Comparative Visualization: Interactive Designs and Algorithms Depending on Data and Tasks

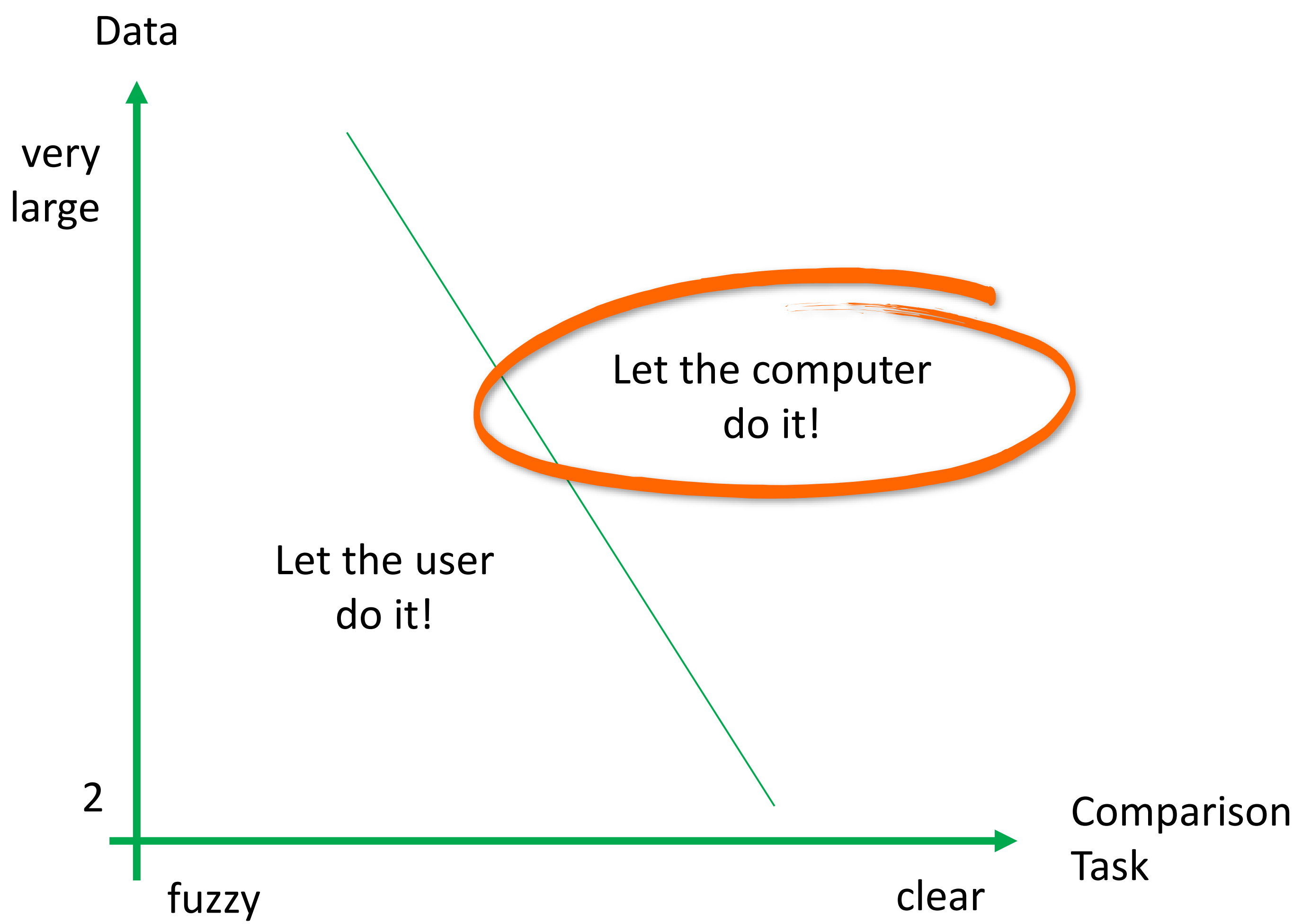
Tatiana von Landesberger[1], Kathrin Ballweg[1], Hans-Jörg Schulz[2], Natalie Kerracher[3], Margit Pohl[4]

VIS Tutorial 2018

1. TU Darmstadt, Darmstadt, Germany
2. Aarhus University, Denmark
3. Edinburgh Napier University, UK
4. TU Wien, Austria

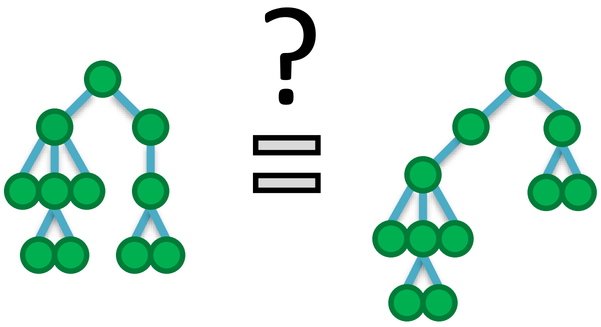
PART II:
ALGORITHMIC COMPARISON

Who performs the comparison?



1-to-1 Comparison

MATCHING



Definition

Matching:

Determination if a data object is related to another in quality, structure, or amount.

Levels of Relational Strictness:

- Identity
- Equivalency
- Similarity

Exact vs. Inexact Matching

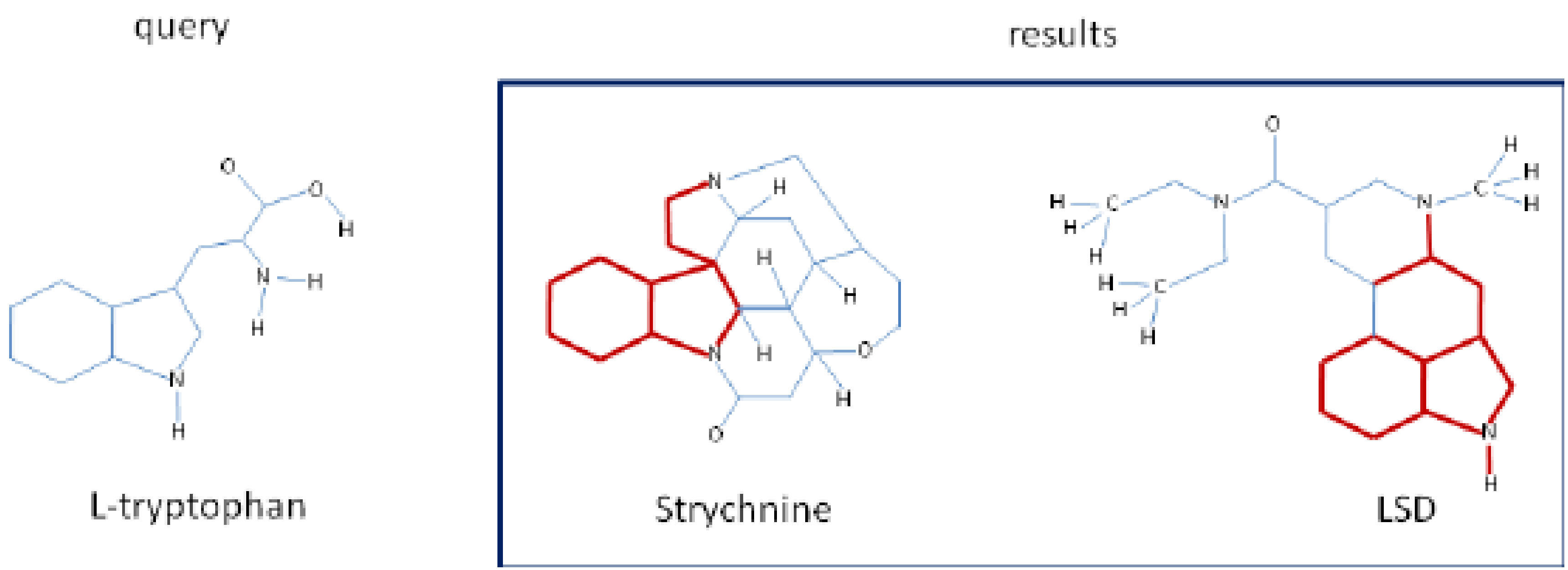


image taken from [Mongiovi et al. 2010]

Principal Pragmatic Approach

Given a Matching Algorithm, e.g.

- String Matching (RegEx, BLAST,…)
- Time Series Matching (DTW, LCSS, DISSIM,…)
- Graph Matching (TALE, SIGMA,…)

Transform data to string/pseudo-time series/ graph and process with given algorithm.

BLAST: [Altschul et al. 1990] DTW: [Berndt & Clifford 1994] LCSS: [Vlachos et al. 2002]
DISSIM: [Frentzos et al. 2007] TALE: [Tian & Patel 2008] SIGMA: [Mongiovi et al. 2010]

Graph -> String Conversion

Canonical Labeling, Canonization

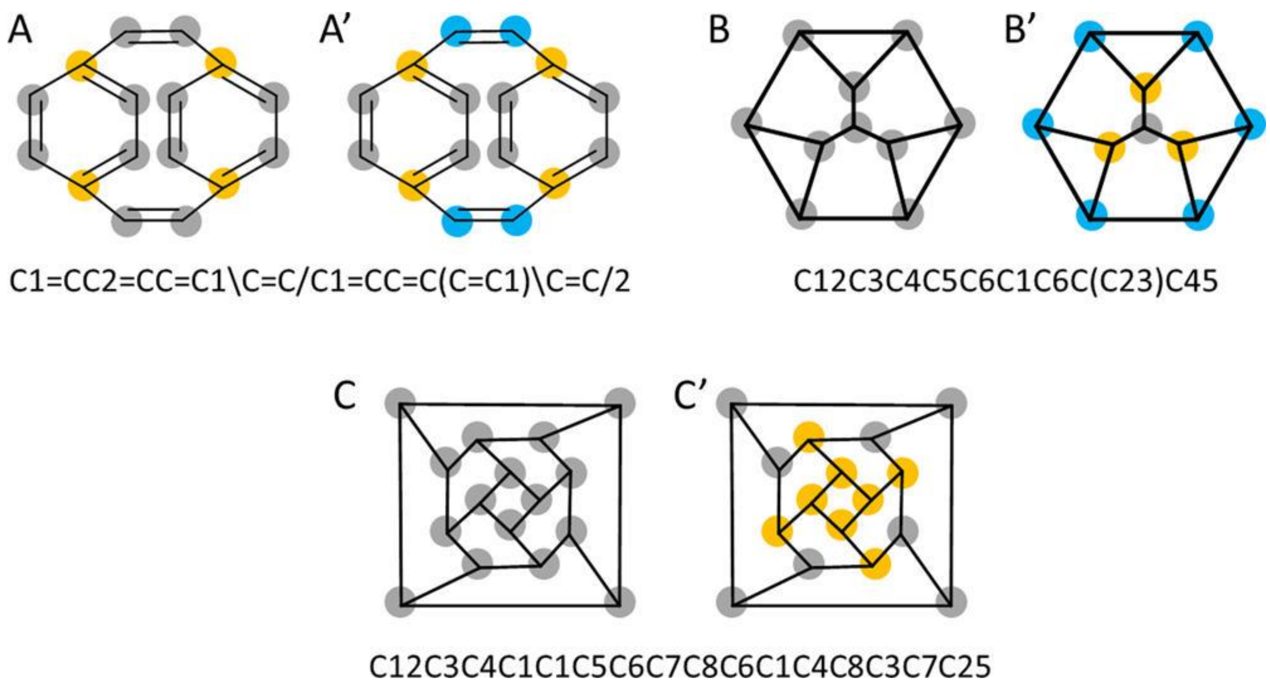
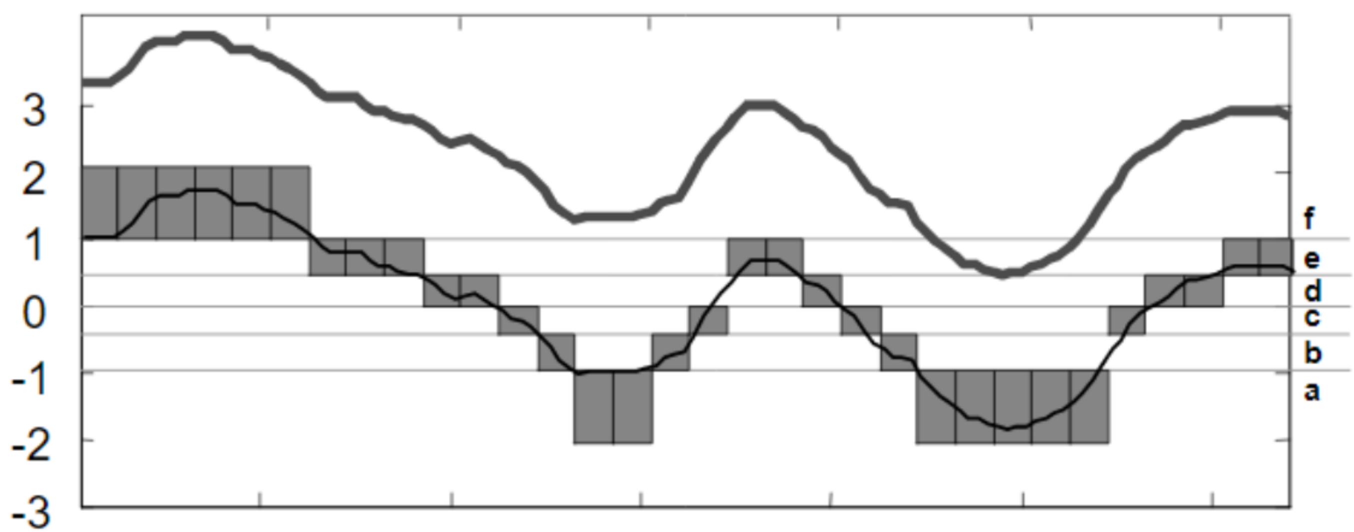


image taken from [Schneider et al. 2015]

Time Series -> String Conversion

Symbolic Aggregate Approximation (SAX)



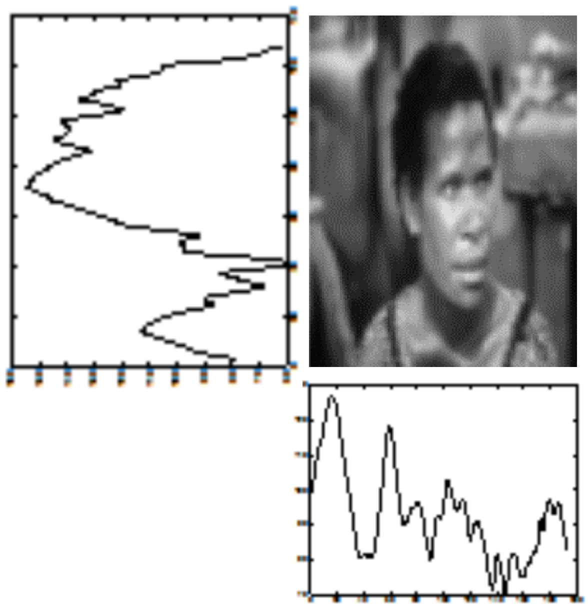
ffffffeeeddcbaabceedcbaaaaacddee

image taken from [Lin et al. 2003]

Image -> Pseudo-Time Series

Column- / Row-wise Aggregates

image taken from [Abdul-Moneim 2013]



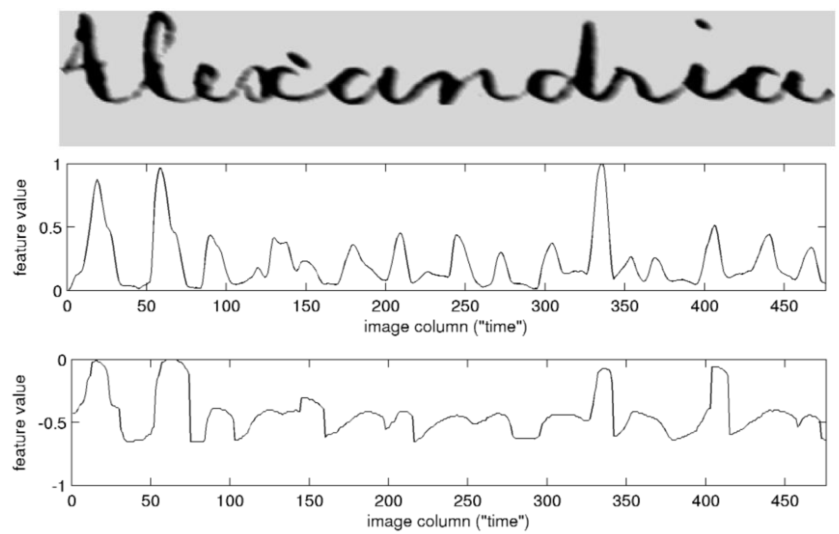


image taken from [Rath & Manmatha 2003]

Shape -> Pseudo-Time Series

Boundary Extraction w.r.t. Centroid

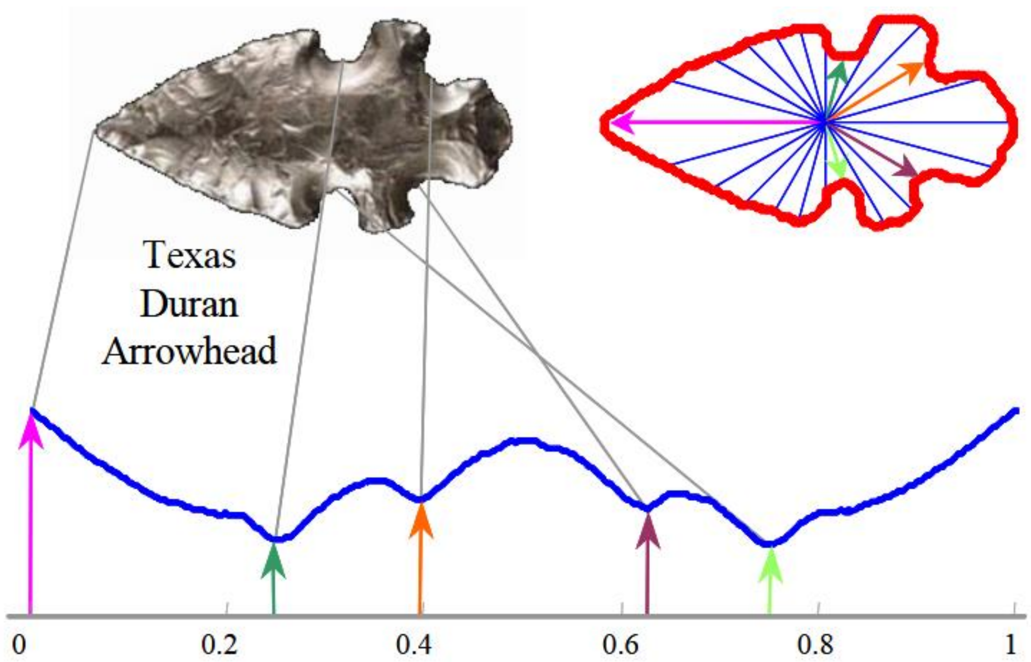
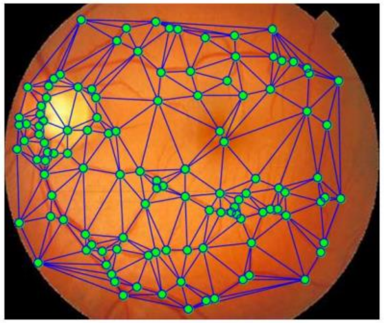


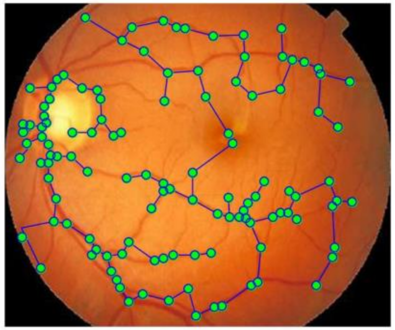
image taken from [Xi et al. 2007]

Image -> Graph Conversion

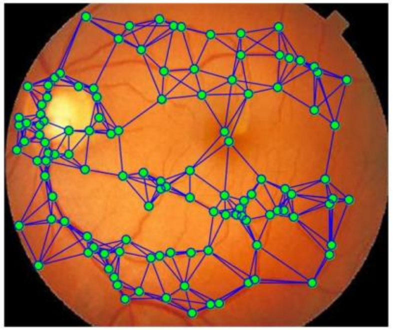
[Deng et al. 2010]



Delauney
Triangulation



Minimum
Spanning Tree



k-nearest
Neighbors

On Distance Metrics

Example: Euclidian Distance

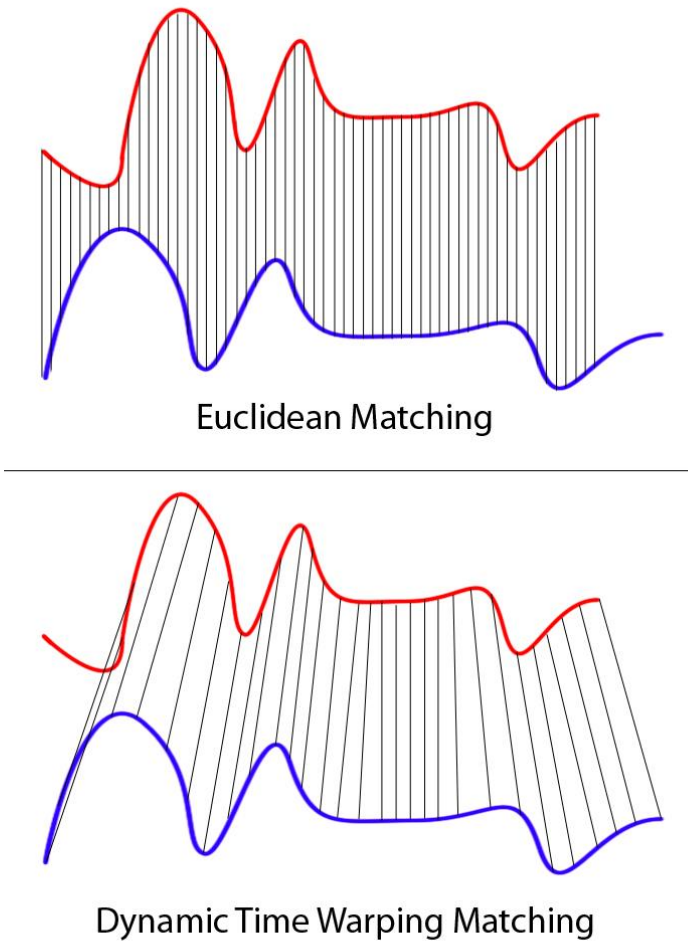


image taken from Wikipedia

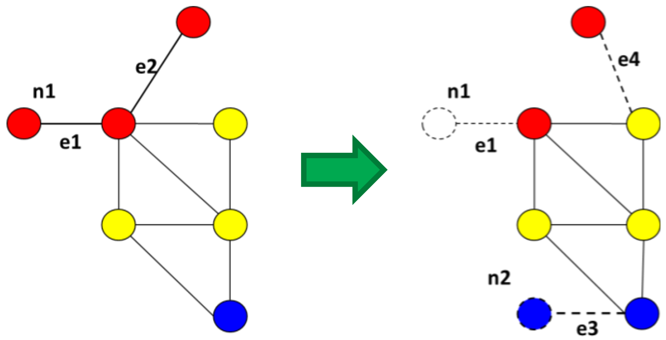
Further Reading:
Ding et al. 2008 – "Querying and mining of time series data: experimental comparison of representations and distance measures"

Aghabozorgi et al. 2015 – "Time-series clustering – A decade review"

Bagnall et al. 2017 – "The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances"

On Distance Metrics

Example: Edit Distance



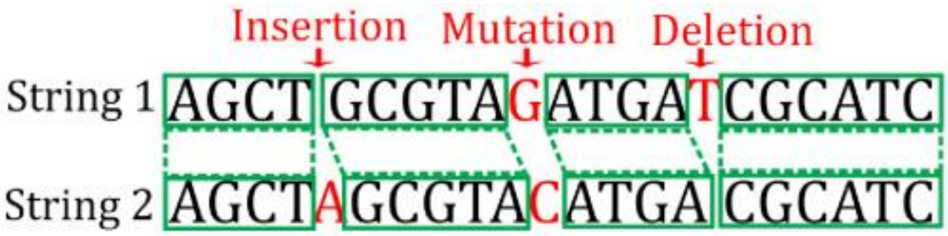
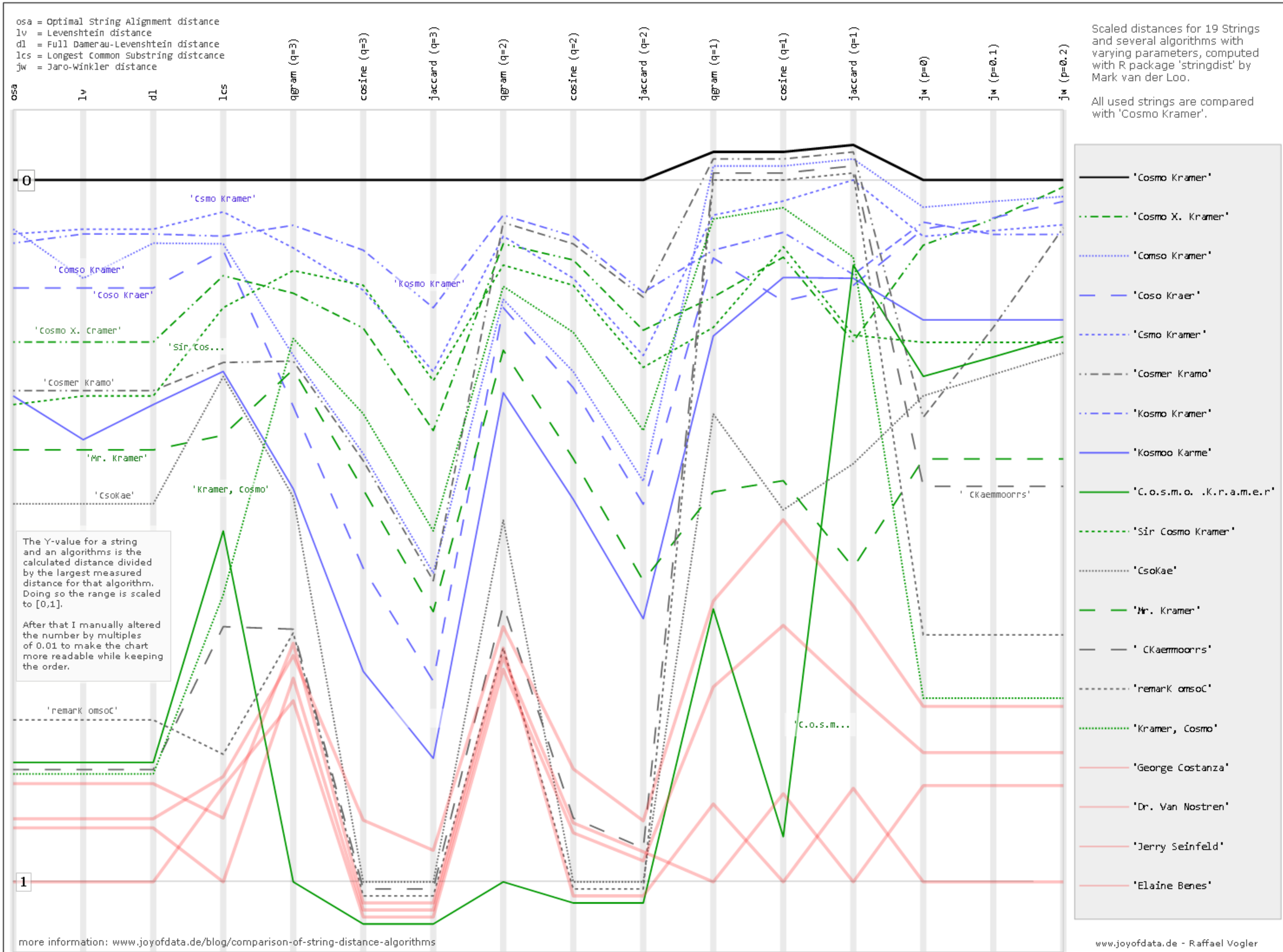


image source: http://gedevo.mpi-inf.mpg.de

image taken from [Maleki et al. 2017]

Further Reading:
Emmert-Streib et al. 2016 – "Fifty years of graph matching, network alignment and network comparison"

Further Reading:
Yu et al. 2016 – "String similarity search and join: a survey"



https://www.joyofdata.de/blog/comparison-of-string-distance-algorithms/

A Survey of Measures and Methods for Matching Geospatial Vector Datasets

EMERSON M. A. XAVIER, Brazilian Army Geographic Service
FRANCISCO J. ARIZA-LÓPEZ and MANUEL A. UREÑA-CÁMARA, Universidad de Jaén

The field of Geographical Information Systems (GIS) has experienced a rapid and ongoing growth of available sources for geospatial data. This growth has demanded more data integration in order to explore the benefits of these data further. However, many data providers implies many points of view for the same phenomena: geospatial features. We need sophisticated procedures aiming to find the correspondences between two vector datasets, a process named *geospatial data matching*. Similarity measures are key-tools for matching methods, so it is interesting to review these concepts together. This article provides a survey of 30 years of research into the measures and methods facing geospatial data matching. Our survey presents related work and develops a common taxonomy that permits us to compare measures and methods. This study points out relevant issues that may help to discover the potential of these approaches in many applications, like data integration, conflation, quality evaluation, and data management.

CCS Concepts: • **Information systems** → **Geographic information systems**; **Similarity measures**

Additional Key Words and Phrases: Spatial data integration, map conflation, geometric algorithms

ACM Reference Format:
Emerson M. A. Xavier, Francisco J. Ariza-López, and Manuel A. Ureña-Cámara. 2016. A survey of measures and methods for matching geospatial vector datasets. ACM Comput. Surv. 49, 2, Article 39 (August 2016), 34 pages.
DOI: http://dx.doi.org/10.1145/2963147

Knowl Inf Syst (2011) 27:1–21
DOI 10.1007/s10115-010-0285-0

SURVEY PAPER

Approximate data instance matching: a survey

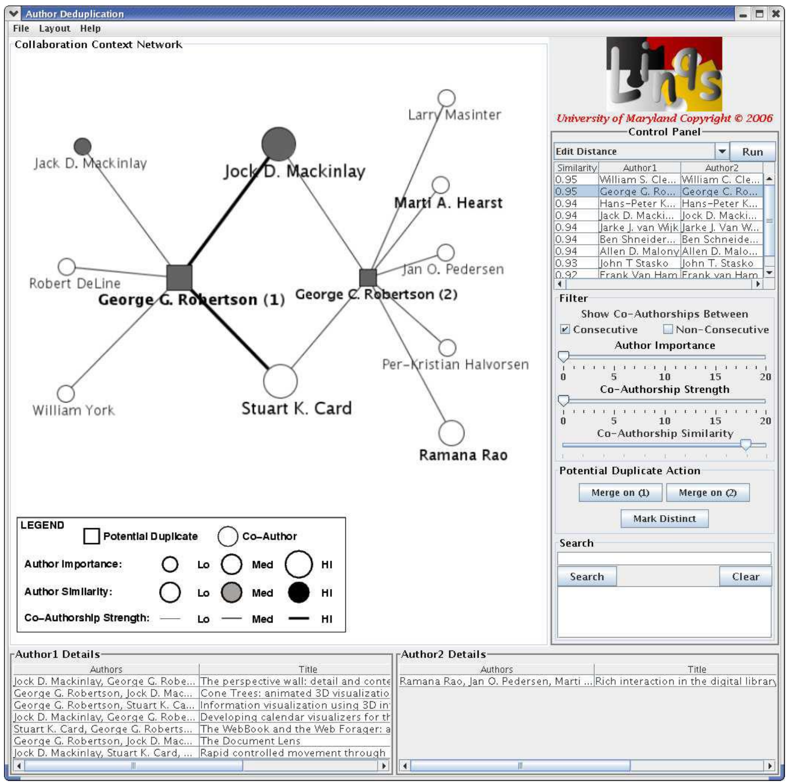
Carina Friedrich Dorneles · Rodrigo Gonçalves ·
Ronaldo dos Santos Mello

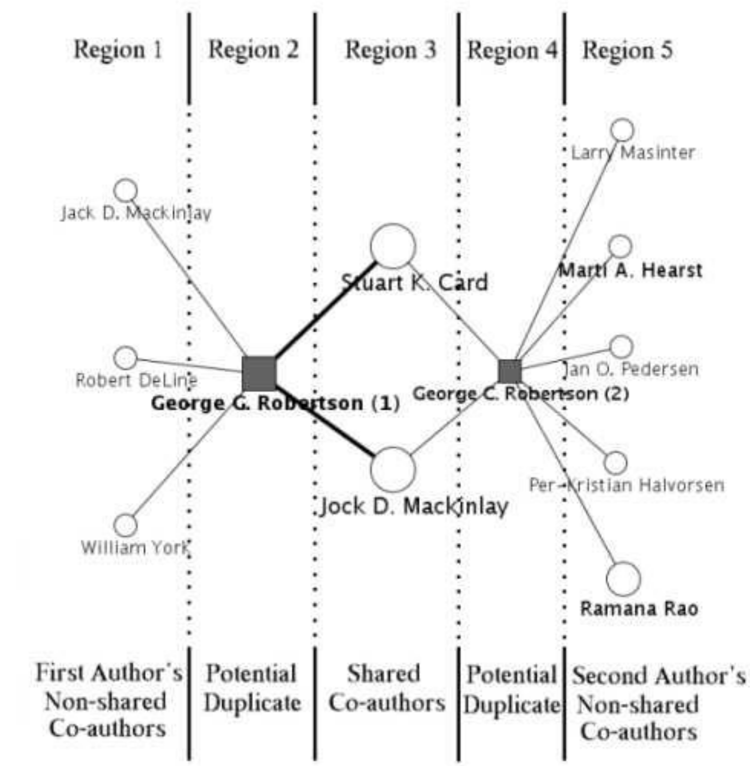
Received: 18 December 2008 / Revised: 20 October 2009 / Accepted: 16 January 2010 /
Published online: 9 April 2010

Abstract Approximate data matching is a central problem in several data management processes, such as data integration, data cleaning, approximate queries, similarity search and so on. An approximate matching process aims at defining whether two data represent the same real-world object. For atomic values (strings, dates, etc), similarity functions have been defined for several value domains (person names, addresses, and so on). For matching aggregated values, such as relational tuples and XML trees, approaches alternate from the definition of simple functions that combine values of similarity of record attributes to sophisticated techniques based on machine learning, for example. For complex data comparison, including structured and semistructured documents, existing approaches use both structure and data for the comparison, by either considering or not considering data semantics. This survey presents terminology and concepts that base approximated data matching, as well as discusses related work on the use of similarity functions in such a subject.

Keywords Instance data matching · Similarity function · Similarity matching · Record linkage · Record matching · Duplicate detection · Object matching · Entity resolution

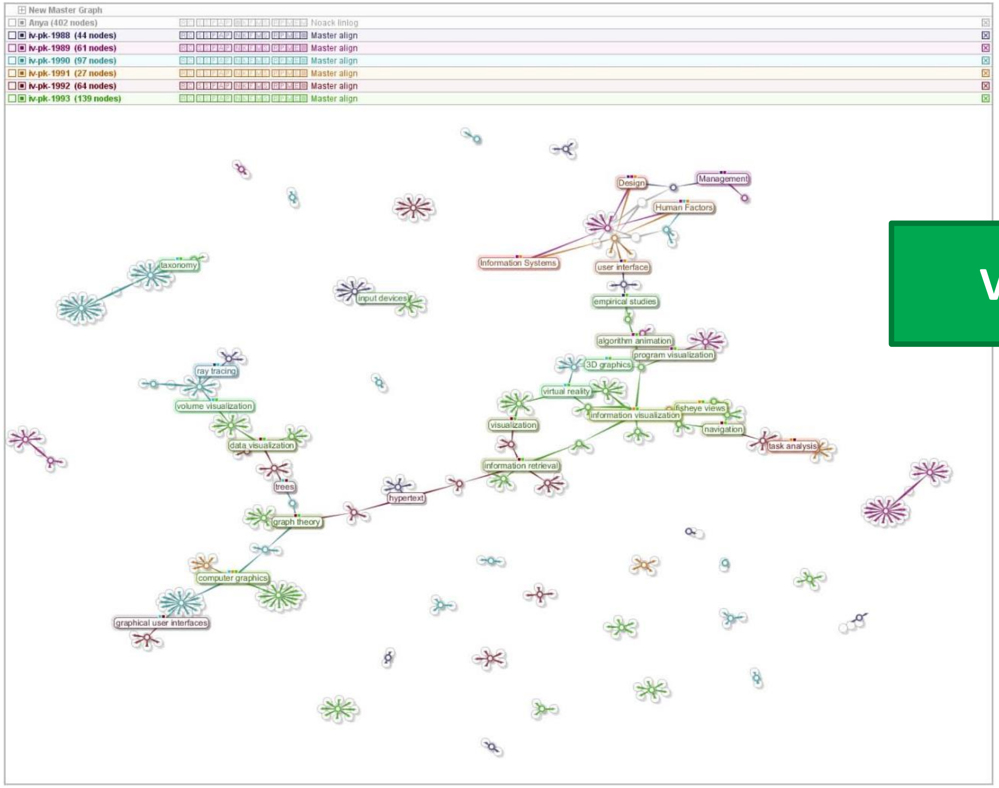
Example: Interactive Entity Resolution with D-Dupe





[Bilgic et al. 2006]

Example: Multi-Layer Comparison for Multiple Graphs

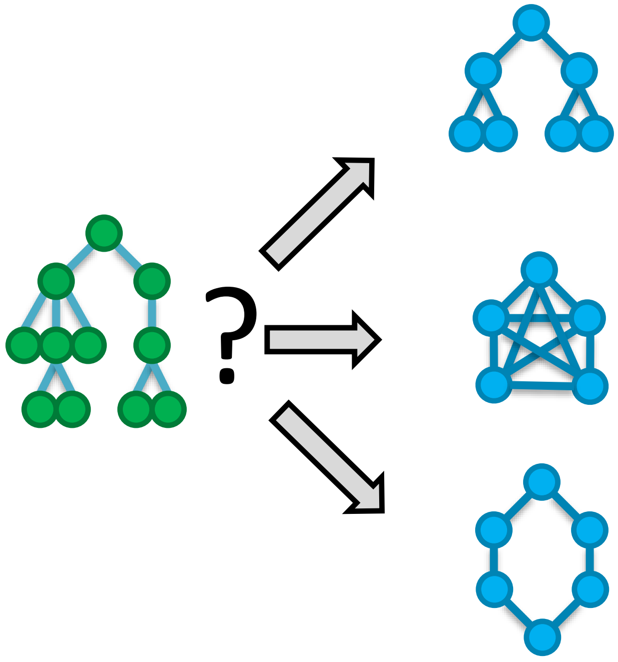


VIDEO

[Hascoët & Dragicevic 2012]

1-to-many Comparison

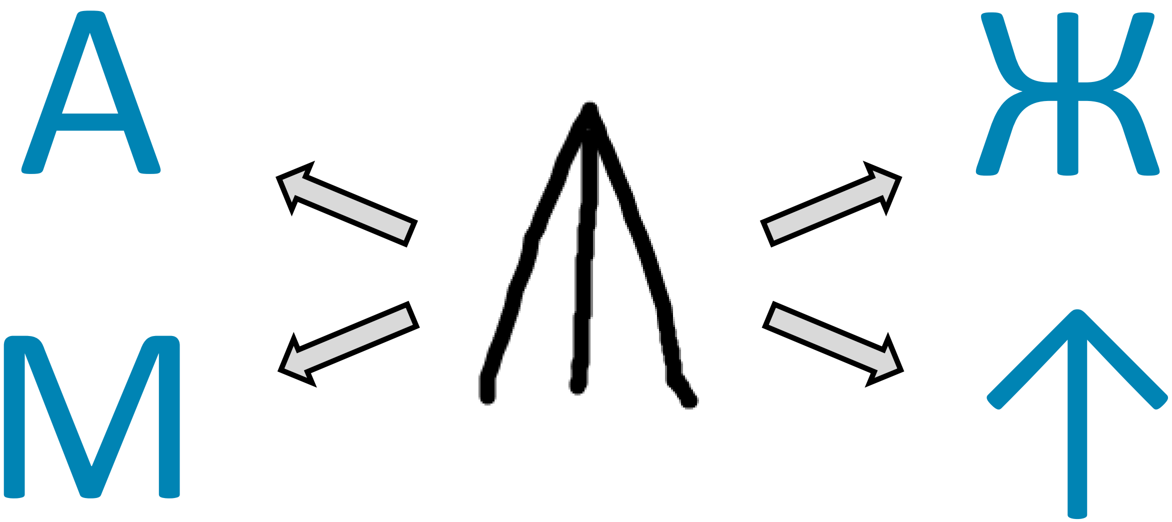
CLASSIFICATION



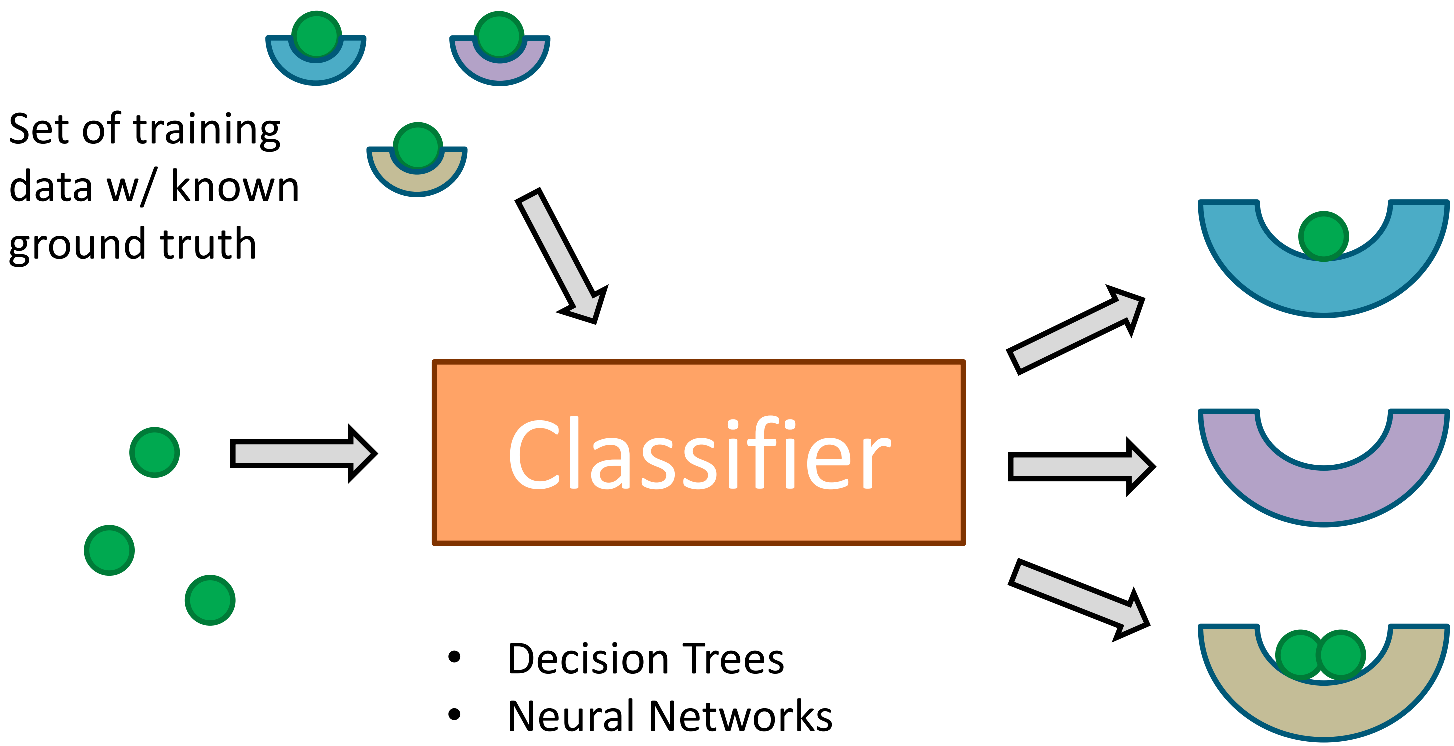
Definition

Classification:

Assignment of a data object to a particular group according to its features, structure, or origin.



Classification as Supervised Learning



- Decision Trees
- Neural Networks
- Support Vector Machines
- Kernel Density Estimation
- …

Decision Trees

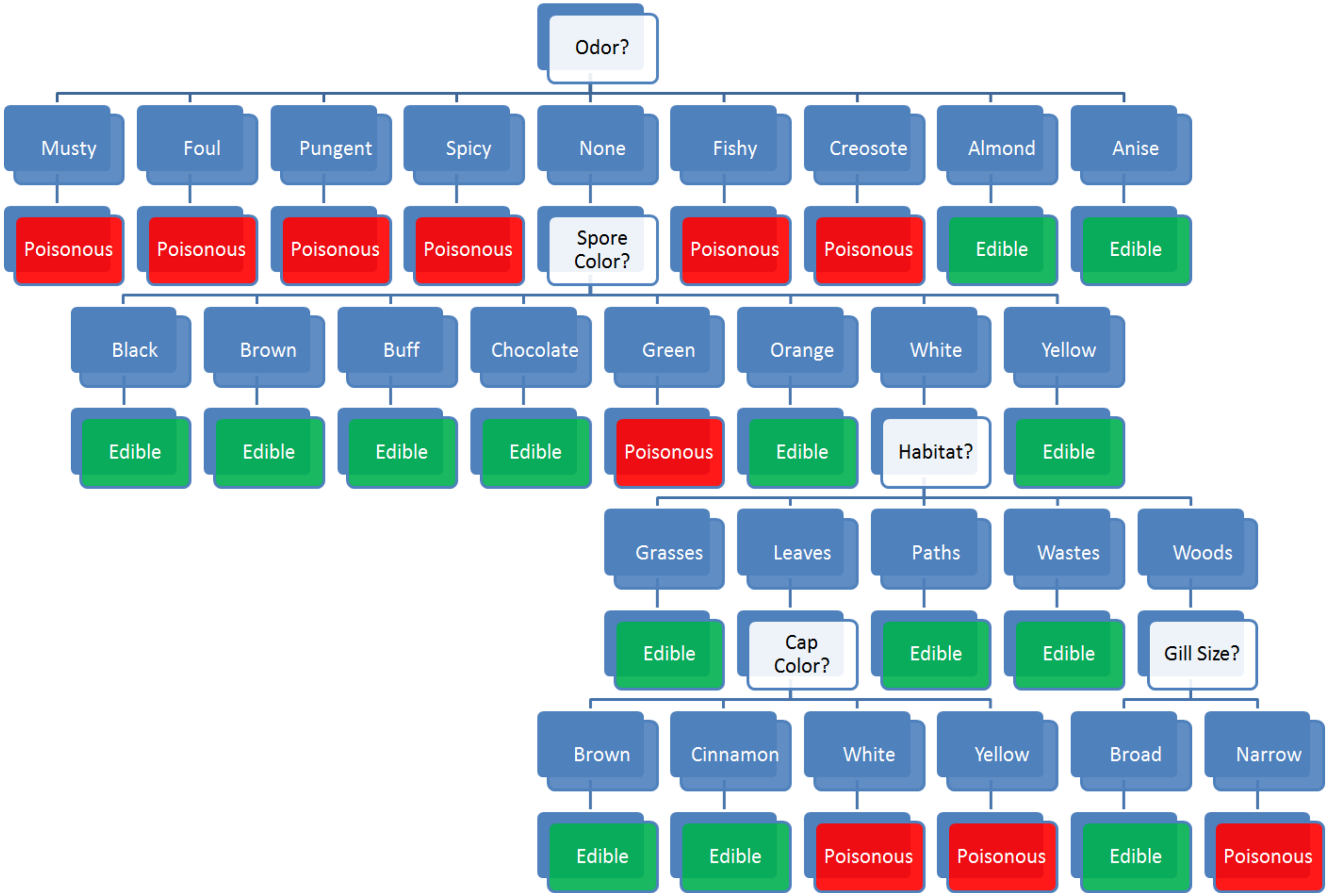


image taken from https://gieseanw.wordpress.com/2012/03/03/decision-tree-learning/

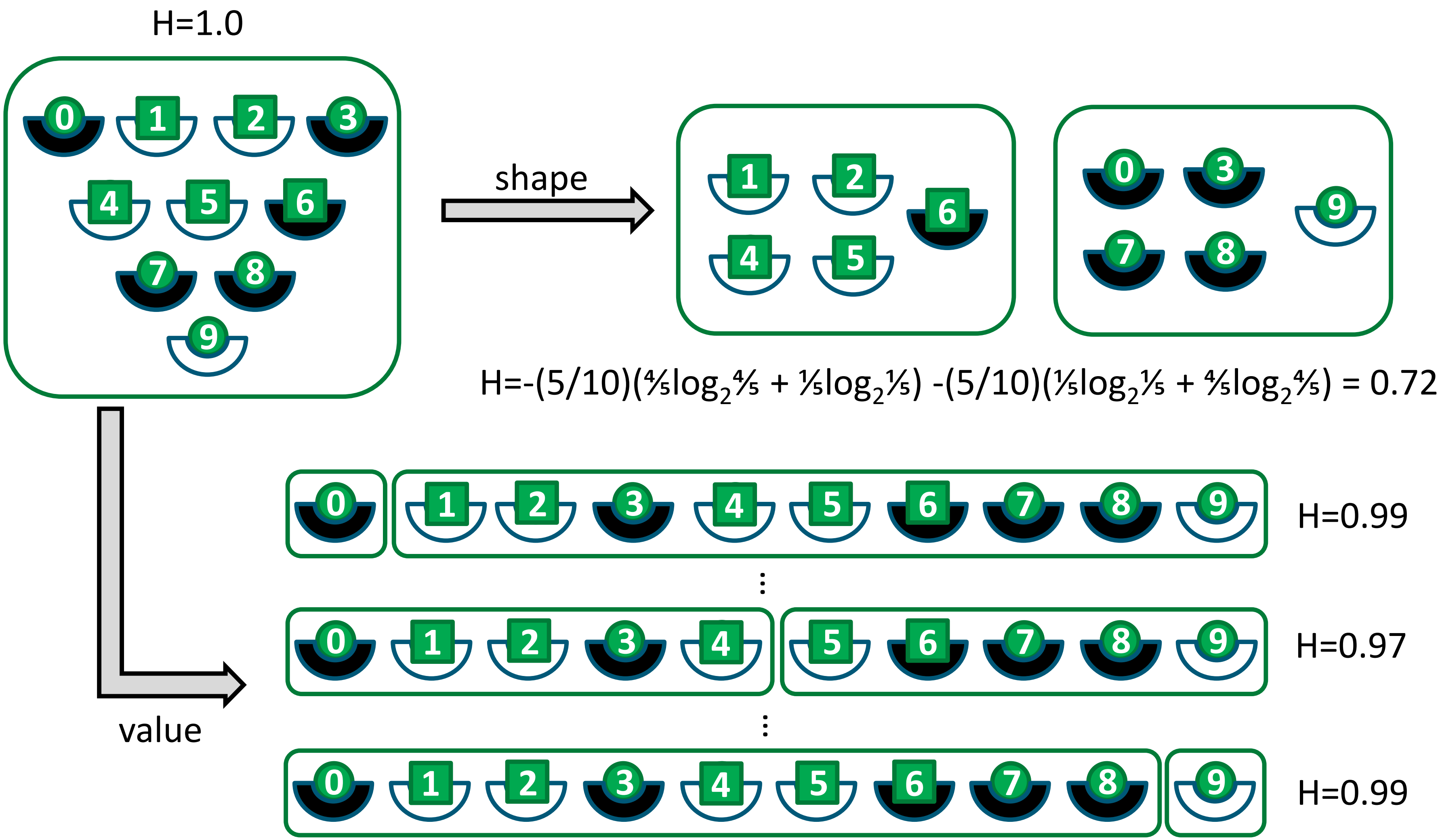
How to Build a Decision Tree?

Split(Dataset):

For all data attributes A in data:

  Compute entropy gain when splitting by A

Decision Tree Example



How to Build a Decision Tree?

Split(Dataset):

For all data attributes A in data:

        Compute entropy gain when splitting by A

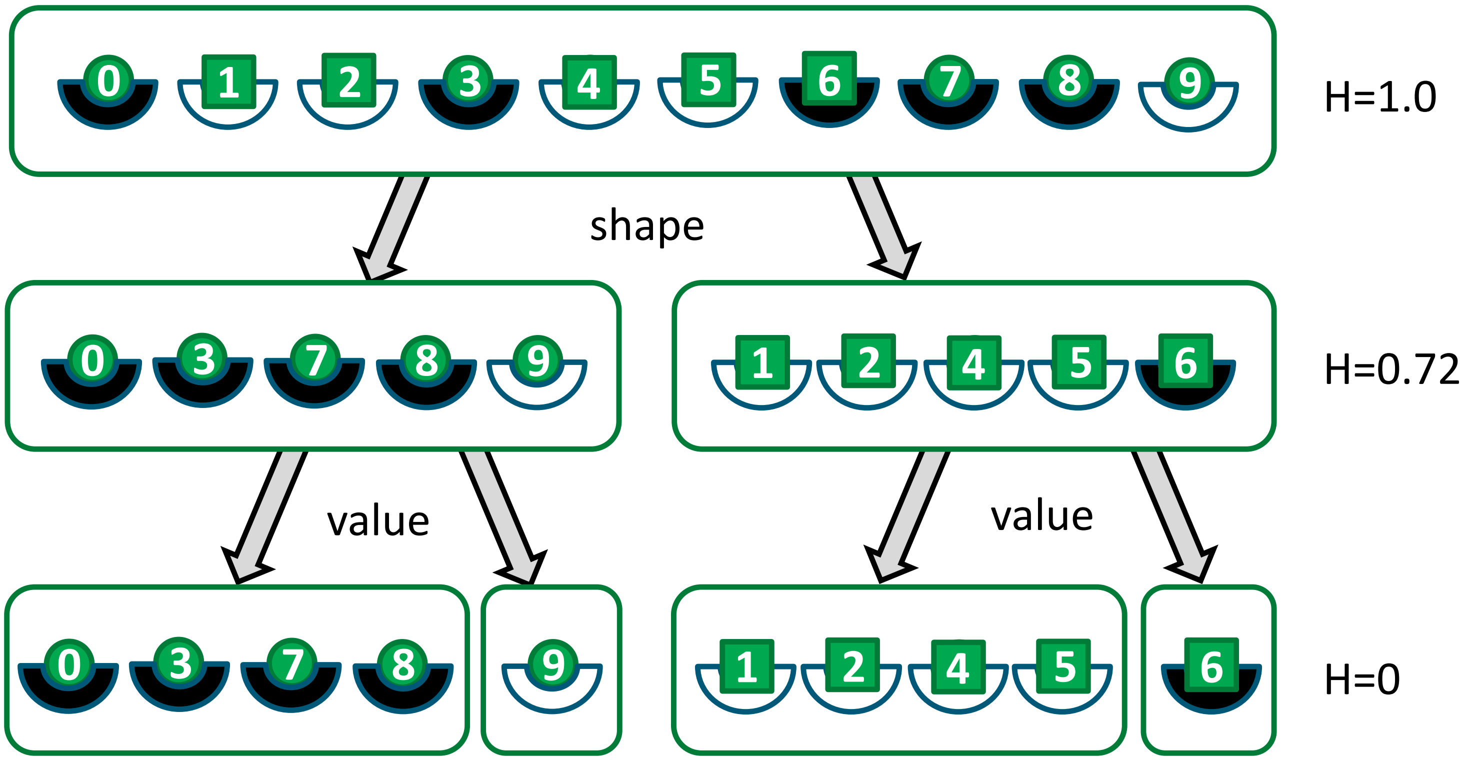
Subdivide Dataset by A w/ highest gain

For all generated data subsets:

        If entropy(Data subset)>0

                Split(Data subset)

Decision Tree Example



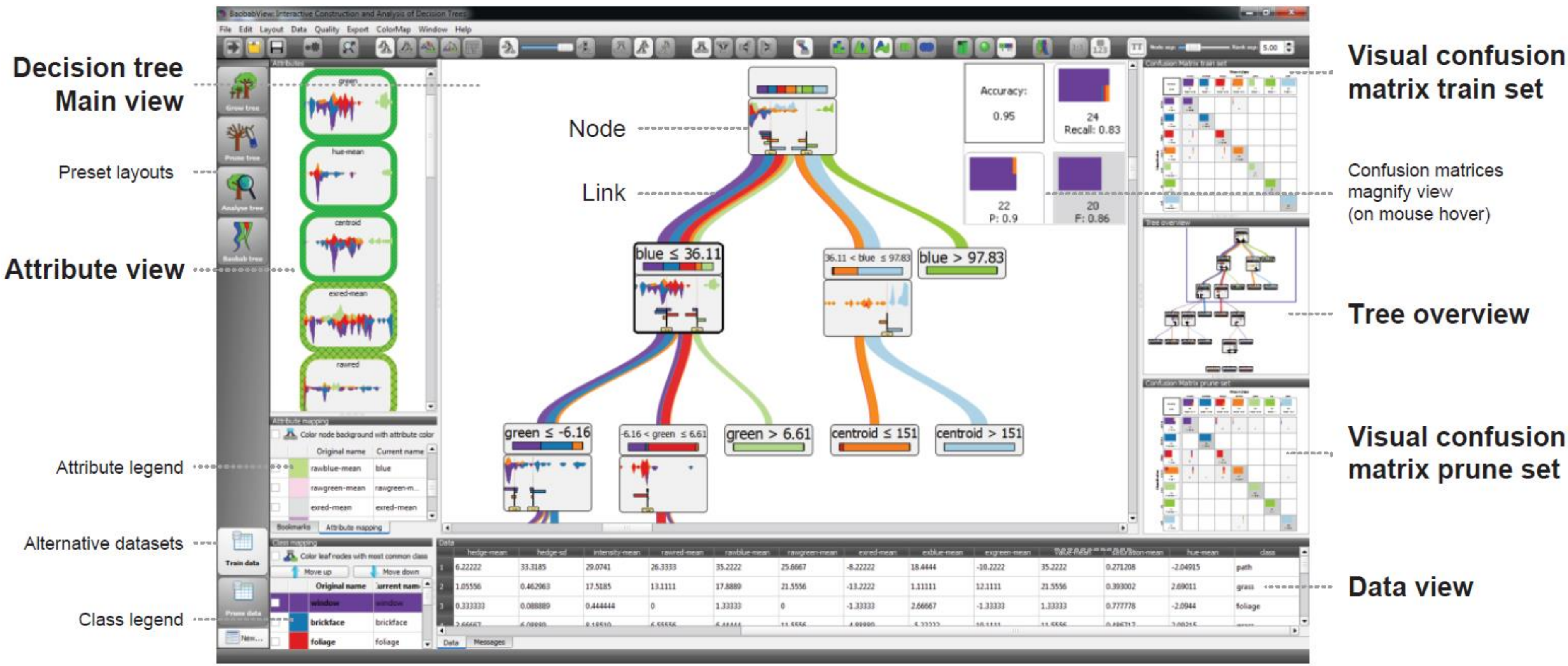
How to Build a Decision Tree?

Algorithms: ID3, C4.5, C5.0 [Quinlan 1993]

Downsides:

- No optimality guarantee (greedy approach)
  --> backtracking to escape local minima
- Possible overfitting (large trees)
  --> pruning of subtrees, or random forests
- Hard splits on continuous data
  --> soft splits that evaluate both branches

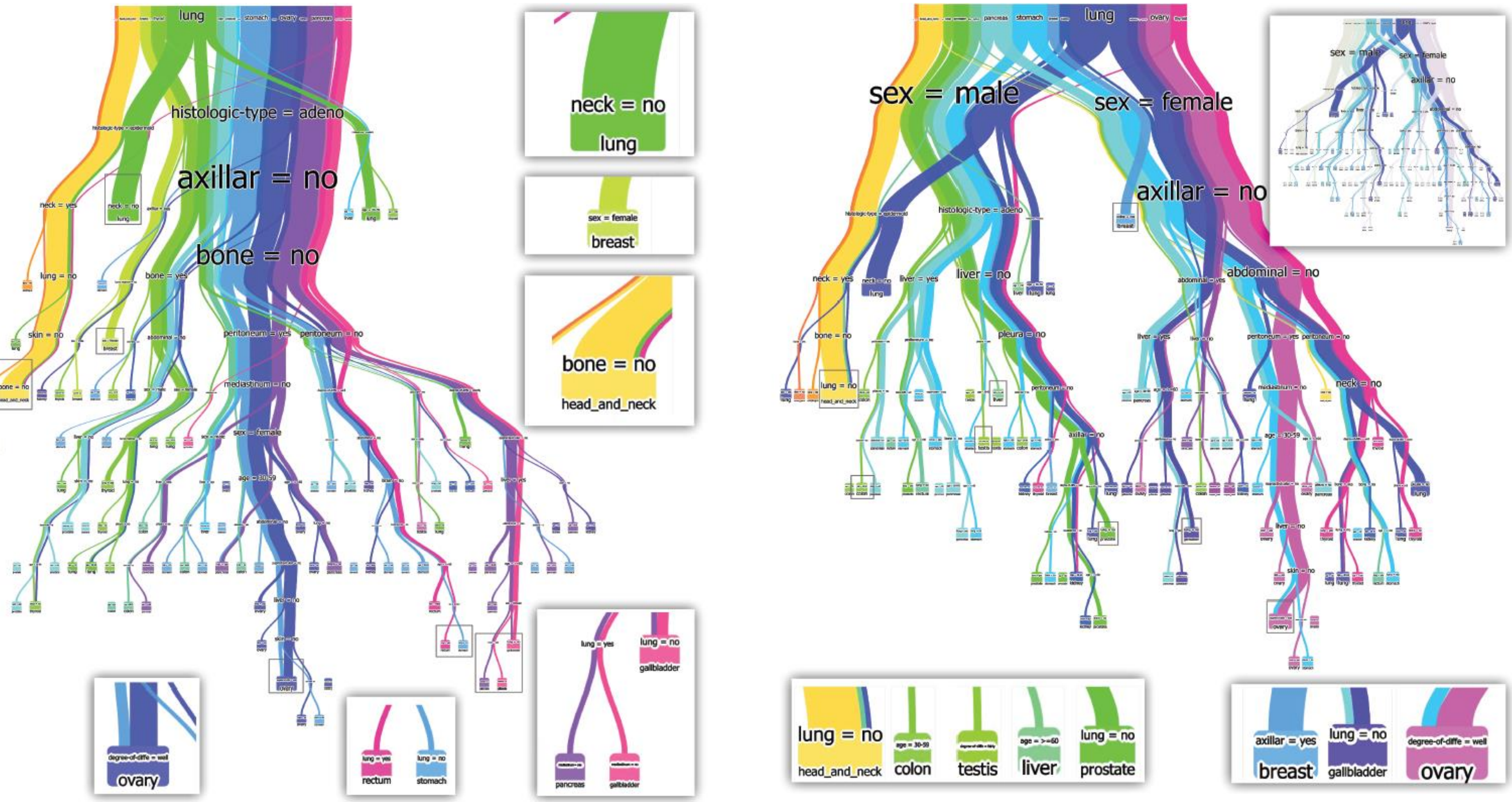
Example: Interactive Decision Tree Construction with BaobabView



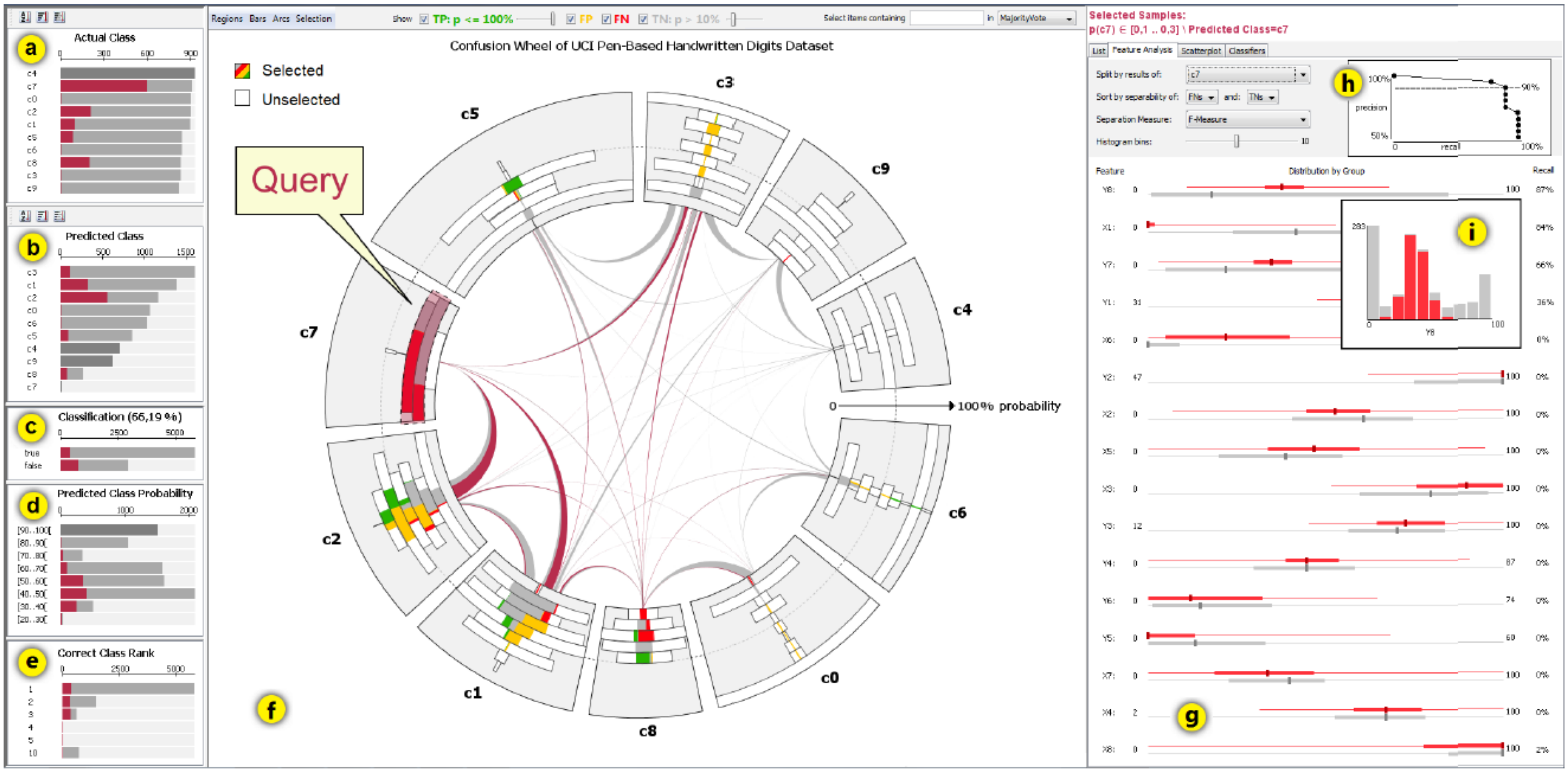
[v.d.Elzen & v.Wijk 2011]

Example: Interactive Decision Tree Construction with BaobabView

[v.d.Elzen & v.Wijk 2011]

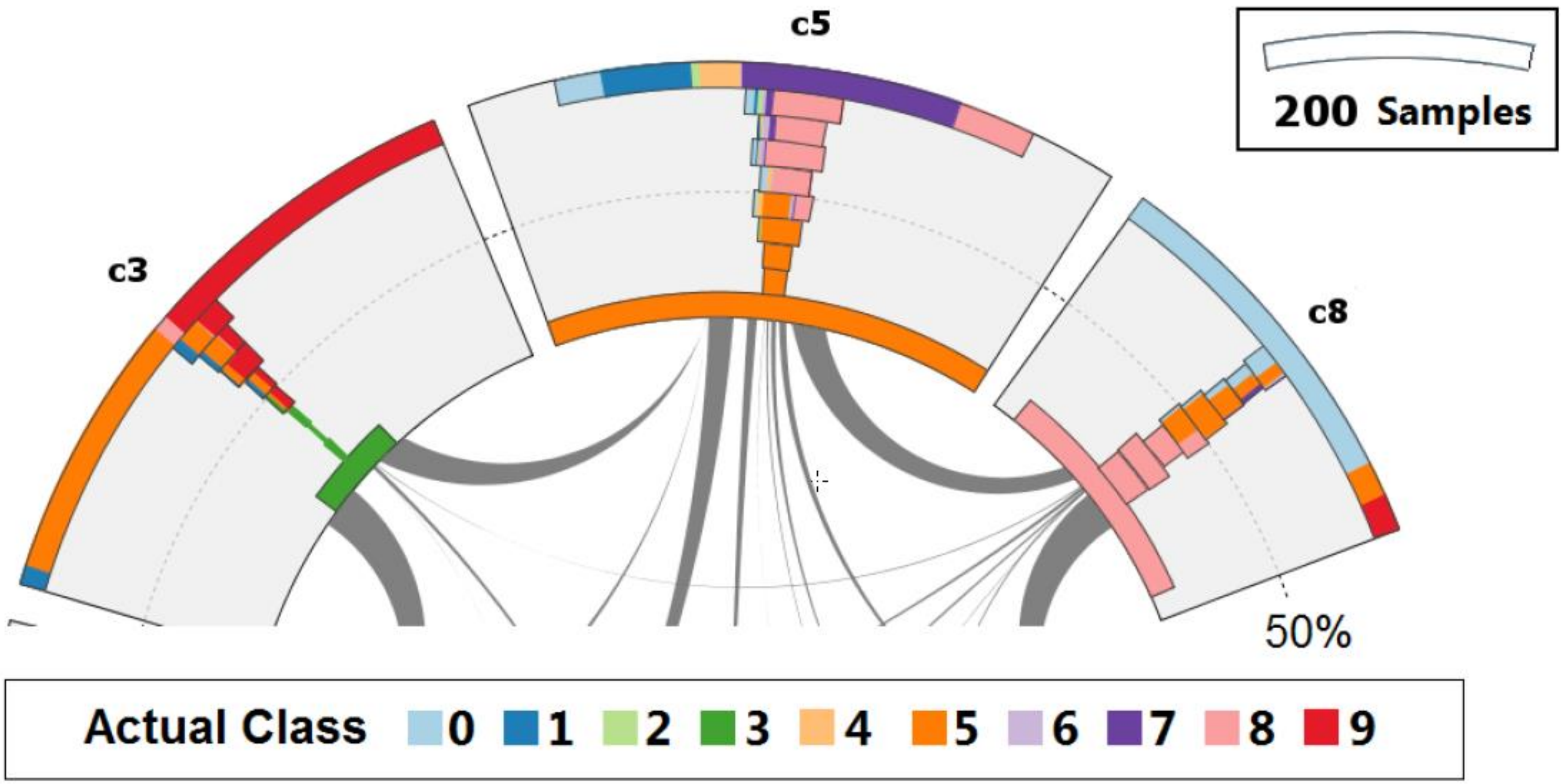


Example: Analysis of Probabilistic Classifiers with Confusion Wheels



[Alsallakh et al. 2014]

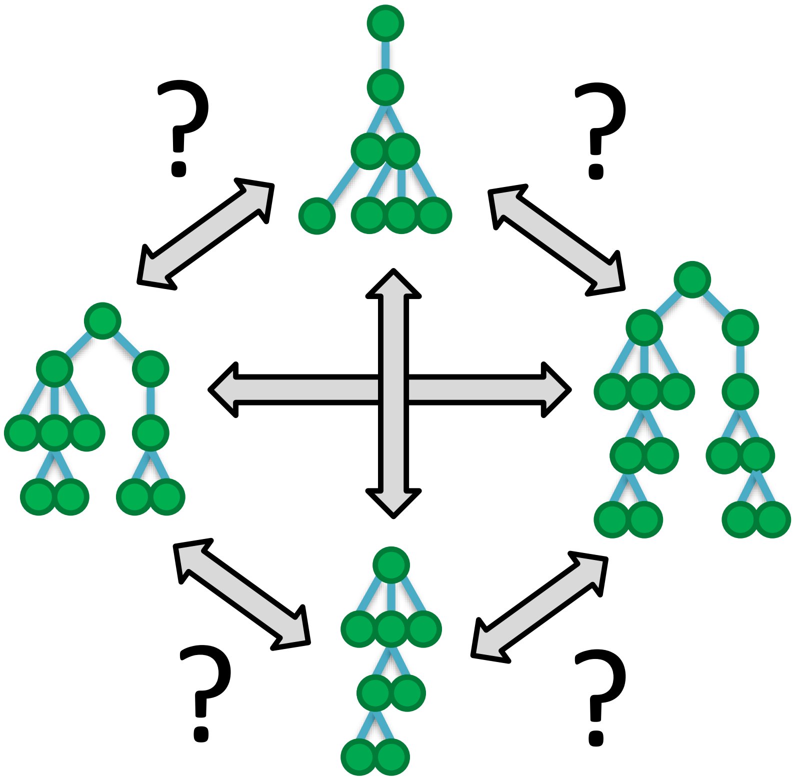
Example: Analysis of Probabilistic Classifiers with Confusion Wheels



[Alsallakh et al. 2014]

many-to-many Comparison

CLUSTERING



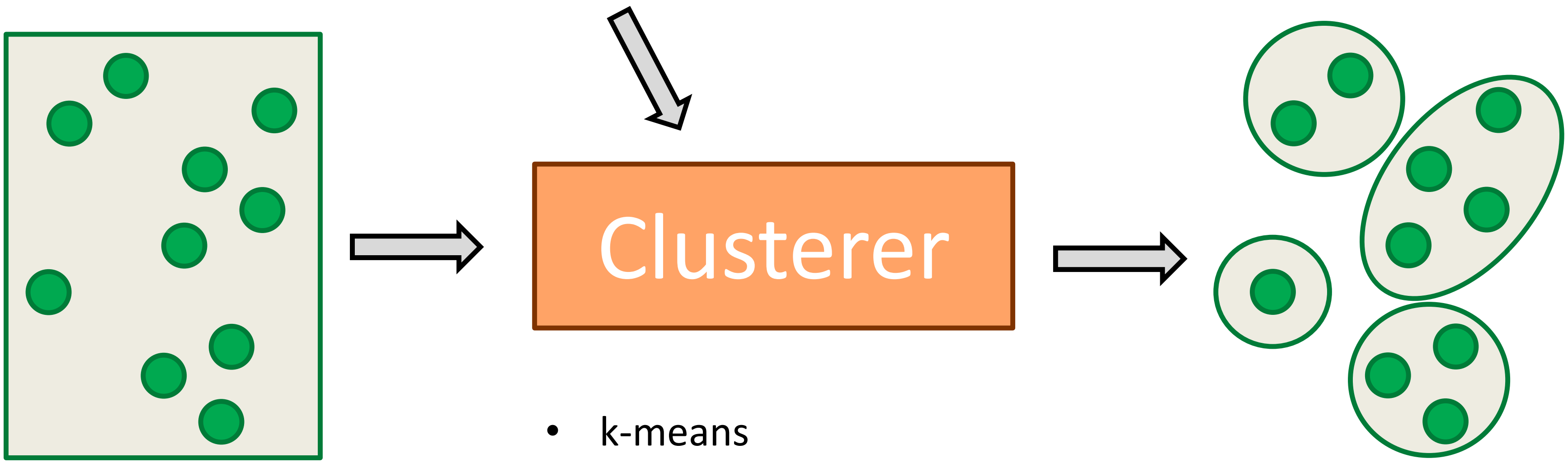
Definition

Clustering: Gathering a number of related data objects into groups, so that data objects within one group are more related to each other than to data objects from other groups.

- **given:** (dis-)similarity measure/matrix
  - n-dimensional, numerical data: Euclidean Distance
  - network data: Graph-theoretic Distance
  - strings of text: Edit Distance
- **sought:** grouping of the data w.r.t. that measure

Clustering as Unsupervised Learning

Similarity metric / matrix



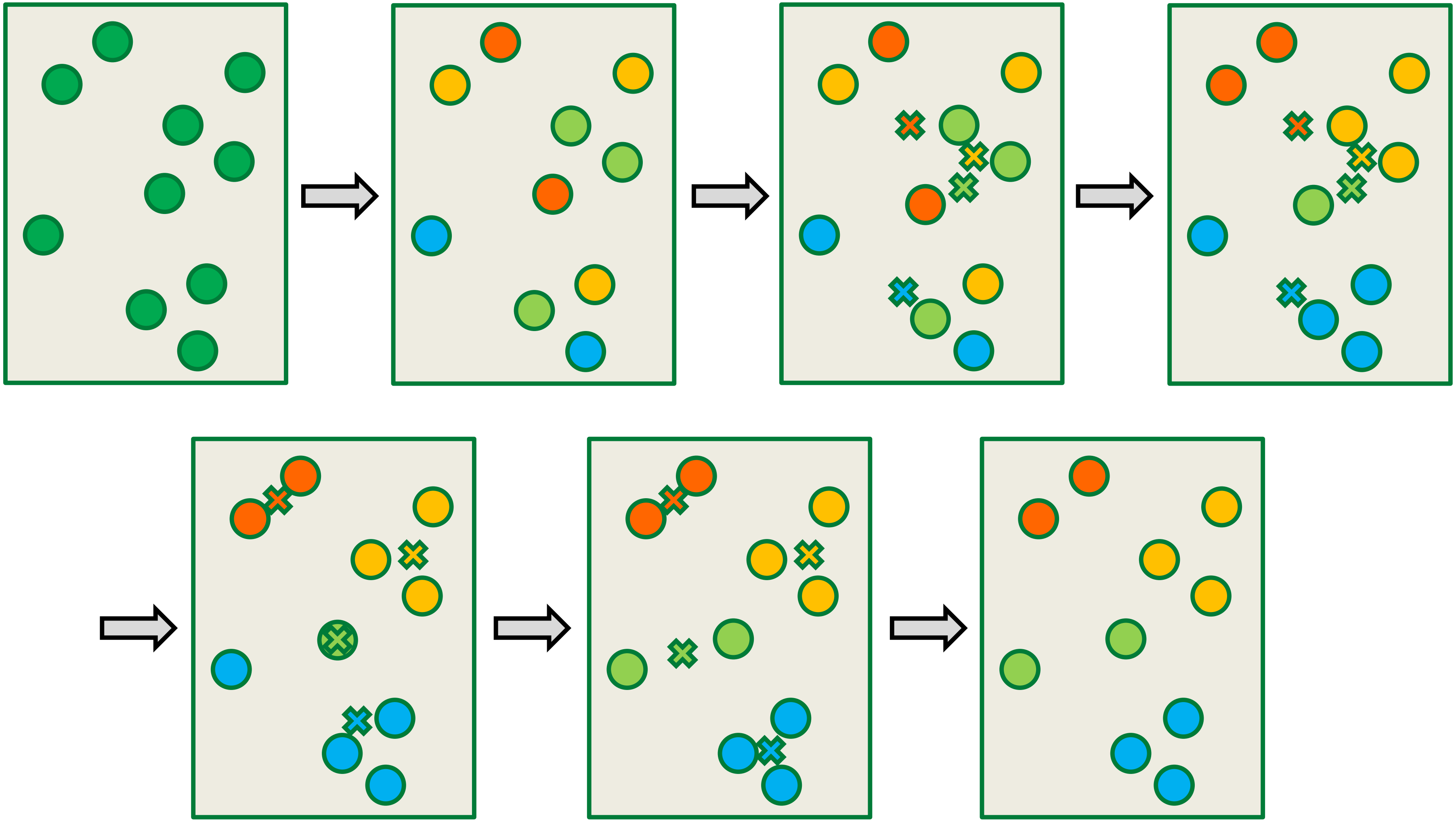
- k-means
- DBSCAN
- Hierarchical Clustering
- Spectral Clustering
- …

What makes a “good” clustering?

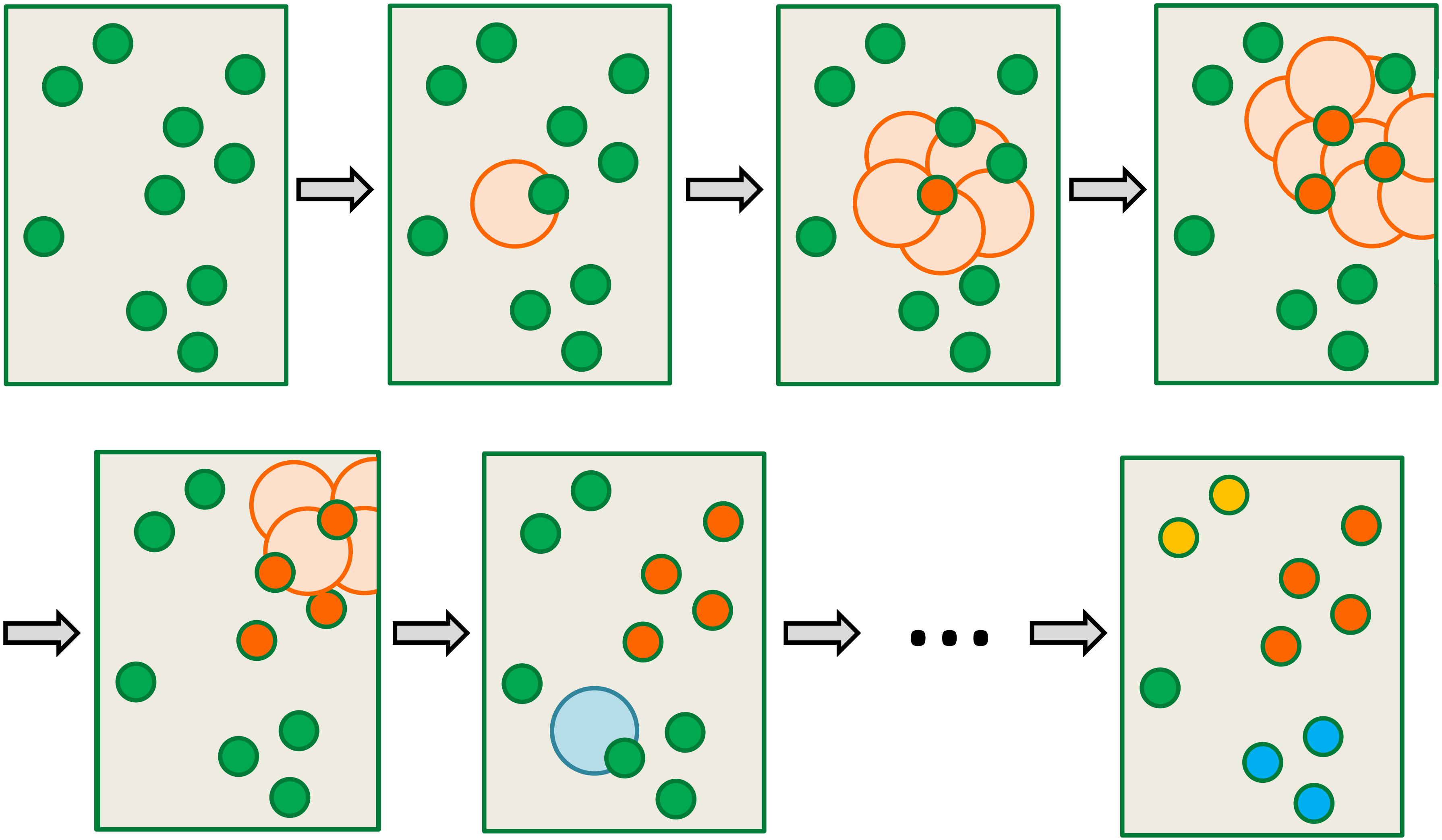
- **Compact:** elements in cluster are similar
- **Separated:** clusters are different
- **Balanced:** cluster membership is equally probable
- **Parsimonious:** much fewer clusters than data objects

Source: Cosma Shalizi (2009)

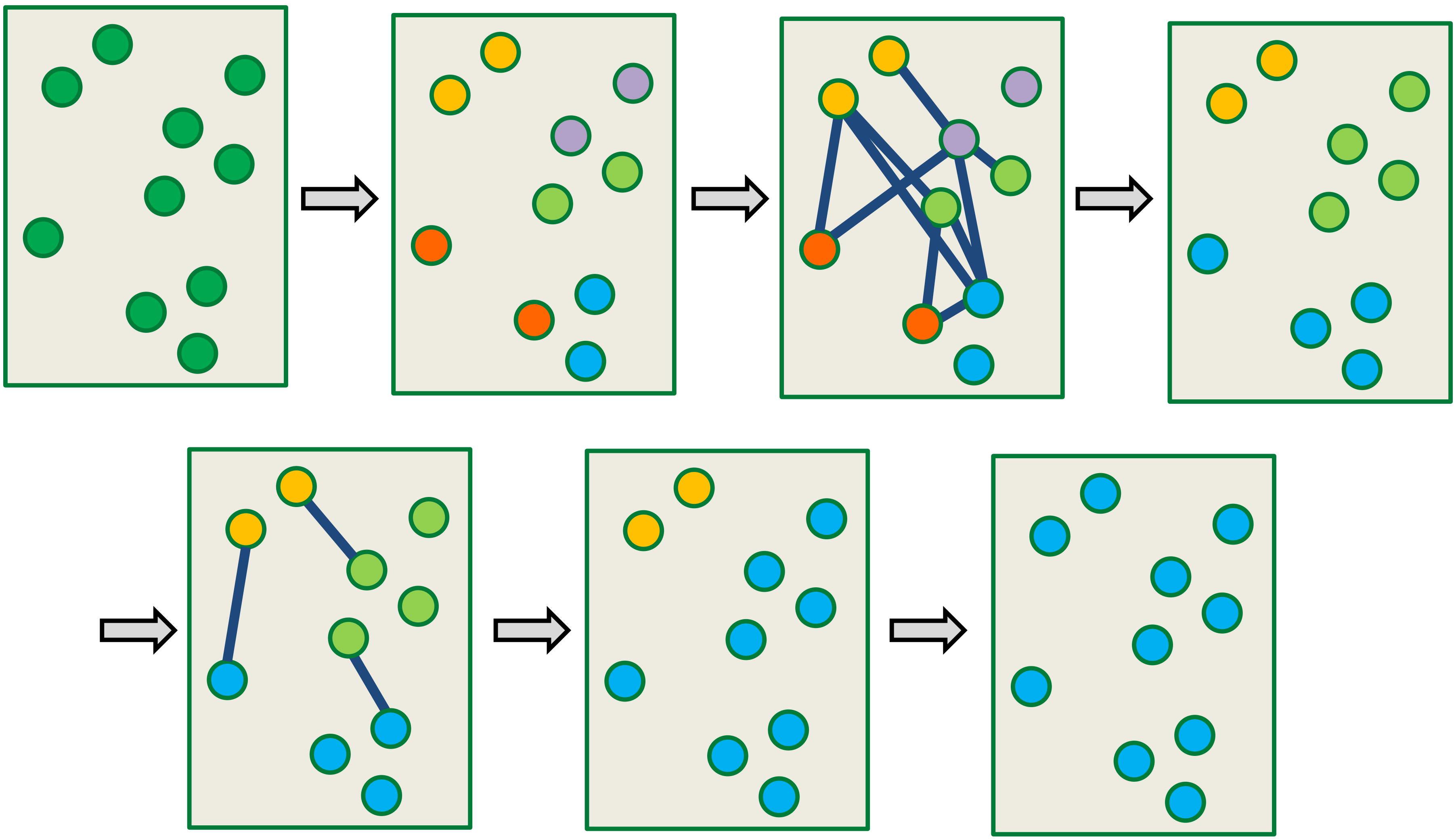
k-Means Clustering



DBSCAN



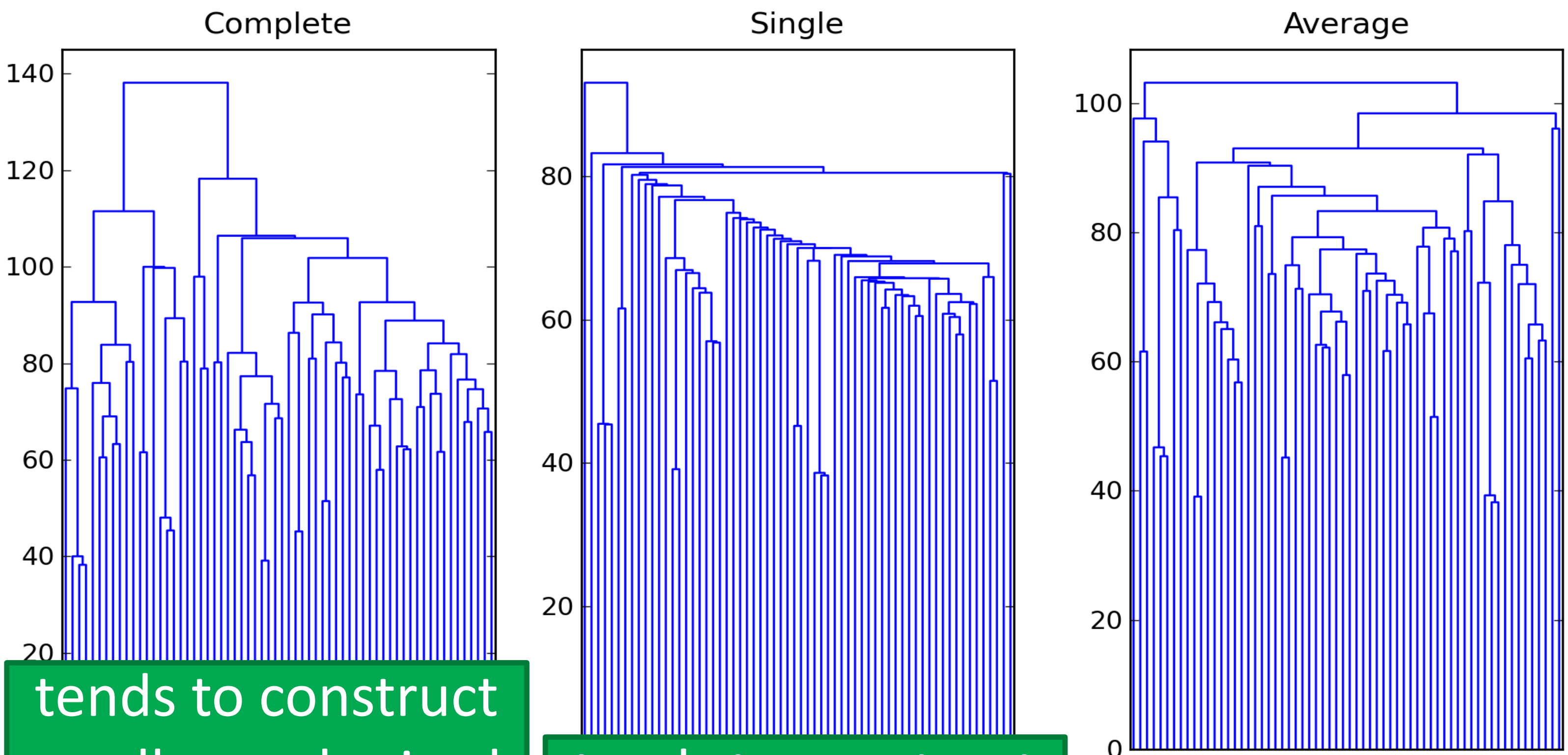
Hierarchical Clustering



Types of Hierarchical Clustering

- Directionality of the clustering:
  - **Top-down:** divisive
  - **Bottom-up:** agglomerative
- Linkage metrics:
  - **Single Linkage:** nearest neighbor
  - **Complete Linkage:** farthest neighbor
  - **Average Linkage:** all neighbors

Effect of different Linkage metrics



tends to construct small, evenly sized clusters

tends to construct chains of clusters

Images taken from Jonathan Taylor (2010)

Comparison

k-Means: O(n) runtime, but requires parameter k

--> elbow method

DBSCAN: O(n*log n) runtime, no k required, robust to outliers, but problematic for uneven density distributions

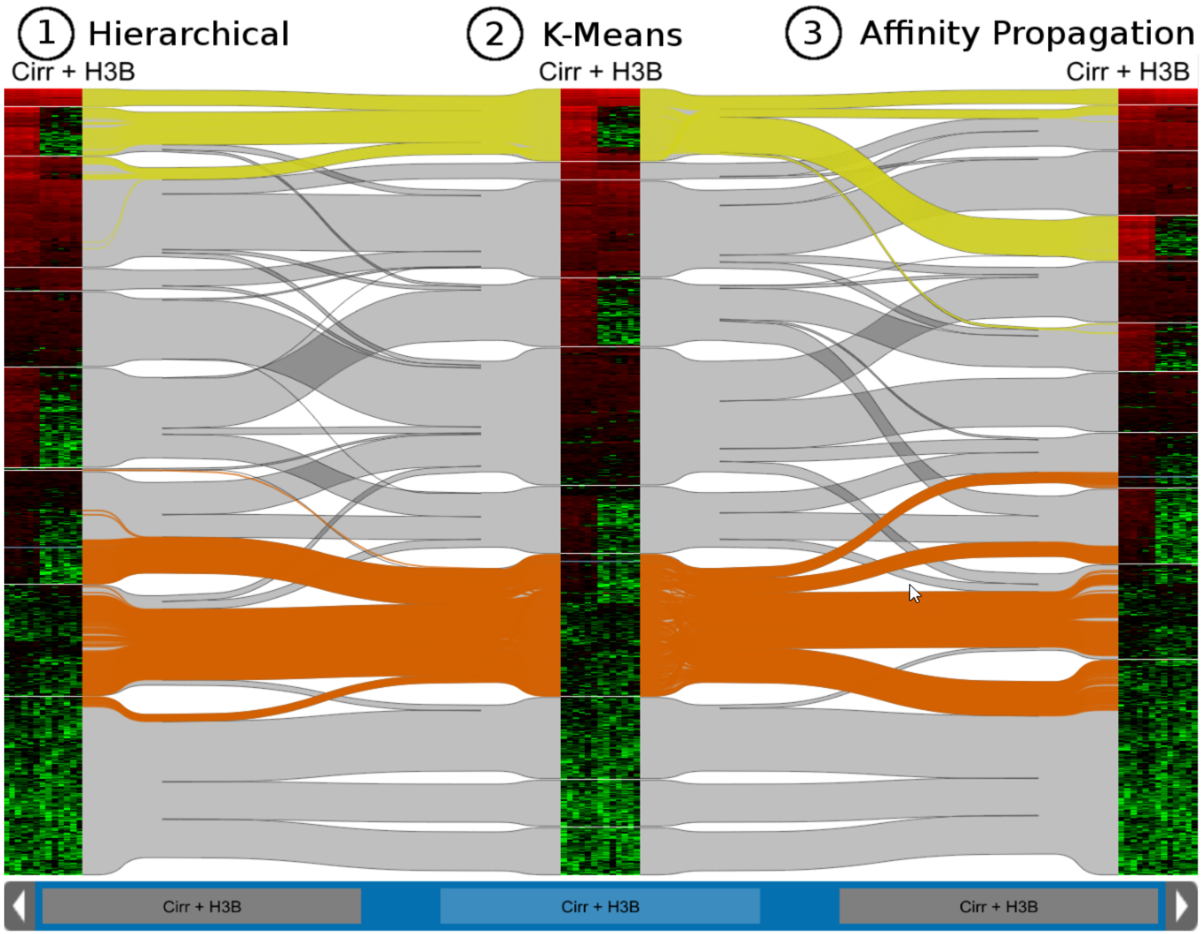
--> hierarchical variant: OPTICS or HDBSCAN

Hierarchical Clustering: O(n^2) runtime, yields multiple cluster granularities in a single run

When there is no obviously “right” way to cluster…

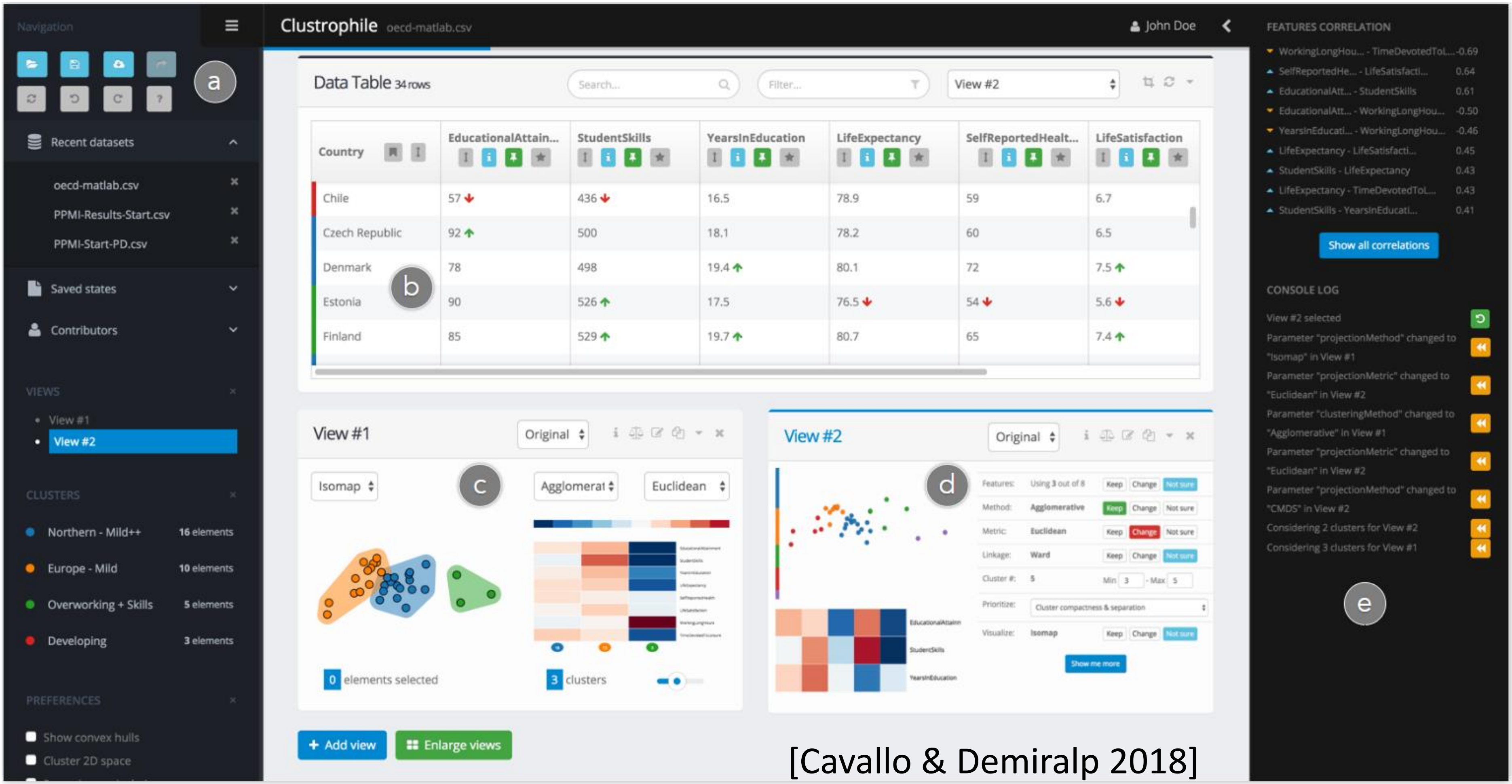
- Consensus Clustering
  - NP complete
- Heuristics:
  - **Quantitative/metric-based: CSPA**
    Cluster-based Similarity Partitioning Algorithm
  - **Structural/graph-based: HGPA**
    Hyper-Graph Partitioning Algorithm

Example: Compare Clusterings with Caleydo Matchmaker



[Lex et al. 2010]

Example: Comparison between Cluster Parameters with Clustrophile



[Cavallo & Demiralp 2018]