Introduction to Data Science

Session 12: Data science ethics

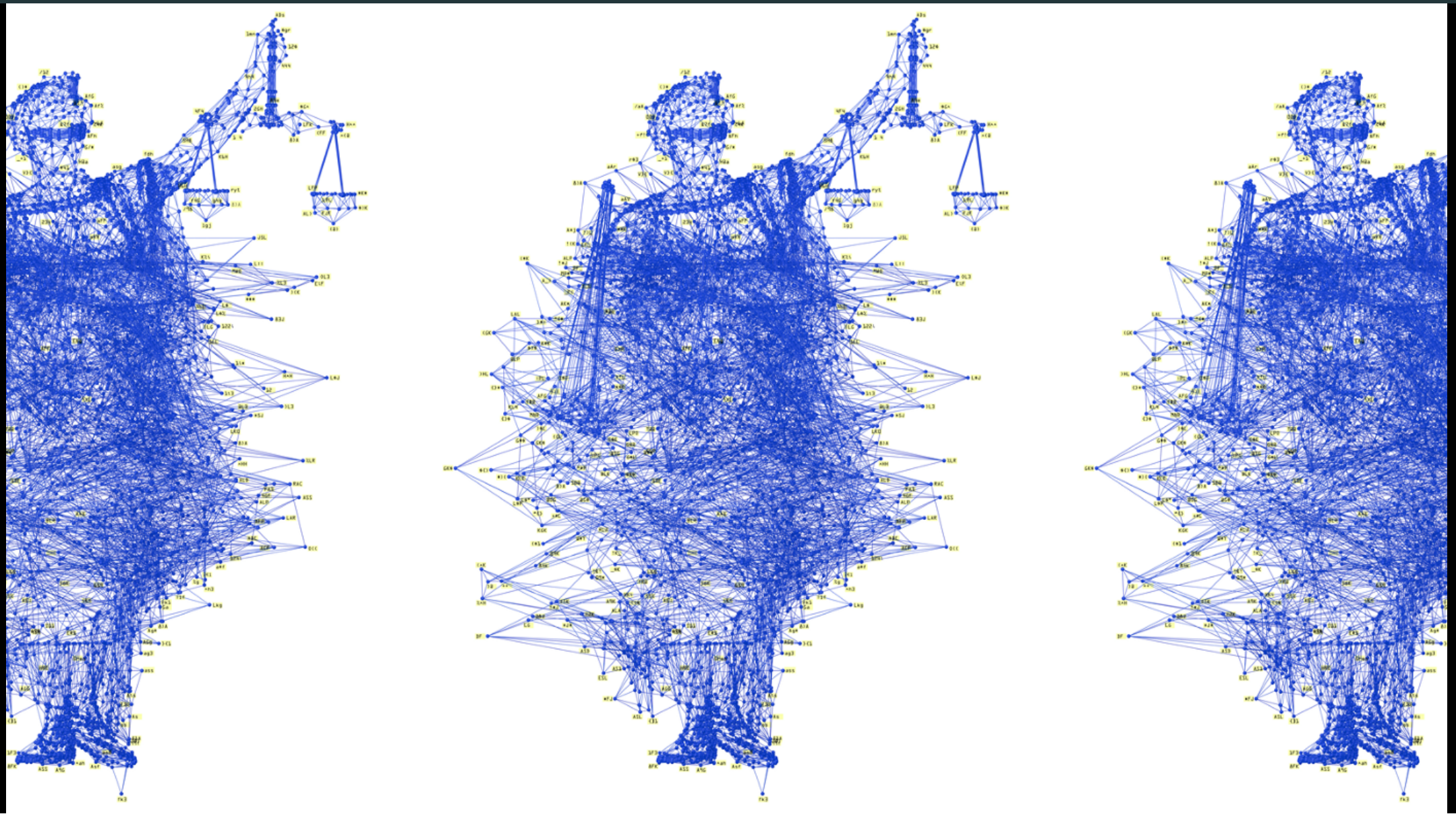
Simon Munzert
Hertie School | GRAD-C11/E1339

Table of contents

Everyday ethics in data science

You came for this...

...and this...



…and this…

...and this...

… but you'll be getting this.

Preparatory work

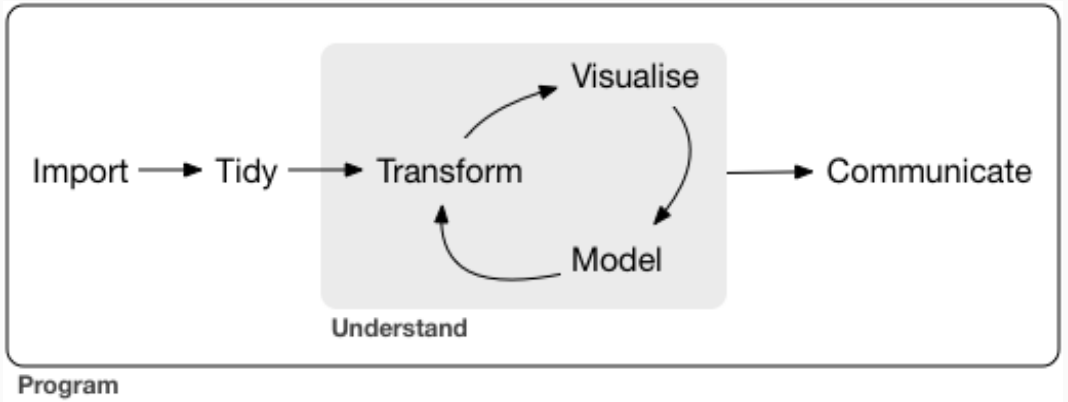
- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

Data operation

- **Wrangle**: import, tidy, manipulate
- **Explore**: visualize, describe, discover
- **Model**: build, test, infer, predict

Dissemination

- **Communicate**: to the public, media, policymakers
- **Publish**: journals/proceedings, blogs, software
- **Productize**: make usable, robust, scalable



… but you'll be getting this.

Preparatory work

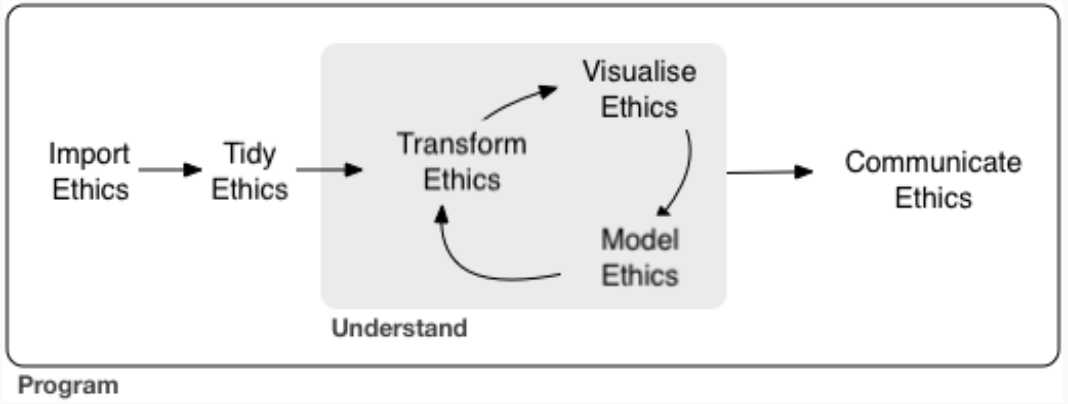
- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

Data operation

- **Wrangle**: import, tidy, manipulate
- **Explore**: visualize, describe, discover
- **Model**: build, test, infer, predict

Dissemination

- **Communicate**: to the public, media, policymakers
- **Publish**: journals/proceedings, blogs, software
- **Productize**: make usable, robust, scalable



Data science ethics

Data science ethics

∋

Classical research ethics[1]

[1]Honesty, objectivity, prudence, openness, respect for intellectual property, social responsibility, ...

Data science ethics

Data science ethics

∋

The ethics of everyday decisions of data scientists

Breakout time! Discuss which ethical issues you may face at various stages of the data science workflow.
Focus on:

Group A: The data collection stage
Group B: The modeling stage
Group C: The communication/dissemination stage

You have **7 minutes** to discuss. Bring concrete examples!

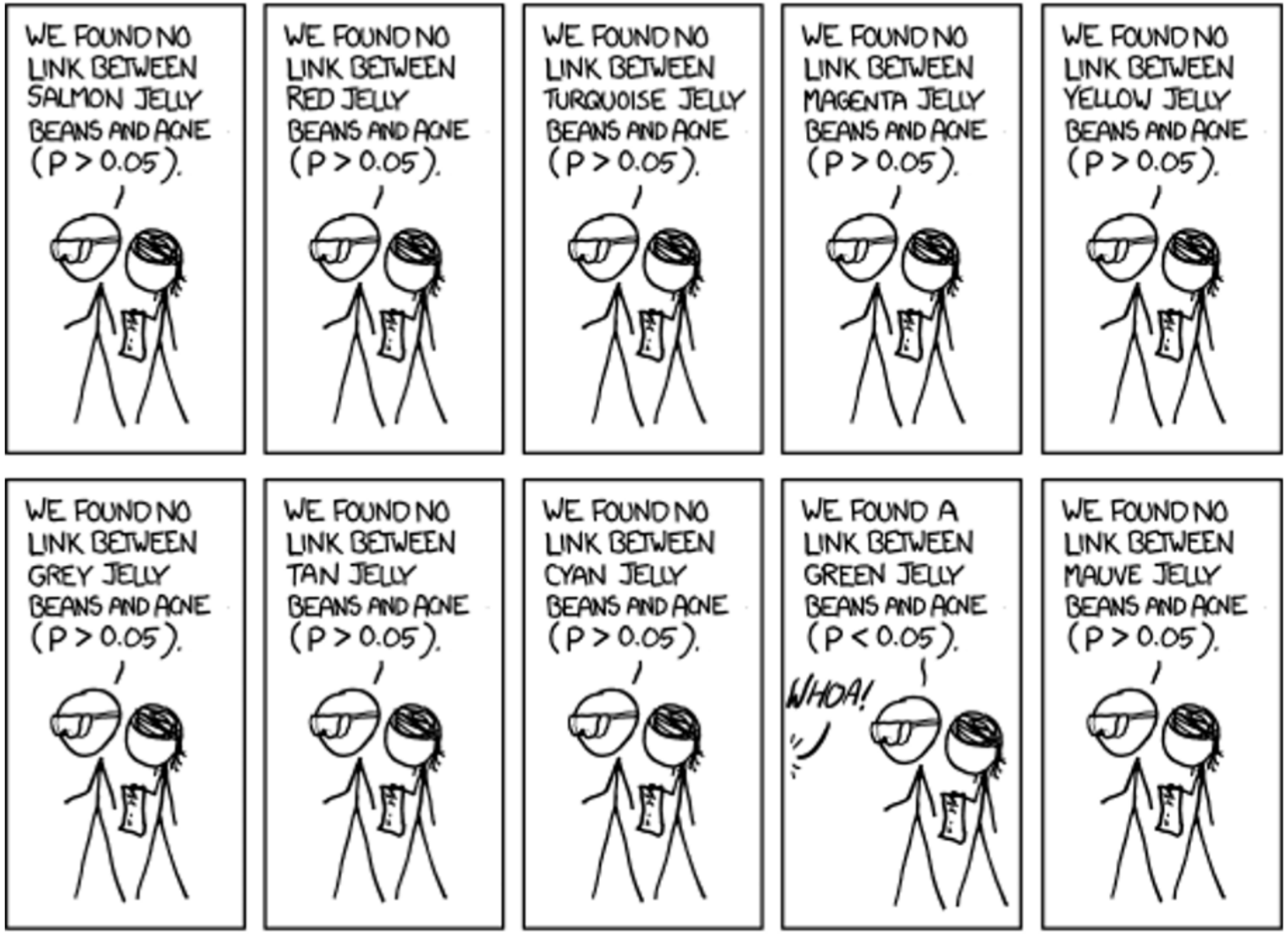
How do I pay clickworkers fairly?

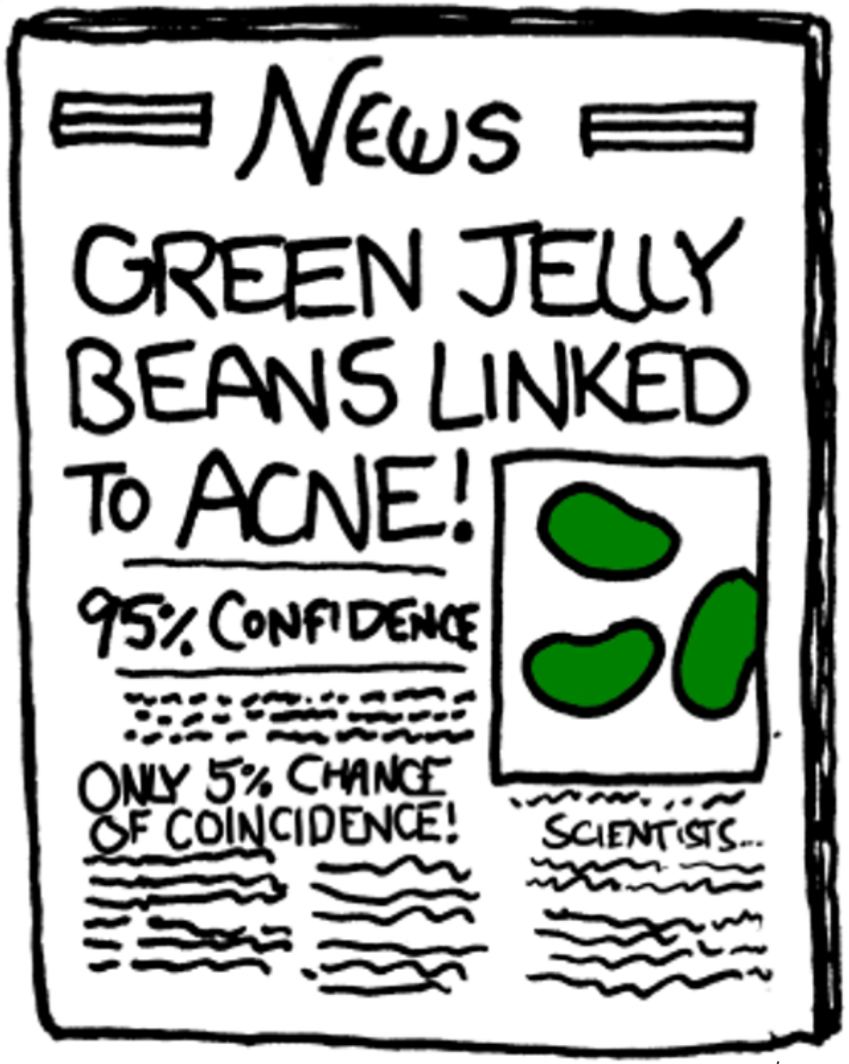
How do I respect intellectual property?

How do I protect the privacy of my research subjects?

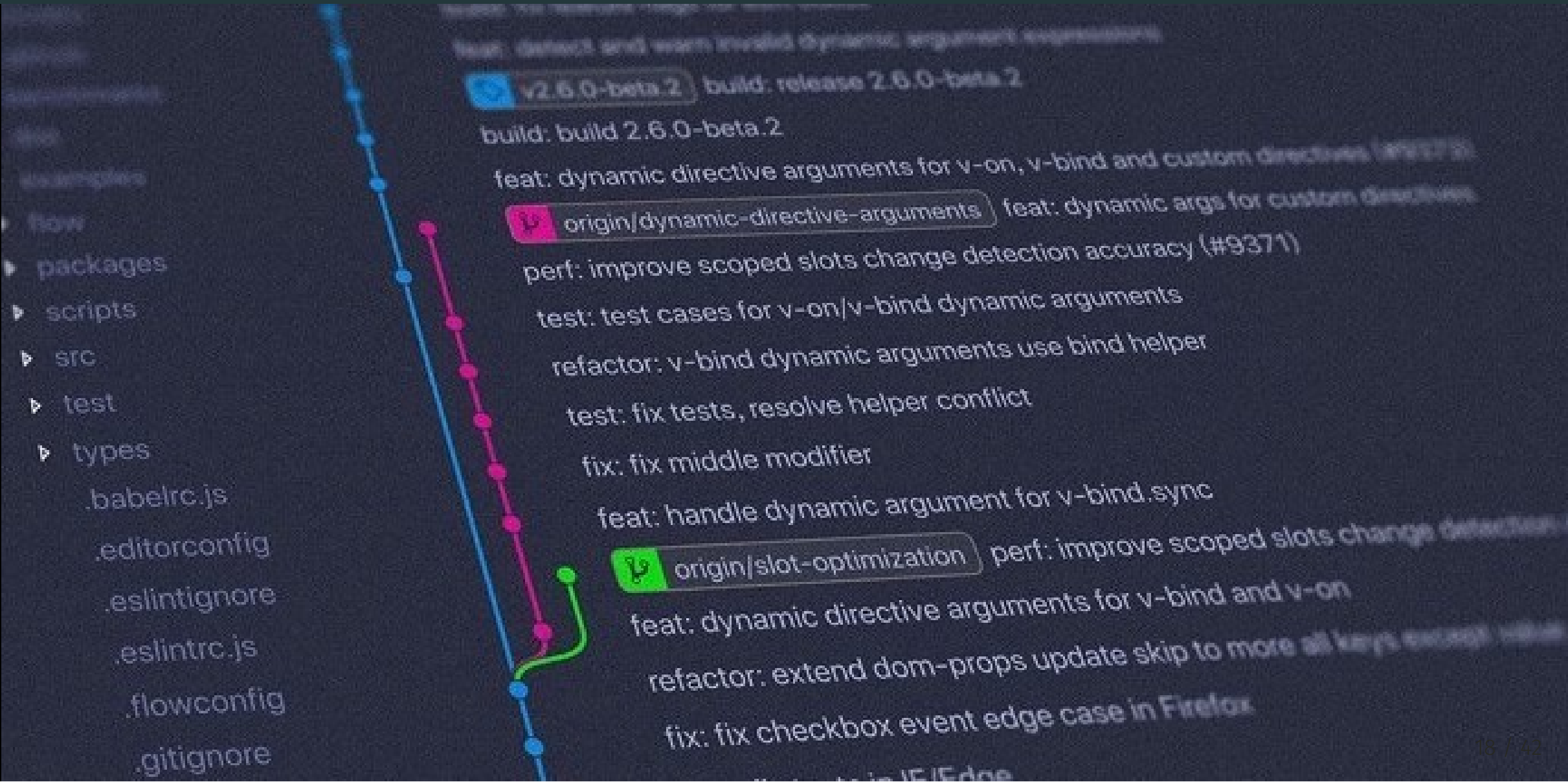
How do I protect the safety of my research subjects?

How do I ensure statistical, measurement validity, etc.?





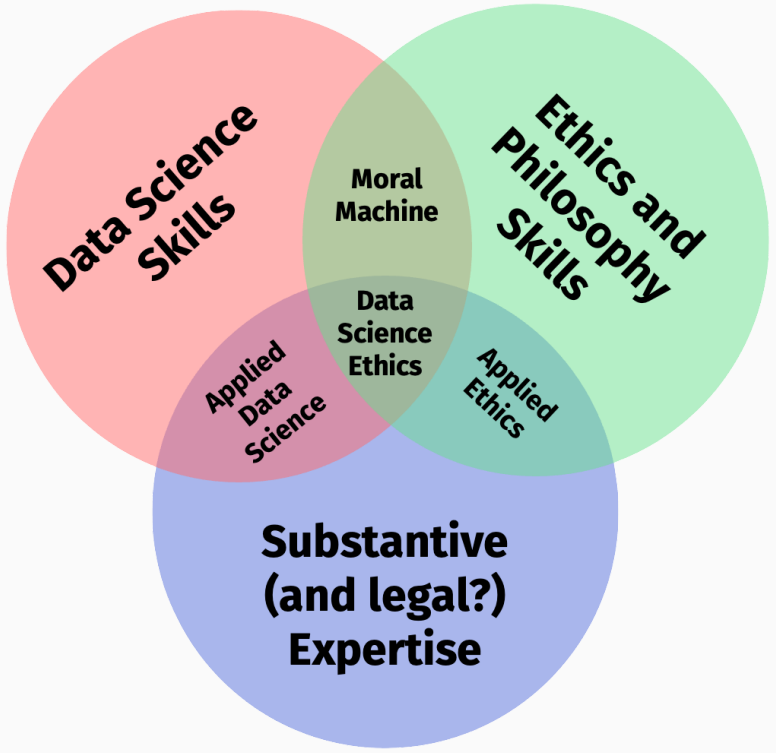
How do I ensure an open science workflow?



How do I communicate results honestly?

Wrap-up

1. **Data science** as a holistic endeavor is more than AI & machine learning.
2. **Data science ethics** is more than automated decision making, algorithm fairness, and privacy preservation.
3. While new data and technologies generate new-ish ethical problems, **you as students of data science** are more likely to be confronted with ethical decisions that pop up in the data science pipeline (i.e., when generating, recording, processing, disseminating data).
4. For **others in your team who are not data scientists**, this implies that a fundamental understanding of the data science pipeline is key to generate ethical insight on relevant problems at the intersection of data science and ethics.



Ethical principles

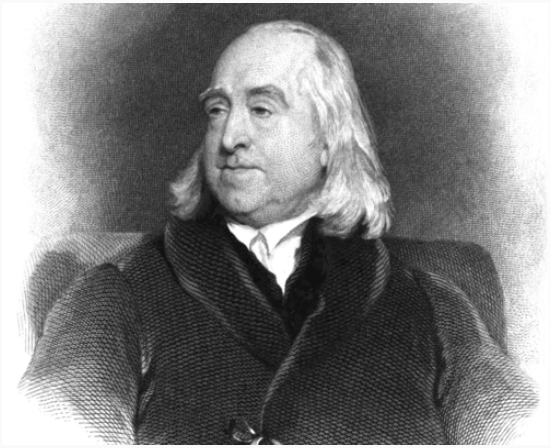
Two ethical frameworks

Deontology

- Follow ethical duties that are derived from a set of rules independent of their consequences.
- Roots in the work of Immanuel Kant.
- The principle of *Respect for Persons* (autonomy) is deeply rooted in deontological thinking.
- Focused on **means**, not **ends**.

Consequentialism

- Take actions that lead to better states in the world.
- Roots in utilitarianism of Jeremy Bentham, John Stuart Mill.
- The principle of *Beneficence* (risk/benefit analysis) is deeply rooted in consequentialist thinking.
- Focused on **ends**, not **means**.

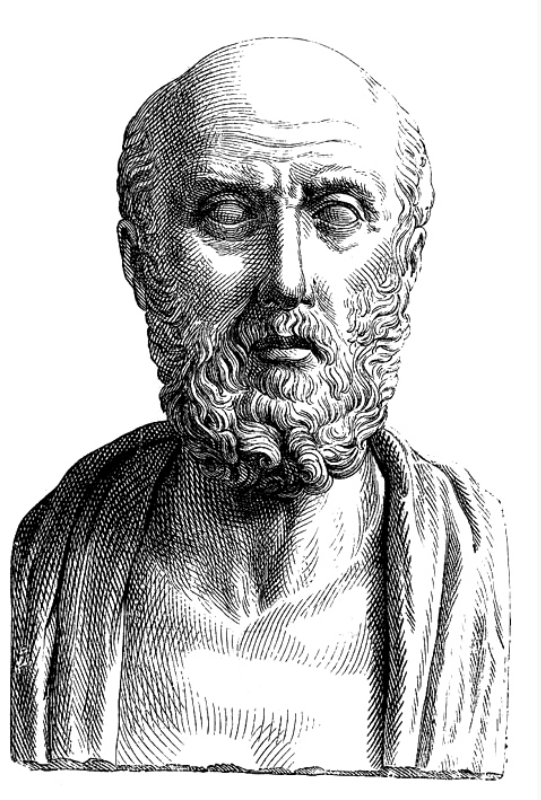


Ethical principles

- Ethical thinking is not an exercise of ticking boxes.
- A set of principles can guide researchers in reflecting about the ethical implications of their research.
- In different contexts these principles can come into conflict with each other.
- In fact, the most interesting cases are when the ethical implications of research involve trade-offs of principles.
- By making principles explicit, those trade-offs can be clarified and decisions better communicated.

- We will focus on the following four principles:

    1. Respect for persons
    2. Beneficence
    3. Justice
    4. Respect for law and public interest

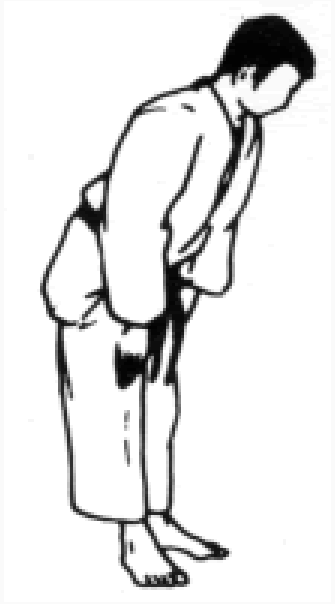


`Source` Wikimedia Commons

1. Respect for persons

Respect for persons is about treating people as autonomous and honoring their wishes.

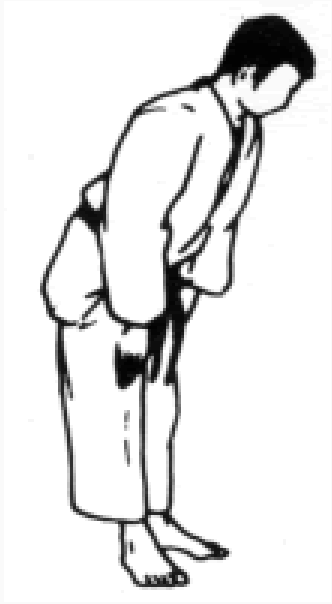
- **Autonomy and consent**: All human subjects should have the right to decide whether to participate in a study, as well as the right to withdraw at any time, without any negative consequences.
- **No coercion**: The decision to participate in a study shall be made freely and without coercion, whether explicit or implied.
- **Protection**: Individuals with diminished capacity to make these decisions must be protected.
- **Orthodox interpretation**: Researchers should not do things to people without their consent.
- **Privacy**: Respect preferences regarding privacy and anonymity.



1. Respect for persons

Respect for persons is about treating people as autonomous and honoring their wishes.

- **Autonomy and consent**: All human subjects should have the right to decide whether to participate in a study, as well as the right to withdraw at any time, without any negative consequences.
- **No coercion**: The decision to participate in a study shall be made freely and without coercion, whether explicit or implied.
- **Protection**: Individuals with diminished capacity to make these decisions must be protected.
- **Orthodox interpretation**: Researchers should not do things to people without their consent.
- **Privacy**: Respect preferences regarding privacy and anonymity.



Questions to reflect on:

- How could coercion look like in practice?
- Under which circumstances is consent ethically problematic? And when could it be practically problematic?

2. Beneficience

Beneficence is about understanding and improving the risk/benefit profile of your study, and then deciding if it strikes the right balance.

- **Do no harm!**
- **Risk/benefit analysis**: Maximize possible benefits, minimize possible harms.
- **Social and scientific value**: Research involving risks or costs to human subjects must have social or scientific value.
- **Wellbeing**: Secure the physical, psychological, and social well-being of research participants and others affected by research.
- **Avoid misuse**: Anticipate the possibility of "Dual Use" (e.g. for military applications), or misuse (e.g. for criminal use) and share findings in ways that minimize these risks.

2. Beneficience

Beneficence is about understanding and improving the risk/benefit profile of your study, and then deciding if it strikes the right balance.

- **Do no harm!**
- **Risk/benefit analysis**: Maximize possible benefits, minimize possible harms.
- **Social and scientific value**: Research involving risks or costs to human subjects must have social or scientific value.
- **Wellbeing**: Secure the physical, psychological, and social well-being of research participants and others affected by research.
- **Avoid misuse**: Anticipate the possibility of "Dual Use" (e.g. for military applications), or misuse (e.g. for criminal use) and share findings in ways that minimize these risks.

Questions to reflect on:

- Under which circumstances could research with a clear net benefit still be impermissible from an ethical point of view?

3. Justice

Justice is about ensuring that the risks and benefits of research are distributed fairly.

- **Protection**: Vulnerable people should be protected from researchers.
- **Equity of access**: Ensure that all groups who could benefit from research should have the chance to participate.
- **Compensation**: Compensate subjects appropriately for their participation.

3. Justice

Justice is about ensuring that the risks and benefits of research are distributed fairly.

- **Protection**: Vulnerable people should be protected from researchers.
- **Equity of access**: Ensure that all groups who could benefit from research should have the chance to participate.
- **Compensation**: Compensate subjects appropriately for their participation.

Questions to reflect on:

- How makes digital research protection and equity of access easier to account for, and how could it become more difficult?
- What does appropriate compensation entail under which circumstances?
- What could be unintended consequences of generous compensation?

4. Respect for law and public interest

Respect for law and public interest extends the principle of beneficience beyond specific research participants to include all relevant stakeholders.

- **Compliance**: Identify and obey relevant laws, contracts, and terms of service.
- **Transparency-based accountability**: Be clear about goals, methods, and results at all stages of research and take responsibility for one's actions.

4. Respect for law and public interest

Respect for law and public interest extends the principle of beneficience beyond specific research participants to include all relevant stakeholders.

- **Compliance**: Identify and obey relevant laws, contracts, and terms of service.
- **Transparency-based accountability**: Be clear about goals, methods, and results at all stages of research and take responsibility for one's actions.

Questions to reflect on:

- Under which circumstances might compliance be impossible to fulfill but the research still ethically acceptable or even desirable?
- Which principle(s) is the transparency-based accountability likely to come in conflict with?

Research ethics in practice

Informed consent

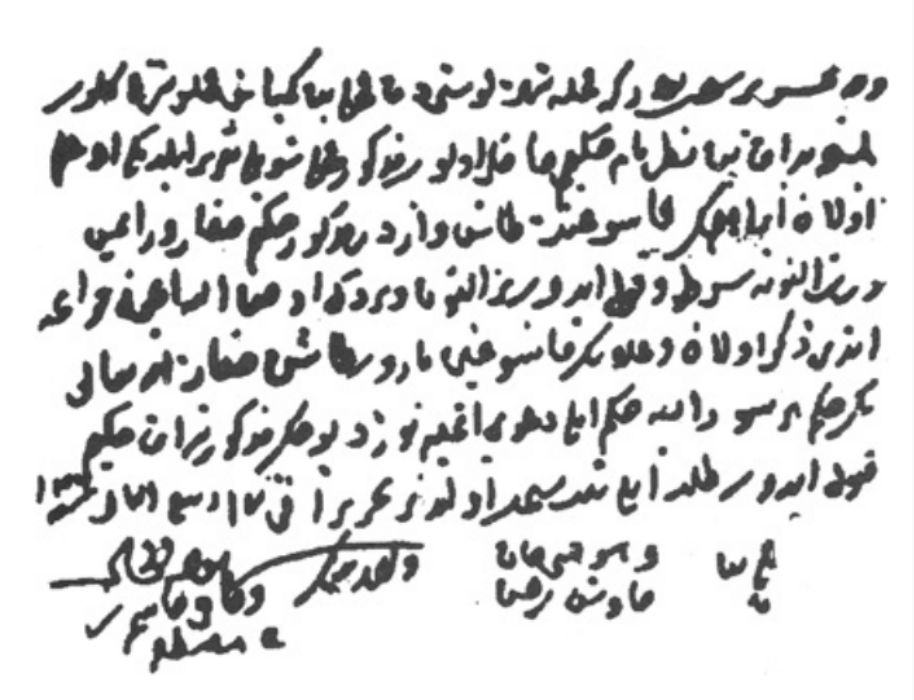
Why informed consent?

- Consent helps ensure individual autonomy.
- Consent should not be seen as a single act of signing a form, but rather a communicative process that extends throughout the course of a research project.

What informed consent should cover

In general, "informed consent" should include communicating

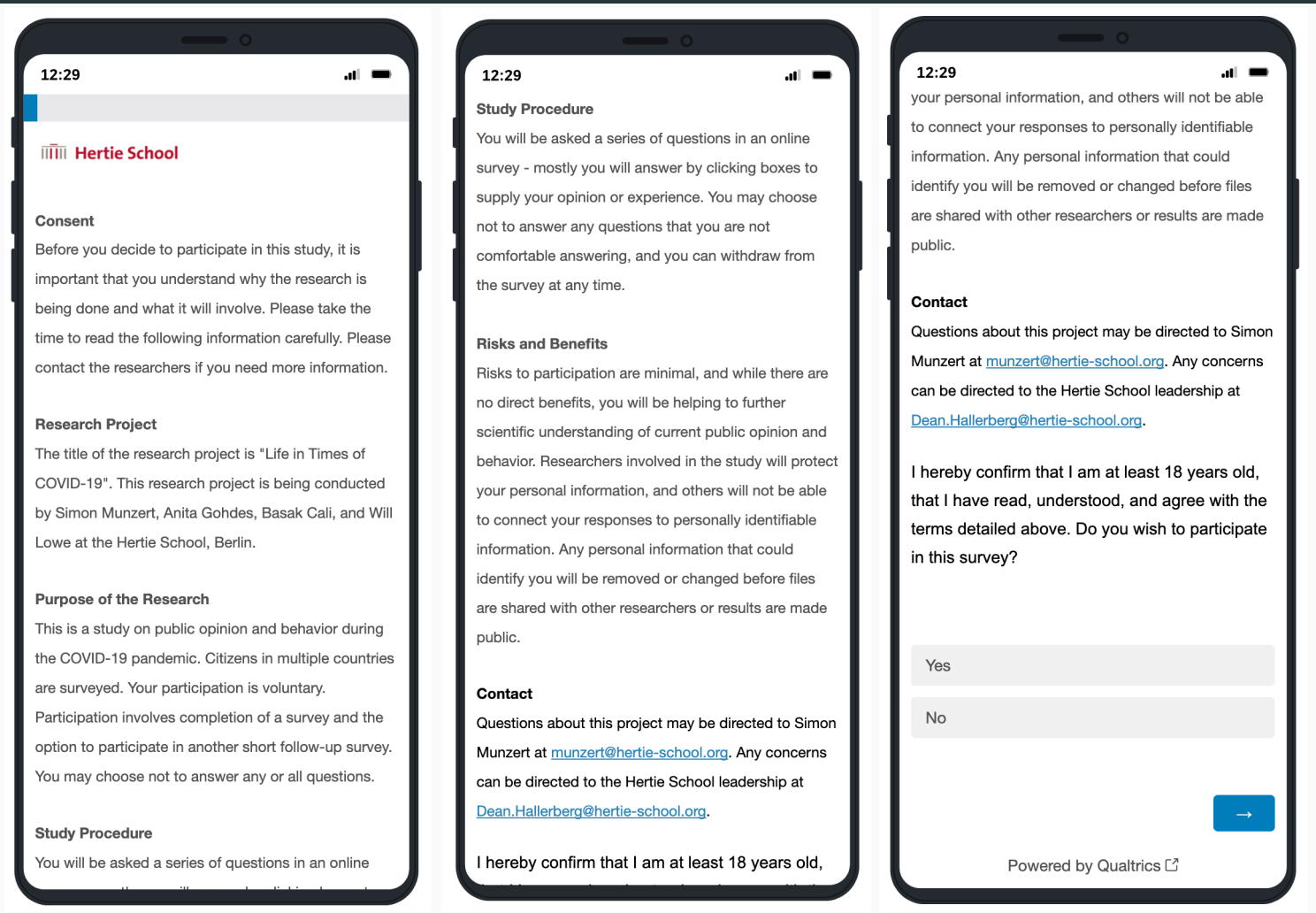
- the research procedure (what),
- the purpose (to what end),
- the stakeholders (by and for whom),
- anticipated risks and benefits, and
- use of data to be collected.



In this Ottoman Empire document from 1539 a father promises to not sue a surgeon in case of death following the removal of his son's urinary stones.

`Source` Salih Selek / Journal of Medical Ethics

Informed consent: example



`Source` Munzert et al. 2021

Informed consent: challenges

When informed consent is not feasible

The **simple rule** "informed consent for everything" is not consistent with ethical principles and research practice. Why?

- Sometimes asking participants to provide informed consent may increase the risks that they face.
- Sometimes having fully informed consent before the study begins could compromise the scientific value of the study.
- Sometimes it is logistically impractical to obtain informed consent from everyone impacted by your study

Informed consent: challenges

When informed consent is not feasible

The **simple rule** "informed consent for everything" is not consistent with ethical principles and research practice. Why?

- Sometimes asking participants to provide informed consent may increase the risks that they face.
- Sometimes having fully informed consent before the study begins could compromise the scientific value of the study.
- Sometimes it is logistically impractical to obtain informed consent from everyone impacted by your study

Towards a better practice

If obtaining full informed consent is not possible, for example, when informed consent would impair the research design, then it must be ensured that

1. there are no undisclosed risks that are more than minimal, and
2. subjects are debriefed whenever possible and appropriate and have the right to withdraw ex post.

Better rule: "The highest possible degree of informed consent must be obtained from research participants."

Informational risk

What is informational risk?

- The potential for harm from the disclosure of information.
- Informational harms could be economic (e.g., losing a job), social (e.g., embarrassment), psychological (e.g., depression), or even criminal (e.g., arrest for illegal behavior).
- Unfortunately, a frequent by-product of research in the digital age (contrary to physical risk).

Informational risk

What is informational risk?

- The potential for harm from the disclosure of information.
- Informational harms could be economic (e.g., losing a job), social (e.g., embarrassment), psychological (e.g., depression), or even criminal (e.g., arrest for illegal behavior).
- Unfortunately, a frequent by-product of research in the digital age (contrary to physical risk).

How to mitigate informational risk?

- Anonymization, i.e. remove obvious personal identifiers such as name, address, telephone number, etc.

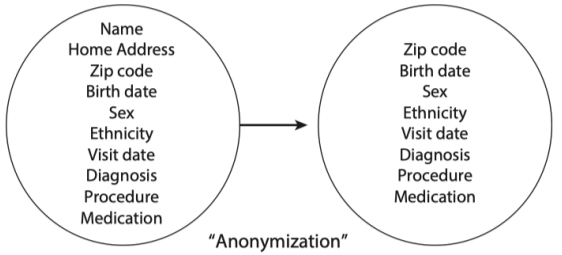


Figure 6.4: "Anonymization" is the process of removing obviously identifying information. For example, when releasing the medical insurance records of state employees, the Massachusetts Group Insurance Commission (GIC) removed names and addresses from the files. I use the quotation marks around the word "anonymization" because the process provides the appearance of anonymity but not actual anonymity.

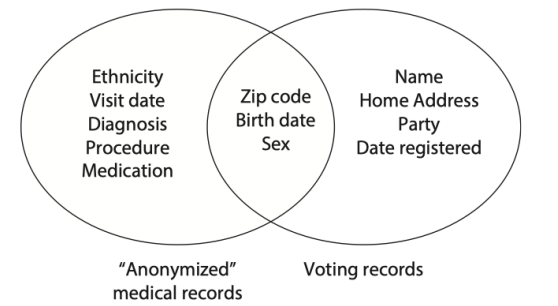


Figure 6.5: Re-idenification of "anonymized" data. Latanya Sweeney combined the "anonymized" health records with voting records in order to find the medical records of Governor William Weld. Adapted from Sweeney (2002), figure 1.

`Source` Matt Salganik, Bit By Bit

Informational risk

What is informational risk?

- The potential for harm from the disclosure of information.
- Informational harms could be economic (e.g., losing a job), social (e.g., embarrassment), psychological (e.g., depression), or even criminal (e.g., arrest for illegal behavior).
- Unfortunately, a frequent by-product of research in the digital age (contrary to physical risk).

Table 6.2: The "Five Safes" are Principles for Designing and Executing a Data Protection Plan (Desai, Ritchie, and Welpton 2016)

Safe	Action
Safe projects	Limits projects with data to those that are ethical
Safe people	Access is restricted to people who can be trusted with data (e.g., people who have undergone ethical training)
Safe data	Data are de-identified and aggregated to the extent possible
Safe settings	Data are stored in computers with appropriate physical (e.g., locked room) and software (e.g., password protection, encrypted) protection
Safe output	Research output is reviewed to prevent accidental privacy breaches

`Source` Matt Salganik, Bit By Bit

How to mitigate informational risk?

- Anonymization, i.e. remove obvious personal identifiers such as name, address, telephone number, etc.
- Safe projects, safe data, safe settings, safe output

Ethics committees and IRBs

Character and purpose

- A (research) ethics committee is an oversight body that ensures that human subject research is carried out in an ethical manner and (depending of the scope) in accordance with the law.
- At US academic institutions Institutional Review Boards (IRBs) take care of this. Their work is regulated by the Common Rule, a rule of ethics that specifies procedures and requirements for ethical research.

The current state of affairs

- In Europe, ethics committees used to be uncommon outside biomedical research.
- As experimentation and other potentially problematic research practices become more common in the social and behavioral sciences, so does ethics oversight.

The mode of operation of ethics committees

1. Universities maintain interdisciplinary or discipline-specific committees.

The mode of operation of ethics committees

1. Universities maintain interdisciplinary or discipline-specific committees.

2. Committees are usually composed of scientists from different disciplines, internal and external persons, and privacy protection officers (at least in Germany 🤷‍♂️)

The mode of operation of ethics committees

1. Universities maintain interdisciplinary or discipline-specific committees.

2. Committees are usually composed of scientists from different disciplines, internal and external persons, and privacy protection officers (at least in Germany 🤷)

3. Researchers submit proposals according to predefined rules that provide information necessary to make judgments about potential ethical issues. This includes:

    - Aim of the research project, timetable, researchers involved, location of the research.
    - Research design: sample, measurement instruments, experiments, etc.
    - Data use and storage concept.
    - Information about features of the project that could be ethically relevant (usually a very long list of questions; see here or here for examples).

The mode of operation of ethics committees

1. Universities maintain interdisciplinary or discipline-specific committees.

2. Committees are usually composed of scientists from different disciplines, internal and external persons, and privacy protection officers (at least in Germany 🤷)

3. Researchers submit proposals according to predefined rules that provide information necessary to make judgments about potential ethical issues. This includes:

    - Aim of the research project, timetable, researchers involved, location of the research.
    - Research design: sample, measurement instruments, experiments, etc.
    - Data use and storage concept.
    - Information about features of the project that could be ethically relevant (usually a very long list of questions; see here or here for examples).

4. The commission meets, consults, and decides; possible outcomes: approval, reject, demand for modification.

The mode of operation of ethics committees

1. Universities maintain interdisciplinary or discipline-specific committees.

2. Committees are usually composed of scientists from different disciplines, internal and external persons, and privacy protection officers (at least in Germany 🤷)

3. Researchers submit proposals according to predefined rules that provide information necessary to make judgments about potential ethical issues. This includes:

    - Aim of the research project, timetable, researchers involved, location of the research.
    - Research design: sample, measurement instruments, experiments, etc.
    - Data use and storage concept.
    - Information about features of the project that could be ethically relevant (usually a very long list of questions; see here or here for examples).

4. The commission meets, consults, and decides; possible outcomes: approval, reject, demand for modification.

5. Depending on the institution, this decision is binding or not. However, more and more journals are requesting the approval of an ethics committee.

Example items from an ethics review questionnaire

1. Does the study involve vulnerable populations (e. g. children <18, prison populations, refugees, ...)?
2. Is it plausible that individuals feel compelled to participate in the study, for instance due to pressure from others, such as management, works council, teachers, traditional or religious leaders, parent, or spouses?
3. Does the research involve individuals who would have difficulty giving meaningful informed consent?
4. Will compensation exceed what is reasonable for time investment and expenses?
5. In the event that the project involves deception, will subjects be debriefed at the end of the study?
6. Is physical pain more than mild discomfort likely to result from participation?
7. Will this project pose any risks to the health and safety of the researchers?
8. Will research involve saving images or audio data from which respondents may be identified?
9. Will you require access to data on research participants held by a third party (physician, school, etc.)?
10. Does this research have potential for misuse (i. e. abuse by criminal or terrorist groups)

Ethics committees in action

Breakout time! Take on the role of an ethics committee and evaluate one of the studies shown on the following pages. Consider these questions:

1. What ethical problems do you see?
2. What questions would you ask the authors to inform your opinion?

You have **10 minutes** to discuss.

An audit study

ARTICLE

https://doi.org/10.1057/s41599-021-00773-2 OPEN

Mapping discrimination in Europe through a field experiment in amateur sport

Carlos Gomez-Gonzalez [1✉], Cornel Nesseler[2] & Helmut M. Dietl[1]

Societies are increasingly multicultural and diverse, consisting of members who migrated from various other countries. However, immigrants and ethnic minorities often face discrimination in the form of fewer opportunities for labor and housing, as well as limitations on interactions in other social domains. Using mock email accounts with typical native-sounding and foreign-sounding names, we contacted 23,020 amateur football clubs in 22 European countries, asking to participate in a training session. Response rates differed across countries and were, on average, about 10% lower for foreign-sounding names. The present field experiment reveals discrimination against ethnic minority groups, uncovering organizational deficiencies in a system trusted to foster social interactions.

`Source` Gomez-Gonzalez et al., 2021

A delicate data set

The OKCupid dataset: A very large public dataset of dating site users

Open Differential Psychology , Nov. 3, 2016, ISSN: 2446-3884

Emil O. W. Kirkegaard, Ulster Institute for Social Research , <emil@emilkirkegaard.dk>
Julius D. Bjerrekær, University of Aalborg , <juliusdb.science@gmail.com>

Abstract

A very large dataset (N=68,371, 2,620 variables) from the dating site OKCupid is presented and made publicly available for use by others. As an example of the analyses one can do with the dataset, a cognitive ability test is constructed from 14 suitable items. To validate the dataset and the test, the relationship of cognitive ability to religious beliefs and political interest/participation is examined. Cognitive ability is found to be negatively related to all measures of religious belief (latent correlations -.26 to -.35), and found to be positively related to all measures of political interest and participation (latent correlations .19 to .32). To further validate the dataset, we examined the relationship between Zodiac sign and every other variable. We found very scant evidence of any influence (the distribution of p-values from chi square tests was flat). Limitations of the dataset are discussed.

Download Paper | Cite

DOI: 10.26775/ODP.2016.11.03

Keywords

intelligence, IQ, cognitive ability, scale construction, Zodiac sign, politics, OKCupid, religiosity, astrology, dating site, big data, open data

Reviewed by

Davide Piffer
Gerhard Meisenberg
Robert L. Williams

Review time 180 days

Review thread
See supplementary materials

`Source` Kirkegaard/Bjerrekær, 2016

Thank you!