Data-driven Analysis of Society for Neuroscience (SfN) Abstracts

Mitchell Wortsman[1], Justin Young[1], Ruobing Xia[2], Shaobo Guan[2]
[1] Department of Computer Science, [2] Department of Neuroscience, Brown University

Introduction

- **Neuroscience:** a fast-growing, interdeciplinary research field.
- **SfN annual meeting:** the largest neuroscience conference.

- **Problem:**
  - Difficult to find a presentation of interest.
  - Difficult to get a big picture of the entire field.
- **Aims of this project:**
  - Explore and visualize the topic structure of presentation abstracts.
  - Test if the research topic can be predicted with a unigram model.
  - Build a recommendation system for attendees.
  - Get more insights: collaboration network & research trend.

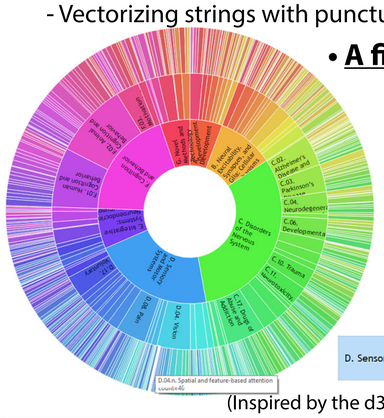
Data

- **Data source:**
  - SfN meeting planner, an online abstract browser. (https://www.sfn.org/annual-meeting/)

- **Data obtaining:**
  - Web scraping using python.
  - HTML cleaning.
  - Localizing each field (title, author...).
- **Data size:** >10,000 entries * 7 years.
- **Data structure:** an example entry.

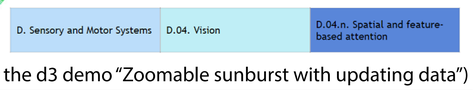
pres_title	topic	authors	institute	abstract
Clonally related interneurons disperse both within and across functional boundaries within the forebrain	A.01.c. Cell lineage	*C. MAYER, X. H. JAGLIN, C. L. CEPKO, S. HIPPENMEYER, G. FISHELL	Physiol. & Neurosci., NYU Neurosci. Inst., New York, NY	The medial ganglionic eminence (MGE) gives rise to the majority of mouse forebrain interneurons. Here, we examine the lineage relationships among MGE-derived interneurons using a replication-defective retroviral library containing a highly diverse set of DNA barcodes. Recovering the barcodes from the mature progeny of infected progenitor cells enabled us to unambiguously determine their respective lineal relationships. We found that clonal dispersion occurs across large areas of the brain and is not restricted by anatomical divisions. As such, sibling interneurons can populate the cortex, hippocampus and striatum. Importantly, we also revealed that the majority of interneurons were generated from asymmetric divisions of MGE progenitor cells,

- **Further cleaning and integration:**
  - Joined with topic names (officially assigned topics).
  - Vectorizing strings with punctuation and stopwords removed.



- **A first look at the data:**

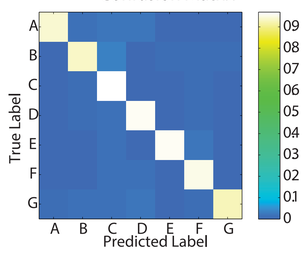
3-layer topics & word cloud by topic.



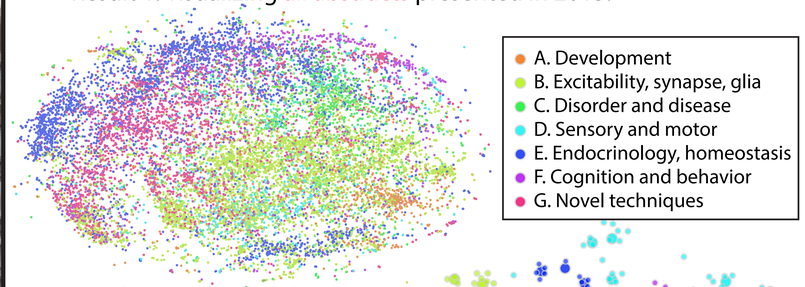
(Inspired by the d3 demo "Zoomable sunburst with updating data")

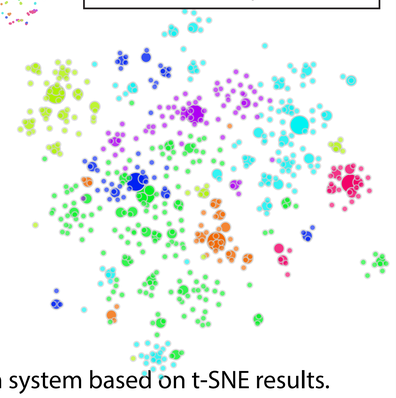
Topic Analysis

- **Supervised learning:**
  - Goal: to understand the extent to which abstracts exhibit patterns in unigram word frequency based on topic.
  - Method: support vector machine (SVM)
    - Training labels: top-level topics.
    - Training size: 70,000.
    - Testing size: 11,000.
    - Performance: 96% accuracy.
  - Conclusion: abstract topics can be well predicted using a unigram model.



- **Unsupervised learning:**
  - Goal: to reduce the dimensionality of the original unigram model so as to visualize and evaluate the topic structure of abstracts.
  - Method: t-distributed stochastic neighbor embedding (t-SNE).
  - Result 1. visualizing all abstracts presented in 2015.



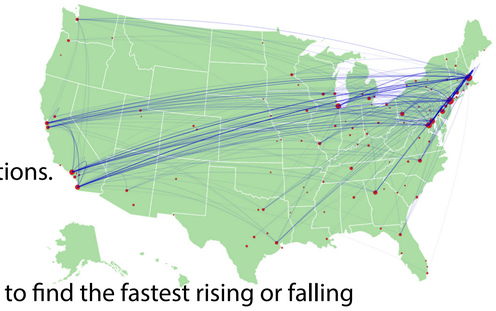


  - Result 2. "Anatomy of topics".
    - Taking averaged input vector for each topic category, and then run t-SNE.
    - Results demonstrated the hierarchical topic structures.
    - Provided more insights to the relationship between topics.
- **Abstract recommendation:**
  - Goal: to build a recommendation system based on t-SNE results.
  - System design: user types in an abstract ID, and the system returns the titles of the most similar/relavant abstracts.
  - Method: calculating the pairwise distance and find the nearest ones.
  - Example results:

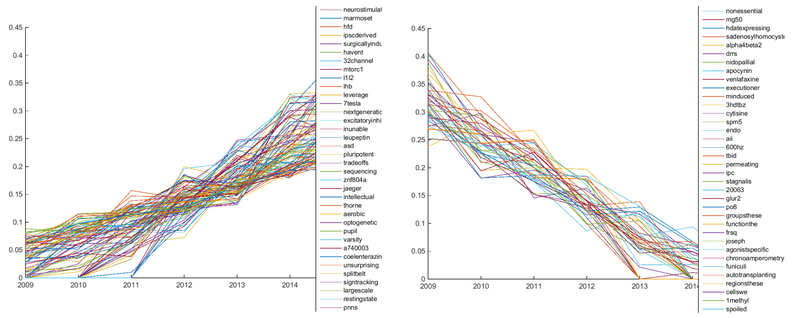
```
[Input]
Feature-based attention regulates long-range neural interactions in...  D.04
[Output]
Motion-direction tuning in the post-saccadic remapped response in m...  D.04
Response modulations by spatial but not by feature-based attention ...  D.04
Spatiotemporal characterization of perisaccadic receptive field str...  F.02
Inter-areal synchronization in visual cortex during a divided atten...  F.02
Maintenance of spatial information modulates the correlated variabi...  D.04
Feature-based attention modulates correlated BOLD activity in the v...  F.01
Distinct patches for gaze following and the passive vision of faces...  F.02
Attentional modulation of V1 neurons depends on physiological respo...  D.04
Distinct computational modules for forming central and peripheral p...  D.04
```

Geography & Trends

- **Collaboration geographic map:**
  - Collaboration: defined as co-authors for the same abstract.
  - Location information: extracted from affiliation lists.
  - Package used: DataMaps d3 library.
  - Bubble size: number of abstracts.
  - Line: substantial collaborations.



- **Research trends:**
  - Using linear regression to find the fastest rising or falling neuroscience terms across years.
  - Helped discover the recently hottest topics: [optogenetics, large-scale recording, 7-tesla, iPSC, resting state...]



Discussion

- **Summary:**
  - We conclude that unigram models can predict the topic of SfN abstracts, and can reveal the natural topic structures.
  - Based on our dimensionality reduction results, we have developed a simple but effective recommendation system.
  - We also gained insights about the geographical collaboration relationships and research trends.
- **Future direction:**
  - Use n-gram instead of unigram.
  - Explore other unsupervised learning tools or topic models.
  - Build an online recommendation system where user is able to find abstracts by putting in an abstract or a few keywords.
  - Build an online trend visualization platform where user can select a term of interest and get the change of its frequency over time.

* See all the details and interactive visualizations @ sfn.metaneuro.com

Acknowledgement:
Many thanks to SfN for the permission of using abstracts.
Thank you Prof. Kraska and TAs!!