A Semantic Map of the last.fm Music Folksonomy

Joseph Biberstine, Russell J. Duhon, Elisha Allgood, Katy Börner
Cyberinfrastructure for Network Science Center
School of Library and Information Science
Indiana University

André Skupin
Department of Geography
San Diego State University

INDIANA UNIVERSITY

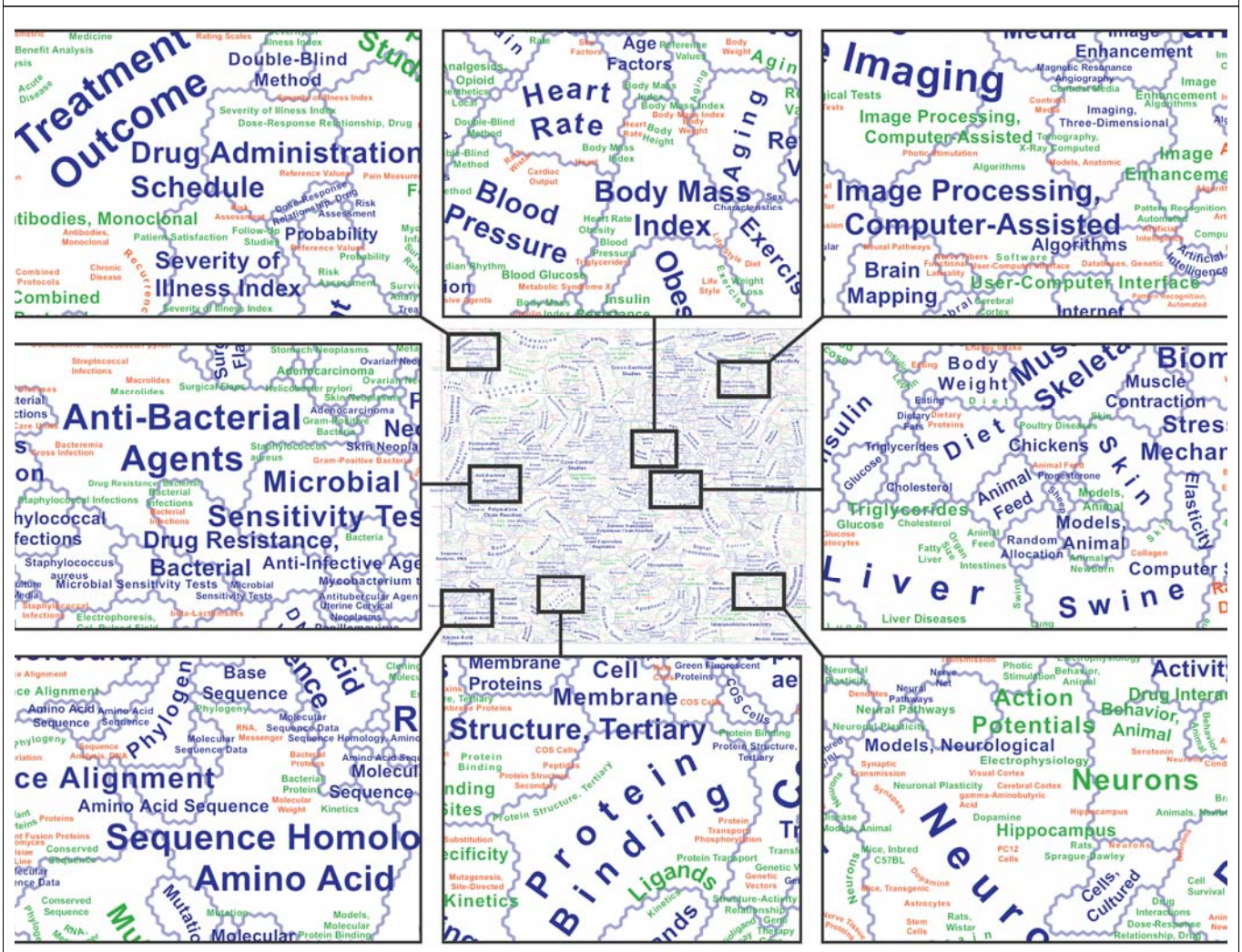
cyberinfrastructure for NETWORK SCIENCE CENTER
cns.iu.edu

SAN DIEGO STATE UNIVERSITY

Finalist, 2010 NSF/Science Visualization Challenge
Finalist, 2011 Intl. Institute for Information Design (IIID) Awards

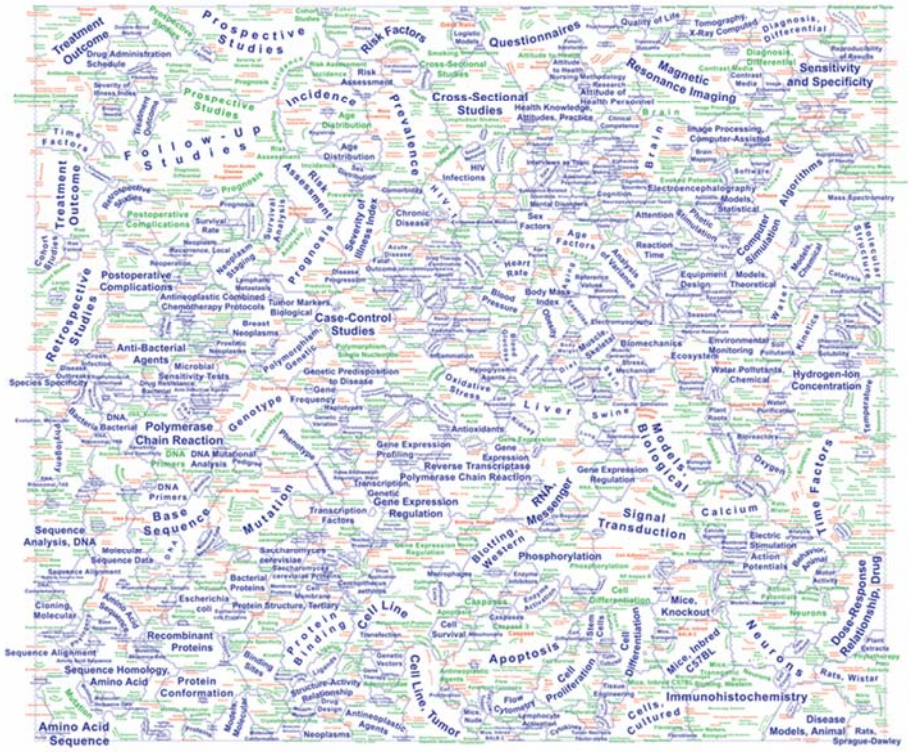
Background: Preceding Project

- "Accuracy of Models for Mapping the Medical Sciences"
  - NIH SBIR (HHSN268200900053C) & James S. McDonnell Foundation
  - Boyack, K.W., Newman, D., Duhon, R.D., Klavans, R., Patek, M., Biberstine, J.R., Schijvenaars, B., Skupin, A., Ma, N. and Börner, K. (2011) Clustering More Than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. *PLoS ONE*. 6(3): e18029.
- Collaborators on SOM portion:
  - Katy Börner, Joseph Biberstine, Russell Duhon (Indiana Univ.)
- Data: 2.15 Mio documents (Medline, 2004-2008)
- Methods:
  - vector-space model (VSM) → self-organizing map (SOM) → GIS
- Challenge:
  - serial run time (estim.): 4 years (!)
- Solution:
  - parallelization & supercomputing: runtime reduced to 6 days



A Semantic Landscape of 2,300 Medical Subject Headings

By **Joseph Biberstine**, **Russell J. Duhon**, **Jennifer R. S. Coffey**, **Katy Börner**, Cyberinfrastructure for Network Science Center, Indiana University, Bloomington and **André Skupin**, San Diego State University



Based on Over 2 Million MEDLINE Publications (2003-2008)

This landscape is a 275 by 275 grid of hexagonal neurons. Regions on the landscape are labeled by the MeSH terms with which their constituent neurons associate most strongly. Light purple borders separate regions defined by each neuron's single strongest term association; those regions are marked with purple labels. Green labels depict regions defined by the second-strongest term association of each neuron, and so on, as shown in the legend below.

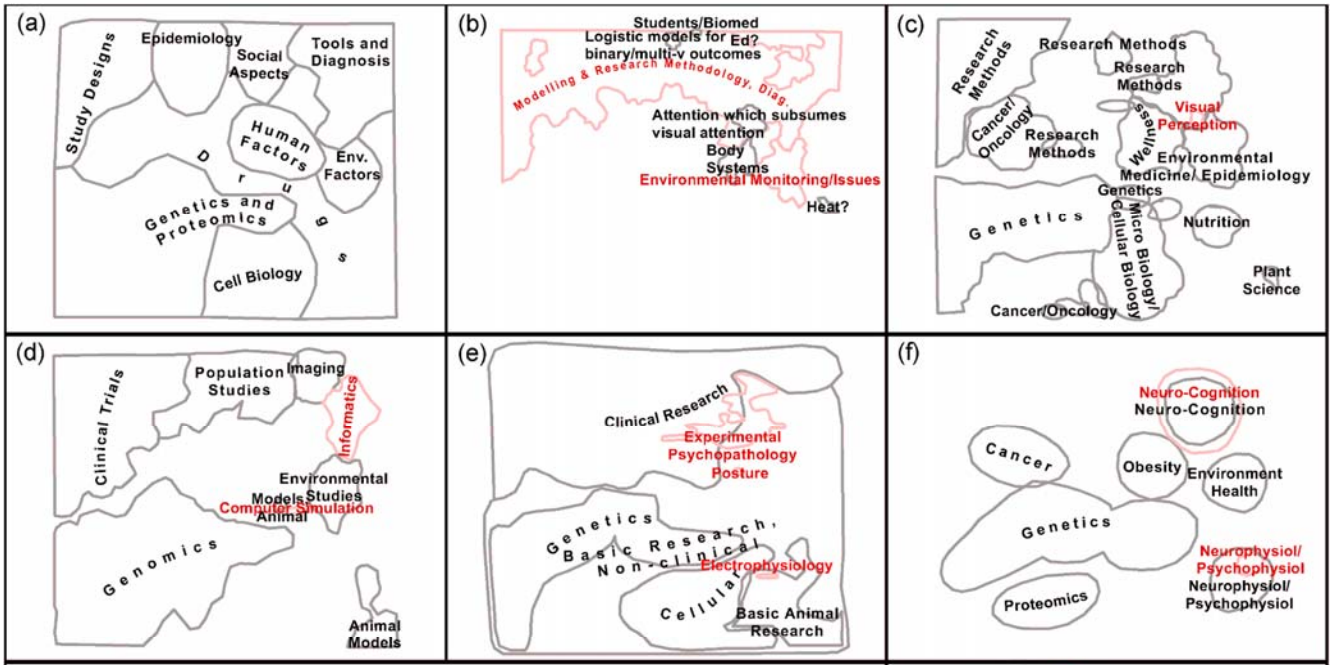
The size and shape of each label roughly reflects the extent of its region. The actual size and shape are somewhat flexible to guard against label overlap.

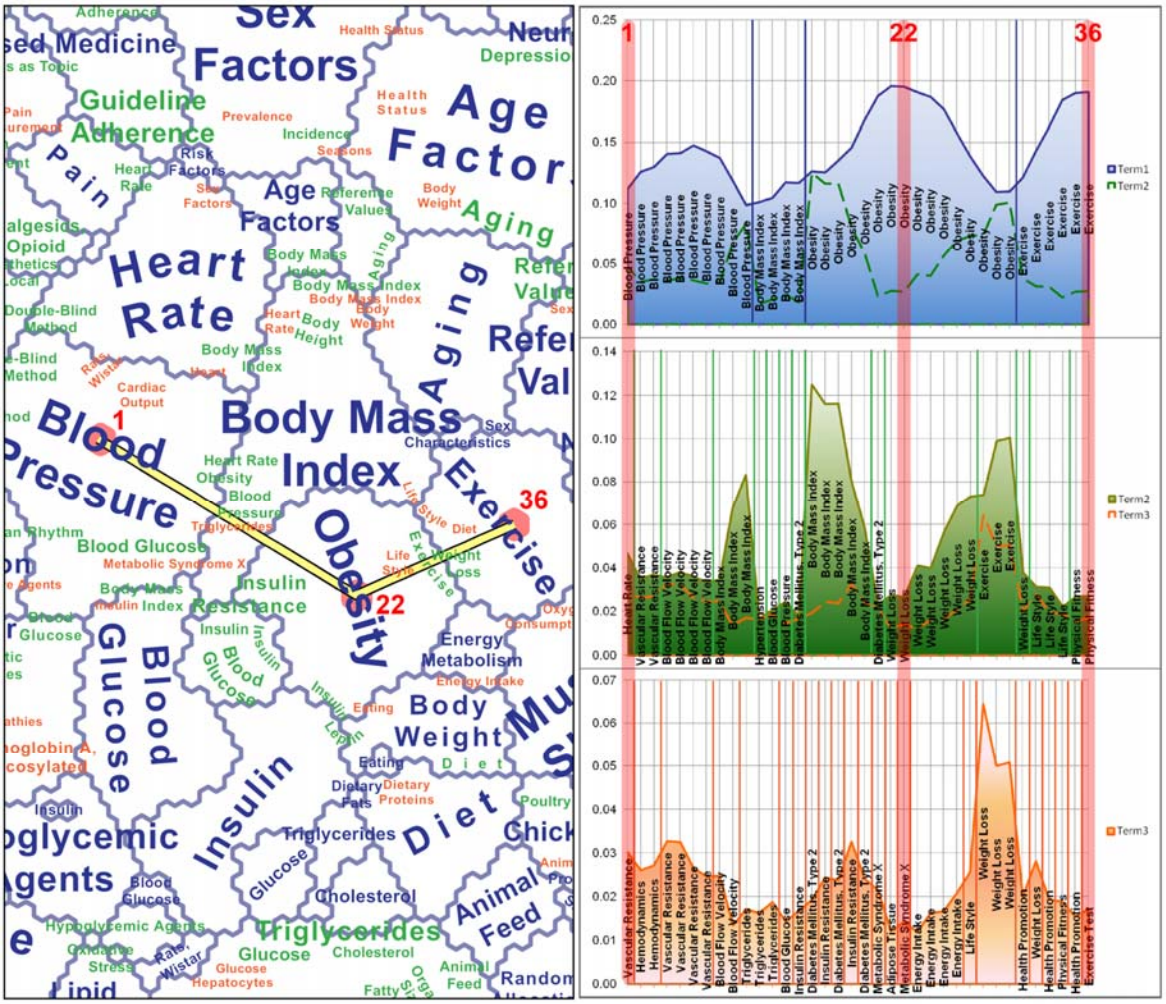
Acknowledgments

This research is funded by the Cyberinfrastructure for Network Science Center at the School of Library and Information Science, Indiana University; the National Science Foundation under award SBE-0738111; and NIH Awards R21DA024259 and IU24RR029822-01. Indiana University's Big Red supercomputer used in this study is supported by the National Science Foundation under Grant No. ACI-03386181, OCI-0451237, OCI-0535258, and OCI-0504075, a Shared University Research grants from IBM, Inc., and the Indiana METACyt Initiative supported in part by Lilly Endowment, Inc.

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Subject Expert Feedback





Motivations

- Pragmatic
  - non-hierarchical SOM now applicable to much larger, more high-dimensional data sets
  - → beyond *small* data
- Contextual
  - expose broader audience to knowledge visualization
  - → find general-interest topic
- Heretic
  - analysis of social media dominated by network concepts
  - → beyond network myopia
- Coincidental
  - existing last.fm data set (Schifanella et al 2010)

→ mapping the collective music mind

Raw Data - Source

- Last.fm is a social Internet radio site
  - users share information about songs they are listening to
  - they can also tag songs
  - with any strings of text they like

Raw Data - Summary

- Gathered during first half of 2009
- 99,405 registered users
  - 52,452 active
- 281,818 tags
- 1,393,559 songs
- 10,936,545 annotations
  - an annotation is a (user, tag, song) triple, a tagging event
- data originally collected for:
  - Schifanella, R., Barrat, A., Cattuto, C., Markines, B., and Menczer, F. (2010). Folks in Folksonomies: Social Link Prediction from Shared Metadata. *Proc. 3rd ACM International Conference on Web Search and Data Mining (WSDM)*.

Top Tags

rock	*singer-songwriter*	*heavy metal*
electronic	*80s*	*chillout*
seen live	*folk*	*dance*
indie	*hard rock*	*british*
alternative	*progressive rock*	*90s*
pop	*indie rock*	*psychedelic*
female vocalists	*electronica*	*blues*
jazz	*punk*	*hip-hop*
classic rock	*instrumental*	*post-rock*
experimental	*soul*	*new wave*
ambient	*black metal*	*soundtrack*
metal	*industrial*	*classical*
alternative rock	*death metal*	*00s*

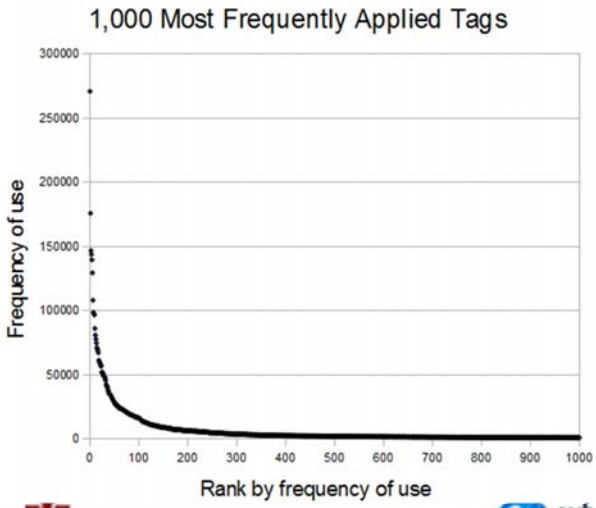
Tags Are More Than Just Genres

- Intensional
  - From recognized genres to simple objective facts
    - rock (rank 1)
    - electronic (2)
    - ..
    - female vocalists (7)
    - female vocalist (64)
    - acoustic (51)
- Extensional
  - A mix of social signals, properties of the user-song experience, and aides to personal categorization
    - seen live (3)
    - beautiful (48)
    - favorites (54)
    - albums i own (97)
    - altar of the metal gods (58)

INDIANA UNIVERSITY cyberinfrastructure for NETWORK SCIENCE CENTER cns.iu.edu SAN DIEGO STATE UNIVERSITY

Raw Data - Thresholding

- SOM method will not scale to 280,000+ tags/dims in raw form
- we consider only the 1,000 most frequently applied tags
- keep only songs annotated with any of these tags by some user
- characterize each song as a vector over each tag dimension summed across users



	Raw	Preprocessed
Tags	281,818	1,000
Songs	1,393,559	1,088,761
Avg Tags/Song	7.8	6.8

INDIANA UNIVERSITY
cyberinfrastructure for NETWORK SCIENCE CENTER cns.iu.edu
SAN DIEGO STATE UNIVERSITY

Self-Organizing Map Algorithm - Parallelized Implementation

- Task completely intractable using typical SOM software
- Preceding project trained on twice as many data and twice as many dimensions, using our own implementation
  - Divide the training data among multiple processes
  - Each process holds a complete copy of the map
  - Periodically synchronize process-local copies of the map to create a new process-global map
- Adapted with several project-specific optimizations from:
  - Lawrence, R.D., Almasi, G.S., Rushmeier, H.E. (1999). A Scalable Parallel Algorithm for Self-Organizing Maps with Applications to Sparse Data Mining Problems. *Data Mining and Knowledge Discovery*.

INDIANA UNIVERSITY
cyberinfrastructure for NETWORK SCIENCE CENTER cns.iu.edu
SAN DIEGO STATE UNIVERSITY

Training the Map

- 2D hexagonal lattice of 32,400 neurons (180x180)
- Input space metric: cosine similarity
  - Induced interpretation: Each training vector (and so consequently each neuron vector) represents a direction in the 1,000-dimensional tag space

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

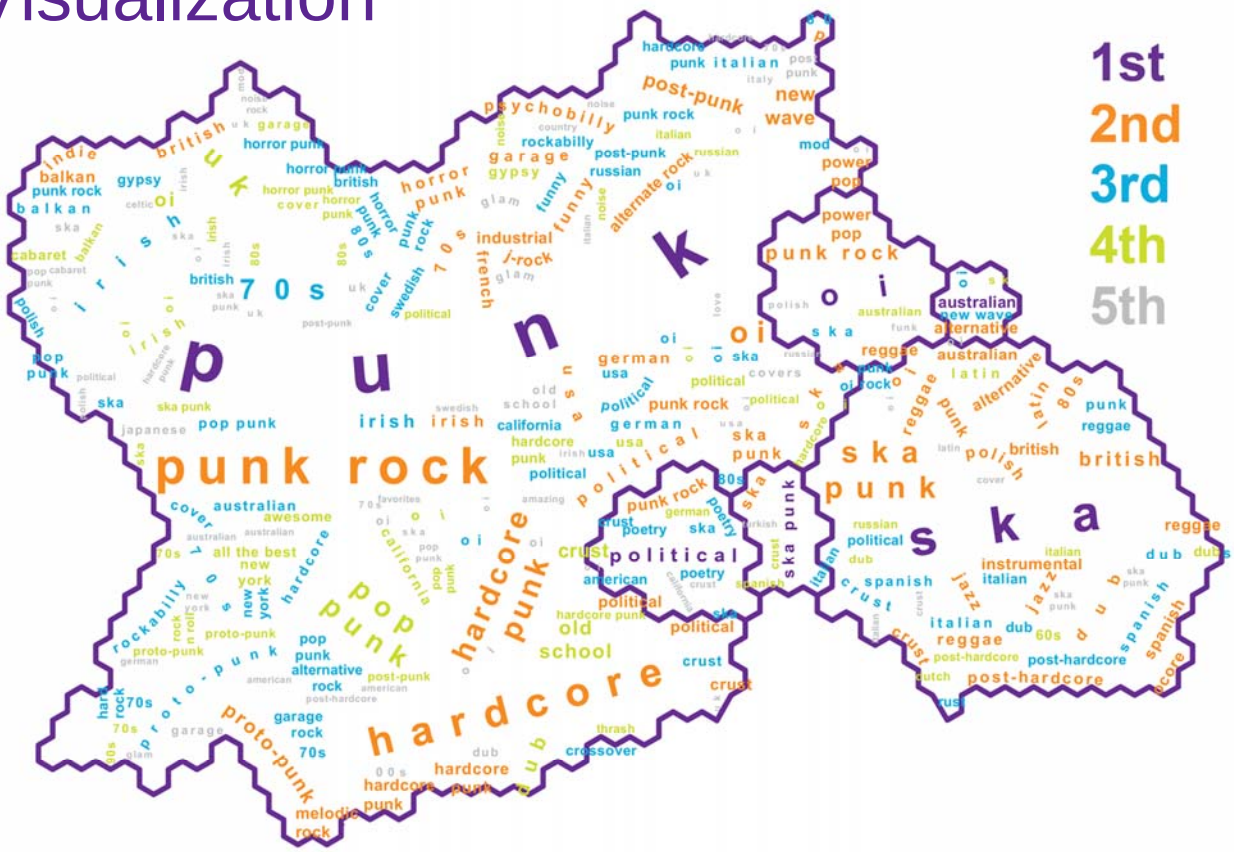
- 50 complete passes over the training data
- 300 processes across 100 compute nodes of Big Red, a supercomputer at Indiana University
  - Parallel runtime = 13 hours
  - Serial equivalent runtime = 5 months

INDIANA UNIVERSITY

cyberinfrastructure for NETWORK SCIENCE CENTER
cns.iu.edu

SAN DIEGO STATE UNIVERSITY

Visualization



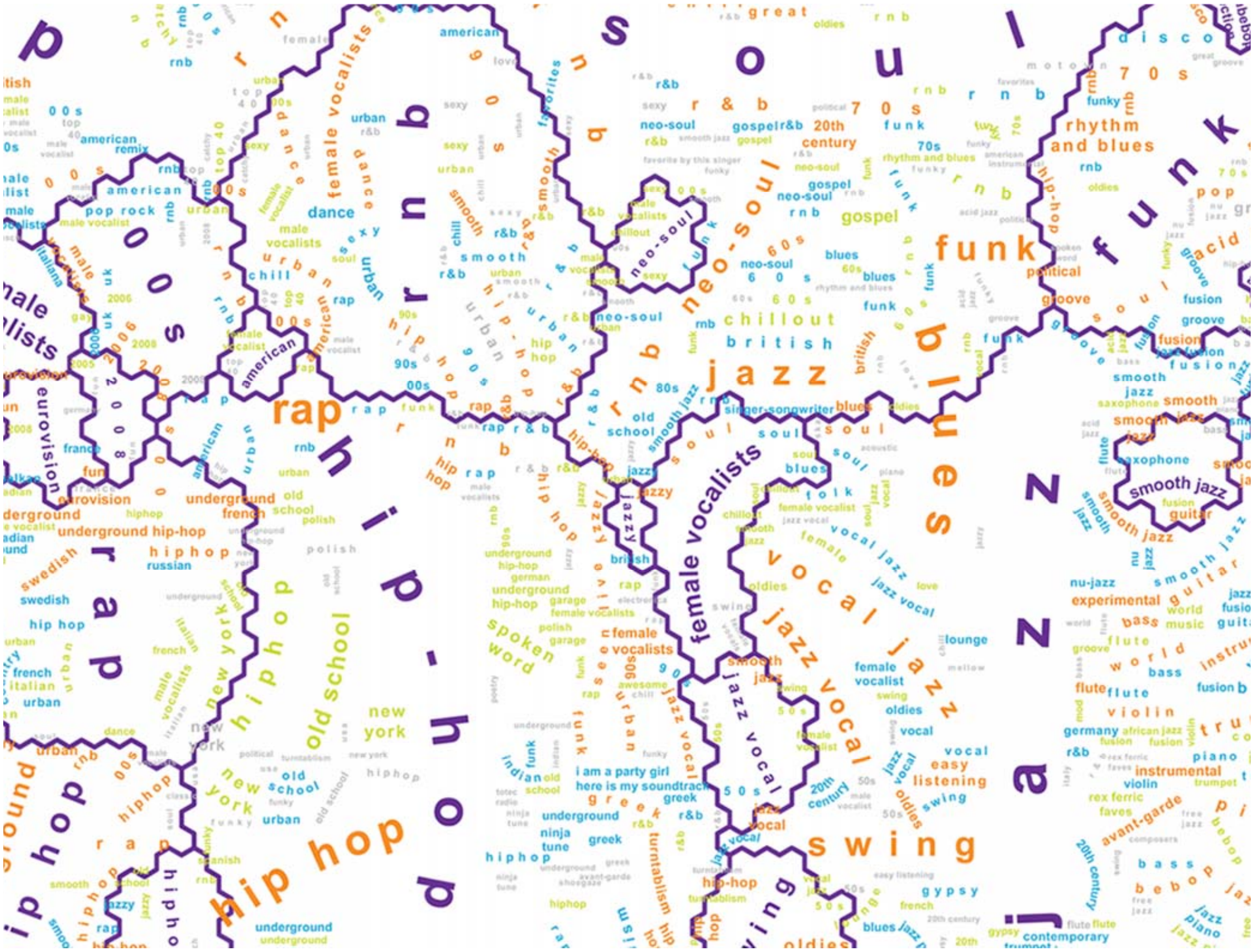
INDIANA UNIVERSITY

cyberinfrastructure for NETWORK SCIENCE CENTER
cns.iu.edu

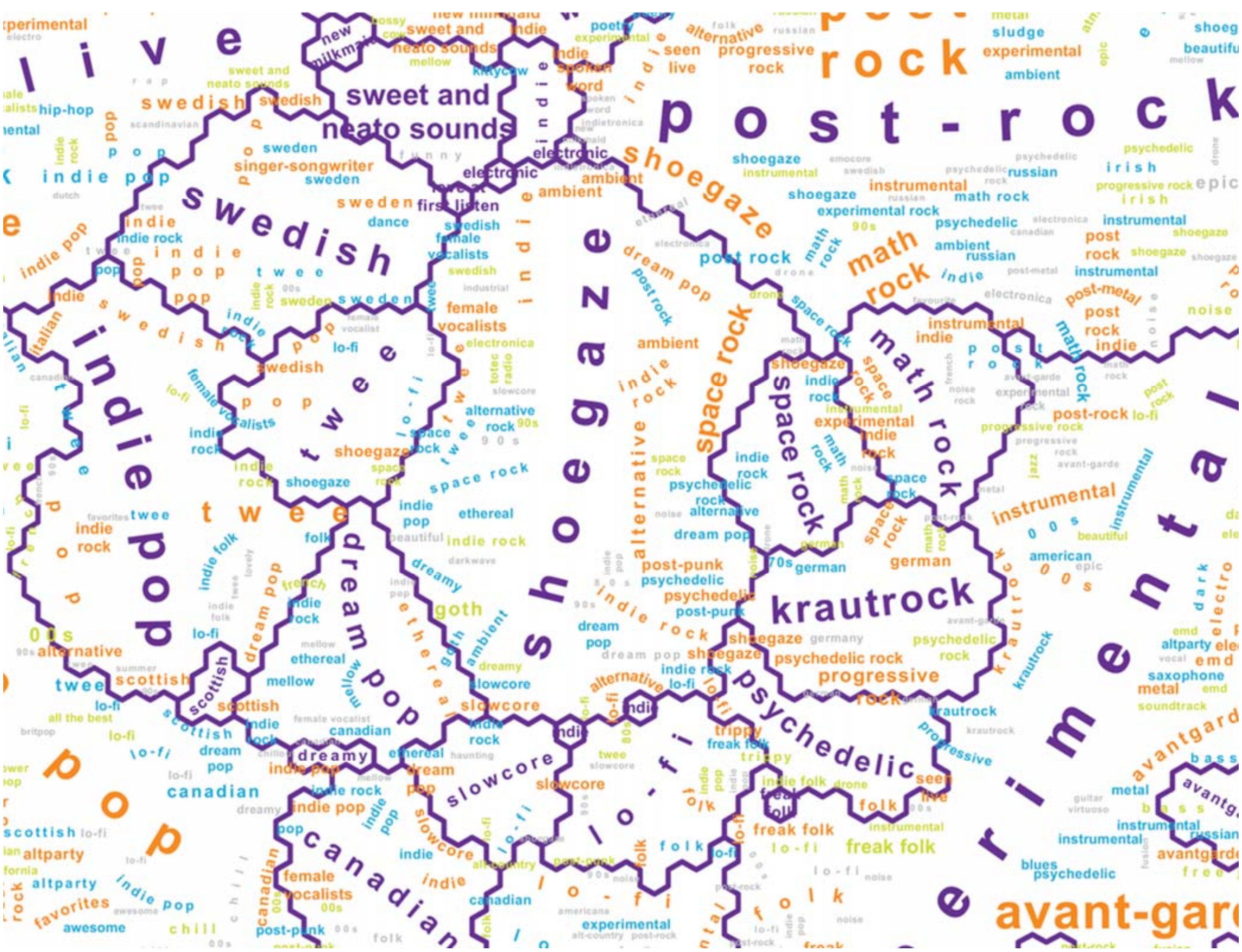
SAN DIEGO STATE UNIVERSITY

Interpretation – What viewers have reported

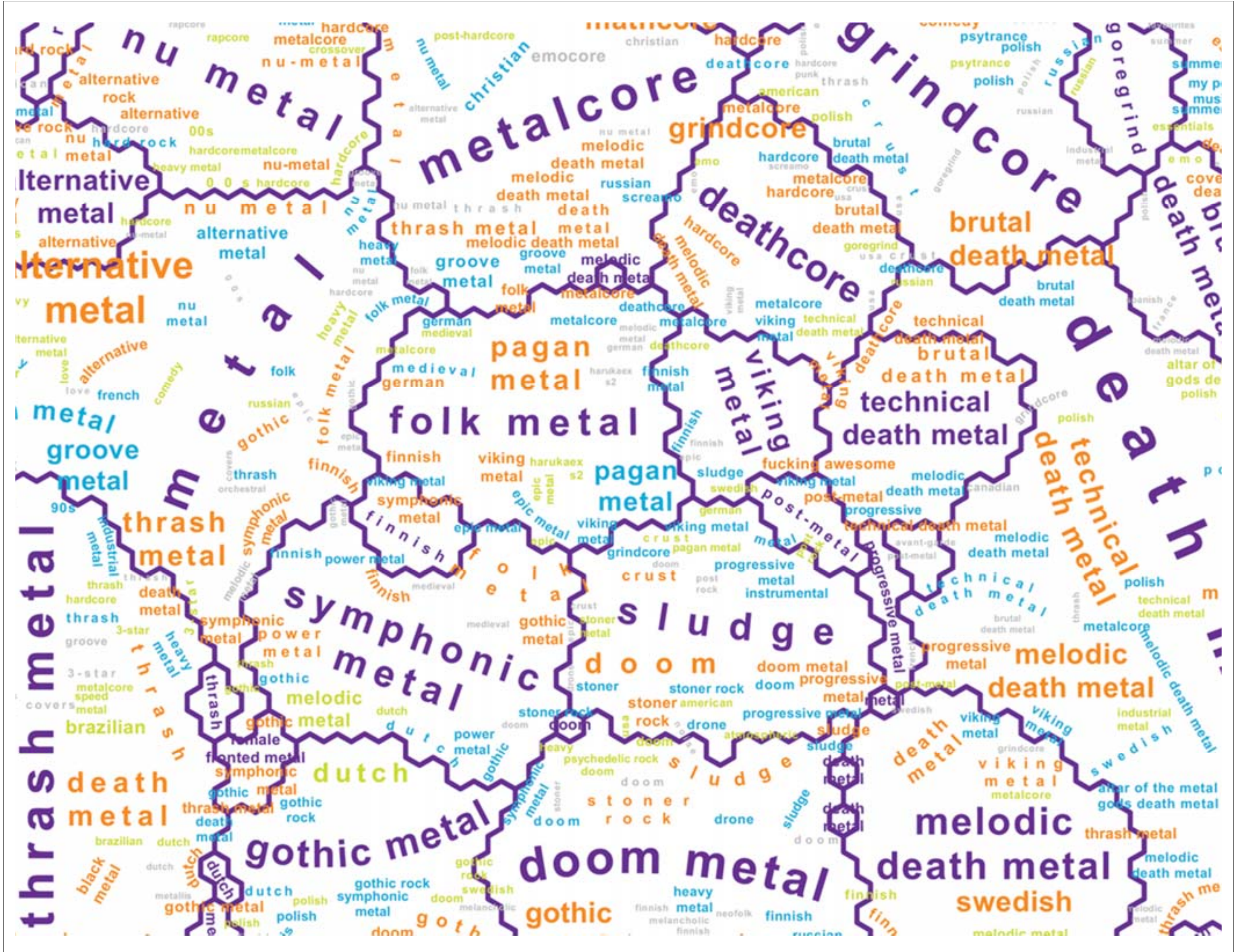
- Recognition
  - Patterns of hierarchical and neighborhood relationships among tags match expectations
    - *rap* & *hip-hop* regions are neighbors
    - *bebop* region is contained in *jazz*
- Discovery
  - Opportunity to find new musical categories
    - *shoegaze*, *lo-fi*, *twee*, …
  - Variety of top-level categories
    - *folk metal*, *viking metal*, *technical death metal*, *thrash metal*, …
- Surprise
  - Relationship between *rock*, *blues*, and *jazz*



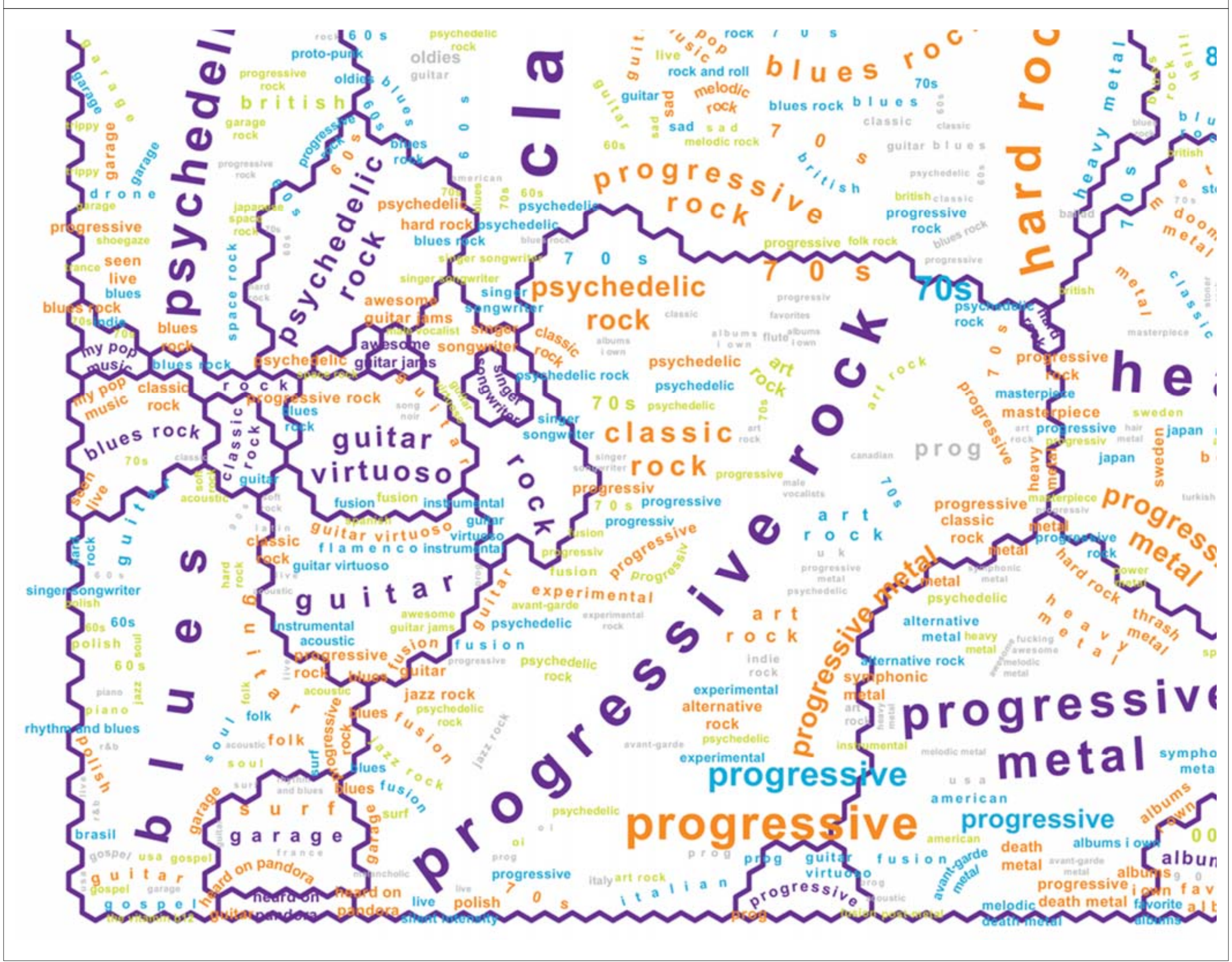
soul
rnb
hip-hop
rap
funk
jazz
blues
neo-soul
female vocalists
vocal jazz
jazz vocal
swing
smooth jazz
hip hop
old school
hiphop
underground
eurovision
male vocalists
00s
american
urban
r&b
70s
rhythm and blues
disco
gospel
chillout
british
oldies
experimental
new york



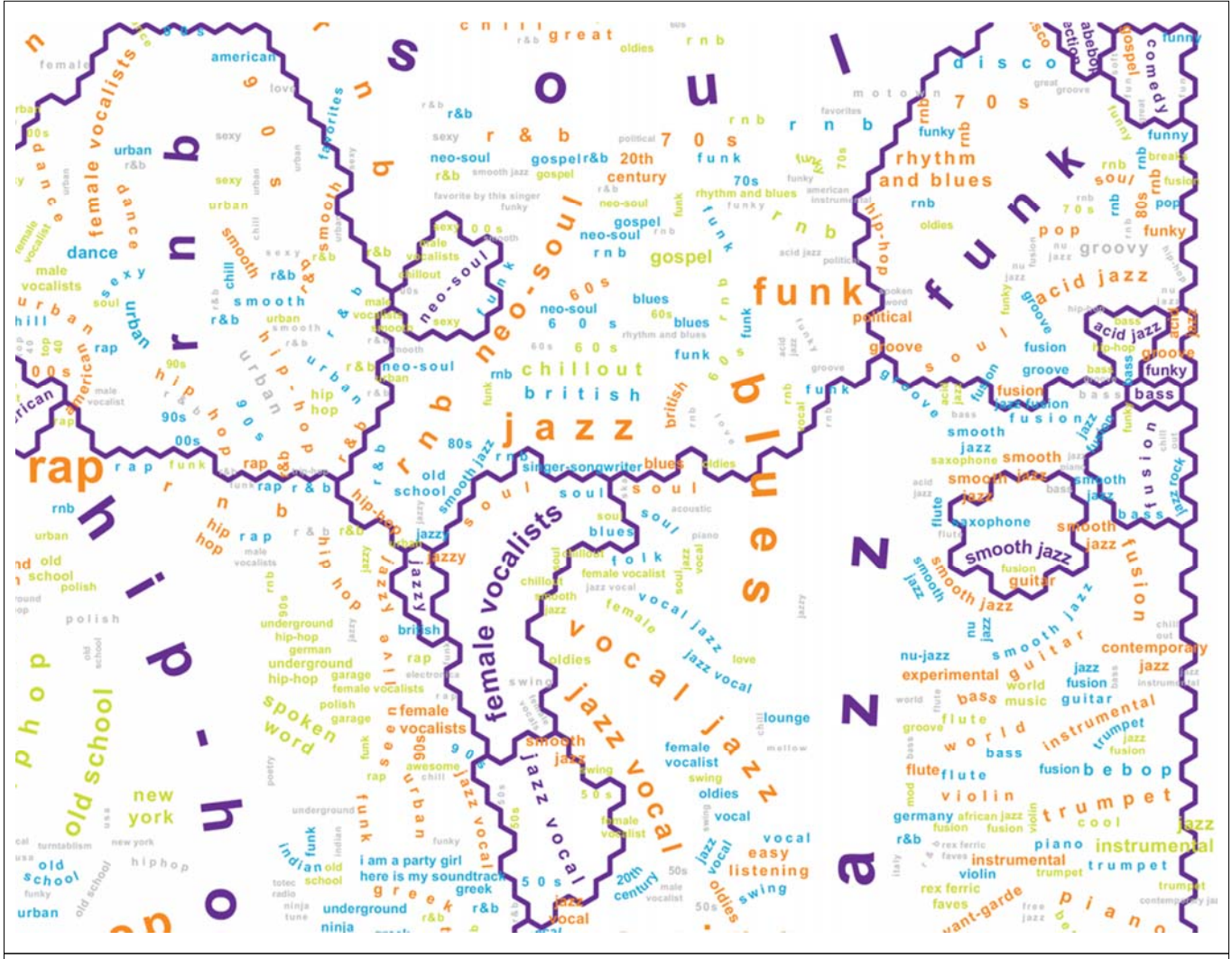
live
post-rock
rock
experimental
sweet and neato sounds
swedish
indie pop
indie
twee
shoegaze
dream pop
math rock
space rock
krautrock
psychedelic
lo-fi
slowcore
canadian
pop
instrumental
avant-garde
freak folk
progressive rock
scottish
electronic
ambient



nu metal
metalcore
grindcore
deathcore
alternative metal
metal
death metal
pagan metal
folk metal
viking metal
technical death metal
thrash metal
symphonic metal
sludge
doom
melodic death metal
gothic metal
doom metal
dutch
gothic
swedish

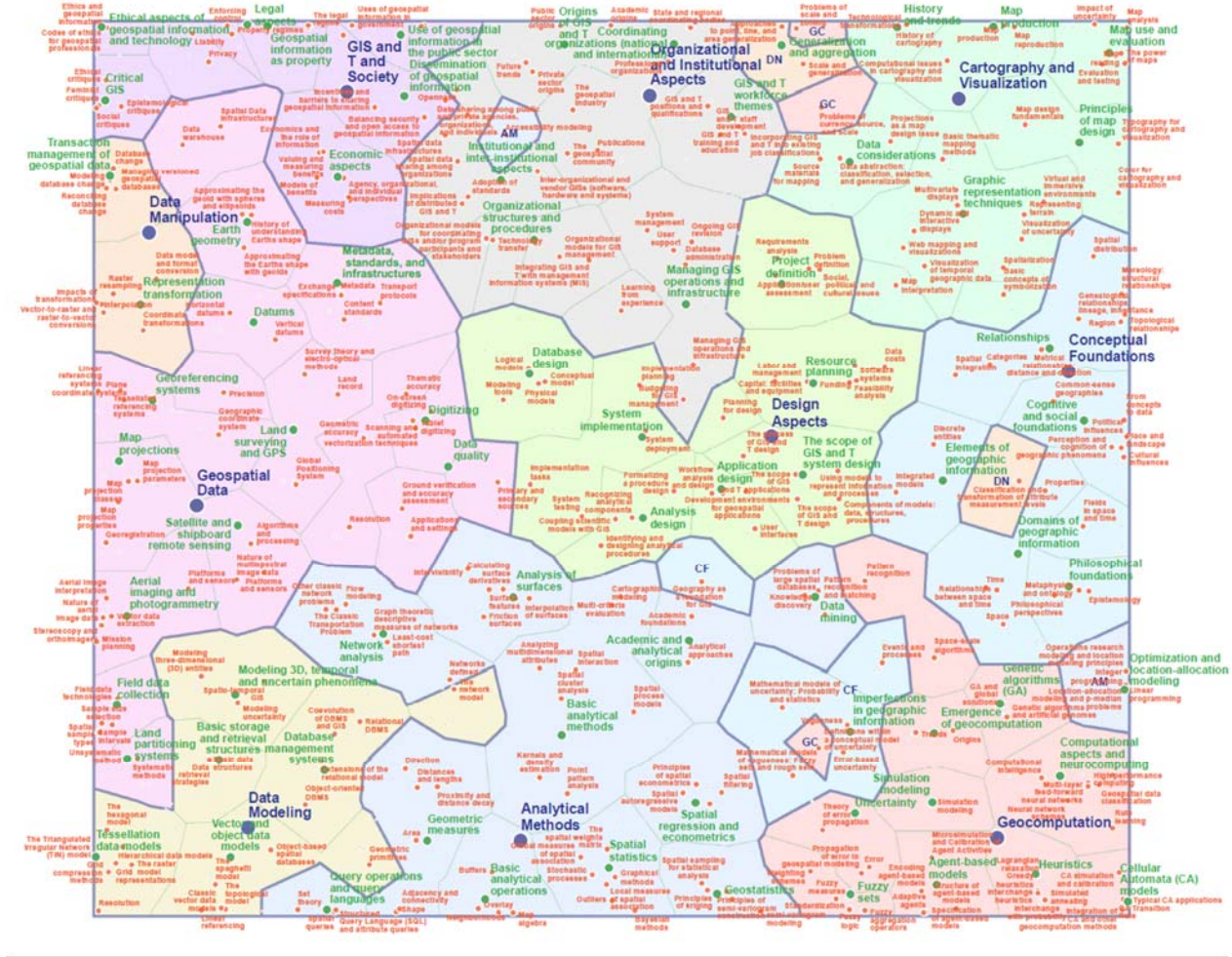


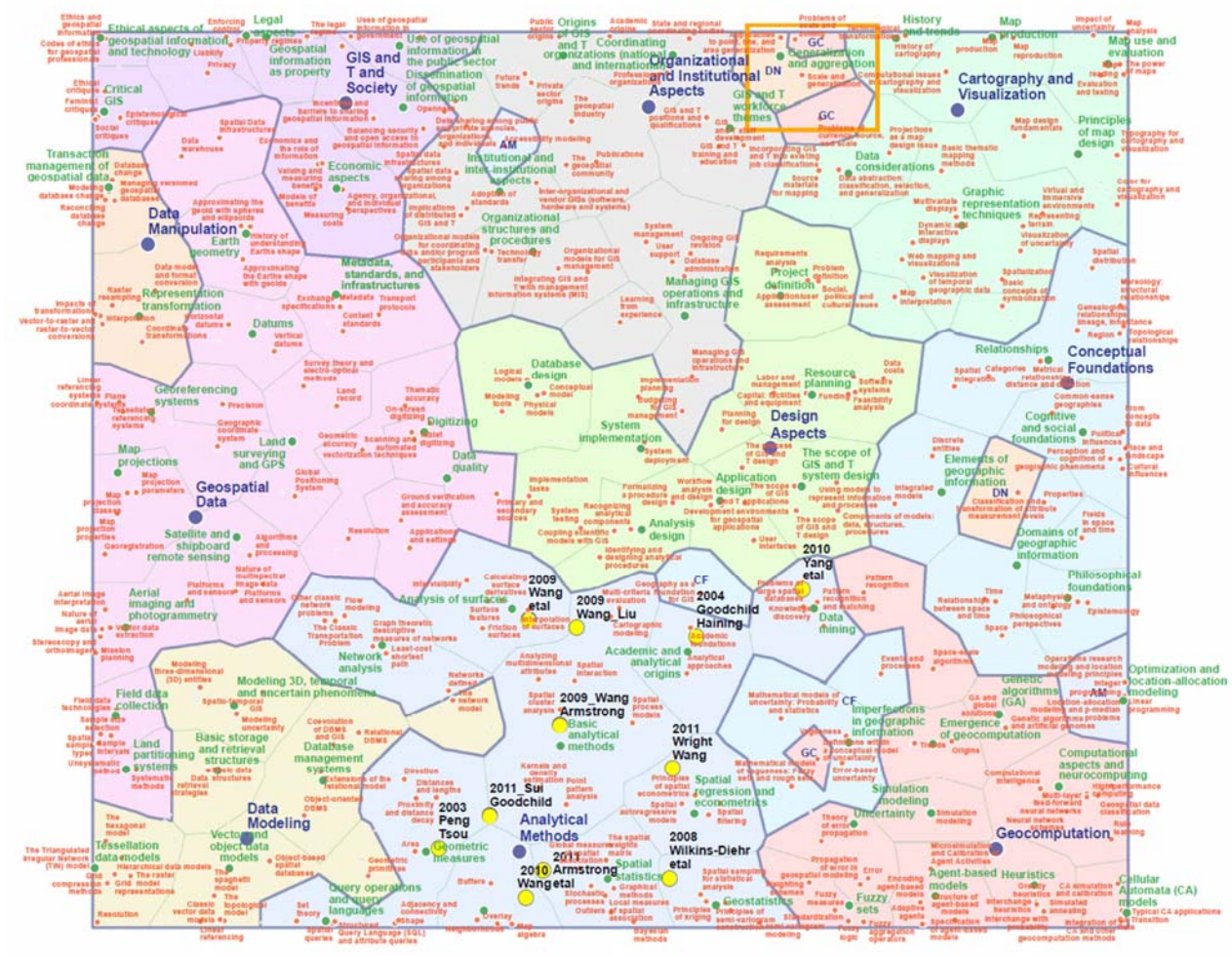
psychedelic
cla
hard roc
blues rock
progressive rock
psychedelic rock
70s
classic rock
guitar virtuoso
guitar
blues
progressive rock
progressive metal
progressive
garage
surf
he

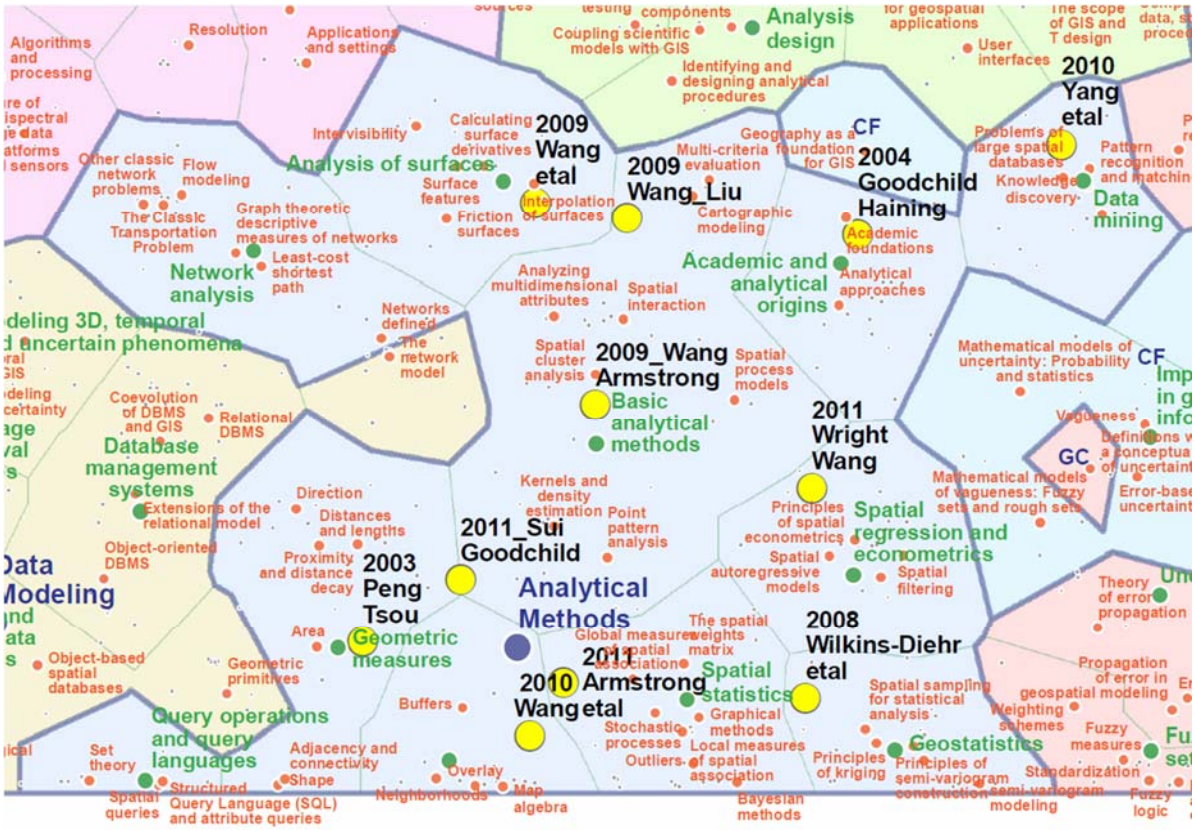


Future Work

- Limited vocabulary and vector sparseness addressed through:
  - memory-efficient computation
  - preprocessing with latent Dirichlet allocation
    - efficient grouping of tags
    - support for topical inference


GIS and T and Society
Organizational and Institutional Aspects
Cartography and Visualization
Data Manipulation
Metadata, standards, and infrastructures
Conceptual Foundations
Geospatial Data
Design Aspects
Data Modeling
Analytical Methods
Geocomputation

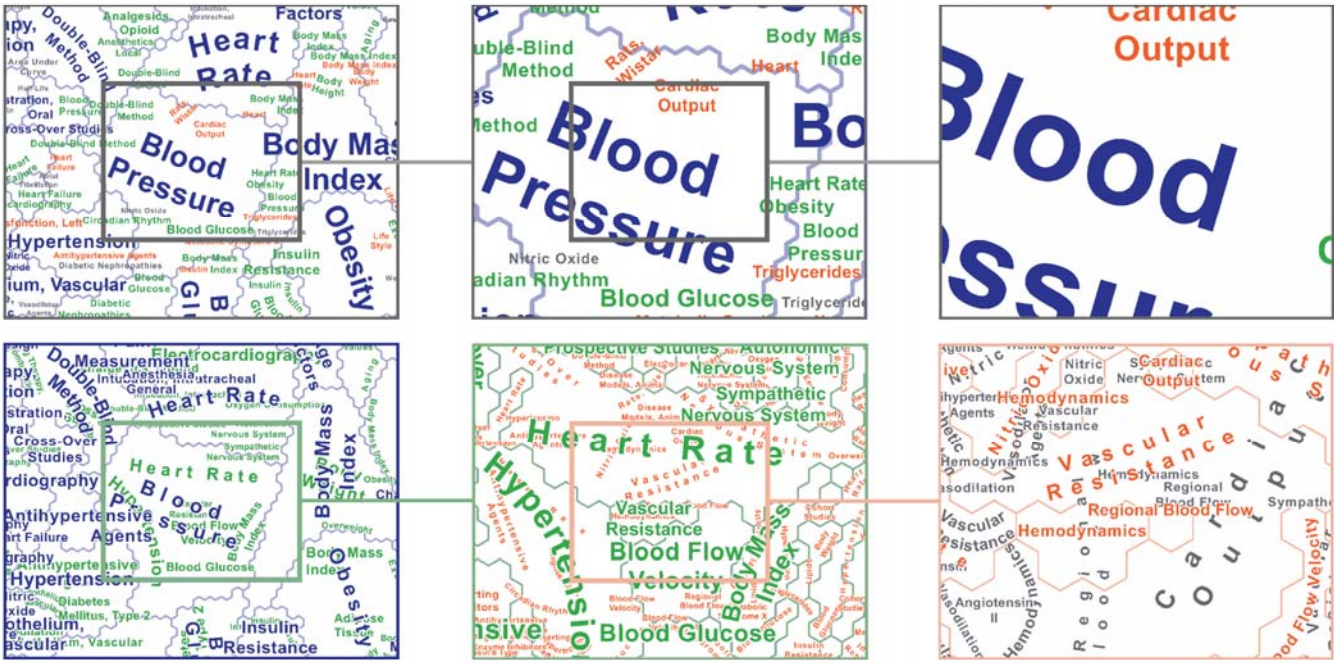

2010 Yang etal
2009 Wang etal
2009_Wang_Liu
2004 Goodchild Haining
2009_Wang Armstrong
2011 Wright Wang
2011_Sui Goodchild
2003 Peng Tsou
2011 Armstrong Wang etal
2008 Wilkins-Diehr etal



Future Work

- Limited vocabulary and vector sparseness addressed through:
  - memory-efficient computation
  - Preprocessing with latent Dirichlet allocation
    - Efficient grouping of tags
    - Support for topical inference
- Static versus interactive map
  - Leverage Web GIS

geometric zoom



semantic zoom

Potential Applications

- Interactive music navigator and playlist generator
- Mapping portfolios as fields of neuronal activation
  - for the set of songs associated with any entity, we can see where in the world they belong
    - user: her/his favorite songs
    - band: their complete work
    - group of users: What is their turf?
  - … or look at derivatives of these fields
    - What is the difference between *The Who* and *The Guess Who*?
    - How has this entity moved through the world of music over time?
    - Where have listeners like me headed next?

see also: Skupin, A. (2009) Discrete and Continuous Conceptualizations of Science: Implications for Knowledge Domain Visualization. *Journal of Informetrics*. 3(3): 233-245.

Contributors

- Joseph Biberstine, Indiana University
- André Skupin, San Diego State University
  - contact: skupin@mail.sdsu.edu
- Russell J. Duhon, IU
- Elisha Allgood, IU
- Katy Börner, IU

Funding

- School of Library and Information Science, Indiana University
- Cyberinfrastructure for Network Science Center, Indiana University
- National Science Foundation