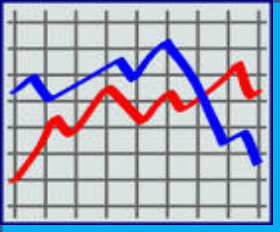


Uses of Big Data for Official Statistics:

Privacy, Incentives, and Statistical Challenges

Discussion Paper by

Steve Landefeld

Senior Advisor to the United Nations Statistics Division for the

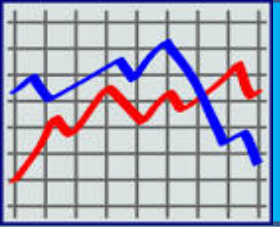
International Conference and Global Working Group Meeting on Big Data for Official Statistics

United Nations Statistics Division

National Bureau of Statistics of China

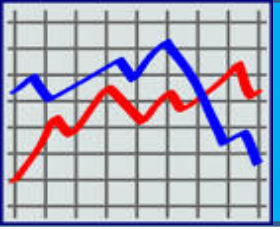
28-31 October 2014

Beijing, China



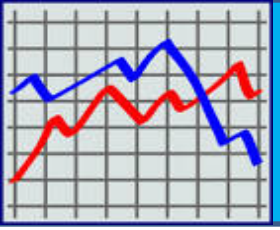
The Growth and Potential Impact of Big Data

- U.S. White House report describes Big Data as "fundamentally reshaping how Americans and people around the World live, work, and communicate.“
- UN Global Working Group formed
  - To provide a strategic vision, direction, and a global programme on Big Data for official statistics.
  - To promote practical use of sources of Big Data for official statistics, while finding solutions to their challenges.
  - To promote capacity building and sharing of experiences.



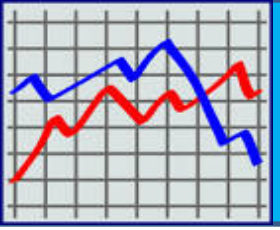
The Growth and Potential Impact of Big Data: Factors Behind Growth

- Advances in information technology that have lowered data collection, storage, and processing costs.
- The development of new sources of data and improved access to existing big data sets, on and off-line,
- The parallel development of creative and powerful new methods to exploit Big Data.
- The recognition that we have massive stores of data collected for such purposes as business and administrative that can be used for a wide array of purposes other than those for which they were originally collected.



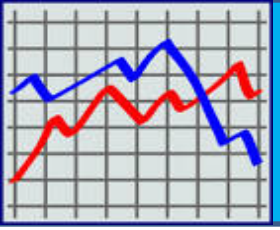
Uses of Big Data for Official Statistics

- Long history of use in National Accounts, mainly as extrapolators to produce more timely and accurate early estimates.
  - Mainly aggregations of business and government micro-data, more recently using internet data scraping, but
  - Micro-data is used in matching of statistical and non-statistical data to develop bias adjustments for survey data, improvements in coverage, and to identify reporting and other problems.
  - Micro-data matching of existing official statistical data sets is also being used to develop new and more detailed data.



Uses of Big Data for Official Statistics

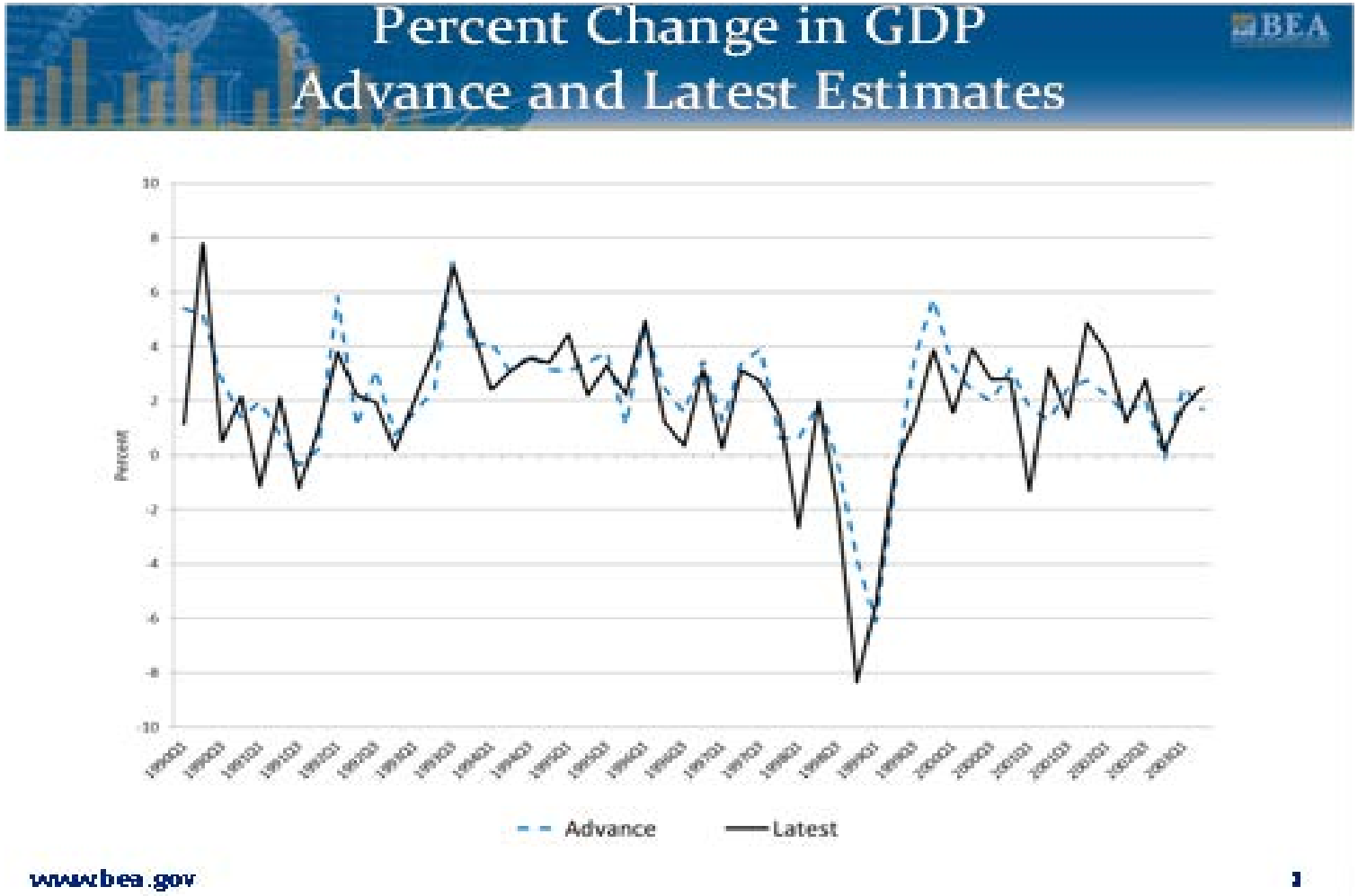
- For privacy, cost, and efficiency reasons this pattern of the use of big data collected for non-statistical purposes -- as extrapolators and methodological research and improvement tools -- is likely to continue.

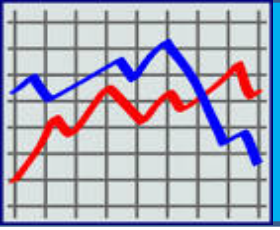


Uses of Big Data for Official Statistics

- Examples of the private source data aggregations included in the accounts are:
  - Ward's/JD Powers/Polk (auto sales/price/registrations)
  - American Petroleum Institute (oil drilling)
  - Airlines for America (airline traffic)
  - *Variety* magazine (motion picture admissions)
  - STR (hotels and motels)
  - Investment Company Institute (mutual fund sales)

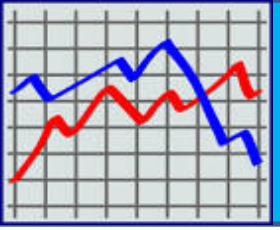
Uses of Big Data for Official Statistics





Uses of Big Data for Official Statistics

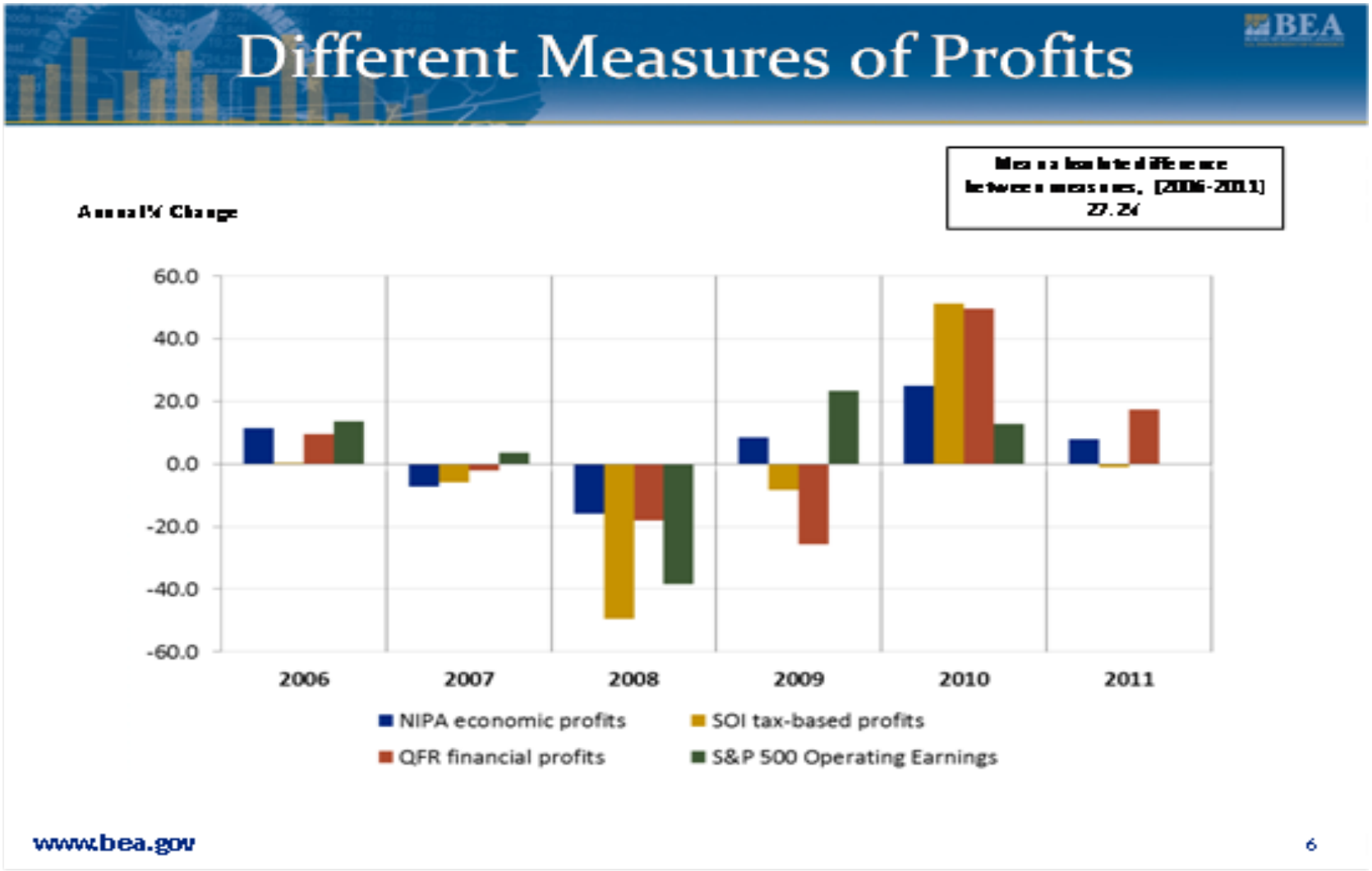
- Examples of the matching of micro-data to improve and expand official estimates
  - Aggregate adjustments by type of income derived from matching of tax and survey data
  - Much more detailed FDI data by matching BEA, BLS, and Census data
  - Construction of disease-specific and aggregate health care price indexes to address using health insurance records

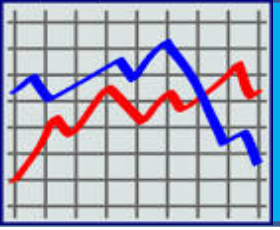


Uses of Big Data for Official Statistics

- Big Data collected for non-statistical purposes works well and lowers costs relative to official surveys but requires a fair amount of work, including adjustments for differences in:
  - Concepts and definitions
  - Representativeness
  - Timing
  - Collection and estimation methods

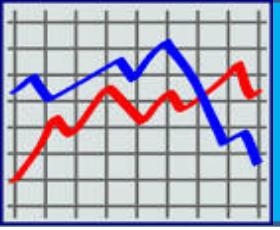
Big Differences in Seemingly Similar Big Data





Process for Evaluating and Using New Big Data

1. Evaluate concepts, definitions, coverage, and performance of extrapolators relative to more comprehensive and consistent annual and benchmark data.
2. Develop new methods to use new extrapolators, including benchmarking, weighting and combining with other indicators, and bias adjustments. (Avoid complex “black boxes”)

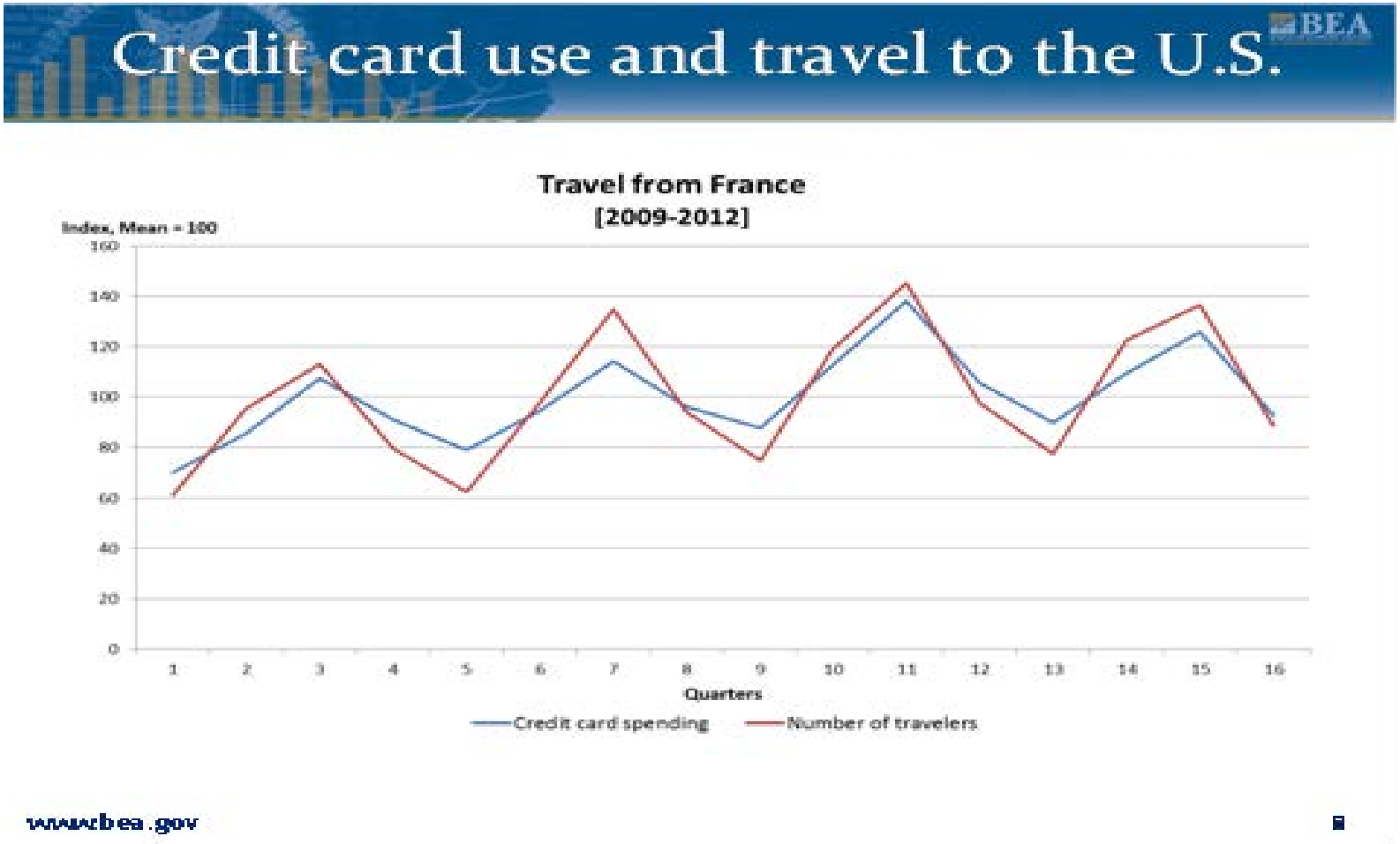


Process for Evaluating and Using New Big Data

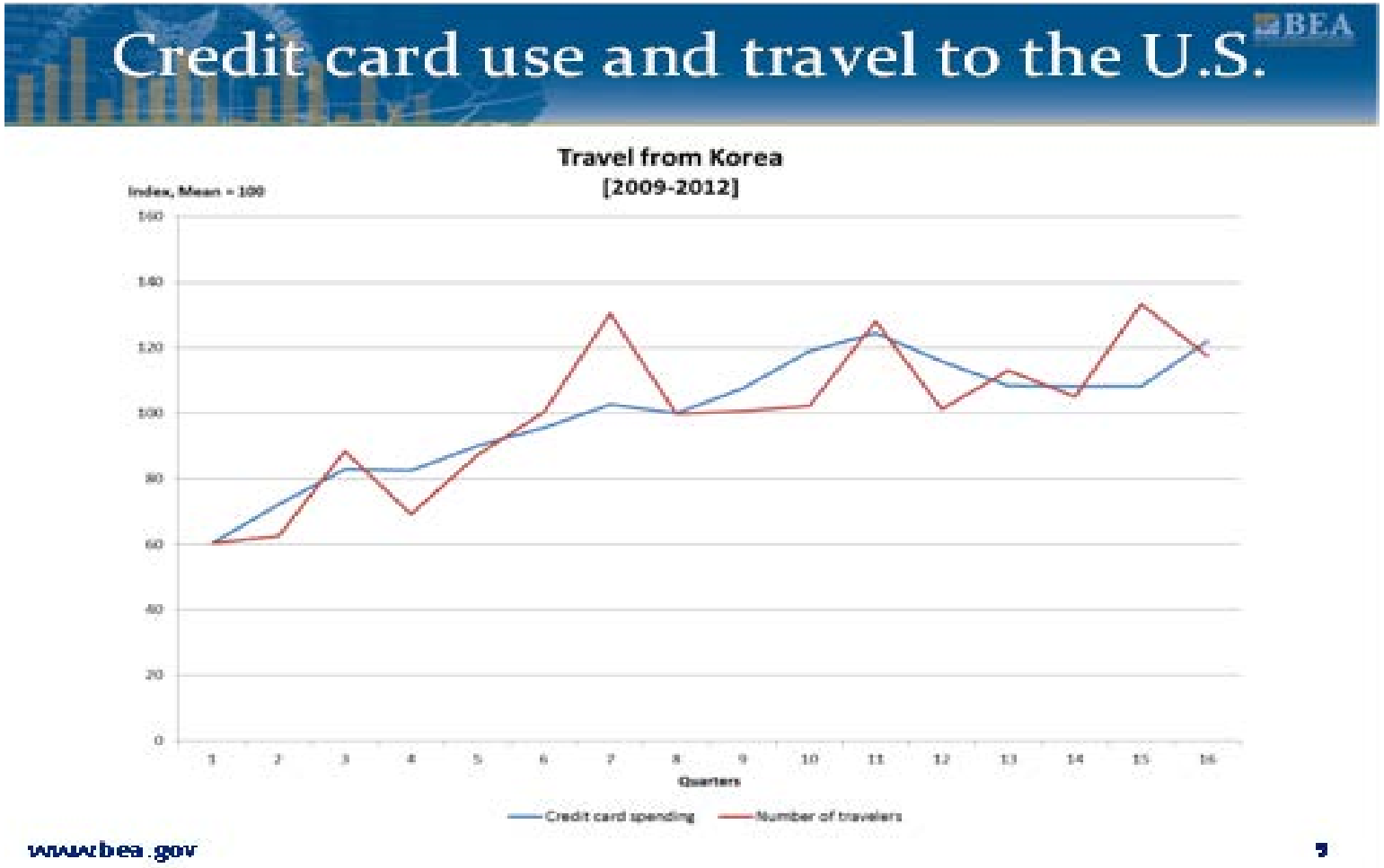
3. Evaluate newly-developed extrapolators relative to existing extrapolators and benchmarks to access their accuracy.

4. Develop seasonal adjustment factors for the new extrapolated data, which is a difficult process given the normally short time series for the new data.

What Works For Some Countries May Not Work for Others



What Works For Some Countries May Not Work for Others



What Works for Some Products May Not Work Well for Others

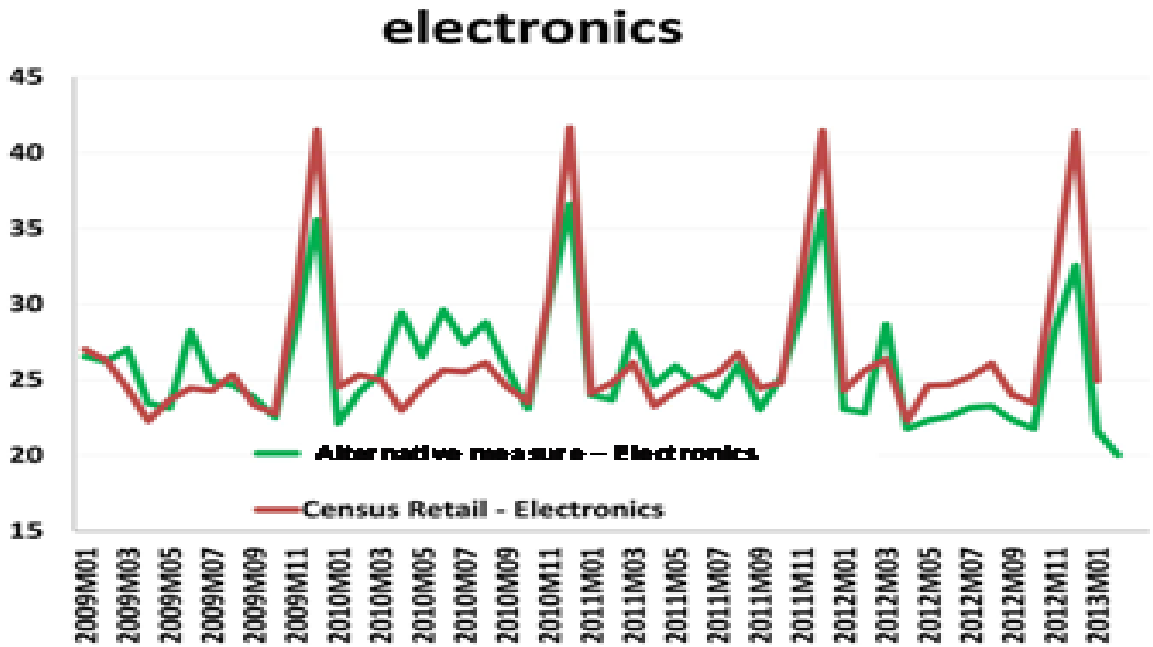
Alternative measures of consumer spending

www.bea.gov

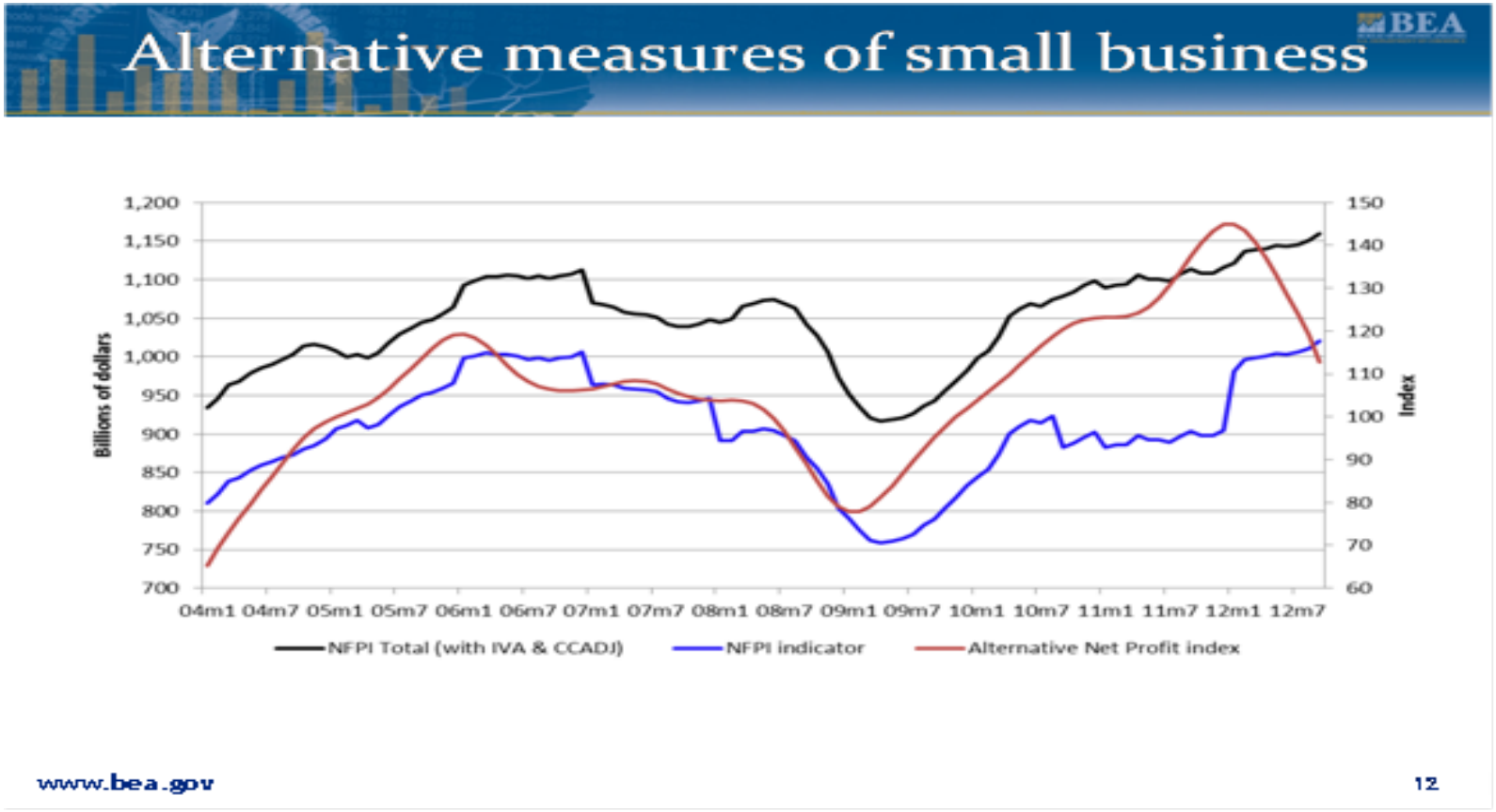
10

What Works for Some Products May Not Work Well for Others

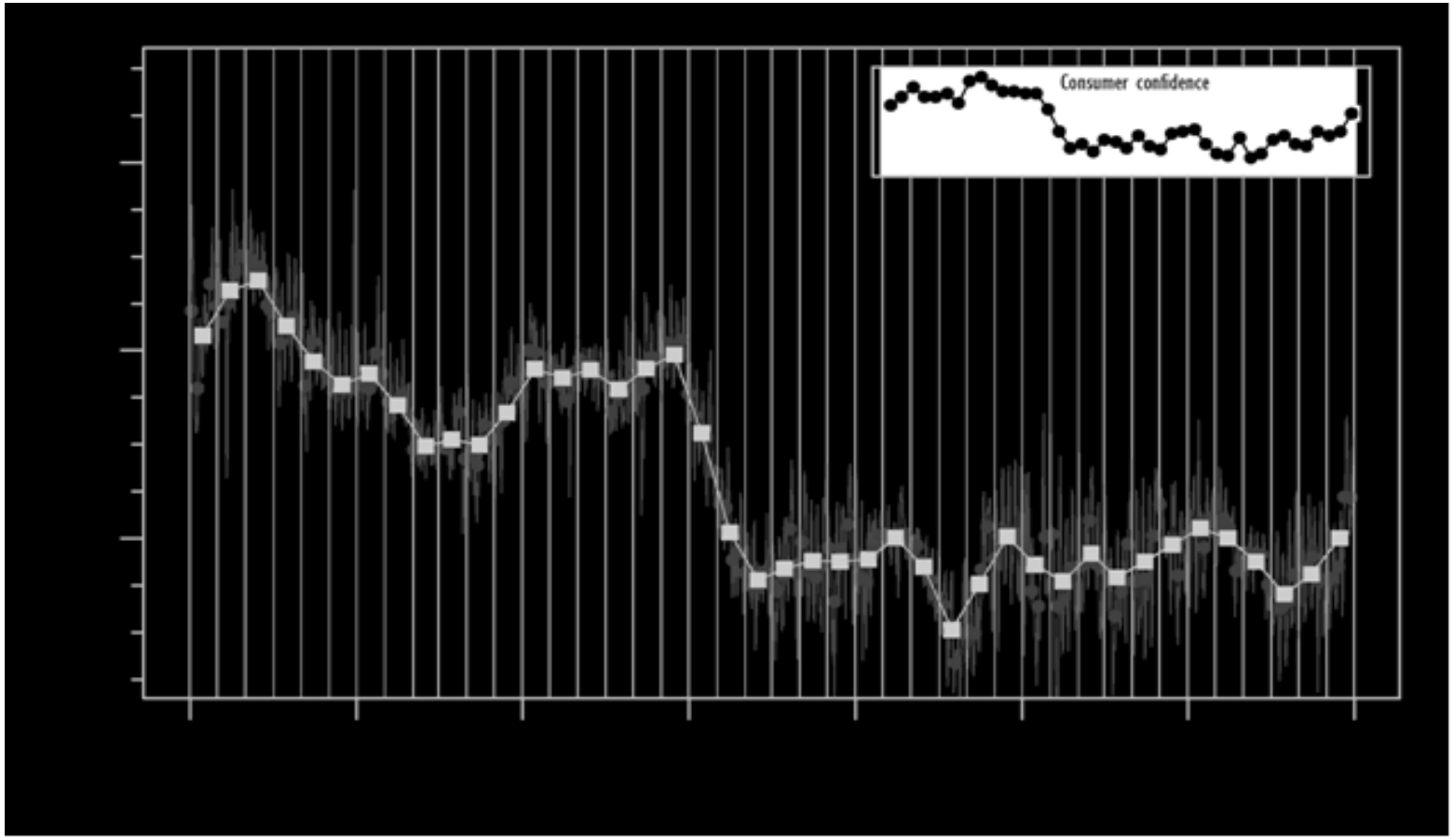
Alternative measures of consumer spending

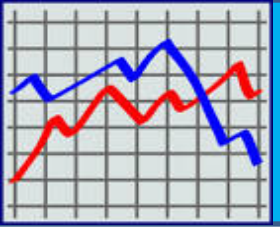


Importance of Consistency over Time



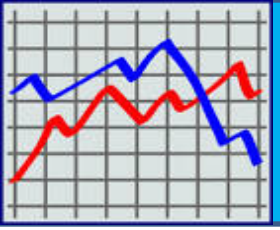
Similar Lessons Apply to Social Media: Netherlands vs. U.S. Experience





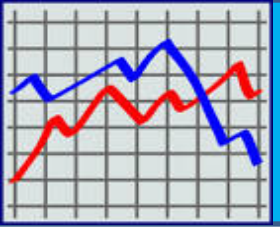
Loss of Control in Using Big Data

- In addition to evaluating the accuracy and cost implications of using Big Data, statisticians need to weigh the potential loss of data and increased dependency associated with use of such data
- Raises the importance of ties to official benchmarks, flexibility, and as noted below attention to incentives and clarity (through protocols) as to the purposes and responsibilities of participants to data agreements



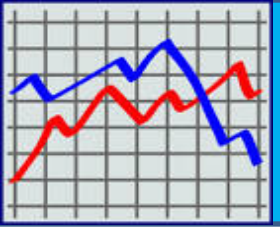
Incentives: Exploiting Public and Private Benefits of Big Data

- Official statistics:
  - More accurate, timely, detailed, and relevant data at significantly lower cost
  - Better branding
  - Increased statistical literacy
- Private Owners of Big Data
  - Benchmarked, more comprehensive, and more accurate data for marketing, location planning, and long and short-term investment decisions
  - Early and accurate data on government indicators is big money!



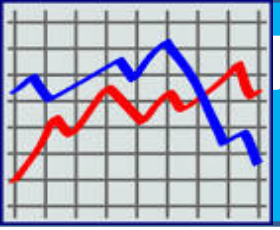
Incentives: Exploiting Public and Private Benefits of Big Data

- Public owners of big data
  - Benchmarked, more comprehensive, and more accurate data for infrastructure planning, provision of services, and designing and assessing tax, regulatory, and health insurance and other programs.
- Search engines
  - More accurate, timely, and relevant data searches
  - Important information to customers on the meaning, and uses of, and match-ups of seemingly similar data



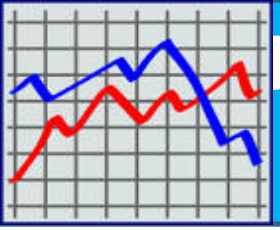
Incentives: Exploiting Public and Private Benefits of Big Data

- Successful public-private collaboration requires:
  - A recognition of the mutual benefits of such a collaboration
  - Transparency between the official statistical agency and owner of the business or administrative data regarding data collection and estimation methods used to produce the data; and
  - Clear and strong rules for protecting the confidentiality of the data and of the proprietary methods used to produce the business or administrative data



Privacy Concerns: Types of Business Data That Must Be Protected

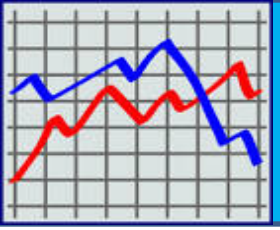
- The data itself is an information product (micro and macro) that is the intellectual property of the firm, has economic value, and can be sold.
  - Government must make sure there is no disclosure that would give it away free.
- Data on details of businesses, such as prices, costs, and market share, that would be useful to competitors.



Privacy Concerns: Types of Business Data That Must Be Protected

- Personal information on customers. Loss of such data through security breaches or hacking undermines the reputation of the firm, and discourages use of electronic transactions that can result in the loss of business.

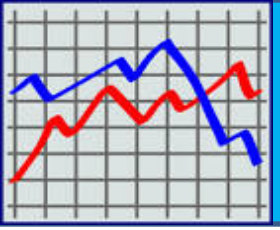
- Proprietary information on the methods and sources used to produce the data are also the intelectual property of the firm and must be protected.



Privacy Concerns: Types of Household Data That Must Be Protected

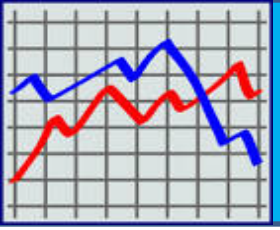
- Name and address or other identifying information that could be used for marketing and other business purposes.

- Intimate personal details, including such information as marital and health status and income.



Privacy Concerns: Types of Household Data That Must Be Protected

- Any information whose use could lead to discriminatory outcomes in such areas as employment, eligibility for loans, or eligibility for government programs.
- Any disclosure to non-statistical agencies that alters the balance of power between individuals and government, including:
  - The use of data collected for statistical purposes for tax, regulatory, investigatory and other non-statistical purposes.



Data Protocols to Protect Privacy Are Essential

- Such protocols to protect privacy are essential and have been used for years to promote trust and address concerns that government may use micro-data for regulatory, tax, and other policies.

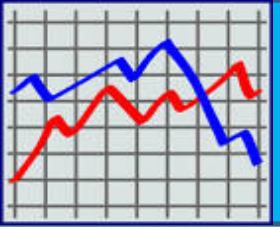
- An erosion of public trust can reduce response rates on official surveys, reduce honesty in reporting, and reduce the overall accuracy of the official statistics collected from the business and household community.

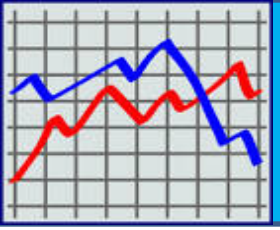
Model Agreement for the Provision of Administrative Records for Statistical Purposes

- Parties to the Agreement
- Legal and Programmatic Authority
- Duration or Period of Agreement
- Purpose
- Use of Data
- Data Quality
- Roles and Responsibilities for Data Protection
  - Confidentiality and Privacy
  - Data Security
  - Data Transfer, Media and Methods for Transmission of Data
  - Record Keeping, Retention, and Disposition of Records

Model Agreement for the Provision of Administrative Records for Statistical Purposes

- Specific Penalties for Unauthorized Disclosure of Information
- Potential Work Constraints
- Breach
- Disclaimers
- Reporting
- Administrative Points of Contact
- Funding Information
- Estimated Costs and Payment
- Resolution of Conflicts
- Modification/Amendment of Agreement
- Cancellation of Agreement
- Periodic Review of Agreement
- Concurrence and Agency Signatory



Conclusions

- Use of Big Data is not new and valuable lessons learned that can be applied now
- Big data has big potential but no panacea and not without costs
- But with careful attention to incentives, privacy, and integration, Big Data can play an important role in improving:
  - Accuracy, timeliness, and relevance at a lower cost than expanding existing data collections