

DATA 605

Ethical & Legal Issues in

Data Science

SPRING 2022

SUNELA THOMAS

APRIL 7, 2022

AGENDA

- Questions?
- **Group Presentation – May 12th**
- NO “live” class on April 21st
- Ethics Training Survey Results Example
- Ten Data Science Ethics Questions
- Key Ethics Principles
- Ethics and Integrity in Data Use and Management
- The 5 C’s
- Breakout

GROUP PRESENTATION

- Groups will be assigned in two weeks
- Criteria:
 - Select your own topic for presentation – something that has an ethical issue in data science (e.g., can ads be banned in a browser, can genetic data be shared for analysis, does ethics differ in cultures, etc.)
 - Formal presentation – 10 minutes
 - Everyone in the team participates
 - Presentation to include:
 - Cover page – title and team members listed
 - Problem Statement/Summary
 - Ethical Issues & relation to the theories learned
 - Proposed Solution
 - References, if any
 - Copy of the presentation will be due to me on May 11th by 11:00pm ET
 - Live presentation to the class on May 12th

Should Data Science Include Ethics Training?			
Region (%Voters)	% Yes	% Not sure	% No
USA/Canada (47%)	82%	5.3%	13%
Europe (31%)	64%	9.9%	26%
Asia (12%)	90%	5%	5%
Australia/NZ (3.4%)	73%	9.1%	18%
Africa/Middle East (3.4%)	45%	9.1%	45%
Latin America (3.1%)	80%	10%	10%

Ethics That Every Data Scientist Should Follow

- Decision Making
- Communicating with the Client
- Confidentiality
- Conflict of Interest
- Potential Clients
- Always being Informative
- Ethics Concerning the Data

Ten Data Science Ethics Questions

1. Which laws and regulations might be applicable to our project?
2. How are we achieving ethical accountability?
3. How might the legal rights of an individual be impacted by our use of the data?
4. How might individual's privacy and anonymity be impinged via our aggregation and linking of data?
5. How do we know that the data is ethically available for its intended use?
6. How do we know that the data is valid for its intended use?
7. How have we identified and minimized any bias in the data or in the model?
8. How was any potential modeler bias identified, and if appropriate, mitigated?
9. How transparent does the model need to be and how is that transparency achieved?
10. What are likely misinterpretations of the results and what can be done to prevent those misinterpretations?

Ethical Principles

Beneficence

- minimize harm
- maximize benefits

Respect of Persons

- informed voluntary consent
- vulnerable subjects must be protected

Justice

- equity in distributing risks and benefits between population
- “fairness” in dealing with research participants
- equity between institutions and research partners

Clinical vs. Research Data: Are The Ethics Different?

Privacy and Confidentiality

Informed Consent vs. Implied Consent

Data Integrity and Data Quality

Data Security and Storage

Ethical Guidelines (for Research)

❖ Declaration of Helsinki

- ethical standard used by the International Committee of Medical Journal Editors
- guidelines govern all medical research

❖ CIOMS Guidelines

- Council for International Organizations of Medical Sciences
- developed guidelines in collaboration with WHO

❖ Belmont Report

❖ National Guidelines (Kenya)

Regulations

- ❖ US: Code of Federal Regulations Title 45, Part 46 (45CFR46)
- ❖ FDA - 21CFR50 and 56
- ❖ NIH
- ❖ HIPAA - related both to clinical records and use of subject data in research

Data Integrity

- The assurance that data is accurate, correct and valid.
- Accuracy and consistency of stored data, indicated by an absence of any alteration in data between two updates of a data record.
- Data integrity is imposed within a database at its design stage using standard rules and procedures and is maintained by error checking and validation routines.
- Exact duplication of the sent data at the receiving end, achieved with error checking and correcting protocols.
- Assurance that the data are unchanged from creation to reception.

Data Integrity (cont'd)

- ❖ Process to maintain data integrity depends on:
 - Collection (accurate representation)
 - Data transfer (accurate recording and transfer of data)
 - Storage and Security (preventing loss of data)
 - Sharing of Data
 - Use of data (analysis)

Data Integrity (cont'd)

- ❖ Fabrication and Falsification of data are one of the most serious challenges to data integrity
- ❖ Human error also contributes to loss of data integrity
- ❖ Concern about research misconduct was a primary motivation for a 1990 conference on data management sponsored by the US Department of Health and Human Services.
- ❖ Conference summarized the many ways in which the conduct of research depends on responsible data management.
- ❖ Responsible research begins with experimental design and protocol approval
- ❖ It involves recordkeeping in a way that ensures accuracy and avoids bias
- ❖ It guides criteria for including and excluding data from statistical analyses
- ❖ It entails responsibility for collection, use, and sharing of data.

Data Integrity (cont'd)

- ❖ Everyone with a role in research has a responsibility to ensure the integrity of the data.
- ❖ The ultimate responsibility belongs to the principal investigator, but the central importance of data to all research means that this responsibility extends to anyone who:
 - helps in planning the study
 - collecting the data
 - analyzing or interpreting the research findings
 - publishing the results of the study
 - maintaining the research records.

Data Collection and Integrity

- ❖ Because data collection can be repetitious, time-consuming, and tedious there is a temptation to underestimate its importance.
- ❖ Those responsible for collecting data must be adequately trained and motivated
- ❖ They should employ methods that limit or eliminate the effect of bias
- ❖ They should keep records of what was done by whom and when

Analysis and Selection of Data

- ❖ The use of statistical methods varies widely among research disciplines and also clinical programs (reporting)
- ❖ It is ideal to analyze and report all data
- ❖ Because it is not possible to report everything that has been done, researchers must make decisions about which studies, data points, and methods of analysis to present.
- ❖ Must critically evaluate the reasons for inclusion or exclusion of data, the measures taken to avoid bias, and possible ways in which bias may nonetheless influence data selection
- ❖ Must clearly document how the data were obtained, selected, and analyzed -- especially if the methods are unusual or potentially controversial

Retention of Data

- ❖ What should be retained?
- ❖ It may be impractical to store extraordinarily large volumes of primary data.
- ❖ At minimum, enough data should be retained to reconstruct what was done.
- ❖ How long should clinical records be retained?

Sharing of Data

- ❖ This is considered an important part of responsible research.
- ❖ De-identified data should be shared so that others can verify your conclusions or analysis
- ❖ Sharing of personal patient information is NOT a good practice as noted in Privacy sections earlier.

Data Security

- ❖ Limiting Access
 - Locked Paper Records Offices
 - Limiting access to Paper or Electronic records to appropriate personnel
 - Password Protection of electronic records
 - Defined privileges for electronic data users
 - Firewalls to prevent outside access
- ❖ Regular Backups and proper archiving

Ownership of Data

- ❖ Who owns the data that is generated?
 - Patient?
 - Institution?
 - Funder?
 - Investigator?
 - Publisher?

Ethics in Publication - General guidelines

- ❖ Research should strive to answer specific questions—not just collect or mine data
- ❖ Statistical issues (sample size) are an important part of design to ensure that the research data is likely to answer the question
- ❖ IRB approval is required when using human subjects, human tissues, or medical records

Publication Ethics and Data Analysis

- ❖ Data should be appropriately analyzed
- ❖ Inappropriate analysis is not necessarily ethical misconduct
- ❖ Fabrication or falsification of data is always ethical misconduct
- ❖ Sources and methods of obtaining and processing data should be disclosed
- ❖ Data exclusions should be explained in full
- ❖ Methods used to analyze data should be explained in detail
- ❖ Post hoc analysis of subgroups is acceptable as long as this is disclosed
- ❖ Data Bias should be discussed in all publications of data or analysis

The 5 C's

Consent

Clarity

Consistency and Trust

Control and Transparency

Consequences

Breakout

ALGORITHMS IN THE OFFICE