Visual Analytics

Intro to Information Visualization

Spring, 2019

Alex Endert

Agenda

- Overview of what the term means and how it relates to information visualization
- Example VA systems

Visual Analytics

Definition and Terminology

Before there was VA

- Growing concern from some that infovis was straying from practical, real world analysis problems
- Infovis typically not applied to massive data sets
  - "show me the data" has scalability limitations
- Infovis "competes" with other computational approaches to data analysis
  - Statistics, data mining, machine learning, broadly speaking "analytics"

Recommendations

- Integrate data mining and information visualization
- Allow users to specify what they are seeking
- Recognize that users are situated in a social context
- Respect human responsibility

The InfoVis Pipeline

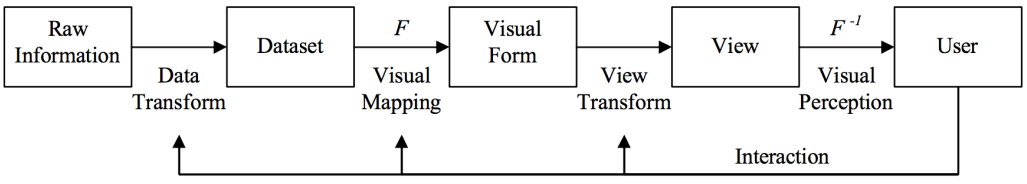
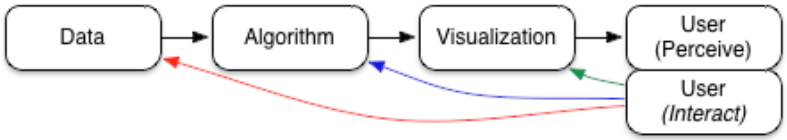
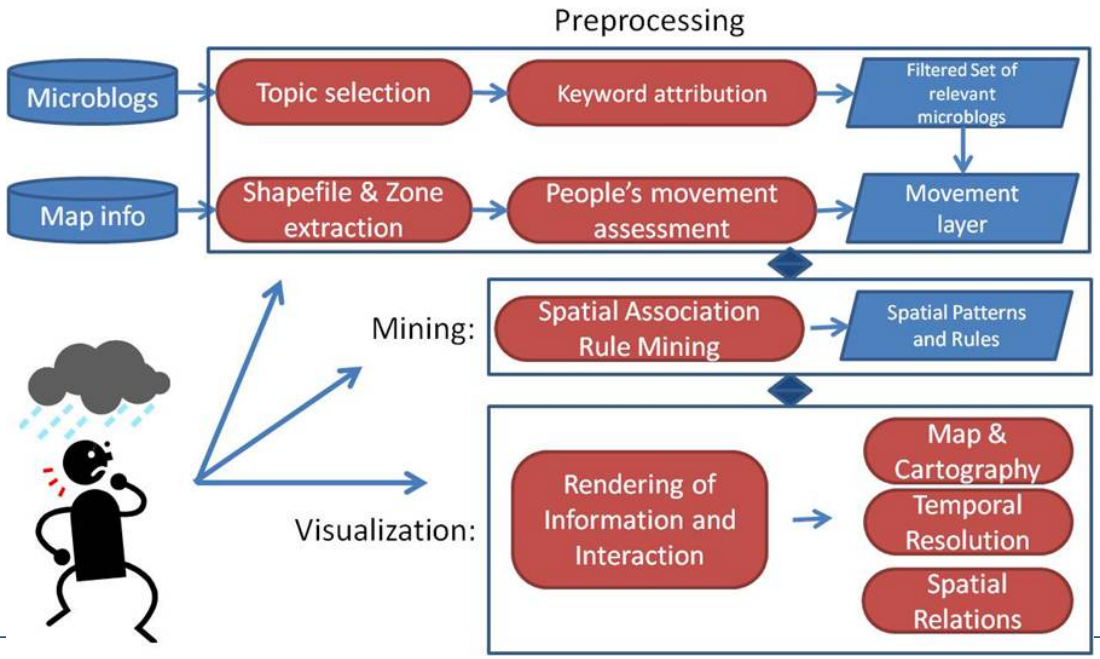


Figure 2: The visualization pipeline, converting information into interactive visual representations (adapted from [Card et al., 1999]).

Visual Analytics Pipeline

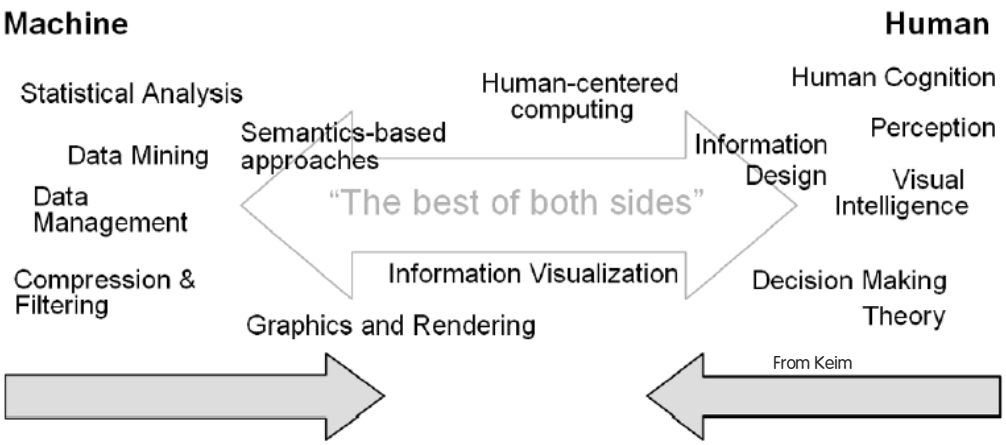


Visual Analytics Pipeline :/



Synergy

- Combine strengths of both human and electronic data processing
  - Gives a semi-automated analytical process
  - Use strengths from each



Integrating Data Mining and ML

- Dimension Reduction
  - PCA, MDS, …
- Feature Selection
  - Entity extraction, …
- Similarity functions
  - "find more like this", …
- Prediction
  - will the stock price go up or down?

More Motivation

- Increasing occurrences of situations and areas with large data needing better analysis
  - DNA, microarrays, sequence mining
  - Business intelligence
  - prediction

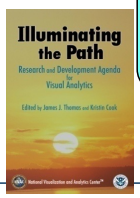
History

- 2003-04 Jim Thomas of PNNL, together with colleagues, develops notion of visual analytics
- Holds workshops at PNNL and at InfoVis ‘04 to help define a research agenda
- Agenda is formalized in book Illuminating the Path, shown on next slide

Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces.

People use visual analytics tools and techniques to

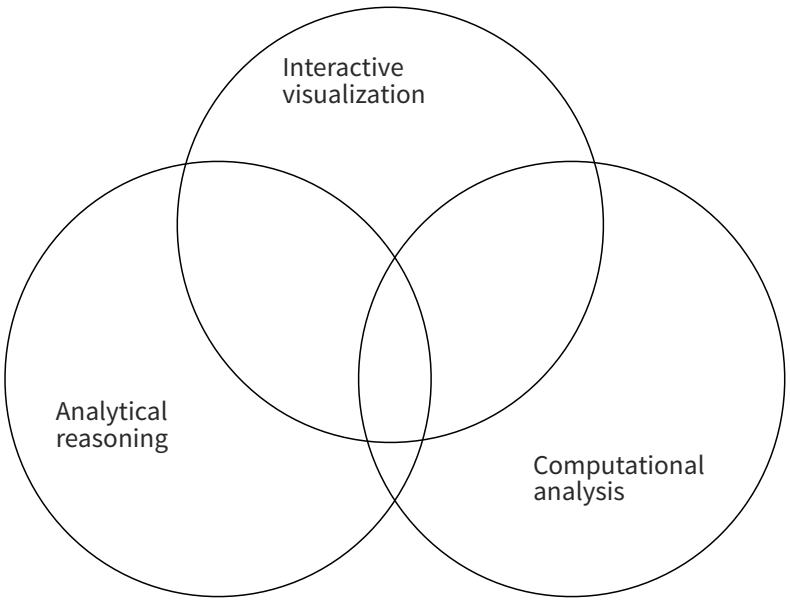
- Synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data
- Detect the expected and discover the unexpected
- Provide timely, defensible, and understandable assessments
- Communicate assessment effectively for action.



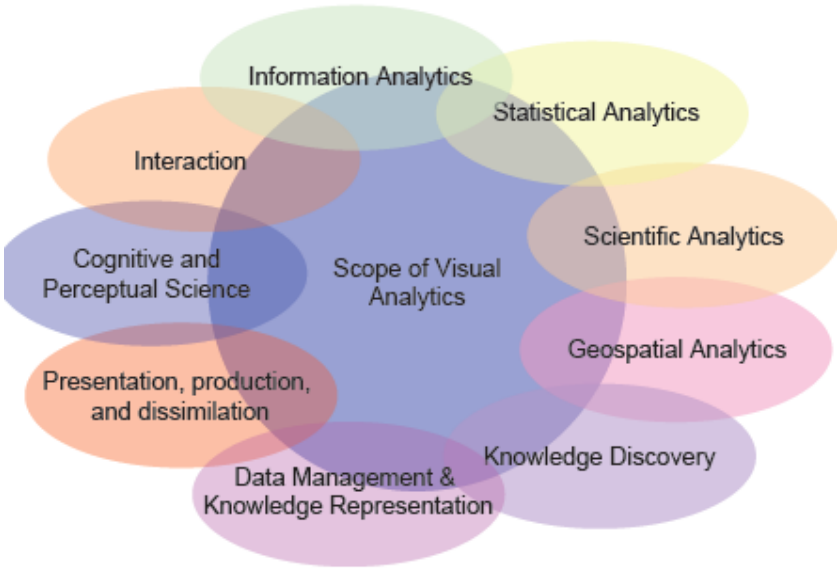
"The beginning of knowledge is the discovery of something we do not understand."
~Frank Herbert (1920 - 1986)

Thomas & Cook
2005

Main Components



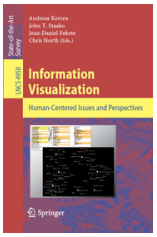
Synergy



From Keim

Alternate Definition

- Visual analytics **combines automated analysis techniques with interactive visualizations** for an effective understanding, reasoning and decision making on the basis of very large and complex data sets



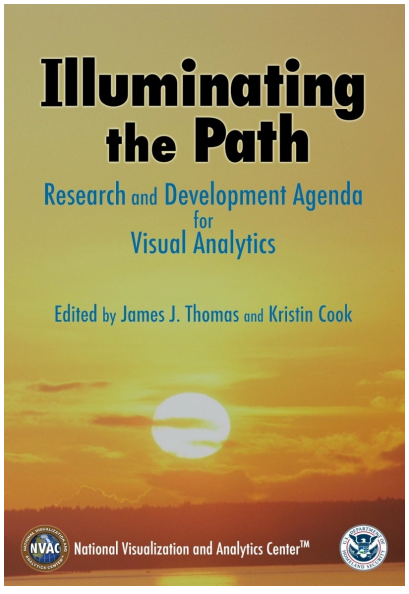
Keim et al, chapter in Information Visualization: Human-Centered Issues and Perspectives, 2008

Visual Analytics

- Encompassing, integrated approach to data analysis
  - Use **computational algorithms where helpful**
  - Use **human-directed visual exploration** where helpful
  - Not just “Apply A, then apply B” though
  - **Integrate the two tightly**

- Available at http://nvac.pnl.gov/ in PDF form
- At IEEE Press in book form

- Special thanks to IEEE Technical Committee on Visualization and Graphics

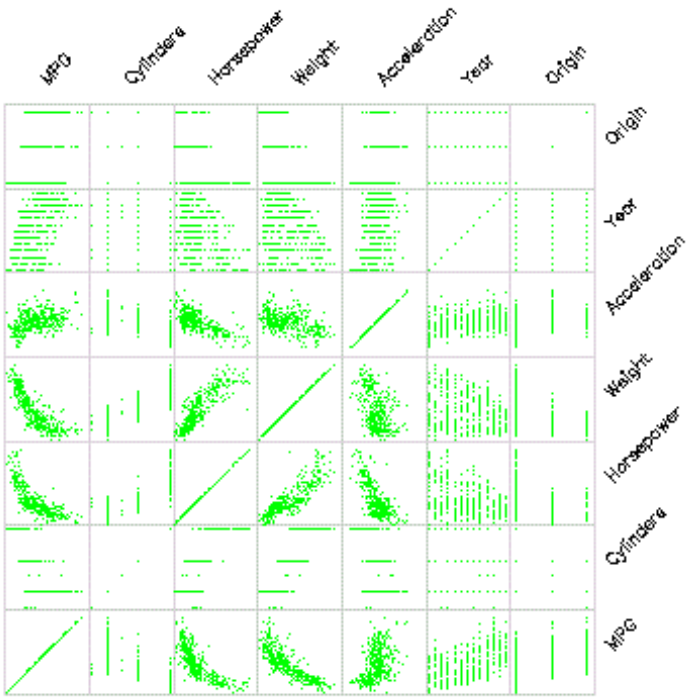


a quick example

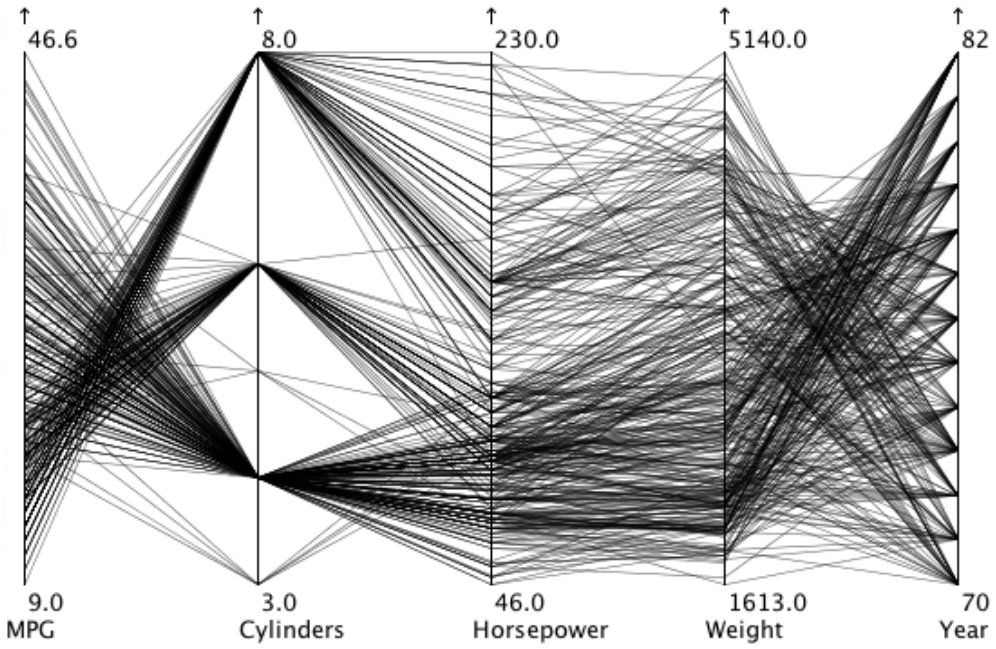
when you have multiple variables:

Represent each possible pair of variables in their own 2-D scatterplot

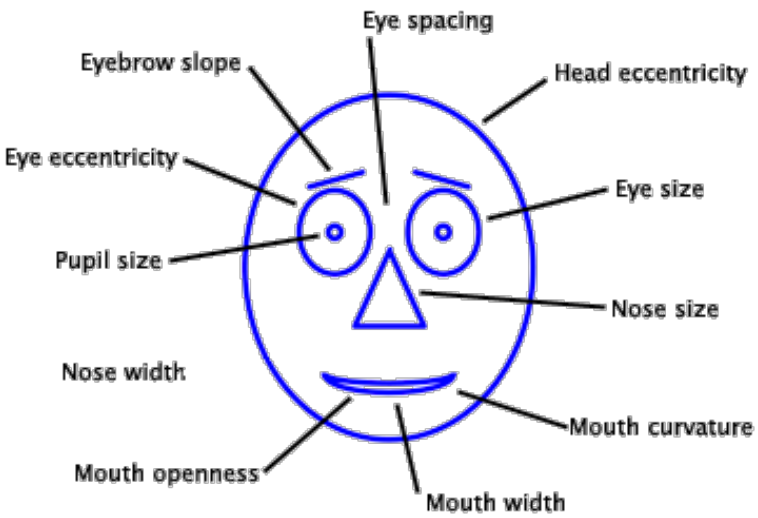
anyone remember the name of this vis technique?



or, Parallel Coordinates



a fun one - Chernov faces



or, use **Dimension Reduction!**

- computational techniques to solve for similar attributes so that the final number or attributes used to describe the dataset are less than the original
  - this is lossy
- for VA, we often reduce from n, to 2
  - we can show 2 dimension in a scatterplot
  - which attributes do we discard? we can't show them all!
  - need to find most salient data attributes and show those

There are loss associated with compressing the dimensionality.

Usually reduce the dimension to 2, because that opens up to a whole bunch to useful visualization tools

Principal Component Analysis (PCA)

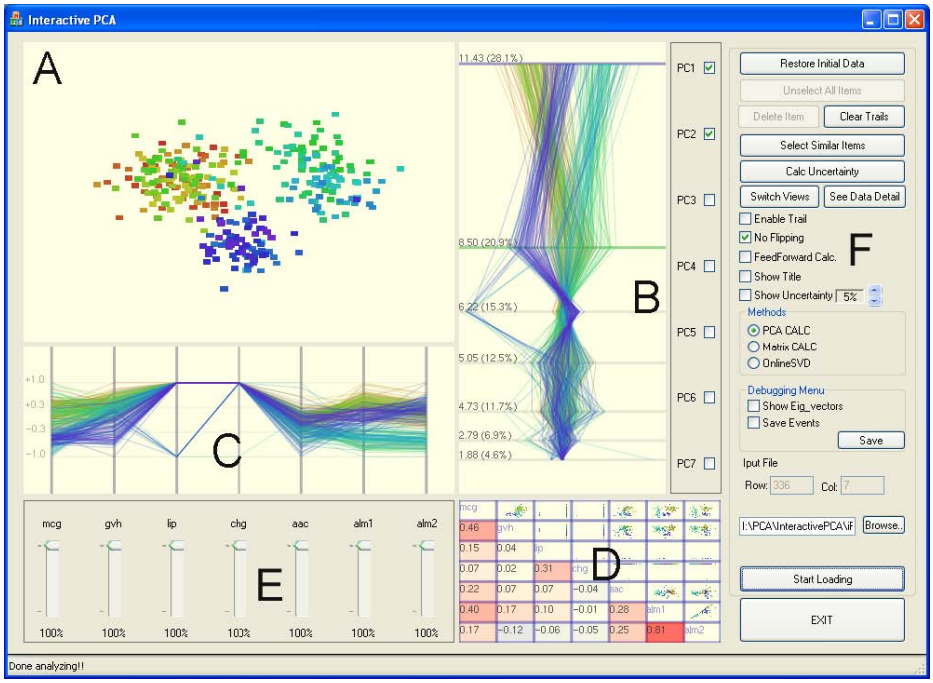
- Frequently used deterministic technique
- transform data table with lots of attributes to principal components, which can be mapped in pairs to x and y axes of scatterplots
  - computed PCs will be <= original attributes of data
- solve for maximum variance
- *i.e., find similar attributes and combine them*

potential PCA limitations

- no guarantee that resulting PCs are “meaningful”, or most important
  - importance is a human, subjective metric
  - max variance may not be most meaningful
- sensitive to outliers
- linear function, what if data/phenomena is non-linear?

iPCA

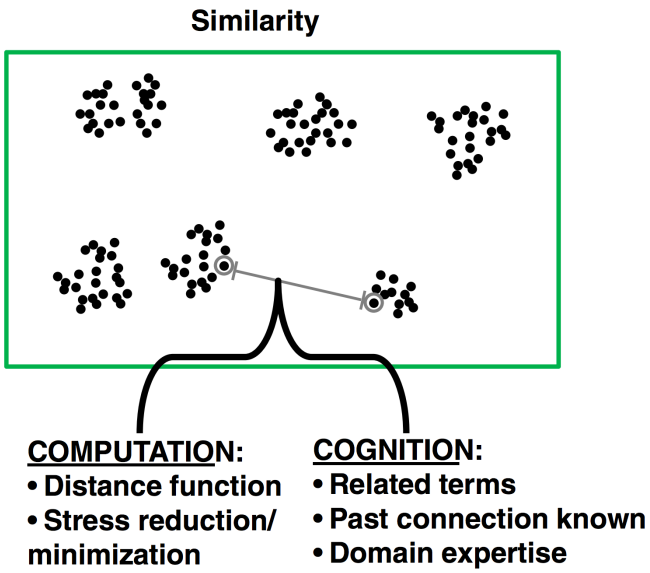
https://youtu.be/oUx-7Hca_i4?t=1m30s



Jeong et al.
CG&A 2009

coupling mental models & data models

- **distance** has meaning to the system and the user
- to the **system**:
  - uses distance functions and similarity metrics
  - preserves distance in high-dimensional feature space in the low-dimensional scatterplot
- to the **user:**
  - proximity approximates similarity
  - we put things nearby if they have some conceptual similarity

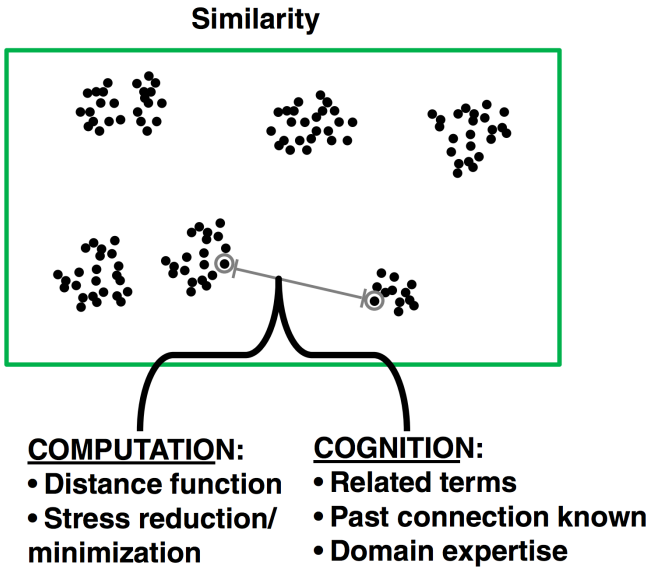


let's try it; InterAxis

- Can control dimension reduction technique using 2 interaction methods
  - directly control the weight of attributes to change the layout of the scatterplot
  - provide example data points on each side of the axes
- http://va.gatech.edu/endert/projects/

coupling mental models & data models

- opens an interesting avenue of research
  - how do we take advantage of this connection?
- we'll get back to this opportunity later



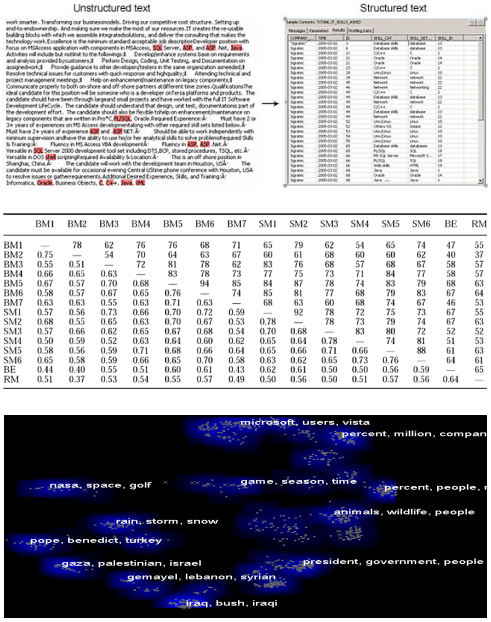
Unstructured Text

Extract Entities

High Dimensional Data

Dimension Reduction Model

Choice of Visual Encodings/ Metaphor



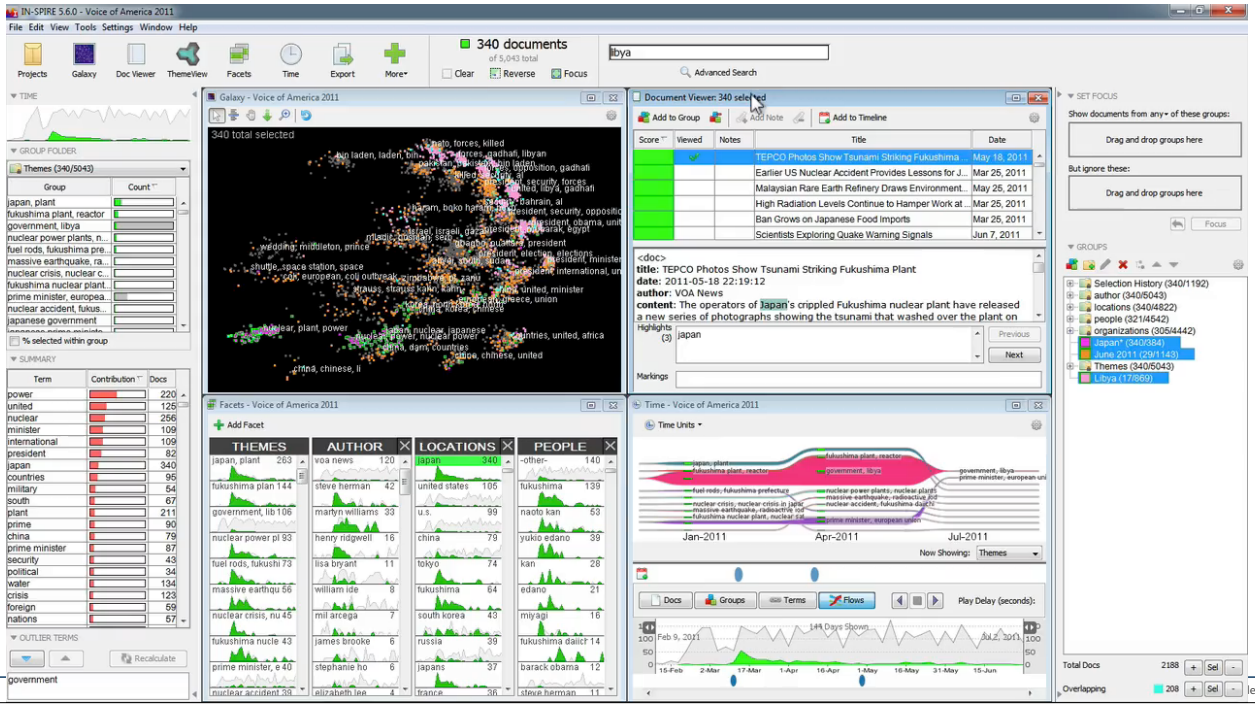
	BM1	BM2	BM3	BM4	BM5	BM6	BM7	SM1	SM2	SM3	SM4	SM5	SM6	BE	RM
BM1	—	78	62	76	76	68	71	65	79	62	54	65	74	47	55
BM2	0.75	—	54	70	64	63	67	60	61	68	60	60	62	40	57
BM3	0.55	0.51	—	72	81	78	62	83	76	68	57	68	67	58	57
BM4	0.66	0.65	0.63	—	83	78	73	77	75	73	71	84	77	58	57
BM5	0.67	0.57	0.70	0.68	—	94	85	84	87	78	74	83	79	68	63
BM6	0.58	0.57	0.67	0.65	0.76	—	74	85	81	77	68	79	83	67	64
BM7	0.63	0.63	0.55	0.63	0.71	0.63	—	68	63	60	68	74	67	46	53
SM1	0.57	0.56	0.73	0.66	0.70	0.72	0.59	—	92	78	72	75	73	67	55
SM2	0.68	0.55	0.65	0.63	0.70	0.67	0.53	0.78	—	78	73	79	74	67	63
SM3	0.57	0.66	0.62	0.65	0.67	0.68	0.54	0.70	0.68	—	83	80	72	52	52
SM4	0.50	0.59	0.52	0.63	0.64	0.60	0.62	0.65	0.64	0.78	—	74	81	51	53
SM5	0.58	0.56	0.59	0.71	0.68	0.66	0.64	0.65	0.66	0.71	0.66	—	88	61	63
SM6	0.65	0.58	0.59	0.66	0.65	0.70	0.58	0.63	0.62	0.65	0.73	0.76	—	64	61
BE	0.44	0.40	0.55	0.51	0.60	0.61	0.43	0.62	0.61	0.50	0.50	0.56	0.59	—	65
RM	0.51	0.37	0.53	0.54	0.55	0.57	0.49	0.50	0.56	0.50	0.51	0.57	0.56	0.64	—

tf-idf for entity selection/extraction

- term frequency–inverse document frequency
  - TF(t) = (num times term t appears in a document) / (num of terms in the document)
  - IDF(t) = log(Total num of documents / num of documents with term t in it)
  - Then, TF x IDF
- commonly used way to apply structure and weighting to unstructured text
  - with structure, we can compute and visualize

IN-SPIRE

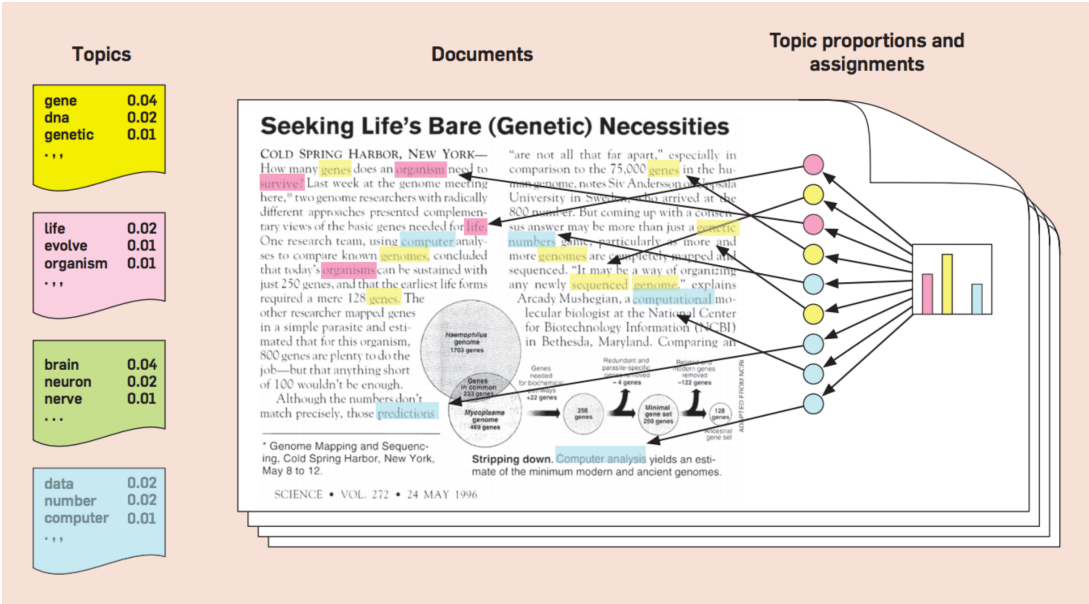
Wise, James A., et al. "Visualizing the non-visual: spatial analysis and interaction with information from text documents." Information Visualization, 1995. Proceedings.. IEEE, 1995.




massive earthquake, ra...
nuclear crisis, nuclear c...
fukushima nuclear plant...
prime minister, europea...
nuclear accident, fukus...
japanese government
% selected within group
SUMMARY
Term | Contribution | Docs
power 220
united 125
nuclear 256
minister 109
international 109
president 82
japan 340
countries 95
military 54
south 67
plant 211
prime 90
china 79
prime minister 87
security 43
political 34
water 134
crisis 123
foreign 59
nations 57
OUTLIER TERMS
Recalculate
government
shuttle, space station, space
cdc, european, ecoli outbreak, zimbabwe, pf, zanu
strauss, strauss kahn, kahn
china, united, minister
greece, union
china, korea, chinese
nuclear, plant, power
japan, nuclear, japanese
nuclear power, nuclear power
countries, united, africa
china, dam, countries
chile, chinese, united
china, chinese, li
<doc>
title: TEPCO Photos Show Tsunami Stri
date: 2011-05-18 22:19:12
author: VOA News
content: The operators of Japan's cripp
a new series of photographs showing
Highlights (3)
japan
Markings
Facets - Voice of America 2011
Add Facet
THEMES
japan, plant 263
fukushima plan 144
government, lib 106
nuclear power pl 93
fuel rods, fukushi 73
massive earthqu 56
nuclear crisis, nu 45
fukushima nucle 43
prime minister, e 40
nuclear accident 39
AUTHOR
voa news 120
steve herman 42
martyn williams 33
henry ridgwell 16
lisa bryant 11
william ide 8
mil arcega 7
james brooke 6
stephanie ho 6
elizabeth lee 4
LOCATIONS
japan 340
united states 105
u.s. 99
china 79
tokyo 74
fukushima 64
south korea 43
russia 39
japans 37
france 36
PEOPLE
-other- 140
fukushima 139
naoto kan 53
yukio edano 39
kan 28
edano 21
miyagi 16
fukushima daiichi 14
barack obama 12
steve herman 11
Time - Voice of America 2011
Time Units
japan, plant
fukushima plant, reactor
fuel rods, fukushima prefecture
nuclear crisis, nuclear crisis in japar
massive earthquake, radioactive iod
fukushima nuclear plant, nuclear cat
Jan-2011
Docs
Groups
Terms
Feb 9, 2011
15-Feb
2-Mar
17-Mar
1-Apr


340 total selected
nato, forces, killed
bin laden, laden, bin
forces, gadhafi, libyan
pakistan
opposition, gadhafi
president, security, forces
united, libya, gadhafi
bahrain, al
haram, boko haram
president, security, oppositio
president, obama, unit
israel, israeli, gaza
barak, egypt
gbagbo, ouattara, president
wedding, middleton, prince
president, election, elections
president, minister
shuttle, space station, space
president, international, ur
cdc, european, ecoli outbreak, zimbabwe, pf, zanu
strauss, strauss kahn, kahn
china, united, minister
greece, union
china, korea, chinese
nuclear, plant, power
japan, nuclear, japanese
nuclear power, nuclear power
countries, united, africa
china, dam, countries
chile, chinese, united
china, chinese, li
Score
Viewed
<doc>
title: TEPCO P
date: 2011-05
author: VOA N
content: The
a new series
Highlights (3)
japan
Markings
Docs
220
125
256
Facets - Voice of America 2011
Add Facet
Time - Voice o
Time Units

topic modeling

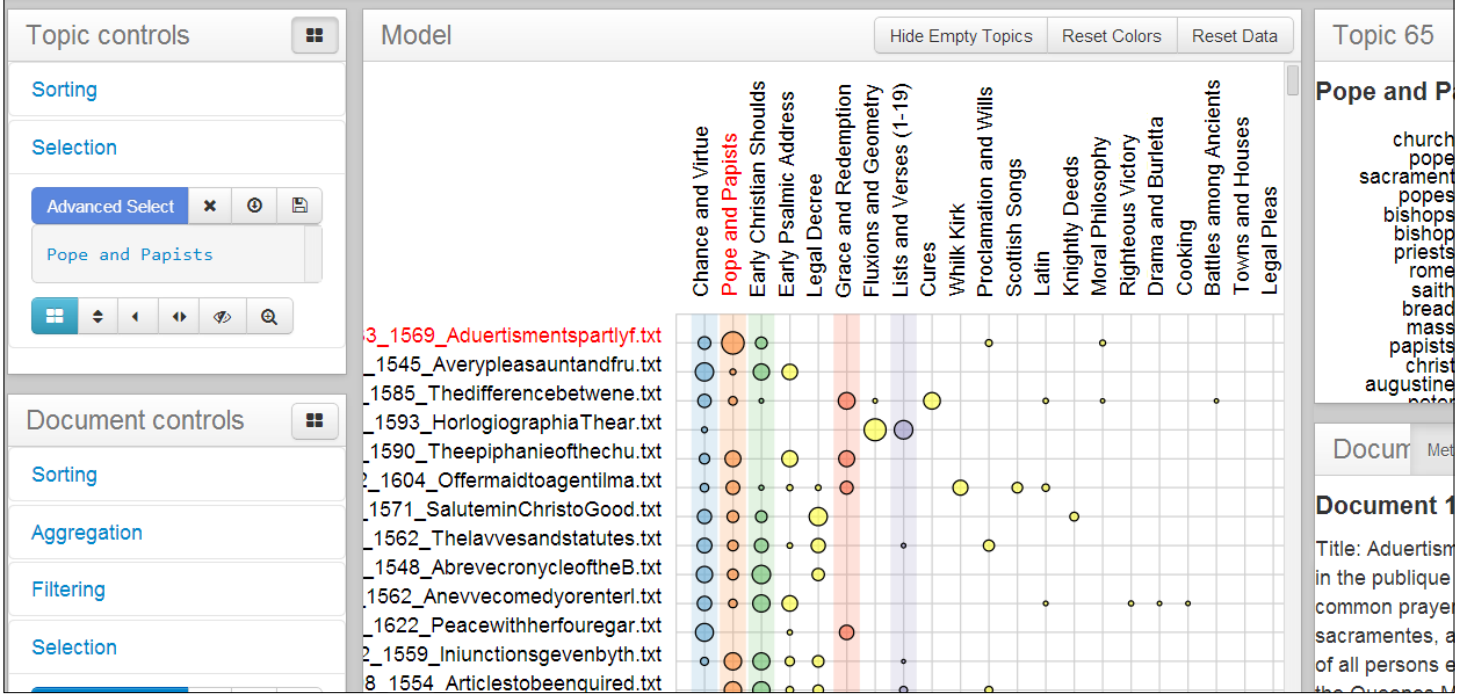


Blei, David M. "Probabilistic topic models." Communications of the ACM 55.4 (2012): 77-84.

Alexander, Eric, et al. "Serendip: Topic model-driven visual exploration of text corpora." Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on. IEEE, 2014.

Serendip

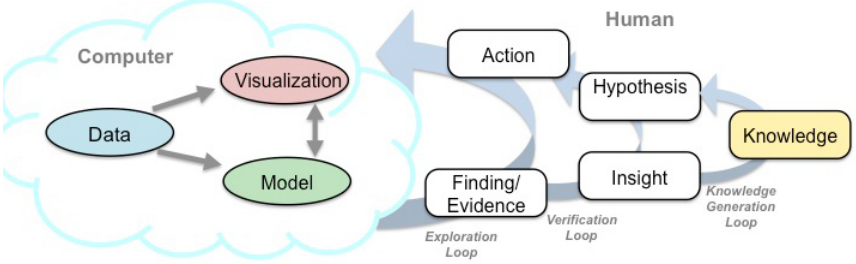
http://vep.cs.wisc.edu/serendip/



Incorporating Human Feedback is
One Central Component

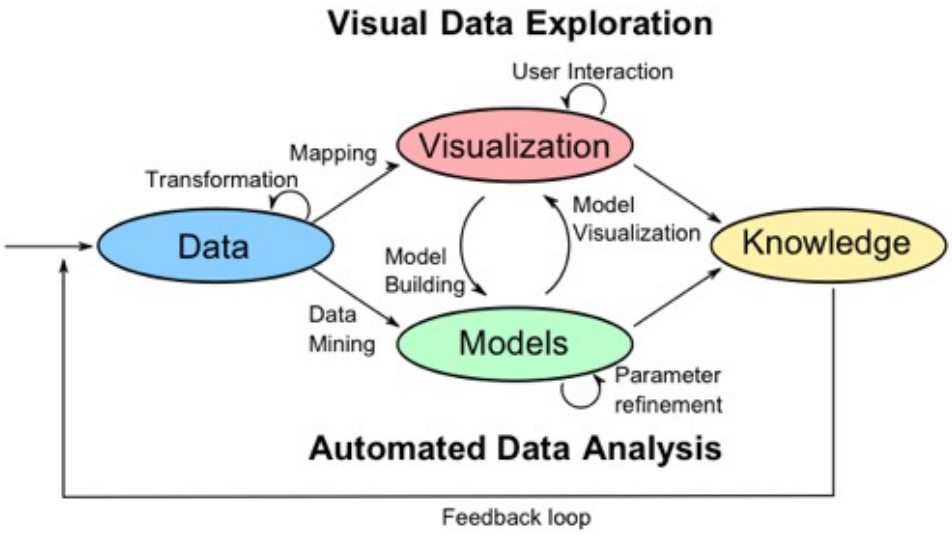
incorporating user feedback

- unsupervised computation is rarely the “final” solution
- “human-in-the-loop” design advocates for including user feedback in the holistic system
- people have domain expertise, can balance subjective tradeoffs, etc.
- let’s look at a few examples of user interaction in VA



Sacha, Dominik, et al. "Knowledge generation model for visual analytics." Visualization and Computer Graphics, IEEE Transactions on 20.12 (2014): 1604-1613.

incorporating user feedback



"Mastering the Information Age Solving Problems with Visual Analytics",edited by Daniel Keim, Jörn Kohlhammer,Geo rey Ellis and Florian Mansmann

Distance = Similarity makes sense to people

Andrews, Christopher, Alex Endert, and Chris North. "Space to think: large high-resolution displays for sensemaking." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2010.

including ML seems to work well

- helps offload computationally-attractive tasks to machines - nice!
- frees up user to focus cognitive energy on other things - also good

- but we can do better!

warning! \begin{alex_rant}

- do not treat user interaction as an easy way out
  - "oh, we'll just give the user a menu for that!"

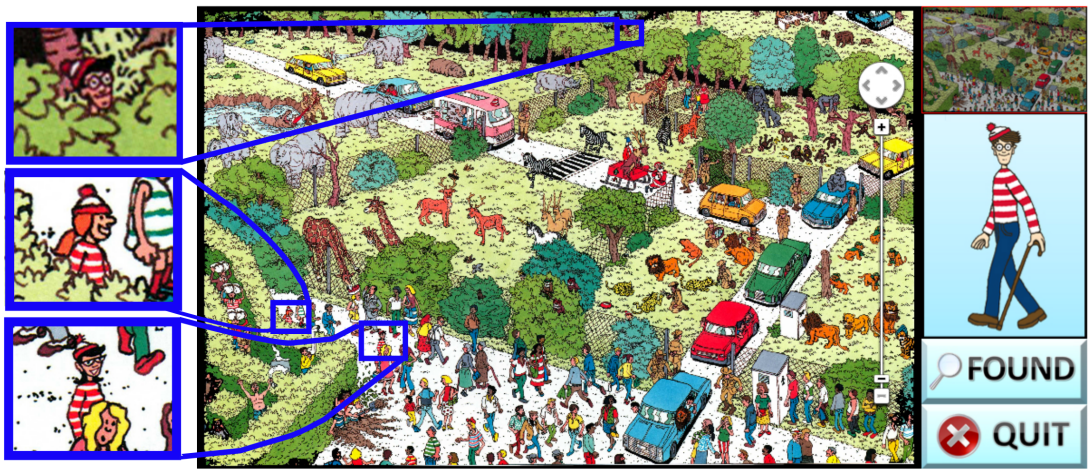
- do not make your problems (as developers and designers) the user's problems

- interaction should be meant to explore data, ask questions, gain insight

user interaction is valuable data

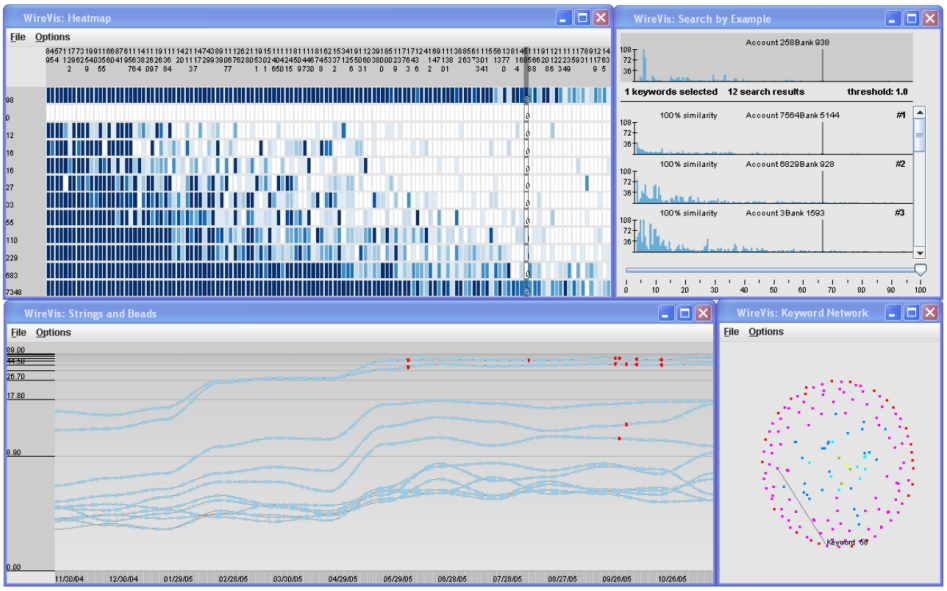
- quantitative record of analysis and exploration process
- what can we learn from this data?
  - data models? sensemaking stages?
- “interaction is the insight”
- let’s look at a few examples

are you good at Where’s Waldo?



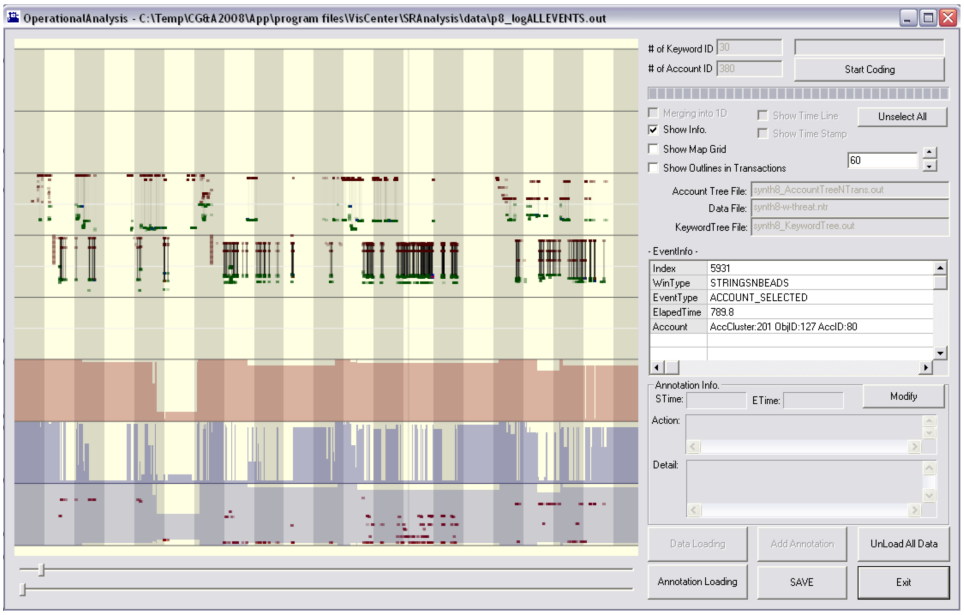
Brown, Eli T., et al. "Finding waldo: Learning about users from their interactions." Visualization and Computer Graphics, IEEE Transactions on 20.12 (2014): 1663-1672.

recover reasoning process: capture



Dou, Wenwen, et al. "Recovering reasoning processes from user interactions." IEEE Computer Graphics and Applications 3 (2009): 52-61.

recover reasoning process: analyze



recover reasoning process: results

- analyzing user interactions after the analysis, researchers can recover:
  - 60% of strategies
  - 60% of methods
  - 79% of findings

- Impressive!
- Can we systematically analyze the user interaction in the system at runtime to steer the models?

another way to look at this

- instead of asking users to directly tune and parameterize models, how can systems learn from their exploratory user interactions?

- How can we use the interactions that people have done to steer the models and the visualizations directly?

let's try it

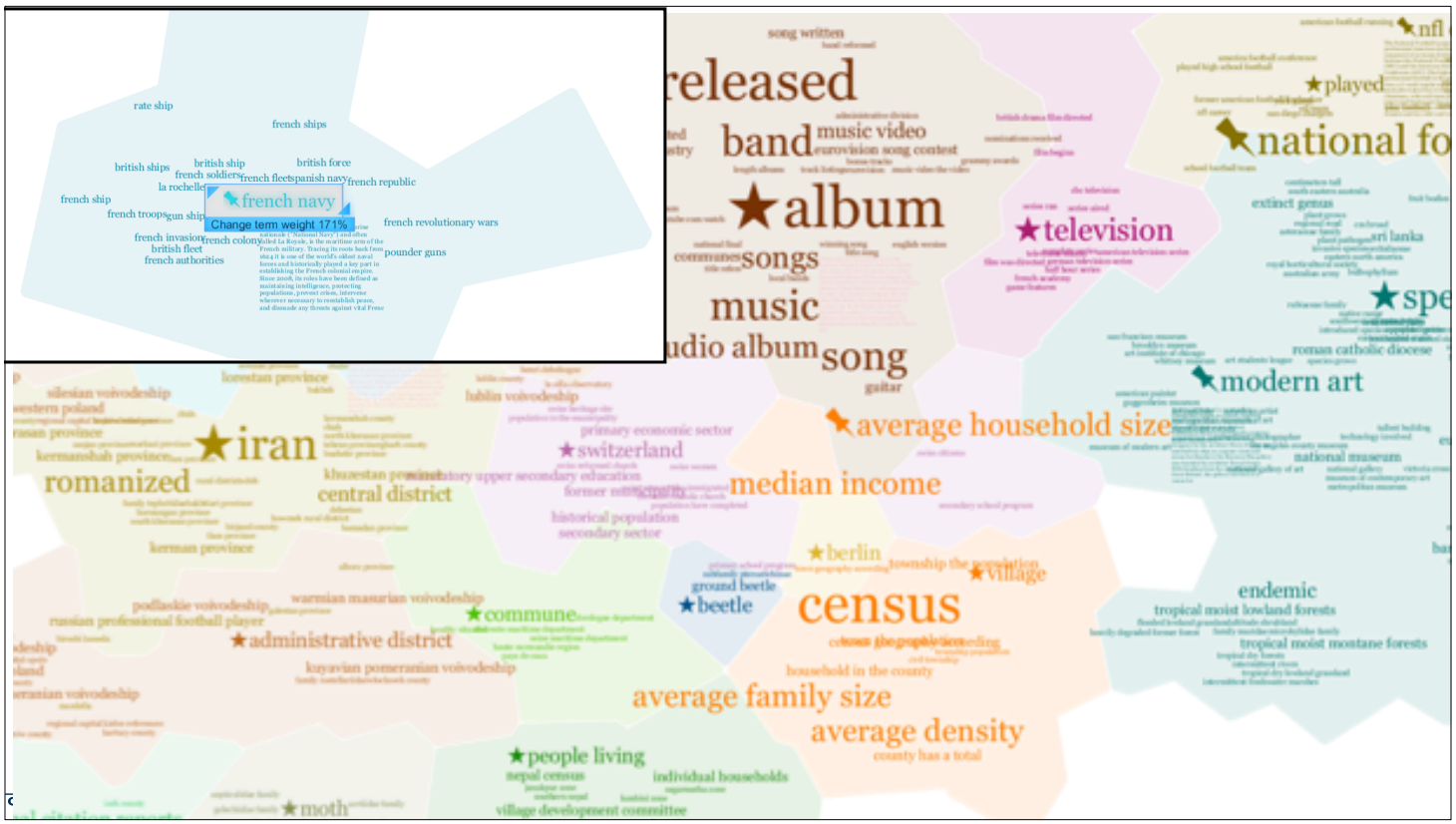
Another example

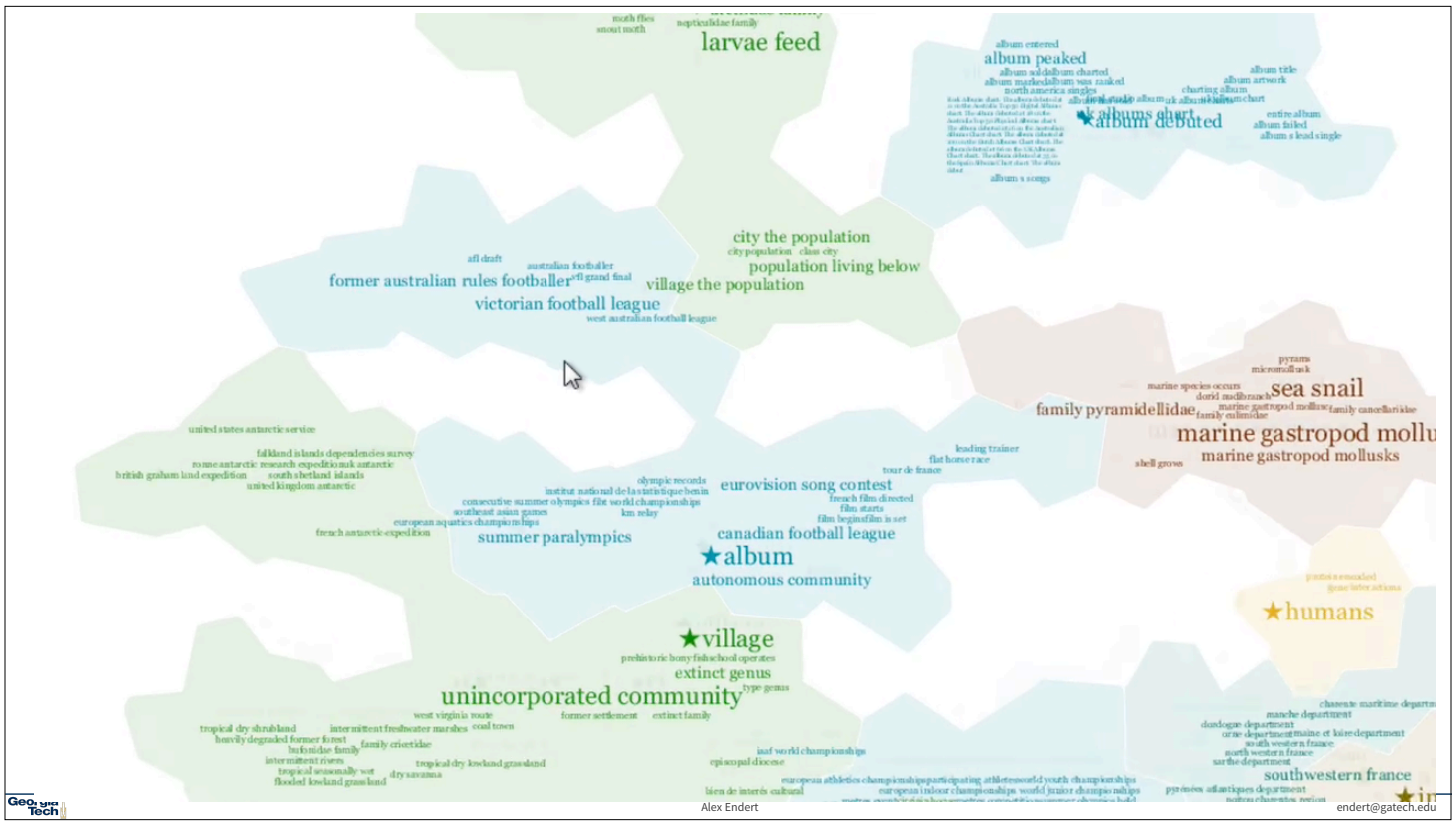
- Wrangler
  - http://vis.stanford.edu/wrangler/
- Visual Analytics for data cleaning

Ranking Data Items

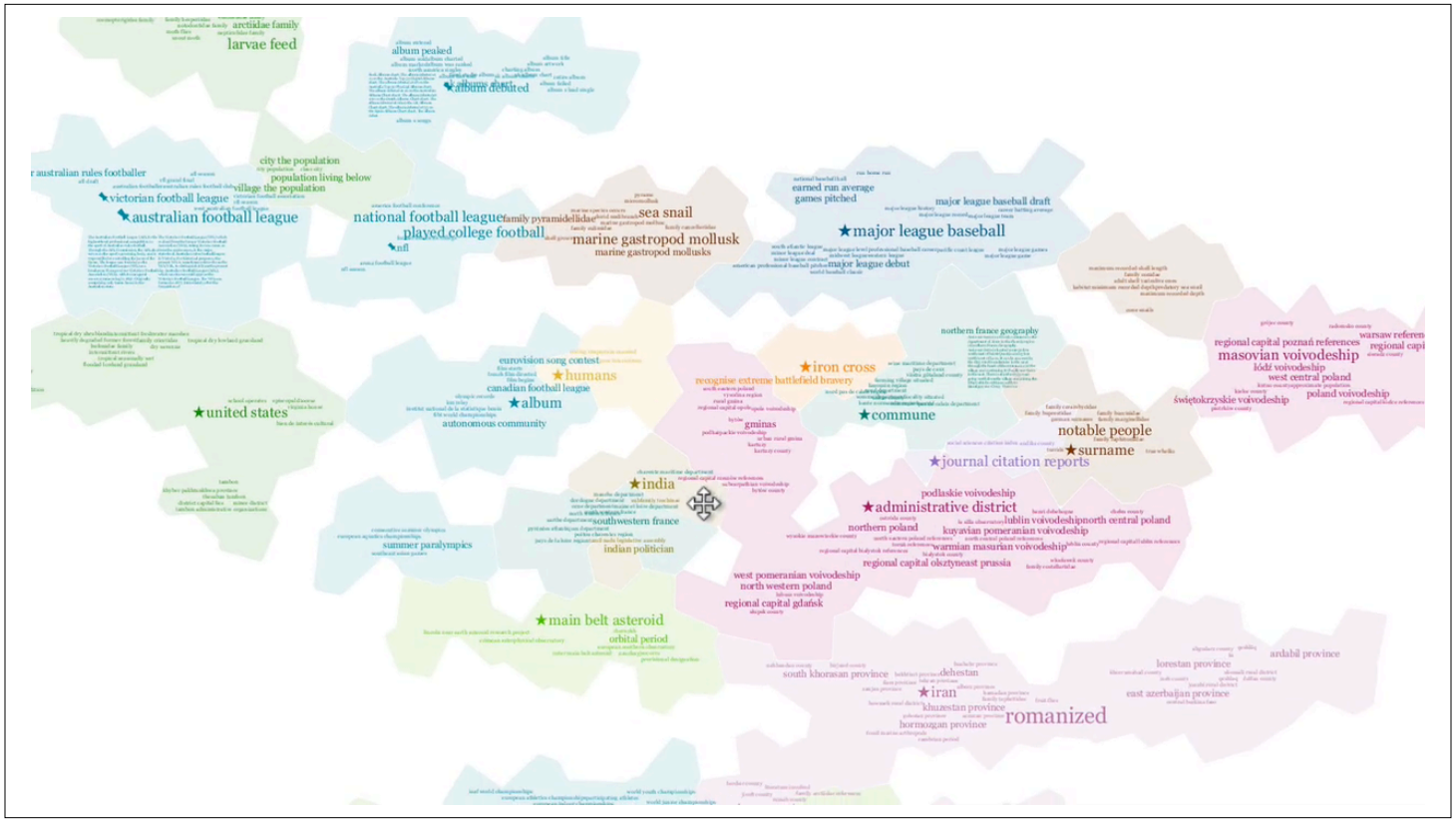
- LineUp

https://lineup.js.org/app/



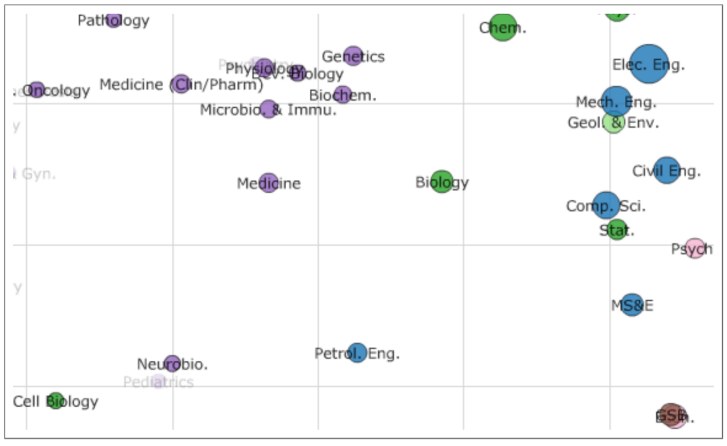


Alex Endert
endert@gatech.edu



potential challenges

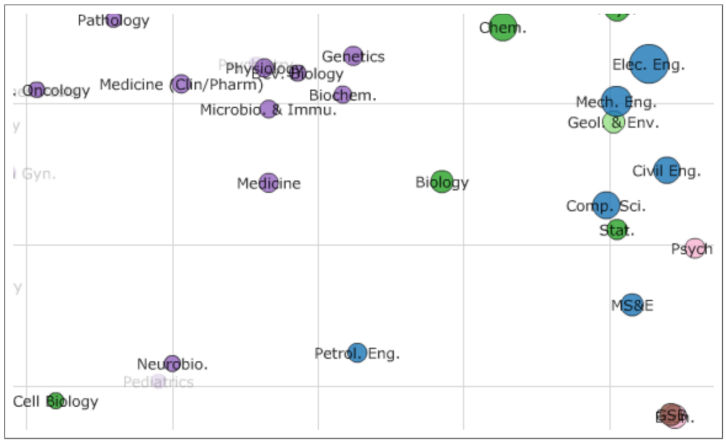
- **interpretability** and **trust**
- computation **approximates** data characteristics and phenomena, but sometimes get is wrong
- good case study by Chuang et al.

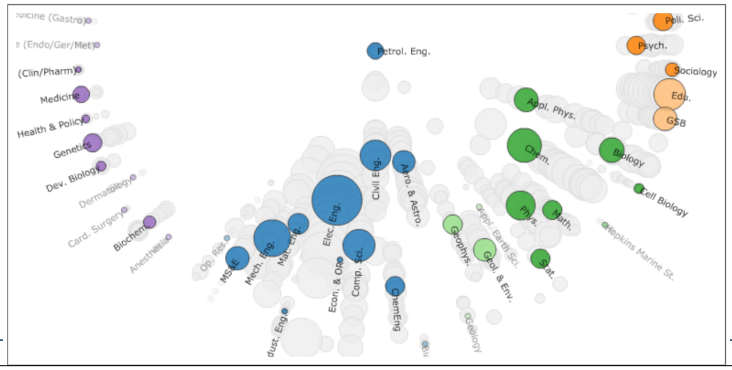


DR-produced view has distortions that can create visual artifacts

Petrol Eng. is similar to Neurobiology?

transformation of view based on user selection can remove these





DR-produced view has distortions that can create visual artifacts

Petrol Eng. is similar to Neurobiology?

transformation of view based on user selection can remove these

endert@gatech.edu