Transforming Science: Cancer Control on caBIG™

PopSciGrid: Using cyberinfrastructure to enable data harmonization, collaboration, and advanced computation of nationally representative behavioral, demographic, and economic data

SONIC

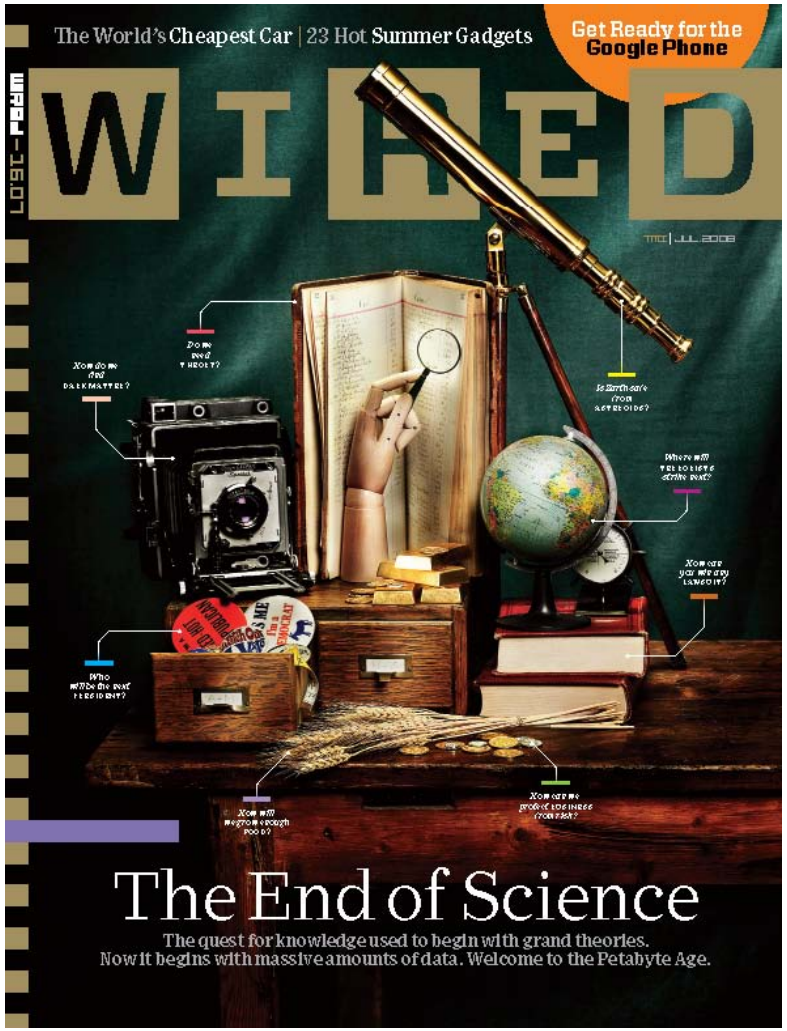
PopSciGrid Team

- Northwestern University, SONIC
  - Noshir Contractor, PhD; York Yao, MS; Yun Huang, PhD
- Health Communication and Informatics Research (NCI/ DCCPS)
  - Bradford Hesse, PhD; Abdul Shaikh, PhD; Glen Morgan, PhD; Richard P. Moser, PhD; Alison Pilsner, MPH; Eric Augustson, PhD
- Booz Allen Hamilton
  - Paul K. Courtney, MS
- Cancer Institute of New Jersey
  - Peter A. Schad, Ph.D
- Science Applications International Corporation (SAIC)
  - Mary Cooper

Outline

- Population sciences research
- caBIG™ as a research platform
- PopSciGrid: a grid for population sciences
- Knowledge discovery on the Grid

The End of Science



"The Petabyte Age: Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grows, so will the opportunity to find answers to fundamental questions. Because in the era of big data, more isn't just more. More is different. "

- Chris Anderson 06.23.08

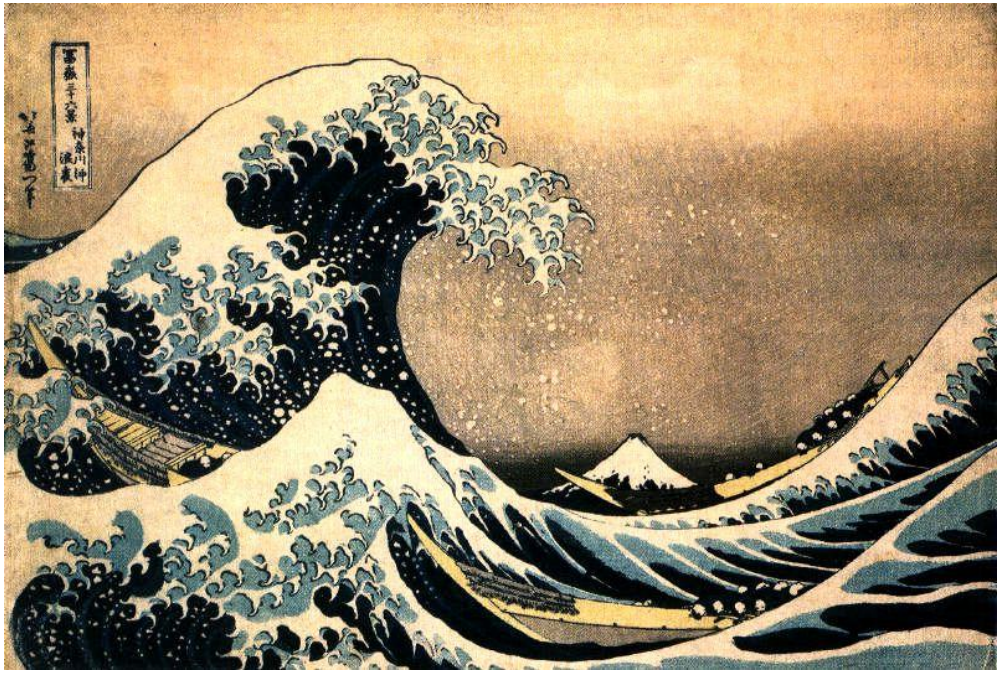
SONIC
Advancing the Science of Networks in Communities

The Current World of Tobacco Research

Source: Peter Schad

- Islands of tobacco-related documents, datasets, analytic tools, and research communities

Information Tsunami



Source: Peter Schad

- Overwhelming volume of data
- Multitude of sources
- Different coding schemes

Informatics Tower of Babel

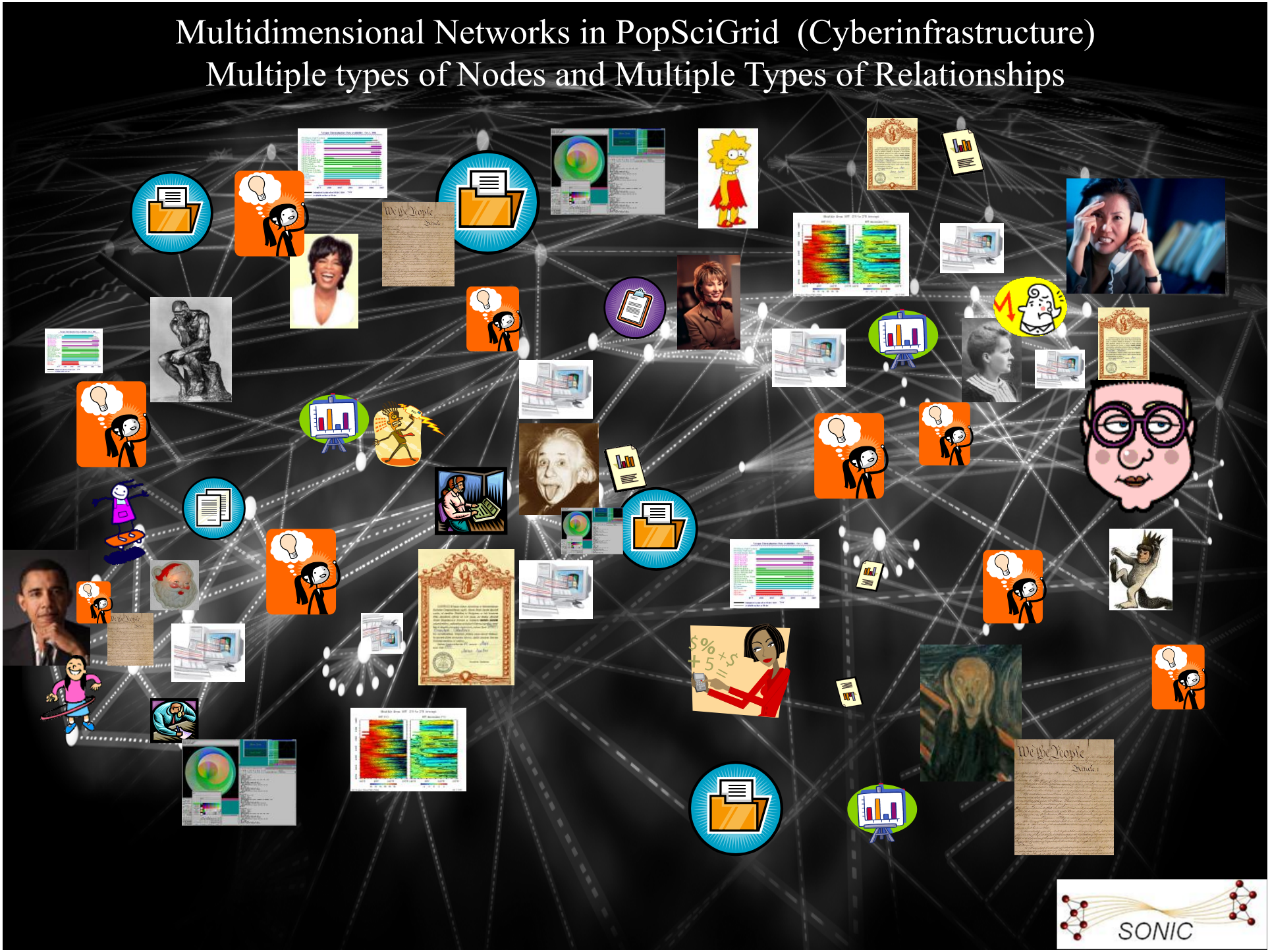
Source: Peter Schad

- Each tobacco research community speaks its own scientific "dialect"
- Integration critical to achieve promise of combating tobacco burden on cancer

Population Sciences Research

- Data for population sciences
  - Data from multiple national surveys/interviews
  - Islands of documents, datasets, analytical tools, and research communities
- Obstacles of data collaboration
  - Difficult to access data
  - Various behavioral measures
  - Different measurement scales
  - Survey-based statistical tools
  - Visualization analytics

Multidimensional Networks in PopSciGrid (Cyberinfrastructure)
Multiple types of Nodes and Multiple Types of Relationships



What is PopSciGrid?

- Proof of concept for cyberinfrastructure in population health and cancer control
- Use state-of-the-science technology to link data, researchers, and resources

Can we transform science?

PopSciGrid Objectives

- Improve access to and usability of population science data
- Real-time integration and analysis of multiple types of data (e.g., population health, economic, geo-spatial)
- Decrease the time it takes to translate research into practice and policy at local and state levels

PopSciGrid Prototype on caBIG™

- Implement sample services on the Grid
  - NHIS 2000-2005, HINTS 2003 & 2005, and tobacco tax 2000-2007 data services
  - Basic statistics, categorical analysis, and prevalence analysis
  - Visualization by region
- Demonstrate the power of the Grid
  - Publish population science data
  - Analyze population data from multiple sources
  - Visualize data on the Grid

Data sets: NHIS, HINTS & Tax

- National Health Interview Survey (NHIS):
  - The principal source of information on the health of the U.S. population
  - 1957-2007: 50 year study
- Health Information National Trends Survey (HINTS)
  - Nationally representative data about the American public's use of cancer-related information
- State-level tobacco tax data 2000-2007
  - Orzechowski & Walker, 2007

Data Sharing through Files

National Health Interview Survey (NHIS)

Celebrating the First 50 years:
1957 - 2007

2005 Data Release

- Readme File
- Notices for Data Users
- Family file
  - Variable summary
  - Variable layout
  - Variable frequencies
  - ASCII data
  - Sample SAS statements
  - Sample SPSS statements
  - Sample Stata statements
- Household file
  - Variable summary
  - Variable layout
  - Variable frequencies
  - ASCII data
  - Sample SAS statements
  - Sample SPSS statements
  - Sample Stata statements

hints Health Information National Trends Survey

How Americans find and use cancer information

View HINTS Findings
Search HINTS Questions
HINTS Briefs

Conduct Research
Dataset
Survey Instrument
Research Using HINTS Data

Learn About HINTS
About HINTS
Contact Us

Need Help?
Contact us by phone, Web, and e-mail
1-800-4-CANCER

Home: Public Use Dataset

HINTS 2005 Dataset, updated June 2006

HINTS data are available for public use. The full dataset (n=5586) includes respondents who completed the entire interview (Completes: n=5394) plus those who completed the Health Communication and General Cancer Questions only (Partial Completes: n=192).

- SAS data and supporting documents: ZIP (4.8 MB)
- SPSS data and supporting documents: ZIP (4.6 MB)

HINTS 2003 Dataset, updated June 2006

The full dataset (n=6369) includes respondents who completed the entire interview (Completes: n=6149) plus those who completed the Health Communication and General Cancer Questions only (Partial Completes: n=220).

- SAS data and supporting documents: ZIP (5.5 MB)
- SPSS data and supporting documents: ZIP (5.4 MB)

Complicated Codebooks

NHIS 2005

2005 NATIONAL HEALTH INTERVIEW SURVEY
Sample Adult Cancer
cancerxx : Tobacco
PUBLIC USE
Document Version Date: 13-Jun-06

Question ID: NAE.148_00.000

Instrument Variable Name: CIGEV50
Final Documentation Name: CIGEV50

Have you smoked at least 50 cigars in your entire life?

Universe: ASTATFLG='1' and (AGE GE '018' and AGE
Description: Sample adults 18+ who have ever smoked a cig

Sources:
Recodes:
Keywords: smoked; cigar
Notes:

Smoked at least 50 cigars

1	Yes
2	No
7	Refused
8	Not ascertained
9	Don't know

Question ID: NAE.150_00.000

HINTS 2005

TOBACCO SCREENER

Next are some questions about your use of cigarettes.

TU-01. Have you smoked at least 100 cigarettes in your entire life?

[IF NEEDED: 5 Packs = 100 Cigarettes.]

TU01Smoke100

YES	1	
NO	2	(TU-17)

TU-02. Do you now smoke cigarettes . . .

TU02SmokeNow

every day,	1	
some days, or	2	(TU-4)
not at all?	3	(TU-11)
DK	9	(TU-17)
RF	8	(TU-17)

TU-03. On the average, how many cigarettes do you now smoke a day?

TU03SmokeDayAlways

[IF NEEDED: 1 Pack = 20 Cigarettes.]

[IF LESS THAN ONE A DAY, ENTER 0. IF 76 OR MORE, ENTER 76.]

|__|__|
NUMBER OF CIGARETTES

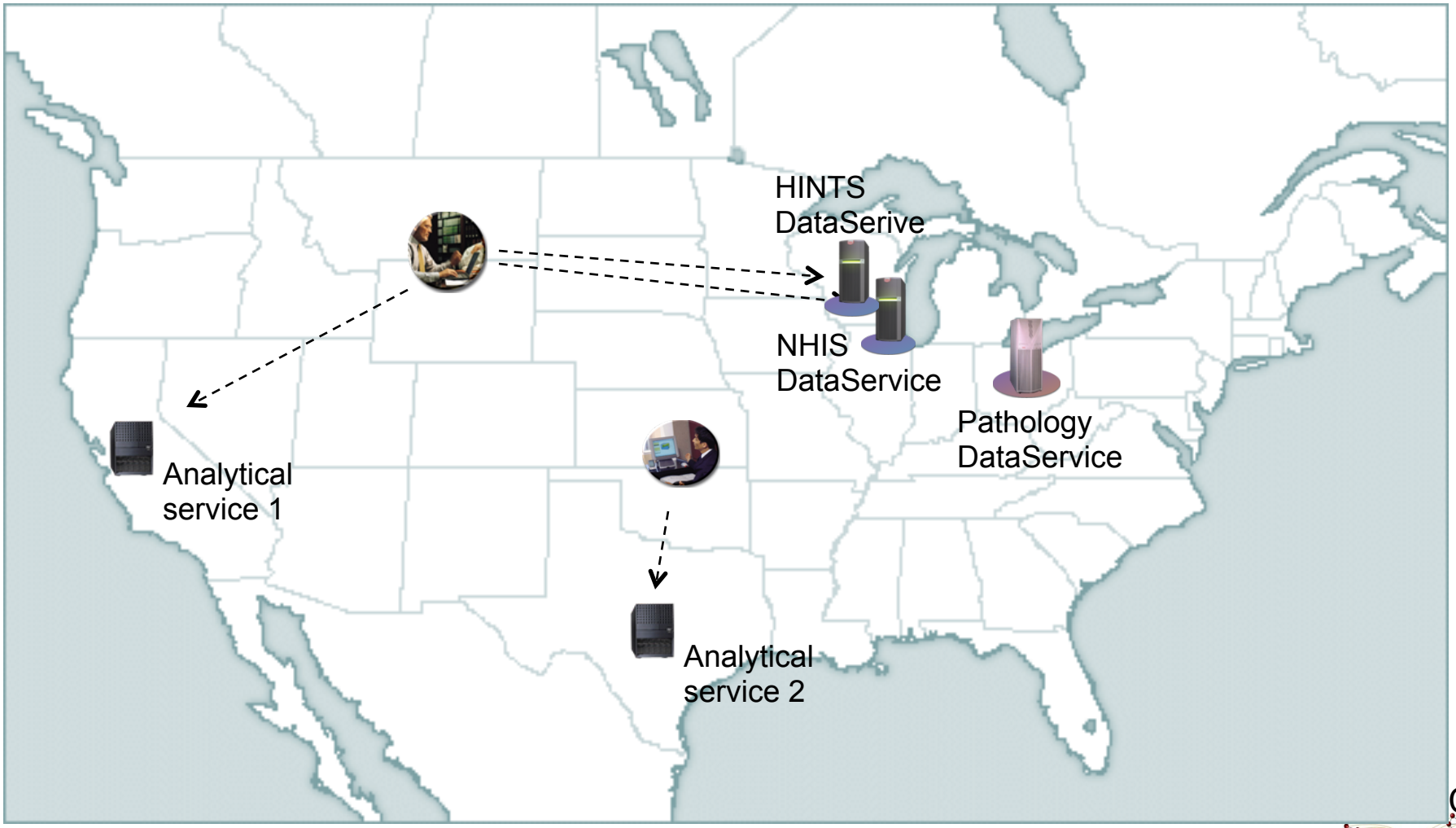
Move to the Grid

- Data collaboration challenge for population science
- Grid enabled services provide a potential solution to move and share data on the Grid.

caBIG™ as a Research Platform

- Sharing resources on the Grid
  - Data services
  - Analytical services
  - Visualization services
- Combining resources on the Grid
  - Integrate data sources
  - Integrate workflows
- Knowledge discovery on the Grid
  - Recommend concepts, datasets, analytical services, users, and service providers based on networks
  - Facilitate collaboration

Synergy among Grid Services



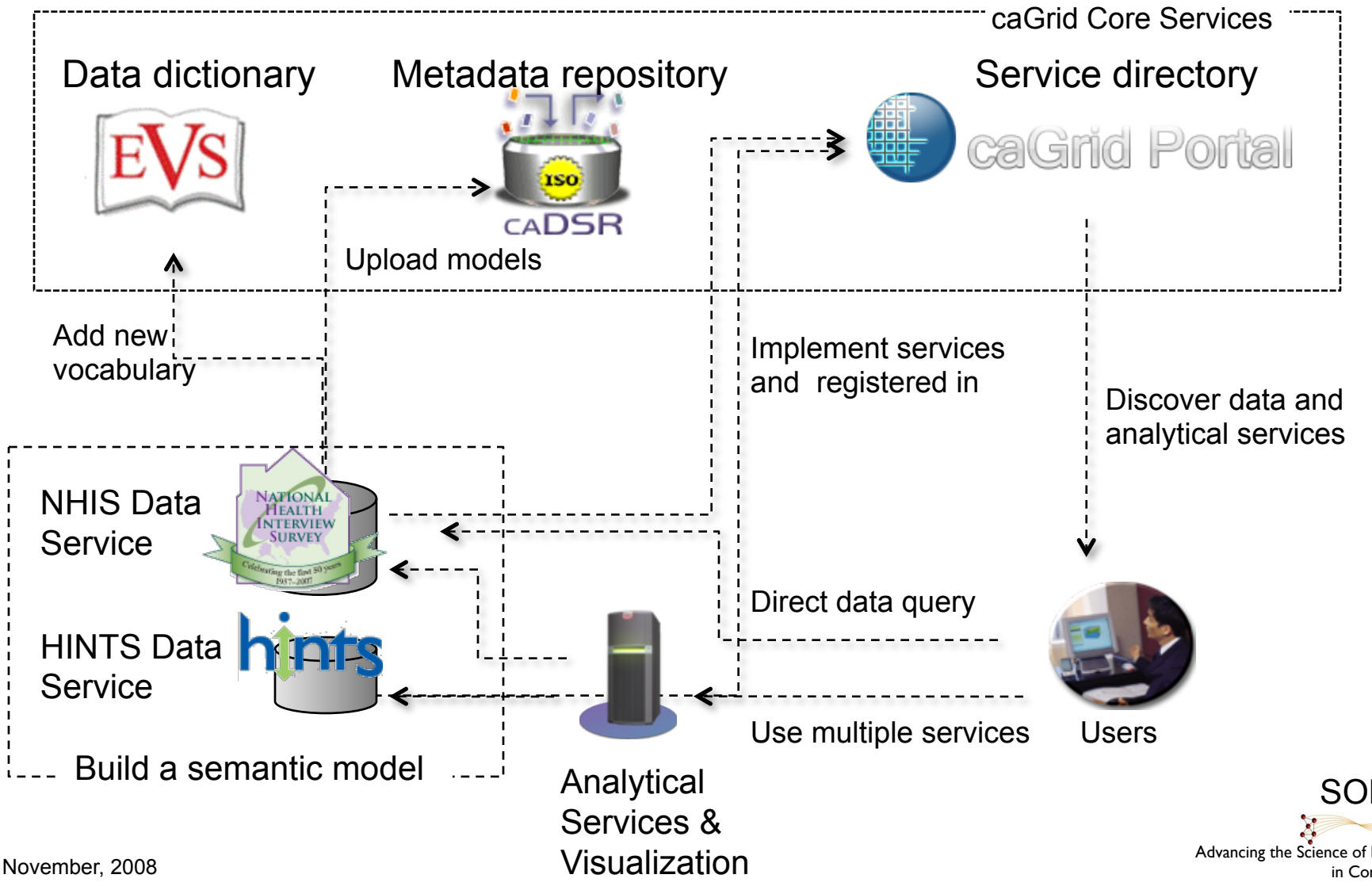
Putting Tobacco Data on the Grid

1. Build a semantic model for each dataset
   - Subject matter experts work with model developers to create detailed, concise definitions for behavioral measures.
2. Register datasets in data dictionary and metadata repository
   - Add new vocabulary to data dictionary
   - Upload semantic models to metadata repository
3. Implement data services
   - Programmers use Grid tools to generate code for data web services.
4. Publish data services on the Grid

Putting Analytical services on the Grid

1. Implement analytical services
   - Design algorithms and application programming interfaces (API)
   - Programmers use Grid tools to generate code for web services.
2. Publish analytical services on the Grid

Overview of Service Integration in PopSciGrid



SONIC
Advancing the Science of Networks in Communities

PopSciGrid Demo

Once data, analytical, and visualization services are published, other programs and web services can use them automatically.

We built the PopSciGrid application to illustrate various methods of data query, analysis, and visualization from three data services.

http://umlmodelbrowser.nci.nih.gov/umlmodelbrowser/

http://cagrid-portal.nci.nih.gov/

http://sonichost-dev.iems.northwestern.edu/GridServer/c/index.html

Analyze Datasets in PopSciGrid

- Data services
  - Publish data in the Grid as web services
- Transformation services
  - Convert different scales
- Analytical services
  - Statistical analysis across datasets
- Visualization services
  - Illustrate datasets by geographical regions (geo-coded data)

Challenges: Lessons Learned

- Technology
  - Common vocabulary and UML modeling
  - Data size/volume and transfer
- Science
  - Team science
  - Data integration and shared measures
  - Transform coding schemes for population sciences to integrate data sources
  - Advanced statistical methods and complex survey sampling
  - Data access regulations and privacy

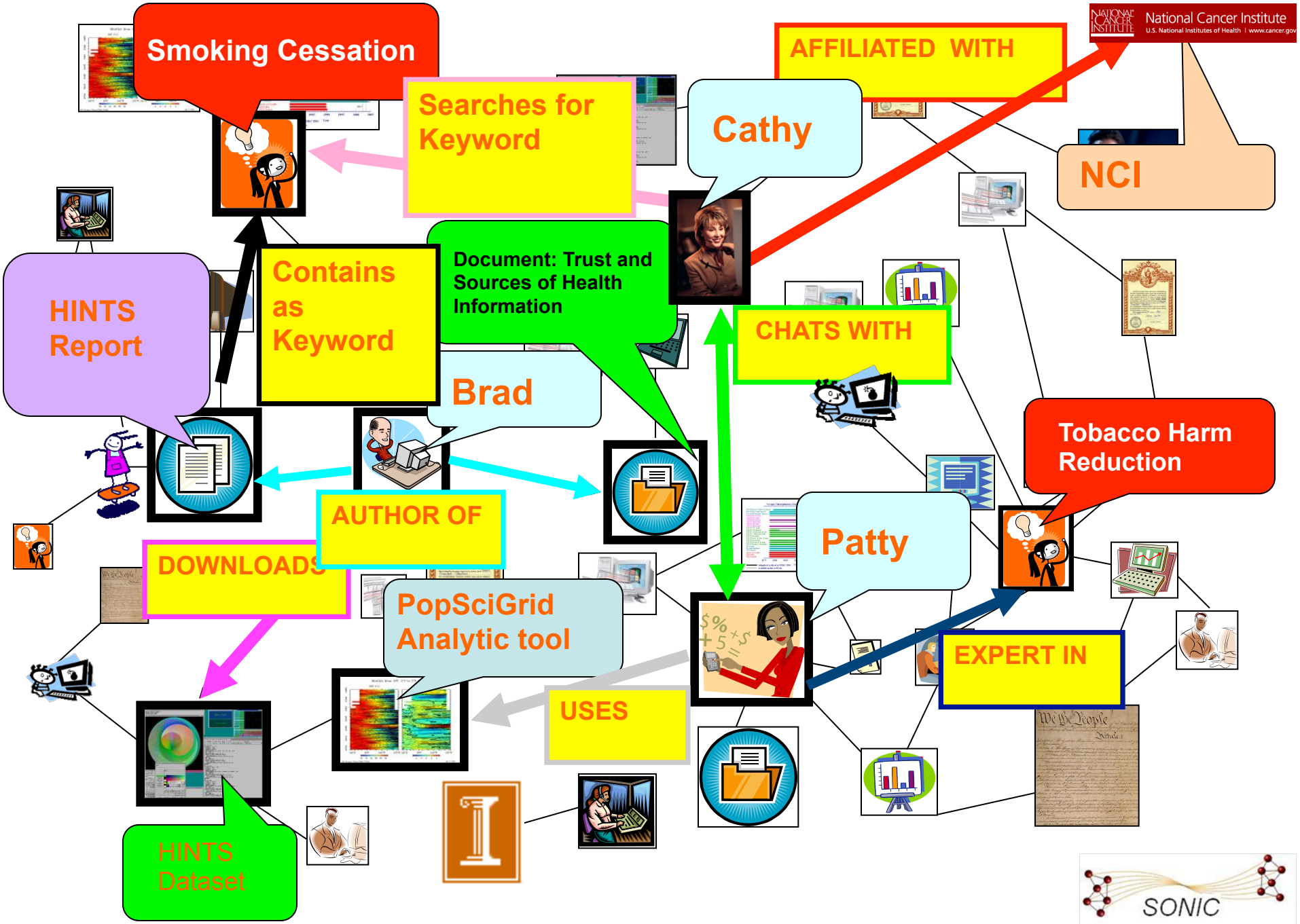
PopSciGrid: What can it do?

- Successful mounting of multiple public health datasets on the grid
  - Dynamic access to 14 datasets spanning 6 years
  - Integration of public health, economic, and geo-spatial data for real-time analyses
  - Multiple ways to overlay this data with geo-spatial data
- Proof of concept that can be built upon by the scientific community
  - Mounting more datasets, different kinds of data (clinical, census, SEER,…), new statistical applications, and user-specific applications
- Potential linkages to Psychosocial Measures and Theories Databases
  - i.e., Rick's database and Sarah's database

Knowledge Networks on the Grid

- Knowledge networks
  - Datasets, analytics, methods
  - Researchers, practitioners, institutions
  - Complex relationship among them
- Discover, diagnose, and design
- Facilitate collaboration
- Cyberinfrastructure Knowledge Networks on the Web (CI-KNOW)

PopSciGrid Community: A Multidimensional Network



CI-KNOW Recommender

- Discover semantic and relation information on the Grid
- Recommend data, analytics, and workflow resources based on networks
- Applications in research portals

Acknowledgements