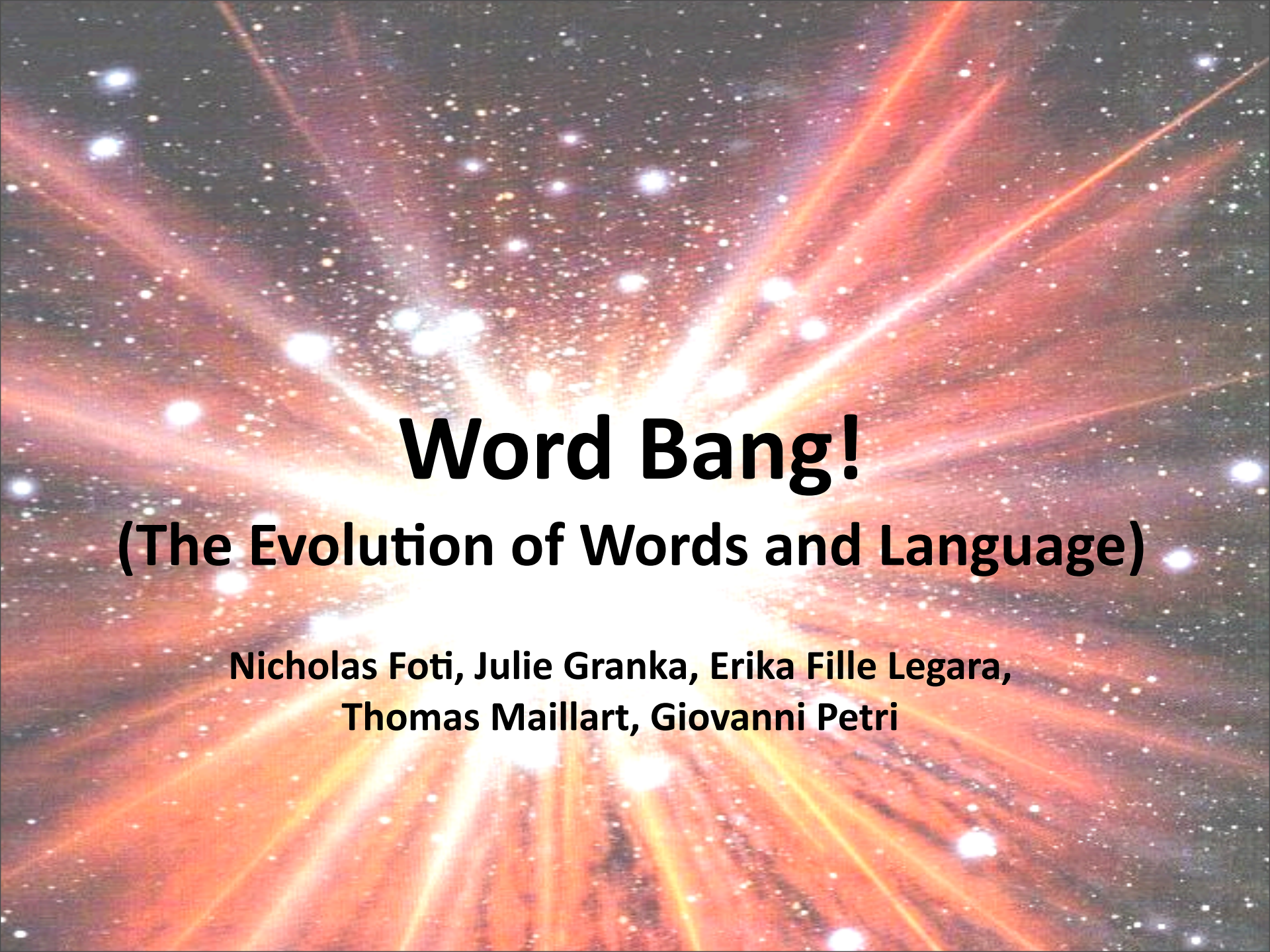


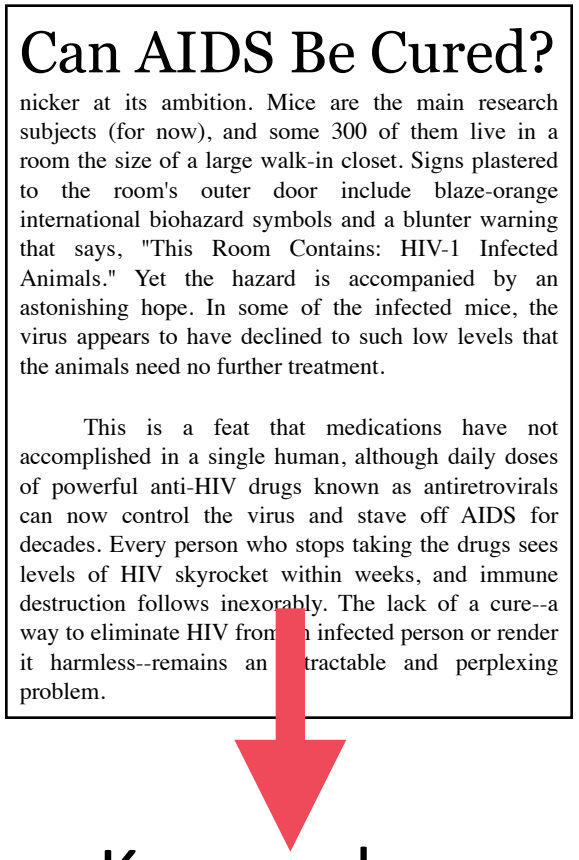
Word Bang!

(The Evolution of Words and Language)

Nicholas Foti, Julie Granka, Erika Fille Legara,
Thomas Maillart, Giovanni Petri

How to choose most pertinent "keyword set" to describe a text?

- Keyword Set:
  - provide a general text understanding
  - differentiate two texts of a similar topic
  - must be limited (<10 keywords)
  - take into account context change over time



Can AIDS Be Cured?

nicker at its ambition. Mice are the main research subjects (for now), and some 300 of them live in a room the size of a large walk-in closet. Signs plastered to the room's outer door include blaze-orange international biohazard symbols and a blunter warning that says, "This Room Contains: HIV-1 Infected Animals." Yet the hazard is accompanied by an astonishing hope. In some of the infected mice, the virus appears to have declined to such low levels that the animals need no further treatment.

This is a feat that medications have not accomplished in a single human, although daily doses of powerful anti-HIV drugs known as antiretrovirals can now control the virus and stave off AIDS for decades. Every person who stops taking the drugs sees levels of HIV skyrocket within weeks, and immune destruction follows inexorably. The lack of a cure--a way to eliminate HIV from an infected person or render it harmless--remains an intractable and perplexing problem.

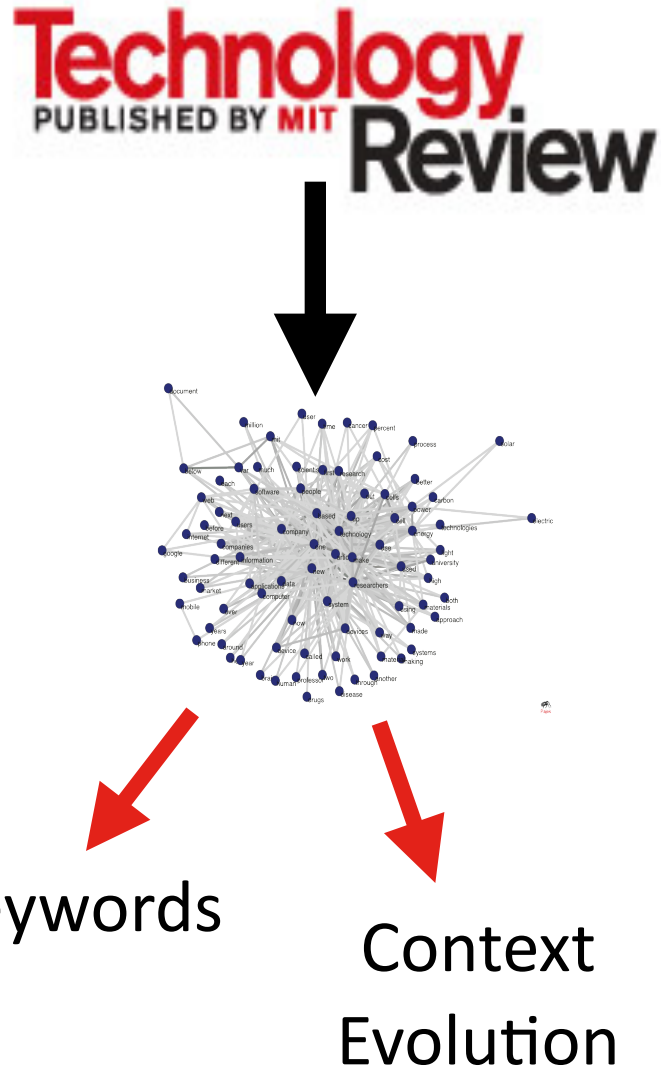
Keywords

{HIV, AIDS,Drugs,etc...}

Experimental Setting

- MIT Technology Review (1997-2010)
  - Easy to crawl
  - Small texts (blog posts 200-400 words)
  - Bounded topics
  - Fast evolution, yet consistent

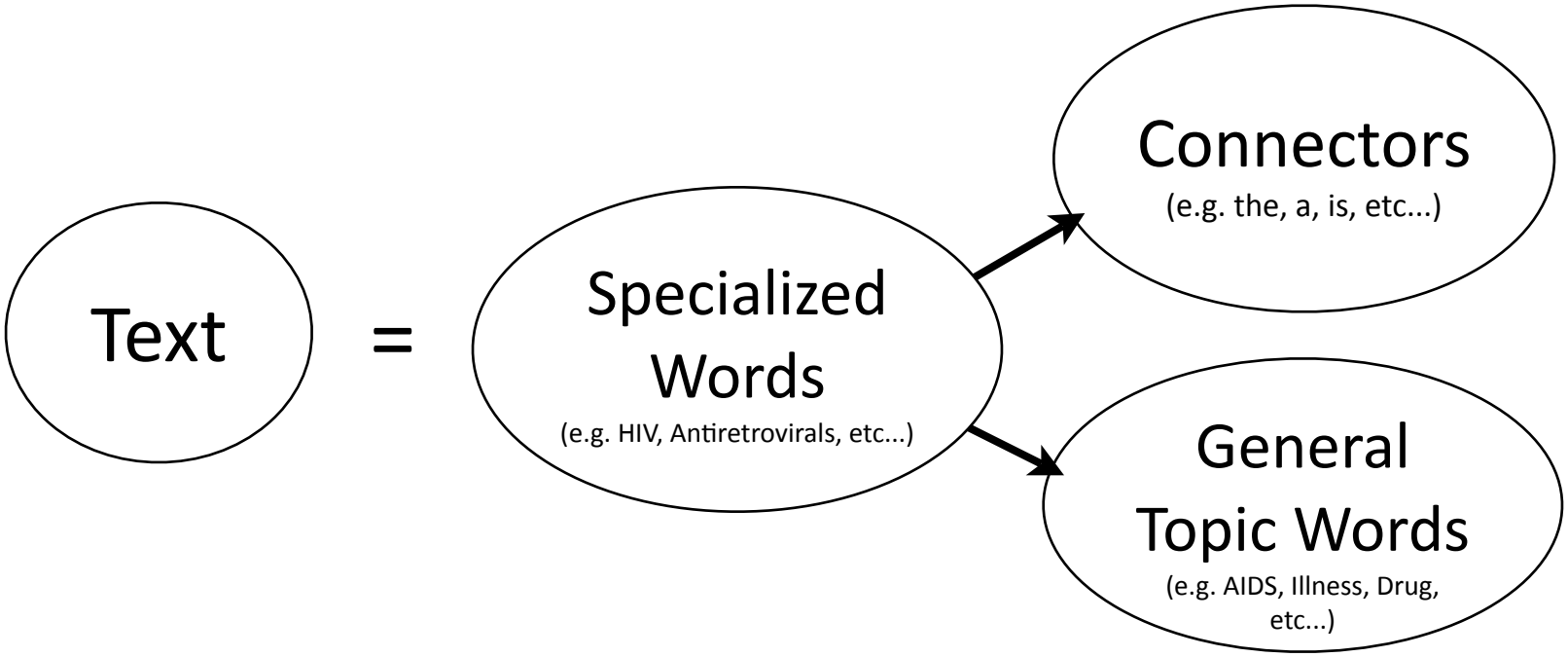
➡ Good benchmark to test differentiation of texts within the same context.



Text Modularity

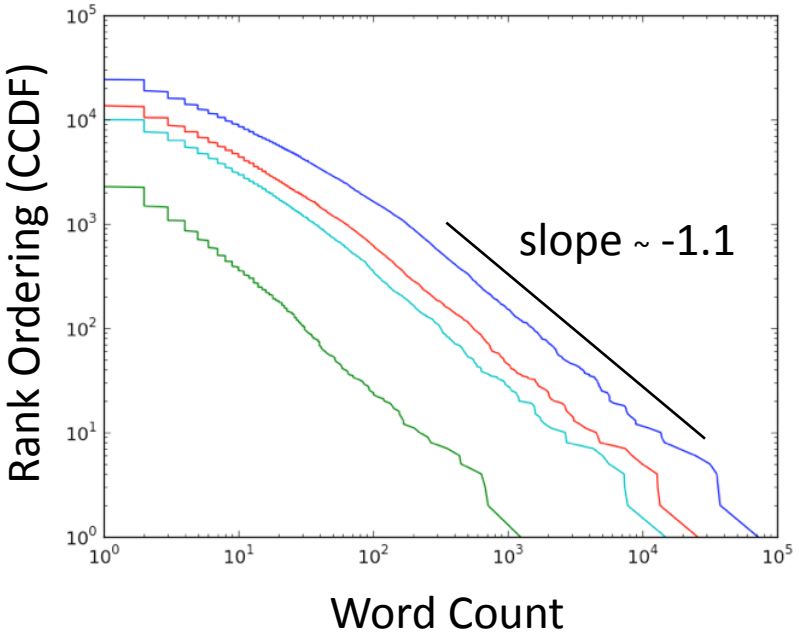
- unique set of more or less specialized words, arranged to convey a unique and precise message

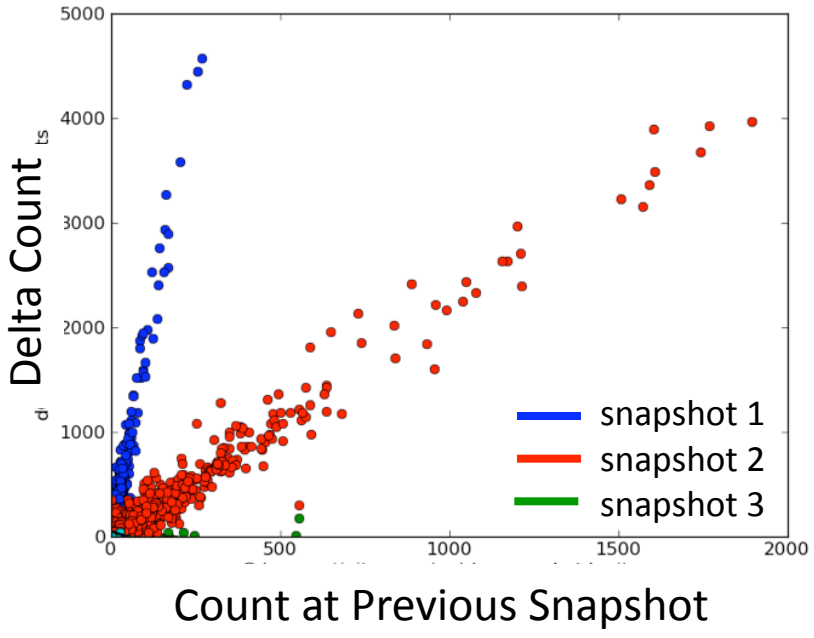
- modular structure (i.e. complex network) of words



Coarse Grained Evidence of Modularity and Self-Organization

Frequency of Word count and Growth between two random timesteps





$$dC = r(C)\ dt + \sigma(C)\ dW\ ,$$

➡ Proportional Growth with Multiplicative Stochastic Process

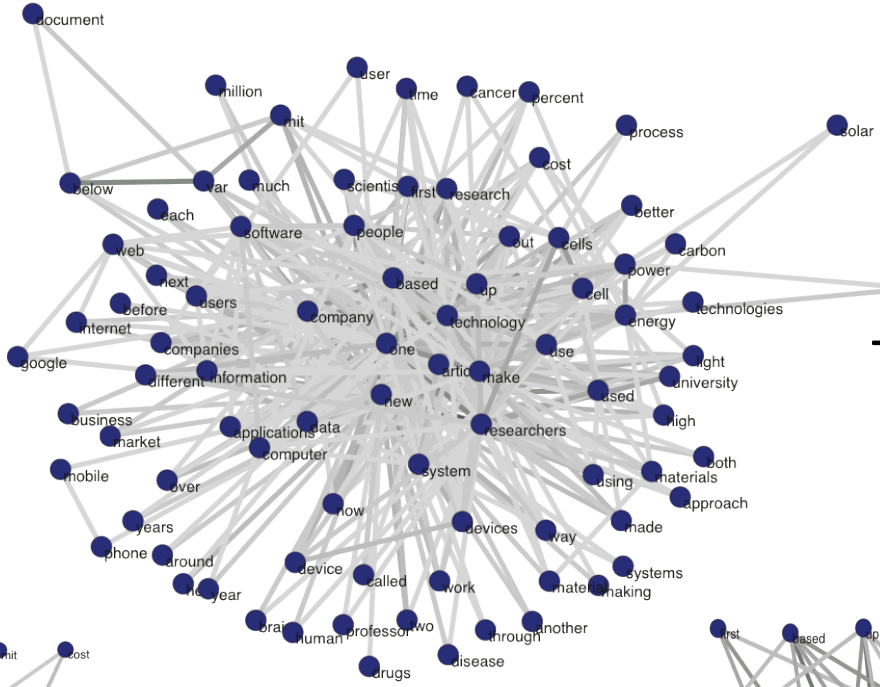
(empirical validation in progress)

Fine Grained Analysis

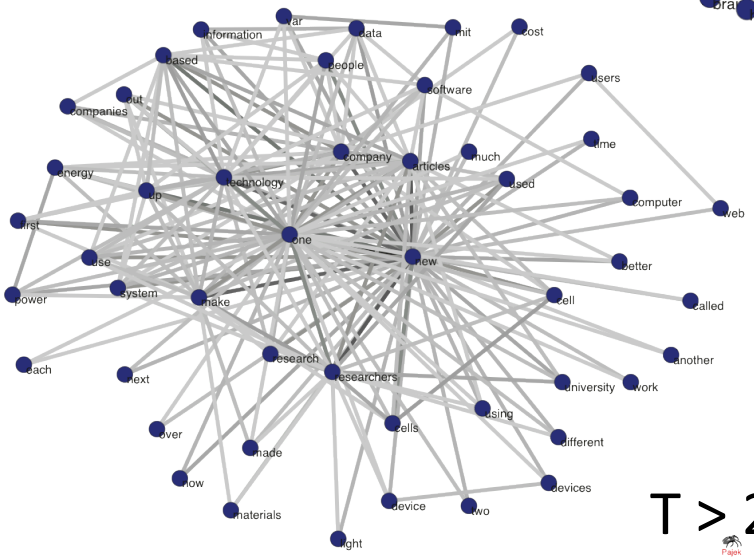
- Data
  - 5000+ blog posts (200-400 words each)
  - 50 most frequent words for each blog post
- Methods
  - Co-occurrence networks
  - Principal Component Analysis

Co-Occurrence Networks

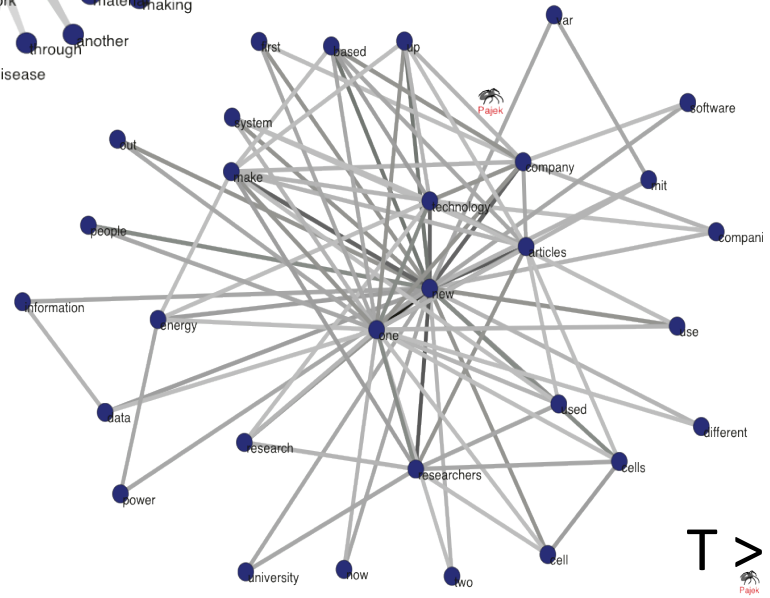
Threshold (T) by number of blogs



T > 150



T > 200



T > 250

Principal Component Analysis

- Rotate data to maximize variance
- Reduce dimension by projecting onto largest principal components (PCs)
- PCs are eigenvectors of data covariance matrix
- Variances of rotated data given by eigenvalues

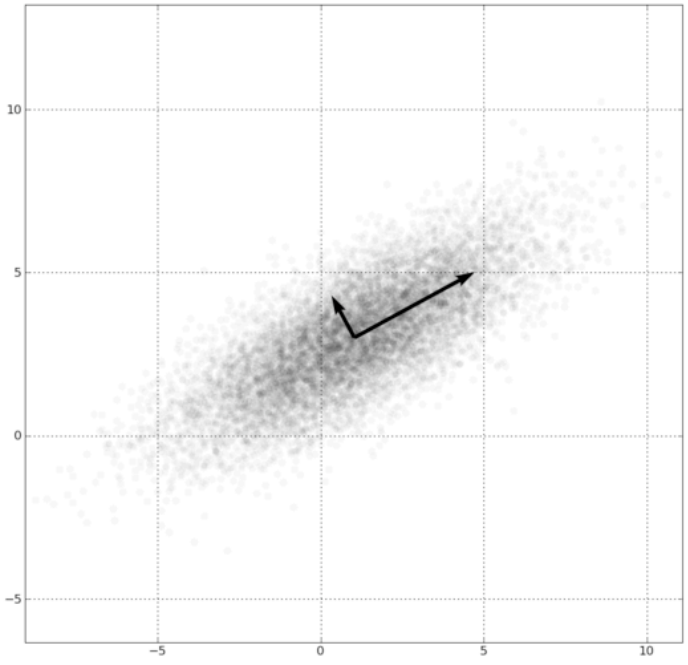
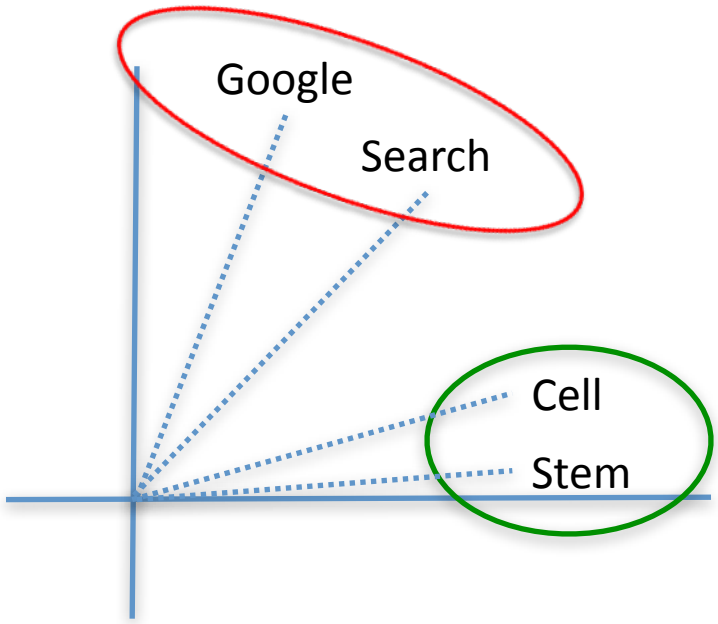


Image: Wikipedia

Clustering Words

- Embed words in R^100 using PCs
- Use k-means to cluster words
- “Distance” between words is angle between them in PC space
- Six clusters gave good results

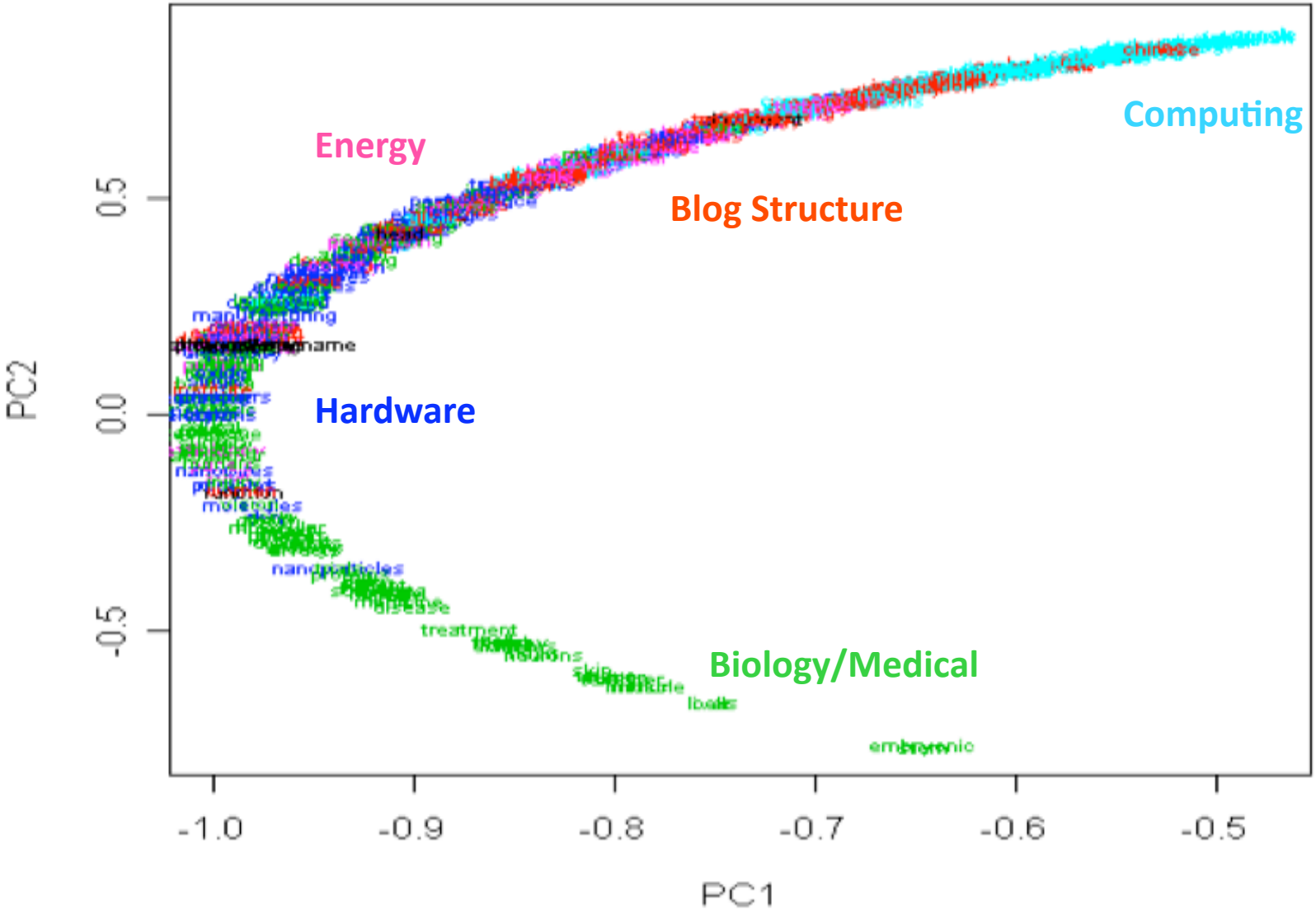


Word Clusters

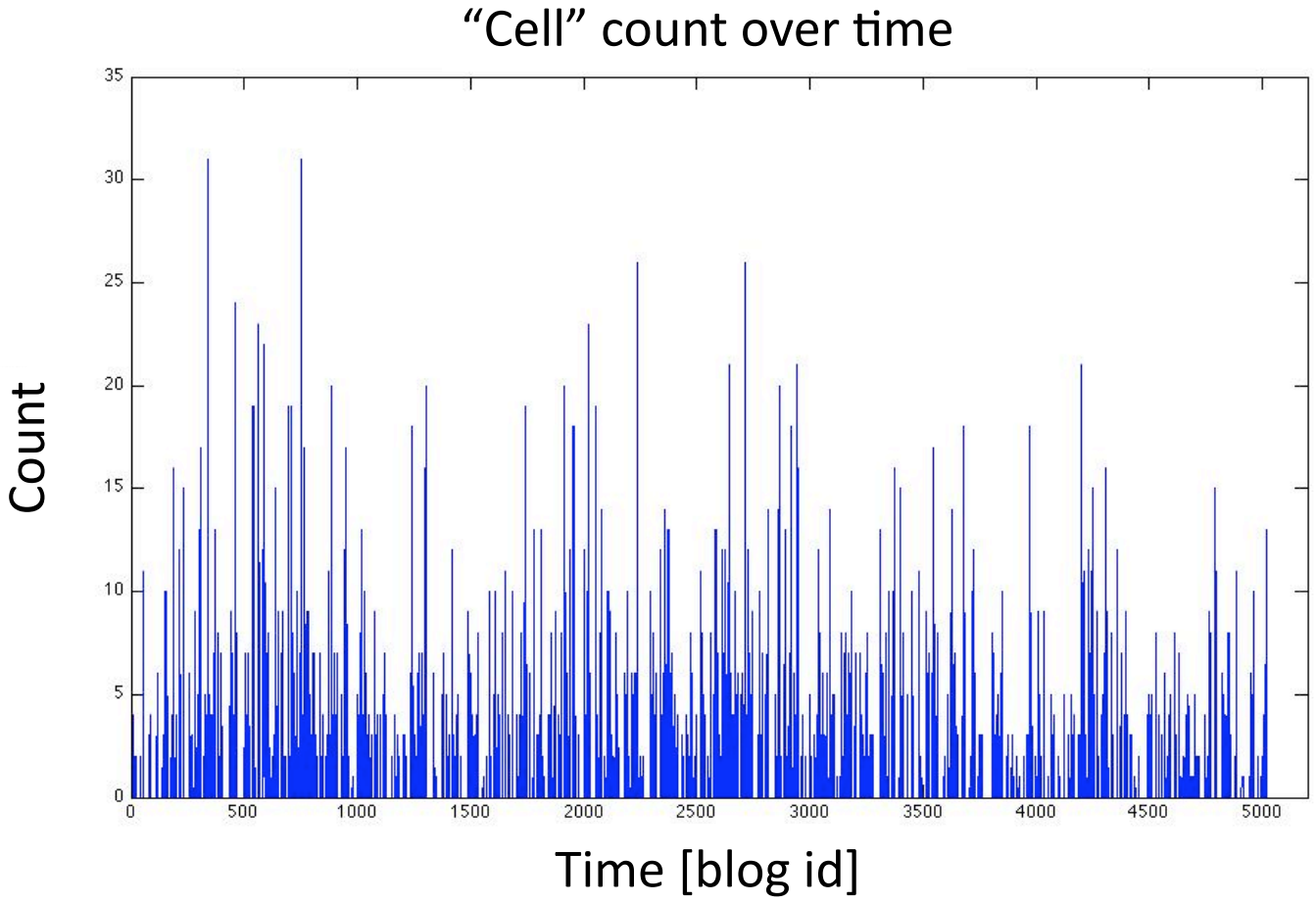
Blog Structure (112)	Biology/Medical (99)	Hardware (94)
mit	cells	silicon
technology	brain	device
research	cancer	nanotubes
engineering	stem	sensors
future	dna	chips

Computing (115)	Energy (66)	Random Code (14)
google	energy	var
software	solar	divhtml
cloud	carbon	getelementsbytagname
web	efficient	function
data	hybrid	hasmore

Word Clusters



Temporal Structure

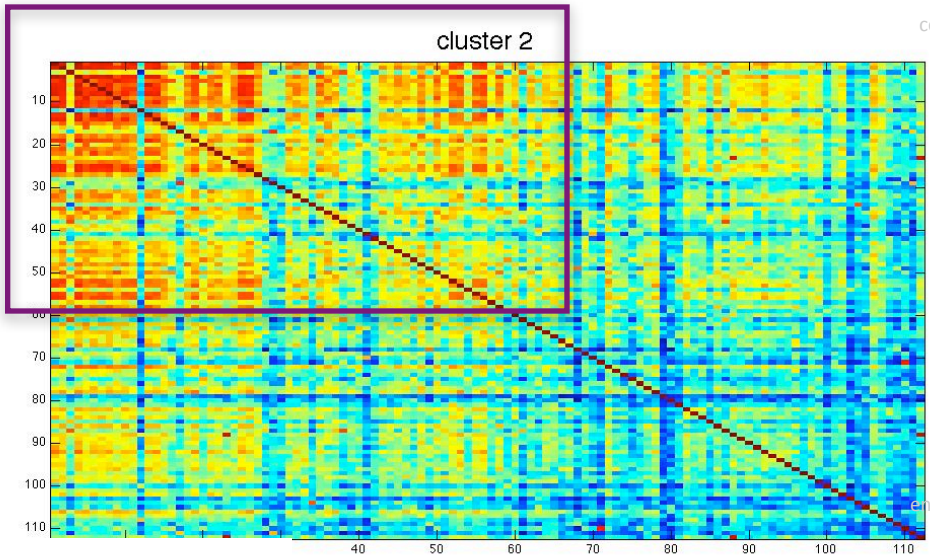


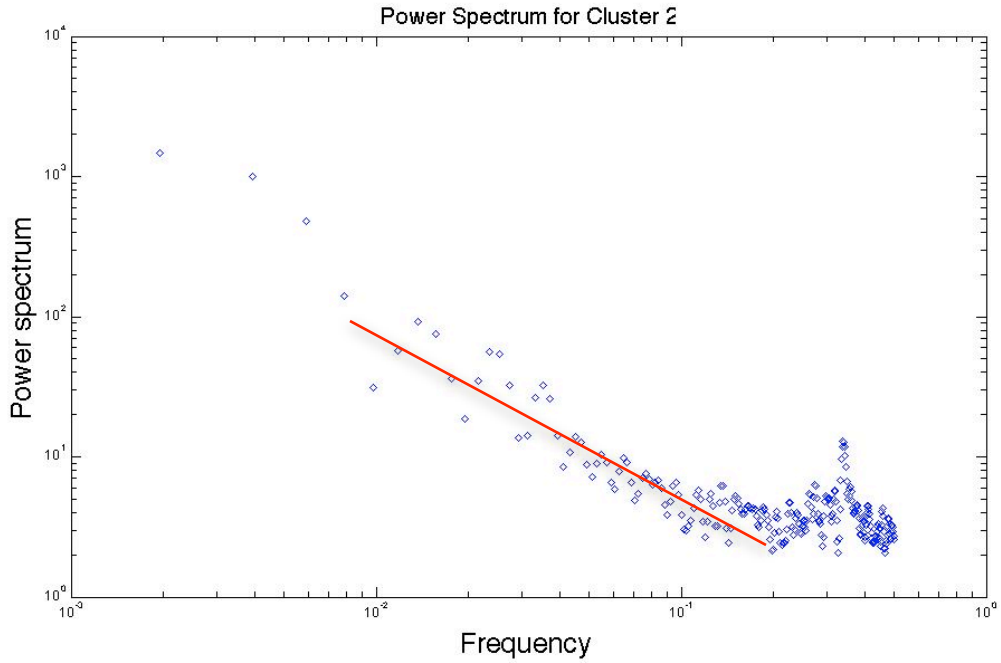
- Goal: find the coarse-graining for which structure appears

Temporal Structure

Single words → noisy series
Many words → memory is washed out

Considering clusters of words allows to reconstruct the structure present in the single words time series





2 regimes:
Short term memory
- every 50 blog posts

Long temporal correlations
- 1/f spectrum tilt

Future Work

1. Isolate words that “differentiate” two texts of the same topic
2. How fine grained dynamics lead to coarse grained self-organization?

- Tools:
    - Temporal latent variable methods (HMMs, Windowed-PCA, etc.)
    - Topic modeling (Bayesian)
    - Apply Kronecker graph framework
    - Spectral clustering
    - Partition Decoupling Method
- Scale up corpus size
    - More blogs
    - More keywords