DPIR SPRING SCHOOL

Classification, clustering, & topic models

14 April 2016

Andy Eggers

Example: Speed ("Do newspapers now give the news?" 1893)

Characterizing content of New York newspapers (based on13 topics) on two Sundays 12 years apart.*

COLUMNS OF READING-MATTER IN NEW YORK NEWSPAPERS, APRIL 17, 1881, AND APRIL 16, 1893.

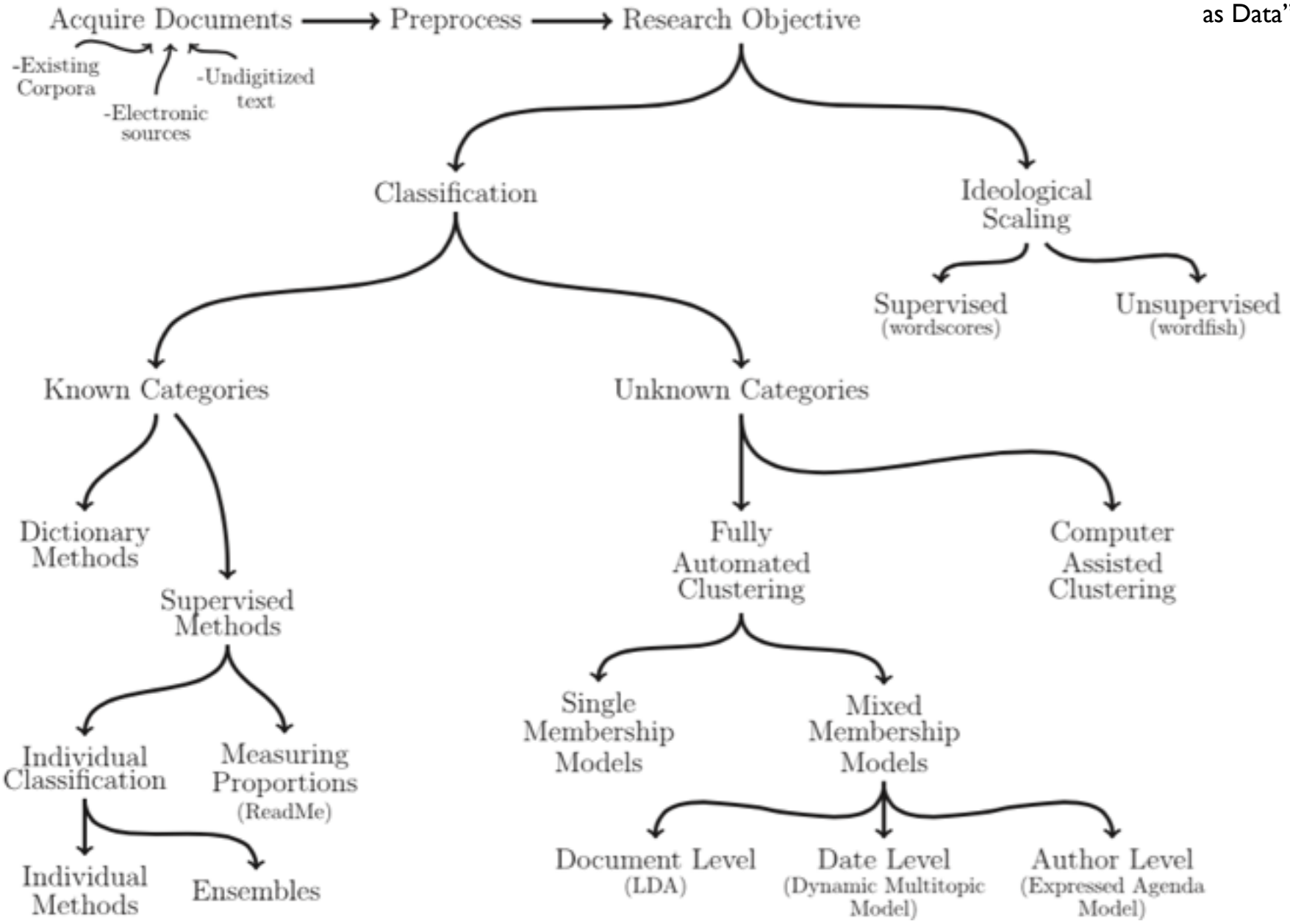
Subject.	Tribune, 1881.	Tribune, 1893.	World, 1881.	World, 1893.	Times, 1881.	Times, 1893.	Sun, 1881.	Sun, 1893.
Editorial	5.00	5.00	4.75	4.00	6.00	5.00	4.00	4.00
Religious	2.00	0.00	0.75	0.00	1.00	0.00	0.50	1.00
Scientific	1.00	0.75	0.00	2.00	1.00	0.00	0.00	2.50
Political	3.00	3.75	0.00	10.50	1.00	4.00	1.00	3.50
Literary	15.00	5.00	1.00	2.00	18.00	12.00	5.75	6.00
Gossip	1.00	23.00	1.00	63.50	.50	16.75	2.00	13.00
Scandals	0.00	1.50	0.00	1.50	1.00	2.50	0.00	2.00
Sporting	1.00	6.50	2.50	16.00	3.00	10.00	0.50	17.50
Fiction	0.00	7.00	1.50	6.50	1.00	1.50	0.00	11.50
Historical	2.50	2.50	2.75	4.00	2.50	1.50	4.25	14.00
Music and Drama	2.50	4.00	1.50	11.00	4.00	7.00	0.00	3.50
Crimes and Criminals.	0.00	0.50	0.00	6.00	0.00	1.00	0.00	0.00
Art	1.00	1.00	3.00	3.00	2.00	0.00	0.25	1.25

Conclusion: "there has been a distinct deterioration and decadence in the New York newspaper press in the last dozen years"

*"I wish to remark here that I selected this date in April merely by chance and not because I was aware of anything in the papers that day making them at all extraordinary."

A classification of "text-as-data" methods

Grimmer and Stewart (2013), "Text as Data"



Key categories of automated classification methods

Unsupervised learning

Classification with **unknown** categories.

"I don't even know where to start with these documents. Can I at least get a summary of what is being discussed?"

Workflow:

- Acquire and process data
- Run classification algorithm
- Try to interpret results (hard)

Supervised learning

Classification with **known** categories, and some classification done by humans.

"I can't classify all these documents. Can I use my classification of this subset to fill in the rest?"

Workflow:

- Acquire and process data
- Decide on classes
- Classify a subset by hand
- Run classification algorithm
- Check accuracy

See Quinn et al (2010)

Like having a robot clean your basement

Unsupervised learning

Tell robot how many piles you want.

Robot tries to put objects in piles with similar objects.

Supervised learning

You put a sample of items into piles.

Robot tries to organize the rest the same way.

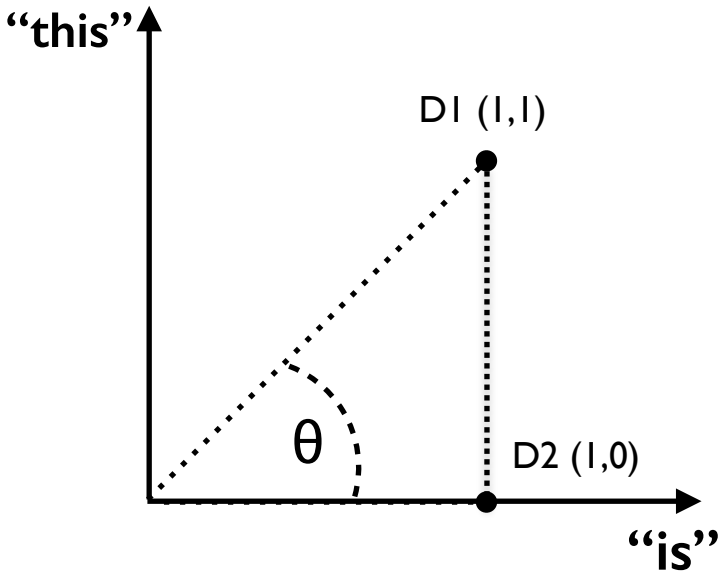
Unsupervised learning: clustering algorithms

How would the robot try to group similar documents together?

First, decide on a measure of similarity (or distance).

Potential measures of similarity/distance between two documents (vectors):

- **Correlation** of column vectors
- **Euclidean distance** between column vectors (in n-dimensional space)
- **Cosine** of angle between column vectors



Term-document matrix

	D1	D2	D3
this	1	1	0
is	1	1	1
a	1	0	0
documen	1	1	0
another	0	1	0
when	0	0	1
lunch	0	0	1

Toy corpus

Document 1: "This is a document."

Document 2: "This is another document."

Document 3: "When is lunch?"

Unsupervised learning: kmeans clustering

Given a measure of similarity/distance, how do we assign documents to groups?

Intuition for k-means clustering:

- **Goal:** Assign documents into k clusters based on similarity
- **Input:** The documents, the number of clusters e.g. k=2
- **Output:** Cluster assignments (e.g. cluster 1: {D1, D2}; cluster 2: {D3})
- **Objective function:** Minimize sum (over documents & terms) of squared distance between document and its cluster's mean location

Augmented TDM (k=2)

	Assign to C1		C1 avg	Assign to C2	C2 avg	Total distance² from cluster means
	D1	D2		D3		
this	1	1	1	0	0	0
is	1	1	1	1	1	0
a	1	0	0.5	0	0	0.5
document	1	1	1	0	0	0
another	0	1	0.5	0	0	0.5
when	0	0	0	1	1	0
lunch	0	0	0	1	1	0
					Sum:	1

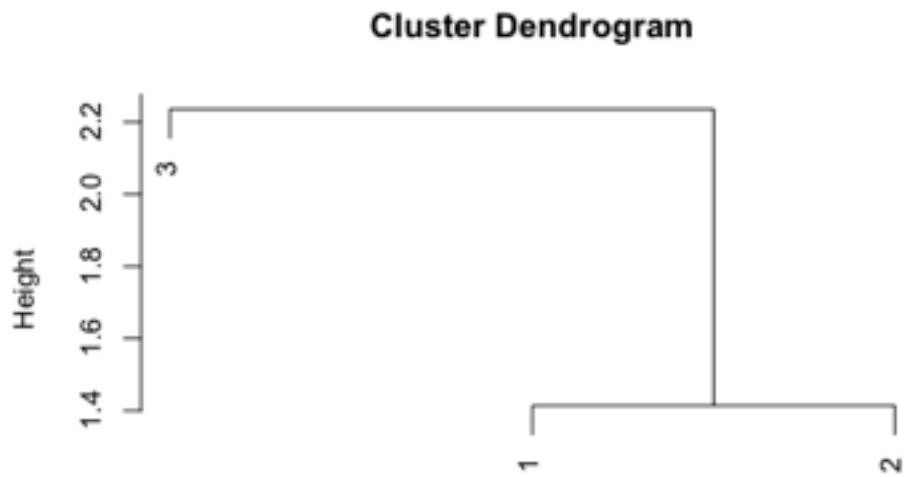
Algorithms do this for us (e.g. kmeans() in R).

Unsupervised learning: hierarchical clustering

Intuition for hierarchical clustering:

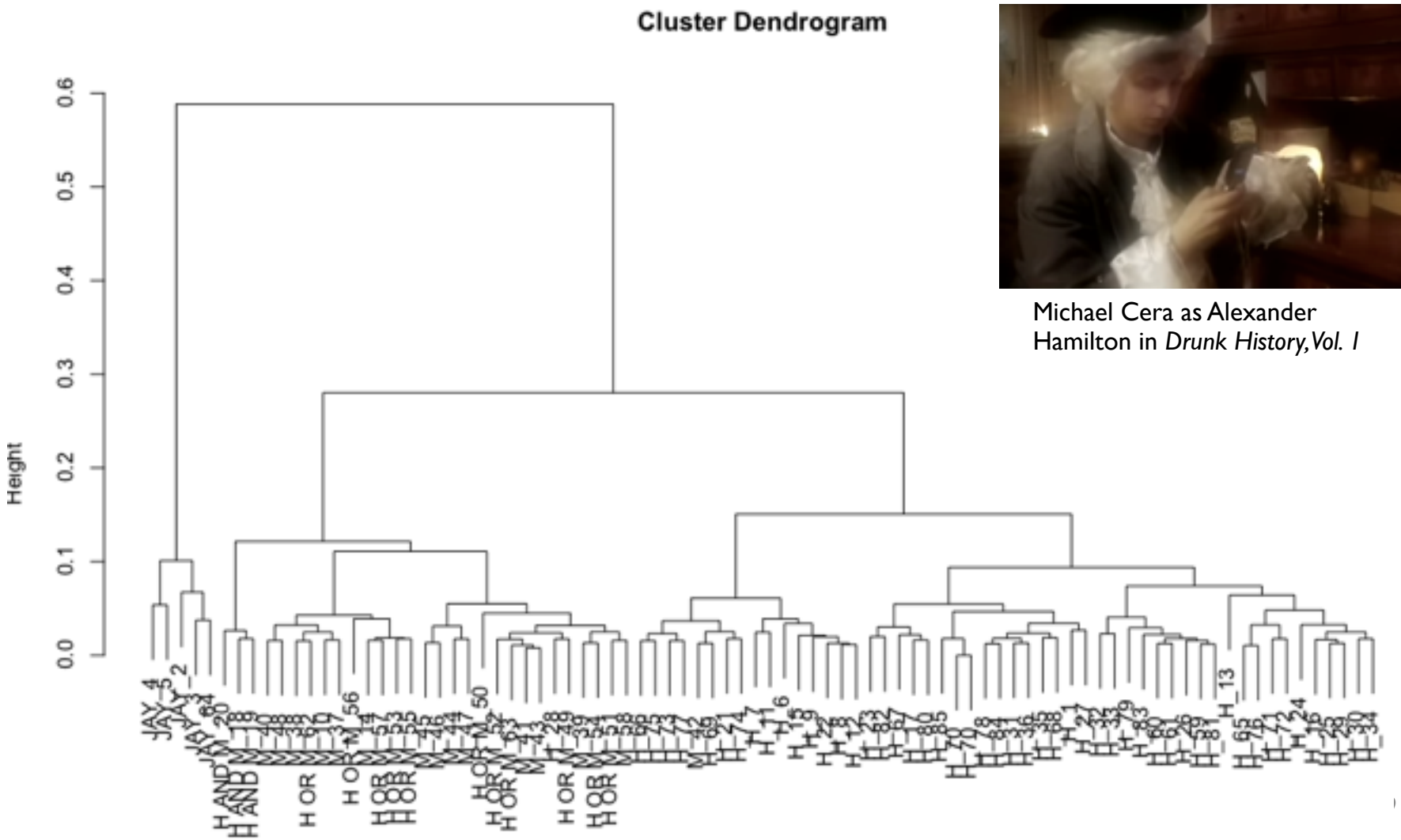
- **Goal:** Assign documents into clusters based on similarity
- **Input:** The distance matrix for the documents
- **Output:** Cluster assignments at each *stage* of the clustering; *cluster dendrogram*
- **Algorithm:** Start with each document in its own cluster. Join the most similar clusters together & recalculate distances. Repeat.

```
> dtm = rbind(c(1,1,1,1,0,0,0), c(1,1,0,1,1,0,0), c(0,1,0,0,0,1,1))
> plot(hclust(dist(dtm)))
```



Unsupervised learning: hierarchical clustering (2)

Hierarchical clustering of *Federalist Papers* based on stop words: solution to an authorship puzzle?



Michael Cera as Alexander Hamilton in *Drunk History, Vol. 1*

Introduction to classification: naive Bayes

Intuition: We have some *labeled* data, with features recorded (word counts).

Then we confront some unlabeled data. Can we use the labeled data to label the unlabeled data?

e.g. spam filter

The Naive Bayes approach labels the unlabeled data based on the word frequencies, which it treats as independent (which is naive).

Introduction to classification: naive Bayes

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})}$$

Bayes Law

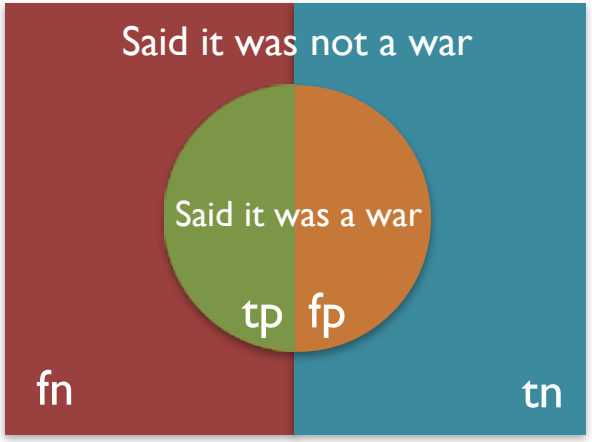
$$\begin{aligned} p(C_k|x_1,\dots,x_n) &\propto p(C_k, x_1,\dots,x_n) \\ &\propto p(C_k)\ p(x_1|C_k)\ p(x_2|C_k)\ p(x_3|C_k)\ \cdots \\ &\propto p(C_k) \prod_{i=1}^{n} p(x_i|C_k)\,. \end{aligned}$$

Naive part

Evaluating a classification model: binary case

Confusion matrix

	Said it was a war	Said it was not a war
Actually war	tp: true positive	fn: false negative
Actually not a war	fp: false positive	tn: true negative



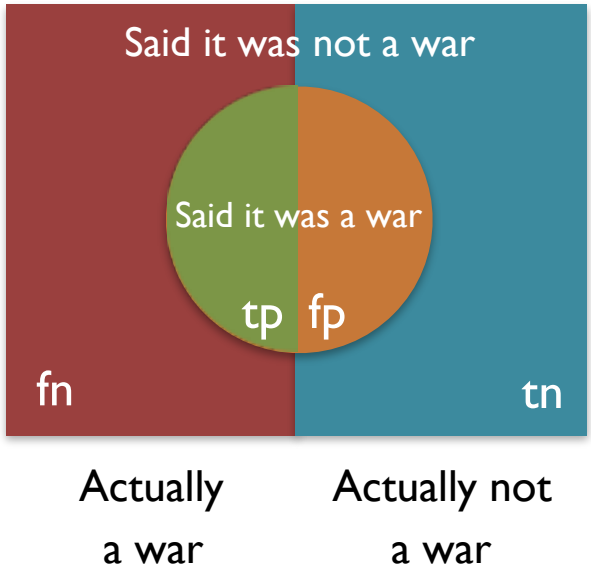
Evaluating a classification model: binary case (2)

Precision: $\frac{tp}{tp + fp}$

Recall/sensitivity: $\frac{tp}{tp + fn}$

Specificity: $\frac{tn}{tn + fp}$

Accuracy: $\frac{tp + tn}{tp + fp + fn + tn}$



Unsupervised learning: model-based approaches

Simplest model-based methods are directly analogous to kmeans clustering: just add statistics (Bayesian/MLE)!

Think of text as having been produced by a data generating process (*generative model*) whose parameters we want to estimate.

- In our usual regressions, parameters are the slope coefficients
- In single-membership topic models, parameters are
  - the word frequencies for each topic
  - the topic membership of each document

Same in kmeans clustering!