

Okay, Now What?

Challenges in Large Data Analysis

July 26, 2011

Andy Wilson

Scalable Analysis and Visualization

Sandia National Laboratories



Analysis Is...



S. HILARI
OPERA

OPERA
LE. NIK MAGN
ET
MAXIMI

S. OPTATI AFR
EPISCOPI
DE
SCHISM. DONAT

LIBRI VII.

S. PROSPER
OPERA

TERTULLIANI
OPERA

INCOGNITUS
IN
PSALMOS



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.

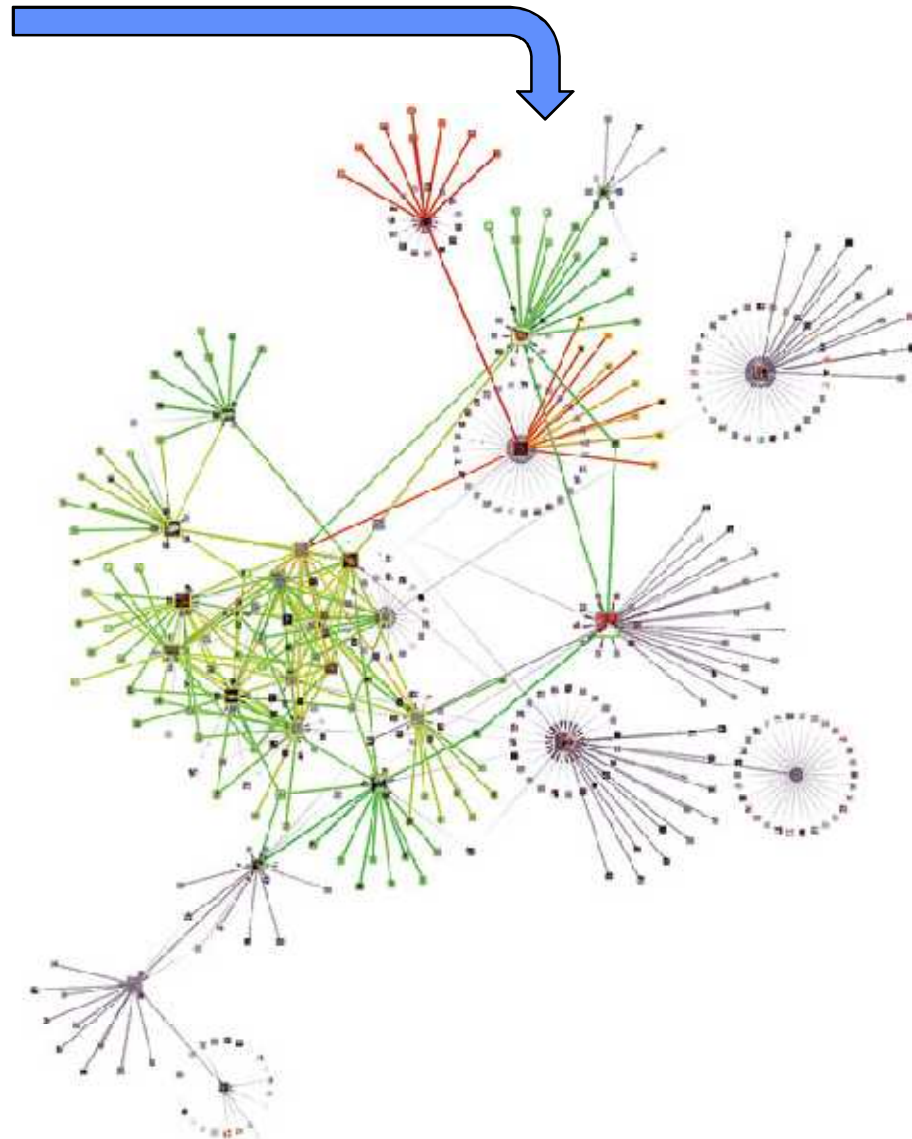
¹¹ Associate Professor, Department of Plant Microbiology and Pathology, University of Missouri, Columbia, MO 65211. Current address: Professor, Department Crop Sciences, University of Illinois, Urbana, IL 61801

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.

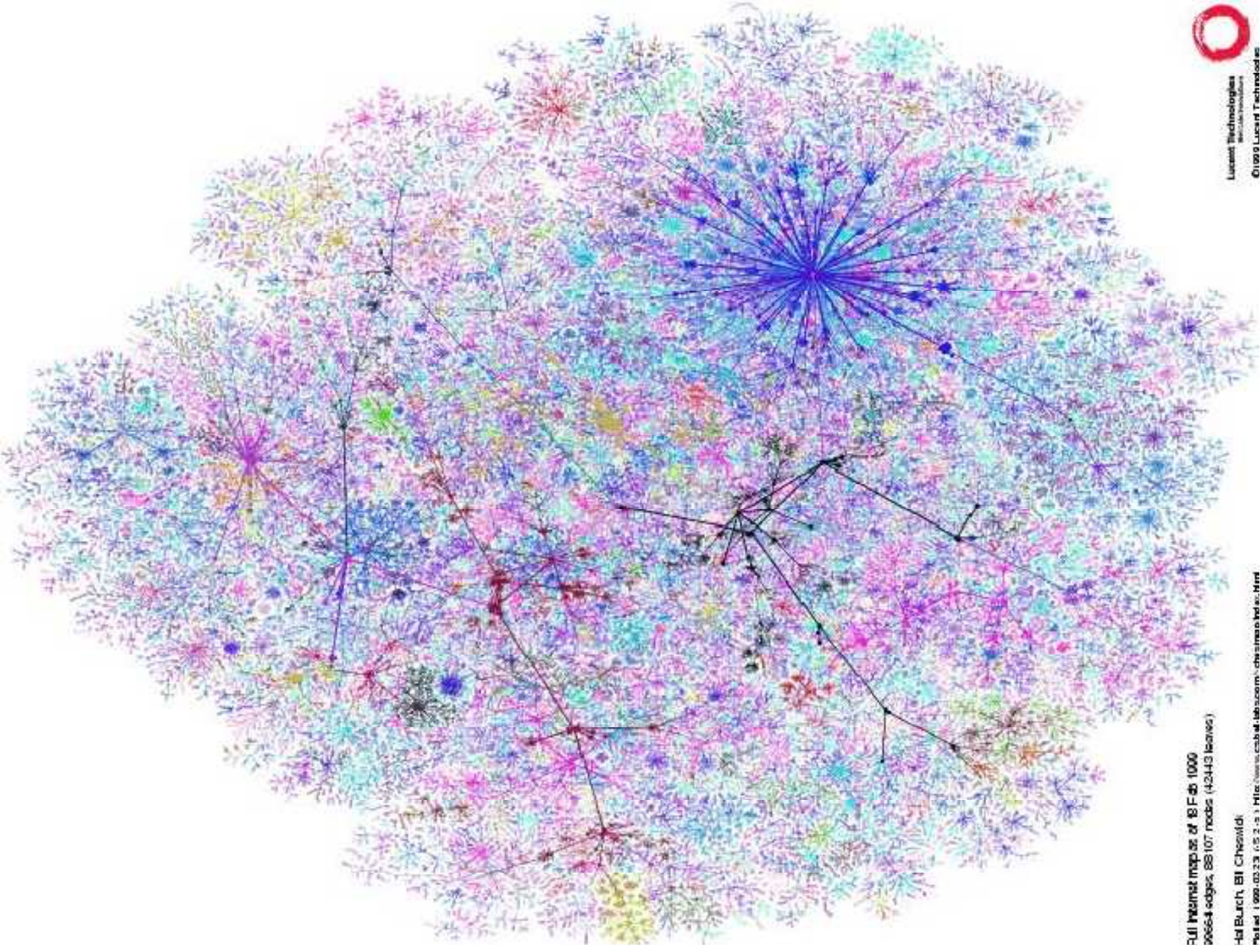
¹⁵ Professor Emeritus, Department of Plant Pathology, University of Nebraska, Lincoln, NE 68583.





So here's what we do...

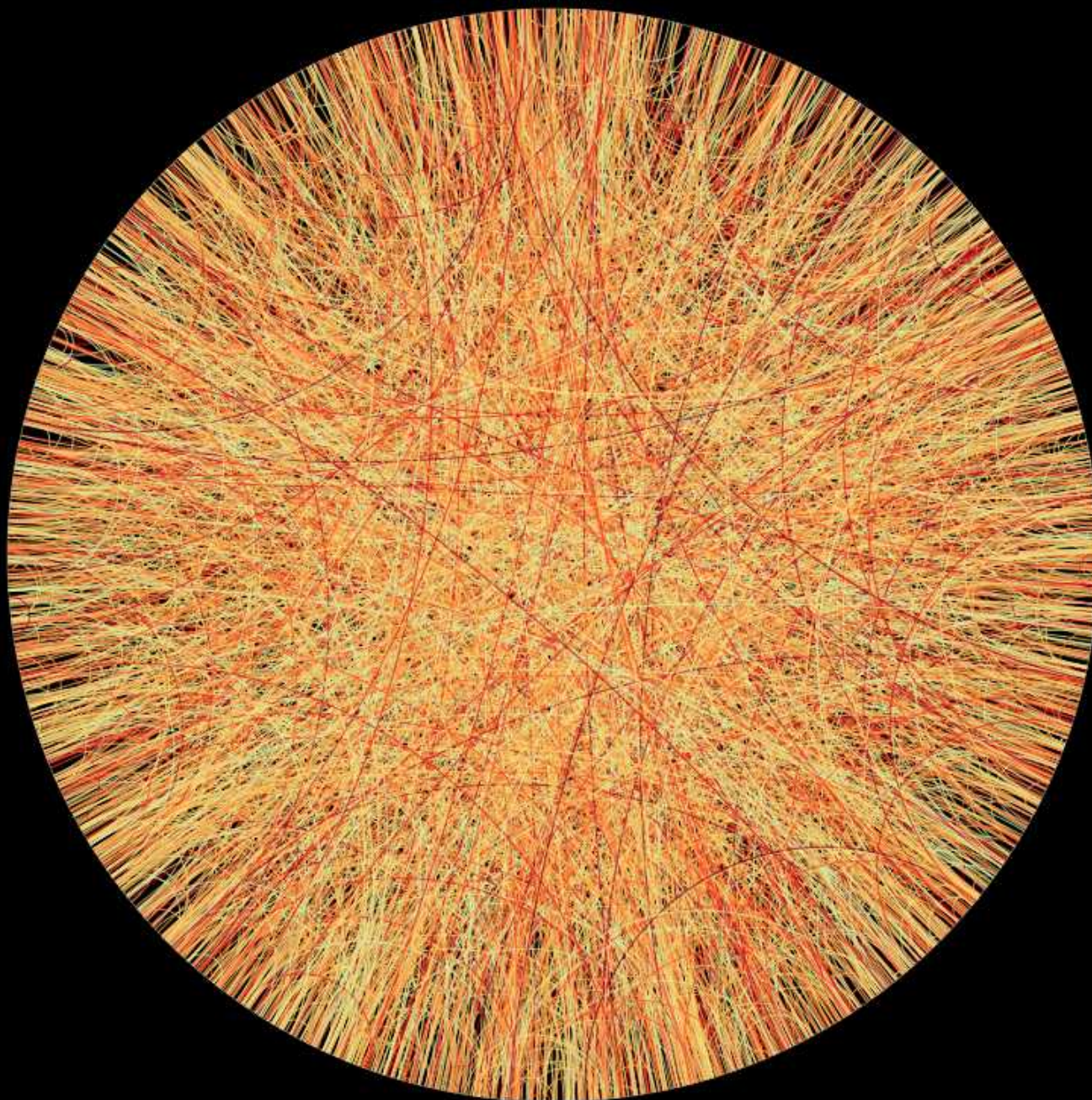
- **If there are two things that computers are good at, it's**
 - **Handling large amounts of data and**
 - **Converting data from one representation to another**
- **Use computers to create a (hopefully manageable) amount of information in usable form. Abstract away the minute detail.**
- **Except...**



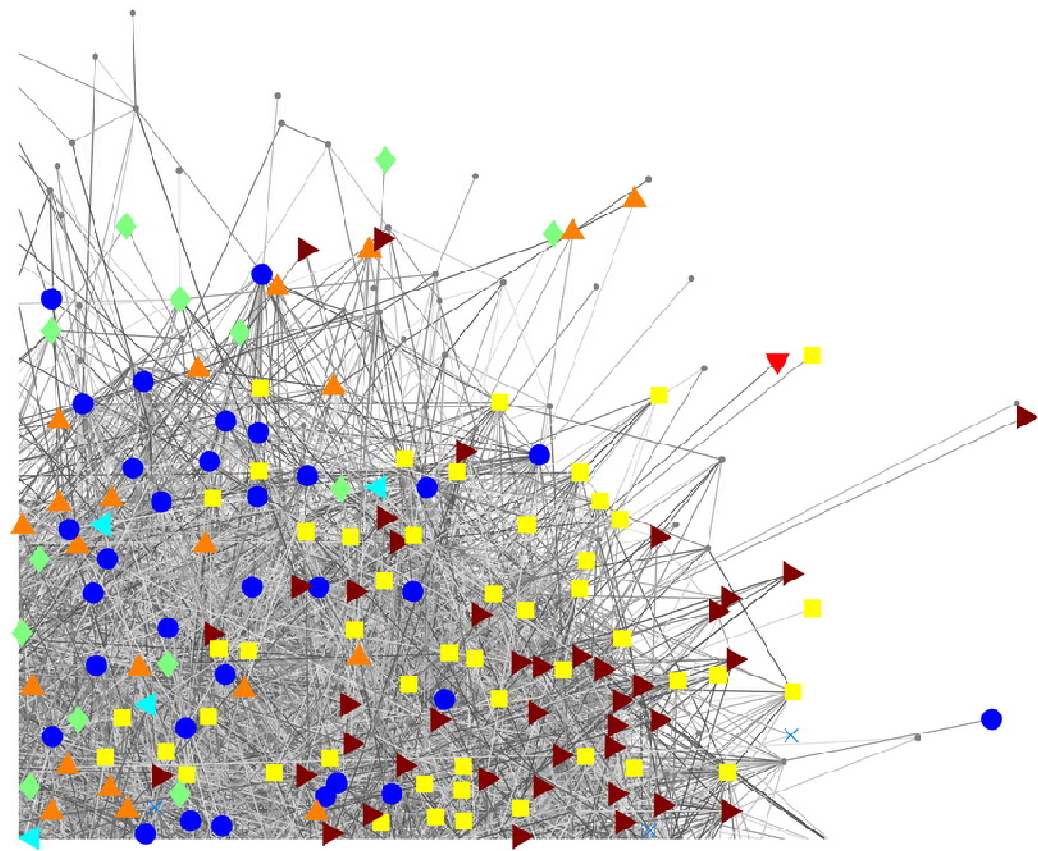
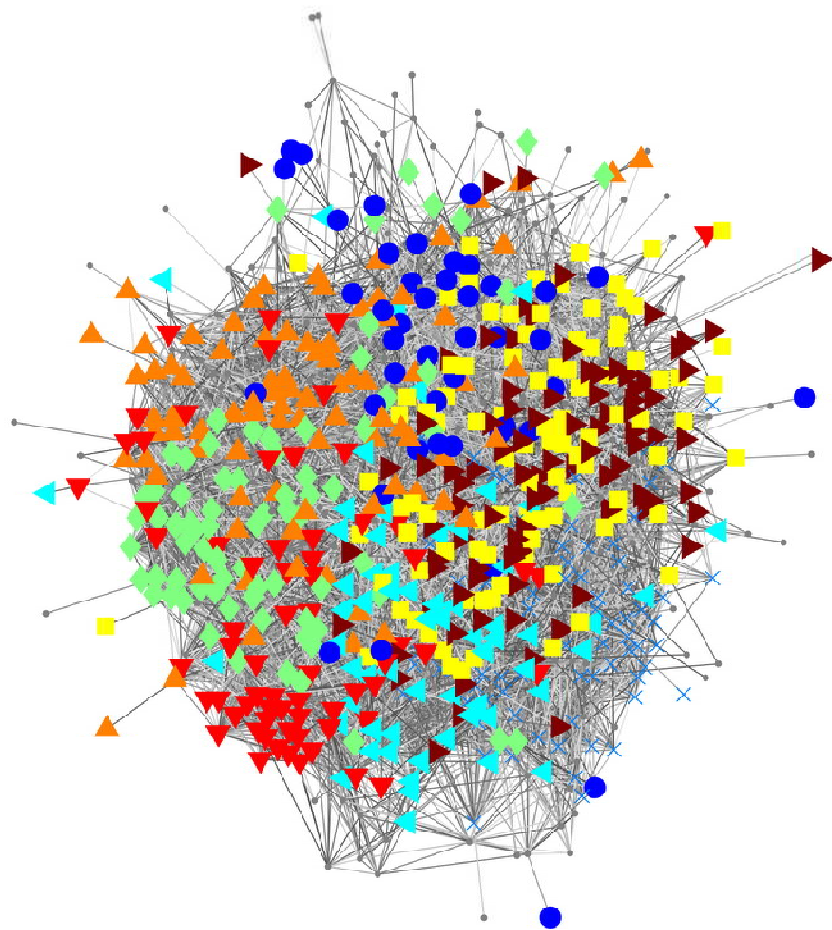
Full Internet map size of 18 Feb 1000
20664 edges, 89 107 nodes (42443 leaves.)

Ed Burch, BII Chesham

Posted: 1999-03-23 (v5.2.1) <http://www.csb.itd.ac.uk/~burch/imap0100.html>



Sandia
National
Laboratories



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



MY THESIS

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



Principle 1: Analyze While Computing

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





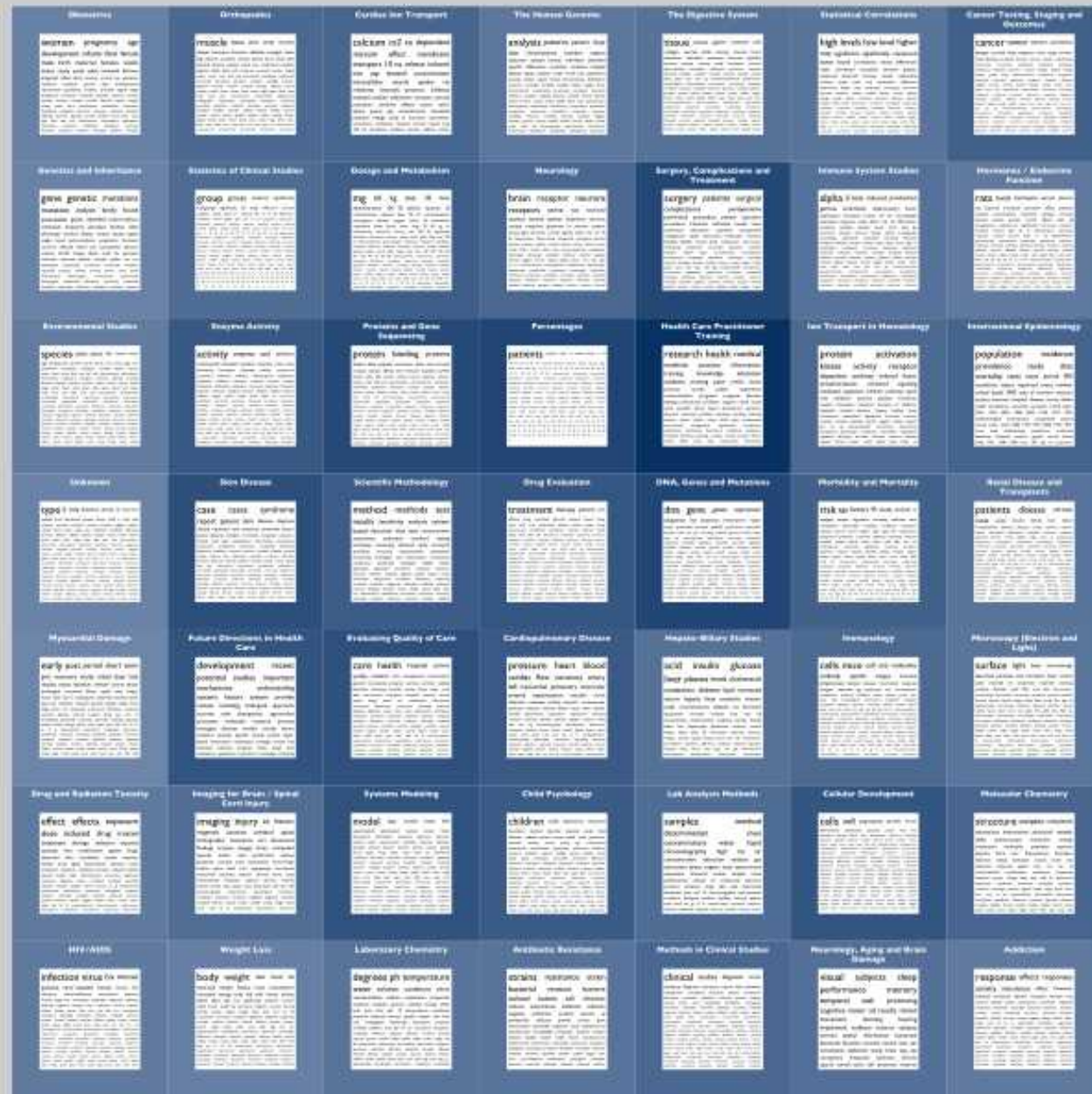
Principle 2: Summarize at Multiple Scales





- Summarize large collections of documents by talking about the topics
- Darker squares = more “important” topics

PubMed: Describing 17 Million Articles with 49 Topics



Colors indicate fraction of topics assigned by each topic

This figure shows the 49 topics generated by the PubMed dataset. The topics are ranked by the fraction of articles assigned to each topic. The topics are ranked by the fraction of articles assigned to each topic. The topics are ranked by the fraction of articles assigned to each topic.

The topics are ranked by the fraction of articles assigned to each topic. The topics are ranked by the fraction of articles assigned to each topic. The topics are ranked by the fraction of articles assigned to each topic.

The topics are ranked by the fraction of articles assigned to each topic. The topics are ranked by the fraction of articles assigned to each topic. The topics are ranked by the fraction of articles assigned to each topic.

Topical Clustering: Example Topics

Cardiopulmonary Disease

pressure heart blood
cardiac flow coronary artery
left myocardial pulmonary ventricular
arterial hypertension vascular aortic
infarction ischemia stroke ischemic cardiovascular
perfusion function arteries failure acute valve systolic
carotid hypertensive reperfusion angiotensin increased
occlusion exercise diastolic stenosis doppler volume mmhg
wall mm hg echocardiography hemodynamic dysfunction
circulation cerebral venous normal vessels bypass hearts mitral
vessel aorta atrial rate end bp st lv cardiopulmonary
cardiomyopathy hypertrophy myocardium angioplasty decreased
resistance reduction measured pressures ventricle systemic increase
reduced regional ejection velocity fraction output infarct angina
index mean peak rest ace cad mi echocardiographic revascularization

Hepato-Biliary Studies

acid insulin glucose
liver plasma levels cholesterol
metabolism diabetes lipid increased
serum hepatic fatty metabolic vitamin
acids concentrations diabetic rat decreased
lipoprotein increase content free rats ldl
concentration mitochondrial oxidative activity density
effect bile hepatocytes glutathione reduced control
lactate lipids total 25 antioxidant synthesis decrease
mellitus normal uptake fasting mmol low hdl mitochondria
peroxidation resistance deficiency oxidation tolerance glycogen
elevated tissue blood level high l4c gsh atherosclerosis
dehydrogenase apolipoprotein incorporation phospholipids

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



Neurology

brain receptor neurons

receptors nerve rat neuronal

nucleus central system dopamine nervous

cortex antagonist glutamate ht selective synaptic

dorsal gaba serotonin cortical agonist spinal rats cns 3h

da hippocampus hippocampal antagonists peripheral agonists

sensory present regions ventral nerves motor lateral neural

nuclei fibers nmda adult area immunoreactivity sympathetic

cholinergic adrenergic cerebellar olfactory cerebral ganglion

induced suggest neuron release ganglia axons cord sites glial

ne ar immunoreactive norepinephrine hypothalamus dopaminergic

degeneration acetylcholine transmission postsynaptic distribution

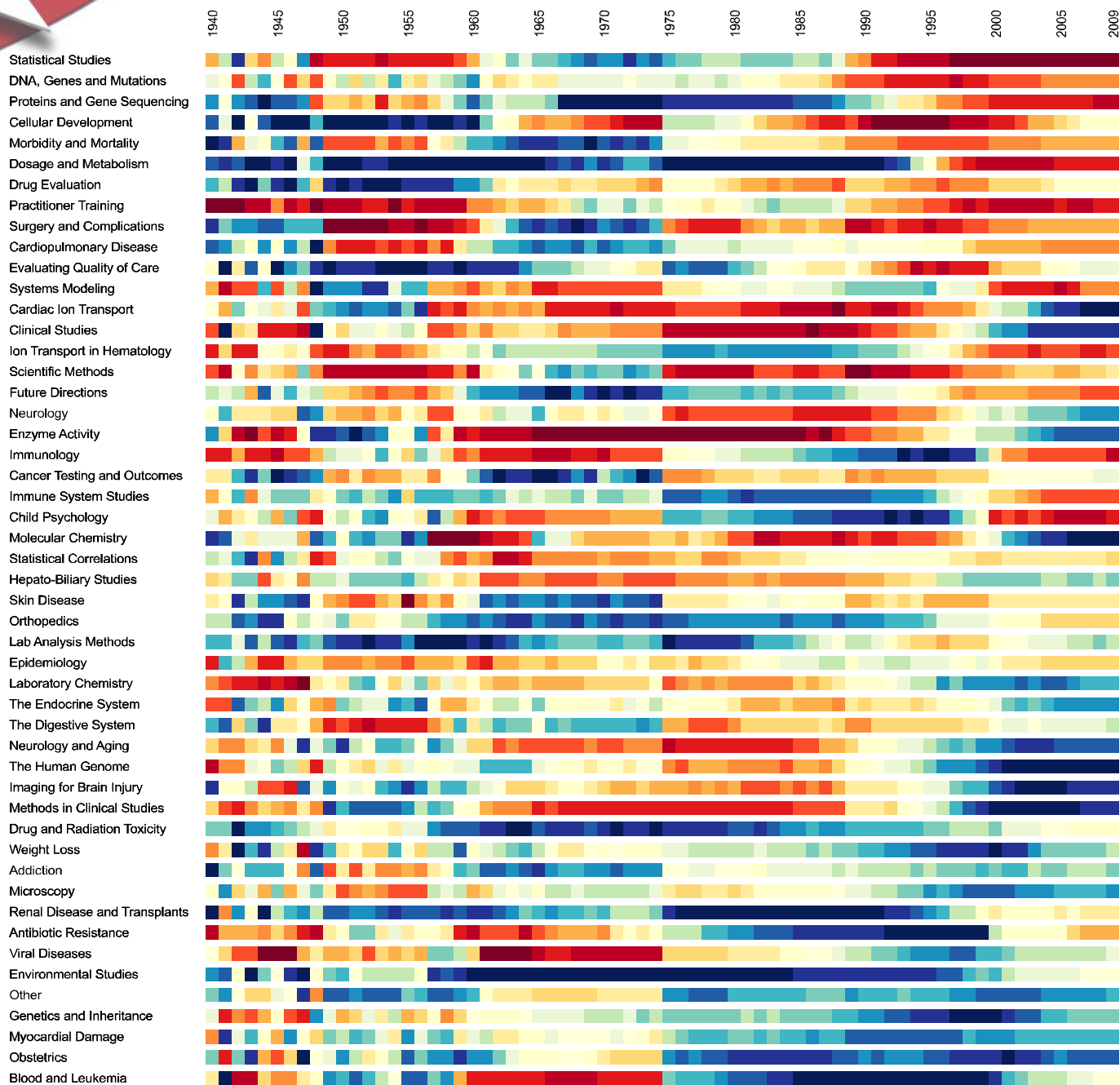
innervation projections cerebellum astrocytes excitatory functional

substance inhibitory examined neurones terminals aspartate

mediated plasticity synapses subtypes injection striatum involved



Principle 3: “Compared to what?”





Principle 4: Familiar Metaphors





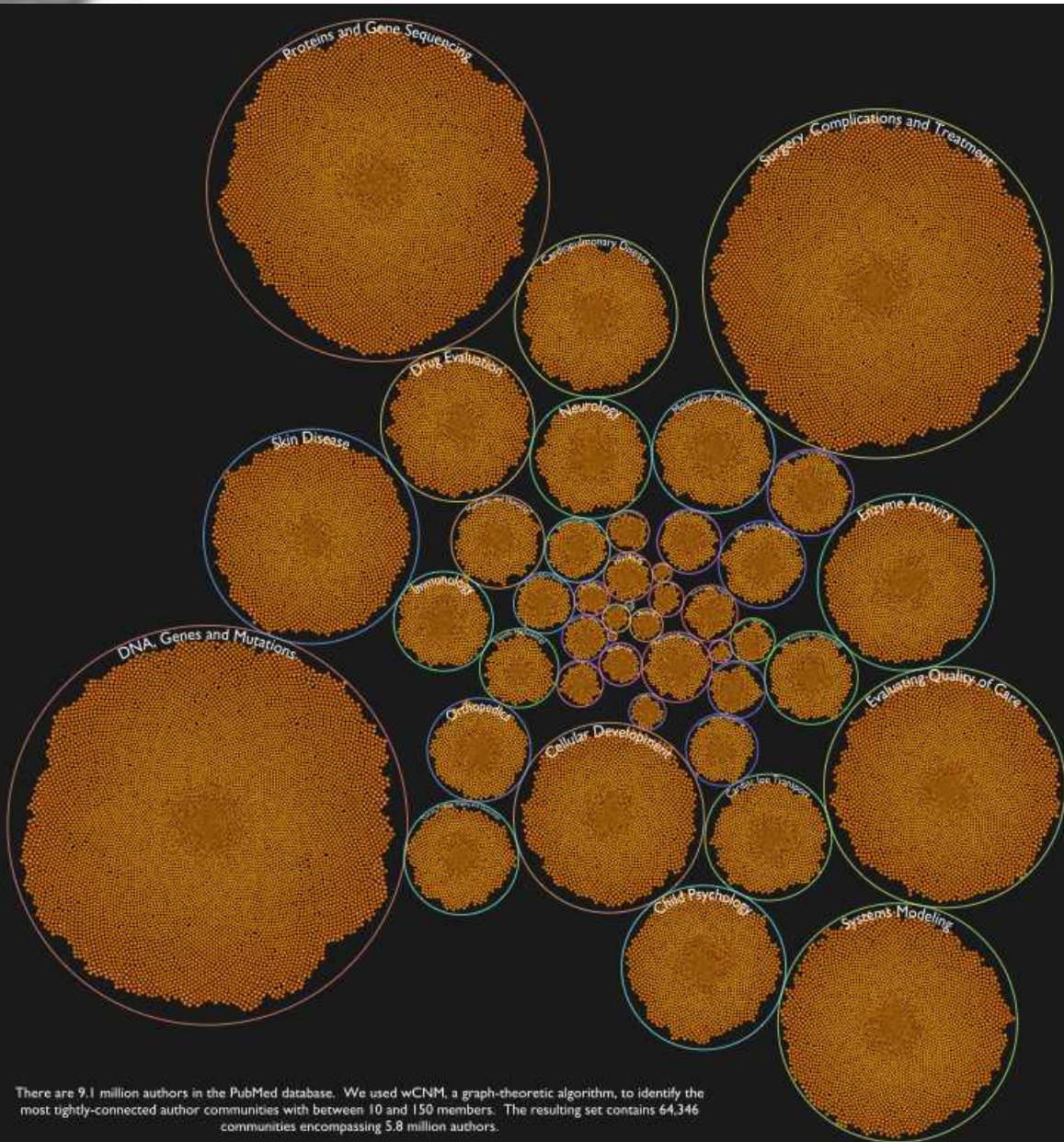


Summary

- **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**



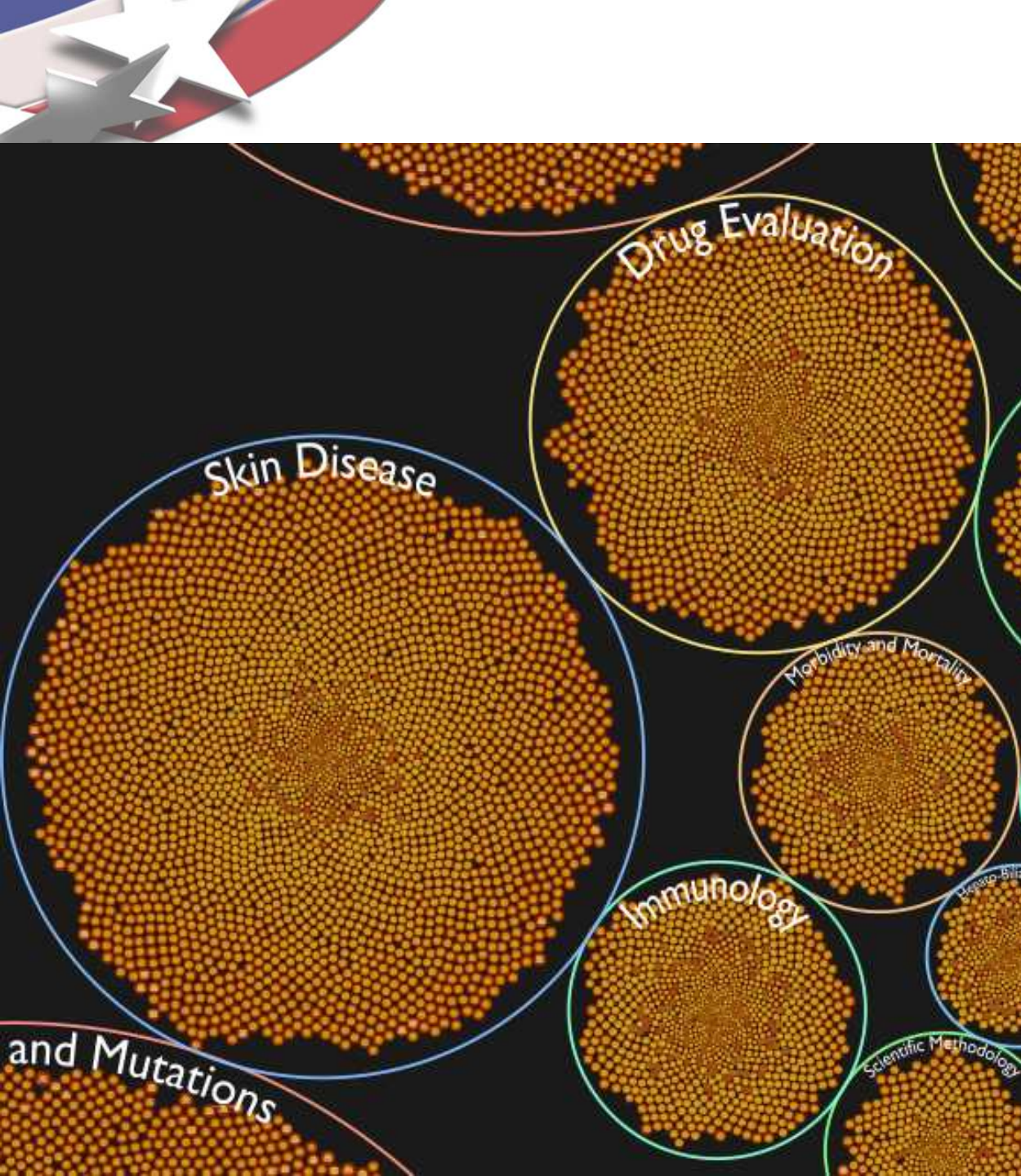
Backup Slides



is Each blob is a topic area

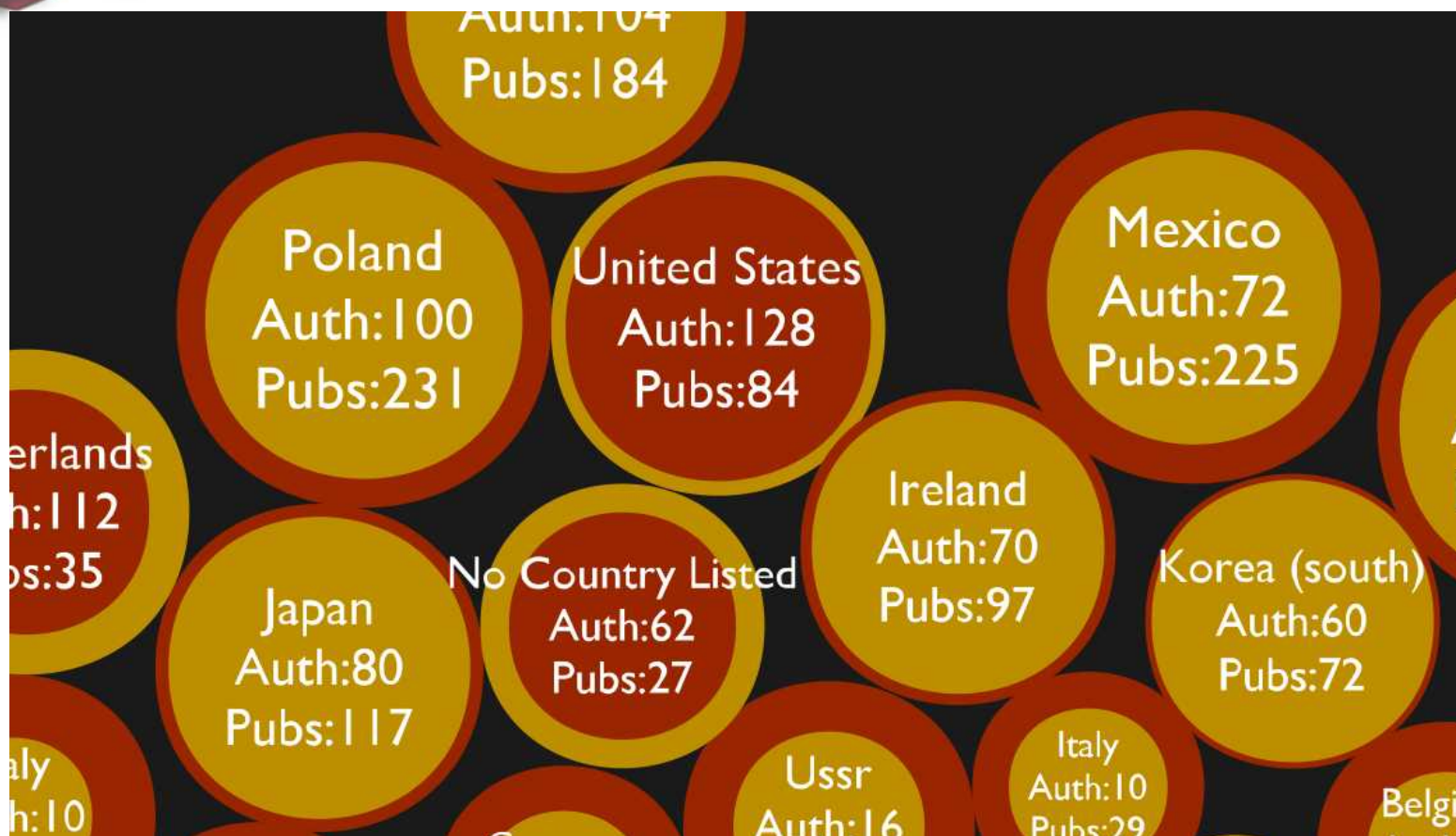
- Each tiny orange dot is a single community
- 44 topics, 64,314 communities
 - 10-150 authors in each community

There are 9.1 million authors in the PubMed database. We used wCNM, a graph-theoretic algorithm, to identify the most tightly-connected author communities with between 10 and 150 members. The resulting set contains 64,346 communities encompassing 5.8 million authors.



Room 1

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



- 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

