**How the fuck does Differential Privacy work?**

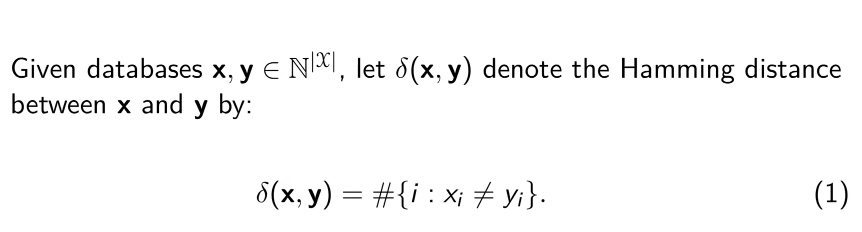
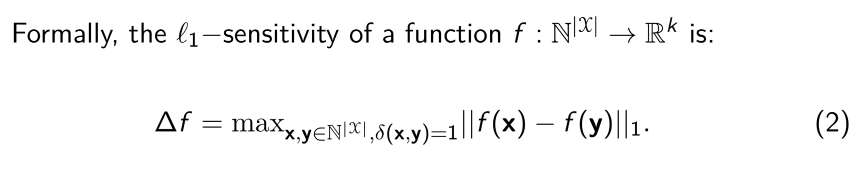
**Differential privacy overview #1: introduction**

<https://www.youtube.com/watch?v=fNHdTcUkC0I>

* Synthetic microdata
  + Record-level data
* Synthetic data is driven by modelling.
* Differential privacy
  + A formal mathematical framework to provide privacy protection guarantees.
  + The main focus is on summary statistics. We are not adding noise to the data itself; we are adding noise to summary statistics. We are making sure that the noise-added-version of the summary statistics are providing privacy protection guarantee.

**Differential privacy overview #1: definitions**

<https://www.youtube.com/watch?v=XSUYLVCu3gg>

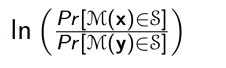
* Key idea: add noise to the output of queries made to databases.
  + Added noise is random; depends on predetermined privacy budget and the type of queries.
* Databases
  + Are data sets that data analysts use for analysis.
  + Are confidential.
  + How can the database holder provide information to the data analyst that is useful and also privacy protected?
* Query
  + Denote numeric queries as functions *f.*  We map databases to *k* real numbers.
  + E.g. how many Native Australians are in this sample of data.
  + Noise is added to the query output for privacy protection.
* Hamming-distance
  + 
  + Under differential privacy, we add noise by considering the scenario where two databases differ by one record i.e. δ(x, y) = 1.
    - Also referred to as adjacent databases.
  + Differential privacy is trying to add noise to an output such that any output from two adjacent databases are indistinguishable.
* L1-sensitivity
  + The magnitude that a single individual’s data can change the l1 norm of the function *f* in the worst case.
  + 
  + “If you have 2 databases that at max differ by 1 entry, what’s the maximum possible change of f(x) (i.e. the query’s output)? It’s the l1-sensitivity.”
  + The l1-sensitivity depends on the query you’re looking at.
  + E.g. 1: suppose x is the confidential sample, y is the database where one data entry is different from x.

Query *f*: count the amount of deaf people in a sample.

What is the l1-sensitivity?

* + - 1 is the maximum change you could get, although there may also be a scenario where there is no change.
    - To get y, we’ve changed a data point in x. In x, a person may have been deaf, and in y this may have been changed to show that the person is not deaf.
  + E.g. 2: query *F*: what is the average income of a sample?

What is the l1-sensitivty; what is the maximum change in income if there is a one record difference between x and y.

* + - We’re thinking about the change in average here, so thinking of the range of possibilities divided by the number (n) of entries is a good start.
    - 
    - Depends on how the database holder is thinking in terms of what the range could be.
  + In summary, the l1-sensitivity depends on the database and the query sent to the database by the analyst.
* Epsilon-differential privacy
  + Want to guarantee that a mechanism behaves similarly (i.e. giving similar outputs) on similar inputs (e.g. when two databases differ by one entry).
  + One approach:
    - Bound the log ratio of the probabilities of the outputs from above.
    - Give an upper bound on the noise added to the output to preserve privacy.
  + The ratio: 
    - Is the log of the ratio of the probability of the output undergone mechanism M from the database x, and that from the database y.
    - This can be considered as the difference in the outputs.
* The privacy budget
  + The term epsilon is the privacy budget that is to be spent by the database holder when answering queries.
  + Pre-determined.
  + If there is only one query, should all the privacy budget be spent?

**Differential privacy overview #1: implications**

<https://www.youtube.com/watch?v=TwJ4RKcmNlk>

* With a given privacy budget, we can add noise according to the e-differential privacy definition to the output.

Sensitivity and added noise

* The l1-sensitivity of a query (function) f is to capture the magnitude a single individual’s data can change the l1 norm of the query f in the worst case.
* l1-sensitivity depends on:
  + the database
  + the query

**Differential privacy overview #2: the Laplace Mechanism**

<https://www.youtube.com/watch?v=H5G0eMZchic&t=394s>

2 Big topics are the Laplace mechanism and properties of differential privacy.

* Key idea: add noise to the output of queries made to databases.
  + Added noise is random; depends on predetermined privacy budget and the type of queries.
    - Privacy budget is what epsilon is used for.
    - The privacy budget is independent of the query.
  + 2 important implications:
    - The added noise is positively related to the sensitivity.
    - The added noise is negatively related to the privacy budget.