LDAExplore

Visualizing Topic Models Generated Using LDA

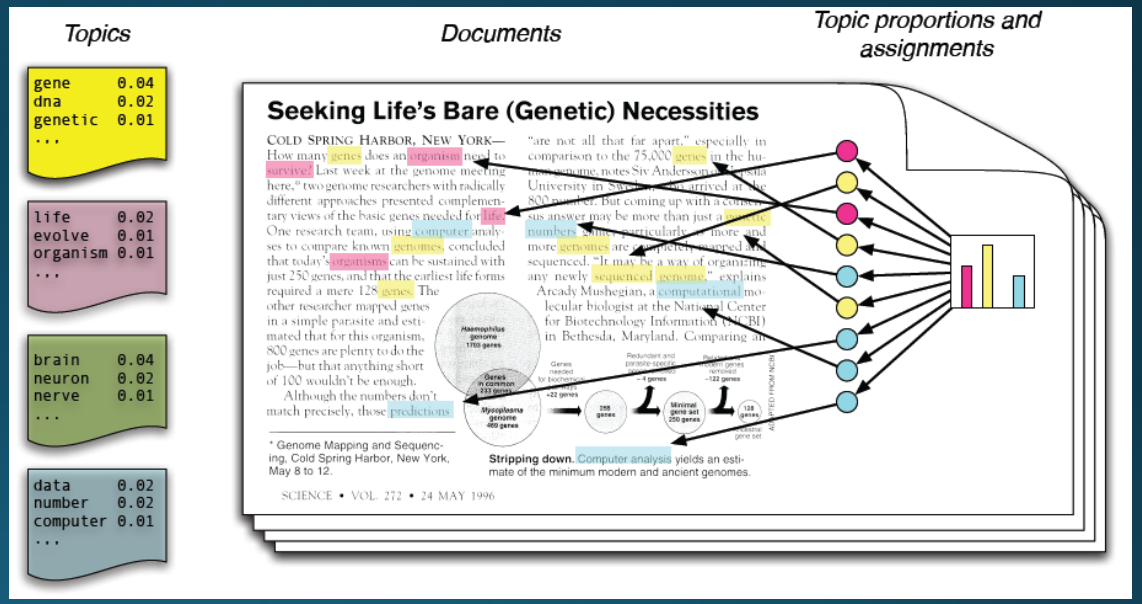
Ashwinkumar Ganesan, Kiante Brantley,

Shimei Pan & Jian Chen

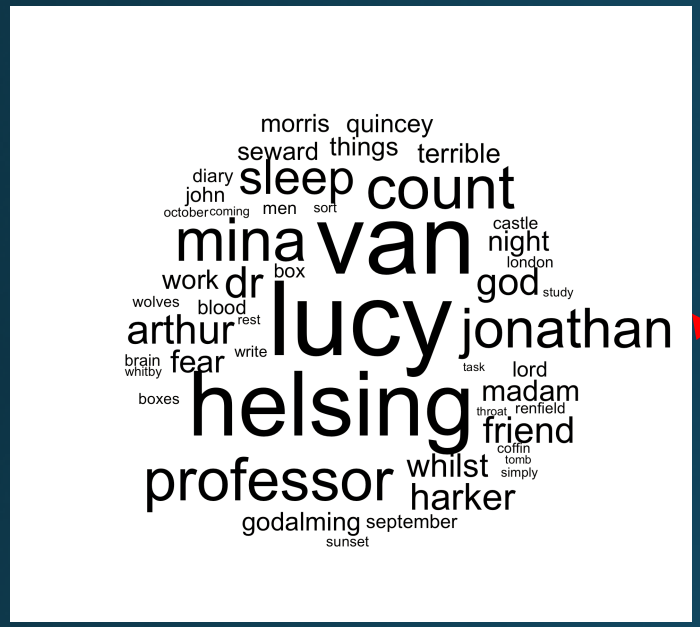
The Gist

What is Topic Modeling ?

- Find hidden topics
- Process Large Sets of Documents
- Group words & information together
- Understand topics & documents



Topic Modeling

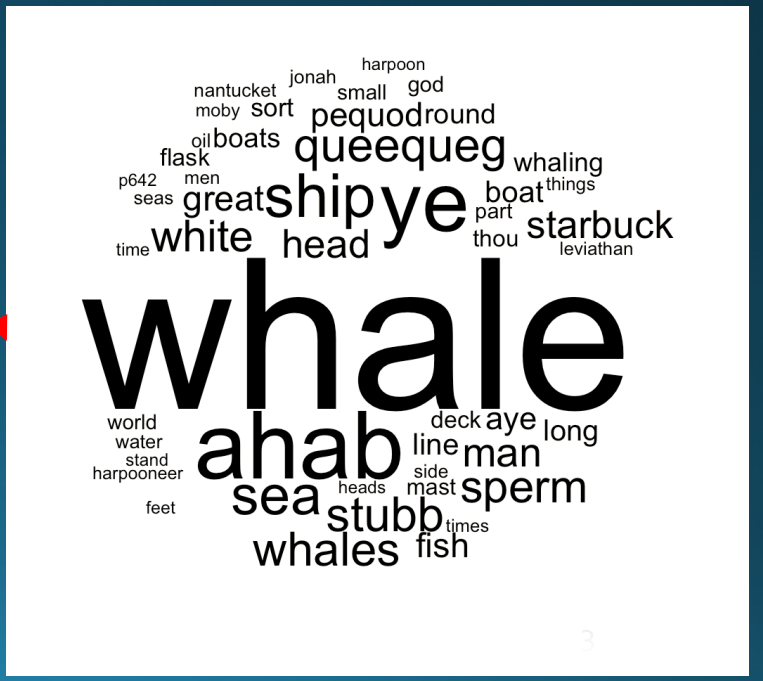


DRACULA

It should not be thought from above that Soviet party line is necessarily disingenuous and insincere on part of all those who put it forward many of them are too ignorant of outside world and mentally too dependent to question (*) self-hypnotism, and who have no difficulty making themselves believe what they find it comforting and convenient to believe. Finally we have the unsolved mystery as to who, if anyone, in this great land actually receives accurate and unbiased information about outside world. In atmosphere of oriental secretiveness and conspiracy which pervades this government, possibilities for distorting or poisoning sources and currents of information are infinite. The very disrespect of Russians for objective truth---indeed, their disbelief in its existence--- leads them to view all stated facts as instruments for furtherance of one ulterior purpose or another. There is good reason to suspect that this government is

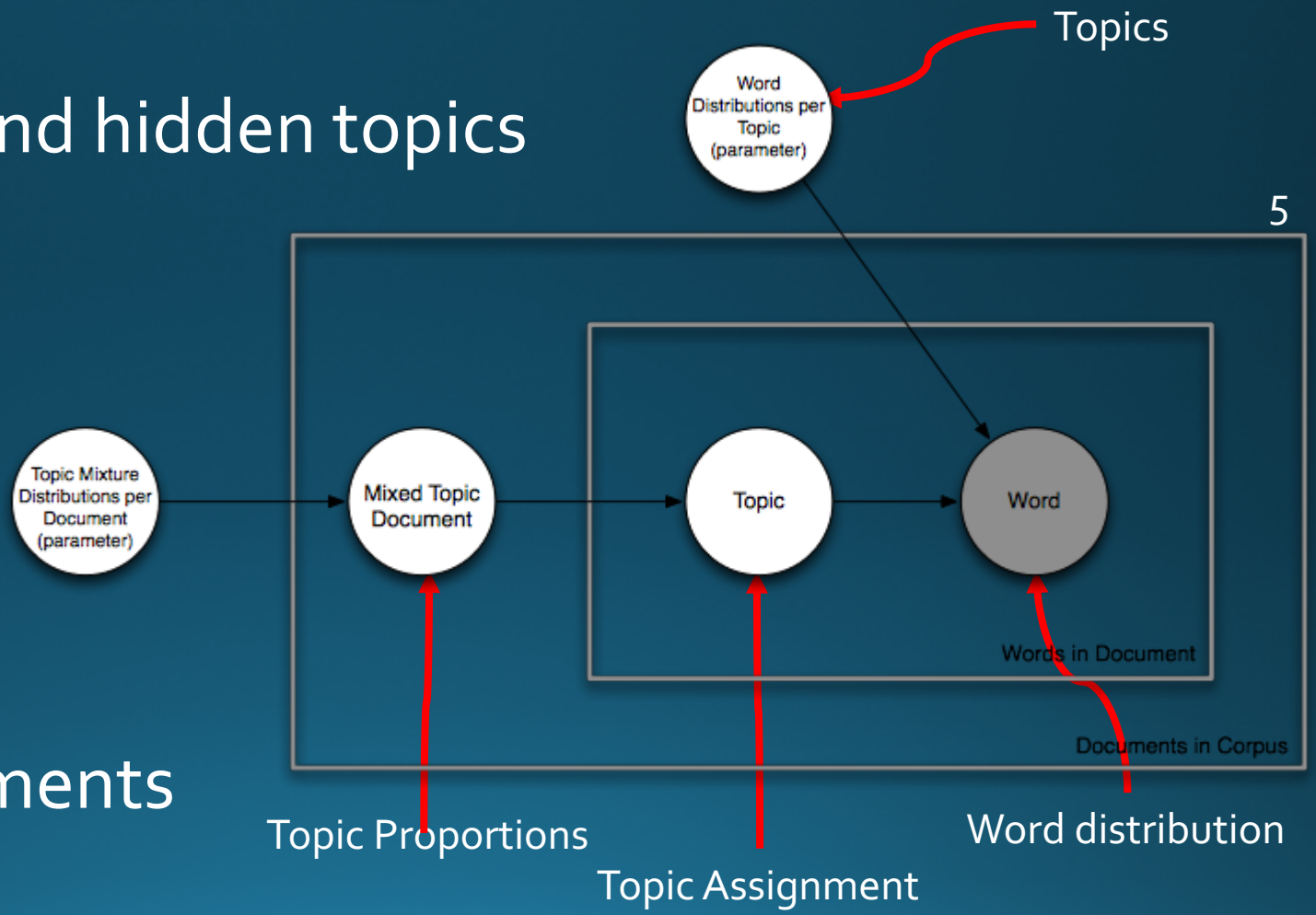
field of international law. This publication has been written with the expectation that the military attorneys making use of it will be provided with a basic understanding of the legal system governing the international community. International law is an area of jurisprudence which challenges. It quite often fails to provide concise "textbook answers" to problems which reach a degree of complexity far greater than that found in any other legal system. Entrusted with the task of regulating the conduct of international sovereign entities, it is a legal framework which develops on a daily basis. Its successes go largely unnoticed, while its failures gain almost instantaneous notoriety and condemnation. It is a jurisprudential system particularly unsuited for complacent personalities and regimented minds. Hopefully, military attorneys will not view the often evident imprecision of international law as a fatal weakness but as an opportunity afforded its practitioner to develop an efficient and viable legal system. Constructive criticism and the ability to apply concepts and rules to practical international legal problems must be based on a working knowledge of the subject matter. The achievement of this end underlies the purpose of this publication.

MOBY DICK



Latent Dirichlet Allocation

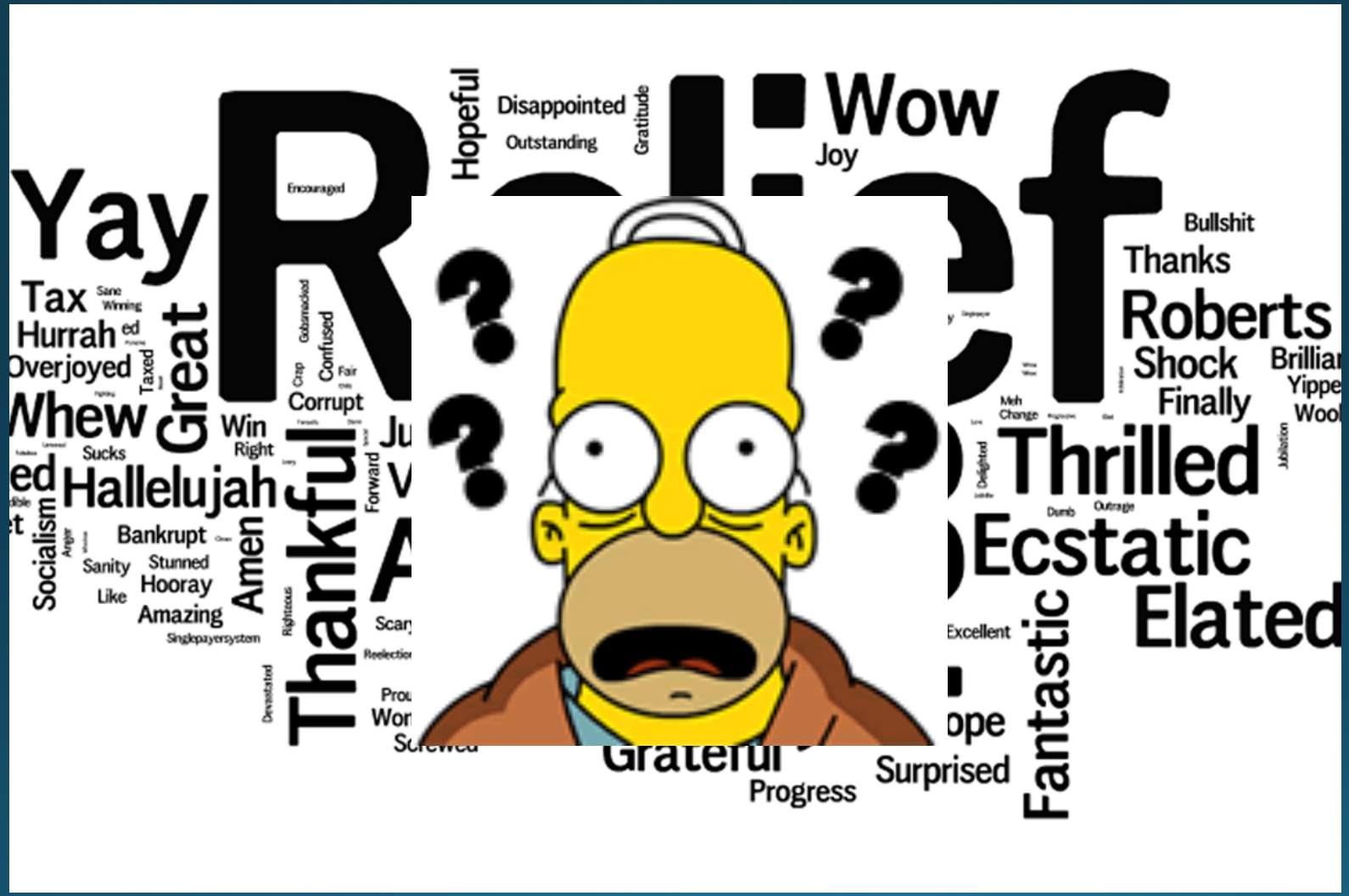
- Probabilistic method to find hidden topics
- Word distributions
- Topic distributions
- Link topic, words & documents



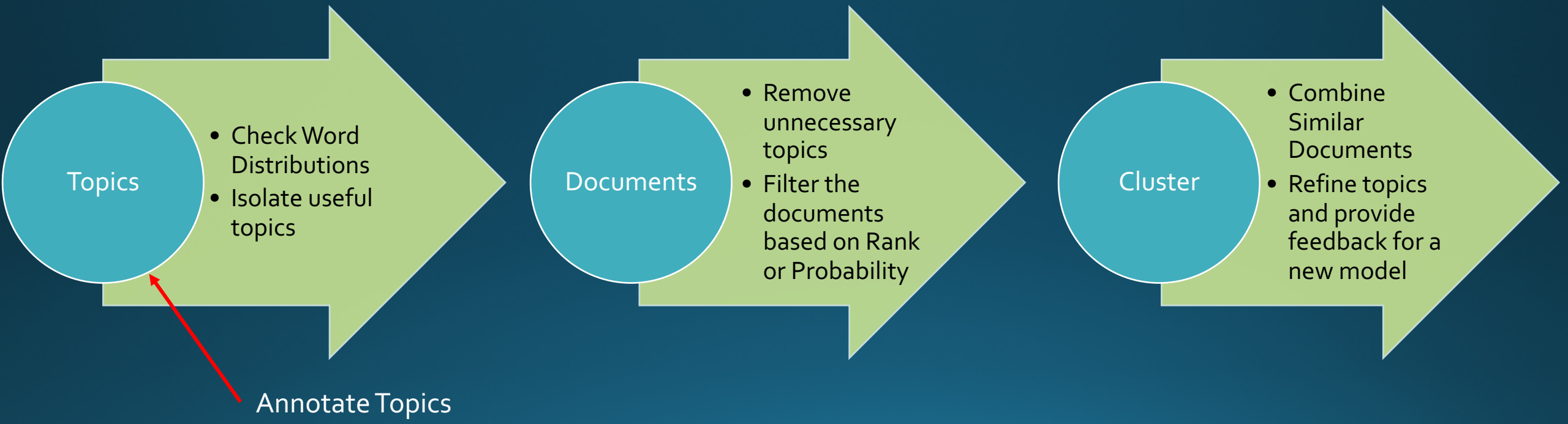
Problems with LDA

- Hidden Parameters
- Number of Topics
- Number of Iterations
- Naïve or Knowledgeable Users
- Understanding Documents

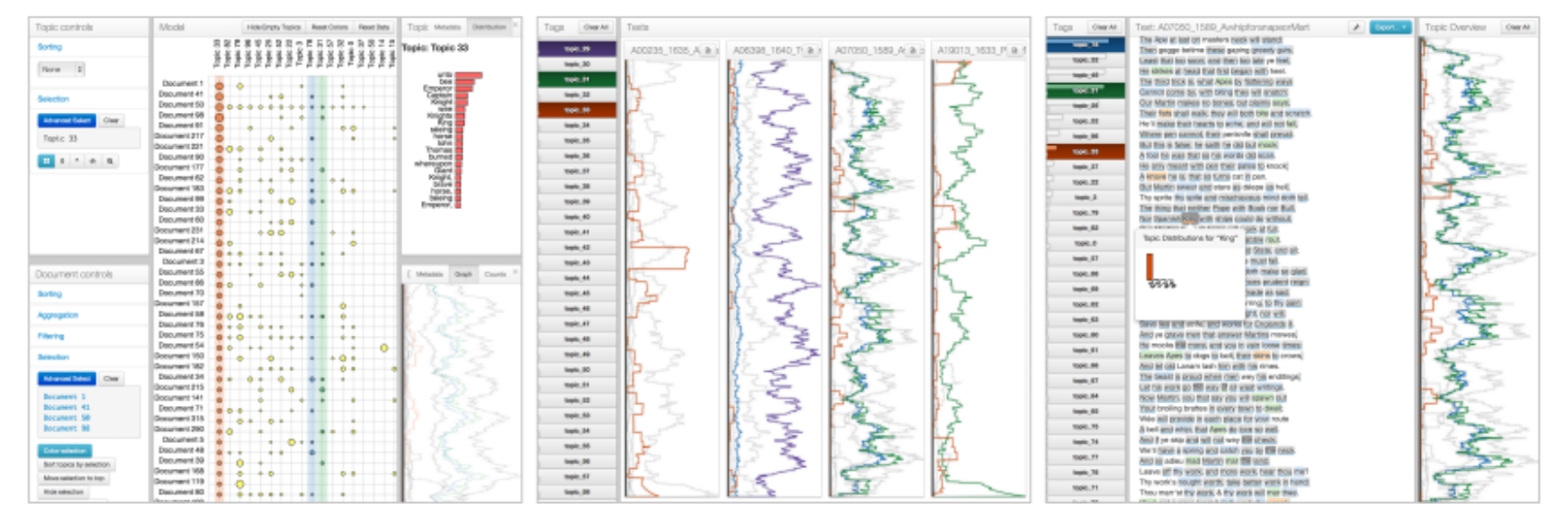
Need For Visualization



Example User Work Flow

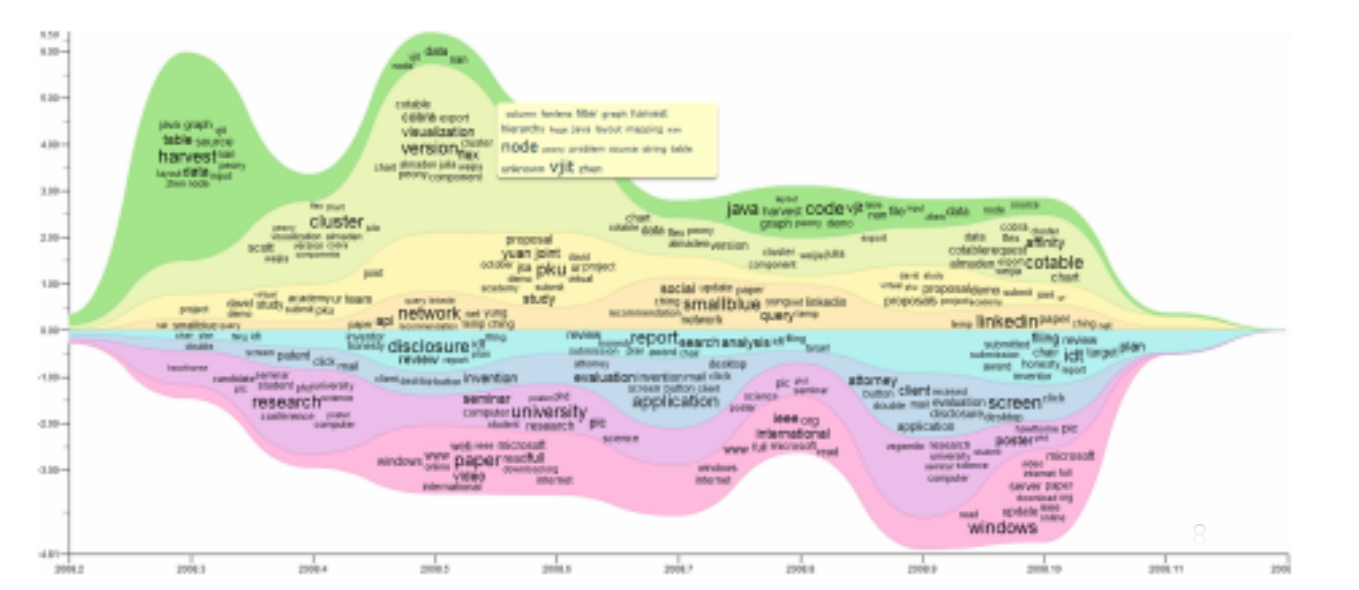


Related Visualizations



SERENDIP [3]

TIARA [4]



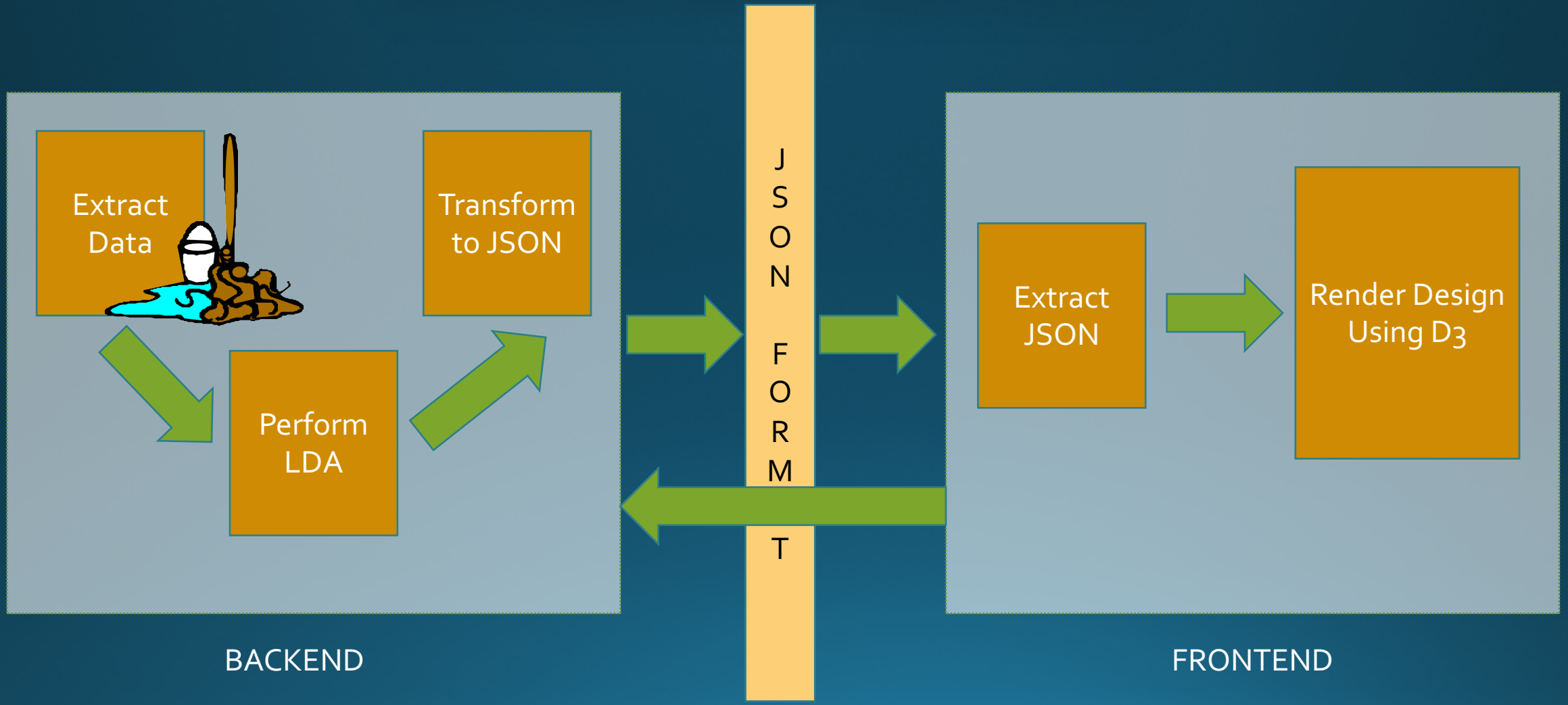
Problems With Visualizations

- Document Exploration without modeling User Feedback
- Hyper parameters that need configuration
- Topics Generated are difficult to understand
- Large Document Corpus require Multiple Views / Scrolling
- Real Time computation of model is difficult

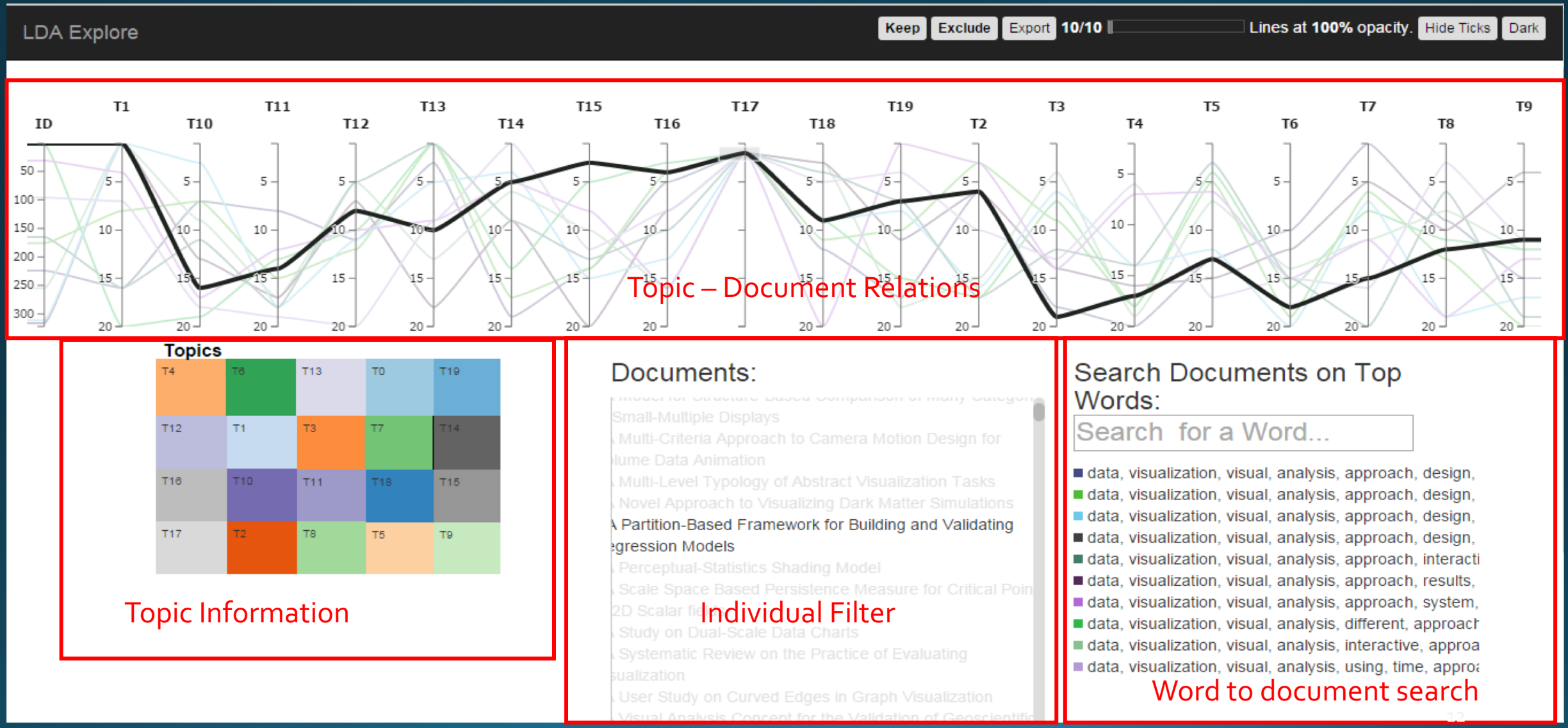
Design Goals

- Visualizing Topics in the document corpus
- Topic Document Relations
- Filtering Documents
- Performing Set Operations
- Clustering Topics & Documents
- Topic Annotations

Visualization Pipeline

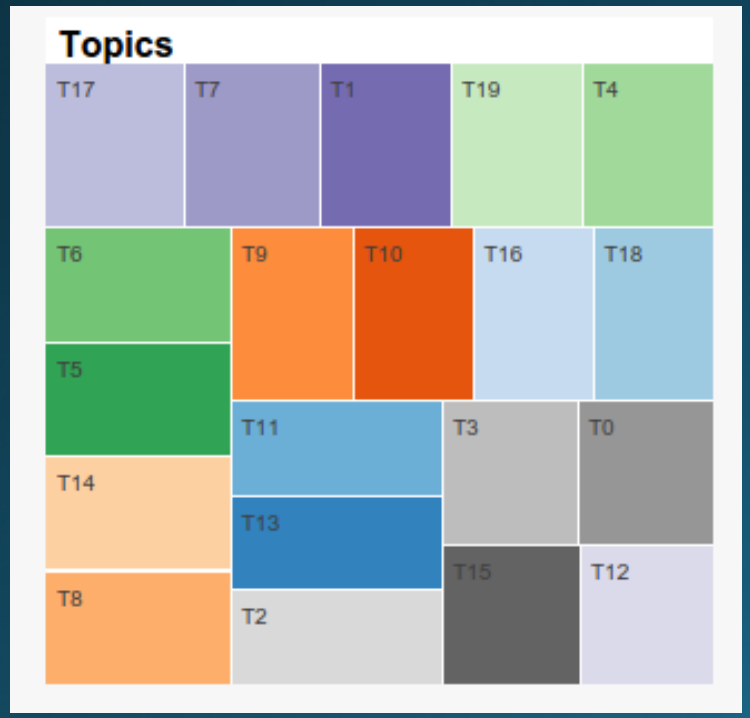


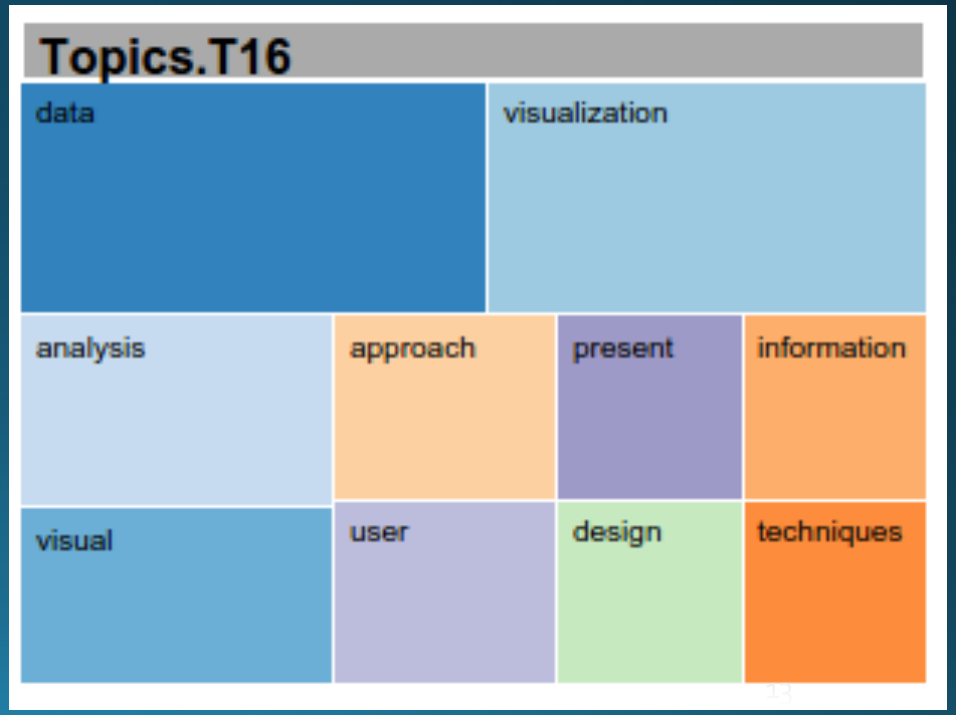
Overall Design



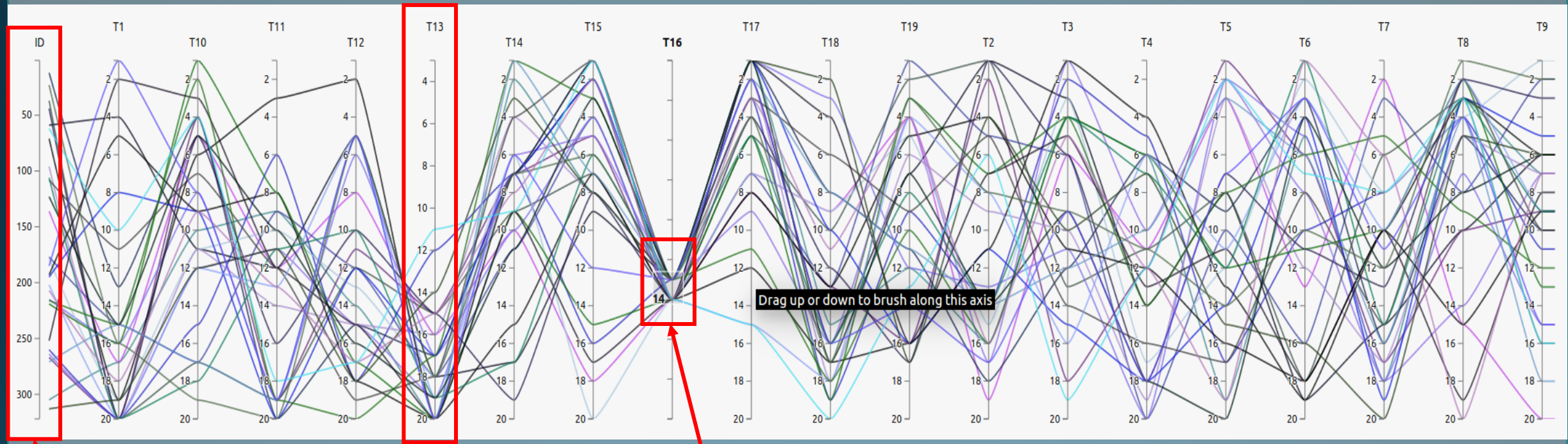
Topic Word Distributions

- Shows the Top 10 Words
- Rectangular sizes show which word is more probable

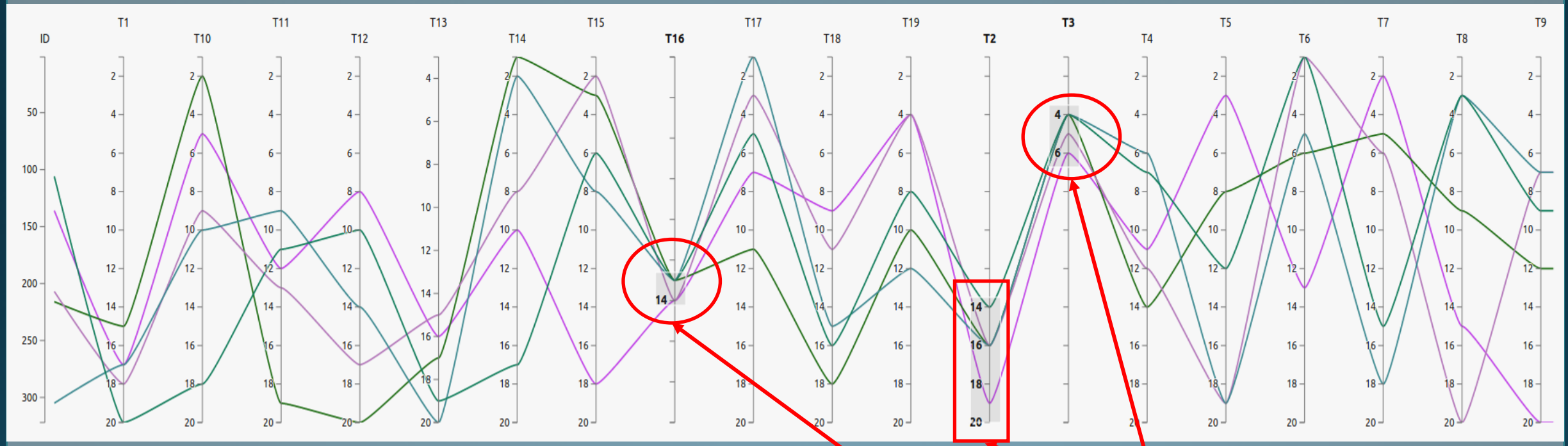




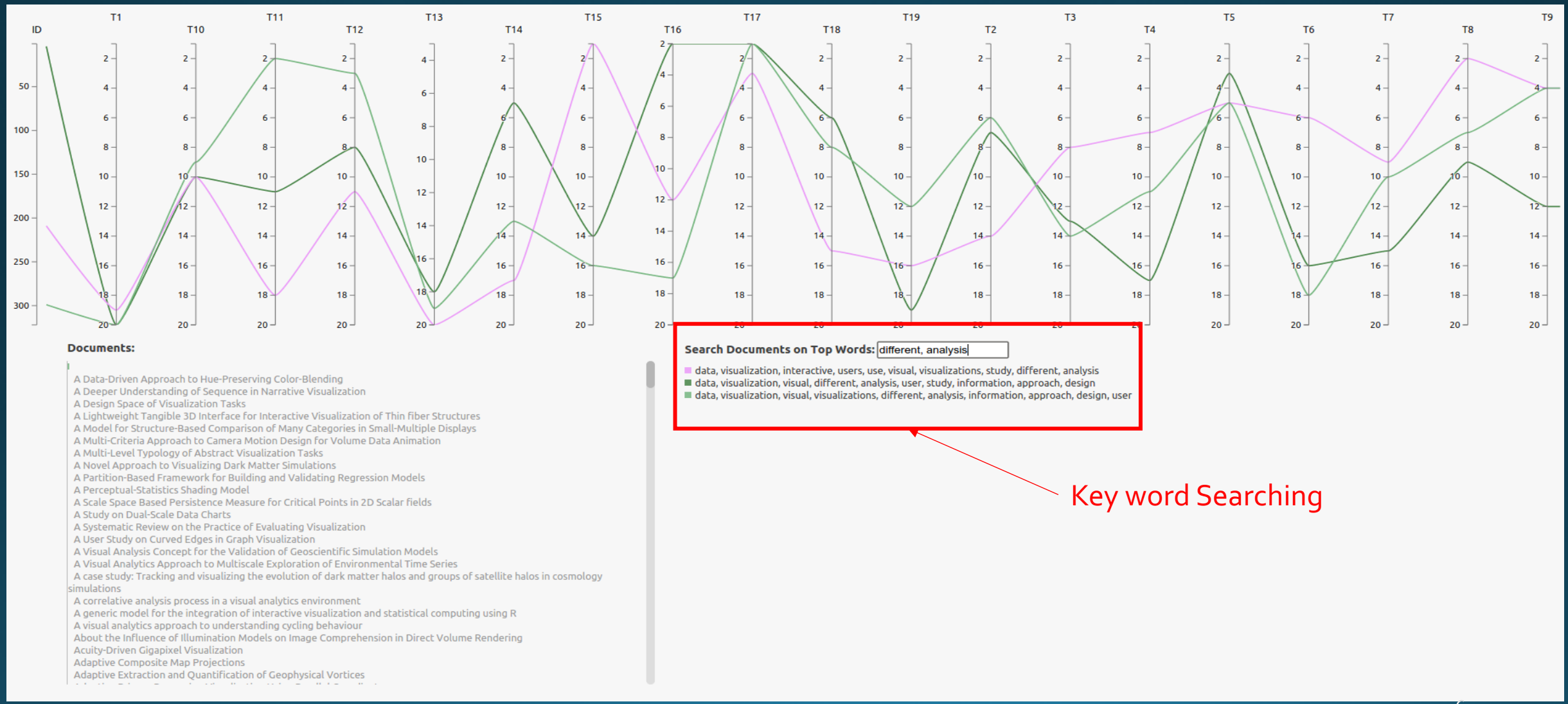
Topic – Document Relations



Topic – Document Relations



Keyword Searches



Generating Top Words & TFIDF

- Using TF-IDF as a measure to clean the data
- There is a preset threshold for eliminating the words

- Word Probabilities:

$$P(w_x, dy) = \sum_{t=1}^{n} \sum_{i=1}^{n} P(w_x, t_i) * P(ti, dy)$$

Pilot Study Details

- There are 5 participants in the survey
- The participants did not have any formal training on the tool
- There is a brief introduction given to the participants at the start of the survey
- Some participants understand topic modeling

Pilot Study Details (2)

Survey Questionnaire

- Overview
- Topics
- Filtering
- Keyword Search

Types Of Questions

- Tasks Execution questions

  How many documents are represented in this visual?

- Understanding questions

  Does the visual have many things on it?

- Reasoning questions

  Which is the least important topic? (Important means highly ranked topic)

- Usability questions

  Are the rankings on the axis visible / readable?

Recommend Changes

- Eliminate Documents
- Connect topics and parallel coordinates
- Combine Searching
- Keep & Exclude can be connected to searching
- Creating a reset button
- Create probability mode & Ranking mode
- Topic – Rank group
- Filtering & Selection memory

Thank You
Questions?

References

[1] Blei, David M. "Probabilistic topic models." Communications of the ACM 55.4 (2012): 77-84.

[3] . Alexander, E., Kohlmann, J., Valenza, R., and Gleicher, M. Serendip: Turning topics back to the text. IEEE Visualization Poster Proceedings.

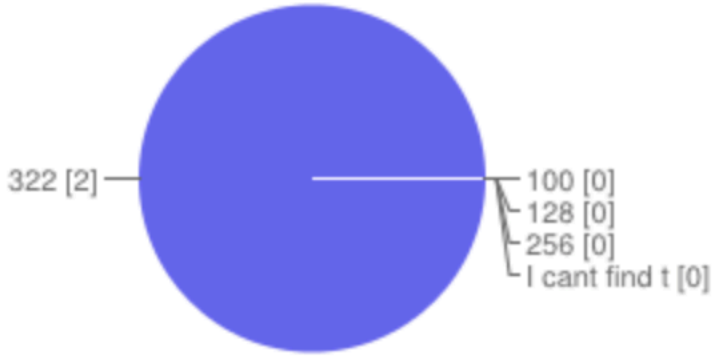
[4] Pan, S., Zhou, M. X., Song, Y., Qian, W., Wang, F., and Liu, S. Optimizing temporal topic segmentation for intelligent text visualization. In Proceedings of the 2013 international conference on Intelligent user interfaces, ACM (2013), 339–350.

[5]http://mcburton.net/blog/joy-of-tm/

[6] Choo, J., Lee, C., Reddy, C. K., and Park, H. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. Visualization and Computer Graphics, IEEE Transactions on 19, 12 (2013), 1992–2001.

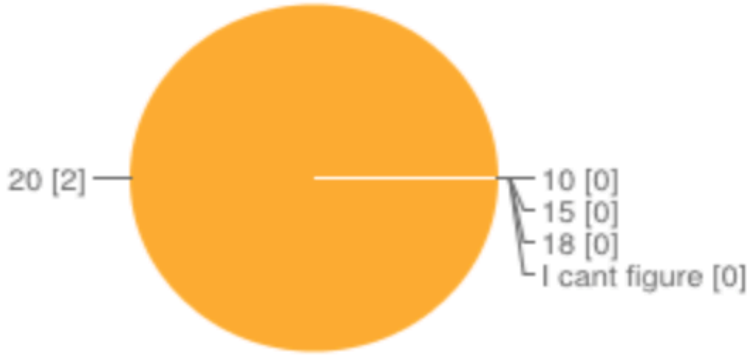
Survey Results

How many documents are represented in this visual?



100	**0**	0%
128	**0**	0%
322	**2**	100%
256	**0**	0%
I cant find the information	**0**	0%

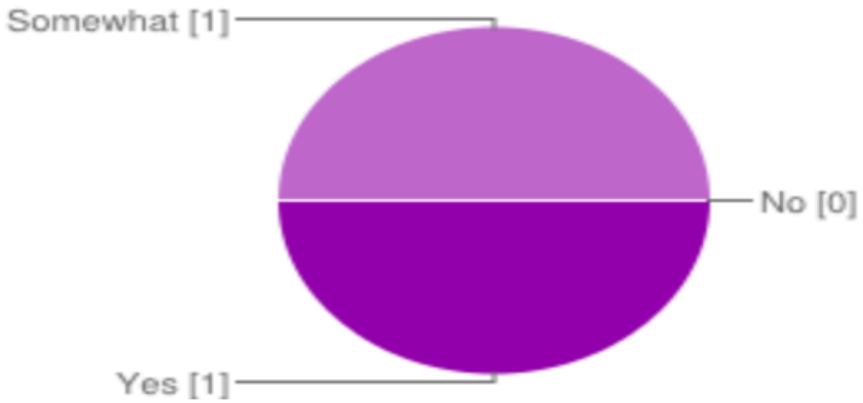
How many topics are there in this visual?



10	**0**	0%
20	**2**	100%
15	**0**	0%
18	**0**	0%
I cant figure it out	**0**	0%

Survey Results

Is it easy to filter using the above method?



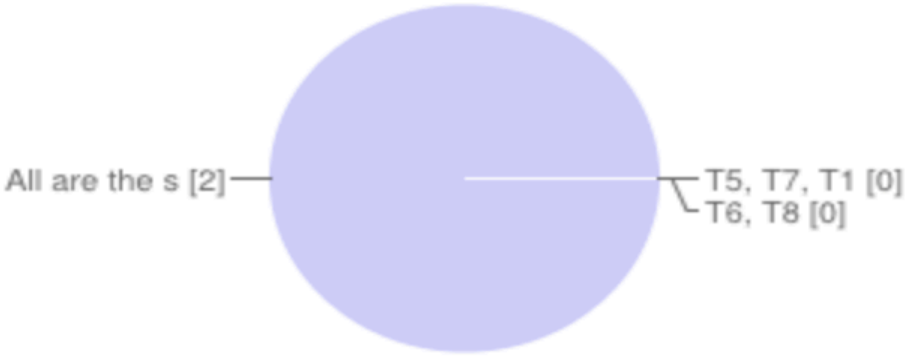
Yes	**1**	50%
Somewhat	**1**	50%
No	**0**	0%

Apply a filter for topic T7 across the range of ranks 1 to 2. What is the relation between topic T7 and T19?

T7 is more important than T19	**2**	100%
T19 is more important than T7	**0**	0%
Can't Say	**0**	0%

Survey Results

Only by looking at the given topics grid, which topics do you feel are important?



T5, T7, T1	**0**	0%
T6, T8	**0**	0%
All are the same	**2**	100%

Move the filter on T19 axis from top to bottom. What do you notice?

Keywords change. Relevant documents in the filter range are highlighted in black while others are grayed out. This is not obvious at first though.
They all are about data, visualization and there are documents seem to be evenly distributed throughout the axis

Survey Issues

- Ambiguity in Questions
- Imprecise human movements
- Filtering issues
- Remote tests
- Adding Ease of Use questions
- Adding a purpose section
- Better Example can be provided
- Data Selection