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

Hye-Chung Kum

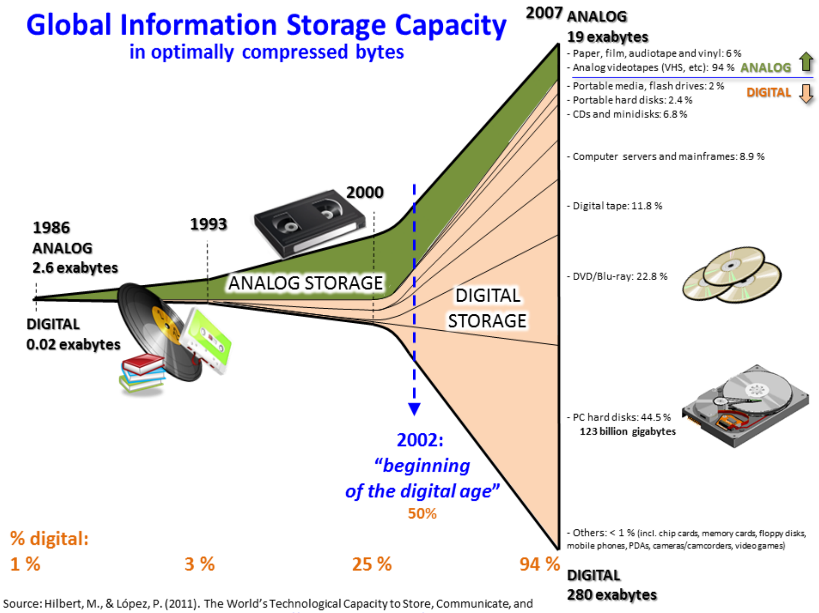
- PhD in computer science (data mining, sequential pattern mining)
- Minor: MSW (policy & management)
- Primary appointment: School of Public Health, HPM (18+ years: CS, SW, HPM, ISEN)
- Joint appointments in Computer Science & Industrial Systems
- Population Informatics Research Group, Texas A&M University
  - Multidisciplinary: CS, HSR, health informatics, SW, sociology, ELSI
  - PCORI: Privacy Preserving Interactive Record Linkage (PPIRL)
  - NSF: A Benchmark Data Linkage Repository (DLRep)
  - TX HHSC: 1115 Medicaid Waiver Evaluation
  - NC DHHS: Management Assistance

Agenda

- Establish the field: Definition
  - What is Data Science ?
  - What is Social Genome ?
  - What is Population Informatics ?
- Traditional social science vs Data Science
- Case Studies

Global Information Storage Capacity

in optimally compressed bytes



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

The Digital Society (~2002)

- Most everything we do are recorded digitally

The Cost of the Digital Society

- There is **no turning back !**
- Personal information is already being used
  - Marketing: Target
  - Campaigning: Cambridge Analytica
  - Intelligence: Edward Snowden

"Facebook users in the US admit to taking steps to reframe their relationship with the social media platform, according to Pew. All told, Pew found that 74 percent of Facebook users say they have taken … actions in the past year. Around 44 percent of younger users between the ages of 18 to 29 said they deleted the Facebook app"

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Snowden Claims NSA Knocked All of Syria's Internet Offline

Why not reap the benefits too ?

- The ability to answer questions about human populations in near real time using distributed datasets that are large, complex, and diverse has the potential to transform social, behavioral, economic, and health sciences (SBEH)
- The results could lead to more informed and effective policy decisions and allocations of public resources
  - What is the long term impact of moving to managed care ?
  - What effect does teacher pay in middle school have on college grades?
- The answers could easily be derived from relevant data sets

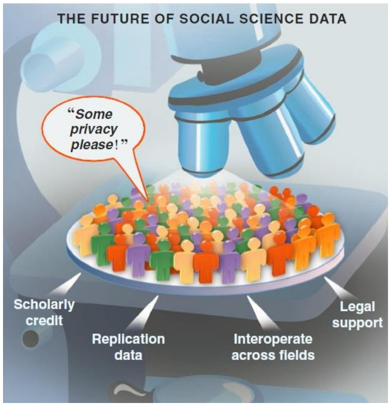
HOW?

Population Informatics = Population Data Science

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

How do we reap the benefits too ?

- Overarching question:
  - How can we use the abundance of existing digital data about people, aka big data, (e.g. government administrative data, electronic health records)
  - to support accurate evidence based decisions for policy, management, legislation, evaluation, and research
  - while protecting the confidentiality of individual subjects of the data?



Source: Gary King. Ensuring the Data-Rich Future of the Social Sciences, *Science*, vol 331, 2011, pp 719-721.

Primary Data: Big Data about People

Properties of BIG DATA : 4V

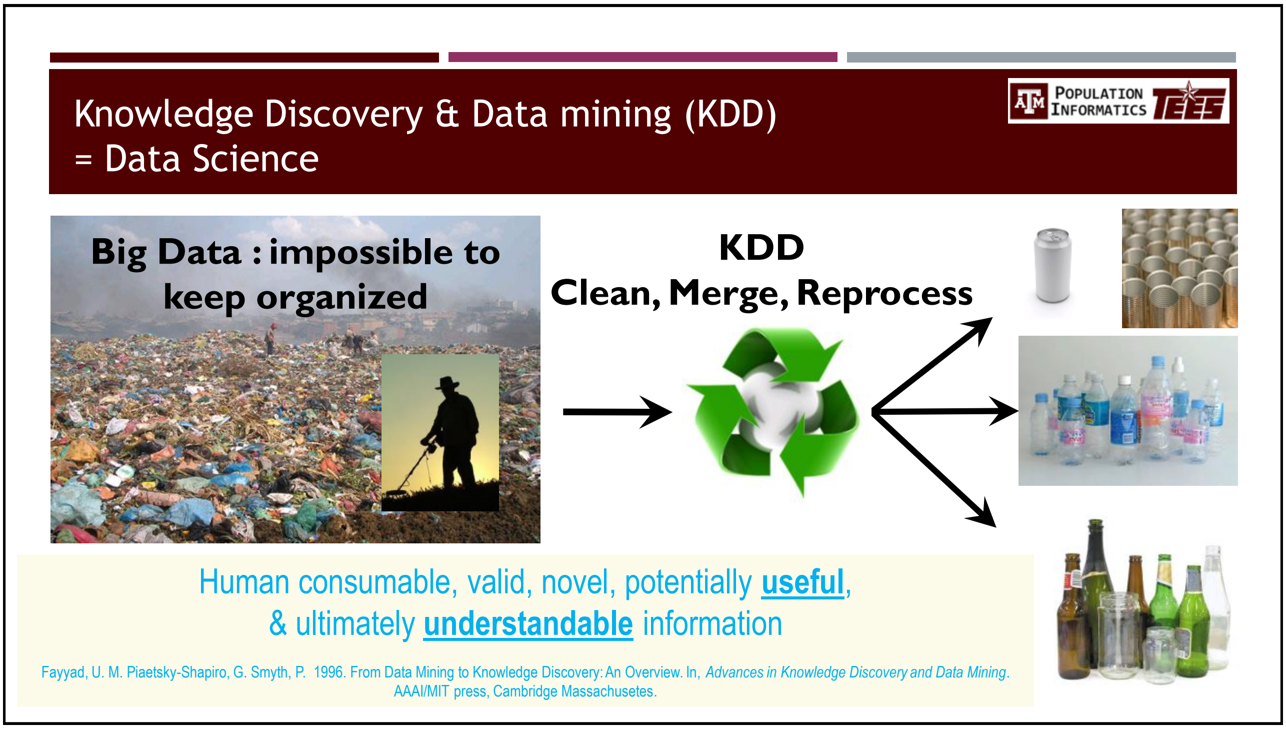
- Volume : lots of data
- Velocity : constantly generating & changing
- Variety : expressed in many ways
- Veracity : lots of errors
- (Value)

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

Primary Methodology: Data Science (KDD)



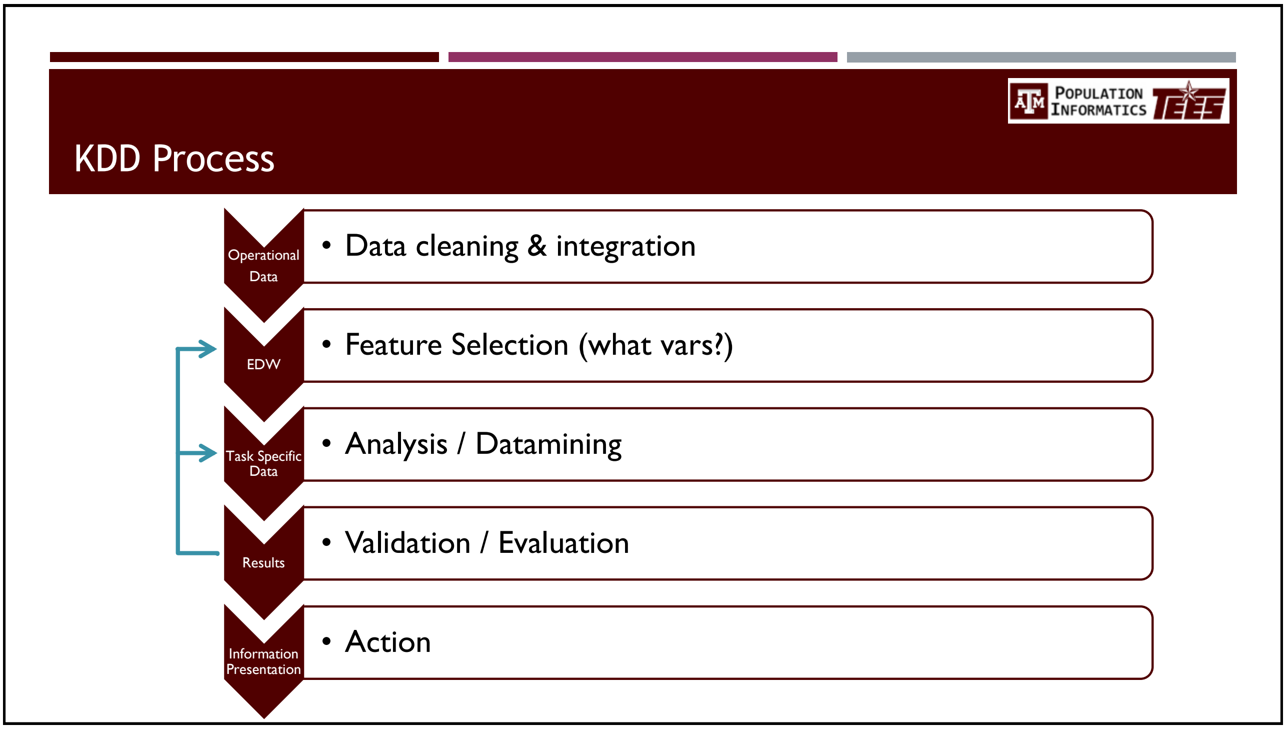
Knowledge Discovery & Data mining (KDD) = Data Science

Big Data : impossible to keep organized

KDD
Clean, Merge, Reprocess

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

Fayyad, U. M. Piaetsky-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.



KDD Process

Operational Data
- Data cleaning & integration

EDW
- Feature Selection (what vars?)

Task Specific Data
- Analysis / Datamining

Results
- Validation / Evaluation

Information Presentation
- Action

What is data science ?
Hye-Chung Kum

POPULATION INFORMATICS TEES

- 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
- Develop agile data pipeline for timely processing that can be iteratively updated to track the dynamic ever changing real world

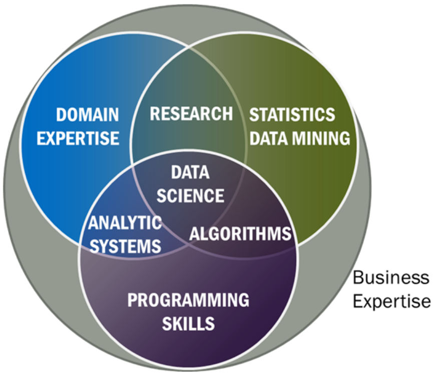
NIST Big Data

POPULATION INFORMATICS TEES

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

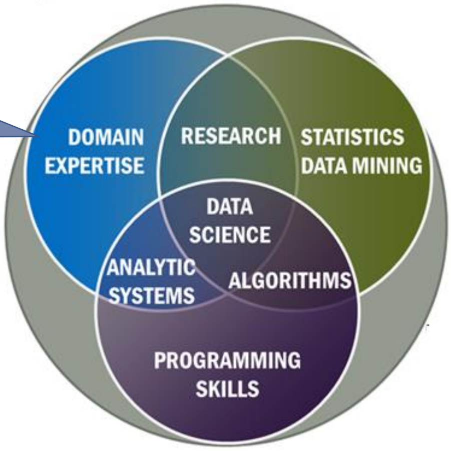


Bioinformatics
Apply Data Science to Human Genome Data

Biology

Human Genome Data

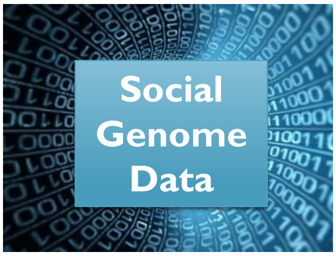
+



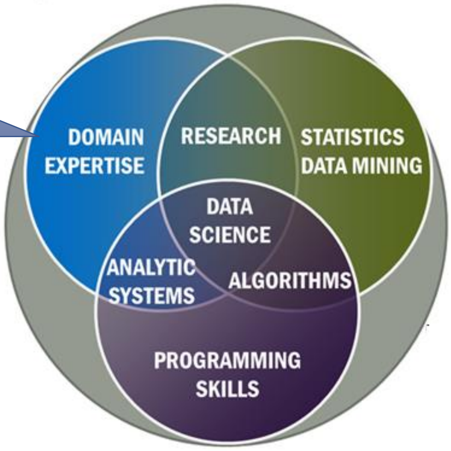
Population informatics
Apply Data Science to Social Genome Data

Studies of society (groups of people)

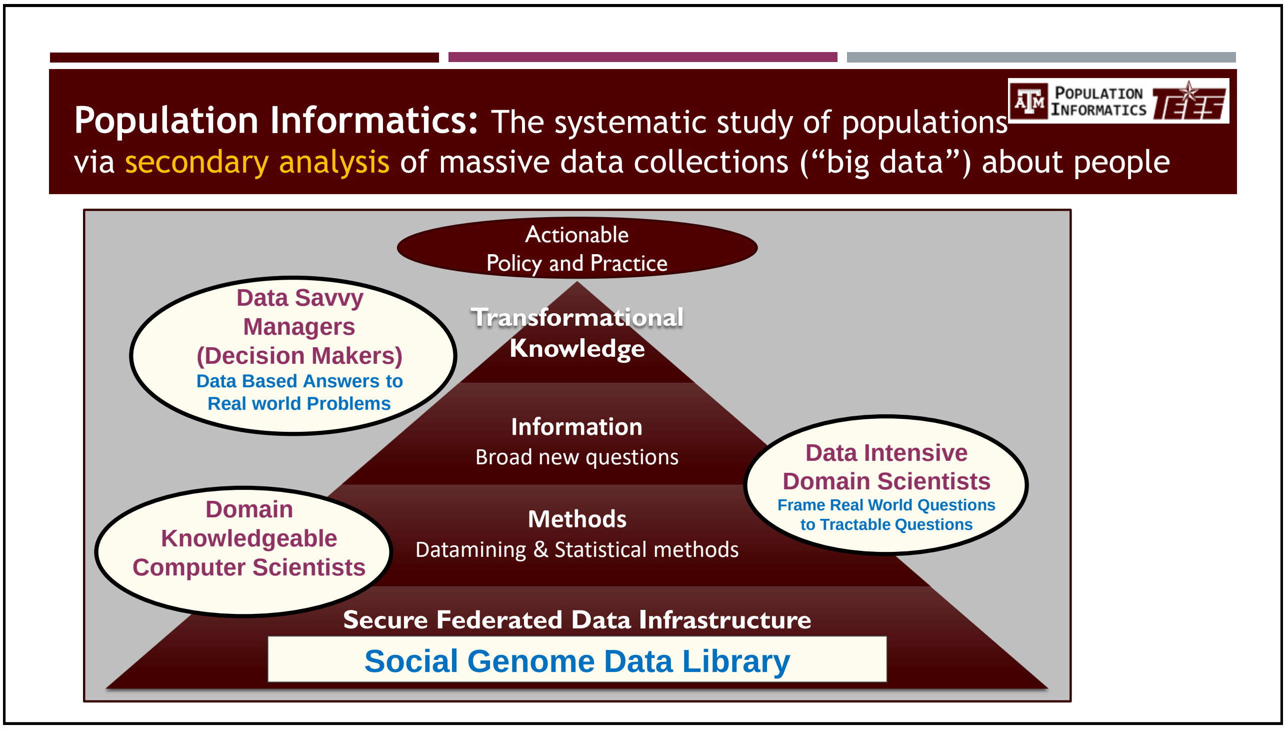
- Social, Behavior, Economic sciences
- Health sciences (population health)


Social Genome Data

+



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



Population Informatics: The systematic study of populations via secondary analysis of massive data collections ("big data") about people

POPULATION INFORMATICS
TEES
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

Other useful perspectives

9/7/2018

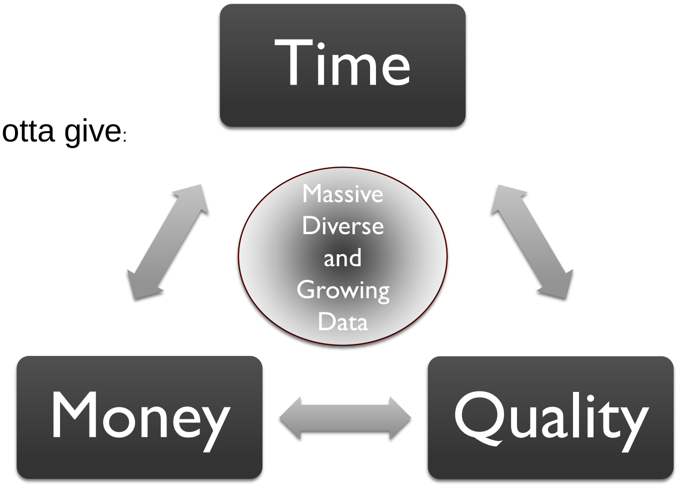
20

New Era in Science : Big Data Science

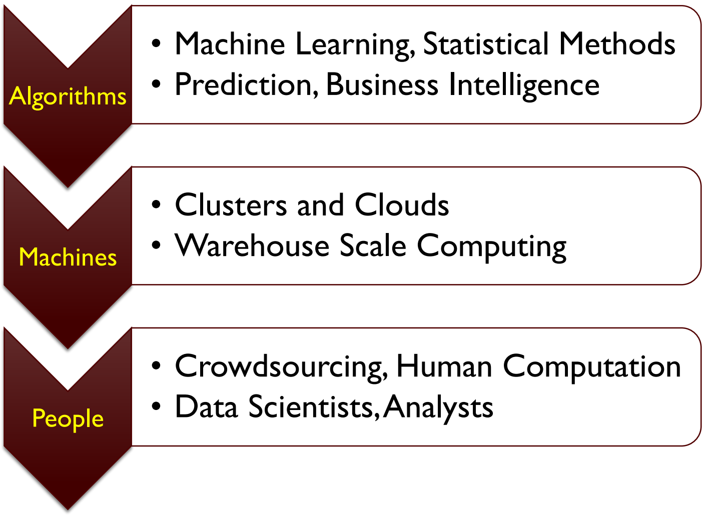
- Data is the new raw material of business: an economic input almost on par with capital and labor.(Microsoft's Craig Mundie)
- Those who can harness the power of data will lead the next century and drive innovation in commerce, scientific discovery, healthcare, finance, energy, government, and countless other fields.
- Students who learn to be a data science will be in high demand.

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

Something's gotta give:



AMPLab: Integrating Three Key Resources Michael Franklin (UC Berkley)

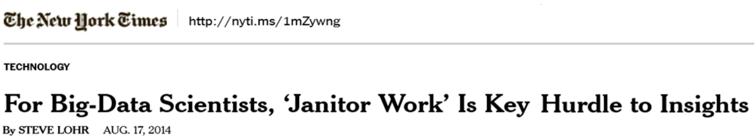


Algorithms
- Machine Learning, Statistical Methods
- Prediction, Business Intelligence

Machines
- Clusters and Clouds
- Warehouse Scale Computing

People
- Crowdsourcing, Human Computation
- Data Scientists, Analysts

Data Wrangling



The New York Times http://nyti.ms/1mZywng

TECHNOLOGY

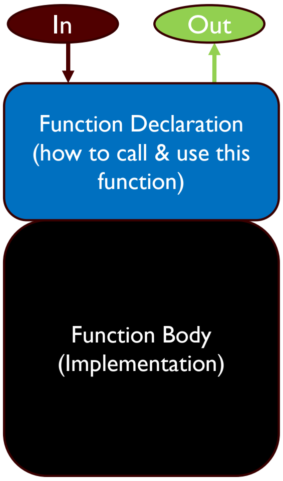
For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

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

Programming

- Code reuse
  - Solve a problem once
  - Reuse your solution for similar problems
- Avoids repetitive typing
  - Consistency
  - Reduce Mistakes
  - Maintenance
    - Easier to fix one function than find and fix all locations of cut & paste code.
- Encapsulation
  - Black box programming
  - Hides internal details of algorithm from users
  - Users typically only care about using the function to get results.
  - Isolates computations, protects variables
    - Interaction through arguments
  - Separates interface and implementation
    - Interface: what a function does
    - Implementation: how a function does it


In
Out
Function Declaration (how to call & use this function)
Function Body (Implementation)

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”

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

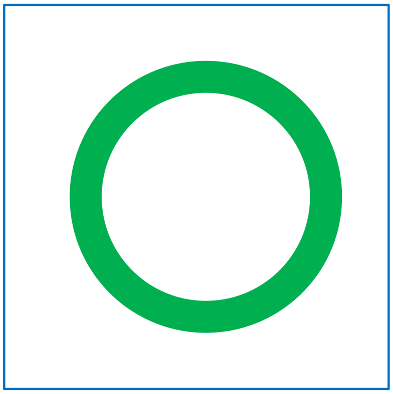
Traditional social science vs Data Science

Is it new ?

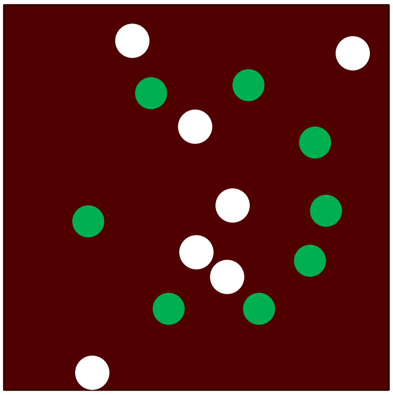
Inflation:
Traditional social science vs Data Science

- Consumer Price Index (CPI)
  - Representative basket of goods and services purchased for consumption by urban households (monthly)
  - This index value has been calculated every year since 1913
  - Bureau of Labor Statistics
- Billion Prices Project : MIT
  - The Billion Prices Project is an academic initiative that uses prices collected from hundreds of online retailers around the world on a daily basis to conduct economic research.
  - Pricing Behavior, Daily Inflation and Asset Prices, Pass-Through (price and exchange rate and international rate), Green Markups (premium for green prod.)

What is the shape of the green line?

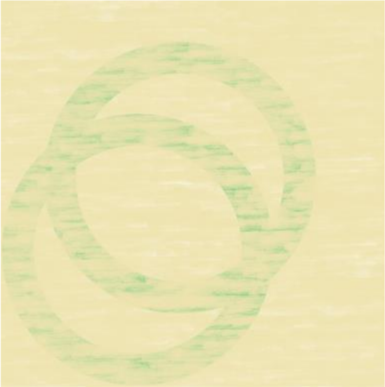


Traditional Science :
Start with nothing – collect data well

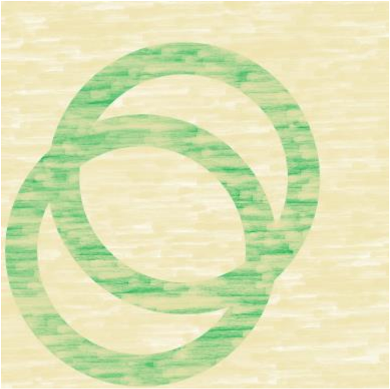


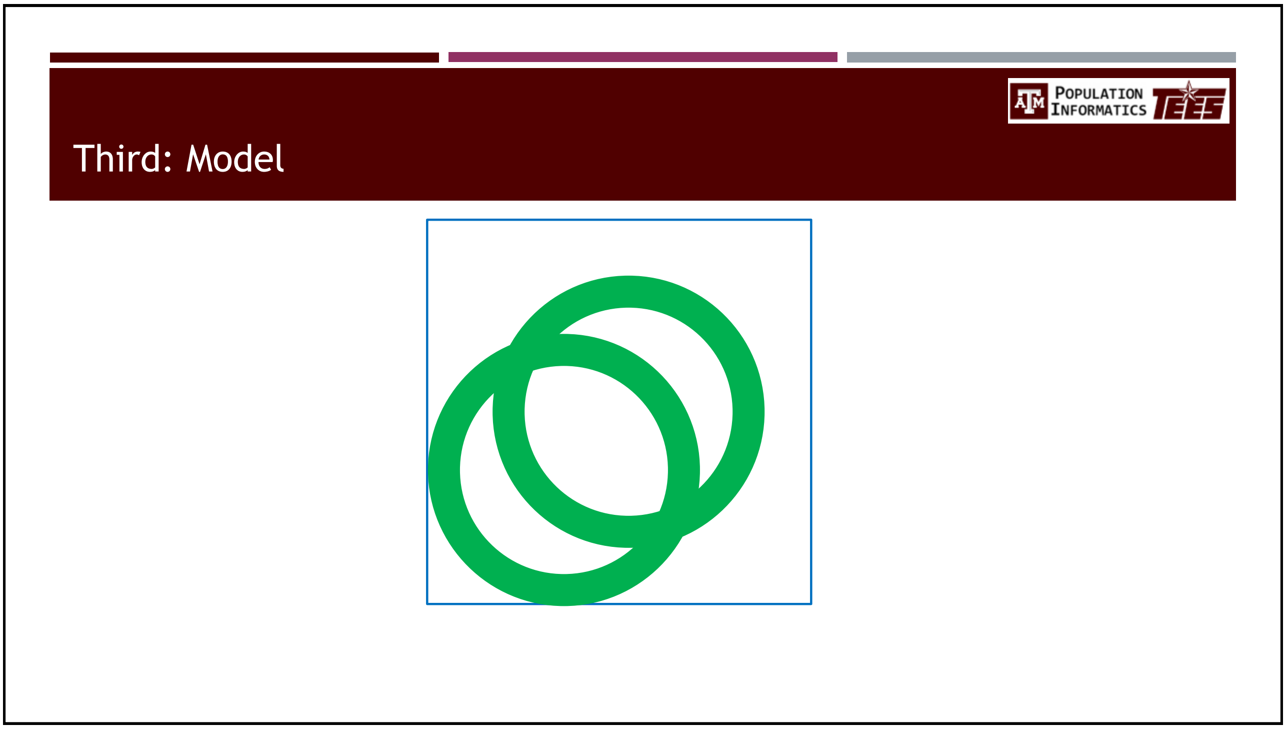
Data Science: EVERYTHING

First, separate out only the relevant data

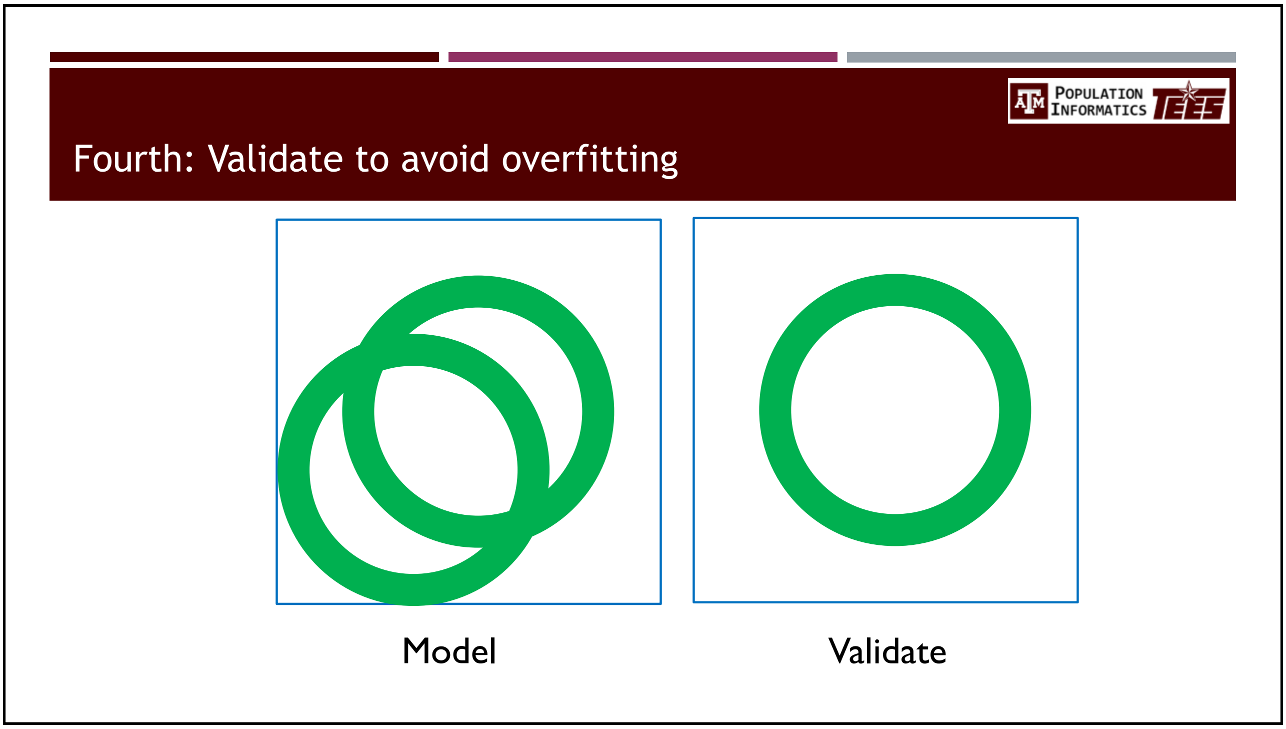


Second, clean noise as much as possible





Third: Model

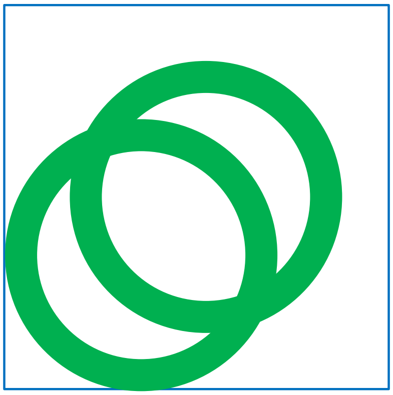


Fourth: Validate to avoid overfitting

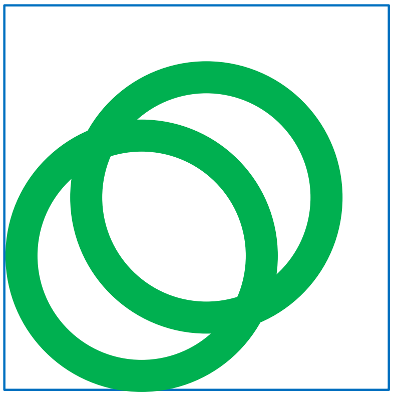
Model

Validate

Sometimes models differ between the two approaches. Why ?



Model



Validate

Comparison

	Traditional Science	Data Science
Common	Use statistics to model from the data points (number of datapoints does matter)	
Focus	Usually more about causation	STRONG correlation
Measurement	• Mostly based on theory (deductive) decide what to measure - green only • Each data collection point is very expensive • With out seeing the other colors • Slow iterative process to discovery	• Iterate between deductive (theory based top down) and inductive (data based bottom up) reasoning to figure out what to measure : can see the other colors, so use existing data to compare • Data is almost free • Different from fishing for results or atheoretical • Faster iteration to discovery
Measurement Error	Reduce/minimize by designing experiments well	Know what it is, adjust for it as best as possible. Usually use data that exist
Bias	Random Points, oversampling	Validation is very important: be careful not to over fit to the data, Know your bias (e.g., lead/length bias)
Main issue	Are there enough points to get the full picture?	Is the data clean enough? Is the data representative ? Sensitivity analysis

Validation

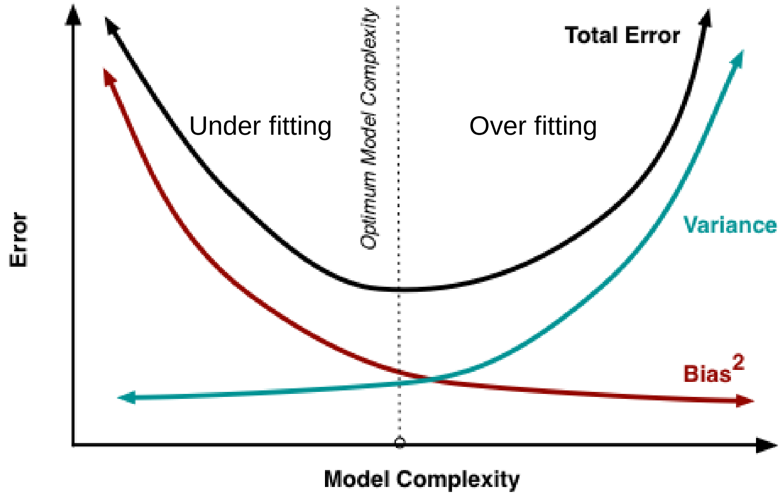
Training Data

Validate Data

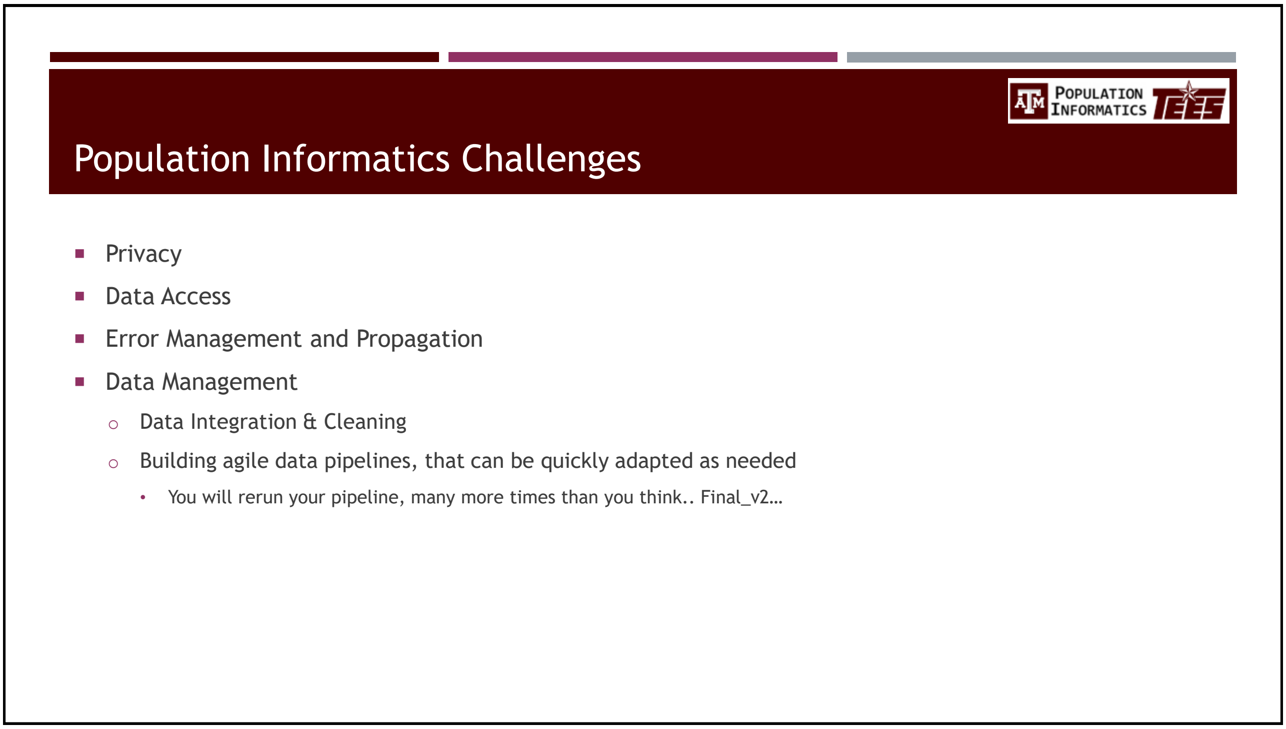
Test Data

- If you have enough data random partition into train/validate/test
- IF what you are trying to identify is a small part of the data, over sample in training data
- If you do not have enough data
  - Cross validation
- Gives confidence to the results when p-value is less meaningful

Validation
http://scott.fortmann-roe.com/docs/BiasVariance.html



Concluding Thoughts



Population Informatics Challenges

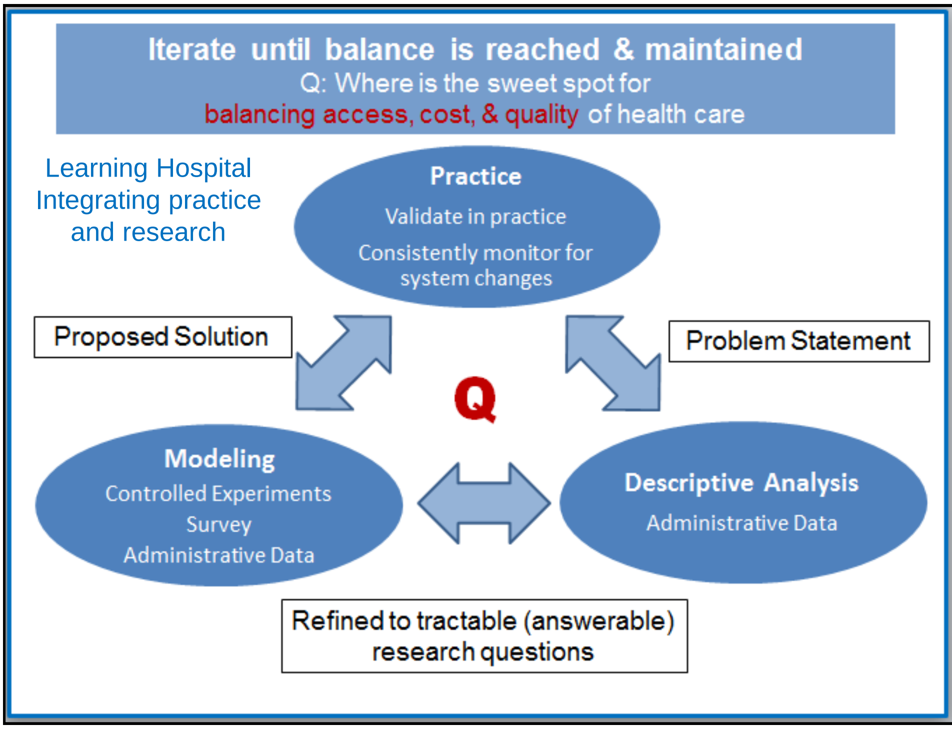
- Privacy
- Data Access
- Error Management and Propagation
- Data Management
  - Data Integration & Cleaning
  - Building agile data pipelines, that can be quickly adapted as needed
    - You will rerun your pipeline, many more times than you think.. Final_v2…

Privacy-by-Design

- A different perspective on privacy and research using personal data
- Personal Data is Delicate/Hazardous/Valuable
- Important to have proper systems in place that give protection but allow for continued research in a safe manner
- All hazardous material need standards
  - Safe environments to handle them in : closed computer server system lab
  - Proper handling procedures : what software are allowed to run on the data
  - Safe containers to store them : DB system

Closing Thoughts

- Overarching question: **How can we use the abundance of existing digital data, aka big data, (e.g. government administrative data, electronic health records) to support accurate evidence based decisions** for policy, management, legislation, evaluation, and research while protecting the confidentiality of individual subjects of the data?
- Preferred approaches: **Data Science** - To build efficient and effective human computer hybrid processes and systems to clean, integrate, and extract actionable information from raw chaotic data and deliver accurate information in a timely secure manner to decision makers (e.g. researchers, policy makers, mangers, clinicians).
- Primary data: **Social Genome data - person level data**, usually identifiable (so we can accurately integrate diverse data) but partitioned data
- Primary issues: **Privacy (safe data access, code of conduct), data integration, error management**,
  - Velocity, variety, veracity, volume (lots of SMALL datasets)



Iterate until balance is reached & maintained

Q: Where is the sweet spot for balancing access, cost, & quality of health care

Learning Hospital
Integrating practice and research

Practice
Validate in practice
Consistently monitor for system changes

Proposed Solution

Problem Statement

Q

Modeling
Controlled Experiments
Survey
Administrative Data

Descriptive Analysis
Administrative Data

Refined to tractable (answerable) research questions

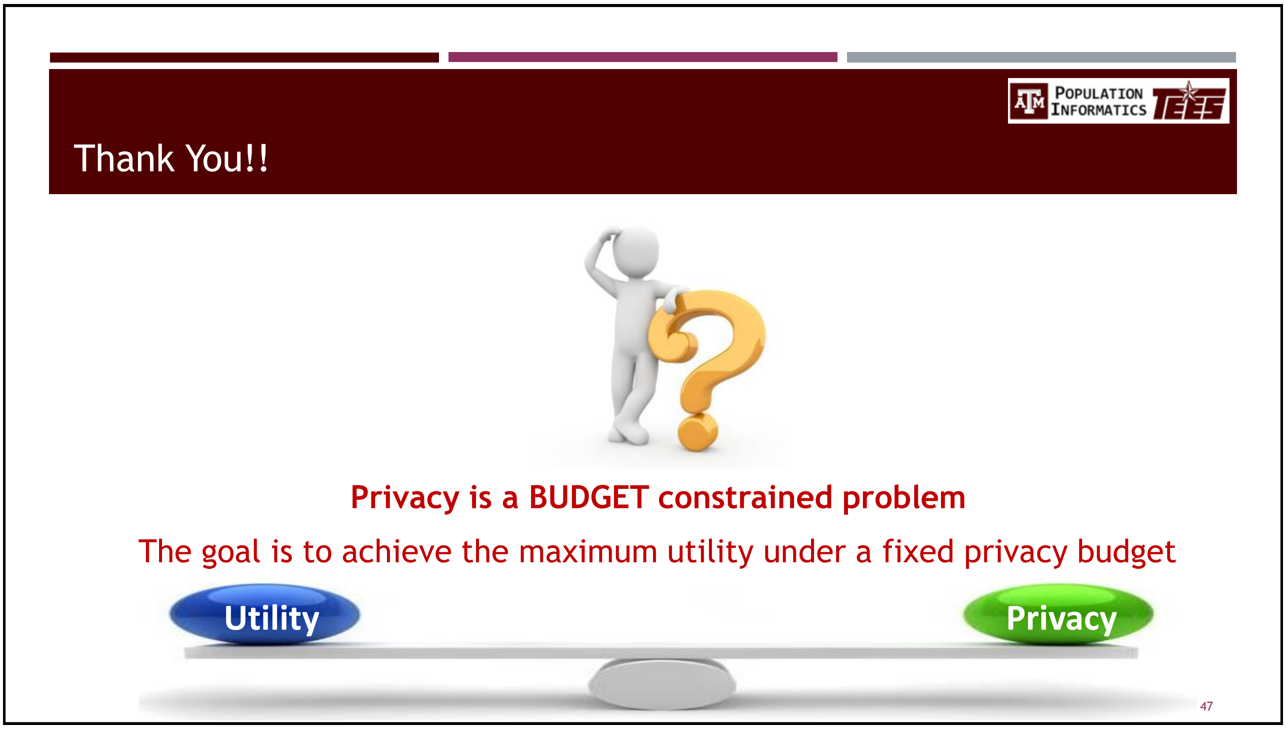
POPULATION INFORMATICS

A data-driven organization acquires, processes, and leverages data in a timely fashion to create efficiencies, iterate on and develop new products, services, and navigate the competitive landscape.

POPULATION INFORMATICS

Learning Hospitals

- Dr. Lenard Berry
- Toussaint JS, Berry LL. The promise of Lean in health care. In Mayo clinic proceedings 2013 Jan 1 (Vol. 88, No. 1, pp. 74-82). Elsevier.
- Cancer care



Thank You!!

Privacy is a BUDGET constrained problem

The goal is to achieve the maximum utility under a fixed privacy budget

47