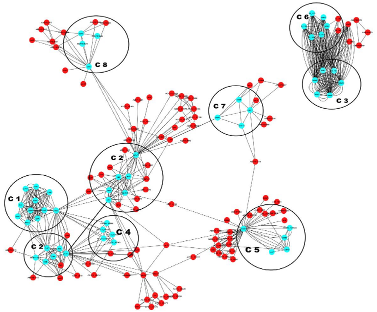
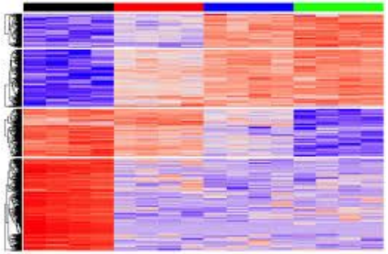


Translator in science

Juliane Siebourg-Polster



Introduction

Today I cover

- How I got to where I am
- What I am doing at Roche
- Two example topics I work on
  - Experiment design
  - Missing value imputation in omics data
- What else it needs to do my job (aka soft skills)

A bit about me

I am Biomathematician / Computational biologist

from Bonn (Germany), living in Basel since 2009

Currently working as a Principal Biostatistician at Roche pRED (pharma research and development)

I enjoy working at the interface of biology, data analysis and mathematical modeling. In pre-clinical drug discovery and development I focus on statistical analysis and modeling of high throughput omics biomarker studies as well as experiment design. In addition I engage in teaching statistical methods and concepts to colleagues.

How I got to where I am

- Studied Biomathematics in Greifswald
  - (2003 - 2009)
  - Focus on Discrete Mathematics
- Study semester abroad in New Zealand
  - (2006 - 2007)
  - Molecular evolution / Phylogenetics
- PhD in Basel at ETH Zürich
  - (2009 - 2014)
  - Statistical modeling in computational biology
- Biostatistician at Roche pRED
  - (since 2014)

UNIVERSITÄT GREIFSWALD
Wissen lockt. Seit 1456

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Roche

What I do at Roche

- Scientific project support
  - Working in small teams with biologists, chemists, medical doctors and lab associates
  - Studies:
    - Biomarker identification
    - Mode of action of a drug / Off targets identification
    - Disease Understanding
  - My tasks:
    - Experiment planning
    - Data analysis
    - Results interpretation with collaborators
- Innovation
  - In a constantly evolving environment, we stay up to date and use / provide cutting edge data analysis tools
  - Literature reviews, conferences, tool development, publications
- Teaching
  - Voluntarily train colleagues on basic concepts of experiment design, statistics, scripting, etc.

Experiment design: The Request

Requestor
Hey! We want to measure proteins in patient's eyes with a new technology. We have 2 disease groups treated with 4 concentrations of a compound at 7 different time points. Are 210 samples enough? Can you analyse this?

Juliane

- What is the question you want to answer?
  - E.g. which groups do you need to compare?
- Do you have an idea about a positive or negative result?
  - Which control samples will you include
- What technical limitations are there?
  - Sample batching to plates / days
  - What do we know about the variability of the platform?
  - Is the number of samples limited?
- How will the experiment be performed?
  - Which steps require sample blocking & randomization

Requestor
Oh …

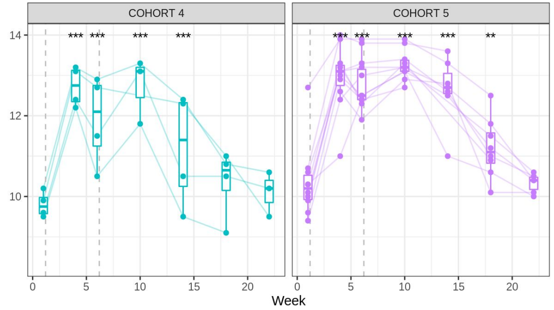
Juliane
Let' have a meeting and discuss!

Experiment design: The Objective

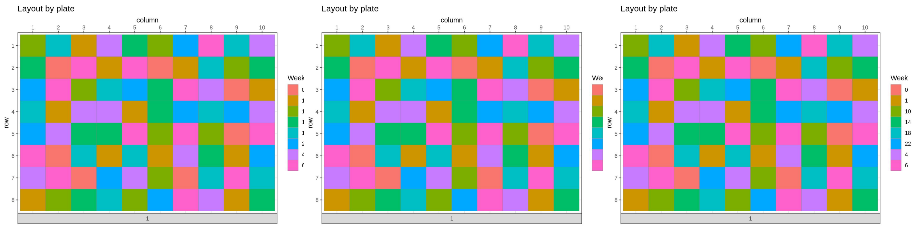
Question:
For a given disease: Are there proteins for which we see
- A treatment related effect in patients at any of the time points.
- A an in- /decrease with increasing dose of the treatment.

Details:
- Controls: untreated measurements at day 1
- The maximum number of patients is limited by the ongoing study
- Samples are distributed on 3 plates (fitting 96 samples), **with strong plate to plate variability**.
- Within plate effects are possible.

Ok! So, the study would look something like this:



And samples would need to be put on plates like this:



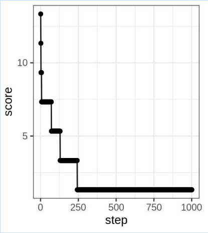
Distribute samples such that dose-group / sex / age effects are not **confounded** with plate effects.

Experiment design: The Solution 1

1) To best compare time points **within** a patient! → Put all samples of a patient on the same plate
2) **Block** patients by dose-group / sex / age E.g. distribute these evenly across plates

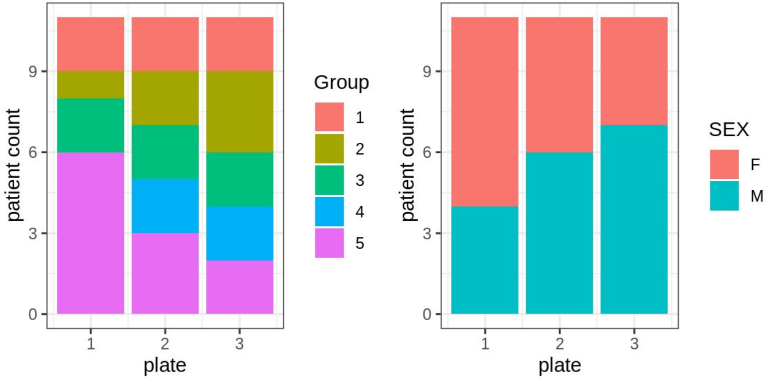
Optimization:
- Greedy iteration
- Objective function:

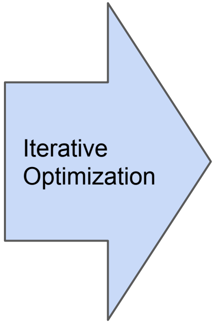
Pearson's Chi-squared test to test independence between *plates* and *sample variables*

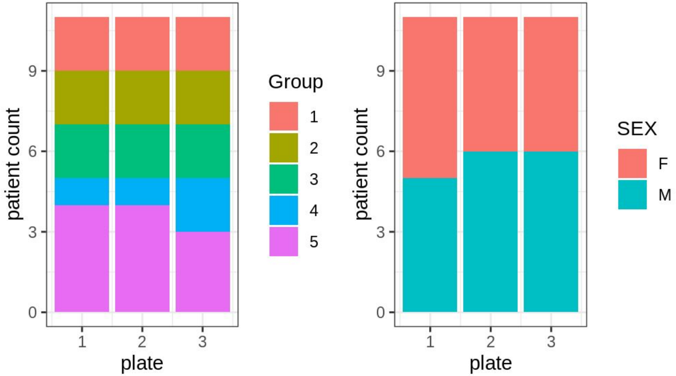


Random distribution of patients to plates

Optimized distribution of patients to plates







Experiment design: The Solution 2

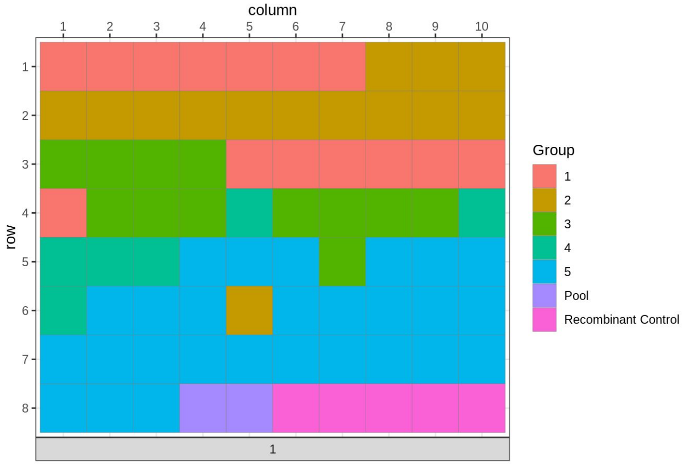
3) Within plate randomization of samples by patient / dose-group / sex / age

Optimization:

- Simulated annealing
- Objective function

Based on maximizing the 2D distance between samples of a group.

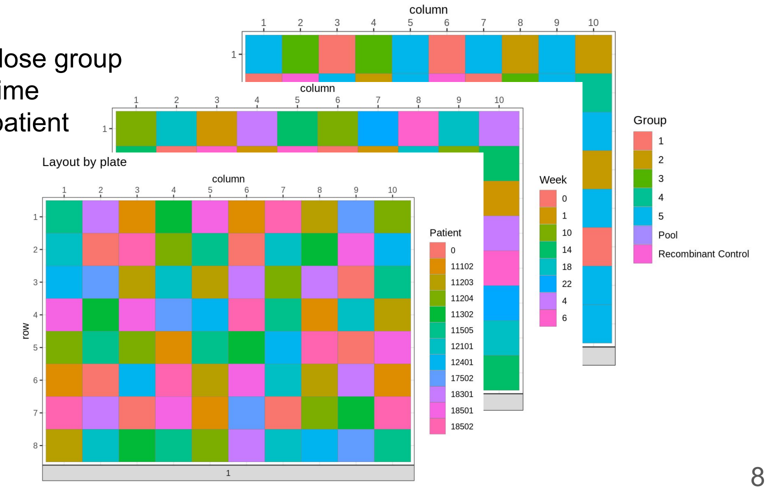
In order assignment of samples within plate 1



Iterative Optimization

Optimized distribution of samples within plate 1

By dose group
By time
By patient

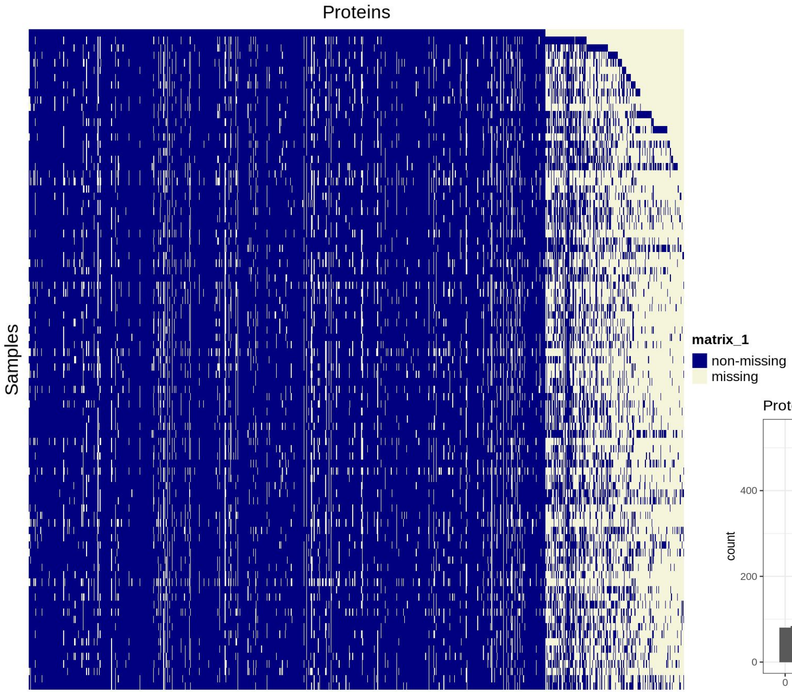


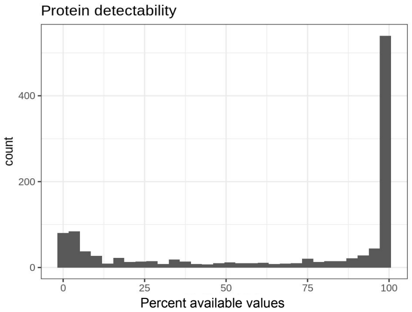
Experiment design: Implementation

- Dedicated team of 4 people
- Literature + existing tool evaluation
- R package development, combining
  - Usage of existing tools
  - Implementation of new optimization strategies
  - Layout examples + markdown templates for larger user community
- … to be published

Missing value imputation: The problem

A *Mass Spec* dataset with high missing rates





Reasons for missingness

- MAR (missing at random): independent of any variable in the data → rare technical failures
- MNAR (missing not at random): missingness process depends on missing values (e.g. below level of detection values (<LOD))

MS data often has a mixture of MNAR and MAR

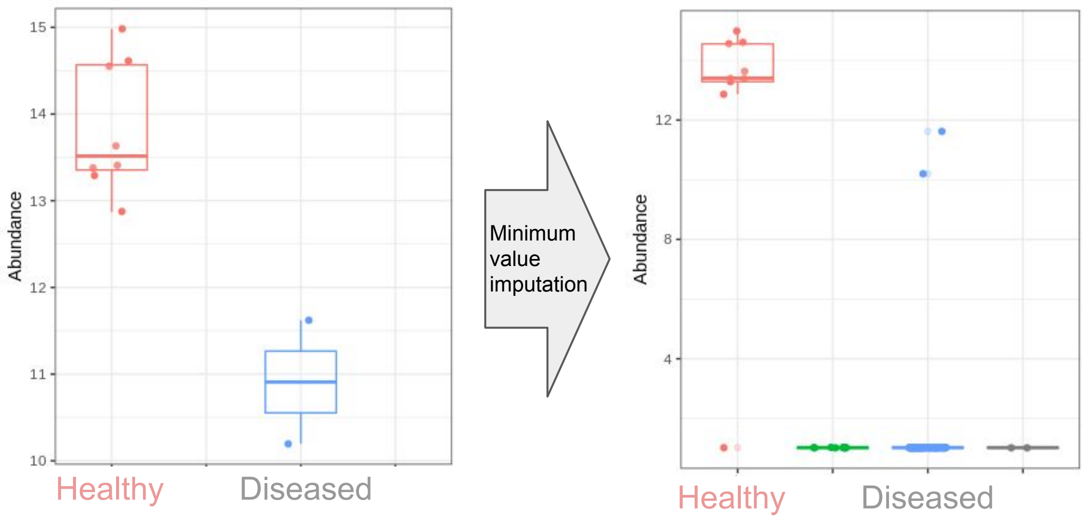
Questions:

- Which proteins to filter out
- Which to keep and
- How to **impute** them

Missing value imputation: Why we care

MNAR patterns can carry valuable information!
E.g. for a protein all patients have missing values, all controls have proper values.

- We don't want to remove such proteins
- But many analysis tools rely on **complete data**



Data imputation
The process of replacing missing data with substituted values, estimated based on other available information

There are many methods available. The choice depends on the context of the data and type of missing value (Random, LOCF, PCA based, KNN based)

Missing value imputation: The Strategy

Strategy to remove junk

1. Presence test for each protein:
   - Fisher test for independence between *missingness* and *disease group*
2. Filter proteins:
   - Remove when missing rate high
   - **except** when passing the presence test
3. Impute remaining missing values as a mixture:
   - For each protein, decide on MAR / MNAR
     - MAR: missingness is low → values close to the median
     - MNAR: missingnes is high → values below the minimum of detection

Results for the example dataset

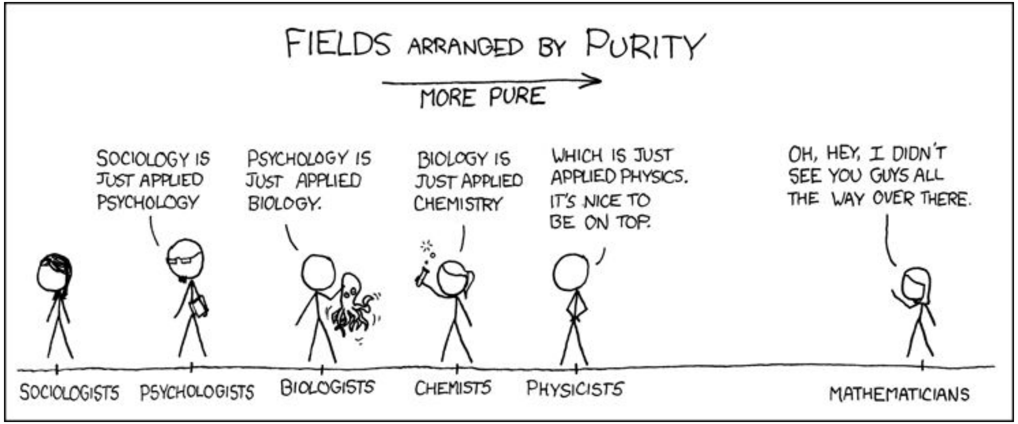
```
Of 1134 features
23 (fdr < 0.05%) / 18 (fdr < 0.01%)
have a differential presence absence
pattern between groups.

843 features are kept as they have more
than 30% (n >= 27) observed points, or
significant differential missing
patterns per group.

Of 843 features
35 have MAR values with max 2% (n >= 2)
missing samples
332 have MNAR values imputed by minimum
for more than 2% (n >= 2) samples.
```

What else it needs to do my job

Like acting as an **interpreter** (biology ↔ statistics/math)

- Translate the question of a biologist into a study design + analysis strategy
- Translate analysis results into human understandable statements

Communication is key!

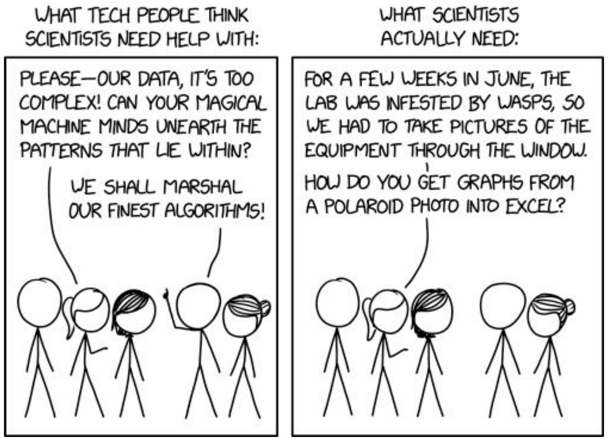


Bridge that gap!

What else it needs to do my job

Be ready to **implement** your ideas

- Programming / scripting is daily work
- e.g. R package development, bash scripting
- Prepare for data wrangling at many levels



What else it needs to do my job

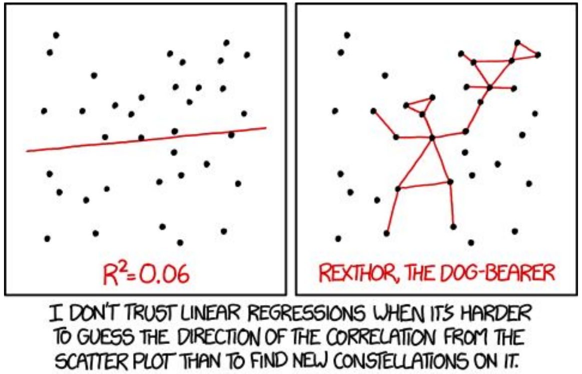
Be **truth-seeking** → sometimes you have to be the *bad guy*

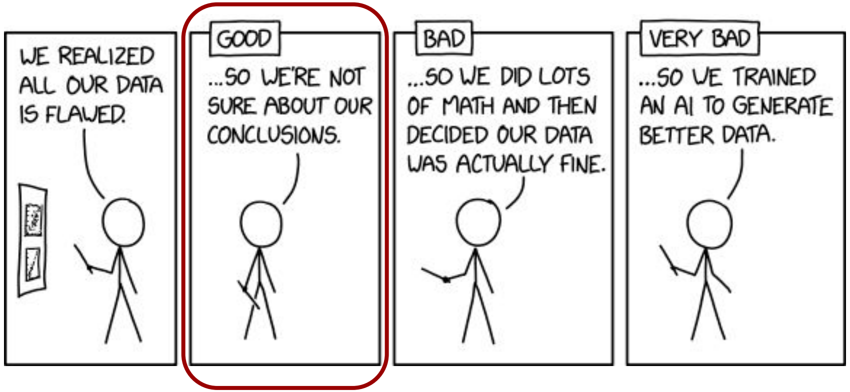
Data analysts are often least biased about a study outcome and can prevent

Hey, we know this protein and it's correlating!

cherry picking or *nicening* results

But can't we still make use of this data?





Questions?

Acknowledgements

- The *designit* R package team (Iakov Davydov, Guido Steiner, Balazs Banfai)
- The BEDA proteomics guild (Balazs Banfai, Javier Gayan, Sabine Wilson, Zhiwen Jiang, Francois Bartolo)
- https://xkcd.com/ for the comics