


 Cornell University

Data access in North America

Current state and future consequences


William C. Block and Lars Vilhuber

 Cornell University

Disclaimer:


The opinions expressed in this presentation are those of the authors and not the National Science Foundation, the U.S. Census Bureau, or any other government agency.

No confidential, restricted-access data was used to prepare this presentation.

 Cornell University

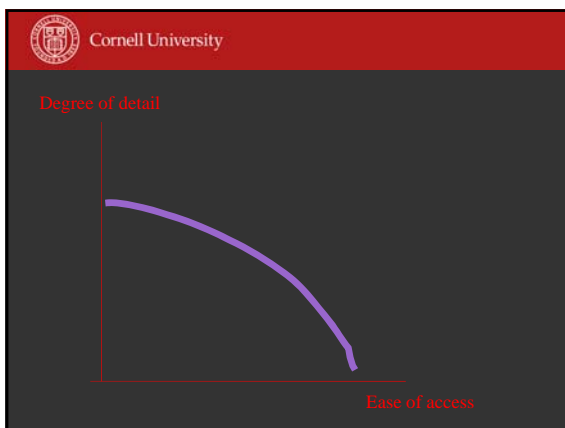
Caveats


- Economist
- Labor Economist
- Micro-data preferred
- US bias

 Cornell University

Classifying North American data

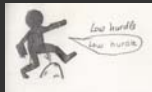
- Access-type
 - Public-use data
 - Contractual access
 - Restricted-access data
- Data source
 - Survey data
 - Administrative data
- Strength of SDL




 Cornell University

Access: Public-use (micro-) data


- Very successful from a usage perspective
- Examples:
 - Current Population Survey (CPS)
 - PSID
 - NLSY
 - Kauffman Firm Survey
- Used in teaching, research, deep scientific corpus




 Cornell University

RA: Contractual restriction


- Examples:
 - NLSY (detailed geo)
 - HRS (additional data)
- Some restrictions on usage in exchange for details
- Few constraints in combining with other data




 Cornell University

RA: Remote controlled access from anywhere


- Examples:
 - CRADC @ Cornell
 - Data enclave @ NORC
 - Synthetic data server @ Cornell
- Typically still cross-dataset access restrictions even within the same environment
- Reduced ability to combine with other data



 Cornell University


RA: Remote execution

- from anywhere
- Examples:
 - NCHS micro data (\$)
 - Statistics Canada
 - (implicit in Synthetic Data Server)
- May be limited in complexity of models that can be estimated

 Cornell University

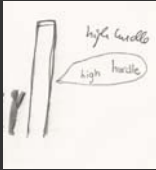
Remote access from controlled location




 Cornell University

Remote access from controlled location

- Examples:
 - Census, BLS, Canadian RDC
 - Even IAB data (from Cornell)
- Limited access (few locations)
- Long application process
- Limited ability to add additional data



 Cornell University


Detail and access

- As detail increases, access restrictions also increase
- What other methods are used?

 Cornell University


**Trade-off:
geographic detail vs. timeliness**

- Decennial Census
 - Tract level
 - Limited characteristics
- American Community Survey
 - More person/household characteristics
 - Precision increases with multi-year estimates

 Cornell University


**Trade-off:
geographic detail vs. timeliness**

- Current Population Survey
 - Monthly estimates
 - No sub-state estimates (exception: 12 large MSAs)

 Cornell University


Data without Boundaries

- Increased access to restricted access data
- Access to data from multiple jurisdictions
- Access to data from multiple “access domains”
- Increasingly detailed public-use data

 Cornell University

Increased access to restricted access data

- Expansion of RDC network
 - USA
 - Canada
- Expansion of data accessible in RDC network
 - Agency for Health Care Research (AHRQ)
 - National Center for Healthcare Statistics (NCHS)

 Cornell University

Access to data from multiple jurisdictions

- Long-standing access
 - IRS, SSA data in Census RDC, can be combined with Census data sources
- New
 - Multi-state access (education-oriented longitudinal data warehouses)

 Cornell University

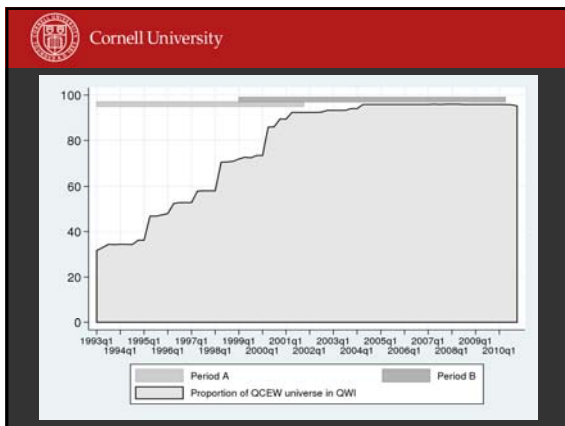



FloridaLightning.com

 Cornell University

Not everything is advancement


- BLS, Census, other agencies remain distinct and separate (despite CIPSEA)
- No cross-border access (Canadian data in US or vice-versa)
- Multi-jurisdiction access may be reduced, not increased (state employment agencies at Census Bureau) for research purposes



 Cornell University


Access to data from multiple “access domains”

- How to get MUCH public-use data into
 - Census RDC
 - CRADC?
- No data curation other than own data
 - > CCBMR (see our presentation at WDA)
- Synthetic data, more detailed geo data
 - Increased ease of combining data

 Cornell University


Other methods

- Increasingly detailed public-use statistics
 - Use of
 - synthetic data
 - new methods of SDL
 - Quarterly Workforce Indicators
 - Business Dynamics Statistics
 - Synthetic SIPP
 - Synthetic LBD

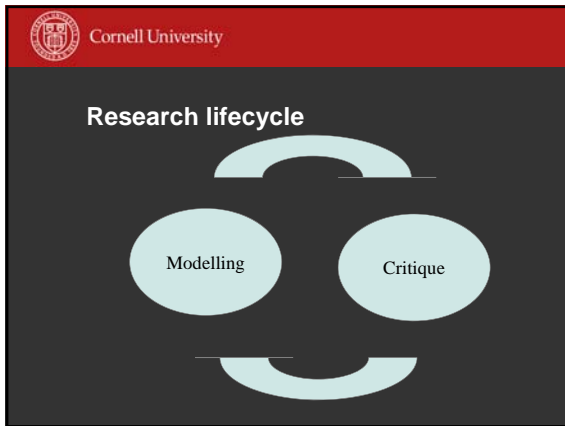
 Cornell University

Example: Abowd and Vilhuber (2012)

- “Did the Housing Price Bubble Clobber Local Labor Market Job and Worker Flows When It Burst?” (AEA, PP, 2012)
- Data sources:
 - FHFA's Housing Price Index
 - BLS' National and Local Unemployment Statistics
 - Census Bureau's Quarterly Workforce Indicators
 - Our own national aggregation of those

 Cornell University

Why do we do this?



- Why?**
- Accelerate the research cycle
 - Increase the body of research for any given data source
 - Improve economic/social/demographic/etc. models through more detailed data

Public-use data very successful

IDEAS Search

Search for: Search

Results: 1,000 of 2,000


Sort by: Relevance

1. Estimating the Return to Education Using the Revised Current Population Survey Education Questionnaire (2009) [Working Paper] by David A. Colander
2. Discussion and Presentation of the Disability Test Results from the Current Population Survey (2009)
3. Do housing wealth and equity savings in the current population survey (2009) [Working Paper] by Long (2009)
4. Family Disagreement on Wealth from the Current Population Survey: A National Survey of Parents' Concerns (2009) [Working Paper]


 Cornell University

Goal of research

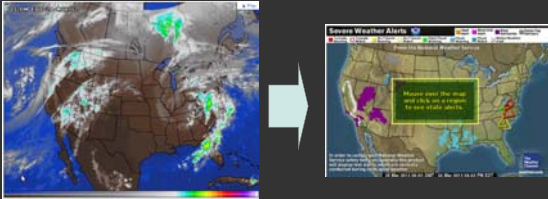
- Understanding of economic and social phenomena
 - Better model-based predictions
 - Better experimental analysis


 Cornell University

Modelling

 Cornell University


Weather modelling



 Cornell University


Behind this:


- A set of models
- Computed using observed data, simulations
- National Centers for Environmental Prediction has two 156-node compute clusters running 24/7
- Precision of predictions?

 Cornell University

Experiments


- Experiments provide useful data under controlled circumstances
- They are sometimes frowned upon...

 Cornell University



Cornell University

Nuclear experiments nowadays



Cornell University


ASC computing environment

- Sequoia next-generation BlueGene/P compute cluster:
 - 98,304 compute nodes
 - 1.6 million processor cores
 - 1.6 PB memory

Cornell University


Bad policy and “experiments” have bad outcomes



 Cornell University


The logical next step?

- If we can simulate...
 - atomic bombs
 - Weather
- Given the right input data (integrated DwB!)
- Can we provide (better) simulations of economic phenomena and policy?

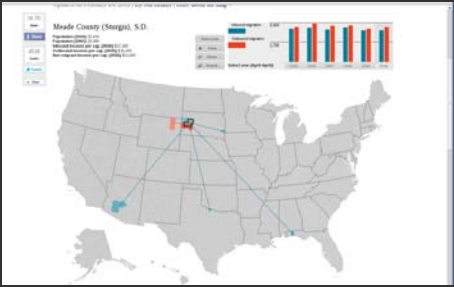
 Cornell University

Let's consider ...

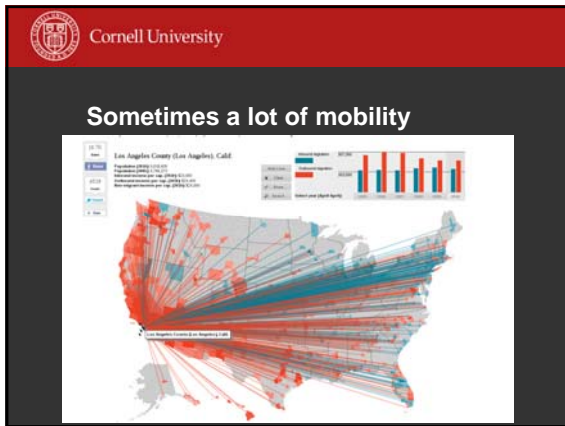
labor market mobility

 Cornell University

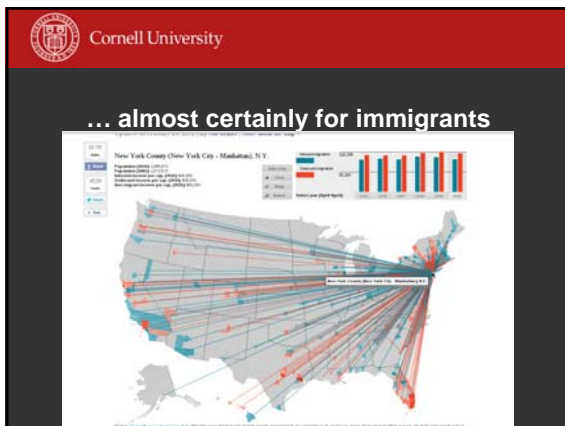
Sometimes only very little mobility



The figure shows a map of the United States with migration flows between states. A legend indicates 'Migration' with a color scale from 0 to 100. A bar chart in the top right corner shows migration flows for 'All states' and 'All states (except South)'. The map highlights a specific migration path from the Midwest to the South.








Cornell University

Presenting

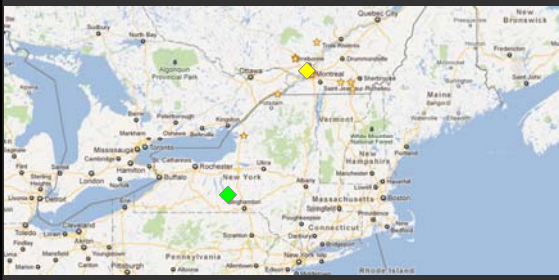
- The bane of integrated data

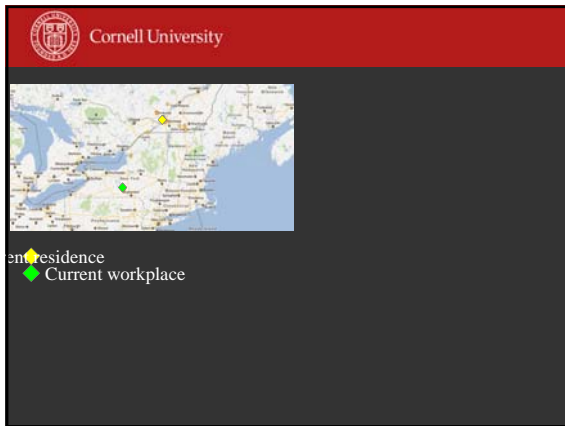
Mr. Data-truncation

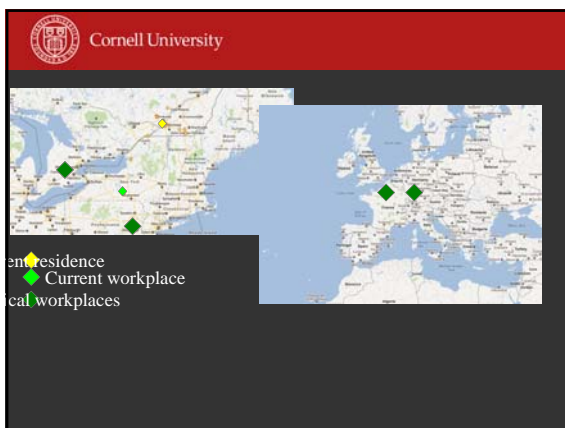
Cornell University

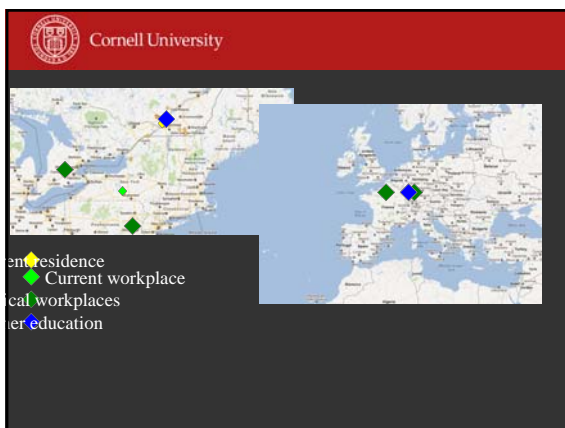


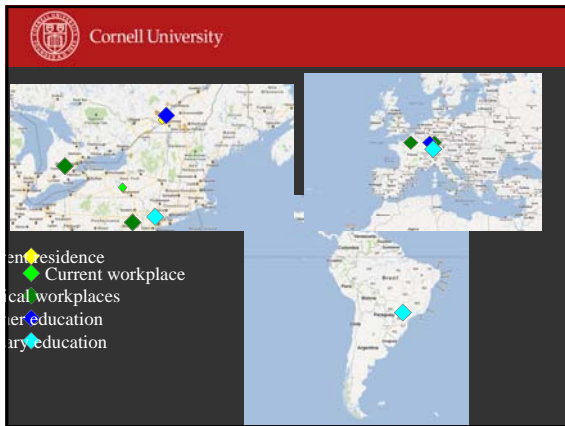
Cornell University





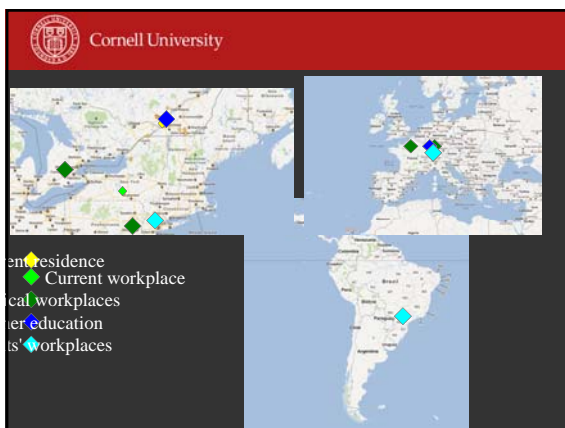






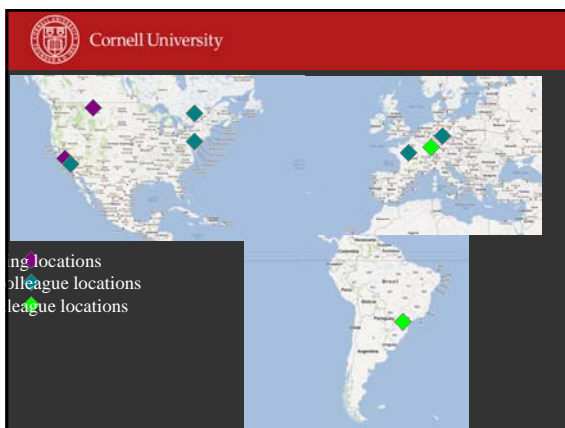
Cornell University

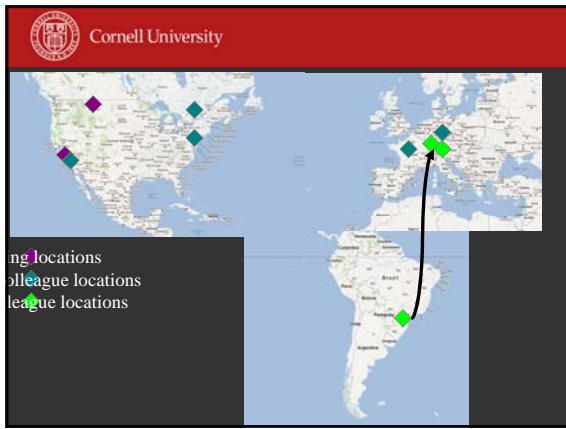
Not just me.

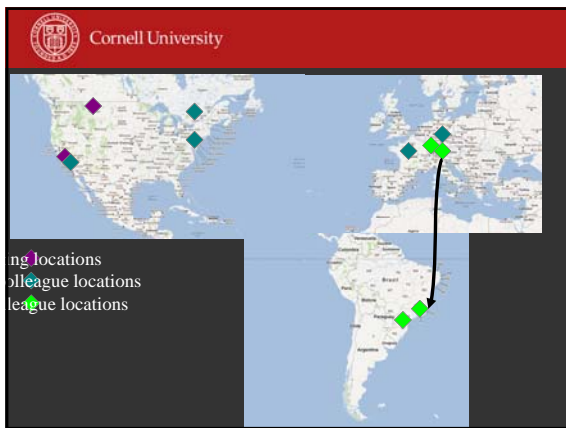


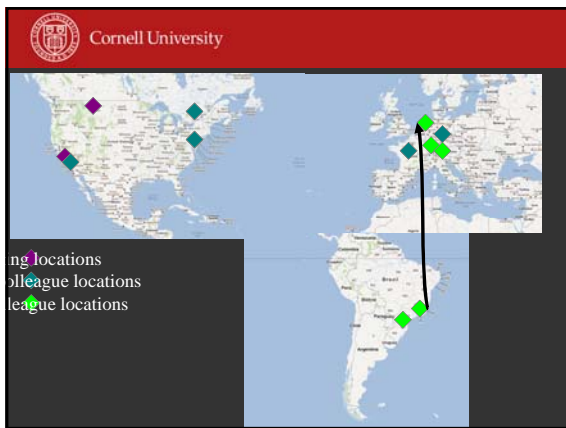


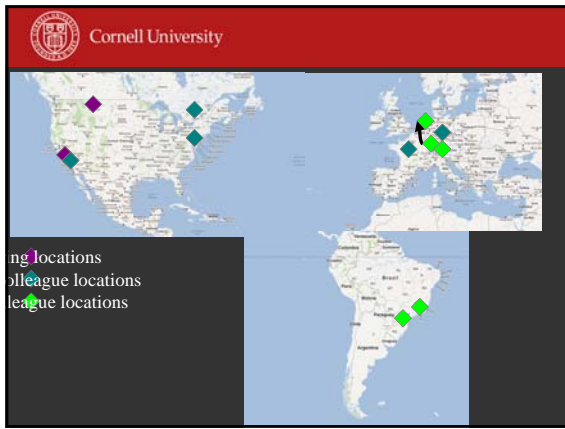


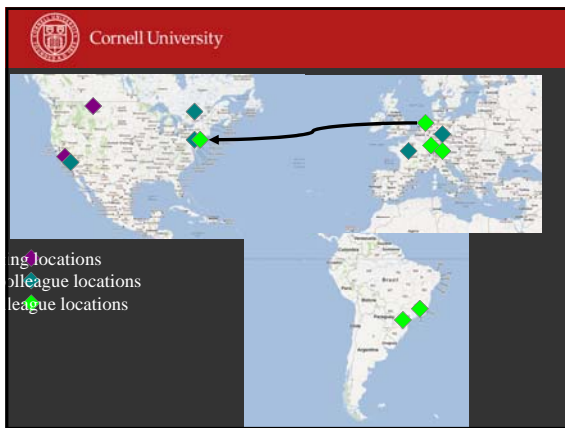


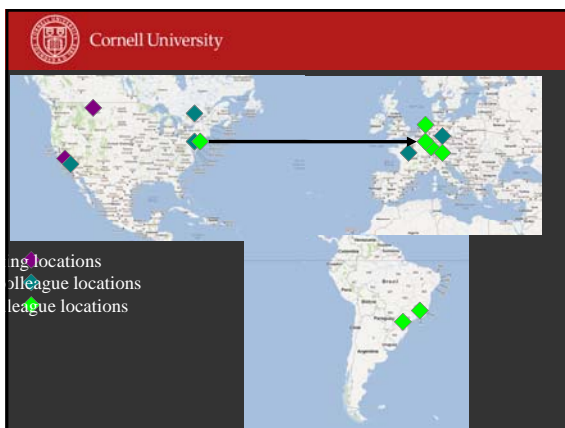













 Cornell University

It gets worse...


- Siblings in Montana (works in Silicon Valley) and Grenoble (used to live in Egypt)
- Parents somewhere in Europe (long live retirement), with retirement income from two state retirement systems (US and Germany)

 Cornell University


Historical data offers some insights



- We can link Tor Janson from Oslo (1880) to his records in the United States
- But we cannot link 21st century Lars Vilhuber

 Cornell University


Hourly data available...



 Cornell University


And I didn't even mention...

- F...b..k
- G....l.
- Tw.....

 Cornell University

This is not the end

- Suppose we solve most of the data access issues
- What kind of data usage models will we see?

 Cornell University


Example mobility

- Kennan and Walker (2003,2011)
- Model determinants of individual location and employment choices along a mobility path
- Computational limitations:
 - 500 HS dropouts
 - State-level choices
 - Only two at any time
 - > 1 day @ 50CPUs to estimate


 Cornell University

Some attempts get close

- “Exploring New Methods for Protecting and Distributing Confidential Research Data” at Michigan (Felicia LeClere) is already working in the cloud
- Census Bureau working with network of researchers, working group on next-generation flexible compute architecture within restricted-access environment

 Cornell University

Outlook

 Cornell University

Consequences of successful DwB

- If you create it (the integrated data environment), they will come
- ... but they may wish for more than you can provide
- Successful data integration must also provide the tools for new (pent-up) modelling strategies



Cornell University

The next frontier



- Tera-scale compute resources for the social sciences, using integrated confidential data



Cornell University
