Clustering

James Sharpnack

Lecture adapted from notes of Sontag, Blei, L. Mackey, R.J. Tibshirani

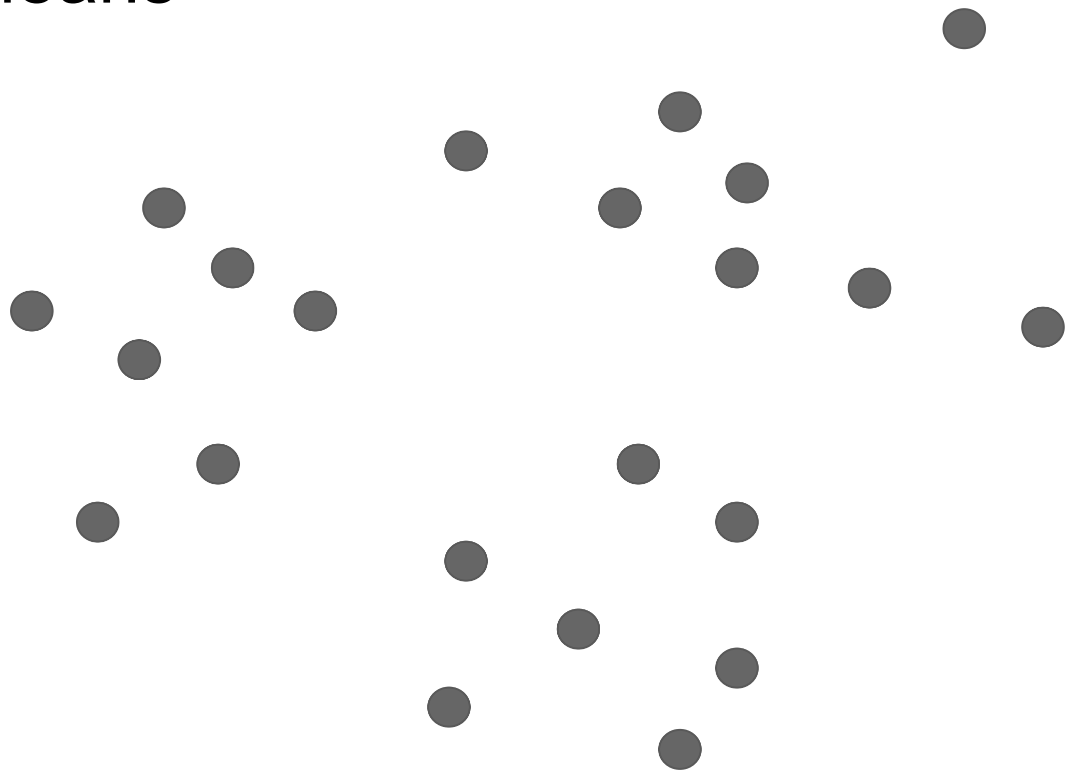
Motivations

- Compressed representations to save storage and computation
- Reduce noise, deal with missingness
- Visualization and exploratory data analysis
- Semi-supervised learning: create features that are used in supervised learning (label propagation)
- Dictionary learning: learning basis elements that provide sparse representations in supervised learning

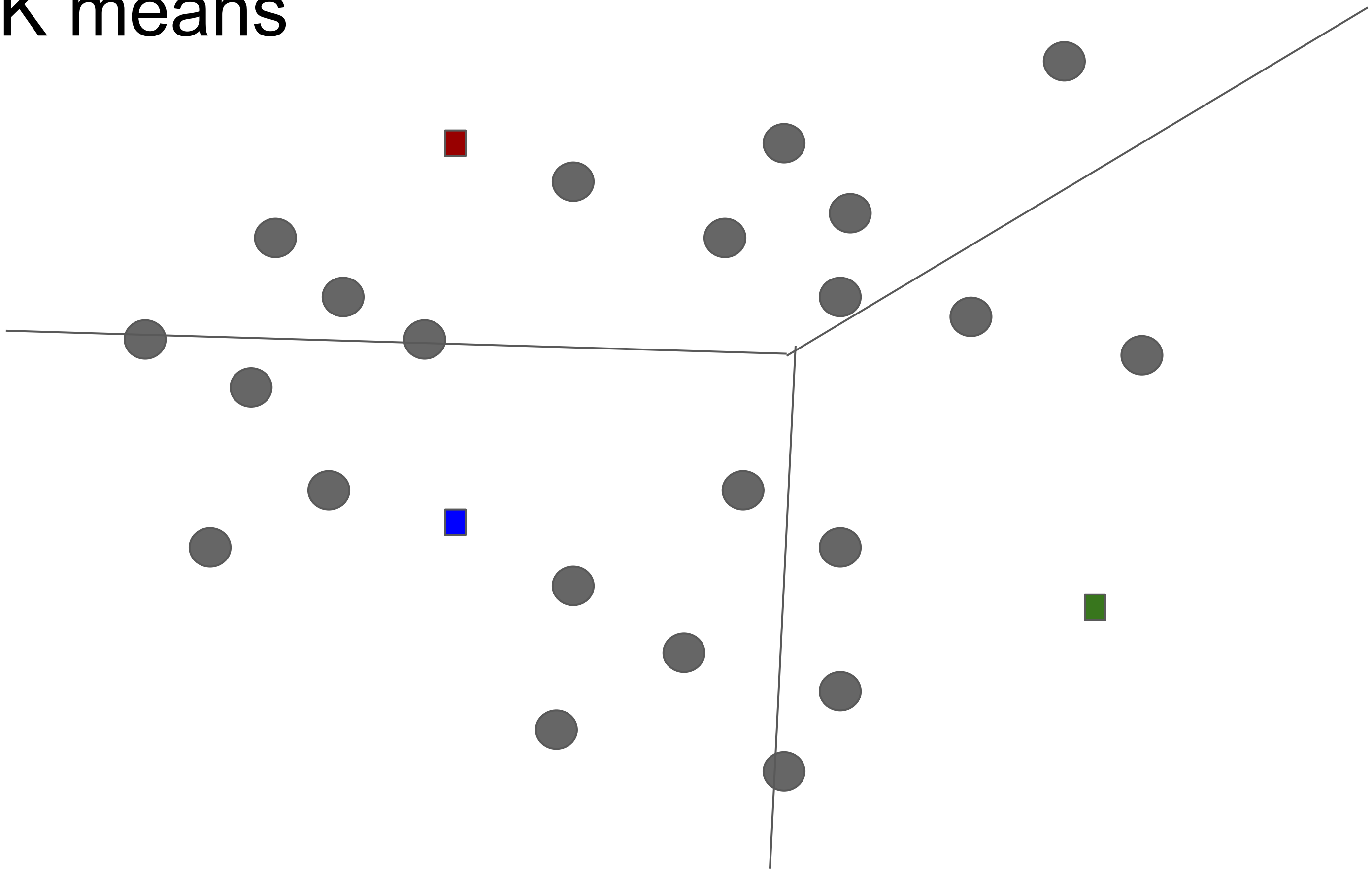
K means

Setting, objective, and algorithm

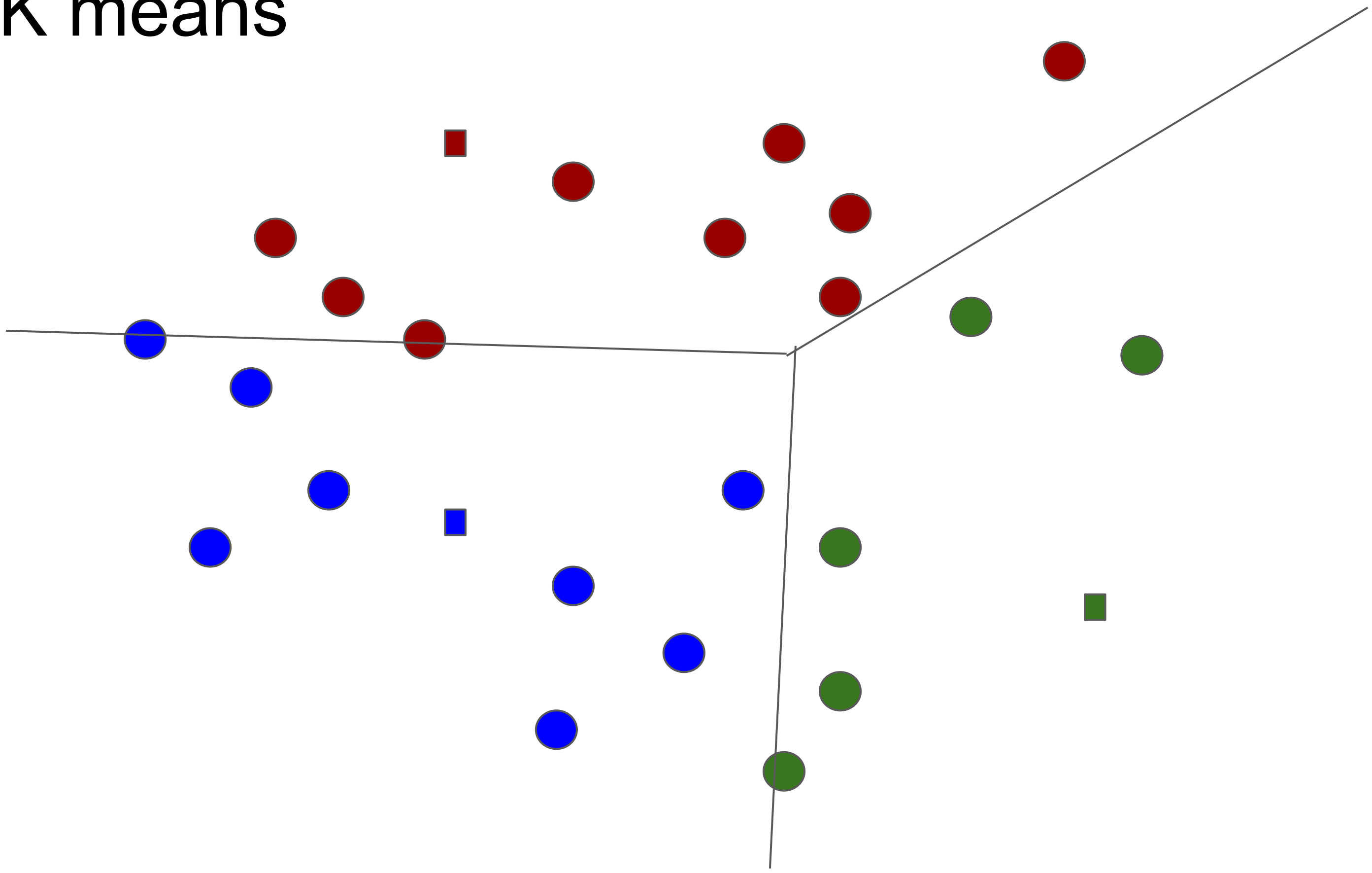
K means



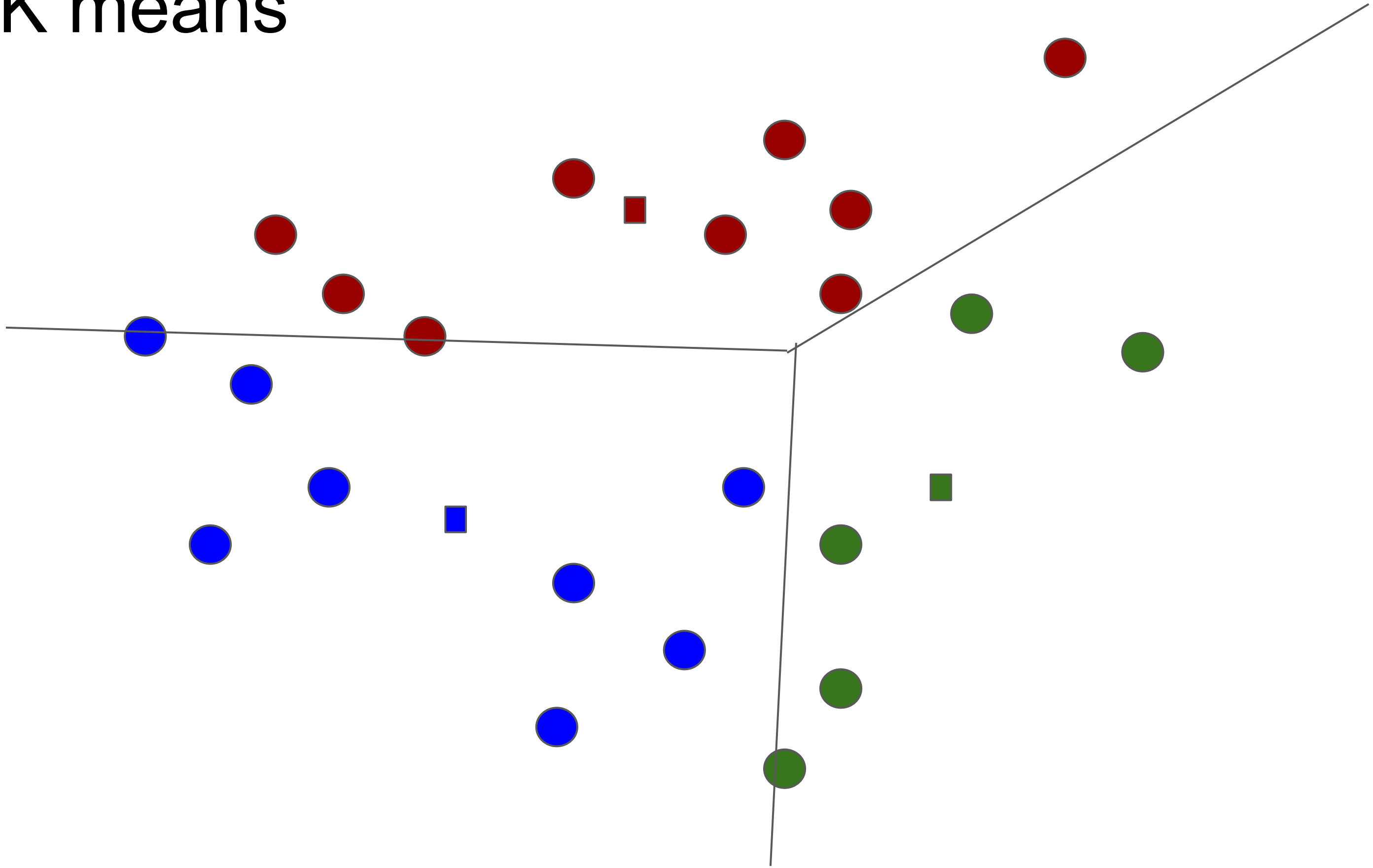
K means



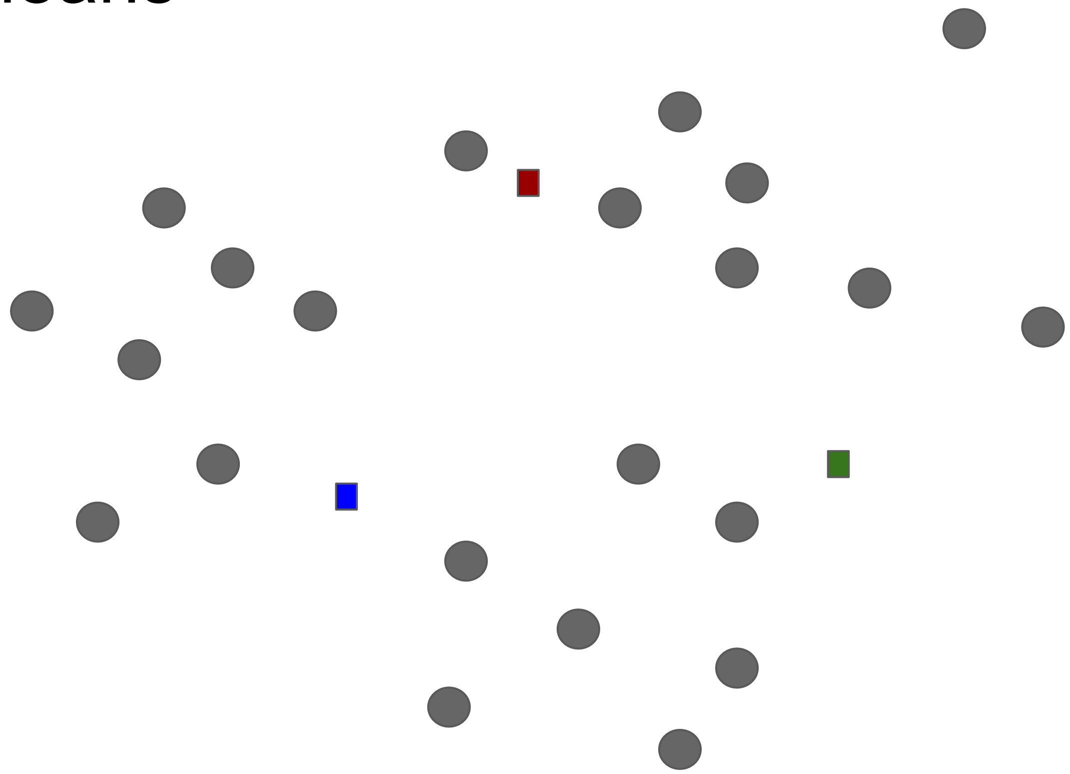
K means



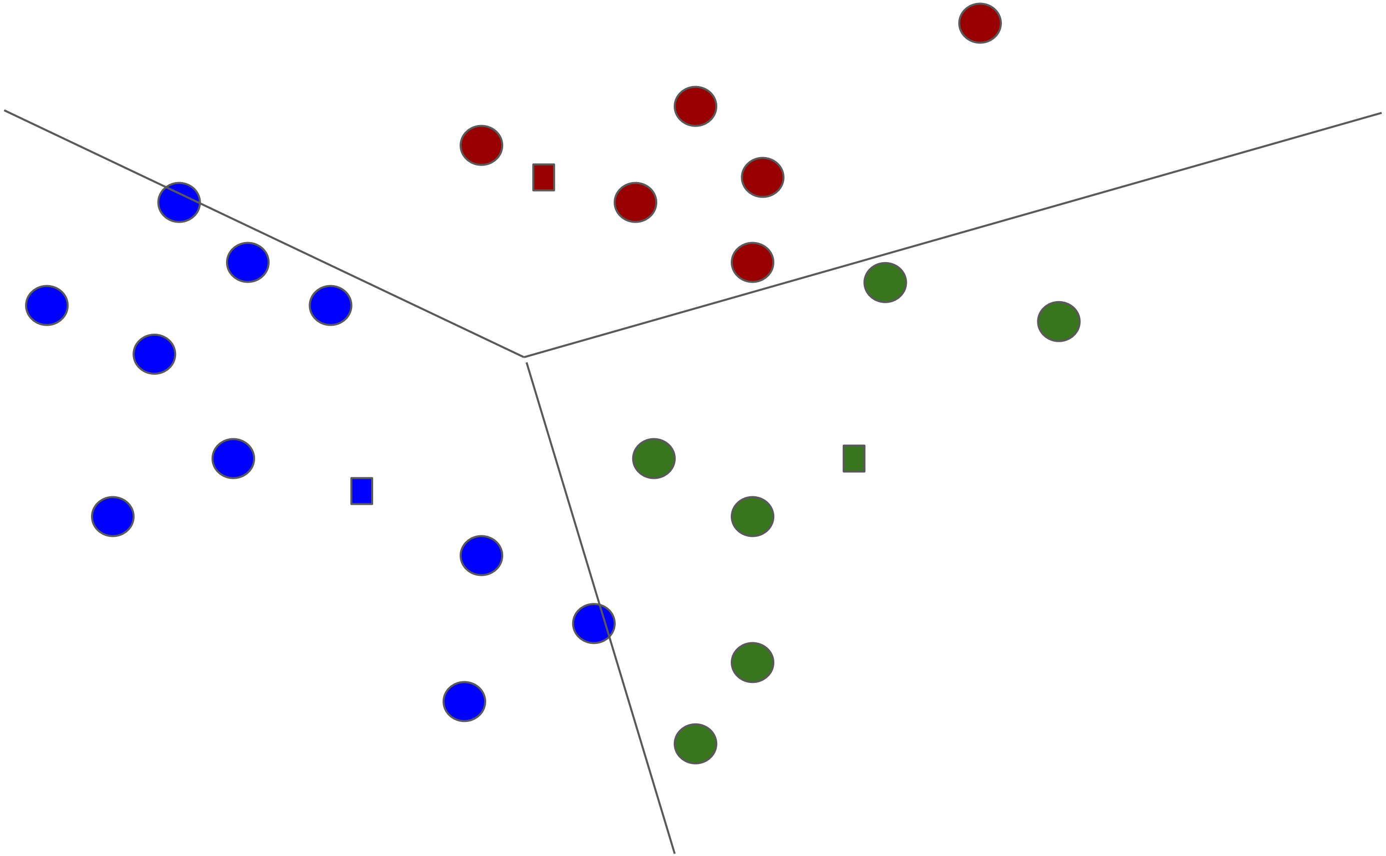
K means



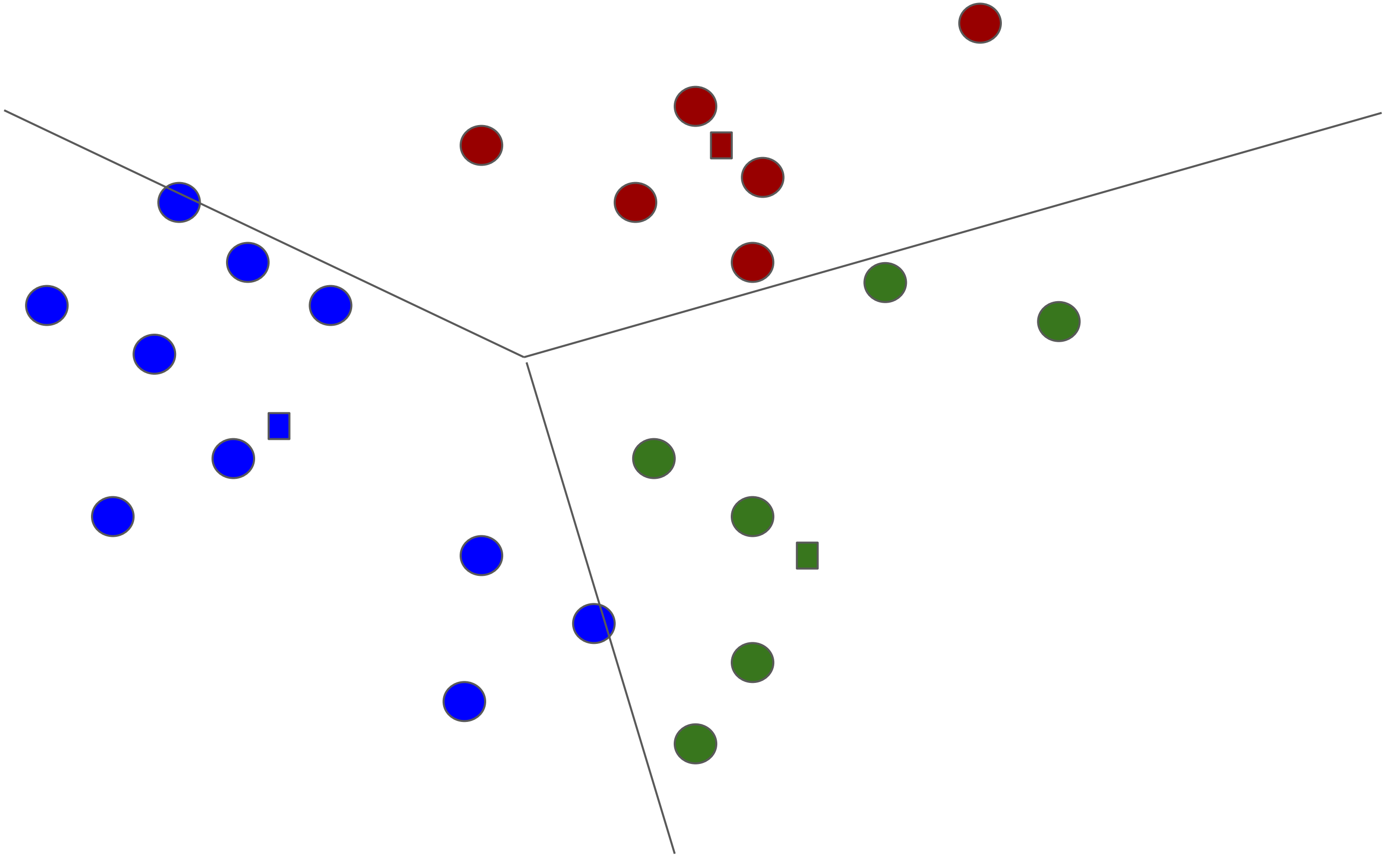
K means



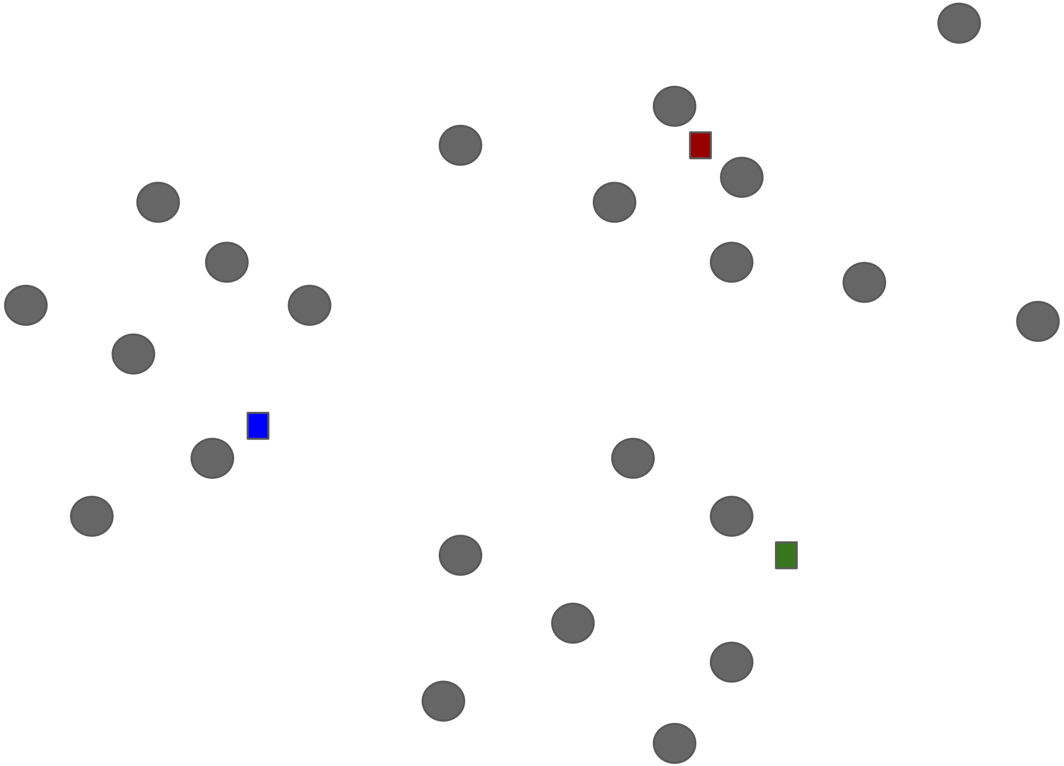
K means



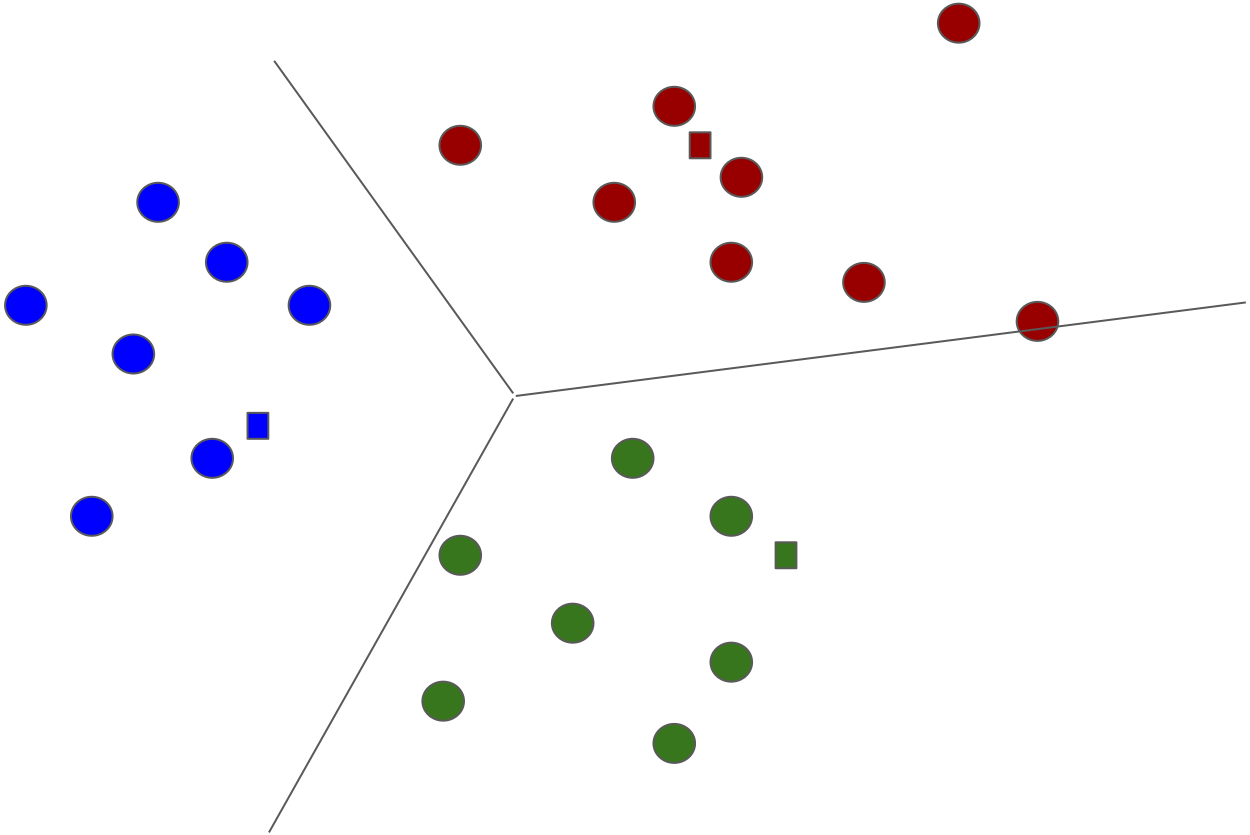
K means



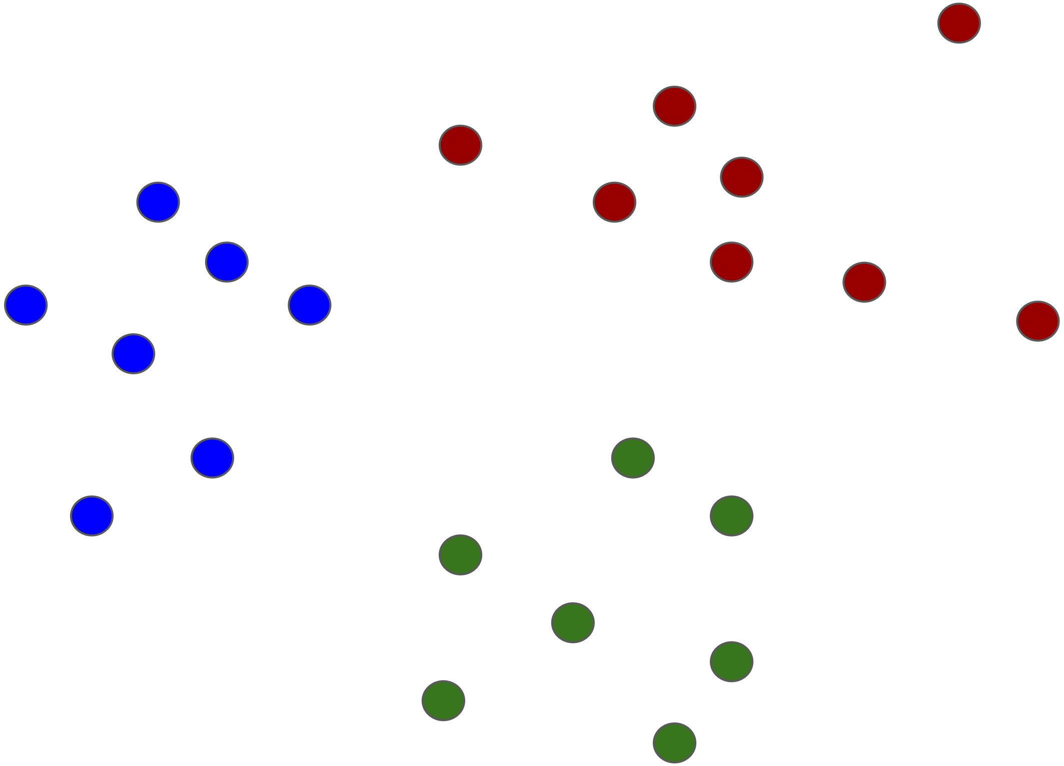
K means



K means



K means



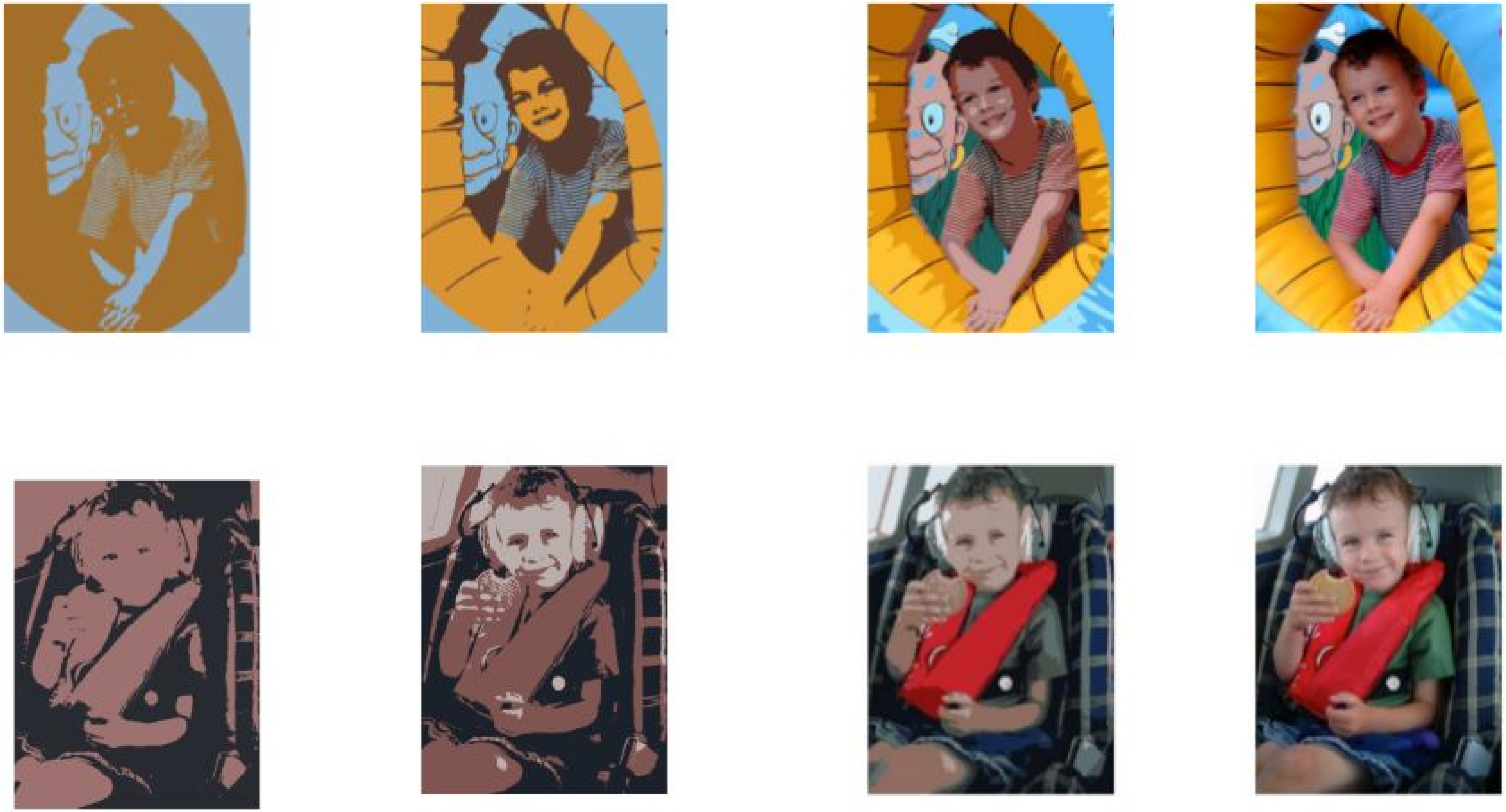
Properties

- Objective always converges (may be exponential time)
- Can use many dissimilarities, L2, L1, hamming distance
- Choosing k is not clear: **Gap statistic**
- Convergence may be slow: **K-means++** run Lloyd's algorithm with random initialization, or use random restarts
- Cluster centers are not usually data point.
- **K-medoids** is kmeans but the cluster centers are chosen to be the data points that minimize distortion
- Can make transformations as before!

Example: Image Segmentation

(1) Each pixel has an RGB value (3 floats)
(2) Calculate color based distance between pixels (far apart pixels can be nearby in color distance!)
(3) Use k-means on pixel features
(4) Pixels are grouped according to colors

K=2 K=3 K=10 Original



4% 8% 17%

FIGURE 14.9. *Sir Ronald A. Fisher (*1890 – 1962*) was one of the founders of modern day statistics, to whom we owe maximum-likelihood, sufficiency, and many other fundamental concepts. The image on the left is a* 1024×1024 *grayscale image at* 8 *bits per pixel. The center image is the result of* 2×2 *block VQ, using* 200 *code vectors, with a compression rate of* 1.9 *bits/pixel. The right image uses only four code vectors, with a compression rate of* 0.50 *bits/pixel*

[Figure from Hastie *et al.* book]

Feature engineering

What do you want distinguishing the clusters?

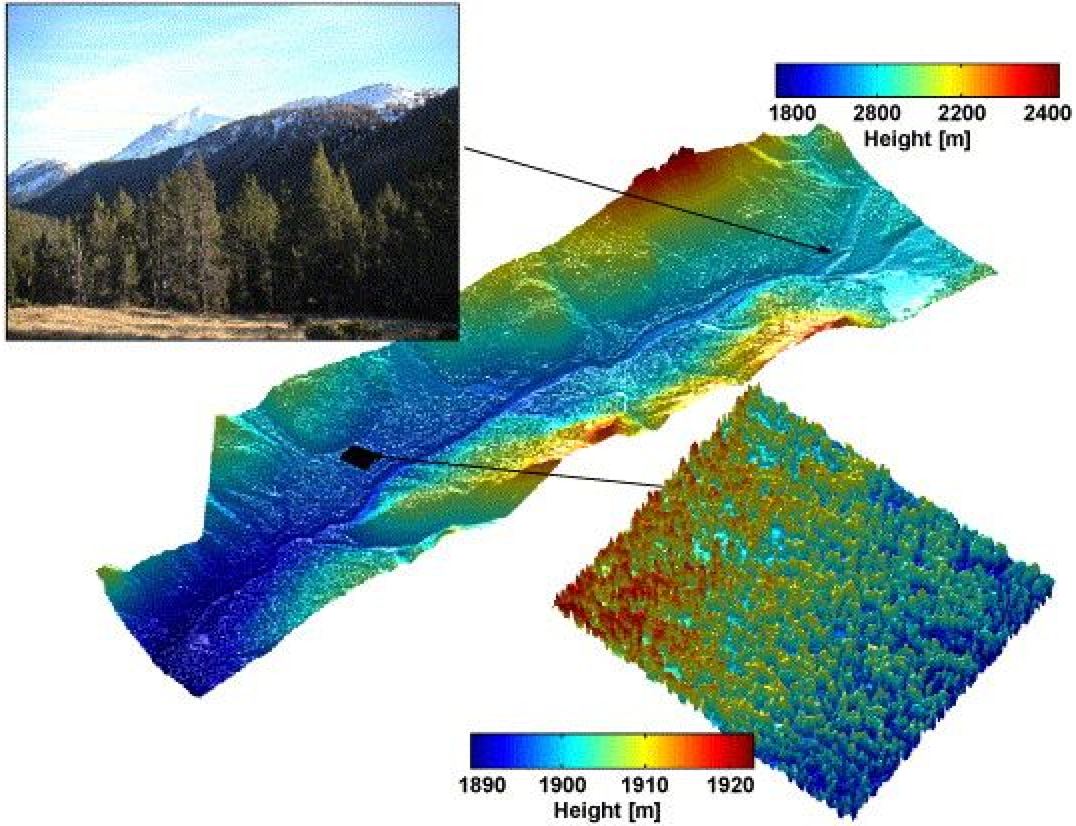
- Cluster pixels based on color, distance, or a combination
- Word content for documents: tf-idf similarity
- Switch role of words and documents and cluster the words based on document counts

Clustering words

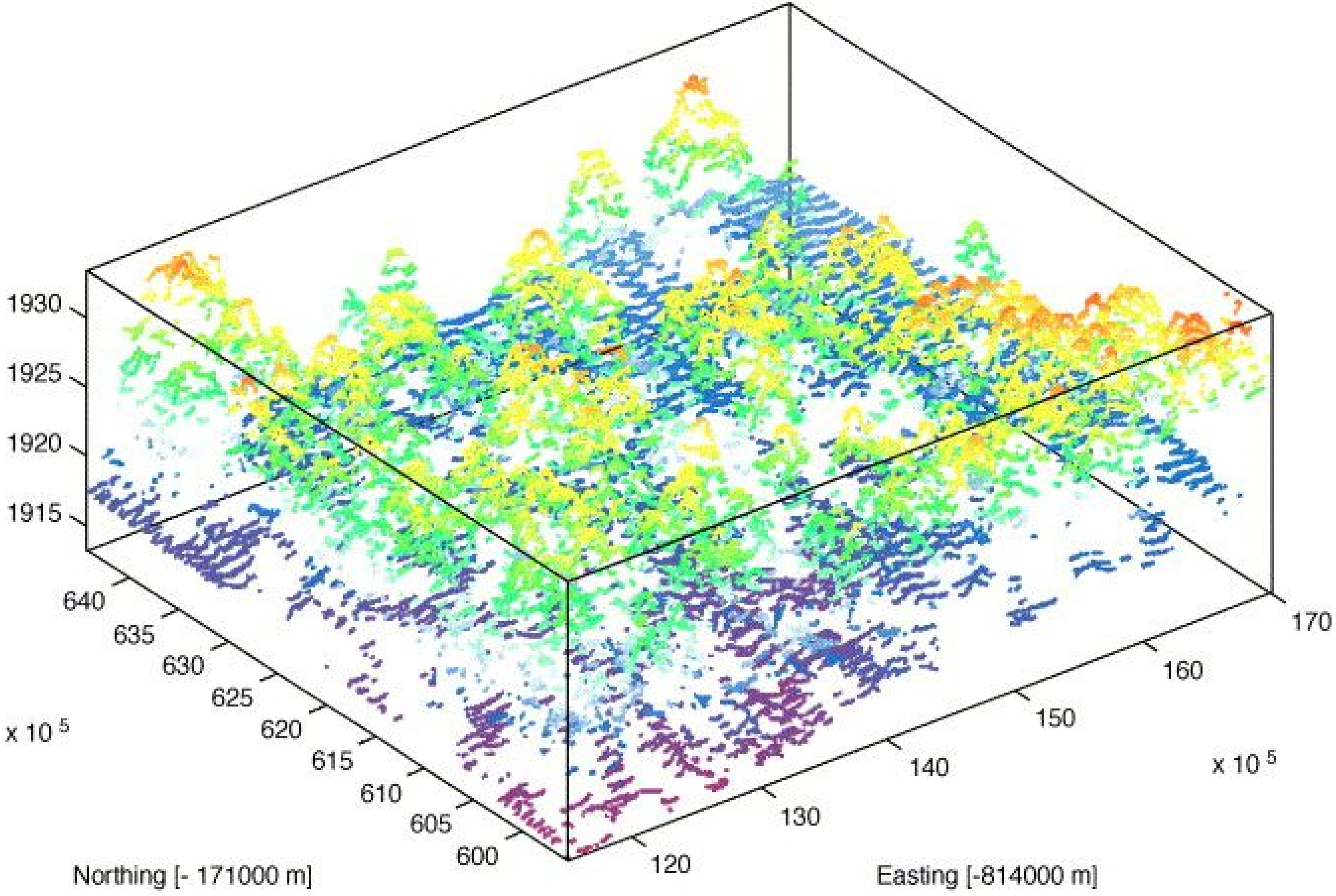
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
employe	applic	action	cadmaz	cfr	amend	anim
fmla	claim	affirm	consult	contain	bankruptci	commod
leav	file	american	copyright	cosmet	code	cpg
	invent	discrimin	custom	ey	court	except
	patent	job	design	hair	creditor	fat
	provision	minor	manag	ingredi	debtor	fe
		opportun	project	label	petition	food
		peopl	sect	manufactur	properti	fruit
		women	servic	product	section	level
				regul	secur	ppm
					truste	refer
						top
						veget

Figure 34. The seven smallest clusters found in the document set. These are stemmed words.

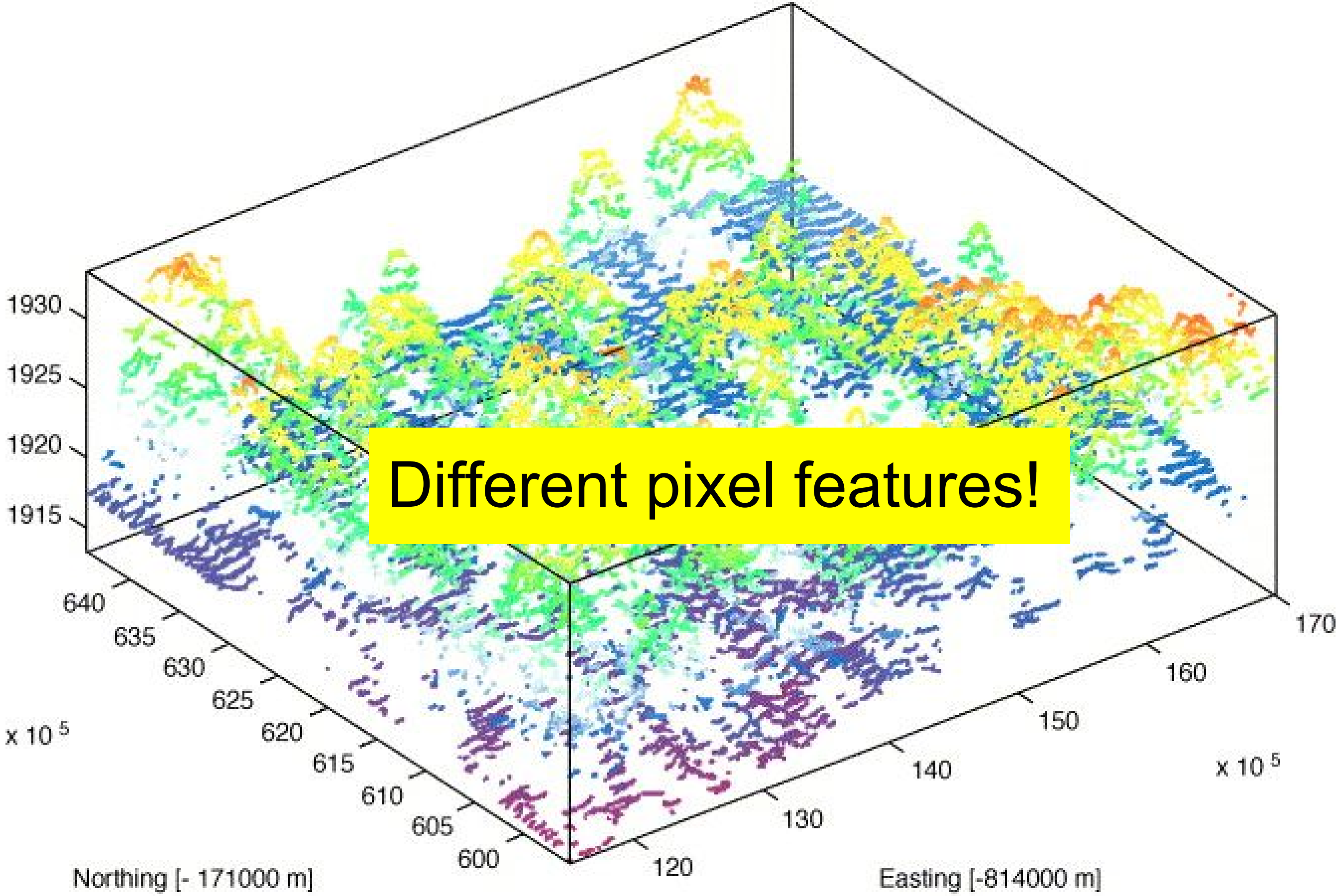
[Jain et al. “Data clustering: a review”, 1999]



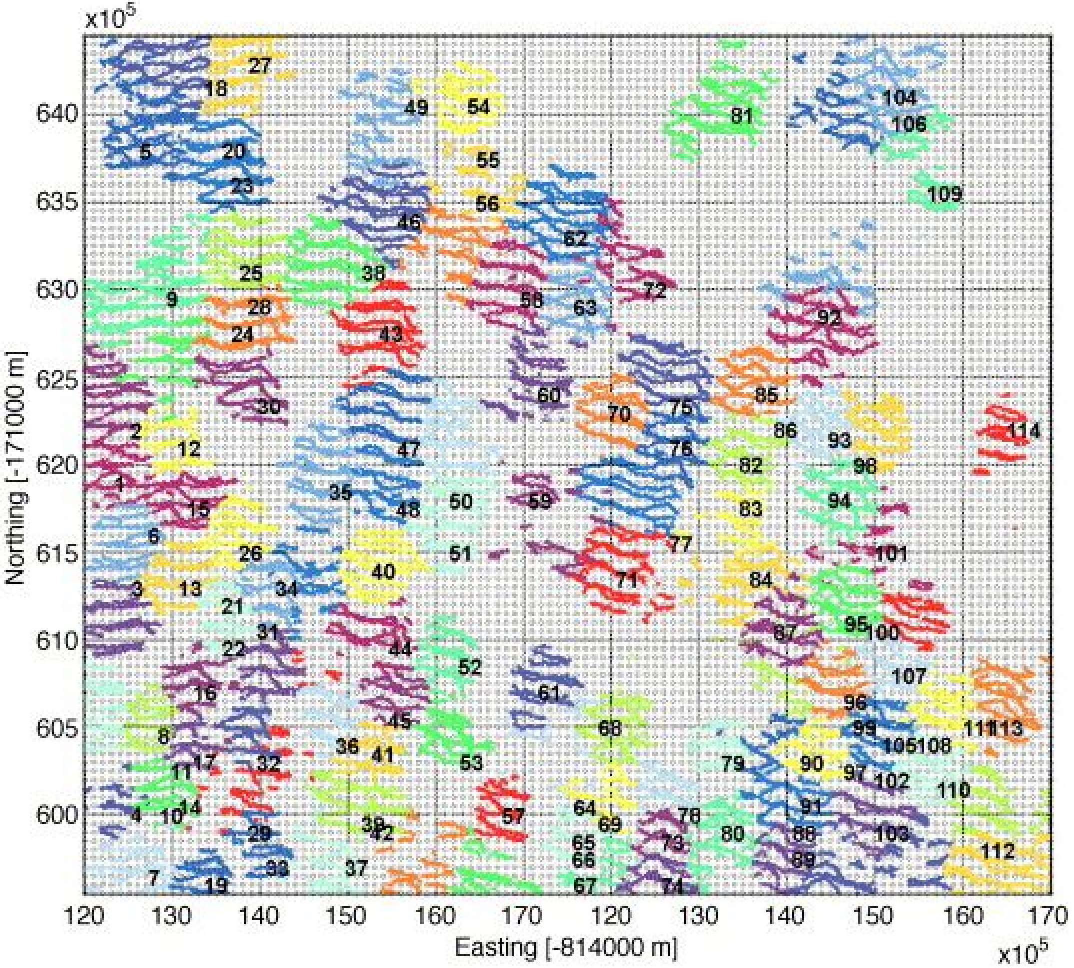
[Morsdorff et al. “LIDAR-based geometric reconstruction…”, 2004]



[Morsdorff et al. “LIDAR-based geometric reconstruction…”, 2004]



[Morsdorff et al. “LIDAR-based geometric reconstruction…”, 2004]



[Morsdorff et al. “LIDAR-based geometric reconstruction…”, 2004]

Agglomerative clustering

- Agglomerative: bottom-up clustering
- Divisive: top-down
- **Dendrograms, cluster similarities, algorithms**

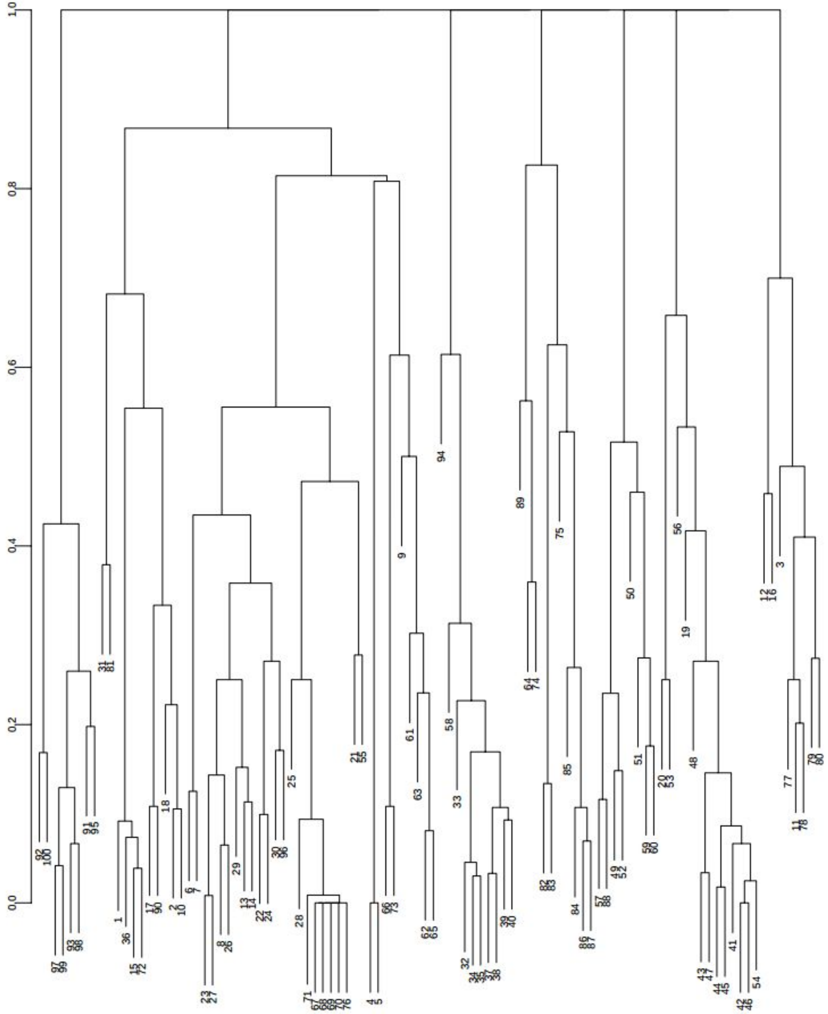
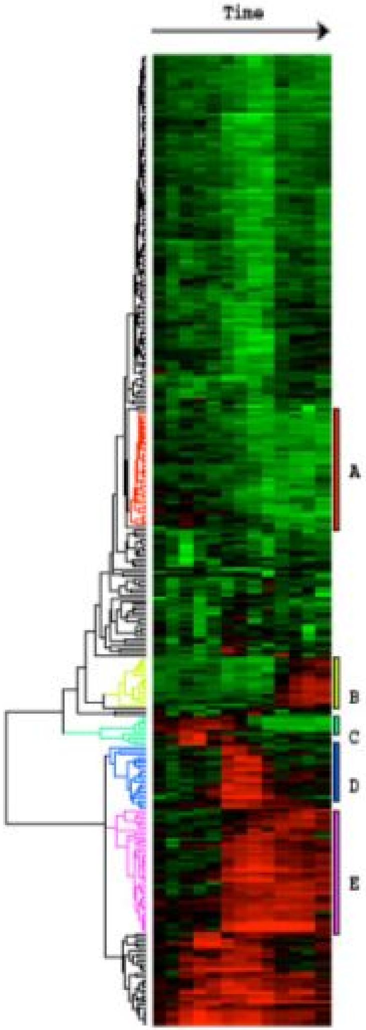
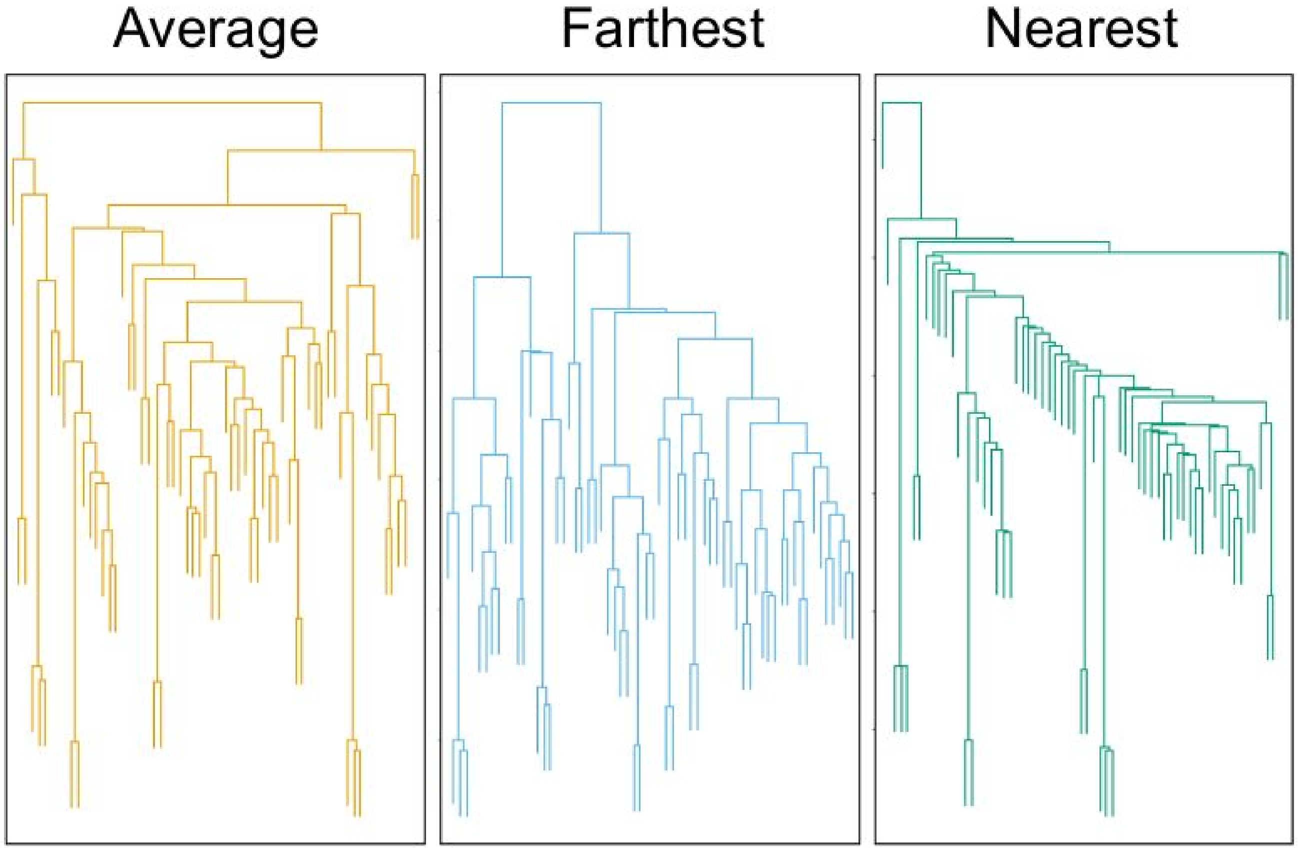


Figure 33. A dendrogram corresponding to 100 books.

Clustering gene expression data



Eisen et al, PNAS 1998



Mouse tumor data from [Hastie *et al.*]

Summary

You should start by asking: what do I want my clustering to do?

This impacts: feature engineering, similarity measure, algorithm selection, visualization!