FIZ Karlsruhe

Leibniz-Institut für Informationsinfrastruktur

Tatyana Skripnikova

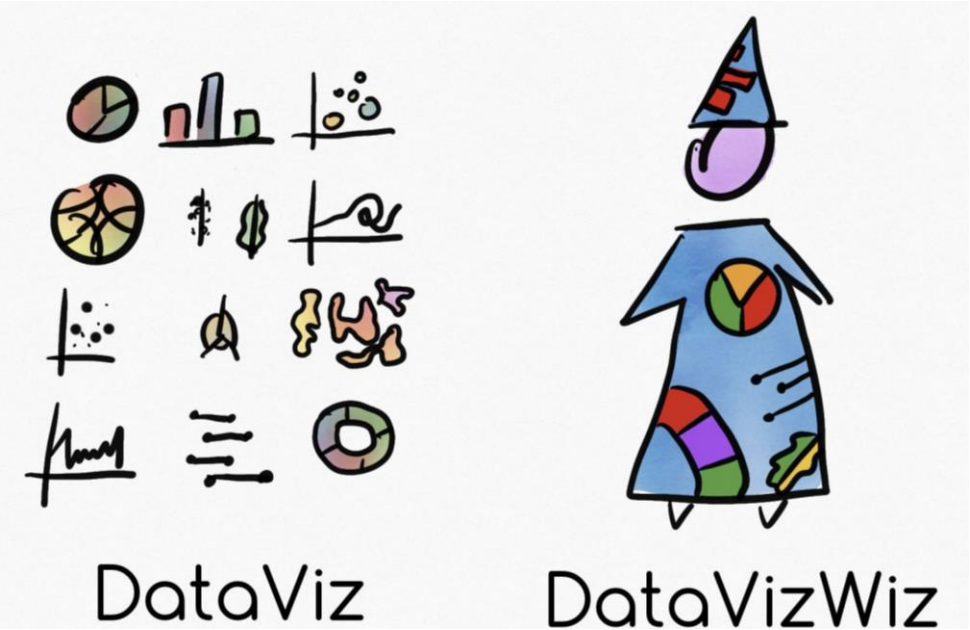
Semantic Views - Interactive Hierarchical Exploration for Patent Landscaping

Tatyana Skripnikova

M.Sc. in computer science at Karlsruhe Institute for Technology

Data scientist at generic de

inovex



Comic source: https://medium.com/nightingale/100-days-of-dataviz-comics-9a24789f3f69

Overview

Background and motivation

Data

Document embeddings

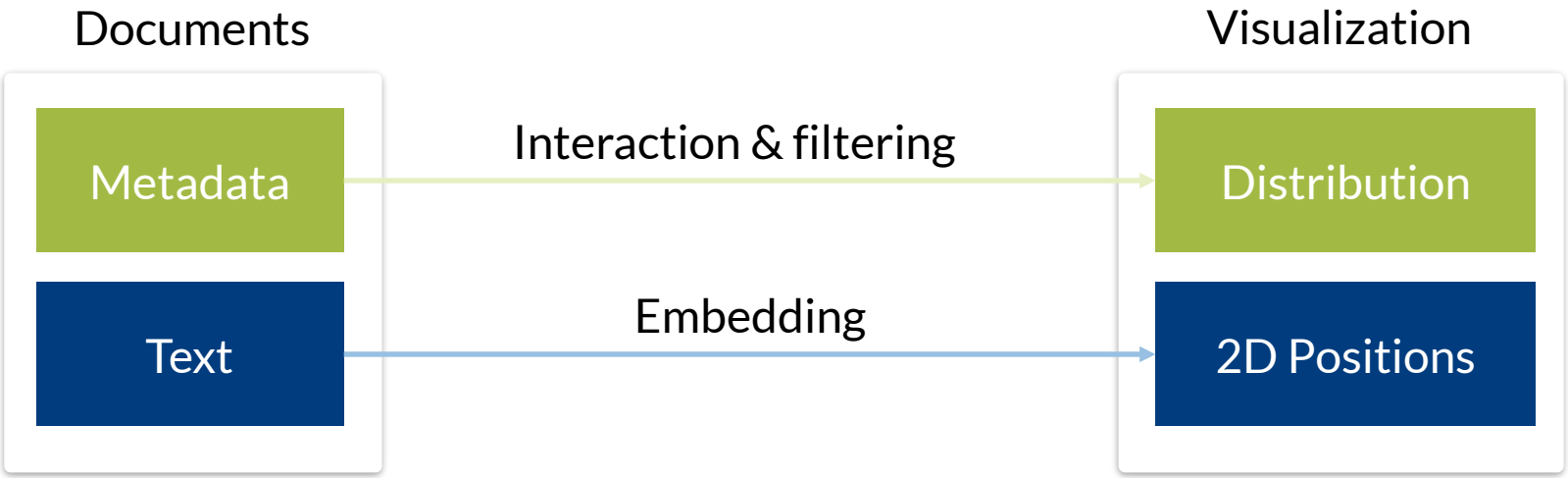
Interaction techniques

Visual elements

Evaluation

Outlook

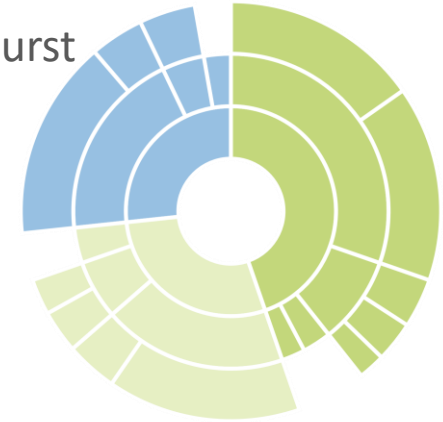
Overview



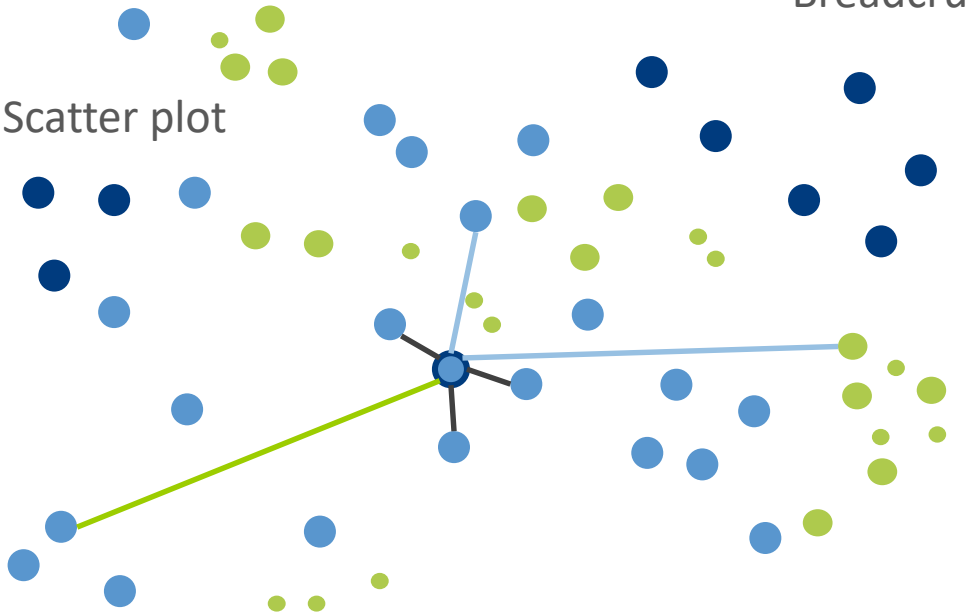
Making semantic similarities tangible for exploration

Breadcrumbs

Sunburst

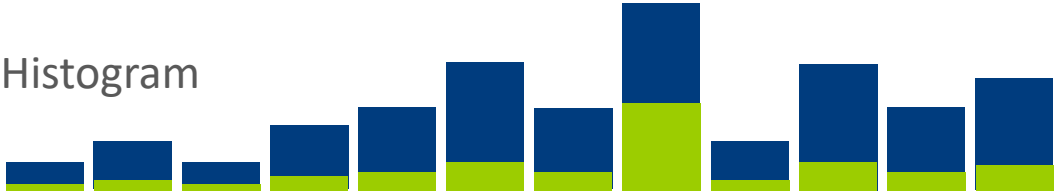


Scatter plot



Details

Histogram



Icon source: Paper by Lil Squid from the Noun Project

Overview

Background and motivation

Data

Document embeddings

Interaction techniques

Visual elements

Evaluation

Outlook

Background and motivation

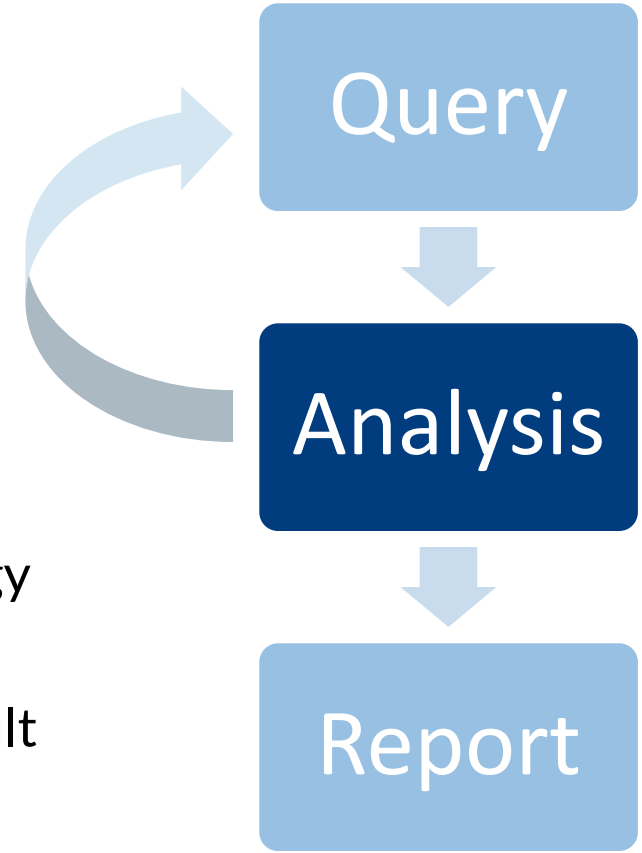
Patent landscaping

Patent situation

of a specific technology or domain

- Research & Development
- Risk assessment - competitors
- New applications for existing technology

- 100s to 1000s of documents in query result
- Experts need help in finding patterns



Overview

Background and motivation

Data

Document embeddings

Interaction techniques

Visual elements

Evaluation

Outlook

Data

Source

- Google Patents Public Datasets
  - Full text for US, abstracts for the rest
    - hair dryer dataset (≈250 docs)
    - 3d printer dataset (≈400 docs)
    - video codec dataset (≈1600 docs)
    - contact lens dataset (≈2600 docs)

- FIZ Karlsruhe
  - diesel engine dataset (≈4700 docs)

https://cloud.google.com/blog/products/gcp/google-patents-public-datasets-connecting-public-paid-and-private-patent-data

Data

Example patent document

US-5448677-A: Electric hair dryer with clogged filter indicator

Priority date: 09.06.1999

Assignees: BRAUN AG, KUECHLER ROBERT

IPC Classes: A45D20/10, A45D20/16

Abstract: A hair dryer has a housing including an air inlet opening and an air exit opening for passage of an air stream ...

Claims: We claim: 1. A hair dryer comprising housing structure, air inlet opening structure in said housing structure …

Kinds of metadata

Categorical attributes

Assignee

- Company or individual (inventor)
  - One or many of both
- Ambiguous
  - Different legal names
  - Spelling
  - Merging with **fuzzy string matching** possible in some cases
  - Otherwise, needs company name thesaurus

Examples

[WIEGAND THOMAS, FRAUNHOFER GES FORSCHUNG]

[FRAUNHOFER GESSELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG E V, WIEGAND THOMAS]

Kinds of metadata

Categorical attributes

Country of registration

- Territorial limits to protection

Family

- Same content, many countries
  - Or diverging content, same country
  - Priority document - first patent in a family

References

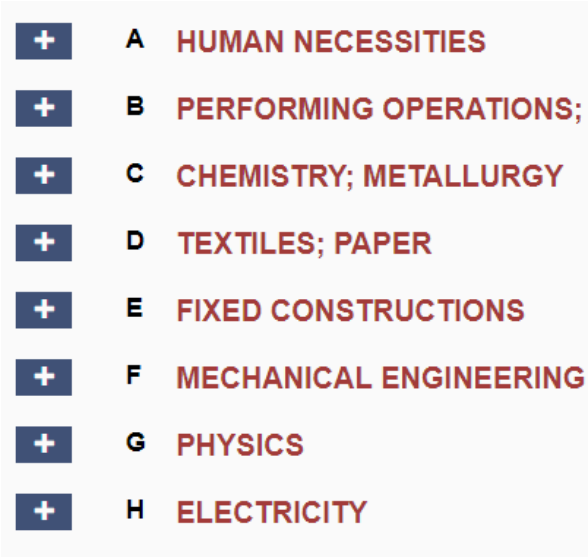
- Forward (citing) and backward (cited by) citations

Kinds of metadata

Categorical attributes

International Patent Classification (IPC) code

- Hierarchical classification of the domain
- 1-20 codes per patent



Section	H	Electricity
Class	H04	Electric communication technique
Subclass	H04N	Pictorial communication, e.g. television
Group	H04N5	Details of television systems
Subgroup	H04N5/202	Gamma control

Kinds of metadata

Date and text attributes

Priority date

Text fields

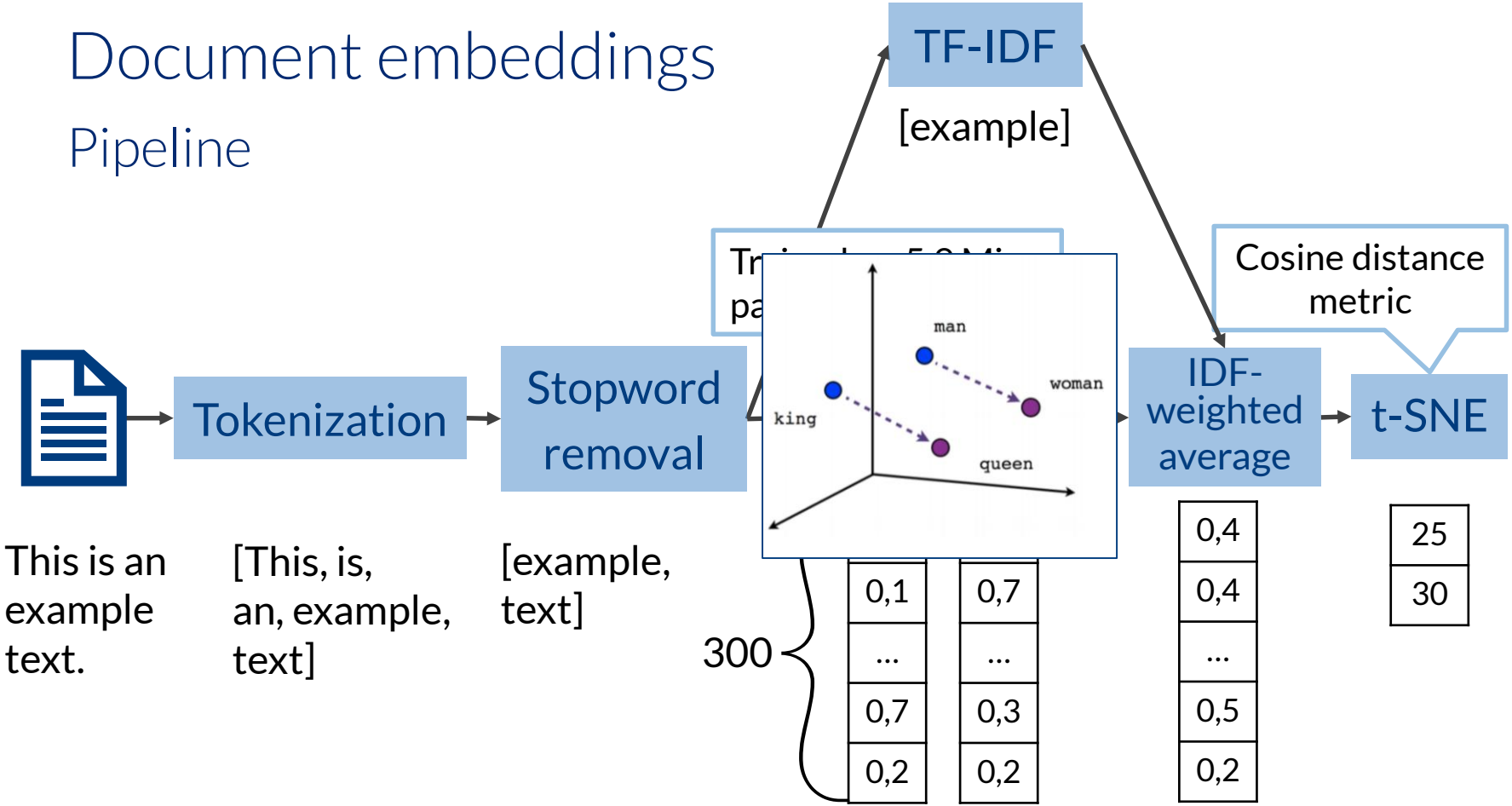
- Title
- Abstract
- Claims

- Mean (title + abstract) ≈ 60 words
- Mean (title + abstract + claims) ≈ 1300 words,
  ≈ 500 words without stopwords

Overview
Background and motivation
Data
Document embeddings
Interaction techniques
Visual elements
Evaluation
Outlook

Document embeddings

Pipeline



Hierarchical clustering

Pipeline

Similar terms

plate: [base, spoon, mounting]

Cluster terms extraction

word2vec

TF-IDF

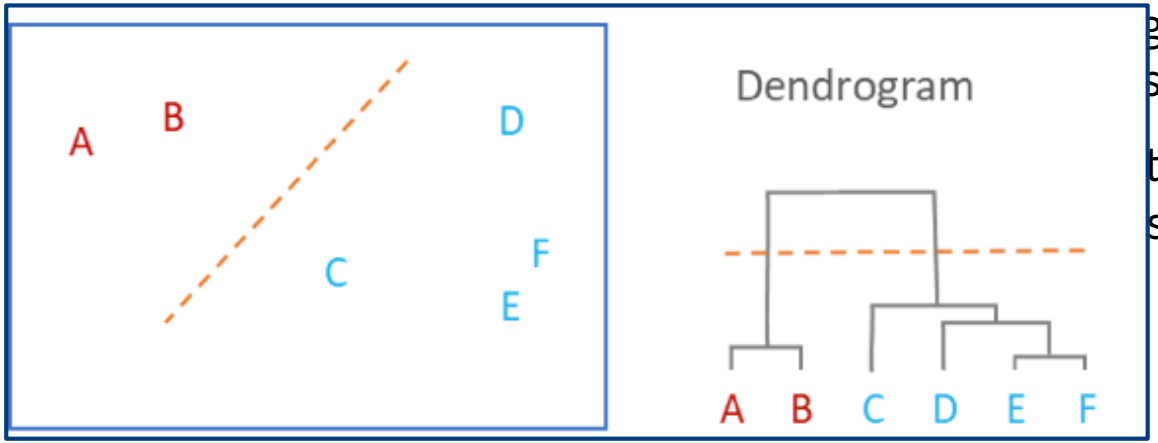
key terms

Most frequent **relevant** document terms

[plate, assembly]

Term augmentation

Dataset



gmented

ster terms

te:

se, mounting]

cluster centers - 3 levels

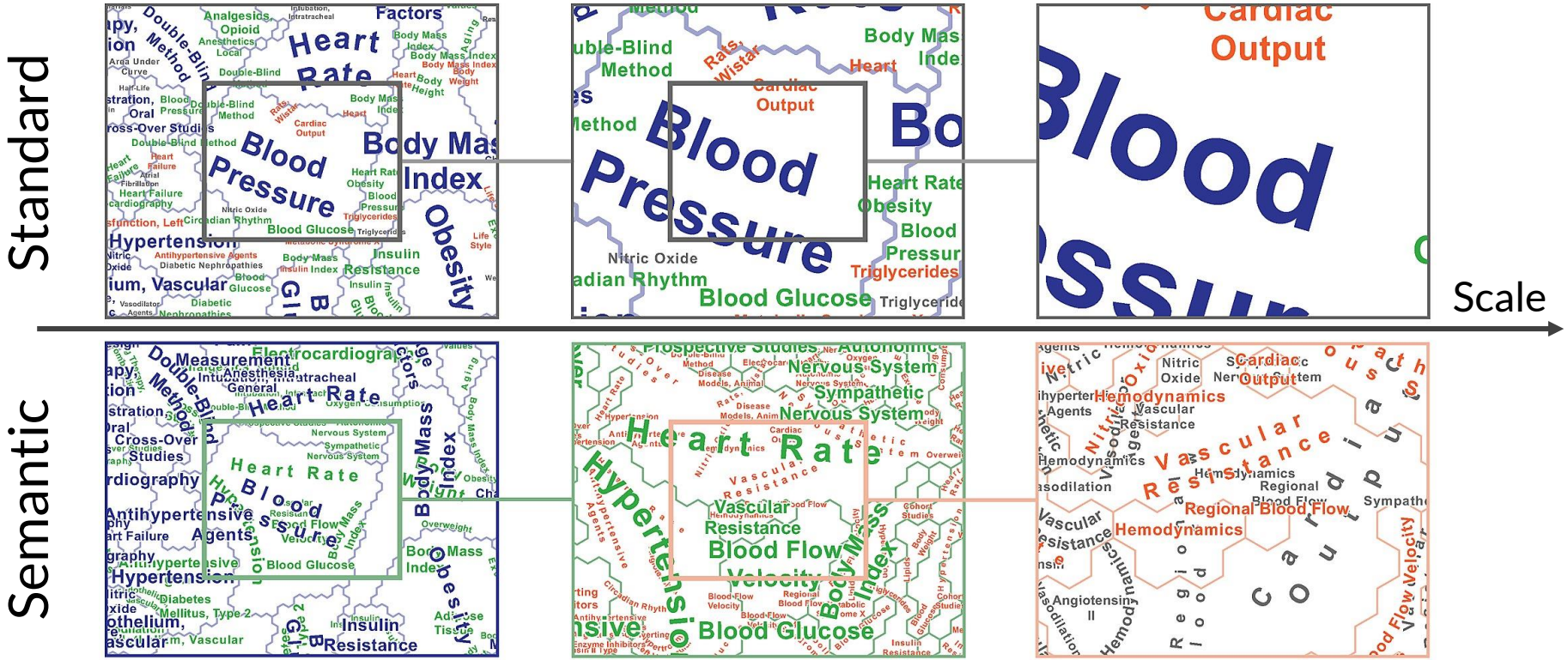
Overview
Background and motivation
Data
Document embeddings
Interaction techniques
Visual elements
Evaluation
Outlook

Visual information seeking mantra

- Overview first
- Zoom and filter
- Details-on-demand

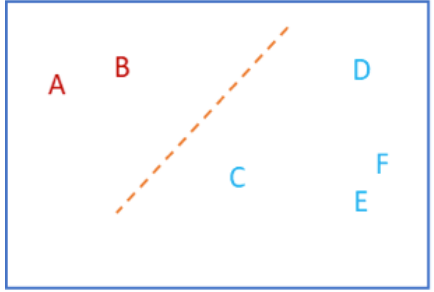
Ben Shneiderman, The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations.
In *Proceedings of the IEEE Symposium on Visual Languages*, pages 336-343, Washington. IEEE Computer Society Press, 1996

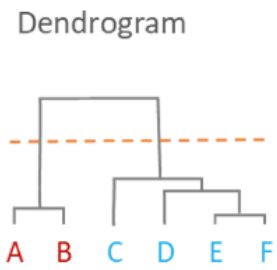
Semantic zoom



Skupin A, Biberstine JR, Börner K (2013) Visualizing the Topical Structure of the Medical Sciences: A Self-Organizing Map Approach. PLoS ONE 8(3): e58779.

Semantic zoom

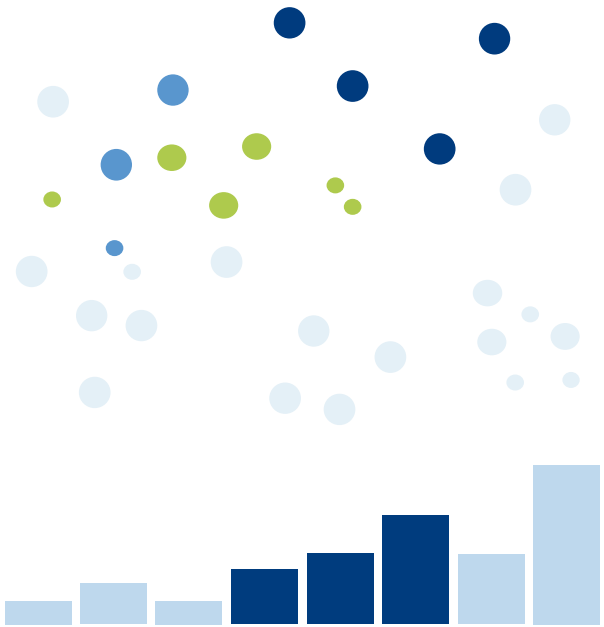




- Different levels of grouping through hierarchical clustering
- More detailed clusters when zooming in

- Semantic approach + interaction techniques allow efficient visual exploration of large datasets. Visual scalability

Brushing and linking

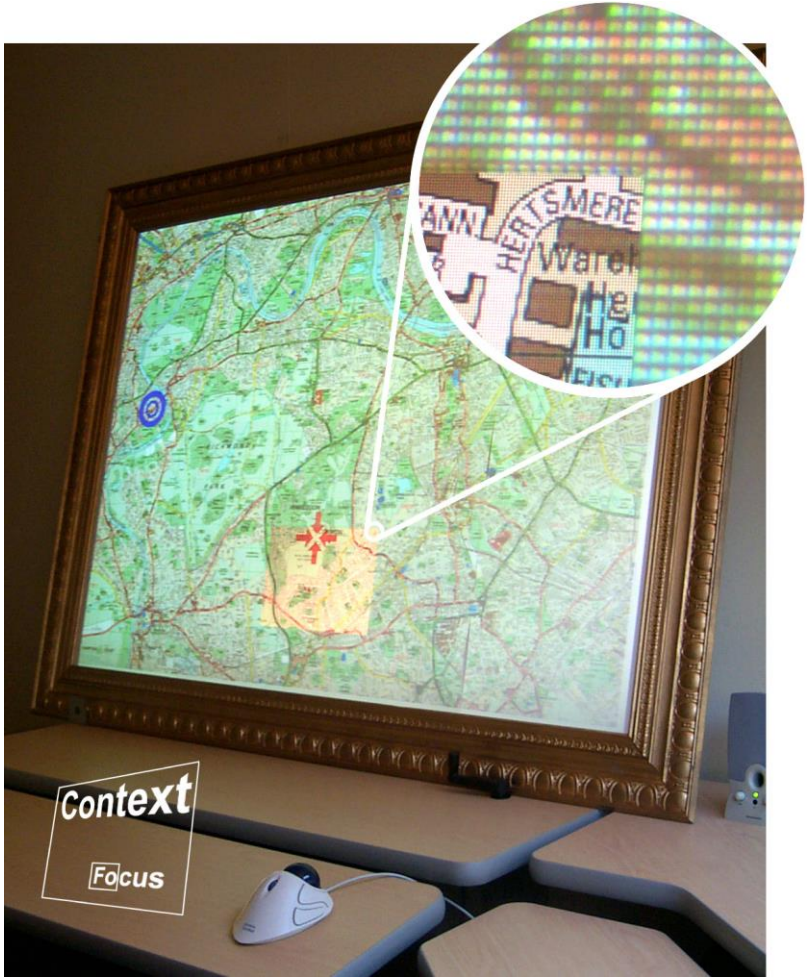


Select data in one view – brushing

Same data highlighted in another view – linking

Focus + context

- Object of interest in detail
- Global view (context) at reduced detail
- Visible simultaneously

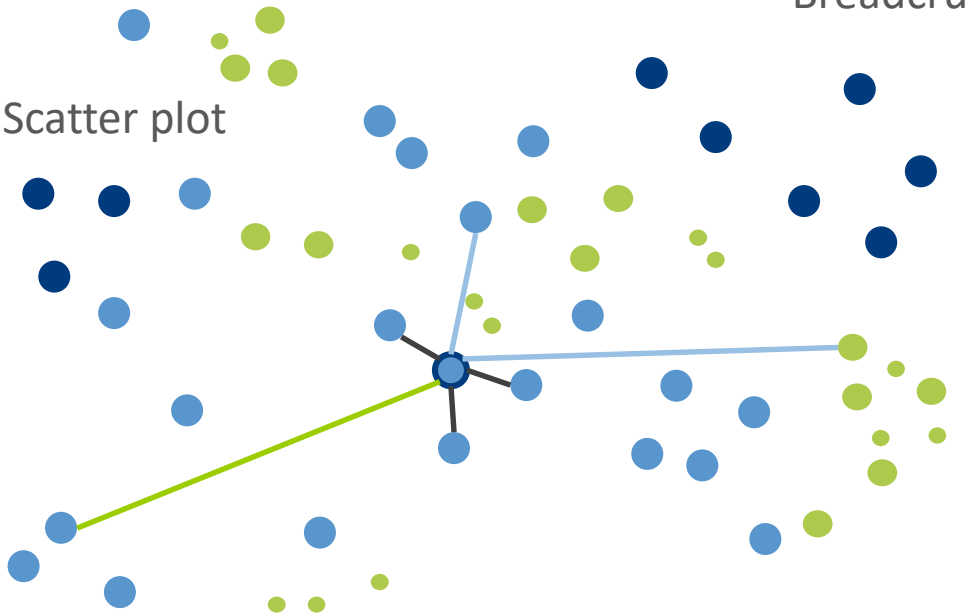


Sources: T. Alan Keahey. Network Visualization Course. Indiana University. 2003
Baudisch, Patrick & Good, Nathaniel & Bellotti, Victoria & Schraedley, Pamela. (2002). Keeping Things in Context: A Comparative Evaluation of Focus Plus Context Screens, Overviews, and Zooming. 10.1145/503376.503423.

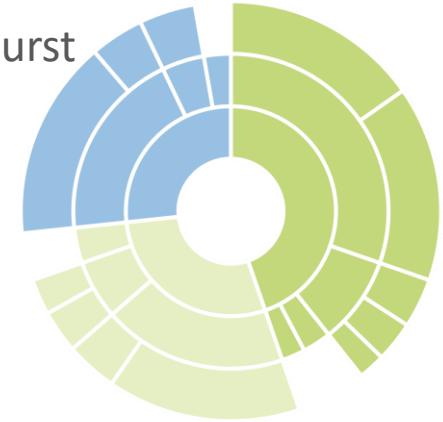
Overview
Background and motivation
Data
Document embeddings
Interaction techniques
Visual elements
Evaluation
Outlook

Breadcrumbs

Scatter plot

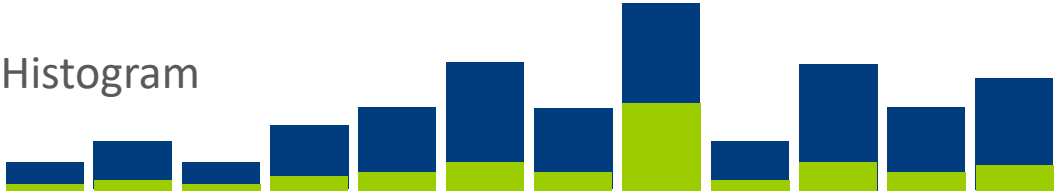


Sunburst



Details

Histogram

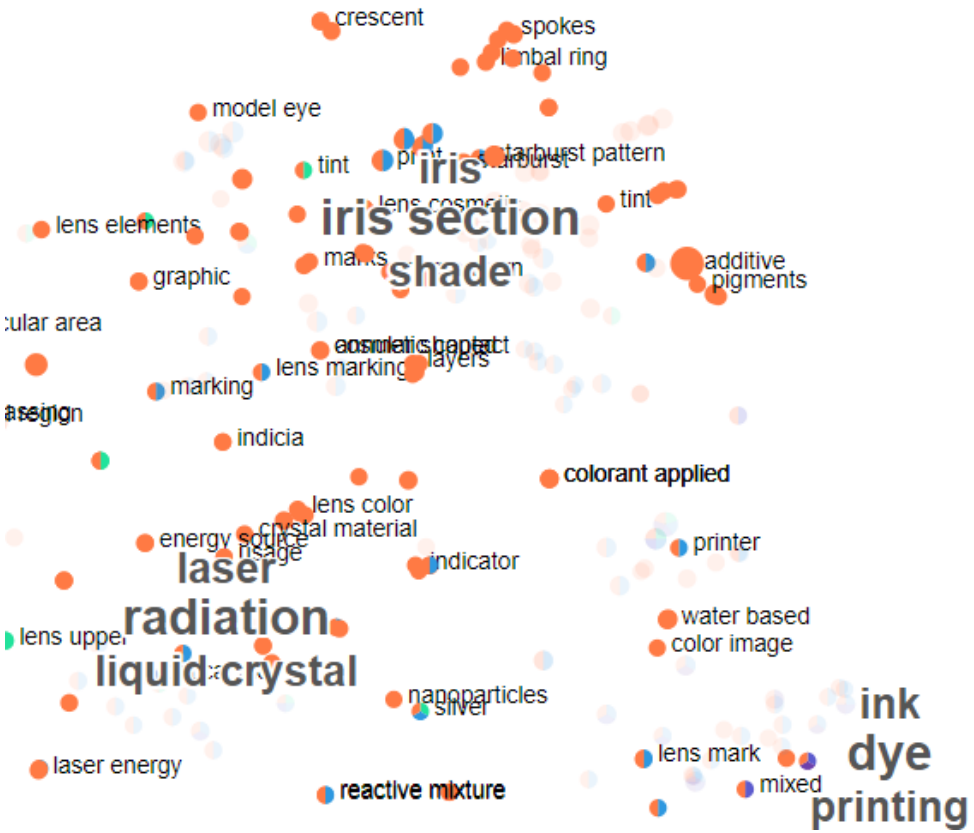


Icon source: Paper by Lil Squid from the Noun Project

Scatterplot

- Depends on other controls
  - Filtering
  - Highlighting

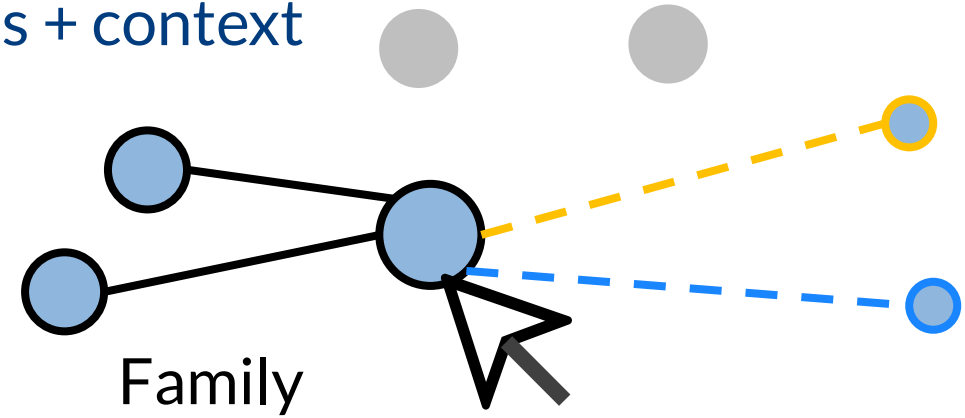
Zoom and filter



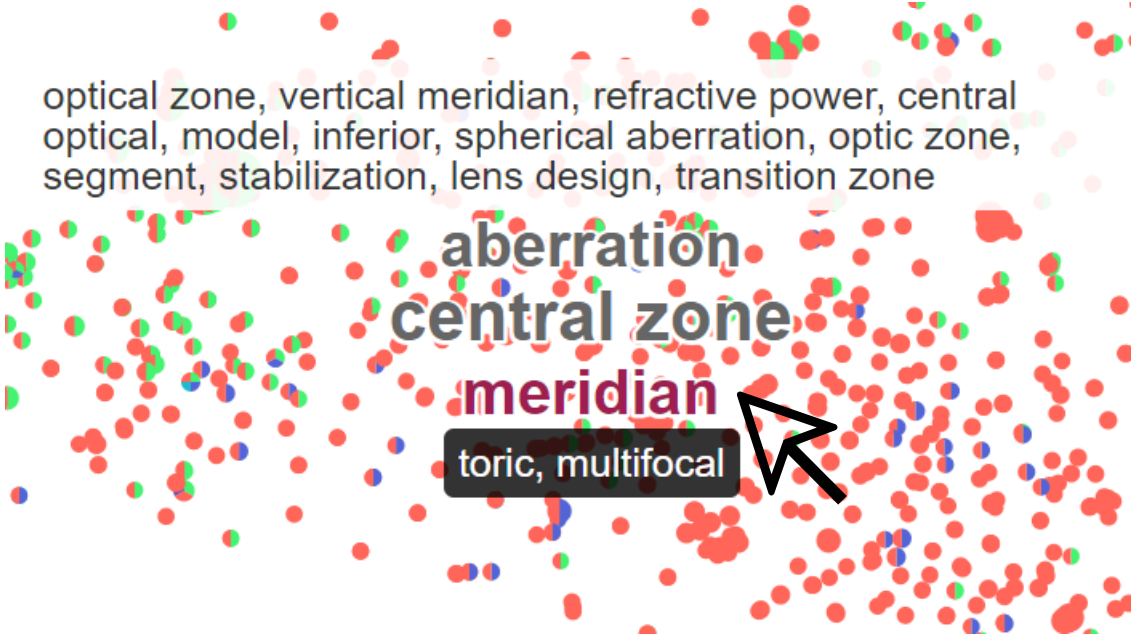
Scatterplot

- Size = number of references
- Color = values within sunburst
- Focus + context

Points belonging to multiple groups:



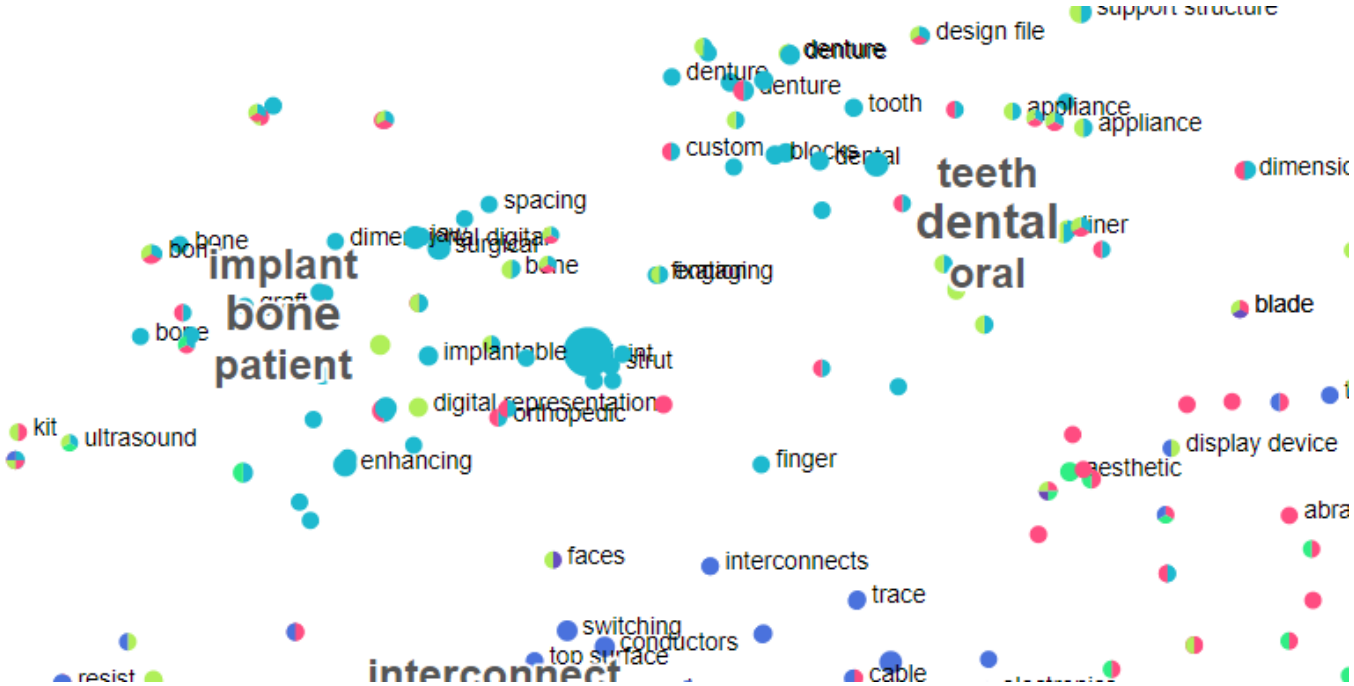
Scatterplot - clusters



Details-on-demand:

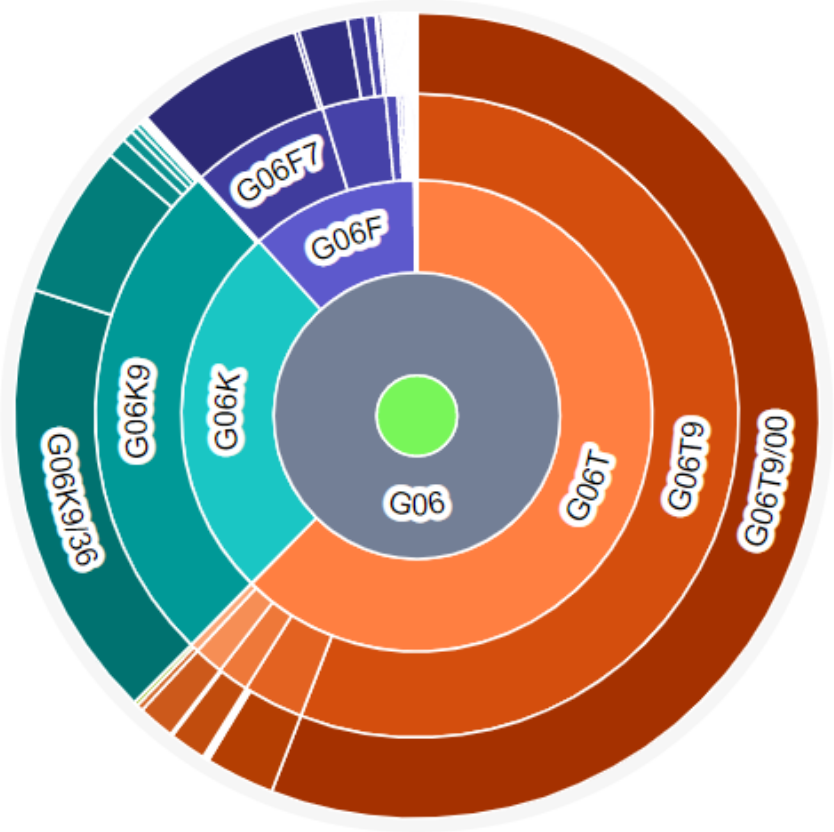
augmented terms

Scatterplot - clusters



Semantic zoom

Sunburst



- Hierarchical pie chart
- Zoomable
- Circular color palette

- Children‘s color: from darker than parent to lighter than parent

- Shows distribution within levels of metadata

Sunburst: variations

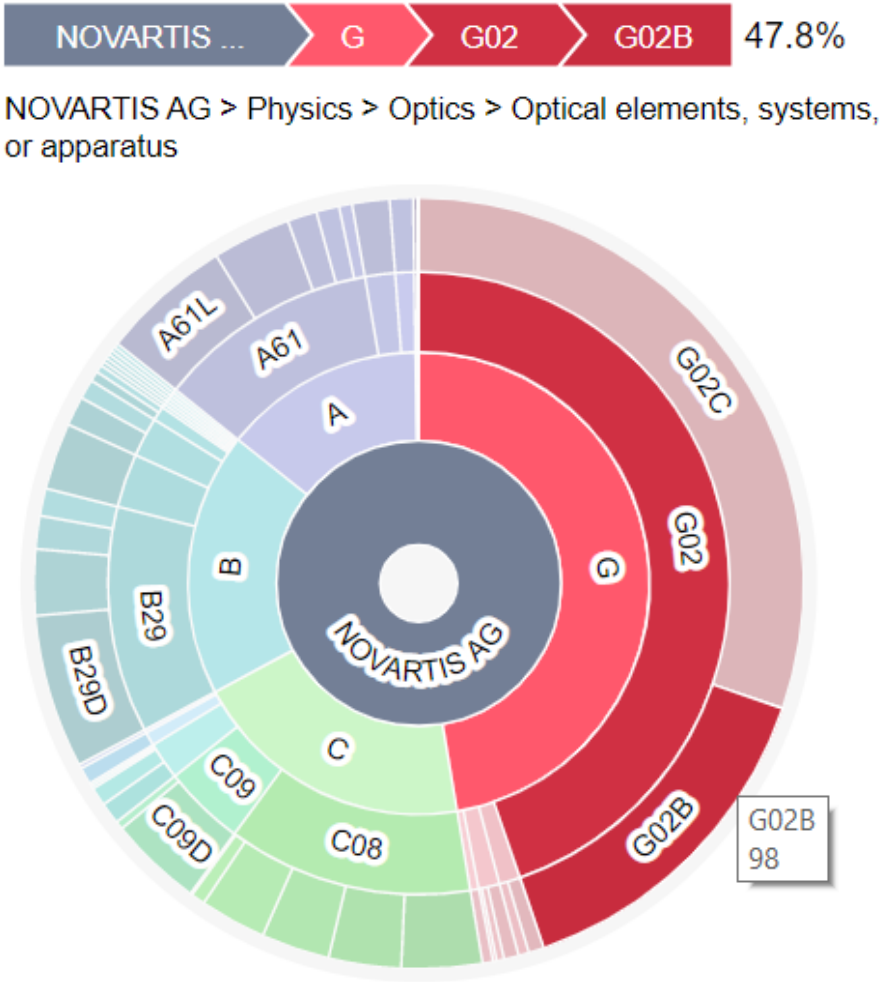
- Competitor analysis:
  - Assignee -> country
  - Assignee -> IPC classes
- Trend analysis
  - Country
  - Country -> IPC classes
  - Country -> assignee

Country -> IPC class

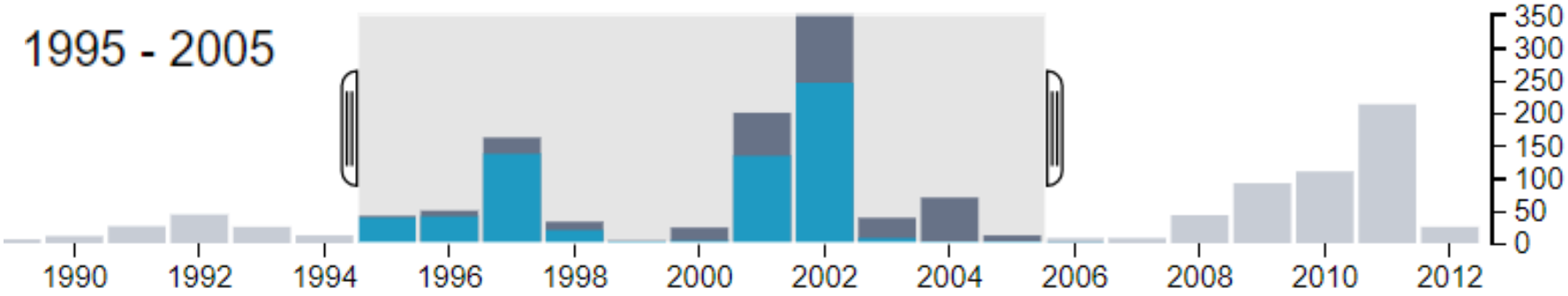
Breadcrumbs

Inseparable from sunburst

- Keep track of previous levels when zoomed in
  - Focus + context
- Hints for IPC code descriptions and long titles
- Show percentage of currently selected group from total
  - Normalization in case of overlaps



Histogram



- Shows whole dataset / current selection in sunburst + current highlight in sunburst
- Allows filtering

- Helps identify trends

Detail view

- All metadata
- Relevant terms
- Persists for a selected patent, temporarily appears for hovered patents
- Details-on-demand

US-9523865-B2 Contact lenses with hybrid power sources

2012.07.26 PLETCHER NATHAN, OTIS BRIAN, VERILY LIFE SCIENCES LLC

Cites 201 Cited by 0 in this dataset

A61B5/00, A61B5/145, G02B7/04, G02C7/04

photovoltaic, types, circuitry, cells, sensing, inductive, supplies, radio frequency, radio, disposed portion

Apparatus, systems and methods of contact lenses with power sources are provided. In some aspects, a contact lens can include a substrate; and a circuit. The circuit can include: one or more sensors disposed on or within the substrate; circuitry disposed on at least a portion of the substrate; one or more photovoltaic cells disposed on at least a portion of the substrate; and a hybrid power component that supplies at least one of two or more different types of power to the circuitry, wherein at least one of the two or more different types of power is radio frequency/inductive power. In various aspects, other types of power can be solar and/or microelectromechanical system power. Additionally, in various aspects, photovoltaic cells can be

Overview
Background and motivation
Data
Document embeddings
Interaction techniques
Visual elements
Evaluation
Outlook

Studies with patent experts

- Formative study: **semi-structured** user interviews

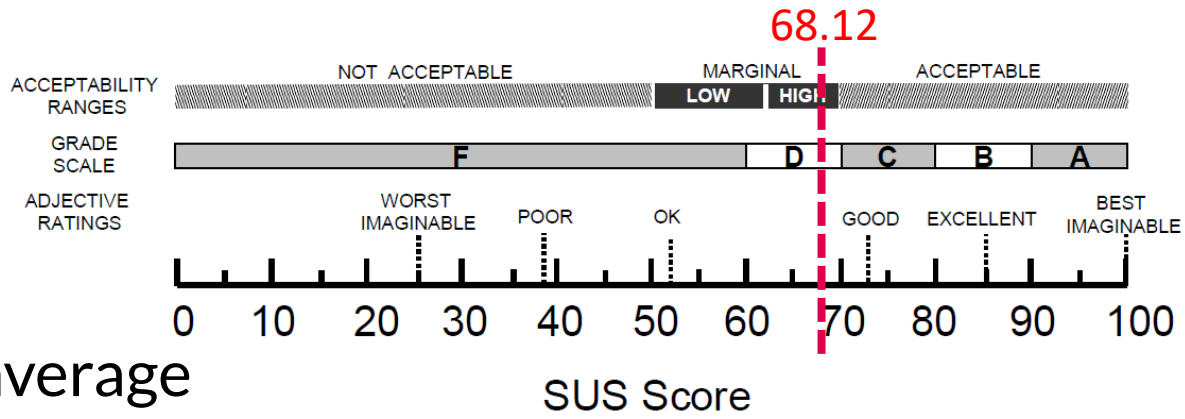
- Summative study:
  - Uncover usability problems
  - Compare semantic embeddings to a traditional approach - TF-IDF document representation

Process
  - Think-aloud tasks
  - System Usability Scale questionnaire
  - Questionnaire for comparison of two approaches

Summative study

Results

- 7 hypotheses
  - 1 refuted
  - 1 partly confirmed
  - 4 confirmed
  - 1 likely confirmed



- SUS score 68.12 – average
  - success for “just” a prototype

Image Source: A. Bangor, P. Kortum, and J. Miller, “Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale,” *J. Usability Stud.*, vol. 4, no. 3, pp. 114–123, 2009.

Summative study

Takeaways

- Minor performance / usability problems uncovered
- Visualization metaphors fit the task and were understood
- Intuitive – requires little training
- Best suited for general overview.
  - Detail-oriented tasks benefit from standard tools: co-occurrence matrices, text search, etc.

Summative study

TF-IDF vs. word2vec document representations

- UI and interactions have much more impact than positions of documents
- Very similar cluster key terms
  - More so for larger clusters
- Semantic embeddings
  - More intuitive placement of clusters
  - Better separation of clusters

  subjective
- **Quantitative** evaluation necessary!

Overview
Background and motivation
Data
Document embeddings
Interaction techniques
Visual elements
Evaluation
Outlook

Outlook

Things to think about

- Automatic detection of suitable hierarchical clustering levels
- Improved performance - graceful degradation
- Other key term extraction methods - TextRank, RAKE etc
- Other embedding methods - paragraph2vec, ELMo, BERT etc
  - Represent different parts of document separately
- Control over parameters of dimension reduction - esp. perplexity
- Hypernyms and hyponyms

Other ideas

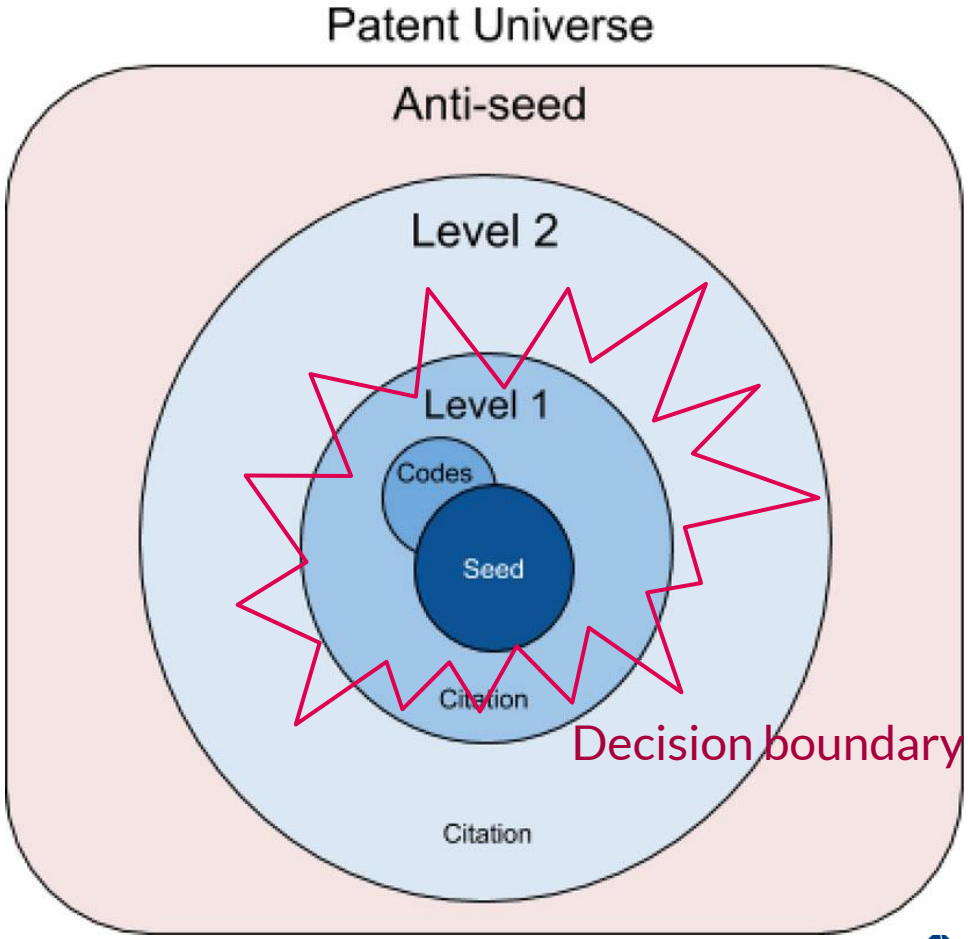
- On-the-fly summary of dynamically selected area
- Prevent overlapping points
- Group small assignees to prevent visual clutter

Icon source: Lasso Tool by icon 54 from the Noun Project

Augment dataset

- Expand
  - Follow reference links and common class codes
- Prune
  - Classify: seed vs. anti-seed
    - Wide-and-deep LSTM
    - Wide: one-hot encoded codes and reference ids
    - Deep: word2vec embeddings

Model not generalizable!



Source: A. Abood and D. Feltenberger, "Automated patent landscaping," *Artif. Intell. Law*, vol. 26, no. 2, pp. 103–125, 2018.

Try it out at

http://patsemtech.fiz-karlsruhe.de

and let’s talk about it

Thank you

Tatyana Skripnikova
Data Scientist

generic.de software technologies AG
Zeppelinstraße 15
76185 Karlsruhe

tatyana.skripnikova@generic.de