UNIVERSITÀ DEGLI STUDI DI MILANO

Maximal Information Coefficient (MIC – search for relationships n a dataset)

Time series analysis and display

Visualizing Categorical data

Multivariate LABELED data visualization/analysis

1. Plot all the features (variables) to identify nonsenses (remove them).

2. Normalized the data (between 0-1 or to have zero mean and unitary std)

3. hypothesis testing for identifying “important” (discriminative) features
   - Continuous variables:
     - t-test for continuous variables (but you need to assume that the underlying distribution is normal)
     - Non-parametric tests if you can’t make any assumption (e.g. Mann-Whitney, Kruskal-Wallis)
   - Categorical data:
     - Fisher exact test (if you get few points)
     - Chi Square (χ^2) test otherwise

4. **For each variable**, and **for each class**, separate the point according to their labels and visualize the feature **density plots (histograms)** and or the boxplots (**with notches) of each class**.

   train decision trees on each feature to effectively assess the variable discrimination capability.

5. Compute pairwise correlations between:
   - each features and the data labels: features that are highly correlated with the label (**should** also have a low p-value) are the most discriminative/should have boxplots with NOT OVERLAPPED notches or different (not overlapping) per class histograms.

   - each features and each other feature: if two features are highly correlated, they are redundant! Remove the one with the highest p-value/highest accuracy/highest correlation with the labels

6. TSNE for reducing the data dimensionality and projecting the data in an (unrolled space) where points in the same class are near. Visualize the 2D data by using scatter plots (of the first 2/3 dimensions computed by TSNE)

4. Compute pairwise correlations between:
   - each features and the data labels: features that are highly correlated with the label (**should** also have a low p-value) are the most discriminative/should have boxplots with NOT
   - eac ... are redu ... rrelation with

Linear correlations with Pearson, Spearman,

Non linear correlations with MIC or statistics of MINE family

5. TSNE for ... ce) where points in ... rst 2/3 dimensions computed by TSNE)

Maximal Information Coefficient

Maximal Information-based Non-parametric Exploration

Generality:

with sufficient sample size the statistic should capture a wide range of interesting associations, not limited to specific function types (such as linear, exponential, or periodic), or even to all functional relationships.

Equitability:

the statistic should give similar scores to equally noisy relationships of different types.

Generability: not only do relationships take many functional forms, but many important relationships—for example, a superposition of functions (composition of functions) —are not well modeled by a unique function.

Equitability: need of giving similar scores to functional relationships with similar $\boldsymbol{R^2}$ values (given sufficient sample size)

coefficient of determination, denoted R^2 or r^2 and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

In 2D Suppose you have a dataset with n points $y_1,\ldots, y_i, \ldots, y_n$ (the dataset is the vector $\mathbf{y} = [y_1,\ldots, y_n]'$), and you fit it with a regression (predicted, fitted) model $f_1,\ldots,f_n$ (known as f_i, or sometimes $\hat{y}_i$, s a vector f).

R^2 is a statistic that will give some information about the goodness of fit of the model f to $\mathbf{y}$. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points. An R^2 of 1 indicates that the regression predictions perfectly fit the data.

Coefficient of determination

If $\bar{y}$ is the mean of the observed data:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

then the variability of the data set can be measured using three sums of squares formulas:

- The total sum of squares (proportional to the variance of the data):

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2,$$

- The regression sum of squares, also called the explained sum of squares:

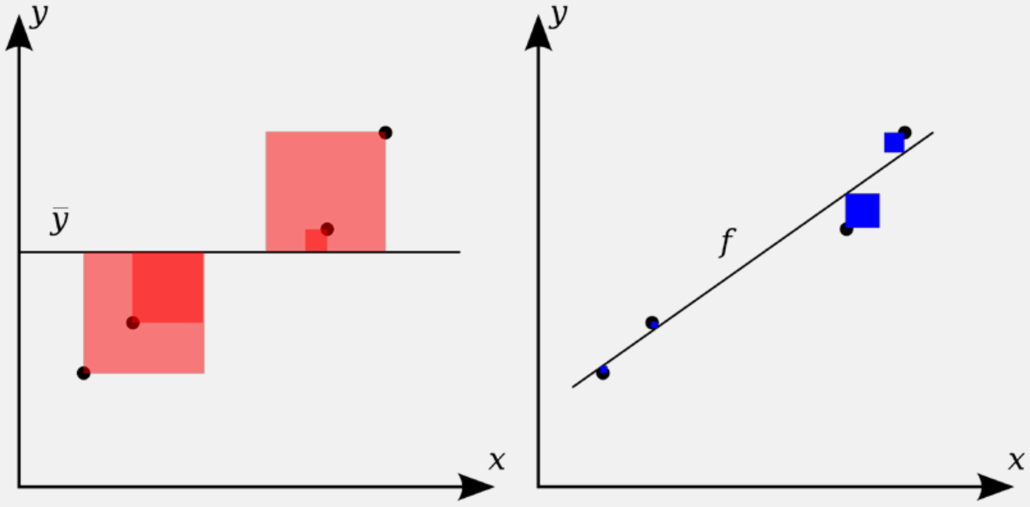
$$SS_{\text{reg}} = \sum_i (f_i - \bar{y})^2,$$

- The sum of squares of residuals, also called the residual sum of squares:

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

The most general definition of the coefficient of determination is

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$



$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

The better the linear regression (on the right) fits the data in comparison to the simple average (on the left graph), the closer the value o R^2 is to 1. The areas of the blue squares represent the squared residuals with respect to the linear regression. The areas of the red squares represent the squared residuals with respect to the average value.

The Idea at the base of MIC:

if a relationship exists between two variables, then a grid can be drawn on the scatterplot of the two variables that partitions the data to encapsulate that relationship.

Therefore if we try all the grids and find a well-fitting grid, the relationship may be estimated in terms of the grid "coverage".

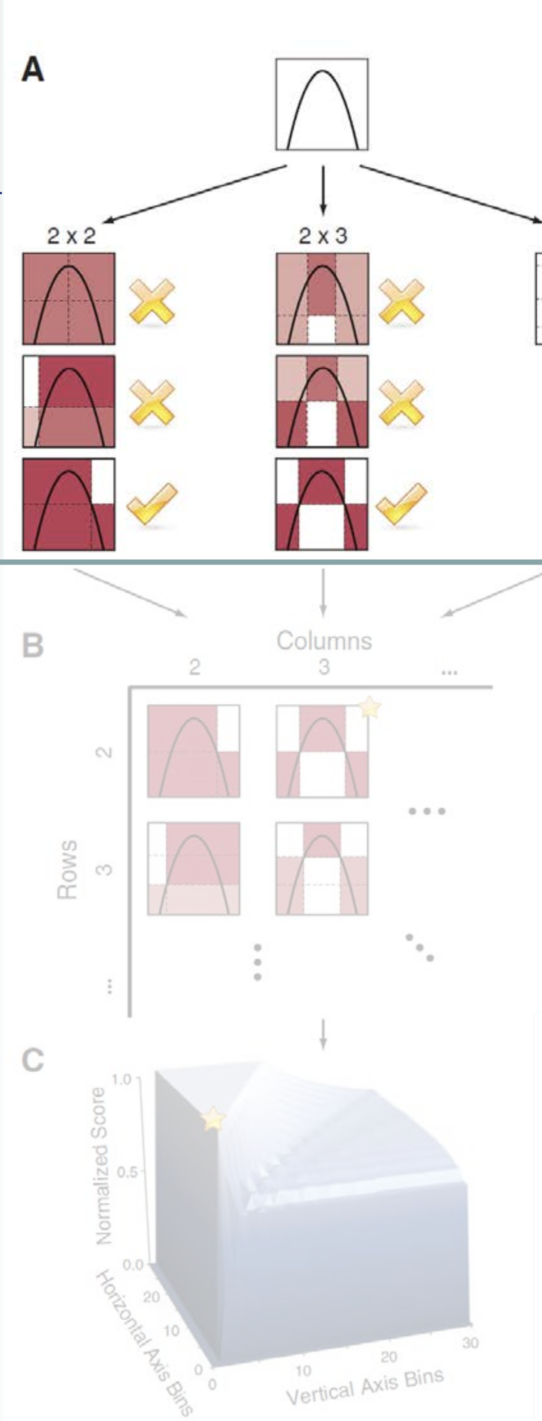


Fig. 1. Computing MIC (A) For each pair (x,y), the MIC algorithm finds the x-by-y grid with the highest induced mutual information. (B) The algorithm normalizes the mutual information scores and compiles a matrix that stores, for each resolution, the best grid at that resolution and its normalized score. (C) The normalized scores form the characteristic matrix, which can be visualized as a surface; MIC corresponds to the highest point on this surface. In this example, there are many grids that achieve the highest score. The star in (B) marks a sample grid achieving this score, and the star in (C) marks that grid's corresponding location on the surface.

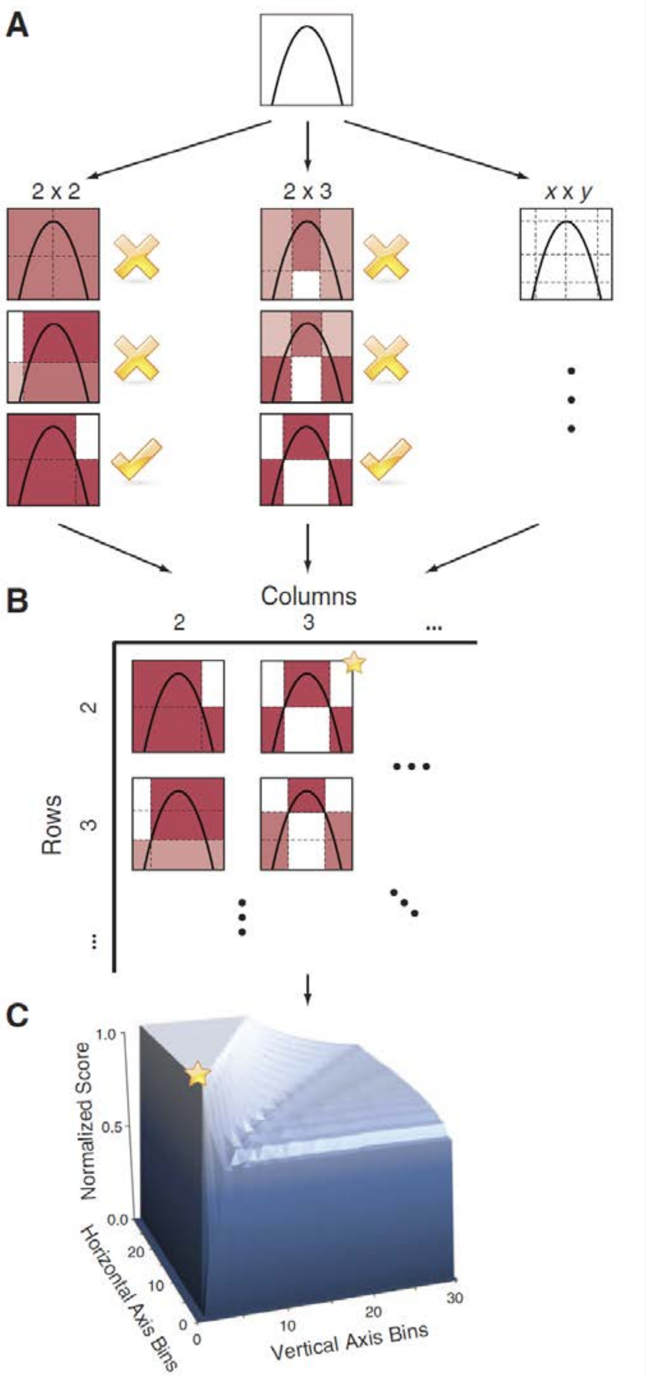
Thus, to calculate the MIC of a set of two-variable data set D:

explore all grids (x, y) [Footnote1] up to a maximal grid size **B(n),** where B(n) depends on the sample size[Footnote2].

From D compute the **characteristic matrix** $\mathbf{M(D)_{x,y}}$ with B(n)*B(n) components as follows.

Given $\mathbf{r < B(n)}$ and $\mathbf{c < B(n)}$:

- Define all the possible grids $\boldsymbol{g_{r,c}} = \mathbf{grid(r, c)}$ that split the image into **r** rows and **c** columns.
- For each of such grids $\boldsymbol{\hat{g}_{r,c}}$ compute its “coverage of the dataset” as the mutual information between the grid and the dataset. $\boldsymbol{mi(\hat{g}_{x,y}, D)}$
- Compute the maximum of the mutual informations on grids **r, c** $\mathbf{m_{x,y}} = \mathbf{max}(\boldsymbol{mi(\hat{g}_{x,y}, D)})$
- $\mathbf{M(D)_{x,y}} = \dfrac{\mathbf{m_{x,y}}}{\mathbf{log(min(x,y))}}$ [Footnote3]



[Footnote1] an (x, y) grid splits the plot into x rows and y columns (x*y rectangles)

[Footnote2] The finest grid (x_{max}, y_{max}) has x_{max}, $y_{max} < B(n) = n^{0.6} = (n^3)^{0.2}$

[Footnote3] normalization factor = log(min(x,y))

Once $M(D)_{x,y}$ has been computed you may compute the MINE statistics (all such that $0 \leq MINE \leq 1$)

$MIC(D) = \max_{x,y<B(n)} \{ M(D)_{x,y} \}$

$MAS(D)$, $MEV(D)$, $MCN(D)$

Before briefly looking at them, how is $\boldsymbol{mi(\hat{g}_{x,y}, D)}$ computed?

Mutual Information Coefficient

$$I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} = E_{P_{XY}} \log \frac{P_{XY}}{P_X P_Y}$$

Mutual Information Coefficient

$$I(X;Y) = \sum_{x,y} \boxed{P_{XY}(x,y)} \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} = E_{P_{XY}} \log \frac{P_{XY}}{P_X P_Y}$$

Number of points that fall inside the box (x,y) divided by the area of the box (x,y)

Mutual Information Coefficient

$$I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{\boxed{P_X(x)} P_Y(y)} = E_{P_{XY}} \log \frac{P_{XY}}{P_X P_Y}$$

Number of points that fall inside the boxes in row x divided by the area of the boxes in row x

Mutual Information Coefficient

$$I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)\boxed{P_Y(y)}} = E_{P_{XY}} \log \frac{P_{XY}}{P_X P_Y}$$

Number of points that fall inside the boxes in column y divided by the area of the boxes in column y

Mutual Information Coefficient

$$I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} = \boxed{E_{P_{XY}}} \log \frac{P_{XY}}{P_X P_Y}$$

Expected value of P_{XY}

Mutual Information interpretation through entropy:

$$H(X) = -\sum_x P_X(x)\log P_X(x) = -E_{P_X}\log P_X$$

entropy is a measure of "uncertainty" – the higher the entropy, the more uncertain one is about a random variable.

$$H(X|Y) = \sum_y P_Y(y)\left[-\sum_x P_{X|Y}(x|y)\log\left(P_{X|Y}(x|y)\right)\right] = E_{P_Y}\left[-E_{P_{X|Y}}\log P_{X|Y}\right]$$

The conditional entropy is the average uncertainty about X after observing a second random variable Y

- It should be maximal when $P_X(x)$ is uniform, and in this case it should increase with the number of possible values X can take;
- It should remain the same if we reorder the probabilities assigned to different values of X;
- The uncertainty about two independent random variables should be the sum of the uncertainties about each of them.

$$I(X;Y) = H(X) - H(X|Y).$$

Mutual information is the *reduction* in uncertainty about variable X after observing Y

Once $M(D)_{x,y}$ has been computed you may compute the MINE statistics

Existing relationship

$MIC(D) = \max_{x,y<B(n)} \{ M(D)_{x,y} \}$ $\qquad$ $0 \leq MIC(D) \leq 1$

From MIC to MINE statistics Family

Maximal Information-based Nonparametric Exploration

Existing relationship

$MIC(D) = \max_{x,y<B(n)} \{ M(D)_{x,y} \}$ $0 \leq MIC(D) \leq 1$

Non-monotonicity of the relationship

$MAS(D) = \max_{x,y<B(n)} \{ | M(D)_{x,y} - M(D)_{y,x} | \}$

Maximum Asymmetry Score, $0 \leq MAS \leq MIC \leq 1$

MAS checks how not symmetric is $M(D)_{y,x}$

Since $M(D)_{y,x}$ is symmetric for monotonic relationships,

$\rightarrow$ **MAS is higher for highly non monotonic relationships**

Once $M(D)_{x,y}$ has been computed you may compute the MINE statistics (all such that $0 \leq MINE \leq 1$)

$MIC(D) = \max_{x,y<B(n)} \{ M(D)_{x,y} \}$

Existing relationship

$MAS(D) = \max_{x,y<B(n)} \{ | M(D)_{x,y} - M(D)_{y,x} | \}$

Non-monotonicity of the relationship

Closeness of the relationship to a function

$MEV(D) = \max_{x,y<B(n)} \{ M(D)_{x,y} \}: x=2, y=2\}$

Maximum Edge Value, $0 \leq MEV \leq MIC \leq 1$

Measures the degree to which the dataset appears to be sampled from a continuous function.

If D passes the "vertical/horizontal" line tests (each vertical or horizontal lines contain only one point of D), then the maximal grids are those for x = 2, y=2.

→ **MEV is higher for Datasets distributed along continuous functions.**

Once $M(D)_{x,y}$ has been computed you may compute the MINE statistics (all such that $0 \le MINE \le 1$)

$$MIC(D) = \max_{x,y<B(n)} \{ M(D)_{x,y} \}$$

Existing relationship

$$MAS(D) = \max_{x,y<B(n)} \{ | M(D)_{x,y} - M(D)_{y,x} | \}$$

Non-monotonicity of the relationship

$$MEV(D) = \max_{x,y<B(n)} \{ M(D)_{x,y} \}: x=2, y=2\}$$

Closeness of the relationship to a function

Complexity of the relationship

$$MCN(D, \varepsilon) = \min_{x,y<B(n)} \{\log(x,y): M(D)_{x,y} \ge (1-\varepsilon)\, MIC(D)$$

Minimum Cell Number, $MIC \le MAS \le MIC \le 1$

Measures the scale of the grids which allow approximating the MIC score (ε controls the level of noise: use higher values of ε for noisy datasets). The highest x and y (the smallest the grid boxes), the highest the complexity of the relationship.

→ **MCN is higher for complex relationships.**

Once $M(D)_{x,y}$ has been computed you may compute the MINE statistics (all such that $0 \leq MINE \leq 1$)

$MIC(D) = \max_{x,y<B(n)} \{ M(D)_{x,y} \}$ **Existing relationship**

$MAS(D) = \max_{x,y<B(n)} \{ | M(D)_{x,y} - M(D)_{y,x} | \}$ **Non-monotonicity of the relationship**

$MEV(D) = \max_{x,y<B(n)} \{ M(D)_{x,y} \}: x=2, y=2\}$ **Closeness of the relationship to a function**

$MCN(D, \varepsilon) = \min_{x,y<B(n)} \{\log(x,y): M(D)_{x,y} \geq (1-\varepsilon) MIC(D)$ **Complexity of the relationship**

Existence of a relationship with power against independence

$TIC(D) = \Sigma_{x,y<B(n)} \{ M(D)_{x,y} \}$

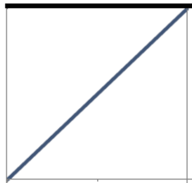
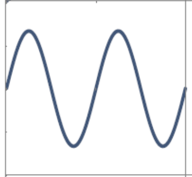
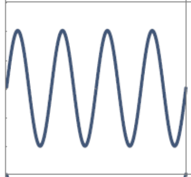
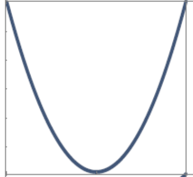
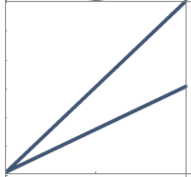
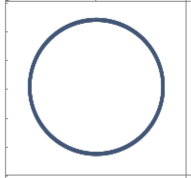
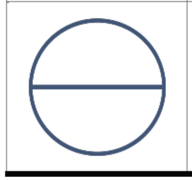
Total Information Coefficient, $MIC \leq 1 \leq TIC$.

While MIC is equitable, TIC achieves power against independence.

Authors suggest to combine MIC and TIC to achieve

- power against independence (by filtering results using TIC)
- equitability (by using MIC on the remaining variable pairs)

when exploring a data set with a large number of nontrivial relationships.

Data	MIC	MAS	MEV	MCN
	1.00	0.00	1.00	2.00
	1.00	0.74	1.00	3.00
	1.00	0.89	1.00	4.00
	1.00	0.69	1.00	2.56
	0.79	0.16	0.70	6.91
	0.71	0.03	0.32	6.87
	0.46	0.19	0.22	6.98

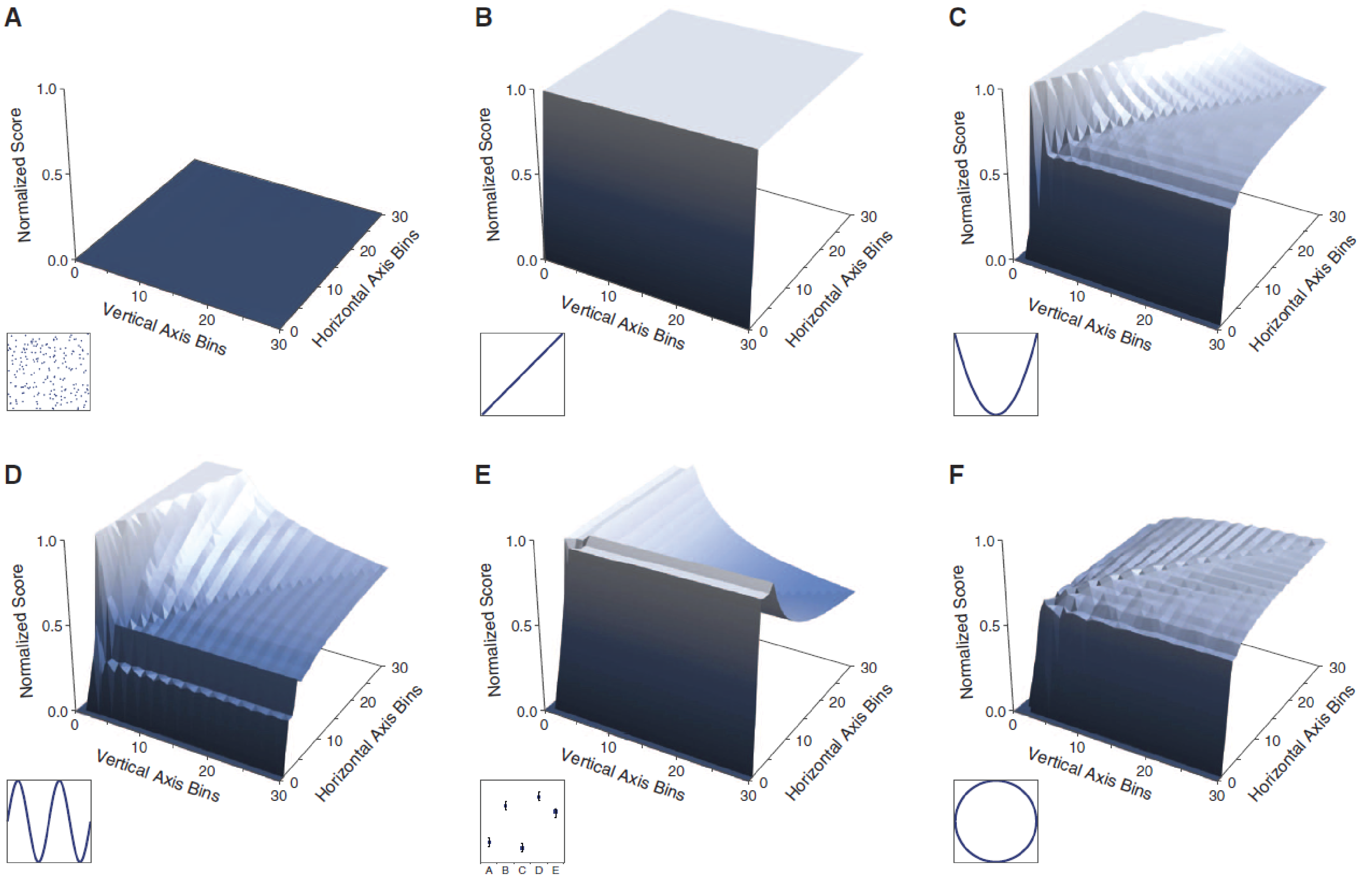
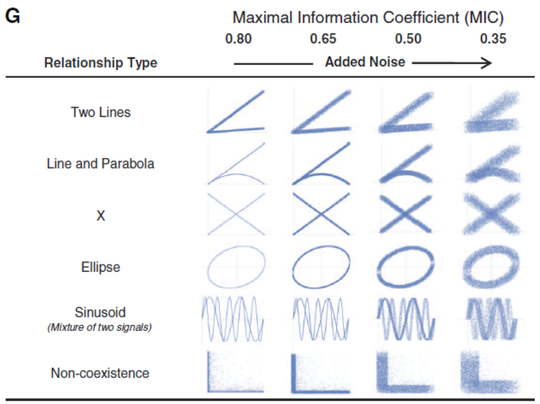


Fig. 3. Visualizations of the characteristic matrices of common relationships. (**A** to **F**) Surfaces representing the characteristic matrices of several common relationship types. For each surface, the *x* axis represents number of vertical axis bins (rows), the *y* axis represents number of horizontal axis bins (columns), and the *z* axis represents the normalized score of the best-performing grid with those dimensions. The inset plots show the relationships used to generate each surface. For surfaces of additional relationships, see fig. S7.

A

Relationship Type	MIC	Pearson	Spearman	Mutual Information (KDE)	Mutual Information (Kraskov)	CorGC (Principal Curve-Based)	Maximal Correlation
Random	0.18	-0.02	-0.02	0.01	0.03	0.19	0.01
Linear	1.00	1.00	1.00	5.03	3.89	1.00	1.00
Cubic	1.00	0.61	0.69	3.09	3.12	0.98	1.00
Exponential	1.00	0.70	1.00	2.09	3.62	0.94	1.00
Sinusoidal (Fourier frequency)	1.00	-0.09	-0.09	0.01	-0.11	0.36	0.64
Categorical	1.00	0.53	0.49	2.22	1.65	1.00	1.00
Periodic/Linear	1.00	0.33	0.31	0.69	0.45	0.49	0.91
Parabolic	1.00	-0.01	-0.01	3.33	3.15	1.00	1.00
Sinusoidal (non-Fourier frequency)	1.00	0.00	0.00	0.01	0.20	0.40	0.80
Sinusoidal (varying frequency)	1.00	-0.11	-0.11	0.02	0.06	0.38	0.76



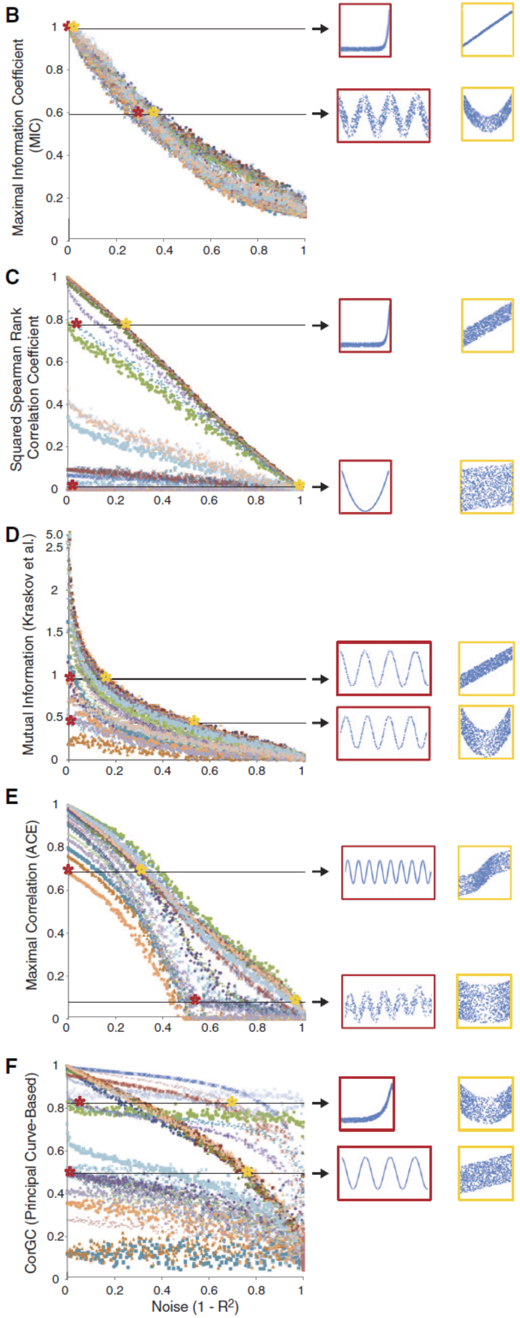


Fig. 2. Comparison of MIC to existing methods (**A**) Scores given to various noiseless functional relationships by several different statistics (*8*, *12*, *14*, *19*). Maximal scores in each column are accentuated. (**B** to **F**) The MIC, Spearman correlation coefficient, mutual information (Kraskov *et al.* estimator), maximal correlation (estimated using ACE), and the principal curve-based CorGC dependence measure, respectively, of 27 different functional relationships with independent uniform vertical noise added, as the R^2 value of the data relative to the noiseless function varies. Each shape and color corresponds to a different combination of function type and sample size. In each plot, pairs of thumbnails show relationships that received identical scores; for data exploration, we would like these pairs to have similar noise levels. For a list of the functions and sample sizes in these graphs as well as versions with other statistics, sample sizes, and noise models, see figs. S3 and S4. (**G**) Performance of MIC on associations not well modeled by a function, as noise level varies. For the performance of other statistics, see figs. S5 and S6.

Once you have computed the characteristics matrix $M(D)_{x,y}$

- **Non-monotonicity**

  The Maximum Asymmetry Score (MAS) is defined by

  $$\mathrm{MAS}(D) = \max_{xy<B} |M(D)_{x,y} - M(D)_{y,x}|$$

  and measures deviation from monotonicity. MAS is never greater than MIC. For an illustration of the intuition behind MAS, see Figure S2.

- **Closeness to being a function**

  The Maximum Edge Value (MEV) is defined by

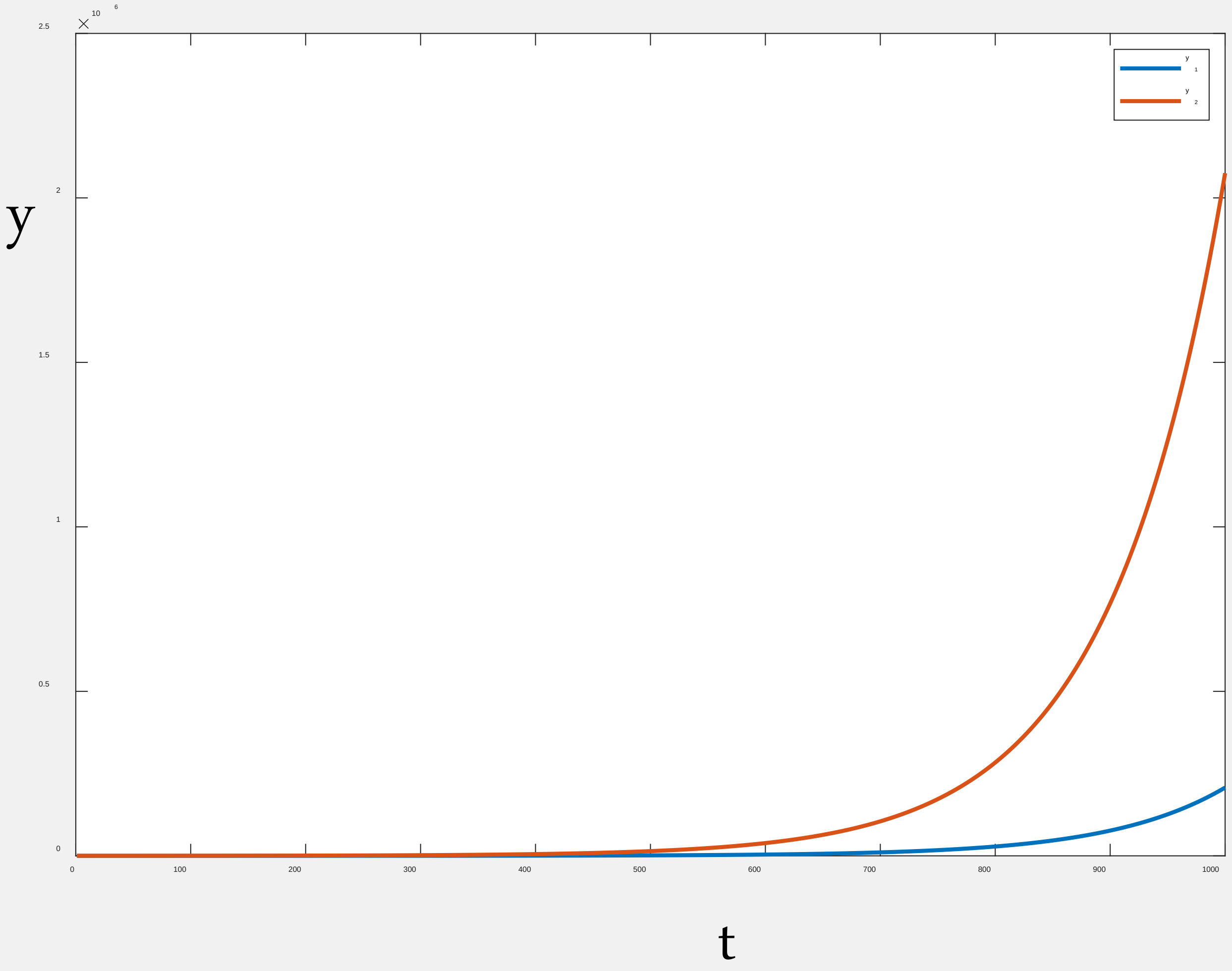
  $$\mathrm{MEV}(D) = \max_{xy<B} \{M(D)_{x,y} : x = 2 \text{ or } y = 2\}$$

PAUSA??

COMPARING DISTRIBUTION TRENDS…

Sometimes you need to change your mindset

Suppose the rate of change for each of the two functions is constant



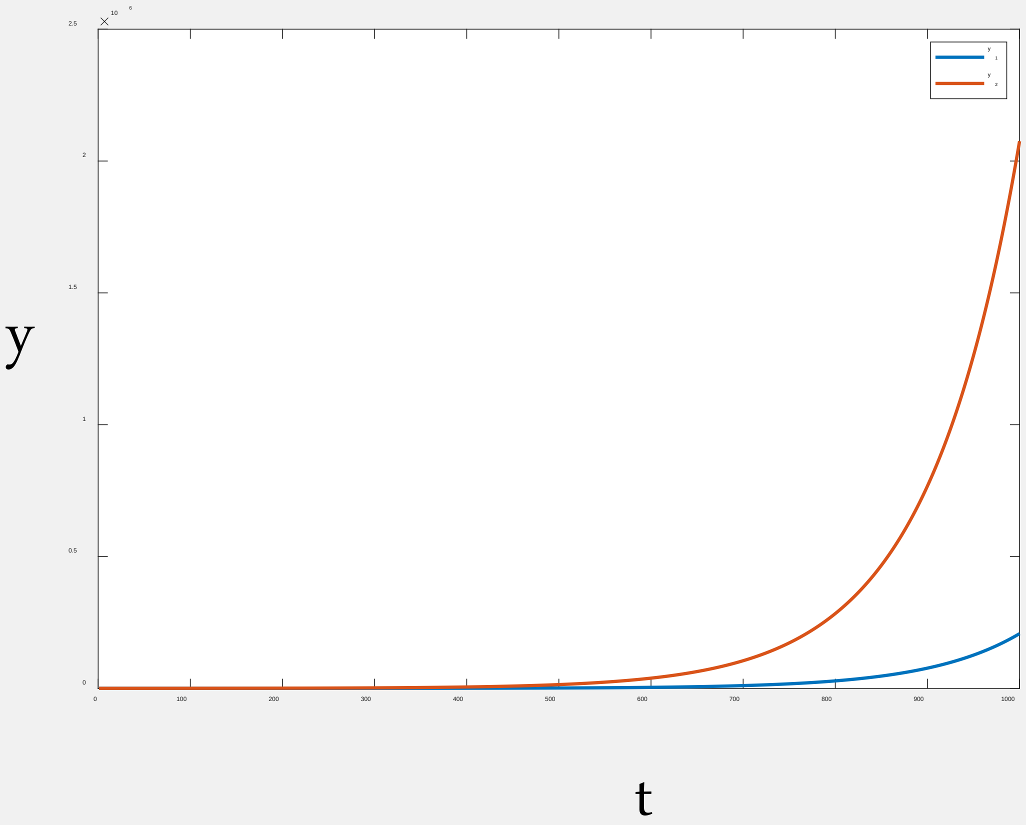
Is the rate of change similar?

What about using logarithmic scales?

$$y_2(t+1) = y_2(t) + y_2(t)^{*}rate_2$$

$$y_1(t+1) = y_1(t) + y_1(t)^{*}rate_1$$

$$rate_1 <= rate_2$$



$y_1(0) = v$ $\quad$ $Log\ (y_1(0)) = Log(\ v\ )$

$y_1(1) = v+v^*\ rate_1$ $\quad$ $Log(y_1(1)) = Log(v\ (1+rate_1)) = Log(\ v\ )+ Log(1+rate_1)$

$y_1(2) = y_1(1) + y_1(1)^*rate_1$ $\quad$ $Log(y_1(2)) = Log(y_1(1)\ )+ Log(1+rate_1) = Log(v) + 2^*\ Log(1+rate_1)$

$y_1(3) = y_1(2) + y_1(2)^*rate_1$ $\quad$ $Log(y_1(3)) = Log(y_1(2)\ )+ Log(1+rate_1) = Log(v) + 3^*\ Log(1+rate_1)$

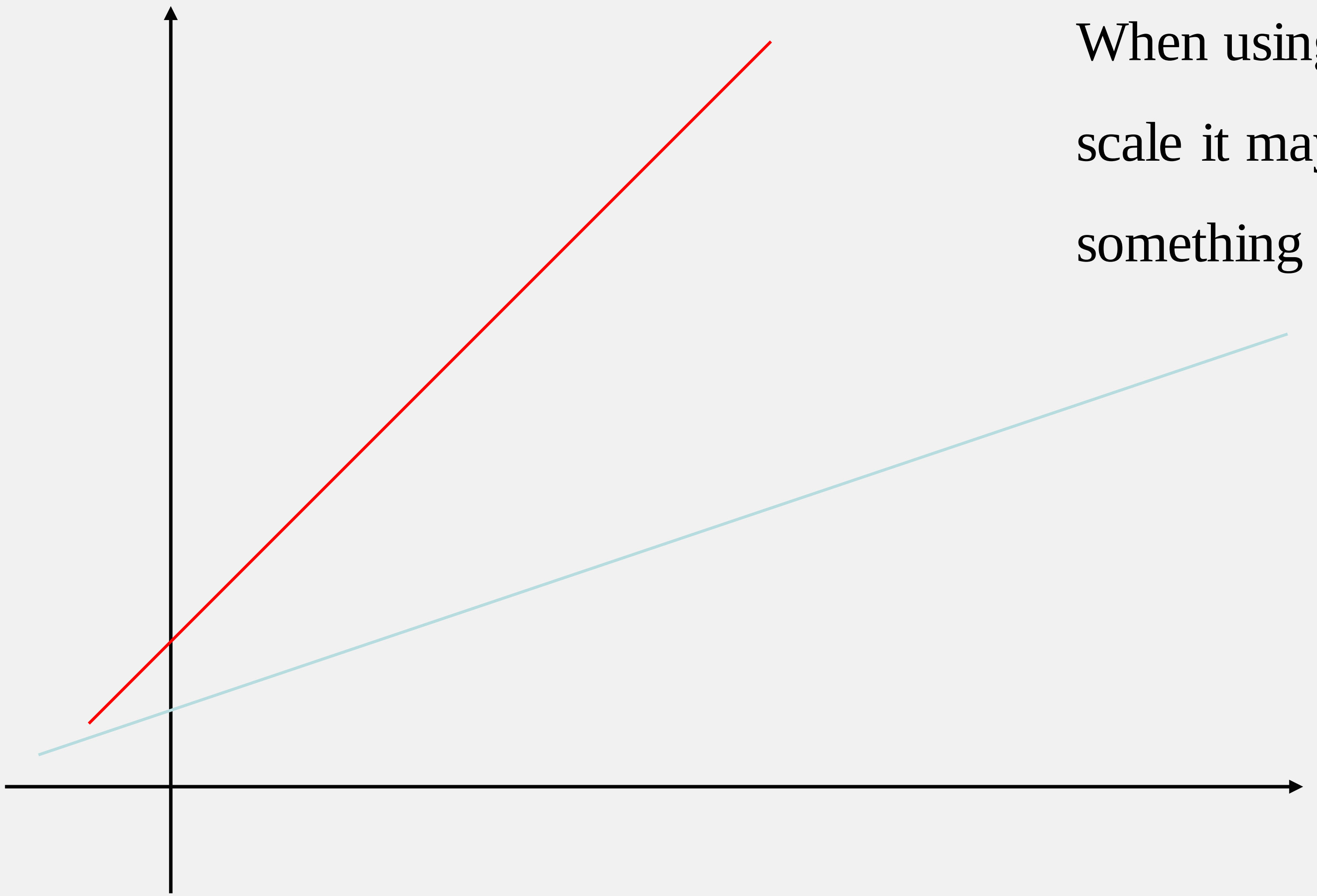
…

$Log(y_1(t+1)) = Log(\ y_1(t) + y_1(t)^*rate_1\ ) = Log(\ y_1(t)\ (1+rate_1\ )\ ) = Log(y_1(t)\ ) + Log(1+rate_2)$ $\mathbf{= Log(\ v\ ) + t^*Log(1+rate_1)}$

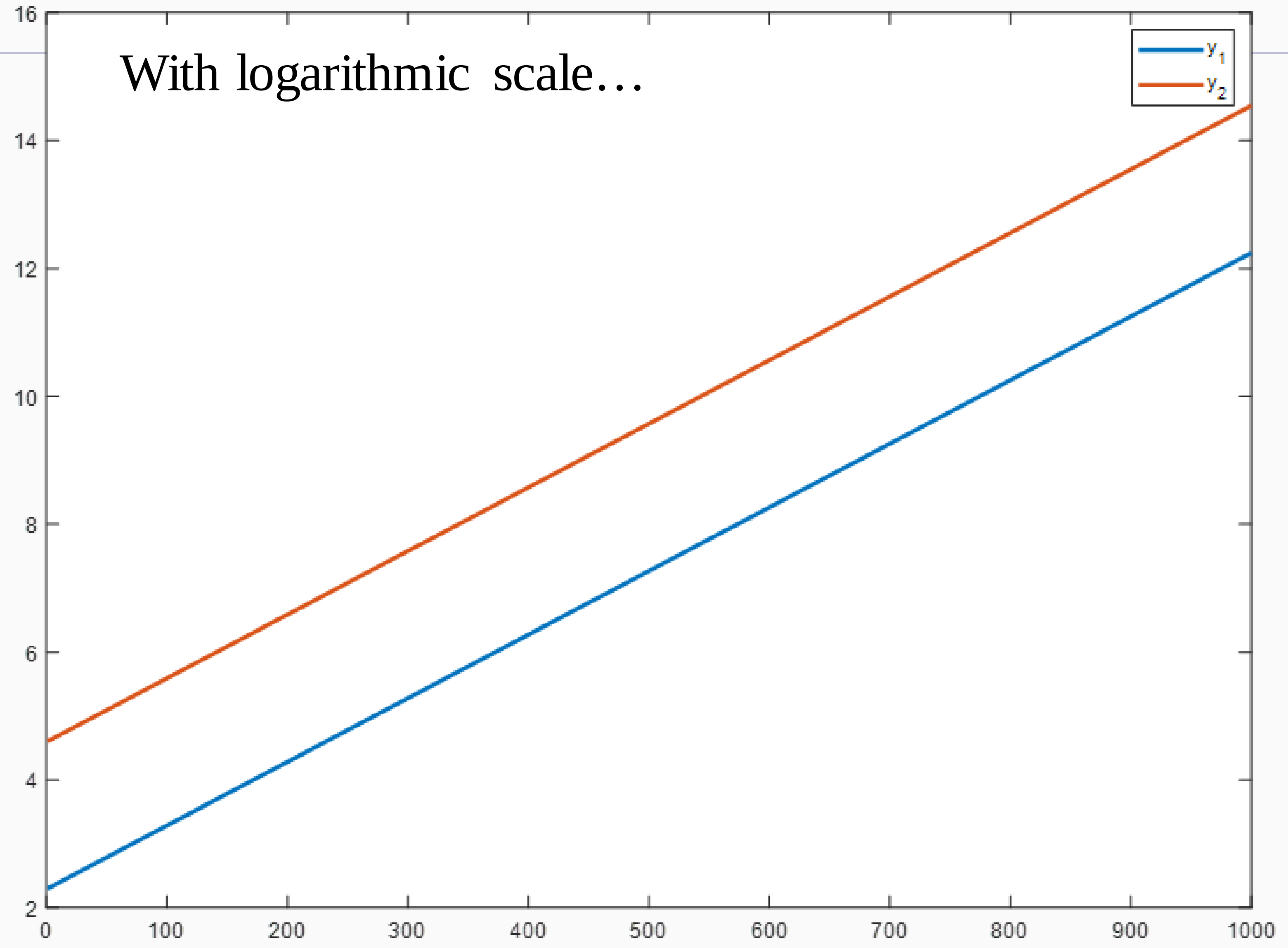
$$\mathbf{Log(y_1(0)\ ) + t^*Log(1+rate_1)}$$

Since $rate_1$ is constant and $Log(y_1(0))$ is also constant we have a line with m =

$Log(1+rate_1)$ and intercept $Log(y_1(0))$

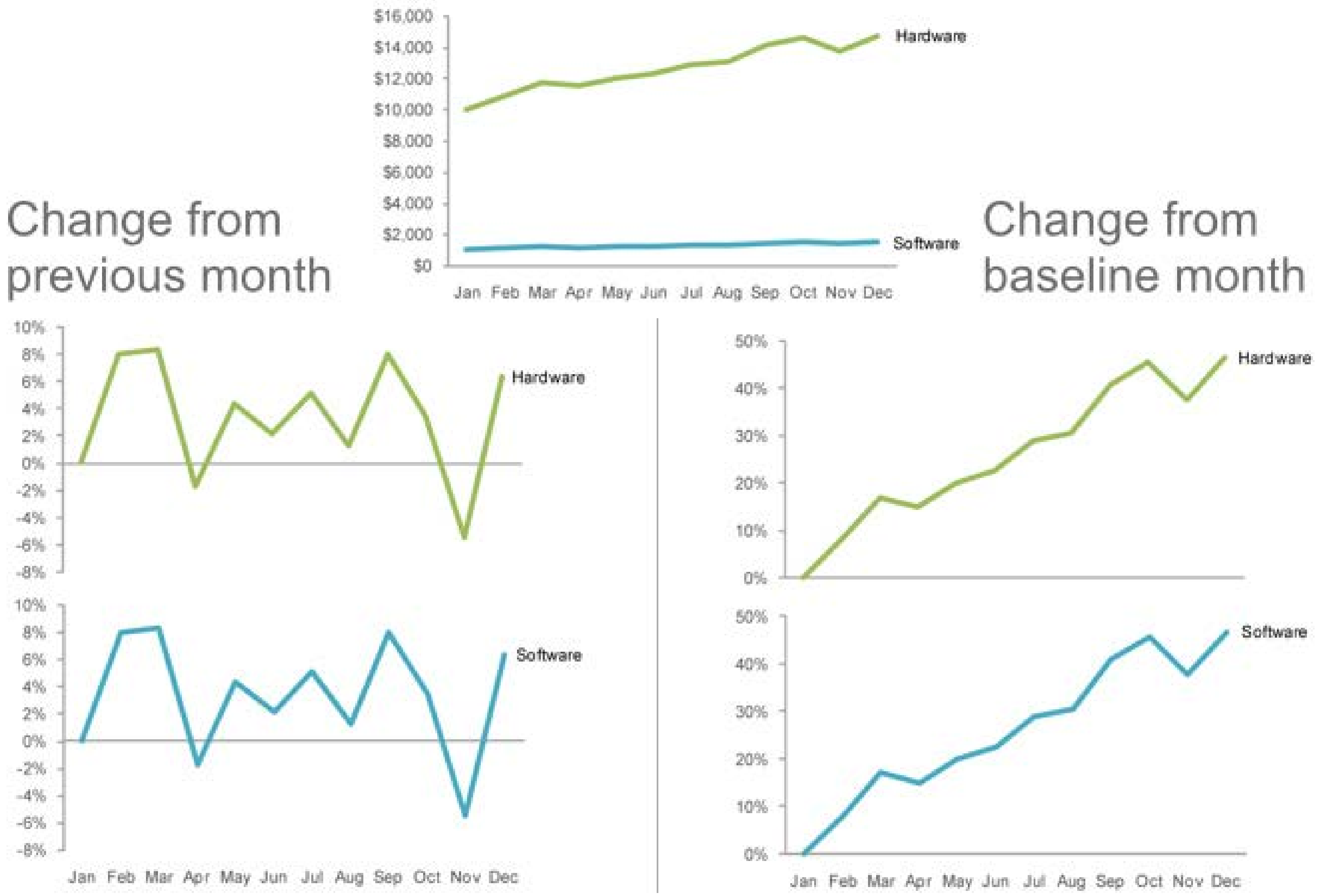


When using logarithmic scale it may be something like that



With logarithmic scale…
y1
y2
16
14
12
10
8
6
4
2
0
100
200
300
400
500
600
700
800
900
1000

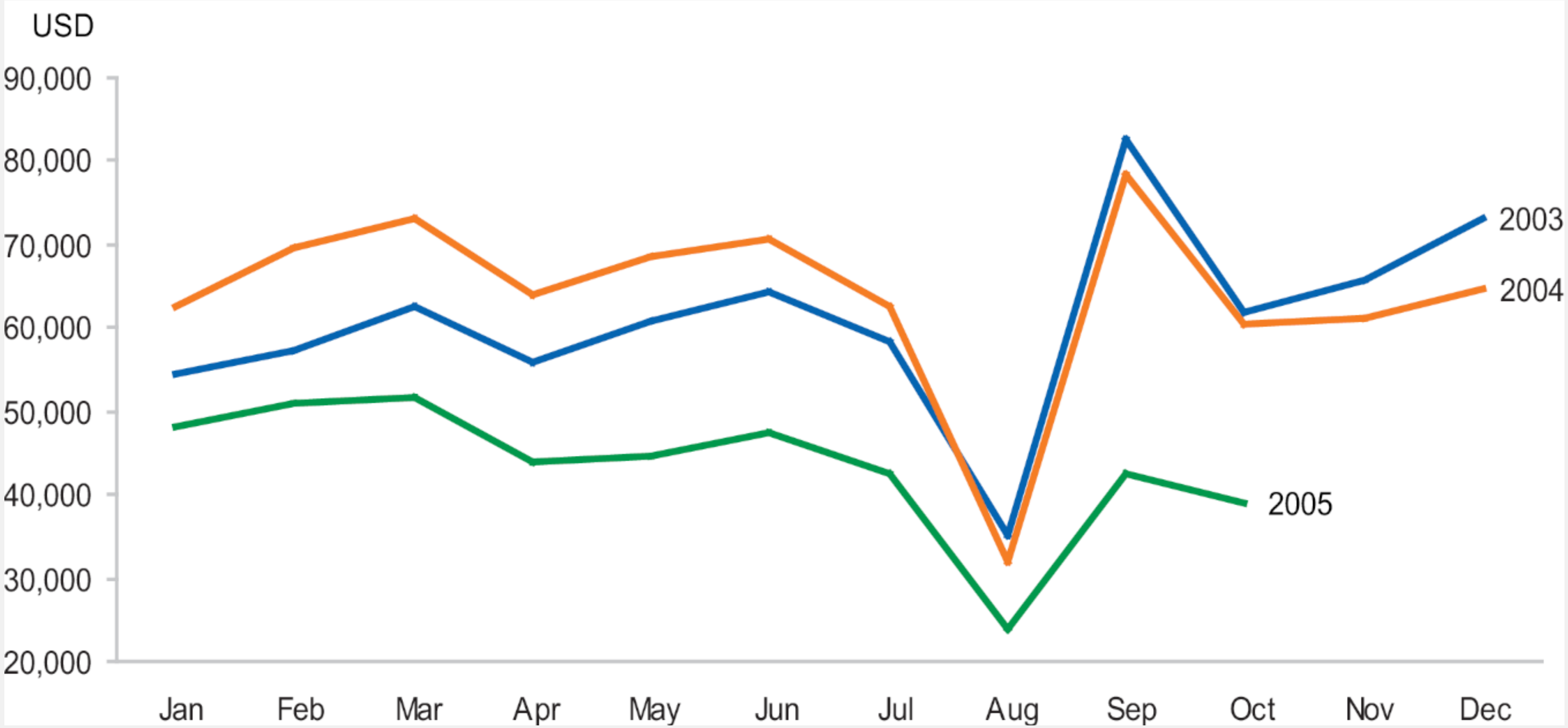
Use percentages to compare rates of change.



Those where Time series

If you look at the whole time-series, to search for differences among different intervals, short-time memory makes you forget when you slide to the next interval

Show them alltoghether in the same plot to allow a straight visual comparison



An EEG time series

Dataset of EEG signals of Open/Close eyes

All data is from one continuous EEG measurement with the Emotiv EEG Neuroheadset.
The duration of the measurement was 117 seconds.
The eye state was detected via a camera during the EEG measurement and added later manually to the file after analysing the video frames.
'1' indicates the eye-closed and '0' the eye-open state.

All values are in chronological order with the first measured value at the top of the data.

INFO AT: http://archive.ics.uci.edu/ml/datasets/EEG+Eye+State#
File with info

For each time step of measurement:

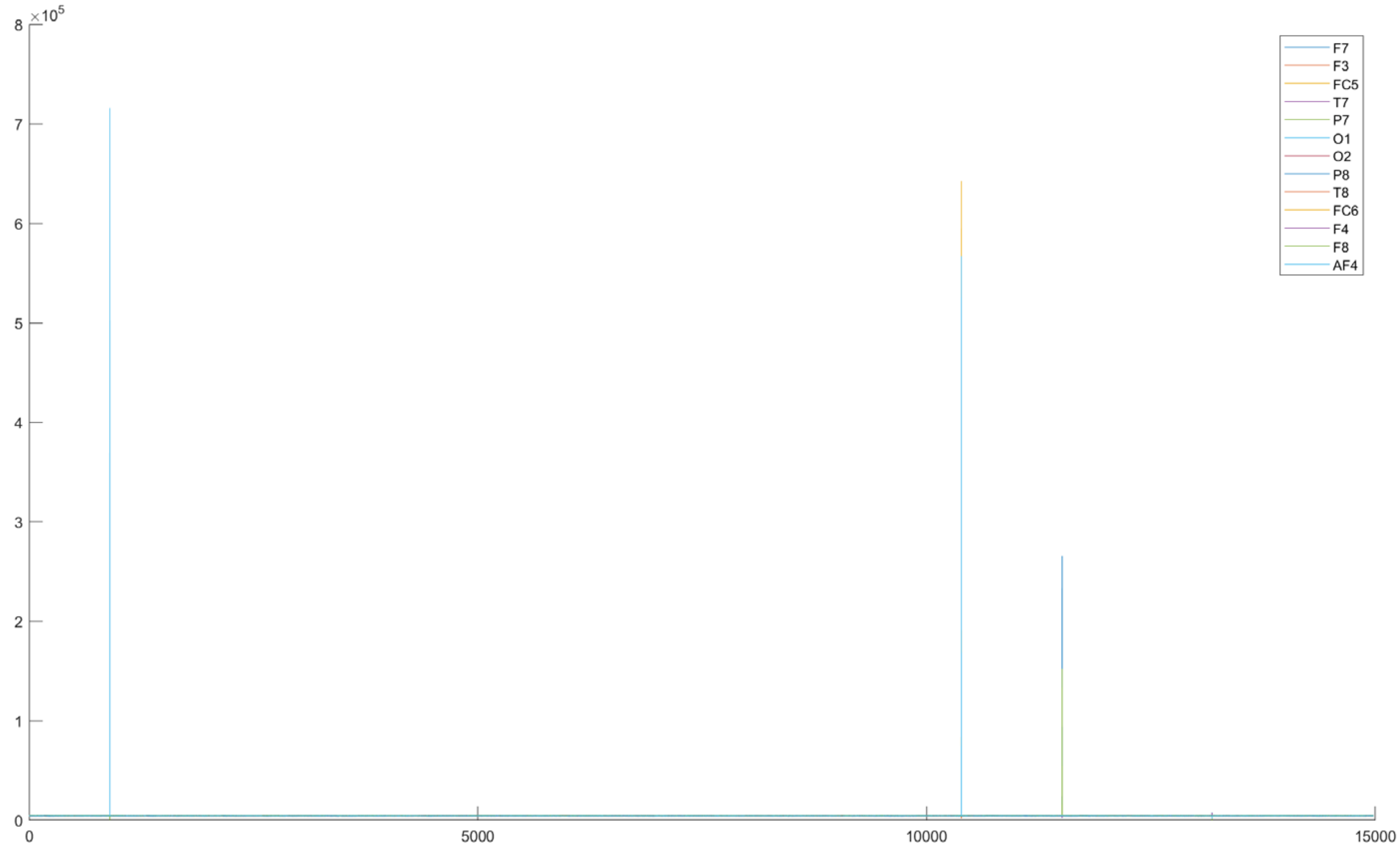
timestep of measurement, 14 different activations, LABEL (0 = open eye/1= close eye)

First step of analysis:

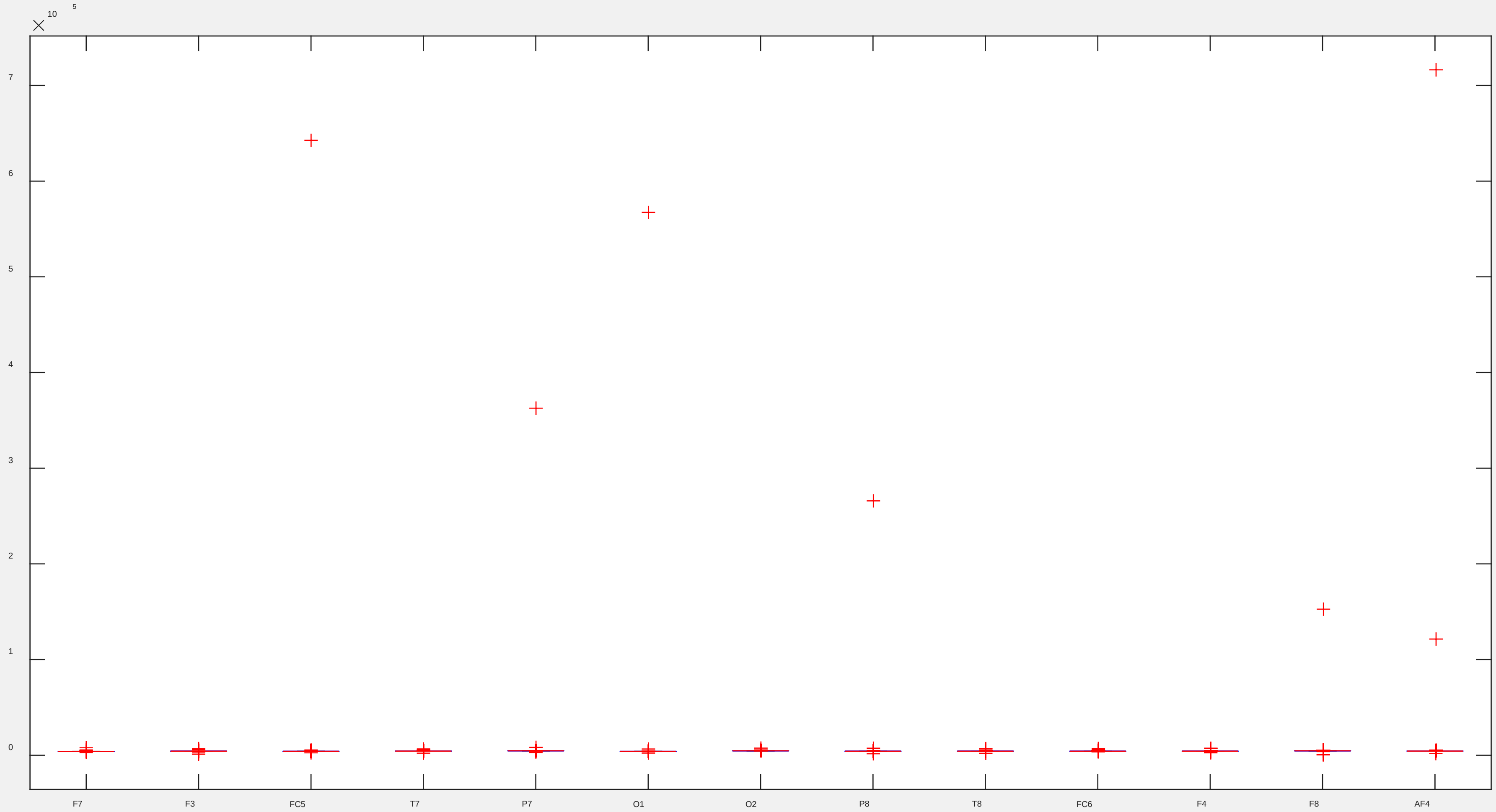
line plot of all the 14 activations in time (regardless of the label)

box-plot of all the 14 activations in time (regardless of the label)

First analysis with MATLAB



×10^5
8
7
6
5
4
3
2
1
0
0
5000
10000
15000
F7
F3
FC5
T7
P7
O1
O2
P8
T8
FC6
F4
F8
AF4

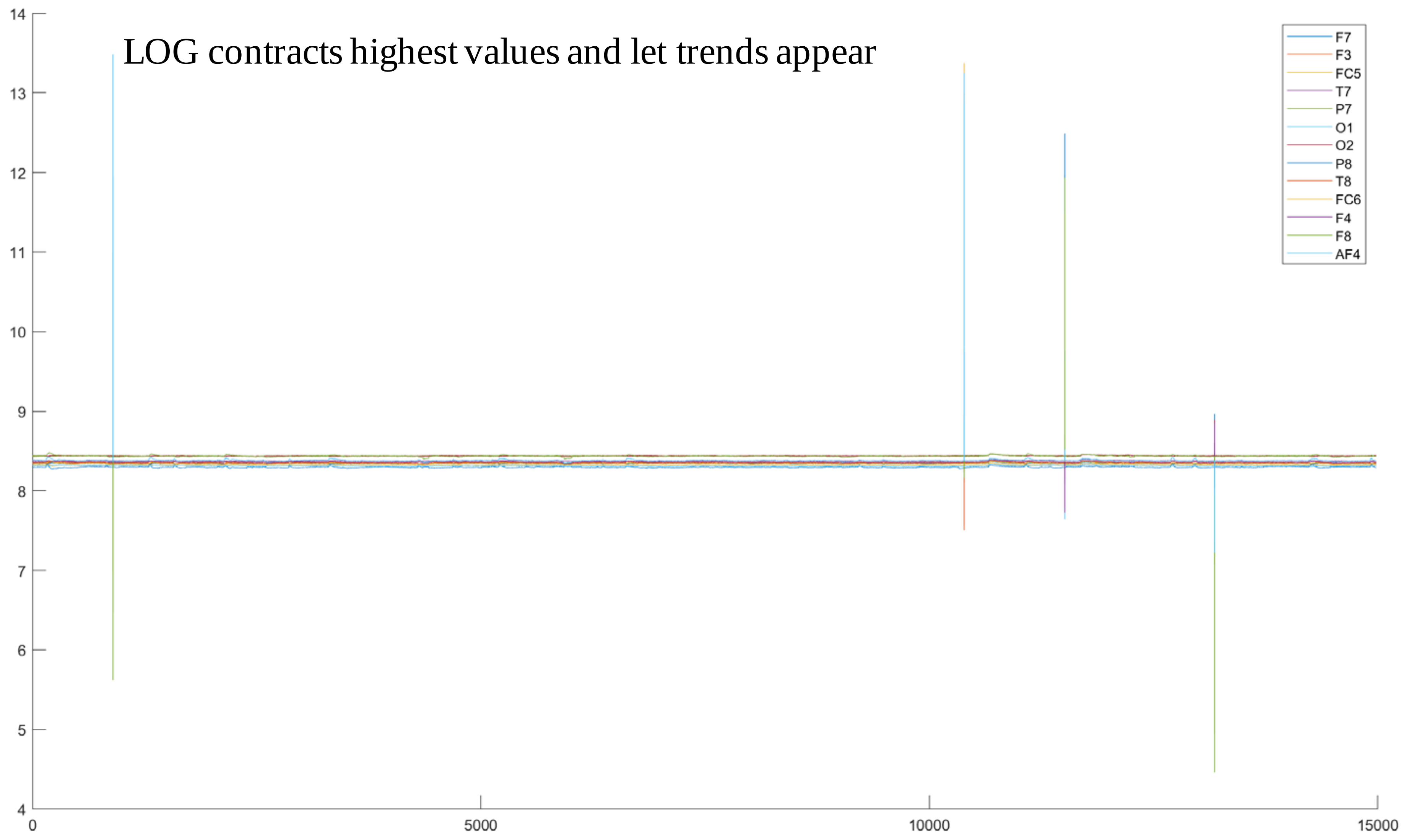


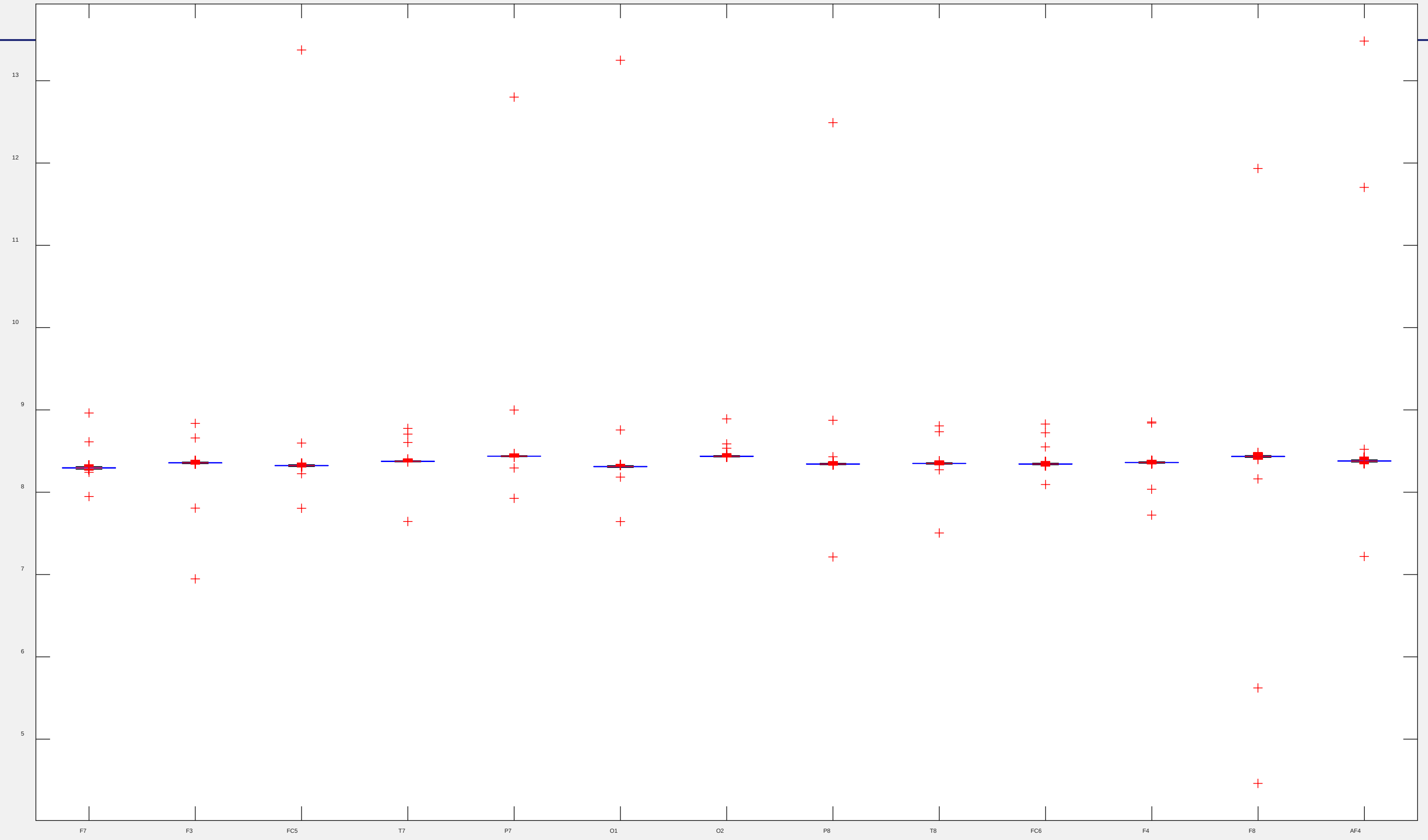
$\times 10^5$
7
6
5
4
3
2
1
0
F7
F3
FC5
T7
P7
O1
O2
P8
T8
FC6
F4
F8
AF4

LOG contracts highest values and increases the scale of small values

LOG shows trends

LOG contracts highest values and let trends appear





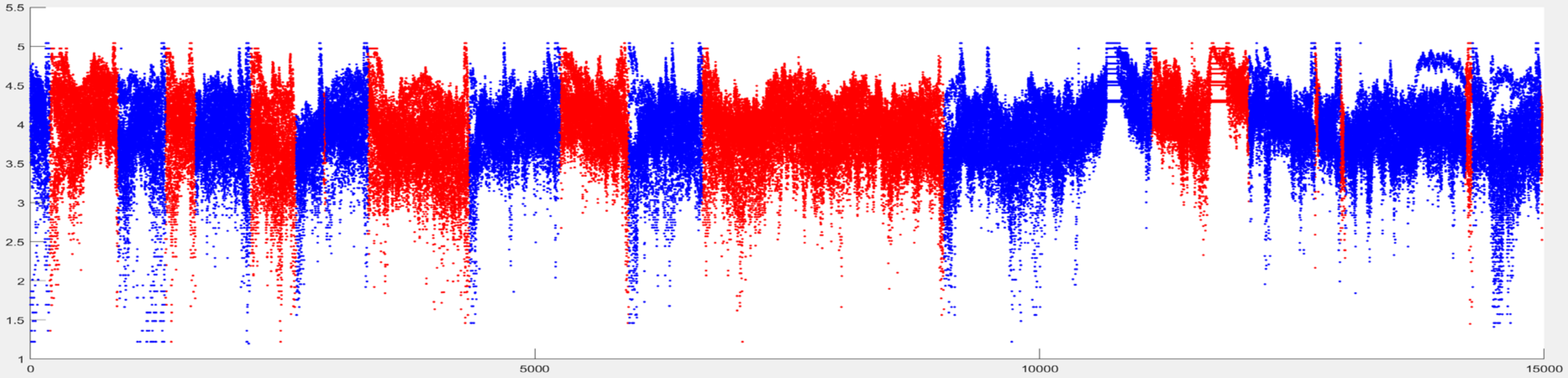
Still highest values hide details: diminish the value of outliers

For each activation (feature):

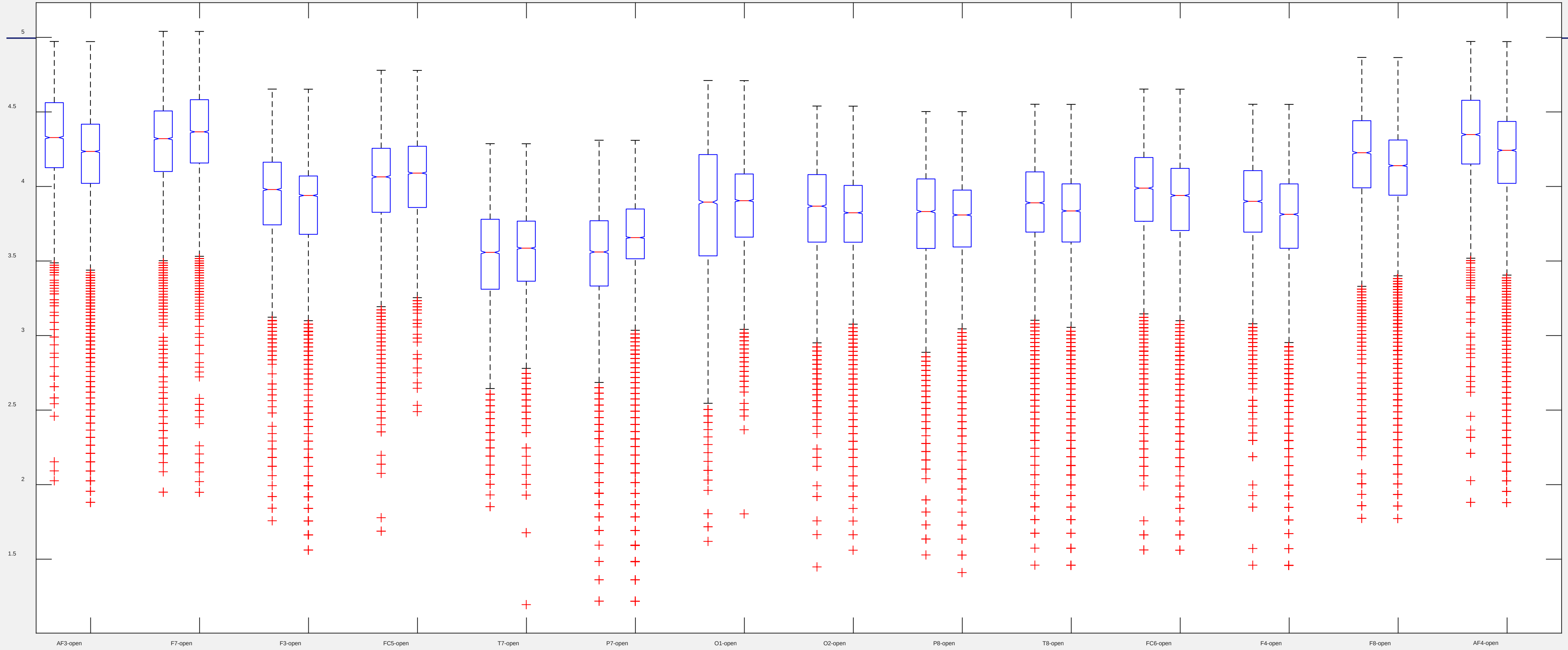
- change outlier values:

```
feature(outliers>max(feature(notOulier))) = max(feature(notOulier)) + range(feature(notOutlier))*0.05
feature(outliers<min(feature(notOulier))) = min(feature(notOulier)) - range(feature(notOutlier))*0.05
```

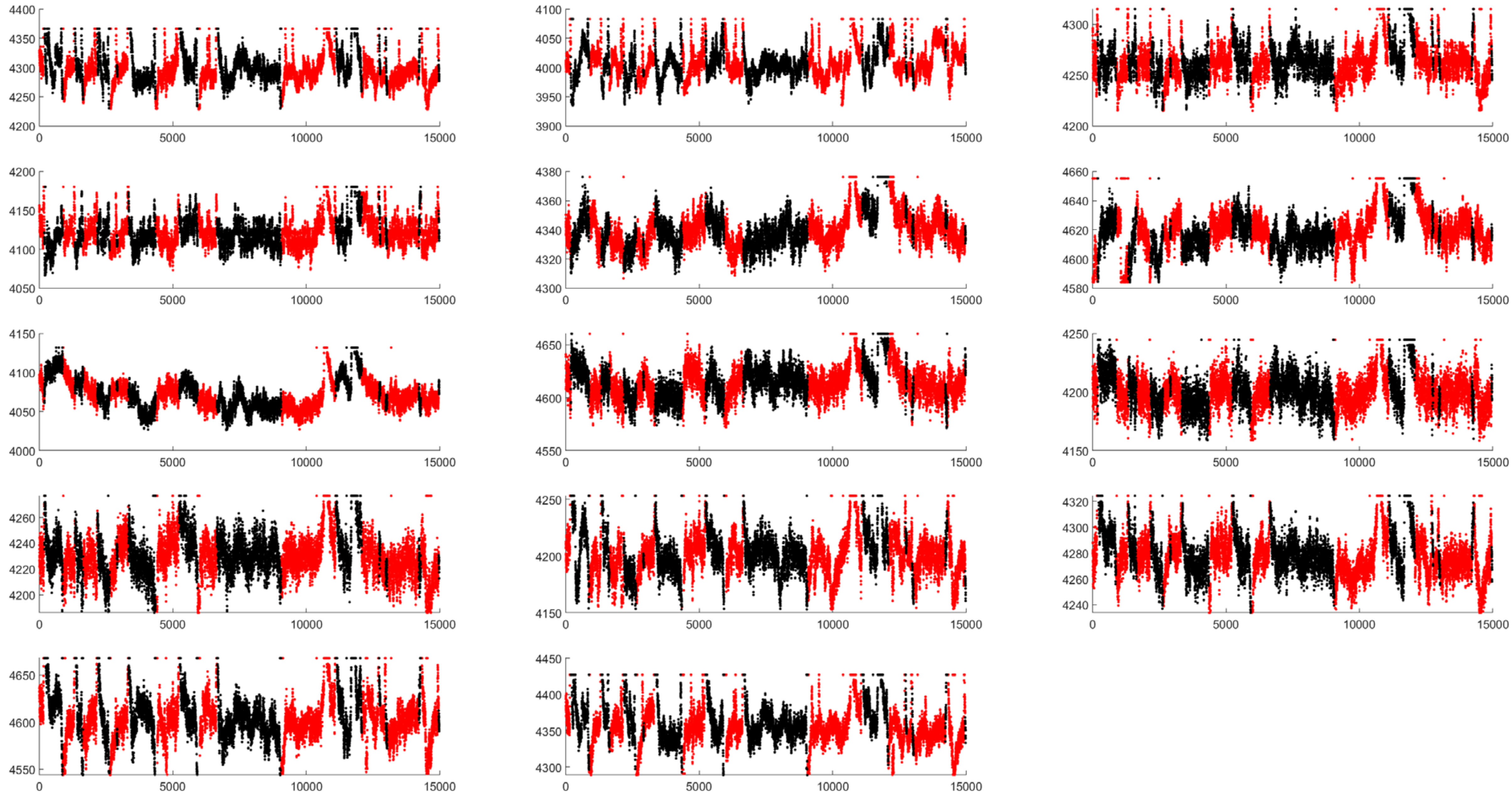
- Translate feature to zero: `feature = feature - min(feature)`

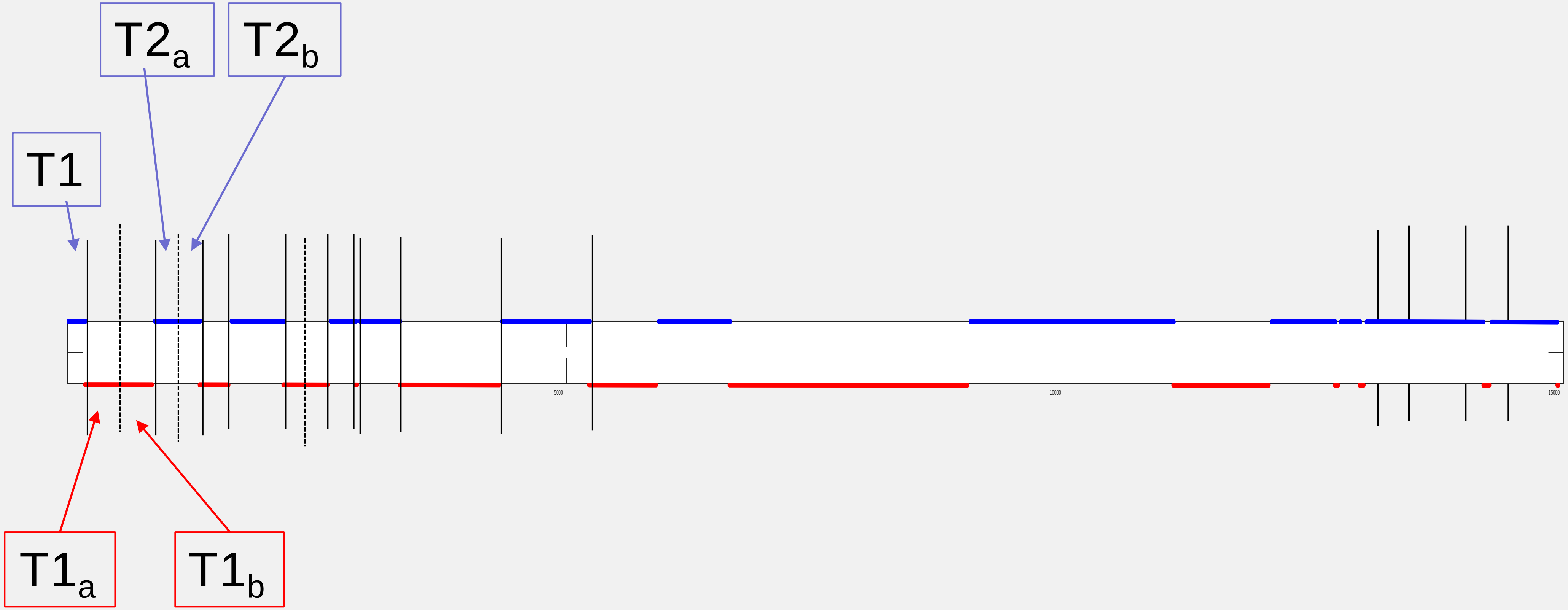


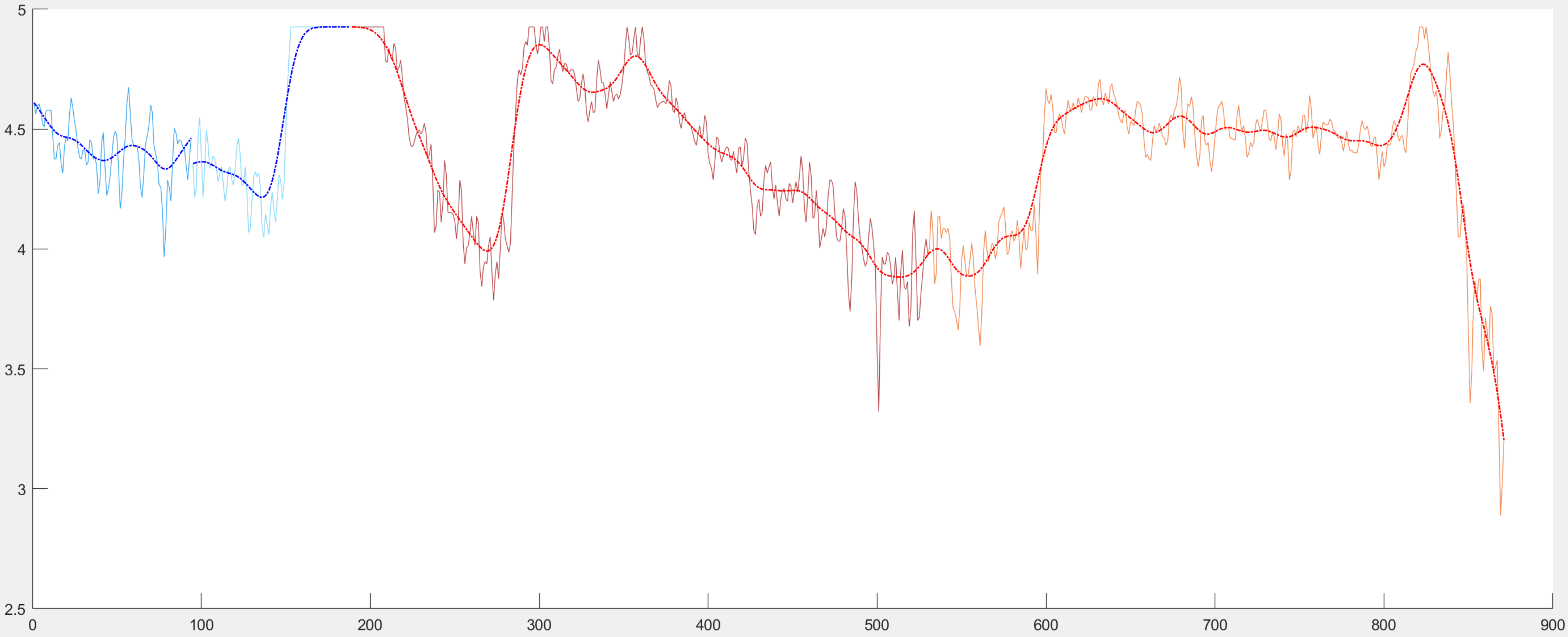
Plotting all the (LOG!!!) feature (blue = open/red = closed)

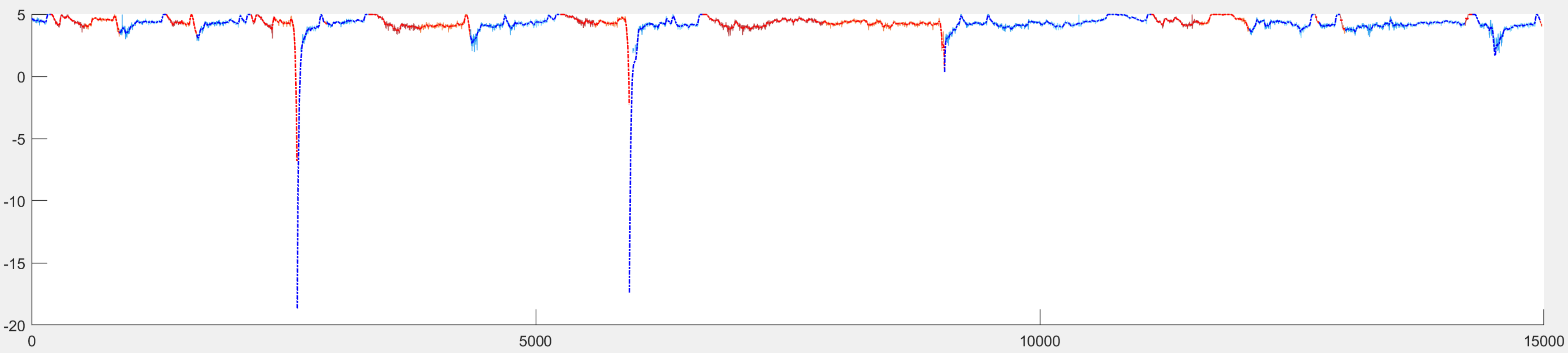


Plotting all the (LOG!!!) feature without distinguishing open and close labels

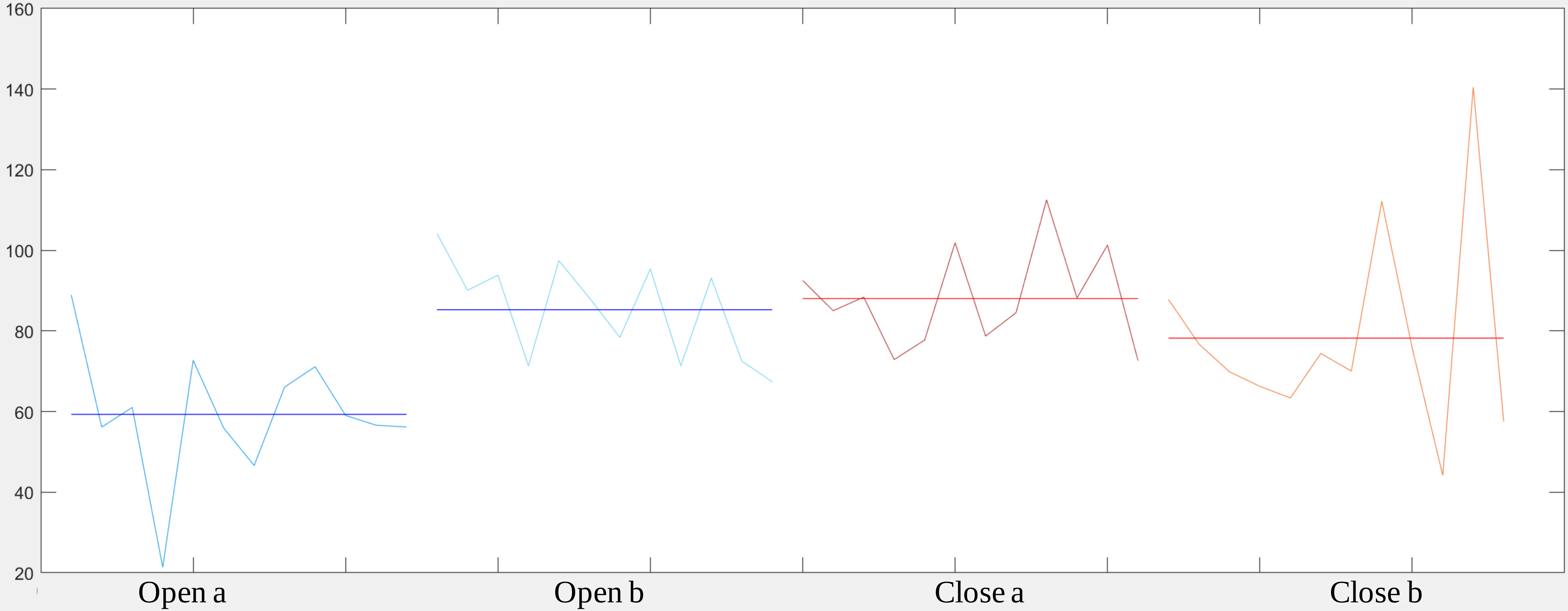

4400 4350 4300 4250 4200
4100 4050 4000 3950 3900
4300 4250 4200
4200 4150 4100 4050
4380 4360 4340 4320 4300
4660 4640 4620 4600 4580
4150 4100 4050 4000
4650 4600 4550
4250 4200 4150
4260 4240 4220 4200
4250 4200 4150
4320 4300 4280 4260 4240
4650 4600 4550
4450 4400 4350 4300
0 5000 10000 15000


T2a
T2b
T1
T1a
T1b
5000
10000
15000

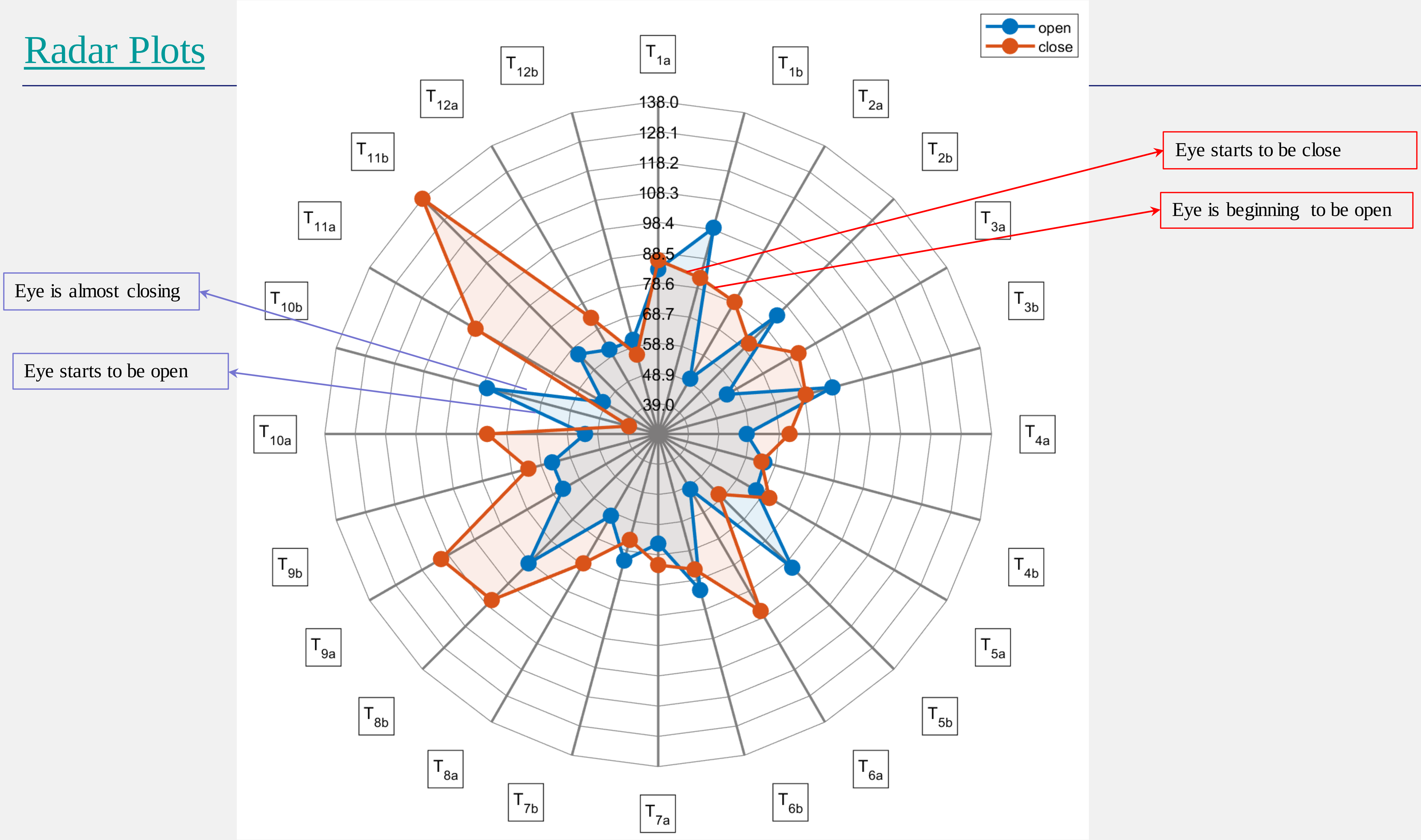

5
4.5
4
3.5
3
2.5
0
100
200
300
400
500
600
700
800
900

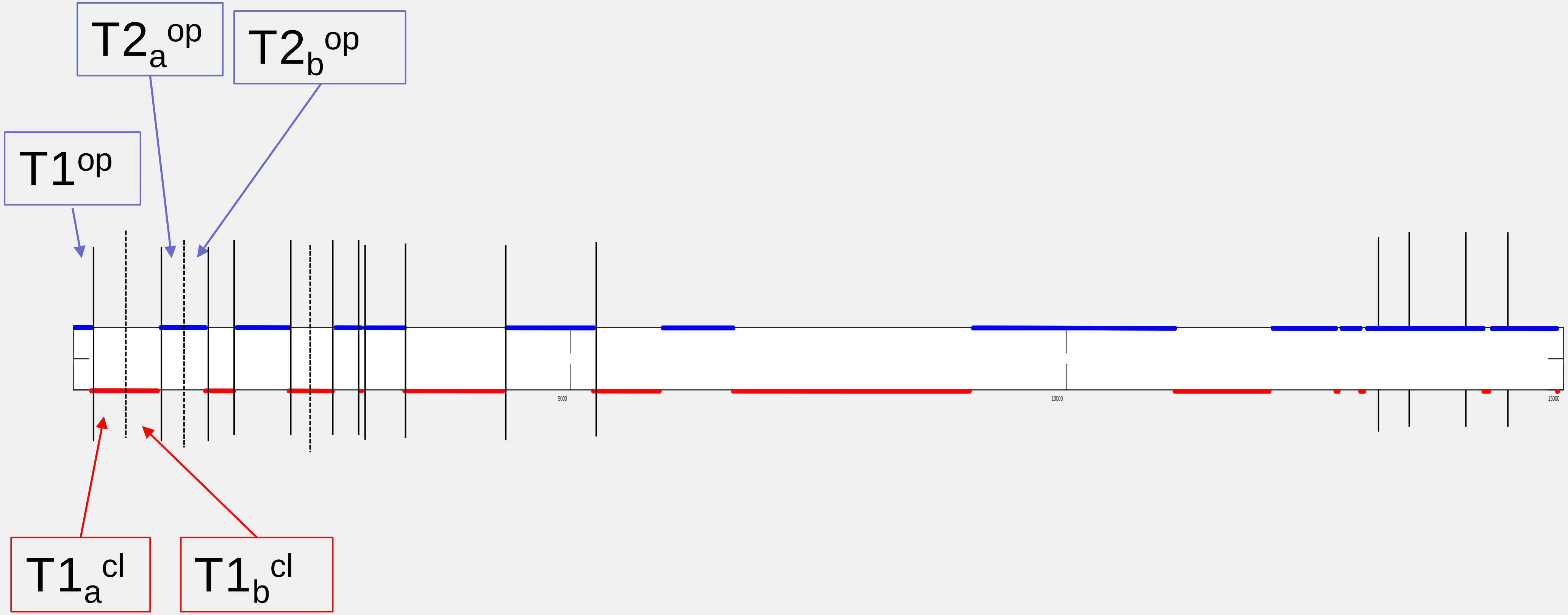

5
0
-5
-10
-15
-20
0
5000
10000
15000

Cycle Plots allow looking at the changing trend in all the periods

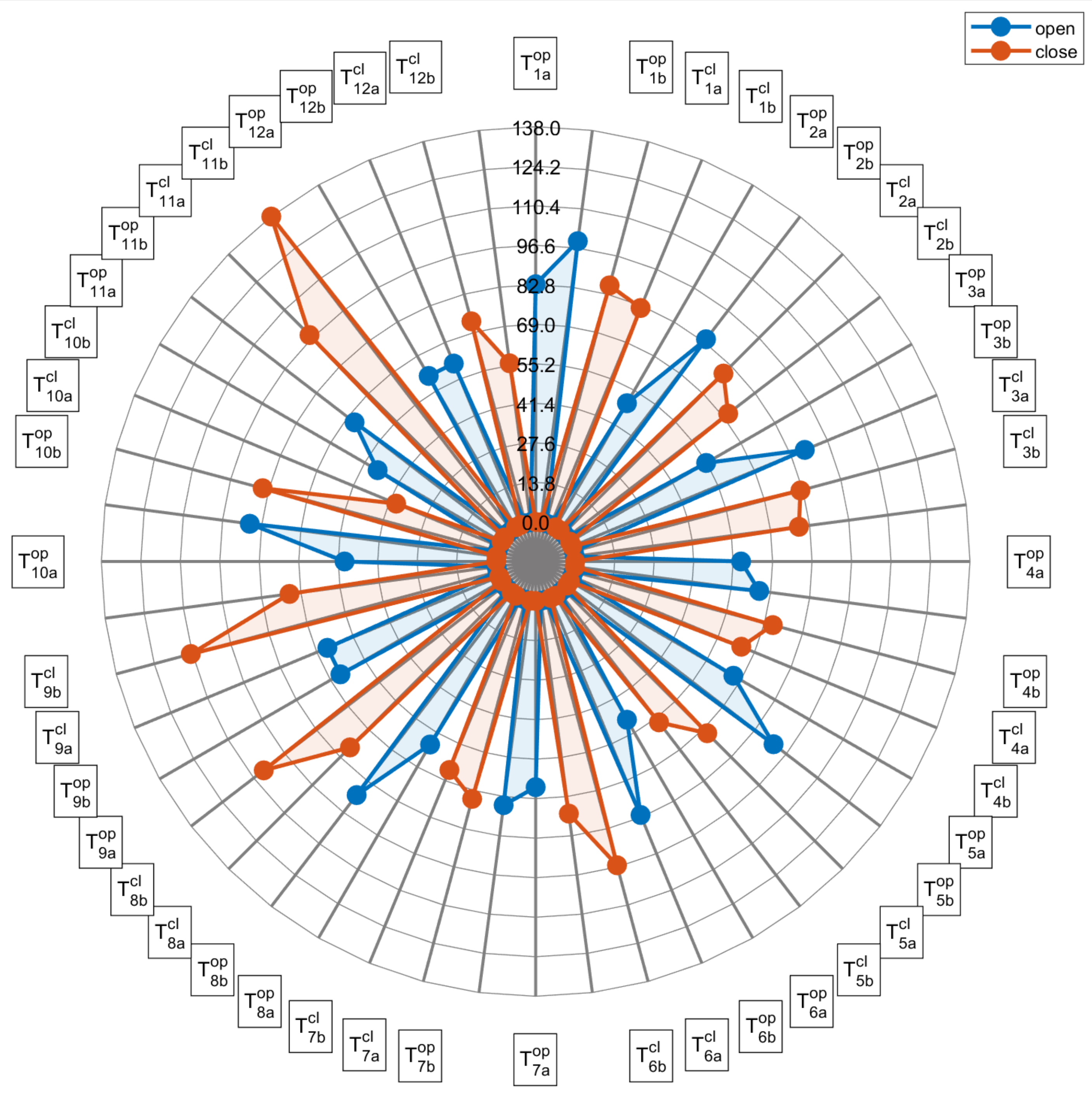


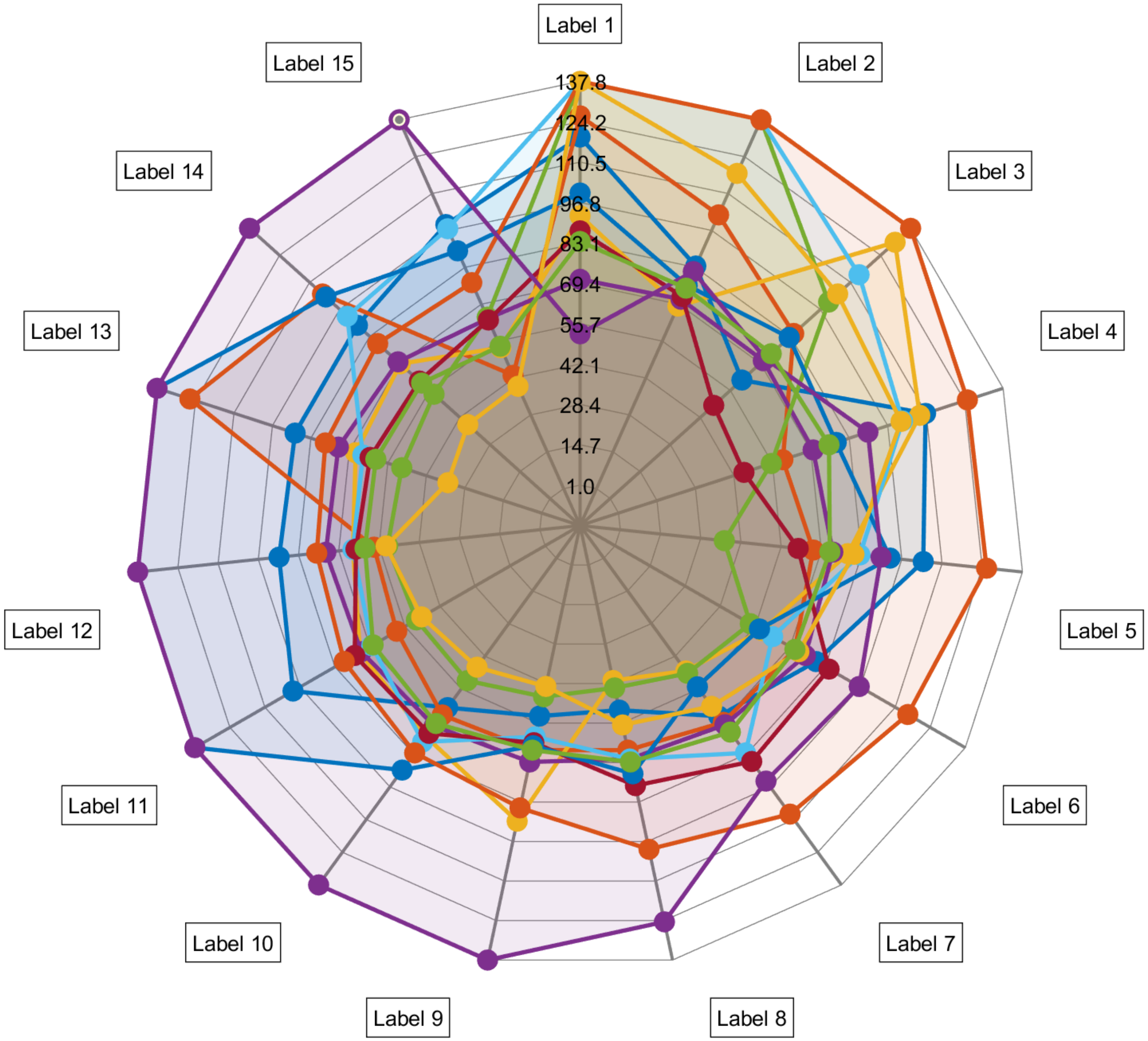
Radar Plots


open
close
T 1a
T 1b
T 2a
T 2b
T 3a
T 3b
T 4a
T 4b
T 5a
T 5b
T 6a
T 6b
T 7a
T 7b
T 8a
T 8b
T 9a
T 9b
T 10a
T 10b
T 11a
T 11b
T 12a
T 12b
138.0
128.1
118.2
108.3
98.4
88.5
78.6
68.7
58.8
48.9
39.0
Eye starts to be close
Eye is beginning to be open
Eye is almost closing
Eye starts to be open

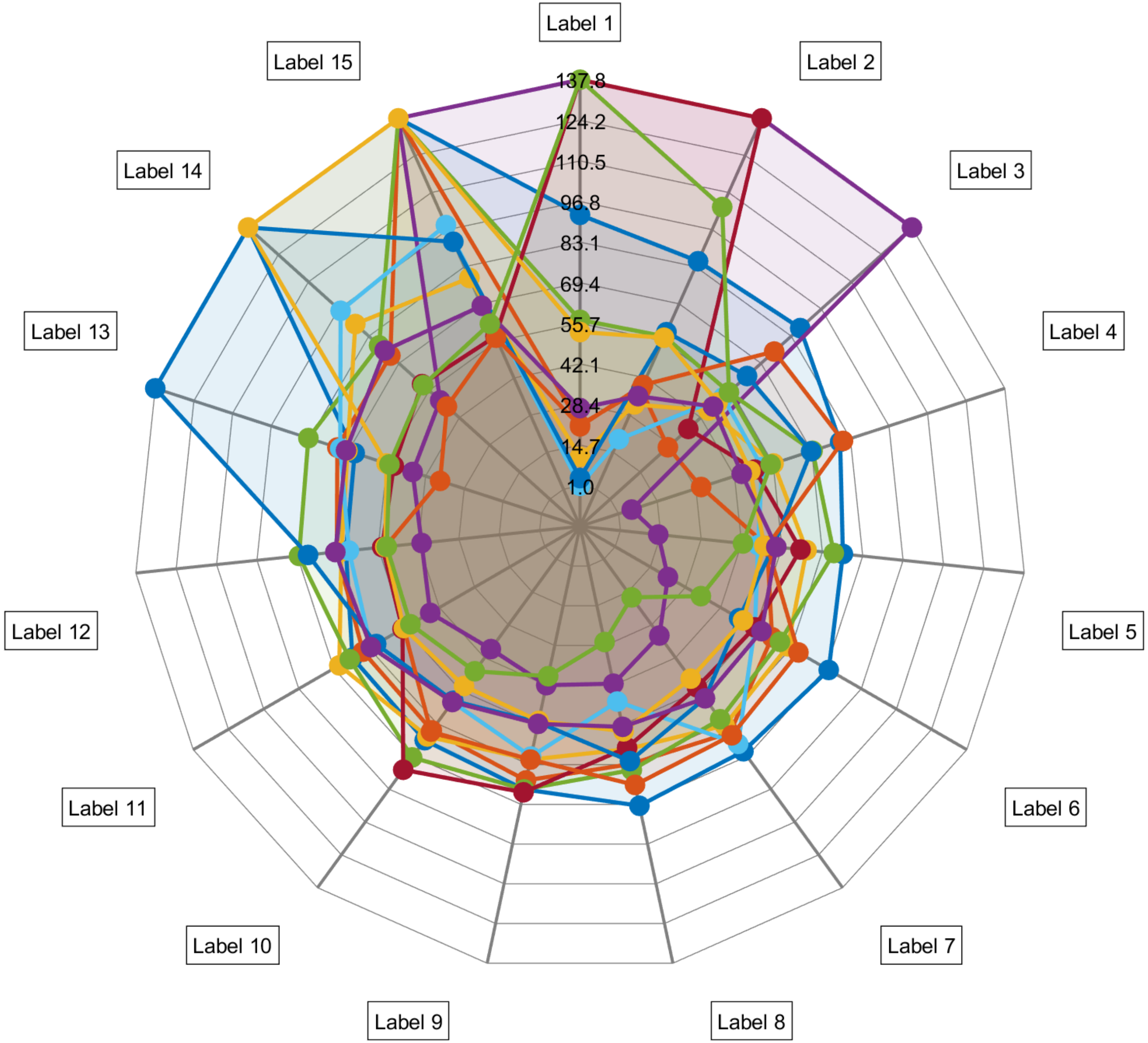


$T2_a^{op}$
$T2_b^{op}$
$T1^{op}$
5000
10000
15000
$T1_a^{cl}$
$T1_b^{cl}$


open
close
138.0
124.2
110.4
96.6
82.8
69.0
55.2
41.4
27.6
13.8
0.0
T^{op}_{1a}
T^{op}_{1b}
T^{cl}_{1a}
T^{cl}_{1b}
T^{op}_{2a}
T^{op}_{2b}
T^{cl}_{2a}
T^{cl}_{2b}
T^{op}_{3a}
T^{op}_{3b}
T^{cl}_{3a}
T^{cl}_{3b}
T^{op}_{4a}
T^{op}_{4b}
T^{cl}_{4a}
T^{cl}_{4b}
T^{op}_{5a}
T^{op}_{5b}
T^{cl}_{5a}
T^{cl}_{5b}
T^{op}_{6a}
T^{op}_{6b}
T^{cl}_{6a}
T^{cl}_{6b}
T^{op}_{7a}
T^{op}_{7b}
T^{cl}_{7a}
T^{cl}_{7b}
T^{op}_{8a}
T^{op}_{8b}
T^{cl}_{8a}
T^{cl}_{8b}
T^{op}_{9a}
T^{op}_{9b}
T^{cl}_{9a}
T^{cl}_{9b}
T^{op}_{10a}
T^{op}_{10b}
T^{cl}_{10a}
T^{cl}_{10b}
T^{op}_{11a}
T^{op}_{11b}
T^{cl}_{11a}
T^{cl}_{11b}
T^{op}_{12a}
T^{op}_{12b}
T^{cl}_{12a}
T^{cl}_{12b}

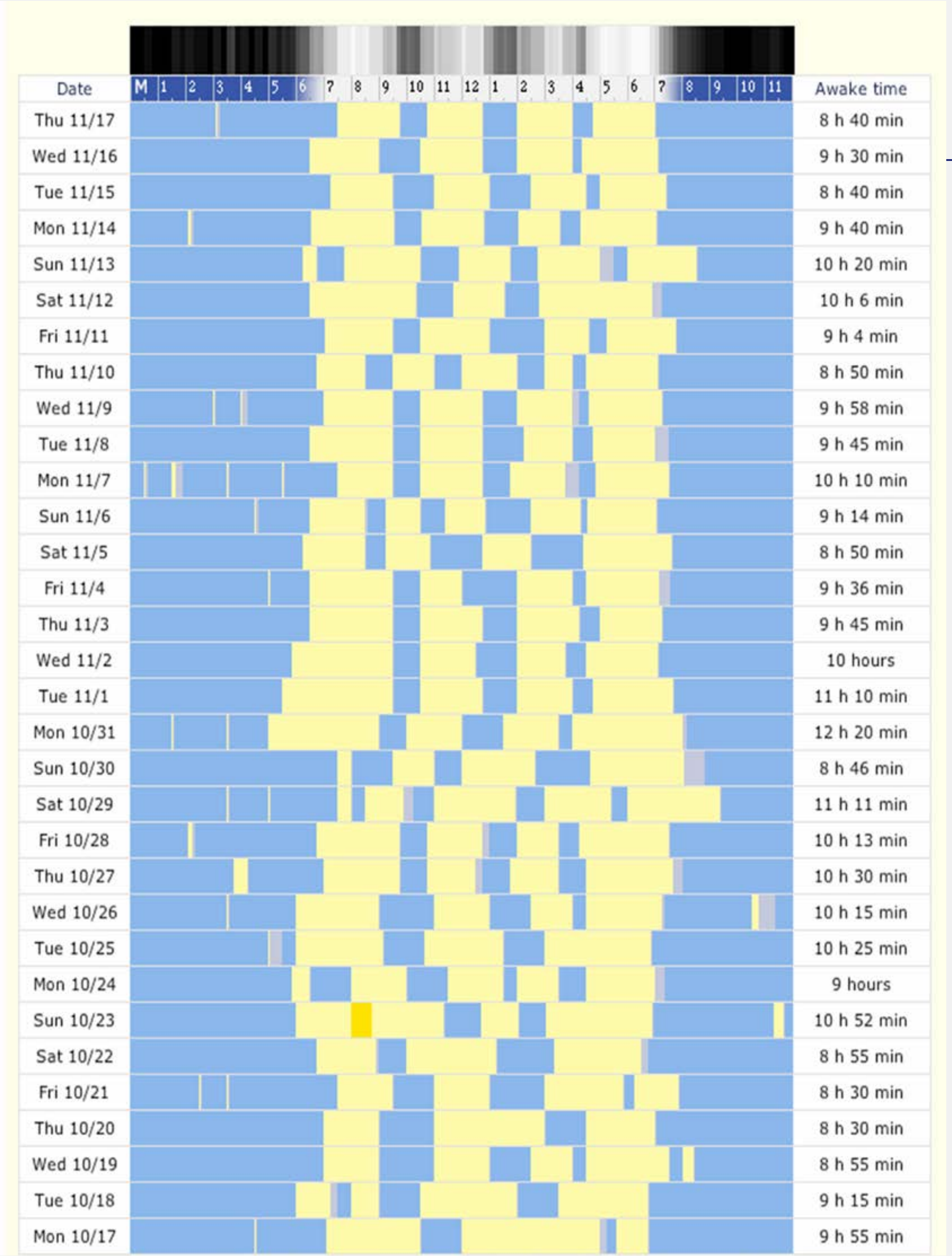


All the T_i^{cl} divided into 15 blocks



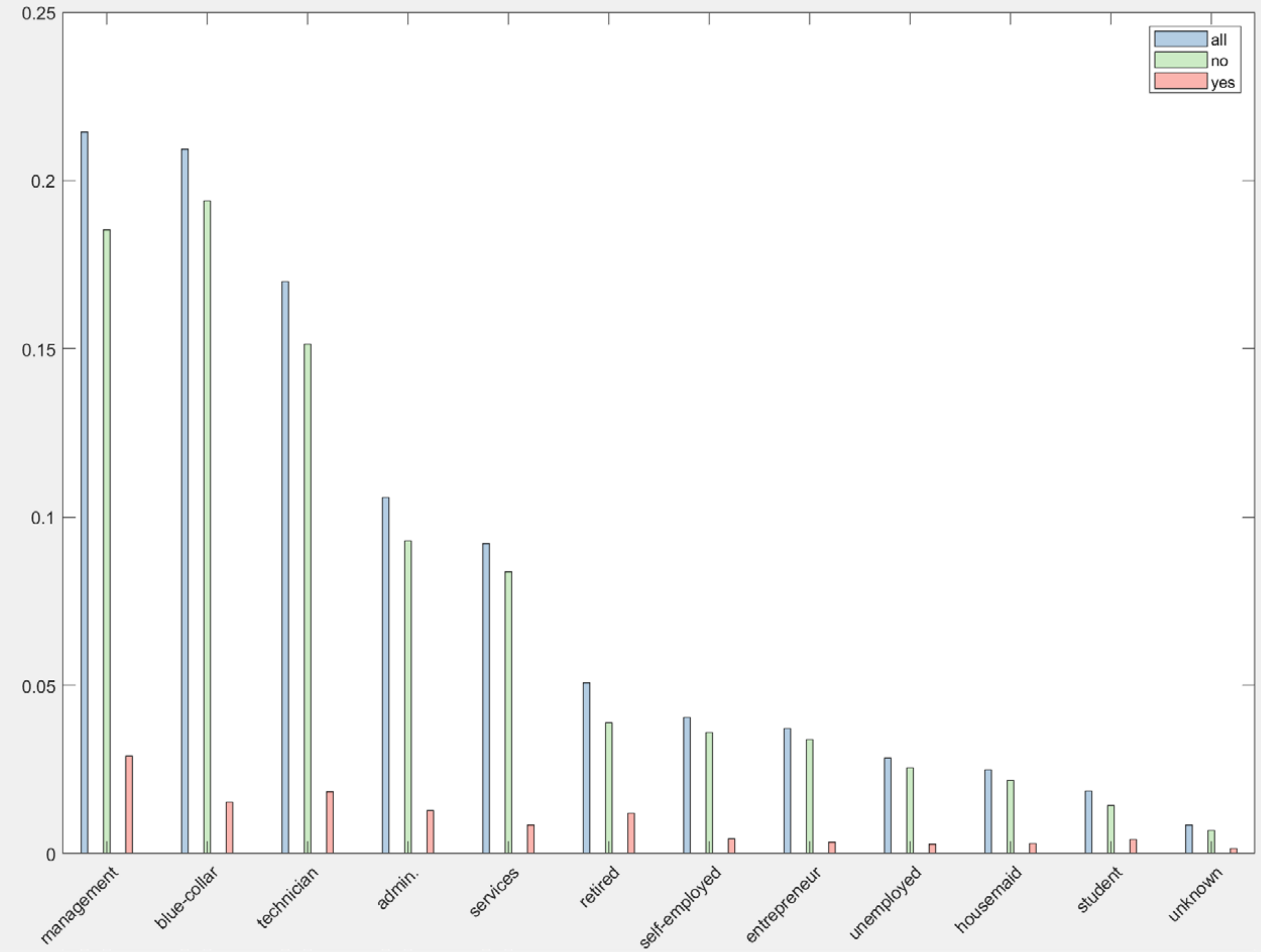
All the T_i^{op} divided into 15 blocks

Alternatively, you may use heatmaps



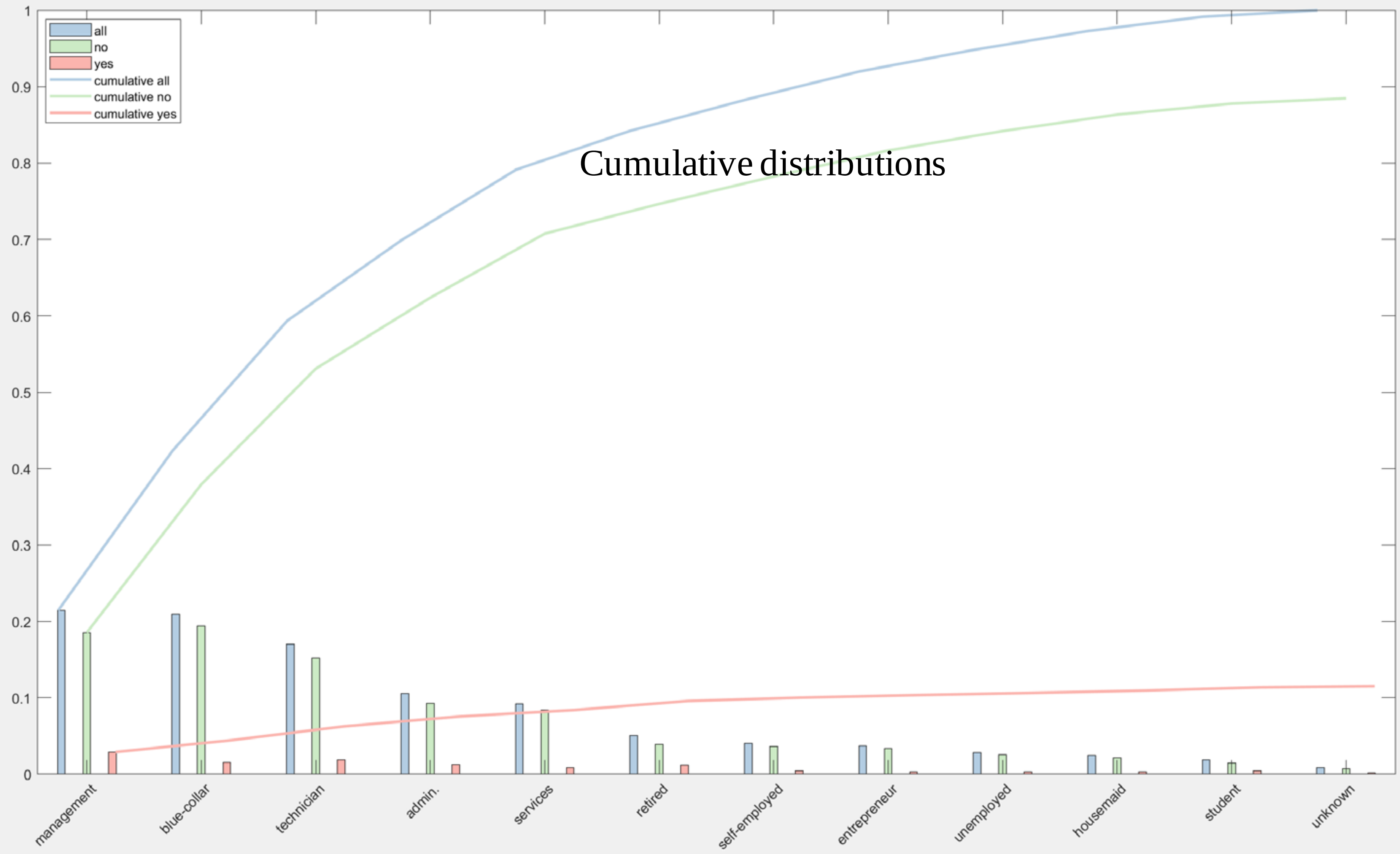
Visualization of categorical data (essentially proportions)

Approved credit-card payments

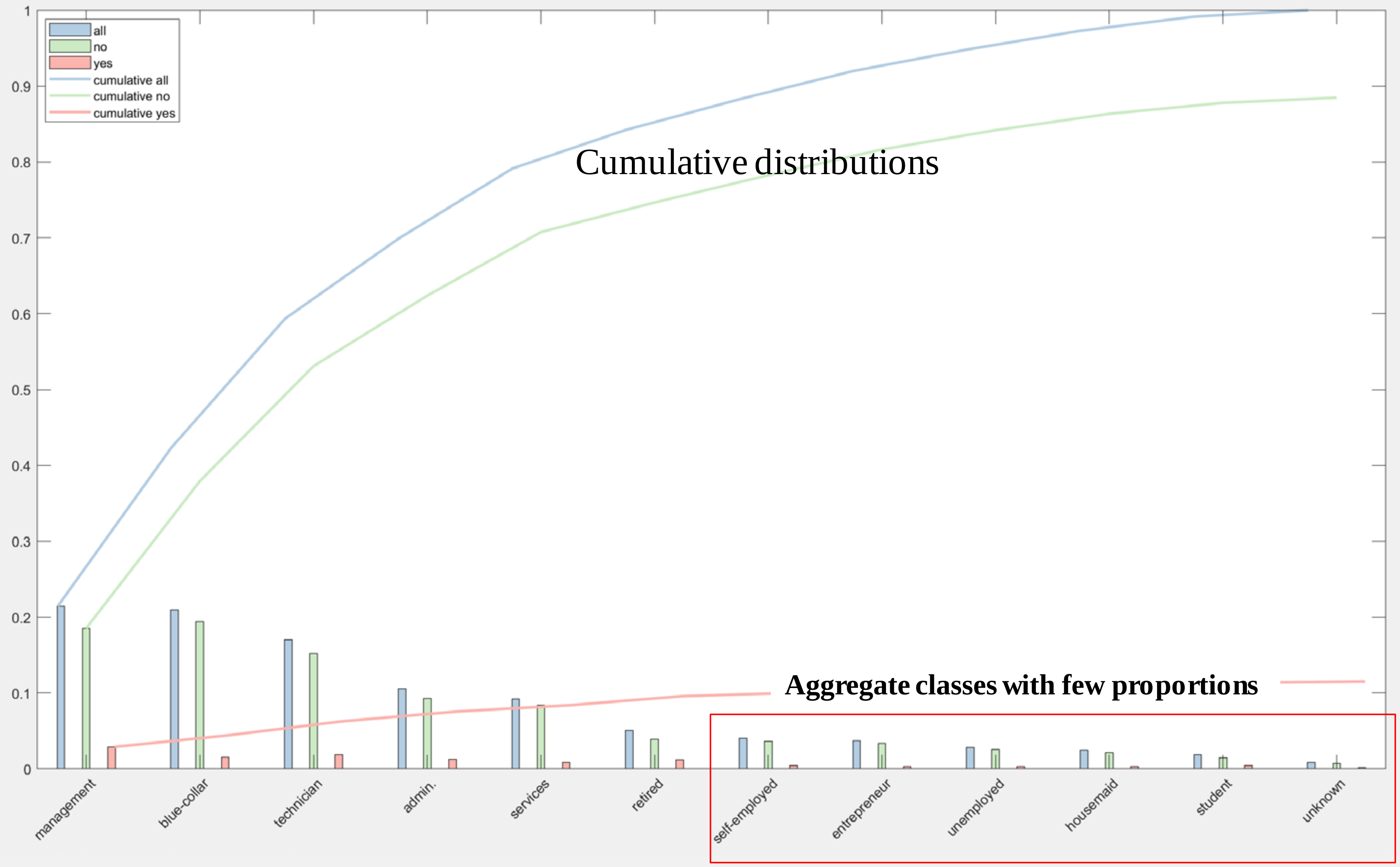


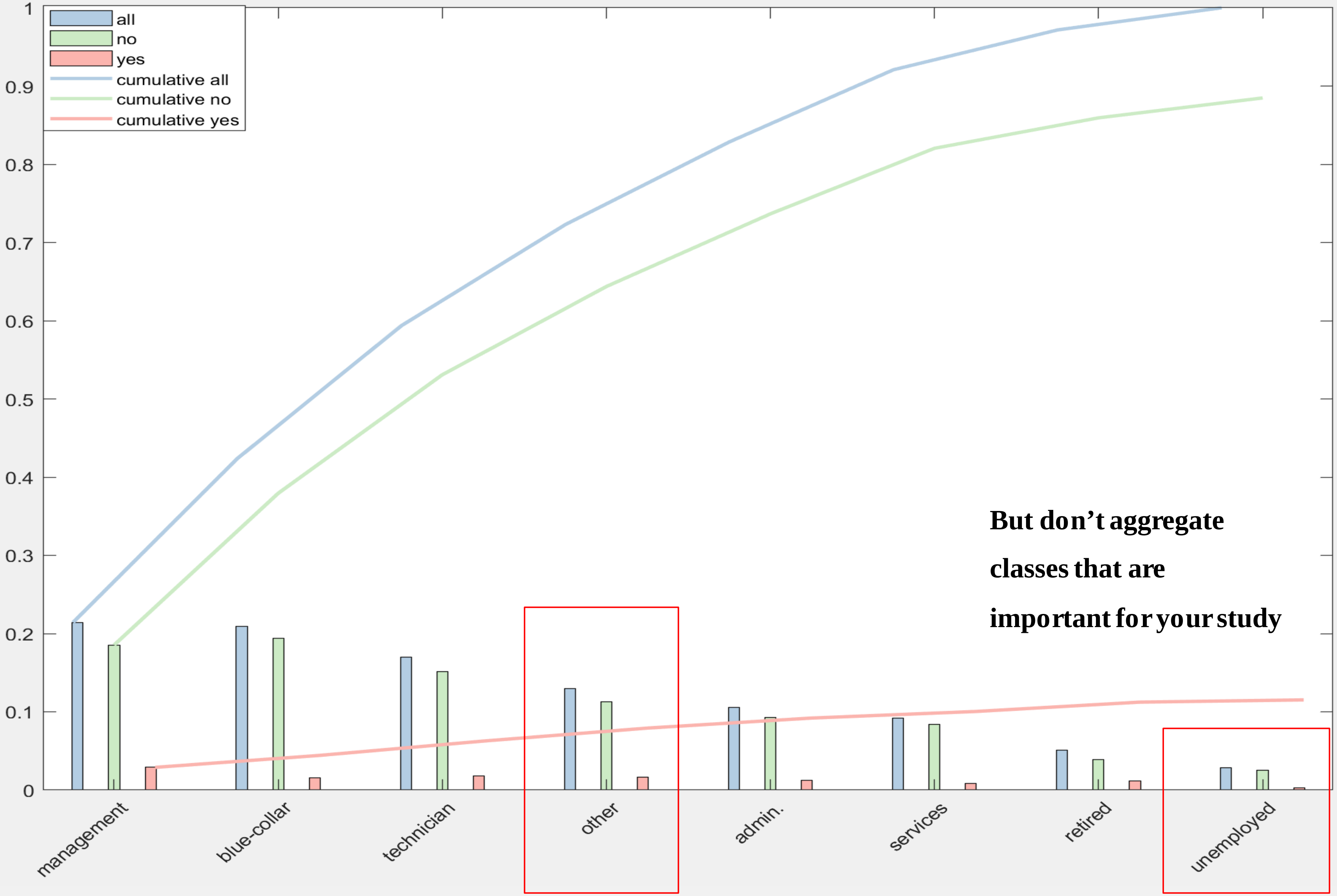
0.25
0.2
0.15
0.1
0.05
0
all
no
yes
management
blue-collar
technician
admin.
services
retired
self-employed
entrepreneur
unemployed
housemaid
student
unknown

A pareto chart might be useful



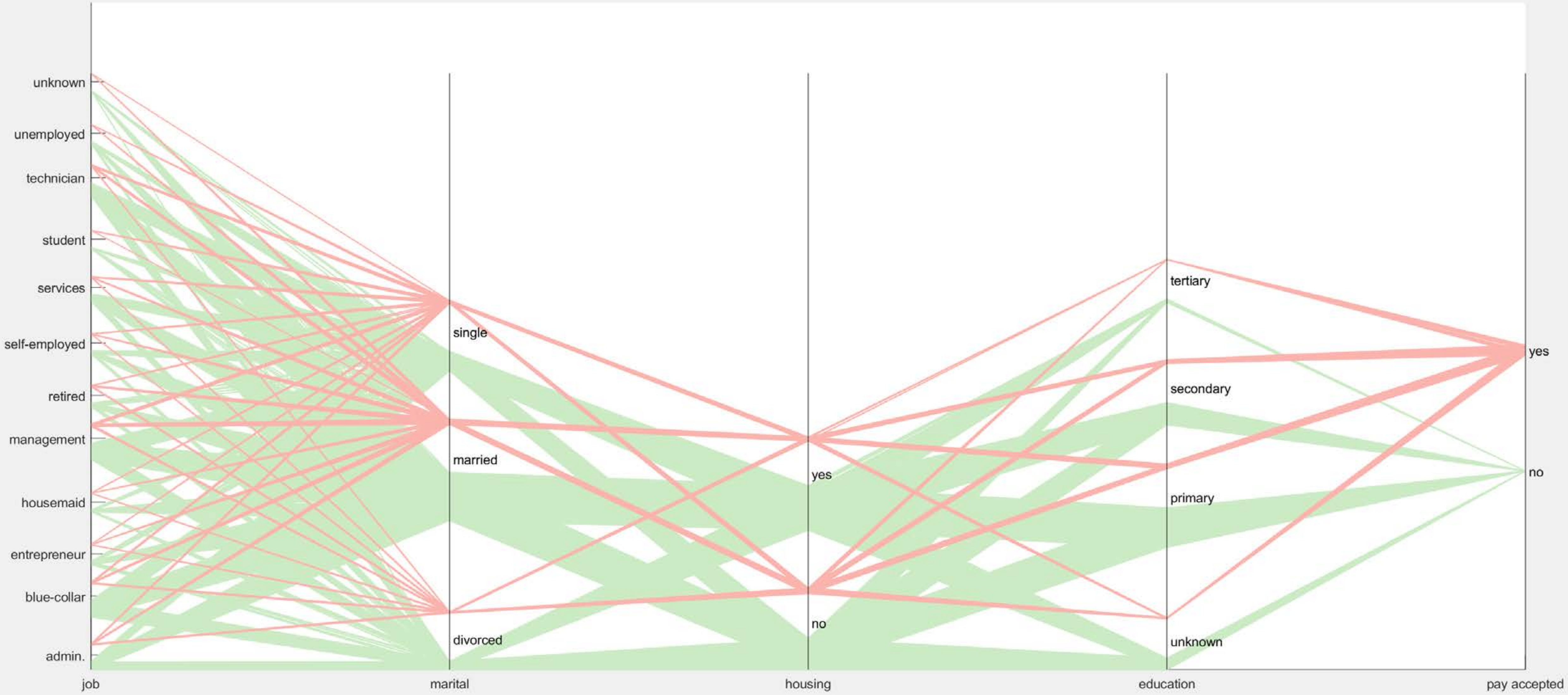
A pareto chart might be useful





But don’t aggregate classes that are important for your study

Parallel sets show the categorical trends



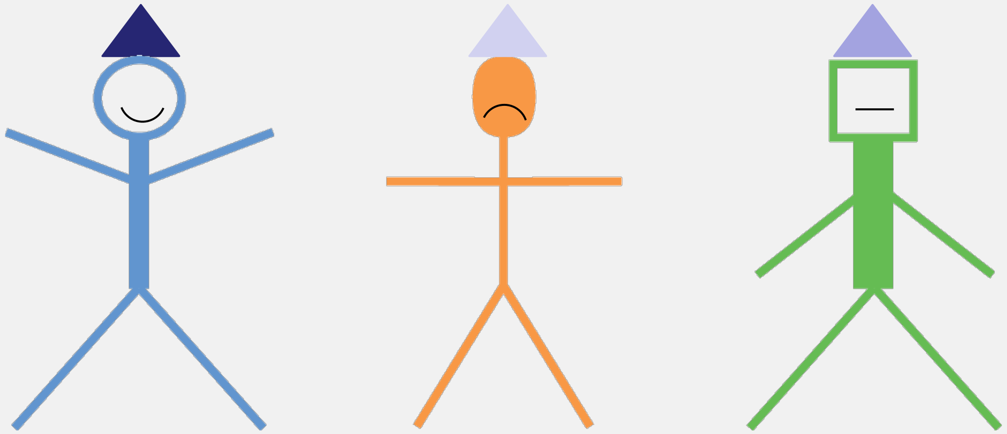
MULTIVARIATE CATEGORICAL DATA

Parallel sets allow visualizing **multivariate** categorical/ordinal data

Otherwise you may use

Glyphs: “a graphical object designed to convey multiple data values”

Information Visualization: Perception for Design, Colin Ware

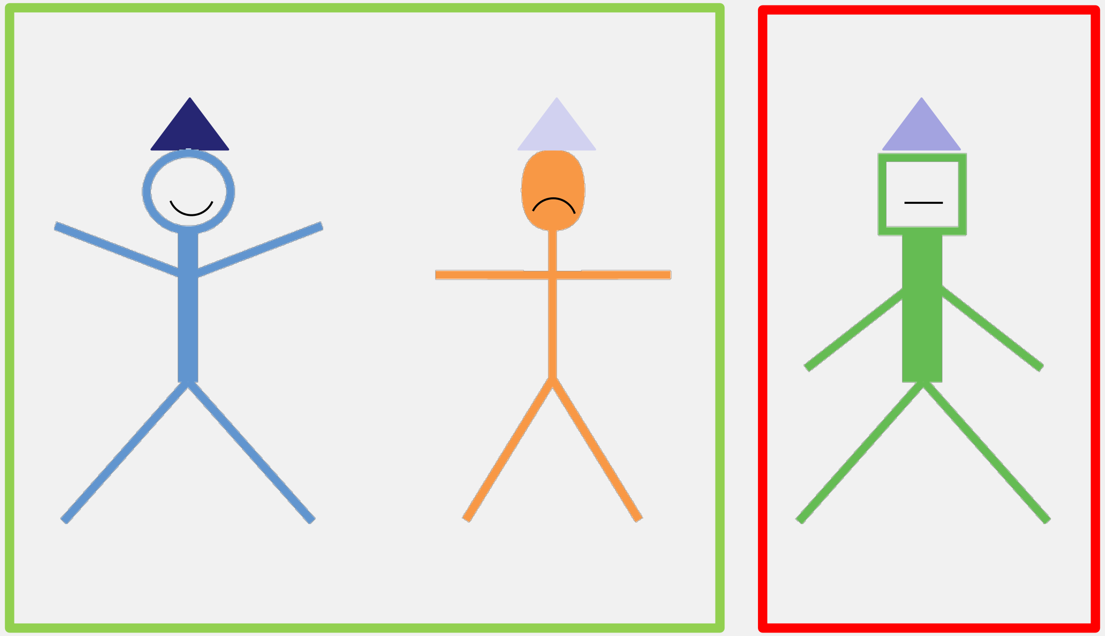


Visual Attribute	Variable
Shape of head + head width	Job* + position in the job**
Marital	Shape of Mouth
Color	Housing***
Color of hat	Education****
Thickness of body	deposit balance
Position of the legs	mean monthly expenses
Position of arm	Expenses of this month

* Jobs clustered to diminish the number of classes
** e.g. CEO, chief administration, manager, employee, intern…
*** Housing could have more classes (private, private with bank loan, under rent, no)

****Education has a sort of ordering

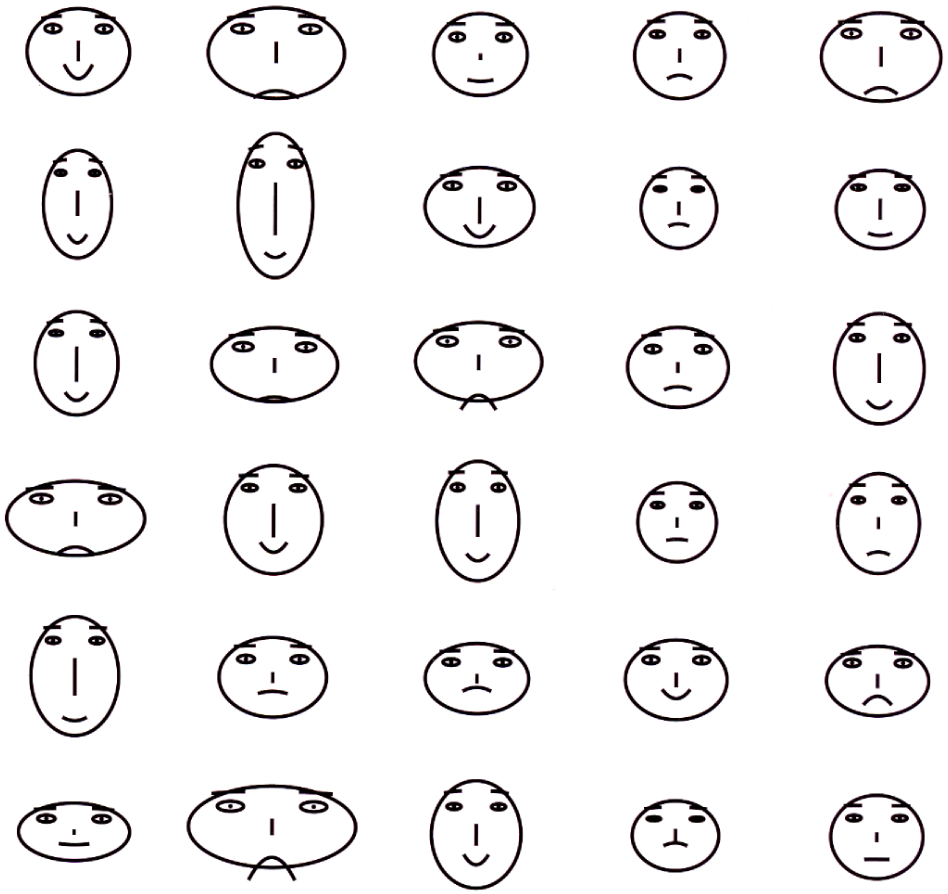
The strongest attribute (enclosure for the LABEL, if any)



Visual Attribute	Variable
Shape of head + head width	Job* + position in the job**
Marital	Shape of Mouth
Color	Housing***
Color of hat	Education****
Thickness of body	deposit balance
Position of the legs	mean monthly expenses
Position of arm	Expenses of this month

* Jobs clustered to diminish the number of classes
** e.g. CEO, chief administration, manager, employee, intern…
*** Housing could have more classes (private, private with bank loan, under rent, no)
****Education has a sort of ordering

Chernoff faces (Herman Chernoff 1972)

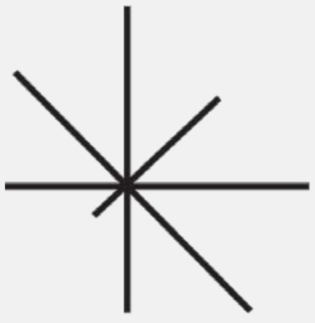


Why faces + expression?

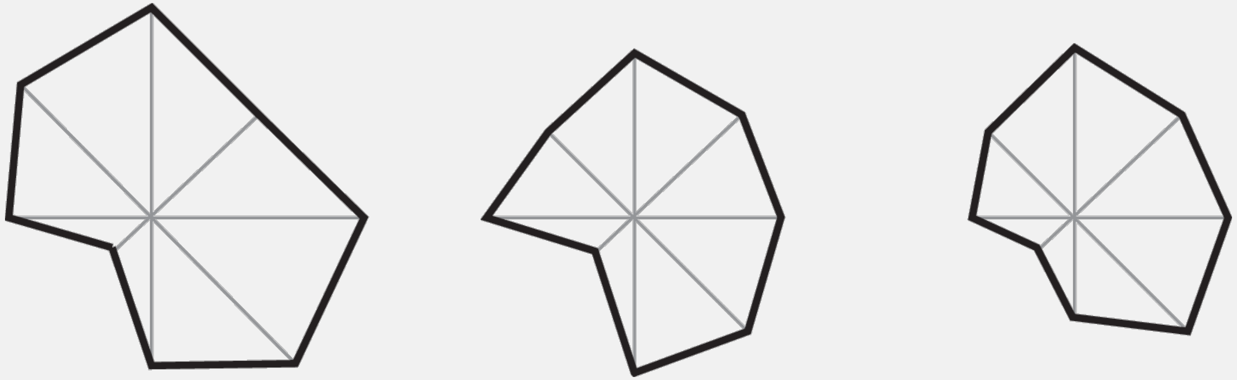
Because we are used to recognize people and interpret their facial expressions

Otherwise, as it they were plotted in a radar plot…

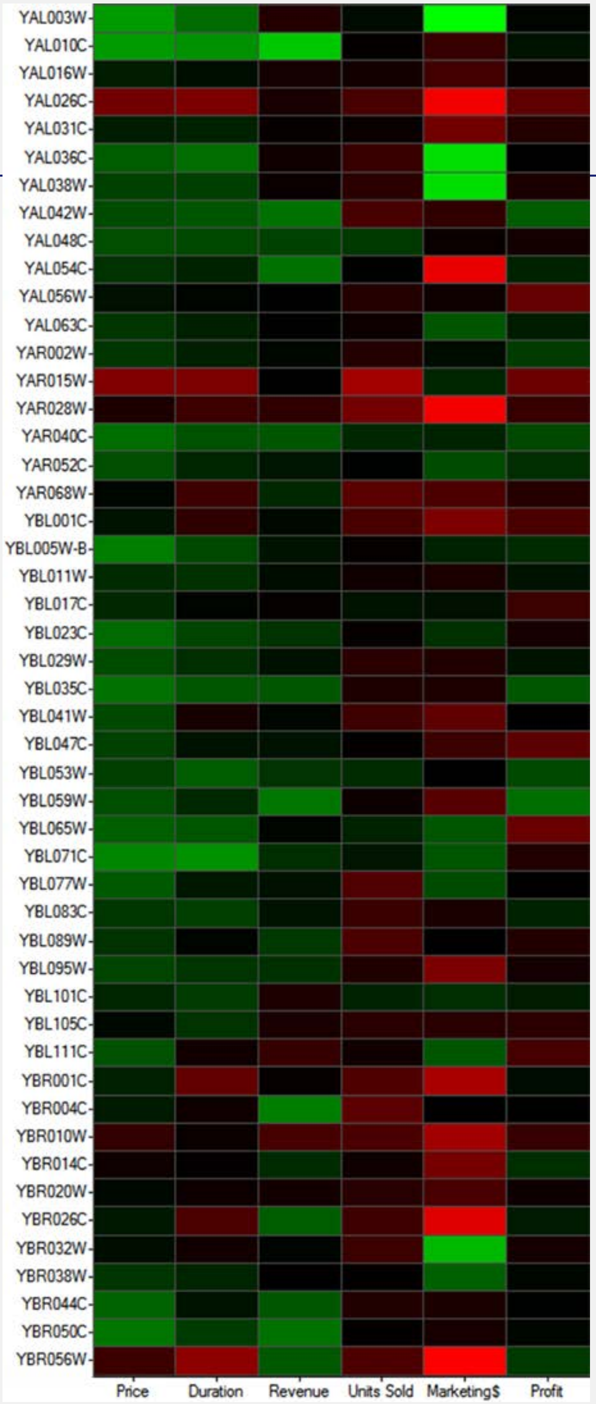
whiskers



Stars



Or categorical heatmaps



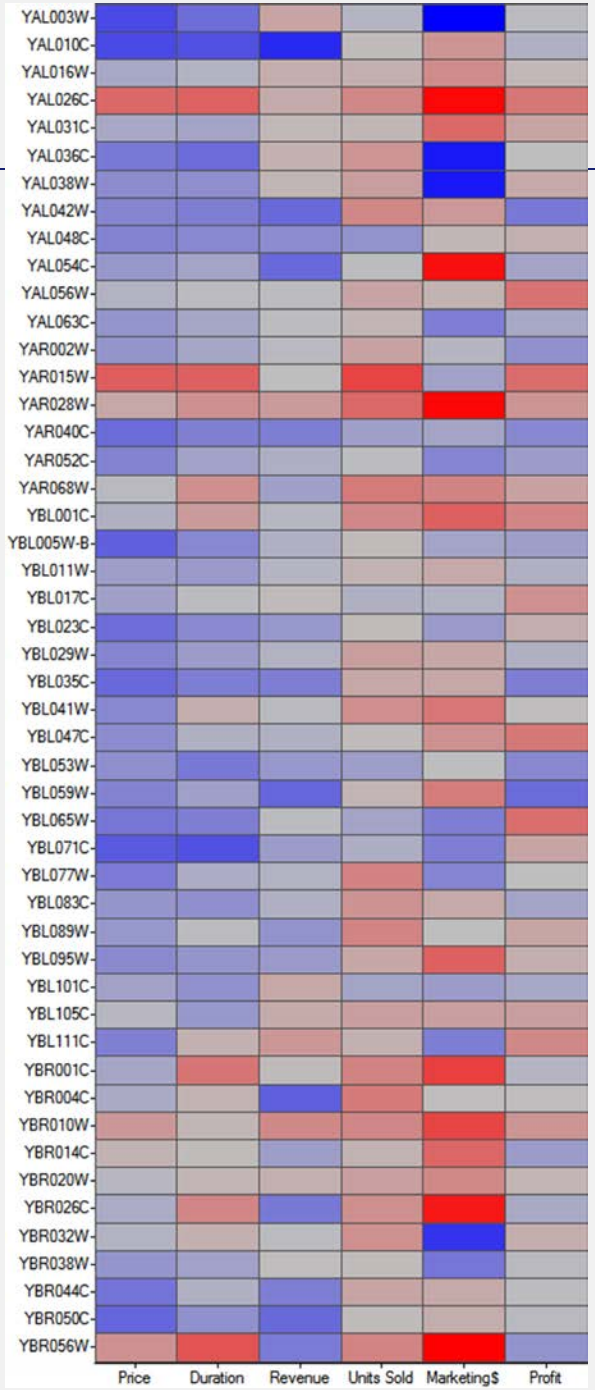
Green = colors over the average (the lighter the higher)

Black = values near the average of the class

Reds = color below average (the ligther the lower)

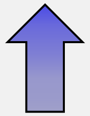
Color Blind?

Black for average is a bad perceptual association (use grey)



Green = colors over the average

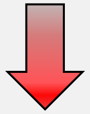
(the more saturated the higher)



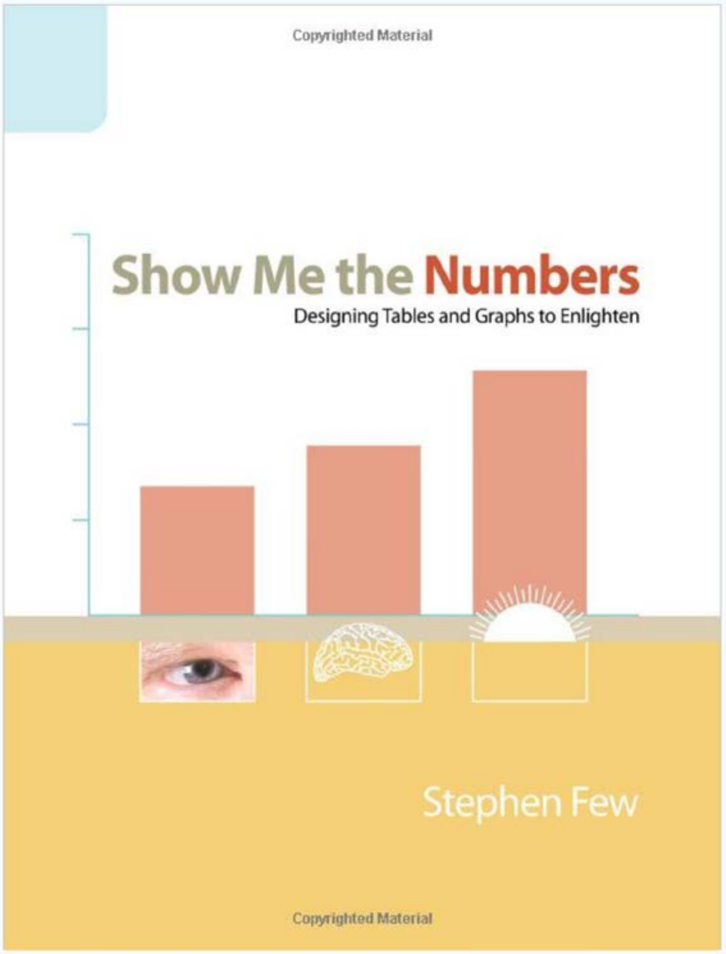
grey = values near the average of the class

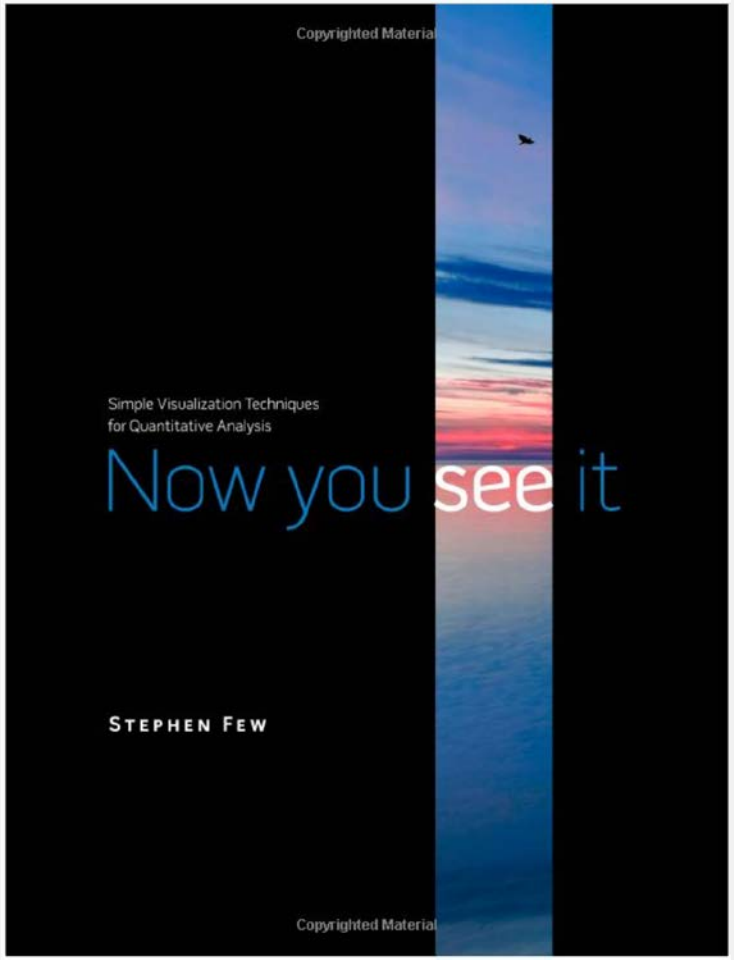
Reds = color below average

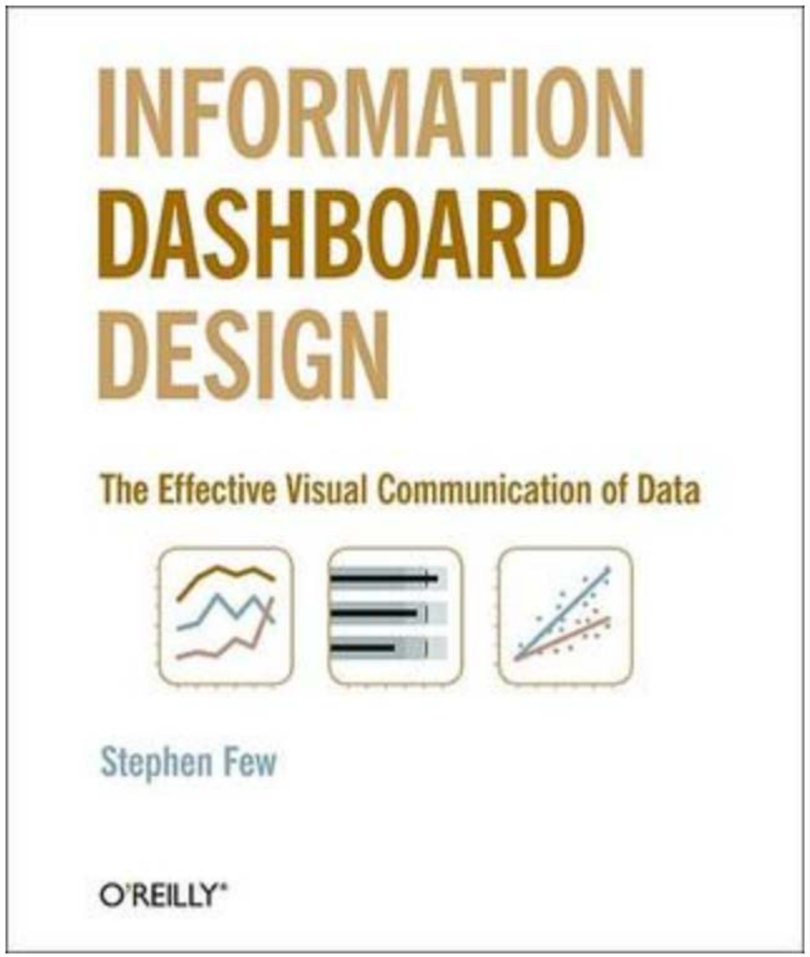
(the more saturated the lower)



Stephen Few







https://www.perceptualedge.com/

Nick Debarats

https://www.practicalreporting.com/about-nick-desbarats

How To Not Accidentally Create Data Visualizations That Lie

A.I. Experiments: Visualizing High-Dimensional Space (with TSNE)


That's all Folks!
Thanks!