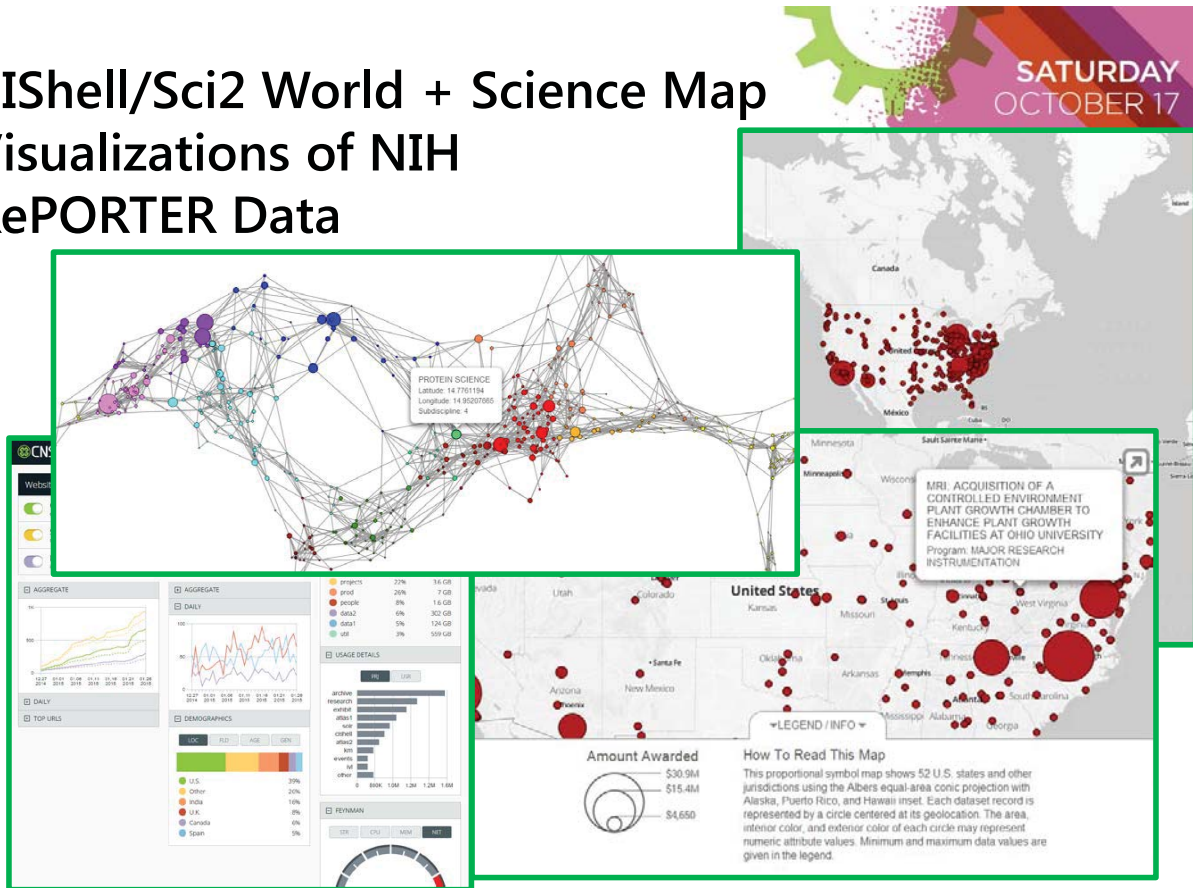SATURDAY
OCTOBER 17

Building Macroscope Workflows
Big Data In -> Big Insights Out

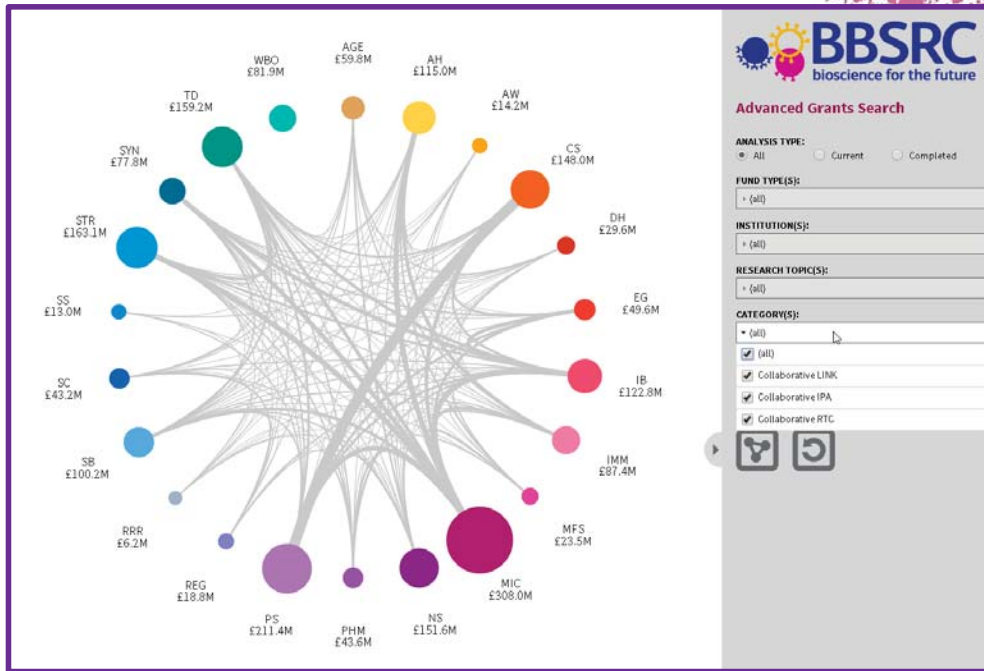**Daniel Halsey**
Cyberinfrastructure for Network Science Center
School of Informatics and Computing, Indiana University, USA

SATURDAY
OCTOBER 17

CIShell/Sci2 World + Science Map Visualizations of NIH RePORTER Data

CNS Cyberinfrastructure for Network Science Center

BBSRC: Visual Interface to Biomedical Funding Data in UK

Illuminated Diagram: Searchable World and Science Maps

Building Macroscope Workflows

- Know your questions
- Know your data
- Know your options

Making Big Data Smaller

- Big data algorithms

- Big iron systems

Making Medium-Sized Data Meaningful

- R
- Tableau
- Sci2
- Data-specific algorithms
- RDBMS
- Small NoSQL

Making Small Data Insightful

- Plotly
- Gephi
- D3
- Leaflet
- Raw

Make It Work: Make It Flow

Identify your insights
- Find needs and desired insights
- Ask questions, know your users
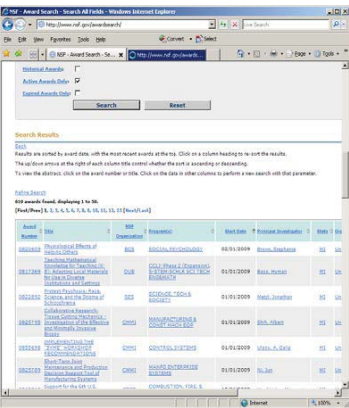- Find supporting datasets, if your original data doesn't have all the answers

THEN: Find how to express this data using a macroscope toolchain
- Determine relevant data processing algorithms
- Identify data analysis algorithms
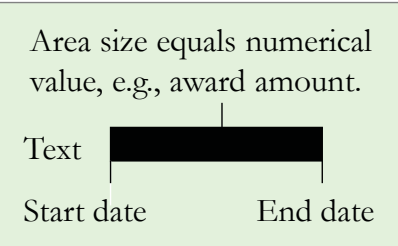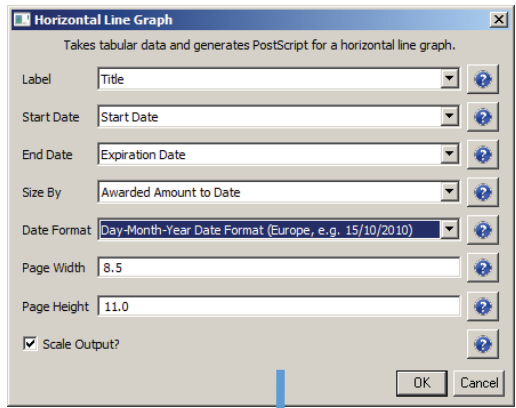- Select data visualization algorithms or tools

CNS Cyberinfrastructure for Network Science Center

Workflow Example:
Horizontal line graph of NSF projects

Download NSF data

Visualize as Horizontal Line Graph

Area size equals numerical value, e.g., award amount.

Text

Start date   End date

**CNS** Cyberinfrastructure for Network Science Center

Workflow Example:
Geomap with Gephi Network Overlay

File with geolocations and linkage info, e.g., an isi biblio-graphy file.

Use Yahoo! Geocoder to identify Latitude, Longitude for each geolocation
Extract attributes per geolocation, e.g., total times cited (TC)
Extract linkages and their attributes, e.g., number of co-occurences
See sample /geo/LaszloBarabasiGeo.net with co-occurrence of "Research Addresses" and full counting of TC per geolocation.
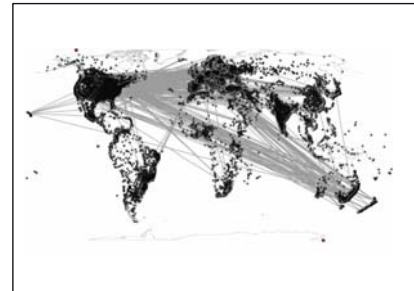
Read into Sci2 Tool to generate geomap and network file → Layout network in Gephi → Combine geomap and network in Photoshop
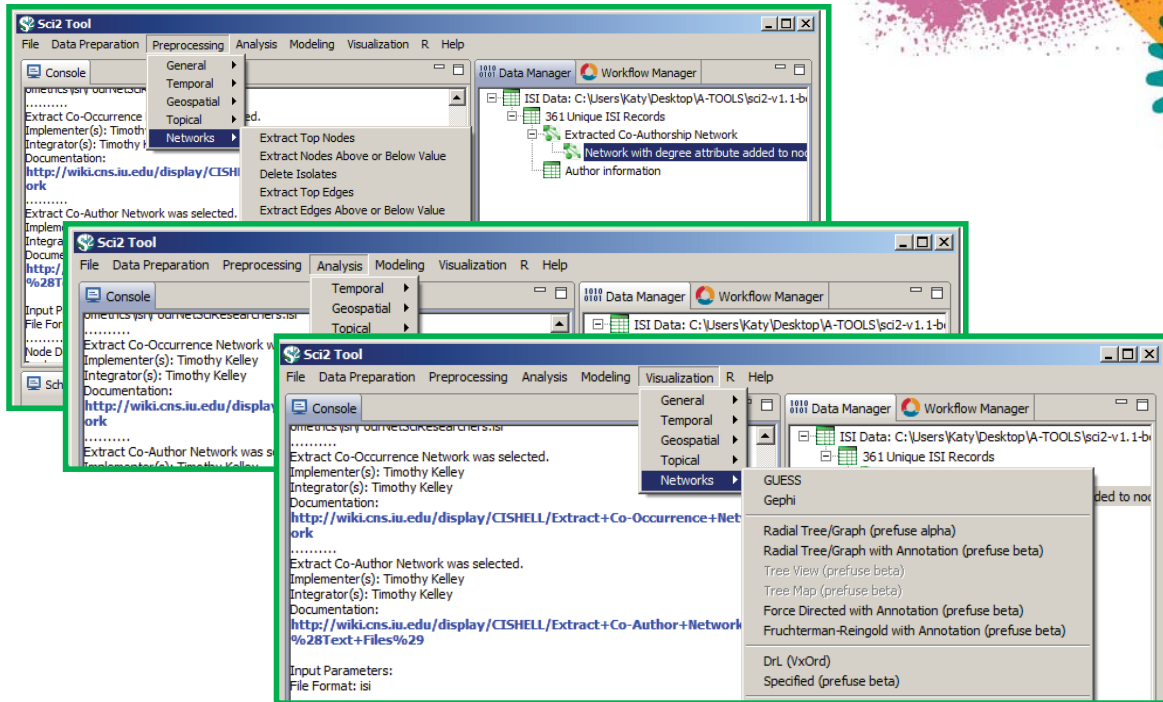

+

=

CNS Cyberinfrastructure for Network Science Center
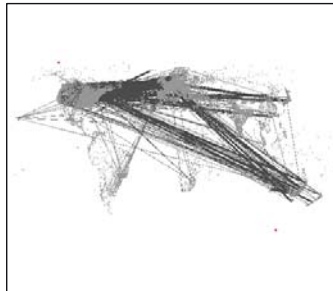
Filling In The Gaps
Building Tools for a Macroscope Workflow

- Know your data structures: Data types coming in, going out, and everything in between
- Link it all: Build the best connector/converter between data preprocessing, analysis and visualization algorithms
- Know your algorithms: Answer
  - When Qs - using temporal analysis & vis algorithms
  - Where Qs - using geospatial analysis & vis algorithms
  - What Qs - using topical analysis & vis algorithms
  - With Whom Qs - using network analysis & vis algorithms
- Use standards: Stand on the shoulders of giants

CNS Cyberinfrastructure for Network Science Center

Sci2 Desktop

CNS Cyberinfrastructure for Network Science Center

Practicing what we preach:
Sci2 *looks* monolithic...but it's not!

- CNS's main desktop product, Sci2 Tool (http://sci2.cns.iu.edu), has a huge set of analysis algorithms, visualizations, and data type converters (180+ algorithms and counting)

- Sci2 is a specific packaging of CIShell (http://cishell.org)—a framework built on top of OSGi (http://osgi.org), the same plugin engine as Eclipse

- Each algorithm is a plugin: what looks like tight integration is the result of plugin discovery and an extensive data file format converter system

- Plugins can do almost anything, in almost any language, as long as they're wrapped correctly

CNS Cyberinfrastructure for Network Science Center