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*

<table>
<thead>
<tr>
<th>Designed Data</th>
<th>Administrative Data</th>
<th>Opportunity Data</th>
<th>Procedural Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Design Icon" /></td>
<td><img src="image2.png" alt="Admin Icon" /></td>
<td><img src="image3.png" alt="Opportunity Icon" /></td>
<td><img src="image4.png" alt="Procedural Icon" /></td>
</tr>
</tbody>
</table>

What are our obligations to use these data ethically?
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

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.


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”
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 - SentriLock
- 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

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

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.
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

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?

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