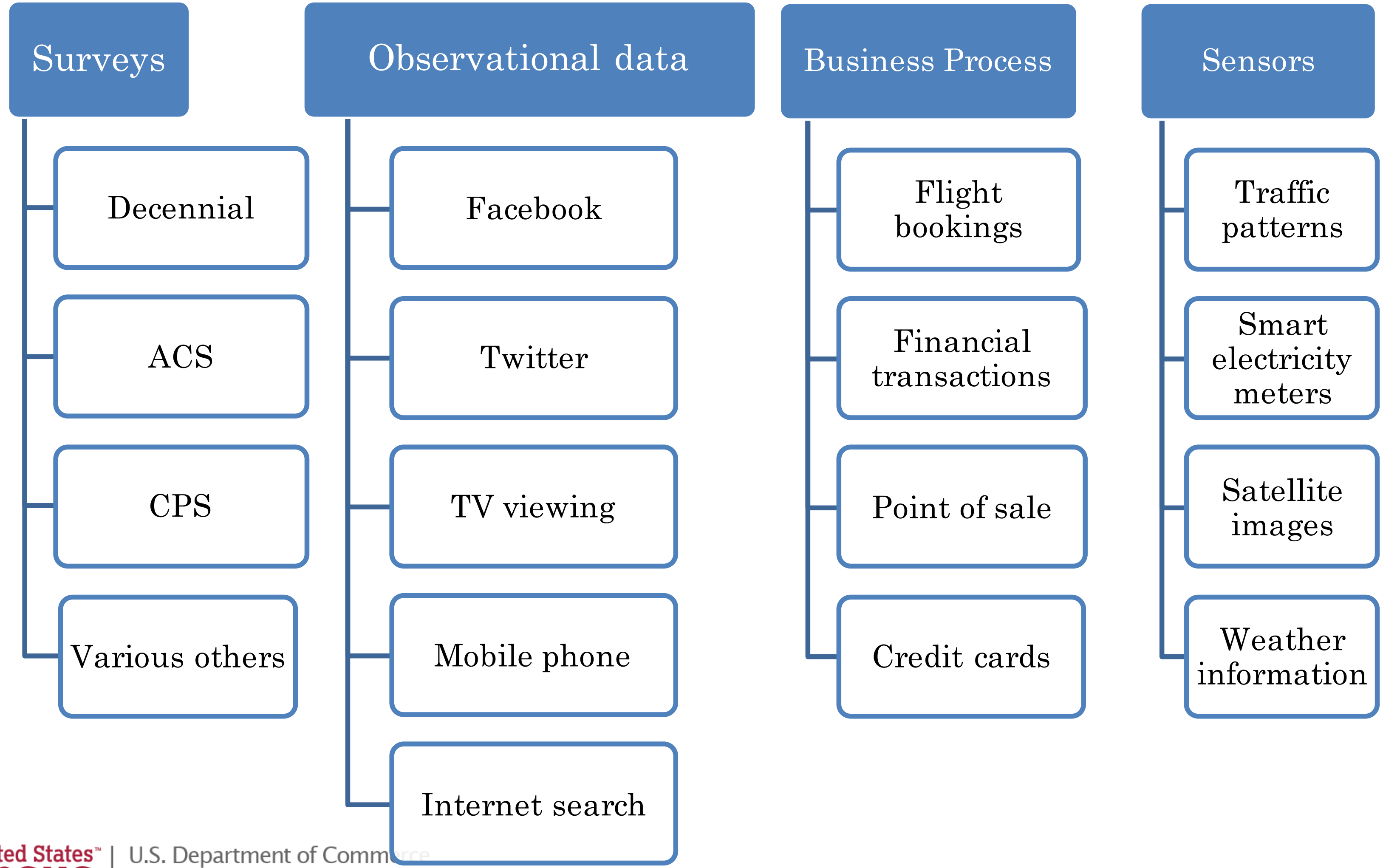
Center for Big Data Research and Applications (CBDRA)

Sudip Bhattacharjee
Chief, Center for Big Data Research and Applications
US Census Bureau

Professor
School of Business
University of Connecticut

Sudip.Bhattacharjee@census.gov

Measure and analyze myriad data types



- Surveys
  - Decennial
  - ACS
  - CPS
  - Various others
- Observational data
  - Facebook
  - Twitter
  - TV viewing
  - Mobile phone
  - Internet search
- Business Process
  - Flight bookings
  - Financial transactions
  - Point of sale
  - Credit cards
- Sensors
  - Traffic patterns
  - Smart electricity meters
  - Satellite images
  - Weather information

Center for Big Data Research

- Use machine learning and Big Data tools and techniques to make current Census products "better, cheaper, faster"
- Research and produce new "products"
- Combine survey data, administrative records, transactions … to improve current products, and produce new ones
  - E.g. UMETRICS – current RDC release

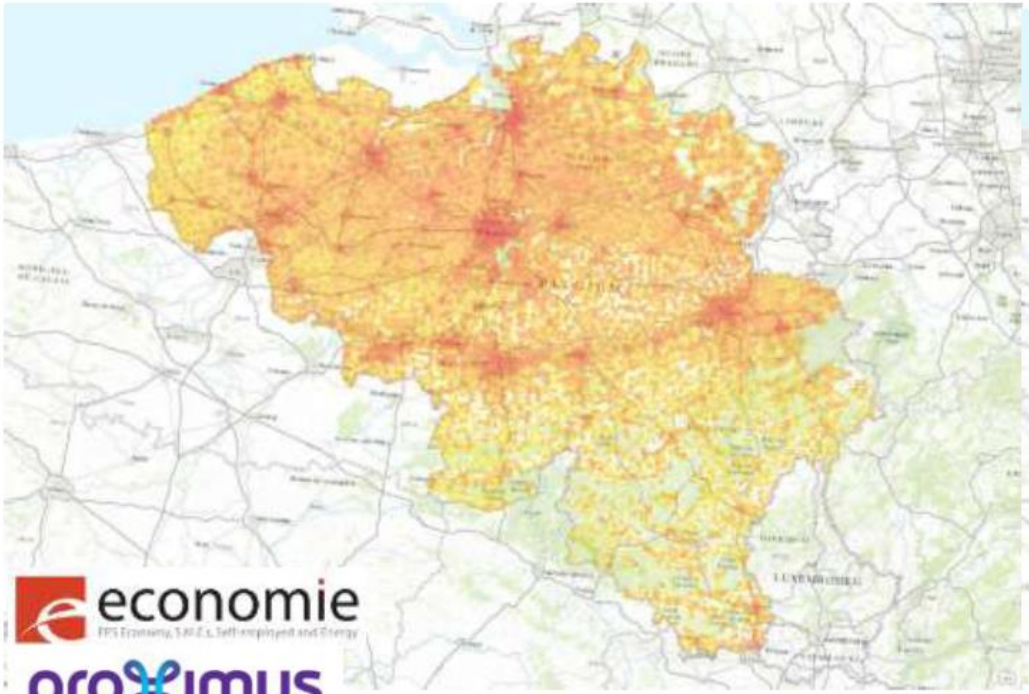
The Vision

Using mobile phone data for national statistics

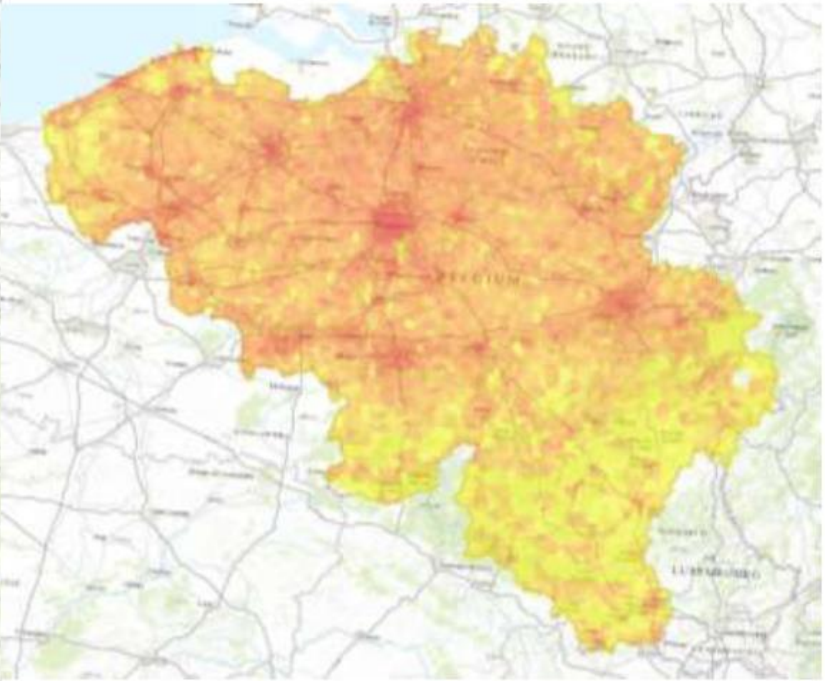
- Experiments in European statistics
  - **Assessing the Quality of Mobile Phone Data as a Source of Statistics (**Proximus Belgium, Statistics Belgium, Eurostat)
  - Population distribution from mobile phone data
  - Complement traditional statistics, capture real time phenomena

Mobile phone data - population

Census 2011



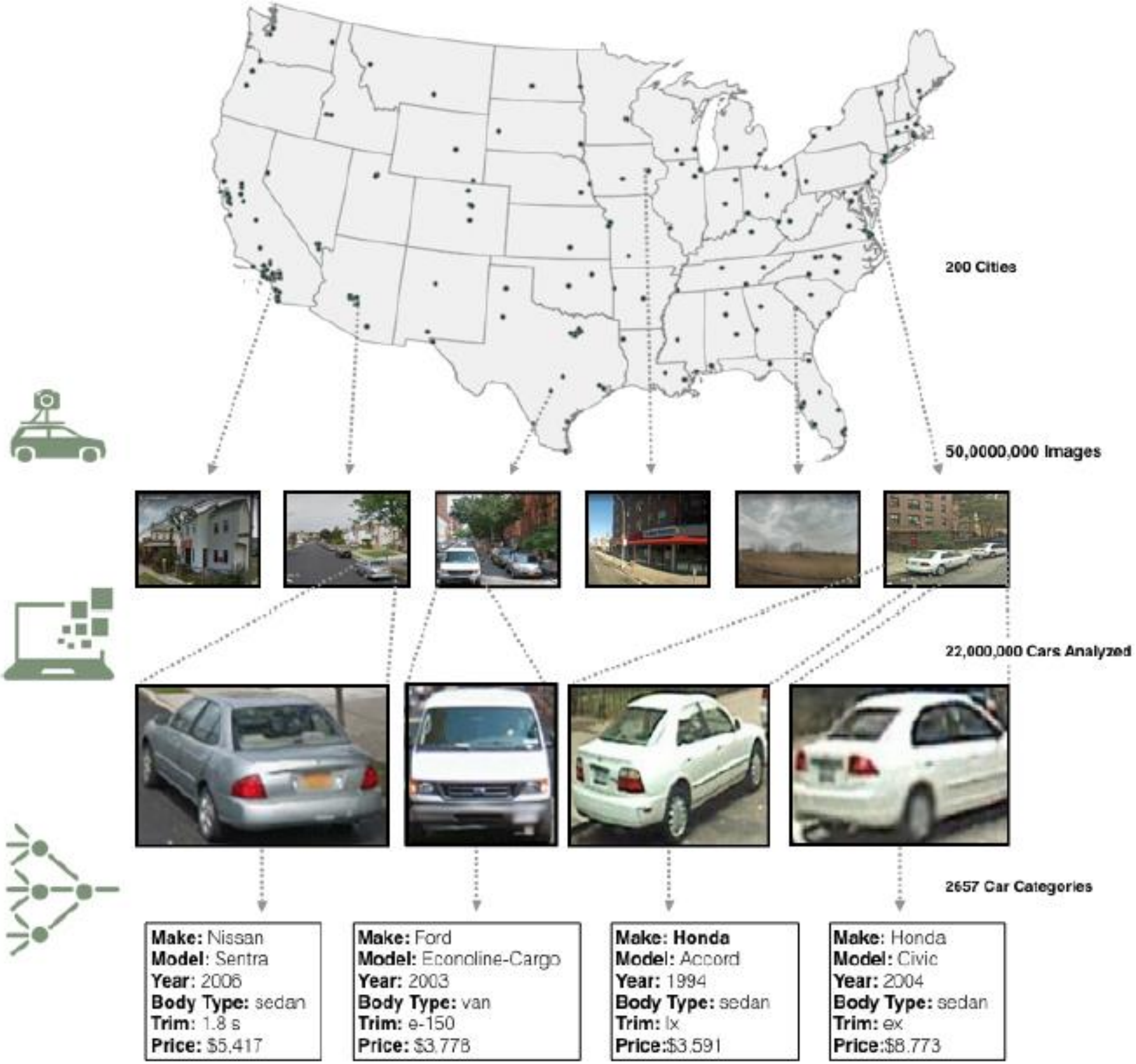
Mobile phones 2015



Eurostat

Using Google images to estimate ACS demographics

- Machine learning of vision
- Using Deep Learning and Google Street View to Estimate the Demographic Makeup of the US
  - Gebru, et al
  - arXiv:1702.06683



200 Cities

50,0000,000 Images

22,000,000 Cars Analyzed

2657 Car Categories

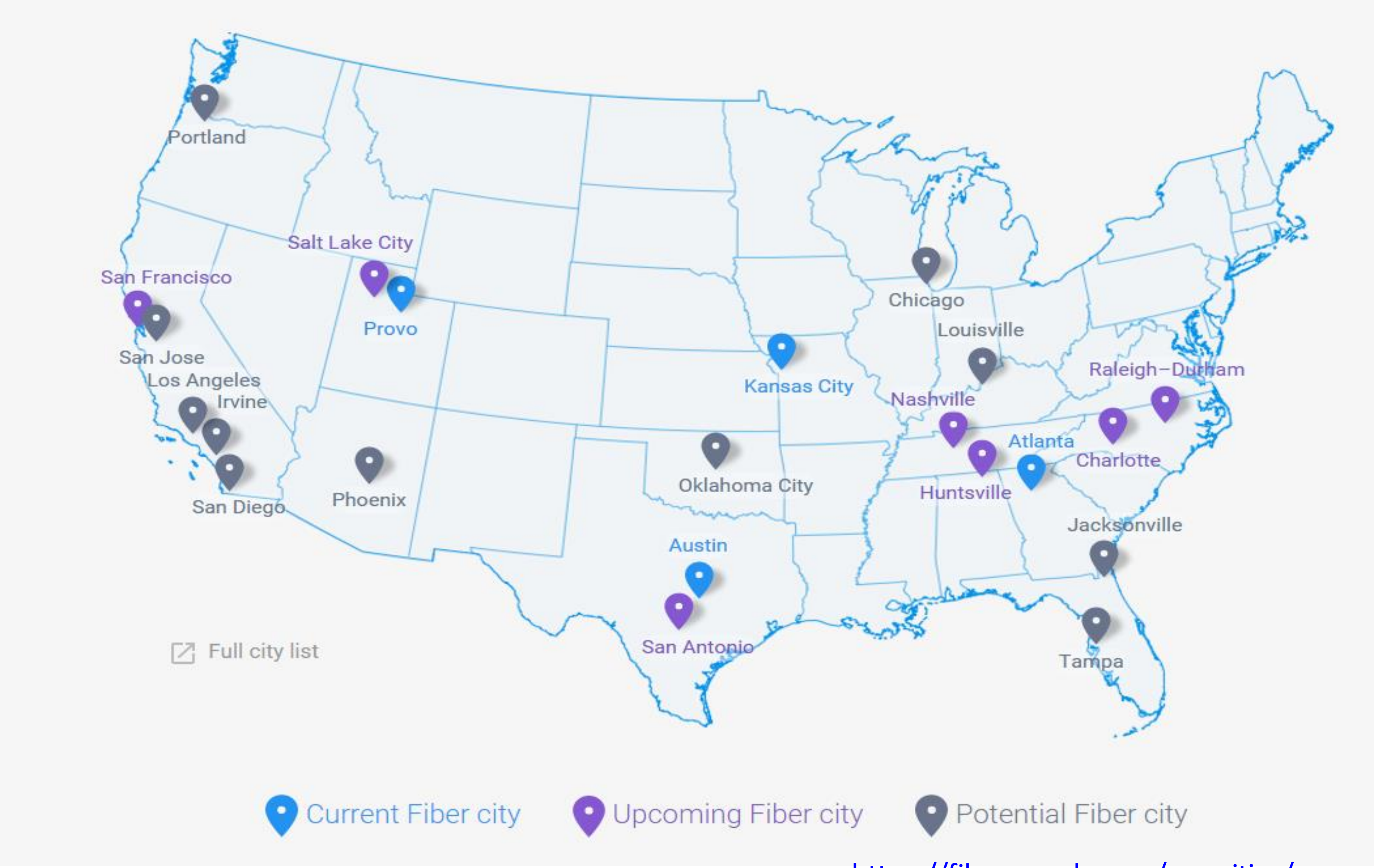
Make: Nissan
Model: Sentra
Year: 2006
Body Type: sedan
Trim: 1.8 s
Price: $5,417

Make: Ford
Model: Econoline-Cargo
Year: 2003
Body Type: van
Trim: e-150
Price: $3,778

Make: Honda
Model: Accord
Year: 1994
Body Type: sedan
Trim: lx
Price:$3,591

Make: Honda
Model: Civic
Year: 2004
Body Type: sedan
Trim: ex
Price:$8,773

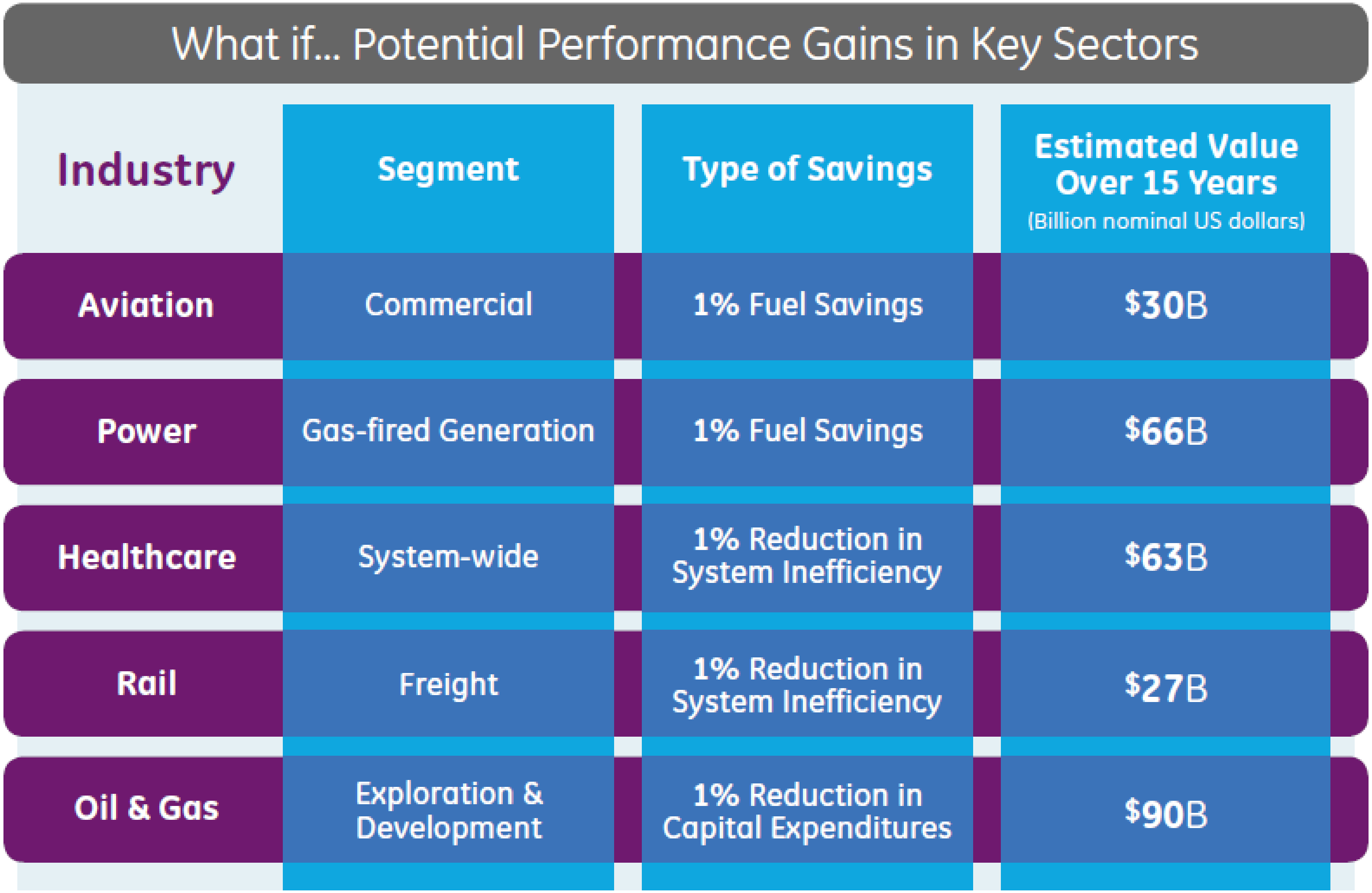
Big Data in Economic Impact of Internet Infrastructure



https://fiber.google.com/newcities/

Google Fiber locations

Big Data in Operations



What if... Potential Performance Gains in Key Sectors

Industry	Segment	Type of Savings	Estimated Value Over 15 Years (Billion nominal US dollars)
Aviation	Commercial	1% Fuel Savings	$30B
Power	Gas-fired Generation	1% Fuel Savings	$66B
Healthcare	System-wide	1% Reduction in System Inefficiency	$63B
Rail	Freight	1% Reduction in System Inefficiency	$27B
Oil & Gas	Exploration & Development	1% Reduction in Capital Expenditures	$90B

(Source: GE)

Big Data in Transportation

- Empty Trailer Backhaul Brokering: Enhancing Revenue and Environmental Sustainability

Big Data in Energy

- Interaction between renewables and traditional electricity generation
- BOTH demand and supply variations
- How to match demand with supply?
  - New market mechanisms
  - Automated agents
  - Smart meters
- Large scale experiments to elicit true preferences
- Causal inference possible

Big Data in Healthcare

- **Medication adherence, opioid use**
- Spatio-temporal analysis
  - e.g. distance to pharmacy
- Household composition
  - e.g. help from family members
- Income effects
- Work condition
- Social networks (phone, FB, etc.)

Several ongoing projects (open to innovative projects)

Sudip Bhattacharjee

Chief, Center for Big Data Research and Applications

US Census Bureau

Professor

School of Business

University of Connecticut

Sudip.Bhattacharjee@census.gov

IRIS Overview

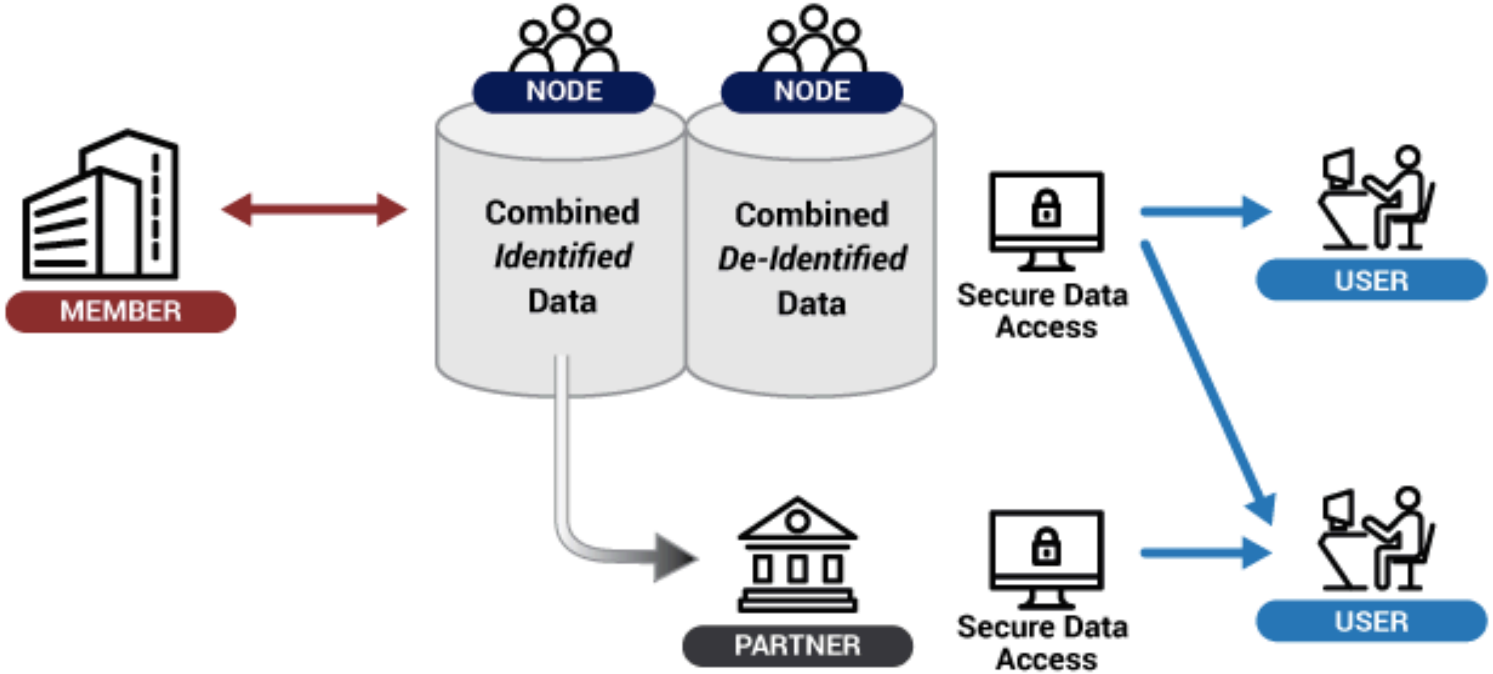
IRIS INSTITUTE FOR RESEARCH ON INNOVATION & SCIENCE

- Intermediary organization to facilitate data sharing for research and reporting among universities, federal agencies, researchers
  - 60 committed university members -- ~$28.7 billion in 2015 expenditures
  - Broad engagement and support from research community/associations
- Platform for fundamental social science research of immediate practical import
  - >70 researchers from 14 institutions have accessed research data to date
  - Sloan support for research community building
- One of the first research/training infrastructures for computational social science
  - Training and courses with ICPSR, JSMP

MEMBERS: Universities contribute data, support infrastructure and receive campus-specific and aggregate reports

NODES: Approved nodes materially improve data, develop products, and expand user communities

USERS: Approved users securely access de-identified aggregate datasets



PARTNERS: Approved partners receive data from IRIS which they improve and make accessible throught their own secure systems

First Research Data Release

- 19 universities
  - $11B in 2014 federal R&D (16% of total)
- Transaction level data
  - 162,694 federal and non-federal sponsored projects
  - 333,565 individuals
    - 28,641 Post-Docs
    - 76,295 Grad Students
    - 87,195 Undergrads
  - $18.1B in vendor spending to 441,796 establishments
  - $6B in subcontracts to other performers
- Links to abstracts etc for federal awards (NIH, NSF, USDA)
- Individual level links to dissertation information
- Title 13 crosswalks to LEHD, LBD, ACS, Decennial Census (available only through the FSRDC system)

Accessing data through the IRIS VDE

- No Census data available in any form, but IRIS data is mirrored in RDCs
- Windows virtual desktop environment, shuts down i/o on your machine.
- Data can be added to your scratch space by IRIS research support staff
- Only aggregate information and statistics such as regression coefficients can be removed
- Export occurs after a privacy disclosure process based on Census procedures
- Restricted access documentation in Wiki format allows user updates
- Online ticket system to report data and software bugs
- No fee now, but a modest fee for researchers who are not affiliated with IRIS member institutions is likely

Accessing data through the IRIS VDE

- Check out background materials and FAQ on IRIS Website (http://iris.isr.umich.edu/research-data/)
- Download and complete application and data use agreement
- All virtual machines are loaded with Windows 7 and the following software, packages, and libraries: Microsoft Office, Stata 14, SAS 9.4, R / RStudio, MATLAB, LaTeX, HeidiSQL, MS SQL Server Management Studio 2014, Gephi, Cytoscape, QGIS, GRASS GIS, Adobe Acrobat Pro,, Notepad++, Python, Anaconda, Jointpoint, PuTTY, WinSCP, and TightVNC Viewer. Researchers can contact IRIS for any questions concerning existing software or to request the installation of additional applications in the VDI.

Research Community Development

- More than **70 researchers** from **14 institutions** have accessed data through either VDE or FSRDC so far
- First research meeting this summer
- Sloan Foundation Support for research grants
    - $15 k Dissertation
    - $30 k early and mid-career grants
    - Call for proposals in Fall 2017
- Constituting a scientific advisory board this summer
- Next data release (target=30 universities) in Winter, 2018.

Sign up for updates on data improvements, funding and training opportunities as well as IRIS news and events via our website's contact page http://iris.isr.umich.edu/contact/

Using Census Data at the Federal Statistical Research Data Centers

Barbara A. Downs

Director, FSRDC

Center for Economic Studies

U.S. Census Bureau

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

FSRDC Environment

Physical Security

- Secure Census facility within host institution
- Census employee on-site at all FSRDCs
- Authorized personnel only
- Researcher Special Sworn Status
  - Requires moderate level background check
  - Oath of confidentiality is for life
- Data accessed via secure connection from thin client device to Census data facility in Bowie, MD
- Printing strictly controlled
- No internet access
- All output reviewed for disclosure risk

FSRDC Environment

Collaboration

- Each project has “home” FSRDC
  - Researchers may collaborate across FSRDCs
  - Projects may move “homes” as researchers relocate
- FSRDC Administrator
  - Coordinates project access across FSRDCs
  - Coordinates review of output

Census Project Proposals

- Three stages of review
  - FSRDC Development and Review
    - Abstract
    - Proposal
    - Predominant Purpose Statement
  - Census Bureau Review
    - 5 concurrent reviews
  - Other Agency Review
    - SSA, BLS, IRS – any agency providing some of the project’s data

Census Project Proposals

Review Criteria

- Scientific merit
- Requires non-public data
- Provides benefit to Census Bureau programs
- Is feasible
- Poses no risk of disclosure of individual or business

Census Project Proposals

Benefits to Census Bureau

- Census-IRS Criteria Agreement
- Helps Census check data it collects, edits, and tabulates
  - Permits rigorous analysis of confidential data
  - Tests validity of data processing rules
  - Evaluates conceptual and processing assumptions
- Prepares new economic or population estimates

Census Project Proposals

Timing

- Census review
  - ~75 days
- Other Agency review
  - 3 to 6 months
- Special Sworn Status
  - Concurrent with Agency review
  - 3 to 5 months

Thank You!

- Links
  - Federal Statistical Research Data Centers

    www.census.gov/fsrdc
- Contact
  - Barbara A. Downs (barbara.a.downs@census.gov)