

# Sharing in a **gray** area

A framework for big data curation



Sara Mannheimer, Montana State University  
Elizabeth Hull, Dryad Data Repository

National Data Integrity Conference 2017, Fort Collins, CO



## Sara Mannheimer

Data Librarian  
Montana State University  
@saramannheimer



## Elizabeth Hull

Operations Manager  
Dryad Digital Repository  
@datadryad

# Today's Talk

Defining “big data”

Benefits of sharing big data

Ethical challenges of big data research and sharing

STRIDE framework for ethical big data sharing

Dryad case studies

Key takeaways and questions

# Defining “big data”

## Practices

“While ‘big data’ is a vague and amorphous term, we use it as a shorthand to refer to practices that draw on largescale datasets and predictive analytics.”

—Metcalf & Crawford, 2016

# Defining “big data”

## Datasets

Extremely large datasets that may be analyzed computationally to reveal patterns, trends, and associations, **especially relating to human behavior and interactions.**

—Oxford Dictionaries

# Some examples of big datasets

Credit card transaction data

Clickstream data tracked by websites

Geospatial data generated from mobile devices

Internet of Things sensor data

Social media posts

# Data sharing is a good thing

Open Data movement

Funder & publisher requirements

# Big data sharing is a good (and ethically challenging) thing

Big data has the potential to reveal insights about people and society on an unprecedented scale.



# Big data sharing is a good (and ethically challenging) thing

Big data has the perceived potential to reveal insights about people and society on an unprecedented scale.

"Big data rich and big data poor" (boyd & Crawford, 2012)

# Big data sharing is a good (and ethically challenging) thing

Big data has the perceived potential to reveal insights about people and society on an unprecedented scale.

"Big data rich and big data poor" (boyd & Crawford, 2012)

Big data represents human subjects and human activity, and must be considered accordingly

# Ethical challenges of big data research

How can data be gathered without people's knowledge or consent and still meet the ethical obligation to treat people with respect, beneficence, and justice, as outlined in the Belmont Report?

—paraphrased from Crawford, Miltner, & Gray, 2014

# The Belmont Report, 1979

**Respect for Persons.** Participants should be fully informed about the research activity, and should opt into the research.

# The Belmont Report, 1979

**Respect for Persons.** Participants should be fully informed about the research activity, and should opt into the research.

**Beneficence.** Research should maximize benefits and minimize possible harms.

# The Belmont Report, 1979

**Respect for Persons.** Participants should be fully informed about the research activity, and should opt into the research.

**Beneficence.** Research should maximize benefits and minimize possible harms.

**Justice.** Participants who bear the risk of possible harm should also receive the benefits of the research.

# Ethical challenges of big data research

Big data are collected under mandatory terms of service rather than responsible research design overseen by university compliance officers.

—paraphrased from Zook et al., 2017

# Ethical challenges of big data research

Big data are collected under mandatory terms of service rather than responsible research design overseen by university compliance officers.

Data are gathered by agents other than the researcher—private software companies, state agencies, and telecommunications firms.

—paraphrased from Zook et al., 2017



# Ethical challenges of big data research

Big data are collected under mandatory terms of service rather than responsible research design overseen by university compliance officers.

Data are gathered by agents other than the researcher—private software companies, state agencies, and telecommunications firms.

Data are only accessible to researchers after their creation, making it impossible to gain informed consent before the research is conducted.

—paraphrased from Zook et al., 2017

# Ethical challenges of big data research

“Scientific research that involves drawing on what is euphemistically known as “passively collected” big data must face difficult questions and develop new ethical frameworks.”

—Crawford, Miltner, & Gray, 2014

# Data Ethics Canvas



<p><b>What are your data sources?</b></p> <p>Name and describe key data sources used in your project, whether you're collecting them yourself or getting access from third parties.</p>	<p><b>Who has rights over your data sources?</b></p> <p>Where did you get the data from? e.g. is it data produced by an organisation or data collected directly from individuals? Do you have permission or another basis on which you're allowed to use this data? What ongoing rights will the data source have?</p>	<p><b>What's your core purpose for using this data?</b></p> <p>What is your primary use case, your business model? Are you collecting more data than is needed for your purpose?</p>	<p><b>Who could be negatively affected?</b></p> <p>Could the manner in which this data is collected, shared, used cause harm? = be used to target, profile, prejudice people = unfairly restrict access (eg exclusive arrangements) Could people "perceive" it to be harmful?</p>	<p><b>Are you communicating potential risks/issues, if any?</b></p> <p>How are limitations and risks being communicated to people affected by your project, and organisations using data? What channels are you using?</p>
<p><b>Are there any limitations in your data sources?</b></p> <p>Which might influence the outcomes of your project, like: = bias in data collection, inclusion, algorithm = gaps, omissions = other sensitivities</p>	<p><b>What policies/laws shape your use of this data?</b></p> <p>Data protection legislation, IP and database rights legislation, sector specific data sharing policies/regulation (e.g. health, employment, taxation) Sector specific ethics legislation?</p>	<p><b>Do people understand your purpose?</b></p> <p>If this is a project/use that could impact on people or more broadly shape/impact society, do people understand your purpose? Has this been clearly communicated to them?</p>	<p><b>How are you minimising negative impact?</b></p> <p>What steps can you take to minimise harm? Are there measures you could take to reduce limitations in your data sources? Could you monitor potential negative impact to support mitigating activities? What benefits will these actions add to your project?</p>	<p><b>When is your next review?</b></p> <p>When will this Data Ethics Canvas be reviewed? How will ongoing issues be monitored?</p>
<p><b>Are you going to be sharing this data with other organisations?</b></p> <p>If so, who?</p>		<p><b>Who will be positively affected by this project?</b></p> <p>What individuals, demographics, organisations? How will they be positively affected? Do they know and understand how they are positively affected?</p>	<p><b>How can people engage with you?</b></p> <p>Can people affected appeal or request changes to the service? To what extent? Are the appeal mechanisms reasonable?</p>	<p><b>What are your actions?</b></p> <p>What steps are you going to take prior to moving forward with this project?</p>



theodi.org

AUGUST 2017

<https://theodi.org/the-data-ethics-canvas>

# Open Data Institute Data Ethics Canvas

What are your data sources?

Who has rights over your data sources?

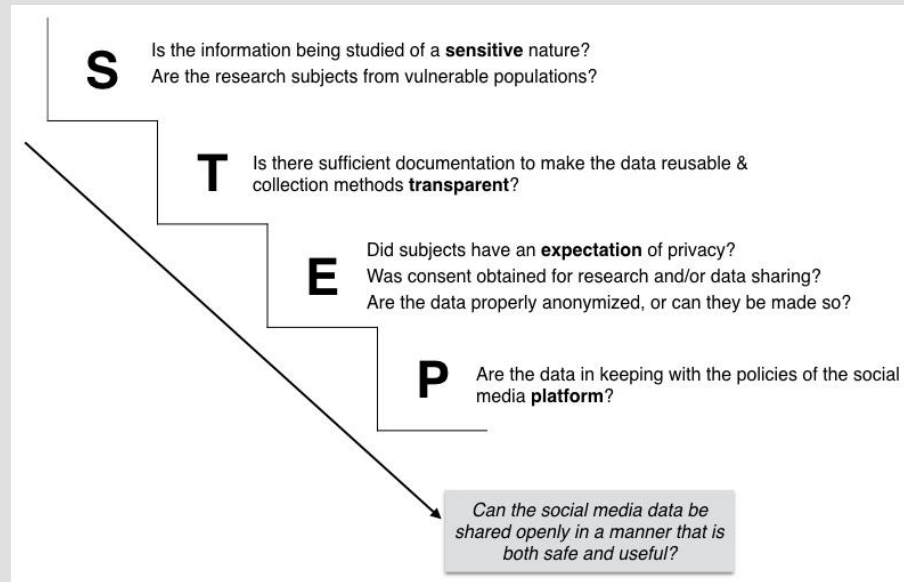
What is your core purpose for using the data?

Who could be negatively affected?

Are you communicating potential risks/issues, if any?

# Our previous research

## The STEP Framework for social media data curation





*Can the data be shared openly in a manner that is both safe and useful?*

**E**xpectation of privacy

**D**ata source

**I**dentification

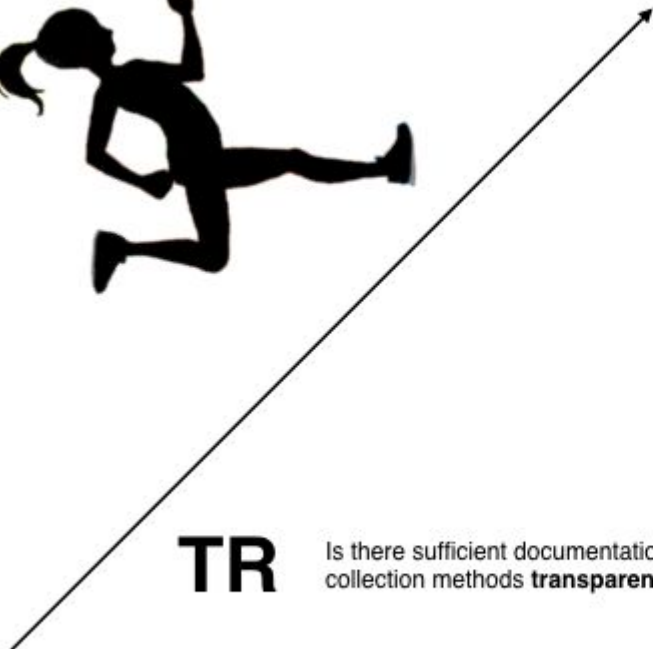
**TR**ansparency

**S**ensitivity



**S**

Is the information being studied of a **sensitive** nature?  
Are the research subjects from vulnerable populations?



**TR**

Is there sufficient documentation to make the data reusable & collection methods **transparent**?





Are the data properly de-identified, or can they be made so?

© 2017 Dryad. All rights reserved. Dryad is a registered trademark of Dryad. All other trademarks are the property of their respective owners.



**D**

What is the **data source**, or the context in which it was collected?



**E**

Did subjects have an **expectation** of privacy?  
Was consent obtained for research and/or data sharing?



*Can the data be shared openly in a manner that is both safe and useful?*

**E**

Did subjects have an **expectation** of privacy?  
Was consent obtained for research and/or data sharing?

**D**

What is the **data source**, or the context in which it was collected?

**I**

Are the data properly de-**identified**, or can they be made so?

**TR**

Is there sufficient documentation to make the data reusable & collection methods **transparent**?

**S**

Is the information being studied of a **sensitive** nature?  
Are the research subjects from vulnerable populations?

# Guiding principles

## STRIDE framework

1. **RISK-BENEFIT ANALYSIS.** When sharing big data, researchers and data curators must measure the benefits of sharing data against the potential risks to human subjects.

# Guiding principles

## STRIDE framework

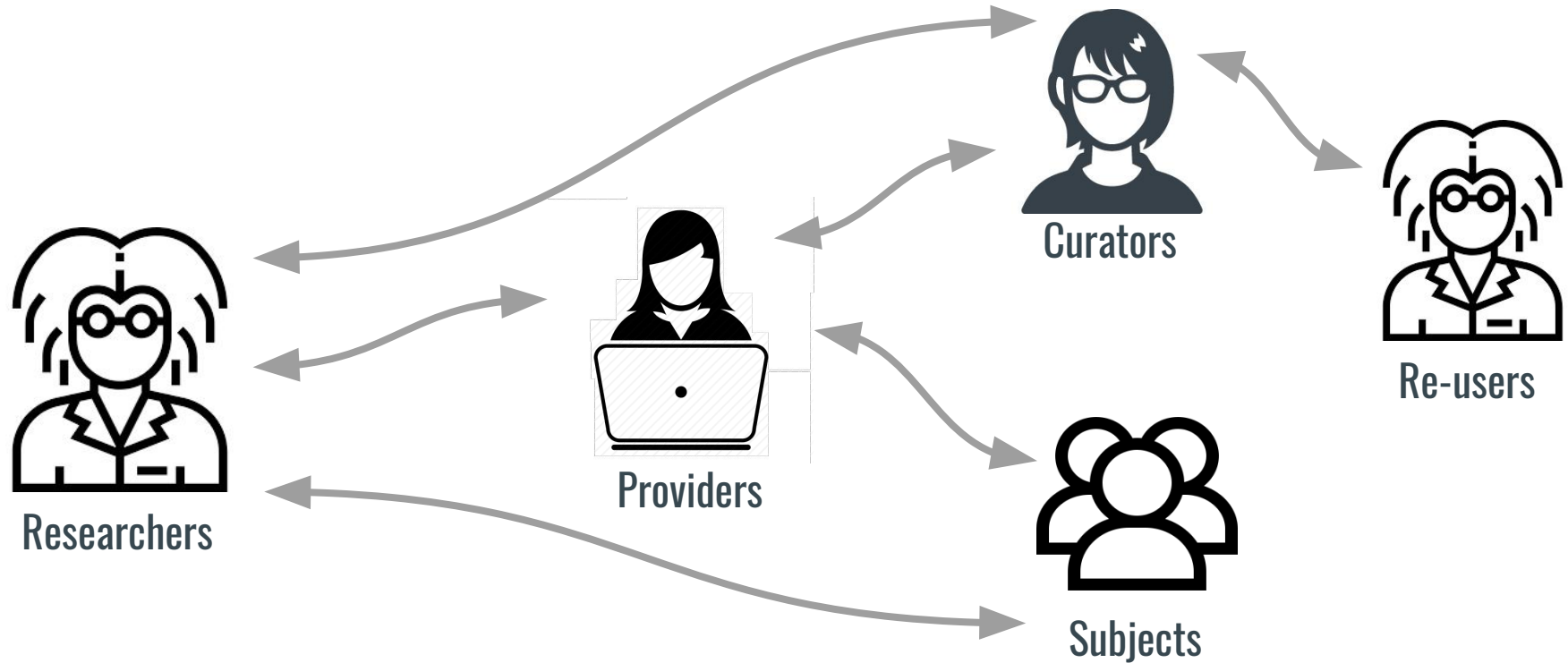
1. **RISK-BENEFIT ANALYSIS.** When sharing big data, researchers and data curators must measure the benefits of sharing data against the potential risks to human subjects.
2. **RESPONSIBILITY.** Data curators can help educate researchers about ethical data sharing, but researchers themselves are ultimately responsible for the data they share.

# Guiding principles

## STRIDE framework

1. **RISK-BENEFIT ANALYSIS.** When sharing big data, researchers and data curators must measure the benefits of sharing data against the potential risks to human subjects.
2. **RESPONSIBILITY.** Data curators can help educate researchers about ethical data sharing, but researchers themselves are ultimately responsible for the data they share.
3. **CONTINUAL INQUIRY.** Ethical practice requires ongoing dialogue and examination.

# Humans in the open data ecosystem







About ▾

For researchers ▾

For organizations ▾

Contact us

Log in

Sign up



**DataDryad.org is a curated general-purpose repository that makes the data underlying scientific publications discoverable, freely reusable, and citable. Dryad has integrated data submission for a growing list of journals; submission of data from other publications is also welcome.**



Submit data now

[How and why?](#)

### Search for data

Enter keyword, author, title, DOI, etc

[Advanced search](#)

### Browse for data

Recently published

Popular

By author

By journal

#### Recently published data

Pracana R, Priyam A, Levantis I, Nichols R, Wurm Y (2017) Data from: The fire ant social chromosome supergene variant Sb shows low diversity but high divergence from SB. *Molecular Ecology* <http://dx.doi.org/10.5061/dryad.js509>

Kondor D, Grauwin S, Kallus Z, Gódor I, Sobolevsky S, Ratti C (2017) Data from: Prediction limits of mobile phone activity modeling. *Royal Society Open Science* <http://dx.doi.org/10.5061/dryad.2t3t7>

Hudgins EJ, Liebhold AM, Leung B (2017) Data from: Predicting the spread of all invasive forest beetles in the United States. *Ecology Letters*

### Latest from @datadryad

Tweets by @datadryad

Dryad Retweeted



**Alexander Naydenov**  
@vremigrant

Five new @Pensoft journals integrated with @datadryad to improve data discoverability [blog.pensoft.net/2017/02/06/fiv...](http://blog.pensoft.net/2017/02/06/fiv...)



<http://datadryad.org>



# Dryad's mission and vision

Dryad is a **curated, general-purpose** home for a wide diversity of data types associated with scientific and medical publications.

Dryad's **vision** is to promote a world where research data is openly available, integrated with the scholarly literature, and routinely re-used to create knowledge.

Our **mission** is to provide the infrastructure for, and promote the re-use of, data underlying the scholarly literature.

<http://datadryad.org>

# Case study #1

Sci-Hub



Elbakyan A, Bohannon J (2016) **Who's downloading pirated papers? EVERYONE**

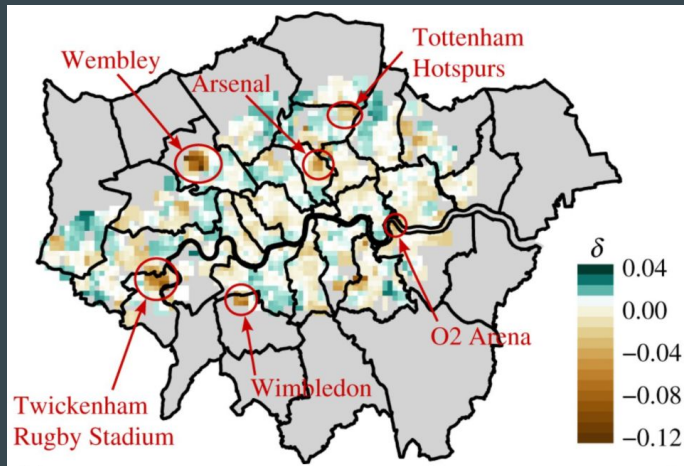
- Data obtained directly from server logs: “To let the world see how Sci-Hub is being used, mapping users at the highest resolution possible while protecting their privacy”
- IP addresses replaced with an arbitrary hex code
- Geographic locations of users aggregated to the nearest cities

[Dryad data; blog post](#)

[Article: Science 352\(6285\): 508-512](#)

# Case study #2

## Spatio-temporal patterns of mobile phone activity



### Kondor et al (2017) **Prediction limits of mobile phone activity modelling**

- 10 months of mobile phone records from London used to investigate the regularity of human telecom activity on urban scales
- Data release agreement with provider
- Used “time averaging and spatial smoothing” to protect privacy

### Dryad data

Article: Royal Society Open Science 4(2): 160900

# Case study #3

All the Twitter feels



Charlton et al (2016) **In the mood: the dynamics of collective sentiments on Twitter**

- Study of “the relationship between the sentiment levels of Twitter users and the evolving network structure that the users created by @-mentioning”
- Data included anonymized user IDs but exact timestamps; no cross-referencing with other identifiable data
- “Private” conversations on a public platform

Dryad data

Article: Royal Society Open Science 3(6): 160162

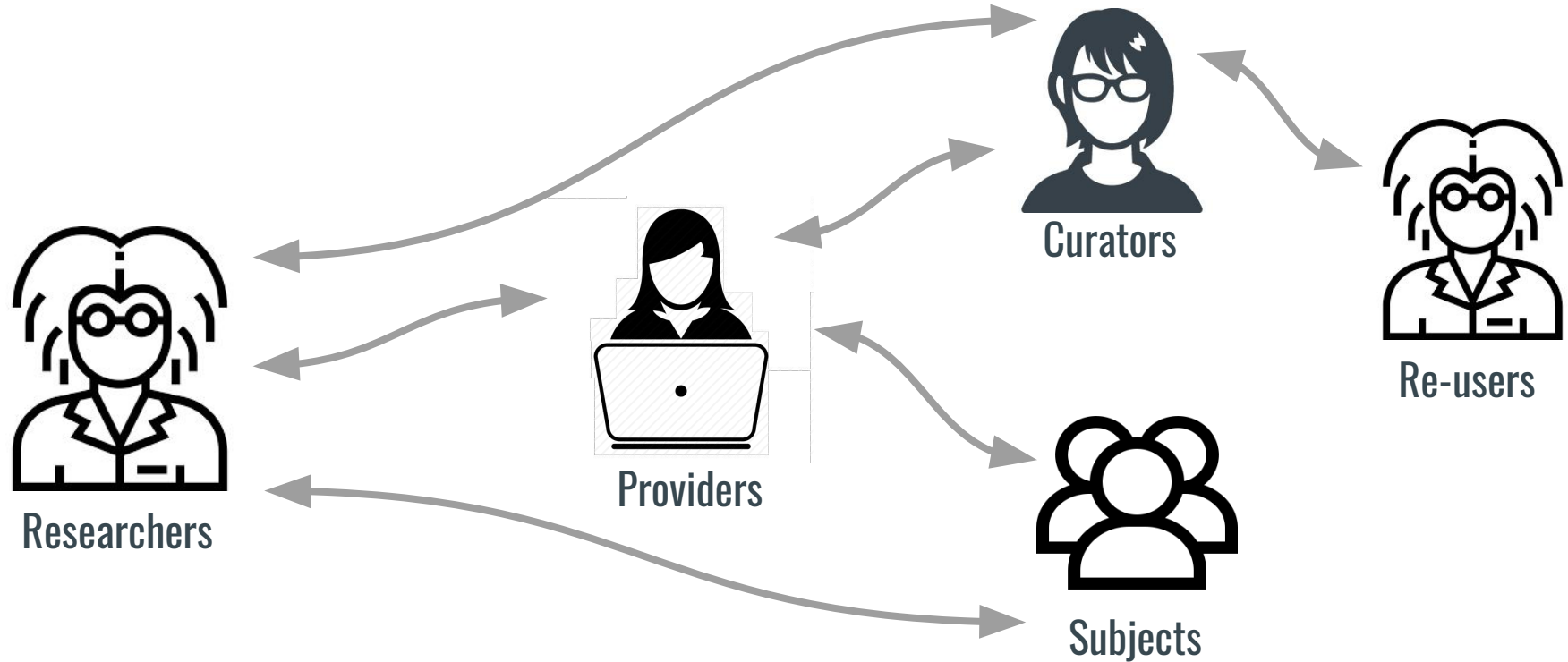
# AREA

# Ethical considerations for big data research

*“One of the most fundamental rules of responsible big data research is the steadfast recognition that most data represent or impact **people**.”*

—Zook et al., 2017

# Humans in the open data ecosystem





# Striding onward



# References

boyd d, Crawford K (2012) Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society* 15(5): 662-679. <https://doi.org/10.1080/1369118X.2012.678878>

Crawford K, Miltner K, Gray ML (2014) Critiquing big data: politics, ethics, epistemology. *International Journal of Communication* 8. <http://ijoc.org/index.php/ijoc/article/view/2167/1164>

Mannheimer S, Hull EA (2017) Sharing selves: developing an ethical framework for curating social media data. *International Data Curation Conference*, Edinburgh, February 20-23. <http://scholarworks.montana.edu/xmlui/handle/1/12661>

Metcalf J, Crawford K (2016) Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society* 3(1). <https://doi.org/10.1177/2053951716650211>

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979) *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects*. Washington, DC: US Government Printing Office. <https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/>

Zook M, Barocas S, Crawford K, Keller E, Gangadharan SP, Goodman A, Hollander R, Koenig BA, Metcalf J, Narayanan A, Nelson A, Pasquale F (2017) Ten simple rules for responsible big data research. *PLOS Computational Biology* 13(3): e1005399. <https://doi.org/10.1371/journal.pcbi.1005399>