Cornell University

Cornell University

Science, Confidentiality, and the Public Interest

John M. Abowd and Lars Vilhuber
Cornell University

Cornell University

Acknowledgements

- Synthetic data development was supported by NSF grants SES-0427889, CNS-0627680, SES-0922005, and SES-0922494 , BCS-0941226, TC-012593, and by the U.S. Census Bureau.
- The Synthetic Data Server is supported by NSF grant SES- 1042181.
- The opinions expressed in this presentation are those of the authors and not Cornell University, the National Science Foundation nor the Census Bureau.

Outline

- Why are the services of statistical agencies public goods?
- Public goods influence what statistical agencies do with confidential data
- Research benefits the public interest and the agency
- Access modalities and synthetic data
- The scientific feedback loop and the public interest
- How well does this work in practice?
- Final thoughts

Why are the services of statistical agencies public goods?

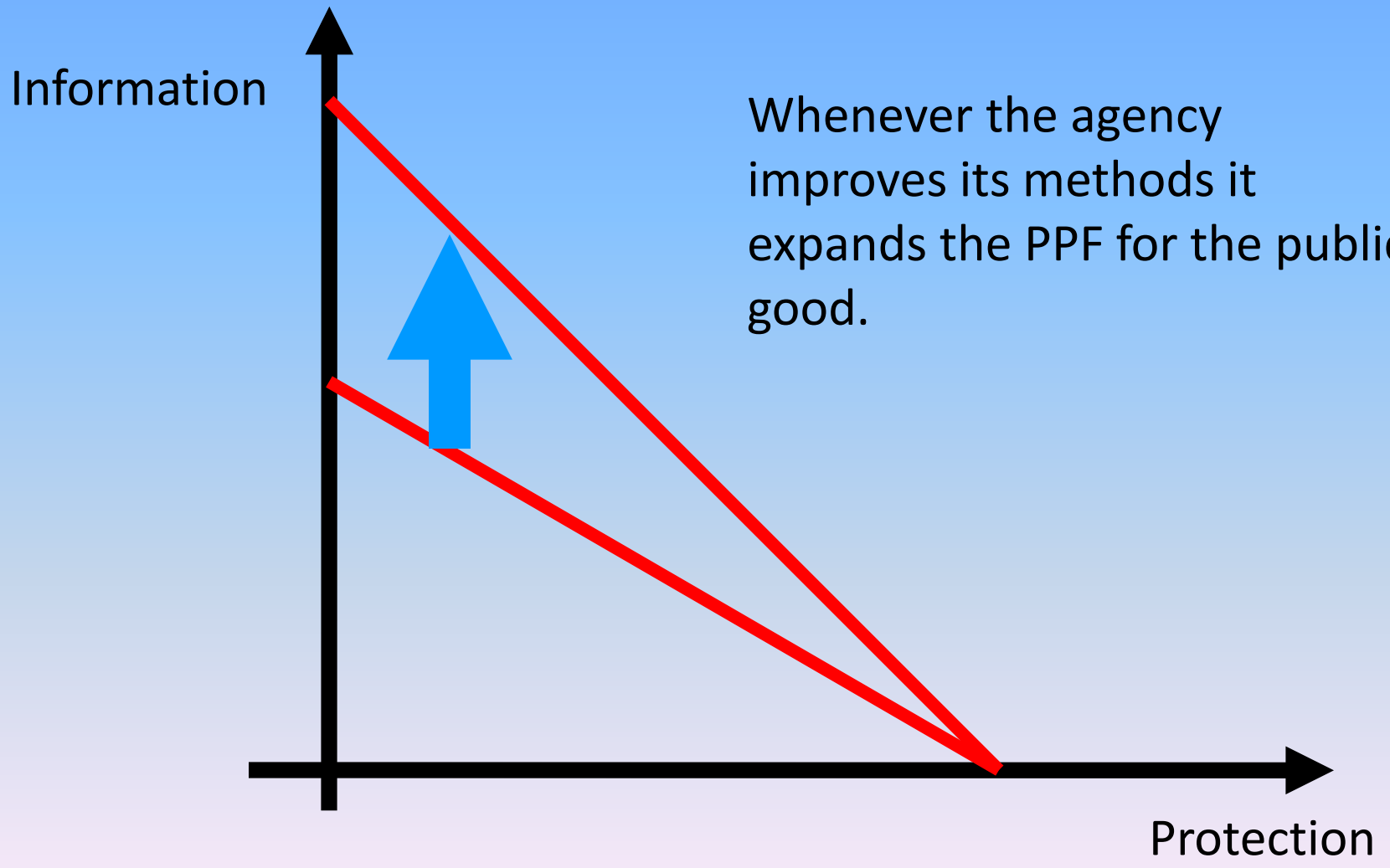
- Many individuals and businesses benefit from the published data without reducing any other user's benefits (inexhaustible)
- Once the data have been published it is extremely difficult to control their use (non-excludable)
- These are the classic characteristics of a public good

Recent Examples of Data as a Public Good

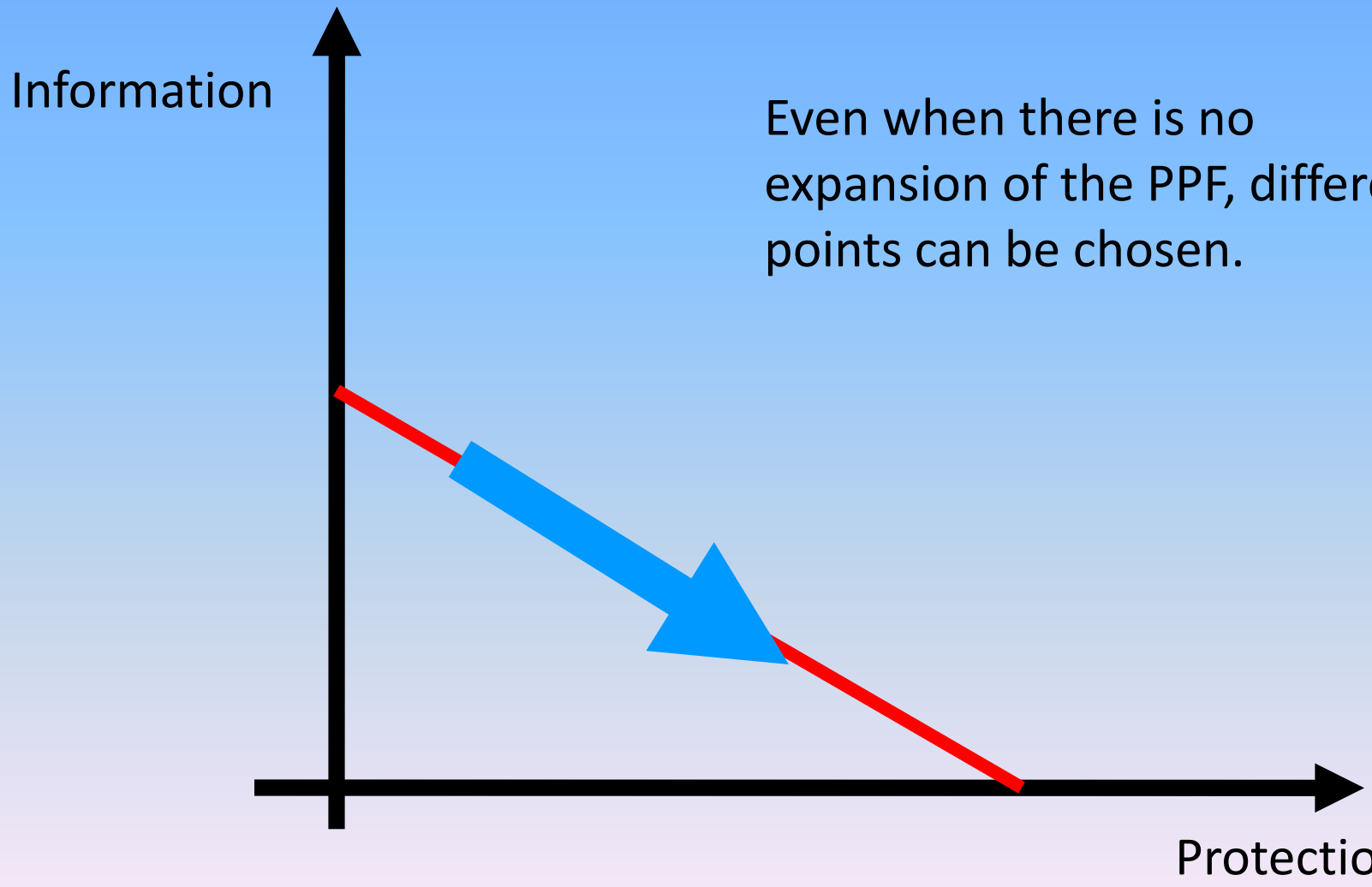
- "Hiring in U.S. Slowed in May With 54,000 Jobs Added" (New York Times, June 3, 2011) – statistics provided by the U.S. Bureau of Labor Statistics
- "Ehec epidemic in Germany: where are the victims, and how many" (Stern, June 15, 2011) – statistics provided by state agencies in Germany
- "Portugal Social Democrats set to win election-exit polls" (Reuters, June 5, 2011, 3:02 PM)

Public Goods Influence What Statistical Agencies Do with the Confidential Data

- Pure statistical agencies don't enforce any laws
- Their only reason for existence is to provide the public good
- They should maximize the output of this good (information)
- But they are constrained by their confidentiality protection statutes and pledges
- Creates a classic Production Possibility Frontier tradeoff (R/U graph in statistics)



Whenever the agency improves its methods it expands the PPF for the public good.



Even when there is no expansion of the PPF, different points can be chosen.

Research Benefits the Public Interest and the Agency

- Research re-uses the existing data looking for relationships, trends and models that were not necessarily anticipated when the data were collected
- The science, placed in the public domain, provides the same public good benefits as other data publication
- And the use directly improves the data

Social Benefits from Research on Confidential Data (biased towards Labor Economics)

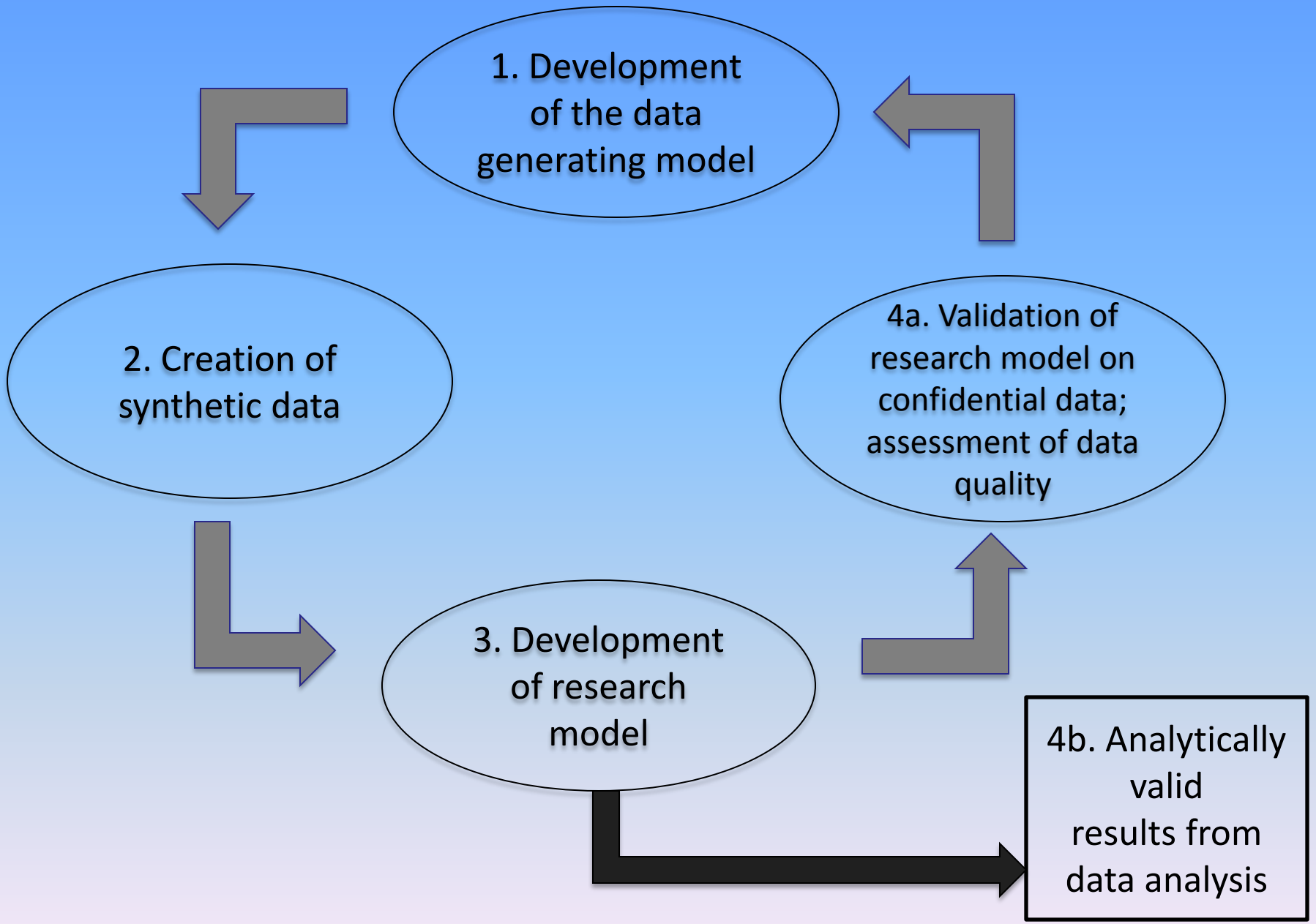
- Gross flows into/out of unemployment (confidential CPS data)
- Worker flows (confidential establishment data)
- Job flows (confidential establishment data)
- Excess flows/churning (linked individual and establishment data)
- Wage structure heterogeneity

Access Modalities

- Physical access: external researcher on site in the agency
- Remote access by proxy: external researcher works in a purpose-built facility
- Remote access: job entry and output delivery
- And now, synthetic data: model development with synthetic data, then analysis of actual data

The Scientific Feedback Loop and the Public Interest

- Users of Synthetic Data Servers pursue normal scientific aims
- Feedback of those models improves the synthetic data
- Making more accurate releases possible
- And helping to improve the protection technology relative to direct access to the confidential data


1. Development of the data generating model
2. Creation of synthetic data
3. Development of research model
4a. Validation of research model on confidential data; assessment of data quality
4b. Analytically valid results from data analysis

Two Major Projects

- Survey of Income and Program Participation linked to IRS and SSA data
  - Now in its second full release as synthetic data
  - Users may have their models fit on the confidential data
- Synthetic Longitudinal Business Database
  - First synthetic establishment micro-data
  - Now in its initial release

SIPP Synthetic Data Projects

- The effect of Social Security benefit rules on the timing of divorce; and wage gaps before and after divorce
- Trends in different measures of earnings volatility and instability; wage flexibility
- Associations between individual characteristics (including wages), uninsured spell lengths, and transitions into and out of health insurance
- Poverty of older women, using the full earnings and marital history
- Transitions over time between segments of the family income distribution and the family earnings distribution
- The male marriage premium
- Optimal extent and timing of pre- and post-tax savings, as well as the timing of Social Security distributions
- The synthetic SIPP data were also analyzed for their data-generating features
  - An illustration of a data sharing procedure that involves multiple parties (leveraging the fact that to put the data together required multiple agencies to share data)
  - Multi-component hypothesis tests for use with partially synthetic data imputed in two stages (leveraging the fairly unique method of generating the synthetic data.

Synthetic LBD Projects

- The five projects that have started to use the Synthetic LBD since its release to the Cornell Synthetic Data Server in May 2011 have all focused on establishment and firm lifecycle issues
- Some have had a stronger focus on using the synthetic data as a collaboration tool (among Census and non-Census researchers)
- Others as a tool to prepare the analysis of the confidential data
- While the use of the synthetic data for its statistical properties dominates, it is not the only use that researchers have found for these data

Cornell University

Search Cornell

VirtualRDC News @ Cornell

Announcement

Synthetic Longitudinal Business Database (SynLBD) now available

March 4th, 2011

The Synthetic LBD Beta Data Product (SynLBD) is now available for access through the VirtualRDC's Synthetic Data Server here at Cornell University. [More »]

Posted in Social Science Gateway, Synthetic Data Server | Comments Off

Social Science Gateway available

May 31st, 2010

The Social Science Gateway @ VirtualRDC is now available and online.

- **If you want to learn about the Social Science Gateway (SSG), we have more information here.**
- **If you already know you want an account, then go straight to the "Requesting a SSG account" page.**

Site search

Search for this text:

Search

Site Navigation

Front page
open all | close all

- General Information
- Data @ VirtualRDC
- Available resources
- Classes and Tutorials
- Documentation
- NSF-Census-IRS Worksh..
- Social Science Gateway
- Synthetic Data Server
- Getting Help

Subscribe to email notifications

Enter your email address:

VirtualRDC News @ Cornell

Synthetic Data Server

The Synthetic Data Server (SDS) was set up to provide early access to new synthetic data products by the U.S. Census Bureau. These datasets are made available to interested researchers in a controlled environment, prior to a more generalized release.
At present, two datasets are made available on this server:

- SIPP Synthetic Beta (SSB) v4 and v5
- Synthetic LBD (SynLBD)

How to request access to the data is described on the next page. Access requests are reviewed for feasibility, but not otherwise restricted.
Once the data provider has signed off on the access, you will receive an account creation message from us with further instructions.

History

The server replaces the previous SIPP Synthetic Beta (SSB) server.

Funding acknowledgement

The SDS is funded through NSF grant SES-1042181.

Site search
Search for this text:
Search

Site Navigation
Front page
open all | close all
General Information
Data @ VirtualRDC
Available resources
Classes and Tutorials
Documentation
NSF-Census-IRS Worksh..
Social Science Gateway
Synthetic Data Server
Getting Help

Subscribe to email notifications
Enter your email address:

Final Thoughts

- Scientific benefits now well established
- Participation of research community in strengthening the confidentiality protections is important
- Because this improves the usefulness of the data and is safer than some current access modalities (physical access and remote access by proxy)