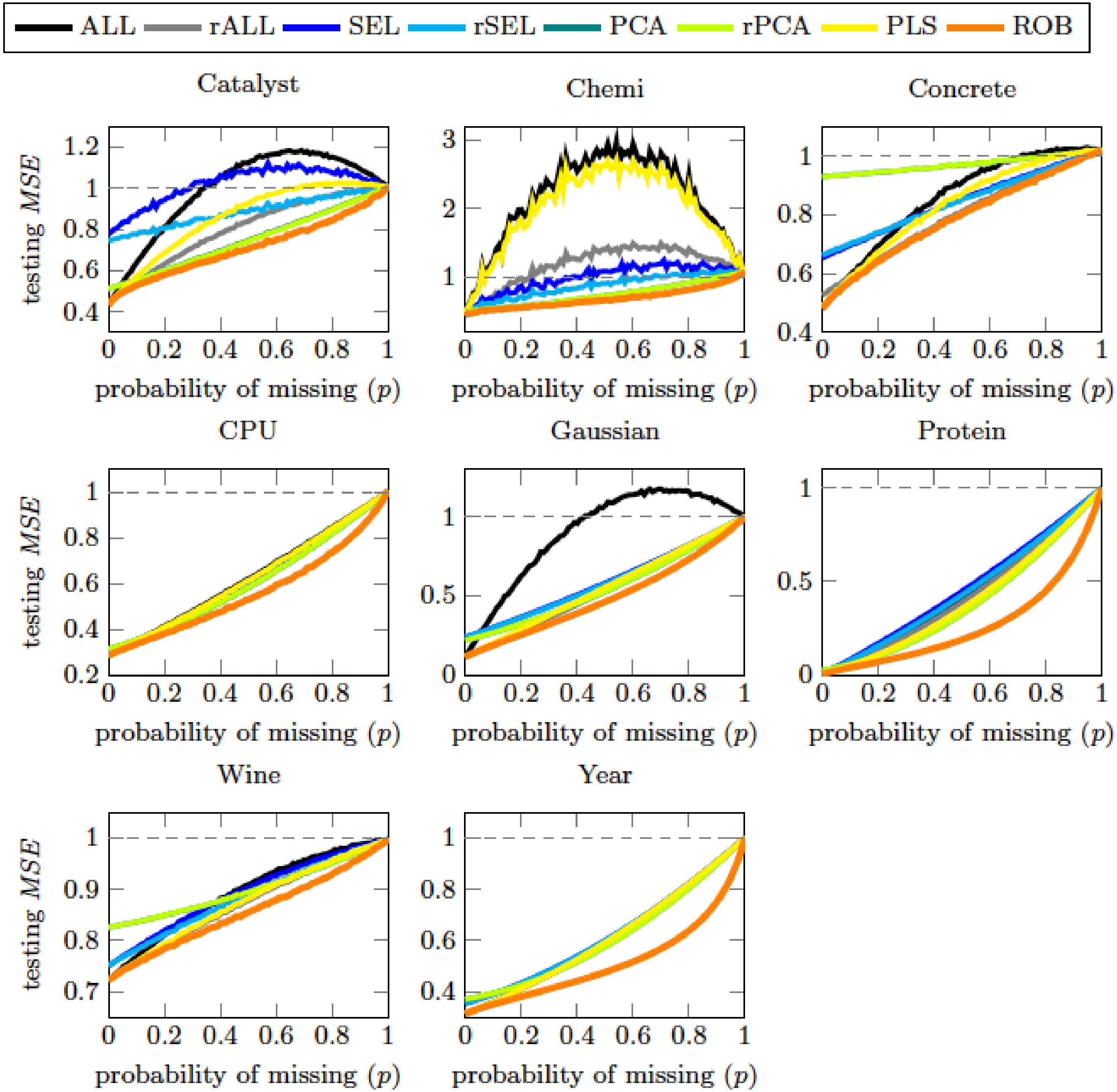
$$\hat{\beta}_{RR} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

is similar to regularized regression

Illustrative example

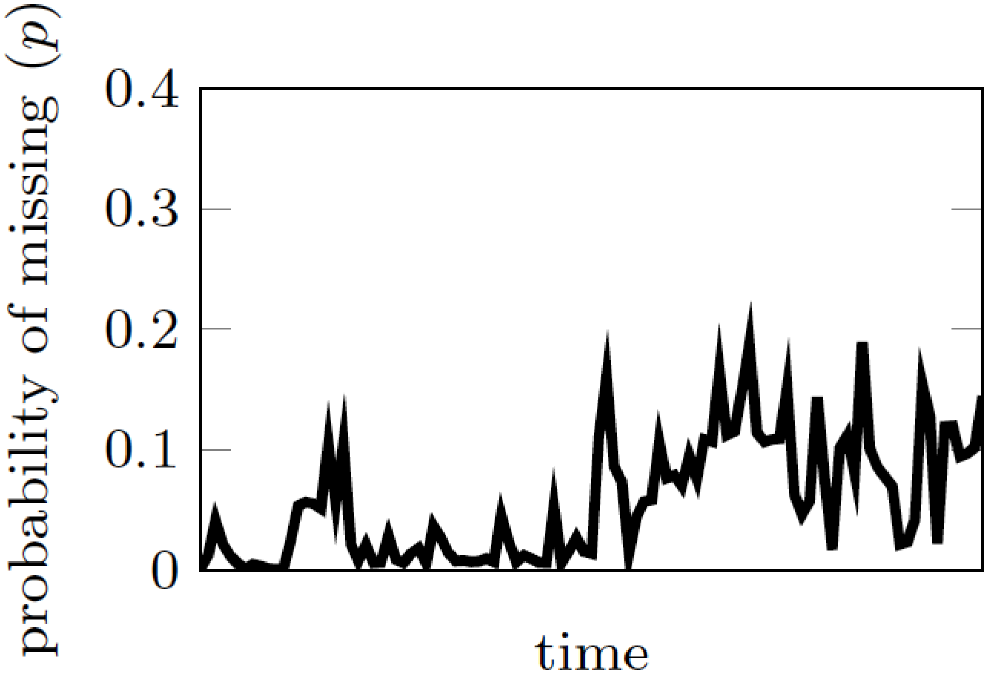


Independent: $y = x_1$
PCA: $y = 0.25x_1 + 0.25x_2 + 0.25x_3 + 0.25x_4$
Overfitted: $y = 2x_1 - 1.5x_2 + x_3 - 0.5x_4$
ROB regression: different model for each value of p


ALL
rALL
SEL
rSEL
PCA
rPCA
PLS
ROB
Catalyst
Chemi
Concrete
CPU
Gaussian
Protein
Wine
Year
testing MSE
probability of missing (p)
0
0.2
0.4
0.6
0.8
1
1.2
2
3
0.5
0.7
0.9

Addressing data stream challenges

- Data evolves over time
  - not only data distribution
  - but also *how* data is missing



Online adaptive ROB algorithm

new observation $\mathbf{x}$ arrives, predict $\hat{y} = \mathbf{x}\beta$

true target value y arrives

update missing value estimate $p \leftarrow \gamma \frac{m}{r} + (1-\gamma)p$

If no missing values

update covariance estimate and model

$$\mathbf{S}_t = \mathbf{S}_{t-1} + \mathbf{x}\mathbf{x}^T - p(\mathbf{x}\mathbf{x}^T - \mathbf{I})$$

$$\beta_t = \beta_{t-1} + \mathbf{S}_t^{-1}\mathbf{x}(y - \mathbf{x}^T\beta_{t-1}) - \mathbf{S}_t^{-1}p(\mathbf{x}\mathbf{x}^T - \mathbf{I})\beta_{t-1}$$

Summary

- We developed
  - an optimization criteria (MSE) for regression being robust to massively missing data
  - a corresponding regression model
  - an algorithm for online operation on streaming data (recursive updates)

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

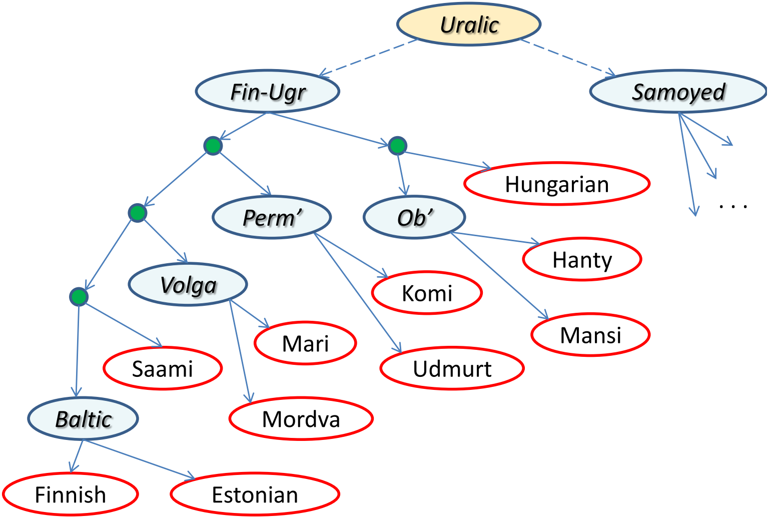
ALGODAN Algorithmic Data Analysis

Modeling inter-linguistic relationships and language evolution

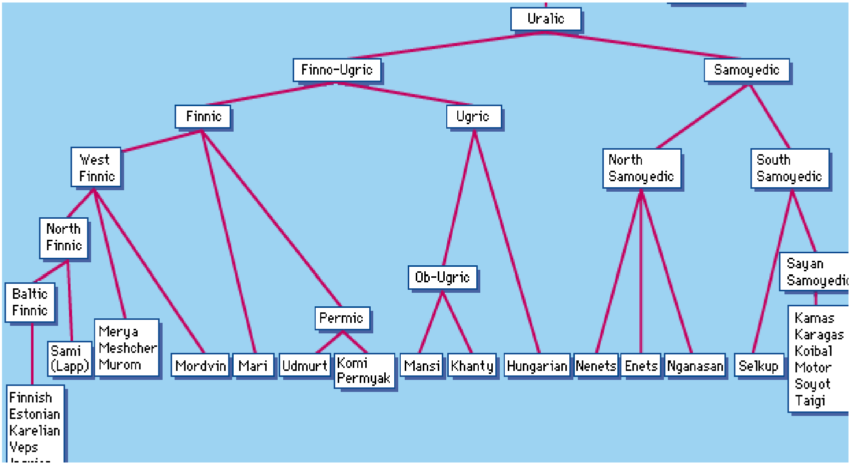
Roman Yangarber
Algodan
March 2014

University of Helsinki, Finland

Uralic Language Family



Uralic Language Family



Data sources

Data is arranged in *Cognate Sets*: set of genetically-related words, from different languages in the language family

→ ... Raw data sample
→ ... Aligned data sample

Central Principle

- **Regularity of sound change:**
  - *Sound change is conditioned only on its phonetic environment*, not on any other factor.
  - Sound change is *deterministically* conditioned

- **NB:** different from, e.g., biological sequence alignment, where mutations are sporadic.

Example sound change: German vs. Germanic

Germanic **t**	English	German
	two	zwei
	ten	zehn
	to	zu
	tell	zähle-n
	tooth	Zahn
	tear	Zähre
	tow	ziehe-n
	tail	Zagel
	heart	Herz
	...	
	tip	Zipf-el
	tide	Zeit
	timber	Zimmer
	...	

stone	Stein
star	Stern
...	

dead	tot
door	Tür
do	tu-n
under	unter
...	

Example sound change: German vs. Germanic

- There are “exceptions” to rules
  - “regular” exceptions?
  - rare/occasional exceptions?

→ probabilistic modeling

→ MDL

code most of the data with rules, then code the exceptions.

Principal Tasks

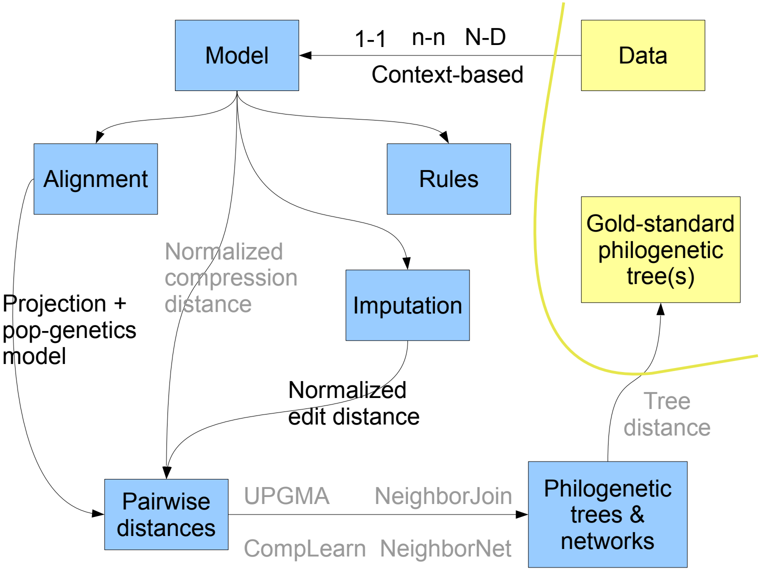
Long-term goal: Determine the origin of everything

- Find cognate sets (from raw language data)
  - difficult to model semantics...
- Find sound-by-sound alignment of all related words
- Find rules of sound correspondence

- Reconstruct philogenetic trees
- Reconstruct proto-forms → at root and internal nodes of the philogeny

- Model borrowing across languages / families
- Model timing → anchor data on absolute time scale

Components


Model
1-1 n-n N-D
Context-based
Data
Alignment
Rules
Gold-standard philogenetic tree(s)
Normalized compression distance
Imputation
Projection + pop-genetics model
Normalized edit distance
Tree distance
Pairwise distances
UPGMA
NeighborJoin
CompLearn
NeighborNet
Philogenetic trees & networks

Problem formulation

Dual problem:

A find the *globally best* **alignment** for the complete data, and

B find the **rules of correspondence**

Chicken and egg...

Approach *in tandem*

Models

Baseline: Initial simplifications

- *Pairwise* alignment: only two languages at a time, "*source*:*target*"
  $\rightarrow$ N-dimensional alignment, $N > 2$ languages
- *1-1* alignment: one source symbol may correspond to only one target symbol—or to empty symbol ϵ (marked "**.**")
  $\rightarrow$ Align n-n symbols (2x2)
- Ignore context
  $\rightarrow$ Model how the *Context* conditions the changes
- Symbols/sounds are treated as ATOMS
  $\rightarrow$ Symbols/sounds analyzed as *vectors of distinctive features*

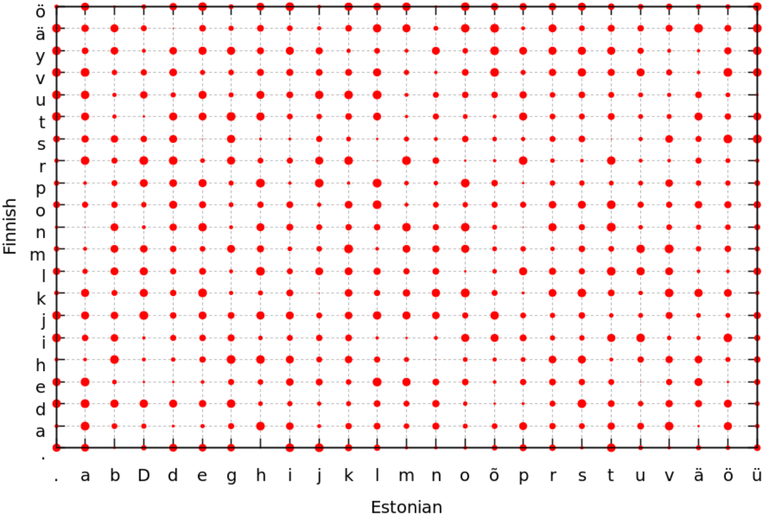
Problem formulation

Alignment → ... Complete data
Rules → in *baseline* model: simply the **counts** of events

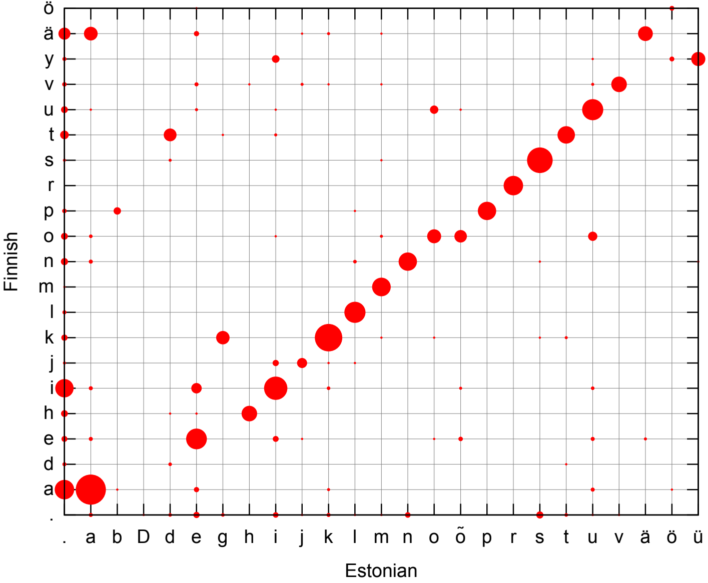
How do we know which rules are better?

(recall, in baseline: rules are 1x1 alingments)

Rules: high entropy



Rules: low entropy



2x2: Aligning Multiple Symbols

Extend the baseline model to a 2x2 model: correspondences of up to two symbols on both sides

The set of admissible *kinds* of events becomes:

$$K = \left\{ \begin{array}{ll} (\# : \#) & (\sigma : .) \\ (. : \tau) & (\sigma : \tau) \end{array} \right\}$$

2x2: Aligning Multiple Symbols

Extend the baseline model to a 2x2 model: correspondences of up to two symbols on both sides
The set of admissible *kinds* of events becomes:

$$K = \left\{ \begin{array}{lll} (\# : \#) & (\sigma : .) & (\sigma\sigma' : .) \\ (. : \tau) & (\sigma : \tau) & (\sigma\sigma' : \tau) \\ (. : \tau\tau') & (\sigma : \tau\tau') & (\sigma\sigma' : \tau\tau') \end{array} \right\}$$

3-D Model

Align more than two languages: e.g., Finnish : Estonian : Mordva

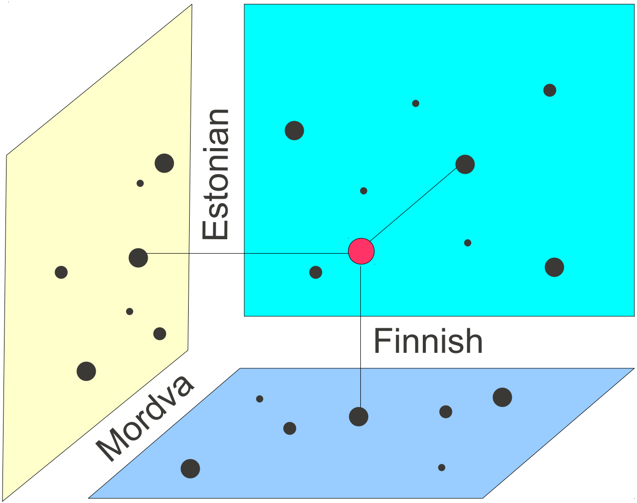
```
y  .  h  d  e  k  s  ä  n
|  |  |  |  |  |  |  |  |
ü  .  h  .  e  k  s  a  .
|  |  |  |  |  |  |  |  |
v  e  χ  .  .  k  s  a  .
```

Model each 3D event as **three** pairwise events

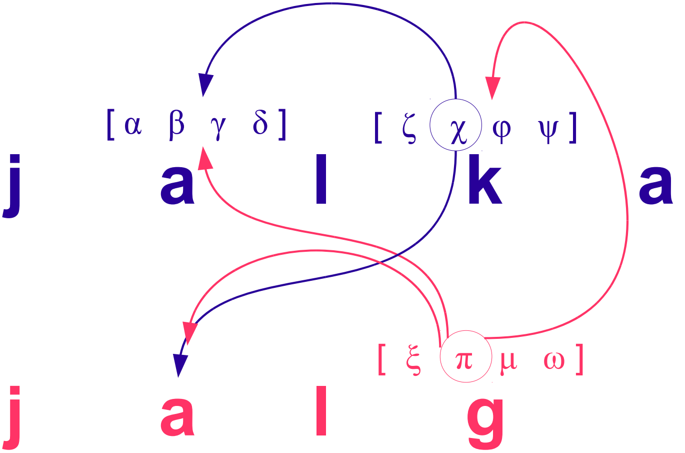
Some examples are *incomplete* – missing data in one language:

```
h  a  a  m  u
|  |  |  |  |
—  —  —  —  —
|  |  |  |  |
č  .  a  m  a
```

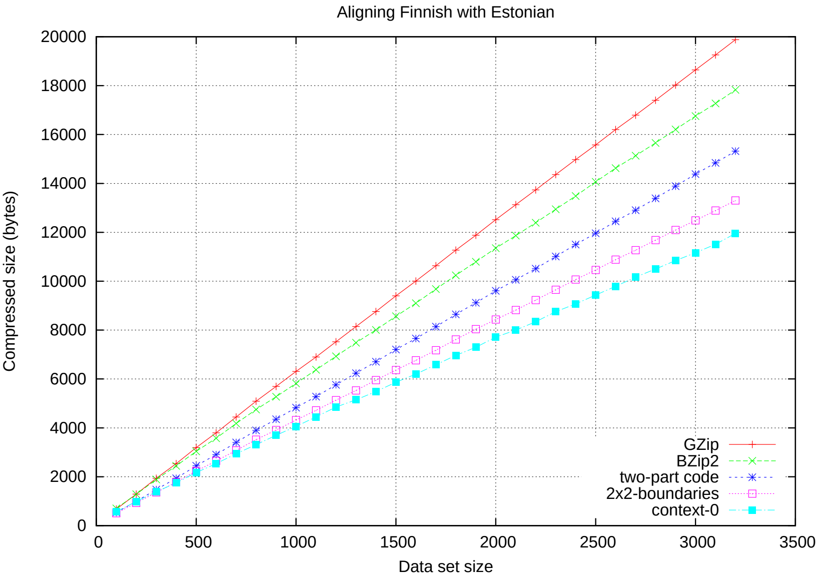
3-D Model



Building decision trees



Performance: Compression rates



Language Distance

Sanity check: Use alignment to measure inter-language distances

- Cost for different language pairs $C(\mathbf{a}, \mathbf{b})$ are not comparable
- *Normalised Compression Distance* (Cilibrasi&Vitanyi, 2005)

$$\delta(\mathbf{a}, \mathbf{b}) = \frac{C(\mathbf{a}, \mathbf{b}) - \min(C(\mathbf{a}, \mathbf{a}), C(\mathbf{b}, \mathbf{b}))}{\max(C(\mathbf{a}, \mathbf{a}), C(\mathbf{b}, \mathbf{b}))}$$

Align all languages in StarLing *pairwise*, e.g., using two-part 1x1 model

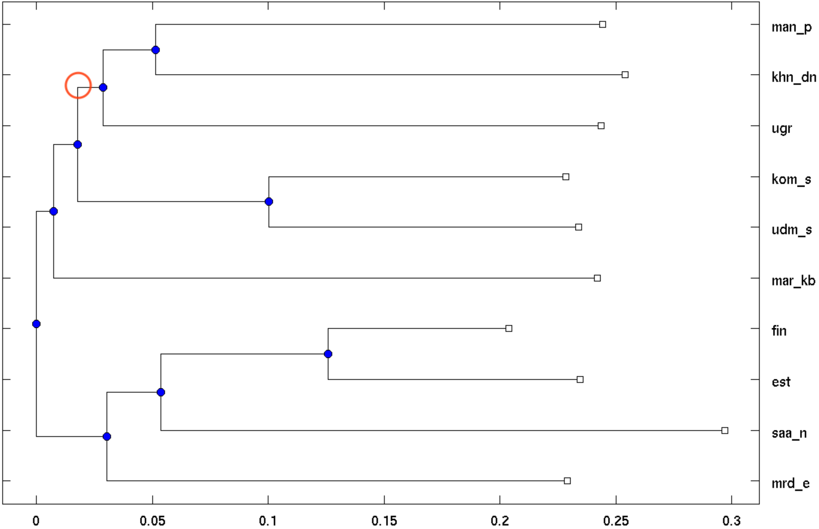
$\rightarrow$ ...

NCD

	fin	*khn*	*kom*	*man*	*mar*	*mrd*	*saa*	*udm*	*ugr*
est	**.372**	.702	.704	.716	.703	.665	.588	.733	.778
fin		.731	.695	.754	.695	.635	.589	.699	.777
khn			.672	**.633**	.701	.718	.668	.712	.761
kom				.675	.656	.678	.700	**.417**	.704
man					.676	.718	.779	.688	.752
mar						**.648**	.671	.674	.738
mrd							.646	.709	.722
saa								.686	.760
udm									.759
ugr									

Table: Pairwise normalised compression distances for Finno-Ugric sub-family of Uralic, StarLing data.

NED with Neighbor Joining


man_p
khn_dn
ugr
kom_s
udm_s
mar_kb
fin
est
saa_n
mrd_e