Better Transnational Access and Data Sharing to Solve Common Questions

Julia Lane
American Institutes for Research
University of Strasbourg
University of Melbourne

Overview

- Common Questions
- New kinds of data
- New kinds of data challenges
- Core access issues
  - Framework
  - Practical Approaches
- Possible next steps

Common Questions

- Employment and Unemployment
- Ageing Populations
- Human Dimensions of Climate Change
- Food Security
- Energy Security
- Science and Innovation

New Kinds of Data

(Lee Giles)

Jim Gray's paradigm

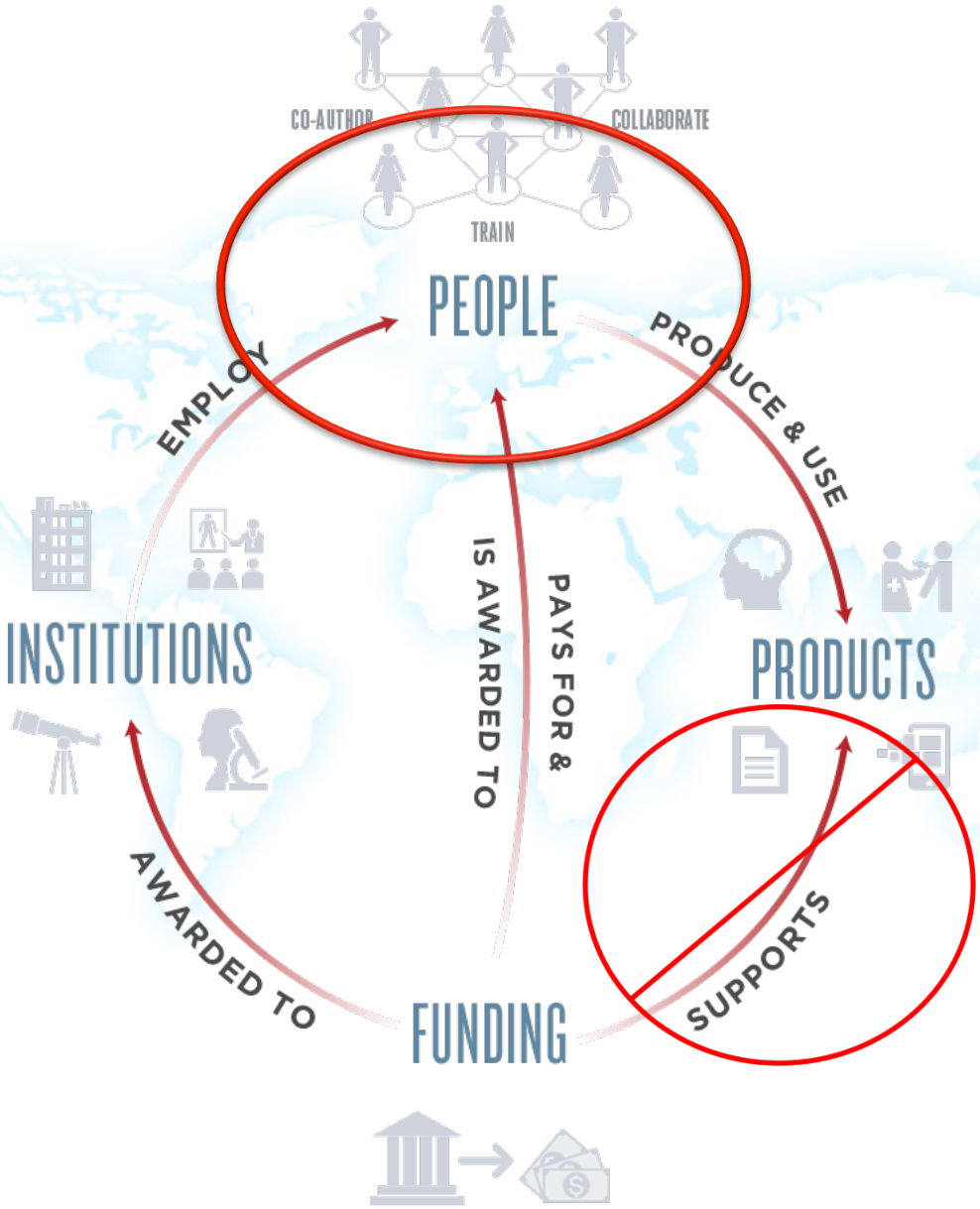
- **Observational Science**
  - Scientist gathers data by direct observation
  - Scientist analyzes data
- **Analytical Science**
  - Scientist builds analytical model
  - Makes predictions.
- **Computational Science**
  - Simulate analytical model
  - Validate model and makes predictions
- **Data Exploration Science**
  - **Data-driven science**
    Data captured by instruments from the web, or data generated by simulation
  - Information extraction
  - Processed by software
  - Placed in a database / files
  - Scientist(s)/scholar(s) analyze(s) database / files
  - *Access crucial*

Training to Climb an Everest of Digital Data

By ASHLEE VANCE
Published: October 11, 2009

MOUNTAIN VIEW, Calif. — It is a rare criticism of elite American university students that they do not think big enough. But that is exactly the complaint from some of the largest technology companies and the federal government.

Big Data in Science and Innovation Policy



Classic Data Issues

- Administrative data (HR records) with sensitive information
- Publicly available data
  - CV
  - USPTO
  - Publications
  - Twitter…..
- Statistical office data
  - Startups
  - Placements (earnings, industry)

Classic data challenges

- What is the legal and regulatory framework within which data are being collected, used, and shared?
- What does informed consent mean in the new environment?
- Do people “own” the data that are collected on them?
- What are researchers’ and institutions’ duties to protect the data held?
- How do we design effective studies within the context of big data with confidentiality concerns?
- How can we facilitate verifiability in findings?
- And what are the practical options? i.e. what can I tell the Institutional Review Boards

Privacy, Big Data and the Public Good: Frameworks for Engagement

Julia Lane
Victoria Stodden
Helen Nissenbaum
Stefan Bender

Book Goals

- Identify ways in which vast new sets of data on human beings can be collected, integrated, and analysed to improve evidence based decision making while protecting confidentiality
  - expand the frontier of research in the legal, statistical and operational aspects of confidentiality protection.
  - provide both a theoretical and practical foundation which federal and municipal authorities can draw from in establishing their data access rules and data security procedures.

Identifying the Need and the Tensions:.

- *Authors: Steve Koonin and Mike Holland, NYU*
- How does understanding and exploiting data
  - create economic value,
  - allow governments to become more effective
  - reduce reporting burdens
- Addresses the inherent tradeoff between data utility and data privacy in the context of the types of data that are now being collected.
- Comment upon institutional contexts, e.g., identical collections and analyses are perceived very differently if done in a governmental versus commercial context.

Information Extraction

- Authors: Andrew Gelman, Columbia and Philip Price, LLNL
- Addresses issues in statistical modeling of modern “big data,”
- Emphasis on new challenges created by dataset size, diversity, and confidentiality issues for the data underlying “big data,”

The Economics of Privacy

- Author: Alessandro Acquisiti, CMU
- Addresses costs, benefits, and incentives.
  - Does behavioral economics shed light on why what people say about their desired level of privacy differs from the levels of information they disclose?
  - As increasingly pervasive technology all but sets the default to disclosure, is the level of effort to protect each disclosure so great that only exceptional individuals are willing to expend that level of effort?
  - Do younger citizens have different norms regarding privacy for their digital footprint?

Cities as data producers and users

- Author: Bob Goerge, University of Chicago
- Role of cities as unit for data collection
- Opportunities for improving information available to city agencies, in the context of the data they have and the pressures they face (internal, external, political, and policy).
  - Do municipalities generally enjoy greater citizen trust than state/provincial or national governments?
  - Do municipalities have the authority to innovate in data collection and privacy protection?
  - If municipalities don't have the authority to set privacy standards, what privacy approaches have cities used in the past and what are their levers for influencing best practices?

II. The Framework

The Social and Ethical context: What does the public know?

- Author: Helen Nissenbaum, NYU
- Lay out differing frameworks for privacy and point out practicalities of where clashes between them arise.
- Where and how public benefits of data capture and analysis (improved scientific understanding or governmental efficiency) are balanced against a negative right to privacy.
- Is the ethical calculus different for non-profit than for-profit entities?
- What challenges cannot be handled by existing law and what should the approach be?

The Legal and Regulatory Framework: What do the rules say about data capture and reuse?

- Author: Katharine Strandburg, NYU
- Do people have an ownership stake in personal data?
- Does the law treat “shedded” data (e.g., location derived from call detail records) or “observed” data (e.g., video, sensor streams) differently than “disclosed” data (e.g., administrative records, social media)?
- What does the Federal Information Processing Standard bring to bear on this?.

The Legal and Regulatory Framework: What do the rules say about data analysis?

- Author: Paul Ohm, University of Colorado
- Covers the duty to protect information held.
  - statutory and follow-on regulatory frameworks dominate,
  - case law trends
  - Discussion of challenges that cannot be handled by existing law and what might be potential approaches.

The Legal and Regulatory Framework: What do the rules say about scientific research and reproducibility?

- Victoria Stodden, Columbia and Stanford
- What are possible solutions that could bridge the gap between open access to data and an access-blocking NDA?
- What might templated data sharing agreements between academic researchers and data producing companies look like?

Section III. Protecting Data while Still Using Data

The Analytical Framework for data release: a statistical view

- Authors: Alan Karr and Jerry Reiter, NISS and DUke
- What is the statistical framework that can be used to determine the risk of reidentification?
- What is the statistical framework that can be used to protect new types of data?
- What are the consequences for the utility of data analysis?
- Which approaches that have been used to limit disclosure of survey data can be applied to “big data”?
- What are the gaps in our knowledge (ie what is the future research agenda)?

The Analytical Framework: Portable approaches to informed consent and open data

Author: John Wilbanks, Sage

- What frameworks are available to permit data reuse?
- How can legal and technical systems be structured to allow people to donate their data to science?
- What are appropriate methods to repurpose traditional consent forms so that user-donated data can be gathered, deidentified and syndicated for use in computational research environments

The analytical framework for data: a cryptographic view

- Author: Cynthia Dwork, Microsoft
- How can differential privacy be applied to help determine the risk of reidentification in big data?
- How can the differential privacy approach help inform data protection in the context of big data?
- What constraints does the differential privacy approach put on the utility of data analysis?
- What are the gaps in our knowledge (ie what is the future research agenda)?

The Operational framework: Institutional Controls

- Authors: Daniel "Dazza" Greenwood; Arek Stopczynski; Brian Sweatt; Thomas Hardjono, Alex “Sandy” Pentland, MIT
- What practical approaches have been used for making big data accessible?
- How do you design institutions to manage the new types of privacy issues? What are the gaps in our knowledge
- What is the future research agenda?

The Operational framework: Engineered Controls

- Author: Carl Landwehr, George Washington
- What can we learn from trustworthy computing approaches to protecting data?
- What is the conceptual framework that is used to understand software vulnerabilities in high security environments?
- How might the lessons learned from cybersecurity research be used to enable access to, but protect the confidentiality of, data on human beings?

Viewing privacy through a different lens

- Author: Peter Elias, Warwick
- An overview of the European experience with big data, and examines the consequences for research and data analysis when countries operate under different legal frameworks.
- There will be a particular focus on the UK example, which has led the way in developing frameworks for access to administrative data, but has different jurisdictions.
- The primary focus is on pragmatic lessons learned from the European experience, what is necessary to be changed and how is change to be achieved

Timeline

- Book Final Draft: January 1
- Conference: March 7, CUSP
- Publication (Cambridge University Press) August
- Presentations: August, Joint Statistical Meetings
- Conference (possible): federal statistical agencies

Next steps: COMMON AGENDA

- Identify ways in which vast new sets of data on human beings can be collected, integrated, and analysed to improve evidence based decision making while protecting confidentiality
  - expand the frontier of research in the legal, statistical and operational aspects of confidentiality protection.
  - provide both a theoretical and practical foundation which INTERNATIONAL authorities can draw from in establishing their data access rules and data security procedures.

Comments and questions?

- Julia Lane
- jlane@air.org