Economics and the economics of privacy: new methods of accessing new data

Lars Vilhuber[1]

[1]Labor Dynamics Institute, ILR, Cornell University, United States

November 2015
UQAM
Montréal, Canada

Disclaimer

Funding

- Vilhuber's work is partially funded by NSF Grants #1042181, #1131848, and #0941226, and by a grant from the Alfred P. Sloan Foundation.

Disclaimer

"

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review by the Census Bureau than its official publications. This report is released to inform interested parties and to encourage discussion. Any findings, conclusions or opinions are those of the authors. They do not necessarily reflect those of the Center for Economic Studies, the U.S. Census Bureau, or the National Science Foundation.

"

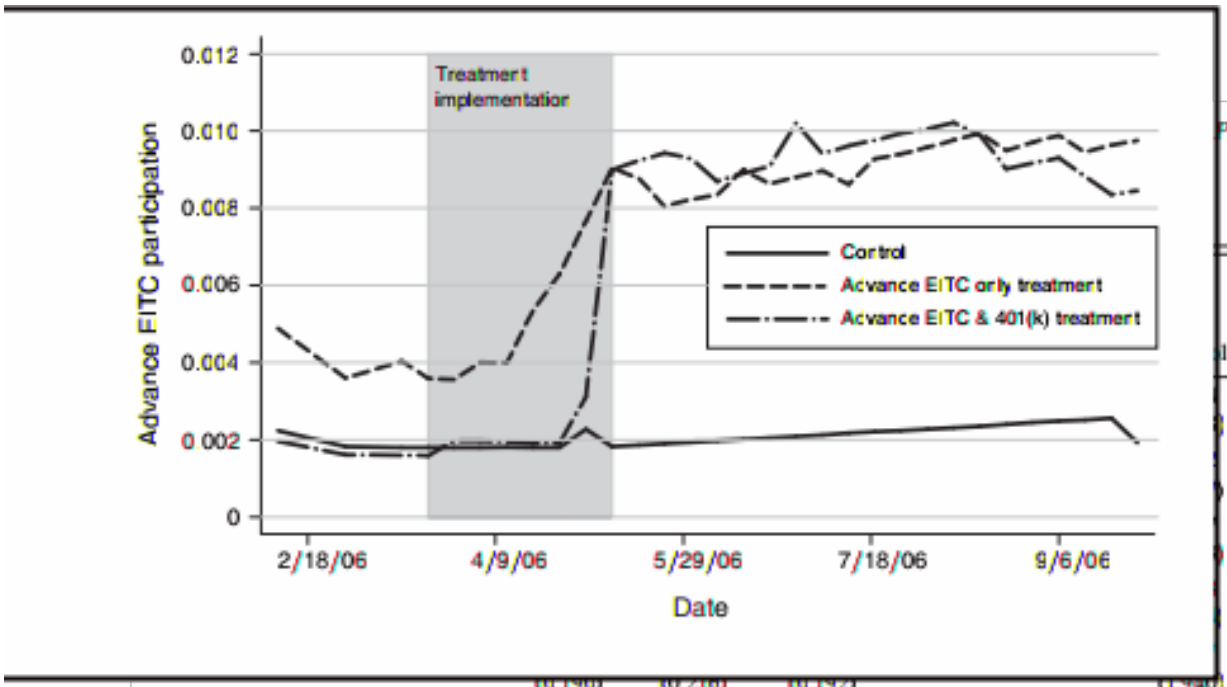
Acknowledgements

This work presents the results of collaborations with

John Abowd, Bill Block, Warren Brown, Ben Perry, Carl Lagoze, Venky Kambhampaty, Ian Schmutte, Kevin McKinney, Javier Miranda, Flavio Stanchi, Hautahi Kingi, Ashwin Machanavajjhala, Mark Kutzbach, Matthew Graham, Samuel Haney, and others.

Context

Economic analysis



						Advance EITC only treatment	Advance EITC & 401(k) treatment
						32.749	28.027
						(11.215)	(5.936)
						19.766	25.883
						(5.631)	(6.481)
						6.334	9.369
						(2.964)	(1.825)
						0.854	0.834
						(0.506)	(0.123)**
						30.503	29.655
						(1.778)	(1.875)
Median wage	7.646	7.257	7.634	Female	80.457	77.448	80.945
	(0.041)	(0.028)***	(0.101)		(1.710)	(2.082)	(1.351)
Weekly hours	24.766	25.374	25.772	Southern region	40.879	67.026	66.217
	(1.022)	(0.914)	(0.587)		(19.268)	(19.752)	(20.250)
W-4 Allowances	1.291	1.676	1.441	Store size	42.019	44.753	44.251
	(0.160)	(0.461)	(0.262)		(3.064)	(3.074)	(2.051)
Age	34.368	33.669	34.782	Weekly turnover rate	0.270	0.361	0.398
	(0.780)	(0.505)	(0.569)		(0.078)	(0.103)	(0.113)
N	2,231	2,226	2,519	*N*	2,231	2,226	2,519

The preponderance of public-use data

Microdata

"... paper uses data from the Current Population Survey..."

The preponderance of public-use data

Microdata

"... paper uses data from the Current Population Survey..."

Macrodata

"We use data downloaded from the Bureau of Economic Analysis..."

source

Yielding...

Administrative data

> *"Our analysis draws on administrative records from the Detroit Work First program linked with unemployment insurance (UI) wage records for the State of Michigan"*
> *Autor/Houseman doi:10.1257/app.2.3.96*

Yielding...

Administrative data

"Our analysis draws on administrative records from the Detroit Work First program linked with unemployment insurance (UI) wage records for the State of Michigan"
Autor/Houseman doi:10.1257/app.2.3.96

Administrative data

"confidential student-level panel dataset provided by the School Board of Alachua County in Florida"
Carrel and Hoekstra doi:10.1257/app.2.1.211

... yielding...

Proprietary data

"This field experiment was made possible by the collaboration of a large-scale, nationwide firm in the retail sector. "
Damon doi:10.1257/app.2.2.147

The Death Knell for Public-use Data

The Death Knell for Public-use Data

- Sounded by young scholars pursuing research programs that mandate inherently identifiable data:
  - Geospatial relations,
  - Exact genome data,
  - Networks of all sorts,
  - Linked administrative records
- These researchers acquire authorized, generally unfettered, restricted access to the confidential, identifiable data and perform their analyses in secure environments.
- But...

...they don't leave behind the scientific trail that has made public-use files so important.

Replication of research results

Critical element of science

- Replication of methods, data inputs, computational environment is a critical element of the scientific approach
- Journals, funding agencies (in the U.S.) have been moving to making archiving of inputs to scientific results more robust, even mandatory

The problem

Good intentions, costly access

"researchers could submit programs that [...] research assistants would run. Alternatively, researchers wishing to work directly with the data could come and work on the Institute's premises. "

The problem

Good intentions, costly access

"researchers could submit programs that [...] research assistants would run. Alternatively, researchers wishing to work directly with the data could come and work on the Institute's premises. "

Uncertain access

"Data [...] is proprietary and owned by the Alachua County, Florida School District. The corresponding author [...] holds the deidentified dataset [...] and will provide copies to authors who receive written permission from the Alachua County Public Schools."

The problem

Good intentions, costly access

"researchers could submit programs that [...] research assistants would run. Alternatively, researchers wishing to work directly with the data could come and work on the Institute's premises. "

Uncertain access

"Data [...] is proprietary and owned by the Alachua County, Florida School District. The corresponding author [...] holds the deidentified dataset [...] and will provide copies to authors who receive written permission from the Alachua County Public Schools."

No access

Some do not provide any information on access.

Not a new problem

Econometrica

"In its first issue, the editor of Econometrica (1933), Ragnar Frisch, noted the importance of publishing data such that readers could fully explore empirical results. Publication of data, however, was discontinued early in the journal's history. [...] The journal arrived full-circle in late 2004 when Econometrica adopted one of the more stringent policies on availability of data and programs.

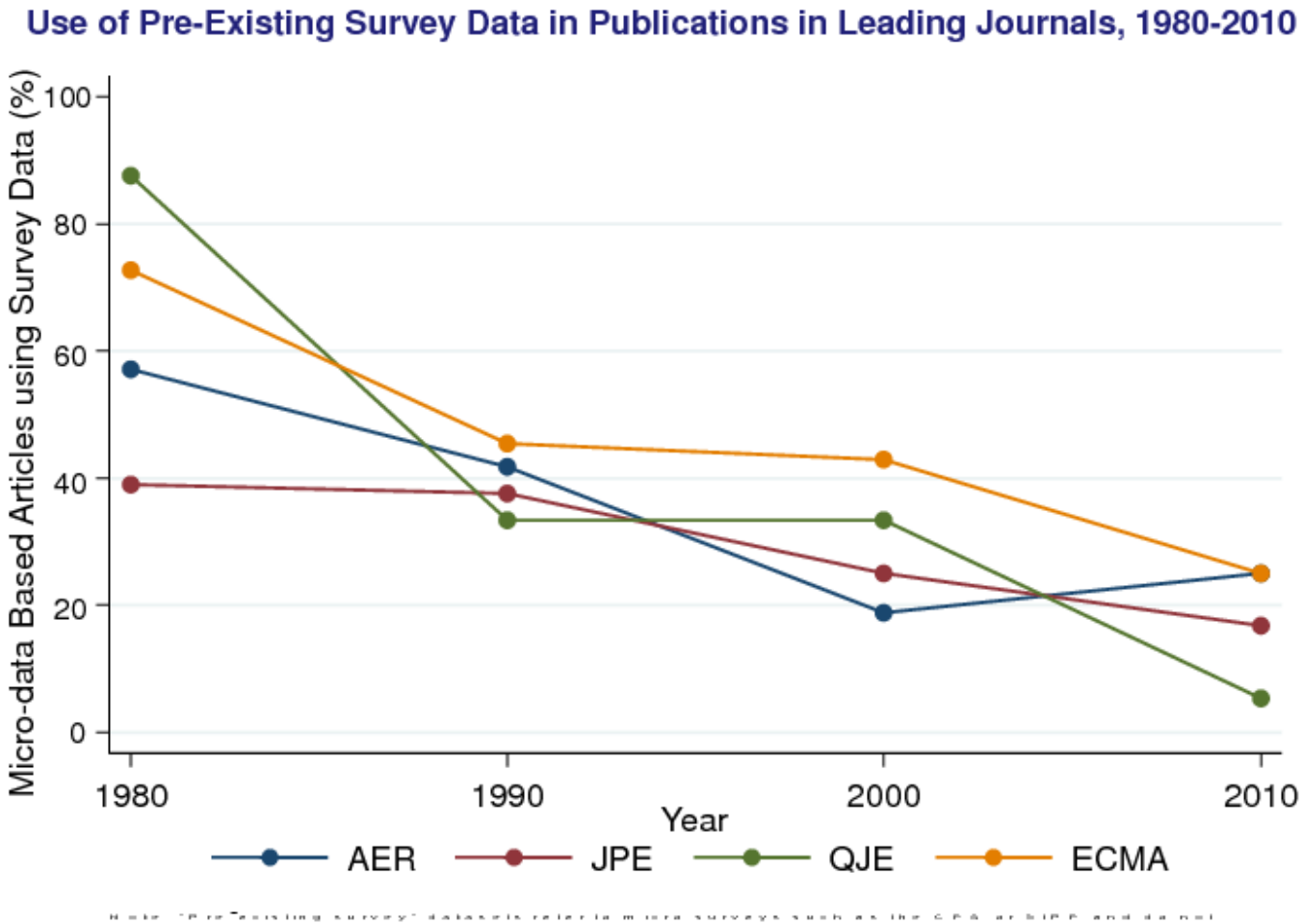
http://www.econometricsociety.org/submissions.asp#4 as cited in Anderson et al (2005)

Problem will become worse

Increased use of restricted-access data

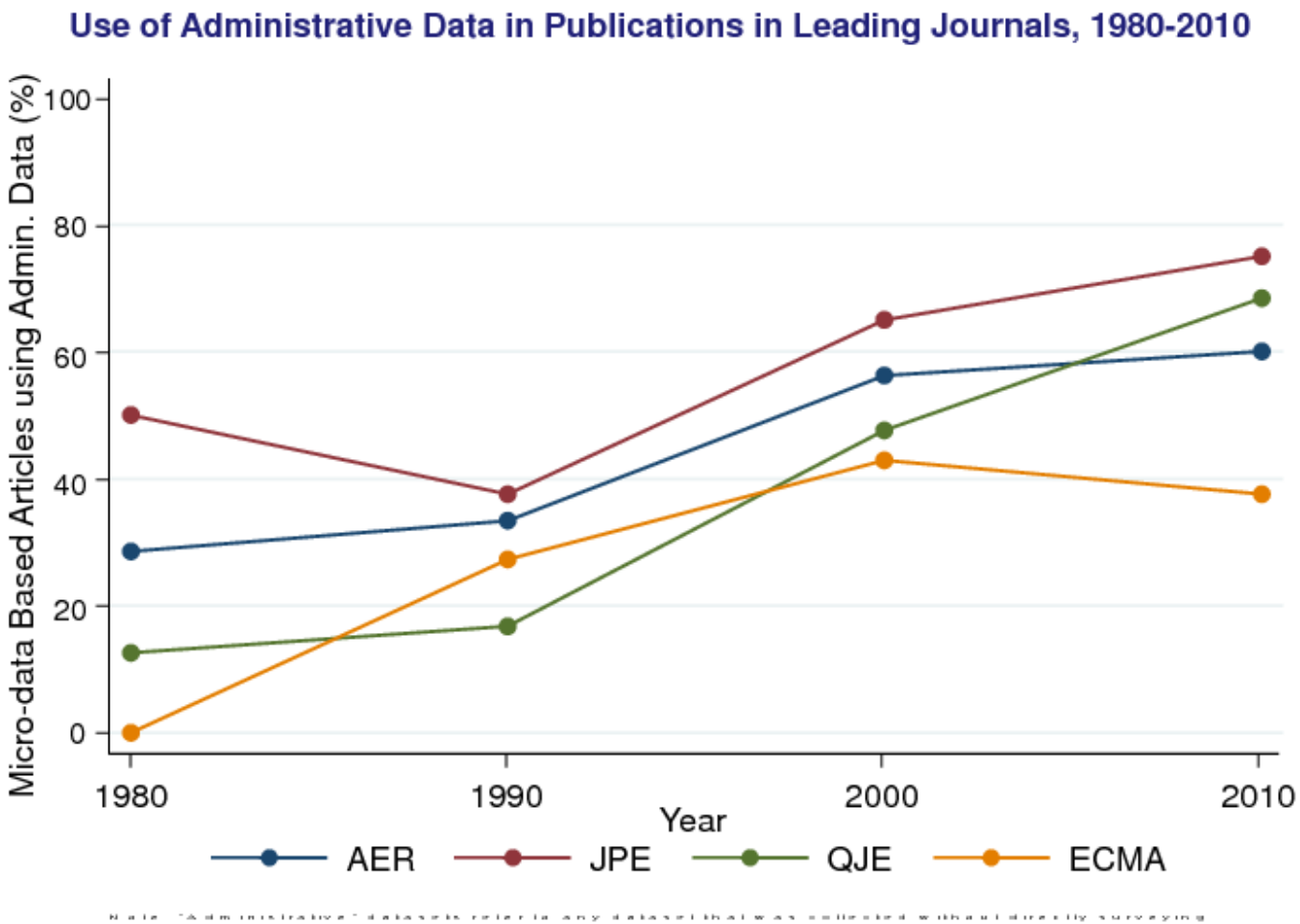
- Archiving (curation) of input data is complicated
- Knowledge discovery is complicated

Decline in the use of classic public-use data



include surveys designed by researchers for their study. Sample excludes studies whose primary data source is from developing countries.

Increase in the use of administrative data in economics



individuals (e.g., scanner data, stock prices, school district records, social security records).
Sample excludes studies whose primary data source is from developing countries.

Results from the LDI Replication Lab

Undergraduate research team

- Census of articles in the American Economic Journal: Applied Economics (2010, 2011, 2013)
- Each article is analyzed for availability of replication archive (as required by journal!)
- If data and programs are available, reproducibility is tested.

Some very preliminary results

Table: Replication Success

	Yes	No	Partial	Sum
2010	10	19	6	35
2011	12	20	4	36
2013	15	12	11	38
Total	37	51	21	109

Some very preliminary results

Table: Reason for Replication Failure

	Missing Data	Corrupted Data	Code Error	Missing Code	Sum
2010	15	1	1	2	19
2011	15	1	1	3	20
2013	12	0	0	0	12
Total	42	2	2	5	51

Some very preliminary results

Table: Reason for Missing Data

	Administrative			Private		
	local	National	Regional	Commercial	Other	Sum
2010	2	8	0	4	3	17
2011	2	8	4	1	0	15
2013	2	2	1	4	2	11
Total	6	18	5	9	5	43

Some very preliminary results

Table: Type of Access to Confidential Data

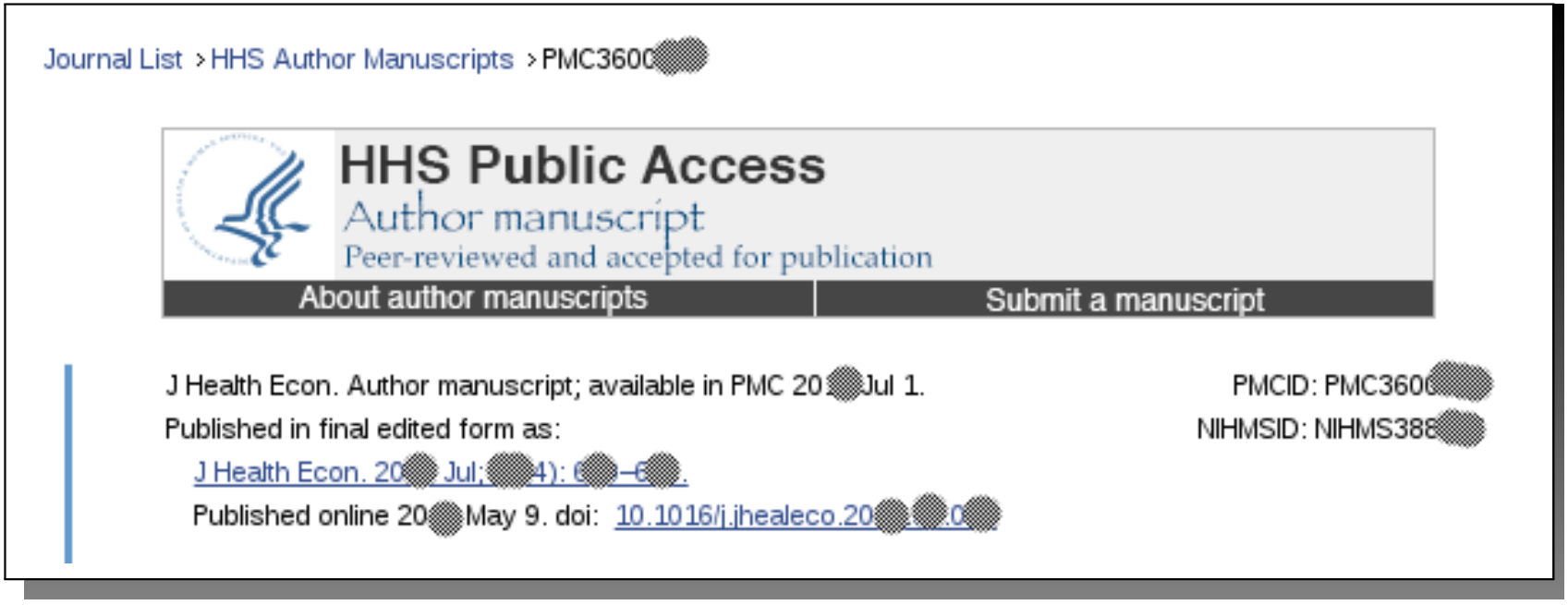
	Formal	Informal w/ Commitment	Informal w/o Commitment	No Info	Sum
2010	2	3	9	3	17
2011	2	0	10	3	15
2013	1	2	8	0	11
Total	5	5	27	6	43

Not limited to one journal

NIH-funded research

- article is open-access
- not clear about data access

A small anonymous example



Journal List > HHS Author Manuscripts > PMC360[redacted]

HHS Public Access

Author manuscript

Peer-reviewed and accepted for publication

About author manuscripts | Submit a manuscript

J Health Econ. Author manuscript; available in PMC 20[redacted] Jul 1.

Published in final edited form as:

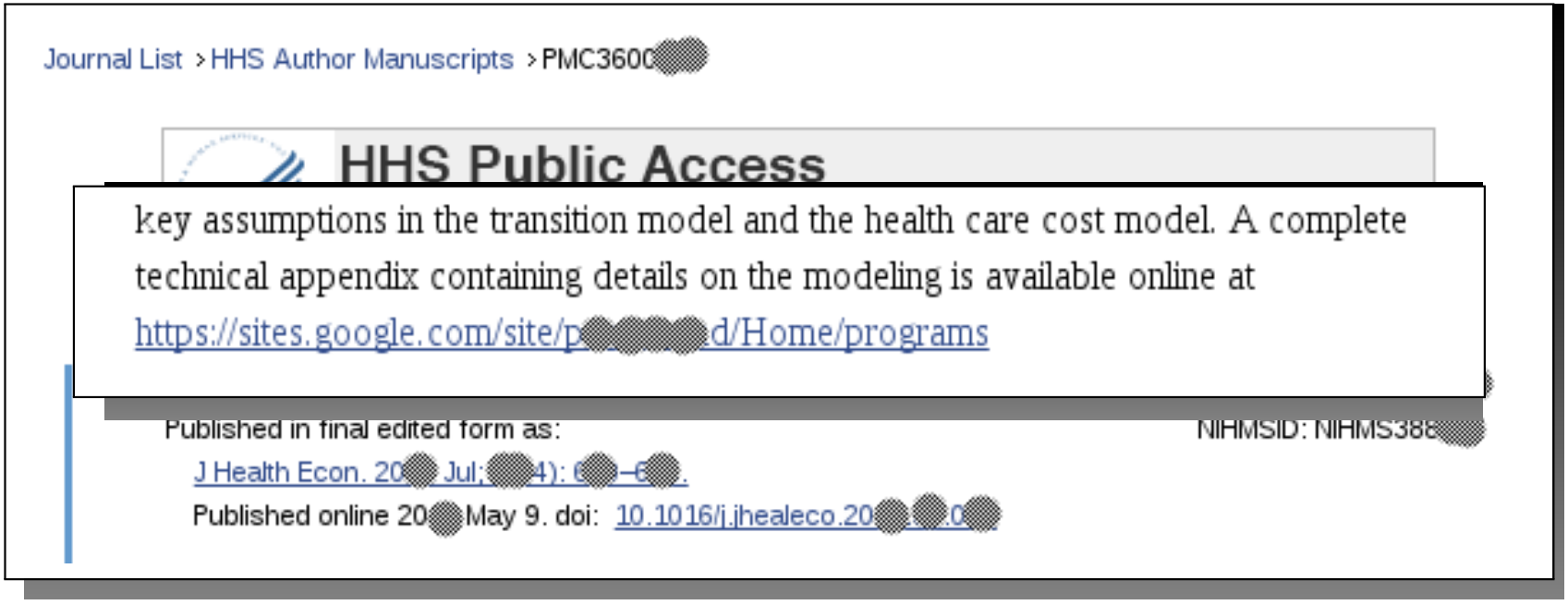
J Health Econ. 20[redacted] Jul;[redacted]4): 6[redacted]–6[redacted].

Published online 20[redacted] May 9. doi: 10.1016/j.jhealeco.20[redacted].0[redacted]

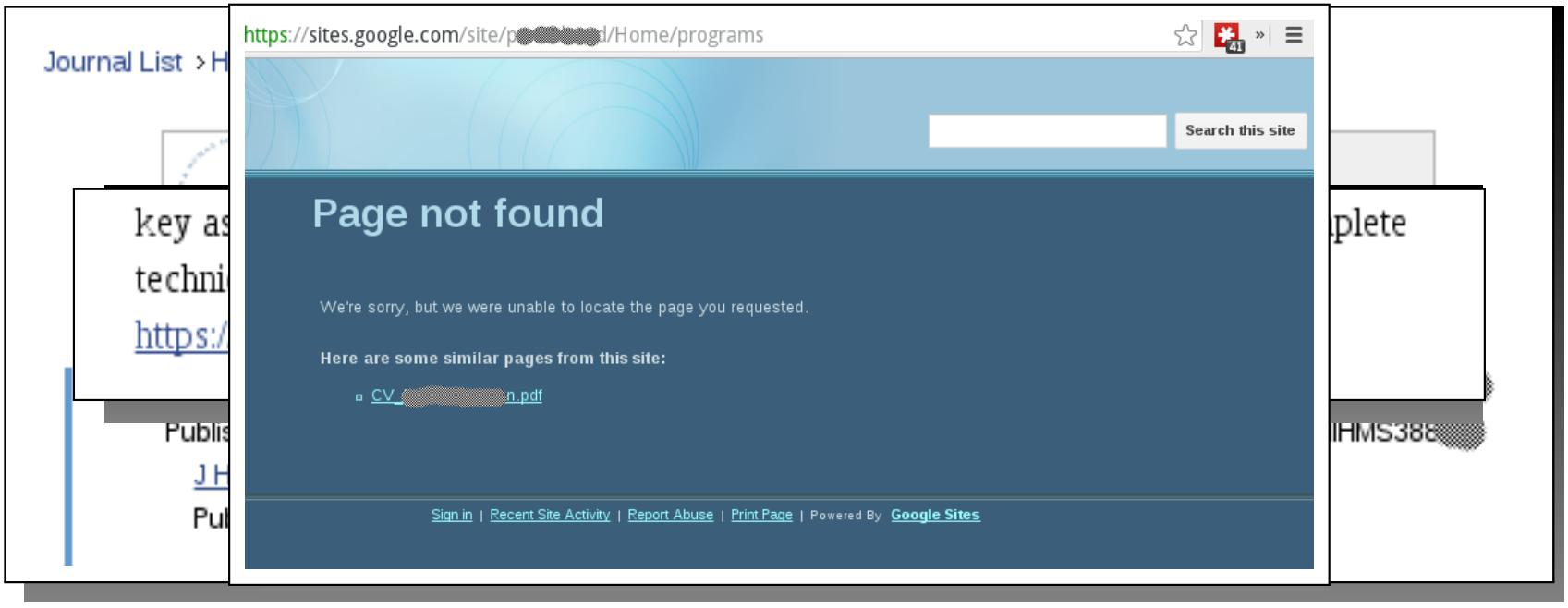
PMCID: PMC360[redacted]

NIHMSID: NIHMS38[redacted]

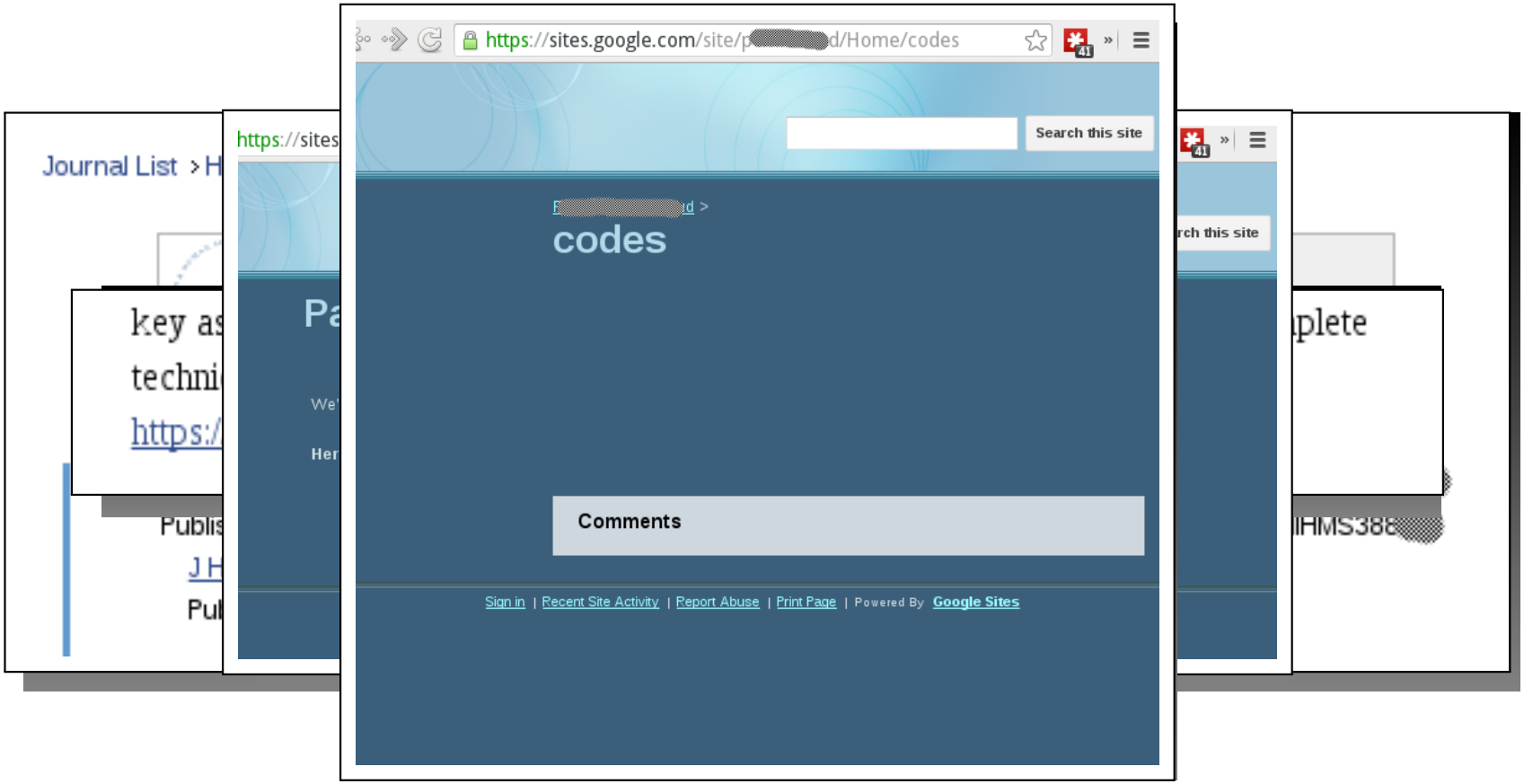
A small anonymous example



A small anonymous example



A small anonymous example



Not limited to economics

Nature, 2012

"Many of the emerging 'big data' applications come from private sources that are inaccessible to other researchers. The data source may be hidden, compounding problems of verification, as well as concerns about the generality of the results."

(Huberman, Nature 482, 308 (16 February 2012) doi:10.1038/482308d)

Other domains

- Biology (genetics data, chemical compounds)
- Computer science (search records, single-firm examples)


2012

A program

Allowing for easier documentation of provenance

- Better documentation about confidential data
- Solving the reproducibility problem

Making data more accessible

- New disclosure limitation techniques
- New data access models

Replicability

Non-federal confidential data

States, school districts, private companies, academic and private surveys: need a place to live to be re-used.

Options

- openICPSR `https://www.openicpsr.org/`
- Harvard Dataverse `https://dataverse.harvard.edu/` (1,315 DV, 59,530 DS)
- Ontario Council of University Libraries: `http://dataverse.scholarsportal.info/dvn/` (64 DV, 5,289 files)

Hinges on compatibility of data deposit rules, laws, regulations, etc.

Can we influence this process?

Data repositories have the technology to receive deposits

- Underutilized
- When integrated into journal workflows, useless (blobs of unstructured ZIP files)

Journals can require data citations

- Review process scrutinizes *article* citations
- Would be easy to enforce *data* citations

Data citations

Examples

Deschenes, Elizabeth Piper, Susan Turner, and Joan Petersilia. ***Intensive Community Supervision in Minnesota, 1990-1992: A Dual Experiment in Prison Diversion and Enhanced Supervised Release*** *[Computer file]. ICPSR06849-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2000. doi:10.3886/ICPSR06849*

Abowd, John M.; Vilhuber, Lars, 2014, ***"Replication data for: National estimates of gross employment and job flows from the Quarterly Workforce Indicators with demographic and industry detail"****, doi:10.7910/DVN/27923, Harvard Dataverse [Distributor], V2*

[src]

Data citations

Examples

Deschenes, Elizabeth Piper, Susan Turner, and Joan Petersilia. ***Intensive Community Supervision in Minnesota, 1990-1992: A Dual Experiment in Prison Diversion and Enhanced Supervised Release*** *[Computer file]. ICPSR06849-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2000. doi:10.3886/ICPSR06849*

Abowd, John M.; Vilhuber, Lars, 2014, ***"Replication data for: National estimates of gross employment and job flows from the Quarterly Workforce Indicators with demographic and industry detail"****, doi:10.7910/DVN/27923, Harvard Dataverse [Distributor], V2*

[src]

So we know how to deposit and cite data...

So we know how to deposit and cite data...

... except nobody does it...

We didn't do it...

Abowd and Vilhuber (2011)

J Econom. Author manuscript; available in PMC 2012 Mar 1.
Published in final edited form as:
J Econom. 2011 Mar 1; 161(1): 82–99.
doi: 10.1016/j.jeconom.2010.09.008
PMCID:
NIHMSID:

National Estimates of Gross Employment and Job Flows from the Quar
Workforce Indicators with Demographic and Industry Detail

John M. Abowd and Lars Vilhuber

Author information ▶ Copyright and License information ▶

Abstract

The Quarterly Workforce Indicators (QWI) are local labor market data produced and released every

We didn't do it...

Abowd and Vilhuber (2011)

J Econom. Author manuscript; available in PMC 2012 Mar 1.
Published in final edited form as:
J Econom. 2011 Mar 1; 161(1): 82–99.
doi: 10.1016/j.jeconom.2010.09.008

Press for the NBER; 2009. pp. 149–230.
5. Abowd JM, Vilhuber L. The sensitivity of economic statistics to coding errors in personal identifiers of Business and Economic Statistics. 2005;23(2):133–152
6. Abowd JM, Zellner A. Estimating Gross Labor Force Flows. Journal of Business and Economic St 1985;3:254–283

Abstract

The Quarterly Workforce Indicators (QWI) are local labor market data produced and released every

We didn't do it...

Abowd and Vilhuber (2011)

J Econom. Author manuscript; available in PMC 2012 Mar 1.
PMCID:
Published in final edited form as:
J Econom. 2011 Mar 1; 161(1): 82–99.
doi: 10.1016/j.jeconom.2010.09.008
NIHMSID:

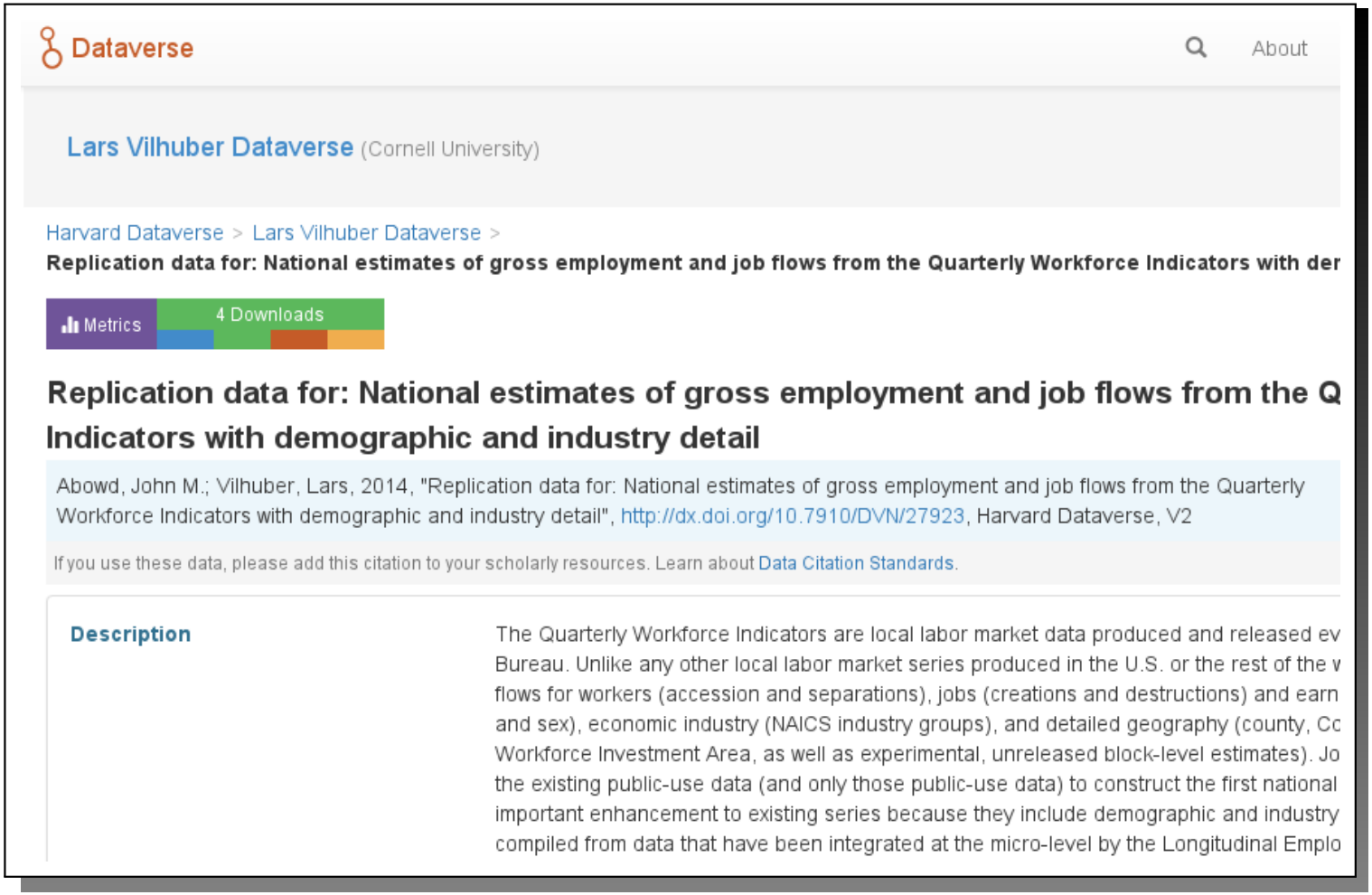
No confidential data were used in this paper. All public-use Quarterly Workforce Indicators data can be acc
http://www.vrdc.cornell.edu/news/data/qwi-public-use-data/. The national indicators developed in this paper
accessed from http://www.vrdc.cornell.edu/news/data/qwi-national-data/. We are grateful for the comments
suggestions of many of our colleagues, past and present, too numerous to list here and thus listed at the w
above and in the working paper version of this article. The opinions expressed in this paper are those of the
not the U.S. Census Bureau nor any of the research sponsors.

Abstract

The Quarterly Workforce Indicators (QWI) are local labor market data produced and released every

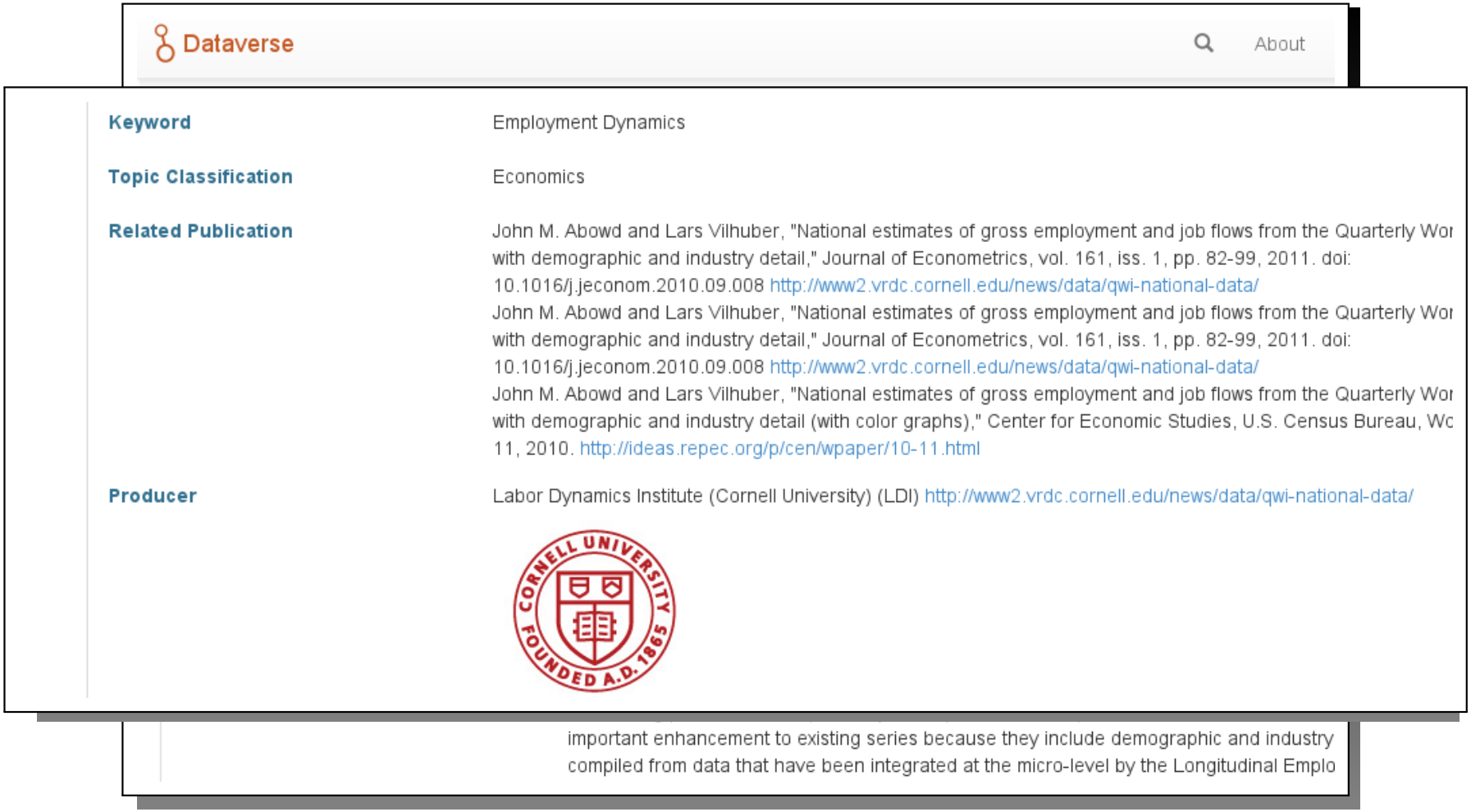
Then we archived it better...

... at Harvard Dataverse



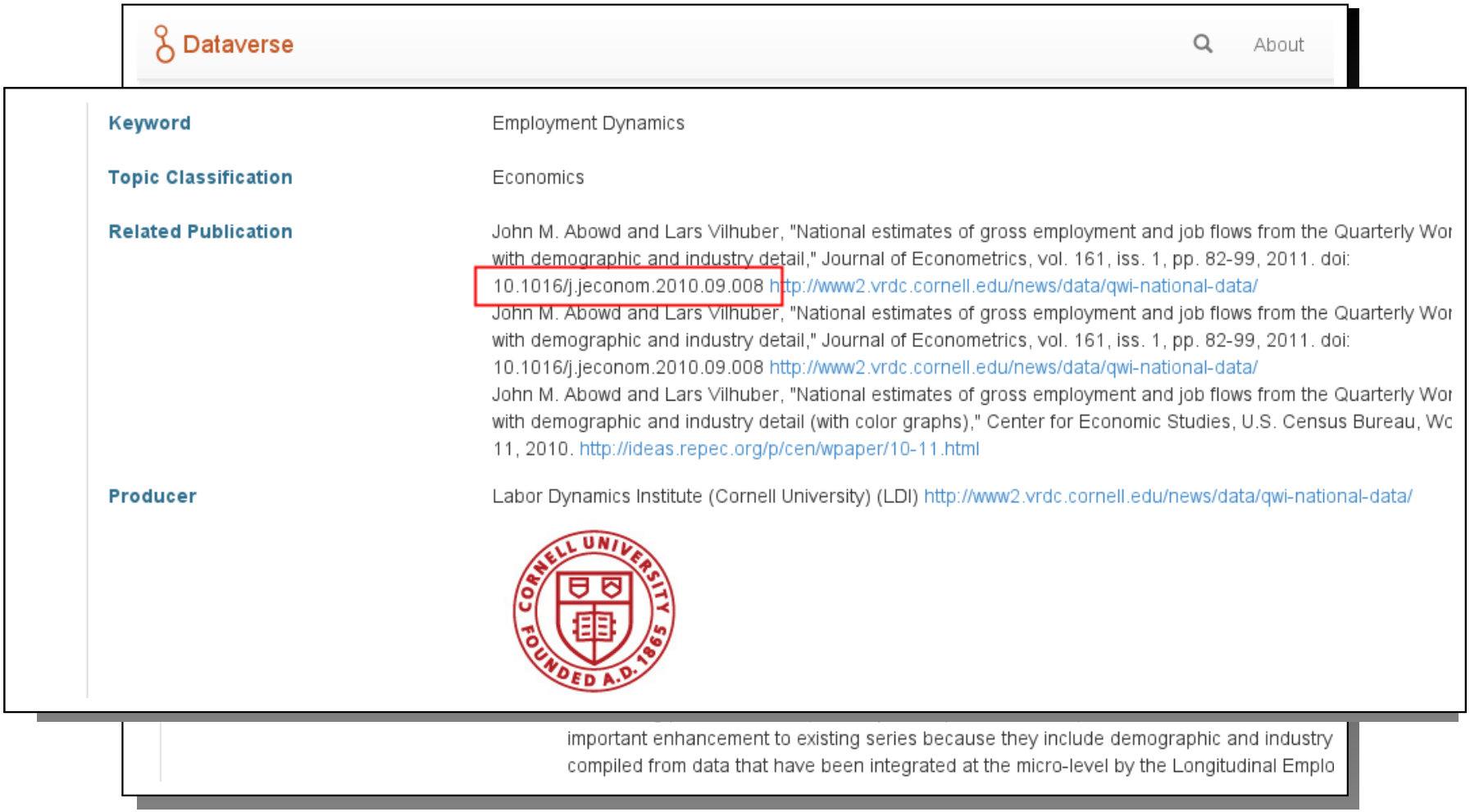
Then we archived it better...

... at Harvard Dataverse



Then we archived it better...

... at Harvard Dataverse



Provenance

The provenance problem

"data provenance, one kind of metadata, pertains to the derivation history of a data product starting from its original sources" [...] "from it, one can ascertain the quality of the data base and its ancestral data and derivations, track back sources of errors, allow automated reenactment of derivations to update the data, and provide attribution of data sources"

Simmhan, Plale, and Gannon, "A survey of data provenance in e-science," ACM Sigmod Record, 2005

Provenance (cont)

PROV model

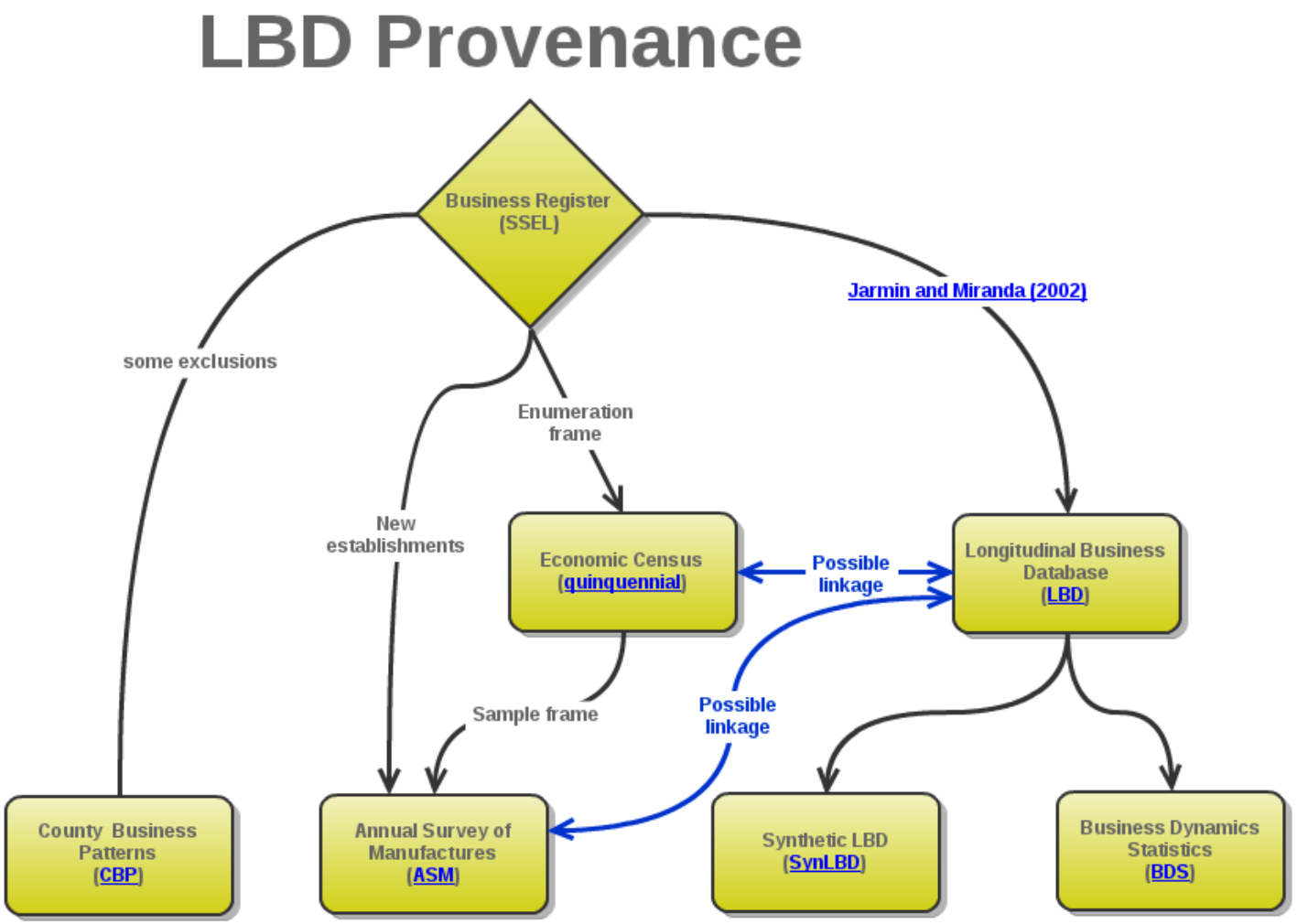
W3C PROV Model based in the notions of

1. **entities** that are physical, digital, and conceptual things in the world;
2. **activities** that are dynamic aspects of the world that change and create entities; and
3. **agents** that are responsible for activities.
4. a set of **relationships** that can exist be- tween them that express attribution,. delegation, derivation, etc.

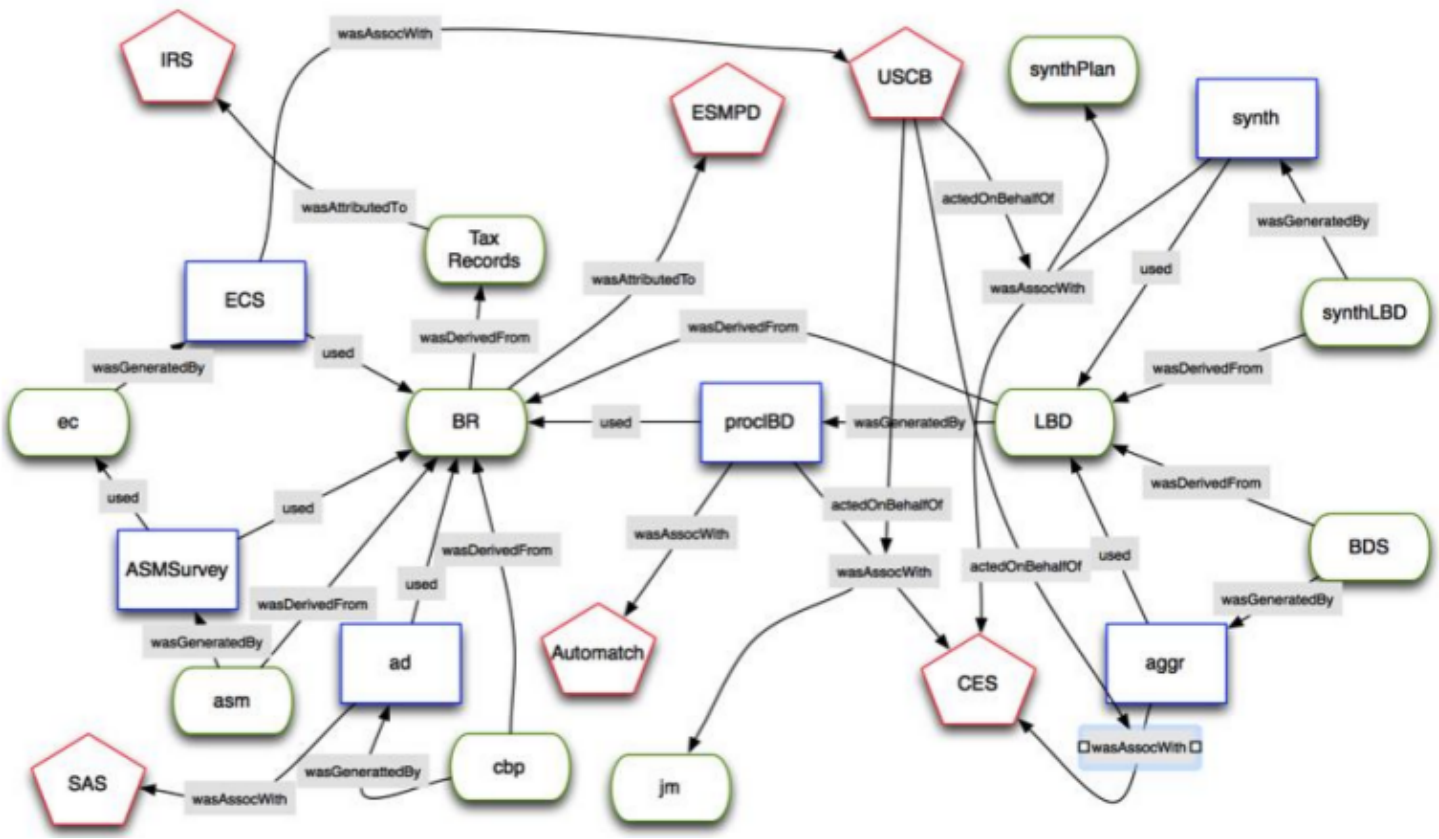
PROV and Metadata

Not (currently) a “native” component of DDI

Incorporating PROV (LBD)



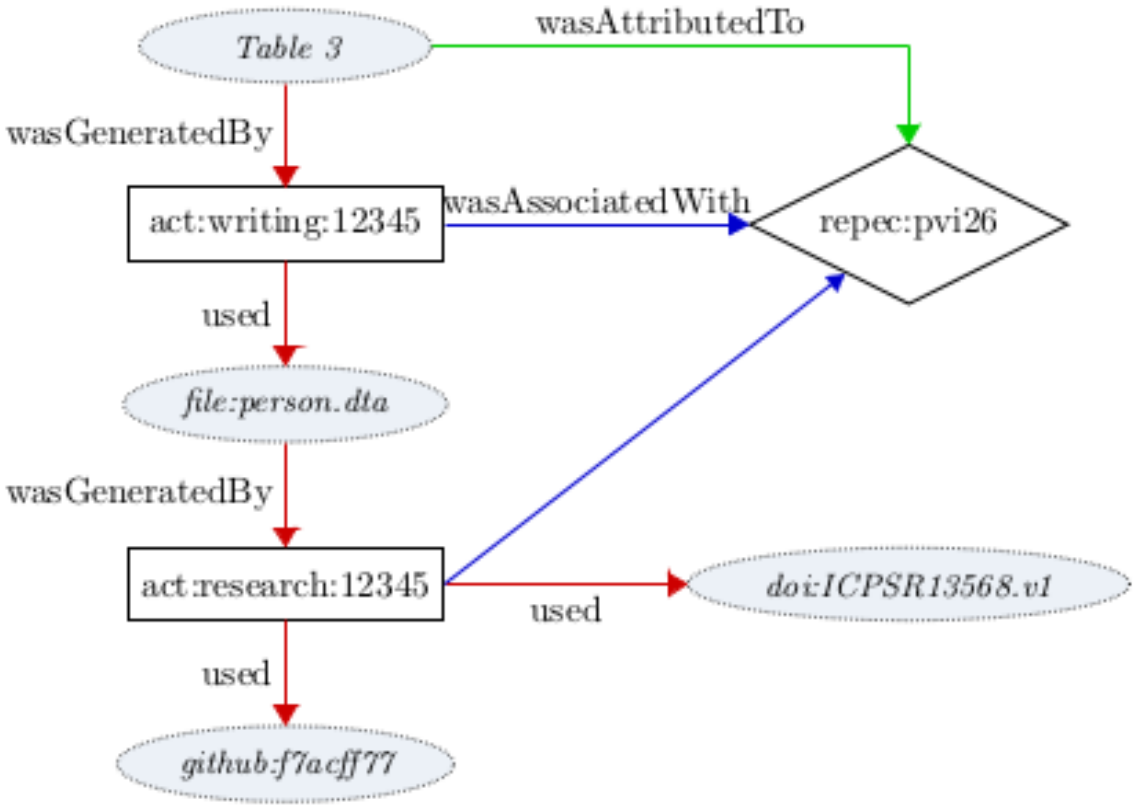
Incorporating PROV (LBD)



Provenance for research

Sample research activity with full provenance

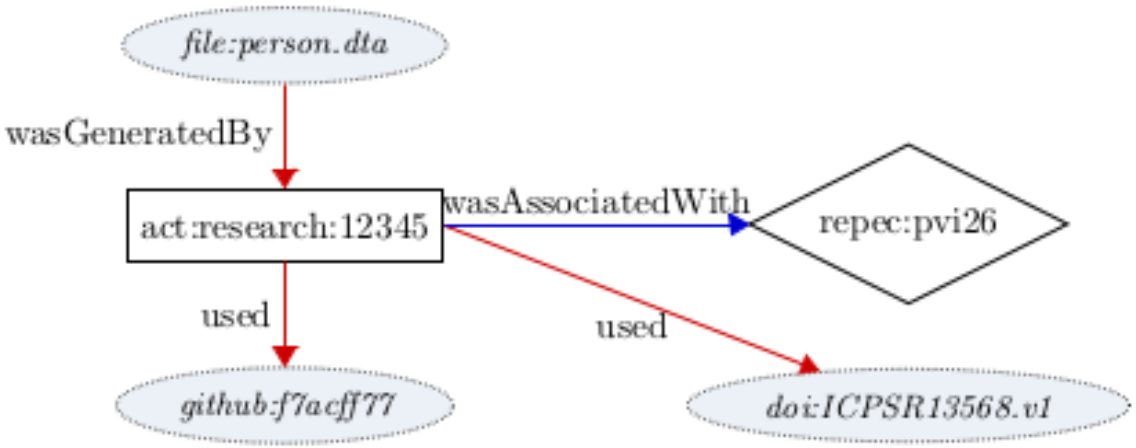
Figure 4: Sample research activity with full provenance



Provenance for research

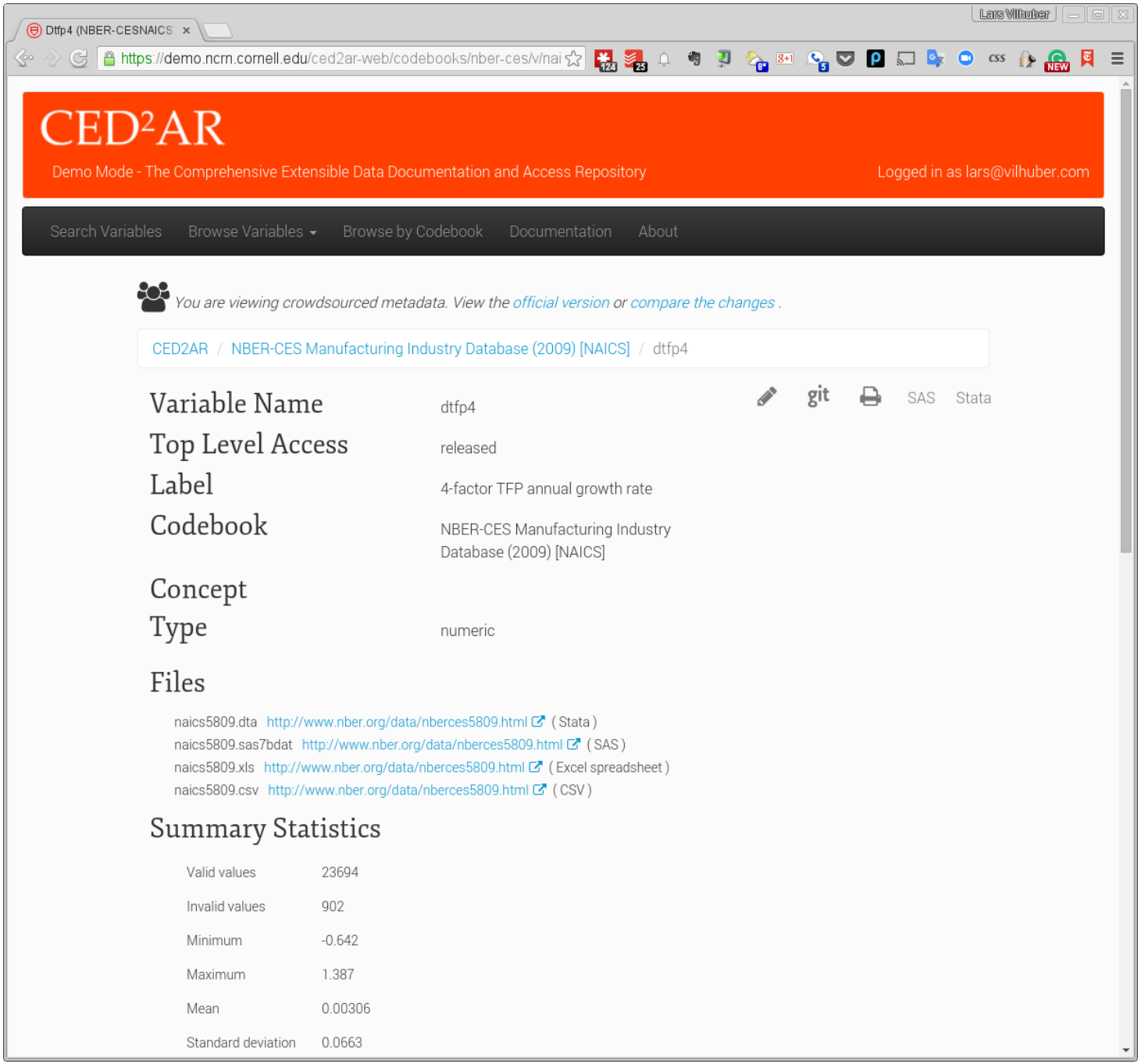
Sample research activity with simple provenance

Figure 5: Sample research activity with simplified provenance



Putting it together

Easy editing of all elements of data description


Dtfp4 (NBER-CESNAICS
Lars Vilhuber
https://demo.ncrn.cornell.edu/ced2ar-web/codebooks/nber-ces/v/nai
CED²AR
Demo Mode - The Comprehensive Extensible Data Documentation and Access Repository
Logged in as lars@vilhuber.com
Search Variables
Browse Variables
Browse by Codebook
Documentation
About
You are viewing crowdsourced metadata. View the official version or compare the changes .
CED2AR / NBER-CES Manufacturing Industry Database (2009) [NAICS] / dtfp4
git
SAS
Stata
Variable Name
dtfp4
Top Level Access
released
Label
4-factor TFP annual growth rate
Codebook
NBER-CES Manufacturing Industry Database (2009) [NAICS]
Concept
Type
numeric
Files
naics5809.dta http://www.nber.org/data/nberces5809.html (Stata)
naics5809.sas7bdat http://www.nber.org/data/nberces5809.html (SAS)
naics5809.xls http://www.nber.org/data/nberces5809.html (Excel spreadsheet)
naics5809.csv http://www.nber.org/data/nberces5809.html (CSV)
Summary Statistics
Valid values 23694
Invalid values 902
Minimum -0.642
Maximum 1.387
Mean 0.00306
Standard deviation 0.0663

Value Ranges

Value Range

Range: [-0.642000019550324 , 1.38699996471405]

Full Description

There are two versions of TFP in the NBER-CES Manufacturing Database: 4-factor and 5-factor. The 5- factor version separates out energy from non-energy materials; the 4-factor uses a single materials input (which includes energy). The TFP calculation requires definitions of the cost shares, the factors, the factor changes, and the output changes. The five cost shares (α_i) vary by industry by year, defined using the variable names in the dataset:

- (α_1) Non-production workers: (pay-prodw)/vship [i.e., their pay divided by shipments]
- (α_2) Production workers: prodw/vship [i.e., their pay divided by shipments]
- (α_3) Energy: energy/vship [i.e., energy expenditure divided by shipments]
- (α_4) Materials: (matcost-energy)/vship [i.e., non-energy materials divided by shipments]
- (α_5) Capital: 1 - (sum of the above shares) [i.e., the residual]

In calculating TFP growth from one year to the next, we use the average of the two years' cost shares: $\overline{\alpha}_{it} = 0.5\alpha_{it} + 0.5\alpha_{it-1}$ The 5 factors (X_i) are defined as follows, using the variable names in the dataset:

- (X_1) Non-production workers: (emp-prode) [i.e., the number of non-production workers]
- (X_2) Production workers: prodh [i.e., production worker hours, not employees]
- (X_3) Energy: (energy/pien) [i.e., real energy expenditures]
- (X_4) Materials: ((matcost/pimat) - (energy/pien)) [i.e., real non-energy materials]
- (X_5) Capital: cap [i.e., total capital stock, already in real terms]

The change in factor usage between one year and the next is defined as the change in natural logs, (for example):
$dX_{it} = \ln(X_{it}) - \ln(X_{it-1})$

We also need the change in real output: (Q) Real output: *vship/piship*

As with factor usage, we express output change in terms of natural logs, hence: $dQ_t = \ln(Q_t) - \ln(Q_{t-1})$

The change in 5-factor TFP (dTFP5) between this year and last is thus defined as: $dTFP5_t = dQ_t - \sum_i (\alpha'_{it} dX_{it}), i = 1, ..., 5$

Given the series of dTFP5 values, one can then "roll up" these changes to form a TFP index (TFP5), by setting the index equal to 1.0 in some initial year t and then growing the index forward by the following equation: $TFP5_{t+1} = \exp[\ln(TFP5_t) + (dTFP5_{t+1})]$

The values of 4-factor TFP growth (dTFP4) and the corresponding TFP index (TFP4) are calculated similarly, but using total materials cost spending rather than separating it into energy and non-energy materials.

Value Ranges

Value Range

Range: [-0.642000019550324 , 1.38699996471405]

Full Description

There are two versions of TFP in the NBER-CES Manufacturing Database: 4-factor and 5-factor. The 5- factor version separates out energy from non-energy materials; the 4-factor uses a single materials input (which includes energy). The TFP calculation requires definitions of the cost shares, the factors, the factor changes, and the output changes. The five cost shares (α_i) vary by industry by year, defined using the variable names in the dataset:

- (α_1) Non-production workers: (pay-prodw)/vship [i.e., their pay divided by shipments]
- (α_2) Production workers: prodw/vship [i.e., their pay divided by shipments]
- (α_3) Energy: energy/vship [i.e., energy expenditure divided by shipments]
- (α_4) Materials: (matcost-energy)/vship [i.e., non-energy materials divided by shipments]
- (α_5) Capital: 1 - (sum of the above shares) [i.e., the residual]

In calculating TFP growth from one year to the next, we use the average of the two years' cost shares: $\overline{\alpha}_{it} = 0.5\alpha_{it} + 0.5\alpha_{it-1}$ The 5 factors (X_i) are defined as follows, using the variable names in the dataset:

- (X_1) Non-production workers: (emp-prode) [i.e., the number of non-production workers]
- (X_2) Production workers: prodh [i.e., production worker hours, not employees]
- (X_3) Energy: (energy/pien) [i.e., real energy expenditures]
- (X_4) Materials: ((matcost/pimat) - (energy/pien)) [i.e., real non-energy materials]
- (X_5) Capital: cap [i.e., total capital stock, already in real terms]

The change in factor usage between one year and the next is defined as the change in natural logs, (for example):
$dX_{it} = \ln(X_{it}) - \ln(X_{it-1})$

We also need the change in real output: (Q) Real output: *vship/piship*

As with factor usage, we express output change in terms of natural logs, hence: $dQ_t = \ln(Q_t) - \ln(Q_{t-1})$

The change in 5-factor TFP (dTFP5) between this year and last is thus defined as: $dTFP5_t = dQ_t - \sum_i (\alpha'_{it} dX_{it}), i = 1, ..., 5$

Given the series of dTFP5 values, one can then "roll up" these changes to form a TFP index (TFP5), by setting the index equal to 1.0 in some initial year t and then growing the index forward by the following equation: $TFP5_{t+1} = \exp[\ln(TFP5_t) + (dTFP5_{t+1})]$

The values of 4-factor TFP growth (dTFP4) and the corresponding TFP index (TFP4) are calculated similarly, but using total materials cost spending rather than separating it into energy and non-energy materials.

Lacking from other implementations

... such as

Dataverse

Lars Vilhuber Dataverse (Cornell University)

Harvard Dataverse > Lars Vilhuber Dataverse >

Replication data for: National estimates of gross employment and job flows from the Quarterly

Metrics 4 Downloads

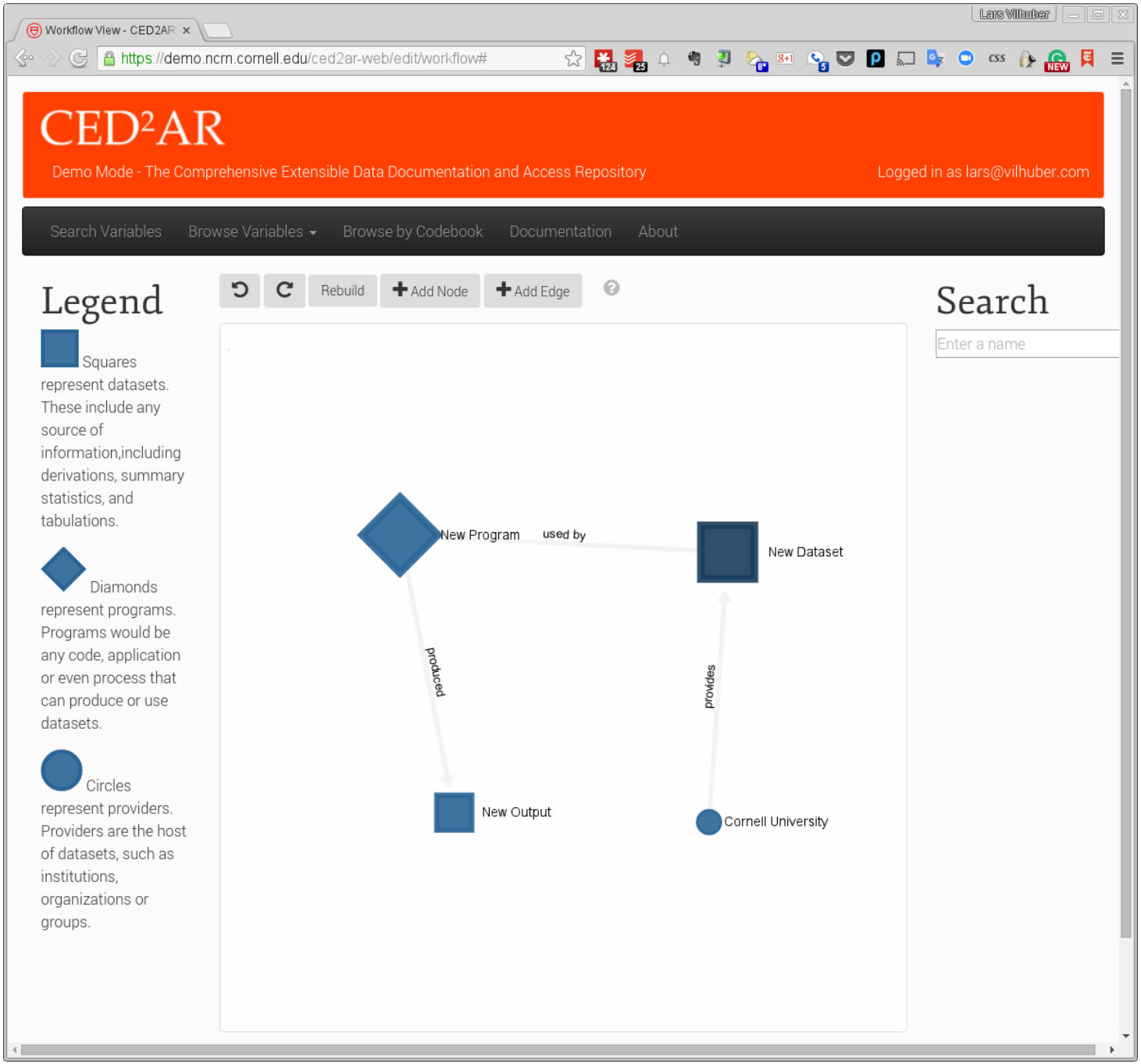
Replication data for: National estimates of gross employment and Indicators with demographic and industry detail

Abowd, John M.; Vilhuber, Lars, 2014, "Replication data for: National estimates of gross employment ar Workforce Indicators with demographic and industry detail", http://dx.doi.org/10.7910/DVN/27923, Harv

If you use these data, please add this citation to your scholarly resources. Learn about Data Citation Standards.

Description

Editing of provenance


Workflow View - CED2AR
Lars Vilhuber
https://demo.ncrn.cornell.edu/ced2ar-web/edit/workflow#
CED²AR
Demo Mode - The Comprehensive Extensible Data Documentation and Access Repository
Logged in as lars@vilhuber.com
Search Variables
Browse Variables
Browse by Codebook
Documentation
About
Legend
Squares represent datasets. These include any source of information,including derivations, summary statistics, and tabulations.
Diamonds represent programs. Programs would be any code, application or even process that can produce or use datasets.
Circles represent providers. Providers are the host of datasets, such as institutions, organizations or groups.
Rebuild
Add Node
Add Edge
New Program
used by
New Dataset
produced
provides
New Output
Cornell University
Search
Enter a name

Possibilities

Enhance journal or working paper archives

- Capture the essential elements of programs, data, and how they are linked

Machine readable!

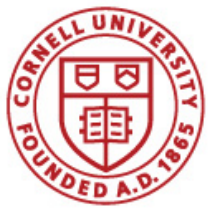
Because the metadata is structured, actionable data ensues

- Reproducible archives!
- Disclosure avoidance requests (Census RDC, German RDC require such documentation, but currently unstructured)

Additional elements

Ex-post linking of articles and data

Keyword	Employment Dynamics
Topic Classification	Economics
Related Publication	John M. Abowd and Lars Vilhuber, "National estimates of gross employment and job flows from the Quarterly Wor with demographic and industry detail," Journal of Econometrics, vol. 161, iss. 1, pp. 82-99, 2011. doi: 10.1016/j.jeconom.2010.09.008 http://www2.vrdc.cornell.edu/news/data/qwi-national-data/
	John M. Abowd and Lars Vilhuber, "National estimates of gross employment and job flows from the Quarterly Wor with demographic and industry detail," Journal of Econometrics, vol. 161, iss. 1, pp. 82-99, 2011. doi: 10.1016/j.jeconom.2010.09.008 http://www2.vrdc.cornell.edu/news/data/qwi-national-data/
	John M. Abowd and Lars Vilhuber, "National estimates of gross employment and job flows from the Quarterly Wor with demographic and industry detail (with color graphs)," Center for Economic Studies, U.S. Census Bureau, Wc 11, 2010. http://ideas.repec.org/p/cen/wpaper/10-11.html
Producer	Labor Dynamics Institute (Cornell University) (LDI) http://www2.vrdc.cornell.edu/news/data/qwi-national-data/



Additional elements

Ex-post linking of articles and data

- Lacking from existing repositories of both data and bibliographies
- Exposure of data providers
- Sometimes manually (labor intensive) performed by data archives (e.g. ICPSR)
- Not currently done on RePEc

Crowd-sourcing data provenance

Let other people contribute

RePEc Genealogy

IDEAS | PAPERS | ARTICLES | ALL AUTHORS | AUTHORS IN GENEALOGY | AMEND GENEALOGY | INSTITUTIONS | DATA (FRED®) |

The RePEc genealogy is a RePEc service hosted by the Research Division of the Federal Reserve Bank of St. Louis

RePEc Genealogy

- Genealogy home
- Listed authors
- FAQ
- Orphan advisors
- Statistics
- **Make additions and changes**
- Links

More services

- IDEAS: Economics research
- EDIRC: Economics institutions
- NEP: New Economics papers by email
- Author registration

RePEc Genealogy page for Lars Vilhuber

This page traces who advised whom during graduate studies for Lars Vilhuber (RePEc Genealogy, EconPapers, IDEAS). You can help amend this and other pages of this project here. You can also look at the page's history.

Graduate studies

Lars Vilhuber got the terminal degree from Département de Sciences Économiques, Université de Montréal, Montréal, Canada in 1999.

Advisor

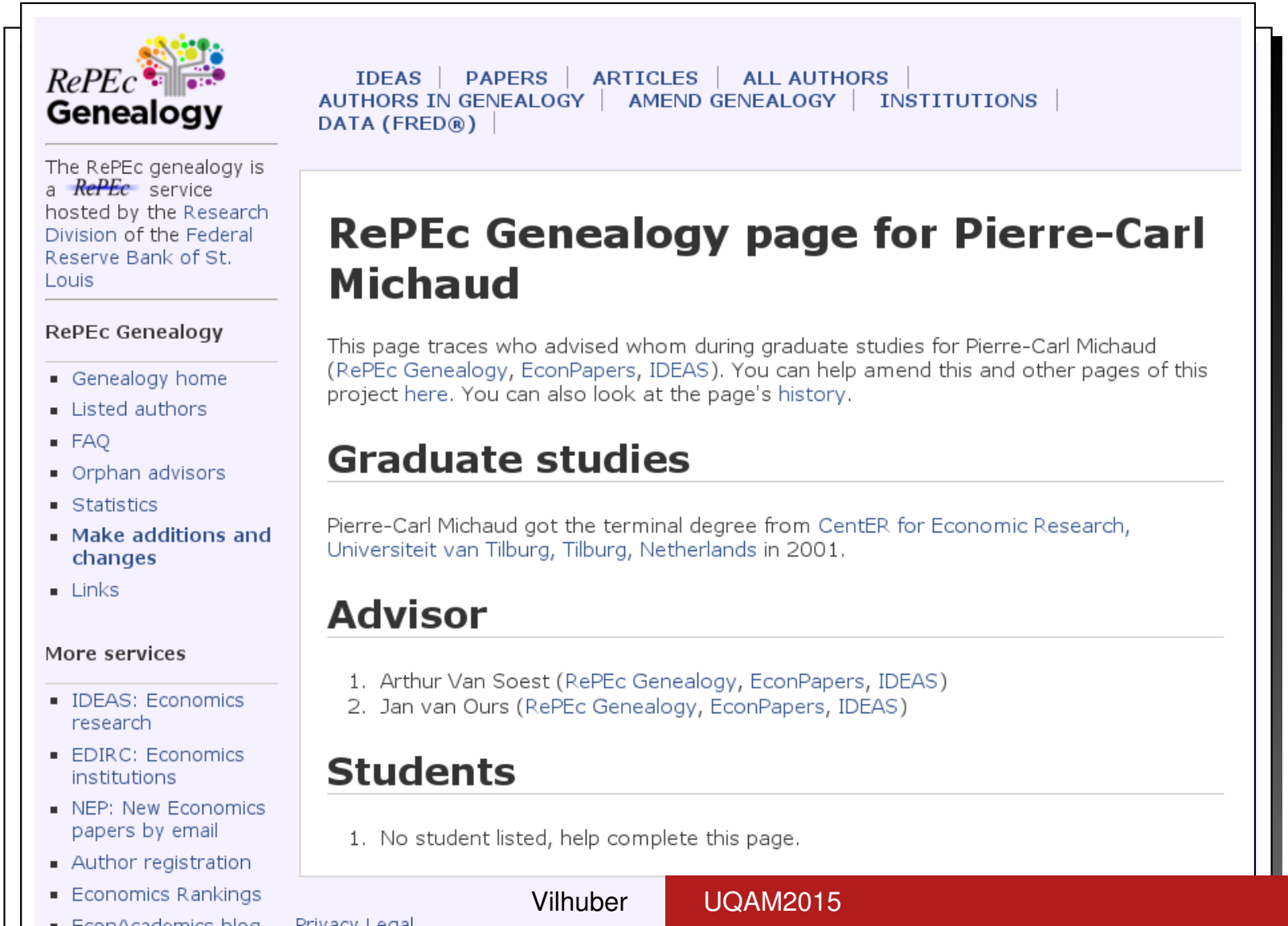
1. David N. Margolis (RePEc Genealogy, EconPapers, IDEAS)
2. Thomas Lemieux (RePEc Genealogy, EconPapers, IDEAS)

Students

1. No student listed, help complete this page.

Crowd-sourcing data provenance

Let other people contribute



RePEc Genealogy

IDEAS | PAPERS | ARTICLES | ALL AUTHORS | AUTHORS IN GENEALOGY | AMEND GENEALOGY | INSTITUTIONS | DATA (FRED®) |

The RePEc genealogy is a RePEc service hosted by the Research Division of the Federal Reserve Bank of St. Louis

RePEc Genealogy

- Genealogy home
- Listed authors
- FAQ
- Orphan advisors
- Statistics
- **Make additions and changes**
- Links

More services

- IDEAS: Economics research
- EDIRC: Economics institutions
- NEP: New Economics papers by email
- Author registration
- Economics Rankings

Privacy Legal

RePEc Genealogy page for Pierre-Carl Michaud

This page traces who advised whom during graduate studies for Pierre-Carl Michaud (RePEc Genealogy, EconPapers, IDEAS). You can help amend this and other pages of this project here. You can also look at the page's history.

Graduate studies

Pierre-Carl Michaud got the terminal degree from CentER for Economic Research, Universiteit van Tilburg, Tilburg, Netherlands in 2001.

Advisor

1. Arthur Van Soest (RePEc Genealogy, EconPapers, IDEAS)
2. Jan van Ours (RePEc Genealogy, EconPapers, IDEAS)

Students

1. No student listed, help complete this page.

Crowd-sourcing data provenance

Work in progress: on RePEc

- Deploy a graphical interface that maps co-author networks, genealogy...
- ... and data provenance
  - incoming: what data did an article use? (LDI Replication workshop scaled up)
  - outgoing: what data did an article create? (Better tracking of replication archives, or the National QWI example)
- Users (or contributors!) can “claim” data, or if hosted on a data repository.

Other methods and efforts

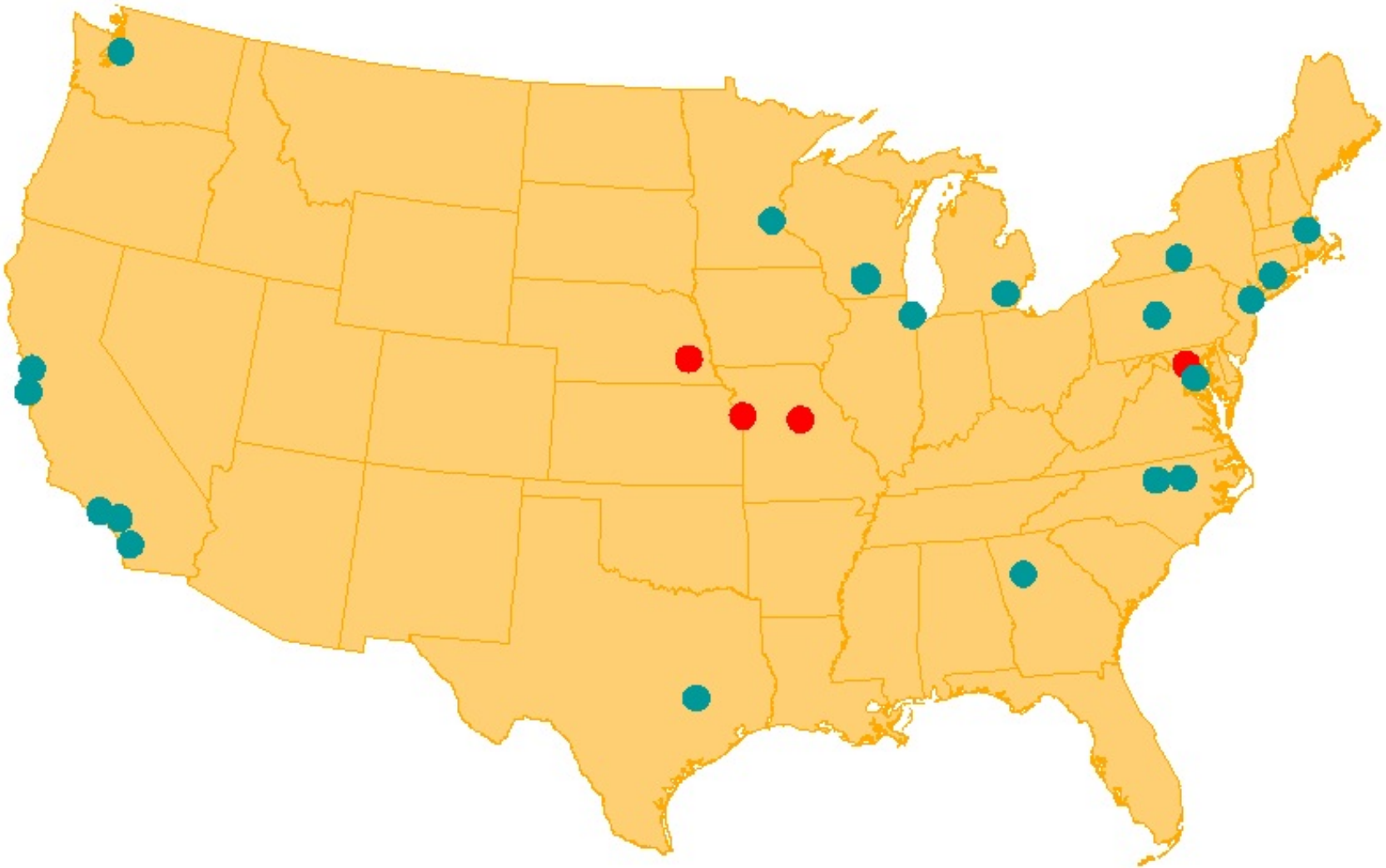
Similar linkage efforts

- RD-Switchboard, based on ORCID IDs
- Direct DataCite/ORCID efforts

... we've only barely started...

Confidentiality

Limitations of restricted data access



Limitations of restricted data access

Users with access to (federal) confidential data in the US

There are **21** (as of 2015-11-09) Federal Research Data Centers (RDCs) in the US. There are approximately 300 researchers with access at any given time. (IRS: 12, BLS: 20?). There are currently **6** servers with total of 200+ CPUs available.

Limitations of restricted data access

Users with access to (federal) confidential data in the US

There are **21** (as of 2015-11-09) Federal Research Data Centers (RDCs) in the US. There are approximately 300 researchers with access at any given time. (IRS: 12, BLS: 20?). There are currently **6** servers with total of 200+ CPUs available.

Users with access to public-use data

There are **20-30** thousand economists in the US. If they each have access to reasonably modern desktop, they have **120k** CPUs. Not counting compute clusters.

Who wants to sit in this?

UK efforts

Who wants to sit in this?

Src: Univ. Edinburgh – Micro, remote, safe settings (safePODS) – extending a safe setting network across a country

Data liberation!

Data curators trade off

- Providing detailed and accurate statistics
- Protecting privacy and confidentiality

Data liberation!

Data curators trade off

- Providing detailed and accurate statistics
- Protecting privacy and confidentiality

What is the optimal tradeoff, given the data have already been collected?

Data curator strategies

Limit access

- Let researchers run wild (with models)...
- ... and limit what can be removed (mostly adhoc)
- RDCs
- remote processing with delay and cost

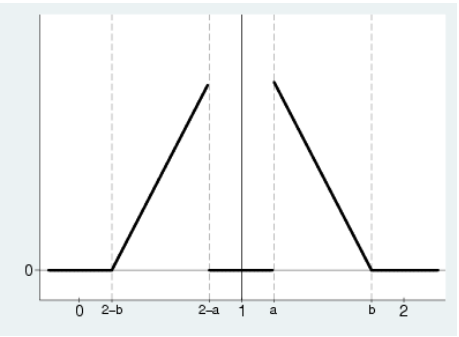
Public-use files

- Disclosure limitation (aggregation, swapping, suppression, etc.)

Some newer methods

Multiplicative Noise Infusion

$$p(\delta_j) = \begin{cases} (b-\delta) \Big/ (b-a)^2, & \delta \in [a,b] \\ (b+\delta-2) \Big/ (b-a)^2, & \delta \in [2-b, 2-a] \\ 0, & \text{otherwise} \end{cases}$$



$$F(\delta_j) = \begin{cases} 0, & \delta < 2-b \\ \left[(\delta+b-2)^2\right] \Big/ \left[2(b-a)^2\right], & \delta \in [2-b, 2-a] \\ 0.5, & \delta \in (2-a, a) \\ 0.5 + \left[(b-a)^2 - (b-\delta)^2\right] \Big/ \left[2(b-a)^2\right], & \delta \in [a,b] \\ 1, & \delta > b \end{cases}$$

where $a = 1 + c/100$ and $b = 1 + d/100$ are constants chosen such that the true value is distorted by a minimum of c percent and a maximum of d percent

Applying noise infusion

Quarterly Workforce Indicators

Published value X_{jt}^* computed from confidential value X_{jt} as

$$X_{jt}^* = \delta_j X_{jt}, \tag{1}$$

See Abowd et al (2009)

Synthetic data (Rubin, 1993; Little, 1993)

Drawing from a posterior predictive distribution

From data $(X, Y))$, where $Y = (Y_{obs}, Y_{nobs})$

$I : i = 0 \iff y \in Y_{nobs}$,

construct PPD as $(Y|X, Y_{obs}, I)$, and

draw Y^*.

Then release $\left(X, Y_k^*\right)$ (k partially synthetic data sets, typically $k > 1$)

Similarity: $\left(X, (Y_{obs}, Y_{nobs}^*)\right)$ (multiply) imputed data

Examples of synthetic microdata

SIPP Synthetic Beta

Survey of Income and Program Participation (SIPP) matched to administrative earnings, then synthesized

Synthetic LBD (SynLBD)

Longitudinal Business Database – longitudinally linked establishment microdata – synthesized

Other uses of synthetic data

American Community Survey tabulations

Group quarters

LEHD Origin-Destination Employment Statistics (LODES)

Synthetic (differentially private) residence information combined with noise-protected establishment counts. (Machanavajjhala et al, 2008)

Key: analytic validity contingent on privacy protection

How well does that work?

LODES

Existing LODES data in OnTheMap application

Employment in Lower Manhattan

Residences of Workers Employed in Lower Manhattan

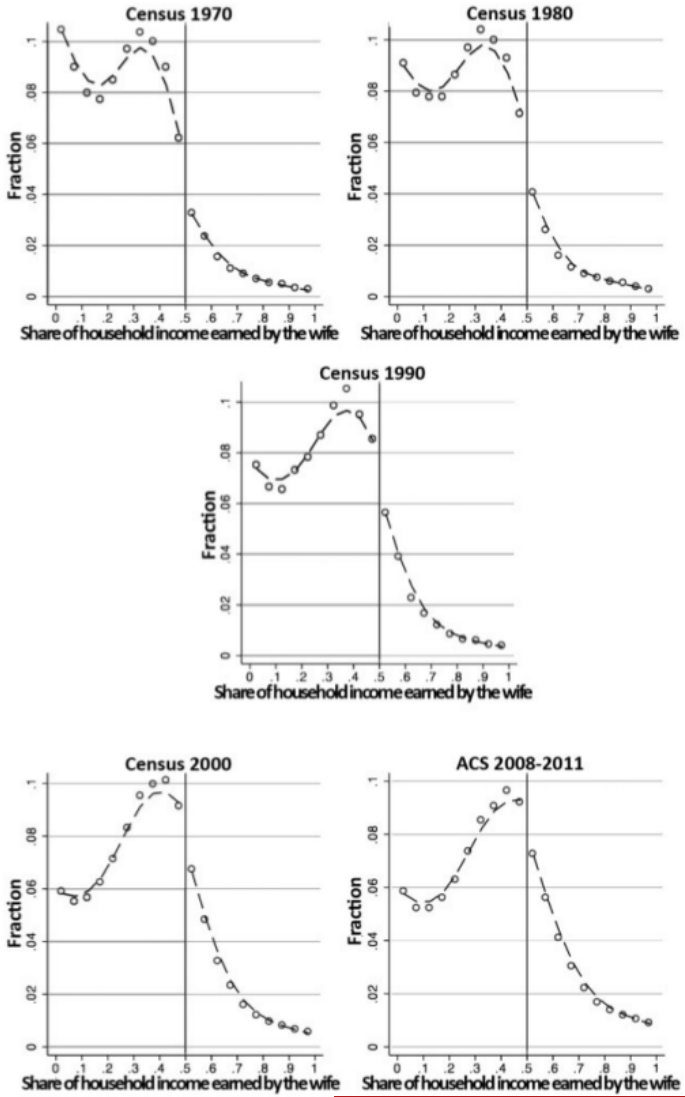
Synthetic Data Server @ Cornell

Open remote access

- Users request account (no restrictions)
- Users run regression on synthetic data
- Users request validation against confidential data

Bertrand et al 2015

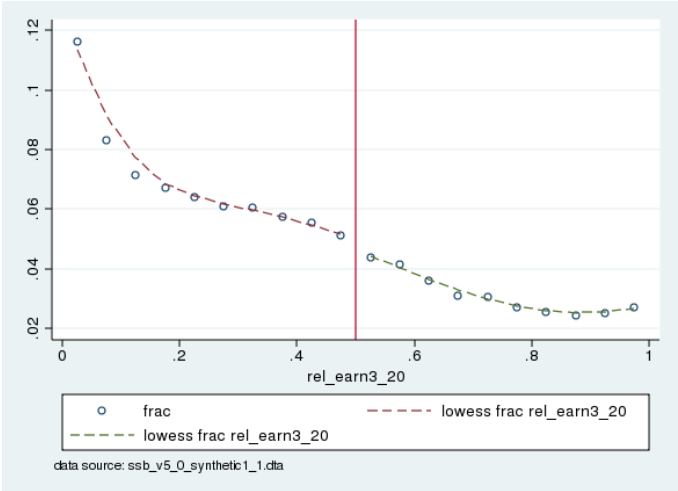
From Bertrand et al (2015)



Bertrand et al 2015

From Bertrand et al (2015), their Figure I

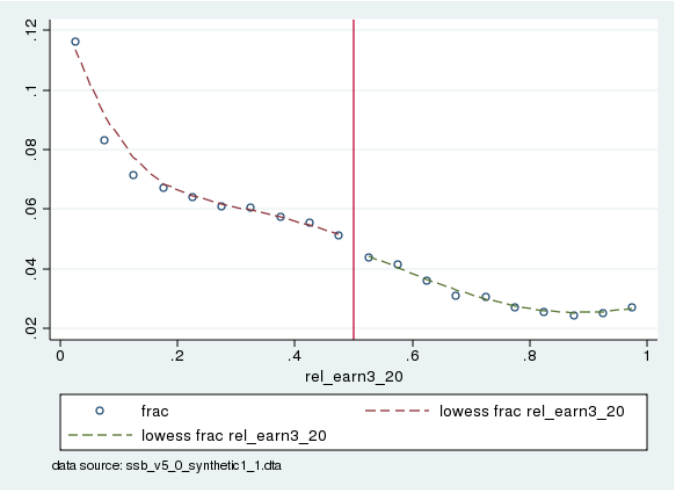
(a) (b)



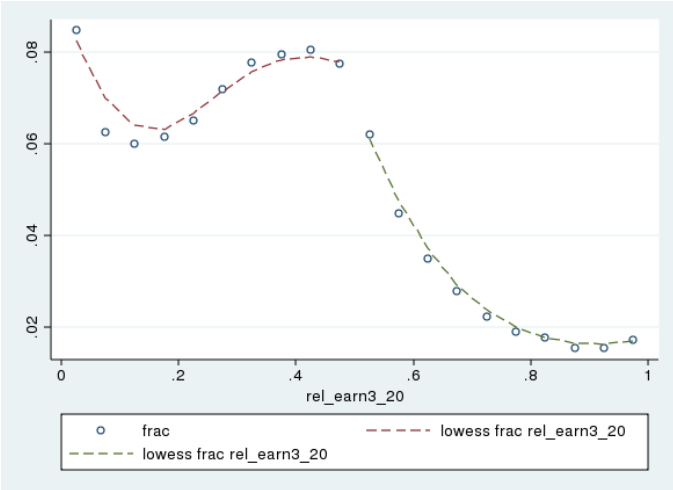
Bertrand et al 2015

From Bertrand et al (2015), their Figure I

(a)



(b)



Synthetic data as a 'blind commitment' device

"Blind analysis: Hide results to seek the truth"

Nature, October 7, 2015 "

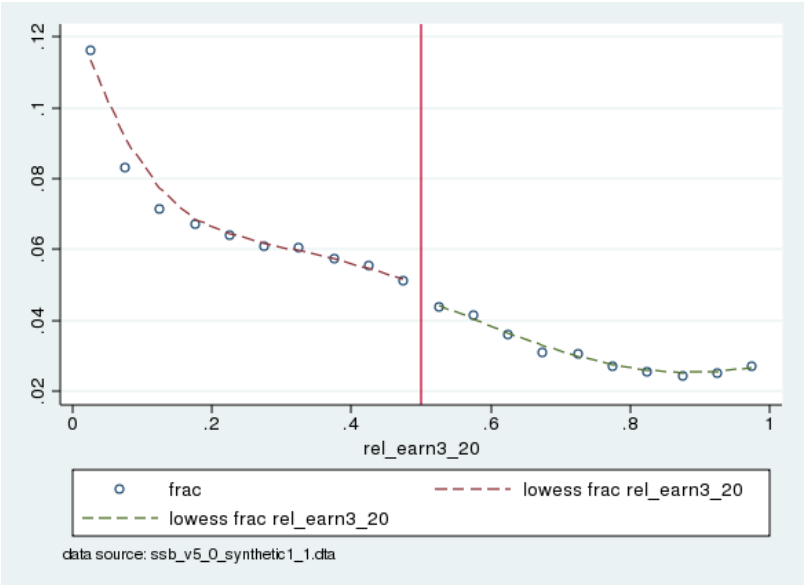
> *temporarily and judiciously removing data labels and altering data values to fight bias and error*

"

Synthetic data together with validation provides such a mechanism.

Bertrand et al 2015

From Bertrand et al (2015), their Figure I

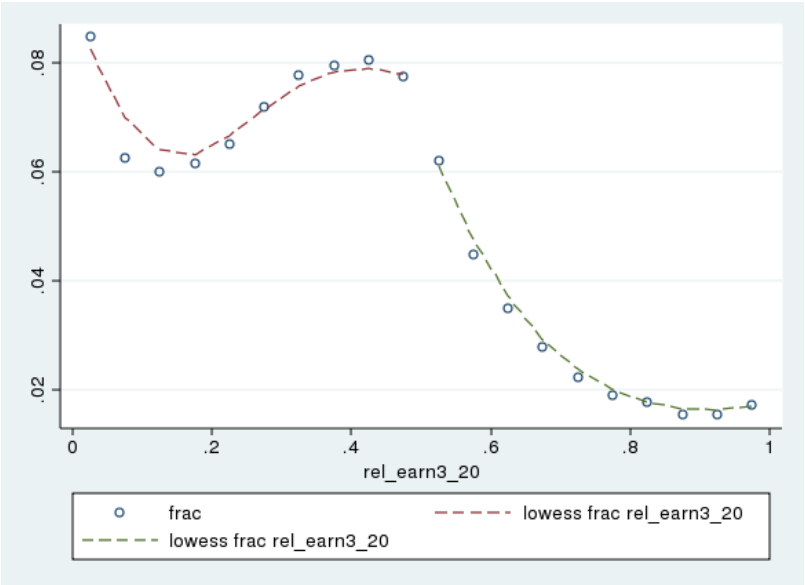


Blind model specification

Bertrand et al 2015

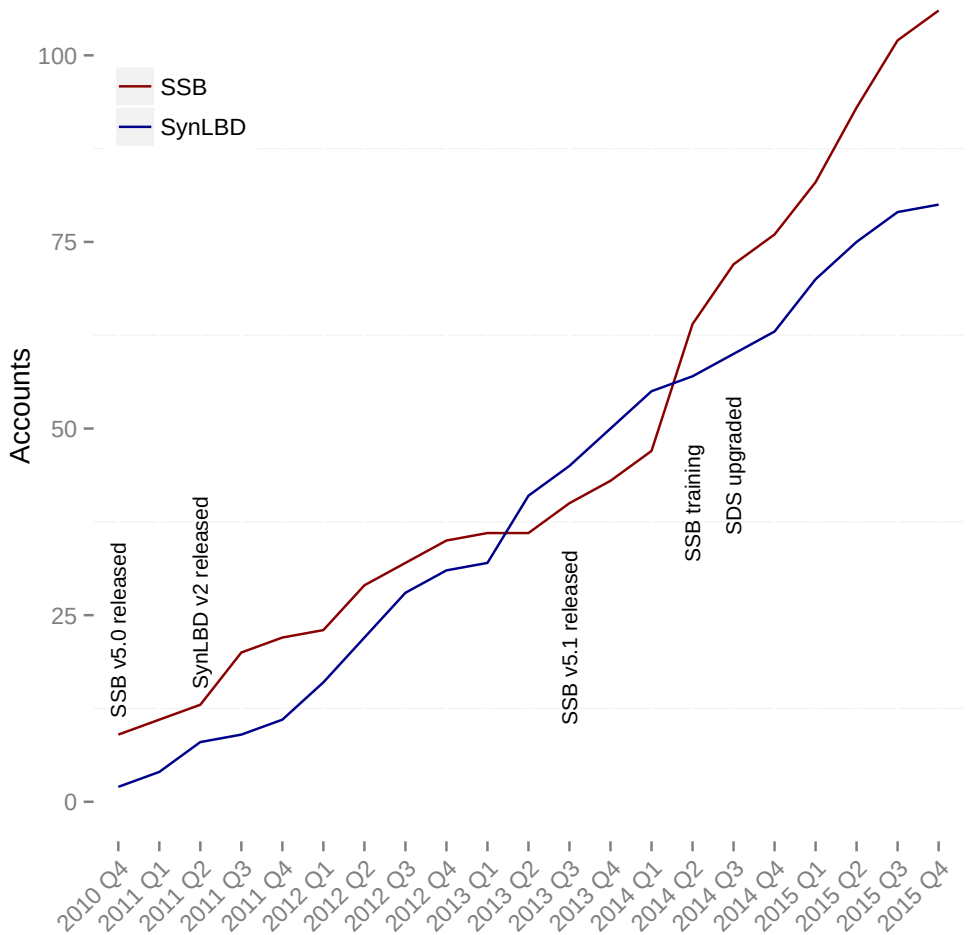
From Bertrand et al (2015), their Figure I

Lifting of veil



Importance of feedback loop

Account creation and events SDS



More general validity results

Consider the overlap of confidence intervals (L, U) for $\beta_{k,m}$ (estimated from the confidential data) and (L^*, U^*) for $\beta^*_{k,m}$ (from the synthetic data).

Confidence interval overlap (Karr et al 2006)

Let $L^{over} = \max(L, L^*)$

Let $U^{over} = \min(U, U^*)$.

Compute $J_{k,m}$ for parameter k in model m.

Then the average overlap in confidence intervals

$$J^*_{k,m} = \frac{1}{2}\left[\frac{U^{over} - L^{over}}{U - L} + \frac{U^{over} - L^{over}}{U^* - L^*}\right]$$

We then average $J^*_{k,m}$ over all estimated models and parameters

Results from 3000 models and 14000 parameters

Table: Confidence interval overlap $J^*_{k,m}$

User	Request	Mean	75th	90th	Max
A	1	0.160	0.246	0.725	0.889
A	2	0.101	0	0.523	0.924
BC	1	0.219	0.509	0.725	0.995

Caution: large number of queries exhaust the “privacy budget”

Protection against all possible queries

Differential privacy

Let $\mathcal{M}$ be a randomized algorithm. Let D and D' be tables that differ in the presence of a single record (*neighbors*). $\mathcal{M}$ satisfies (ϵ, δ)-differential privacy if for all $S \subseteq \mathit{range}(\mathcal{M})$,

$$\log \frac{\Pr[\mathcal{M}(D) \in S]}{\Pr[\mathcal{M}(D') \in S] + \delta} \leq \epsilon$$

δ allows for the ratio of probabilities to be unbounded with a small failure probability. To avoid algorithms that disclose individual records, δ should be set smaller than $1/n$.

Information content is limited

Sequence of queries matters

- Order matters!
- Data custodian must decide which queries (=tables) to release first
- Then leave remaining privacy budget to researchers (?)

No free lunch

No information can be released without some privacy loss.

Accuracy

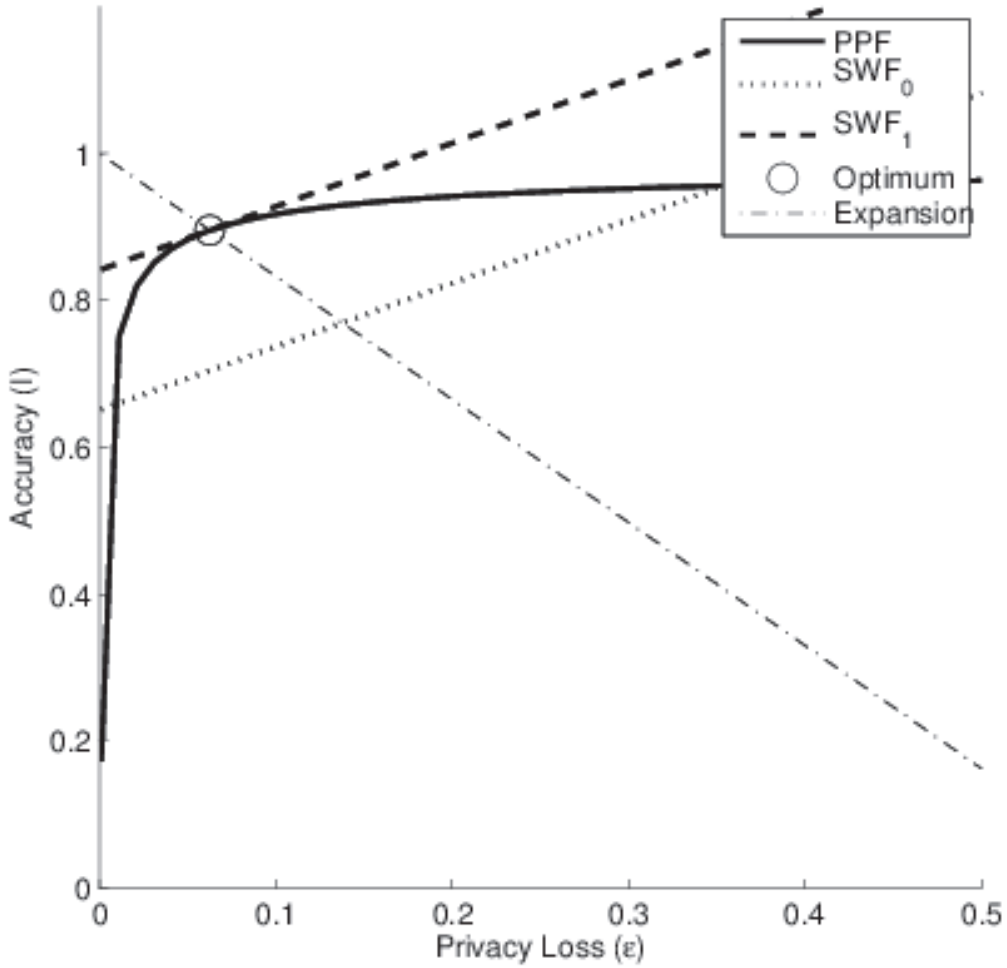
Definition ((α, β)-accuracy)

A query release mechanism M satisfies (α, β)-accuracy for query sequence $\{f_1, f_2, \ldots, f_k\} \in \mathcal{F}^k$, $0 < \alpha \leq 1$, and $0 < \beta \leq 1$, if

$$\min_{1 \leq i \leq k} \{\Pr[|a_i - f_i(x)| \leq \alpha]\} \geq 1 - \beta.$$

Abowd and Schmutte

Model the demand for accuracy (social welfare function SWF)



Technology for Anonymization

Intuition: Online Query Mechanism

1. User sends query

2. Mechanism returns random output conditional on
   - database
   - history

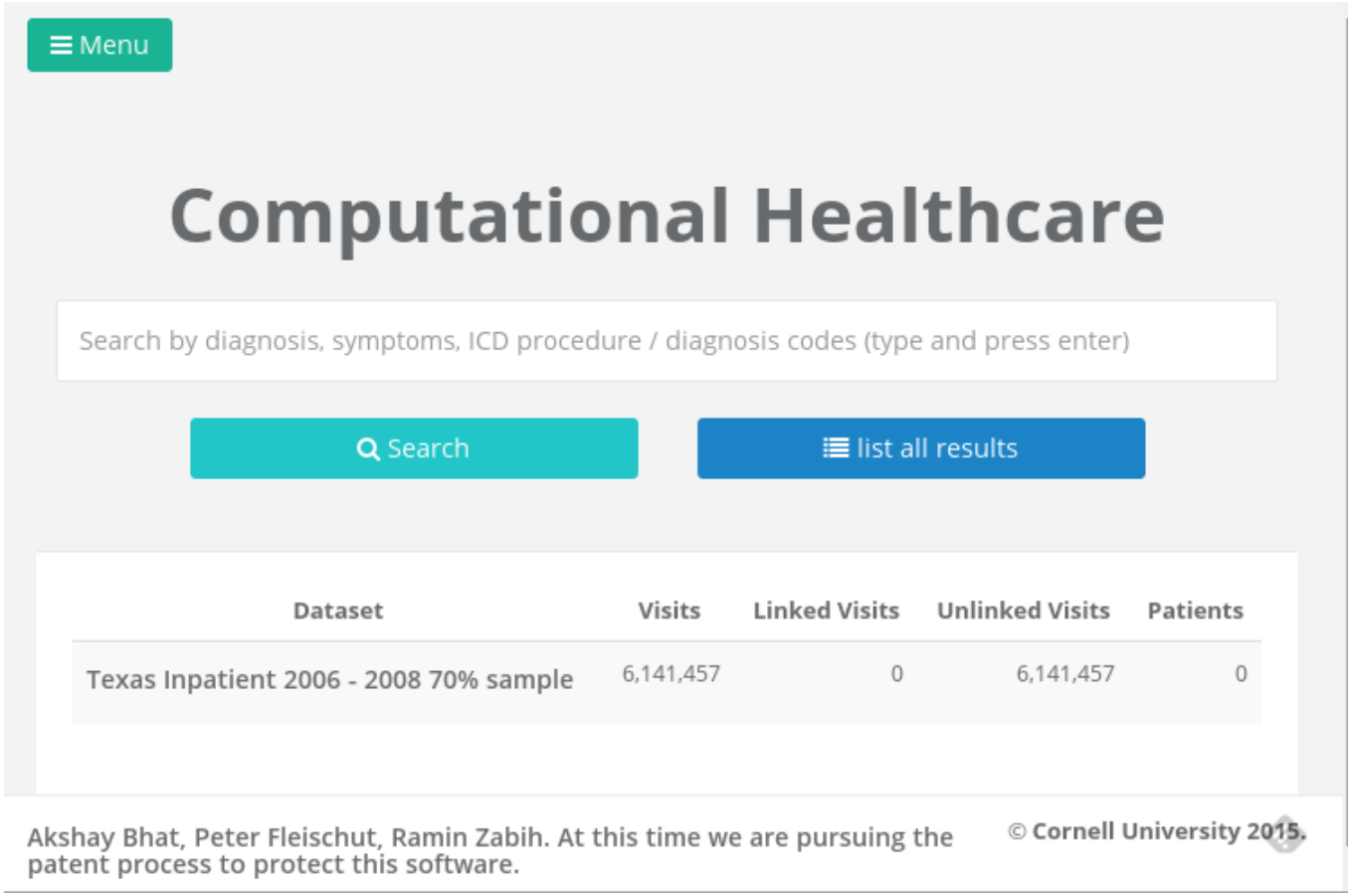
3. Use mechanisms that are provably *differentially private*

Relevancy to medical applications

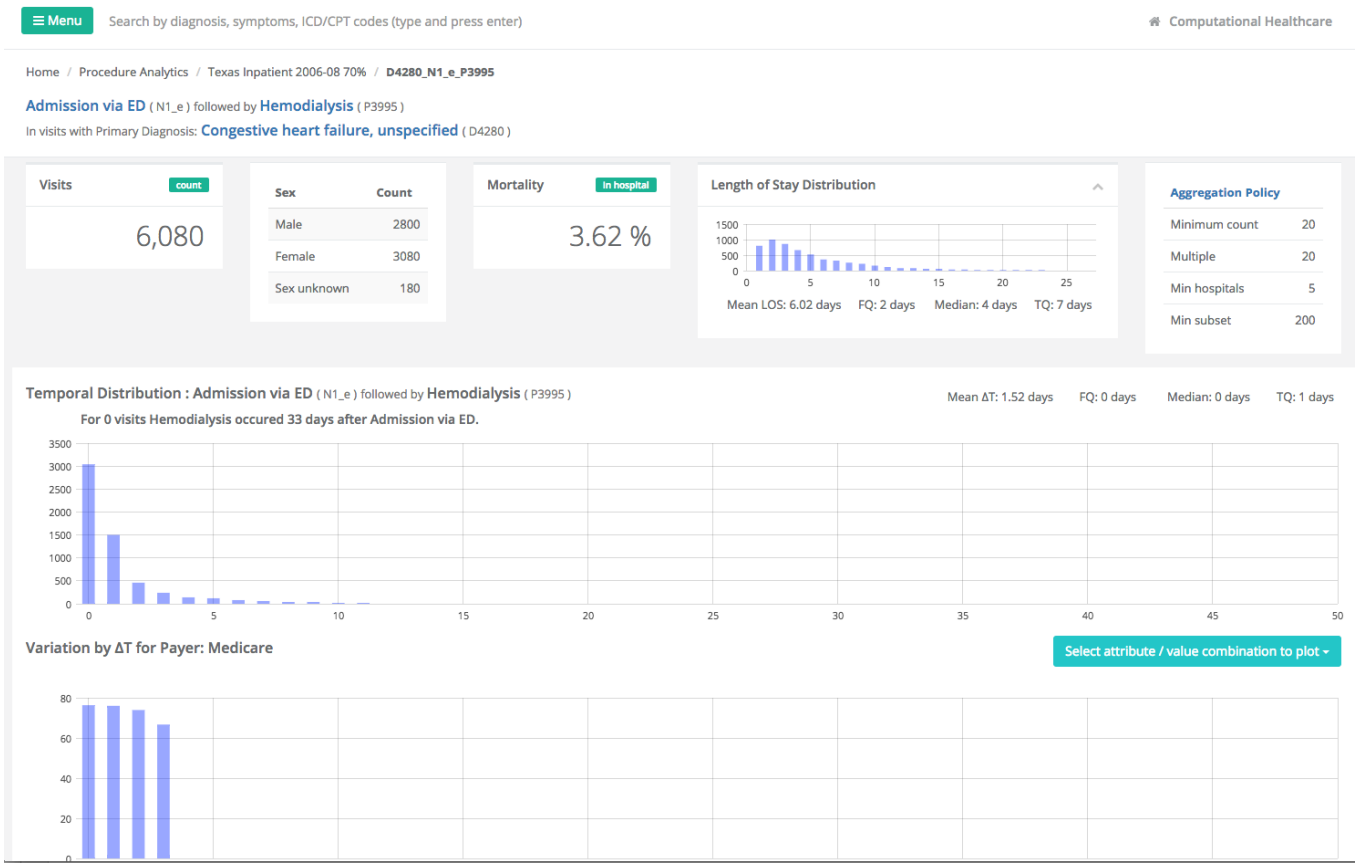
Confidentiality and socio-medical data

- Restricted-access: e.g. Health and Retirement Survey (HRS) biomarkers (same level of confidentiality as other more detailed data)
- Restricted remote access (remote data enclave): health insurance (“all-payer”) claims data (APCDs) [Health Care Cost Institute (HCCI)]
- Trade-off: Midlife in the United States (MIDUS) coarsens geography, but does not modify biomarkers

Relevancy to medical applications



Relevancy to medical applications



Interactively exploring the technological frontier

Active use is critical

Provide users with an online query frontend that interfaces directly with the confidential data, providing differentially private answers. This may still require that all users be authorized users (TBD), and may be appropriate for certain research hospital settings. The benefit would come from agency-signoff on the mechanisms, obviating the need for each user to be an authorized user.

Exhaustion of information content

Once the privacy budget is exhausted through the sequence of queries, any additional queries are rejected (yield a null set), because answering them is no longer possible without decreasing somebody's privacy beyond the allowed limit.

Silver lining

How limiting the mechanism is...

Analyses that are provably dependent upon only the query set used to generate the current generation of the synthetic data are provably analytically valid with accuracy that is a function of the (α,β)-accuracy used to generate the synthetic data.

Conclusion

Tying it together

Guiding light

Make data more accessible, by first-time users and by re-users.

Provenance and synthetic data

Reproducible analysis is key

- In order to simulate Iterative Database Construction (IDC), we need to be able to re-run a suite of analysis.
- Structure imposed by Synthetic Data Server (SDS) is useful
- Actionable metadata is critical for scalability


Merci

Extra slides

Acronyms

HCCI	Health Care Cost Institute
HRS	Health and Retirement Study
HRS	Health and Retirement Survey
MIDUS	Midlife in the United States
SDS	Synthetic Data Server, see `http://www.vrdc.cornell.edu/sds/`