Tree Clustering

Valeriy Khakhutskyy

Technische Universität München
Adviser: Prof. Dr. Dr. Fabian Theis

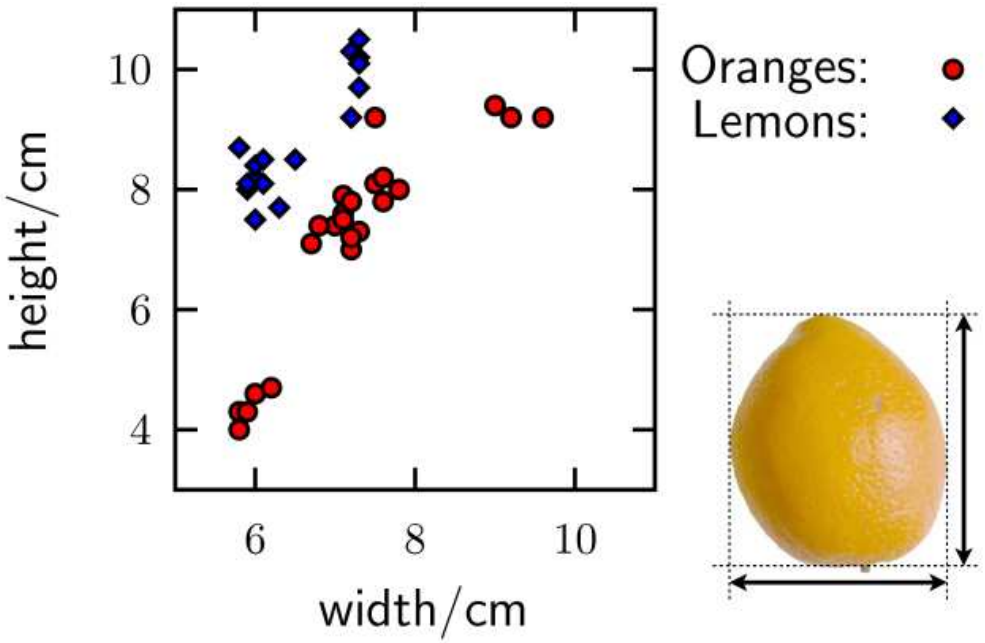
March 21, 2011

Agenda

Clustering

Clustering is a method of unsupervised learning to discover

- natural partitions
- new patterns
- hidden properties



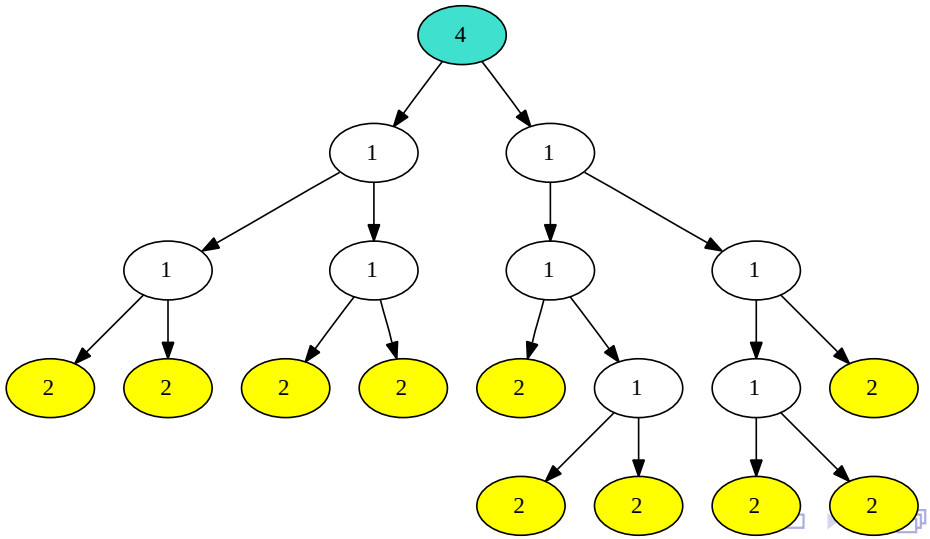
Source: [1]

Tree Data-Structures

Trees are used to describe hierarchical data in different scientific areas, i.e.

- linguistics (XML)
- image analysis (JPEG)
- compiler optimisation (syntax trees)
- computational biology (cell differentiation)

Interesting for us: unordered labelled trees



Overview


Tree Structured Data
Metric Definition
Centroid Tree
Clustering Algorithm
Data partitioning
Characteristic trees
Pattern recognition

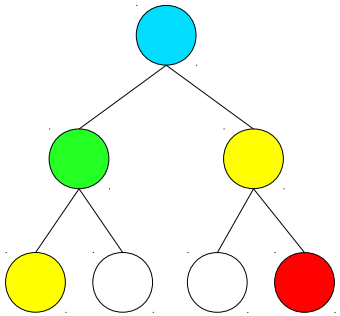
Metrics: Constrained Tree Edit Distance [3]

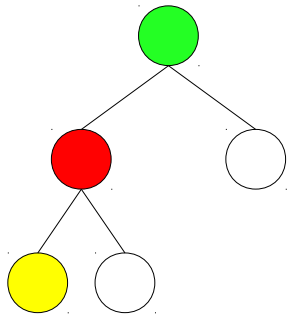
Operations:

- Change label
- Add node
- Remove node

Constraints:

- Proper edit distance
- Subtrees mapped to subtrees





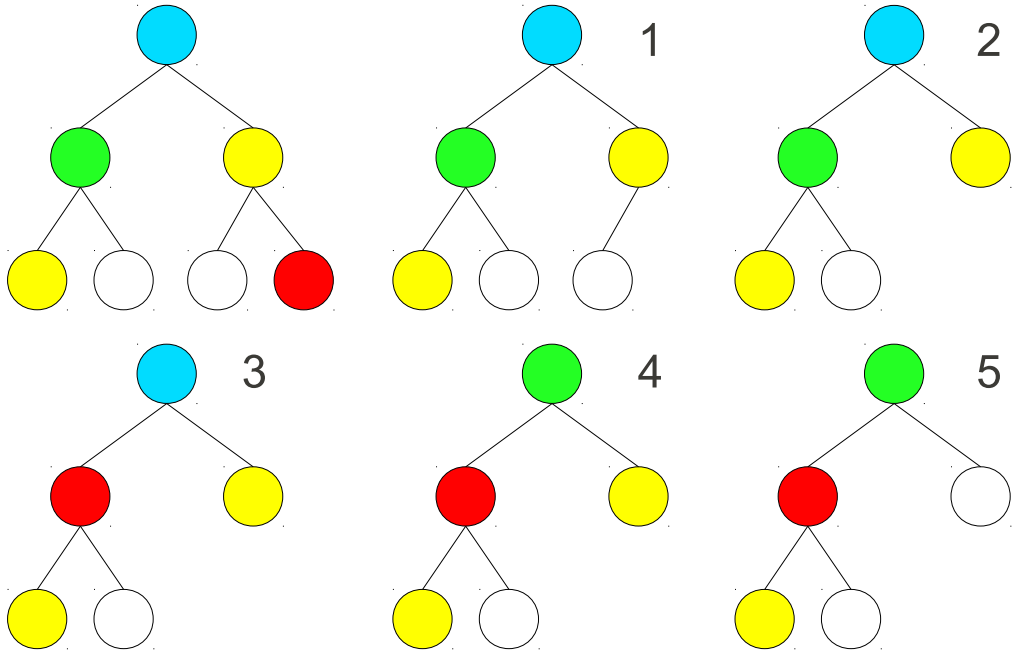
Metrics: Constrained Tree Edit Distance [3]

Operations:

- Change label
- Add node
- Remove node

Constraints:

- Proper edit distance
- Subtrees mapped to subtrees



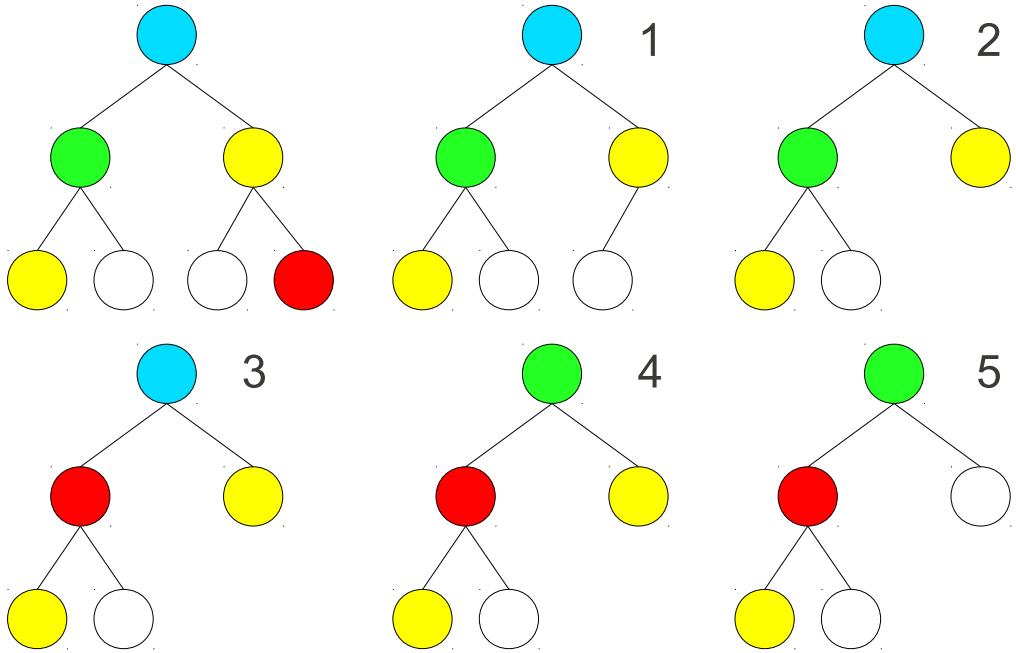
Metrics: Constrained Tree Edit Distance [3]

Operations:

- Change label
- Add node
- Remove node

Constraints:

- Proper edit distance
- Subtrees mapped to subtrees



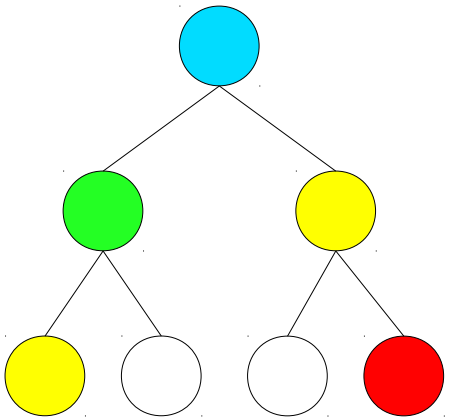
Algorithm complexity: $\mathcal{O}(b \cdot |T_1||T_2| \cdot \log_2(b))$

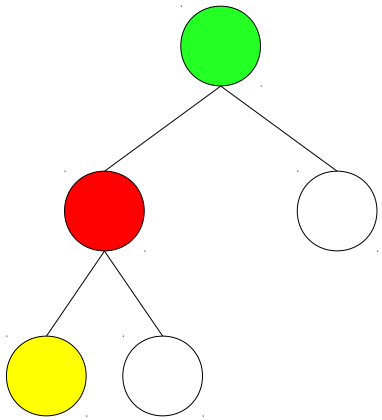
Metrics: Maximal Similarity Common Subtree [2]

Find maximal common subtree

Constraints:

- Ancestor-descendant relationship remains



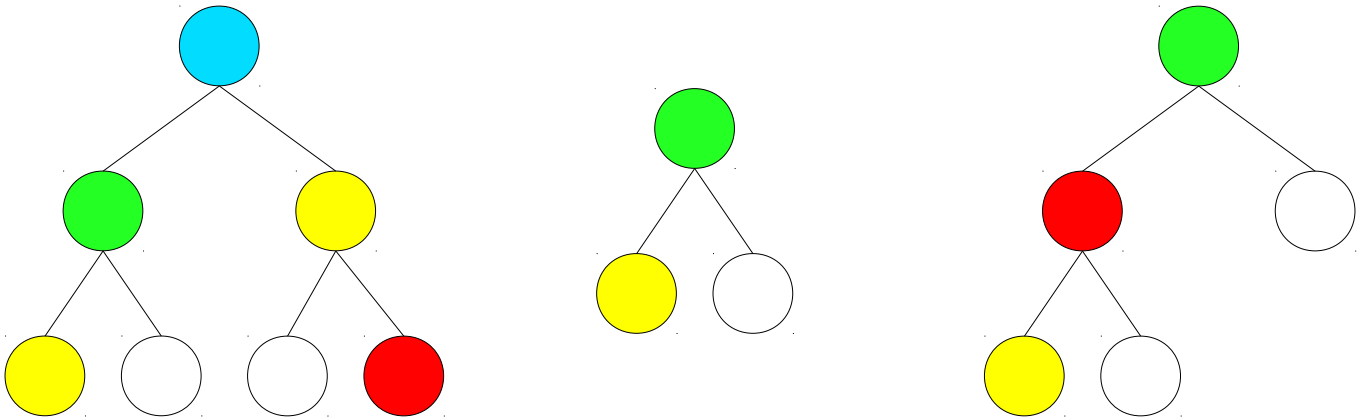


Metrics: Maximal Similarity Common Subtree [2]

Find maximal common subtree

Constraints:

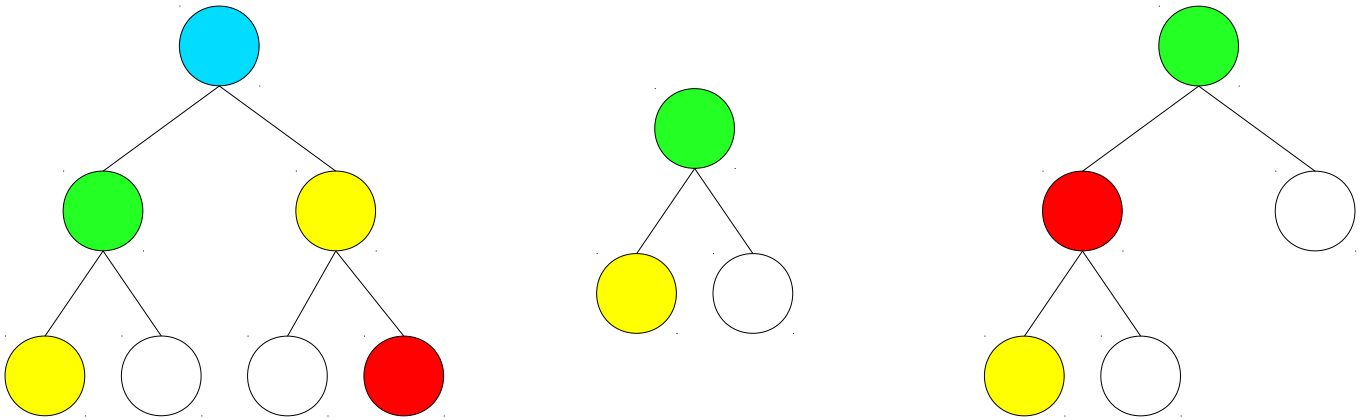
- Ancestor-descendant relationship remains



Metrics: Maximal Similarity Common Subtree [2]

Find maximal common subtree
Constraints:

- Ancestor-descendant relationship remains



In general: *isomorphic* subtrees
Algorithm complexity: $\mathcal{O}(b \cdot |T_1||T_2|)$

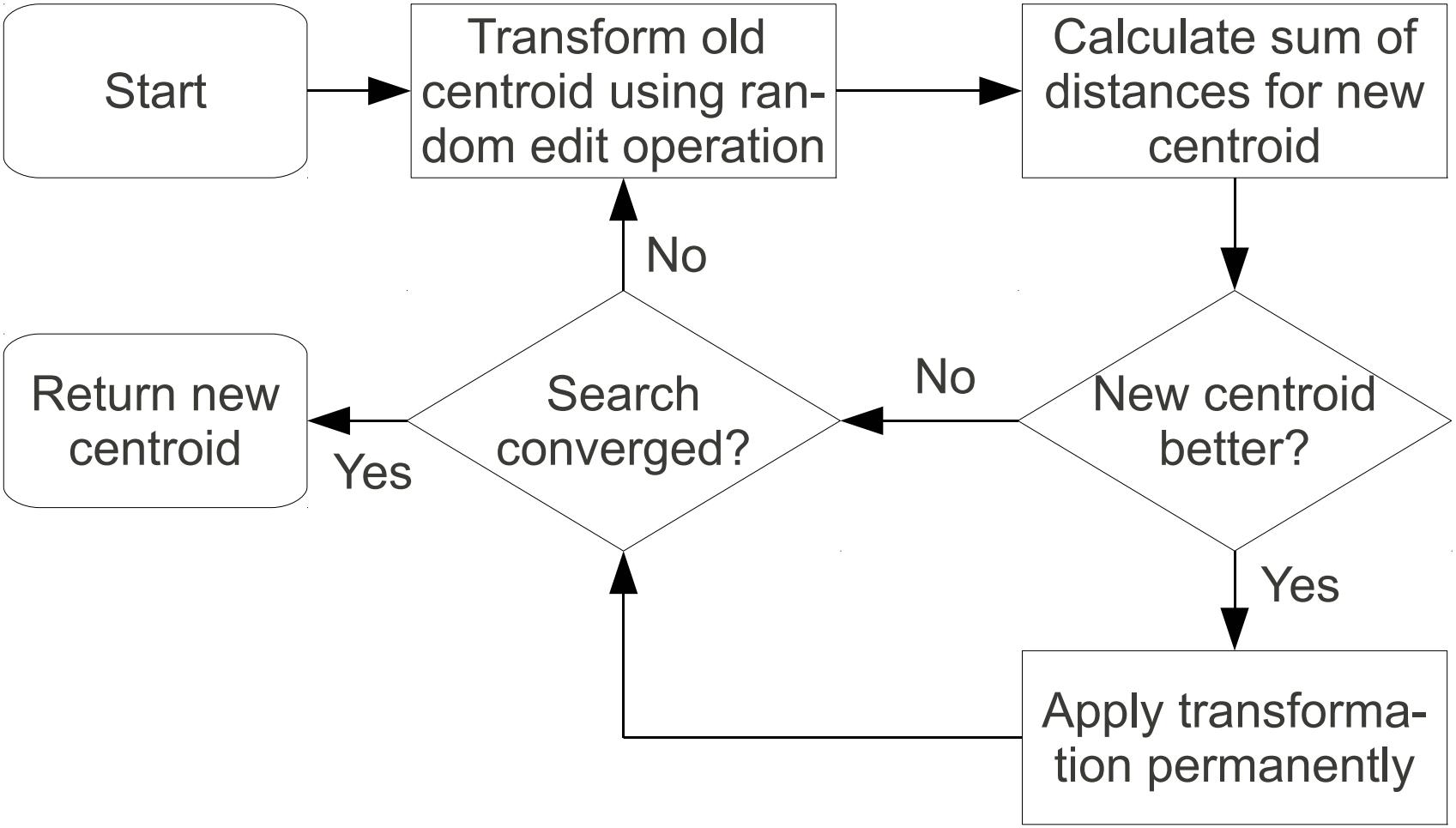
Mean Tree, Median Tree, and Medoid Tree

- For *mean tree* p=2, for *median tree* p=1:

$$\arg \min_{\bar{T} \in \mathcal{T}} \sum_{i=1}^{n} d(\bar{T}, T_i)^p$$

- For *medoid tree* additional constraint: $\bar{T} \in \{T_1, ..., T_2\}$

Centroid Search: Flow Chart



k-Means Clustering: Formal

$$\min E(\mathbf{W}, \mathbf{C}) := \sum_{i=1}^{k} \sum_{j=1}^{N} w_{ij} d(a_j, c_i)^p \tag{1}$$

subject to

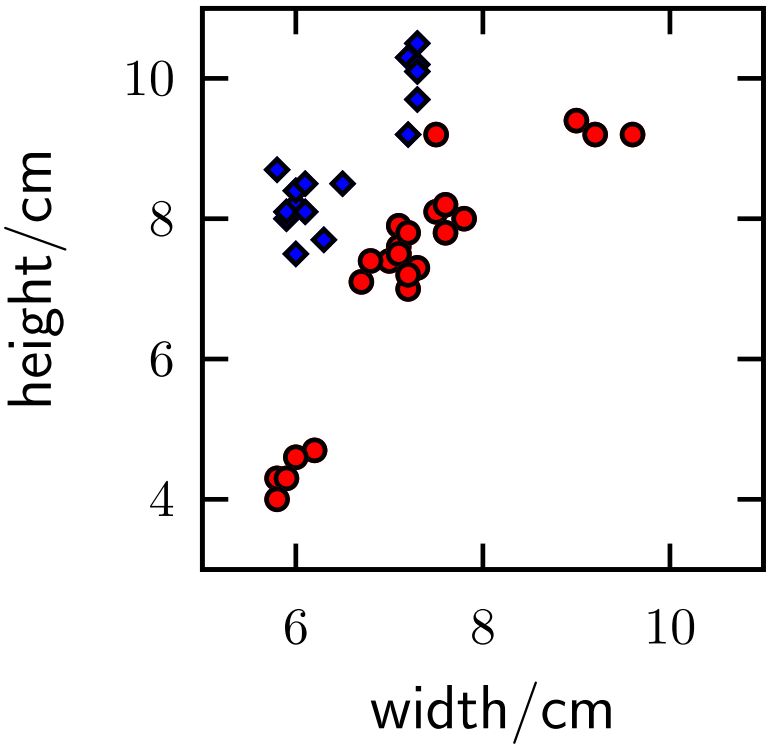
$$w_{ij} \in \{0, 1\}, \qquad \text{for} \quad 1 \leq i \leq k, 1 \leq j \leq N, \tag{2}$$

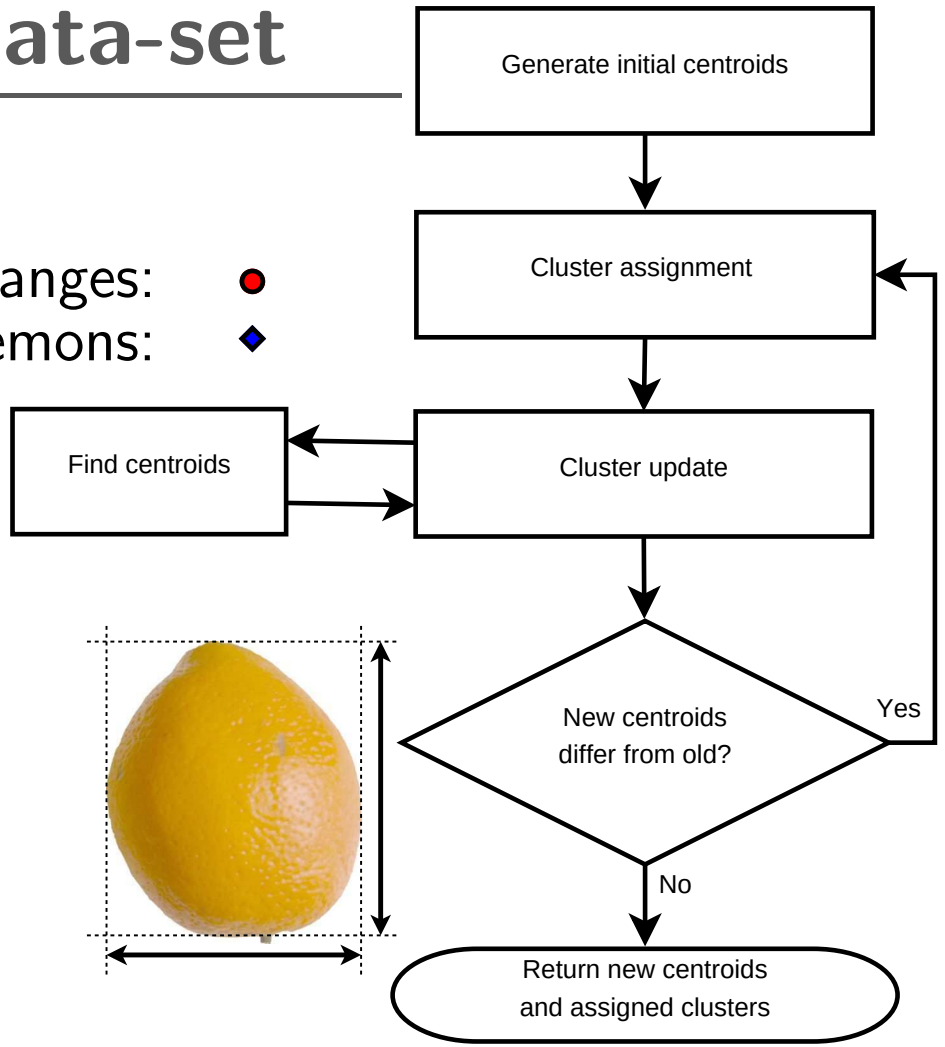
$$\sum_{i=1}^{k} w_{ij} = 1 \qquad \text{for} \quad 1 \leq i \leq k, 1 \leq j \leq N. \tag{3}$$

k-Means Clustering: Algorithm

A two-dimensional data-set

Source: [1]

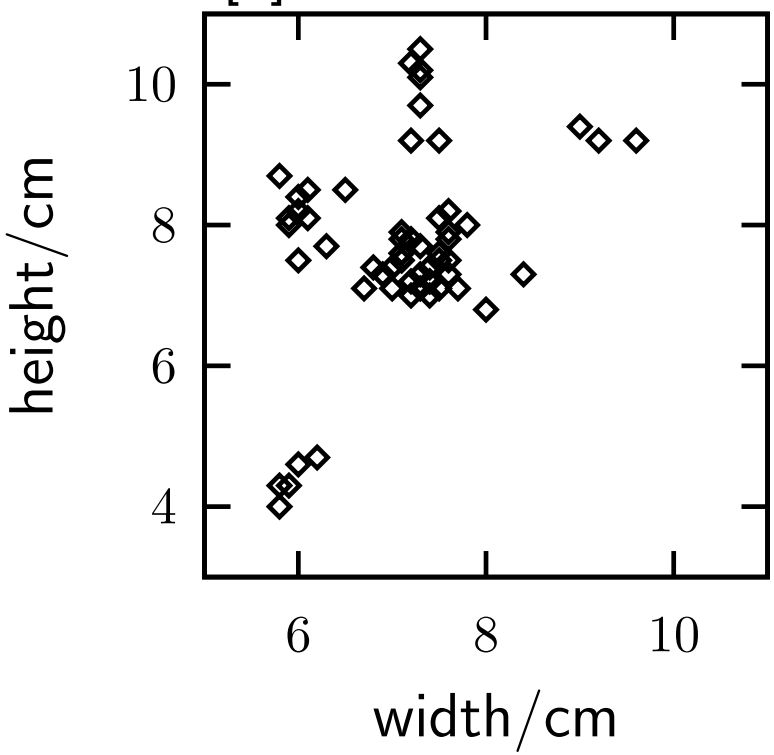


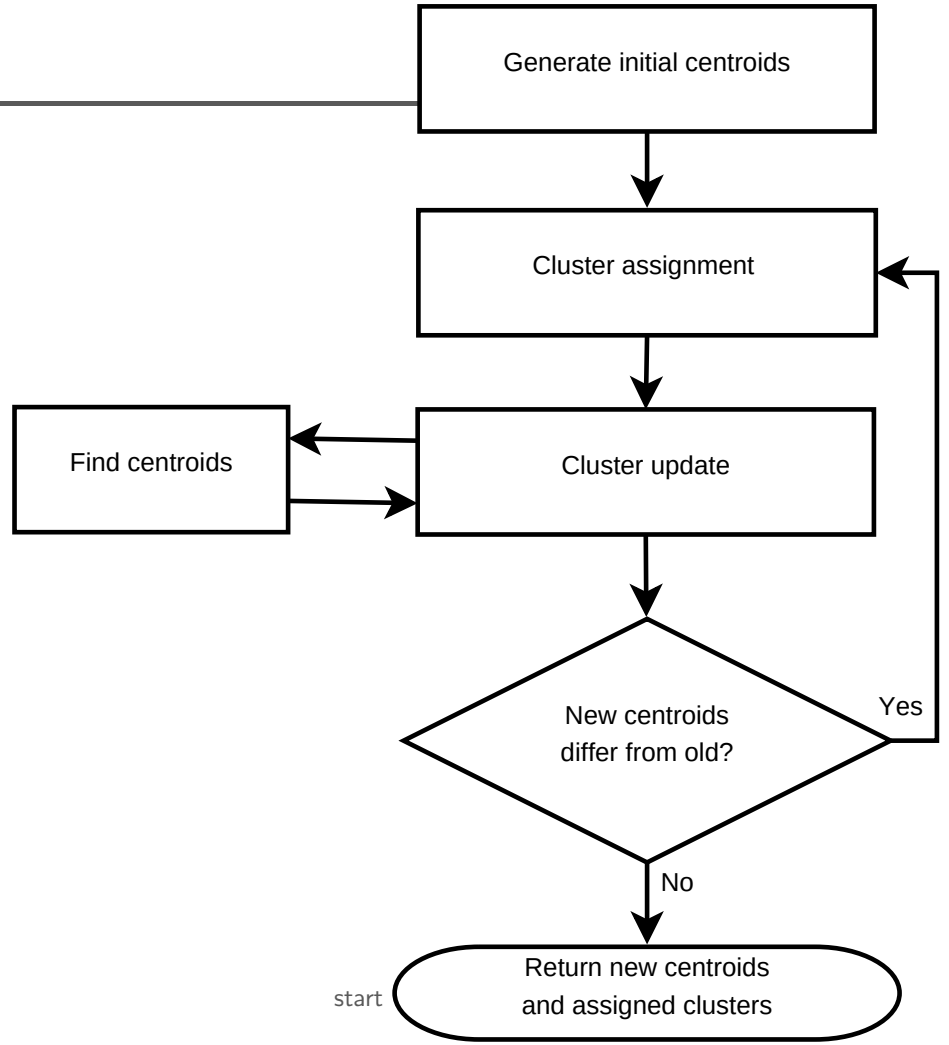


k-Means Clustering: Algorithm

Clustering

Source: [1]

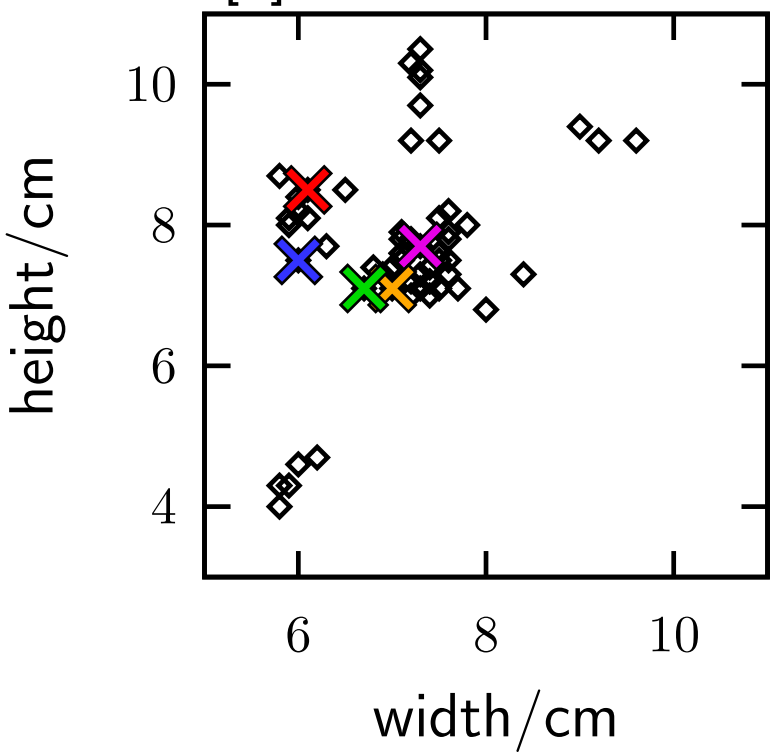


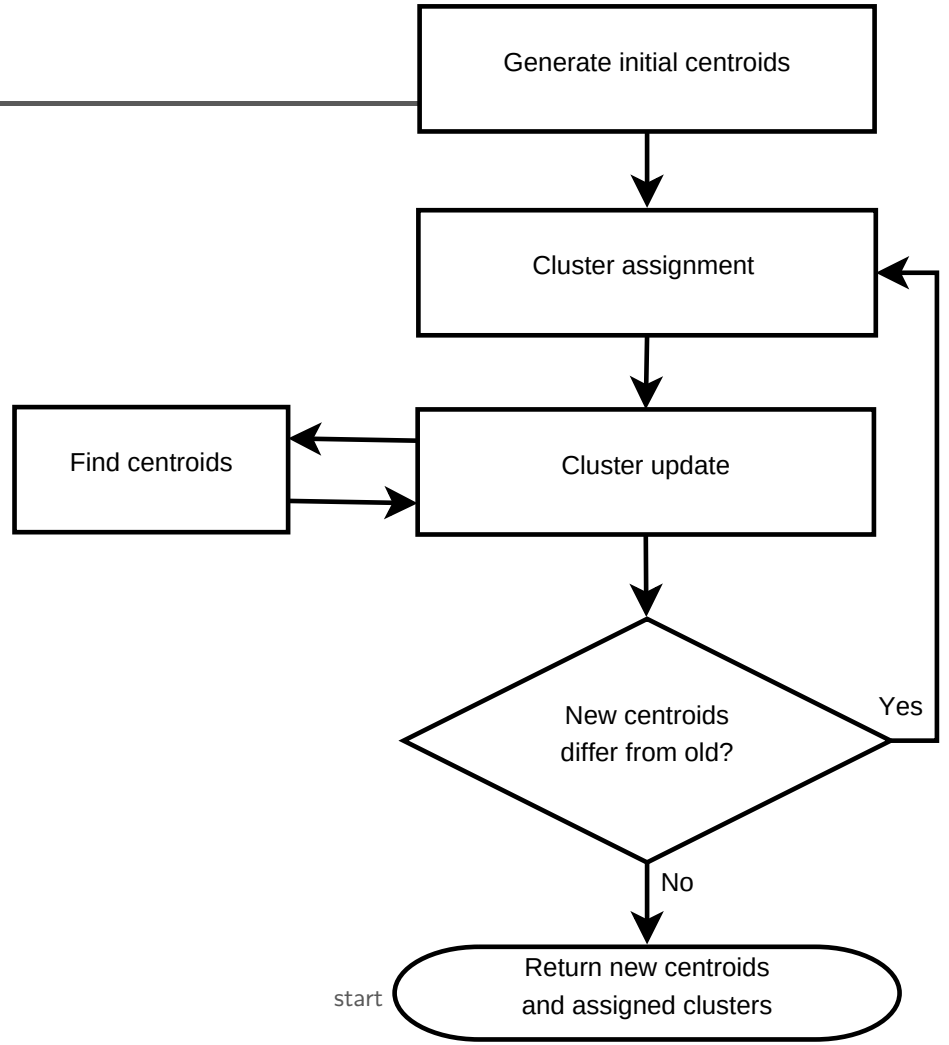


k-Means Clustering: Algorithm

Clustering

Source: [1]

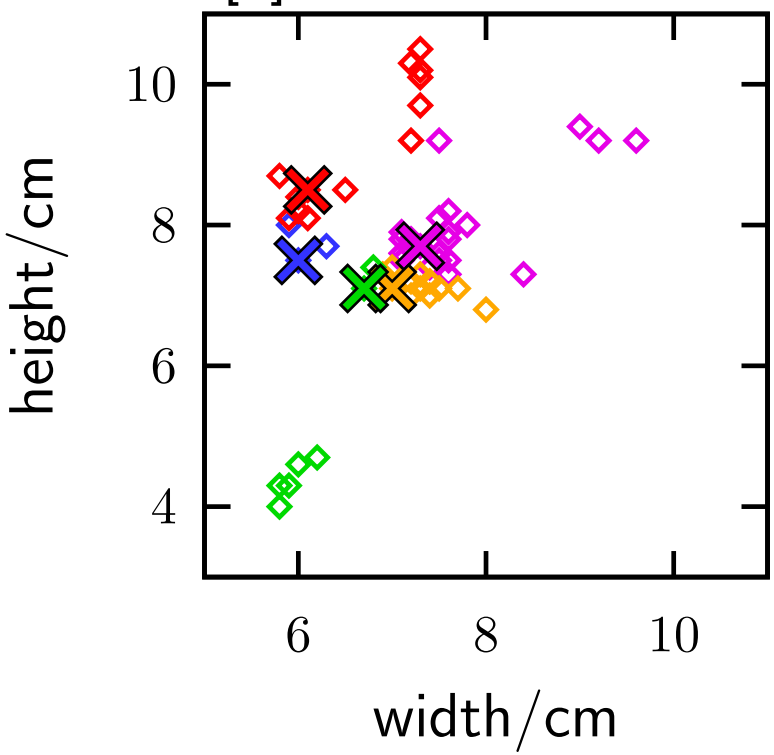


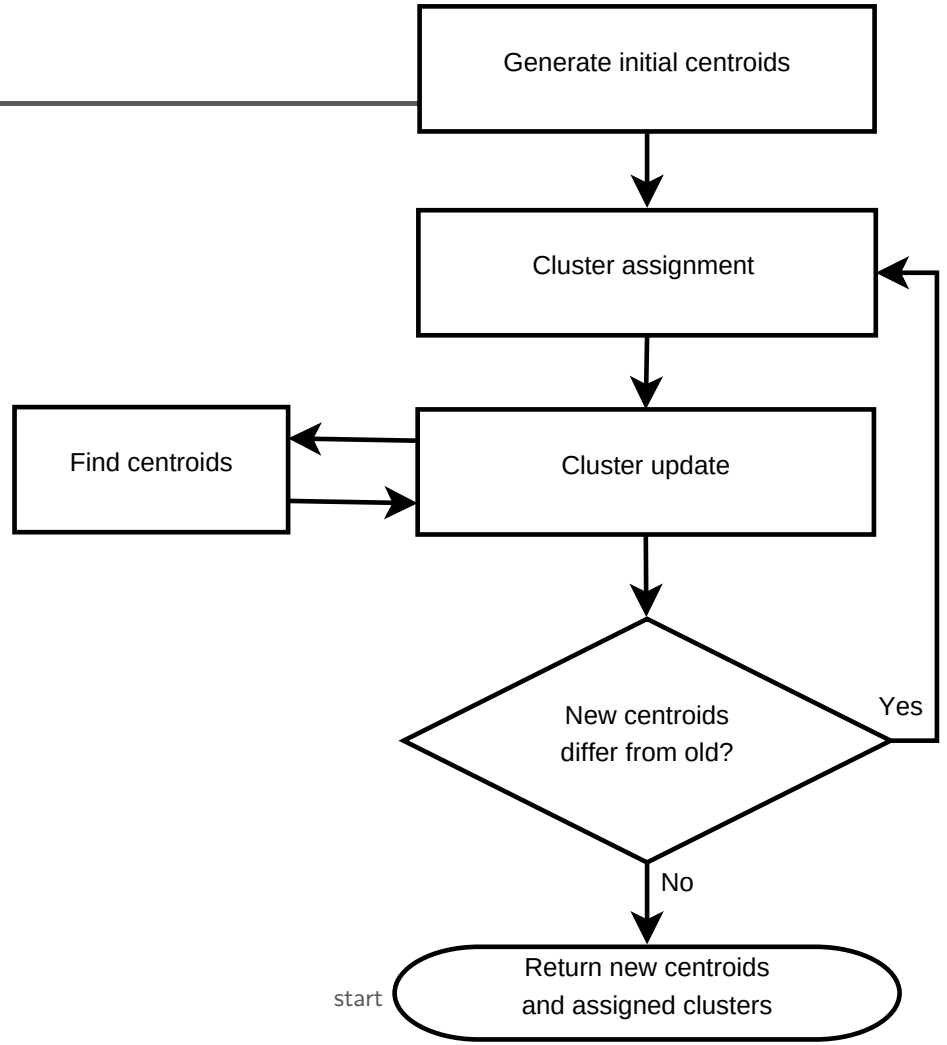


k-Means Clustering: Algorithm

Clustering

Source: [1]

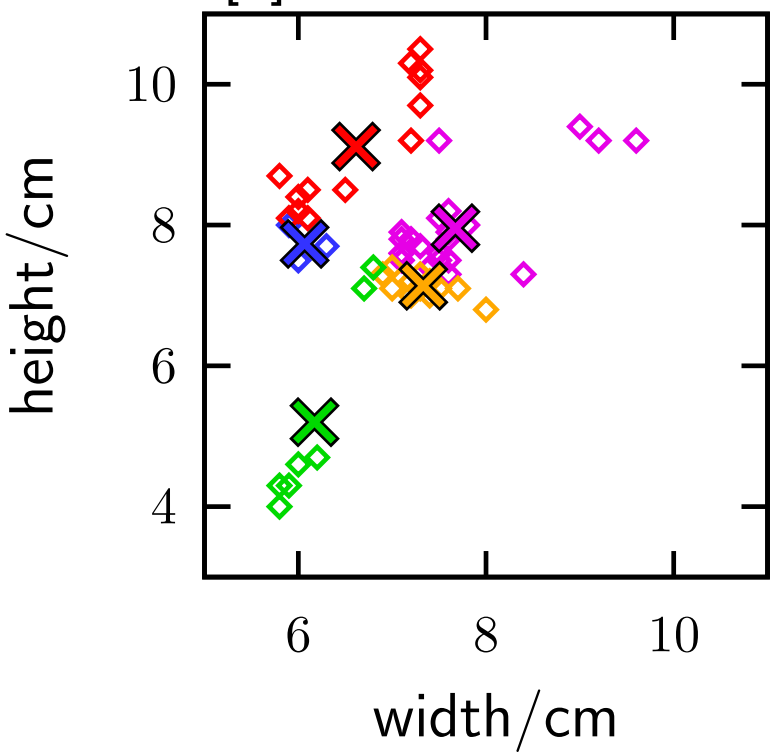


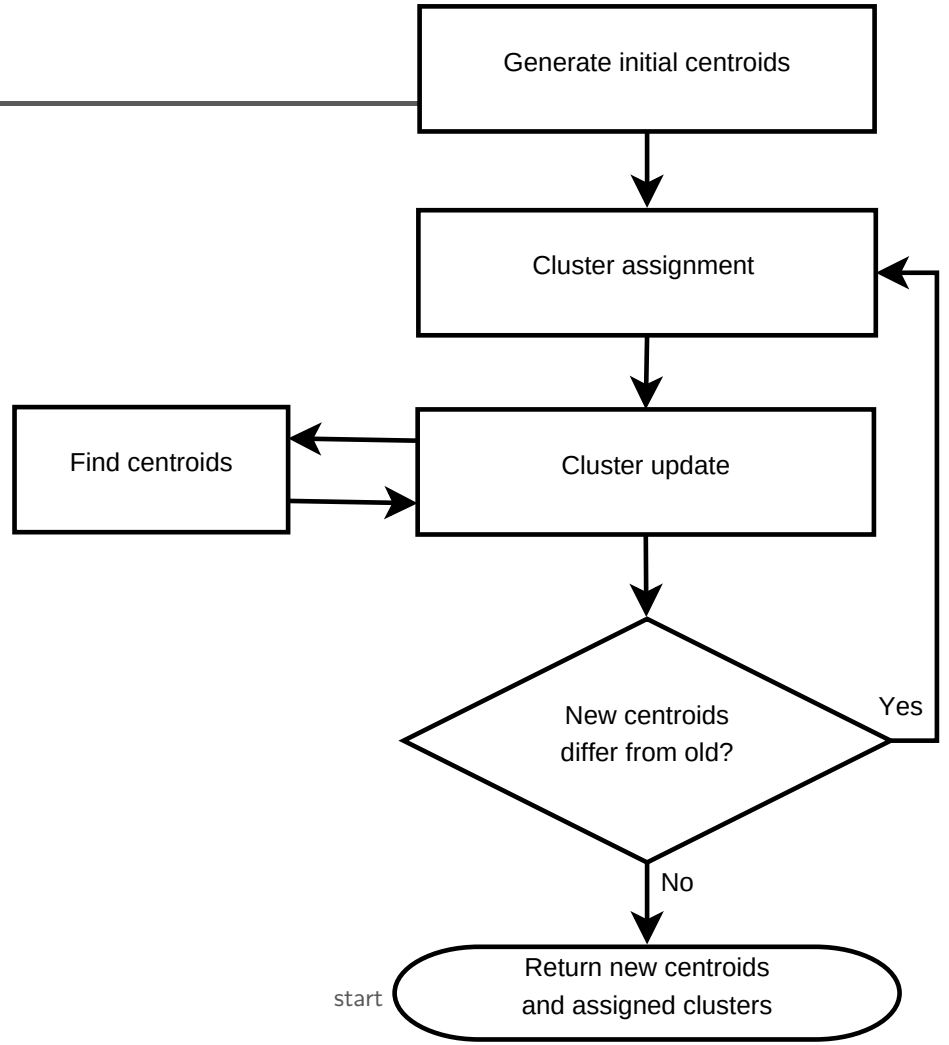


k-Means Clustering: Algorithm

Clustering

Source: [1]

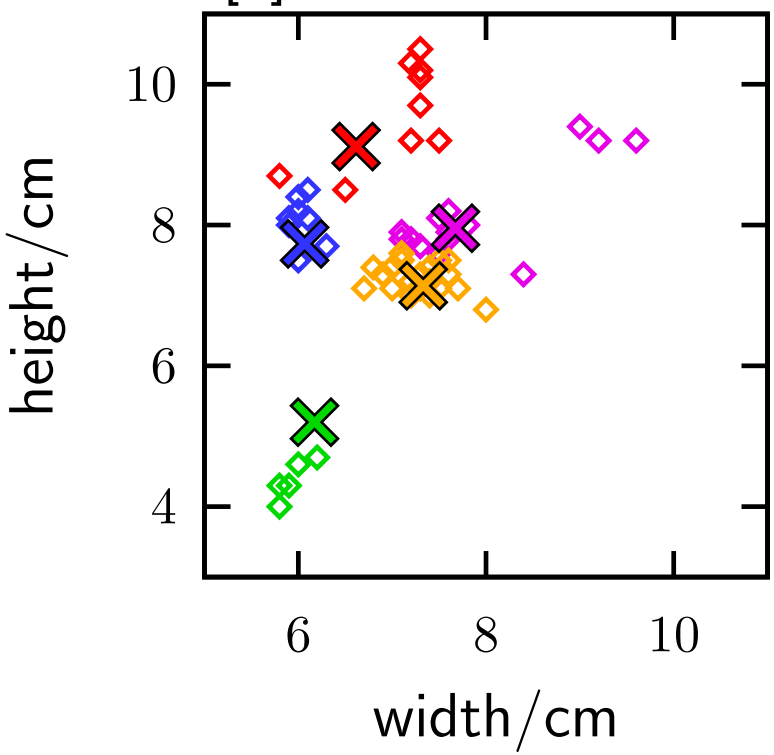


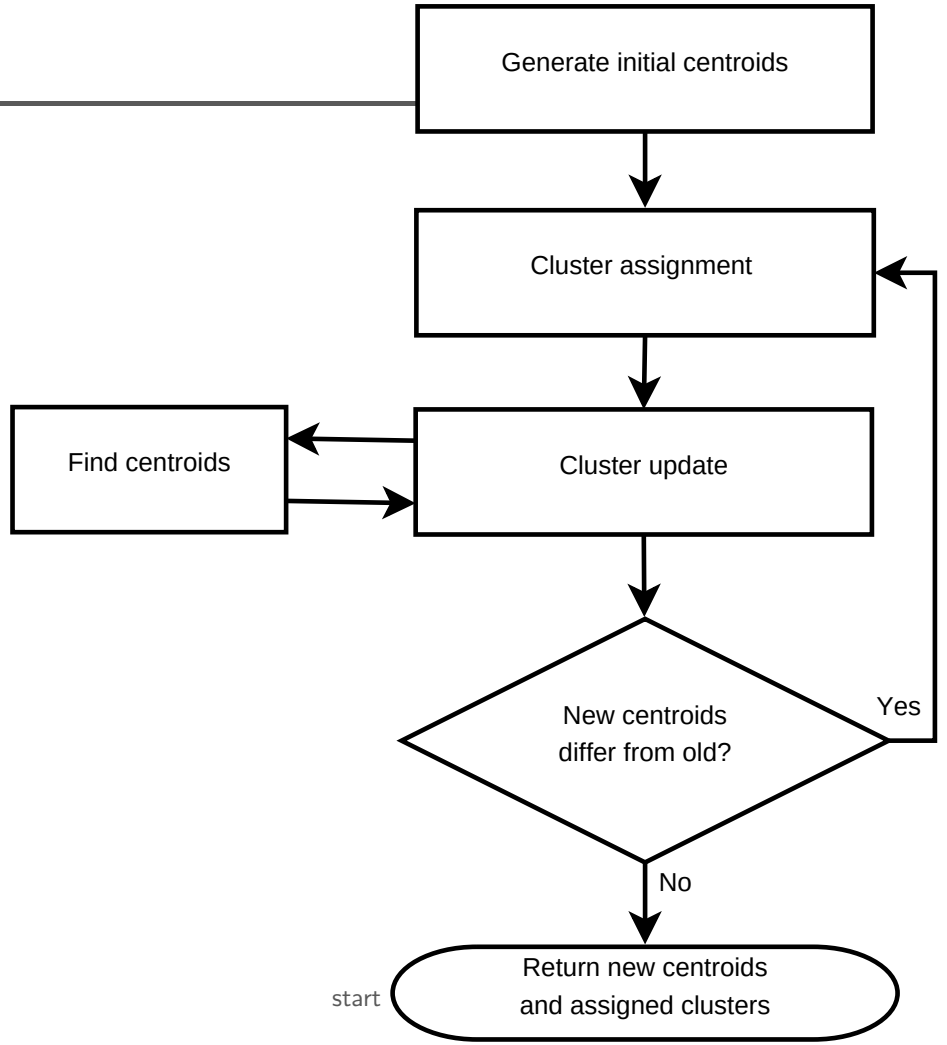


k-Means Clustering: Algorithm

Clustering

Source: [1]

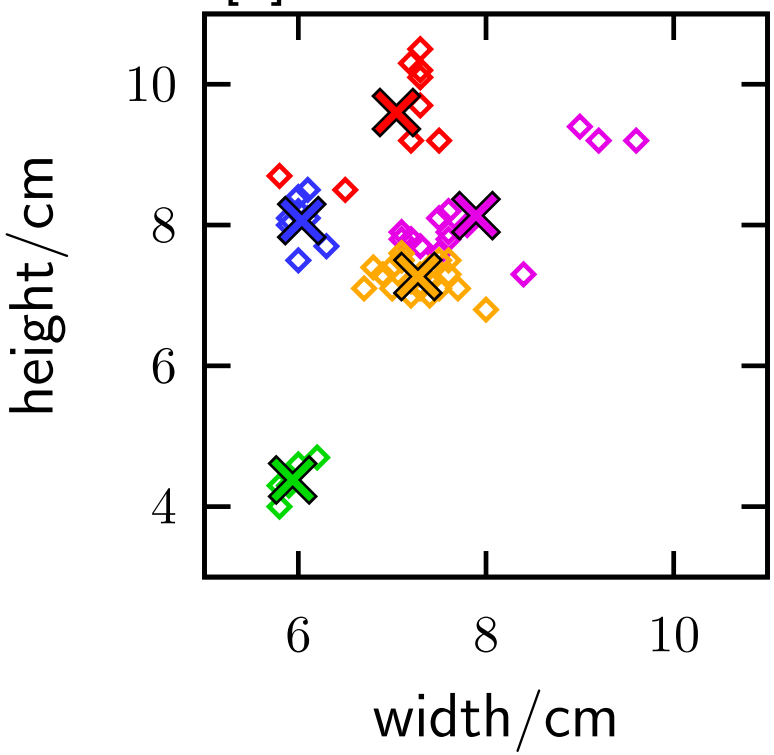


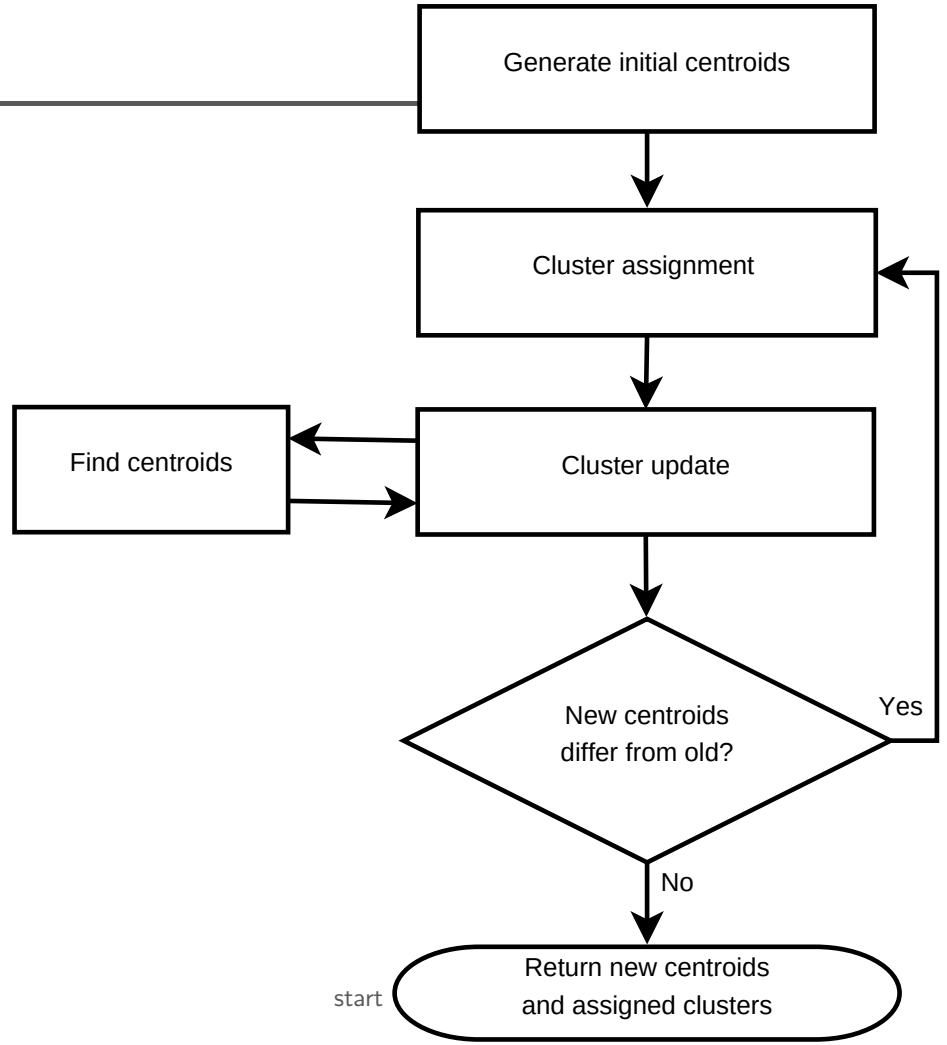


k-Means Clustering: Algorithm

Clustering

Source: [1]

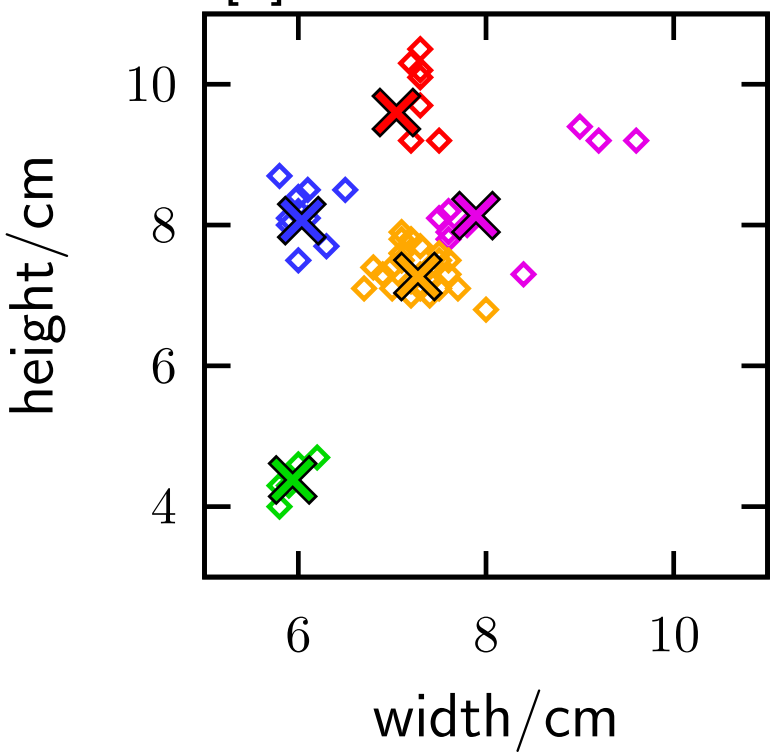


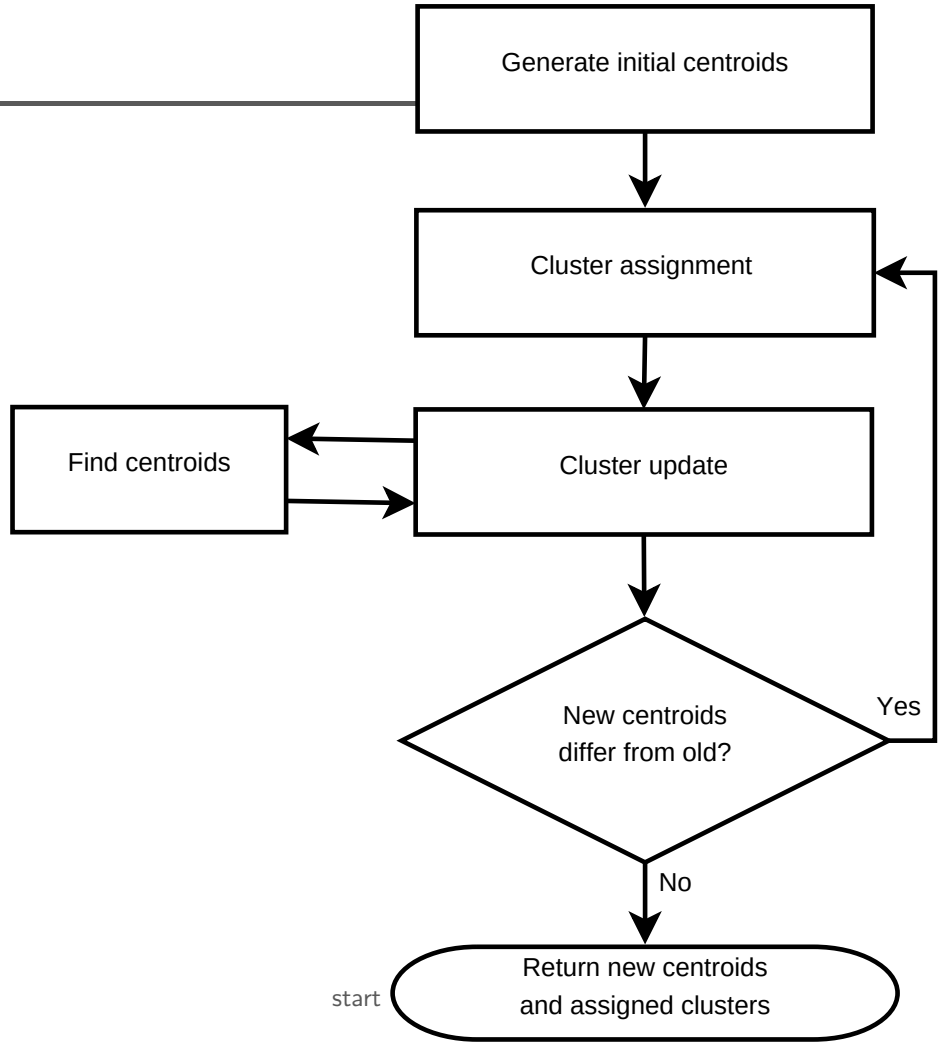


k-Means Clustering: Algorithm

Clustering

Source: [1]

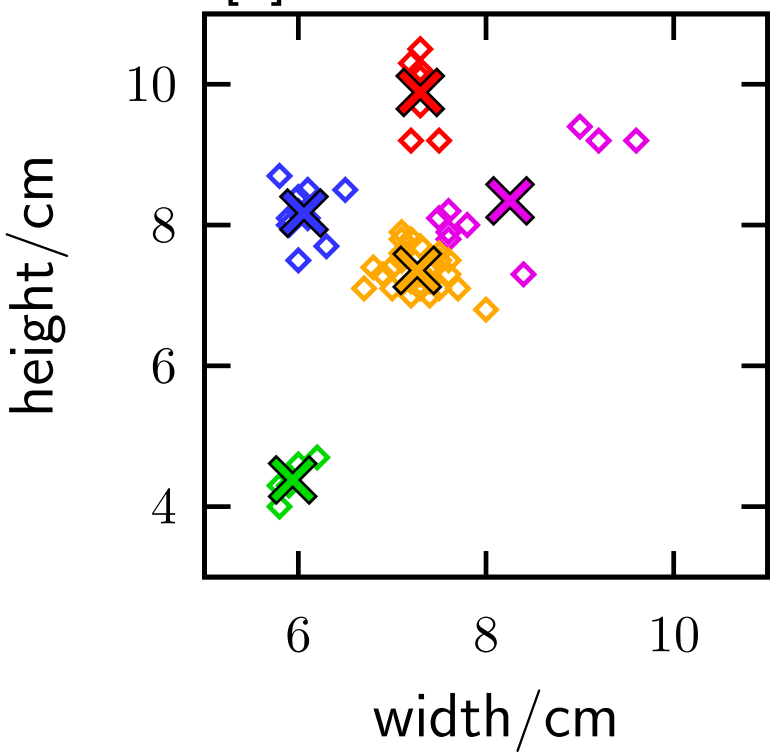


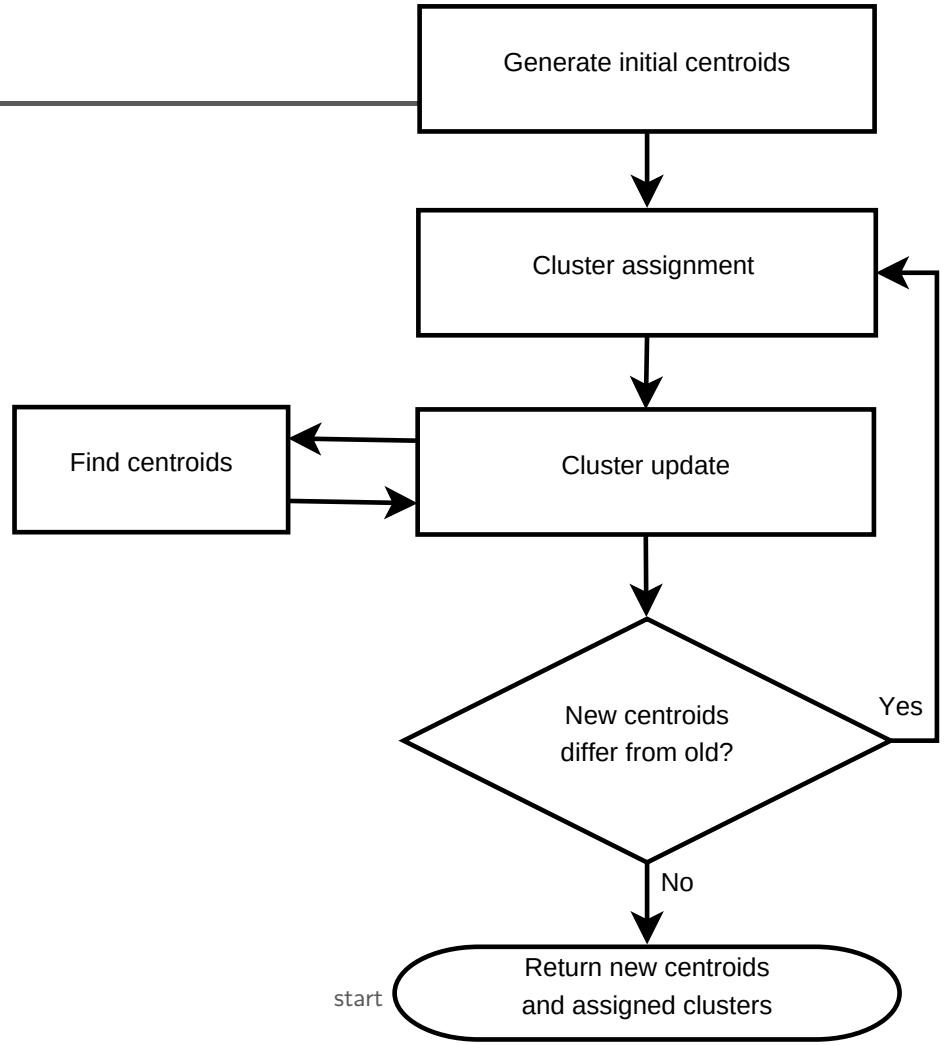


k-Means Clustering: Algorithm

Clustering

Source: [1]

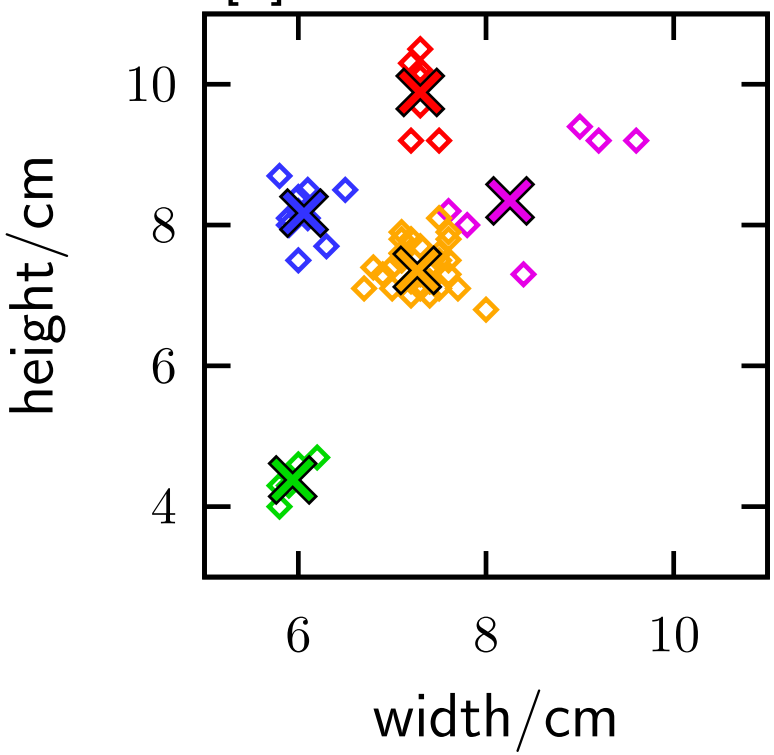


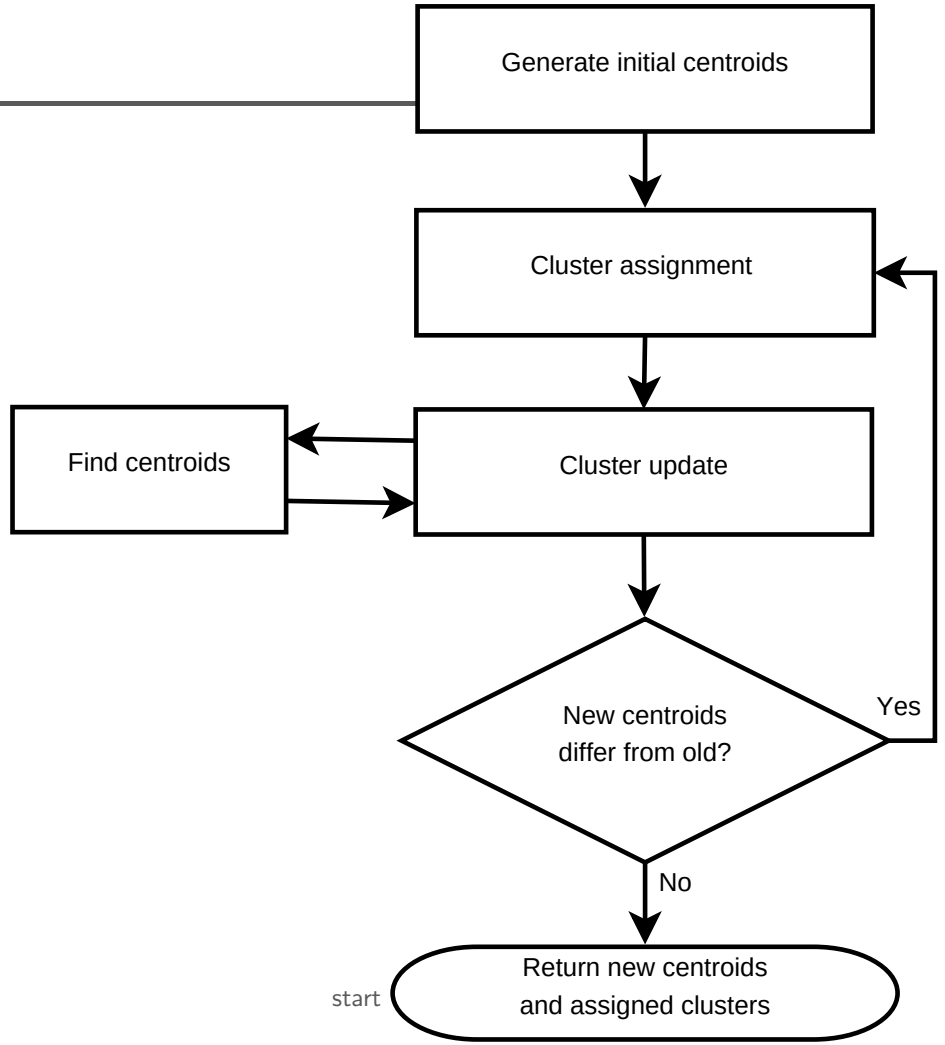


k-Means Clustering: Algorithm

Clustering

Source: [1]

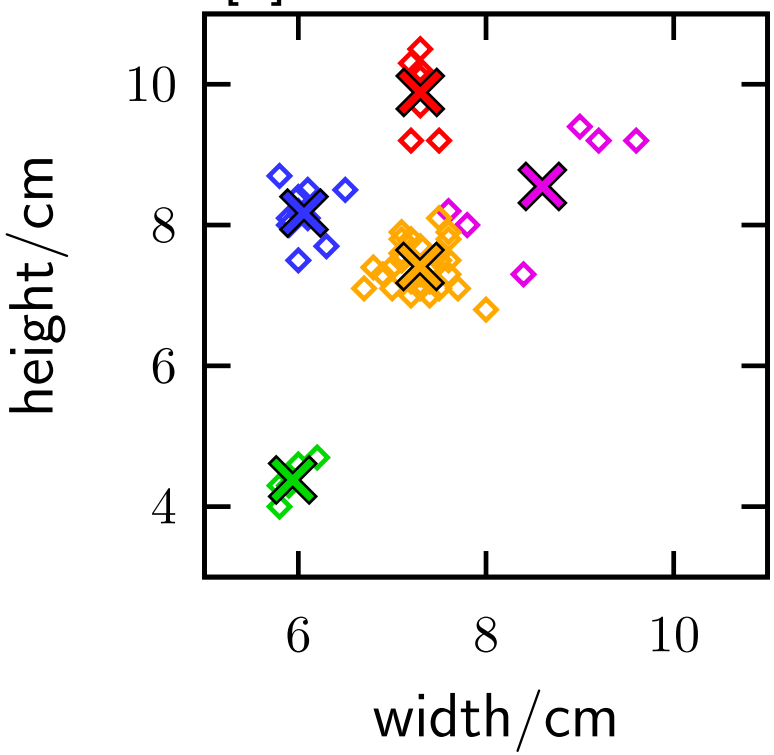


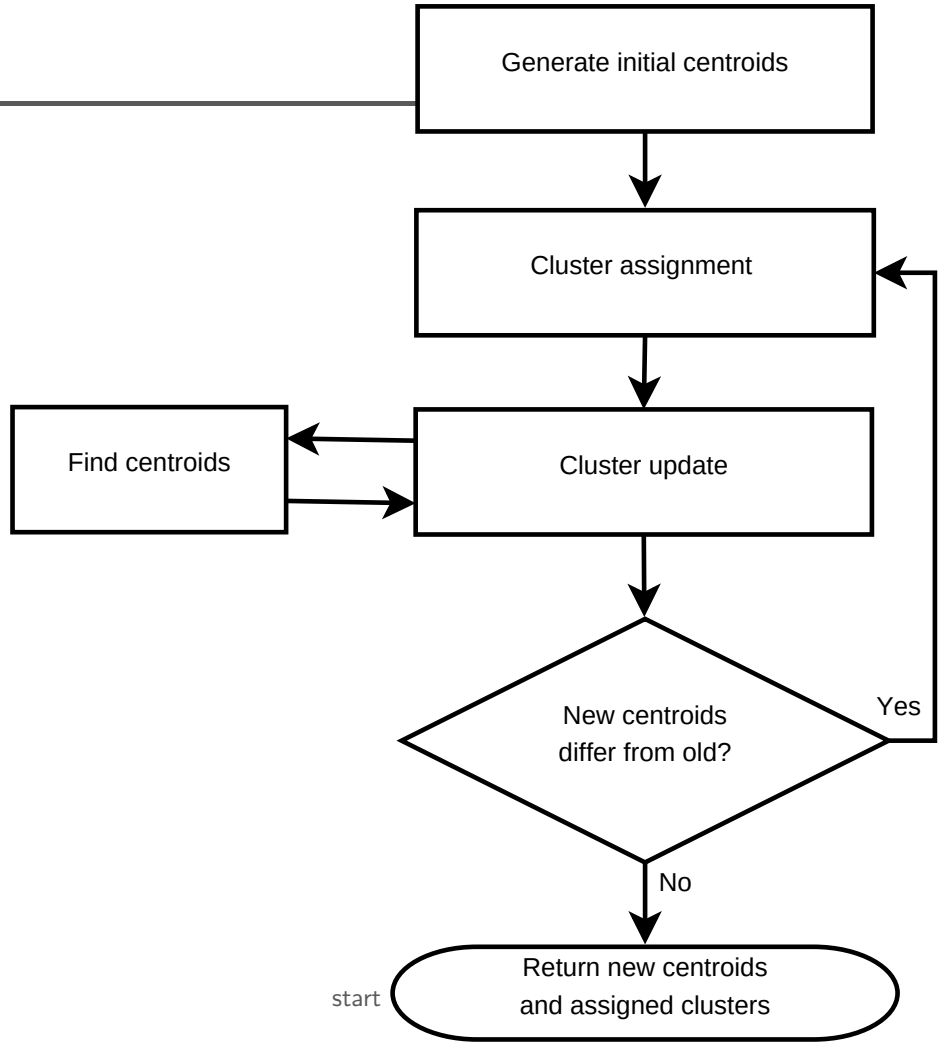


k-Means Clustering: Algorithm

Clustering

Source: [1]

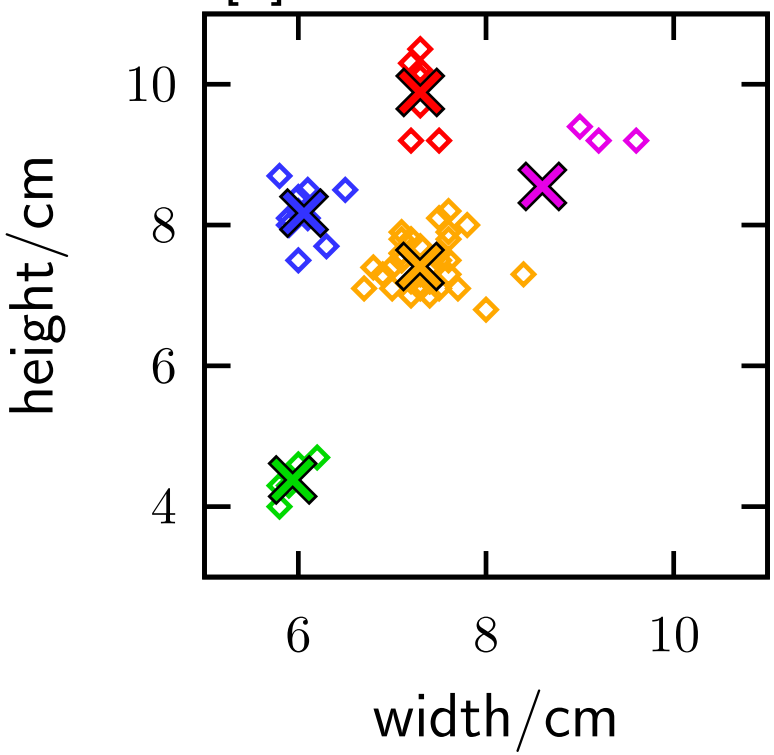


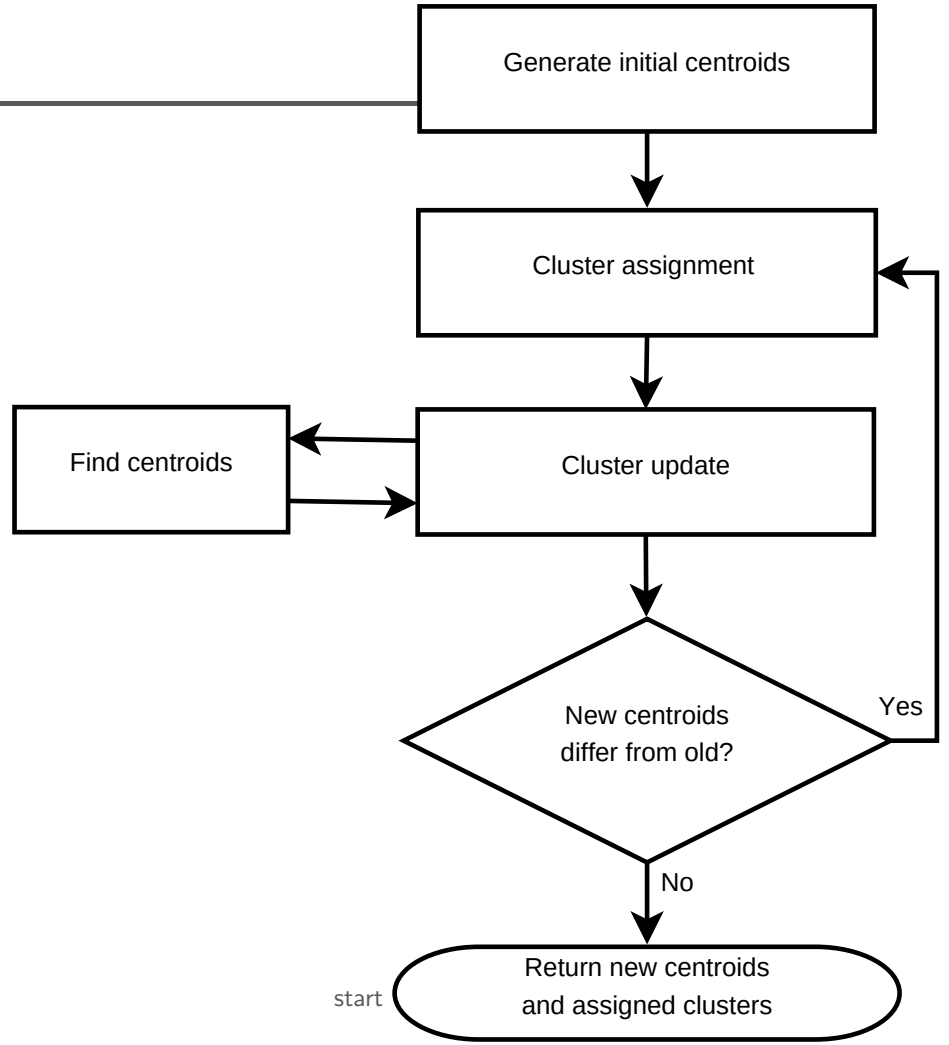


k-Means Clustering: Algorithm

Clustering

Source: [1]

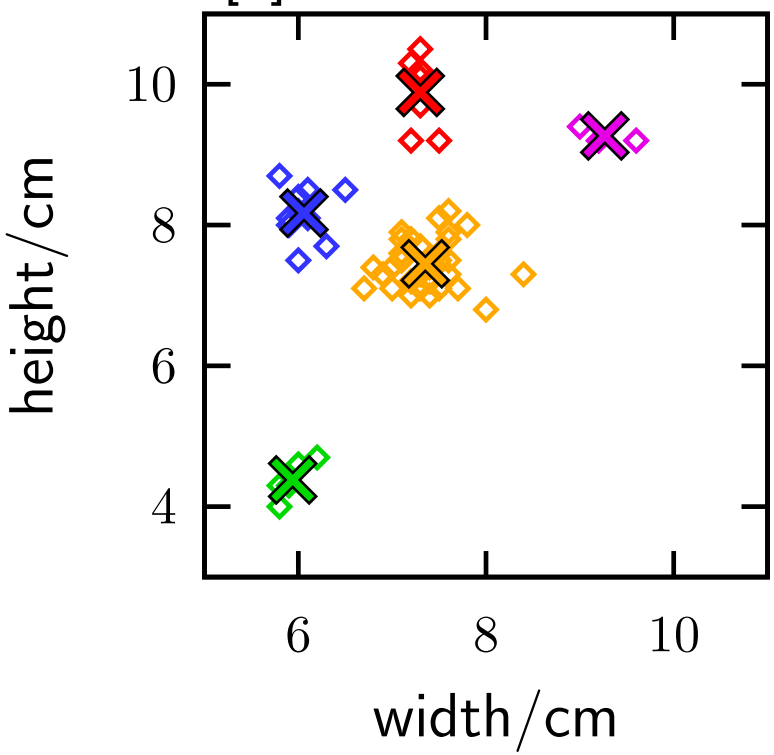


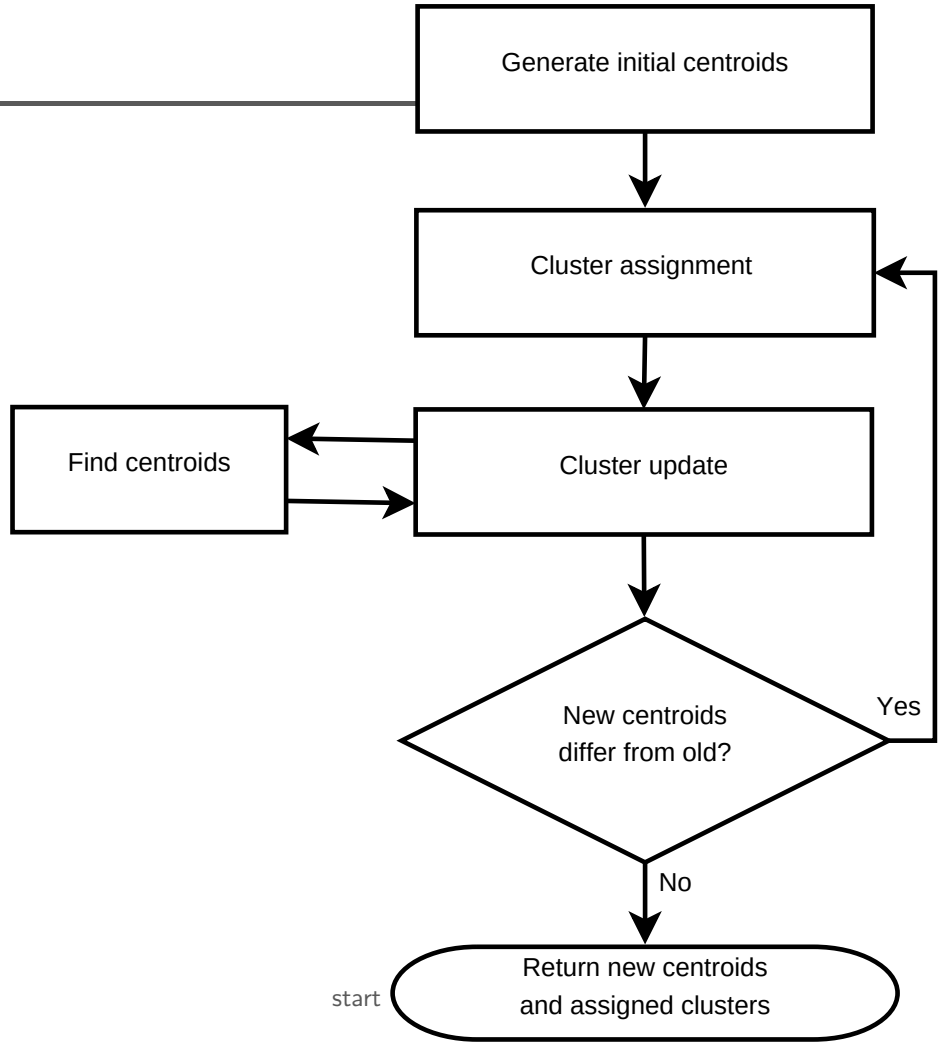


k-Means Clustering: Algorithm

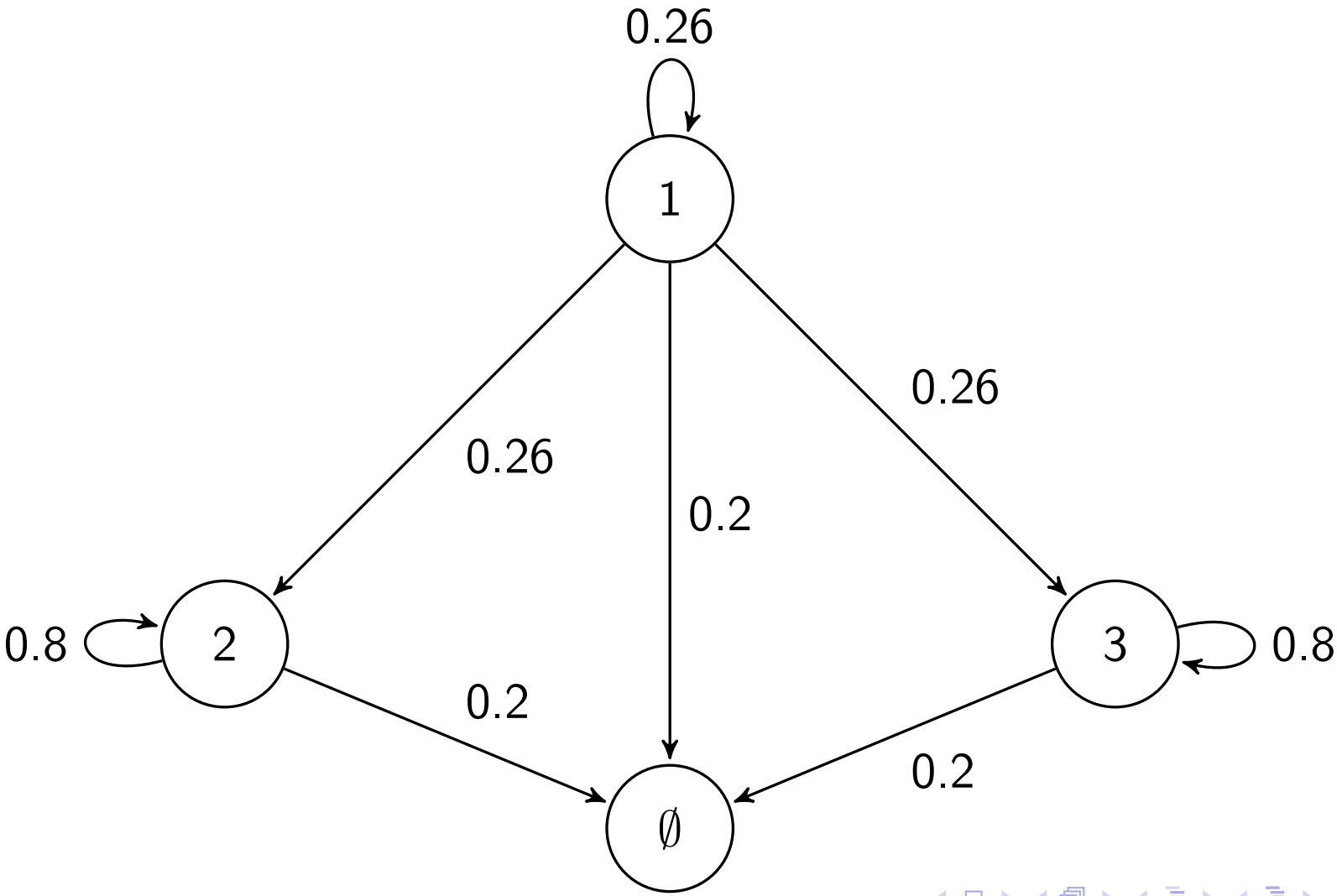
Clustering

Source: [1]

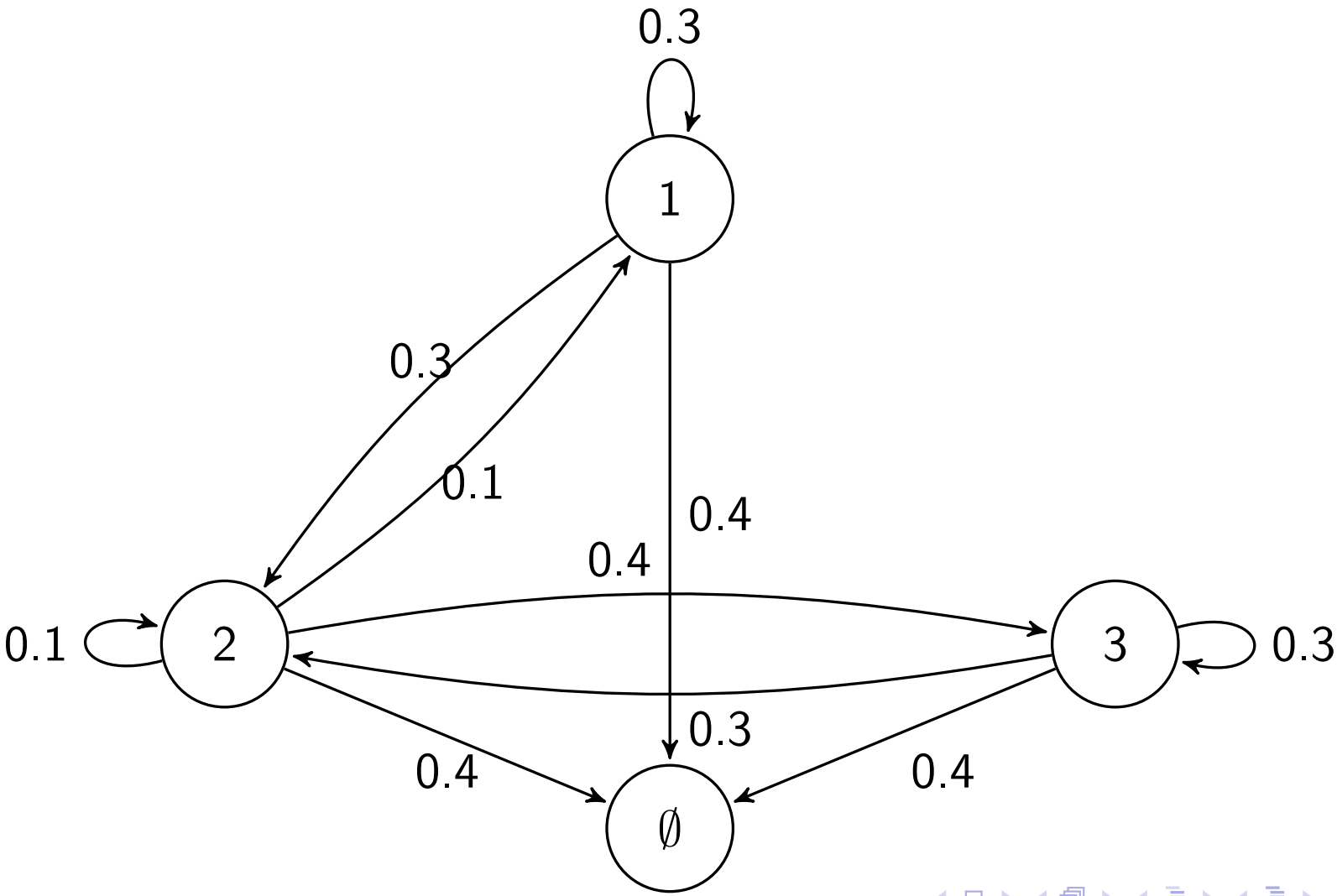




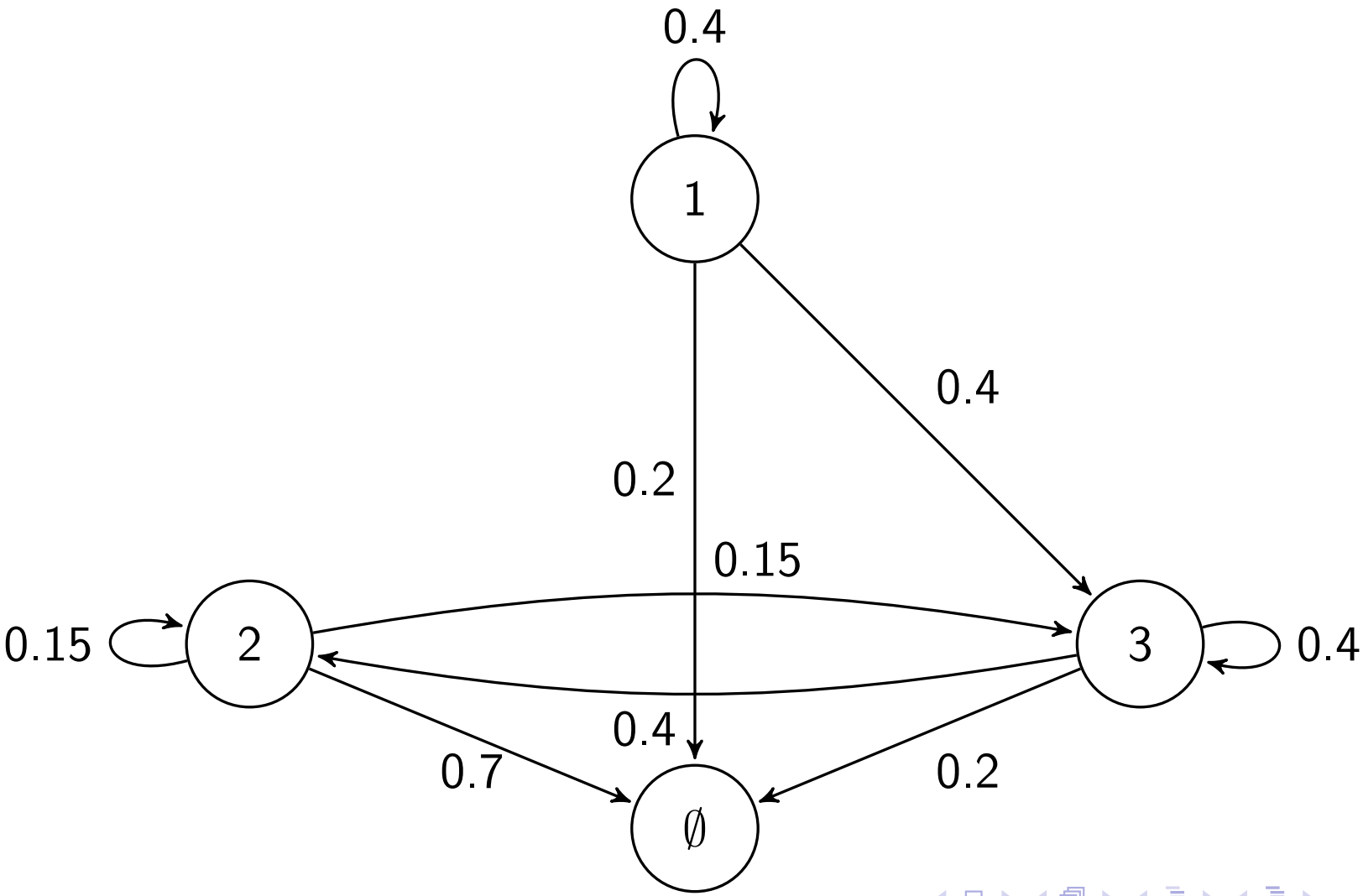
Artificial Data: Type 1



Artificial Data: Type 2



Artificial Data: Type 3

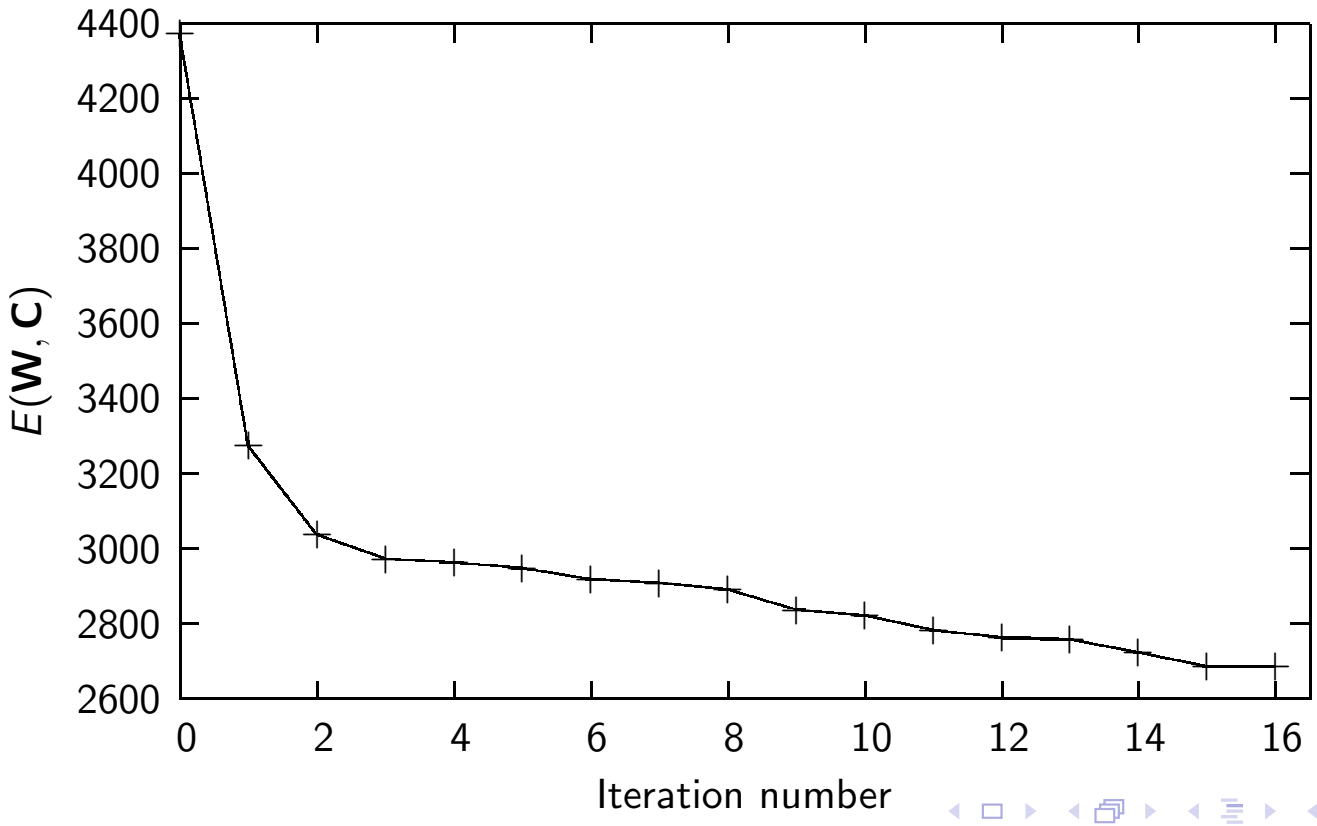


Artificial Data: Results Summary

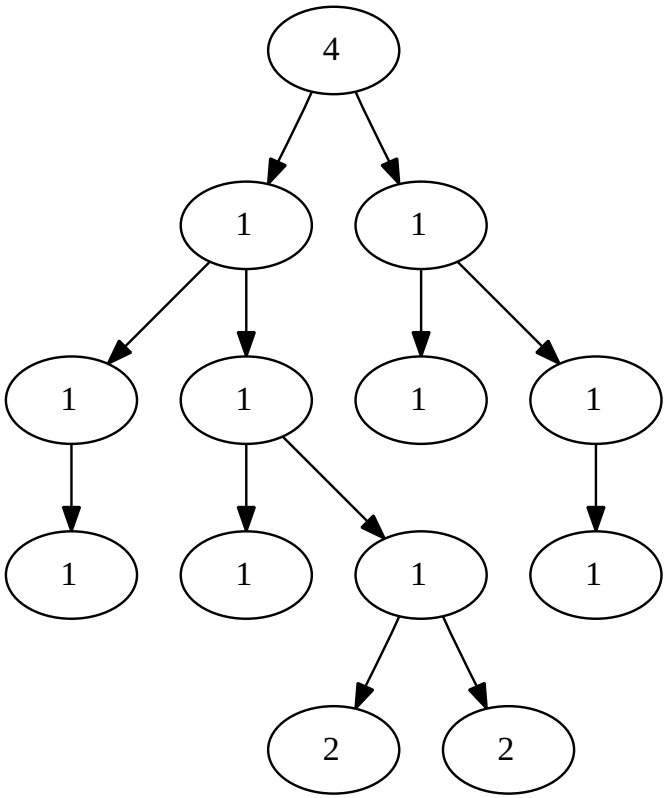
Metric	Centroid	k	Distances to type 1	Distances to type 2	Distances to type 3	Cluster sizes
original data			0.397	0.258	0.250	
MaxSim 1	mean	4	0.644	0.271	0.300	12,7,5,6
MaxSim 3	mean	4	0.299	0.302	0.599	12, 0, 10, 6
MaxSim 4	median	5	0.658	0.249	0.382	0, 3, 1, 13, 11
cTED	medoid	4	0.710	0.233	0.577	27,1,1,1
MaxSim 3	medoid	4	0.573	0.271	0.285	6,4,10,10
MaxSim 4	medoid	5	0.657	0.269	0.267	6,4,5,6,9

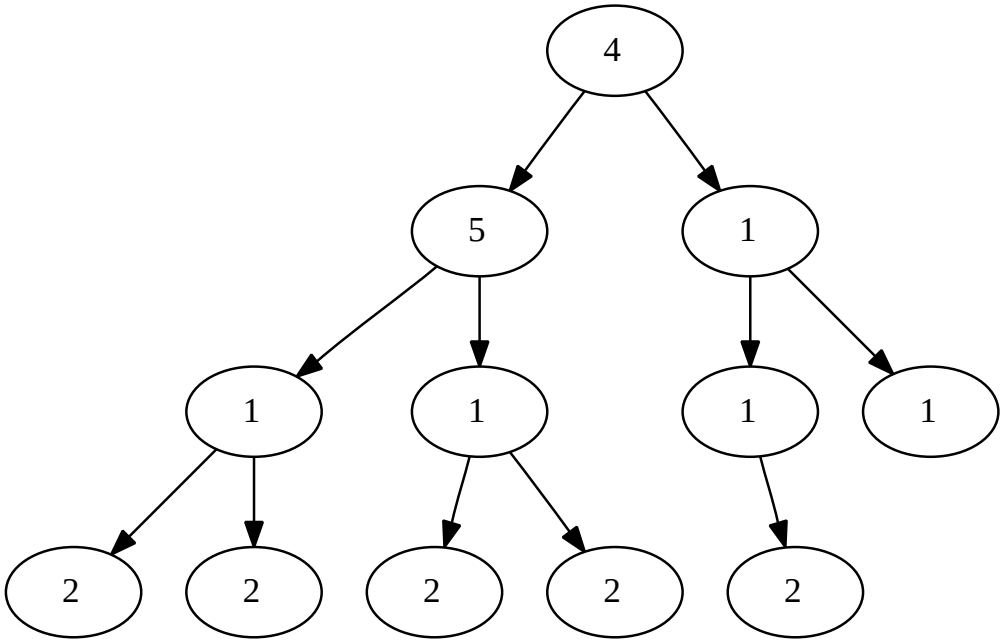
Real Data: Overview

Dataset: Cell differentiation from Granulocyte-Macrophage-Progenitor (GMP)
Labels: General type (1-None, 2-Myeloid, 4/5-HSC)



Real Data: Biological Meaning





Conclusions

- Clustering of tree data structures is possible using described metrics
- Maximal similarity metrics showed better results and higher efficiencythan cTED
- The centroids can be obtained using random search algorithm
- k-medoids shows good performance and higher efficiency than k-means and k-medians
- Experiment results for real data are biologically consistent

Further Research

- Results evaluation by experts in the domain
- Simulated annealing or variable neighbourhood search instead of random search
- Definition of subgradient on the tree space to avoid random search

Bibliography


I. Murray.
Oranges and lemons.
http://homepages.inf.ed.ac.uk/imurray2/.


A. Torsello, D. Hidović-Rowe, and M. Pelillo.
Polynomial-time metrics for attributed trees.
IEEE transactions on pattern analysis and machine intelligence, 27(7):1087–99, Jul 2005.


K. Zhang.
A constrained edit distance between unordered labeled trees.
Algorithmica, 15(3):205–222, 1996.