

# INFO 747

## Social and Economic Data

John M. Abowd  
January 2007

# Outline

- Introductions and preliminaries
- Protocols
- A little background
- Course objectives
- Course requirements

# Introductions

- Alternate instructor and technologies coordinator (Cornell): Lars Vilhuber
- Alternate instructor (Cornell): Warren Brown
- Alternate instructor (Baruch): Sandy Korenman
- Alternate instructor (Boston): Wayne Gray
- Alternate instructor (Boston): Jim Davis
- Alternate instructor (CES): Ron Jarmin
- Alternate instructor (StatsCan): Gustave Goldman
- Alternate instructor (Montreal): Jean Poirier
- Cornell video producer: Jason Kenyon
- INFO 447 (undergraduate version) instructor: John Abowd (meets MWF 10:10-11:00 at Cornell)
- [CISER Social and Economic Data Workshops](#) instructor: Warren Brown

# Protocols

- You will always see me in the distance learning frame
- You should see the lecture presentation in the VNC session active in the classroom
- If you want to ask a question, hold up the sign “QUESTION: Site” in your camera window. When I can take a question, I’ll call on you by site.

# Tools

- Course web page:  
<http://vrdc.ciser.cornell.edu/info747/>
- Virtual RDC:
  - You will need software for this
  - For setup information see  
[http://vrdc.ciser.cornell.edu/news/?page\\_id=9](http://vrdc.ciser.cornell.edu/news/?page_id=9)

# The Information Technologies Research Grant from NSF

- A program that encourages innovative, high-payoff IT research and education
- This project directly supports the many research studies and data products created by researchers in the Research Data Center network and the Longitudinal Employer-Household Dynamics Program at the Census Bureau (all RDC administrators, including the one at headquarters received three years of partial salary support from this grant).

# 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

# Principal Investigators

- Ron Jarmin, Center for Economics Studies
- Trivellore Raghunathan, University of Michigan
- Stephen Roehrig, Carnegie Mellon University
- Matthew Shapiro, University of Michigan
- I am the coordinating PI

# Project Teams

- SIPP-SSA-IRS Public Use File
  - John Abowd (Cornell and Census), Martha Stinson (Census), Gary Benedetto (Maryland and Census), Karen Masken (IRS), Lisa Dragoset (Cornell and Census), Bryan Ricchetti (Cornell and Census), Simon Woodcock (Simon Fraser), Sam Hawala (Census), Lars Vilhuber (Cornell and Census)
- Longitudinal Business Database
  - Jerry Reiter (Duke), Saki Kinney (Duke), Ron Jarmin (Census), Javier Miranda (Census), Arnie Resnick (Census)
- Longitudinal Employer-Household Dynamics Infrastructure Micro-data
  - Jeremy Wu (Census), Simon Woodcock (Simon Fraser), Gary Benedetto (Census), John Abowd (Cornell and Census)
- American Community Survey
  - Sam Hawala (Census), Rolando Rodriguez (Census), Jerry Reiter (Duke)
- Quarterly Workforce Indicators
  - Data Synthesis Team: John Abowd (Cornell and Census), Kaj Gittings (Cornell and Census), Fredrik Andersson (Cornell and Census), Matt Freedman (Maryland and Census), Lars Vilhuber (Cornell and Census)
- On The Map
  - Data Synthesis Team: Fredrik Andersson (Cornell and Census), Matt Freedman (Maryland and Census), John Abowd (Cornell and Census), Kaj Gittings (Cornell and Census), Jeremy Wu (Census)
  - Mapping software (not NSF-supported): John Carpenter (ExCensus)
- Methodology
  - Jerry Reiter (Duke), Trivellore Raghunathan (Michigan), John Abowd (Cornell), Lars Vilhuber (Cornell)
- Software
  - Stephen Roehrig (CMU), Lars Vilhuber (Cornell), Bryan Ricchetti (Census), Josep Domingo-Ferrer (Barcelona), Vicenc Torra (Barcelona)

# This Course

- Was sponsored by this NSF grant to develop a new generation of researchers who can use and produce these tools.
- Now taught by a team of instructors from around the RDC network.
- Now launched in coordination with the Montreal RDC (part of the StatsCan RDC network)

# Course Objectives

- Teach researchers to use the confidential micro-data that underlie public use products
- Cover the legal, ethical, statistical and computational issues
- Develop the tools for editing, linking, analyzing those data
- Integrate research on the confidential data with improvement of the public use data

# Course Requirements

- Lectures
- Labs
- Familiarity with Census RDCs
- Familiarity with Statistics Canada RDCs
- Virtual RDC use
- For a grade: negotiate the grading requirements with the instructor at your location