Collaborations between Official Statistics and Academia in the Era of Big Data

World Statistics Day
October 20-21, 2015
Budapest

Vijay Nair
University of Michigan

Past-President of ISI
vnn@umich.edu

What is “Big Data”?

Latest buzzword used to describe the “data tsunami” and the “information revolution”

McKinsey Global Institute




June 2011

Big data: The next frontier for innovation, competition, and productivity

BIG DATA: Is it really new?

Large Datasets → Massive Datasets →Big Data

1996: Massive Data Sets: Proceedings of a Workshop. US National Academy Press, Washington, D.C.

Background

Beginning in the 1980s:

- Rapid advances in technologies for:
  - Data capture
  - Data storage
  - Data transmission

  → Increase in the amount of data that are collected routinely

- Rapid advances in computing power

  → Increase in ability to analyze large data sets quickly and efficiently

SO … WHAT IS NEW WITH BIG DATA?

Tremendous Increases in SCALE in …

VOLUME, VARIETY, VELOCITY (3 V's) + Veracity and Validity

VOLUME:

- Massive increases in data volume (from terabytes to petabytes to exabytes)
- Lots and lots of "smaller" data sets too …

- Science: Climate, Environment, Space, Astronomy, Genomics
- Health: Electronics medical records, Personalized Medicine
- Transportation, computer and communication networks
- Search engines
- Customer transactional data
- Social media and social networks (Facebook, Twitters, Smart phones, …)

Big data—a growing torrent

Source:
McKinsey Report

5 billion mobile phones in use in 2010

30 billion pieces of content shared on Facebook every month

15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress

235 terabytes data collected by the US Library of Congress by April 2011

Decreasing cost of measurement, storage, and computing technologies

$600 to buy a disk drive that can store all of the world's music

What's new with BIG DATA (contd.)

- VARIETY (heterogeneity):

  "Unstructured" data:

  – text, speech, video, images, network structure, hierarchical structure, …

  – Issues with storage, data compression, modeling, and analysis

- VELOCITY (Timeliness):

  – Streaming data: VOIP (skype, google hangout), online games, …

  – Can't store all the data → implications for analysis and algorithms

DEFINITION OF BIG DATA? (McKinsey Report)

Box 1. What do we mean by "big data"?

"Big data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data—i.e., we don't define big data in terms of being larger than a certain number of terabytes (thousands of gigabytes). We assume that, as technology advances over time, the size of datasets that qualify as big data will also increase. Also note that the definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry. With those caveats, big data in many sectors today will range from a few dozen terabytes to multiple petabytes (thousands of terabytes).

INCOMPLETE: REFRS ONLY TO SIZE (VOLUME) OF DATASETS

What else is new with Big Data?

- Investment in Funding and Infrastructure

OBAMA ADMINISTRATION UNVEILS "BIG DATA" INITIATIVE: ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS

- Advance state-of-the-art core technologies needed to collect, store, preserve, manage, analyze, and share huge quantities of data.
- Harness these technologies to accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning; and
- Expand the workforce needed to develop and use Big Data technologies.

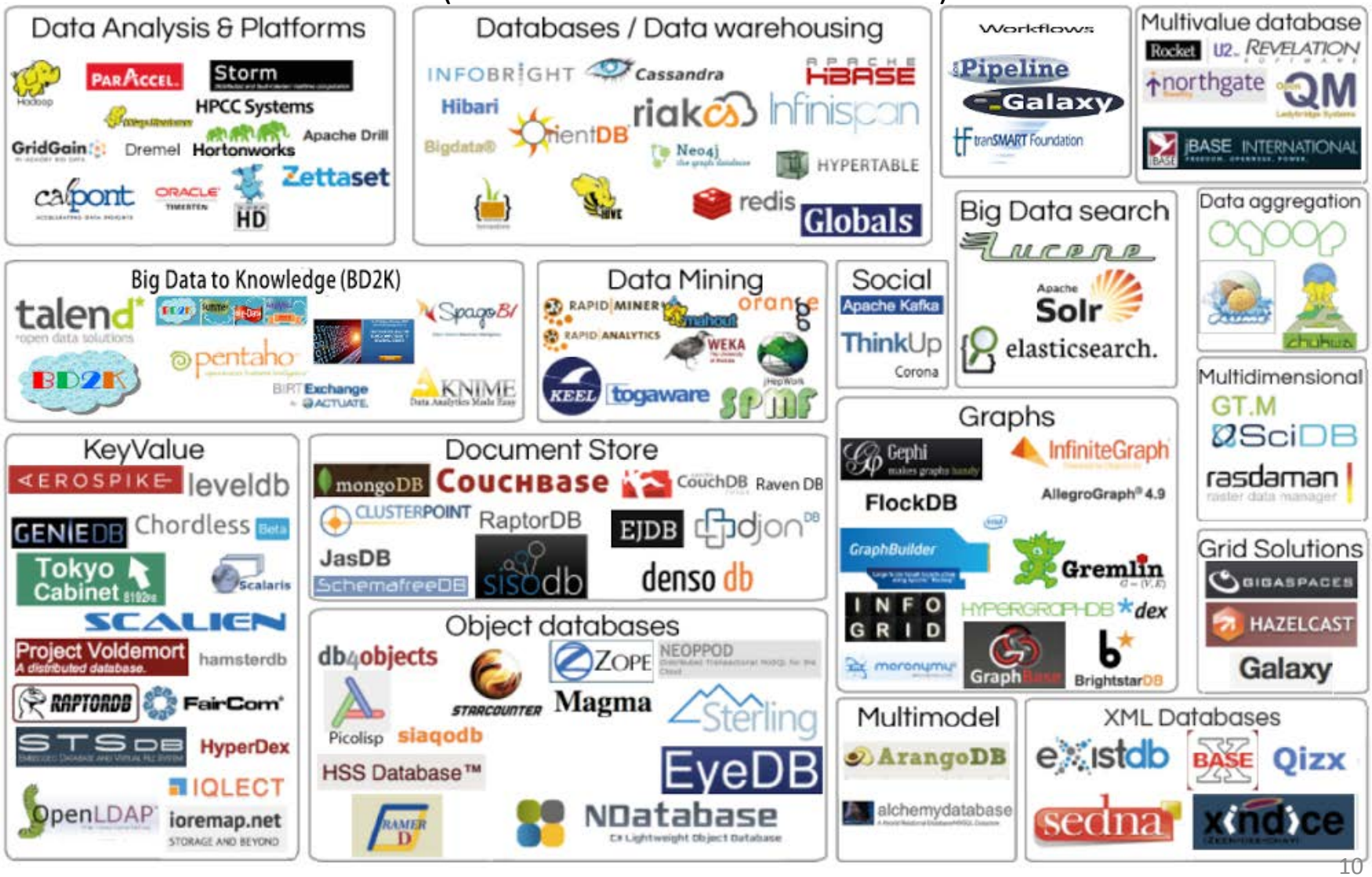
- Office of US Chief Data Scientist
- National Institutes of Health – BD2K Initiative: Big Data to Knowledge
- **Parallel initiatives by governments in other countries, UN, World Bank, …**

- **Investments by business and industry in infrastructure**

Infrastructure and Computing Platforms

(www.bd2k.org/BigDataResourceome.html)

(Slide from Ivo Dinov and MIDAS)



Opportunities

Science, Health, and Engineering

- Climate, Environment, Genomics, Space, Astronomy
- Transportation, Computer and Communication Networks
- Health: Electronics medical records, Personalized Medicine

Business, Industry, and Public Sector

- Customer transactional data
- Search engines
- Social media and social networks
  - (Facebook, Twitters, Smart phones, …)
- Community detection → National and International Security

Opportunities in Official Statistics

- Sources of Data
  - Surveys … statistically well-designed studies

    BUT … becoming increasingly costly, non-response for surveys

    (interestingly … people are still willing to provide information on social media)
  - Web surveys are less expensive … response bias … and non-response is still problem

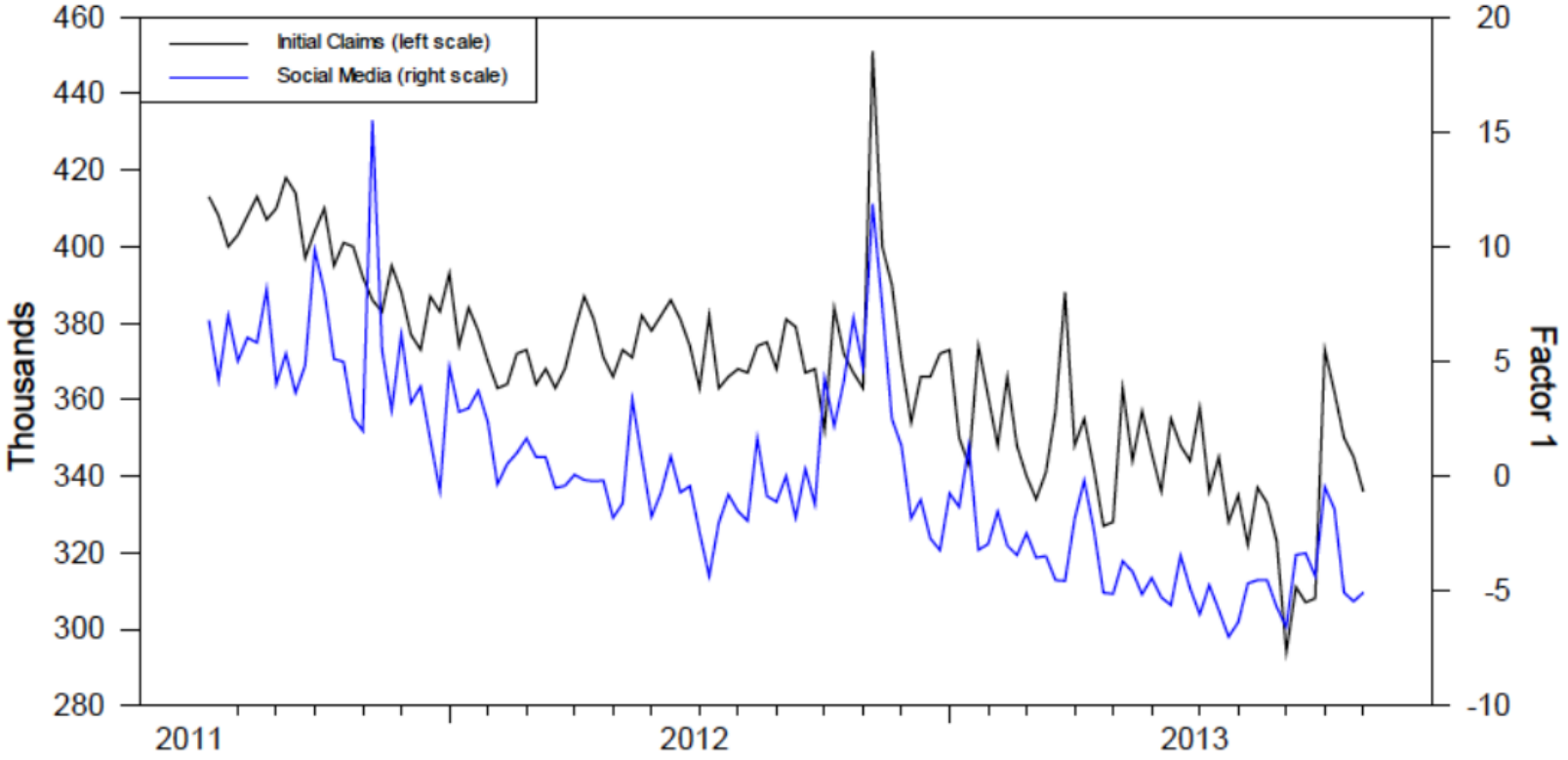
- Challenges – survey information not timely … forecasting based on past … need to know what is happening now – "nowcasting"

- Alternative – look at secondary sources of data to "mine" information
  - examples: Health epidemics→ Google flu trends;
  - use of mobiles for traffic patterns and jams

An Example

- Using Social Media to Measure Labor Market Flows – Antenucci et al. UM Tech Report
- Measure economic activity + analyze economic behavior in real time, using information independent from surveys and admin data …
- Use twitter data to create indices for job flows: job loss, job search, and job posting …
- Based on job-related tweets such "lost my job" or "pink slip" or "canned" or "fired" …
- Analysis based on 10% of all twitter data from July 2011 to Nov 2013 – 19.3 billion tweets and ~44 terabytes of data
- Search twitter data for relevant features – k-grams (based on subject matter knowledge, post-processing to "clean" the data …
- Better performance on economic data on initial job unemployment insurance claims … more timely information →

Figure 2. Initial Claims for Unemployment Insurance and Job Loss and Unemployment Factor 1



Note: Figure shows the Department of Labor's Initial Claims for Unemployment Insurance (left scale, revised data, seasonally adjusted) and the Social Media Factor 1 (right scale). The factor is estimated as described in the text and is no way fit to the initial claims data.

Challenges

- Secondary (readily available) data … may contain useful information
- BUT … hard to know up from how useful … the degree of usefulness varies a lot
- How to decide upfront if it worth investing?

- Lots of heterogeneity, missing data, not right variables of interest, sparse information in some areas than others; duplication … signal to noise can be small
- Lots of biases in the data … observational data … so statistical inference about population of interest or prediction can be tricky
- Correlation vs Causation → does not tell you why

- Data access – data "owned" by someone else (Google, Facebook, etc.) – restrictions on access
- Concerns about privacy and security

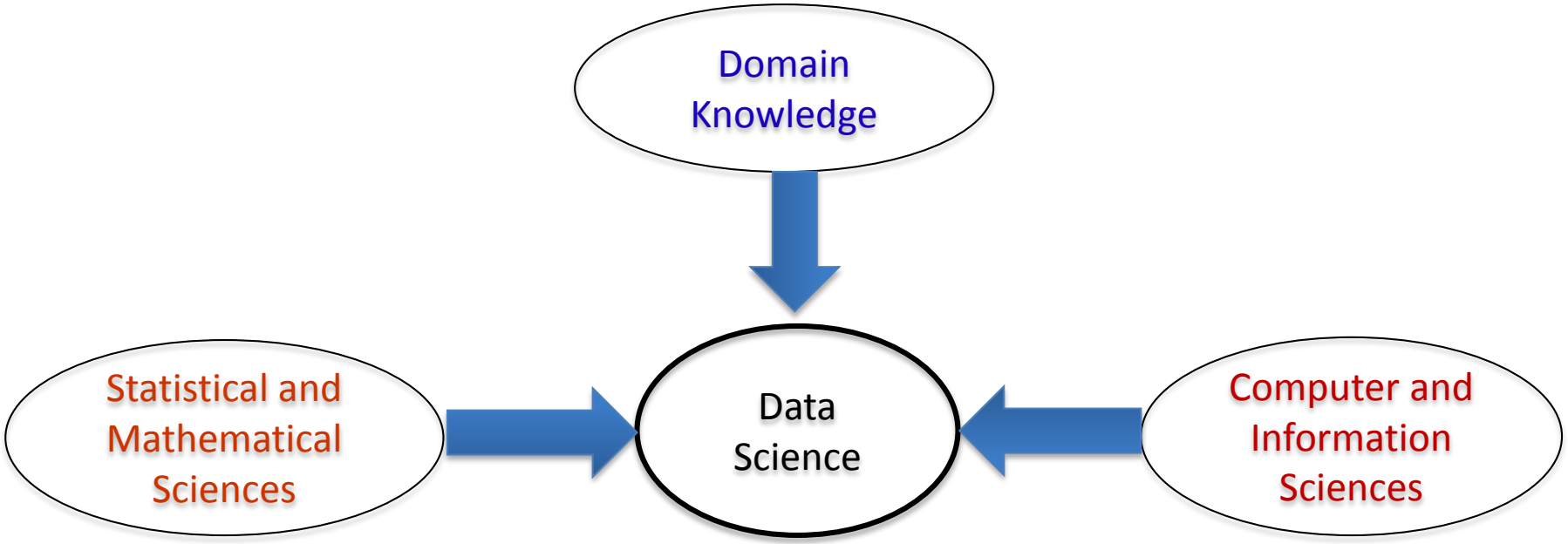
Academia and Official Statistics: Areas for Collaboration

Extensive Research in Academia

- Data management, storage, and retrieval
- Computing platforms
- Data provenance, curation, …
- Privacy, security, confidentiality, …
- Visualization
- Data analytics
  - Algorithms
  - Scalability for large datasets, divide and conquer approaches, …)
  - Analyzing complex data – satellite images, graphs, networks
  - Combining heterogeneous data from multiple sources
  - DEALING WITH BIASES IN OBSERVATIONAL DATA

Emergence of Data Science as a "new field"

Wikipedia … "study of the generalizable extraction of knowledge from data"



Concepts, methods and algorithms involved in collecting, curating, managing, analyzing, and transforming *data* into information …
to enable knowledge creation and decision-making in a variety of application domains

Education and Training

- Huge demand for data scientists – "supply" lagging …
- US – expected shortfall of people trained in data science about 200,000 by 2018 (McKinsey Report)
- Even bigger need for managers with appreciation for data
- Need for retraining of professionals
- Many DS Initiatives in Universities and Industry and Funding Agencies
- Span research, technology and infrastructure, applications, education and training

University of Michigan
Data Science Initiative – $100 mill

- MIDAS – institute to foster research, training, and collaborations
  - 35 new faculty
  - Support key research areas
  - Industry outreach
- Infrastructure
  - Big data computing and data storage
  - Consulting and Support
- Educational programs
  - DS Undergraduate Program (joint between Statistics and CS)
  - Certificate Program
  - Plans for Master's and PhD programs

Models for Collaboration

- Partnerships between Statistical Agencies, Academia, and Industry
- Some of this is happening but not nearly enough
- **Role of the International Statistical Institute**
  - Membership includes National and International Statistical Agencies, Academia, Central Banks, and Private Sector
  - Extensive expertise in relevant areas among individual members from academia and other areas
  - Considerable focus on capacity building
  - Can bring together relevant people from ISI and other organizations and help form teams to support initiatives

Contact the ISI Office, President, or me … vnn@umich.edu

Does Big Data mean the end of scientific theories?

Chris Anderson (2008 Wired essay):

Big Data renders the scientific method obsolete:

Throw enough data at an advanced machine-learning technique, and all the correlations and relationships will simply jump out. We'll understand everything.

OPPOSING VIEWS:

- **You can't just go fishing for correlations and hope they will explain the world. If you're not careful, you'll end up with spurious correlations.**
- **Even more important, to contend with the "why" of things, we still need ideas, hypotheses and theories.**
- **If you don't have good questions, your results can be silly and meaningless.**
- **Having more data won't substitute for thinking hard, recognizing anomalies and exploring deep truths.**

Thank you!