

ETHICAL PRINCIPLES AND DATA SCIENCE – REPURPOSING ADMINISTRATIVE & OPPORTUNITY DATA

Stephanie Shipp

Sallie Keller

Aaron Schroeder

SOCIAL AND DECISION ANALYTICS DIVISION
BIOCOMPLEXITY INSTITUTE

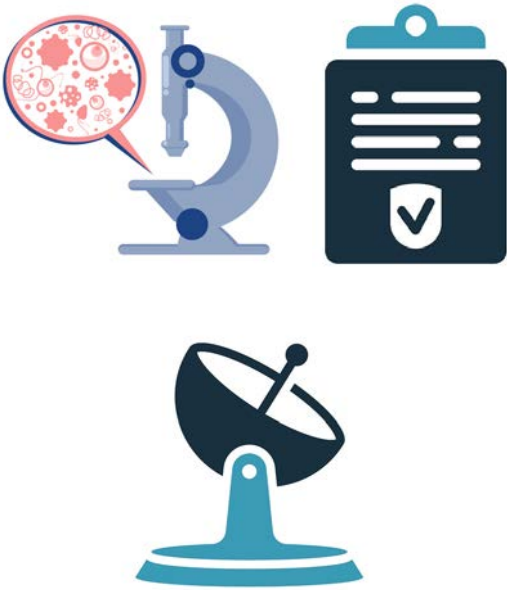
FCSM Fall Conference
21 September 2020



Repurposing *all* data sources

Local, State, and Federal

Designed Data



Administrative Data




Opportunity Data



Procedural Data





What are our
obligations to use
these data ethically?

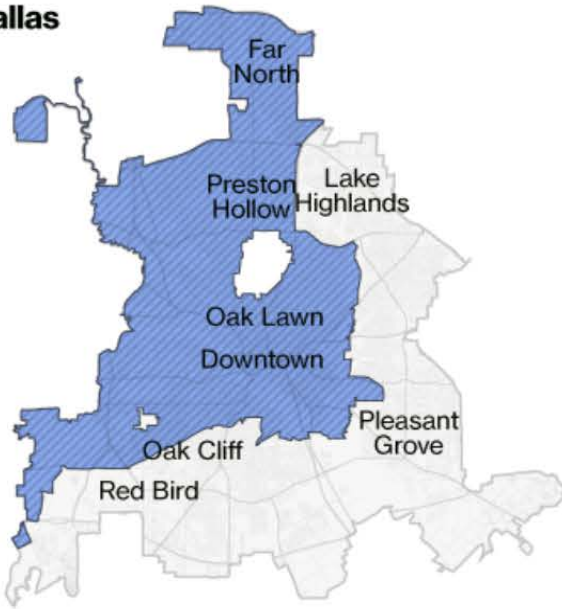
Atlanta



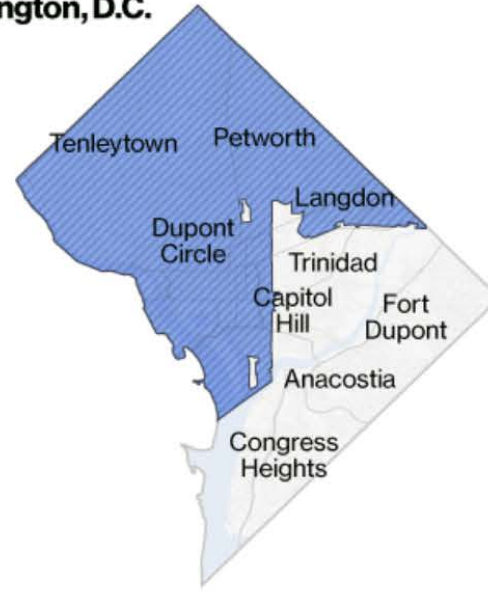
New York City



Dallas



Washington, D.C.



Responsible analyses A cautionary note

**Amazon Doesn't Consider
the Race of Its Customers.
Should It?**

Bloomberg

April 21, 2016

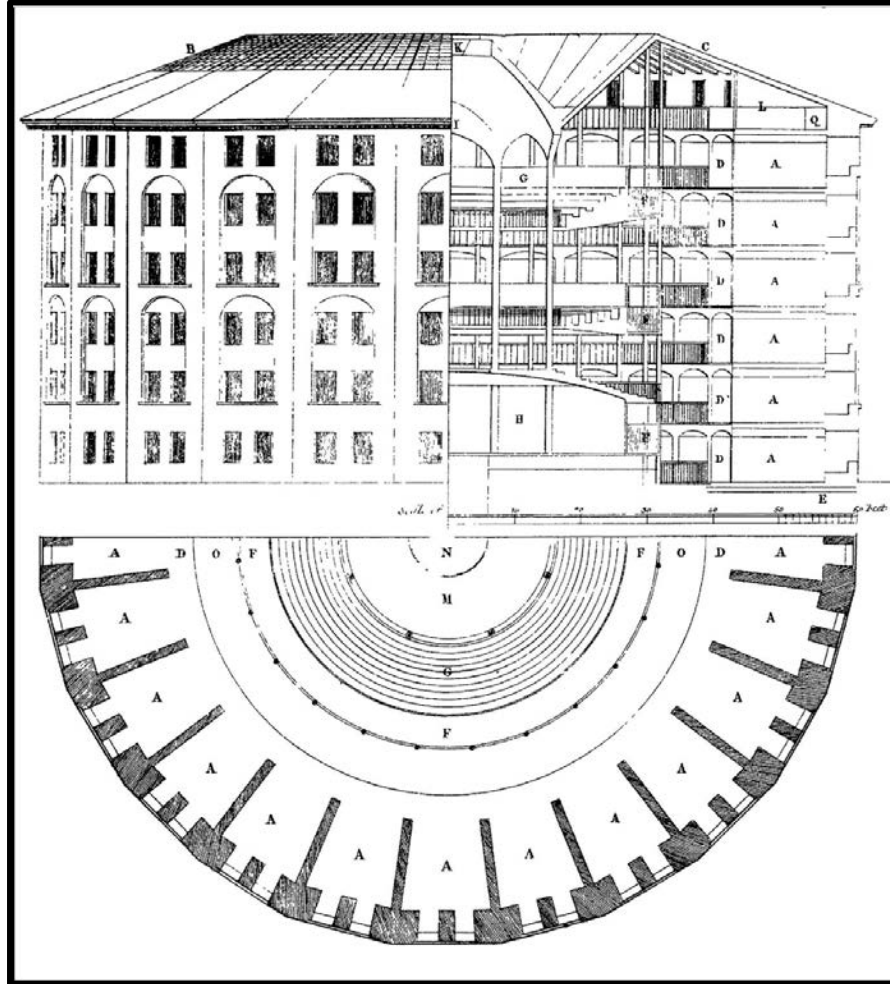
Hospital risk scores prioritize white patients



When health risk prediction algorithms focus on cost rather than illness, racial bias can creep in

Source: https://www.sciencemag.org/news/2019/10/hospital-risk-scores-prioritize-white-patients?utm_campaign=news_daily_2019-10-24&et rid=475169293&et_cid=3044305

Data Science – Repurposing Data



Panoptican Prison Metaphor

- Today, we can observe behavior based on repurposing existing data without consent or awareness by those providing the data
- Requires a broader view of ethics that adapts to the spirit as well as the rules

Source: The Works of Jeremy Bentham, 1791, described in Salganik, Bit by Bit (2018)

Source: Keller, S. A., Shipp, S., & Schroeder, A. (2016). Does big data change the privacy landscape? a review of the issues. *Annual Review of Statistics and Its Application*, 3.

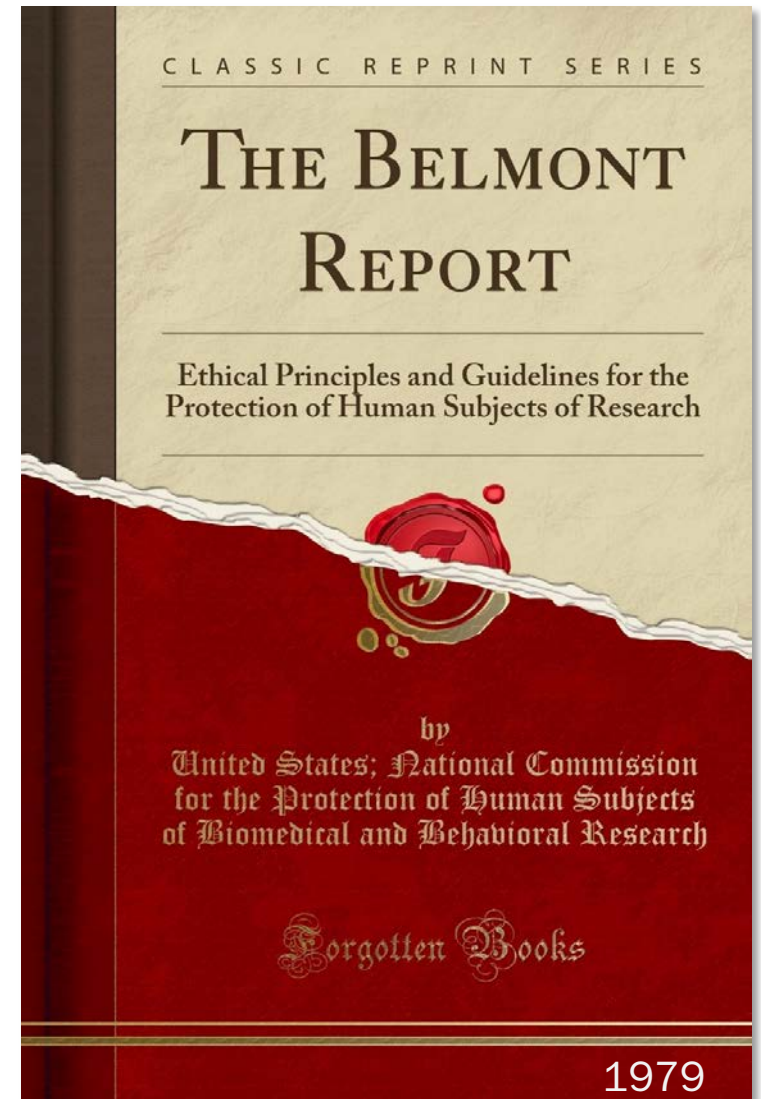
The Belmont Report

The **Belmont Principles** provide the foundation for the Institutional Review Board (IRB) process and for a principles-based approach to ethics.

- **Respect for persons** – honoring individual wishes
- **Beneficence** – weighing risks and benefits of study
- **Justice** – fair distribution of risks and benefits

Led to the **Common Rule** which governs U.S. Government research

Focus is on
“research involving human subjects”



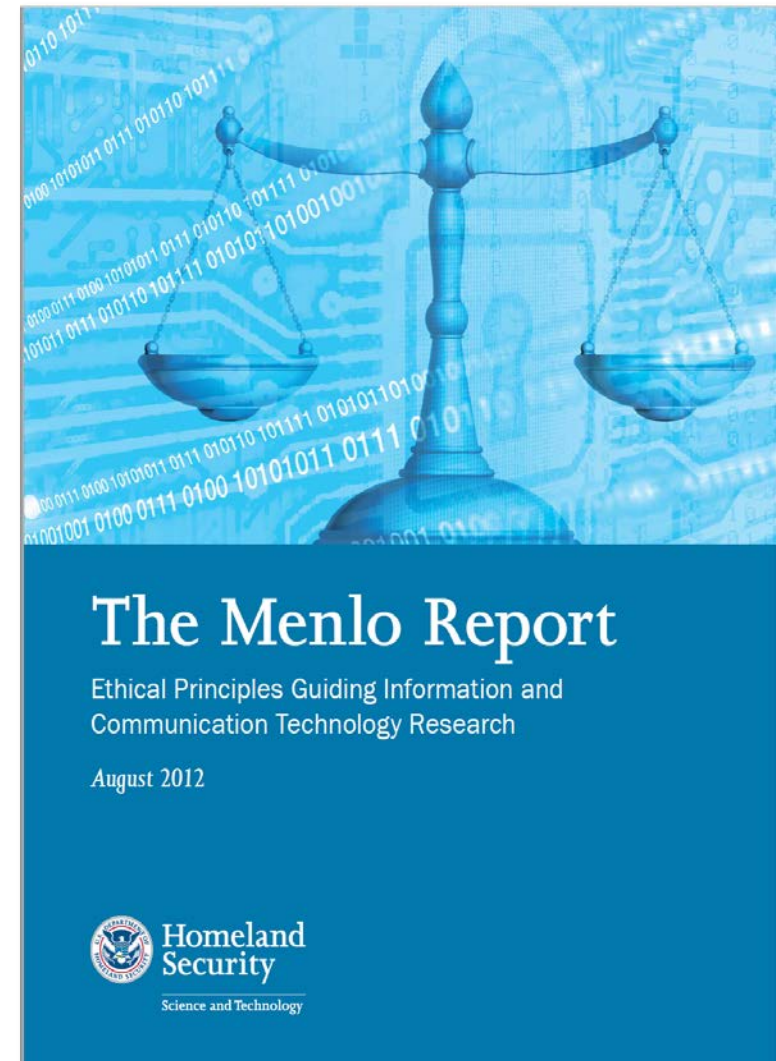
The Menlo Report

Department of Homeland Security commissioned a report to introduce **Belmont principles** in ICT (Information, Communication, Technology)

The **Menlo Report** extends the **Belmont Principles** to include **Respect for Law and Public Interest**

Focus is on
“research with human-harming potential”

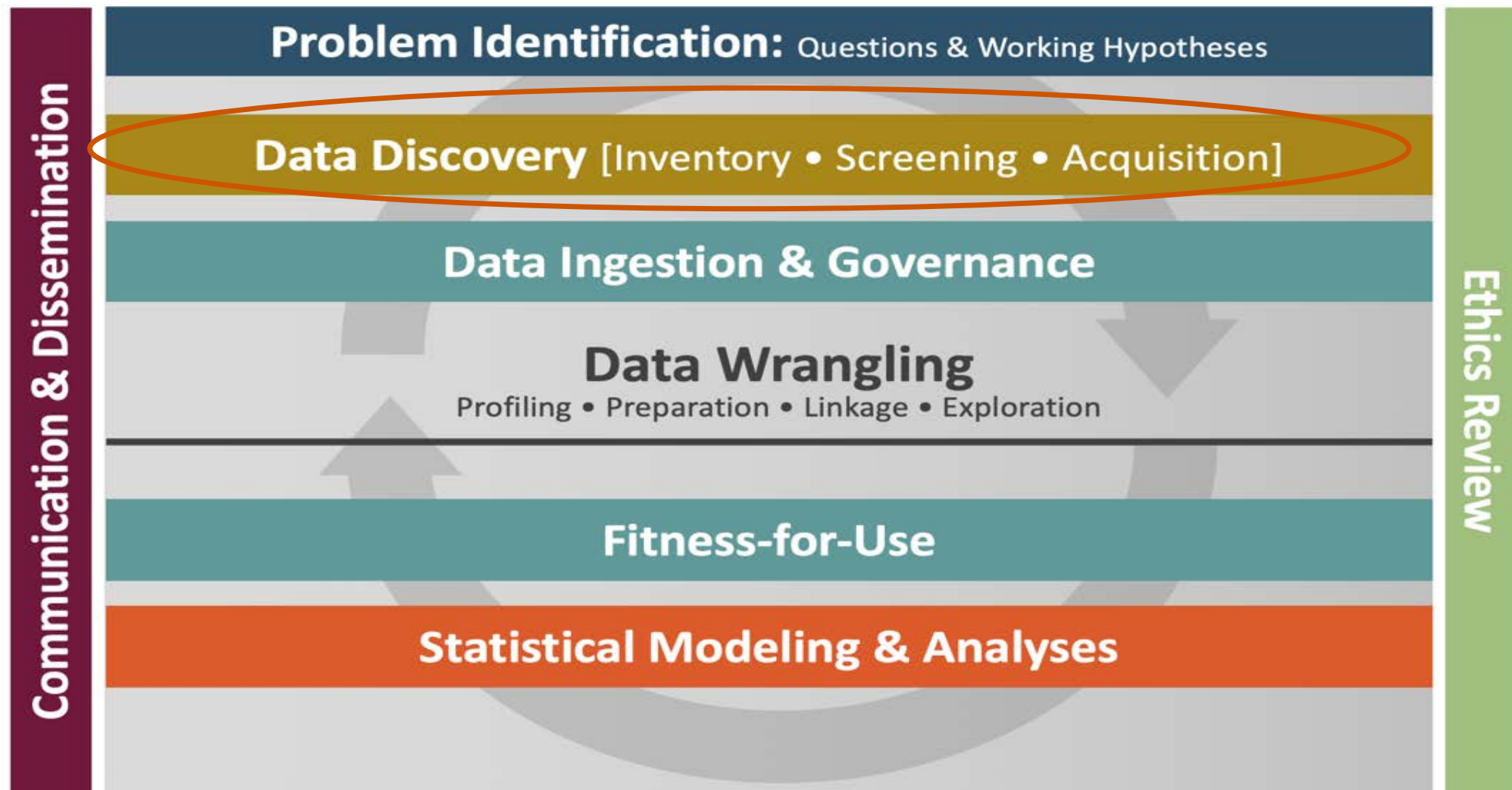
Important in the digital age where the use of digital technologies and repurposing of data can expose people to risks





How does this apply
to our research using
all data?

Data pipeline **starts** with data discovery



Example: Data Discovery around Local Housing

Commerical Data

[Black Knight Financial Services](#)
[MPF Research \(RealPage\)](#)
[National Association of REALTORS](#)
[Real Capital Analytics](#)
[Zillow](#), [Redfin](#)
[Mortgage Bankers Association](#)
[CoreLogic](#)
[MLS Data](#)
[MRIS](#)
[RealtyTrac](#)
[WegoWise](#)
[Equifax Credit Scores](#)
[TransUnion Credit Data](#)
[Experian](#)
[Foot Traffic - SentiLock](#)
[Axiometrics, Inc](#)
[Planet Labs](#)
[Blackbridge](#)
[CoStar](#)

Local - data sharing agreements

[Community Planning & Housing Development](#)
[Permitting](#)
[Real Estate Assessments](#)
[Economic Development](#)
[GIS or Mapping Center](#)
[Crime data](#)
[Fire and EMS](#)
[Building Energy Report Cards](#)
[Bicycle & Pedestrian Counters](#)
[Resident Poll Results](#)
[Alerts](#)

State – data sharing agreements

[Housing Virginia](#)
[Northern Virginia Association of Realtors](#)
[VHDA Housing Analysis](#)
[Virginia Housing Coalition](#)

Other – mixed access

[National Change Database \(NCDB\)](#)
[Community Commons Maps](#)
[Crime Reports](#)
[IPUMS-USA](#)
[National Council on Real Estate Investment and Fiduciaries](#)
[Location Inc \(Neighborhoodscout\)](#)
[USDA Forest](#)
[Maponics](#)
[Center for Regional Analysis](#)
[Urban Tree Canopy](#)
[Yelp](#)
[Walk Score](#)
[RS Metrics](#)
[AirBnB](#)
[TripAdvisor](#)
[InfoUSA Mailing List](#)

Screening & Acquisition

Data Source	Geography
American Community Survey data (Census), 2012-2016	Block Groups
American Time Use Survey (BLS), 2017	National
Youth Risk Behavior Surveillance System, 2015	State
County Health Rankings, 2017	County
Built Environment, e.g., Grocery stores, SNAP retailers, recreation centers, community gardens	Address Level
Fairfax real estate tax assessment data (CoreLogic)	Address Level
Fairfax Open data: Zoning, Environment, water, Parks, Roads	Shapefiles
Fairfax County Youth Survey, 2016 8 th , 10 th , 12 th graders	High School Attendance Area
Virginia Department of Education, 2017	High School
National Center for Education Statistics, 2014-2015	High School
Center for Disease Control, 2014-2015	High School
Regional electronic health records (in-process)	Individual

Example from project in Fairfax, Virginia

- Teen obesity and physical activity
- Data can cross HIPPA, FERPA, Commercial, open

Definitions

Privacy refers to the amount of personal information individuals allow others to access about themselves

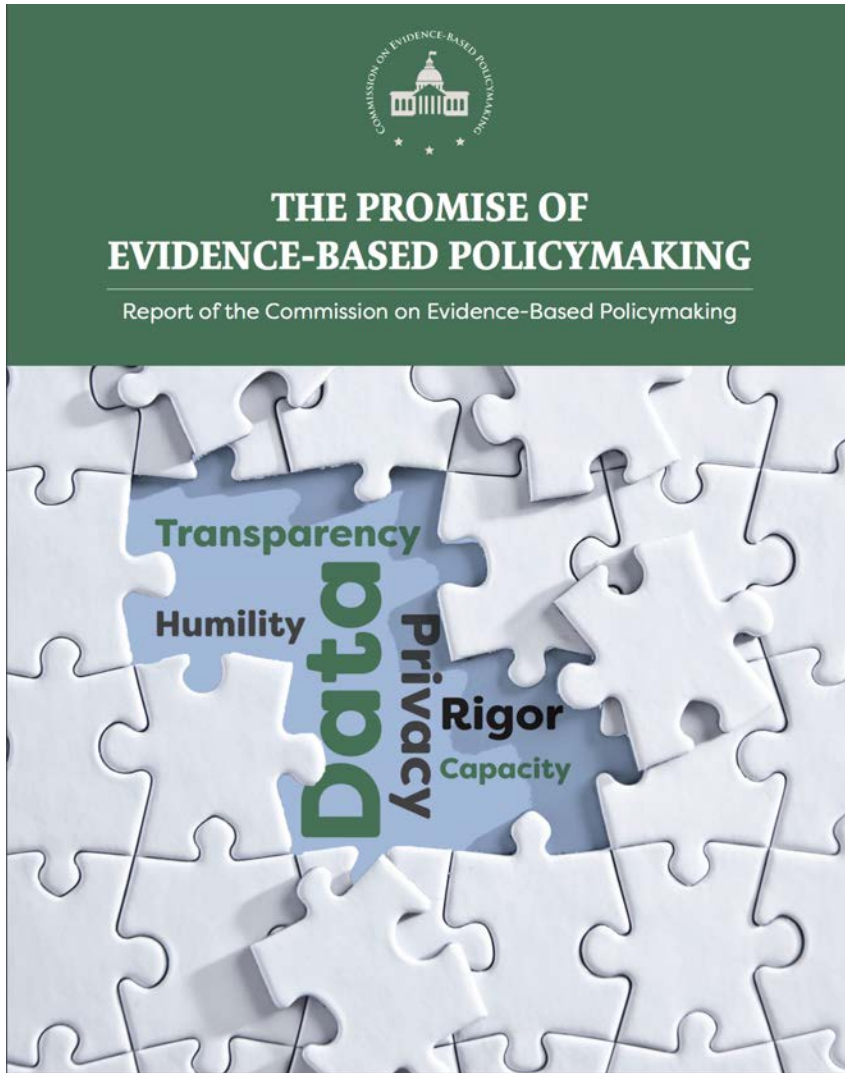


Confidentiality is the process that data producers and researchers follow to keep individuals' data private

Security applies to data storage and transport

Privatization of Data is the collection, aggregation, and (re)processing of personal data to sell to consumers

What about informed consent?



“Access to data held by the government should occur only in service to the public interest.”

Murray/Ryan Commission on evidence-based policy making, 2017

IRB: Waivers of the elements of consent

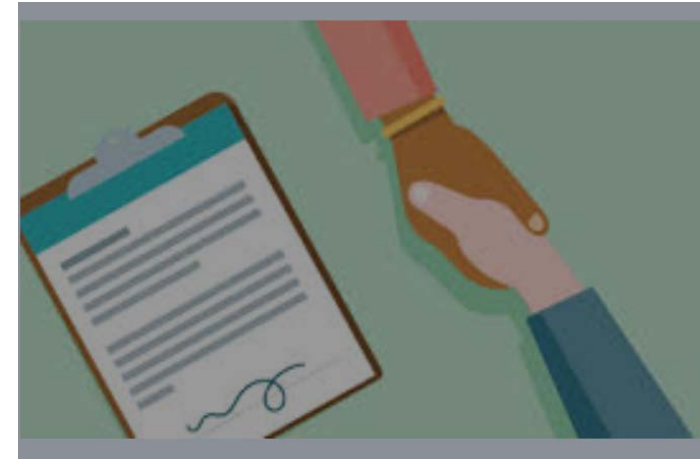
Federal regulations allow IRBs to **authorize researchers to modify the consent process ... only if these criteria are met.**

- The research involves **no more than minimal risk to the subjects.**
- **The research could not practicably be carried out without the requested waiver or alteration.**
- **The waiver or alteration will not adversely affect the rights and welfare of the subjects.**

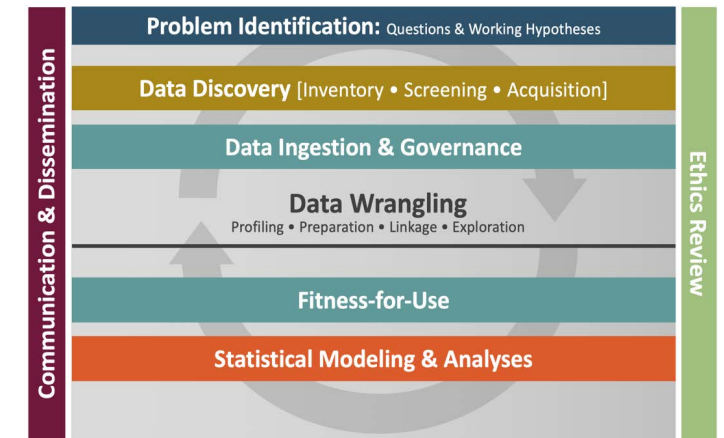
Local government data sharing agreements

Typical language in "local" data sharing agreements

- WHEREAS the County and ENTITY NAME seeks to improve the County and its citizens' quality of life and services while accelerating the County's efficiency and resiliency through the use of County data and combining County community planning and management skills with ENTITY NAME's analytical and data science expertise



Ethics Checklist for Data Science Lifecycle



- 1 **Project Initiation and Problem Identification**
- 2 **Data Discovery, Inventory, Screening, and Acquisition**
- 3 **Data Ingestion and Governance**
- 4 **Data Wrangling**
- 5 **Fitness-for-Use Assessment**
- 6 **Statistical Modeling and Analysis**
- 7 **After-Project Debriefing**

Understanding Bias

Data Science for Public Good project

Asked students:

What is life like in rural America?

What is the best thing about living in a rural community?

What is the worst thing about living in rural America?

What is the most pressing problem facing rural communities today?

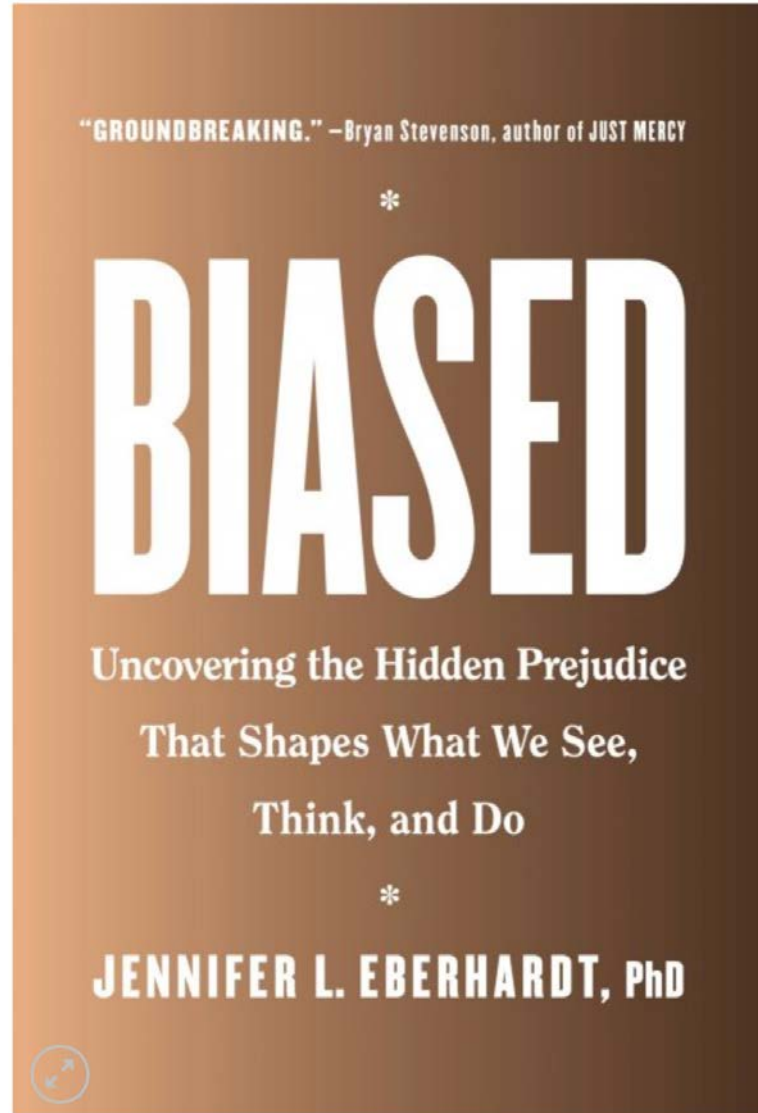
How can a data scientist help rural communities?

What is life like in rural America?



Source: Shawn Dorius, Associate Professor, Social Demography, Iowa State University,
Data Science for Public Good, Understanding Rural Bias, Summer 2020

Practical Solutions – Introduce Friction



Social networking company Nextdoor saw that too many “suspicious character” postings on its online bulletin boards were based solely on race,

Eberhardt. Stanford psychologist, suggested they create a checklist so people had to specify suspicious behavior before describing appearance

Friction - people have to evaluate their reasoning before making bias-based assumptions

Incidence of racial profiling fell by 75%

Use of these data in research before change would perpetuate bias

Incorporating ethics into Data Science Lifecycle

- ✓ Adapt criteria to ensure **implementation of ethical principles**
- ✓ Make **ethical considerations** and **discussion of implicit biases an active and continuous part of the project**
- ✓ **Seek expert help** when ethical questions at any stage cannot be answered
- ✓ Incorporate **ethical guidelines from relevant professional societies** (e.g., American Statistical Association, American Physical Society)
- ✓ **Introduce friction** by “interrogating ourselves and being aware when we’re beginning to make stereotypic associations.” (Eberhardt 2020)
- ✓ A **principles-based approach** allows researchers to make decisions and communicate their decision process for cases where ethical rules do not yet exist (Salganik 2018)

