Chapter XII: Data Pre and Post Processing

1. Data Normalization
2. Missing Values
3. Curse of Dimensionality
4. Feature Extraction and Selection
    4.1. PCA and SVD
    4.2. Johnson–Lindenstrauss lemma
    4.3. CX and CUR decompositions
5. Visualization and Analysis of the Results
6. Tales from the Wild

Zaki & Meira, Ch. 2.2, 2.4, 6 & 8

XII.5: Visualization and Analysis

1. Visualization techniques
    1.1. Projections onto 2D or 3D
    1.2. Other visualizations
2. Analysis of the Results
    2.1. Significance
    2.2. Stability
    2.3. Leakage

Visualization Techniques

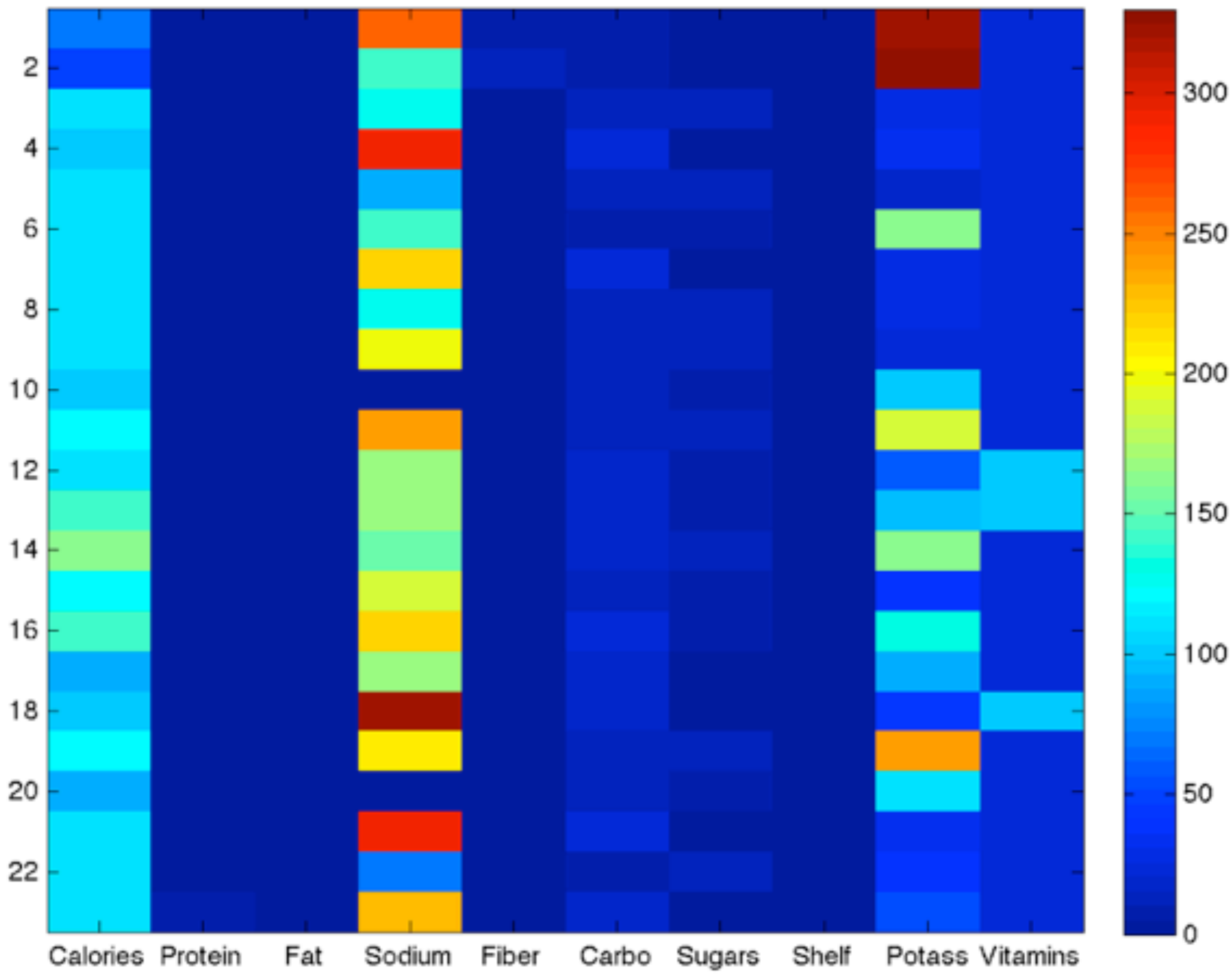
- **Visualization** is an important part of the analysis of the data and the results
  - Good visualization can help us see patterns in the data and verify whether our found results are valid
  - Visualization also helps us to interpret the results
- Visualization can also lead us seeing patterns that are not (significant) in the data
  - Visualization alone can never be the basis of analysis

Projecting multi-dimensional data

- The most common visualization takes *n*-dimensional data and projects it into 2 or 3 dimensions for plotting
  - Different methods retain different type of information
- We've already seen few projections
  - SVD/PCA can be used in multiple ways
    - Either project the data in the first singular vectors
    - Or do a singular vector scatter plot
- Creating good projections is an on-going research topic

Example: Cereal data

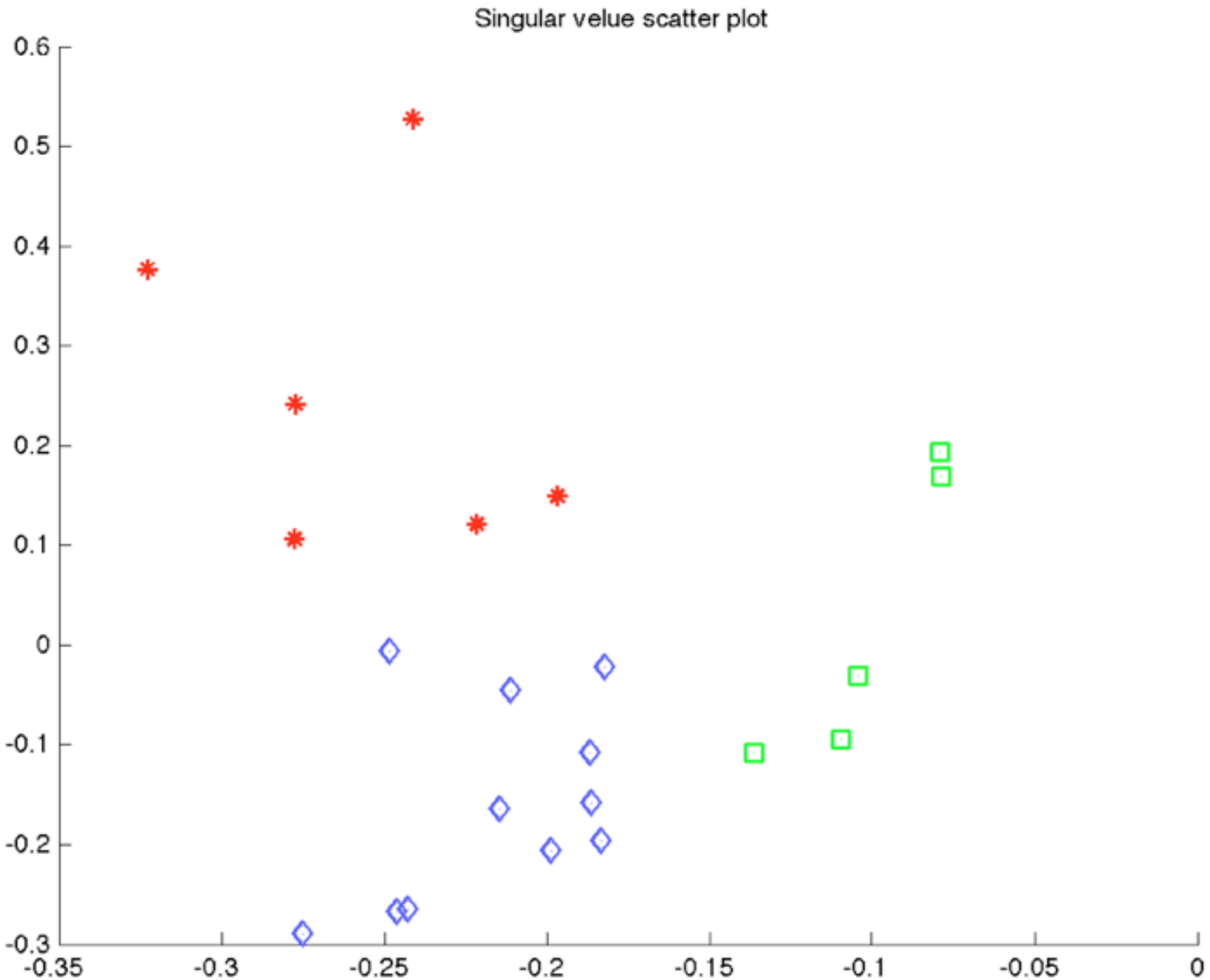
- Data of 77 different cereals
  - http://lib.stat.cmu.edu/DASL/Datafiles/Cereals.html
  - We use only 23 Kellogs manufactured cereals in the examples



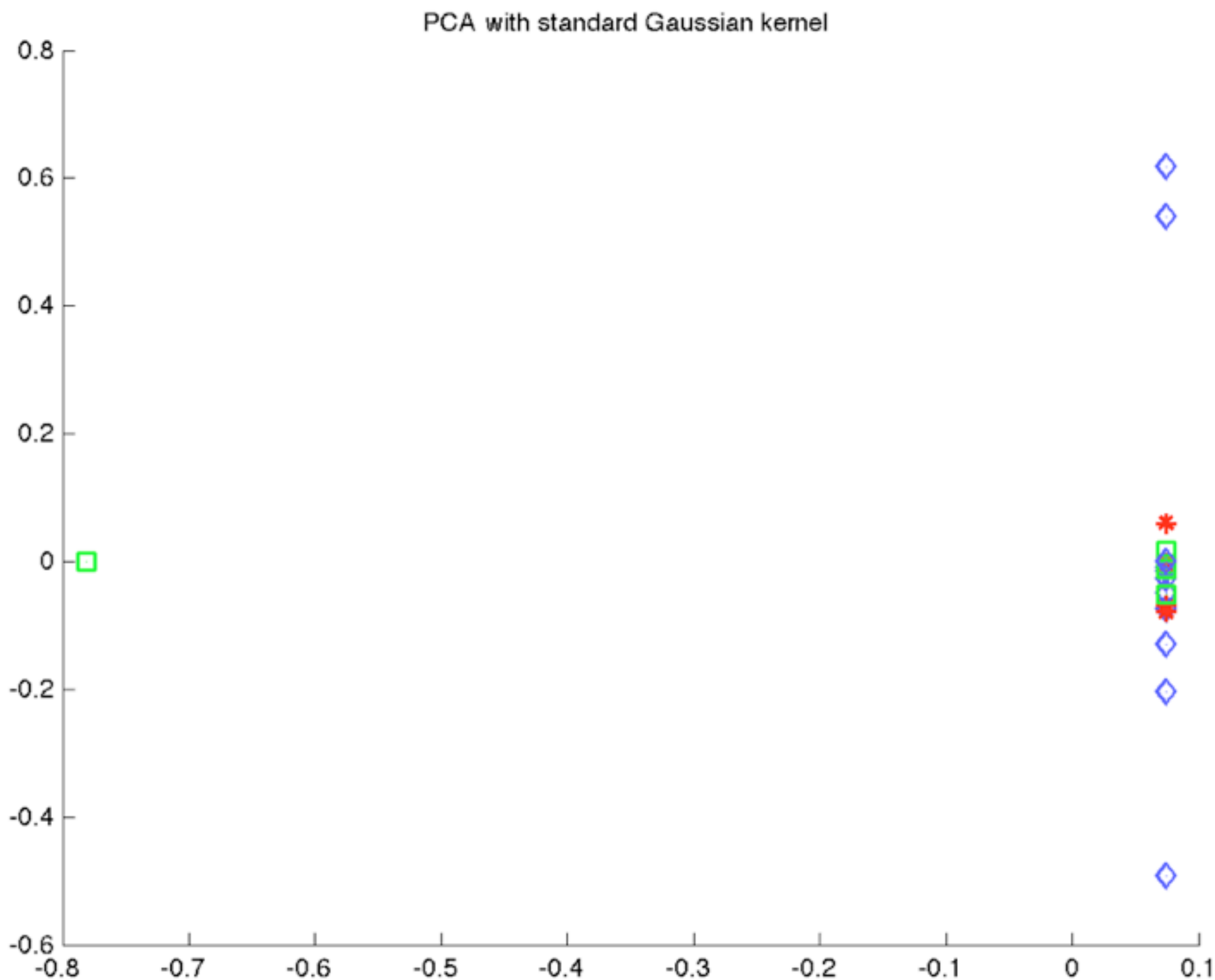
Example: Clustering

- We clustered the Cereal data using *k*-means
  - But is the clustering meaningful?
  - How do we plot a clustering?
- One idea: project the data into 2D and mark which point belongs to which cluster
  - Question: will we see the clustering structure?

Cereals in SVD Scatter

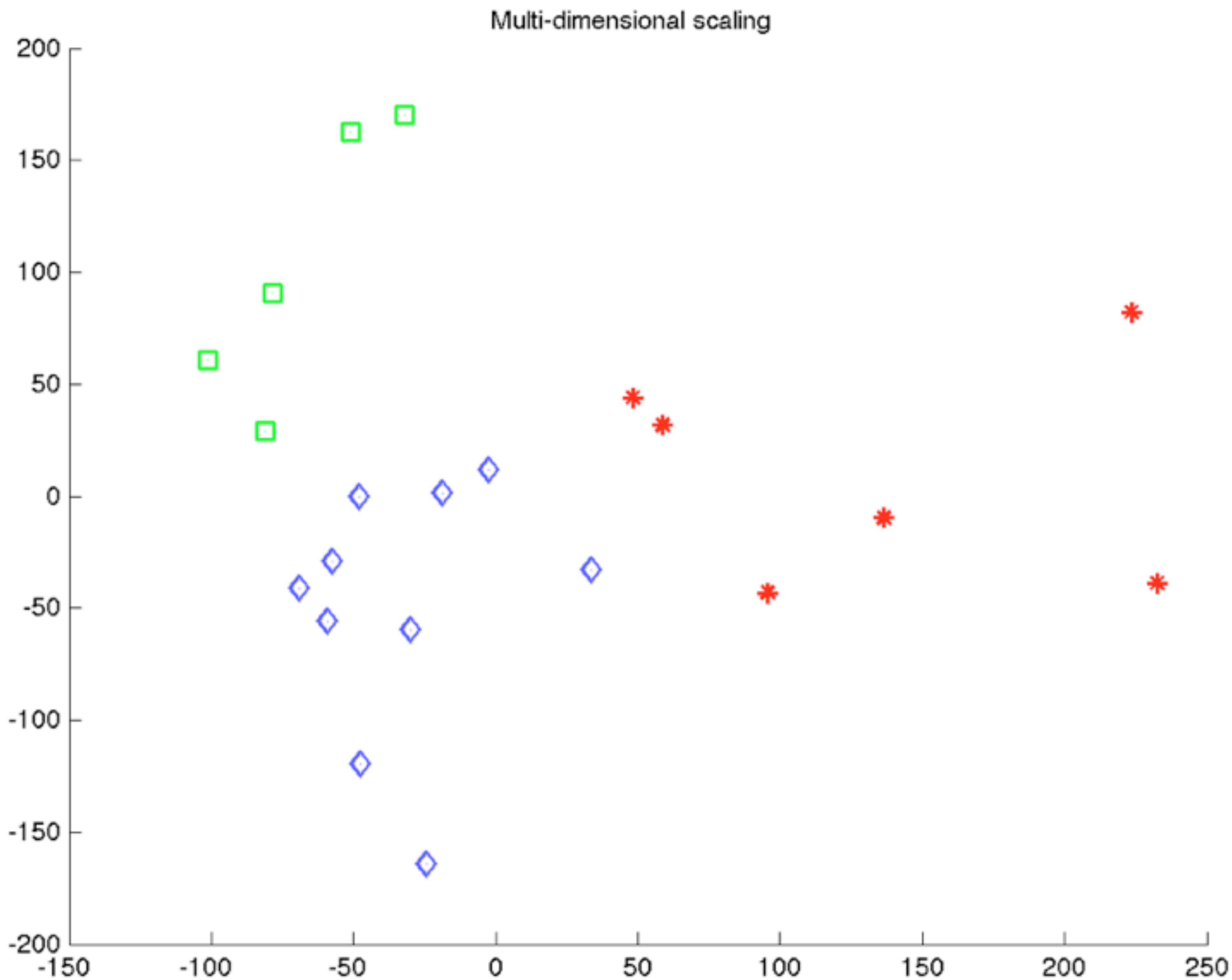


Cereals in PCA w/ Gaussian kernel

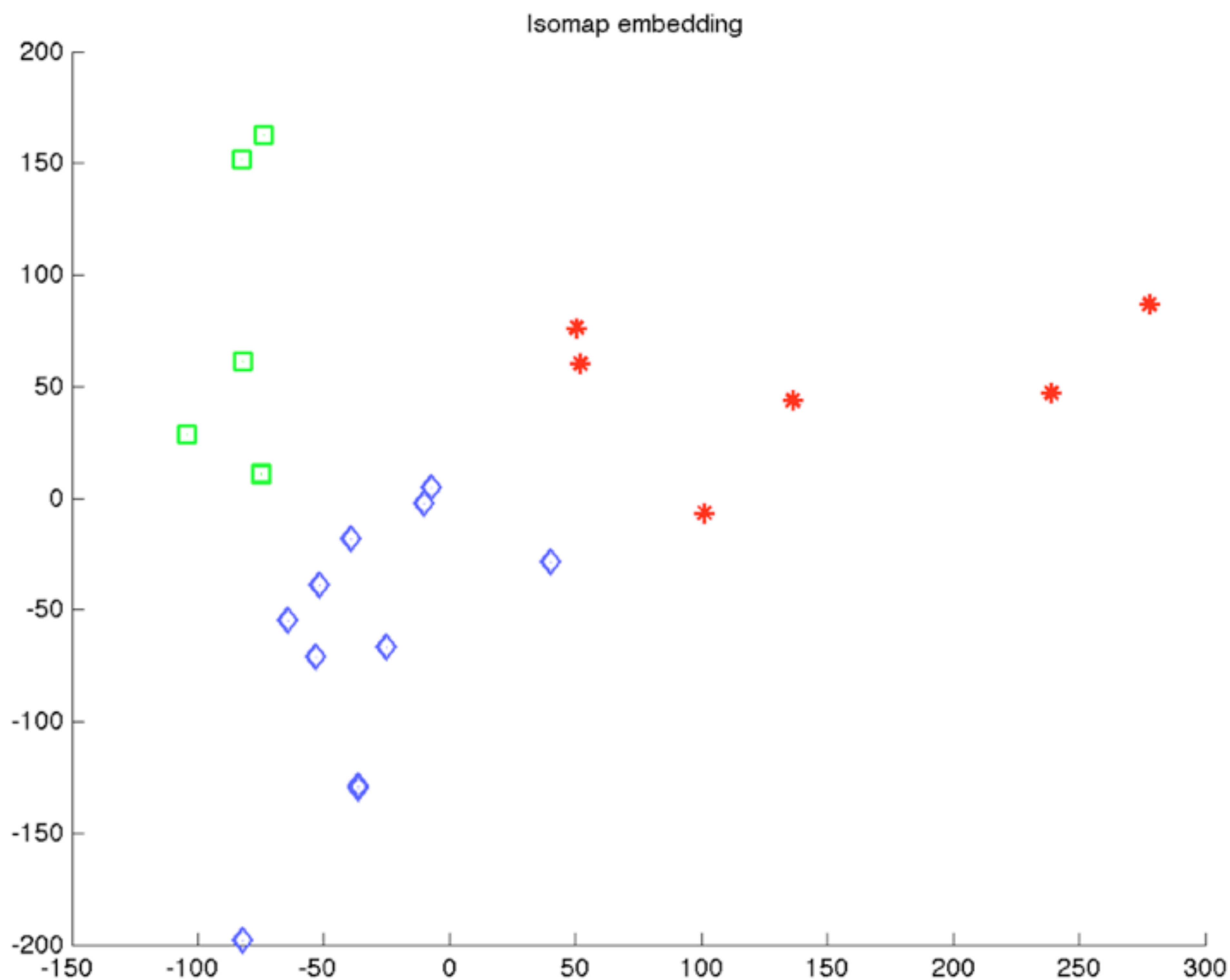


Zaki & Meira Ch. 7.3

Cereals and multidimensional scaling

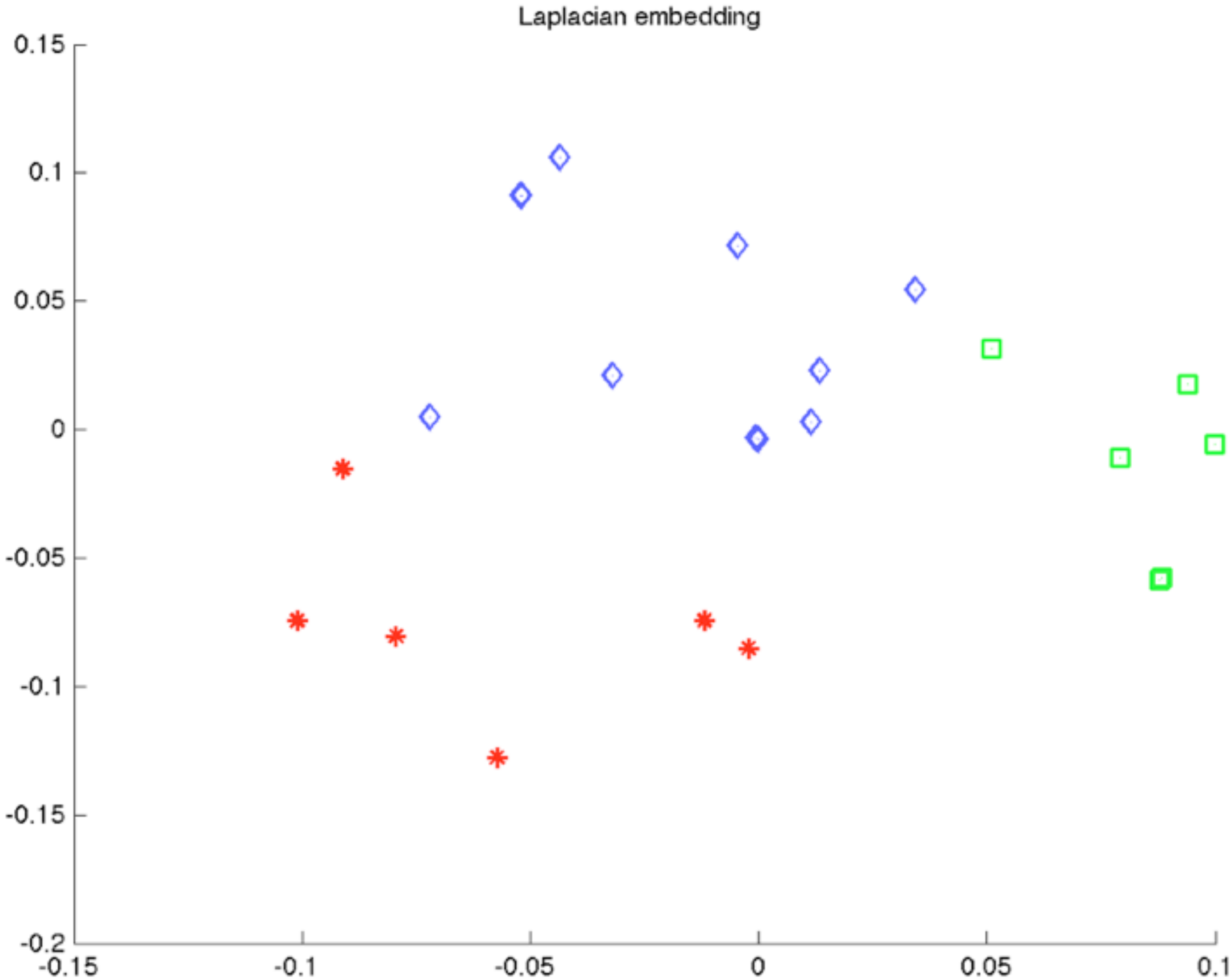


Cereals and Isomap



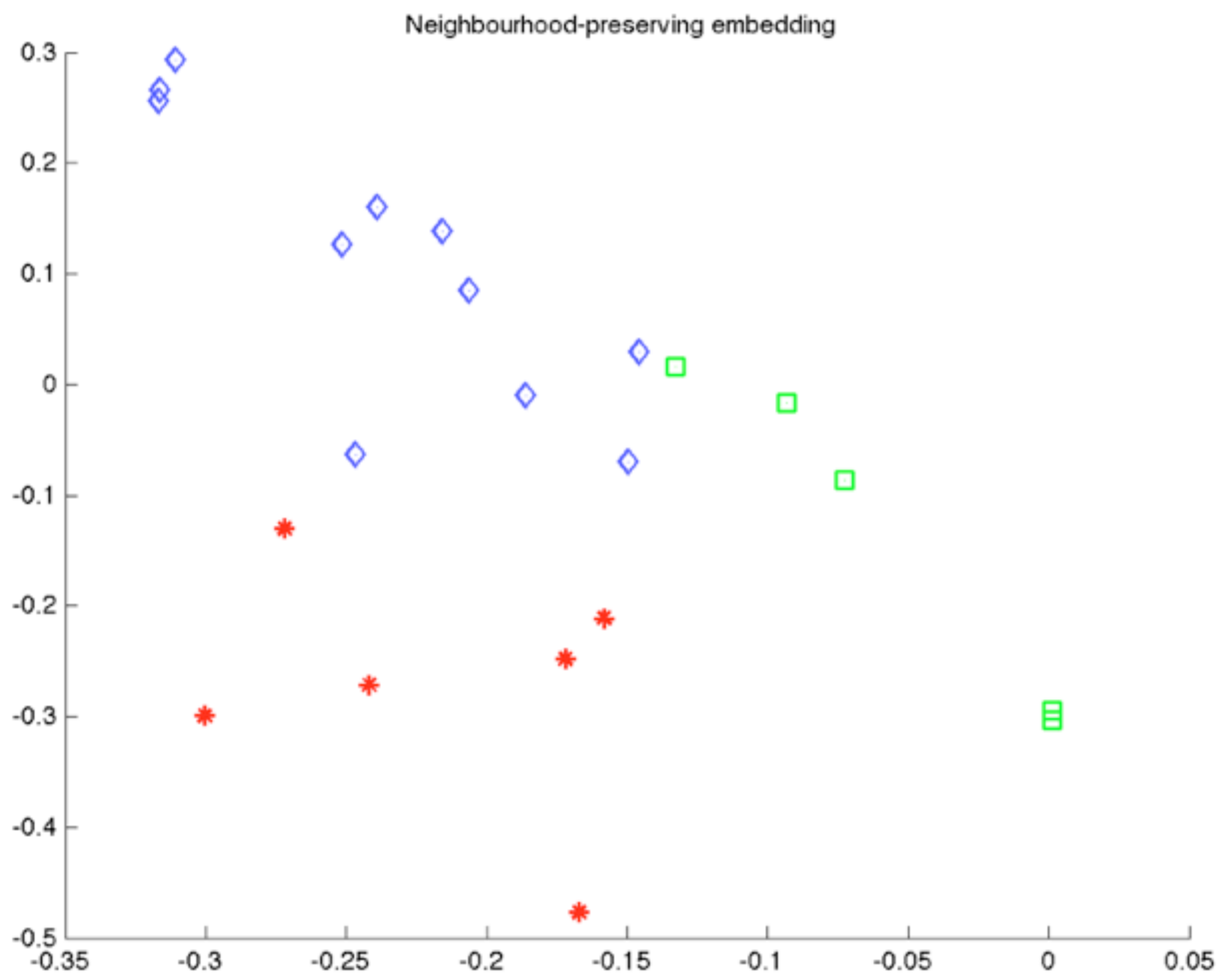
Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, *290*(5500), 2319–2323. doi:10.1126/science.290.5500.2319

Cereals and Laplacian eigenmaps



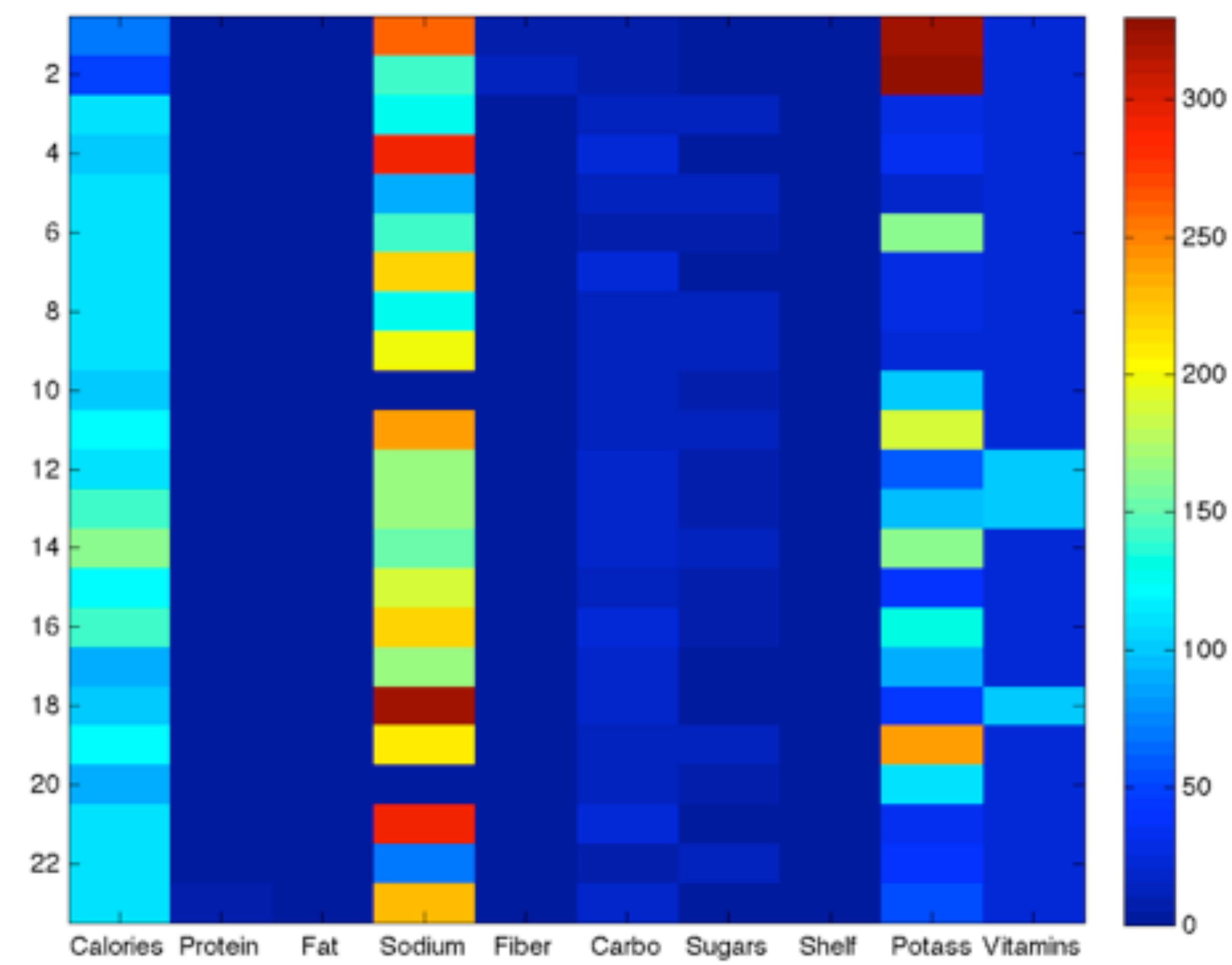
Belkin, M., & Niyogi, P. (2003). Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computing*, *15*(6), 1373–1396. doi:10.1162/089976603321780317

Cereals and neighbourhood-preserving embedding

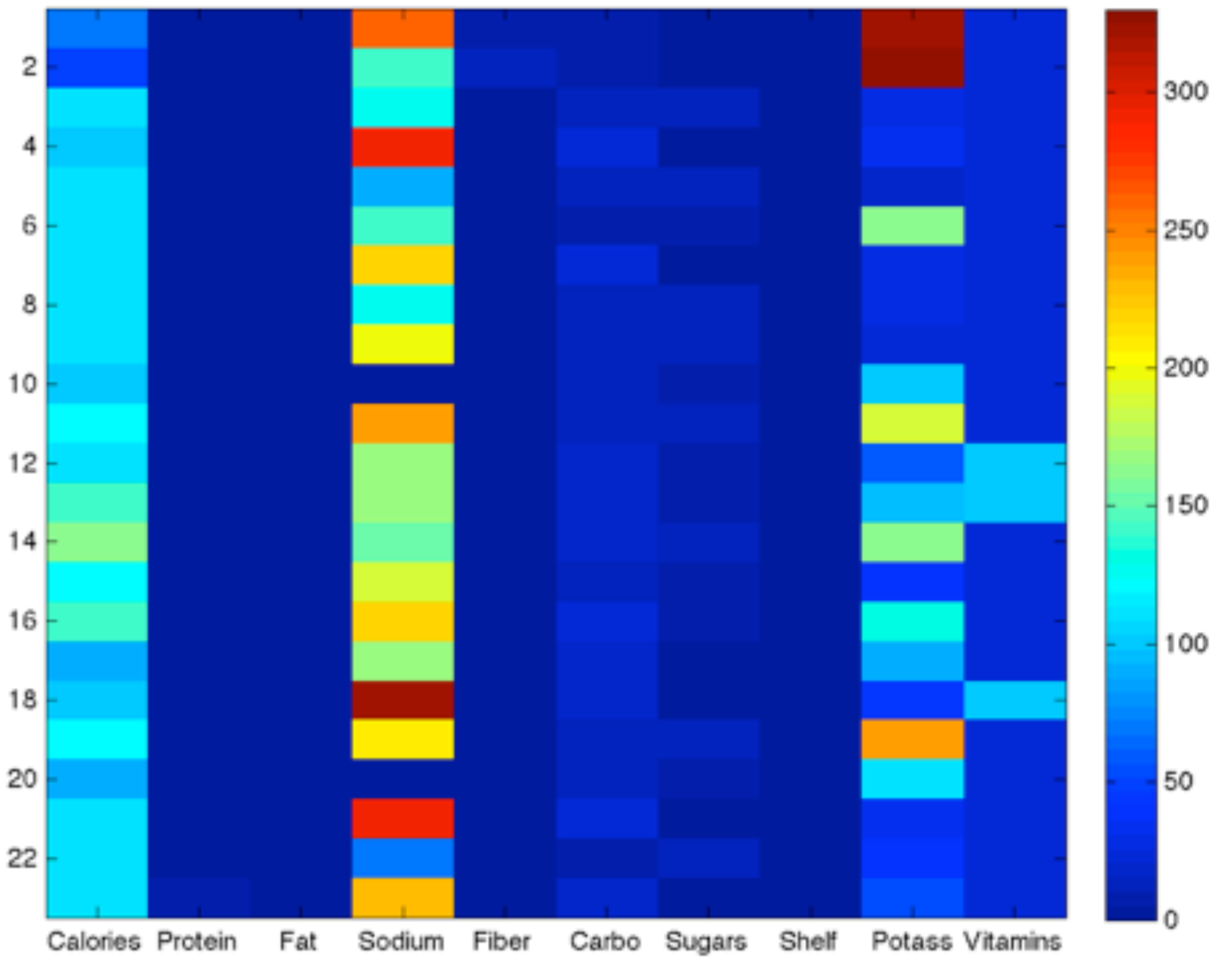


Non-projection visualizations

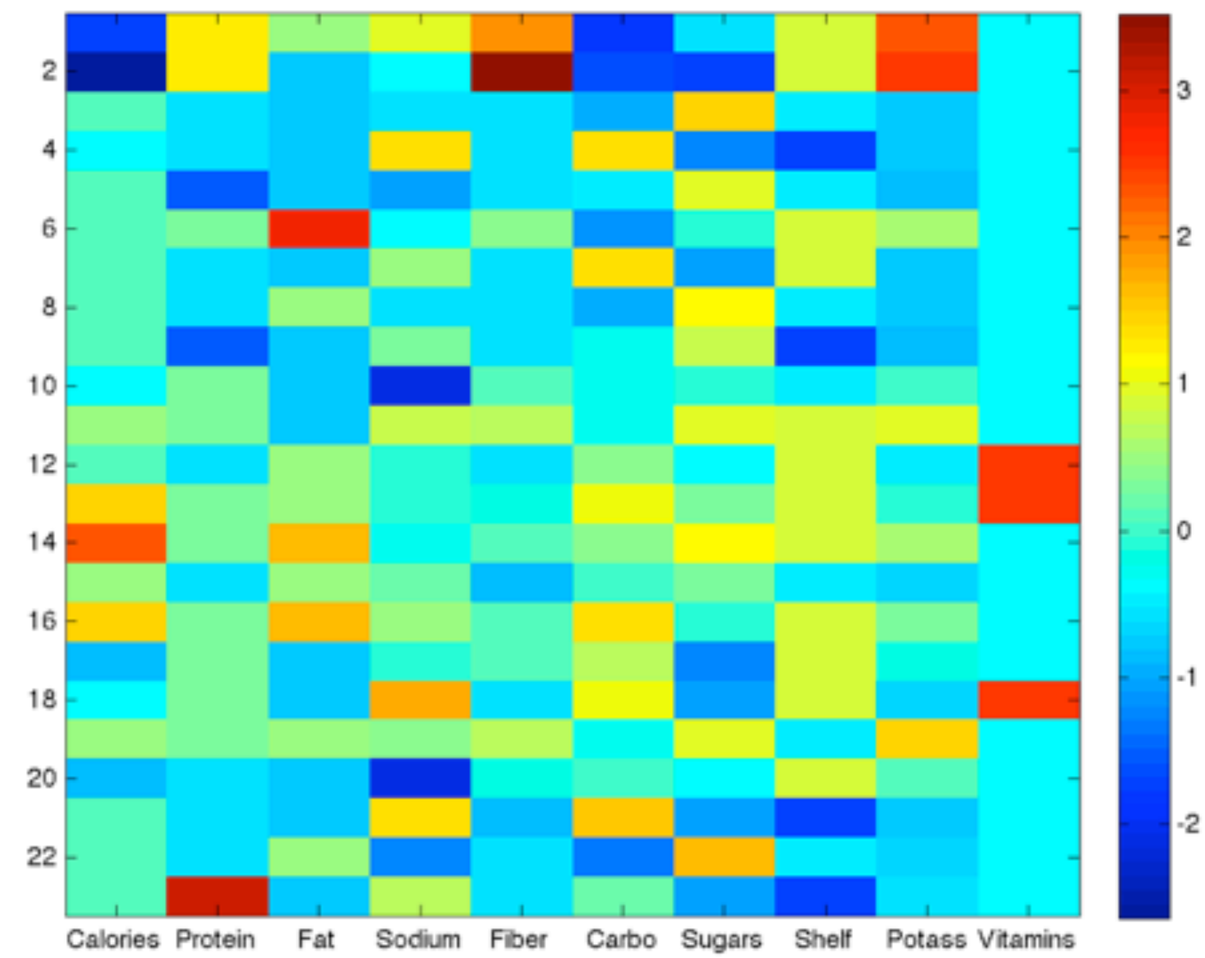
- Projections are not the only type of visualizations
  - Again, we have seen other visualizations before
  - These are often a bit more specific
    - But not always…



Heat maps

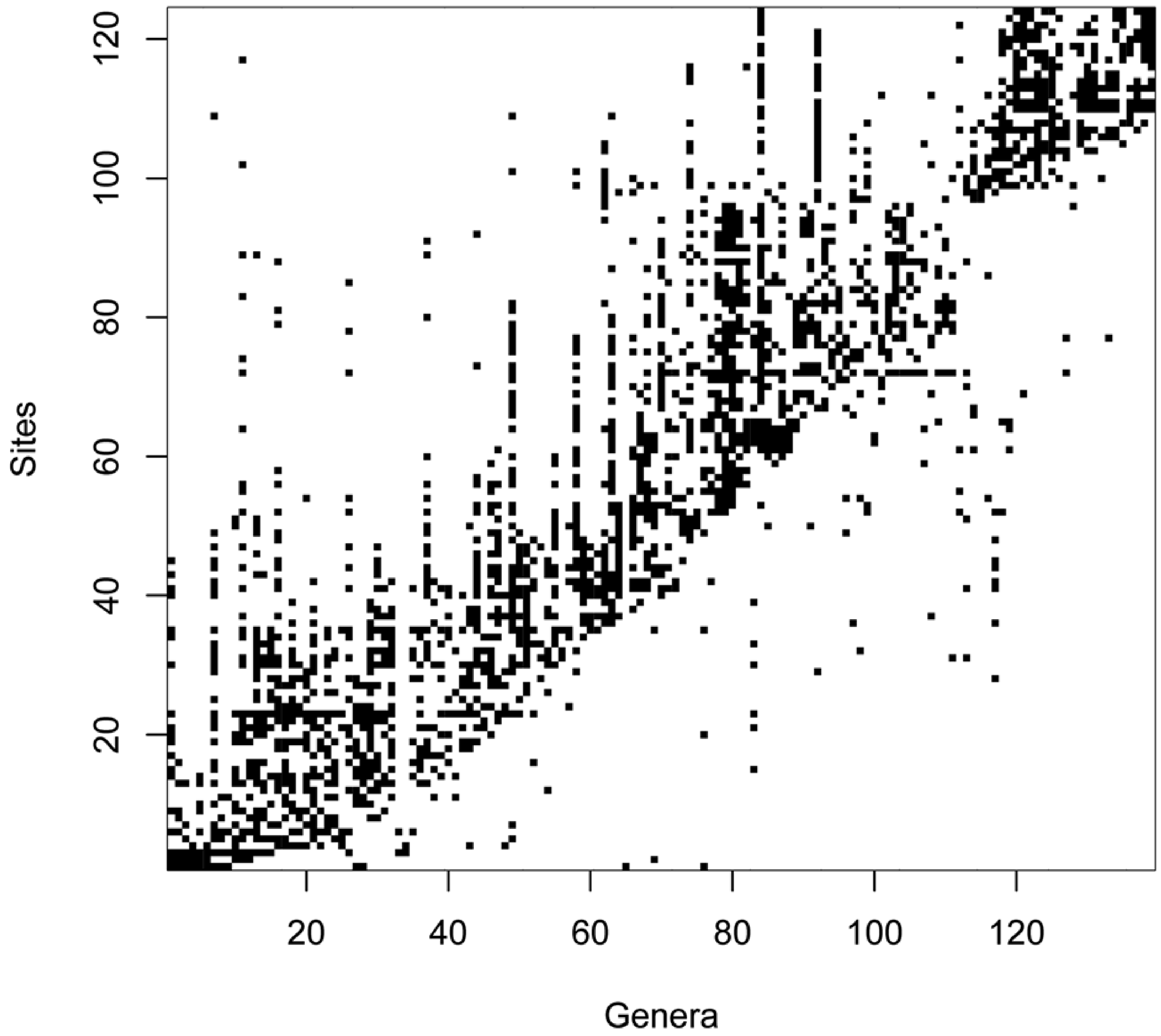


Original

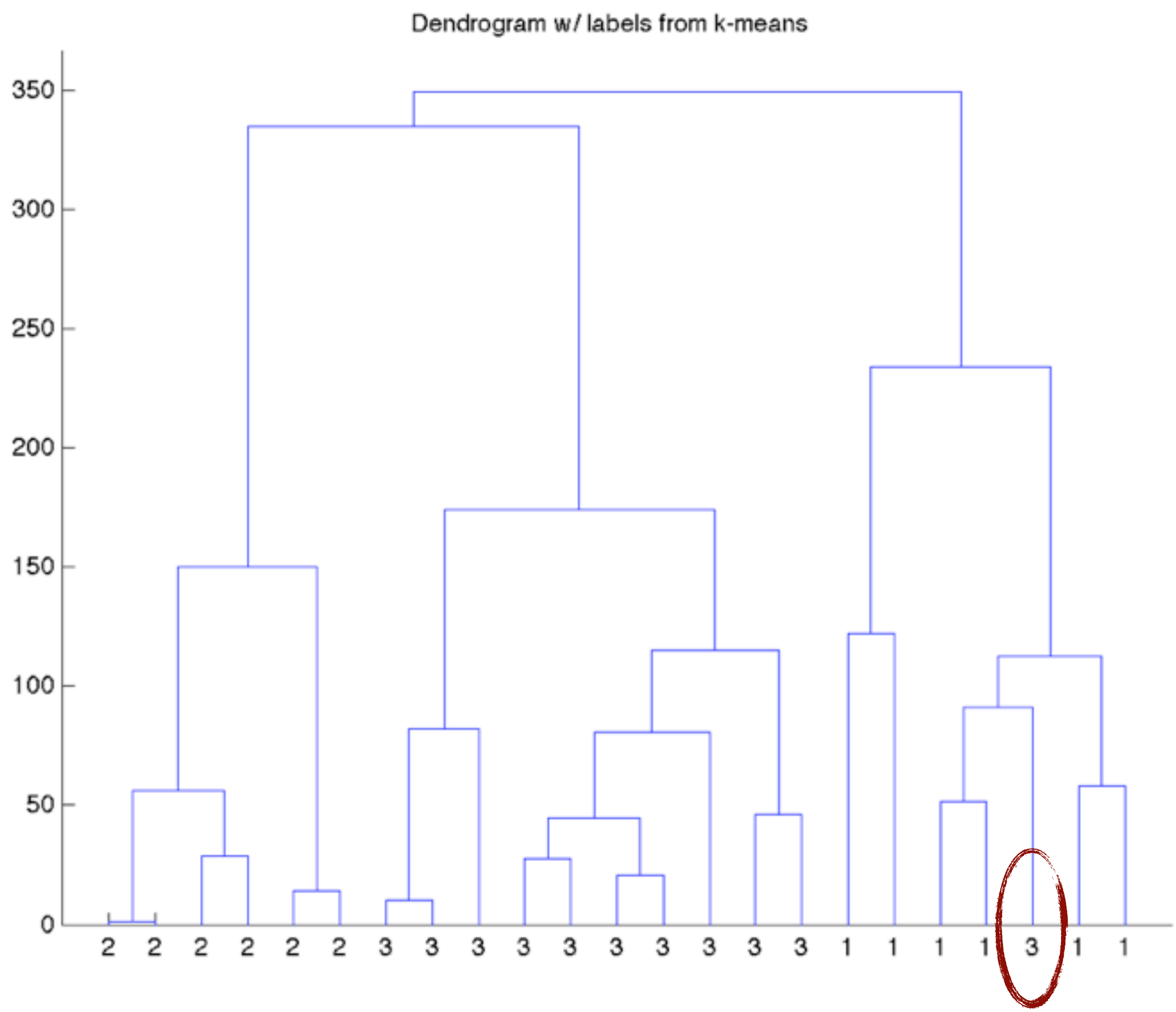


Normalized

Heat maps with sorting



Dendrograms



Heat maps with dendrograms

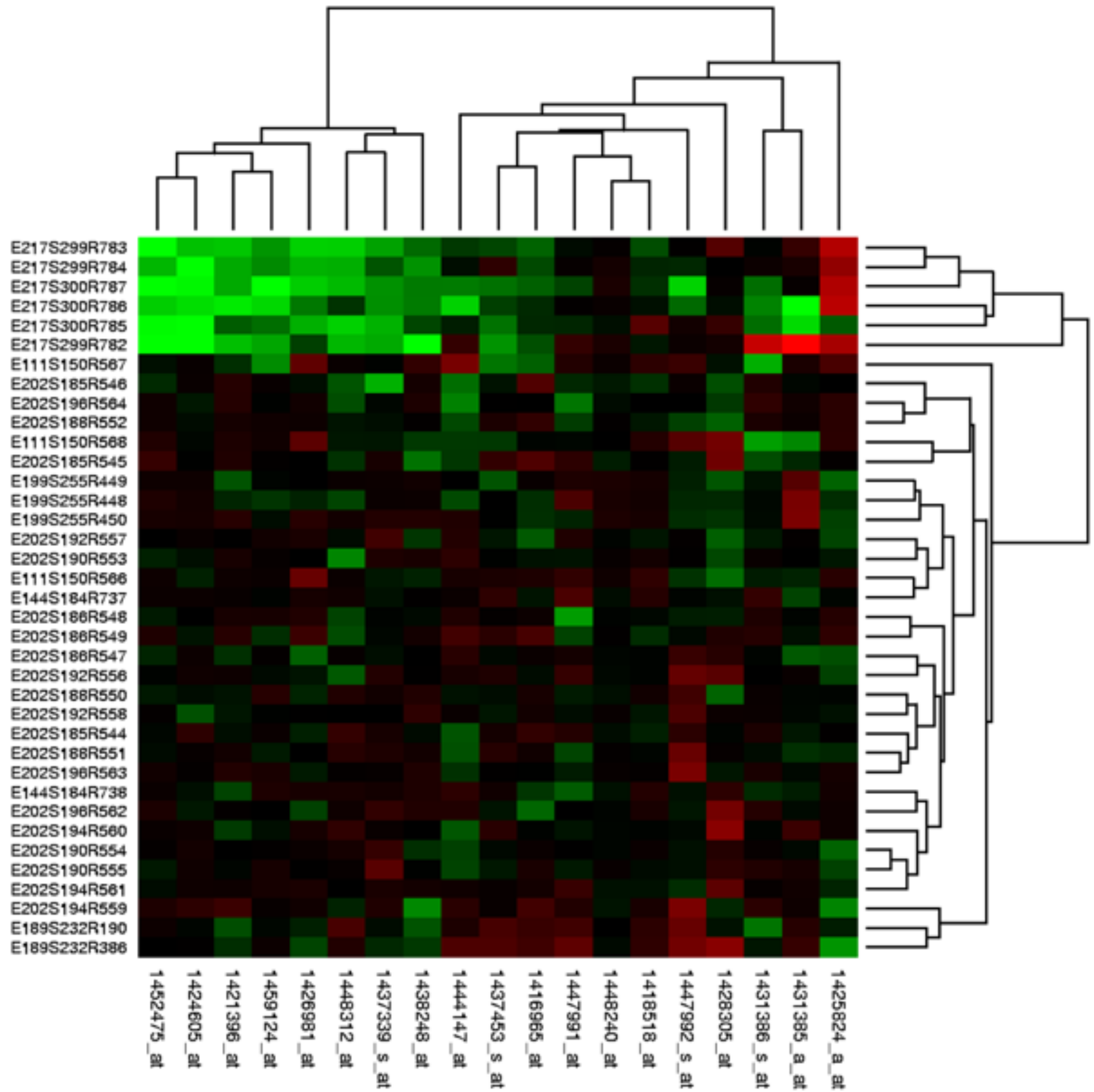
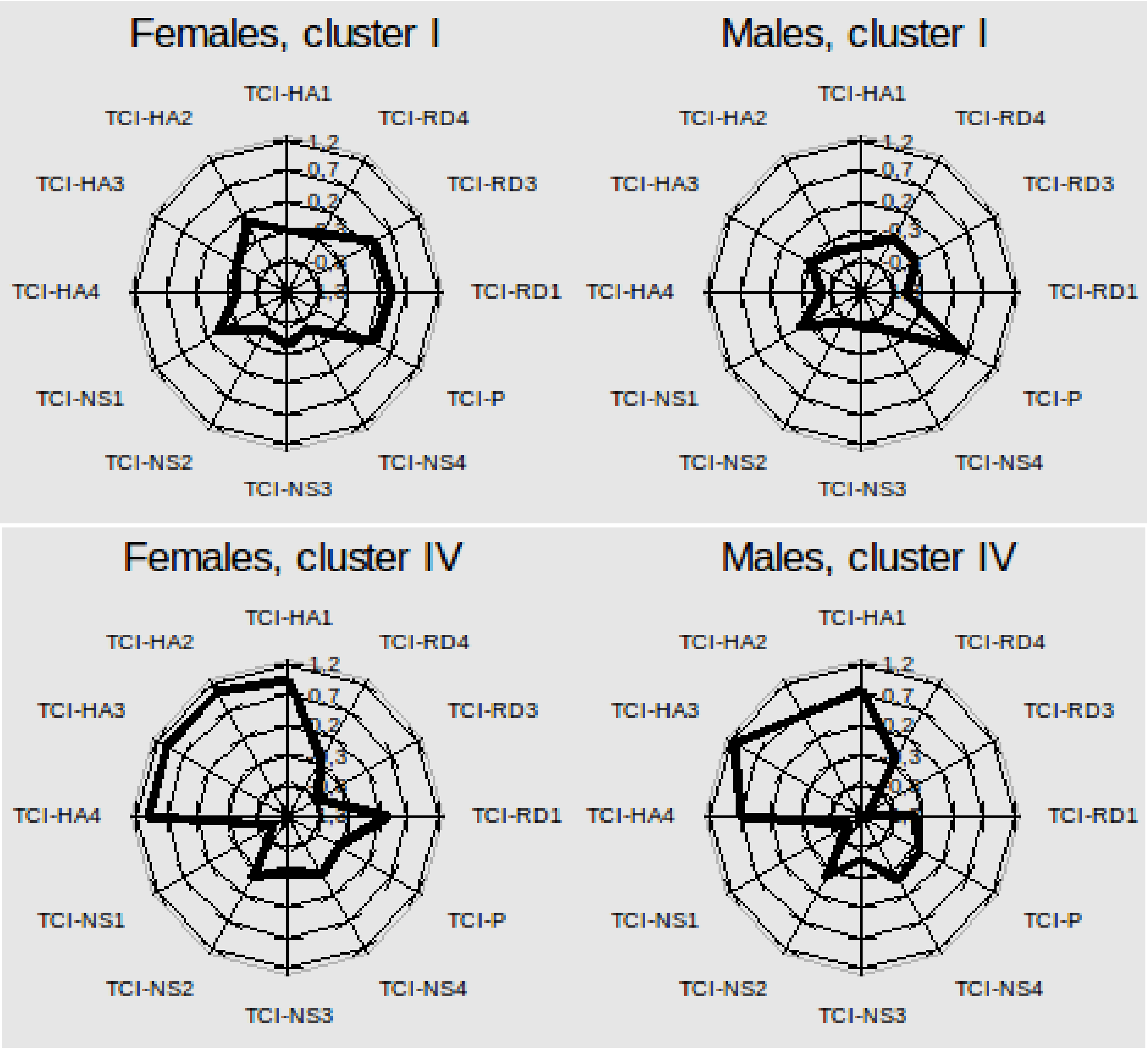
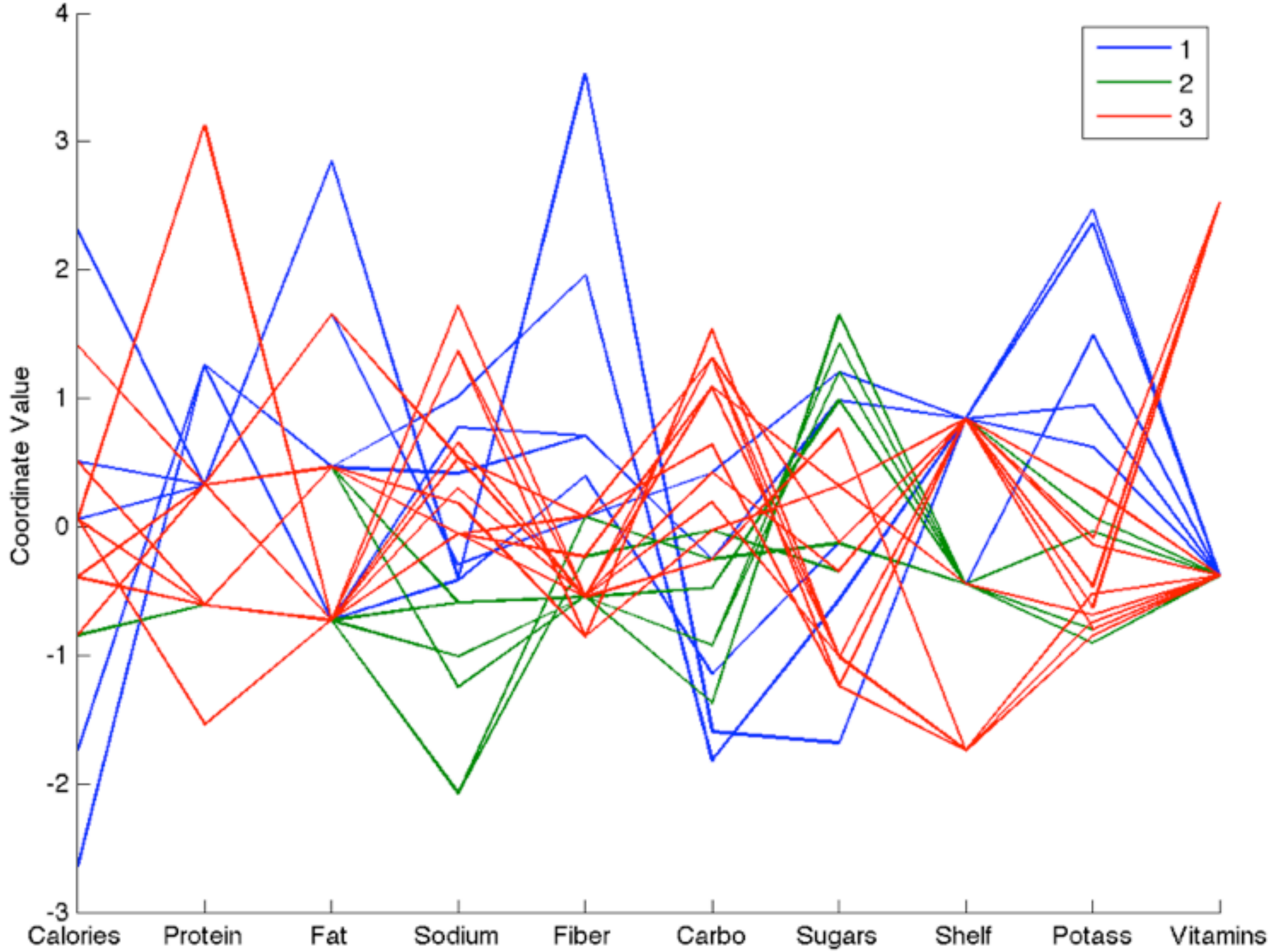


Image: Wikipedia

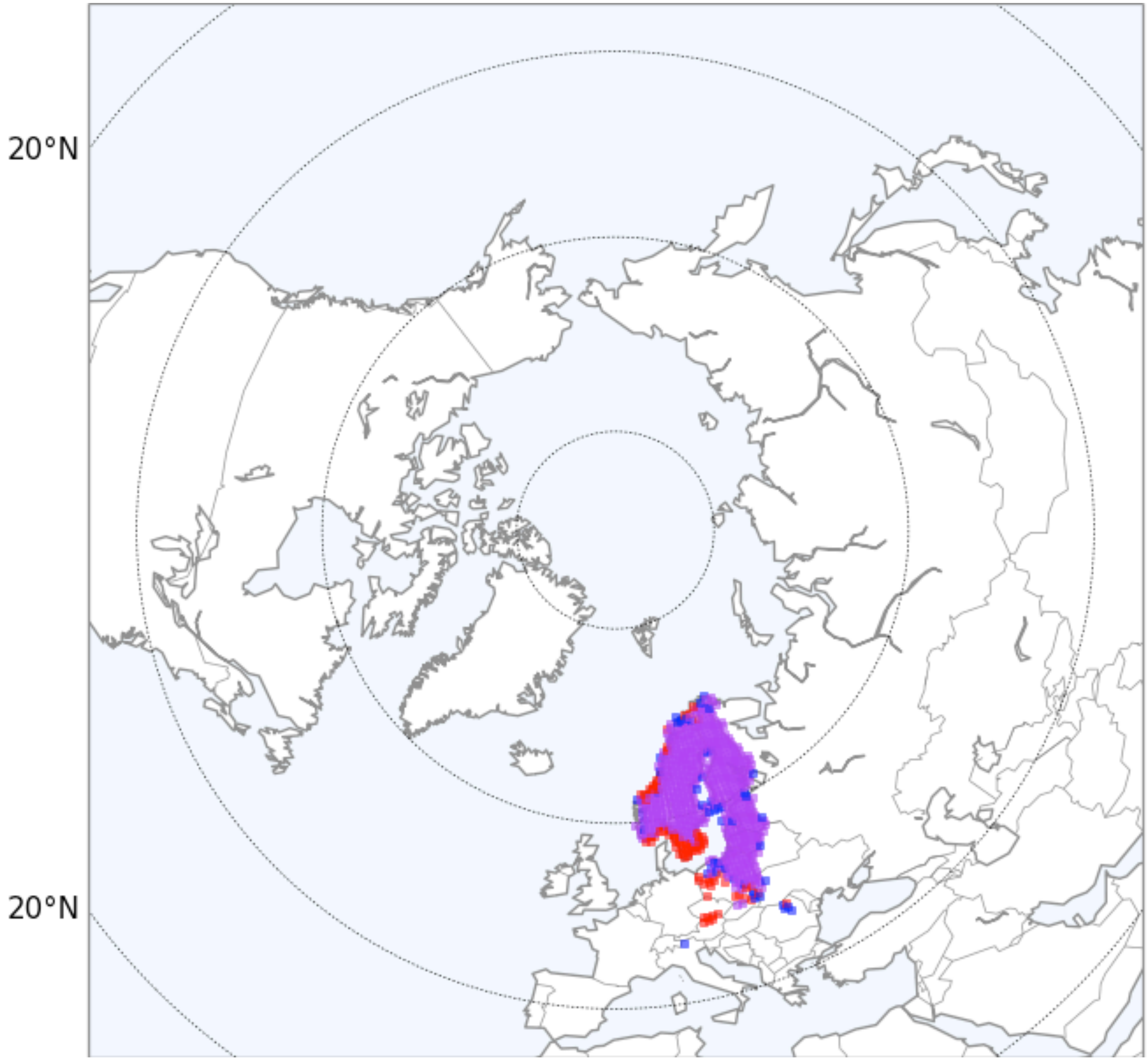
Radar charts



Parallel coordinates



Maps…



Analysis of the results

- Without analysis, there's not much point in doing data mining
- The analysis should be done by domain experts
  - People who know what the data contains and how to interpret the results
- Data mining is about finding surprising things…
  - … so domain experts are needed to
    - tell if the results really are surprising
    - verify that the surprising results are meaningful in the context

Significance of the results

- Statistical significance tests can be applied to the results
  - But they require forming the null hypothesis
- Too weak null hypothesis $\Rightarrow$ even significant results are not necessarily significant at all
  - But strong null hypotheses are harder to test
- We rarely can use (full-blown) exact tests
- Sometimes we can use asymptotic tests
- In other times we can use permutation tests

Significance testing example (1)

- We want to test the significance of association rule $X \rightarrow Y$ in a data with n rows
- Null hypothesis 1: Itemsets X and Y both appear in the data but their tidsets are independent random variables
  - Each transaction contains X with probability $supp(X)/n$
- The probability for $supp(XY)$ is a tail of a binomial distribution for $p = supp(X)supp(Y)/n^2$

$$\sum_{s=supp(XY)}^{n} \binom{n}{s} p^s (1-p)^{n-s}$$

Significance testing example (2)

- Null hypothesis 2: $X \to Y$ does not add anything over a generalization $W' \to Y$, where $W \subsetneq X$ assuming the row and column marginals are fixed
- The odds ratio measures the odds of X occurring with Y versus the odds of W (but not other parts of X) occurring with Y
  - For any W, we can consider the null hypothesis that odds ratio = 1 ($X \setminus W$ is independent of Y given W)
  - We can compute the p-value for this hypothesis using hypergeometric distribution
  - We can test null hypothesis 2 by computing the p-values for all generalizations of X

Z&M Ch. 12.2.1

Significance testing example (3)

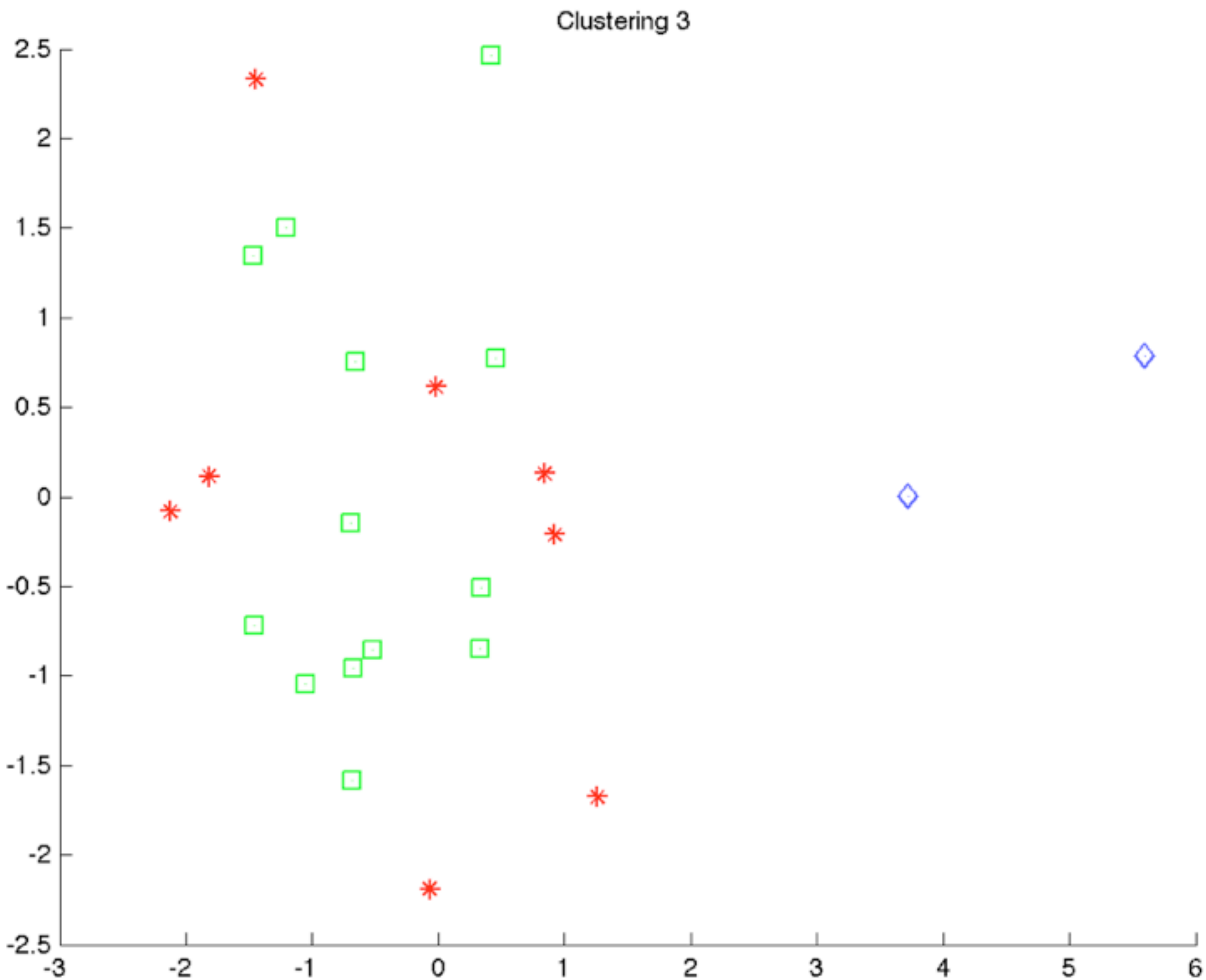
- Null hypothesis 3: The confidence of the rule is explained merely by the row and column marginals of the data
  - Confidence can be replaced with any other interest measure
- This we can test by generating new data sets with same row and column marginals
  - If many-enough of them contain rules with higher confidence, we cannot reject our null hypothesis
  - Generating such data can be done e.g. with swap randomization
- This is called **permutation test**

Z&M Ch. 12.2.2

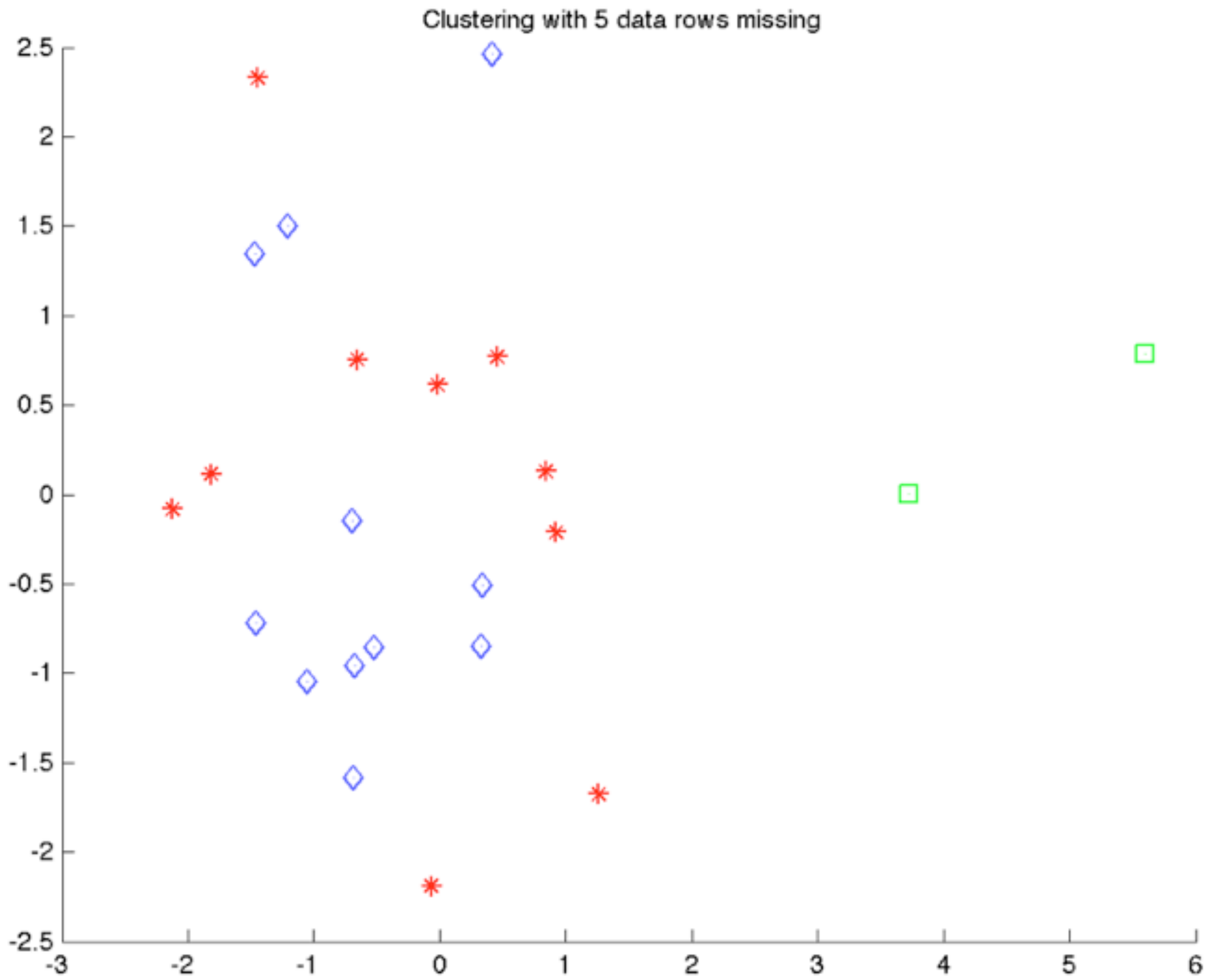
Stability

- The **stability** of a data mining result refers to its robustness under perturbations
  - E.g. if we change all the numerical values a bit, the clusterings shouldn't change a lot
  - We can also remove individual rows/columns or make more data unknown
- Stability should be tested after the results have been obtained
  - Run the same analysis with perturbed data

Stability example (1)



Stability example (2)



Leakage

- **Leakage** in data mining refers to the case when prediction algorithm learns from data is should not have access to
  - Problem as the quality is assessed using already-historical test data
  - E.g. INFORMS'10 challenge: predict the value of a stock
    - Exact stock was not revealed
    - But "future" general stock data was available!
      $\Rightarrow$ 99% AUC (almost perfect prediction!)
  - More subtle one's exist
    - E.g. removing a crucial feature creates a new type of correlation

XII.6: Tales from the Real World

1. Working with non-CS folks

Talk their language!

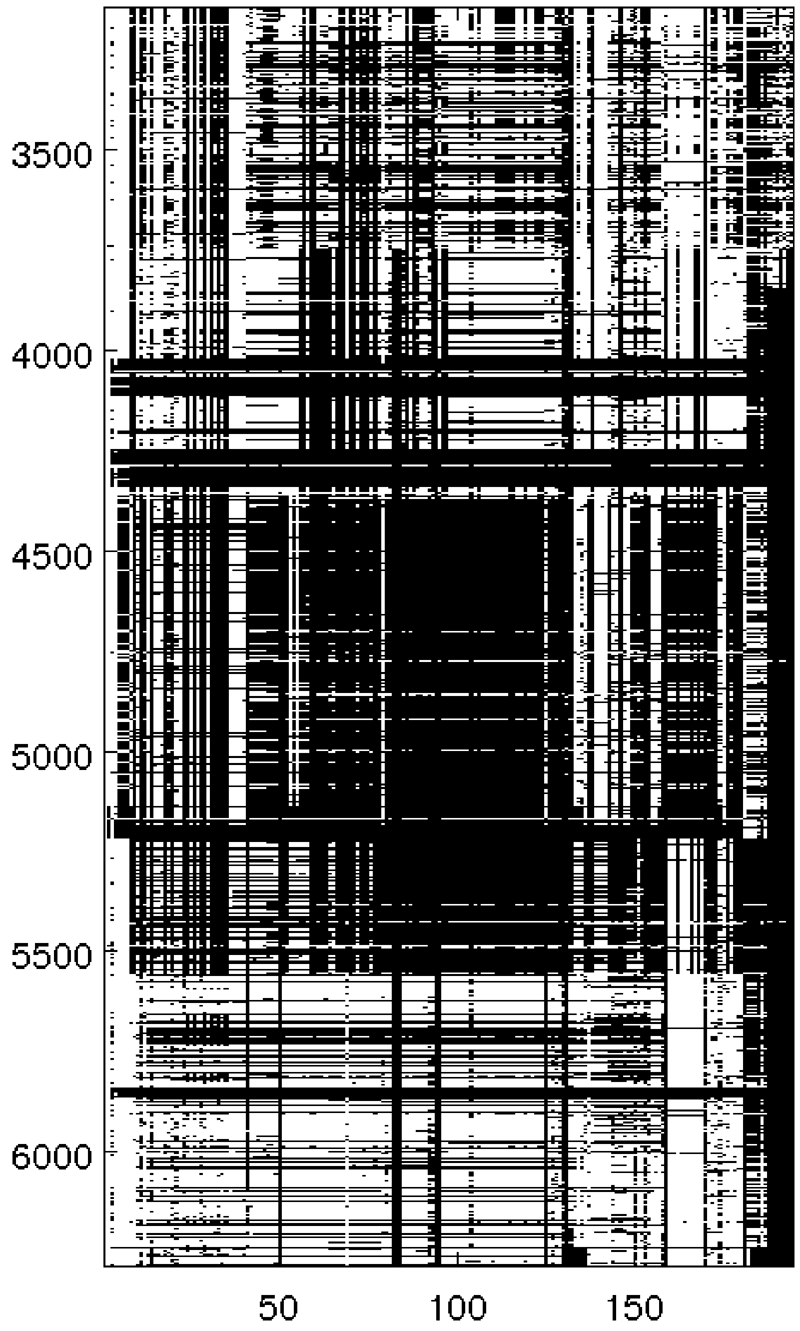
Archeotype

Voronoi tesselation

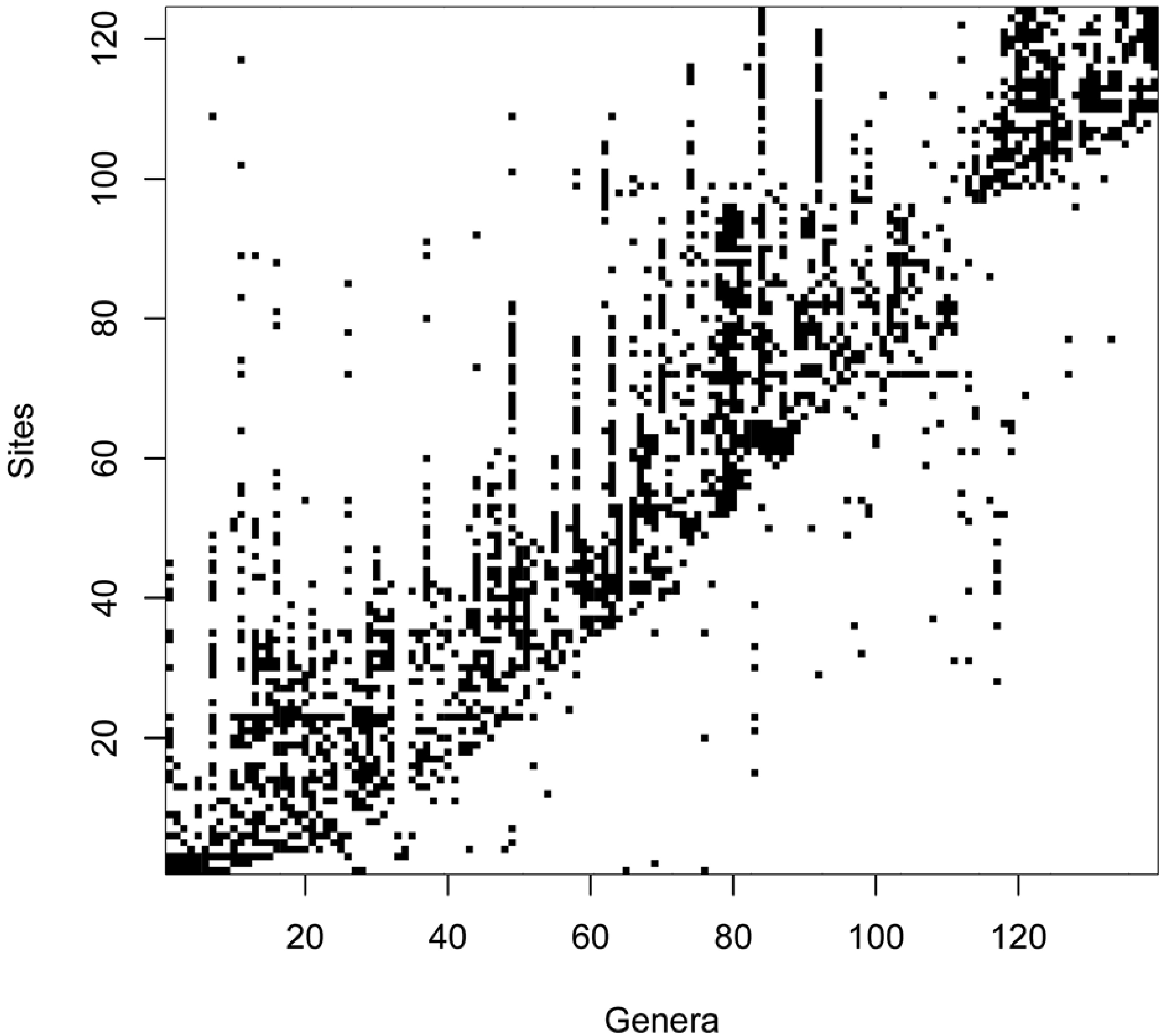
Red queen's problem

NP-hard

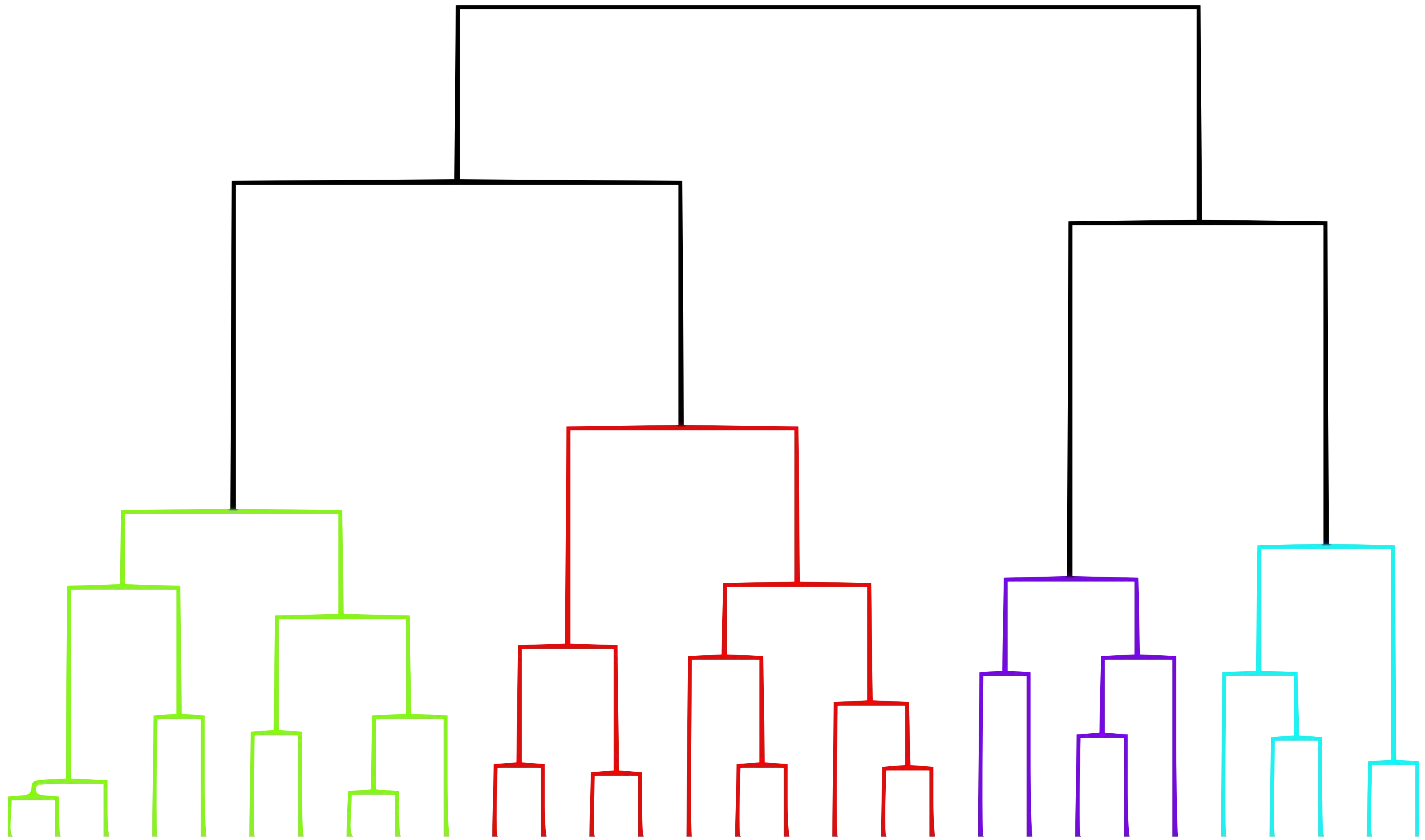
Data is dirty



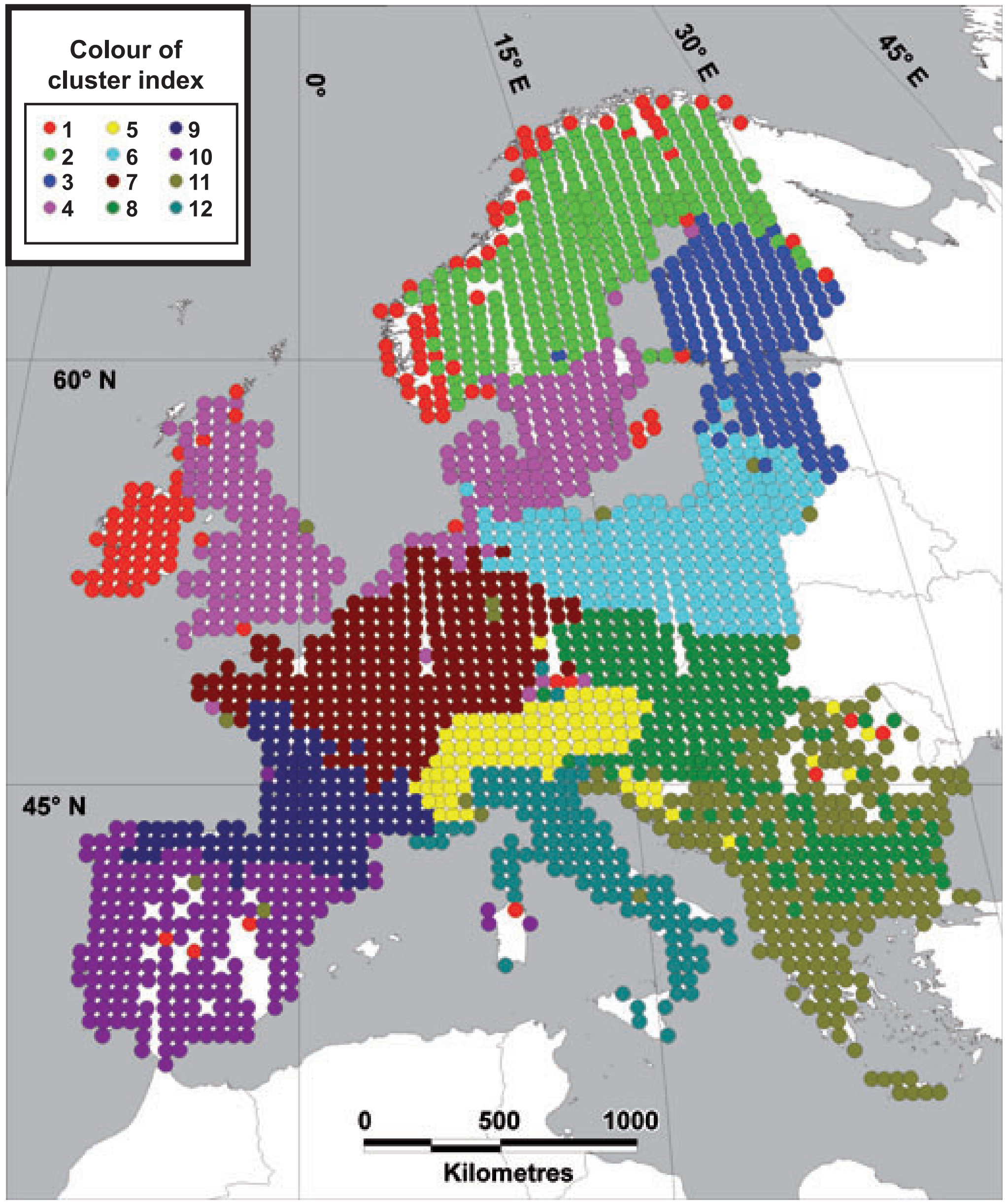
Not all data is BIG
It's all just constants

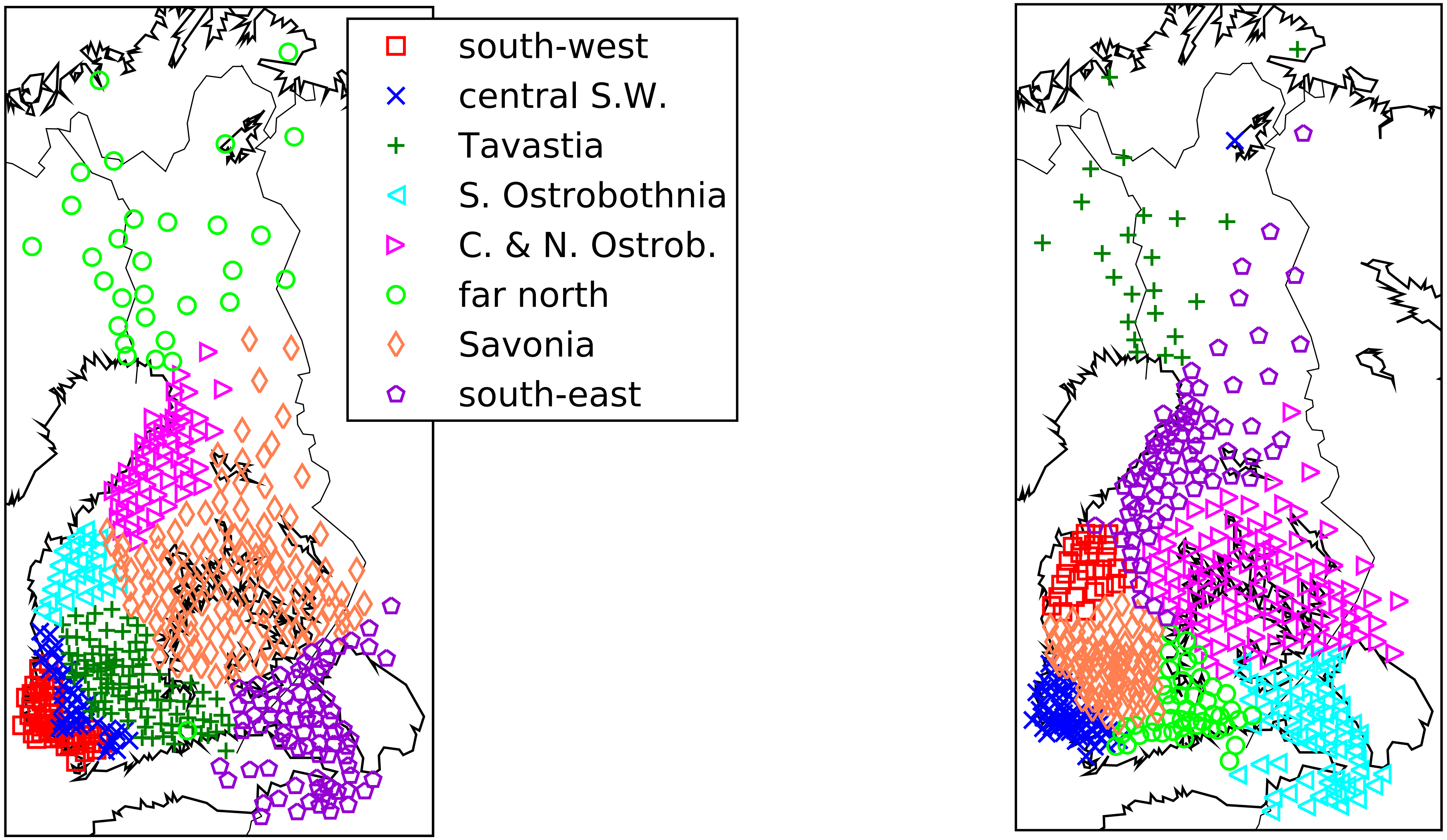


The best algorithm is the algorithm you have with you



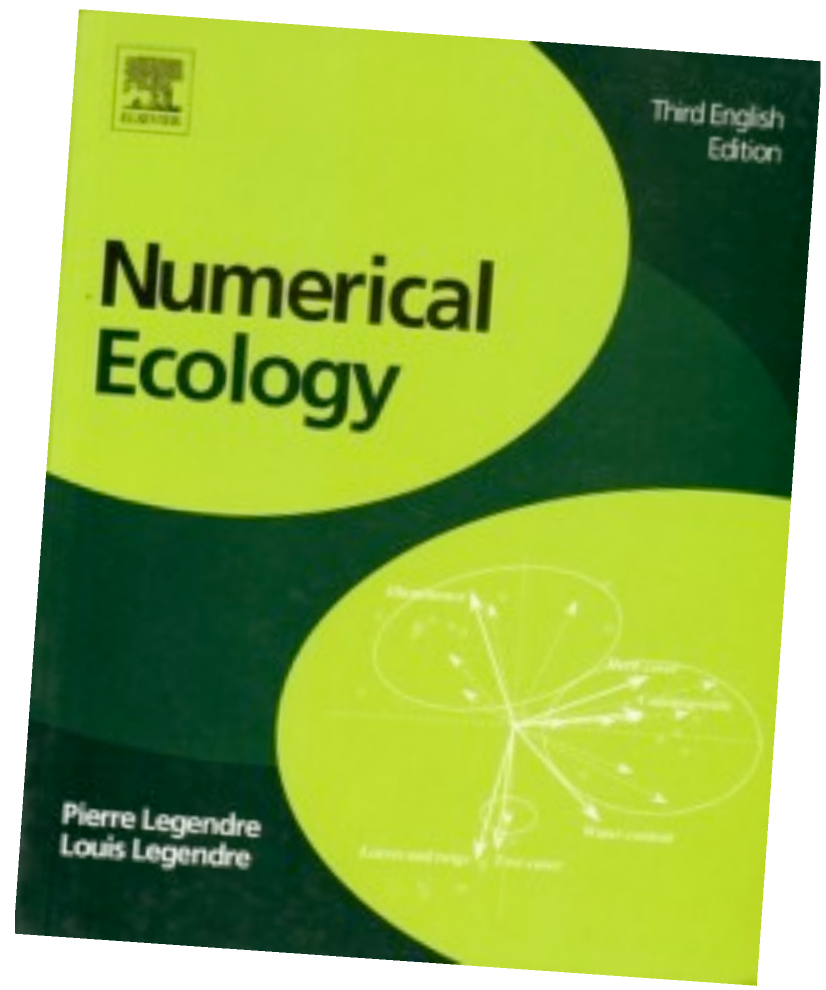
Beware the analysis

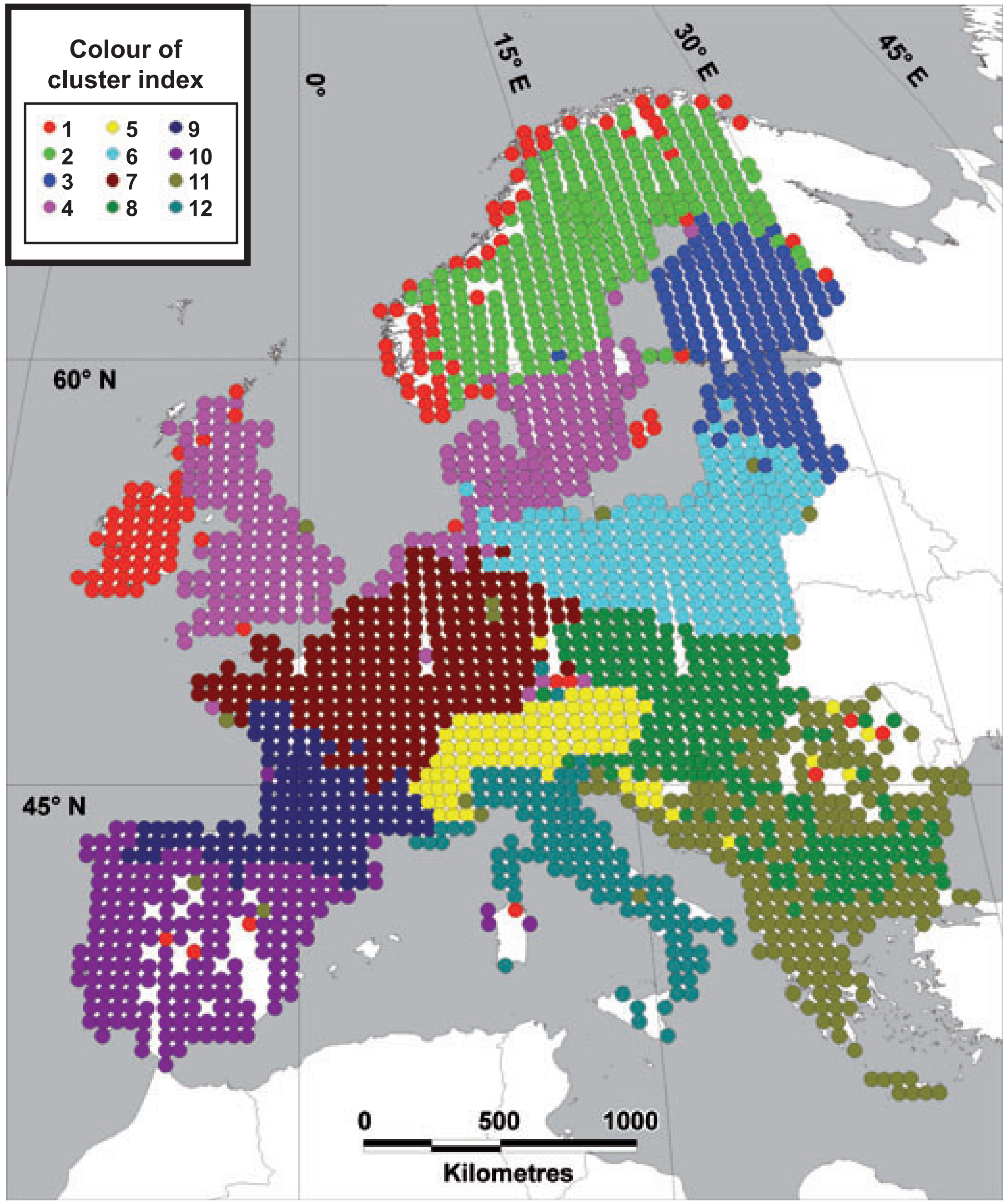


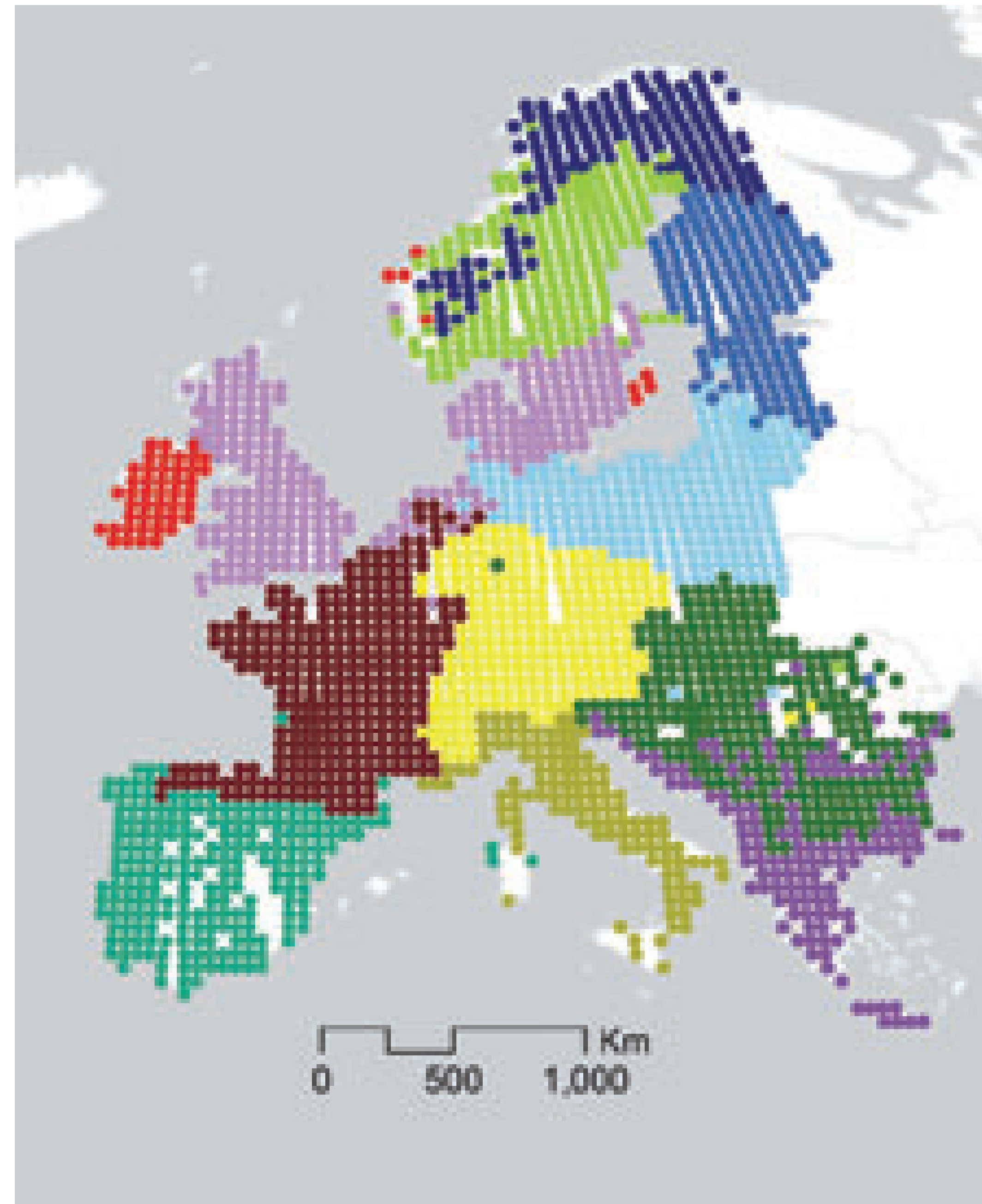


Itkonen: *Proto-Finnic Final Consonants: Their history in the Finnic languages with particular reference to the Finnish dialects, part I: 1, Introduction and The History of -k in Finnish*, 1965

Know the math of the domain




Colour of
cluster index
1
2
3
4
5
6
7
8
9
10
11
12
0°
15° E
30° E
45° E
60° N
45° N
0
500
1000
Kilometres


0
500
1,000
Km

Data mining = voodoo science

The response from several social scientists has been rather unappreciative along the following lines: “Where is your hypothesis? What you’re doing isn’t science! You’re doing DATA MINING !”

http://andrewgelman.com/2007/08/a_rant_on_the_v/

The clash of paradigms

- Form a hypothesis
- Design a test
- Collect the data
- Test hypothesis
- Rinse and repeat

- Take somebody else's data
- Pick an algorithm
- Run the algorithm
- Analyse the results
- Rinse and repeat

Summary

- Think before you do
- Think while you do
- Think what you just did

- Real-world data analysis requires care and expertise
- Visualizations are powerful tools in data analysts toolbox
  - With great power comes great responsibility
- Data mining might be voodoo science
  - But who wouldn’t want to know the voodoo?