Visual Analysis of Textual and non-Textual Documents

Prof. Dr. Tobias Schreck
Institute for Computer Graphics and Knowledge Visualization

UX Day Graz 2015, 14.11.2015

Large, Complex Data

Technological progress

- Acquisition, production, storage
- Data integration, data mining

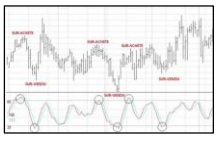
→ Large and increasing amounts of data

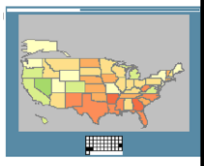
Data-intensive application domains

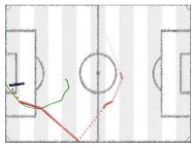
- Business, Research
- Biomedical Engineering
- Social Media

User Tasks

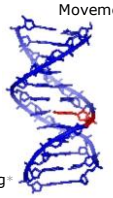
- Search for data items
- Explore for patterns of interest
- Communicate


Finance and Business*


Spatiotemporal Data

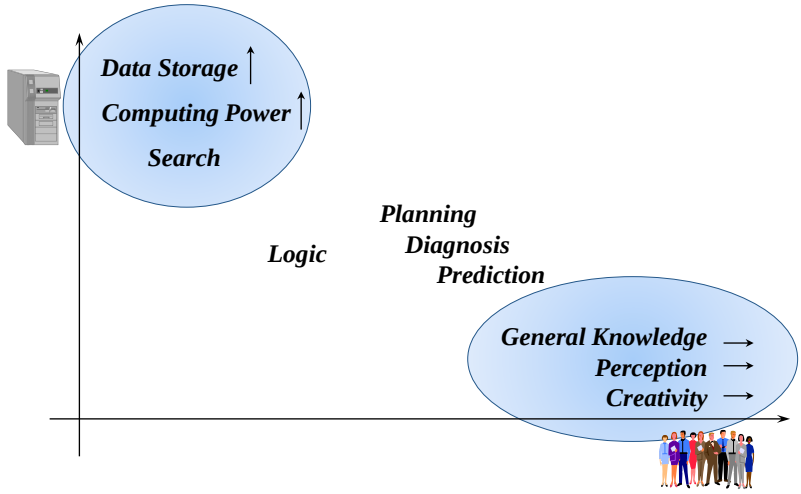

Movement Data

Social media data


Biomedical Engineering*

(*) Wikimedia Commons

Abilities of Humans and Computers



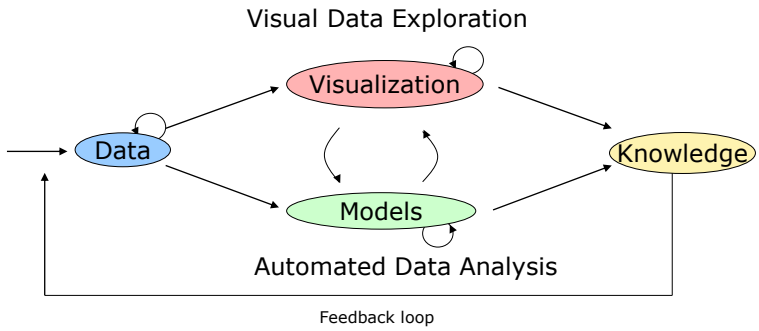
Slide courtesy of Daniel Keim

Why Interactive Visualization ?

- Automated techniques not sufficient
  - Data ambiguous and incomplete
  - Complex relationship
  - Semantic gaps
  - Limited Accuracy
- Visual-interactive access central for
  - Exploration of Data
  - Generation of Hypotheses
  - Interpretation of Results
  - Steering of the Analysis

Visual Analytics Process

Tight Integration of Visual and Automatic Data Analysis Methods for Information Exploration and Scalable Decision Support



- Introduction
- Visual Analysis of Textual Data
- Social Media Data
- Visual Data Retrieval
- Conclusions

Sentiment Analysis

- Opinion score derived from adjectives, nouns, and verbs
- Identifies positive and negative sections

→ Overview over large document corpora

→ Find articles which suit the mood of the reader

D. A. Keim, F. Mansmann, D. Oelke and H. Ziegler. **Visual Analytics: Combining Automated Discovery with Interactive Visualizations**. Proceedings of the 11th International Conference on Discovery Science (DS 2008), Springer-Verlag, pages 2-14, 2008

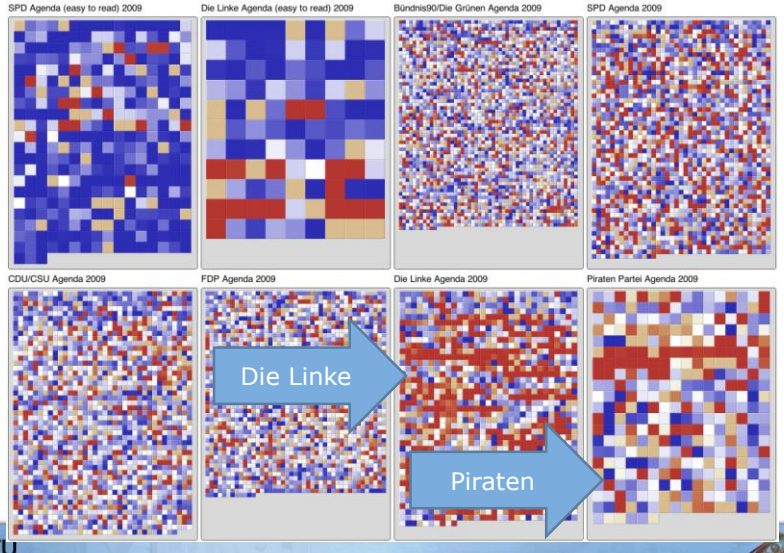
Sentiment Analysis: News Overview

Readability Features

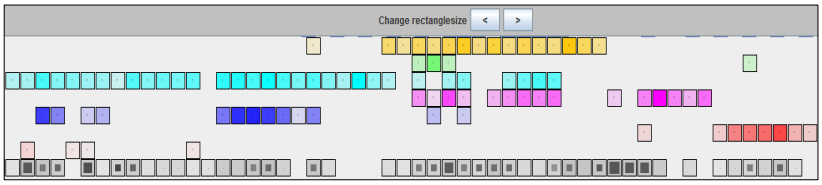
	Voc. Difficulty	Word Length	Nominal Forms	Sent. Length	Compl. Sent. Struc.
Analysis of word frequencies: Large document collections such as the Project Gutenberg (http://www.gutenberg.org/) or Wikipedia (http://www.wikipedia.com) allow to calculate the average usage frequency of a word 1 We exploited those resources to determine how common the words of a text sample on average are.					
This measure is related to the one already proposed in [16], following the assumption that parts of the sentence that are interrupted by subordinate sentences or parenthesis have to be stored in a temporary memory which increases the mental complexity of processing the sentence.					
The implementation of 141 different simple text features allows us an unbiased search for text features with high expressiveness with respect to readability.					
Die Literaturangabe in der Bibtex Datei muss noch vervollständigt werden!					
Among the most popular ones are the Flesch-Kincaid Readability Test [12], Flesch Reading Ease [7], SMOG [13], the Coleman-Liau-Index [4], and Gunning Fog [8].					

D. Oelke, D. Spretke, A. Stoffel and D. A. Keim. **Visual Readability Analysis: How to make your writings easier to read**. Proceedings of IEEE Conference on Visual Analytics Science and Technology (VAST '10), pages 123 - 130, 2010.

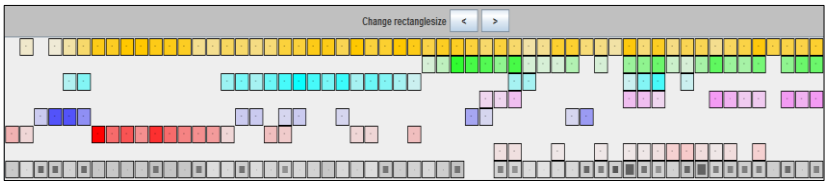
Vocabulary Difficulty (2009 German Election Programs)



Attribute-based: Story, Character Complexity



King's *IT*

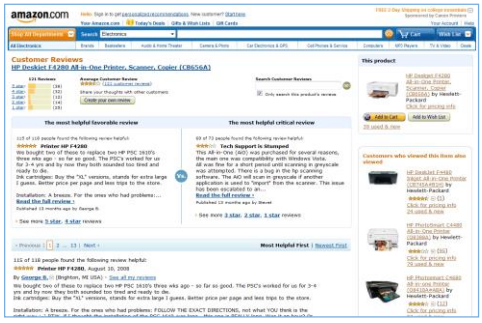


Rowling's *Harry Potter*

F. Wanner, J. Fuchs, D. Oelke and D. A. Keim. **Are my Children Old Enough to Read these Books? Age Suitability Analysis**. POLIBITS - Research journal on Computer science and computer engineering with applications, 2011.

Attribute-based: Visual Review Analysis

- User opinions abundantly available
  - Forums, Blogs
  - E-commerce
  - …
- Many application possibilities
  - Product reviews for customers
  - Market analysis
  - Customer relationship management



Amazon customer reviews
(amazon.com)

Attribute-based: Visual Review Analysis

- Basic method
  - Identify product attributes
  - Identify positive/negative opinions
  - Calculate weighted attribute vector
- Visual comparison of sets of reviews
  - Glyph matrix approach
  - Cluster analysis
- Applied to printer product reviews

I feel obligated to counter the bad reviews.

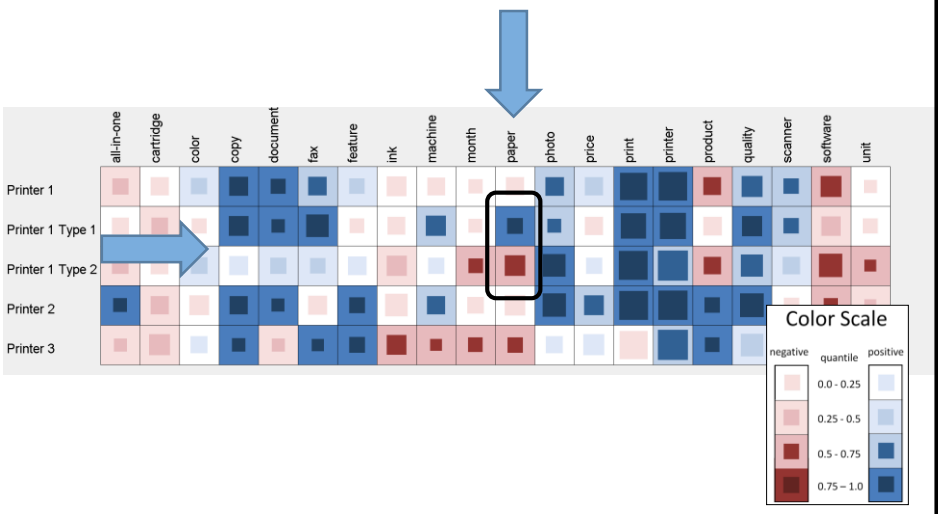
This printer is just fine.

I don't know what people are complaining about regarding the software but it installed seamlessly and is intuitive in its operation.

Even though the paper tray jams sometimes altogether I am happy that I bought this wonderful printer.

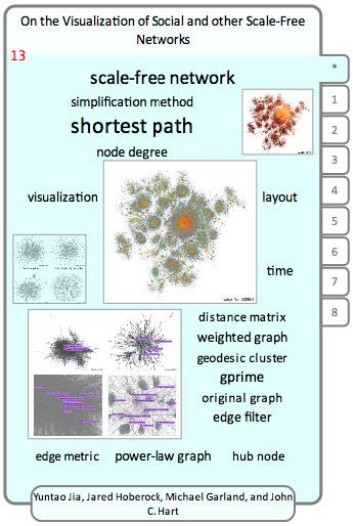
cartridge	paper tray	price	printer	scanner	software
0	-1	0	+1	0	+1

Attribute-based: Visual Review Analysis



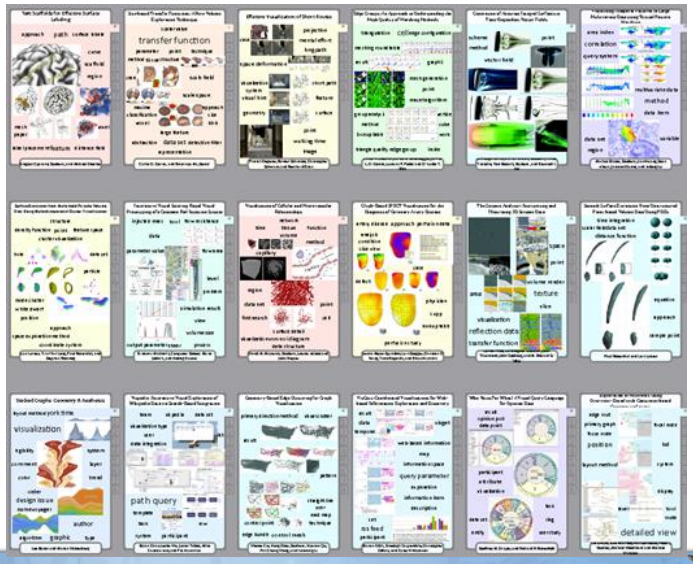
Visual Content Overviewing

- Visual abstract for scientific articles
  - Extraction of important figures and keyword
  - Layout of elements in generalized word cloud
- Overviewing
- Navigation
- Comparison

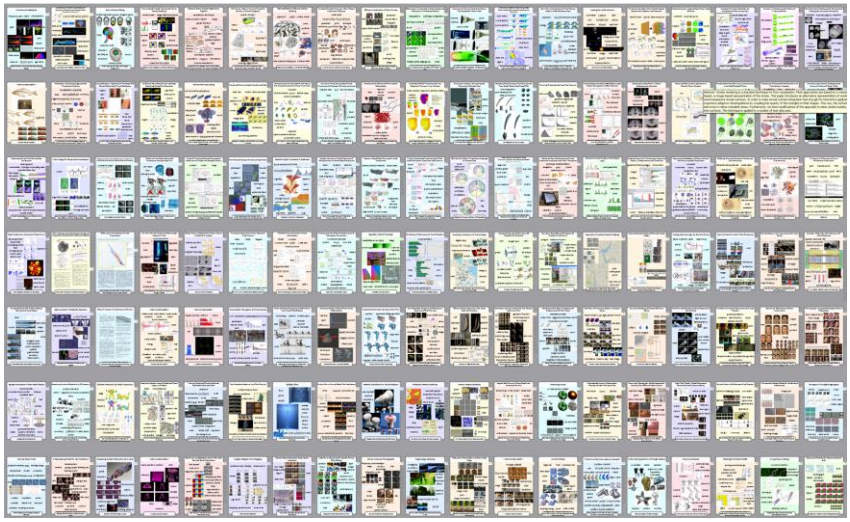


H. Strobelt, D. Oelke, C. Rohrdantz, A. Stoffel, D. A. Keim and O. Deussen. **Document Cards: A Top Trumps Visualization for Documents**. IEEE Transactions on Visualization and Computer Graphics, 15(6):1145-1152, 2009.

Visual Content Overviews



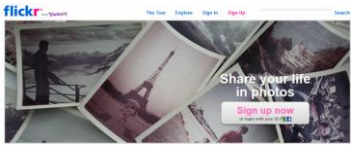
Visual Content Overviews



- Introduction
- Visual Analysis of Textual Data
- Social Media Data
- Visual Data Retrieval
- Conclusions

Visual Analysis of Social Media Data

- Social media important data
  - Content (text, visual, geospatial, …)
  - Metadata (location, time, user, …)
- Cases for analysis
  - Marketing
  - Sentiment mining
  - Disaster response, security

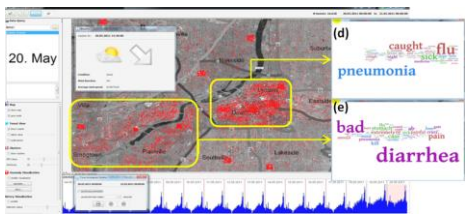


T. Schreck and D. Keim. ***Visual analysis of social media data***. IEEE Computer, Special Issue on Cutting-Edge Research in Visualization, 46(5):68–75, May 2013.

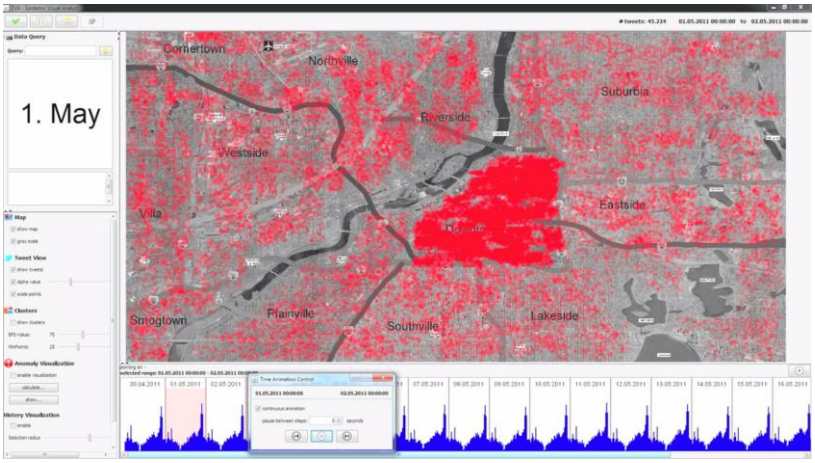
T. Schreck and D. Keim. ***Visual analysis of social media data***. IEEE Computer, Special Issue on Cutting-Edge Research in Visualization, 46(5):68–75, May 2013.

VAST Challenge 2011: Epidemic Spread

- **Scenario**
  - Metropolitan city with street network and POIs
  - 1M geo-located microblog messages
  - Fictitious hidden epidemic scenario
- **Task**
  - Find possible epidemic spread and its characteristics

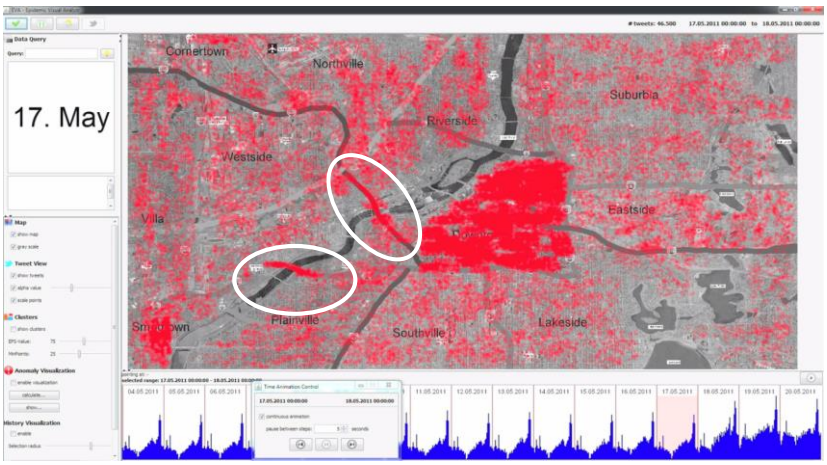


VAST Micro Blogging Challenge

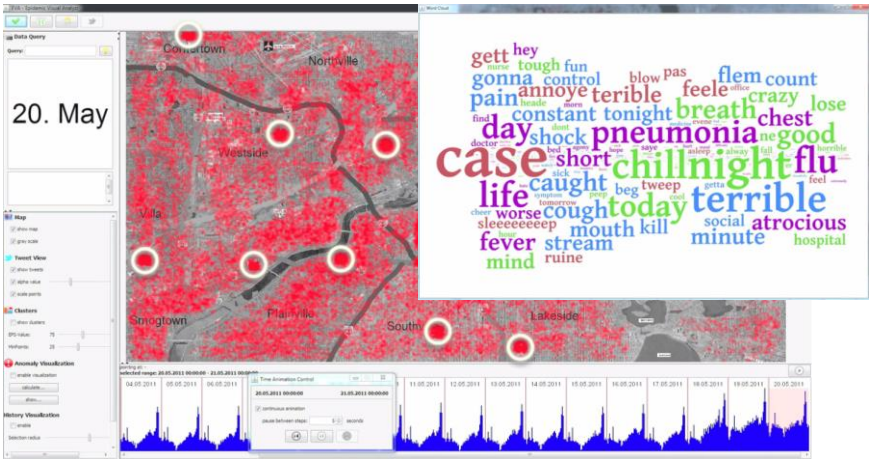


[BBF+2011]

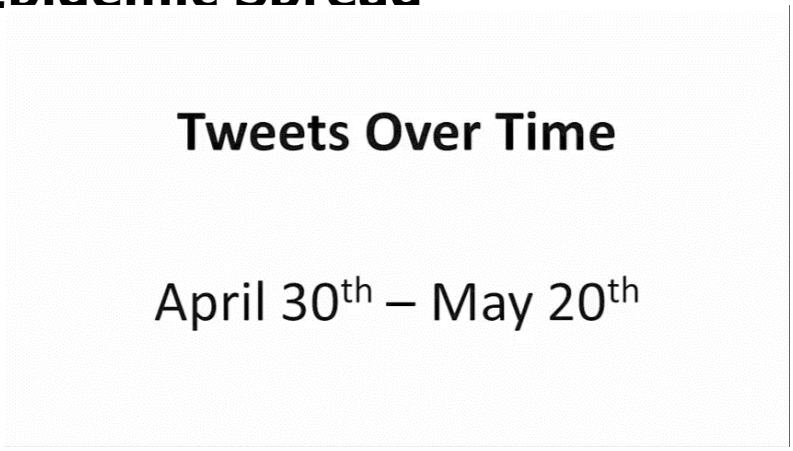
Concentration on Bridges



Concentration in Hospitals

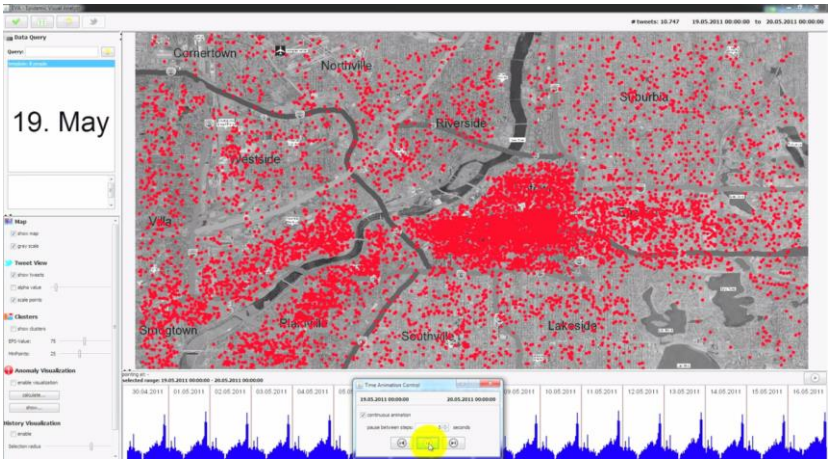


VAST Challenge 2011: Epidemic Spread

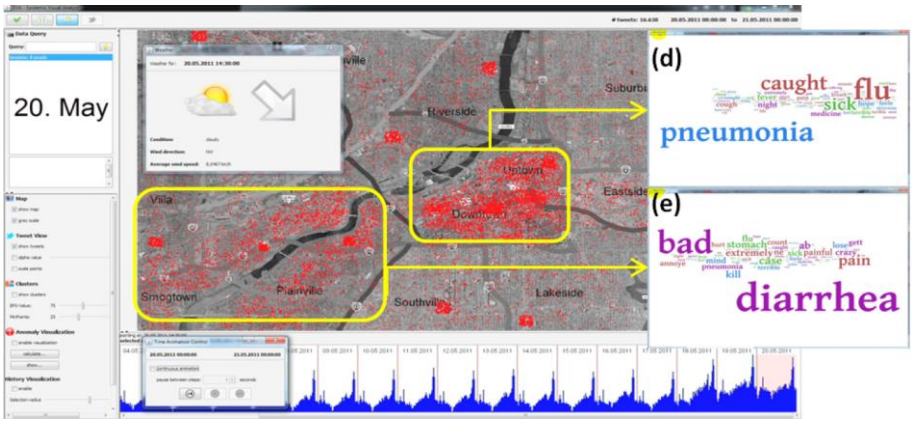


E. Bertini, J. Buchmüller, F. Fischer, S. Huber, T. Lindemeier, F. Maaß, F. Mansmann, T. Ramm, M. Regenscheit, C. Rohrdantz, C. Scheible, T. Schreck, S. Sellien, F. Stoffel, M. Tautzenberger, M. Zieker, and D. Keim. ***Visual analytics of terrorist activities related to epidemics (VAST 2011 Grand Challenge Award)***. In Proc. IEEE Symposium on Visual Analytics Science and Technology, pages 329–330, 2011. Contest report paper.

Message Distribution (19.05.)
– Filtered for Symptom Keywords



VAST Micro Blogging Challenge

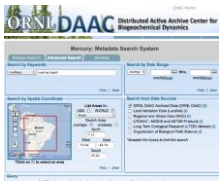


- Introduction
- Visual Analysis of Textual Data
- Social Media Data
- Visual Data Retrieval
- Conclusions

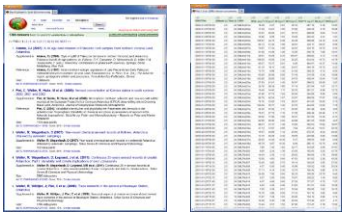
Visual Search and Analysis in Research Data

- Data generation in scientific process
- Data libraries
  - Serve and preserve
  - Transparency and reproducibility
  - 4th Paradigm [Gray]
- User access
  - Mostly, meta-data based
  - Meta data expensive to curate
  - Lack of content-based search

Sloan Digital Sky Survey Repository



Oakridge DAAC data archive

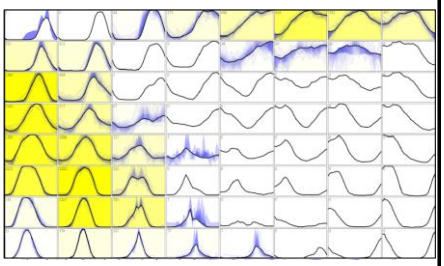


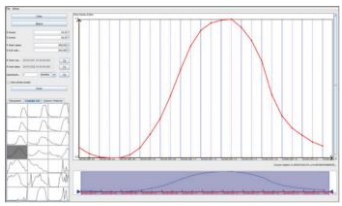
AWI PANGAEA web data repository

Visual Search in Research Data

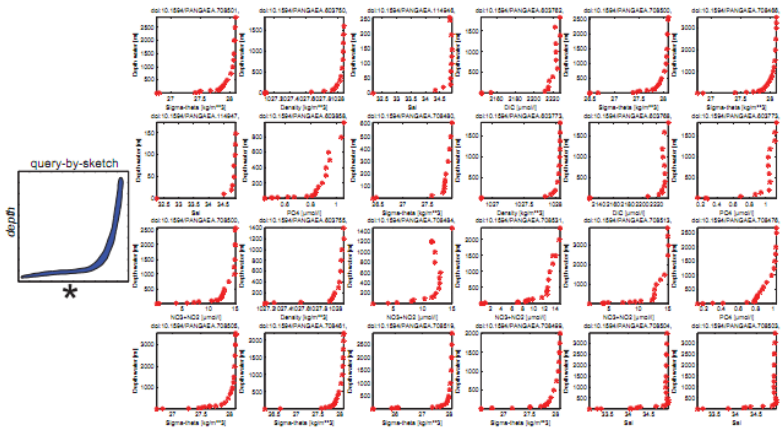
VisInfo
VISUELLER ZUGANG ZU FORSCHUNGSDATEN

- **Visual Exploration and Search in Time Series**
  - Feature-based similarity function
  - Overviews using Self-Organizing Map algorithm
  - Retrieval using example data or query editor
  - Prototype operated by TIB Hannover for Pangaea data



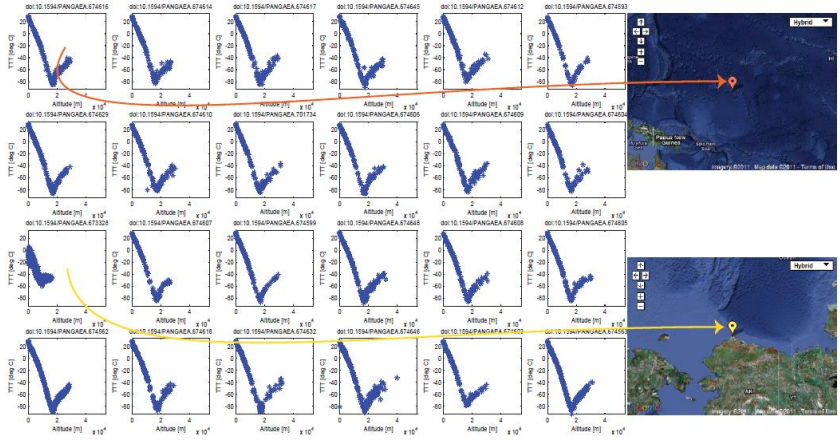


Example Query-by-Sketch



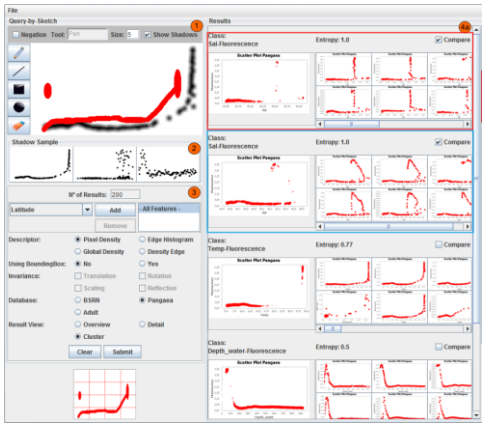
Explorative Search Interface

Adding geospatial reference



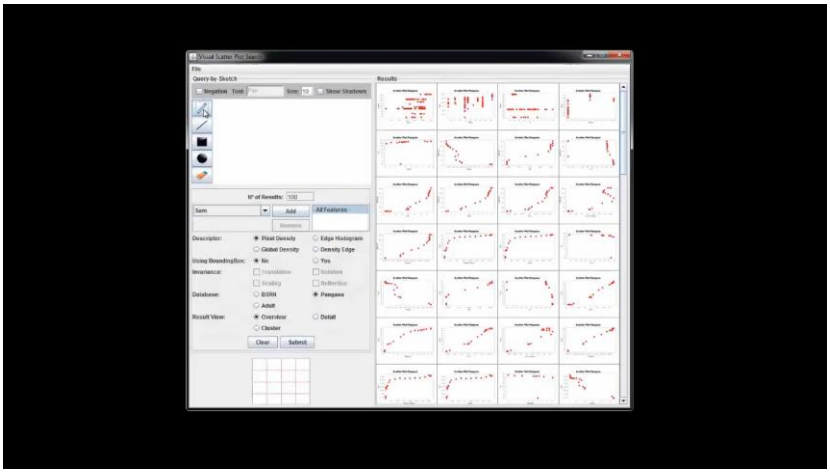
Support for Query-as-you-Sketch

- **Visual retrieval in scatter plot spaces**
  - Sketch interface to query for patterns of interest
  - Online matching searches candidates after each stroke
  - Shape suggestions on the fly



L. Shao, M. Behrisch, T. Schreck, T. von Landesberger, M. Scherer, S. Bremm and D. A. Keim. **Guided Sketching for Visual Search and Exploration in Large Scatter Plot Spaces**. Proc. EuroVA International Workshop on Visual Analytics, 2014.

Support for Query-as-you-Sketch



- Introduction
- Visual Analysis of Textual Data
- Social Media Data
- Visual Data Retrieval
- Conclusions

Conclusions

Summary

- Large and complex data arising
  - Social media, web
  - Open data
  - Etc. etc.
- Explore, retrieve, make sense, communicate
- Visual data analysis
  - Data mining/analysis
  - Visualization
  - Interaction

Challenges

- Widening the user base
  - Expert systems vs. visualization for the masses
- Visual literacy
- (Visual/online) journalism
- User experience and data visualization

Acknowledgments

TU Graz
Fraunhofer IGD
Universität Konstanz
Data Analysis and Visualization
Baden-Württemberg.de
DFG
Graduate School Explorative Analysis and Visualization of Large Information Spaces
SEVENTH FRAMEWORK PROGRAMME

+ all collaborators and students