Reproducibility, Privacy in Practice

UNIVERSITETET
I OSLO

A (slight) change of plans:

- Lecture 8 - Reproducibility and Privacy in Practice
- Lecture 9 - Evaluating Models, Scarcity and Abundance of Data

Reproducibility

What could make data science non-reproducible?

UNIVERSITETET
I OSLO

Key Challenges

- Changing the data
- Changing the model
- Changing the training

Changing the data

- As discussed last time, virtually all data is dynamic
  - Surveys, user preferences, time series, …
- Underlines the importance of data engineering
- Many ways used to keep track of data
  - Write-only databases
  - Check-pointing
  - Finger-printing

Important: Need to keep track what data your model was trained on (as MLFlow does!)

Changing the model

- Lots of models use randomness
  - Random Forests
  - SDG
  - K-Means
  - Neural Networks (initialization, dropout, ++)

Need to set (and record) a random seed for reproducibility

Changing the training

- Many procedures use randomness
  - Train/test split
  - Cross-validation (minus leave-one-out)
  - Bootstrapping

Again, need to set (and record) a random seed for reproducibility

Privacy in Practice

UNIVERSITETET
I OSLO

The Industry Perspective

- With GDPR, few companies will touch ‘anonymizing’ data
  - Potential penalties are too high
  - Either data has been collected for given use, or it hasn’t
  - More on that in the GDPR lecture
- However, for some e.g. governmental tasks this might be of interest

Randomization

- We'll cover only randomization
  - Laplace and similar methods are less interesting computationally
- We'll see the trade-off between anonymity and model performance

UNIVERSITETET
I OSLO