Visually Exploring Documents: Topic Modeling

Working with text

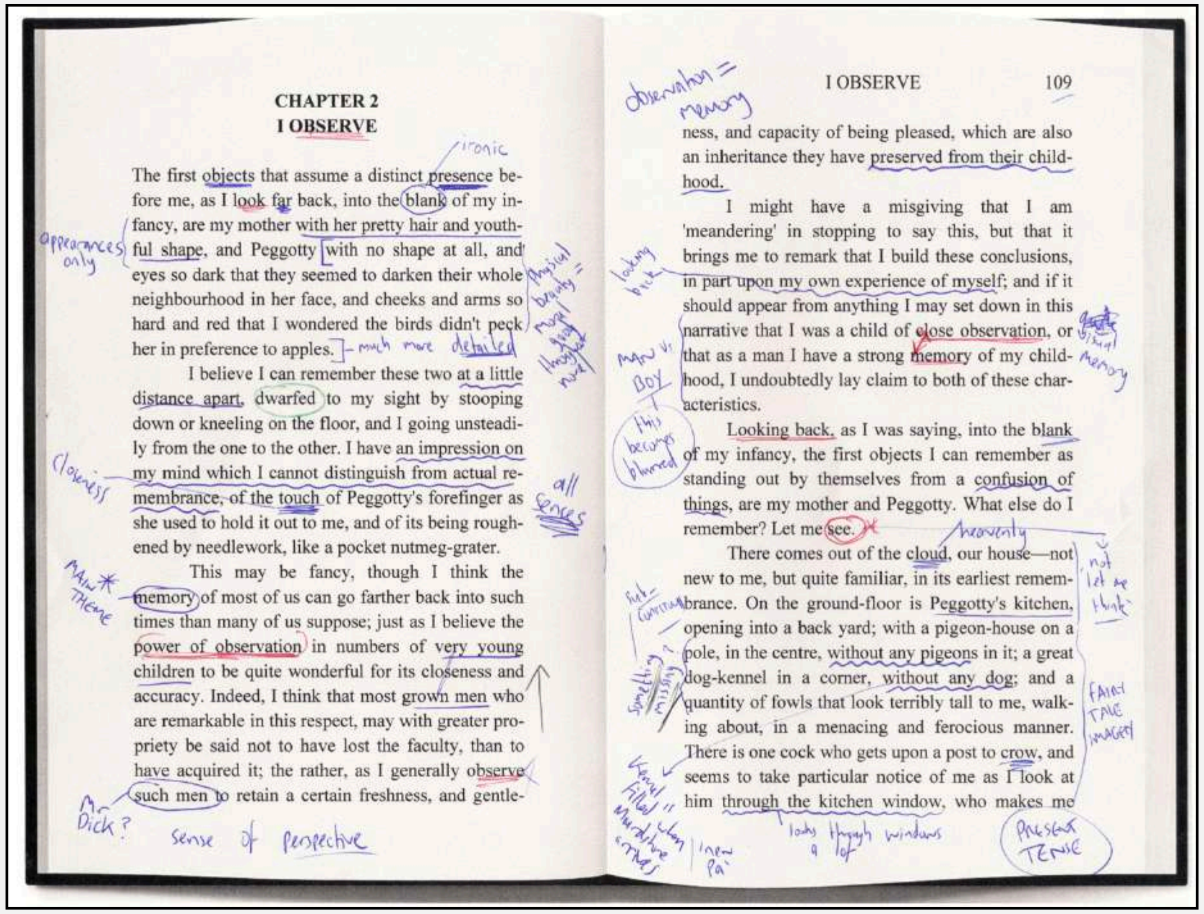
- “Text visualization” can mean many things
    - fonts, sizes, colors, kerning, typography in general
- We should consider text visualization from the more general definition of visualization: *amplify cognition*
    - Help people improve comprehension of a piece of text
    - Understand themes in text corpora
    - Locate/search relevant textual documents

Tasks in Text Visualization

- The way we approach text visualization is highly dependent on tasks.

- <u>Close reading</u>: deeply comprehending text, going beyond the words on the page

[Jänicke et al. 2015]



CHAPTER 2
I OBSERVE

The first objects that assume a distinct presence before me, as I look far back, into the blank of my infancy, are my mother with her pretty hair and youthful shape, and Peggotty with no shape at all, and eyes so dark that they seemed to darken their whole neighbourhood in her face, and cheeks and arms so hard and red that I wondered the birds didn't peck her in preference to apples.

I believe I can remember these two at a little distance apart, dwarfed to my sight by stooping down or kneeling on the floor, and I going unsteadily from the one to the other. I have an impression on my mind which I cannot distinguish from actual remembrance, of the touch of Peggotty's forefinger as she used to hold it out to me, and of its being roughened by needlework, like a pocket nutmeg-grater.

This may be fancy, though I think the memory of most of us can go farther back into such times than many of us suppose; just as I believe the power of observation in numbers of very young children to be quite wonderful for its closeness and accuracy. Indeed, I think that most grown men who are remarkable in this respect, may with greater propriety be said not to have lost the faculty, than to have acquired it; the rather, as I generally observe such men to retain a certain freshness, and gentleness, and capacity of being pleased, which are also an inheritance they have preserved from their childhood.

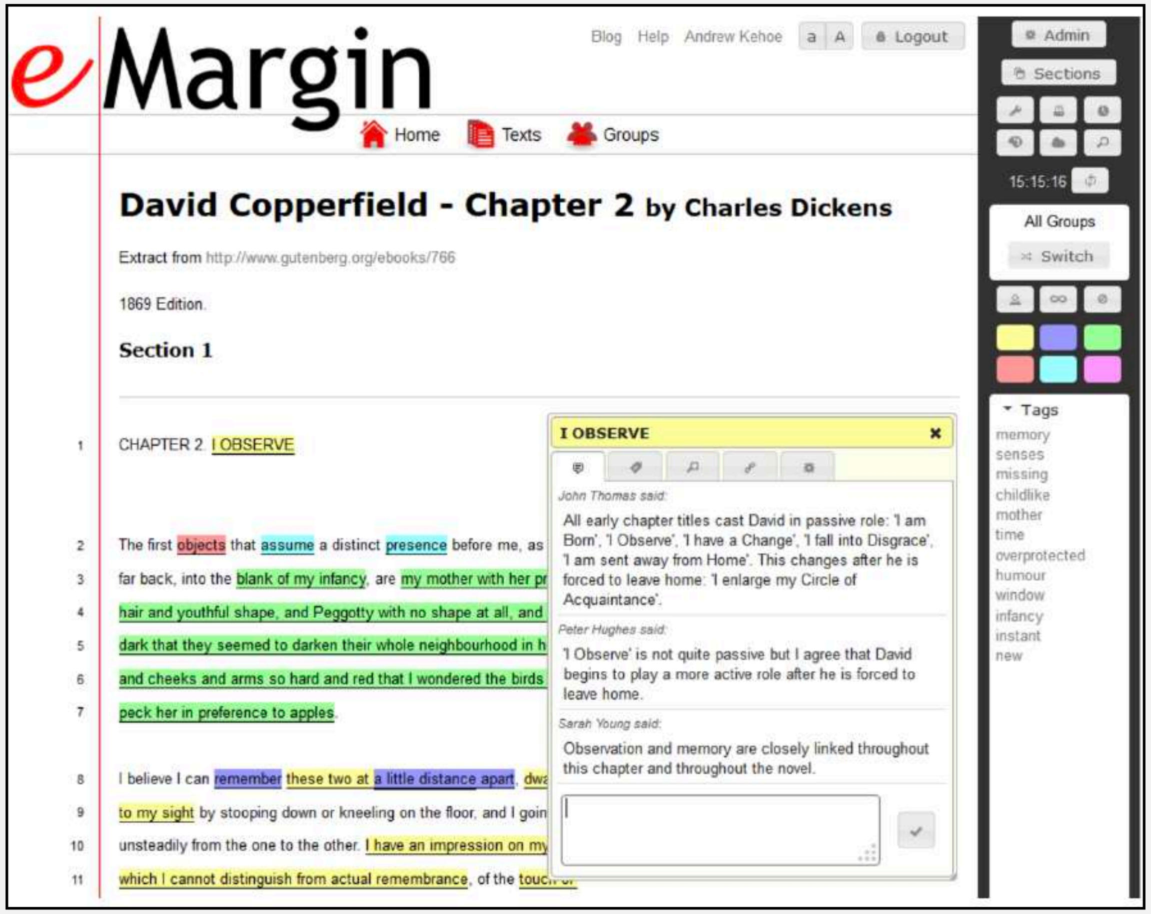
I might have a misgiving that I am 'meandering' in stopping to say this, but that it brings me to remark that I build these conclusions, in part upon my own experience of myself; and if it should appear from anything I may set down in this narrative that I was a child of close observation, or that as a man I have a strong memory of my childhood, I undoubtedly lay claim to both of these characteristics.

Looking back, as I was saying, into the blank of my infancy, the first objects I can remember as standing out by themselves from a confusion of things, are my mother and Peggotty. What else do I remember? Let me see.

There comes out of the cloud, our house—not new to me, but quite familiar, in its earliest remembrance. On the ground-floor is Peggotty's kitchen, opening into a back yard; with a pigeon-house on a pole, in the centre, without any pigeons in it; a great dog-kennel in a corner, without any dog; and a quantity of fowls that look terribly tall to me, walking about, in a menacing and ferocious manner. There is one cock who gets upon a post to crow, and seems to take particular notice of me as I look at him through the kitchen window, who makes me

I OBSERVE 109

ironic
appearances only
physical beauty = moral goodness
much more detailed
closeness
all senses
MAIN * THEME
Mr Dick?
sense of perspective
observation = memory
looking back
MAN vs BOY
this becomes blurred
visual memory
heavenly
not let me think
fast-forward
something missing?
FAIRY TALE IMAGERY
Kennel filled // when Murdstone enters
new Pa
looks through windows a lot
PRESENT TENSE



David Copperfield - Chapter 2 by Charles Dickens

Extract from http://www.gutenberg.org/ebooks/766

1869 Edition.

Section 1

CHAPTER 2. I OBSERVE

The first objects that assume a distinct presence before me, as
far back, into the blank of my infancy, are my mother with her pr
hair and youthful shape, and Peggotty with no shape at all, and
dark that they seemed to darken their whole neighbourhood in h
and cheeks and arms so hard and red that I wondered the birds
peck her in preference to apples.

I believe I can remember these two at a little distance apart, dwa
to my sight by stooping down or kneeling on the floor, and I goin
unsteadily from the one to the other. I have an impression on my
which I cannot distinguish from actual remembrance, of the touch of

I OBSERVE

John Thomas said:

All early chapter titles cast David in passive role: 'I am Born', 'I Observe', 'I have a Change', 'I fall into Disgrace', 'I am sent away from Home'. This changes after he is forced to leave home: 'I enlarge my Circle of Acquaintance'.

Peter Hughes said:

'I Observe' is not quite passive but I agree that David begins to play a more active role after he is forced to leave home.

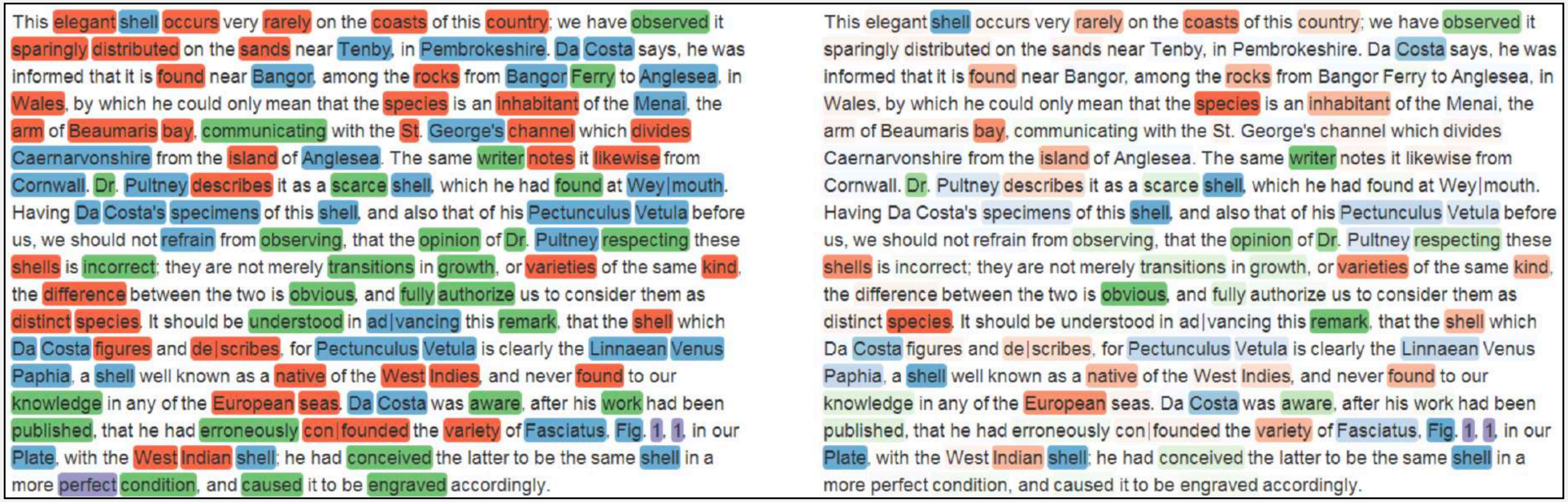
Sarah Young said:

Observation and memory are closely linked throughout this chapter and throughout the novel.

Close Reading Designs

Color

[Alexander et al. 2014]

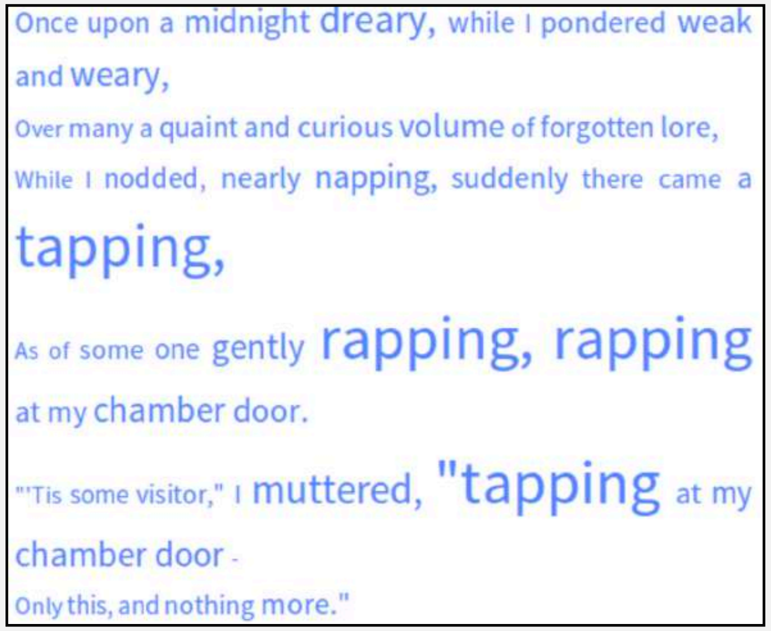


This elegant shell occurs very rarely on the coasts of this country; we have observed it sparingly distributed on the sands near Tenby, in Pembrokeshire. Da Costa says, he was informed that it is found near Bangor, among the rocks from Bangor Ferry to Anglesea, in Wales, by which he could only mean that the species is an inhabitant of the Menai, the arm of Beaumaris bay, communicating with the St. George's channel which divides Caernarvonshire from the island of Anglesea. The same writer notes it likewise from Cornwall. Dr. Pultney describes it as a scarce shell, which he had found at Wey|mouth. Having Da Costa's specimens of this shell, and also that of his Pectunculus Vetula before us, we should not refrain from observing, that the opinion of Dr. Pultney respecting these shells is incorrect; they are not merely transitions in growth, or varieties of the same kind, the difference between the two is obvious, and fully authorize us to consider them as distinct species. It should be understood in ad|vancing this remark, that the shell which Da Costa figures and de|scribes, for Pectunculus Vetula is clearly the Linnaean Venus Paphia, a shell well known as a native of the West Indies, and never found to our knowledge in any of the European seas. Da Costa was aware, after his work had been published, that he had erroneously con|founded the variety of Fasciatus, Fig. 1, 1, in our Plate, with the West Indian shell; he had conceived the latter to be the same shell in a more perfect condition, and caused it to be engraved accordingly.

This elegant shell occurs very rarely on the coasts of this country; we have observed it sparingly distributed on the sands near Tenby, in Pembrokeshire. Da Costa says, he was informed that it is found near Bangor, among the rocks from Bangor Ferry to Anglesea, in Wales, by which he could only mean that the species is an inhabitant of the Menai, the arm of Beaumaris bay, communicating with the St. George's channel which divides Caernarvonshire from the island of Anglesea. The same writer notes it likewise from Cornwall. Dr. Pultney describes it as a scarce shell, which he had found at Wey|mouth. Having Da Costa's specimens of this shell, and also that of his Pectunculus Vetula before us, we should not refrain from observing, that the opinion of Dr. Pultney respecting these shells is incorrect; they are not merely transitions in growth, or varieties of the same kind, the difference between the two is obvious, and fully authorize us to consider them as distinct species. It should be understood in ad|vancing this remark, that the shell which Da Costa figures and de|scribes, for Pectunculus Vetula is clearly the Linnaean Venus Paphia, a shell well known as a native of the West Indies, and never found to our knowledge in any of the European seas. Da Costa was aware, after his work had been published, that he had erroneously con|founded the variety of Fasciatus, Fig. 1, 1, in our Plate, with the West Indian shell; he had conceived the latter to be the same shell in a more perfect condition, and caused it to be engraved accordingly.

Size

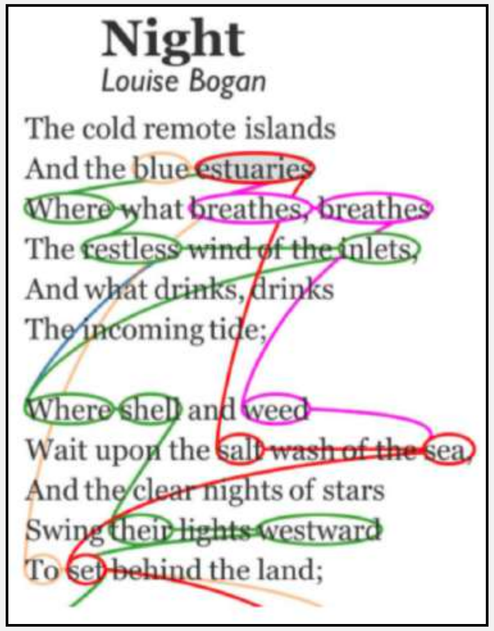
[Walsh et al. 2014]



Once upon a midnight dreary, while I pondered weak and weary,
Over many a quaint and curious volume of forgotten lore,
While I nodded, nearly napping, suddenly there came a tapping,
As of some one gently rapping, rapping at my chamber door.
"'Tis some visitor," I muttered, "tapping at my chamber door -
Only this, and nothing more."

Node-link Diagrams

[Coles et al. 2014]



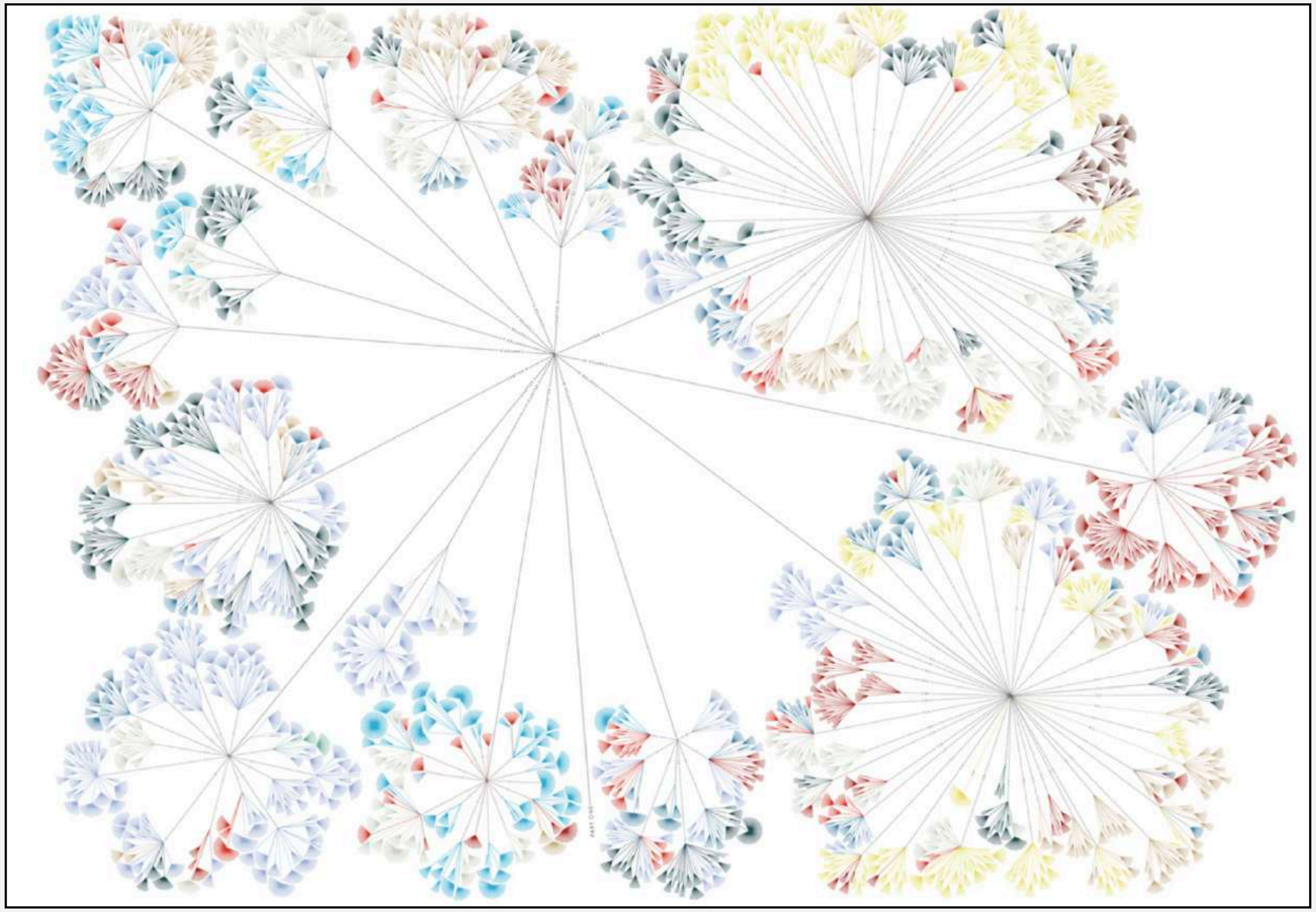
Night
Louise Bogan

The cold remote islands
And the blue estuaries
Where what breathes, breathes
The restless wind of the inlets,
And what drinks, drinks
The incoming tide;

Where shell and weed
Wait upon the salt wash of the sea,
And the clear nights of stars
Swing their lights westward
To set behind the land;

Distant Reading

- Understand structure, relationships, themes, connections within a document.



[Posavec 2007]

Document Exploration

- A “document”: sequence of words.
    - A book, a wikipedia article, a paper abstract, a tweet
- Can view as generic high-dimensional data, apply techniques we’ve discussed so far for visual exploration
- But documents are unique:
    - Really high-dimensional (e.g. dimensionality = size of English vocabulary)
    - Usually sparse
    - Yet, each dimension - word - is interpretable

Document Exploration Tasks

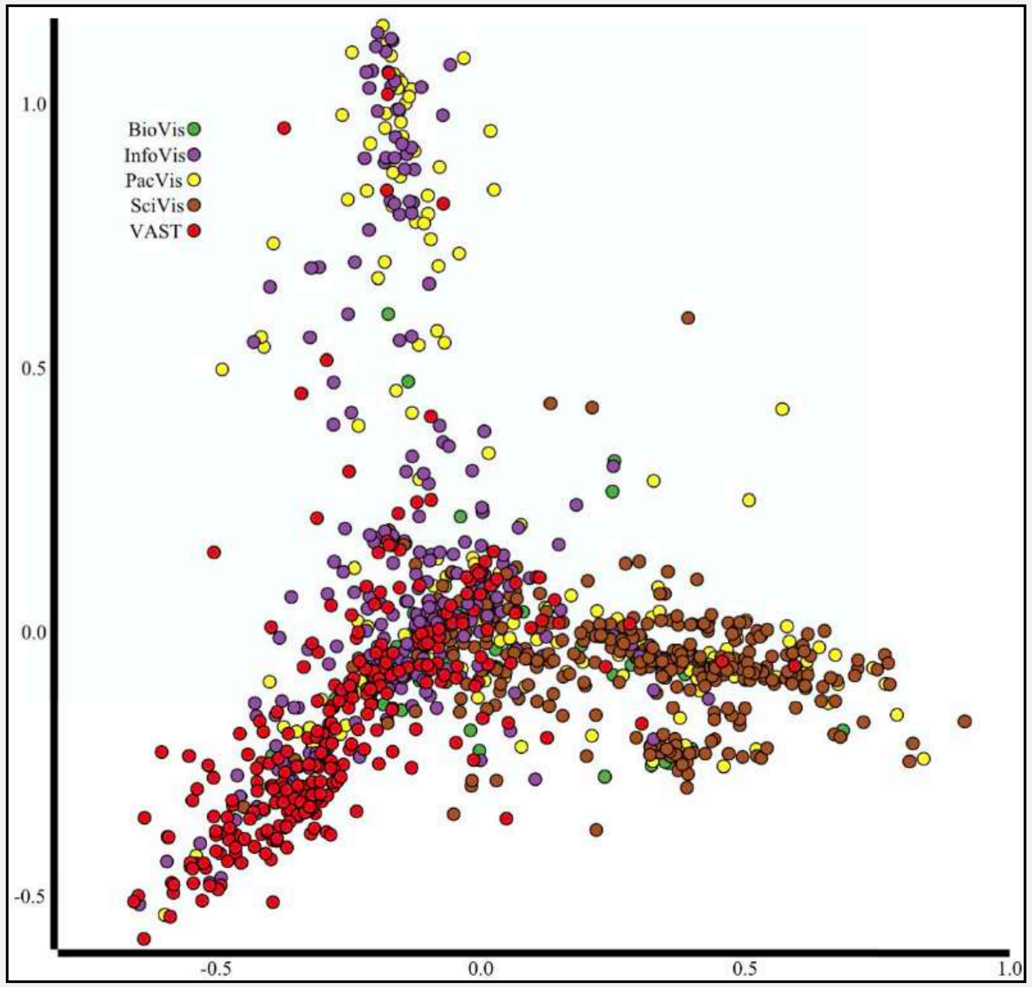
- Suppose you were provided hundreds of articles relevant to your respective research backgrounds.

- What tasks are relevant to you for gaining an understanding - and advancing your research - on the document corpus?

Exploring Documents: Dimensionality Reduction

- We represent each document as a point in a high-dimensional space.
  - Each dimension is a word
  - The value of a dimension is the word count: the number of times the word appeared in the document

[Anderson et al. 2014]



Limitations with this approach?

Matrix Factorization

- Dimensionality reduction is … too reductive. We seek better **representations** of documents.

- Let's consider **matrix factorization**.

$W \in \mathbb{R}^{n\times d} \longrightarrow \|W - \tilde{U}\tilde{V}^T\|$ $\quad\quad$ $\tilde{U} \in \mathbb{R}^{n\times k}$ $\quad$ $\tilde{V} \in \mathbb{R}^{d\times k}$

$\mathbf{w}_i \approx \tilde{\mathbf{u}}_i\tilde{V}^T$ a reconstruction of word counts for document *i*

$\tilde{\mathbf{u}}_i$ assigns importance to each column of $\tilde{V}$

$\tilde{V}$ each column vector: a set of weights, one for each word → **shared across documents**

k determines approximation quality

$k \ll \min(n, d)$ we share limited information across documents for reconstruction

Ideally: columns represent meaningful, latent factors shared across documents

Optimistic, but unrealistic, example

- We set *k* to 2

$$\tilde{V} = [\tilde{\mathbf{v}}_1 \, \tilde{\mathbf{v}}_2] \qquad \tilde{\mathbf{v}}_1 \in \mathbb{R}^d \qquad \tilde{\mathbf{v}}_2 \in \mathbb{R}^d$$

- Then we can represent each document with 2 numbers

$$\tilde{\mathbf{w}}_i = [w_{i1}, w_{i2}]$$

$\tilde{\mathbf{w}}_i = [1,0]$ take on words from $\tilde{\mathbf{v}}_1$

$\tilde{\mathbf{w}}_i = [0,1]$ take on words from $\tilde{\mathbf{v}}_2$

- Does this look familiar? What are we doing with documents in this manner?

- Clustering! (kind of)

$\tilde{\mathbf{v}}_1 \in \mathbb{R}^d$ **set of words that describe one cluster**

$\tilde{\mathbf{v}}_2 \in \mathbb{R}^d$ **set of words that describe other cluster**

Matrix Factorization: SVD

- Clustering interpretation only meaningful if the reconstruction is good!

$$\|W - \tilde{U}\tilde{V}^T\|_F$$

- We can find the global minimum of this energy via the singular value decomposition (SVD)

$$W = U\Lambda V^T \, , U \in \mathbb{R}^{n\times n} \, , \Lambda \in \mathbb{R}^{n\times d} \, , V \in \mathbb{R}^{d\times d} \qquad U^TU = I \, , V^TV = I$$

- Truncate the SVD to the top k singular values

$$\tilde{U} = U_{1:k}\sqrt{\Lambda_{1:k}} \qquad \tilde{V} = V_{1:k}\sqrt{\Lambda_{1:k}}$$

- Can write approximation as the following expansion:

$$W \approx \sum_{i=1}^{k} \lambda_i \mathbf{u}_i \mathbf{v}_i^T$$ **best rank-k approximation**

Nonnegative Matrix Factorization

- Limitation: the document and latent factors can be negative - not easily interpretable! Bad for visualization.

- So, then, let's enforce nonnegativity!

$$\|W - UV^T\|_F^2 , \quad s.t.\ U, V \geq 0$$

- We can then say "latent factors *V* contribute *u* amount to the reconstruction"

- Challenge: fixing *U*, energy is convex in *V*. Fixing *V*, energy is convex in *U*. But not convex in both! (due to nonnegativity constraint - cannot apply Eckart-Young theorem)

Algorithm: Multiplicative Updates

$$\|W - UV^T\|_F^2 \,, \quad s\,.\,t\,.\;U, V \geq 0$$

- Alternate between the following updates:

$$V_{ab} \leftarrow V_{ab} \frac{(U^T W)_{ab}}{(U^T U V^T)_{ab}} \qquad U_{ab} \leftarrow U_{ab} \frac{(WV)_{ab}}{(UV^T V)_{ab}}$$

- Update can be seen as a particular (per-element) step size chosen for gradient descent:

$$\begin{aligned} E(V) = \|W - UV^T\|_F^2 &= tr((W - UV^T)^T (W - UV^T)) \\ &= tr(W^T W - 2VU^T W + VU^T UV^T) \end{aligned}$$

- Take derivative of trace:

$$\frac{\mathbf{d}E}{\mathbf{d}V^T} = 2(U^T U V^T - U^T W)$$

Algorithm: Multiplicative Updates continued…

$$V_{ab} \leftarrow V_{ab} - \eta_{ab}(U^TUV^T - U^TW)_{ab} \quad , \quad \eta_{ab} = \frac{V_{ab}}{(U^TUV^T)_{ab}}$$

$$V_{ab} \leftarrow V_{ab} - \frac{V_{ab}}{(U^TUV^T)_{ab}}(U^TUV^T - U^TW)_{ab} = \cancel{V_{ab} - V_{ab}} + \underline{V_{ab}\frac{(U^TW)_{ab}}{(U^TUV^T)_{ab}}}$$

- Starting from an initial guess for *U* and *V* that are both nonnegative, this scheme ensures:
  - Both will remain nonnegative
  - Energy decreases at each iteration
  - Will arrive at *some* fixed-point solution (local minimum)

[Lee & Seung 2001]

Still, some limitations

- Weights are unbounded:
  - Some latent factors could dominate others.
  - Thus, document weights become hard to interpret.
- At this point, might start introducing regularization terms.
- However, consider the following probabilistic interpretation:

$$w_{ij} \approx \mathbf{u}_i^T \mathbf{v}_j = \sum_{l=1}^{k} u_{il} v_{jl} \longrightarrow \sum_{l} \underline{p(z_l \,|\, \theta)}\,\underline{p(w \,|\, z_l, \beta)}$$

probability of latent factor, given a document

probability of word, given latent factor

Latent Dirichlet Allocation

- Next, let's consider *all* words in a document:

$$p(\theta, \mathbf{w} \mid \alpha, \beta) = \underline{p(\theta \mid \alpha)} \prod_{n} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta)$$

Probability of latent factors - or **topics**

- We marginalize out θ

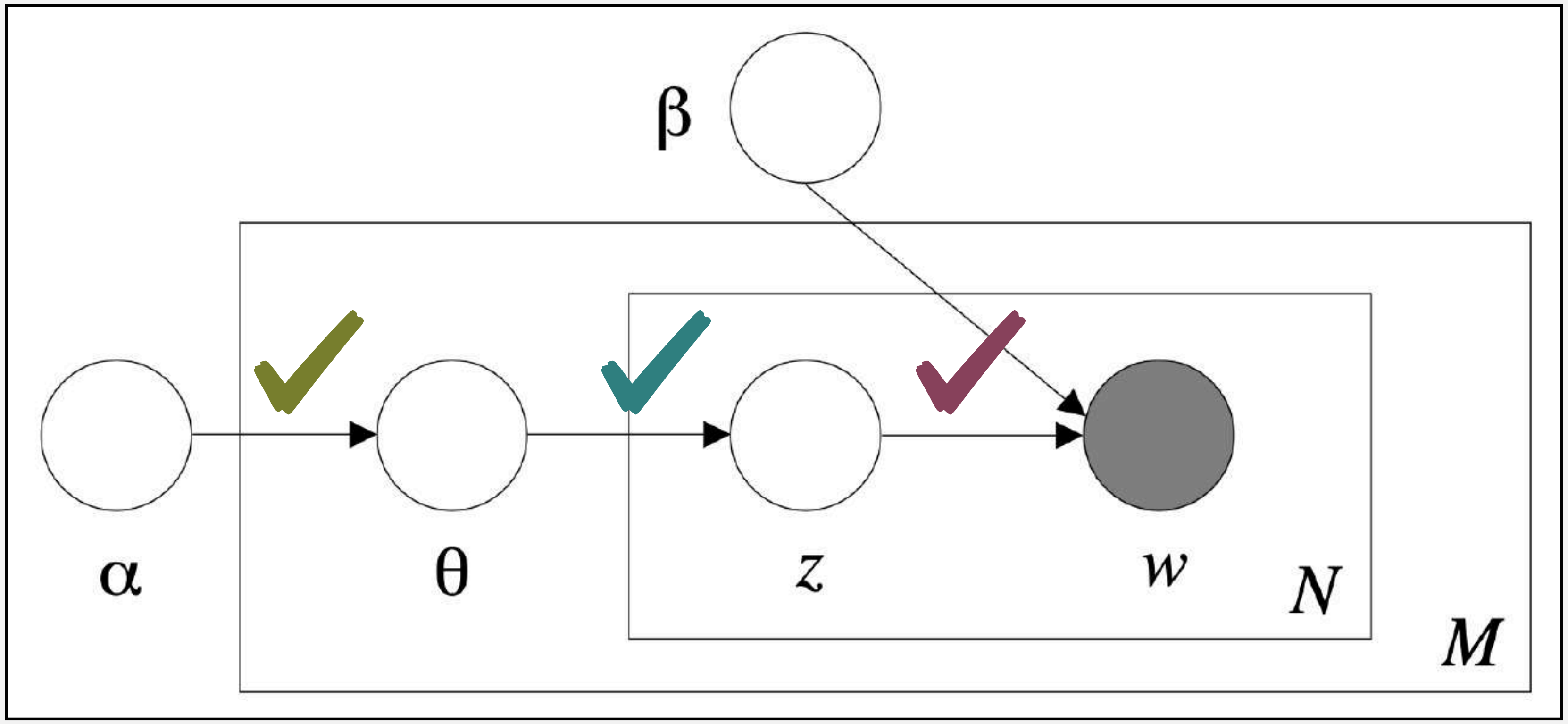
$$p(\mathbf{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta$$

- Last, consider *all* documents:

$$p(\mathbf{w} \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d$$

Probabilistic Model

$$p(\mathbf{w} \mid \alpha, \beta) = \prod_{d=1}^{M} \int \underline{p(\theta_d \mid \alpha)} \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} \underline{p(z_{dn} \mid \theta_d)} \underline{p(w_{dn} \mid z_{dn}, \beta)} \right) d\theta_d$$

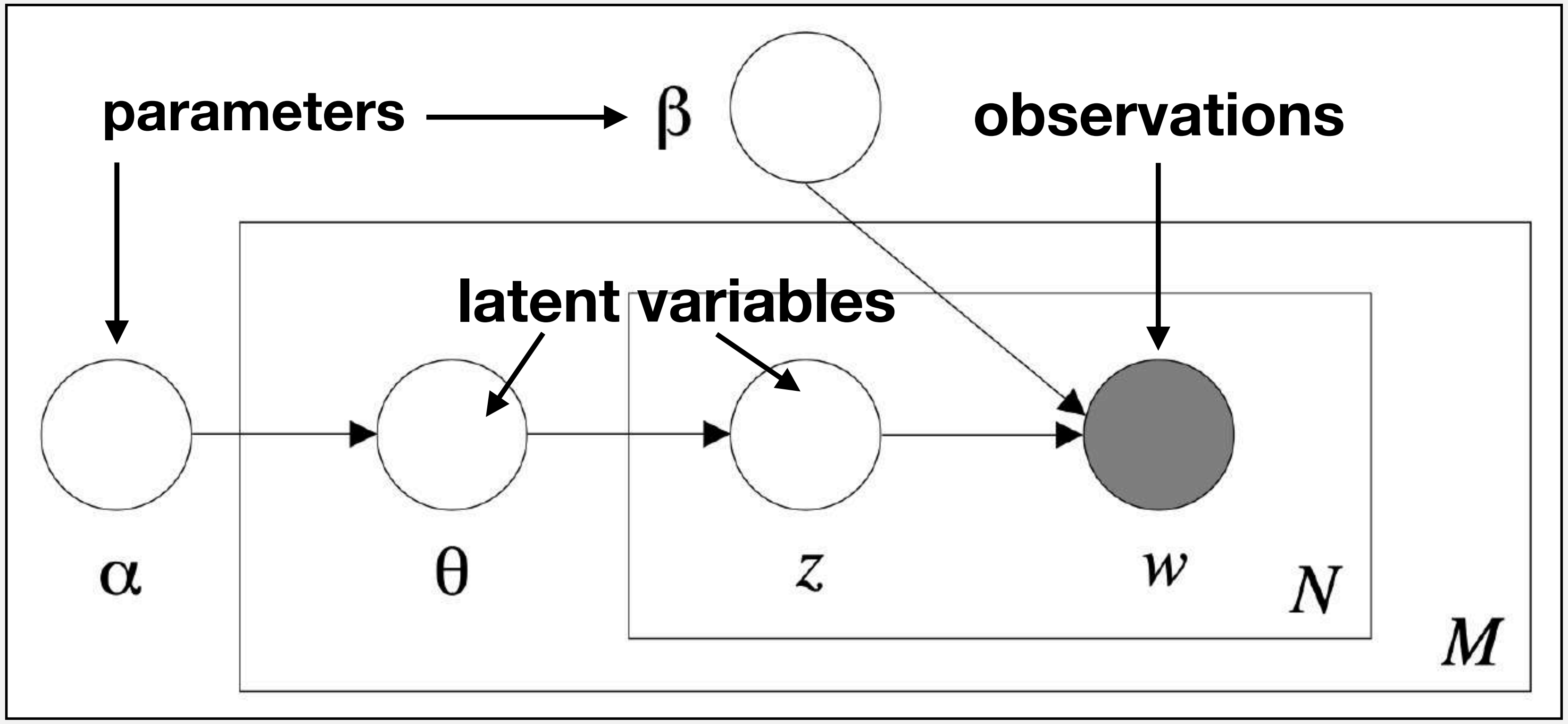


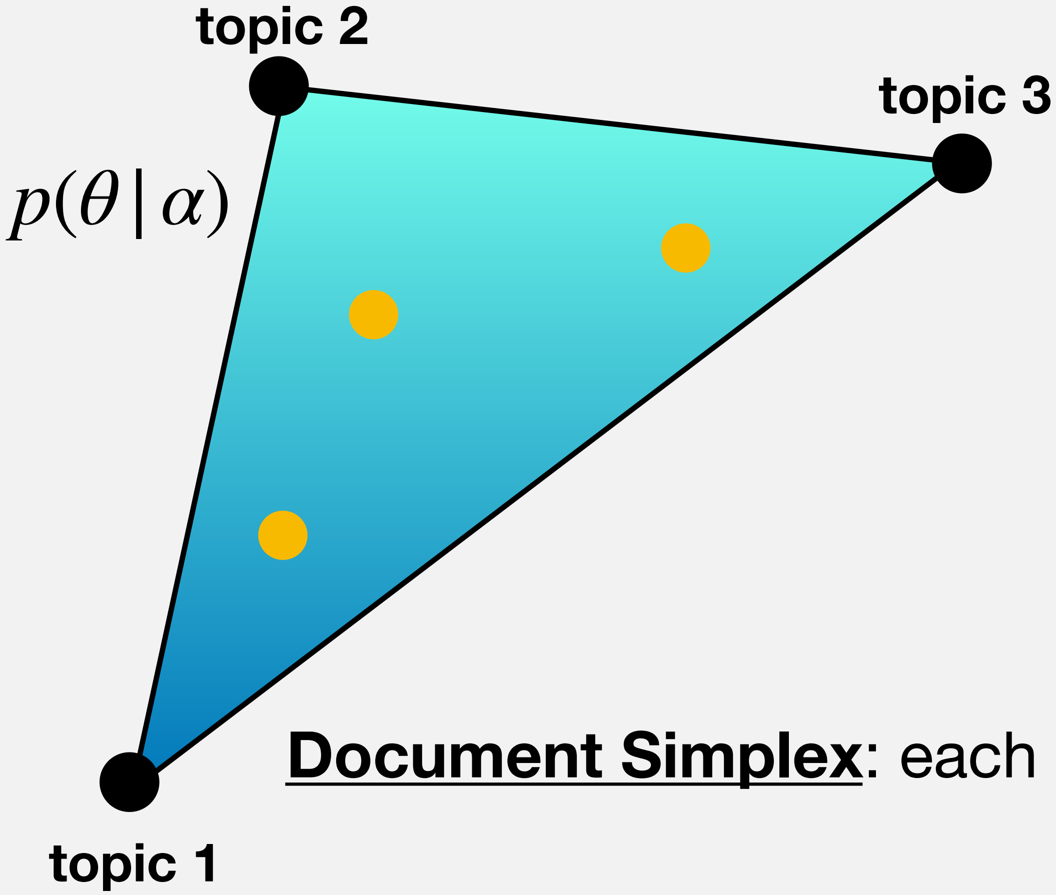
For a given document, draw a mixture of topics

To generate a word, first draw a random topic, given the mixture

Last, sample a word from the topic

Probabilistic Model



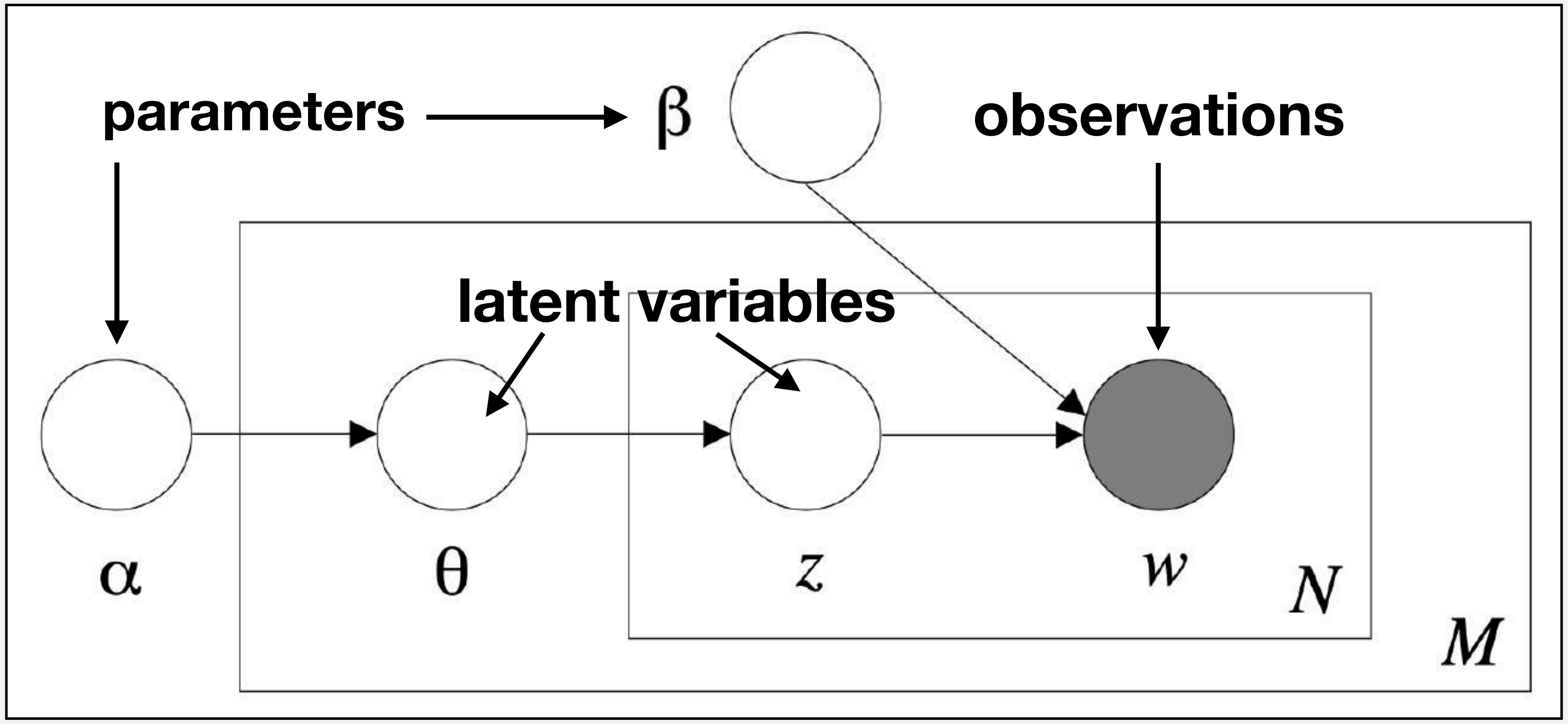


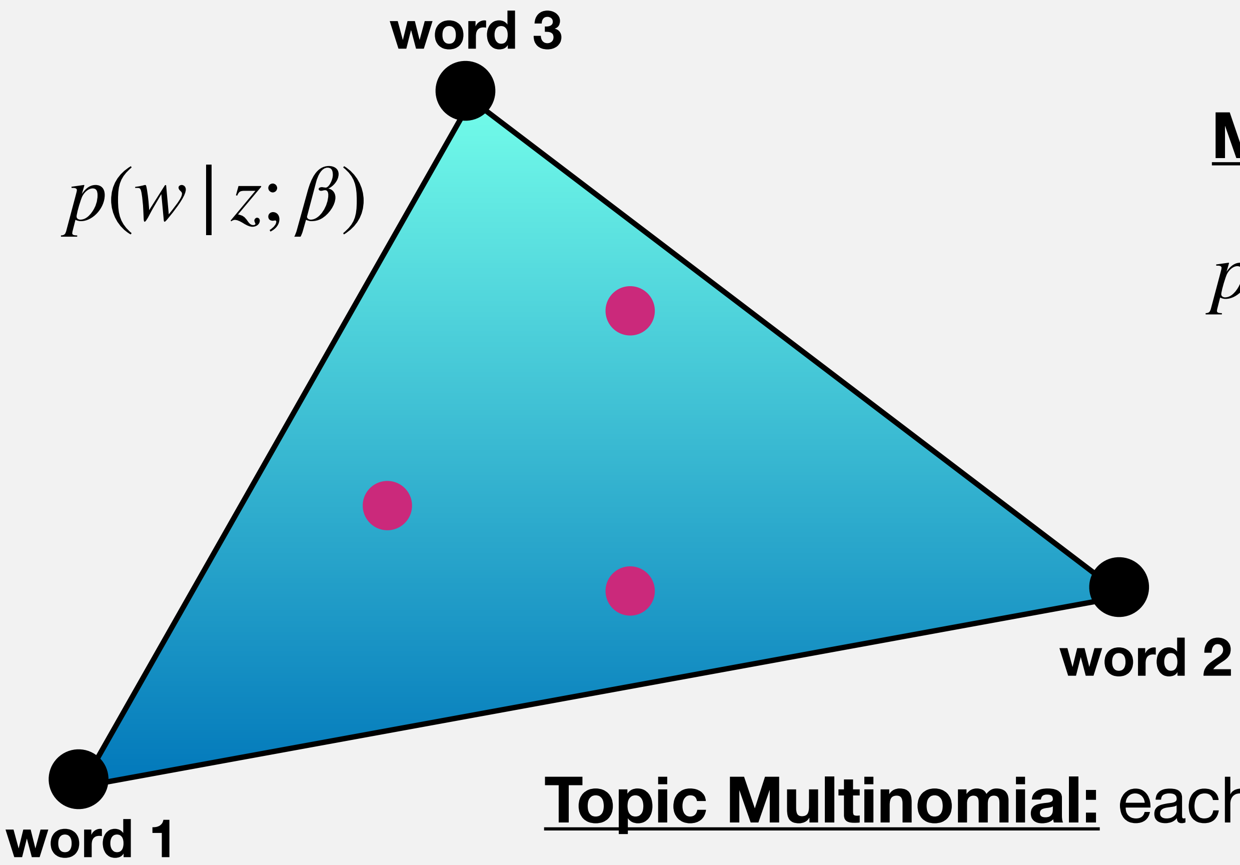
Dirichlet distribution

$$p(\theta | \alpha) = Z(\alpha) \prod_i \theta_i^{\alpha_i - 1}$$

Document Simplex: each point is a mixture over topics

Probabilistic Model



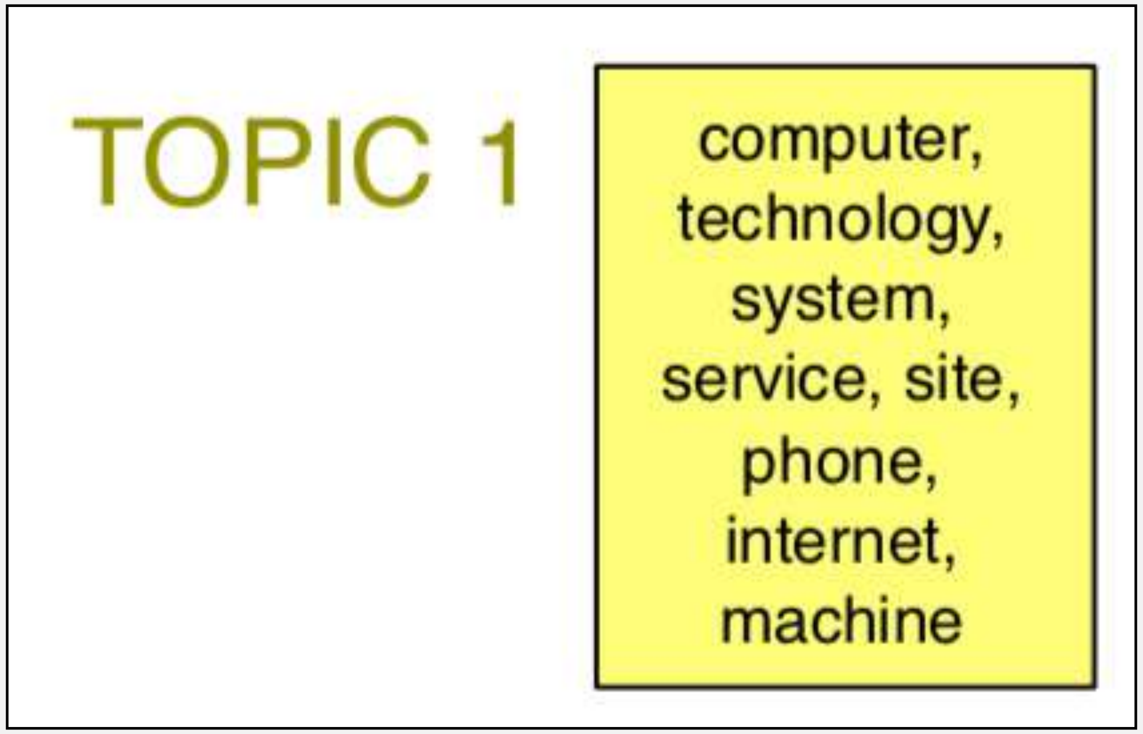


Multinomial distribution

$$p(w \mid z; \beta) = Z(\beta_{z_i}) \prod_i w_i^{\beta_{z_i}}$$

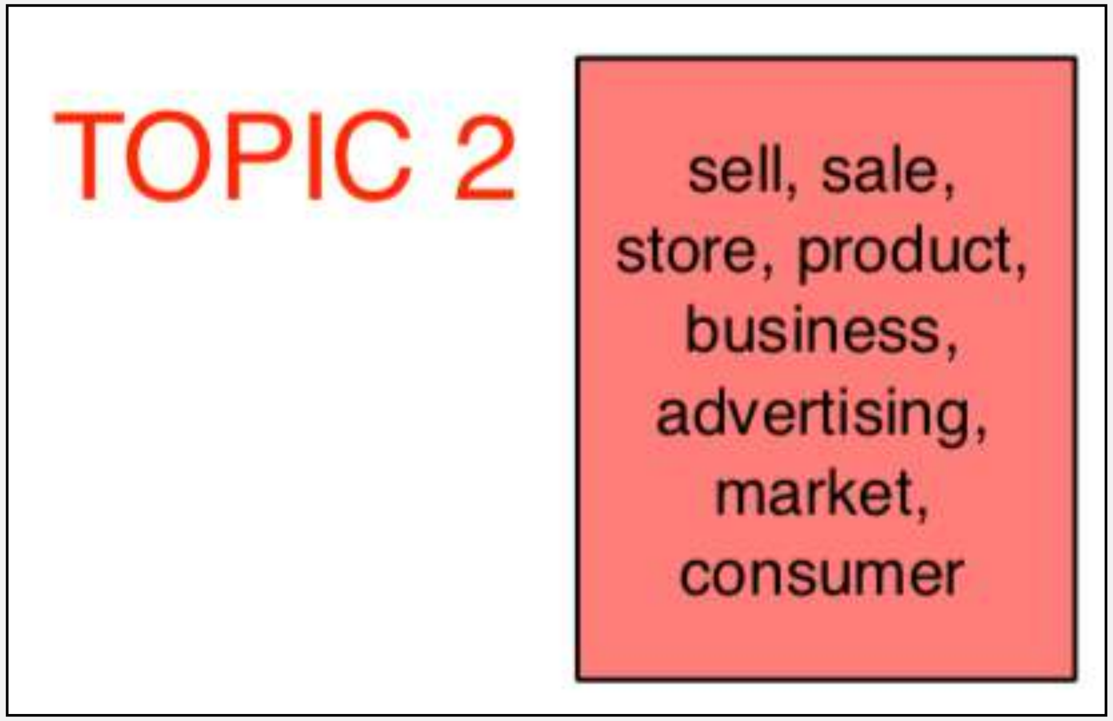
Topic Multinomial: each point is a distribution over words

Generative Model



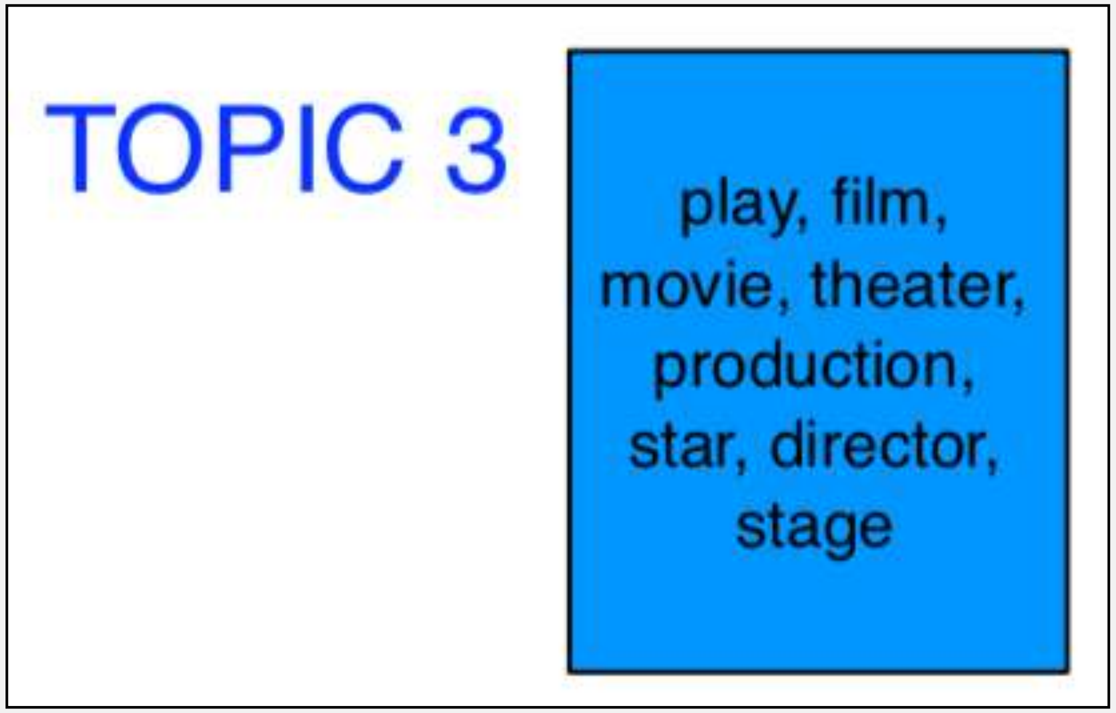
TOPIC 1

computer, technology, system, service, site, phone, internet, machine



TOPIC 2

sell, sale, store, product, business, advertising, market, consumer

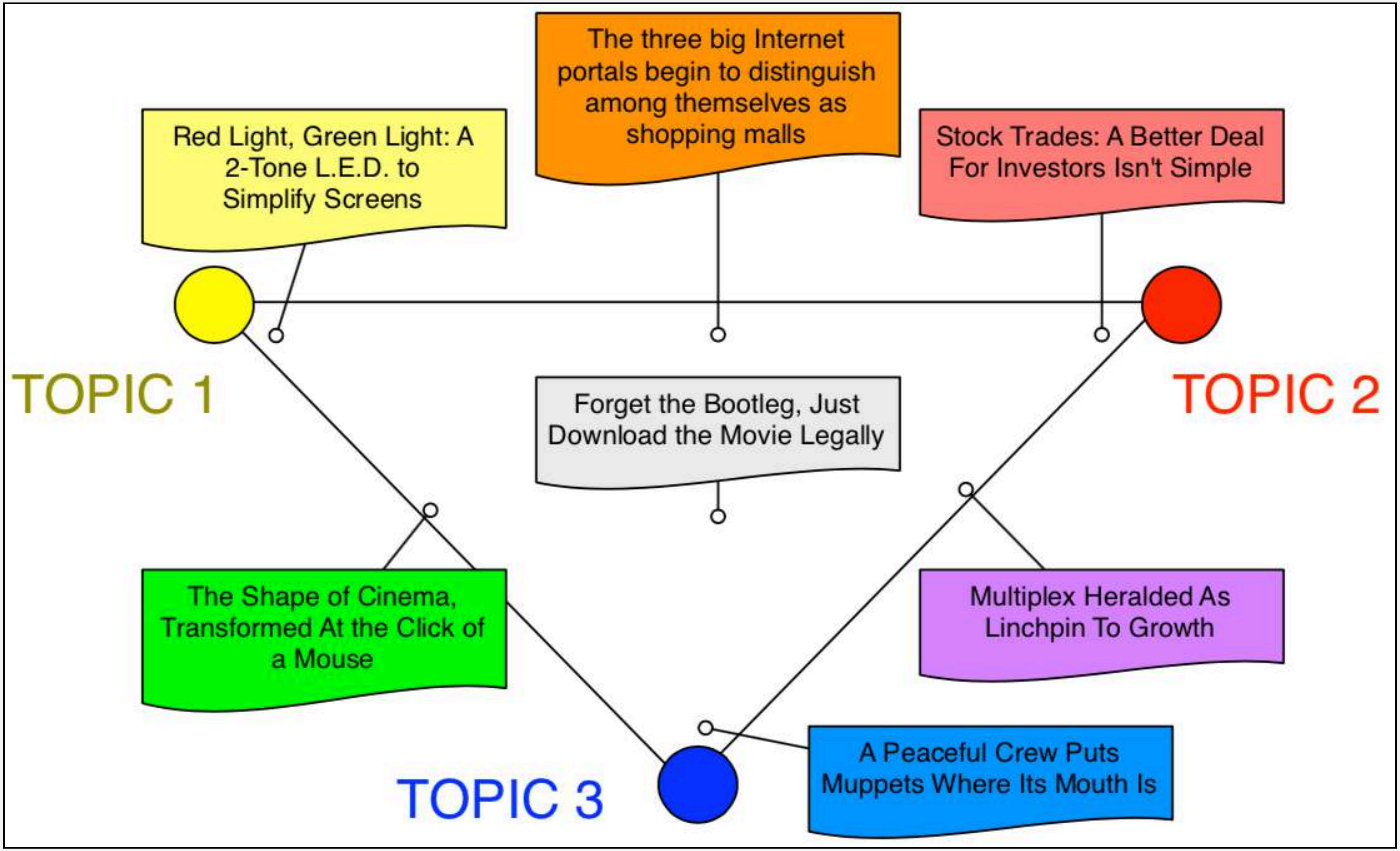


TOPIC 3

play, film, movie, theater, production, star, director, stage

(slide from David Mimno)

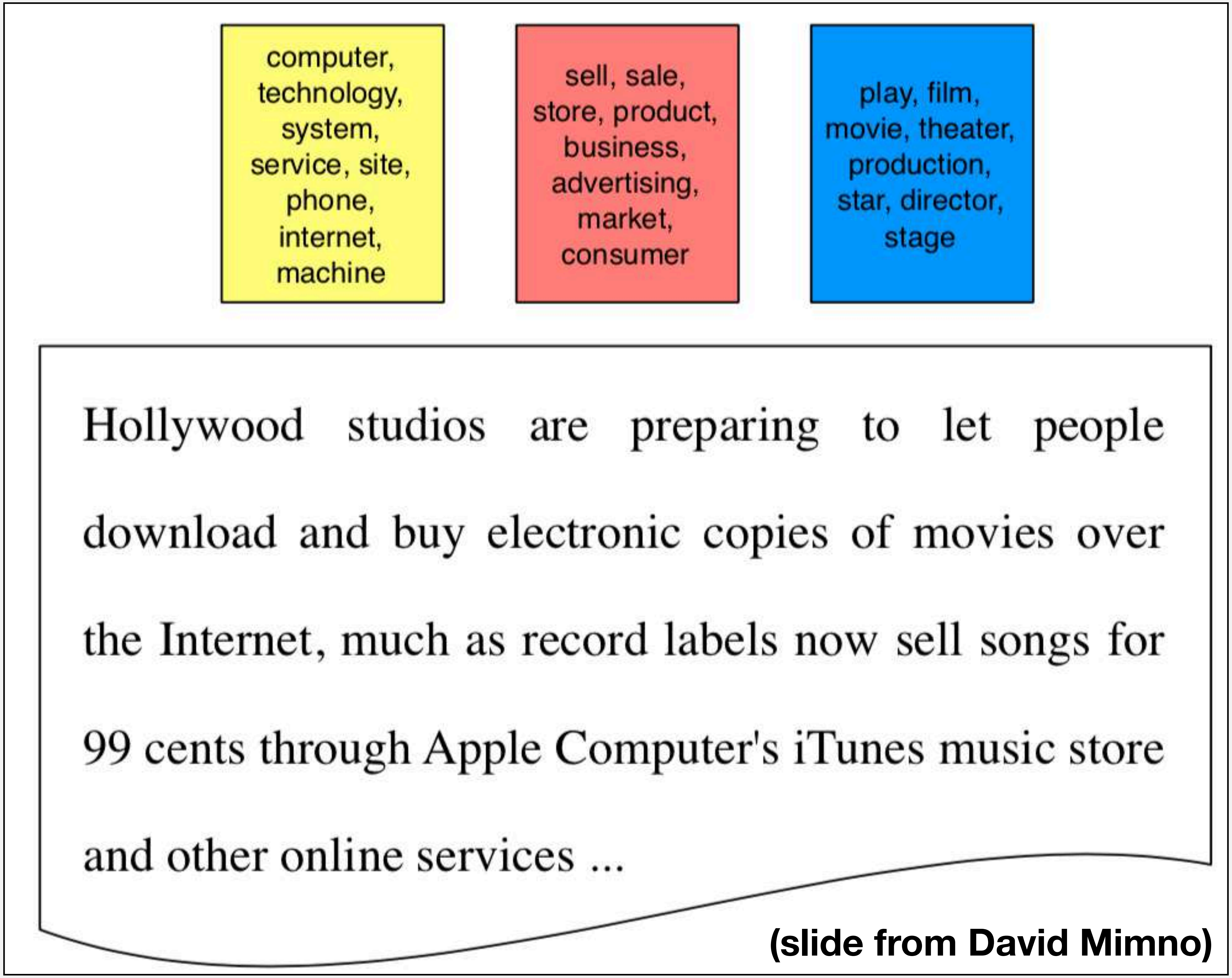
Generative Model



The three big Internet portals begin to distinguish among themselves as shopping malls
Red Light, Green Light: A 2-Tone L.E.D. to Simplify Screens
Stock Trades: A Better Deal For Investors Isn't Simple
TOPIC 1
Forget the Bootleg, Just Download the Movie Legally
TOPIC 2
The Shape of Cinema, Transformed At the Click of a Mouse
Multiplex Heralded As Linchpin To Growth
A Peaceful Crew Puts Muppets Where Its Mouth Is
TOPIC 3

(slide from David Mimno)

Generative Model



computer, technology, system, service, site, phone, internet, machine

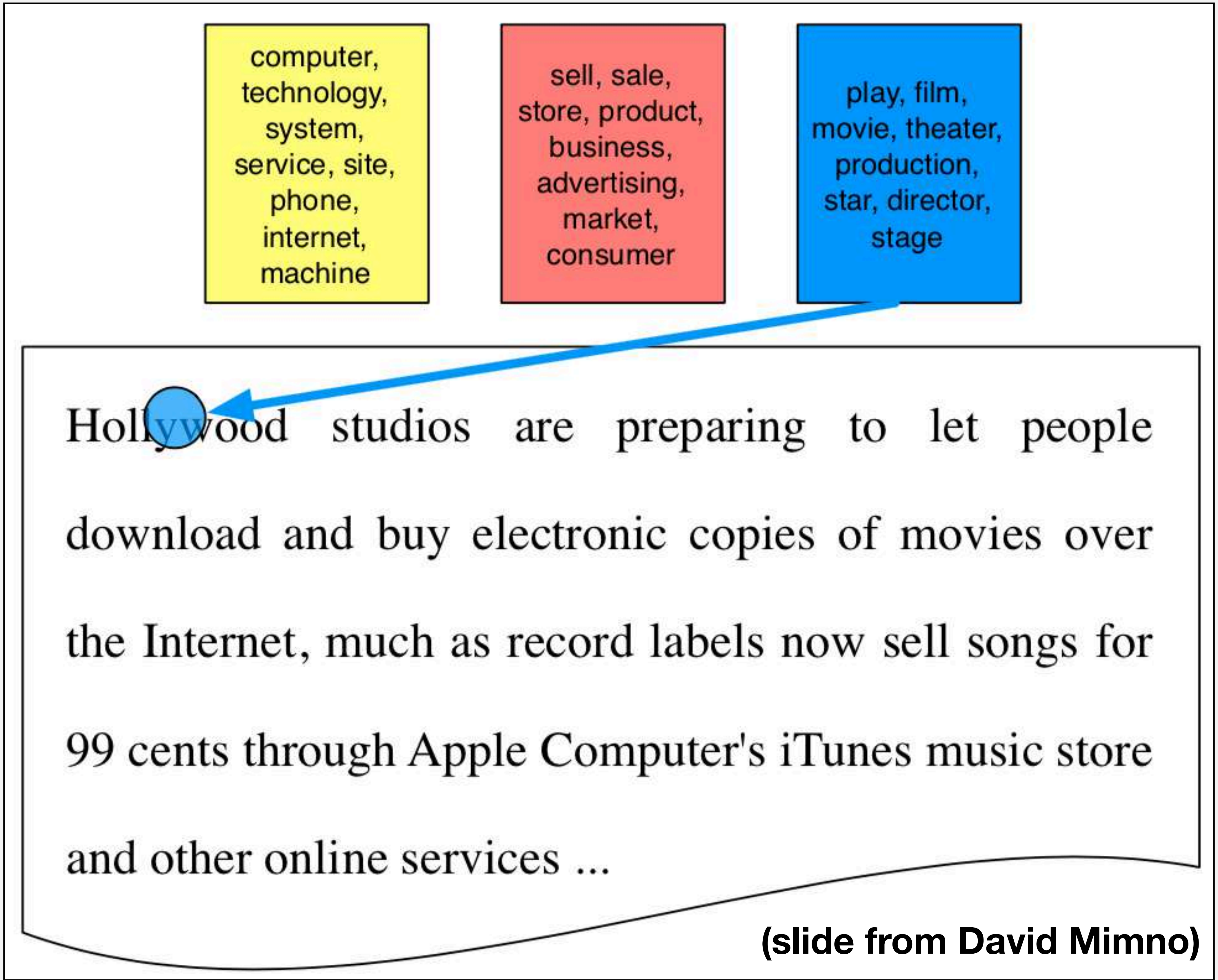
sell, sale, store, product, business, advertising, market, consumer

play, film, movie, theater, production, star, director, stage

Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

(slide from David Mimno)

Generative Model



computer, technology, system, service, site, phone, internet, machine

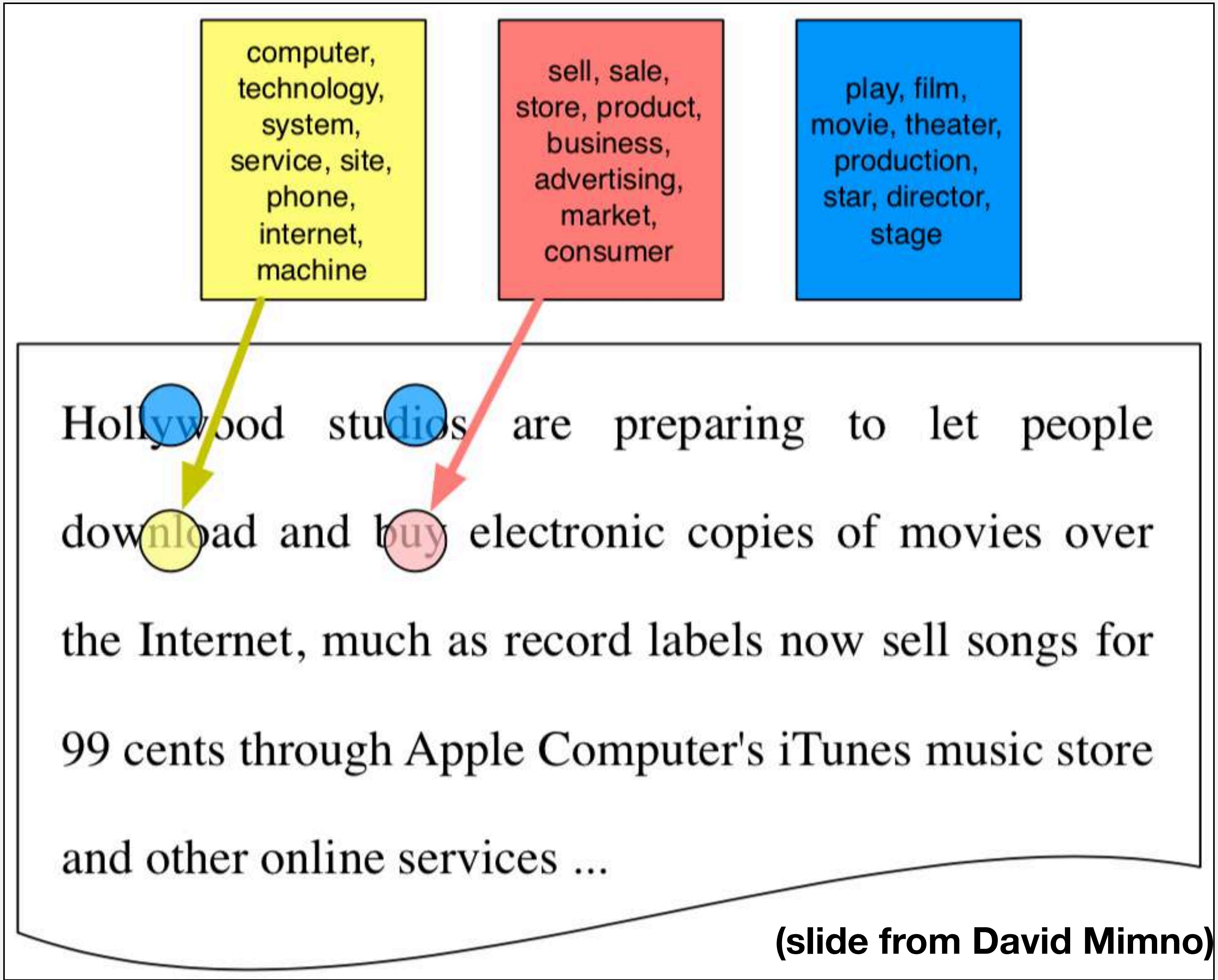
sell, sale, store, product, business, advertising, market, consumer

play, film, movie, theater, production, star, director, stage

Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

(slide from David Mimno)

Generative Model



computer, technology, system, service, site, phone, internet, machine

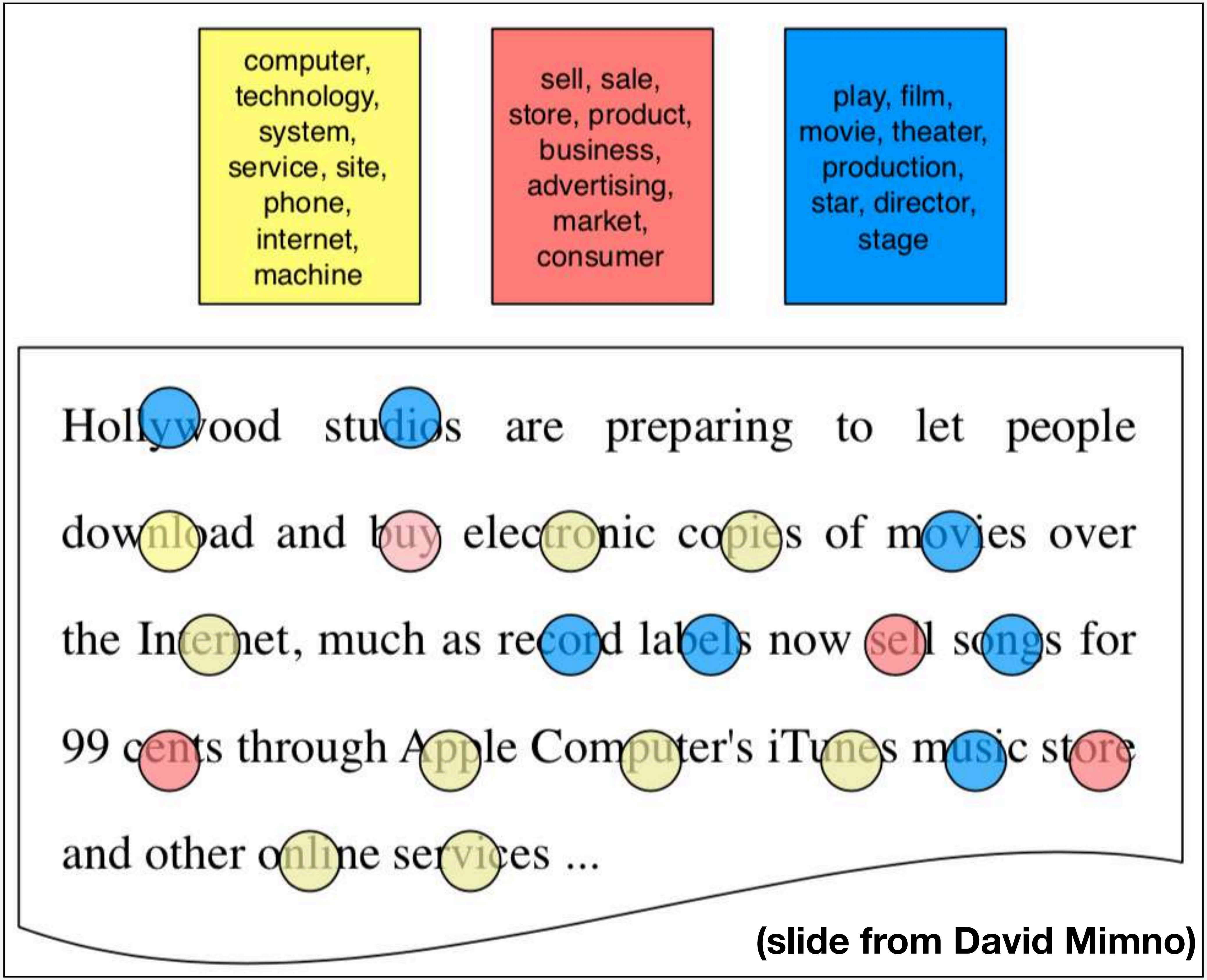
sell, sale, store, product, business, advertising, market, consumer

play, film, movie, theater, production, star, director, stage

Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

(slide from David Mimno)

Generative Model



computer, technology, system, service, site, phone, internet, machine

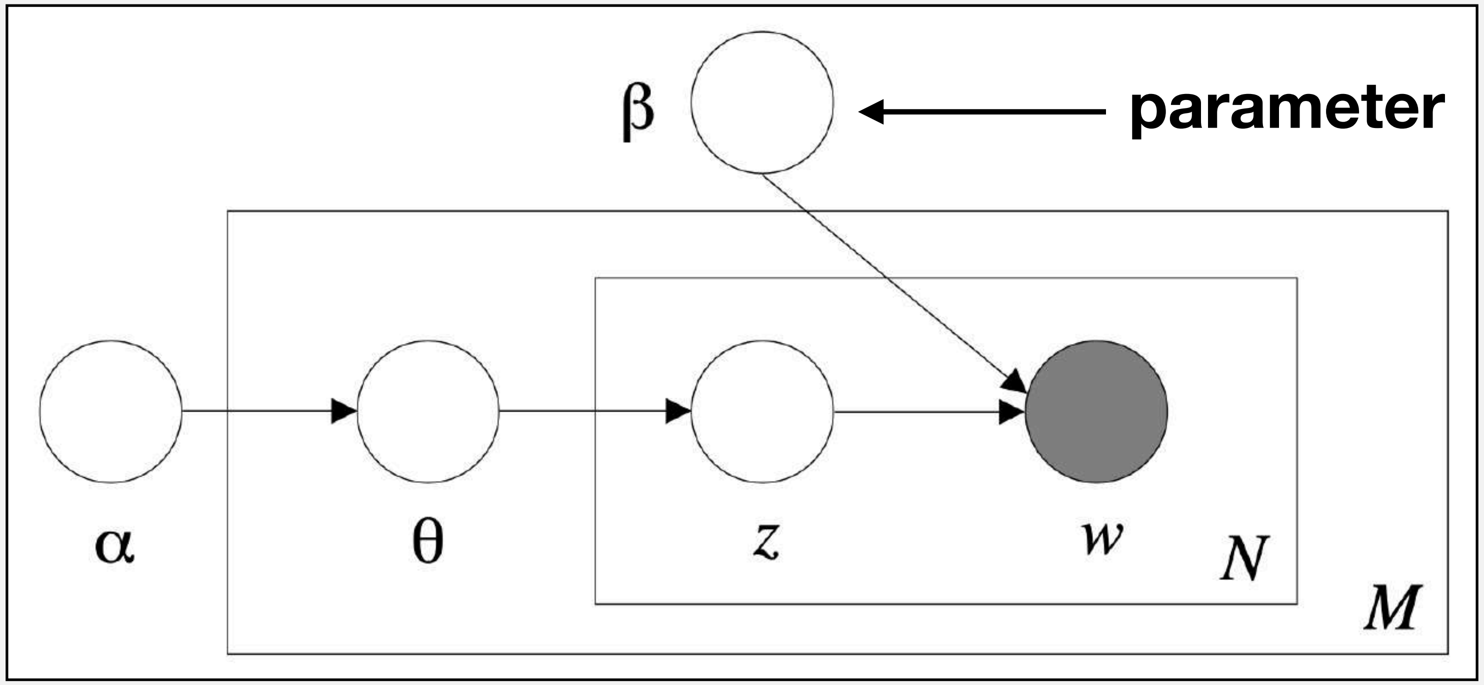
sell, sale, store, product, business, advertising, market, consumer

play, film, movie, theater, production, star, director, stage

Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

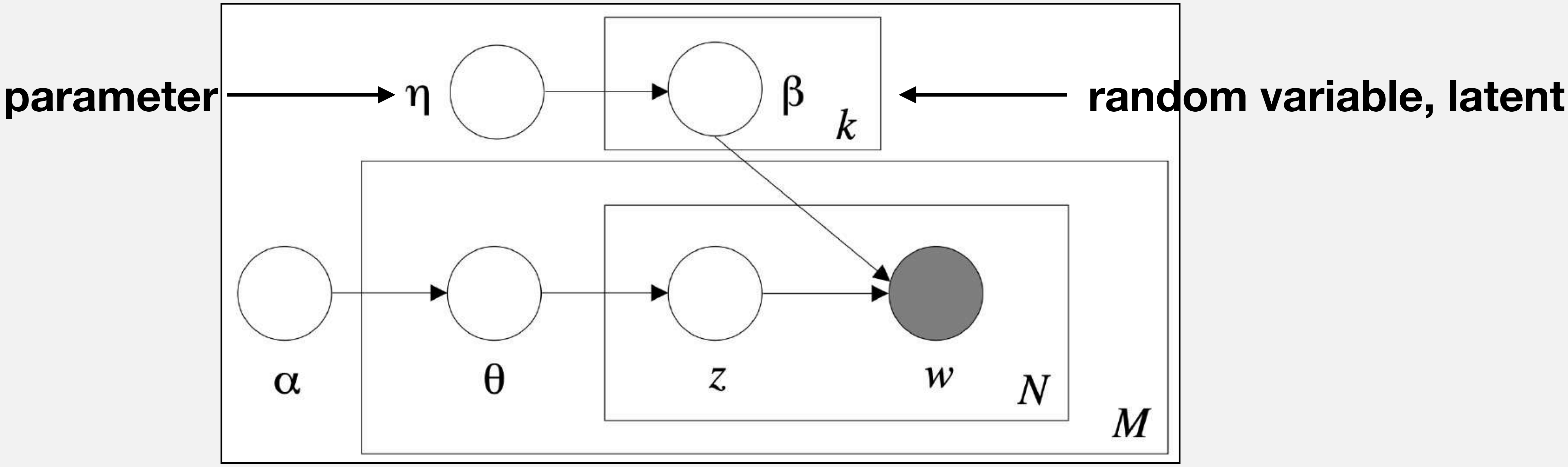
(slide from David Mimno)

Problem with Sparsity



parameter

- Documents might be similar, but have no words in common!



parameter

random variable, latent

Inference in LDA

- Need to compute the posterior distribution of hidden variables - intervals, point estimates, etc..

- However, this is intractable: need to evaluate a really complicated integral

- **Variational Inference**: we approximate the distribution we would like, with one that permits tractable optimization

$$q(\beta, \mathbf{z}, \theta \mid \lambda, \phi, \gamma) = \prod_{i=1}^{k} \underline{\textbf{Dir}(\beta_i \mid \lambda_i)} \prod_{d=1}^{M} \underline{\textbf{Dir}(\theta_d \mid \gamma_d)} \prod_{n=1}^{N} \textbf{Mult}(z_n \mid \phi_n)$$

Probability distribution for a topic (over words)

Probability distribution for a document (over topics)

Relating *q* and *p*

- We maximize the Evidence Lower BOund (ELBO):

$$\underline{\log p(\mathbf{w} \mid \alpha, \eta)} \geq \mathscr{L}(\mathbf{w}, \lambda, \phi, \gamma) = \mathbb{E}_q[\log p(\mathbf{w}, \underline{\beta, \mathbf{z}, \theta} \mid \underline{\alpha, \eta})] - \mathbb{E}_q[\log q(\underline{\beta, \mathbf{z}, \theta} \mid \underline{\lambda, \phi, \gamma})]$$

- Log-likelihood of document

- Latent variables are shared between distributions

- Differences: Dirichlet and Multinomial parameters

- A consequence of Jensen’s inequality: a concave (e.g. *log*) function of an expectation is lower-bound by the expectation of the concave function

Main Optimization

$$\begin{aligned}\mathscr{L} = \sum_d \mathbb{E}_q[\log p(\mathbf{w}_d \,|\, \beta_d, \mathbf{z}_d, \theta)] + \mathbb{E}_q[\log p(\mathbf{z}_d \,|\, \theta_d)] + \mathbb{E}_q[\log p(\theta_d \,|\, \alpha)] + \mathbb{E}_q[\log p(\beta \,|\, \eta)] \\ - \mathbb{E}_q[\log q(\mathbf{z}_d \,|\, \phi_d)] - \mathbb{E}_q[\log q(\theta_d \,|\, \gamma_d)] - \mathbb{E}_q[\log q(\beta \,|\, \lambda)]\end{aligned}$$

- Can solve for variational parameters via coordinate ascent:

$$\phi_{dwk} \propto \exp\{\mathbb{E}_q[\log \theta_{dk}] + \mathbb{E}_q[\log \beta_{kw}]\}$$ **topic-relevance**

if *k* is related to *d and* w, then the topic, document, and word are all related

$$\gamma_{dk} = \alpha + \sum_w n_{dw}\phi_{dwk}$$

a topic is relevant to a document, if document-conditioned words are topic-relevant

$$\lambda_{kw} = \eta + \sum_d n_{dw}\phi_{dwk}$$

a topic is relevant to a word, if documents that contain the word are topic-relevant

What does inference give us?

- We have now estimated our variational parameters.
  - One describes a probability distribution for each topic
  - The other describes a probability distribution for each document

Topics	**Documents**
$\textbf{Dir}(\beta_i \mid \lambda_i)$	$\textbf{Dir}(\theta_d \mid \gamma_d)$

- We can draw random samples from each distribution … or, if we just want a single realization, we take expectations:

$$\mathbb{E}[\textbf{Dir}(\beta_i \mid \lambda_i)] = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_{ij}} \qquad \mathbb{E}[\textbf{Dir}(\theta_d \mid \gamma_d)] = \frac{\gamma_d}{\sum_{j=1}^{k} \gamma_{dj}}$$

Tasks in Visualizing Topic Models

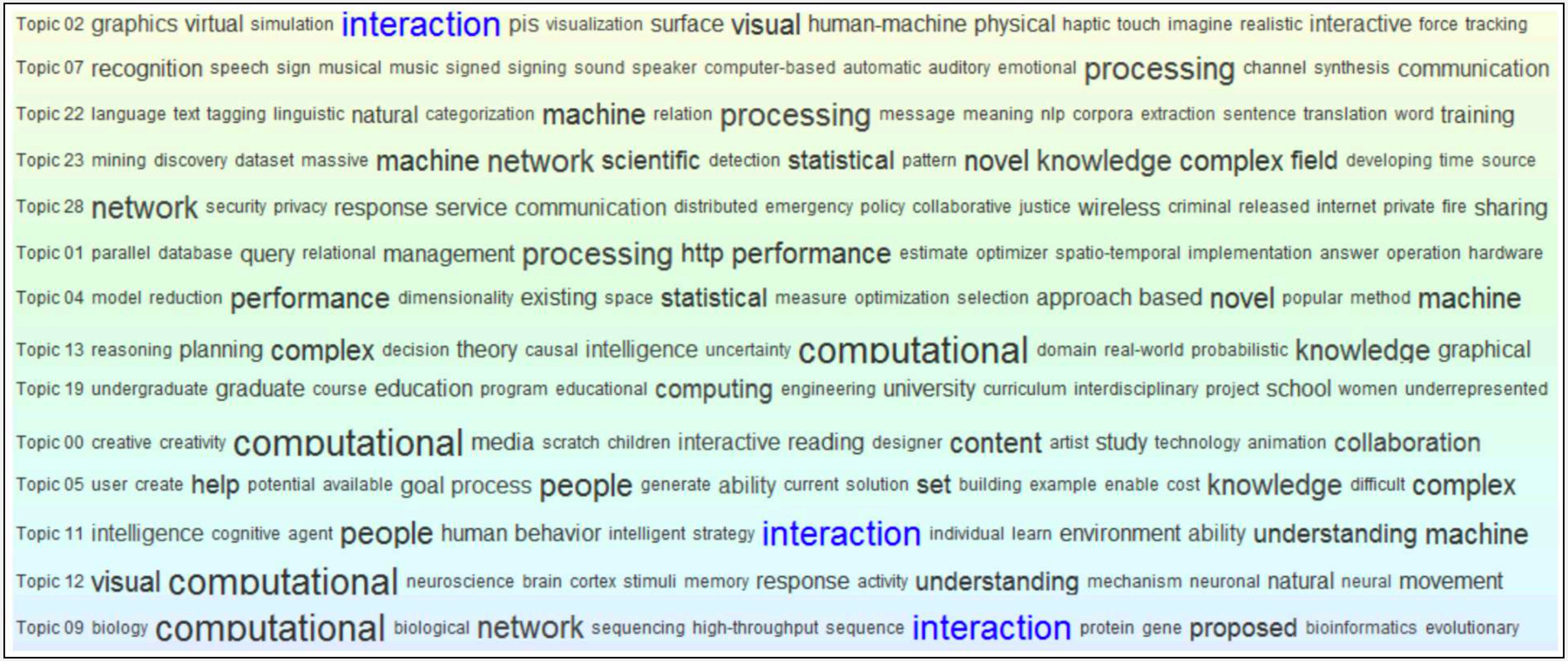
- Comparing documents
- Comparing topics
- Understanding a topic
- Understanding a document, *in terms of* topics
- Other data:
    - Time? Document Categories?

Tasks determine how we prioritize visual encodings and interactions!

Visualizing a topic?

- Easy? choose its highest-probability words.

- **Visualizing multiple topics**: some decisions need to be made…



Topic 02 graphics virtual simulation interaction pis visualization surface visual human-machine physical haptic touch imagine realistic interactive force tracking

Topic 07 recognition speech sign musical music signed signing sound speaker computer-based automatic auditory emotional processing channel synthesis communication

Topic 22 language text tagging linguistic natural categorization machine relation processing message meaning nlp corpora extraction sentence translation word training

Topic 23 mining discovery dataset massive machine network scientific detection statistical pattern novel knowledge complex field developing time source

Topic 28 network security privacy response service communication distributed emergency policy collaborative justice wireless criminal released internet private fire sharing

Topic 01 parallel database query relational management processing http performance estimate optimizer spatio-temporal implementation answer operation hardware

Topic 04 model reduction performance dimensionality existing space statistical measure optimization selection approach based novel popular method machine

Topic 13 reasoning planning complex decision theory causal intelligence uncertainty computational domain real-world probabilistic knowledge graphical

Topic 19 undergraduate graduate course education program educational computing engineering university curriculum interdisciplinary project school women underrepresented

Topic 00 creative creativity computational media scratch children interactive reading designer content artist study technology animation collaboration

Topic 05 user create help potential available goal process people generate ability current solution set building example enable cost knowledge difficult complex

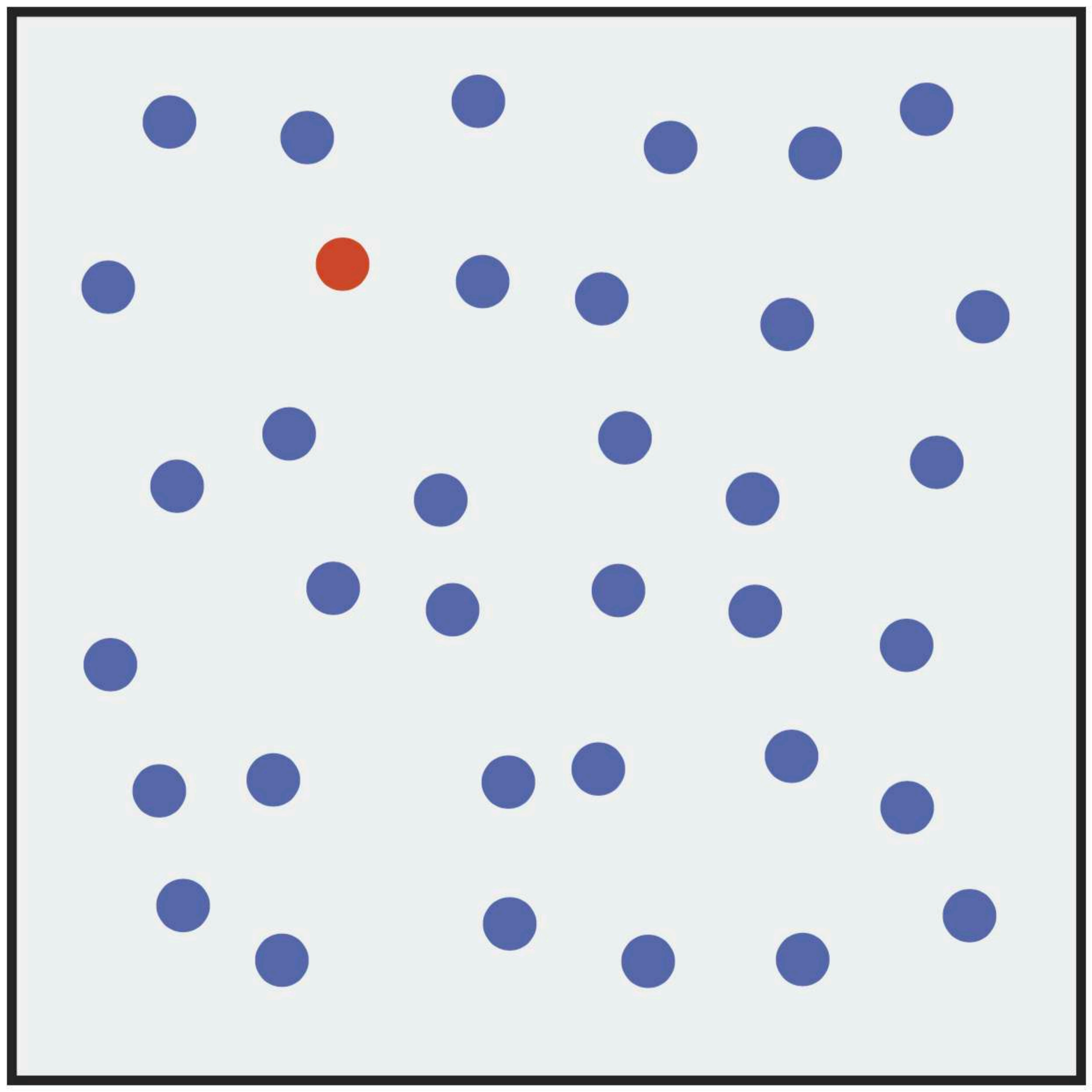
Topic 11 intelligence cognitive agent people human behavior intelligent strategy interaction individual learn environment ability understanding machine

Topic 12 visual computational neuroscience brain cortex stimuli memory response activity understanding mechanism neuronal natural neural movement

Topic 09 biology computational biological network sequencing high-throughput sequence interaction protein gene proposed bioinformatics evolutionary

[Dou et al. 2011]

Preattentive Processing



Implications for Text Visualization

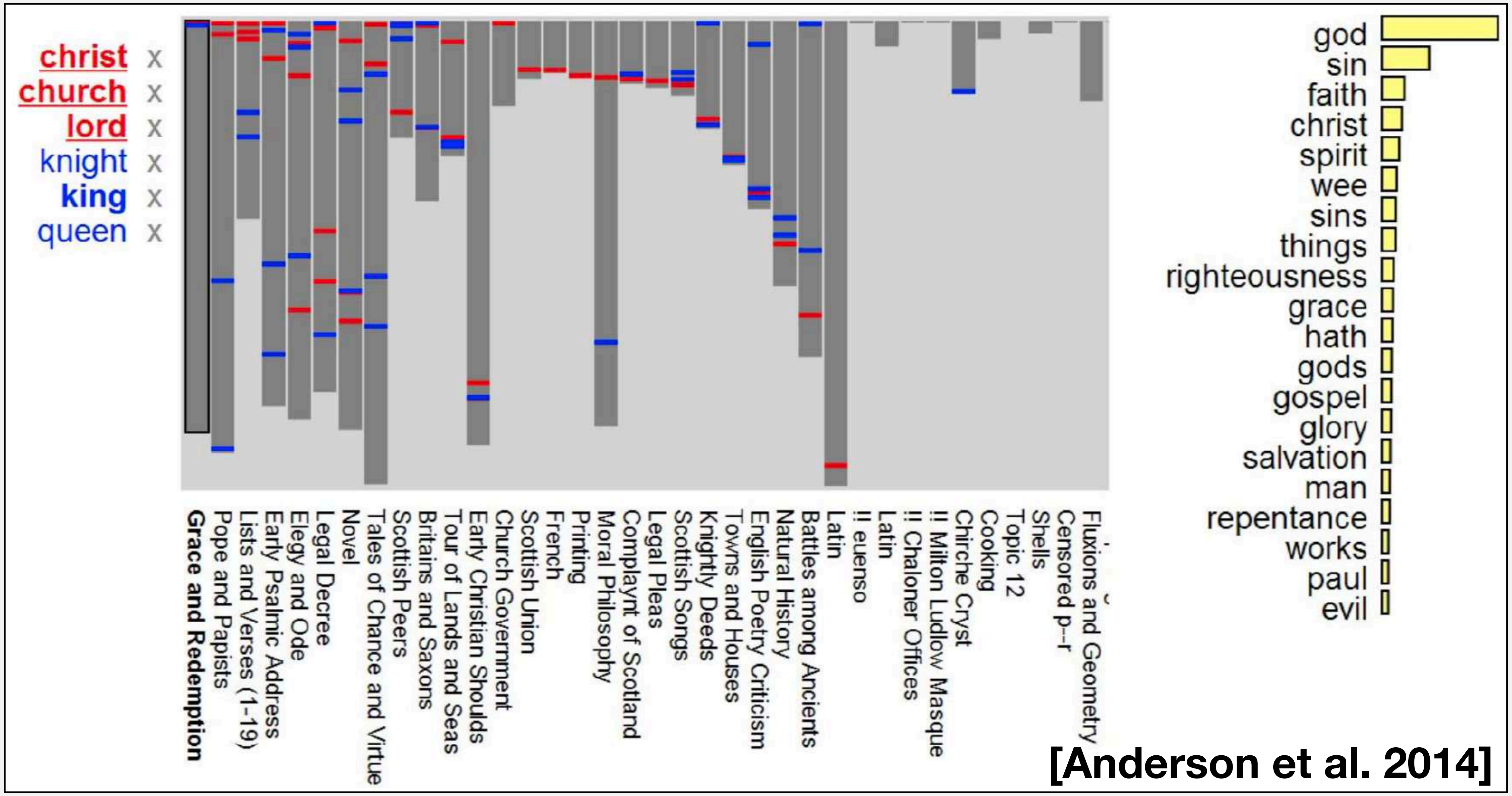
- Showing a list/scatterplot of text, without consideration of visual channels, can result in serial processing - *slow*!

- If possible, use other visual channels to style text.

- Topic modeling gives us additional information which we may use to visually style text data - **so we should use it**.

Topic Vis, Pt. 2

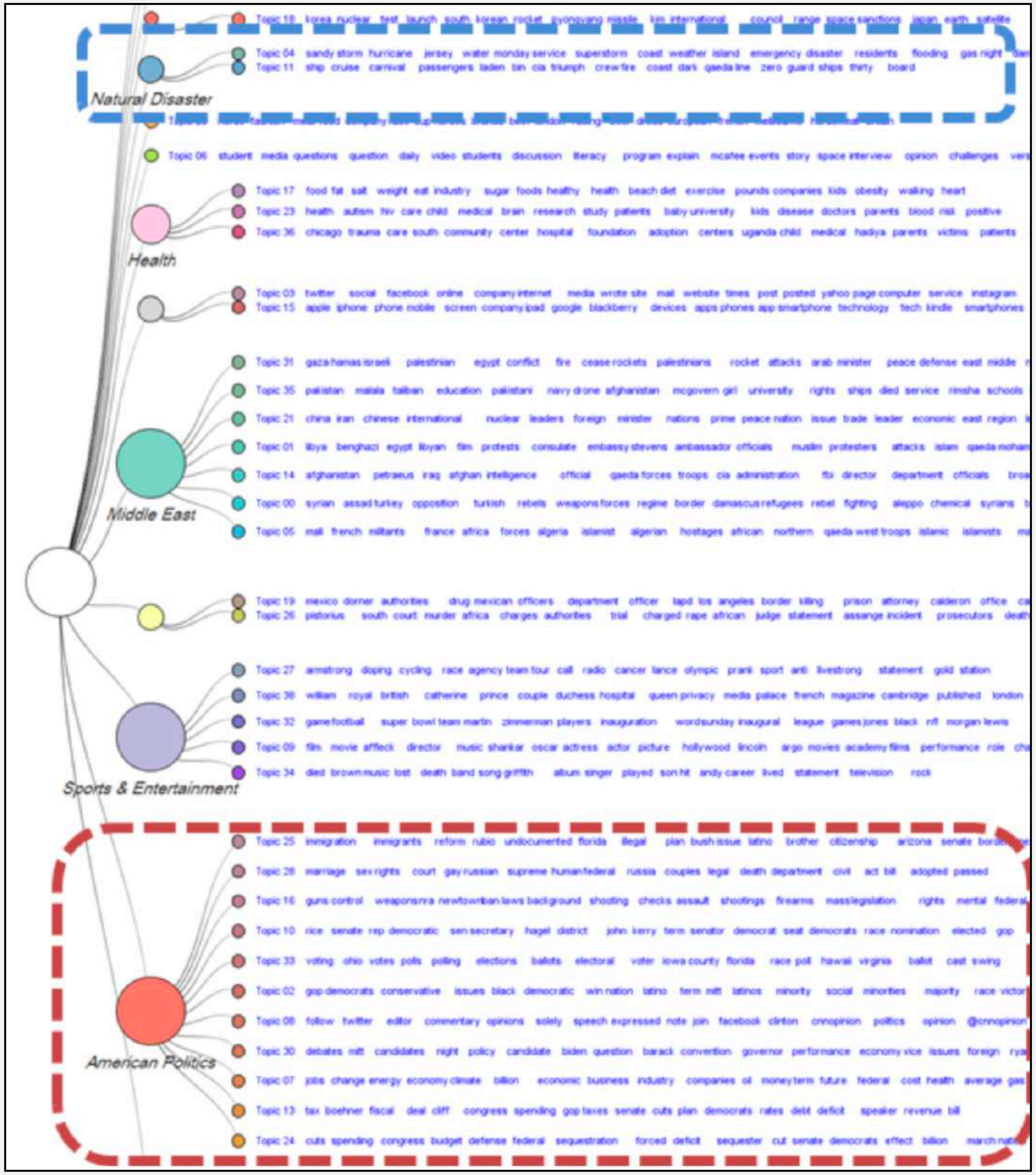
- An alternative:



[Anderson et al. 2014]

Topic Vis, Pt. 3

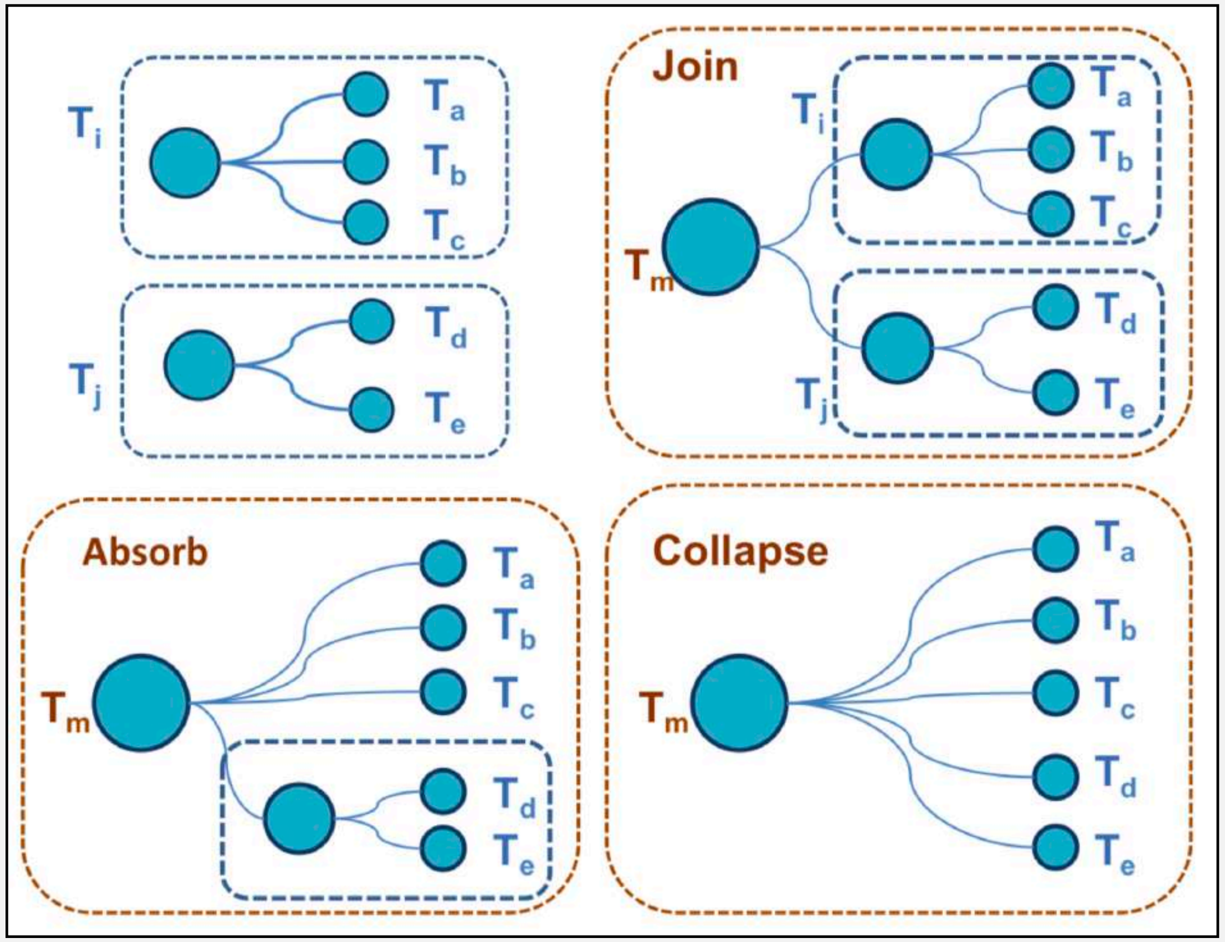
- Lots of topics?
- Hierarchies!



[Dou et al. 2013]

Building the Hierarchy

- Start from a set of topics, treat each as leaf nodes in a tree, repeat:
  - Consider the following types of operations for a pair of subtrees:



Building the Hierarchy

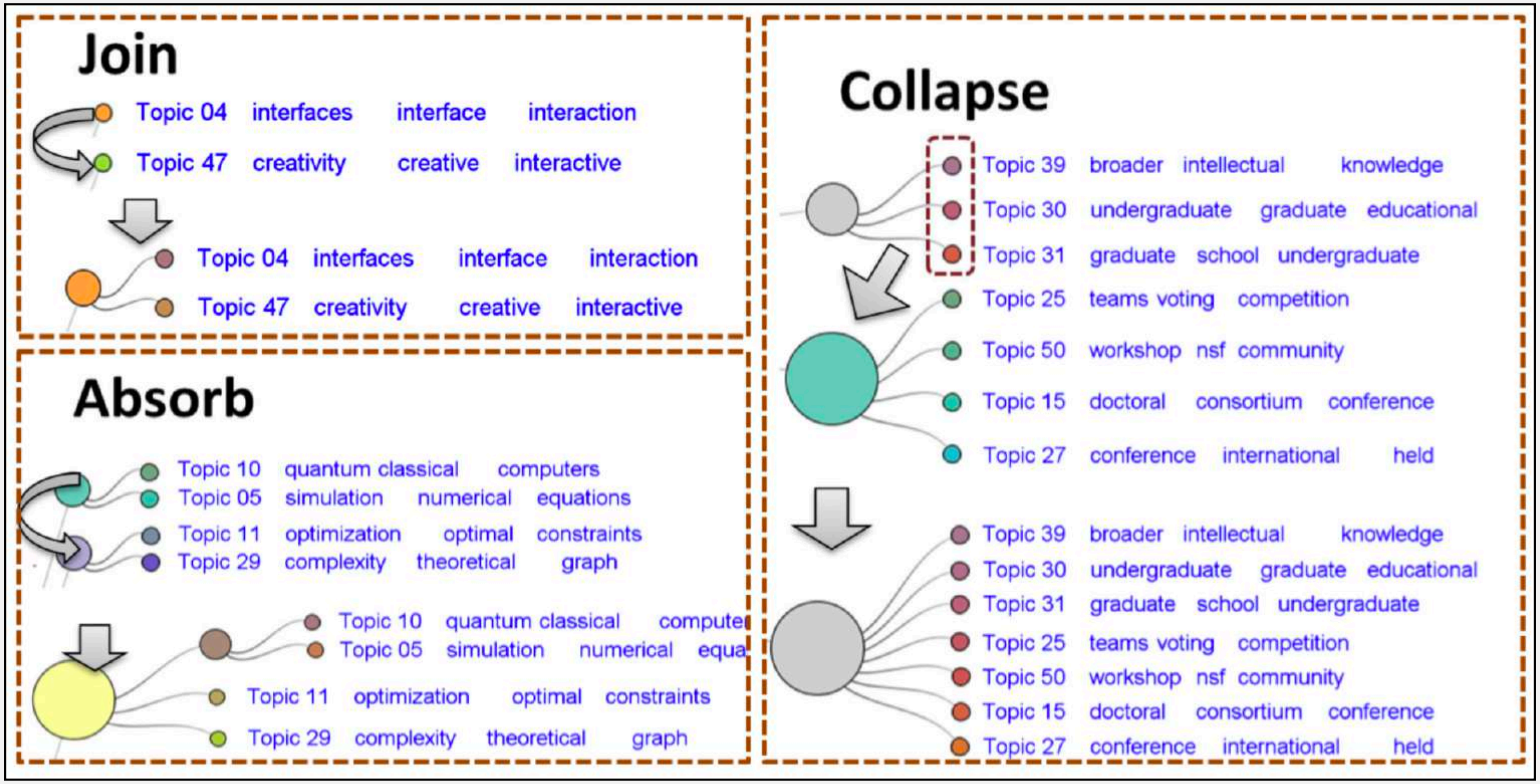
- Find the operation that gives a pair of subtrees the “lowest cost”. Cost? Distance between topics?

$$d_H(t_i, t_j) = \sum_{v=1}^{N} (\sqrt{t_{i,v}} - \sqrt{t_{j,v}})^2$$

- Non-leaf nodes? Average their distributions…
- Sidebar: any potential issues with this distance?
- Algorithm continued: merge the two nodes with lowest cost, repeat until we reach the root!

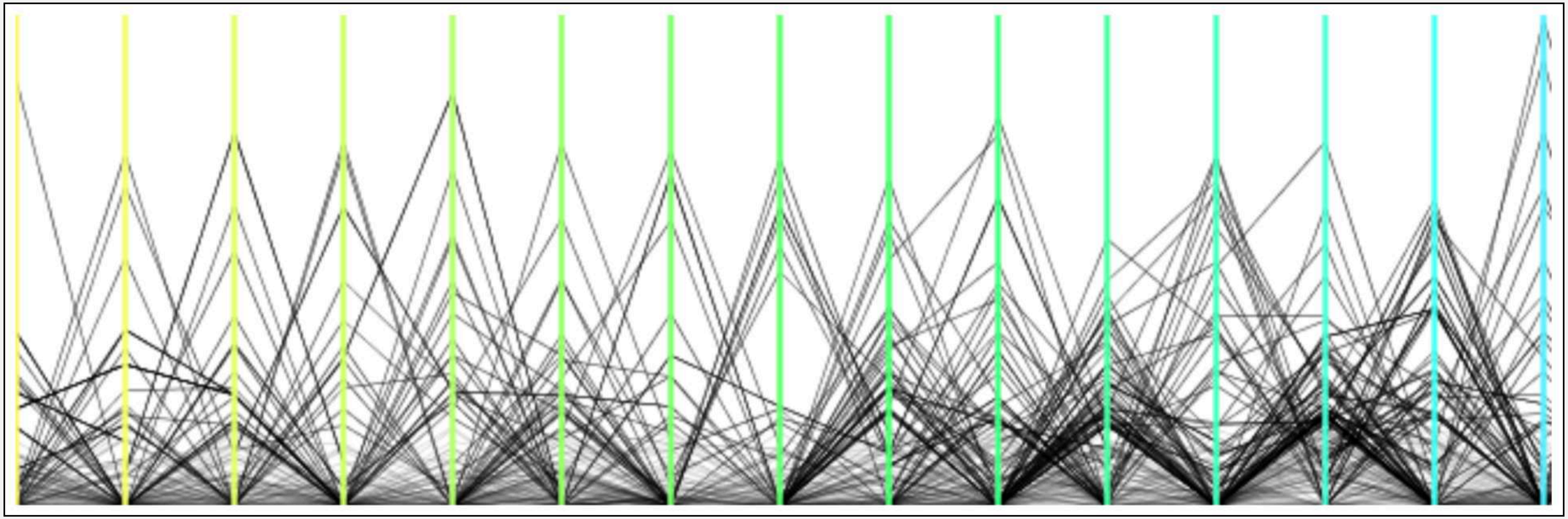
Hierarchy: Good?

- Might be imperfect, so allow the user to adjust by exposing these operations:


Join
Topic 04 interfaces interface interaction
Topic 47 creativity creative interactive
Topic 04 interfaces interface interaction
Topic 47 creativity creative interactive
Absorb
Topic 10 quantum classical computers
Topic 05 simulation numerical equations
Topic 11 optimization optimal constraints
Topic 29 complexity theoretical graph
Topic 10 quantum classical compute
Topic 05 simulation numerical equa
Topic 11 optimization optimal constraints
Topic 29 complexity theoretical graph
Collapse
Topic 39 broader intellectual knowledge
Topic 30 undergraduate graduate educational
Topic 31 graduate school undergraduate
Topic 25 teams voting competition
Topic 50 workshop nsf community
Topic 15 doctoral consortium conference
Topic 27 conference international held
Topic 39 broader intellectual knowledge
Topic 30 undergraduate graduate educational
Topic 31 graduate school undergraduate
Topic 25 teams voting competition
Topic 50 workshop nsf community
Topic 15 doctoral consortium conference
Topic 27 conference international held

Visualizing a document?

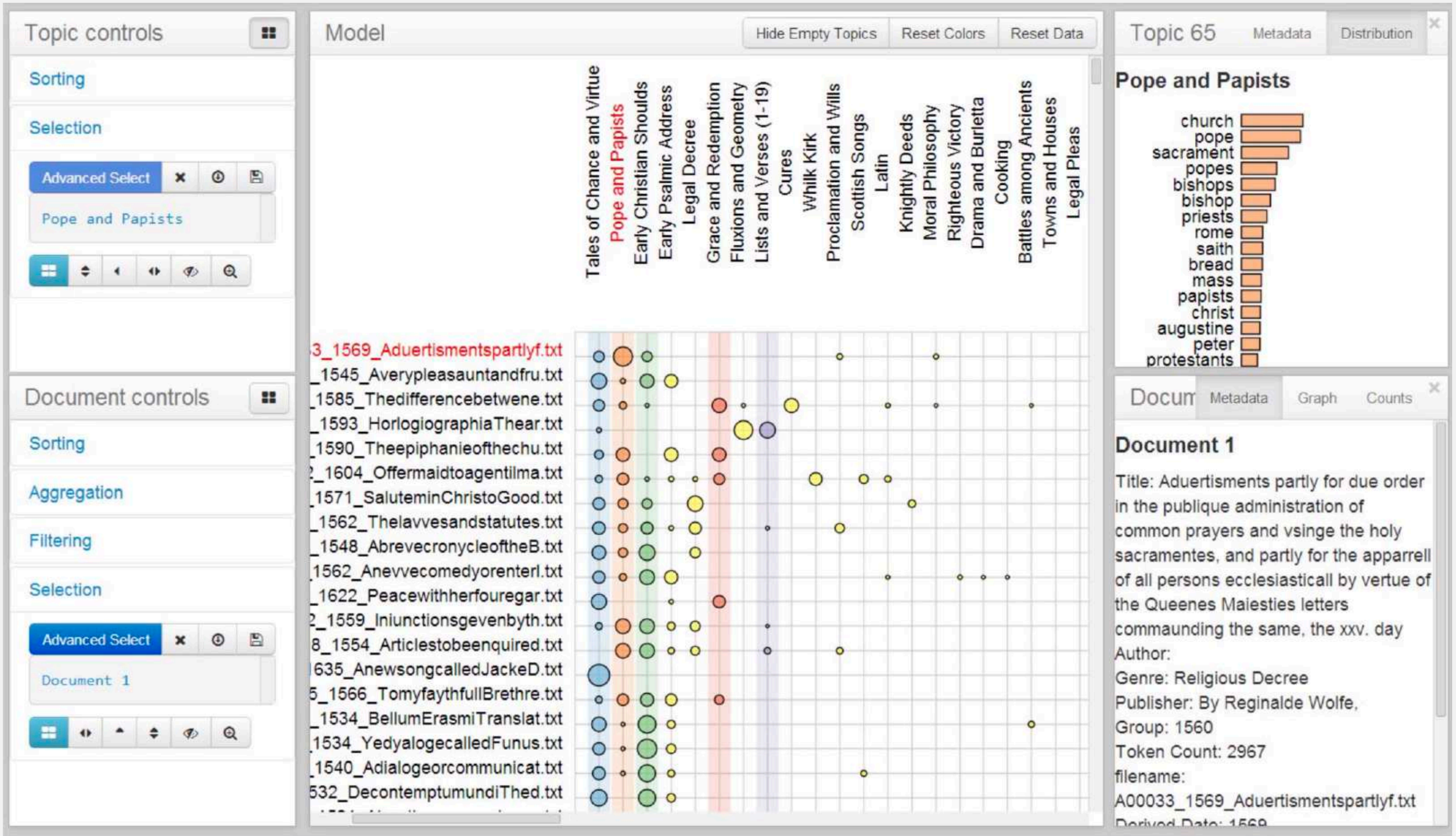
- Easy? Show topic assignments.
- Visualizing multiple topics? Parallel coordinates!



[Dou et al. 2011]

Document Vis, Pt. 2

- Poly-lines can quickly become a source of clutter!
- An alternative:



Document 1

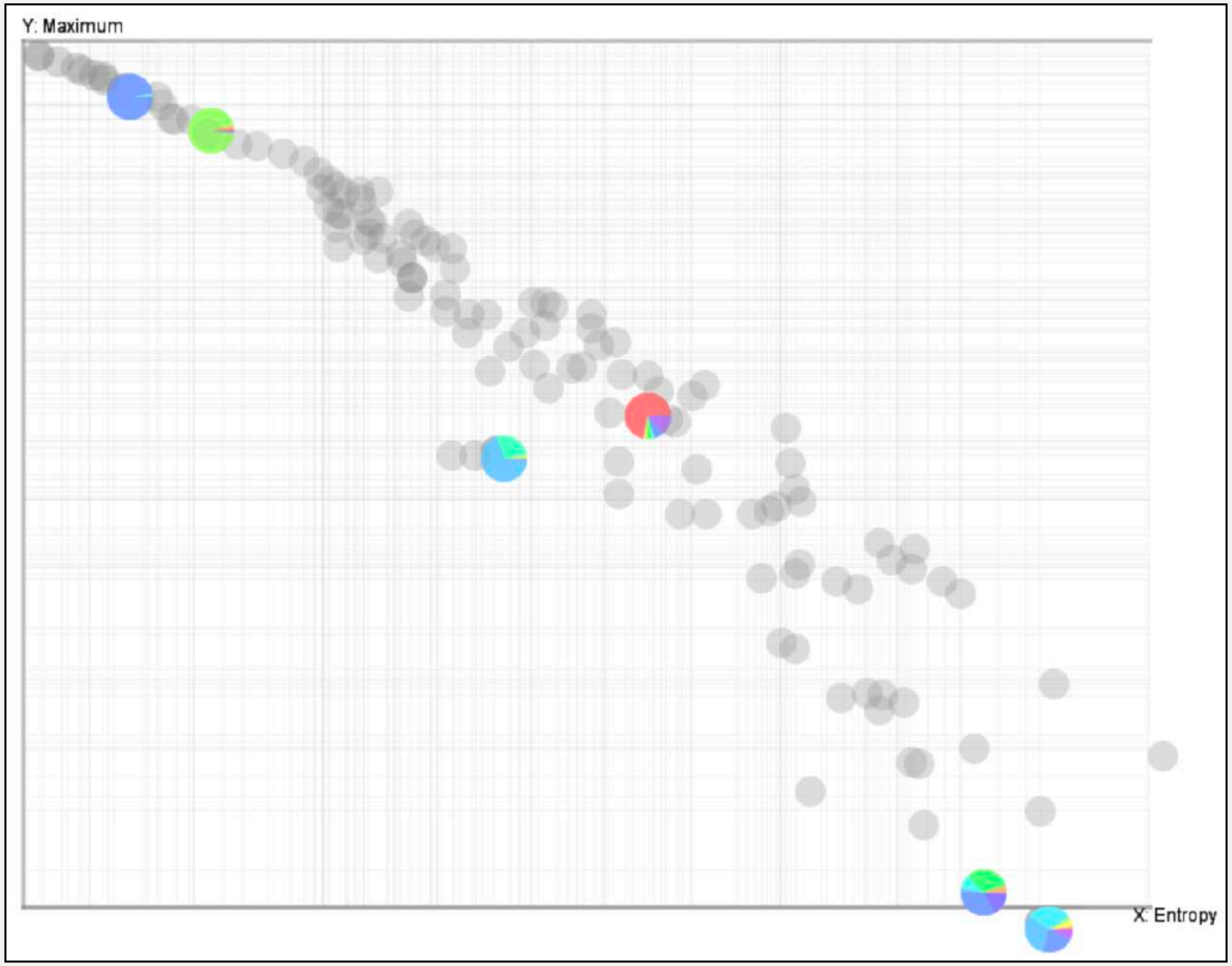
Title: Aduertisments partly for due order in the publique administration of common prayers and vsinge the holy sacramentes, and partly for the apparrell of all persons ecclesiasticall by vertue of the Queenes Maiesties letters commaunding the same, the xxv. day
Author:
Genre: Religious Decree
Publisher: By Reginalde Wolfe,
Group: 1560
Token Count: 2967
filename:
A00033_1569_Aduertismentspartlyf.txt
Derived Date: 1569

[Anderson et al. 2014]

Document Vis, Pt. 3

- Order matters!

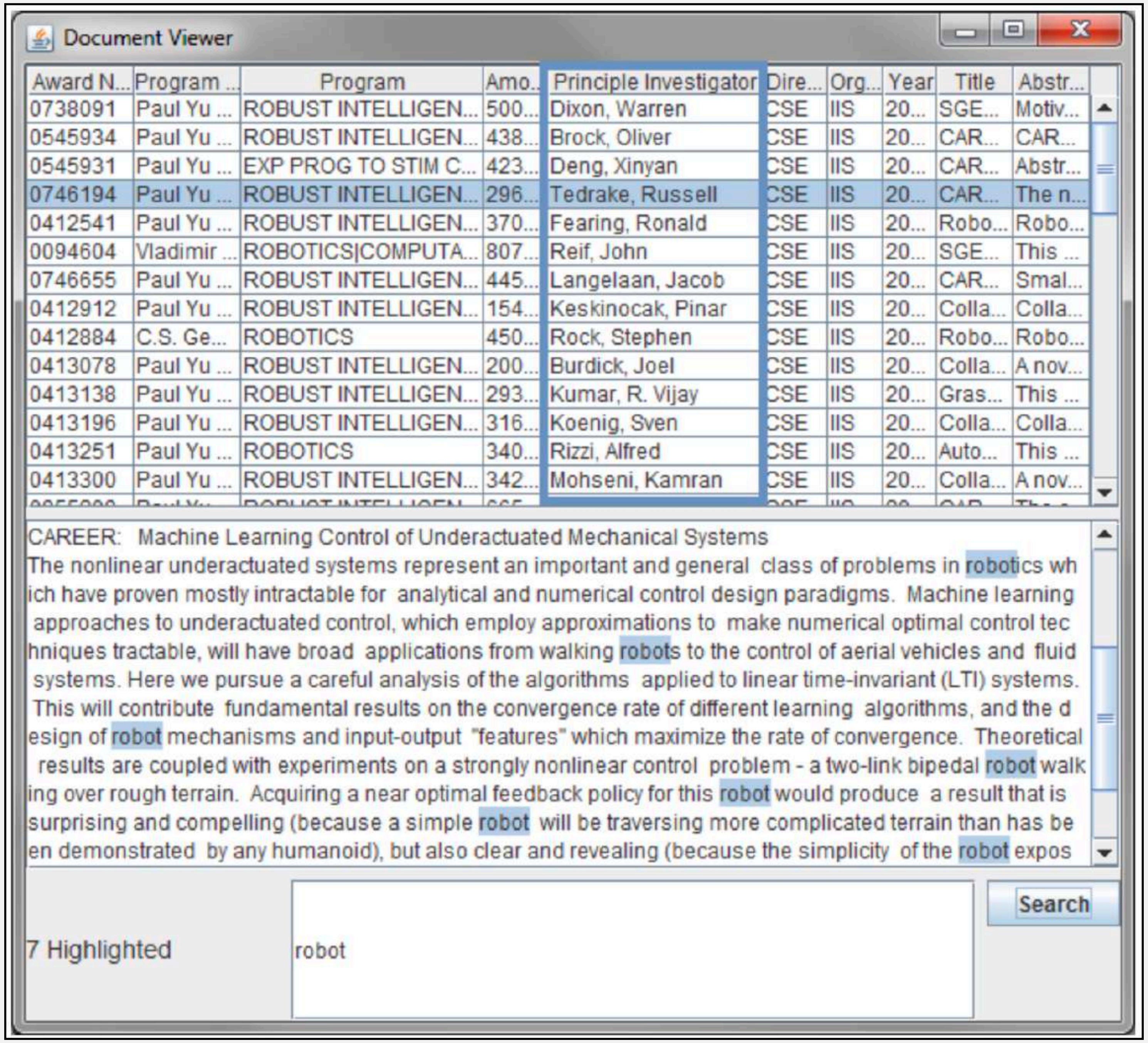
- An alternative:



[Dou et al. 2011]

Text Vis?

- Show the document directly:



Document Viewer

Award N...	Program ...	Program	Amo...	Principle Investigator	Dire...	Org...	Year	Title	Abstr...
0738091	Paul Yu ...	ROBUST INTELLIGEN...	500...	Dixon, Warren	CSE	IIS	20...	SGE...	Motiv...
0545934	Paul Yu ...	ROBUST INTELLIGEN...	438...	Brock, Oliver	CSE	IIS	20...	CAR...	CAR...
0545931	Paul Yu ...	EXP PROG TO STIM C...	423...	Deng, Xinyan	CSE	IIS	20...	CAR...	Abstr...
0746194	Paul Yu ...	ROBUST INTELLIGEN...	296...	Tedrake, Russell	CSE	IIS	20...	CAR...	The n...
0412541	Paul Yu ...	ROBUST INTELLIGEN...	370...	Fearing, Ronald	CSE	IIS	20...	Robo...	Robo...
0094604	Vladimir ...	ROBOTICS|COMPUTA...	807...	Reif, John	CSE	IIS	20...	SGE...	This ...
0746655	Paul Yu ...	ROBUST INTELLIGEN...	445...	Langelaan, Jacob	CSE	IIS	20...	CAR...	Smal...
0412912	Paul Yu ...	ROBUST INTELLIGEN...	154...	Keskinocak, Pinar	CSE	IIS	20...	Colla...	Colla...
0412884	C.S. Ge...	ROBOTICS	450...	Rock, Stephen	CSE	IIS	20...	Robo...	Robo...
0413078	Paul Yu ...	ROBUST INTELLIGEN...	200...	Burdick, Joel	CSE	IIS	20...	Colla...	A nov...
0413138	Paul Yu ...	ROBUST INTELLIGEN...	293...	Kumar, R. Vijay	CSE	IIS	20...	Gras...	This ...
0413196	Paul Yu ...	ROBUST INTELLIGEN...	316...	Koenig, Sven	CSE	IIS	20...	Colla...	Colla...
0413251	Paul Yu ...	ROBOTICS	340...	Rizzi, Alfred	CSE	IIS	20...	Auto...	This ...
0413300	Paul Yu ...	ROBUST INTELLIGEN...	342...	Mohseni, Kamran	CSE	IIS	20...	Colla...	A nov...

CAREER: Machine Learning Control of Underactuated Mechanical Systems
The nonlinear underactuated systems represent an important and general class of problems in robotics wh ich have proven mostly intractable for analytical and numerical control design paradigms. Machine learning approaches to underactuated control, which employ approximations to make numerical optimal control tec hniques tractable, will have broad applications from walking robots to the control of aerial vehicles and fluid systems. Here we pursue a careful analysis of the algorithms applied to linear time-invariant (LTI) systems. This will contribute fundamental results on the convergence rate of different learning algorithms, and the d esign of robot mechanisms and input-output "features" which maximize the rate of convergence. Theoretical results are coupled with experiments on a strongly nonlinear control problem - a two-link bipedal robot walk ing over rough terrain. Acquiring a near optimal feedback policy for this robot would produce a result that is surprising and compelling (because a simple robot will be traversing more complicated terrain than has be en demonstrated by any humanoid), but also clear and revealing (because the simplicity of the robot expos

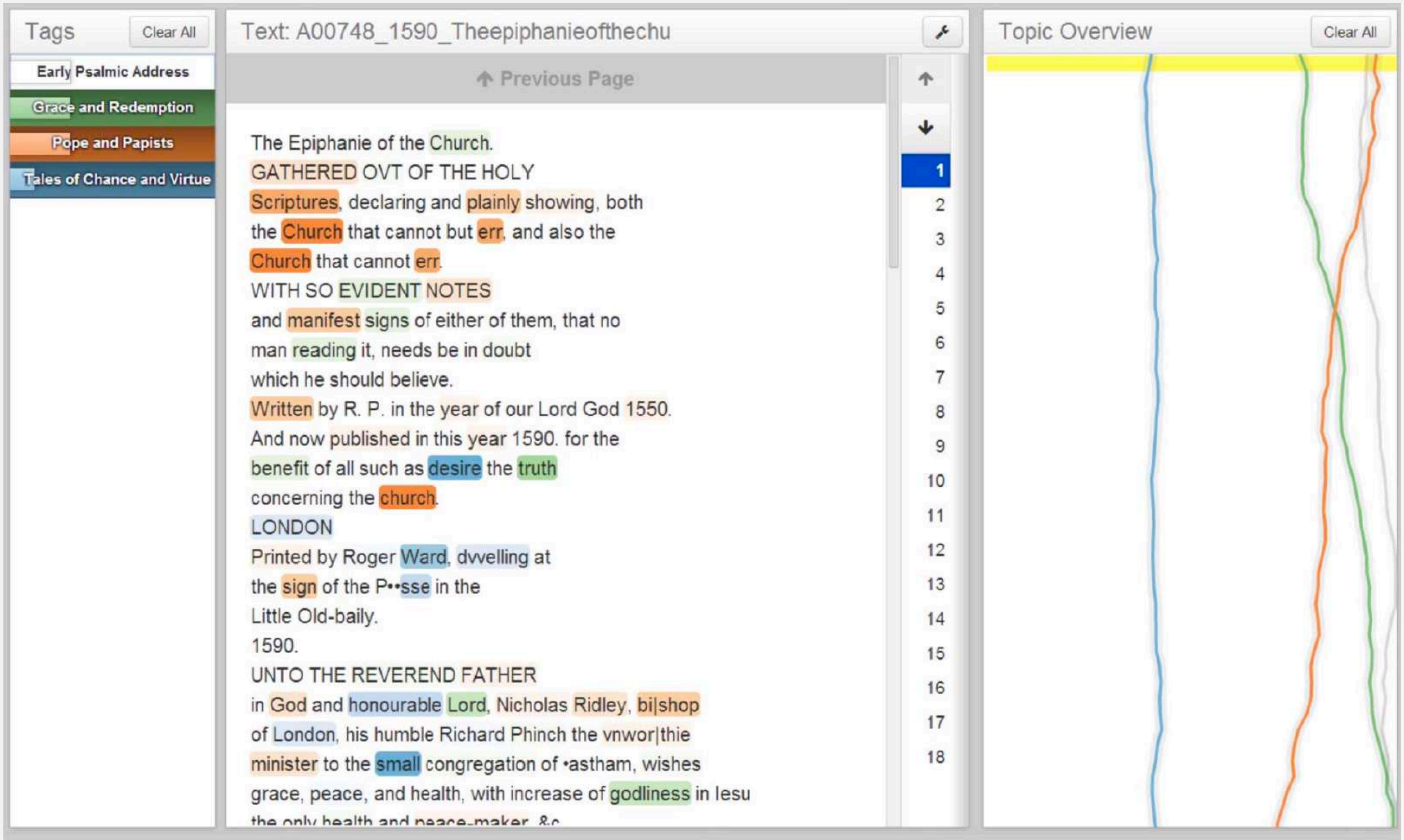
7 Highlighted

robot

[Dou et al. 2011]

Text Vis, Pt. 2

- Use topic model to aid in showing text!

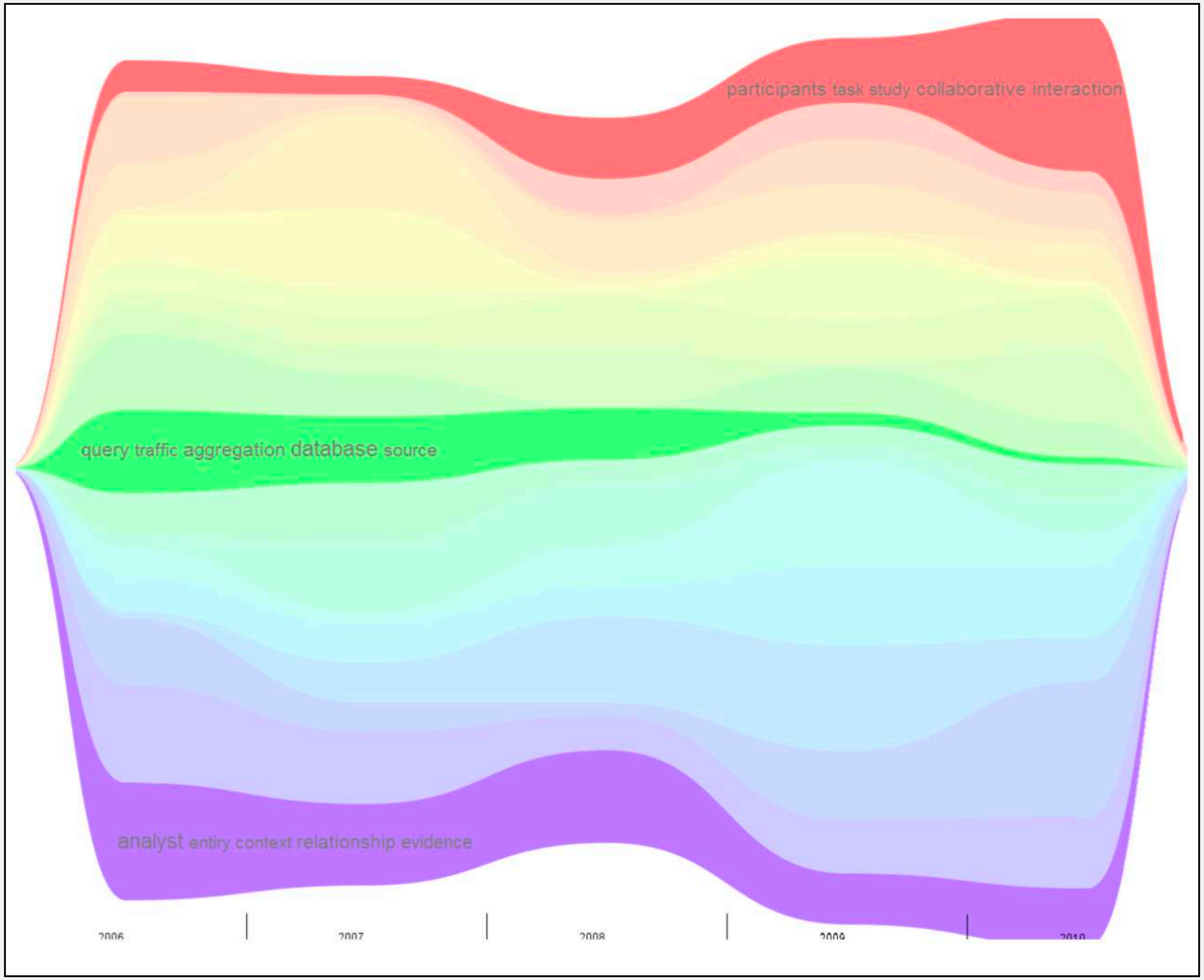


The Epiphanie of the Church.
GATHERED OVT OF THE HOLY
Scriptures, declaring and plainly showing, both
the Church that cannot but err, and also the
Church that cannot err.
WITH SO EVIDENT NOTES
and manifest signs of either of them, that no
man reading it, needs be in doubt
which he should believe.
Written by R. P. in the year of our Lord God 1550.
And now published in this year 1590. for the
benefit of all such as desire the truth
concerning the church.
LONDON
Printed by Roger Ward, dvvelling at
the sign of the P••sse in the
Little Old-baily.
1590.
UNTO THE REVEREND FATHER
in God and honourable Lord, Nicholas Ridley, bi|shop
of London, his humble Richard Phinch the vnwor|thie
minister to the small congregation of •astham, wishes
grace, peace, and health, with increase of godliness in Iesu
the only health and peace-maker, &c.

[Anderson et al. 2014]

Handling Time

- How do topics vary over time?



[Dou et al. 2011]

What is this time series?

- Need to derive a time-dependent measure of **topic relevance**.

- Method:

  - Fix a time interval unit

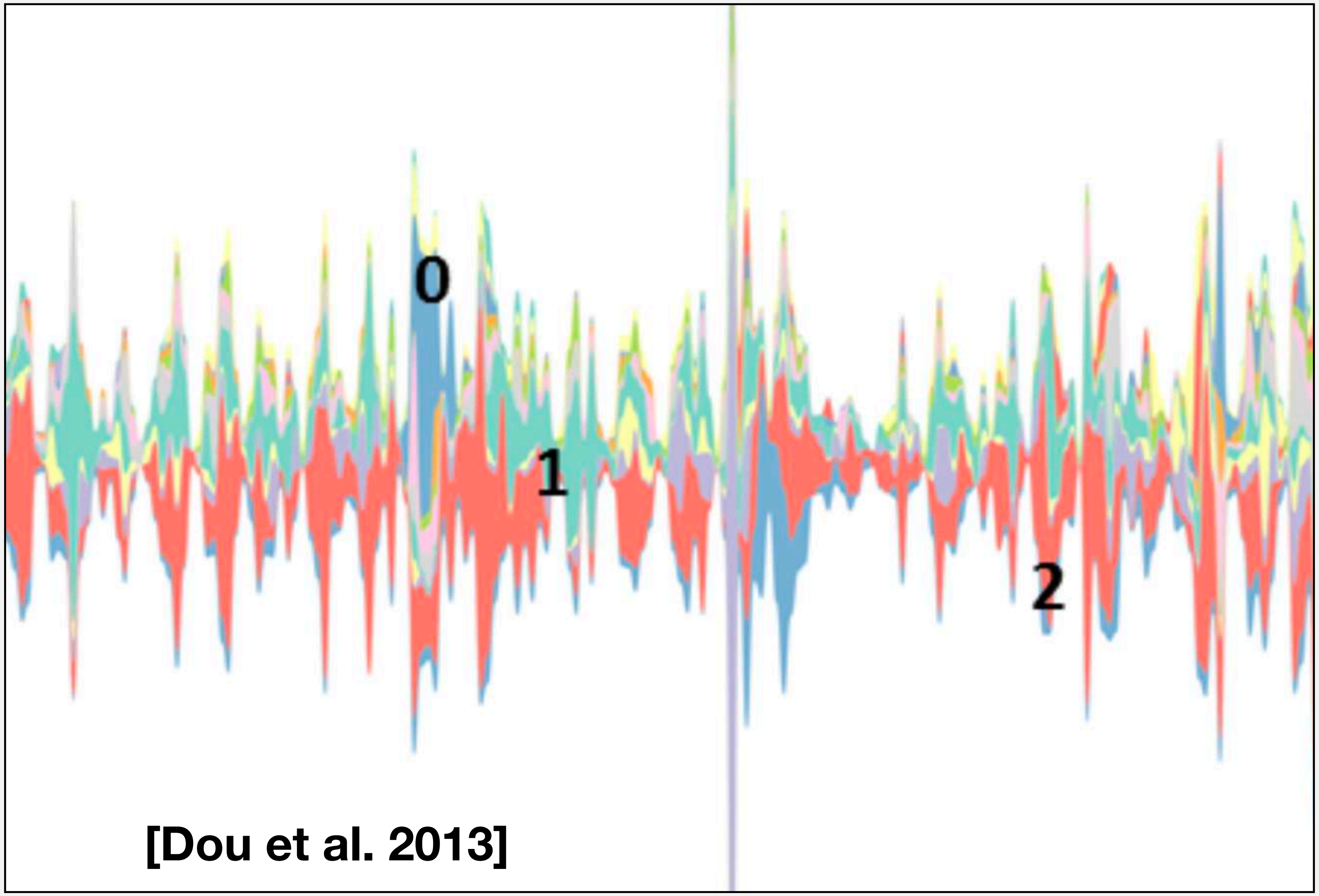
  - Take a temporal sliding window over all time steps, where the window length is this interval.

  - For a given temporal window, gather all documents:

$$\tau_D(i) = \sum_{d \in D} d_i$$

Hierarchical Temporal Evolution

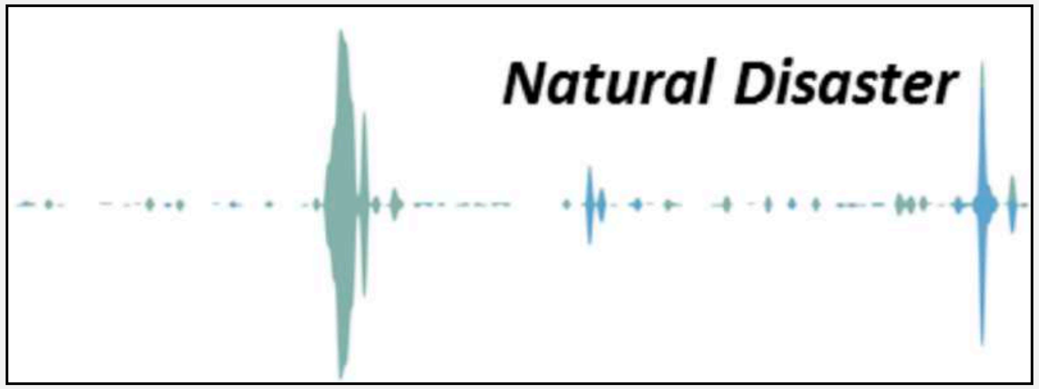
- Top-level nodes:

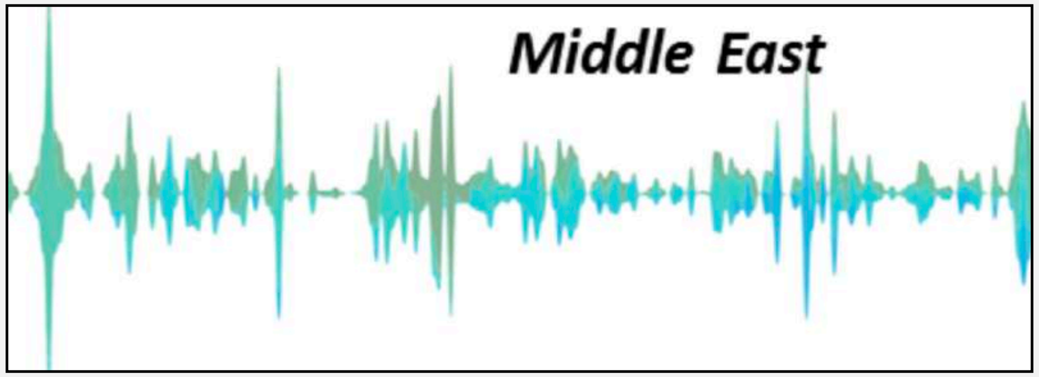


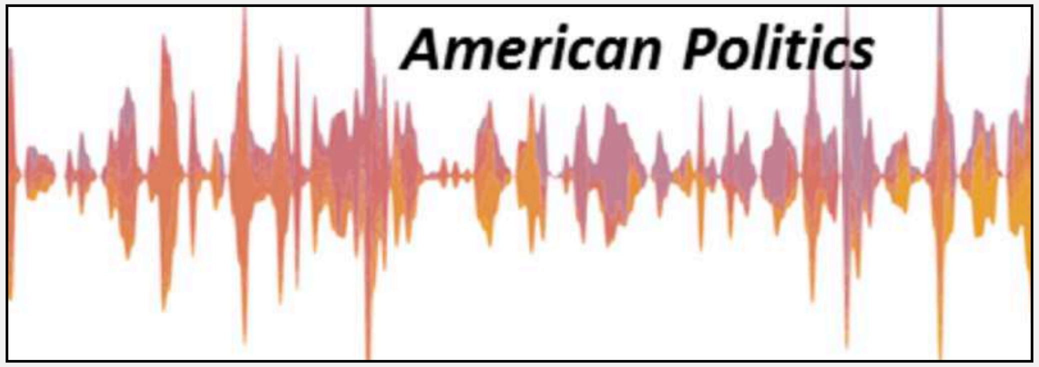
[Dou et al. 2013]

Hierarchical Temporal Evolution

- Interactive! Click on a node, expand to its children:







Topic Modeling for Exploring Conversations

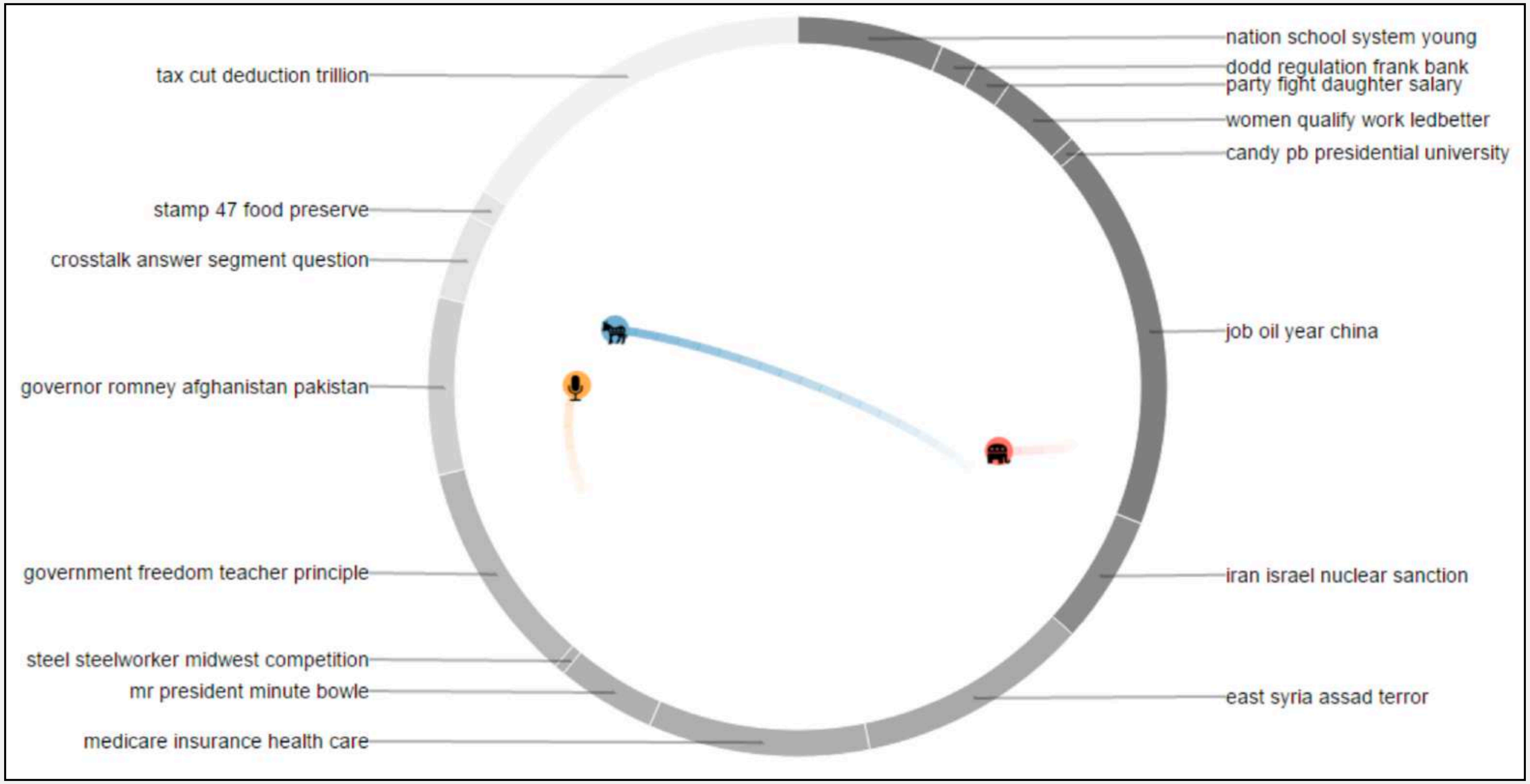
- Conversation data
  - Set of speakers
  - Each speaker delivers an utterance: a set of statements, at a particular time.
  - Treat each utterance as a document: topic modeling!

ConToVi: Multi-Party Conversation Exploration

- Objective: show the progression of a conversation, in the context of topic membership

- Necessitates different views:

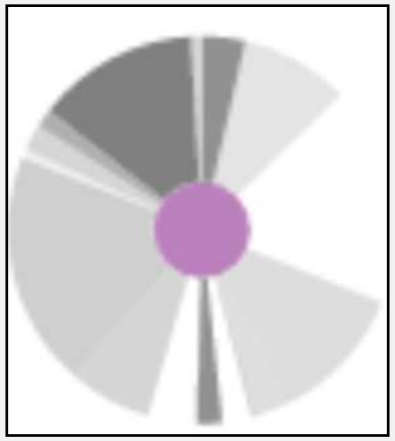
[El-Assady et al. 2016]



nation school system young
dodd regulation frank bank
party fight daughter salary
women qualify work ledbetter
candy pb presidential university
job oil year china
iran israel nuclear sanction
east syria assad terror
tax cut deduction trillion
stamp 47 food preserve
crosstalk answer segment question
governor romney afghanistan pakistan
government freedom teacher principle
steel steelworker midwest competition
mr president minute bowle
medicare insurance health care

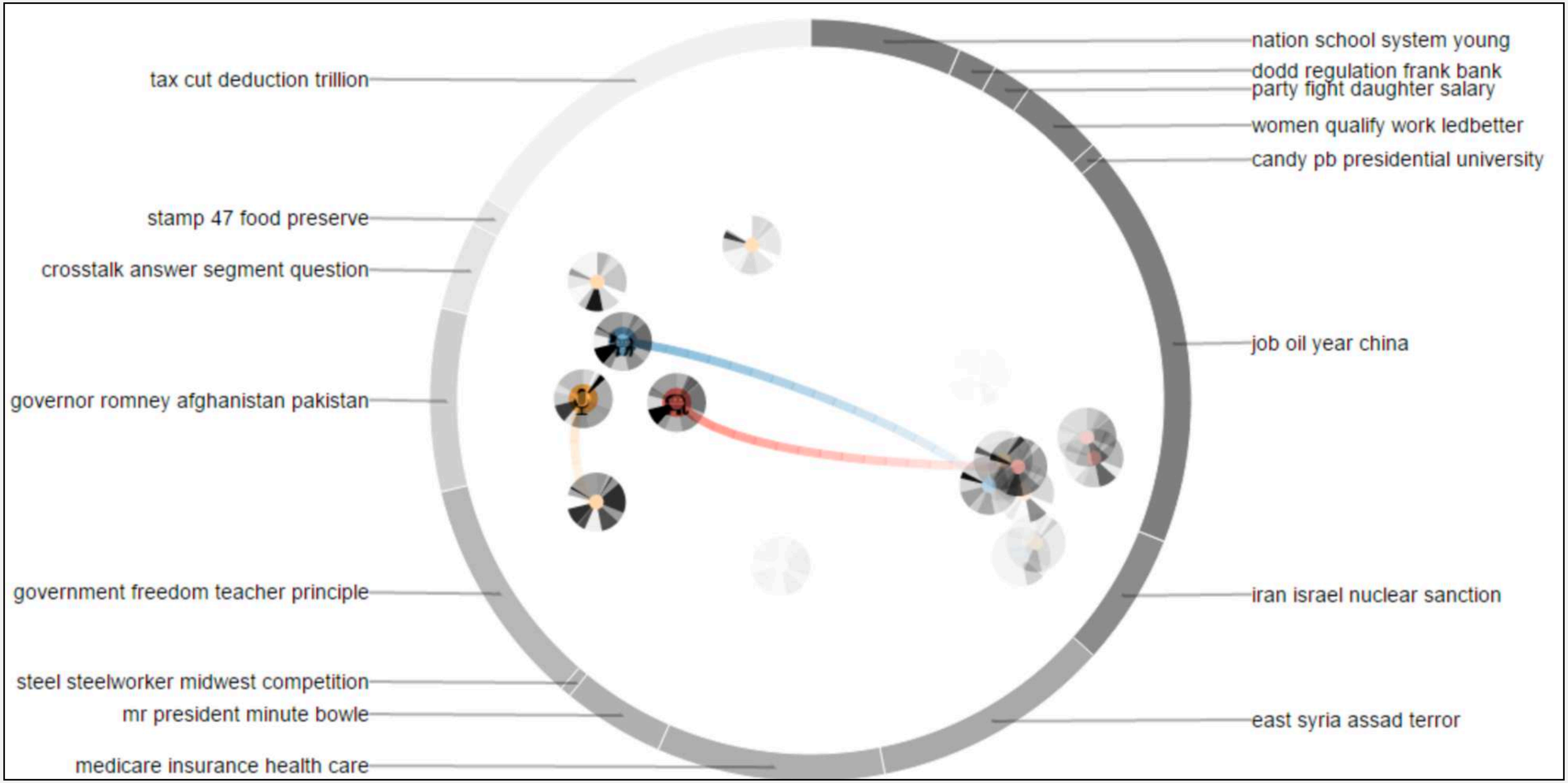
Topic Glyph

- Explicitly visually encode topic membership:



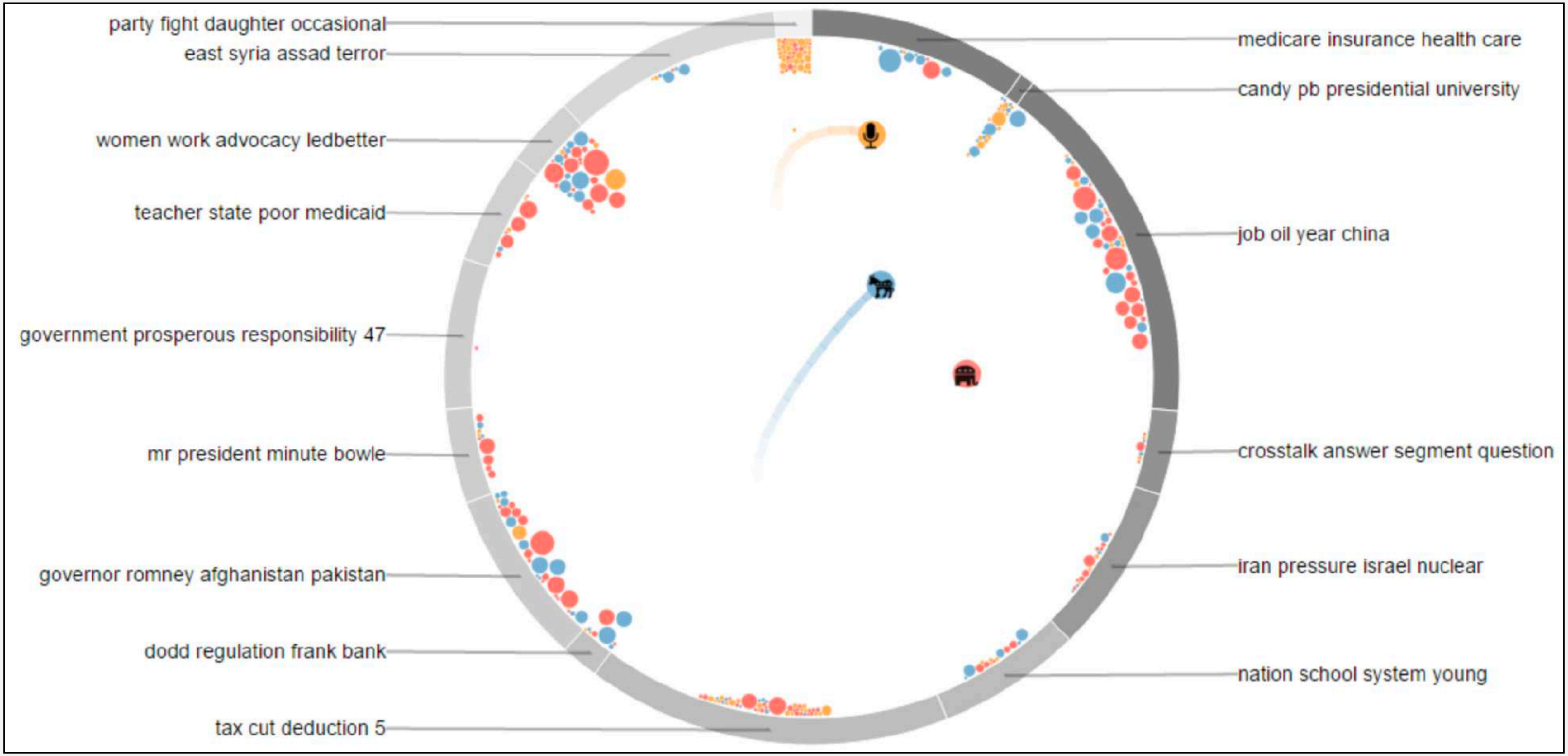
- Arc angles: map to the global view (previous slide)
- Brightness: document-topic weight
- Other visual encodings?

View 2: Glyphs


nation school system young
dodd regulation frank bank
party fight daughter salary
women qualify work ledbetter
candy pb presidential university
job oil year china
iran israel nuclear sanction
east syria assad terror
tax cut deduction trillion
stamp 47 food preserve
crosstalk answer segment question
governor romney afghanistan pakistan
government freedom teacher principle
steel steelworker midwest competition
mr president minute bowle
medicare insurance health care

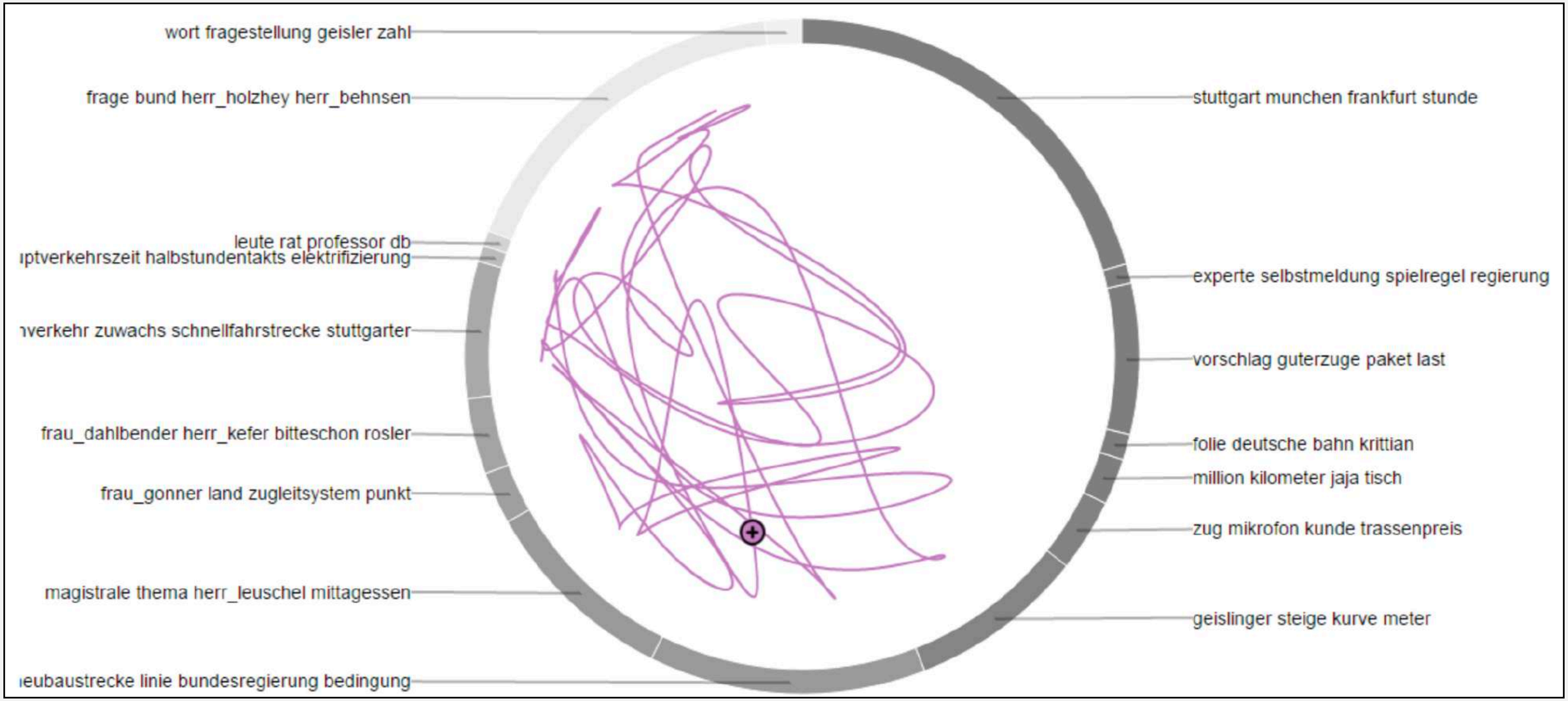
Animation with Context!

View 3: “Sedimentation”



Animation with Context!

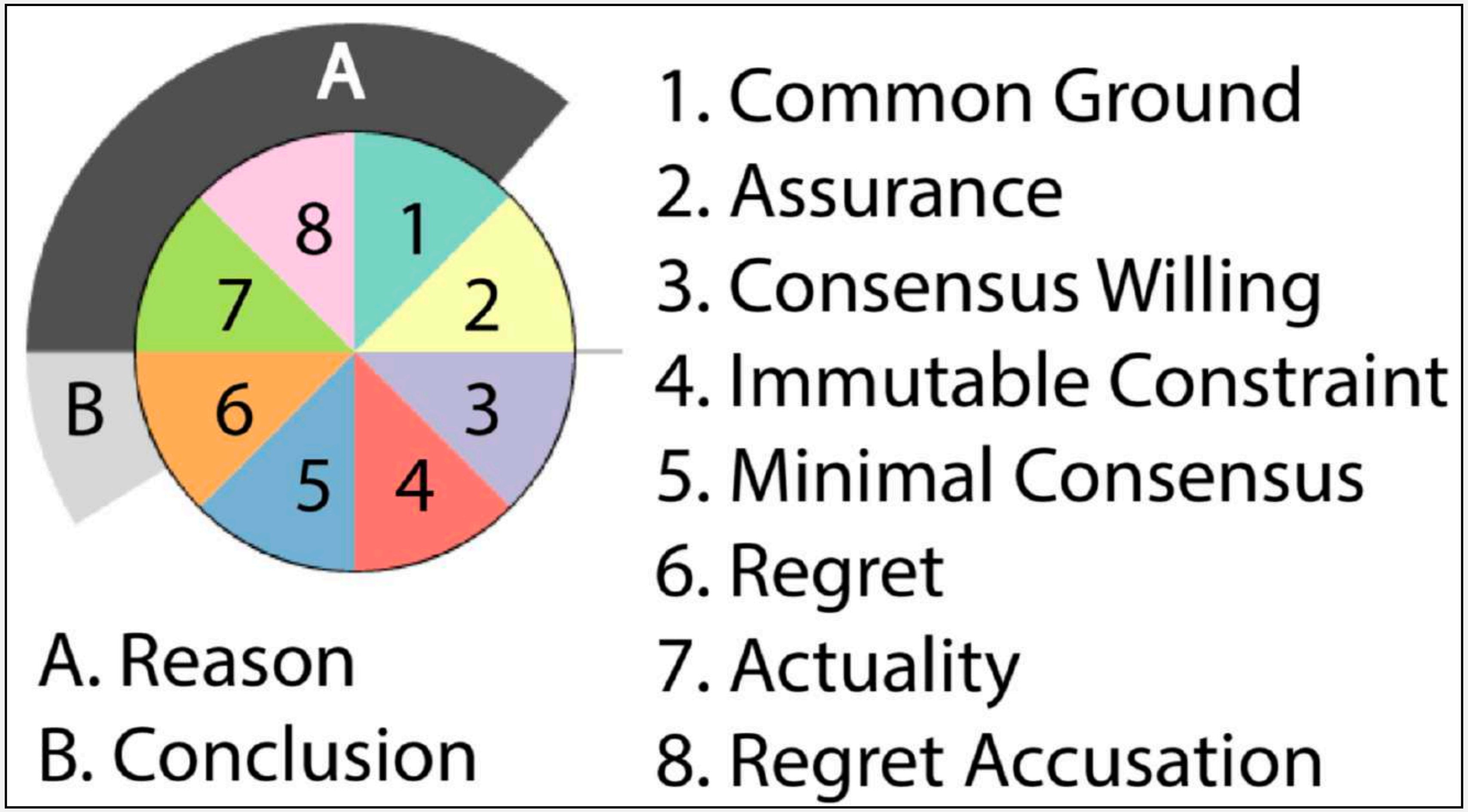
View 4: Speaker Paths


wort fragestellung geisler zahl
frage bund herr_holzhey herr_behnsen
leute rat professor db
ıptverkehrszeit halbstundentakts elektrifizierung
ıverkehr zuwachs schnellfahrstrecke stuttgarter
frau_dahlbender herr_kefer bitteschon rosler
frau_gonner land zugleitsystem punkt
magistrale thema herr_leuschel mittagessen
ıeubaustrecke linie bundesregierung bedingung
stuttgart munchen frankfurt stunde
experte selbstmeldung spielregel regierung
vorschlag guterzuge paket last
folie deutsche bahn krittian
million kilometer jaja tisch
zug mikrofon kunde trassenpreis
geislinger steige kurve meter

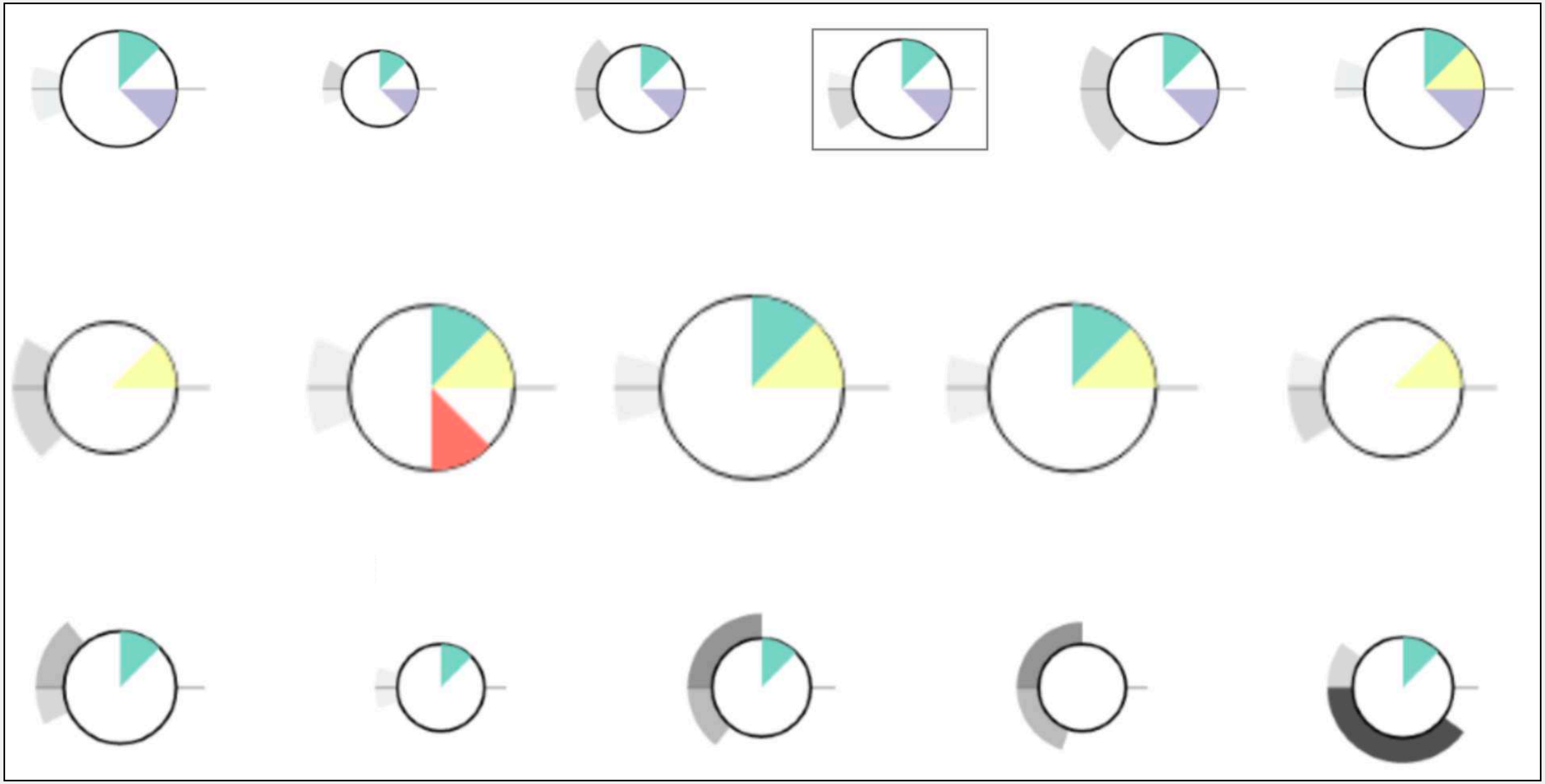
Potential Issues? Improvements?

Argumentation Patterns

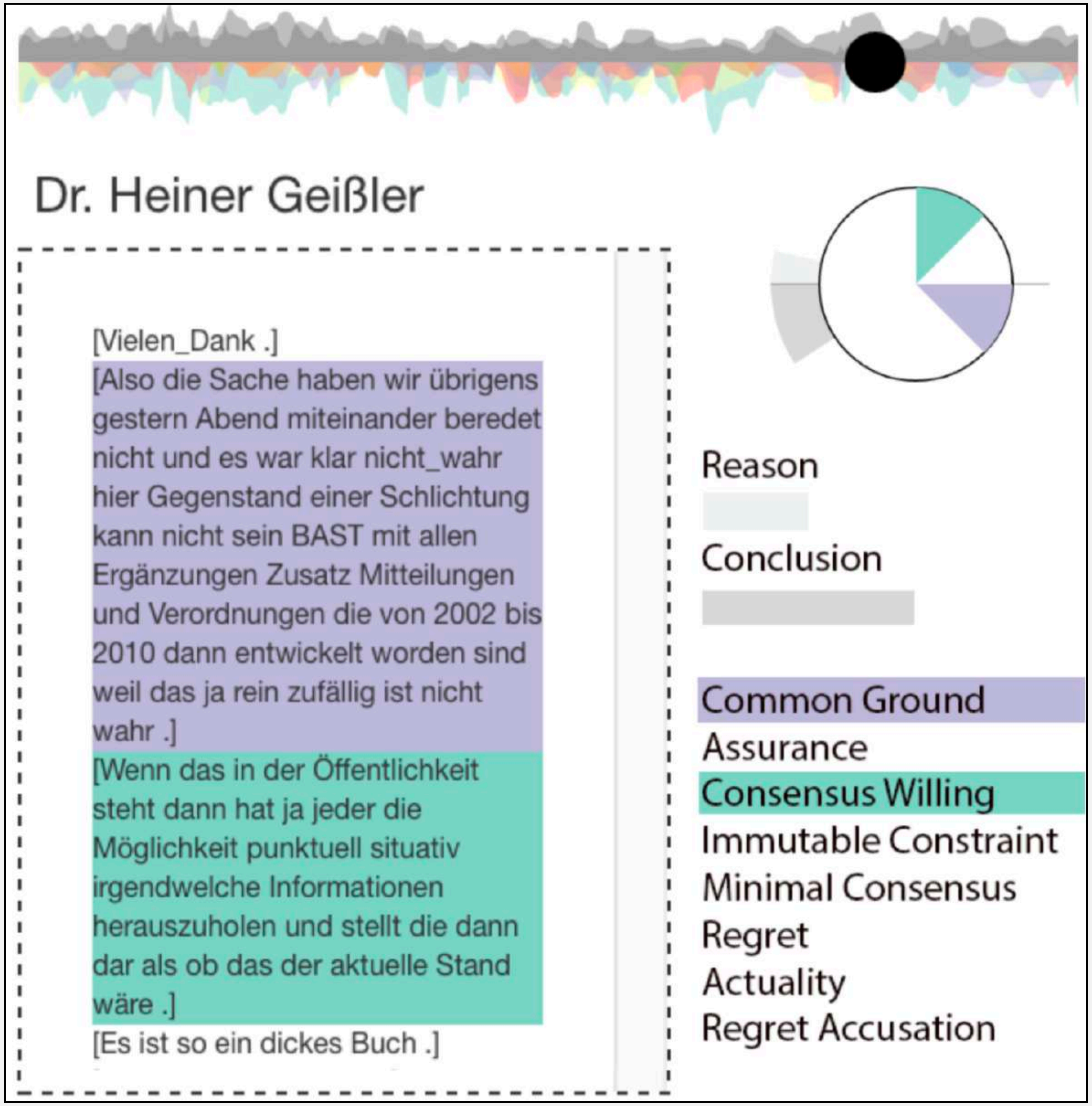
- Speaker patterns are extracted, used as additional semantics to understand utterances:



Comparing Stances



Detail-on-Demand



Dr. Heiner Geißler

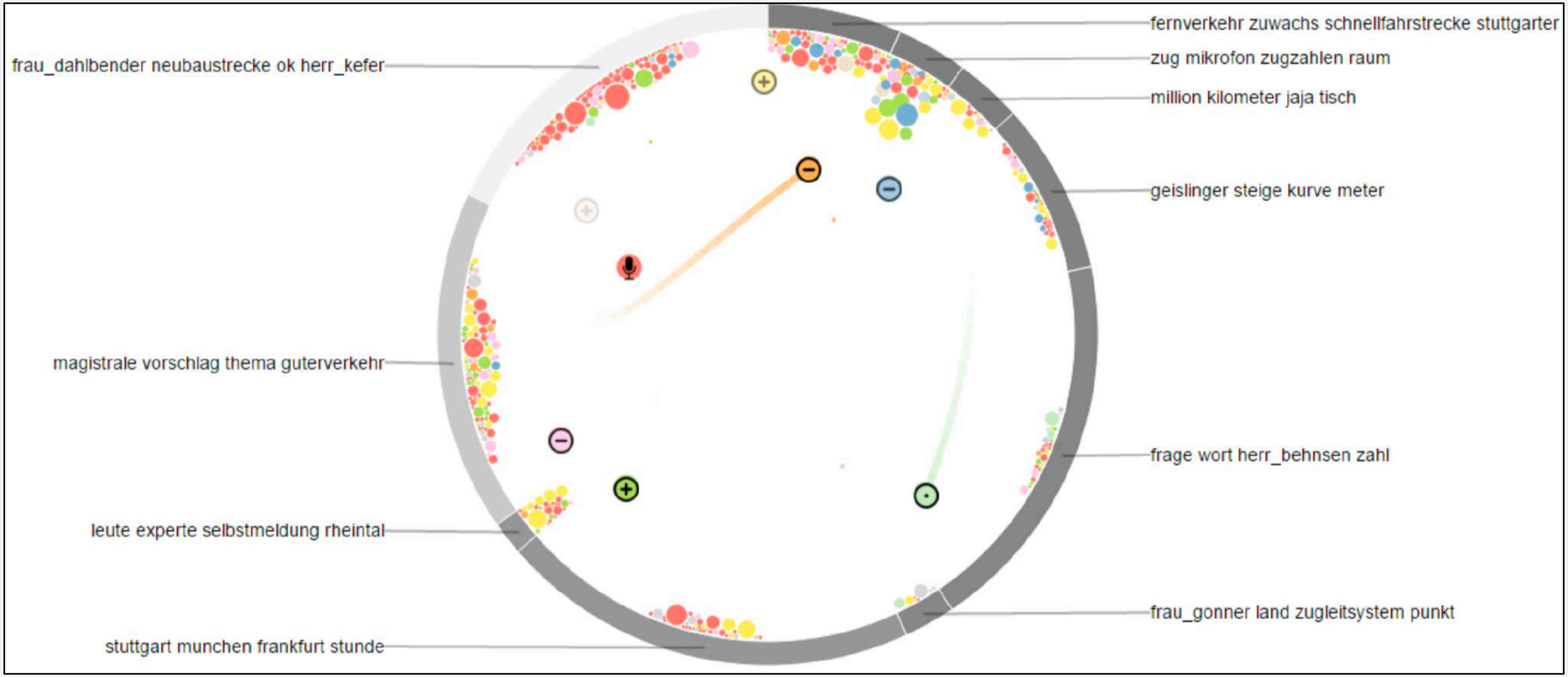
[Vielen_Dank .]

[Also die Sache haben wir übrigens gestern Abend miteinander beredet nicht und es war klar nicht_wahr hier Gegenstand einer Schlichtung kann nicht sein BAST mit allen Ergänzungen Zusatz Mitteilungen und Verordnungen die von 2002 bis 2010 dann entwickelt worden sind weil das ja rein zufällig ist nicht wahr .]

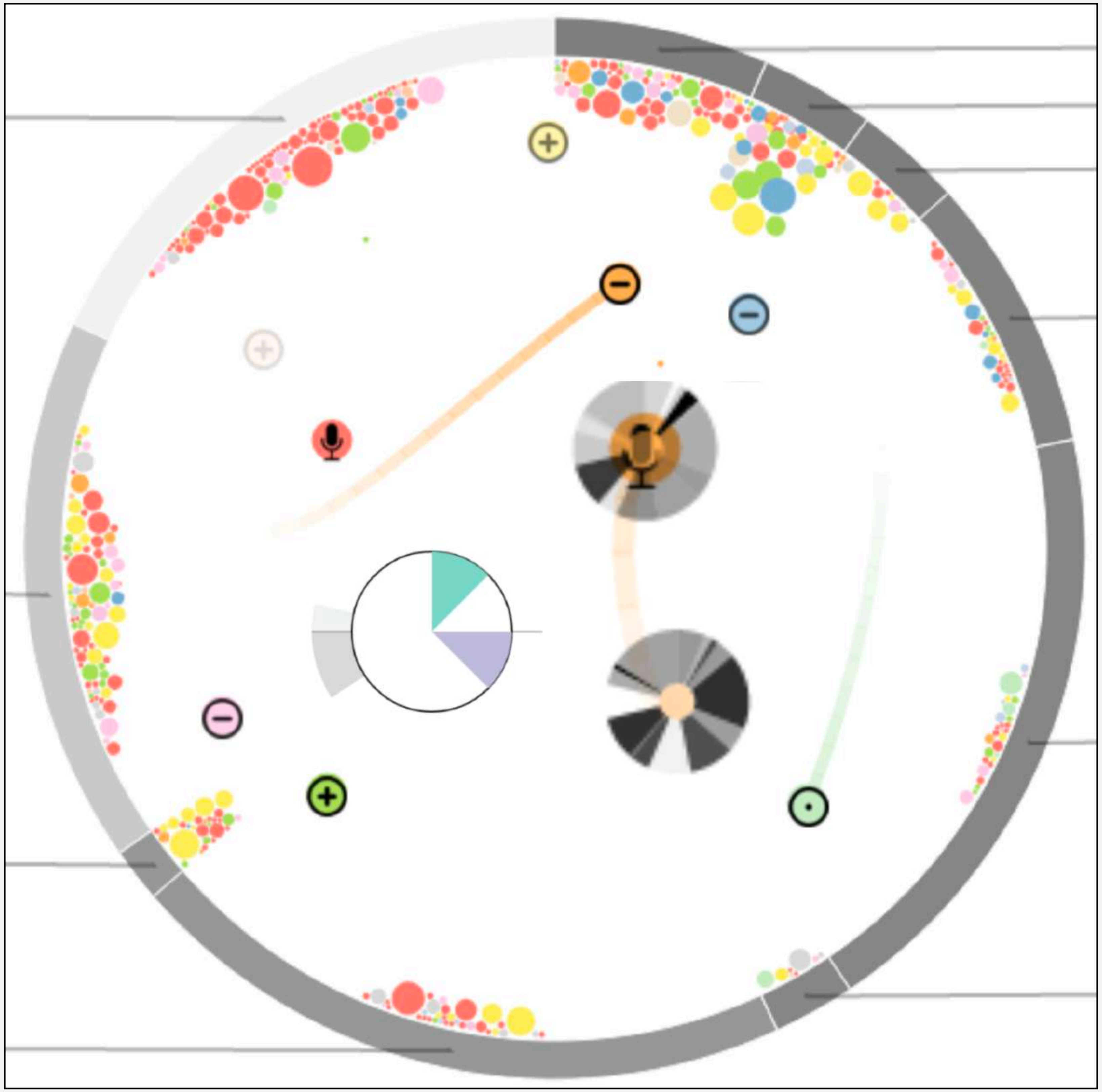
[Wenn das in der Öffentlichkeit steht dann hat ja jeder die Möglichkeit punktuell situativ irgendwelche Informationen herauszuholen und stellt die dann dar als ob das der aktuelle Stand wäre .]

[Es ist so ein dickes Buch .]

Putting it together


frau_dahlbender neubaustrecke ok herr_kefer
fernverkehr zuwachs schnellfahrstrecke stuttgarter
zug mikrofon zugzahlen raum
million kilometer jaja tisch
geislinger steige kurve meter
magistrale vorschlag thema guterverkehr
frage wort herr_behnsen zahl
leute experte selbstmeldung rheintal
frau_gonner land zugleitsystem punkt
stuttgart munchen frankfurt stunde

Sidebar: Encoding Overload?



Sidebar: Encoding Overload?

- Equally important to designing visual encodings is the ability for us (humans) to *decode* the visualization.

- **Decode**: going from visual channel back to data

- Be careful about *prioritizing* visual encodings:

  - Most important aspects of data should get mapped to channels, or combination of channels, that a human can easily decode (e.g. *channel effectiveness*)