Privacy issues with gathering and sharing measurement data

Scott Jordan
University of California, Irvine

* nothing in here is legal advice

Two discussion questions

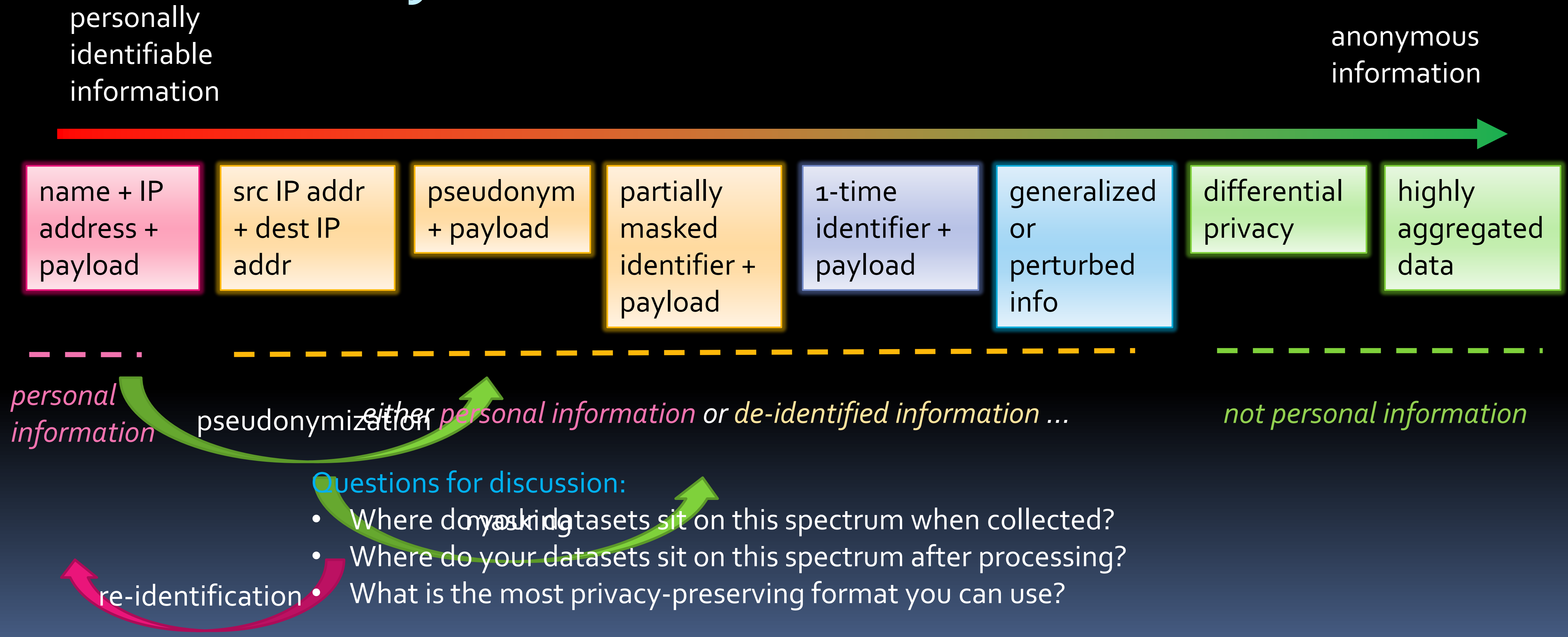
[1] What is the most privacy-preserving form of data you can use and still be able to answer your research questions?

[2] What is a reasonable code of conduct?

Is privacy an issue?

- Does the data relate to a person?
  - if not, it's not personal information
    - example: traceroute data, if you (not the person) chose the destinations (active)
- Is the information publicly available?
  - if so, it's not private information
    - example: DNS TLD zone files, if they have been made available to the general public

What is the most privacy-preserving form of data you can use?



personally identifiable information

anonymous information

name + IP address + payload

src IP addr + dest IP addr

pseudonym + payload

partially masked identifier + payload

1-time identifier + payload

generalized or perturbed info

differential privacy

highly aggregated data

personal information

either personal information or de-identified information …

not personal information

pseudonymization

masking

re-identification

Questions for discussion:

- Where do your datasets sit on this spectrum when collected?
- Where do your datasets sit on this spectrum after processing?
- What is the most privacy-preserving format you can use?

What is a reasonable code of conduct?

- Model #1 [de-identification]
    - the information is stored *solely* in a form in which it cannot reasonably be linked to a particular consumer
    - you are contractually obligated to
        - maintain and use the information in de-identified form,
        - not to attempt to re-identify the information
- Question for discussion:
    - what research can't be done without identifiable information?

What is a reasonable code of conduct?

- Model #2 [IRB]
  - for qualified research only
  - the information is stored in two forms:
    - *for most researchers*, in a form in which it cannot reasonably be linked to a particular consumer
    - *for a few select researchers*, in an identifiable form
  - you are contractually obligated to
    - use it solely for research purposes,
    - store it in the most privacy-preserving form that enables the research, and
    - limit access to the means of identification

- Question for discussion:
  - are these the "correct" obligations?

Privacy: form of data

What is the most privacy-preserving form of data you can use and still be able to answer your research questions?

- Presumes identification of research questions
  - Sometimes, measurement researchers start with a dataset and then try to pose and answer questions
- Datasets span the identifiability spectrum:
  - PII:
    - datasets that include identities are likely rare
  - Maybe PII & maybe de-identified:
    - datasets most commonly do not include identified data but include reasonably identifiable data, e.g.
      - passive data with IP addresses
      - partially masked data
      - masked & pseudonymized data (crytopan)

Privacy: form of data

What is the most privacy-preserving form of data you can use and still be able to answer your research questions?

- Datasets span the identifiability spectrum (continued):
  - Likely de-identified:
    - generalized: synthetic data
      - issues: error, bias, research questions that can be answered
    - differential privacy:
      - issues: setup, adaptation to new types of queries, research questions that can be answered
      - (more from Simson)
  - Non-personal information:
    - active measurements often produce non-personal information

- Most research probably requires two forms of the same data
  - Raw version that does not include identified data but is reasonably linkable
  - Processed version that has privacy-preserving techniques applied

Privacy: code of conduct

What is a reasonable code of conduct?

At least two models:

- IRB:
  - most familiar
  - paradigm is a good match
    - for using two forms of data (raw, processed)
    - for implementing a code of conduct:
      - use it solely for research purposes,
      - store it in the most privacy-preserving form that enables the research, and
      - limit access to the means of identification
  - however, IRBs often don’t have the prerequisite expertise (risks, techniques, law)

Privacy: code of conduct

What is a reasonable code of conduct?

At least two models (continued):

- de-identified data:
  - less familiar
    - what qualifies?
    - doesn’t allow for keeping more detailed raw data
  - but captures the promise not to attempt to re-identify the information
- DRB (Simson)?

Privacy: removing impediments

- Cross-fertilization
    - between measurement researchers, privacy researchers, and folks who understand laws/regs
        - measurement researchers need to know what privacy-preserving algorithms to use and what this allows them to do
            - without learning whole new fields

Privacy: removing impediments

- Standardizing a code of conduct
  - code of conduct should
    - align research needs, privacy laws/regs, IRB concerns
    - make it easier to share
    - lessen re-identification attacks
  - standardization would lessen impediments:
    - IRBs
    - campus counsel
    - company counsel
    - incorrect perceptions of privacy laws/regs