Special Lecture 7

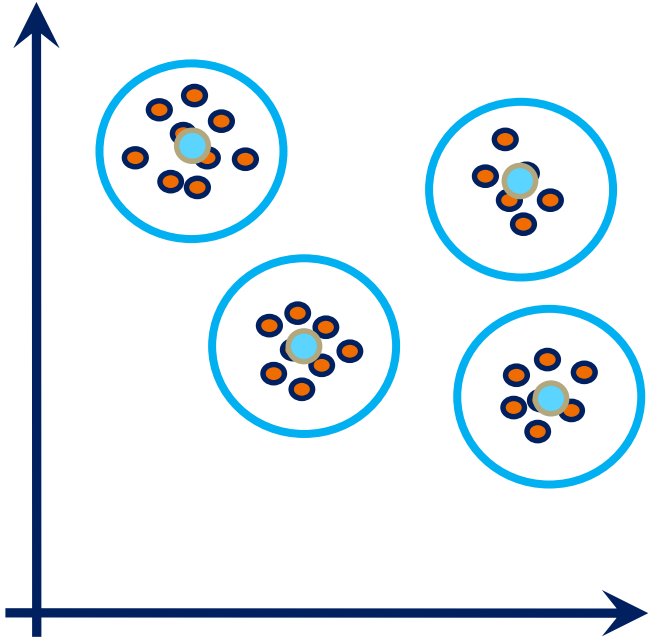
Unsupervised Learning

Clustering

Identifying Similar Patterns

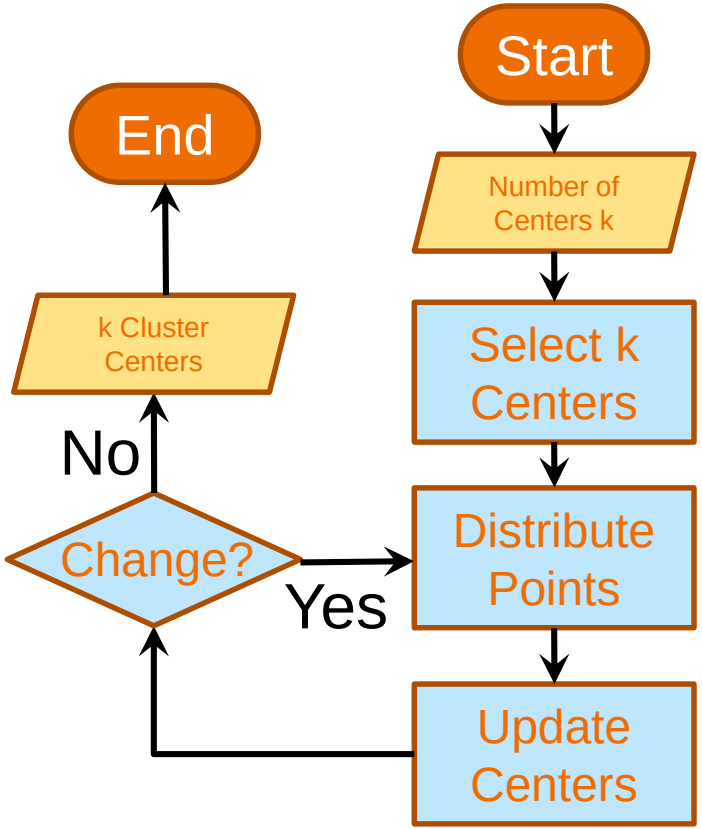
K-Means

- You are given N points
- How do we find k clusters?
  - What if we know the cluster centers?
- How do we find the cluster
- centers?
  - What if we know the k clusters?

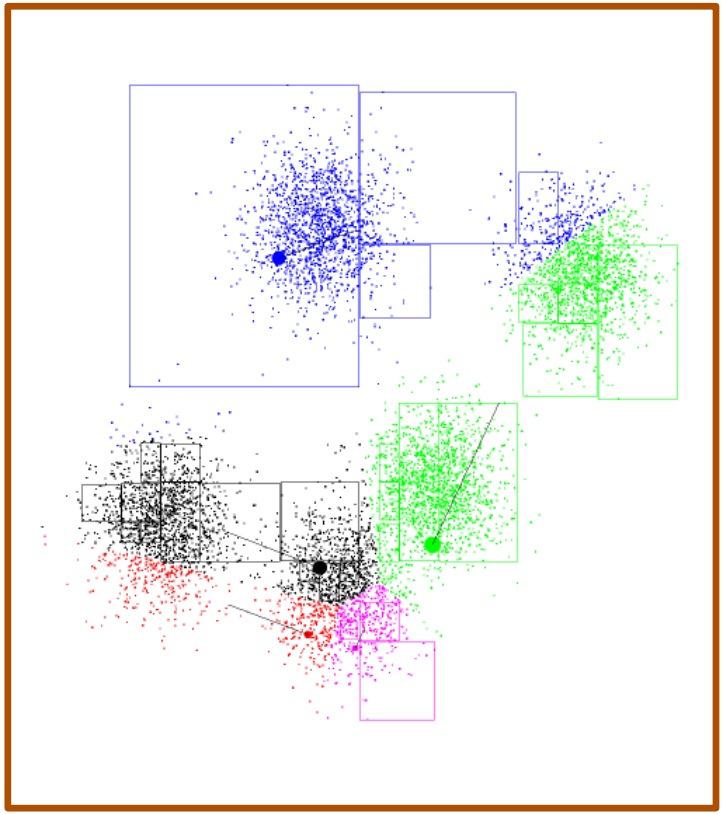


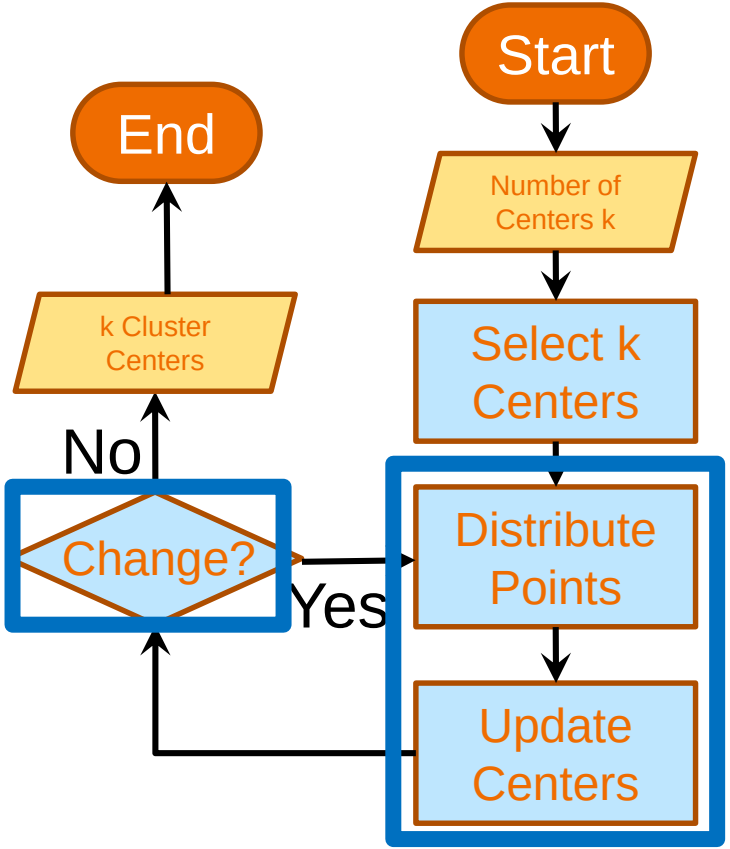
K-Means

1. Input: k (number of clusters)
2. Randomly select k centers
    3. Distribute Points
    4. Update Centers
    5. Repeat 3,4 till convergence
6. Output: Cluster centers

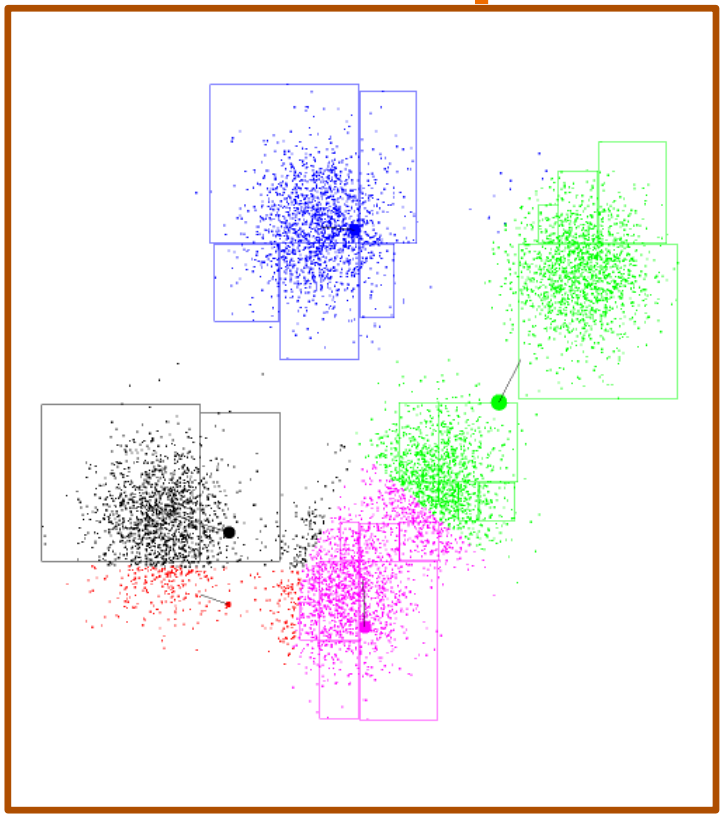


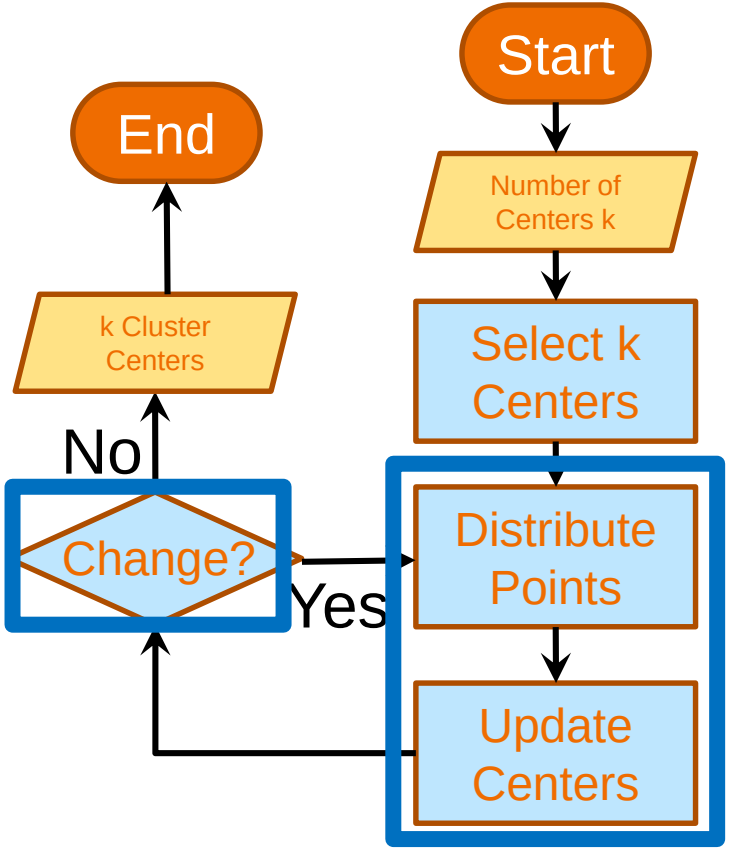
K-Means



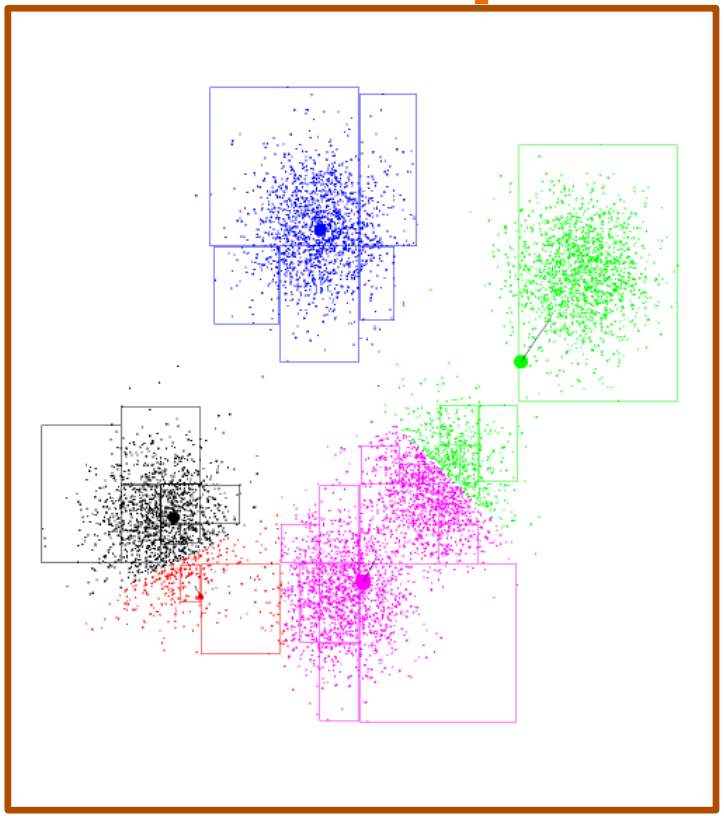

Start
Number of Centers k
Select k Centers
Distribute Points
Update Centers
Change?
Yes
No
k Cluster Centers
End

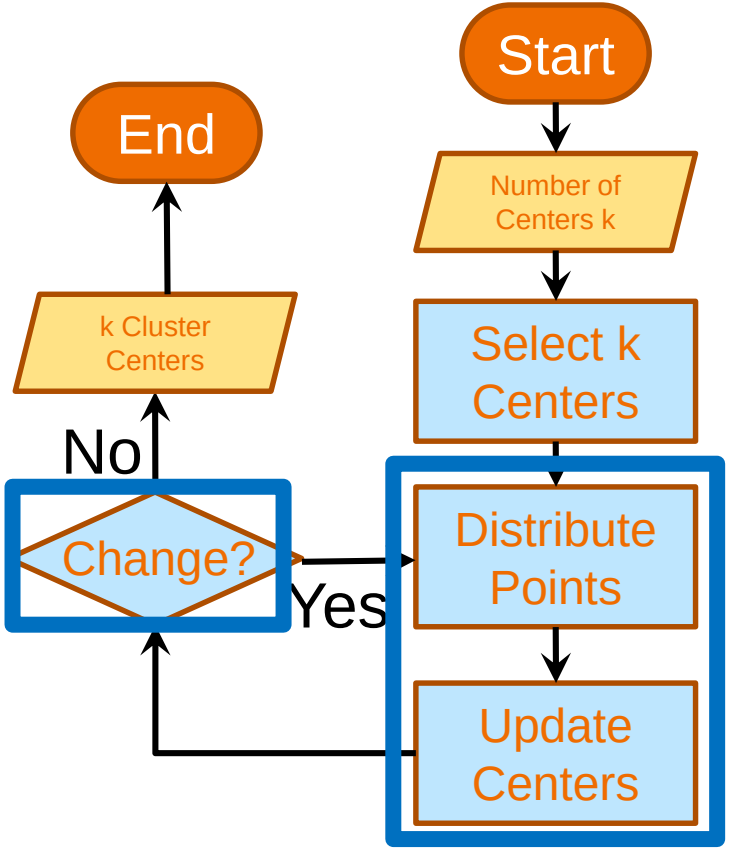
K-Means: Update 1



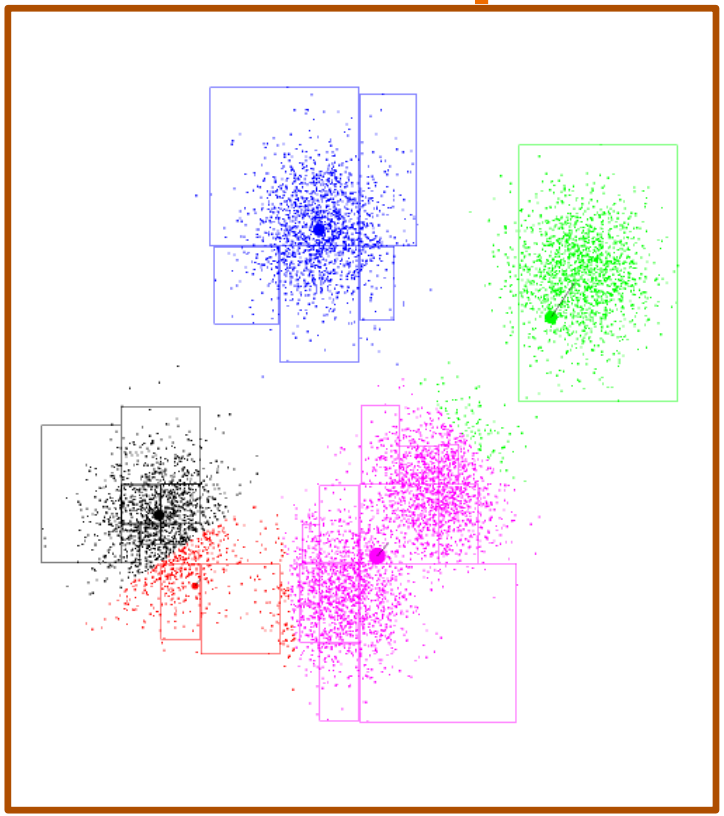


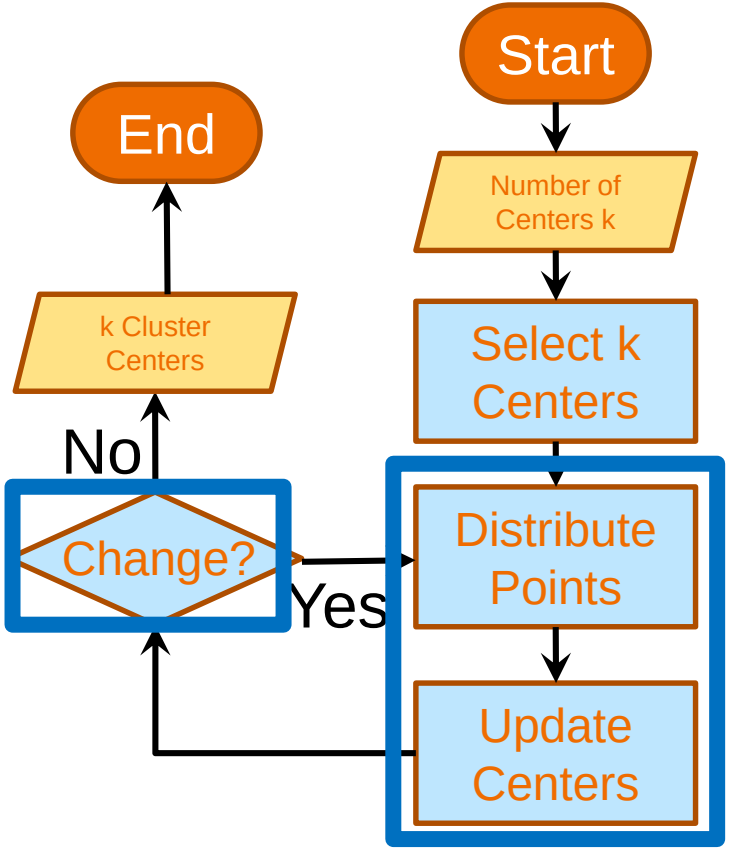
K-Means: Update 2



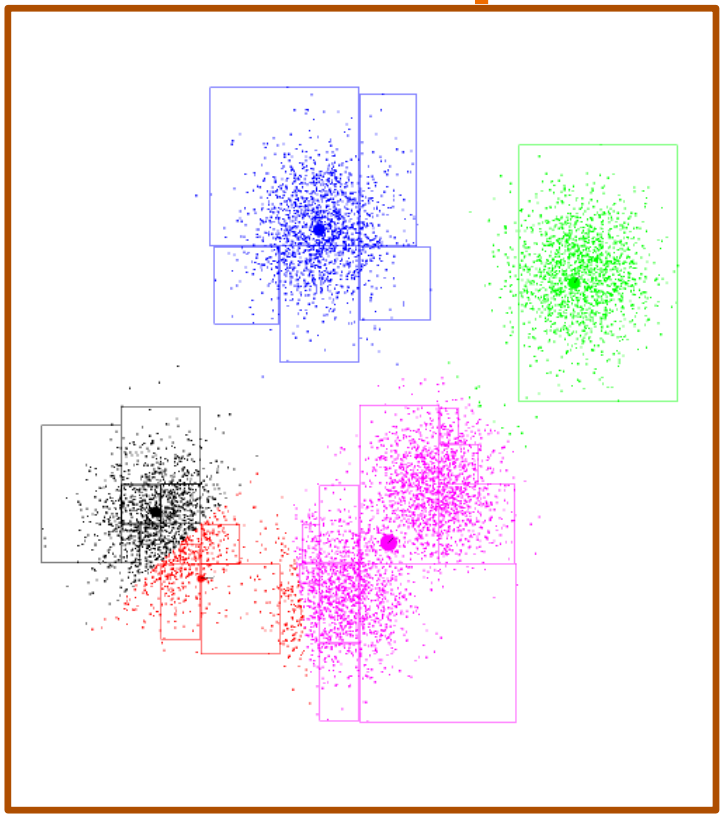


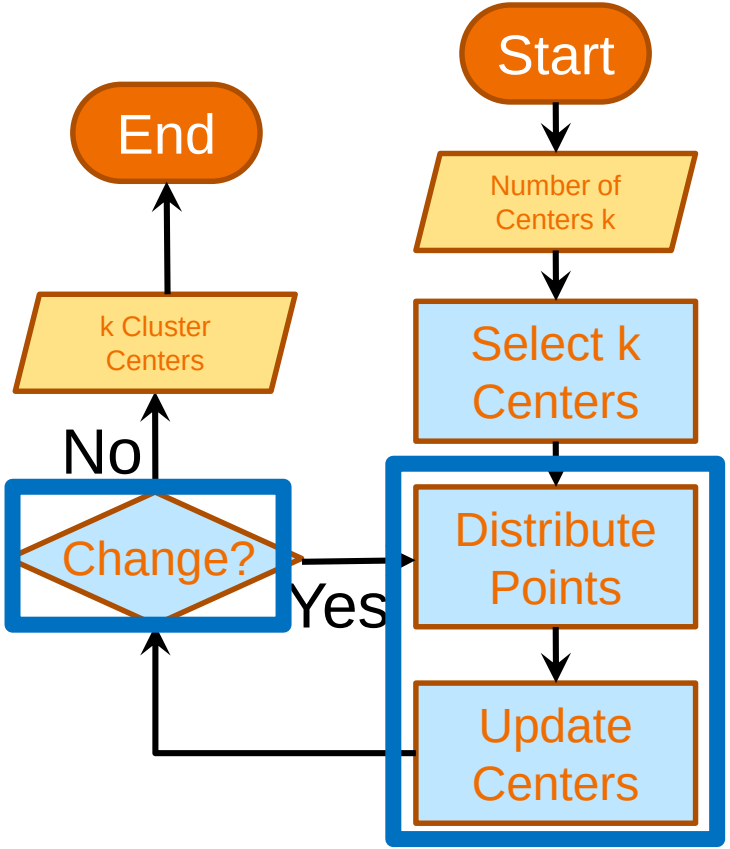
K-Means: Update 3



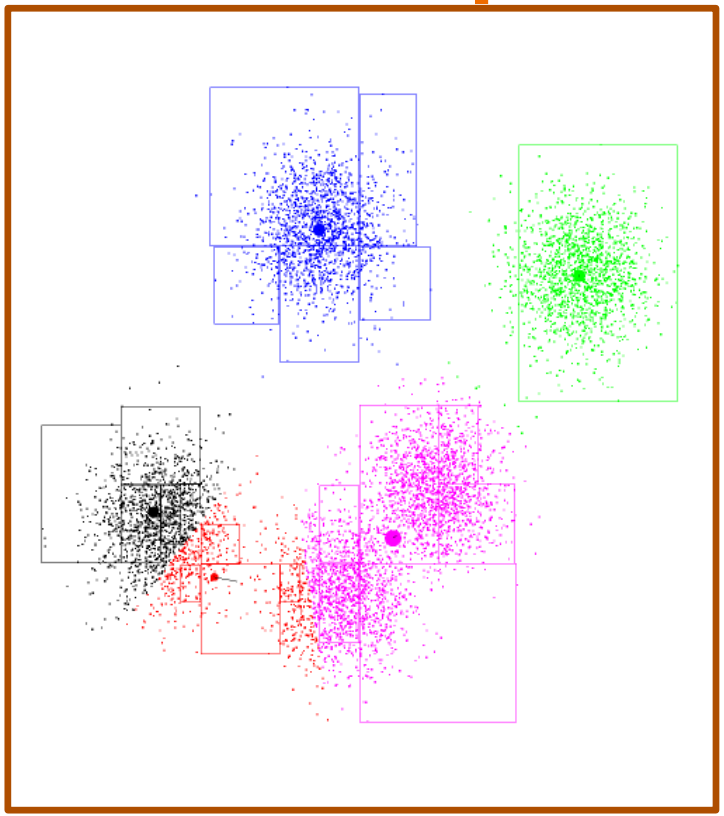


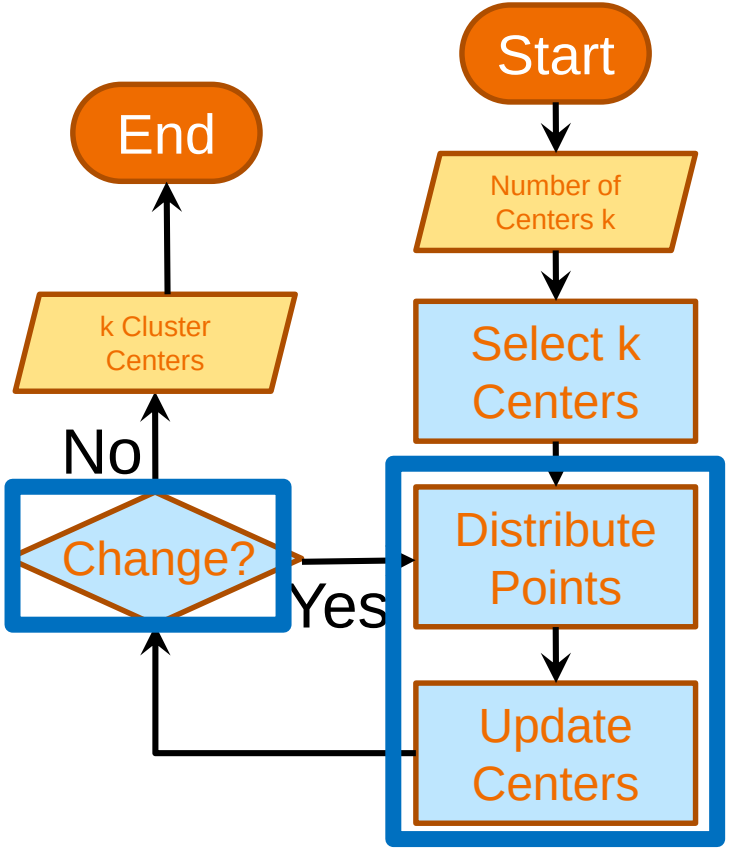
K-Means: Update 4



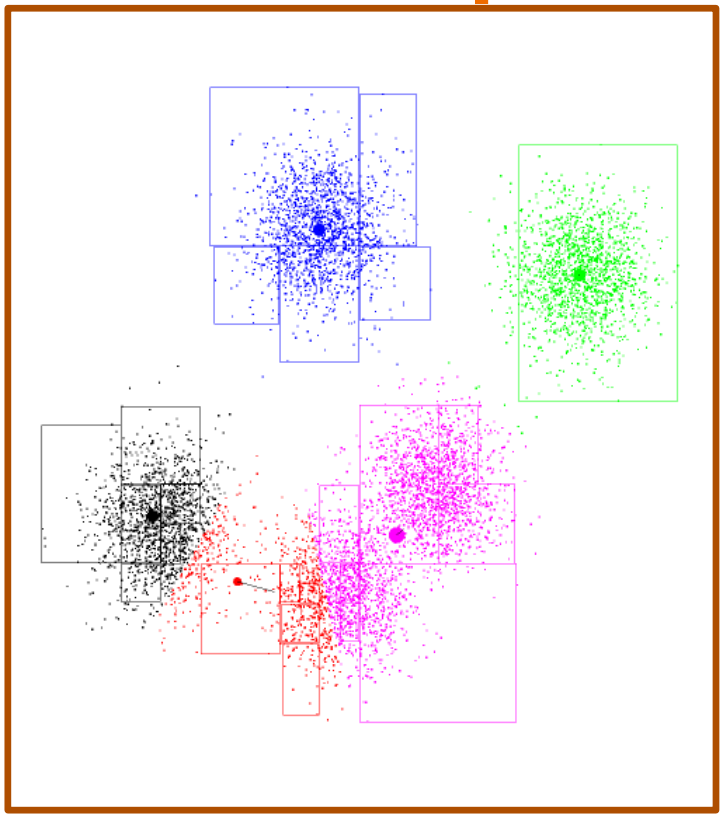


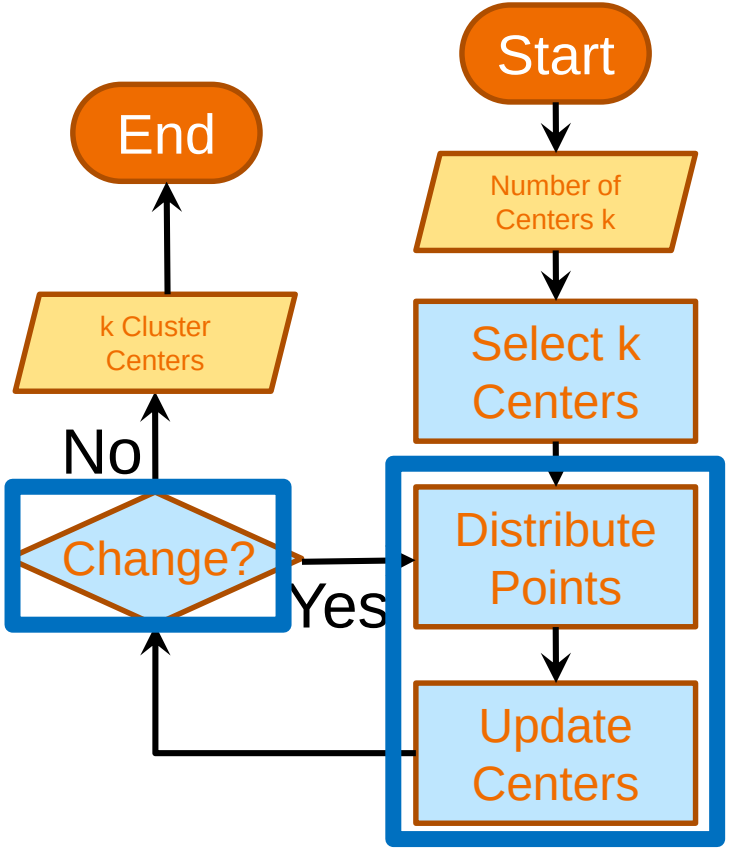
K-Means: Update 5



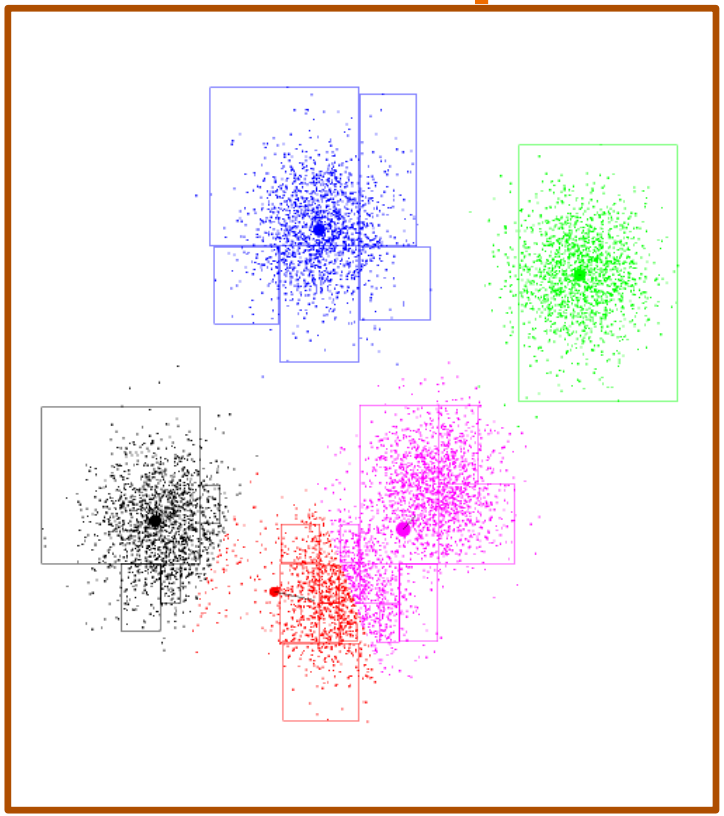


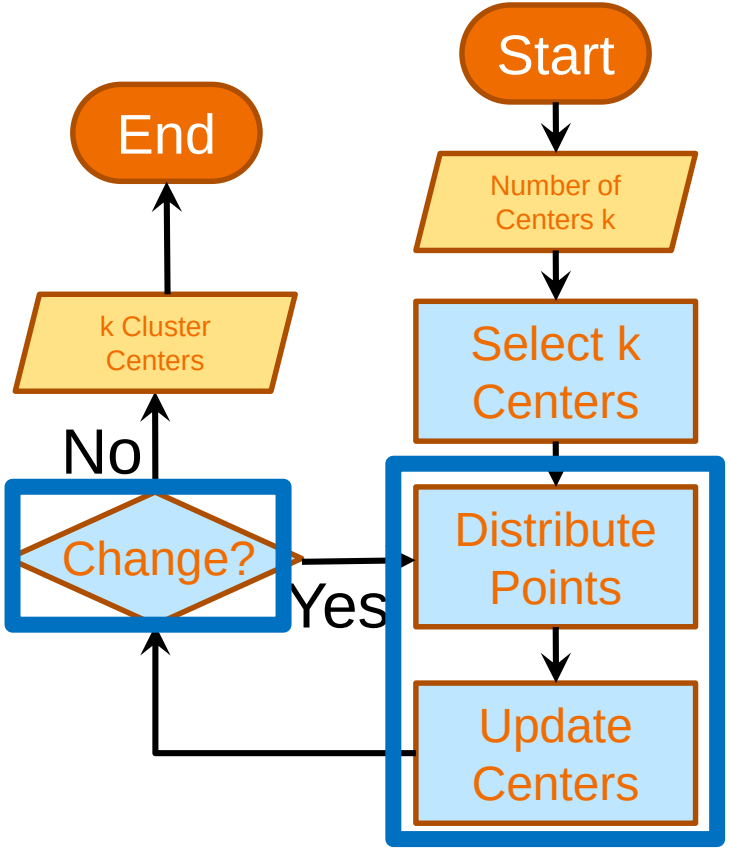
K-Means: Update 6



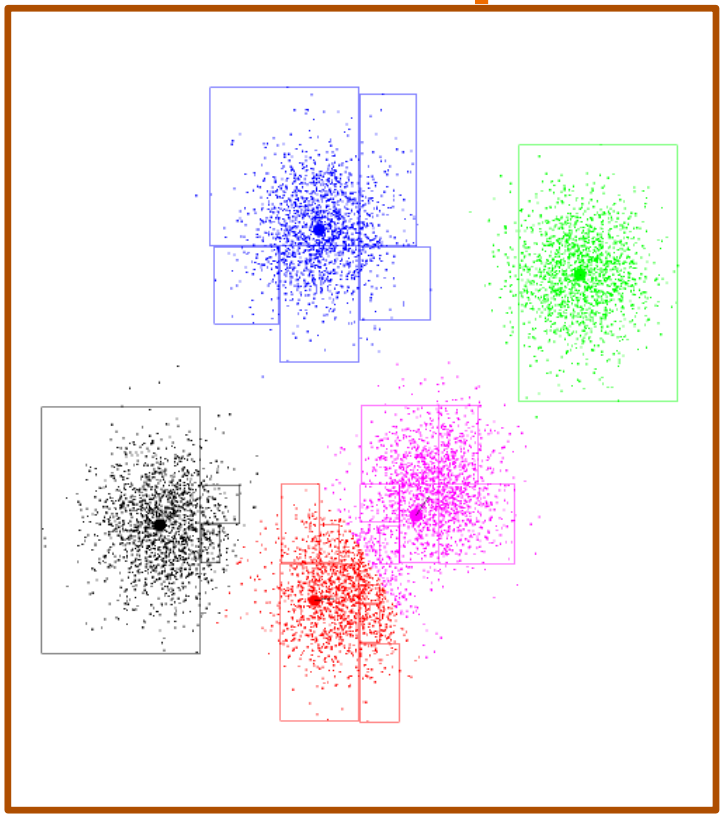


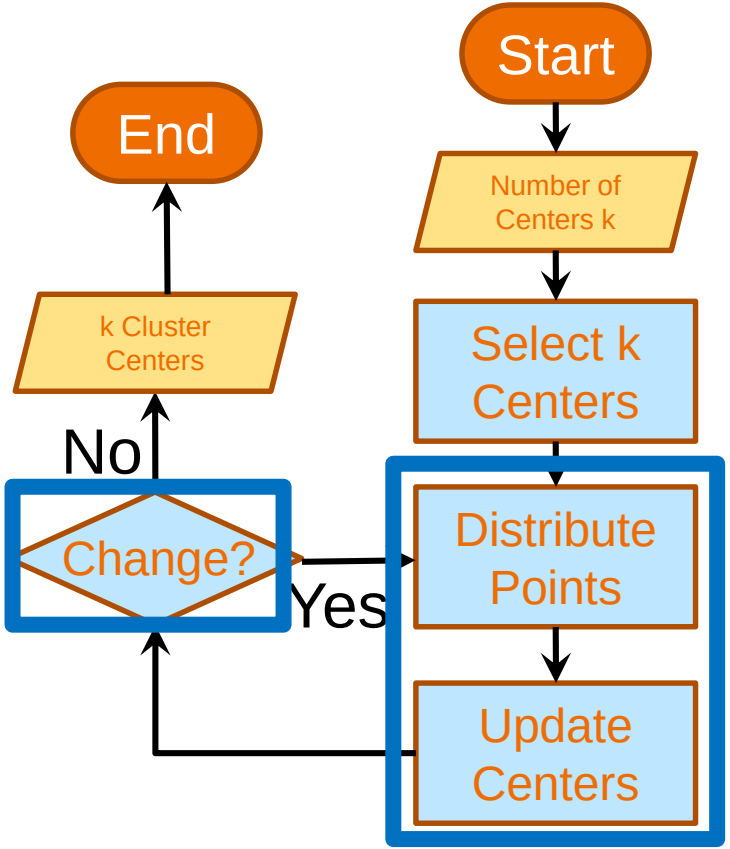
K-Means: Update 7



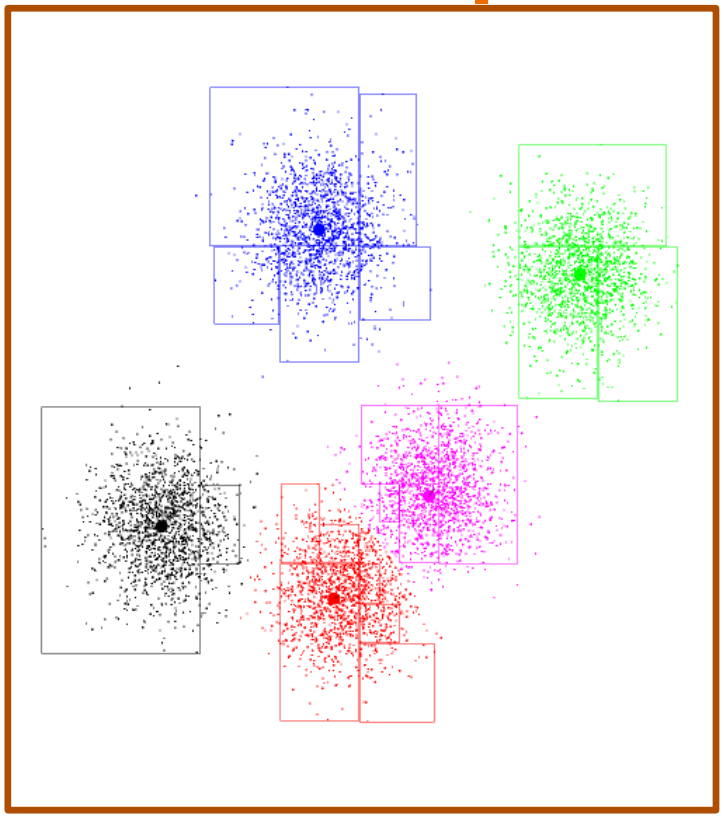


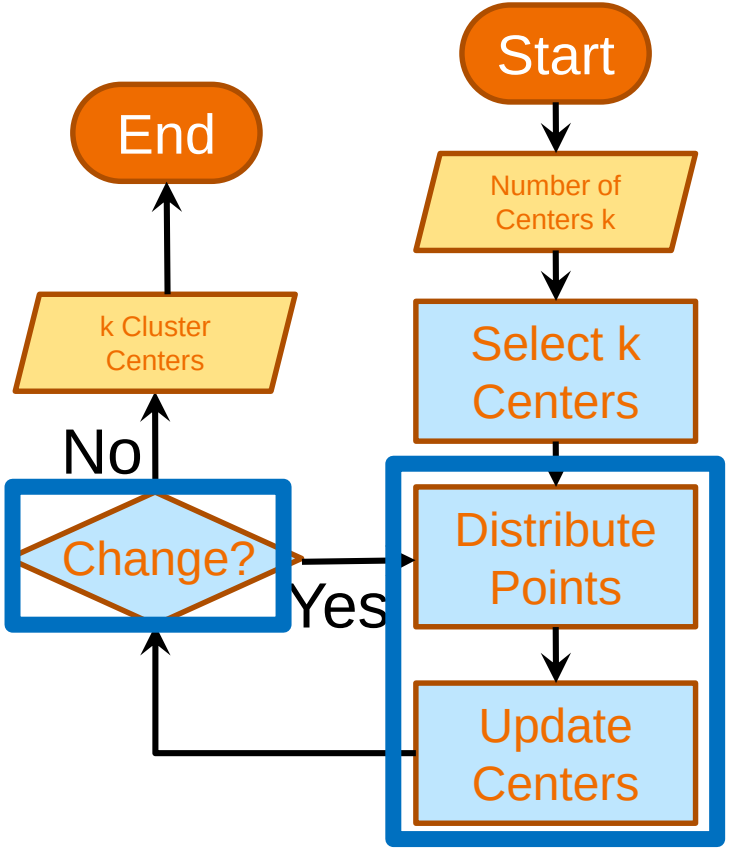
K-Means: Update 8





K-Means: Update 9





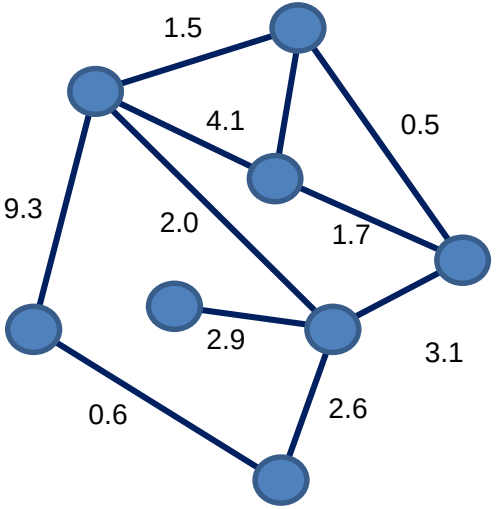
Hierarchical Clustering

Hierarchical Clustering

- Data in the world is not flat
  - Animals, Trees, Birds, Fish, Rocks
  - Types of Animals, Species, Subspecies, Types of rocks,...
- Can we recover the hierarchical structure from clustering
  - Agglomerative vs. Divisive
  - Bottom-up vs. Top-down

Graph Theory: A short review

- A graph is a collection of
  - Vertices, and
  - Edges
- Edges can have weights on them
- Tree is a special kind of graph

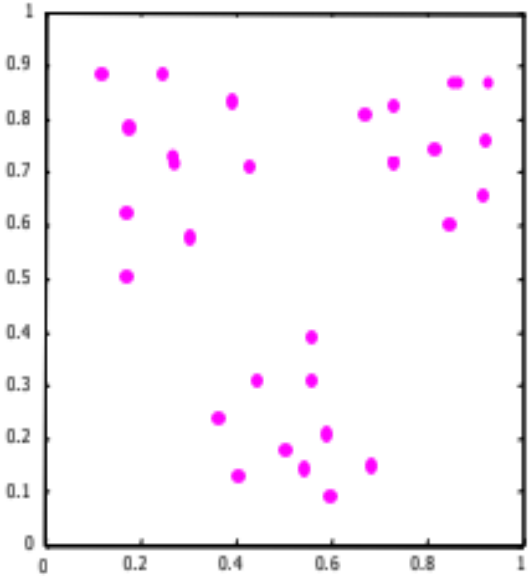


Single-Link Algorithm

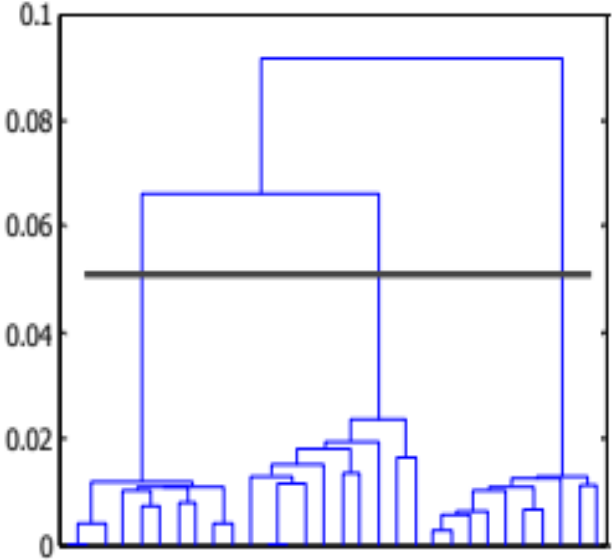
- Form a hierarchy for the data points (dendrogram), which can be used to partition the data

- The “closest” data points are joined to form a
- cluster at each step
- Closely related to the minimum spanning tree-based clustering

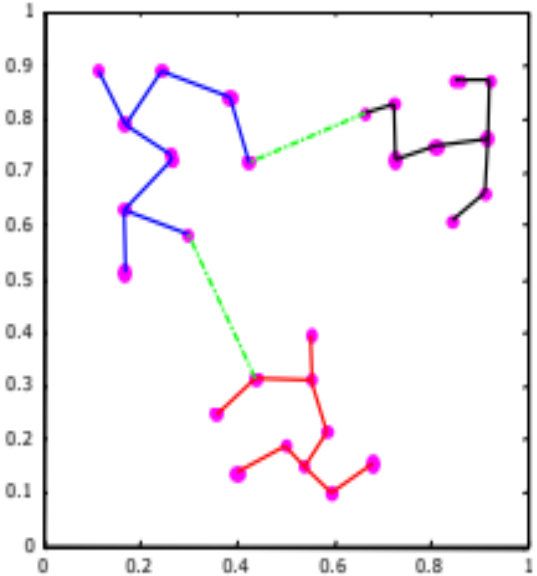
Single-Link Algorithm



Data



Dendrogram



Single-link, k=3

User's Dilemma!

- Which similarity measure and features to use?
- How many clusters?
- Which is the "best" clustering method?
- Are the individual clusters and the partition valid?
- How to choose algorithmic parameters?

Data Clustering: Jain and Dubes.

Association Rules

Identifying Co-occuring Patterns

Association Rules

$\mathcal{D}$:

Transaction ID	Items
1	Tomato, Potato, Onions
2	Tomato, Potato, Brinjal, Pumpkin
3	Tomato, Potato, Onions, Chilly
4	Lemon, Tamarind

Rule: Tomato, Potato $\rightarrow$ Onion (confidence: 66%, support: 50%)

Support(X) = |transactions containing X| / $|\mathcal{D}|$

Confidence(R) = support(R) / support(LHS(R))

Problem proposed in [AIS 93]: Find all rules satisfying user given minimum support and minimum

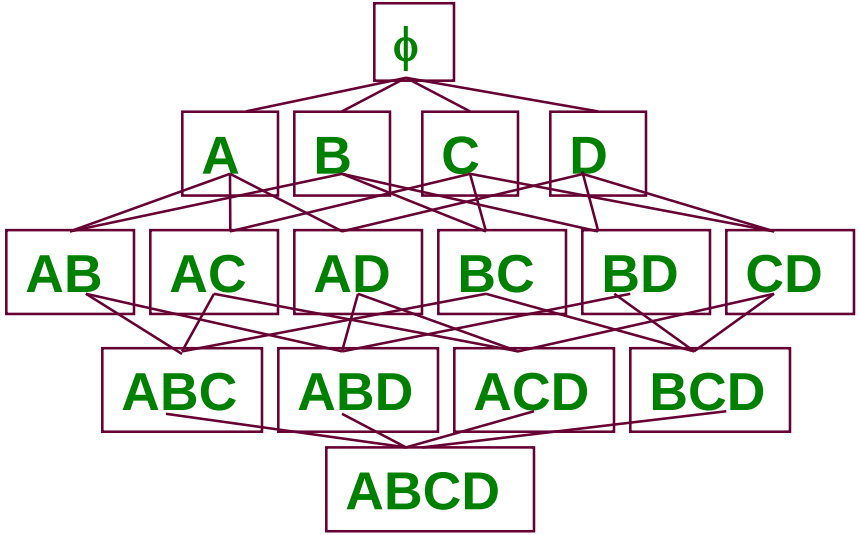
Typical Solution Strategy

- STEP 1: Find all *frequent* itemsets
  (computationally expensive)
  - Itemset X is frequent iff $support(X) \geq minsup$

- STEP 2: Find *rules* from the frequent itemsets
  (computationally inexpensive)
  - Rule quantity: *too many rules* are usually

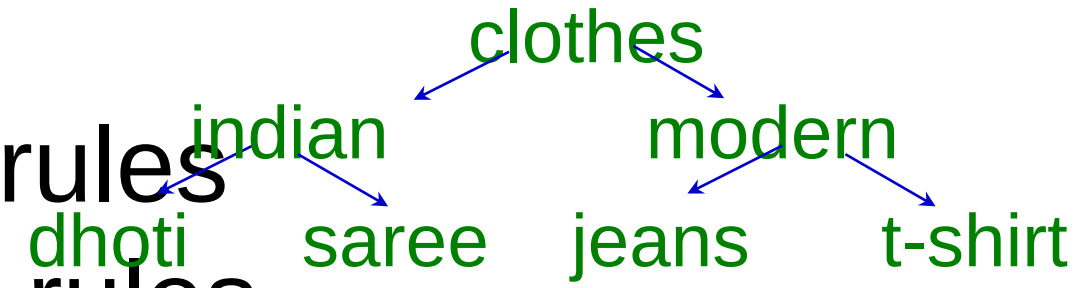
Difficulty

- Extremely computationally expensive
- Naïve solution
  - exponential time and memory w.r.t. $|I|$
  - linear time w.r.t. $|\mathcal{D}|$
- Typically, $|I|$ is in thousands, $|\mathcal{D}|$ is in billions



Types of Association Rules

- Boolean association rules
- Hierarchical rules



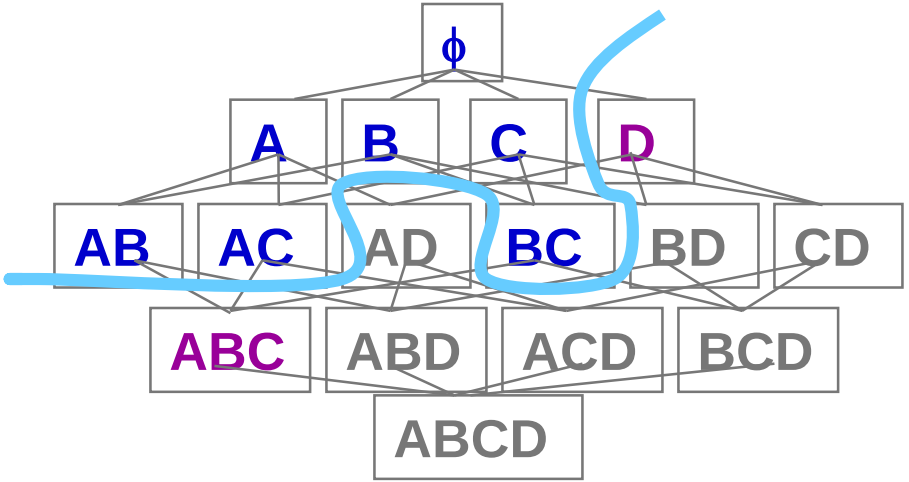
dhoti, saree → t-shirt

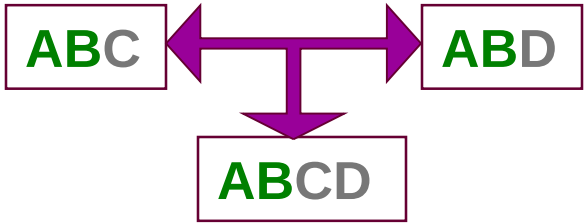
More Types

- Cyclic / Periodic rules
  - Sunday → vegetables
  - Christmas → gift items
  - Summer, rich, jobless → ticket to Hawaii
- Constrained rules
  - Show itemsets whose average price > Rs.10,000
  - Show itemsets that have television on RHS
- Sequential rules
  - Star wars, Empire Strikes Back → Return of the Jedi

The Apriori Algorithm

Idea: An itemset can be frequent only if all its subsets are frequent.



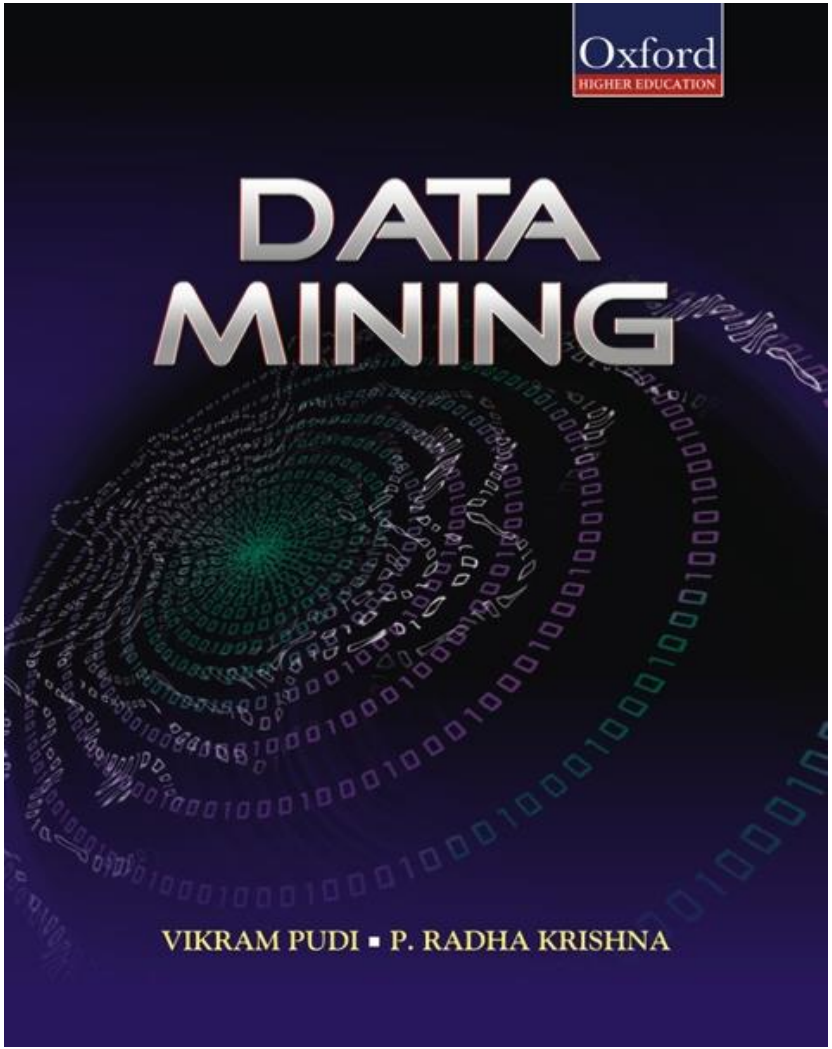



```
Apriori( DB, minsup ):
C = {all 1-itemsets}
            // candidates = singletons
while ( |C| > 0 ):
    make pass over DB, find counts of C
    F = sets in C with count ≥ minsup*|DB|
    output F
    C = AprioriGen(F) // gen. candidates
```

```
AprioriGen( F ):
for each pair of itemsets X, Y in F:
   if X and Y share all items, except last
      Z = X ∪ Y // generate candidate
      if any imm. subset of Z is not in F:
         prune Z // Z can't be frequent
```

Privacy Issues

- Users provide inaccurate data to protect their privacy.
- How can inaccurate data be effectively mined?
- How can data be modified in such a way as to ensure data privacy and rule accuracy?
- How can data be modified in such a way as to ensure rule privacy? – hide sensitive rules
- Can mined results be used to retrieve original data transactions?



Thanks!

Questions?