Cluster analysis of time series data

Tomáš Bartoň

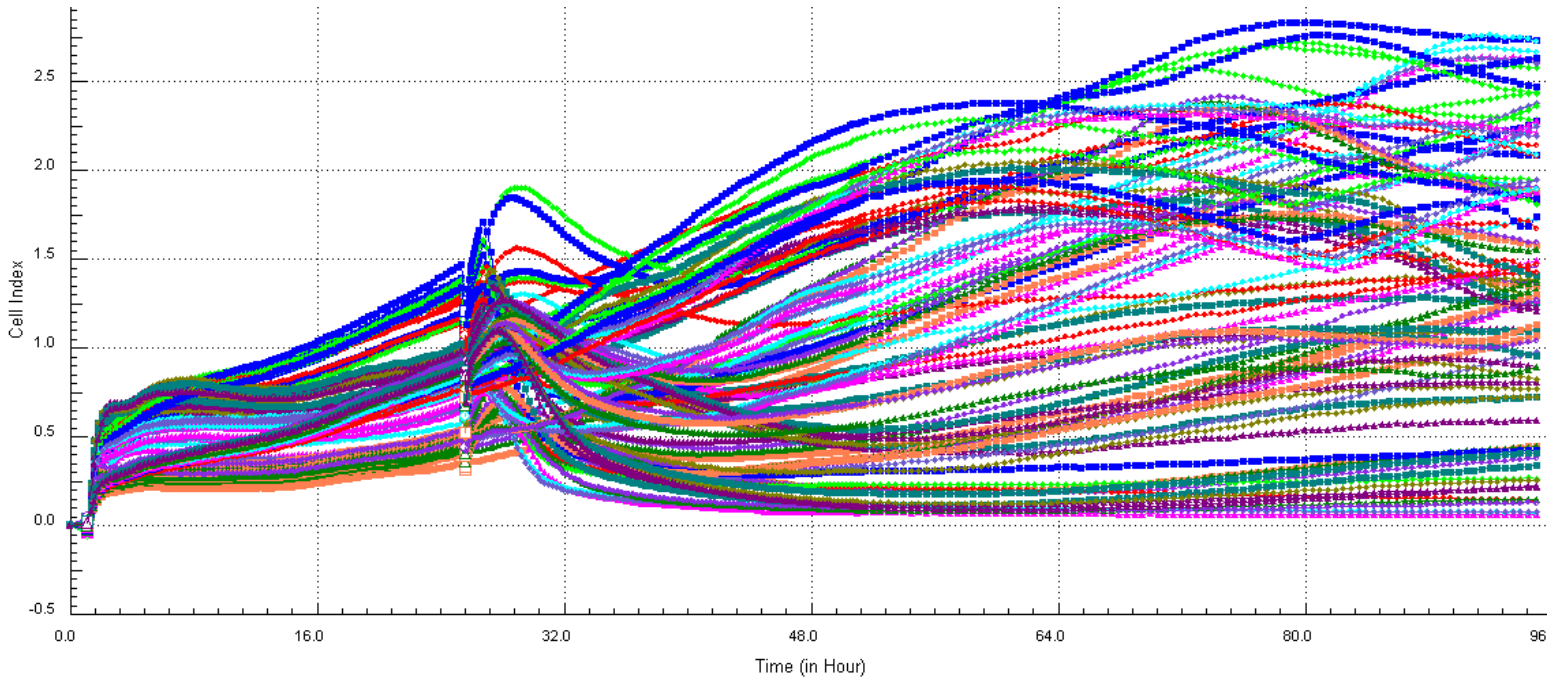
Supervisor: **Ing. Pavel Kordík, Ph.D.**

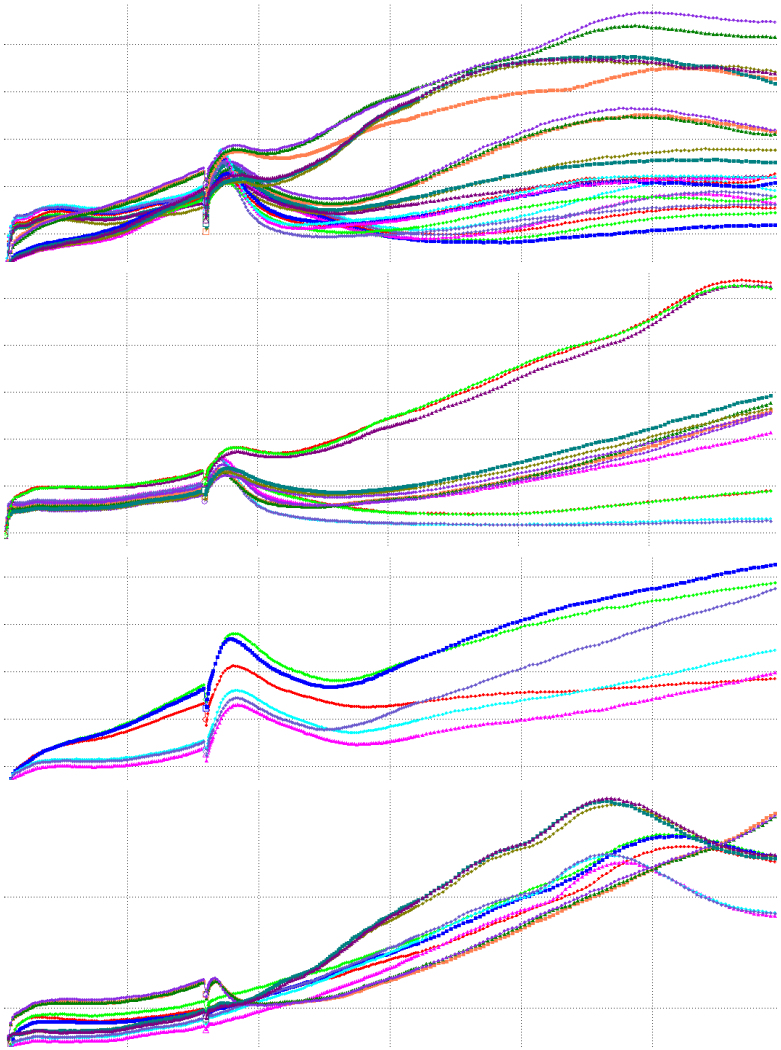
Department of Theoretical Computer Science
Faculty of Information Technology
Czech Technical University in Prague

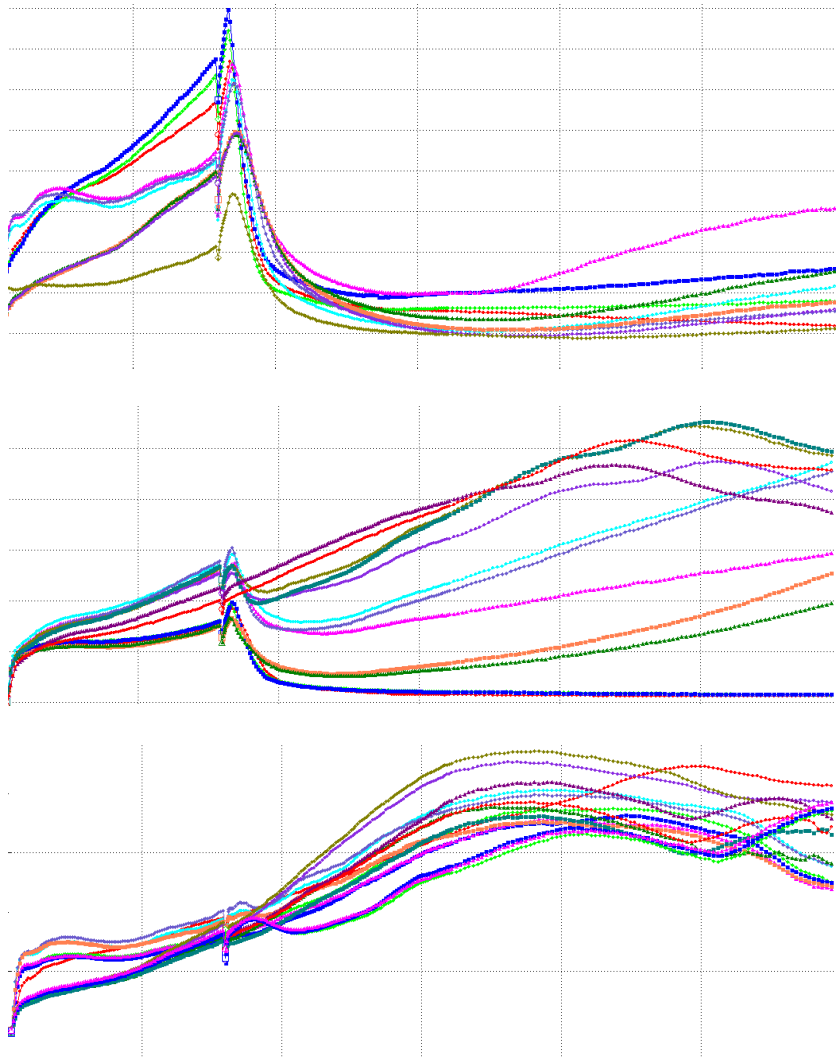
January 5, 2012

The Problem

- find suitable method for identification of patterns
- assign samples into (unknown) groups

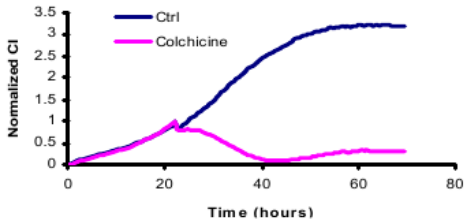




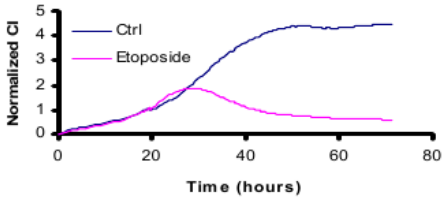


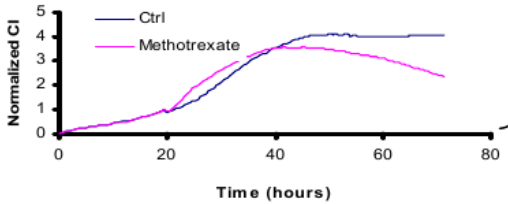
Goals

- capture global trends
- absolute values (sometimes) doesn't matter
- signals are not periodical
- discover unknown patterns

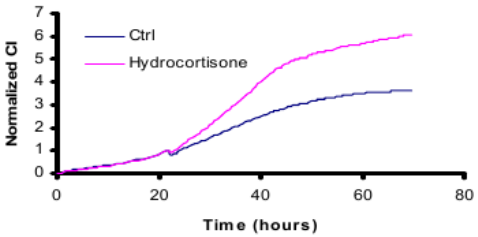


Anti-mitotic

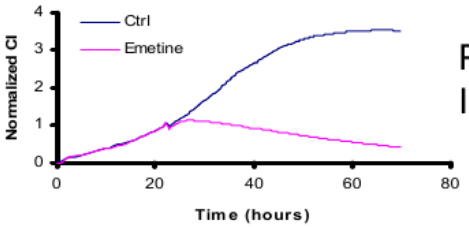




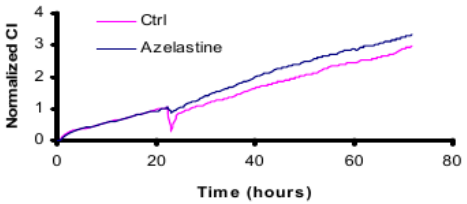
DNA Damaging

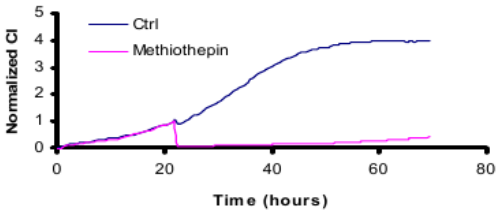


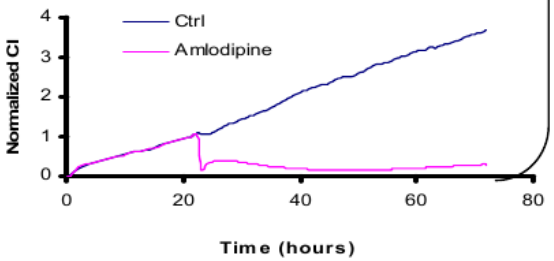
Nuclear Receptor modulator



Protein Synthesis Inhibitors







Calcium Modulators

Phases of clustering process

1. data cleaning
2. data integration
3. data selection
4. data transformation
5. clustering
6. pattern evaluation
7. knowledge representation

Clustering

No “correct” clustering exists

Definition

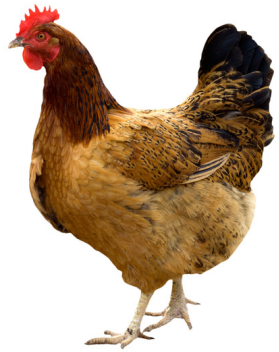
"Those methods concerned in some way with the identification of homogeneous groups of objects"
[Arabie et al., 1996]

Definition

"A cluster is a set of entities that are alike, and entities from different clusters are not alike" [Everitt, 1993]

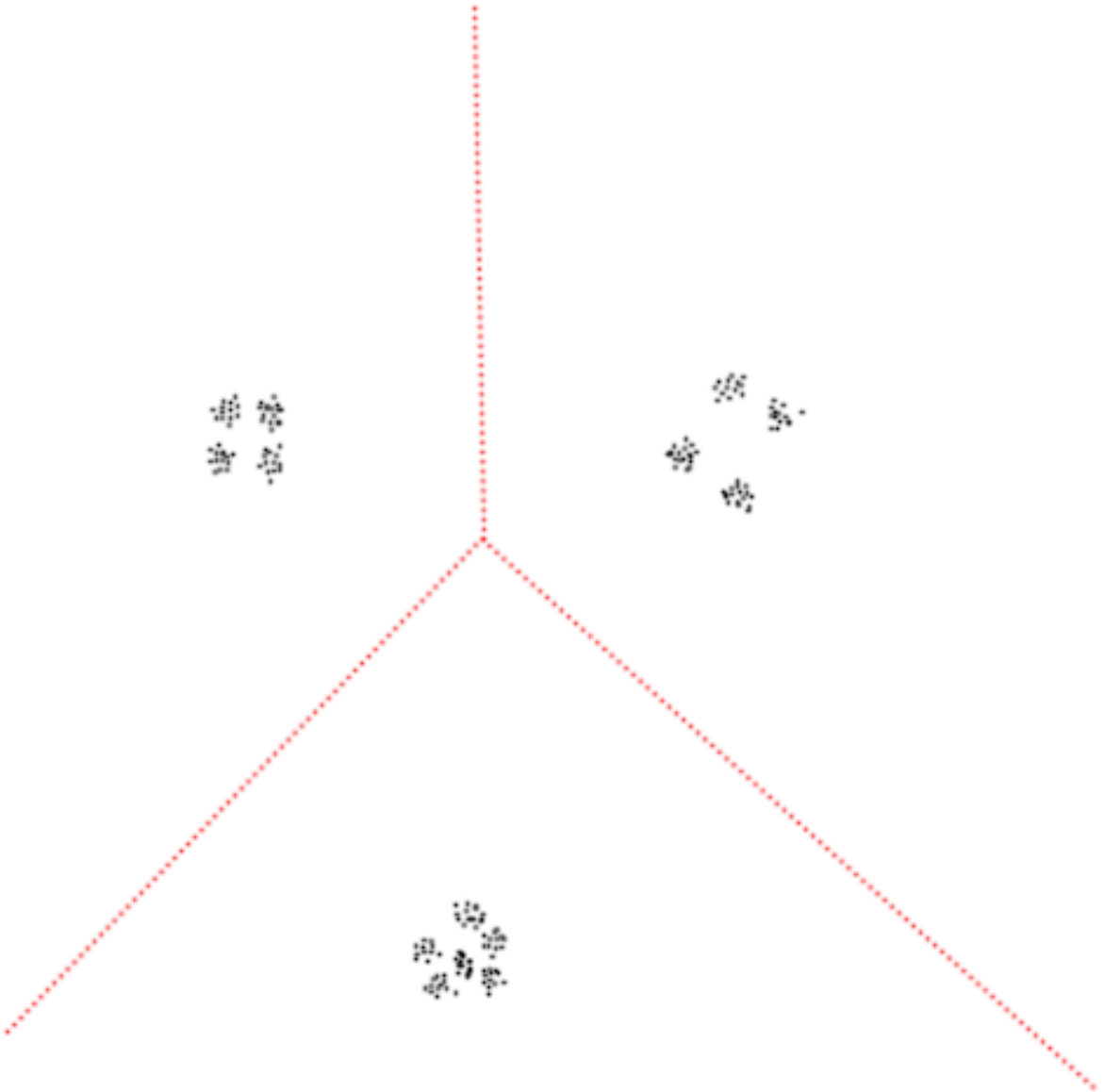
- clustering can be used for understanding data

- to perform clustering you need to understand data

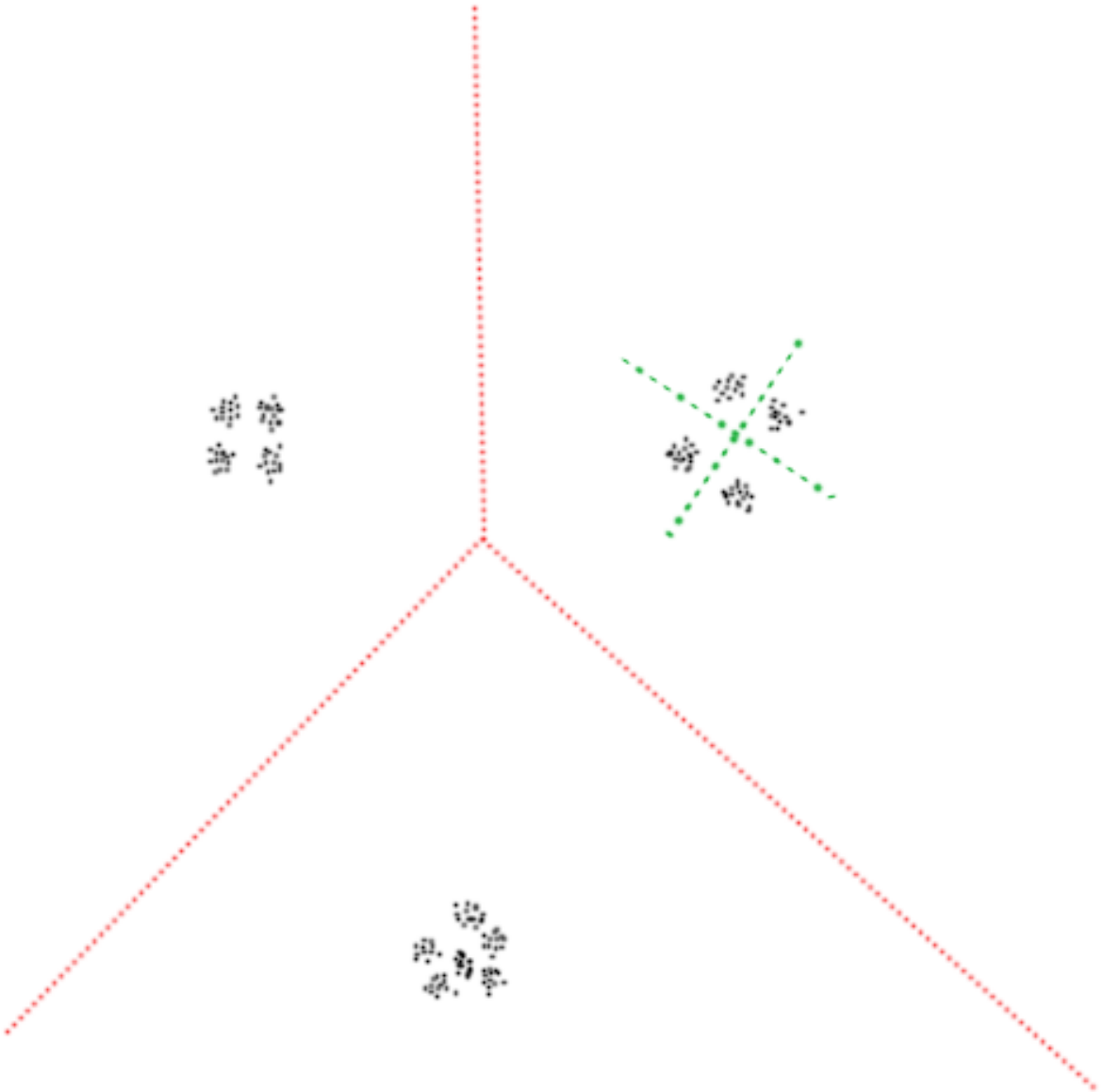


Determine number of clusters

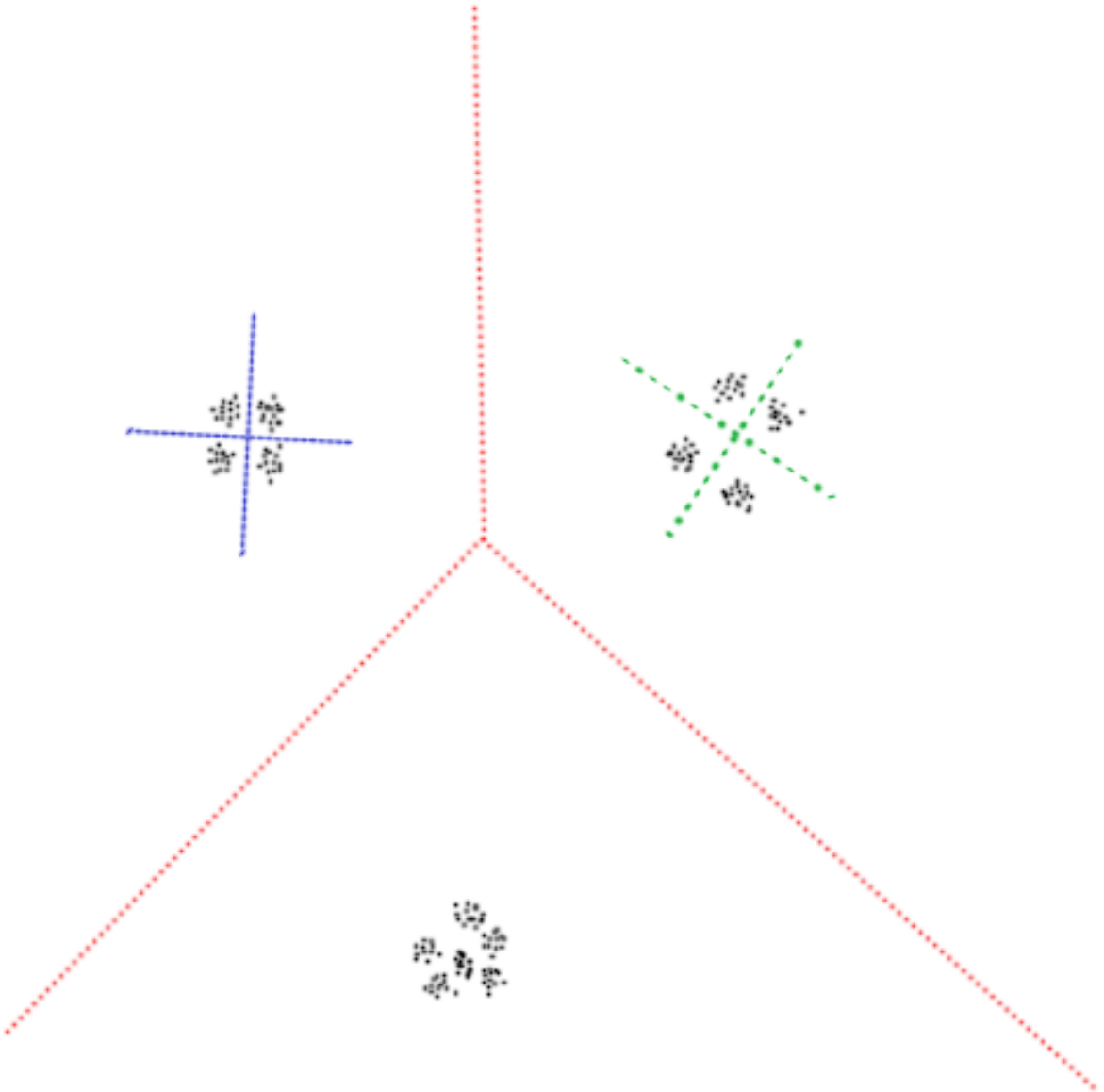
$k = 3$



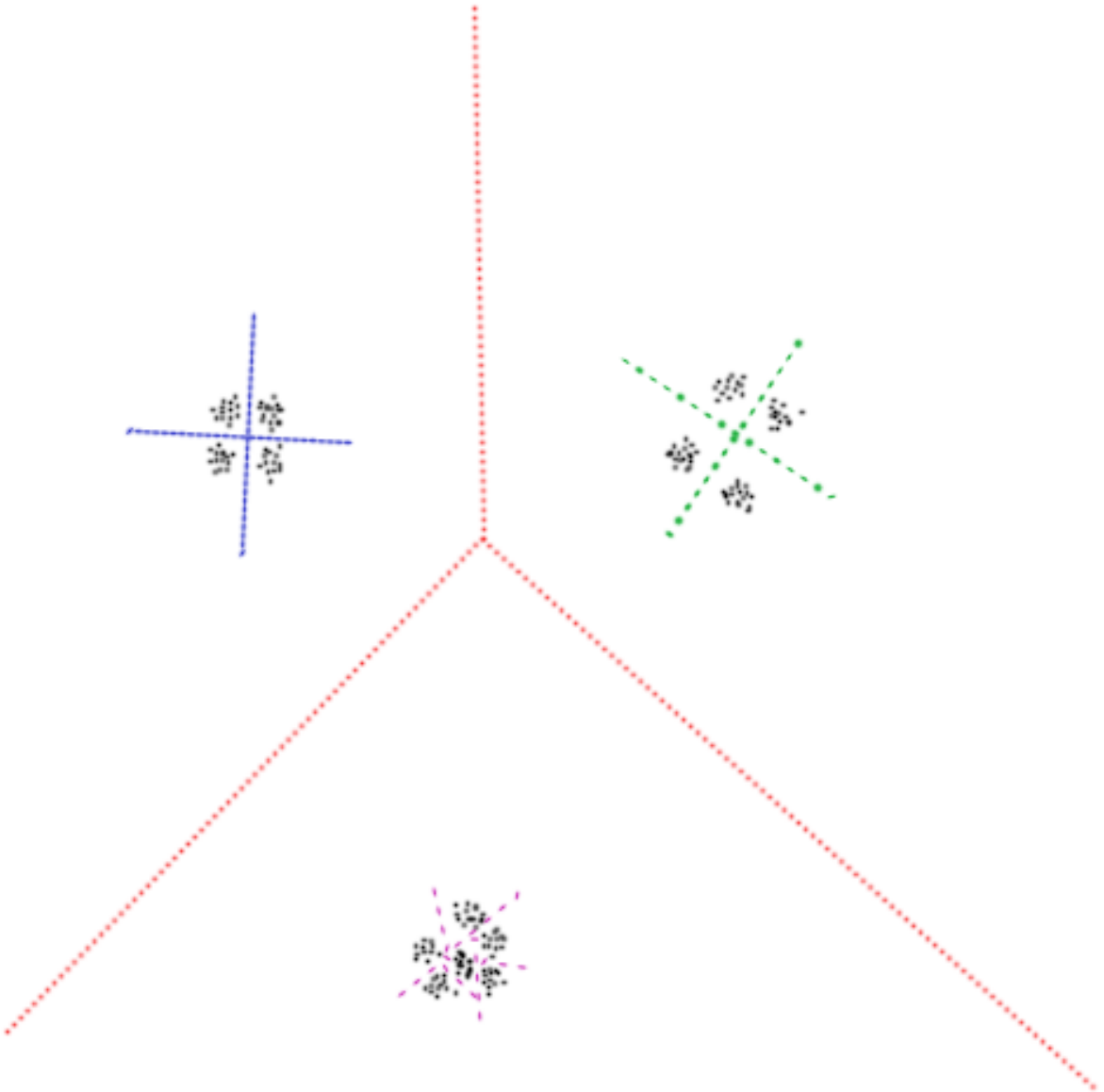
$k = 6$



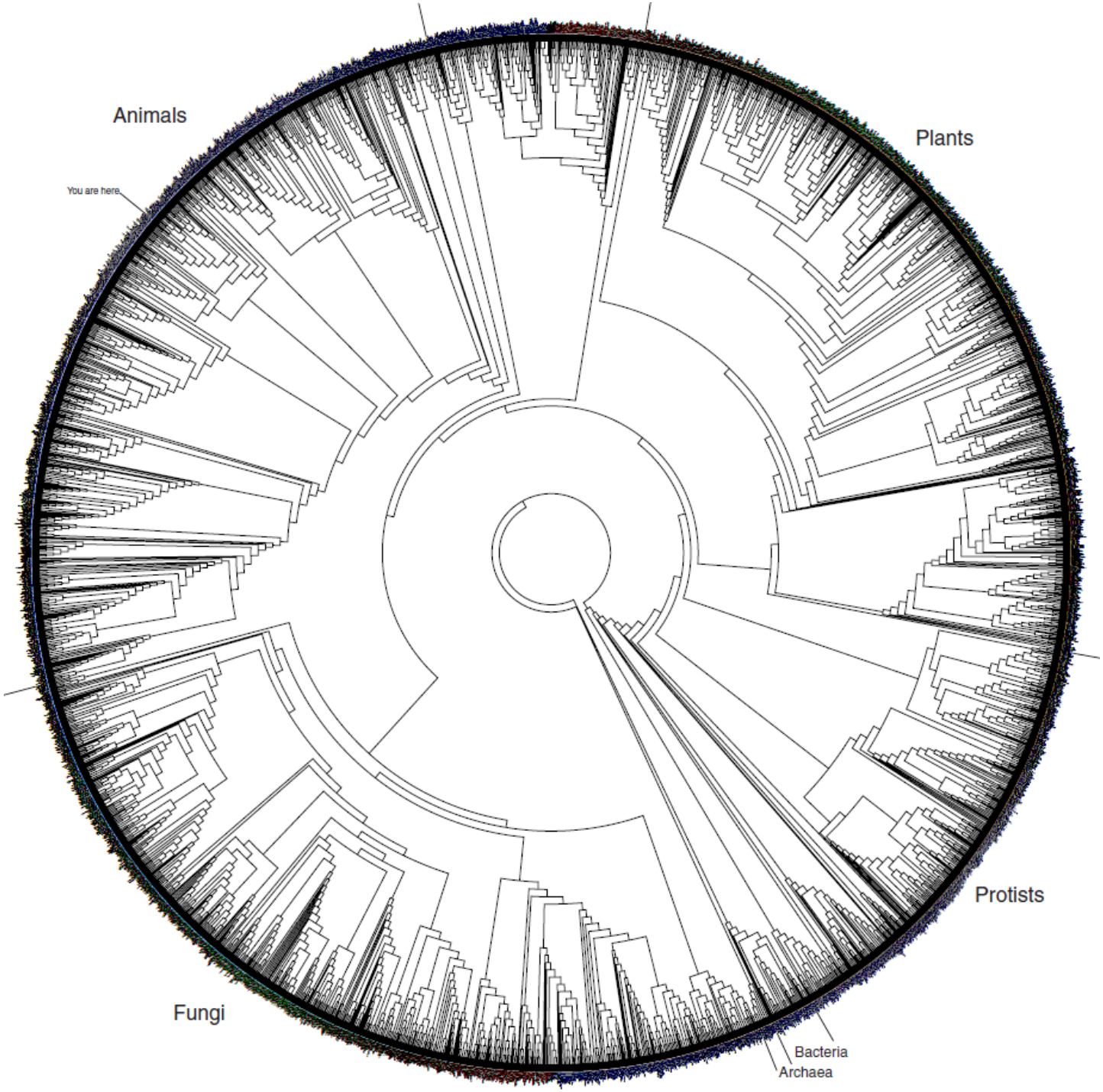
$k = 9$



$k = 14$



Determining the number of clusters in a data set is challenging [Mufti et al., 2005]


Animals
Plants
You are here
Protists
Fungi
Bacteria
Archaea

www.zo.utexas.edu/faculty/antisense/tree.pdf

from Chinese encyklopedia Heavenly Emporium of Benevolent Knowledge. Animals are divided into [Borges, 1952]:

- those that belong to the emperor
- embalmed ones
- those that are trained
- suckling pigs
- mermaids
- fabulous ones
- those that are included in this classification
- innumerable ones
- etcetera

Clustering is ill-defined [Caruana et al., 2006]

All we care about is the "usefulness" of the clustering for achieving our final goal [Guyon et al., 2009]

Time series

Problem

- sensitive to small changes
- sum of distance does not capture shape of curve
- computationally expensive
- redundant information

Autoregressive model

- predict an output of a system based on the previous outputs

$$X_t = c + \sum_{i=1}^{p} \varphi_i \cdot X_{t-i} + \epsilon_t$$

- φ_i – parameters of the AR model
- X_t – amplitude of the signal
- ϵ_t – white noises

Moving average

$$X_t = \mu + \epsilon_t + \sum_{i=1}^{q} \phi_i \cdot \epsilon_{t-i}$$

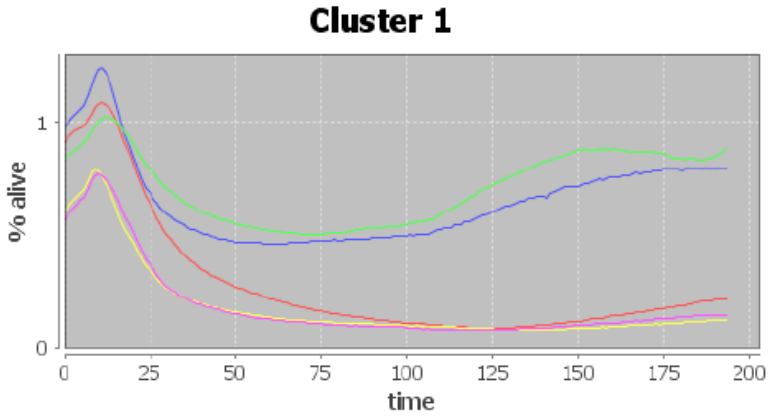
- ϕ_i – parameters of the AR model
- μ – expectations of X_t (often assumed to equal 0)
- ϵ_t – white noises

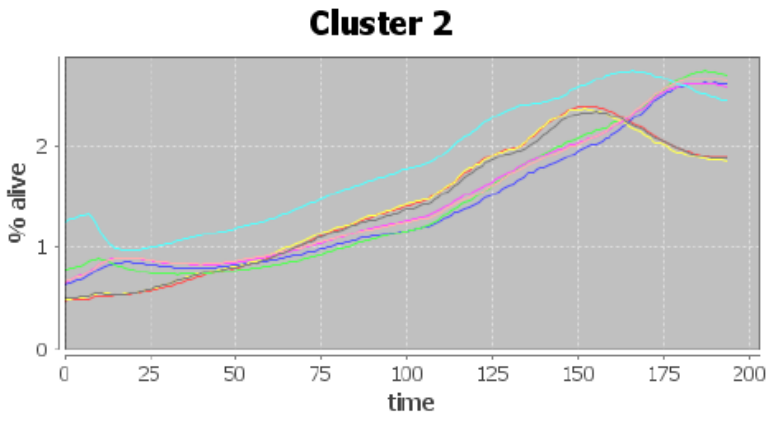
Autoregressive–moving-average model

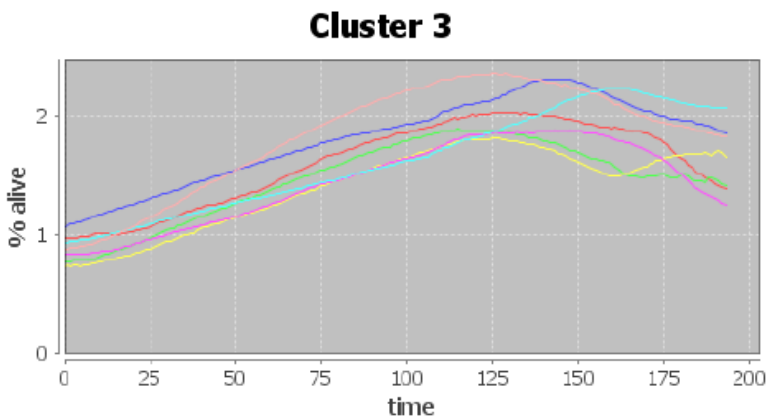
putting all together:

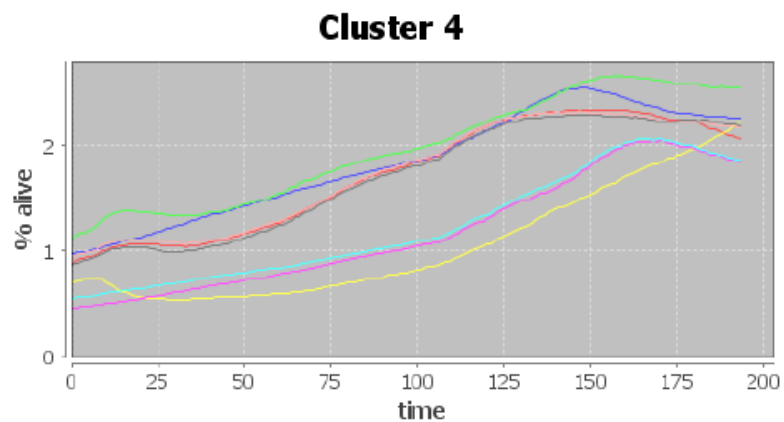
$$X_t = c + \epsilon_t + \sum_{i=1}^{p} \varphi_i \cdot X_{t-i} + \epsilon_t + \sum_{i=1}^{q} \phi_i \cdot \epsilon_{t-i}$$

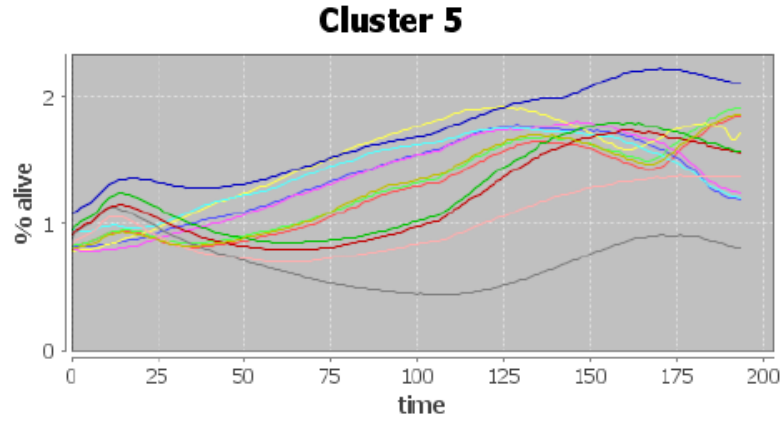
- `ARMA(p, q)` refers to the model with `p` autoregressive terms and `q` moving-average terms
- in Matlab function `armax[Time-domain, data object]`

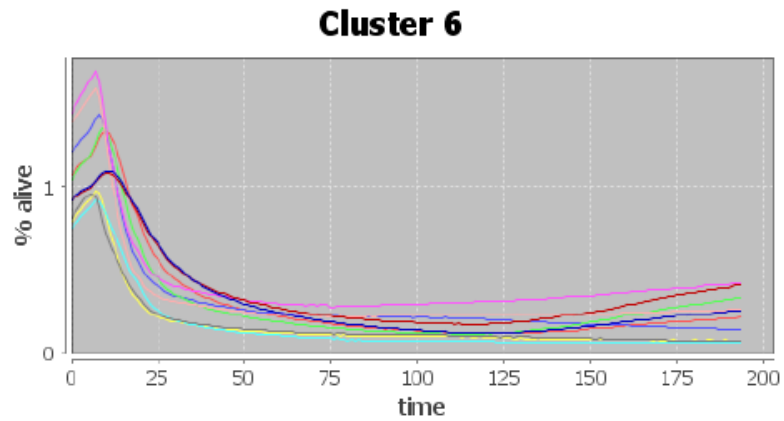

Cluster 1
% alive
0
1
0
25
50
75
100
125
150
175
200
time

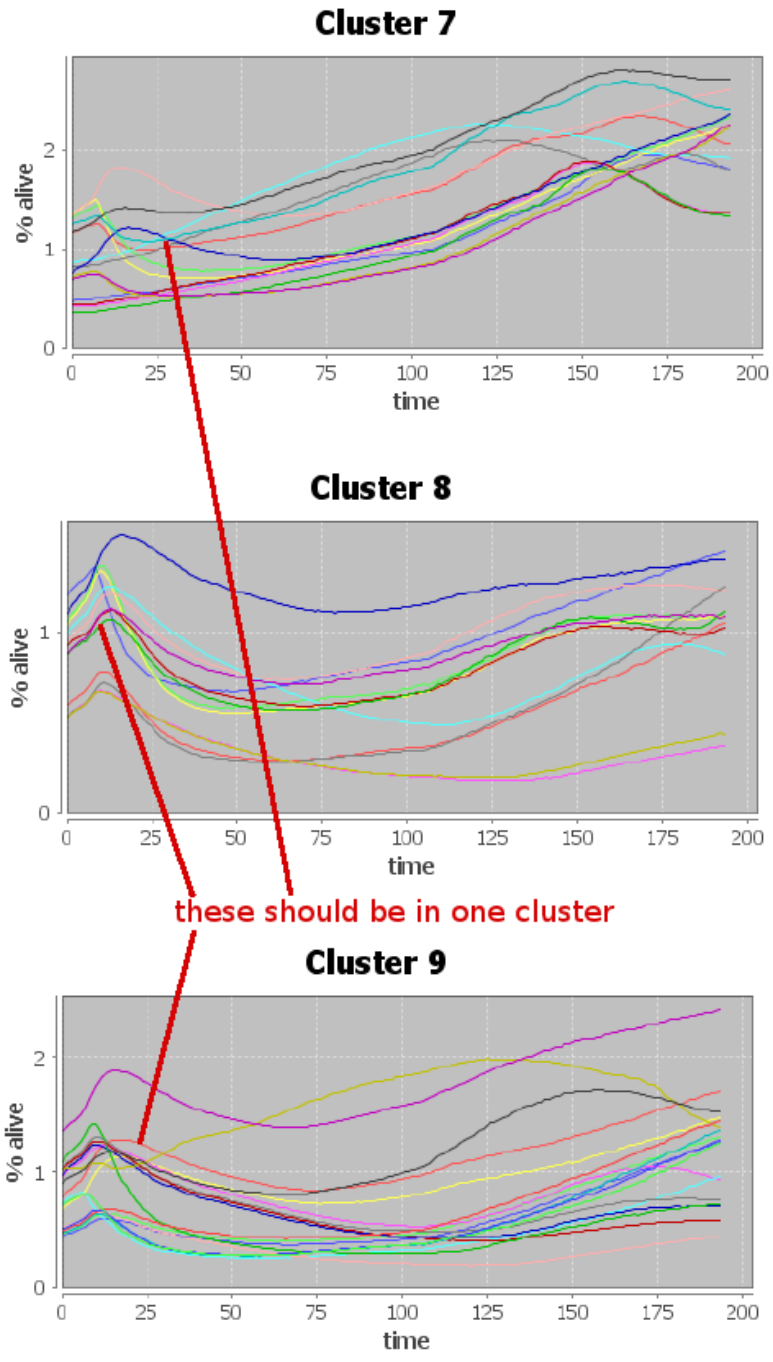

Cluster 2
% alive
0
1
2
0
25
50
75
100
125
150
175
200
time


Cluster 3
% alive
0
1
2
0
25
50
75
100
125
150
175
200
time

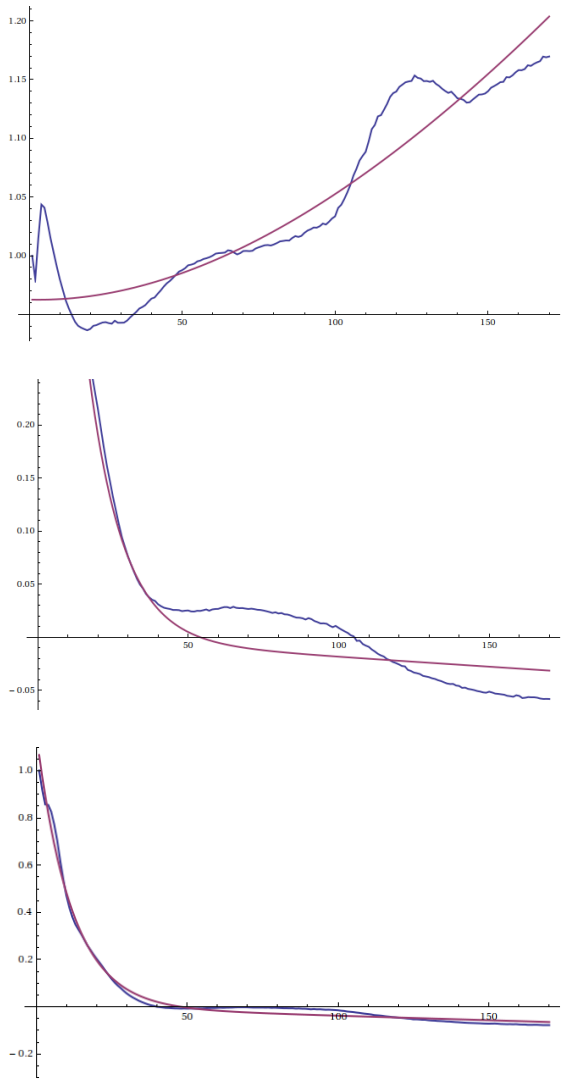

Cluster 4
% alive
0
1
2
0
25
50
75
100
125
150
175
200
time


Cluster 5
% alive
0
1
2
0
25
50
75
100
125
150
175
200
time

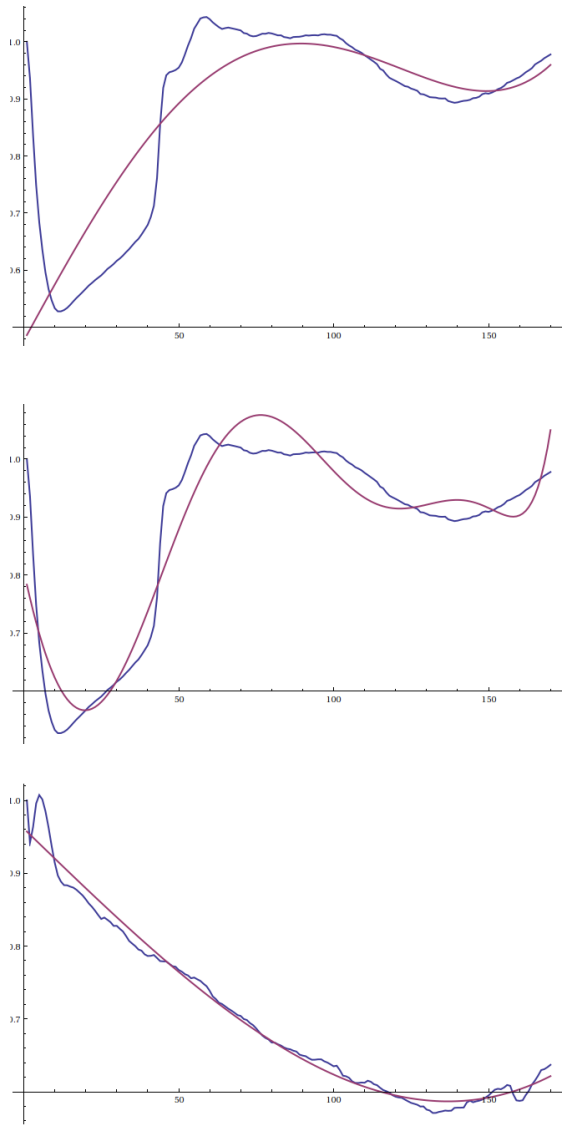

Cluster 6
% alive
0
1
0
25
50
75
100
125
150
175
200
time


Cluster 7
% alive
0
1
2
0
25
50
75
100
125
150
175
200
time
Cluster 8
% alive
0
1
0
25
50
75
100
125
150
175
200
time
these should be in one cluster
Cluster 9
% alive
0
1
2
0
25
50
75
100
125
150
175
200
time

Exponential

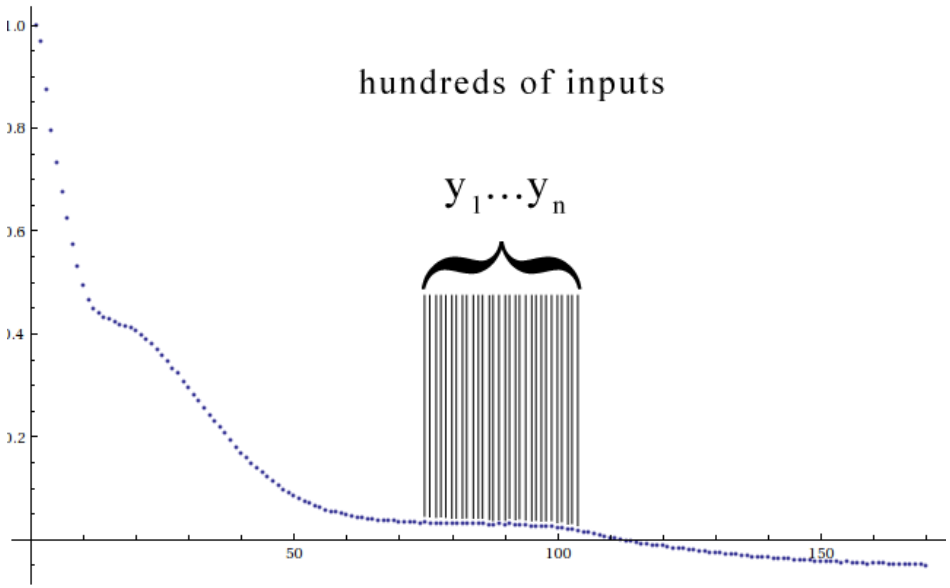


Polynomial



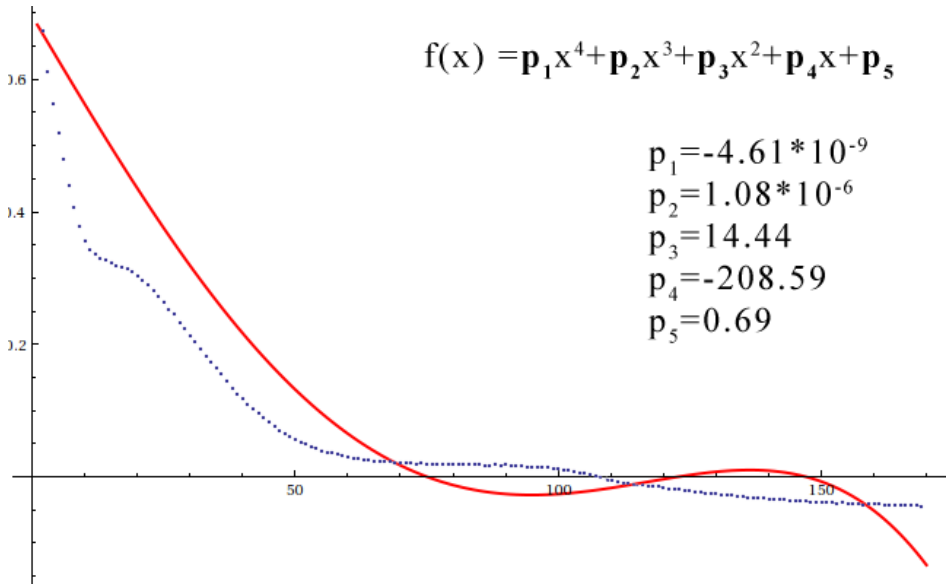
Representation of inputs

Measured values



- too many inputs
- does not represent patterns

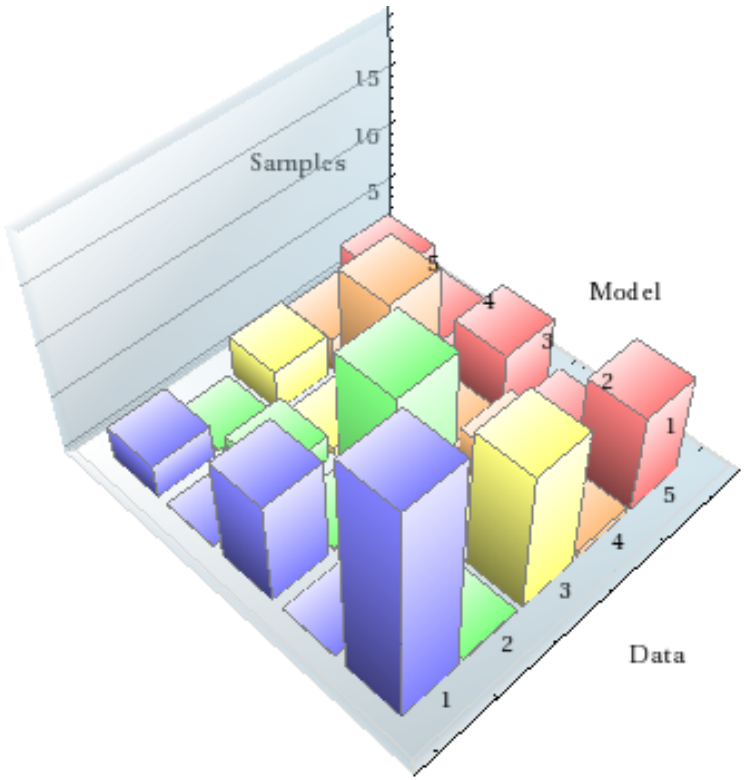
Approximated model



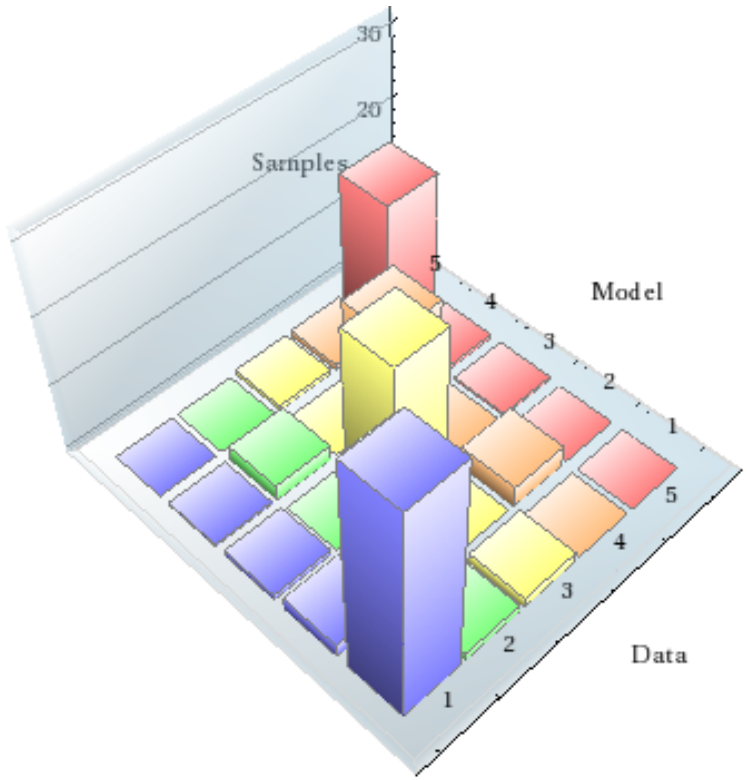
- only 5 parameters describing whole curve
- represent patterns

How many parameters do we need?

2 parameters



5 parameters



Which parameters to choose?

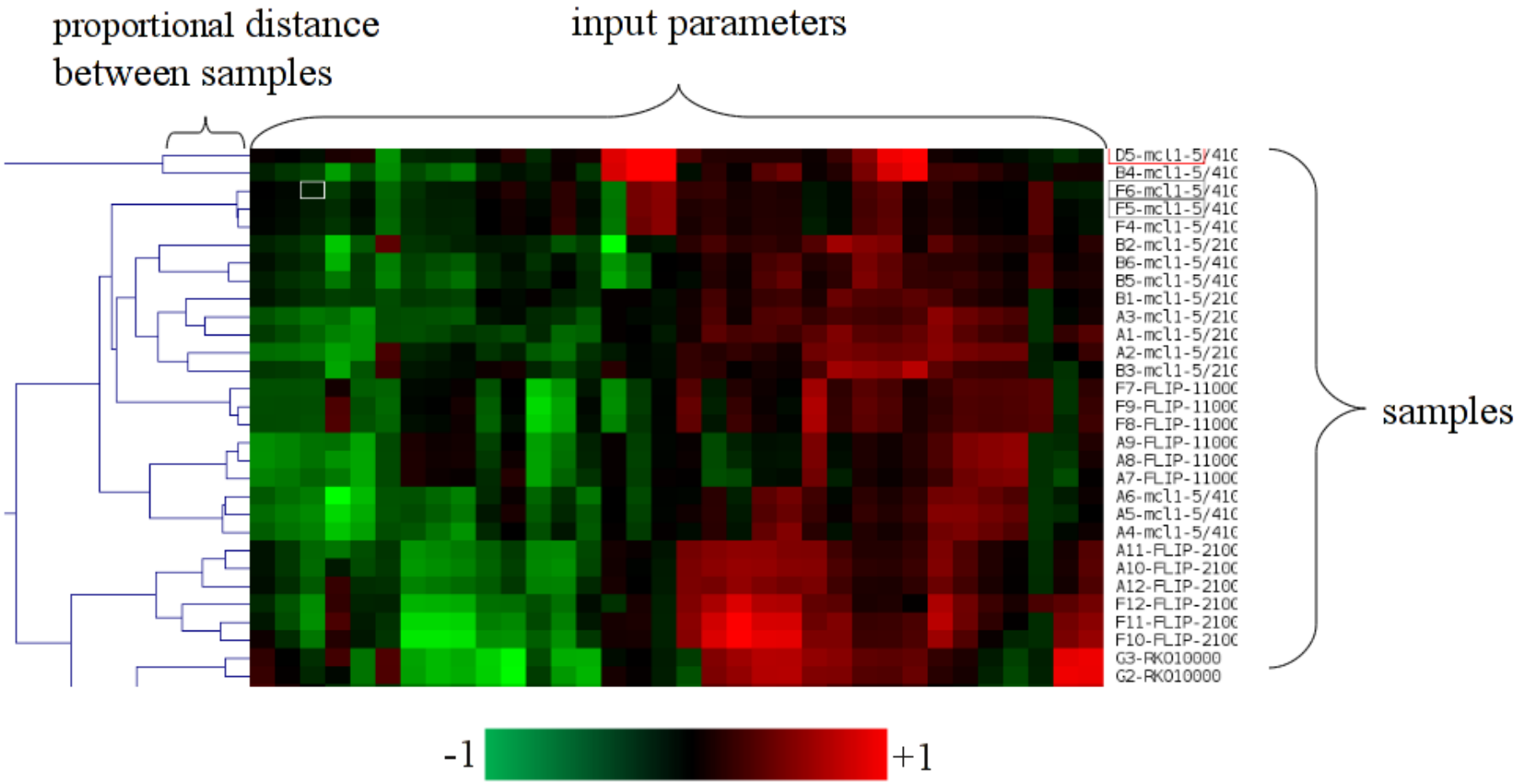
on previous slide input parameters were following:

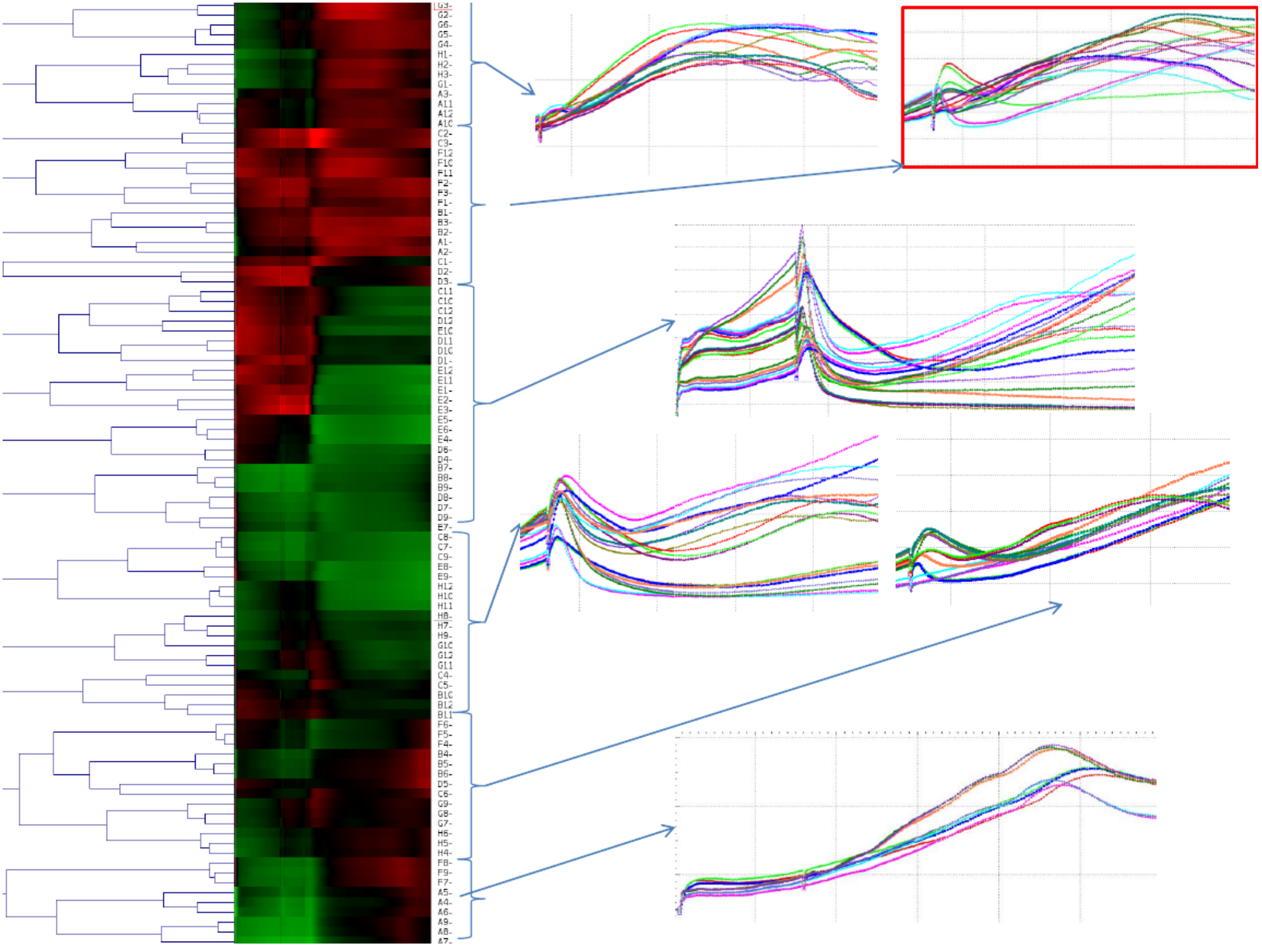
- mean
- minimum
- maximum
- linear coefficient
- quadratic coefficient

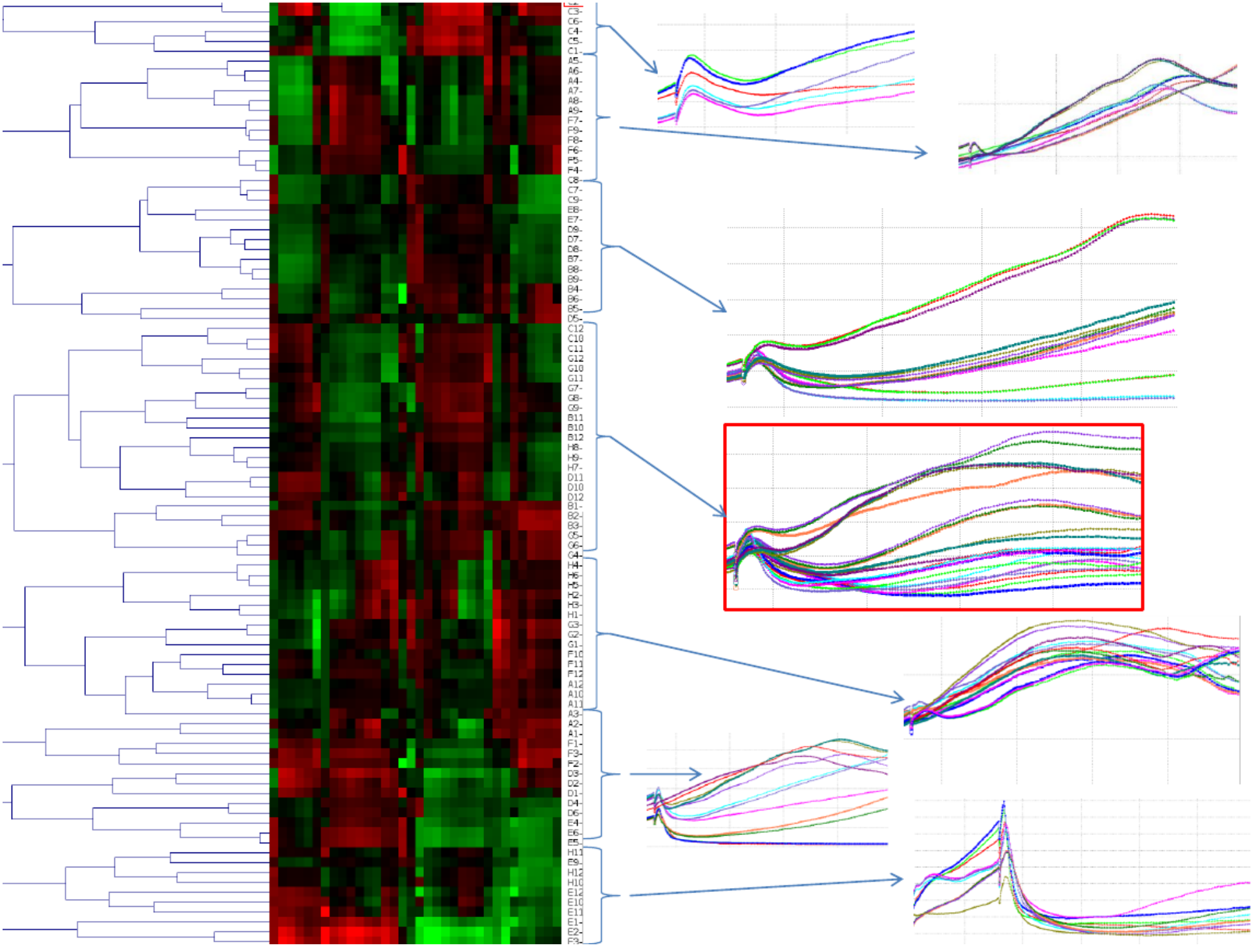
for EEG clustering is [Siuly et al., 2011] using:

- minimum
- maximum
- mean
- median
- modus
- first quartile
- third quartile
- inter-quartile range
- standard deviation

Dendrogram


proportional distance
between samples
input parameters
D5-mcl1-5/41C
B4-mcl1-5/41C
F6-mcl1-5/41C
F5-mcl1-5/41C
F4-mcl1-5/41C
B2-mcl1-5/21C
B6-mcl1-5/41C
B5-mcl1-5/41C
B1-mcl1-5/21C
A3-mcl1-5/21C
A1-mcl1-5/21C
A2-mcl1-5/21C
B3-mcl1-5/21C
F7-FLIP-1100C
F9-FLIP-1100C
F8-FLIP-1100C
A9-FLIP-1100C
A8-FLIP-1100C
A7-FLIP-1100C
A6-mcl1-5/41C
A5-mcl1-5/41C
A4-mcl1-5/41C
A11-FLIP-210C
A10-FLIP-210C
A12-FLIP-210C
F12-FLIP-210C
F11-FLIP-210C
F10-FLIP-210C
G3-RKO10000
G2-RKO10000
samples
-1
+1


G3-
G2-
G6-
G5-
G4-
H1-
H2-
H3-
G1-
A3-
A11
A12
A10
C2-
C3-
F12
F10
F11
F2-
F3-
F1-
B1-
B3-
B2-
A1-
A2-
C1-
D2-
D3-
C11
C10
C12
D12
E10
D11
D10
D1-
E12
E11
E1-
E2-
E3-
E5-
E6-
E4-
D6-
D4-
B7-
B8-
B9-
D8-
D7-
D9-
E7-
C8-
C7-
C9-
E8-
E9-
H12
H10
H11
H8-
H7-
H9-
G10
G12
G11
C4-
C5-
B10
B12
B11
F6-
F5-
F4-
B4-
B5-
B6-
D5-
C6-
G9-
G8-
G7-
H6-
H5-
H4-
F8-
F9-
F7-
A5-
A4-
A6-
A9-
A8-
A7-


C3-
C6-
C4-
C5-
C1-
A5-
A6-
A4-
A7-
A8-
A9-
F7-
F9-
F8-
F6-
F5-
F4-
C8-
C7-
C9-
E8-
E7-
D9-
D7-
D8-
B7-
B8-
B9-
B4-
B6-
B5-
D5-
C12
C10
C11
G12
G10
G11
G7-
G8-
G9-
B11
B10
B12
H8-
H9-
H7-
D11
D10
D12
B1-
B2-
B3-
G5-
G6-
G4-
H4-
H6-
H5-
H2-
H3-
H1-
G3-
G2-
G1-
F10
F11
F12
A12
A10
A11
A3-
A2-
A1-
F1-
F3-
F2-
D3-
D2-
D1-
D4-
D6-
E4-
E6-
E5-
H11
E9-
H12
H10
E12
E10
E11
E1-
E2-
E3-

C-Index

The C-index was reviewed in Hubert and Levin [1976]

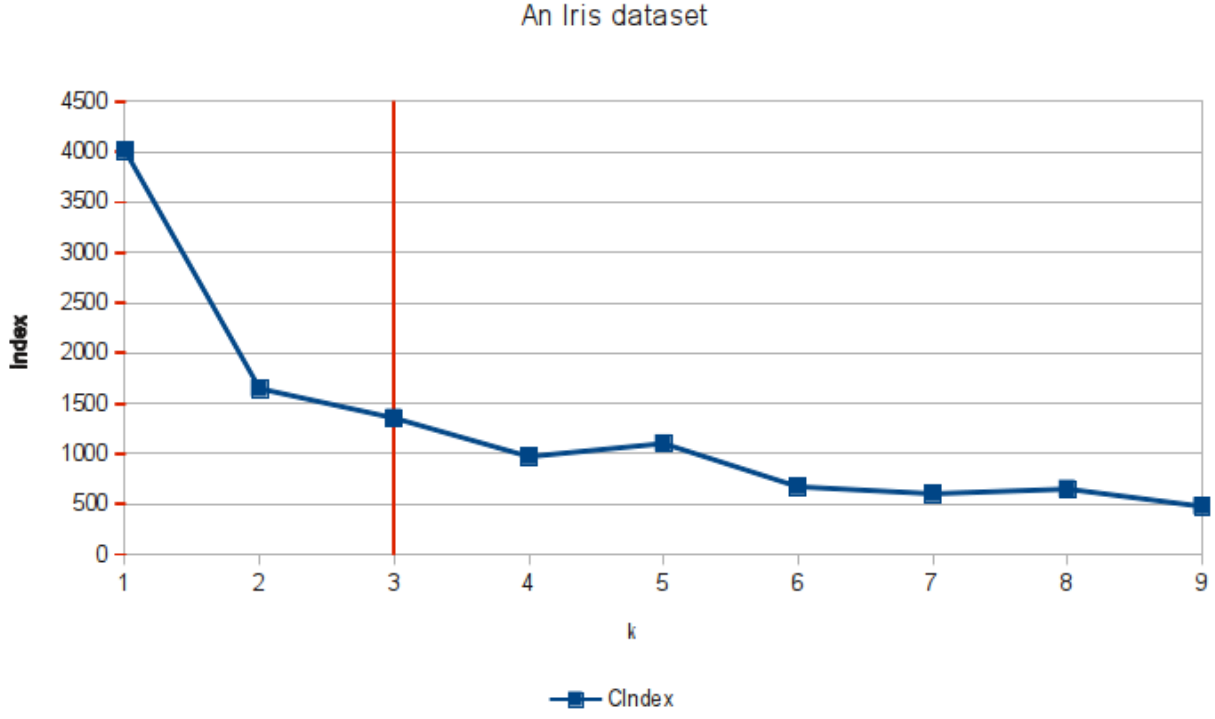
$$p_{c-index} = \frac{d_w - \min(d_w)}{\max(d_w) - \min(d_w)}$$

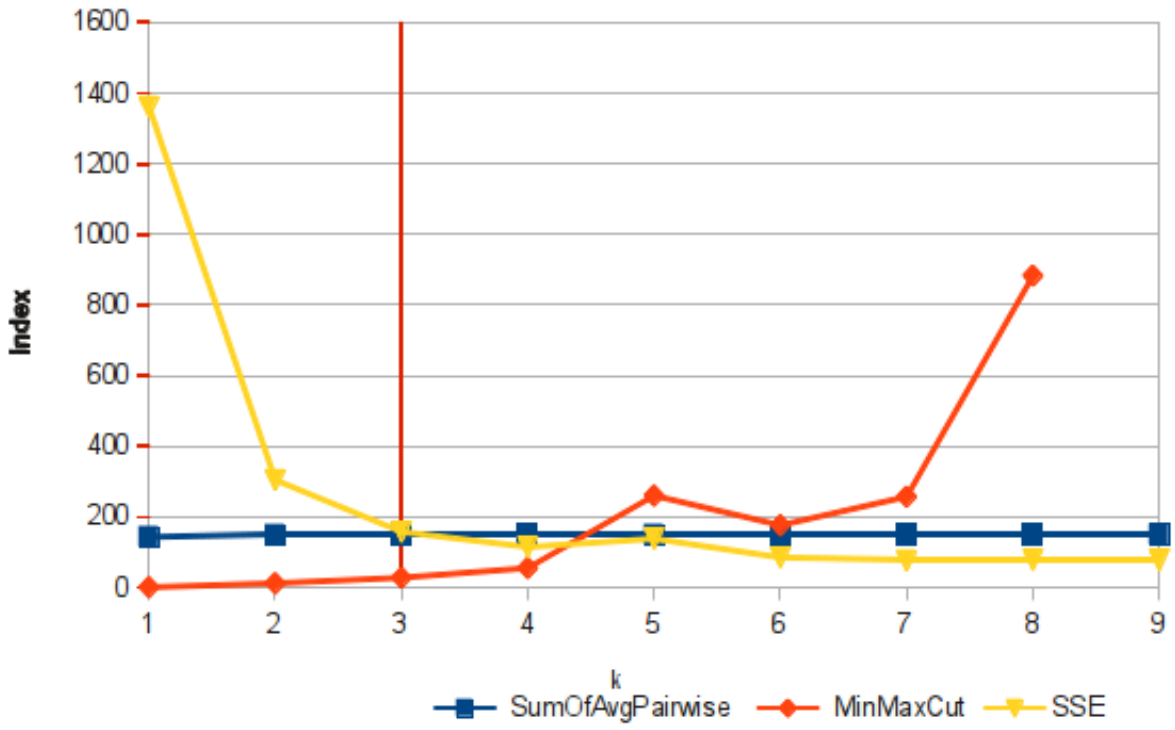
where d_w is the sum of the within cluster distances.

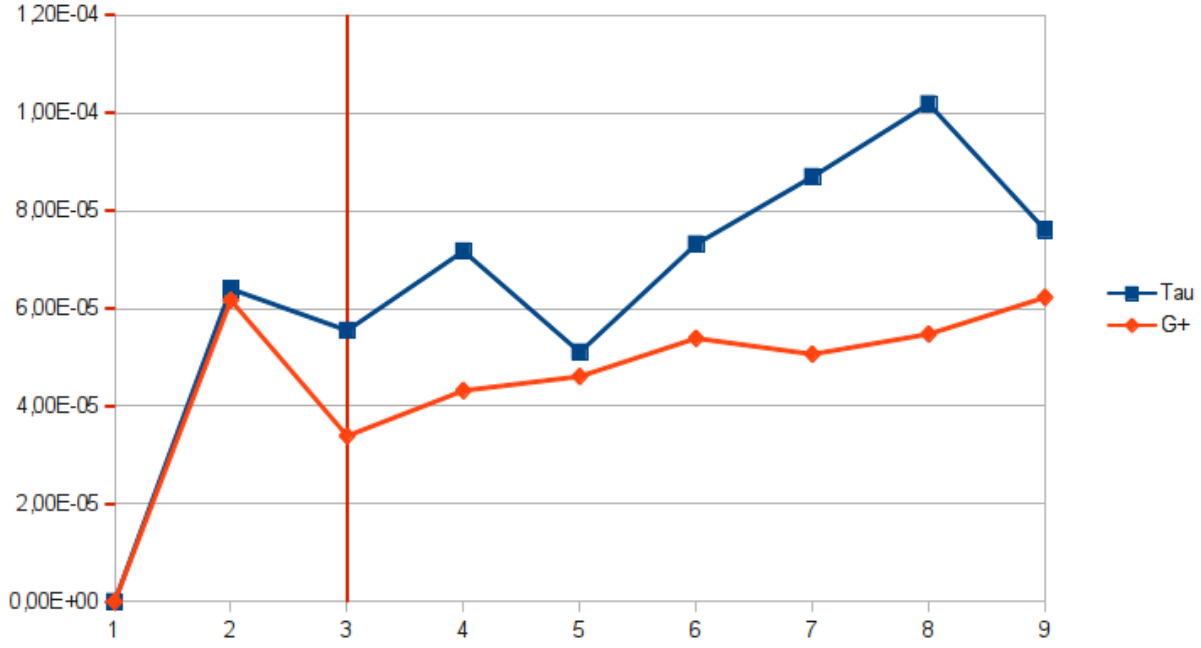
Gamma

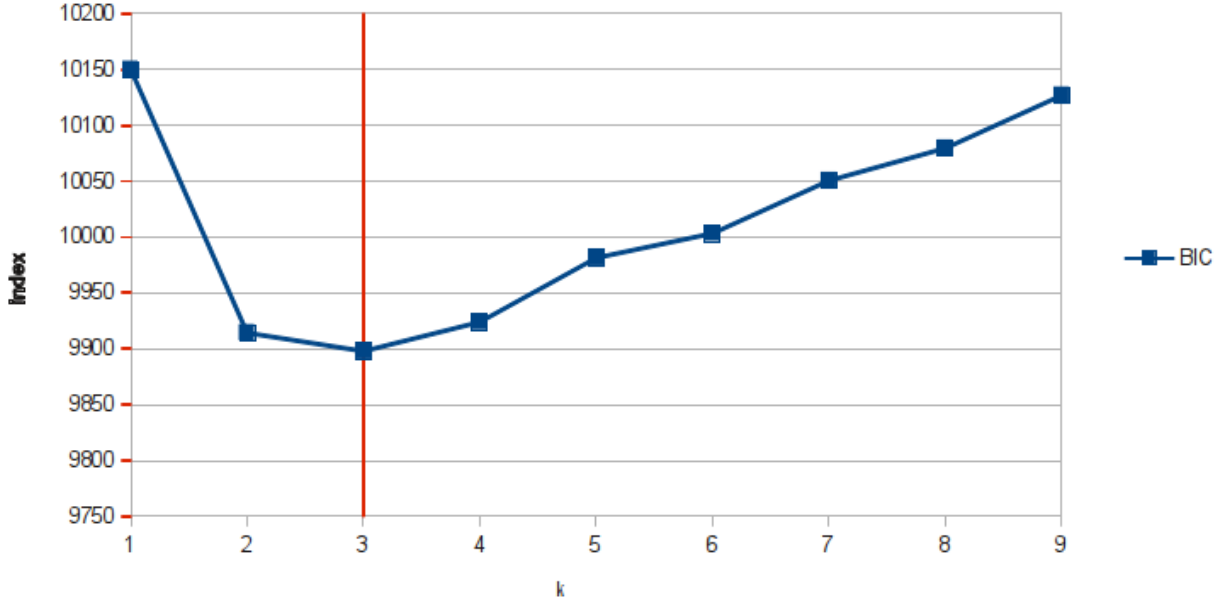
$$p_{gamma} = \frac{s(+) + s(-)}{s(+) - s(-)}$$

where $s(+)$ represents the number of consistent comparisons involving between and within cluster distances, and $s(-)$ represents the number of inconsistent outcomes Milligan and Cooper [1985]


An Iris dataset
4500
4000
3500
3000
2500
2000
1500
1000
500
0
Index
1
2
3
4
5
6
7
8
9
k
CIndex


1600
1400
1200
1000
800
600
400
200
0
Index
1
2
3
4
5
6
7
8
9
k
SumOfAvgPairwise
MinMaxCut
SSE


1,20E-04
1,00E-04
8,00E-05
6,00E-05
4,00E-05
2,00E-05
0,00E+00
1
2
3
4
5
6
7
8
9
Tau
G+


10200
10150
10100
10050
10000
9950
9900
9850
9800
9750
Index
1
2
3
4
5
6
7
8
9
k
BIC

The strive for objectivity, repeatability, testability etc. is perfectly right attitude as long as their proper place in the "hierarchy of aims" is maintained, but becomes very harmful if these tools dominate over the purpose of scientific research. [Holynski, 2005, p. 487]

Questions?

Thanks for your attention!

tomas.barton@fit.cvut.cz

References I

P. Arabie, L. J. Hubert, and G. D. Soete. Clustering and classification. World Scientific, 1996.

J.L. Borges. El idioma analítico de john wilkins. Obras completas, 1952.

Rich Caruana, Mohamed Elhawary, Nam Nguyen, and Casey Smith. Meta clustering. In Proceedings of the Sixth International Conference on Data Mining, ICDM '06, pages 107–118, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2701-9. doi: 10.1109/ICDM.2006.103. URL http://dx.doi.org/10.1109/ICDM.2006.103.

B. S. Everitt. Cluster Analysis. Edward Arnold, 1993.

References II

I. Guyon, U. Von Luxburg, and R.C. Williamson. Clustering: Science or art. In NIPS 2009 Workshop on Clustering Theory, 2009.

Roman B. Holynski. Philosophy of science from a taxonomist's perspective. Genus, 16(4):469–502, 2005.

L.J. Hubert and J.R. Levin. A general statistical framework for assessing categorical clustering in free recall. Psychological Bulletin, 83(6):1072, 1976.

Glenn W. Milligan and Martha C. Cooper. An examination of procedures for determining the number of clusters in a dataset. Psychometrika, 50(2):159–179, June 1985.

References III

G. Bel Mufti, P. Bertrand, and L. El Moubarki. Determining the number of groups from measures of cluster stability. In Proceedings of International Symposium on Applied Stochastic Models and Data Analysi, pages 404–412, 2005.

Siuly, Yan Li, and Peng (Paul) Wen. Clustering technique-based least square support vector machine for eeg signal classification. Computer Methods and Programs in Biomedicine, 104(3): 358–372, 2011. URL http://dblp.uni-trier.de/db/journals/cmpb/cmpb104.html#SiulyLW11;http://dx.doi.org/10.1016/j.cmpb.2010.11.014;http://www.bibsonomy.org/bibtex/2976ac83c8e51dd3ff108ee52da22902d/dblp.