# Session 2

## Exploratory data analysis, training models

UNIVERSITETET
I OSLO

# The plan

- Business context of data
- Feature exploration
  - Categoricals
    - Support
  - Scaling
- Outliers, missing data
- Correlations
- Fitting a model in scikit-learn
- Evalutating a model

# Business context of data

- Always need subject matter expertise, not obvious from data alone
  - Subject matter experts (SMEs) usually part of a project team
- In the lectures, we need to ‘invent’ the business context

**It is essential to get this right.**

**If not, you will solve the wrong problem!**

# Exploratory data analysis

- This makes sense only with business context in mind
- Fishing for signals works in very few cases
  - Beware of ‘we give you data, you see what you can find’
  - You don’t know what you’re looking for!

# Categoricals

- Python DS framework (pandas, sklearn) doesn't handle them well
  - Pandas has a `Categorical`
  - Generally needs one-hot encoding (drop first not standard)
- Beware of categories with small support
  - Can give extra insight, but can cause trouble

# Outliers

- Generic strategy: Calculate z-score, label high-z data as outlier
  - Won’t work for multi-modal distributions
- Always ask why there are outliers
  - Measurement error?
  - Faulty sensor?
  - Survivor bias?
  - Data transfer error?