PAUL SCHERRER INSTITUT

PSI

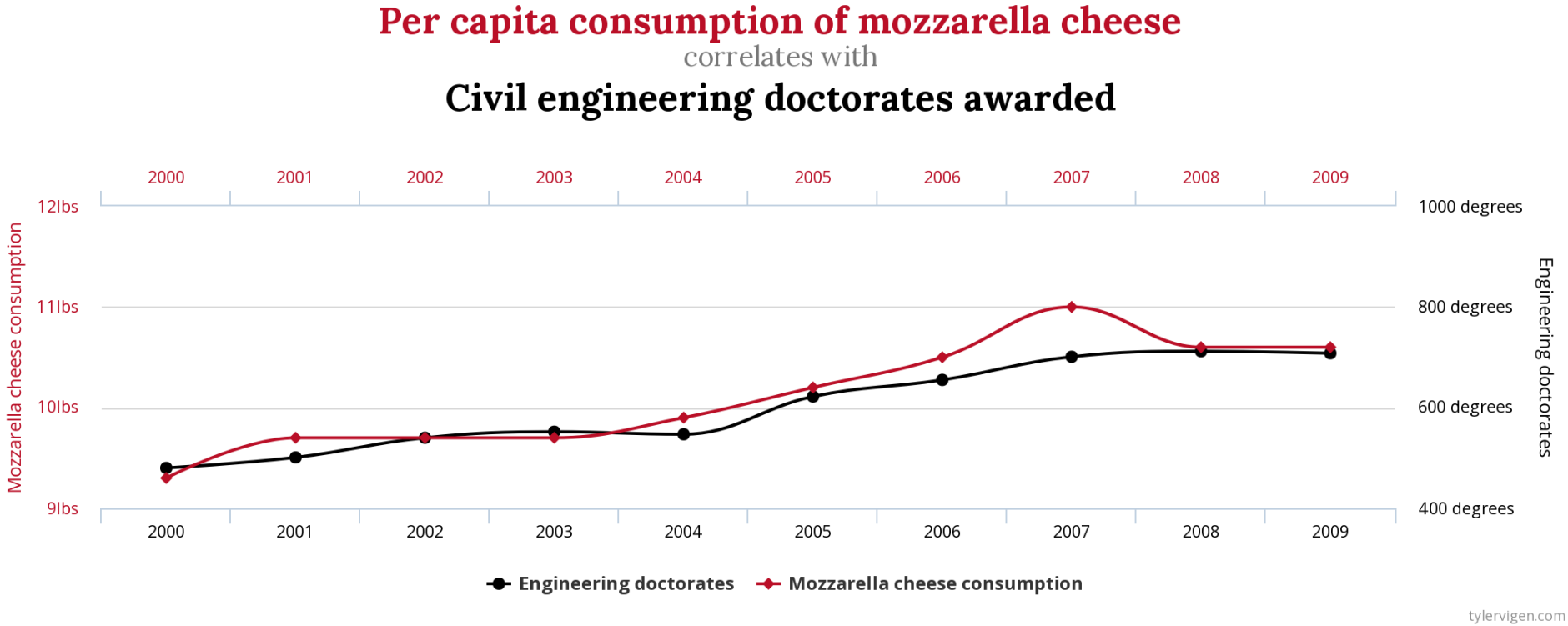
Data mining experience at HIPA

Jochem Snuverink, Andreas Adelmann

Contents

- Data mining
- HIPA
- Data mining at HIPA
- Summary

Data mining / analysis



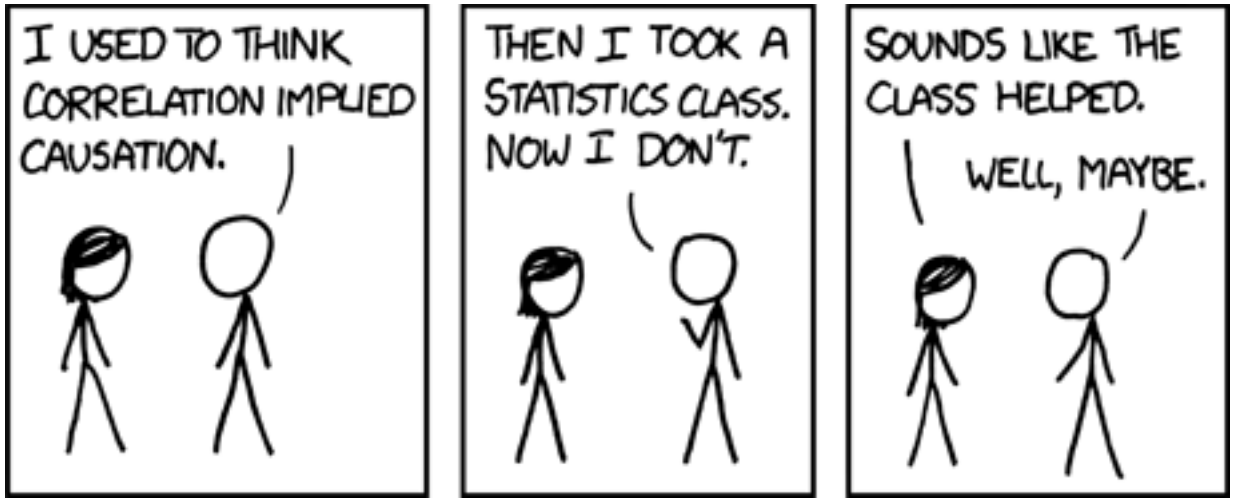
Data mining

- Most common types of data mining[1]
  - Anomaly detection (outlier/change/deviation detection)
  - Dependencies (finding relationships between variables)
  - Clustering
  - Classification
  - Regression (finding a model function)
  - Summarisation (compactifying dataset, visualisation)

1: "From Data Mining to Knowledge Discovery in Databases" – Fayyad, Piatetsky-Shapiro

PAUL SCHERRER INSTITUT

Data mining

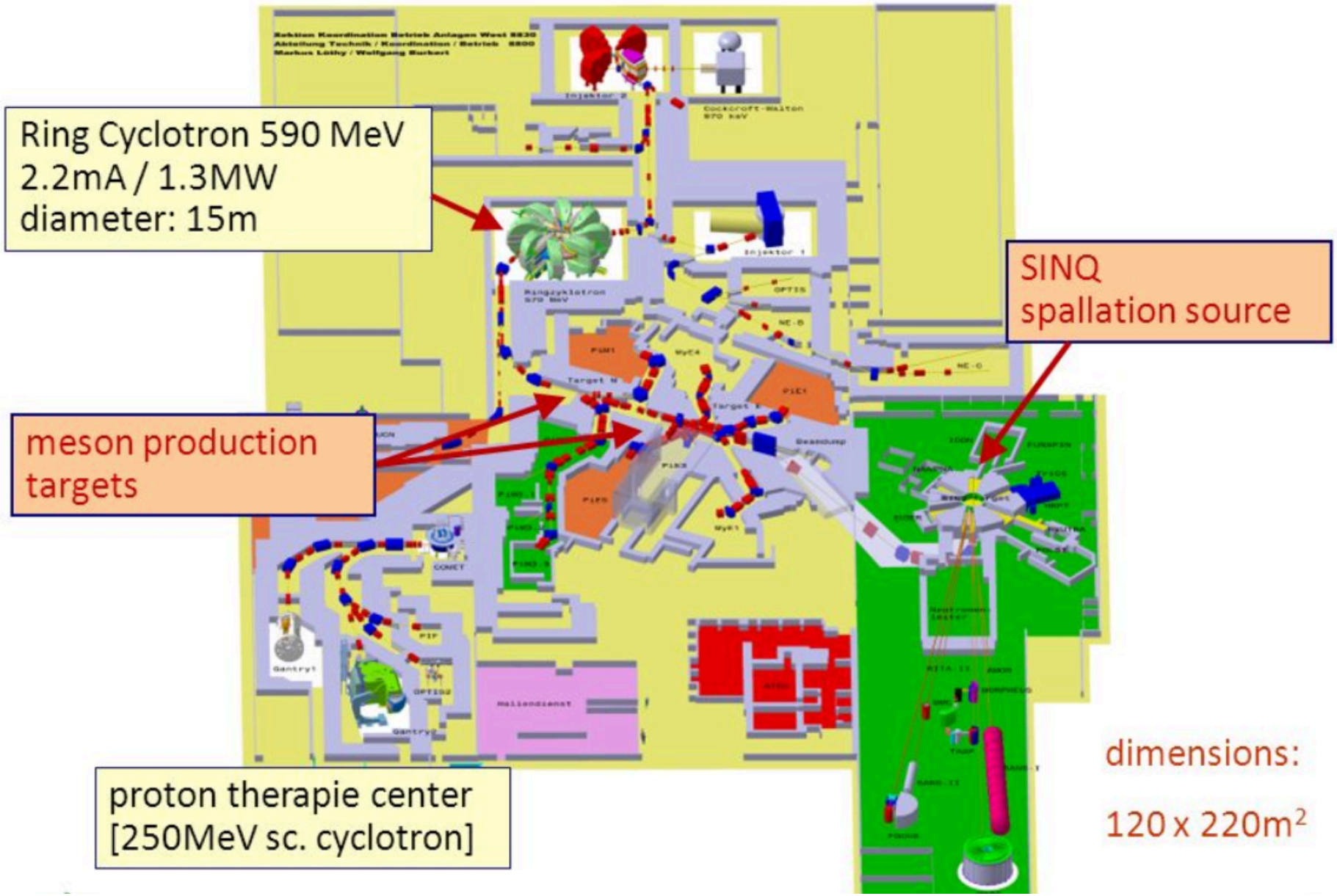


Data mining

Six phases (according to CRISP-DM)

1. Problem understanding
2. Data understanding
3. Data preparation
4. Modeling ← here Machine Learning could be used
5. Evaluation
6. Deployment

Problem understanding: HIPA



ML ideas

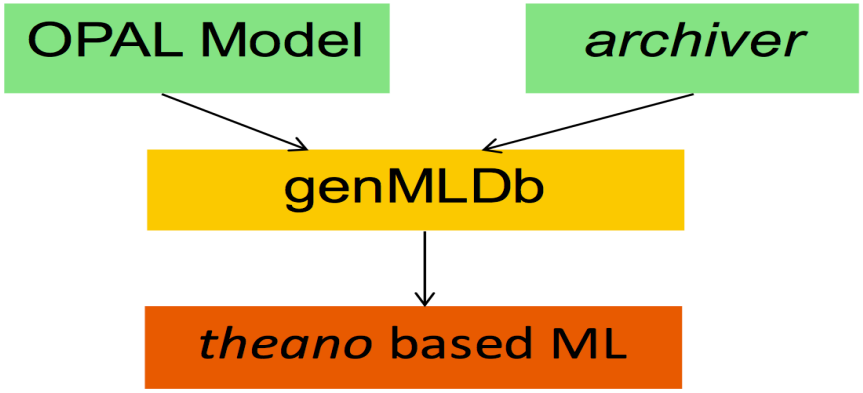
- Interlocks prediction and prevention (see talk Andreas)
  - Benefit: Increased uptime, reduced damage and activation
  - Why ML: Many different interlocks, large amount of archived data for training
  - Which type: Supervised learning with time-step memory (RNN)
- Minimise beam losses
  - Benefit: Less activation, possibly reach higher current
  - Why ML: Many tuning knobs in complicated machine
  - Which type: Reinforcement learning (Policy Gradients)
- Target spot size optimisation
  - Benefit: Improved machine protection and neutron yield
  - Why ML: Direct measurement (optical monitor) relatively slow. Predict and interpret the optical image based on machine diagnostics
  - Which type: Supervised learning with image recognition / prediction(?)
- Stable isotope production
  - Benefit: Improved isotope production and less operator supervision
  - Why ML: Product quality only afterwards. No good predictive model
  - Which type: Supervised learning with some sort of data reduction

Data understanding

- About 80 GB data / year
  - 30+ years of data
- 3 machine databases with different set of variables:
  - Short term (<10 days)
  - Medium term (<10 months)
  - Long term
- Variables are logged differently
  - monitored versus different fixed rates
- Interlock database
- Isotope production quality (excel sheet)

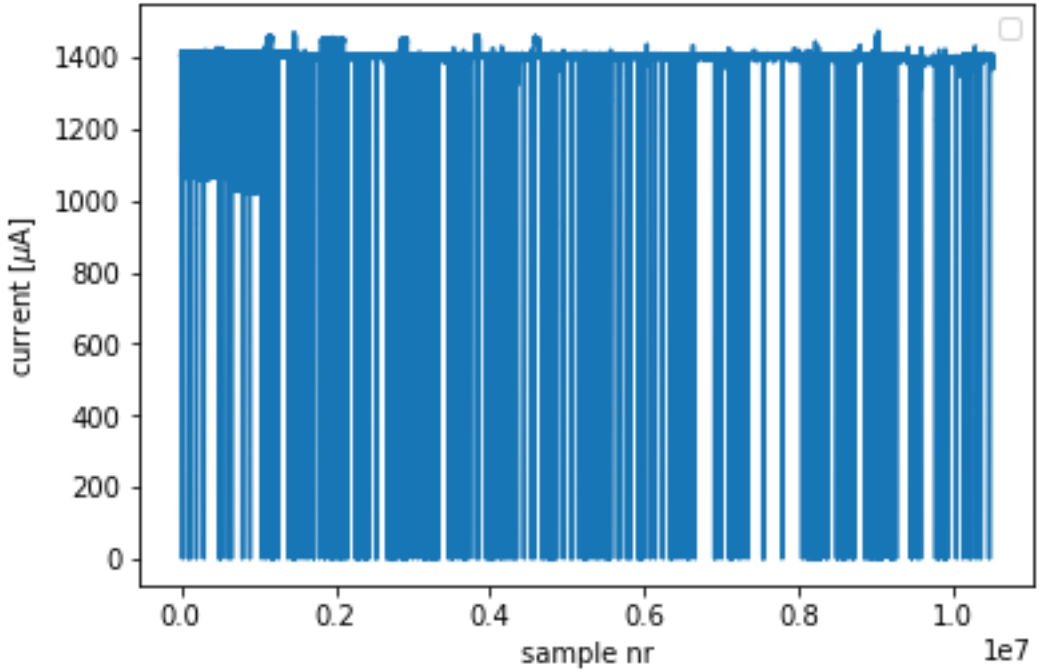
Data preparation

- Select list of variables: magnet currents, diagnostics (BPMs, loss monitors), etc.
  - problem dependent
  - for interlock study only ring cyclotron and high energy line
- Get data from archive with command line tool:
  - Variables missing ☹ -> adjust list, about 90 left
  - Simple usage puts a new line for every parameter change
    - Unmanageable for more than a few variables
    - Fixed data rate of 10 Hz (intrapolated values)
  - Crashes when trying to limit significant values
  - 120MB / day (binary format), 1e7 samples
  - Missing values -> remove whole sample
- Combine interlock data
  - add interlocks to closest timestamp (are we sure clocks synchronised?)

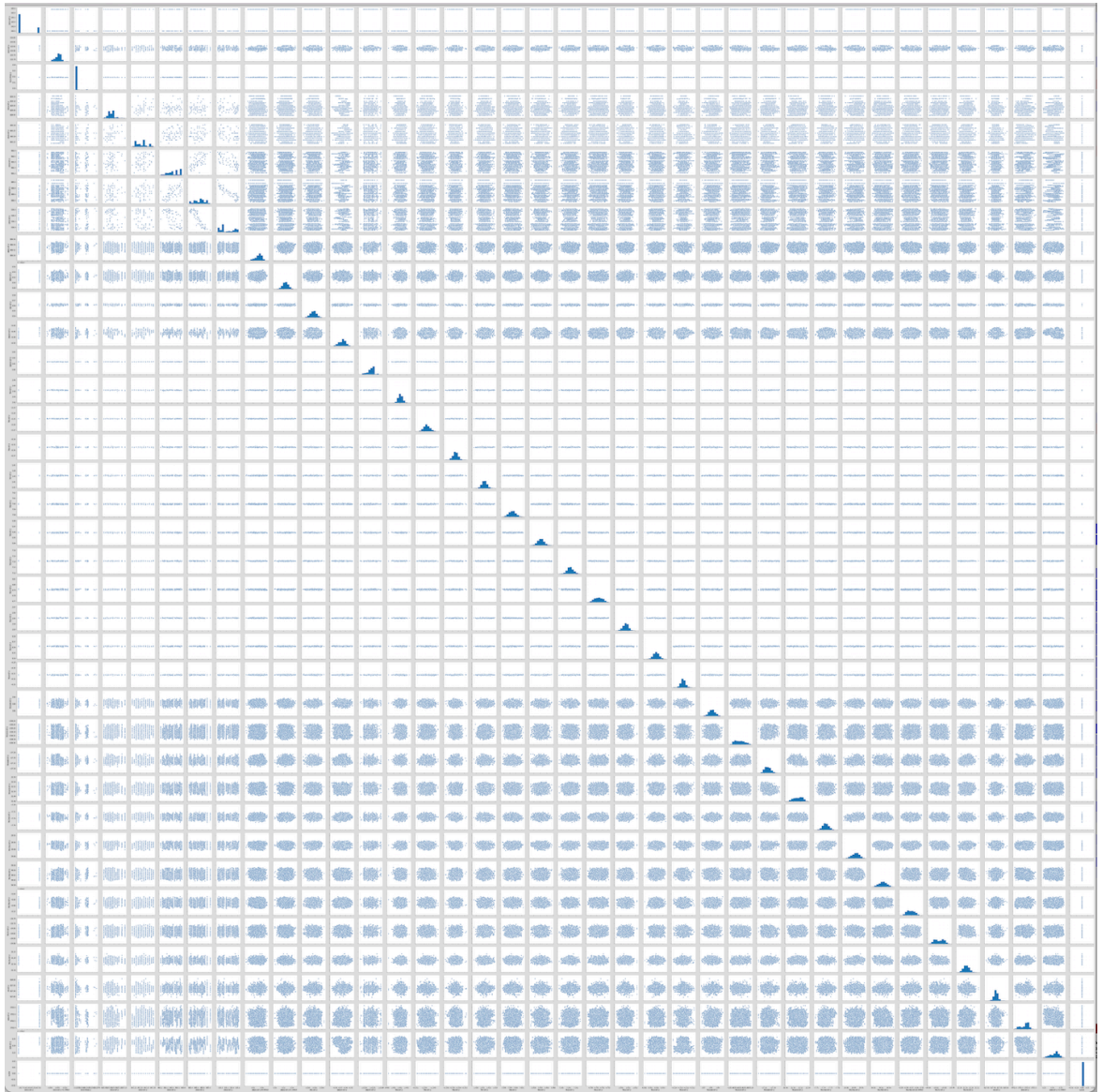


Data selection

- Clean variables
  - variables not changed significantly (from physics point of view)
  - remove variables that have correlation higher than 0.95 with another
    - will probably not add much to a model
- Clean samples
  - Remove bad machine conditions
    - machine off or ramping: current < 1300 uA

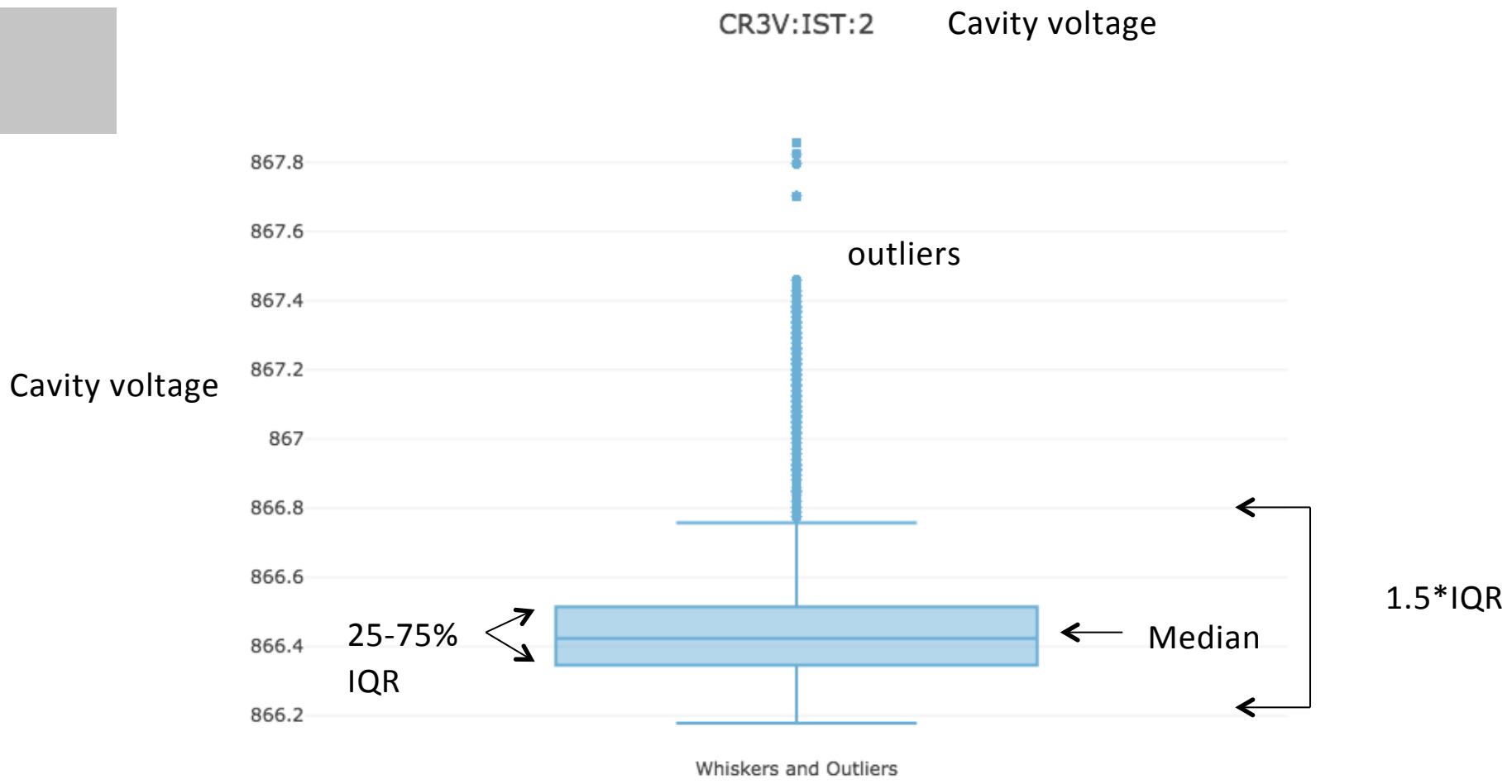


Visualisation – scatter plot matrix



Probably not so useful

Visualisation – box and whisker plot



Visualisation – correlation matrix

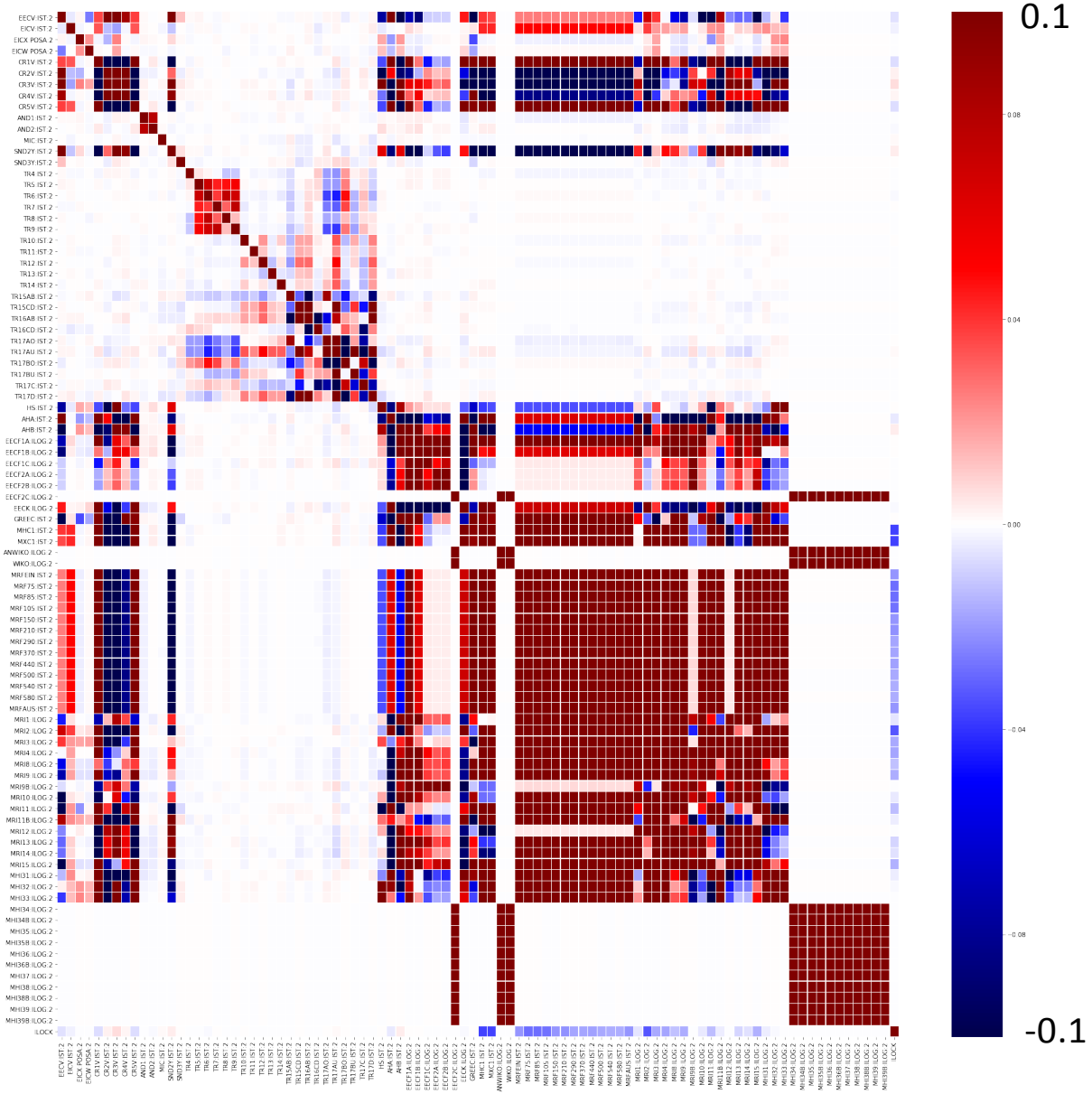
extraction and RF
in ring cyclotron

trim coils
in ring cyclotron

phase and loss
monitors
in ring cyclotron

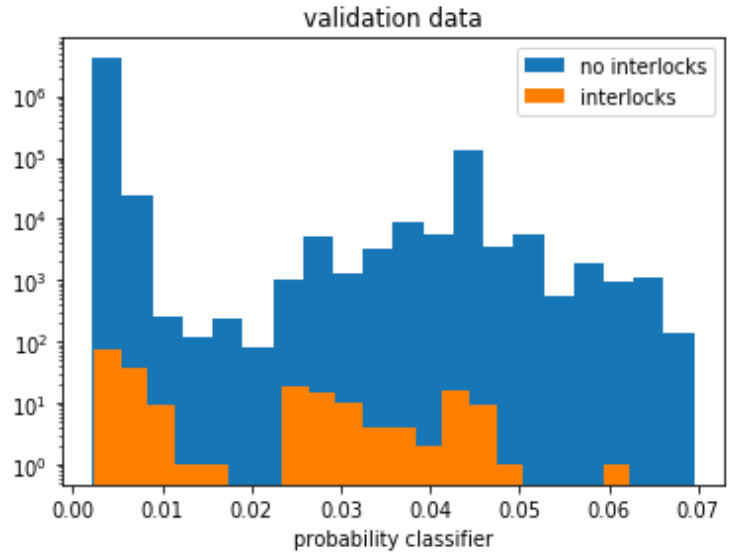
loss monitors in
targets beamline

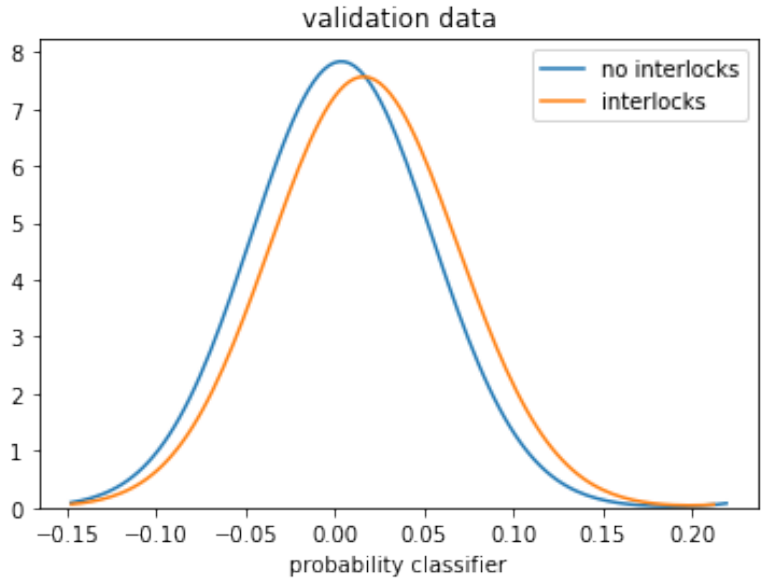
Interlocks →



Modeling – binary classifier

see talk Andreas

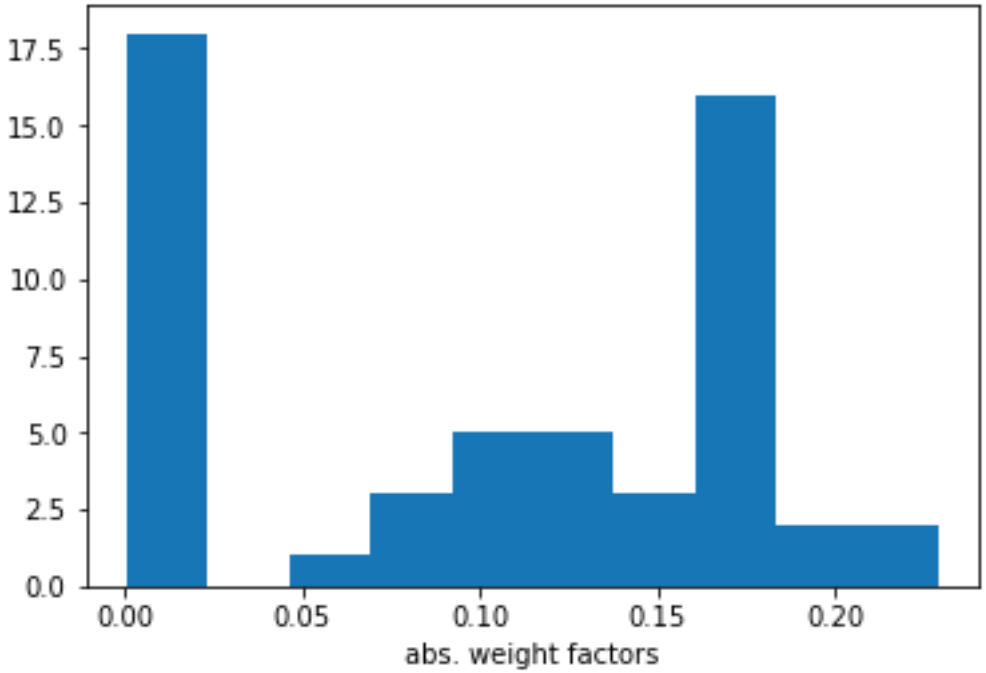




density plot
some discrimination power
but many false positives

Evaluation

- Network weights



- As expected the variables with significant weight correspond to those with high correlation.

Summary and Outlook

- Simple methodology for data mining
- Personal experience on HIPA data shown
  - Data preparation step most tricky
    - Discuss with controls group how this can be improved
  - Data normalisation needed for ML
- Some simple visualisation plots that can guide for large amounts of data
- Simple regression model
  - Reduce false positive rate
  - Add predictive power (RNN)