Data Clustering

—— A Survey

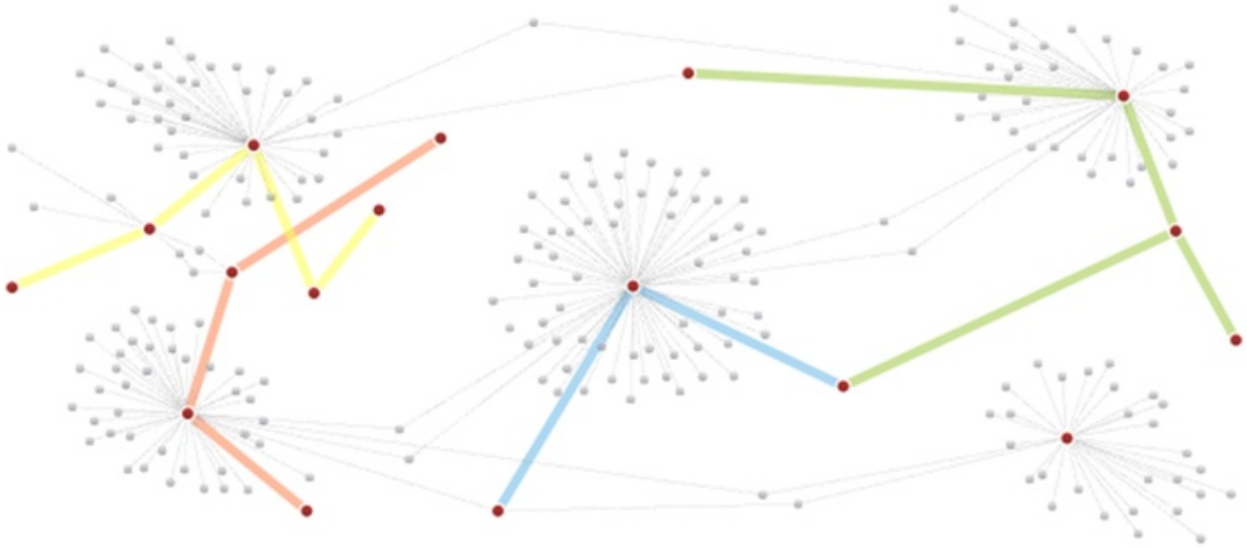
QIUYI ZHANG 12330402

Sun Yat-Sen University · Dec. 20, 2014

CONTENTS

Organize
Analyze
Summerize

The explosion of data

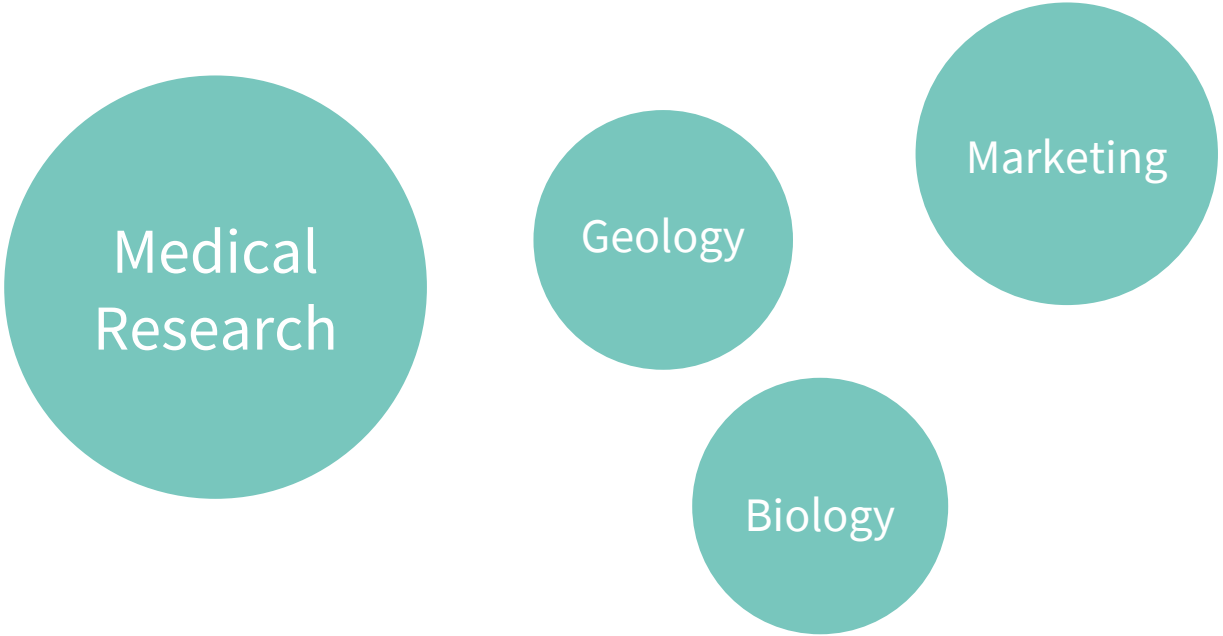


What is data clustering?

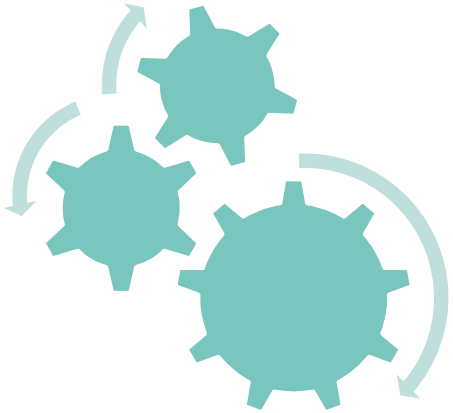
Computer science?

Well...

Not just computer science.



The development of data clustering



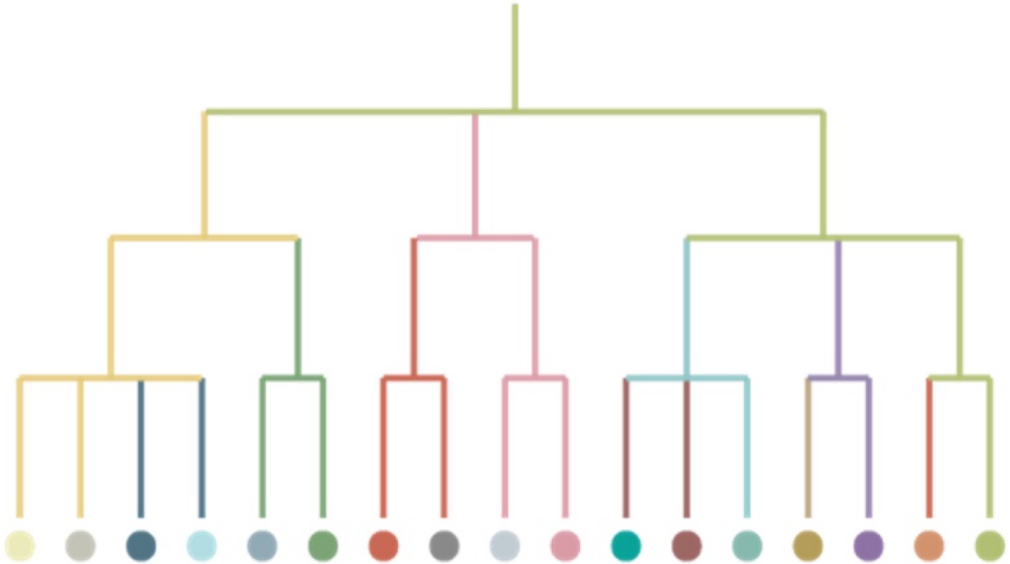
Algorithm

Criteria

The development of data clustering

- Image processing
- Object and feature recognition
- Information retrieval
- And more...

Applications of data clustering



Hierarchy
= Tree
= Dendrogram

Hierarchical clustering

Agglomerative
(bottom-up)

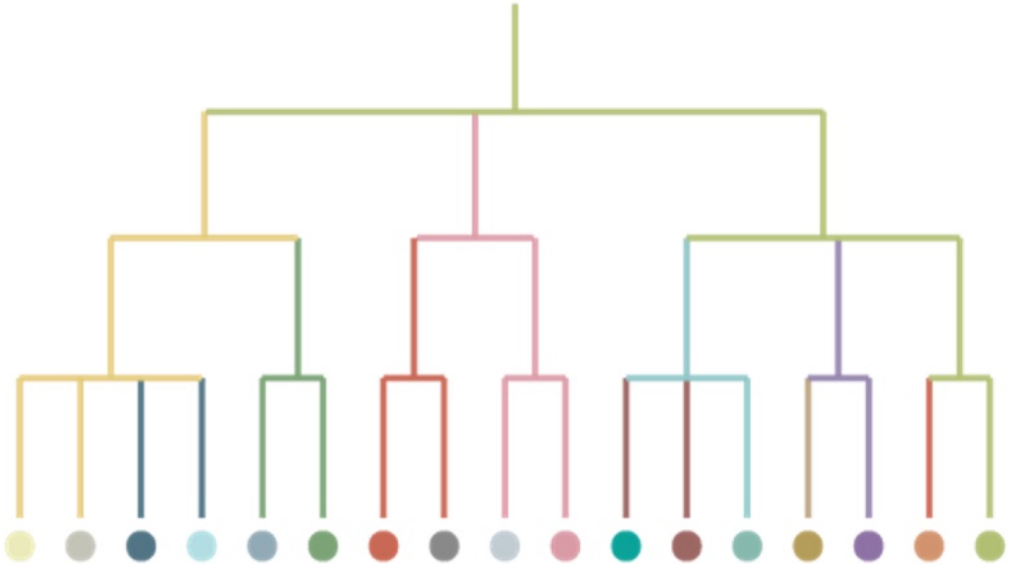
Hierarchical clustering

Agglomerative
(bottom-up)

Hierarchical clustering

Agglomerative
(bottom-up)

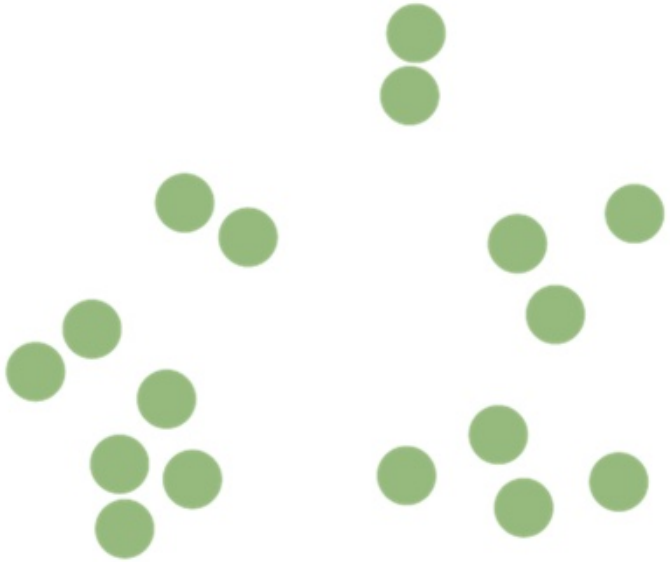
Hierarchical clustering



Hierarchy
= Tree
= Dendrogram

Hierarchical clustering

Divisive
(top-down)



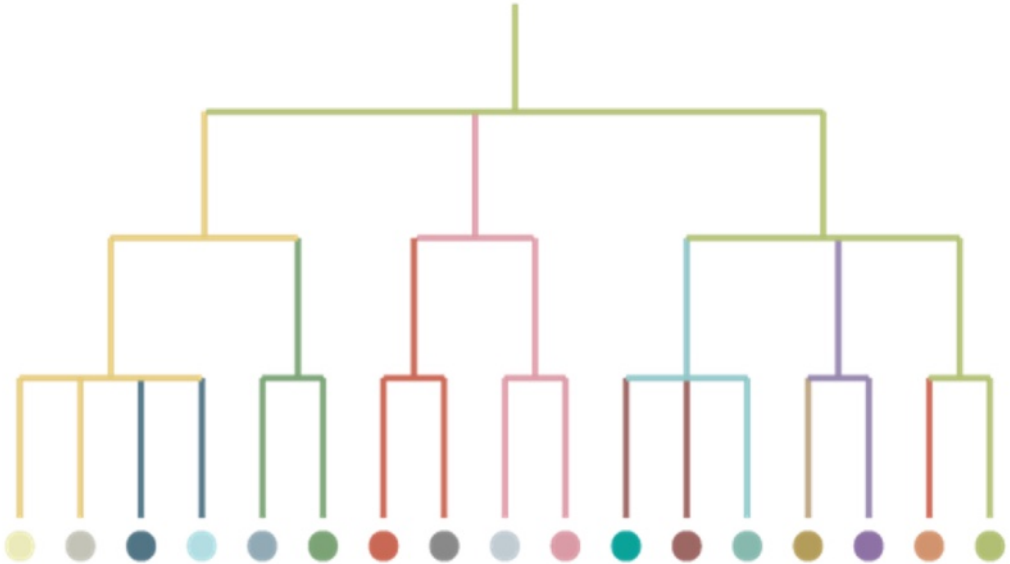
Hierarchical clustering

Divisive
(top-down)

Hierarchical clustering

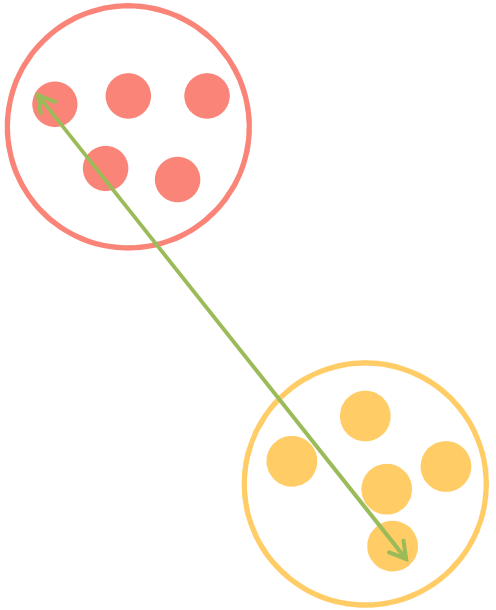
Divisive
(top-down)

Hierarchical clustering



Hierarchy
= Tree
= Dendrogram

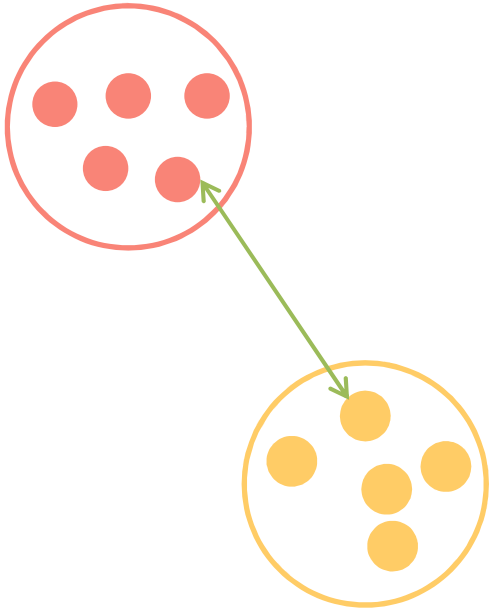
Hierarchical clustering



CLINK (complete-linkage)

Smallest maximun pairwise distance

CLINK and SLINK(Agglomerative)



SLINK (single-linkage)

Smallest minimun pairwise distance

Typically

$O(n^3)$

(n x 10^3 in a minute)

Complexity of hierarchical clustering

Not good...
even for a 100 x 100 image!

Complexity of hierarchical clustering

Can be optimized to

$O(n^2)$

(n x 10^5 in a minute)

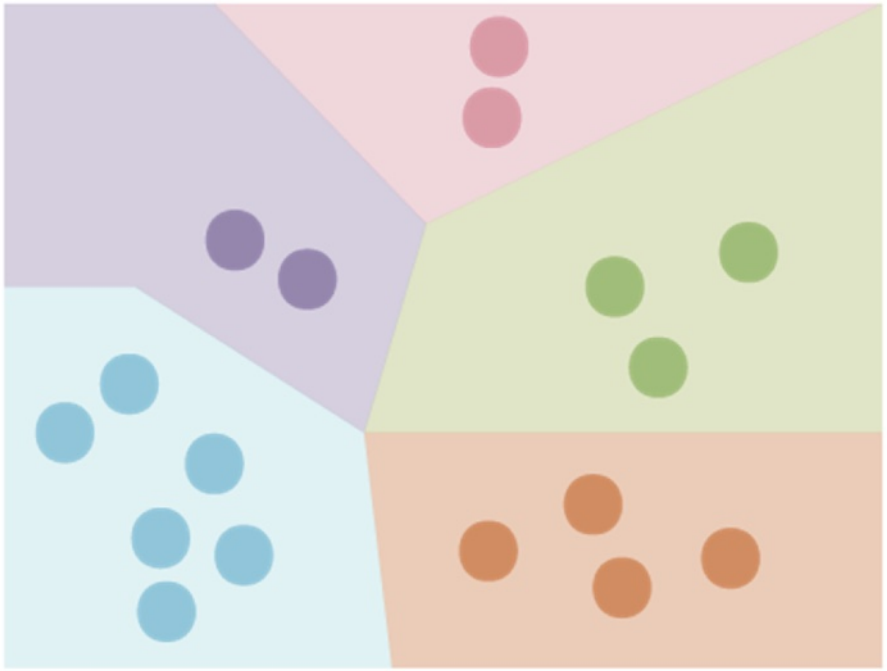
Complexity of hierarchical clustering

Well,
still not good enough.

Complexity of hierarchical clustering

- BIRCH
  - Balanced Iterative Reducing and Clustering using Hierarchies
- Chameleon
  - dynamic modeling
- ROCK
  - RObust Clustering using linKs

Improvements



Partition,
Not hierarchy.

Partitional clustering

Original data

The classic k-means

Initial k means

The classic k-means

Assign

The classic k-means

Update

The classic k-means

Assign

The classic k-means

Update

The classic k-means

Assign

The classic k-means

Complexity of k-means

$O(nkt)$

n = number of objects
k = number of clusters
t = number of iterations

($n \times 10^9$ in a minute)

Efficent,
But not stable.

Drawbacks of k-means

Drawbacks of k-means

- Use medians or menoids
- k-means++
  - chooses the initial values carefully
- Fuzzy C-Means algorithm
  - allows data points belong to more than one cluster
  - associate each point with a membership level

Improvements

Clusters

II

Regions with higher density

Original data

DBSCAN

Pick a point,
Connect
based on desity.

DBSCAN

Pick another unvisited point, connect.

DBSCAN

Pick & connect.

DBSCAN

Pick & connect.

DBSCAN

Pick & connect.

DBSCAN

O(nlogn)

(n x 10^7 in a minute)

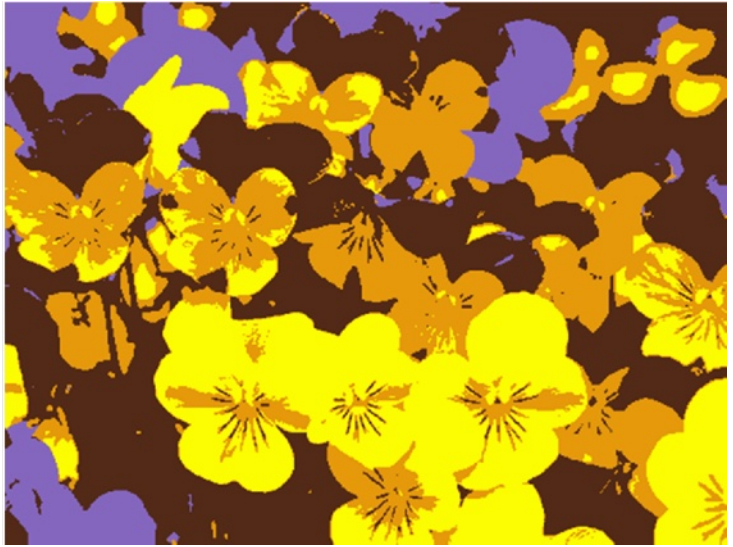
Complexity of DBSCAN

- GDBSCAN
  - Generalized, allow non-spatial attributes
- PDBSCAN
  - Parallel
- BRIDGE
  - combines the k-means algorithm with DBSCAN

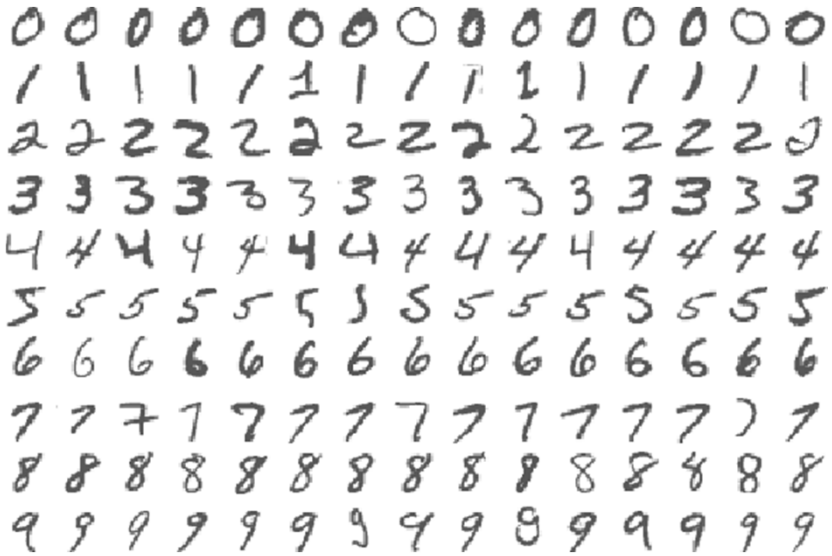
Improvements

Image Segmentation

Color Reduction



Characters in handwriting

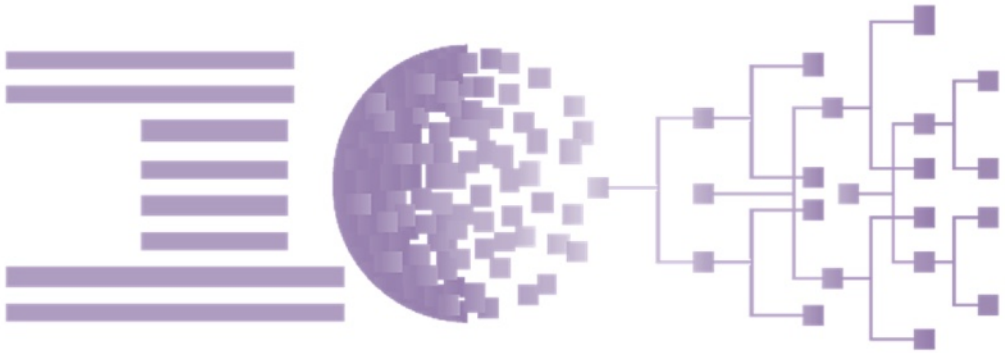


Object and feature recognition

Recognize scenes, events, instruments, ...

Object and feature recognition

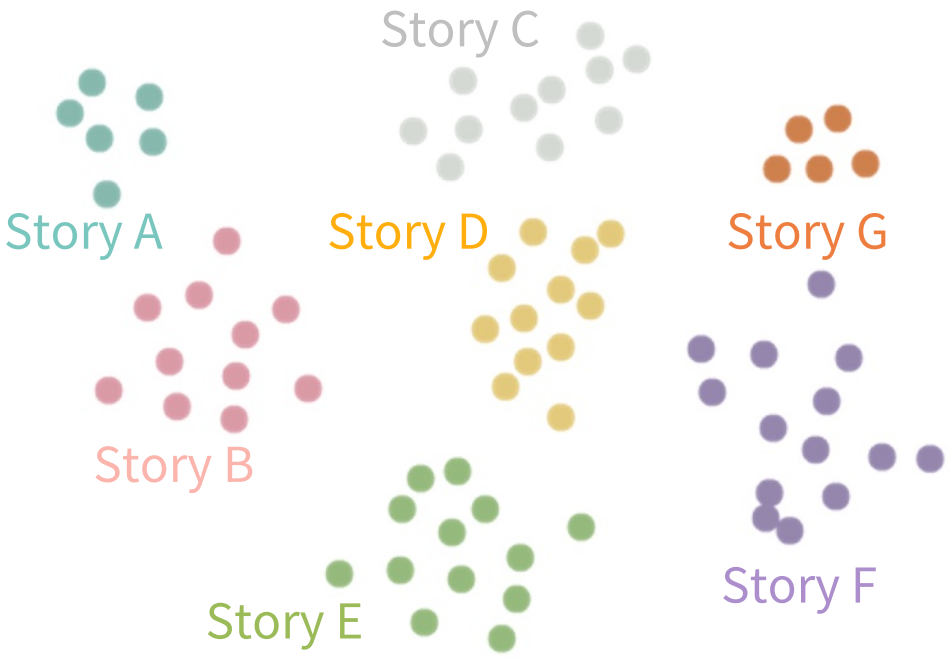
Document Classification



Information retrieval

Better ranking
for search results.

Information retrieval



Cluster as representation itself.

Information retrieval

Both works.

No one is perfect.

THANKS