INTERACTIVE
Data Analysis

Jeffrey Heer @jeffrey_heer
Univ. of Washington + Trifacta

Data Analysis & Statistics, Tukey & Wilk 1966

Four major influences act on data analysis today:

1. The formal theories of statistics.
2. Accelerating developments in computers and display devices.
3. The challenge, in many fields, of more and larger bodies of data.
4. The emphasis on quantification in a wider variety of disciplines.

Data Analysis & Statistics, Tukey & Wilk 1966

While some of the influences of statistical theory on data analysis have been helpful, others have not.

Data Analysis & Statistics, Tukey & Wilk 1966

Exposure, the effective laying open of the data to **display the unanticipated**, is to us a major portion of data analysis…

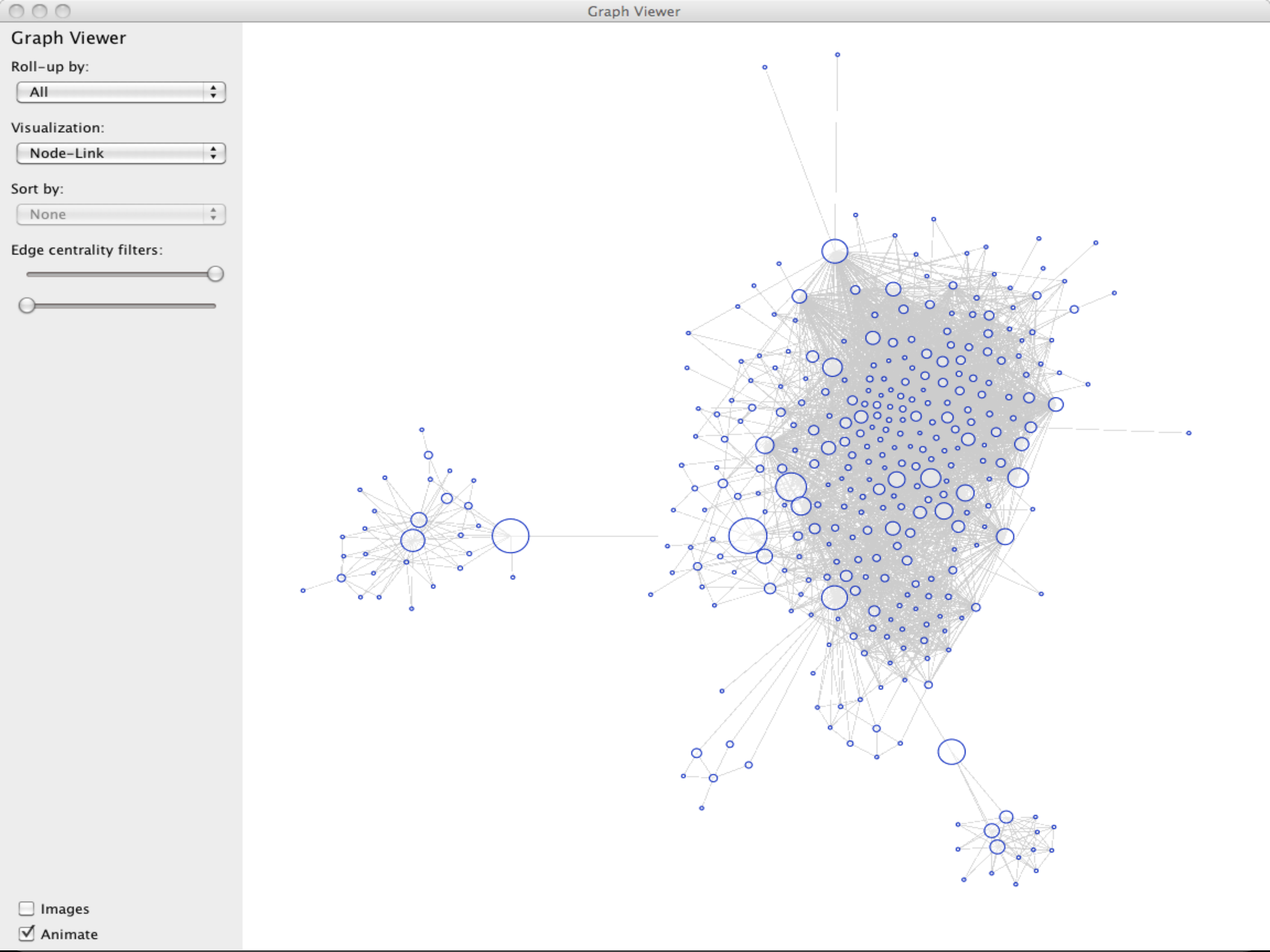
It is not clear how the **informality** and **flexibility** appropriate to the **exploratory character** of exposure can be fitted into any of the structures of formal statistics so far proposed.

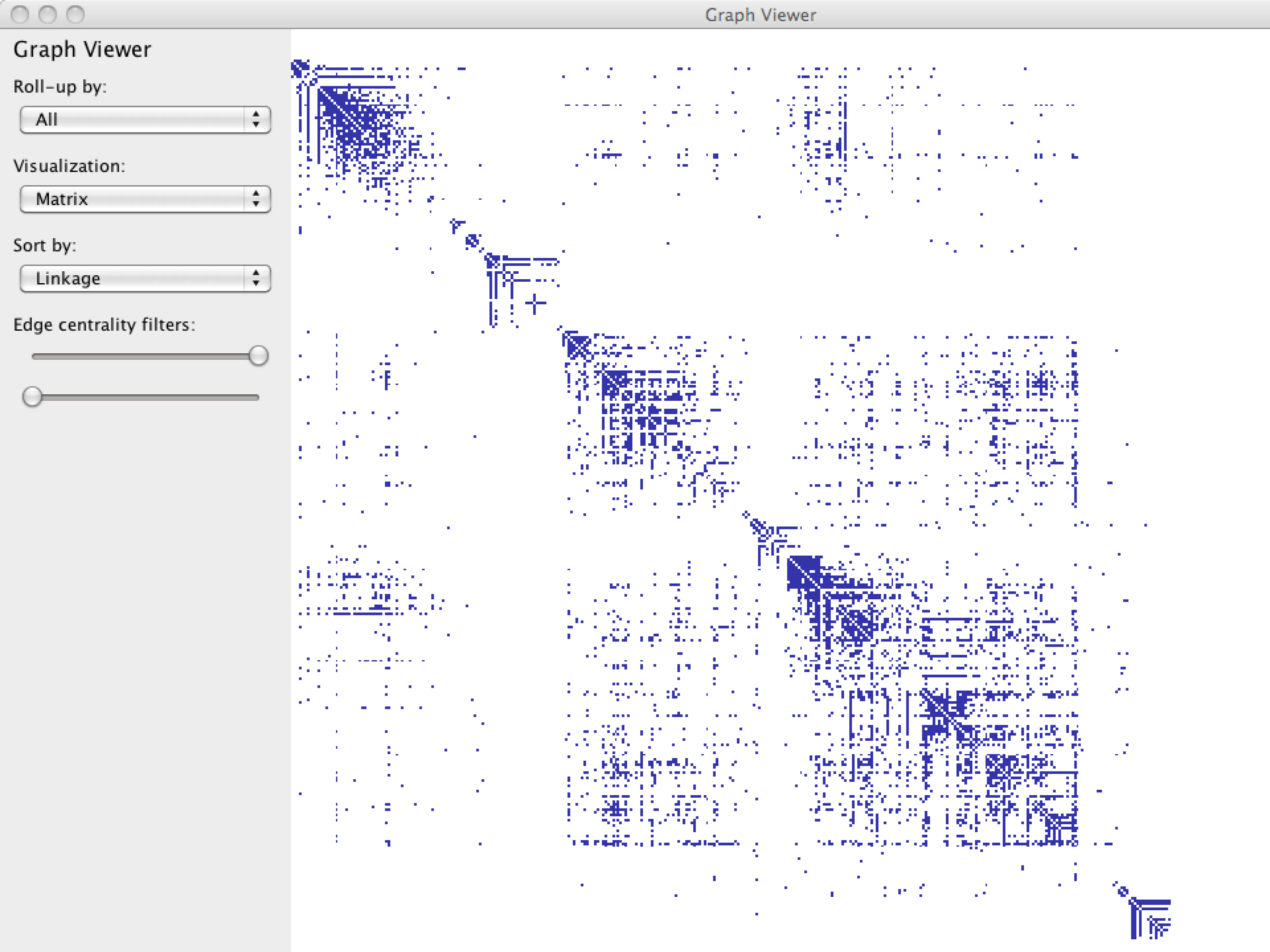
Data Analysis & Statistics, Tukey & Wilk 1966

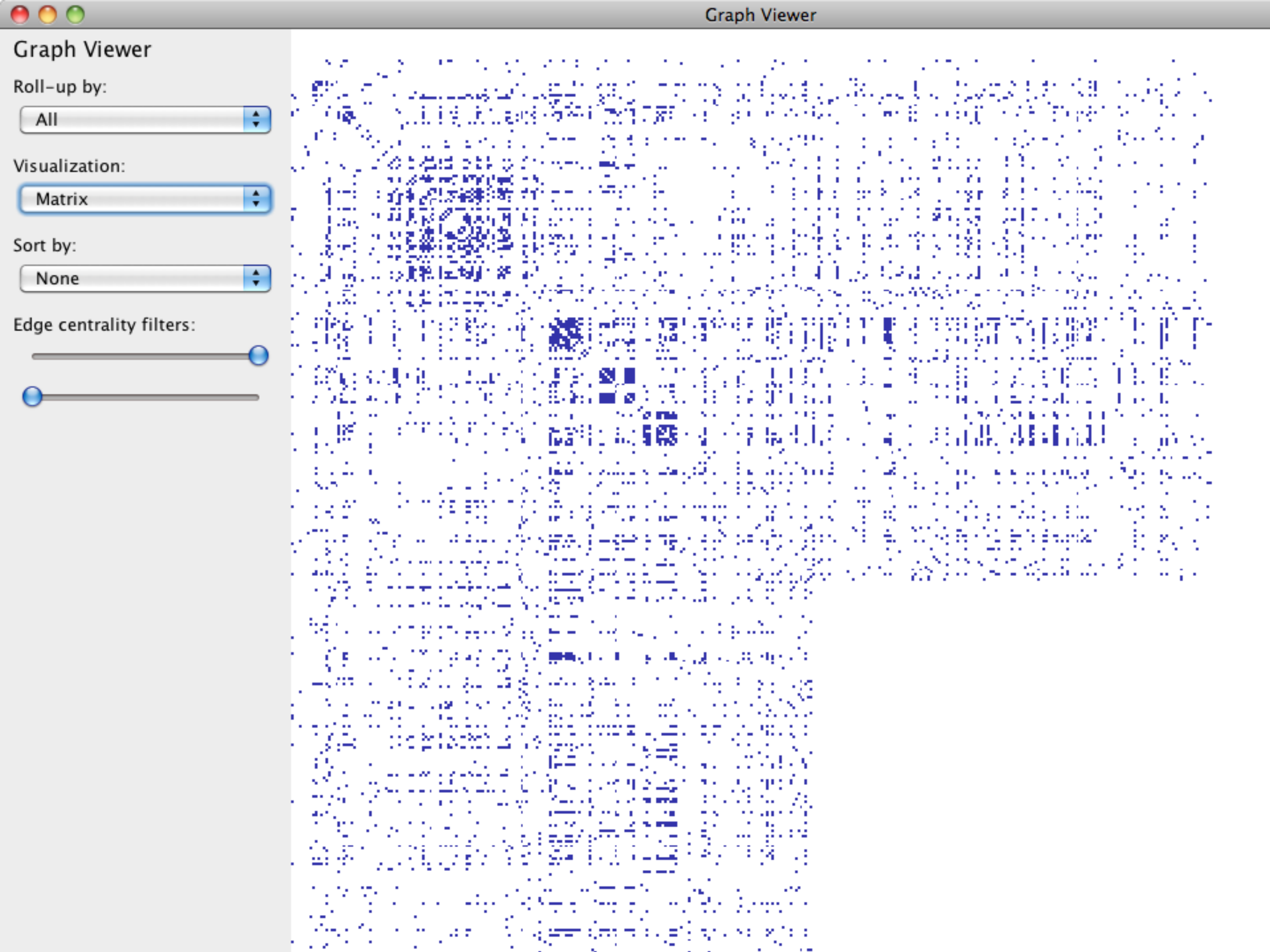
Accordingly, both approaches and techniques need to be structured so as to **facilitate human involvement and intervention**.

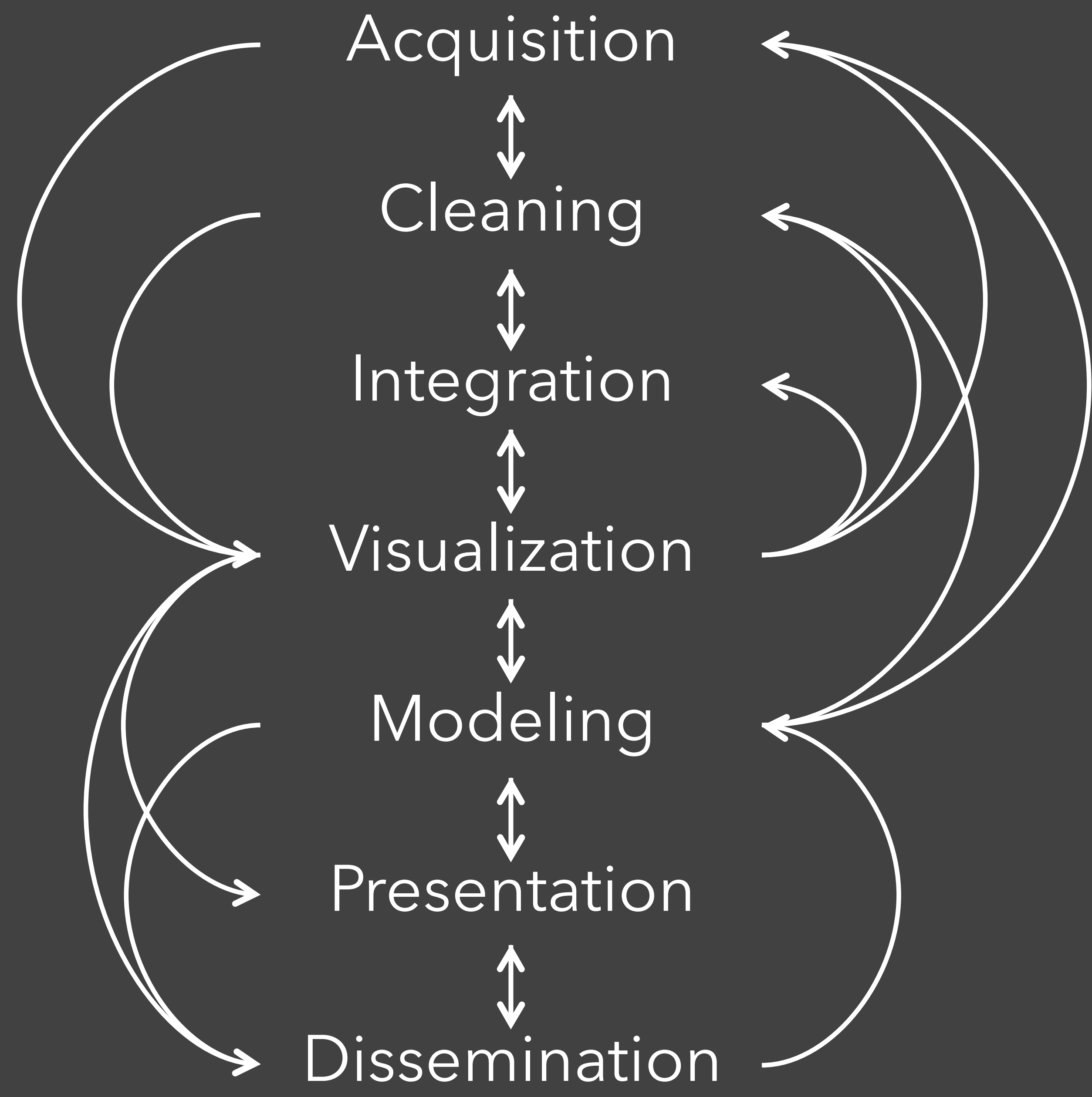
Some implications for effective analysis are: (1) it is essential to have convenience of **interaction of people and intermediate results** and (2) at all stages of data analysis, the outputs need to be **matched to the capabilities of the people who use it and want it.**

Data Analysis & Statistics, Tukey & Wilk 1966


Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Node-Link
Sort by:
None
Edge centrality filters:
Images
Animate


Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Matrix
Sort by:
Linkage
Edge centrality filters:


Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Matrix
Sort by:
None
Edge centrality filters:


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination

How can we **transform data** without programming?

I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I’m lucky if I get to do any “analysis” at all.

Anonymous Data Scientist

from our interview study [VAST’12]

Bureau of Justice Statistics - Data Online
http://bjs.ojp.usdoj.gov/

Reported crime in Alabama

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	4525375	4029.3	987	2732.4	309.9
2005	4548327	3900	955.8	2656	289
2006	4599030	3937	968.9	2645.1	322.9
2007	4627851	3974.9	980.2	2687	307.7
2008	4661900	4081.9	1080.7	2712.6	288.6

Reported crime in Alaska

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	657755	3370.9	573.6	2456.7	340.6
2005	663253	3615	622.8	2601	391
2006	670053	3582	615.2	2588.5	378.3
2007	683478	3373.9	538.9	2480	355.1
2008	686293	2928.3	470.9	2219.9	237.5

Reported crime in Arizona

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	5739879	5073.3	991	3118.7	963.5
2005	5953007	4827	946.2	2958	922
2006	6166318	4741.6	953	2874.1	914.4
2007	6338755	4502.6	935.4	2780.5	786.7
2008	6500180	4087.3	894.2	2605.3	587.8

Reported crime in Arkansas

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	2750000	4033.1	1096.4	2699.7	237
2005	2775708	4068	1085.1	2720	262
2006	2810872	4021.6	1154.4	2596.7	270.4
2007	2834797	3945.5	1124.4	2574.6	246.5
2008	2855390	3843.7	1182.7	2433.4	227.6

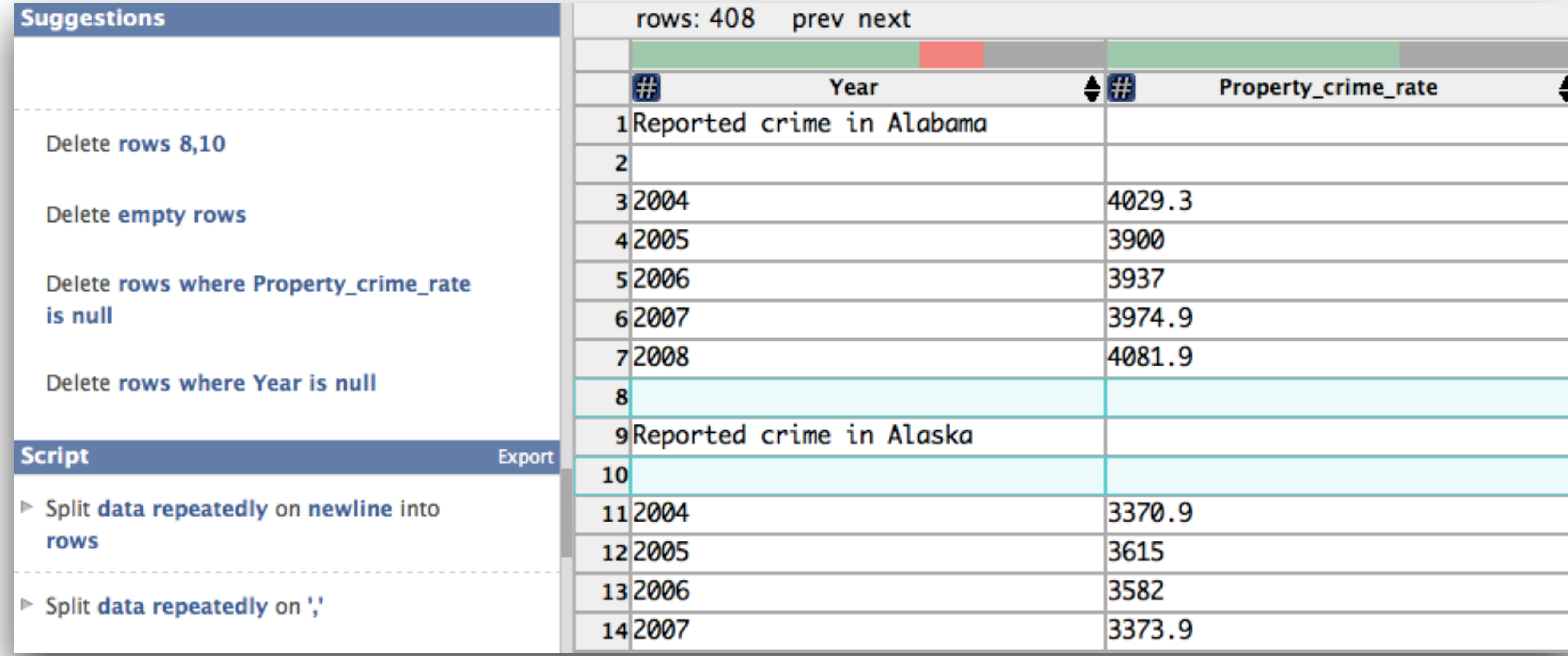
Reported crime in California

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	35842038	3423.9	686.1	2033.1	704.8
2005	36154147	3321	692.9	1915	712
2006	36457549	3175.2	676.9	1831.5	666.8
2007	36553215	3032.6	648.4	1784.1	600.2
2008	36756666	2940.3	646.8	1769.8	523.8

Reported crime in Colorado

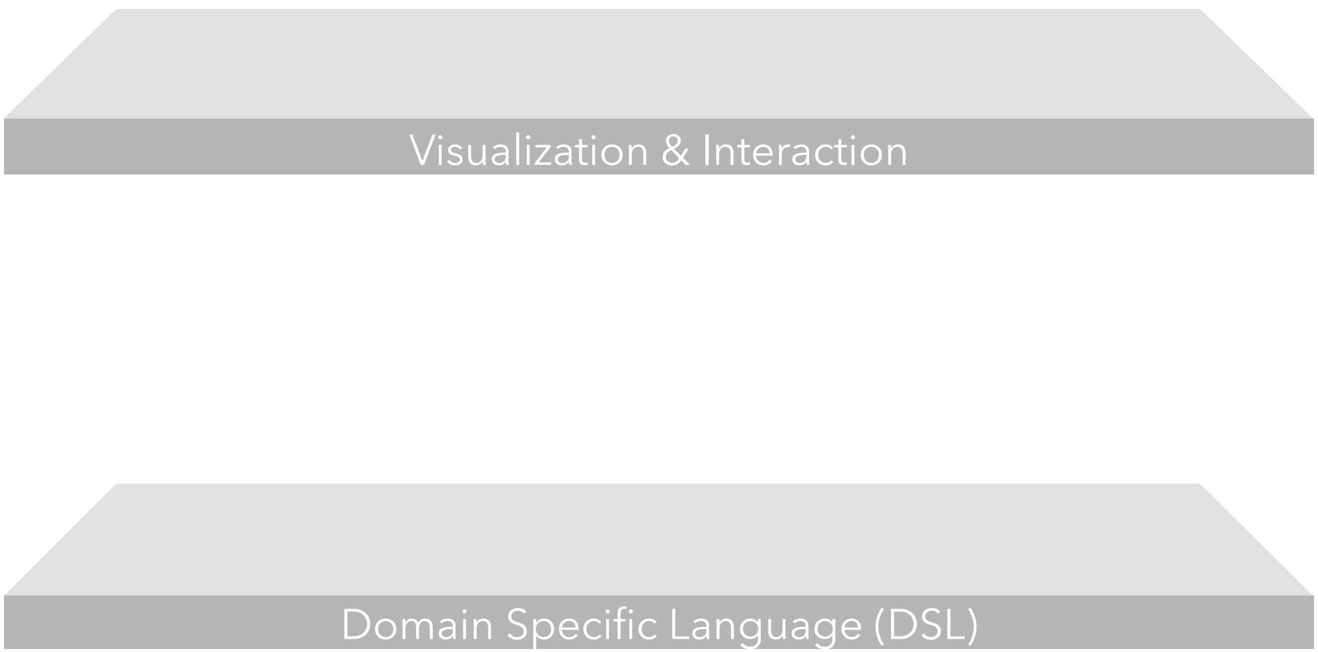
Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	4601821	3918.5	717.3	2679.5	521.6

DataWrangler

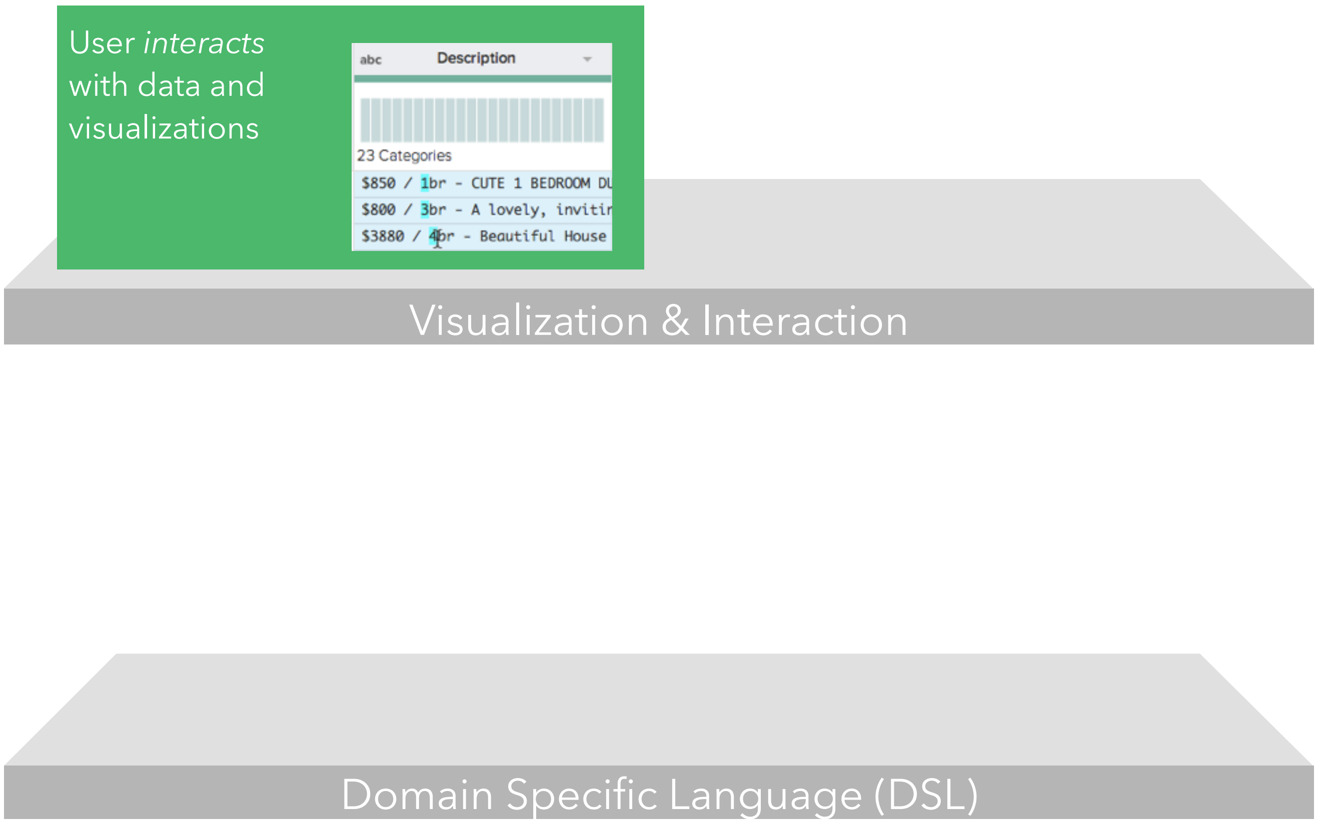


with **S. Kandel**, P. Guo, A. Paepcke & J. Hellerstein [CHI'11]

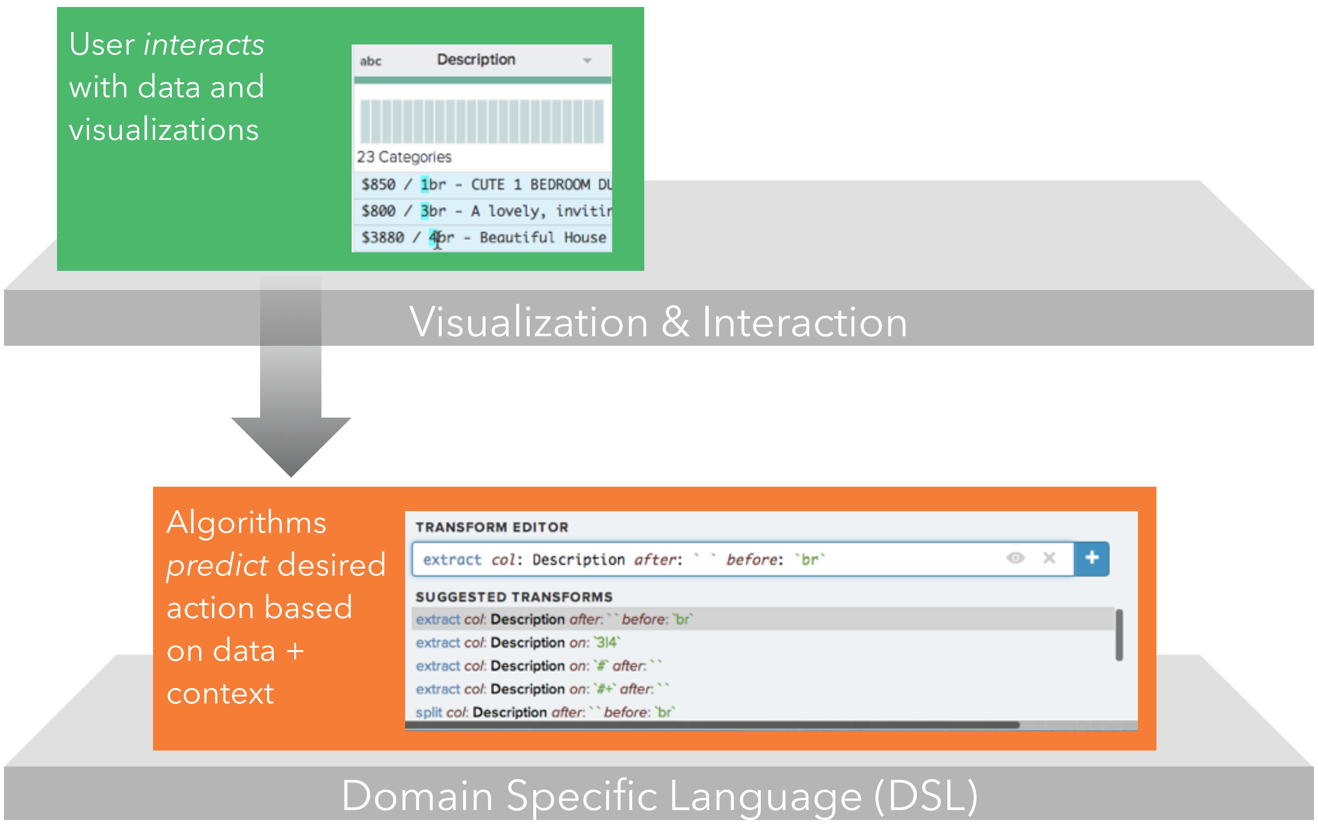
Predictive Interaction



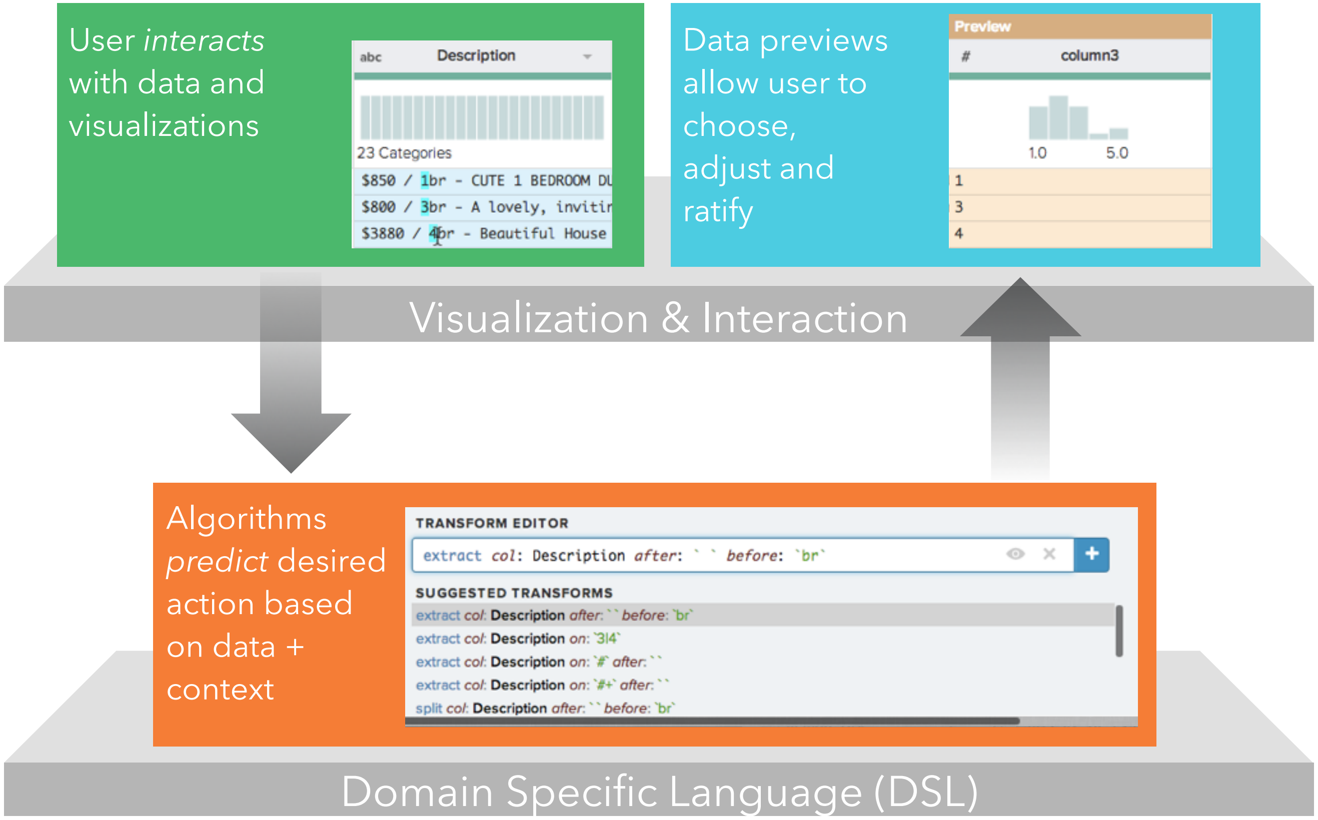
Predictive Interaction



Predictive Interaction



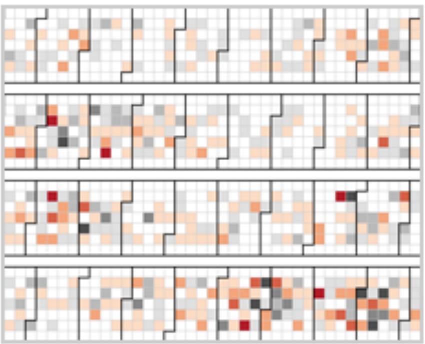
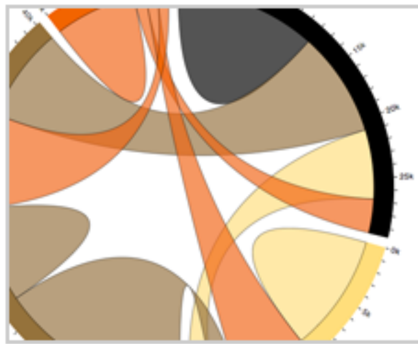
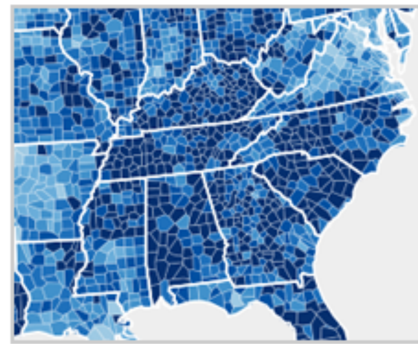
Predictive Interaction

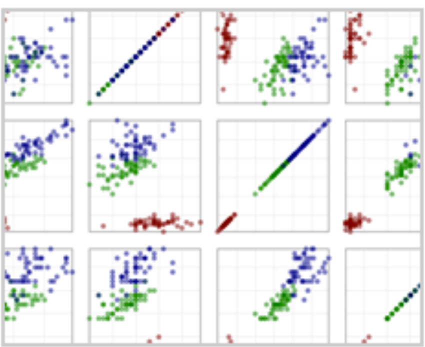
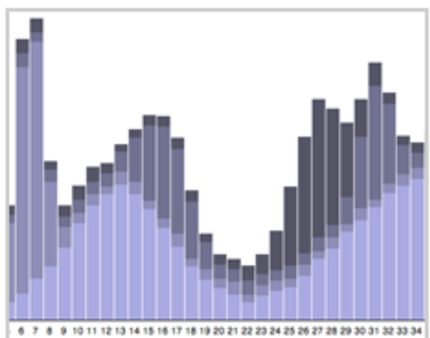


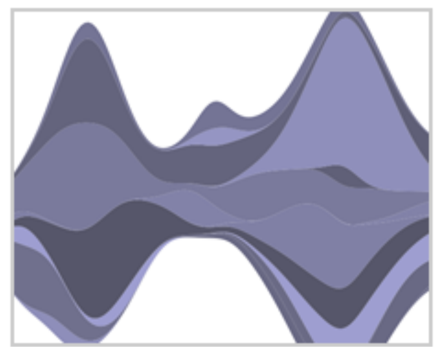
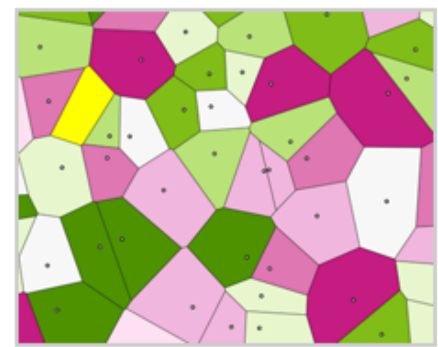
TRIFACTA

How might we support **expressive** and **effective** visualization designs?

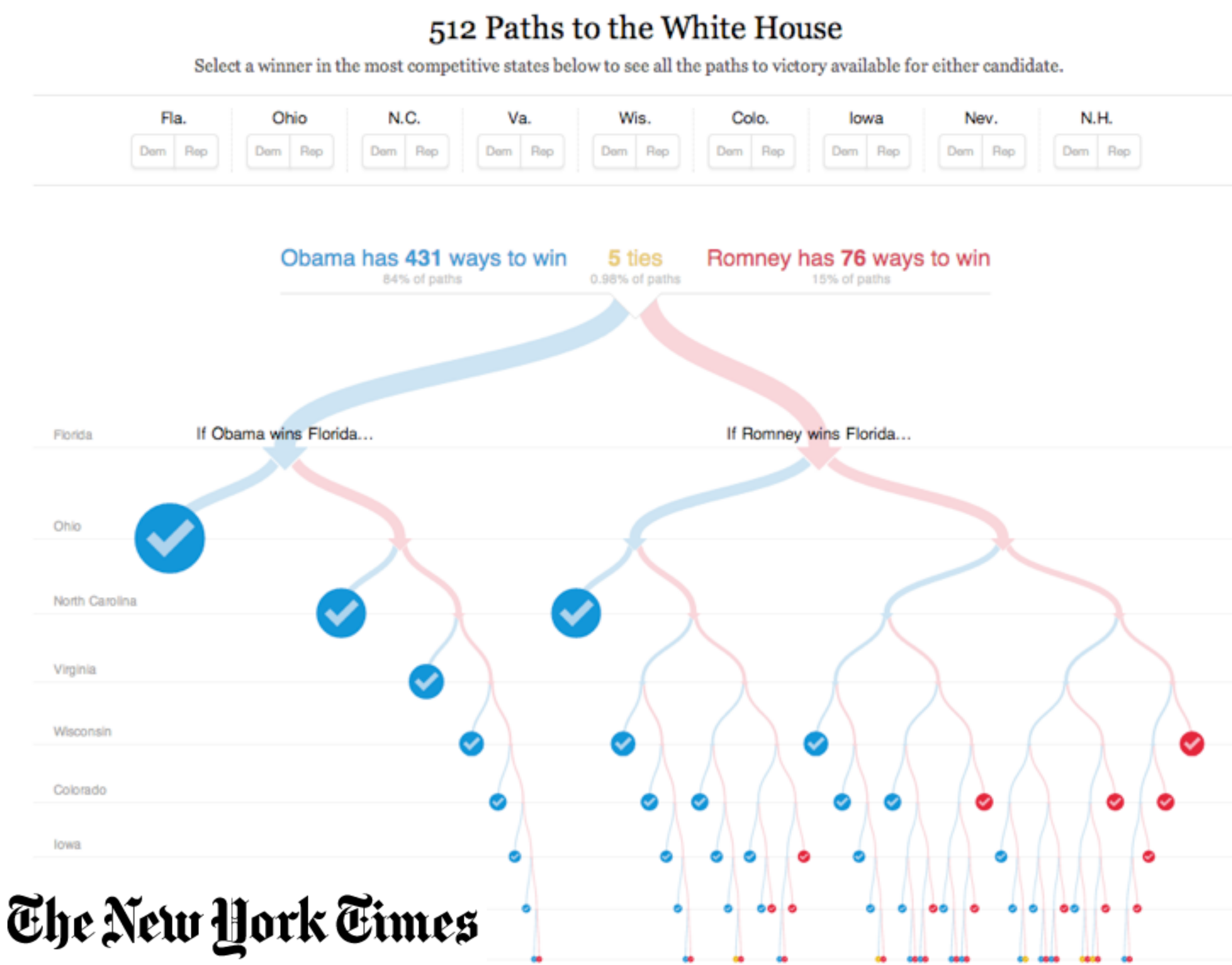
d3.js Data-Driven Documents

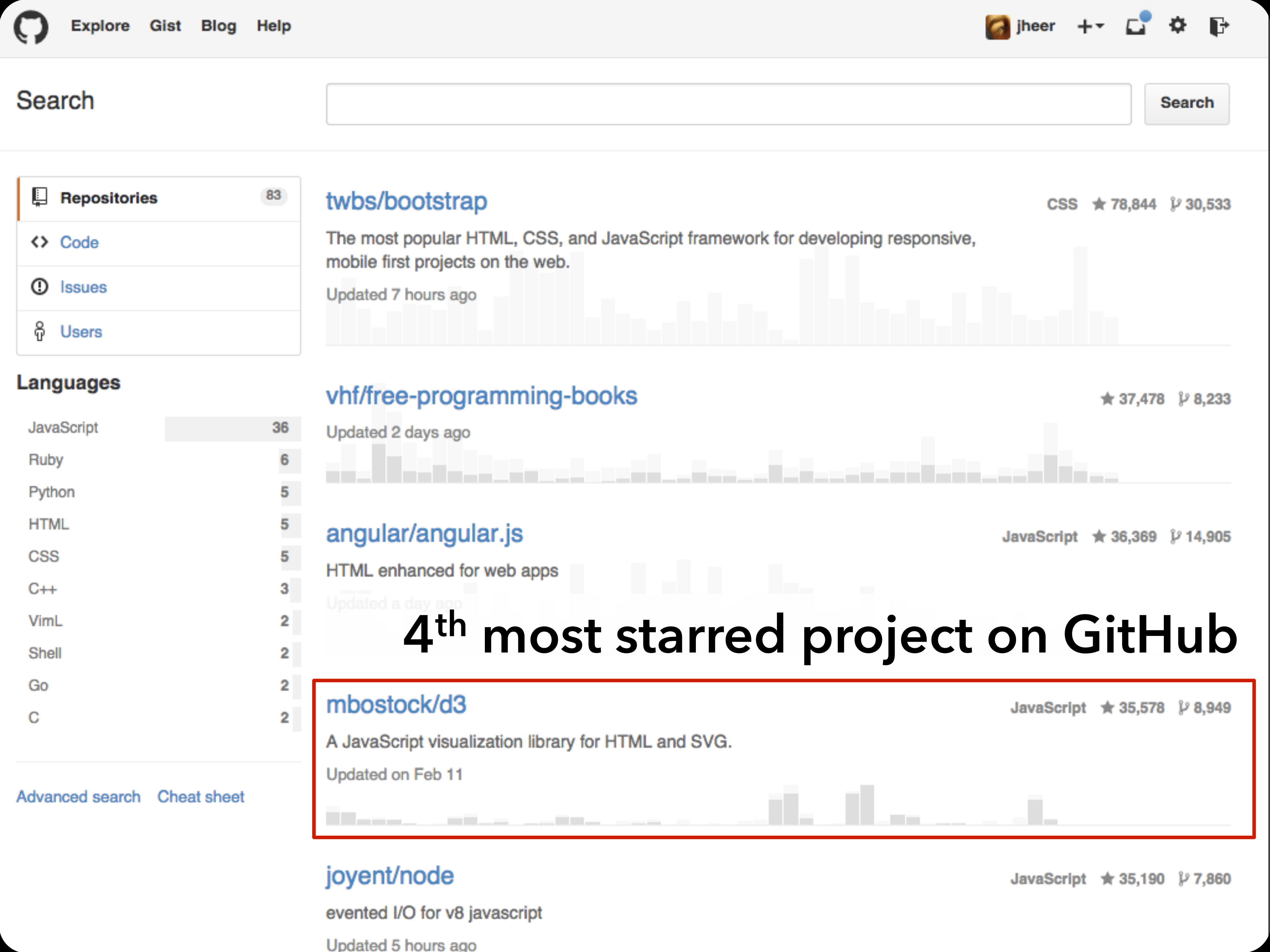
with **M. Bostock**, V. Ogievetsky [InfoVis '11]


512 Paths to the White House
Select a winner in the most competitive states below to see all the paths to victory available for either candidate.
Fla.
Ohio
N.C.
Va.
Wis.
Colo.
Iowa
Nev.
N.H.
Dem
Rep
Obama has 431 ways to win
84% of paths
5 ties
0.98% of paths
Romney has 76 ways to win
15% of paths
Florida
If Obama wins Florida...
If Romney wins Florida...
Ohio
North Carolina
Virginia
Wisconsin
Colorado
Iowa
The New York Times


stamen design

stamen design

4th most starred project on GitHub



Vega https://vega.github.io

Vega A VISUALIZATION GRAMMAR

Vega is a declarative format for creating, saving, and sharing visualization designs. With Vega, visualizations are described in JSON, and generate interactive views using either HTML5 Canvas or SVG.

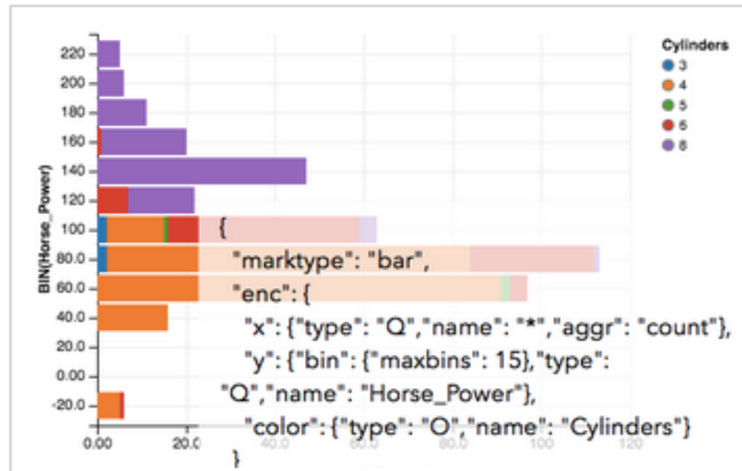
TOOLKITS

VEGA offers a full declarative visualization grammar, suitable for expressive custom interactive visualization design and programmatic generation.

Tutorial | Documentation | Discussion Forum

v1.5 (stable): download, examples, github

NEW v2.0 (dev): download, examples, github

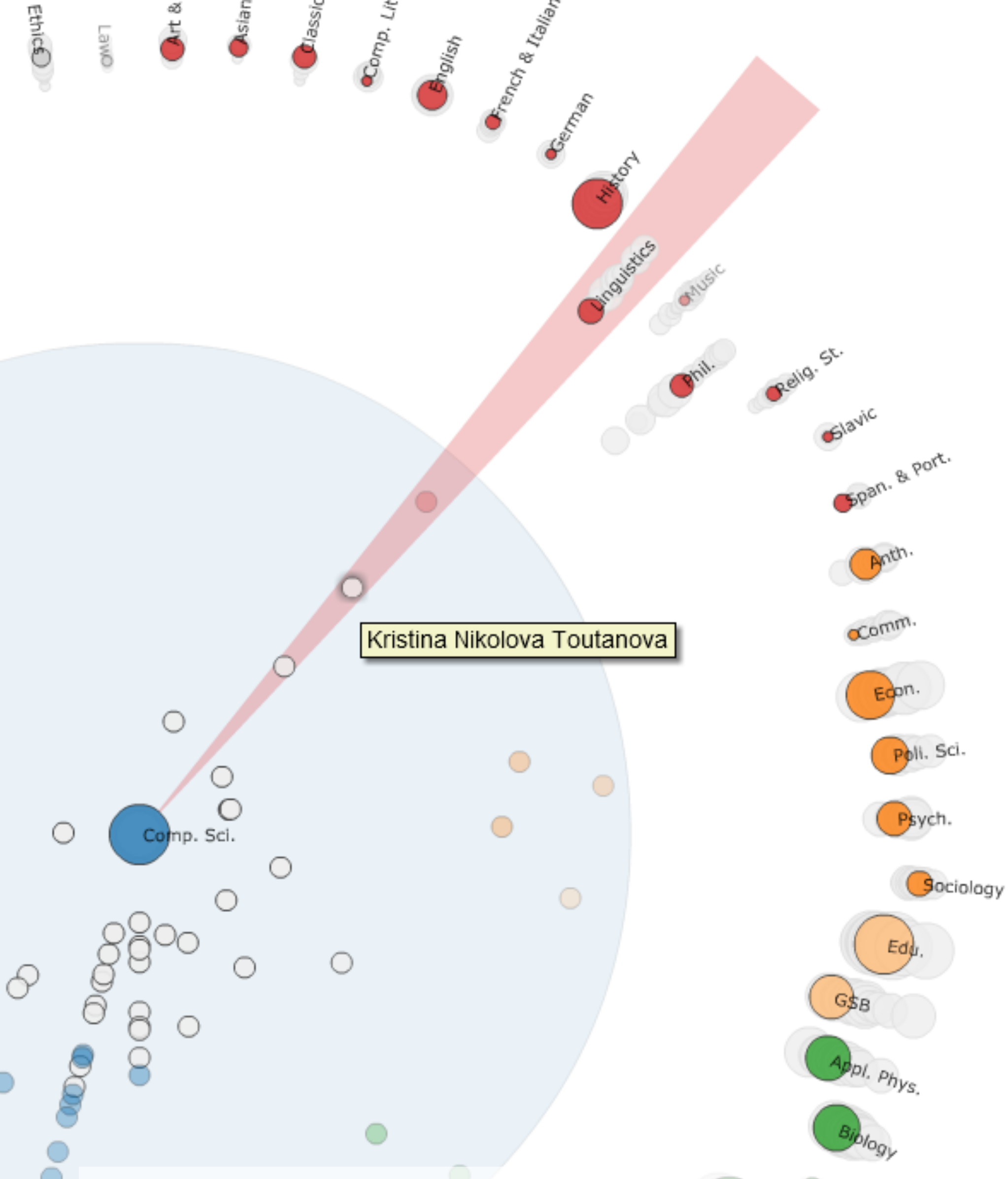


NEW **VEGALITE** provides a higher-level grammar for visual analysis, comparable to ggplot or Tableau, that generates complete Vega specifications.

Online Editor | GitHub

vega.github.io

How might we support model **interpretation** and **refinement**?



Effective statistical models for syntactic and semantic disambiguation

Student: Kristina Nikolova Toutanova
Advisor: Christopher D. Manning

Computer Science (2005)

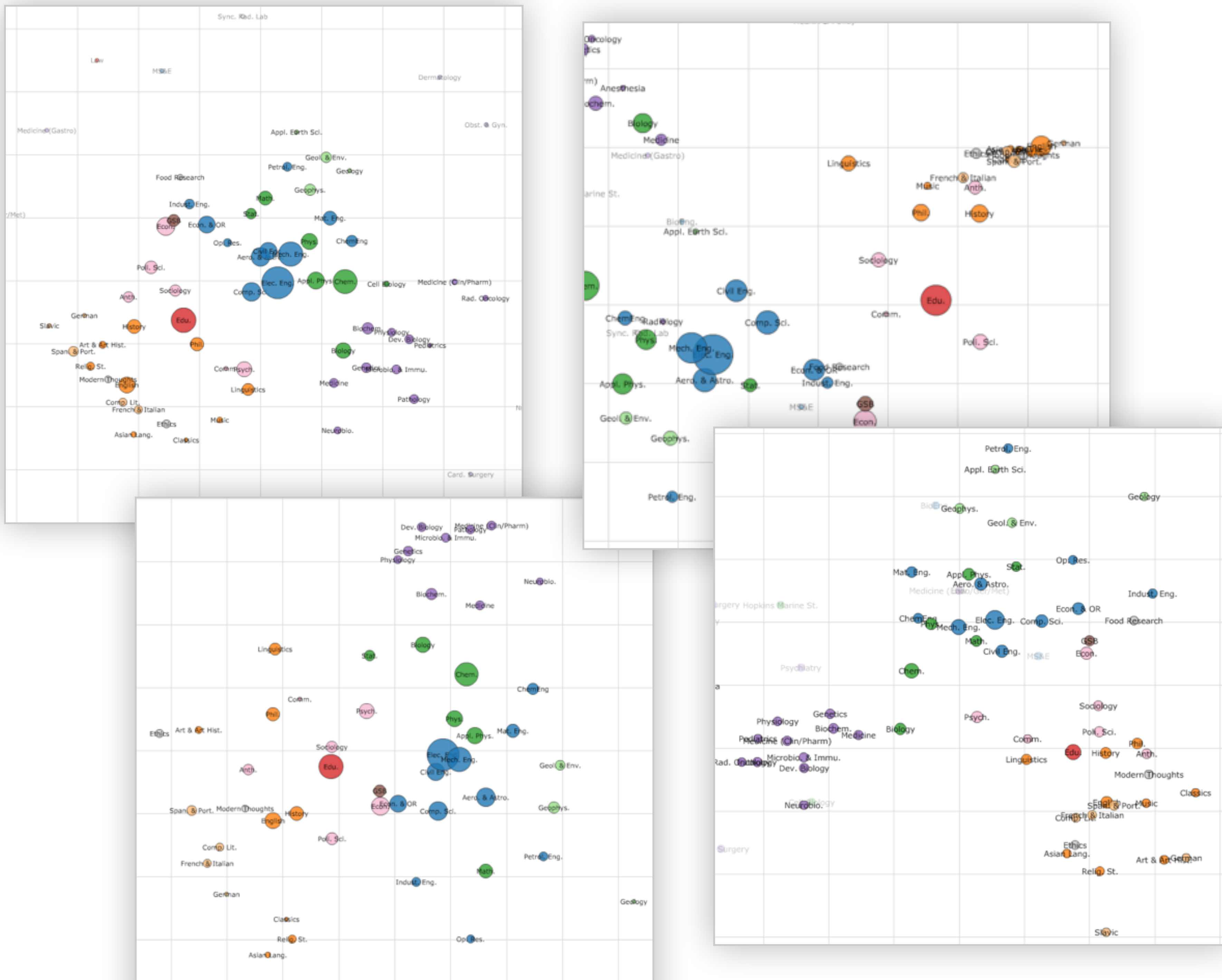
Keywords: Syntactic, Semantic, Tree kernels, Parsing

Abstract:

This thesis focuses on building effective statistical models for disambiguation of sophisticated syntactic and semantic natural language (NL) structures. We advance the state of the art in several domains by (i) choosing representations that encode domain knowledge more effectively and (ii) developing machine learning algorithms that deal with the specific properties of NL disambiguation tasks--sparsity of training data and large, structured spaces of hidden labels. For the task of syntactic disambiguation, we propose a novel representation of parse trees that connects the words of the sentence with the hidden syntactic structure in a direct way. Experimental evaluation on parse selection for a Head Driven Phrase Structure Grammar shows the new representation achieves superior performance compared to previous models. For the task of disambiguating the semantic role structure of verbs, we build a more accurate model, which captures the knowledge that the semantic frame of a verb is a joint structure with strong dependencies between arguments. We achieve this using a Conditional Random Field without Markov independence assumptions on the sequence of semantic role labels. To address the sparsity problem in machine learning for NL, we develop a method for incorporating many additional sources of information, using Markov chains in the space of words. The Markov chain framework makes it possible to combine multiple knowledge sources, to learn how much to trust each of them, and to chain inferences together. It achieves large gains in the task of disambiguating prepositional phrase attachments.

Stanford Dissertation Browser

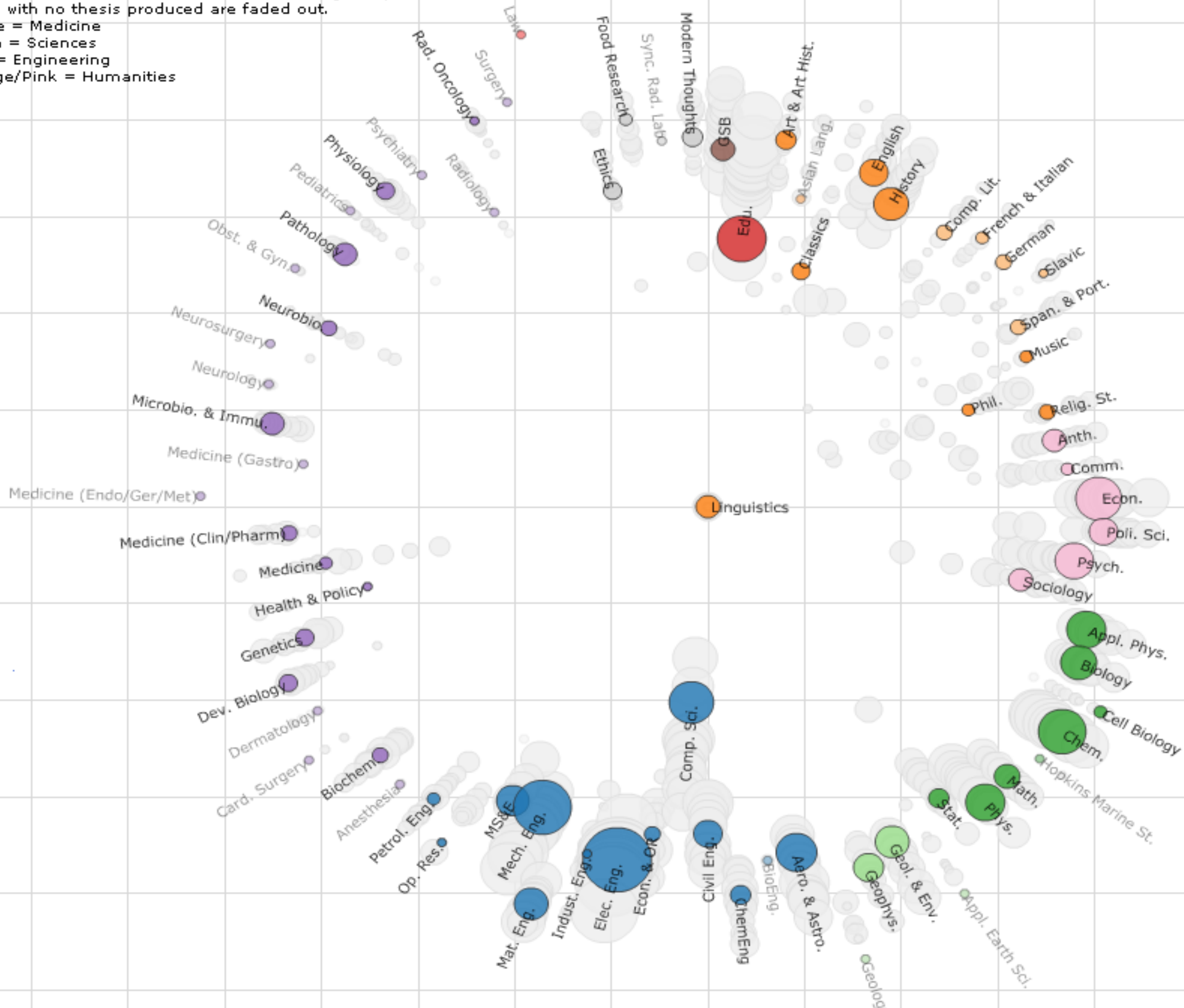
with **Jason Chuang**, Dan Ramage & Christopher Manning



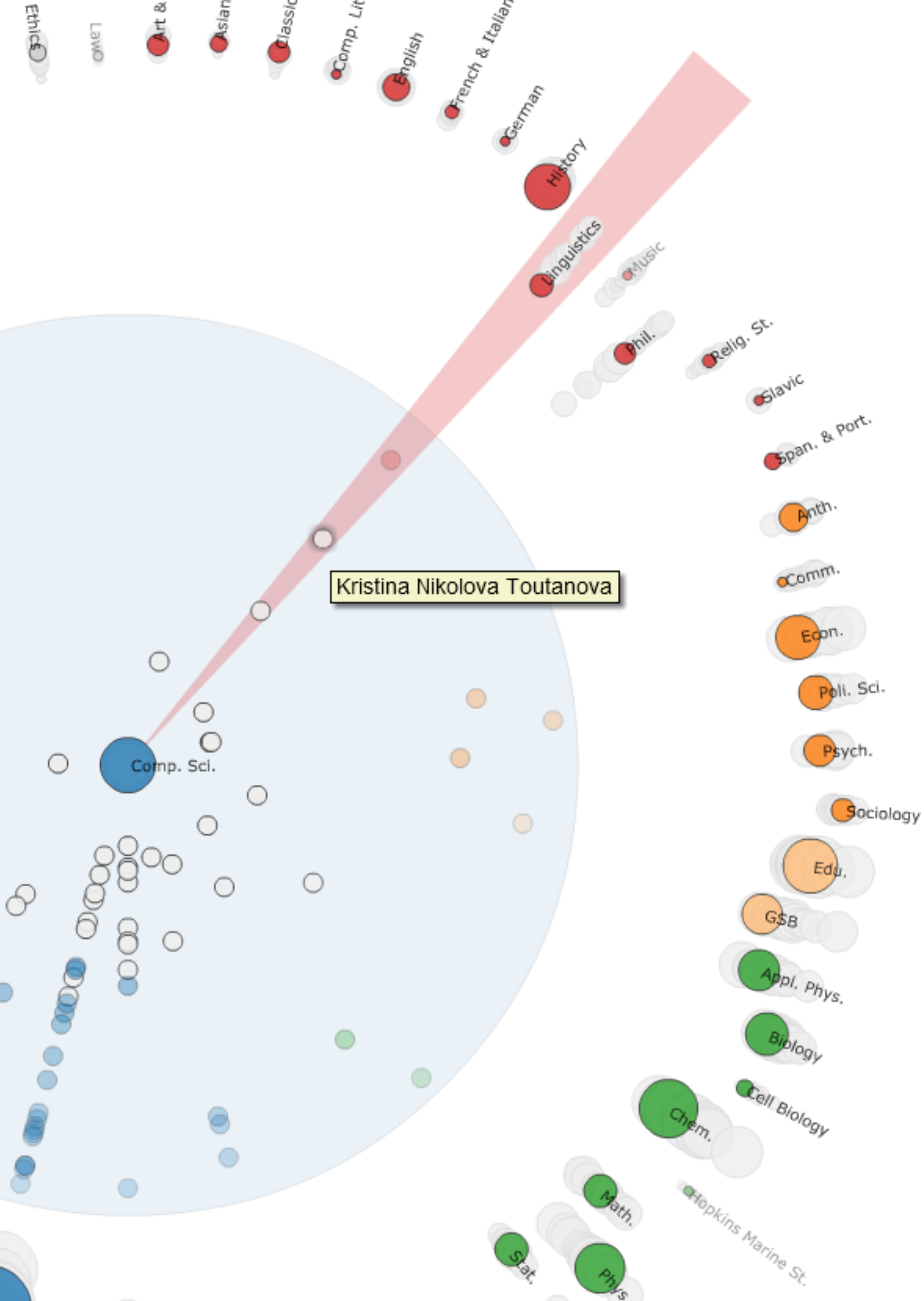
Sync. Rad. Lab
Law
MS&E
Dermatology
Medicine (Gastro)
Appl. Earth Sci.
Obst. & Gyn.
Geol. & Env.
Petrol. Eng.
Geology
Food Research
Geophys.
Indust. Eng.
Math.
Stat.
Econ.
Econ. & OR
Mat. Eng.
Op. Res.
Chem. Eng.
Civil Eng.
Mech. Eng.
Aero. & Astro.
Phys.
Poli. Sci.
Elec. Eng.
Appl. Phys.
Chem.
Cell Biology
Medicine (Clin/Pharm)
Sociology
Comp. Sci.
Rad. Oncology
Anth.
German
Slavic
History
Edu.
Art & Art Hist.
Biochem.
Physiology
Dev. Biology
Pediatrics
Span. & Port.
Phil.
Biology
Relig. St.
Comm.
Psych.
Genetics
Microbio. & Immu.
Modern Thoughts
English
Linguistics
Medicine
Comp. Lit.
French & Italian
Pathology
Ethics
Music
Asian Lang.
Classics
Neurobio.
Card. Surgery
Oncology
Anesthesia
Biochem.
Medicine (Gastro)
Marine St.
Bio. Eng.
Radiology
Sync. Rad. Lab
Geol. & Env.
GSB
Petrol. Eng.
Stat.

Topic Distance Between Stanford Depts

Area of circles denote number of theses in a given year.
Depts with no thesis produced are faded out.
Purple = Medicine
Green = Sciences
Blue = Engineering
Orange/Pink = Humanities



Oh, the humanities!



Effective statistical models for syntactic and semantic disambiguation

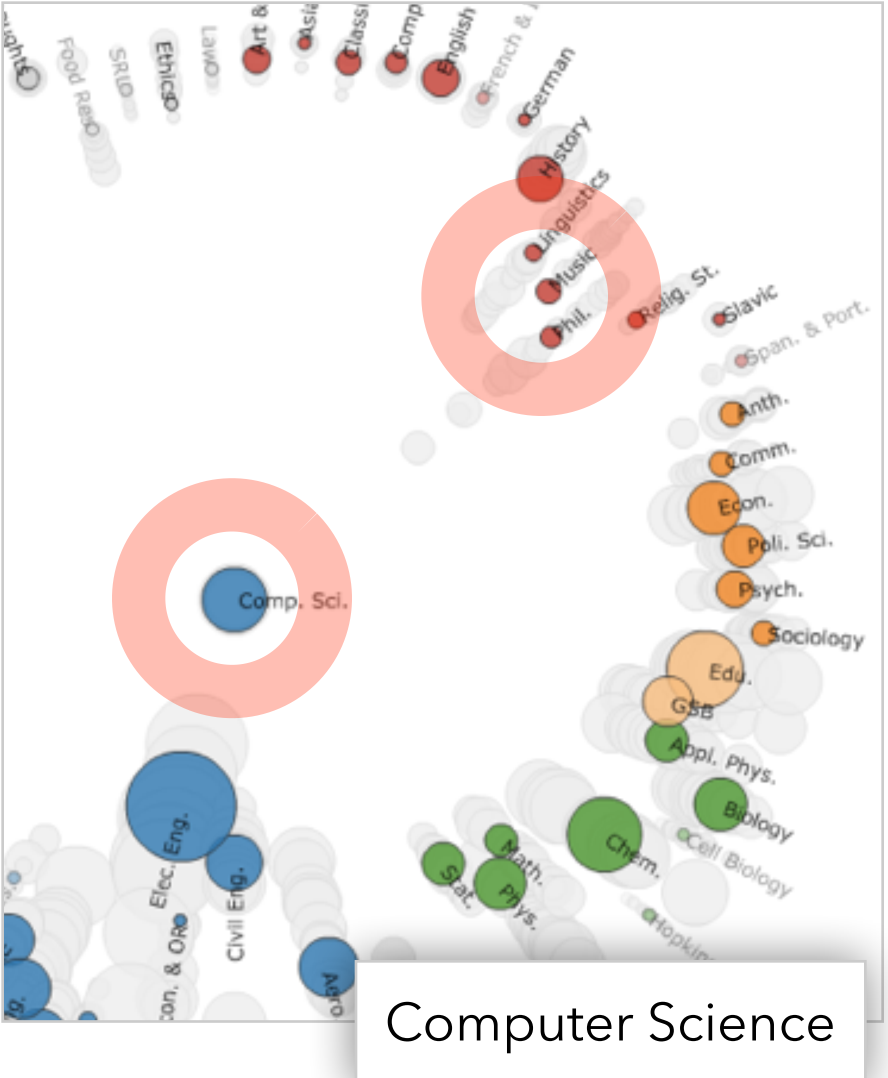
Student: Kristina Nikolova Toutanova
Advisor: Christopher D. Manning

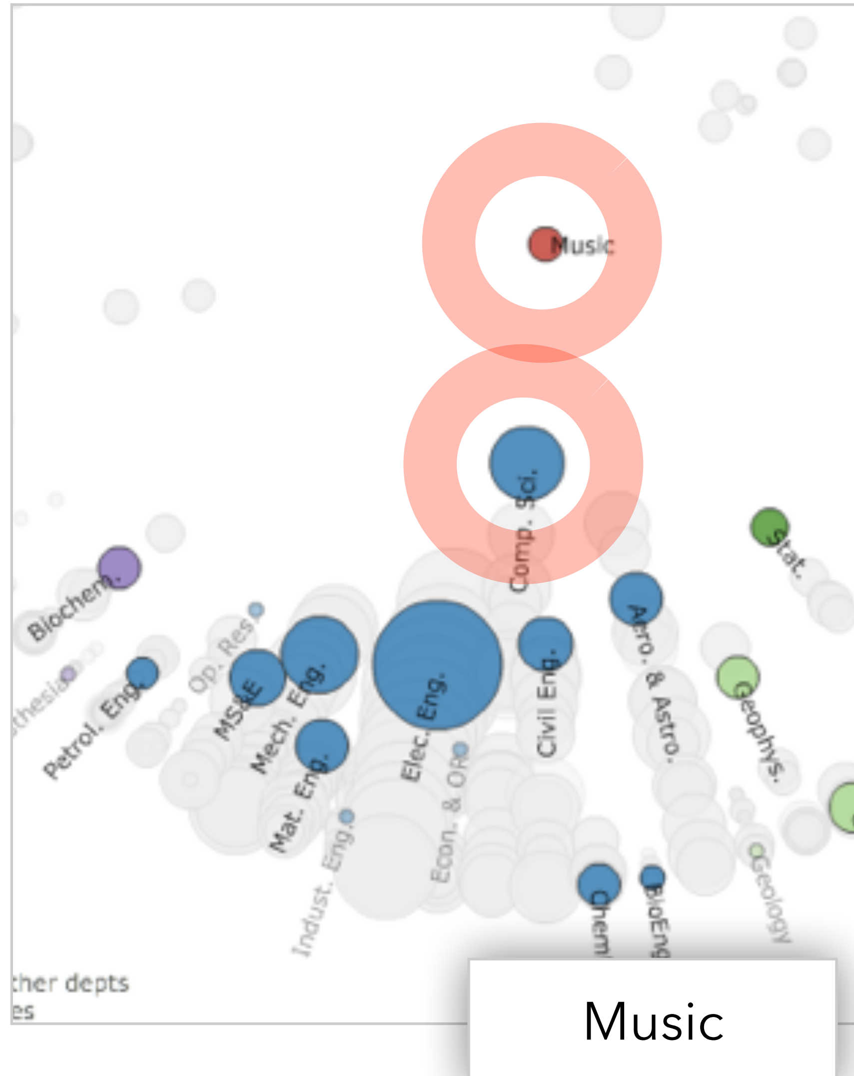
Computer Science (2005)

Keywords: Syntactic, Semantic, Tree kernels, Parsing

Abstract:

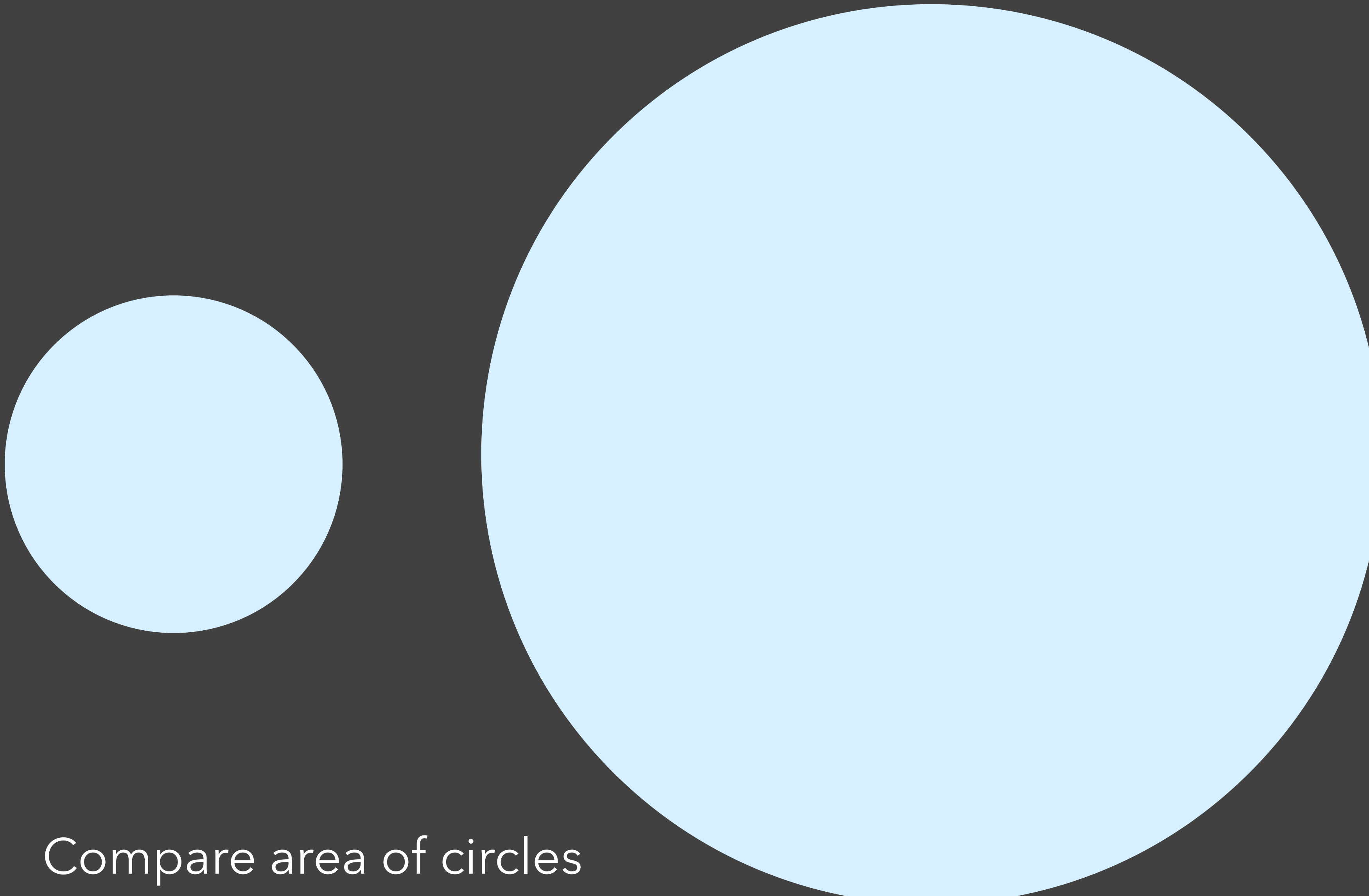
This thesis focuses on building effective statistical models for disambiguation of sophisticated syntactic and semantic natural language (NL) structures. We advance the state of the art in several domains by (i) choosing representations that encode domain knowledge more effectively and (ii) developing machine learning algorithms that deal with the specific properties of NL disambiguation tasks--sparsity of training data and large, structured spaces of hidden labels. For the task of syntactic disambiguation, we propose a novel representation of parse trees that connects the words of the sentence with the hidden syntactic structure in a direct way. Experimental evaluation on parse selection for a Head Driven Phrase Structure Grammar shows the new representation achieves superior performance compared to previous models. For the task of disambiguating the semantic role structure of verbs, we build a more accurate model, which captures the knowledge that the semantic frame of a verb is a joint structure with strong dependencies between arguments. We achieve this using a Conditional Random Field without Markov independence assumptions on the sequence of semantic role labels. To address the sparsity problem in machine learning for NL, we develop a method for incorporating many additional sources of information, using Markov chains in the space of words. The Markov chain framework makes it possible to combine multiple knowledge sources, to learn how much to trust each of them, and to chain inferences together. It achieves large gains in the task of disambiguating prepositional phrase attachments.



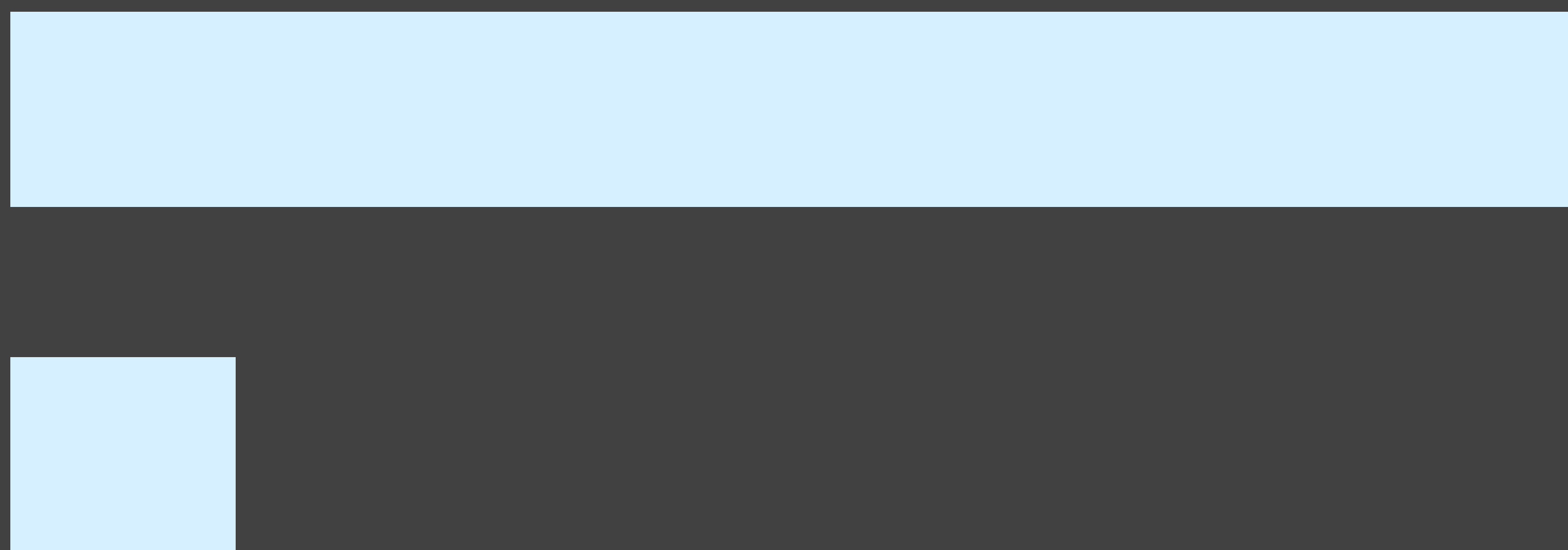


“Word Borrowing” via Labeled LDA

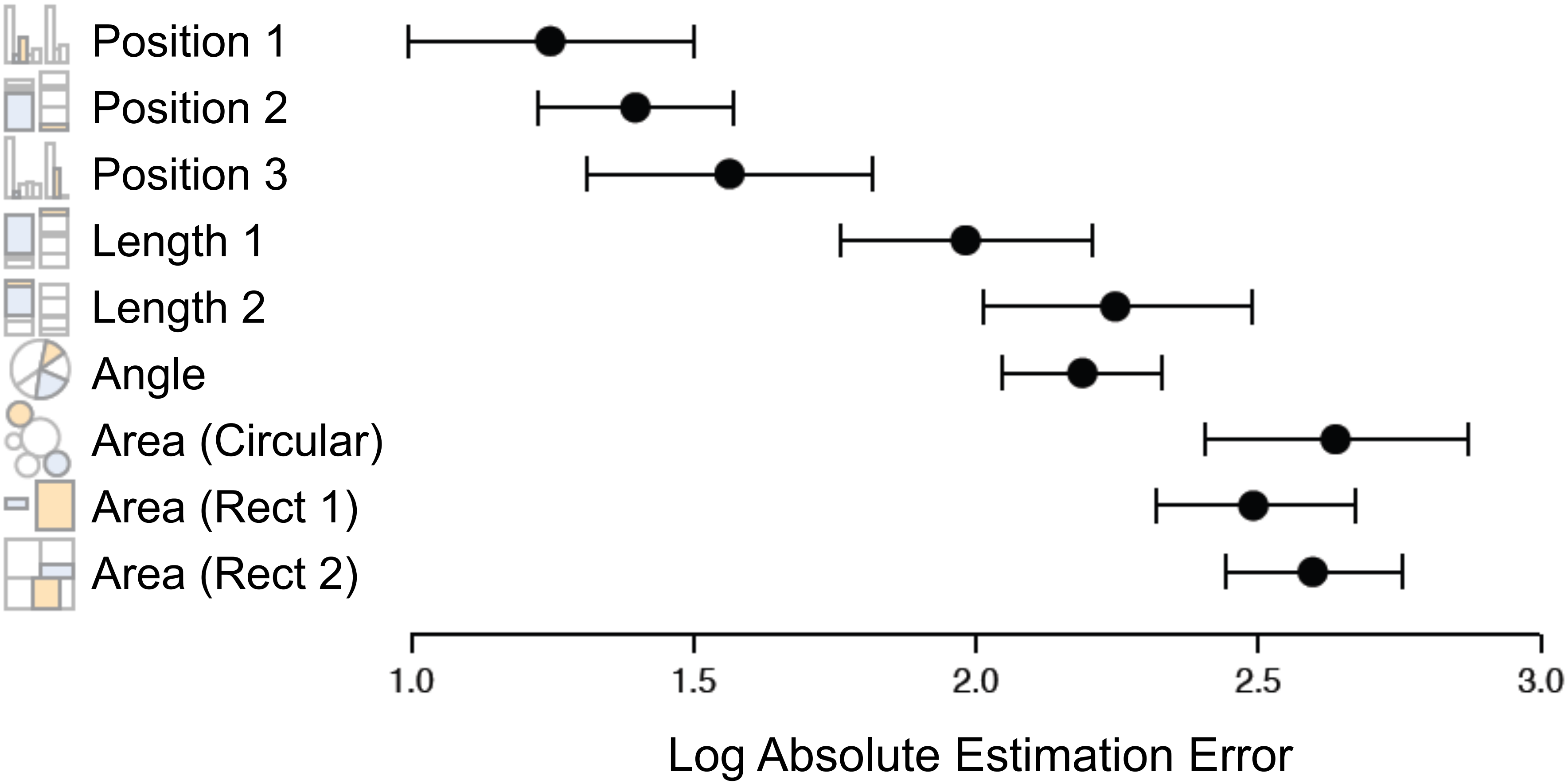
What makes a visualization **“good”**?



Compare area of circles



Compare length of bars



Graphical Perception Experiments

Empirical estimates of encoding effectiveness

Estimating Proportions

Most accurate

Least accurate

Position (common) scale

Position (non-aligned) scale

Length

Slope

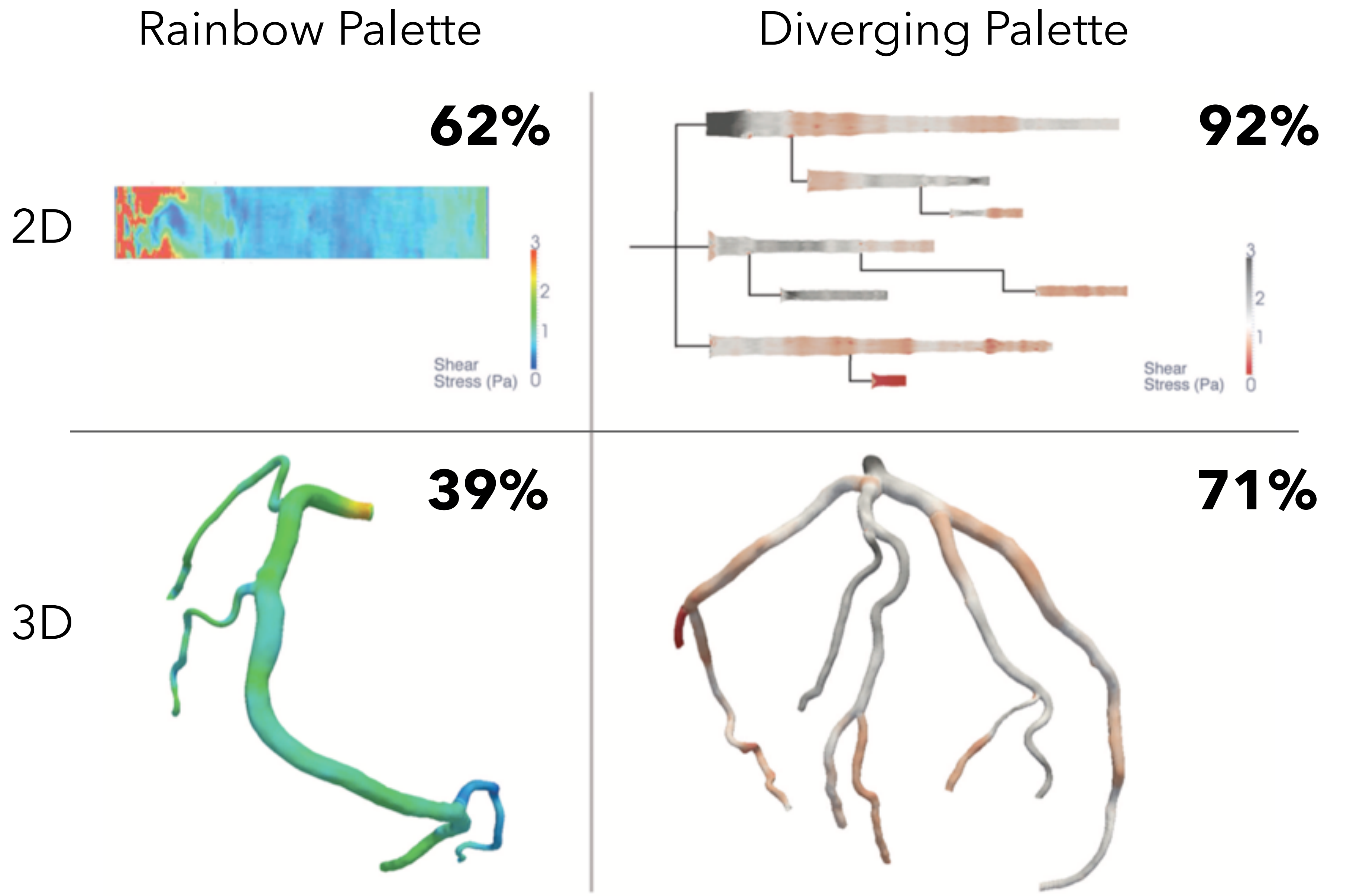
Angle

Area

Volume

Color hue-saturation-density

Artery Visualization [Borkin et al. '11]



INTERACTIVE
Data Analysis

Jeffrey Heer @jeffrey_heer
Univ. of Washington + Trifacta