SAND2011-5070C

Okay, Now What?

Challenges in Large Data Analysis

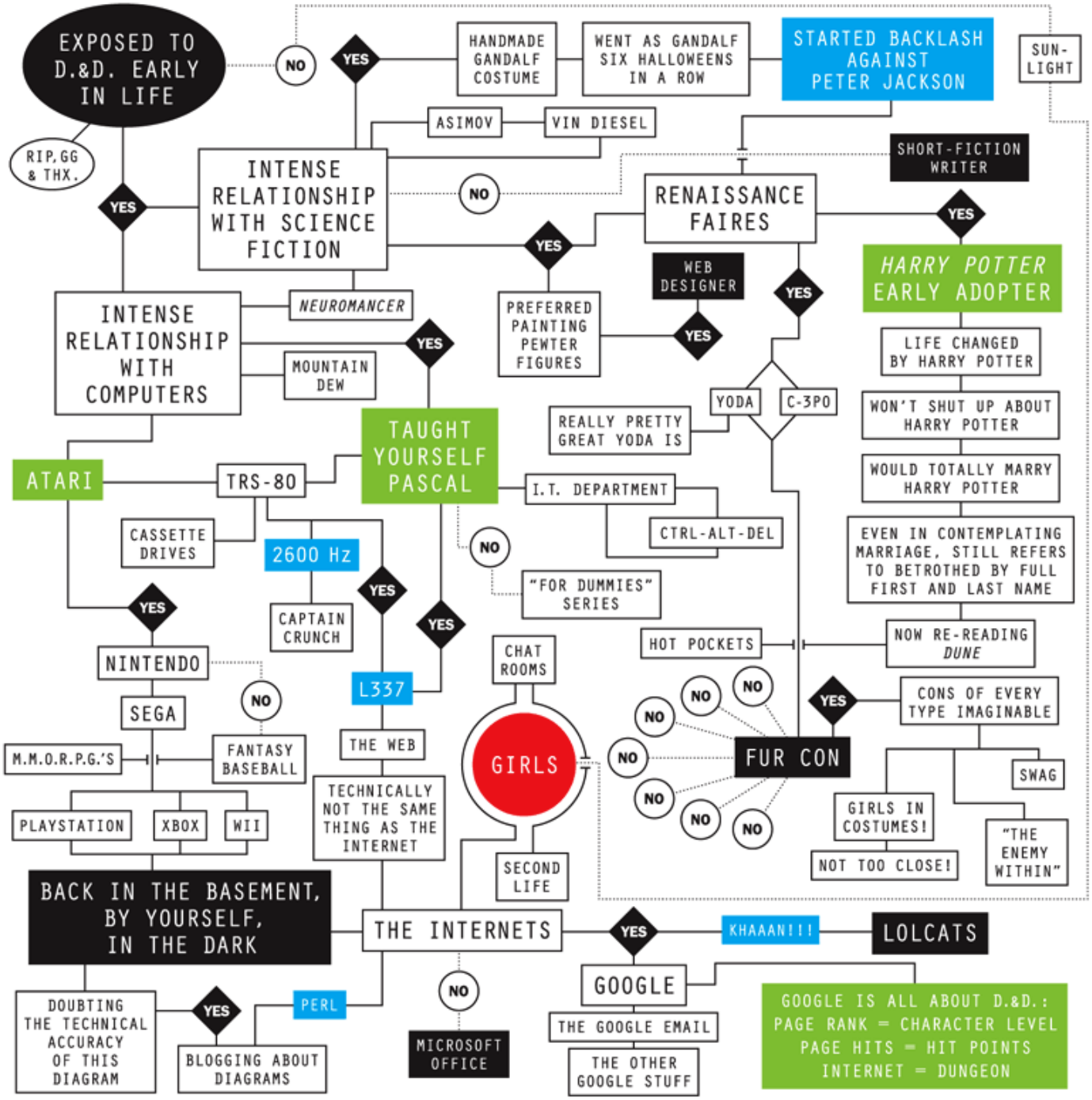
July 26, 2011

Andy Wilson
Scalable Analysis and Visualization
Sandia National Laboratories

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Analysis Is…

Sandia National Laboratories



Sandia National Laboratories

Sandia National Laboratories

Problem #2: The Semantic Gap

[1] Research Associate Professor, Department of Plant Microbiology and Pathology, University of Missouri, Columbia, MO 65211. Current address: Research Plant Pathologist, USDA ARS Crop Genetics and Production Unit, Jackson, TN 38301.

[2] Research Associate, Department of Plant Pathology, The Ohio State University, Columbus, OH 43210. Current address: Product Labels Manager, Monsanto, St. Louis, MO 63167.

[3] Professor Agronomist, Department of Horticulture and Crop Science, The Ohio State University, Columbus, OH 43210.

[4] Research Associate, Department of Plant Pathology, Iowa State University. Current address: Coordinator, Master Gardener Program, Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN 47907.

[5] Research Plant Pathologist, USDA, ARS Soybean/Maize Germplasm, Pathology, Genetics Research Unit, and Professor, Department of Crop Sciences, University of Illinois, Urbana, IL 61801.

[6] Professor, Department of Plant Pathology, University of Wisconsin, Madison, WI 53706.

[7] Research Extension Nematologist, Department of Entomology, Purdue University, West Lafayette, IN 47907.

[8] Professor, Department of Entomology, Purdue University, West Lafayette, IN 47907.

[9] Professor, Department of Plant Pathology, Kansas State University, Manhattan, KS 66506.

[10] Associate Professor, Department of Entomology, Michigan State University, East Lansing, MI 48824.

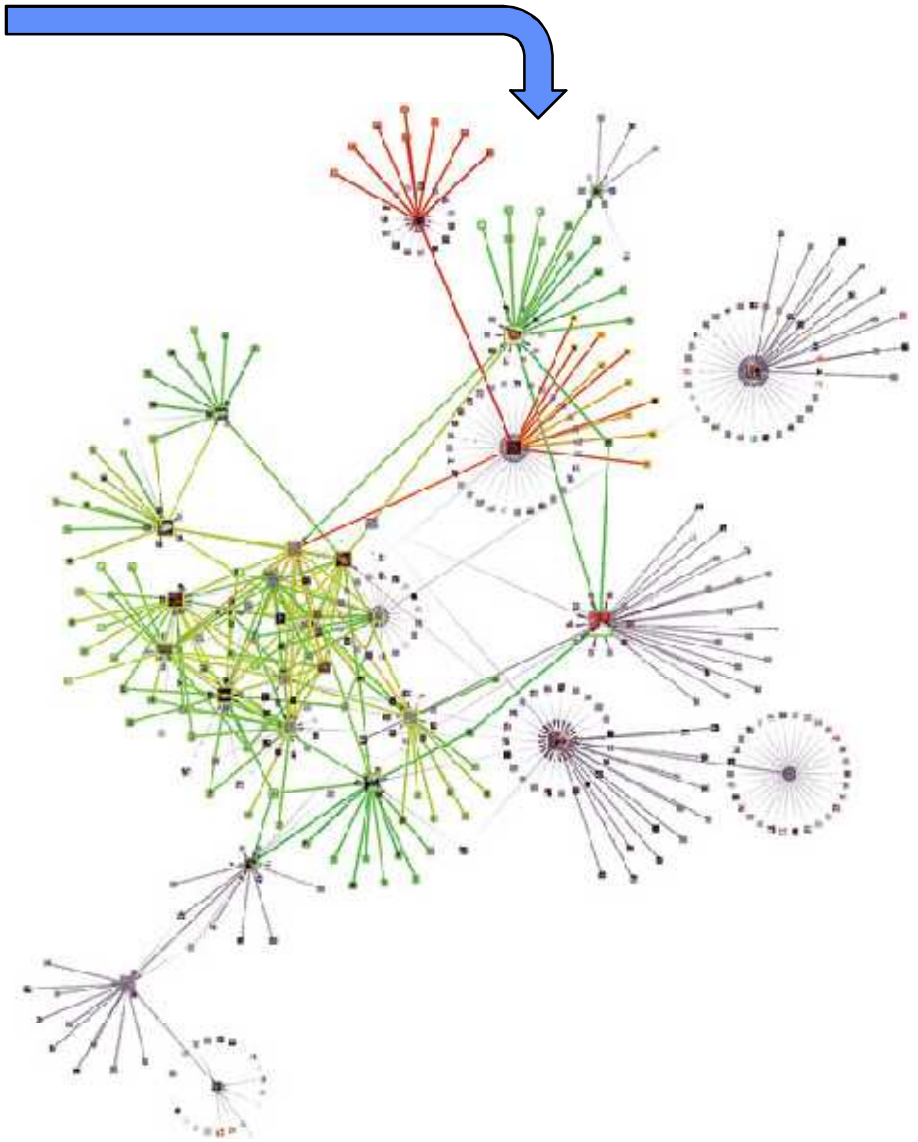
[11] Associate Professor, Department of Plant Microbiology and Pathology, University of Missouri, Columbia, MO 65211. Current address: Professor, Department Crop Sciences, University of Illinois, Urbana, IL 61801

[12] Professor Emeritus, Department of Plant Pathology, University of Minnesota, St. Paul, MN 55108.

[13] Professor, Department of Plant Pathology, Iowa State University, Ames, IA 50011.

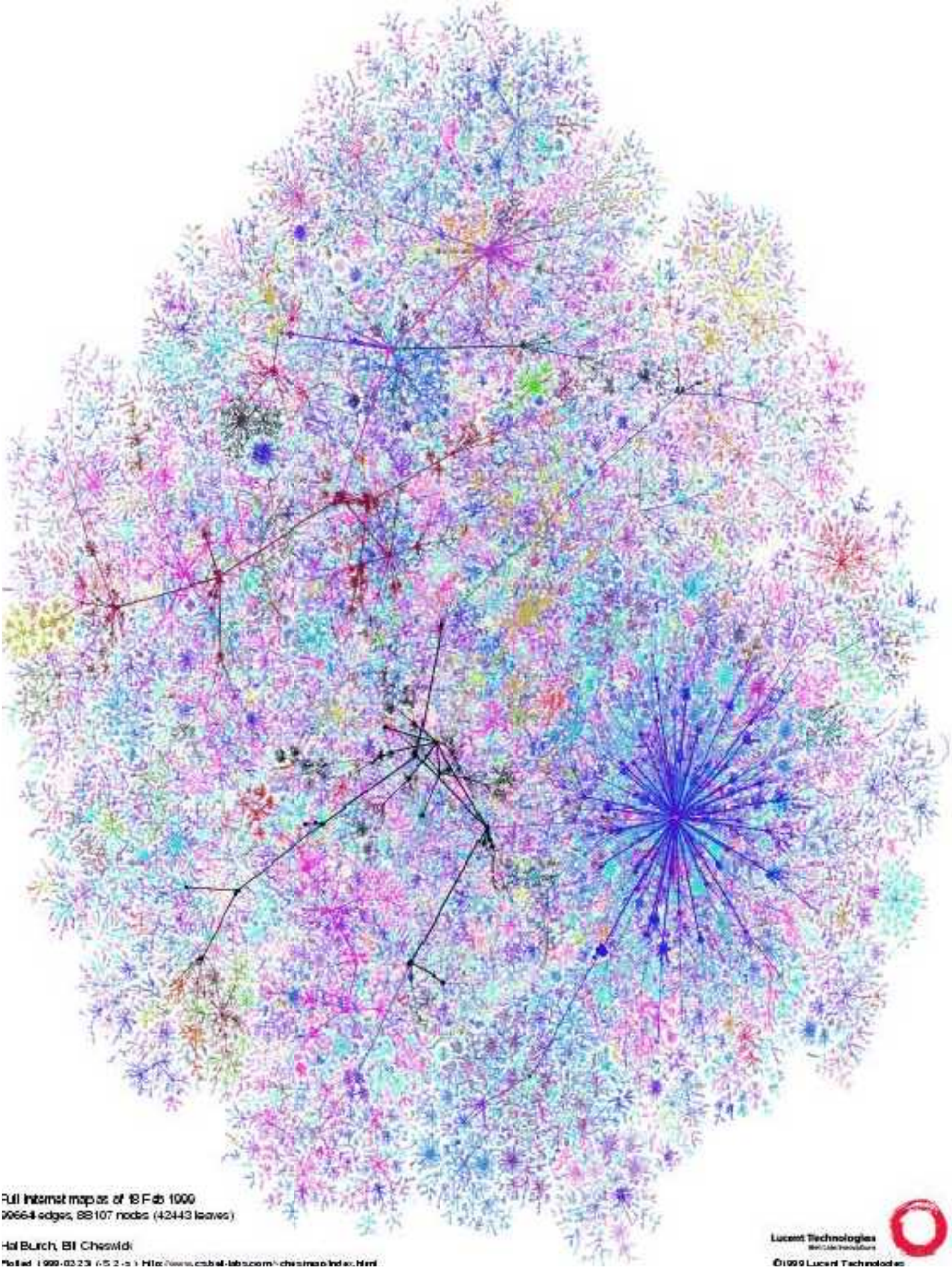
[14] Research Associate, Department of Plant Pathology, The Ohio State University, Columbus, OH 43210. Current address: Associate Professor, Texas Agricultural Experimental Station, Rt. 3, Box 219, Lubbock, TX 79401.

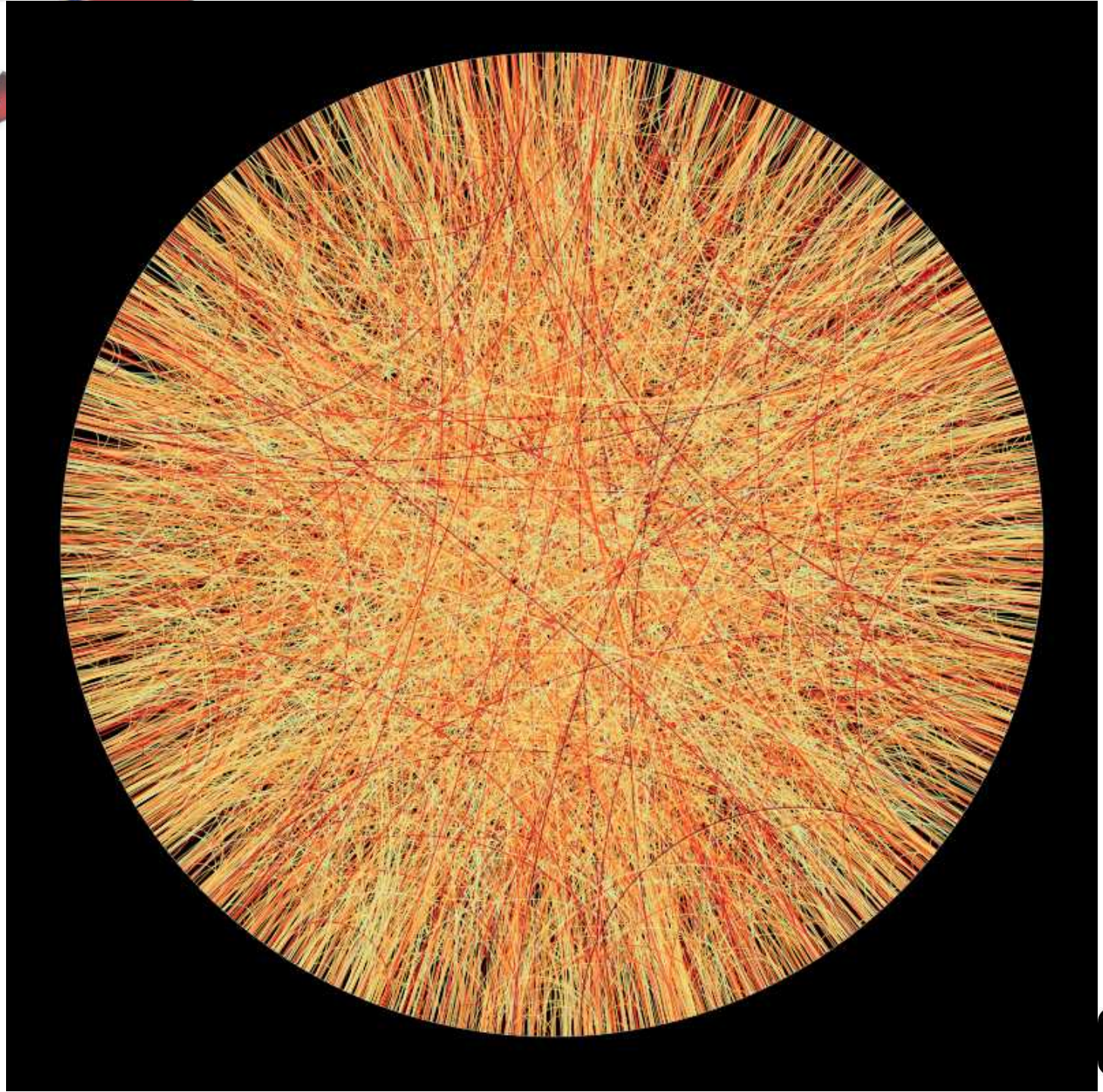
[15] Professor Emeritus, Department of Plant Pathology, University of Nebraska, Lincoln, NE 68583.



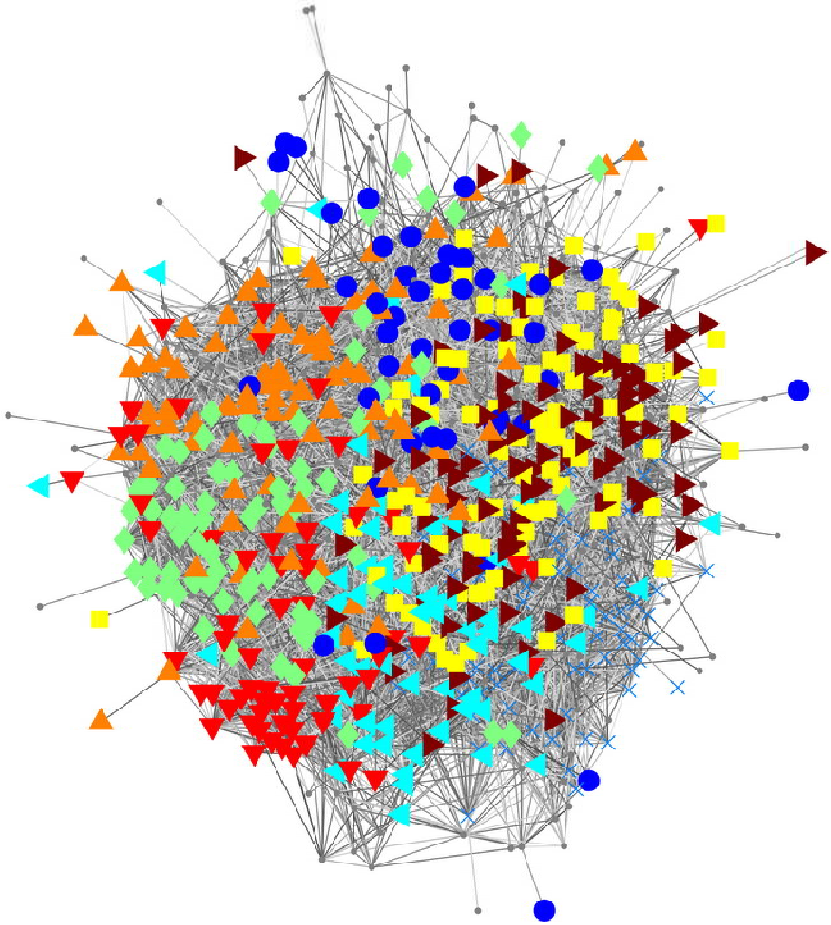
So here's what we do…

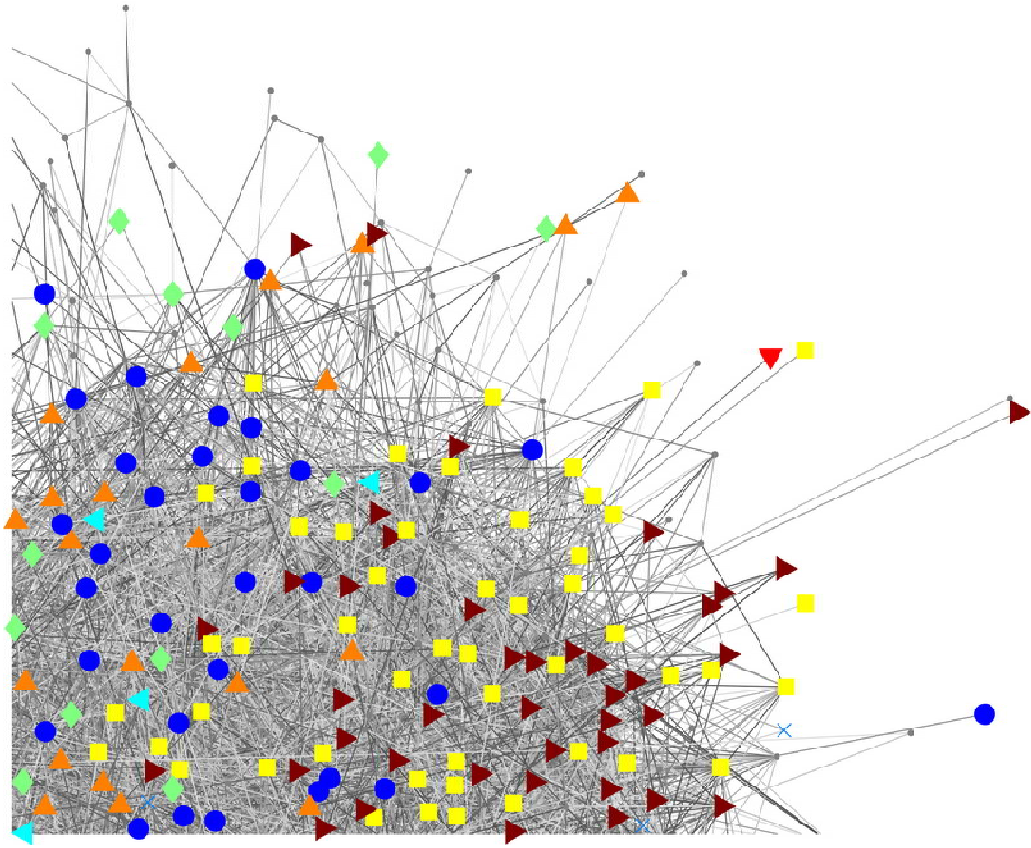
- If there are two things that computers are good at, it's
  - Handling large amounts of data and
  - Converting data from one representation to another
- Use computers to create a (hopefully manageable) amount of information in usable form. Abstract away the minute detail.
- Except…


Full Internet map as of 18 Feb 1999
99664 edges, 88107 nodes (42443 leaves)
Hal Burch, Bill Cheswick
Mapped 1999-02-23 (-5 2 -1) http://www.cs.bell-labs.com/~ches/map/index.html
Lucent Technologies
©1999 Lucent Technologies



Sandia National Laboratories





"Comparing Community Structure to Characteristics in Online Collegiate Networks", Traud et al.

Sandia National Laboratories

MY THESIS

Sensemaking and narrative formation are best aided by tools that provide multiscale summaries, familiar metaphors and expose more detail on request.

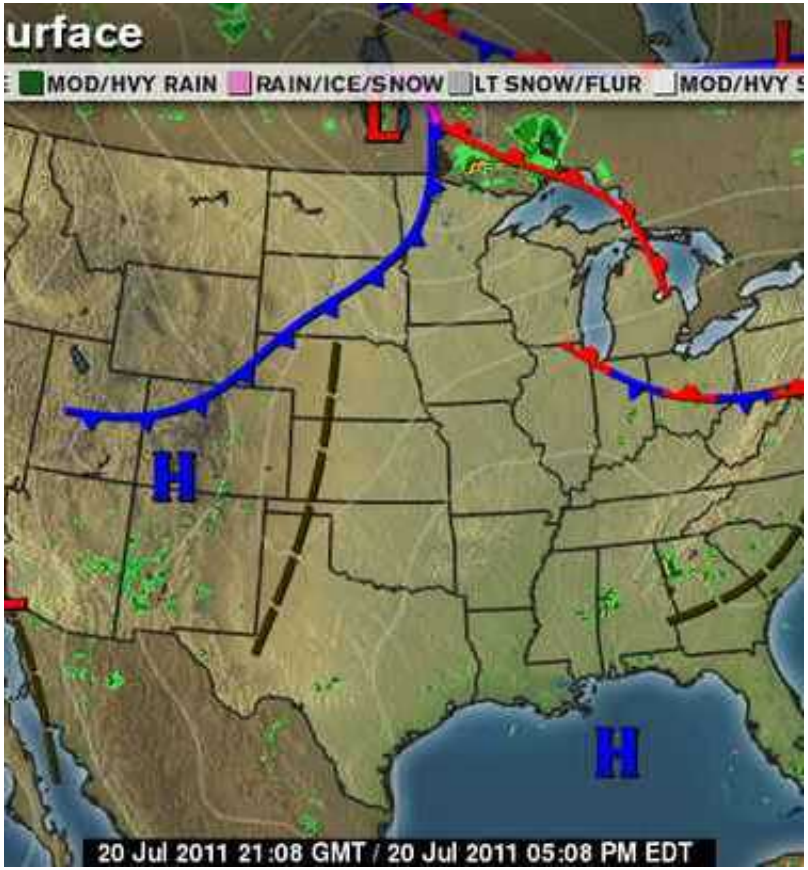
Sandia National Laboratories

Principle 1: Analyze While Computing

Sandia National Laboratories

Connect analysis code to the running simulation to extract exactly the structures you want with maximum fidelity or even guide the computation.

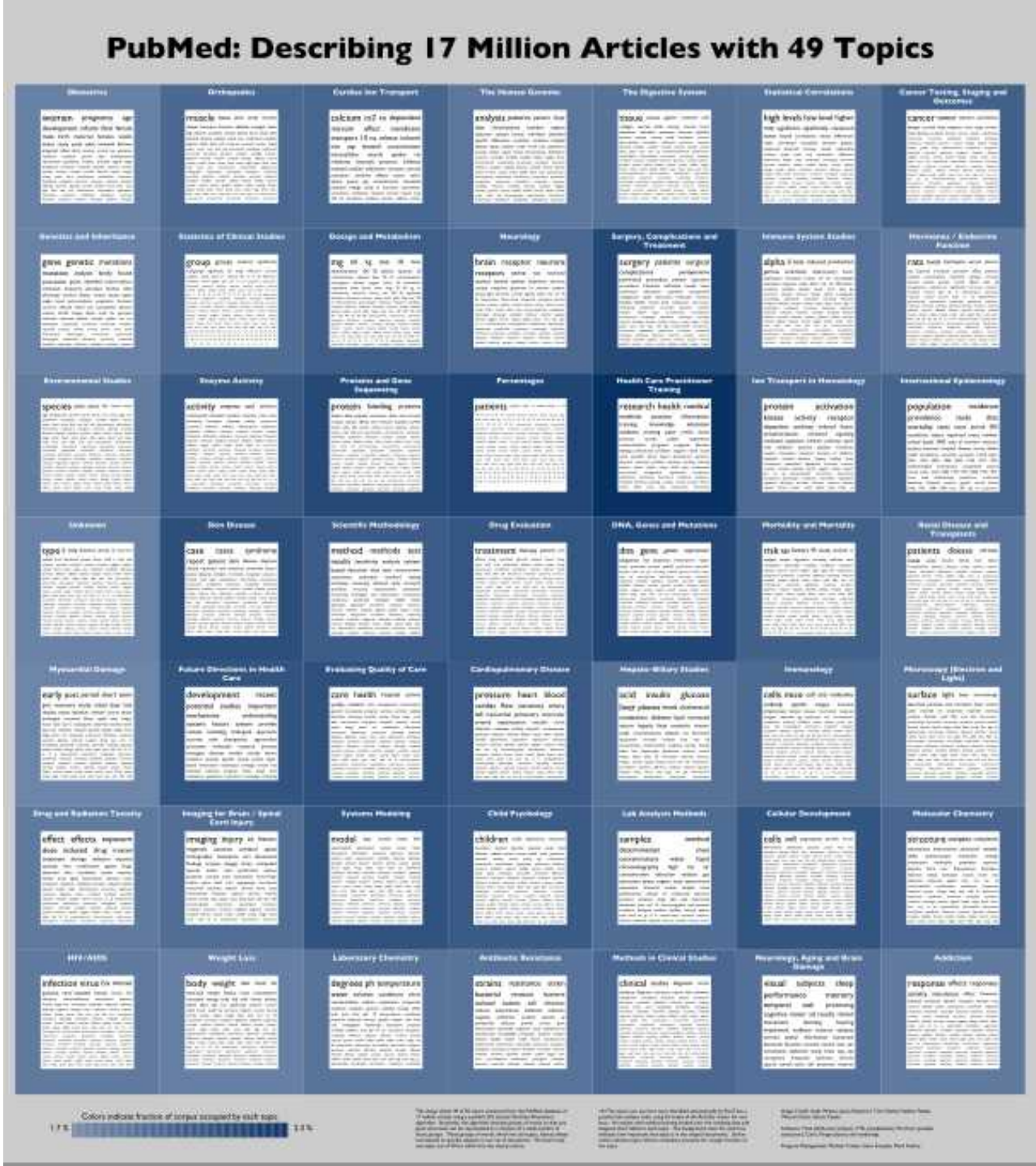
Principle 2: Summarize at Multiple Scales



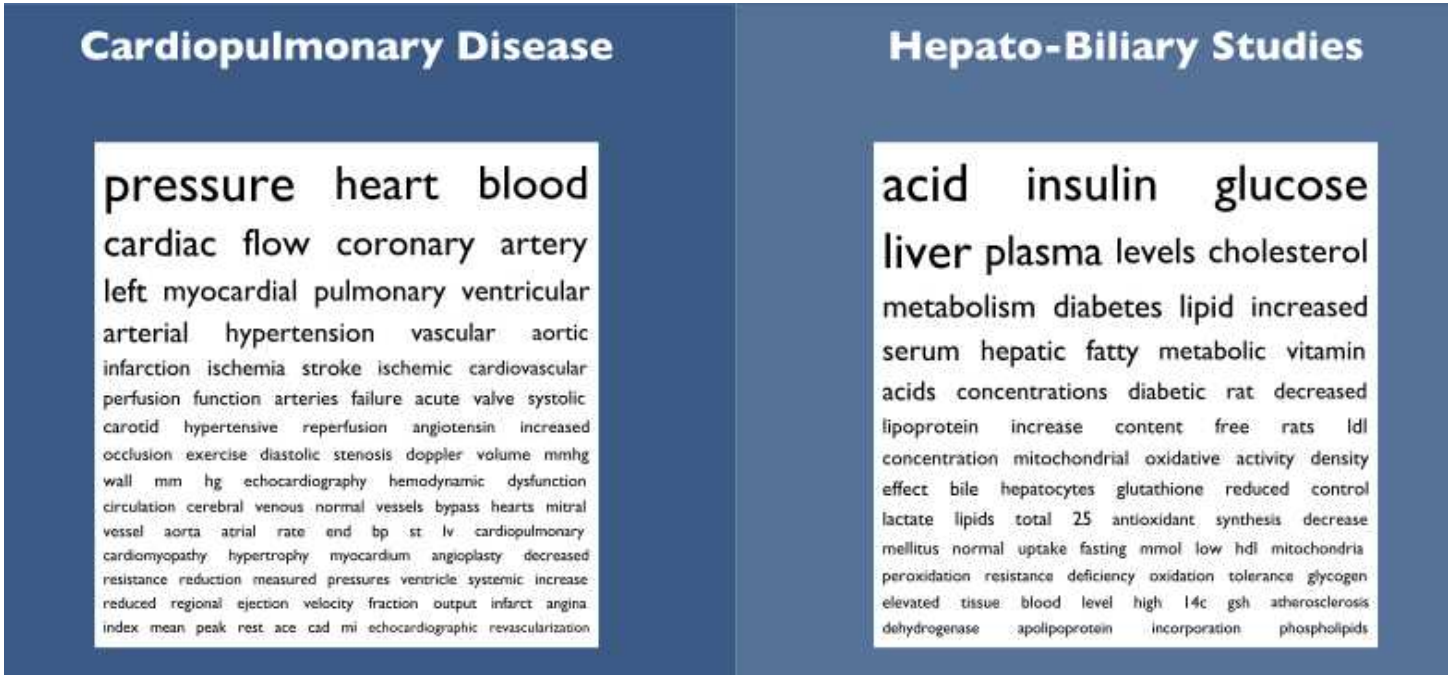
Sandia National Laboratories

- Summarize large collections of documents by talking about the topics

- Darker squares = more “important” topics

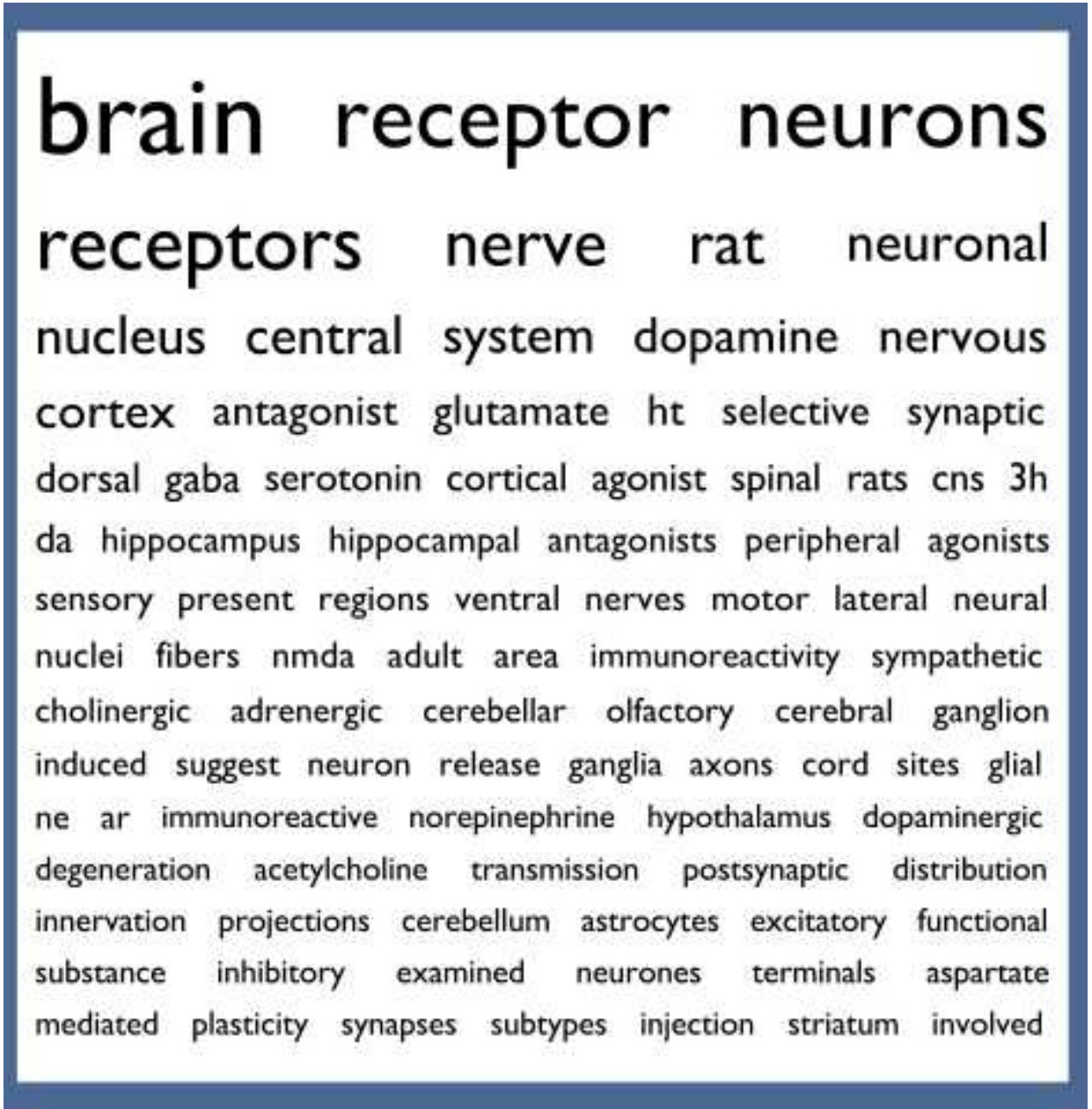


Topical Clustering: Example Topics



- Topic labels come from a subject-matter expert
- Words in each cloud sized by importance in topic

Neurology



Sandia National Laboratories

Principle 3: “Compared to what?”

Sandia National Laboratories

Principle 4: Familiar Metaphors

Sandia National Laboratories

Sandia National Laboratories

Summary

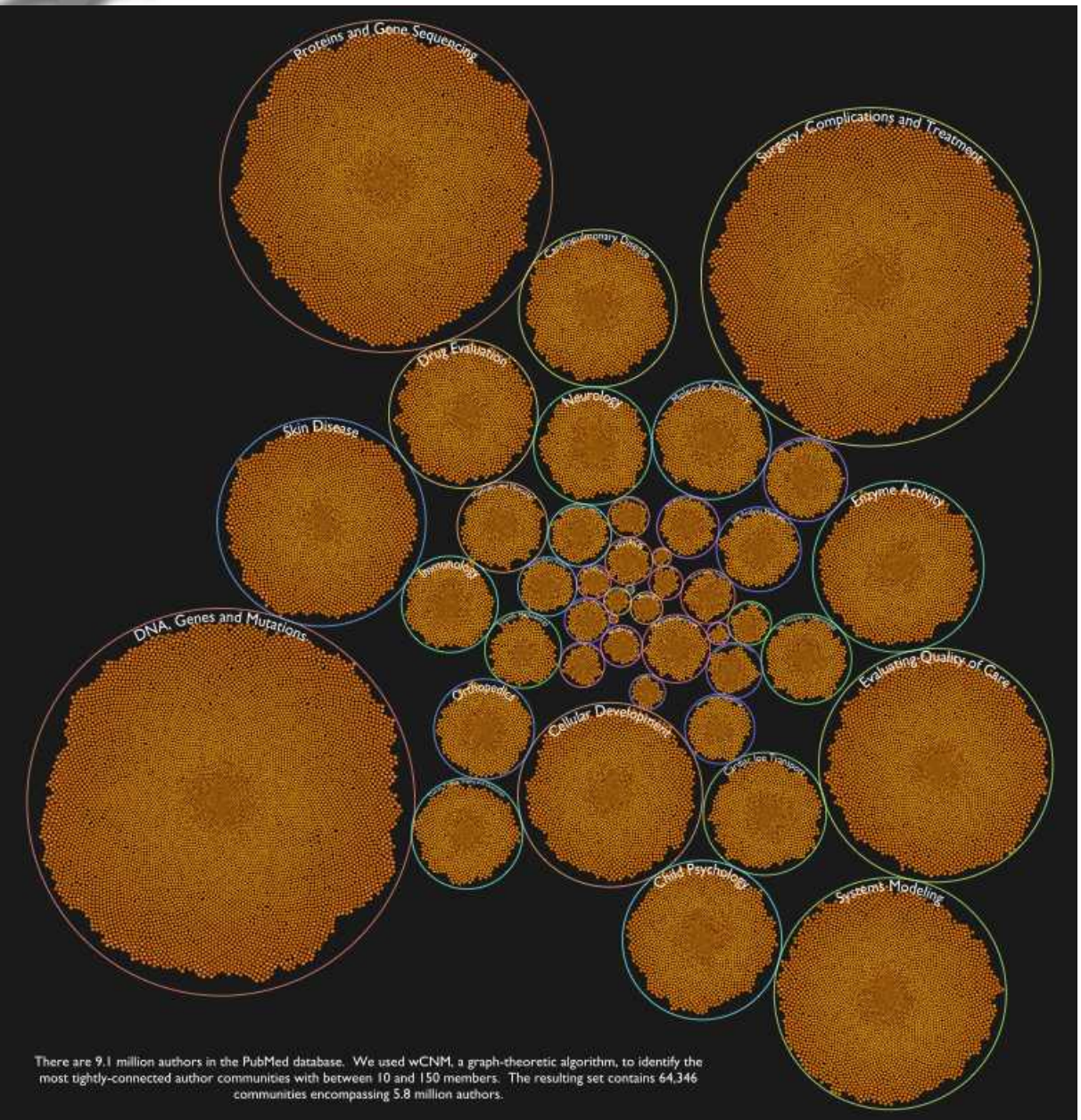
- Abstraction and metaphor are necessary for large data but carry cognitive load

- Principles for Sensemaking Tools
  - Analyze While Computing
  - Summarize at Multiple Scales
  - Use Familiar Metaphors
  - “Compared To What?”

- Tie together Analyst, Toolsmith and Researcher

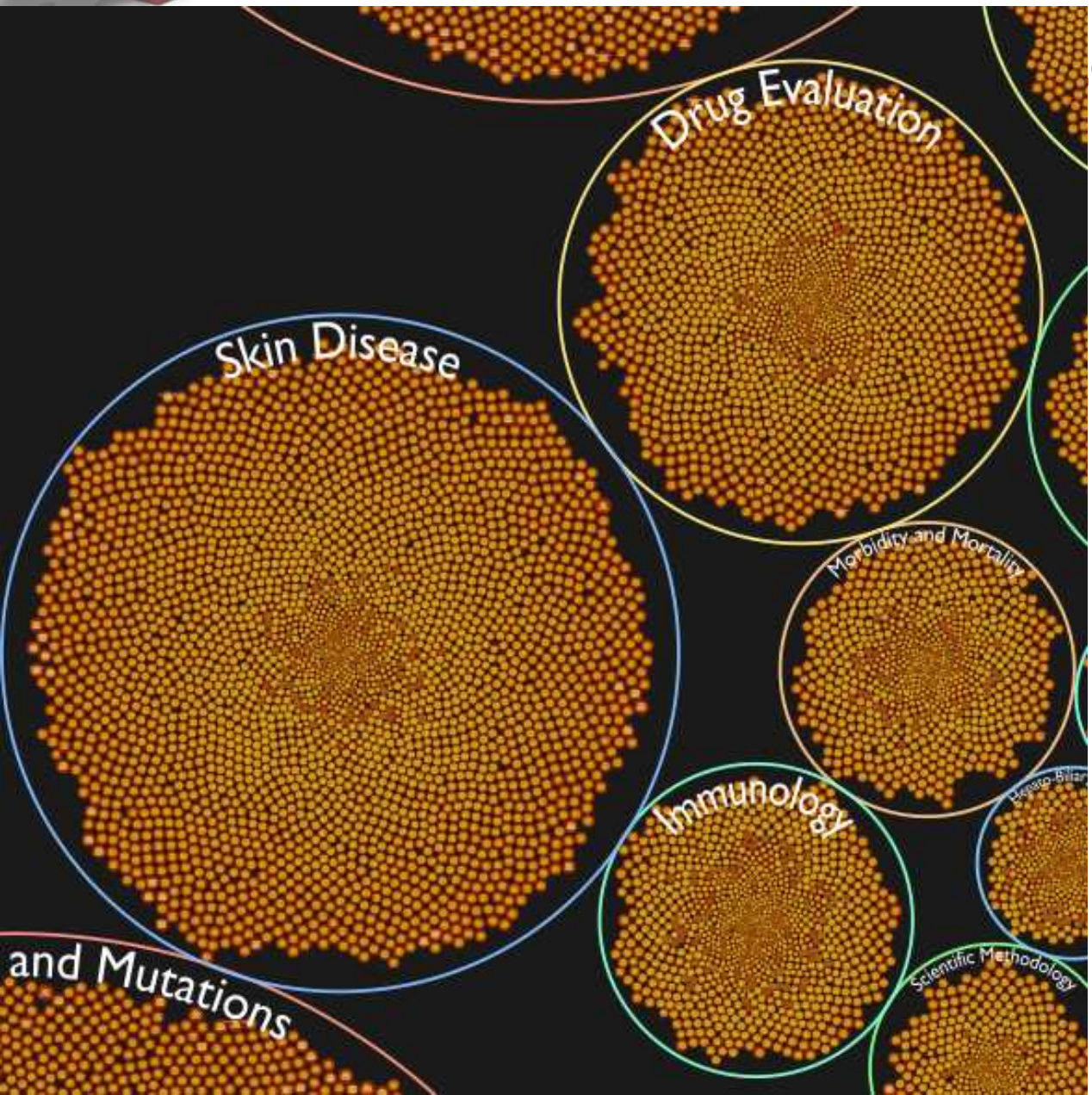
Backup Slides

Sandia National Laboratories



- Each blob is a topic area
- Each tiny orange dot is a single community
- 44 topics, 64,314 communities
  - 10-150 authors in each community

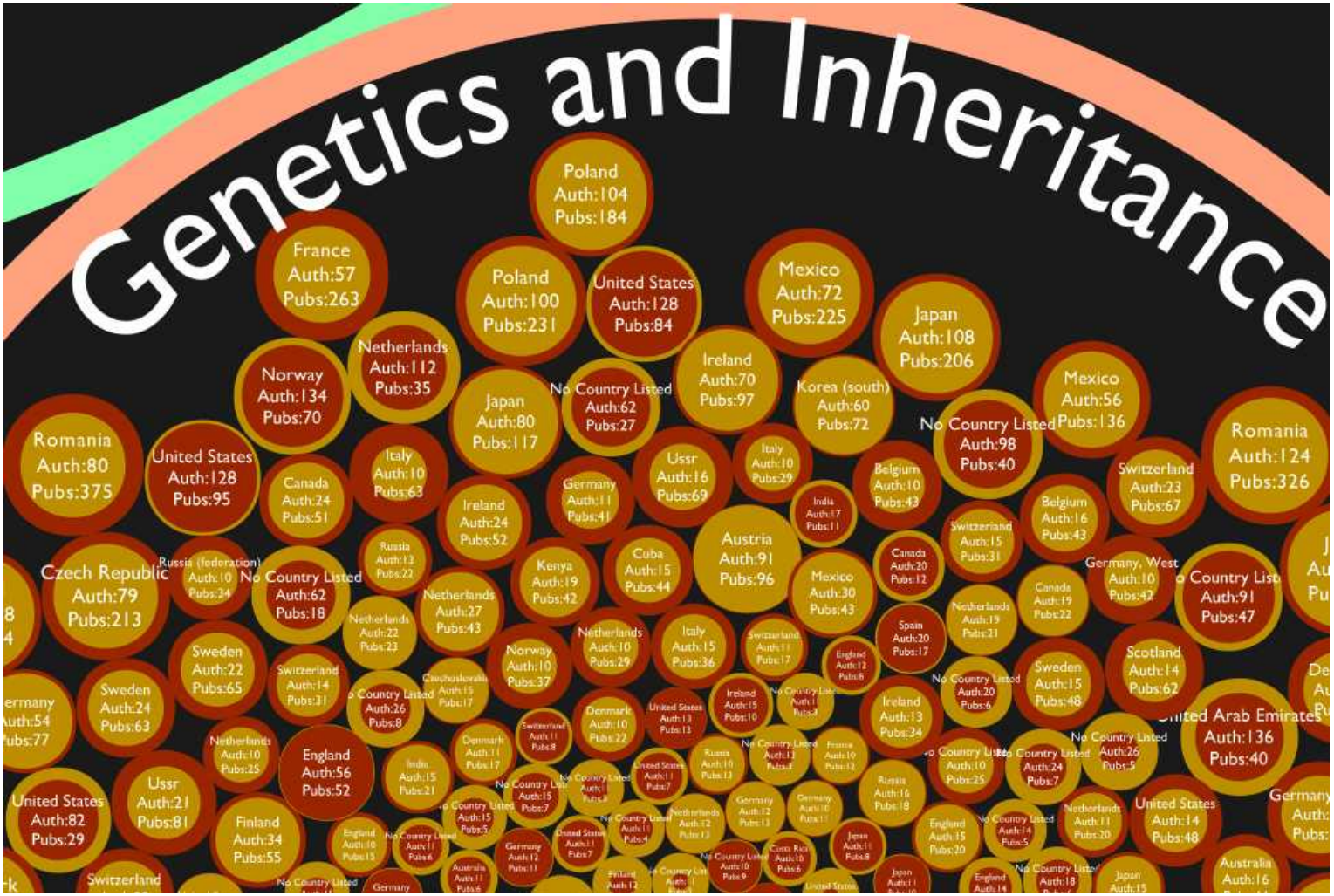
Sandia National Laboratories



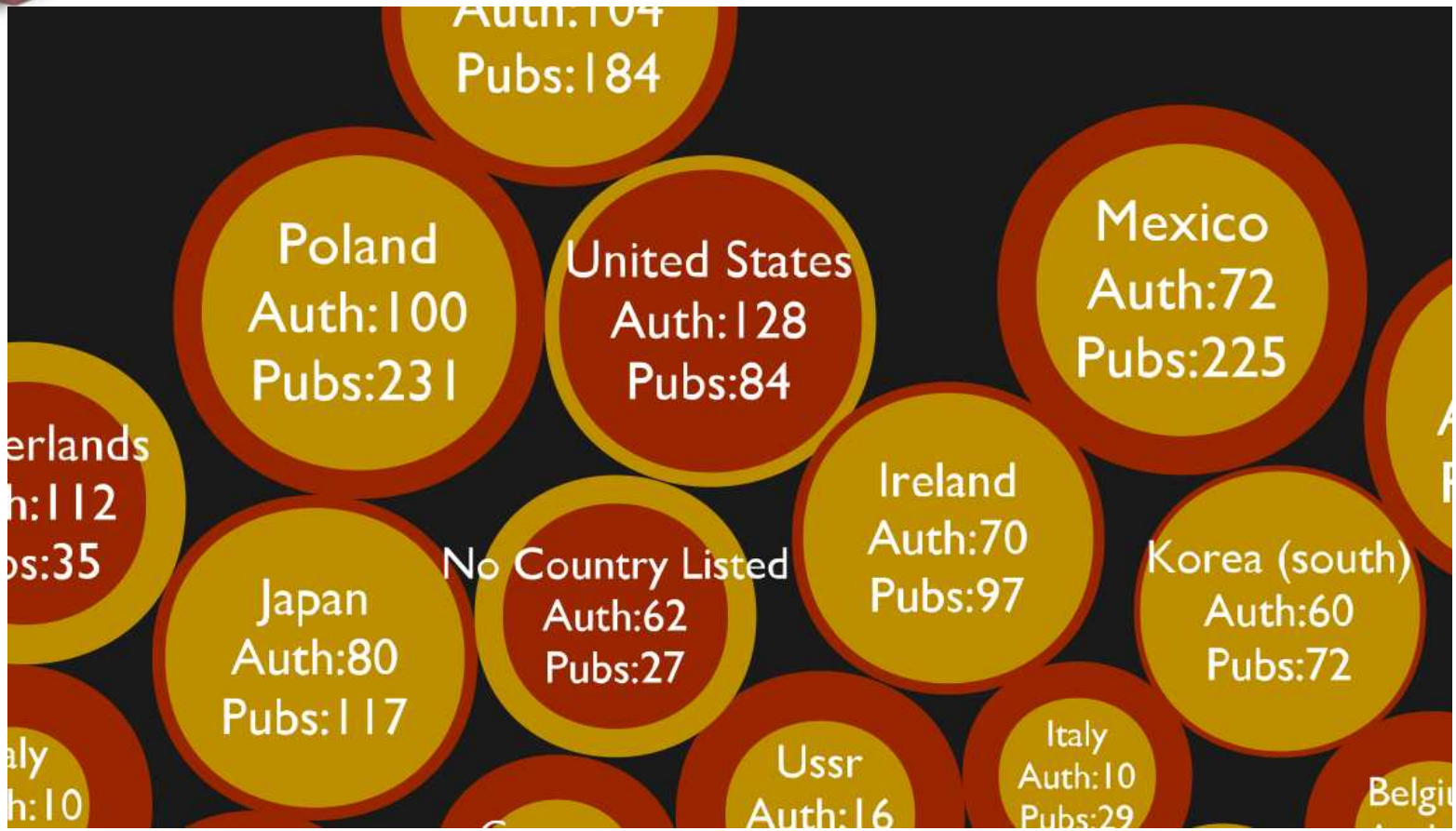
oom 1

- Topic areas contain between 5 and 9800 communities
- Each community is sized according to its author/article count

Sandia National Laboratories



- Finer-grained topics (and diagrams) are possible
- Hierarchical layout possible with advanced clustering algorithms



- Community glyphs indicate relative proportion of authors to articles
- Red = articles, yellow = authors
- Red outside yellow means more articles than authors (and vice versa)
- Labels show country where most papers are published

Sandia National Laboratories