ASV Automatische Sprachverarbeitung

UNIVERSITÄT LEIPZIG

Exploratory Search Through Visual Analysis of Topic Models

- Göttingen Dialog in Digital Humanities -

Patrick Jähnichen, Patrick Oesterling, Tom Liebmann, Gerhard Heyer, Gerik Scheuermann and Christof Kuras

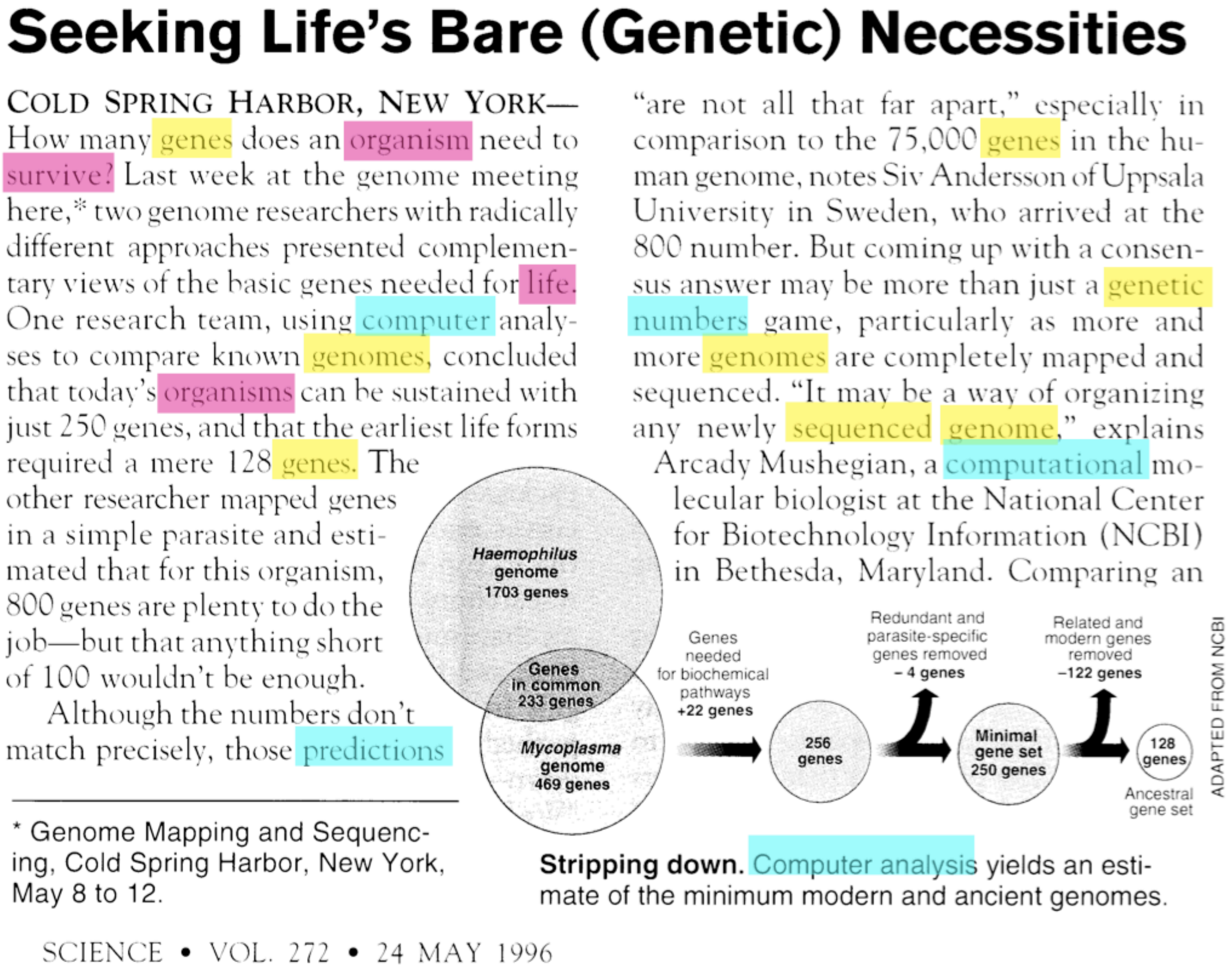
Agenda

- Topic Models (Latent Dirichlet Allocation)
- Exploratory Search and Visual Analysis
  - Task 1 – Inspecting a topic
  - Task 2 – Overview over topics
  - Task 3 – Ambiguity Resolution
  - Task 4 – Document retrieval
  - Task 5 – Finding topics transitively

Topic Models

- Who has heard of topic models?
- Who knows what they do?
- Who knows how they do it?

Topic Models



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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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

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

SCIENCE • VOL. 272 • 24 MAY 1996

[5], p. 19

Topic Models

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

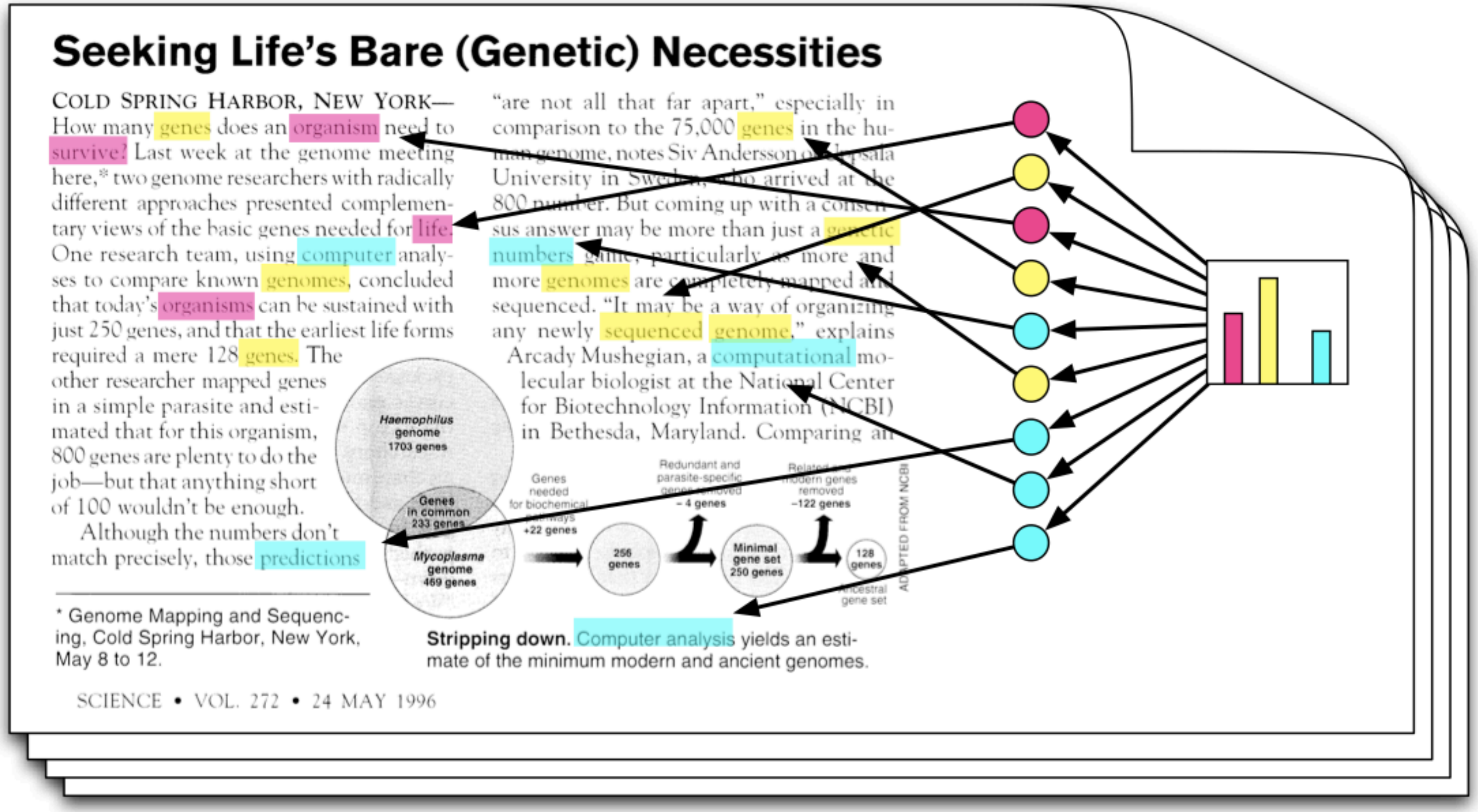
life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Topic proportions and assignments



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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed –4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

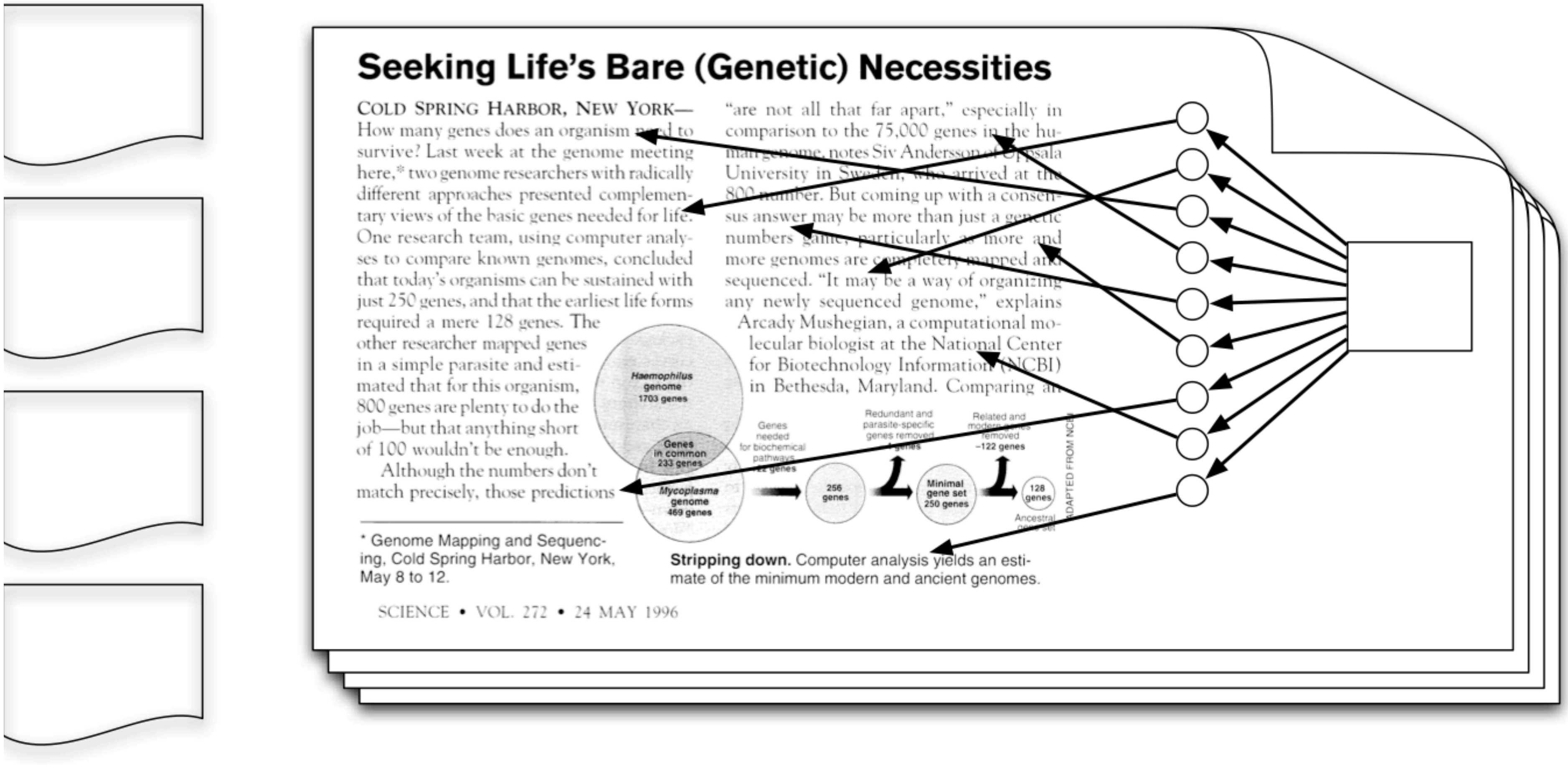
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

[5], p. 20

Topic Models



Topics

Documents

Topic proportions and assignments

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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed -4 genes
Minimal gene set 250 genes
Related and modern genes removed -122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

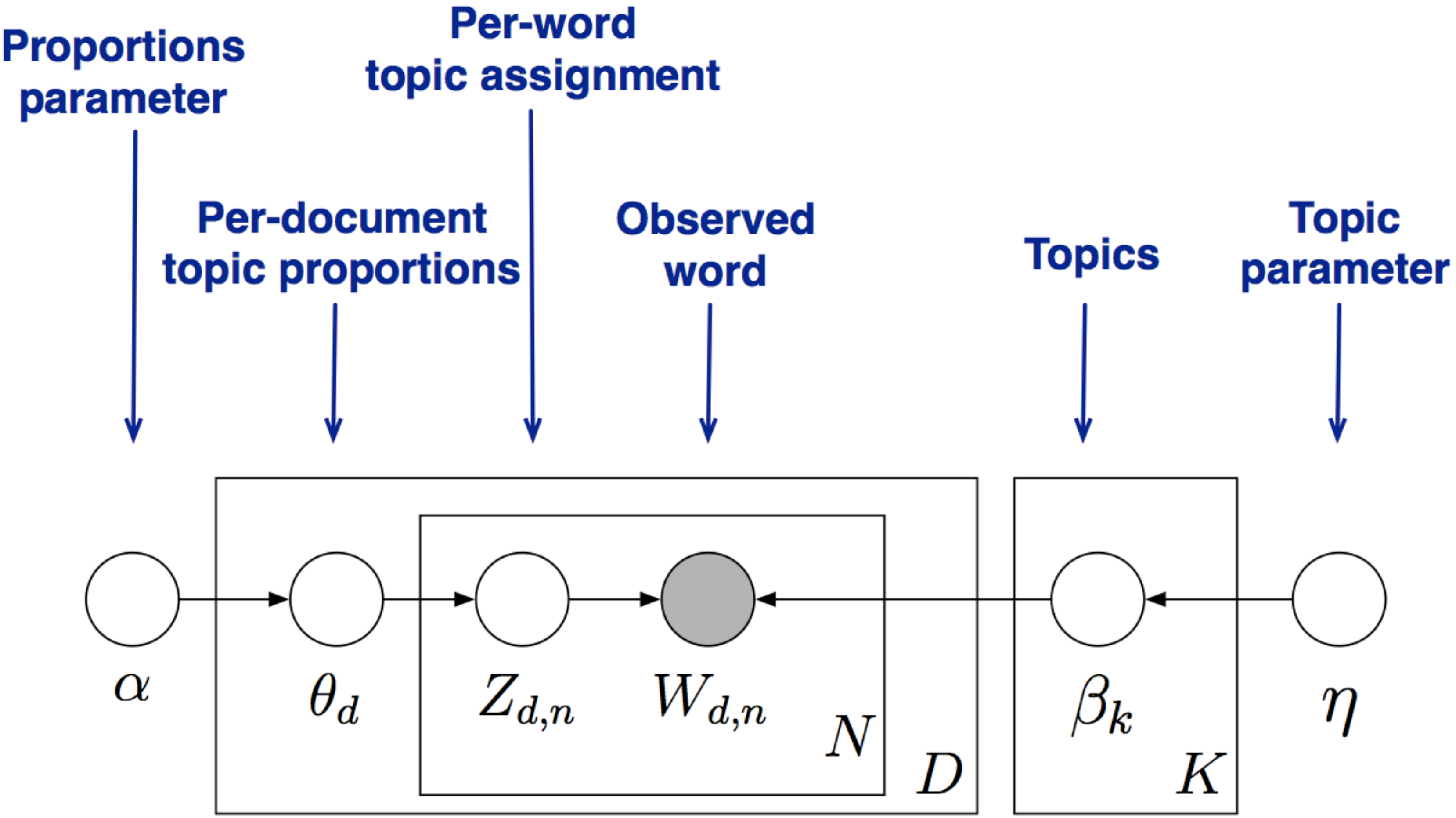
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

[5], p. 21

Topic Models



Proportions parameter

Per-word topic assignment

Per-document topic proportions

Observed word

Topics

Topic parameter

[5], p. 23

Topic Models

- Bayes‘ law

$$p(\text{parameters}|\text{data}) = \frac{p(\text{data}|\text{parameters})p(\text{parameters})}{p(\text{data})}$$

- parameters: the documents‘ topic proportions, the topics, the word topic assignments
- data: the words

Topic Models

$$p(\beta,\theta,z,w)=\prod_{k=1}^{K}p(\beta_k|\eta)\prod_{d=1}^{D}\left\{p(\theta_d|\alpha)\prod_{n=1}^{N_d}p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta_{1:K})\right\}$$

- easy-to-read-off joint probability
- (don‘t worry too much about it)
- helps to determine p(w) because

$$p(w)=\int_{\beta}\int_{\theta}\int_{z}p(\beta,\theta,z,w)$$

Topic Models

- finding p(w) is intractable
- learn an approximation: 2 main methods
- Sampling
  - Initialize randomly
  - Repeatedly reassign each word to a topic (conditioned on all other assignments)
  - measure the likelihood each time
- Variational inference
  - Find probable solution by optimization

Topic Models

- example analysis of some of our wortschatz data (http://wortschatz.uni-leipzig.de)
- 100 million sentences from 2010
- the following results are taken from an LDA model with 100 topics

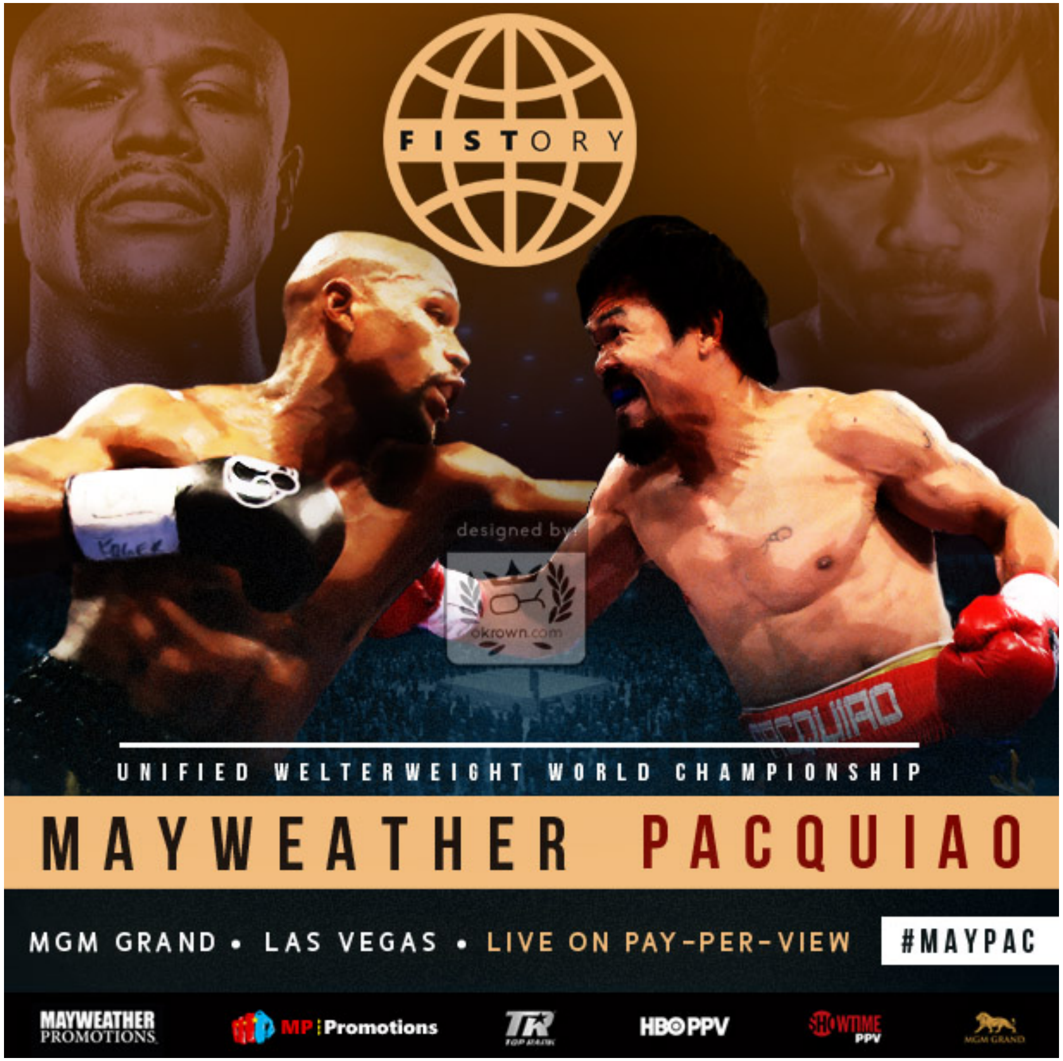
Topic Models

people
haiti
air
flight
day
hours
airport
weather
morning
port
night
earthquake
early
thursday
fire
area
officials

“Peacekeeping – MINUSTAH“ by Marco Domino/The United Nations Development Programme , as used in [1], CC BY 2.0

Topic Models

fight
vegas
las
round
ring
boxing
match
world
wrestling
champion
fighter
ufc
pacquiao
title
continues
mayweather
fights



“Floyd Mayweather vs Manny Pacquiao“ by oDOTkrown [2], CC BY-ND 3.0

Topic Models

oil
bp
gulf
spill
gas
company
coast
water
mexico
drilling
million
day
disaster
louisiana
barrels
damage
hurricane

“Deepwater Horizon offshore drilling unit on fire“, US Coast Guard [3], CC0, Public Domain

Topic Models

space
nasa
station
shuttle
program
earth
planet
center
rocket
moon
international
mission
launch
mars
star
search
science

Space Shuttle launch [4], CC0 Public Domain

Exploratory Search

- only looked at topics by now
- what about those small bar charts
- can use information about topics in documents *and* about words in topics

Exploratory Search

- define tasks for exploratory search
    - inspect a topic
    - get an overview over all of them
    - find words that appear in multiple topics (i.e. have different meanings)
    - get the documents that talk about a topic
    - having a document, find other related ones (that have similar topic distributions)

Exploratory Search

- Task 1 – Inspecting a topic
  - difficult, a topic is a probability distribution
  - problem 1: find prominent words in the topic
    - this is easy: sort them by probability is one option
  - problem 2 – find overall topic significance
    - use information on topics in documents
    - compute the average usage of a topic across documents

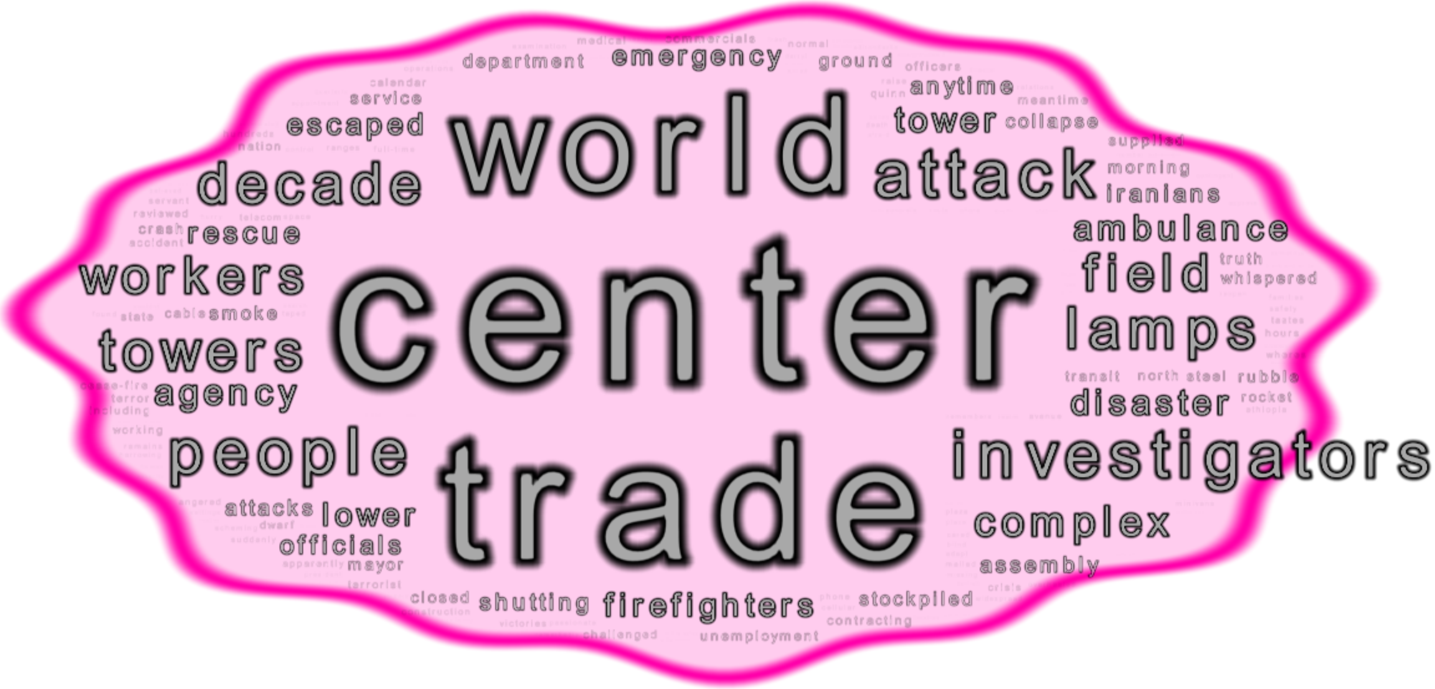
Visual Analysis

- Task 1 – Inspecting a topic
  - topics are presented as word clouds
  - most prominent term in the middle, other follow in a spiral around it (inside -> outside)
  - first portion of list has high probability -> display only those (parameter for minimum probablity)
  - cloud is scaled according to topic relevance

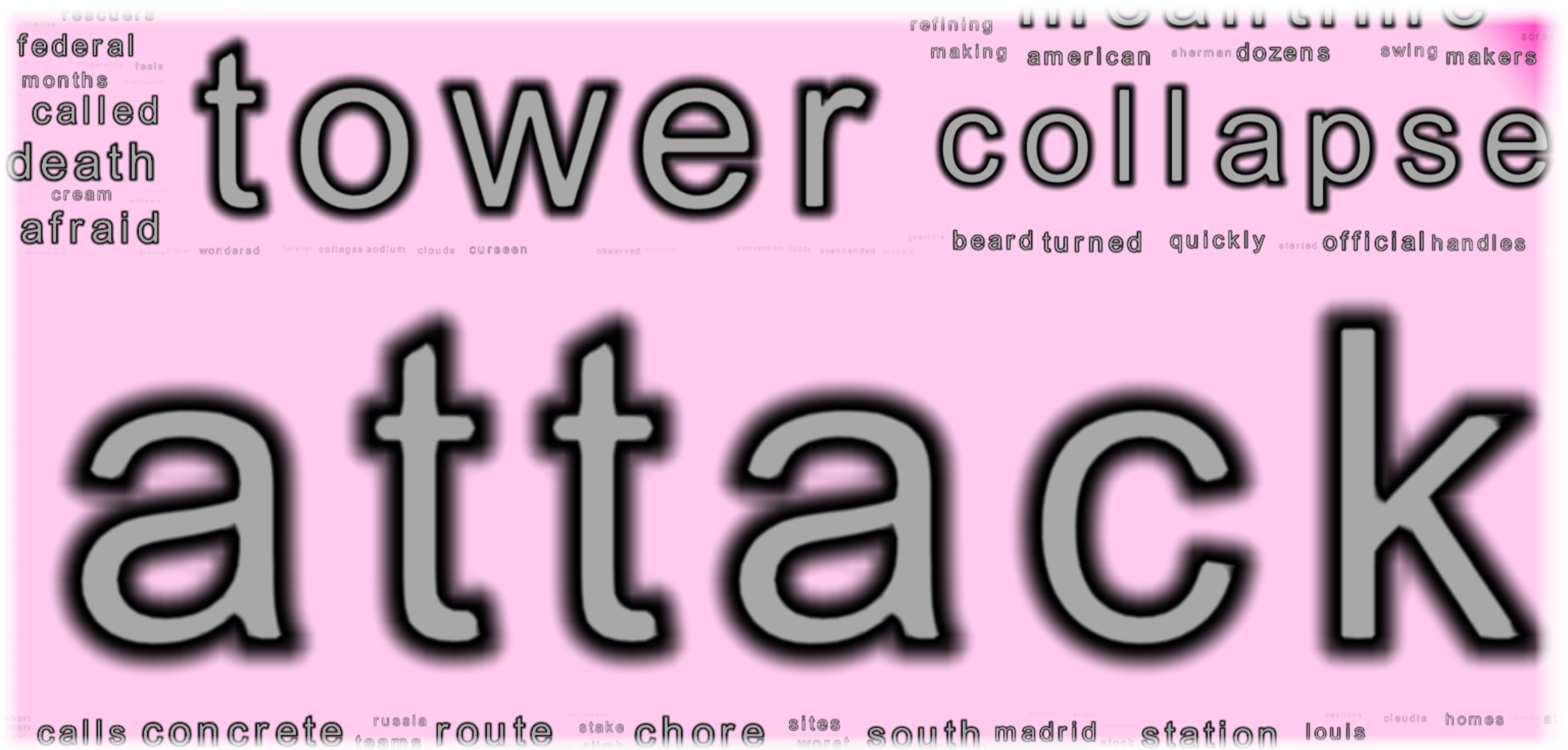
Visual Analysis

emergency
anytime
escaped
world
tower
decade
attack
ambulance
rescue
workers
center
field
lamps
towers
agency
disaster
people
trade
investigator
attacks
lower
complex
officials
shutting
firefighters

Visual Analysis



Visual Analysis



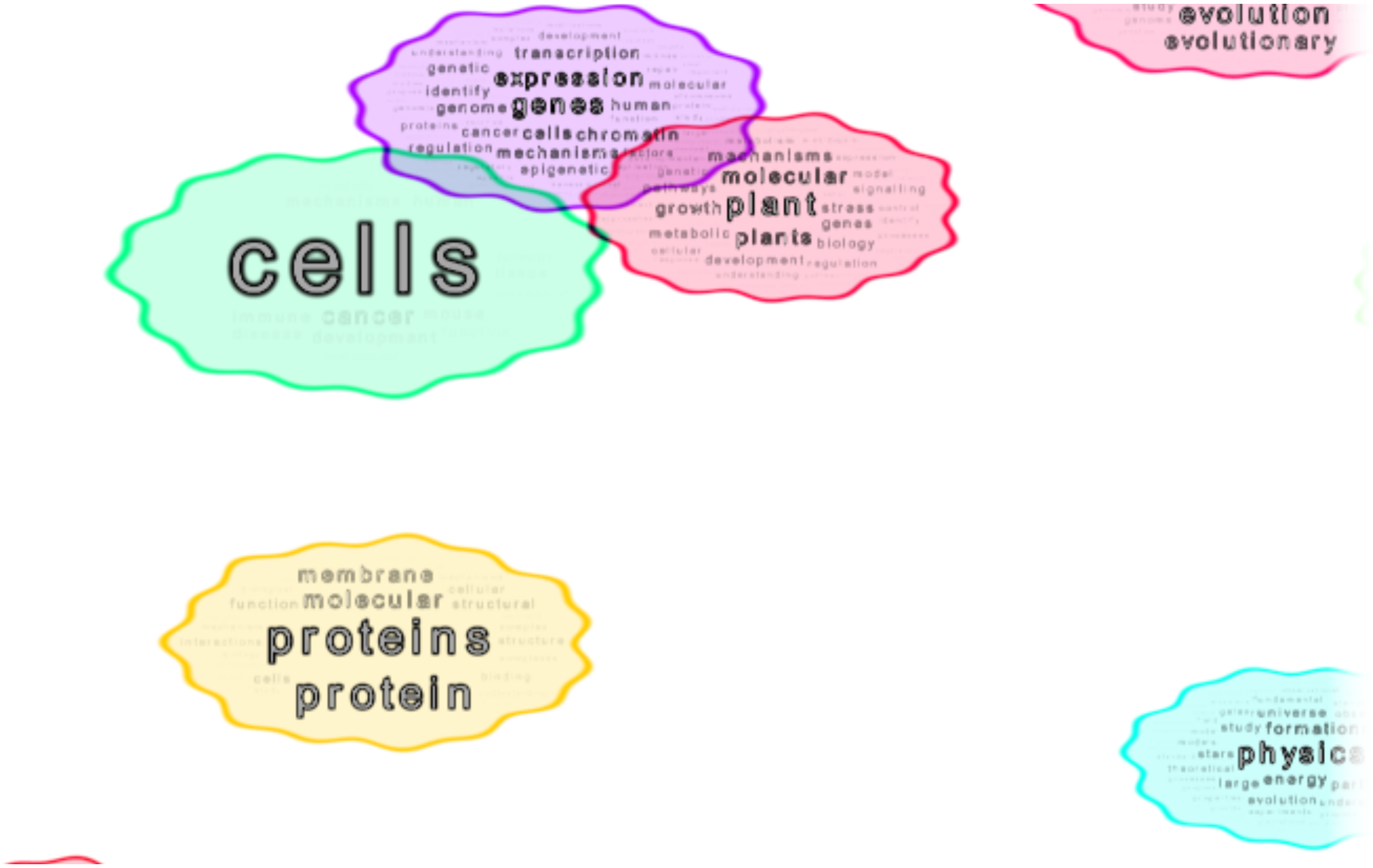
Exploratory Search

- Task 2 – Overview over Topics
  - use information about topic relevance from last task
  - problem: find similarities between topics
    - interprete topics as distributions -> posterior log-odds, Jensen-Shannon divergence
    - interprete topics as vectors in R^V -> cosine distance, euclidean distance etc.

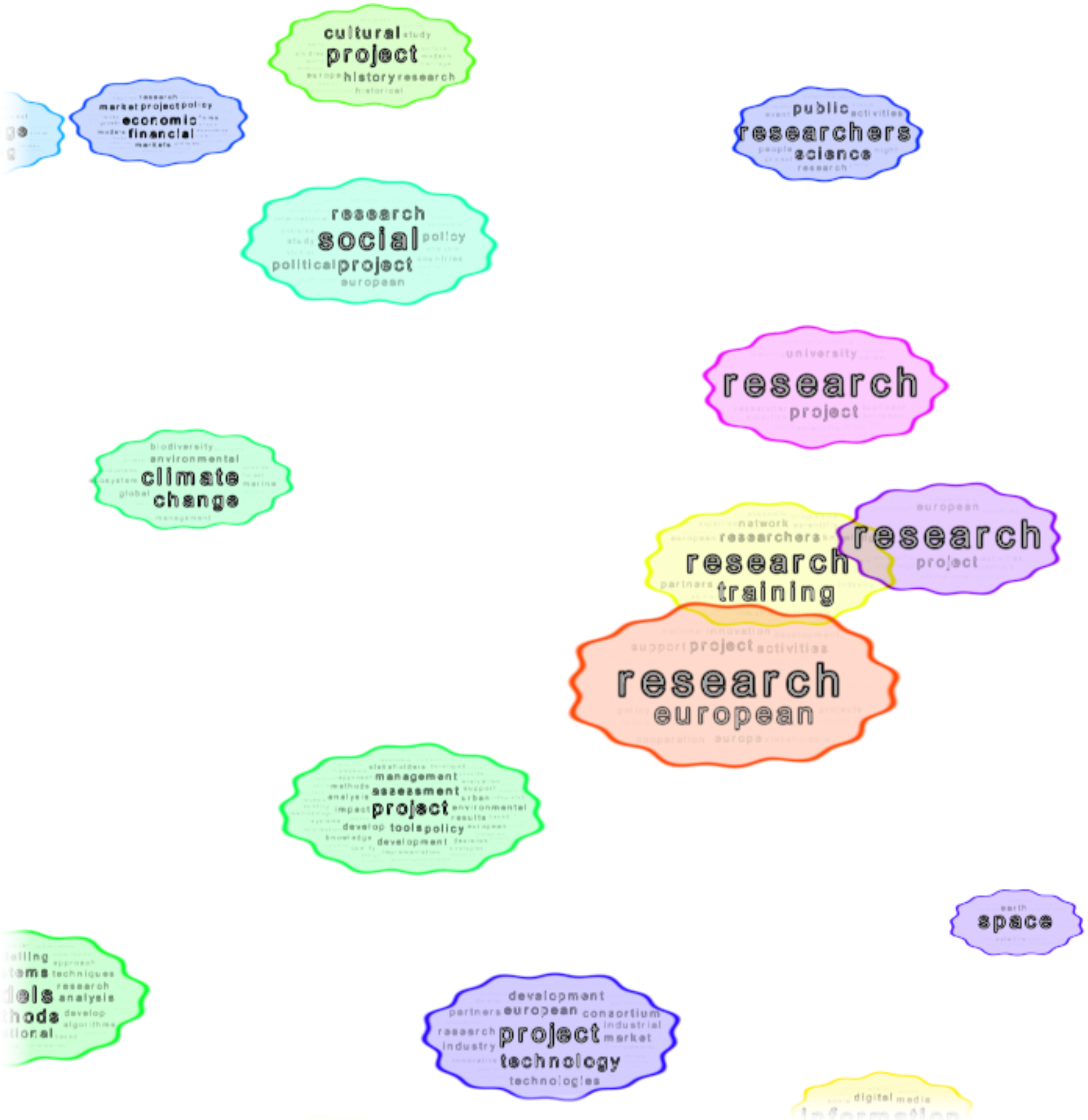
Visual Analysis

- Task 2 – Overview over Topics
  - after computing similarities, we have a similarity matrix
  - find a mapping into two dimensions given this matrix
  - implementation with Force Directed and Sammon‘s mapping

Visual Analysis



Visual Analysis



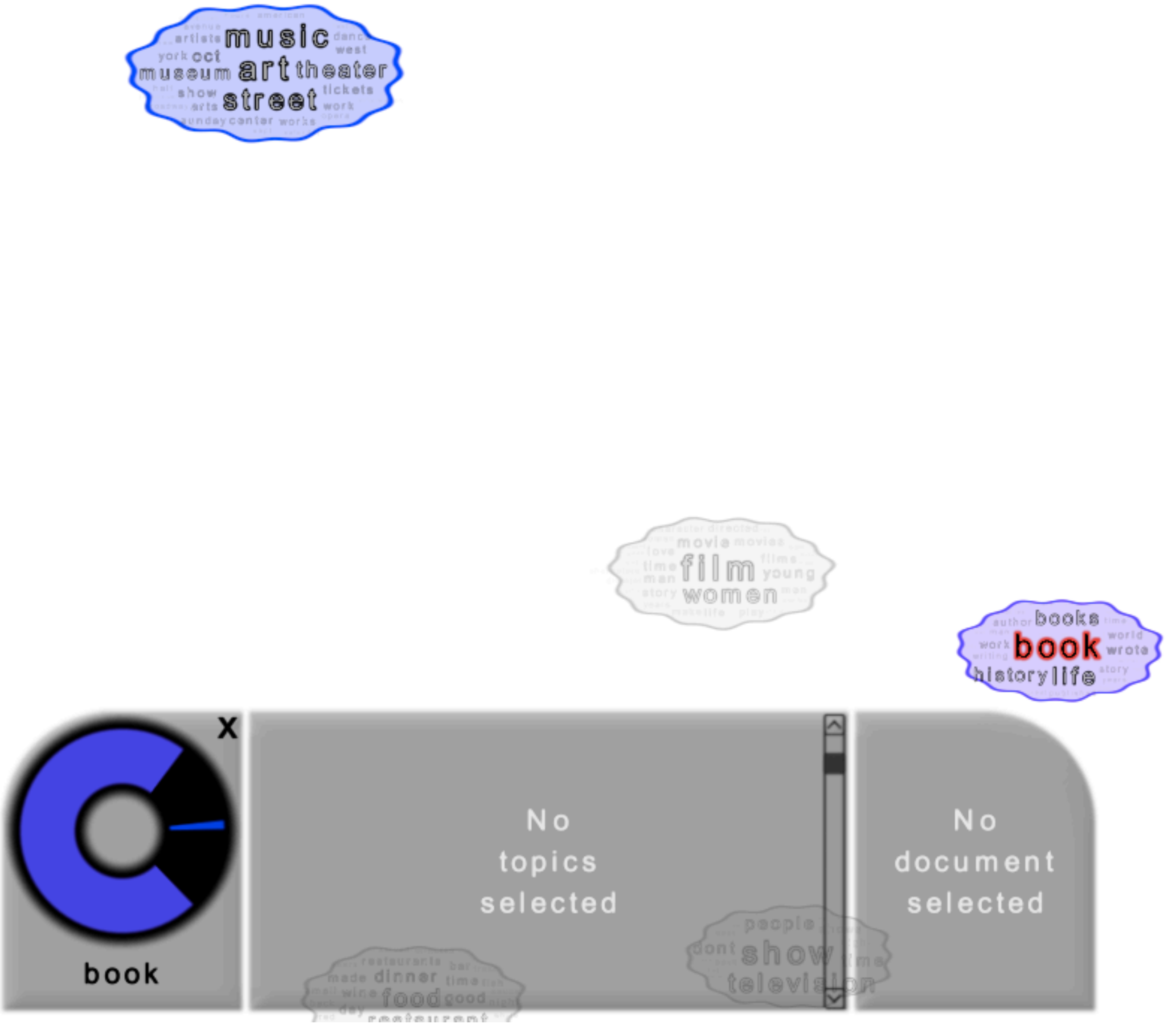
Exploratory Search

- Task 3 – Ambiguity resolution
  - based on structural semantics
  - if terms appear in different contexts, they have different meanings
  - meanings can be related (polysemous terms) or not (homonyms)

Visual Analysis

- Task 3 - Ambiguity resolution
  - words are clickable and highlighted when selected
  - also highlighted in other topics (when they are above a certain threshold)
  - a pie chart shows how high the probability is in other topics
  - other topics are de-colored

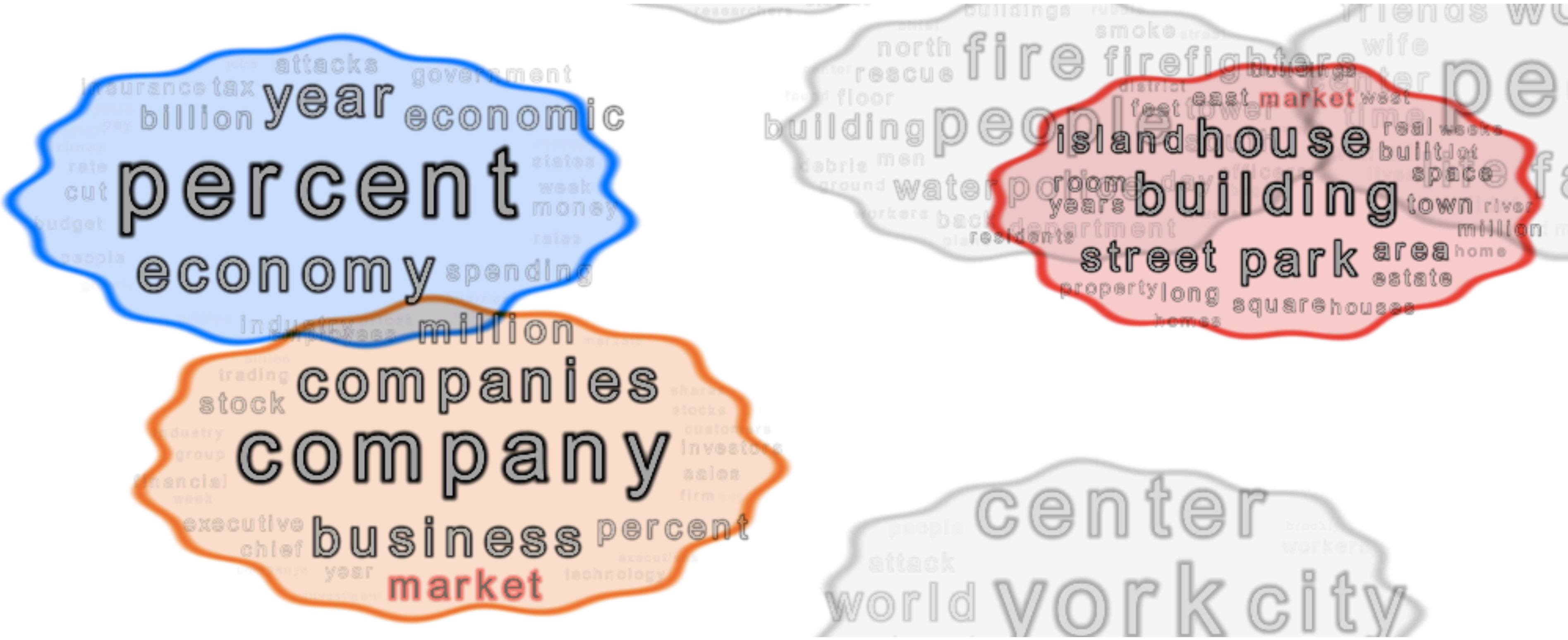
Visual Analysis



No topics selected

No document selected

Visual Analysis



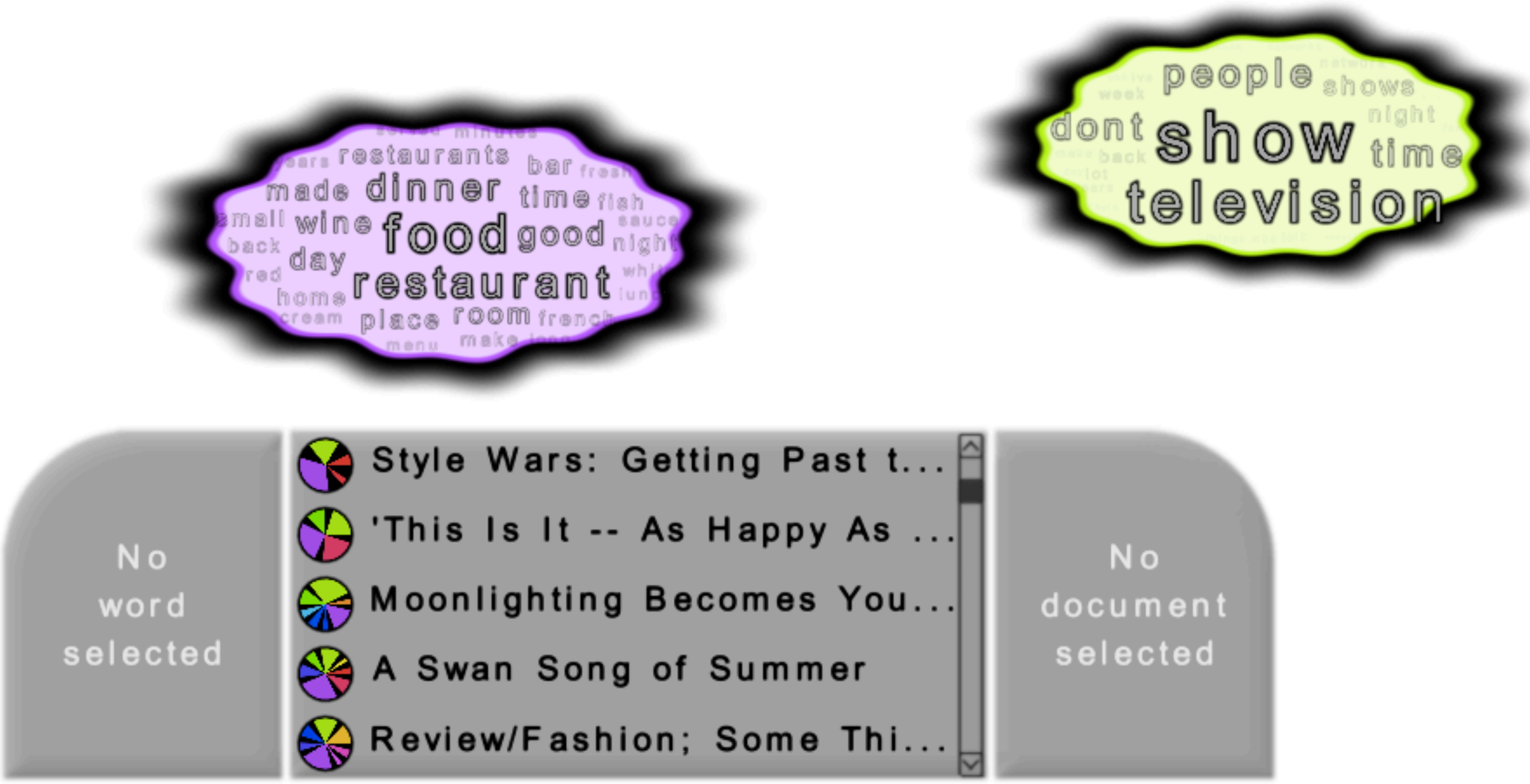
Exploratory Search

- Task 4 – Document retrieval
  - find documents that are related to topics
  - makes use of documents‘ topic distributions (the bar charts)
  - again sort by probabilities
  - also, select multiple topics -> gets documents that are related to topic combination

Visual Analysis

- Task 4 – Document retrieval
  - Documents are also clickable (with right click) and then framed
  - multiple topics are clickable in that way
  - according to the probability of the selected topics, a list of document is displayed

Visual Analysis



No word selected

Style Wars: Getting Past t...

'This Is It -- As Happy As ...

Moonlighting Becomes You...

A Swan Song of Summer

Review/Fashion; Some Thi...

No document selected

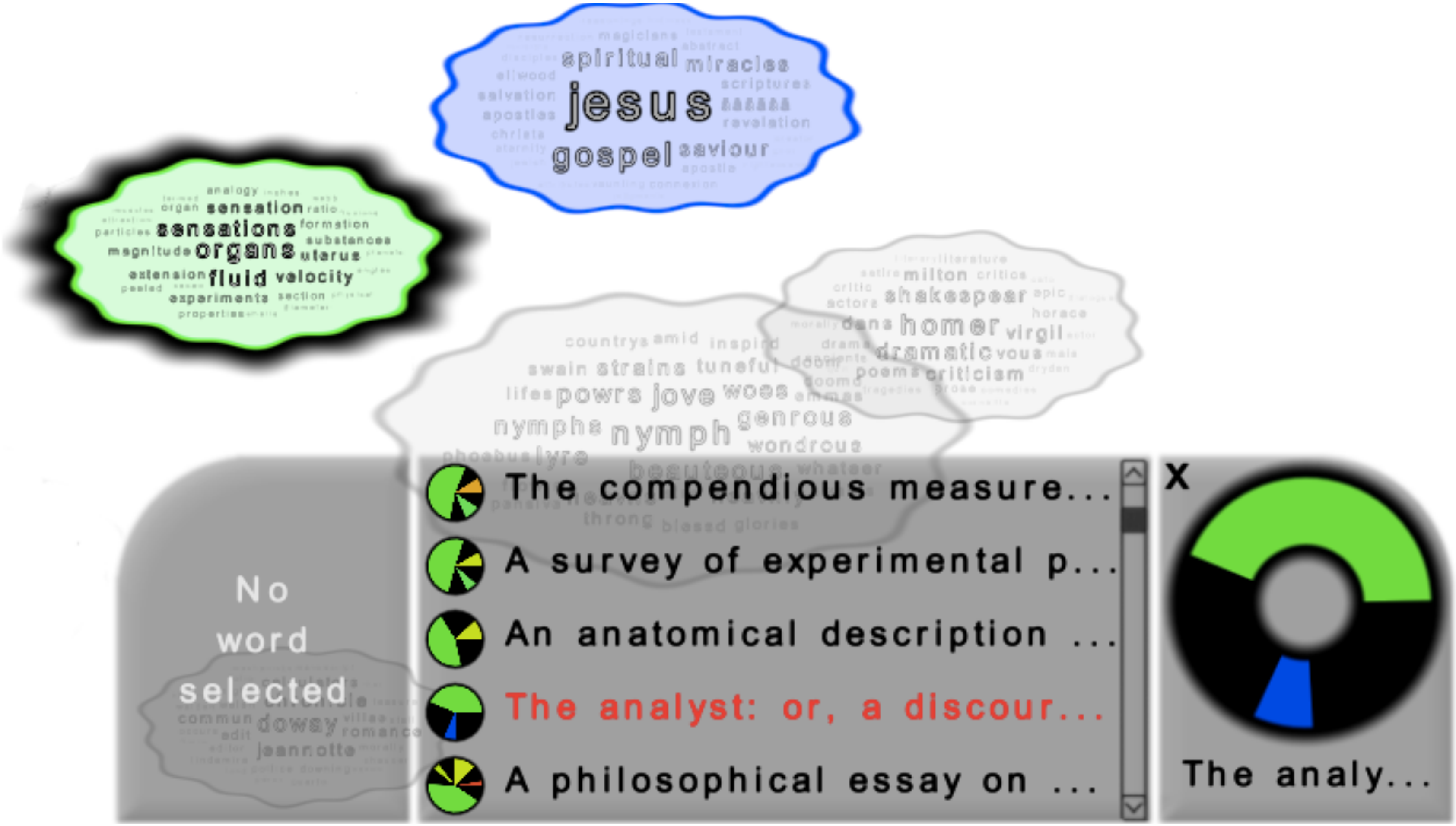
Exploratory Search

- Task 5 – Finding topics transitively
  - having identified an interesting document find other topics related to it
  - easy task, but can help immensely to uncover relationships in the data

Visual Analysis

- Task 5 – Finding topics transitively
  - click documents in the list highlights them
  - shows a second pie chart with the topic distribution of the selected document
  - pie chart elements are clickable and select other topics
  - in consequence, this also finds related documents

Visual Analysis



The compendious measure...

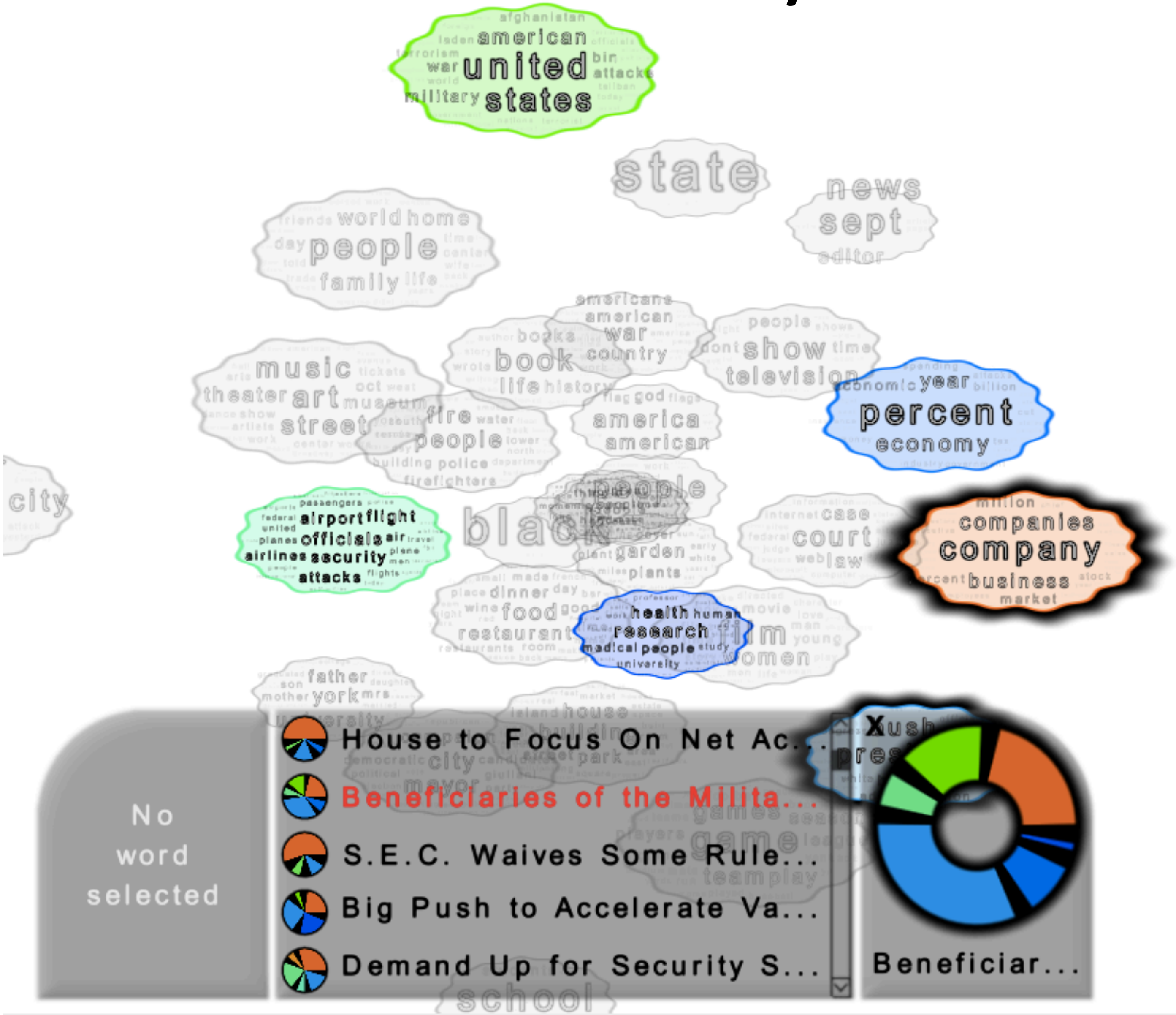
A survey of experimental p...

An anatomical description ...

The analyst: or, a discour...

A philosophical essay on ...

Visual Analysis



No word selected

House to Focus On Net Ac...

Beneficiaries of the Milita...

S.E.C. Waives Some Rule...

Big Push to Accelerate Va...

Demand Up for Security S...

Beneficiar...

References

[1] http://en.wikipedia.org/wiki/2010_Haiti_earthquake

[2] http://odotkrown.deviantart.com/art/floyd-mayweather-manny-pacquiao-by-oDOTkrown-514560496

[3] US Coast Guard – 100421-G-XXXXL-Deepwater Horizon fire, http://cgvi.uscg.mil/media/main.php?g2_itemId=836285

[4] http://pixabay.com/en/rocket-launch-rocket-take-off-nasa-67643/

[5] David M. Blei, „Probabilistic Topic Models“, talk at Machine Learning Summer School Kyoto, 2012, http://www.cs.columbia.edu/~blei/talks/Blei_MLSS_2012.pdf