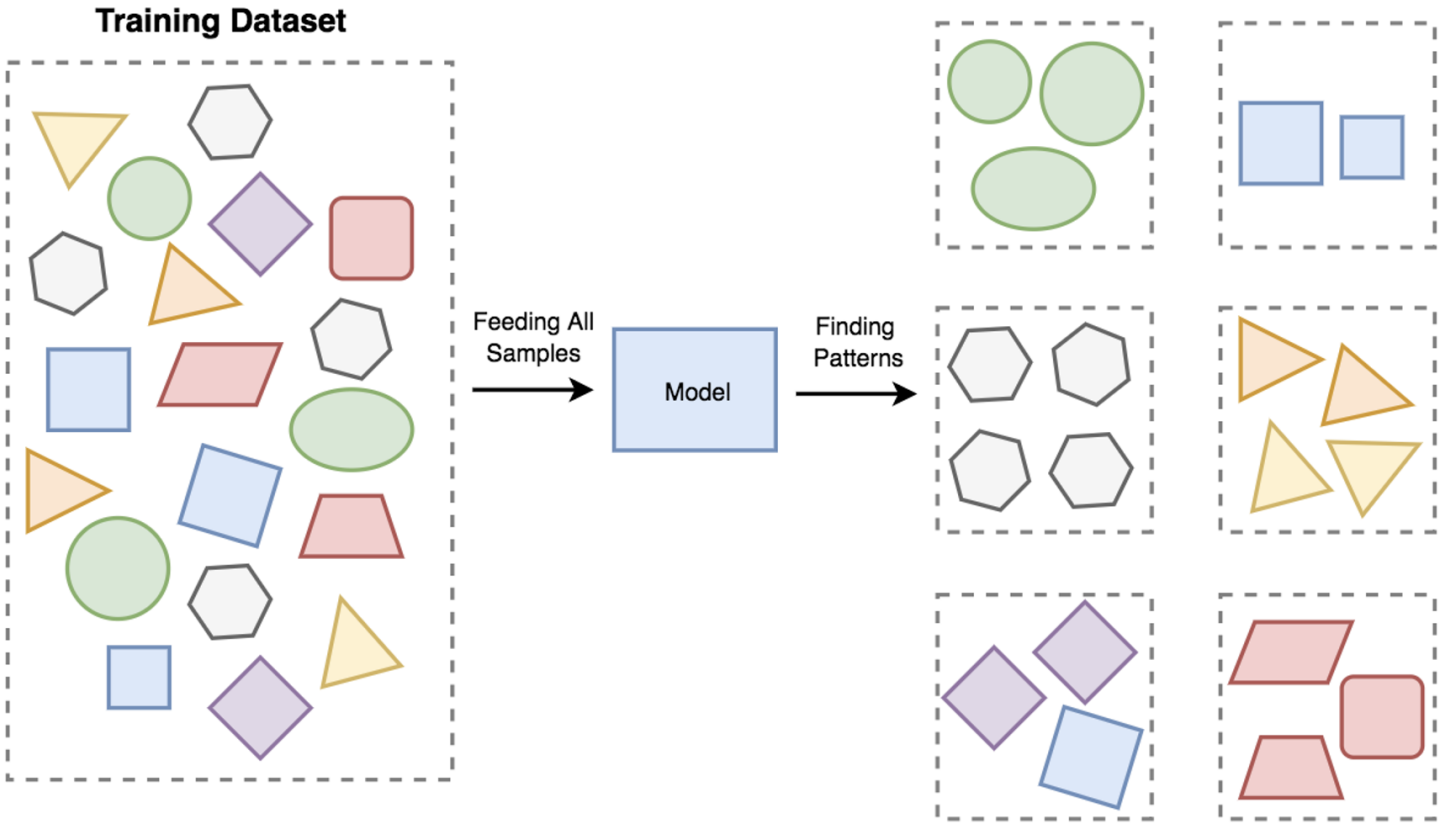
Topic Modeling

Ayoub Bagheri

Lecture plan

1. Text clustering
2. Probabilistic topic modeling
3. Latent Dirichlet allocation

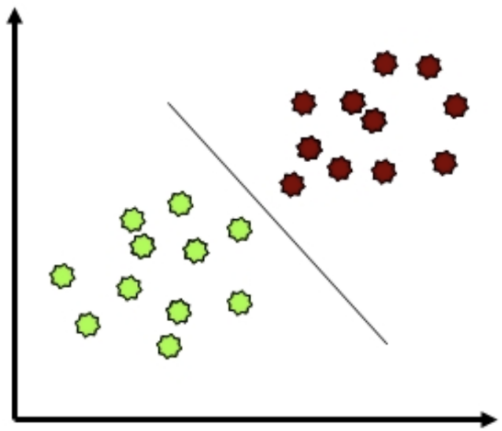
Unsupervised learning



Clustering versus classification

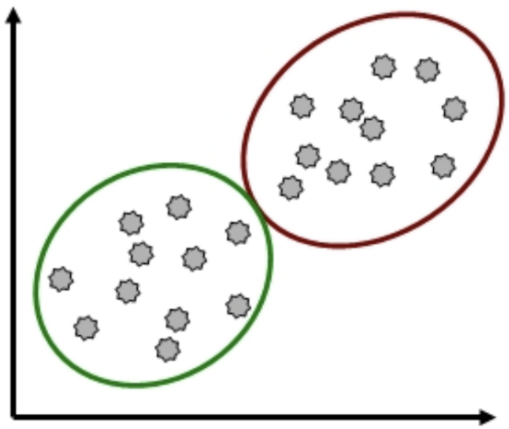
CLASSIFICATION

- Labeled data points
- Want a "rule" that assigns labels to new points
- Supervised learning



CLUSTERING

- Data is not labeled
- Group points that are "close" to each other
- Identify structure or patterns in data
- Unsupervised learning



->

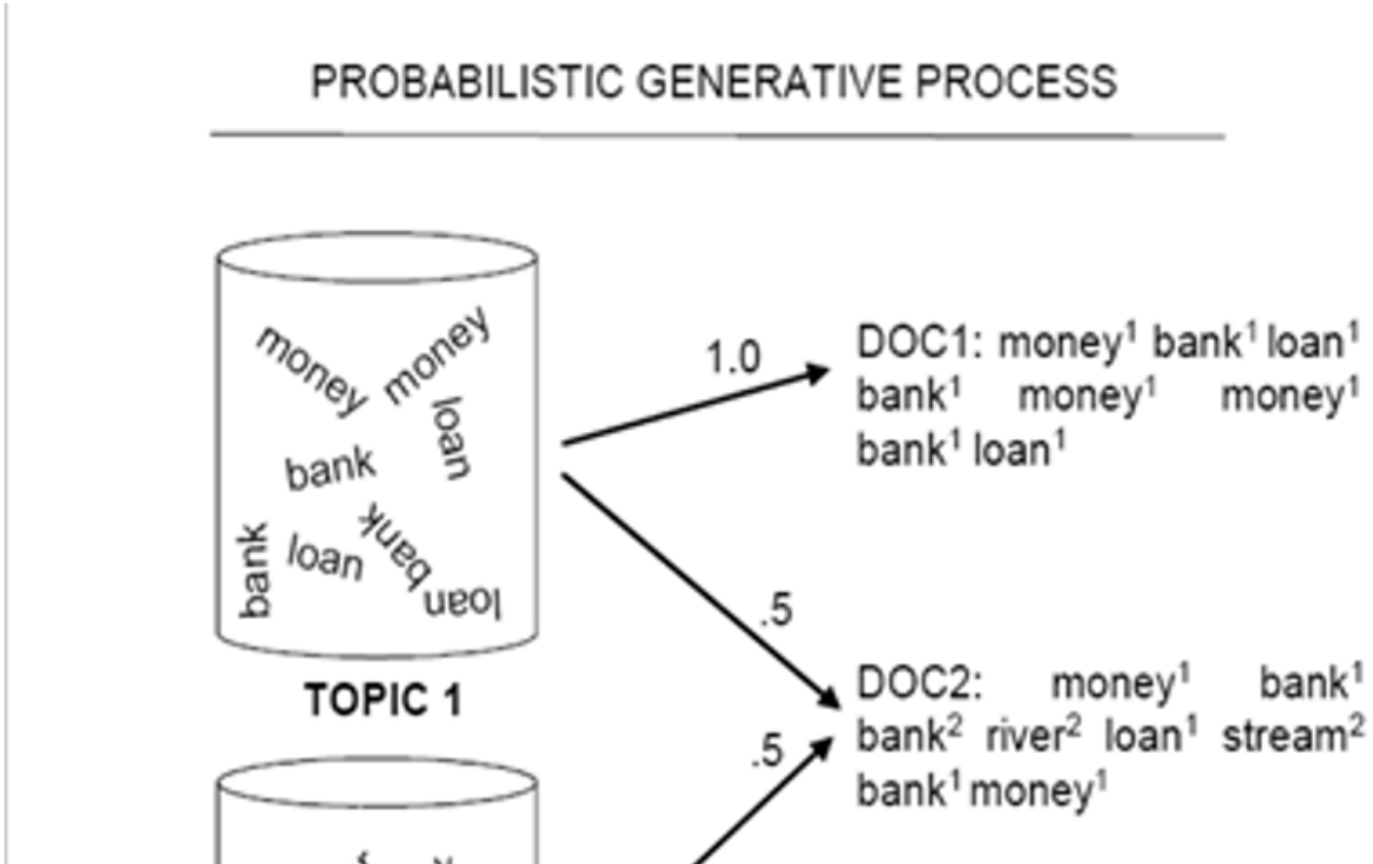
-> ->

-> -> ->

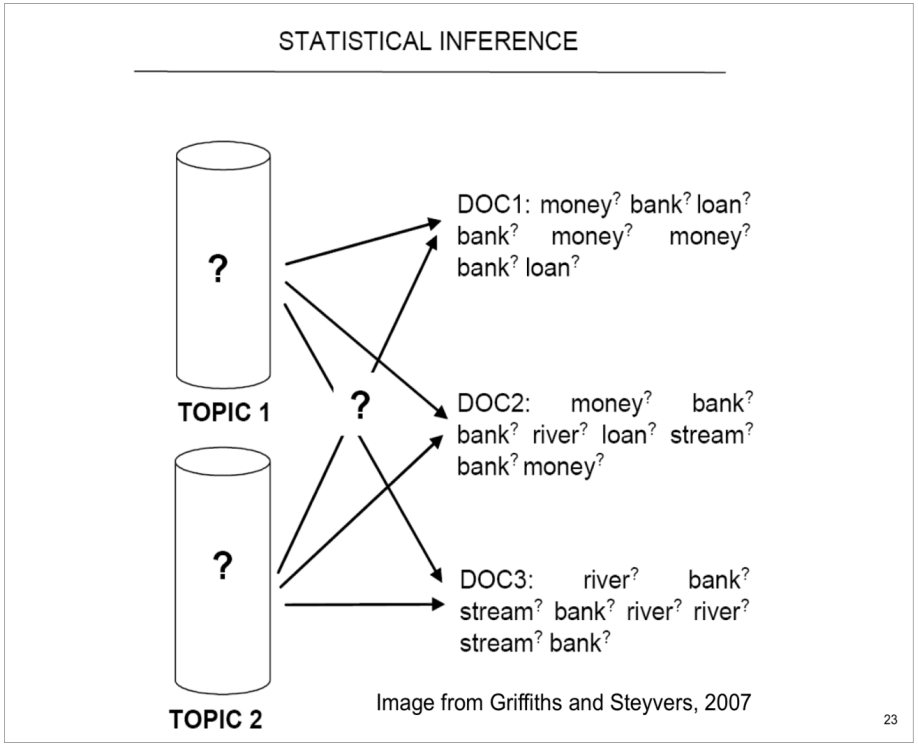
Topic Modeling

Topic models

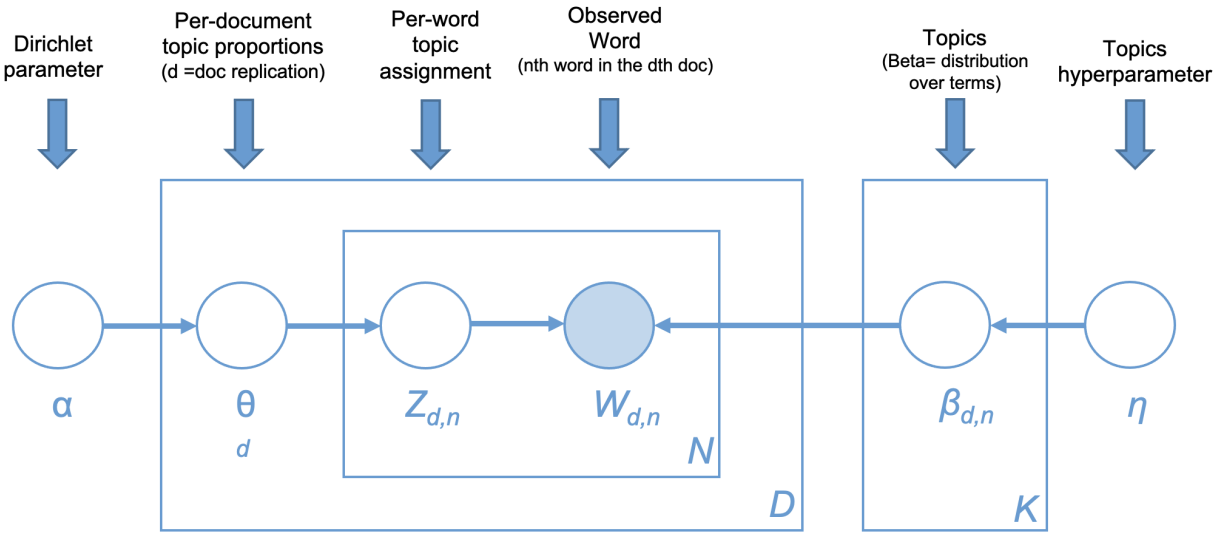
- Three concepts: words, topics, and documents
- Documents are a collection of words and have a probability distribution over topics
- Topics have a probability distribution over words
- Model:
  - Topics made up of words used to generate documents



Topic models | Reality: Documents observed, infer topics



LDA graphical model



Graphical model representation of LDA. The boxes are "plates" representing replicates.
The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

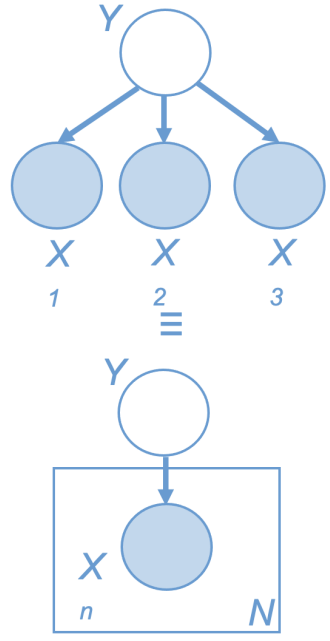
- Nodes are random variables
- Edges denote possible dependence
- Observed variables are shaded
- Plates denote replicated structure

K	specified number of topics
k	auxiliary index over topics
V	number of words in vocabulary
v	auxiliary index over topics
d	auxiliary index over documents
N_d	document length (number of words)
i	auxiliary index over words in a document
α	positive K-vector
β	positive V-vector
$Dir(\alpha)$	a K-dimensional Dirichlet
$Dir(\beta)$	a V-dimensional Dirichlet
z	Topic indices: $z_{d,i} = k$ means that the i-th word in the d-th document is assigned to topic k

Plates

D = docs
N = words
K = topics

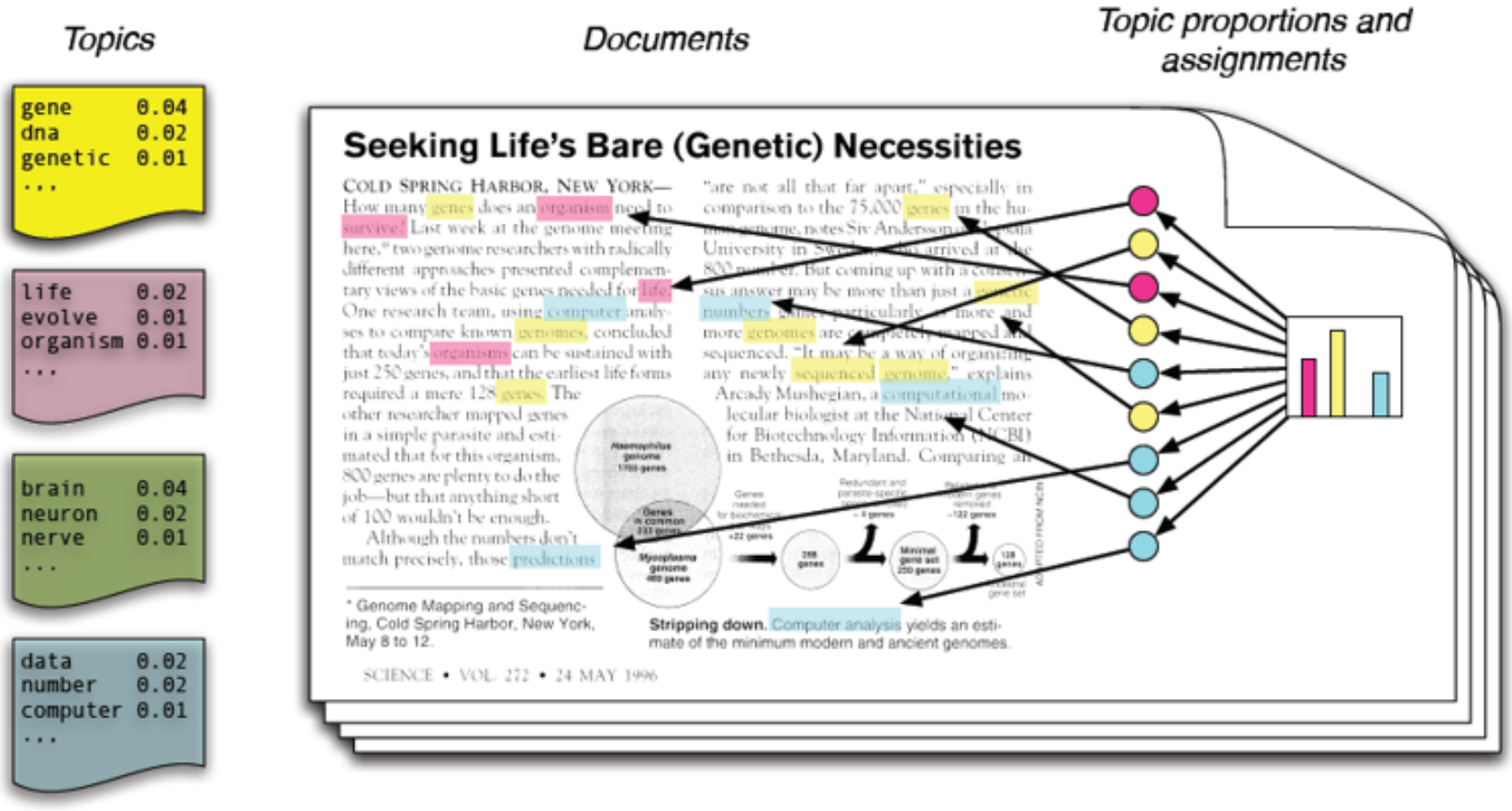
Graphical models



Probabilistic modeling

1. Treat data as observations that arise from a generative probabilistic process that includes hidden variables: For documents, the hidden variables reflect the thematic structure of the collection.
2. Infer the hidden structure using posterior inference: What are the topics that describe this collection?
3. Situate new data into the estimated model: How does this query or new document fit into the estimated topic structure?

LDA: Identifying structure in text



Cluster Validation

Desirable properties of clustering algorithms

- Scalability
  - Both in time and space
- Ability to deal with various types of data
  - No/less assumption about input data
  - Minimal requirement about domain knowledge
- Interpretability and usability

What is a good clustering?

- Internal criterion: A good clustering will produce high quality clusters in which:
  - the intra-class (that is, intra-cluster) similarity is high
  - the inter-class similarity is low
  - The measured quality of a clustering depends on both the document representation and the similarity measure used

Cluster validation

- Criteria to determine whether the clusters are meaningful
  - Internal validation
    - Stability and coherence
  - External validation
    - Match with known categories

Internal validation

- Coherence
  - Inter-cluster similarity v.s. intra-cluster similarity
  - Davies–Bouldin index

We prefer smaller DB-index!

External criteria for clustering quality

- Quality measured by its ability to discover some or all of the hidden patterns or latent classes in gold standard data
- Assesses a clustering with respect to ground truth ... requires labeled data
- Assume documents with C gold standard classes, while our clustering algorithms produce K clusters, $\omega_1, \omega_2, \ldots, \omega_K$ with n_i members.

Summary

Summary

- Text clustering
- In clustering, clusters are inferred from the data without human input (unsupervised learning)
- Topic modeling

Practical 6