Understanding Big Data Analytics for Research

Hye-Chung Kum

Texas A&M Health Science Center, Dept. of Health Policy & Management
University of North Carolina at Chapel Hill, Dept. of Computer Science
(kum@tamhsc.edu)
http://research.tamhsc.edu/pinformatics

PUBLIC HEALTH
TEXAS A&M HEALTH SCIENCE CENTER

POPULATION INFORMATICS RESEARCH GROUP

Agenda

- What is Big Data ? What is Data Science ?
- What is Population Informatics & the Social Genome ?

Agenda

- What is Big Data ? What is Data Science ?
- What is Population Informatics & the Social Genome ?

Properties of BIG DATA : 4V

- Volume : lots of data
- Velocity : constantly generating & changing
- Variety : expressed in many ways
- Veracity : lots of errors
- (Value)

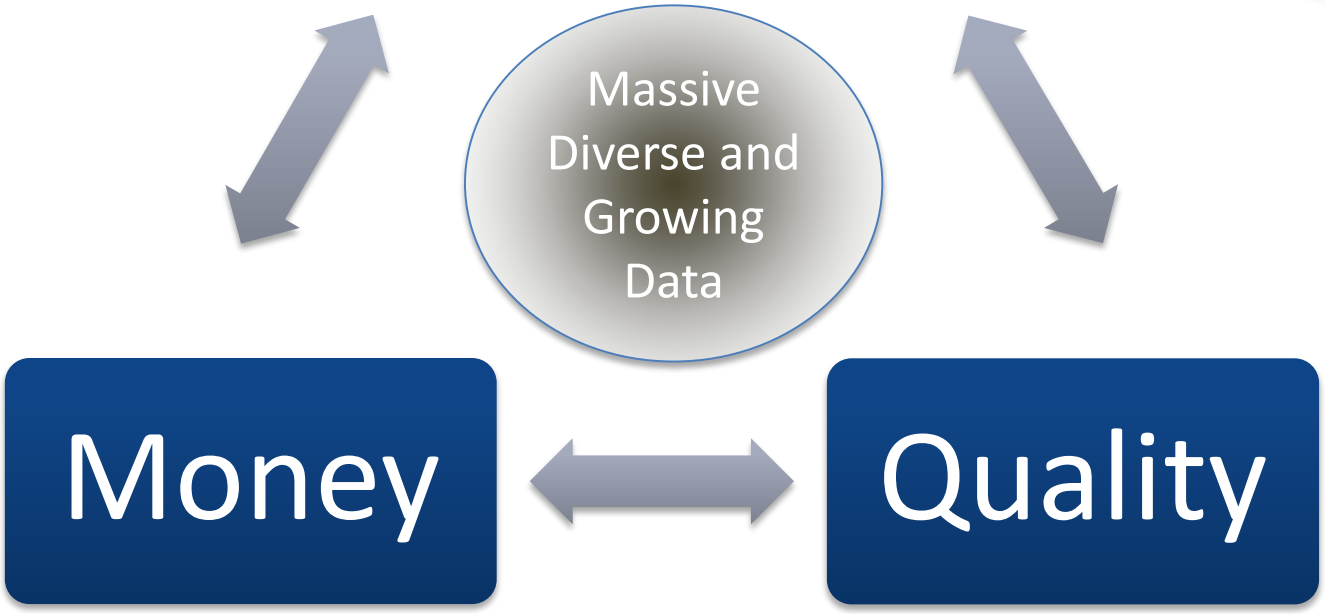
EXAMPLE: the INTERNET!
What do you do to find information/knowledge on the Internet?

Finding actionable information on the Internet

- Figure out your question (refine as you find out more)
  - Descriptive: what is X?
  - Hypothesis: Does X do Y?
- Ontology/Taxonomies: Knowledge representation about the world (synonyms, relationship between concepts)
- Information integration
- Triangulation / validation
- Map: Zoom In / Zoom Out

The Big Data Problem – Nutshelled Michael Franklin (UC Berkley)

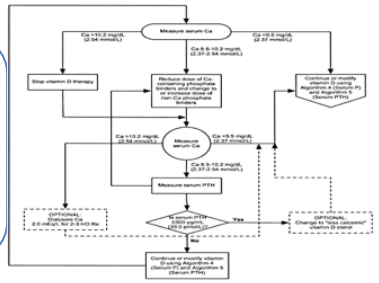
Something's gotta give:



AMPLab: Integrating Three Key Resources

Algorithms

- Machine Learning, Statistical Methods
- Prediction, Business Intelligence



Machines

- Clusters and Clouds
- Warehouse Scale Computing

People

- Crowdsourcing, Human Computation
- Data Scientists, Analysts

NIST Big Data Public Working Group (NBD-PWG)

- NIST: National Institute of Standards and Technology (HIPAA security standard)
- Leaders of activity
  - Wo Chang, NIST
  - Robert Marcus, ET-Strategies
  - Chaitanya Baru, UC San Diego
- http://bigdatawg.nist.gov/home.php

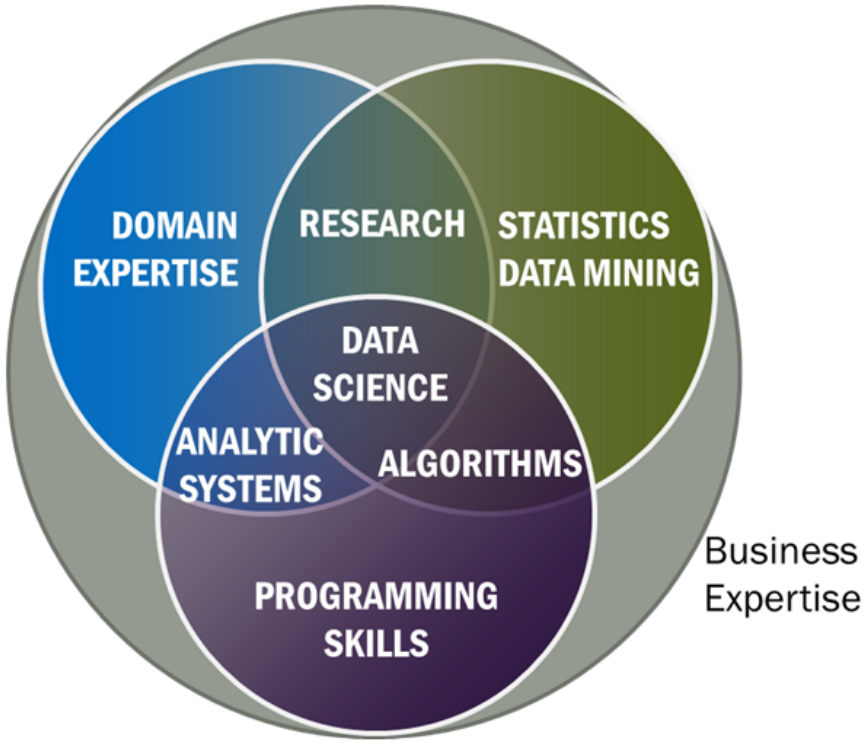
NBD-PWG Subgroups & Co-Chairs

- Requirements and Use Cases SG
  - Geoffrey Fox, Indiana U.; Joe Paiva, VA; Tsegereda Beyene, Cisco
- Definitions and Taxonomies SG
  - Nancy Grady, SAIC; Natasha Balac, SDSC; Eugene Luster, R2AD
- Reference Architecture SG
  - Orit Levin, Microsoft; James Ketner, AT&T; Don Krapohl, Augmented Intelligence
- Security and Privacy SG
  - Arnab Roy, CSA/Fujitsu Nancy Landreville, U. MD Akhil Manchanda, GE
- Technology Roadmap SG
  - Carl Buffington, Vistronix; Dan McClary, Oracle; David Boyd, Data Tactic

Data Science Definition (Big Data less consensus)

- **Data Science** is the extraction of actionable knowledge directly from data through a process of discovery, hypothesis, and analytical hypothesis analysis.
- A **Data Scientist** is a practitioner who has sufficient knowledge of the overlapping regimes of expertise in business needs, domain knowledge, analytical skills and programming expertise to manage the end-to-end scientific method process through each stage in the big data lifecycle.

Big Data refers to digital data volume, velocity and/or variety whose management requires scalability across coupled horizontal resources



NIH: Big Data to Knowledge (BD2K)

- http://bd2k.nih.gov/
- NIH Names Dr. Philip E. Bourne First Associate Director for Data Science
  - December 9, 2013
- NIH commits $24 million annually for Big Data Centers of Excellence
  - July 22, 2013
- Bioinformatics – DNA/RNA data
- http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-14-020.html

NIH Definition

- The term 'Big Data' is meant to capture the opportunities and challenges facing all biomedical researchers in accessing, managing, analyzing, and integrating datasets of diverse data types [e.g., imaging, phenotypic, molecular (including various '–omics'), exposure, health, behavioral, and the many other types of biological and biomedical and behavioral data] that are increasingly larger, more diverse, and more complex, and that exceed the abilities of currently used approaches to manage and analyze effectively.
- **Data Scientist**: Development of a sufficient cadre of researchers skilled in the science of Big Data, in addition to elevating general competencies in data usage and analysis across the behavioral research workforce.

NIH: 4 Big Data Issues

- Data Compression/Reduction
  - Data Compression refers to the algorithm-based conversion of large data sets into alternative representations that require less space in memory. Data Reduction refers to the reduction of data volume via the systematic removal of unnecessary data bulk.
- Data Visualization
  - Data Visualization refers broadly to human-centric data representation that aids information presentation, exploration, and manipulation. This is typically performed via the use of visual and graphical techniques; however, these can be augmented with sound and other sensory cues to create deeper experiences.
  - SEE the DATA: Zoom In / Zoom Out (mapquest)
- Data Provenance (replicable science – tractable processes)
  - Data Provenance refers to the chronology or record of transfer, use, and alteration of data that documents the reverse path from a particular set of data back to the initial creation of a source dataset. Provenance of digital scientific data is useful for determining attribution, enabling data citation, identifying relationships between objects, tracking back differences in similar results, guaranteeing the reliability of the data, and to allow researchers to determine whether a particular dataset can be used in their research by providing lineage information about the data.
  - Good programming practice
- Data Wrangling (data cleaning/integration)
  - Data Wrangling is a term that is applied to activities that make data more usable by changing their form but not their meaning. Data wrangling may involve reformatting data, mapping data from one data model to another, and/or converting data into more consumable forms.
  - http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

NIH and Biomedical "Big Data"

http://acd.od.nih.gov/Big-Data-to-Knowledge-Initiative.pdf

COMMENT

A vision for data science

To get the best out of big data, funding agencies should develop shared tools for optimizing discovery and train a new breed of researchers, says **Chris A. Mattmann**.

PEOPLE POWER

To solve big-data challenges, researchers need skills in both science and computing — a combination that is still all too rare. A new breed of 'data scientist' is necessary.

***Nature* 2013**

Thomas Davenport
Competing on Analytics

- Skill set for good data scientists
  - IT & Programming skills
  - Statistical skills
  - Business skills:
    - Understand pros/cons of decisions & actions
    - Communication skills
    - Excel / PowerPoint
  - Intense curiosity: the most important skill or trait. "a desire to go beyond the surface of a problem, find the question at its heart, and distill them into a very clear set of hypothesis that can be tested"

Data science teams need people with the skills and curiosity to ask the big questions (oreilly)

- Technical expertise: the best data scientists typically have deep expertise in some scientific discipline.
- Curiosity: a desire to go beneath the surface and discover and distill a problem down into a very clear set of hypotheses that can be tested.
- Storytelling: the ability to use data to tell a story and to be able to communicate it effectively.
- Cleverness: the ability to look at a problem in different, creative ways.
- Health is a very important domain
  - Team lead: good questions, good interpretation & implications
- http://radar.oreilly.com/2011/09/building-data-science-teams.html

Thank you!

Questions?

Population Informatics Research Group

http://research.tamhsc.edu/pinformatics/

http://pinformatics.web.unc.edu/

PUBLIC HEALTH
TEXAS A&M HEALTH SCIENCE CENTER

POPULATION INFORMATICS RESEARCH GROUP