

# Introduction to open-sourcing data

Matthew J. Salganik  
Department of Sociology  
Princeton University

Summer Institutes in Computational Social Science  
2020

The Summer Institutes in Computational Social Science is supported by grants from the Russell Sage Foundation and the Alfred P. Sloan Foundation.



To wrap-up activity:

- ▶ Pay your MTurk workers
- ▶ Document and release your data to your local organizer



[https://www.youtube.com/watch?v=66oNv\\_DJuPc](https://www.youtube.com/watch?v=66oNv_DJuPc)

Brief introduction into open-sourcing your data:

- ▶ Store your data in a simple format

Brief introduction into open-sourcing your data:

- ▶ Store your data in a simple format
- ▶ Provide documentation

Brief introduction into open-sourcing your data:

- ▶ Store your data in a simple format
- ▶ Provide documentation
- ▶ Beware of privacy

Beware of privacy

# Beware of privacy

Risks come from combining data sources

$\underbrace{\text{Baking soda}}_{\text{Safe}} + \underbrace{\text{Vinegar}}_{\text{Safe}} =$





# Beware of privacy

Remove personally identifying information and/or information that could be used for linking

## NIST definition [\[ edit \]](#)

---

The following data, often used for the express purpose of distinguishing individual identity, clearly classify as PII under the definition used by the [National Institute of Standards and Technology](#) (described in detail below):<sup>[15]</sup>

- [Full name](#) (if not common)
- [Face](#) (sometimes)
- [Home address](#)
- [Email address](#) (if private from an association/club membership, etc.)
- [National identification number](#) (e.g., [Social Security number](#) in the U.S.)
- [Passport number](#)
- [Vehicle registration plate](#) number
- [Driver's license](#) number
- [Face](#), [fingerprints](#), or [handwriting](#)
- [Credit card numbers](#)
- [Digital identity](#)
- [Date of birth](#)
- [Birthplace](#)
- [Genetic information](#)
- [Telephone number](#)
- [Login name](#), [screen name](#), [nickname](#), or [handle](#)

[https://en.wikipedia.org/wiki/Personally\\_identifiable\\_information](https://en.wikipedia.org/wiki/Personally_identifiable_information)

# Privacy and Security

## Myths and Fallacies of “Personally Identifiable Information”

<http://dx.doi.org/10.1145/1743546.1743558>

In this case, we recommend:

- ▶ Removing PII (name, email address, etc)
- ▶ Removing TurkID
- ▶ Coarsen age, geography, and race/ethnicity
- ▶ Coarsen timestamp
- ▶ Anything else?

For more on coarsening, see this code:

[https://github.com/compsocialscience/summer-institute/blob/master/2019/materials/day4-surveys/activity/mturk\\_data\\_cleaning.Rmd](https://github.com/compsocialscience/summer-institute/blob/master/2019/materials/day4-surveys/activity/mturk_data_cleaning.Rmd)

For more about de-identification, see

- ▶ *Bit by Bit*, Sec 6.6.2 “Understanding and managing informational risk”
- ▶ Lundberg, Levy, Narayanan, Salganik (2019) “Privacy, ethics, and data access: A case study of the Fragile Families Challenge”

In this case, you should share your data with your local organizer who can post it here:  
<https://github.com/compsocialscience/summer-institute/tree/master/2020/materials/day4-surveys/datasets>

When you start your projects next week

- ▶ plan to release your data
- ▶ plan to release your code

# Questions