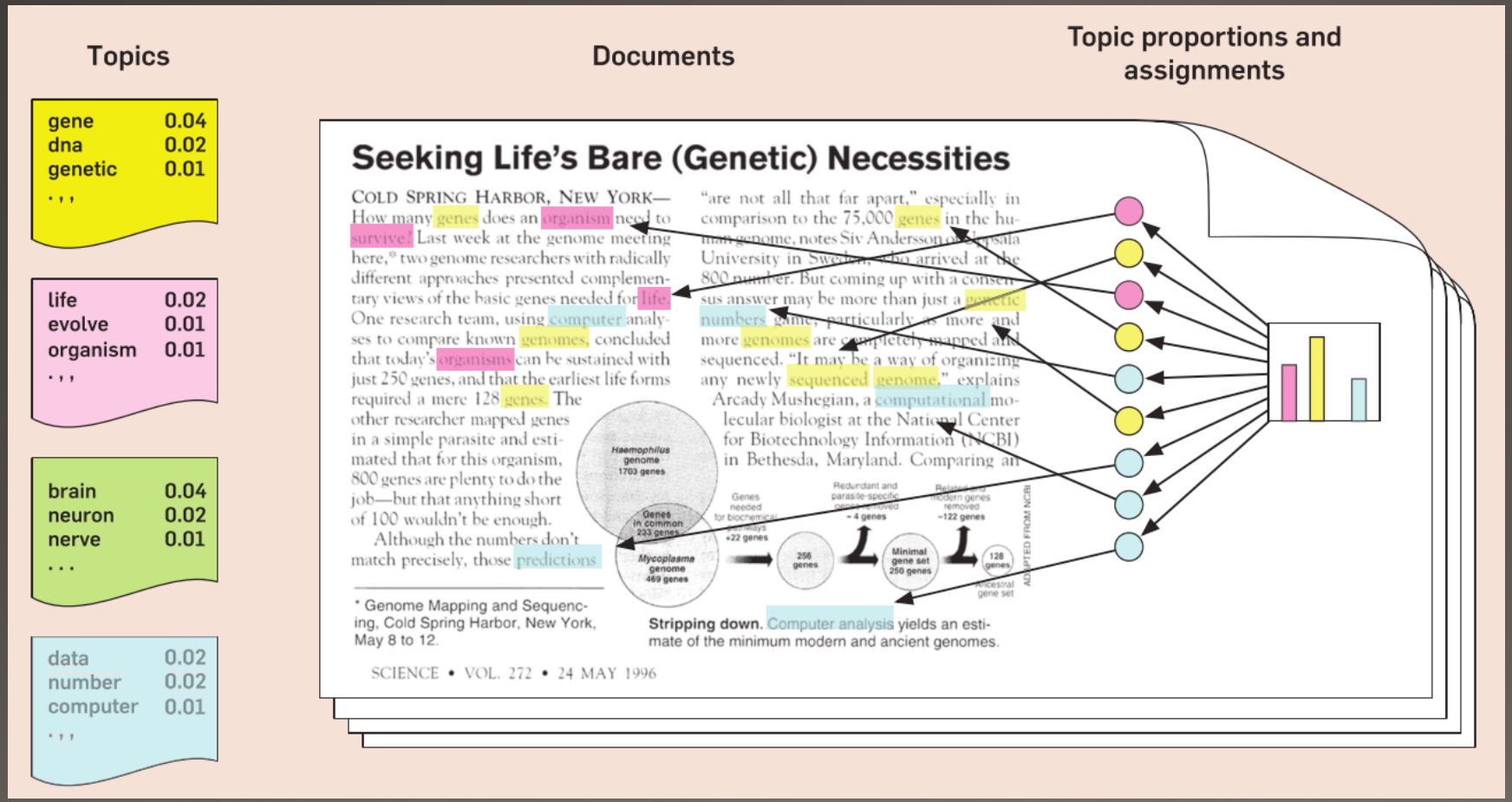
Mohamed Amin
mohamedamin.ca

Topic Modeling

What is Topic Modeling?

- A set of statistical techniques and algorithms that allow the thematic discovery of information in archives of documents
- A type of unsupervised machine learning – No prior labels
- Corpus is composed of documents
- Each document is a distribution over topics
- Each topic is a distribution over words

Generative Process



Types

- Latent Semantic Indexing/Analysis (LSA/LSI)
- Probabilistic Latent Semantic Indexing/Analysis (pLSA/pLSI)
- Latent Dirichlet Allocation (LDA)
- Dynamic Topic Models (DTM)
- Supervised Latent Dirichlet Allocation (sLDA)
- Relational Topic Models (RTM)
- Hierarchical Dirichlet Process (HDP)

Use cases

- Research publications
  - Citation influence impact, history of ideas, recommending scientific articles
- News articles
  - Summarizing news articles, comparing multiple news sources, topic trends
- Author analysis
  - Finding influential authors, ranking of web content based on author
- Web content
  - Analyzing online reviews, sentiment analysis
- Software Source Code
  - Evolution of code, developer contributions using author topic analysis
- Other types of data
  - Patterns in voice, genetic data, images, social networks

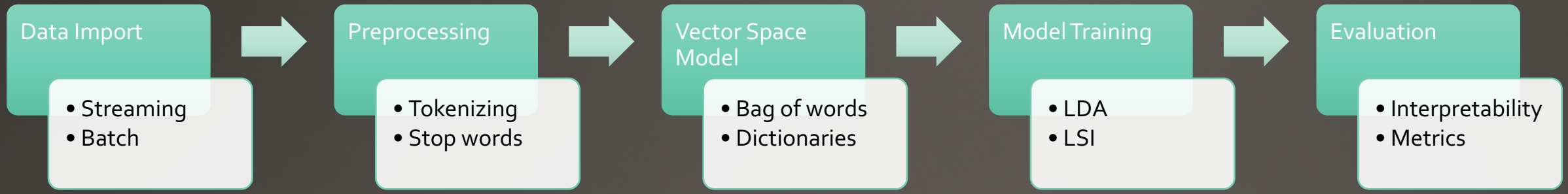
Tools

- Python – Gensim
- R – topicmodels package
- R– lda package
- Java – Mallet
- C – lda-c
- Matlab – Topic Modeling Toolbox

Evaluation

- Human in the loop
  - Word intrusion
  - Topic intrusion
  - Interpretability
- Metrics
  - Perplexity
  - Cosine similarity
  - Tokens per topic
  - Document Entropy
  - Coherence

Pipeline


Data Import
• Streaming
• Batch
Preprocessing
• Tokenizing
• Stop words
Vector Space Model
• Bag of words
• Dictionaries
Model Training
• LDA
• LSI
Evaluation
• Interpretability
• Metrics

Gensim – topic modeling for humans

- Open Source – GNU LGPLv2
- Document streaming
- Memory independent
- Implements LDA, LSI, DTM, HDP,
- Deep learning using word2vec and doc2vec
- Wrappers around other implementations such as Mallet