Training and Capacity Building for the Data Revolution: Partnerships with Academia, Business and Industry

High-Level Forum

UNSC

March 2, 2015

Vijay Nair
International Statistical Institute
University of Michigan, Ann Arbor, USA
isi-president@umich.edu

BIG DATA

Large Datasets → Massive Datasets →Big Data

1996: Massive Data Sets US National Academy Report

What is New?

Increase in Scale → in VOLUME, VARIETY, VELOCITY (3 V's)

Volume: Massive increase → Digital age

Gigabytes to Terabytes to … Petabytes … Exabytes

BUT … definition of BIG (volume) varies by area

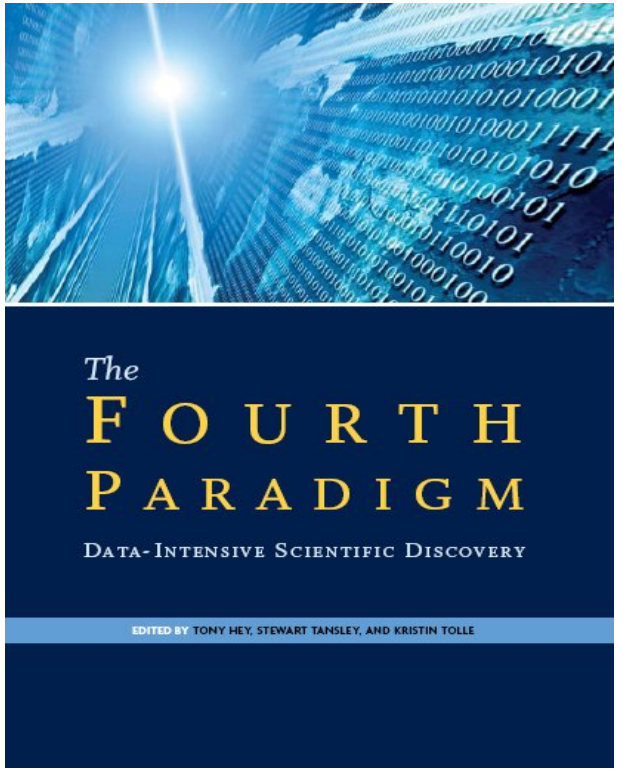
Variety → text, images (satellite and others), social networks, …

Velocity → streaming data

Developments in academia, business and industry

- Huge increase in the awareness of the potential of "DATA"
- **Data-based discoveries** in science, medicine, engineering, and technology
- **Competitive advantage** for business and industry
- **Timely information** for governments and society

"Data science [has emerged] as a fourth mode of scientific discovery … on par with experimental; theoretical; and computational analysis."



Big Data Infrastructure Resources

www.bd2k.org/BigDataResourceome.html

Related Initiatives

Federal agencies: Several hundred million dollars a year in research funding: in US by NSF, NIH, …

Foundations: Moore-Sloane → more than $ 30 million last year on creating an environment for Data Science

Corporations: Many … Google → Google Correlate, Google Trends, Google Consumer Surveys …

Universities: Many …
U Michigan … $42 million on Data Science Initiatives
Infrastructure, consulting and service
Research, Faculty hiring, Industrial outreach,
Education and Training in Data Science → demand for Data Scientists (Statisticians or …?)

Types of "Big Data" and Issues

- Primary: Data collected to study specific scientific, health, engineering or public sector problem
  - Biology (genomic, proteomic, …)
  - Electronic medical records
  - Transportation Networks
  - Satellite Image
  - Census or Large-Scale Surveys

- Secondary (data that are readily available or collected for different reasons and that potentially has useful information for other purposes)
  - Customer transactional data
  - Web searches
  - Social media (facebook, LinkedIn, Twitter)
  - Sensor data (automobiles, smart phones, …)

Types of "Big Data" and Issues

- Primary data
  - Typically collected with some specific goals and hypotheses in mind
  - Data access – people who collect, analyze, and make conclusions are typically owners of the data
  - Issues related to data security, sharing may have been thought out
  - Often, population and inference issues are clear

Challenges related to volume, variety or velocity of data:

Storage, management, privacy, analysis, algorithms …

Secondary Data Sources: Challenges

- Data (such as social media data) collected for different reasons
- May contain useful information for your purpose (official statistics)
- BUT … hard to know up front how useful … signal to noise can be small
- Lots of problems … heterogeneity, missing , duplication …

- Lots of biases in the data … observational data … so statistical inference about population of interest or prediction can be tricky
- Correlation vs Causation

- **More useful for finding interesting information, patterns and trends, downturns and upturns in a timely manner ("nowcasting") than formal inference**
- **Data access – data "owned" by someone else (Google, Facebook, etc.) – restrictions on access, concerns about privacy and security**
- **Data provenance …**

Data Sources: Surveys, …

- 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
  – Administrative Data

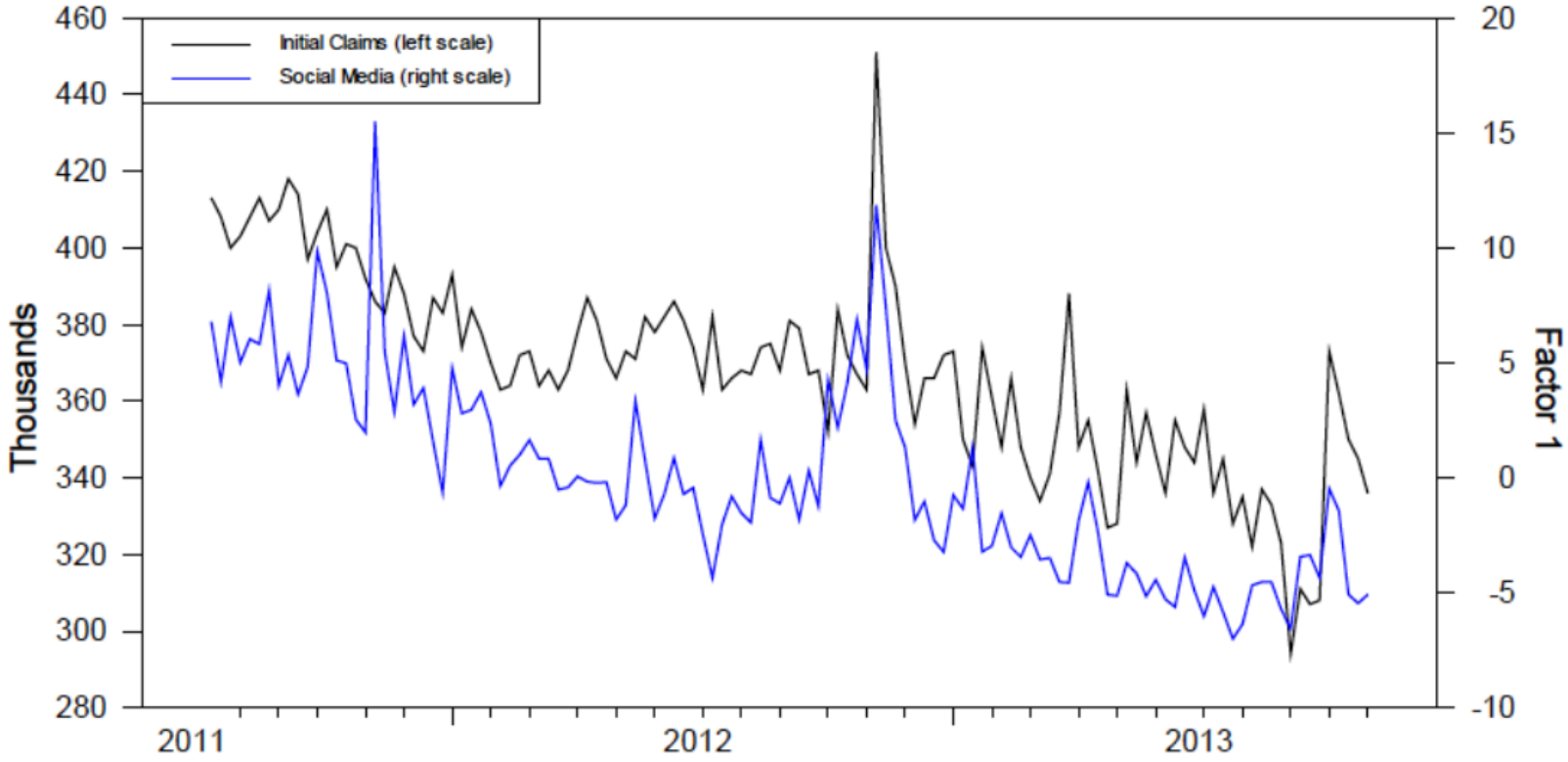
- Challenges – information not timely … forecasting based on past … need to know what is happening now – "nowcasting"

- Alternative – looking at secondary sources of data to "mine" information –
- Examples: Health epidemics → Google flu trends;

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.

Antenucci et al. University of Michigan

Partnerships with Academia, Business and Industry:

Areas

- Expertise in Development of Infrastructure:
  - Cloud Computing and Storage
  - Database Management, Provenance, Curation
- Data Privacy and Security

- Data Quality
- Statistical Analysis (Analytics)
  - Data Mining
  - Combining Heterogeneous Data Sources
  - Complex Data Structures
  - Statistical Learning
  - Causal Inference from secondary sources of data

Collaboration between different disciplines

International Statistical Institute

Promote the understanding, development and good practice of statistics worldwide

- Celebrating 130 years
- Covers all areas of statistics
- Includes 7 international societies under its umbrella
- Almost 5,000 individual members (more than 140 countries) in the ISI family
- 100+ institutional members (most NSOs), central banks, etc.
- Connections to other professional societies

- **Broad, world class technical expertise among members**
- **Members eager to contribute to capacity building worldwide**

ISI: Contributions

- Statistical Capacity Building
  - Leadership and good governance workshops in official statistics: Adis Ababa, Senegal, S. Korea, **Tanzania (April, 2005)**
  - Training in methodology and computing (Africa, India, Latin America, SE Asia …)
  - Support for developing-country participants to attend WSC, short courses, regional and association conferences, etc.
  - Funded by World Bank, African Development Bank, etc.
- Promoting the careers of women and young statisticians

  Committees, networking, workshops, awards, …
- ISI Declaration of Professional Ethics …

ISI is fully committed to be a partner in the Data Revolution for Sustainable Development

- **Harness its members, societies, and international network**
- **Help to engage partners from academia, business and industry**