

Daniel Willim: 8 h

Erik Jergéus: 8 h

Part 1: Privacy and history and limitations of AI

1. Statistical disclosure limitation (SDL) is a collection of methods designed to protect the privacy of individuals whose data has been used in analyses that will be released to the public. Which of the following specifically compromise the usefulness of SDL to protect the privacy of individuals

- ☒ Increasing computational power
- ☐ A critical bug in broadly used SDL softwares
- ☐ Fast internet connectivity to mobile devices including LTE and 5G.
- ☒ Availability of personal data on the internet and other places
- ☐ Increased public concern about privacy following revelations by Edward Snowden

2. Which of the following are ethical arguments *for* protecting subjects' private data in commercial and research applications?

- ☐ If subjects do not want their data to be disclosed publicly they should not share it.
- ☒ Protecting the privacy of subjects is a matter of safeguarding their dignity and welfare.
- ☐ Private data should only be protected insofar that it does not interfere with a company's ability to maximize economic growth, because human well-being is associated with increasing economic prosperity.
- ☐ Protecting private data is only important when it concerns highly private information such as sexual orientation, religion, or political beliefs.
- ☐ A failure to protect the privacy of subjects could render them vulnerable to identity theft and extortion.

3. The following analysis techniques or data acquisition techniques are *not* differentially private

- ☐ In a movie rating data-set, removing the names of subjects and randomizing some of their ratings.
- ☒ In personal health records including hospital admission, removing names, blood type, annual income, information about HIV status, and social security numbers of subjects.
- ☒ In a financial transaction data-set, removing the name and address of subjects excluding postal code, and randomly exchanging labels of credit card transactions within subjects transactions, such that individual payment amounts cannot be tracked to payment of particular services or goods.
- ☐ In an online survey measuring adolescents and young adults recreational use of illicit drugs using a yes or no question and registering their age. The subjects answers to the yes or no question are stored with a probability of 50%. If a subjects answer is not stored it is sampled randomly with a 50-50 chance of yes or no. Similarly, the age is either registered as their correct age with a probability of 50%, else a fake age is sampled from a uniform distribution of integers from 14 to 29 years.

2/ 6

- ☒ A series of street-level cameras in down-town Gothenburg are set up to disincentivize crime and simultaneously to collect data to reduce crowding and improve citizens shopping experiences. The cameras use a facial recognition algorithm to track pedestrians faces and count how often subjects pass by a given camera. No images are stored, each camera generates a coded representation of each face using the same algorithm on each camera. The coded representation of a face, along with the GPS coordinates of the camera, date, and time are stored.

4. Fisher made important contributions to statistical significance testing. He saw it as a way of communicating 'statistical findings that was as unassailable as a logical proof'. Statistical significance testing can *not*:


- ☐ Tell us whether a sample is likely to have occurred under the null-hypothesis.
- ☒ Tell us whether what we are measuring is important with regards to the conclusion we wish to draw.
- ☐ Tell us whether a sample is unlikely, up to some probability threshold, to have occurred under the null-hypothesis.
- ☒ Infer the underlying cause of the measured data.
- ☒ Objectively relate humans ethnicity's relationship to their personality traits.

5. What were important early drivers of the statistical research of Galton, Fisher, and Pearson?
- ☐ The study of coin-flips
 - ☒ Eugenics
 - ☐ Association of genetic variants observed in large human cohorts to disease phenotypes from next-generation sequencing technologies.
 - ☒ Biological evolution.
 - ☐ Fault detection of machines in factories powering the industrial revolution.
6. In a study from 1925 published in the scientific journal which is now known as “Annals of Human Genetics,” Pearson claimed to have shown that Jewish children, in particular girls, were less intelligent on average than their non-Jewish counterparts. He further claimed that intelligence was not significantly correlated with any environmental factor that could be improved, such as health, cleanliness, or nutrition. What aspects of the study may influence these conclusions?
- ☐ Pearson did not use the right statistical significance test.
 - ☒ Pearson did not use a certified intelligence test.
 - ☒ Pearson made use of self-reported data (survey data) of the conditions of students home lives.
 - ☒ Pearson made use of teachers assessment of students intelligence.
 - ☒ Pearson did not account for environmental factors that could not be improved, e.g. the students’ refugee status.
7. In the Fragile Families Challenge researchers from 160 teams were tasked to predict a number of social outcomes — child grade point average, child grit, household eviction, household material hardship, primary caregiver layoff, and primary caregiver participation in job training — using extensive longitudinal data. Is this a difficult prediction task? Why? Why not?

3/ 6

-
- ☒ Yes, because the challenge includes predicting the future (“wave 6”), and only selected data is available for the time-point (“wave 6”) to be predicted.
 - ☐ Yes, because the challenge includes predicting the future (“wave 6”), and no data is available for the time-point (“wave 6”) to be predicted.
 - ☒ Yes, because social outcomes depends on many parameters. Many of which may not be captured well by the dataset.
 - ☐ No, because the researchers can use any machine learning method and they have access to a big data set.
 - ☐ No, because social outcomes are known to be easy to predict even with simple models.


8. The authors of the Fragile Families Challenge study argue that '*predictability*' is poor measure for understanding the mechanisms of social outcomes. What alternatives do the authors propose?

 Descriptions.

☐ Using statistical significance testing.


☐ Using Foucault's analysis of power.


☒ The current understanding is correct but incomplete because it lacks theories that explain why outcomes are difficult to predict even with high-quality data.

 Causal inference.

9. What important *ethical* implications may a poor '*predictability*' of social outcomes have?

☐ It will be more difficult for employers to use AI to screen job applicants, leading to an increased work load on Human Resource employees.


 Using predictive modeling in the criminal legal system may lead to wrongful convictions and arrests.


 It will put legal strain on the police and military because the most performant models for identifying criminals and terrorists use deep learning that are difficult to interpret.


 Using predictive modeling in the child-protective services may lead to wrongful removal of children from their families.

☐ It will require larger machine learning models to improve predictions, which in turn require more compute power and electricity to run, putting strain on the environment.

10. What reasons can lead to unfair and unethical AI and Data Science models and algorithms?

 Sampling bias.

 Systemic bias in society.

 Poorly designed surveys.

☐ Low memory capacity of embedded devices.

☐ Overly rigorous testing and use of the common task method.

Case 1: Wearable health-technologies and private health insurance

1.1 How can the training pipeline affect the predictions of Vivaksa's health status prediction algorithm when deployed on the general public?

At the moment they have only trained their model on data from their employees who are not in any way a representative sample of the world's population. Even if there are no statistical differences in accuracy of heart rate monitoring due to skin color [1], many other factors are affecting a person's health that may differ between people of different ages, genders and/or cultures. Hence the model might be accurate on the 3 employees and people similar to them, but the model would not be accurate for the whole population. Another problem is that they only trained on one month of data, but health and stress changes are often slow and might depend on many other factors not tracked by Vivaksa hence it would be easy to make false correlations between their data and the health and stress of a person.

1.2 Can Vivaksa improve their predictions? if so how? and if not, why

In short, by fixing many of the problems discussed in q1.1 they would improve their algorithms predictions ability and mitigate some of the inherent bias of the algorithm. For example, if they instead trained their model on a more representative sample of the world's population and also tracked some data like biological gender, age, country of origin etc. and increased the timespan to a couple of years and not a couple of months the model might give a better prediction.

The study [1] also showed that the accuracy of wearable technologies may differ a lot depending on the model, hence by using data from different wearables the noise in the data will resemble the real-world data more and therefore improve the predictions.

If the only input to the model would be data from the wearable technologies then by fixing the problems discussed here the predictions might be better and less biased but they will never be super accurate since a person's health depends on many more factors not tracked by a watch hence a prediction will never be perfect.

We know that some of the proposed changes are not technically or economically viable but here we discuss it theoretically as if the world was a magic land where everything was possible. 🙄

[1]: Investigating sources of inaccuracy in wearable optical heart rate sensors, Brinnae Bent, Benjamin A. Goldstein, Warren A. Kibbe and Jessilyn P. Dunn

2.1 Can the roll-out of this pilot program undermine the insurance company's own value of fairness? – explain why or why not.

Vivaksa's algorithm isn't fair, according to discussions in q1.1 and q1.2, hence any decisions made from information by this algorithm would be biased in the same way and are therefore dubious at best.

But even if we assume that Vivaksas algorithm is equally accurate on all people, there are still several problems related to fairness. Firstly, to get this discount you need to have some sort of smartwatch/fitness tracker, so people who do not want to have a smartwatch or can't afford one are now paying more for the same insurance, even if they are in perfect health. Another problem is that you also open up for tech-savvy persons to mod their wearable technologies to report better health status, while doing this would be illegal it is a rather easy fraud scheme that has a low chance to be detected giving people with questionable morals and a bit of tech knowledge an advantage.

But there is an argument to be made that this change would be fairer towards healthy people since they have a lower risk of falling ill and therefore requiring their health insurance and unhealthy people could just go to another company that does not do this profiling. (assuming that the rest of the health insurance business doesn't apply similar methods)

Case 2: AI-enabled human personality prediction

How would we have to adjust our notions of inalienable individual rights once we had the ability to identify people as criminals before they ever acted?

1. (1pt) What are the limitations of Selfie2Personality's technology - if any?

The most glaring problem with this technology would be that the algorithm could at best be as good as the reference material. For example, the technology is based upon the research by Xiaolin Wu and Xi Zhang which has multiple problems in this department: they don't get the data sets for labelling criminals vs non-criminals from the same objective source (biased data sets), they don't have the metadata to distinguish between other factors such as socioeconomic status (potential unknown factors) and they take court conviction at its face value (labelling issues). The other main reference material that they use has the important disclaimer "*The scores computed by the algorithm approximate real participants' subjective evaluations of perceived trustworthiness on the processed images based on the pose taken by the individual. It provides no meaningful information whatsoever about individuals' actual trustworthiness*"[1]. This tells us that the trustworthiness score just mediates the cultural and racial biases of the participants and is therefore at best useful for finding biases in the participants, not for finding traits radiating trustworthiness in the image set.

[1]: <https://osf.io/j68xu/>

2. Following **protests** in a city that caused significant damages to public and private property law-enforcement is looking for suspects. Analysing social media posts prior to the protests, the law-enforcement identify a pattern of certain **political leanings**, however without being able to pin-point specific suspects as to who caused the damages. To narrow their possibilities, law-enforcement reach out to Selfie2Personality who agree to cooperate, as they see it as their civic duty to help bring justice in society. (2pt) Selfie2Personality **share the social media accounts and a psychological profile** of users from the city metropolitan area who fit the political stance (as predicted from their app) that law-enforcement found to be sympathetic to the protests. Are there any **ethical and privacy concerns** with this **collaboration** and **how may they manifest?**

First of all, the app must state in its terms of service how the data profiling will be used, and if the subjects live in the EU (other locations might have similar laws) they must follow the

laws of GDPR, which among other things prohibits data profiling on children and gives all users protected by the law the right to be forgotten, informed and halt the profiling (“*unless the controller demonstrates that the objection overrides the interests, rights and freedoms of the data subject*”) [1]. The sharing of social media accounts together with the psychological profile grants the law-enforcement access to information which does not expire and can be exploited for malicious purposes outside of the protests (for example targeting with political advertisements or in a more extreme scenario can be killed or harmed by their political opposition). Selfie2Personality did not just share the names of people suspected of political leanings but released their entire psychological profile, which can be used for both targeting and profiling. This knowledge can manifest in good things, such as shorter queues on air-planes by classifying people by risk of terrorist activity, but it is still discrimination and should be carefully considered before applied [2]. All of these conclusions were made under the assumption that the data given was a good unbiased psychological profile of the users, which is unlikely considering their research material and what was discussed in part 1. Making assumptions based on potentially biased or inaccurate data in this scenario can potentially lead to groups of people being interrogated without proof nor basis. An example of when this could go wrong is if many people from an ethnic group are deemed potentially violent by the biased psychological profile and as such are incorrectly identified as likely suspects. This could increase the ethnic group’s disdain for the law-enforcement and would validate some peoples’ stereotypes, potentially leading to for example wider social inequalities.

[1] *GDPR and profiling*, Majlet
<https://www.mailjet.com/gdpr/profiling/>

[2] *E-Government, Targeting and Data Profiling*, Dr. P Henman, doi:
10.1300/J399v02n01_05
https://www.tandfonline.com/doi/abs/10.1300/J399v02n01_05?casa_token=eUw1GkqGd9YAAAAA%3Ag5fL_rbvocsVzXliy_XBRrbXeJaAVf7Mczlz31ahEtKw1a1kKqatJad1XjQ7qOgdg-4VQIrlNsiw&