14. TOPIC MODELING

LEARNING OBJECTIVES

- Topic Modeling
- Latent Semantic Analysis
- Latent Dirichlet Allocation

REVIEW OF LESSON 13

LAST LESSON REVIEW

- Feature extraction from documents
- Bag of words
- TF-Idf
- CountVectorizer, Tf-Idf Vectorizer, HashingVectorizer
- Scikit Pipeline

LAB REVIEW

Text Classification - Lab 20 mn

LATENT VARIABLE MODELS - TOPIC MODELING

LATENT VARIABLE MODELS

Attempting to uncover structure or organization inherent in the text.

Unsupervised learning techniques

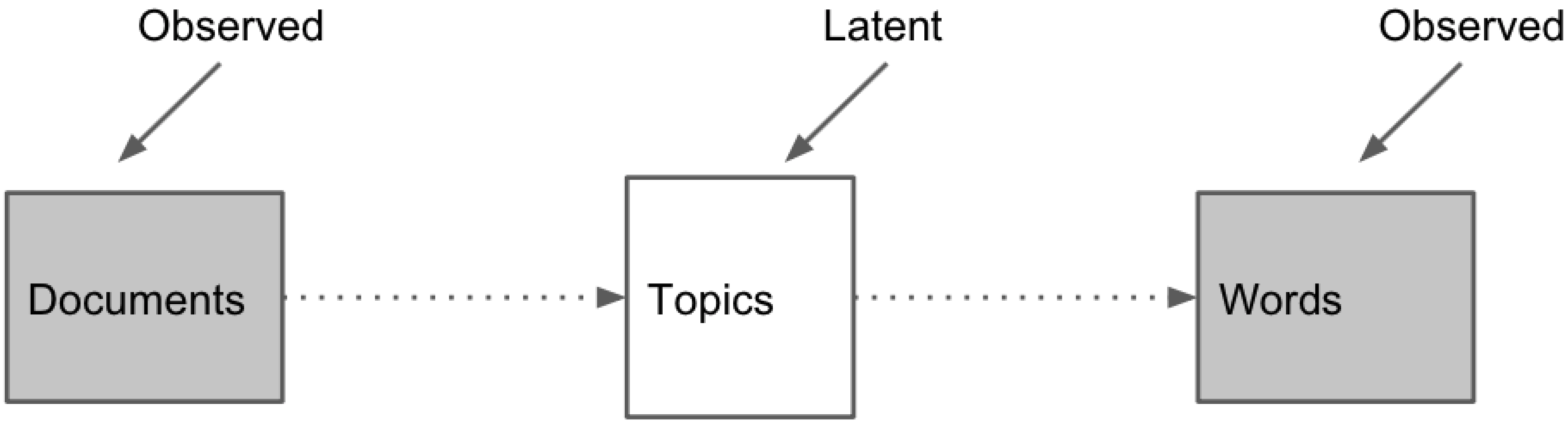
APPLICATION

TOPIC MODELING

These techniques are commonly used for recommending news articles or mining large troves of data data and trying to find commonalities.

Topic modeling, is used in the NY times recommendation engine by mapping the NYT articles to a **latent space of topics**.

LATENT SPACE OF TOPICS



- Documents are about several topics at the same time. Topics are associated with different words.
- Topics in the documents are expressed through the words that are used

GOAL OF TOPIC MODELING

Fast and easy birds eye view of the large datasets.

- What are the documents about?
- What are the key themes?

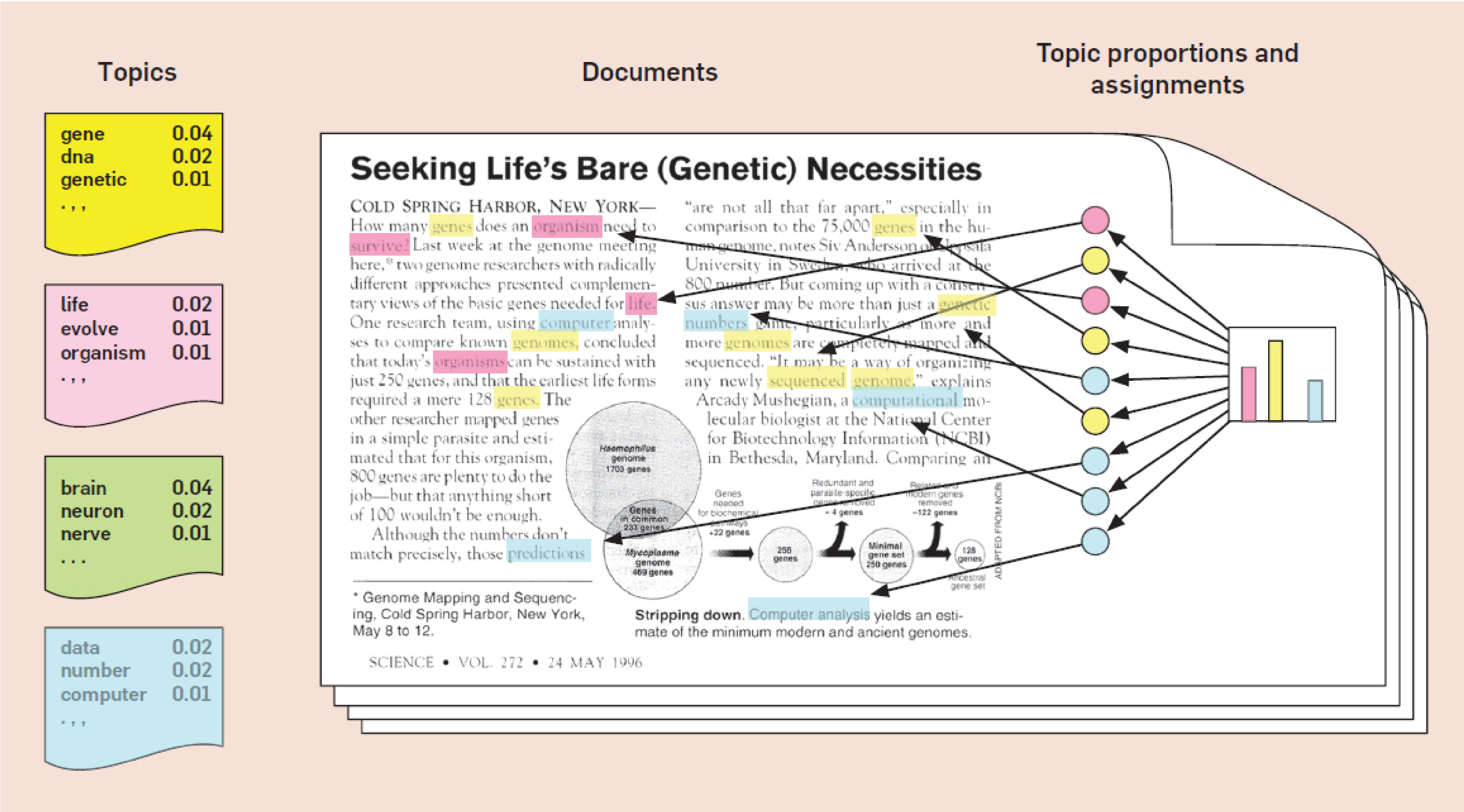
Very powerful when coupled with different covariates: year of publication, author...

- Longitudinal analysis: How the key themes change over time?
- Focus of discussion: Who is focussing on one topic

Examples:

- Topic Modeling in Presidential Debates

MIXTURE MODEL



TECHNIQUES

Vector-based techniques:

- Latent Semantic Analysis (LSA) (a.k.a Latent Semantic Indexing - LSI)

Probabilistic techniques

- Probabilistic Latent Semantic Analysis (pLSA)
- Latent Dirichlet Allocation (LDA)
  - Many LDA extensions
  - Hierachical Dirichlet Process

LATENT SEMANTIC ANALYSIS (LSA)

This is our corpus

- D1: *modem the steering linux. modem, linux the modem. steering the modem. linux!*
- D2: *linux; the linux. the linux modem linux. the modem, clutch the modem. petrol.*
- D3: *petrol! clutch the steering, steering, linux. the steering clutch petrol. clutch the petrol; the clutch.*
- D4: *the the the. clutch clutch clutch! steering petrol; steering petrol petrol; steering petrol!!!!*

PREPROCESSED

- D1: *modem the steering linux modem linux the modem steering the modem linux*
- D2: *linux the linux the linux modem linux the modem clutch the modem petrol*
- D3: *petrol clutch the steering steering linux the steering clutch petrol clutch the petrol the clutch*
- D4: *the the the clutch clutch clutch steering petrol steering petrol petrol steering petrol*

DOCUMENT TERM MATRIX

	D1	D2	D3	D4
linux	3	4	1	0
modem	4	3	0	1
the	3	4	4	3
clutch	0	1	4	3
steering	2	0	3	3
petrol	0	1	3	4

$$\begin{array}{cc} & \begin{array}{ccc} & \mathbf{d}_j & \\ & \downarrow & \end{array} \\ \mathbf{t}_i^T \rightarrow & \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{bmatrix} \end{array}$$

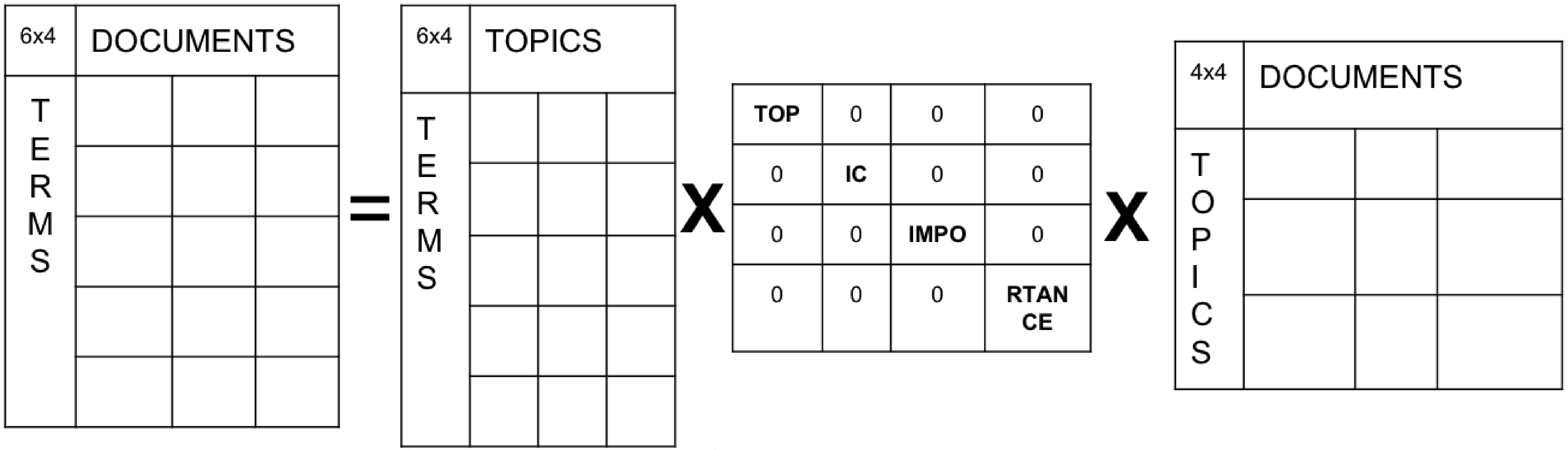
This matrix can be huge

How can we reduce it and at the same time uncover the topics?

LATENT SEMANTIC ANALYSIS

Singular value decomposition (SVD) of the document-term matrix:

Find three matrices U, Σ, V so that: $X = U\Sigma V^t$



- Cool Linear Algebra: Singular Value Decomposition

LATENT SEMANTIC ANALYSIS

DIMENSION REDUCTION

For example with 5 topics, 1000 documents and 1000 word vocabulary

- Original Document Term matrix: $1000 \times 1000 = 10^6$
- LSA representation: $5 \times 1000 + 5 + 5 \times 1000$ 10^4
  - -> 100 times less space

LATENT SEMANTIC ANALYSIS

$$
\begin{bmatrix} 3 & 4 & 1 & 0 \\ 4 & 3 & 0 & 1 \\ 3 & 4 & 4 & 3 \\ 0 & 1 & 4 & 3 \\ 2 & 0 & 3 & 3 \\ 0 & 1 & 3 & 4 \end{bmatrix}
=
\begin{array}{c c}
 & \begin{array}{cccc} \text{To1} & \text{To2} & \text{To3} & \text{To4} \end{array} \\
\begin{array}{c} \text{Te1} \\ \text{Te2} \\ \text{Te3} \\ \text{Te4} \\ \text{Te5} \\ \text{Te6} \end{array} &
\begin{bmatrix} -0.33 & -0.53 & 0.37 & -0.14 \\ -0.32 & -0.54 & -0.49 & 0.35 \\ -0.62 & -0.10 & 0.26 & -0.14 \\ -0.38 & 0.42 & 0.30 & -0.24 \\ -0.36 & 0.25 & -0.68 & -0.47 \\ -0.37 & 0.42 & 0.02 & 0.75 \end{bmatrix}
\end{array}
\times
\begin{array}{c}
\text{Topic Importance} \\
\begin{bmatrix} 11.4 & & & \\ & 6.27 & & \\ & & 2.22 & \\ & & & 1.28 \end{bmatrix}
\end{array}
\times
\begin{array}{c c}
 & \begin{array}{cccc} \text{D1} & \text{D2} & \text{D3} & \text{D4} \end{array} \\
\begin{array}{c} \text{To1} \\ \text{To2} \\ \text{To3} \\ \text{To4} \end{array} &
\begin{bmatrix} -0.42 & -0.48 & -0.57 & -0.51 \\ -0.56 & -0.52 & 0.45 & 0.46 \\ -0.65 & 0.62 & 0.28 & -0.35 \\ -0.30 & 0.34 & -0.63 & 0.63 \end{bmatrix}
\end{array}
$$

LATENT SEMANTIC ANALYSIS

Keep the 2 most important Eigenvalues (i.e topic importance)

3	4	1	0
4	3	0	1
3	4	4	3
0	1	4	3
2	0	3	3
0	1	3	4

=

	To1	To2
Te1	-0.33	-0.53
Te2	-0.32	-0.54
Te3	-0.62	-0.10
Te4	-0.38	0.42
Te5	-0.36	0.25
Te6	-0.37	0.42

X

Topic Importance

11.4			
	6.27		
		2.22	
			1.28

X

	D1	D2	D3	D4
To1	-0.42	-0.48	-0.57	-0.51
To2	-0.56	-0.52	0.45	0.46

LATENT SEMANTIC ANALYSIS

3	4	1	0
4	3	0	1
3	4	4	3
0	1	4	3
2	0	3	3
0	1	3	4

=

Word assignment to topics

	IT	cars
linux	-0.33	-0.53
modem	-0.32	-0.54
the	-0.62	-0.10
clutch	-0.38	0.42
steering	-0.36	0.25
petrol	-0.37	0.42

X

Topic Importance

11.4	
	6.27

X

Topic distribution acr documents

	D1	D2	D3	D4
IT	-0.42	-0.48	-0.57	-0.51
cars	-0.56	-0.52	0.45	0.46

LAB LSA

LSA WITH SCIKIT

Latent Semantic Analysis - LAB

PROBABILISTIC LSA

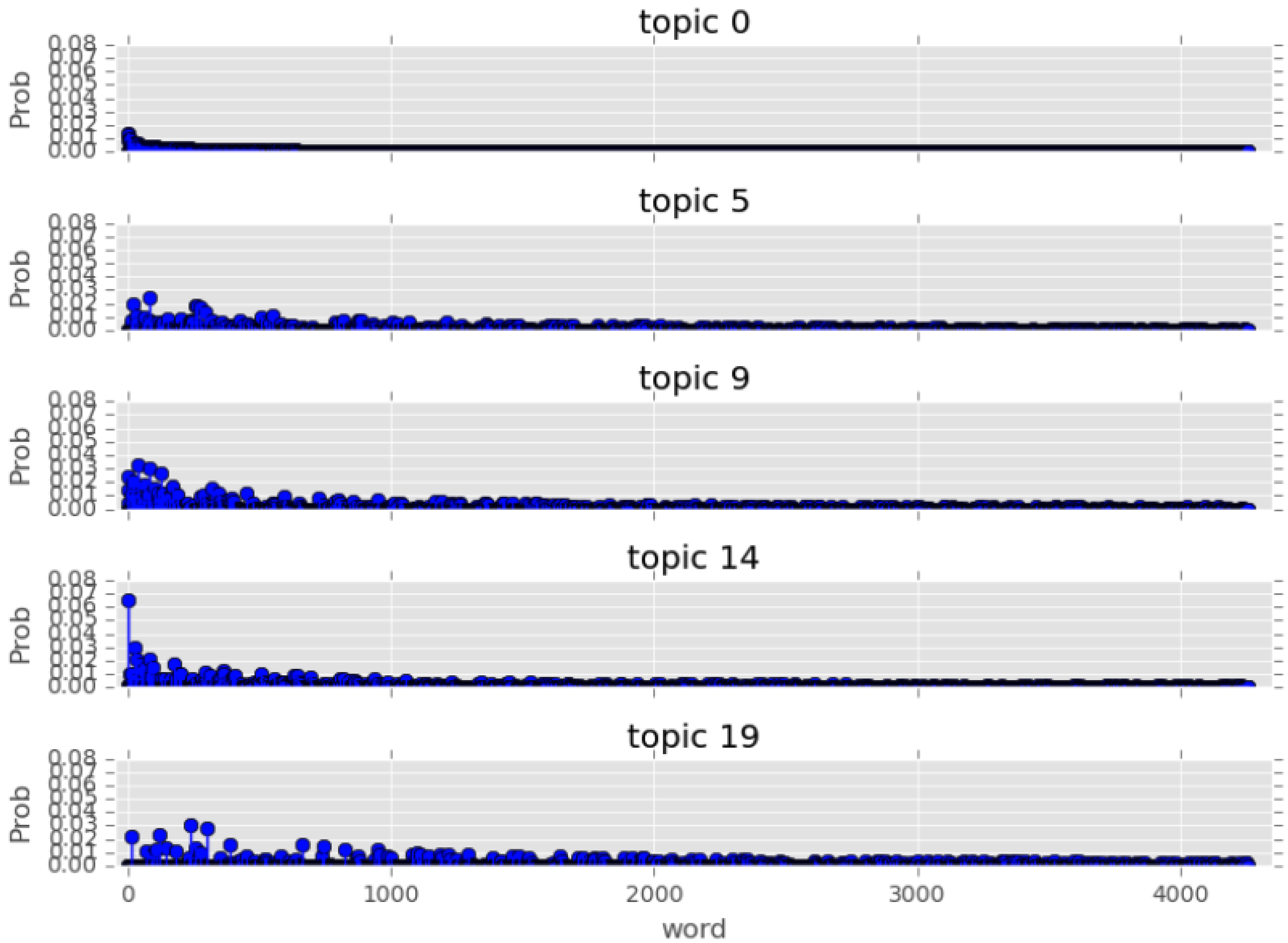
What is a topic?

A list of probabilities for each of the possible words in a vocabulary.

Example topic:

- dog: 5%
- cat: 5%
- hamster: 3%
- turtle: 1%
- calculus: 0.000001%
- analytics:
  0.000001%

PROBABILISTIC LSA



PROBABILISTIC LSA

Instead of finding lower-ranked matrix representation, we can try to find a **mixture** of *word -> topic* & *topic -> documents* distributions that are most likely given the observed documents.

- We define a statistical model of how the documents are being made (generated).
- Then we try to find parameters of that model that best fit the observed data

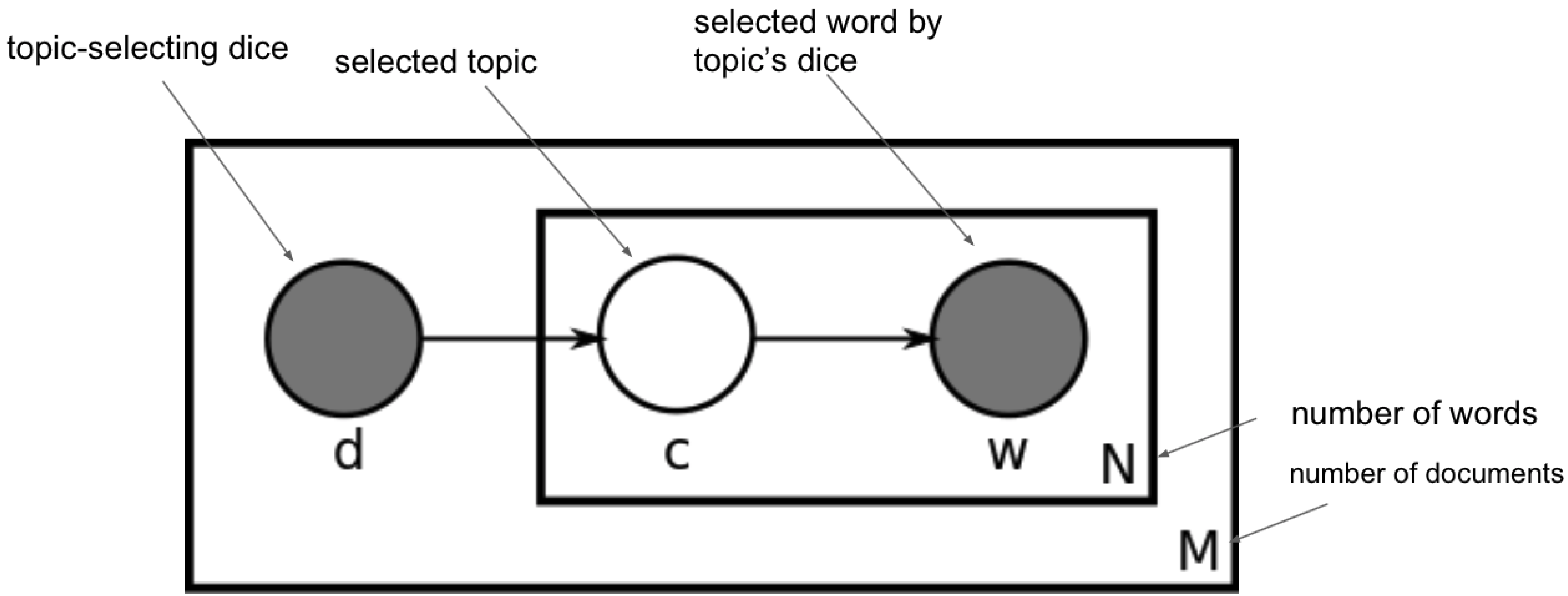
This is called a **generative process** in topic modeling terminology.

GENERATIVE PROCESS

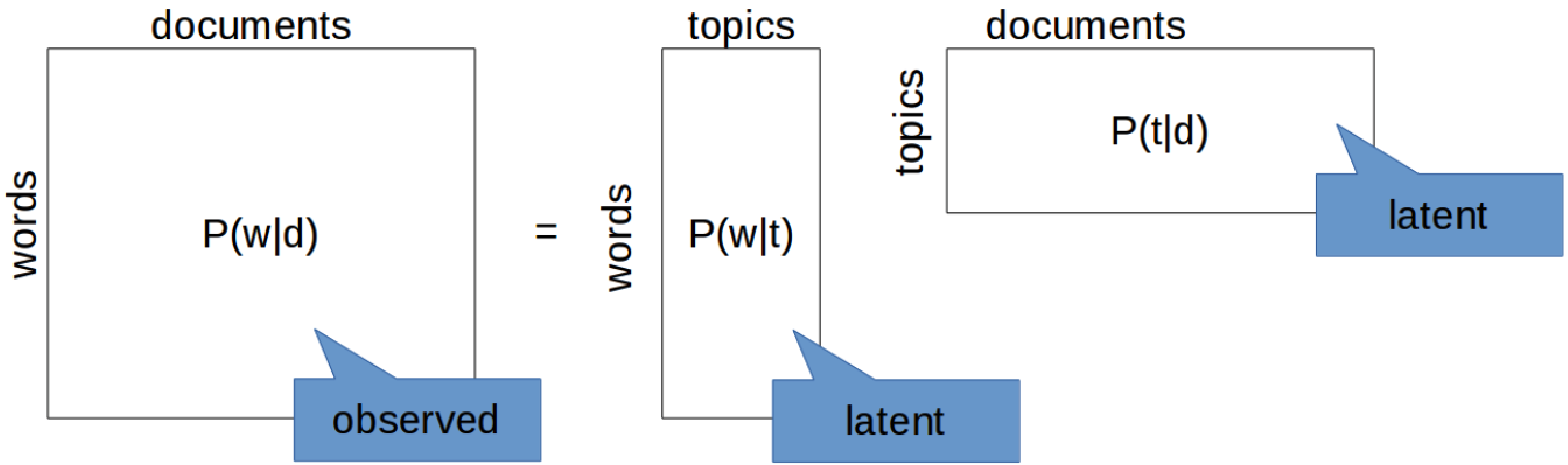
We received a 50 word long document by our reporter John Doe. He is allowed to write only about one of the 6 possible topics, using only 6 words.

- For the first word, he throws a dice that tells him what is the topic of the first word. Say it is topic 1 (IT)
- Then he throws another dice to pick which word to use to describe topic 1. Say it is word 1 (linux)
- The process is repeated for all 50 words in the document.
- Dices are weighted!!!
  - The first dice for picking topics puts more weight on IT topic that on the other 5 topics.
  - Also, dice for IT topic, puts more weight on words 'linux' and 'modem'.
  - Likewise dice for topic 2 (cars) puts more weight on word 'petrol' and 'steering'

GENERATIVE PROCESS



PROBABILISTIC LSA DECOMPOSITION



$$P(word/document) = \sum_{topics} p(topic/document).\, p(word/topic)$$

LDA: AN EXTENSION TO PLSA

- pLSA: Binomial distribution
- LDA: Dirichlet distribution

LDA ASSUMPTIONS

In LDA, we encode our assumptions about the data. Two important assumptions:

1. On average, how many topics are per document? more or less?
2. On average, how are words distributed across topics? Are topics strongly associated with more or less words?

Those assumptions are defined by two vectors α and β:

- α: K dimensional vector that defines how K topics are distributed across documents. Smaller **αs favor fewer topics** strongly associated with each document.
- β: V dimensional vector that defines how V words are associated across topics. **Smaller βs favor fewer words** strongly associated with each topics

LDA

We set K the number of topics

We work backwards from the documents to the find the α and β

Latent Dirichlet Allocation - Gensim

HOT TECHNOLOGY TOPICS

https://github.com/alexperrier/gads/blob/master/14_topic_modeling/py/Hot%20Tech%20T

LINKS

- Building the Next New York Times Recommendation Engine
- Topic Modeling in historical Newspapers
- Dissecting the Presidential Debates with an NLP Scalpel
- Clustering text documents using k-means
- Topic Modeling of Twitter Followers
- Dirichlet Distribution
- Topic Modeling in historical Newspapers
- Topic Modeling for the Social Sciences