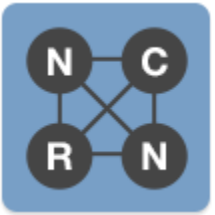
Can Government-Academic Partnerships Help Secure the Future of the Federal Statistical System? Examples from the NSF-Census Research Network

John M. Abowd and Stephen E. Fienberg
Cornell University and Carnegie Mellon University

May 8, 2015 CNSTAT Public Seminar

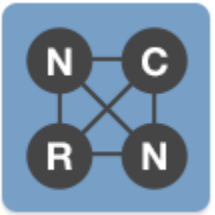
The 21st Century Statistical Agency

- "… the current Census Bureau survey and census methods are unsustainable. Changes must occur in the acquisition of data and construction of statistical information for the Census Bureau to succeed." Robert Groves, Director, Census Bureau, September 8, 2011
- "All parties need to work together to better use all the information we have, whether survey data or big data. Indeed, blending these two types of data creatively will produce new and better ways to inform sound decision making by our nation's businesses, families, and policymakers." Erica Groshen, Commissioner, BLS, March 18, 2015
- "In essence, statistical agencies need to define their primary business as that of providing relevant, accurate, and timely statistical information rather than continuing long-standing data collection and estimation programs for their own sakes." CNSTAT, *Principles and Practices*, 2013 (pp. 42-43)



The Future Is Now

- Paradigm changes are occurring in what citizens and scientists view as the appropriate statistical products of agencies.
  - Sample surveys with tabular summaries, the backbone of the system in the post-WWII era, are being so thoroughly augmented by technology-intensive tools that the resulting outputs are qualitatively different.
  - Consistent time series are being challenged by organic data look-alikes.
- University-based research teams and the federal statistical agencies have a strong mutual interest in navigating this paradigm shift together, just as they did in the 1960s when sample surveys took off.
  - Modern computational tools play the same role now that survey design and implementation did in the 1960s.



NSF-Census Research Network

Innovative, interdisciplinary research in theory, methodology and computational tools

NCRN Coordinating Office
Carnegie-Mellon University
Cornell University
Duke University / National Institute of Statistical Sciences (NISS)
Northwestern University
University of Colorado at Boulder / University of Tennessee
University of Michigan
University of Missouri
University of Nebraska

Latest Seminar

Calendar of events

Example 1: The 369kg (800 lb.) Gorilla. The Decennial Census

- The 2010 US decennial census cost approximately $13.5 billion, nearly double what was spent on the 2000 census, which in turn doubled the 1990 costs.
- The $21 million investment in the NCRN could be recovered many times over from a major process breakthrough.
- The fastest path to a less expensive 2020 census is through the use of more and more up-to-date technology to drastically shorten the timeline between design and implementation, including
  - Human capital of respondents.
  - Human capital of agency.

Traditional Census Bureau Approach

- Starts with the Master Address File (MAF) and works toward people:

  Physical locations→Households→Families→Persons

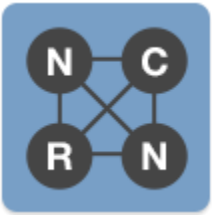
- The process may need to be reversed:

  Physical locations←Households←Families←Persons

  at least in part to make efficient use of online forms to increase accuracy and reduce costs of mail out mail back.
  - **A major focus of CMU node.**
- Two related areas of NCRN collaboration:
  - Record linkage
  - Online survey methods.

Record Linkage

- Record linkage & de-duplication use same statistical techniques.
- Record linkage (or matching) occurs at virtually every stage of census operational and experimental census designs, e.g.:
  - Updating MAF with other address lists – deduplication.
  - Integrating non mail[out mail back forms using MAF.
  - Role in non-response follow-up.
  - Accuracy assessment using matched records for post-enumeration survey.
- Privacy protection: Understanding intruder linkage attacks.
- **New Tools**: CMU, Duke, Michigan and Cornell nodes exploring multiple list generalizations of Fellegi-Sunter and more direct Bayesian methods; one of the major areas of collaboration
  - **Need to propagate uncertainty from matching process into subsequent analyses**.



Role of Online Census and Survey Forms

- Online forms and concerns regarding privacy.
- More far-reaching approach would involve using record linkage and administrative records to prepopulate online census forms.
  - “Record Linkage on the fly.”
- Allow building MAF in new ways, e.g., using modern real-time technology, including remote sensing.
- Major advances in adaptive online surveys (Nebraska node)

Example 2: The 181kg Sibling. The American Community Survey

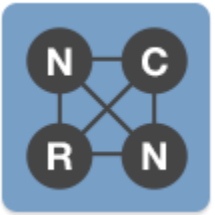
- American Community Survey (ACS) *raison d'être* is the production of timely statistics for detailed subpopulations, including small geographic block-groups.
- Diverse NCRN work on updating many different aspects of ACS design, data collection, and reporting – adapting processes to digital area.
  - multiple online modes, and paradata analysis enter here.
  - **CMU, Nebraska nodes.**
- Reporting at low levels of geography confronts traditional confidentiality restrictions on ACS data.
  - **CMU, Cornell, Duke, Missouri nodes.**

Example 3: Reconceptualizing the Role of Geography in Census and Survey Data

- The geography of census-taking doesn’t correspond to the geography of interest to most users
- Two examples illustrate how an agency can produce more reliable estimates for the geography of interest using inputs from the ACS, including other integrated data

Direct Method

- Based on Missouri node work on spatial-temporal modeling
- Takes the predefined Census geographies and timeline as the observable inputs
- Uses Bayesian hierarchical modelling to produce estimates at the geospatial and temporal points of interest to the analyst
- Can be done with public-use data
- Even more powerful when applied using confidential inputs
- Can be directly integrated with disclosure avoidance procedures and formal privacy protection models



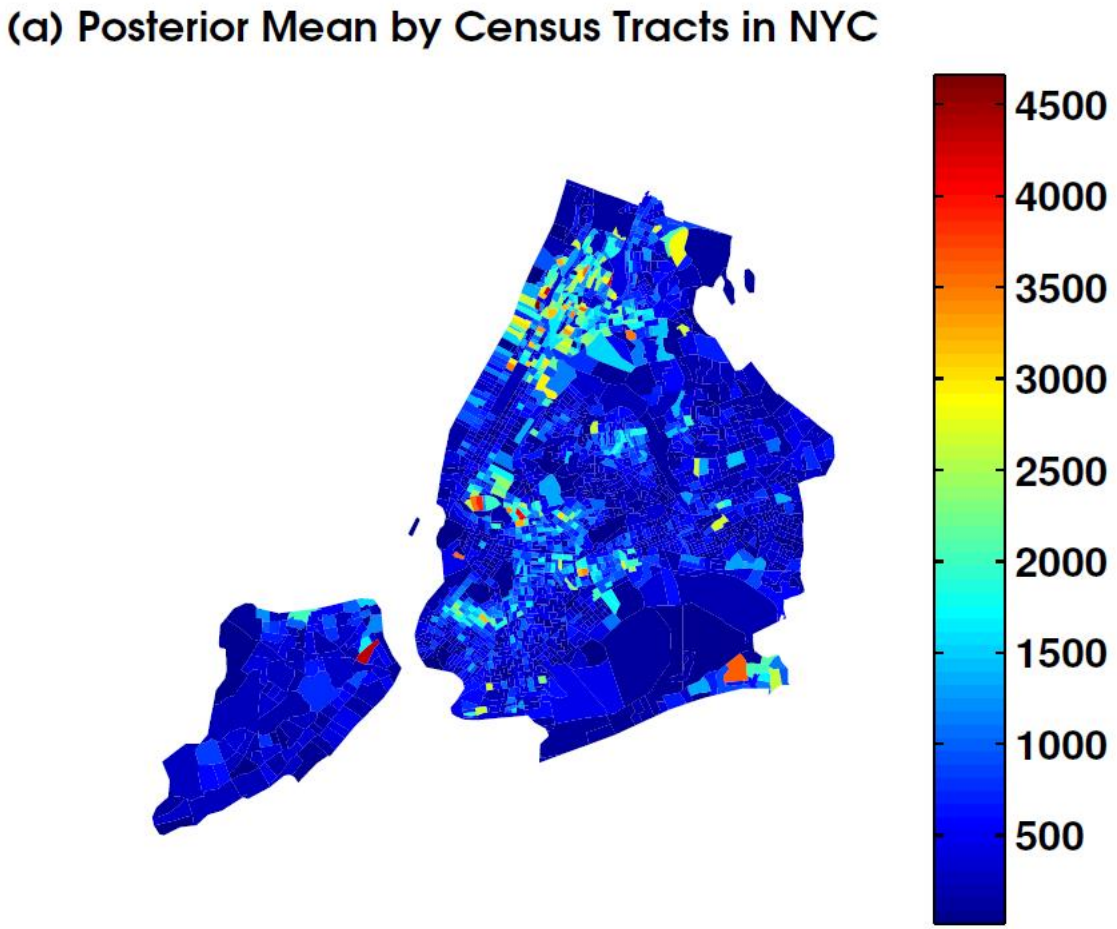
(a) Community District Boundaries in NYC

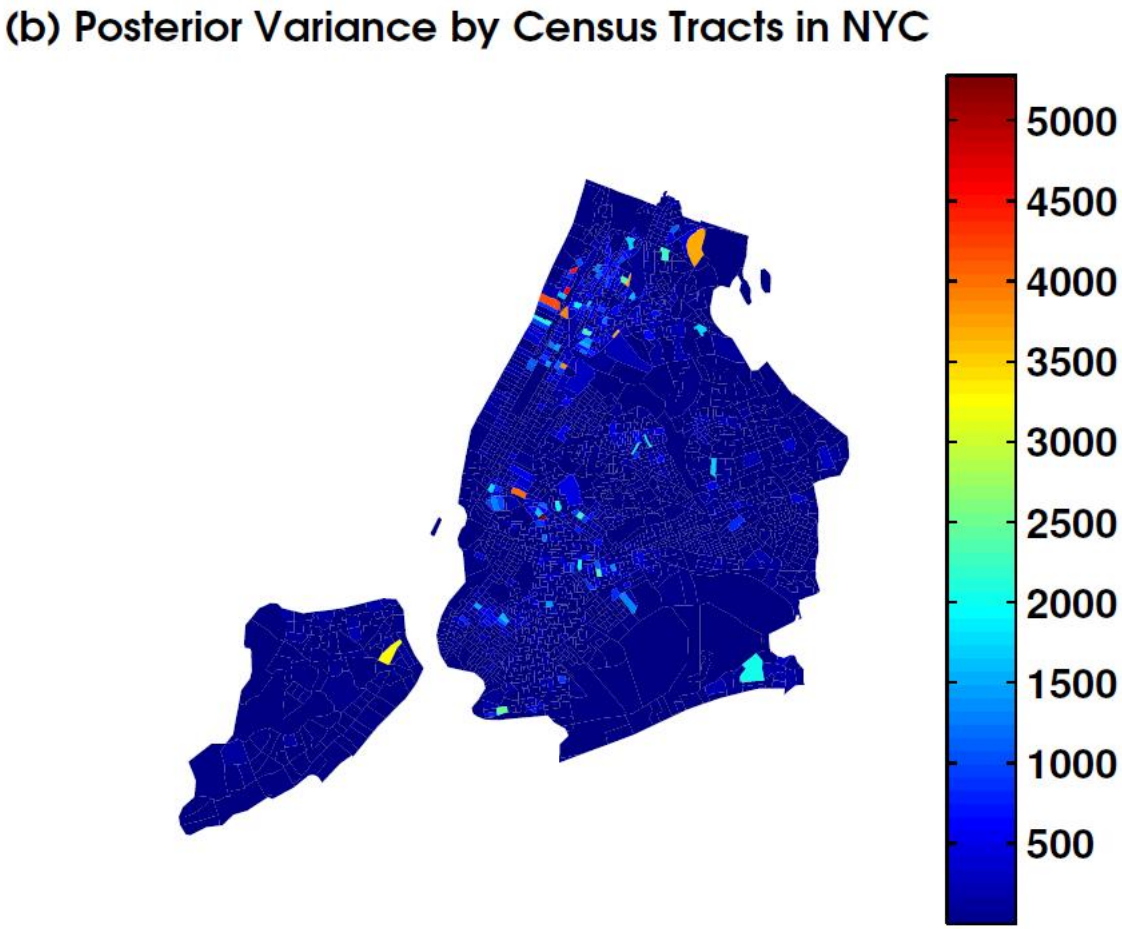
(b) Census Tract Boundaries in NYC

(c) NYC PUMA/Community District Overlap

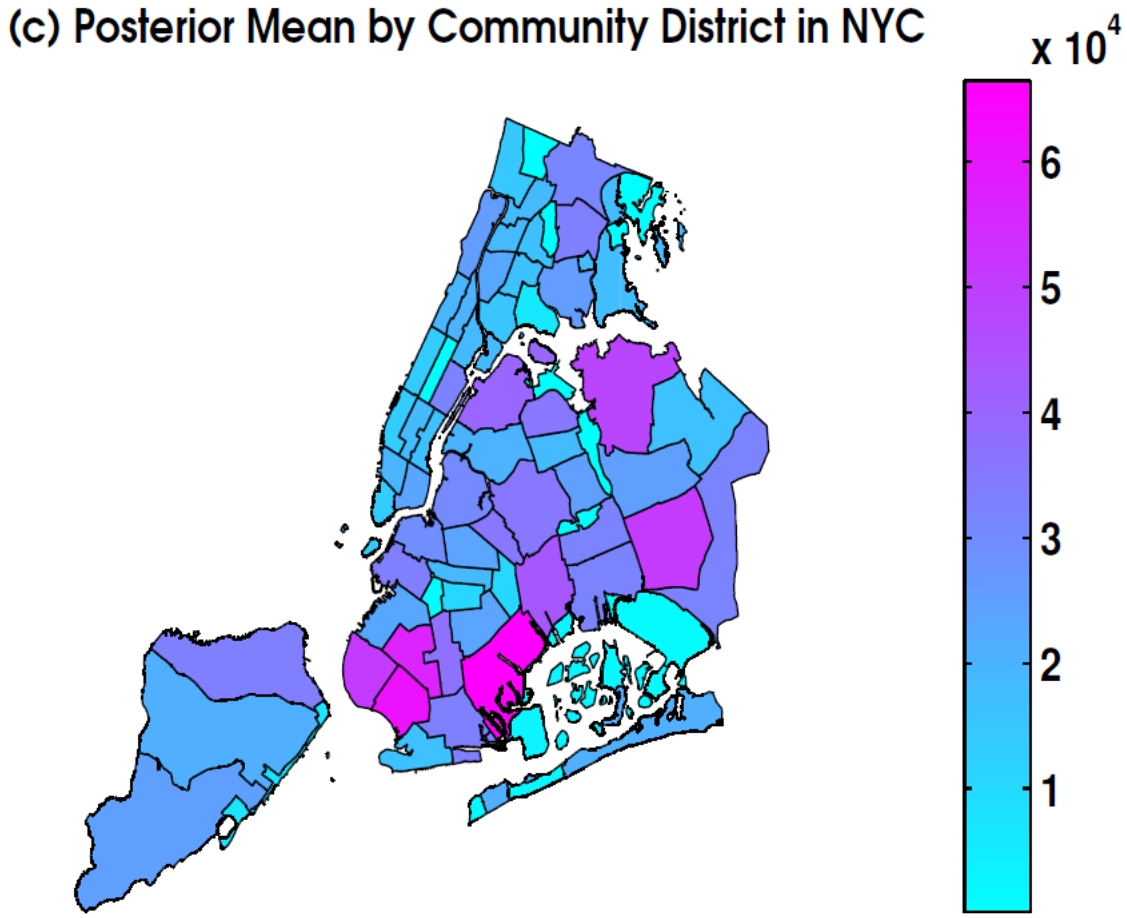
(a) Desired geography
(b) Published geography
(c) Discrepancy w.r.t. PUMAs

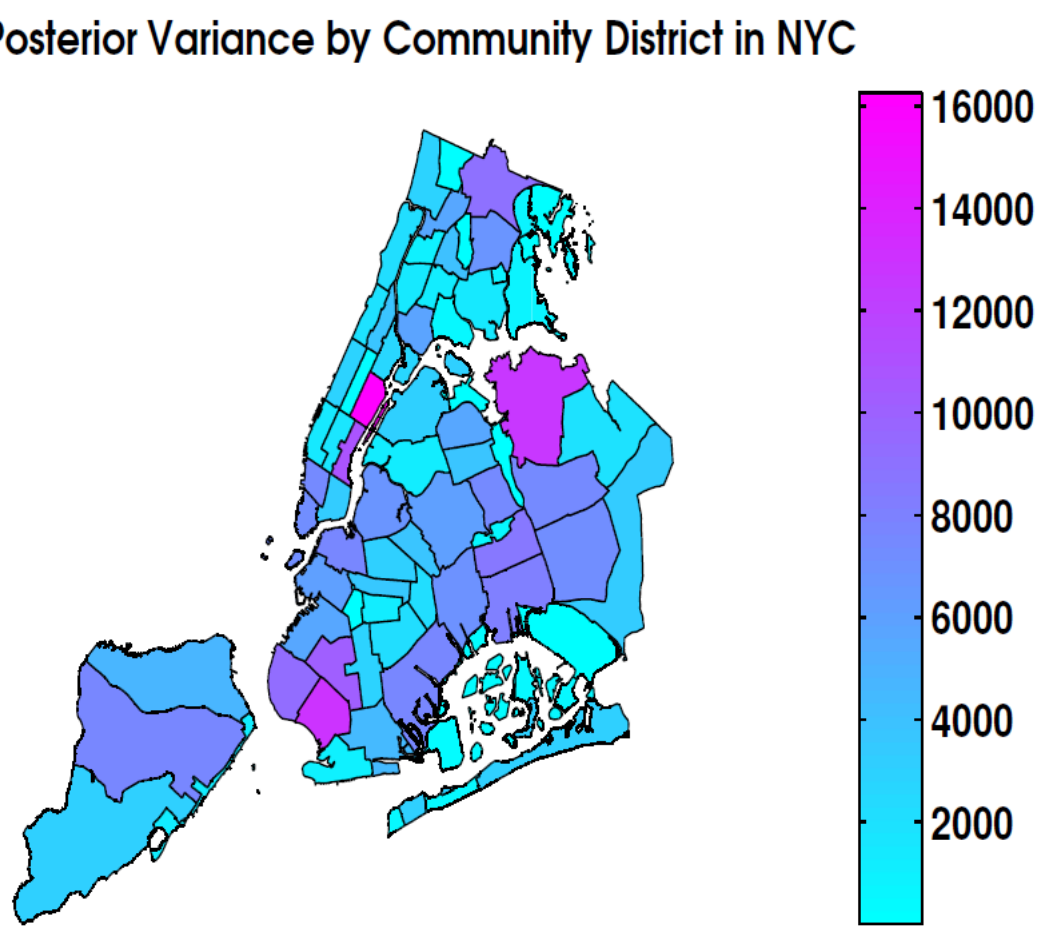
5-year Estimates of Poverty Population: NYC
(model output, not publication tables)





5-year Estimates at Target Geography
(output from the same model with redefined geography)



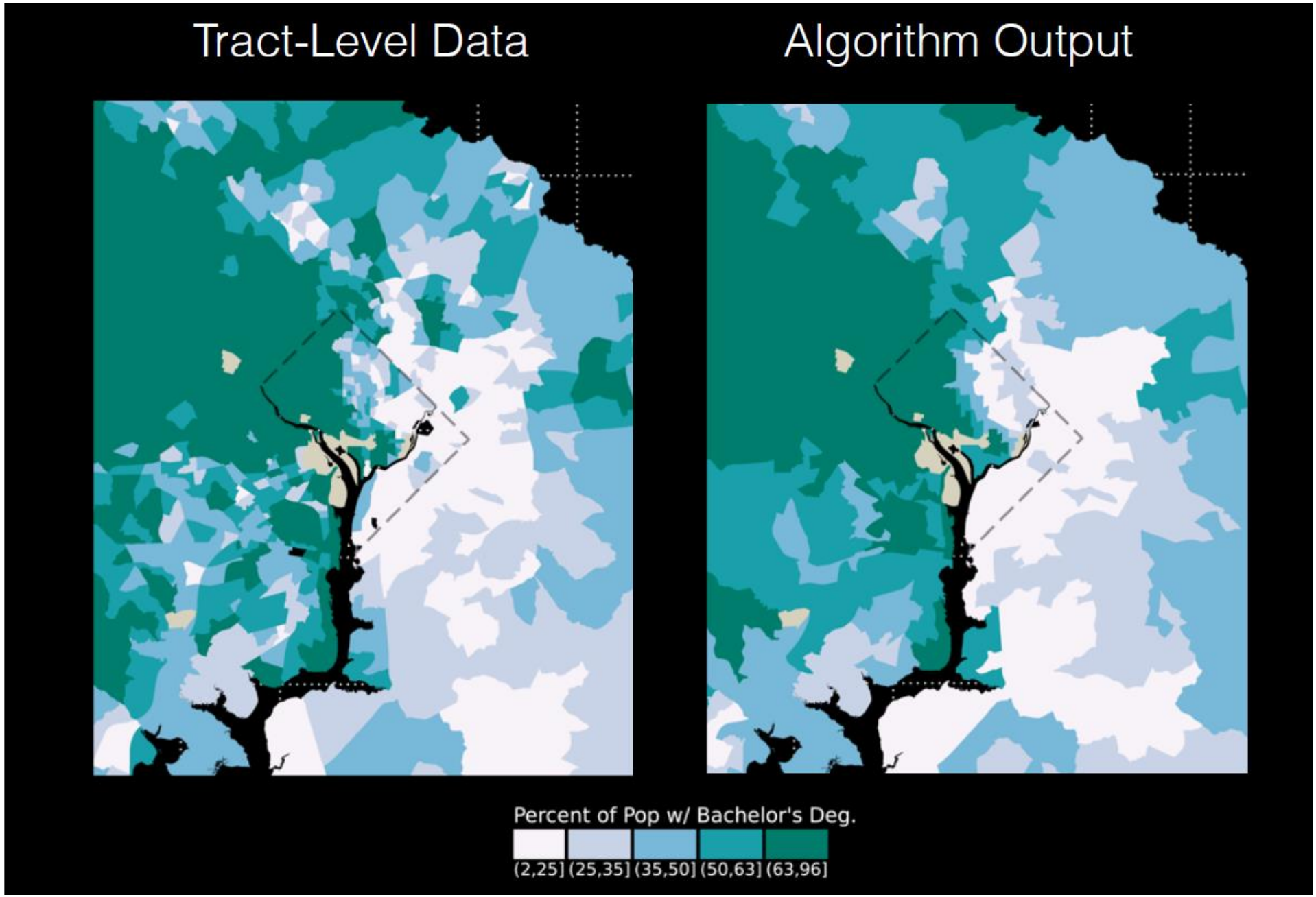


Indirect Method

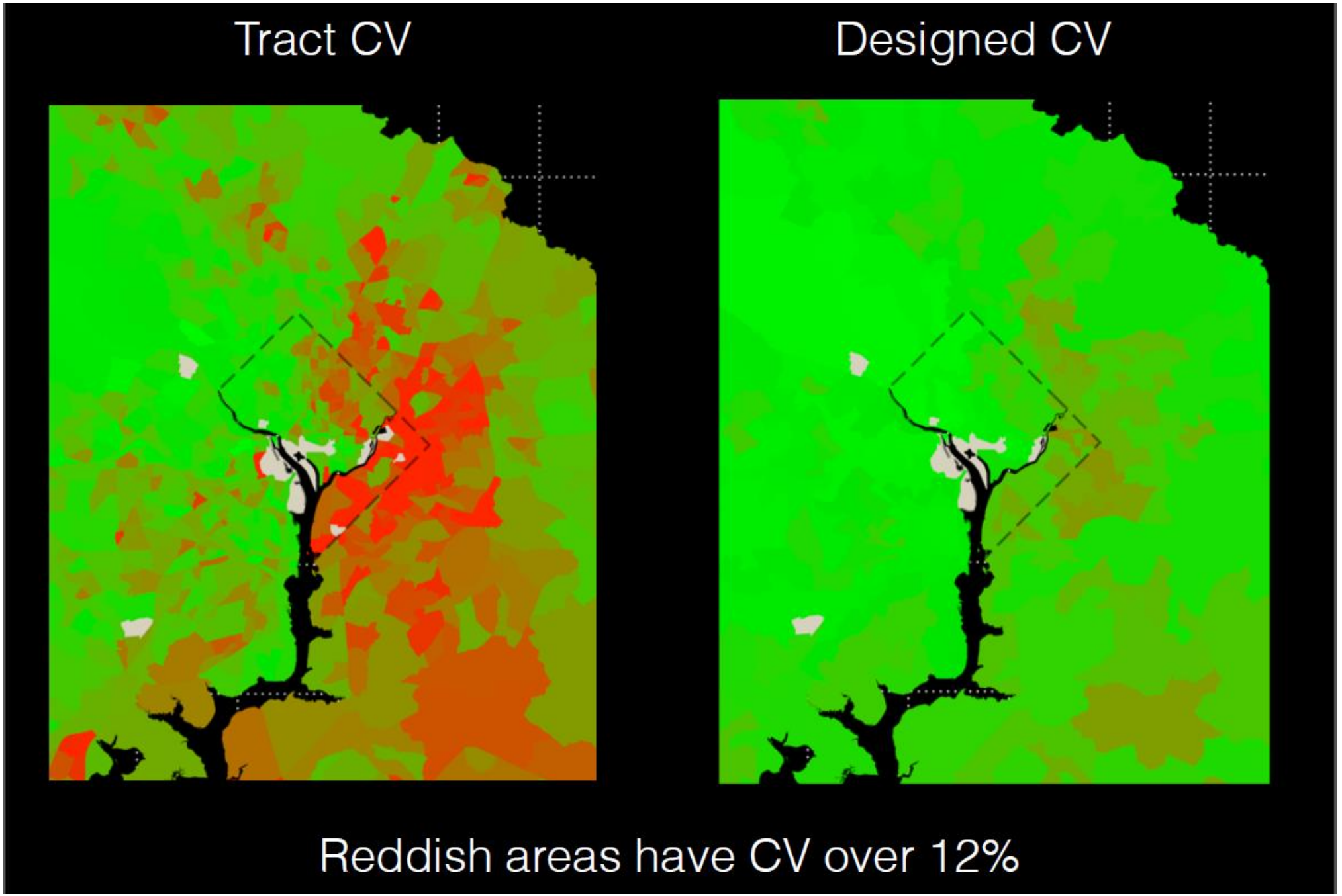
- Work of the Colorado/Tennessee node
- Goal is to produce a choropleth map of a characteristic of interest with controlled precision
- Combine the published geographies to produce a more accurate map

Left: Choropleth map of DC area as published in 5-year tract estimates

Right: Choropleth map with controlled coefficient of variation

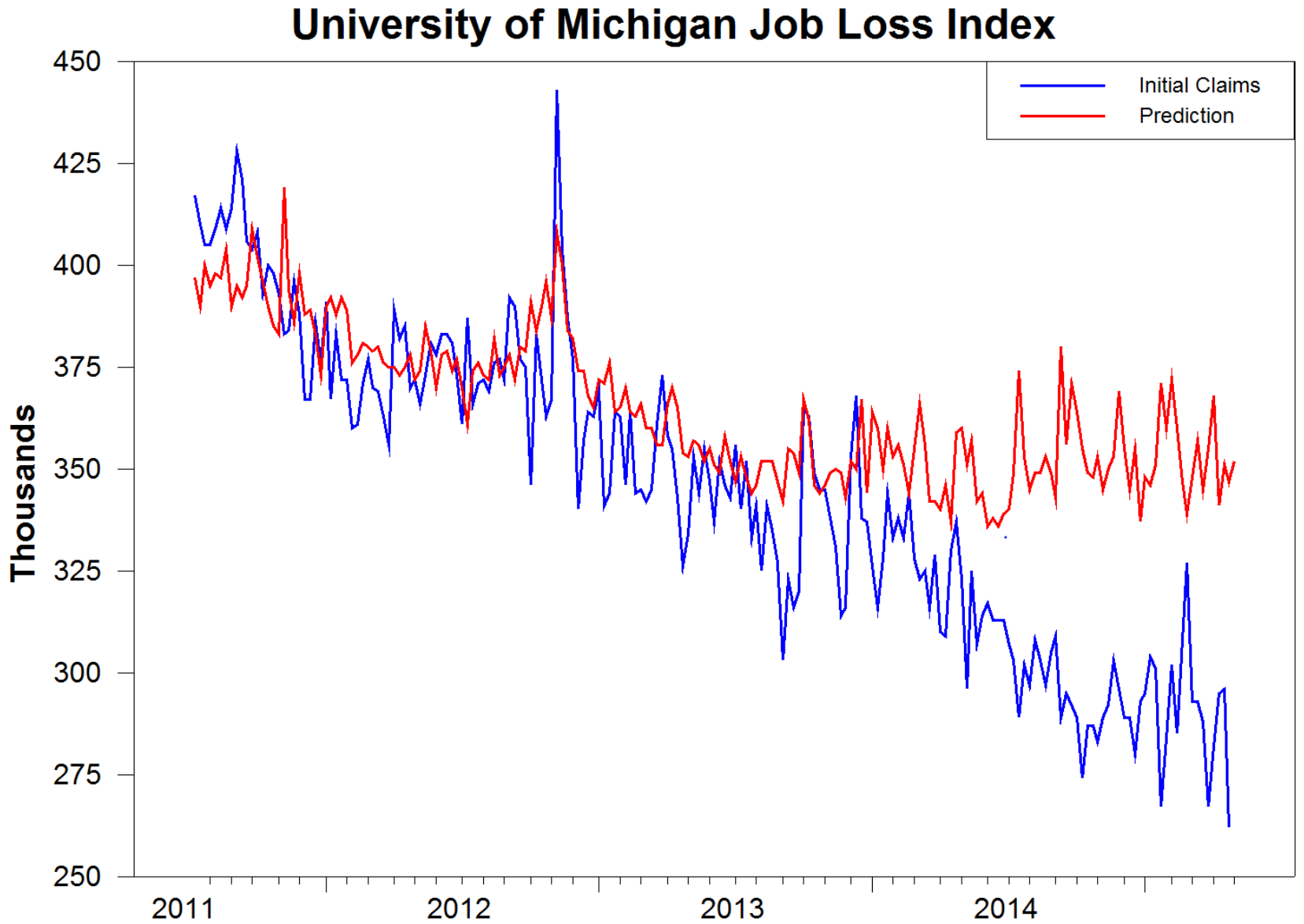


Even though the maps are very similar, the map with the controlled CV is much more accurate overall.



Example 4: Alternative Approaches to Gathering Data

- Data from organic sources can be used in combination with traditional survey and census data in much more direct ways
- By using the data as part of a formal estimation model (Missouri and Cornell nodes)
- By using the data to form new time series indicators, and study their properties (Michigan node)
- Example here is from the Michigan node's analysis of Tweets about job loss

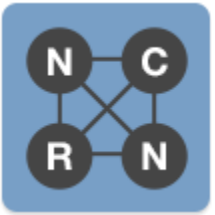

University of Michigan Job Loss Index
Initial Claims
Prediction
Thousands
450
425
400
375
350
325
300
275
250
2011
2012
2013
2014

Why Might UM Index Diverge from Initial Claims?

- Measurement issues:
  - Normalization: Shift in total Tweet activity
  - Gray swans: Occasional, non-germane spikes in job phrases cumulate over time
- Economic issues:
  - Job loss delinked from UI claiming in stronger job market
- There are far fewer actual data points than the “big data” tag implies
- For an organic data index to stand on its own, we need to understand how the relation between its inputs and other indicators evolves

Example 5: Challenges to Privacy and Confidentiality

- **NCRN teams at CMU, Cornell, and Duke** had long-standing ties to the privacy and confidentiality research community.
- Three approaches to confidentiality protection span work of the nodes:
  - Data swapping (the Census Bureau method of choice for both the decennial census and the ACS)
  - Multiple imputation (involving the preparation of replicate synthetic data sets)
  - Differential privacy (emanating from cryptography and computer science) which offers the strongest possible privacy guarantees but doesn't scale well
- Much larger social issue:
  - What are the appropriate settings for data privacy and accuracy?
  - How to pick an appropriate point on the Risk-Utility tradeoff

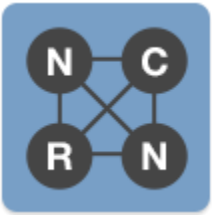


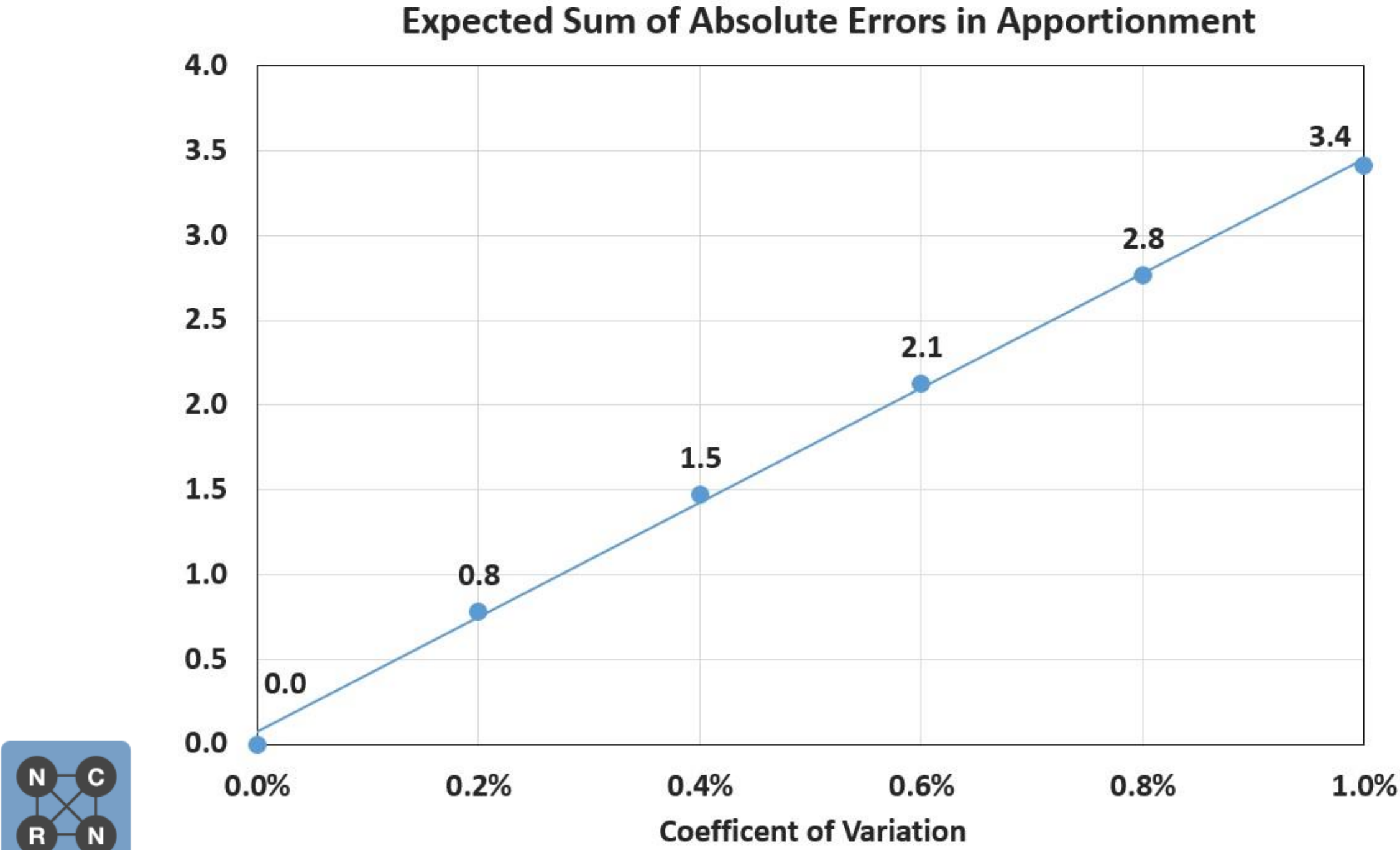
Example 6: Assessing the quality of government statistics

- How important is reducing the error variance in the decennial census?
- How can we improve the transparency of agency statistical products based on many integrated data sources?

How Important Is Controlling the Error in Census Population Counts?

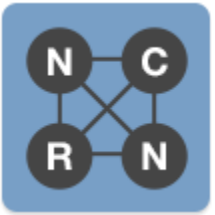
- We all understand that the methods used to enumerate a state's population for the allocation of seats in the House of Representatives have many sources of variation
- Reducing this variation is expensive
- How much accuracy does it buy?
- Northwestern node studied this problem




Expected Sum of Absolute Errors in Apportionment
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.8
1.5
2.1
2.8
3.4
0.0%
0.2%
0.4%
0.6%
0.8%
1.0%
Coefficent of Variation

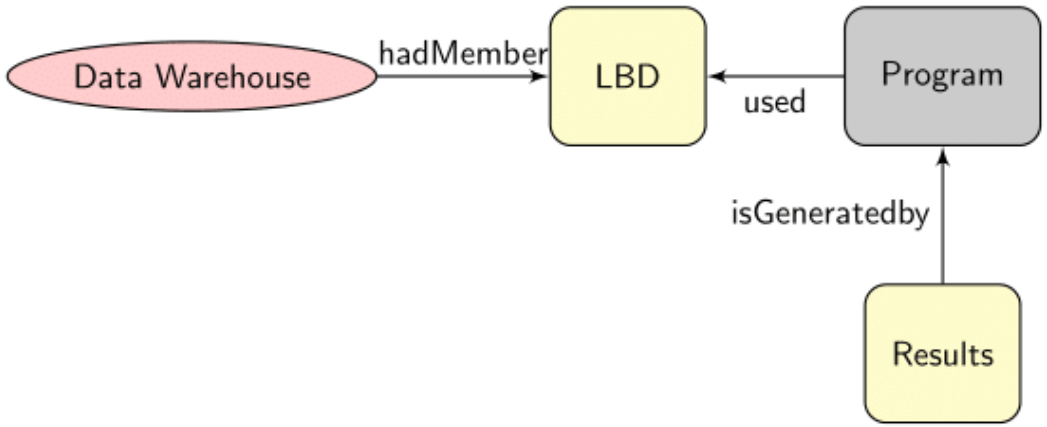
How Do You Keep Official Statistics Transparent?

- When an agency's estimates are derived from many sources with complicated interactions in the editing and computation, the process needs automated documentation
- Cornell node solution to this problem is to implement provenance metadata standards as part of the metadata database for a product
- The following maps were generated automatically from that system for a relatively straightforward product
- Dynamic maps for more complicated products are also produced automatically



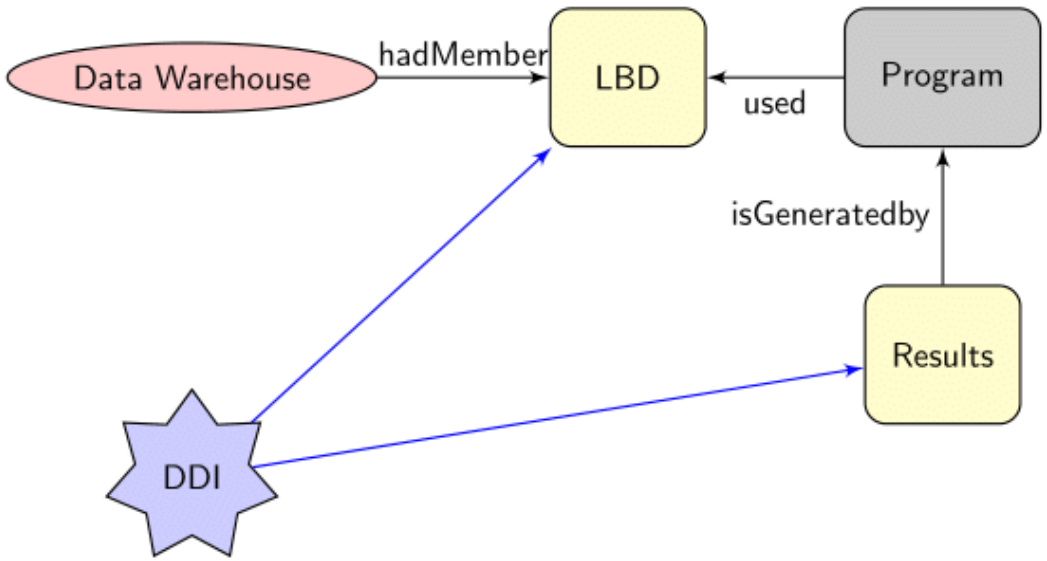
DDI+PROV for workflow

Schematic graph of the PROV enhancement of the Data Documentation Initiative protocol



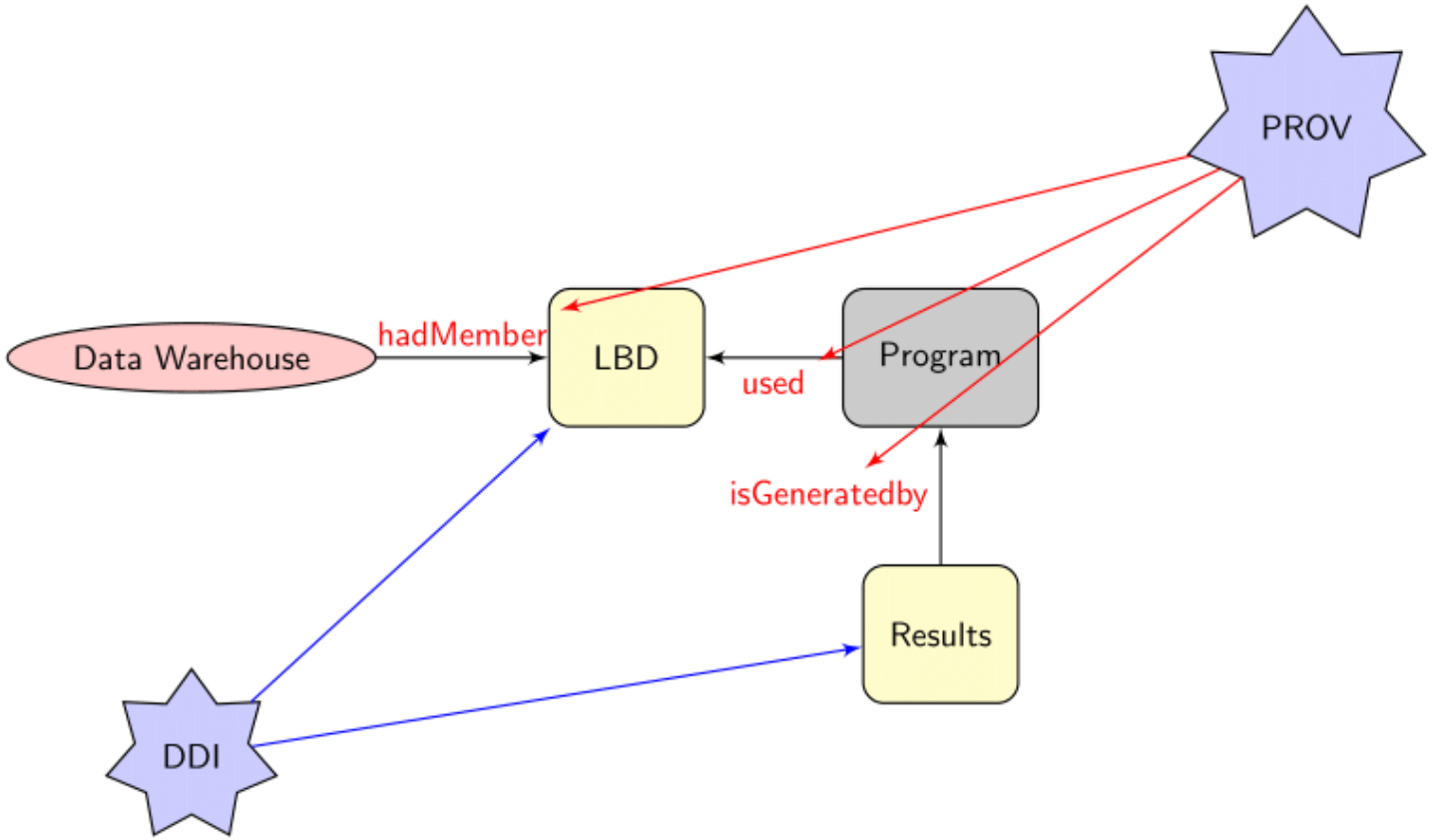
DDI+PROV for workflow

Schematic graph of the PROV enhancement of the Data Documentation Initiative protocol



DDI+PROV for workflow

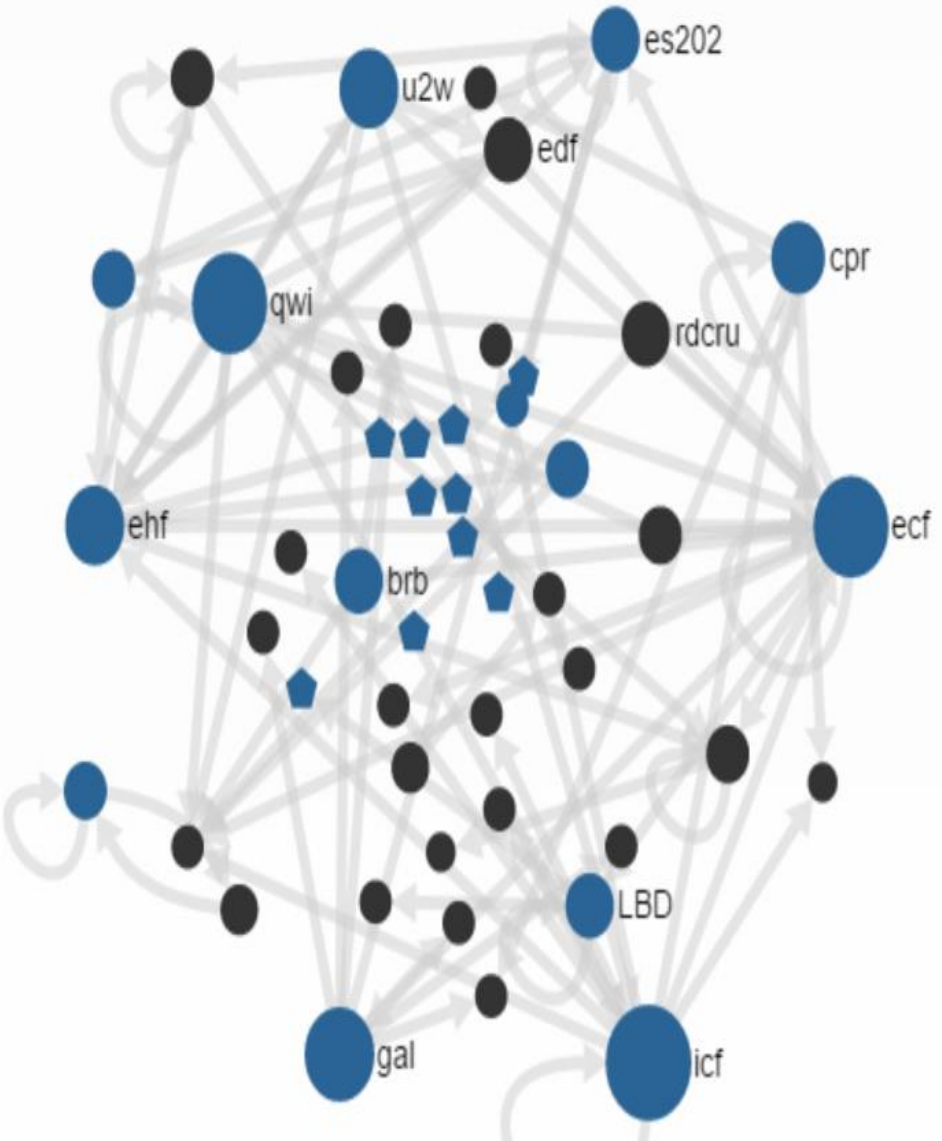
Schematic graph of the PROV enhancement of the Data Documentation Initiative protocol



Workflow Graph

Redraw Graph

--Show All Nodes--



Example 7: Educating a new generation of researchers versed in official statistics

- More 180 unique co-authors on published output from NCRN, many completely new to the world of official statistics
- New collaboration networks that have active agency participants
- Post-doc, graduate, and undergraduate training

What's next?

"We would like to suggest that if this program is continued in some fashion in the future, serious thought be given to implementing a variety of new models for facilitating the movements of researchers between academia and the Federal Statistical System."

- Recommendation of the NSF reverse site visit review committee
- Can be implemented by temporary assignment of agency professionals to node teams for visits
- Cost and disruption to agency work can be controlled through the use of new telework systems, which support full access to the work environment and video linking to colleagues

Thank you.