HUDK 4051: LEARNING ANALYTICS: PROCESS & THEORY

3/3/20 2:36 PM

Events

https://www.eventbrite.com/e/data-for-good-phebe-vayanos-usc-center-for-ai-in-society-tickets-92768526159?utm_source=sendinblue&utm_campaign=DSI_Newsletter_2&utm_medium=email

https://www.eventbrite.com/e/data-science-day-2020-columbia-university-tickets-86128772477?utm_source=sendinblue&utm_campaign=Events_Weekly_February_25&utm_medium=email

Dashboards

Phone Usage

https://taylormwang.shinyapps.io/Phone-Usage-Time/.

Long Jump

https://starsg123.shinyapps.io/interactive-visualization/

Age vs. Social Media Popularity

https://timlxq.shinyapps.io/shiny_activity_2/

Word Cloud

https://cindyzhou.shinyapps.io/WordCloud/

Natural Language Processing

3/3/20 2:36 PM

NLP

Analyses of language produced by humans (by computers)

- Treats language as a varied pool of information sources
- In order to:
  - Understand language (Cognitive Science)
  - Respond to the speaker appropriately (AI)

- Examples
  - Translation
  - Automated feedback (education, shopping)
  - Study linguistics, cognition, development, etc.

Methodological History

1930s

Understanding

Rule based

- Complex sets of rules (grammar/syntax)
- Chomsky

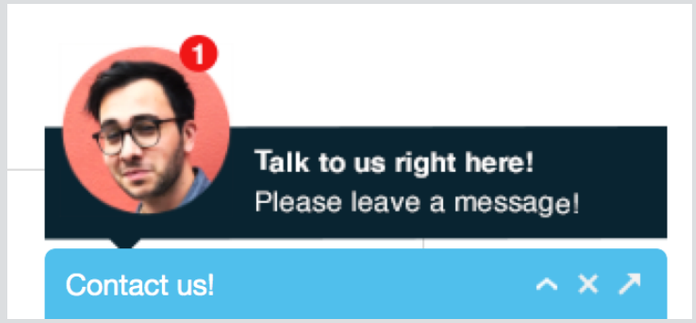
1980s

Processing

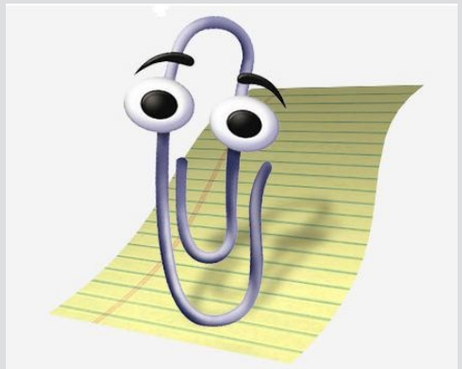
Statistical

- Infer rules from data
- IBM

Industry



Education

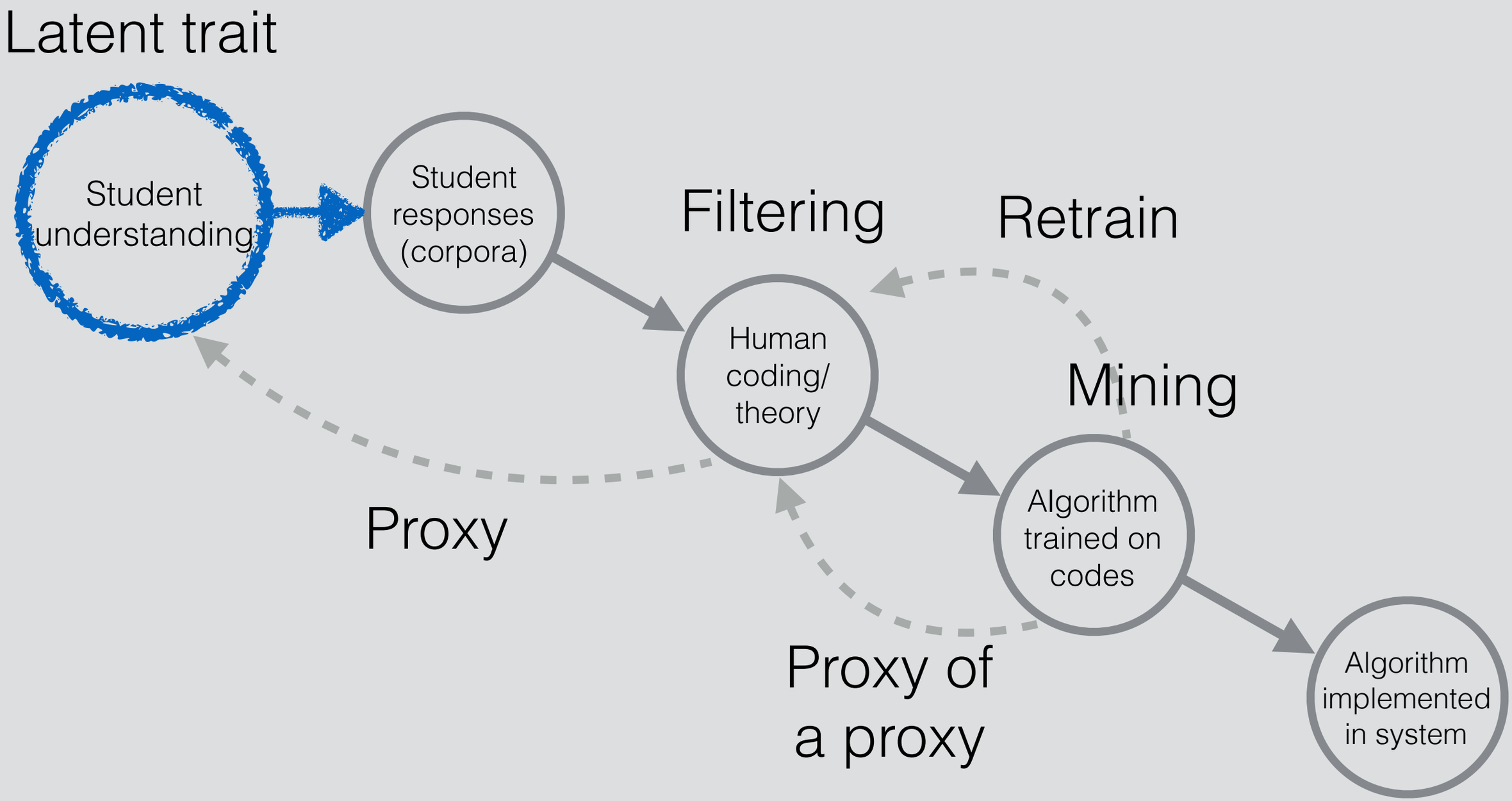


Essential Problem

- Heterogeneity
- We get rid of this by asking MCQ questions - but we also throw out a lot of information when we do that

- Collect more data and more complex data through written answers

Overall Method



Coding

Word counting

Google books Ngram Viewer

Types of Expressions

"I don't know…"
"I dunno…"

Stemming

Take the root of the word:
educate, education, educating

Tokenization (bag of words)

Chopping word/phrase into *tokens*

- Remove punctuation
- Find best number of letters to represent a word/meaning
- Consider all possible versions of word
- Stop word removal

↓

Features

Supervised Learning

Sentiment Analysis

Computationally identifying and categorizing opinions from text/audio/video

- Tokenize: cut text into useful chunks (paragraph into statement, statement into words)
    - “I love this class!” -> 
- Clean: remove stuff you don’t think is useful
    - “I love this class!” -> 
- Remove stop words
    - “I love this class!” -> 
- Classification
    - Positive (+1)/Negative (-1)/Neutral (0)
        - Train a model
        - Use a lexicon/dictionary

Algorithms

Feature selection

- Not all tokens are useful, which ones can we scrap?

Feature extraction

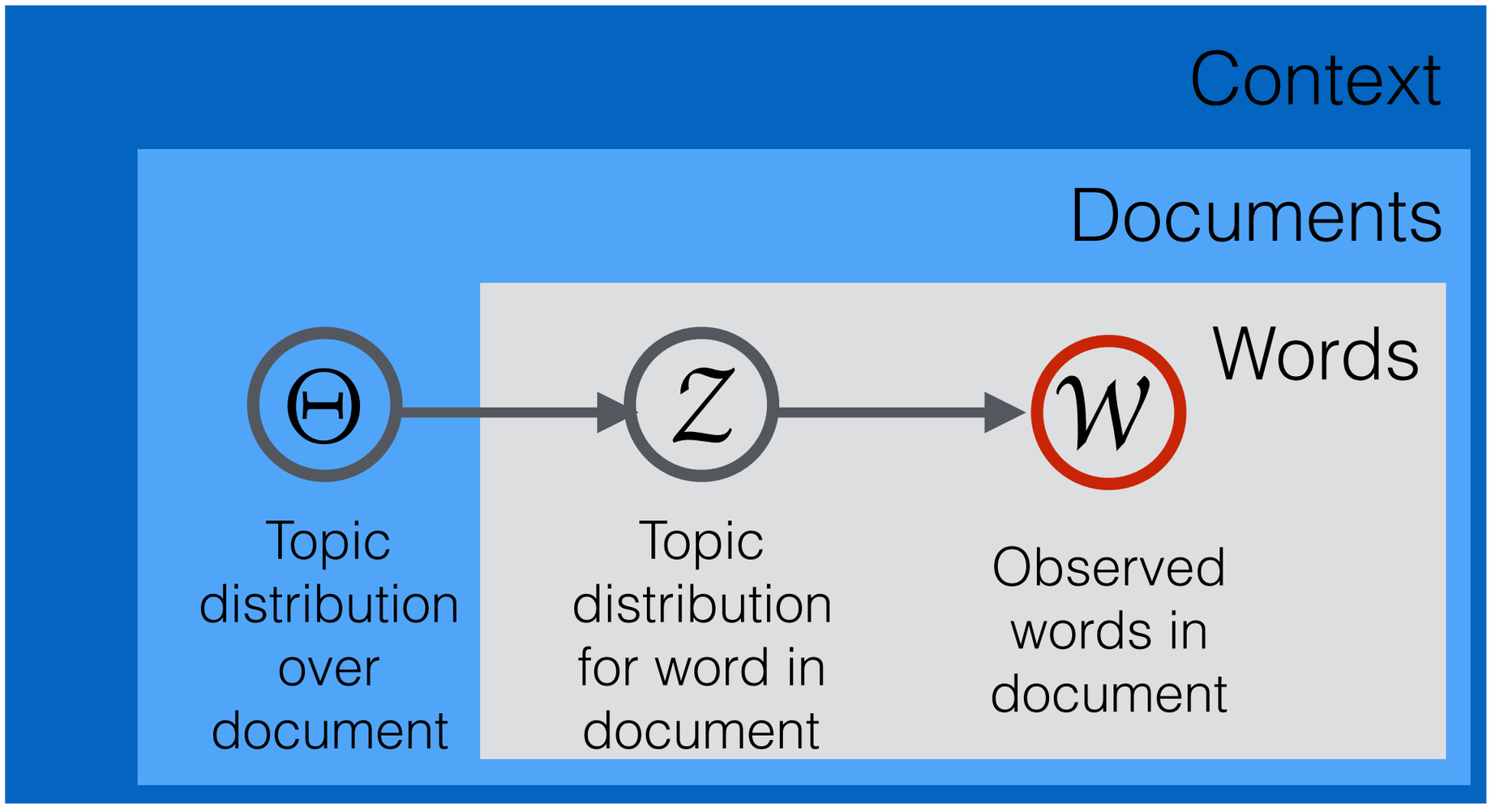
- Extracting features from combining tokens

Topic Modeling with Latent Dirichlet Analysis (LDA)

Topic Modeling

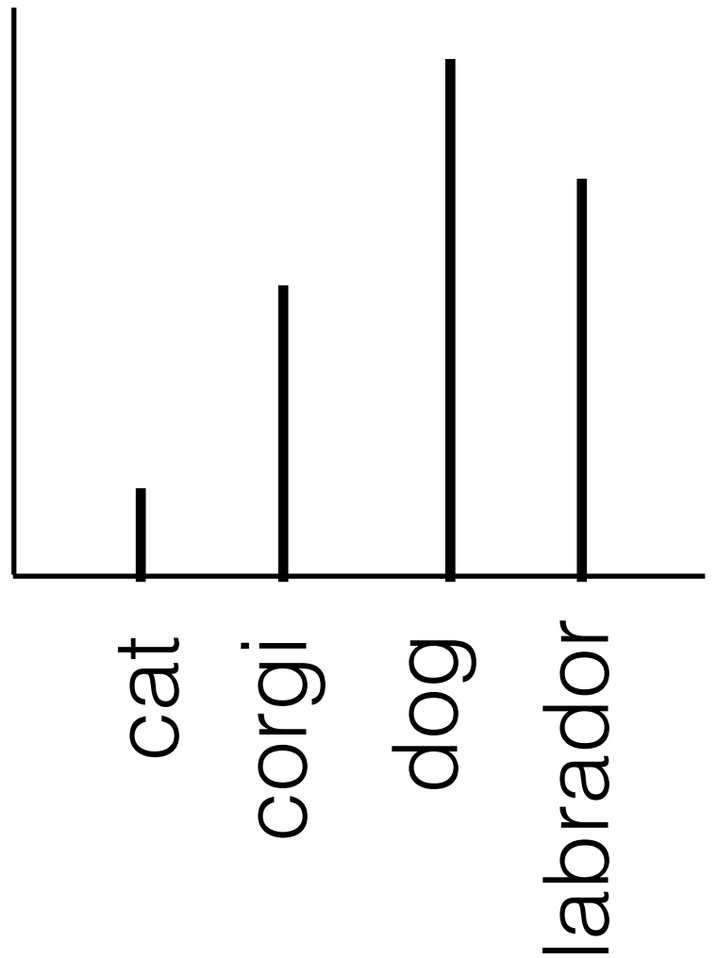
A topic model is a type of statistical model for discovering the abstract topics that occur in a collection of documents

Organizing Words

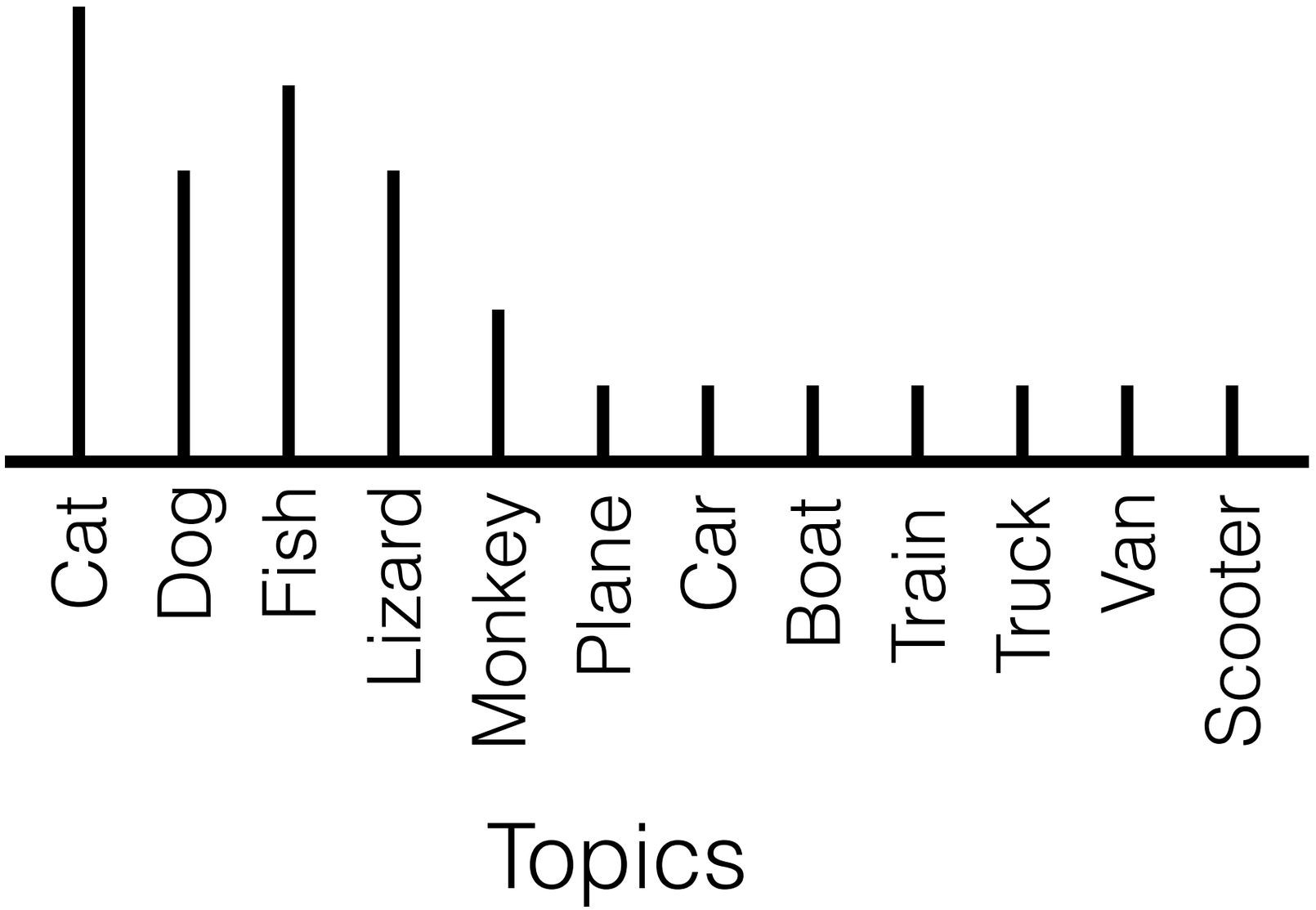


Topics (Z)

- 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

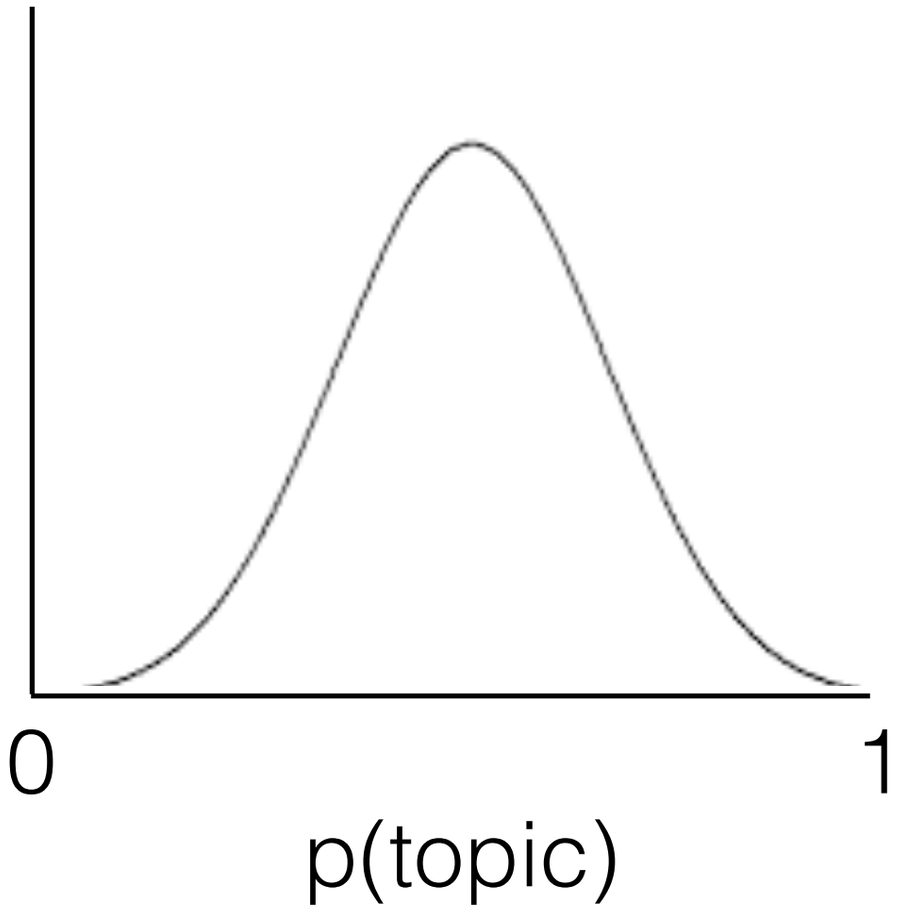
Documents

- A document is a probability distribution over topics

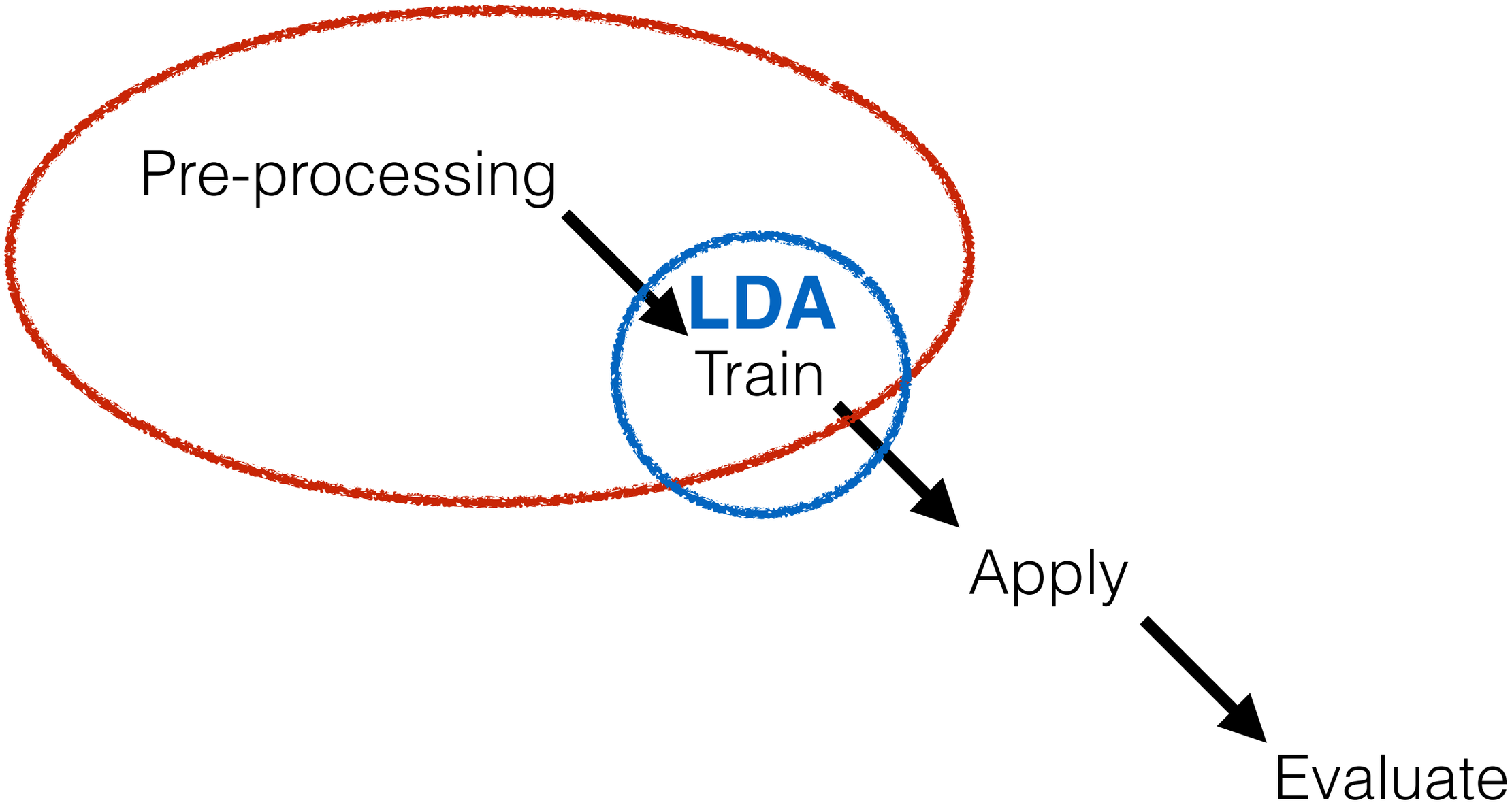
Document

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

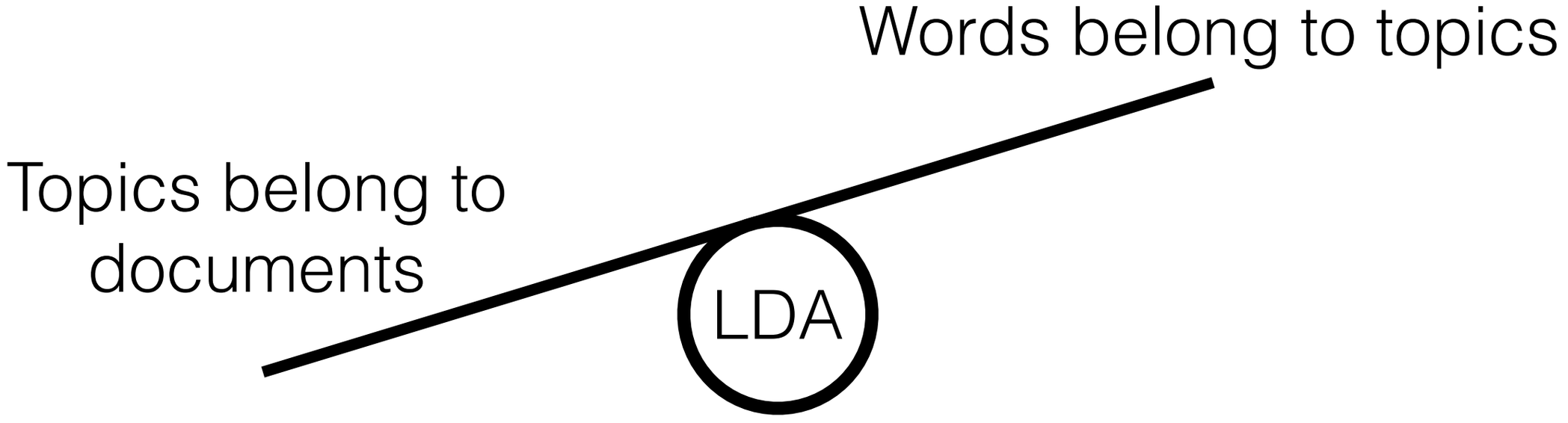
Topic 2

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

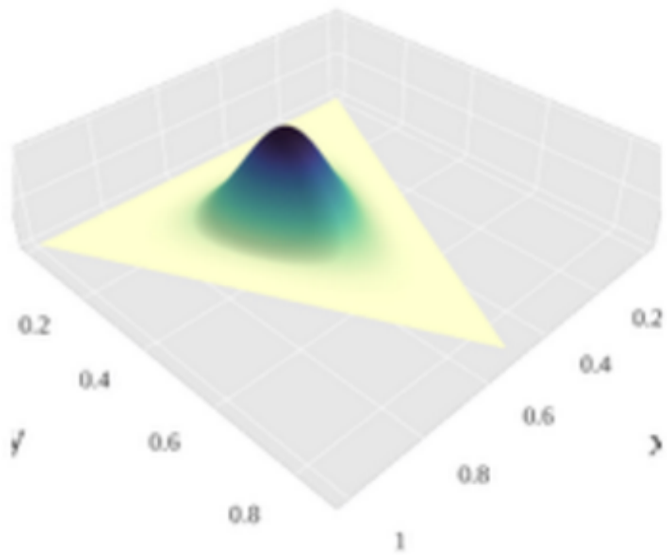
LDA

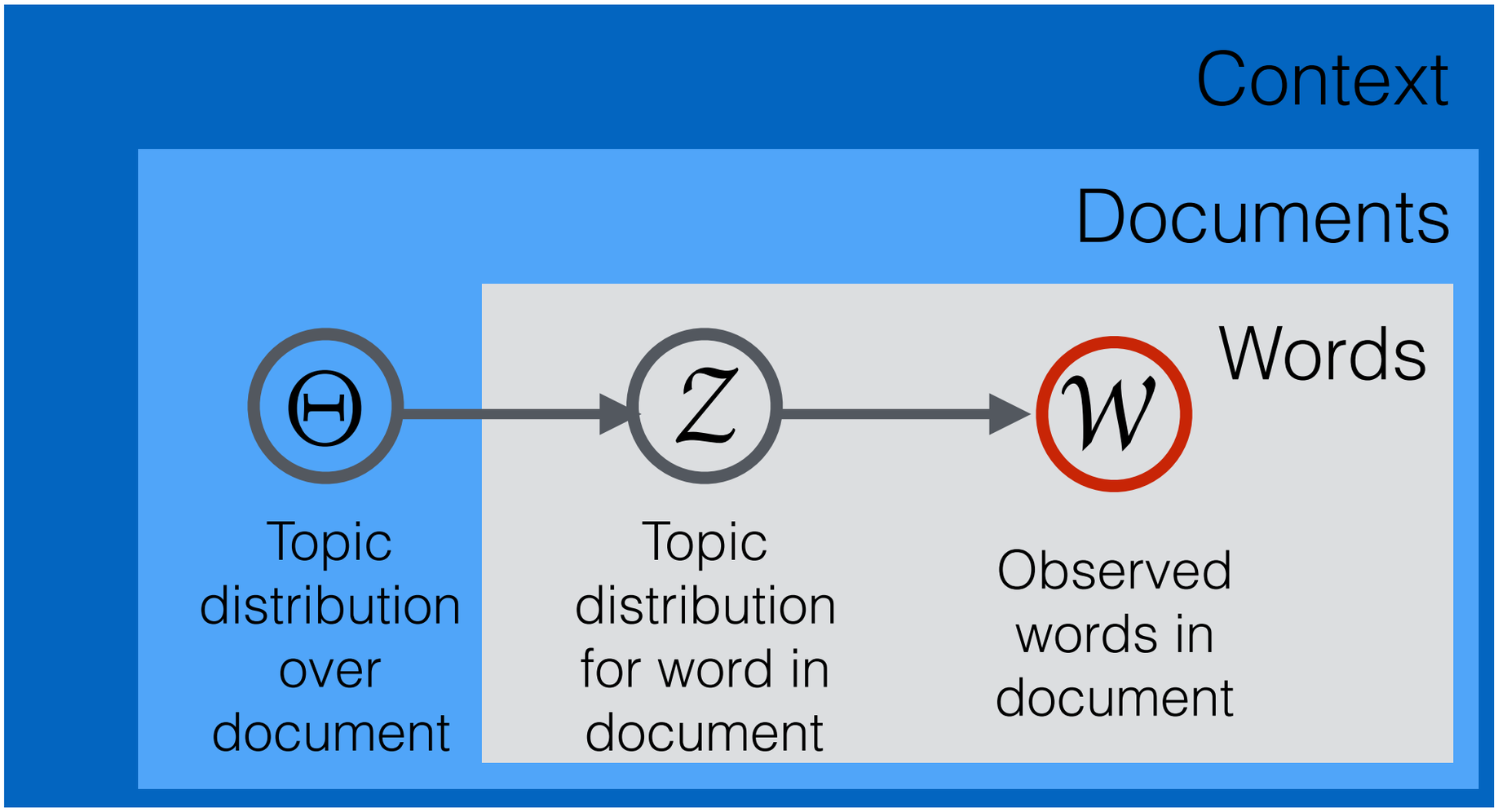


Dirichlet Distribution

- 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

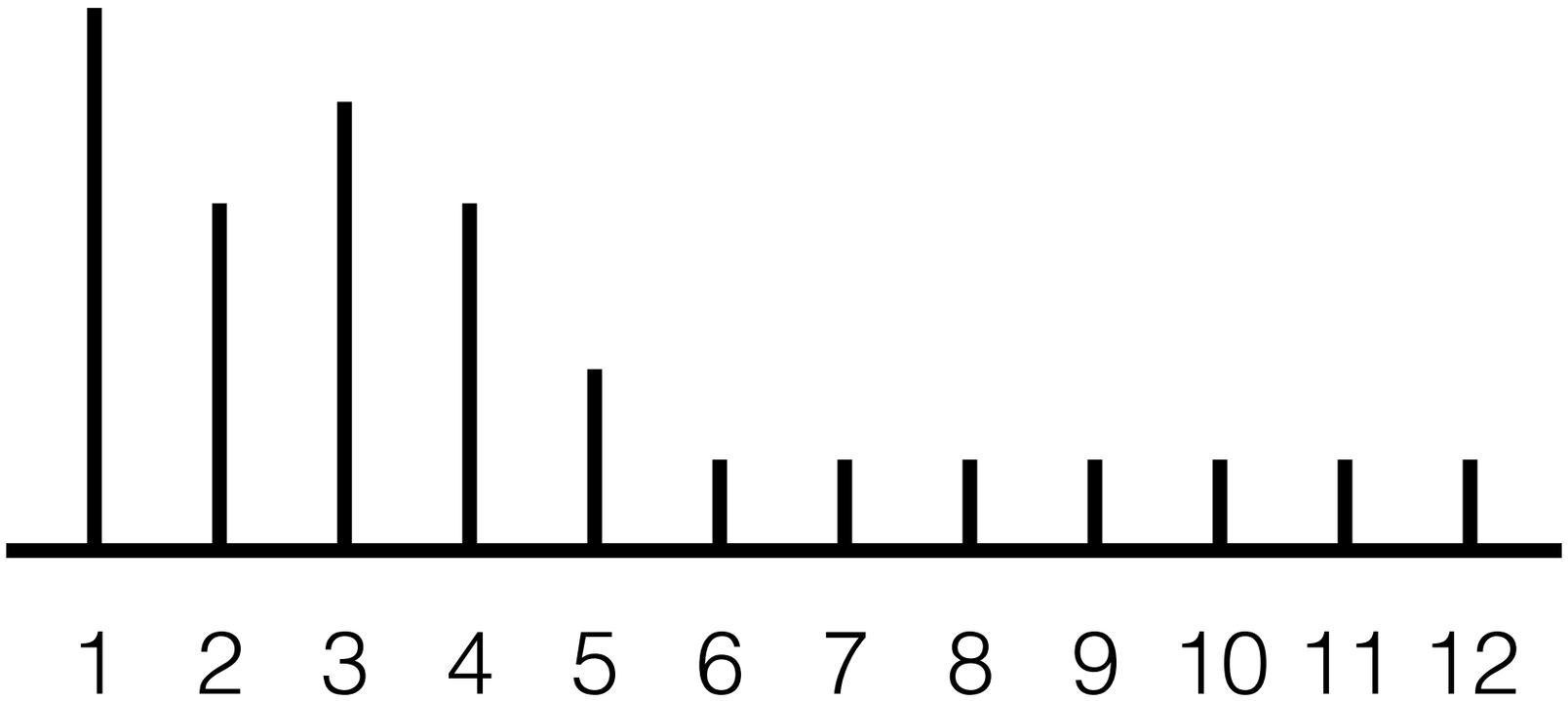
	Term1	Term2	Term3
Doc1			
Doc2			
Doc3			

	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