Population Informatics: Applying Data Science To Advance the Health and Welfare of Populations

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

Who am I

- PhD in computer science (data mining)
- MSW (policy & management)
- 11 years: Appointment in CS, SW, HPM
- RESEARCH FOCUS
  - **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? This question focuses on the data science of using massive secondary datasets, a step before the traditional statistical methods can be applied to the data to address specific questions to improve public health.
  - **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).
- Population Informatics Research Group (TAMU & UNC)

Agenda

- What is Big Data ? What is Data Science ?
- What is Population Informatics & the Social Genome ?
- How is Data Science different from traditional science?
- Doing research with Big Data

PUBLIC HEALTH
TEXAS A&M HEALTH SCIENCE CENTER
POPULATION INFORMATICS RESEARCH GROUP

Optional Agenda

- Hands on
  - Data integration
  - Programming/Debugging
- Privacy
- How does a computer work?

Agenda

- What is Big Data ? What is Data Science ?
- What is Population Informatics & the Social Genome ?
- How is Data Science different from traditional science?
- Doing research with Big Data

Properties of BIG DATA : 4V

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

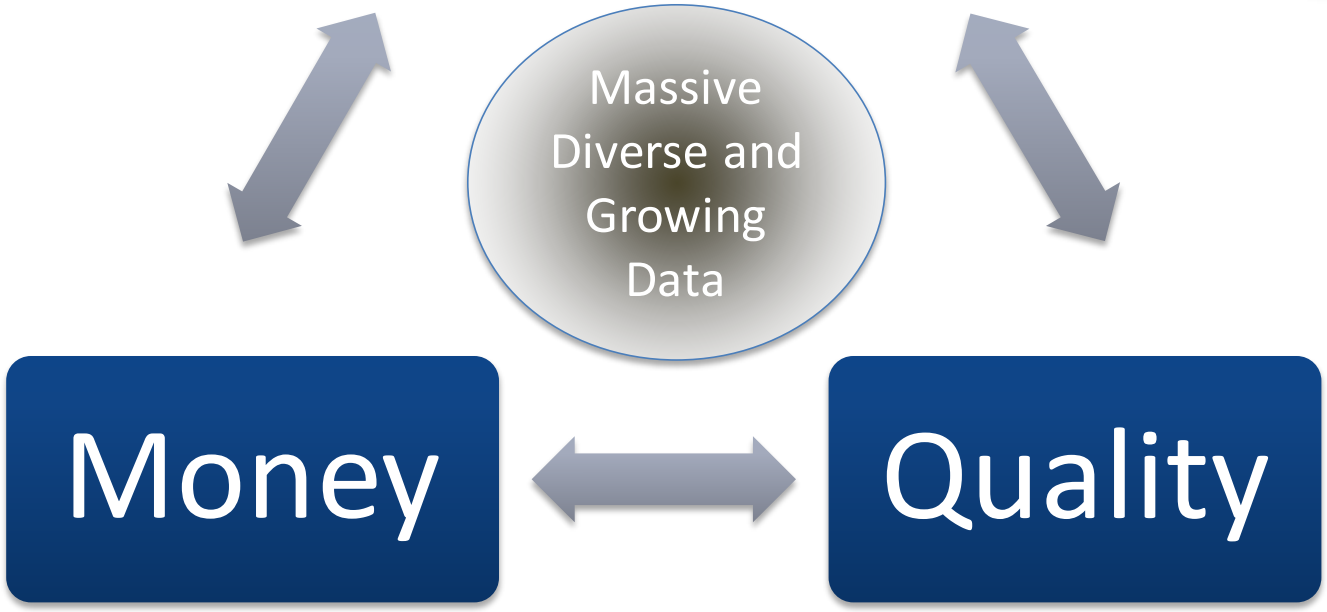
EXAMPLE: the INTERNET!
What do you do to find information/knowledge on the Internet?

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

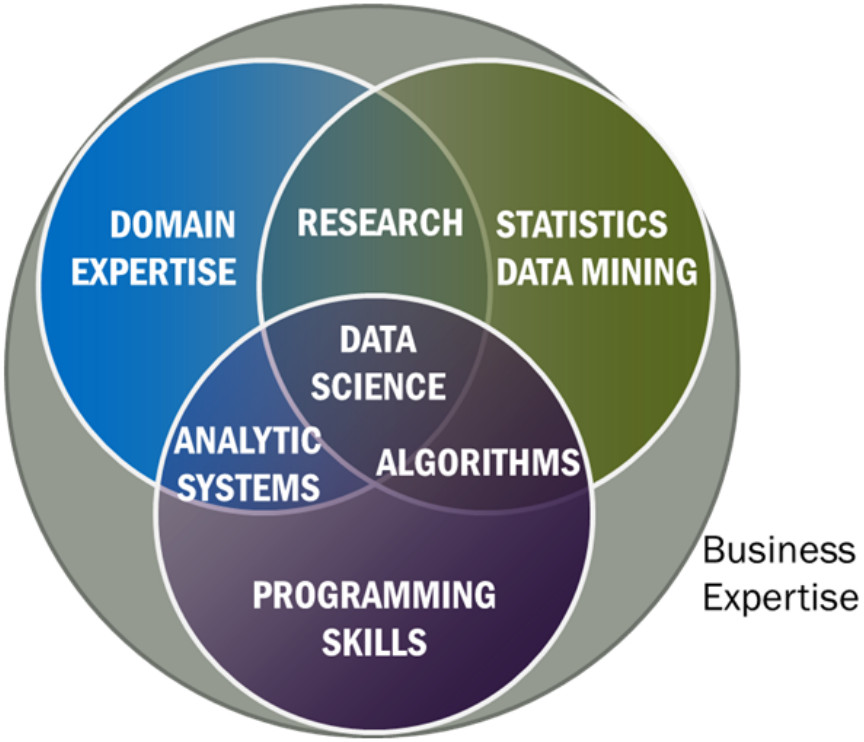
NBD-PWG Subgroups & Co-Chairs

- 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.

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**

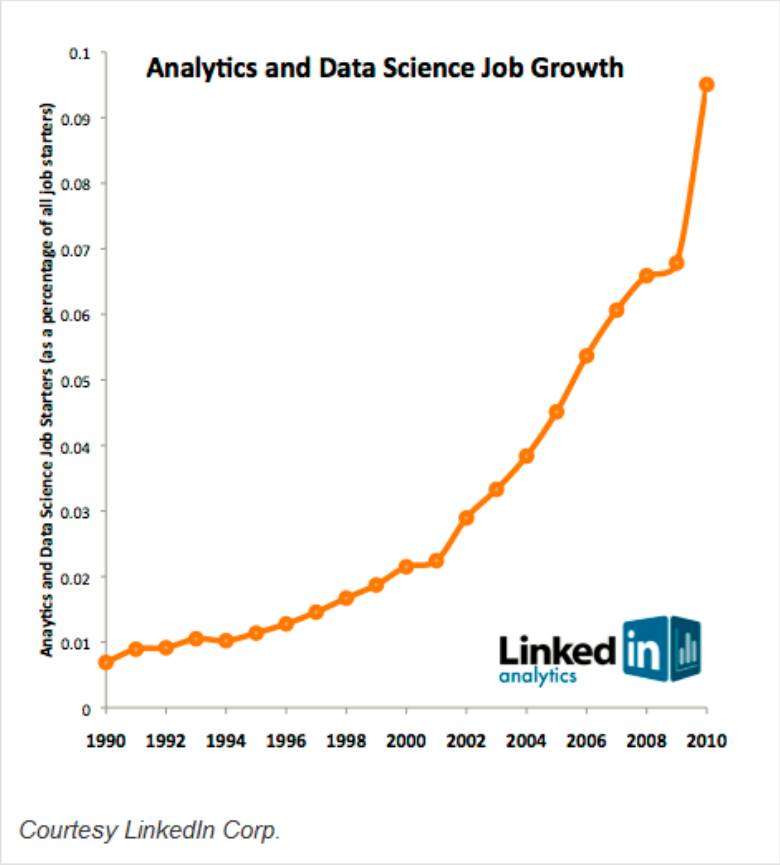
Thomas Davenport
Competing on Analytics

- 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

Job market of data scientists



Courtesy LinkedIn Corp.

- statisticians will be the next sexy job
  - Google Chief Economist Hal Varian
- shortage of 190,000 data scientists by the year 2019
  - McKinsey Global Institute

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.

High International Interest Doing Good Research with Big Data

- Figuring how to do good research with Big Data
- EUDAT (Oct 2011): EU
  - €16-million initiative to develop an international data management infrastructure over 3 years
- White house (Mar 2012) : US
  - a national effort to fund Big Data research across the federal agencies including NSF, NIH, DOD (Dept of Defense), and DOE (Dept of Energy)

International Population Health Informatics Research

- US : LEHD (Census Bureau) – 2010 Nobel Prize in economics
- Australia & New Zealand
  - National Centre for Epidemiology and Population Health (NCEPH), The Australian National University
  - Australian Institute of Health and Welfare
  - Centre for Health Record Linkage
  - Centre for the Study of Assessment and Prioritisation in Health, School of Medicine and Health Science (NZ)
- EU
  - Health Information Research Unit, School of Medicine, Swansea University, Wales, UK
  - Health Services Research Unit, University of Aberdeen, Scotland
- Canada
  - Canadian Institute for Health Information
  - Child and Youth Data Lab, Alberta Centre for Child, Family and Community Research

PUBLIC HEALTH
TEXAS A&M HEALTH SCIENCE CENTER

POPULATION
INFORMATICS
RESEARCH GROUP

Take away: What is Data Science ?

4 Vs of Big Data

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

- Time
- Money
- Quality (Precision)

AMP

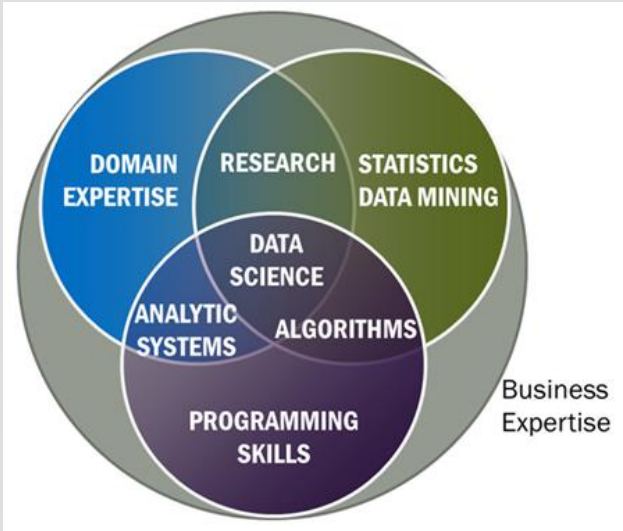
- Algorithm
- Machine
- People

PUBLIC HEALTH
TEXAS A&M HEALTH SCIENCE CENTER

POPULATION INFORMATICS RESEARCH GROUP

Take away: What is Data Science ?

Data Science is a team science: Need to know enough to communicate with statisticians and programmers



PUBLIC HEALTH
TEXAS A&M HEALTH SCIENCE CENTER

POPULATION INFORMATICS RESEARCH GROUP

Take away: What is Data Science ?

The most important skill or trait in a data scientists

Intense curiosity: “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”

Agenda

- What is Big Data ? What is Data Science ?
- What is Population Informatics & the Social Genome ?
- How is Data Science different from traditional science?
- Doing research with Big Data

Social Genome: Putting Big Data to Work to Advance Society

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

The Cost of the Digital Society

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

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

TARGET

Snowden Claims NSA Knocked All of Syria's Internet Offline

Why not reap the benefits too ?

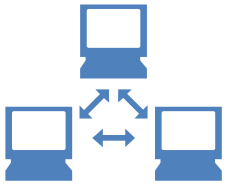
- Allocations of resources for education
  - What is the long term impact of moving to managed care ?
  - What effect does teacher pay in middle school have on college grades?
- The answers could easily be derived from relevant data sets

PUBLIC HEALTH
TEXAS A&M HEALTH SCIENCE CENTER
POPULATION INFORMATICS RESEARCH GROUP

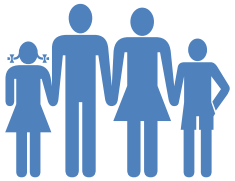
The Problem

- Higher privacy standard and accountability
  - Difficult to **access**: Privacy/confidentiality
  - No **neutral safe place** to put it together
- Requires more accurate results compared to recommending books
  - **Error management**
- Lack Data management tools and expertise
  - Difficult to **integrate**: silo/scattered data
  - Difficult to clean data

The Power of the Social Genome

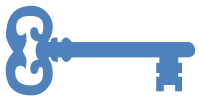


Our activities from birth until death leave **digital traces** in large databases



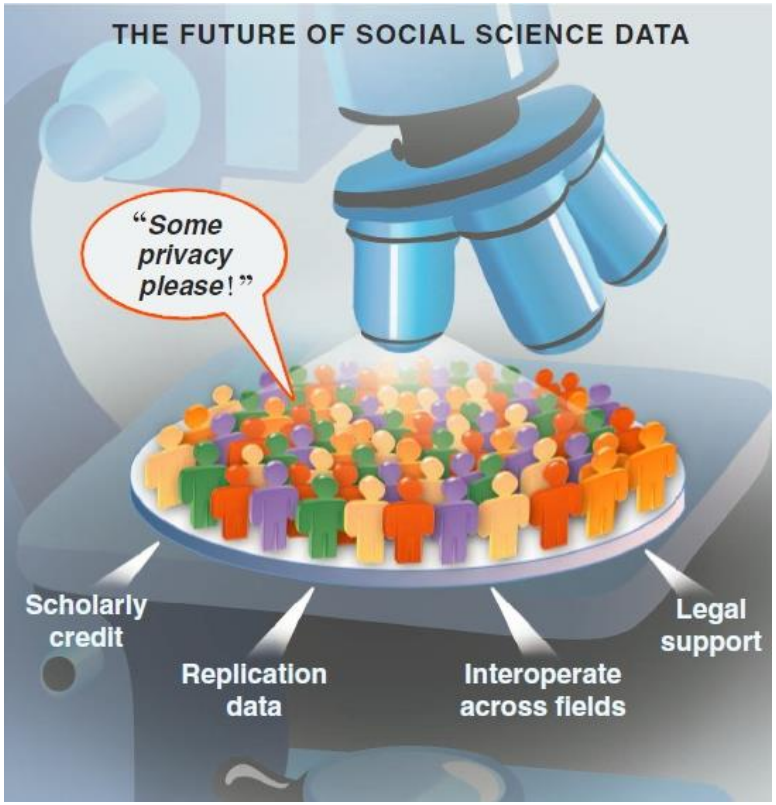
Digital traces capture our **social genome**, the footprints of our society

The social genome data are **buried** in the **massive** and **chaotic data**



It holds **crucial insights** into many of the most challenging problems facing our society (e.g. healthcare, education, economics)

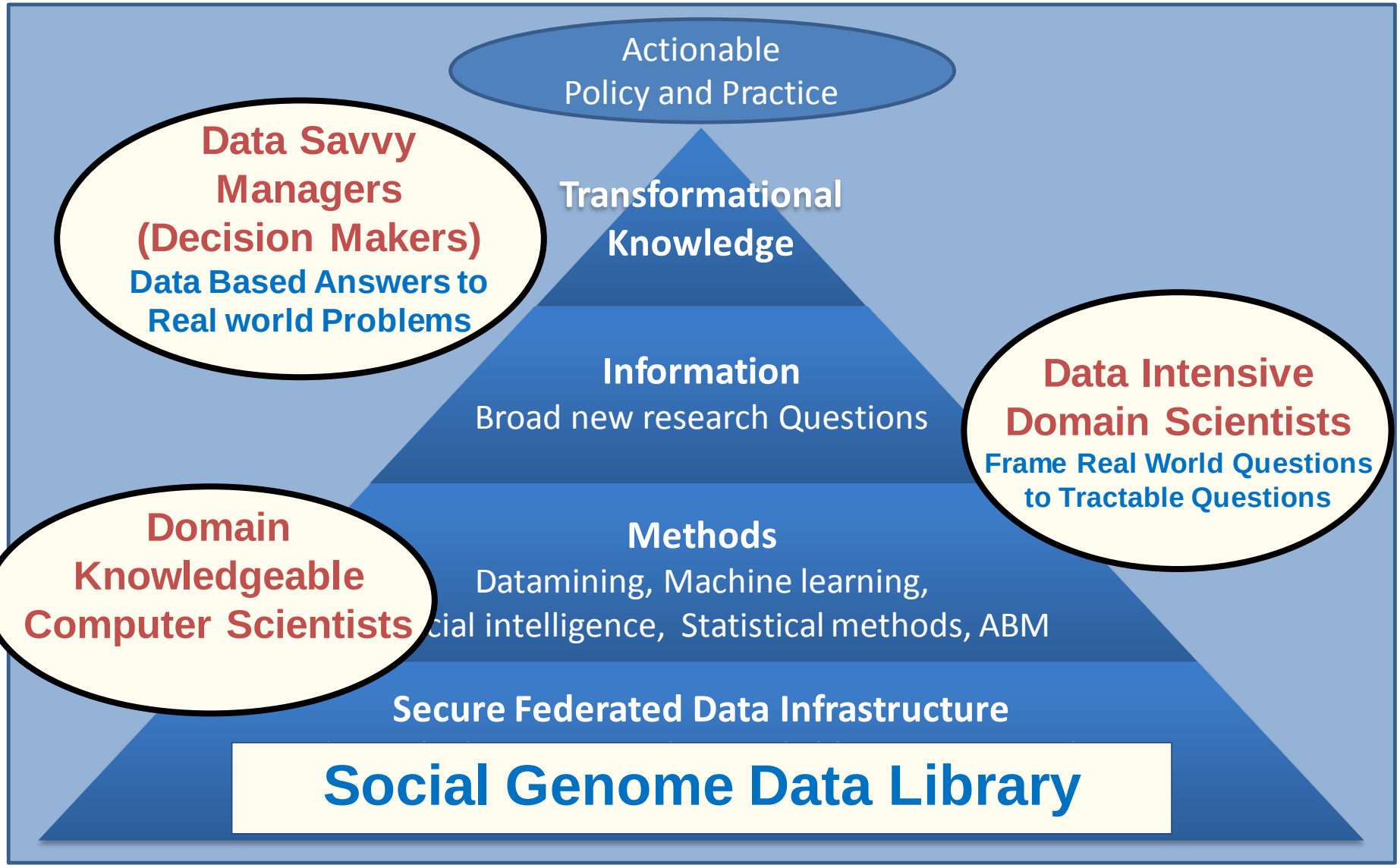
The Impact: Population Informatics



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

- **Game changing research**
  - Transform health and social sciences
  - Major impact on public policy and management
- **Easier and safer access** to **better data and tools**
  - Improved security of sensitive data already on campus
  - Improved integration of data
- At the **price of running a public library**

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. Jan 2014



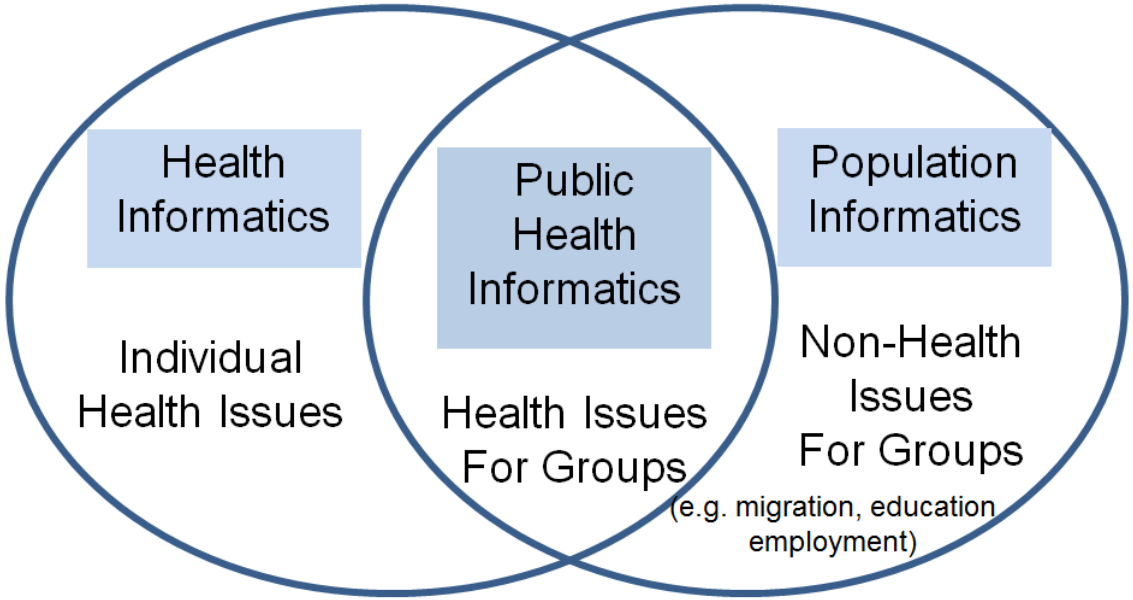
POPULATION INFORMATICS RESEARCH GROUP

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

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 (termed “big data”) about people.



North Carolina

- Child Welfare
- Medicaid
- TANF
- SNAP (foodstamps)
- Income (UI)
- Education
- Juvenile Justice

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. Jan 2014

LEHD : US Census Burea

- 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)

The Social Genome Data Library
Privacy platform for doing safe research

- Personal Data is
  - Delicate/Hazardous/Valuable
- Important to have proper systems in place that give protection (opt out)
- But allow for continued research in a safe manner (deidentified when possible)
- 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

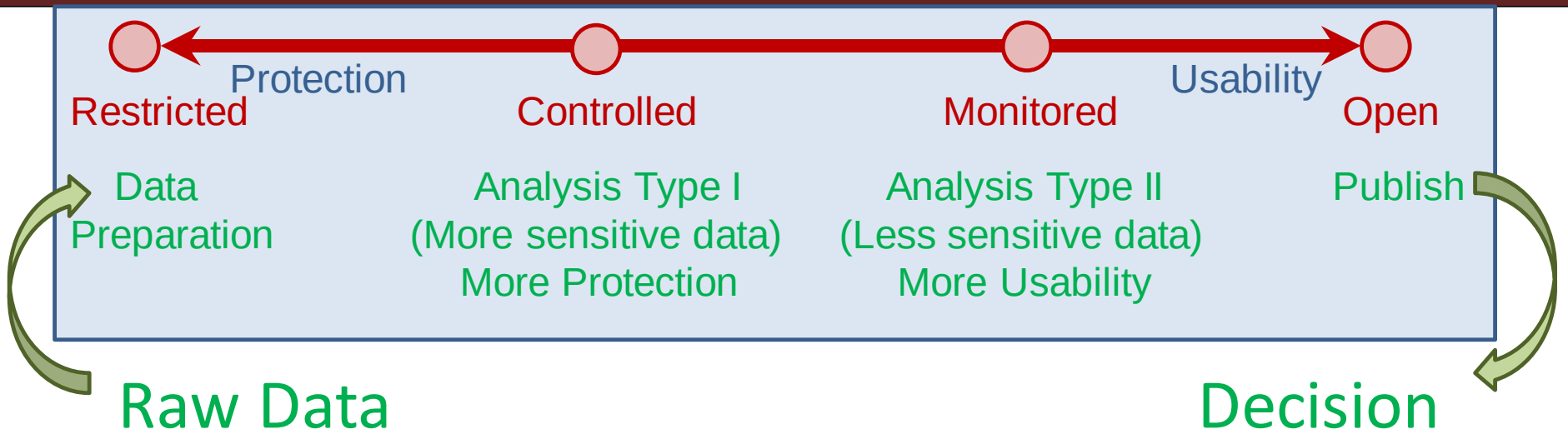
PUBLIC HEALTH
TEXAS A&M HEALTH SCIENCE CENTER

POPULATION INFORMATICS RESEARCH GROUP

WORKFLOW

Safe Platform for Data to Decision

System of Access Models



- Goal: To design an information system that can enforce the varied continuum from one end to the other such that one can balance privacy and usability as needed to turn data into decisions for a given task

The start …

- Write up a research plan on
  - What data you need
  - What you want to do with them
  - Determine access levels for each data
- Submit to IRB process

IRB: Risk of privacy violation vs. Benefit to Society

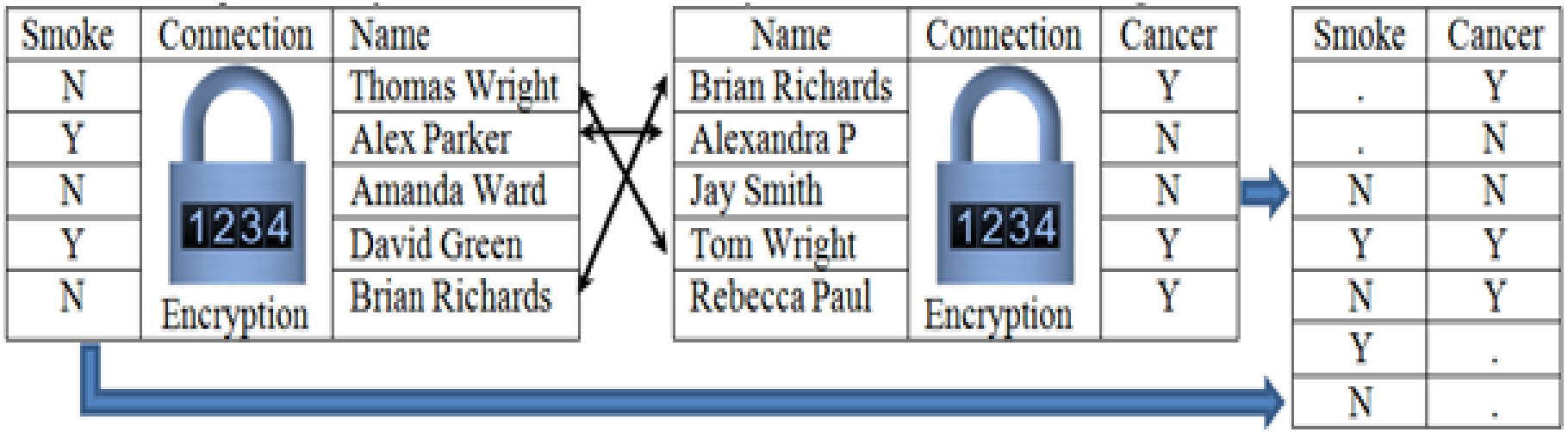
- Risk of attribute disclosure
  - Group disclosure
  - Linkage attack using auxiliary information
- Risk of identity disclosure
- Given?
  - Kinds of data elements used in the study
    - Name/dob/cancer status/ etc… (are there $$)
  - What system the data resides in : HW/SW
    - Risk of outsiders intruding / insider attack / negligence
  - What can users do with the data on the system
    - Take data off / look at everything / only do limited queries

Restricted Access : Prepare the customized data

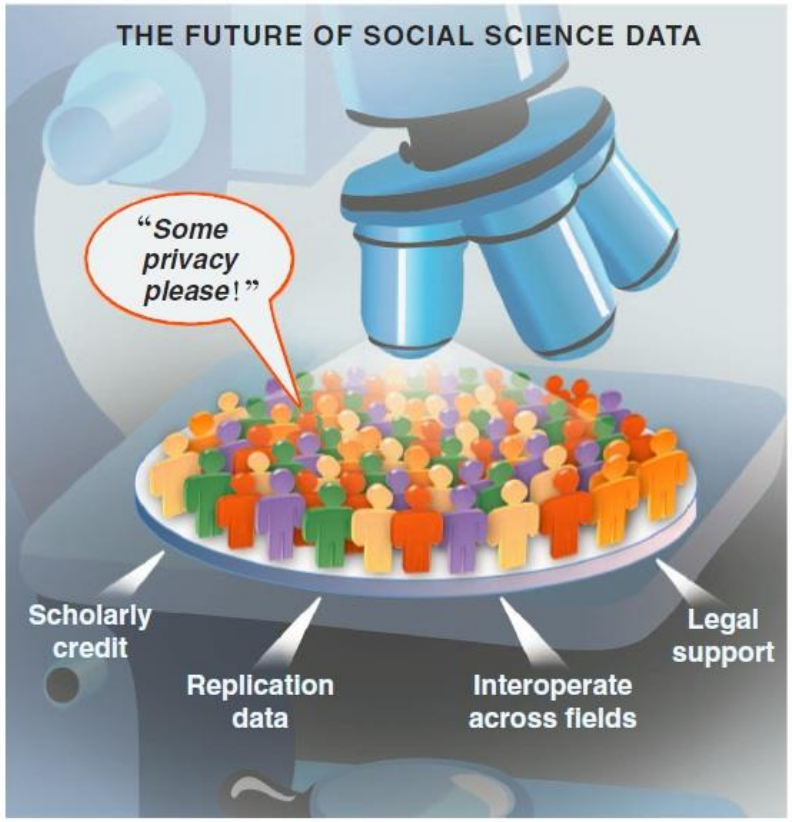
- Decoupled Data (Kum 2012)
  - Automated Honest Broker SW
- Sample selection
- Attribute selection
- Data integration (access to PII)
- Some data cleaning
- Full IRB
- Example: RDC

Privacy Preserving Interactive Record Linkage

- Decouple data via encryption
- Automated honest broker approach via computerized third party model
- Chaffe to prevent group disclosure
- Kum et al. 2012



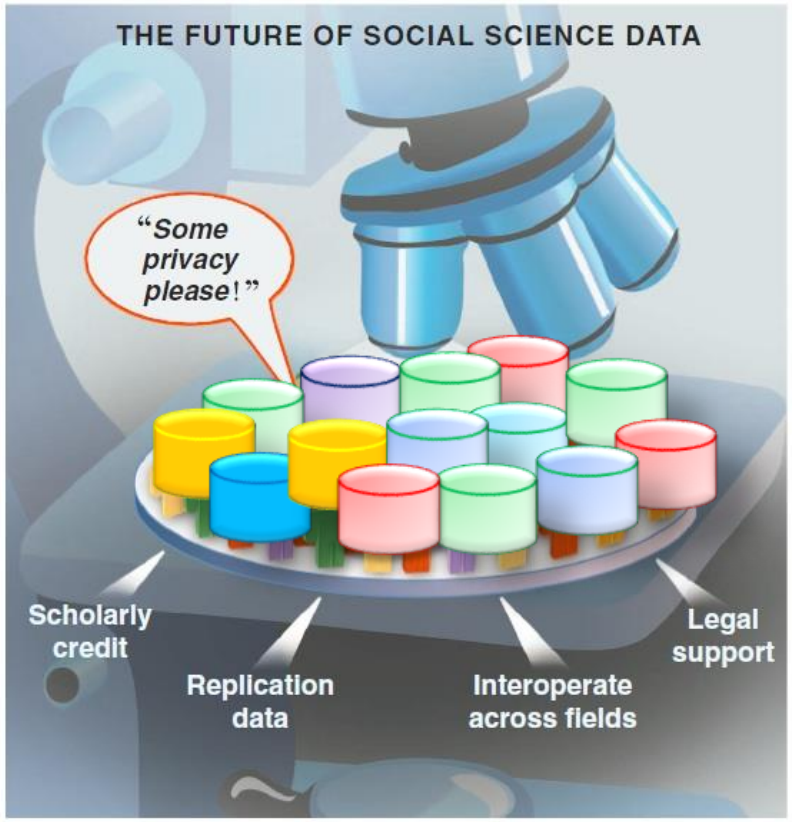
Controlled Access : Model using given tools



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

- With approved deidentified data
- Locked down VM: customized appliances
- only approved software
- Remote access via VPN
- Very effective for threats from HBC
- Full IRB
- U Chicago-NORC , UNC-Tracs (CTSA), UCSD-iDASH, SAIL

Monitored Access : Freely Repurpose

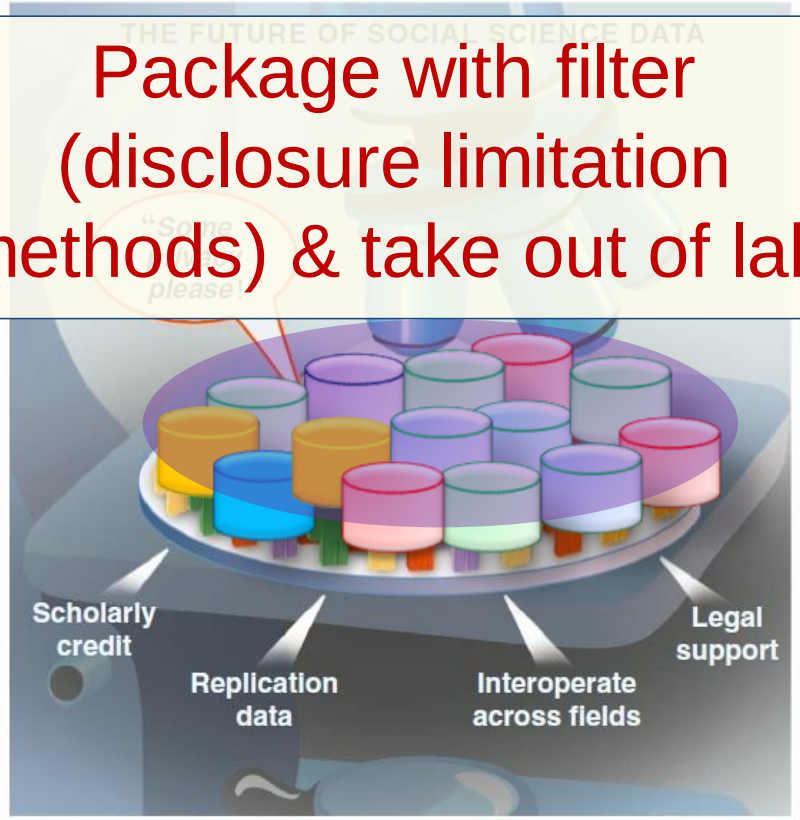


Gary King. Ensuring the Data-Rich Future of the Social Sciences, Science, vol 331, 2011, pp 719-721.

- Information Accountability model
- **Exempt IRB: Explicit data use agreement (5 big Q)**
  - **Public online (crowdsource)**
- **Any software & auxiliary data**
- Remote Access via VPN
- Less sensitive data (e.g. Aggregate data)
- SHRINE, Secure Unix servers

Open Access : No restriction on use

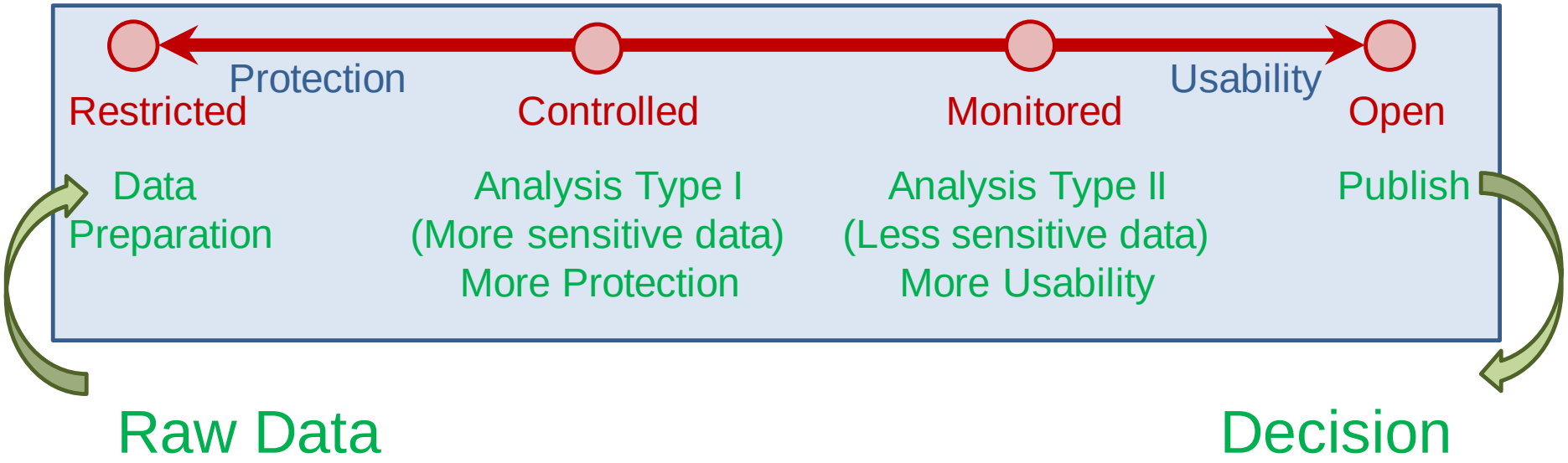
Package with filter (disclosure limitation methods) & take out of lab



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

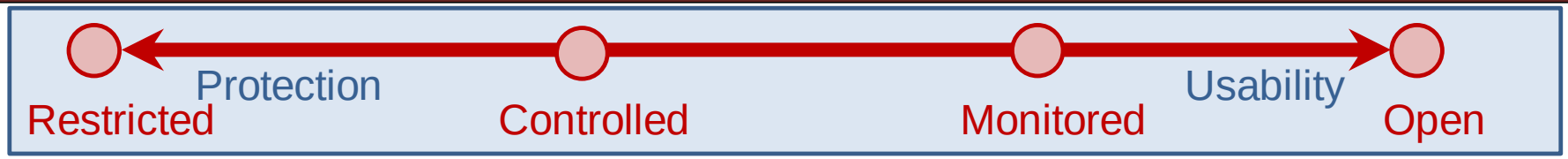
- **Anyone : Publish information for others**
- No IRB
- No monitoring use
- Disclosure Limitation Methods (filter)
- Sanitized data
- Public websites, publications
- Publish data use terms

Use Published Data for Good Decision Making



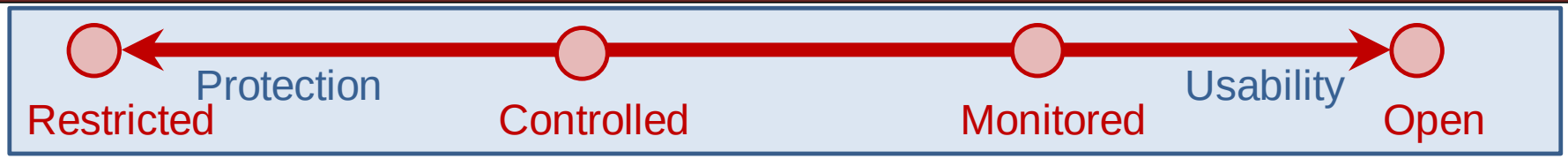
Deployed together the four data access models can provide a comprehensive system for privacy protection, balancing the risk and usability of secondary data in population informatics research

Privacy Protection Mechanism



Access	Restricted Access	Controlled Access	Monitored Access	Open Access
Protection Approach	Physical restriction to access	Lock down VM (limit what you can do on the system)	Information accountability	Disclosure Limitation
Monitoring Use	All use on & OFF the computer is monitored	All use on the computer is monitored		Trust
IRB	Full IRB approved	Full IRB approved	IRB Exempt (register)	Terms of Use
R1:Crypto-graphic Attack	Very Low Risk	Low Risk. Would have to break into VM	High Risk	NA
R2: Data Leakage	Very Low Risk. Memorize data and take out	Physical data leakage (Take a picture of monitor)	Electronically take data off the system.	

Comparison of risk and usability



		Restricted Access	Controlled Access	Monitored Access	Open Access
Usability	U1.1: Software (SW)	Only preinstalled data integration & tabulation SW. No query capacity	Requested and approved statistical software only	Any software	Any software
	U1.2: Data	No outside data allowed But PII data	Only preapproved outside data allowed	Any data	Any data
	U2: Access	No Remote Access	Remote Access	Remote Access	Remote Access
Risk	R1:Crypto-graphic Attack	Very Low Risk	Low Risk. Would have to break into VM.	High Risk	NA
	R2: Data Leakage	Very Low Risk. Memorize data and take out	Physical data leakage (Take a picture of monitor)	Electronically take data off the system.	NA

Data privacy and confidentiality

- Critical for population informatics
- Data governance
- Security
- Ethics of data use

Abuse of Data

- Too Often
  - Sufficient time is needed to give anything a proper chance and then to evaluate with data
- Too Punitive (score card)
  - Remember goal of using data is to improve
  - Can back fire by becoming too conservative or short term focused
- Too Rigid
  - Reality change over time
  - Data has LOTS of issues. Only an ESTIMATE

Conclusion

- There is a lot you can do with digital data now
- BUT, lots of data is not the answer
  - You have to learn to use data properly
  - You have to learn to handle data if you want to do good research using massive secondary data
  - Massive secondary data requires as much or more preprocessing as primary data collection
    - Research design, data cleaning, data preparation
  - Nothing replaces common sense (critical thinking) and curiosity in research

PUBLIC HEALTH
TEXAS A&M HEALTH SCIENCE CENTER

POPULATION
INFORMATICS
RESEARCH GROUP

vocab

- Privacy vs confidentiality
- Privacy vs security

Take Away I
Information Accountability Works

- Secrecy : Hiding information
  - In reality, has limited power to protect privacy
  - Severe Consequences related to
    - Accuracy of data and decisions, use of data for legitimate reasons, transparency & democracy
- **Information Accountability** (Credit Report)
  - Very clear transparency in the use of the data
  - Disclosure : Declared in writing, so when something goes wrong the right people are held accountable (data use agreements)
  - IT WORKS! Primary method used to protect financial data
  - Internet : crowdsourced auditing (public access IRB)
  - Logs & audits : what to log, how to keep tamperproof log

D.J. Weitzner et al., Information Accountability, Comm. ACM, vol. 51, no. 6, 2008, pp. 82–87.

Ethical & Controlled Repository

Very clear transparency in the use of the data

- Open Access: Summary statistics (open data)
- Monitored Access: Register your research plan
- Controlled Access: Approval by IRB
  - Risk of privacy violation vs. benefit to society
- Restricted Access: Link data accurately

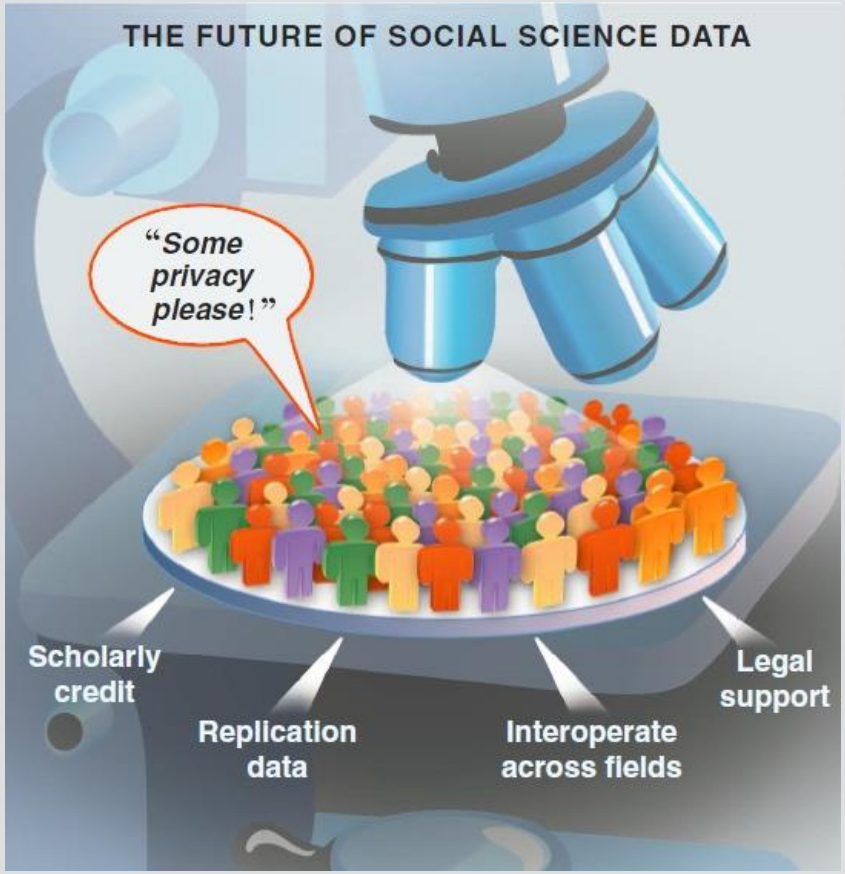
Take Away II
Privacy is a BUDGET constrained problem

- Differential Privacy proves each query leads to some privacy loss while providing some utility in terms of data analysis.
- The goal is to achieve the maximum utility under a fixed privacy budget

Take Away III
Privacy Expectation for **Doing Research**

- Consider the **RISK of HARM** versus **BENEFIT to SOCIETY**
- Taking into account the **COST**
  - Individual privacy
  - Cost of integrity of data : bad data can lead to wrong decisions
  - Lost opportunity cost of no access to data
  - Organization transparency & accountability (democracy)
  - Value gained through obtaining timely, accurate, appropriate information for good decision making
  - Financial cost of data security measures
- **Transparent and accountable use of data**

Will you donate your data to find a cure for cancer?



Practical Example

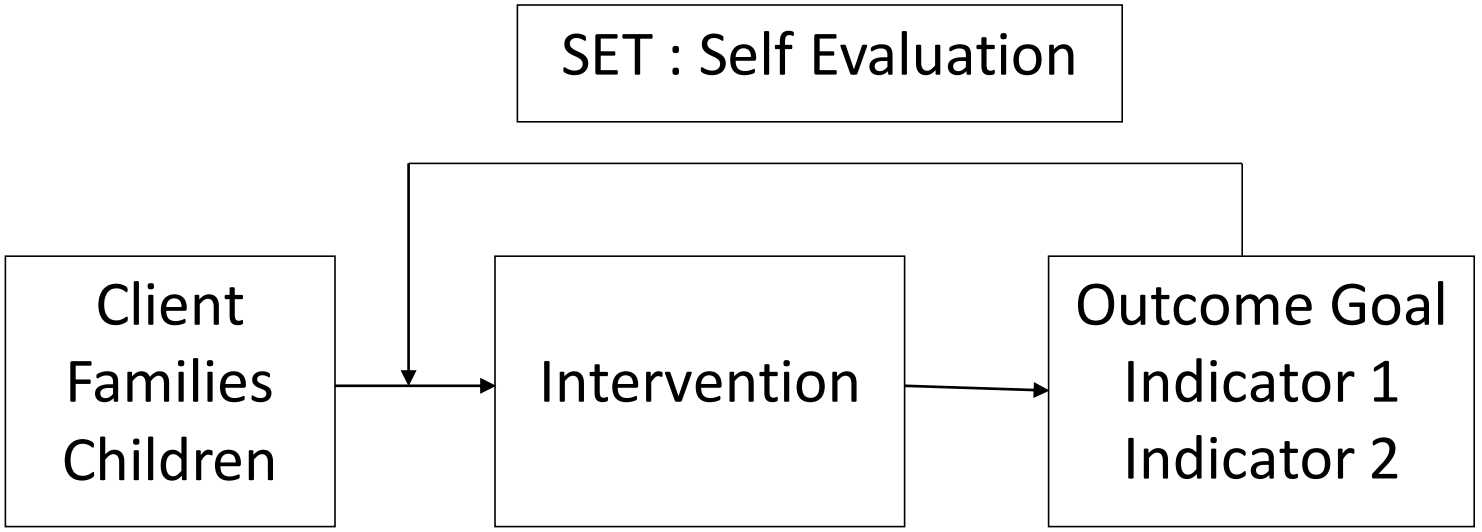
Self Evaluation

Traditional evaluation

Traditional evaluation

- Hallmark: Detachment of the evaluator from policymakers and program managers
- Evaluation becomes adversarial
- Can become ill-informed due to the lack of communication between evaluators and staff
- Scientific objectivity : results in tasks being defined in such narrow terms that their contribution to informed policy debate is minimized

Alternative : Self Evaluation



- is a form of empowerment evaluation
- that is collaborative and participatory
- an ongoing process
- as long at it is technically strong, a viable alternative

Self-evaluation : who ?

- SET (self-evaluation team)
  - a team of diverse people
  - gathered locally for their expertise
  - Should involve stake holders
  - evaluators may work for an independent organization
  - or be employed by the agency administering the program

Guilford County : Leading By Results

- Goal : At-risk children and families should be safe and healthy in stable environments
  - Indicator 1: We will decrease the rate of children placed away from their home from 7% in FY 2003 to 5% by the end of FY 2005.
  - Indicator 2: We will decrease the rate of children entering foster care who are initially placed in shelter or group home care from 13% to 11% by the end of FY 2005.
  - Indicator 3: We will reduce the rate of children re-entering care from 10% in FY 2003 to 8% by the end of FY 2005.
  - Indicator 4: We will reduce the number of children in care with four or more placement moves from 14% in FY 2003 to 12% by the end of FY 2005.
  - Indicator 5: We will maintain the percent of children substantiated/in need of services that are not repeat victims of substantiated maltreatment at 92% by the end of FY 2005.
  - Indicator 6: We will continue to work on addressing disparities associated with race/ethnicity as evidenced by a decrease in the percentage of African American children in care from 57% by the end of FY 2005.

Things to watch out for in self-evaluation

- Outcomes is only one part of the picture
  - You have to monitor process
  - You have to pay attention to context
- Don't lose sight of the forest
  - Remember one indicator is just a tree
- Balance changes to the intervention with solid research

Key ingredients for self-evaluation

- Outcomes for clients are clearly defined and disseminated throughout the agency
- It is a collaborative process that brings together individuals with different kinds of expertise to discuss the data used to measure outcomes and agency processes
- It is an ongoing process with regular meetings to discuss agency status on outcomes
- It requires timely and accessible data that appropriately measure outcomes and other indicators of interest
- It should include ongoing attention to implementation progress of the core strategies to improve the outcomes
- It requires technical expertise to ensure defensibility and adaptability

Key ingredients for self-evaluation

- Outcomes for clients are clearly defined and disseminated throughout the agency
- It is a collaborative process that brings together individuals with different kinds of expertise to discuss the data used to measure outcomes and agency processes
- It is an ongoing process with regular meetings to discuss agency status on outcomes
- **It requires timely and accessible data that appropriately measured outcomes and other indicators of interest**
- It should include ongoing attention to implementation progress of the core strategies to improve the outcomes.
- **It requires technical expertise to ensure defensibility and adaptability**

→ **KDD Technology**

Self Evaluation and KDD

- **as long at it is technically strong**
  - Usually requires technical assistance
  - Outcomes data analysis
  - Knowledge Discovery and Datamining (KDD)

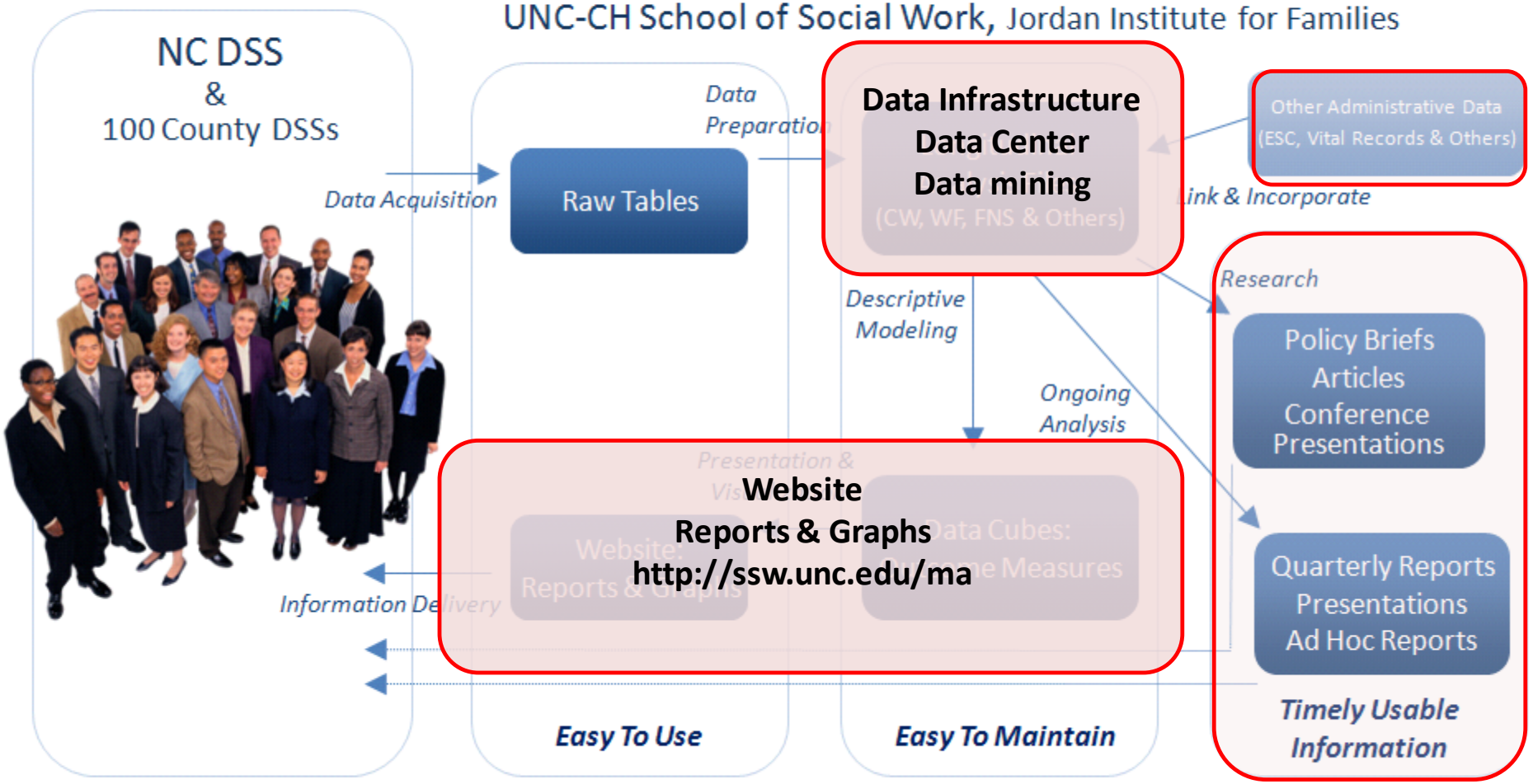
- Usher, C. L., Wildfire, J. & Schneider, S. (2001)
- Usher, C.L. (1995)

KDD Architecture for SW Administrative Data

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

UNC-CH School of Social Work & NC DSS

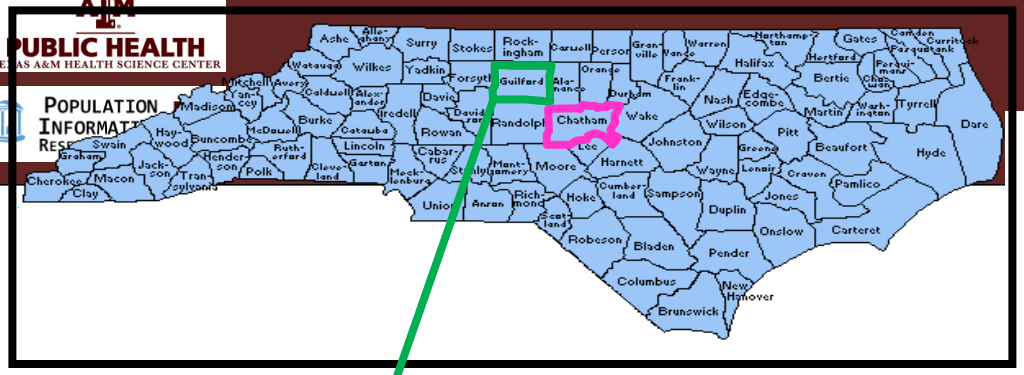
Jordan Institute for Families, http://ssw.unc.edu/ma

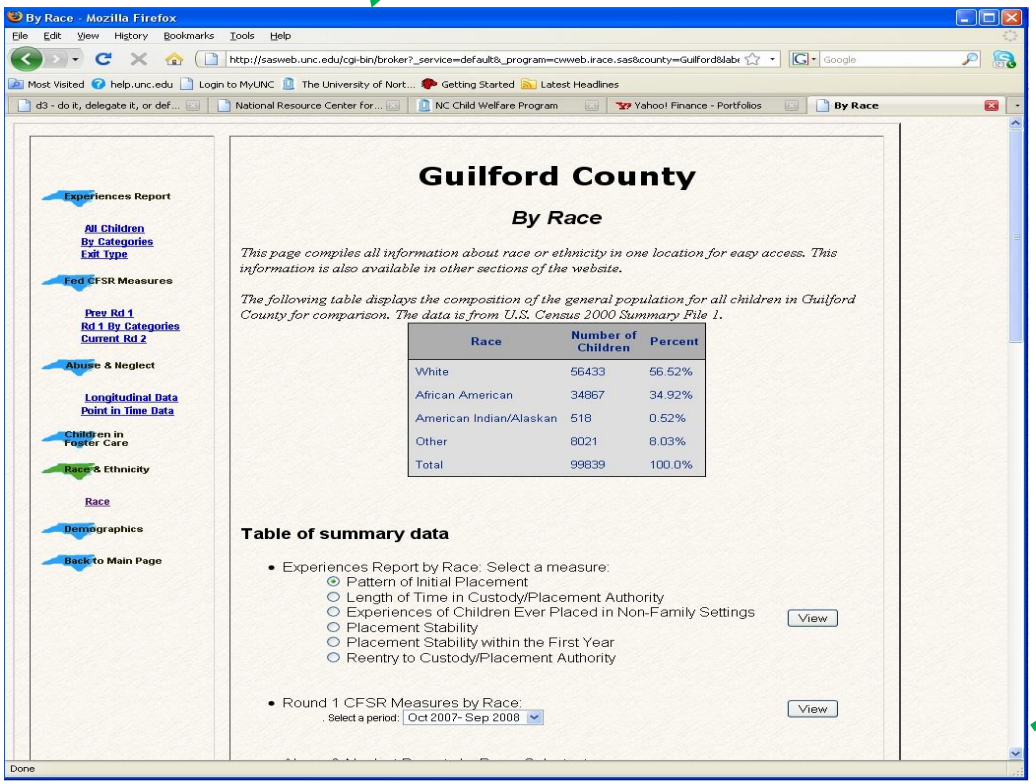


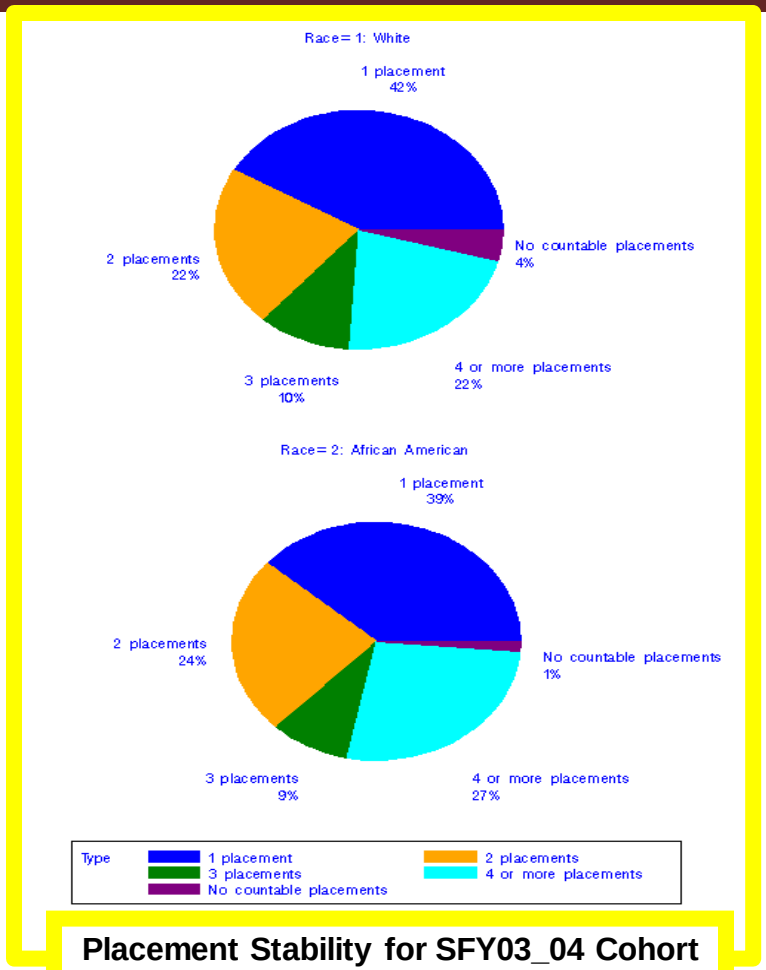
Duncan, D., Kum, H., Flair, K., & Stewart, C.J. (2009). Management Assistance for Child Welfare, Work First, and Food & Nutrition Services in North Carolina. Jordan Institute for Families website: http://ssw.unc.edu/ma

PUBLIC HEALTH
TEXAS A&M HEALTH SCIENCE CENTER

POPULATION INFORMATION RESEARCH



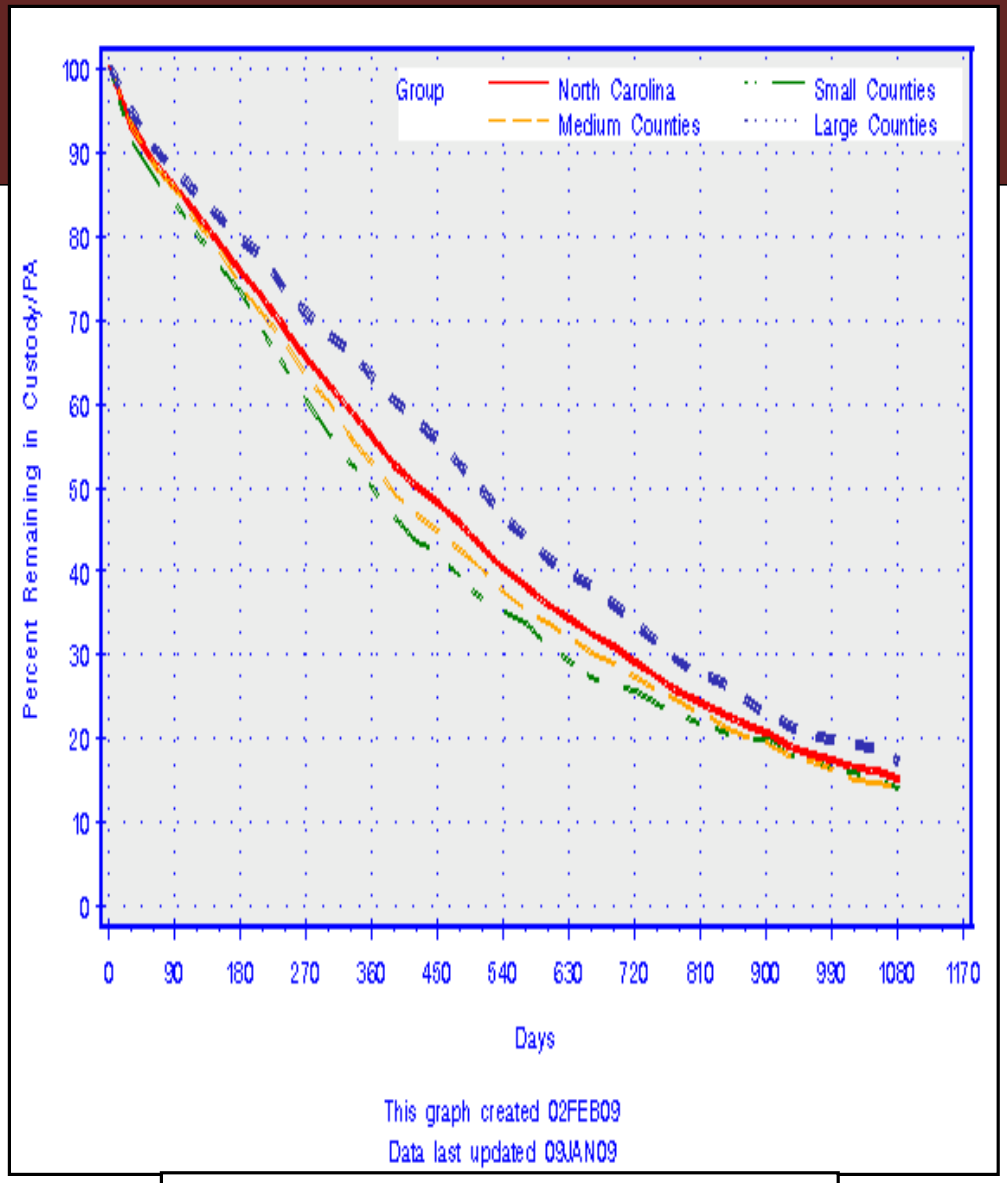




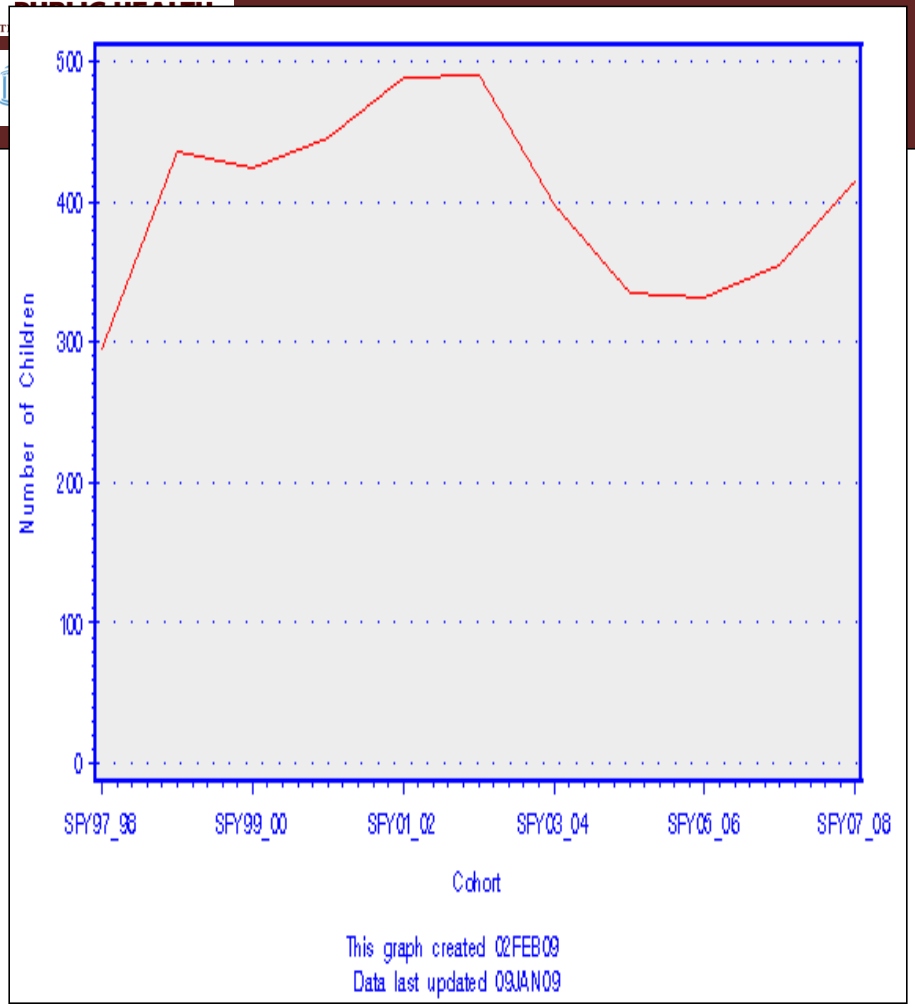
Placement Stability for SFY03_04 Cohort In Guilford County by Race

Explanation for this Chart

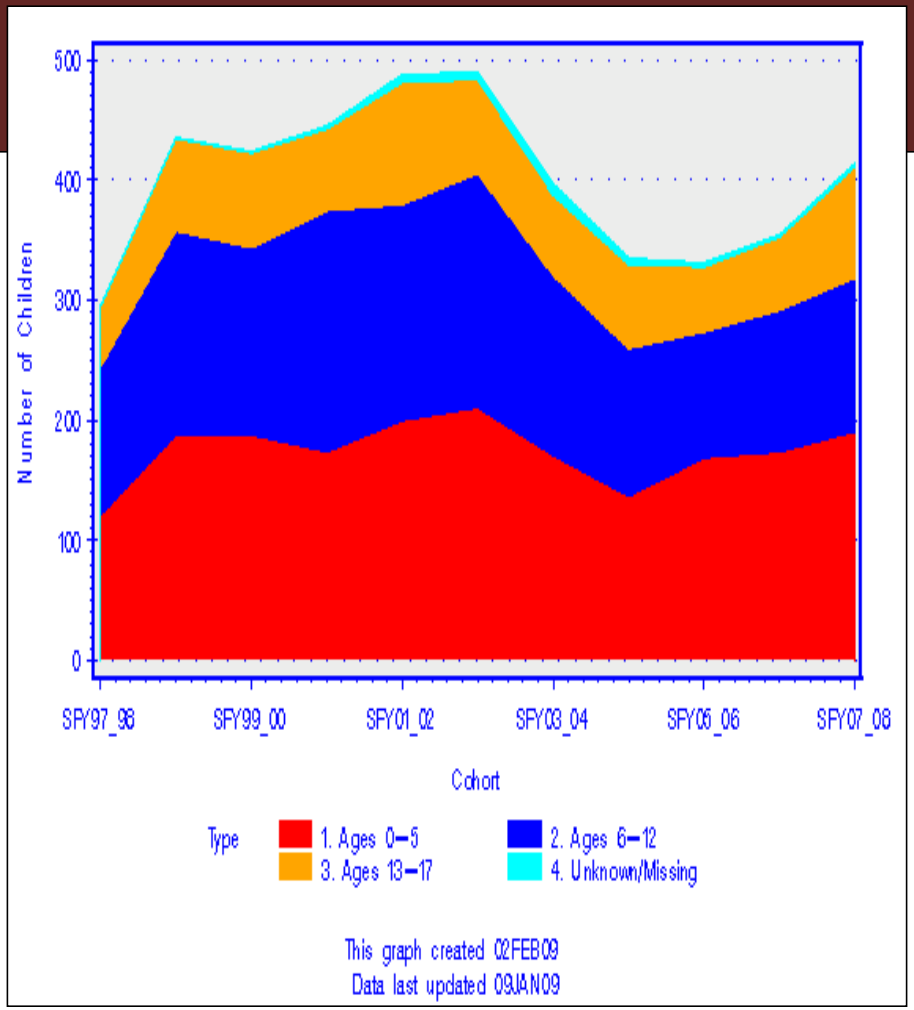
This chart depicts how many days children spent in their first Custody/Placement Authority. Where the line crosses the 50 percent line, half of the children in that cohort are no longer in custody. Length of stay is a longstanding concern regarding children's experiences in out-of-home care …



Rate of Leavaing Custody for the Children in SFY05_06 Cohort for North Carolina



Reports of Abuse and Neglect
In Chatham County
Unique Number of Children
by First Ever Report Cohort



Reports of Abuse and Neglect by Age
In Chatham County
Unique Number of Children
by First Ever Report Cohort

Stanly County
Stokes County
Surry County
Swain County
Transylvania County
Tyrrell County
Union County
Vance County
Wake County
Warren County
Washington County
Watauga County
Wayne County
Wilkes County
Wilson County
Yadkin County
Yancey County
Judicial District 1
Judicial District 2

Child Welfare
Experiences Report
All Children
By Categories
Summary Data
CFSR Measures
Prev Rd 1
Rd 1 By Categories
Current Rd 2
Abuse & Neglect
Longitudinal Data
Point in Time Data
Children in Foster Care
All Children
Age Out
Race & Ethnicity
Work First
Food & Nutrition Services
Papers & Reports
Additional Information
Help

[North Carolina] : Reports of Abuse and Neglect Type of Finding on Most Severe Report by Categories

Unique Number of Children by First Ever Report: Longitudinal Data

State Fiscal Year=SFY1997_1998

Type Found	Total	White	African American	American Indian/Alaskan	Other Races	Hispanic	Non_Hispanic	Male	Female	Ages 0 to 5	Ages 6 to 12	Ages 13 to 17	Missing Age Information
Abuse and Neglect	587	370	184	10	23	55	532	233	354	226	209	148	4
Abuse	1145	712	370	11	52	78	1067	434	711	352	465	328	0
Neglect	13587	7525	5297	258	507	1062	12525	6896	6691	7358	4480	1653	96
Dependency	248	111	118	3	16	39	209	121	127	134	52	59	3
Services Recommended	3	0	3	0	0	0	3	2	1	2	1	0	0
Unsubstantiated	43215	26297	14725	878	1315	3256	39959	21880	21334	20243	15658	7079	235
Services Not Recommended	3	0	3	0	0	0	3	2	1	1	0	0	2

State Fiscal Year=SFY1998_1999

Type Found	Total	White	African American	American Indian/Alaskan	Other Races	Hispanic	Non_Hispanic	Male	Female	Ages 0 to 5	Ages 6 to 12	Ages 13 to 17	Missing Age Information
Abuse and Neglect	606	394	191	4	17	56	550	237	369	228	240	137	1
Abuse	988	627	304	12	45	78	910	349	639	319	388	275	6
Neglect	13776	7689	5250	250	587	1126	12650	7219	6557	7270	4628	1779	99
Dependency	242	129	101	4	8	30	212	118	124	100	63	76	3
Unsubstantiated	45463	27246	15720	924	1573	3421	42042	23043	22420	21199	16415	7619	230
Services Not Recommended	2	0	2	0	0	0	2	2	0	1	0	0	1

Research

- Barth, R. P., Duncan, D. F., Hodorowicz, M.T., and **Kum, H.C.**, **Felonious Arrests of Former Foster Care and TANF-Involved Youth**, *Journal of the Society for Social Work and Research,* 1:pp 104-123, 2010.
- **Kum, H. C.**, Barth, R., Stewart, C. J., Lee, C. K. (2012). **Coming of Age: Employment Outcomes for Youth Who Age Out of Foster Care Through Their Middle to Late Twenties**.
- **Kum, H.C.**, Duncan, D.F., & Stewart, C. J., **Supporting Self-Evaluation in Local Government via KDD**, *Government Information Quarterly: Building the Next-Generation Digital Government Infrastructures,* 26(2):pp 295-304, April 2009.
- Various state reports
  - 09 North Carolina's Children's Index, Action for Children
  - BTC (NC Budget & Tax Center) Brief, NC Justice Center
  - BTC (NC Budget & Tax Center) Brief, NC Justice Center

KDD system

- Easy to use and maintain
- Timely information
- Difficulties :counting!
  - What to measure? How to measure?
  - Measure child maltreatment in county ?
  - When you have a solid count of important things, easy to apply sophisticated statistics on the count
- IT Difficulties
  - Project management : testing, help pages
  - Data management

Lessons : common sense counting

- What to count?
- What is the context (denominator)?
- Understand exactly what you have
  - What is the unit of rows
  - What is being counted in variables (columns)
- ALWAYS check for face validity of counts
  - Data that is not used goes bad
- Cross check whenever you can
- Sensitivity analysis
- Look at your data

Agenda

- What is Big Data ? What is Data Science ?
- What is Population Informatics & the Social Genome ?
- How is Data Science different from traditional science?
- Doing research with Big Data

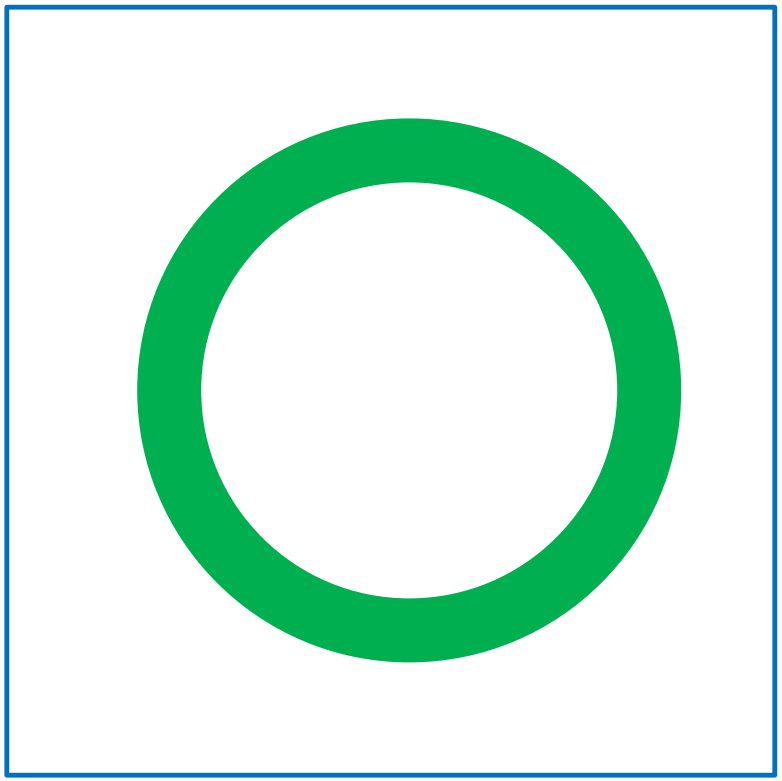
Video

- http://research.tamhsc.edu/pinformatics/data-science/

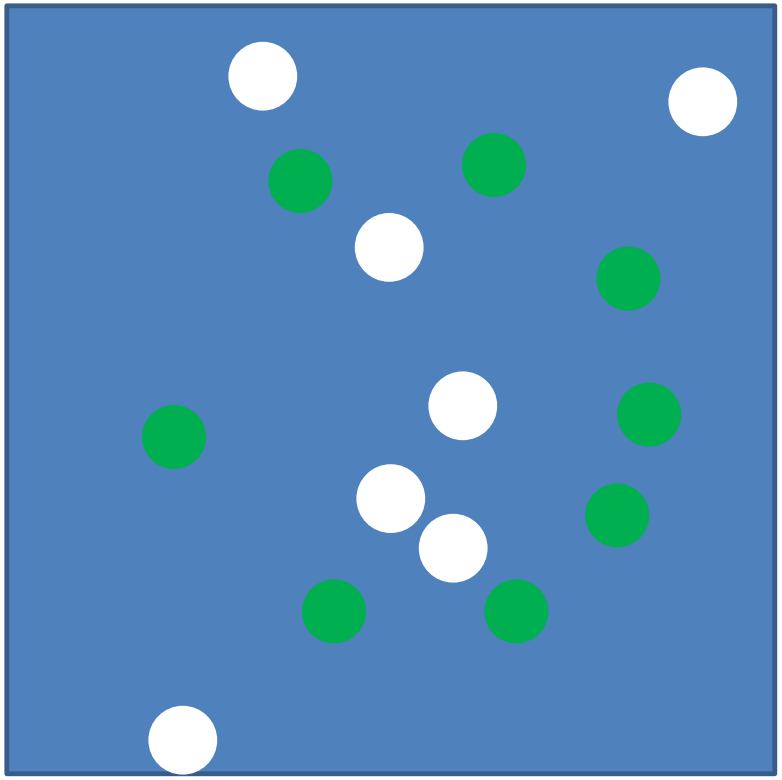
Inflation:
Traditional social science vs Data Science

- 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.)

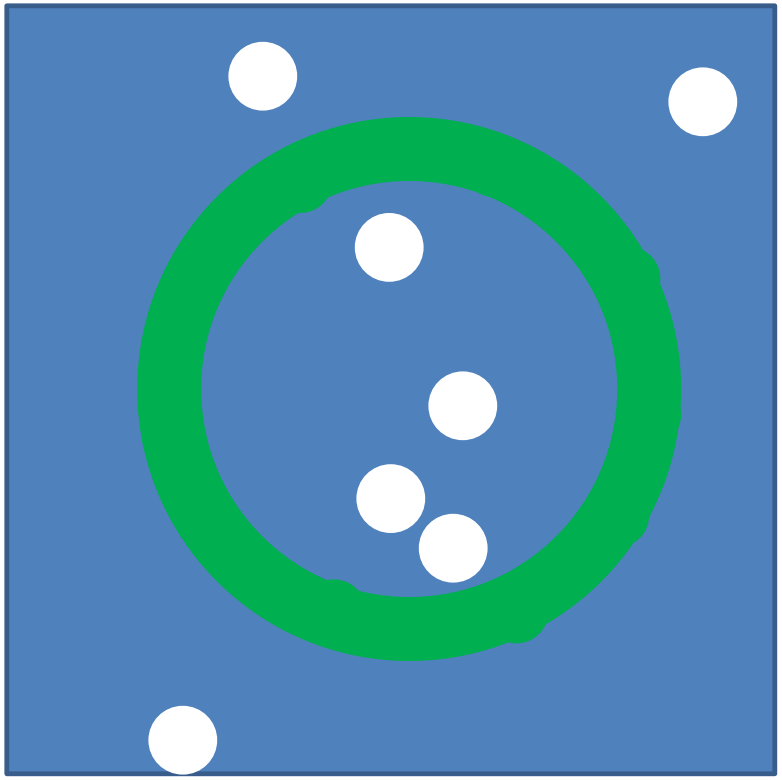
What is the shape of the green line?



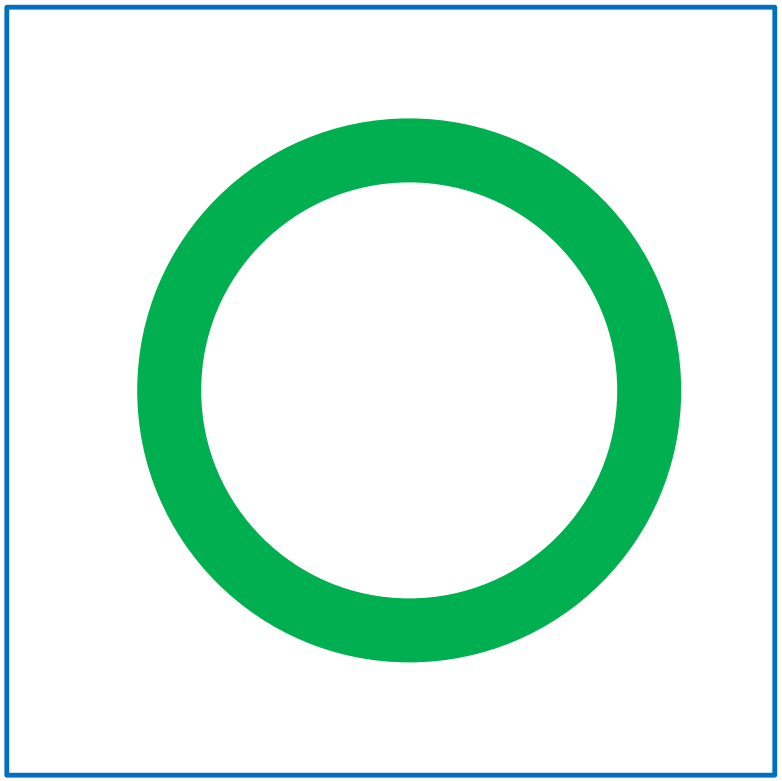
Traditional Science :
Start with nothing – collect data well



Traditional Science :
Start with nothing – collect data well



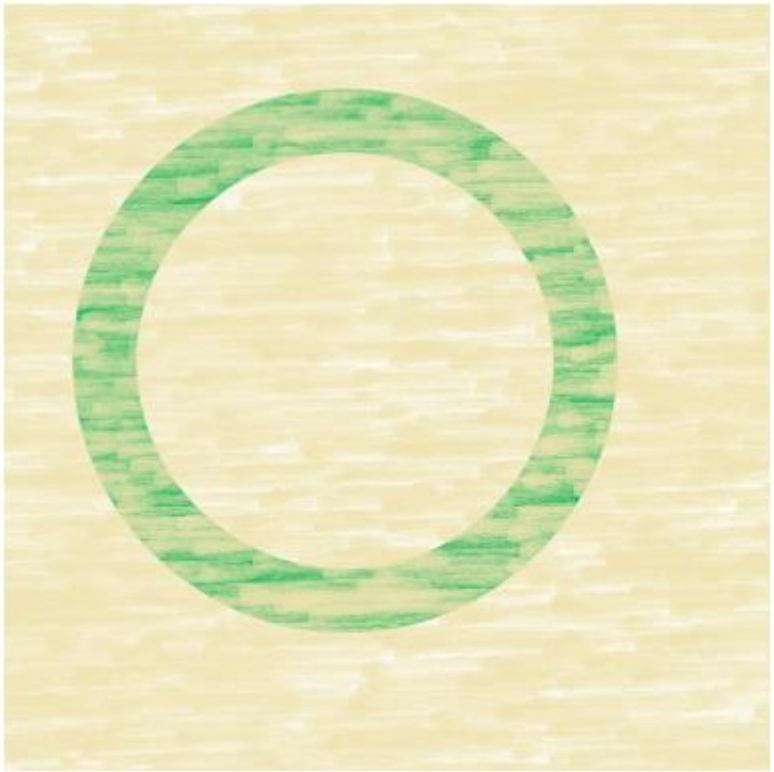
Data Science : Start with ALL the data



PUBLIC HEALTH
TEXAS A&M HEALTH SCIENCE CENTER

POPULATION INFORMATICS RESEARCH GROUP

Noise



PUBLIC HEALTH
TEXAS A&M HEALTH SCIENCE CENTER

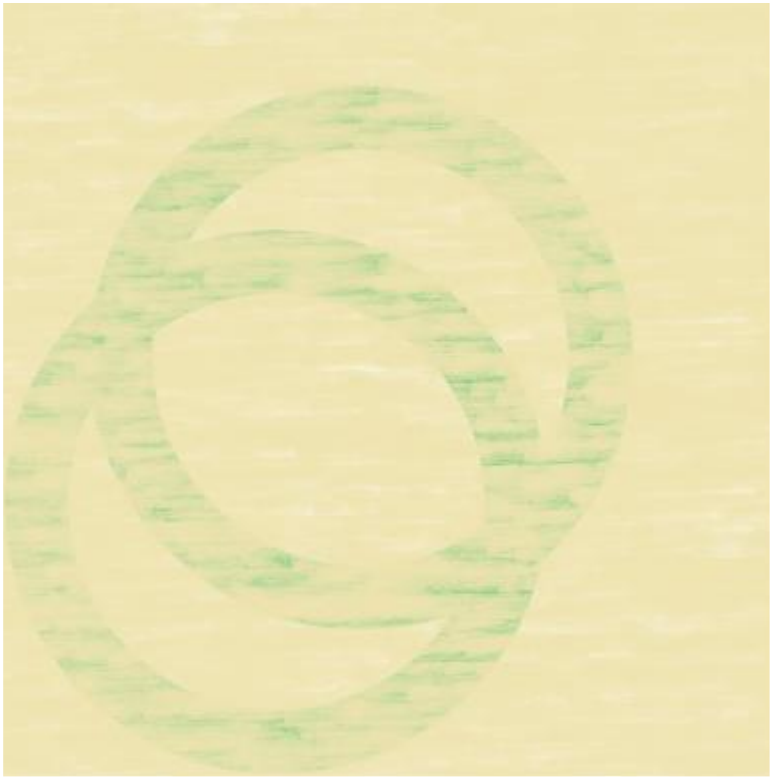
POPULATION INFORMATICS RESEARCH GROUP

Data Science: EVERYTHING

PUBLIC HEALTH
TEXAS A&M HEALTH SCIENCE CENTER

POPULATION INFORMATICS RESEARCH GROUP

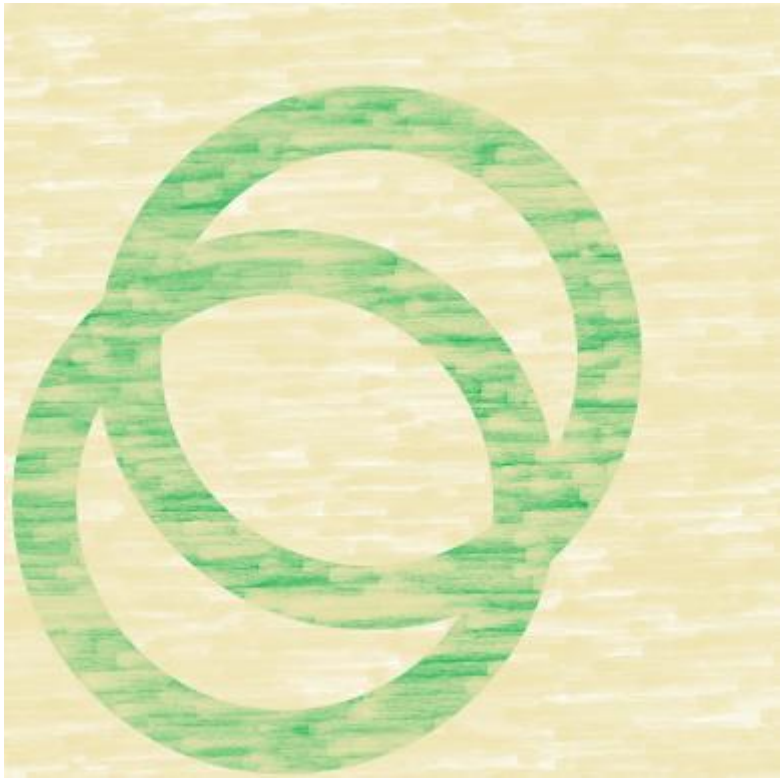
First, separate out only the relevant data



PUBLIC HEALTH
TEXAS A&M HEALTH SCIENCE CENTER

POPULATION INFORMATICS RESEARCH GROUP

Second, clean noise as much as possible



Data Science
Knowledge Discovery & Data mining (KDD)

Big Data : impossible to keep organized

KDD
Clean, Merge, Reprocess

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

Analytic dataset

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

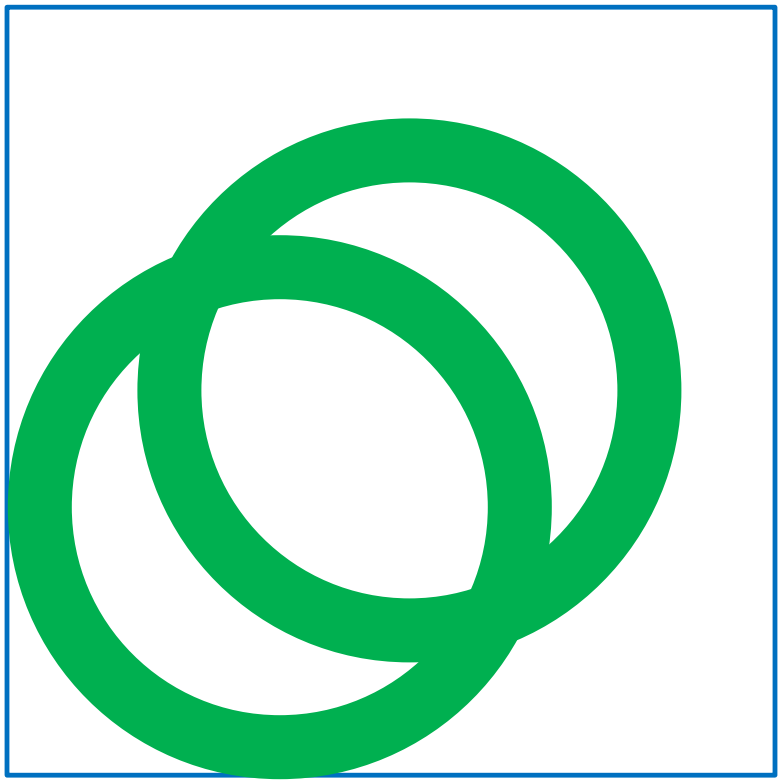
- Data Wrangling is a term that is applied to activities that make data more usable by changing their form but not their meaning
  - reformatting data: MDY vs YMD
  - mapping data from one data model to another: ICD9 vs CPT code
  - and/or converting data into more consumable forms: to graphs
- 30-80% 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

Tools for data wrangling

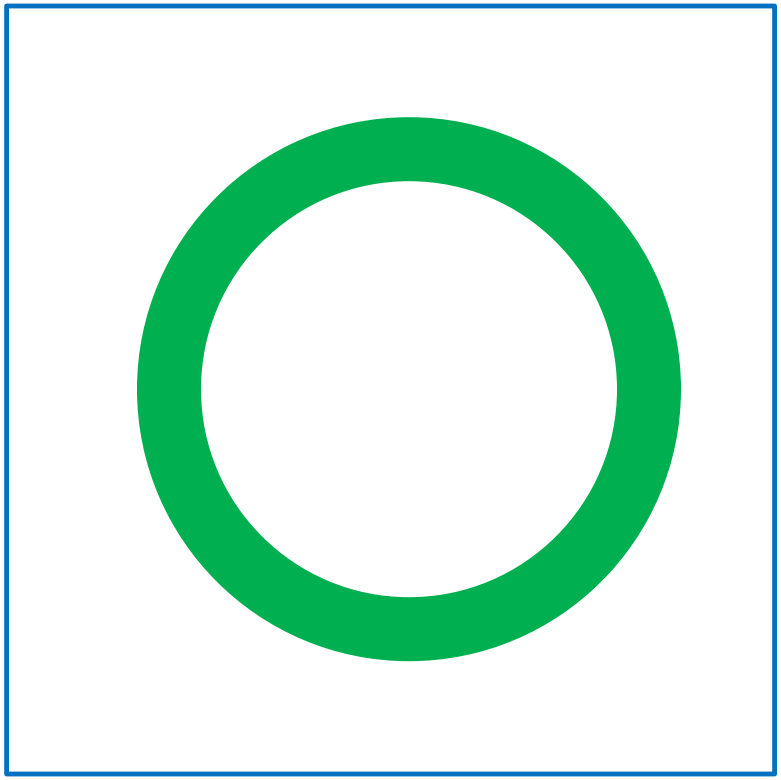
- Reusable code
  - SAS macros, stata & R: user defined
  - Write your own
  - Download others: Read others and modify
  - Per data, accumulate custom tools for the data
    - Keep it organized so you can find it when you need it
    - Keep it general enough so it can be used more widely
  - Learn basic
    - Programming concepts: loops, conditionals, flow diagram
    - Software Engineering

Third: Model

Fourth: Validate to avoid overfitting



Model



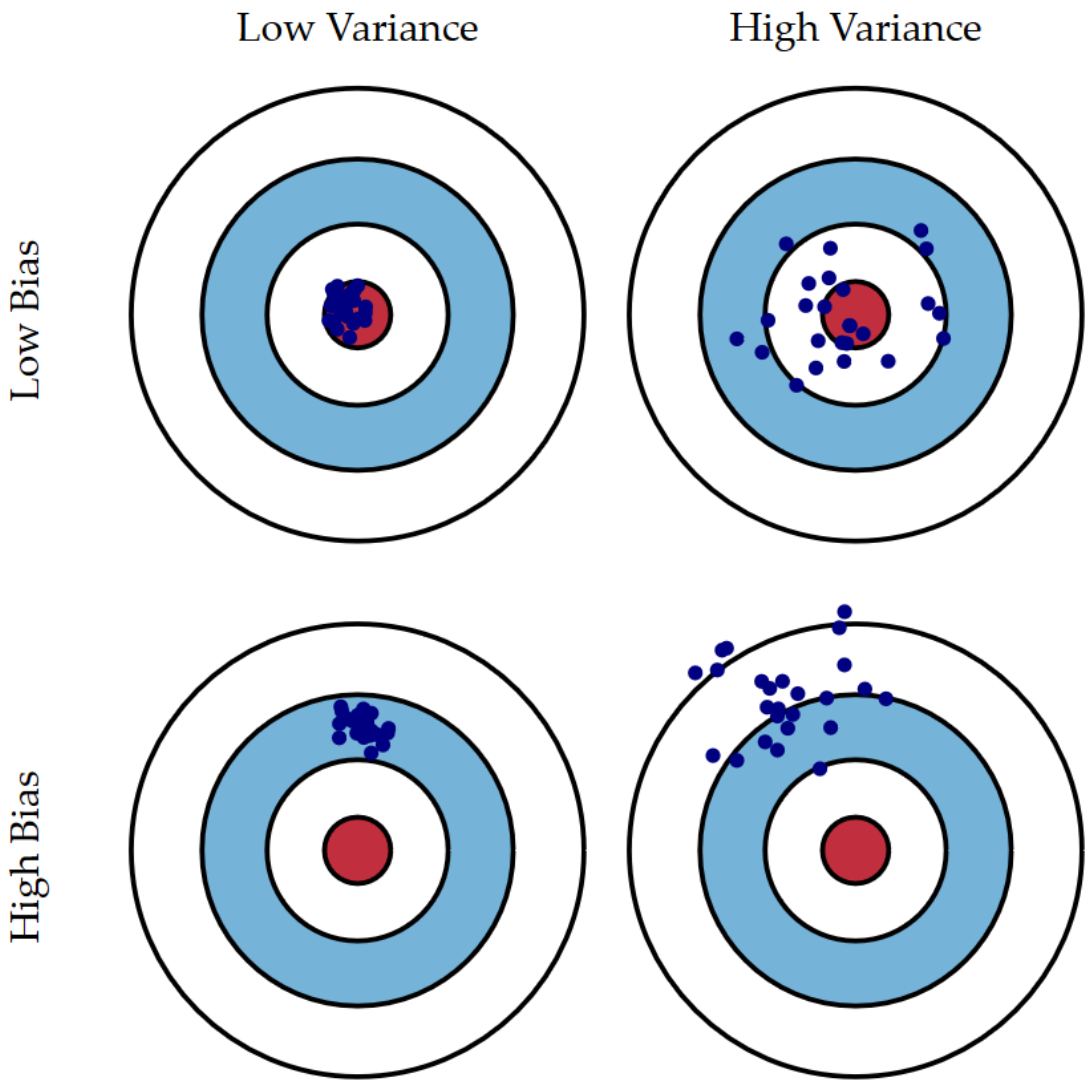
Validate

Validation

- If you have enough data random partition into train/validate/test
- If you do not have enough data
  - Cross validation
- Gives confidence to the results when p-value has little meaning

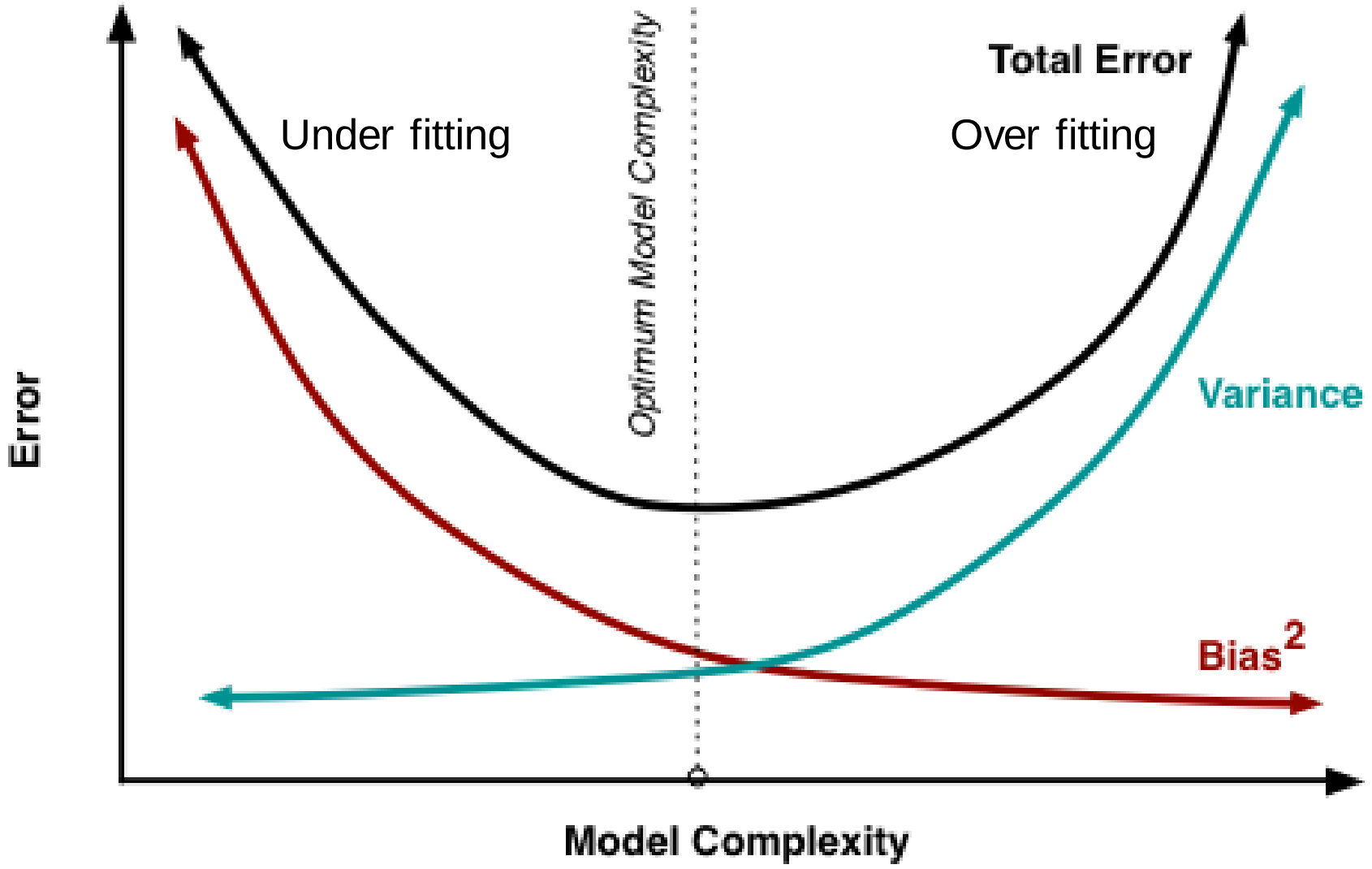
Bias and Variance

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

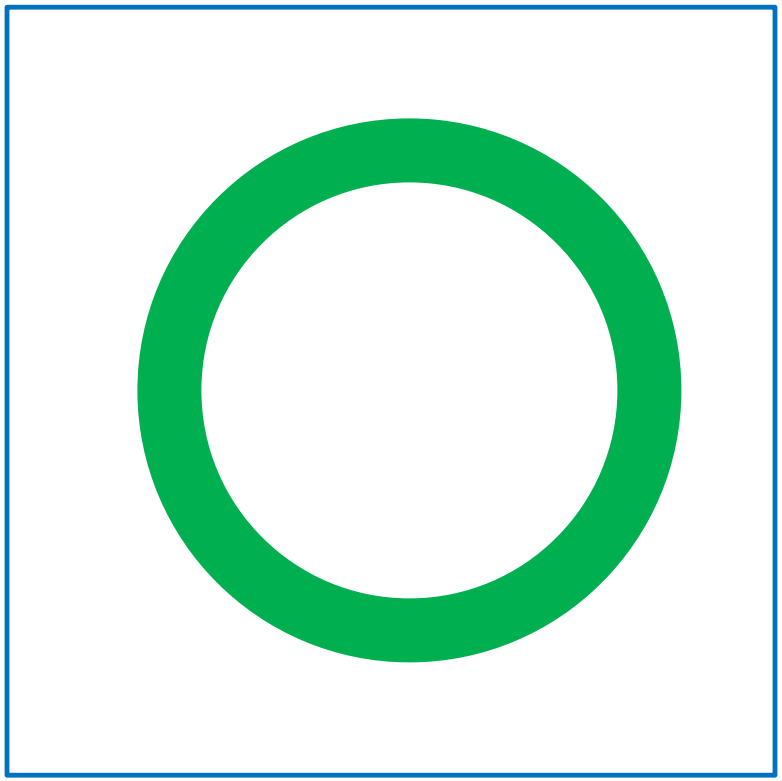


Validation

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



Sometimes models differ between the two approaches. Why ?



Traditional Science Model

Validated from Data Science Model

Comparison

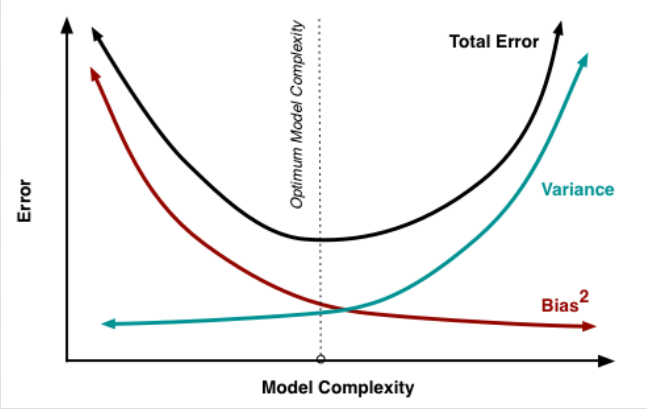
	Traditional Science	Data Science
Common	Use statistics to model from the data points (number of data does matter)	
Focus	Usually more about causation	STRONG correlation
Measurement	• Mostly Based on theory (deductive) decide what to measure - green only • With out seeing the other colors • Slow iterative 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 • 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 the bias
Main issue	Are there enough points to get the full picture?	Is the data clean enough? Is the data representative ? Sensitivity analysis

PUBLIC HEALTH
TEXAS A&M HEALTH SCIENCE CENTER

POPULATION
INFORMATICS
RESEARCH GROUP

Take away in Data Science and Statistics

Both are about understanding/developing general models from many data points (variance & bias). The critical part is good measurement.



PUBLIC HEALTH
TEXAS A&M HEALTH SCIENCE CENTER

POPULATION INFORMATICS RESEARCH GROUP

Take away in Data Science and Statistics

Statistics: Hypothesis driven

Data Science: Iterate between deductive (theory based top down) and inductive reasoning (data based bottom up).

MUST validate. Sensitivity analysis is important

Take away in Data Science and Statistics

Use data science first to develop analytic data sets from secondary data

Then use statistics to run traditional models (causation) OR

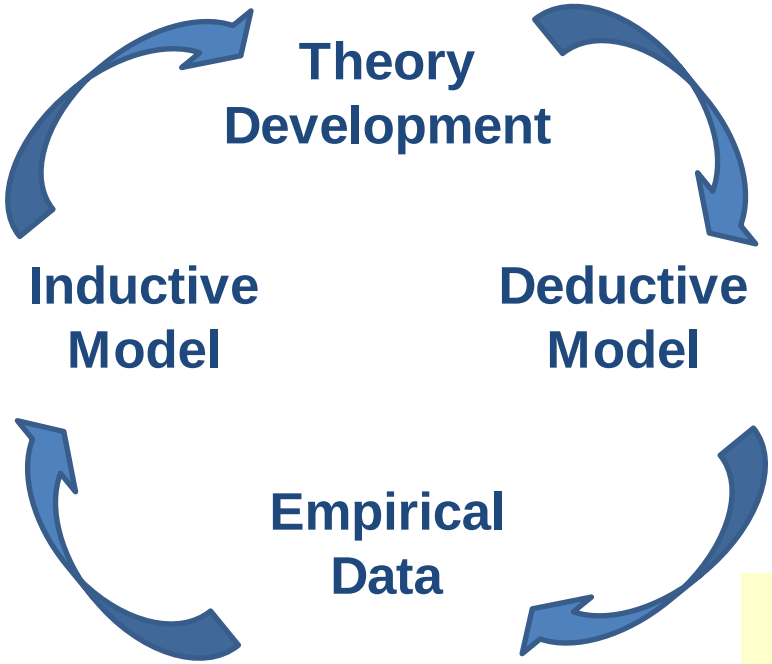
more descriptive models (correlation)

Agenda

- What is Big Data ? What is Data Science ?
- What is Population Informatics & the Social Genome ?
- How is Data Science different from traditional science?
- Conclusion: Doing research with Big Data

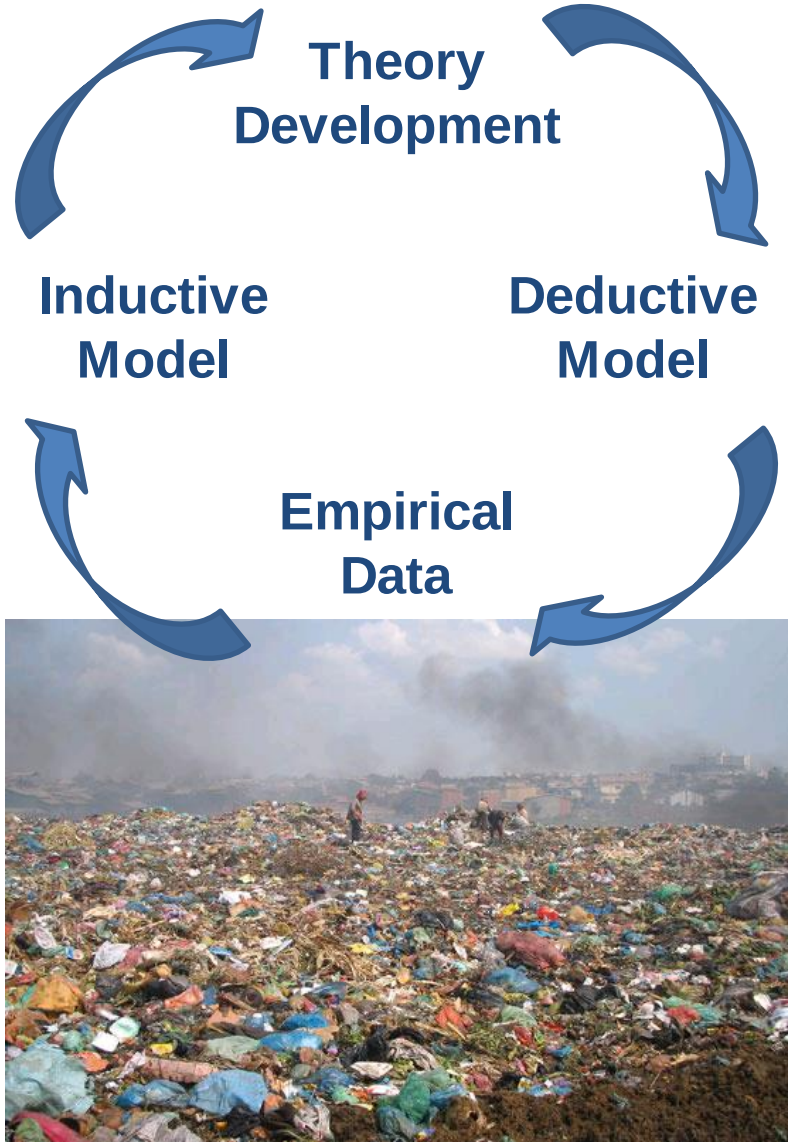
Scientific Process: Traditional

- One research project: one cycle
- Iteratively build on prior work : spiral - SLOW



Minor iterations
Pilot / pretest Data

Scientific Process: Data Science



- One research project: numerous cycles
- Iteratively refine question until satisfied: spiral – FAST progress
  - Agile & accurate movement around the data
- Final answer
  - if required data exist: Data Science
  - if need to collect new data: Traditional Science
  - If possible validate via Traditional Science
  - Use data science to figure out which question should be studied in depth

Big Data Modeling

- NLP: natural language processing (java)
  - Medical notes
  - Social network
- Iterative:
  - Hill Climbing: random start
  - Local vs global max
  - Convergence: converged enough
- Predictive modeling
  - Model drift over time
- Machine learning
- SVM : support vector machine
  - Boundary cases
- Kernel: transform high dimensional space
- Hadoop / map-reduce
  - Parallel processing

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

Big Data
Data Science

Proposed Solution

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.

Problem Statement

Modeling

Controlled Experiments
Survey
Administrative Data

Descriptive Analysis

Administrative Data

Data Science
And / Or
Traditional Science

Refined to tractable (answerable) research questions

Data Science

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