Broadening data access through synthetic data

Lars Vilhuber[1]

[1] Labor Dynamics Institute, ILR, Cornell University

NCRN Meetings Spring 2015 @ NAS

with input from John Abowd (NCRN, Cornell), Jerry Reiter (NCRN, Duke), Luke Shaefer (NCRN, Michigan), Saki Kinney (NISS), Jörg Drechsler (IAB Germany), Javier Miranda, Martha Stinson, Gary Benedetto, Lori Reeder (Census Bureau). Support through NSF Grant SES-0820349, SES-0922005, SES-1042181, **SES-1131848**, and Alfred P. Sloan Foundation grant G-2015-13903.

Disclaimer

- Part of the research results were obtained while Vilhuber was a Special Sworn Status researcher of the U.S. Census Bureau at the Center for Economic Studies. All results have been screened to insure that no confidential data are revealed.
- Research results and conclusions expressed are those of the authors and do not necessarily reflect the views of the Census Bureau.

Outline

Setting the stage

Contributions

Conclusion

Scope of presentation

For multitaskers: **goo.gl/zJprV**

Two synthetic datasets...

- **Survey of Income and Program Participation (SIPP) Synthetic Beta** (v4 released in 2009, v5 2010, v6 in March 2015) [SSB]
- Synthetic Longitudinal Business Database (v2 released 2011) [**SynLBD**]

... or methods ...

- SynLBD methodology applied to US, German, Canadian data (ongoing)

... lessons learned

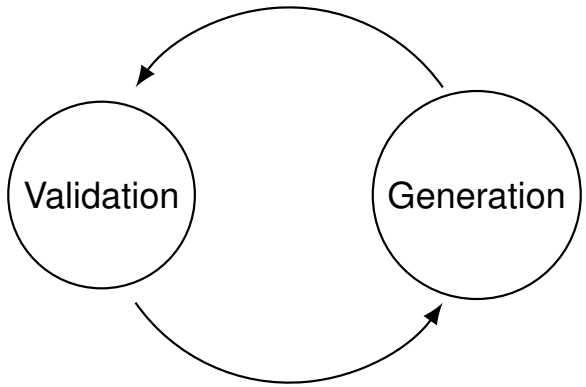
- from **Synthetic Data Server** [**SDS**] at **Cornell** (since 2010)

Background

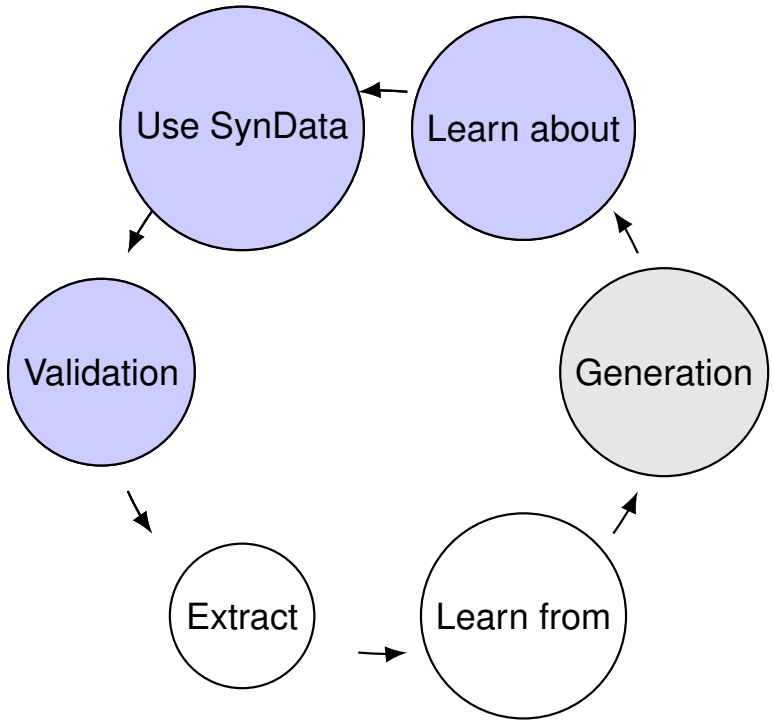
Creation of analytically valid synthetic data relies on

Synthetic data feedback loop

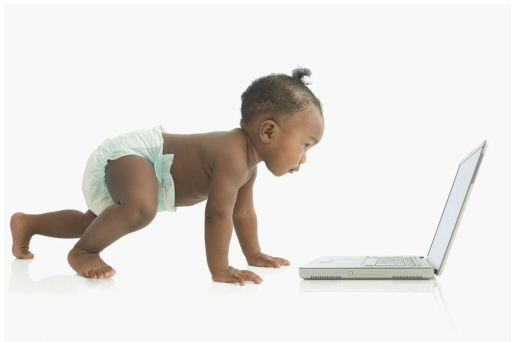
- Create synthetic data
- Models estimated on synthetic data
- Models validated on confidential data
- Lessons learned incorporated into next generation



Background


Use SynData
Learn about
Validation
Generation
Extract
Learn from

Launching synthetic data



→

→

Contributions

I will outline the following efforts, including of NCRN nodes:

Contributions

- Learning about synthetic data
- Encouraging use of synthetic data
- Facilitating validation
- International expansion

Contributions

Focus on learning

In order to get researchers to use the data, they need to know about.

Contributions

- Learn about the data
  - Data documentation
  - Provenance

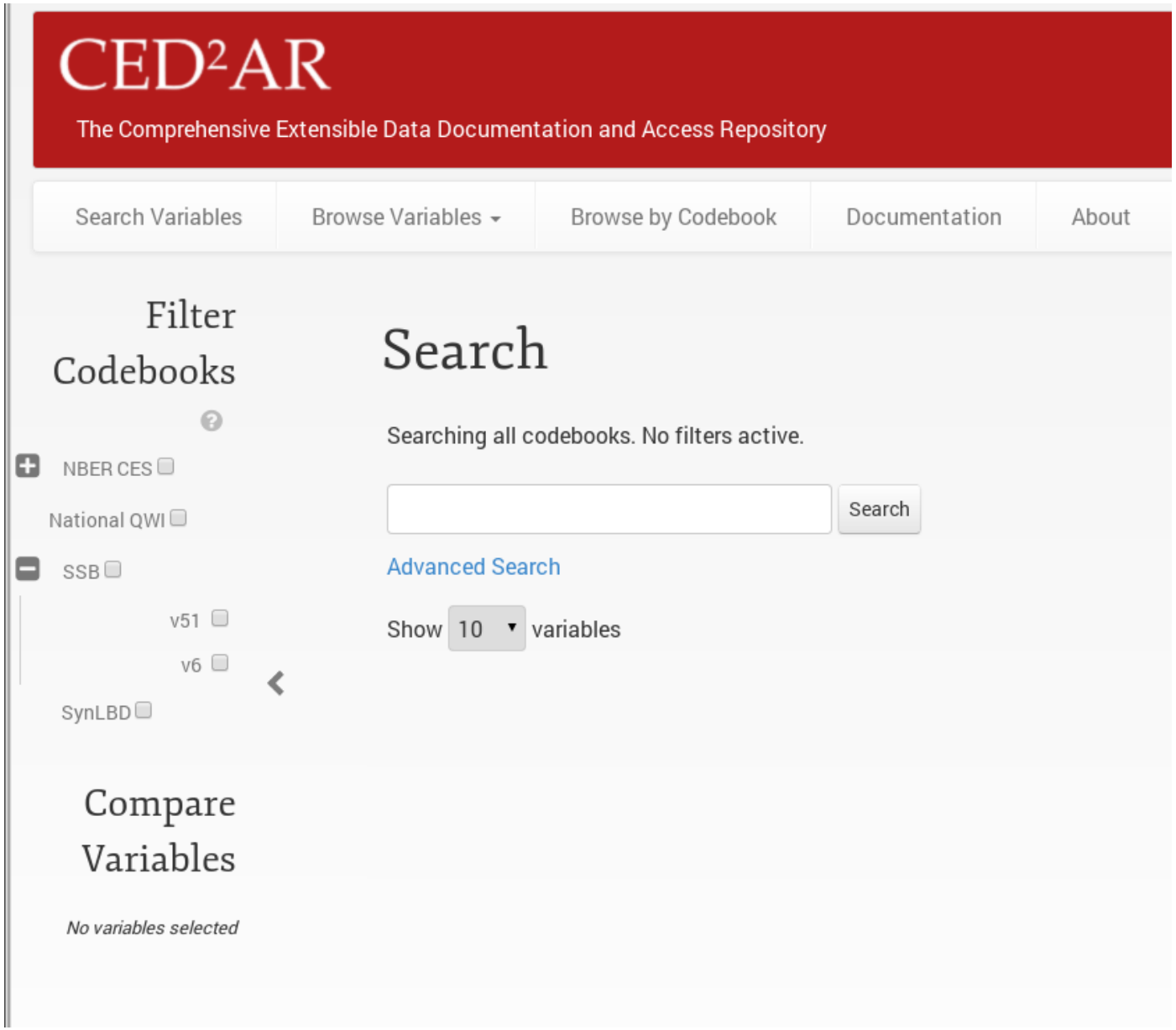
  $\rightarrow$ **CED^{2}AR codebooks (Cornell NCRN)**
- Focussed dissemination
  - Training$\rightarrow$ **NCRN: Michigan, Duke, Census**
  - Use in published research$\rightarrow$ see Cornell website
  - Presentations$\rightarrow$ many people

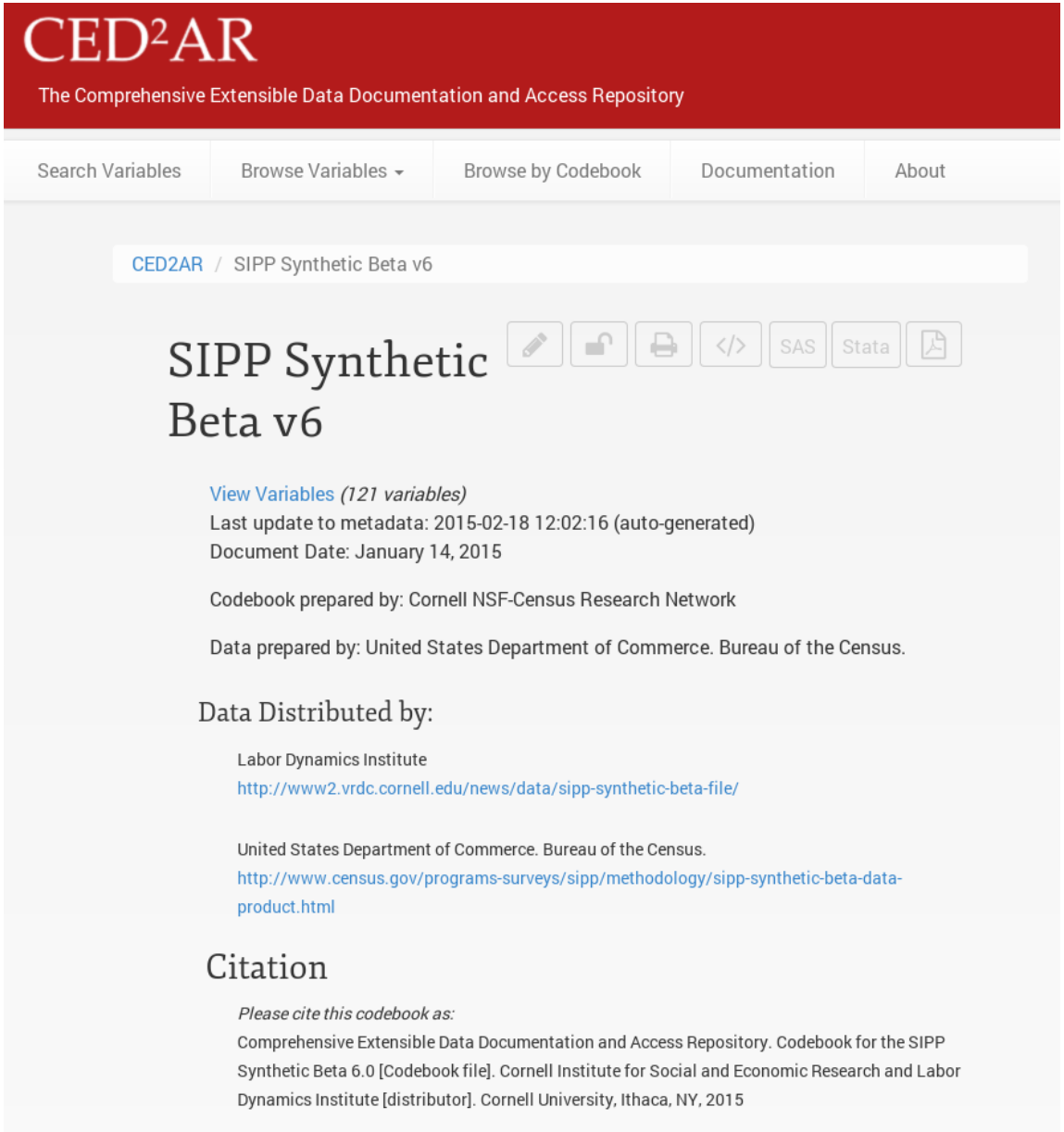
Documentation

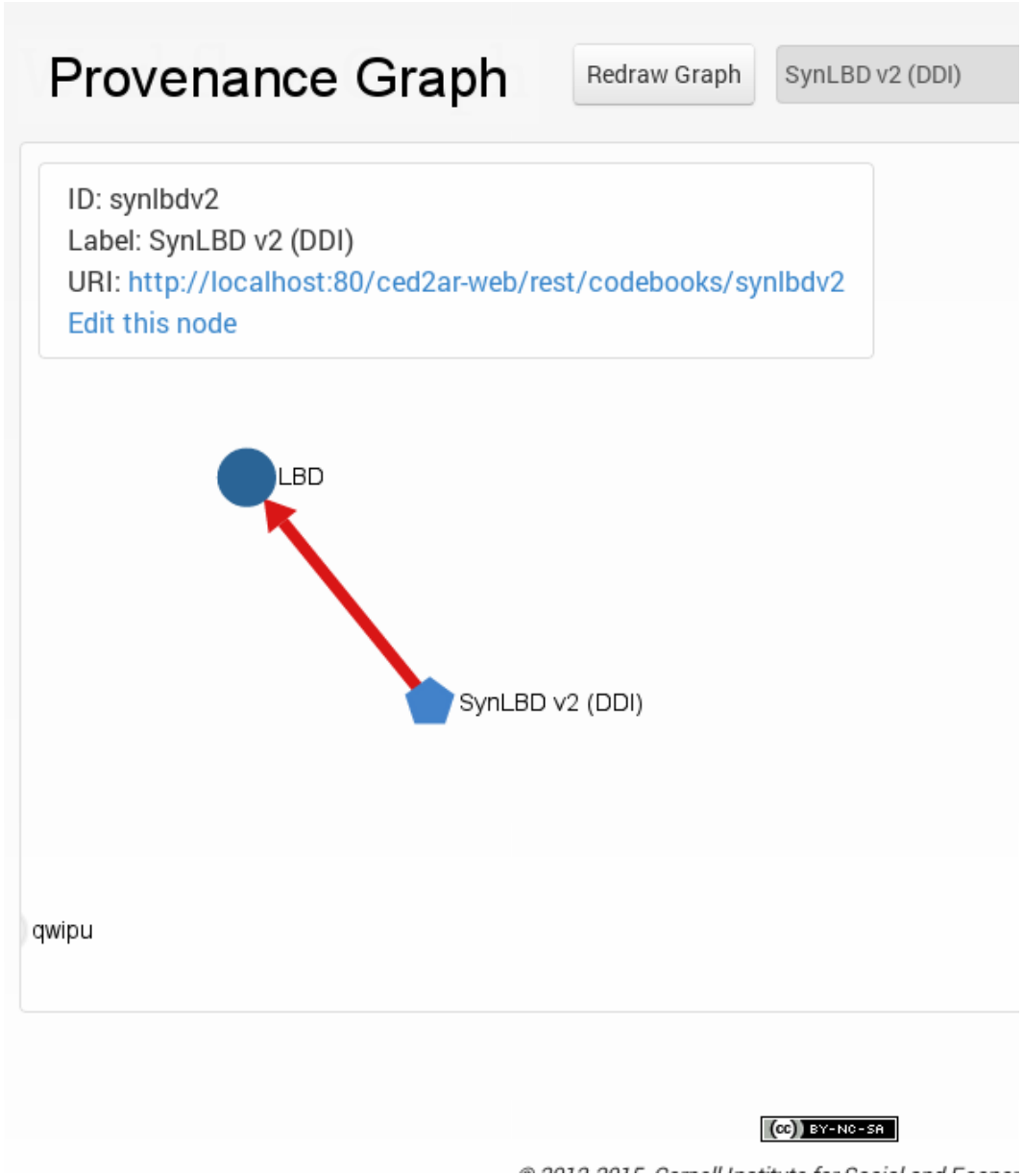
Documentation

Improving documentation

- Part of **Cornell NCRN** mission
- Improve overall availability of documentation
- Improve controlled availability of documentation on confidential data
- Maintain interoperability with other systems (use/expand/influence metadata standards)

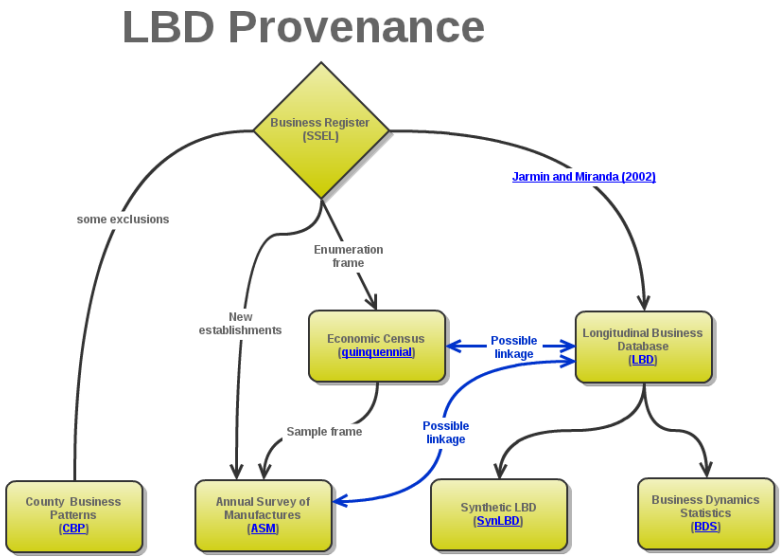

CED²AR
The Comprehensive Extensible Data Documentation and Access Repository
Search Variables
Browse Variables
Browse by Codebook
Documentation
About
Filter Codebooks
NBER CES
National QWI
SSB
v51
v6
SynLBD
Compare Variables
No variables selected
Search
Searching all codebooks. No filters active.
Search
Advanced Search
Show 10 variables


CED²AR
The Comprehensive Extensible Data Documentation and Access Repository
Search Variables
Browse Variables
Browse by Codebook
Documentation
About
CED2AR / SIPP Synthetic Beta v6
SAS
Stata
SIPP Synthetic Beta v6
View Variables (121 variables)
Last update to metadata: 2015-02-18 12:02:16 (auto-generated)
Document Date: January 14, 2015
Codebook prepared by: Cornell NSF-Census Research Network
Data prepared by: United States Department of Commerce. Bureau of the Census.
Data Distributed by:
Labor Dynamics Institute
http://www2.vrdc.cornell.edu/news/data/sipp-synthetic-beta-file/
United States Department of Commerce. Bureau of the Census.
http://www.census.gov/programs-surveys/sipp/methodology/sipp-synthetic-beta-data-product.html
Citation
Please cite this codebook as:
Comprehensive Extensible Data Documentation and Access Repository. Codebook for the SIPP Synthetic Beta 6.0 [Codebook file]. Cornell Institute for Social and Economic Research and Labor Dynamics Institute [distributor]. Cornell University, Ithaca, NY, 2015


Provenance Graph
Redraw Graph
SynLBD v2 (DDI)
ID: synlbdv2
Label: SynLBD v2 (DDI)
URI: http://localhost:80/ced2ar-web/rest/codebooks/synlbdv2
Edit this node
LBD
SynLBD v2 (DDI)
qwipu
(cc) BY-NC-SA

Goal

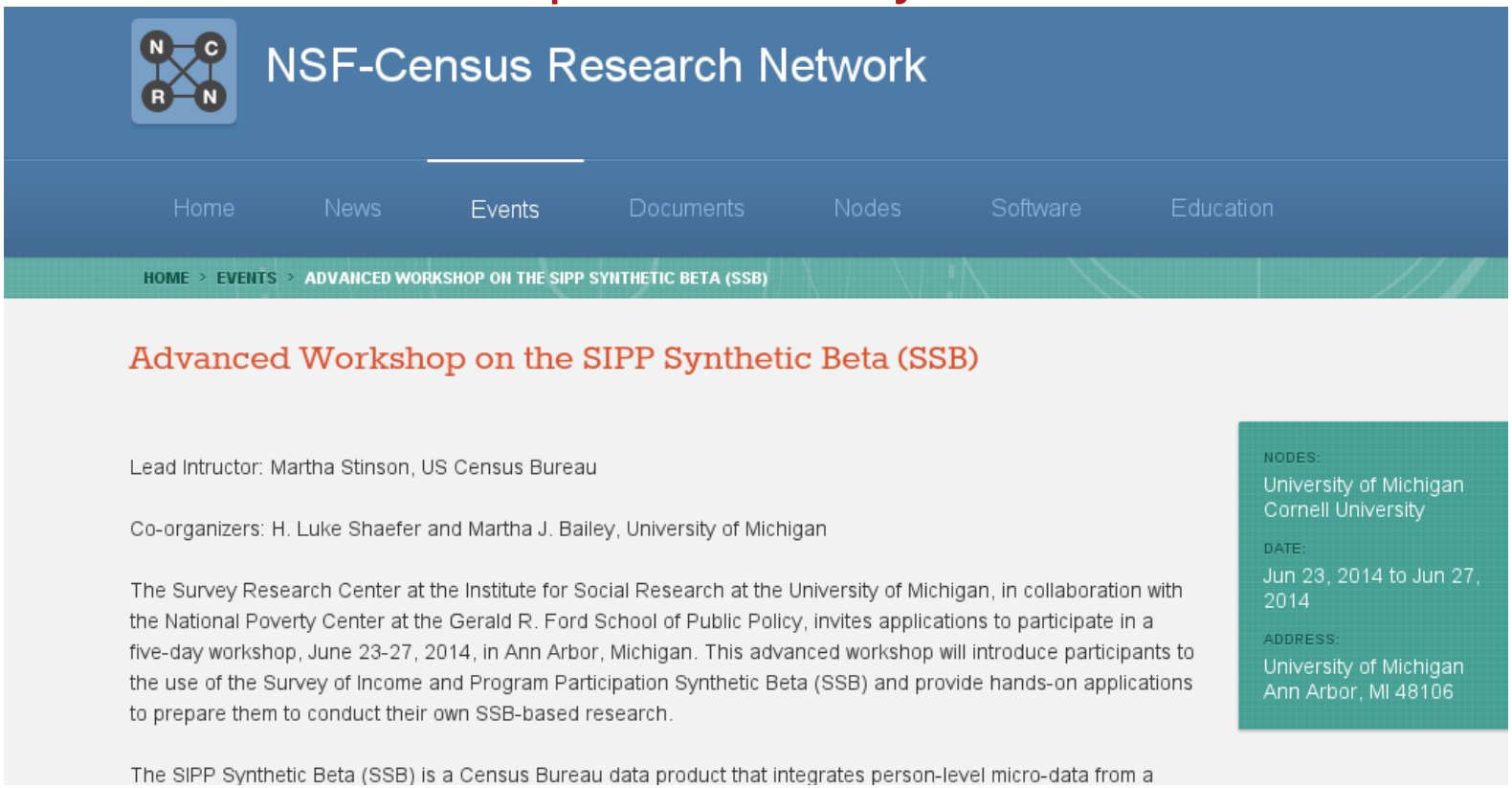
- **Derive** the provenance graph from existing information
- **Simplify** the procedures
- Work within **existing** standards



Teaching

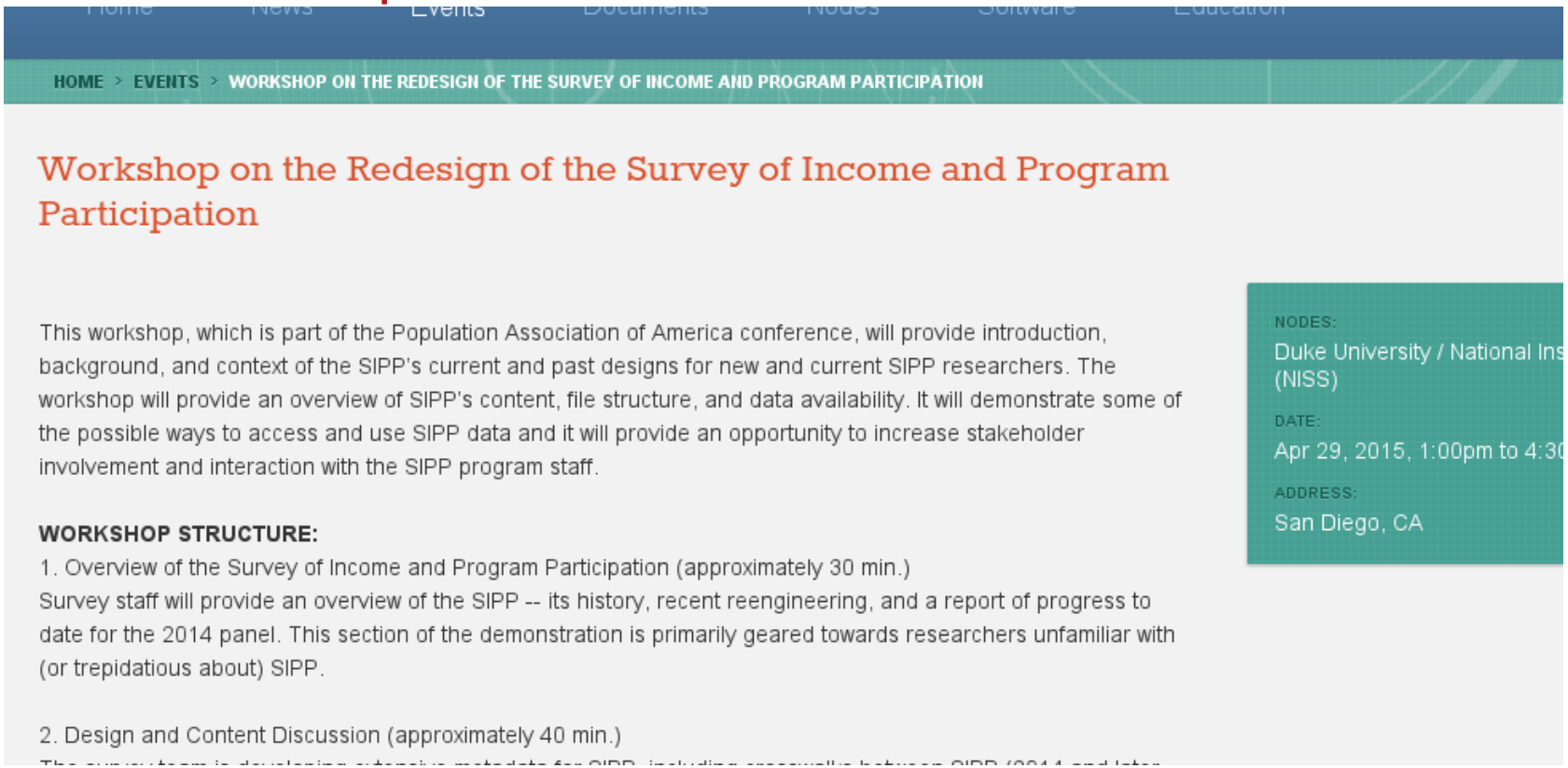
Teaching

Advanced Workshop on SIPP Synthetic Beta



Teaching

PAA Workshop on SIPP



Home News Events Documents Nodes Software Education

HOME > EVENTS > WORKSHOP ON THE REDESIGN OF THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

Workshop on the Redesign of the Survey of Income and Program Participation

This workshop, which is part of the Population Association of America conference, will provide introduction, background, and context of the SIPP's current and past designs for new and current SIPP researchers. The workshop will provide an overview of SIPP's content, file structure, and data availability. It will demonstrate some of the possible ways to access and use SIPP data and it will provide an opportunity to increase stakeholder involvement and interaction with the SIPP program staff.

NODES:
Duke University / National Ins
(NISS)

DATE:
Apr 29, 2015, 1:00pm to 4:30

ADDRESS:
San Diego, CA

WORKSHOP STRUCTURE:

1. Overview of the Survey of Income and Program Participation (approximately 30 min.)
Survey staff will provide an overview of the SIPP -- its history, recent reengineering, and a report of progress to date for the 2014 panel. This section of the demonstration is primarily geared towards researchers unfamiliar with (or trepidatious about) SIPP.

2. Design and Content Discussion (approximately 40 min.)

Teaching

Workshops on SIPP Synthetic Beta

- Part of the activities of **Michigan and Triangle NCRN node**
- Additional support from **Cornell**
- Integrated into workshops, summer schools, conferences
- SSB-oriented workshop: about 12 participants, June 2014, hands-on teaching using the **Synthetic Data Server** with core support from the Census Bureau team developing SSB.
- Demonstration of SSB during PAA on April 29, 2015

Teaching

Use of synthetic data for graduate research

- Wait times for thesis projects using confidential data may be long
- Wait times for SDS accounts substantially shorter
- Census Bureau has fast turnaround on validation
- Anecdotal evidence of substantial use of SDS projects by students
- Two theses using the synthetic data, several others in progress

Encouraging use of synthetic data

Contributions

Use Focus on use

In order to get researchers to use the data, it needs to be convenient and useful.

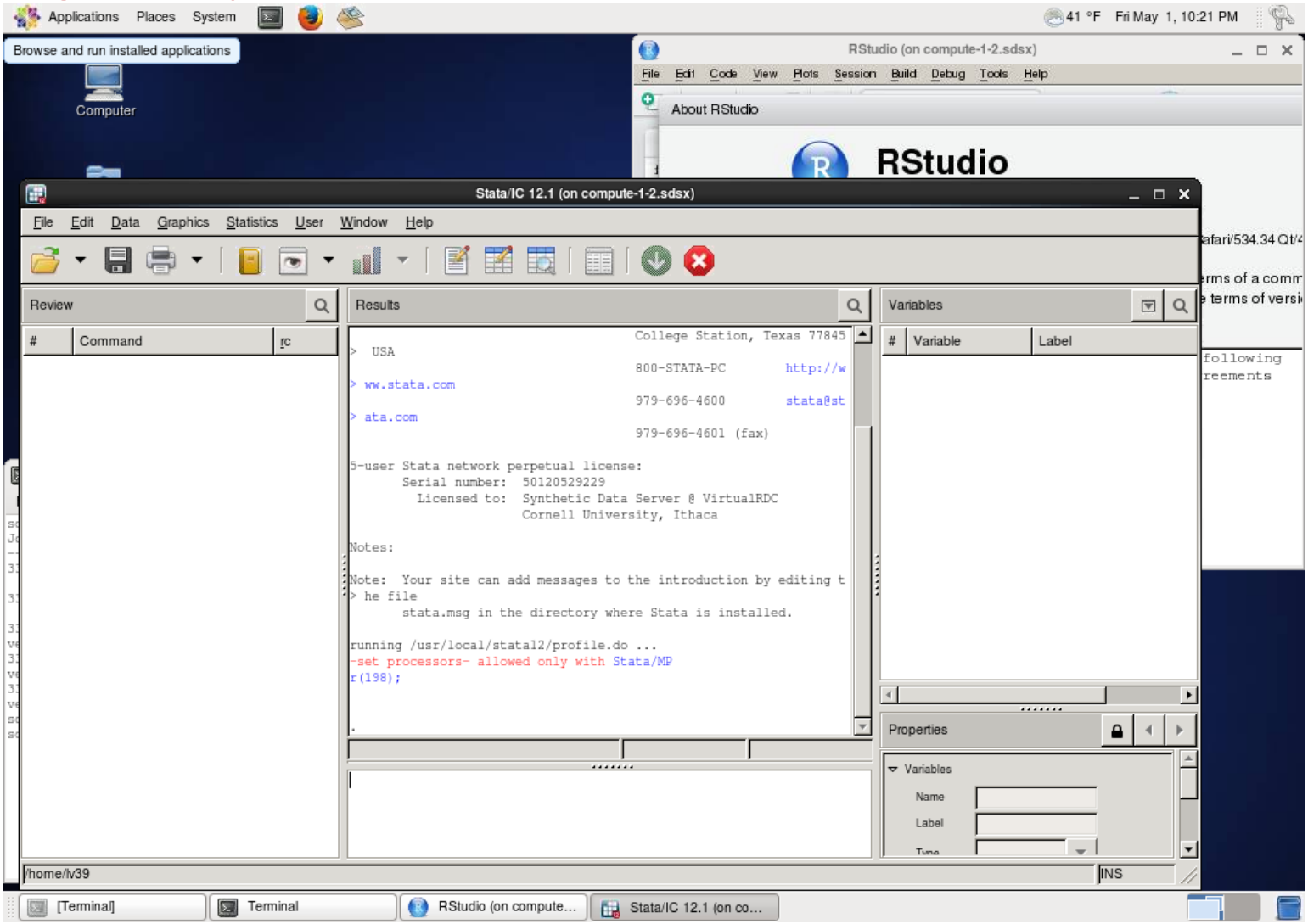
Contributions

- Allow researchers to work as close as possible to their regular workflow
  - Ideally, downloadable data (desktop paradigm)
  - If not, server-based desktop paradigm with easy access

$\rightarrow$ Synthetic Data Server (Cornell)

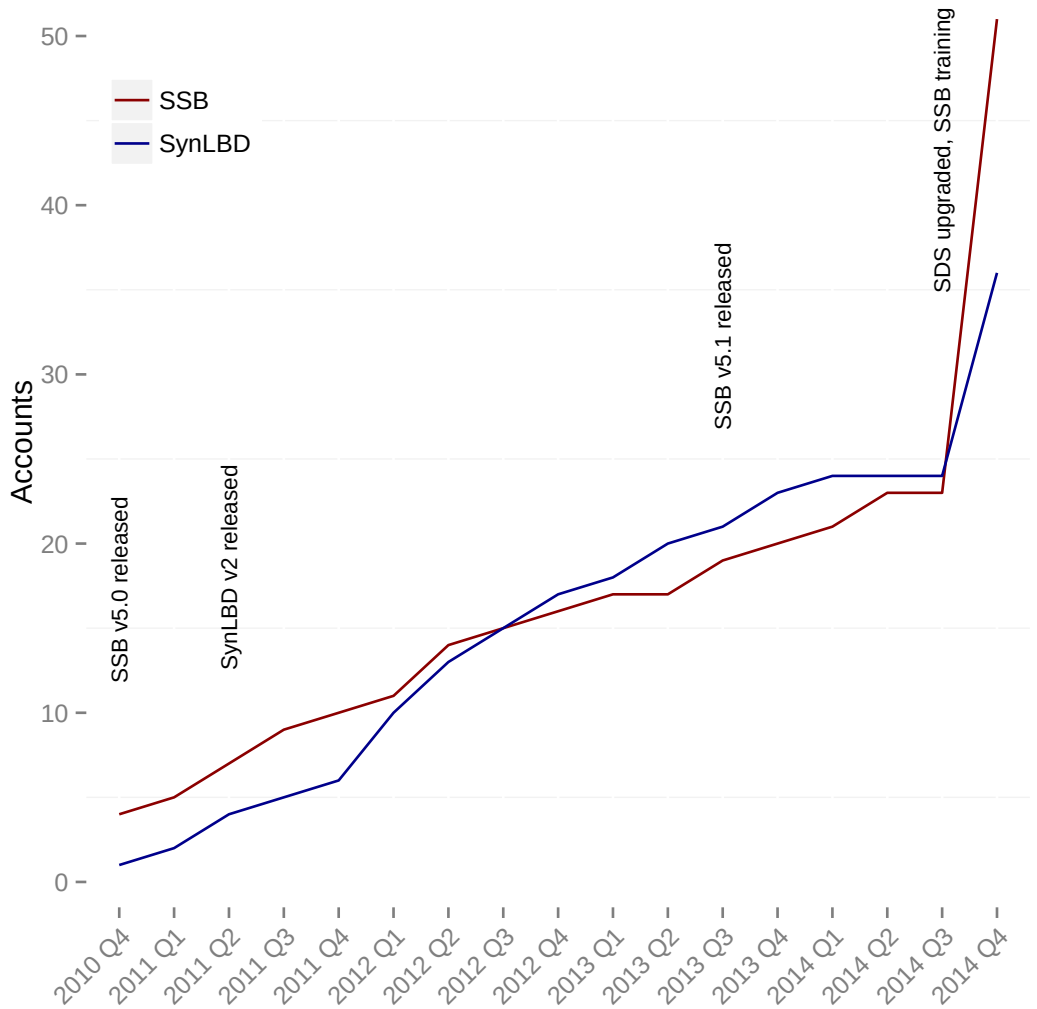
What does that mean?

Using the Synthetic Data Server



How has that worked?

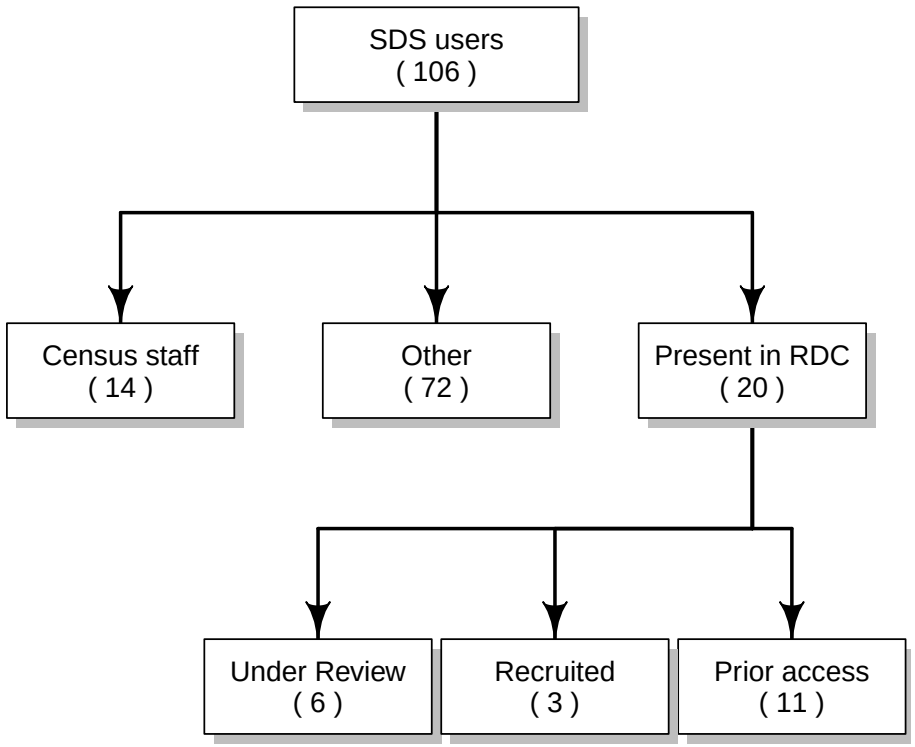
Usage of Synthetic Data Server



Remember...

How has that worked?

User profile



How has that worked?

Additional pathways

- Not only a valid data analytic tool ...
- ... additional pathway to confidential data
- ... with better utility than other “test” data

Facilitating validation

Contributions

Focus on validation methods

Validation for statistical agencies

- Validation is a cost, to be balanced against alternate access mechanisms
- Cheaper is better

Validation for researchers

- Validation is a cost, to be balanced against alternate access mechanisms
- Faster is better

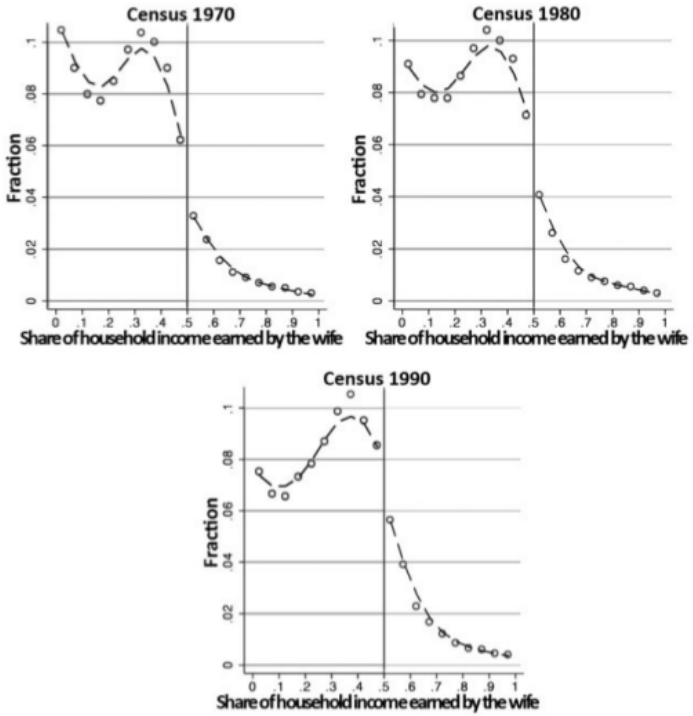
Statistics

Hard metrics are hard to come by

In December 2012, **30 users** of the **SynLBD** had generated **5 validation** requests.

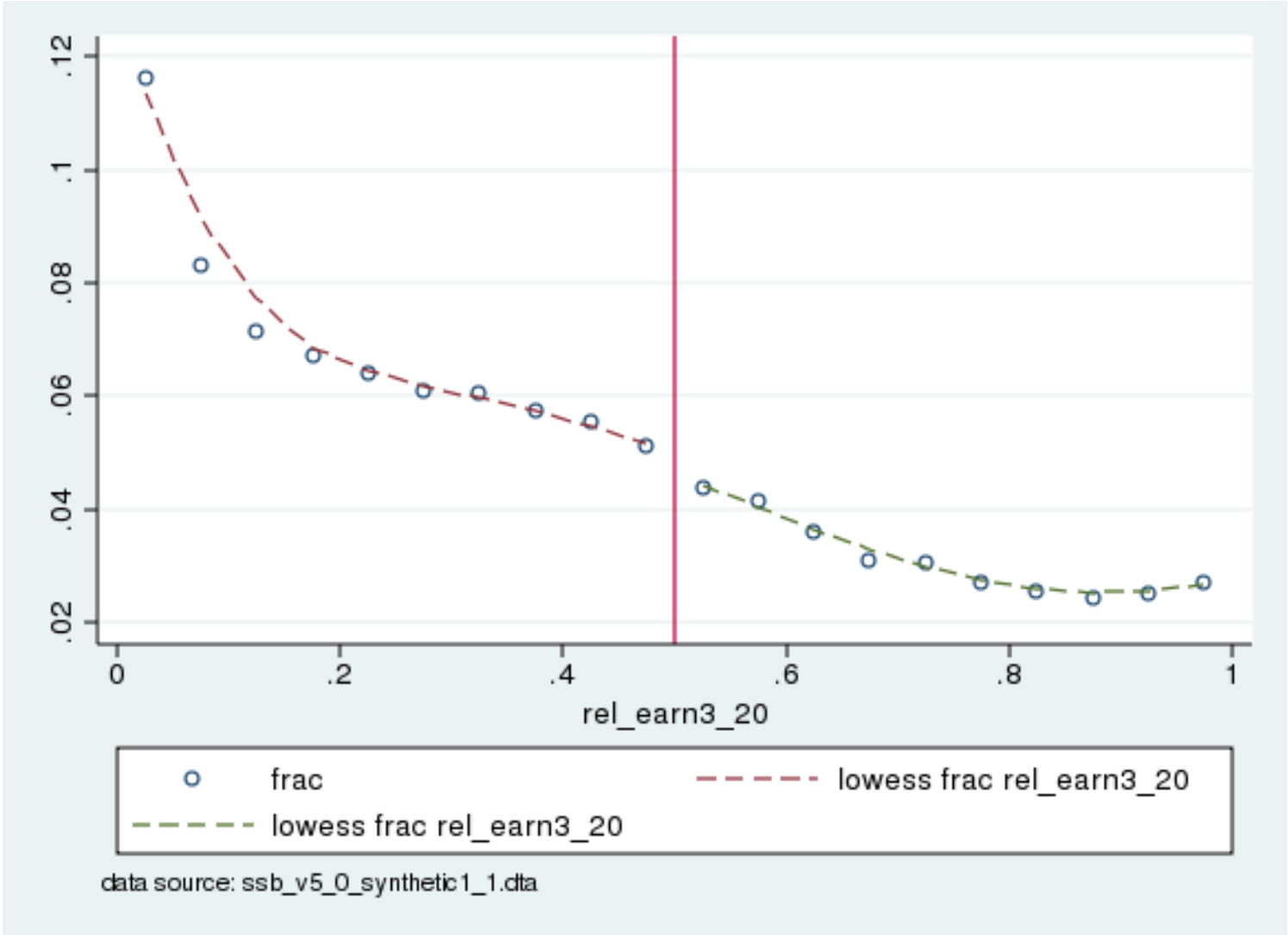
An example: Bertrand, Kamenica, Pan (2015)

Study on relative income within households



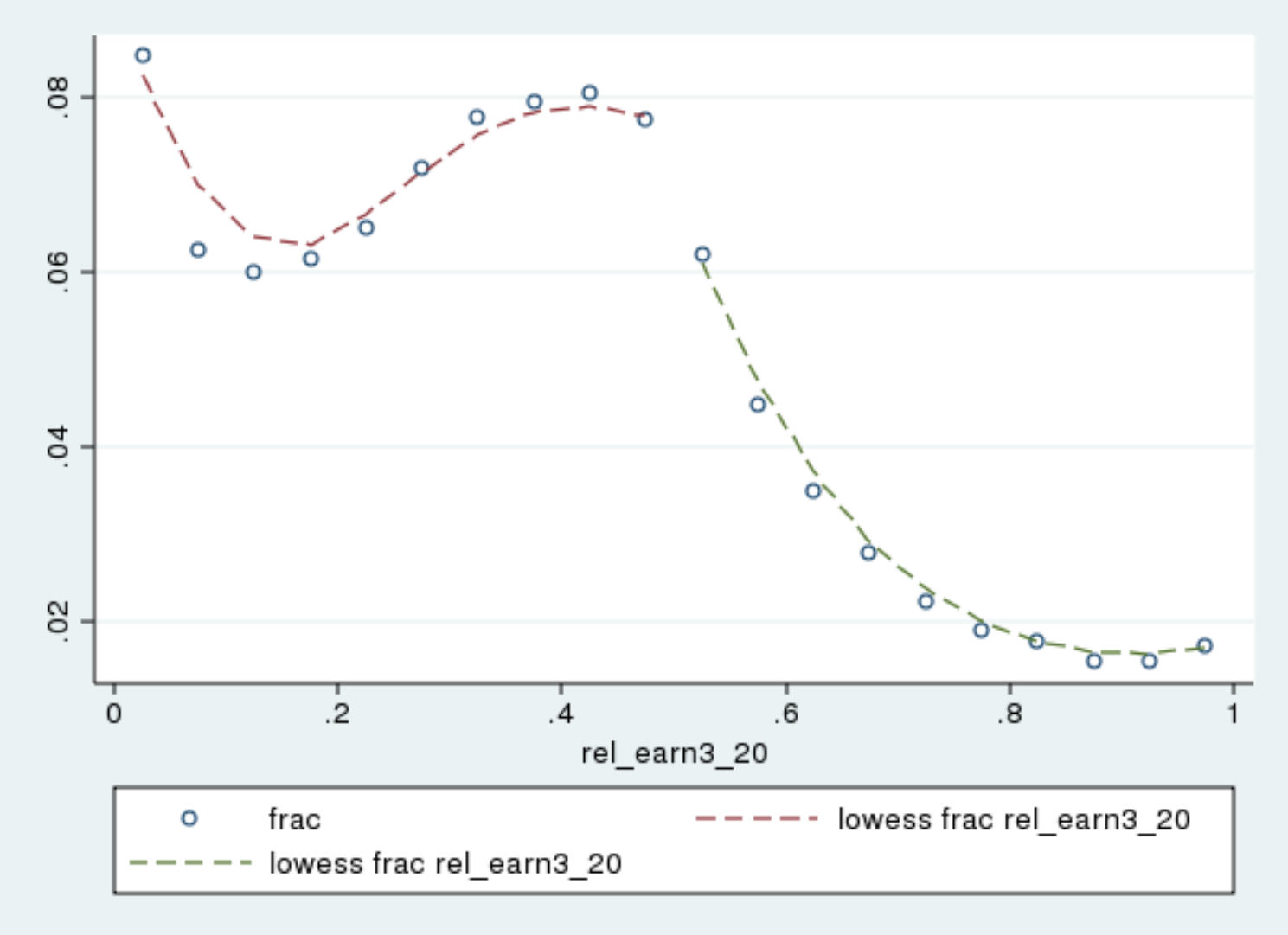
An example: Bertrand, Kamenica, Pan (2015)

Project with SIPP Synthetic Beta



An example: Bertrand, Kamenica, Pan (2015)

Validation against Gold Standard SIPP file



Making validation easier

Key insight

validation = replication

Solution

Use workflow tools

Key problem

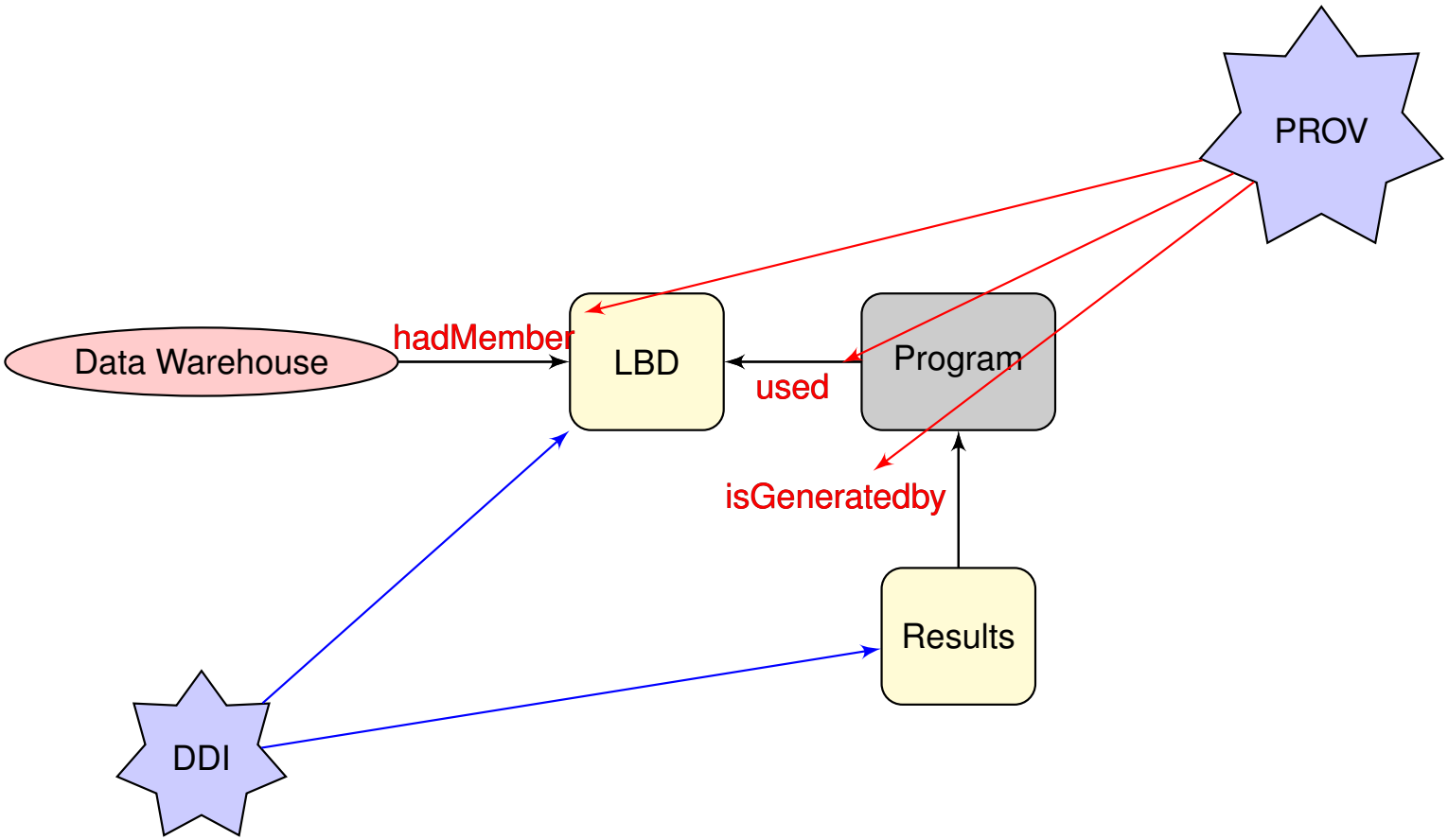
social scientists don't like workflow tools

Our solution

Ex-post workflow documentation

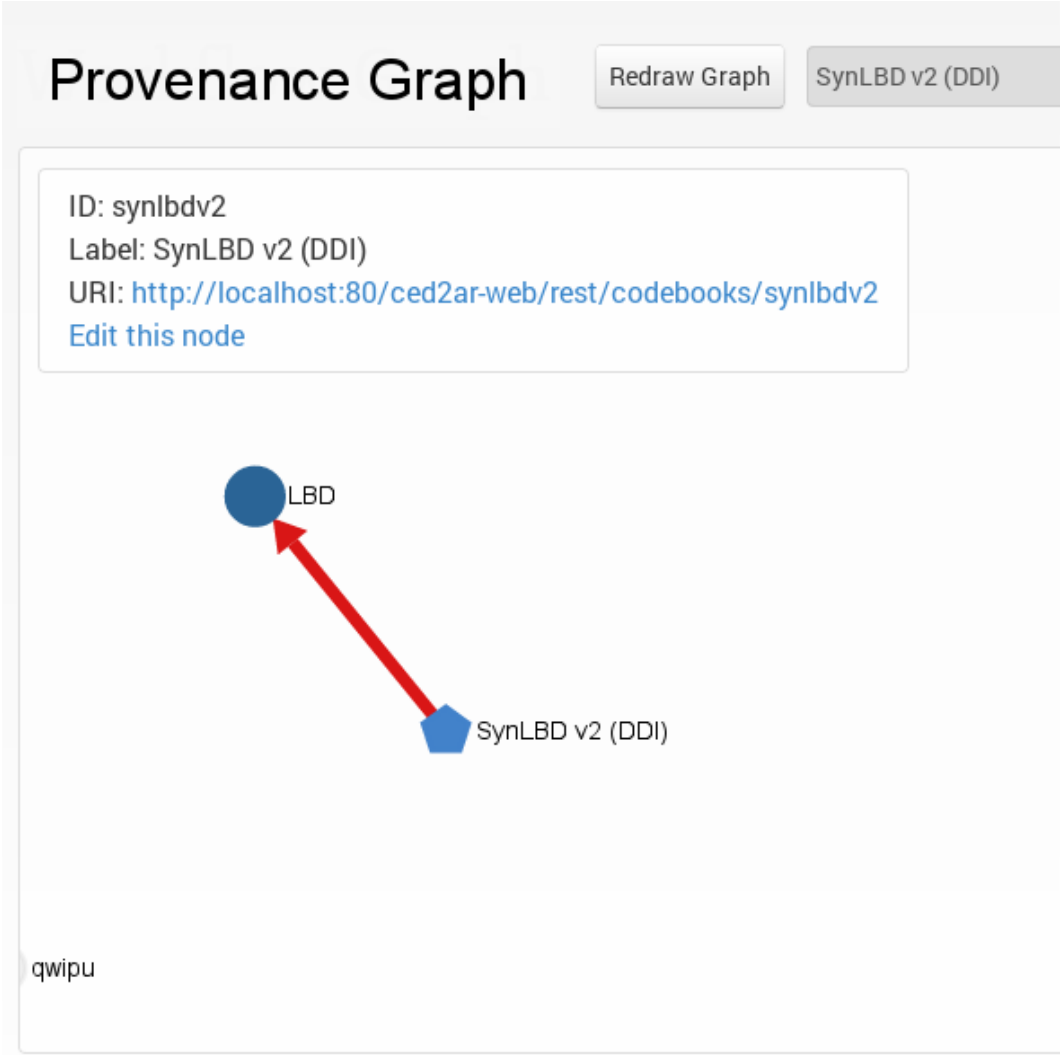
- Within a restricted-access environment (RDC) or a validation-requirement environment (SDS), user is required to document end results
- Already includes (i) description of variables (ii) description of programs used to generate variables (iii) description of transformations
- **Cornell NCRN**: same data documentation standard as used for codebook generation
- ... with one twist: addition of PROV language

DDI+PROV for workflow

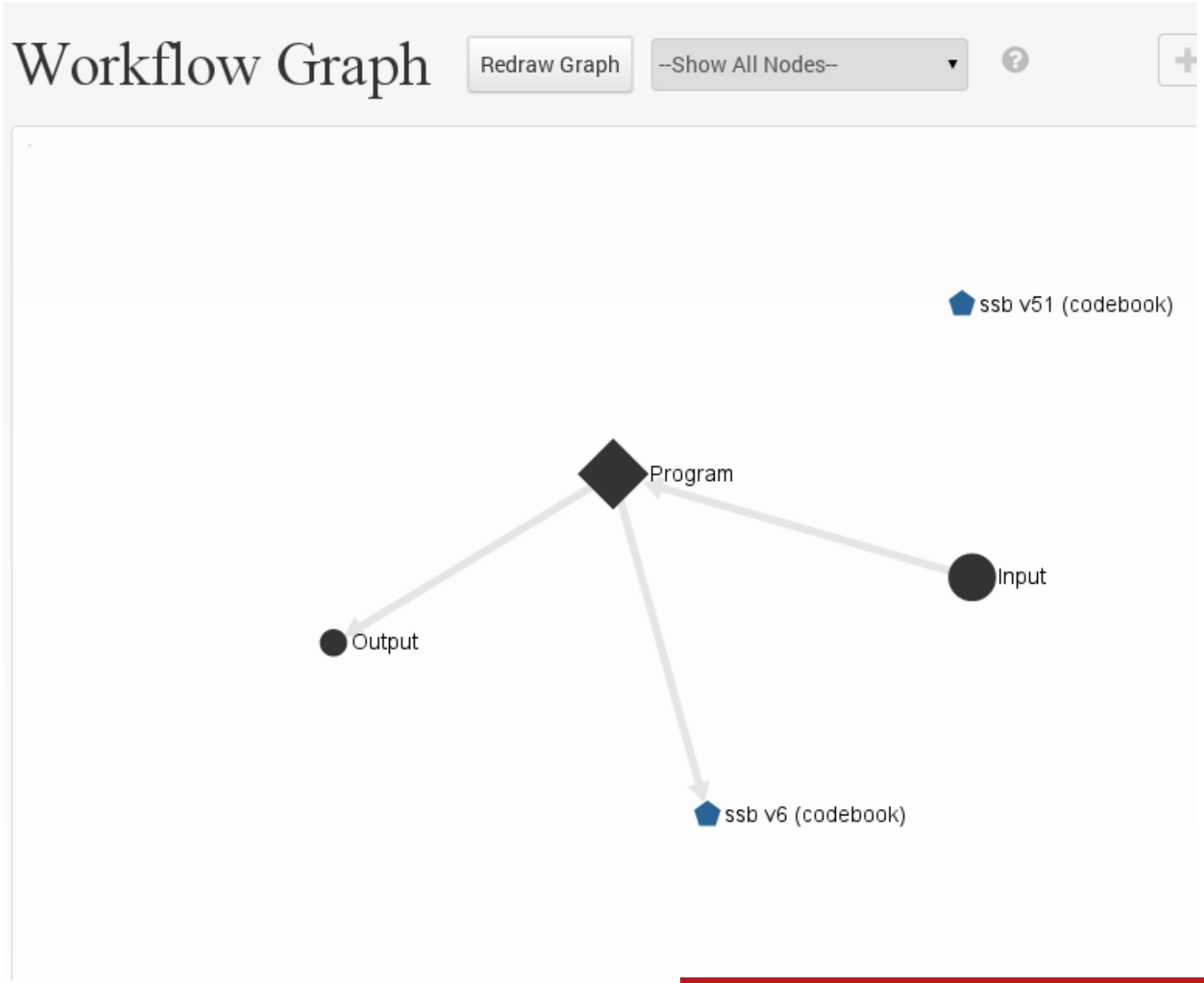


How to create this?

You've seen something like this before:



How to create this?



Why does this matter?

Machine-actionable documents

From the same document...

- ... generate result release request *(which the user needed to generate anyway)*
- ... generate programs for validation *(which the statistical agency needs anyway)*
- ... database the analyses for later meta-analysis *(helping in the extraction of models and results)*

Example release request

Appendix D: Clearance Request Memo

REQUEST FOR CLEARANCE OF RESEARCH OUTPUT

Center for Economic Studies and Research Data Centers

****

* Project #:

* Submitted by:

*

* For CES Reviewer to complete:

* Cleared for release:

* Cleared by:

****

1. GENERAL INFORMATION

a. Name of this request's subdirectory under the project's main clearance directory:

b. Please provide a general description of the outputs you wish to clear:

c. Please state how the outputs are part of the research project as approved (You may summarize or copy descriptions from your proposal, with page references.)

2A. DESCRIPTIONS OF RESEARCH SAMPLES:

Describe your Research sample(s) or "cuts" of data used in research output. For each sample, please describe your selection criteria and how the research sample differs from the samples underlying survey publications or other samples you have used. Take as much space as you need for each; add samples as needed.

SAMPLE 1:

SAMPLE 2:

SAMPLE 3:

2B. RELATIONSHIP BETWEEN SAMPLES

Describe how your samples relate to each other (e.g., if you have two samples, is one a subsample of another?) In the cases of samples and subsamples, there is an implicit third sample, the difference between the two. Please describe this sample above. We probably will need to examine any implicit samples as well.

2C. RELATIONSHIP TO OTHER PUBLICATIONS

Describe how your samples may relate to similar samples from other projects or from survey publications. (e.g., how your sample of an industry in the LRD differs from the Census of Manufactures or Annual Survey of Manufactures files in the LRD).

Appendix D: Clearance Request Memo

Request for Clearance of Research Output

Center for Economic Studies and Research Data Centers

Project # :

Submitted by :

For CES reviewer to complete

Cleared for release :

Cleared by :

1. General information

a. Name of this request's subdirectory under the project's main clearance directory:

http://localhost:8080/ced2ar-web/codebooks/synlbdv2

b. Please provide a general description of the outputs you wish to clear:

In most countries, national statistical agencies do not release establishment-level business microdata, because doing so represents too large a risk to establishments' confidentiality. One approach with the potential for overcoming these risks is to release synthetic data; that is, the released establishment data are simulated from statistical models designed to mimic the distributions of the underlying real microdata. The Synthetic Longitudinal Business Database (SynLBD) is the synthetic data version of the Longitudinal Business Database (LBD), an annual economic census of establishments in the United States comprising more than 20 million records dating back to 1976. More information is available at https://www.census.gov/ces/dataproducts/synlbd/index.html. In this codebook, variables are noted as "blanked" if they are available on the confidential version but have been removed from the synthetic version; "synthetic" if the confidential values have been synthesized and released on the synthetic version.

c. Please state how the outputs are part of the research project as approved (You may summarize or copy descriptions from your proposal, with page references.)

Where should this come from?

2. Research

Contributions

Use Focus on use

In order to get researchers to use the data, it needs to be convenient and useful.

Contributions

- Expand applications and data coverage is one way of enhancing utility to researchers, and possibly statistical agencies
  - *German Synthetic LBD* (Work by Drechsler and Vilhuber)

→ **Cornell**, IAB Germany

  - *Canada* (tentative, grant proposal)

→ **Cornell, Duke**, others

  - other countries (some expression of interest)

Contributions

Use Focus on use

In order to get researchers to use the data, it needs to be convenient and useful.

Contributions

- Expanding to international context

→ contributes to

(robustness of synthetic data models)

→ greater utility for researchers: existence of cross-national comparable confidential data files (cross-country analysis)

Thank you.