Lecture 6: Text Clustering with Unsupervised Learning

Text Re-Use

- Text Re-Use algorithms (like “Smith-Waterman”) measure similarity by finding and counting shared sequences in two texts above some minimum length, e.g. 10 words.
    - useful for plagiarism detection, for example.
- precise but slow
    - shortcut: look at proportion of shared (hashed) 5-grams across texts

Smith-Waterman Algorithm

Smith-Waterman Scoring

D	E	-	S

		-	D	E	S	I	G	N
	-	0	0	0	0	0	0	0
	I	0	0	0	0	5	4	3
D	D	0	5	4	3	4	4	3
E	E	0	4	10	9	8	7	6
A	A	0	3	9	9	8	7	6
S	S	0	2	8	14	13	12	11

Match = +5
Mismatch = -1
Gap = -1

1: DESIGN
2: IDEAS

Aligned:
1: DE-S
2: DEAS

TF-IDF Weighting

- TF/IDF: "Term-Frequency / Inverse-Document-Frequency."
- The formula for word w in document k:

$$\underbrace{\frac{\text{Count of } w \text{ in } k}{\text{Total word count of } k}}_{\text{Term Frequency}} \times \underbrace{\log\left(\frac{\text{Number of documents in } D}{\text{Count of documents containing } w}\right)}_{\text{Inverse Document Frequency}}$$

- The formula up-weights relatively rare words that do not appear in all documents.
  - These words are probably more distinctive of topics or differences between documents.

```
Example: A document contains 100 words, and the word appears 3 times in the document.
The TF is .03. The corpus has 100 documents, and the word appears in 10 documents.
the IDF is 0log(100/10) ≈ 2.3, so the TF-IDF for this document is .03×2.3 = .07.
Say the word appears in 90 out of 100 documents:
Then the IDF is 0.105, with TF-IDF for this document equal to .003.
```

Cosine Similarity

- Representatio nof document i as a vector x_i
  - for example x_i = term counts or x_i = IDF-weighted term frequencies
- Each document is a non-negative vector in an n_x-space, where n_x = vocabulary size
  - documents are rays, and similar documents have similar vectors
- Can measure similarity between documents i and j by the cosine of the angle between x_i and x_j:
  - With perfectly collinear documents (that is, $x_i = \alpha x_j, \alpha > 0$), $cos(0) = 1$
  - For orthogonal documents (no words in common), $cos(\pi/2) = 0$

Cosine Similarity

- Cosine similarity is computable as the normalized dot product between the vectors:

$$cos\_sim(x_1, x_2) = \frac{x_1 \cdot x_2}{\|x_1\|\|x_2\|}$$

```
from sklearn.metrics.pairwise import cosine_similarity
# between two vectors:
sim = cosine_similarity(x, y)[0,0]
# between all rows of a matrix:
sims = cosine_similarity(X)
```

Cosine Similarity Example

The Three Documents and Similarity Metrics



Considering only the 3 words from the above documents: 'sachin', 'dhoni', 'cricket'

Doc Sachin: Wiki page on Sachin Tendulkar
Dhoni - 10
Cricket - 50
Sachin - 200

Doc Dhoni: Wiki page on Dhoni
Dhoni - 400
Cricket - 100
Sachin - 20

Doc Dhoni_Small: Subsection of wiki on Dhoni
Dhoni - 10
Cricket - 5
Sachin - 1

Document - Term Matrix (Word Counts)

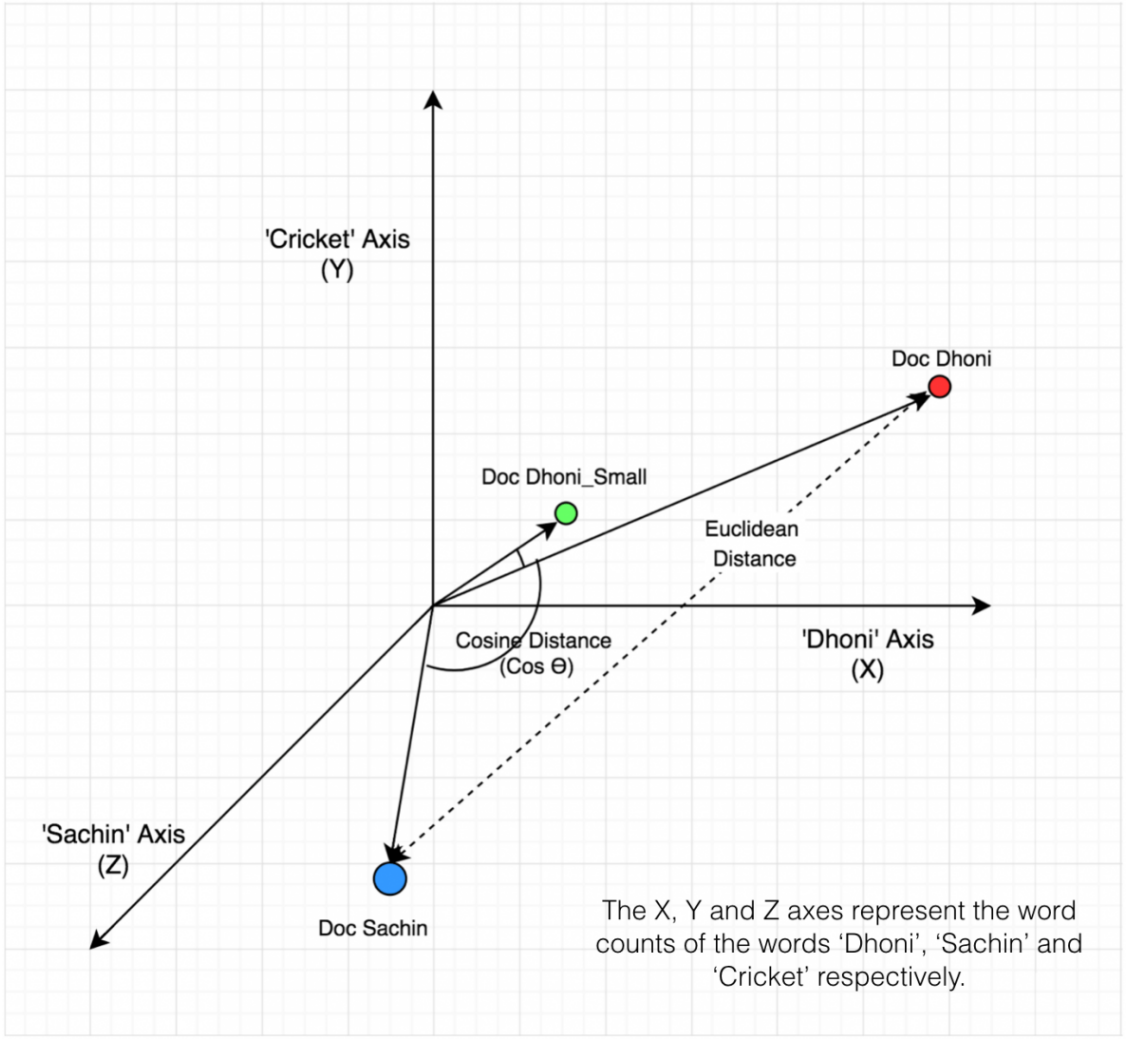
Word Counts	*"Dhoni"*	*"Cricket"*	*"Sachin"*
Doc Sachin	10	50	200
Doc Dhoni	400	100	20
Doc Dhoni_Small	10	5	1

Similarity Metrics

Similarity or Distance Metrics	*Total Common Words*	*Euclidean distance*	*Cosine Similarity*
Doc Sachin & Doc Dhoni	10 + 50 + 10 = **70**	**432.4**	**0.15**
Doc Dhoni & Doc Dhoni_Small	20 + 10 + 7 = **37**	**204.0**	**0.23**
Doc Sachin & Doc Dhoni_Small	10 + 10 + 7 = **27**	**401.85**	**0.77**

Cosine Similarity Example

Projection of Documents in 3D Space



Burgess et al, “Legislative Influence Detectors”

The two largest interest group associations: ALEC (on the conservative side) and ALICE (on the liberal side)

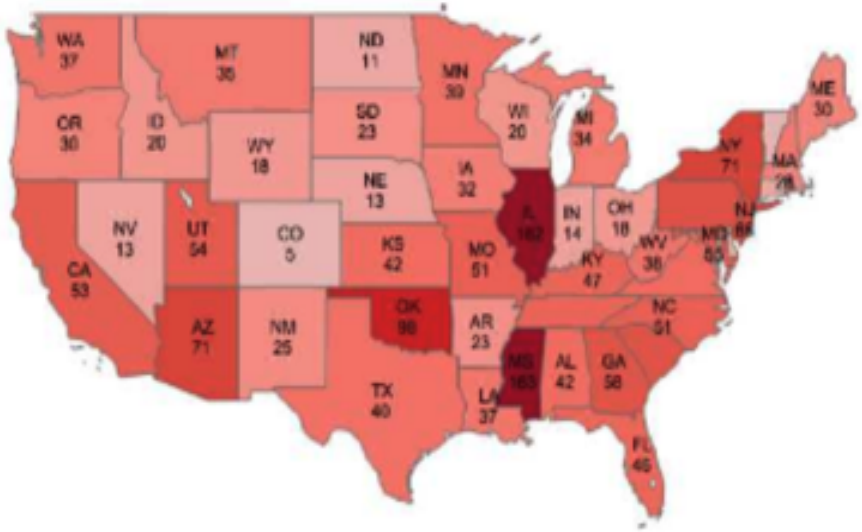


Figure 7: Introduced bills by state from ALEC model legislation

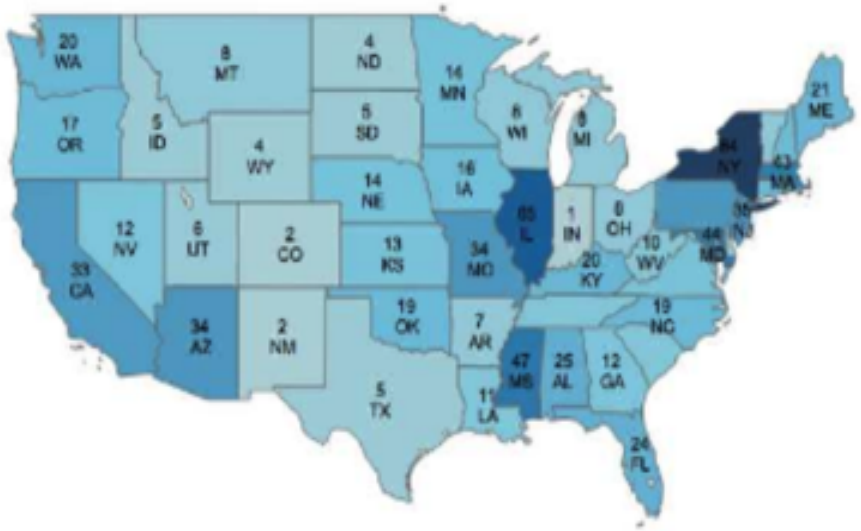
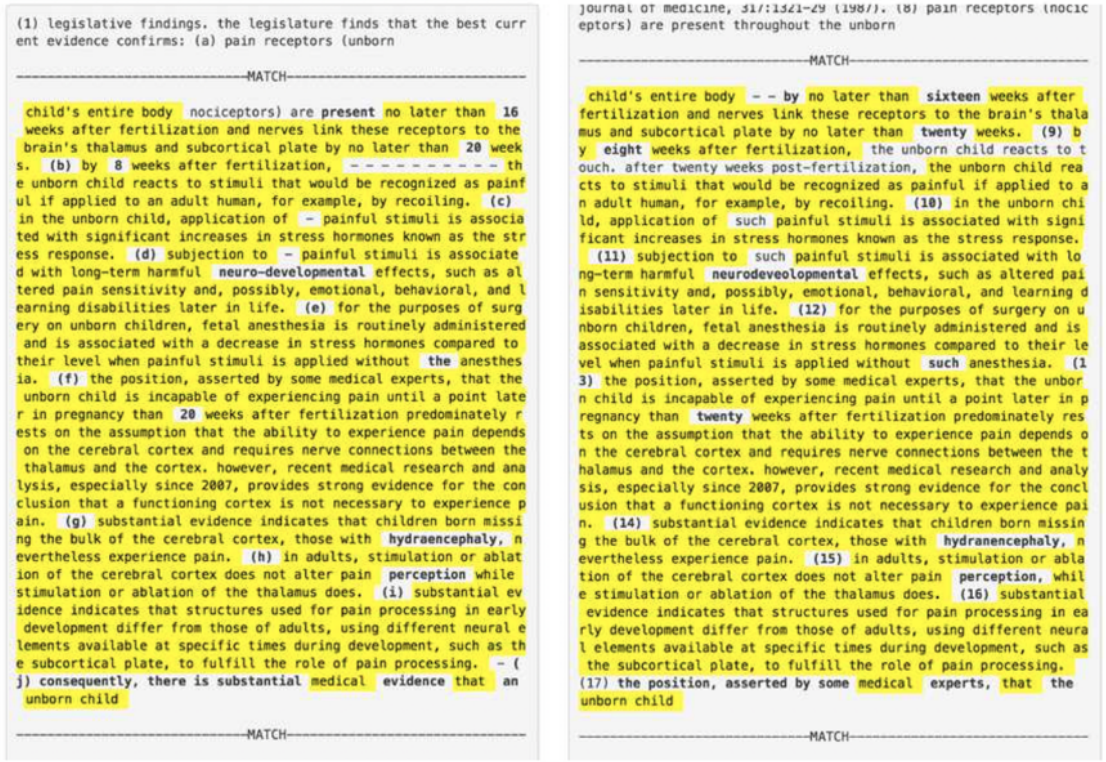


Figure 8: Introduced bills by state from ALICE model legislation

Burgess et al, “Legislative Influence Detectors”



(1) legislative findings. the legislature finds that the best curr
ent evidence confirms: (a) pain receptors (unborn

----------MATCH----------

child's entire body nociceptors) are present no later than 16
weeks after fertilization and nerves link these receptors to the
brain's thalamus and subcortical plate by no later than 20 week
s. (b) by 8 weeks after fertilization, - - - - - - - - - - th
e unborn child reacts to stimuli that would be recognized as painf
ul if applied to an adult human, for example, by recoiling. (c)
in the unborn child, application of - painful stimuli is associa
ted with significant increases in stress hormones known as the str
ess response. (d) subjection to - painful stimuli is associate
d with long-term harmful neuro-developmental effects, such as al
tered pain sensitivity and, possibly, emotional, behavioral, and l
earning disabilities later in life. (e) for the purposes of surg
ery on unborn children, fetal anesthesia is routinely administered
and is associated with a decrease in stress hormones compared to
their level when painful stimuli is applied without the anesthes
ia. (f) the position, asserted by some medical experts, that the
unborn child is incapable of experiencing pain until a point late
r in pregnancy than 20 weeks after fertilization predominately r
ests on the assumption that the ability to experience pain depends
on the cerebral cortex and requires nerve connections between the
thalamus and the cortex. however, recent medical research and ana
lysis, especially since 2007, provides strong evidence for the con
clusion that a functioning cortex is not necessary to experience p
ain. (g) substantial evidence indicates that children born missi
ng the bulk of the cerebral cortex, those with hydraencephaly, n
evertheless experience pain. (h) in adults, stimulation or ablat
ion of the cerebral cortex does not alter pain perception while
stimulation or ablation of the thalamus does. (i) substantial ev
idence indicates that structures used for pain processing in early
development differ from those of adults, using different neural e
lements available at specific times during development, such as th
e subcortical plate, to fulfill the role of pain processing. - (
j) consequently, there is substantial medical evidence that an
unborn child

----------MATCH----------

journal of medicine, 317:1321-29 (1987). (8) pain receptors (nocic
eptors) are present throughout the unborn

----------MATCH----------

child's entire body - - by no later than sixteen weeks after
fertilization and nerves link these receptors to the brain's thala
mus and subcortical plate by no later than twenty weeks. (9) b
y eight weeks after fertilization, the unborn child reacts to t
ouch. after twenty weeks post-fertilization, the unborn child rea
cts to stimuli that would be recognized as painful if applied to a
n adult human, for example, by recoiling. (10) in the unborn chi
ld, application of such painful stimuli is associated with signi
ficant increases in stress hormones known as the stress response.
(11) subjection to such painful stimuli is associated with lo
ng-term harmful neurodeveolopmental effects, such as altered pai
n sensitivity and, possibly, emotional, behavioral, and learning d
isabilities later in life. (12) for the purposes of surgery on u
nborn children, fetal anesthesia is routinely administered and is
associated with a decrease in stress hormones compared to their le
vel when painful stimuli is applied without such anesthesia. (1
3) the position, asserted by some medical experts, that the unbor
n child is incapable of experiencing pain until a point later in p
regnancy than twenty weeks after fertilization predominately res
ts on the assumption that the ability to experience pain depends o
n the cerebral cortex and requires nerve connections between the t
halamus and the cortex. however, recent medical research and analy
sis, especially since 2007, provides strong evidence for the concl
usion that a functioning cortex is not necessary to experience pai
n. (14) substantial evidence indicates that children born missin
g the bulk of the cerebral cortex, those with hydranencephaly, n
evertheless experience pain. (15) in adults, stimulation or abla
tion of the cerebral cortex does not alter pain perception, whil
e stimulation or ablation of the thalamus does. (16) substantial
evidence indicates that structures used for pain processing in ea
rly development differ from those of adults, using different neura
l elements available at specific times during development, such as
the subcortical plate, to fulfill the role of pain processing.
(17) the position, asserted by some medical experts, that the
unborn child

----------MATCH----------

Figure 10: Match between Scott Walker’s bill and a highly similar bill from Louisiana. For a detailed view, please visit http://dssg.uchicago.edu/lid/.

Notes on Cosine Similarity

- For a corpus with n rows, the pairwise similarities give $n \times (n-1)$ similarity scores.
- tf-idf down-weights terms that appear in many documents, usually gives better results.

> Alternative distance metrics:

- dot product (sensitive to document length)
- Euclidean distance, $||v_1 - v_2||$
- Jensen-Shannon Divergence
- etc.

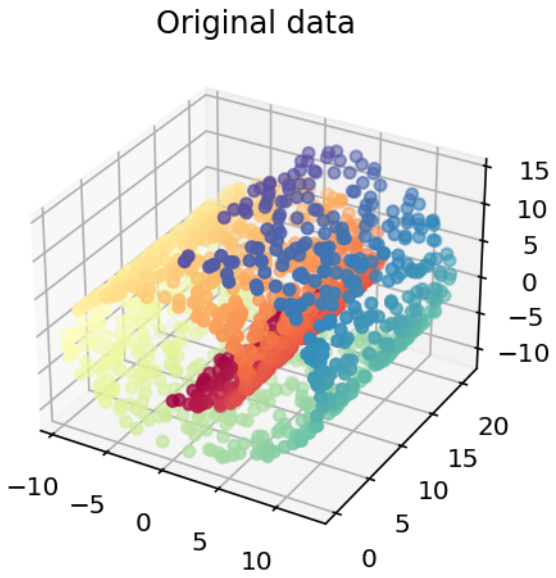
Dimensionality Reduction

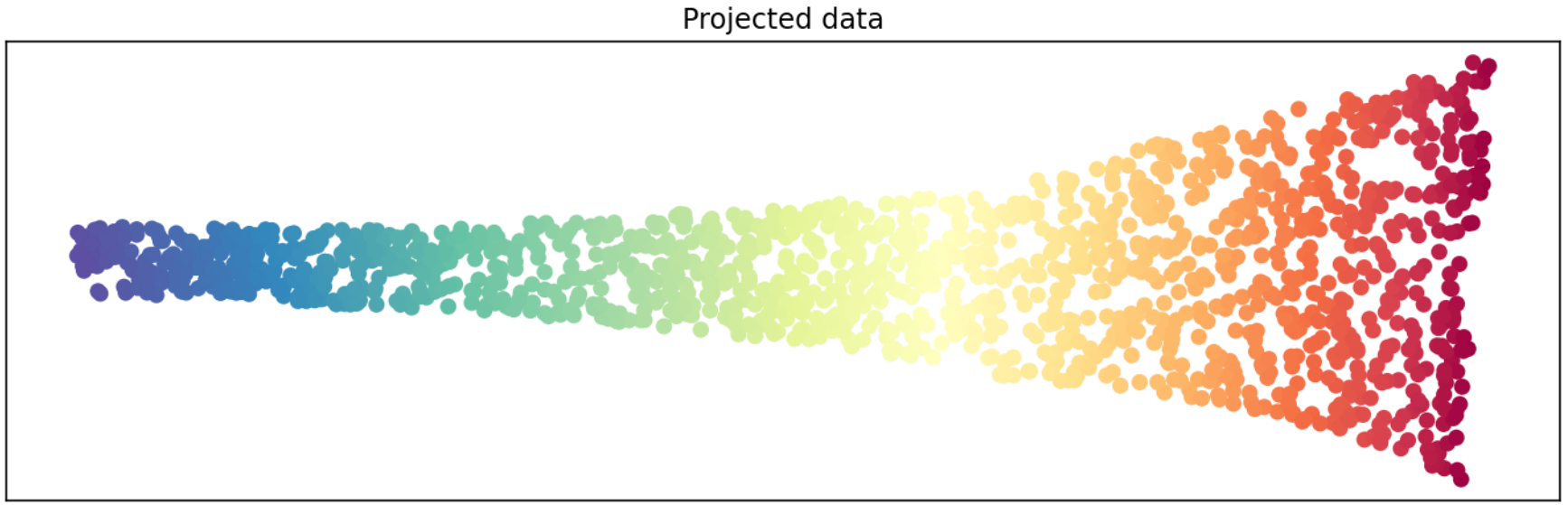
- Datasets are not distributed uniformly across the feature space.
- They have a lower-dimensional latent structure – a manifold – that can be learned.
- `Dimensonality reduction` makes data more interpretable – for example by projecting down to two dimensions for visualization.
- improves computational tractability.
- can improve model performance.

Swiss Roll reduction with LLE

```
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D

Axes3D
# ----------------------------------------------------------------------
# Locally linear embedding of the swiss roll

from sklearn import manifold, datasets

X, color = datasets.make_swiss_roll(n_samples=1500)

print("Computing LLE embedding")
X_r, err = manifold.locally_linear_embedding(X, n_neighbors=12, n_components=2)
print("Done. Reconstruction error: %g" % err)

# ----------------------------------------------------------------------
# Plot result

fig = plt.figure()

ax = fig.add_subplot(211, projection="3d")
ax.scatter(X[:, 0], X[:, 1], X[:, 2], c=color, cmap=plt.cm.Spectral)

ax.set_title("Original data")
ax = fig.add_subplot(212)
ax.scatter(X_r[:, 0], X_r[:, 1], c=color, cmap=plt.cm.Spectral)
plt.axis("tight")
plt.xticks([]), plt.yticks([])
plt.title("Projected data")
plt.show()
```

The Swiss Roll



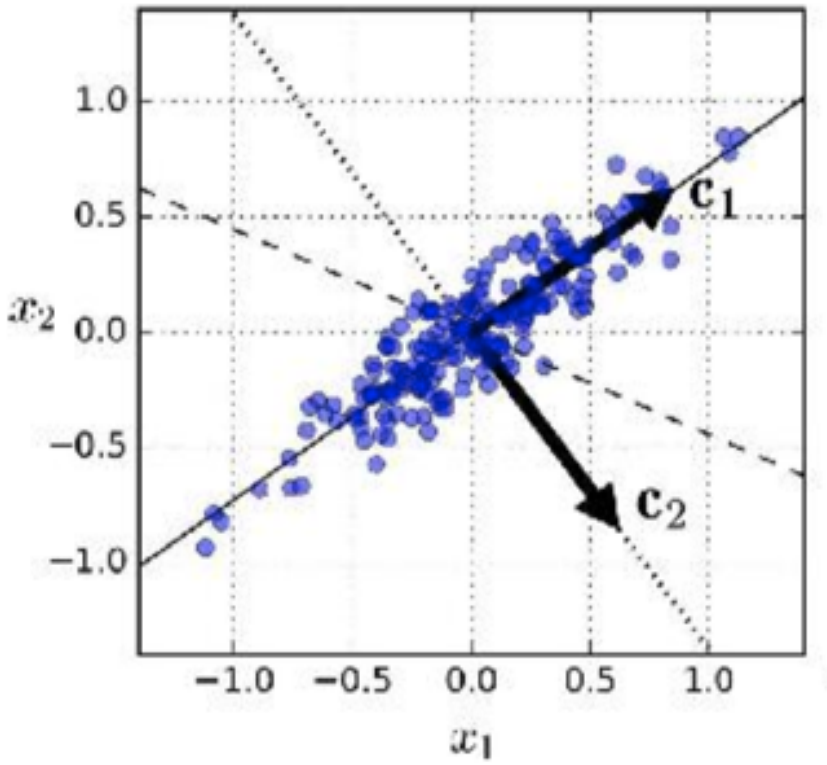


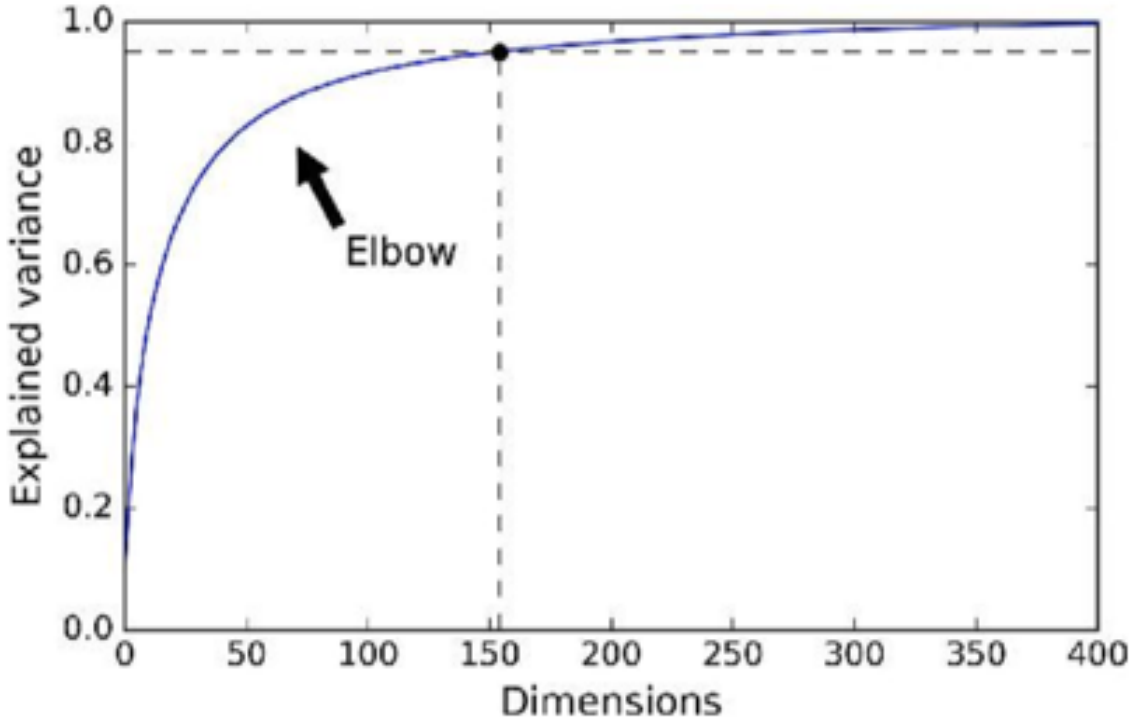
PCA (principal component analysis) / SVD (singular value decomposition)

- PCA computes the dimension in data explaining most variance.
- after the first component, subsequent components learn the (orthogonal) dimensions explaining most variance in dataset after projecting out first component.

```
from sklearn.decomposition import PCA
pca = PCA(n_components=10)
X_train_pca = pca.fit_transform(X_train)
```

PCA (principal component analysis) / SVD (singular value decomposition)





PCA/NMF for Dimension Reduction

Data can be reduced by projecting down to first principal component dimensions.

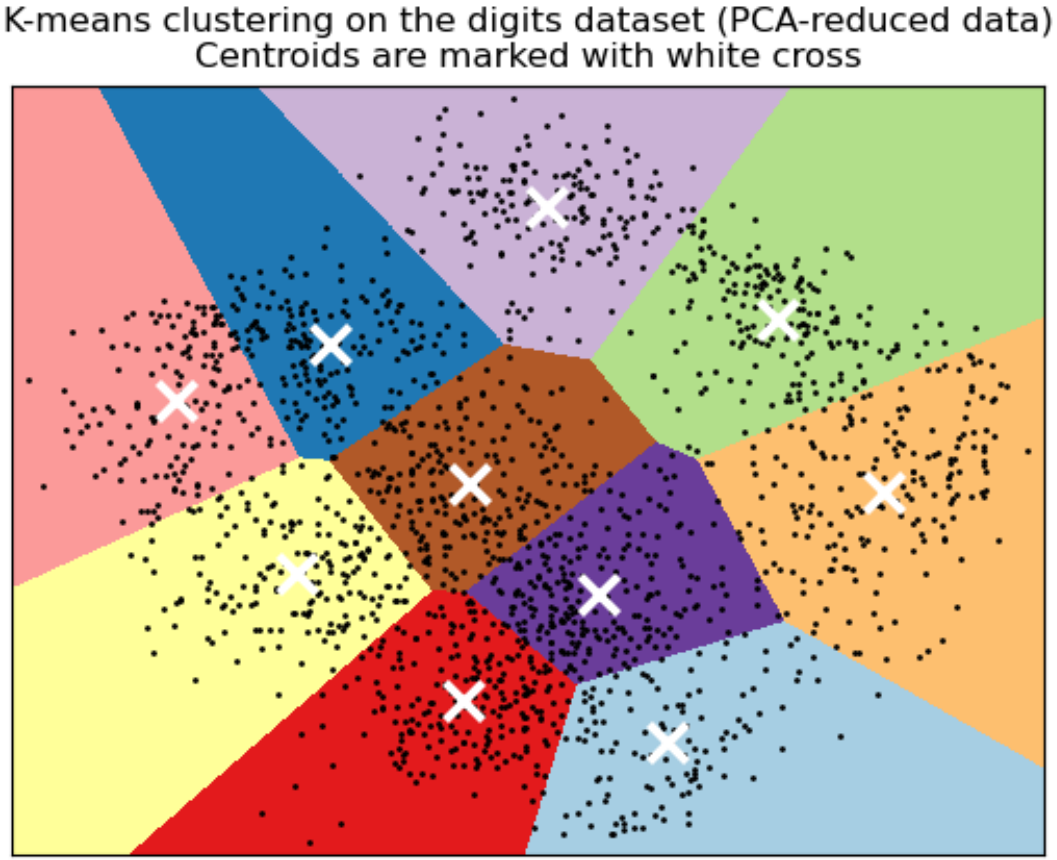
- Distance metrics between observations (e.g. cosine similarity) are approximately preserved.
- For supervised learning, reduced matrix be used as predictors instead of the original matrix.
    - but might destroy (a lot of) predictive information in your dataset.
    - compromise: use feature selection to keep strong predictors, and take principal components of weak predictors.
- PCA dimensions are not interpretable.
- For non-negative data (e.g. counts or frequencies), Non-negative Matrix Factorization (NMF) provides more interpretable factors than PCA.

k-means clustering

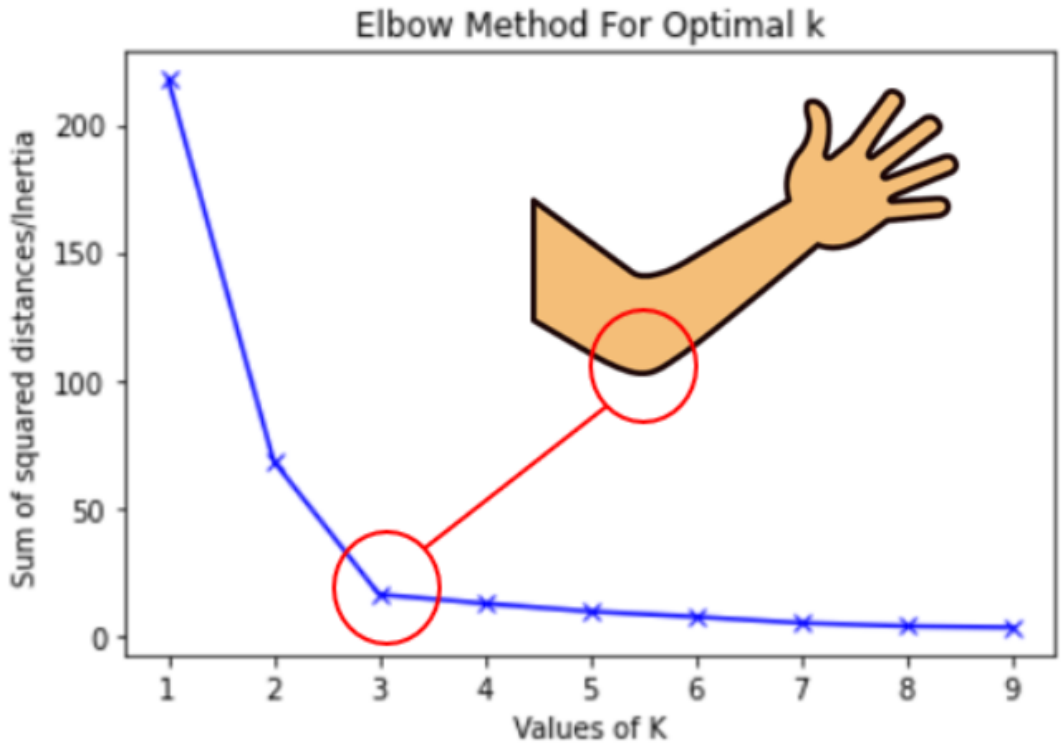
- Matrix of predictors treated as a Euclidean space (should standardize all columns)
- algorithm: initialize cluster centroids randomly, then shift around to minimize sum of within-cluster squared distance

```
from sklearn.cluster import Kmeans
kmeans = KMeans(n_clusters=10)
kmeans.fit(X)
assigned_cluster = kmeans.labels_
```

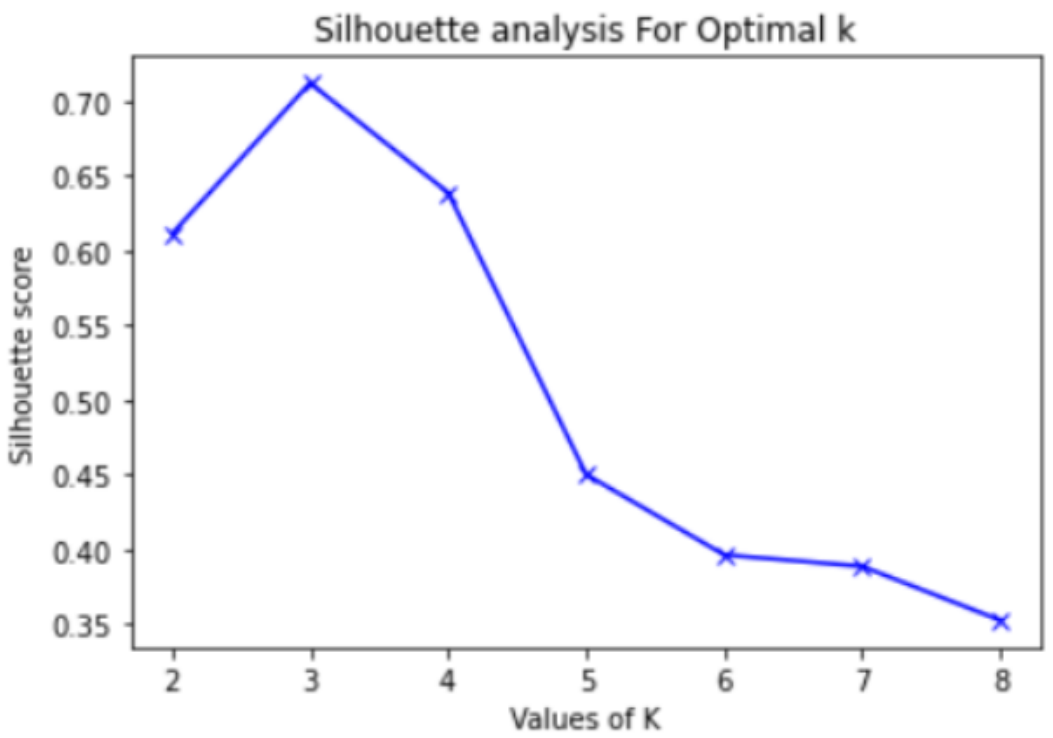
k-means clustering



Selecting k (number of clusters)



Line plot between K and inertia

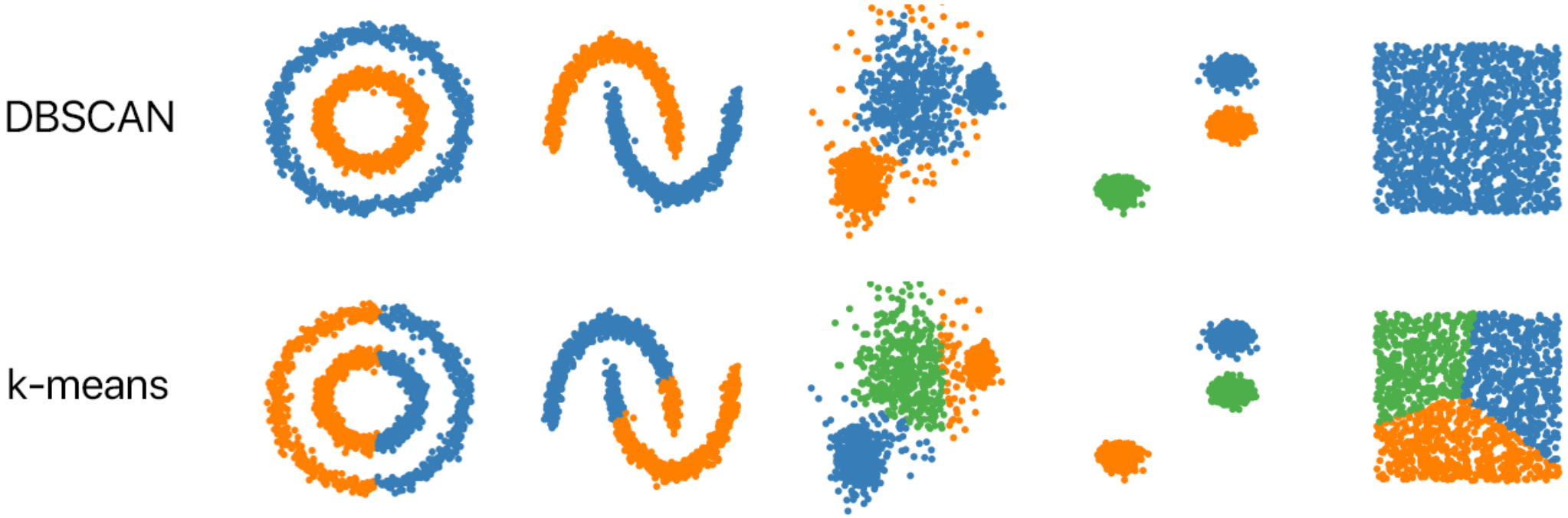


Line plot between K and Silhouette score

Other clustering algorithms

- “k-medoid” clustering use L1 distance rather than Euclidean distance; produces the “medoid” (median vector) for each cluster rather than “centroid” (mean vector).
  - less sensitive to outliers, and medoid can be used as representative data point.
- DBSCAN defines clusters as continous regions of high density.
  - detects and excludes outliers automatically
- Agglomerative (hierarchical) clustering makes nested clusters.

DBSCAN vs k-means



Applications

Hoberg and Phillips, "Text-Based Network Industries and Endogenous Product Differentiation"

- "business description" section from annual regulatory filings, preprocessed by extracting nouns, drop words appearing in more than 25% of documents.
- vector represetnation: binary for whether word appears (rather than counts)
- clusters of these vectors are "industries" – sets of firms with similar lists of nouns in their business descriptions.

Topic Models

- summarize unstructured text
- use words within document to infer subject
- useful for dimension reduction
- topic models are more interpretable than other dimension reduction methods, such as PCA.

Latent Dirichlet Allocation (LDA)

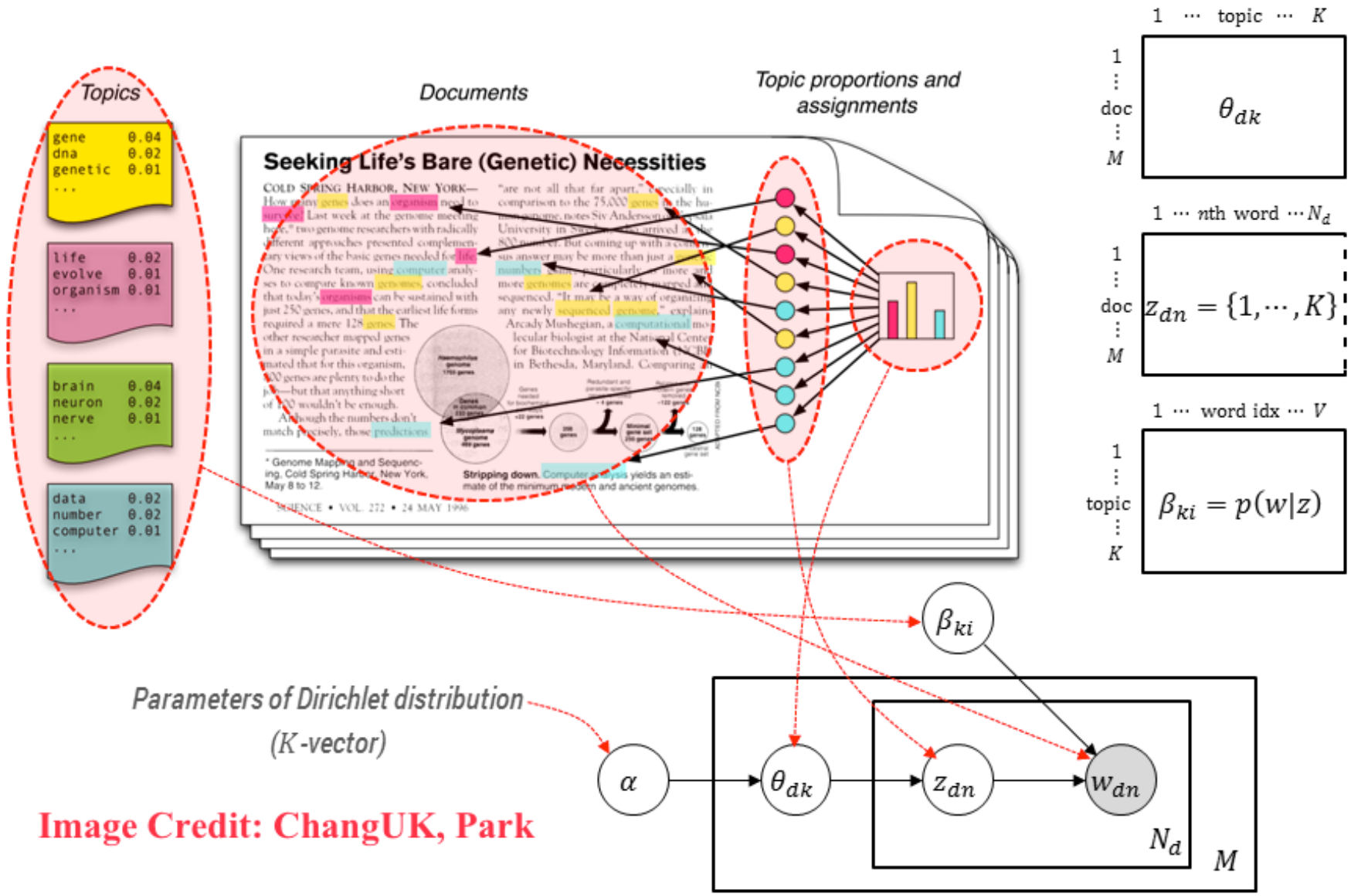


Image Credit: ChangUK, Park

Latent Dirichlet Allocation (LDA)

`Latent` means hidden, `Dirichlet` is a type of probability distribution. Latent Dirichlet Allocation means that we are trying to find all the probability distributions and they are hidden.

Latent Dirichlet Allocation (LDA):

- Each topic is a distribution over words.
- Each document is a distribution over topics.

Input: $N \times M$ document-term count matrix X

Assume: there are K topics (tunable hyperparameter, use coherence).

Like PCA or NMF, LDA works by factorizing X into:

- an $N \times K$ document-topic matrix
- an $K \times M$ topic-term matrix

Latent Dirichlet Allocation

As humans, we know exactly what each question is about.

question	question	question	question
How Do Football Players Stay Safe?	What is the most hated NFL football team of all time	Who is the greatest political leader in the world and why?	Why do people treat politics like it's a football team or some kind of sport?

The 1st question is about football, the 3rd question is about politics, and the 4th question is about politics and football.

Latent Dirichlet Allocation

When we fit these 4 questions to LDA, it will give us back something like this:

question	question	question	question
How Do Football Players Stay Safe?	What is the most hated NFL football team of all time	Who is the greatest political leader in the world and why?	Why do people treat politics like it's a football team or some kind of sport?
100% Topic A	90% Topic A 10% Topic C	100% Topic B	40% Topic A 60% Topic B

The 1st question is 100% of Topic A, the 3rd question is 100% of Topic B, and the last question is split of Topic A and Topic B.

Latent Dirichlet Allocation

The word “football” has the highest weight in Topic A, followed by “NFL” followed by “player”. So we could infer that this topic is about sport.

The word “politics” has the highest weight in Topic B, followed by “leader”, followed by “world”. So we could infer that this topic is about politics. As shown below:

- Topic A: 40% football, 30% NFL, 10% player ... Sport
- Topic B: 30% political, 20% leader, 10% world ... Politics

Latent Dirichlet Allocation

Then we go back to our original questions, here are the topics!

question	question	question	question
How Do Football Players Stay Safe?	What is the most hated NFL football team of all time	Who is the greatest political leader in the world and why?	Why do people treat politics like it's a football team or some kind of sport?
Sport	Sport	Politics	Sport + Politics

Latent Dirichlet Allocation

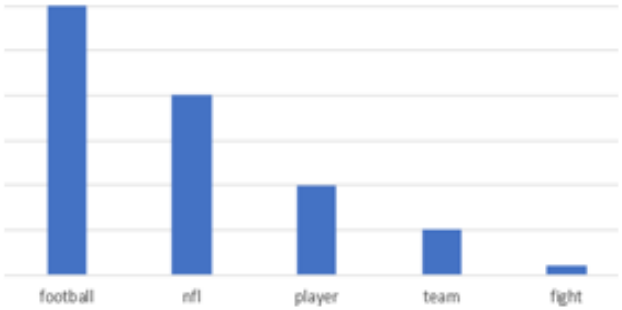
To dive a little deeper, every questions is a mix of topics, and every topic is a mix of words.

Every **question** consists of a mix of **topics**

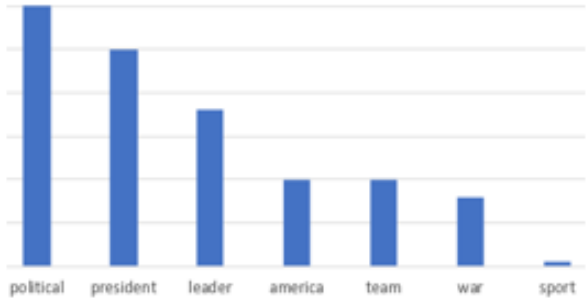
question	question	question	question
How Do Football Players Stay Safe?	What is the most hated NFL football team of all time	Who is the greatest political leader in the world and why?	Why do people treat politics like it's a football team or some kind of sport?
100% Topic A	90% Topic A 10% Topic C	100% Topic B	40% Topic A 60% Topic B

Every **topic** consists of a mix of **words**

Topic: Sport



Topic: Politics



Latent Dirichlet Allocation

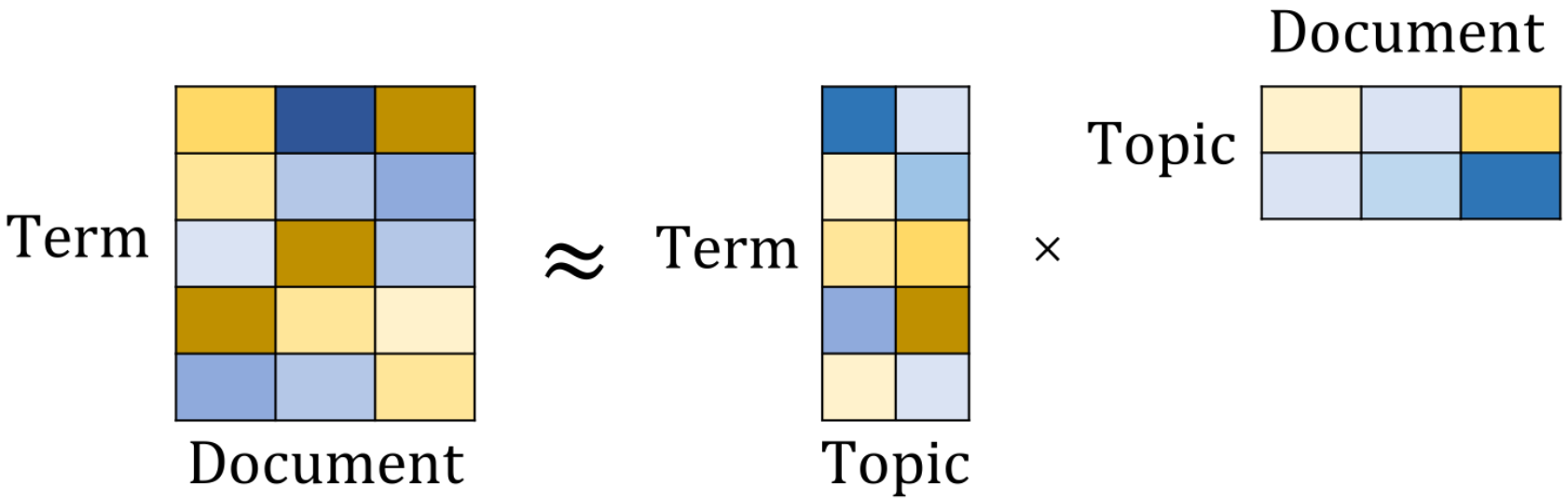
A question is a probability distribution of topics, and every topic is a probability distribution of words.

What LDA does is that when you fit it with all those questions, it is trying its best to find the best topic mix and the best word mix.

Non-negative Matrix Factorization (NMF)

- Family of linear algebra algorithms for identifying the latent structure in data represented as a non-negative matrix.
- NMF can be applied for topic modeling, where the input is term-document matrix, typically TF-IDF normalized.
- Input: Term-Document matrix, number of topics.
- Output: Two non-negative matrices of the original n words by k topics and those same k topics by the m original documents.
- Basically, we are going to use linear algebra for topic modeling.

Non-negative Matrix Factorization (NMF)

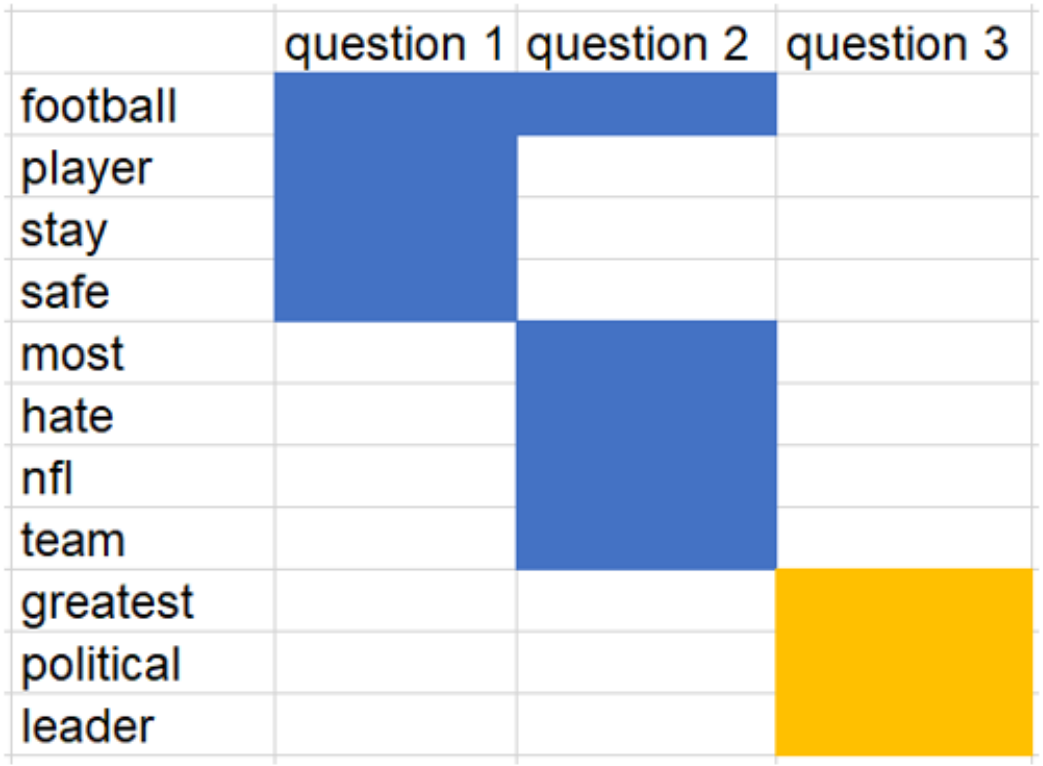


$$X \approx [\ \hat{W} \times \hat{H}\ ]_+$$

We take the term-document matrix, decompose to two matrices, first one has every topic and what terms in it, and 2nd one has every document and what topics in it.

Non-negative Matrix Factorization (NMF)

- How do football players stay safe?
- What is the most hated NFL football team?
- Who is the greatest political leader?



On the left we have 3 questions, on the right we have term-document matrix for these 3 questions. We choose k=2 topics.

Non-negative Matrix Factorization (NMF)

W
Weights for terms

	Topic A	Topic B
football		
player		
stay		
safe		
most		
hate		
nfl		
team		
greatest		
political		
leader		

H
Weights for questions

	question 1	question 2	question 3
Topic A			
Topic B			

After decomposition, we got two non-negative matrices of the original n words by k topics and those same k topics by the m original documents.

Using an LDA Model

Once trained, can easily get topic proportions for a corpus.

- or any document – doesn't have to be in training corpus.
- main topic is the highest-probability topic
- documents with highest share in a topic work as representative documents for the topic.

Can then use the topic proportions as variables in a social science analysis.

- e.g., Catalinac (2016) shows that after a Japanese political reform that reduced intraparty competition, candidate platforms reduced local pork and increased national policy.

From Pork to Policy

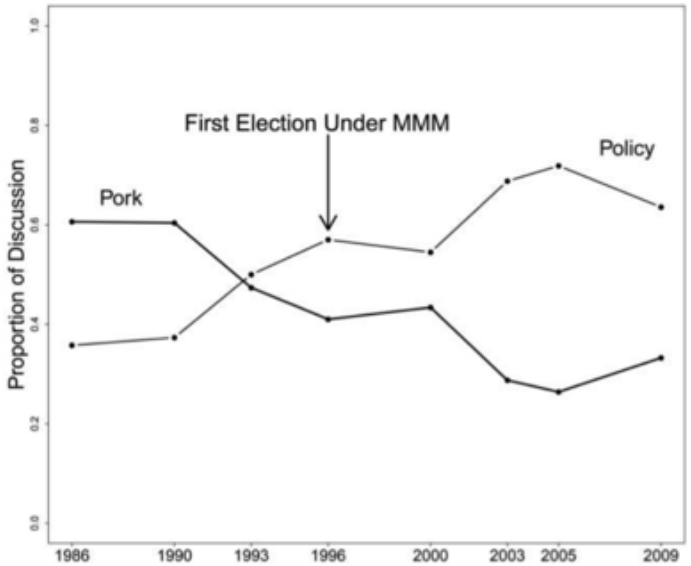


TABLE 1 A Summary of Common Assumptions and Relative Costs Across Different Methods of Discrete Text Categorization

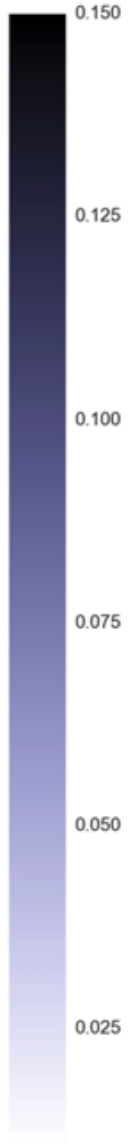
	Method				
A. Assumptions	***Reading***	***Human Coding***	***Dictionaries***	***Supervised Learning***	***Topic Model***
Categories are known	No	Yes	Yes	Yes	No
Category nesting, if any, is known	No	Yes	Yes	Yes	No
Relevant text features are known	No	No	Yes	Yes	Yes
Mapping is known	No	No	Yes	No	No
Coding can be automated	No	No	Yes	Yes	Yes
B. Costs					
Preanalysis Costs					
Person-hours spent conceptualizing	Low	High	High	High	Low
Level of substantive knowledge	Moderate/High	High	High	High	Low
Analysis Costs					
Person hours spent per text	High	High	Low	Low	Low
Level of substantive knowledge	Moderate/High	Moderate	Low	Low	Low
Postanalysis Costs					
Person-hours spent interpreting	High	Low	Low	Low	Moderate
Level of substantive knowledge	High	High	High	High	High

Recommended: read this part of Quinn, Monroe, Colaresi, Crespin, and Radev (2010)

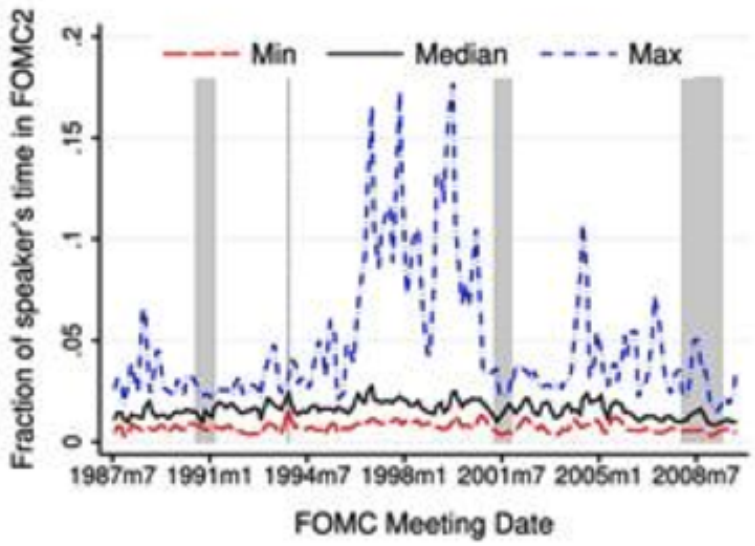
Topic modeling Federal Reserve Bank transcripts

- Analyze speech transcripts from FOMC (Federal Open Market Committee).
  - private discussions among committee members at Federal Reserve (U.S. Central Bank)
  - 150 meetings, 20 years, 26,000 speeches, 24,000 unique words.
- Pre-processing:
  - drop stopwords, stems; vocab = 10,000 words
- LDA:
  - K = 40 topics selected for interpretability / topic coherence.

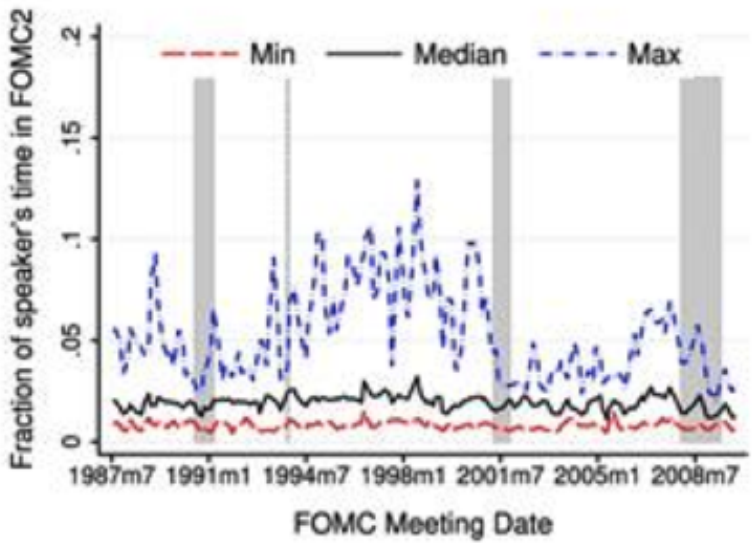
													Pro-cyclicality
Topic0[1]	product	increas	wage	price	cost	labor	rise	acceler	inflat	pressur	trend	compens	0.024
Topic1[1,2]	growth	slow	economi	continu	expans	strong	trend	inflat	will	recent	slowdown	moder	0.023
Topic2[2]	inflat	expect	core	measur	higher	path	slack	gradual	continu	remain	view	suggest	0.017
Topic3[1]	percent	year	quarter	growth	month	rate	last	next	state	averag	california	employ	0.007
Topic4	number	data	look	chang	measur	use	point	show	revis	estim	gdp	actual	0.007
Topic5[1,2]	polici	inflat	monetarpol	need	time	can	monetari	move	tighten	view	action	believ	0.005
Topic6[2]	rate	term	expect	real	lower	increas	rise	level	declin	short	nomin	year	0.005
Topic7	statement	word	chang	meet	languag	discuss	issu	want	read	sentenc	view	use	0.005
Topic8[2]	chairman	support	mr	direct	recommend	agre	asymmetr	prefer	symmetr	move	toward	favor	0.004
Topic9[1]	employ	continu	growth	job	nation	region	seem	state	manufactur	greenbook	busi	bit	0.004
Topic10	dollar	unitedstates	export	countri	import	foreign	japan	growth	abroad	trade	develop	currenc	0.003
Topic11	model	use	simul	shock	effect	scenario	nairu	differ	rule	chang	baselin	altern	0.003
Topic12[2]	risk	may	balanc	seem	side	uncertainti	possibl	economi	probabl	reason	upsid	much	0.003
Topic13	forecast	greenbook	staff	project	differ	assumpt	littl	assum	somewhat	lower	end	period	0.002
Topic14	period	committe	consist	econom	run	maintain	futur	read	slightli	stabil	expect	develop	0.002
Topic15	invest	incom	spend	capit	household	consum	busi	hous	consumpt	sector	stock	stockmarket	0.002
Topic16[1]	month	report	increas	survey	expect	indic	remain	continu	last	recent	data	activ	0.002
Topic17[1]	project	forecast	year	quarter	expect	will	percent	revis	anticip	growth	next	recent	0.002
Topic18	question	ask	issu	let	want	answer	rais	discuss	don	start	without	okay	0.001
Topic19	peopl	talk	lot	much	comment	around	differ	number	realli	look	thing	hear	0.001
Topic20	presid	ye	governor	parri	stern	vice	hoenig	minehan	kelley	jordan	moskow	mcteer	0.001
Topic21	move	can	evid	signific	stage	inde	will	issu	economi	may	quit	clearli	0.001
Topic22[2]	chairman	thank	mr	time	meet	laughter	comment	let	will	point	call	may	0.0
Topic23[1]	year	panel	line	shown	right	chart	expect	project	percent	middl	left	next	0.0
Topic24	district	nation	area	continu	sector	construct	manufactur	report	activ	region	economi	remain	0.0
Topic25	know	someth	happen	right	thing	want	look	sure	can	realli	anyth	els	0.0
Topic26[1,2]	polici	might	committe	market	may	tighten	eas	risk	action	staff	possibl	potenti	-0.001
Topic27	year	continu	product	price	level	industri	will	sale	increas	auto	last	district	-0.001
Topic28[1]	inventori	product	sale	level	order	will	sector	come	good	quarter	much	adjust	-0.001
Topic29	price	oil	increas	energi	effect	import	suppli	product	demand	will	market	oilprices	-0.002
Topic30	term	might	point	can	sens	run	short	probabl	time	longer	tri	someth	-0.002
Topic31	seem	may	time	certainli	bit	littl	quit	much	far	perhap	better	might	-0.003
Topic32	money	aggreg	borrow	seem	rang	reserv	rate	target	time	altern	suggest	million	-0.003
Topic33[2]	move	market	point	will	fundsrate	rate	basispoints	need	fed	today	basi	time	-0.004
Topic34[1]	report	busi	compani	year	contact	firm	sale	worker	expect	plan	director	industri	-0.004
Topic35	will	fiscal	ta	budget	cut	govern	effect	billion	state	spend	deficit	year	-0.005
Topic36	will	economi	world	rather	problem	believ	can	situat	much	seem	view	good	-0.008
Topic37	realli	look	side	thing	lot	problem	concern	littl	pretti	situat	kind	much	-0.012
Topic38	bank	credit	market	loan	financi	debt	lend	fund	concern	financ	problem	spread	-0.018
Topic39[1,2]	economi	weak	recoveri	recess	confid	eas	neg	econom	will	turn	declin	period	-0.059



Pro-Cyclical Topics

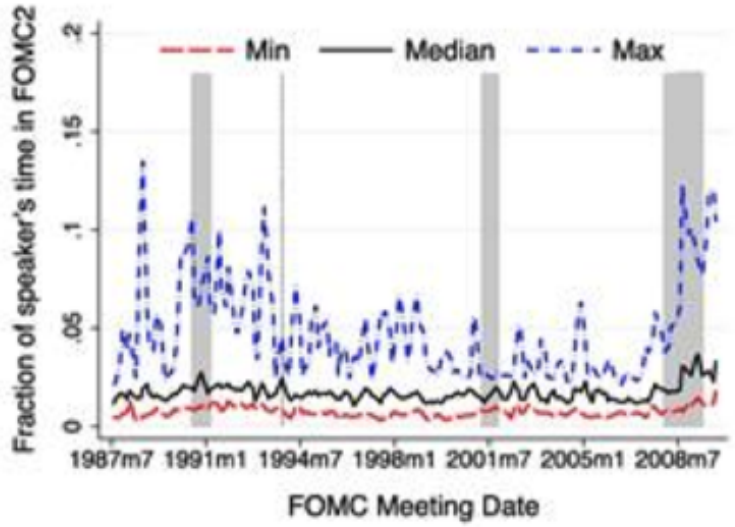


(A) Topic 0 'Productivity'

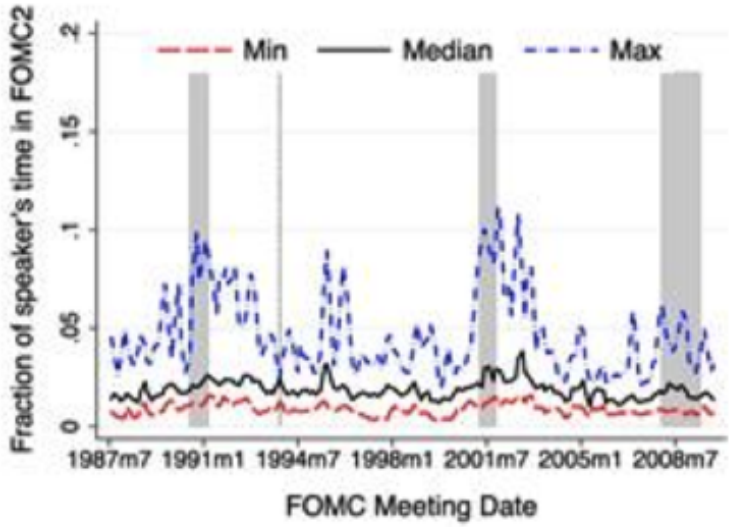


(B) Topic 1 'Growth'

Counter-Cyclical Topics



(A) Topic 38 'Financial sector'



(B) Topic 39 'Economic weakness'

Effect of Transparency

TABLE IV

SUMMARY OF COMMUNICATION MEASURES (MEETING-SECTION-SPEAKER LEVEL)

Count measures		Topic measures	
Name	Description	Name	Description
Words	The count of words spoken	Concentration	The Herfindahl index applied to distribution over policy topics
Statements	The count of statements made	Quant	Percentage of time on data topics
Questions	The count of questions asked	Avg Sim (X) $X \in \{B, D, KL\}$ B = Bhattacharyya D = dot product KL = Kullback − Leibler	The similarity between a speaker's distribution over policy topics and the FOMC average, computed using metric X
Numbers	The count of numbers spoken	Pr (no dissent)	The fitted value for no voiced dissent from the LASSO for policy topic selection (only FOMC2)

Effect of Transparency

- In 1993, there was an unexpected transparency shock where transcripts became public.
- Increasing transparency results in:
  - higher discipline / technocratic language (probably beneficial)
  - higher conformity (probably costly)
- Highlights tradeoffs from transparency in bureaucratic organizations.