

Natural Language Processing

Natural Language Processing (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

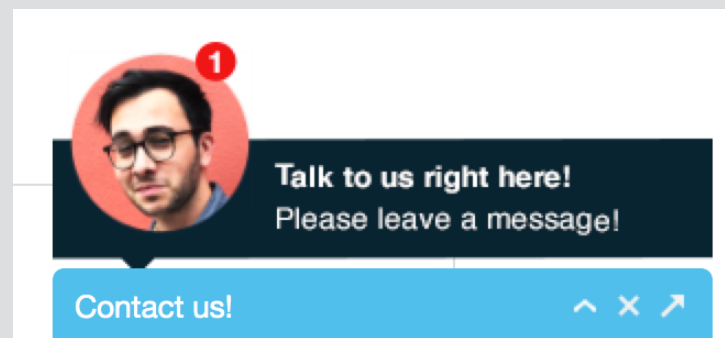
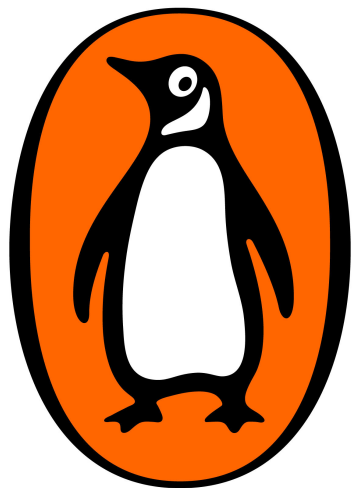
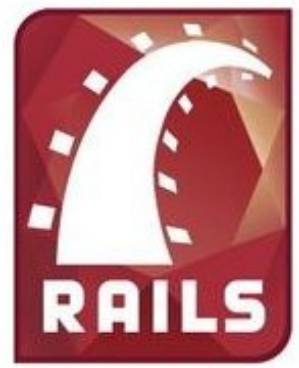


Processing

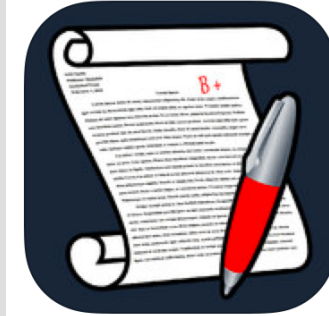
Statistical

- Infer rules from data
- IBM

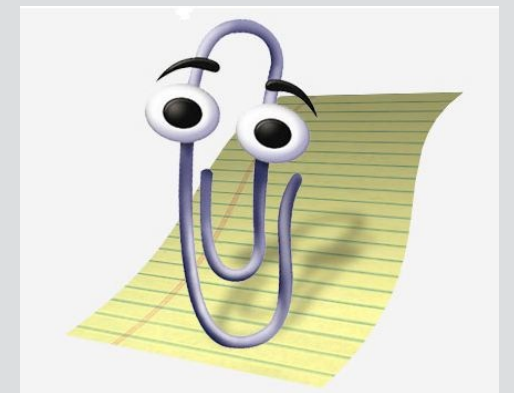
Industry



Education



iSTART:



GLENCOE ONLINE ESSAY GRADER
powered by Bookette SkillWriter™

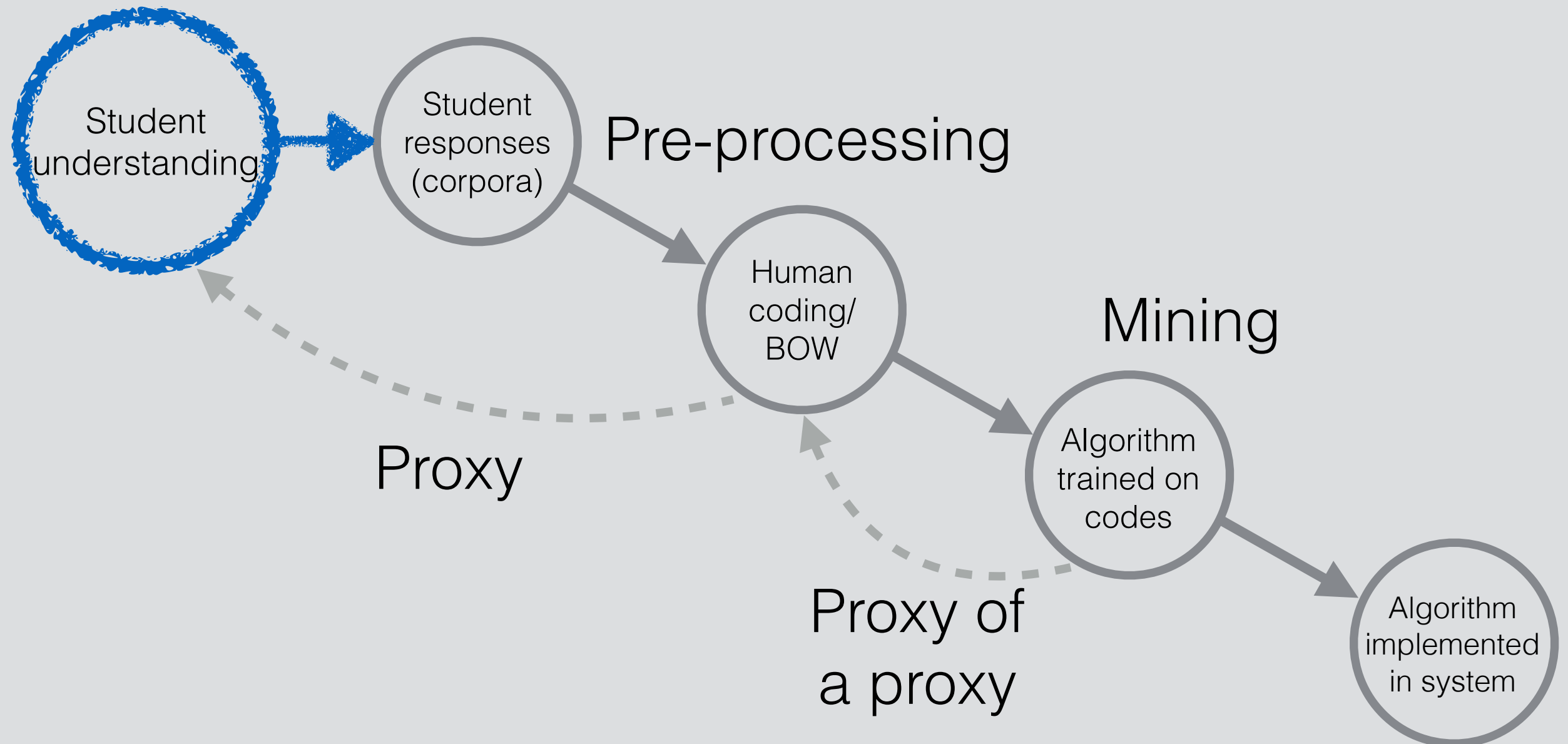


Essential Problem

- 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

Latent trait



Coding

Word counting



Google books Ngram Viewer

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

Types of Expressions

“I don’t know...”

“I dunno...”

Stemming

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


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

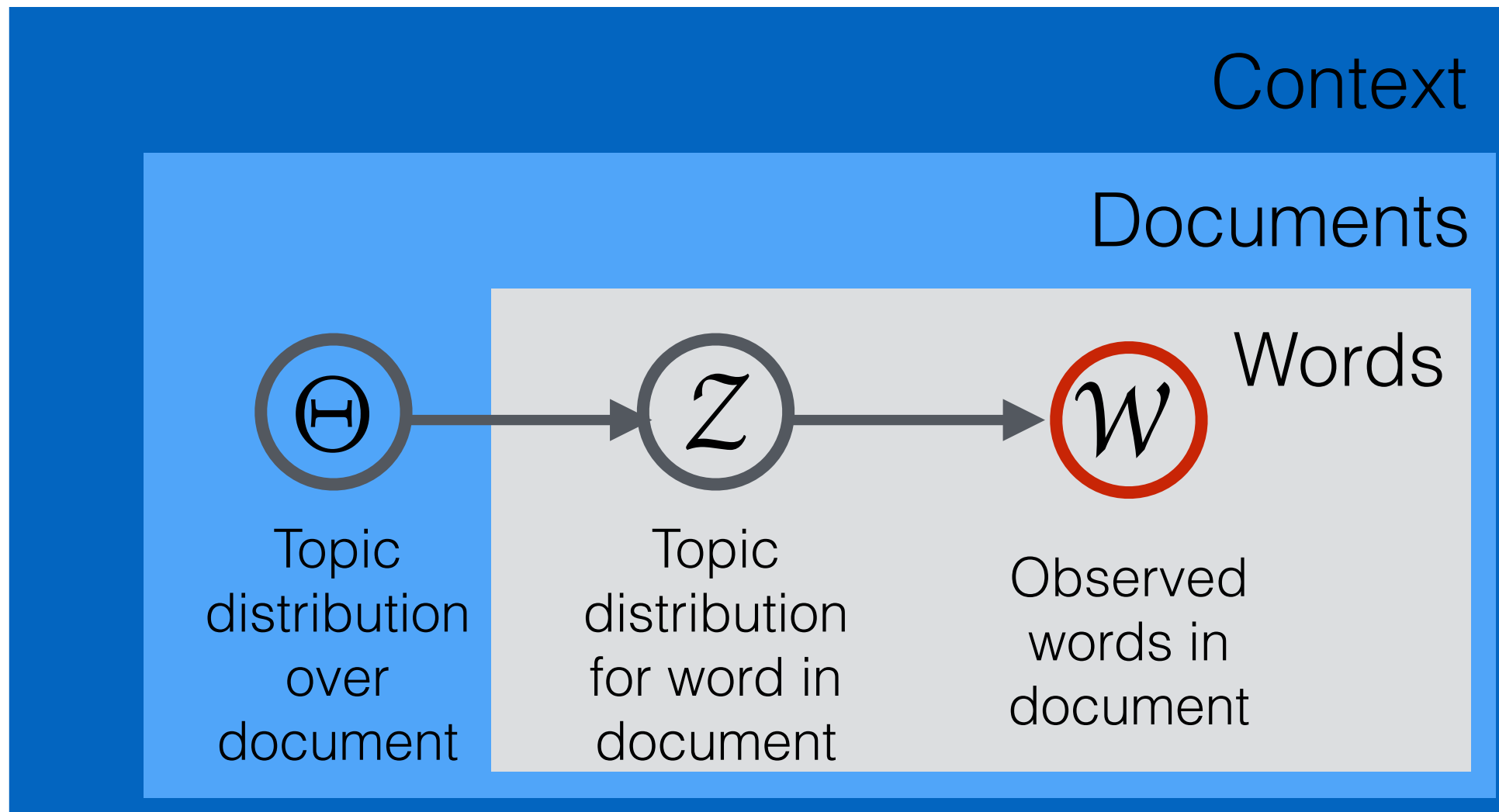
Topic Modeling with Latent Dirichlet Allocation (LDA)

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



Organizing Words

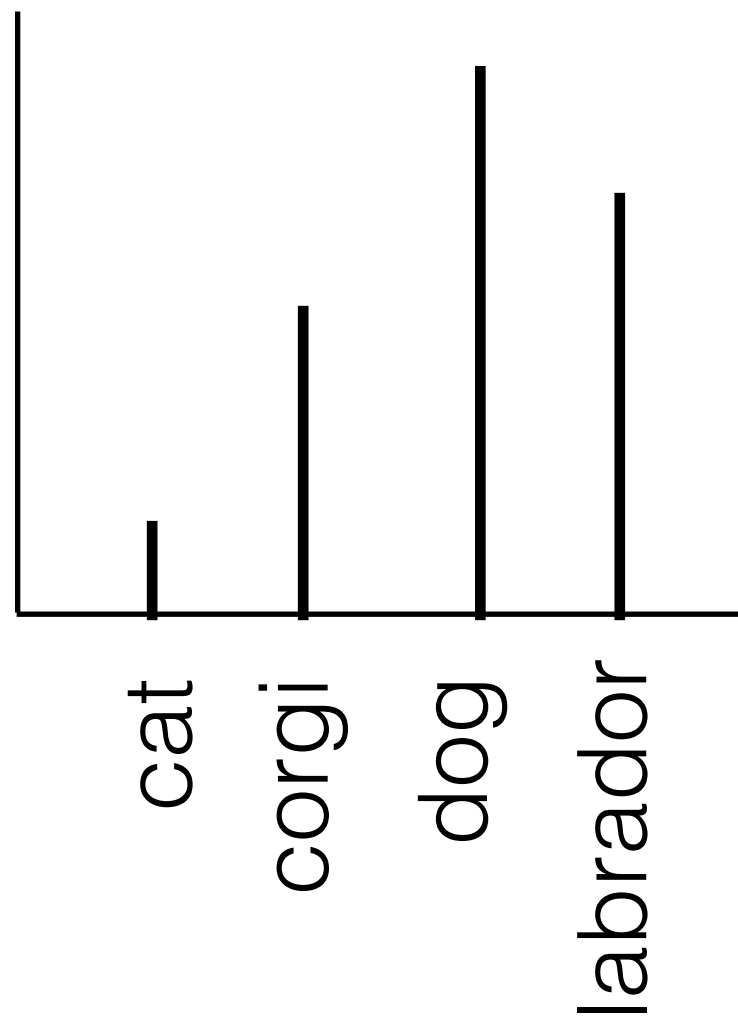
	cat	dog
doc 1	0.01	0.02
doc 2	0.01	0.03
doc 3	0.00	0.00

	cat	dog	corgi
cat		0.2	0.0
dog	0.2		0.5
corgi	0.0	0.5	

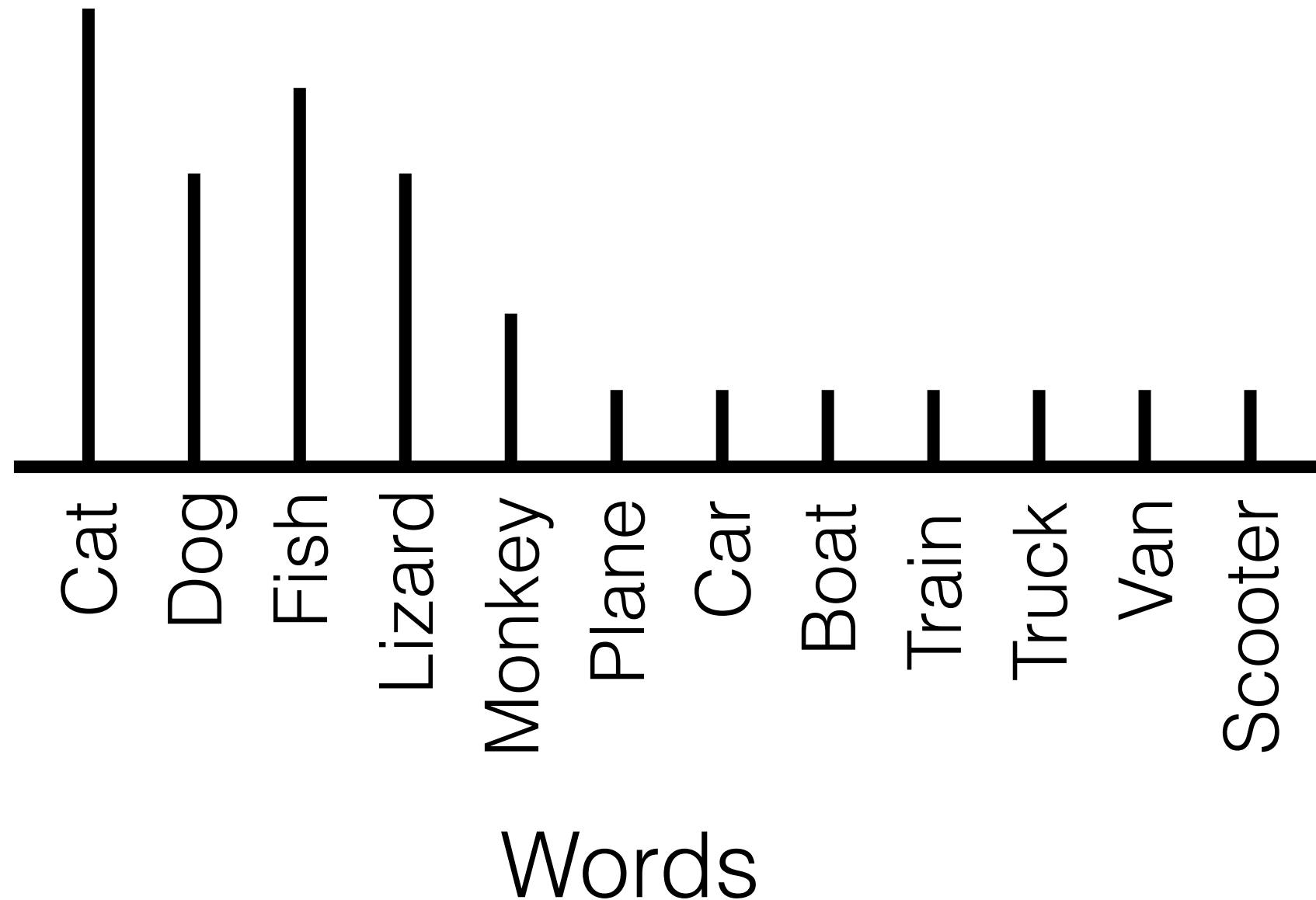
	cat	dog	corgi
doc 1	0.01	0.02	0.02
doc 2	0.01	0.03	0.03
doc 3	0.00	0.00	0.02

Topics (Z)

- A topic is a probability distribution over words



Topic Distribution for a Document



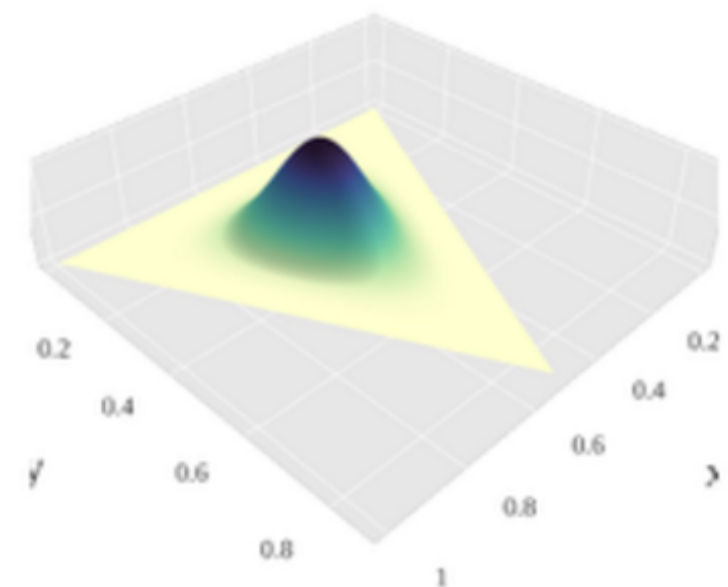
A document can be described by a recipe of topics and “how much” of each topic it contains

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

Dirichlet Distribution

- Peter Gustav Lejeune Dirichlet
- 1805 - 1859
- German mathematician
- Helped develop the definition of the *word function*
- Distribution on probability distributions
- Distribution over words, over documents



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

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

We can predict the topic of a
document from the words it
contains

Project

To use student notes to
process text data, map
sentiment and then discover
topics