

# Population Informatics: Applying Data Science to Big Data about People to Advance Population Health

Hye-Chung Kum, Associate Professor (kum@tamu.edu)

Population Informatics Lab (<https://pinformatics.org/>)

Department of Health Policy and Management, School of Public Health

Department of Computer Science and Engineering

Department of Industrial and Systems Engineering

The Center for Remote Health Technologies and Systems (CRHTS)

Texas A&M University



4/14/2020

2

2

## Primary Methodology: Data Science (KDD)

3

NIST Big Data

POPULATION INFORMATICS

## 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

9/29/13

IEEE BigData Overview October 9 2013

8

4

## Bioinformatics

### Apply Data Science to Human Genome Data

POPULATION INFORMATICS

5

## Population informatics

### Apply Data Science to Social Genome Data

Studies of society (groups of people)

- Social, Behavior, Economic sciences
- Health sciences (population health)

+

Kum, H.C., Krishnamurthy A., Machanavajjhala A., and Ahalt S. Social Genome: Putting Big Data to Work for Population Informatics. *IEEE Computer Special Outlook Issue*. pp 56-63. Jan 2014

6

## Population Informatics: The systematic study of populations

via **secondary analysis** of massive data collections (“big data”) about people

**Actionable Policy and Practice**

**Transformational Knowledge**

**Information**  
Broad new questions

**Methods**  
Datamining & Statistical methods

**Secure Federated Data Infrastructure**

**Social Genome Data Library**


**Data Savvy Managers (Decision Makers)**  
Data Based Answers to Real world Problems

**Data Intensive Domain Scientists**  
Frame Real World Questions to Tractable Questions


**Domain Knowledgeable Computer Scientists**

7


## Knowledge Discovery & Data mining (KDD) = Data Science




**Big Data : impossible to keep organized**



**KDD  
Clean, Merge, Reprocess**






Human consumable, valid, novel, potentially useful,  
& ultimately understandable information

Fayyad, U. M. Pietsky-Shapiro, G. Smyth, P. 1996. From Data Mining to Knowledge Discovery: An Overview. In, *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT press, Cambridge Massachusetes.

8

## KDD Process



Operational Data

EDW

Task Specific Data

Results

Information Presentation

- Data cleaning & integration
- Feature Selection (what vars?)
- Analysis / Datamining
- Validation / Evaluation
- Action

9

## Data Wrangling



The New York Times | <http://nyti.ms/1mZywnG>

TECHNOLOGY

### For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR | AUG. 17, 2014

- Data Wrangling is a term that is applied to activities that make data more usable by changing their form but not their meaning
  - reformatting data: MDY vs YMD
  - mapping data from one data model to another: ICD9 vs CPT code
  - and/or converting data into more consumable forms: to graphs
- 30-80% of the work in using big data
- Once raw data is “wrangled” into the correct analytic data
  - Running statistics models are fairly simple and similar to what you do traditionally
  - There are new methods but, usually requires a LOT of data

10

## Thomas Davenport: *Competing on Analytics*



- Skill set for good data scientists
  - IT & Programming skills: Very basic programming concepts in SAS
    - <https://pinformatics.tamhsc.edu/phpm672/>
  - 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”

11

## 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>

12

## What is data science ? Hye-Chung Kum



- **Measurement (=features):** Smart/clever counting of real things (meaningful to people) in the digital data
- **Information generation:** Then modeling using those measures (features)
- **Delivery of information:** Storytelling with data
  - Careful: Abuse of data
- **Develop agile data pipeline** for timely processing that can be iteratively updated to track the dynamic ever changing real world
- Skills
  - Curiosity
  - Tenacity
  - Good judgement & critical thinking
- For more information, check out lab website on data science menu
  - Videos I show in class are there.

13

## Regional Record Linkage Centers Population Data Linkage Network



### Canada

- Alberta Centre for Child, Family and Community Research, The Child and Youth Data Lab (CYDL)
- The Canadian Institute for Health Information (CIHI)
- Population Data BC
- Health Services Analysis Unit, Alberta Health Services
- Institute for Clinical Evaluative Sciences-- Ontario
- Manitoba Centre for Health Policy
- Statistics Canada

### Germany

- German Record Linkage Center

### New Zealand

- Department of Population Health, University of Otago, Christchurch

### Australia

- Australian Institute of Health and Welfare
- Australian Bureau of Statistics
- Centre for Health Record Linkage
- Health LinQ
- SA NT DataLink
- Western Australia Data Linkage Branch
- Population Health Research Network
- The Centre for Data Linkage
- The Tasmanian Data Linkage Unit (TDLU)

### United Kingdom

- Avon Longitudinal Study of Parents and Children (ALSPAC)
- Oxford Record Linkage Group
- Information Services Division, Scotland
- SAIL Databank
- UK Biobank

14

## Case Study 1: Research LEHD: US Census Bureau



- Vertically integrated in one domain
  - Wage : UI (Unemployment Insurance) Data
- Decision support : LEHD website
- By building an integrated data that “permits the real world of the US economy to be interrogated by the models of unemployment dynamics” Peter Diamond, Dale Mortense, and Christopher Pissarides shared the Nobel Prize in economics 2010 (David Warsh, economicprinciple.com)
  - Modeling the churning behavior in the labor market

15