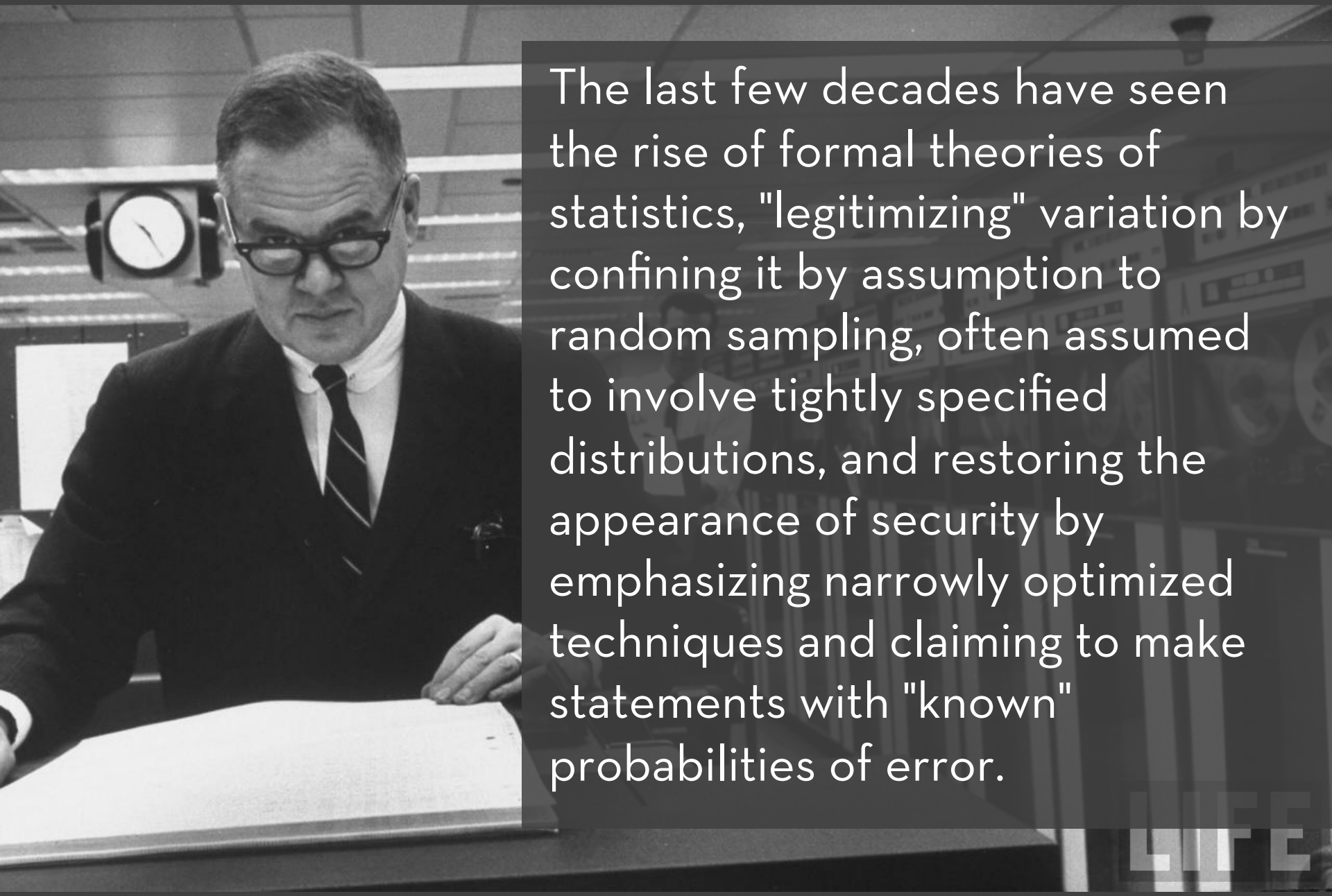
PEOPLE + DATA + COMPUTATION

Working Productively with Data

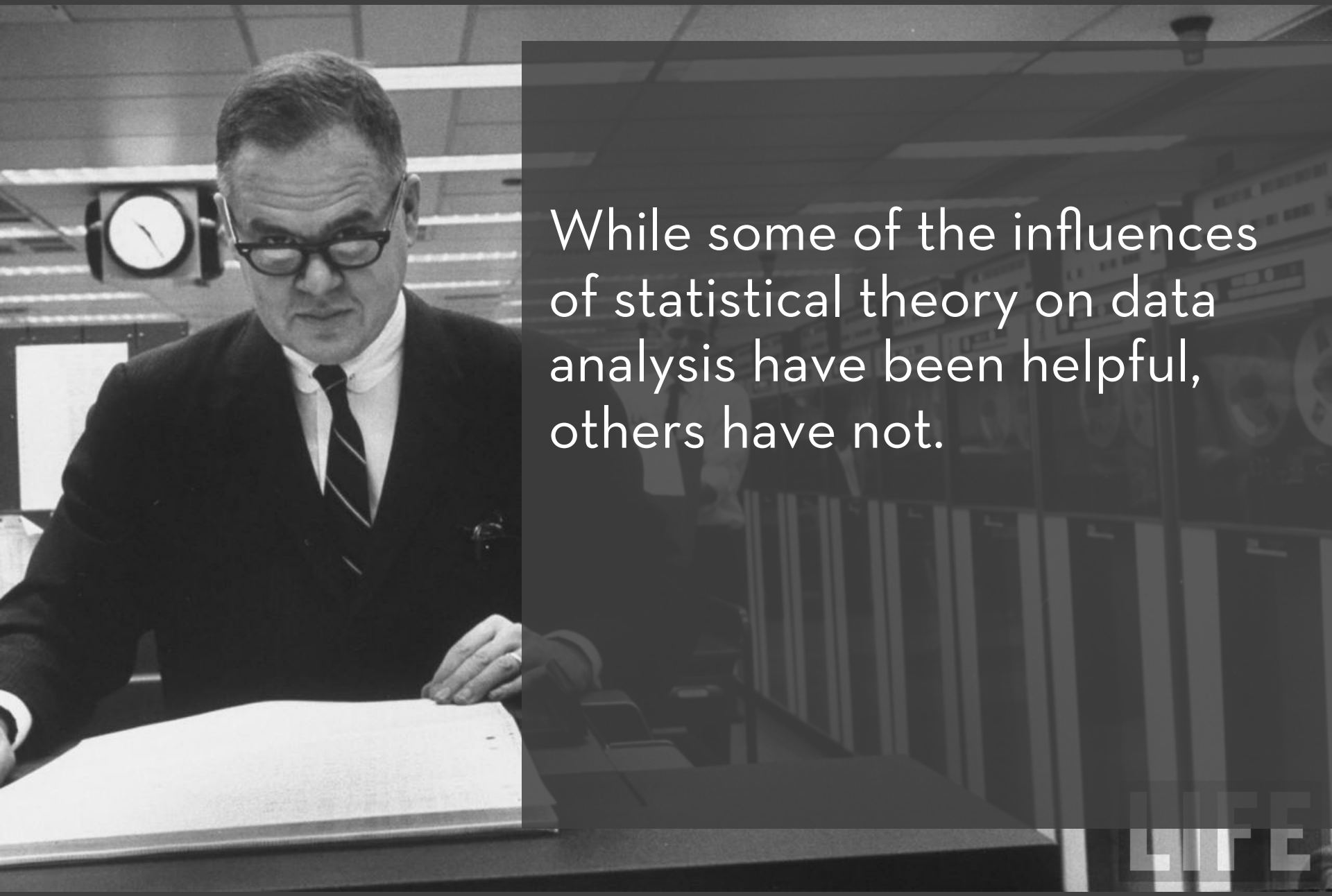
Jeffrey Heer
Joe Hellerstein

TRIFACTA

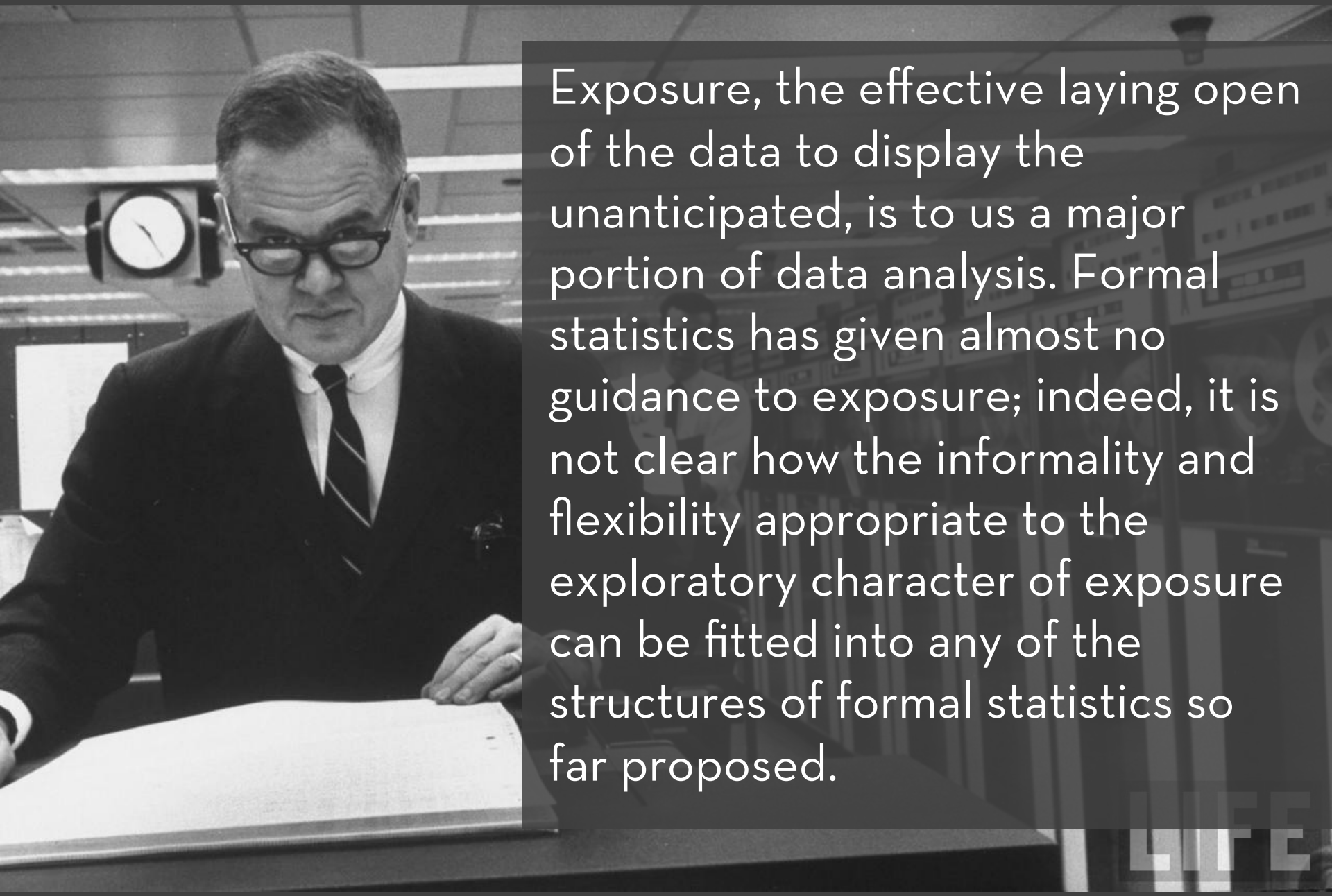
The Future of Data Analysis, Tukey 1962



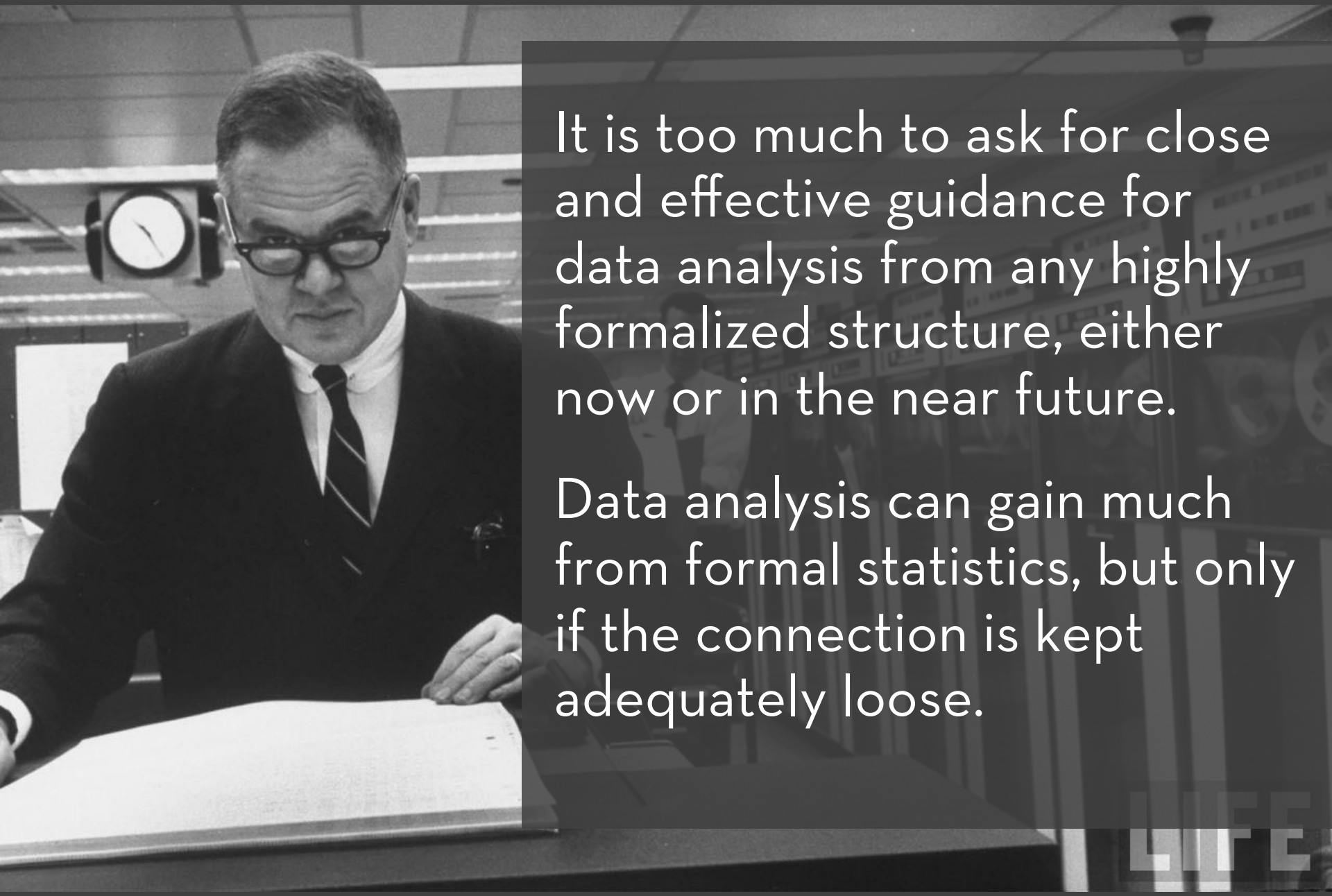
The last few decades have seen the rise of formal theories of statistics, "legitimizing" variation by confining it by assumption to random sampling, often assumed to involve tightly specified distributions, and restoring the appearance of security by emphasizing narrowly optimized techniques and claiming to make statements with "known" probabilities of error.



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



Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis. Formal statistics has given almost no guidance to exposure; indeed, 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.



It is too much to ask for close and effective guidance for data analysis from any highly formalized structure, either now or in the near future.

Data analysis can gain much from formal statistics, but only if the connection is kept adequately loose.

The Future of Data Analysis, Tukey 1962

Visualization

Acquisition

Cleaning

Integration

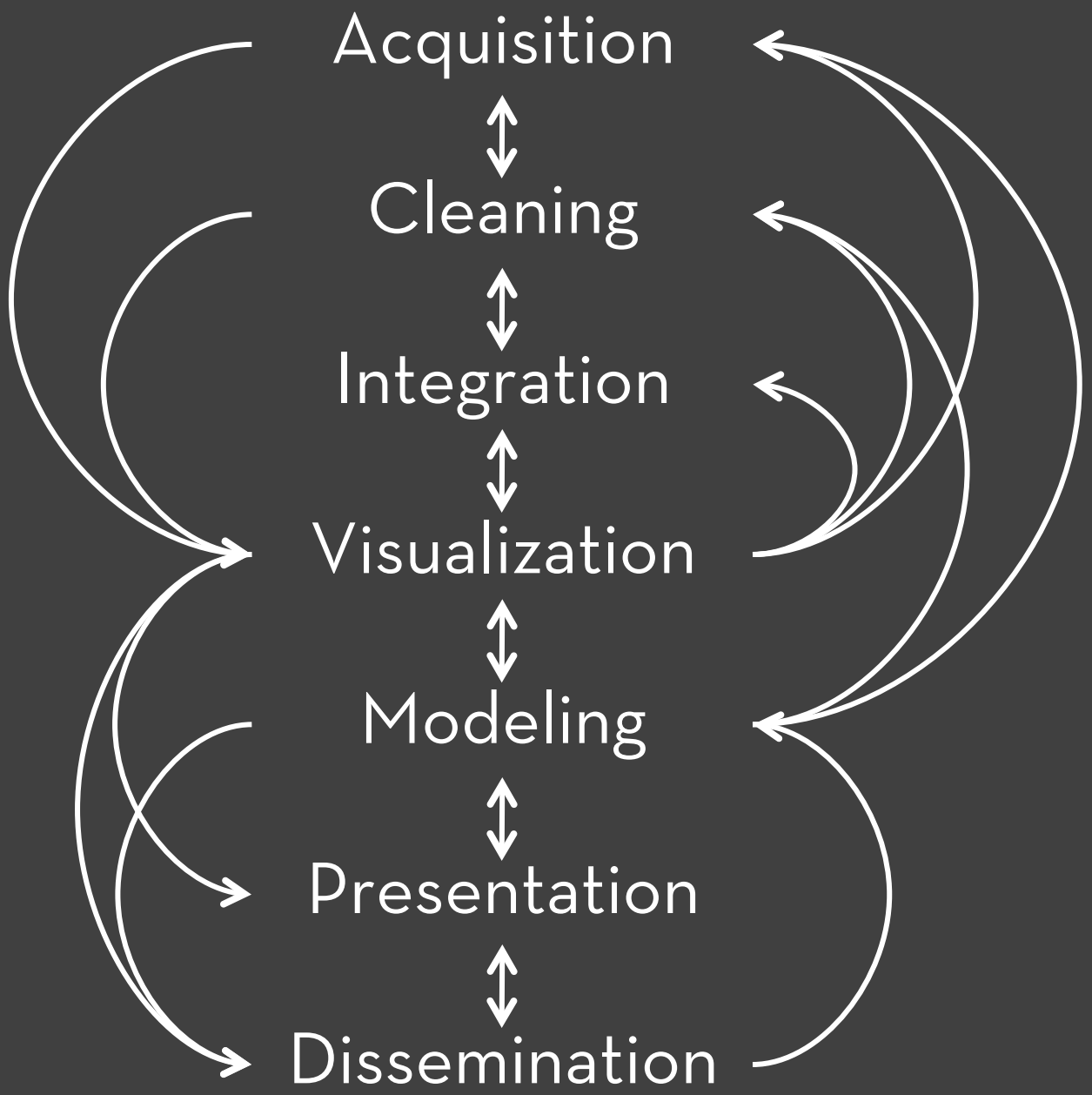
Visualization

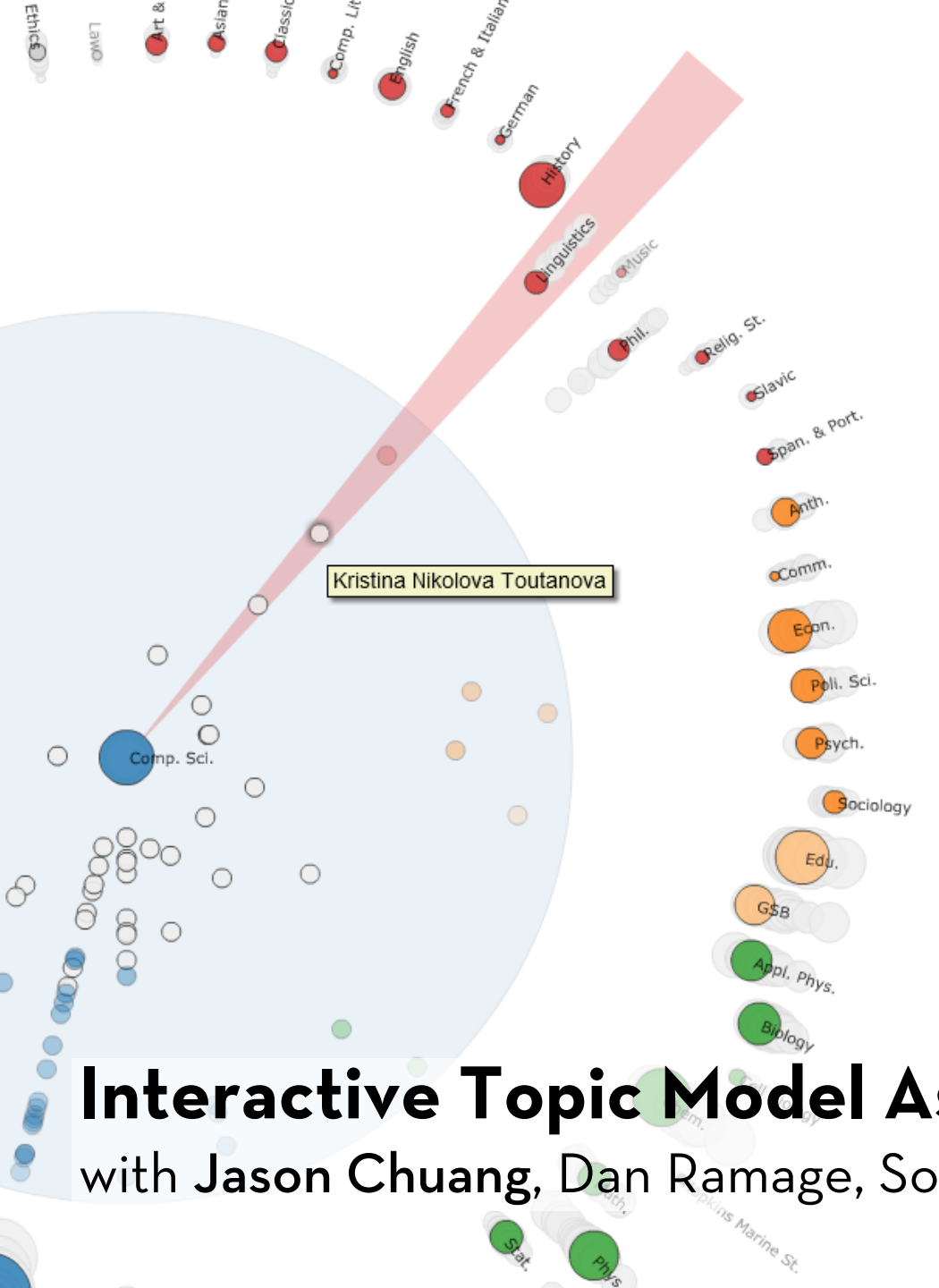
Modeling

Presentation

Dissemination


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination



Effective statistical models for syntactic and semantic disambiguation

Student: Kristina Nikolova Toutanova
Advisor: Christopher D. Manning

Computer Science (2005)

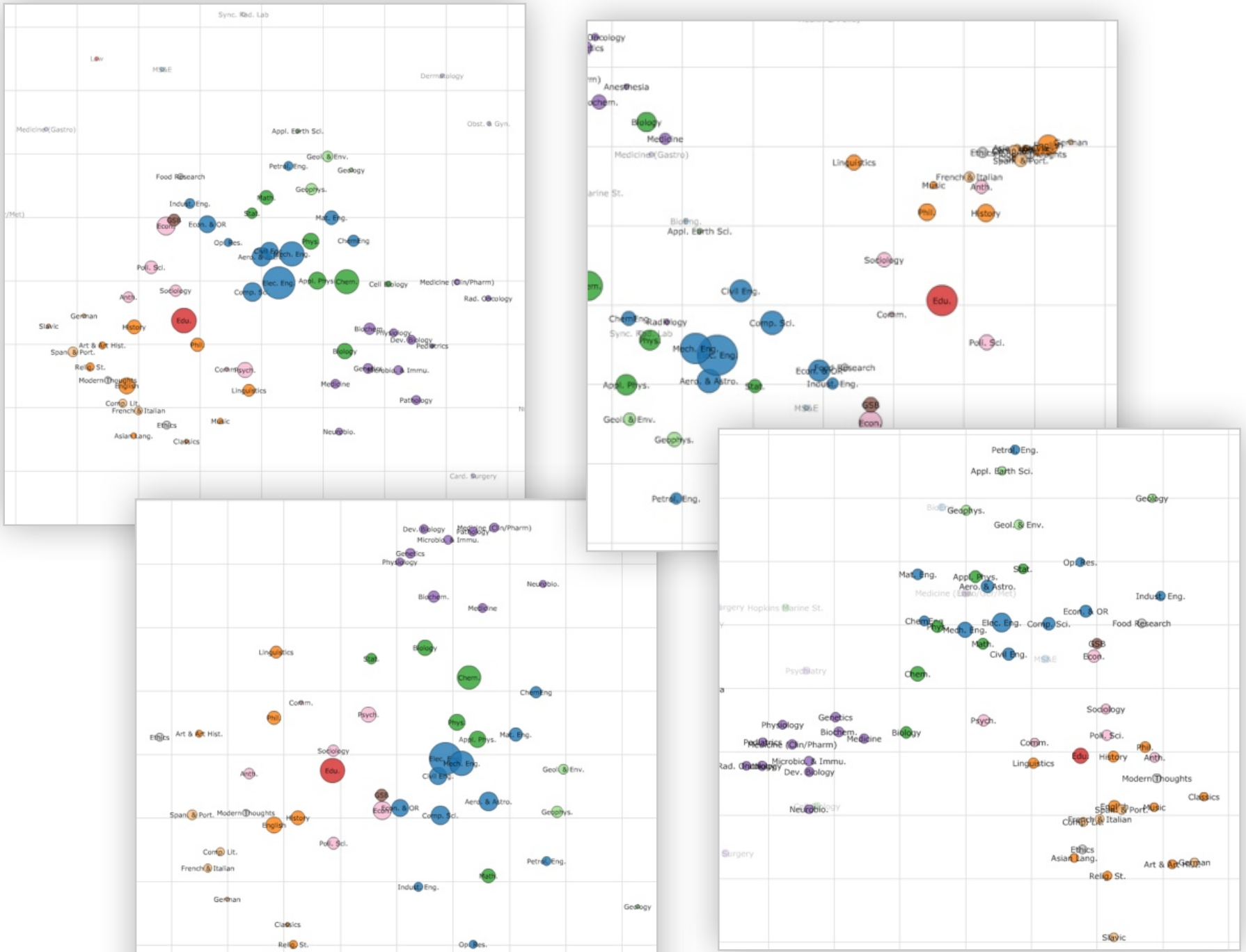
Keywords: Syntactic, Semantic, Tree kernels, Parsing

Abstract:

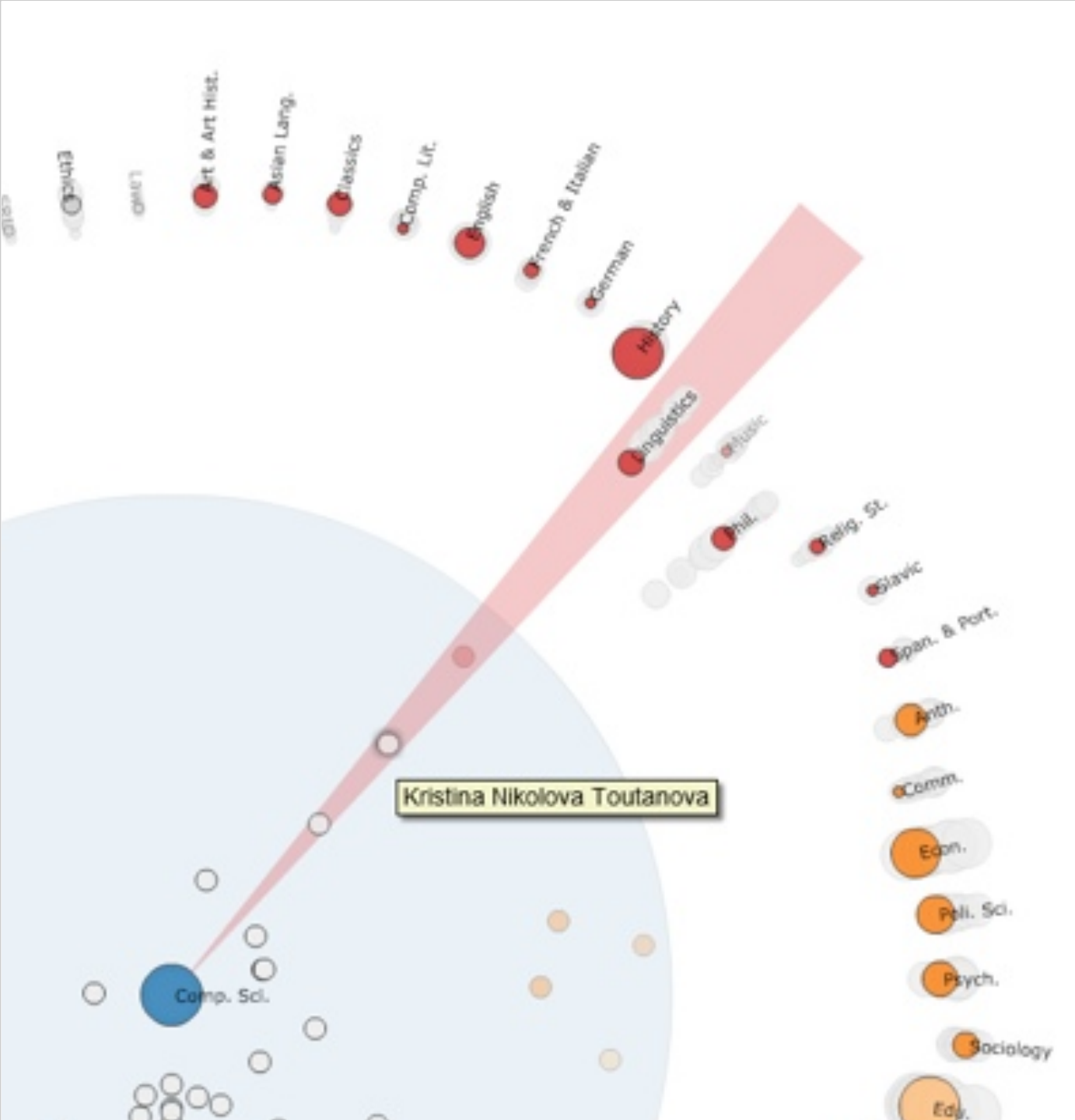
This thesis focuses on building effective statistical models for disambiguation of sophisticated syntactic and semantic natural language (NL) structures. We advance the state of the art in several domains by (i) choosing representations that encode domain knowledge more effectively and (ii) developing 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.

Interactive Topic Model Assessment

with **Jason Chuang**, Dan Ramage, Sonal Gupta & Chris Manning


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

Oh, the humanities!



Effective statistical models for sy

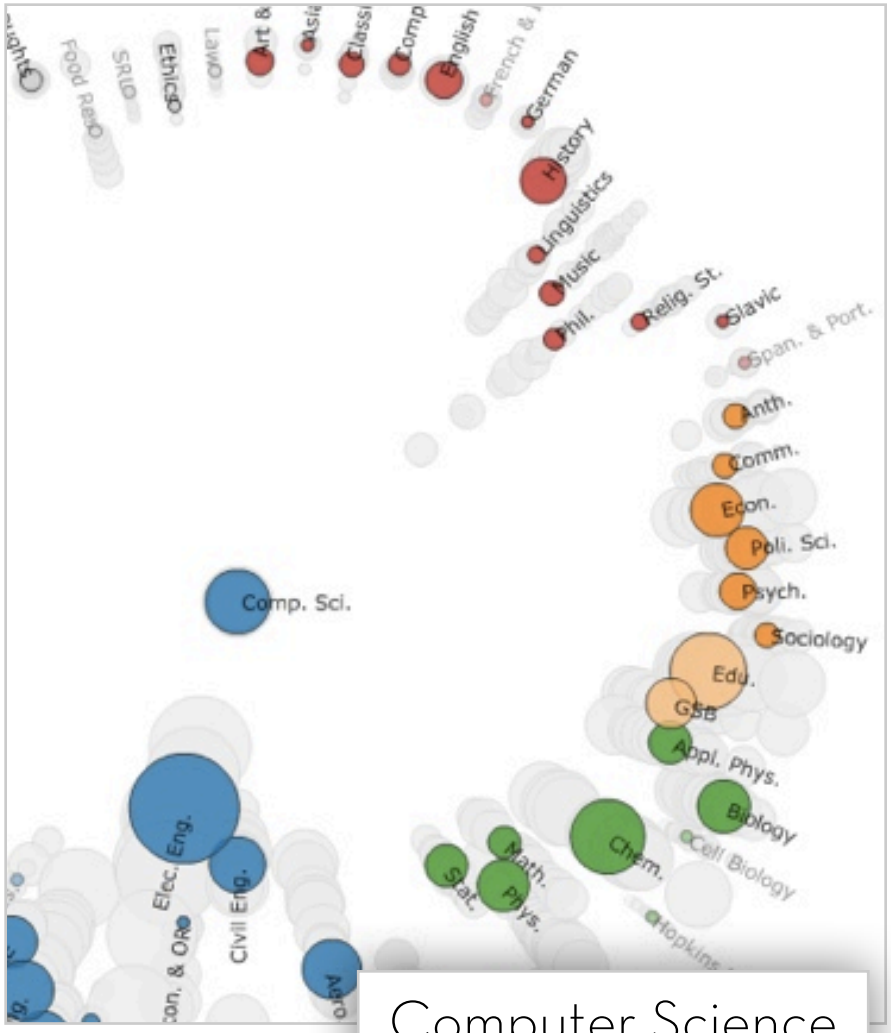
Student: Kristina Nikolova Toutanov
Advisor: Christopher D. Manning

Computer Science (2005)

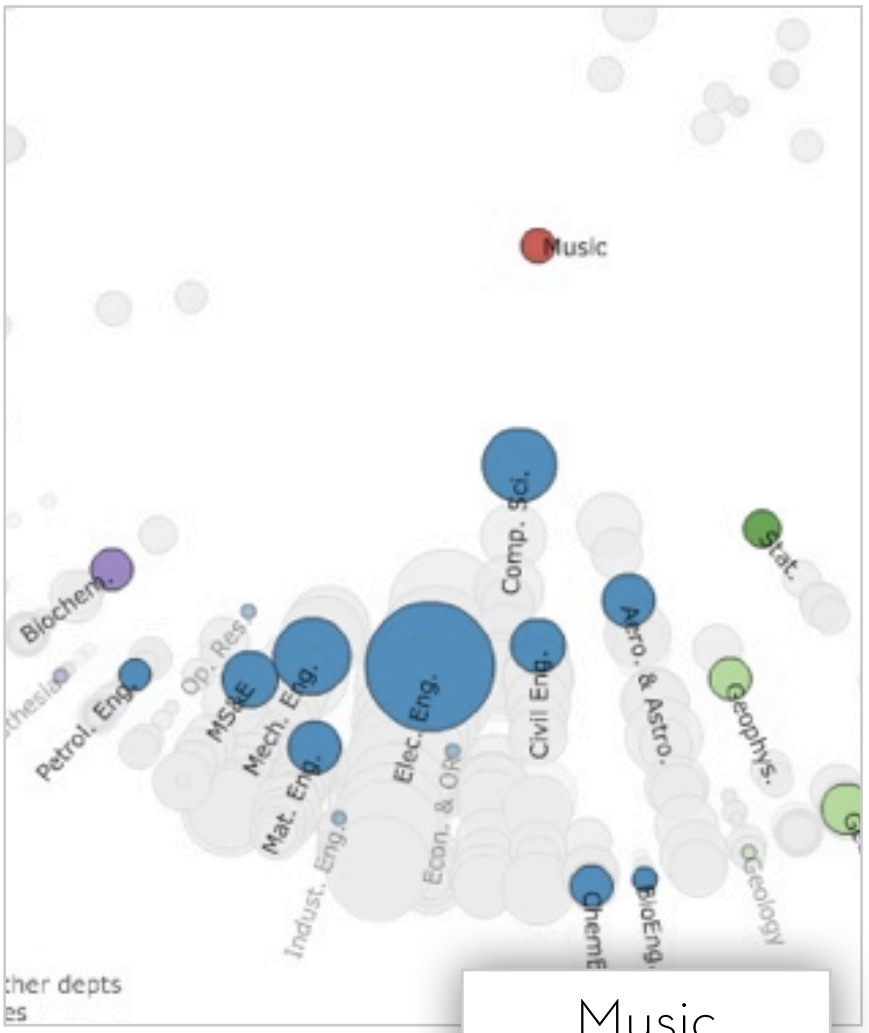
Keywords: Syntactic, Semantic, Tre

Abstract:

This thesis focuses on building effe
disambiguation of sophisticated sy
(NL) structures. We advance the st
(i) choosing representations that e
effectively and (ii) developing machi
the specific properties of NL disamb
data and large, structured spaces o
syntactic disambiguation, we propo
trees that connects the words of th
structure in a direct way. Experimen
a Head Driven Phrase Structure Gra
achieves superior performance com
task of disambiguating the semanti
more accurate model, which capture
frame of a verb is a joint structure w
arguments. We achieve this using a
Markov independence assumptions
labels. To address the sparsity prob
develop a method for incorporating
information, using Markov chains in
chain framework makes it possible t
sources, to learn how much to trust
inferences together. It achieves lar
disambiguating prepositional phras

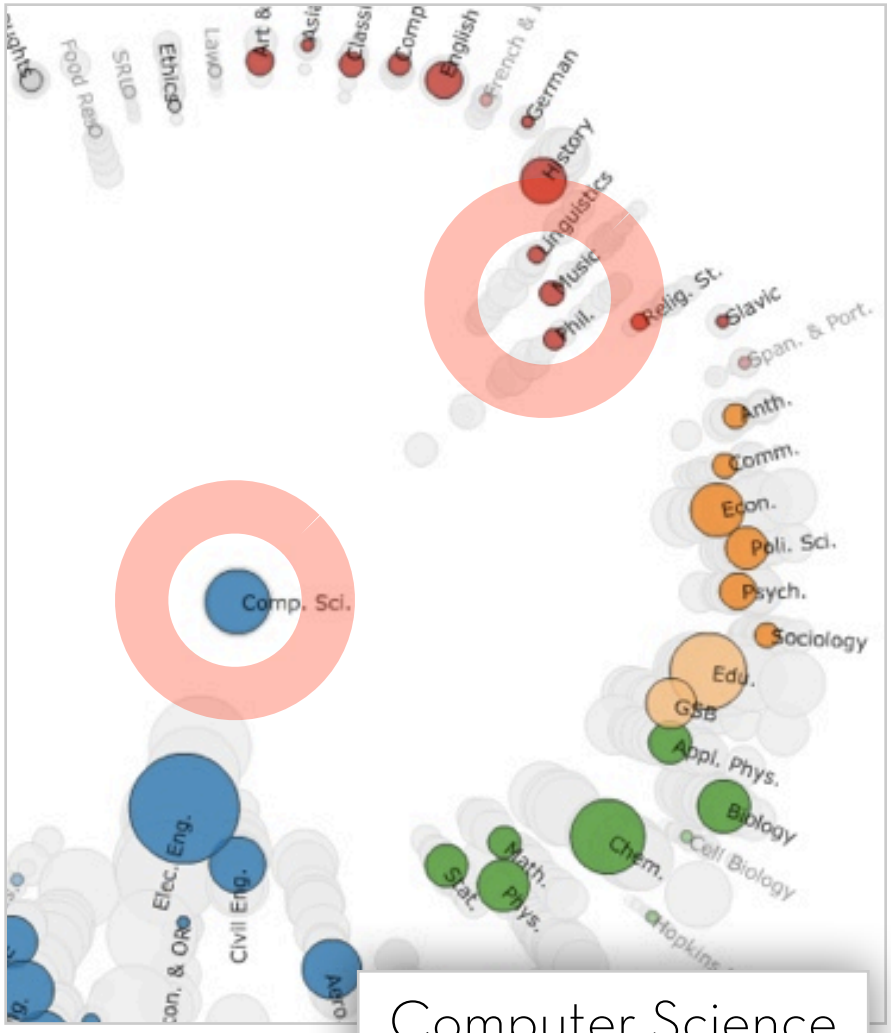


Computer Science

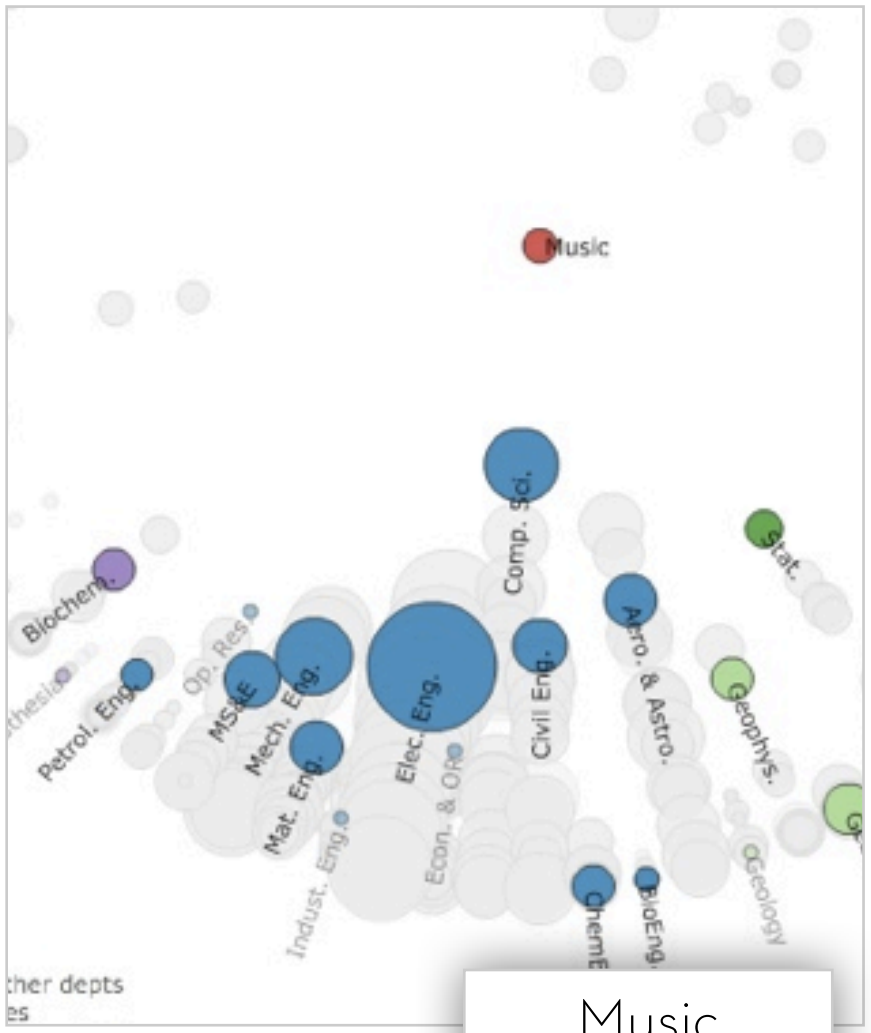


Music

“Word Borrowing” via Labeled LDA

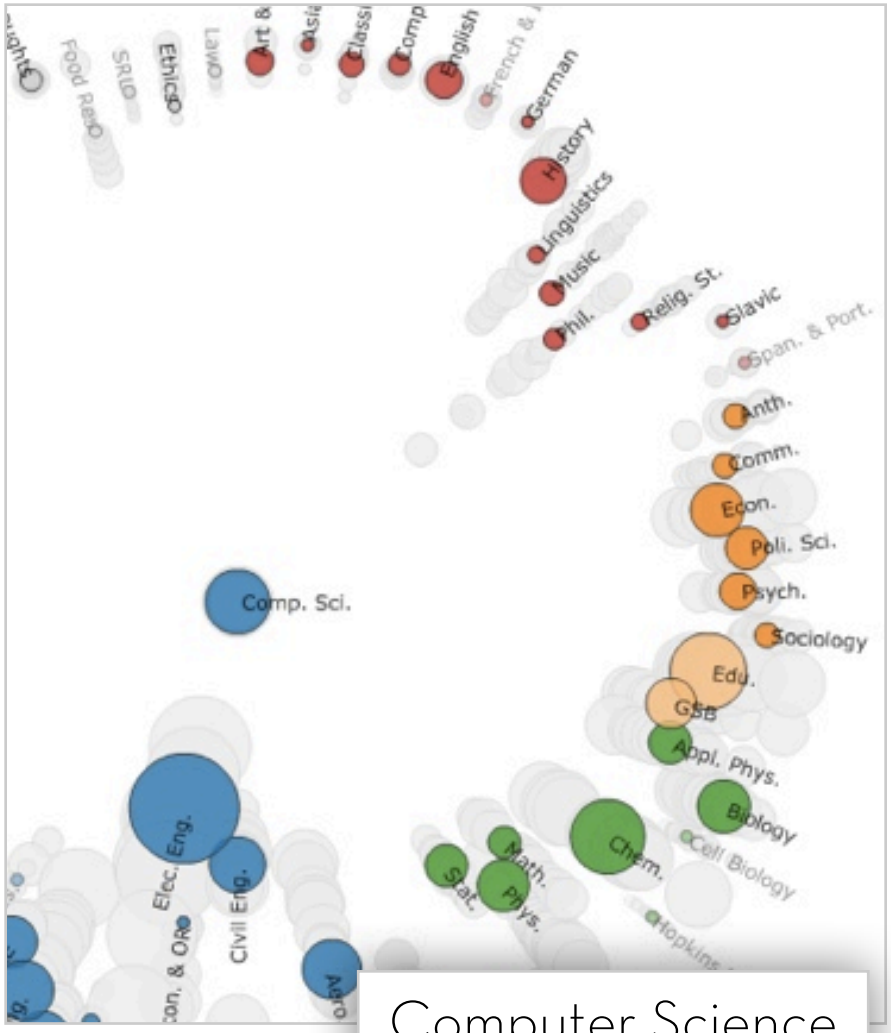


Computer Science

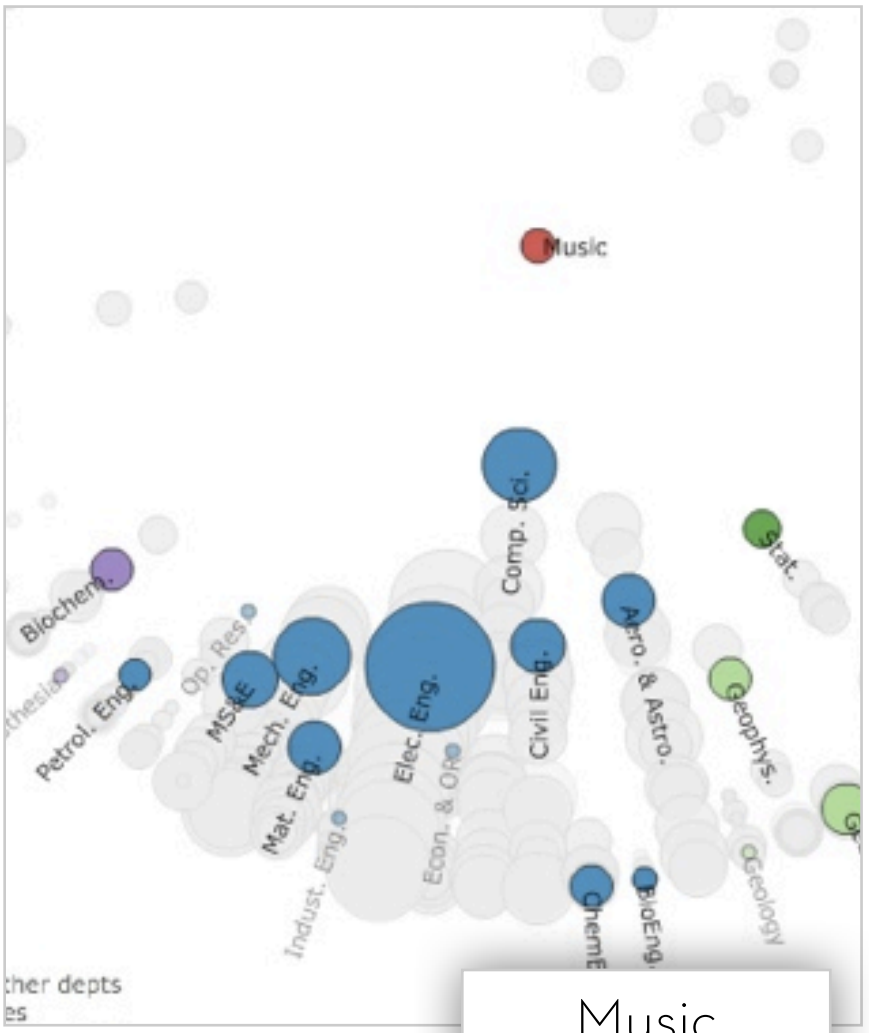


Music

“Word Borrowing” via Labeled LDA

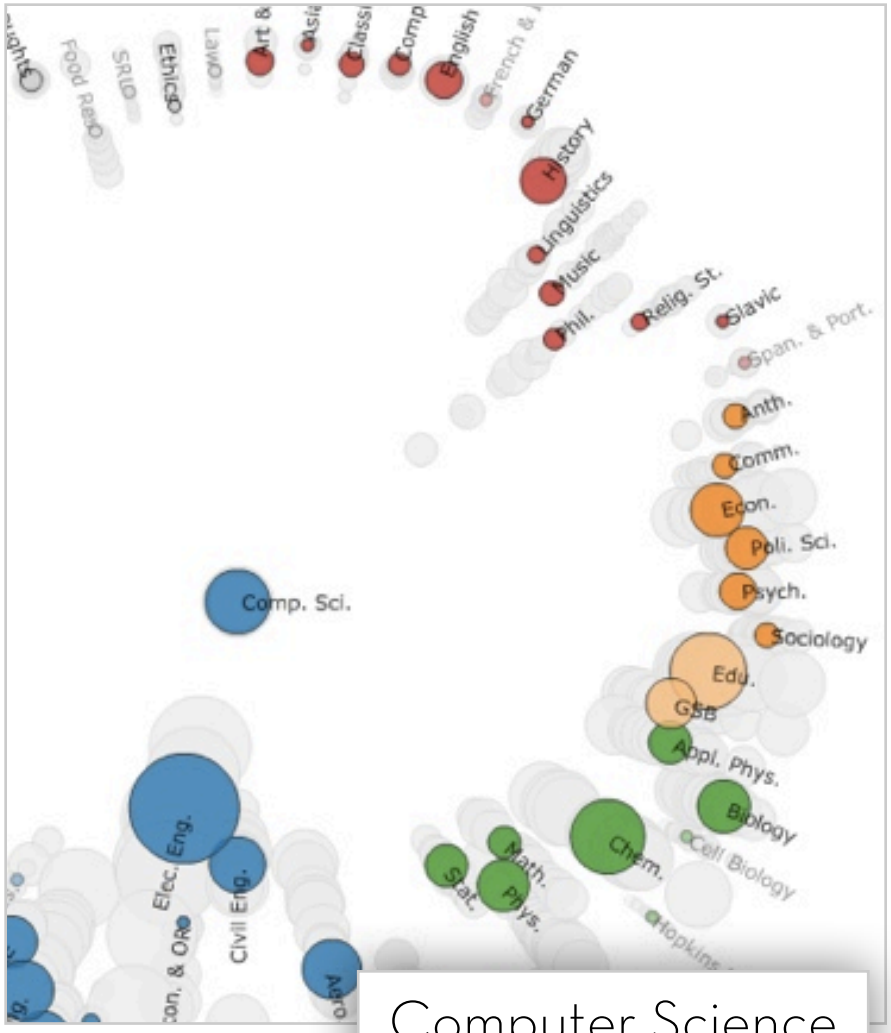


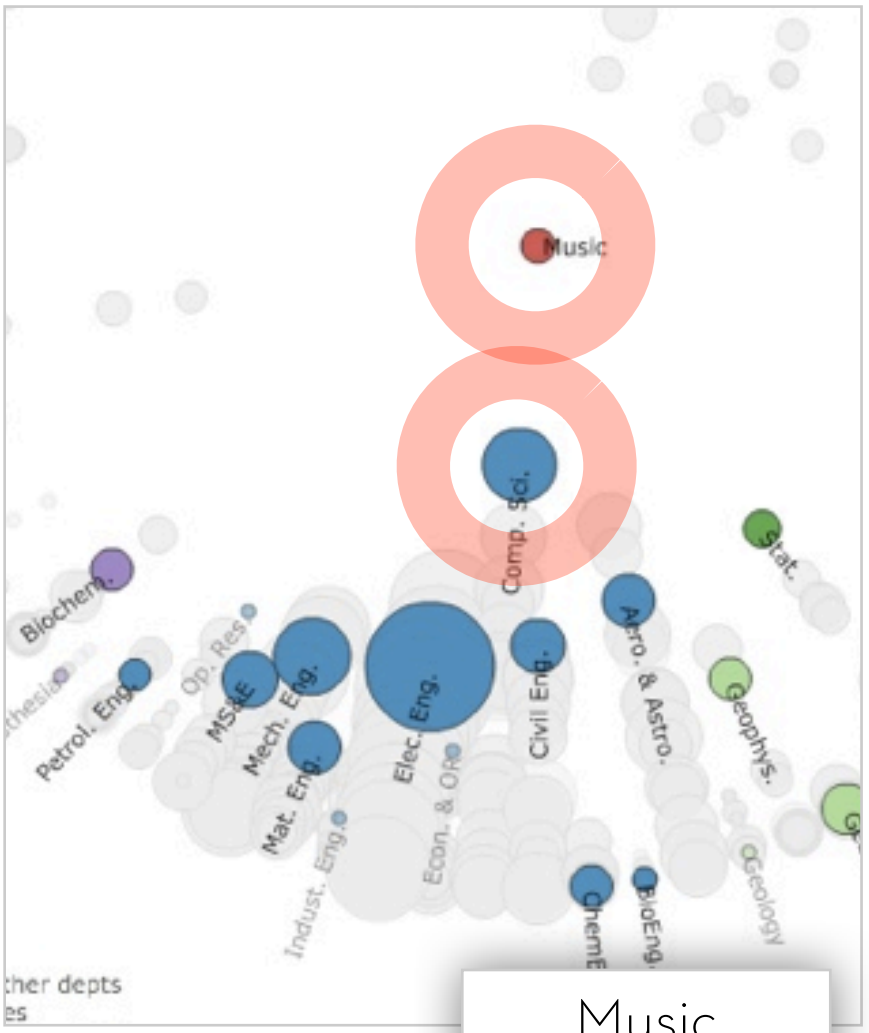
Computer Science



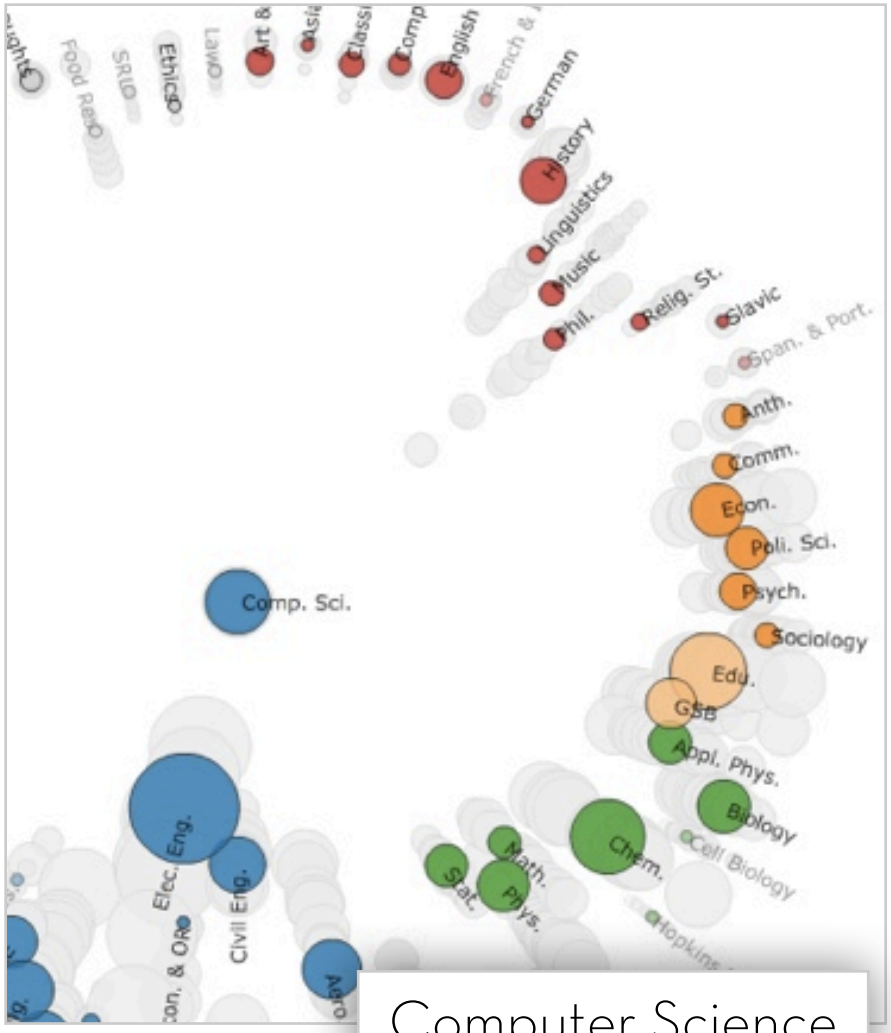
Music

“Word Borrowing” via Labeled LDA

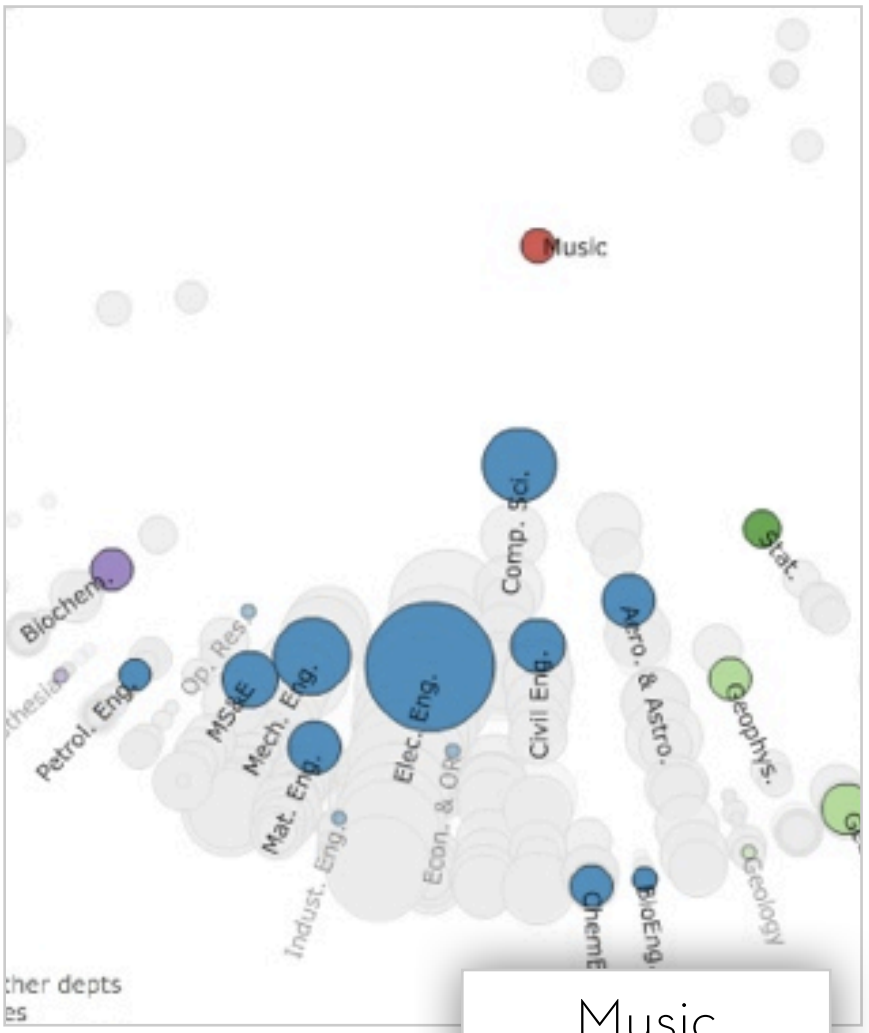




“Word Borrowing” via Labeled LDA



Computer Science

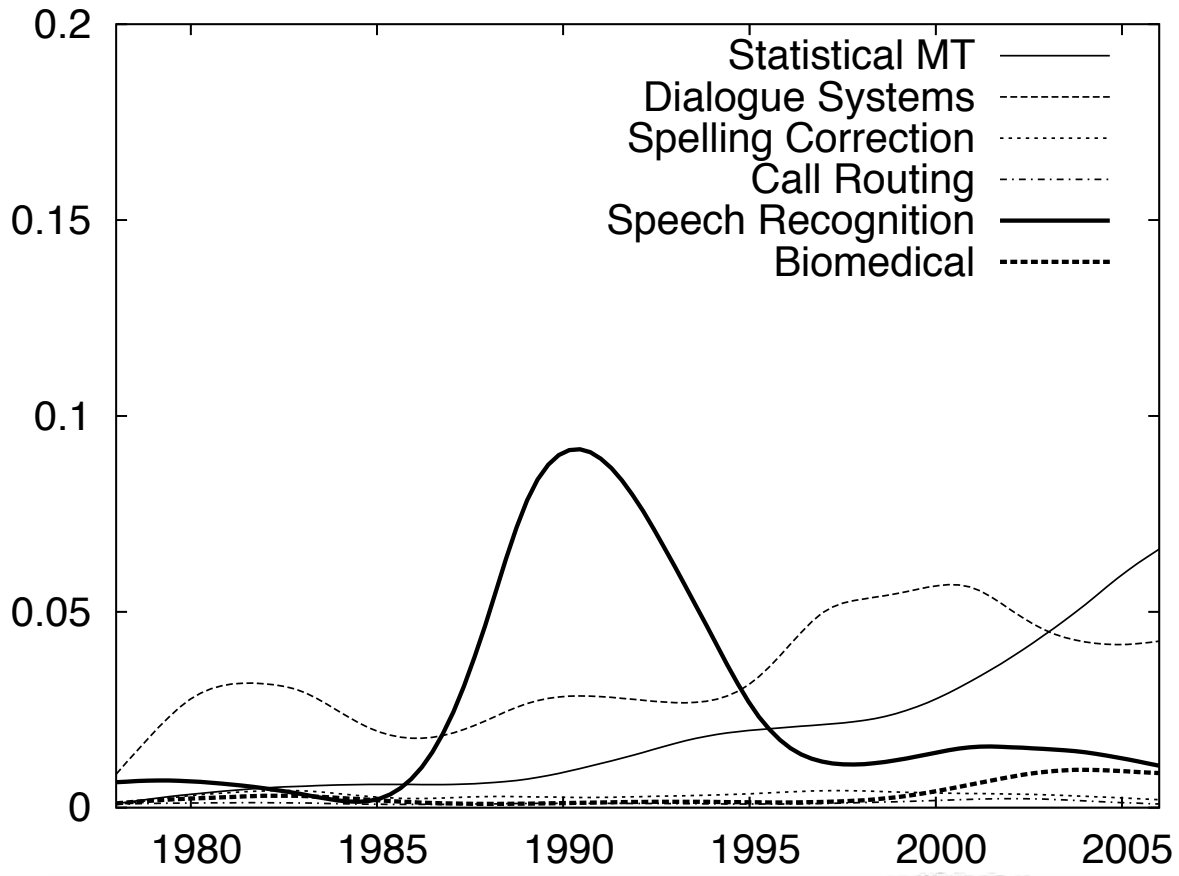


Music

“Word Borrowing” via Labeled LDA

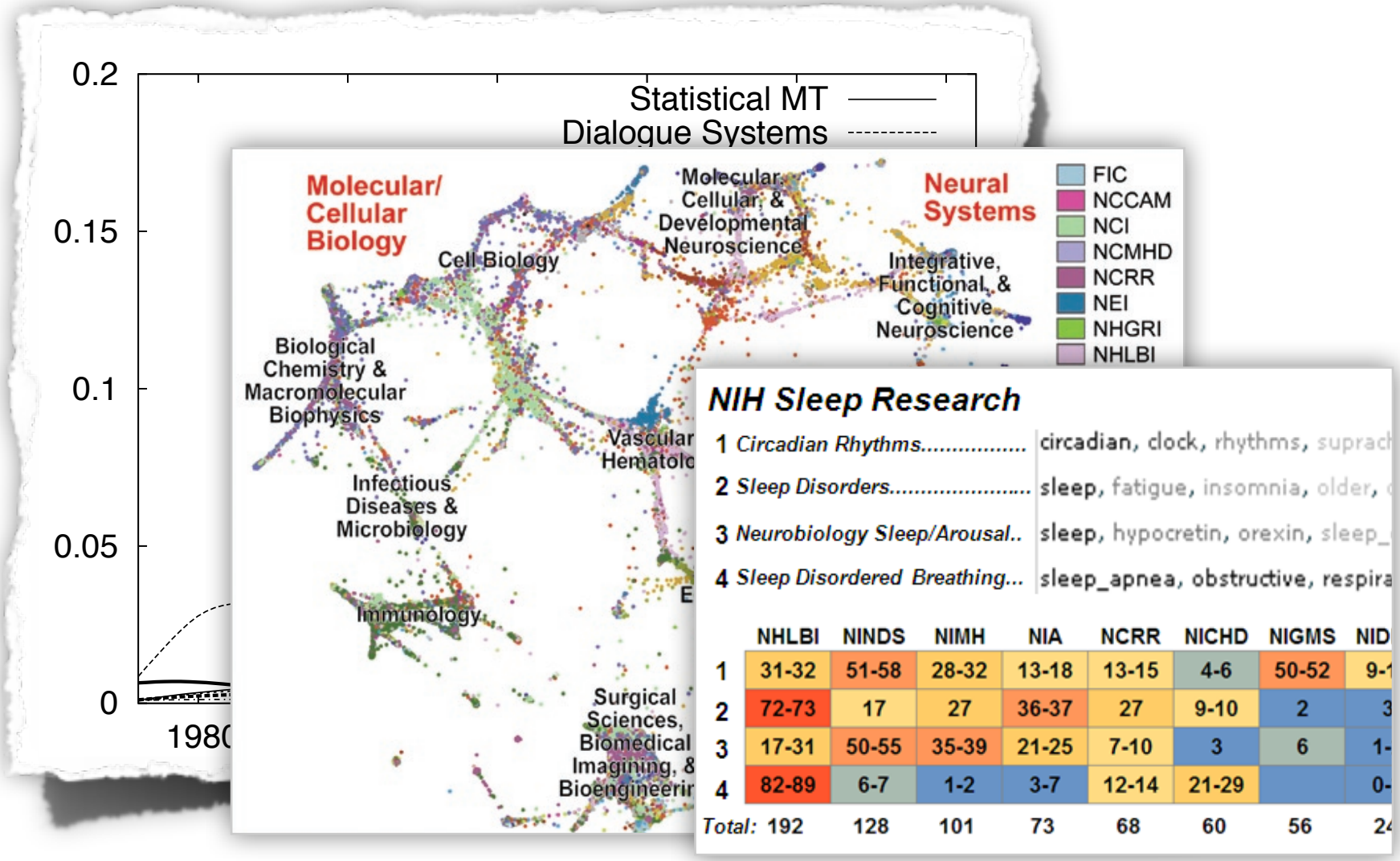
Real-World Topical Analysis

Real-World Topical Analysis



[Hall et al. 2008]

Real-World Topical Analysis



[Talley et al. 2011]

Current Practices

Current Practices

Topic Words: vegf angiogenesis vascular_endothelial_growth_factor angiogenic end
antiangiogenic anti_angiogenic vegf_a tumor_angiogenesis vegfr2 growth signaling
Title Words: angiogenesis, vegf, vascular_endothelial_growth_factor, angiogenic, tu
neovascularization, angiopoietin, signaling, vegfr, vascular, human
Phrases: vascular_endothelial_growth_factor vegf, vegf angiogenesis, vegf receptor,

[Talley et al. 2011]

Current Practices

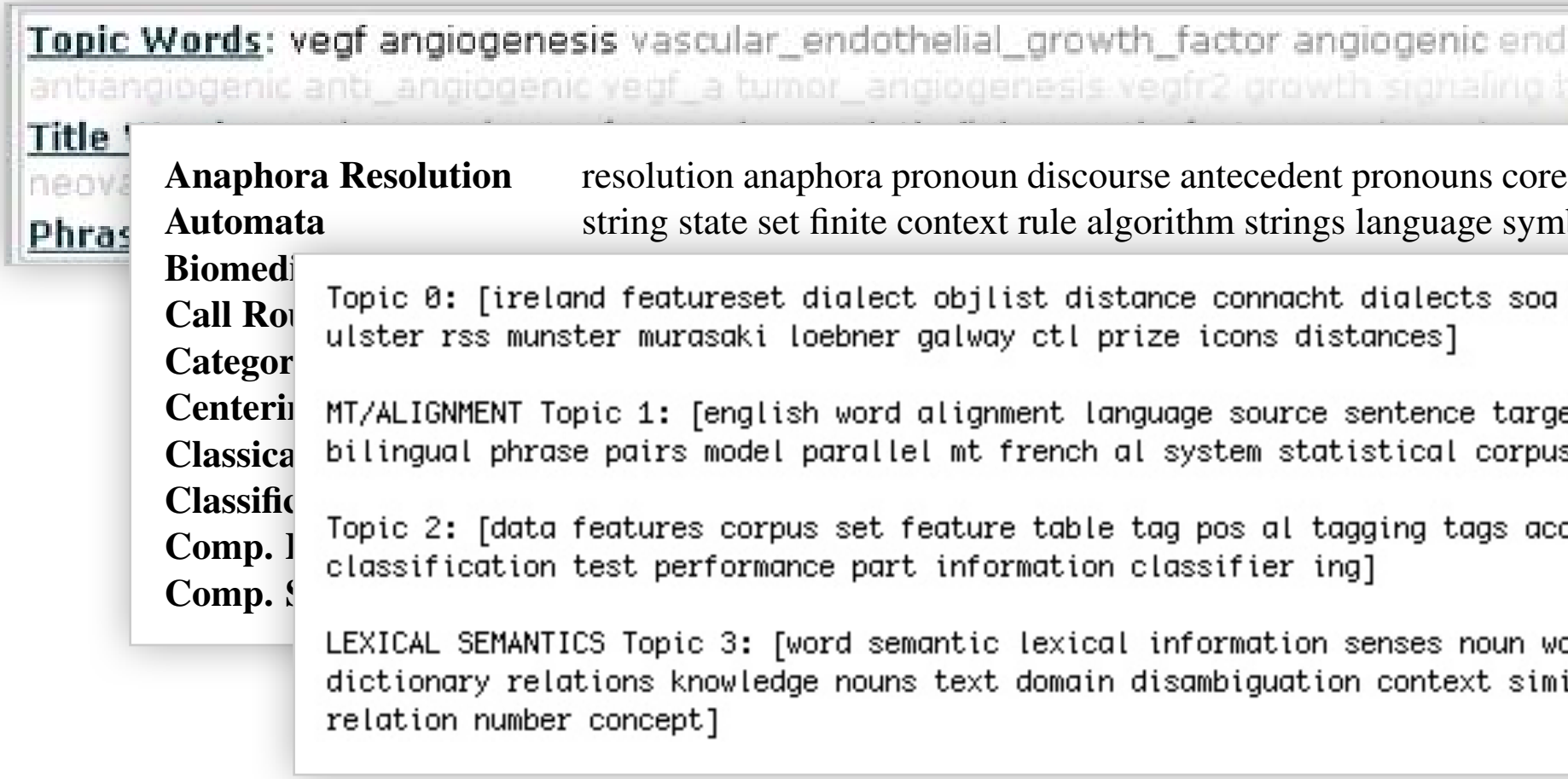
Topic Words: vegf angiogenesis vascular_endothelial_growth_factor angiogenic end
antiangiogenic anti_angiogenic vegf_a tumor_angiogenesis vegfr2 growth signaling

Title
neova
Phras

Anaphora Resolution	resolution anaphora pronoun discourse antecedent pronouns core
Automata	string state set finite context rule algorithm strings language symb
Biomedical	medical protein gene biomedical wkh abstracts medline patient c
Call Routing	call caller routing calls destination vietnamese routed router desti
Categorial Grammar	proof formula graph logic calculus axioms axiom theorem proofs
Centering*	centering cb discourse cf utterance center utterances theory coher
Classical MT	japanese method case sentence analysis english dictionary figure
Classification/Tagging	features data corpus set feature table word tag al test
Comp. Phonology	vowel phonological syllable phoneme stress phonetic phonology
Comp. Semantics*	semantic logical semantics john sentence interpretation scope log

[Hall et al. 2008]

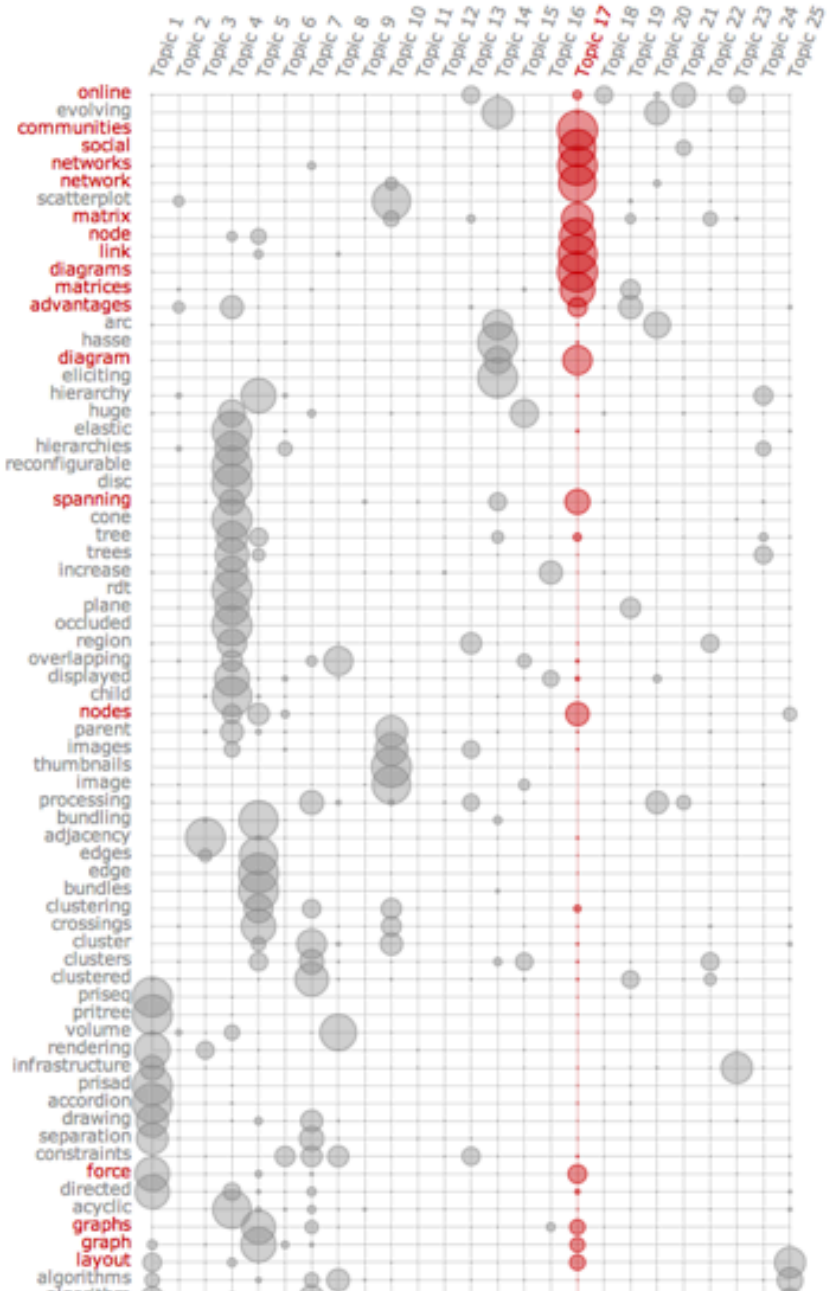
Current Practices

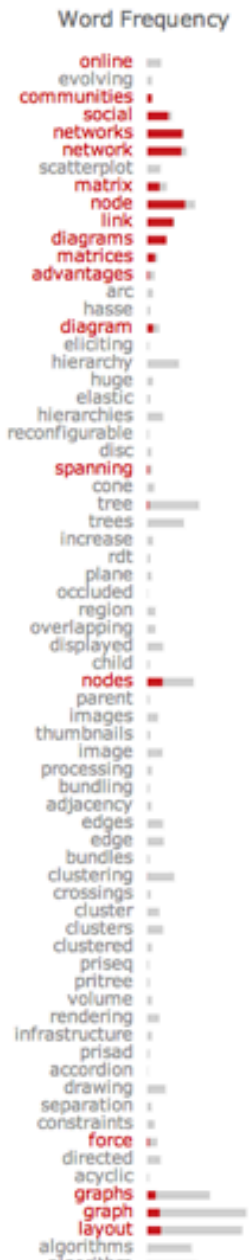


[Hall et al. 2008]

How might we better support **human-in-the-loop verification** of topic models?

Termite | Topic Model Visualization





Representative Docu

A Comparison of the Readability of Graphs Using Node-Link and Ma
Mohammad Ghoniem Jean-Daniel Fekete Philippe Castagliola

Using Multilevel Call Matrices in Large Software Projects
Frank van Ham

Improving the Readability of Clustered Social Networks using Node
Nathalie Henry Anastasia Bezerianos Jean-Daniel Fekete

MatrixExplorer: a Dual-Representation System to Explore Social Ne
Nathalie Henry Jean-Daniel Fekete

NodeTrix: a Hybrid Visualization of Social Networks
Nathalie Henry Jean-Daniel Fekete Michael J. McGuffin

The need to visualize large social networks is growing as hardware
and many new data sets become available. Unfortunately, the visu
resolve the basic dilemma of being readable both for the global str
of local communities. To address this problem, we present NodeTri
combines the advantages of two traditional representations: node-l
of a network, while arbitrary portions of the network can be shown
analysis of communities. A key contribution is a set of interaction t
NodeTrix visualization by dragging selections to and from node-link
NodeTrix representation to explore the dataset and create meaning
Finally, we present a case study applying NodeTrix to the analysis
illustrate the capabilities of NodeTrix as both an exploration tool an

Visualizing Causal Semantics using Animations
Nivedita R. Kadaba Pourang P. Irani Jason Leboe

Balancing Systematic and Flexible Exploration of Social Netw
Adam Perer Ben Shneiderman

Social network analysis (SNA) has emerged as a powerful method
networks. However, interactive exploration of networks is currently
patterns and comprehend the structure of networks with many node
medley of statistical methods and overwhelming visual output whic
explore in an orderly manner. This results in exploration that is lar
to help structural analysts understand social networks more effecti
attribute ranking and coordinated views to help users systematicall
flexibly iterate through visualizations of measures to gain an overv
networks using link structure, find cohesive subgroups, and focus o
networks by viewing different link types separately, or find pattern
overview. For each operation, a stable node layout is maintained in
comparisons. SocialAction offers analysts a strategy beyond opport
techniques for exploring social networks.

Causality Visualization Using Animated Growing Polygons
Niklas Elmqvist Philippas Tsigas

SpicyNodes: Radial Layout Authoring for the General Public

Display **442 of 442** papers containing regular expression [ ] in the ☑ title, ☑ author, or ☑ abstract.

1 Graph Visualization

Graph, network, node-link diagram, layout, adjacency matrix, reordering

- Asymmetric Relations in Longitudinal Social Networks
- Multi-Level Graph Layout on the GPU
- Balancing Systematic and Flexible Exploration of Social Networ
- Parallel Edge Splatting for Scalable Dynamic Graph Visualizatic

2 Text Visualization

Text, topics, sentiment analysis

- The Shape of Shakespeare: Visualizing Text Using Implicit Surf
- ThemeRiver: visualizing theme changes over time
- From Metaphor to Method: Cartographic Perspectives on Infor
- Participatory Visualization with Wordle
- FacetAtlas: Multifaceted Visualization for Rich Text Corpora
- Mapping Text with Phrase Nets

3 Multidimensional visualization

Parallel coordinates, small multiples, splom, scatterplot matrix, multidimensional projections, embeddings, MDS, PCA

- Rolling the Dice: Multidimensional Visual Exploration using Sc
- Scattering Points in Parallel Coordinates
- Multidimensional Detective
- Improved Similarity Trees and their Application to Visual Data
- Steerable, Progressive Multidimensional Scaling

4 Tree Visualization

Treemap, node-link diagram, hierarchies

- SpaceTree: supporting exploration in large node link tree, desig
- Browsing Zoomable Treemaps: Structure-Aware Multi-Scale Na
- Interactive Visualization of Genealogical Graphs

5 Software Visualization

algorithm animation, traces, logs

- code_swarm: A Design Study in Organic Software Visualizatior
- The Visual Code Navigator: An Interactive Toolset for Source Cc
- Using Multilevel Call Matrices in Large Software Projects

6

Topic: Significant and coherent area of research

Exemplary terms: techniques, methods, systems, people...
Separate the terms by commas or semicolons

Exemplary documents: 3 or more papers
Drag and drop from InfoVis proceedings

2011 InfoVis Conference

Providence, Rhode Island

Theory and Foundations

Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization
Enrico BERTINI Andrada TATU Daniel KEIM

Benefitting InfoVis with Visual Difficulties
Jessica HULLMAN Eytan ADAR Priti SHAH

Product Plots
Hadley WICKHAM Heike HOFMANN

Visualization Rhetoric: Framing Effects in Narrative Visualization
Jessica HULLMAN Nick DIAKOPOULOS

Adaptive Privacy-Preserving Visualization Using Parallel Coordinates
Aritra DASGUPTA Robert KOSARA

Techniques

Context-Preserving Visual Links
Markus STEINBERGER Manuela WALDNER Marc STREIT Alexander LEX Dieter SCHMALSTIEG

Design Study of LineSets, a Novel Set Visualization Technique
Basak ALPER Nathalie RICHE Gonzalo RAMOS Mary CZERWINSKI

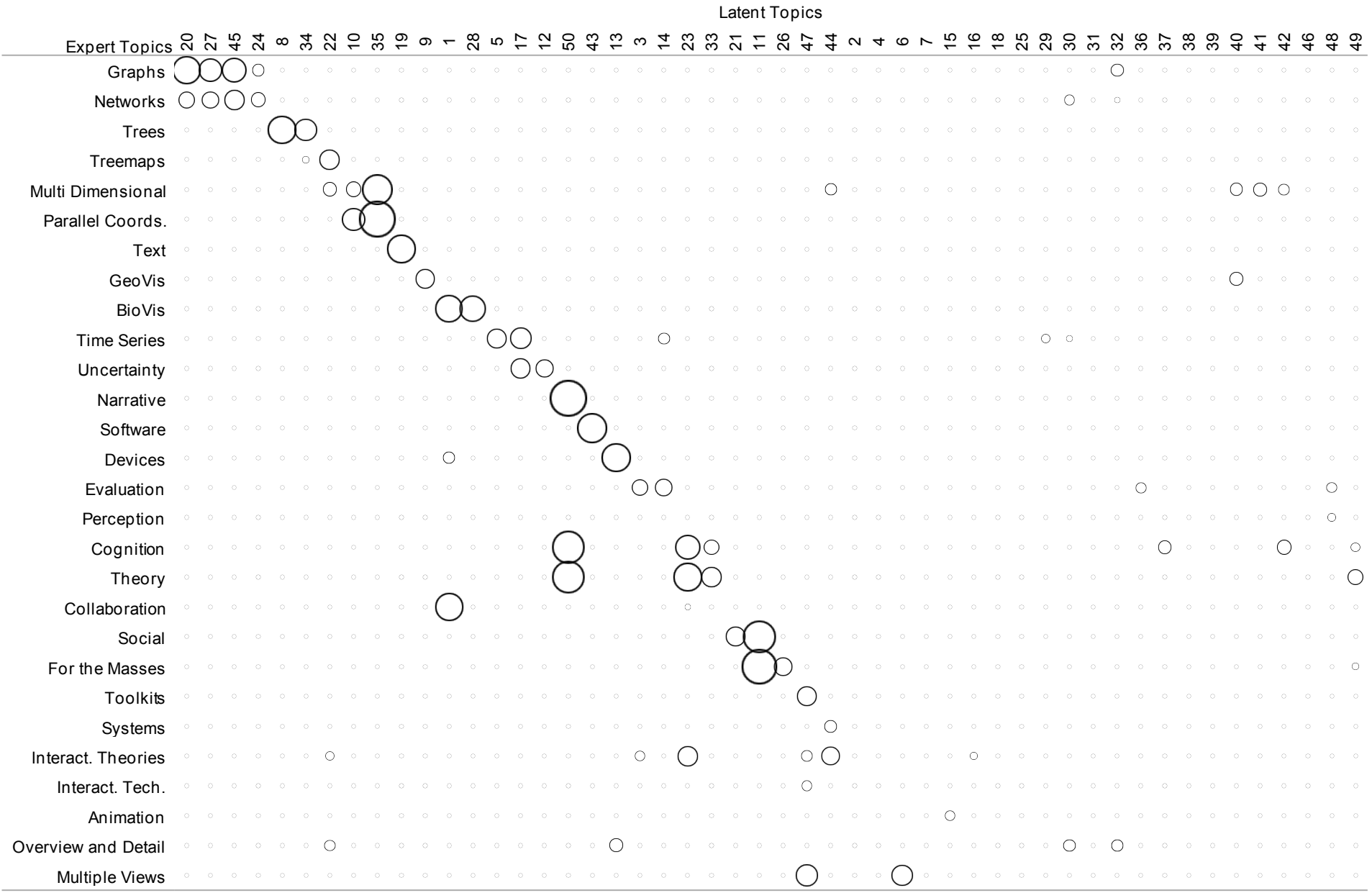
Developing and Evaluating Quilts for the Depiction of Large Layered Graphs
Juhee BAE Benjamin WATSON

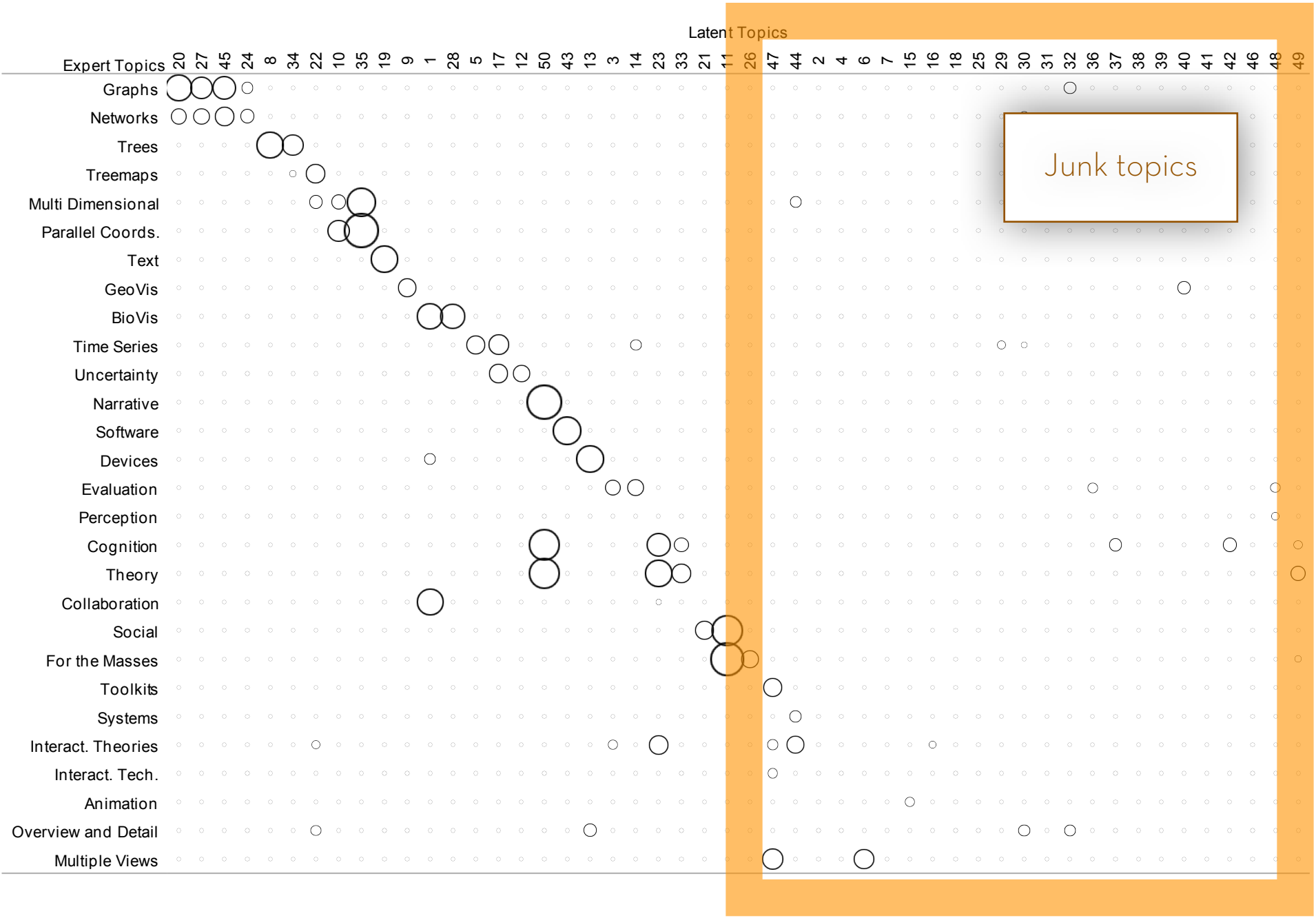
Arc Length-based Aspect Ratio Selection
Justin TALBOT John GERTH Pat HANRAHAN

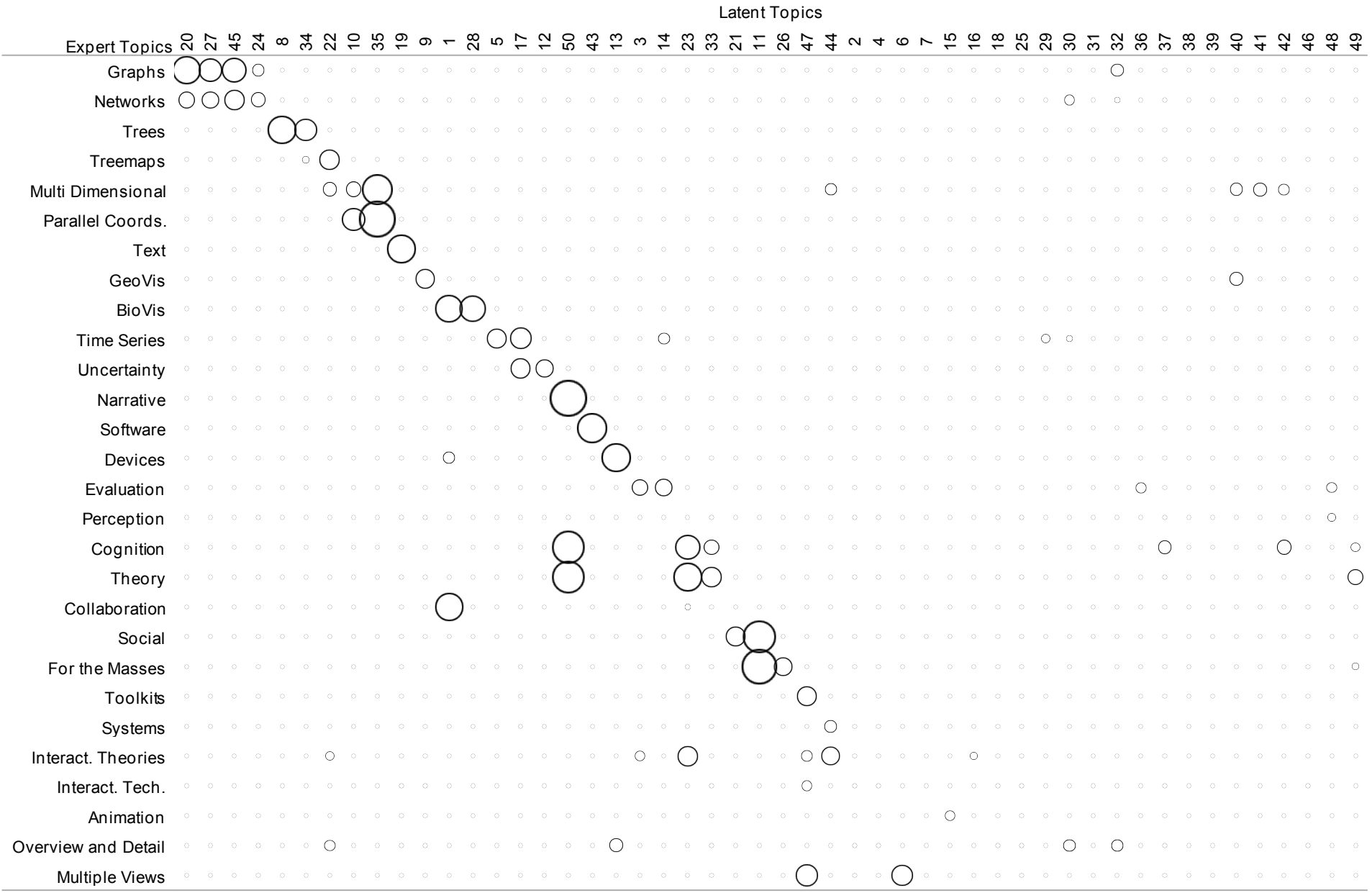
Asymmetric Relations in Longitudinal Social Networks
Ulrik BRANDES Bobo NICK

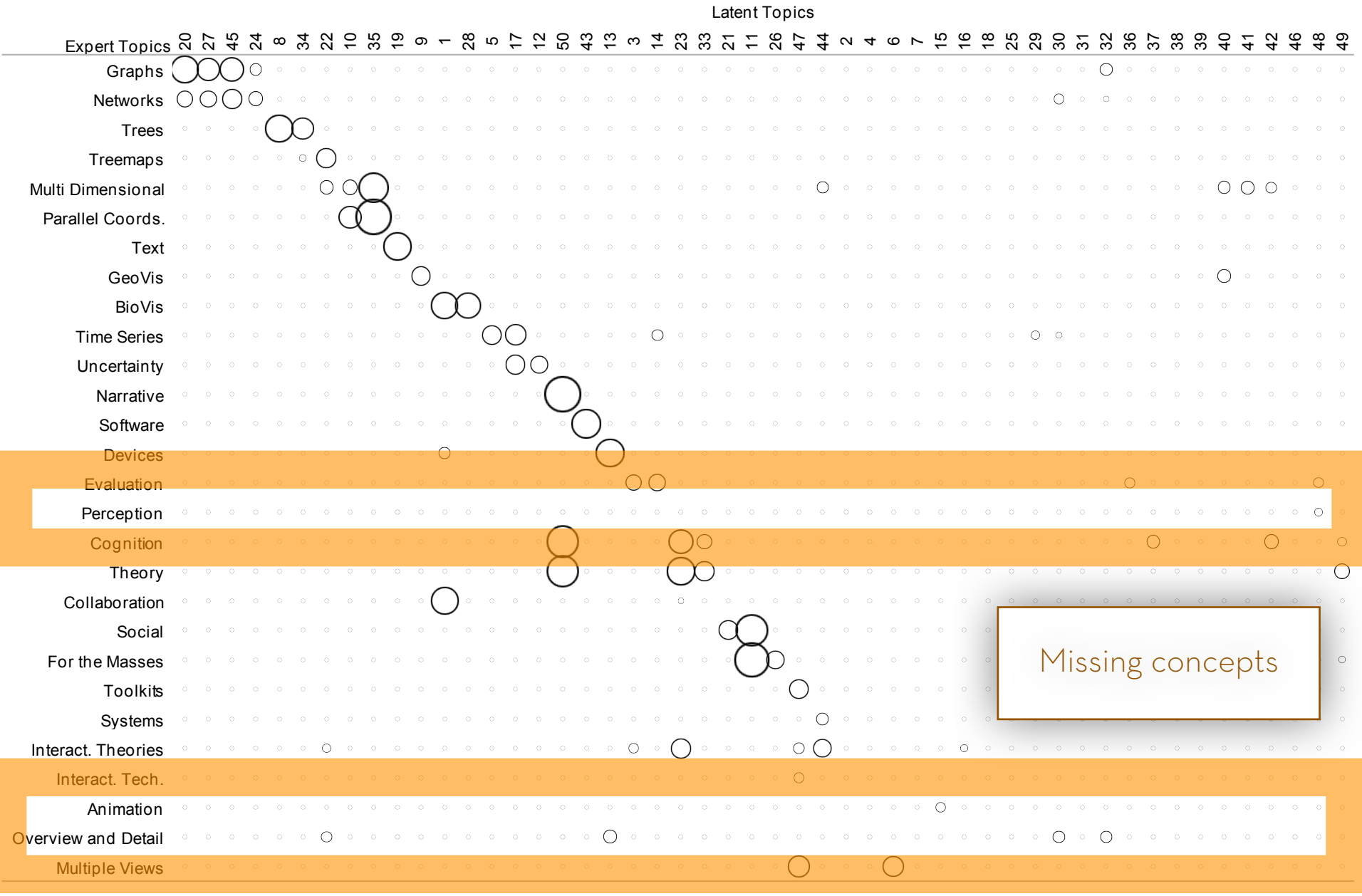
Systems and Frameworks

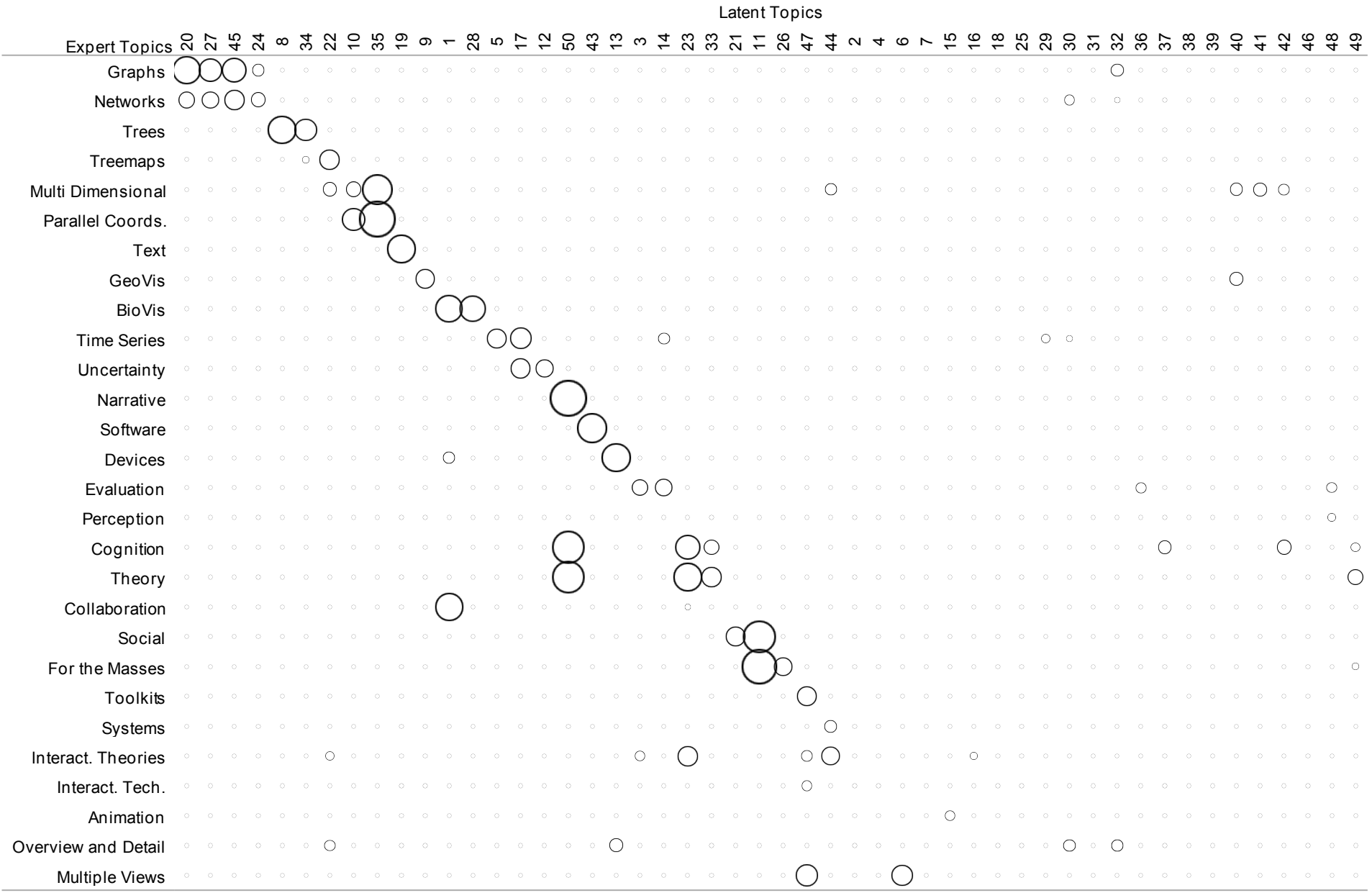
VisBricks: Multiform Visualization of Large, Inhomogeneous Data
Alexander LEX Hans-Joerg SCHULZ Marc STREIT Christian PARTL Dieter SCHMALSTIEG

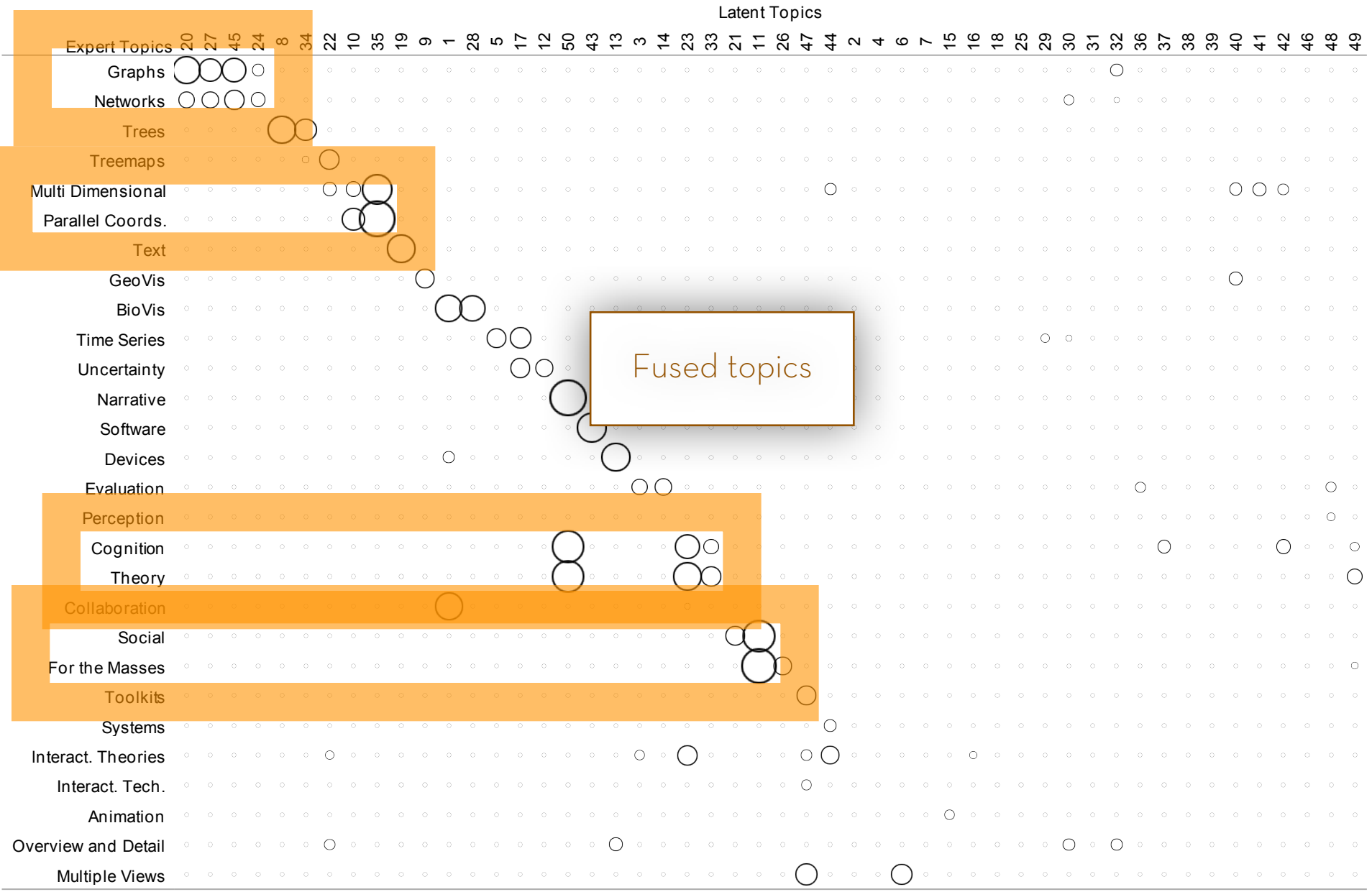

Latent Topics
Expert Topics
20 27 45 24 8 34 22 10 35 19 9 1 28 5 17 12 50 43 13 3 14 23 33 21 11 26 47 44 2 4 6 7 15 16 18 25 29 30 31 32 36 37 38 39 40 41 42 46 48 49
Graphs
Networks
Trees
Treemaps
Multi Dimensional
Parallel Coords.
Text
GeoVis
BioVis
Time Series
Uncertainty
Narrative
Software
Devices
Evaluation
Perception
Cognition
Theory
Collaboration
Social
For the Masses
Toolkits
Systems
Interact. Theories
Interact. Tech.
Animation
Overview and Detail
Multiple Views

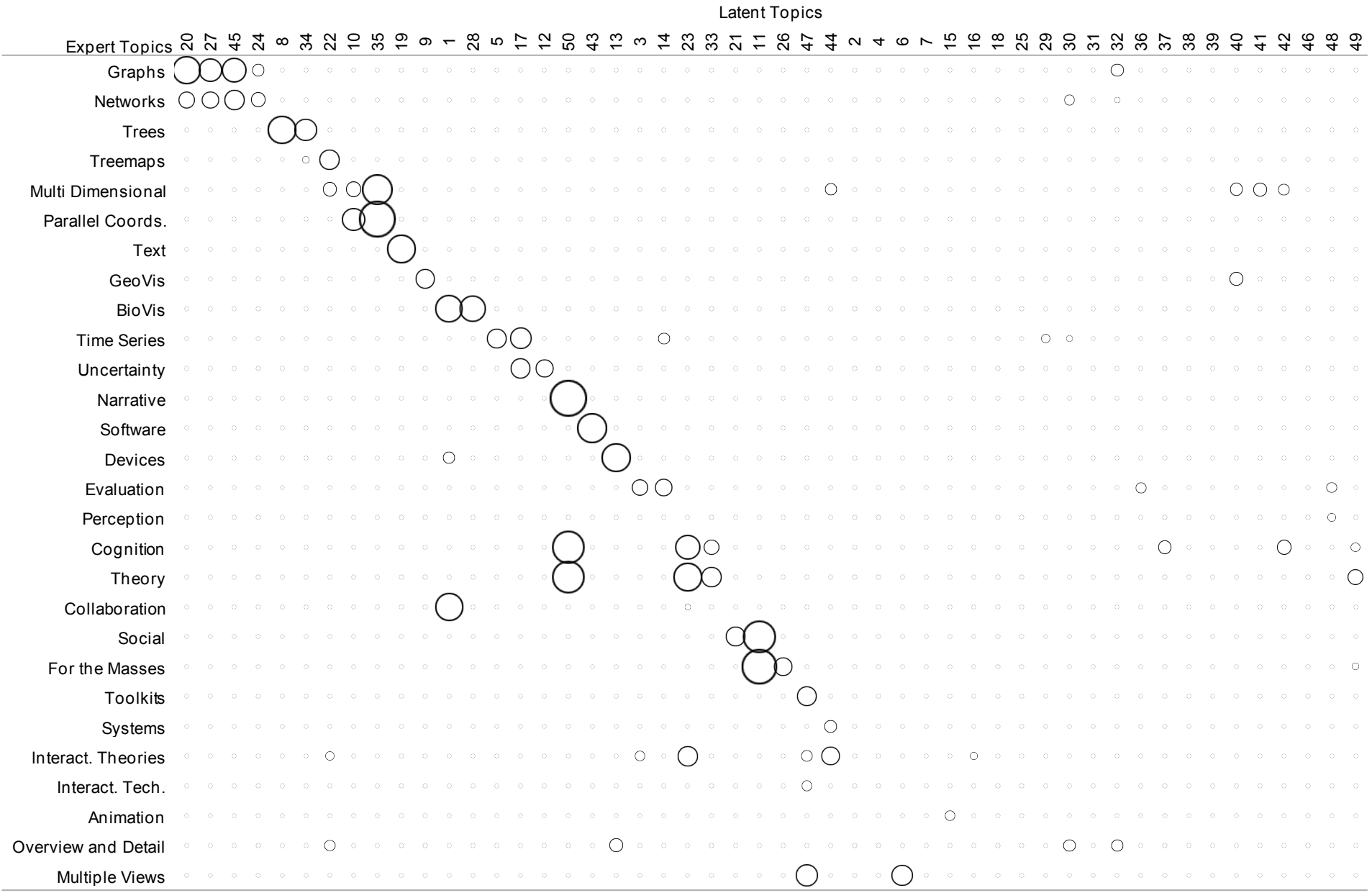

Latent Topics
Expert Topics
20 27 45 24 8 34 22 10 35 19 9 1 28 5 17 12 50 43 13 3 14 23 33 21 11 26 47 44 2 4 6 7 15 16 18 25 29 30 31 32 36 37 38 39 40 41 42 46 48 49
Graphs
Networks
Trees
Treemaps
Multi Dimensional
Parallel Coords.
Text
GeoVis
BioVis
Time Series
Uncertainty
Narrative
Software
Devices
Evaluation
Perception
Cognition
Theory
Collaboration
Social
For the Masses
Toolkits
Systems
Interact. Theories
Interact. Tech.
Animation
Overview and Detail
Multiple Views
Junk topics

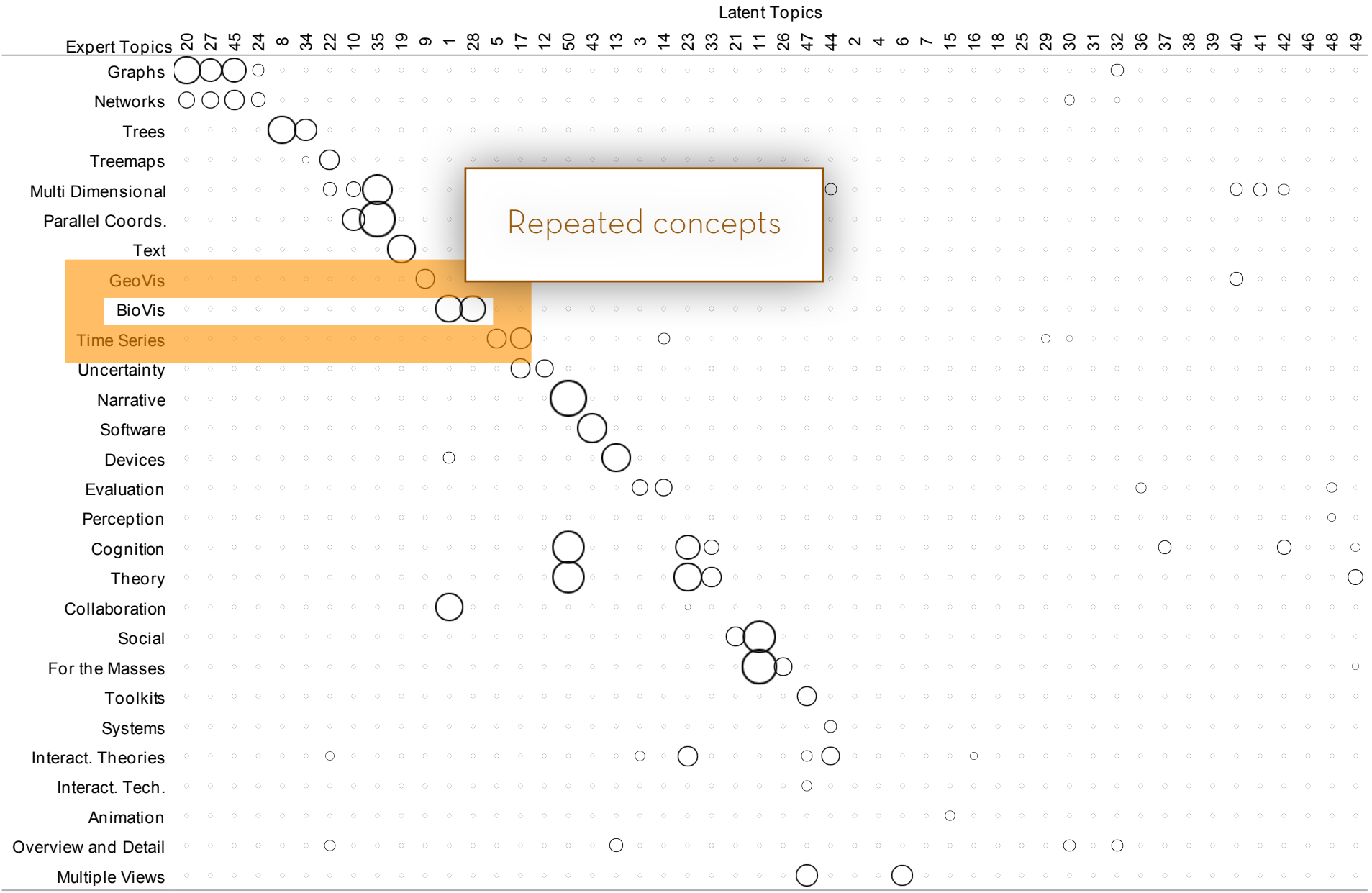

Latent Topics
Expert Topics
20 27 45 24 8 34 22 10 35 19 9 1 28 5 17 12 50 43 13 3 14 23 33 21 11 26 47 44 2 4 6 7 15 16 18 25 29 30 31 32 36 37 38 39 40 41 42 46 48 49
Graphs
Networks
Trees
Treemaps
Multi Dimensional
Parallel Coords.
Text
GeoVis
BioVis
Time Series
Uncertainty
Narrative
Software
Devices
Evaluation
Perception
Cognition
Theory
Collaboration
Social
For the Masses
Toolkits
Systems
Interact. Theories
Interact. Tech.
Animation
Overview and Detail
Multiple Views

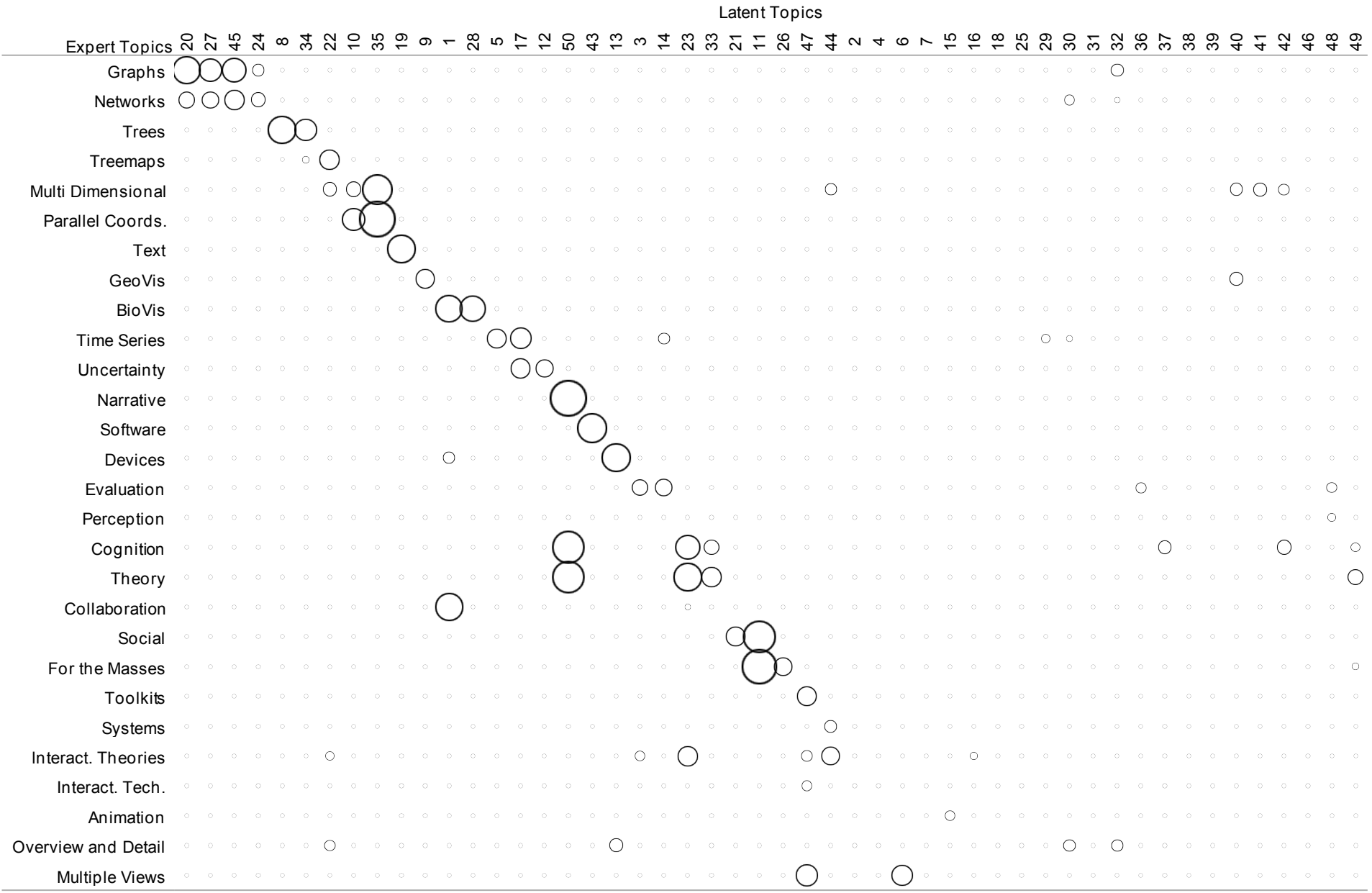

Latent Topics
Expert Topics
20 27 45 24 8 34 22 10 35 19 9 1 28 5 17 12 50 43 13 3 14 23 33 21 11 26 47 44 2 4 6 7 15 16 18 25 29 30 31 32 36 37 38 39 40 41 42 46 48 49
Graphs
Networks
Trees
Treemaps
Multi Dimensional
Parallel Coords.
Text
GeoVis
BioVis
Time Series
Uncertainty
Narrative
Software
Devices
Evaluation
Perception
Cognition
Theory
Collaboration
Social
For the Masses
Toolkits
Systems
Interact. Theories
Interact. Tech.
Animation
Overview and Detail
Multiple Views
Missing concepts

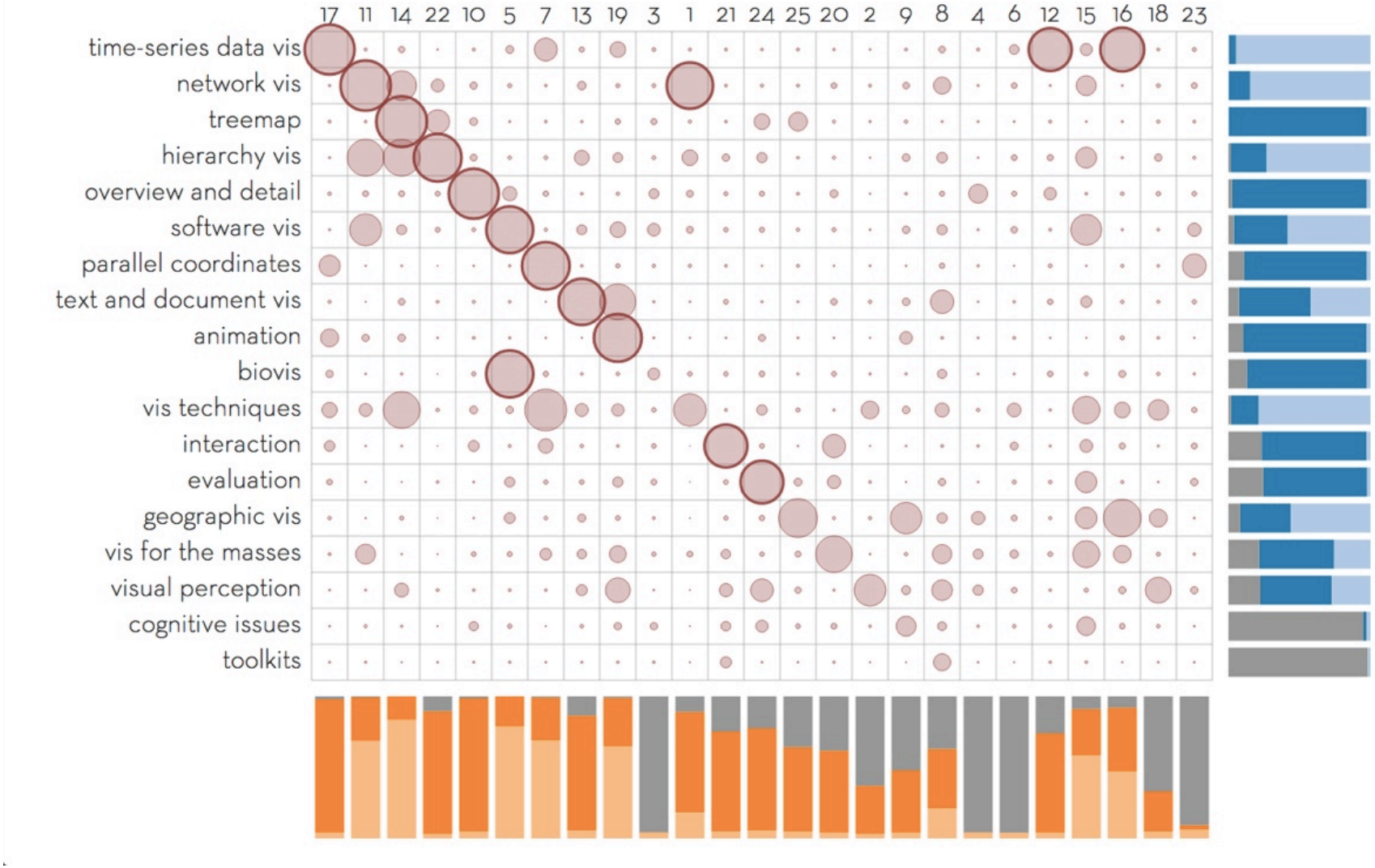

Latent Topics
Expert Topics
20 27 45 24 8 34 22 10 35 19 9 1 28 5 17 12 50 43 13 3 14 23 33 21 11 26 47 44 2 4 6 7 15 16 18 25 29 30 31 32 36 37 38 39 40 41 42 46 48 49
Graphs
Networks
Trees
Treemaps
Multi Dimensional
Parallel Coords.
Text
GeoVis
BioVis
Time Series
Uncertainty
Narrative
Software
Devices
Evaluation
Perception
Cognition
Theory
Collaboration
Social
For the Masses
Toolkits
Systems
Interact. Theories
Interact. Tech.
Animation
Overview and Detail
Multiple Views

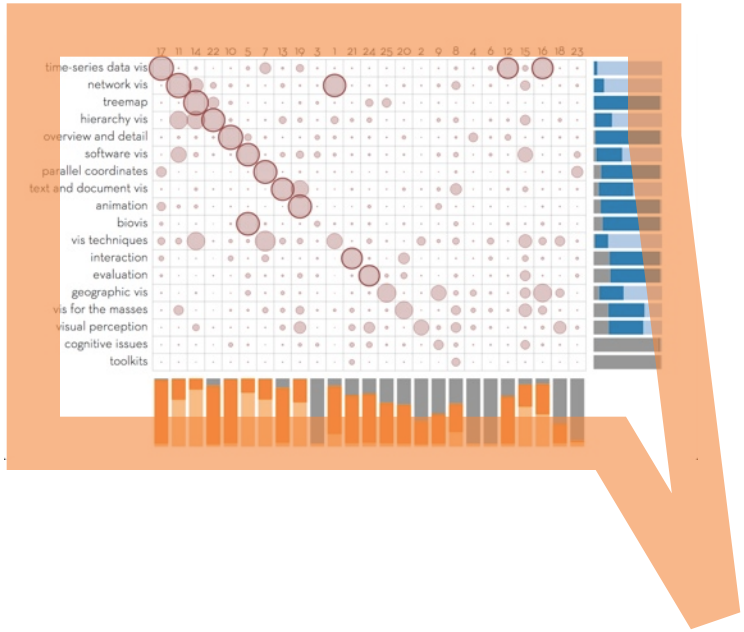

Latent Topics
Expert Topics
20 27 45 24 8 34 22 10 35 19 9 1 28 5 17 12 50 43 13 3 14 23 33 21 11 26 47 44 2 4 6 7 15 16 18 25 29 30 31 32 36 37 38 39 40 41 42 46 48 49
Graphs
Networks
Trees
Treemaps
Multi Dimensional
Parallel Coords.
Text
GeoVis
BioVis
Time Series
Uncertainty
Narrative
Software
Devices
Evaluation
Perception
Cognition
Theory
Collaboration
Social
For the Masses
Toolkits
Systems
Interact. Theories
Interact. Tech.
Animation
Overview and Detail
Multiple Views
Fused topics

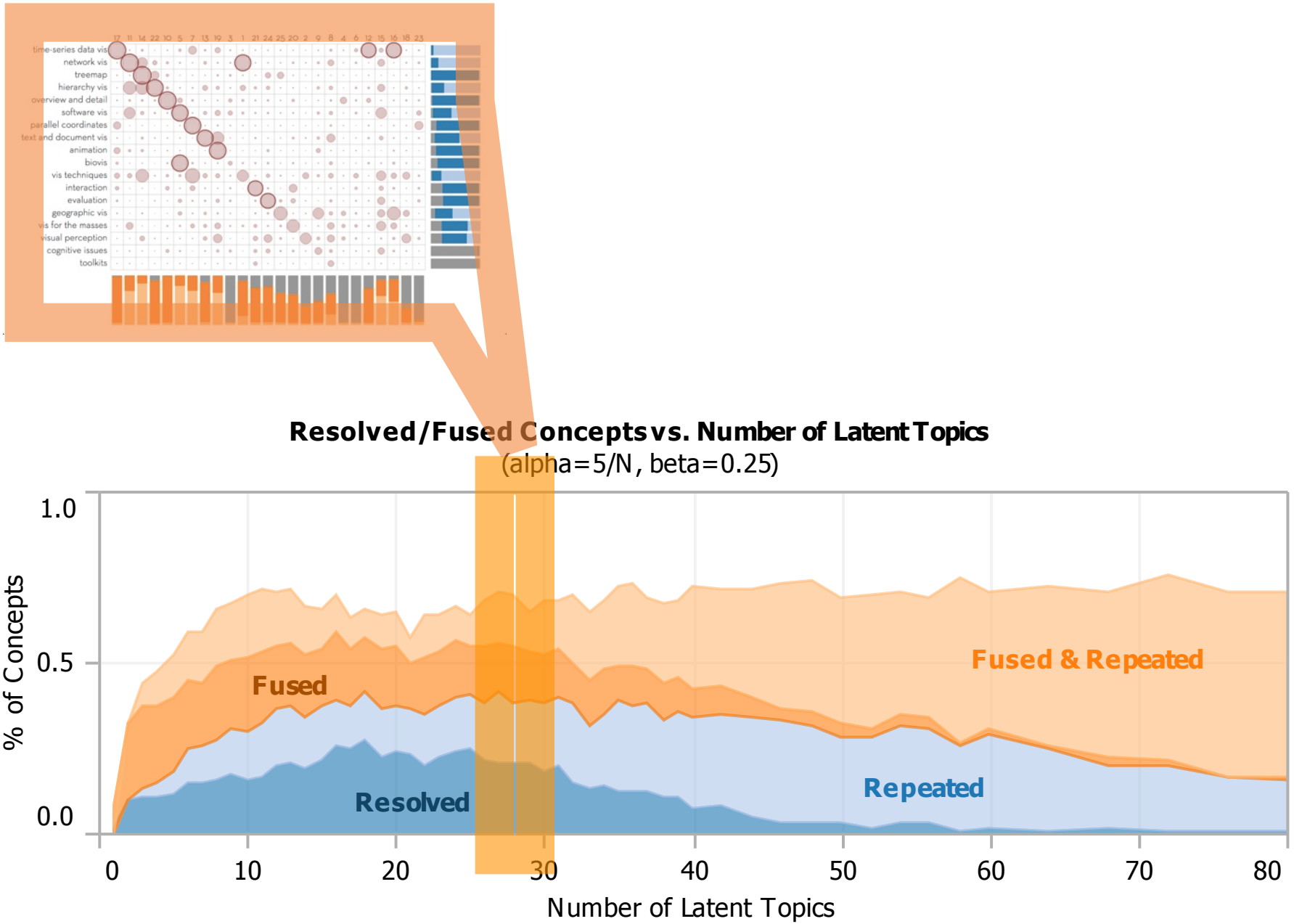

Latent Topics
Expert Topics
20 27 45 24 8 34 22 10 35 19 9 1 28 5 17 12 50 43 13 3 14 23 33 21 11 26 47 44 2 4 6 7 15 16 18 25 29 30 31 32 36 37 38 39 40 41 42 46 48 49
Graphs
Networks
Trees
Treemaps
Multi Dimensional
Parallel Coords.
Text
GeoVis
BioVis
Time Series
Uncertainty
Narrative
Software
Devices
Evaluation
Perception
Cognition
Theory
Collaboration
Social
For the Masses
Toolkits
Systems
Interact. Theories
Interact. Tech.
Animation
Overview and Detail
Multiple Views

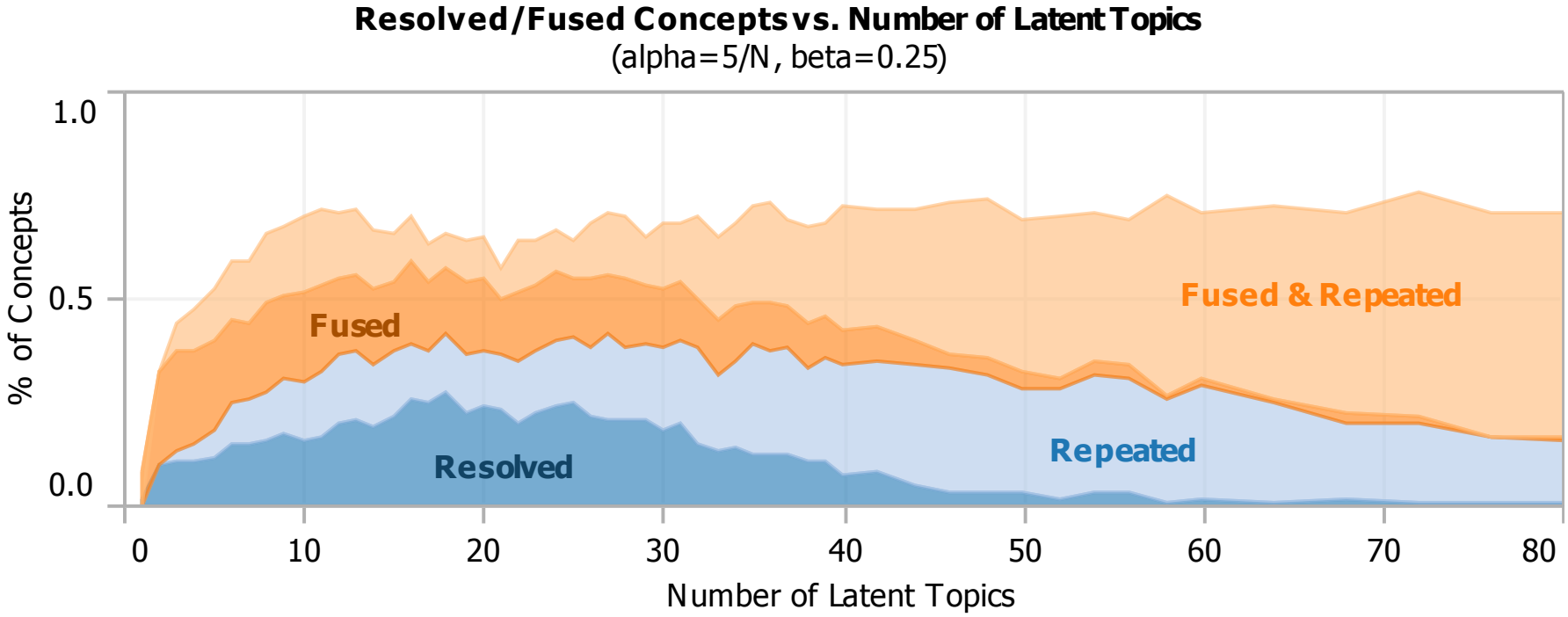

Latent Topics
Expert Topics
20 27 45 24 8 34 22 10 35 19 9 1 28 5 17 12 50 43 13 3 14 23 33 21 11 26 47 44 2 4 6 7 15 16 18 25 29 30 31 32 36 37 38 39 40 41 42 46 48 49
Graphs
Networks
Trees
Treemaps
Multi Dimensional
Parallel Coords.
Text
GeoVis
BioVis
Time Series
Uncertainty
Narrative
Software
Devices
Evaluation
Perception
Cognition
Theory
Collaboration
Social
For the Masses
Toolkits
Systems
Interact. Theories
Interact. Tech.
Animation
Overview and Detail
Multiple Views
Repeated concepts

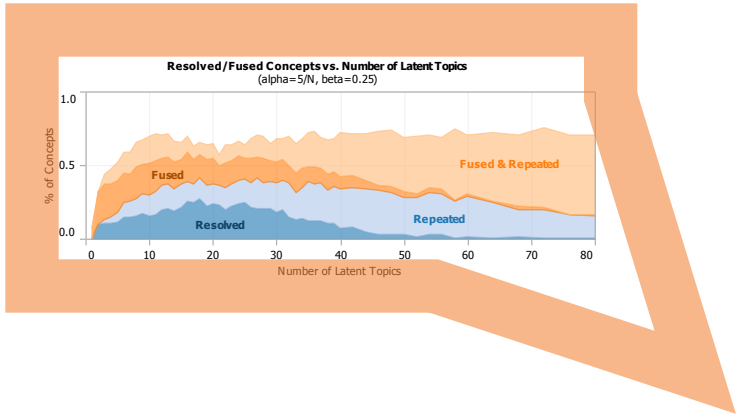

Latent Topics
Expert Topics
20 27 45 24 8 34 22 10 35 19 9 1 28 5 17 12 50 43 13 3 14 23 33 21 11 26 47 44 2 4 6 7 15 16 18 25 29 30 31 32 36 37 38 39 40 41 42 46 48 49
Graphs
Networks
Trees
Treemaps
Multi Dimensional
Parallel Coords.
Text
GeoVis
BioVis
Time Series
Uncertainty
Narrative
Software
Devices
Evaluation
Perception
Cognition
Theory
Collaboration
Social
For the Masses
Toolkits
Systems
Interact. Theories
Interact. Tech.
Animation
Overview and Detail
Multiple Views

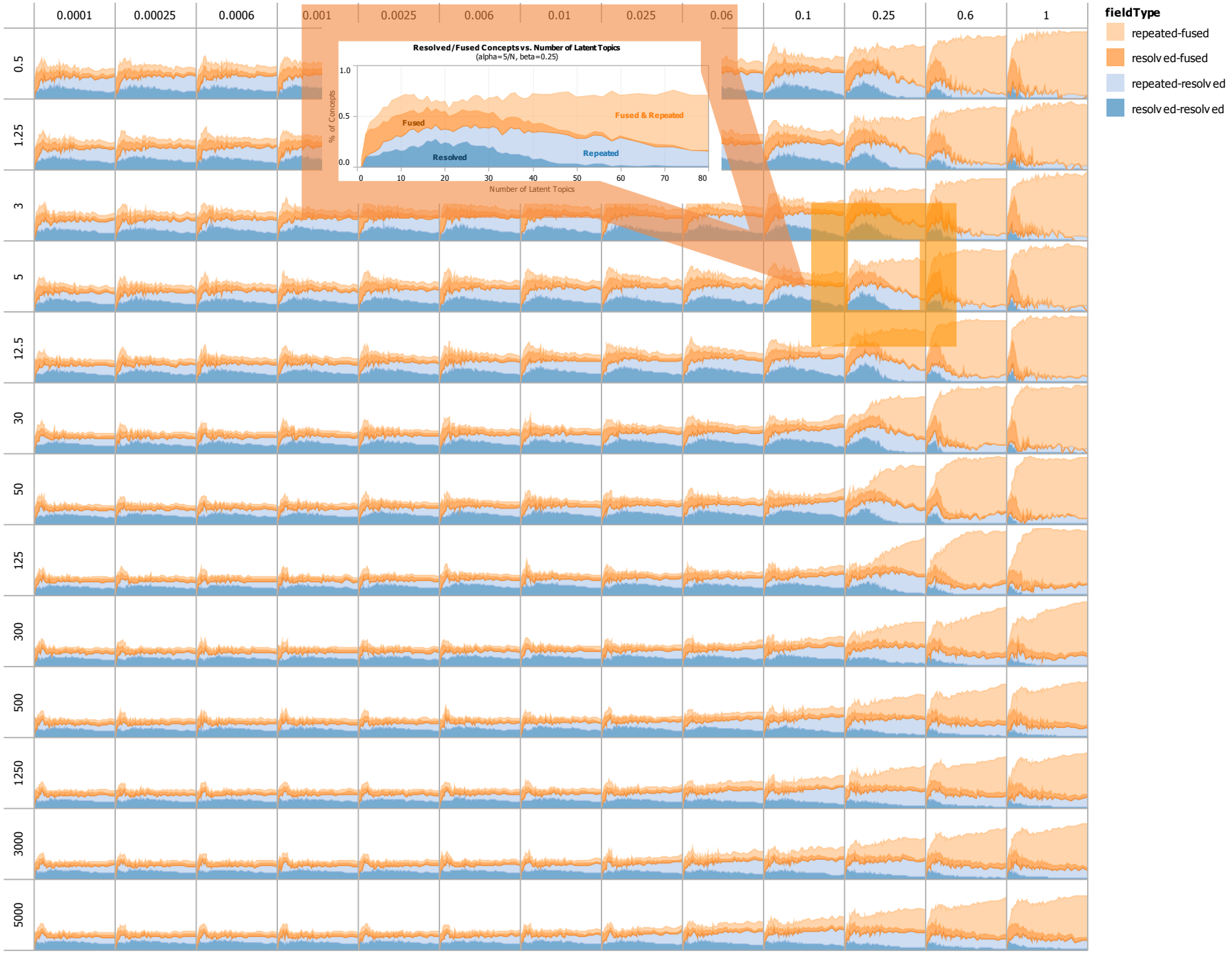

17 11 14 22 10 5 7 13 19 3 1 21 24 25 20 2 9 8 4 6 12 15 16 18 23
time-series data vis
network vis
treemap
hierarchy vis
overview and detail
software vis
parallel coordinates
text and document vis
animation
biovis
vis techniques
interaction
evaluation
geographic vis
vis for the masses
visual perception
cognitive issues
toolkits

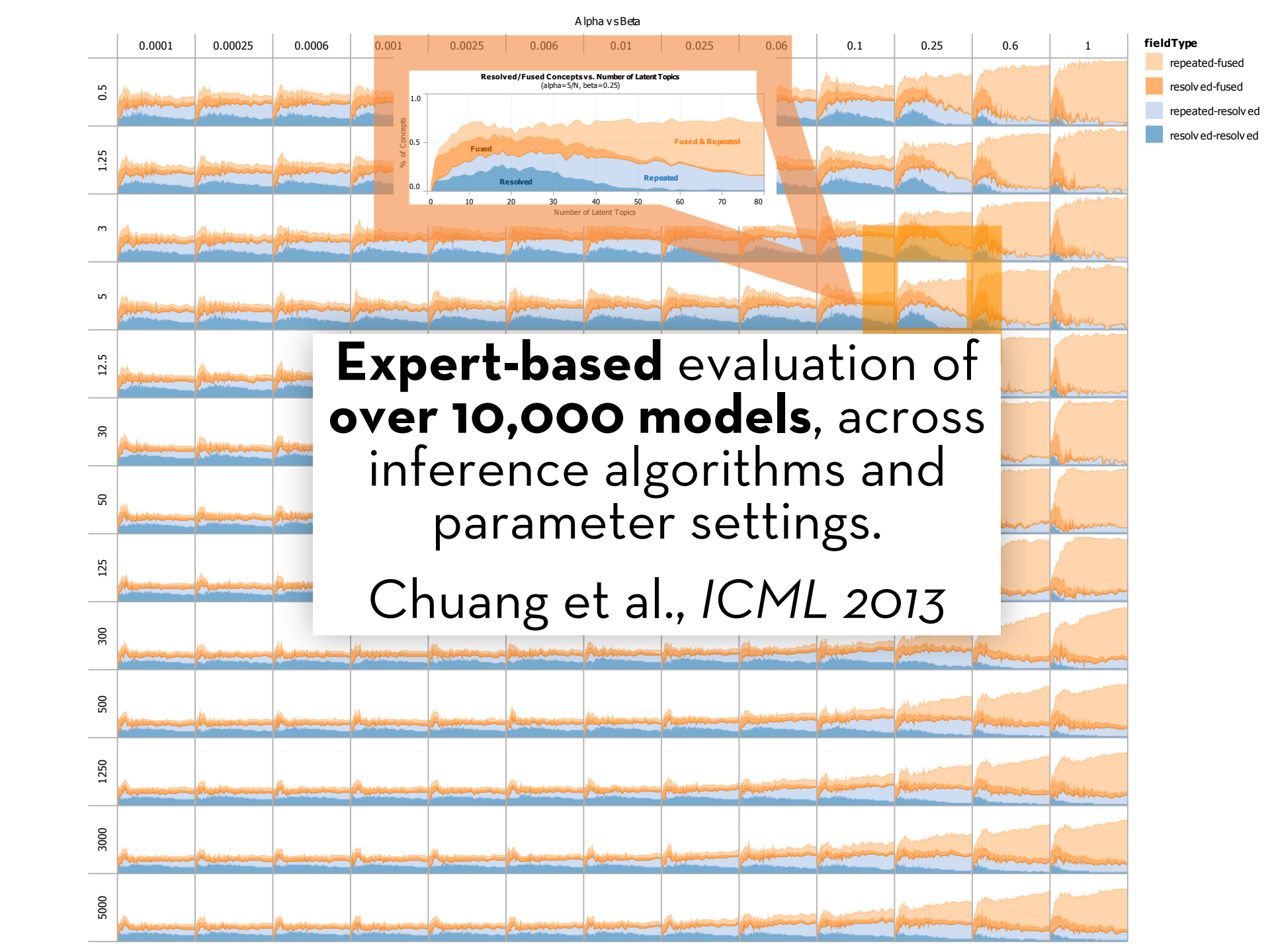

17 11 14 22 10 5 7 13 19 3 1 21 24 25 20 2 9 8 4 6 12 15 16 18 23
time-series data vis
network vis
treemap
hierarchy vis
overview and detail
software vis
parallel coordinates
text and document vis
animation
biovis
vis techniques
interaction
evaluation
geographic vis
vis for the masses
visual perception
cognitive issues
toolkits


time-series data vis
network vis
treemap
hierarchy vis
overview and detail
software vis
parallel coordinates
text and document vis
animation
biovis
vis techniques
interaction
evaluation
geographic vis
vis for the masses
visual perception
cognitive issues
toolkits
Resolved/Fused Concepts vs. Number of Latent Topics
(alpha=5/N, beta=0.25)
1.0
0.5
0.0
% of Concepts
Fused
Fused & Repeated
Resolved
Repeated
0
10
20
30
40
50
60
70
80
Number of Latent Topics


Resolved/Fused Concepts vs. Number of Latent Topics
(alpha=5/N, beta=0.25)
1.0
0.5
0.0
% of Concepts
Fused
Fused & Repeated
Repeated
Resolved
0
10
20
30
40
50
60
70
80
Number of Latent Topics


Resolved/Fused Concepts vs. Number of Latent Topics
(alpha=5/N, beta=0.25)
1.0
0.5
0.0
% of Concepts
Fused
Fused & Repeated
Resolved
Repeated
0
10
20
30
40
50
60
70
80
Number of Latent Topics


Alpha vs Beta
0.0001
0.00025
0.0006
0.001
0.0025
0.006
0.01
0.025
0.06
0.1
0.25
0.6
1
0.5
1.25
3
5
12.5
30
50
125
300
500
1250
3000
5000
Resolved/Fused Concepts vs. Number of Latent Topics
(alpha=5/N, beta=0.25)
% of Concepts
1.0
0.5
0.0
Fused
Fused & Repeated
Resolved
Repeated
0
10
20
30
40
50
60
70
80
Number of Latent Topics
fieldType
repeated-fused
resolved-fused
repeated-resolved
resolved-resolved


Alpha vs Beta
0.0001
0.00025
0.0006
0.001
0.0025
0.006
0.01
0.025
0.06
0.1
0.25
0.6
1
0.5
1.25
3
5
12.5
30
50
125
300
500
1250
3000
5000
fieldType
repeated-fused
resolved-fused
repeated-resolved
resolved-resolved
Resolved/Fused Concepts vs. Number of Latent Topics
(alpha=5/N, beta=0.25)
% of Concepts
1.0
0.5
0.0
Fused
Fused & Repeated
Resolved
Repeated
0
10
20
30
40
50
60
70
80
Number of Latent Topics
Expert-based evaluation of over 10,000 models, across inference algorithms and parameter settings.
Chuang et al., ICML 2013

STEPPING BACK

STEPPING BACK:
A POST-GRADUATE EXERCISE

Your chosen field's biggest blind spot?

STEPPING BACK:
A POST-GRADUATE EXERCISE

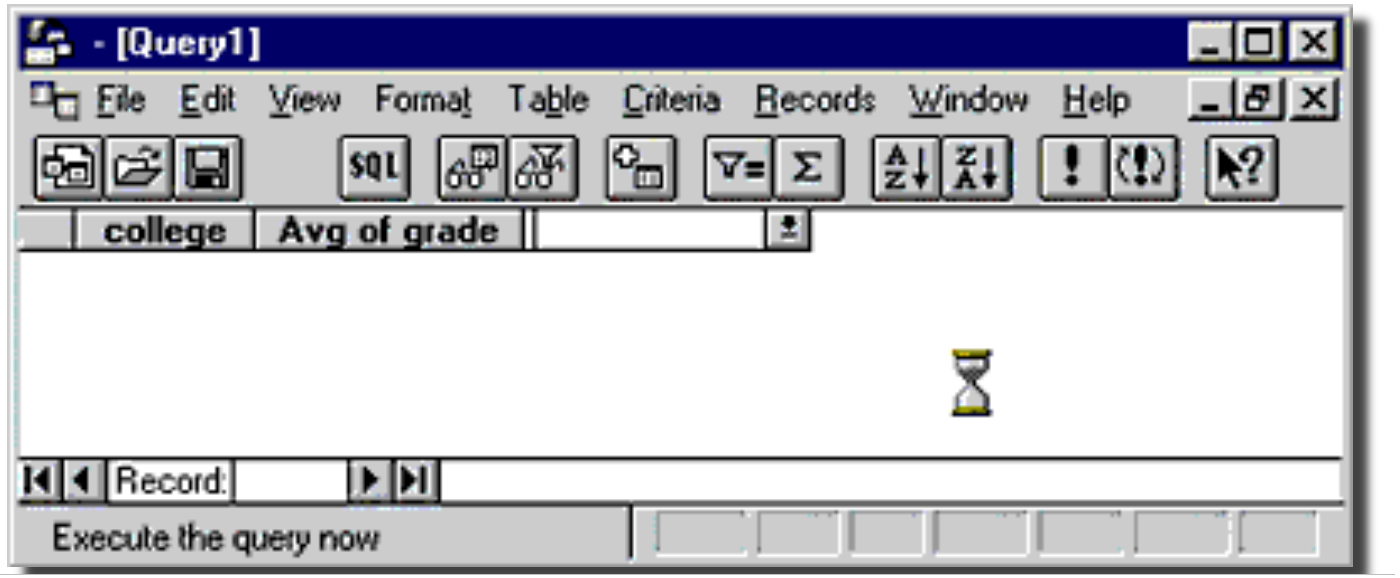
Your chosen field's biggest blind spot?

STEPPING BACK:
A POST-GRADUATE EXERCISE

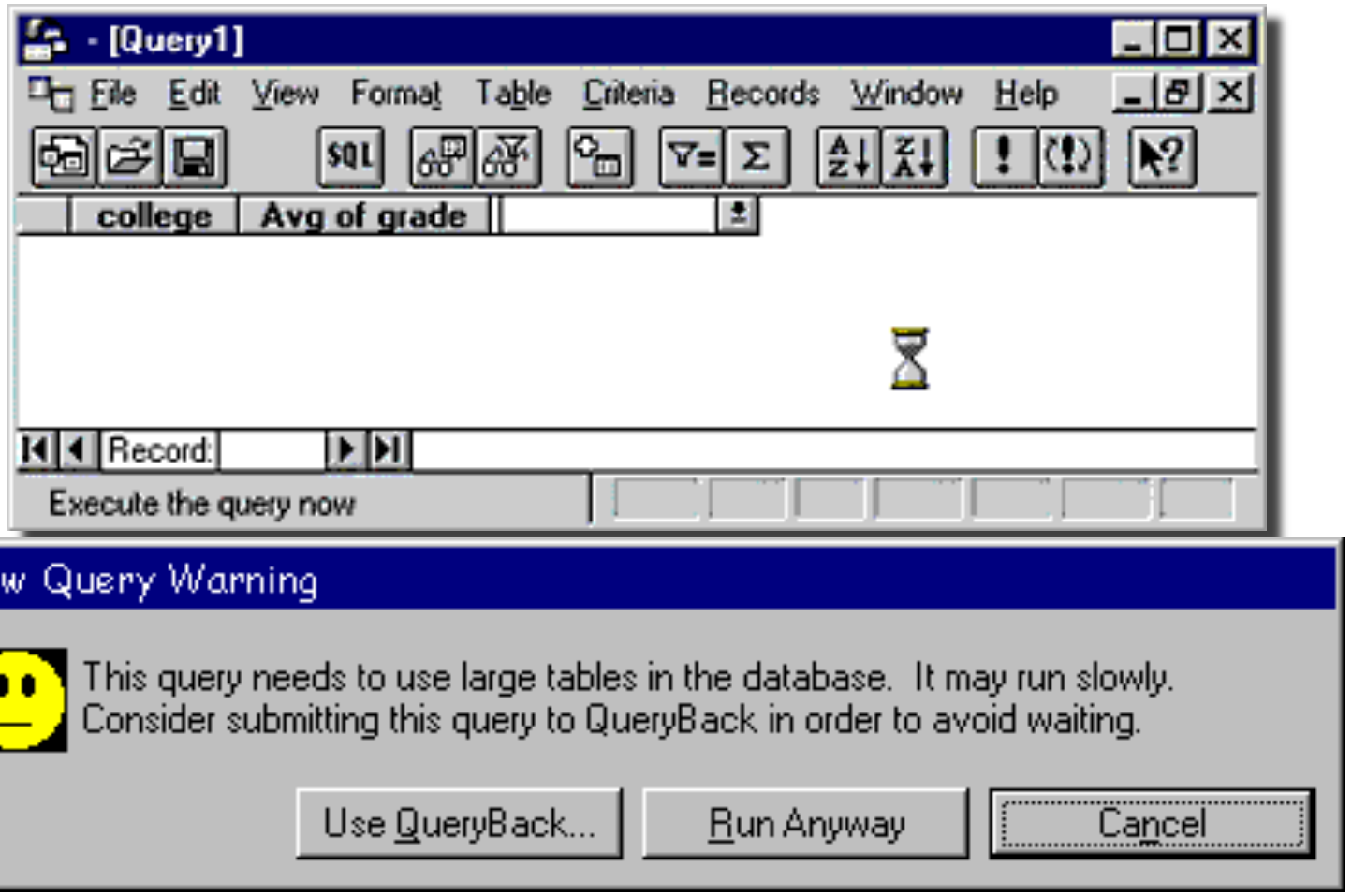
Your chosen field's biggest blind spot?

"A little disdain is not amiss;
a little scorn is alluring."
— William Congreve, 1670–1729

BATCH PROCESSING AND UI

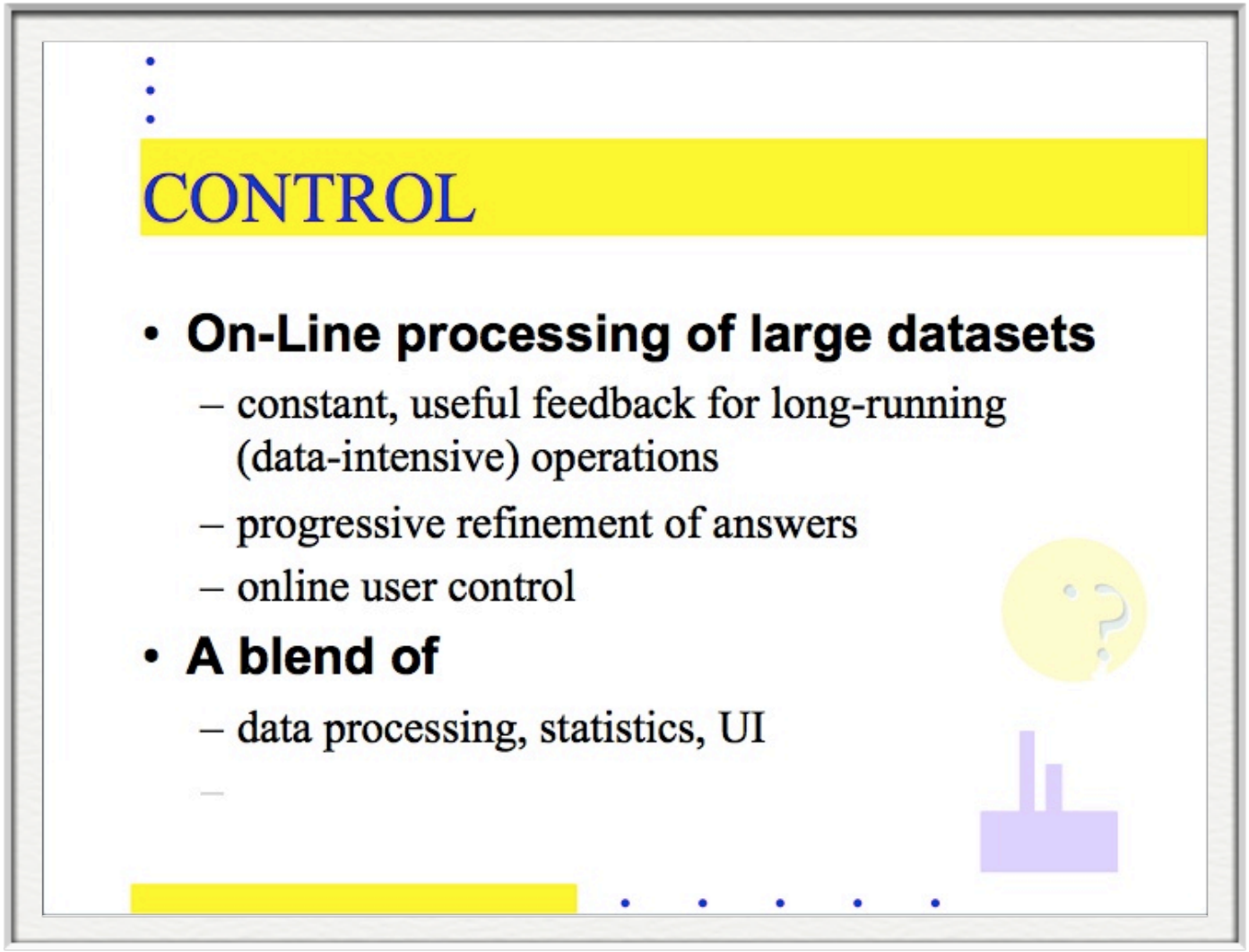


BATCH PROCESSING AND UI



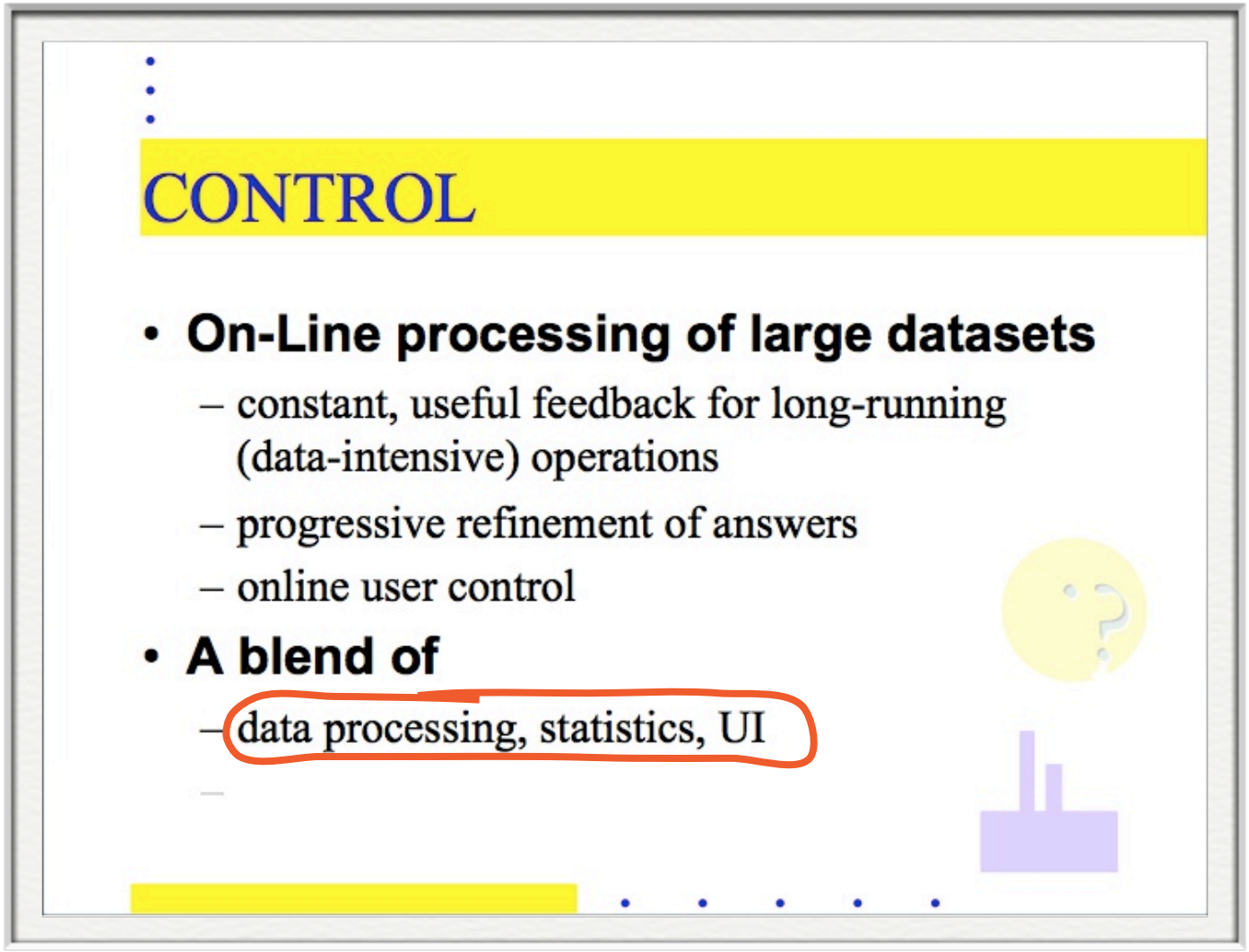
THE CONTROL PROJECT: 1996-2001

http://control.cs.berkeley.edu



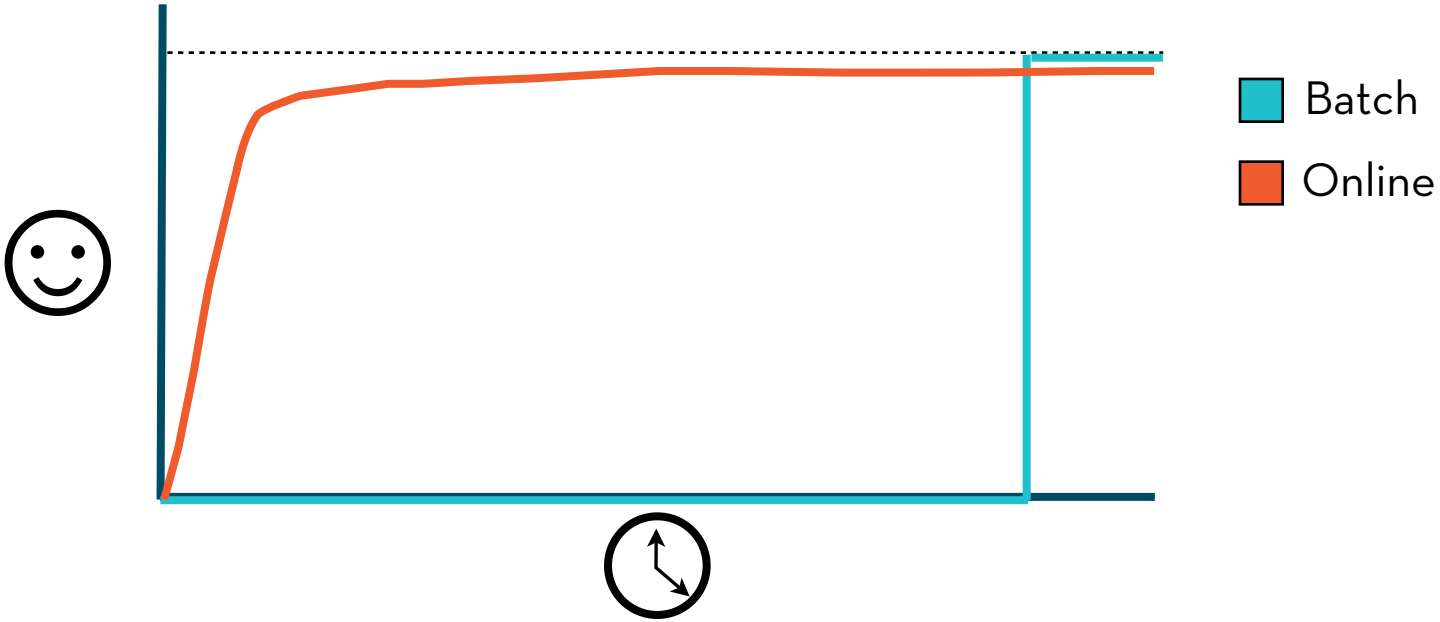
THE CONTROL PROJECT: 1996-2001

http://control.cs.berkeley.edu

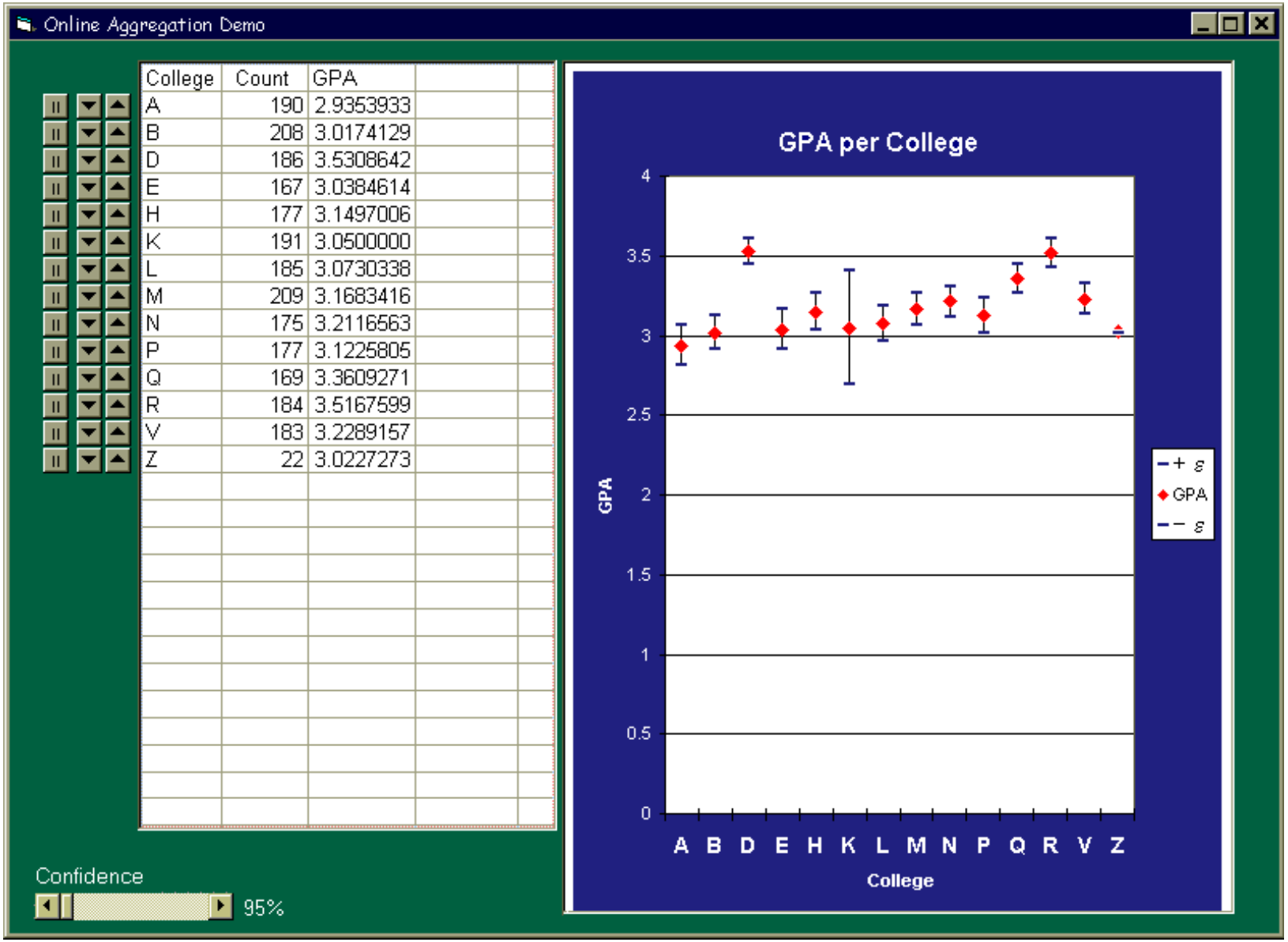


GOALS FOR ONLINE PROCESSING

- Maximize 1st derivative of the “mirth index”
- Mirth subject to dynamic redefinition
- Need FEEDBACK and CONTROL

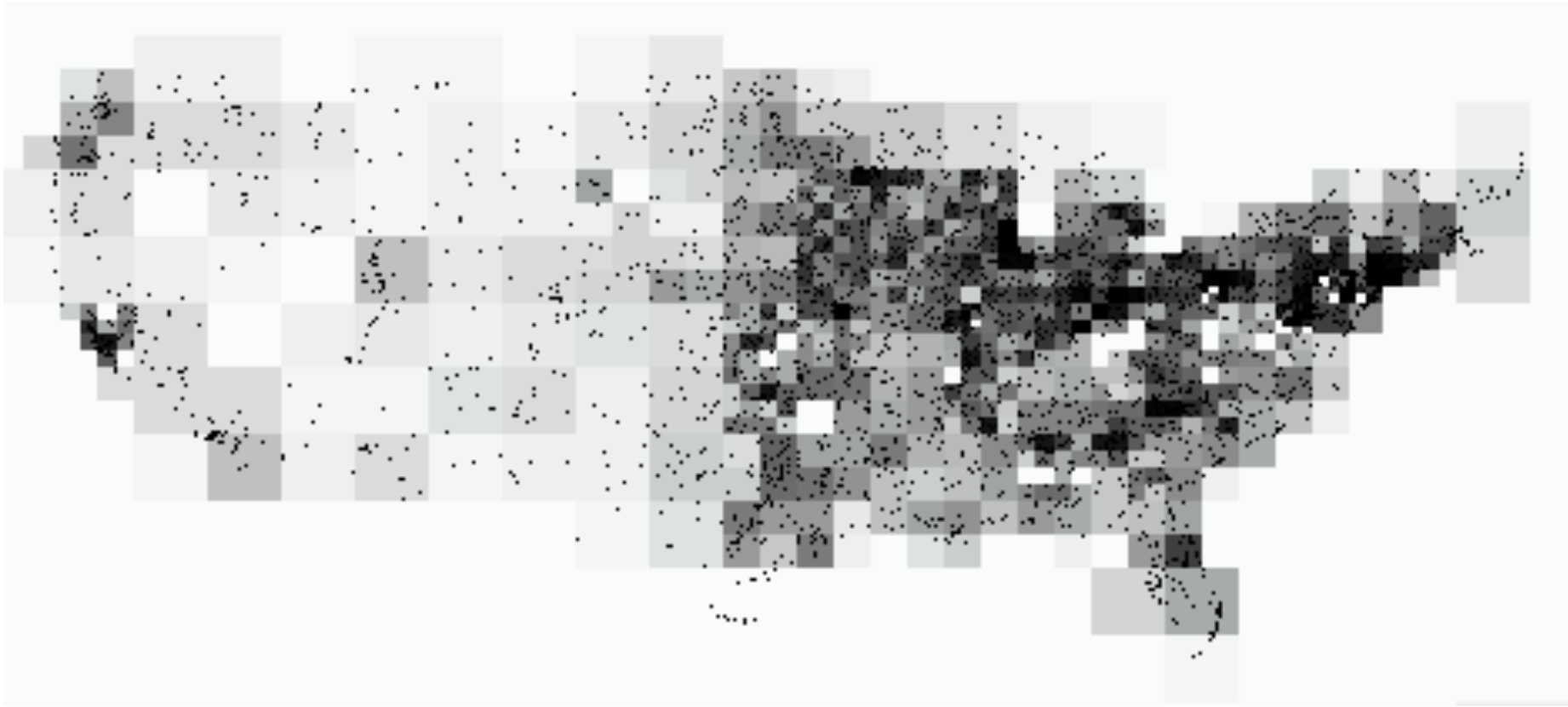


ONLINE AGGREGATION [SIGMOD '97, '99]

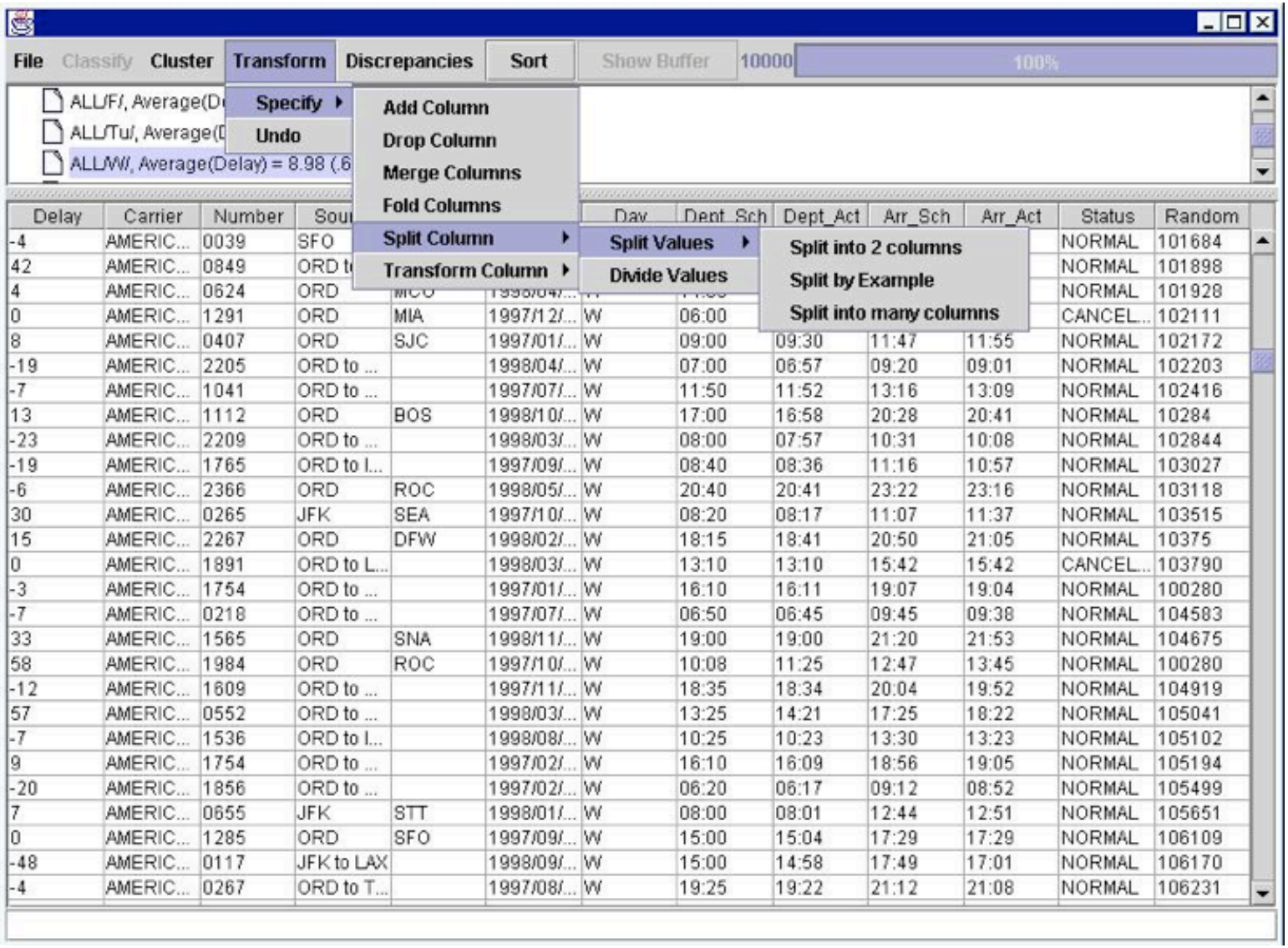


CLOUDS [IEEE COMPUTER 8/99]

CLOUDS [IEEE COMPUTER 8/99]



POTTER'S WHEEL [VLDB 01]



HADOOP WINTER

HADOOP WINTER AND A NEW SPRING

HADOOP WINTER AND A NEW SPRING

Scalable Approximate Query Processing With The DBO Engine

Christopher Jermaine, Subramanian Arumugam, Abhijit Pol, Alin Dobra
CISE Department, University of Florida
Gainesville, FL, USA
{cjermain, sa2, apol, adobra}@cise.ufl.edu

Incremental, Approximate Database Queries and Uncertainty for Exploratory Visualization

Danyel Fisher
Microsoft Research

MapReduce Online

Tyson Condie, Neil Conway, Peter Alvaro, Joseph M. Hellerstein
UC Berkeley

Khaled Elmeleegy, Russell Sears
Yahoo! Research

Online Aggregation for Large MapReduce Jobs

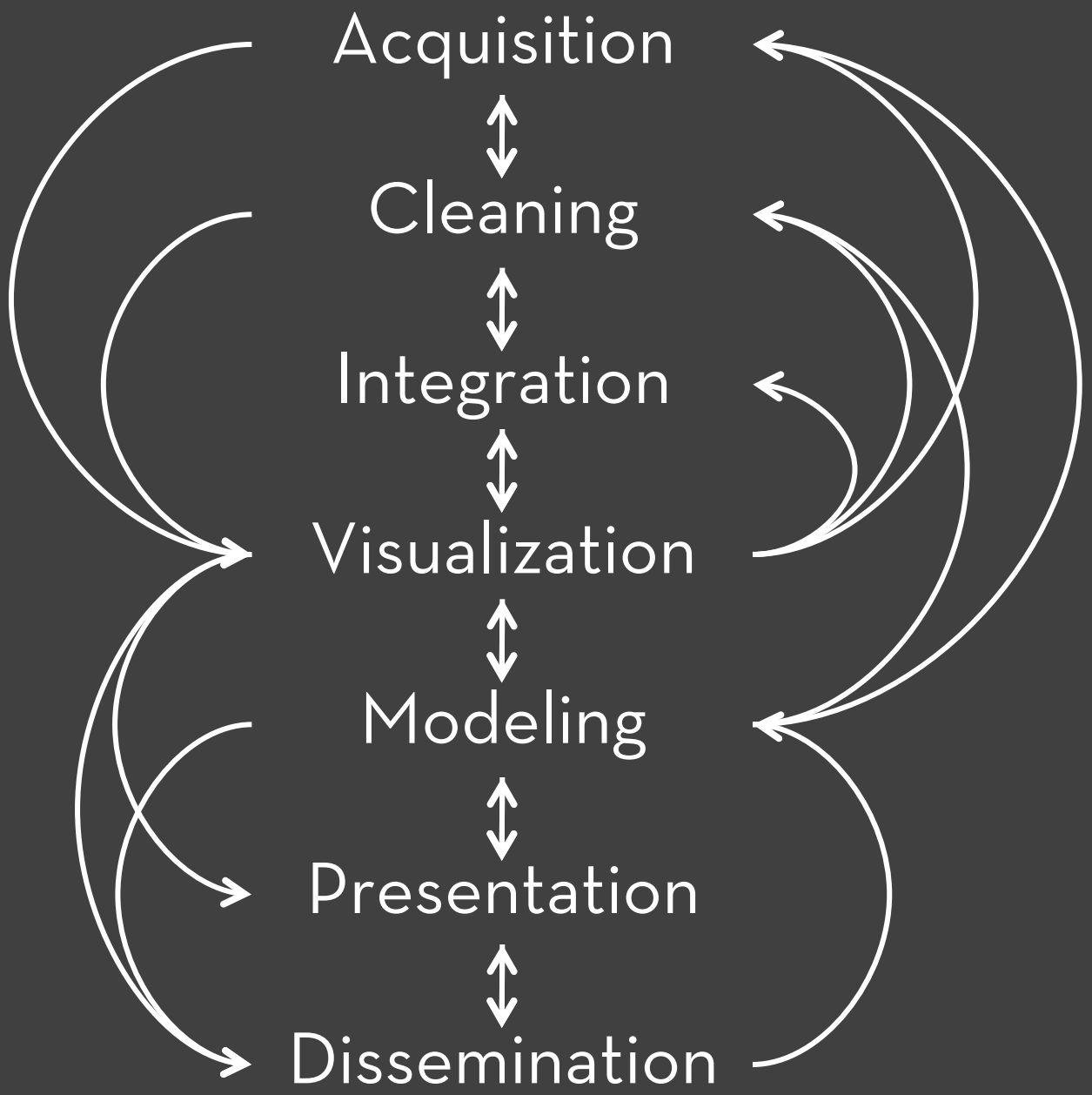
Niketan Pansare[1], Vinayak Borkar[2], Chris Jermaine[1], Tyson Condie[3]
[1]Rice University, [2]UC Irvine, [3]Yahoo! Research
np6@rice.edu, vborkar@ics.uci.edu, cmj4@rice.edu, tcondie@yahoo-inc.com

Stat! – An Interactive Analytics Environment for Big Data

Mike Barnett[1], Badrish Chandramouli[1], Robert DeLine[1], Steven Drucker[1], Danyel Fisher[1], Jonathan Goldstein[1], Patrick Morrison*[2], John Platt[1]

[1]Microsoft Research
Redmond, Washington, USA
{mbarnett, badrishc, rdeline, sdrucker, danyelf, jongold, jplatt}@microsoft.com

[2]North Carolina State University
Raleigh, North Carolina, USA
pjmorris@ncsu.edu


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination

ENGAGING PRACTITIONERS

To gain insight, we interviewed **35 analysts**:

ENGAGING PRACTITIONERS

To gain insight, we interviewed **35 analysts**:

25 Companies
Healthcare
Retail, Marketing
Social networking
Media
Finance, Insurance

Various titles
Data analyst
Data scientist
Software engineer
Consultant
Chief technical officer

ENGAGING PRACTITIONERS

To gain insight, we interviewed **35 analysts**:

25 Companies
Healthcare
Retail, Marketing
Social networking
Media
Finance, Insurance

Various titles
Data analyst
Data scientist
Software engineer
Consultant
Chief technical officer

Enterprise Data Analysis and Visualization: An Interview Study
Sean Kandel, Andreas Paepcke, Joseph Hellerstein, Jeffrey Heer
IEEE Visual Analytics Science & Technology (VAST), 2012

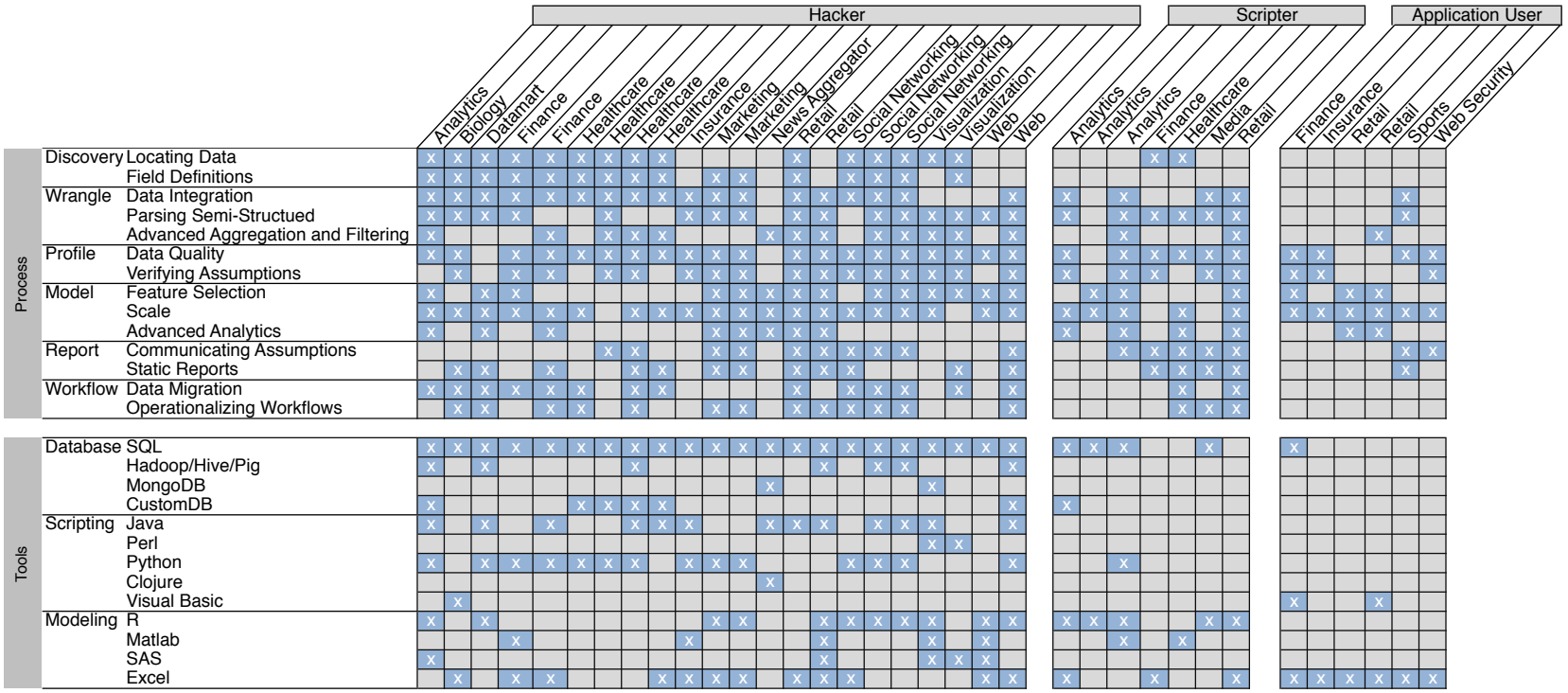


Fig. 1. Respondents, Challenges and Tools. The matrix displays interviewees (grouped by archetype and sector) and their corresponding challenges and tools. *Hackers* faced the most diverse set of challenges, corresponding to the diversity of their workflows and toolset. *Application users* and *scripters* typically relied on the IT team to perform certain tasks and therefore did not perceive them as challenges.

“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.”

“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.”

Lost productivity

“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.”

Lost productivity

“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.”

Lost productivity

“Most of the time once you transform the data … the insights can be scarily obvious.”

“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.”

Lost productivity

“Most of the time once you transform the data … the insights can be scarily obvious.”

Lost accessibility

“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.”

Lost productivity

“Most of the time once you transform the data … the insights can be scarily obvious.”

Lost accessibility

“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.”

Lost productivity

“Most of the time once you transform the data … the insights can be scarily obvious.”

Lost accessibility

“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.”

Lost productivity

“Most of the time once you transform the data … the insights can be scarily obvious.”

Lost accessibility

“It’s easy to just think you know what you are doing and not look at data at every intermediary step.

An analysis has 30 different steps. It’s tempting to just do this then that and then this. You have no idea in which ways you are wrong and what data is wrong.”

Interactivity and Visualization

TRIFACTA

PEOPLE + DATA + COMPUTATION

TRIFACTA

PEOPLE + DATA + COMPUTATION

Analytic Productivity

Remove drudgery, restore time

TRIFACTA

PEOPLE + DATA + COMPUTATION

Analytic Productivity

- Remove drudgery, restore time

Data Accessibility

- Enable self-service data manipulation

TRIFACTA

PEOPLE + DATA + COMPUTATION

CONTROL

- **On-Line processing of large datasets**
  - constant, useful feedback for long-running (data-intensive) operations
  - progressive refinement of answers
  - online user control
- **A blend of**
  - data processing, statistics, UI

TRIFACTA

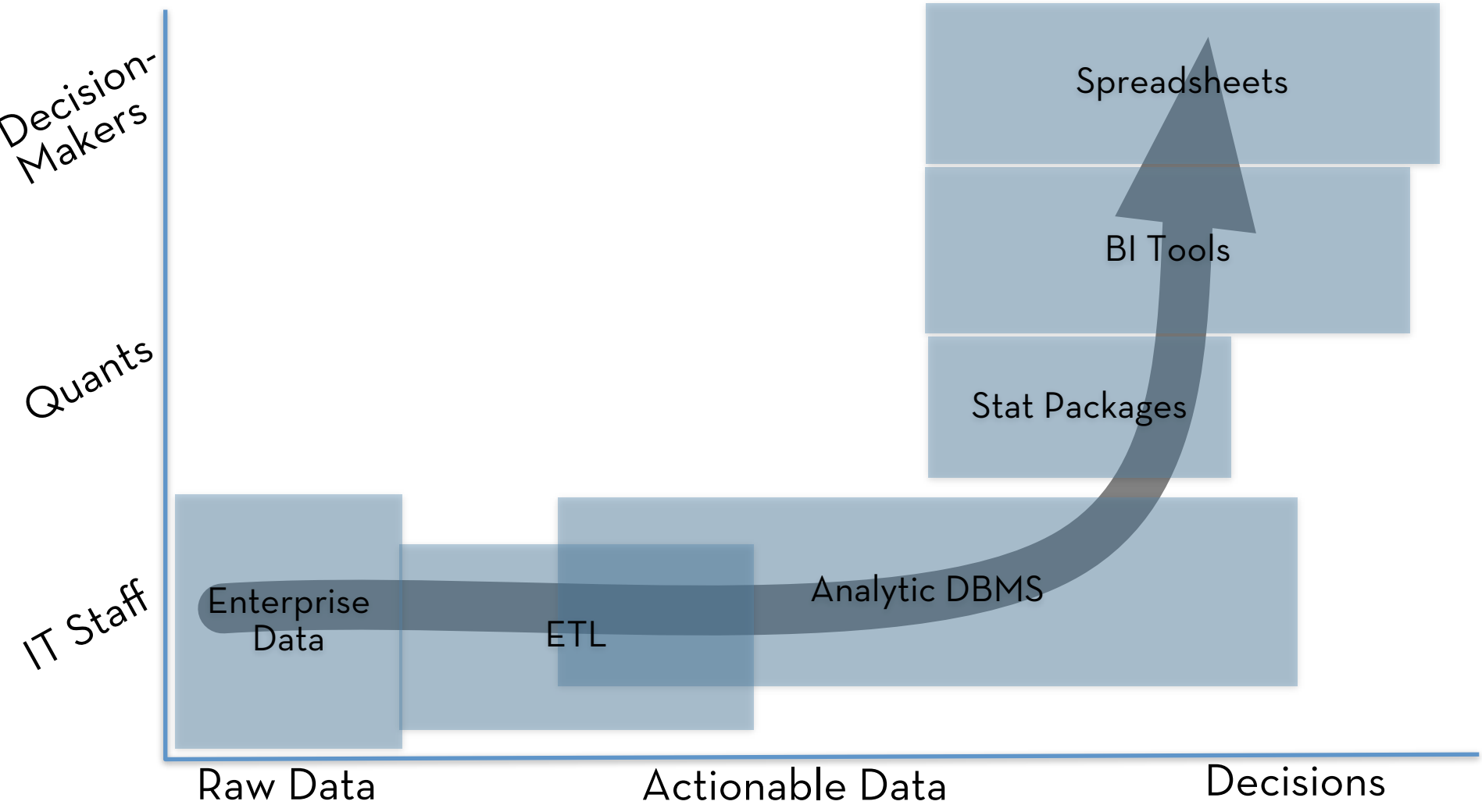
PEOPLE + DATA + COMPUTATION

CONTROL

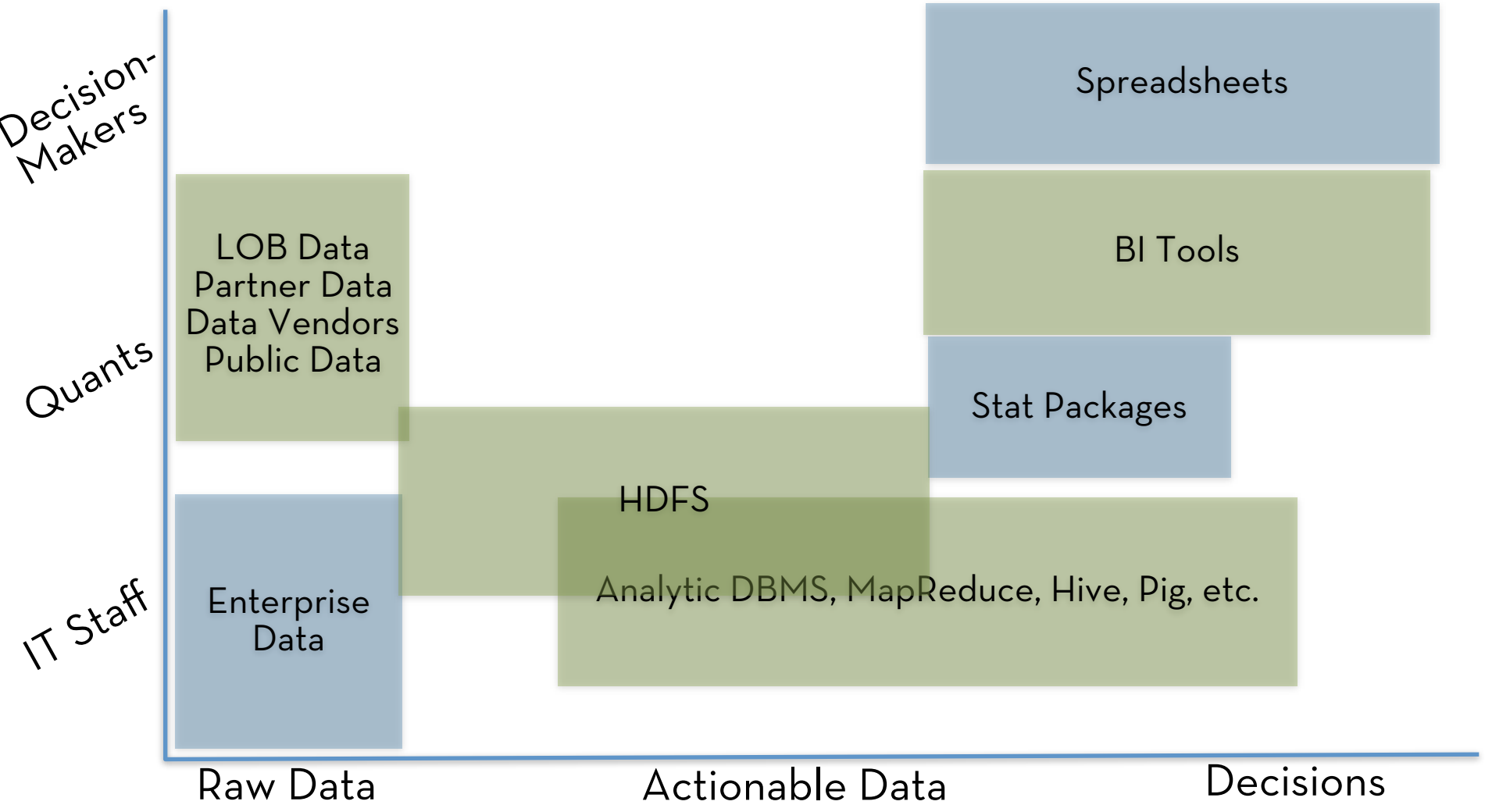
- **On-Line processing of large datasets**
  - constant, useful feedback for long-running (data-intensive) operations
  - progressive refinement of answers
  - online user control
- **A blend of**
  - data processing, statistics, UI

TRADITIONAL LANDSCAPE

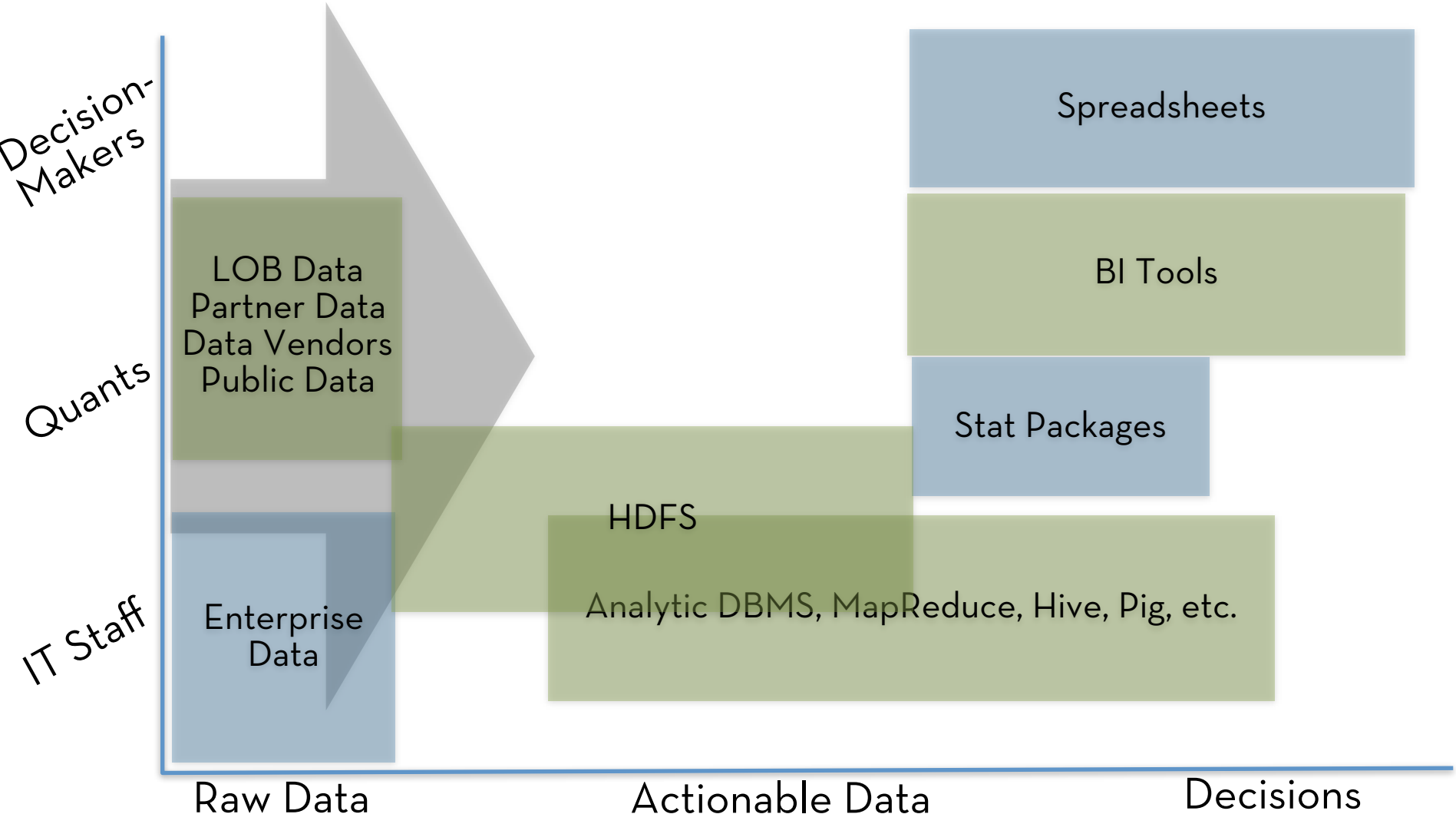
TRADITIONAL LANDSCAPE



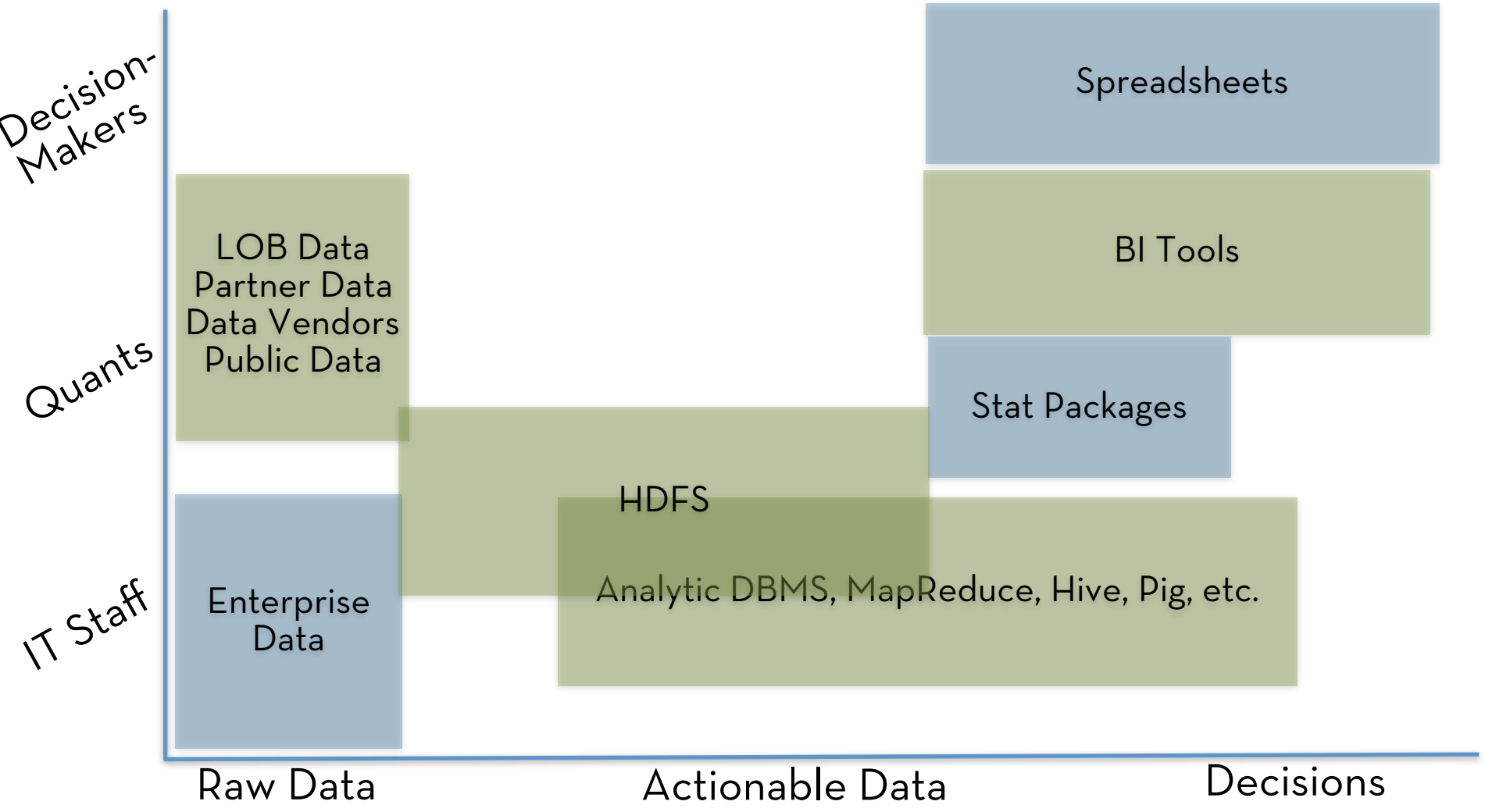
NEW LANDSCAPE



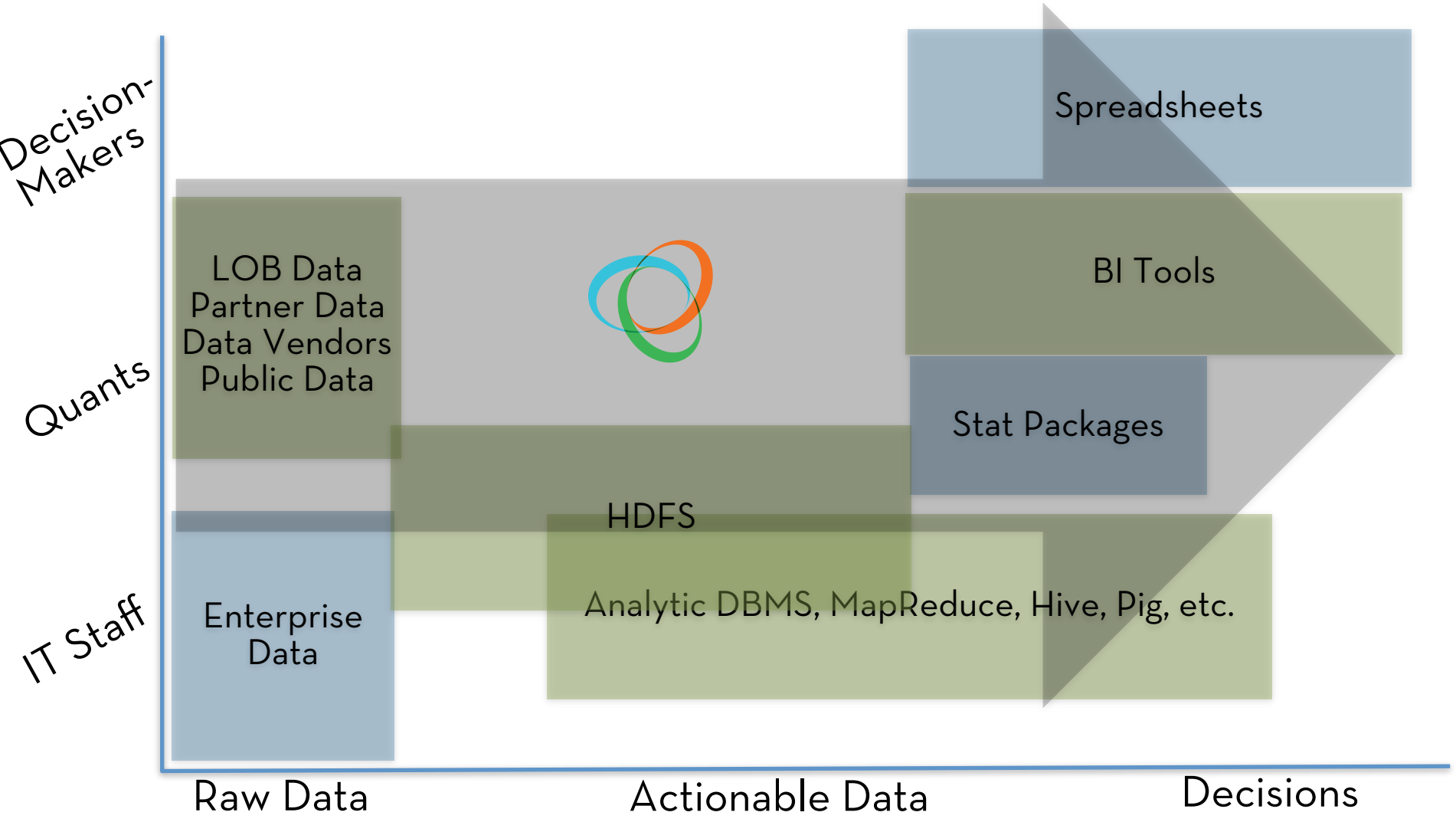
NEW LANDSCAPE



NEW LANDSCAPE



NEW LANDSCAPE



PEOPLE + DATA + COMPUTATION