Synthetic Data – A New Future for Public Use Micro-Data?

John M. Abowd
December 7, 2004

Overview

What is the new NSF-Census research project?

Who are the research teams?

How do the RDC research projects and the synthetic data system work together?

What resources and opportunities does the new project bring to Census?

How can we best collaborate?

NSF-ITR Grant

Goals and Overview

The Information Technologies Research Grant from NSF

A program that encourages innovative, high-payoff IT research and education

Our grant proposal cited the many research studies and data products created by previous NSF support for the Research Data Center network and the Longitudinal Employer-Household Dynamics Program

What Is It?

$2.9 million 3-year grant to the RDC network (Cornell is the coordinating institution)

To provide core support for scientific activities at the RDCs

To develop public use, analytically valid synthetic data from many of the RDC-accessible data sets

To facilitate collaboration with RDC projects that help design and test these products

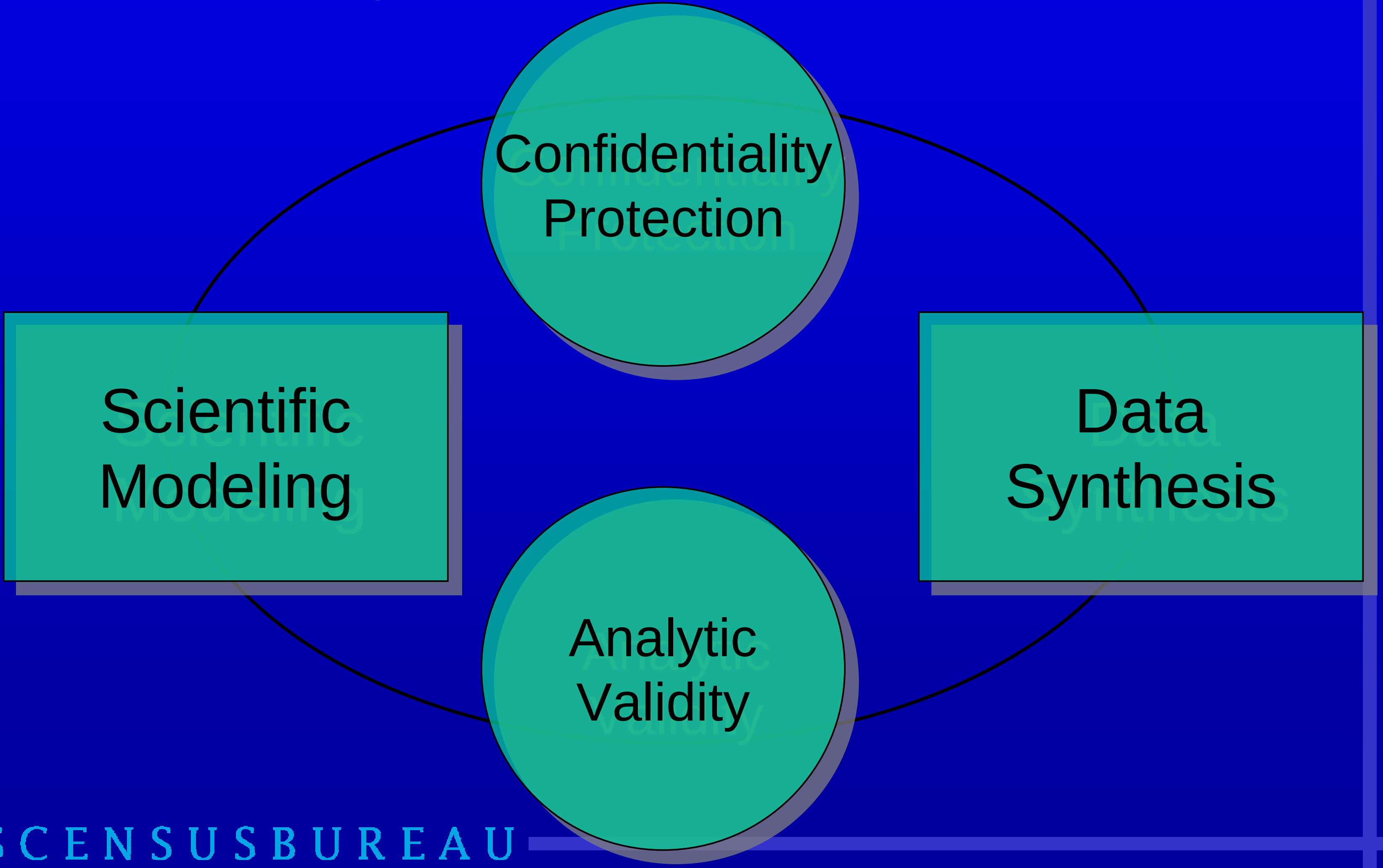
Public Use Data Products Are the Lifeblood of Statistical Agencies

- RDC-based teams understand the public use data products produced at Census and how they relate to the underlying confidential data products
- In the demographic area there are many public use micro data products
  - – But, their confidentiality protection is increasingly challenging
- In the economic area there are very few public use micro-data products
  - – But all the data are used for public use aggregate products

Public Use Data Products Are the Lifeblood of Statistical Agencies

- Integrated data products, like LEHD, produce public use summary data (QWIs) but no micro data products
  - But, synthetic data offers the possibility of releasing customized micro data products

The Research – Synthetic Data Feedback Cycle



Research Teams

Who is working with us

Principal Investigators

Ron Jarmin, Center for Economics Studies

Trivellore Raghunathan, University of Michigan

Stephen Roehrig, Carnegie Mellon University

Matthew D Shapiro, University of Michigan

I am the coordinating PI

Project Teams and Coordinators

Neil Bennett, CUNY Baruch

Gail Boyd, Argonne National Laboratories

Marjorie McElroy, Duke University

Wayne Gray, Clark University

John Haltiwanger, University of Maryland

Andrew Hildreth, UC Berkeley

Margaret Levenstein, University of Michigan

Jerome Reiter, Duke University

Jeremy Wu, LEHD Census

Ray Bair, Argonne National Laboratories

Lars Vilhuber, Cornell University

Team Locations

At Census, in the Center for Economic Studies and the Longitudinal Employer-Household Dynamics Program

At the RDCs in Washington Plaza, Boston (NBER, Cambridge), California (UCLA and Berkeley), Chicago (Consortium administered by Northwestern), Ann Arbor (University of Michigan), Research Triangle (Consortium administered by Duke), New York (Cornell and Baruch)

Research Projects and Synthetic Data

How they work together

The Multi-layer System

Basic confidential data

- Fundamental product of virtually all Census programs
- Leads to the publication of public-use products (summary data, micro data, narrative data)

Gold-standard confidential data

- Edited, documented and archived research versions of confidential data
- Used in internal Census research and at Research Data Centers

More Layers

Partially-synthetic micro data

- Preserves the record structure or sampling frame of the gold standard micro data
- Replaces the data elements with synthetic values sampled from an appropriate probability model

Fully-synthetic micro data

- Uses only the population or record linkage structure of the gold standard micro data
- Generates synthetic entities and data elements from appropriate probability models

Example: SIPP-SSA-IRS Project

Links IRS detailed earnings records and Social Security benefit data to public use SIPP data

Basic confidential data: SIPP (1990-1993, 1996); W-2 earnings data; SSA benefit data

Gold standard: completely linked, edited version of the data with variables drawn from all of the sources

Partially-synthetic data: created using the record structure of the existing SIPP panels with all data elements synthesized using Bayesian bootstrap and sequential regression multivariate imputation methods.

Example: LBD

Longitudinal Business Database: longitudinal integration of Census Business Registers developed at CES

Basic confidential data: Business Registers, including co-mingled IRS data

Gold standard: LBD and associated metadata at CES and accessible via RDCs

Fully synthetic micro data: one of the major R&D tasks of the ITR grant

More Examples

CES/NBER productivity series

Decennial Census and PUMS/ACS

LEHD and QWI-Online

Geo-spatial matching of establishment and household data

Linked versions of CPS data similar to SIPP-SSA-IRS project

The Desired Result

Implement the feedback loop of developing releasable synthetic data and testing models on the confidential and synthetic data

Promote active collaboration between Census and RDC researchers

Reinforce the role of the RDCs as an important research and development arm of the Census Bureau

Focus on clear public-use products—a direct Title 13, Chapter 5 benefit

Using data makes them better

RDC Implementation

Some details

Structure of Support

- Explicit RDC and internal projects to produce synthetic data
- Support for basic RDC expenses via subcontract to Census with an account for each RDC
- Direct support for research teams that are developing the supercluster and virtual RDC

Expectations of Researchers

- All projects at RDCs are part of the NSF-ITR and will benefit from it
  - RDC administrators are directly supported (about 1/3 of all grant funds are used for this)
- All projects will database results and programming files per metadata specs that the ITR will develop
- The individual research projects will provide a laboratory for developing synthetic data and testing the analytic validity of the new products

Distance Learning Class

Social and Economic Data

To begin Spring 2005

Wednesdays 7:00-9:30pm beginning January 26

Census and RDCs invited to enroll students

Meets via distance-learning facilities intermediated by a professional producer

Examples drawn from many Census data products

Exercises developed on the Virtual RDC

Computer Resources

The RDC Supercluster and the Virtual RDC

Two Computer Systems

The supercluster will be installed at Census facilities at Bowie

- fully-integrated part of the RDC computing system and subject to all security and operating requirements of that system
- ANL team to help design and optimize

Secondary system is a "Virtual RDC"

- Configured in much the same way as the cluster is set up in the RDC environment
- WITHOUT any confidential data (operated at Cornell)
- WITH secure access from anywhere
- Automated file release (simulating disclosure review)

Supercluster Specifications

4 x 64-way Itanium-based thick cluster

Internal memory 768GB (192 in each node)

Operating System: SuSE Linux Enterprise Server (SLES) version 9

Integrated into the RDC SAN for disk storage

Software

- SAS 9.1.3 for Itanium/Linux
- GAUSS
- Stata
- MPI and similar programming interfaces
- Fortran and C compilers from Intel

Virtual RDC Will Provide

A close approximation of the main cluster environment

A close approximation of RDC thin client

Disclosure-proofed metadata and synthetic data

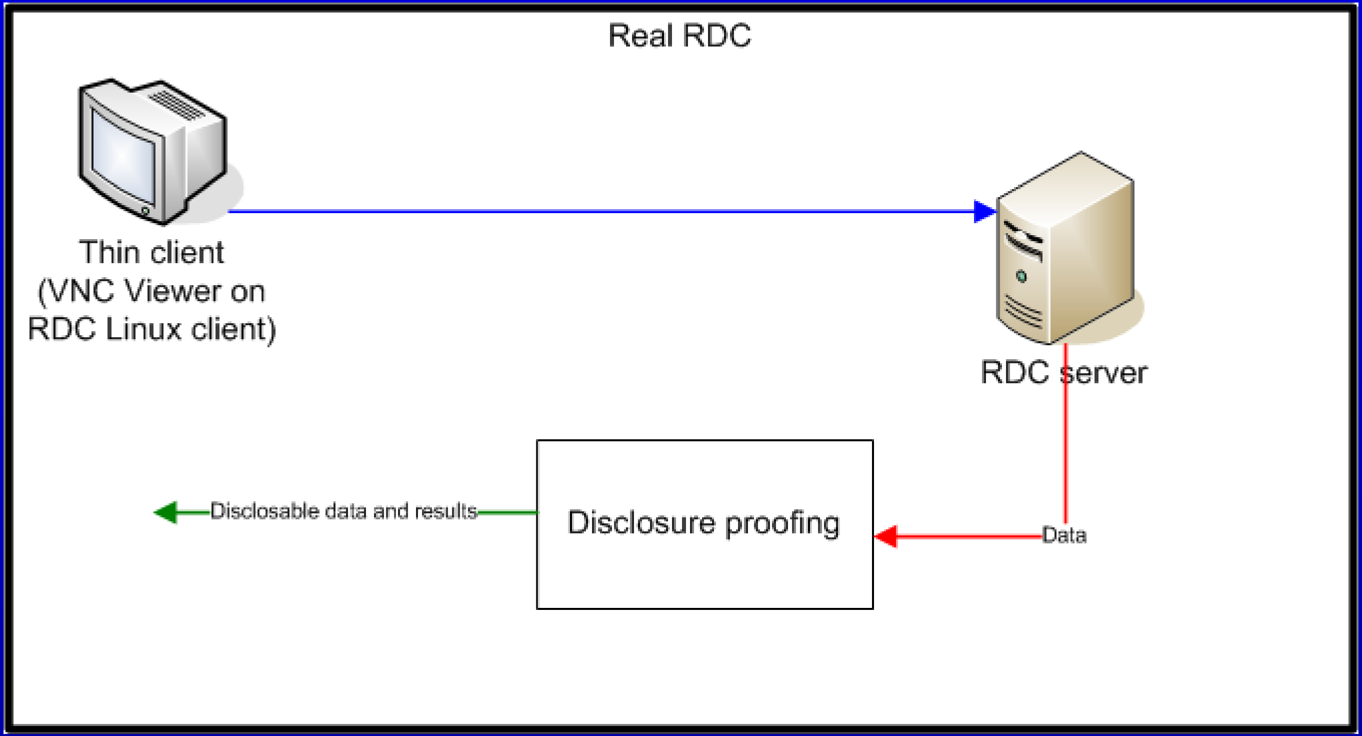
Virtual RDC Goals

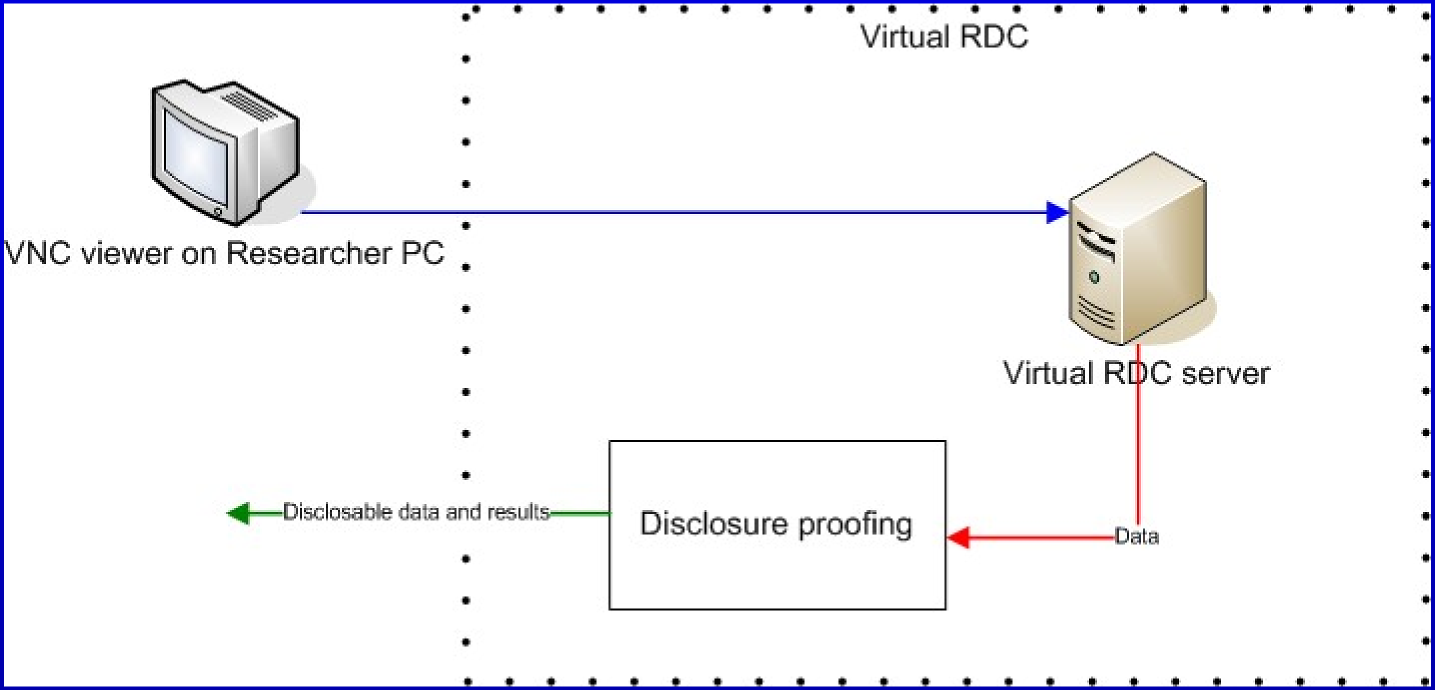
TESTING: initially test programs in a virtual RDC environment before deploying them in the RDC network

TRAINING: allow researchers to experience the “limitations” of the RDC environment (work through the thin client)

RESEARCH: once synthesized and disclosure-proofed data are available, actual research can be done on the virtual RDC server, though that may not be the only place where this occurs

FEEDBACK: run research, and provide output in a file format that is immediately transferable to the RDC network to feed back into the synthesizer


Real RDC
Thin client (VNC Viewer on RDC Linux client)
RDC server
Disclosure proofing
Disclosable data and results
Data


Virtual RDC
VNC viewer on Researcher PC
Virtual RDC server
Disclosable data and results
Disclosure proofing
Data

Collaboration

Opportunities and Suggestions

Opportunities

The grant team consists of many Census-based researchers and long-term Census collaborators.

We are openly seeking more internal Census collaborators.

The ANL supercluster team, funded through the grant, brings enormous experience in designing and running unix-based superclusters.

Feedback

Please suggest important areas where you would like to develop projects

Recruit internal and university-based research teams to collaborate via the RDC network

Other ideas?

Further Information

To request a copy of the NSF proposal, send e-mail to John.Abowd@census.gov (same as John.Abowd@cornell.edu).

To participate in the distance learning course, contact Ron Jarmin at CES e-mail to ron.s.jarmin@census.gov.

Thank you for your efforts.