Topic Evolution in Life Sciences Research

Jingfei Cai
September 7th, 2017

Agenda

- Objectives
- Data Extraction
- Data Analysis
- Next Steps
- Key Takeaways

Objectives

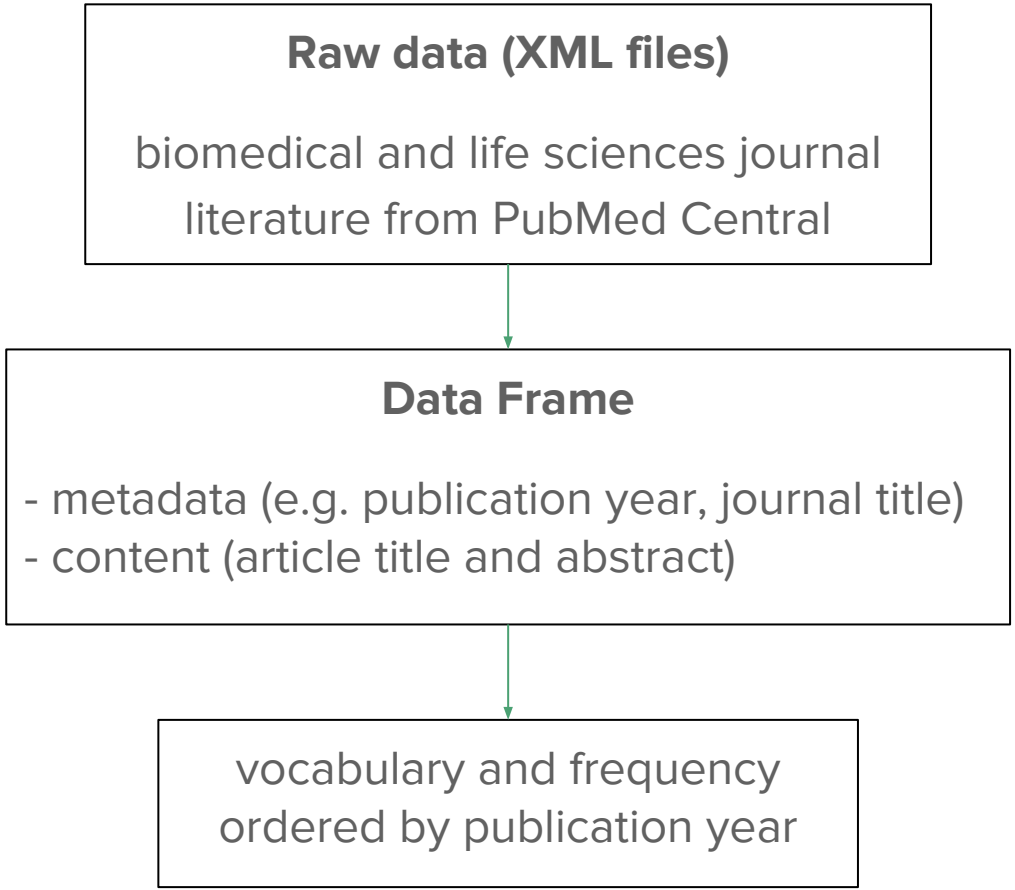
Use text from biomedical and life science literature to gain insights on research topic trends over time

- Discover underlying themes
- Track changes over time

Agenda

- Objectives
- **Data Extraction**
- Data Analysis
- Next Steps
- Key Takeaways

Data Extraction



Agenda

- Objectives
- Data Extraction
- **Data Analysis**
  - Exploratory Data Analysis
  - Dynamic Topic Modeling
  - Interpretation of Results
- Next Steps
- Key Takeaways

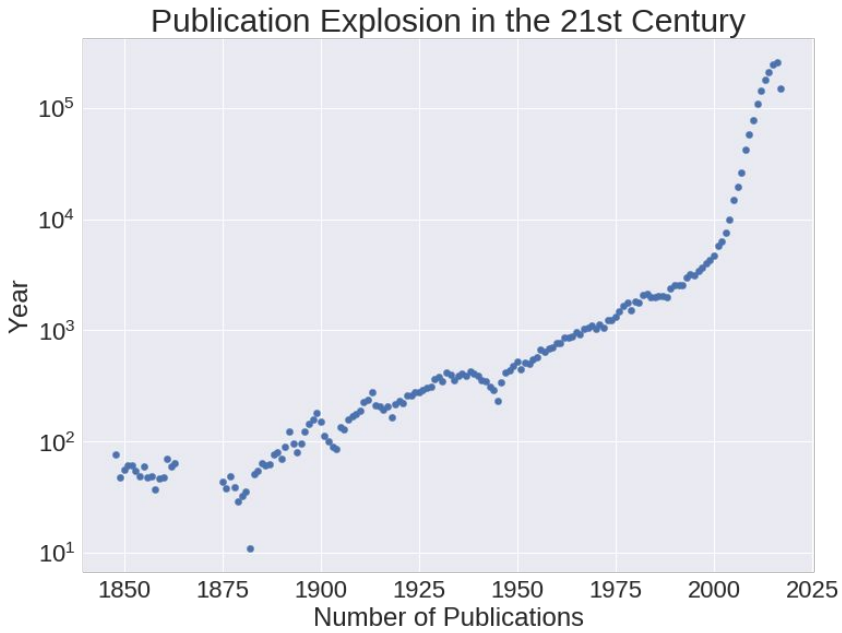
Exploratory Data Analysis

Comprehensive collection:

- 1,669,759 articles
- 8,462 journals
- 1848 – 2017 (159 unique years)

Explosion of electronic document archives

- A wealth of information!
- But how to process them?



Dynamic Topic Modeling

Probabilistic time series models

Capture evolution of topics in sequentially organized corpus

Assumptions of **static topic model**
(e.g. latent Dirichlet allocation (LDA)):

- Words of each document are independently drawn from a mixture of “topics”
- Mixing proportions are randomly drawn for each document
- The topics are shared by all documents (!!)

Assumptions of **dynamic topic model**:

- Data is divided by time slice (e.g. by year)
- Documents of each slice has *k*-component topics, which are evolved from the topics associated with the previous time slice

Interpretation of Results

Articles from *The Journal of Cell Biology*

23,896 articles from 1962 through present (2017)

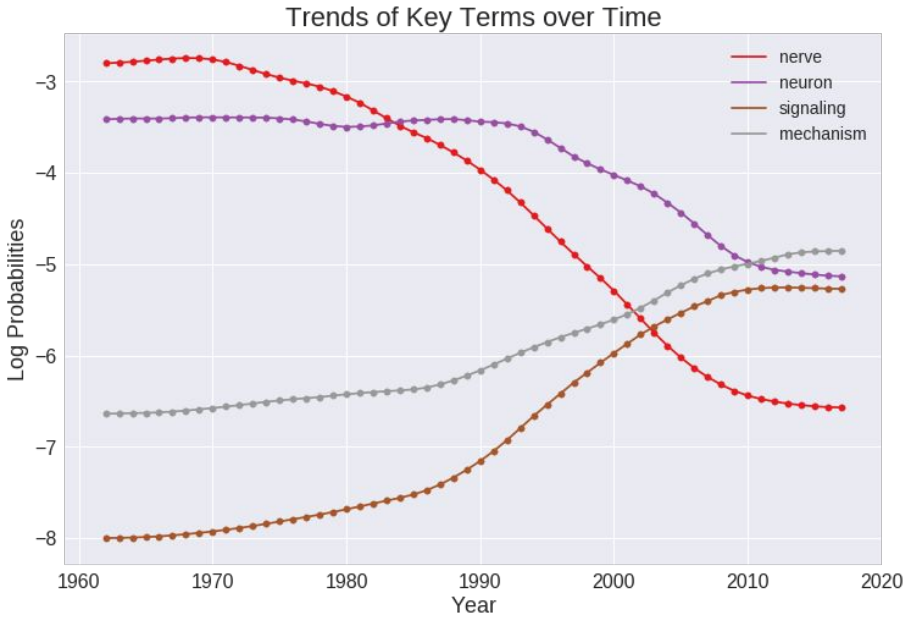
2.2 million words

6,619 words in vocabulary after pruning

Estimated 10-component dynamic topic model:

- cytoskeletal systems
- inter-cell communications
- nucleus, cell replications and cycles
- inter- and intra-cell transport
- neuroscience
- cell signaling
- imaging techniques (esp. microscopy)
- gene transcription and translation
- cell/tissue cultures, cancer research
- mitochondria

“neuroscience”



1965: 'nerve', 'axon', 'neuron', 'myelin', 'sheath'

1975: 'nerve', 'axon', 'neuron', 'synaptic', 'terminal'

1985: 'cam', 'neuron', 'nerve', 'cell', 'axon', 'ngf'

1995: 'neuron', 'cell', 'axon', 'growth', 'apoptosis'

2005: 'cell', 'apoptosis', 'protein', 'neuron', 'death'

2015: 'cell', 'protein', 'function', 'cellular', 'mechanism'

Interpretation of Results

Articles from *The Journal of Experimental Medicine*

23,246 articles from its inception through present (1896–2017)

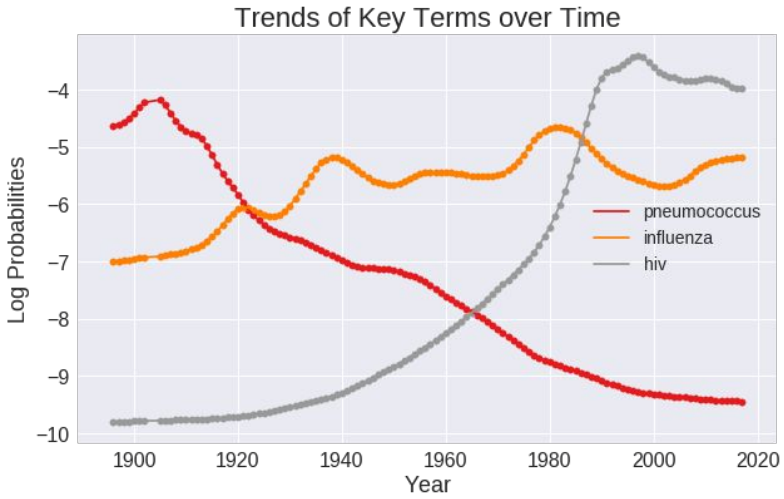
2.5 million words

6,790 words in vocabulary after pruning

Estimated 12-component dynamic topic model:

- infectious disease
- cancer
- inflammation
- immune response & immunization
- organ transplants
- development of immune cells
- mechanism of immune response
- serum composition
- mixed topics (??)
- circulatory system
- genetics
- cardiovascular disease

“infectious disease”



1900: 'bacillus', 'case', 'organism', 'culture', 'pneumococcus'

1910: 'bacillus', 'case', 'organism', 'infection', 'culture'

1920: 'infection', 'bacillus', 'case', 'organism', 'virus'

1930: 'virus', 'infection', 'disease', 'bacillus', 'inoculation'

1940: 'virus', 'mouse', 'infection', 'monkey', 'inoculation'

1950: 'virus', 'mouse', 'infection', 'strain', 'poliomyelitis'

1960: 'mouse', 'virus', 'infection', 'strain', 'infected'

1970: 'virus', 'mouse', 'infection', 'infected', 'strain'

1980: 'virus', 'mouse', 'infected', 'infection', 'strain'

1990: 'virus', 'infection', 'infected', 'hiv', 'mouse'

2000: 'infection', 'virus', 'hiv', 'infected', 'viral'

2010: 'infection', 'virus', 'viral', 'hiv', 'response'

Agenda

- Objectives
- Data Extraction
- Data Analysis
- **Next Steps**
- Key Takeaways

Next Steps

- Expanding Scope:
  - Extend analysis to the whole text mining collections from PubMed Central
  - Extend to full text of the articles
- Optimization:
  - Better ways to parse HTML elements to get cleaner text
  - Inspect XML files to extract as much relevant info as possible (less missing information)
  - Further pruning the vocabulary:
    - Domain-specific stop words
    - Domain-specific stemming
  - Domain-specific knowledge is needed to better interpret results/fine-tune models
- Comparison with other topic models
  - e.g. gensim LDA

Agenda

- Objectives
- Data Extraction
- Data Analysis
- Next Steps
- **Key Takeaways**

Key Takeaways

- Dynamic topic modeling technique, combined with Natural Language Processing, is a powerful tool for organizing and exploring a large collection of text documents
- When applied to biomedical and life science literature, it can aid researchers and curious laypersons alike to discover interesting themes and trends

Questions?