SPATIAL DATA ANALYSIS

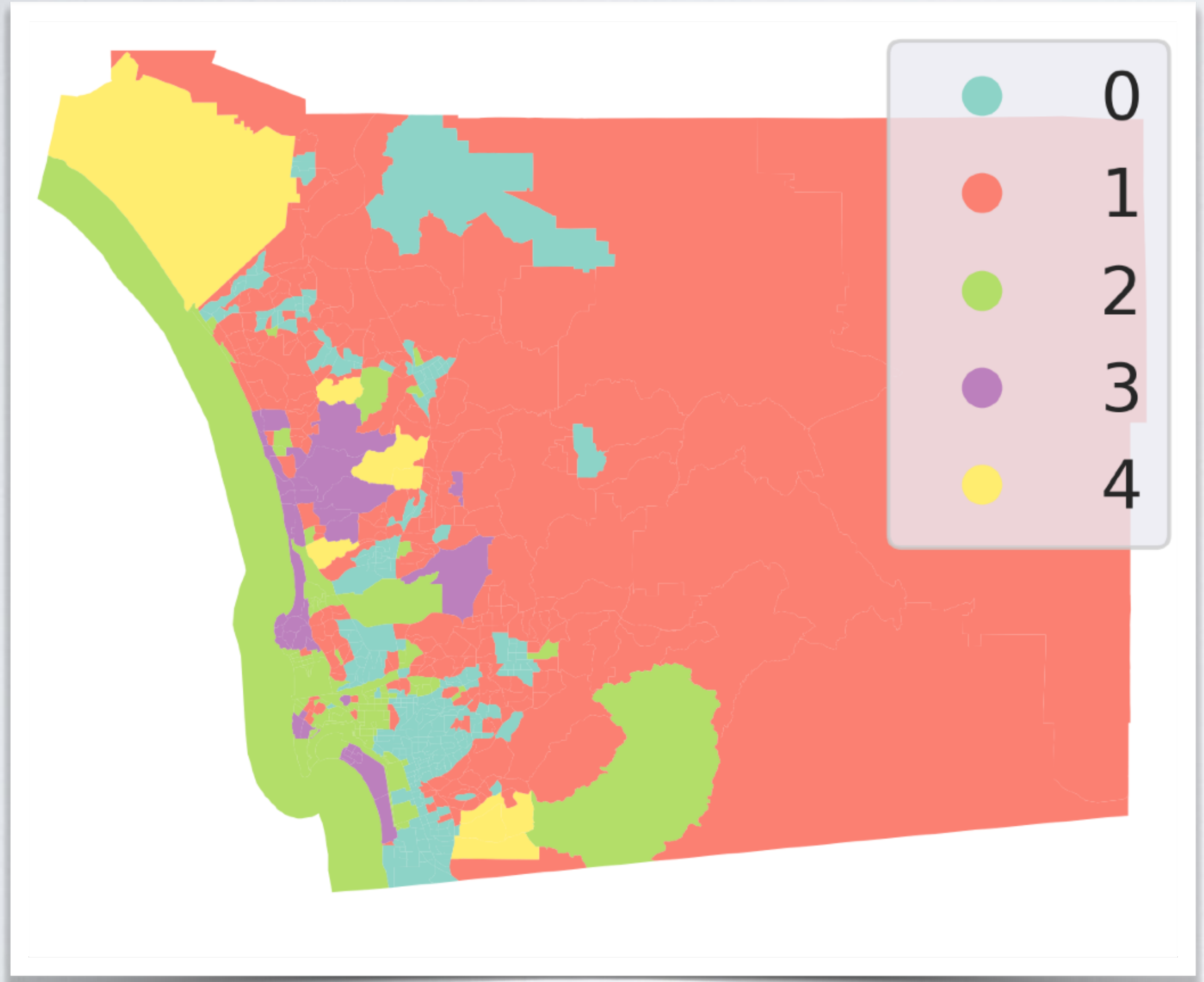
REGIONALIZATION

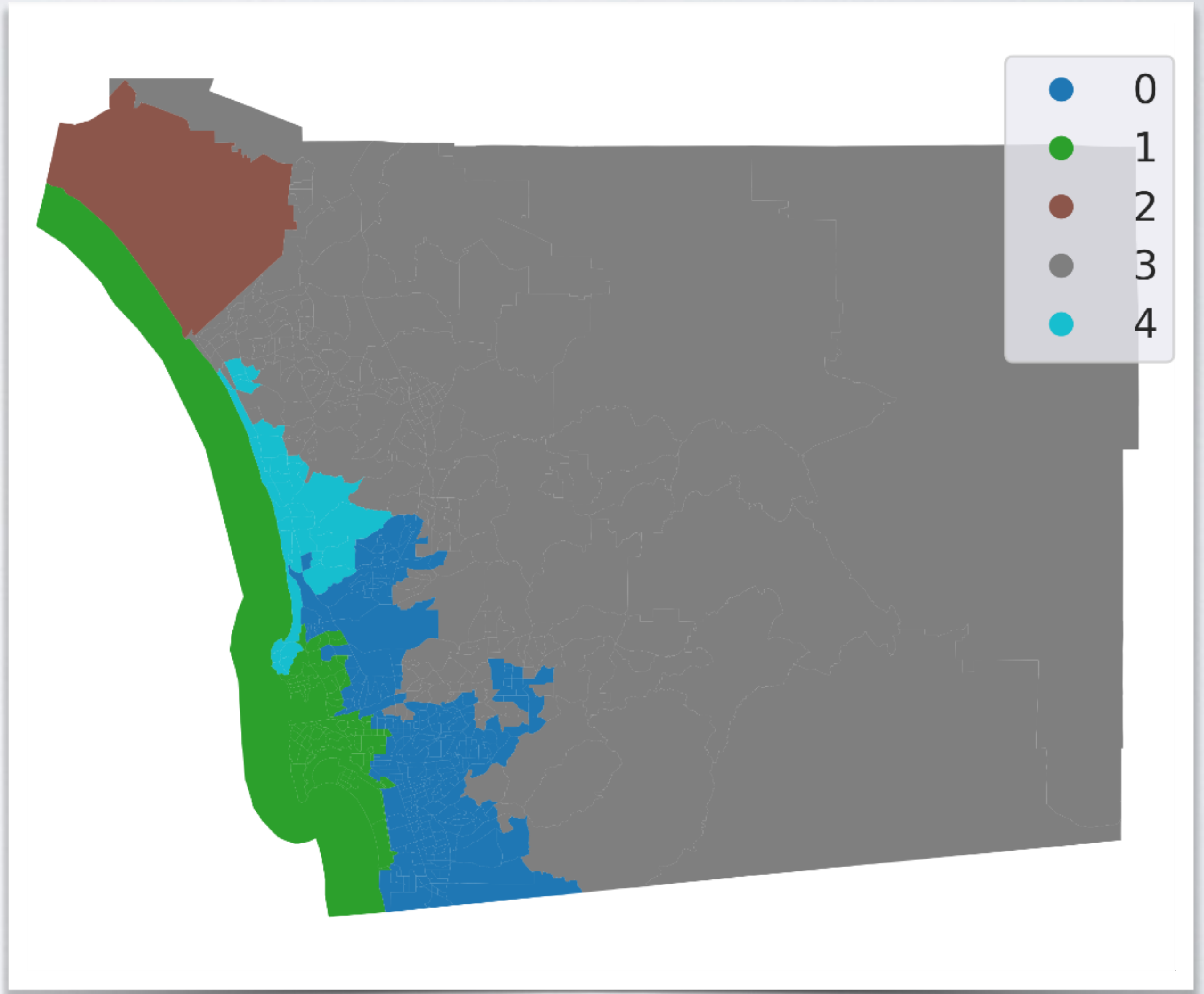
Spatial clustering

REGIONALIZATION

- Clustering: finding groups of similar observations
- If the data has a spatial structure, we might want the clusters to be contiguous in space
- =>Add a spatial constraint

REGIONALIZATION



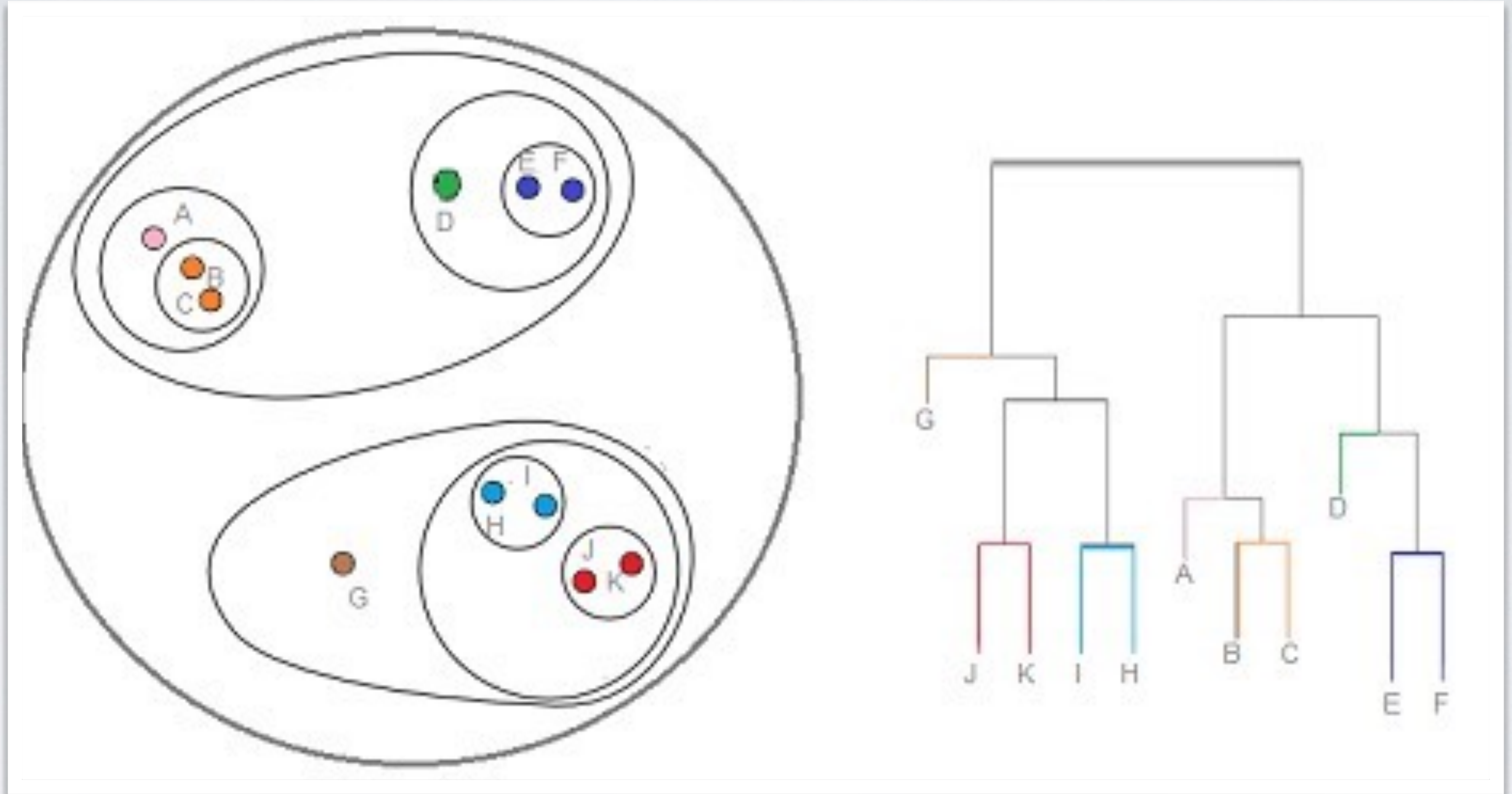


https://geographicdata.science/book/notebooks/10_clustering_and_regionalization.html

AGGLOMERATIVE CLUSTERING

- Define a notion of distance between two sets of points, e.g.
  - Minimal distance between sets elements
  - Average distance between elements
  - …
- Start with each item in its own cluster
- **While** nb_cluster > 1
  - Merge the two closest cluster

DENDROGRAM



https://www.statisticshowto.com/hierarchical-clustering/

CLUSTER DISTANCES

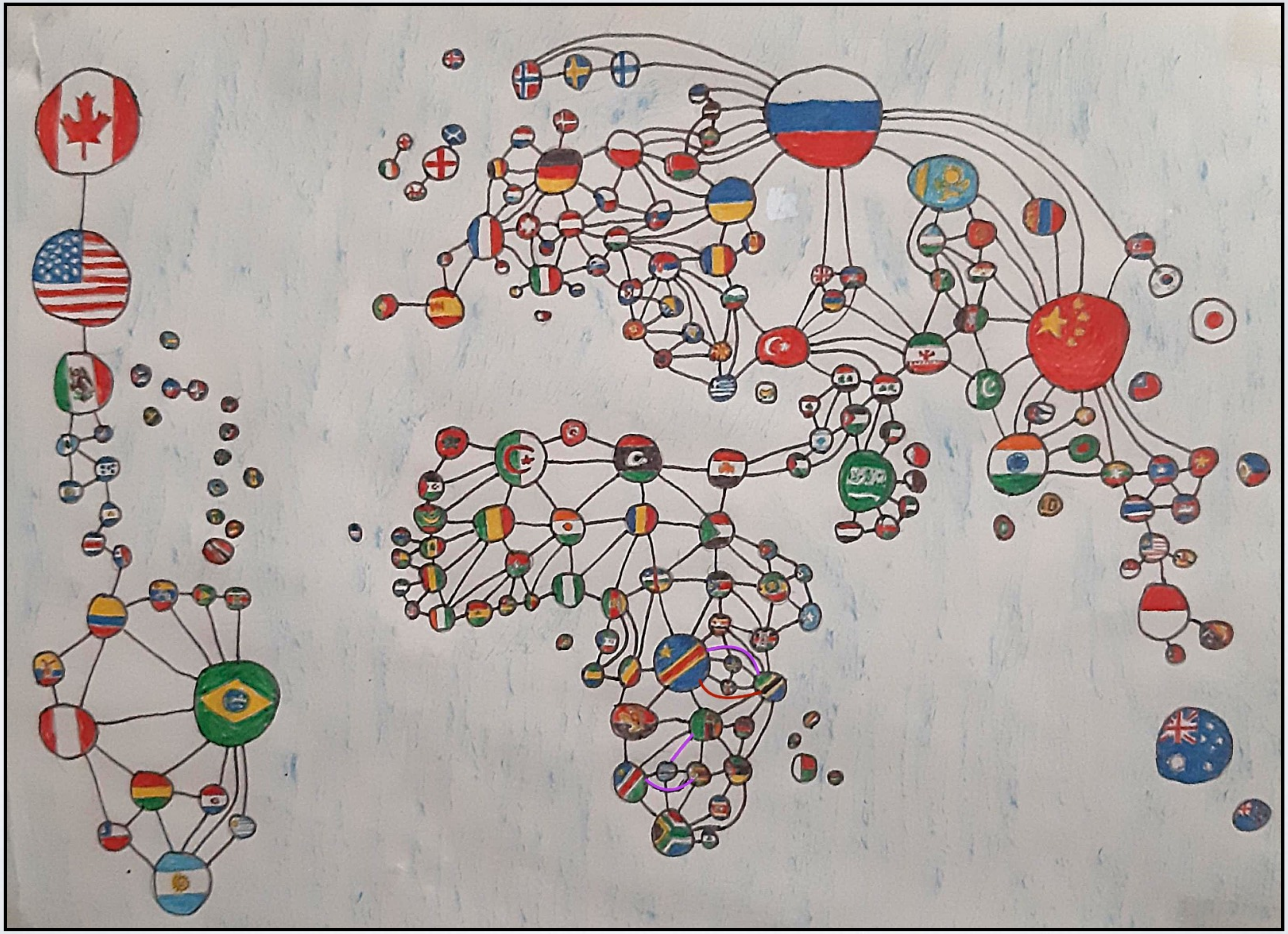
- Choose a distance function
  - Euclidean distance
  - Cosine distance
  - …

- Choose a cluster distance strategy
  - **single** uses the minimum of the distances between all observations of the two sets.
  - **complete** or 'maximum' linkage uses the maximum distances between all observations of the two sets.
  - **average** uses the average of the distances of each observation of the two sets.
  - **ward** minimizes the variance of the clusters being merged. (Within-Cluster Sum of Squares)
    - $\Delta \text{WCSS} = \text{WCSSnew} - (\text{WCSS}C_1 + \text{WCSS}C_2)$
    - Similar objective than k-means, but more greedy

REGIONALIZATION

- To discover spatial clusters, we want to allow merging only **spatially contiguous** clusters
- Solution: Connectivity matrix
  - A **graph** describing what element is a **neighbor** of another element.
  - Can merge only clusters with at least one edge between clusters

REGIONALIZATION



REGIONALIZATION

- Connectivity matrix (Binary graph)
  - Contiguity:
    - Contact between surface
    - Distance < threshold
  - KNN (K-nearest-neighbors)

- Spatial Weights Matrix (Weighted graph)
  - Put weights on edges
    - Inverse of the distance
    - Inverse of the squared distance…
  - Row normalized: sum of weights of neihgbors=1

REGIONALIZATION

- Other methods
  - K-means with constraints
    - Multiple variants
  - DBSCAN: principle of a graph with threshold…

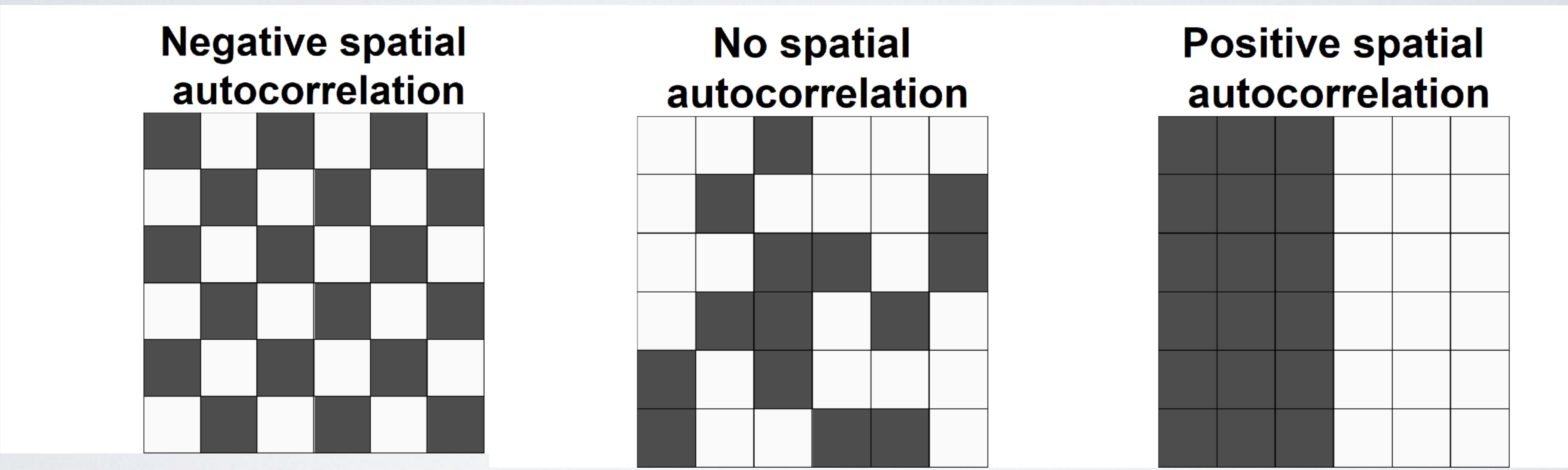
SPATIAL AUTOCORRELATION

Global

INTUITION

- Suppose you have attributes on observations
  - Binary (vote FOR/AGAINST, has covid cases or not, etc.)
  - Multi-label (candidate, type of apartments, etc.)
- Are those points distributed randomly/independently?
  - Or is there a correlation between the position of a point and the ones close to it
- Correlation between a variable and itself in space
  - =>Spatial autocorrelation

INTUITION


Negative spatial autocorrelation
No spatial autocorrelation
Positive spatial autocorrelation

INTUITION

- Using a Spatial Weights Matrix
  - w_{ij}: weight of edge (i, j)
- Spatial lag: $y_i^{sl} = \sum_j w_{ij} y_j$
  - With y_j the variable of interest
- Weighted average of neighbors

LINEAR SPATIAL AUTOCORRELATION

- Compute Pearson's linear correlation between
  - Value for observation x
  - Spatial lag for observation x
- In practice, people rather use Moran's I
  - Generalization to take into account:
    - Different # of neighbors
    - Different weights

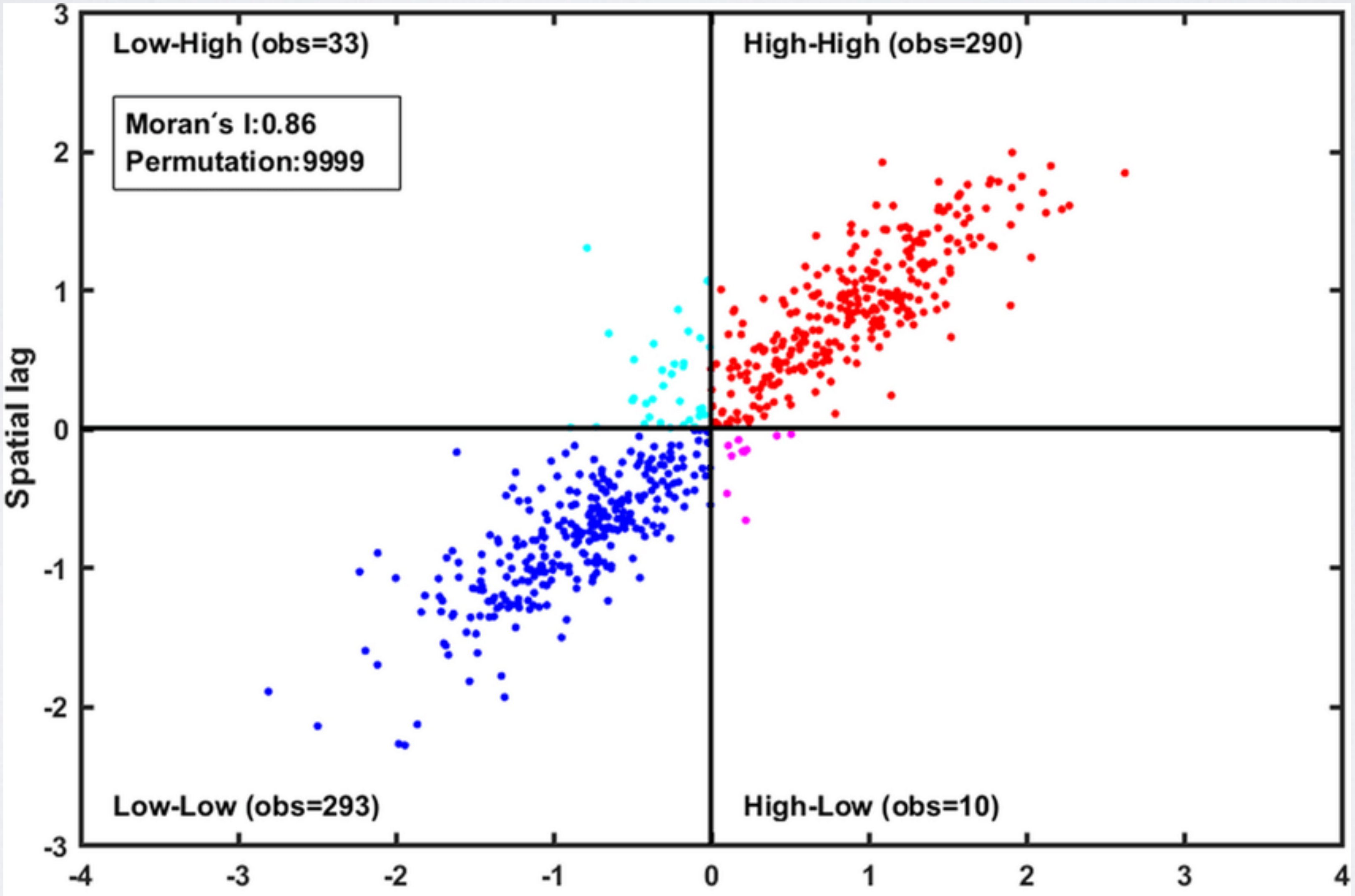
MORAN'S I

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} z_i z_j}{\sum_i z_i^2}$$

- w_{ij}: weight of edge (i, j)
- z_i: value at i, <u>standardized</u>
- n: nb. of observations

MORAN'S PLOT

Plot relation between standardized values



Moran's I is the slope of a linear regression on this plot

SPATIAL AUTOCORRELATION

Local

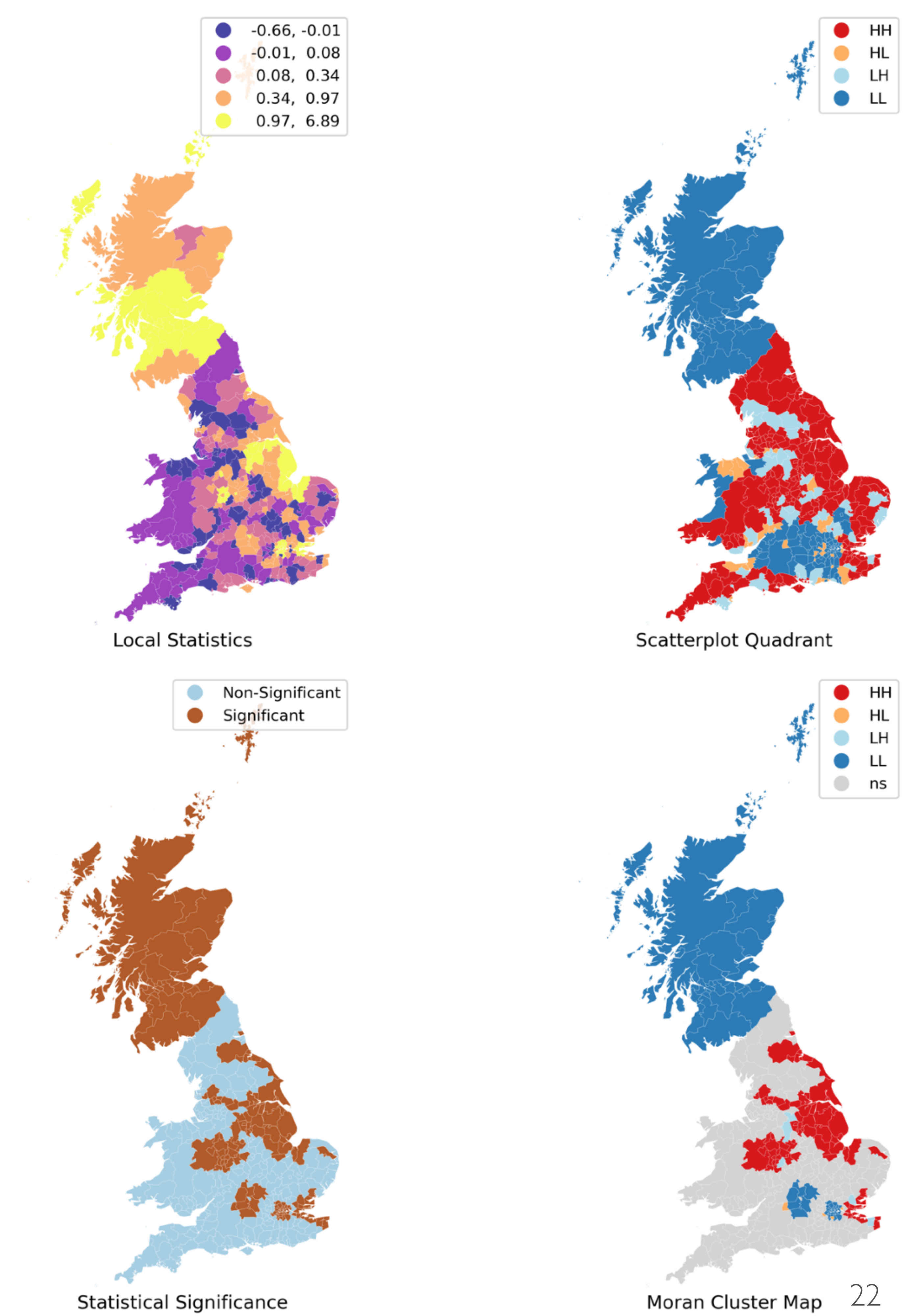
INTUITION

- Single scores are often misleading
- We can look at the details:
  - Where are positive/negative autocorrelations?
  - Where is the autocorrelation significant?
- Introduce LISA
  - Local Indicators of Spatial Association

LISA

- 1)Compute significance: Moran's Ii
  - $I_i = \frac{z_i}{m_2} \sum_j w_{ij} z_j \; ; \; m_2 = \frac{\sum_i z_i^2}{n}$
    - m_2: variance of the variable of interest
    - z_i: standardized value
  - Positive value: positive spatial correlation at this point
  - Negative value: negative spatial correlation at this point
  - 0 or close to 0: no significant spatial autocorrelation

Brexit vote example (Support for Brexit)



HH: Hot spots
LL: Cold spots
LH: doughnuts
HL: diamonds in the rough

https://geographicdata.science/book/notebooks/07_local_autocorrelation.html

TEMPORAL DATA ANALYSIS

TIME SERIES

- Consider a time series
  - A variable evolving with time
    - Price of something, etc.

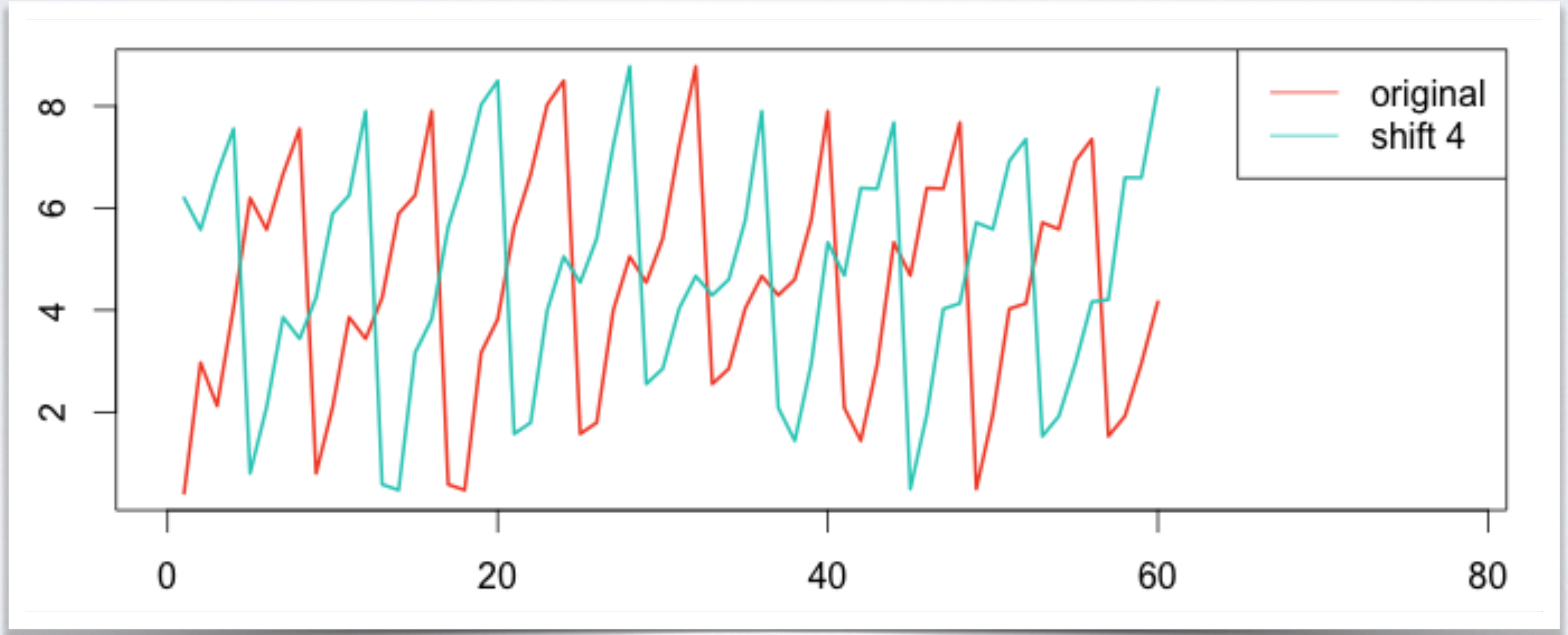
- Multivariate time series
  - Multiple time series for multiple variables
    - Price of multiples cryptocurrencies
    - For a pro-player, statistics of game-performance…
    - Etc.

AUTOCORRELATION

- Intuition: are values at time t correlated with values at $t + \Delta_t$
  - With Δ_t a shift
- Objective a bit different from spatial
  - Not an evaluation of similar to "neighbors"
  - But is there a typical "lag" at which we observe repeated patterns

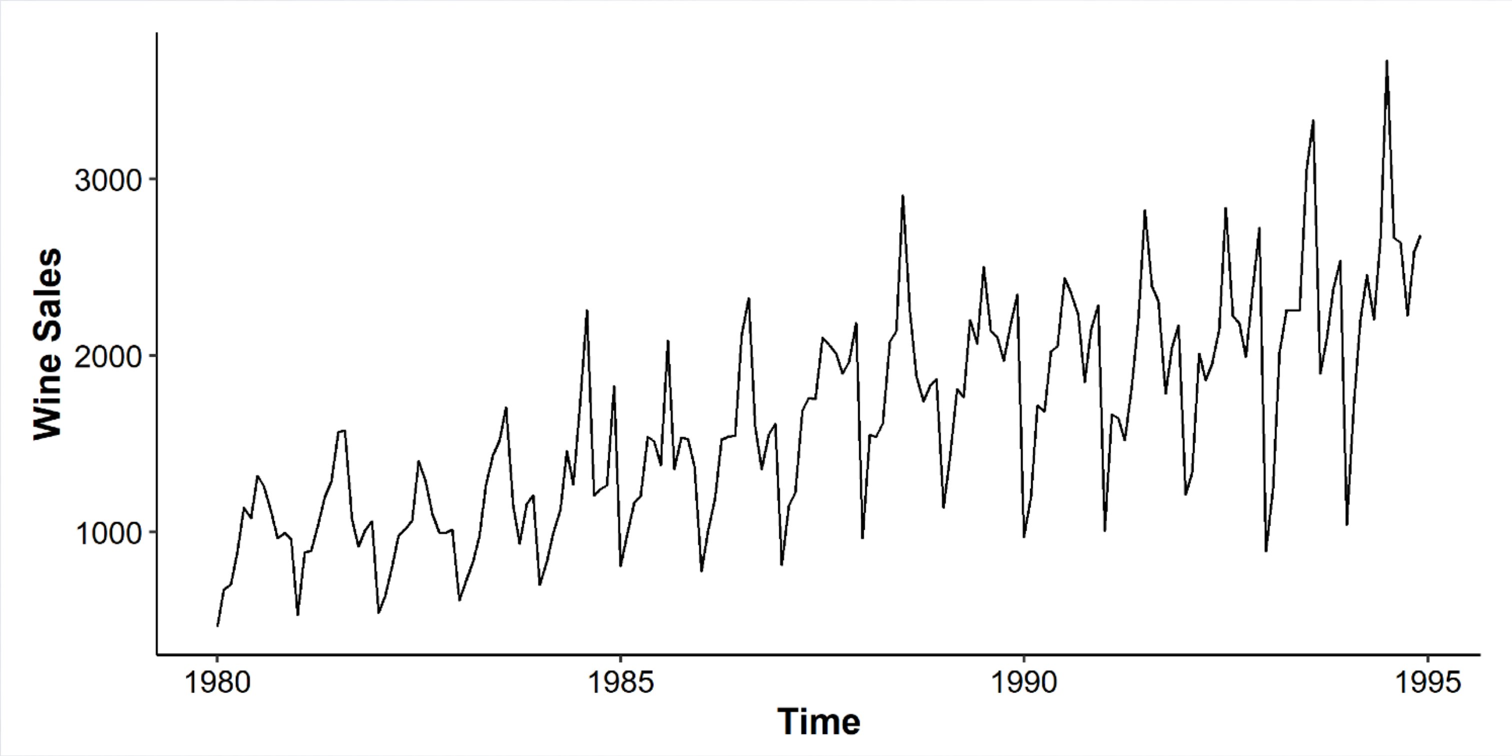
AUTOCORRELATION

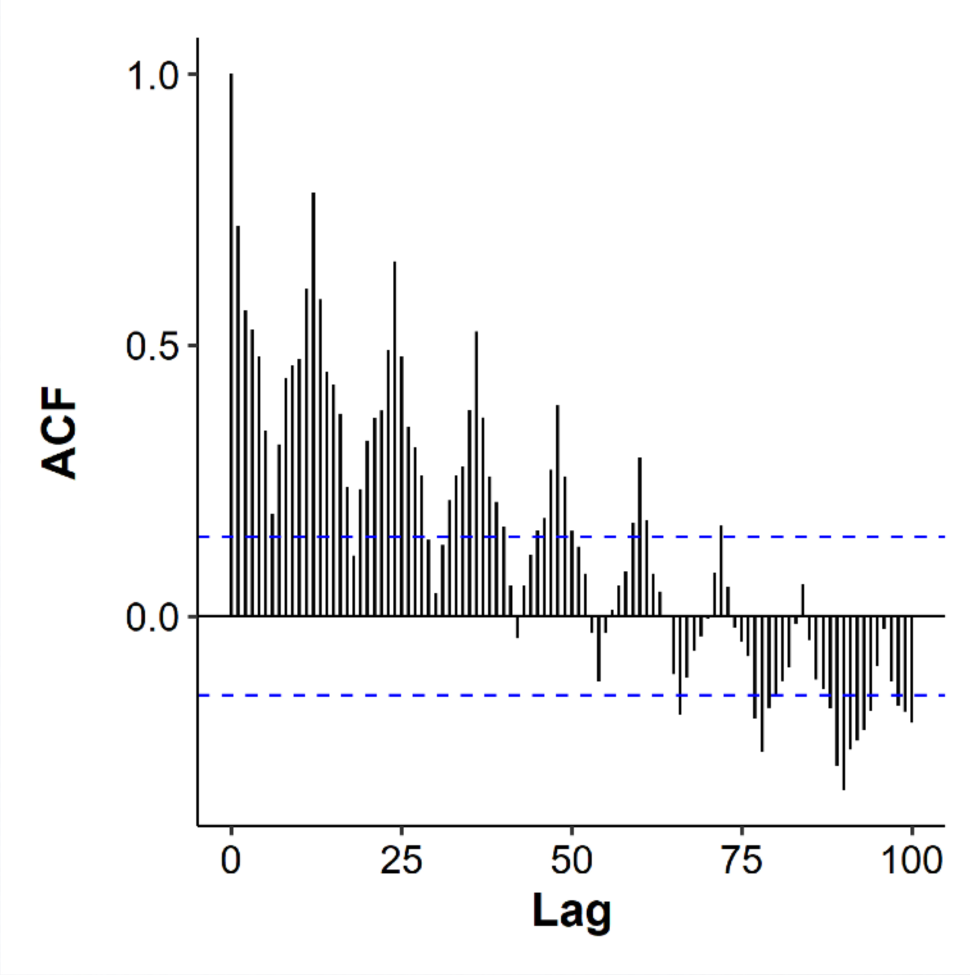
- Typical approach: linear correlation (Pearson) between
  - The time series
  - The shifted time series, with shift Δ_t



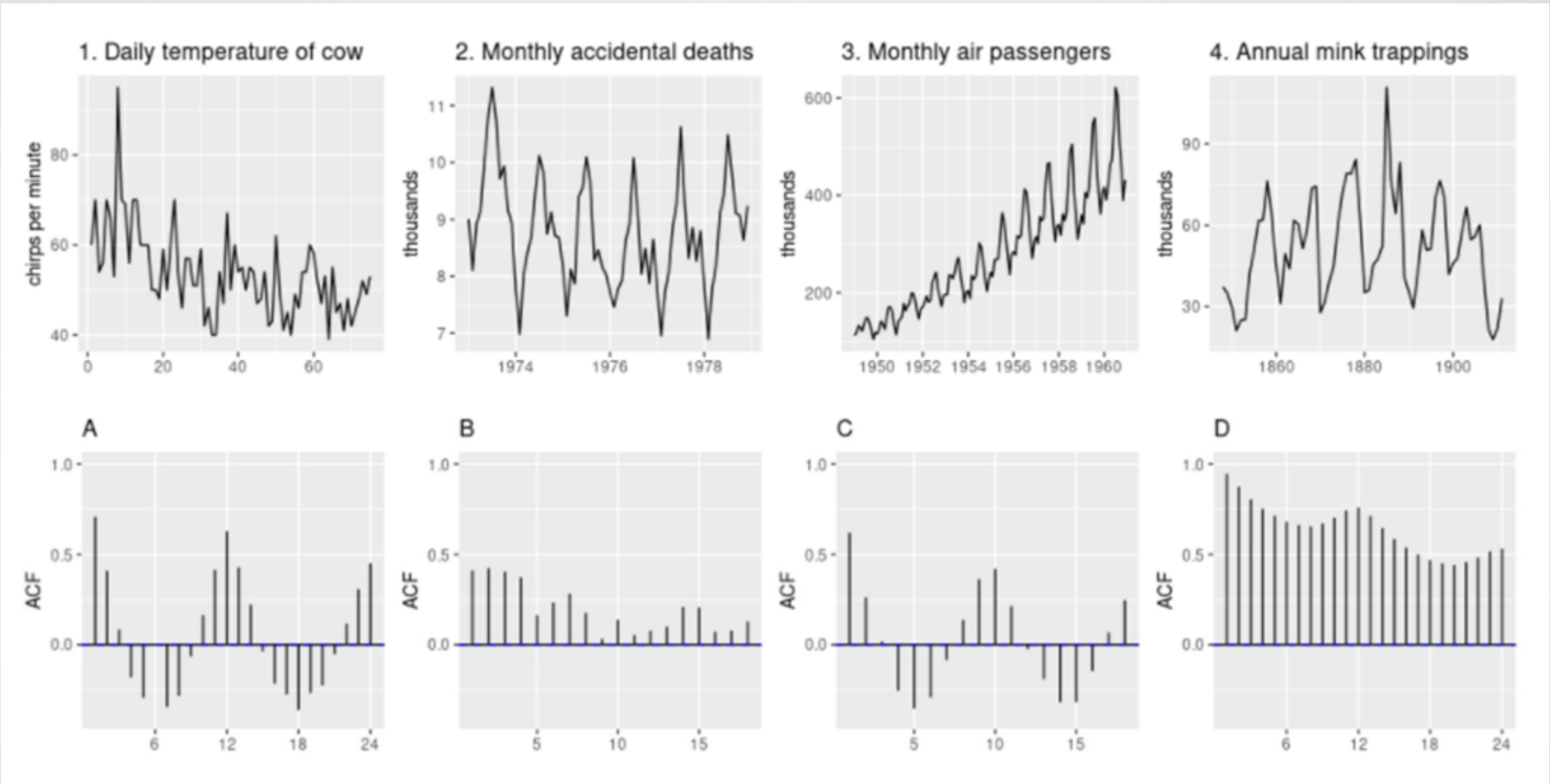
AUTOCORRELATION

- Finding seasonal/periodic patterns:
  - ACF: AutoCorrelation function: autocorrelation score for each lag





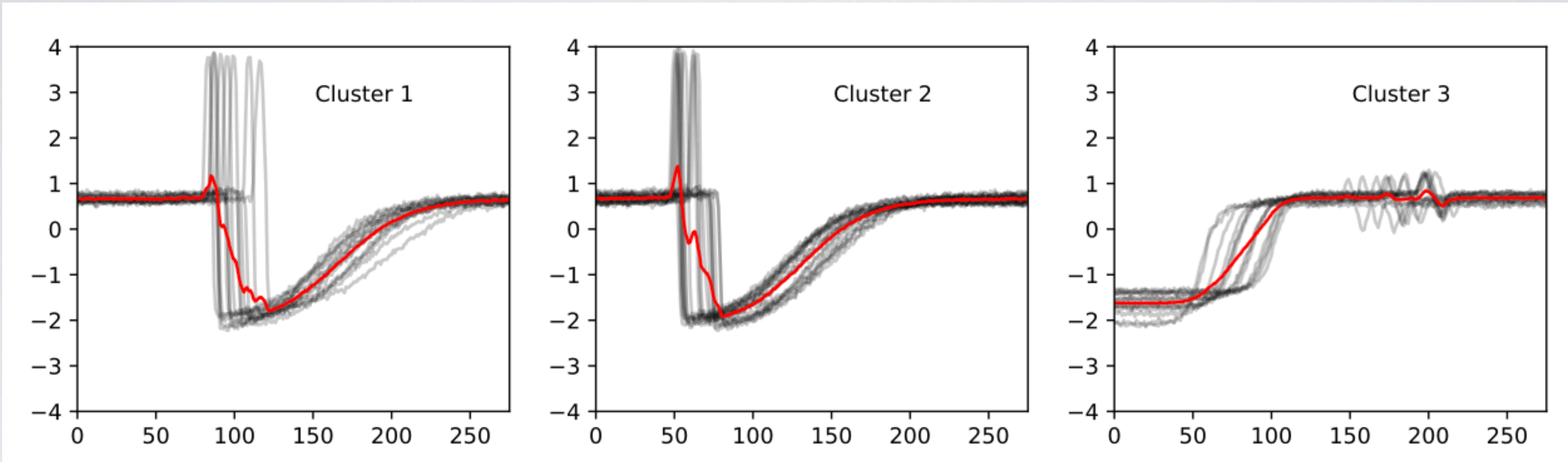
AUTOCORRELATION



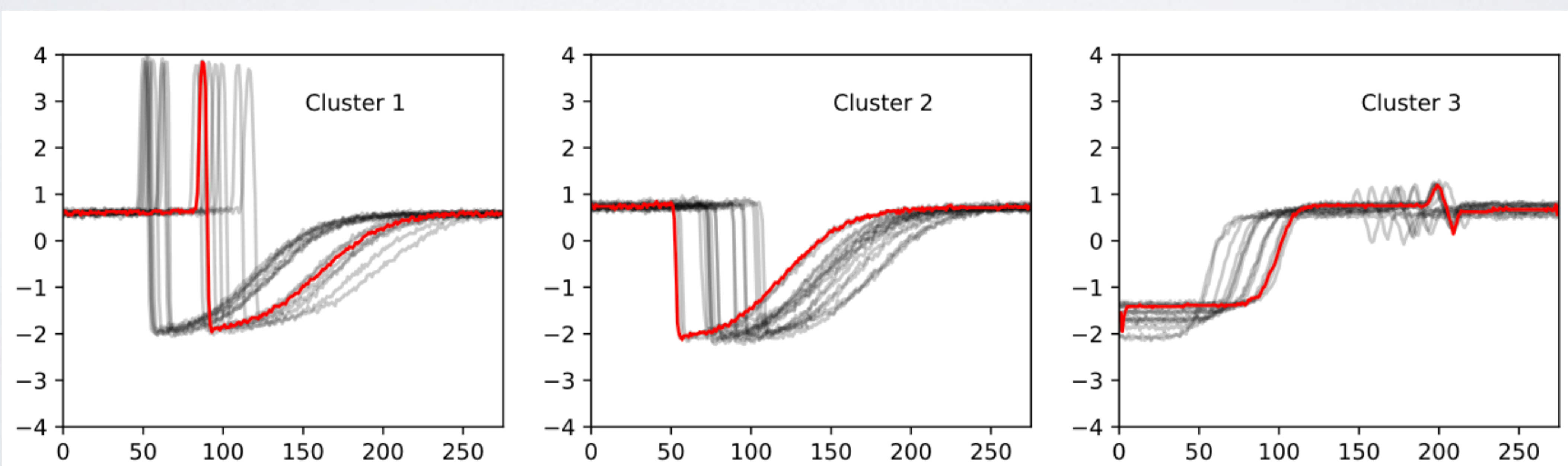
CLUSTERING

- Clustering multiple time series
  - Number of items sold per week for different products
    - Find products with a similar selling lifecycle
- If time series are well-aligned
  - Each time series is a vector
    - Use k-means. Time series having similar values at the same time will be clustered together
- Problem if some time series start at a different time, or last longer

Without time warping

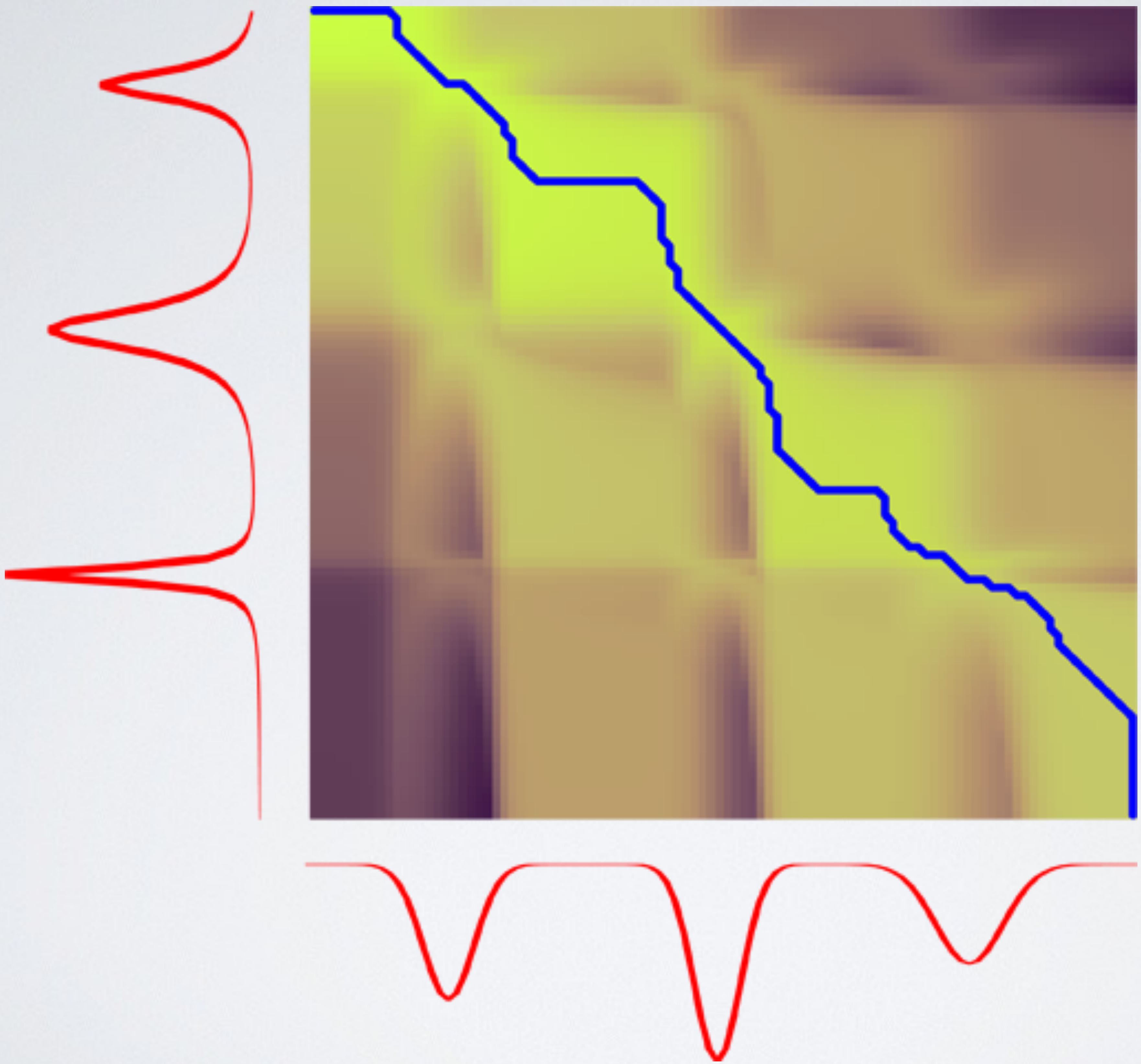


With time warping



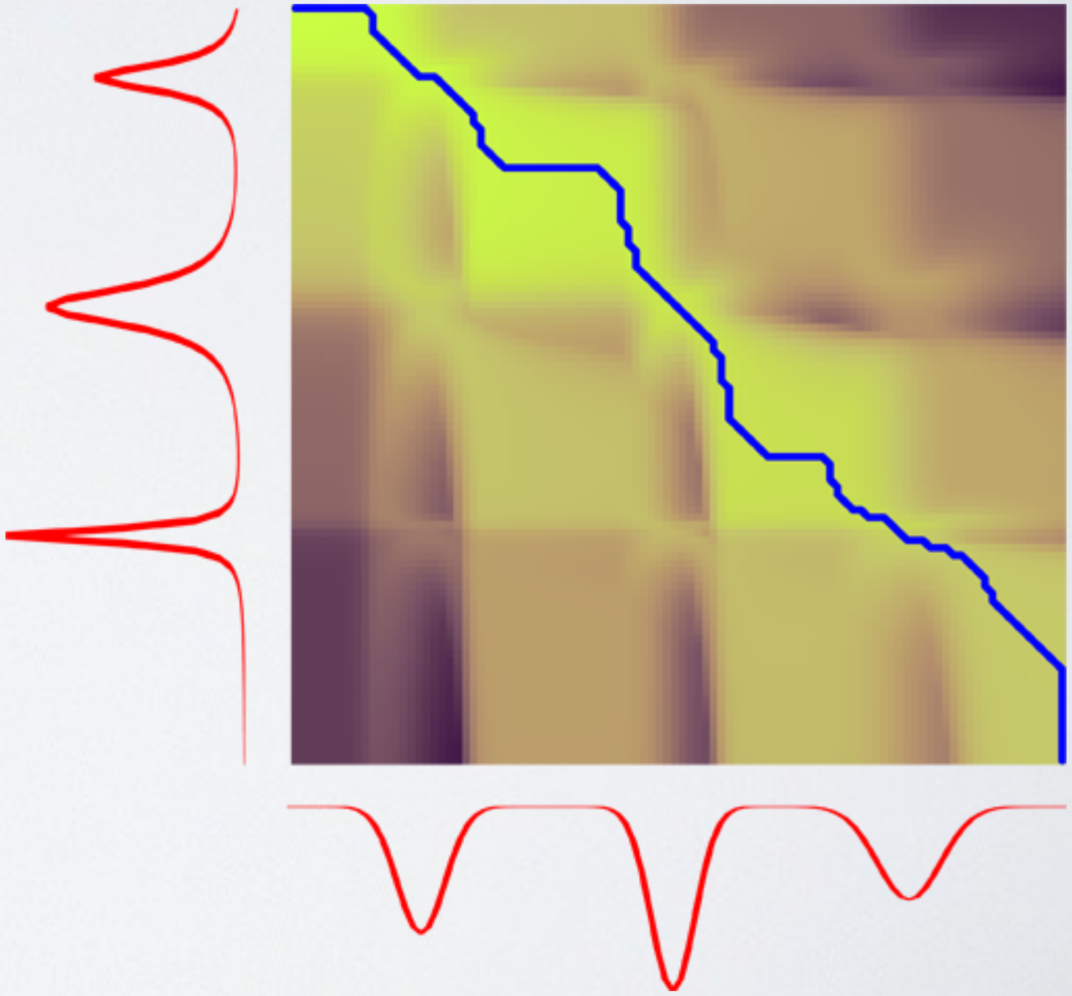
TIME WARPING

- Find an optimal alignment
  - Non-linear transformation
- Step 1: build a matrix of distance between each timestep in each time series
  - Times series of length m and n
    - Matrix of size $m \times n$

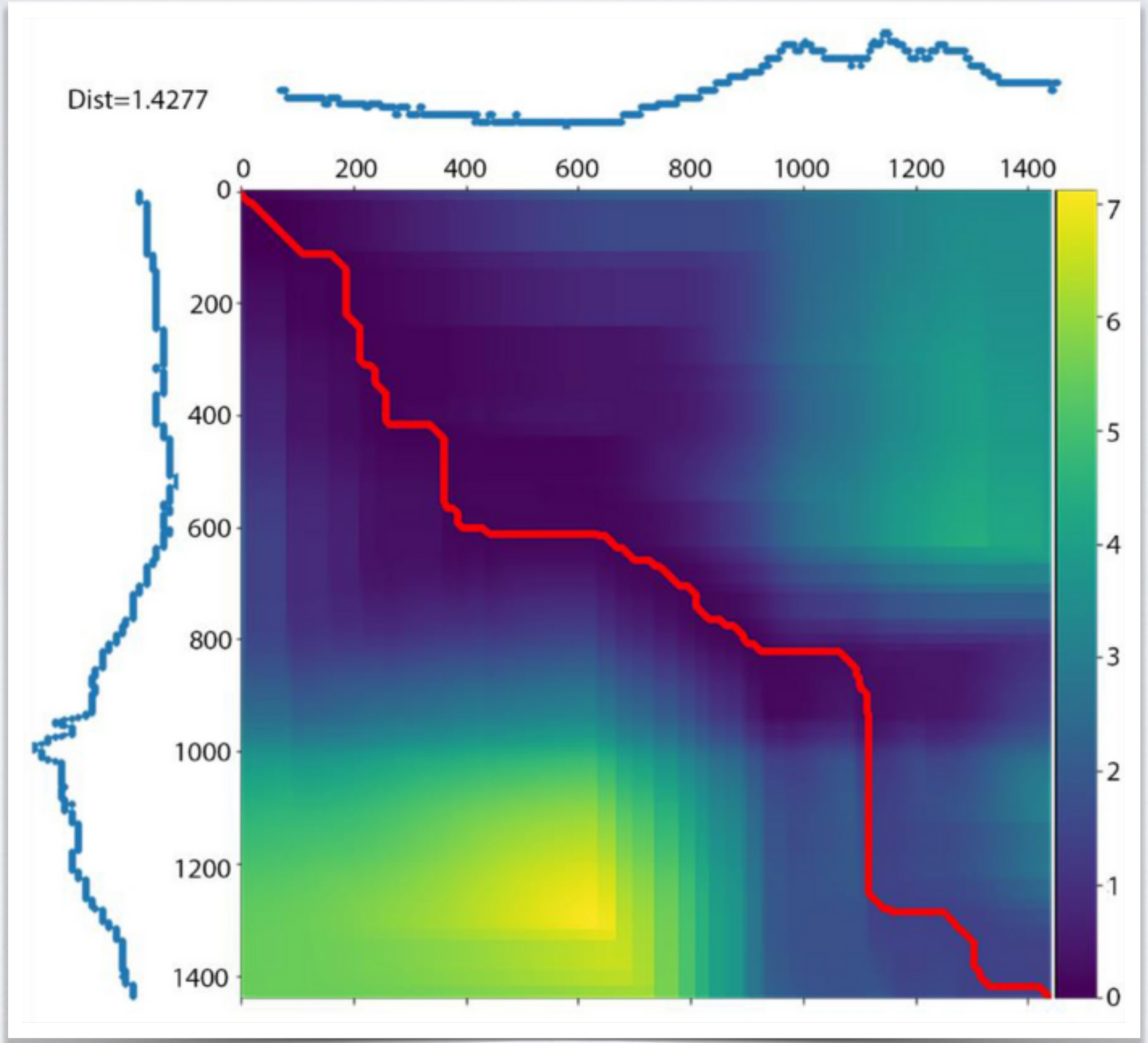


TIME WARPING

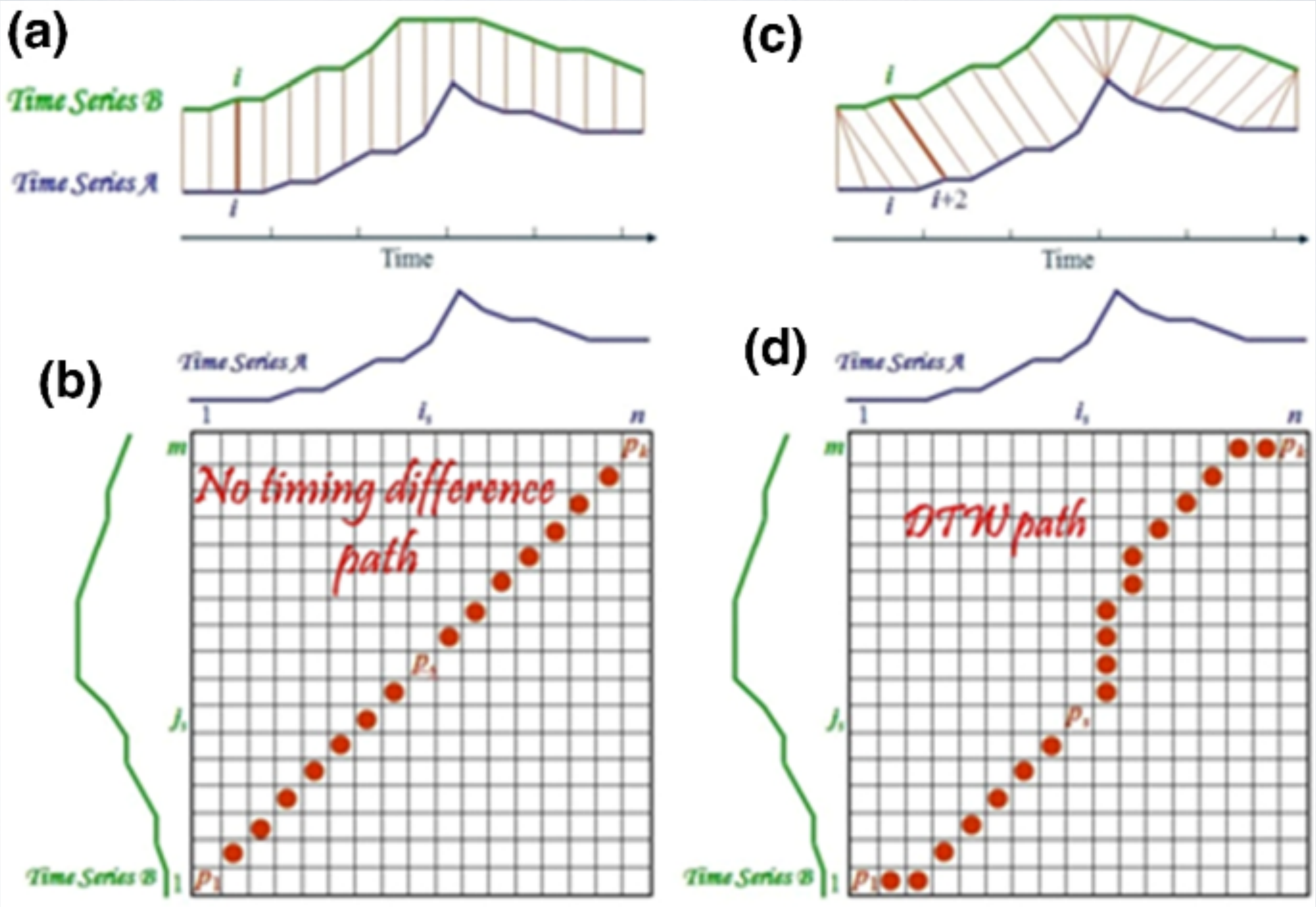
- Values in the matrix are "penalties"
- Find an optimal path in this matrix:
  - 1)Minimize the sum of penalties
  - 2)continuous line
  - 3)monotonous (never go up)



TIME WARPING



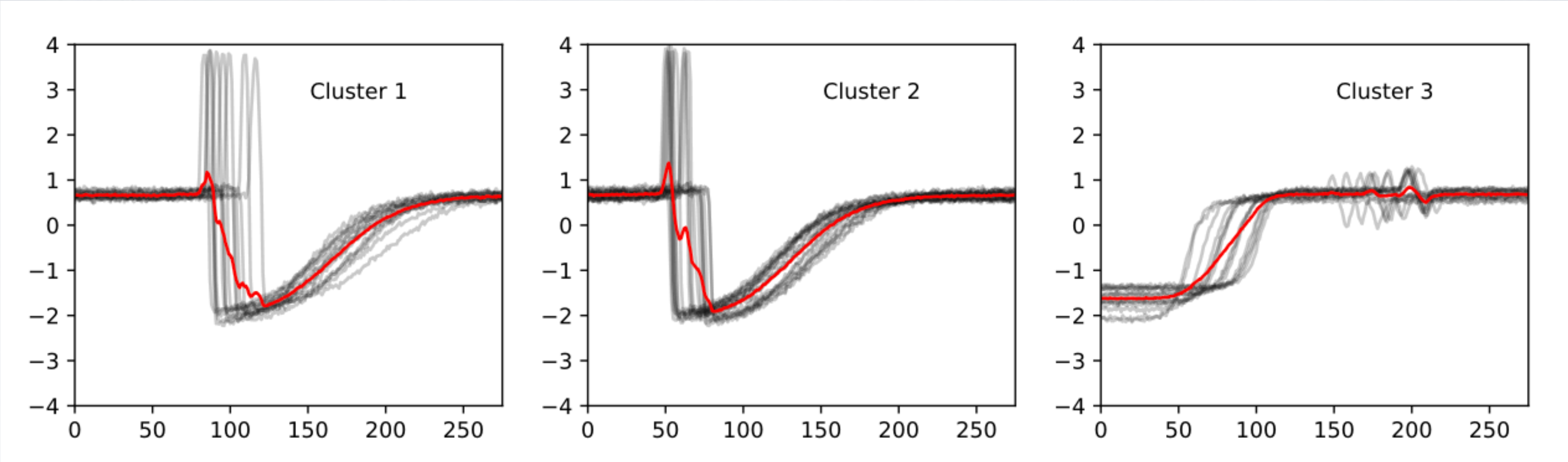
TIME WARPING



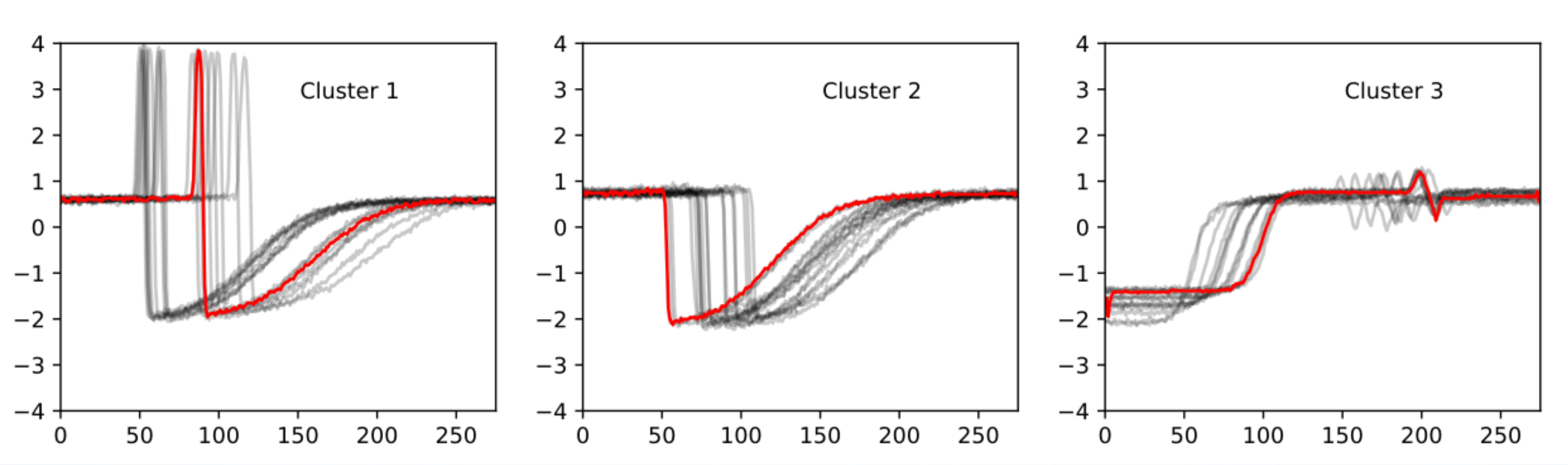
FINDING OPTIMAL PATH

- Finding an optimal path is costly for long time-series
- Exact approach: Dijkstra algorithm formulation
  - Improved by pruning
- Greedy approaches: FastDTW
  - Add constraint, acceptable lost, coarsening…

Without time warping



With time warping

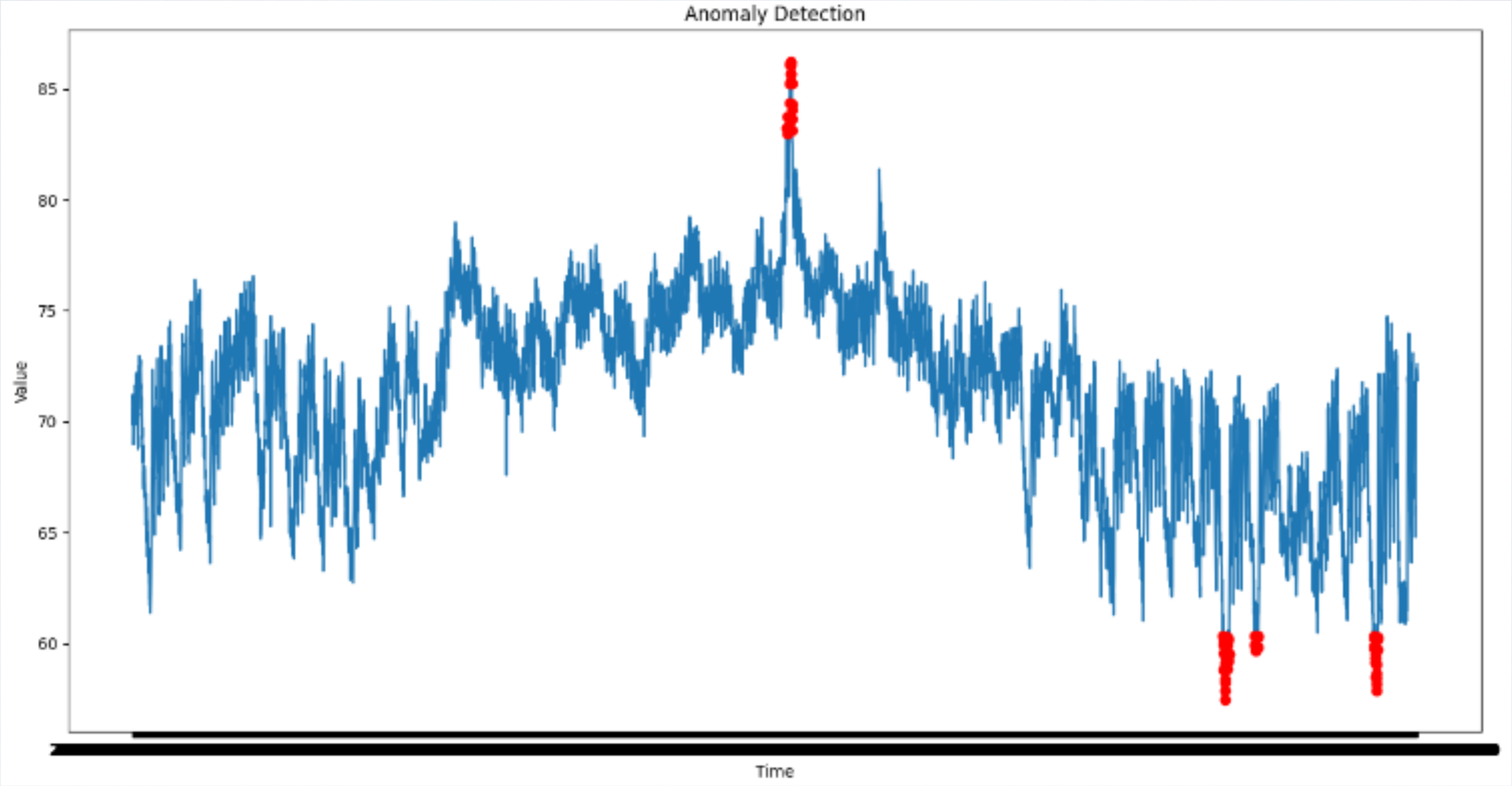


ANOMALY DETECTION

INTUTION

- We would like to find anomalous points in a time series
- General principle of anomaly detection:
  - Make a "prediction" of the expected value
  - An anomaly is a point that differ strongly from a prediction
- Simplest approach: moving average

EXAMPLE

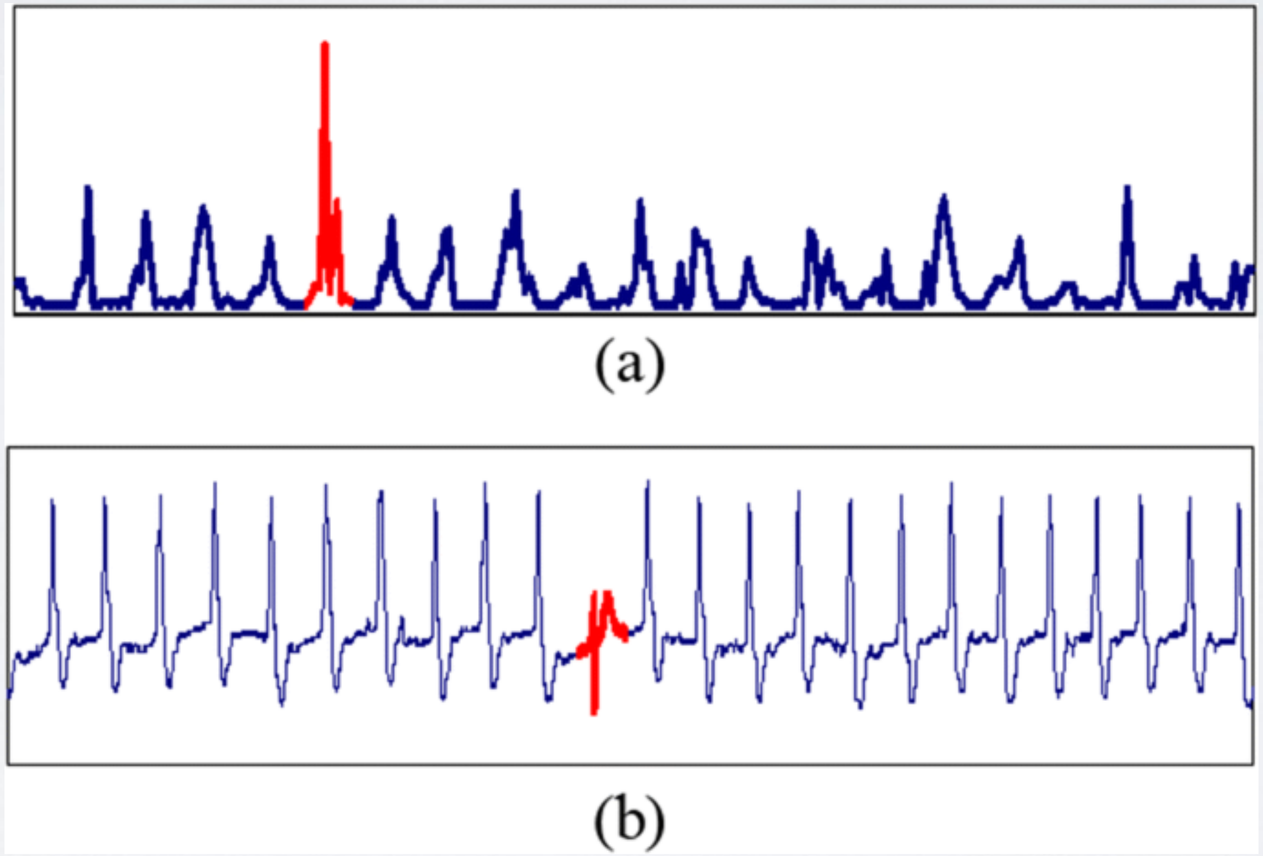


MOVING AVERAGE ANOMALY DETECTION

- 1)Compute a **moving average** to smooth the time series
  - Choose an appropriate time window Δ_t…
- For a point at t, we have a reference: all points in $[t - \frac{\Delta_t}{2}, t + \frac{\Delta_t}{2}]$
- Use a statistical test to evaluate exceptionality
  - For instance, 3 standard deviations from the mean, assuming normality…

MOVING AVERAGE ANOMALY DETECTION

Do not work in complex cases



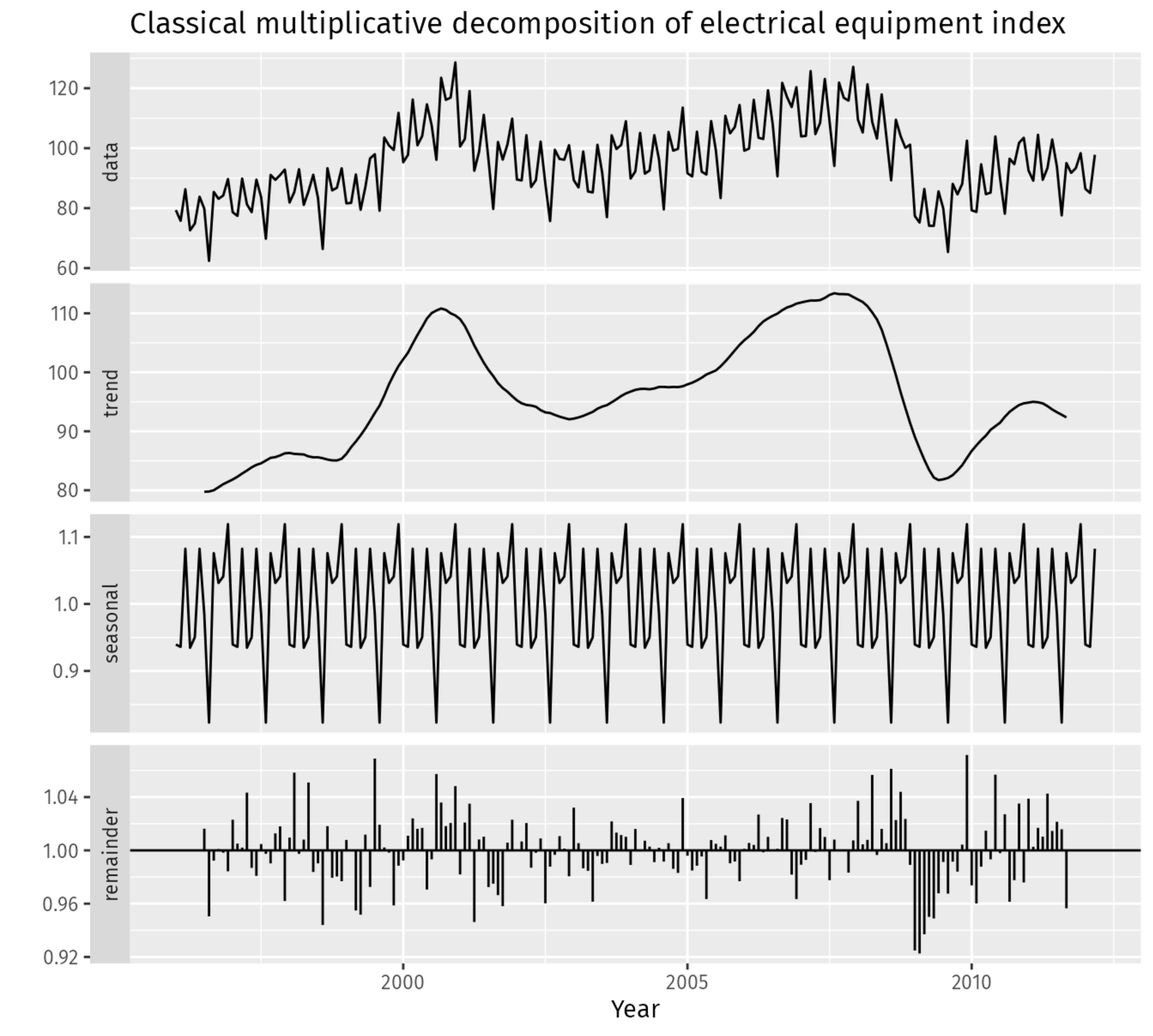
Needs better estimate of expected value

TIME SERIES DECOMPOSITION

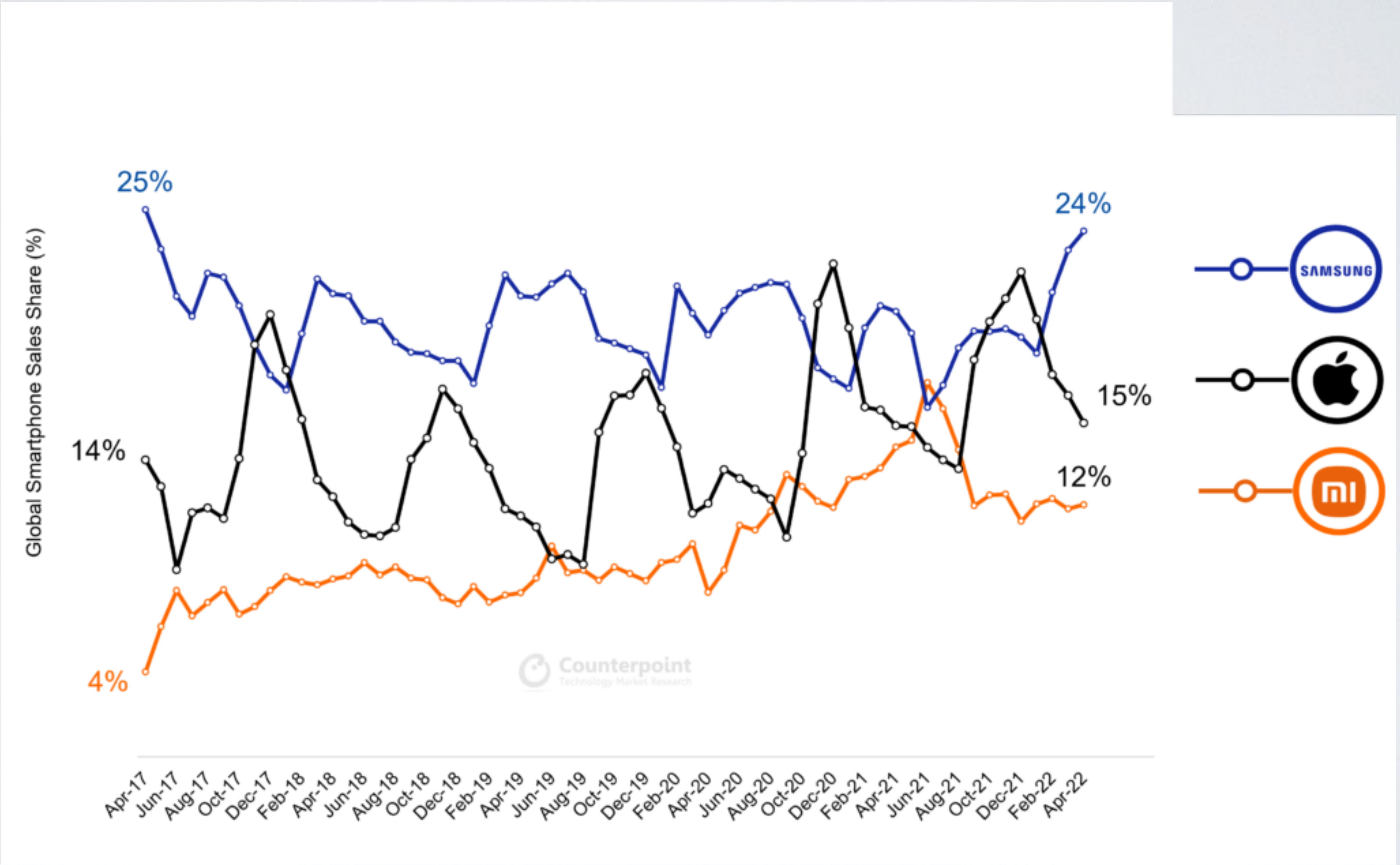
INTUITION

- We assume that a time series is the addition of 3 factors:
  - 1)A **trend**. This is the main global change of the variable
    - e.g.: smartphone brand sales: adoption by more people, more or less popular, etc.
  - 2)A **seasonal** component
    - e.g.: every Christmas, people buy more smartphones
  - 3)A **reminder**: what is not explained by those two factors

INTUITION


Classical multiplicative decomposition of electrical equipment index
data
trend
seasonal
remainder
120
100
80
60
110
100
90
80
1.1
1.0
0.9
1.04
1.00
0.96
0.92
2000
2005
2010
Year

INTUITION



Global Smartphone Sales Share (%)
25%
24%
14%
15%
12%
4%
SAMSUNG
Counterpoint
Technology Market Research
Apr-17
Jun-17
Aug-17
Oct-17
Dec-17
Feb-18
Apr-18
Jun-18
Aug-18
Oct-18
Dec-18
Feb-19
Apr-19
Jun-19
Aug-19
Oct-19
Dec-19
Feb-20
Apr-20
Jun-20
Aug-20
Oct-20
Dec-20
Feb-21
Apr-21
Jun-21
Aug-21
Oct-21
Dec-21
Feb-22
Apr-22

HOW TO

- Classical decomposition of time series
  - Choose a relevant time scale Δ_t, e.g., year, month… (e.g., Using ACF plot)
- 1)Compute trend using a sliding window Δ_t
- 2)Compute the detrended time series
  - Time series - trend
- 3)Compute the average season, i.e., average values on each window Δ_t
- 4)Remove the average season from the detrended time series
  - What remains is the **reminder/residuals**

HOW TO

- Classical decomposition of time series
  - One can evaluate the relevance of the Δ_t period by computing the similarity between seasons
- We can replace the additive model with a multiplicative model
  - $y_t = T_t + S_t + R_t$
  - $y_t = T_t \times S_t \times R_t$

HOW TO

- More advanced approaches exist
  - STL decomposition
  - SARIMA (ARIMA with seasonality)
  - Facebook Prophet
    - $y_t = T_t + S_t + H_t + R_t$
    - H_t corresponds to holidays or special events
    - T is a linear/logistic function with change points, to predict the future
    - S is a Fourier series, i.e., a sum of sinusoidal signals
    - The model parameters are fitted using a method similar to likelihood maximization (remember Gaussian mixtures?)