Data Science Project Survey Outline:

**Goal:**

* Answer these questions,

**Research Questions:**

Are users comfortable with the service’s privacy defaults?

Are users aware of the data Company X collects about them?

Suppose users were not aware. Would their concerns be more serious after explaining to them that Company X does, indeed, collect data about them?

Will users’ privacy concerns grow if they see the data Company X collects about each of them?

Survey Form:

Prior to handing over the survey form, users were asked if they would consent to taking a survey concerning the data privacy associated with Amazon accounts.

Q0: Do you have an amazon account? How regularly do you use Amazon or its products?

* Frequently
* Infrequently
* Not at all

Q1: How familiar are you with the privacy terms surrounding your Amazon account?

Range from 1 (not at all), 2(vague details), 3(a general idea), 4(mostly informed), 5(very familiar)

Q2: Do you think that these terms are sufficient to protect your personal information?

Range from 1 (I have or want to submit complaints), 2 (I am worried, but not enough to act), 3 (I’m not sure), 4 (I am unconcerned), 5 (My data is safe with Amazon)

Q2.5: Do you know that Amazon collects data about you?

* Yes
* No

Q3: Do you know how to access your Amazon data?

* No
* Basic Account Information
* Advanced Information

Q4: What kind of data do you think Amazon collects about you?

Tick boxes:

* Website interaction
* Search history
* Order/return history
* ~~Alexa/Echo/Ring~~
* Kindle books/downloads
* ~~Advertising~~
* Wishlist
* ~~Amazon Media (books, games, movies, etc.)~~
* Physical Store interactions
* ~~Addresses~~
* Phone Numbers
* ~~Customer Reviews~~
* Prime Membership
* ~~Email Interactions~~
* Photos

Example: Share list with them and a sample of data from my Amazon Kindle (ie. Reading times)

Q4: Knowing this, does that change your answer to Q2? And if so, to what?

* Range 1 – 5

Q5: Do you think you will engage with Amazon content differently after this survey?

* Yes
* No

Differential privacy:

* How to leak useful information without disclosing private information
* Fundamental Law of information Recovery states it is impossible to release useful information without leaking private information
* Need to Quantify and bound information disclosed about an individual
* **Goal**: perform data analysis over a dataset without producing harm to any individual whose record is in the dataset
  + Identification is key to harm -> they can’t harm you if they do not know who you are
  + This definition is unworkable because:
* Better Goal: Nothing about an individual is learned from a dataset, D, that cannot be learned from the same dataset but without the individual’s data, D
* Differential Privacy def:
  + It is not possible to tell if the input to an algorithm, A, contained an individual’s data or not just by looking at the output of A
  + Ie. No one can learn much about one individual from the dataset
  + Including your data in a dataset does not increase your chances of being harmed
* For Every pair of input datasets, D1, D2, that differ in one row, the presence or absence of that single record does not change the result
* You use a noise formula to limit information shared, but the more adjusted it is, the less useful it is