When the hay looks like needles

Statistical challenges in omics data mining

Patrik Edén
2015-02-25

Theoretical Physics:

Najmeh Abiri, pdh student
Patrik Edén, researcher
Mattias Ohlsson, researcher
Carsten Peterson, professor

Immunotechnology:

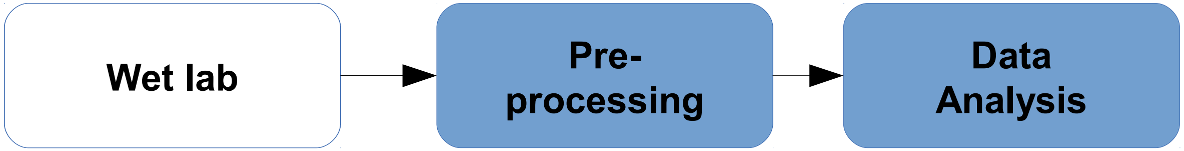
Payam Delfani, phd student
Christer Wingren, professor

Wet lab:

(all up to scanning)

Data analysis:

E.g. supervised feature selection
Biomarker search, profiles, classifiers
for diagnosis, prognosis, personalized medicine



Pre-processing:

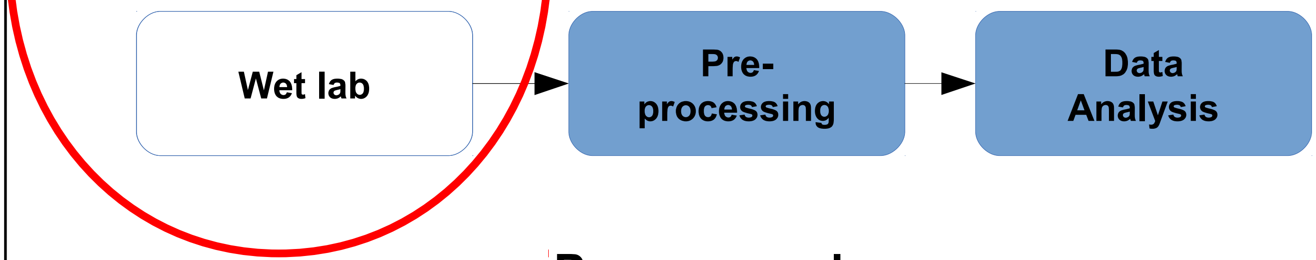
Quality control
Correction for technical effects (e.g. slide-to-slide effects)
Noise reduction (filter low-variance reporters)

Wet lab:

(all up to scanning)

Data analysis:

E.g. supervised feature selection
Biomarker search, profiles, classifiers
for diagnosis, prognosis, personalized medicine

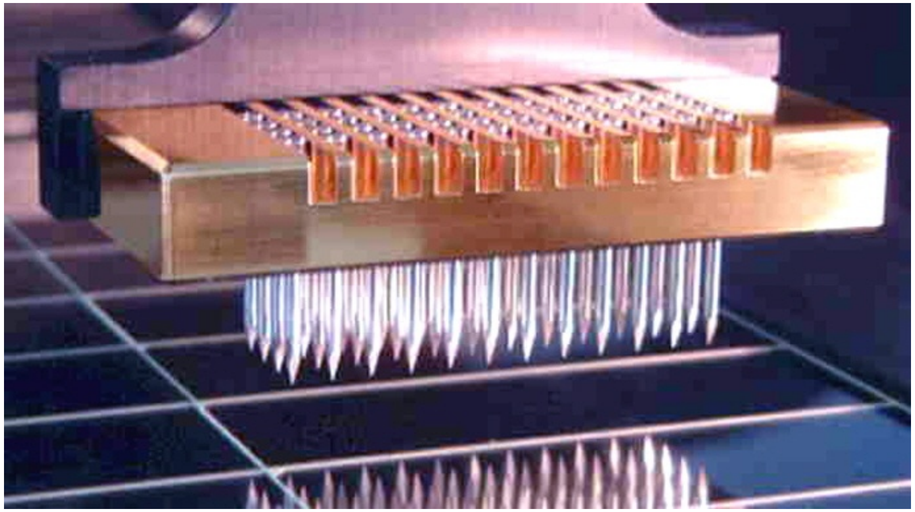


Pre-processing:

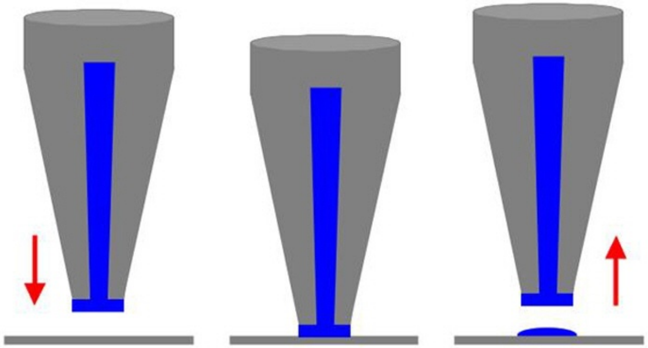
Quality control
Correction for technical effects (e.g. slide-to-slide effects)
Noise reduction (filter low-variance reporters)

Experiment:

I: Print dots on surface. Each dot with specific “probes” (antibodies)



ArrayIt.com

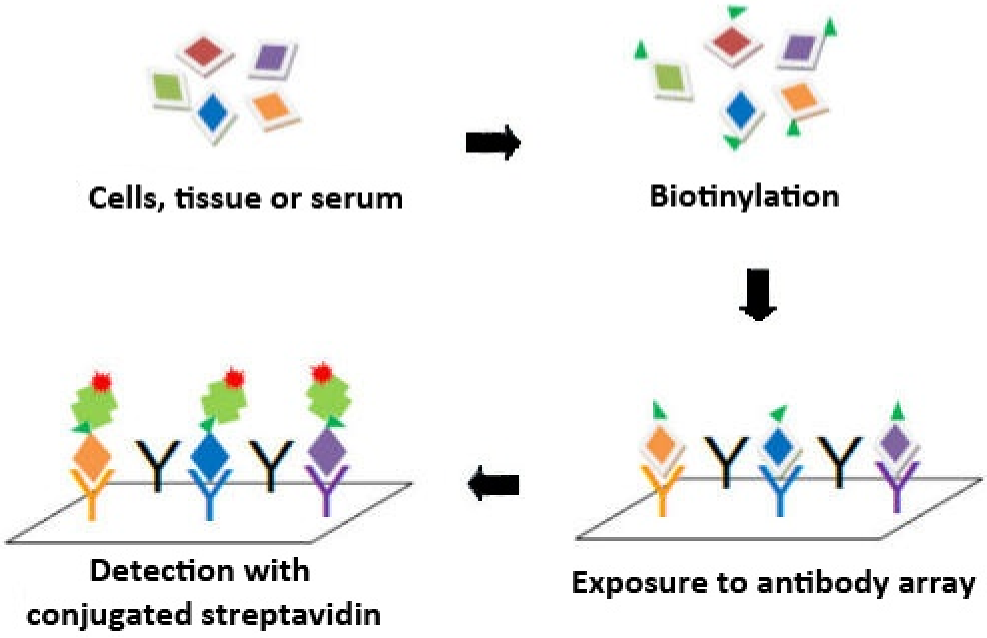


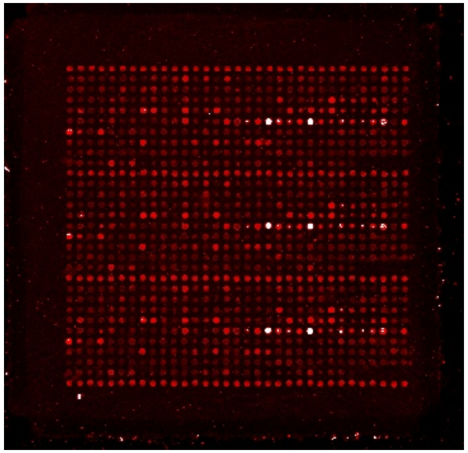
Experiment:

II: Make molecules in sample flourescent (or binding flourescence)

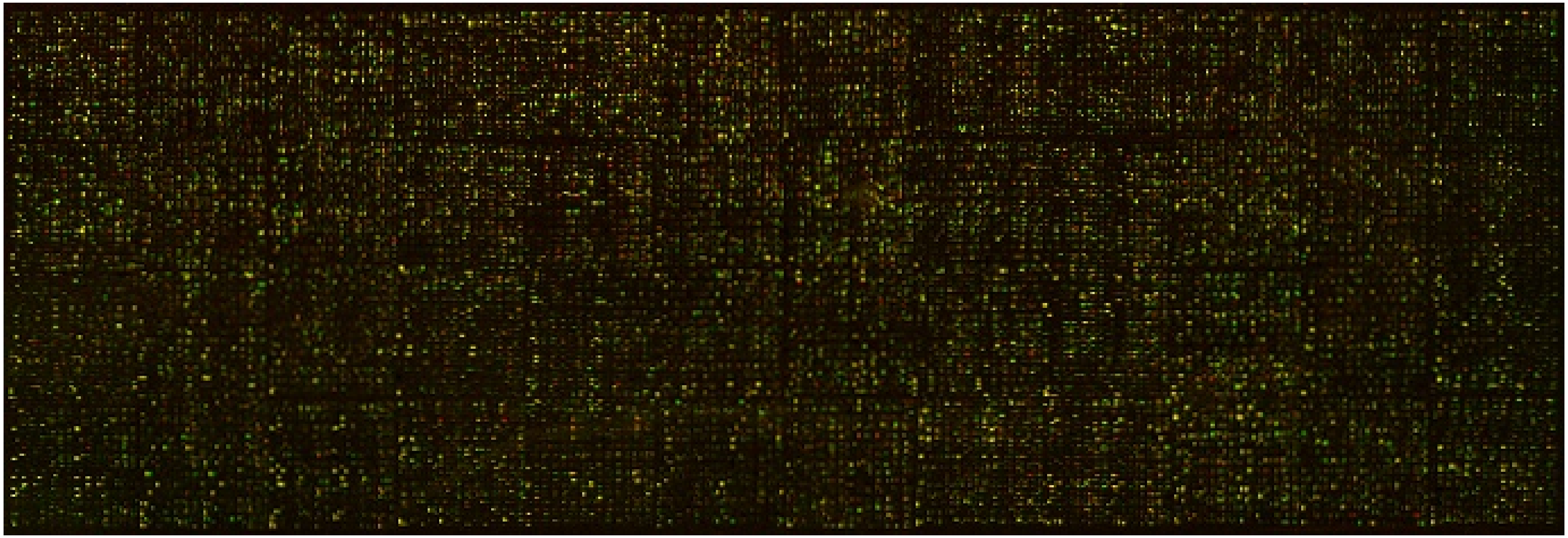
III: Pour onto surface

IV: Scan. Intensity=multiplicity.


Cells, tissue or serum
Biotinylation
Exposure to antibody array
Detection with conjugated streptavidin



Borrebaeck, Wingren, et al.



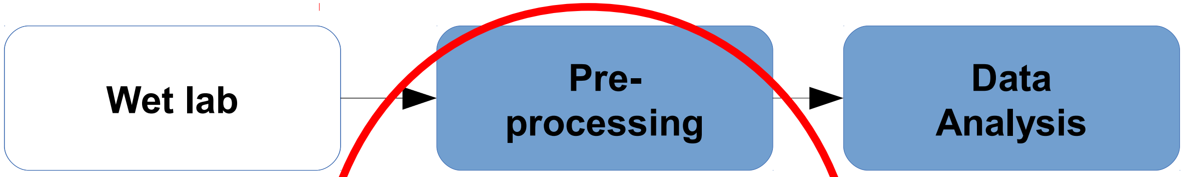
30.000 simultaneous measurements of mRNA.

Wet lab:

(all up to scanning)

Data analysis:

E.g. supervised feature selection
Biomarker search, profiles, classifiers
for diagnosis, prognosis, personalized medicine



Wet lab

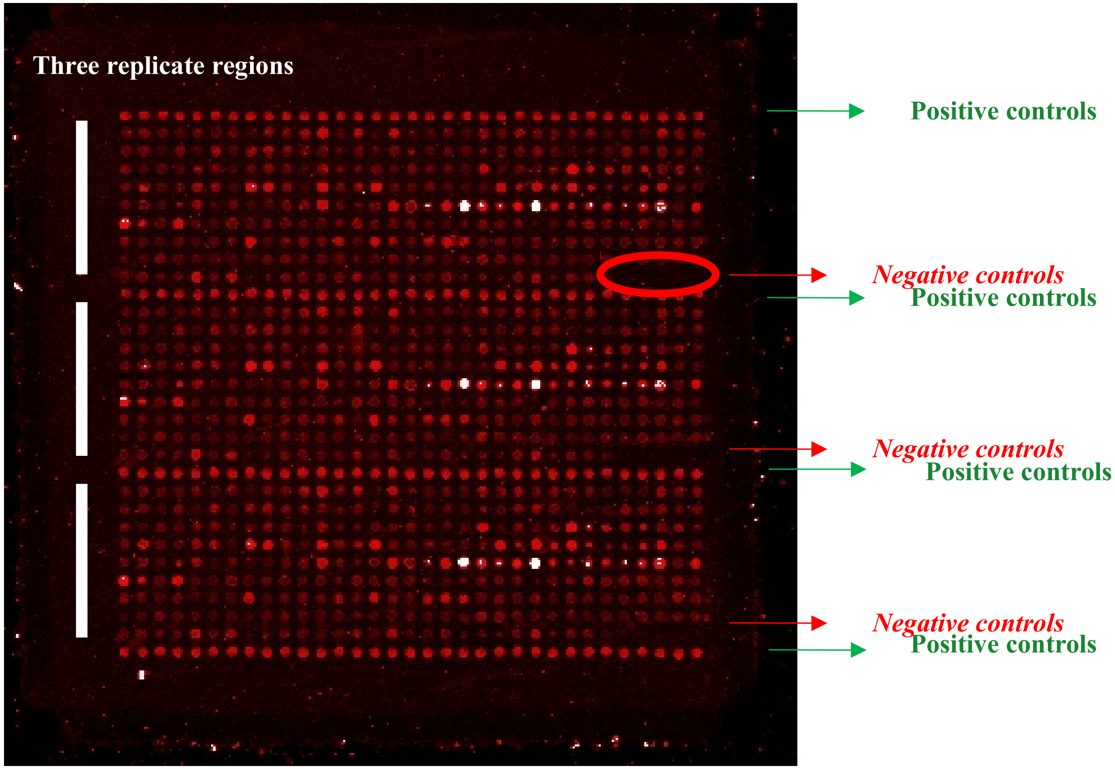
Pre-processing

Data Analysis

Pre-processing:

Quality control
Correction for technical effects (e.g. slide-to-slide effects)
Noise reduction (filter low-variance reporters)

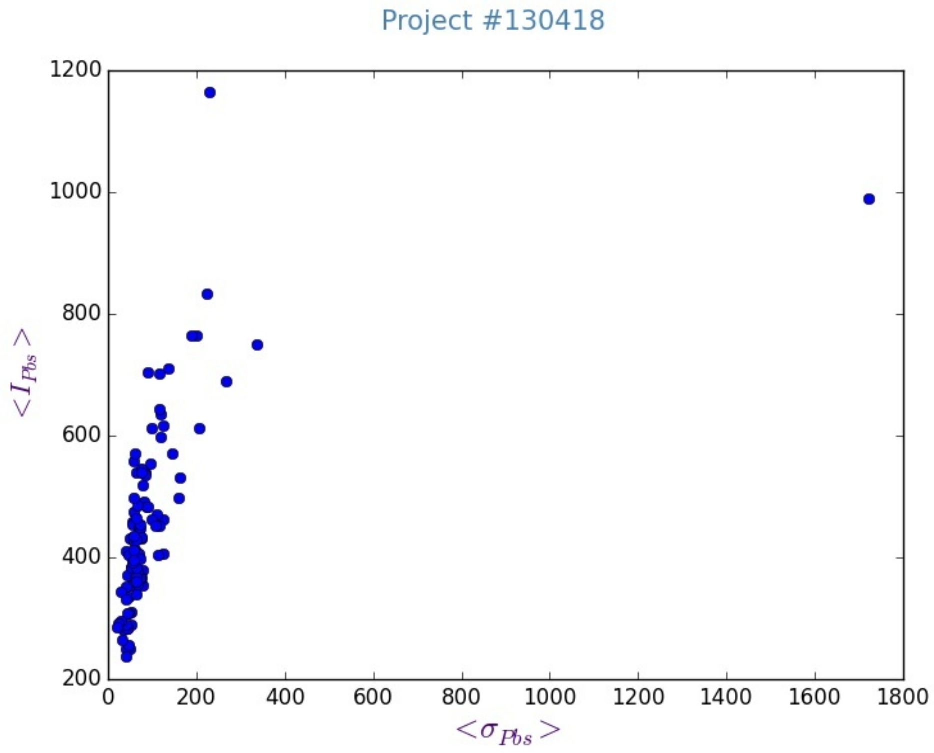
Pre-processing



Example of open pre-processing question

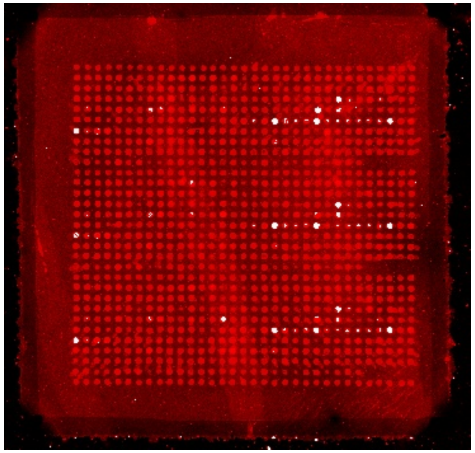
(*N.Abiri, ongoing work*)

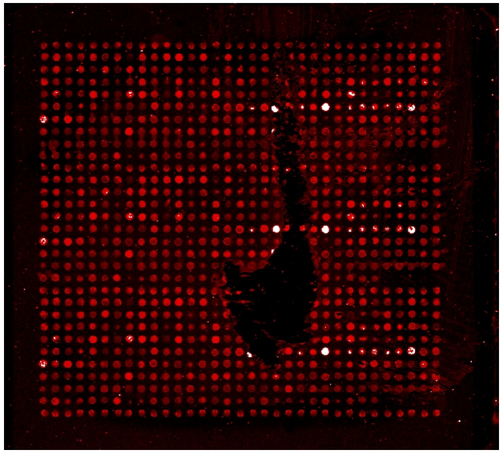
Mean intensity of negative controls vary from one array to another.
How is that best compensated?



Example of open pre-processing question

Quality control is important. What are the criteria?





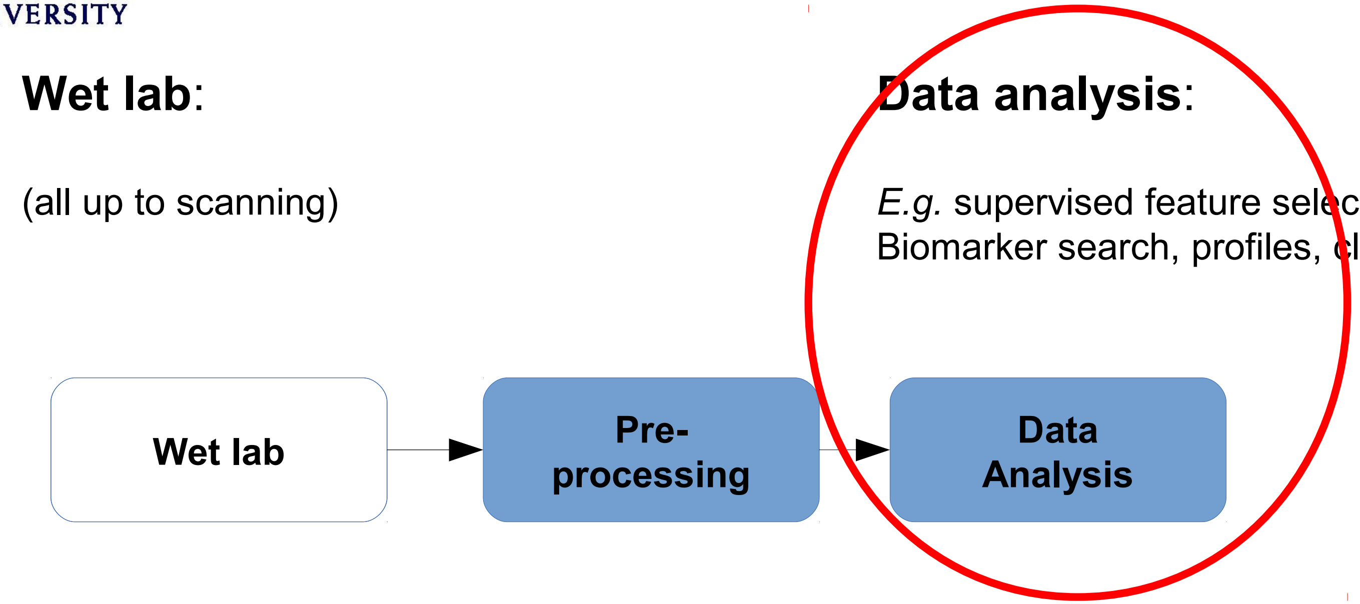
Pre-processing “necessary evil”.
These slides illustrated “necessary”
Soon: and “evil”

Wet lab:

(all up to scanning)

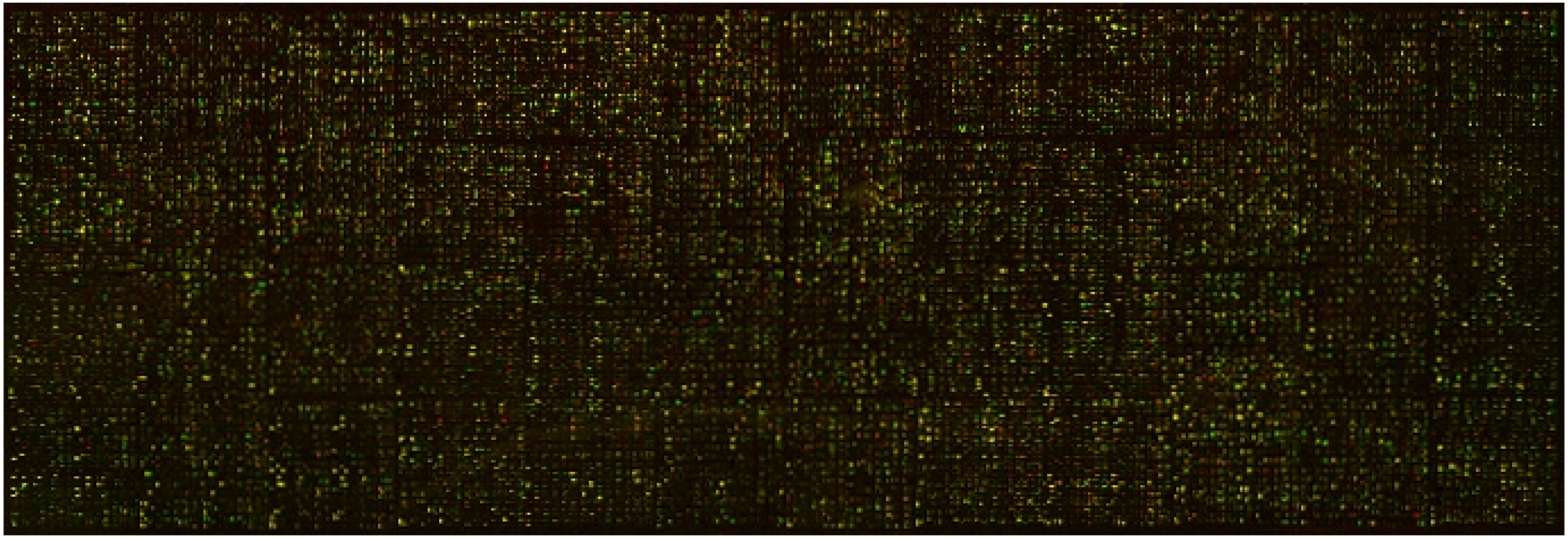
Data analysis:

E.g. supervised feature selection
Biomarker search, profiles, classifiers



Pre-processing:

Quality control
Correction for technical effects (e.g. slide-to-slide effects)
Noise reduction (filter low-variance reporters)

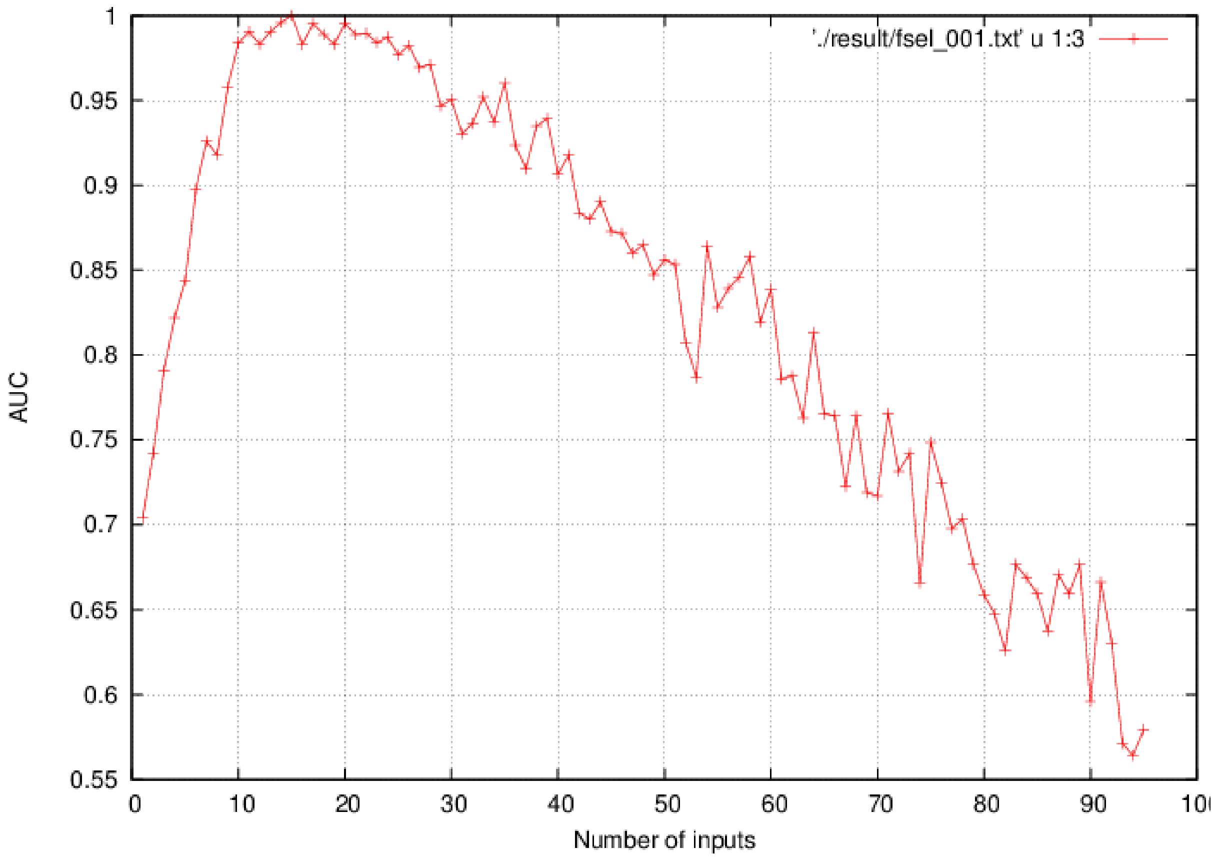


30.000 simultaneous measurements of mRNA.

Typically, 100-1000 samples.

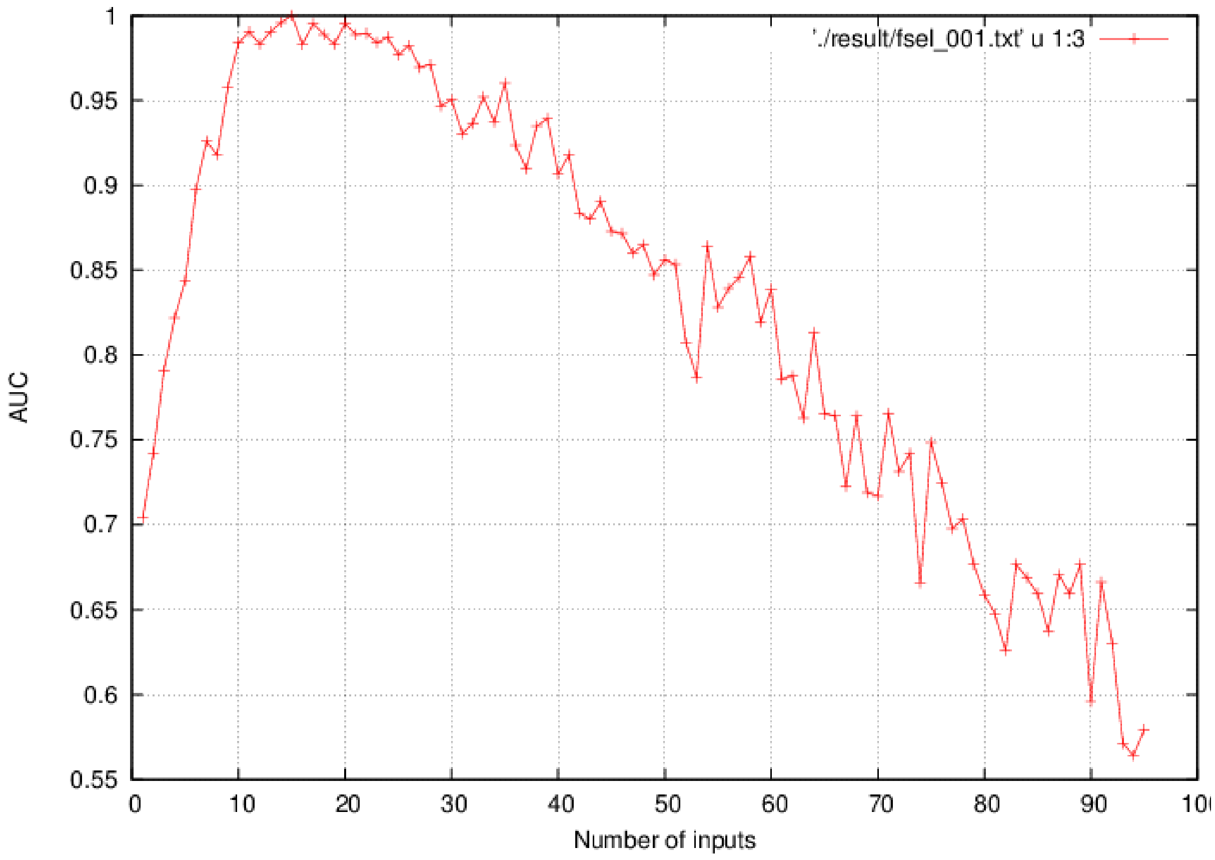
Many measurements will fit perfectly!
Machine learning allows for arbitrarily complex combinations!
Solution: set aside **validation** set of samples

Mattias Ohlsson:
The ”curse of dimensionality”



Feature selection using standard backward elimination and a standard classification model

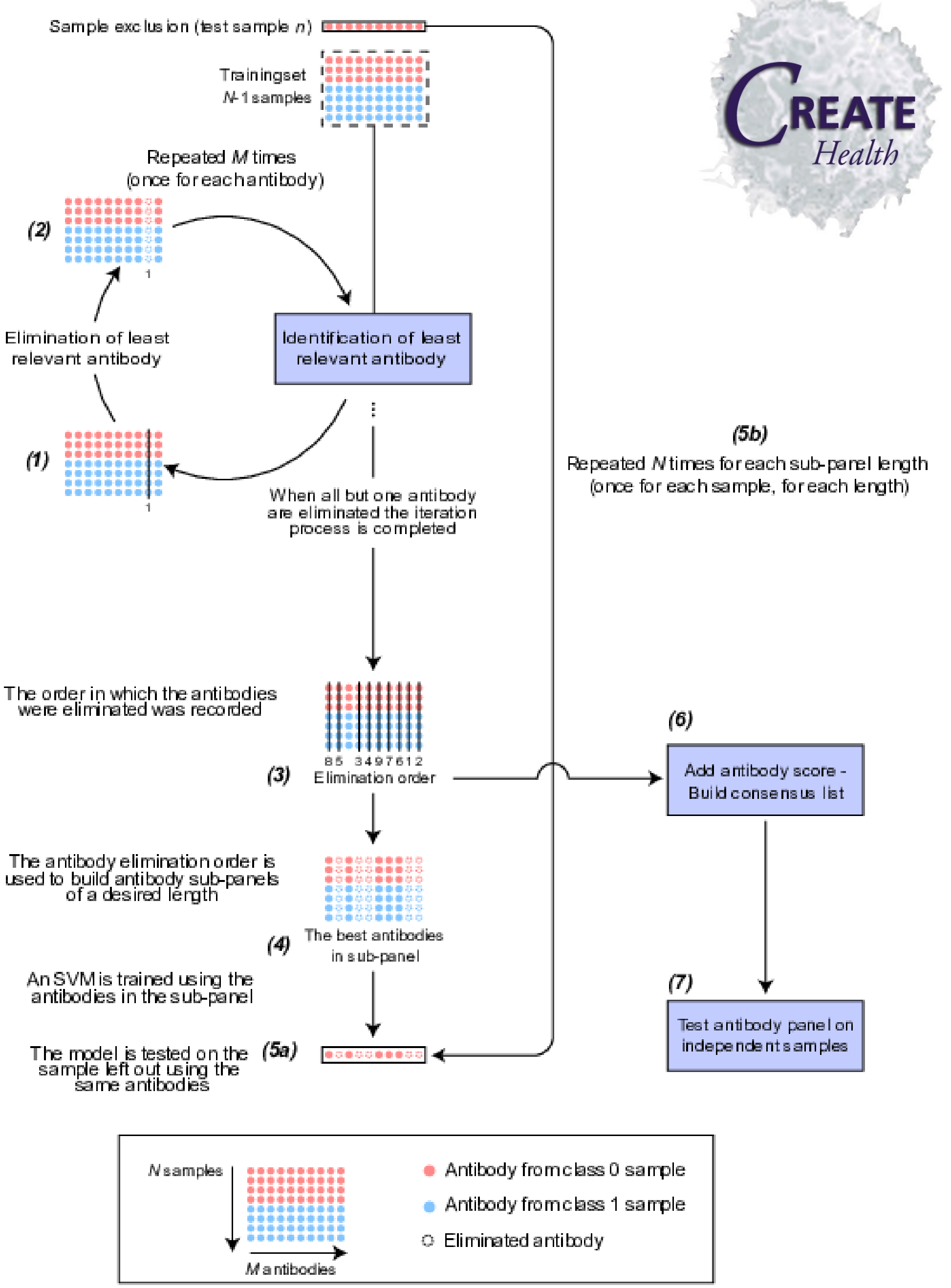
Mattias Ohlsson:
The ”curse of dimensionality”



Feature selection using standard backward elimination and a standard classification model

This is artificial and completely random data!!!

Solution strategy applied to a feature selection problem using support vector machines


Sample exclusion (test sample *n*)
Trainingset
N-1 samples
Repeated *M* times
(once for each antibody)
(2)
1
Identification of least relevant antibody
Elimination of least relevant antibody
(1)
1
When all but one antibody are eliminated the iteration process is completed
(5b)
Repeated *N* times for each sub-panel length
(once for each sample, for each length)
The order in which the antibodies were eliminated was recorded
85 3497612
(3)
Elimination order
(6)
Add antibody score - Build consensus list
The antibody elimination order is used to build antibody sub-panels of a desired length
(4)
The best antibodies in sub-panel
An SVM is trained using the antibodies in the sub-panel
(7)
Test antibody panel on independent samples
The model is tested on the sample left out using the same antibodies
(5a)
N samples
M antibodies
Antibody from class 0 sample
Antibody from class 1 sample
Eliminated antibody

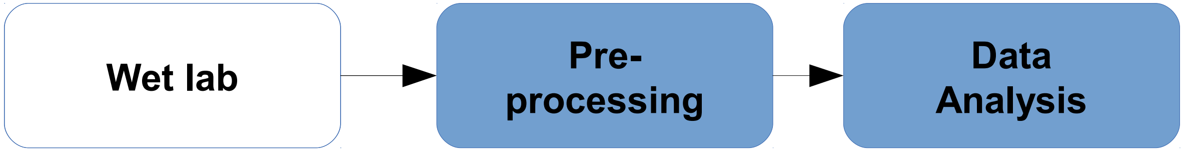
A. Carlsson et al. *PNAS* **108,** (2011)

Wet lab: *cannot be redone*

(all up to scanning)

Data analysis: *Sample annotations used*

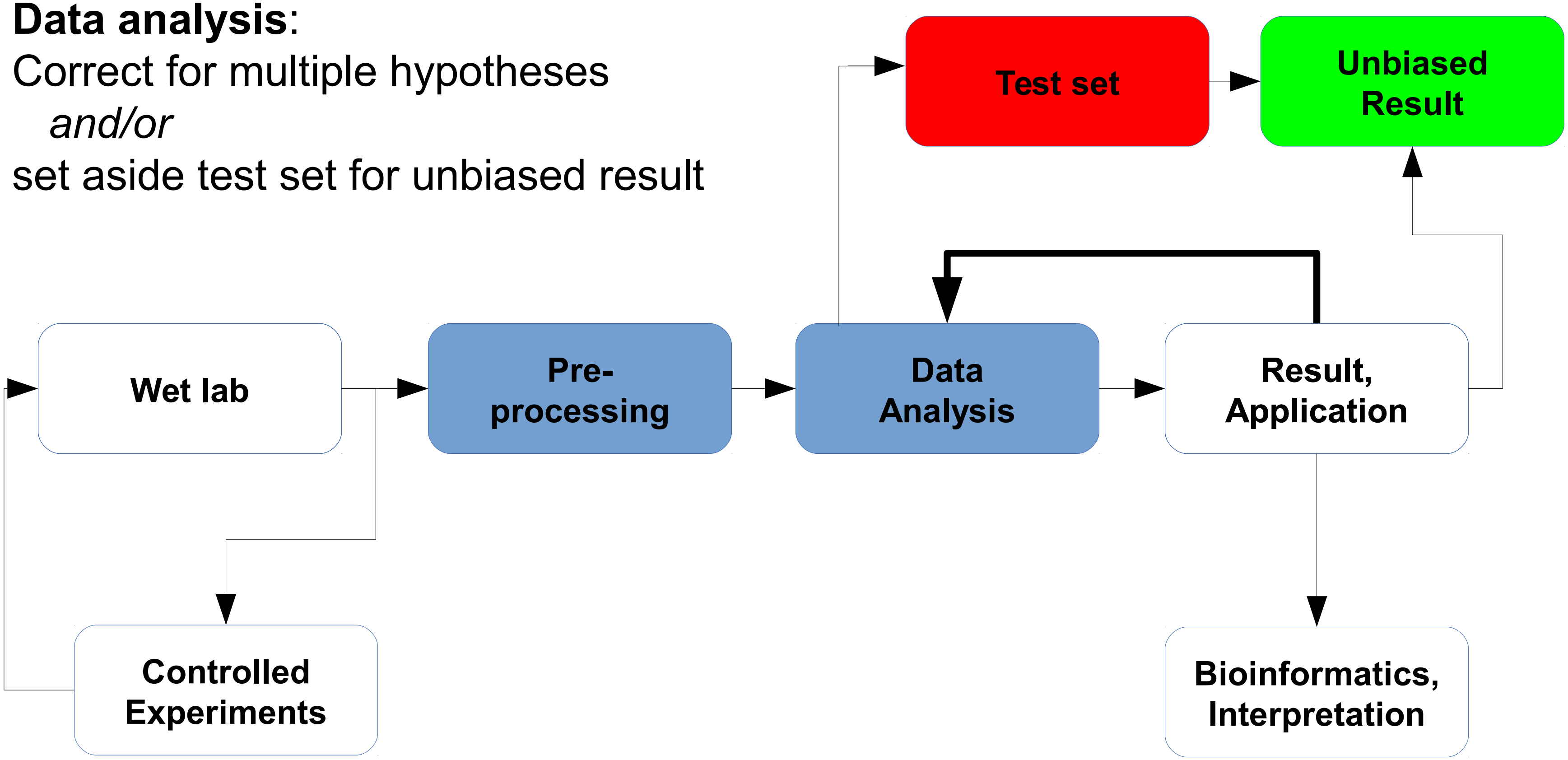
E.g. supervised feature selection
Biomarker search, profiles, classifiers



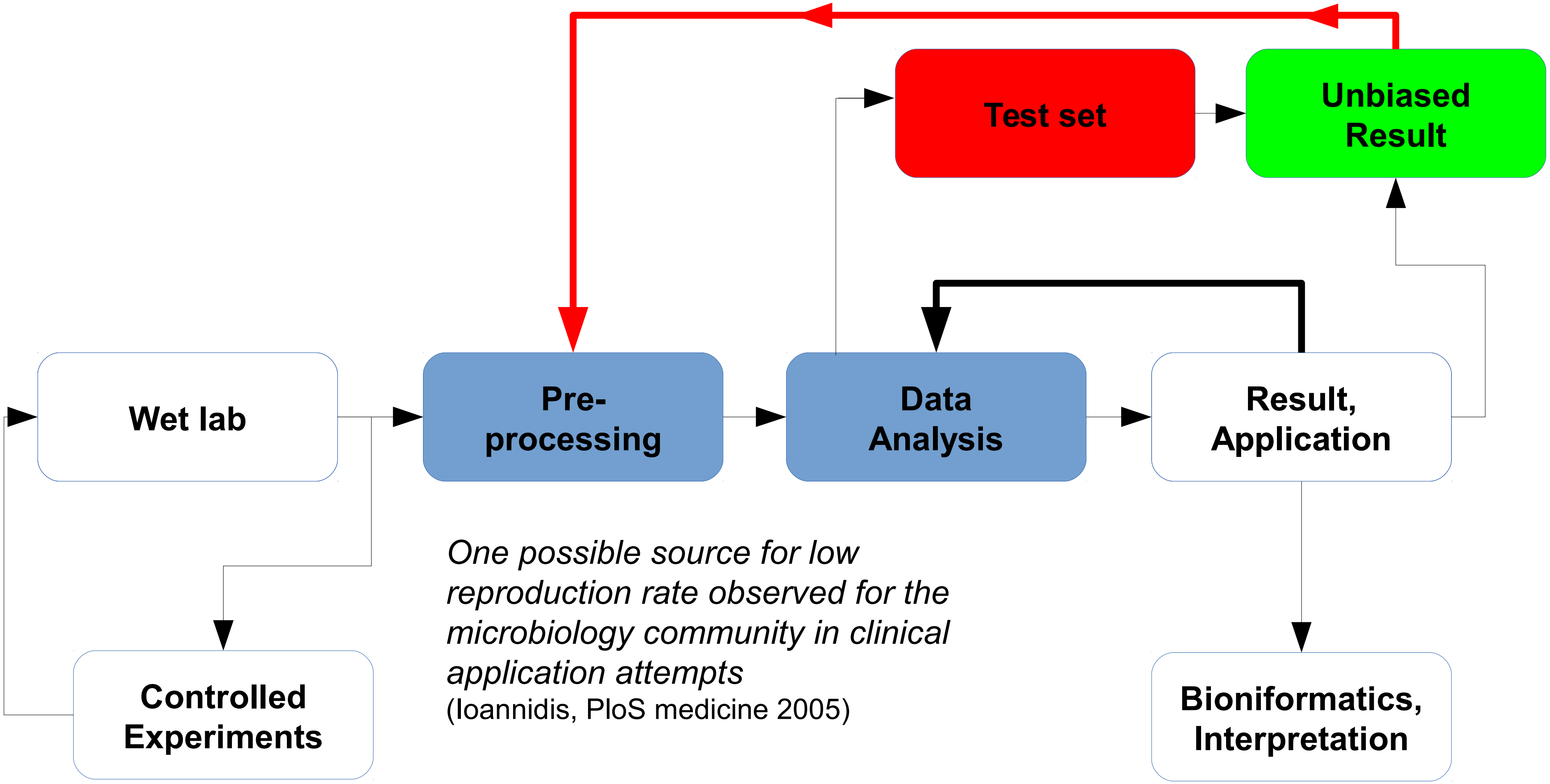
Pre-processing:
Independent of sample annotations

Quality control
Correction for technical effects (e.g. slide-to-slide effects)
Noise reduction (filter low-variance reporters)

Data analysis:
Correct for multiple hypotheses
and/or
set aside test set for unbiased result



If anything before test set selection is reconsidered, the final result is no longer unbiased



One possible source for low reproduction rate observed for the microbiology community in clinical application attempts
(Ioannidis, PloS medicine 2005)

THE NEW YORKER
“THE TRUTH WEARS OFF”
December 13, 2010

The Economist

October 19th–25th 2013

Economist.com

Washington’s lawyer surplus

How to do a nuclear deal with Iran

Investment tips from Nobel economists

Junk bonds are back

The meaning of Sachin Tendulkar

HOW SCIENCE GOES WRONG

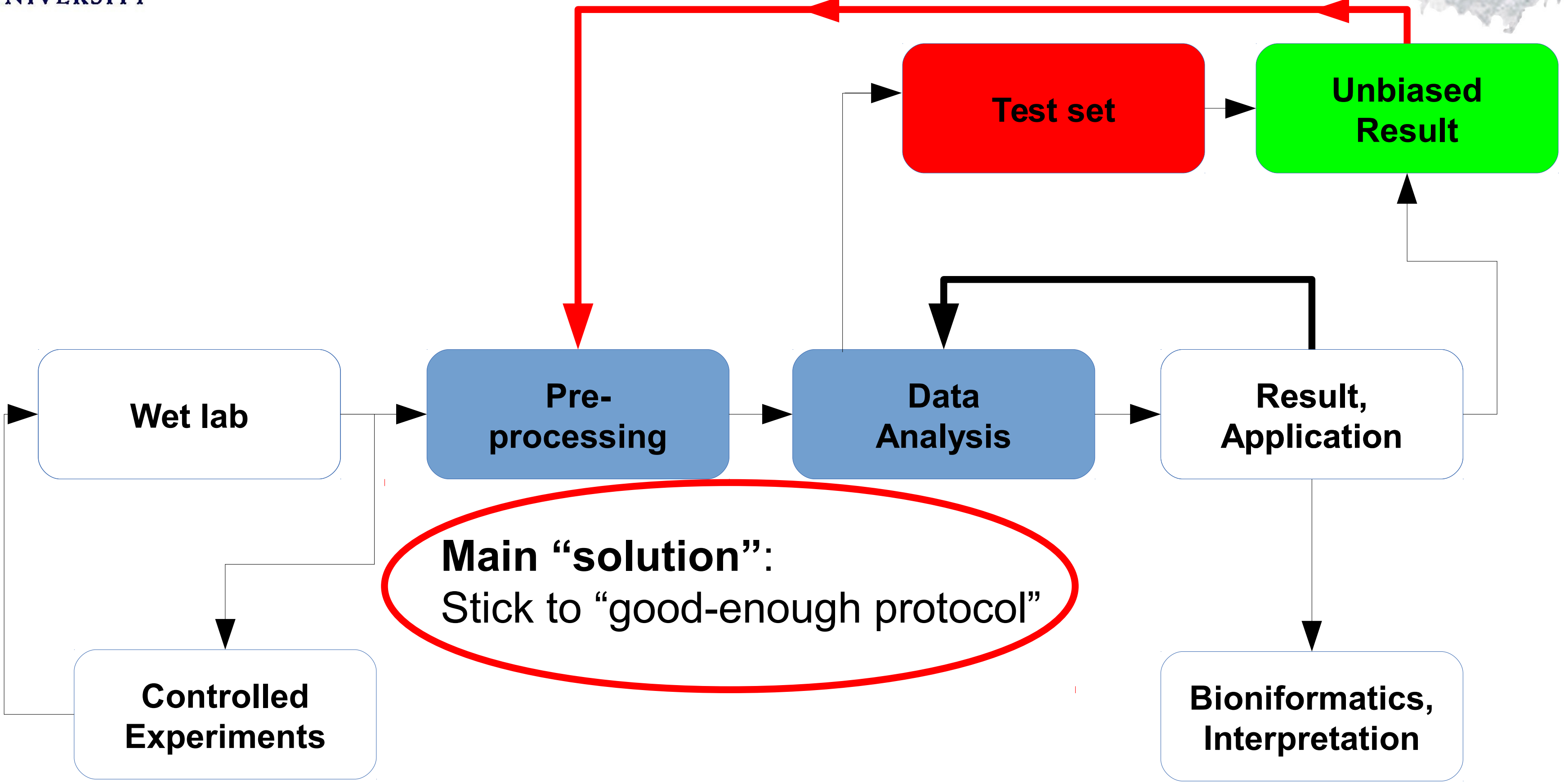
Mainly:
*Ioannidis, J: **“Why Most Published Research Findings Are False”** PloS Medicine, 2 e124 (2005).*

Community effects:

- *Lack of funding / career paths for reproducing experiments*
- *Lack of publications of negative results*

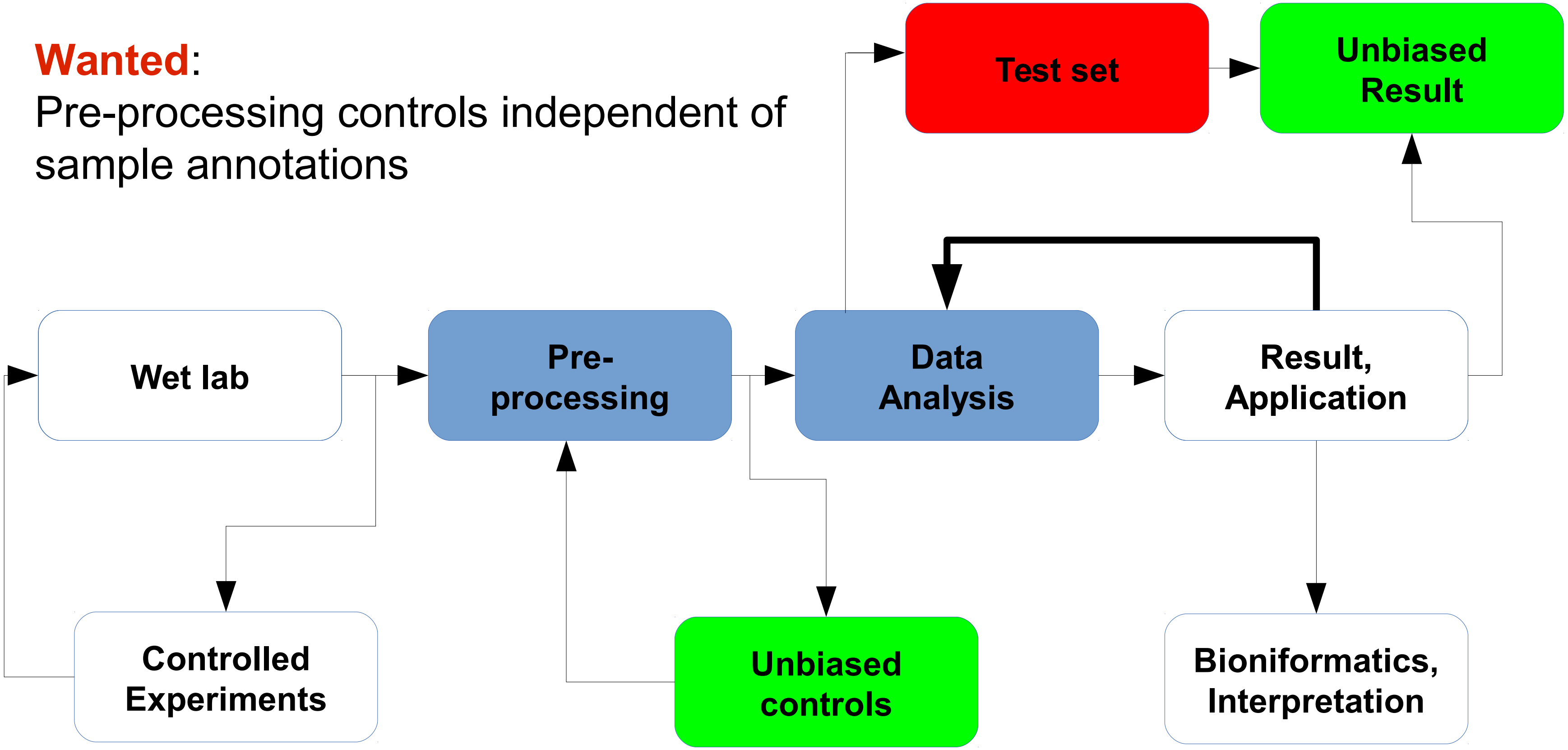
Also:

- ***Re-analysis of data to find good p-value***



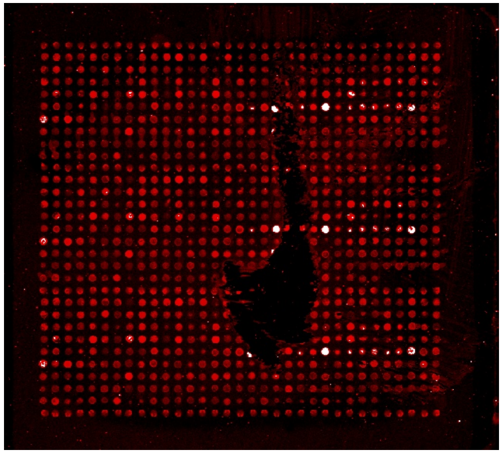
Main “solution”:
Stick to “good-enough protocol”

Wanted:
Pre-processing controls independent of sample annotations



Current project:
- More use of replicate information
- **Validated Imputation**

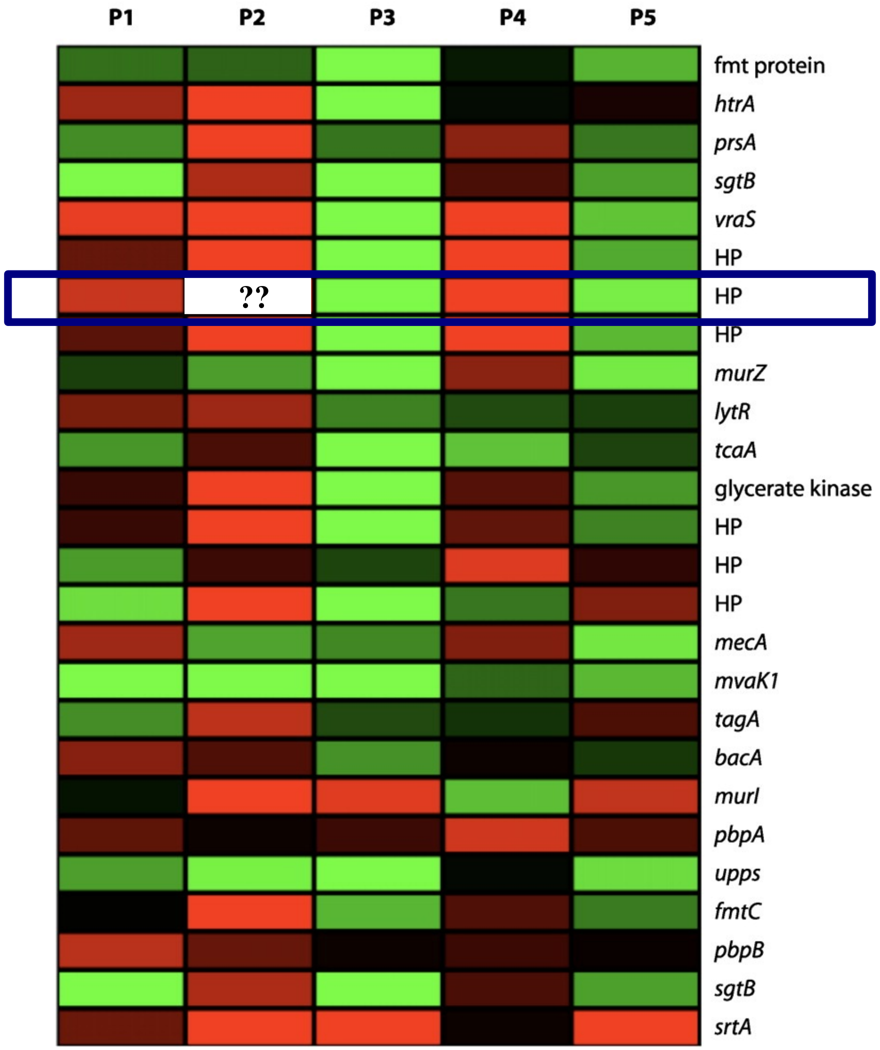
Imputation: Missing value estimation



There can be missing values

Many analysis tools require complete data matrices

Simple: estimate with row average
Better: use data correlations



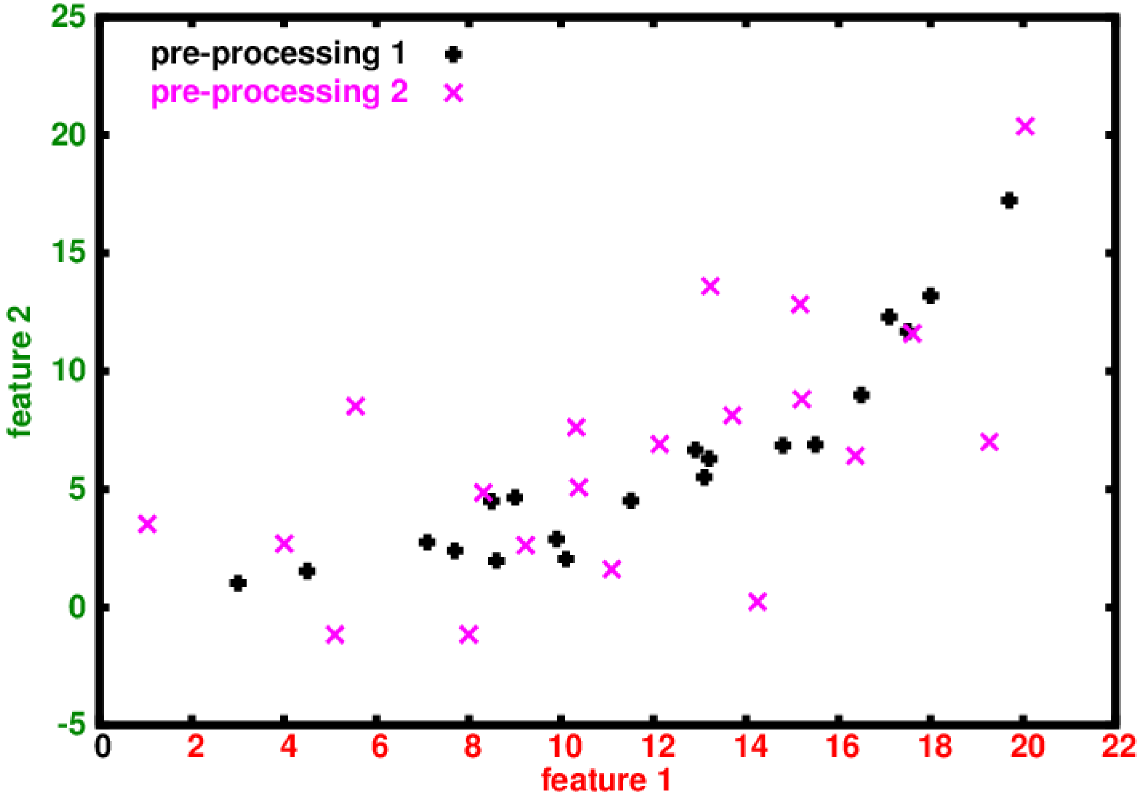
Shown with **Validated Imputation:**

Artificially remove data. Compare imputed value with known answer.

Standard use:
Use bench-mark preprocessed data to test imputation algorithms

Our use:

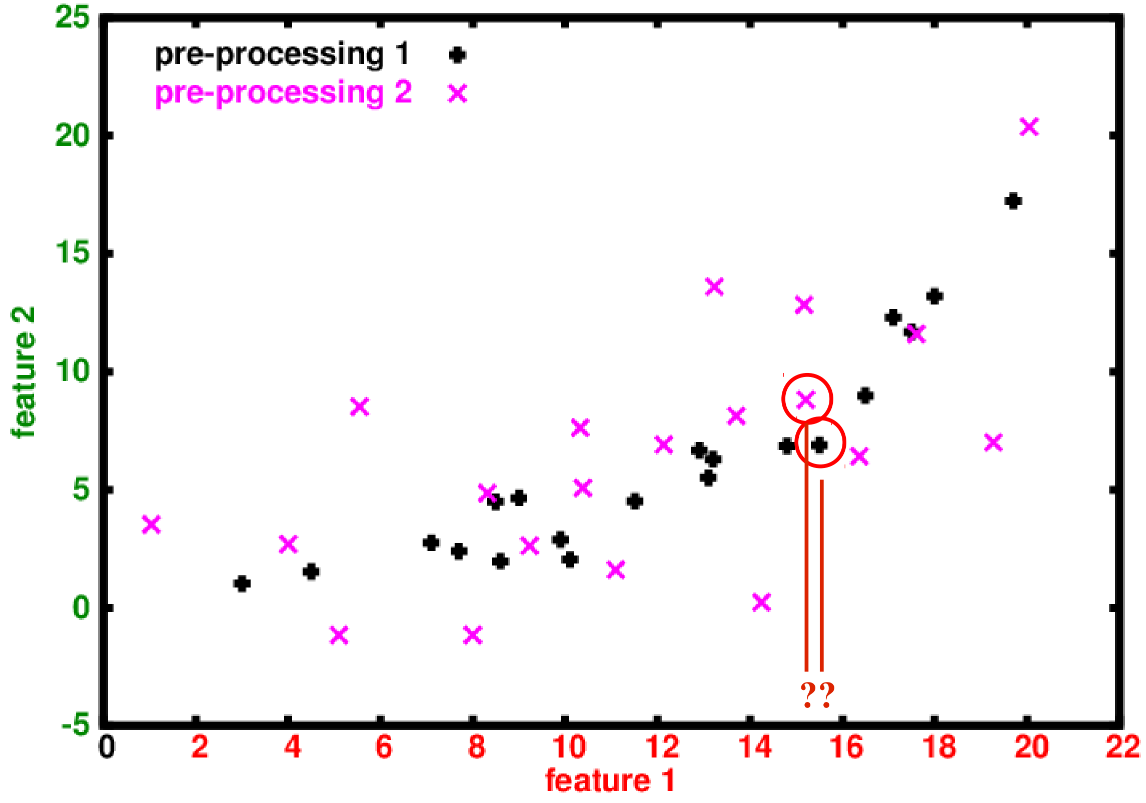
Use bench-mark imputation to test pre-processing



Same data, two pre-processings:
***Black**: good noise reduction*
***Purple**: very noisy*

Our use:

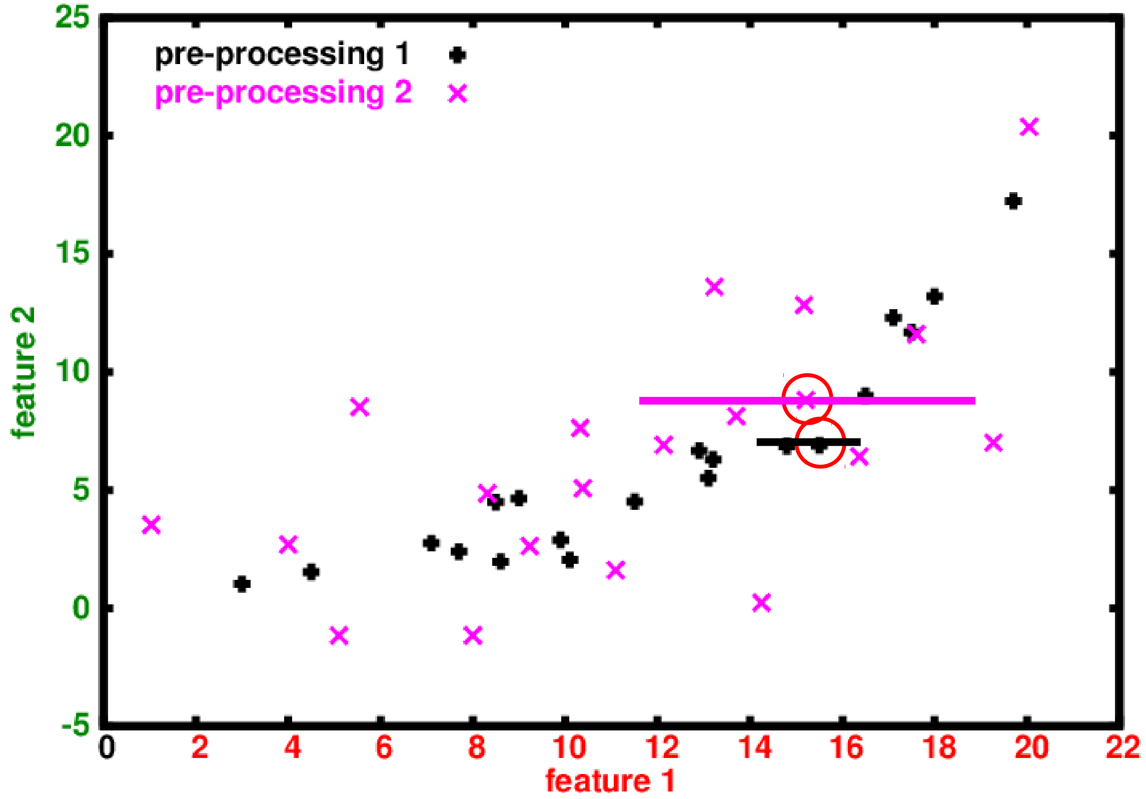
Use bench-mark imputation to test pre-processing



Artificially remove ***feature 1*** *value for a sample (in both pre-processings)*

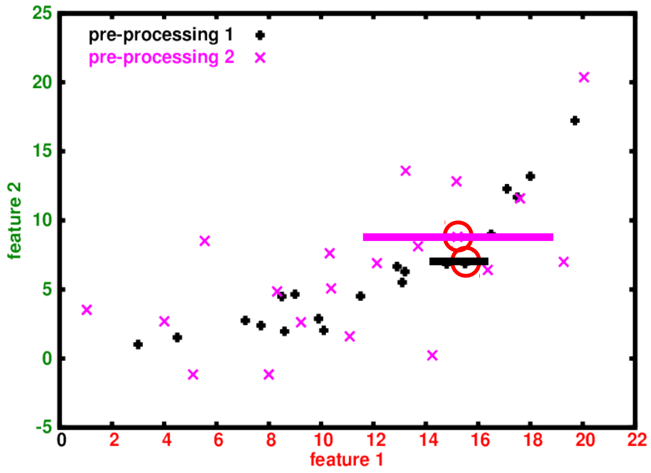
Our use:

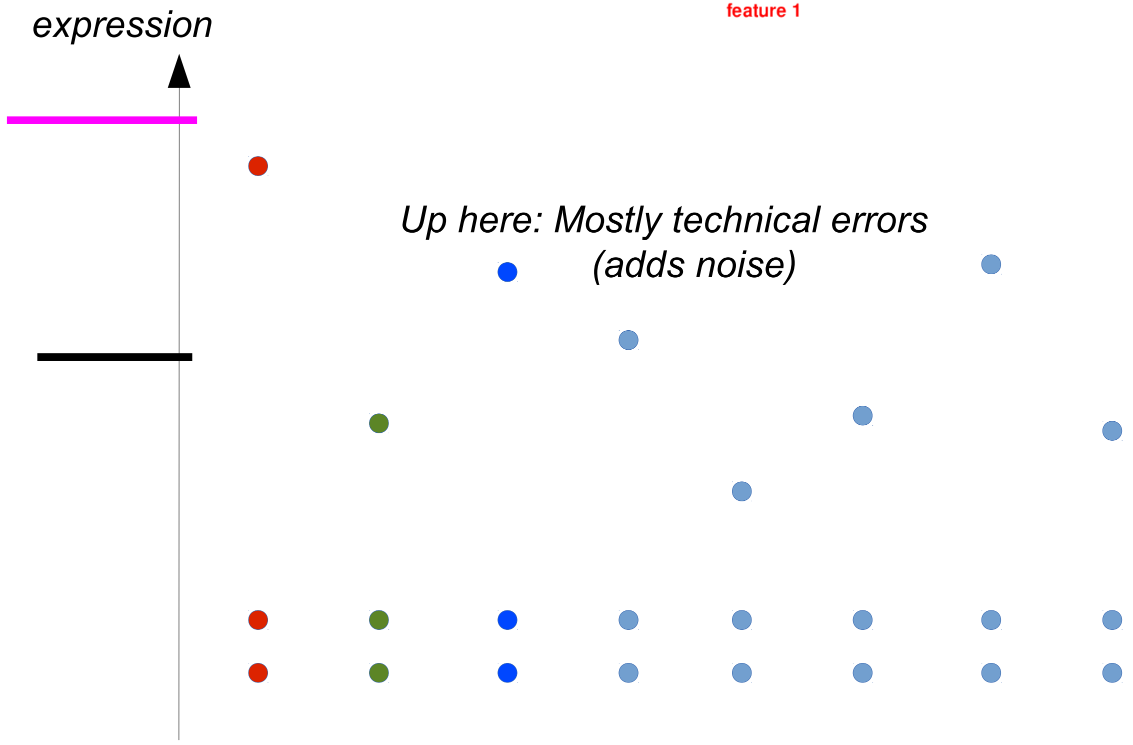
Use bench-mark imputation to test pre-processing



*Use **feature 2** (and others) to estimate missing value*
*Good (**black**) pre-processing: somewhere close*
*Noisy (**purple**) pre-processing: almost anywhere*

Example:
Outliers among triplicates or not?





Purple outlier cutoff

Black outlier cutoff

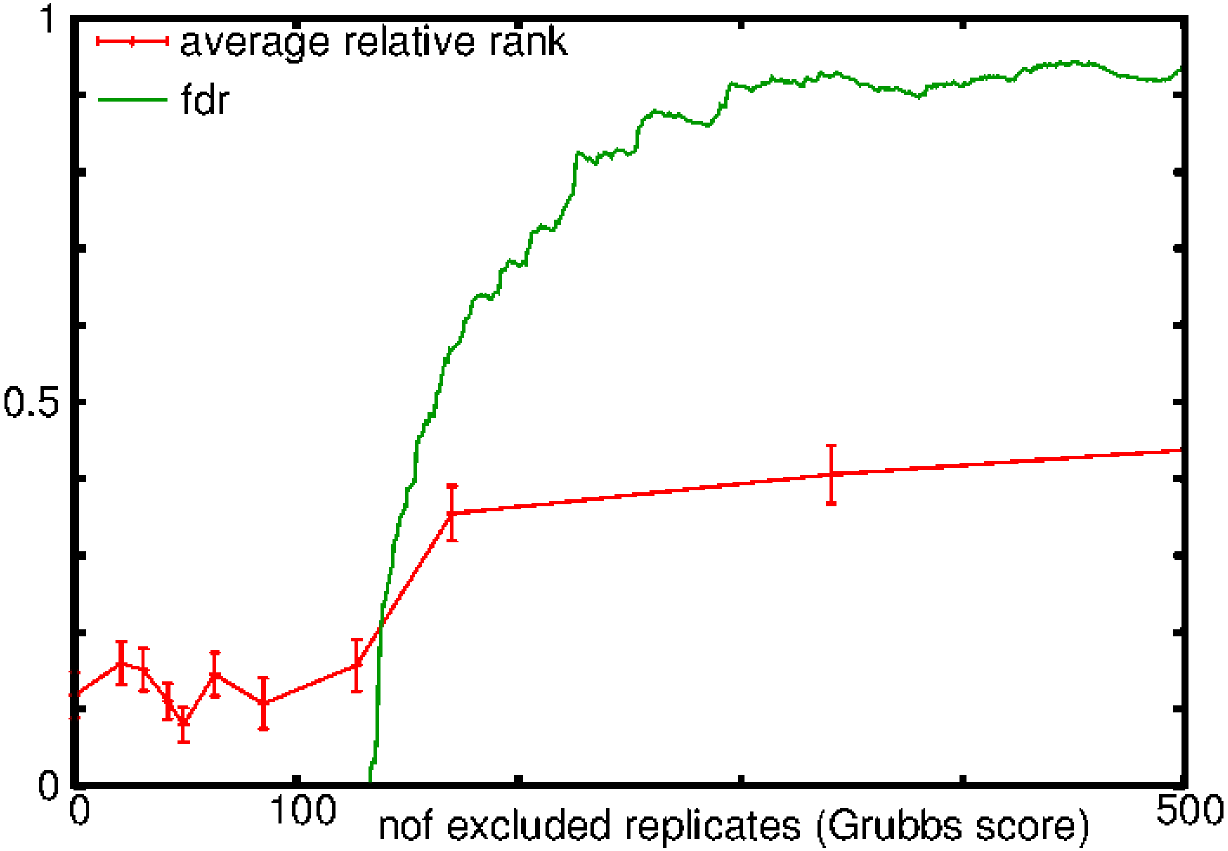
Up here: Mostly technical errors (adds noise)

Sample s:

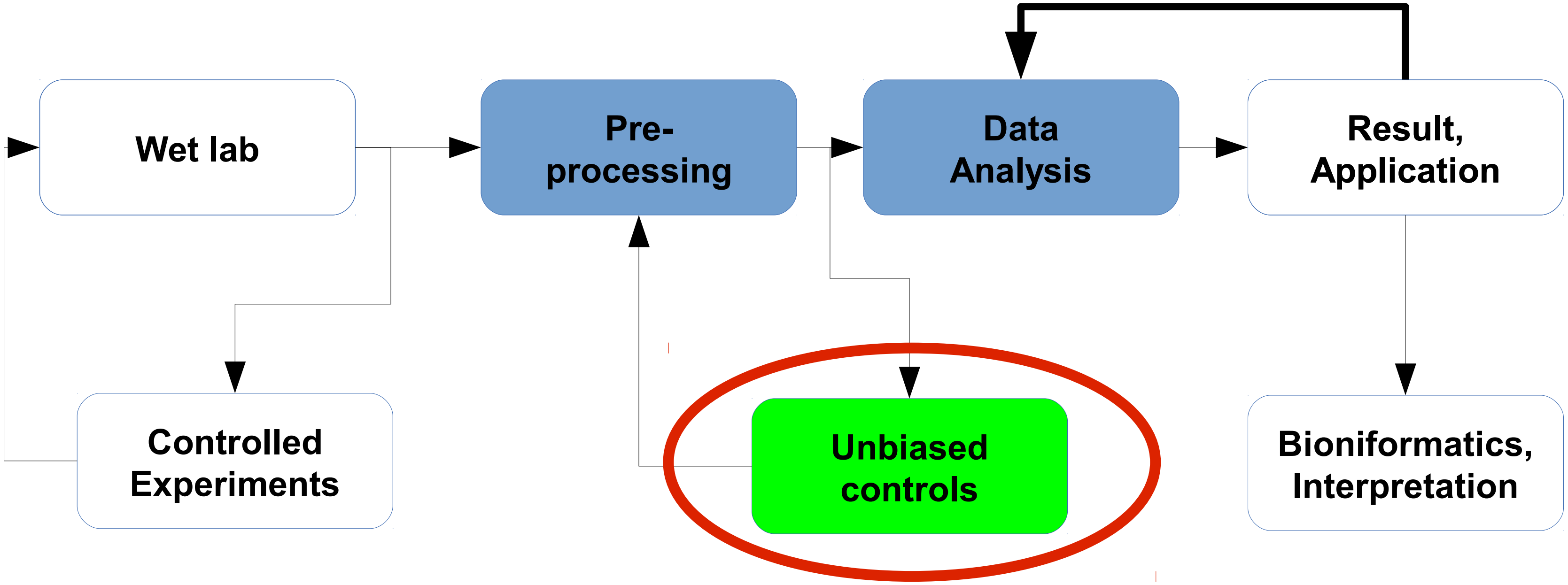
feat.1 ***feat.2*** ***feat.3*** *...*

Outlier detection among spot replicates:

Statistical model and Validated Imputation agree!



The microarray community is in need of pre-processing controls ignorant of sample annotations



Success requires close collaboration between experimental and computational expertise

While co-developing pre-processing protocols for the protein antibody array, we have found promising methods of high relevance for many microarray platforms.