



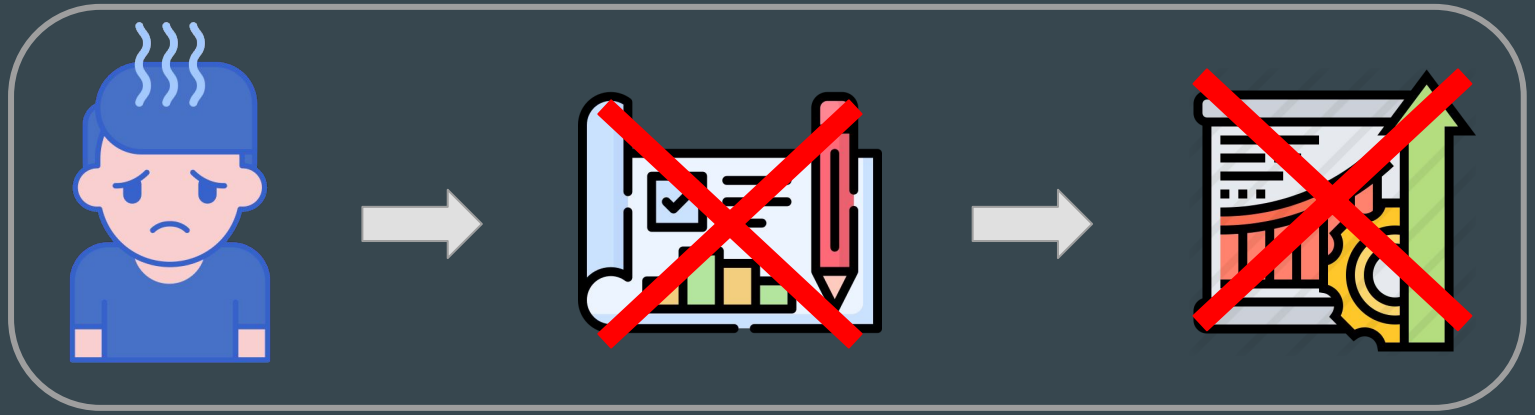
# Defining Corporate Health Classes

...

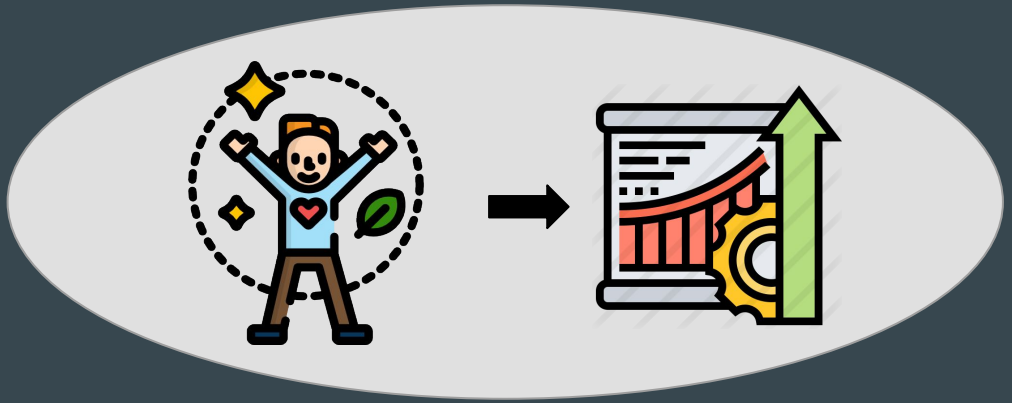
Data Innovation Lab
29th of July 2020

# Content

# Project Overview



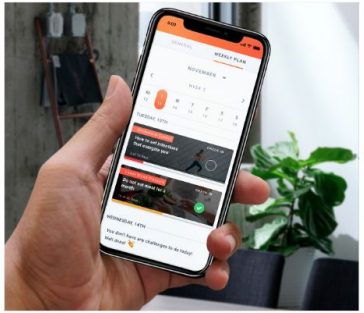
wellabe



Mobile diagnostics

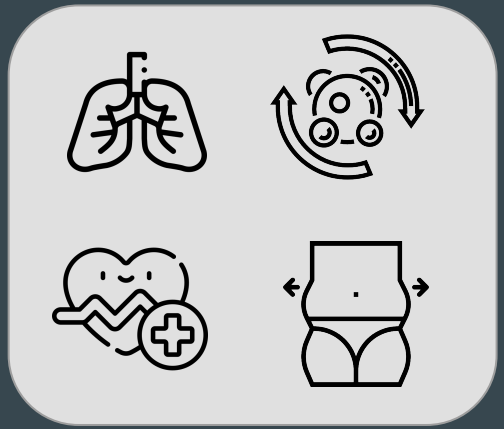
Video consultation

Personalized intervention



Annual retention


# The Dataset :



+

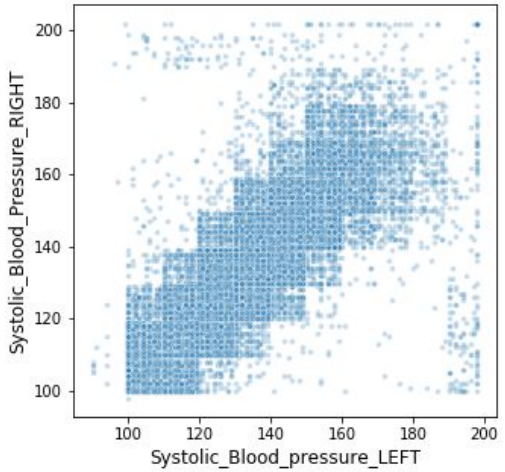

Step 1
Data Understanding & Handling
Step 2
Find Indicators for General Health
Step 3
Define Health Classes


Step 1
Data Understanding & Handling
Step 2
Find Indicators for General Health
Step 3
Define Health Classes

# Synthetic Data

- Data Privacy
- Generated from original data
- Replica of specific properties of real data

Synthetic Patterns



# Statistical & Medical Analysis

50 000 people
- 57.5 % male
- 42.5 % female

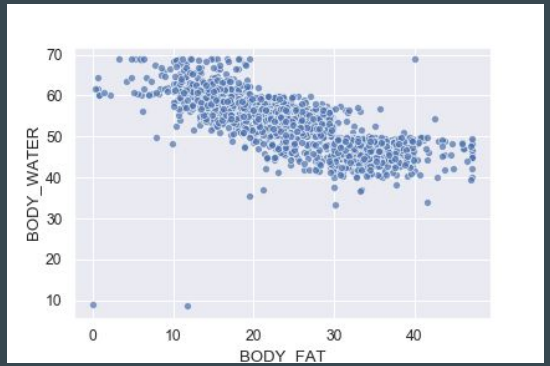
Age: 18-65 years

- Medical understanding of the given features
  - blood sugar -> concentration of glucose present in the blood

- Possible ranges of features
- Known diseases

| Normal | Prediabetes | T2 diabetes | Hypoglycemia |
|---|---|---|---|
| 70 - 99 | 100 - 125 | >126 | <70 |

- Summary of the data
  -> Mean, Standard deviation
- Distributions of features
  -> Histogram, Normal probability plot
- Correlation between features
  -> Scatter plot, Pearson Correlation Coefficient



# Data Cleaning

**Outlier Detection**

- Define thresholds
- Relationship between features
- Replace with NaNs

**Estimation of Missing Values**

- Estimate NaNs
- Median
- Formulas
- Linear Regression


Step 1
Data Understanding & Handling
Step 2
Find Indicators for General Health
Step 3
Define Health Classes

# Indicators for General Health

## Review Score

indicator of general health given by doctors

## Biological Age

- predicted chronological age
- difference between predicted and real age indicates general health

# Regression Methods

Multiple Linear Regression (MLR):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

- manual feature transformation
- good interpretability

Generalized Additive Model (GAM):

$$g(\mathbb{E}[y|X])) = \beta_0 + f_1(X_1) + f_2(X_2) + \ldots + f_M(X_N)$$

- automatic feature transformation via B-splines *f*
- still good interpretability due to additive structure

Multivariate Adaptive Regression Splines (MARS):

$$\tilde{f}(X) = \beta_0 + \sum_{k=1}^{K} \beta_k h_k(X)$$

- non-linear transformations via hinge functions *h* of the form max(0, *x* - *t*) or max(0, *t* - *x*) with *t* as threshold
- good interpretability

# General Setup

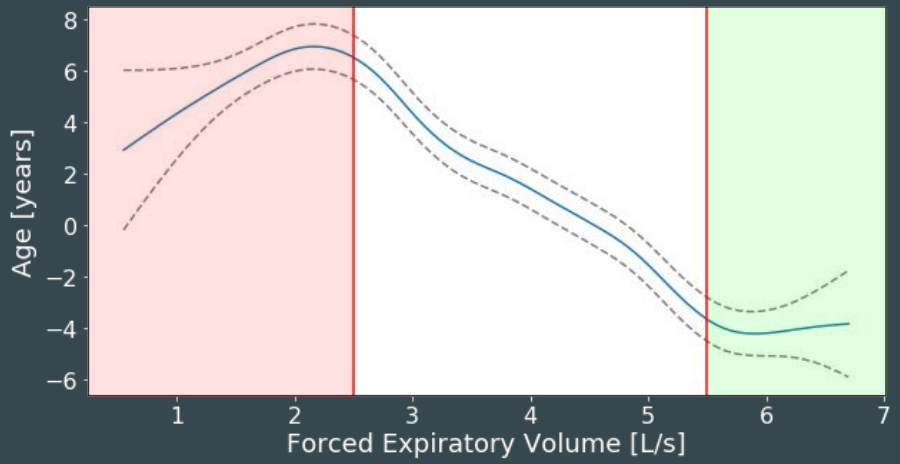
Data:
- Cleaned dataset
- Dropped rows with a NaN value within one of the used features

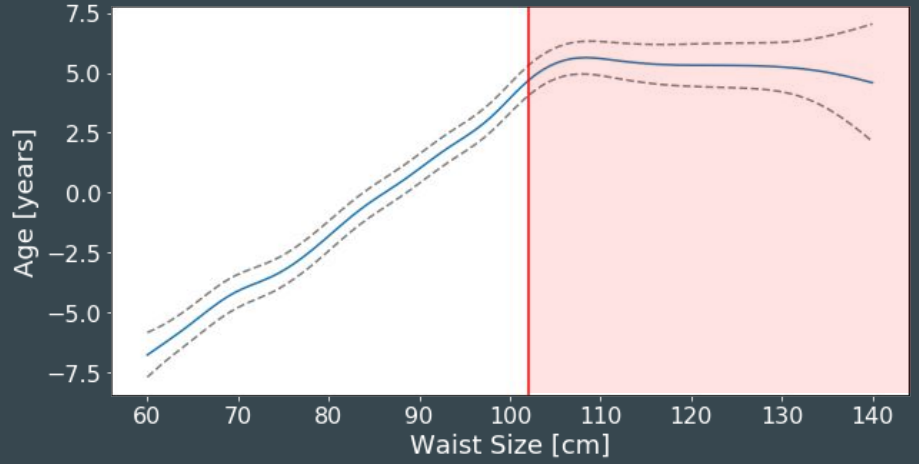
Comparison of Methods:
- Mean Average Error (MAE) for accuracy
- Check relations to other features  (indicator of general health)

# Results Age Prediction

| Method | Baseline | MLR | GAM |
|---|---|---|---|
| MAE | 9.38 | 7.84 | 7.21 |

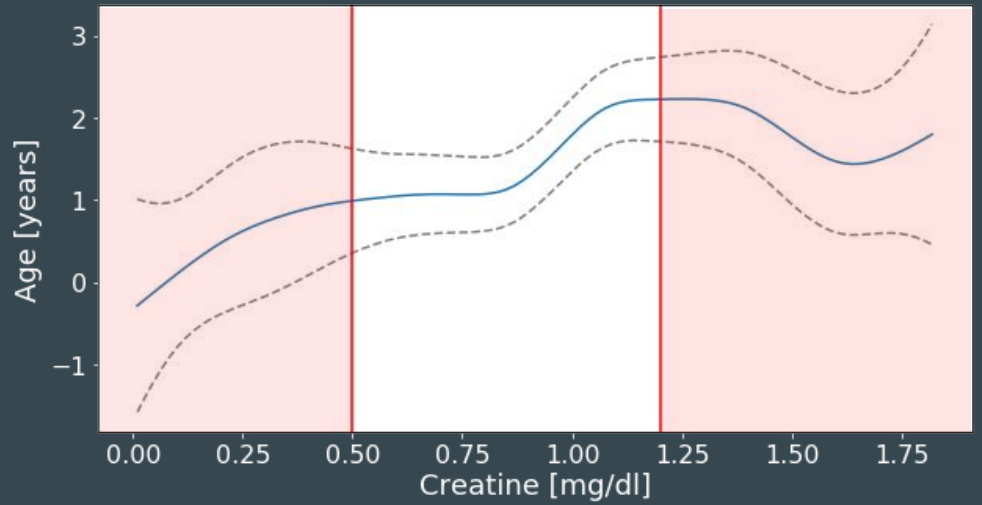
GAM Example Feature Functions:



**Feature functions not always linear in normal range**

Feature function of creatine



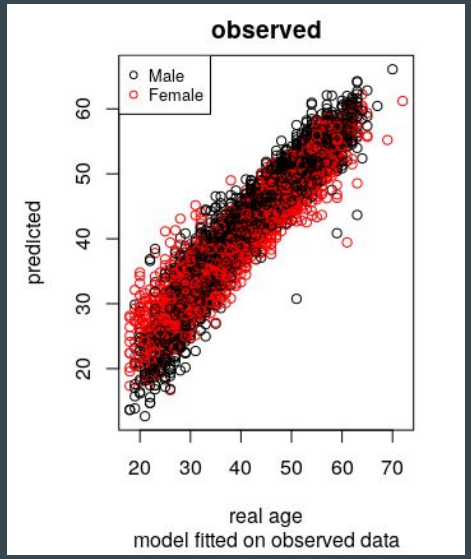

>MLR gives better health indication for some of the features

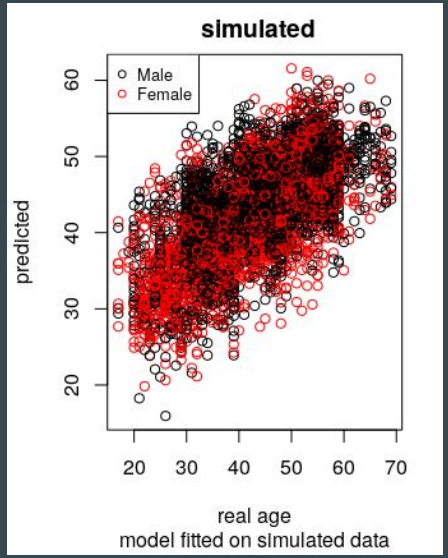
# Age Difference

Age Difference = Biological Age - Chronological Age

Result:
No clear correlation between worse medical values and Age Difference
-> not a meaningful indicator of general health

# Comparison to Original Dataset



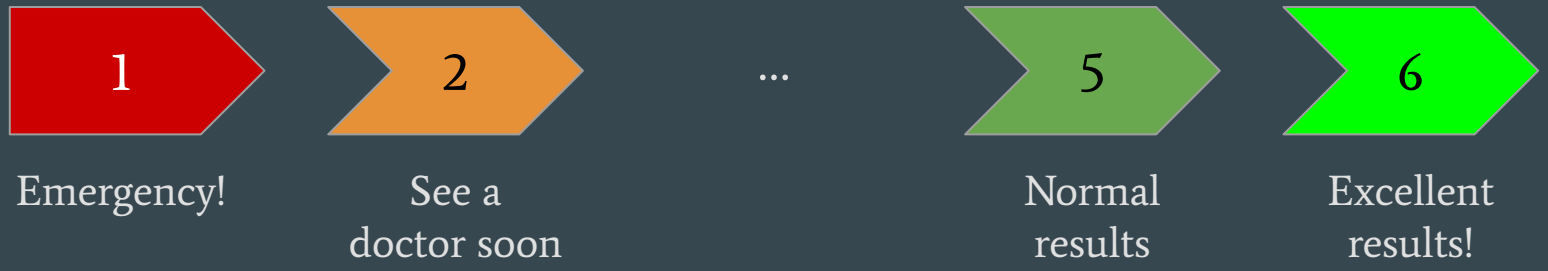




MAE: 2.30 6.82

> Loss of information in synthetic data

# Review Score

The score given by a doctor




Motivation:

- finding features relevant for general health prediction
- additionally, giving a review score automatically

# Methods & Results

Results: MAE

| baseline | GAM | MARS |
|---|---|---|
| 1.21 | 0.85 | 0.82 |

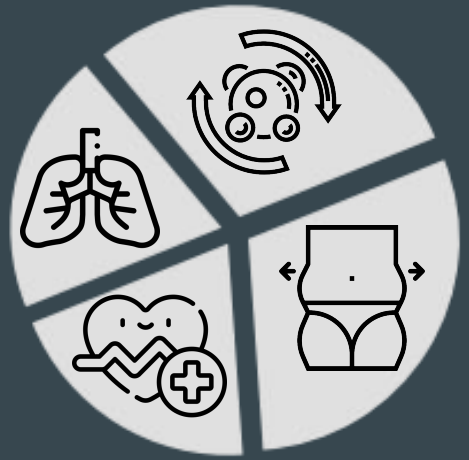
Selected by MARS features:

- fatty liver index (FLI) → fatty liver
- blood pressure → hypertension
- body fat → overweight, obesity
- visceral fat → overweight, obesity
- cholesterol → high blood lipids


Step 1
Data Understanding & Handling
Step 2
Find Indicators for General Health
Step 3
Define Health Classes

# Methodology: Preprocessing

- Feature selection
  - started from the list of features relevant for general health
  - ensured to construct a versatile set
- Estimation of missing values/outliers
- Standardization



# Methodology: Clustering

- K-Means
  - roughly same size of clusters
  - no medical interpretation
- Density Clustering (DBSCAN)
  - could not detect more than one cluster
  - possible reason: health is continuous => no distinct clusters
- **Hierarchical Clustering**
  - hierarchical structure of clusters
  - resulting clusters were most interpretable

# Female Clusters

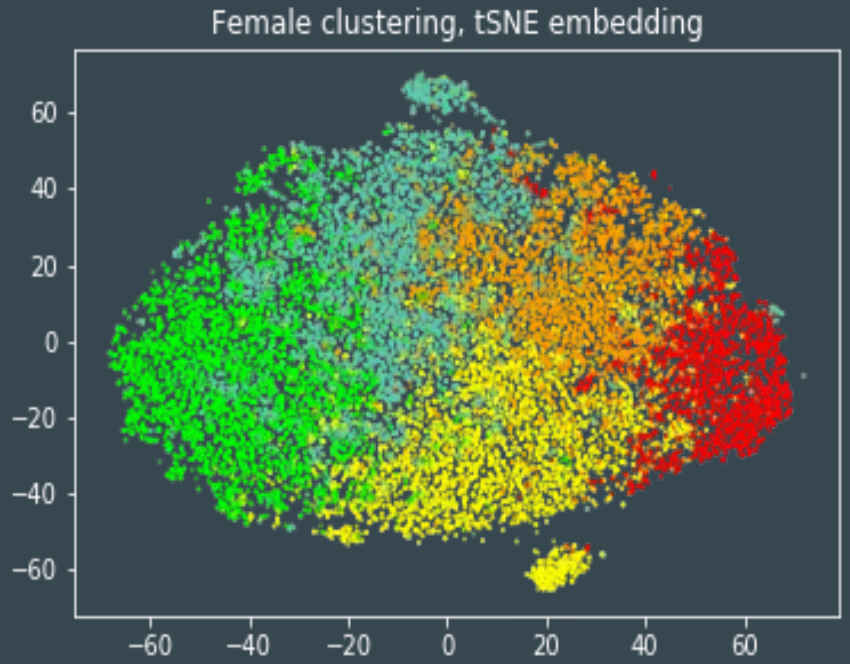
Group 1: young, low fatty liver index (FLI), low BMI, high forced expiratory volume in 1 sec;

Group 2: values in normal range;

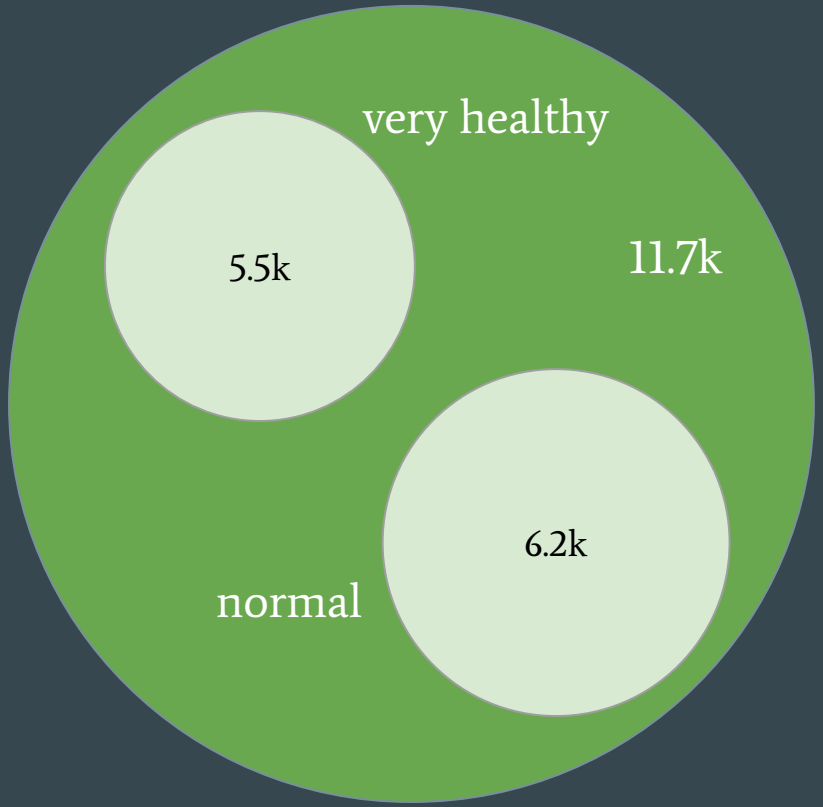
Group 3: high body fat (low body water);

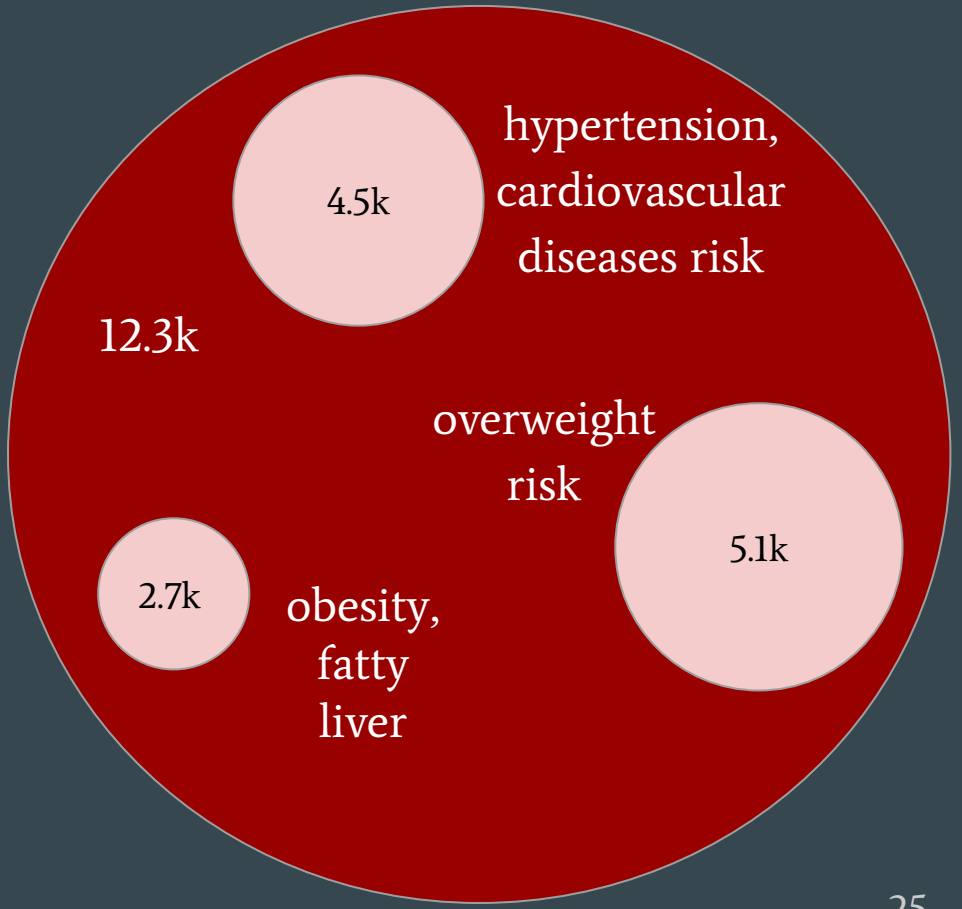
Group 4: older, high body fat, elevated blood pressure;

Group 5: high body fat, high FLI, high blood pressure.



# Female Clusters



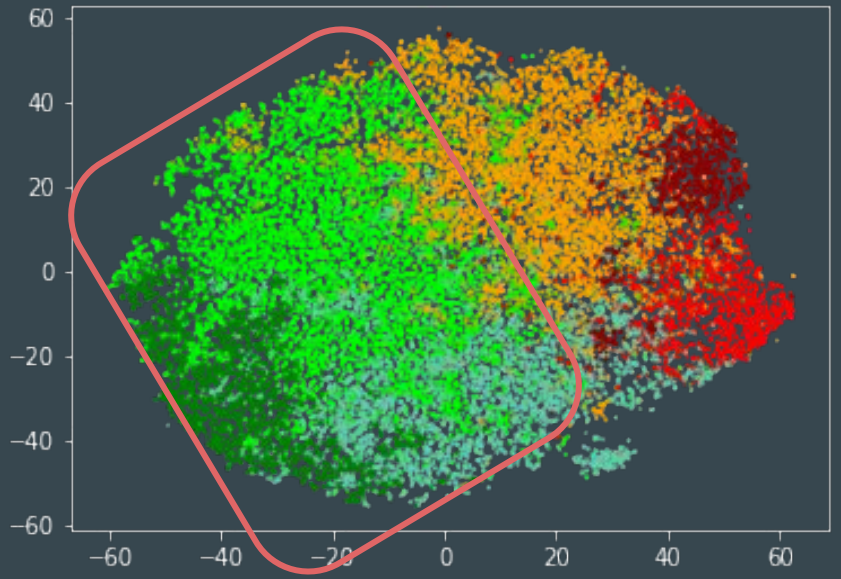



# Male Clusters

**Low risk groups 20.7k (63%)**

- Group 1: 3.5k (11%)
  25-35y, healthy with fit body figure and good lung functions
- Group 2: 6.4k (19%)
  30-40y, slightly high risk of fatty liver and obesity
- Group 3: 10.7k (33%)
  >40y, low risk of fatty liver, slightly higher blood pressure

Male clustering, t-SNE embedding




Mainly received high review scores

Group 1 stick to healthy habits

Group 2 and 3 need regular checkups

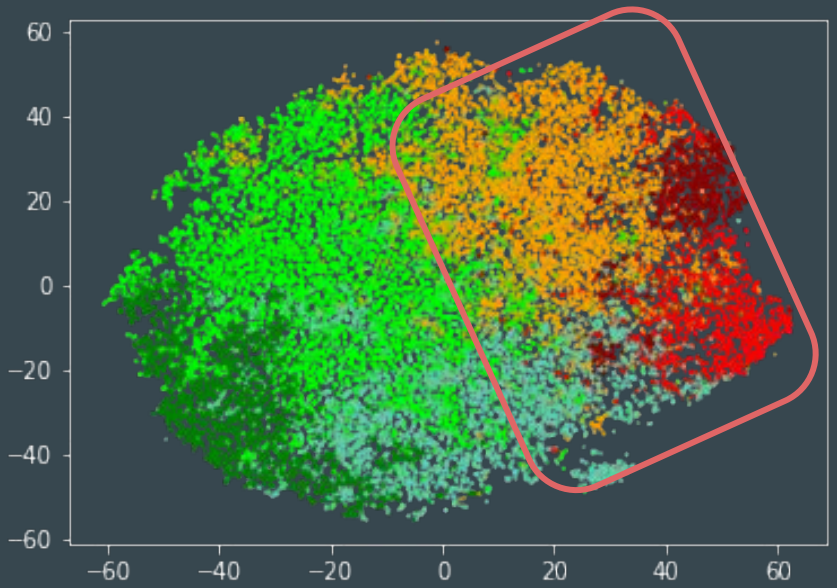
# Male Clusters

**High risk groups 12.2k (37%):**

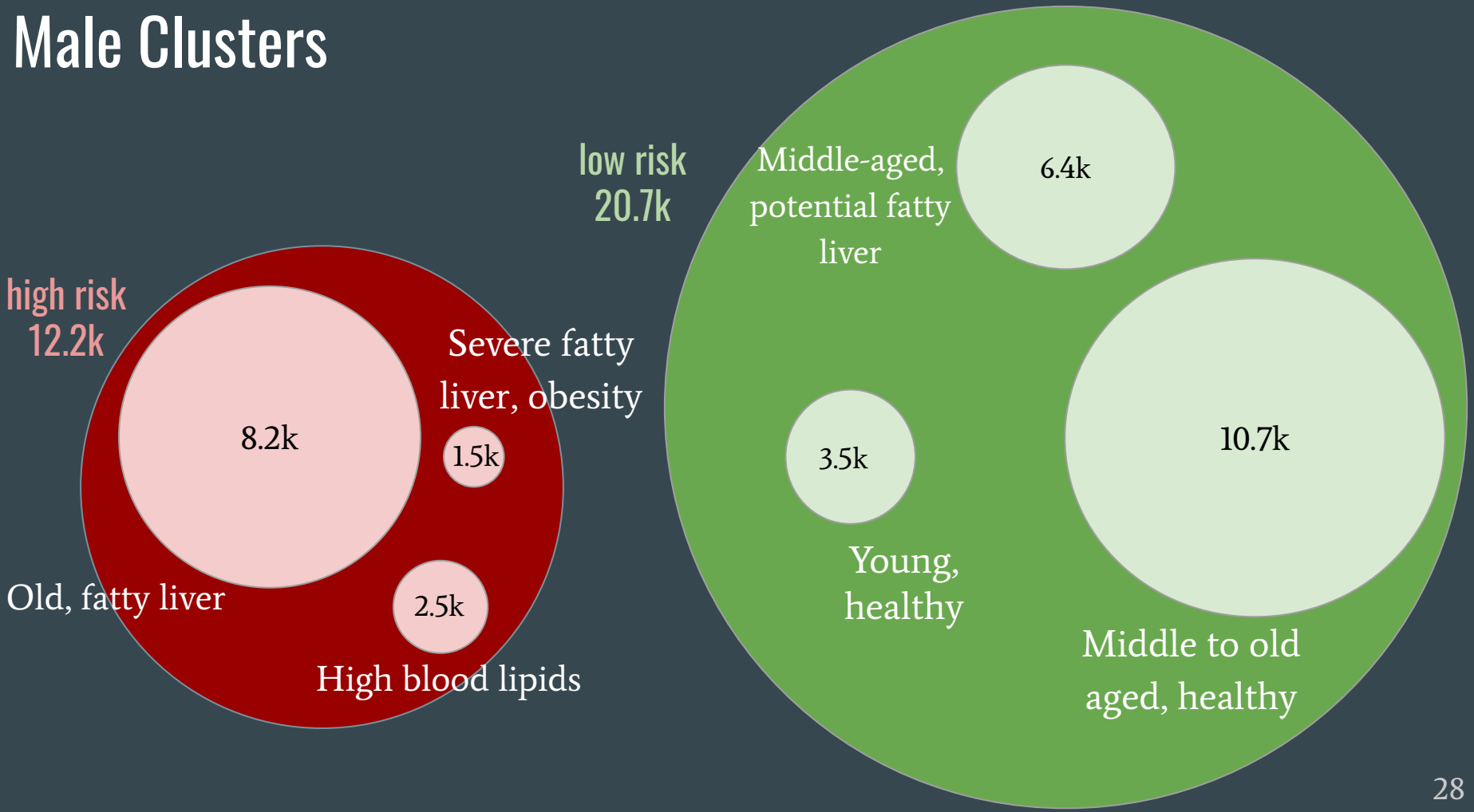
- Group 4: 8.2k (25%)
  >50y, relatively high fatty liver risk
- Group 5: 2.5k (8%)
  40-50y, very high blood lipids (fat)
- Group 6: 1.5k (4%)
  40-50y, high body fat, very high risk of fatty liver and obesity

Consult doctors
Control diet and alcohol assumption
More physical exercises

Male clustering, t-SNE embedding




# Male Clusters



# Discussions

- Interpretable groups consistent **with review scores**
  - with **synthetic data**
- Provide **medical interpretation** of each group
- Results consistent with the **high risk diseases** we wanted to detect
- Other common diseases not considered
  - Diabetes (since not fasting results)
  - Osteoporosis (since we only have bone mineral mass instead of density)

# Conclusion

# Wrap Up


- Synthetic data exploration
- Medical understanding
- Outlier handling

- Biological age prediction
- Review score prediction

- Hierarchical clustering
- Health classes definition

# Conclusion and Value of our Work

wellabe

**Define health classes**

- Interpretive clusters
- Individual recommendations

**Synthetic health data**

- Medical domain knowledge
- Limitations and potentials

# Thank you for listening!