Alan Meeting Notes

Week 9

* Laplace noise: how is noise drawn from it, what about gaussian noise, why use Laplace and gaussian as opposed to others?
* Composition: break this down into stuff that makes sense in a basic way; avoid the high level language unless it’s explained really well.
* Extensive research; how would I explain it? How would I implement it in the context of research data analysis; this forms part of the design of the project:
  + “here’s a cool theoretical framework, I understand w a normal example, now I’m going to apply what I understand to designing a system that can be implemented for research data because it meets xyz criteria”.
    - Thinks I’m tackling something very difficult for an undergraduate student 😊.
* Alan’s questions/things he was confused about:
  + “Reverse the randomised algorithm? If you want the actual data out of the database (as a trustworthy user) you need to reverse the randomised noise
    - My thoughts: maybe only add the randomised noise to statistical analysis? Keep the accurate database for high-level authorised users and use differential privacy for data analysts that just want to analyse the population
  + “if you had 1 million peoples’ bank account details, and you inject noise into the result to make it differentially private. But the actual bank balances and other bank details need to be accurate. So how do you recover the actual details?”
    - The noise is only added to statistical analyses (?) so the actual details are untouched, whereas queries on the data set are altered.
  + “looking for something that we can replicate and implement that is easy to understand.”
    - Have to make this into English.

Week 10

* Explain DP in a high-level to someone who is smart, but at the same time be able to dumb down the jargon.
* “starting to like the research aspects of DP. Was a little bit worried it would be trash.”
* Mock database of HIPAA identifiers.
  + Already completed.
  + Generate a mock database to muck around with, make a few queries and get familiar with the Matlab-SQLite connection.
* Build confidence with open-source examples; if someone during Q & A were to ask a question about any aspect of how it works, know how to answer it (at least to a basic extent).
* Markers will most likely be 2 people that know absolutely nothing about this subject.
  + “come off as knowledgeable so that someone smart will understand what I’m talking about. Also so if they ask a question, I know how to give a knowledgeable answer”.
* For presentation:
  + 12 minute presentation, have roughly 10-20 slides prepared. Prepare slides as I’m researching, have it follow the same structure as thesis.

Week 12

* Thesis draft due Tuesday afternoon/night. Alan said he’d find time Wednesday morning to assess it. Specify which sections.
* Thesis document due Friday.
* Alan’s questions/things he was confused about:
  + Smartnoise example: “epsilon is the value of privacy loss, so if there is lower privacy loss why does the error get higher?”
* Neither of my markers will be Alan (fucking rip)
* Remember to pick up the meeting log from Alan
* Make the thesis enjoyable to read. Tell a story. If it’s interesting to read, the markers won’t want to cut themselves while marking it.
* Create a Gantt chart for the 13 weeks of next semester. Include achievable milestones.
  + “Week 1, I plan to do this, week 2 I plan to do this, etc.”
  + DON’T JUST SAY “Week 1 I plan to write code”, that’s vague as fuck, be specific.
  + SMART acronym for planning ahead.
  + No but seriously, with your timeline of measurable milestones, don’t even think of writing “will write code”.
* Other student: 40+ pages, included 3 mini experiments, results section and discussion. Rock solid HD.
* Logical flow of progression in thesis document.
  + First sentence should define the paragraph. First sentence should also lead from the last sentence of the previous paragraph.
* Chunks of information do not form a good story. Want to tell a story of the 13 weeks of research this semester:
  + Here’s the problem I’m trying to solve for this thesis
  + This is why it’s interesting
  + Why it’s a challenging problem
  + Why it’s an engineering problem
  + Here’s the research I’ve performed on how other people have tried to solve it
    - I’ve come across differential privacy, it looks promising because it does xyz.
  + Method section:
    - Here’s some code that has allowed me to explore DP
    - I’ve inputted some fake data that is relevant and here are the results that I got from it.
    - I’m going to use matlab to
    - I’m going to use the open source code to explore DP, set up the parameters that I need for my actual implementation. “what is an acceptable amount of privacy loss. I’m going to explore it for a data set that is like what I’m going to be using. I’m going to be writing most of my code in matlab using a link to SQLite, here’s my exploration of this to make sure that matlab can interact with SQLite and that I know how to perform queries. So I’ve done the ground work to, in the first week of the next semester, get myself ready to hit the ground running in doing part B of the thesis.”
  + All prelim work should show that I’ve set myself up to go.
  + “This is what I’ve learned, why it’s good and I can apply the theory in this open-source software to a dataset that is close to what I intend to do.”
* Working dataset to play with is the ‘500 entries mock dataset’.

Auxiliary notes:

* Alan’s example of credit scores in the US with Experion hack; “credit score essentially determines if you can get a loan from a bank. Credit score is computed from financial information. Database was hacked. If you can get a hold of someone’s name, social security number, date of birth and address you can apply for credit on their behalf”.
  + If you had a good credit score, they could really mess that up for you.
  + Real-world case study on the importance of data privacy.
* The definition of differential privacy with a solid example.
* “How would you explain this all to your grandmother?”