

Exploration of Text Data: Topic Modeling, Information Retrieval, and their Visualizations

Jaakko Peltonen

Tampere University,

Faculty of Information Technology and Communication Sciences



Jaakko.Peltonen@tuni.fi

www.sis.uta.fi/~tojape/

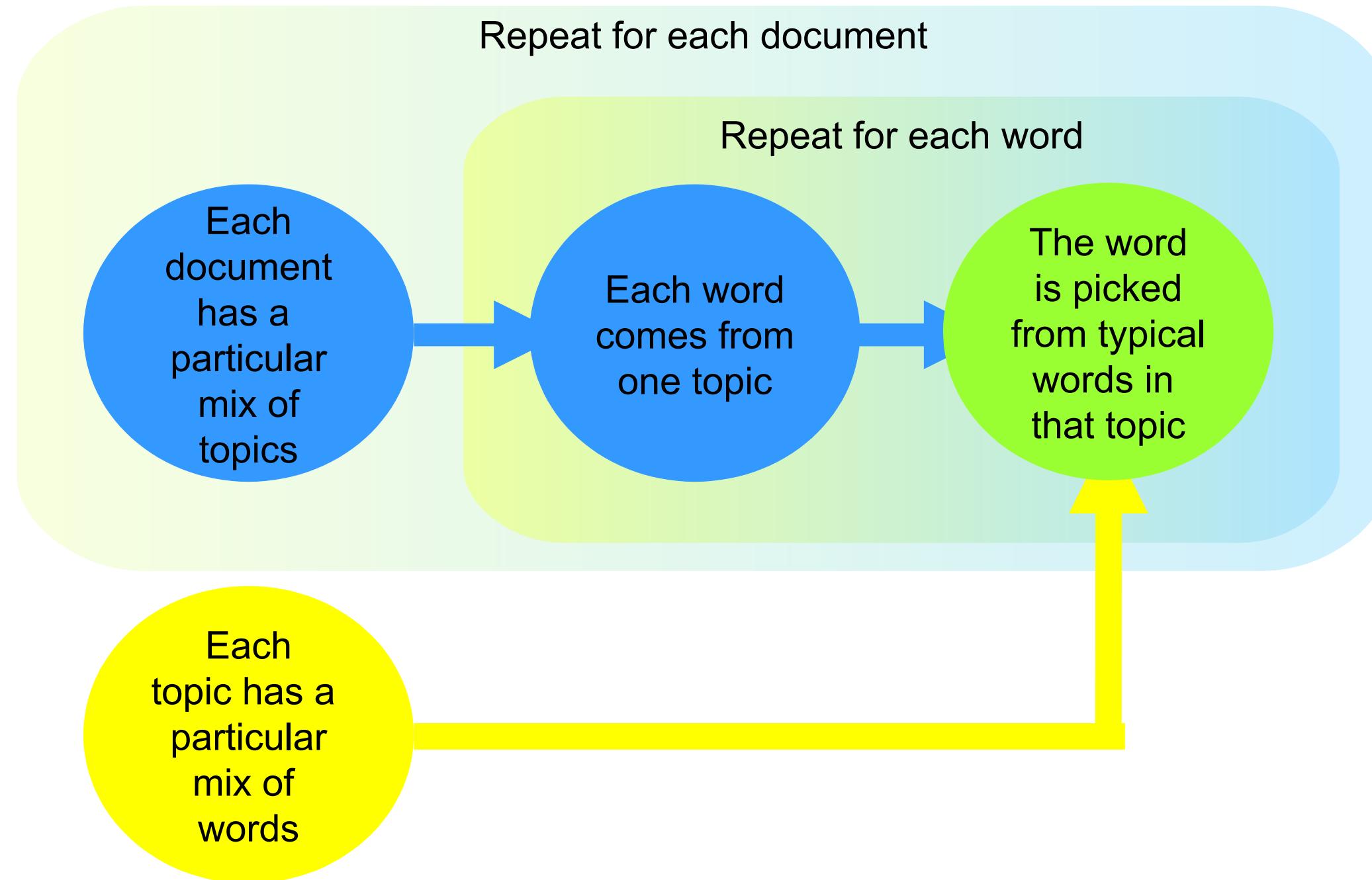
Types of text data

- Literature: fiction, nonfiction in multiple genres.
Online and digitized
- News: online news, digitized newspapers
- News comments
- Text content of webpages
- Search result snippets
- Online product descriptions
- Reviews: online and digitized
- Questionnaire answers
- Scripts and closed-caption tracks of movies and TV
- Scripts, closed-caption tracks, and other transcripts of online video
- Social media discussion
- Question-answer sites
- Instructions (e.g. recipes, instruction manuals, online how-tos)
- Online and digitized encyclopedias
- Online and digitized textbooks
- Scientific research articles
- Textual annotations for various data (e.g. biological experiment databases; RDF databases)
- Laws
- Court case records
- Patents
- Customer service records
- Service records, e.g. patient records

Topic Models - Idea

- Represent document content as bags of words
- Two-step process per word:
“choose what to talk about” (a topic), then
“choose what to say” (a word from the topic)
- Fit the model to data: learns the topics in the data
- Basic example: Latent Dirichlet Allocation
- Nonparametric version: Hierarchical Dirichlet Process topic model
- Deep hierarchies: Tree-structured Hierarchical Dirichlet Process, Author Tree-structured Hierarchical Dirichlet Process

Topic Models - Graphical representation



Latent Dirichlet Allocation - Mathematics

Latent Dirichlet Allocation (for machine learning: David Blei, Andrew Ng, Michael Jordan, JMLR 2003):

very popular probabilistic model for underlying themes in text data collections

Latent Dirichlet Allocation - Mathematics

Generative process:

For each topic $z = 1, \dots, K$: let there be a distribution of words $p(w|z; \beta)$

For each document $d = 1, \dots, M$:

1. Choose the number of words

$$N_d \sim \text{Poisson}(\xi)$$

2. Choose proportions of topics in this document: $\theta \sim \text{Dirichlet}(\alpha)$

3. For each word $n = 1, \dots, N$:

1. **Choose a topic** $z \sim \text{Multinomial}(\theta)$

2. **Choose the word** w from $p(w|z; \beta)$

Latent Dirichlet Allocation (for machine learning: David Blei, Andrew Ng, Michael Jordan, JMLR 2003):

very popular probabilistic model for underlying themes in text data collections

Latent Dirichlet Allocation - Mathematics

Generative process:

For each topic $z = 1, \dots, K$: let there be a distribution of words $p(w|z; \beta)$

For each document $d = 1, \dots, M$:

1. Choose the number of words

$$N_d \sim \text{Poisson}(\xi)$$

2. Choose proportions of topics in this document: $\theta \sim \text{Dirichlet}(\alpha)$

3. For each word $n = 1, \dots, N$:

1. **Choose a topic** $z \sim \text{Multinomial}(\theta)$

2. **Choose the word** w from $p(w|z; \beta)$

Likelihood of data:

$$\prod_{d=1}^M \int_{\theta_d} p(\theta_d | \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}=1}^K p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}; \beta) \right) d\theta_d$$

For each document

For each word in the document

Latent Dirichlet Allocation - Mathematics

Generative process:

For each topic $z = 1, \dots, K$: let there be a distribution of words $p(w|z; \beta)$

For each document $d = 1, \dots, M$:

1. Choose the number of words

$$N_d \sim \text{Poisson}(\xi)$$

2. Choose proportions of topics in this document: $\theta \sim \text{Dirichlet}(\alpha)$

3. For each word $n = 1, \dots, N$:

1. **Choose a topic** $z \sim \text{Multinomial}(\theta)$

2. **Choose the word** w from $p(w|z; \beta)$

Likelihood of data:

$$\prod_{d=1}^M \int_{\theta_d} p(\theta_d | \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}=1}^K p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}; \beta) \right) d\theta_d$$

For each document

For each word in the document

Parameters to be optimized:

$$\alpha, \beta, (\text{and } \xi)$$

Optimization:

Variational Bayes: approximate posterior distributions of parameters

Gibbs sampling: draw samples from the posterior

Latent Dirichlet Allocation - results

- For each document: topic proportions
e.g. [Topic1: 0.2, Topic2: 0.4, Topic3: 0.3, Topic4: 0.1]
- For each topic: word distribution, e.g.

Topic1:

visualization	0.15
plot	0.13
graph	0.11
algorithm	0.10
method	0.09
view	0.08
interface	0.08
interaction	0.07
experiment	0.06
layout	0.05
overview	0.05
user	0.03

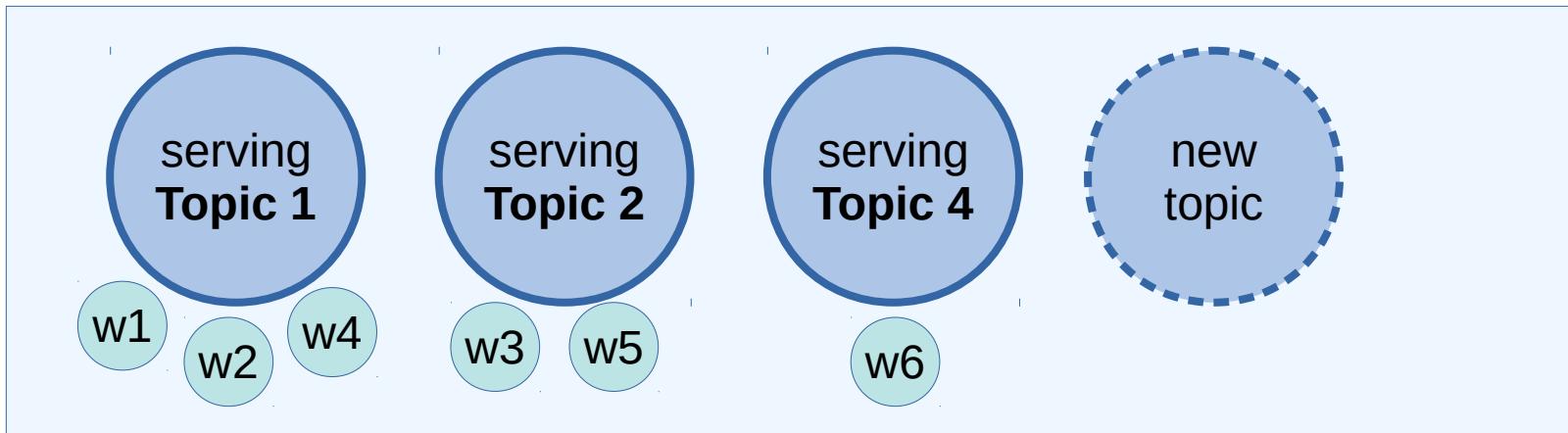
Topic2:

graph	0.16
edge	0.15
node	0.13
vertex	0.11
layout	0.10
drawing	0.09
crossing	0.09
marker	0.07
bundle	0.04
link	0.03
diagram	0.02
adjacency	0.01

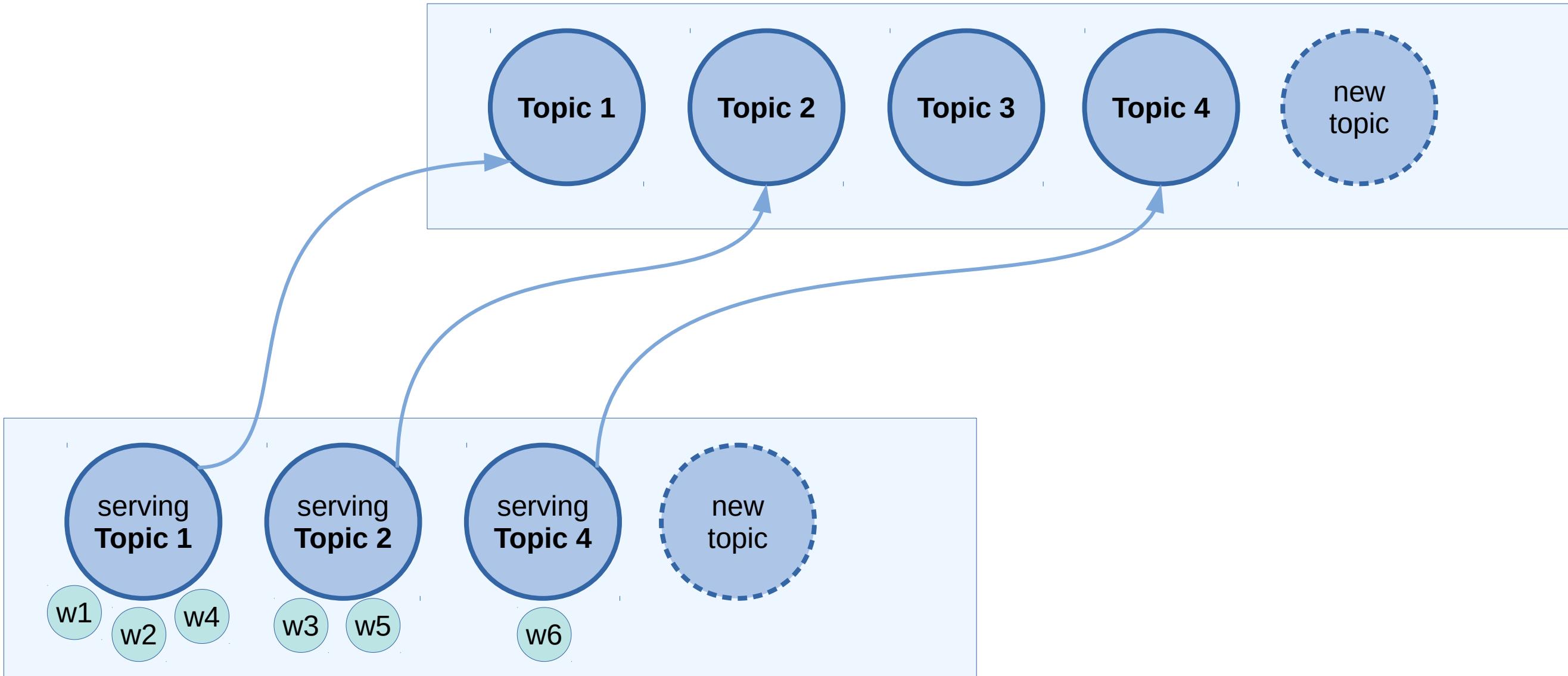
Nonparametric topic models: Hierarchical Dirichlet Process

- Latent Dirichlet Allocation assumes the number of topics is known and fixed.
- Nonparametric modeling by Dirichlet processes insteads learns the needed number of topics from the data.
- A "Dirichlet process" (DP) is a prior over multinomial distributions with varying numbers of (topic) possibilities with no upper limit.
- Each sample from a DP is a distribution with some finite number of possibilities.
- In DP topic modeling, you don't need to actually sample the distributions: it is enough to be able to decide which word came from which topic
- "Blackwell-McQueen urn", "stick-breaking process", "Chinese restaurant process": different representations for the data generation process

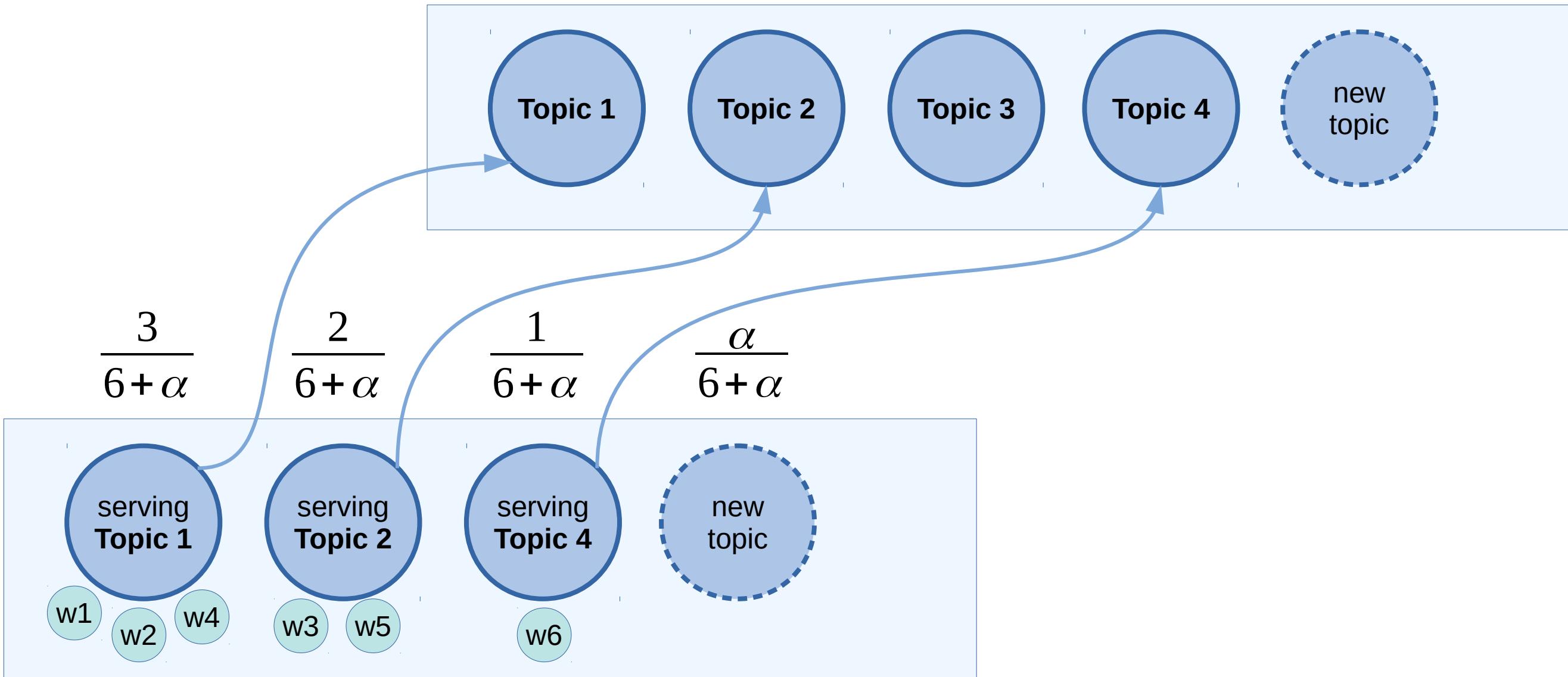
HDP topic model inference



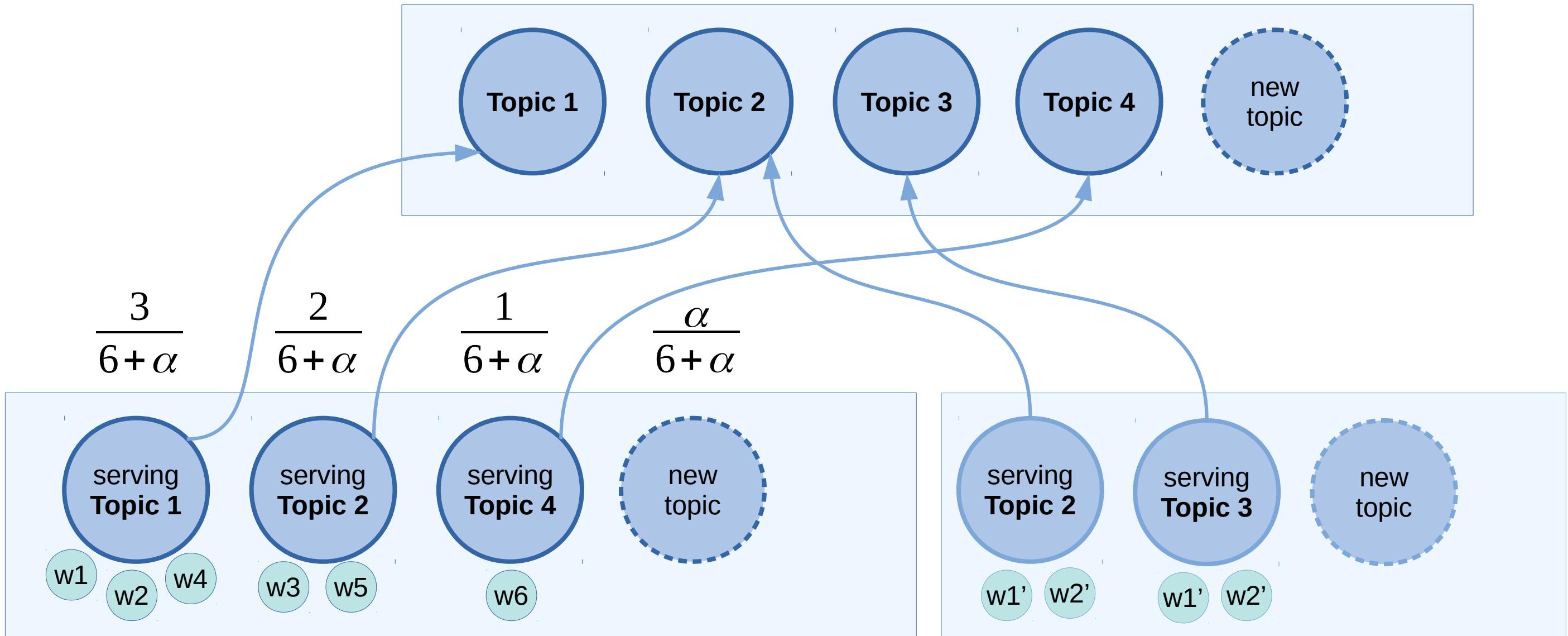
HDP topic model inference



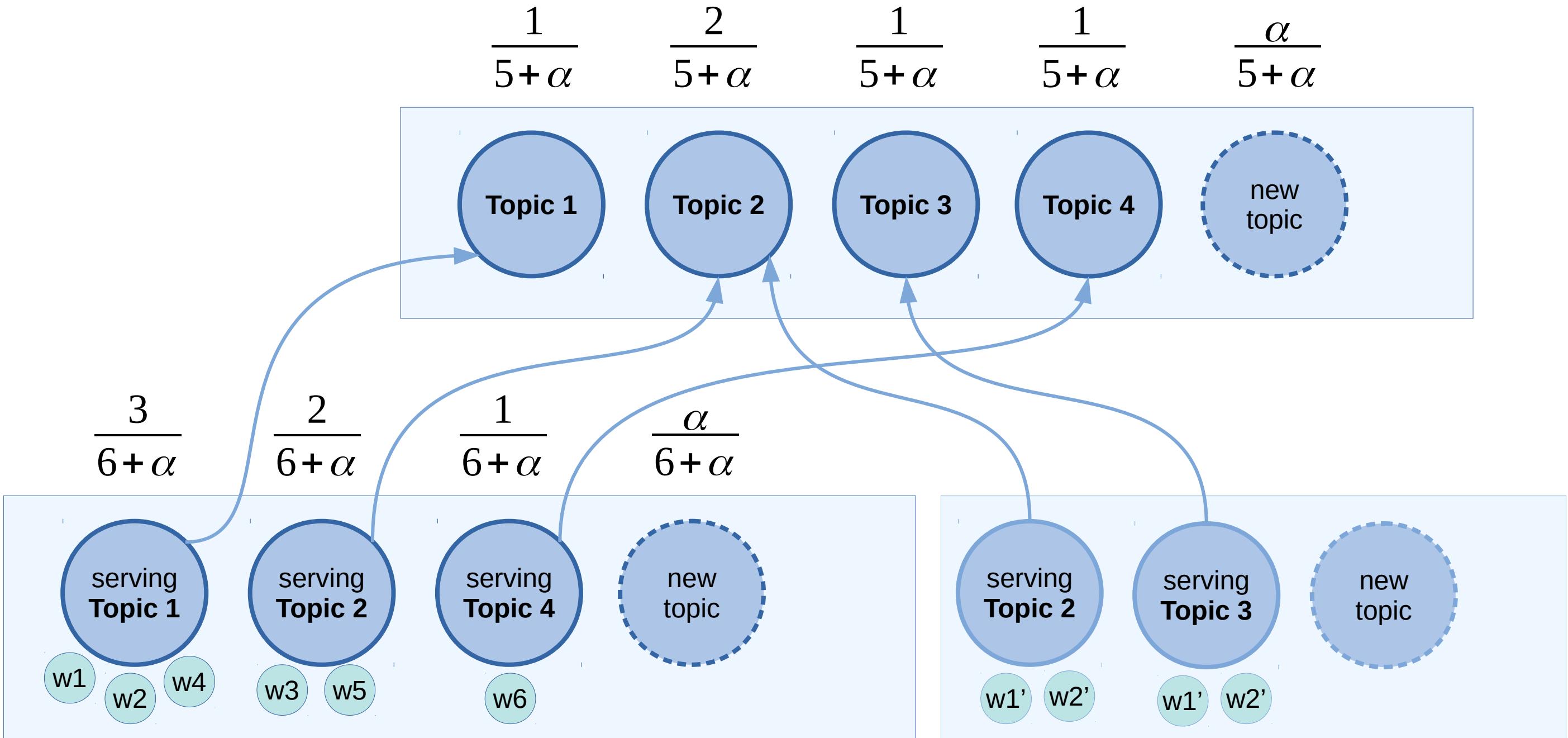
HDP topic model inference



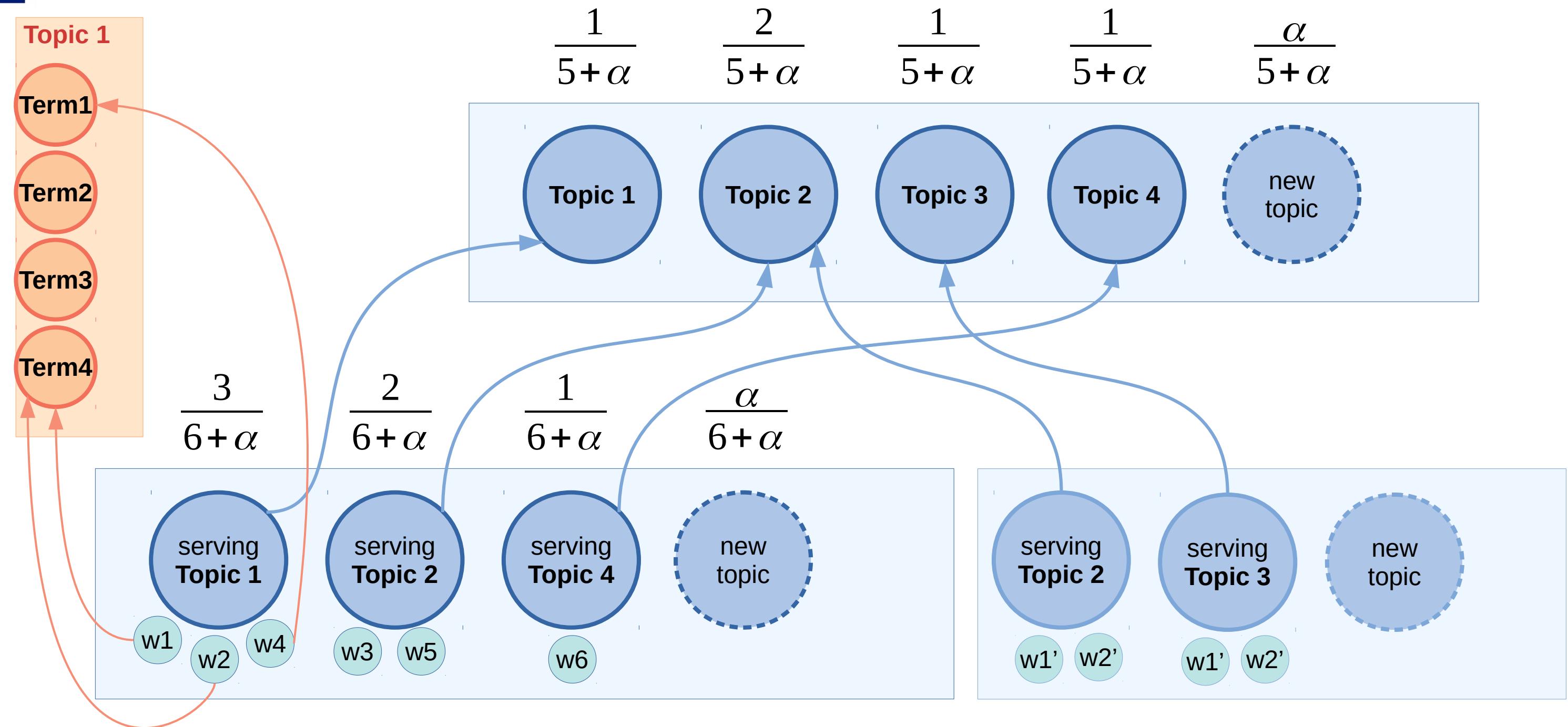
HDP topic model inference



HDP topic model inference

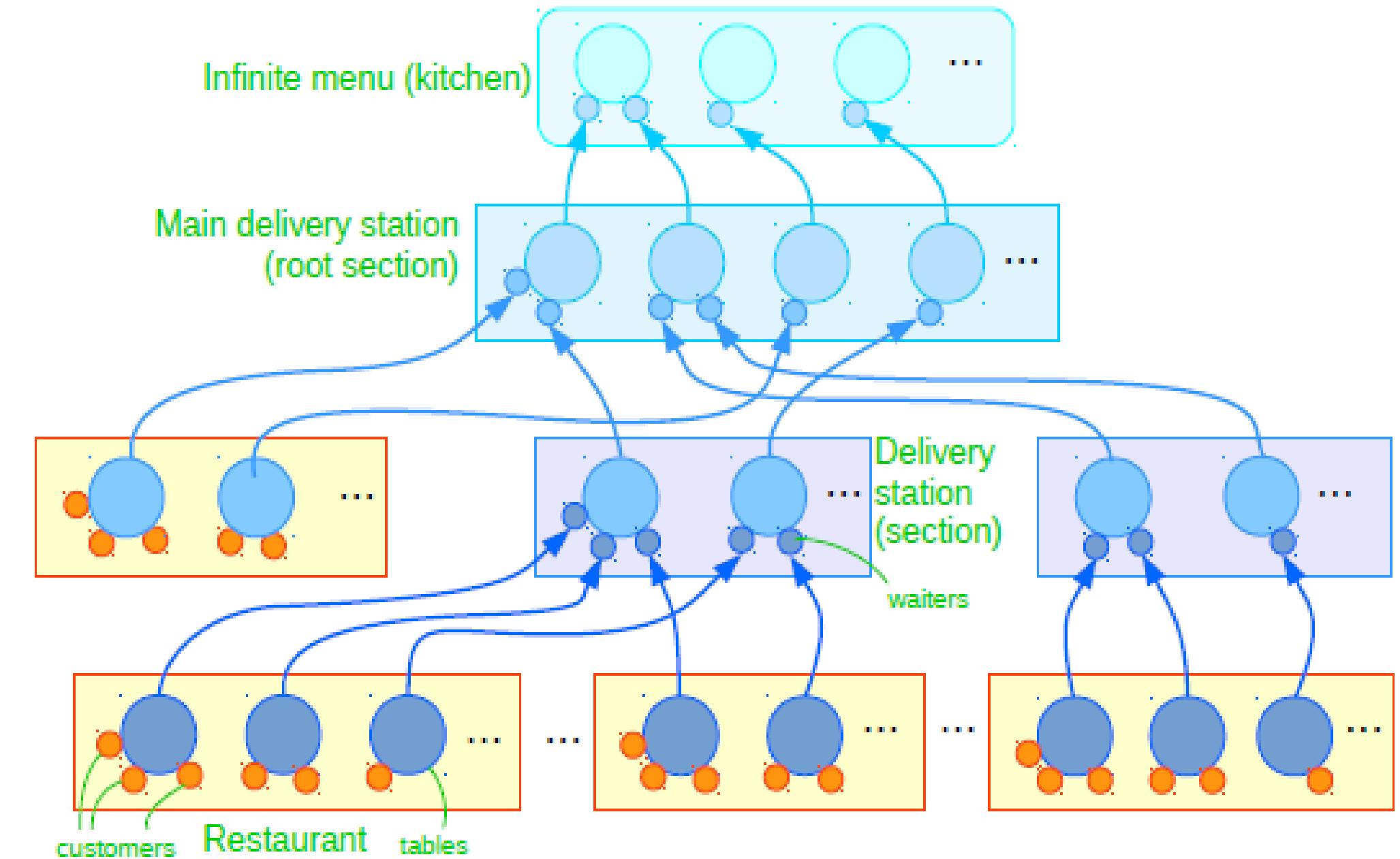


HDP topic model inference



Extension to deep hierarchies: THDP

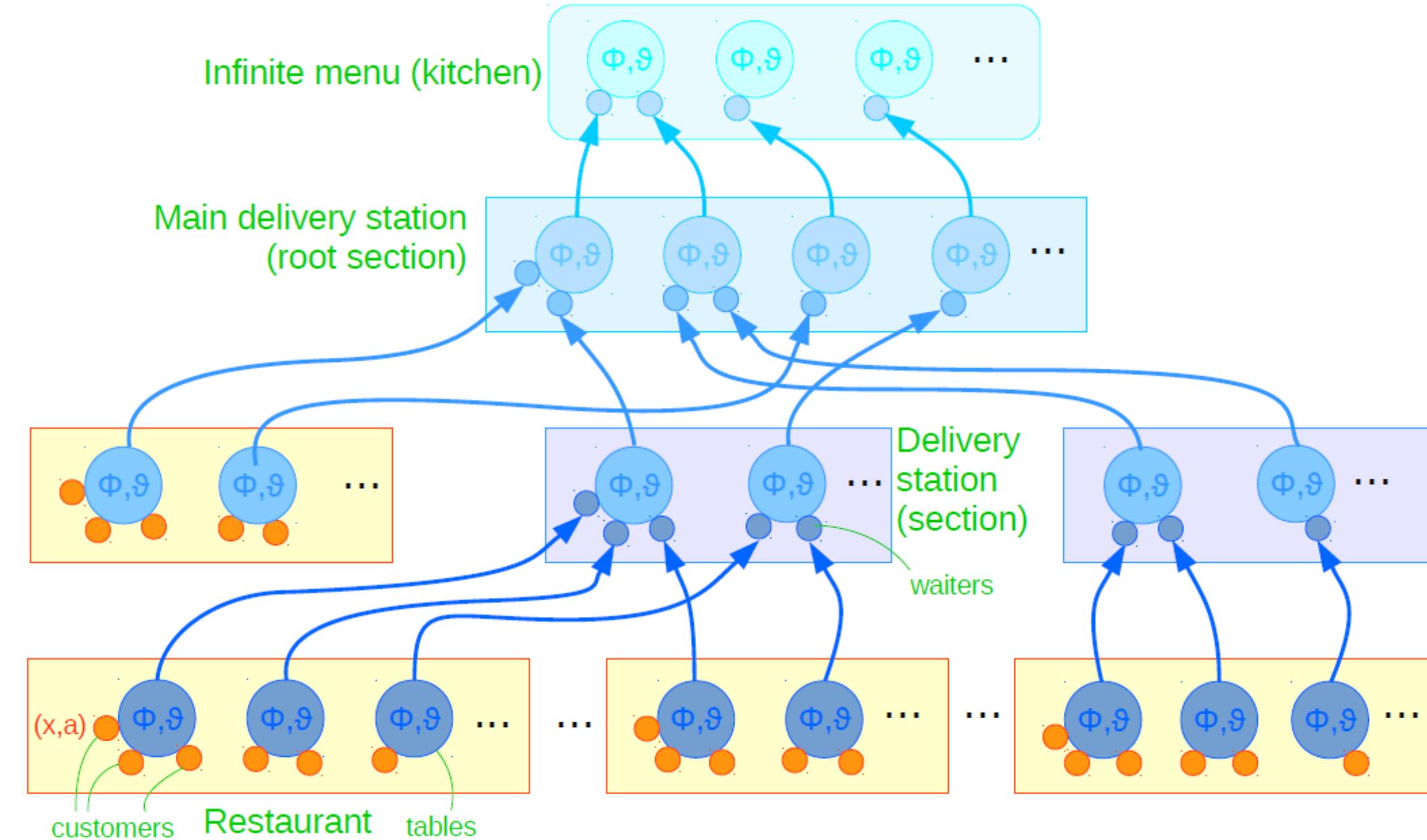
- THDP (Alam et al., DCAI 2018) extends the nonparametric inference to deep hierarchies like multi-level conversation forums



Picture from: H. Alam, J. Peltonen, J. Nummenmaa, and K. Järvelin. Tree-structured Hierarchical Dirichlet Process. In proceedings of DCAI 2018, Springer, 2018.

Extension to deep hierarchies and authors: ATHDP

- ATHDP (Alam et al., DS 2018) also models contribution of different authors



Picture from: H. Alam, J. Peltonen, J. Nummenmaa & K. Järvelin. Author Tree-structured Hierarchical Dirichlet Process. In proceedings of DS 2018, Springer, 2018.

THDP topic model - results

- Number of active topics
- For each discussion area and document: topic proportions
e.g. [Topic1: 0.2, Topic2: 0.4, Topic3: 0.3, Topic4: 0.1]
- For each topic: word distribution, e.g.

Topic1:

visualization	0.15
plot	0.13
graph	0.11
algorithm	0.10
method	0.09
view	0.08
interface	0.08
interaction	0.07
experiment	0.06
layout	0.05
overview	0.05
user	0.03

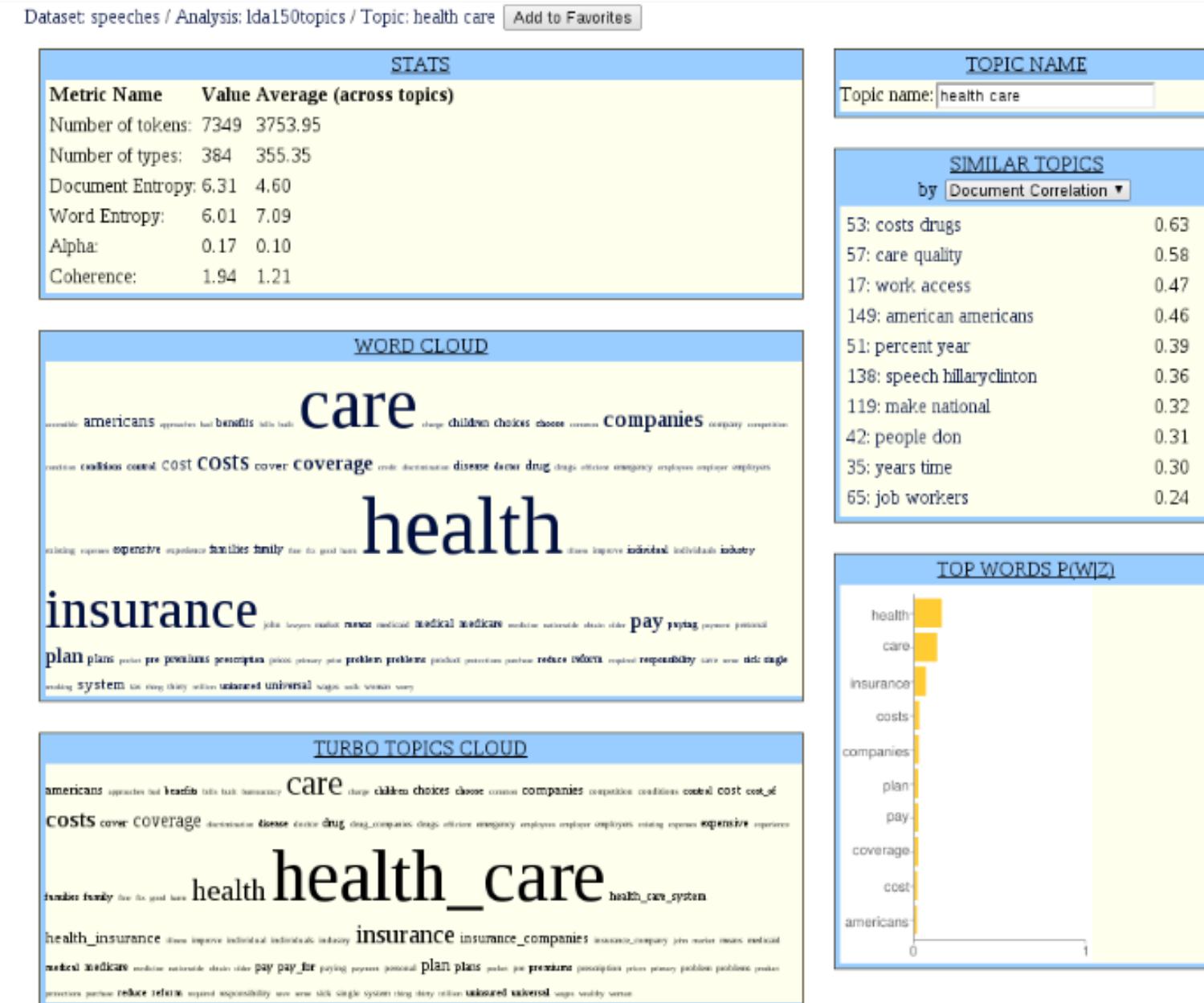
Topic2:

graph	0.16
edge	0.15
node	0.13
vertex	0.11
layout	0.10
drawing	0.09
crossing	0.09
marker	0.07
bundle	0.04
link	0.03
diagram	0.02
adjacency	0.01

Topic model visualization

Topic Browser
(Gardner et al. 2010):
sorts/filters topics e.g.
by dispersion; shows

- top words per topic
- context in random document,
- top documents per topic,
- similar topics, and
- category prevalences



Picture from: M.J. Gardner, J. Lutes, J. Lund, J. Hansen, D. Walker, E. Ringger, and K. Seppe. The Topic Browser: An Interactive Tool for Browsing Topic Models. In NIPS workshop on challenges of data visualization, 2010.

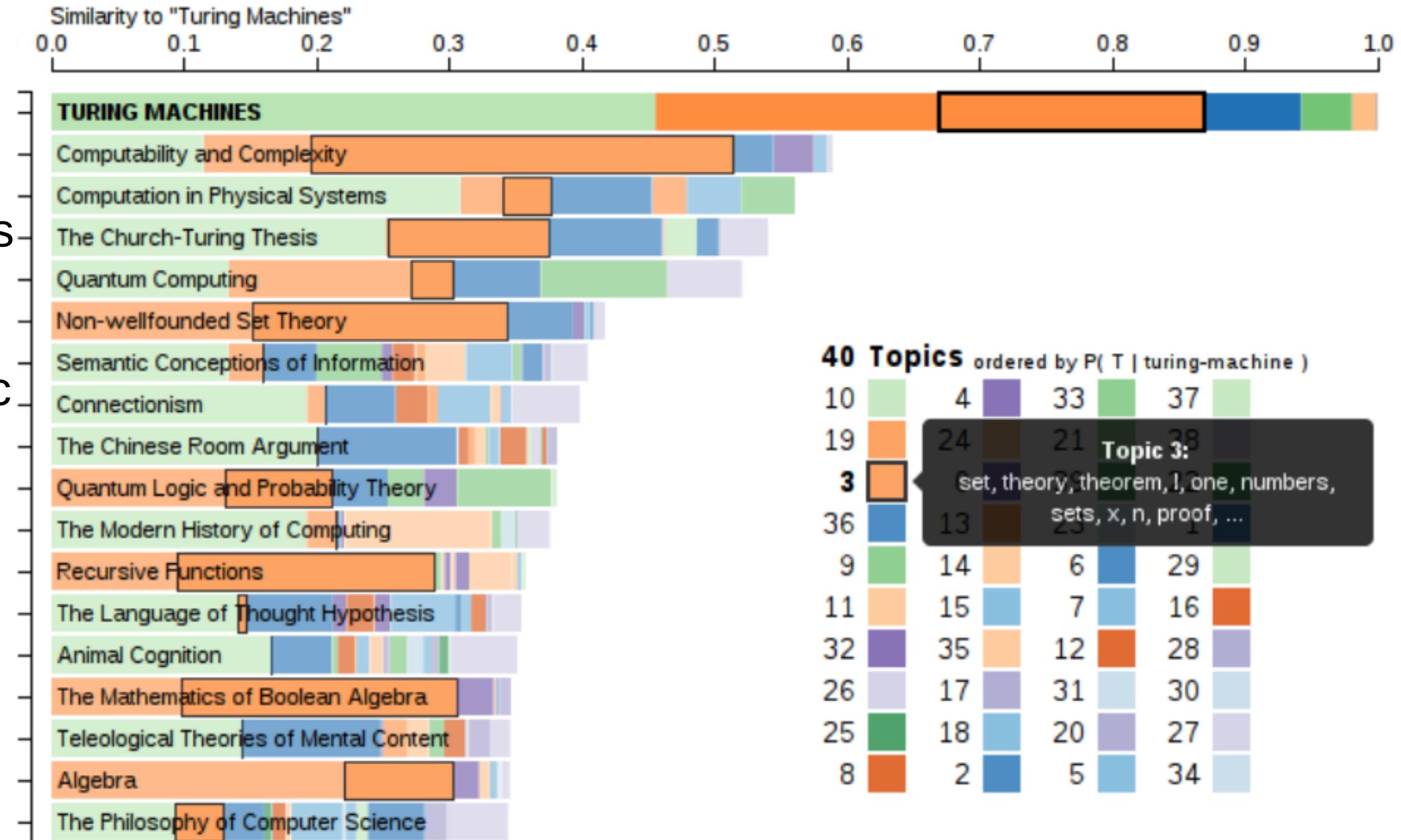
Topic model visualization



Picture from: A.J.B. Chaney and D.M. Blei. Visualizing Topic Models. In AAAI Conference on Weblogs and Social Media, 2012.

Topic model visualization

- Topic Explorer (Murdock and Allen, AAAI'15): orders documents by similarity to selected document or topic shows topic distribution in documents.

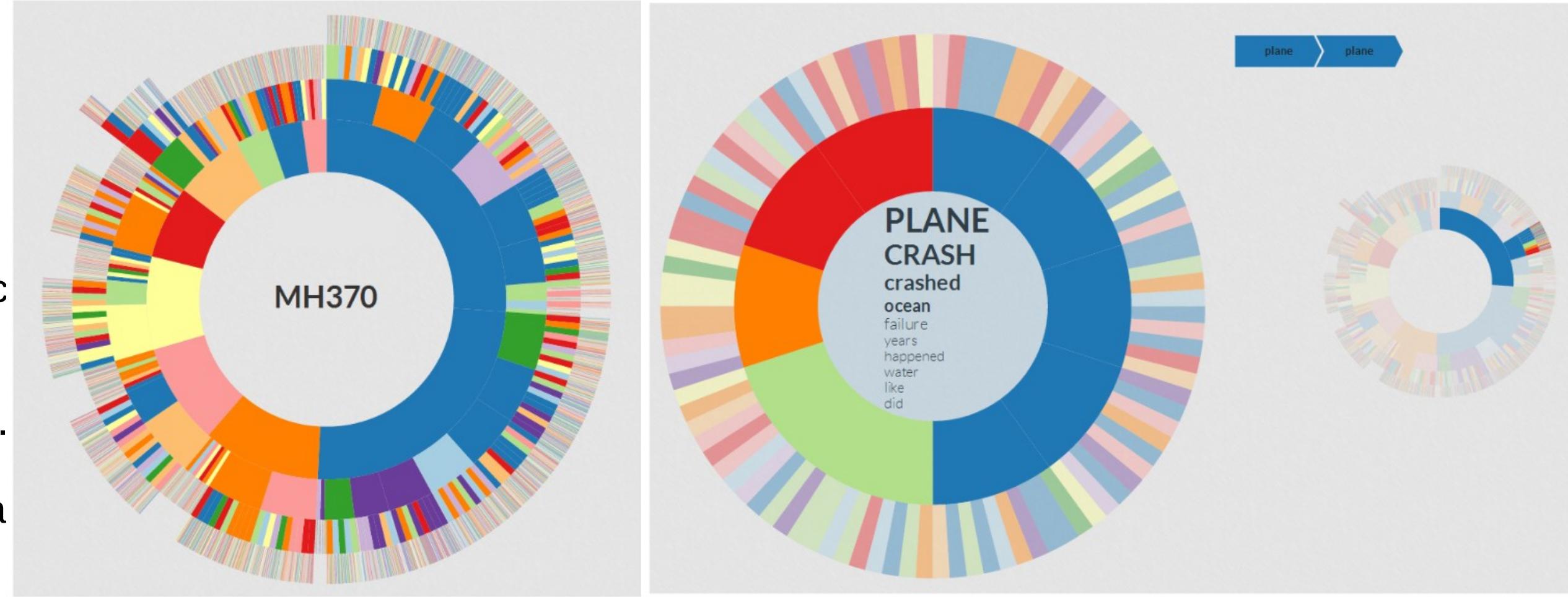


Picture from: J. Murdock and C. Allen. Visualization Techniques for Topic Model Checking. In AAAI'15, AAAI, 2015.

Topic model visualization

Hiérarchie
(Smith et al.
2014):

- splits each topic into subtopics using synthetic documents.
- Shown in a sunburst chart.
- User can click to zoom in to a topic and its subtopics.

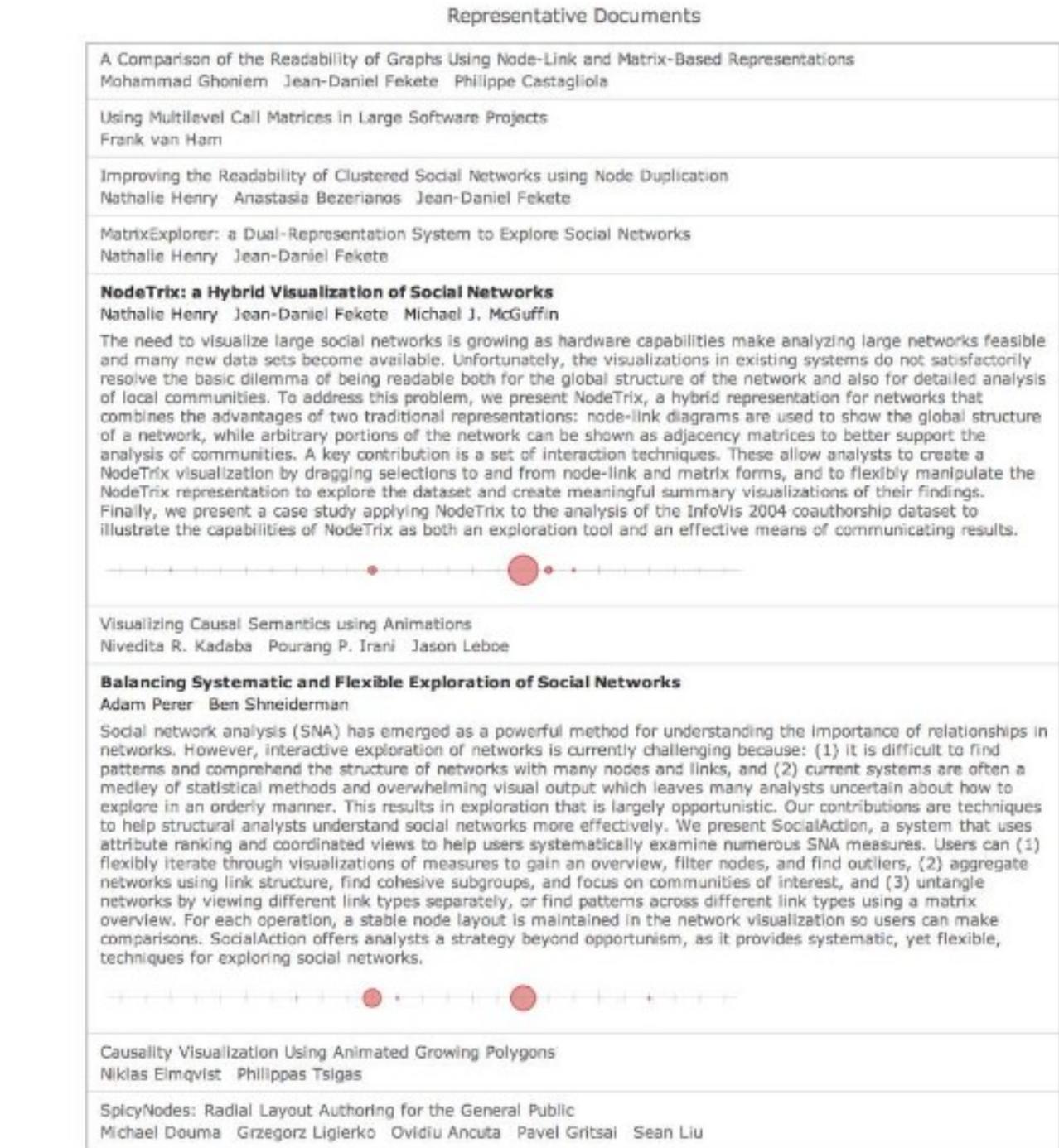
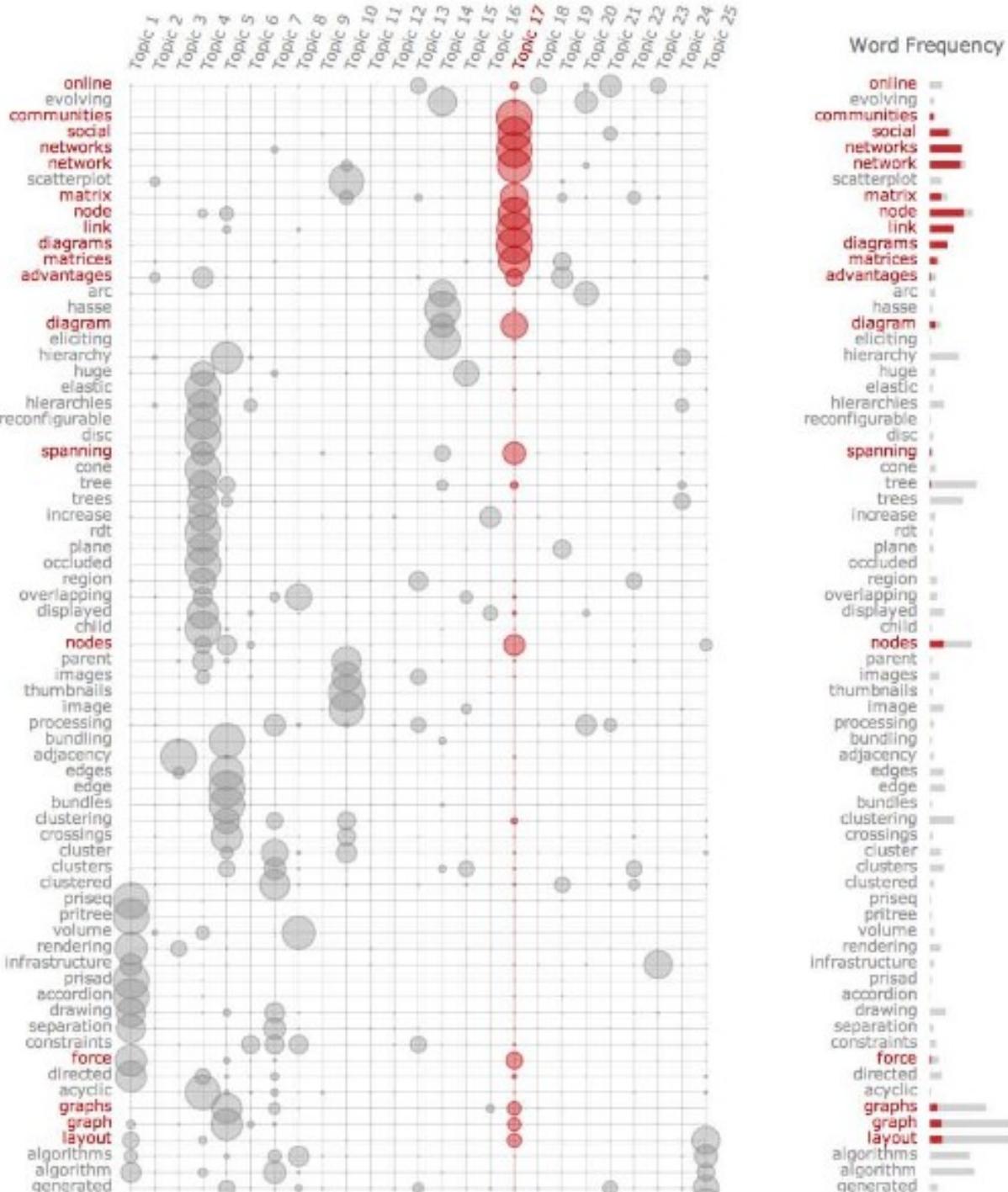


Pictures from: A. Smith, T. Hawes, and M. Myers. Interactive Visualization for Hierarchical Topic Models. Workshop on Interactive Language Learning, Visualization, and Interfaces, 2014.

Topic model visualization

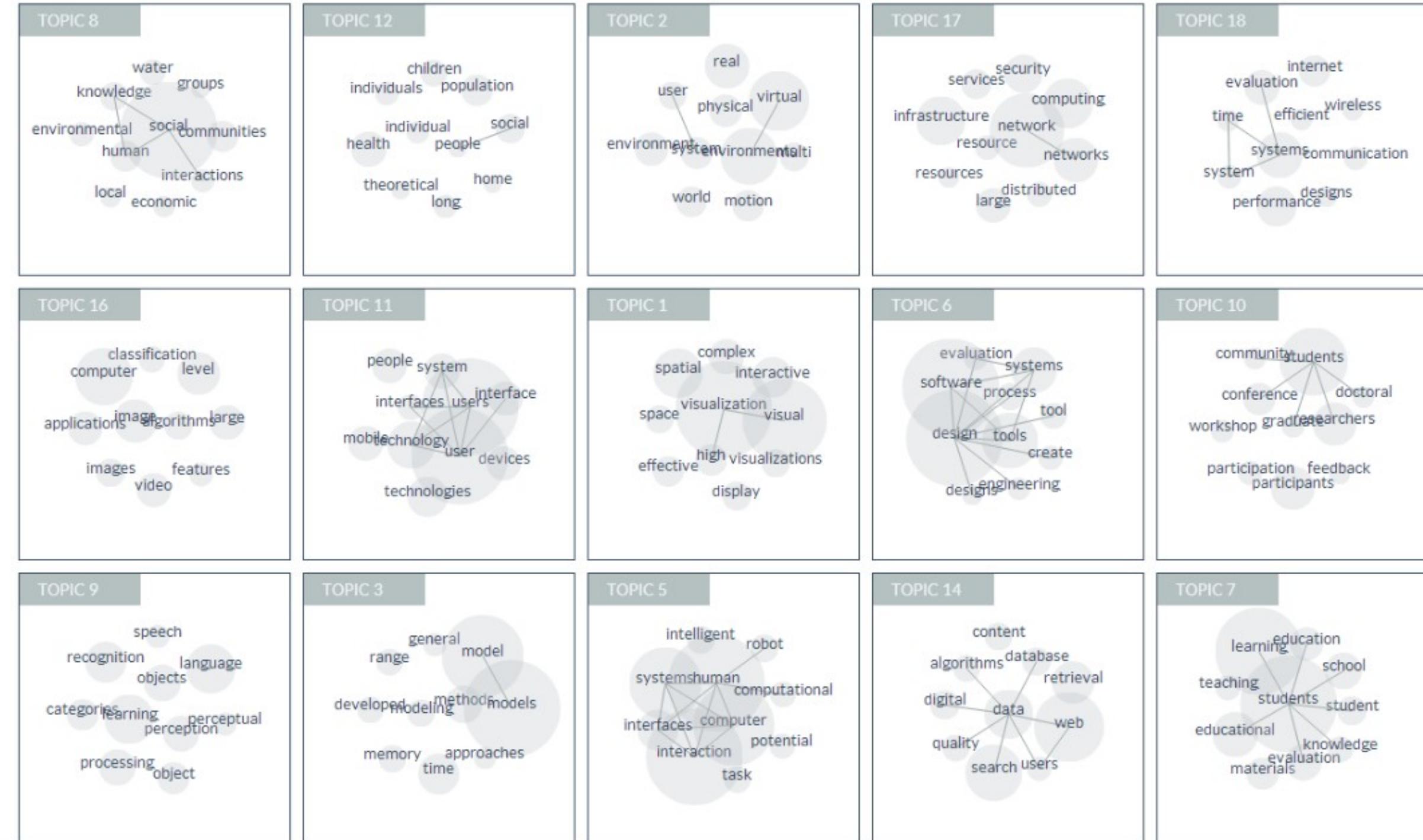
Termite
(Chuang et al., 2012):
term vs topic
matrices
with seriation

Picture from: J. Chuang, C.D. Manning, J. Heer. Termite: Visualization Techniques for Assessing Textual Topic Models. In AVI'12, ACM, 2012.



Topic model visualization

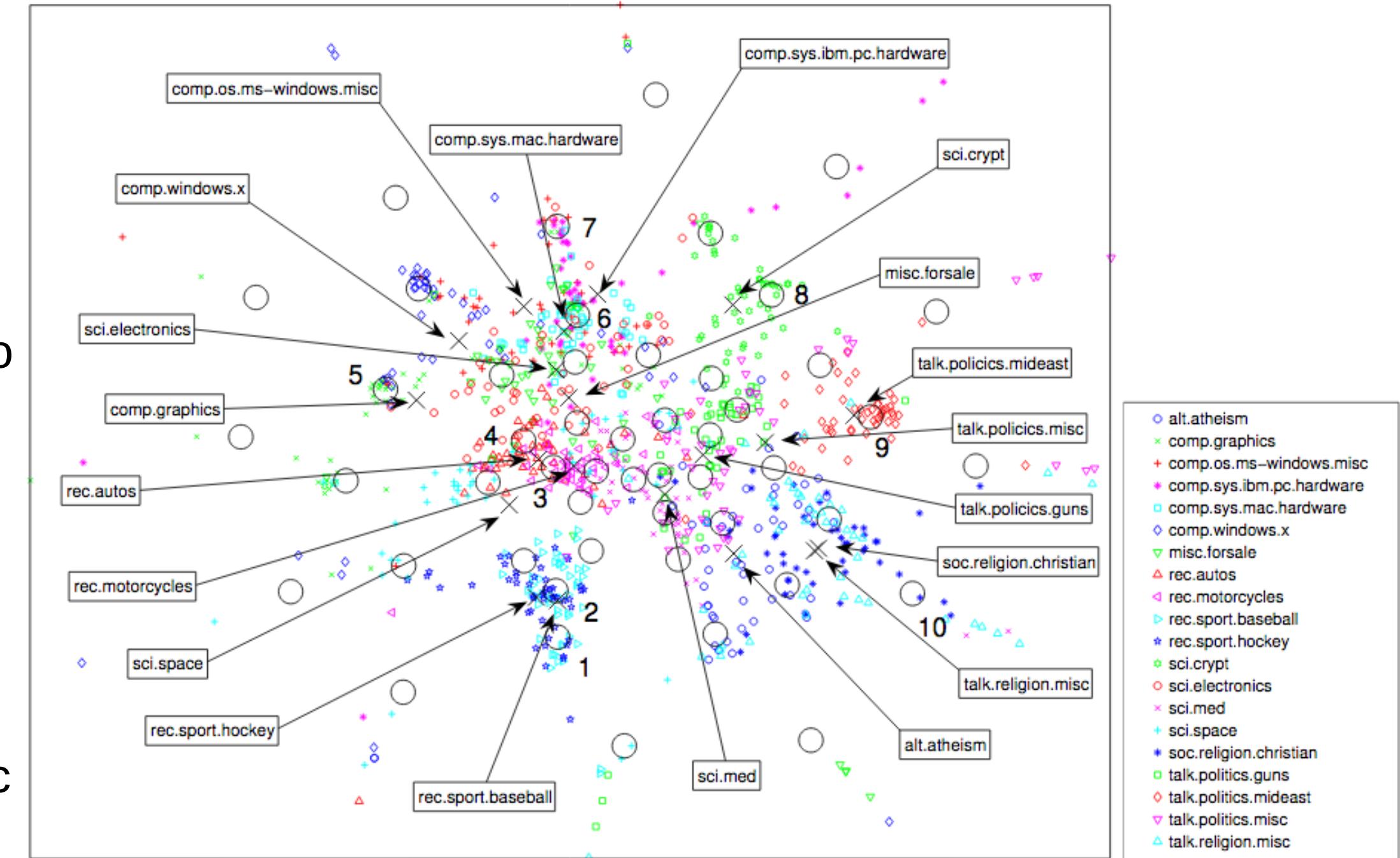
Group-in-a-box layout for topic models (Smith et al., 2014): boxes organized by connectivity (most connected in center), graph per topic shows term co-occurrence



Picture from: A. Smith, J. Chuang, Y. Hu, J. Boyd-Graber, L. Findlater. Concurrent Visualization of Relationships between Words and Topics in Topic Models. In Workshop on Interactive Language Learning, Visualization, and Interfaces, 2014.

Topic model visualization

- Probabilistic Latent Semantic Visualization (Iwata et al., KDD'08): topic model extended to have document and topic coordinates, topic probabilities depend on closeness of document coordinate to topic coordinate.



Picture from: T. Iwata, T. Yamada, N. Ueda. Probabilistic Latent Semantic Visualization: Topic Model for Visualizing Documents. In KDD'08, ACM, 2008.



Suomi24

- One of Finland's most popular message forums
 - 18 years of discussion 2001-2018
 - 2434 conversation sections
 - Over 5M threads, 16M usernames
 - Administrator created sections do not describe true content variation
 - What topical content exists, and how does it vary over the forum hierarchy?

Katkerus
Luottamus
Intohimo
Ikävä
Häpeä
Järki ja tuntee
Onnellisuus
Kateus
Viha
Haluttomus
Rakkaus ja rakastaminen
Mustasukkaisuus
Ihastuminen

Avioehdot
Hääkampaus
Sulhanen
Häälahjat
Häämatka ja hääyö
Vihkiminen
Pukeutuminen
Hääjuhla
Kaaso ja bestman
Sormukset
Kosiminen ja kihlaus
Peittarit
Häävalmisteet

SUOMI24 Etusivu Keskustelu Treffit Posti Chat Alennuskoodit Suomi24 yrityksille Opaus KIRJAUDU

Etsi keskustelusta □

keskustelu²⁴

Keskustelu24 □ Muoti ja kauneus □ Miesten muoti □ Farkkuhaalarit

Farkkuhaalarit

Haalarirakkauks 23.7.2014 18:24

Tuo hauska ja ihana vaate tai joidenkin mielestä vihattu vaate eli farkkulappuhaalarit. Nykyisin enää näkee harvojen naisten ja varsinkaan miesten pällä farkkuhaalarita. Itse olen n. 30 v. nainen ja pidän joskus pälläni farkkuhaalarita. Nyt onnistuin bongaamaan miehen pällä siniset farkkuhaalarit lauantaina 19.7.2014 Sijian Symphony laivalla Tukholman risteilyllä ja kuin sattumaa niin samaisen miehen huomasin tiistaina illalla 22.7.2014 Helsingissä Linnanmäellä farkkuhaalarareissa kävelemässä.

Tätyy kyllä todeta, että ihastuin tähän mieheen farkkuhaalareissaan. Hän näytti aikas söttiltä ja hausalta mieheltä. Ehkä hän asuu kenties pääkaupunkiseudulla. Jos tämä mies tunnistaa itsensä näistä paikoista, niin voisi läittää tälle palstalle viestiä, jos näkisimme toisemme ja tietysti farkkuhaalareissa.

Toivoisin näkeväni enemmänkin samanhenkisiä ihmisiä sekä naisia että miehiä farkkuhaalarit pällä ja samalla voitaisiin kokoontua johonkin farkkuhaalareissa. Farkkuhaalarit on kivat ja rennot pitää pällä sekä ihanat miestenkin pällä. Mitä mieltä muut olette farkkuhaalareista? Laittakaan kommenttia, niin voitaisiin keskustella farkkuhaalareista.

Jaa Ilmianna

46 Vastausta

asdsda 3.8.2014 09:00

itse mies ja käytän farkkuhaalareita. on shortsihaalareita ja ihan pitkälahkeisia farkkuhaalareita.

Jaa Ilmianna

epäselvä 4.8.2014 15:52

Monesko farkkuhaalariketu tämä on?

Komentoi

1 VASTAUS:

Pitkät housut 4.8.2014 17:57

Ei pysy enää laskuissa mukana, mutta farkkuhaalarit näyttäisi olevan suosittu aihe. Farkkuhaalarit on jes housut.

Komentoi lainaten

Jaa Ilmianna

Haalaritylliä 8.8.2014 23:17

Enpäs ole bongaamasi mies mutta vaatekaapistani löytyy neljät lappuhaalarit :)

Vaikka eivät ole kovin muodissa nykyisin, niin kyllä ne ylläni näkyvät. Milloin missäkin. Ostareilla, ystävien luona vieraillessa, joskus jopa baarissa. Jonkinmoinen fetissi noihin, siksipä innoissaan näistä :)

Komentoi

3 VASTAUSTA:

Lappupöksy 10.8.2014 23:15

Useammat farkkulappuhaalarit minulta löytyy ja vaimolla on muutamat myös. Ihan huippu housut miehille. Sopii myös naisille, jos ovat naisellista mallia.

Komentoi lainaten

Jaa Ilmianna

henkselman 11.8.2014 0:13

Lappupöksy kirjoitti:

Useammat farkkulappuhaalarit minulta löytyy ja vaimolla on muutamat myös. Ihan huippu housut mi... Näytä lisää

E-kontakti.fi

25 v. Pohjanmaa 31 v. Uusimaa 26 v. Uusimaa

40 v. Uusimaa 49 v. Uusimaa 38 v. Pirkanmaa

LÖYDÄ SEURAA TOSITARKOITUKSELLA!

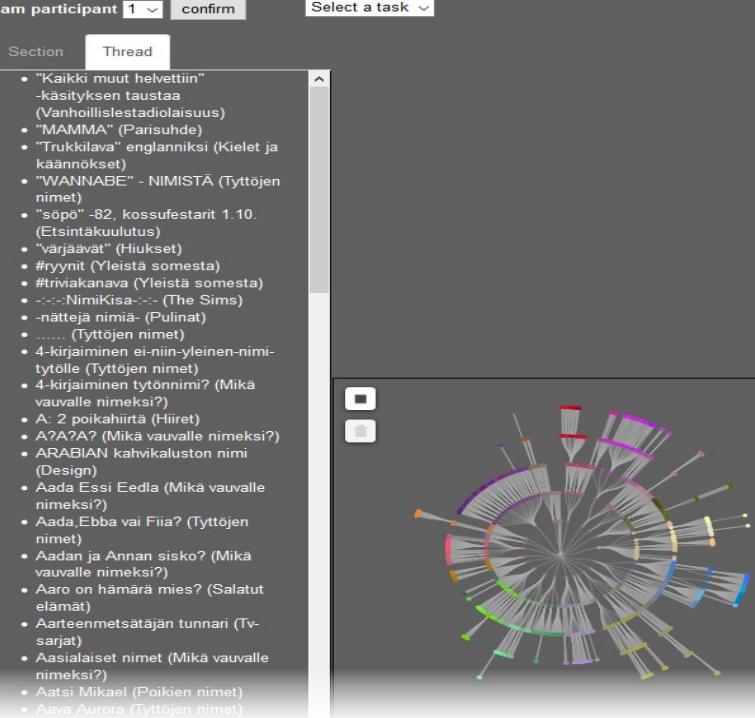
Haen naista: 25-39 v. 40-54 v. 55+v.

Haen miestä: 25-39 v. 40-54 v. 55+v.

Nyt voi yrityksesi näkyä tassä! Alkaen 199€/kk

Aller media Lisätieto Klikkaa tästä >

ADON news Klikkaa tästä

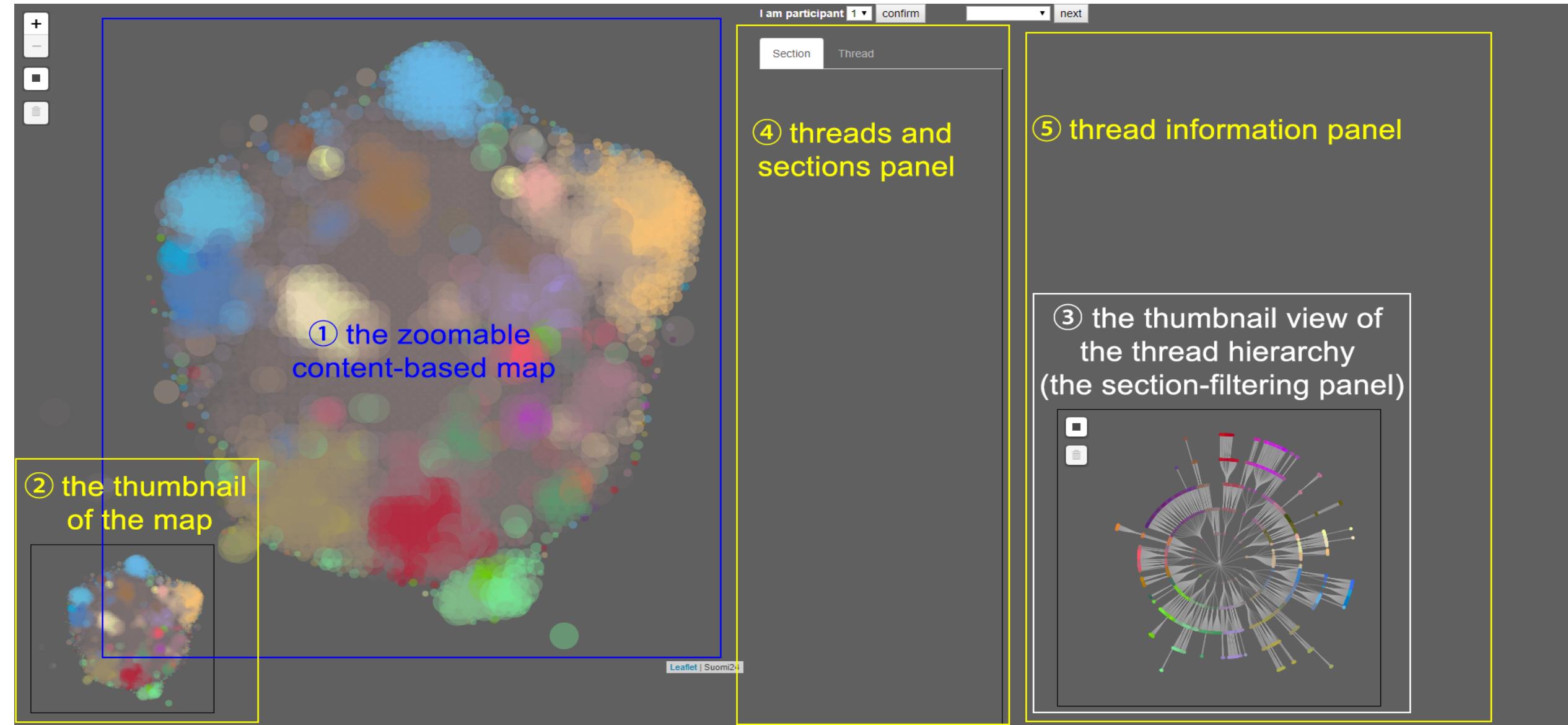


- PIHVI: interactive system for visualizing and exploring a large **hierarchical text corpus of online forum postings**.
- The main view shows a **large-scale scatter plot**, created by flexible nonlinear dimensionality reduction based on text contents of the postings.
- We couple it with a **coloring optimized to represent the forum hierarchy** by a second dimensionality reduction.

Pictures from: J. Peltonen, Z. Lin, K. Järvelin, J. Nummenmaa. PIHVI: Online Forum Posting Analysis with Interactive Hierarchical Visualization. In ESIDA 2018.

PIHVI interface

Pictures from: J. Peltonen, Z. Lin, K. Järvelin, J. Nummenmaa. PIHVI: Online Forum Posting Analysis with Interactive Hierarchical Visualization. In ESIDA 2018.



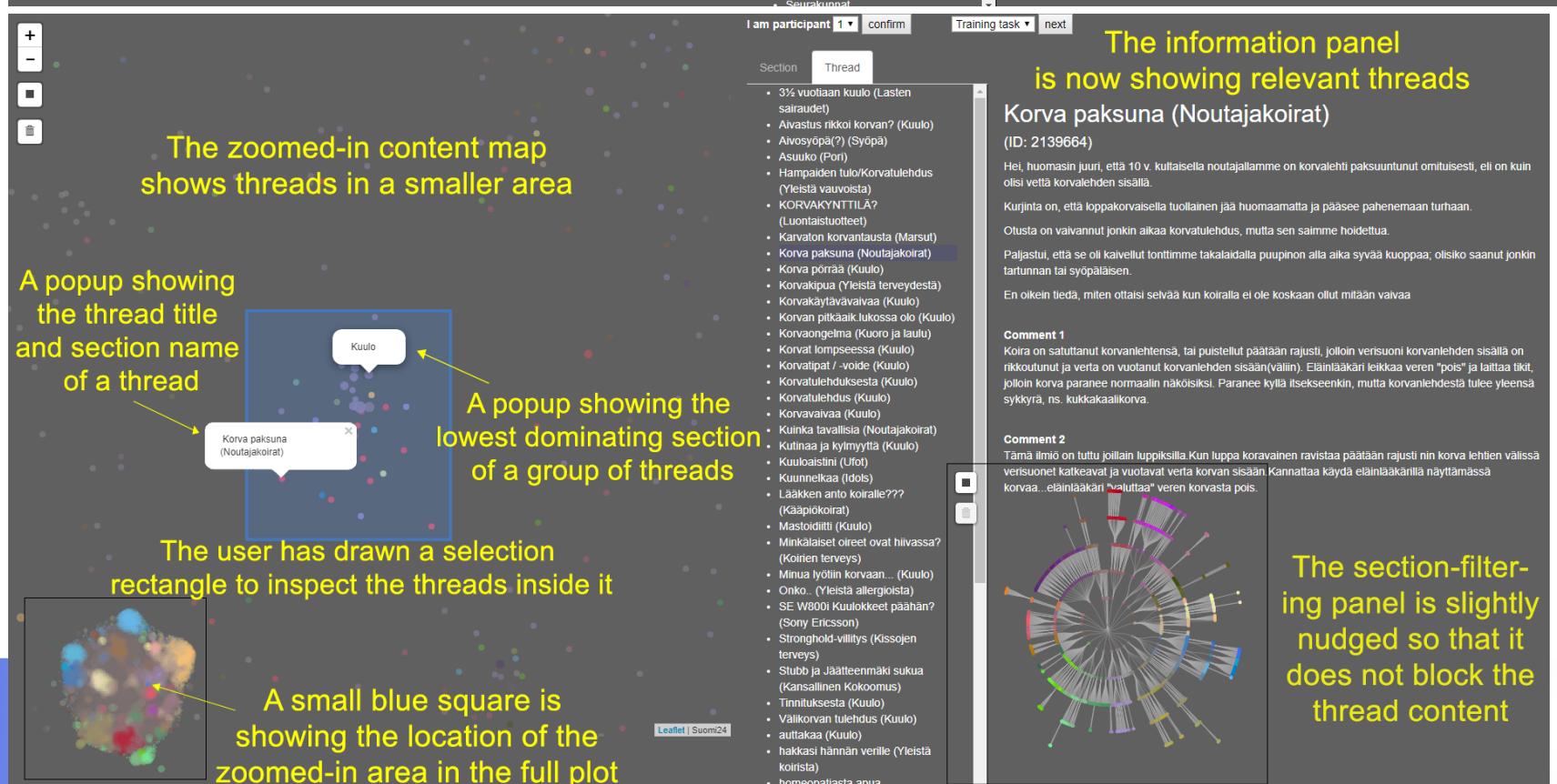
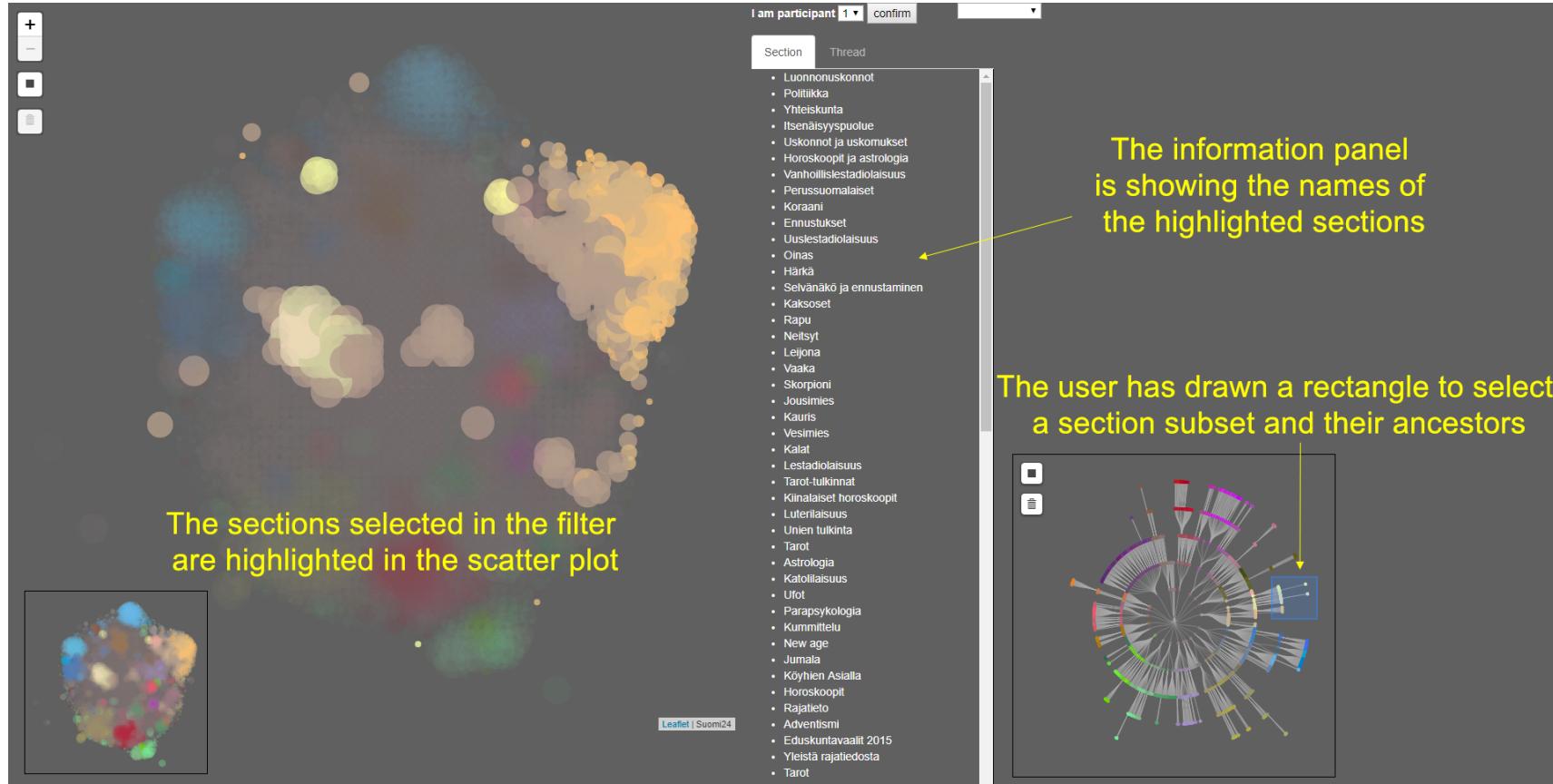
Five linked views

1. Content-based map = Interactive scatter plot of the thread collection, created by dimensionality reduction.
Threads **organized by content similarity**: similar threads are shown nearby
3. Section hierarchy graph: **Colors represent section similarity**, nearby sections have similar colors.
Colors linked to content plot.

Filtering content by section

Zooming,
Selecting content by similarity,
Content details on demand

Pictures from: J. Peltonen, Z. Lin, K. Järvelin, J. Nummenmaa. PIHVI:
Online Forum Posting Analysis with Interactive Hierarchical
Visualization. In ESIDA 2018.

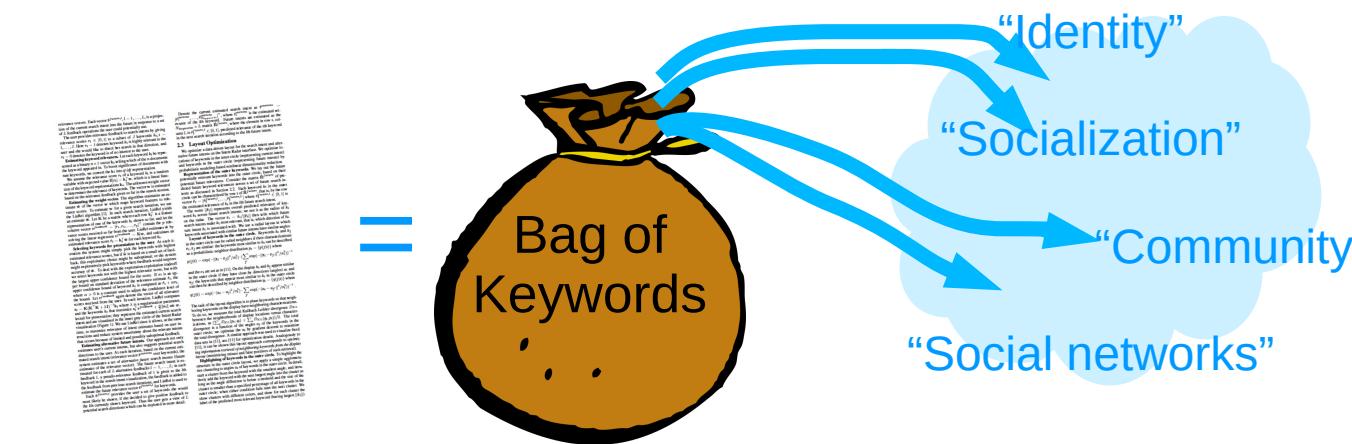


Information retrieval

- Rank candidate documents by how well they match the query phrase.
- Language model approach:
 - Query represents a fragment of a desired (ideal) document.
 - Compute probability that each candidate document can produce the query.

- Unigram language model: each document is a multinomial distribution (bag of keywords), document score is

$$p(query|document) = \prod_{term=1}^{N_{vocabulary}} p(term|document)^{count(term|query)}$$



- More advanced models include sentence structure and document connectedness in the ranking
- Interactive methods also include relevance feedback

Interfaces for exploratory search

Facets: filters items by a metadata attribute (Yee et al. 2003)

Picture from: K.-P. Yee, K. Swearingen, K. Li, and M.A. Hearst. Faceted metadata for image search and browsing. In ACM CHI 2003.

Flamenco

Refine your search further within these categories:

Media (group results)
costume (3), drawing (2),
lithograph (1), woodcut (6), woven object (2)

Location: all > Asia
Afghanistan (1), China (4), China or Tibet? (3), India (2), Japan (13), Russia (1), Turkey (3), Turkmenistan (1)

Date (group results)
17th century (3), 18th century (3), 19th century (10), 20th century (3), date ranges spanning multiple centuries (7), date unknown (2)

Themes (group results)
music, writing, and sport (5), nautical (1), religion (2)

Objects (group results)
clothing (5), food (1), furnishings (4), timepieces (1)

Nature (group results)
bodies of water (3), fish (1), flowers (2), geological formations (1), heavens (3), invertebrates and arthropods (1), mammals (2), plant material (3), trees (1)

Places and Spaces (group results)
bridges (1), buildings (1), dwellings (1)

These terms define your current search. Click the to remove a term.

Location: Asia

Shapes, Colors, and Materials: fabrics

Search all items within current results

28 items (grouped by location) [view ungrouped items](#)

Afghanistan 1


Girl's Ceremonial...
no artist
20th century

China 4

 4 boats on lake,...
Anonymous
post World War II

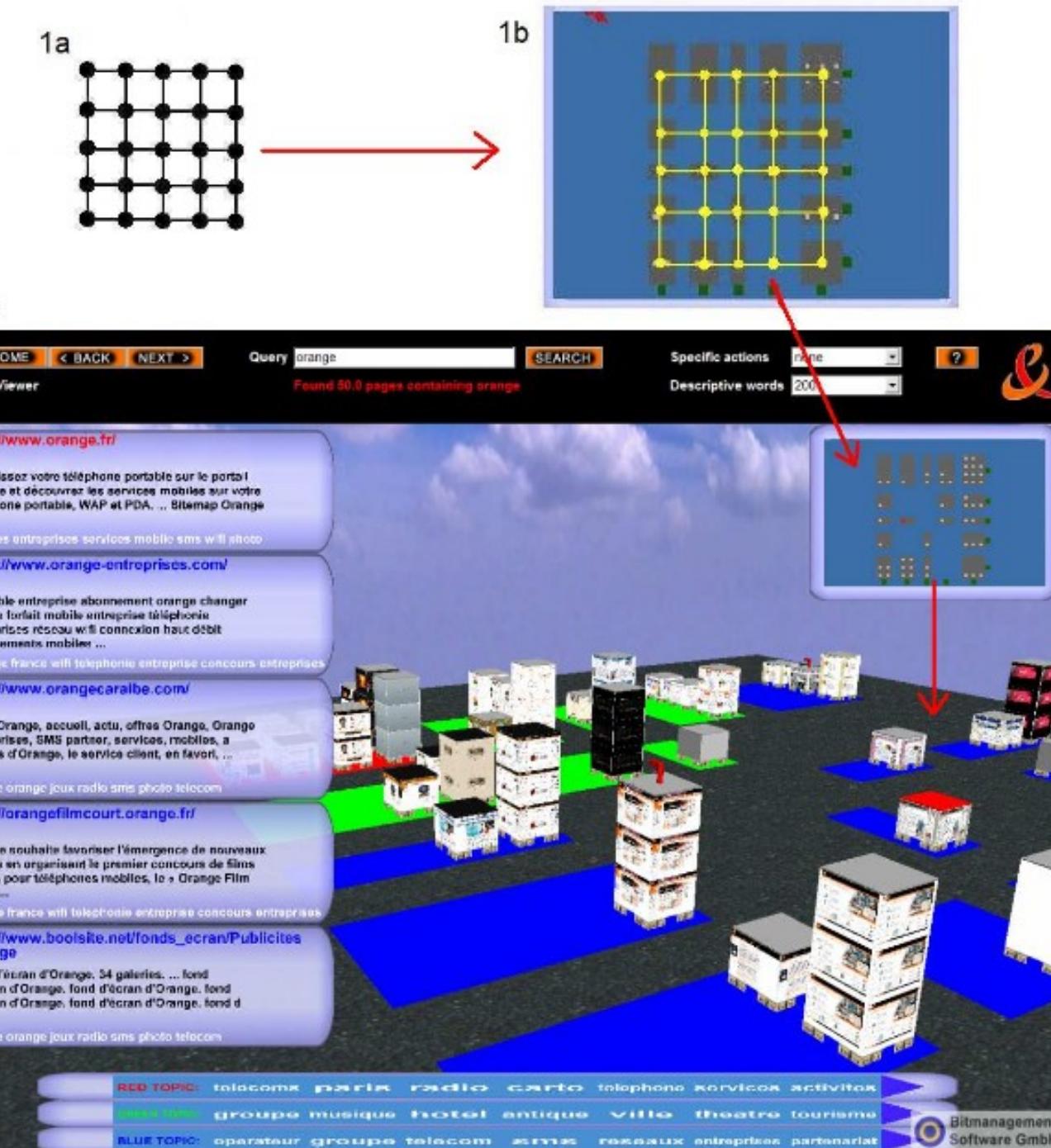
 Embroidery
no artist
19th century

 Embroidery
no artist
19th century

 Embroidery :
no artist
19th century

Interfaces for exploratory search

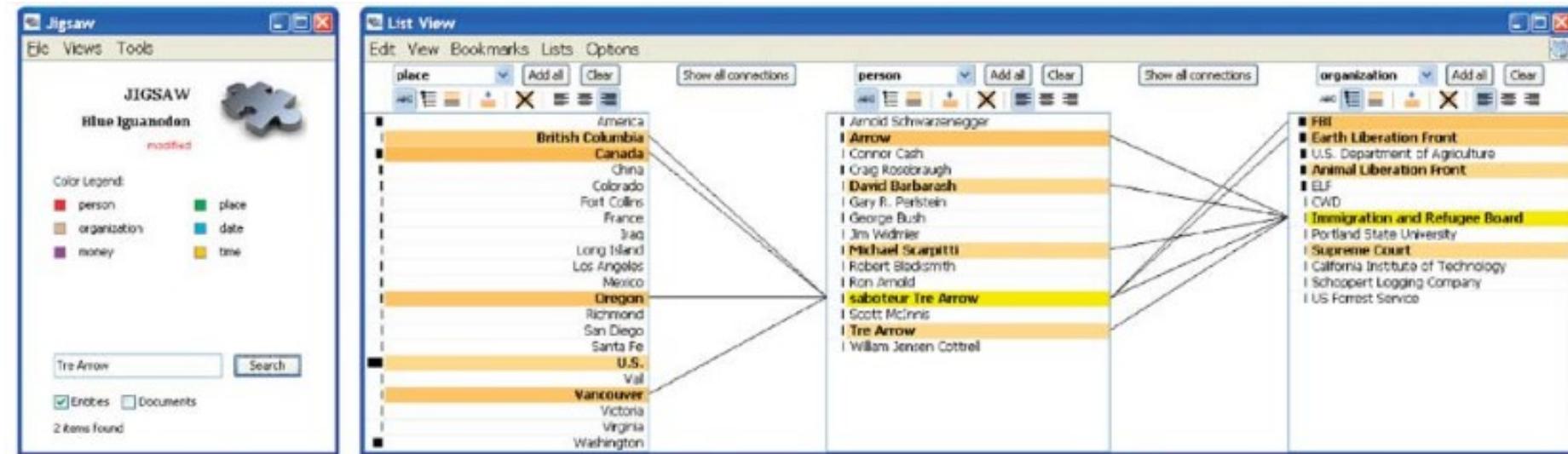
Clusters:
documents in
clusters on a
map (Bonnel et
al. 2006)



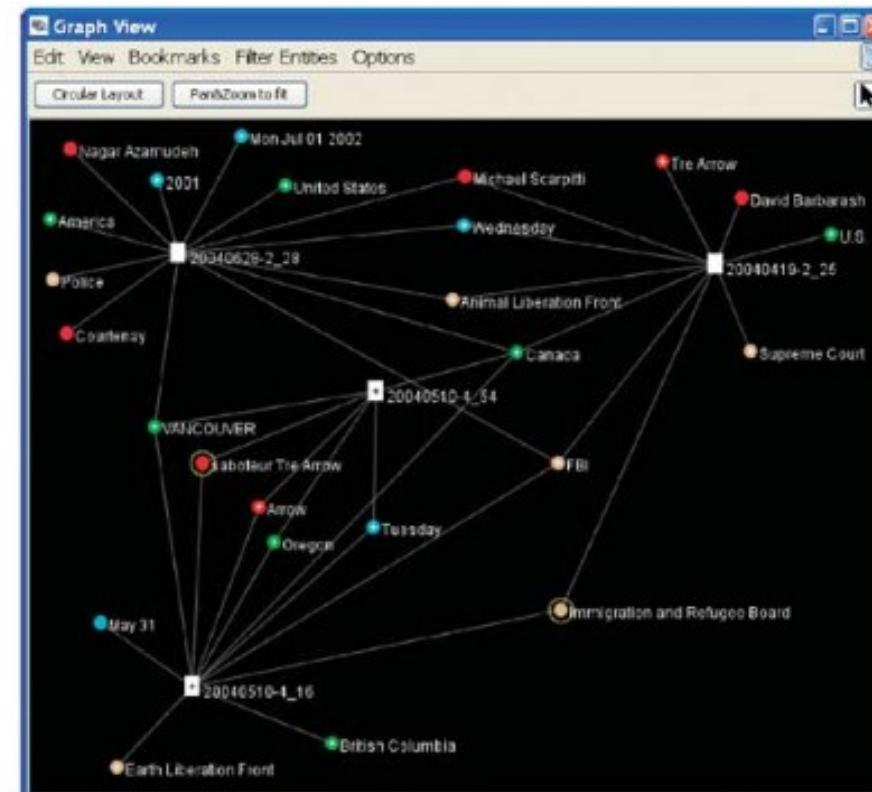
Picture from: N. Bonnel, V. Lemaire, A. Cotarmanac'H, A. Morin. Effective Organization and Visualization of Web Search Results. In EurolMSA'06, IASTED, 2006.

Interfaces for exploratory search

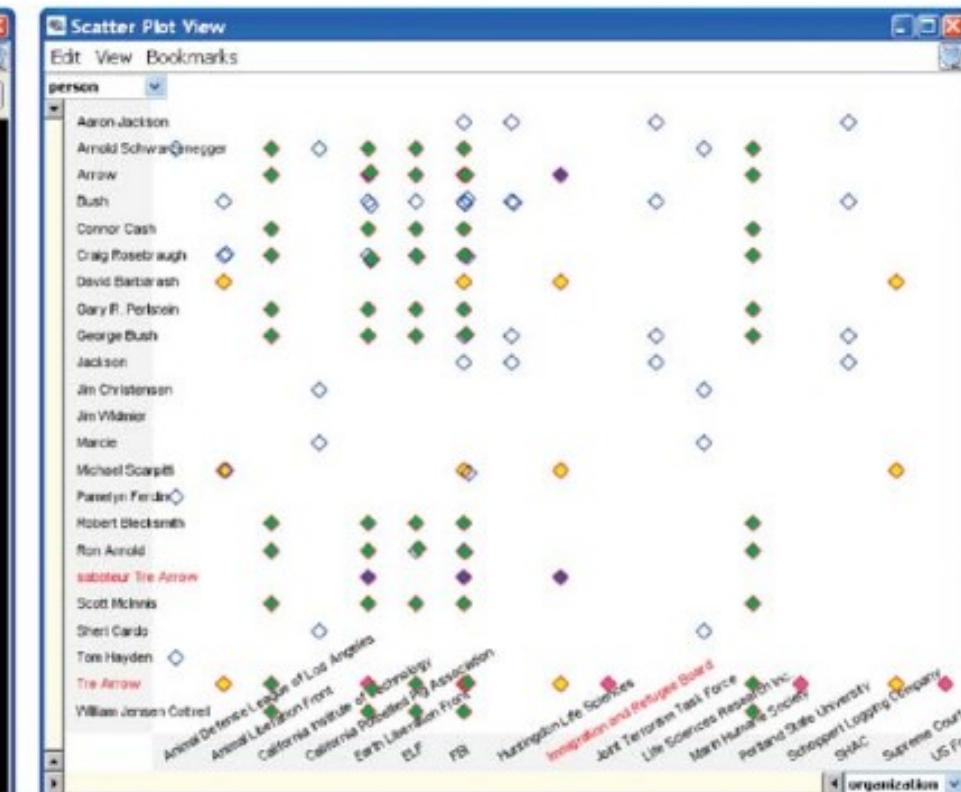
Jigsaw: interface with views such as document-entity graphs (Stasko et al. 2008)



C



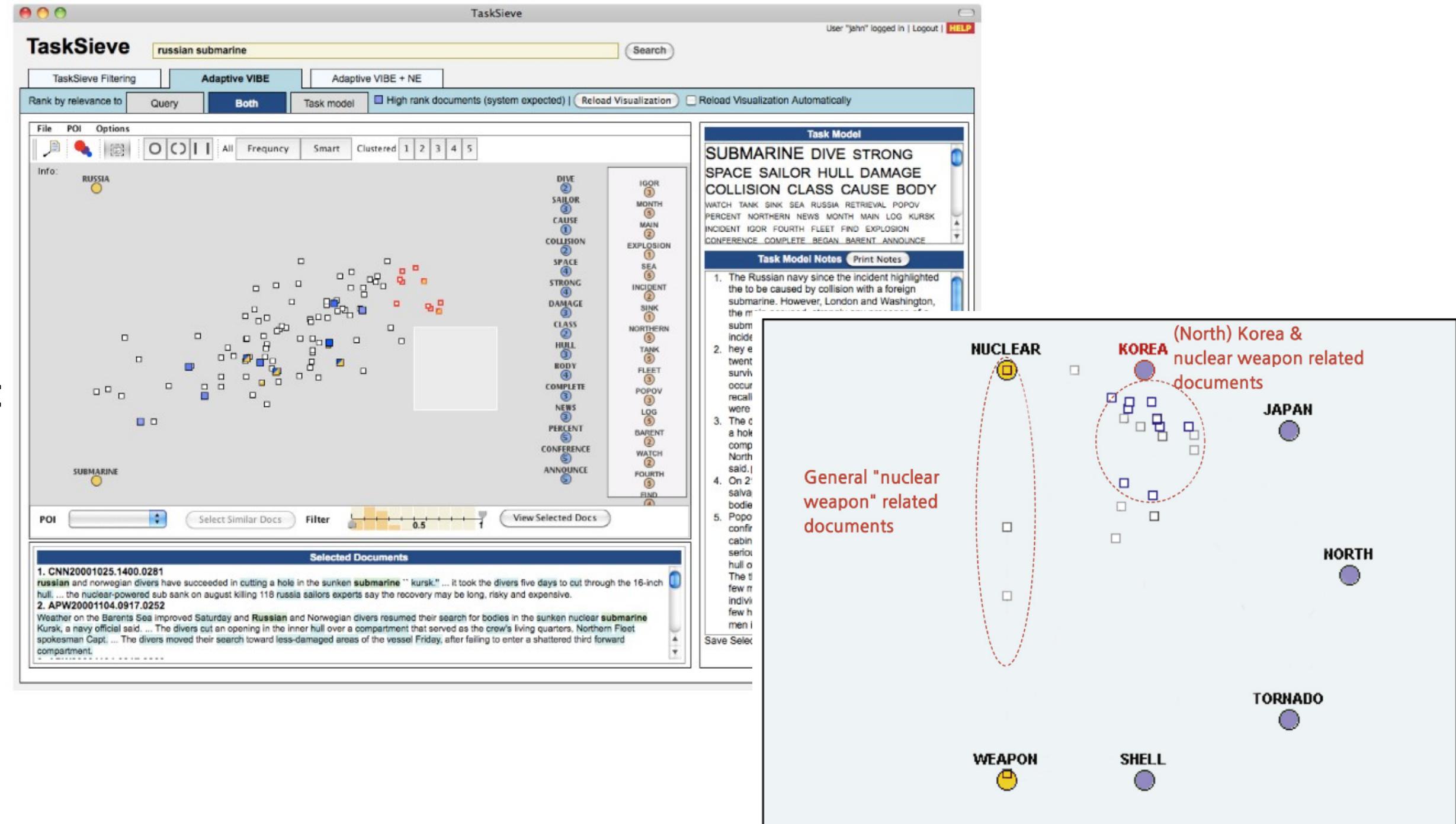
D



Picture from: J. Stasko, C. Görg, and Z. Liu. Jigsaw: Supporting investigative analysis through interactive visualization. Information Visualization, 7(2), 118-132, 2008.

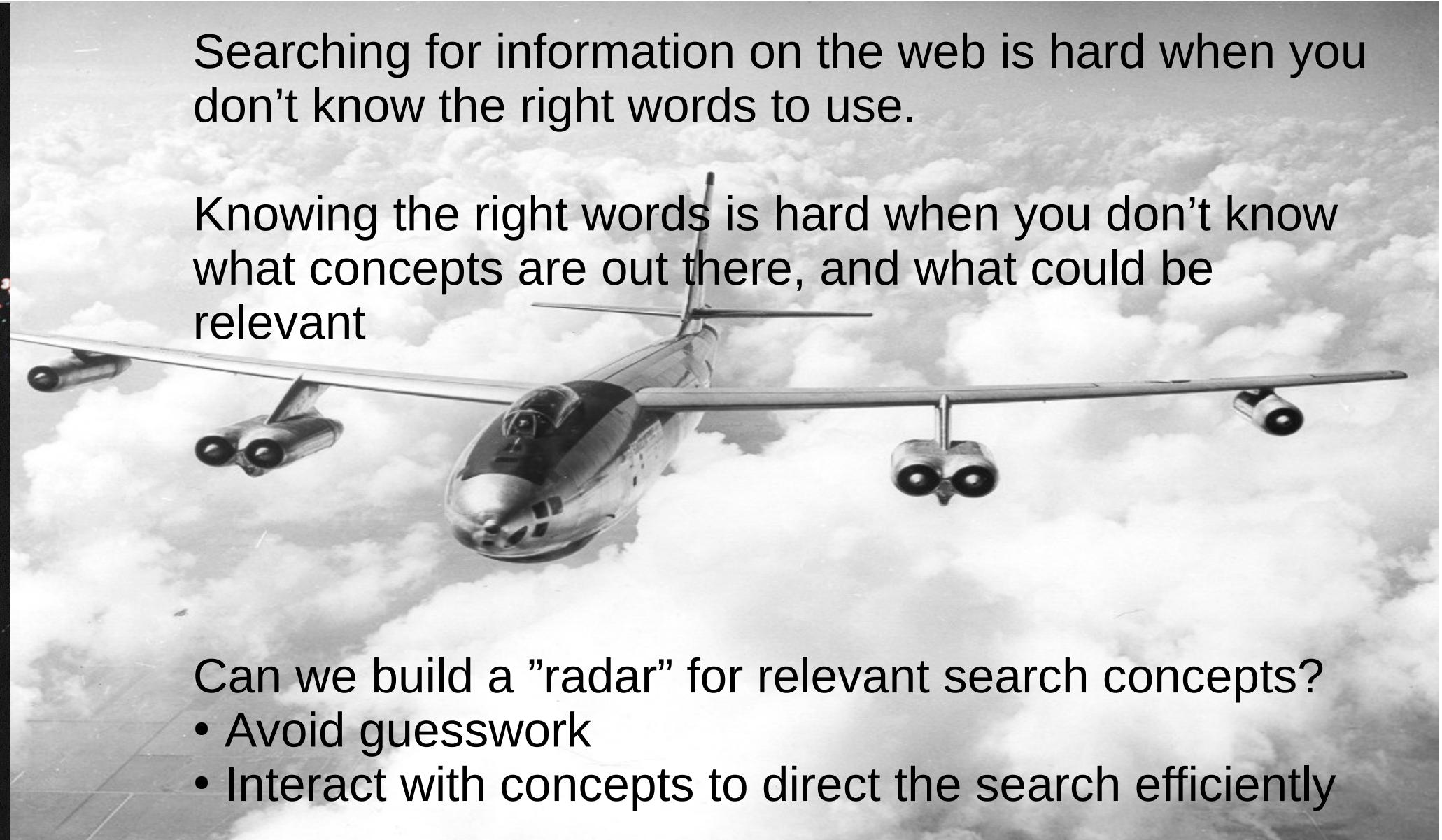
Interfaces for exploratory search

Adaptive VIBE
(Ahn and Brusilovsky 2013): interface arranging documents by similarity to reference points:
1) query terms,
2) terms from a user model



Pictures from: J.-w. Ahn and P. Brusilovsky. Adaptive visualization for exploratory information retrieval. Information Processing and Management 49:1139-1164, 2013.

SciNet: Dimensionality reduction for the search information space



Searching for information on the web is hard when you don't know the right words to use.

Knowing the right words is hard when you don't know what concepts are out there, and what could be relevant

Can we build a "radar" for relevant search concepts?

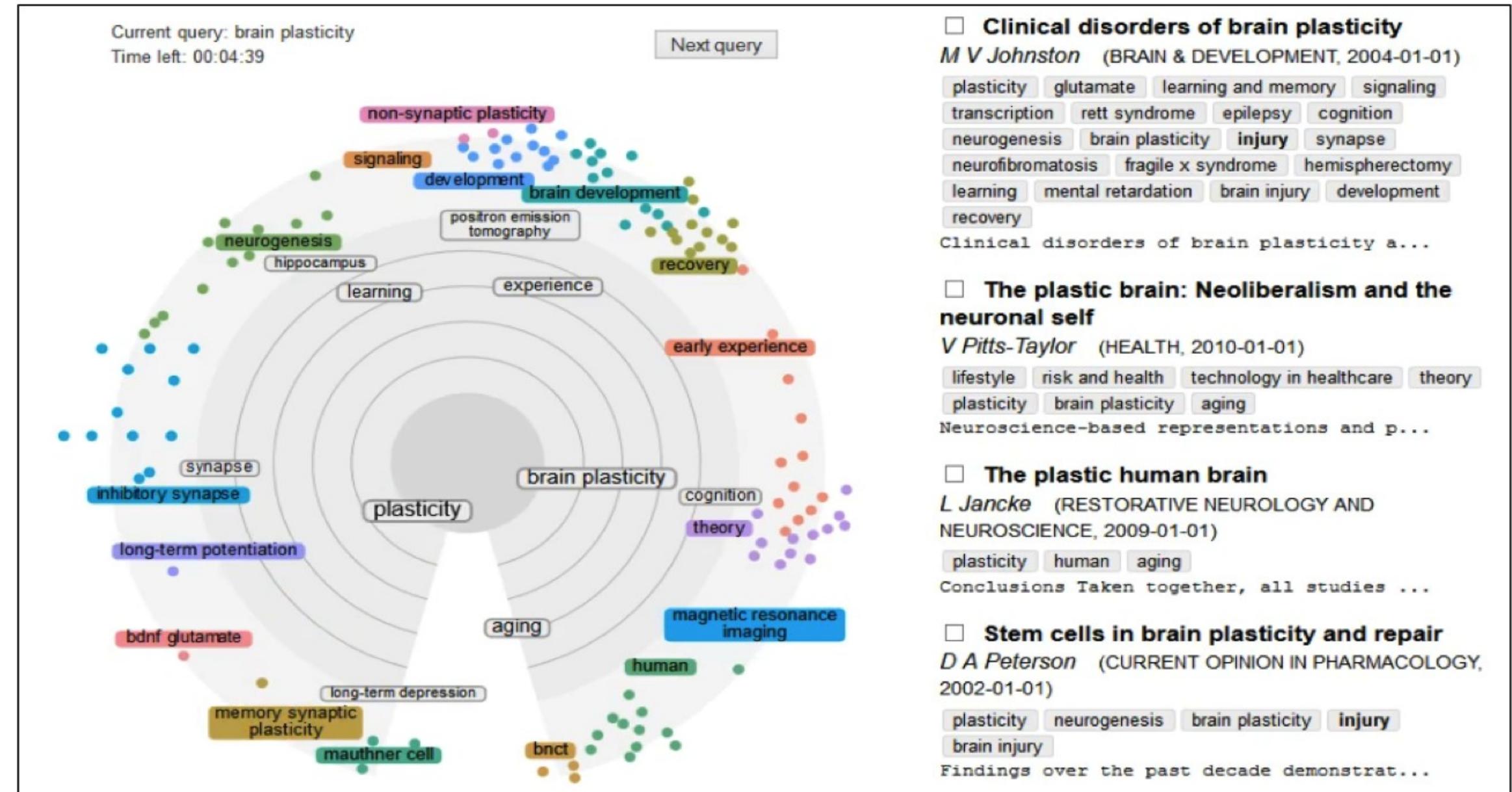
- Avoid guesswork
- Interact with concepts to direct the search efficiently

SciNet: Dimensionality reduction for the search information space

keywords = concepts

radius = relevance
(predicted from document content and relevance feedback on keywords)

angles = dimensionality reduction result, keywords that respond similarly to feedback get similar angles

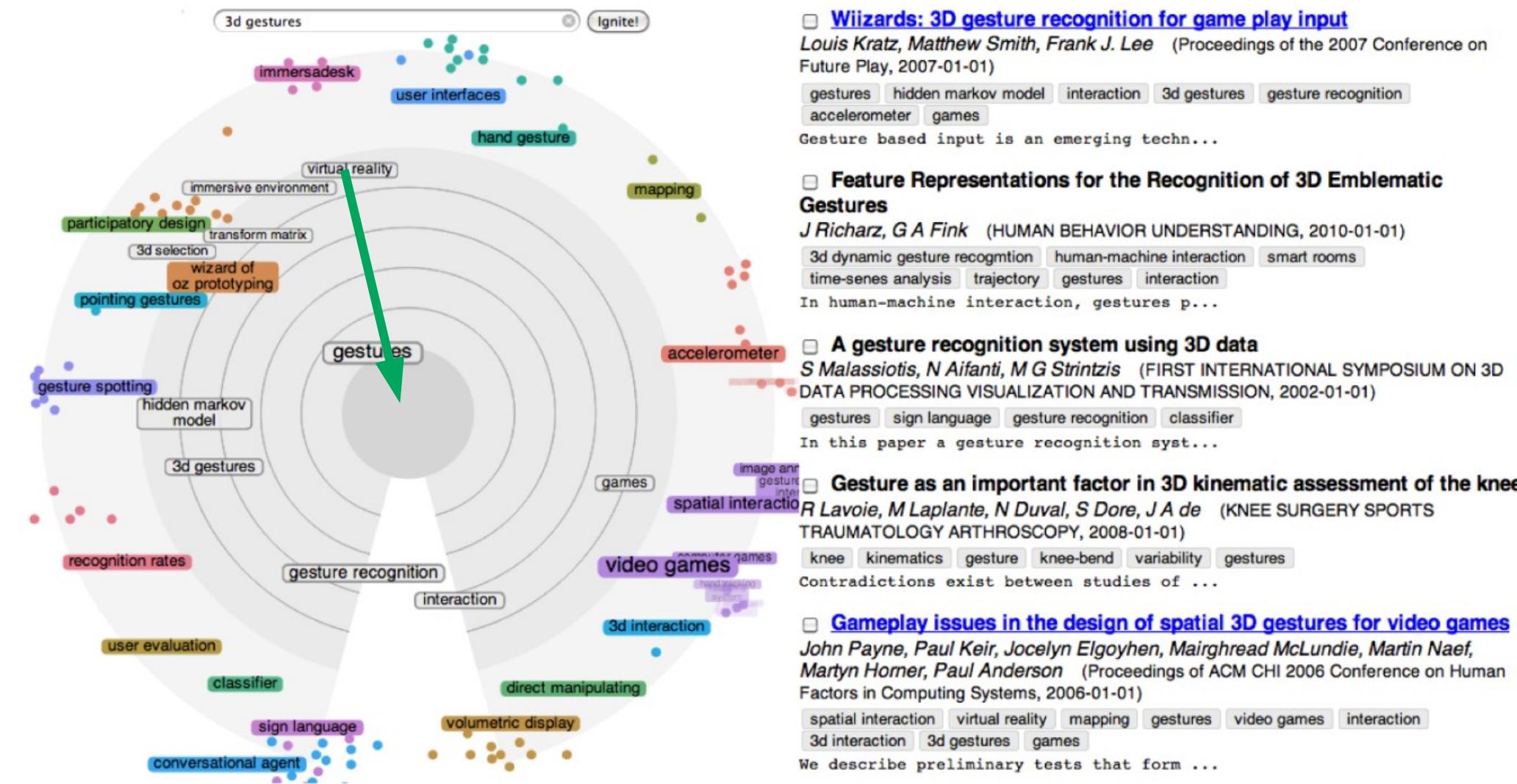


References:

- T. Ruotsalo, J. Peltonen, M.J. A. Eugster, D. Glowacka, P. Floréen, P. Myllymäki, G. Jacucci, S. Kaski. Interactive Intent Modeling for Exploratory Search. ACM TOIS, 2018
- J. Peltonen, K. Belorustceva, and T. Ruotsalo. Topic-Relevance Map: Visualization for Improving Search Result Comprehension. In IUI 2017.
- T. Ruotsalo, J. Peltonen, M. Eugster, D. Glowacka, K. Konyushkova, K. Athukorala, I. Kosunen, A. Reijonen, P. Myllymäki, G. Jacucci, S. Kaski. Directing Exploratory Search with Interactive Intent Modeling. In CIKM 2013.

SciNet: Dimensionality reduction for the search information space

The user can give **feedback** by dragging concepts towards the center



SciNet clearly improves performance in exploratory search:

- Users see more relevant content during search
- Users direct search better --> better essay answers
- Users comprehend the search space better, in information comprehension experiments

References: T. Ruotsalo, J. Peltonen, M.J. A. Eugster, D. Glowacka, P. Floréen, P. Myllymäki, G. Jacucci, and S. Kaski. Interactive Intent Modeling for Exploratory Search. ACM Transactions on Information Systems, 36(4), article 44, October 2018.

J. Peltonen, J. Strahl, and P. Floreen. Negative Relevance Feedback for Exploratory Search with Visual Interactive Intent Modeling. In IUI 2017.

J. Peltonen, K. Belorustceva, and T. Ruotsalo. Topic-Relevance Map: Visualization for Improving Search Result Comprehension. In IUI 2017.