Corporate Data Access and Sharing: Private Data Resources for the Common Good

Steve Eglash, Executive Director
Data Science Programs, Stanford University

WORKSHOP ON THE CURRENT PRACTICES OF PRIVATE COMPANIES AND THEIR USE OF BIG DATA AND KEY ISSUES AND CHALLENGES WITH PRIVACY AND CONFIDENTIALITY

Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods

February 25, 2016; Stanford University; Stanford, CA

The Goal

Accessing statistical data,
from the private sector,
to be used in combining multiple data sets,
for the common good

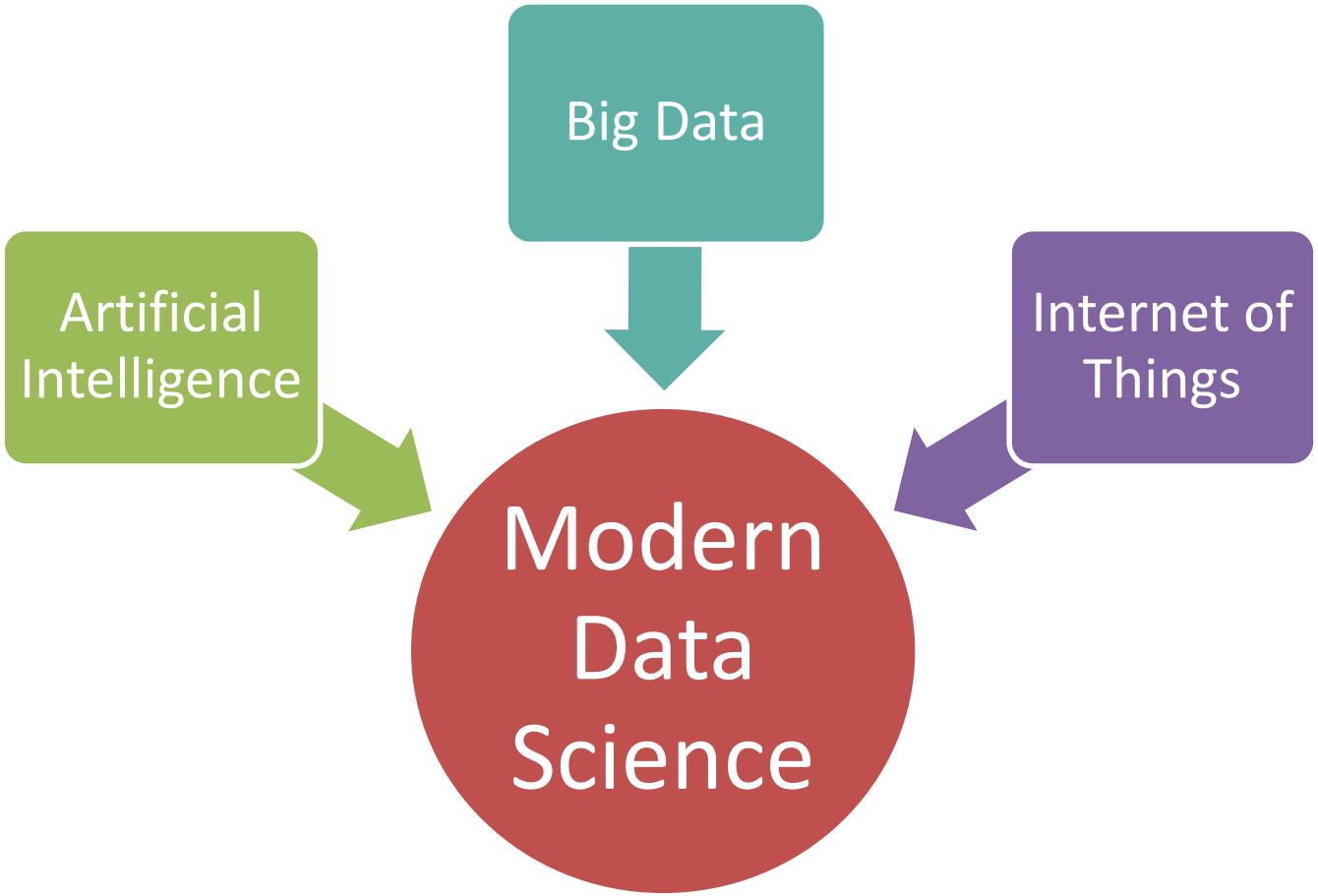
Steve Eglash Personal Introduction

- My role at Stanford
  - Research administration, connect faculty with corporations and government agencies, structure relationships
  - Executive Director of Stanford Data Science Initiative, Artificial Intelligence Lab, Secure Internet of Things Project, Stanford AI Lab-Toyota Center for Artificial Intelligence Research
  - Establish a repository of data for research: Twitter, others
- My background
  - BS UC Berkeley, MS and PhD Stanford, all in Electrical Engineering
  - MIT Lincoln Lab, SDL Inc. (JDSU), Worldview (VC), US DOE, Cyrium Technologies, Stanford

The Data Science Revolution

- The data science revolution
  - More data
  - More sources of data
  - Cheaper storage
  - More powerful compute
  - Advanced algorithms
  - Mobile devices

Convergence



Meta Themes

1. Ability to act on individuals rather than averages
2. Machine learning and deep learning
3. Ability to search and extract information from unstructured, semi-structured, and structured data
4. Opportunity to transition from retrospective analysis to prediction and “what if” scenario planning
5. Statistical and probabilistic approaches
6. Contextual and human-centered interactions
7. Automated decision-making

Data as a Strategic Resource

- Important for policy, societal needs, research / scholarship, business, humanitarian purposes
- Analogous to other valuable resources like human resources, intellectual property, capital equipment, customers, and brand
- Includes proprietary, personal, and public data
- Data and data analytics

Data Commons Concept

- Data commons—an open repository of raw and curated data for research
- Discussion initiated by Google with Stanford
- Stanford as a trusted 3rd party, developer, and administrator
- Stanford addressing barriers
  - Hiring staff in very competitive job market
  - The most valuable data may be proprietary
    - Competitive advantage
    - Liability
  - Funding

Multiple Data Sets

- Various kinds of data: consumer purchases, housing, transportation, e-commerce, utilities & energy, social media, credit & financial, medical transactions
- Value in combining data from diverse sources such as government agencies, private companies, Internet, social media, Internet of Things
- Rising cost of national data collection is not sustainable; use of multiple data sets can reduce need for expensive data collection
- Can mitigate effect of decreasing participation rates in conventional surveys and resultant increased risk of poor quality

Industry Data

- Industry data is different from statistical agency data but can still be used to
  - Improve timeliness of preliminary estimates
  - Provide leading indicators of social change

(from Constance F. Citro, *From multiple modes for surveys to multiple data sources for estimates,* Survey Methodology, vol. 40, pp. 137-161, December 2014)

Issues

- Company proprietary (competition-sensitive) data
- Personal identifiable information
- Skilled data scientists are needed for very large data sets and advanced analytical techniques
  - Data integration, entity resolution, data cleansing
  - Streams of real-time data (accuracy / precision, missing data)
- Domain experts are needed for supervised machine learning and intelligent application

Different Models and Sources

- **Zillow**—aggregator of public data
- **LinkedIn**, **Facebook**, **Twitter**—aggregators of user-supplied data (by agreement)
- **Google**, **Amazon**—proprietary and personal data generated by search, email, etc.
- **Target**
  - Uses federal statistics for store location
  - Adds confidential information on product mix
  - Net result is creation of economic value and IP

Survey Goals and Methodology

- **Goals**
  - Gain understanding of circumstances under which the company might share their data and permit it to be used in conjunction with other data sources in the construction of national statistics
  - Identify issues that would have to be addressed
- **Interviews**—private and anonymized, unless permission is granted for public release
- **Informants**—senior executives, knowledgeable regarding their company's criteria for data policies
- **Scope**—data resources and how they are used within (and if applicable outside) the company

Interview Subjects

Companies from the following industries:

Automobile & transportation	Finance	Retailing
Communications	Insurance	Semiconductor
Consumer electronics	Internet & search	Social media
Data & information technology	Research	Technology

Possible Interview Questions (page 1 of 3)

1. What are your current procedures for permitting research use of your data resources?
   a. How many outside researchers have accessed your data?
   b. Are there standardized procedures for other interested researchers?
   c. What are the benefits to the company of this?
   d. What are the burdens on the company of this?
2. What kinds of agreements do you have with data suppliers in terms of use of the data and confidentiality of the data? How do you protect your confidentiality pledges to data suppliers?
3. How do you handle any IP issues in data use from external analysts?
4. Do you treat access requests for old data (e.g., data collected in the past) differently than requests for contemporaneous data?
5. How do you handle requests for statistical combinations of your data with other data?

Possible Interview Questions (page 2 of 3)

6. Under what circumstances would you consider partnerships with Federal statistical agencies for use of your data in constructing national statistics?
   a. Could you see benefits to your company by doing this?
   b. What burdens do you see for your company doing this?
   c. What are the biggest concerns or risks to your company that you see by doing this?
   d. Are you aware of the strong legal protections that statistical agencies have covering the use and disclosure of data they acquire for statistical purposes? Does this aid your decision?
7. What are the impediments seen in your company for the joint analysis of your data with other data sets?
8. Do you see legal impediments for such sharing?
9. Do you see impediments arising from fear of competitors for such sharing?
10. Do you see impediments from the costs to your company for such sharing?

Possible Interview Questions (page 3 of 3)

11. If there are concerns about providing direct government access to the company's data, would an independent, trusted third party be acceptable as a means of providing access to data for research and statistical purposes in a secure environment for use by federal statistical agencies and academic researchers?
    a. Are you aware of any such arrangements currently?
    b. What are the key factors that would influence your decision as to whether to enter into an arrangement like this?