Population Informatics, Data Science, & Big Data

Hye-Chung Kum

Population Informatics Research Group

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

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

Course URL:

http://pinformatics.tamhsc.edu/phpm672

Agenda

- Introduction
  - What is Big Data
  - What is Data Science
  - What is Population Informatics
- Data Science
  - Data vs Theory
  - Doing Analytics Right
  - Challenges
- Examples of population informatics
  - Management Decision Support
  - Evaluation
  - Research

Agenda

- Introduction
  - What is Big Data
  - What is Data Science
  - What is Population Informatics
- Data Science
  - Data vs Theory
  - Doing Analytics Right
  - Challenges
- Examples of population informatics
  - Management Decision Support
  - Evaluation
  - Research

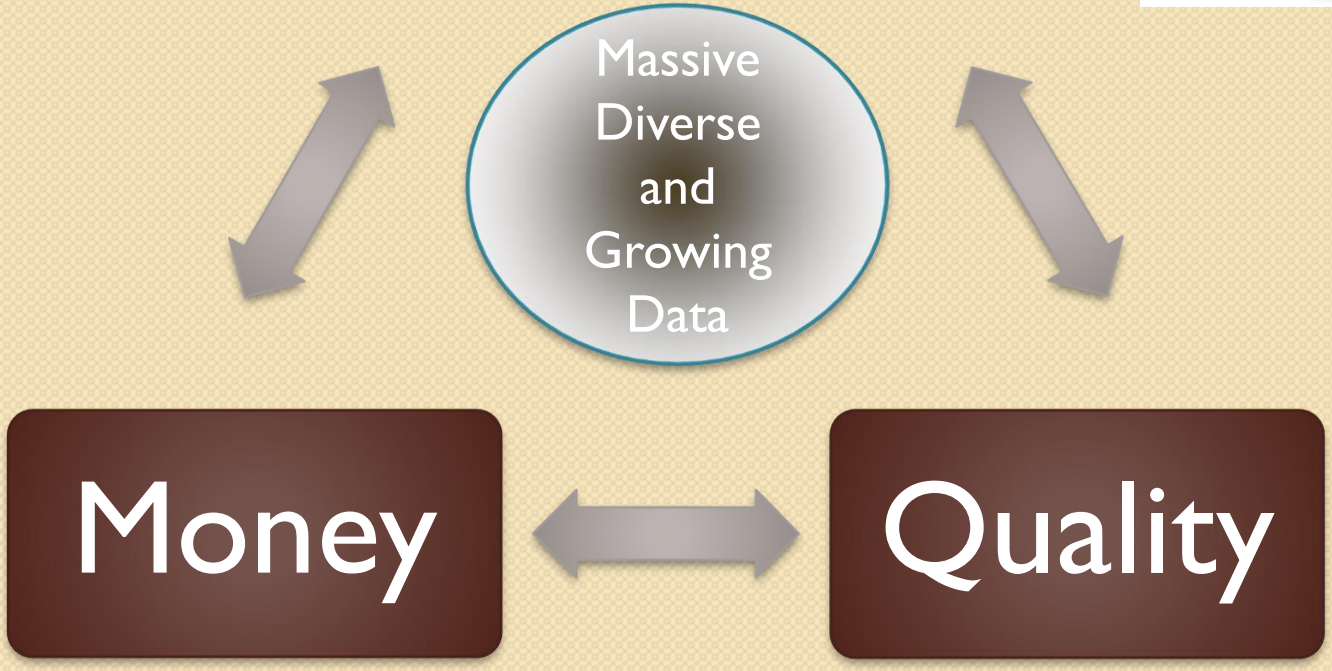
Properties of BIG DATA : 4V

- Volume : constantly generating
- Velocity : constantly changing
- Variety : expressed in many ways
- Veracity : lots of errors

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

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

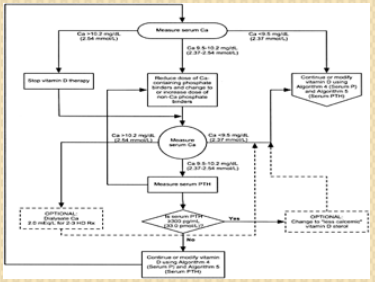
Something's gotta give:



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

Agenda

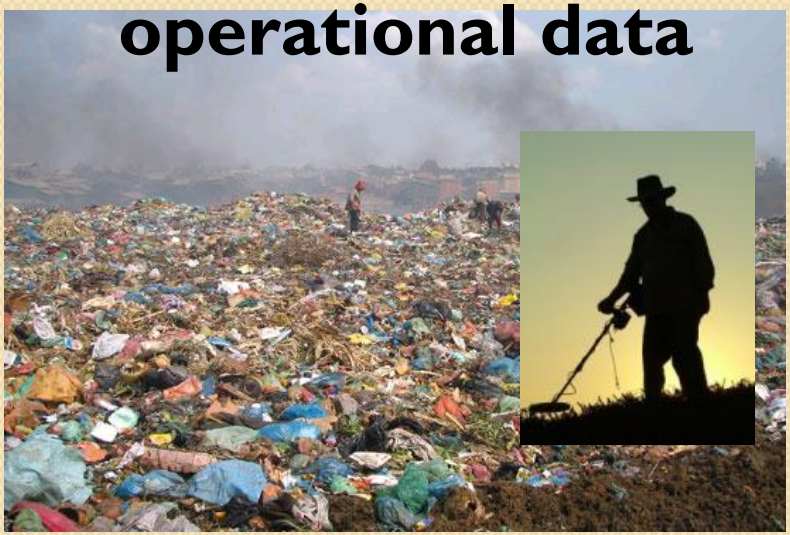
- Introduction
  - What is Big Data
  - What is Data Science
  - What is Population Informatics
- Data Science
  - Data vs Theory
  - Doing Analytics Right
  - Challenges
- Examples of population informatics
  - Management Decision Support
  - Evaluation
  - Research

What is Data Science?

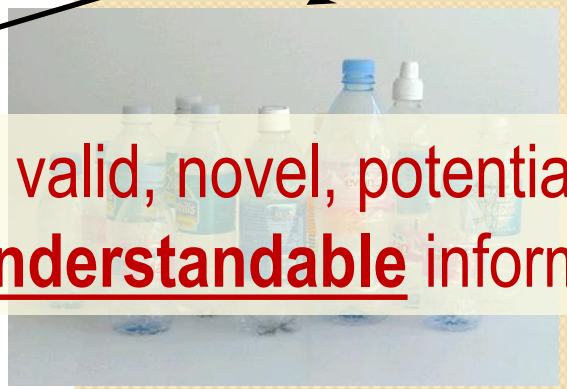
- Other words
  - Knowledge Discovery & Data mining (KDD)
  - Business Intelligence / Business Analytics
- Collecting and refining information from many sources
- Analyzing and presenting the information in useful ways
- Iterating between inductive and deductive reasoning to get to the truth
- So people can make better business decisions

Data Science
Knowledge Discovery & Data mining (KDD)

Big Data :
operational data

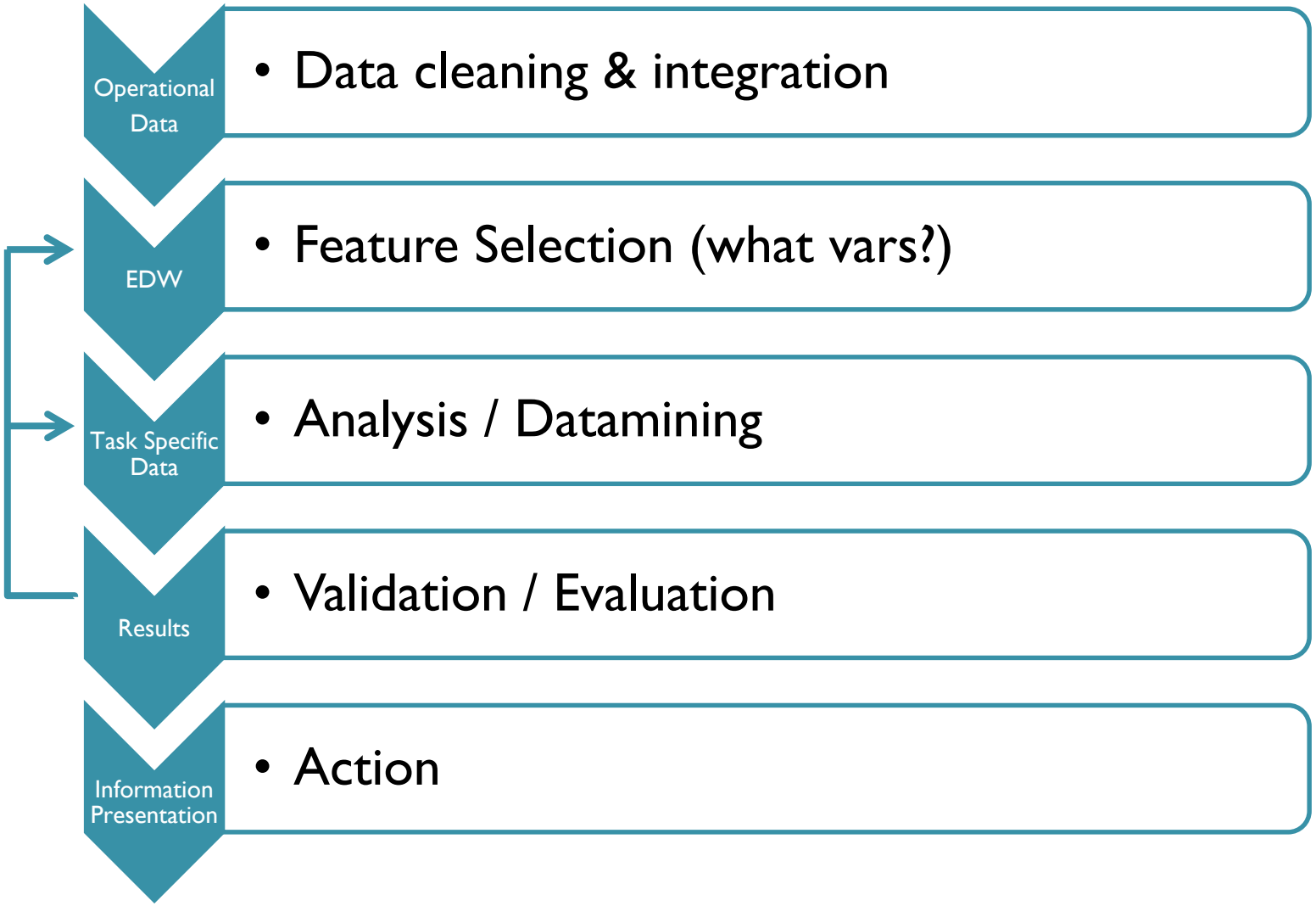


KDD
Clean, Merge, Reprocess

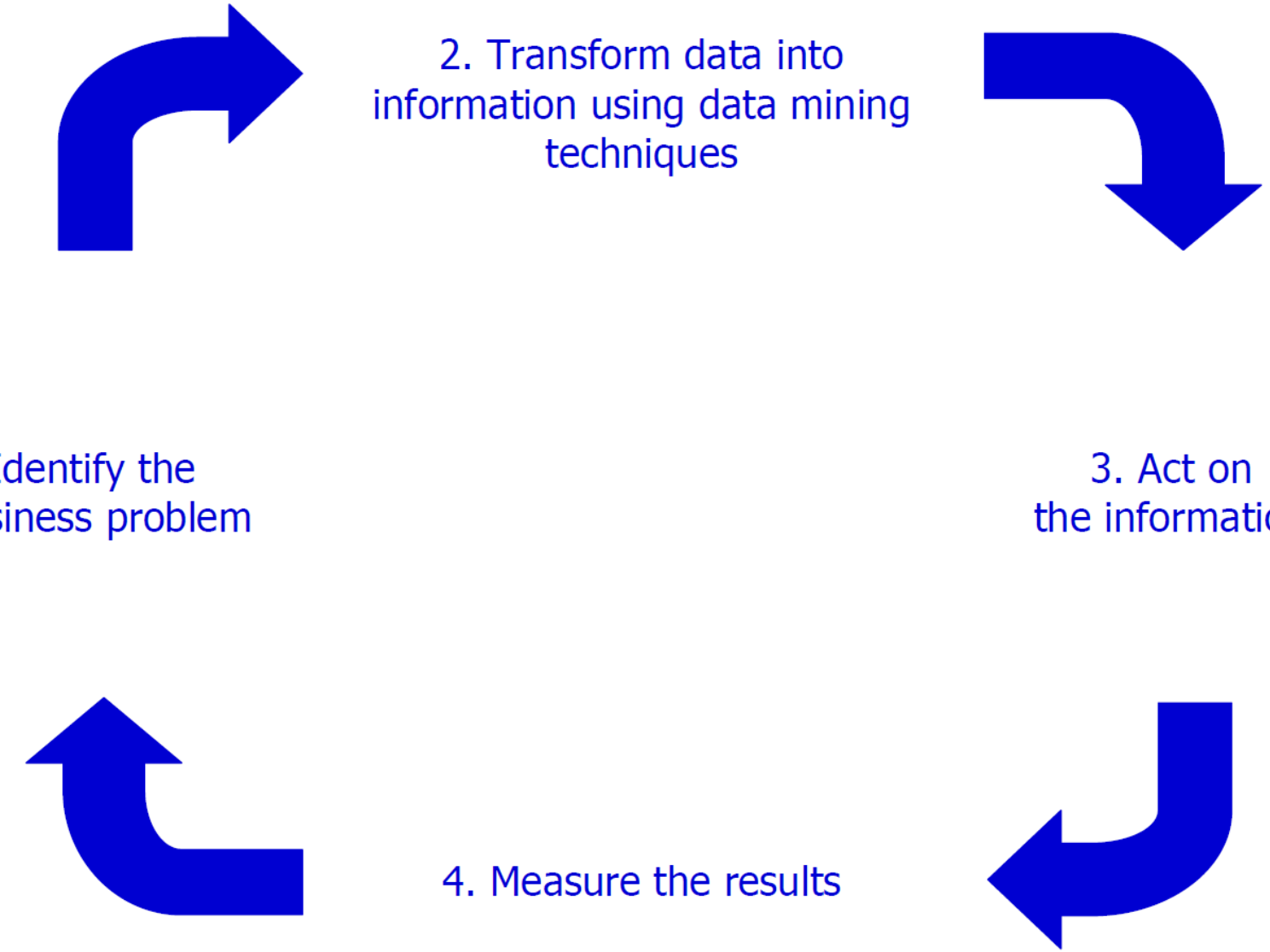


Human consumable, valid, novel, potentially **useful**,
and ultimately **understandable** information

KDD Process



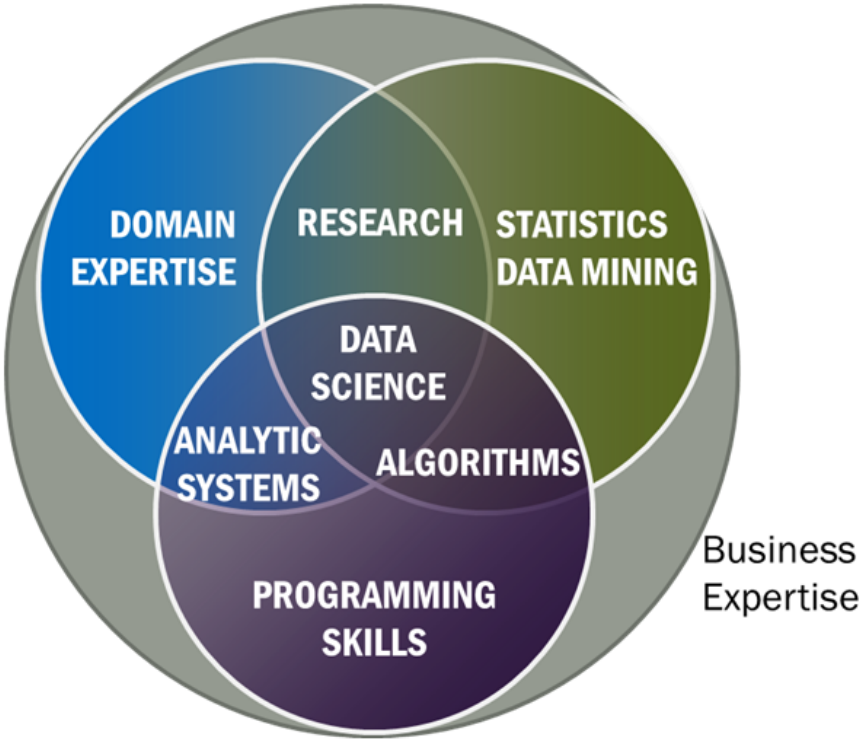
The Virtuous Cycle of Data to Decision & Action



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



Agenda

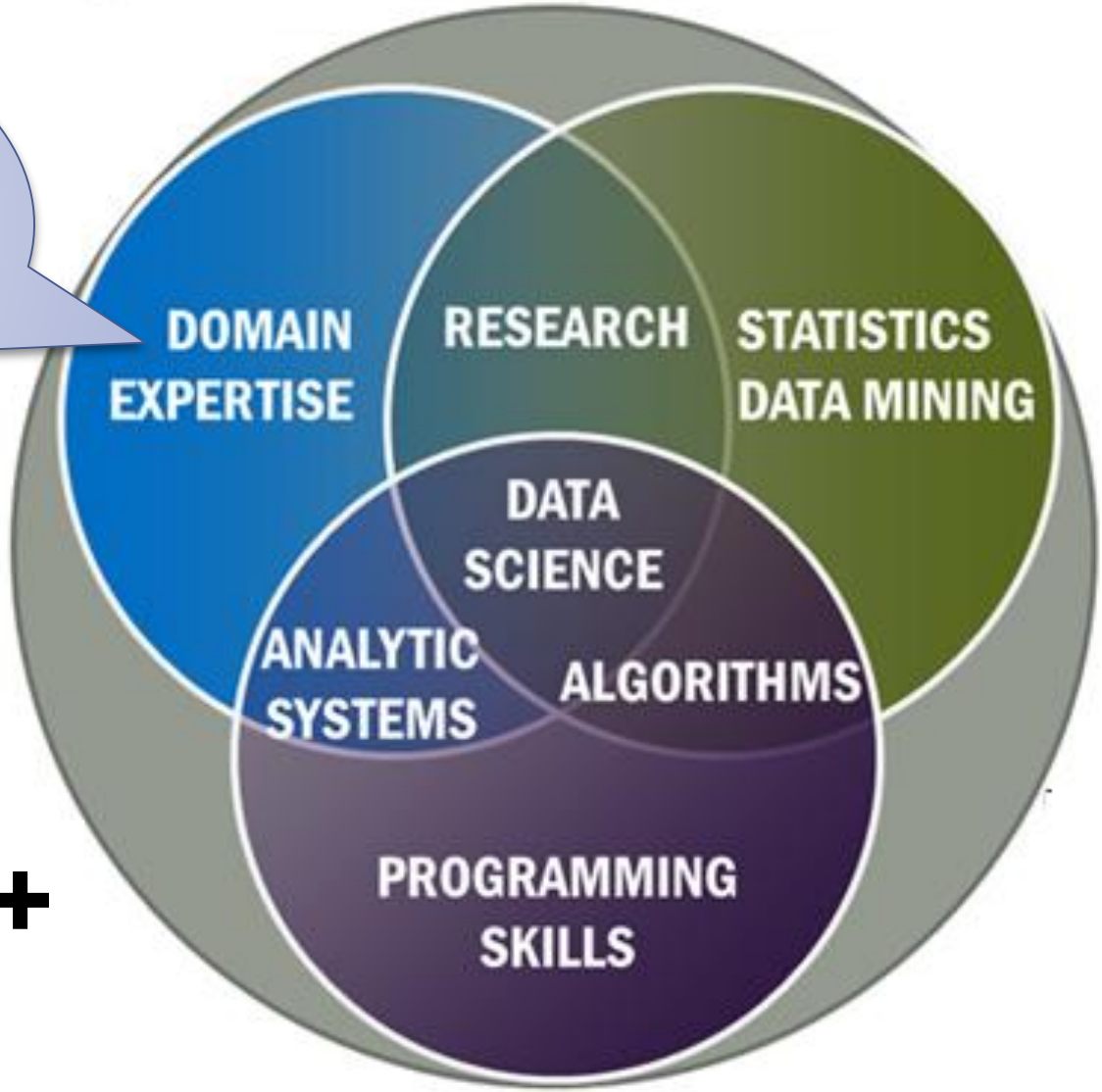
- Introduction
  - What is Big Data
  - What is Data Science
  - What is Population Informatics
- Data Science
  - Data vs Theory
  - Doing Analytics Right
  - Challenges
- Examples of population informatics
  - Management Decision Support
  - Evaluation
  - Research

Bioinformatics

Apply Data Science to Human Genome Data

Biology

+

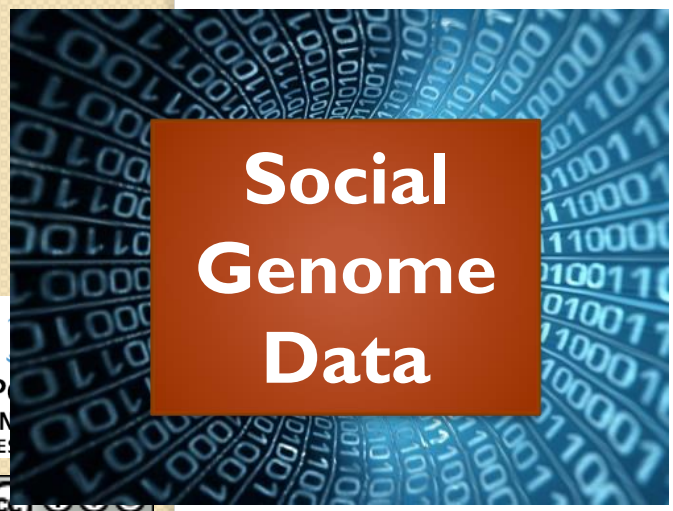


Population informatics

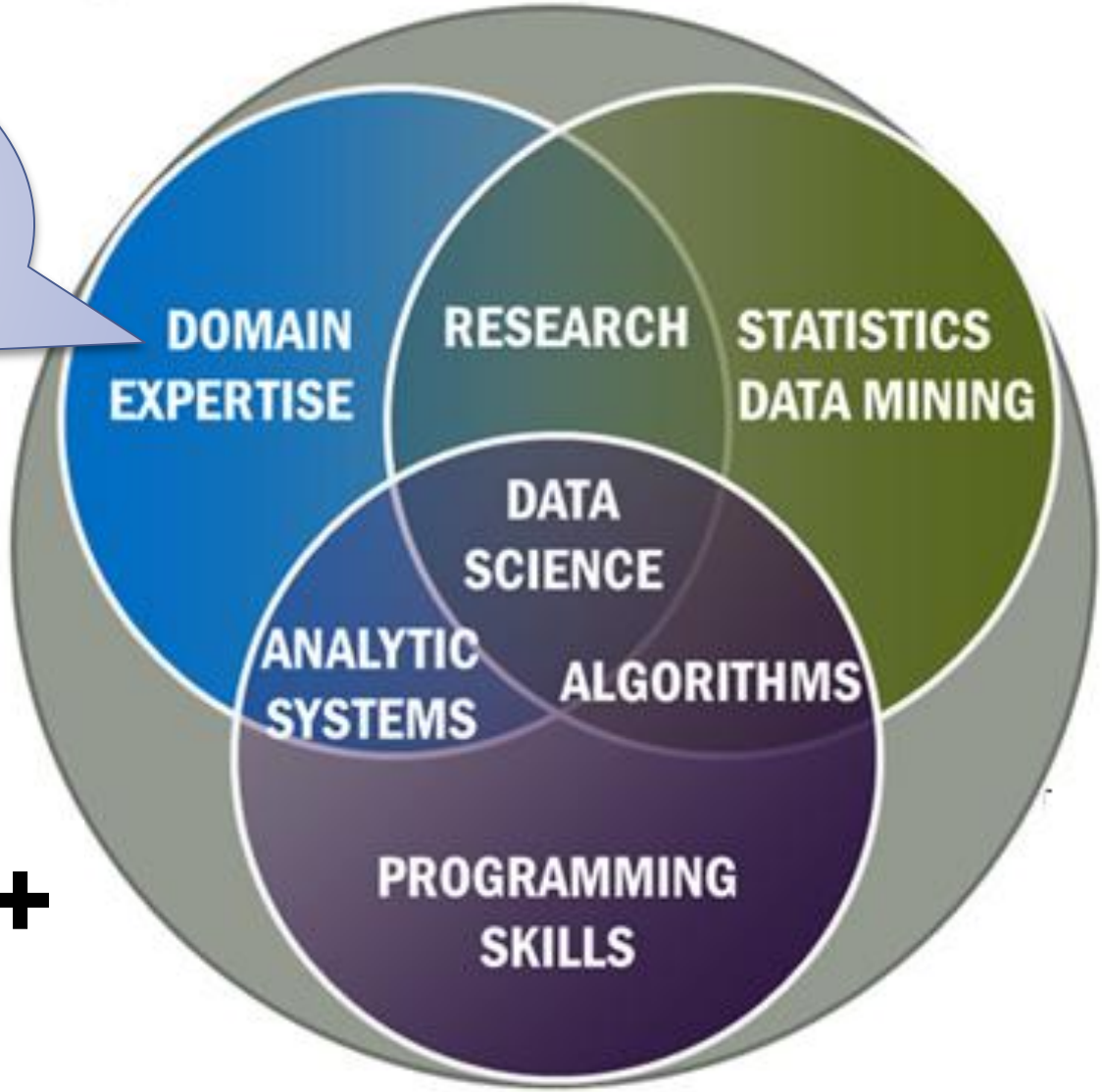
Apply Data Science to Social Genome Data

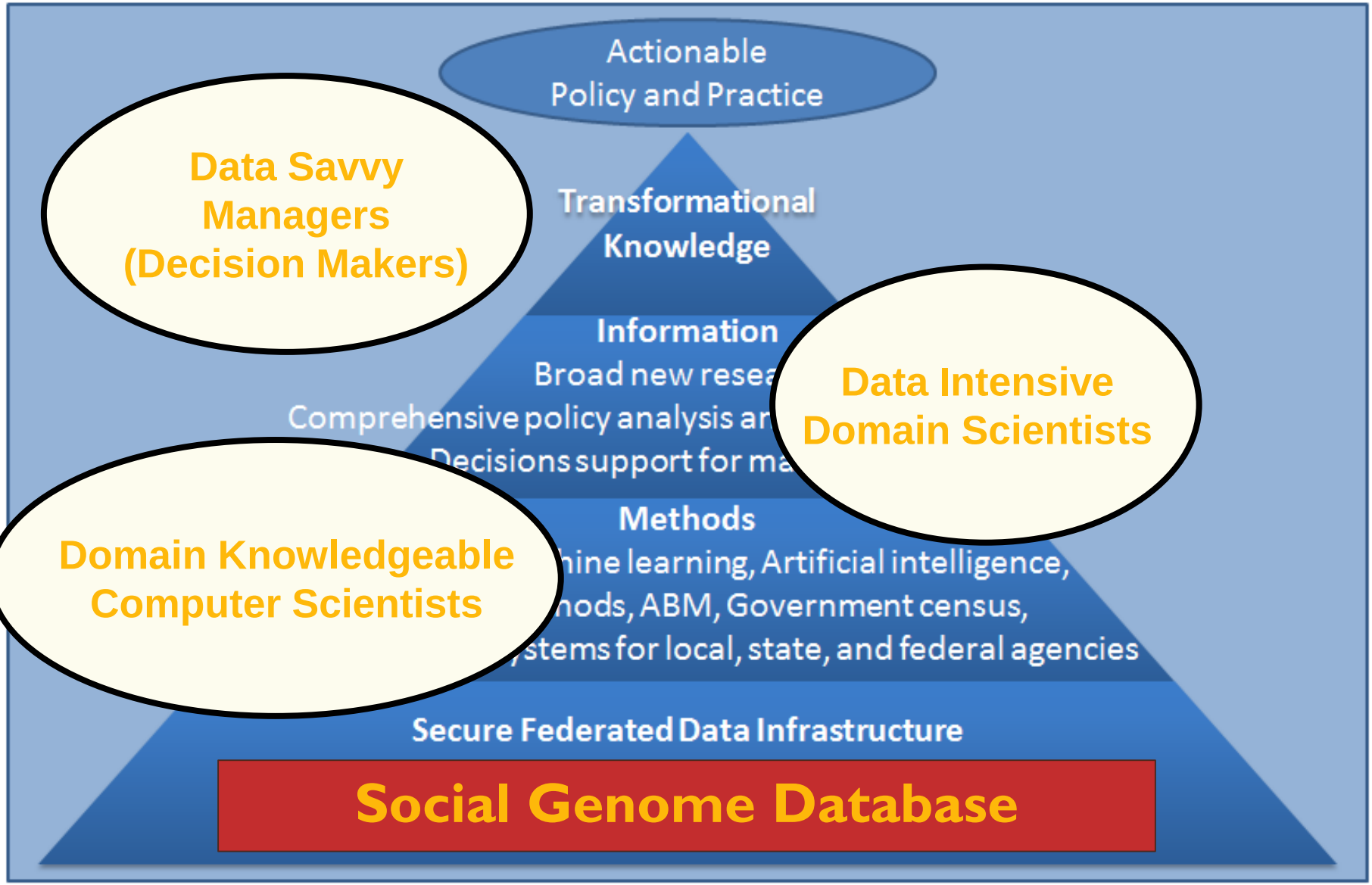
Studies of society (groups of people)

- Social sciences
- Health sciences (population health)



+



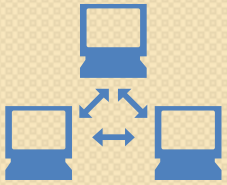


UNC POPULATION INFORMATICS RESEARCH GROUP

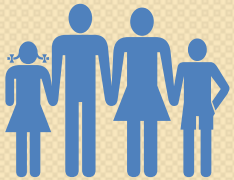
Population Informatics

Hye-Chung Kum, Population Informatics Research Group
Dept. of Computer Science, UNC-CH, http://pinformatics.web.unc.edu/

The Power of the Social Genome

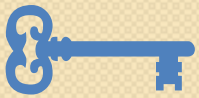


Our activities from birth until death leave **digital traces** in large databases



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

The social genome data are **massive** and **chaotic**



It holds **crucial insights** into many of the most challenging problems facing our society (i.e. affordable and accessible quality healthcare, economics, education, employment, and welfare)

Social Genome

- Data-intensive research using distributed, federated, person-level datasets in near real time has the potential to transform social, behavioral, economic, and health sciences—but issues around privacy, confidentiality, access, and data integration have slowed progress in this area. When technology is properly used to manage both privacy concerns and uncertainty, big data will help move the growing field of population informatics forward.

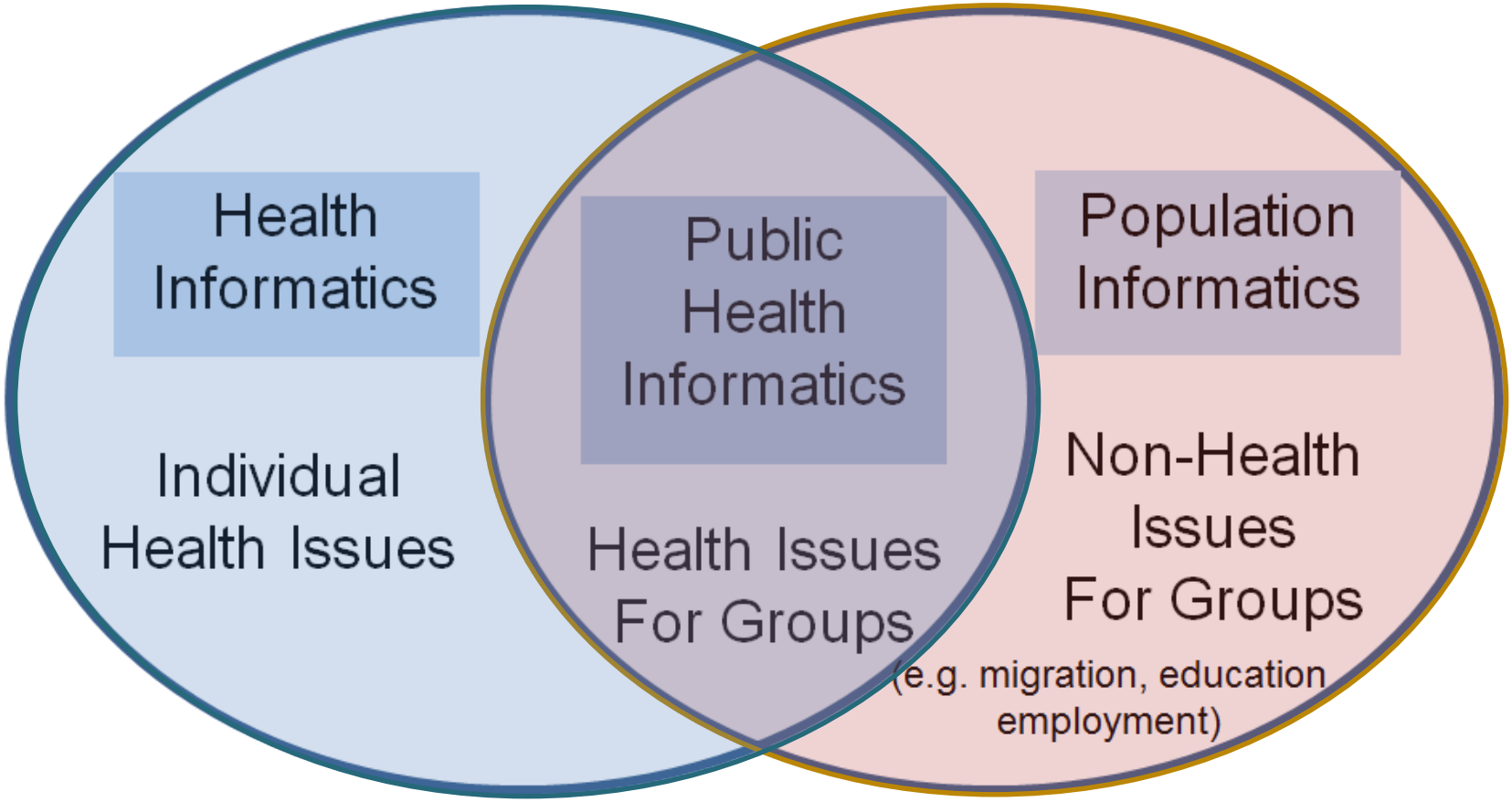
Thomas Davenport Competing on Analytics

- 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

?? Informatics



Public Health Informatics

- Yasnoff et al. (2000): Public health informatics is defined as the systematic application of information and computer science and technology to public health practice, research, and learning

Public Health Informatics Competencies

- Public Health Informatics Competencies Working Group, CDC
- Report: Informatics competencies for public health professionals
  - *Class 1: Effective use of information*
  - *Class 2: Effective use of information technology*
  - *Class 3: Effective management of information technology projects*
  - http://www.nwcphp.org/docs/phi/comps/phic_web.pdf

Public Health Informatics Competencies

- Describe at a basic level the fundamentals of a computer network
- Describe at a basic level the Internet and World Wide Web
- Describe at a basic level technologies employed to ensure computer systems' security
- Public Health Informatics Competencies Working Group

- Cunningham et al. 2007. Baseline Assessment of Public Health Informatics Competencies in Two Hudson Valley Health Departments. Public Health Reports. May-Jun 2007. V122

Agenda

- Introduction
  - What is Big Data
  - What is Data Science
  - What is Population Informatics
- Data Science
  - Data vs Theory
  - Doing Analytics Right
  - Challenges
- Examples of population informatics
  - Management Decision Support
  - Evaluation
  - Research

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

- "All models are wrong but some are useful"
- The Petabyte Age is different because more is different
  - Google translate
- Google's founding philosophy is that we don't know why this page is better than that one: If the statistics of incoming links say it is, that's good enough. No semantic or causal analysis is required. Correlation is enough.

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

- But faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete.
- Biology: Hutchinson disease
- Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.
- What can science learn from Google?

Agenda

- Introduction
  - What is Big Data
  - What is Data Science
  - What is Population Informatics
- Data Science
  - Data vs Theory
  - Doing Analytics Right
  - Challenges
- Examples of population informatics
  - Management Decision Support
  - Evaluation
  - Research

There are ways to do analytics right Value in Big Data

- 63 percent of healthcare executives in the federal government believe that big data will improve population health management.
- The Mckinsey report on big data valued integrated data on patient services at $300 billion.
- But so few have proper goals and strategies for their data

There are ways to do analytics right It’s not the data, but the people

- It's not about the data, but how you're going to manage it.
- The focus, according to Hughes, should be on information management, including data governance, stewardship and quality.
  - If you are just about grabbing data, you will be on a data grab forever.
- It may sound impressive to say that your organization has access to terabytes of patient information, but without robust technology and smart people to manipulate it, that data is simply words and numbers without context
- A severe shortage of analytics pros makes navigating this landscape all the more difficult
- “It's also a mistake to think you can staff up on this easily"
- lack of qualified data engineers

There are ways to do analytics right Data Management is Key

- Raw data from claims or from an EMR database are not suitable for analysis.
- Turning raw data into usable information requires preparation, including normalization and validation.
- Only then can an organization gain trustworthy insights from the information and put it to use in maximizing patient care, reducing risk and strengthening a business's bottom line

There are ways to do analytics right EDW: Organized Data Library

- Hughes says that organizations have been spending too much time and money on enterprise data warehouses, which he sometimes refers to as "data landfills."
- An EDW isn't where data goes to die. An EDW is a staging point for analytics.
- An EDW needs to be easy for clinicians to understand and interpret, and also needs to interoperate with and push data back out to other systems
- Sometimes this is done in too fragmented a fashion
- My thoughts: Art
  - Appropriate size task, data
  - Balance organizing data with actual using data

Agenda

- Introduction
  - What is Big Data
  - What is Data Science
  - What is Population Informatics
- Data Science
  - Data vs Theory
  - Doing Analytics Right
  - Challenges
- Examples of population informatics
  - Management Decision Support
  - Evaluation
  - Research

Challenges

- Privacy
- Data Access
- Data Management
  - Data Integration & Cleaning
- Error Management and Propagation

What do we know about information privacy

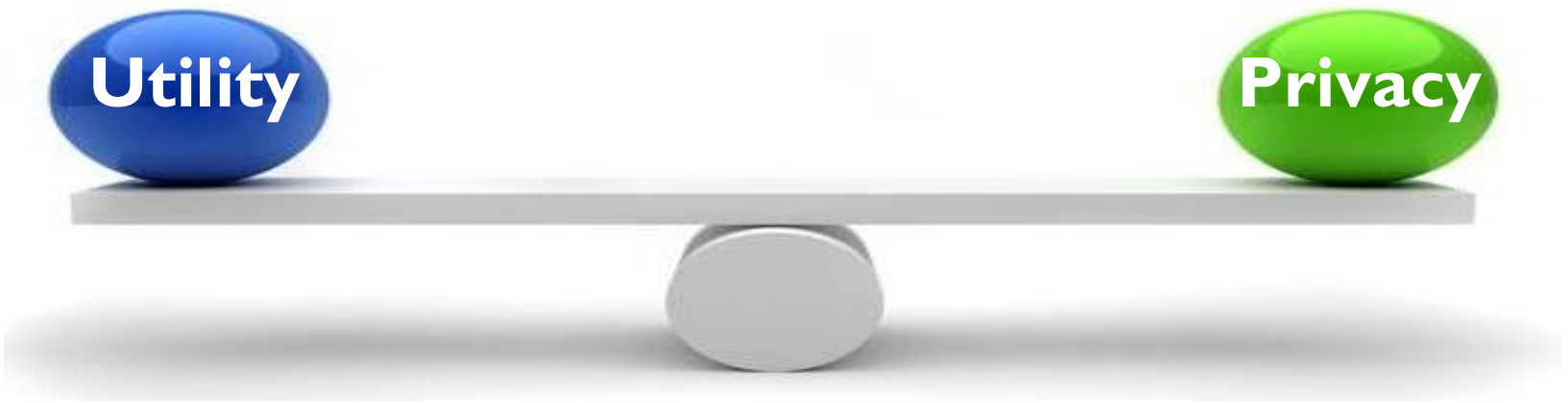
Information Accountability Works

- 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.

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



Social Issues: Balance between

- Individual privacy
  - Secrecy does not work very well : accuracy of data
- Cost of integrity of data
  - Bias in research: selective sampling
  - Incorrect analysis that can lead to wrong decisions
- Organization transparency & accountability
- Freedom of speech
  - Marketing is freedom to express why one should prescribe certain drugs
  - Marketing is freedom to send junk mail & call
  - Thus, getting more information to better target is acceptable and should be allowed

Privacy as contextual integrity (legal)

- Helen Nissenbaum (NYU Law School)
- *Washington Law Review, Vol. 79, No. 1, 2004*
- a conceptual framework for understanding privacy expectations and their implications developed in the literature on law, public policy, and political philosophy
- Privacy Protection / Violation
  - Social norms of expectation (on use, sharing etc)
  - Due diligence
  - Quantifying harm : loss of job

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**

Privacy-by-Design

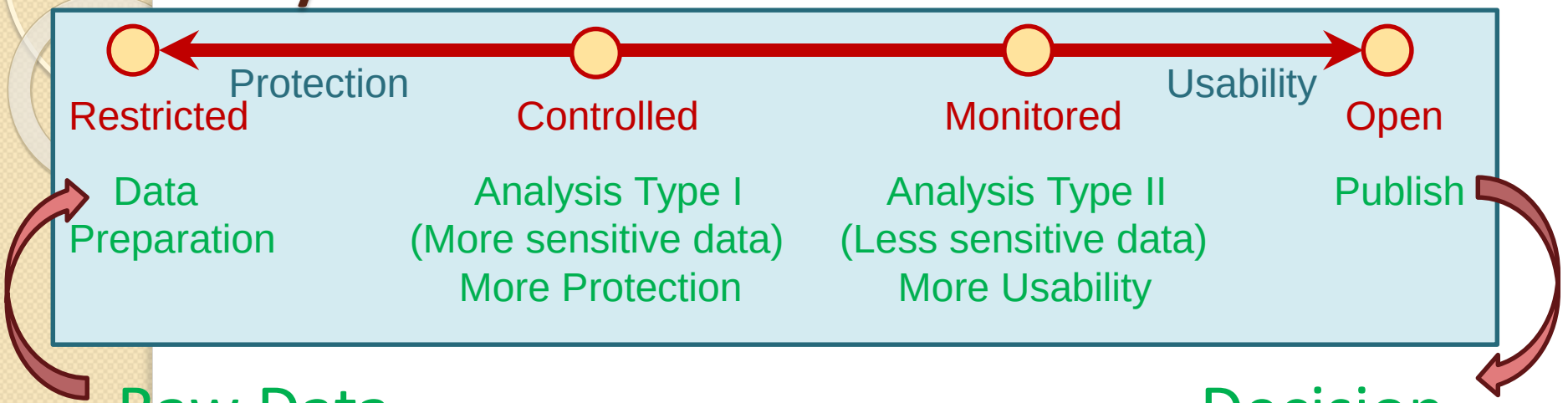
- A different perspective on privacy and research using personal data
- Personal Data is Delicate/Hazardous/Valuable
- Important to have proper systems in place that give protection but allow for continued research in a safe manner
- All hazardous material need standards
  - Safe environments to handle 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

WORKFLOW (Data Access)

Safe Platform for Data to Decision

POPULATION INFORMATICS RESEARCH GROUP

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

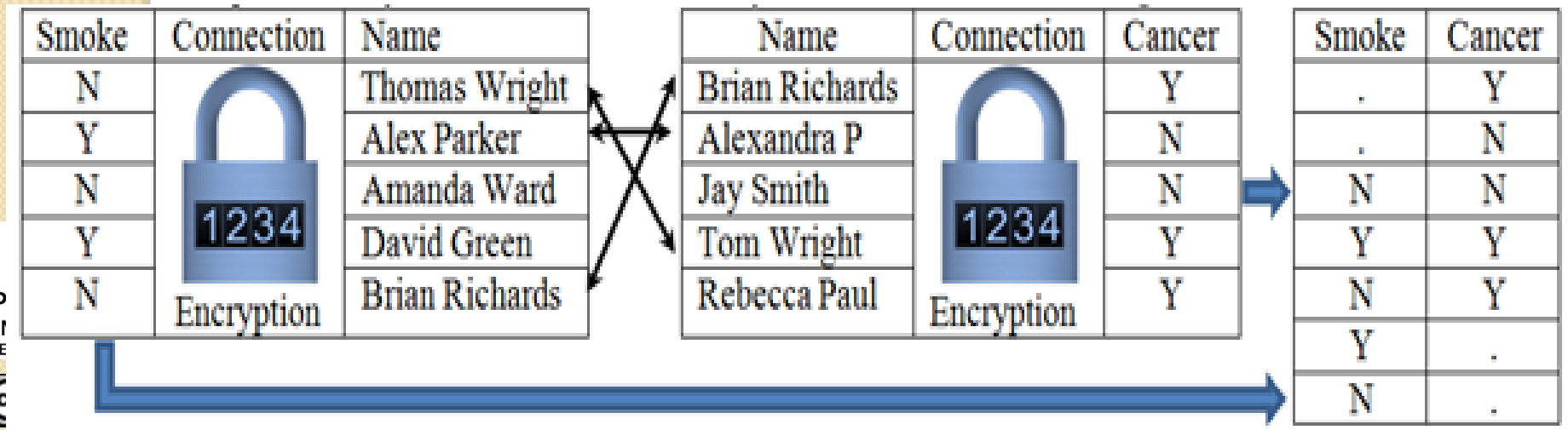
Restricted Access : Prepare the customized data

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

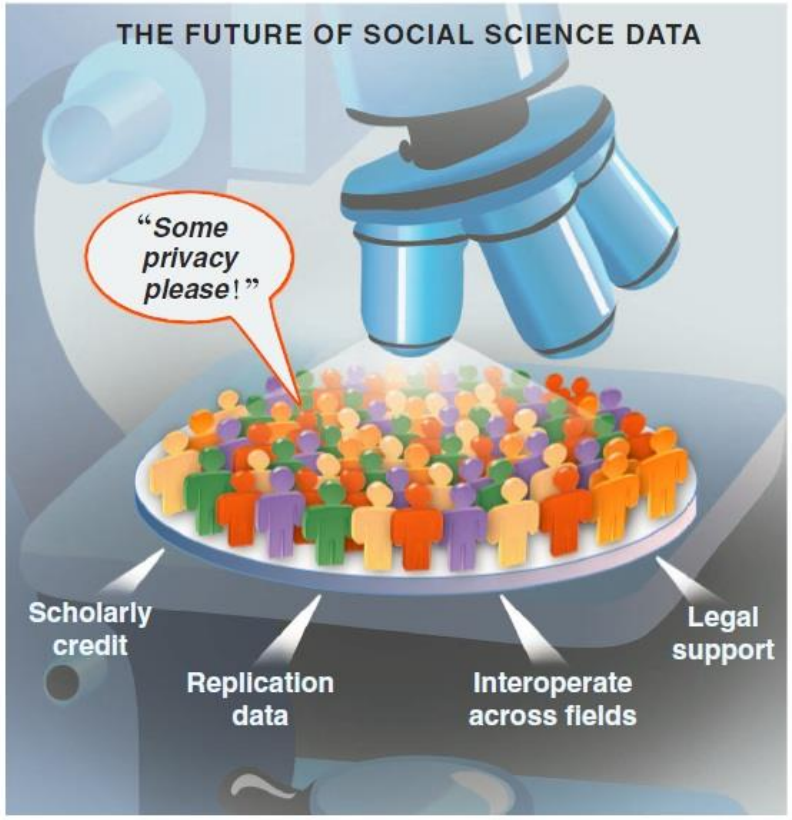
Privacy Preserving Interactive Record Linkage

- Decouple data via encryption
- Automated honest broker approach via computerized third party model
- Chaffe to prevent group disclosure

Kum, H.C., Krishnamurthy A., Machanavajjhala A., Reiter M., and Ahalt S. **Privacy Preserving Interactive Record Linkage (PPIRL).** J Am Med Inform. Assoc. 2014;21:212–220. doi:10.1136/amiajno-2013-002165



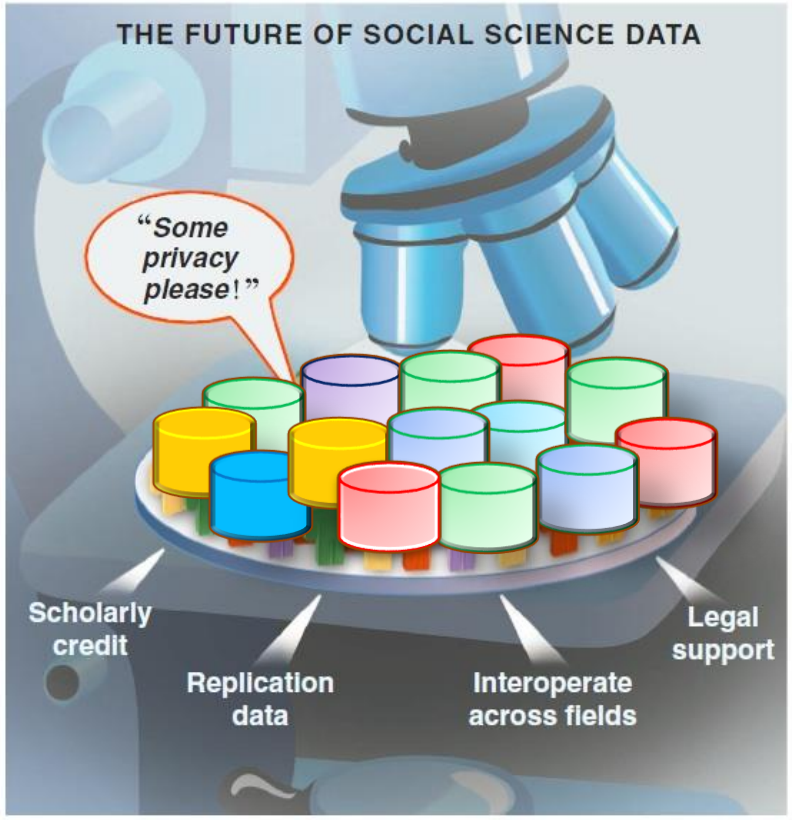
Controlled Access : Model using given tools



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

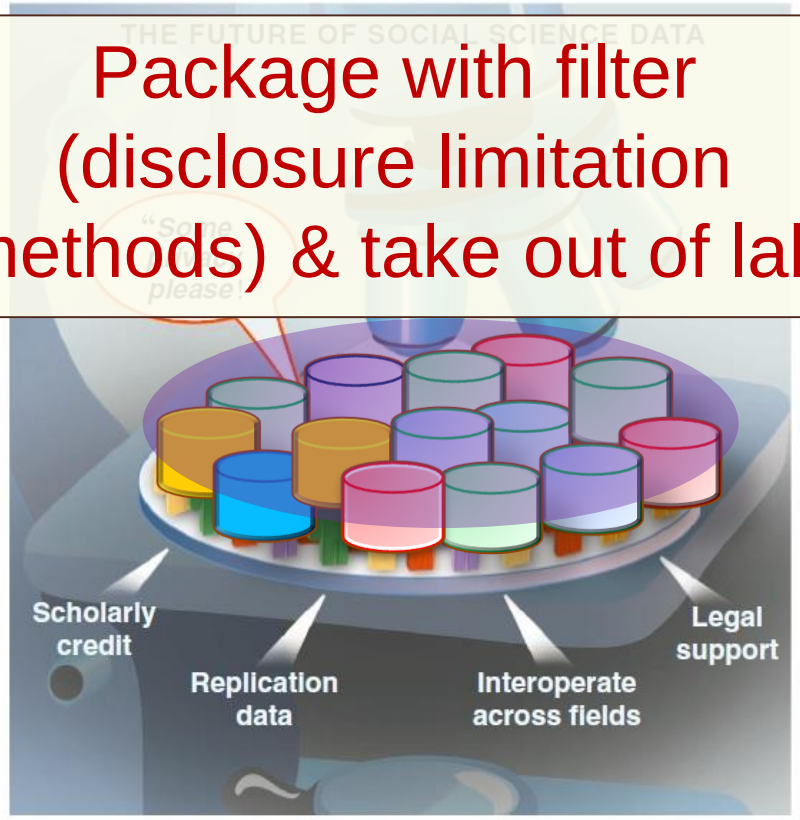


Gary King. Ensuring the Data-Rich Future of the Social Sciences, Science, vol 331, 2011, pp 719-721.

- 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

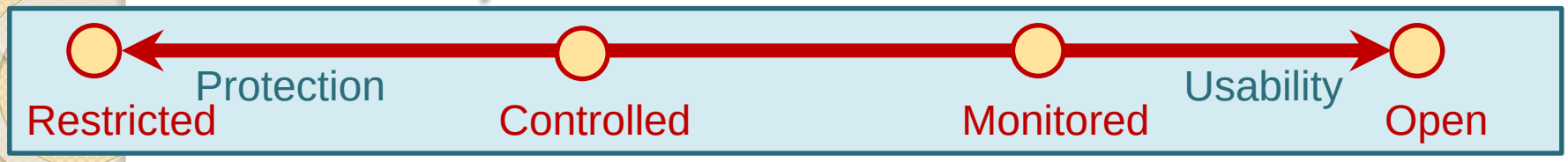
Package with filter (disclosure limitation methods) & take out of lab



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

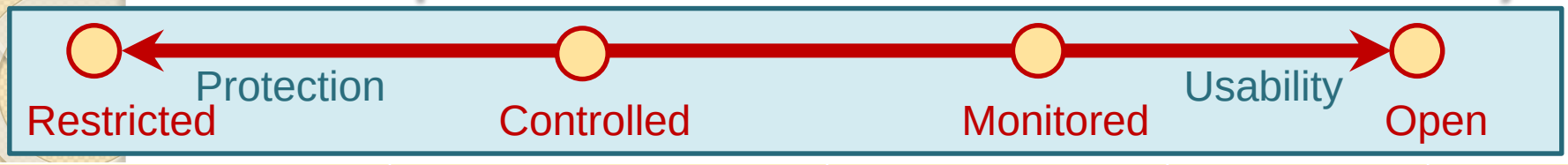
- **Anyone : Publish information for others**
- No IRB
- No monitoring use
- Publish data use terms
- Disclosure Limitation Methods (filter)
- Sanitized data
- Public websites, publications

Privacy Protection Mechanism



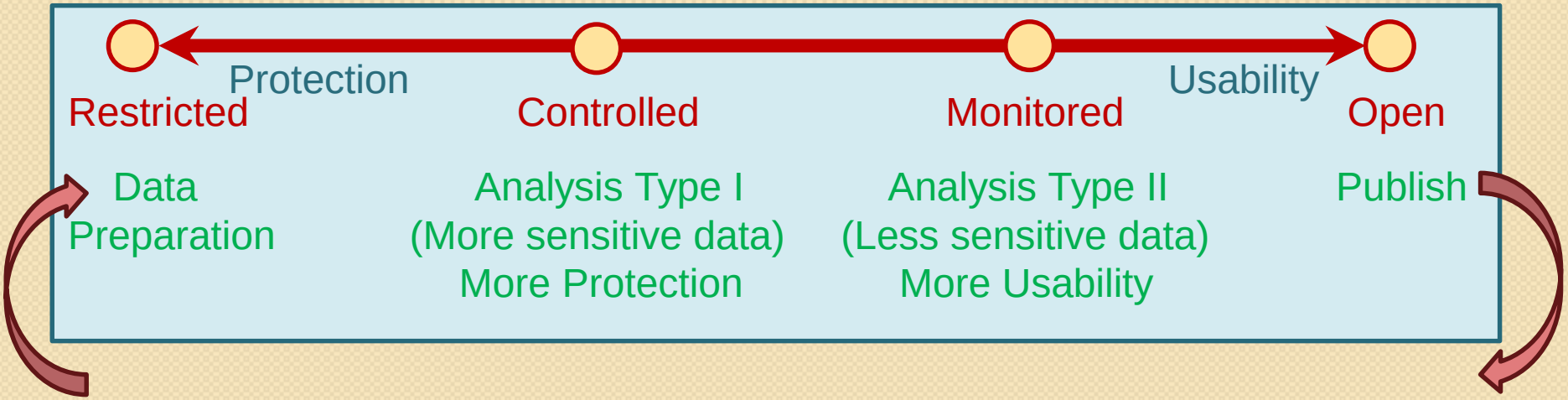
Access	Restricted Access	Controlled Access	Monitored Access	Open Access
Protection Approach	Physical restriction to access	Lock down VM (limit what you can do on the system)	Information accountability	Disclosure Limitation
Monitoring Use	All use on & OFF the computer is monitored	All use on the computer 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

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

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) at some point
- Primary issues: **Privacy (safe data access, code of conduct), data integration, error management**,
  - Velocity, variety, veracity, volume (lots of SMALL datasets)

Agenda

- Introduction
  - What is Big Data
  - What is Data Science
  - What is Population Informatics
- Data Science
  - Data vs Theory
  - Doing Analytics Right
  - Challenges
- Examples of population informatics
  - Management Decision Support
  - Evaluation
  - Research

Examples of population informatics

- Management Decision Support
  - Online Dashboards
  - Open data
- Evaluation
- Research
- Infographics
  - http://research.tamhsc.edu/pinformatics/data-science/
  - video