“Discovering the Ebb and Flow of Ideas in Text Corpora”

Published in IEEE Computer, February 2012

Casey Klippel

Friday, April 13, 2012

In collaboration with:
Justin Jee
Shahriar Hossain
Naren Ramakrishnan
Bud Mishra

Columbia University | New York University | Virginia Tech

So, what does:

“Discovering the Ebb and Flow of Ideas in Text Corpora”

mean?

Finding meaning in Big Data

- Data mining

  borrowing a definition from Naren Ramakrishnan:

  "the process of extracting non-trivial and actionable insights from data."

- Wide Variety of Uses:

  Commercial (e.g., targeted ads)

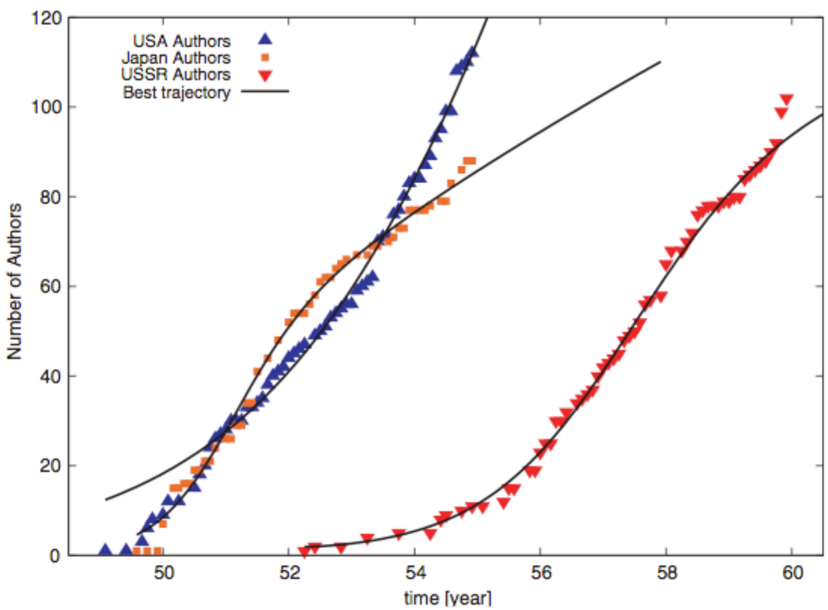
  Academic (e.g., gene expression)

  "Fun" (e.g., Google Trends)

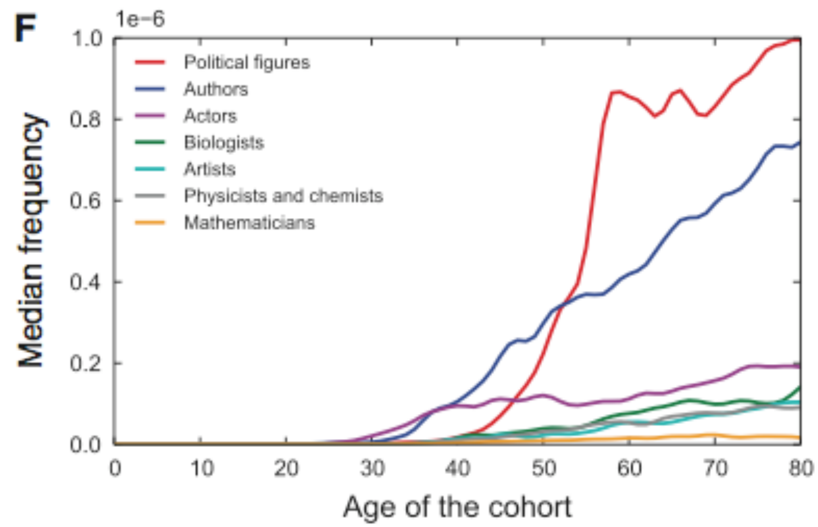
The Question

- Investigate the ideas surrounding events in recent Western history with major cultural implications
  - AIDS epidemic (1980s)
  - Einstein's theory of relativity (early 1900s)
  - Semmelweis's discovery of hand sanitation (1840s)

How have other researchers investigated the spread of ideas?



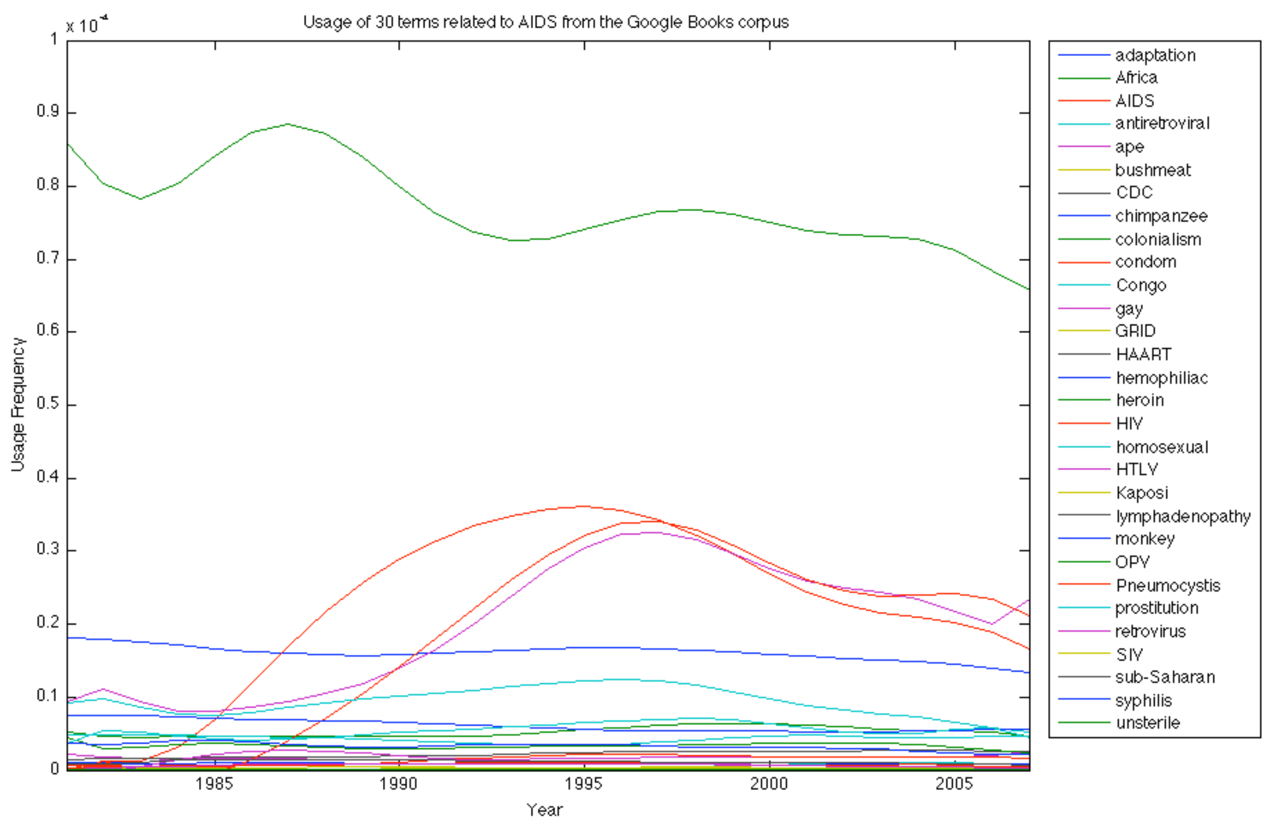
L.M.A. Bettencourt et al., "The Power of a Good Idea: Quantitative Modeling of the Spread of Ideas from Epidemiological models," *Physica A*, May 2006, pp.513-536.



J. Michel et al., "Quantitative Analysis of Culture Using Millions of Digitized Books," *Science*, 14 Jan 2011, pp. 176-182.

The Databases

- Google Books, New York Times, PubMed
- Large amounts of data but difficult to interpret



The Algorithm: GOALIE

- Cluster-based temporal segmentation
- Main idea: embed a clustering algorithm in a segmentation algorithm.

<u>Data</u>: multiple measurement vectors $W=\{w_1,w_2,\ldots,w_N\}$, where each w_i is a time series over $T=\{t_1,t_2,\ldots,t_l\}$.

- Each w_i represents a term (eg, AIDS) and its corresponding time series t_i represents usage data over the given years (eg, 1981-2008)

The Algorithm: GOALIE

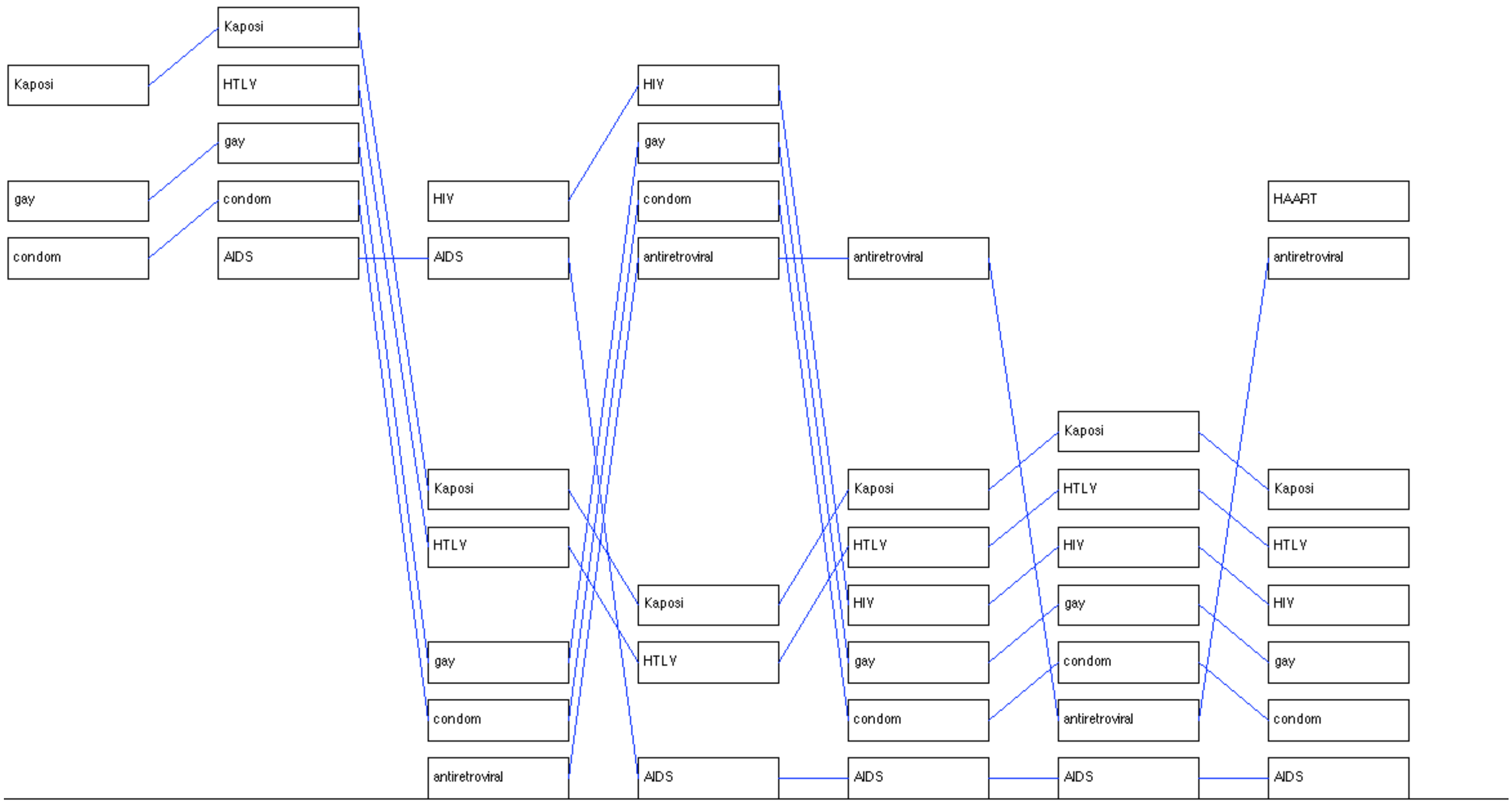
Segmentation Goal: express T as a sequence of segments: $s_{t1}^{ta}, s_{ta+1}^{tb}, \ldots, s_{tk}^{tl}$ where each segment s_{ti}^{te}, $t_i \leq t_e$, is a set of consecutive time points.

Clustering Goal: Given two clusterings of terms, e.g. W_{ta}^{tb} for left time segment s_{ta}^{tb} and W_{tb+1}^{tc} for right time segment s_{tb+1}^{c}, obtain clusters W that are local within each segment but are highly dissimilar to neighboring clusters.

Constraints: Define segment length constraints l_{min} and l_{max}, and maximum number of clusters.

The Algorithm: GOALIE

Temporal Segmentation Visualization



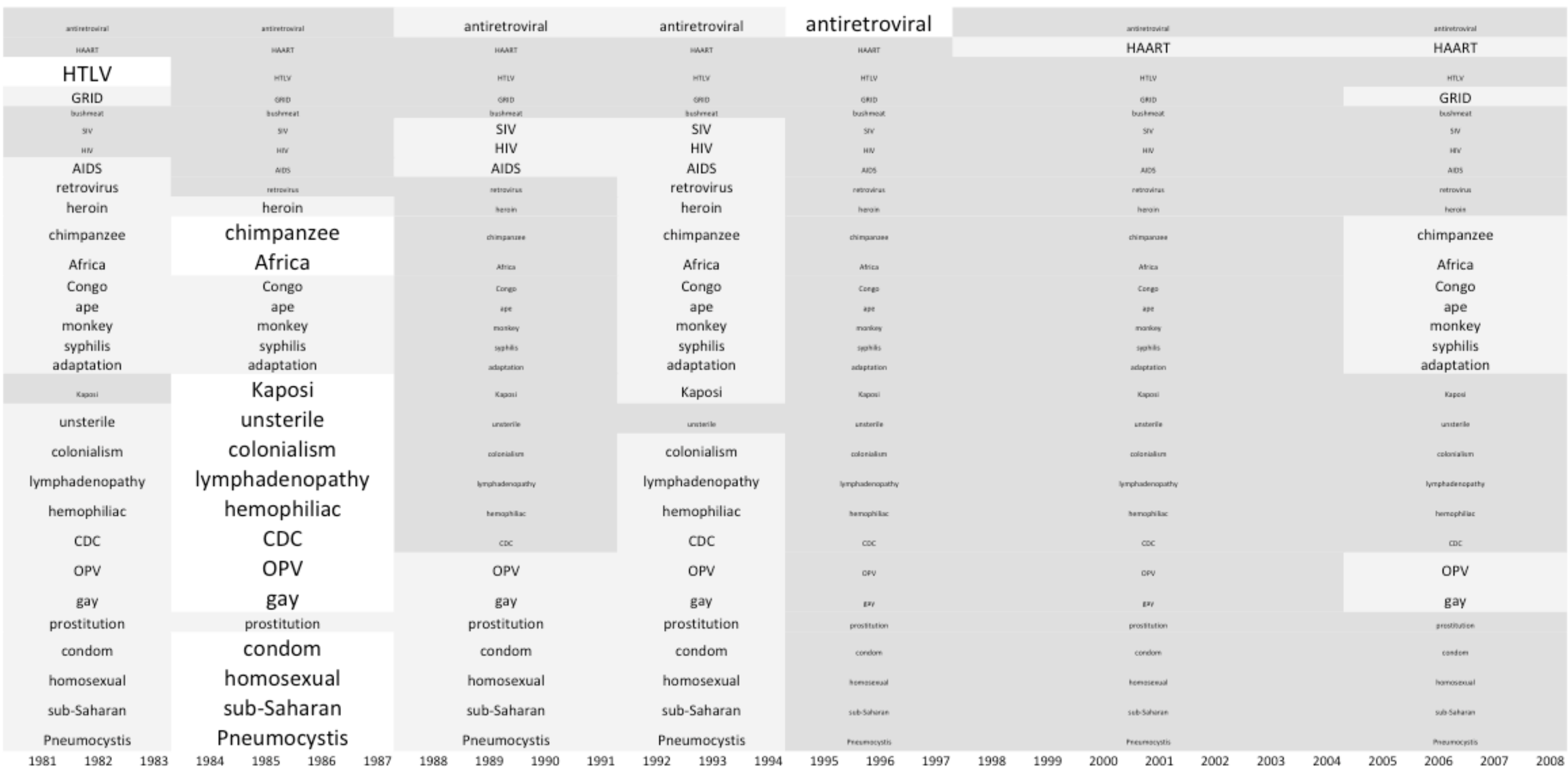
The Algorithm: GOALIE

1. Given two clusters W_a and W_b, before and after a given time point, define cluster random variables α ={1,…,r} and β = {1,…,c}.
2. Measure the clusters' similarity using an r x c contingency table. Entry n_{ij} represents the number of one-to-one relationships between the elements of the *i*-th cluster of W_a and the *j*-th cluster of W_b.
   - Ideally a highly disparate pair of clusterings result in a total of (r+c) uniform distributions across the row and columns.
3. Define *r* random variables R_i (i=1,…,r) w.p. $p_{Ri}(j)=n_{ij}/n_i$ corresponding to each row, and *c* random variables R_j (j=1,…,c) w.p. $p_{Cj}(i)=n_{ij}/n_j$ corresponding to each column.
4. **Minimize divergence from the uniform distributions over the rows U(1/c) and columns U(1/r) by minimizing:**

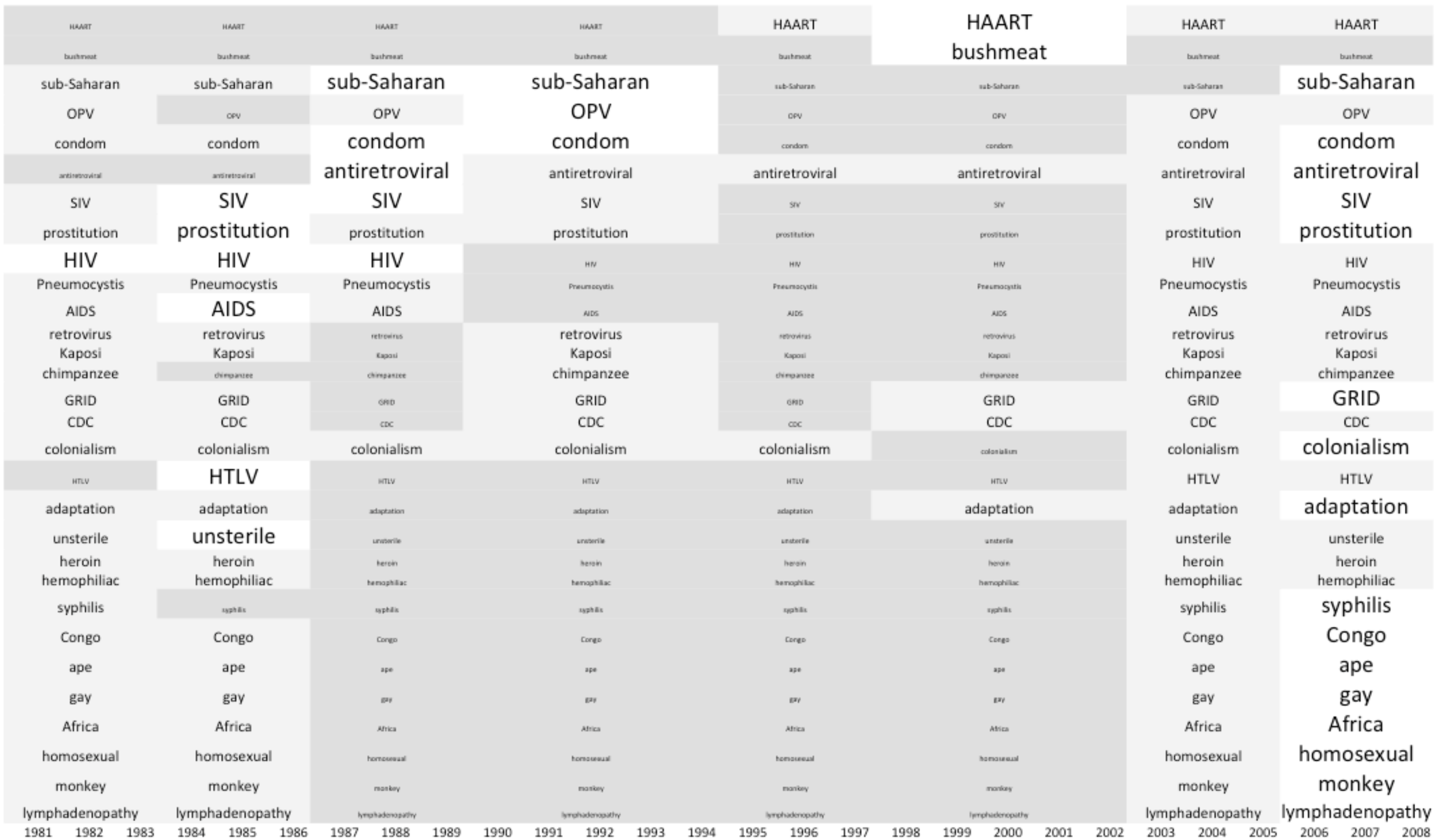
$$F = \frac{1}{r}\sum_{i=1}^{r} D_{KL}\left(p_{R_i}||U\left(\frac{1}{c}\right)\right) + \frac{1}{c}\sum_{j=1}^{c} D_{KL}\left(p_{C_j}||U\left(\frac{1}{r}\right)\right)$$

where $D_{KL}(p||q) = \sum_x p(x)\log_2 \frac{p(x)}{q(x)}$ is the Kullback-Leibler (KL) divergence.

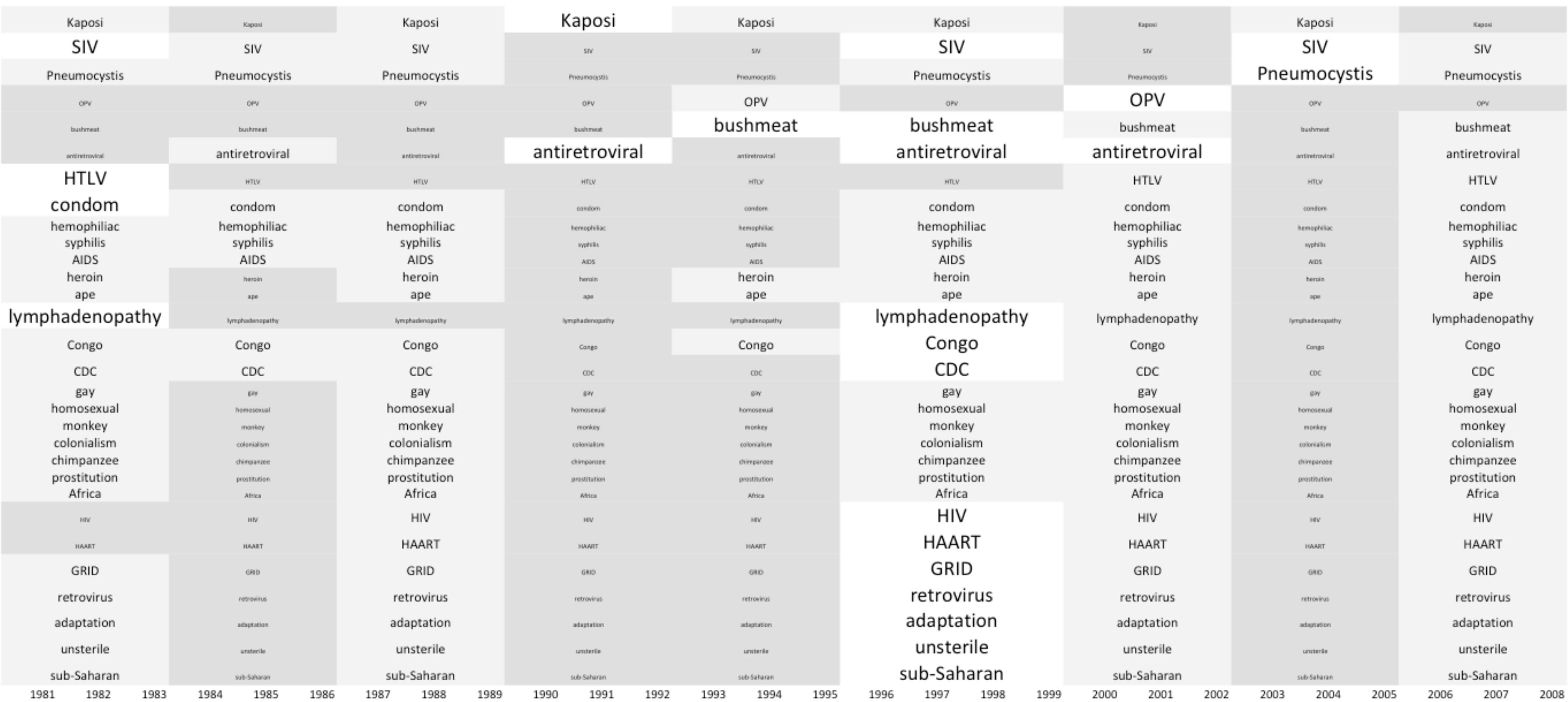
Results: AIDS, Google Books



Results: AIDS, PubMed

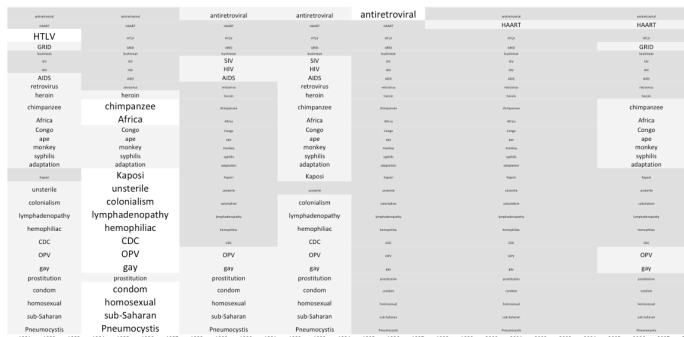


Results: AIDS, New York Times

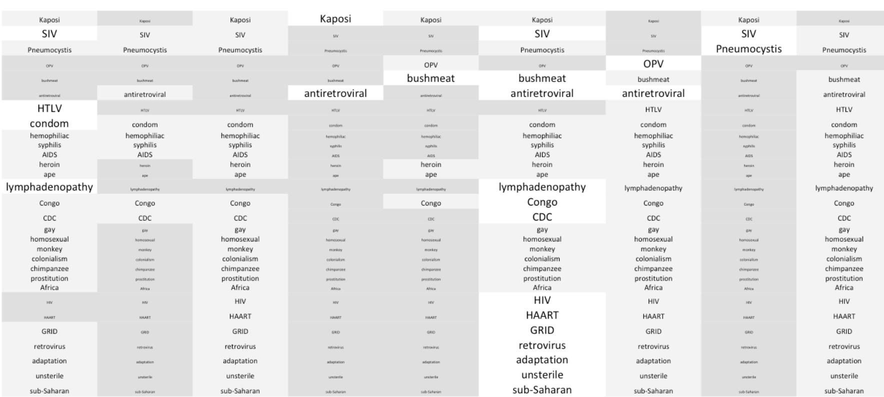


Results: AIDS, all three databases

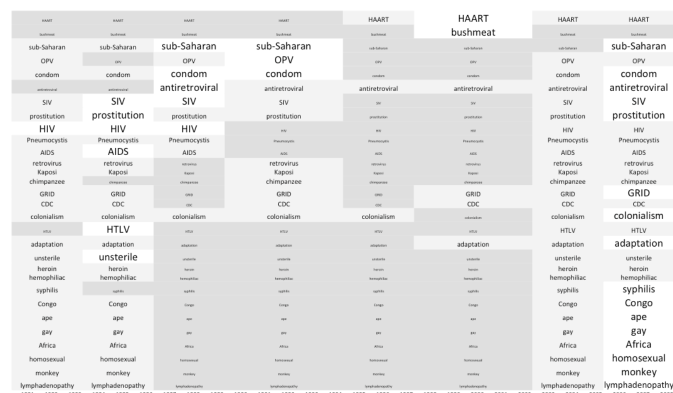
Google Books



New York Times



PubMed



Results: Einstein

Results: Semmelweis