Can Big Data Motivate New Theories and Methods?

Noshir Contractor
Jane S. & William J. White Professor of Behavioral Sciences
Northwestern University, USA
nosh@northwestern.edu
Twitter:@noshir

Supported by following National Science Foundation grants:
IIS-0838564, IIS-0841583**,** BCS- 0940851, OCI-0904356, CNS-1010904
Army Research Institute W91WAW-08-C-0106/P00004**;**
Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053
& NIH/NCRR & NIH/NCRR grant to Northwestern University
Clinical and Translational Sciences Institute (NUCATS)

NORTHWESTERN
UNIVERSITY

SONIC
advancing the
science of networks in communities

Outline

- We are in the midst of a perfect storm for understanding and enabling social phenomenon because of recent developments in:
  - Theories: Theoretical advances about existing and emerging socio-technical phenomena
  - Data: Developments in Semantic Web/Web 2.0 provide the technological capability to capture, store , merge, and query relational metadata needed to more effectively understand and enable networks.
  - Methods: An ensemble of innovative designs and analytic techniques
  - Computational infrastructure: Cloud computing and petascale are critical to face the computational challenges in observing and analyzing the data

Theory

- Theories we have not been able to test at scale (e.g. brokerage and mobility)
- Theories that invite discussion of planned obsolescence (e.g., Media Richness)
- Theories that need redrawing of boundary conditions (e.g. Social identity in virtual contexts)
- Theories that focus attention on new – or at least increasingly prevalent – phenomena (e.g., collective intelligence , team assembly – self assembly/self-selection in Wikipedia)
- Theories that invite consideration of new combinations of variables (e.g., include neuroscience explanations in attention management; directed nonverbal data such as eye-gaze in group performance, activity sensor data in network formation)

Theory

- Theories we have not been able to test at scale (e.g. brokerage and mobility)
- Theories that invite discussion of planned obsolescence (e.g., Media Richness)
- Theories that need redrawing of boundary conditions (e.g. Social identity in virtual contexts)
- Theories that focus attention on new – or at least increasingly prevalent – phenomena (e.g., collective intelligence , team assembly – self assembly/self-selection in Wikipedia)
- Theories that invite consideration of new combinations of variables (e.g., include neuroscience explanations in attention management; directed nonverbal data such as eye-gaze in group performance, activity sensor data in network formation)

Theory

- Theories we have not been able to test at scale (e.g. brokerage and mobility)
- Theories that invite discussion of planned obsolescence (e.g., Media Richness)
- Theories that need redrawing of boundary conditions (e.g. Social identity in virtual contexts)
- Theories that focus attention on new – or at least increasingly prevalent – phenomena (e.g., collective intelligence , team assembly – self assembly/self-selection in Wikipedia)
- Theories that invite consideration of new combinations of variables (e.g., include neuroscience explanations in attention management; directed nonverbal data such as eye-gaze in group performance, activity sensor data in network formation)

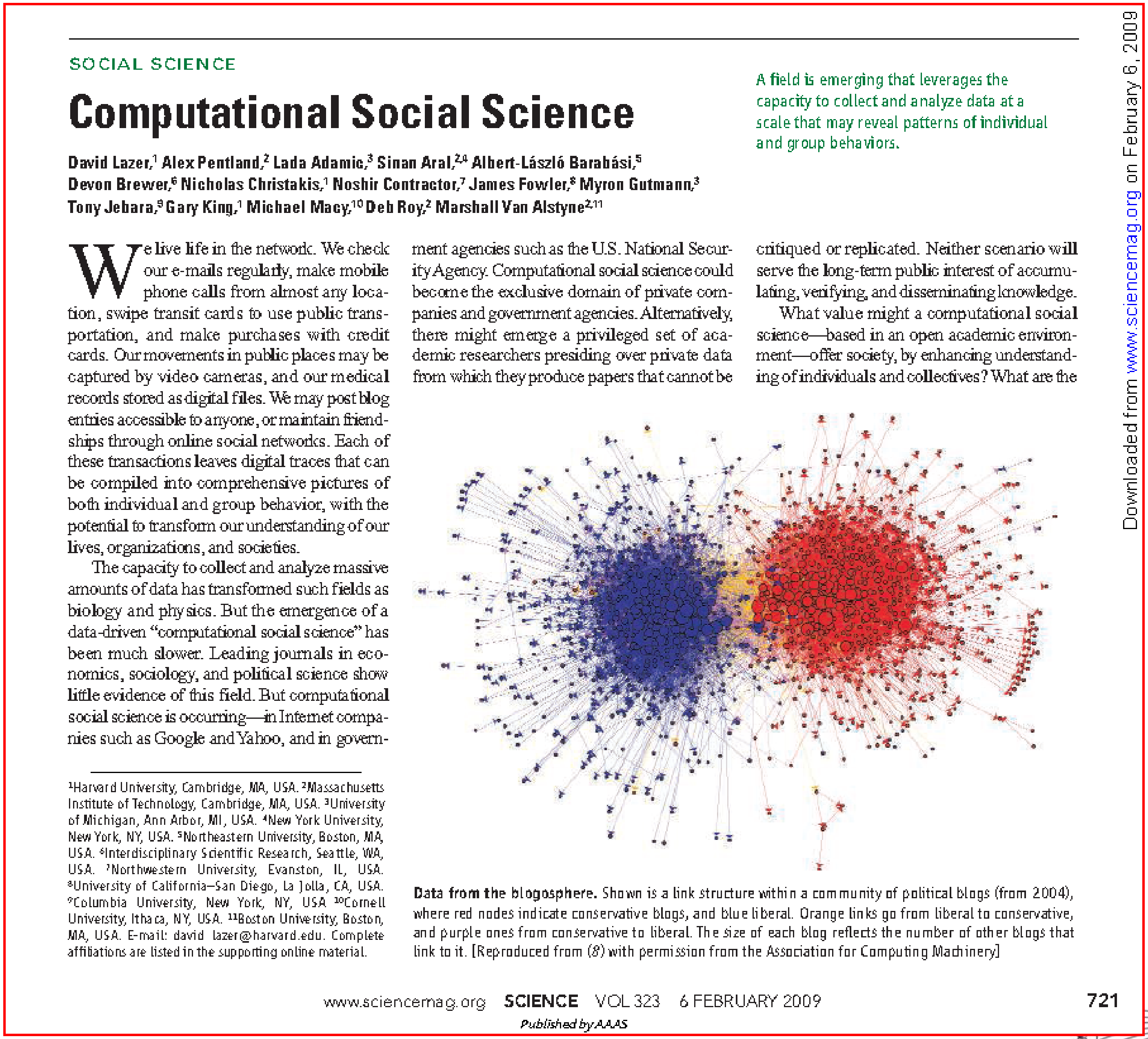
Theory

- Theories we have not been able to test at scale (e.g. brokerage and mobility)
- Theories that invite discussion of planned obsolescence (e.g., Media Richness)
- Theories that need redrawing of boundary conditions (e.g. Social identity in virtual contexts)
- Theories that focus attention on new – or at least increasingly prevalent – phenomena (e.g., collective intelligence , team assembly – self assembly/self-selection in Wikipedia)
- Theories that invite consideration of new combinations of variables (e.g., include neuroscience explanations in attention management; directed nonverbal data such as eye-gaze in group performance, activity sensor data in network formation)

Theory

- Theories we have not been able to test at scale (e.g. brokerage and mobility)
- Theories that invite discussion of planned obsolescence (e.g., Media Richness)
- Theories that need redrawing of boundary conditions (e.g. Social identity in virtual contexts)
- Theories that focus attention on new – or at least increasingly prevalent – phenomena (e.g., collective intelligence , team assembly – self assembly/self-selection in Wikipedia)
- Theories that invite consideration of new combinations of variables (e.g., include neuroscience explanations in attention management; directed nonverbal data such as eye-gaze in group performance, activity sensor data in network formation)

Challenges of empirically testing, extending, and exploring theories … until now



SOCIAL SCIENCE

Computational Social Science

David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3] Tony Jebara,[9] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven "computational social science" has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring—in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

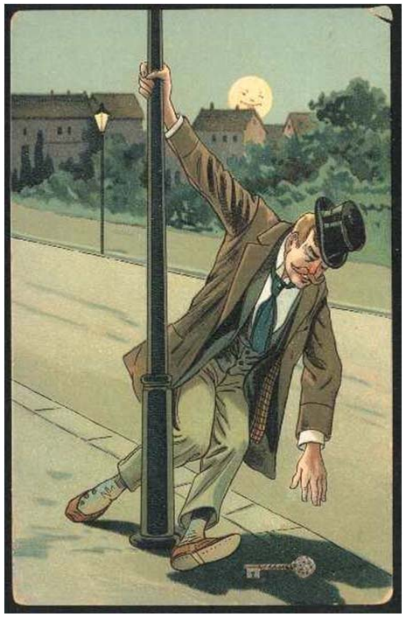
What value might a computational social science—based in an open academic environment—offer society, by enhancing understanding of individuals and collectives? What are the

Data from the blogosphere. Shown is a link structure within a community of political blogs (from 2004), where red nodes indicate conservative blogs, and blue liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it. [Reproduced from (*8*) with permission from the Association for Computing Machinery]

[1]Harvard University, Cambridge, MA, USA. [2]Massachusetts Institute of Technology, Cambridge, MA, USA. [3]University of Michigan, Ann Arbor, MI, USA. [4]New York University, New York, NY, USA. [5]Northeastern University, Boston, MA, USA. [6]Interdisciplinary Scientific Research, Seattle, WA, USA. [7]Northwestern University, Evanston, IL, USA. [8]University of California–San Diego, La Jolla, CA, USA. [9]Columbia University, New York, NY, USA [10]Cornell University, Ithaca, NY, USA. [11]Boston University, Boston, MA, USA. E-mail: david_lazer@harvard.edu. Complete affiliations are listed in the supporting online material.

Published by AAAS

Looking under the lamppost



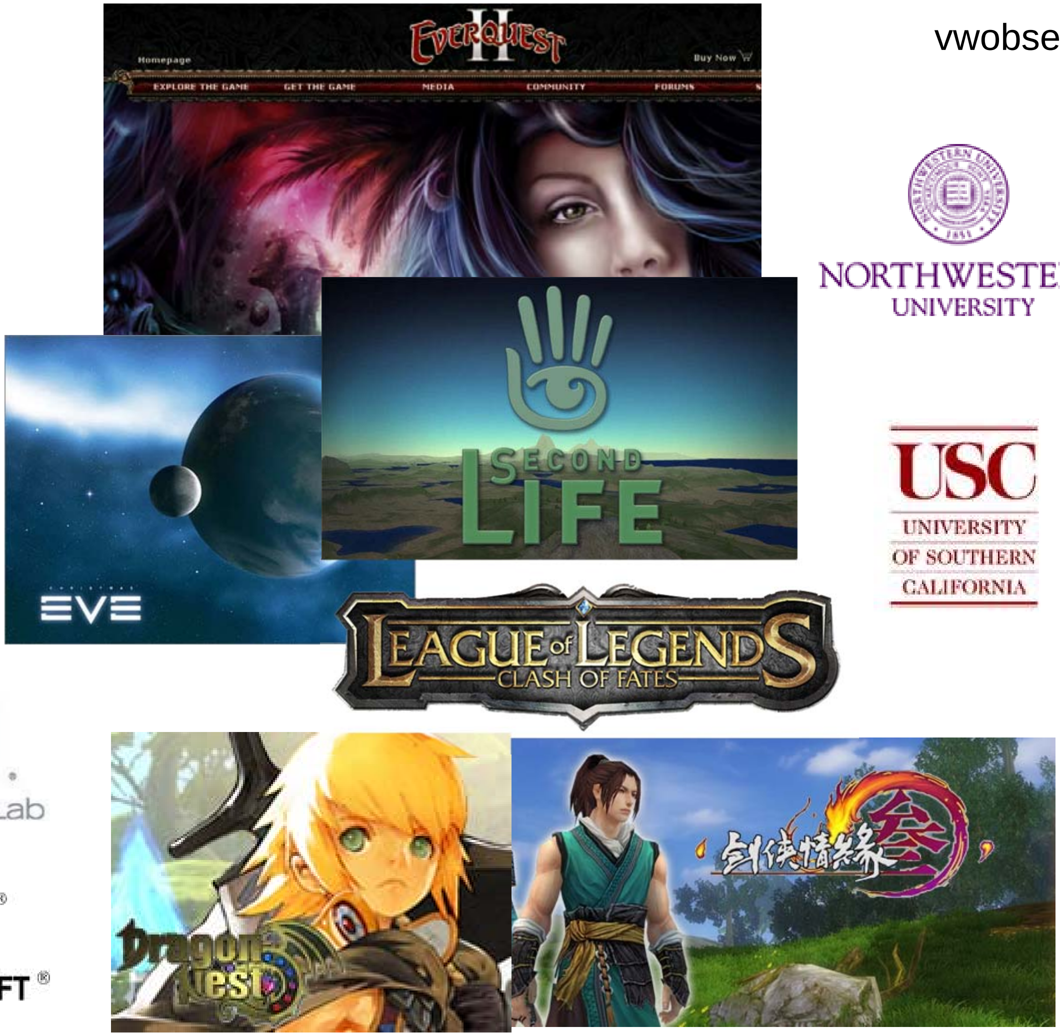
- A drunken man is crawling around on his hands and knees under a lamp-post. His friend asks him “what are you doing crawling around under that lamp-post? The drunk responds that he has lost his keys and is looking for them. His friend responds “your car is over here, you have not been near that lamp-post”. The drunk responds “it is very dark and this is the only place where there is some light”.

Data: Sources

- Large digital trace data repositories: actions, interactions, and transactions (e.g. in a virtual world)

Virtual World Exploratorium

vwobservatory.org



NSF

AIR FORCE RESEARCH LABORATORY

SOE SONY ONLINE ENTERTAINMENT

Linden Lab

SHANDA GAMES

KINGSOFT

NORTHWESTERN UNIVERSITY

REGENTS OF THE UNIVERSITY OF MINNESOTA

USC UNIVERSITY OF SOUTHERN CALIFORNIA

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

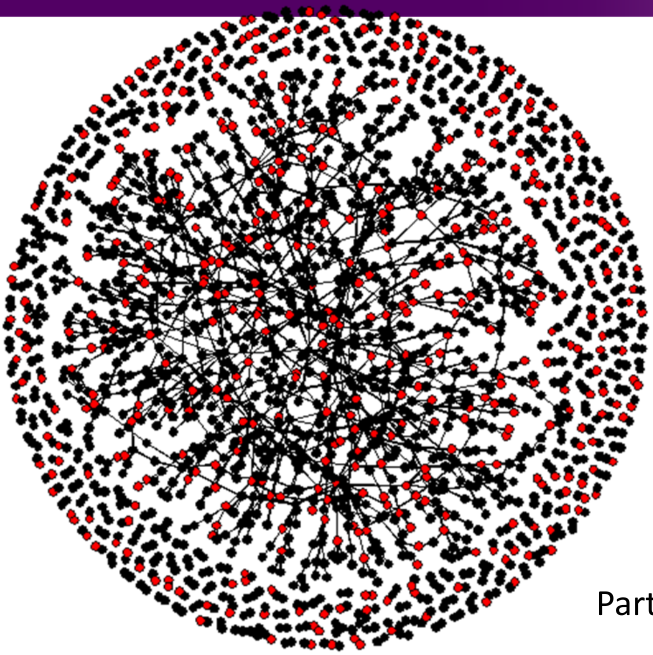
HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

FUDAN UNIVERSITY 1905
復旦大學

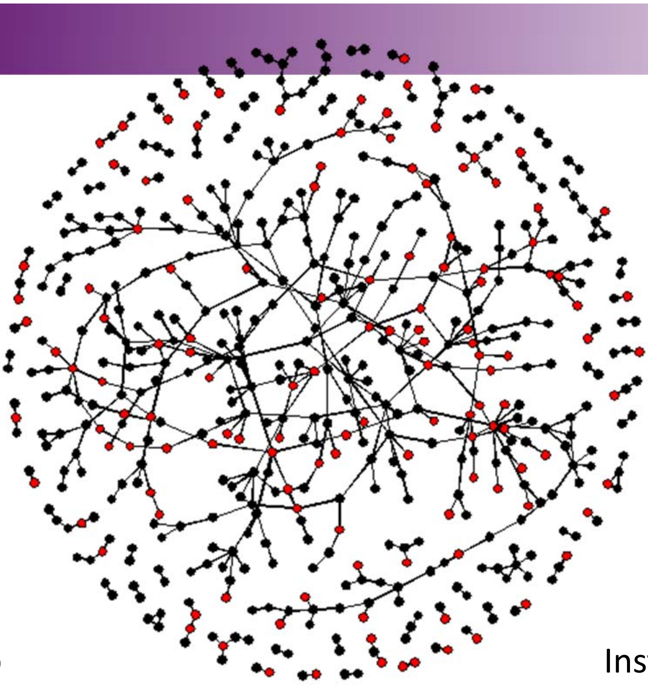
NORTHWESTERN UNIVERSITY

advancing the
science of networks in communities

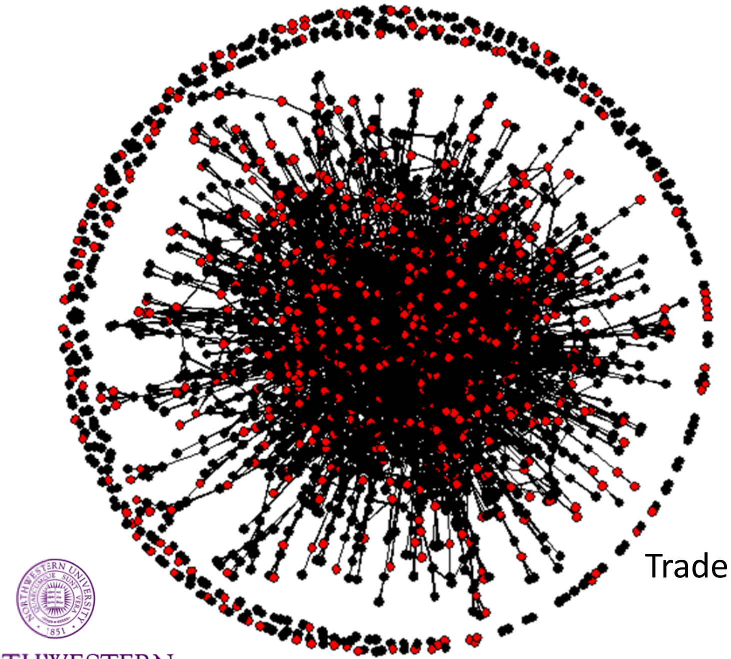
Black: male
Red: female



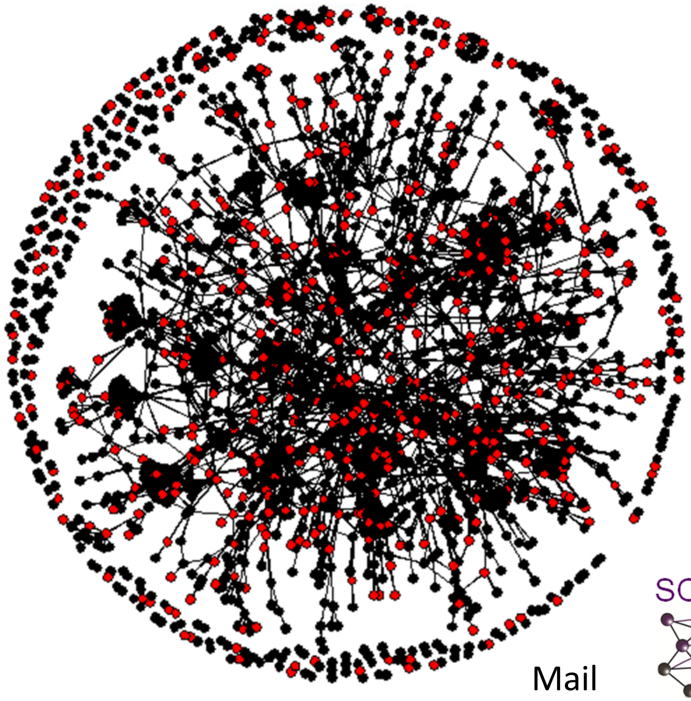
Partnership



Instant messaging



Trade



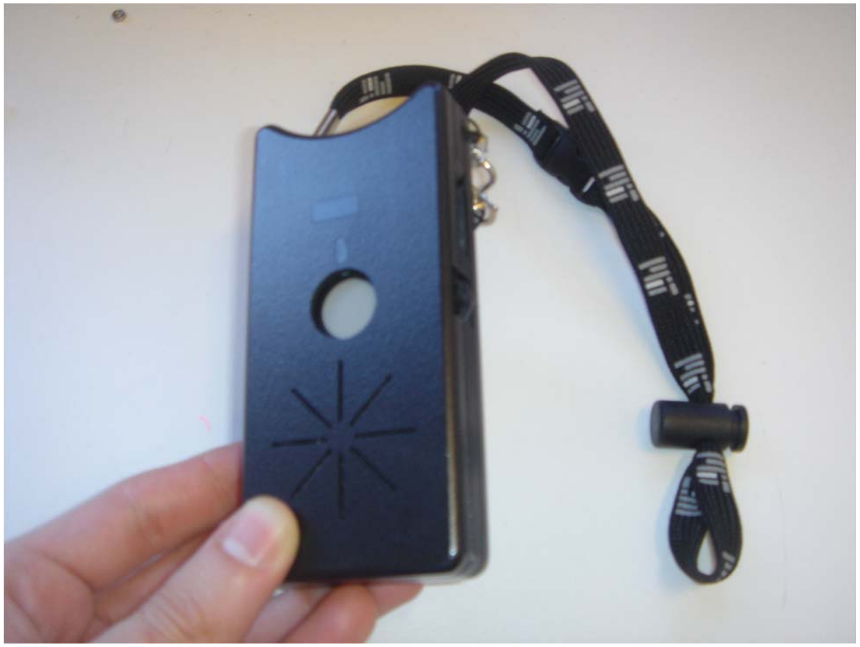
Mail

Data: Sources

- Large digital trace data repositories: actions, interactions, and transactions (e.g. in a virtual world)

- Data from Sensors: human activity, neuro and physiological measures of attitudes

MIT Media Lab



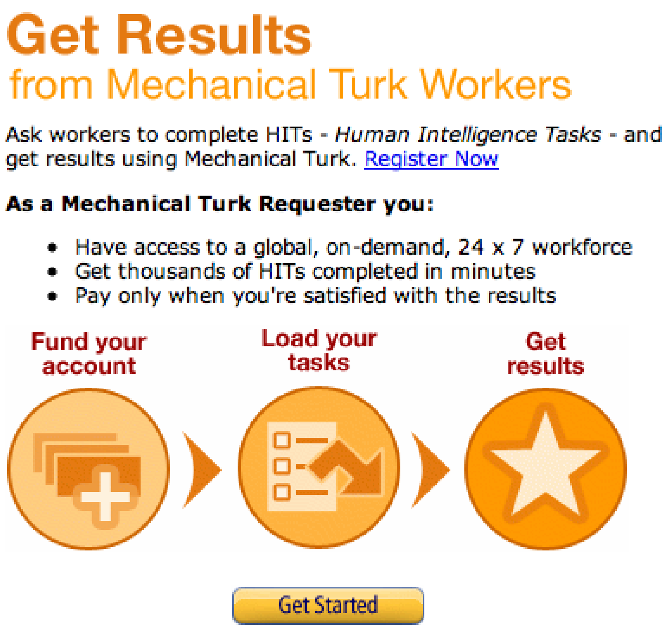
- A sociometric badge (commonly known as a "sociometer") is a wearable electronic device capable of automatically measuring the amount of face-to-face interaction, conversational time, physical proximity to other people, and physical activity levels using social signals derived from vocal features, body motion, and relative location.

http://hd.media.mit.edu/badges/

Data: Sources

- Large digital trace data repositories: actions, interactions, and transactions (e.g. in a virtual world)
- Data from Sensors: human activity, neuro and physiological measures of attitudes
- Crowdsourcing data from people intentionally: Mechanical Turk experiments ….

amazonmechanical turk
beta
Artificial Artificial Intelligence



Get Results
from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. Register Now

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

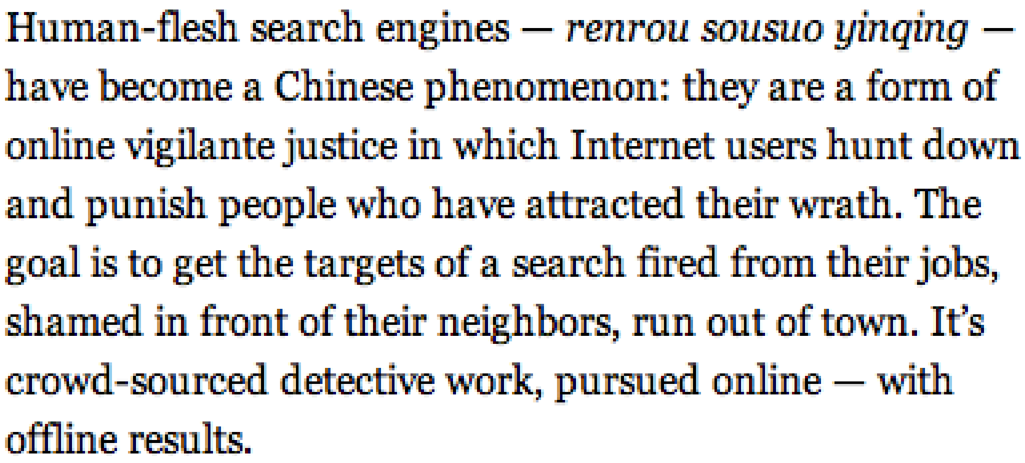
Data: Sources

- Large digital trace data repositories: actions, interactions, and transactions (e.g. in a virtual world)

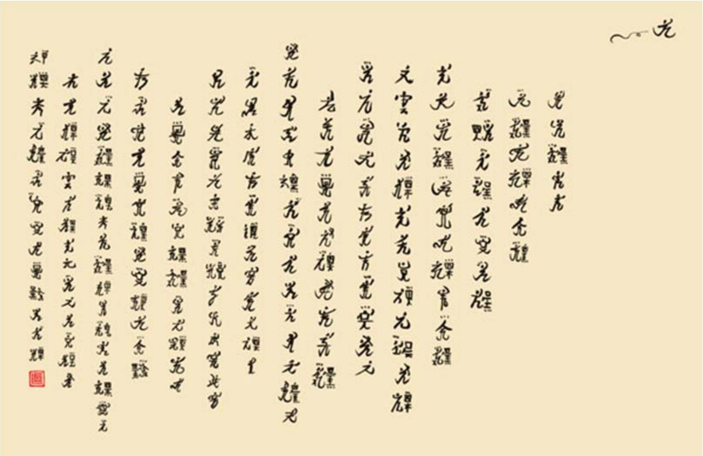
- Data from Sensors: human activity, neuro and physiological measures of attitudes

- Crowdsourcing data from people intentionally (Mechanical Turk experiments, Human Flesh Search) ……

“Human-Flesh” Search



Human-flesh search engines — *renrou sousuo yinqing* — have become a Chinese phenomenon: they are a form of online vigilante justice in which Internet users hunt down and punish people who have attracted their wrath. The goal is to get the targets of a search fired from their jobs, shamed in front of their neighbors, run out of town. It’s crowd-sourced detective work, pursued online — with offline results.



http://www.nytimes.com/2010/03/07/magazine/07Human-t.html

NORTHWESTERN
UNIVERSITY

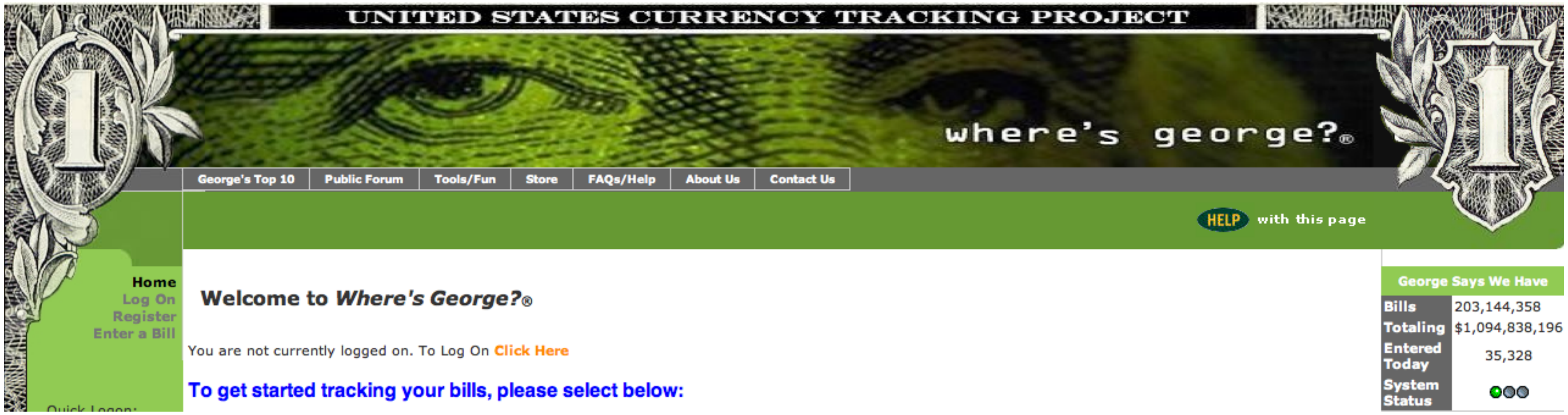
SONIC
advancing the
science of networks in communities

Data: Sources

- Large digital trace data repositories: actions, interactions, and transactions (e.g. in a virtual world)

- Data from Sensors: human activity, neuro and physiological measures of attitudes

- Crowdsourcing data from people intentionally (Mechanical Turk experiments Pearl Jam concert) or unintentionally (Where’s George?)

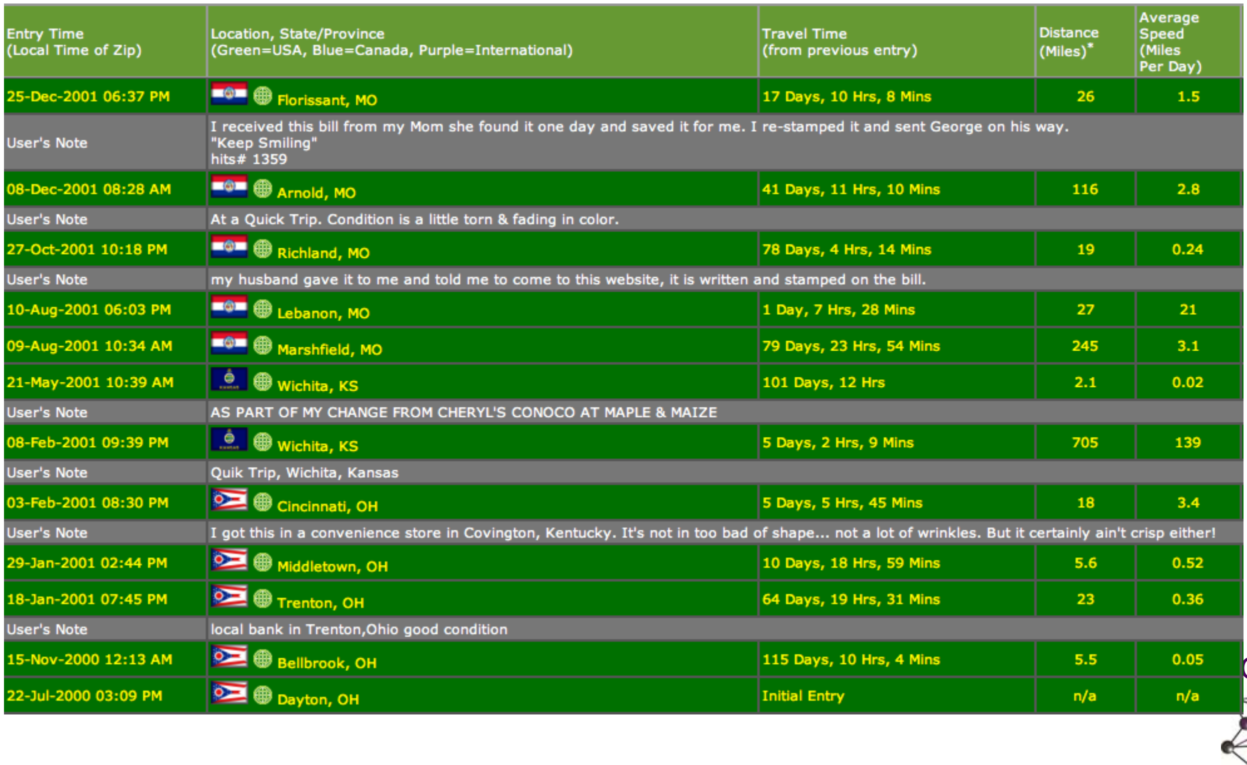


Welcome to *Where's George?*®

You are not currently logged on. To Log On Click Here

To get started tracking your bills, please select below:

George Says We Have	
Bills	203,144,358
Totaling	$1,094,838,196
Entered Today	35,328
System Status	●●●



Entry Time (Local Time of Zip)	Location, State/Province (Green=USA, Blue=Canada, Purple=International)	Travel Time (from previous entry)	Distance (Miles)*	Average Speed (Miles Per Day)
25-Dec-2001 06:37 PM	Florissant, MO	17 Days, 10 Hrs, 8 Mins	26	1.5
User's Note	I received this bill from my Mom she found it one day and saved it for me. I re-stamped it and sent George on his way. "Keep Smiling" hits# 1359			
08-Dec-2001 08:28 AM	Arnold, MO	41 Days, 11 Hrs, 10 Mins	116	2.8
User's Note	At a Quick Trip. Condition is a little torn & fading in color.			
27-Oct-2001 10:18 PM	Richland, MO	78 Days, 4 Hrs, 14 Mins	19	0.24
User's Note	my husband gave it to me and told me to come to this website, it is written and stamped on the bill.			
10-Aug-2001 06:03 PM	Lebanon, MO	1 Day, 7 Hrs, 28 Mins	27	21
09-Aug-2001 10:34 AM	Marshfield, MO	79 Days, 23 Hrs, 54 Mins	245	3.1
21-May-2001 10:39 AM	Wichita, KS	101 Days, 12 Hrs	2.1	0.02
User's Note	AS PART OF MY CHANGE FROM CHERYL'S CONOCO AT MAPLE & MAIZE			
08-Feb-2001 09:39 PM	Wichita, KS	5 Days, 2 Hrs, 9 Mins	705	139
User's Note	Quik Trip, Wichita, Kansas			
03-Feb-2001 08:30 PM	Cincinnati, OH	5 Days, 5 Hrs, 45 Mins	18	3.4
User's Note	I got this in a convenience store in Covington, Kentucky. It's not in too bad of shape... not a lot of wrinkles. But it certainly ain't crisp either!			
29-Jan-2001 02:44 PM	Middletown, OH	10 Days, 18 Hrs, 59 Mins	5.6	0.52
18-Jan-2001 07:45 PM	Trenton, OH	64 Days, 19 Hrs, 31 Mins	23	0.36
User's Note	local bank in Trenton,Ohio good condition			
15-Nov-2000 12:13 AM	Bellbrook, OH	115 Days, 10 Hrs, 4 Mins	5.5	0.05
22-Jul-2000 03:09 PM	Dayton, OH	Initial Entry	n/a	n/a

NORTHWESTERN
UNIVERSITY

ONIC
advancing the
science of networks in communities

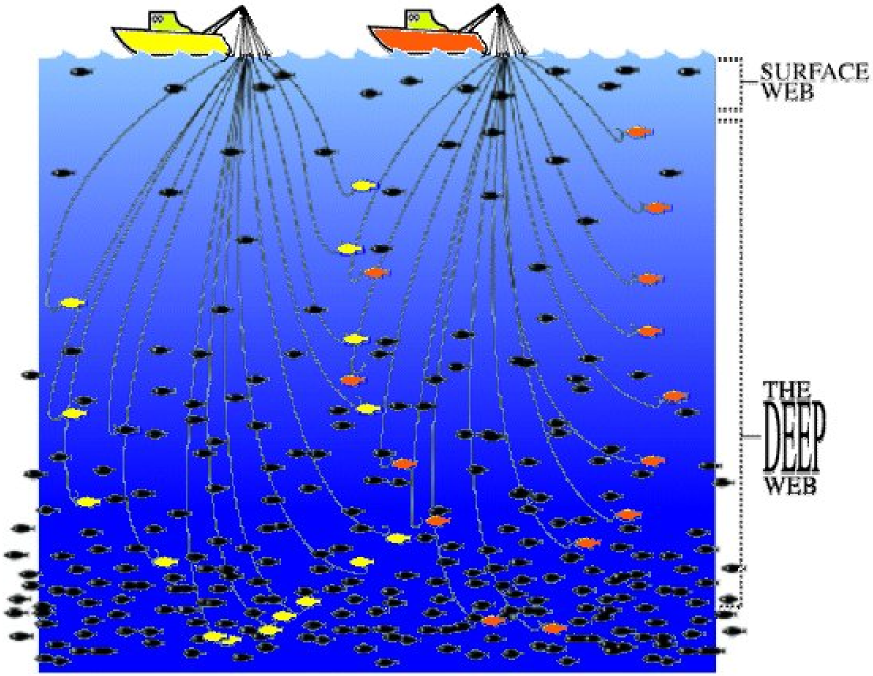
Data: Fusion

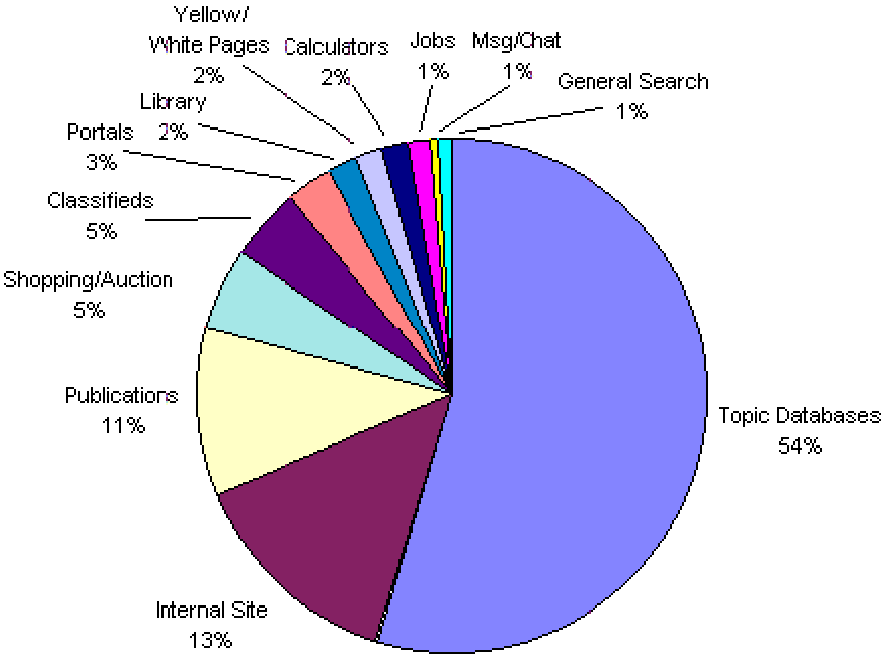
- The Era of Big Data is so 2011!

- Enter the Era of “Broad Data” (Hendler, 2012)

The “Deep Web”

- Data behind web services
- Data behind query interfaces (databases or files)

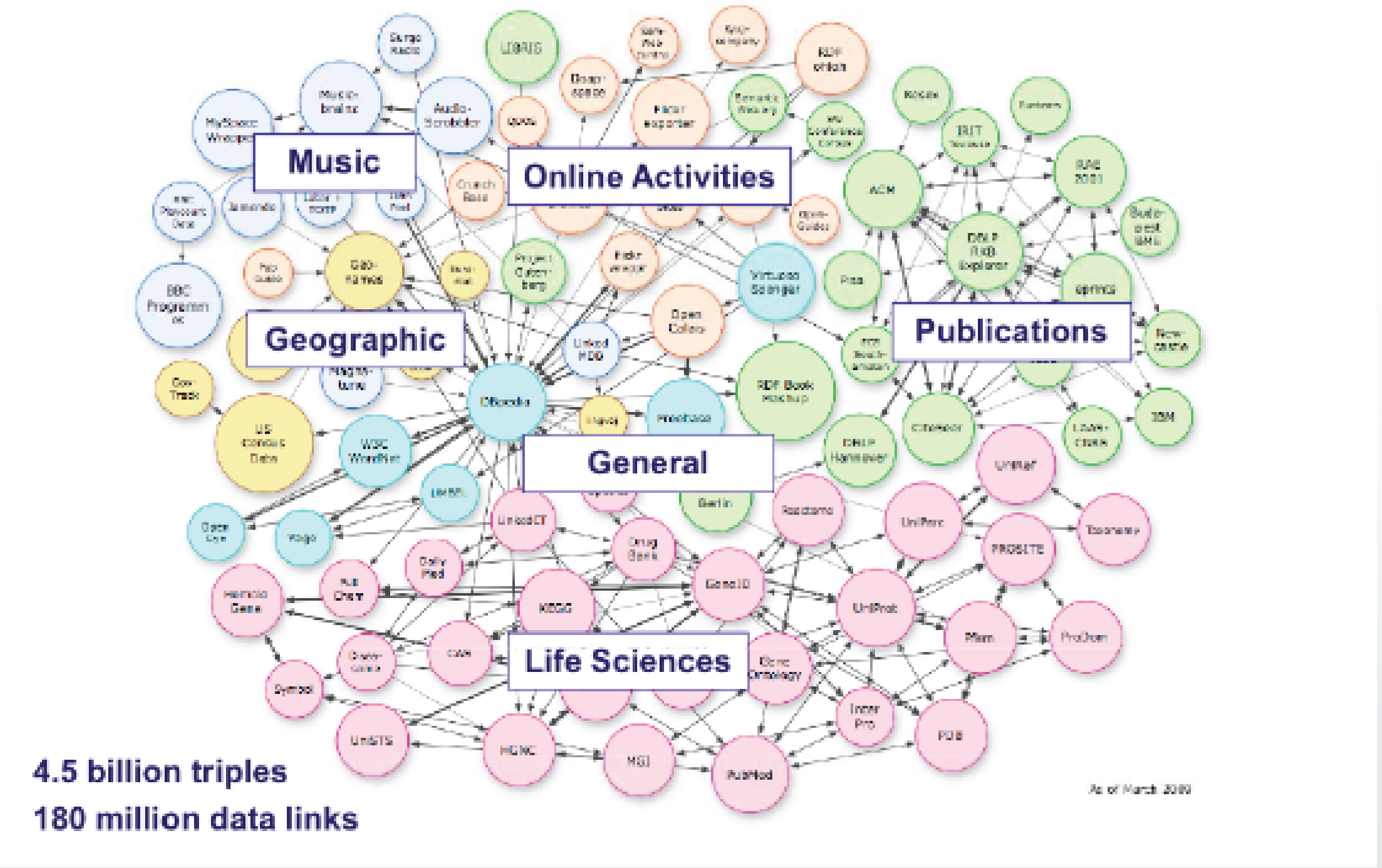




(Hendler, 2012)

Exposing the Deep Web: Linked Open Data

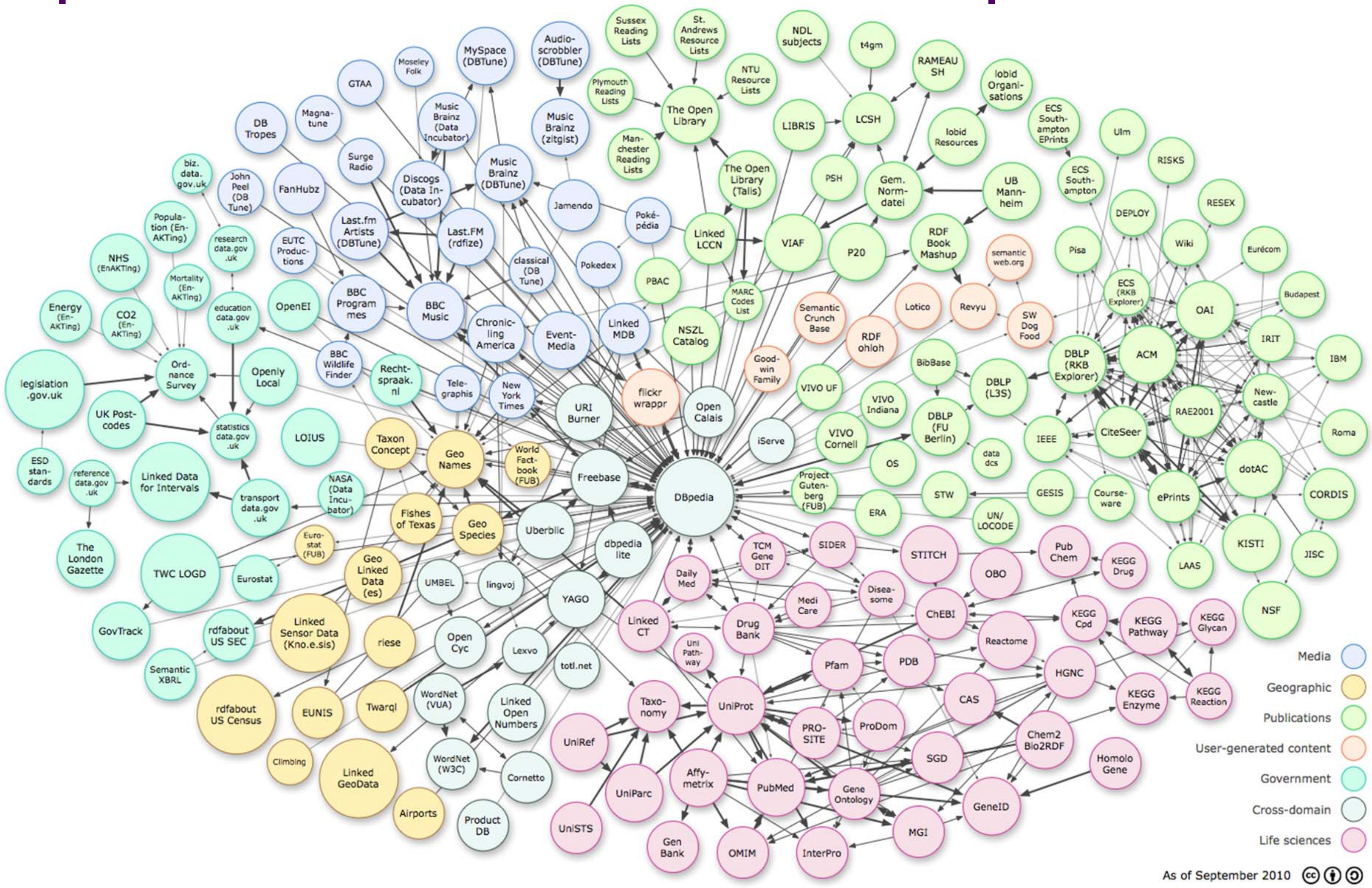
LOD Datasets on the Web: March 2009



4.5 billion triples

180 million data links

Exponential Growth in Linked Open Data



http://richard.cyganiak.de/2007/10/lod/

NORTHWESTERN UNIVERSITY

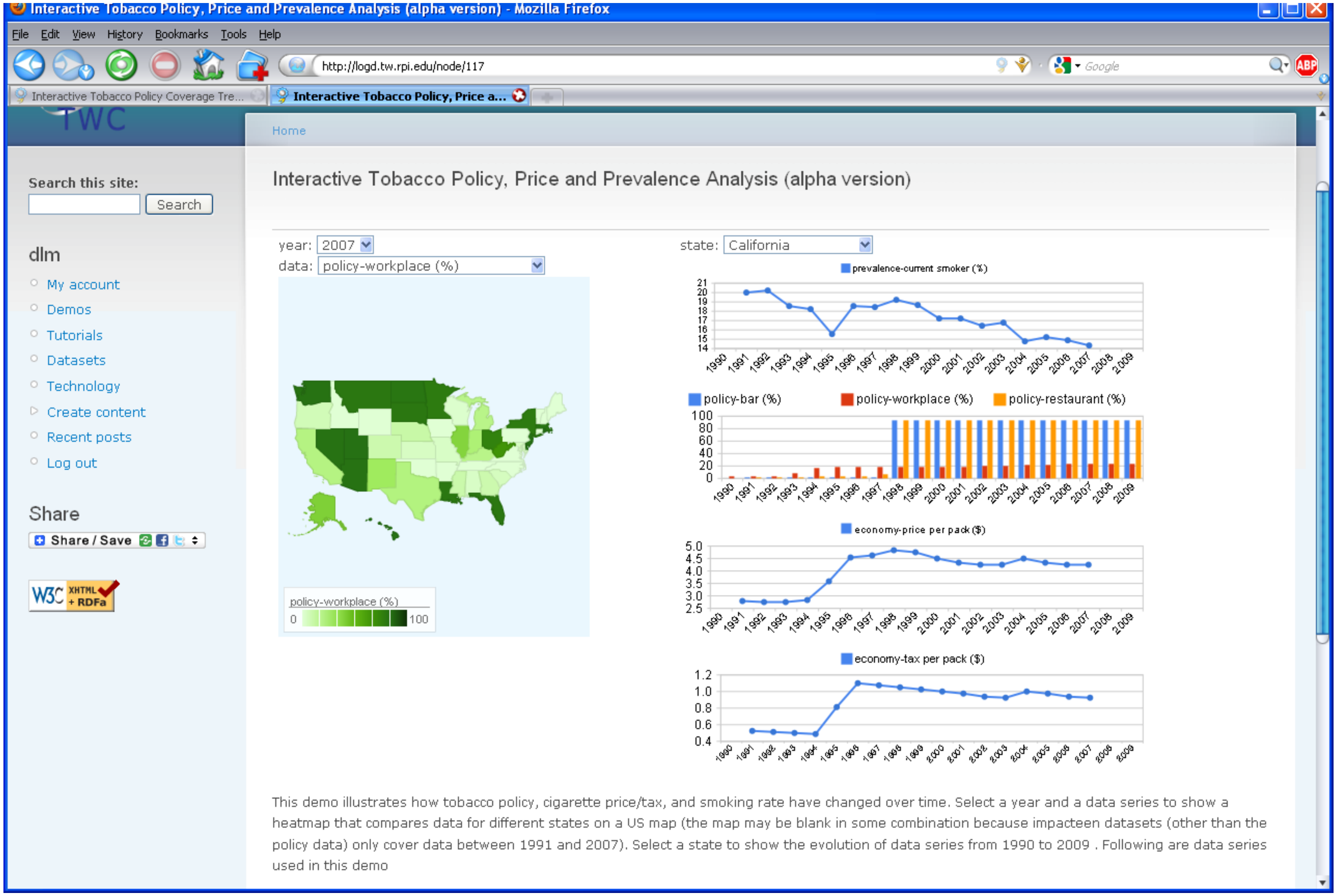
SONIC

advancing the science of networks in communities

PopSci Grid: Population Sciences Grid

- Convey complex health-related information to consumer and public health decision makers for community health impact
- Leverage the growing evidence base for communicating health information on the Internet
- Inform the development of future research opportunities effectively utilizing cyberinfrastructure for cancer prevention and control.

PopSciGrid Example State - California



Interactive Tobacco Policy, Price and Prevalence Analysis (alpha version) - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://logd.tw.rpi.edu/node/117

Interactive Tobacco Policy Coverage Tre... | Interactive Tobacco Policy, Price a...

Home

Interactive Tobacco Policy, Price and Prevalence Analysis (alpha version)

year: 2007

data: policy-workplace (%)

state: California

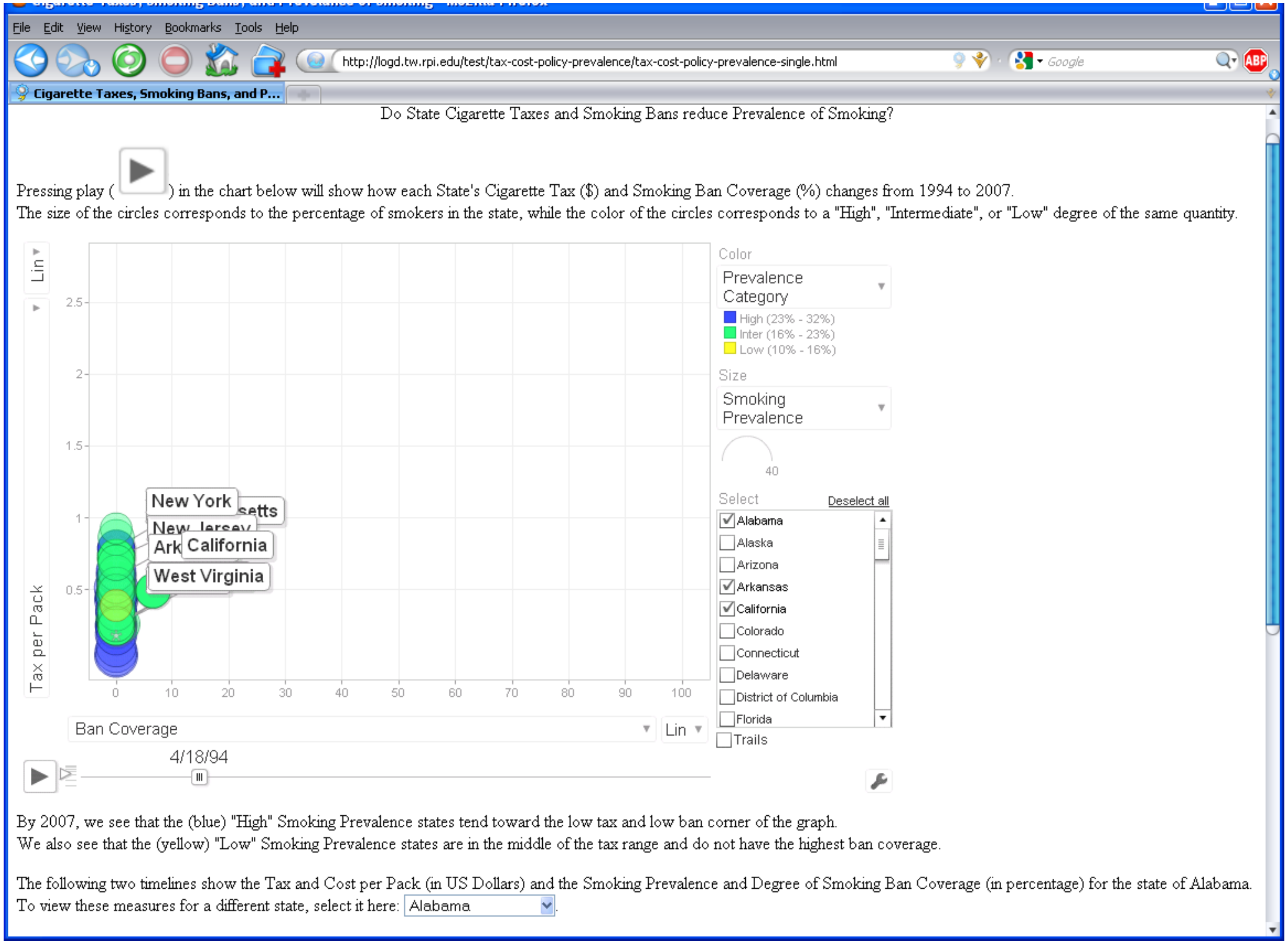
This demo illustrates how tobacco policy, cigarette price/tax, and smoking rate have changed over time. Select a year and a data series to show a heatmap that compares data for different states on a US map (the map may be blank in some combination because impacteen datasets (other than the policy data) only cover data between 1991 and 2007). Select a state to show the evolution of data series from 1990 to 2009 . Following are data series used in this demo

NORTHWESTERN UNIVERSITY

SONIC

advancing the
science of networks in communities

PopSciGrid in Action (inspired by Hans Rosling)


File
Edit
View
History
Bookmarks
Tools
Help
http://logd.tw.rpi.edu/test/tax-cost-policy-prevalence/tax-cost-policy-prevalence-single.html
Google
Cigarette Taxes, Smoking Bans, and P...

Do State Cigarette Taxes and Smoking Bans reduce Prevalence of Smoking?

Pressing play (▶) in the chart below will show how each State's Cigarette Tax ($) and Smoking Ban Coverage (%) changes from 1994 to 2007.
The size of the circles corresponds to the percentage of smokers in the state, while the color of the circles corresponds to a "High", "Intermediate", or "Low" degree of the same quantity.

Color: Prevalence Category

- 🟦 High (23% - 32%)
- 🟩 Inter (16% - 23%)
- 🟨 Low (10% - 16%)

Size: Smoking Prevalence

40

Select   Deselect all

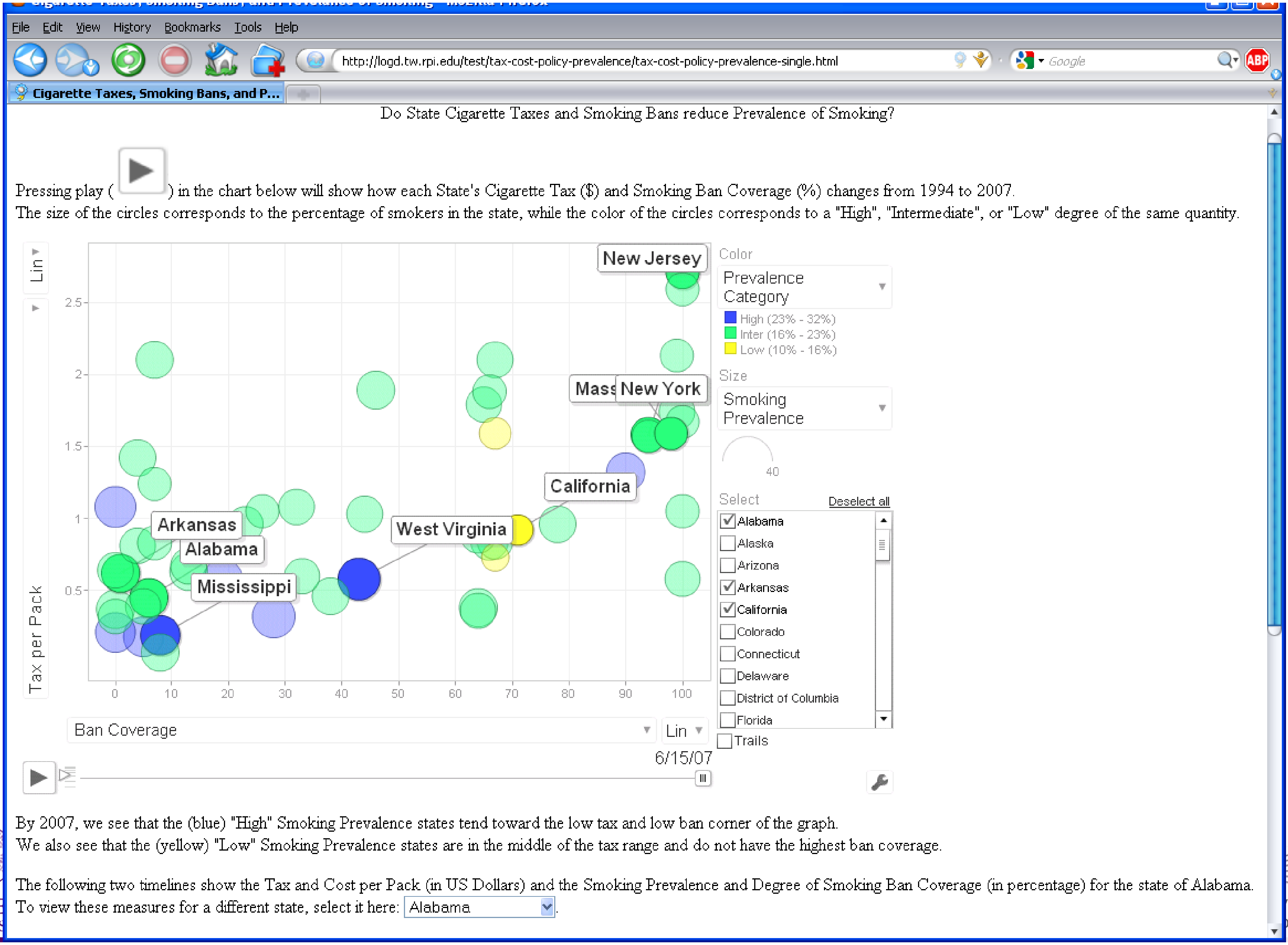
☑ Alabama
☐ Alaska
☐ Arizona
☑ Arkansas
☑ California
☐ Colorado
☐ Connecticut
☐ Delaware
☐ District of Columbia
☐ Florida

☐ Trails

By 2007, we see that the (blue) "High" Smoking Prevalence states tend toward the low tax and low ban corner of the graph.
We also see that the (yellow) "Low" Smoking Prevalence states are in the middle of the tax range and do not have the highest ban coverage.

The following two timelines show the Tax and Cost per Pack (in US Dollars) and the Smoking Prevalence and Degree of Smoking Ban Coverage (in percentage) for the state of Alabama.
To view these measures for a different state, select it here: Alabama.

PopSciGrid II (inspired by Hans Rosling)


File Edit View History Bookmarks Tools Help
http://logd.tw.rpi.edu/test/tax-cost-policy-prevalence/tax-cost-policy-prevalence-single.html
Google
ABP
Cigarette Taxes, Smoking Bans, and P...
New Jersey
Mass
New York
California
West Virginia
Arkansas
Alabama
Mississippi
Lin
Tax per Pack
2.5
2
1.5
1
0.5
0
10
20
30
40
50
60
70
80
90
100
Ban Coverage
Lin
6/15/07

Do State Cigarette Taxes and Smoking Bans reduce Prevalence of Smoking?

Pressing play (▶) in the chart below will show how each State's Cigarette Tax ($) and Smoking Ban Coverage (%) changes from 1994 to 2007.
The size of the circles corresponds to the percentage of smokers in the state, while the color of the circles corresponds to a "High", "Intermediate", or "Low" degree of the same quantity.

Color: Prevalence Category
- 🟦 High (23% - 32%)
- 🟩 Inter (16% - 23%)
- 🟨 Low (10% - 16%)

Size: Smoking Prevalence

40

Select — Deselect all
- ☑ Alabama
- □ Alaska
- □ Arizona
- ☑ Arkansas
- ☑ California
- □ Colorado
- □ Connecticut
- □ Delaware
- □ District of Columbia
- □ Florida

□ Trails

By 2007, we see that the (blue) "High" Smoking Prevalence states tend toward the low tax and low ban corner of the graph.
We also see that the (yellow) "Low" Smoking Prevalence states are in the middle of the tax range and do not have the highest ban coverage.

The following two timelines show the Tax and Cost per Pack (in US Dollars) and the Smoking Prevalence and Degree of Smoking Ban Coverage (in percentage) for the state of Alabama.
To view these measures for a different state, select it here: Alabama.

Methods

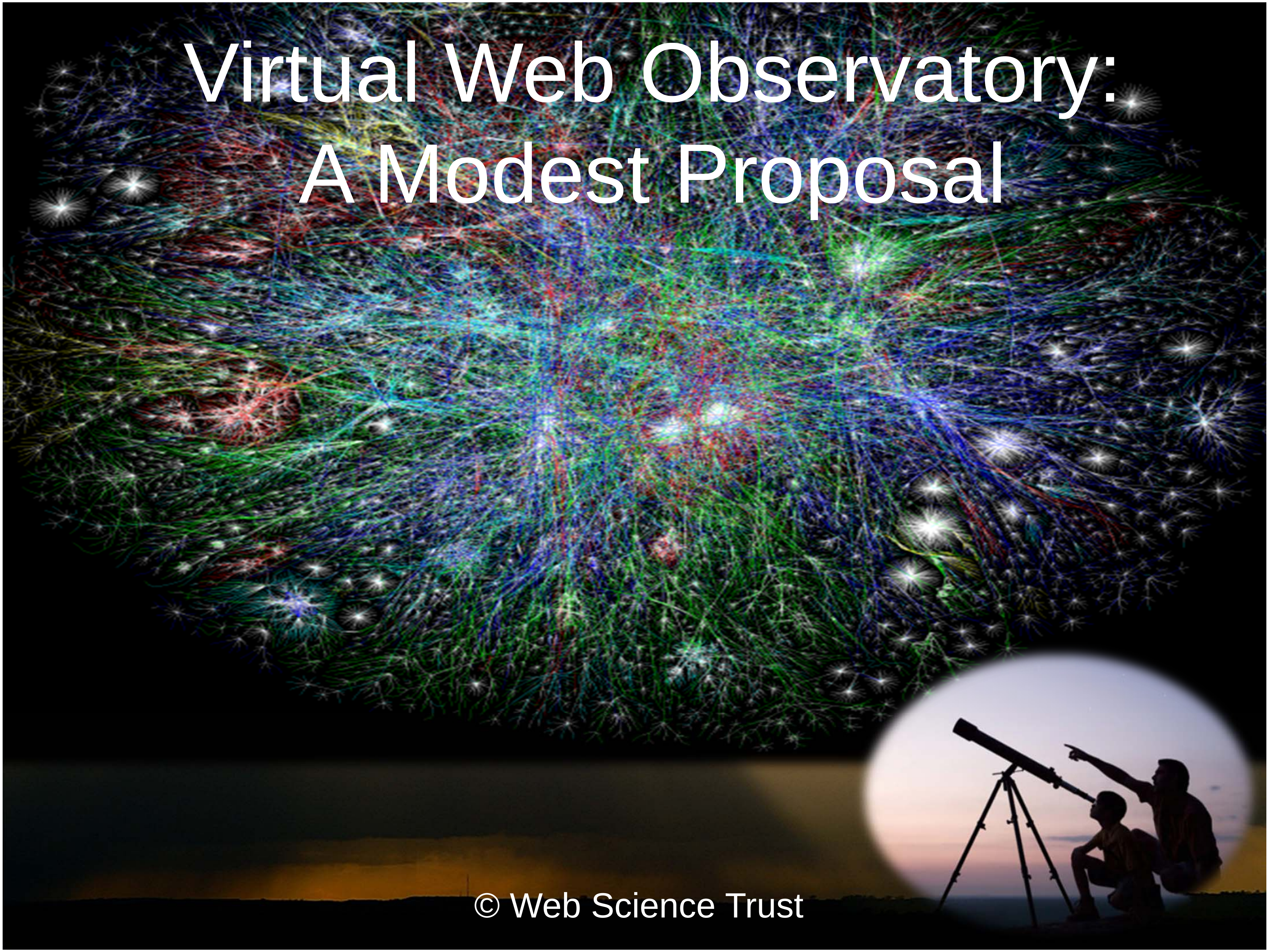
- Study Design: Interventions at scale to study behavioral changes (e.g., Amazon)
- Confirmatory methods: New focus on event-based longitudinal analysis and hypergraphs.
- Exploratory methods: Data/text mining – from “Dustbowl empiricism” to the “Superbowl of novel insights”

Data: Fusion

- We need a Google for Data! Web links pages - but not data.

- A “Virtual Web observatory”: connecting people, data, documents, instruments (surveys, web tools),…

Virtual Web Observatory:
A Modest Proposal



The Web as a repository

- Micro scale
- of documents
- of links
- of data

Macro scale

- of social activity
- of business activity
- of citizen activity

Do we have what is needed to:

- Understand its evolution?
- Analyze its impact on human behavior?
- Anticipate future developments?

in 1000 years, how will scientists study how we lived on the web?

Virtual Web Observatory

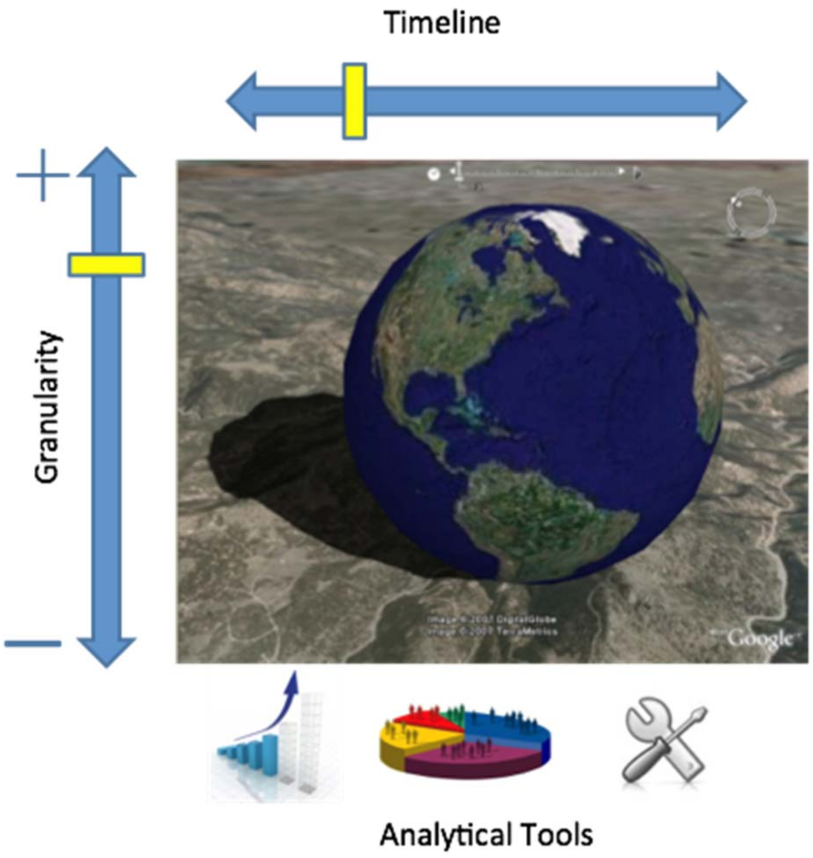
Visualisation of activity on the Web

Analysis of the past developments

Projections and simulations of its evolution

© Web Science Trust

Virtual Web Observatory



Timeline

Granularity

Analytical Tools