Navigating the Data Privacy Maze

Privacy Considerations for Data Scientists

Katharine Jarmul - KIProtect
PyData Edinburgh July 2018

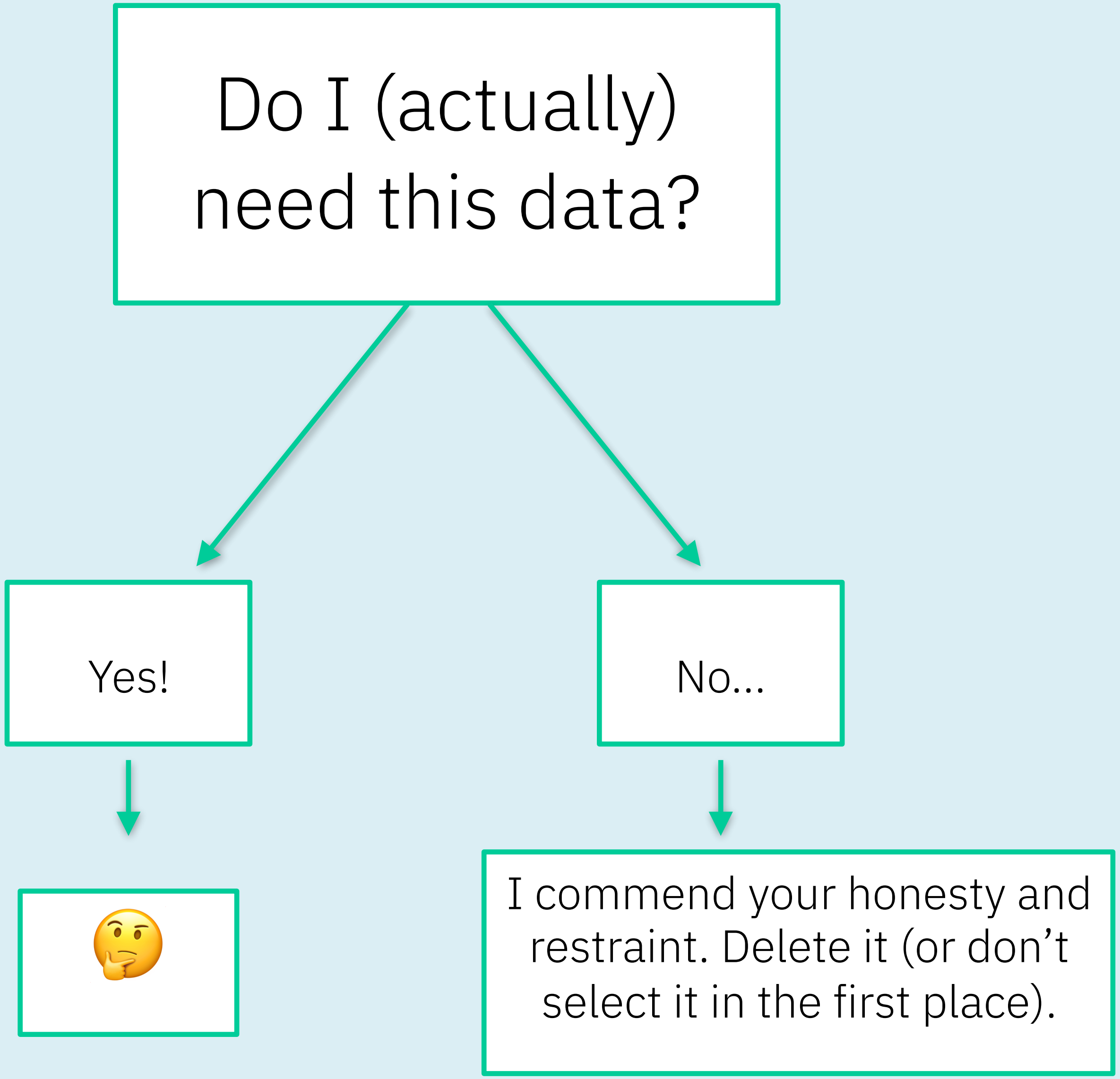
Obligatory GDPR Slide

Credit:
http://memecollection.net/updated-privacy-policy/

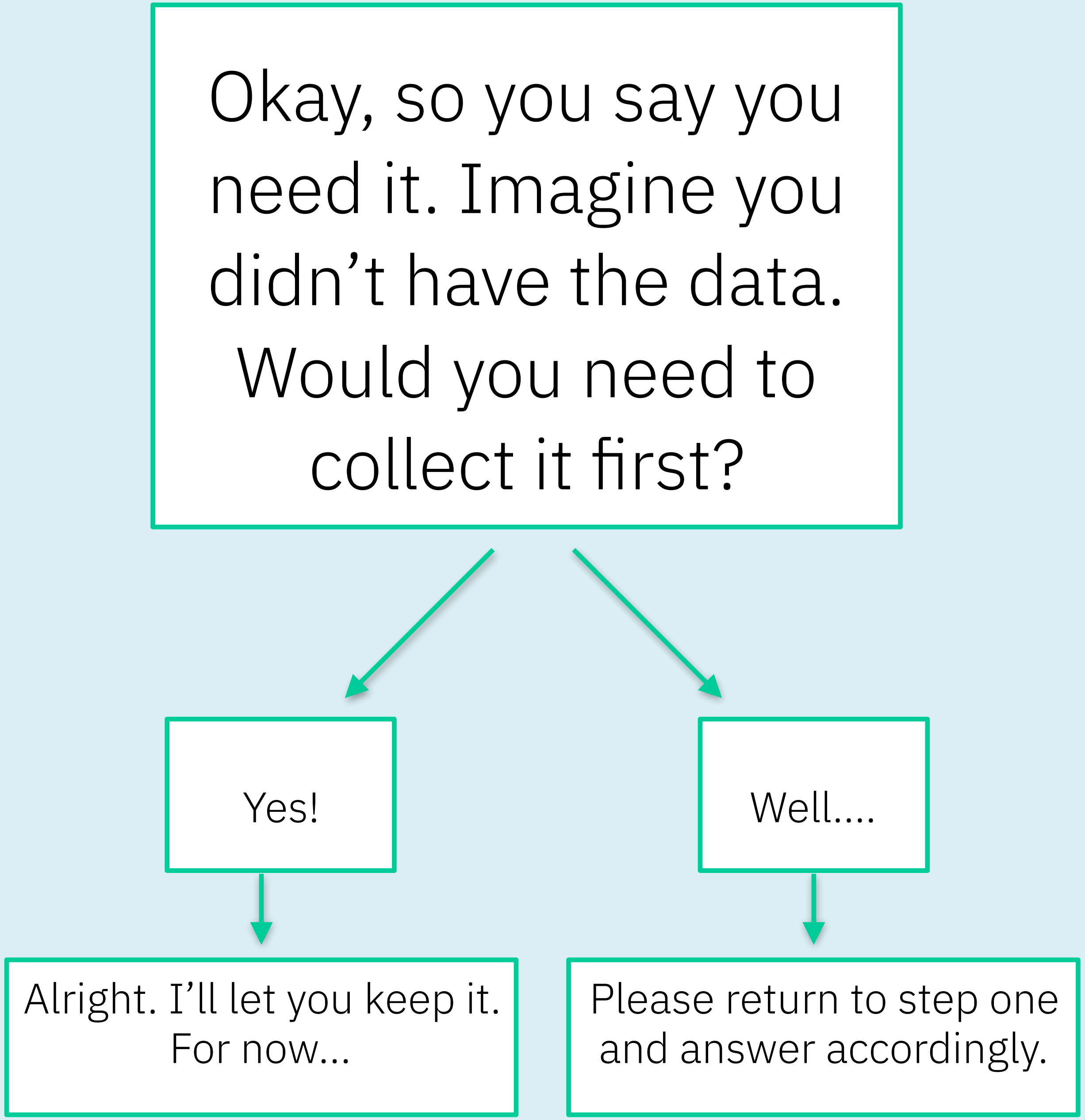
What is Privacy for Data Scientists?

Photo Credits:
Left: (Lily Martin/CBC)
Right: (Thor Swift/NYT)

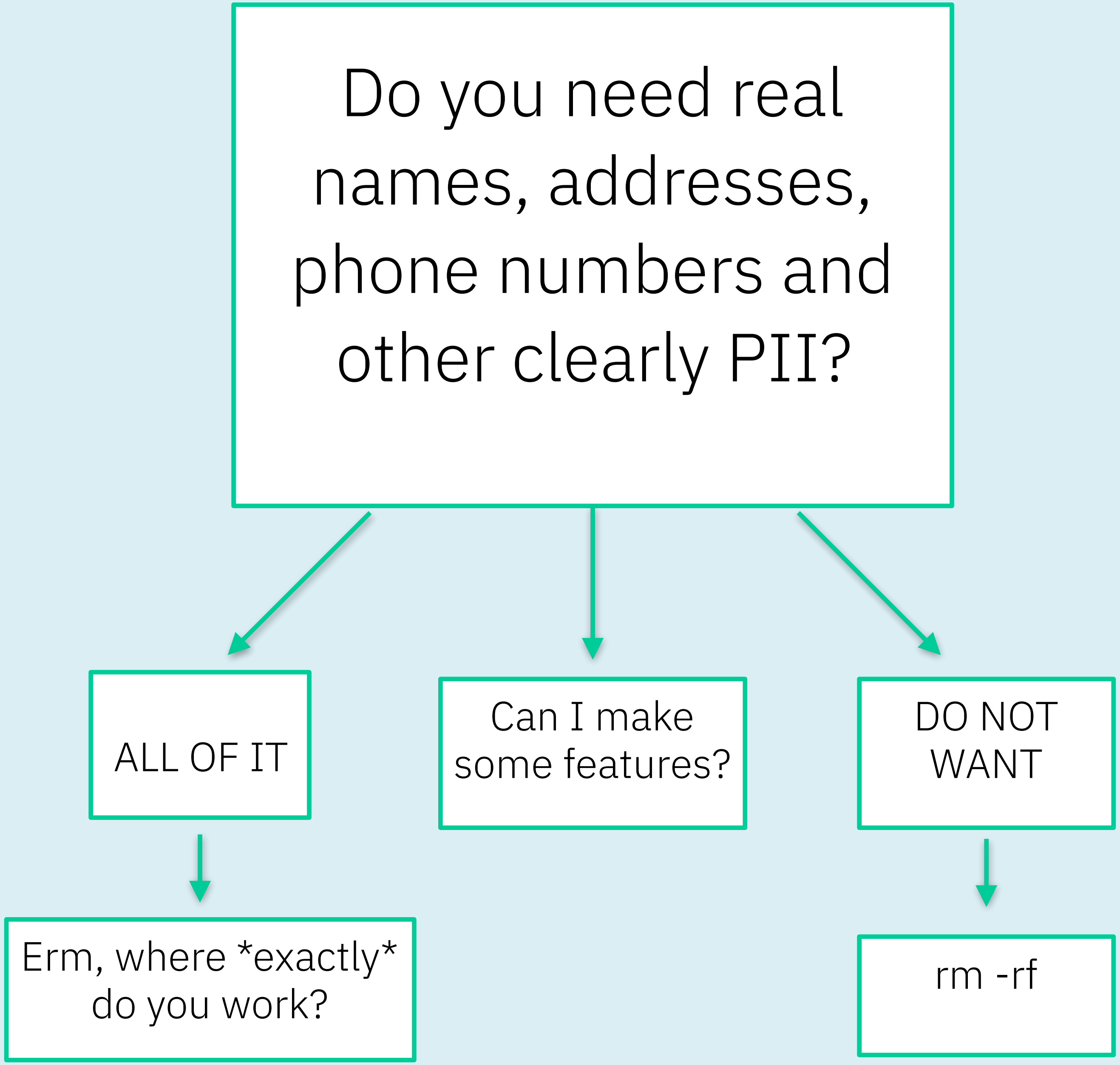
Navigating Data Privacy



Gimme the Data (cont.)

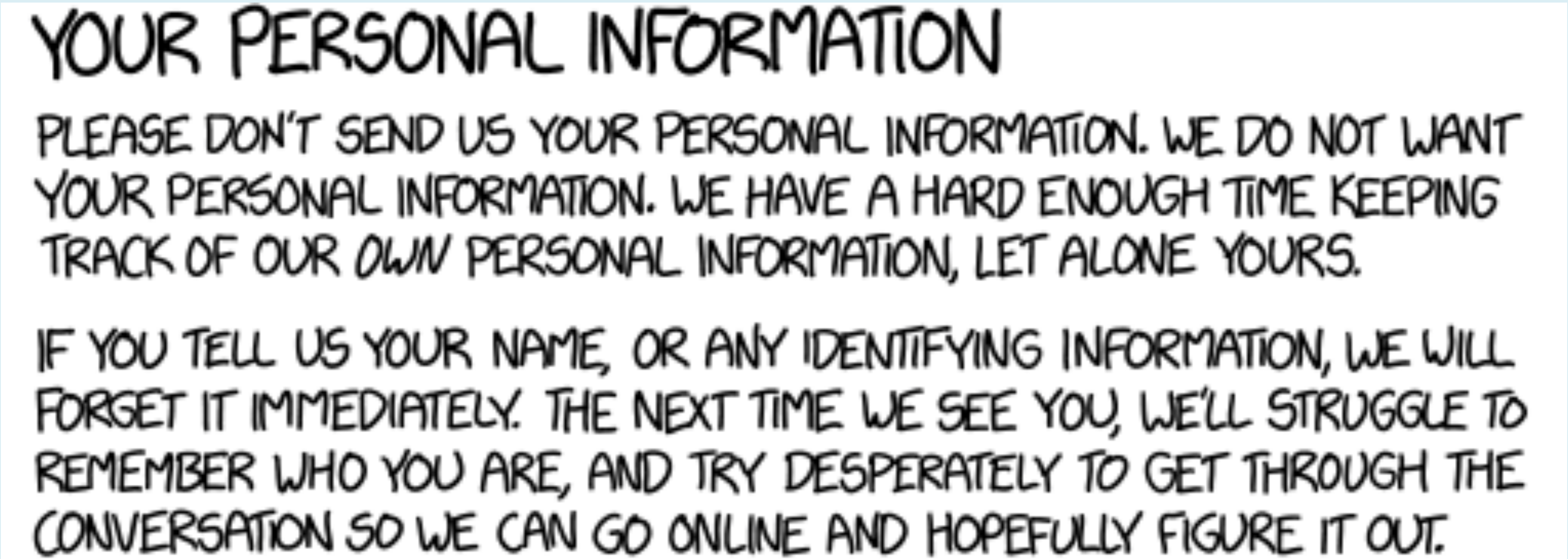


Actually Required Sensitive Data



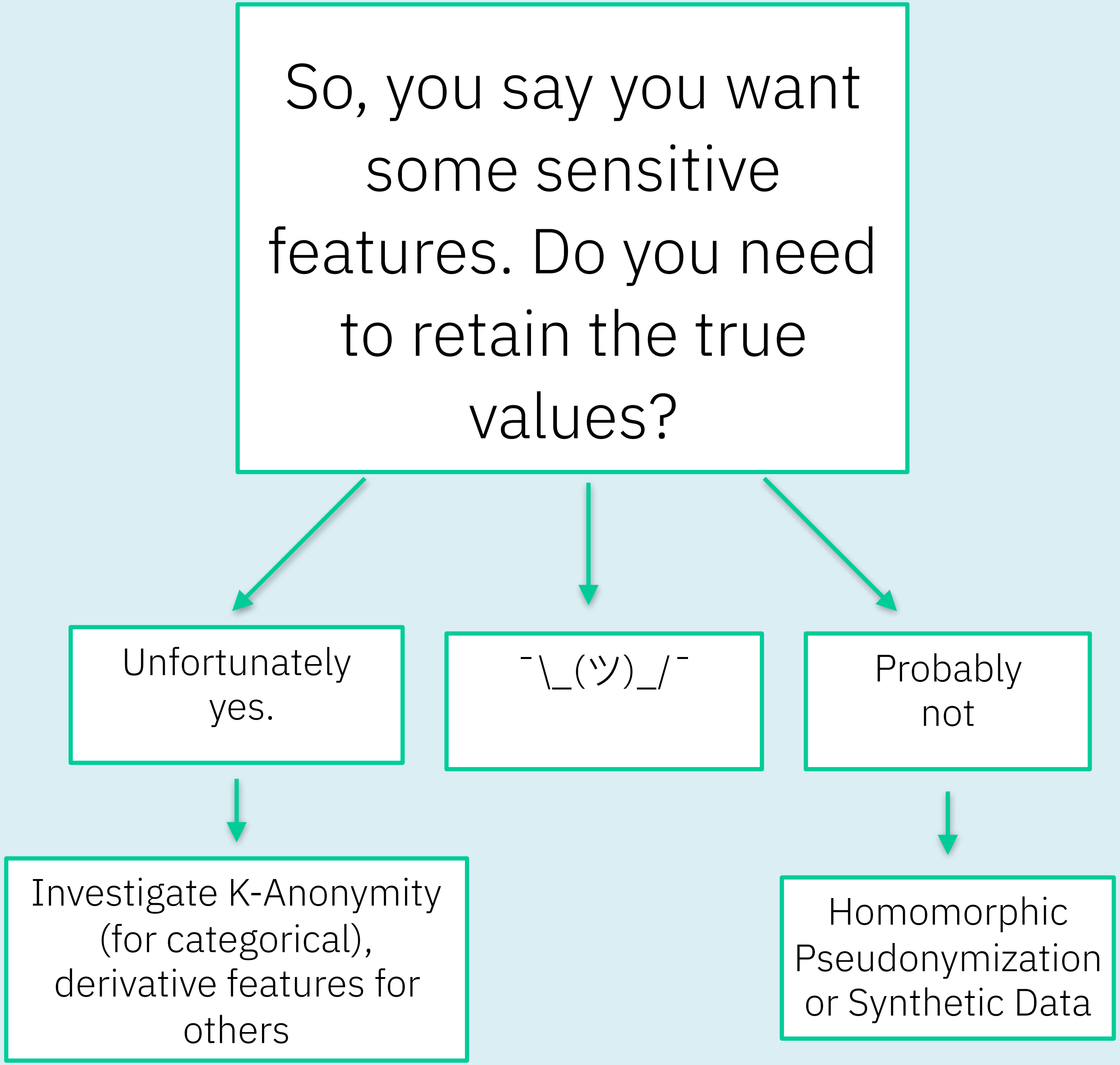
The Case for Deletion

- Best possible protection
- Determine if derivative features (i.e. “uses free email” or “name ends with a vowel”) can convey enough info (and still preserve some privacy)



https://xkcd.com/1998/

Actually Required PII



The Case for K-Anonymity

- Create groups / blocks which allow us to cluster similar individuals together, preserving individual privacy for the grouping (i.e. age: 30-40, larger region vs zip code)
- Usefulness depends on the diversity of your overall dataset and the diversity of the groups
- Does not automatically guarantee against information disclosure about a person or group
  - l-diversity and t-closeness can help with this

Determining Your Data Needs

So you're not sure you need 100% accurate sensitive data…

Are you using it for testing, a demo, exploration?
(or anything else that may not make it to production?)

Yes. I am a researcher.

When possible, use realistic synthetic or pseudonymized data ***first***.

No. I am an engineer.

And you aren't testing? 😭

The Case for Pseudonymization

- Pseudonymization (depending on method) can allow you to preserve individual privacy while still retaining information based on the attribute
- KIProtect Pseudonymization API allows for homomorphic pseudonymization (structure-preserving mechanism)
- Does not protect against statistical attacks or linkage attacks (using outside information to determine identity)

Sharing Your Data / Models

Okay, so it’s not for testing or your experimentation has moved onto the next stage. You need the real data.

Will this data or the model it helps build be shared with other teams?

Yes. Of course.

For ML, you should look into privacy-preserving models or training on anonymized data. For analysis, look at aggregated anonymization.

No. It’s for my team only.

The Case for Privacy-Preserving ML

- If you must learn on private data, please utilize privacy-preserving mechanisms
- Determine what the attack vector is:
  - Do you trust your MLaaS provider?
  - Do you trust your API users? (i.e. black box access)
  - Do you trust people with access to the model or training mechanism itself? (i.e. white box access)
- Active area of research!

Data and Model Security, Anyone?

So you had an internal-only model or dataset that the company now wants to "monetize," so this means it might go to partners, other teams or in production...

Can you fix it before it hits production / publish?

Yes. I can retrain or edit the dataset.

For ML, you should look into privacy-preserving models or training on anonymized data. For analysis, look at aggregated anonymization.

Erm...

Please make sure to lock down the API or access.

The Case for Anonymization

- If you are releasing data to an untrusted audience or public audience, please employ anonymization!
- Deletion of sensitive fields can help, but is not enough (long-tail distribution / linkage)
- Differential privacy can be used to regulate the amount of information w.r.t. a single variable or attribute
- Aggregated anonymization can help preserve privacy but still allow for group insights (i.e. Apple Differential Privacy for Emoji Use)

End Story: Treat Your Data Like Radiation

- Try to remove as much extraneous data as possible.
- Think: Do I really need this much radiation? Probably not, let’s just have the minimal amount of radiation we need.
- Use techniques to make the radiation less harmful — understanding there are a wide variety of options and levels and you can use to determine the range of exposure.
- (raw data —> pseudonymization —> anonymization —> deletion)

Thank you for your time!

Questions? I'd Love to hear them!

Or reach out anytime:

info@kiprotect.com
@KIProtect (Twitter)
https://github.com/kiprotect

Katharine Jarmul
katharine@kiprotect.com
@kjam (Twitter)

7scientists GmbH
KIProtect
Bismarckstr. 10-12
10625 Berlin