INNOVATIONS IN PUBLIC DATA

Julia Lane
American Institutes for Research
Overview

• Key ideas
• Definitions and motivation
• Application: Science of Science and Innovation Policy
  • Text analysis
  • Linking
  • Units of analysis
• Privacy and Confidentiality
Key ideas

• Opportunity
  • New units of analysis
  • New sources of data
  • New types of analysis

• Challenges
  • New sets of skills
  • Linking
  • Privacy and confidentiality

• What needs to be done by APDU community
  • Training and Analysis
  • Trust
  • Use
Overview

• Key ideas
• Definitions and motivation
• Application: Science of Science and Innovation Policy
  • Text analysis
  • Linking
  • Units of analysis
• Privacy and Confidentiality
Big Data definition

• “Big Data” is an imprecise description of a rich and complicated set of characteristics, practices, techniques, ethics, and outcomes all associated with data. (AAPOR)
• No canonical definition
• By characteristics: Volume Velocity Variety (and Variability and Veracity)
• By source: found vs. made
• By use: professionals vs. citizen science
• By reach: datafication
Motivation: New paradigm

- New analytical paradigm
  - Not just data – but whole approach
- New business paradigm
  - Cost
  - Coverage
  - Timeliness
  - Competition
Motivation: Value in other fields
Motivation: Need analysis, use and trust
Overview

• Key ideas
• Definitions and motivation
• Application: Science of Science and Innovation Policy
  • Text analysis
  • Linking
  • Units of analysis
• Privacy and Confidentiality
How much should a nation spend on science? What kind of science? How much from private versus public sectors? Does demand for funding by potential science performers imply a shortage of funding or a surfeit of performers? A new “science of science policy” is emerging, and it may offer more compelling guidance for policy decisions and for more credible advocacy.
The ITG undertook a literature review to determine the state of the science to date. A questionnaire was also circulated to Federal agencies to ascertain what methods are currently in use for programmatic investment decision making, as well as to ask what tools and resources are needed by Federal agencies that are currently unavailable. The ITG found that:

- There is a well developed body of social science knowledge that could be readily applied to the study of science and innovation.
- Although many Federal agencies have their own communities of practice, the collection and analysis of data about the science and scientific communities they support is heterogeneous and unsystematic.
- Agencies are using very different models, data and tools to understand their investments in science and technology.
- The data infrastructure is inadequate for decision-making.
We don’t know WHAT is funded

- Two surveys… provide some of the most significant data available to understand research and development (R&D) spending and policy in the United States. Budget officials at science agencies, Congress, and interest groups representing scientists, engineers, and high technology industries, among others, constantly cite the survey results—or studies based on those results—in making public policy arguments. However, the survey data are of insufficient quality and timeliness to support many of the demands put on them. The information provided to SRS is often a rough estimate, frequently based on unexamined assumptions that originated years earlier.” (National Research Council 2010, p. 1, emphases added)
We don’t know WHO is funded

Information Collection, Analysis and Dissemination

The working group was frustrated and its activities were held back throughout its study by the lack of comprehensive data regarding biomedical researchers. The timeframe and resources of the study did not allow for comprehensive data collection or the implementation of a comprehensive model of the biomedical workforce. It is evident from the data-gathering and analyses undertaken by the working group that there are major gaps in the data currently being collected on foreign-trained postdoctoral researchers and those who work in industry.

In general, there are two fundamental types of data that are valuable – aggregate-level data and individual-level longitudinal data – and two broad sources for each type of data – administrative records and surveys. Aggregate level data is necessary to determine the number of people in various positions, but individual-level data and especially longitudinal individual data, in which individuals are tracked over time, makes it possible to identify the characteristics and trajectories of individuals and is important for rigorous modeling and evaluation. Survey data is valuable for addressing targeted questions, especially questions that do not leave accessible administrative records, but is costly to collect and therefore usually collected on limited samples and frequently it takes a long time until these data are ready for analysis. Administrative data can cover an entire population and be obtained through near real-time feeds. Our ability to use administrative data is improving rapidly and NIH already has and is continuing to develop a wealth of such data (e.g. the STAR METRICS project).
And we don’t know the RESULTS

Our Economy

NIH research is a powerful economic engine, investing more than $31 billion annually in medical research for the American people. In fiscal year 2011, NIH-funded research supported an estimated 432,000 jobs all across the United States.

The economic impact of NIH does not end there. It has been estimated that every $1 of NIH funding generates about $2.21 in local economic growth. Also, discoveries arising from NIH-funded research serve as a foundation for the entire U.S. biomedical industry. Long considered the world’s leader in innovation, that vital sector exports an estimated $90 billion in goods and services annually and employs 1 million U.S. citizens with wages totaling an estimated $84 billion.

Consider the economic payoff of just one NIH-supported research initiative: the successful effort to read all the letters in the human DNA instruction book. The U.S. government’s $4 billion investment in the Human Genome Project spurred an estimated $796 billion in economic growth from 2000-2010—a 141-fold return on investment, after adjusting for inflation.
A conceptual framework
Using Big Data

Source: Ian Foster, University of Chicago
What skills are needed

- **Web Scraping**
- **Text Analysis**
- **Networks**
- Link to Census and USPTO Admin data
- **Statistical Inference**
- **Longitudinal Database**
- **Visualization**

**Data Capture**
- Structured Data (different ways of collecting data)
- Text data (making use of new types of data)
- Relationship data (creating new units of economic activity)
- Output data (creating innovation measures)

**Data Analysis**
- Sense making (from the population)
- Analysis (capturing social and economic dynamics)
- Conveying information (in statistically valid context)
Overview

• Key ideas
• Definitions and motivation
• Application: Science of Science and Innovation Policy
  • Text analysis
  • Linking
  • Units of analysis
• Privacy and Confidentiality
Different Text analytics paradigms: Best is to use statistical machine learning augmented with lexicons and linguistics (Rayid Ghani)

<table>
<thead>
<tr>
<th>Description</th>
<th>Do it by hand</th>
<th>Hire linguists</th>
<th>Do it the right way</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rules based on lists of words</td>
<td>Rules using words and linguistic operators (parts of speech for example)</td>
<td>Statistical approaches that can be trained and learn over time. Can incorporate lexicons and linguistics as well</td>
<td></td>
</tr>
<tr>
<td>Ease of creation &amp; maintenance</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td>Context Sensitiveness</td>
<td>Low</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Interpretability</td>
<td>High (unless the rules get large)</td>
<td>Medium</td>
<td>Medium</td>
</tr>
</tbody>
</table>
Overview

- Key ideas
- Definitions and motivation
- Application: Science of Science and Innovation Policy
  - Text analysis
  - Linking
  - Units of analysis
- Privacy and Confidentiality
Institution

Agency Budget

Agency

Award

Record

Institution

Award Funding

Endowment Funding

Financial System

Disbursement

Research Project

Start-Up

Papers

Patents

HR System

Procurement System

Subcontracting System

Financial System

Personnel

Vendor

Contractor

Hire

Buy

Engage

Existing Institutional Reporting

State

Download
Overview

• Key ideas
• Definitions and motivation
• Application: Science of Science and Innovation Policy
  • Text analysis
  • Linking
  • Units of analysis
• Privacy and Confidentiality
Unit of analysis: Networks

Figure 6: Ego network for a single faculty member

Width of ties represents number of grants in common

Figure 7. Sample Walktrip Community

Grad Student 17638759 – Ego network includes 18 total partners of whom 1/3 are not members of this community.

Grad Student 29908482 – Ego network includes 4 total partners of whom half are not members of this community.

Size of nodes is proportional to an individual’s total network degree
Unit of analysis: Projects

Research group composition: average per PI
Most Purdue matches were Faculty members who performed Consulting and Educational Services.
Purdue Summary Statistics

By State, Purdue employees were more diverse than California, but majority remained in Indiana
Within Indiana, most chose to remain near Lafayette (not Indianapolis)
Overview

- Key ideas
- Definitions and motivation
- Application: Science of Science and Innovation Policy
  - Text analysis
  - Linking
  - Units of analysis
- Privacy and Confidentiality
Access for Research is Critical
Core Questions

- What is the legal framework?
- What is the practical framework?
- What is the statistical framework?
Privacy, Big Data, and the Public Good
Frameworks for Engagement

Edited by Julia Lane
Victoria Stodden
Stefan Bender
Helen Nissenbaum
Key ideas

• Opportunity
  • New units of analysis
  • New sources of data
  • New types of analysis

• Challenges
  • New sets of skills
  • Linking
  • Privacy and confidentiality

• What needs to be done by APDU community
  • Training and Analysis
  • Analysis
  • Use
And a reminder of why