ICPSR

SHARING DATA TO ADVANCE SCIENCE

Opportunities and Obstacles Big Data, Data Sharing and the Future of Social Science

Margaret C. Levenstein

ICPSR Director

ICPSR INSTITUTE FOR SOCIAL RESEARCH
UNIVERSITY OF MICHIGAN

Society for Prevention Research
San Francisco, California
May 30, 2019

Big data challenge and opportunity

- Opportunities
  - More timely
    - Heart rate streamed rather than measured at periodic visits
  - More granular
    - Individuals, transactions, locations, embedded in networks
  - Digital trace data created automatically
    - Survey response not necessary
- Challenges
  - Consent? Privacy? Privately owned?
  - It's big – storage and computation, having been conquered, have now re-emerged

Data Sharing

- Opportunities
  - Leverage large investment in data collection
  - Increase transparency and reproducibility of research
    - Increase trust in science
  - Facilitate knowledge building
    - Science is inherently incremental, explaining what came before as well as what is novel
- Challenges
  - It's hard, takes real resources
    - Requires more than taking down a paywall to make data FAIR
      - Findable, Accessible, Interoperable, Reusable
  - Requires protecting private interests
    - Subjects, PIs, data owners

Sharing is Caring

- If no one else can access the data, it's not science
- We know how to protect privacy – and private property
- Research subjects, and most of the general public, want to contribute to scientific progress
  - Regulation can protect from harm and profiteering while allowing scientific progress

What is to be done?

- Rules and tools
  - Standards are more effective than mandates
    - Lower costs, create new norms
  - Standards make tools easier to design
    - Tools make it possible for researchers to analyze and share
      - Lower barriers to entry
      - Lower incremental cost

What is ICPSR doing?

- What is ICPSR?
  - Preserving and accessing shared data and data-related content
    - Journal repositories and journal-related deposits
  - Training in data analysis and data stewardship
- Three new initiatives
  - LinkageLibrary
  - SOMAR
  - Researcher passport

ICPSR

- Founded in 1962 by 22 universities, now consortium of 800 institutions world-wide
- Focus on social and behavioral science data, broadly defined
- Current holdings
  - 10,000 studies, quarter million files
  - 1500 are *restricted studies*, almost always to protect confidentiality
  - Bibliography of Data-related Literature with 75,000 citations
- Approximately 60,000 active MyData (“shopping cart”) accounts
- Thematic data collections
  - Drug addiction, aging, arts, child care, education, criminal justice, demography, health and medical care, and minorities
  - Data Lumos
- Summer Program in Quantitative Methods of Social Research

Preserving and accessing shared data and data-related content

- Make data sharing feasible
  - ICPSR's General Archive
    - Anyone can deposit
    - Curated and preserved
  - Guidance over data life cycle
    - Templates for consent, IRB, DMP consistent with transparent and reproducible access
- Incentivize data sharing
    - Standard citation
    - Bibliography
    - Usage statistics

Data linkage challenges

- Data in the Wild
  - Often requires linking data from different sources
- Linkage more accurate with more detailed information
  - Need standards for safe, ethical ways to enhance data with new linkages
- Linked data easier to re-identify, even after removing unique identifiers
  - Need safe places to analyze linked data
- Linkage strategies introduce differences in datasets that are often not well documented


LINKAGE LIBRARY
Maintaining datasets to support the data linkage community

LINKAGE LIBRARY

- Enable researchers to share linked (or linkable) data and linkage strategies
  - Algorithms, code
- Compare approaches across projects, datasets, disciplines
  - Improve linkage practices
  - Improve transparency
- Build data community
  - Threaded commenting among community members

Private data and data privacy

- Researchers increasingly make use of private data
  - Private because it *belongs* to a company that asserts control over it
  - Private because it contains information about individuals that they might not want to be public
- Academic journals in economics
  - Required data sharing, for transparency and reproducibility
  - Found 1/3 of empirical articles requested waiver
    - Data belonged to someone else
    - Data contained confidential information

Public data and privacy

- Increasing concern over risk that “anonymized” will be re-identified
- Driving factor in Census Bureau announcements re changes in production of public data products
  - Increasing computational power and availability of information about individuals and households
- Confidentiality protection through noise infusion rather then swapping, aggregating, suppressing
  - Noise infusion is more transparent
  - How much noise? Who gets hidden? What relationships get obscured?

Access to private data

- Long-standing arrangements
  - Each involves both a technological and a social component
  - Limit collaboration and very expensive to scale
- Local computing on secure, stand alone computers
  - Data use agreements
    - Enumerate researcher and institutional responsibilities and consequences
  - Encrypted CDs or download
  - Researcher responsible for disclosure review
- Physical enclaves
  - Data use agreements
    - Enumerate researcher and institutional responsibilities and consequences
  - Controlled computing environment
  - Third party disclosure review

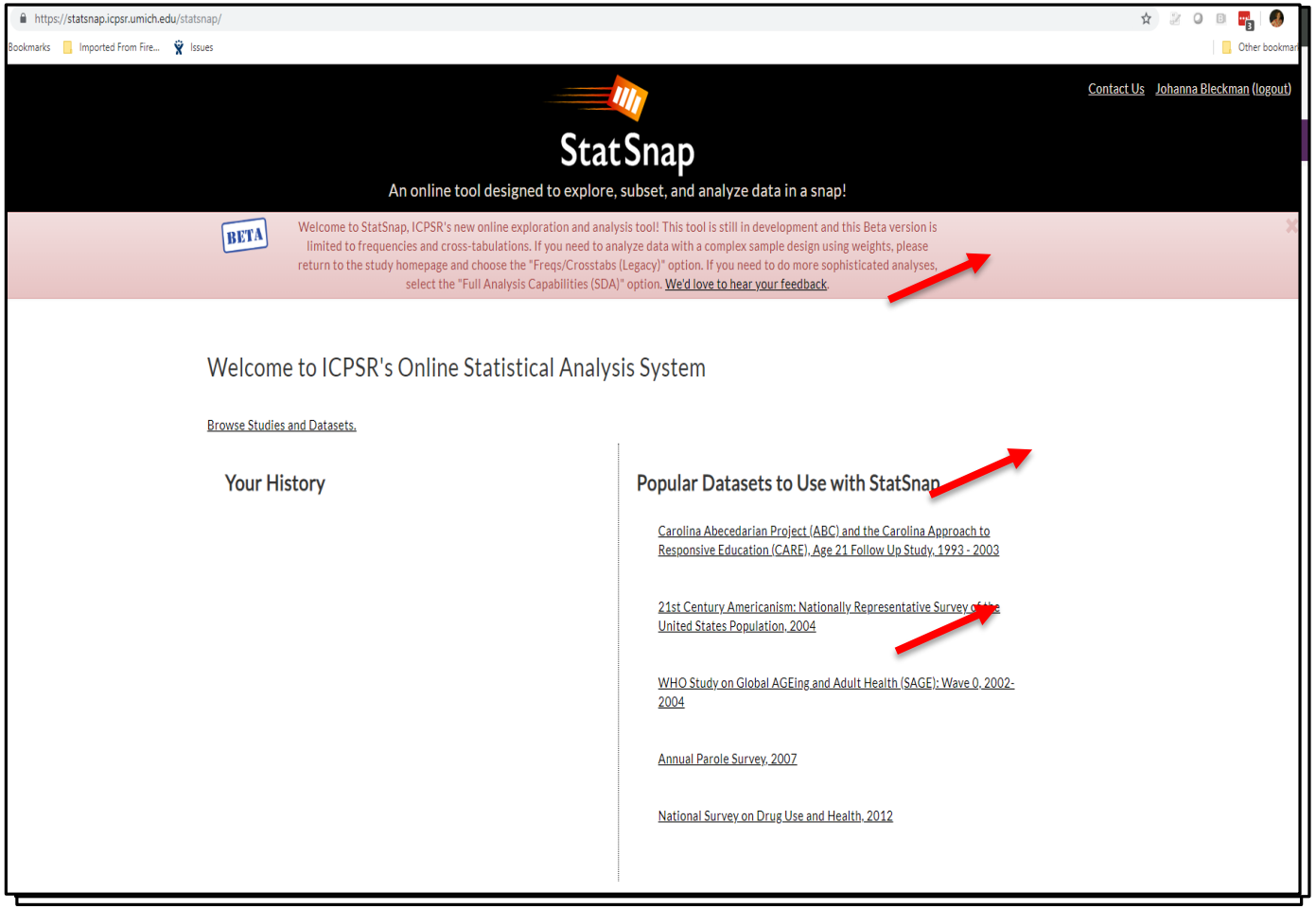
Emerging arrangements for accessing confidential data

- Virtual data enclaves
  - Data use agreements
    - Researcher Passport
  - Controlled computing environment accessed from local computer
  - Third party disclosure review
- Secure on-line computing
  - Analysis of data that the researcher cannot see
  - Automated disclosure review, with minimally necessary noise infusion
  - Secure multi-party computing
    - Computationally very intensive
  - Requires highly processed and interoperable data
    - Difficult to use with non-designed data without large up-front investment whose appeal is
      - Digital traces of human activity
      - Available essentially immediately

Researcher Passport

- Researcher Passport: Improving Data Access and Confidentiality Protection
  - ICPSR's Strategy for a Community-normed System of Digital Identities of Access
  - https://deepblue.lib.umich.edu/handle/2027.42/143808
  - Identifies inconsistent language and policies that impede access
- Passports for safe people
  - Verified identities, institutional affiliation
  - Training
  - Experience (good and bad)
- Visas to control access
  - Permission to "enter" (access) specific data specifying
    - Passport holder
      - Project, Place, Period

Researcher Passport and Radius




SOMAR
Social Media Archive @ ICPSR

SOMAR: Social Media Archive

- Addresses 4 communities who:
  - Study social media use specifically
  - Leverage social media data to understand people and society
  - Study social science methods
  - Investigate new methods for curation, publication, confidentiality and quality assessment, and long-term management of research data
- Archive enables historical and longitudinal analyses often missing from rapidly changing social medial platforms

SOMAR: Social Media Archive

- Archive data where possible
- Archive workflows and code where data sharing is prohibited
  - Eg: Twitter IDs and code for rehydrating
- Curation and metadata
  - Provenance, dates, hashtags, confidentiality protection

SOMAR Challenges

- Technical infrastructure
- Ethical and legal infrastructure
- Metadata enhancements
- Adoption

Building models of access to data

- Trusted intermediaries
  - Credentialed researchers
  - Privacy protecting technologies
  - Cooperation from data custodians?
- Public sector
  - Foundations for Evidence-Based Policy Act of 2018
    - Federal Statistical Research Data Center network
  - State and local governments
    - Patchwork of arrangements
- Private sector

Solutions?

- Templates and standards
  - For agreements
  - For data and meta-data
  - For transmission
  - Universities, funders, learned societies, journals must support standards
- Credible burden reduction by leveraging business information systems
- Trusted intermediaries
  - Archive and access stale data for research

Lessons

- Be not afraid
  - Be creative in your use of data
- Do the right thing
  - Be ethical in your use of data
- Sharing is caring