HUDK 4051: LEARNING ANALYTICS: PROCESS & THEORY

2/22/17 11:56 AM

https://thedigitalservice.org

In the news

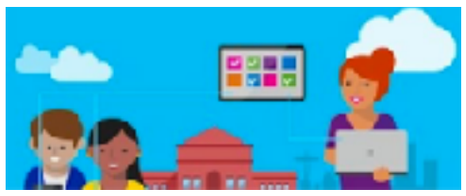
What Betsy DeVos means for edtech

This early-stage edtech startup wants to gamify cramming for tests

"Our primary goal is to curate the content that exists, but also to really present it in a way that isn't terrifying to kids."

The Zuckerberg Ed-Tech Connection

Survey Finds Some Ed Tech Software and Applications Are Lacking Proper Security



Nonprofits help school districts make good decisions on ed tech

Survey Says 86 Percent of Schools Expect to Spend More on Digital Curriculum

Microsoft Launches New Cloud Tool and Convertible Devices to Ease Classroom Tech Use

Is higher ed ready for the big edtech explosion?

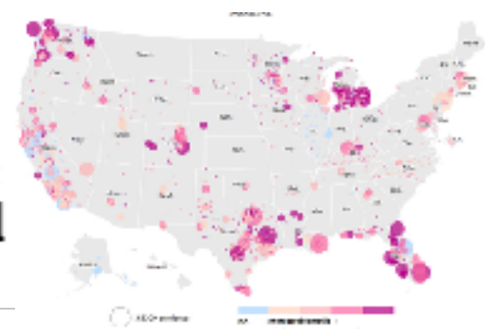
$252 billion to be spent by colleges and universities on campus edtech by 2020

Area school systems frustrated with the Tennessee Department of Education's data mistakes

DOE emails hundreds of teacher's assistants' social security

UNESCO education data puts Mauritius ahead of Nigeria

'Alternative' Education: Using Charter Schools to Hide Dropouts and Game the System



Events

- Learning Analytics Seminar Series, March 9 Andrew Gibson Writing Analytics (http://laseries.pressible.org/)
- NYC School of Data 2017, March 4 (https://www.eventbrite.com/e/nyc-school-of-data-2017-tickets-32191968043)
- Big Data Hub Workshop, Feb 24 (http://nebigdatahub.org/event/2017-nebdih-annual-workshop/)
- 2017 Art+Feminism Wikipedia Edit-a-thon, March 11 (https://www.eventbrite.com/e/2017-artfeminism-wikipedia-edit-a-thon-tickets-31462938496)
- 2017 Tri-State Education Career Fair, March 4 (Cowin Conference Center. Students should register on TCCS LINK)

Opportunities

- Ford Foundation Technology Fellow (https://ford-foundation.forms.fm/technology-fellow?utm_content=bufferbb868&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer#pub_banner)

- Montefiore Life Sciences TODAY 4-5pm Thompson 136 (https://www.myinterfase.com/columbia-tc/job_view.aspx?token=2MphGL58+iOvrB%2fyBD0cdQ%3d%3d)

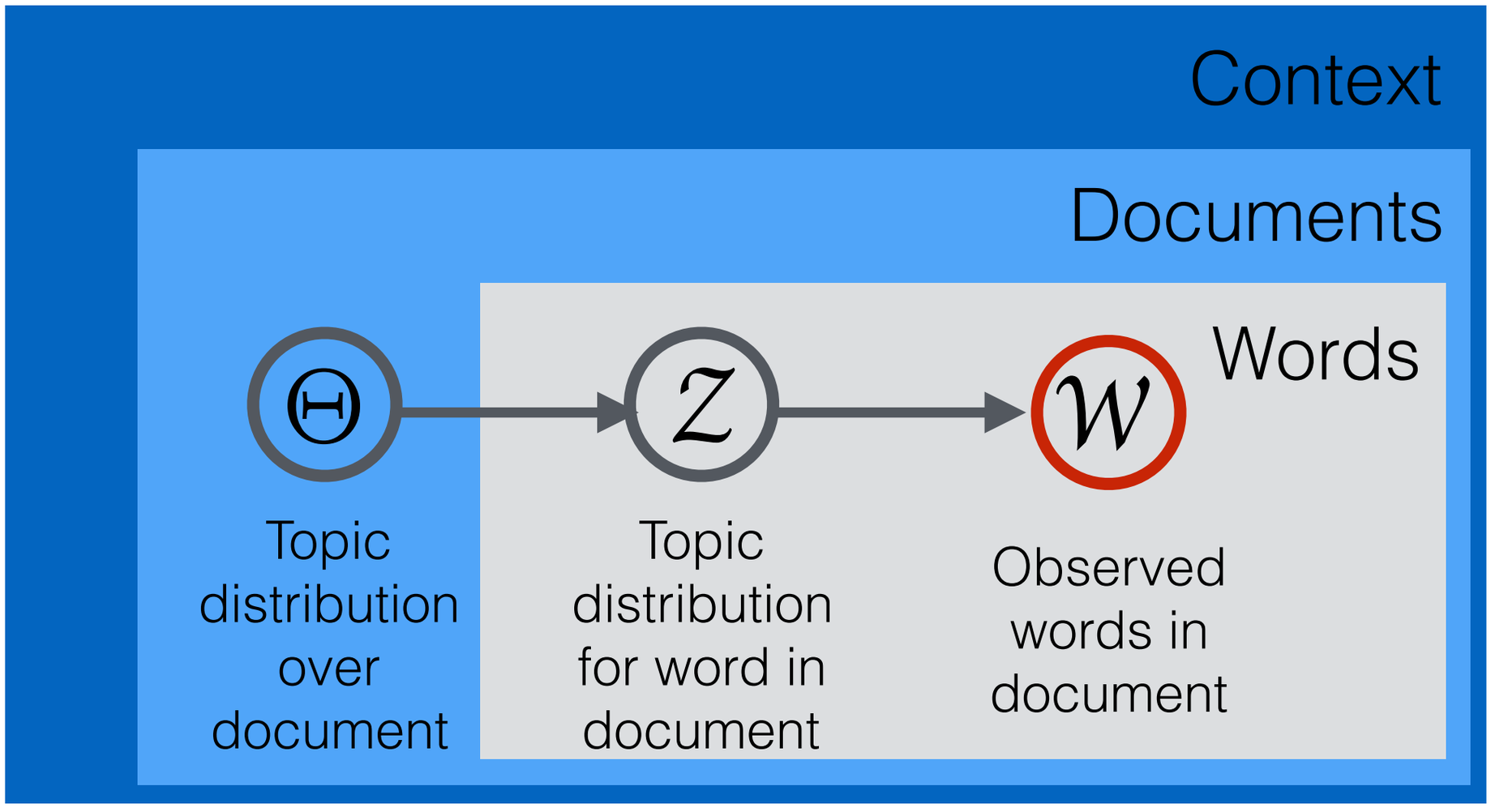
- Student Success Analytics network under the Educause (https://docs.google.com/document/d/1AYScI5H950hBiW1BH9w1D2qZvgHkGfacJxqsuQ2Me44/edit)

Topic Modeling with Latent Dirichlet Analysis (LDA)

Topic Modeling

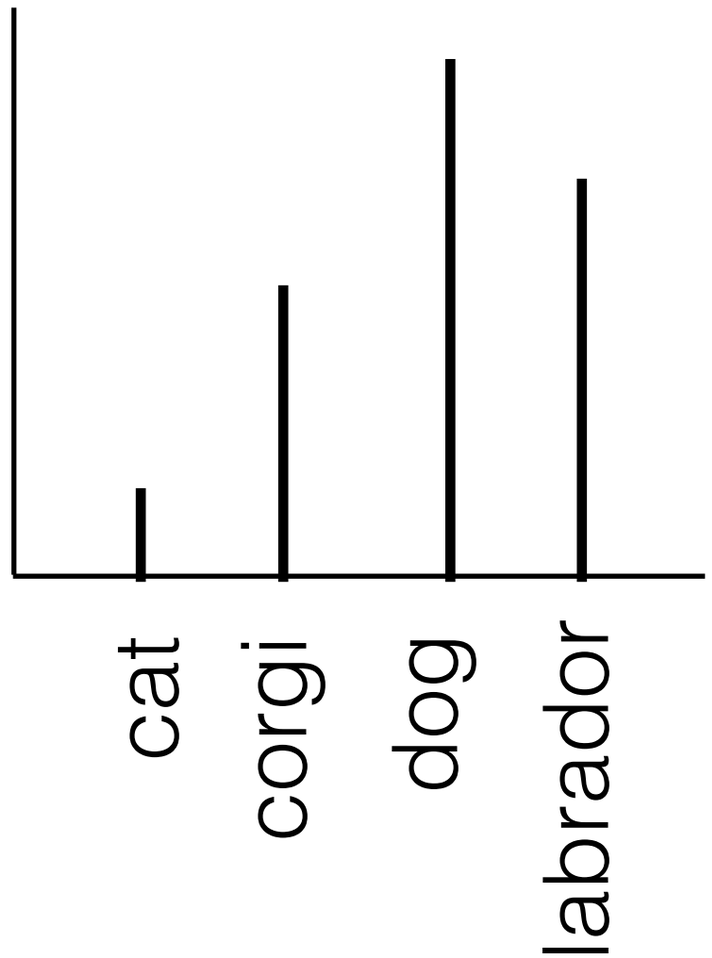
A topic model is a type of statistical model for discovering the abstract topics that occur in a collection of documents

Organizing Words

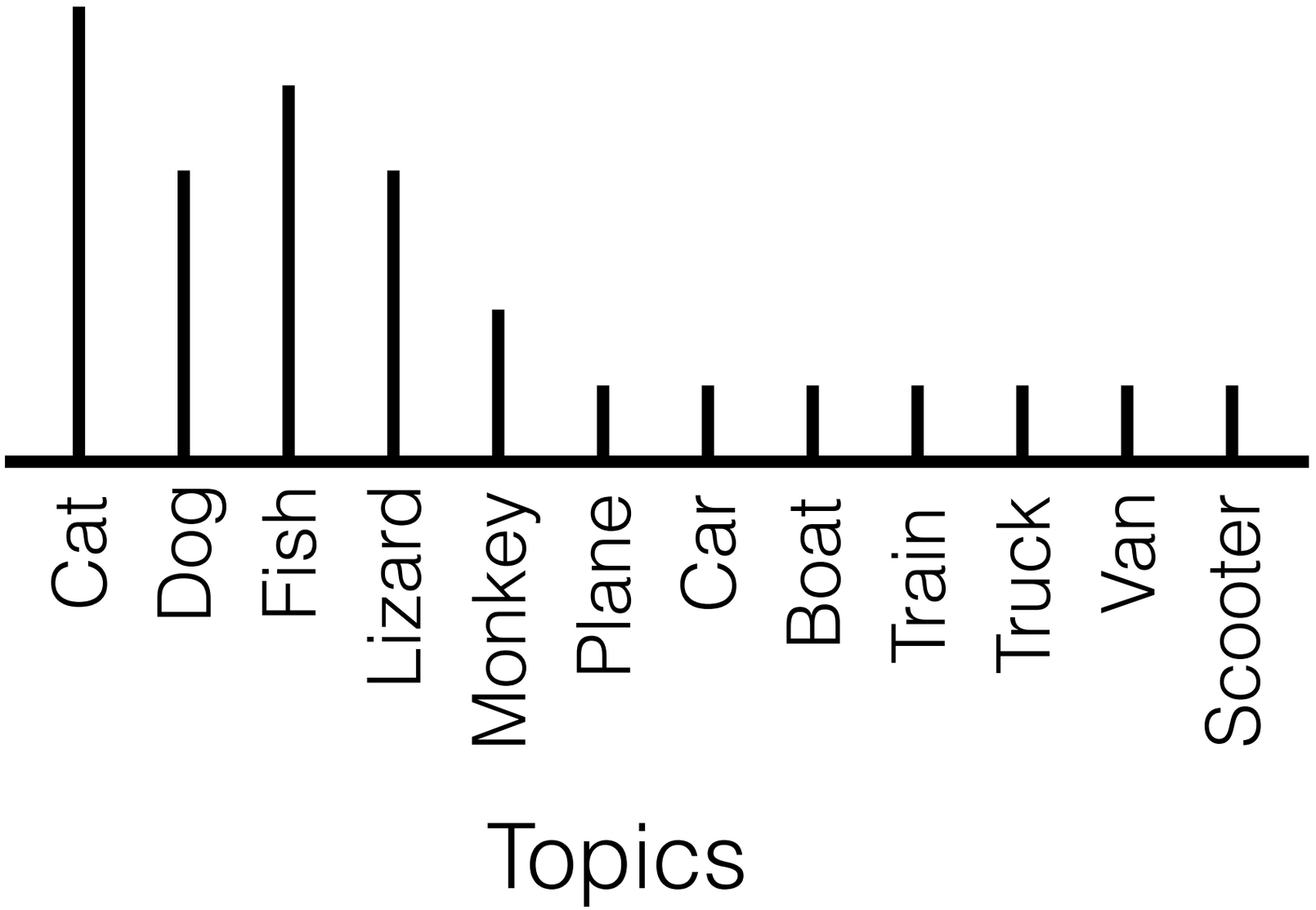


Topics (Z)

- A topic is a probability distribution over words



Topic Distribution for a Document



A document can be described by a recipe of topics and “how much” of each topic it contains

Documents

- A document is a probability distribution over topics

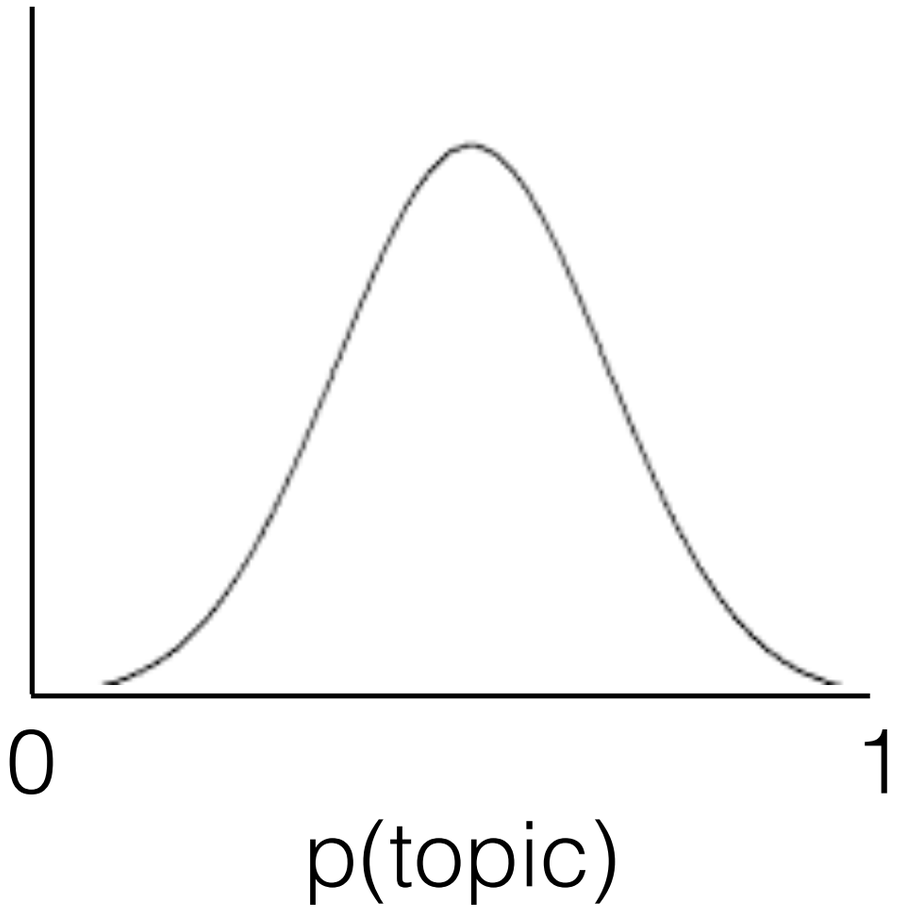
Document

word word word
word word word
word word word
word word word
word word word
word word word
word word word

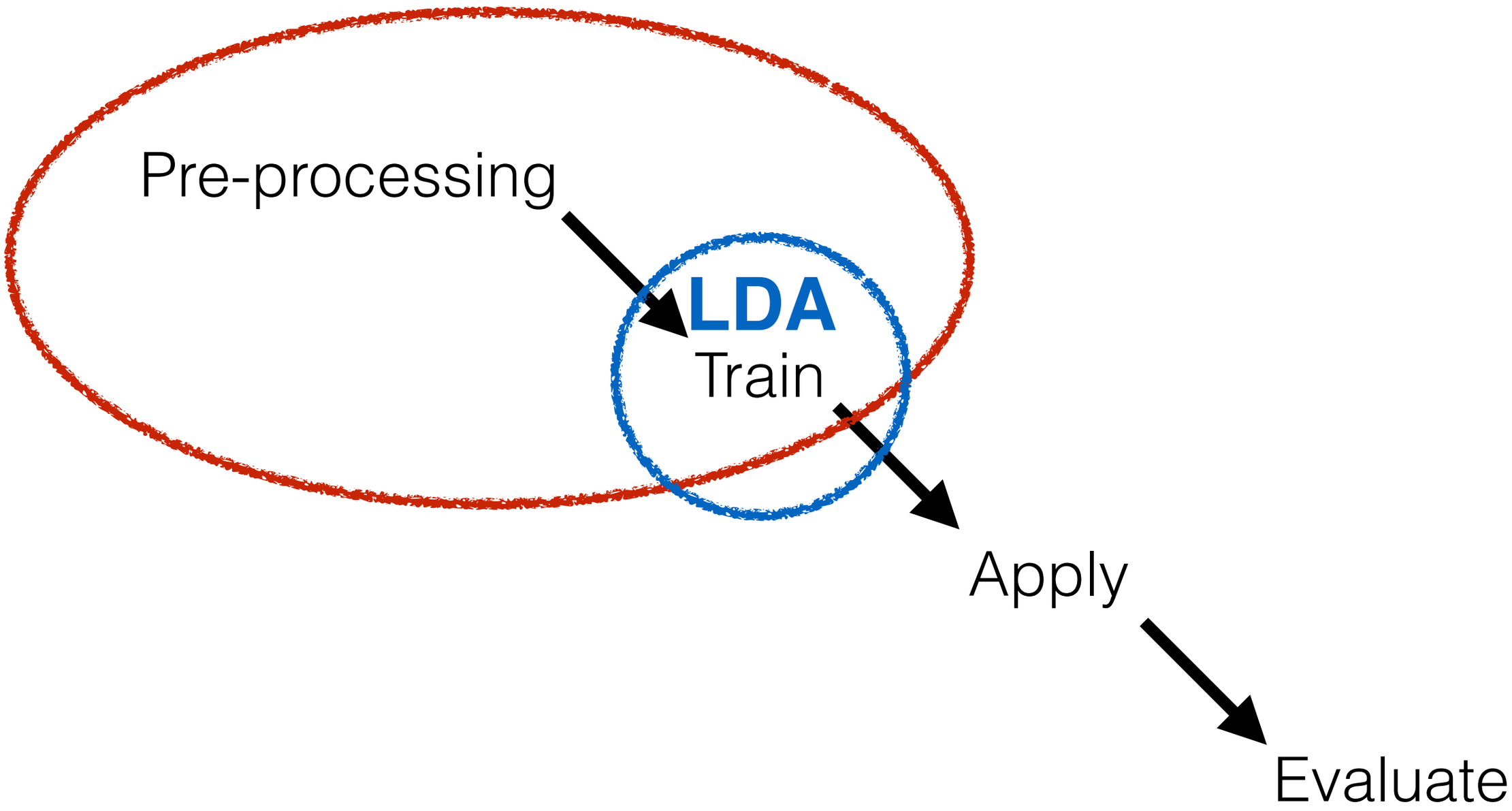
Topic 1

Topic 2

Topic 3



Process



What does LDA do?

- Assumes that documents cover particular topics and particular topics are covered by particular words

- Therefore, can group similar documents by their word profiles which represent topics

- LDA calculates those distributions

- Like cluster analysis we need to supply the number of topics

Logic of Process

Document

word word word
word word word
word word word
word word word
word word word
word word word
word word word

Topic 1

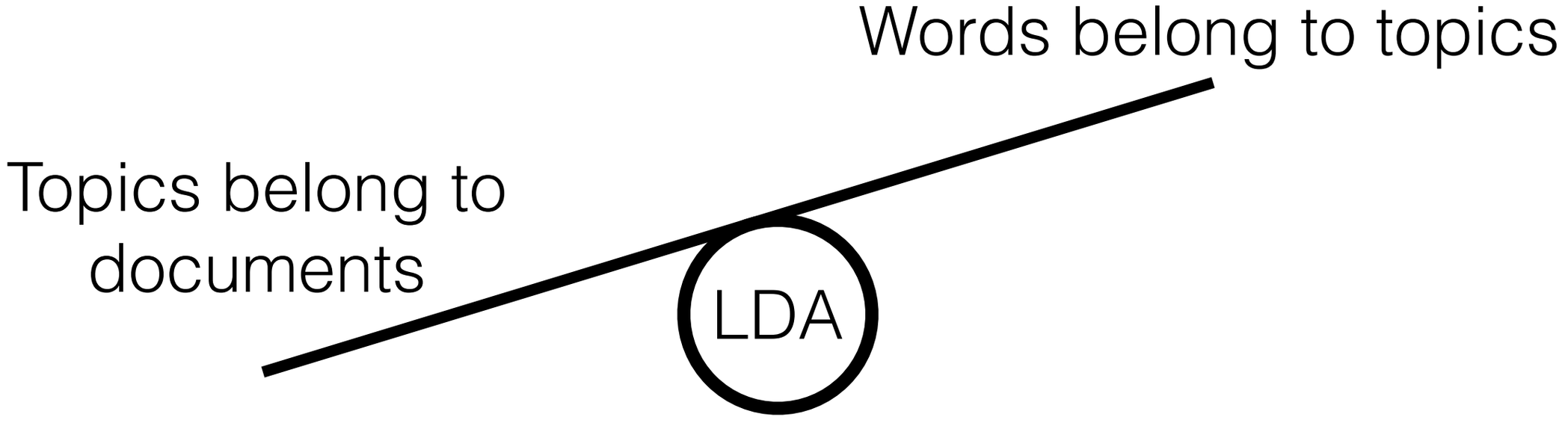
Topic 2

Topic 3

Basic Idea

- Documents are made up of words that belong (with some probability) to topics
- So…We can just reverse engineer these words to learn what a document is about

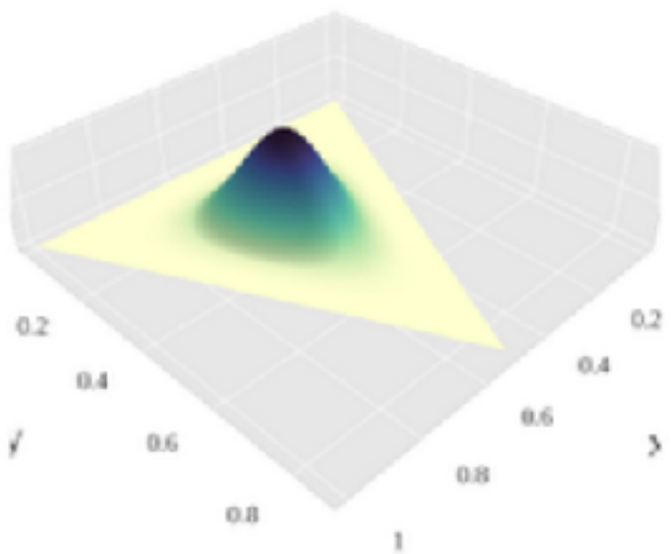
LDA

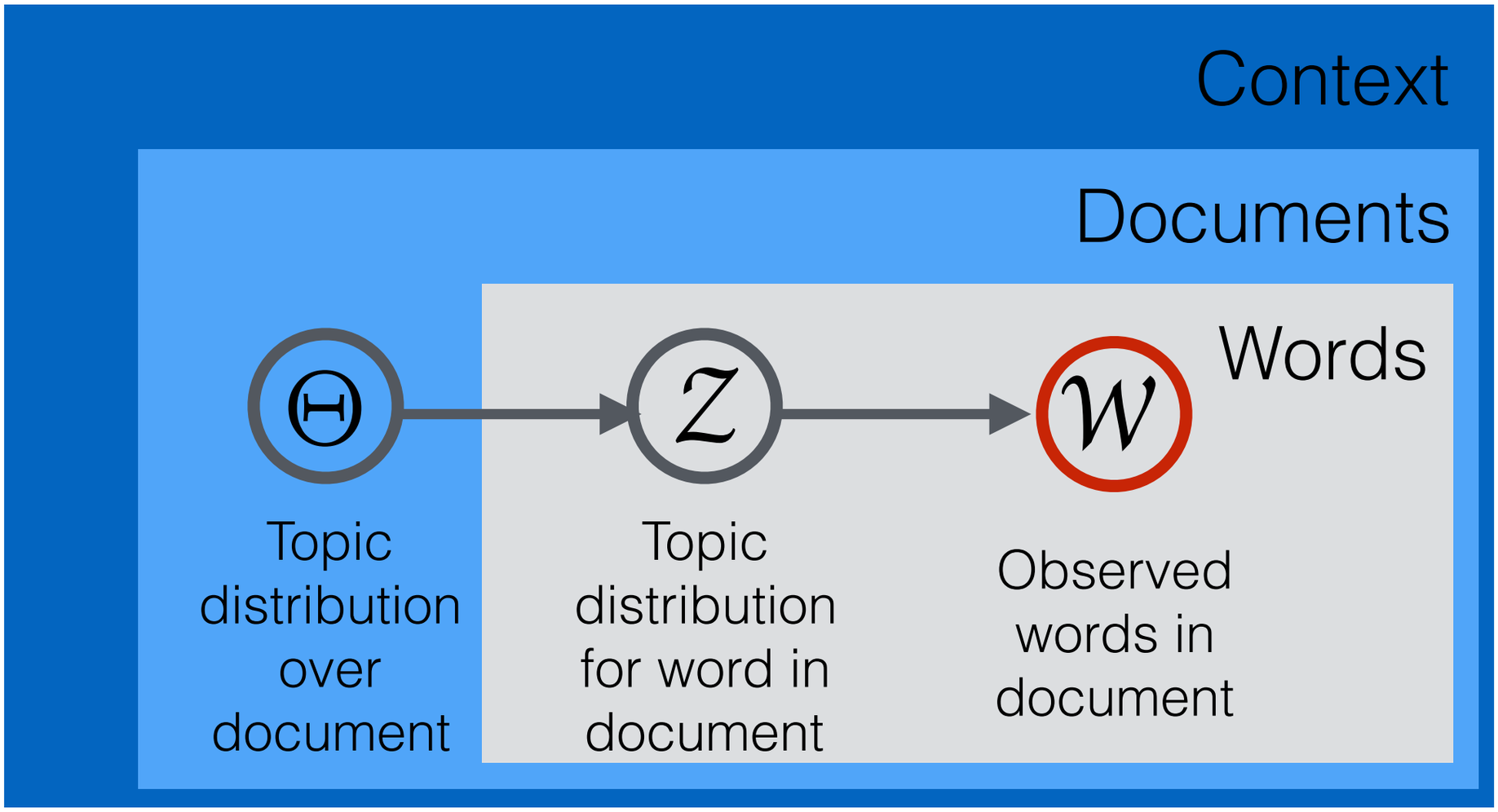


Dirichlet Distribution

- Peter Gustav Lejeune Dirichlet
- 1805 - 1859
- German mathematician
- Helped develop the definition of the word *function*

- Distribution on probability distributions




Context
Documents
Words
Θ
Z
W
Topic distribution over document
Topic distribution for word in document
Observed words in document

Term Document vs. Document Term Matrices

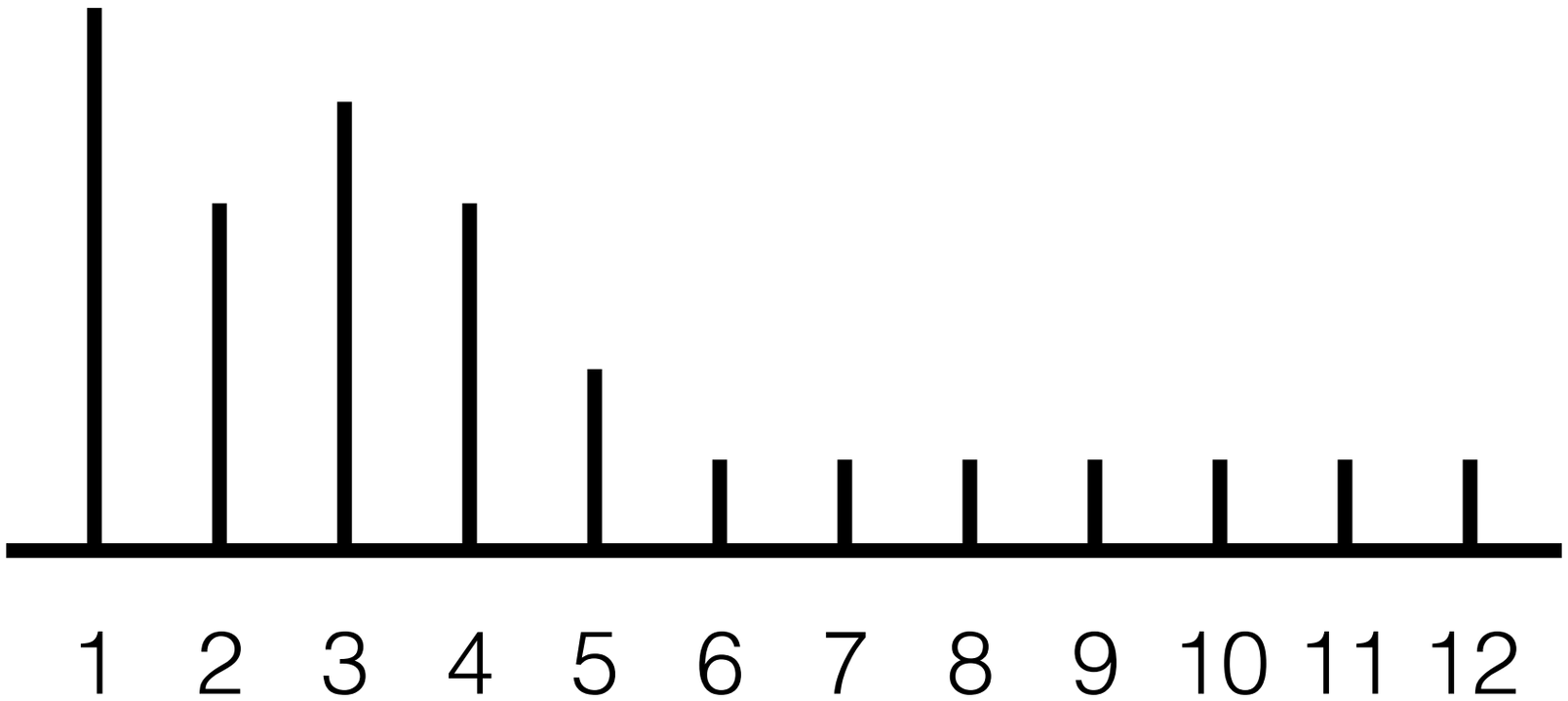
	Term1	Term2	Term3
Doc1			
Doc2			
Doc3			

	Doc1	Doc2	Doc3
Term1			
Term2			
Term3			

Term Frequency = Number of times a word appears in a document

Inverse Document Frequency = number of documents in the corpus which contain a term

Topic Distribution for a Document



If we have both of those pieces of information & the model…

We can predict the topic of a document