Differential Privacy

# Introduction

Data Science’s core resource is data, and the need to gather more of it to improve the models is constant. But data is a double edged sword, it’s like the new oil: it’s very valuable but leaking it causes a lot of damage. Even if the damage is usually not environmental (even if an incorrect model might trigger the wrong chain of event leading to damages to the environment), it can cause a lot of damage to the individual at the source of the data. For example, the Equifax breach dating back to the summer 2017 leaked various key information (names, dates of birth, Social Security numbers and addresses) of about 143 millions Americans; these information are also used as unique identifier across many industries like banking or utilities. The immediate consequence of this breach has been the skyrocketing sale of Identity Protection services, but the long term consequences are still unknown. Another form of leakage, which is usually more insidious, is the possibility to infer the identity of individuals or entities through key markers in data aggregates. An example can be found with the Netflix Prize challenge: even though the data was anonymized, by matching the data provided in this dataset (date of liking a certain movie, rating of the movie, etc.) with publicly available data, it was possible to deanonymize it with a high level of confidence.

This raise the question of how to protect individuals from potential data leakages, what techniques are available to us today, what are their advantages and drawbacks, and ……… FIND OUT THE LAST PART …….

We are going to focus in this paper on Differential Privacy, a mathematical method developed in the 2005 by Cynthia Dwork and Frank McSherry[[1]](#endnote-1), and largely in use by major technology companies like Apple and Google.

# Why? How? What?

Data aggregation can leak information from the original data, exposing the entities constituting the original dataset. This has been observed numerous time, and even though the complexity of these data models hasn’t ceased to increase, the method to avoid them hasn’t evolved much: peoples are in charge of picking and choosing which information is released to the public.

Differential Privacy is not a silver bullet, it doesn’t fix all privacy issues and doesn’t protect everyone from data leakage. It

1. [US 7698250](https://worldwide.espacenet.com/textdoc?DB=EPODOC&IDX=US7698250), Cynthia Dwork & Frank McSherry, "Differential data privacy", published 2010-04-13 [↑](#endnote-ref-1)