A scalable tool for longitudinal Twitter analysis: understanding the impact of COVID-19 on public discourse

Walter Dempsey[25] Alfred Hero[134] Conrad Hougen[3] Brandon Oselio[2]
Wayne Wang[4]

[1]Department of Biomedical Engineering

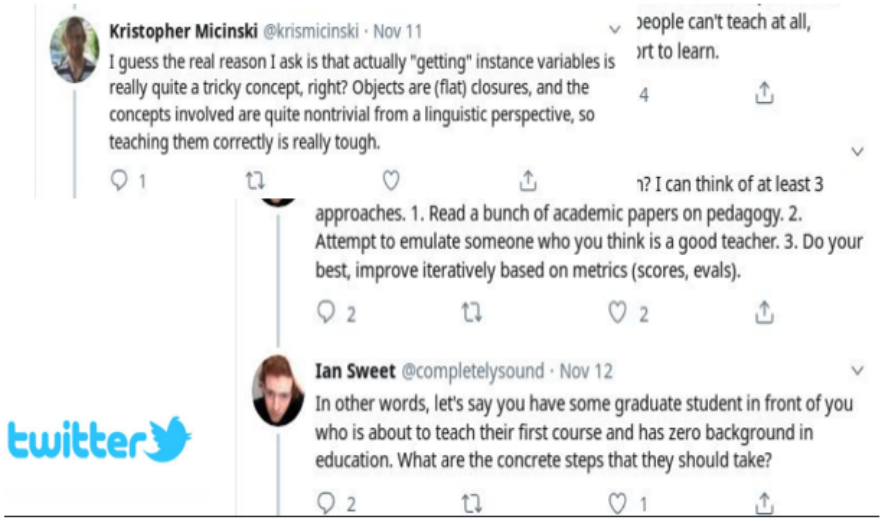
[2]Department of Biostatistics

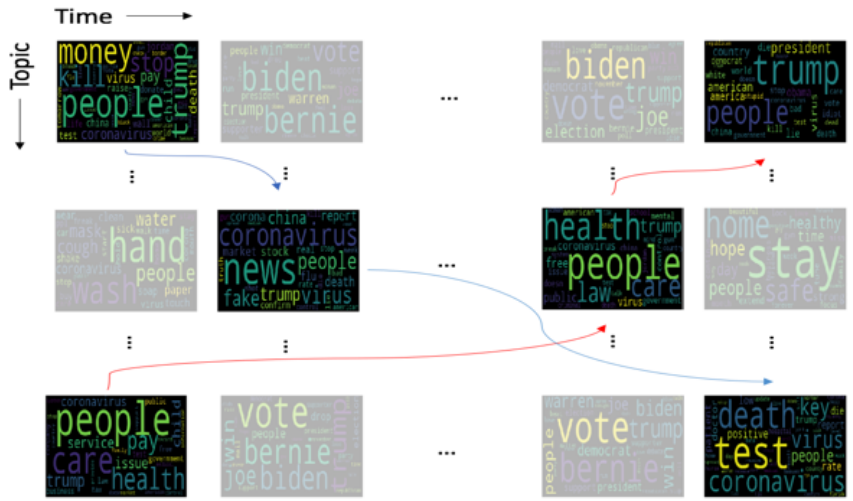
[3]Department of EECS

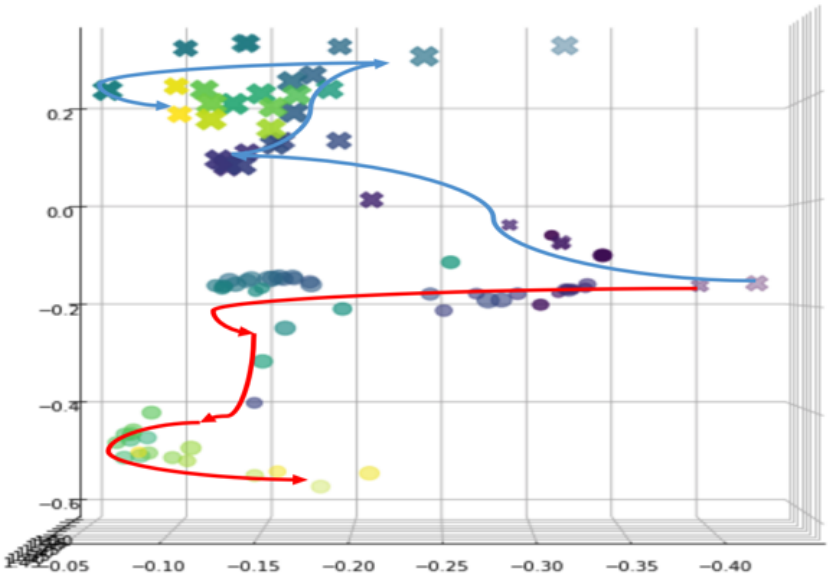
[4]Department of Statistics

[5]Institute of Social Research

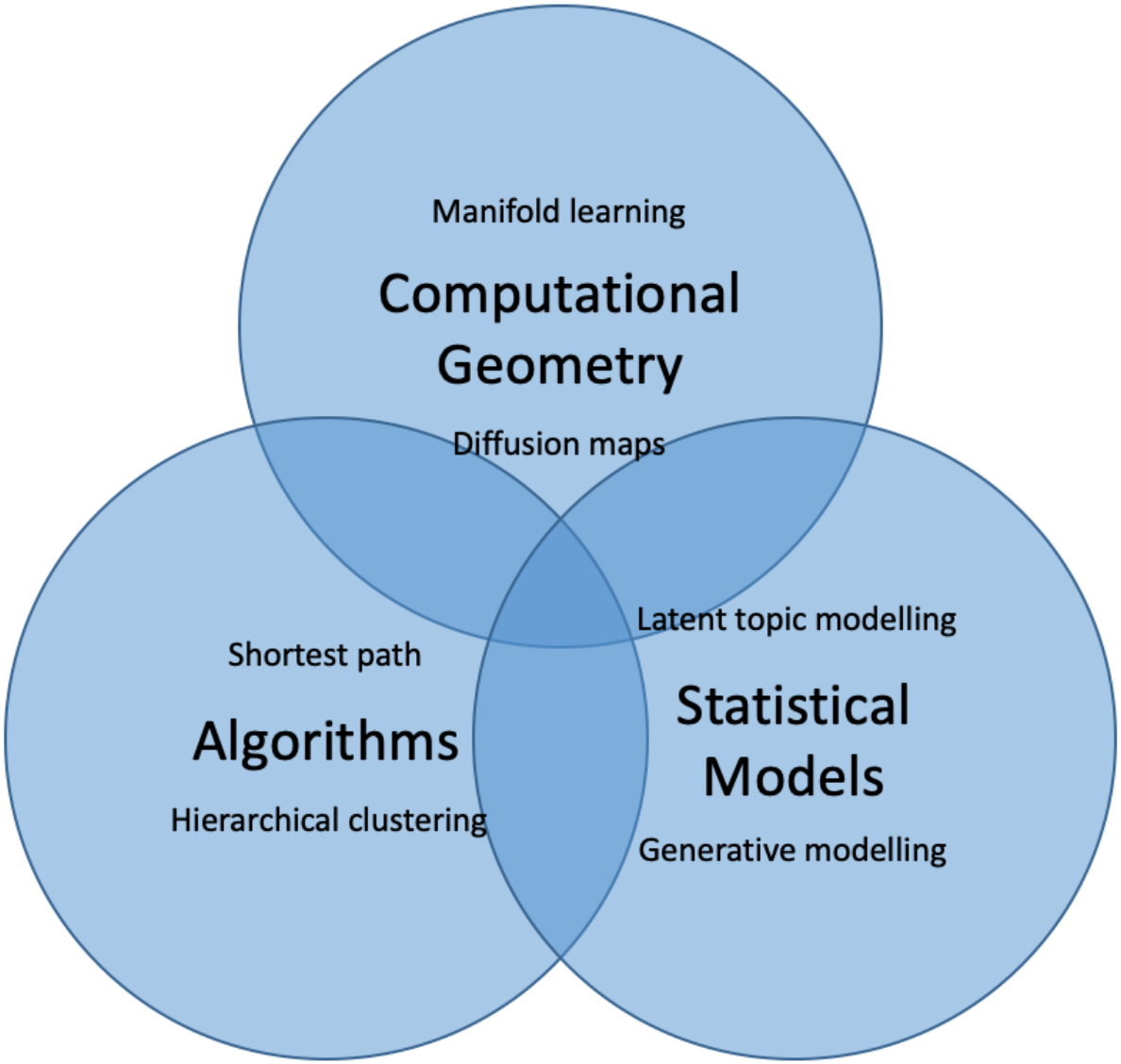
Overview of results



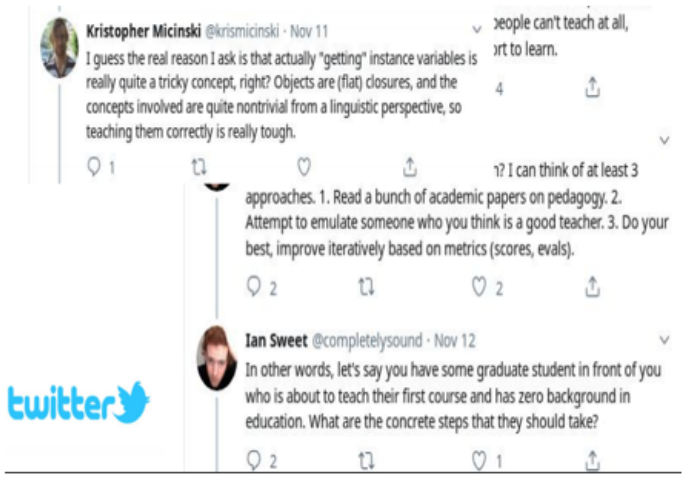




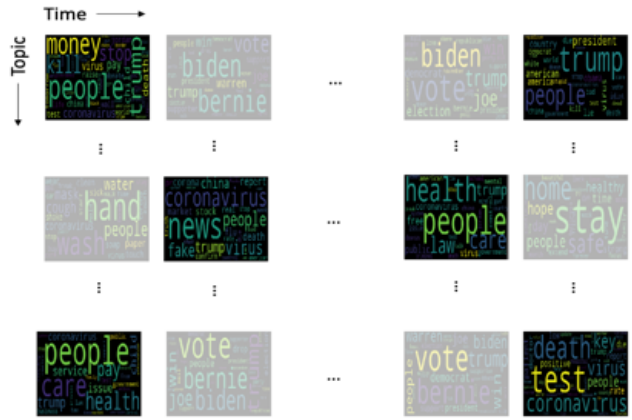
Data science tools



Roadmap

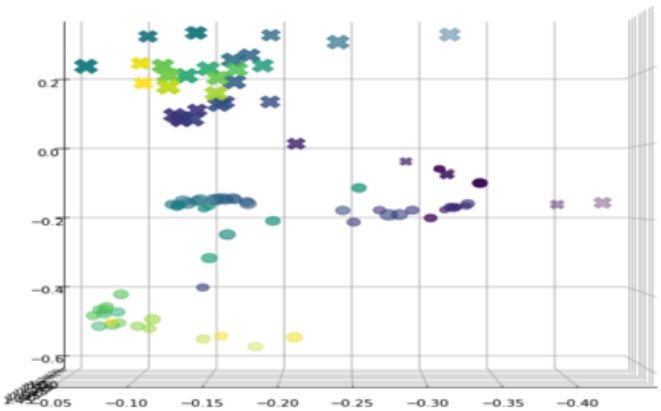


1. LDA on temporally smoothed corpus



2. Hellinger distances

3a. PHATE embedding



3b. Neighborhood graph and shortest path

4. Visualize and interpret multiple paths and clusters

Outline

Twitter decahose stream [1]

Decahose stream

The Decahose delivers a $\sim 10\%$ **random sample** of the realtime Twitter Firehose (300 to 500 million tweets per day) through a streaming connection. This is accomplished via a realtime sampling algorithm which randomly selects the data, while still allowing for the expected low-latency delivery of data as it is sent through the firehose by Twitter. One of the features available with Decahose: enhanced reliability - **geographic diversity** ($\sim 0.1\% - 0.5\%$ of sampled tweets contain geo-location info) of backend systems[a].

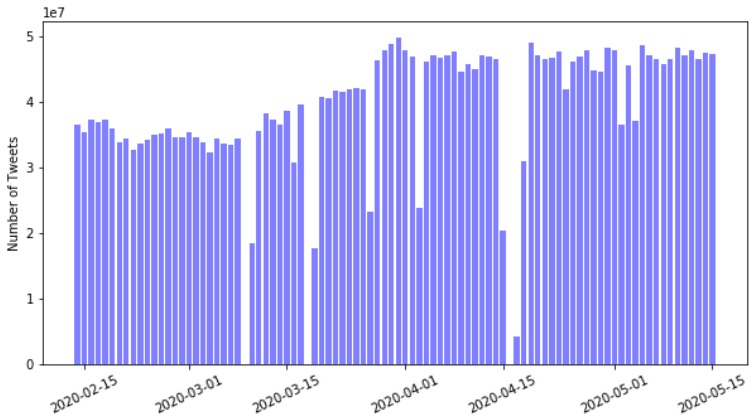
[a] *Decahose stream*. https://developer.twitter.com/en/docs/tweets/sample-realtime/overview/decahose. Accessed: 2020-04-20.

[1] Made available to us by MIDAS.

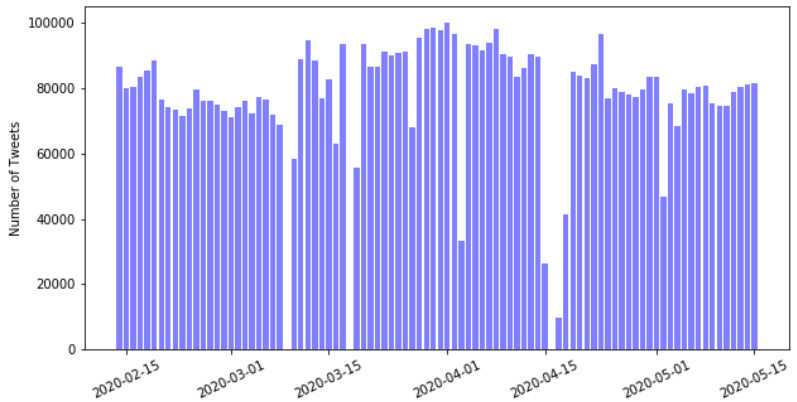
Subsampled tweets from Feb 15 to May 15

Three levels of subsampling:

- Geotagged US tweets: to study spatial variations of tweeting behavior.
  - English tweets: to avoid confusion to topic models and to allow easy interpretation.
    - Non-retweets: to study the influence of the original tweets.



(a) Raw decahose tweets volume from Feb 15 to May 15.



(b) Geotagged US, Non-retweet, English decahose tweets volume from Feb 15 to May 15.

Limitations of subsampled tweets

The data is noisy and subject to selection bias, and for it to be useful for general purposes one needs:

- People in the US who tweets about their opinions.
  - People in the US who tweets about their opinions + have Twitter location service on (so the data is geotagged).
- People who are truthful about their tweets.
  - People who are truthful about their tweets and use proper languages.
- ...

But we are *NOT* doing predictions, *NOR* making strong inferential decisions based on the data.

Overall, these issues do not affect our study of "public discourse" on COVID-19, and we are interested in understanding this subpopulation.

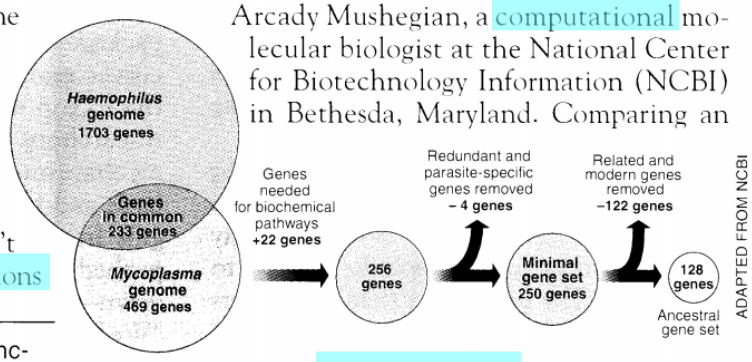
Outline

Brief introduction to topic modelling

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

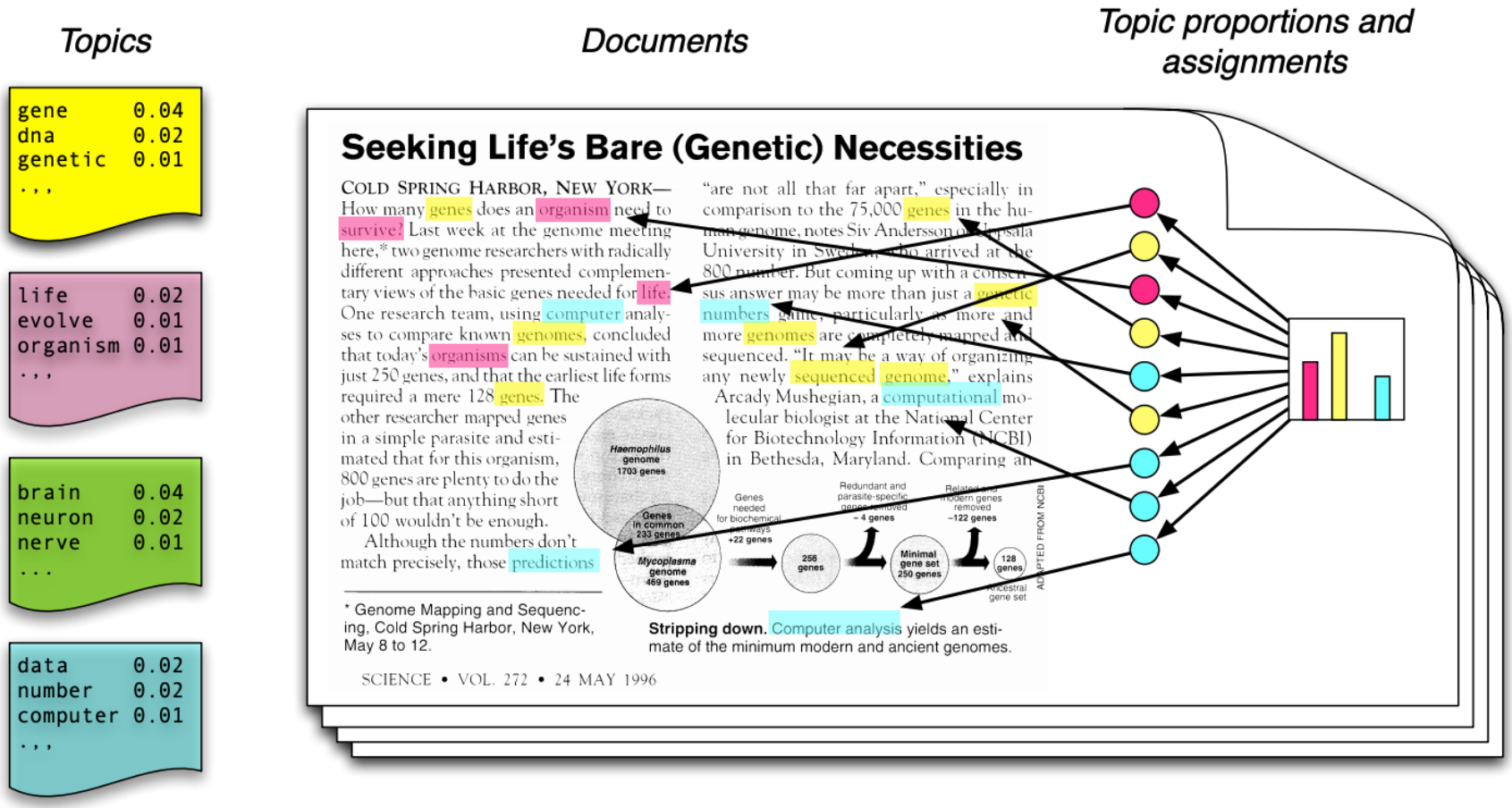
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Documents exhibit multiple topics[2].

[2] David M Blei. "Probabilistic topic models". In: *Communications of the ACM* 55.4 (2012), pp. 77–84.

Latent Dirichlet Allocation[3]



[3]David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

LDA and Twitter LDA[4] as generative processes

LDA

For each topic t

- $\phi^t \sim \text{Dir}(\beta)$

For each doc d

- $\theta^d \sim \text{Dir}(\alpha)$
- For each word i in doc d
  - $z_{d,i} \sim \text{Multi}(\theta^d)$
  - $w_{d,i} \sim \text{Multi}(\phi^{z_{d,i}})$

But, tweets (Micro-text) concentrate on *single topics*, and *aggregation/pooling* of tweets is needed.

Twitter LDA

For each topic t

- $\phi^t \sim \text{Dir}(\beta)$

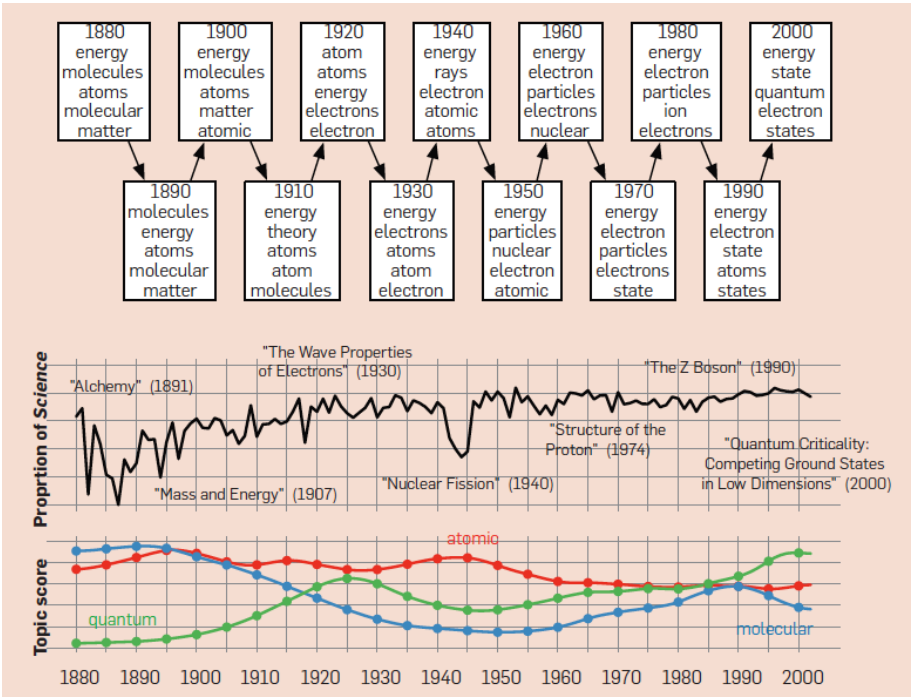
For each user u

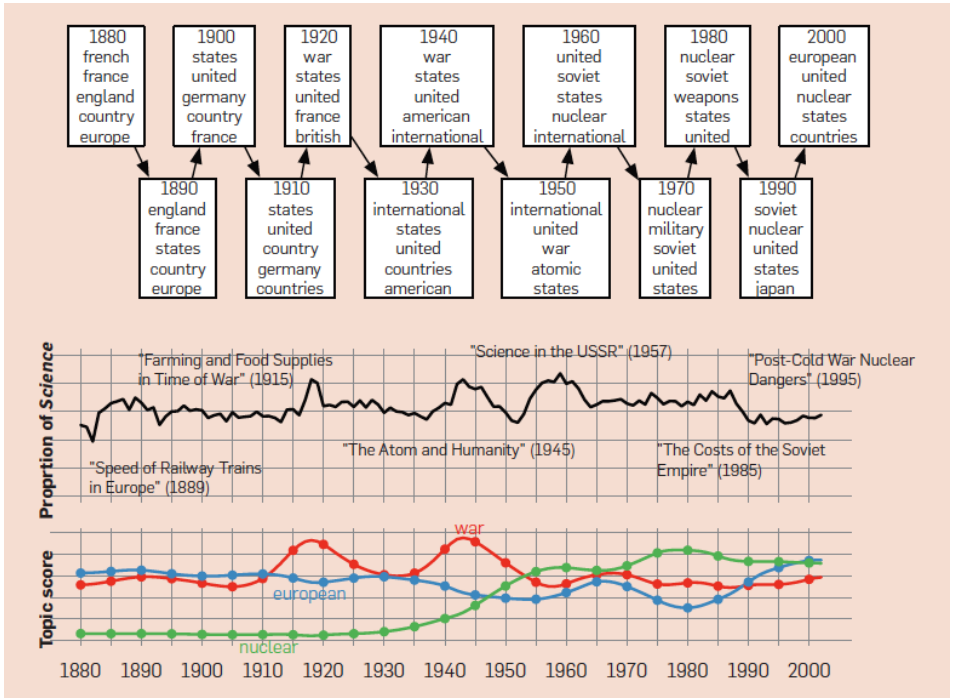
- $\theta^u \sim \text{Dir}(\alpha)$
- *For each tweet* s
  - $z_{u,s} \sim \text{Multi}(\theta^u)$
  - For each word i
    - $w_{u,s,i} \sim \text{Multi}(\phi^{z_{u,s}})$

But, we are interested in *time evolution* of the topics as well.

[4] Wayne Xin Zhao et al. "Comparing twitter and traditional media using topic models". In: *European conference on information retrieval*. Springer. 2011, pp. 338–349.

Dynamic topic model





Two topics from a dynamic topic model, which was fit to the Science journal from 1880 to 2002. Top words at each decade were illustrated[5][6].

Idea: chaining together the topic and topic proportions distributions through random walk stochastic processes and *jointly fit the model*.

[5] David M Blei and John D Lafferty. "Dynamic topic models". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 113–120.

[6] David M Blei. "Probabilistic topic models". In: *Communications of the ACM* 55.4 (2012), pp. 77–84.

Why not dynamic LDA?

Why not dynamic LDA?

- Pros: dynamics explicitly built into the generative model.
- Cons:
  - Sensitive to model assumptions (e.g., stationarity)
  - Computationally unstable and expensive (relying on approximate inference algorithms)
  - Reliable existing software for dynamic short-document LDA

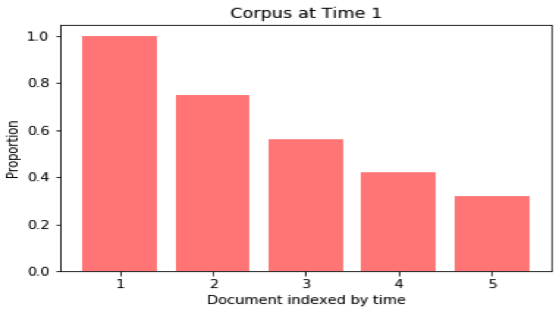
Goal: provide a simple suite of tools for general use, using robust and widely available software.

Idea: mimic the idea of dynamic topic models while taking simple, modular, and interpretable approaches.

Emulation of dynamic model: temporally smoothed corpus

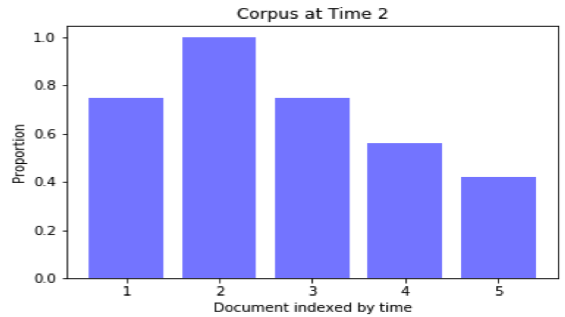
Doc 1	Doc 2	Doc 3	Doc 4	Doc 5
Tweet 1	Tweet 1	Tweet 1	Tweet 1	Tweet 1
Tweet 2	Tweet 3	Tweet 3	Tweet 3	
Tweet 3	Tweet 4	Tweet 4		
Tweet 4	Tweet 5			
Tweet 5				

$5\times1 = 5 \rightarrow 5\times0.75^{1} \approx 4 \rightarrow \cdots \rightarrow 5\times0.75^{4} \approx 1$



Doc 1	Doc 2	Doc 3	Doc 4	Doc 5
Tweet 1	Tweet 1	Tweet 1	Tweet 1	Tweet 1
Tweet 2	***Tweet 2***	Tweet 3	Tweet 3	***Tweet 2***
~~***Tweet 3***~~	Tweet 3	Tweet 4	***Tweet 4***	
Tweet 4	Tweet 4	***Tweet 5***		
Tweet 5	Tweet 5			

$5\times0.75^{1} \approx 4 \rightarrow 5\times1 = 5 \rightarrow \cdots \rightarrow 5\times0.75^{3} \approx 2$



Smoothed subsampling using examples of 5 documents each containing 5 tweets. Doc 1 aggregates tweets from day 1, Doc 2 aggregates tweets from day 2, etc.

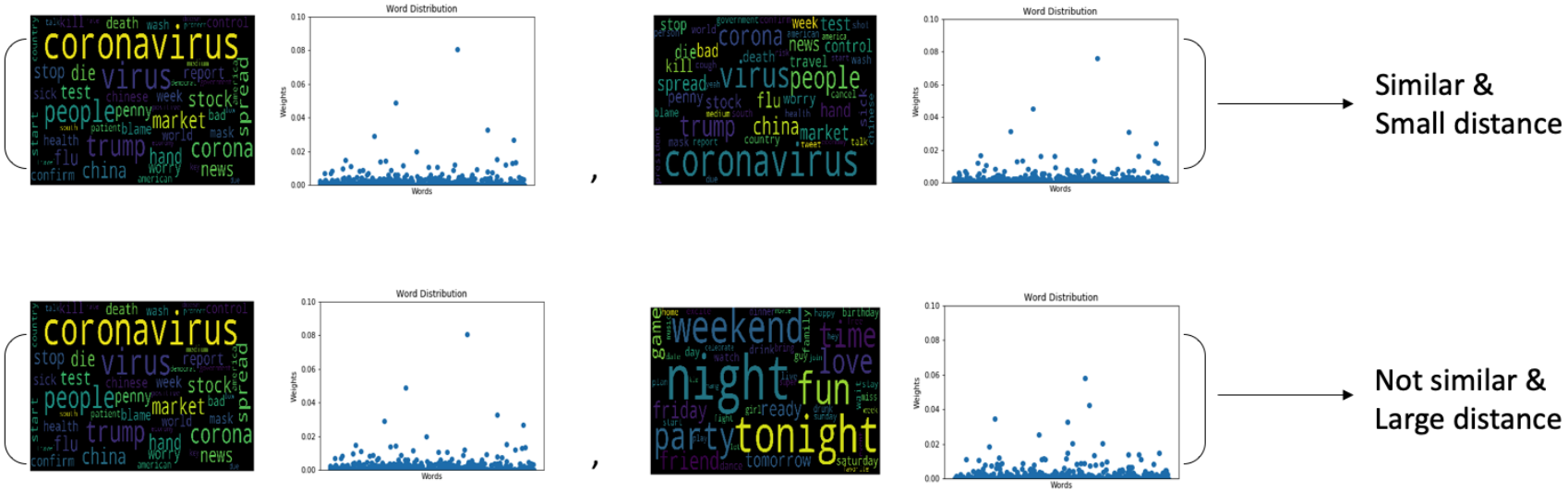
T-LDA applied *independently in parallel* to each temporally smoothed corpus from Feb 15 to May 15 ($\approx$ 90 time points) with 50 topics $\Rightarrow$ *marginal models fitted time by time*.

Outline

Fusing marginal results

Goal: "stitch" together marginal LDA results, i.e., topics, by connecting all pairs of topics between any consecutive time points. Assign "weights" to each connection that measures similarity.

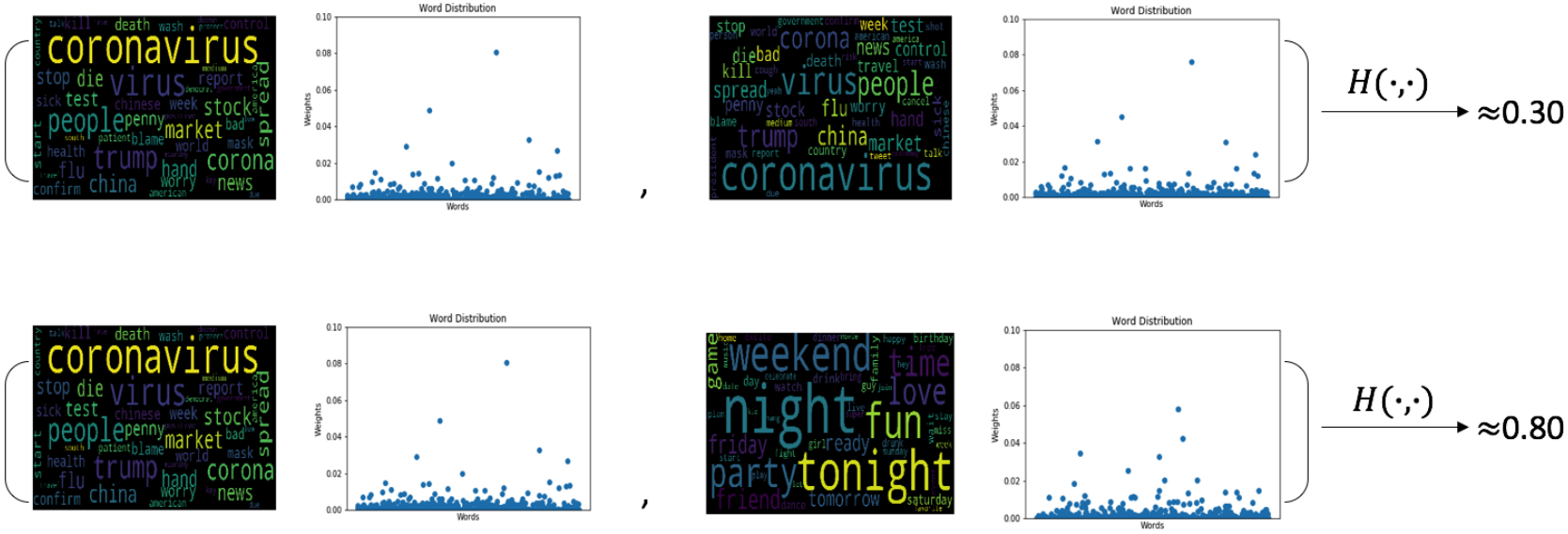
Idea: topics are represented by distributions over words $\Rightarrow$ measure similarity between two discrete distributions.



Hellinger distances between two topics

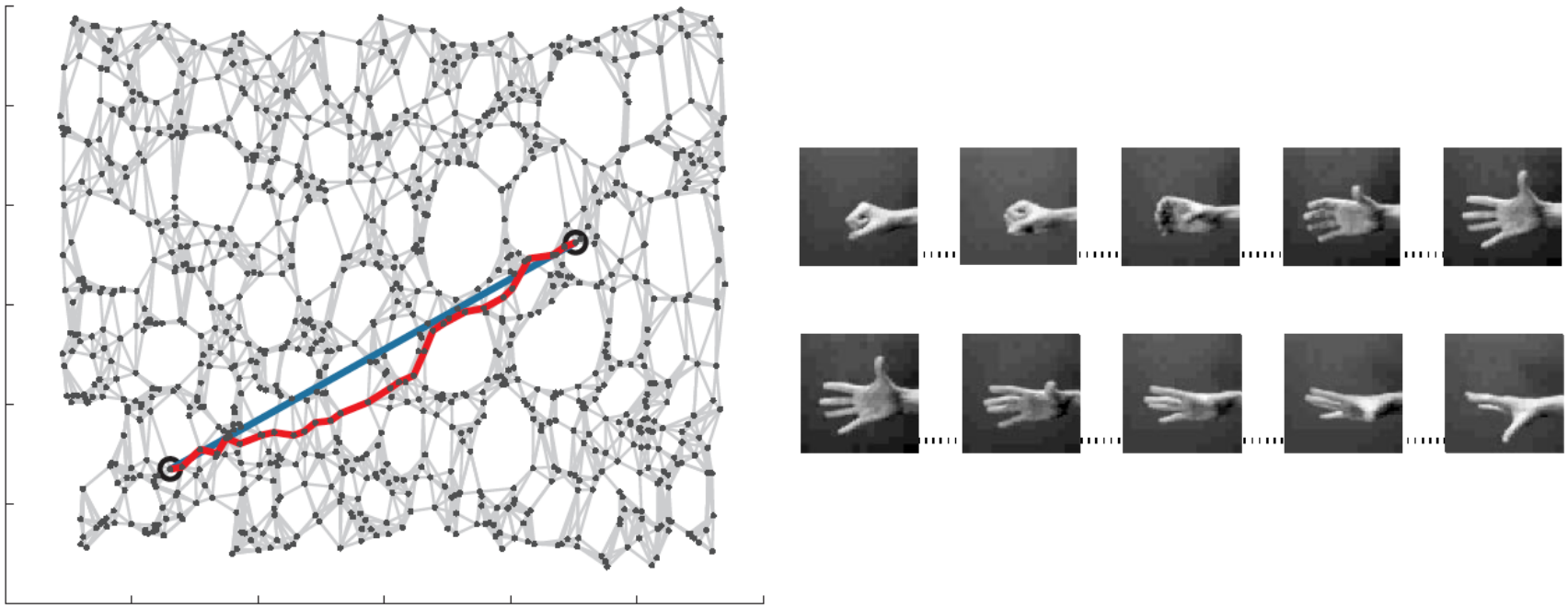
Hellinger distance [7] between discrete probability distributions $P = (p_1, \ldots, p_N)$ and $Q = (q_1, \ldots, q_N)$:

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{n=1}^{N} (\sqrt{p_n} - \sqrt{q_n})^2}, \quad 0 \leq H(\cdot, \cdot) \leq 1.$$



[7] A case of f-divergence that measures distance between probability distributions: `https://en.wikipedia.org/wiki/F-divergence`.

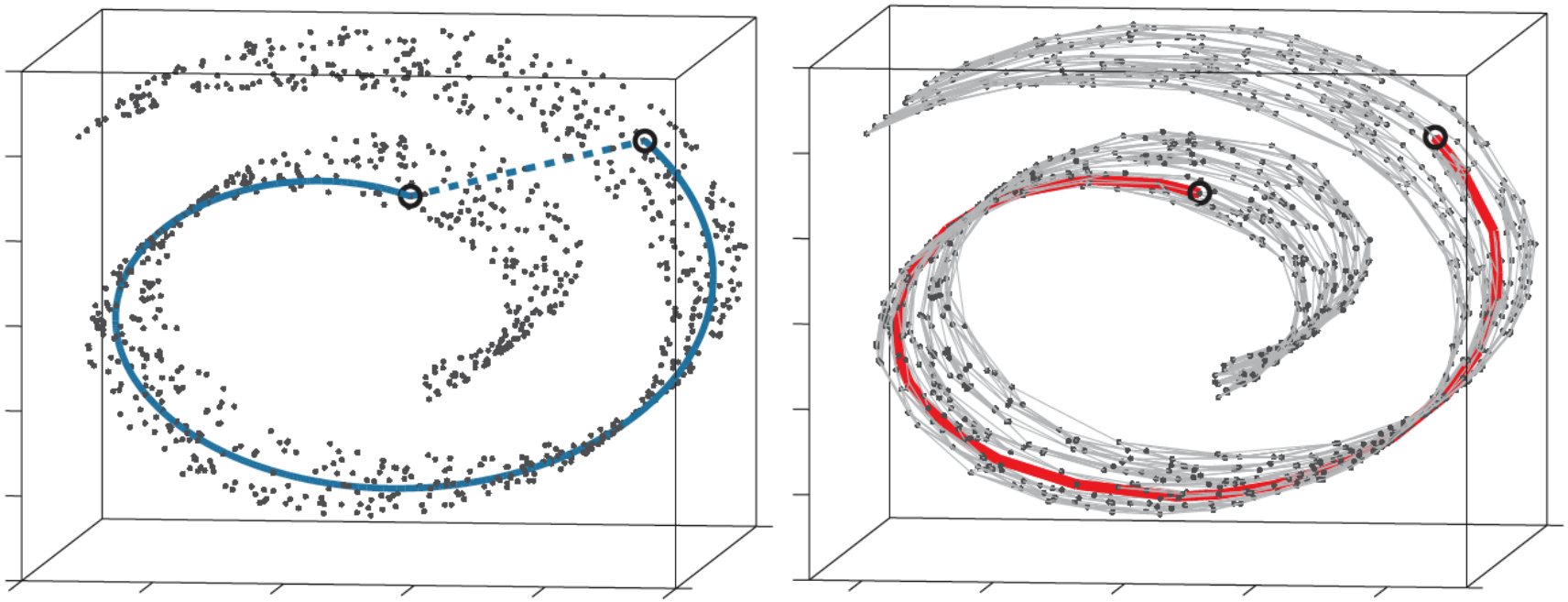
Shortest path on neighborhood graph



Shortest distance on neighborhood graph captures perceptually natural but highly nonlinear morphs of the corresponding high-dimensional observations by transforming them approximately along geodesic paths (solid curve on the left plot)[8].

[8] Joshua B Tenenbaum, Vin De Silva, and John C Langford. "A global geometric framework for nonlinear dimensionality reduction". In: *science* 290.5500 (2000), pp. 2319–2323.

Full vs. neighborhood weighted graph of observations

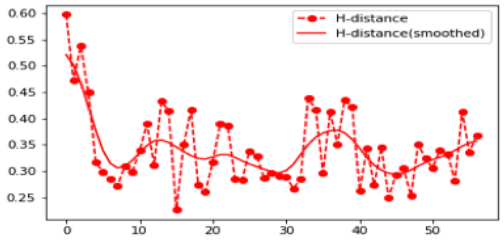


The "Swiss roll" data set, illustrating how nearest neighborhood graph (7 nearest neighbors in this case) exploits geodesic paths for nonlinear dimensionality reduction[9].

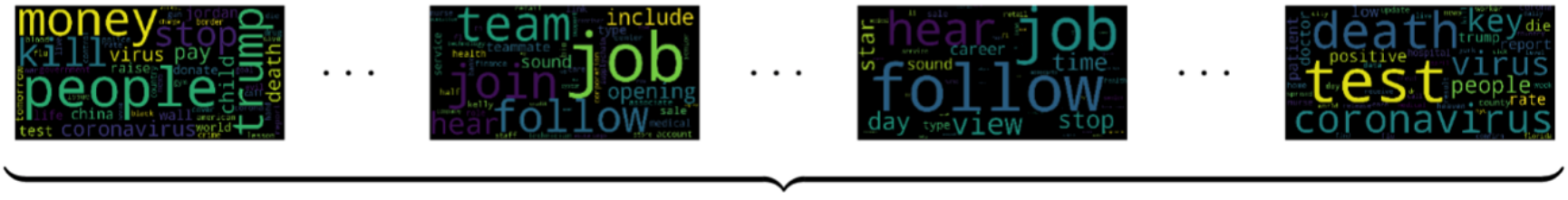
[9] Joshua B Tenenbaum, Vin De Silva, and John C Langford. "A global geometric framework for nonlinear dimensionality reduction". In: *science* 290.5500 (2000), pp. 2319–2323.

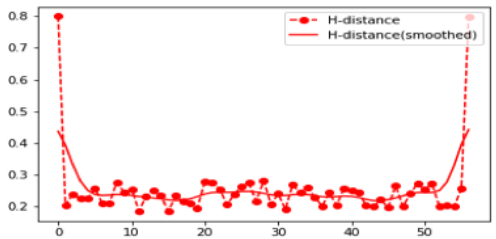
Topic transformation on full vs. 10-nearest neighbor graph

10-NN graph shortest path



Full graph shortest path

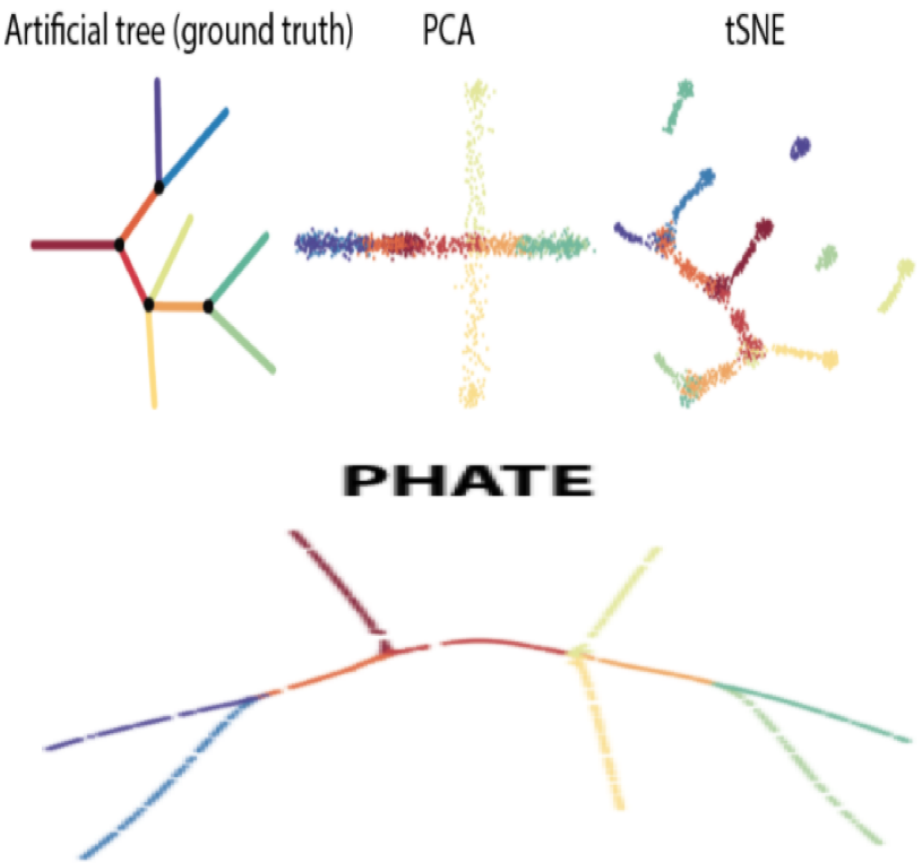




Outline

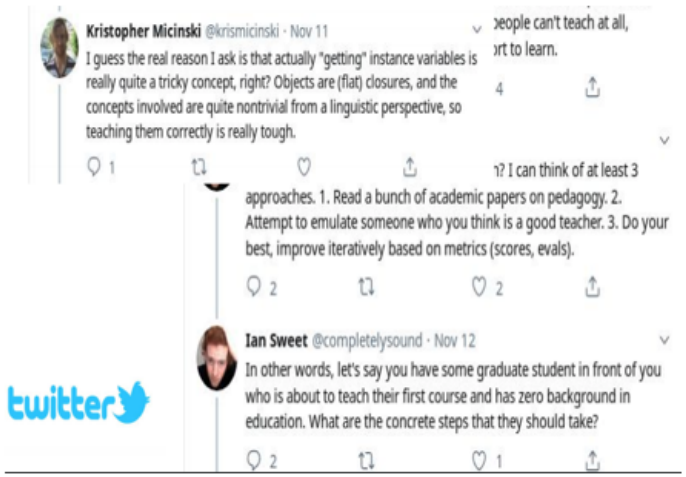
PHATE[10]

- To visualize and interpret high dimensional word distributions, need *lower-dimensional embedding* that capture the intrinsic high-dimensional *trajectory structure* of the data.
- Traditional methods like PCA assumes linearity.
- Nonlinear methods like t-SNE DO NOT naturally exhibits trajectory or progression.
- PHATE (Potential of Heat-diffusion for Affinity-based Trajectory Embedding) is designed explicitly to preserve progression structure in data.

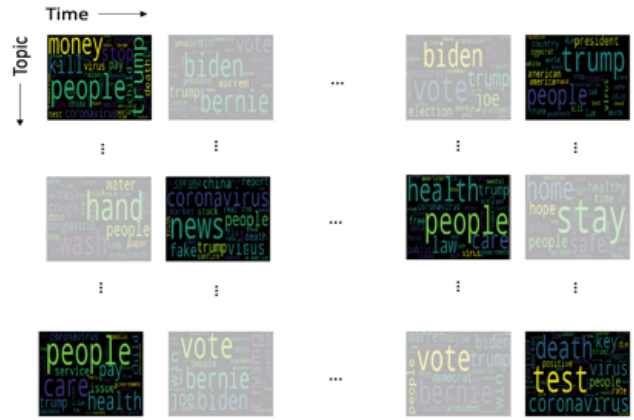


[10] Kevin R Moon et al. "Visualizing structure and transitions in high-dimensional biological data". In: *Nature Biotechnology* 37.12 (2019), pp. 1482–1492.

Summary of the procedure

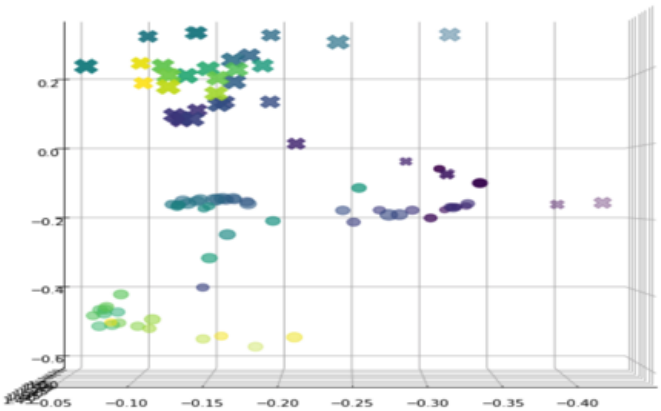


1. LDA on temporally smoothed corpus



2. Hellinger distances

3a. PHATE embedding

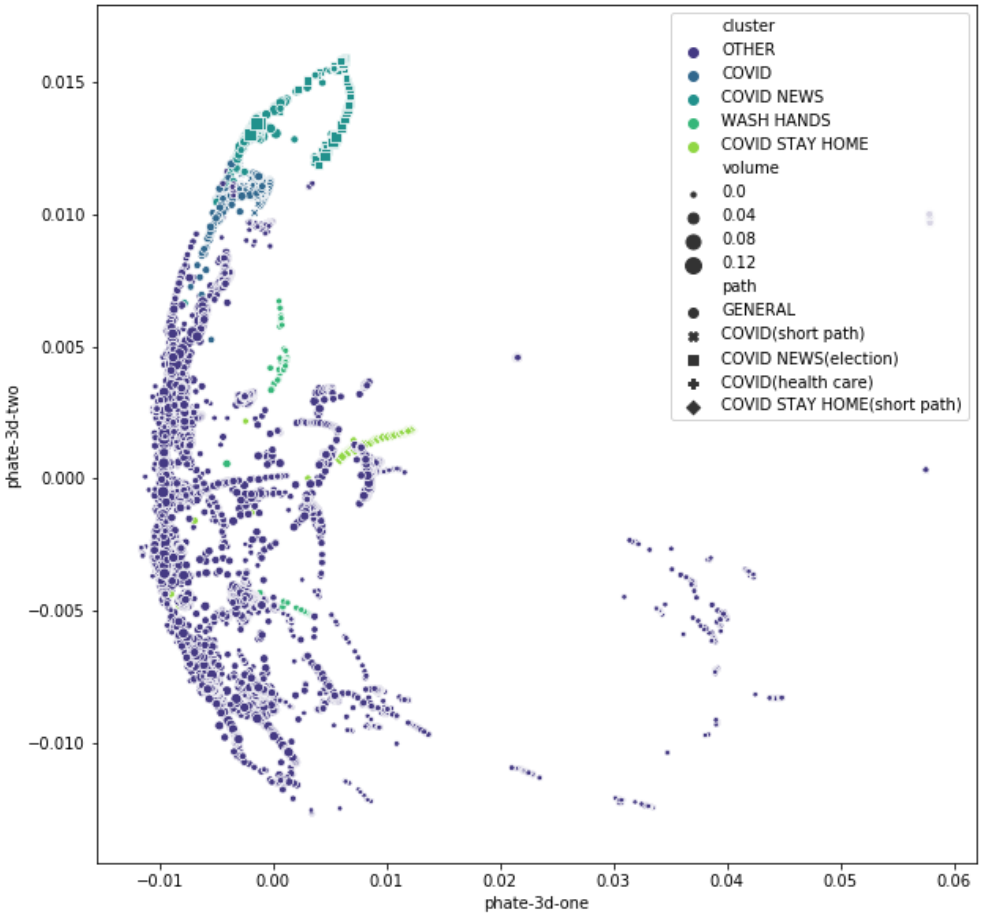


3b. Neighborhood graph and shortest path

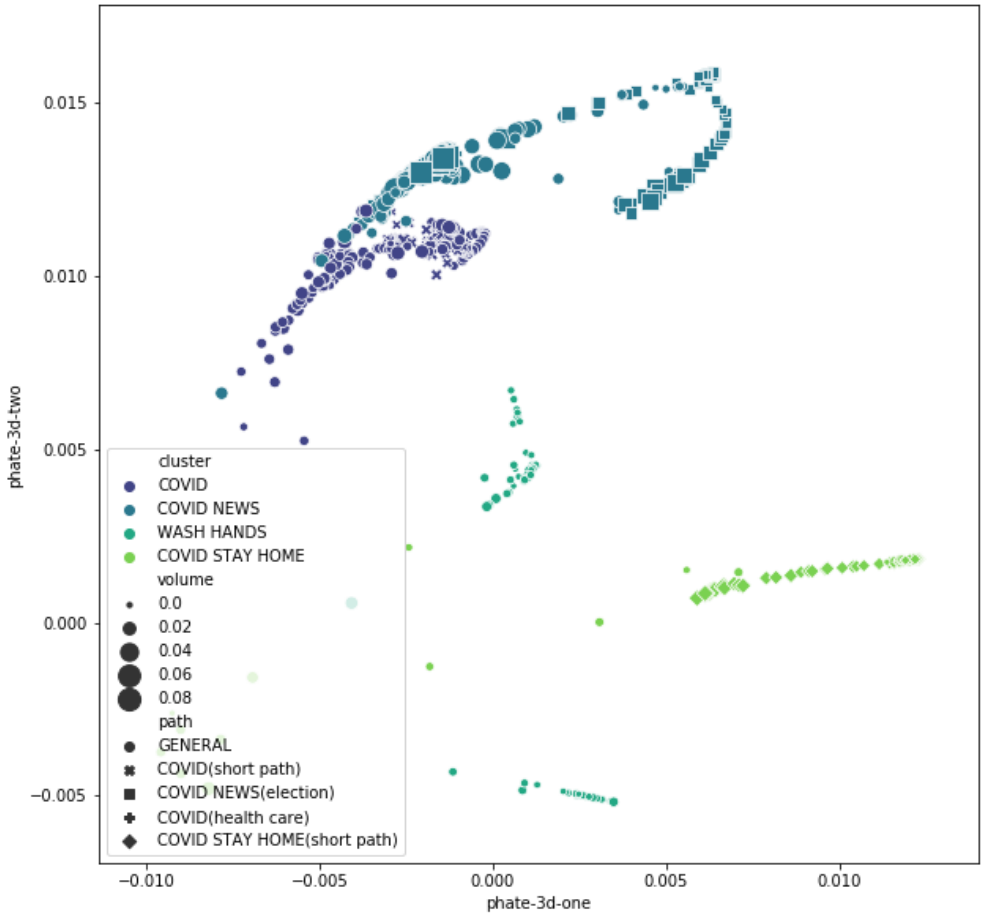
4. Visualize and interpret multiple paths and clusters

Outline

PHATE embedding of 4500 word distributions

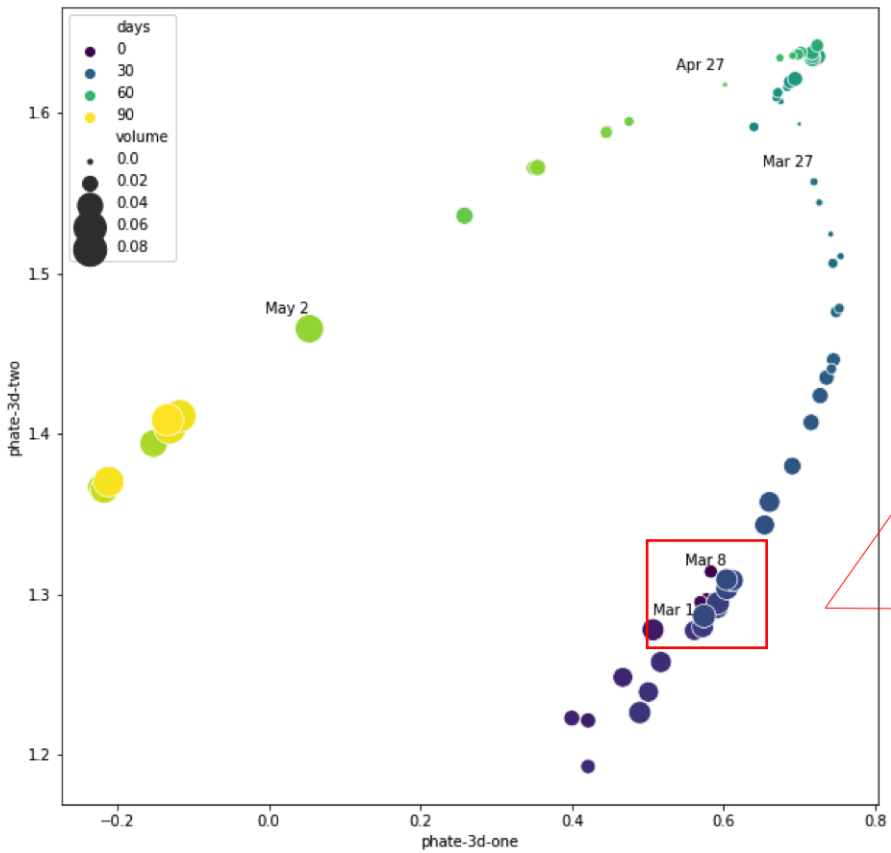


PHATE embedding of selected COVID topics

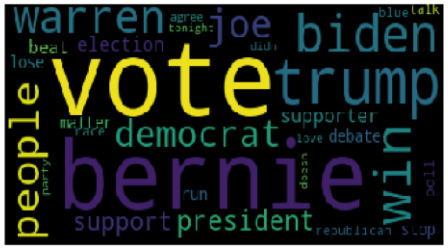


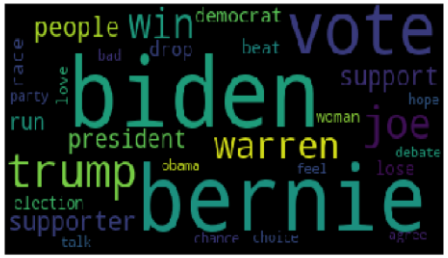
Case study I: presidential election topic

Idea: use real events we know about to understand the trajectories.



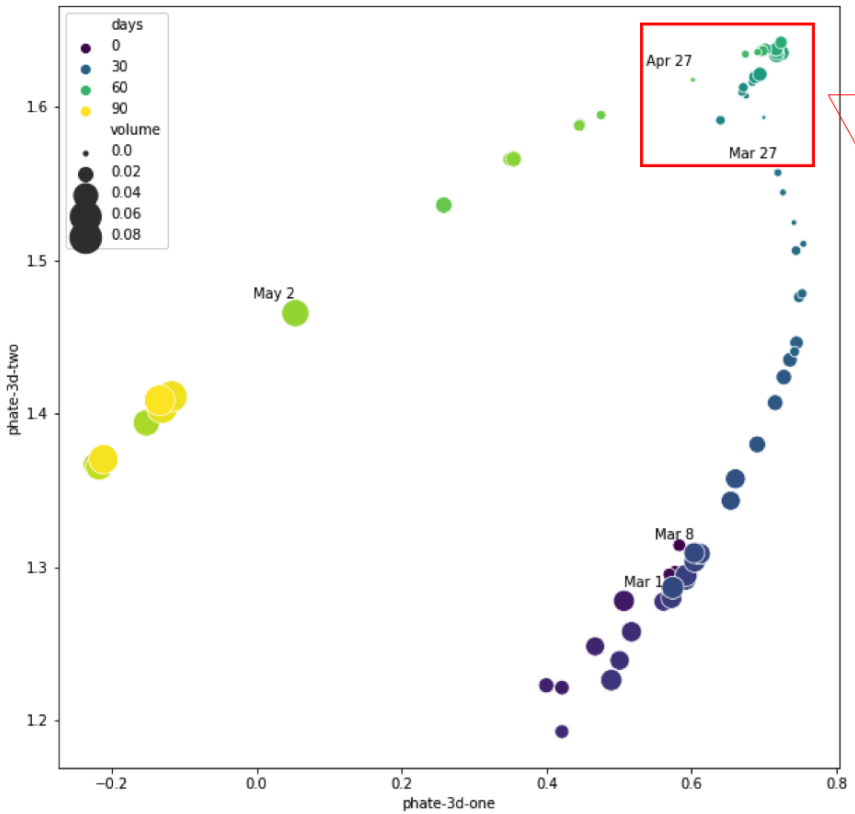
Super Tuesday, Mar 3

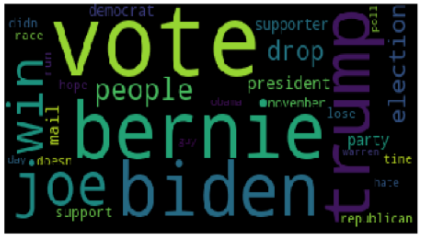




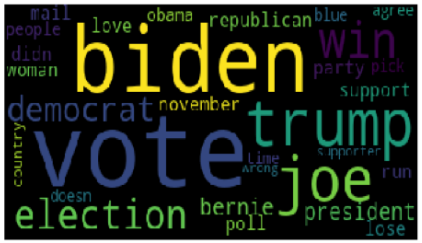
Super Tuesday week clustered with high volume.

Case study I cont'd: April subcluster





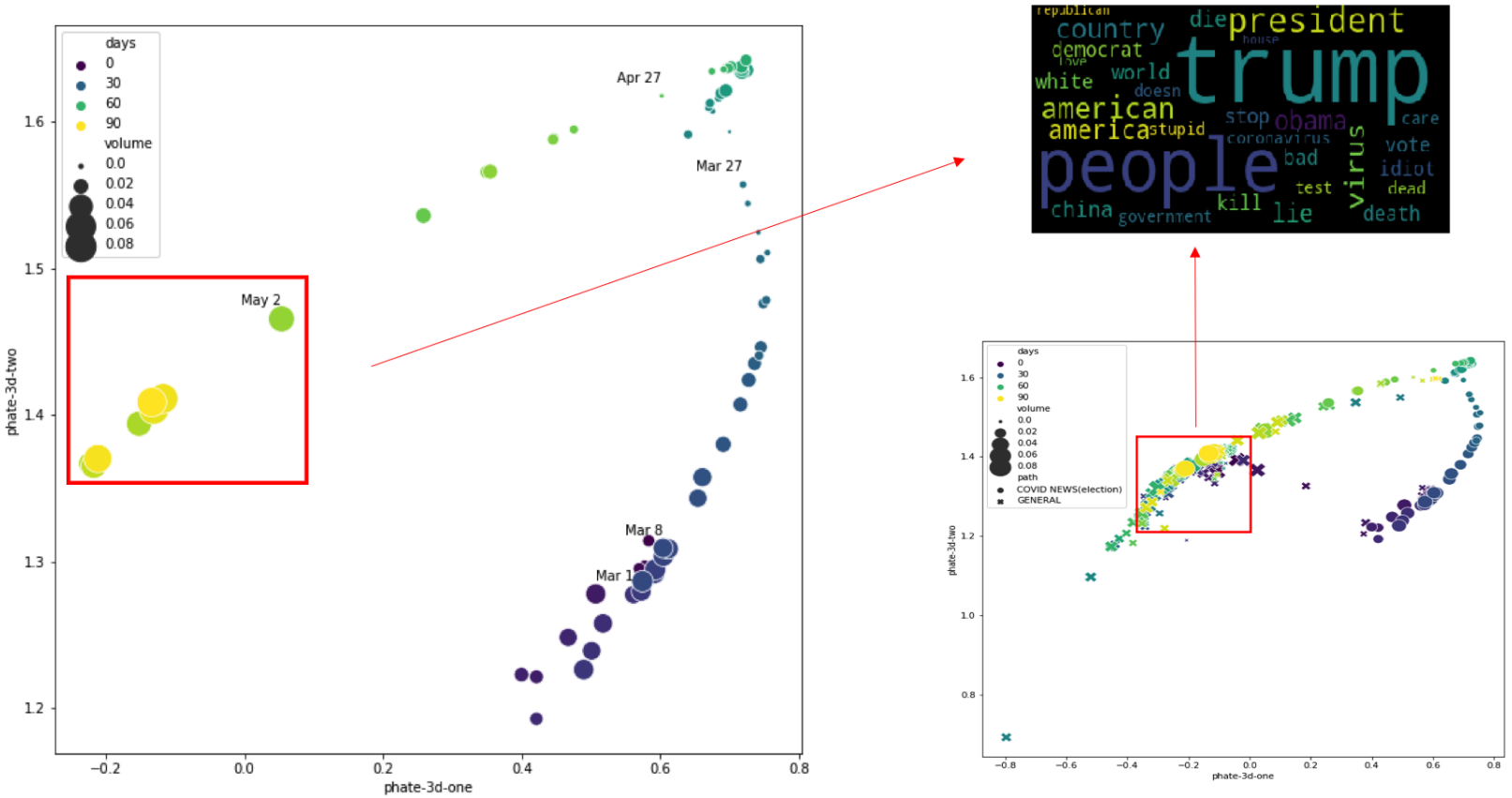
Bernie Sanders dropped out, April 8



People endorsed Biden, who wins Wisconsin & Ohio, April 27

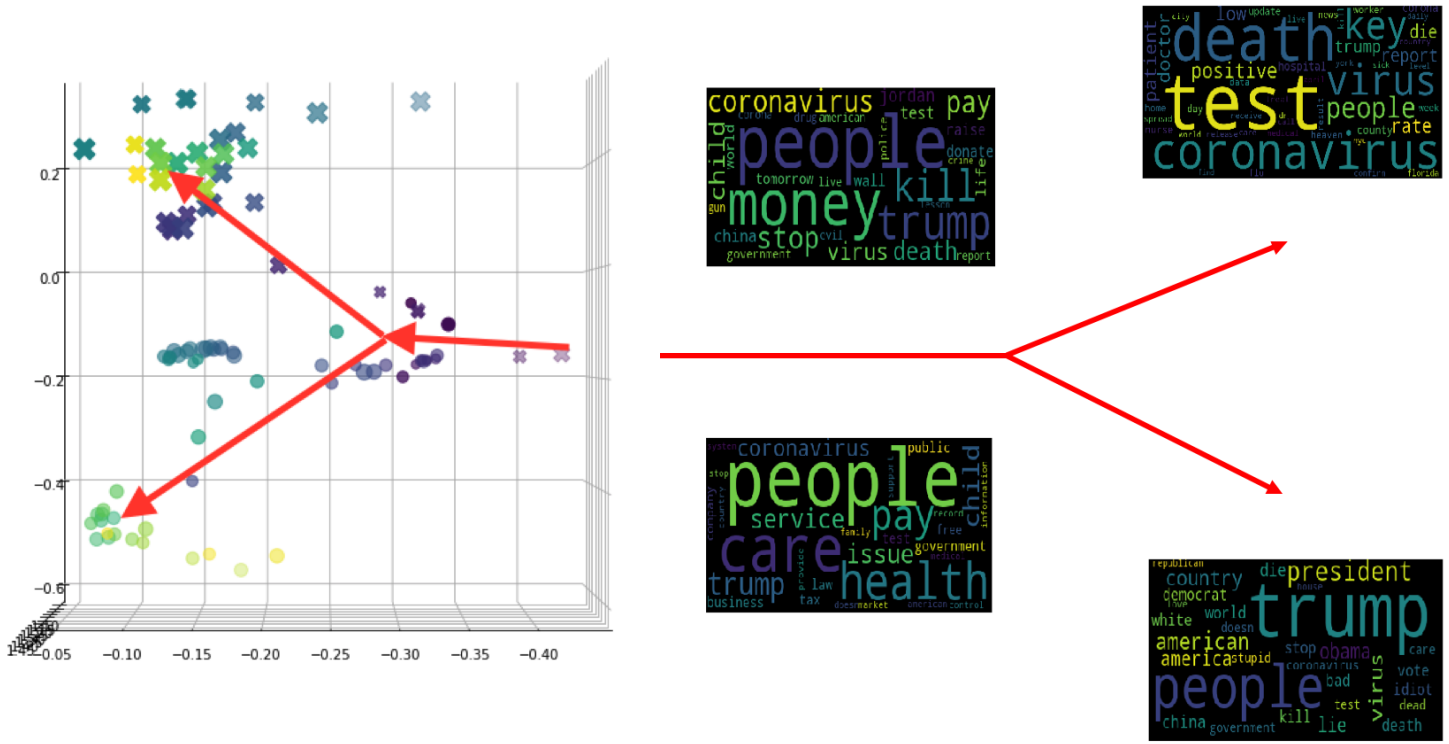
Tweets volume dies out in late March but re-surged and clustered in April because Bernie Sanders dropped out and more election-related activities.

Case study I cont'd: May subcluster



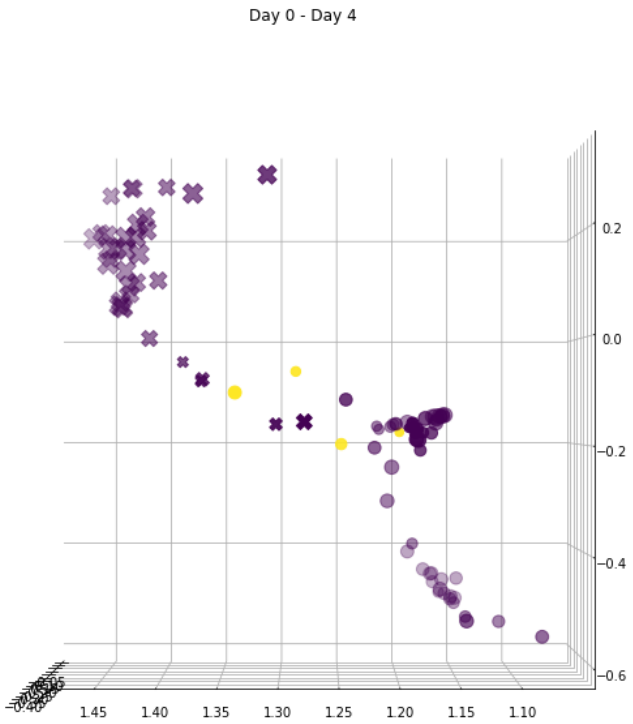
Election tweets in May converged to general COVID-related news tweets.

Case study II: general COVID topic



Divergence of two similar COVID-related conversations that ended up as a COVID-testing conversation and a COVID-politics conversation, respectively.

Case study II cont'd: dynamics



Conclusion

This talk:

- Generative statistical model & Computational geometry & Algorithms collectively provide efficient ways for understanding high-dimensional noisy Twitter data $\Rightarrow$ visualization and interpretation of public discourse around COVID-19.
- Longitudinal analysis of Twitter data done without complicated models $\Rightarrow$ understanding the impact of COVID-19 on various societal aspects.

Future work:

- Wrapping various tools into a single re-usable implementation for the public.
- Apply such tools globally or to tweets from other regions other than the US.
- Spatial analysis of Twitter data.