Machine Learning for Mortality Risk Prediction: A Study on Cohort Data

MEB Aging Epidemiology Group
Jim Jakobsson
Supervisors: Sara Hägg , Patrik Edén, & Anders Rantzer

Introduction and Background

Mortality risks increase with age, the complex interactions of other factors relative to age are less clear

Analysing comprehensive datasets could clarify those factors

Traditional statistical approaches, and humans, are limited in their ability to process high-dimensional data and detect non-linear patterns

- **Machine learning offers powerful advantages:**
  - Analyse all records simultaneously
  - Identify key predictors of mortality
  - Data-driven insights for personalised interventions

Aim of Study

The aim of the study is to apply machine learning to analyse SATSA, to identify patterns and predictors associated with mortality in aging adults

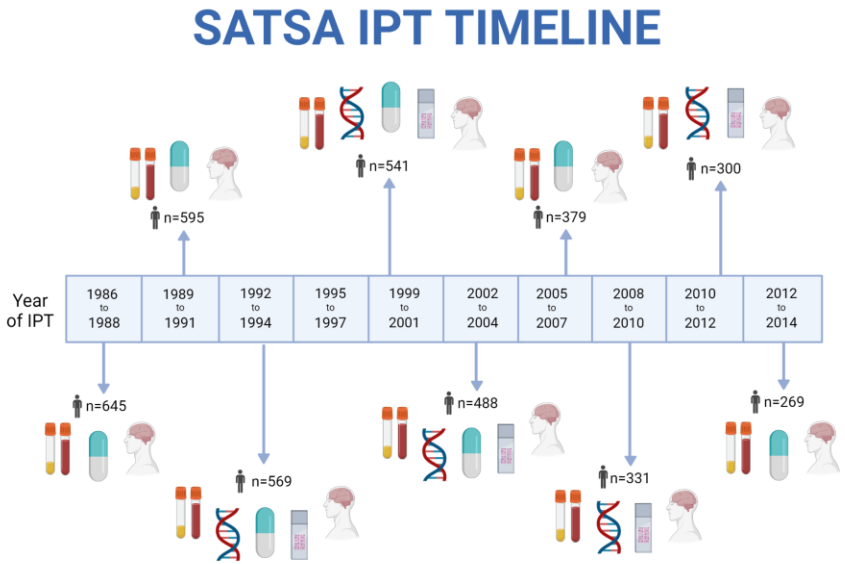
Research questions

1 What are the strongest predictors for a 10 and 20 year mortality prediction among aging adults?

2 Can tree-based supervised machine learning models accurately predict mortality risk using cohort data, without using synthetic imputations?

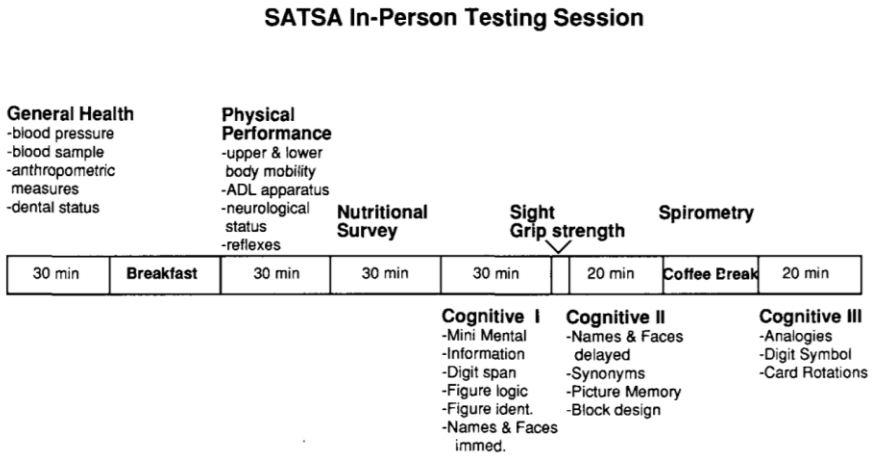
SATSA: The Swedish Adoption/Twin Study of Aging

- Longitudinal study of twins started in 1978
- Unique focus on twins reared apart and together
- Comprehensive health and cognitive assessments over 10 measurements, 30+ years
- Data collected through:
  - → Questionnaires
  - → In-person tests
- **Strength**: multiple domains included



Description of In-Person Testing

- Eligibility: Pairs where both responded to the first questionnaire, age 50+
- First wave used: n=654, 423 variables
- Comprehensive assessments:
  - General health
  - Physical performance
  - Cognitive testing
- Mean age: 63.9



Examples of Variables Collected in the Study

Cognitive Performance		Cardiovascular measurements		Respiratory function		Functional Abilities	
Variable name	Variable description	Variable name	Variable description	Variable name	Variable description	Variable name	Variable description
asymb bsymb	Perceptual speed - identify corresponding digits for a series of symbols	bpsres	Blood pressure	fv1	Forced expiratory volume, 1 second	strength	Grip strength
face1	Correctly identify previously viewed faces	tg1	Triglycerides	fvc	Forced vital capacity	activities of daily living	Screw in a lightbulb, rotary telephone, put coins into slots
arottot	Spatial ability – card rotations. Determine if card is rotated compared to reference	sedim	Erythrocyte sedimentation rate	kvotvc	Ratio fv1/fvc		

Main Challenges with SATSA Dataset

SATSA offers a valuable opportunity to discover contributions to aging, however:

- It is a dataset containing vast amounts of data
- Complex interactions between 400+ variables
- Several missing values
- Identification and handling of continuous/categorical variables
- Traditional statistics require pre-defined assumptions
- Scale too large to handle manually

Advantages of Supervised Machine Learning

Algorithms that can **make predictions or decisions** without being **explicitly programmed** for every scenario

Automatically discovers patterns in complex data, comparing attributes of different classes: survived/deceased

Can handle **high-dimensiona**l datasets

Identify **subtle interactions** between variables

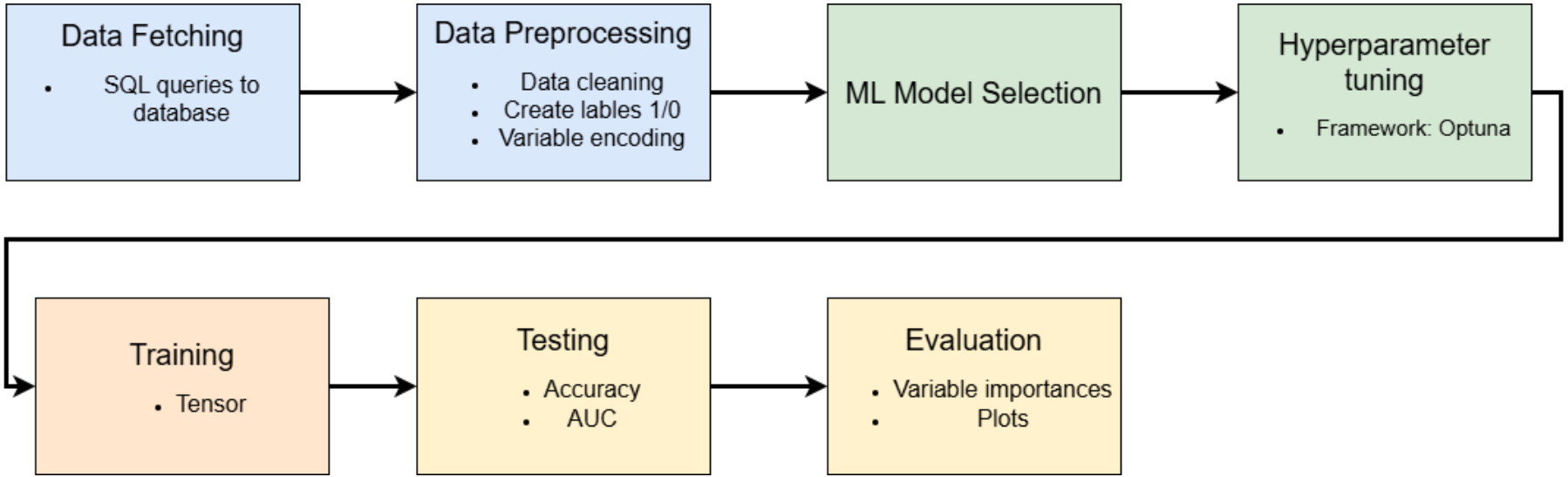
Prioritises the **most important factors**

Learns which **combinations of factors** are linked to longer or shorter lifespan

Methodology

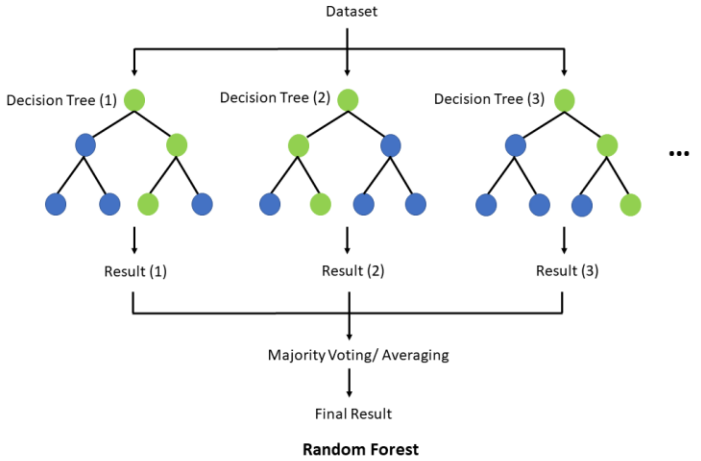
	Step	Description
1	**Created a labeled dataset from SATSA**	• Tracked survival status at 10/20 years after IPT1 • Labeled each participant: survived(0)/deceased(1)
2	**Transformed missing data into insights**	• Missing tests may signal health status • Binary indicator variables introduced (missing/not missing)
3	**No synthetical imputation of missing values**	• Common practice to impute, sensitivity analysis needed • All results are grounded in actual observations
4	**Standardised variables**	• Converted continous values to range between 0 and 1 • Categorical variables encoded with one-hot encoding
5	**Applied machine learning models**	• Used tree-based models that can handle missing data • Combined multiple models for better predictions
6	**Trained the models on the data using Tensor**	• Computationally expensive, cannot be done on a laptop • Using 96 cores, Tensor required 10-45 minutes
7	**Interpreted results**	• Performance metrics used: AUC, accuracy, F1 score • Quantified importance of each variable using the framework SHAP

Flowchart of Data



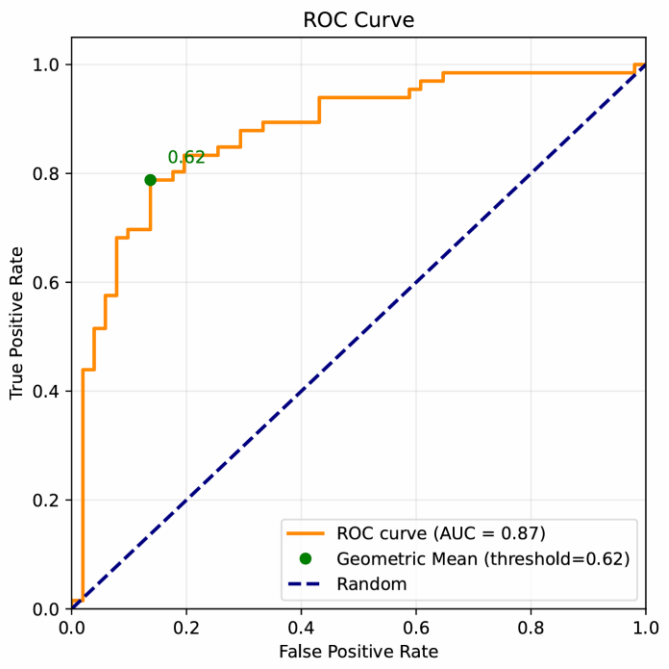
Random Forest Classifier

- **Combines many decision trees**
  - → Each tree learns from different health measurements
  - → Trees vote on mortality probability
  - → Final prediction based on fraction of positive outcomes
- **Great performance**
  - → Handles missing data natively
  - → Can capture complex, non-linear relations
  - → Robust to outliers
  - → Robust to overfitting



Model Performance

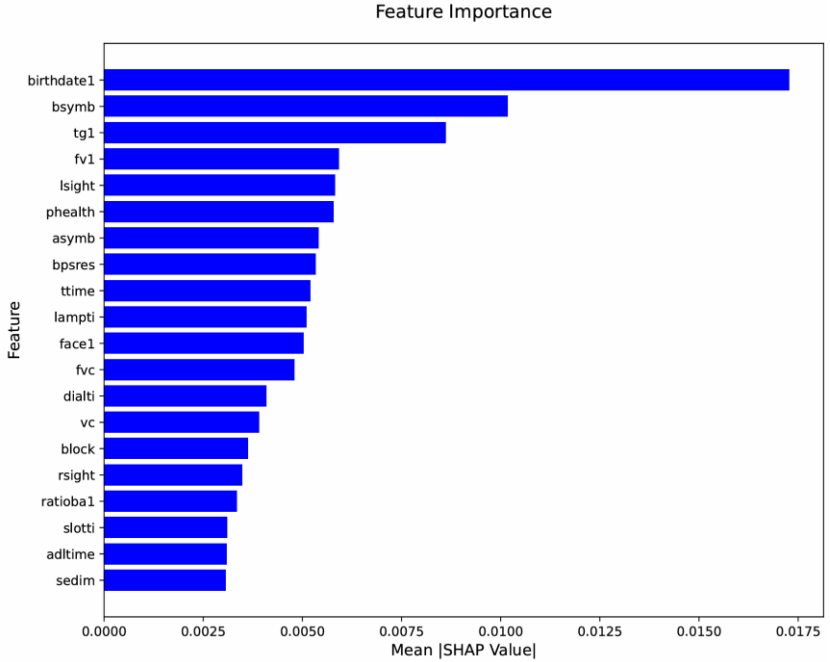
- **20-year prediction**
  - → Balanced dataset
    - 56% deceased, 44% survived
  - → Accuracy: 0.81, AUC: 0.87
- **10-year prediction**
  - → Unbalanced dataset
    - 27% deceased, 73% survived
  - → Accuracy: 0.73, AUC: 0,79
- **Geometric mean and class weights used to balance data**
  - → No synthetic data used



ROC and AUC for a 20-year prediction, using a Random Forest

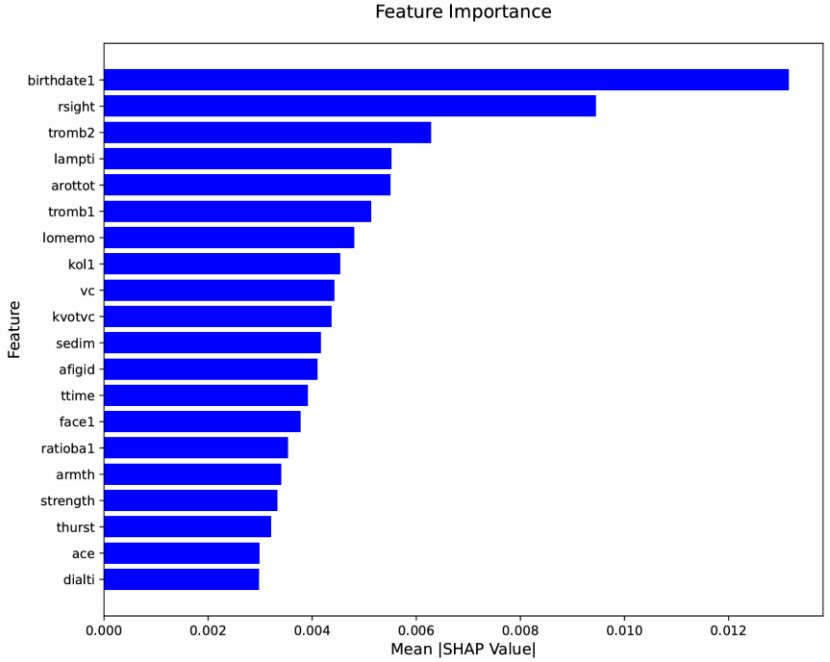
Key Variables 20-year Mortality Prediction

- **Cognitive performance**
  - → Perceptual speed, **asymb** & **bsymb**
- **Triglyceride levels**
  - → **Tg1**
  - → From excess calories
  - → Elevates risk for diabetes, strokes
- **Respiratory functionality**
  - → Forced expiratory volume, **fv1**
  - → Forced vital capacity, **fvc**
  - → Vital capacity, **vc**
- **Vision assessments**
  - → **lsight** & **rsight**
- **Blood pressure**
  - → Systolic, **bpsres**



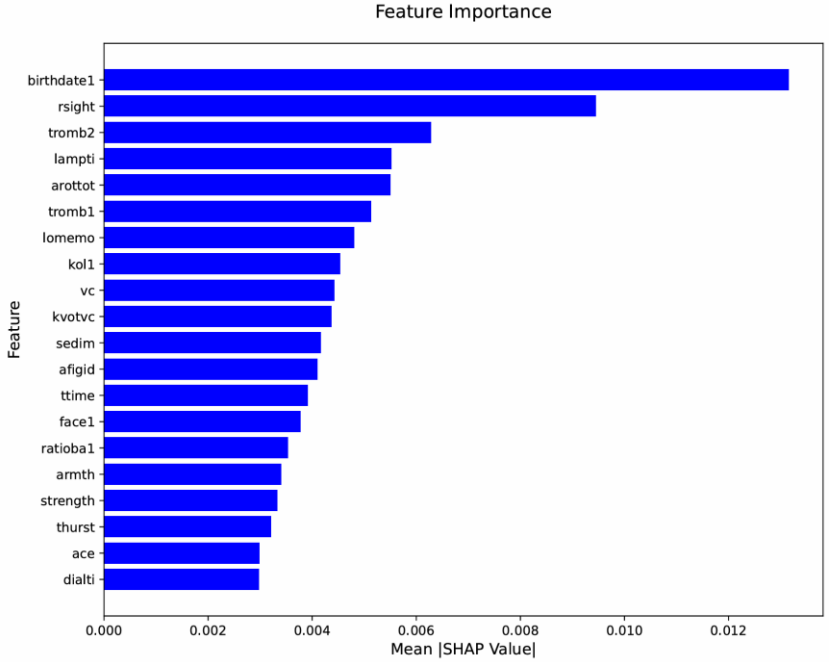
Key Variables 10-year Mortality Prediction

- **Focused on cardiovascular measurements**
  - → Thrombocyte numbers, **tromb1** & **tromb2**
  - → Cholesterol levels, **kol1**
  - → Erythrocyte (red blood cells) sedimentation rate, **sedim**
  - → Apolipoprotein B/A, **ratioba1**
- **Functional assessments**
  - → Fine motor skills, **lampti**
  - → Grip strength, **strength**
- **Long-term memory difficulties**
  - → **lomemo**



Common Key Predictors 10/20 years

- **Age**
  - → birthdate1
- **Ratio apolipoprotein B/A**
  - → Ratioba1
  - → Carries cholesterol in the blood
- **Vision assessments**
  - → lsight & rsight
- **Face recognition**
  - → face1
- **Vital capacity**
  - → vc



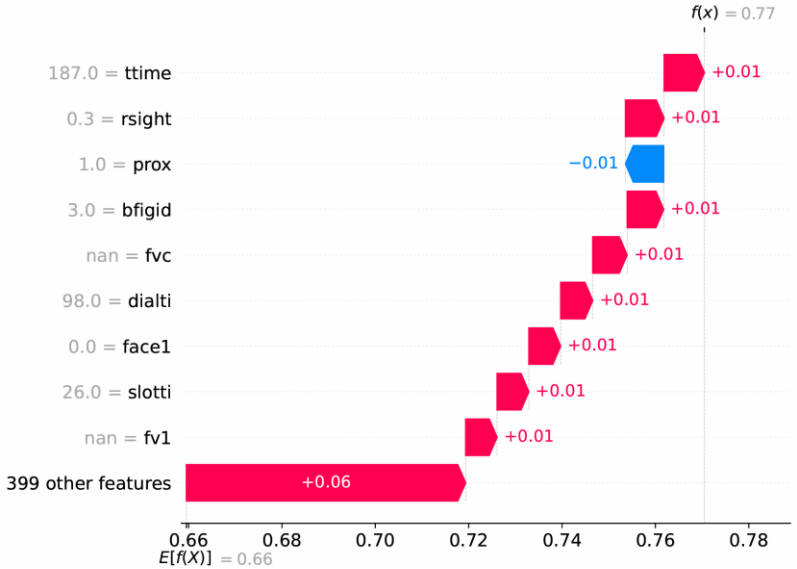
Key variables 10-year mortality prediction

Individualised Risk Assessment

Variable contributions for one person

- **Age not a significant factor**
- **Slow at the functional tests**
  - → Ttime, dialti, slotti
- **Poor vision**
  - → rsight
- **Good proximal muscle function**
  - → Prox
  - → Lowers mortality risk
- **Missing values for lung measurements**
  - → Fvc, fv1
- **Low scores in cognitive test**
  - → Bfigid: figure identification
  - → Face1: face recognition

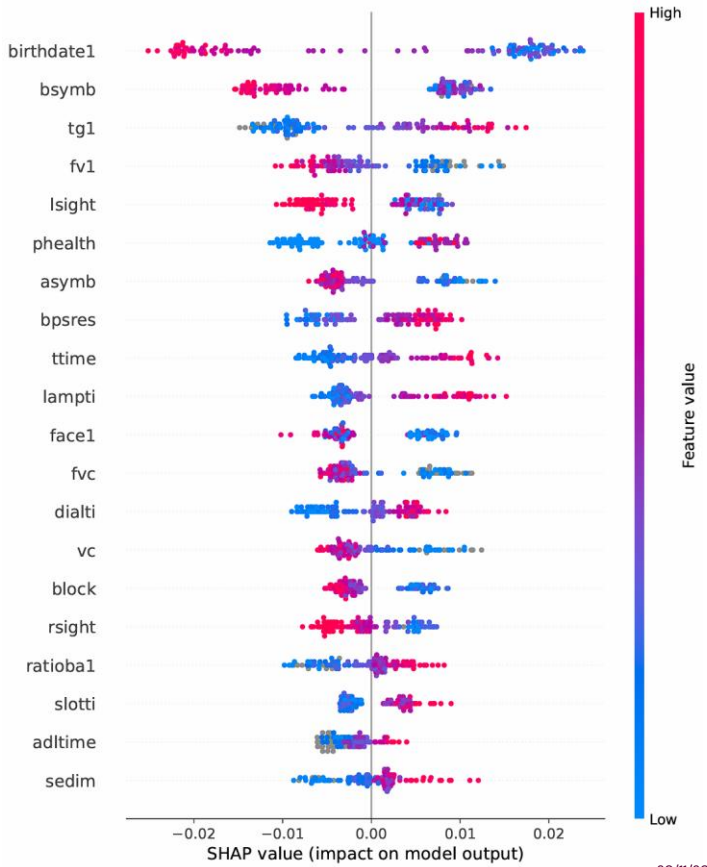
Individual 10-year Mortality Risk Assessment: 77%



Value Distribution

- Shows a measurement's impact on mortality risk
  - → Red = high values, blue = low values
- Position shows if factor increases (right) or decreases (left) risk
- High sedimentation rate (sedim) as significant as blood pressure
- Low vital capacity (vc) significant
- Low scores in asymb & bsymb significant
- High time for lamp insertion (lampti) significant

Summary plot for a 20-year mortality prediction



Conclusions

Significant for a 20-year mortality prediction*

- Cognitive performance:
  - Perceptual speed
  - Spatial ability
- Metabolic measurement:
  - Triglyceride levels
- Blood pressure

** Symbolises degenerative factors*

Significant for a 10-year mortality prediction

- Cardiovascular measurements:
  - Thrombocyte counts
  - Cholesterol levels
  - Erythrocyte sedimentation rate
- Functional assessments:
  - Grip strength
  - Fine motor skills
- Long-term memory difficulties

Significant across both prediction windows

- Age
- Vision
- Apolipoprotein
- Facial recognition
- Respiratory function

Limitations to Study

- **Single time-point measurement**
  - → Missing health trajectories over time
  - → Future work: include all measurements from all IPTs
- **No imputation or use of synthetic data**
  - → Limited to tree-based ML models
  - → Unable to balance data sets using synthetic data
  - → Future work: imputation with sensitivity analysis
- **Little feature engineering**
  - → Almost every variable is in the training data
  - → Future work: reduce to domain specific features to improve accuracy


KAROLINSKA INSTITUTET
ANNO 1810
Karolinska
Institutet