Approaches to enhancing privacy when conducting data science about people

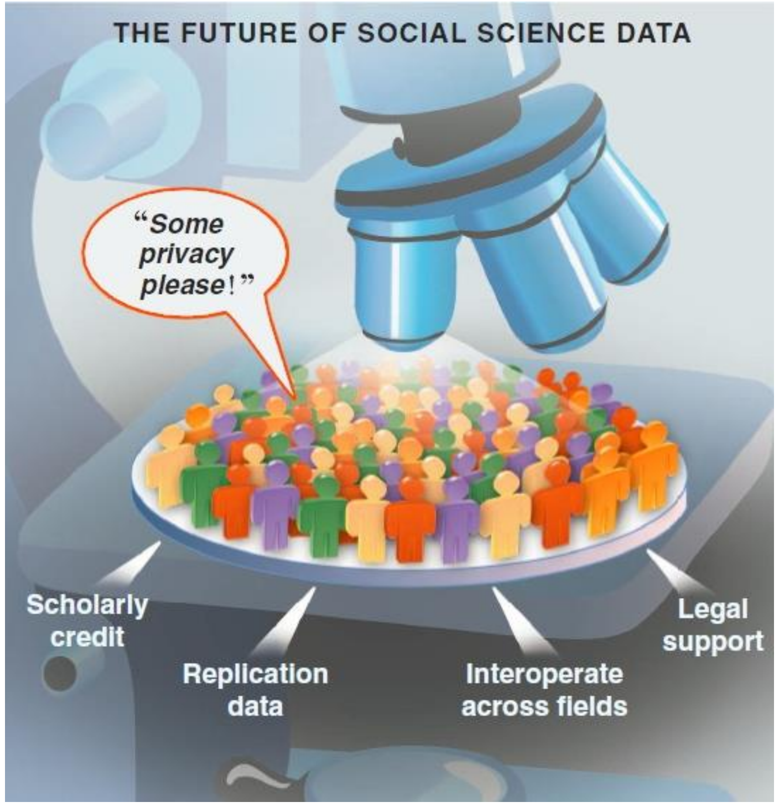
Case Study using Record Linkage

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

Data Science & Information Privacy Barrier: Data Sharing

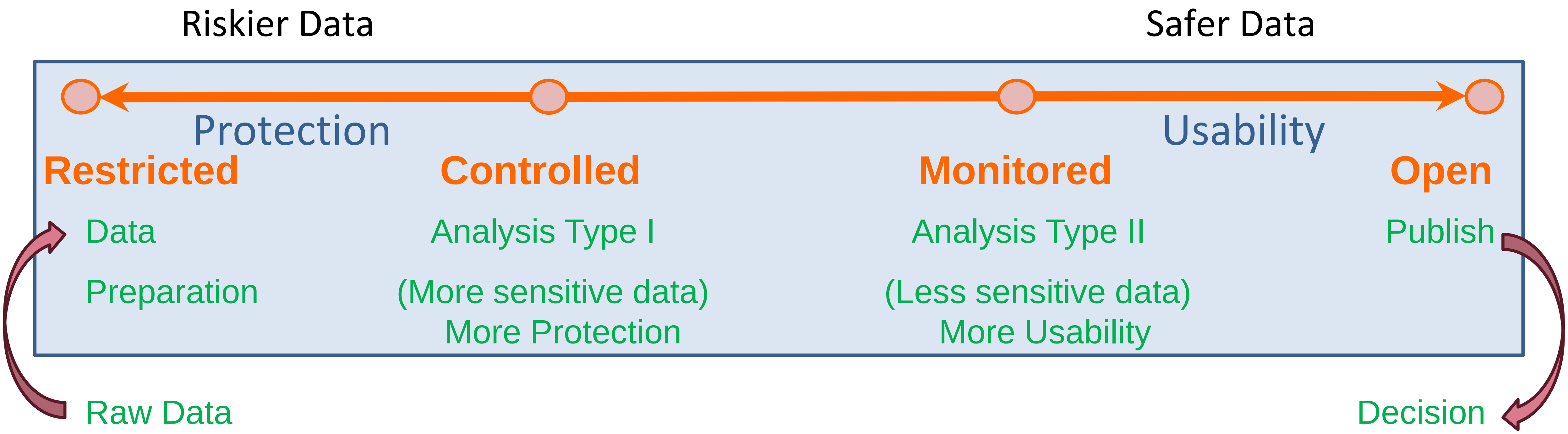
How do we conduct responsible research?

- Human interaction required for high quality data
  - Concerns about privacy
- A holistic approach
  - Data governance
  - Public engagement
  - Technical



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

System of Access Models



Riskier Data

Safer Data

Protection

Usability

Restricted

Data

Preparation

Controlled

Analysis Type I

(More sensitive data)
More Protection

Monitored

Analysis Type II

(Less sensitive data)
More Usability

Open

Publish

Raw Data

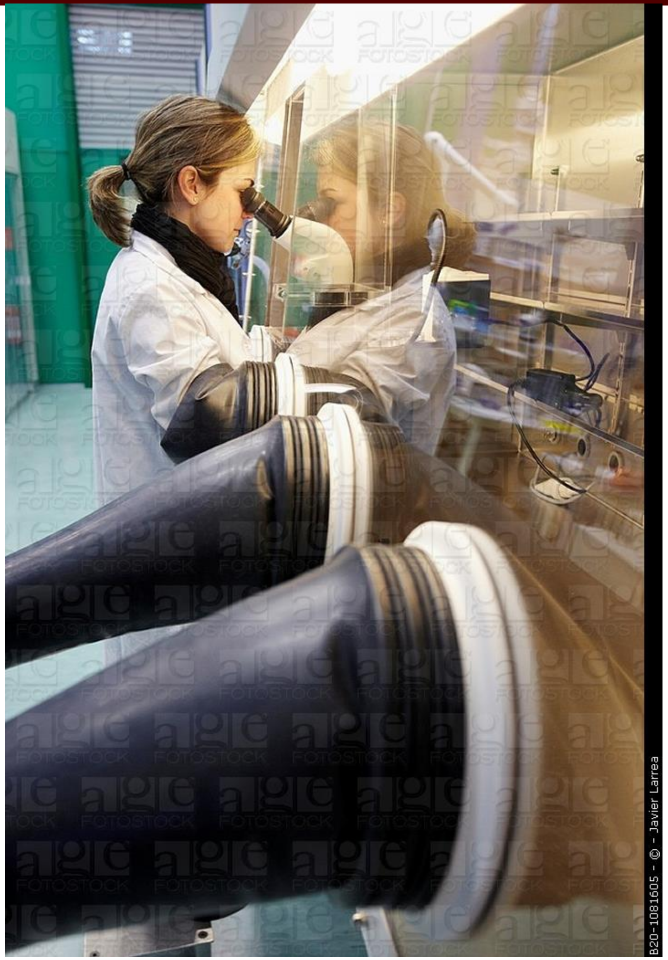
Decision

- 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 do you want to do with them
  - Determine access levels for each data
- Submit to IRB process

Restricted Access : Prepare the customized data
Limit access to only records that need manually review



B20-1081605 - © - Javier Larrea

- 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)

Controlled Access :
Model using given tools

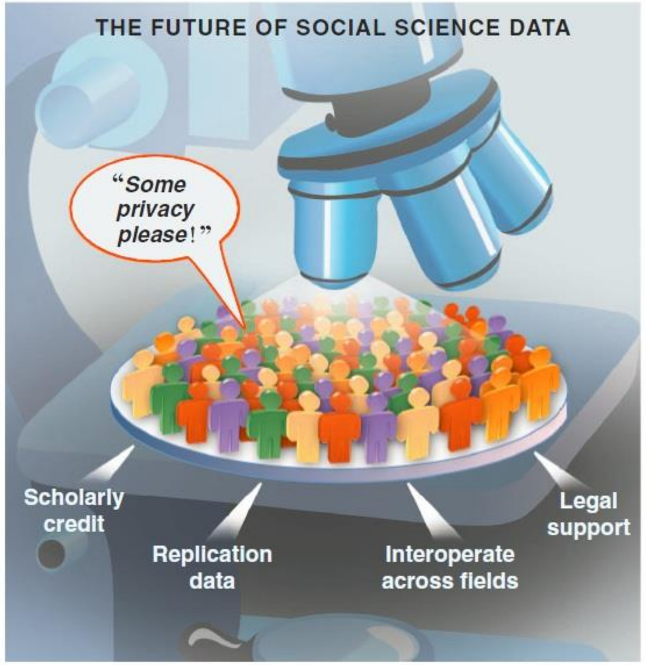


Fig. 1. New types of research data about human behavior and society pose many opportunities if crucial infrastructural challenges are tackled.

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

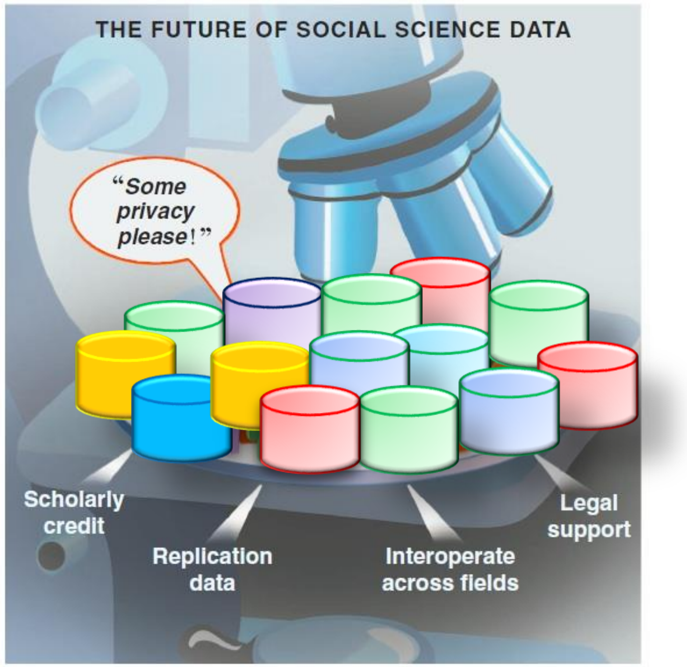


Fig. 1. New types of research data about human behavior and society pose many opportunities if crucial infrastructural challenges are tackled.

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

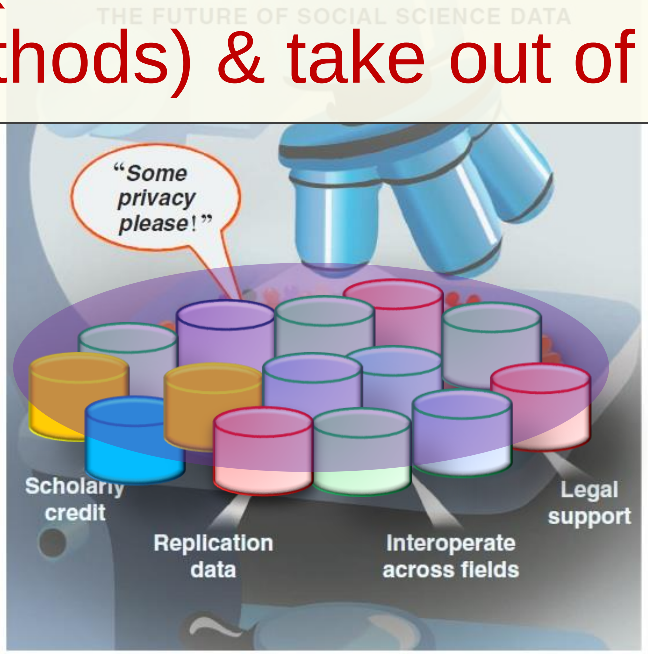
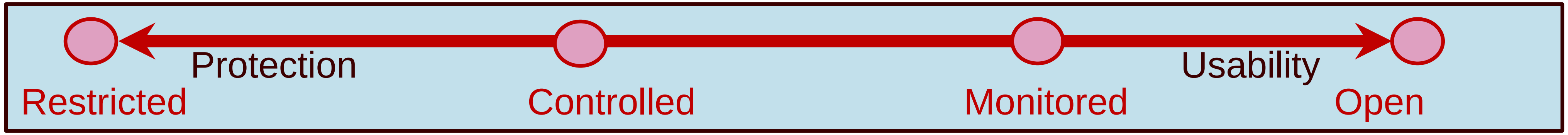


Fig. 1. New types of research data about human behavior and society pose many opportunities if crucial infrastructural challenges are tackled.

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

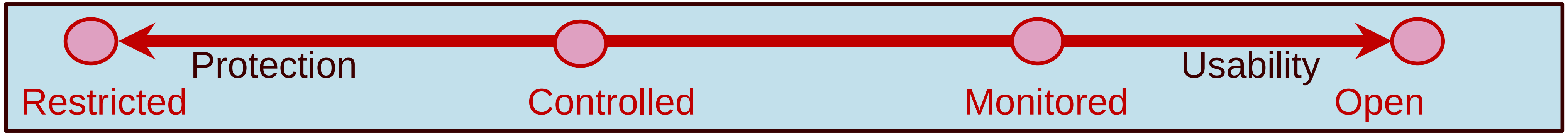
- Anyone : Publish information for others
- No IRB
- No monitoring use
- Publish data use terms
- Disclosure Limitation Methods (filter)
  - Be careful of incorrect use
- Sanitized data
- Public websites, publications

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 usability



Protection

Usability

Restricted

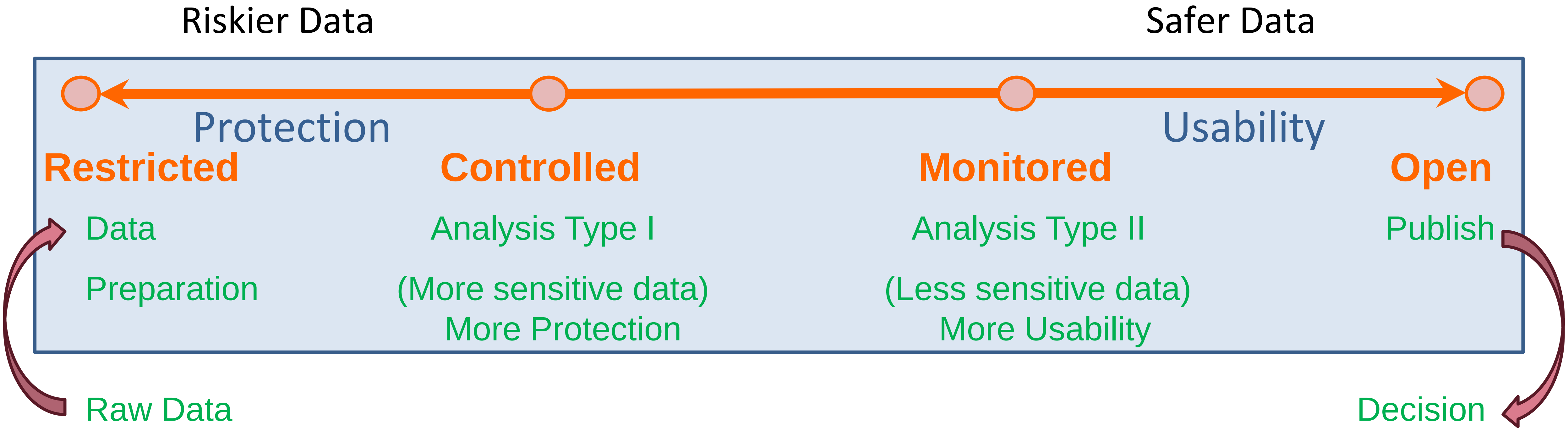
Controlled

Monitored

Open

	Restricted Access	Controlled Access	Monitored Access	Open Access
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

Use 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

Kum, H.C., and Ahalt, S. (2013). Privacy by Design: Understanding Data Access Models for Secondary Data, American Medical Informatics Association (AMIA) joint summits on translation science: clinical research informatics

Vocab: Information Privacy

- What is *information* privacy?
- Privacy vs confidentiality
  - don’t ask vs don’t tell
- Privacy vs security
- PHI: Protected Health Information
  - Covered entity, covered function
- PII: Personally Identifiable Information
- Coded data

Vocab: Informed Consent

- Opt in
- Opt out
- Blanket consent
- Revised Common Rule: Broad consent
  - Once opt out, must be able to respect.
  - Waiver is not possible

Vocab: Disclosure

- Identity disclosure
- Attribute disclosure
- Harm from disclosure
  - Identity theft: SSN, Name, DOB
  - HIV status
- Group disclosure
- Partial disclosure
- Incremental disclosure
- Minimum necessary standard
  - Cost of implementation?

Information Privacy 101: Point One
Privacy is a BUDGET constrained problem

- Differential privacy literature proves each query leads to some privacy loss while providing some utility in terms of data analysis
- Current protection mechanism in database research is not effective
  - de-identified data cannot be linked
  - Not sharing enough details: leads to bias, and invalid results
- The goal is to achieve the maximum utility under a fixed privacy budget

Information Privacy 101: Point two
Information Accountability (Transparency) Works

- Secrecy : Hiding information does not support legitimate use
  - 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 support effective use (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.

Information Privacy 101: Point three
Privacy is contextual

- Helen Nissenbaum (NYU Law School): contextual integrity
- *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

Information Privacy: Myths and fallacies

- "There is no silver bullet to privacy preserving computation"
  - Narayanan A, Shmatikov V. Myths and fallacies of personally identifiable information. Communications of the ACM. 2010 Jun 1;53(6):24-6.
- Manage risk by knowing how to handle the tools appropriately
- Privacy by Design: a well orchestrated system to enhance privacy
  - Good IRB approval guidelines
  - Well designed systems to conduct analysis
    - Minimum necessary
    - Fine grained access control
  - Education & Training
  - Regular privacy audits

Thank you

- Secure access models available at TAMU: Virtual Data Library (ViDaL)
  - https://vidal.tamu.edu/
- Questions?
  - Hye-chung Kum, kum@tamu.edu
- Population Informatics Lab
  - https://pinformatics.org/