SciLifeLab

# Bayesian methods for the interpretation of label-free proteomics data

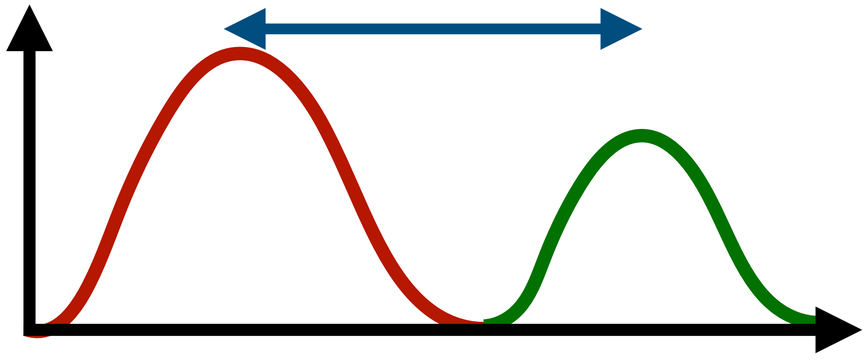
Lukas Käll
Royal Institute of Technology - KTH
School of Biotechnology
Stockholm, Sweden

KTH VETENSKAP OCH KONST

http://percolator.ms
http://kaell.org

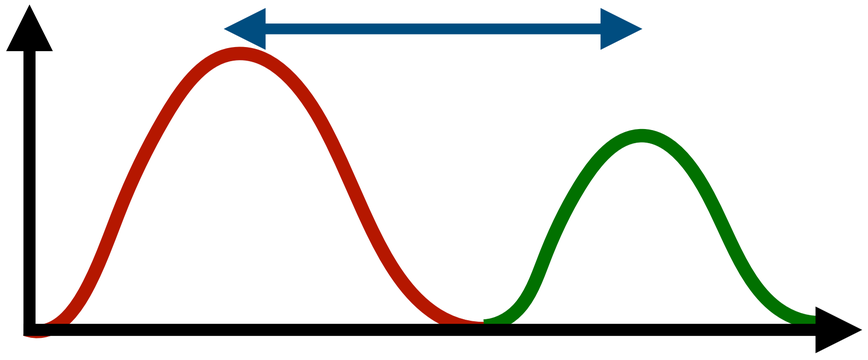
# Outline

1. Background on label-free quantification
2. Combined identification and quantification error rates — Triqler
3. Clustering and Quantification MS/MS data — Quandenser

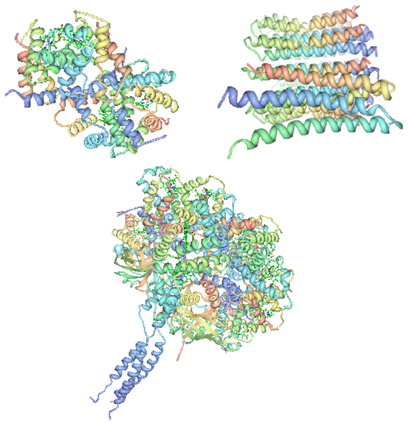


# Outline

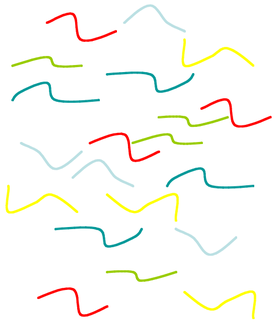
1. Background on label-free quantification
2. Combined identification and quantification error rates — Triqler
3. Clustering and Quantification MS/MS data — Quandenser



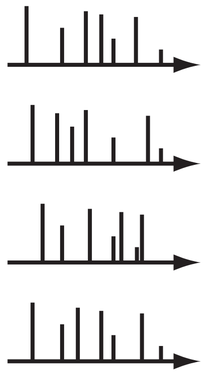
# In shotgun proteomics the analytes are peptides, not proteins



Digestion



Detection



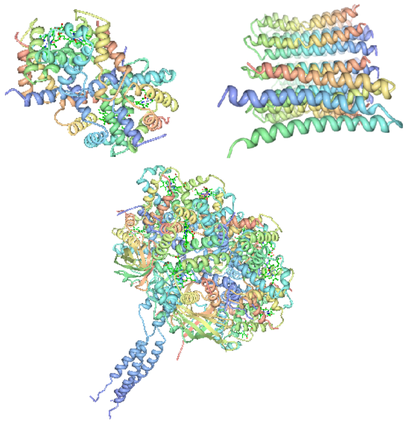
IAMAPEPTIDER 38.2
MGREATMATCHK 12.5
ATRYIDENTIFYMEK 21.7
LMAKEMYDAYR 63.1

Spectral matching and abundance integration

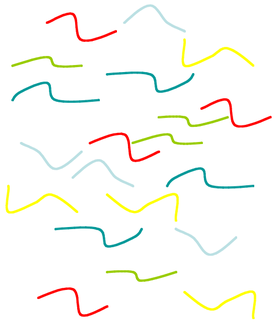
GLB4_LUMTE 2.3
COX1_LUMTE 1.5
ATP6_LUMTE 18.2

Protein inference and summarization

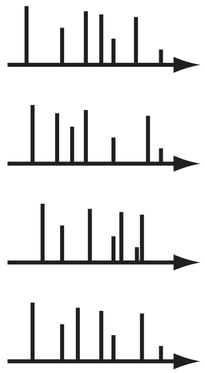
# In shotgun proteomics the analytes are peptides, not proteins



Digestion



Detection



IAMAPEPTIDER 38.2

MGREATMATCHK 12.5

ATRYIDENTIFYMEK 21.7

LMAKEMYDAYR 63.1

Spectral matching and abundance integration

Protein inference and summarization

GLB4_LUMTE 2.3

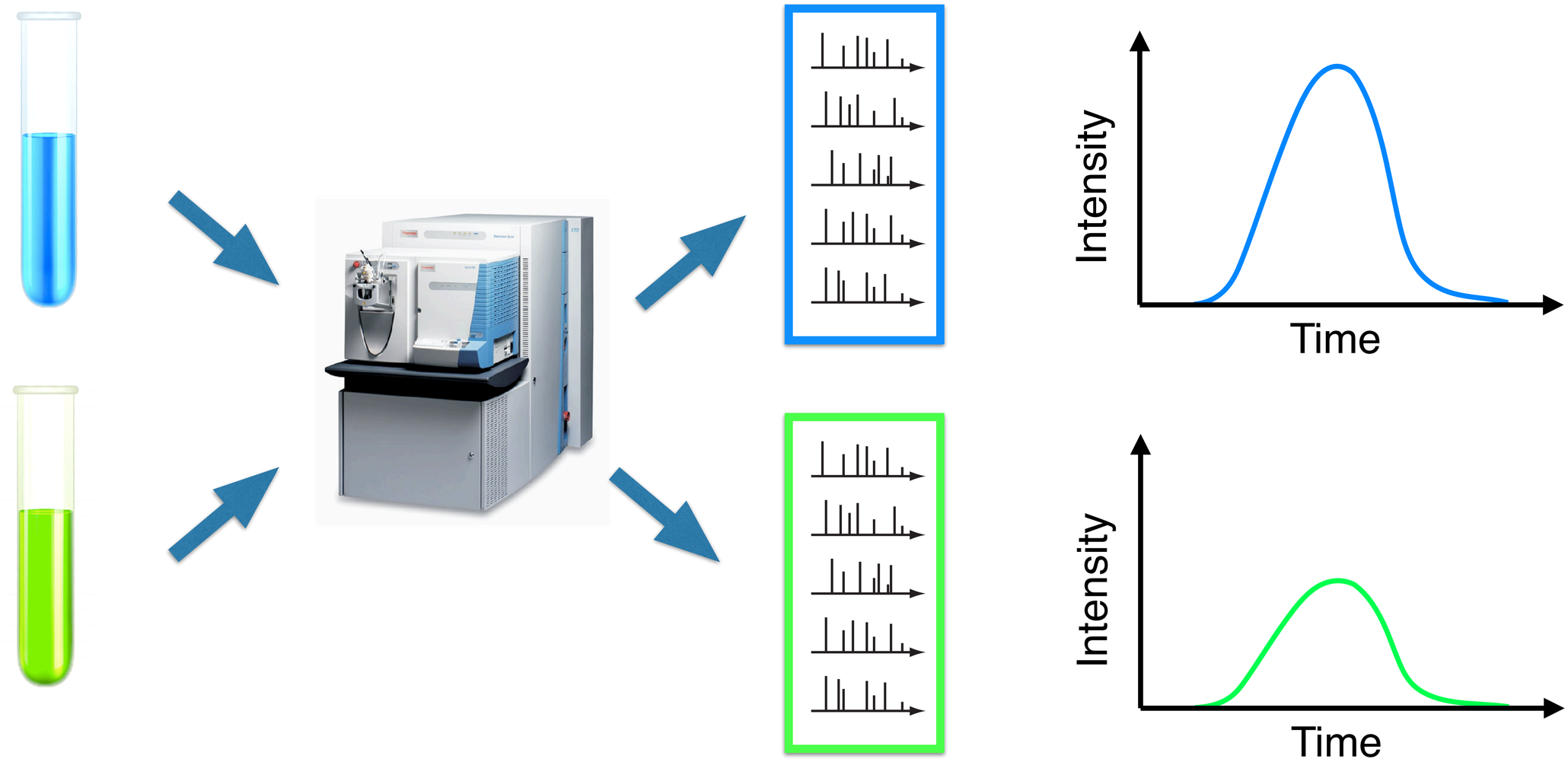
COX1_LUMTE 1.5

ATP6_LUMTE 18.2

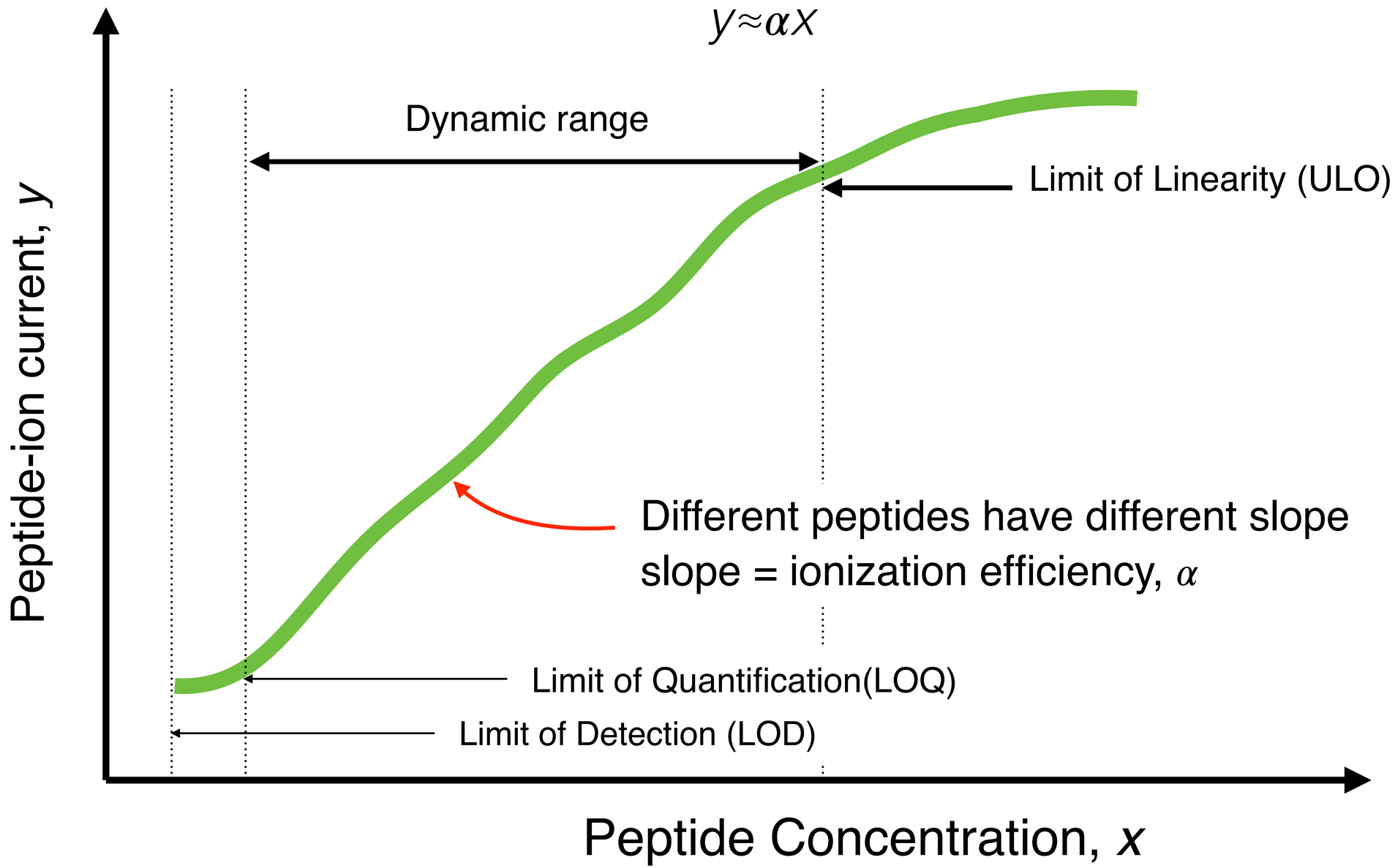
Protein Summarization:

The procedure that derives protein relative concentration from its peptides' abundances

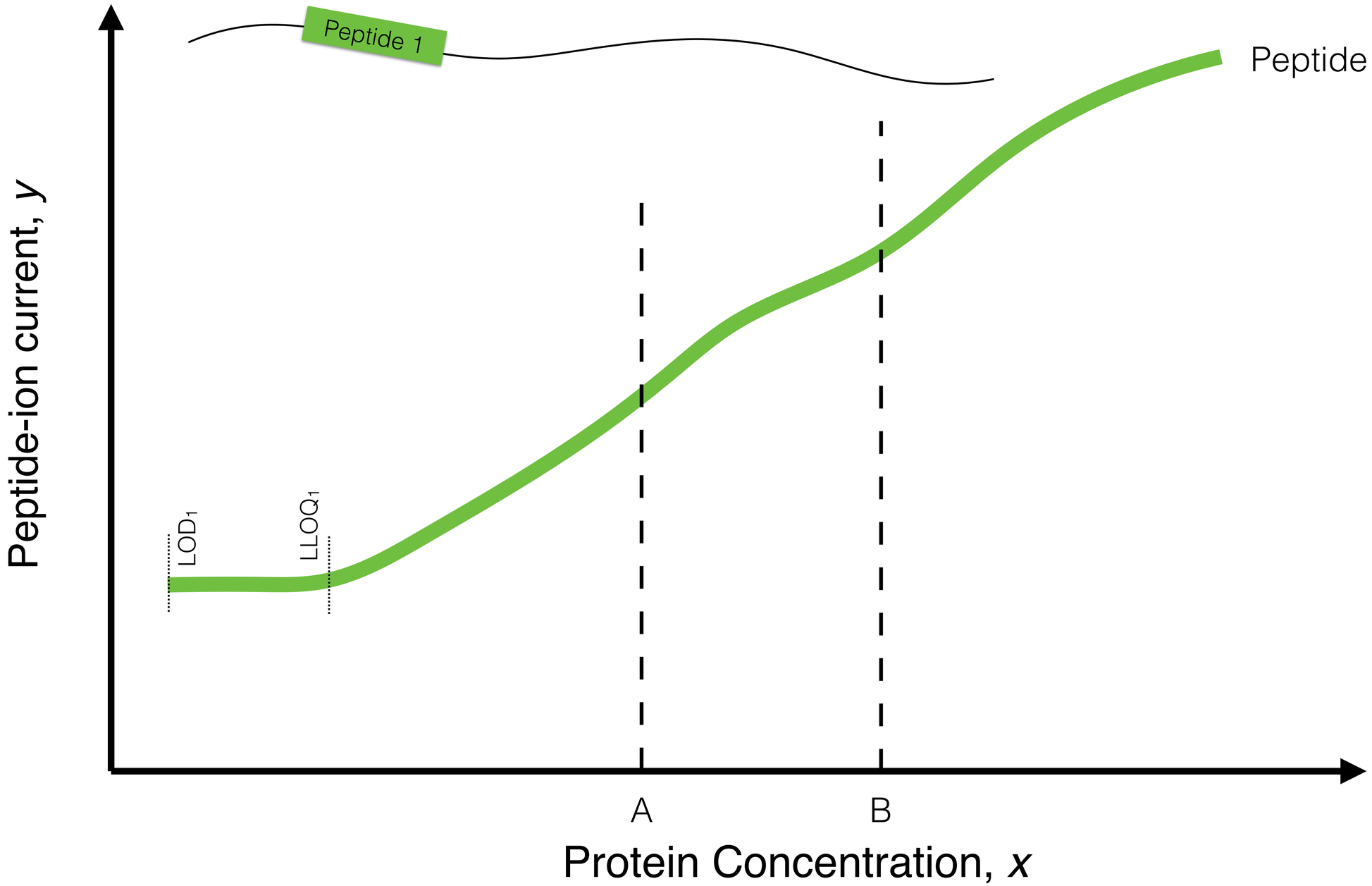
# Label-free quantification: Separate MS analysis of each sample



# A peptide's ion-current is linearly proportional to its concentration

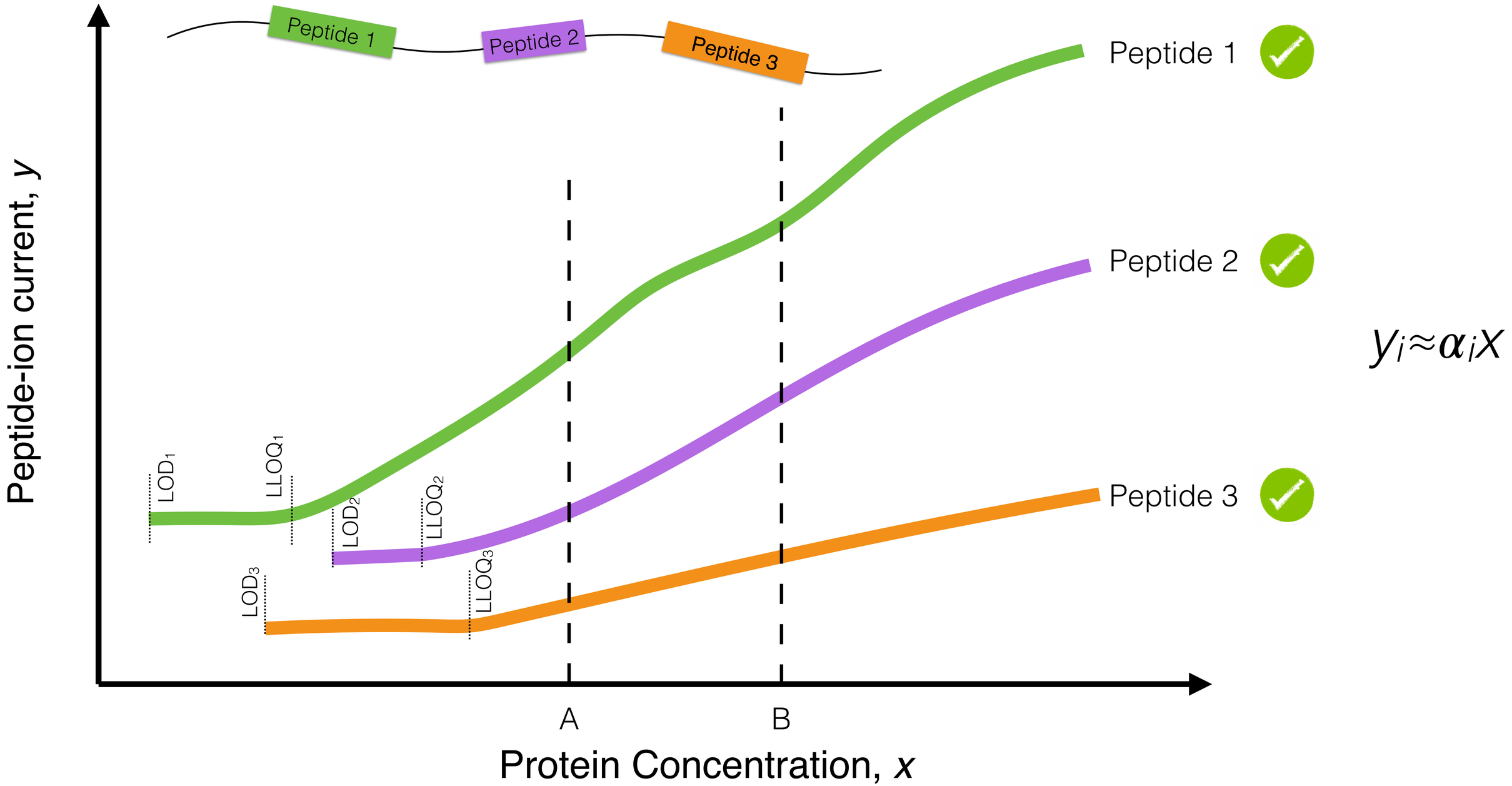


# A protein's peptides' ion-currents are proportional to the protein's concentration

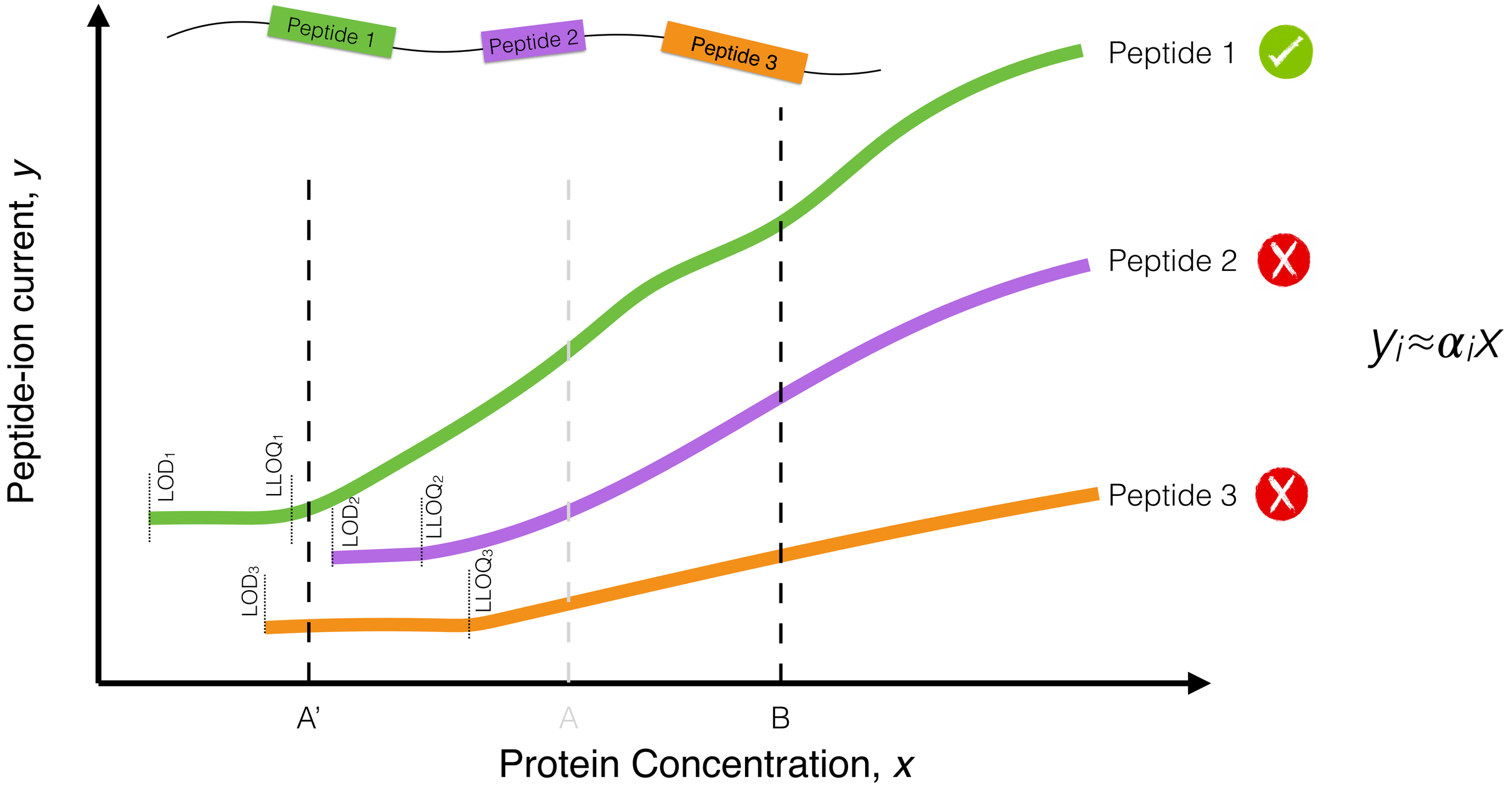



$$y_i \approx \alpha_i X$$

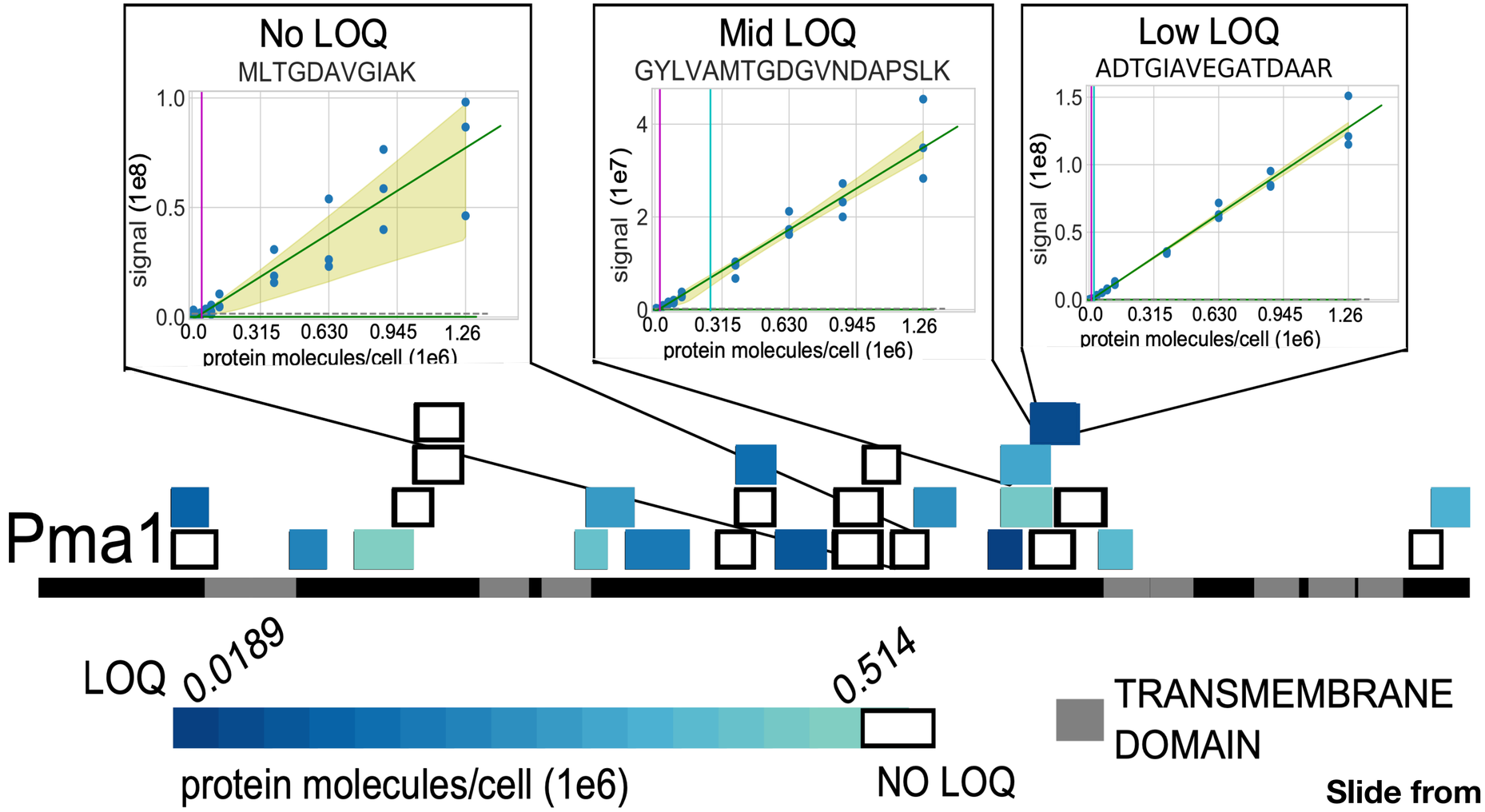
# A protein's peptides' ion-currents are proportional to the protein's concentration



# A protein's peptides' ion-currents are proportional to the protein's concentration



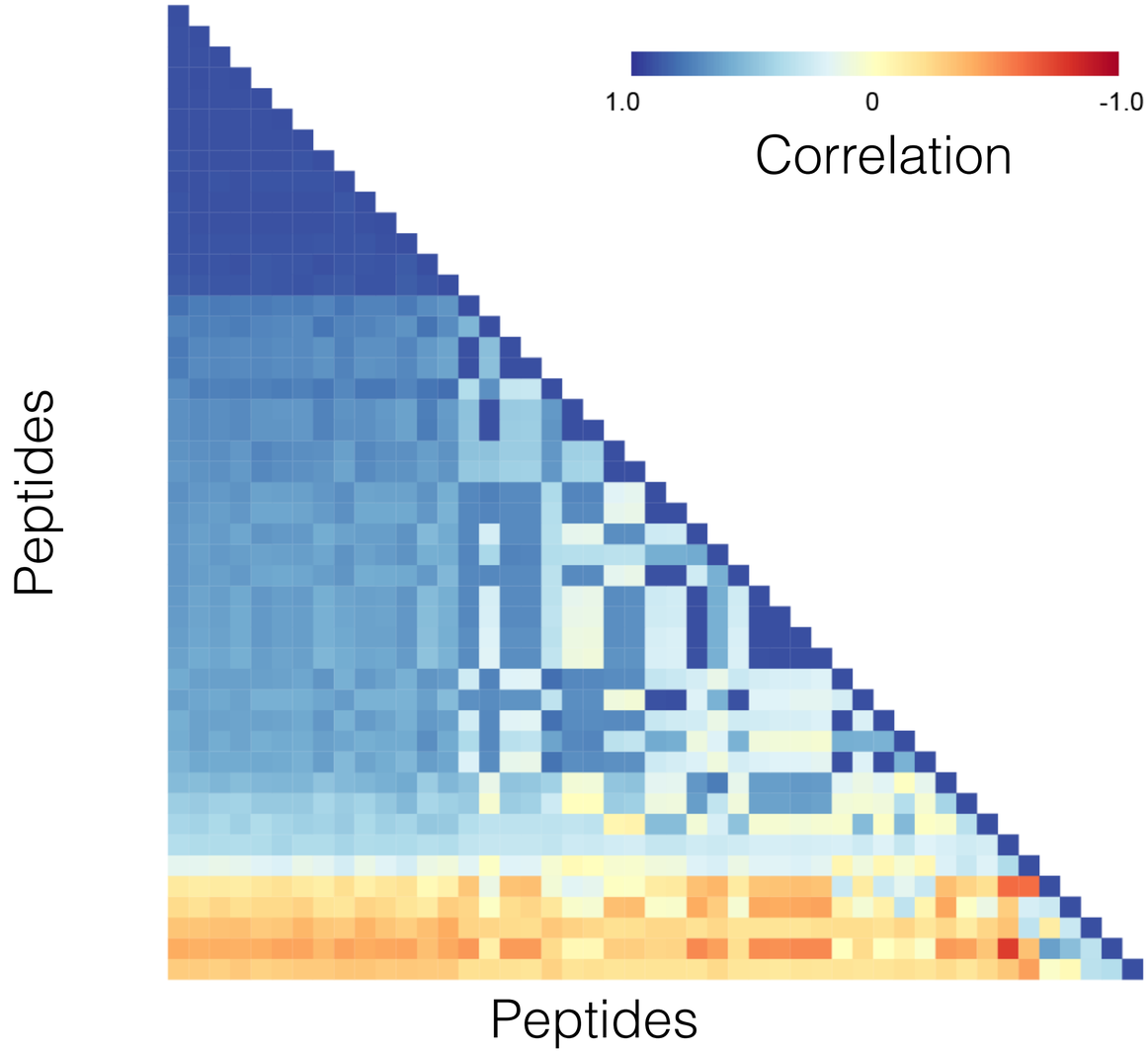
# Protein's quantitative accuracy depends on which peptides you observe




**Slide from Lindsay Pino, UoW**

**[bioRxiv, https://doi.org/10.1101/719179]**

# The constituent peptides' ion-currents of a protein are not all co-varying



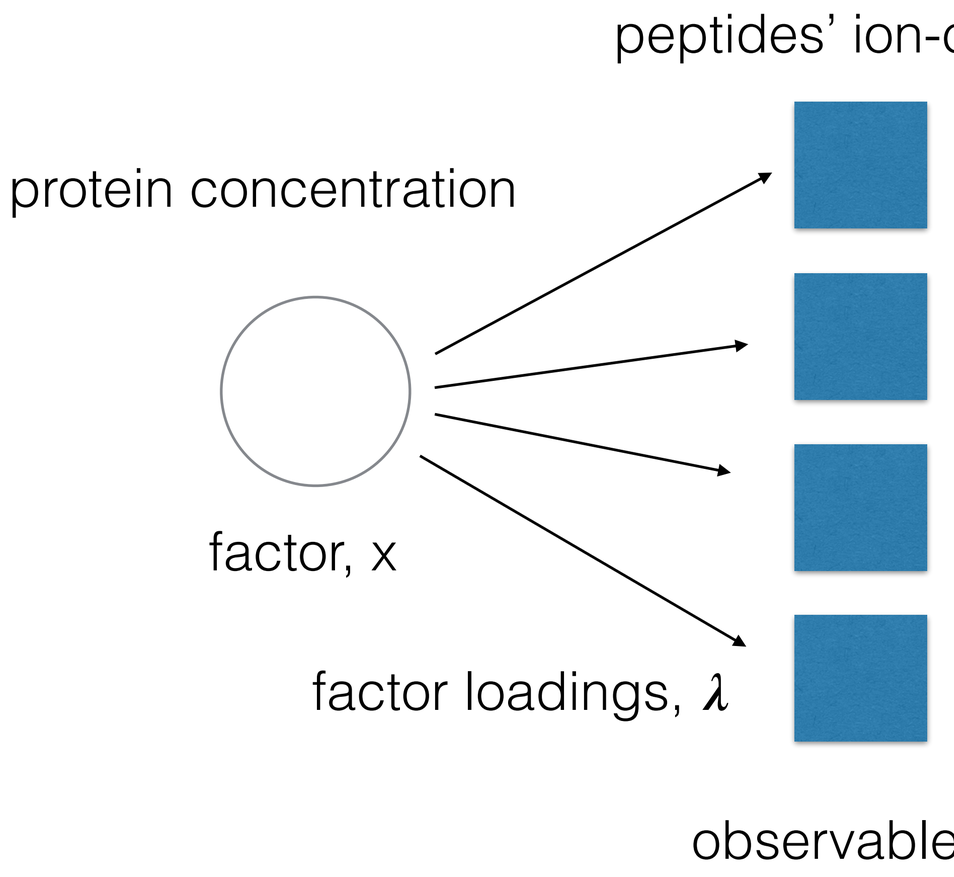

**Covariance matrix of the XICs of 47 peptides from the same protein PYGM_RABIT in 4 concentrations**

**Data: iPRG2015**
**6 spiked in proteins in yeast background, 4 concentrations, triplicates**

**[Zhang et al, MCP 16:936]**

# Factor Analysis of peptides' ion-current — Diffacto

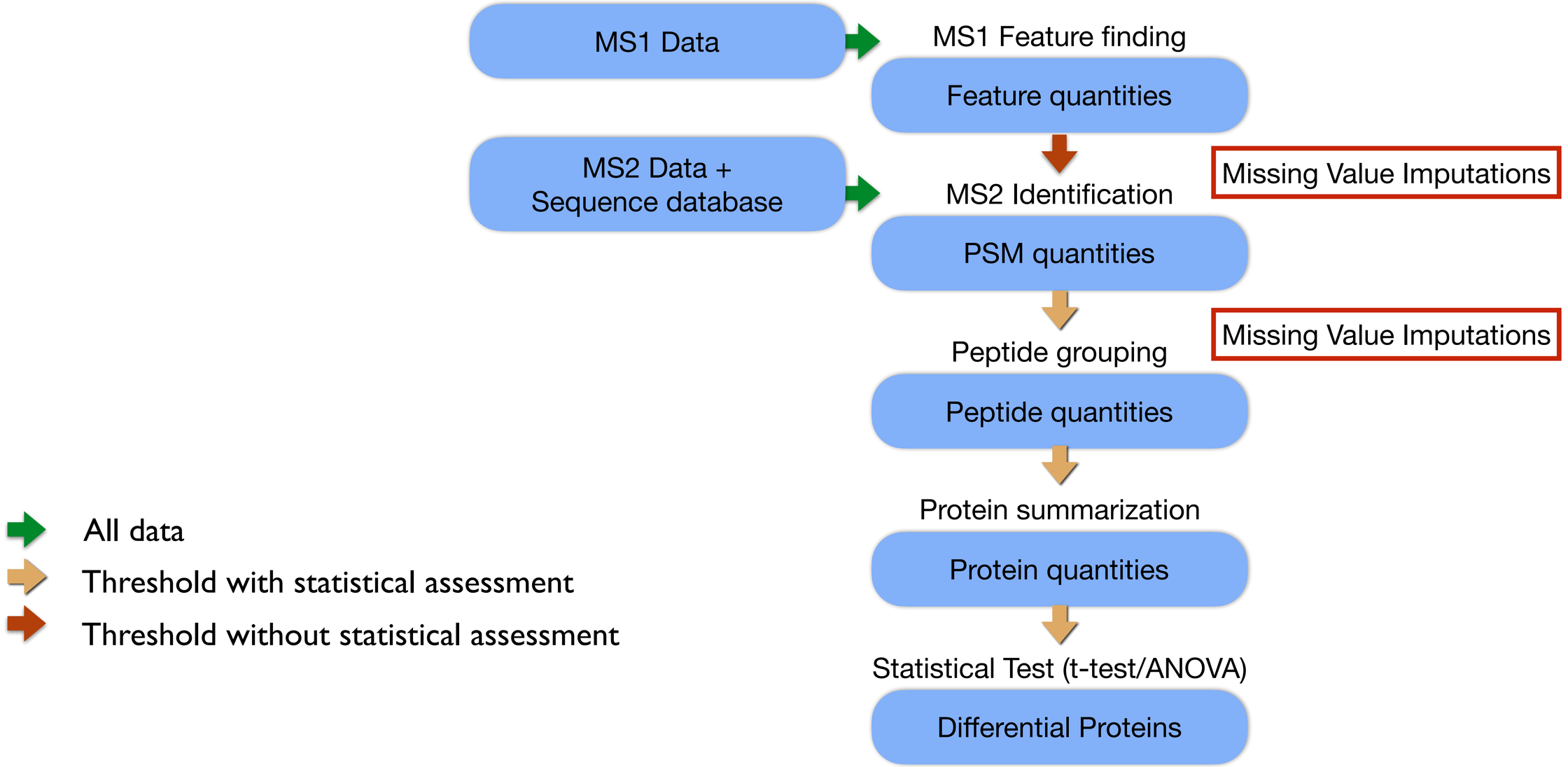
Factor analysis searches for joint variations of observed variables in response to unobserved latent variables




$$\mathbf{y}=\lambda x+\varepsilon$$

**[Zhang et al, MCP 16:936]**

# The quantitative mass spectrometry-based proteomics data processing cascade

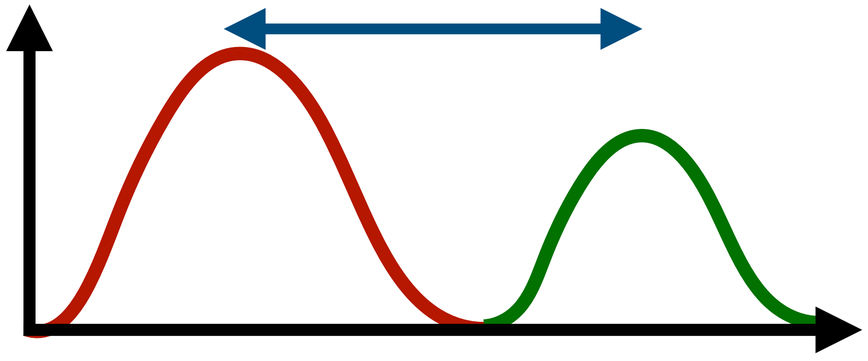


# Outline

1. Background on label-free quantification

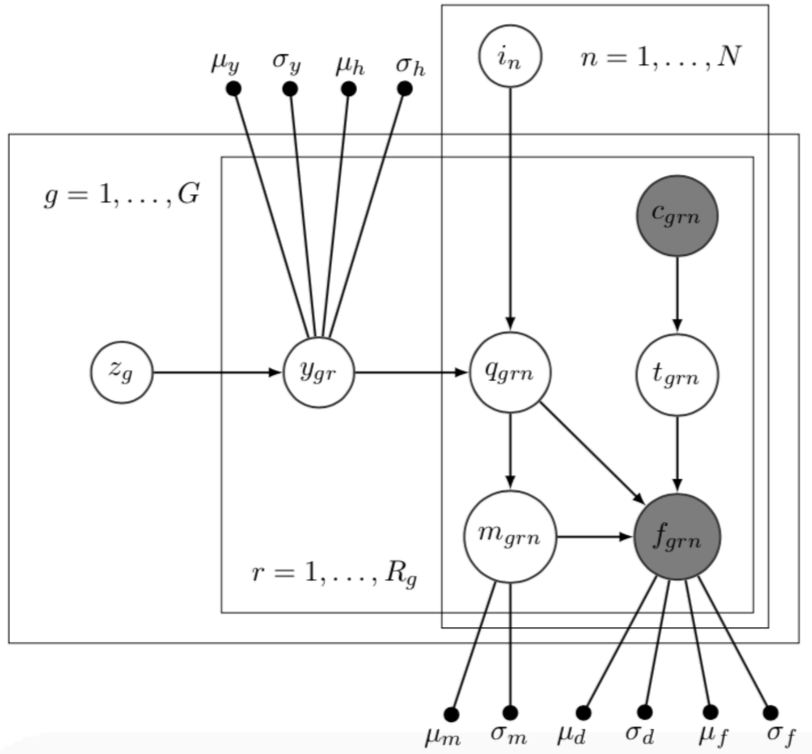
2. Combined identification and quantification error rates — Triqler

3. Clustering and Quantification MS/MS data — Quandenser



# Triqler:

## A unified quantification error model for proteomics




Input: PSMs with score & intensity

Output: proteins with *Pr(diff. exp.)*

Runtime: couple of minutes

Python package: pip install triqler

Paper: The & Käll MCP (2018)

Matthew The

# Why do we need another method?

- Current methods lack proper FDR control
  - Several (arbitrary) thresholds
  - Neglecting error propagation
  - Rely on missing value imputations
- Often only diff. exp. at p-value cutoff

# What *is* proper FDR control in protein quantification?

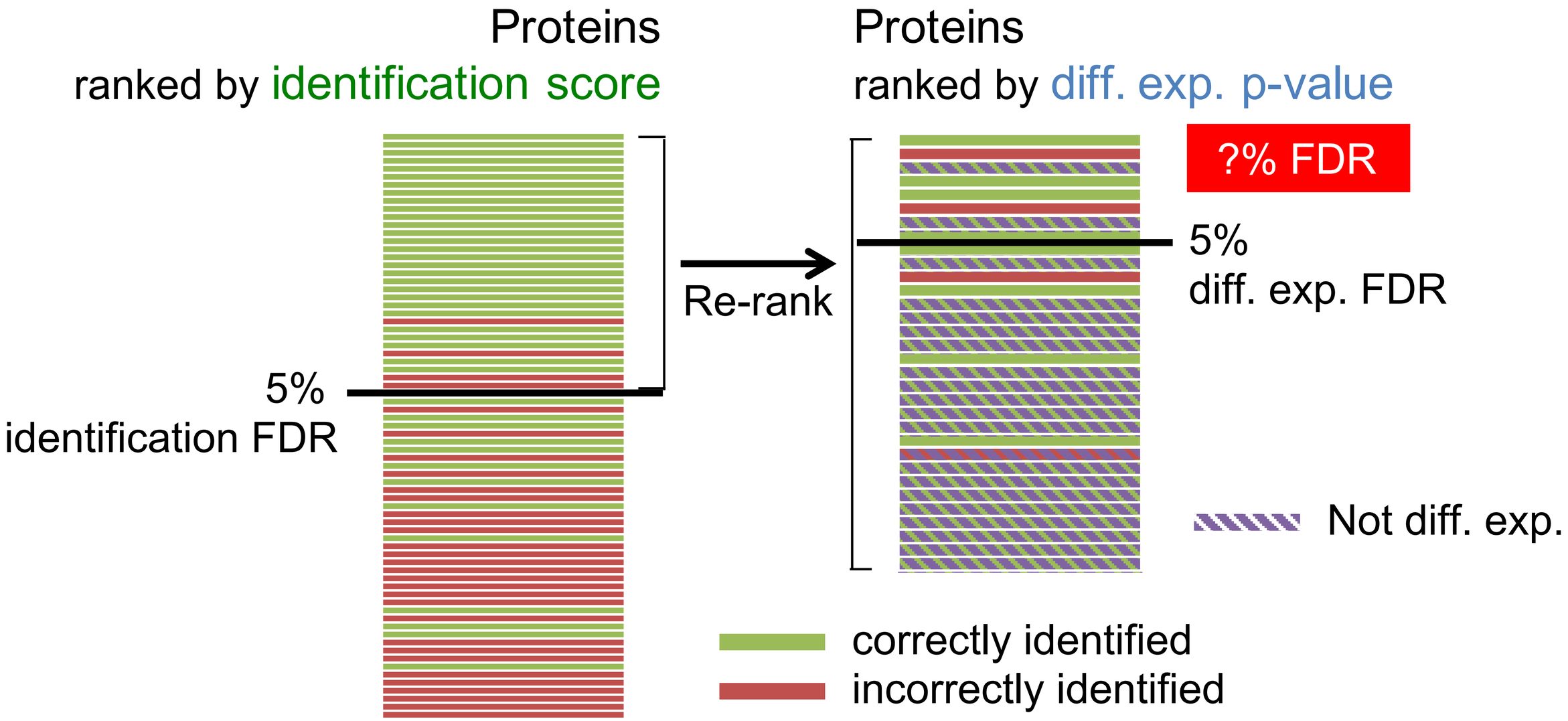
A false positive is a protein that …

… is incorrectly identified

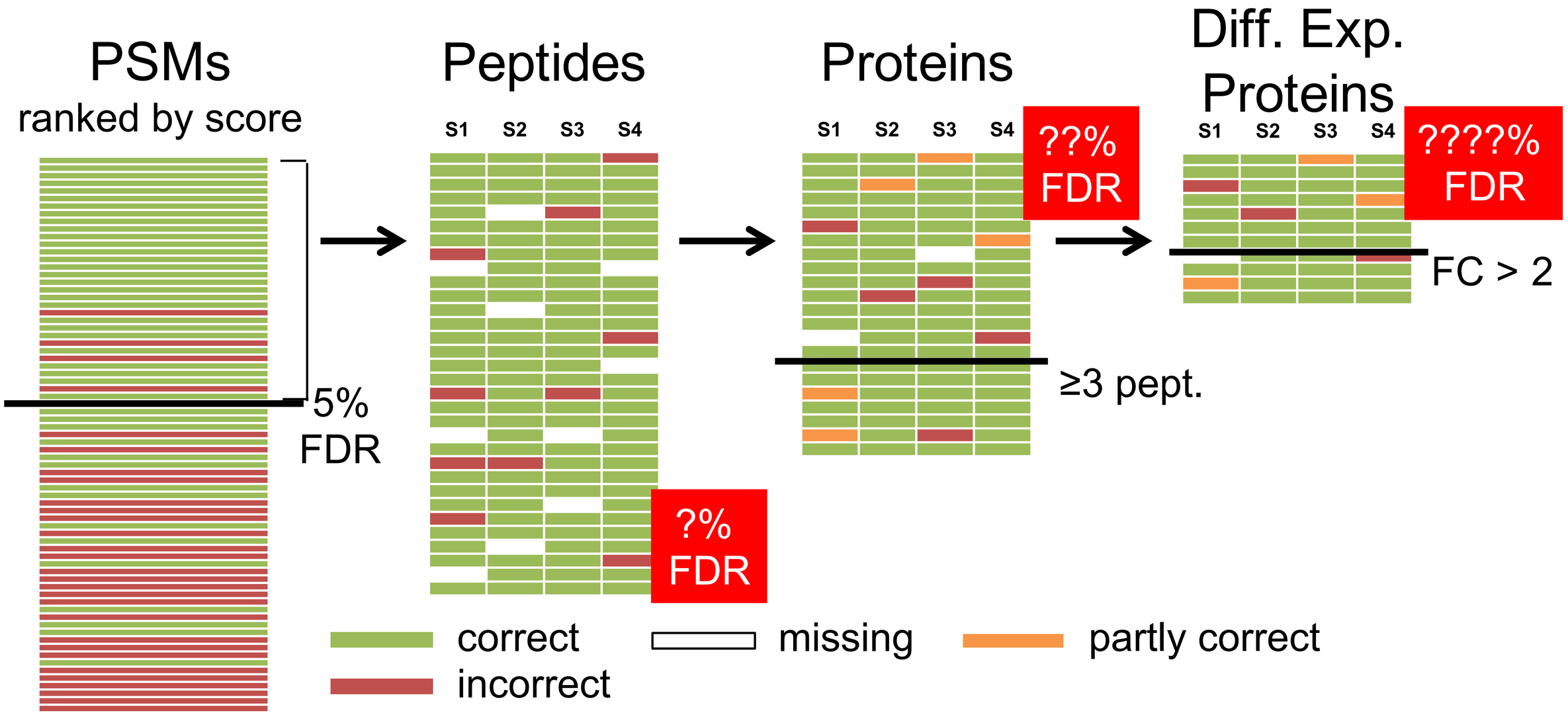
… **or** is incorrectly quantified

… **or** is not differentially expressed

# Thresholds: wolves in sheep's clothing

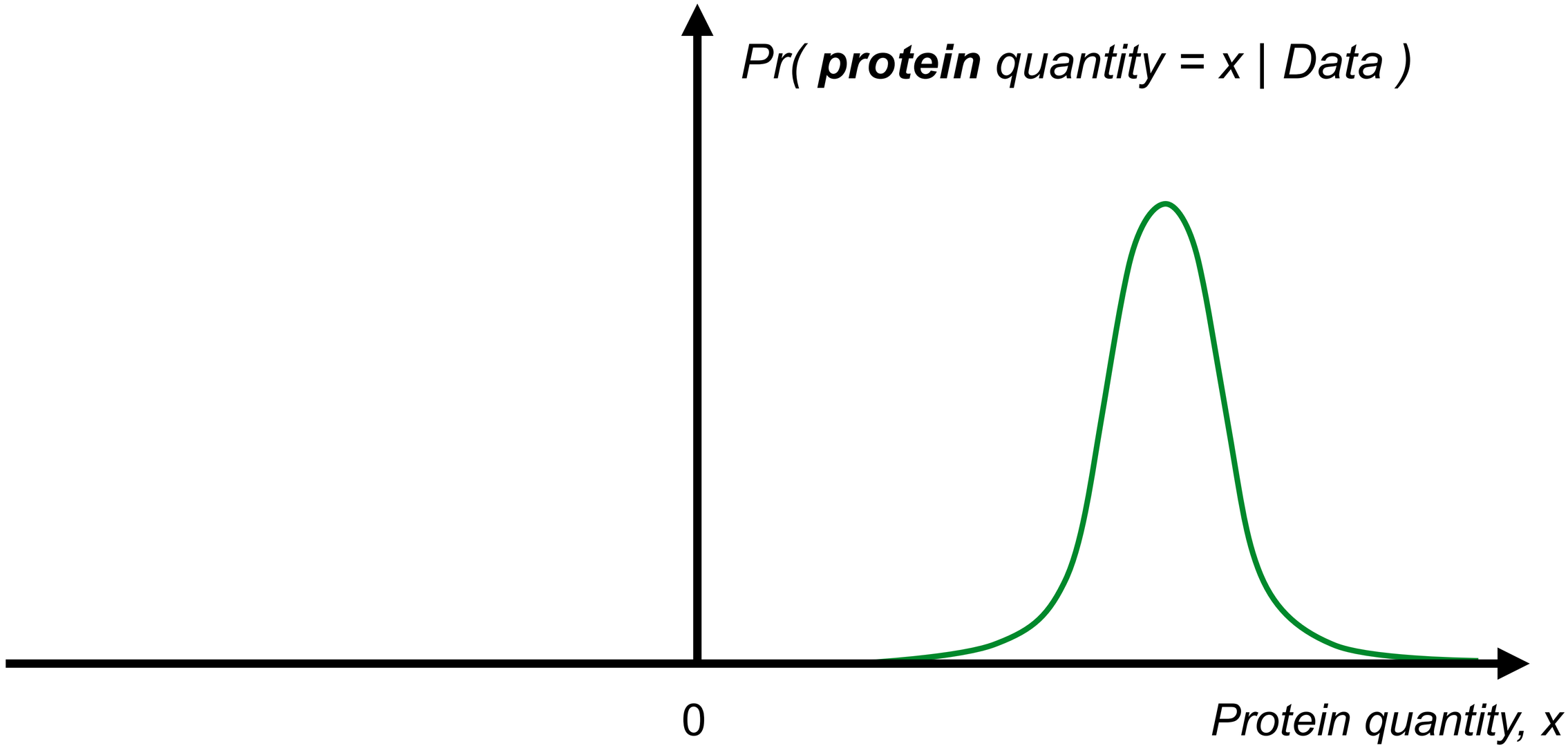


# The road to hell is paved by well-intended thresholds

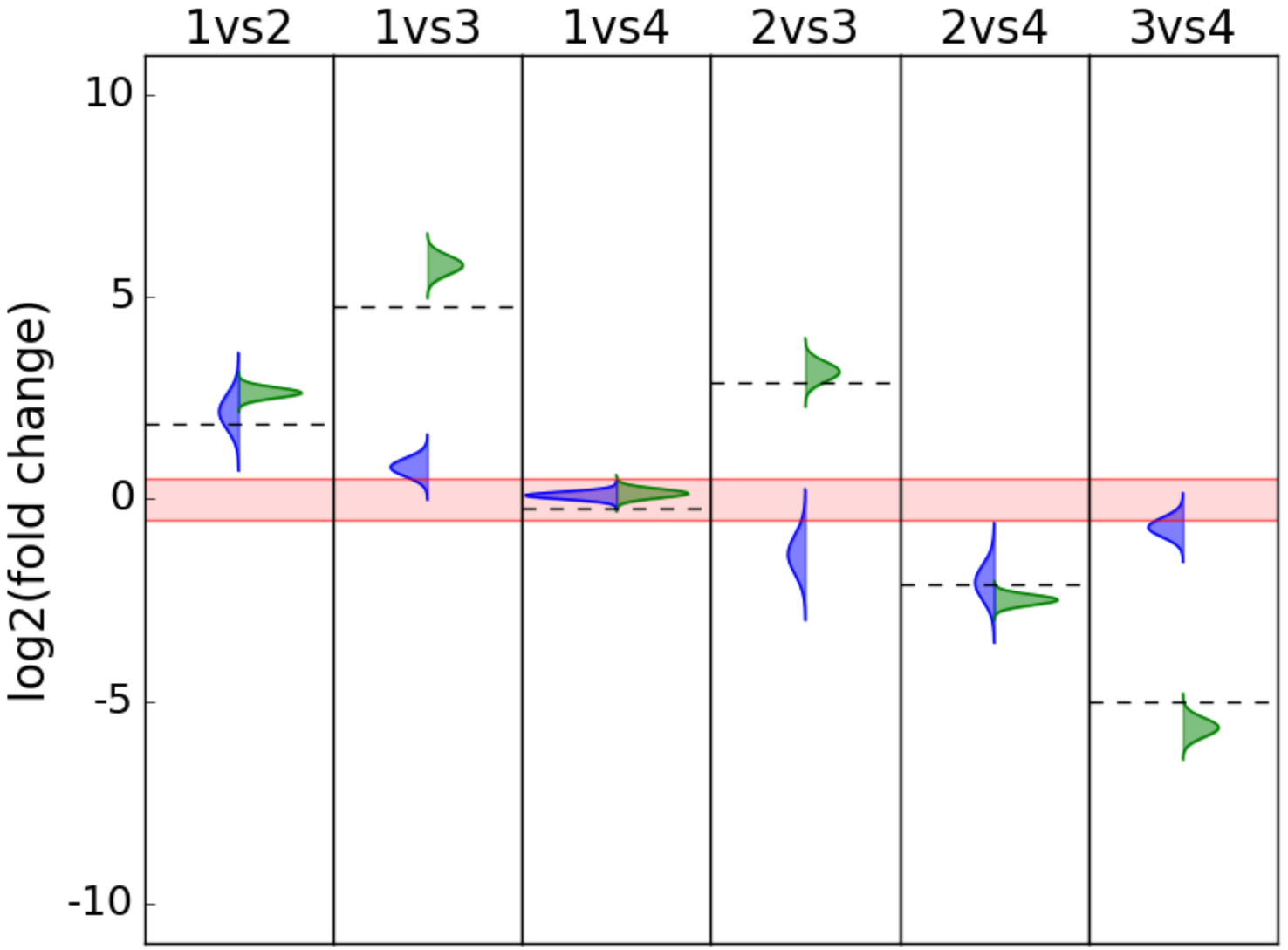


# Bayesian modeling:
# An era beyond multiple thresholds, imputation and p-values

# Quantitative Posterior Distribution are easy to understand

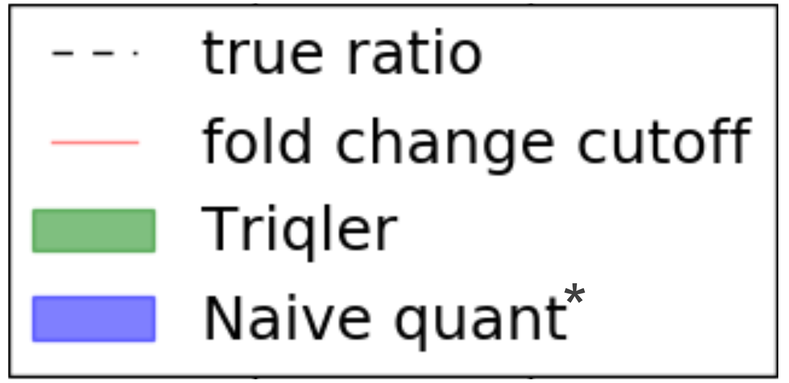


# Triqler outputs posterior distributions




## iPRG2015 study

MYG_HORSE spiked into yeast background




* 5% FDR PSM-id, Top3, row-mean imputation

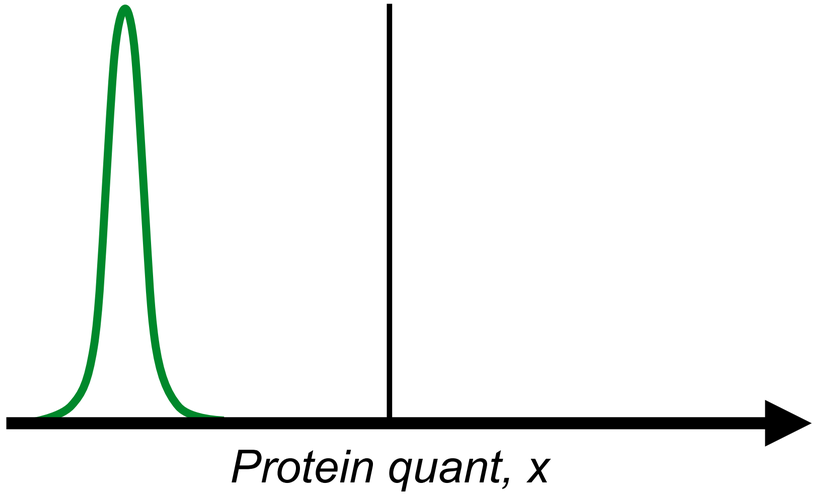
# Certainty in data manifested as certainty in quantity

*Pr( **protein** quant = x | Data )*

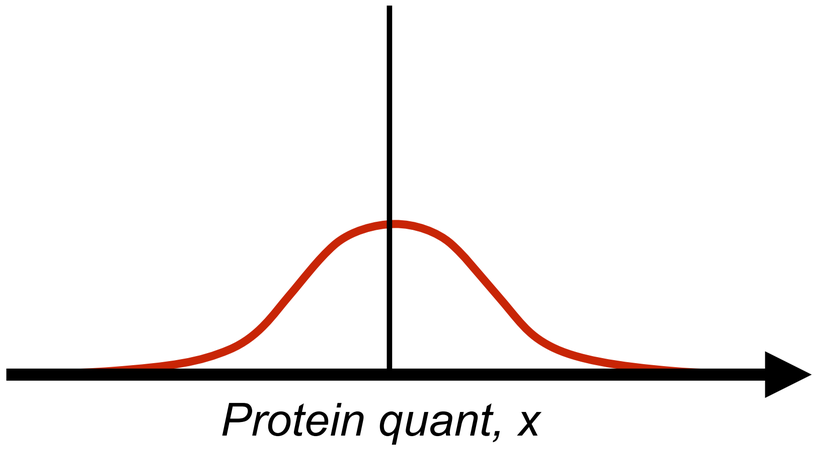
Certainty of:

- Feature identifications
- Peptide-spectrum matches
- Missing value Imputations
- Agreement between replicates

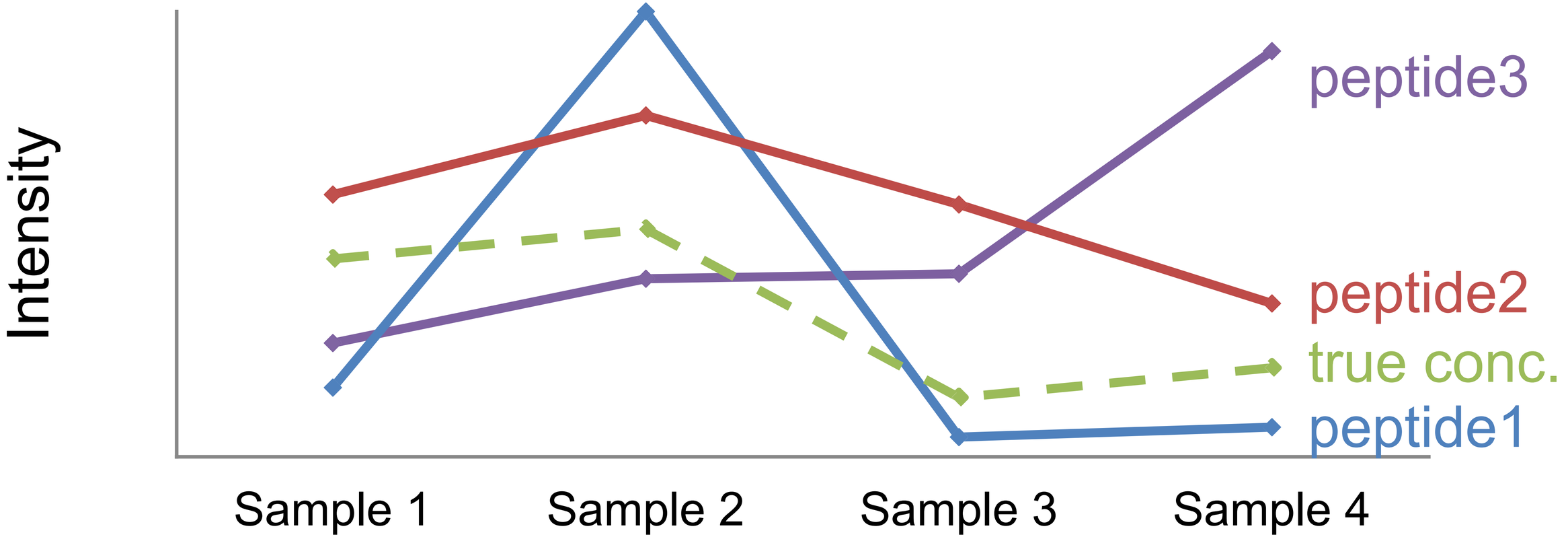
Certain



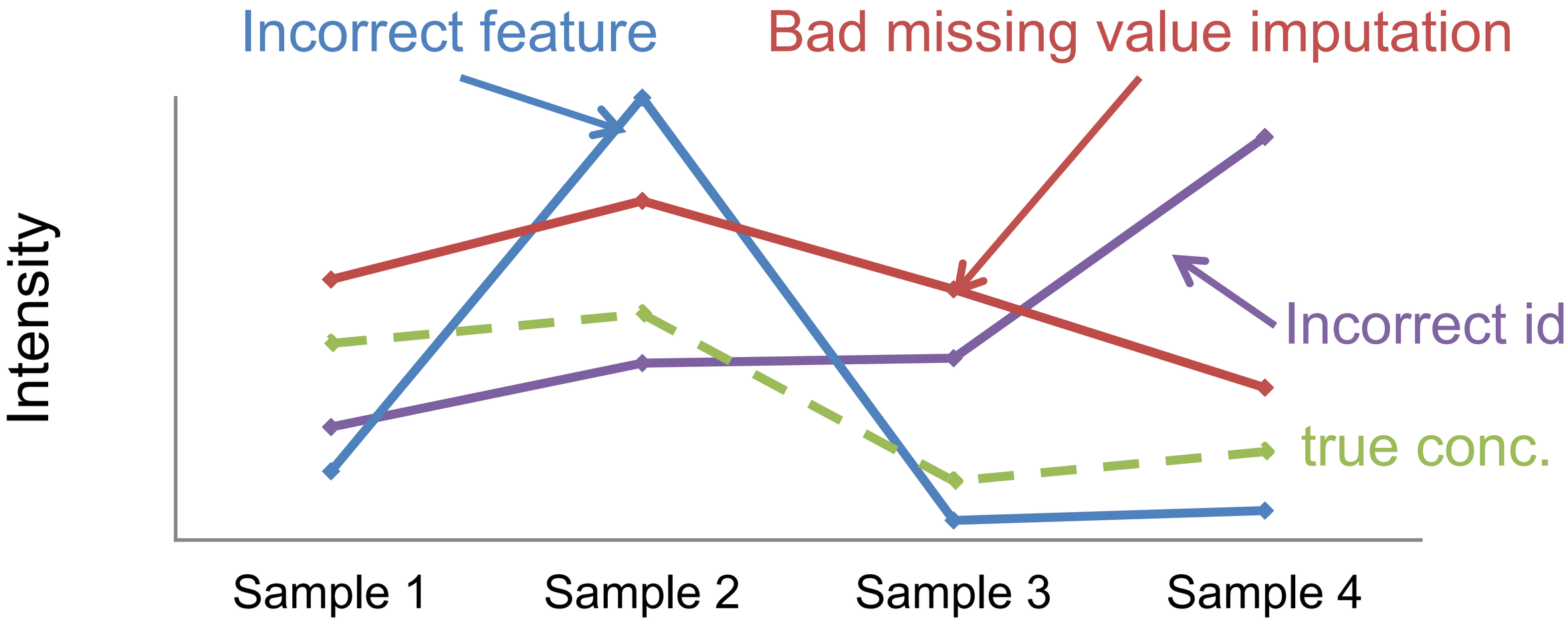

Uncertain



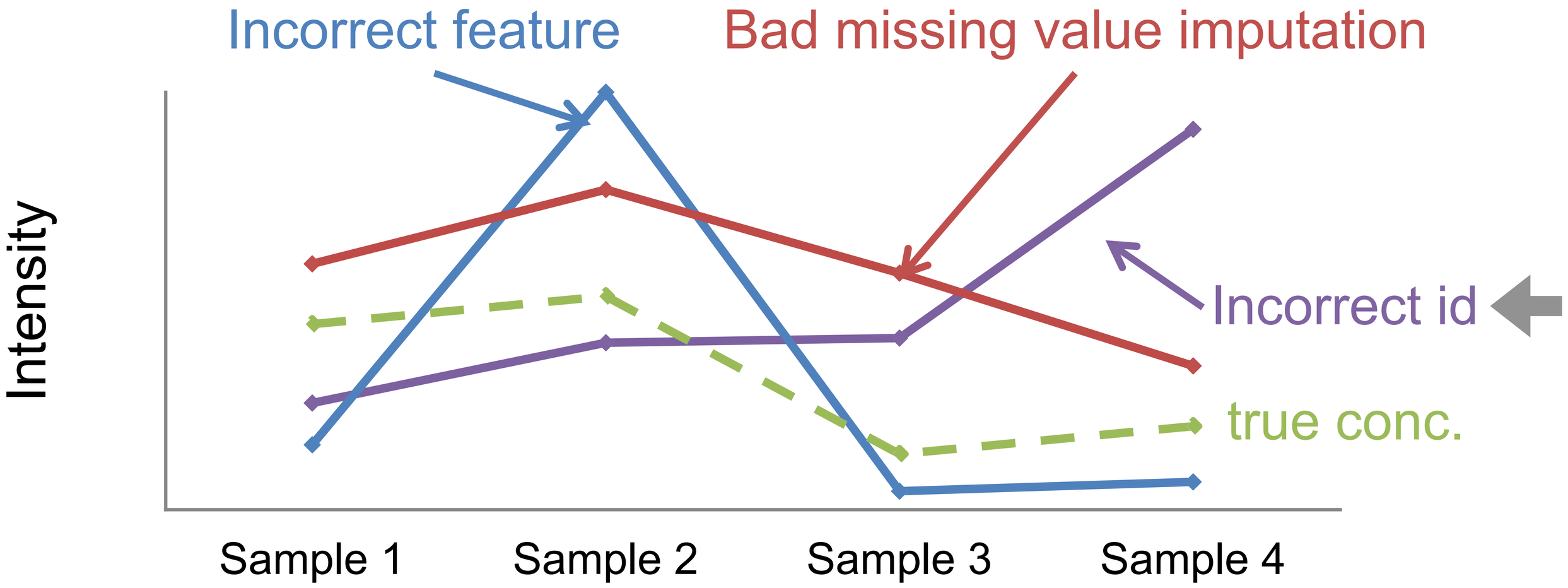
# Quantification: easy to find error source, hard to get right



# Quantification: easy to find error source, hard to get right

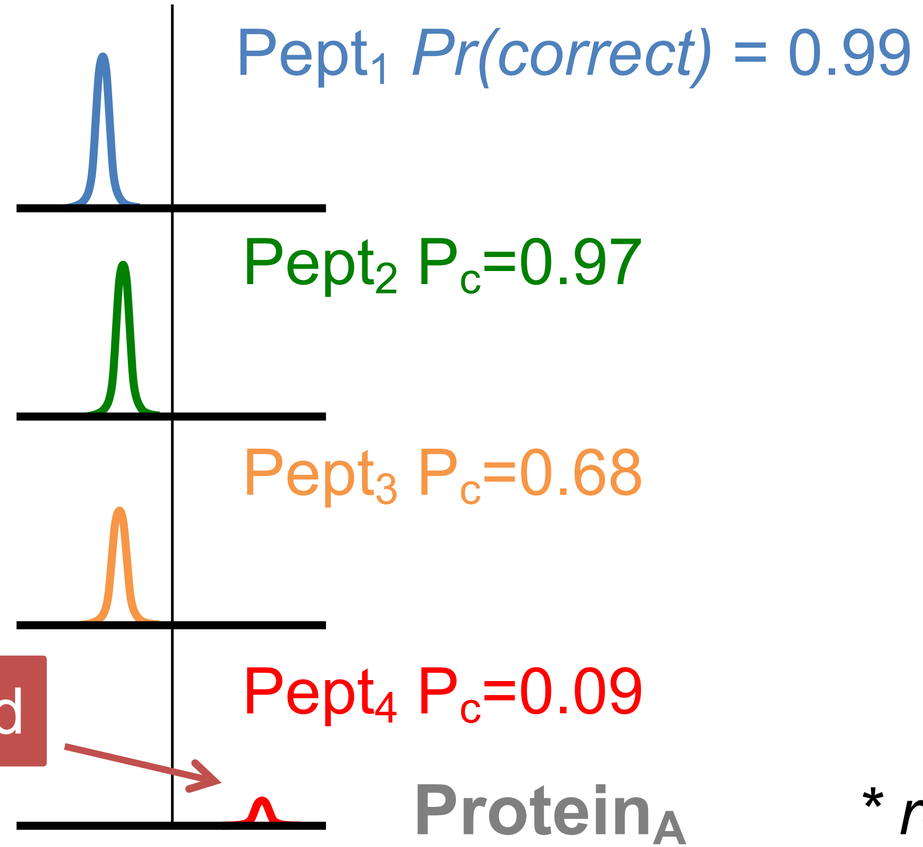


# Quantification: easy to find error source, hard to get right



# Forget thresholds, use *posterior* probabilities

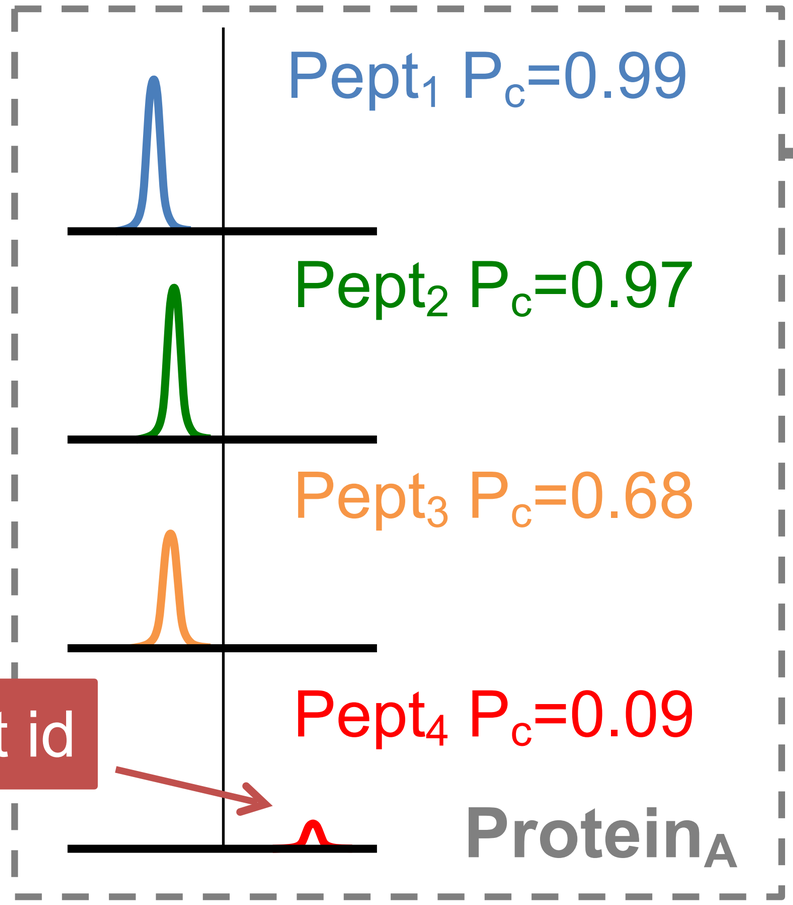
*Pr ( Data* | ***peptide*** *quant* = x )*




** relative to other samples; log2 transformed*

# Forget thresholds, use *posterior* probabilities

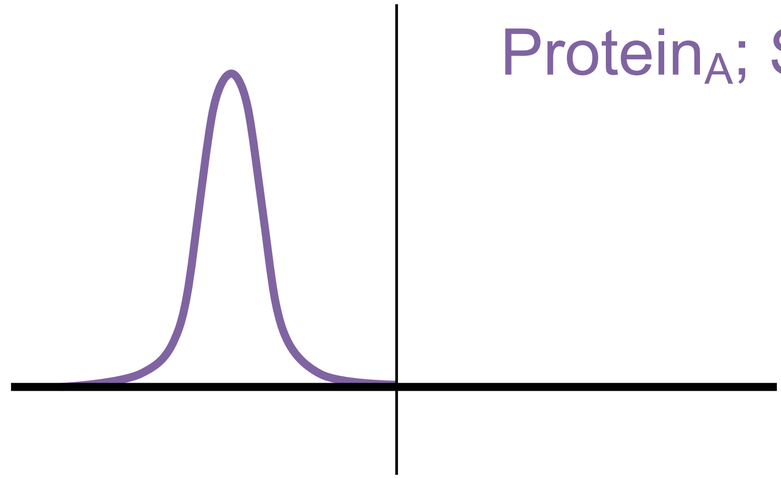
*Pr ( Data | **peptide** quant* = x )*




Incorrect id

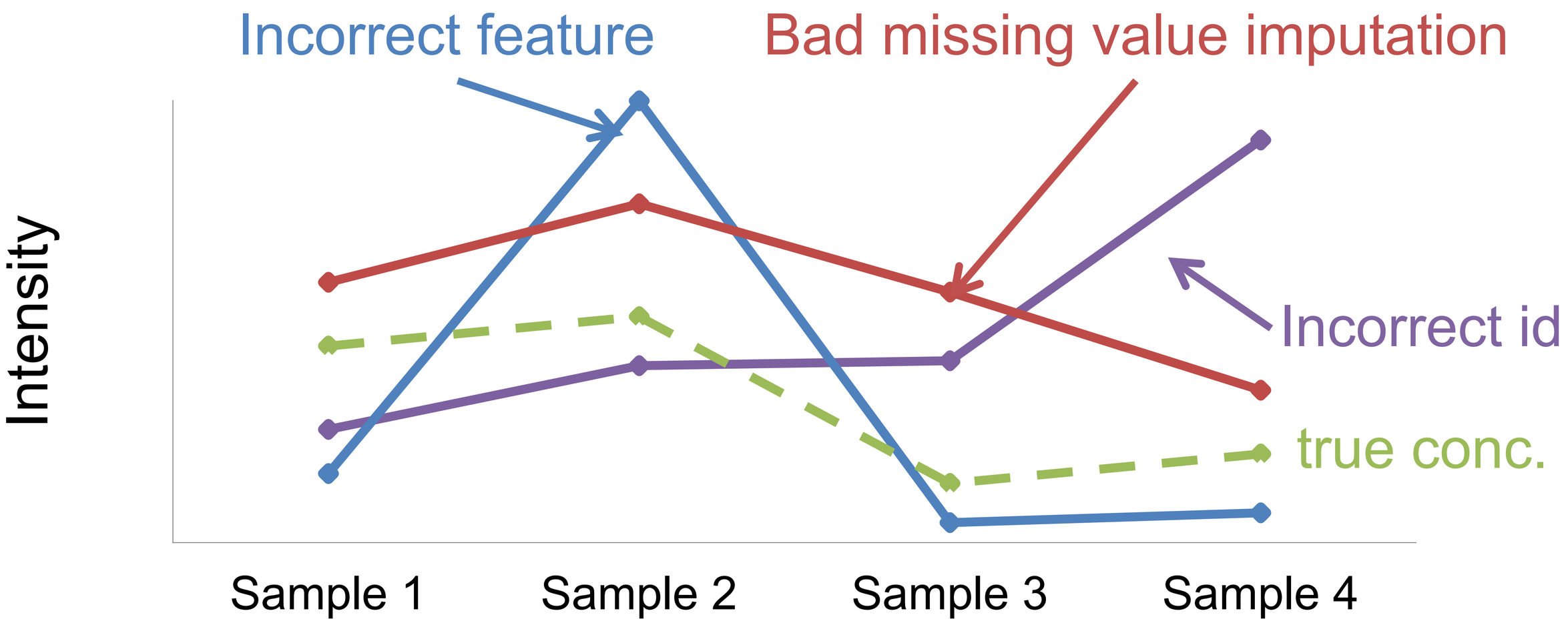
+prior

*Pr( **protein** quant* = x | Data )*

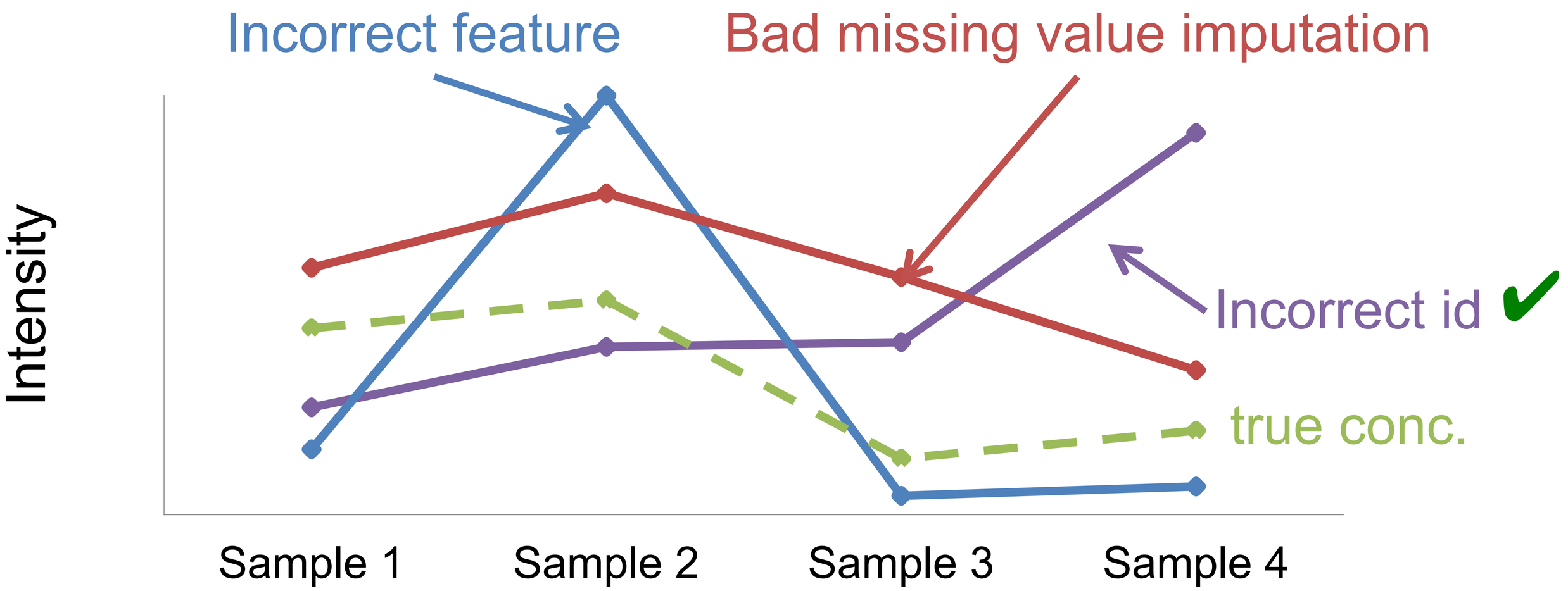



** relative to other samples; log2 transformed*

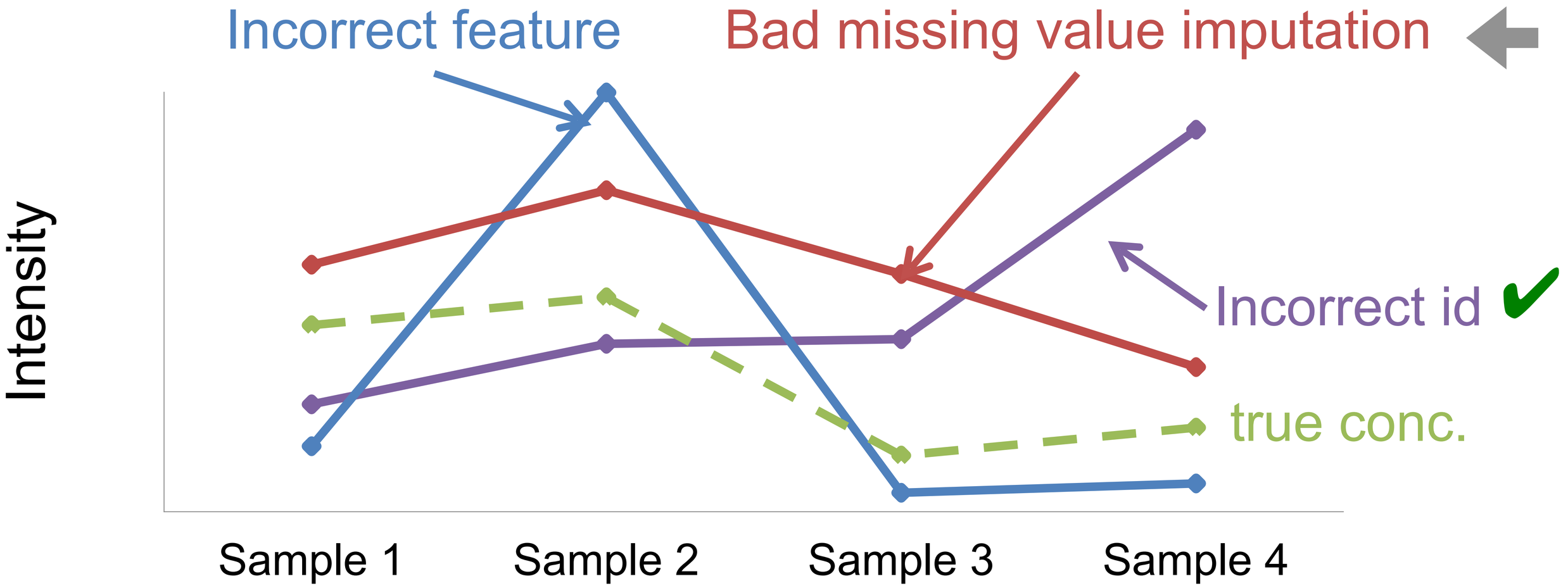
# Quantification: easy to find error source, hard to get right



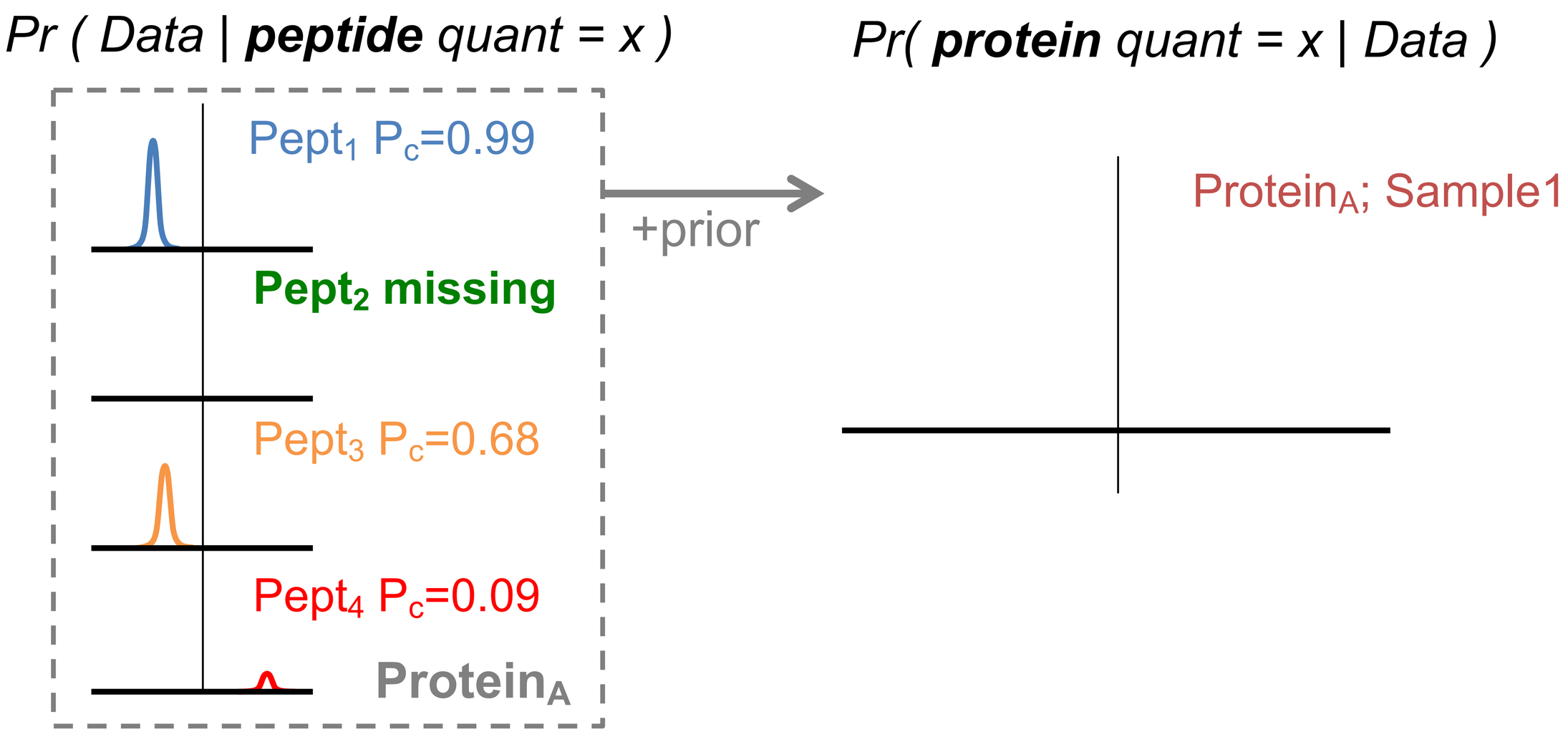
# Quantification: easy to find error source, hard to get right



# Quantification: easy to find error source, hard to get right

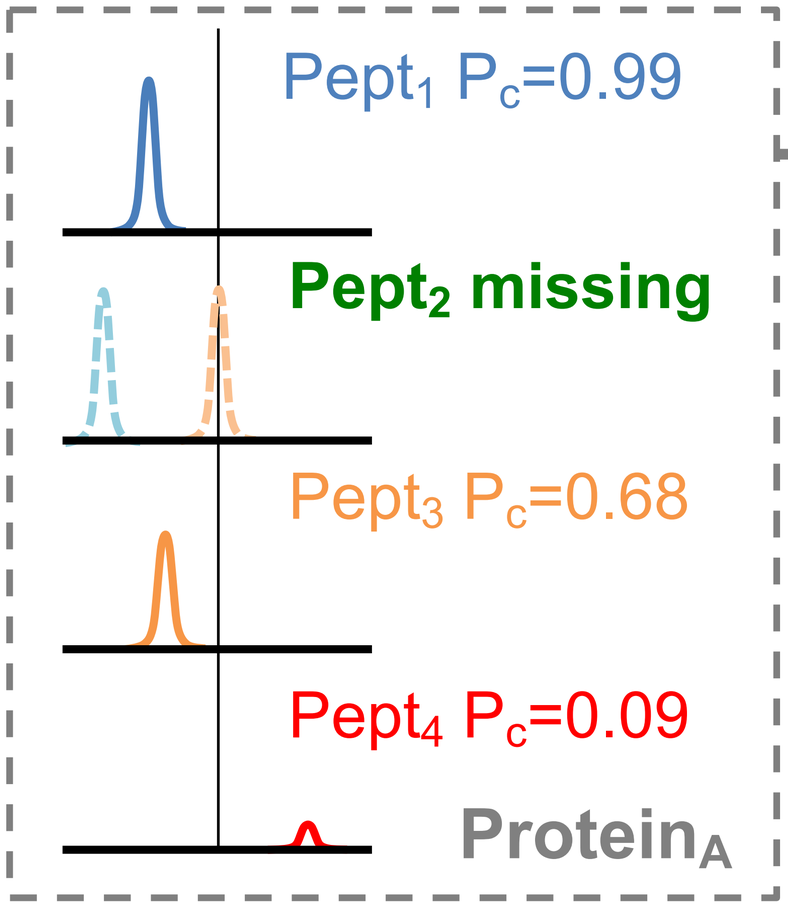


# Imputed values are less reliable than observed values



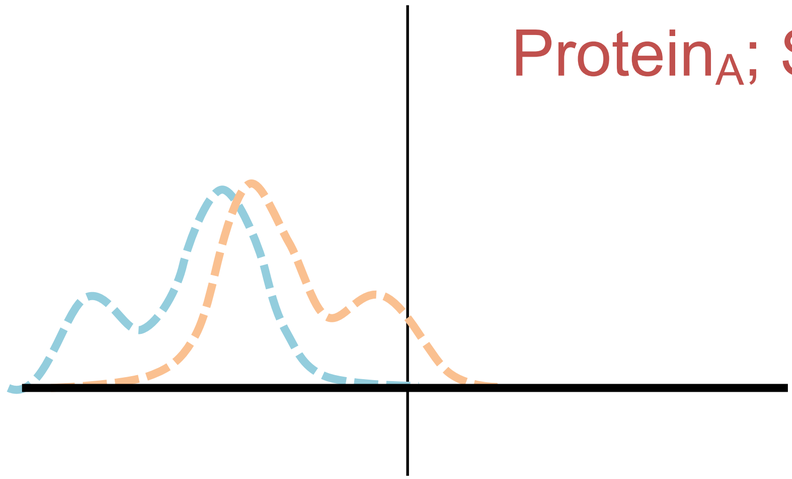
# Imputed values are less reliable than observed values

*Pr ( Data | **peptide** quant = x )*



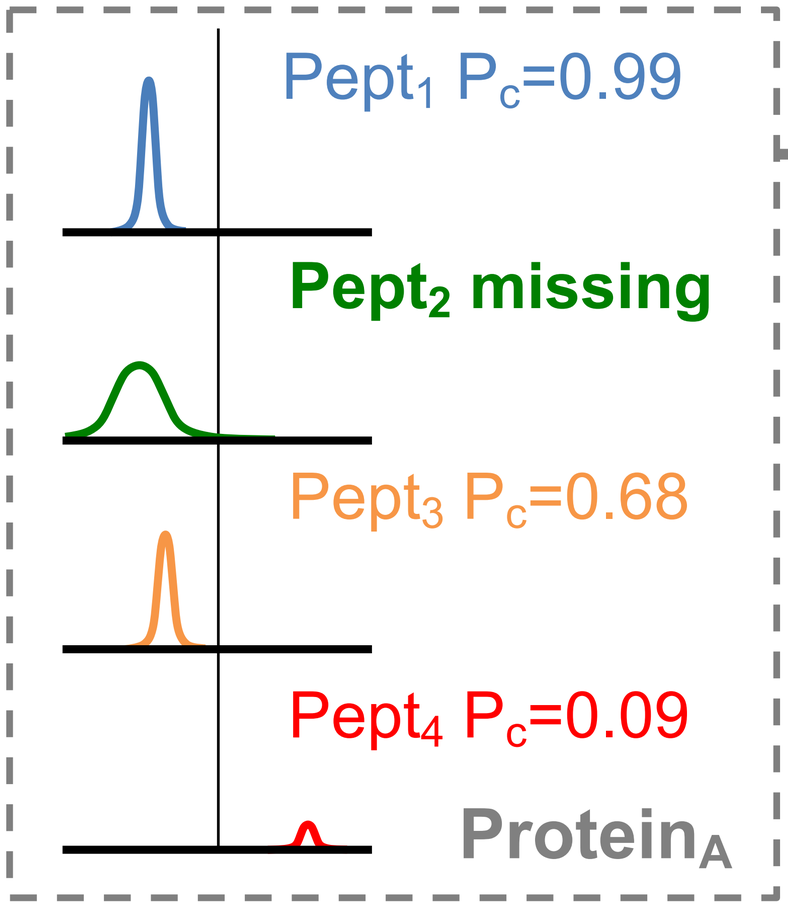

+prior

*Pr( **protein** quant = x | Data )*



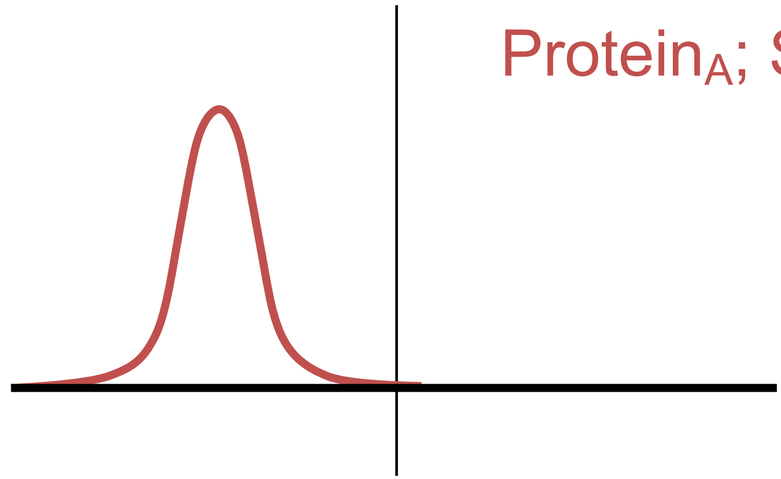
# Imputed values are less reliable than observed values

*Pr ( Data | **peptide** quant = x )*

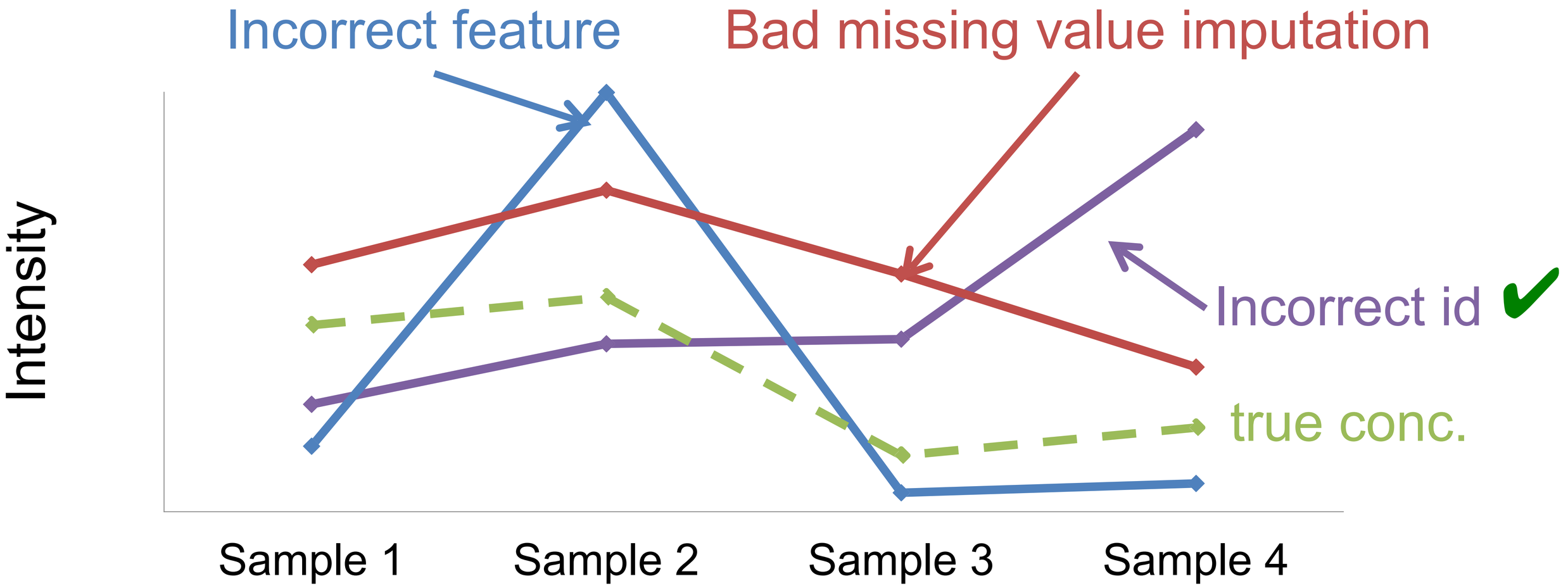



+prior

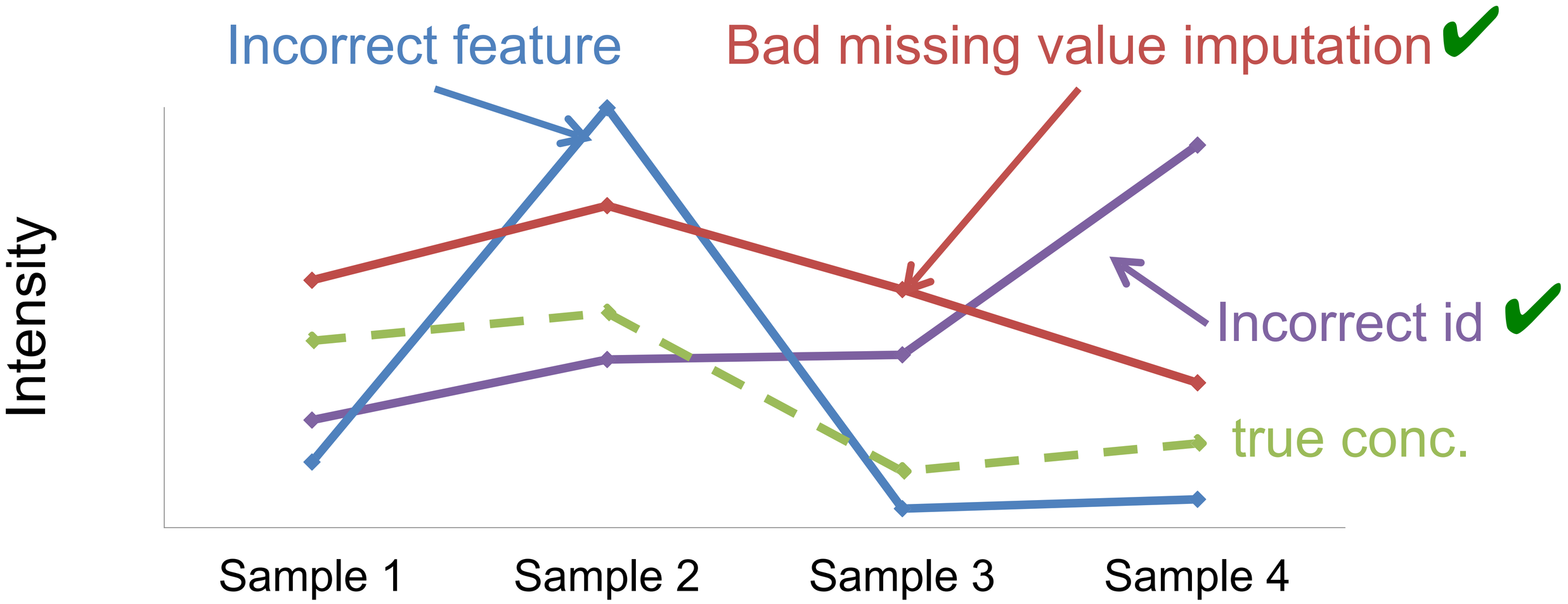
*Pr( **protein** quant = x | Data )*



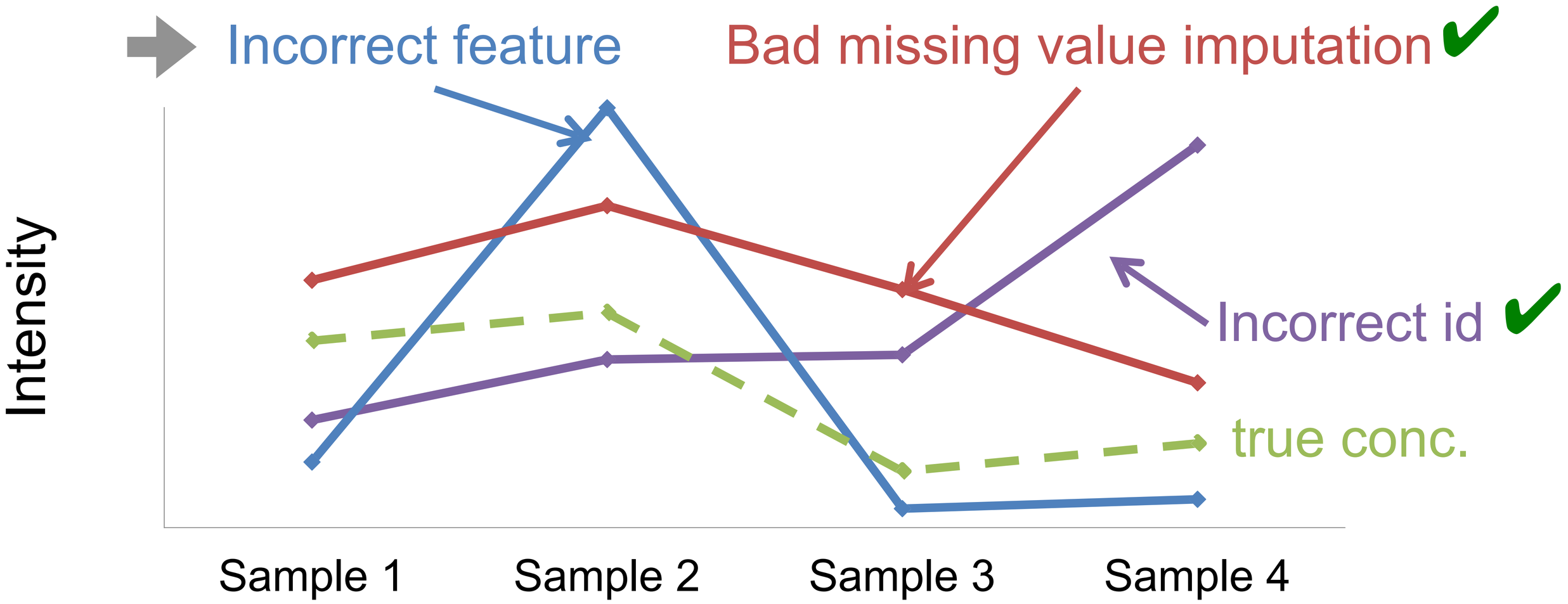
# Quantification: easy to find error source, hard to get right



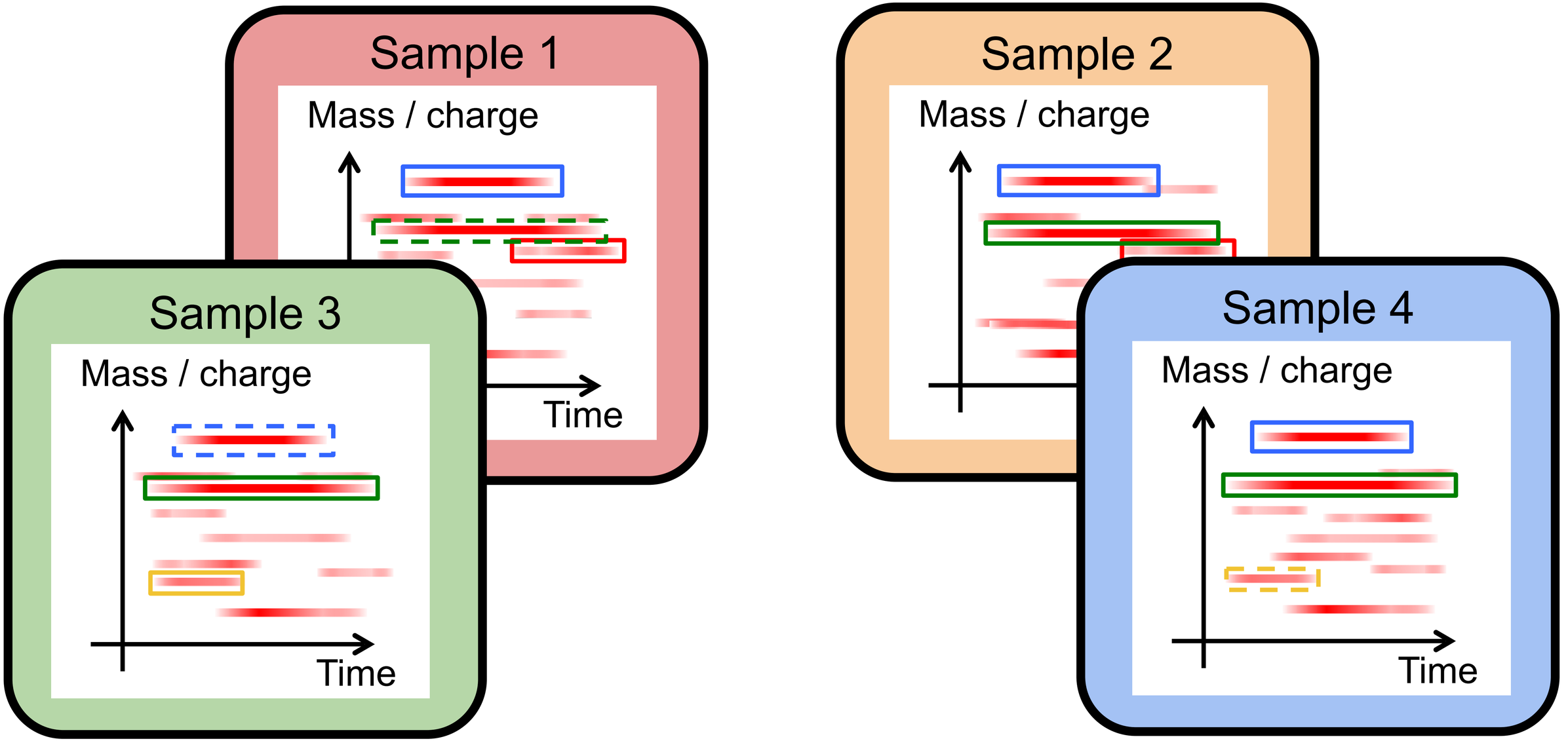
# Quantification: easy to find error source, hard to get right



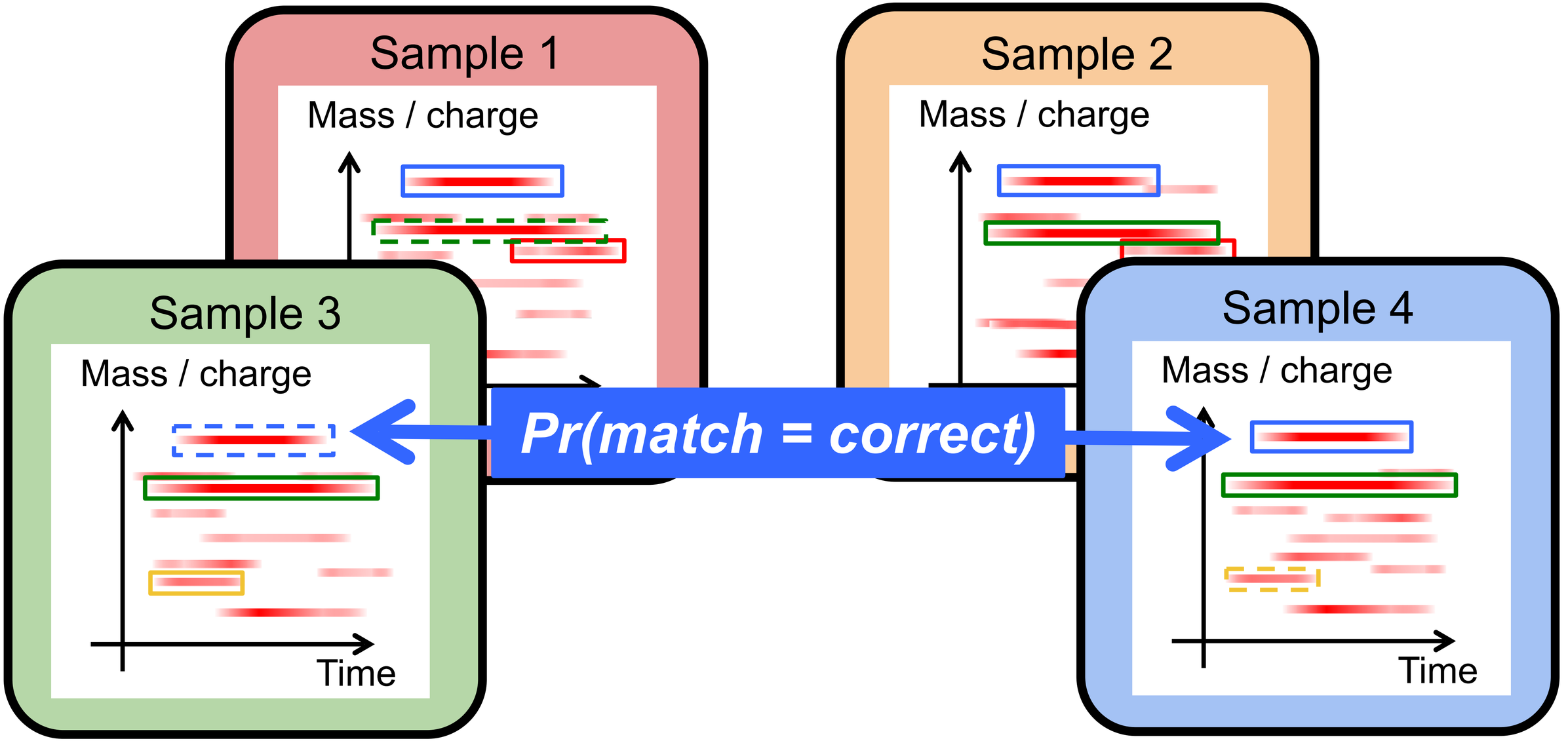
# Quantification: easy to find error source, hard to get right



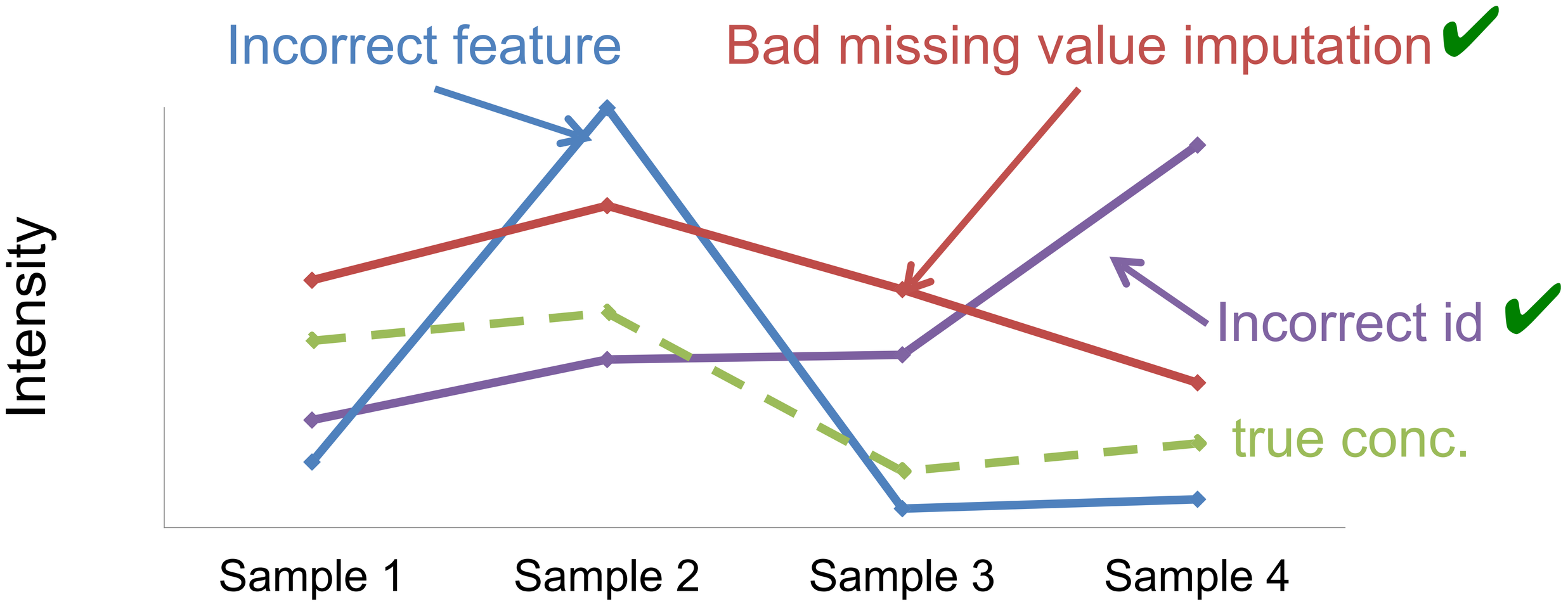
# Match-between-runs leveled up



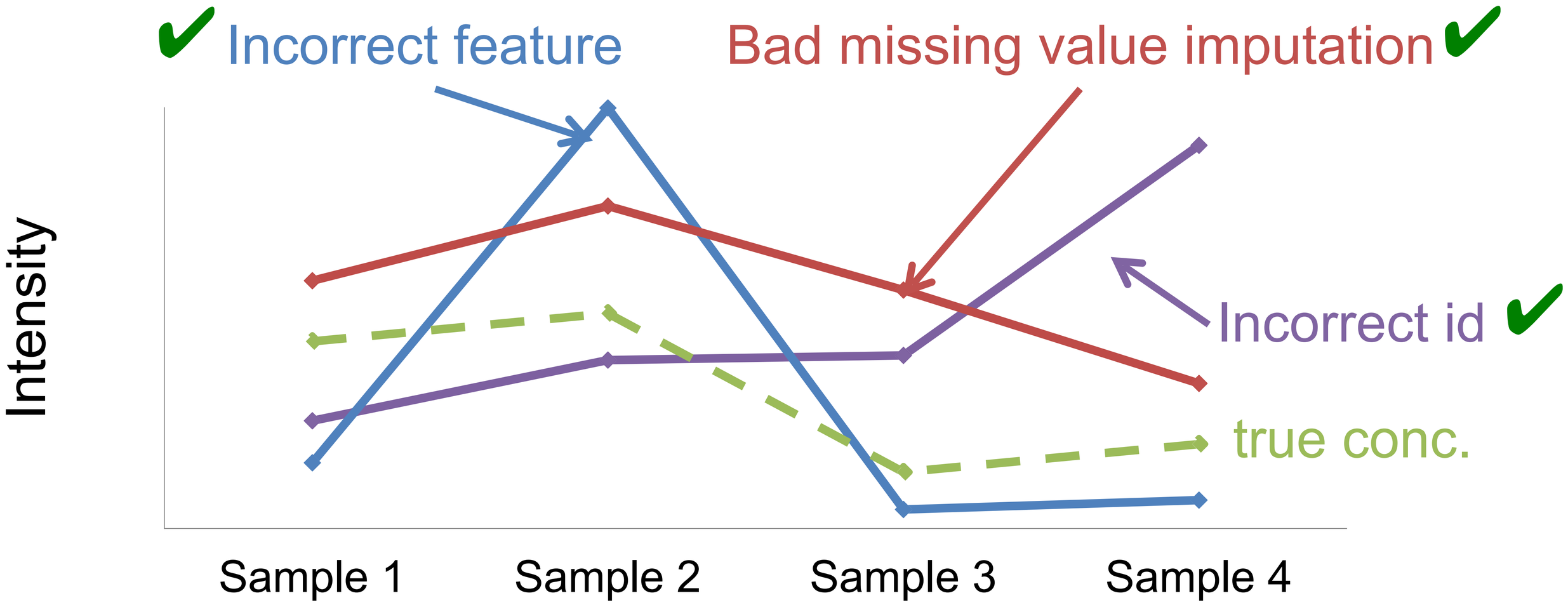
# Match-between-runs leveled up



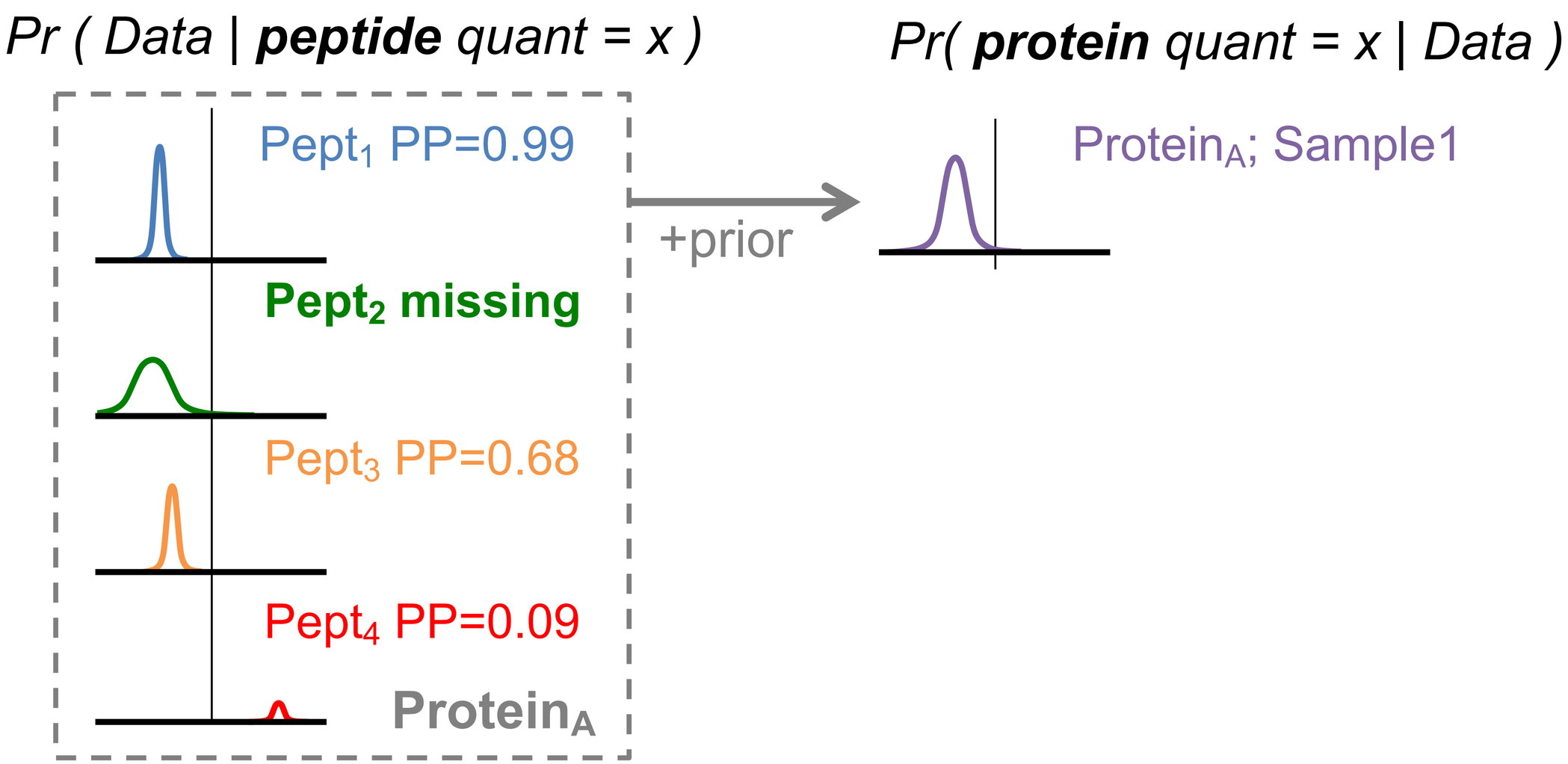
# Quantification: easy to find error source, hard to get right



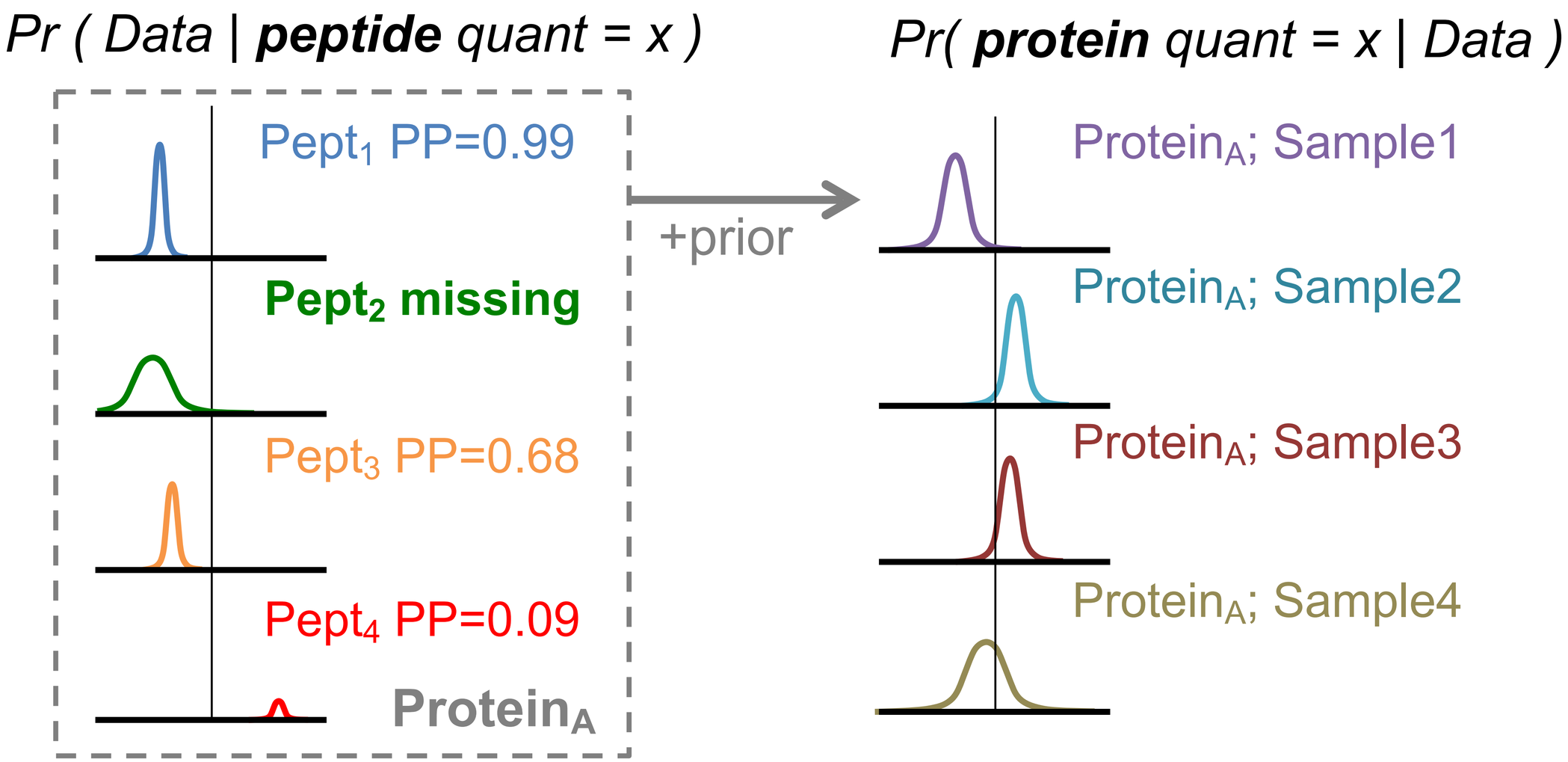
# Quantification: easy to find error source, hard to get right



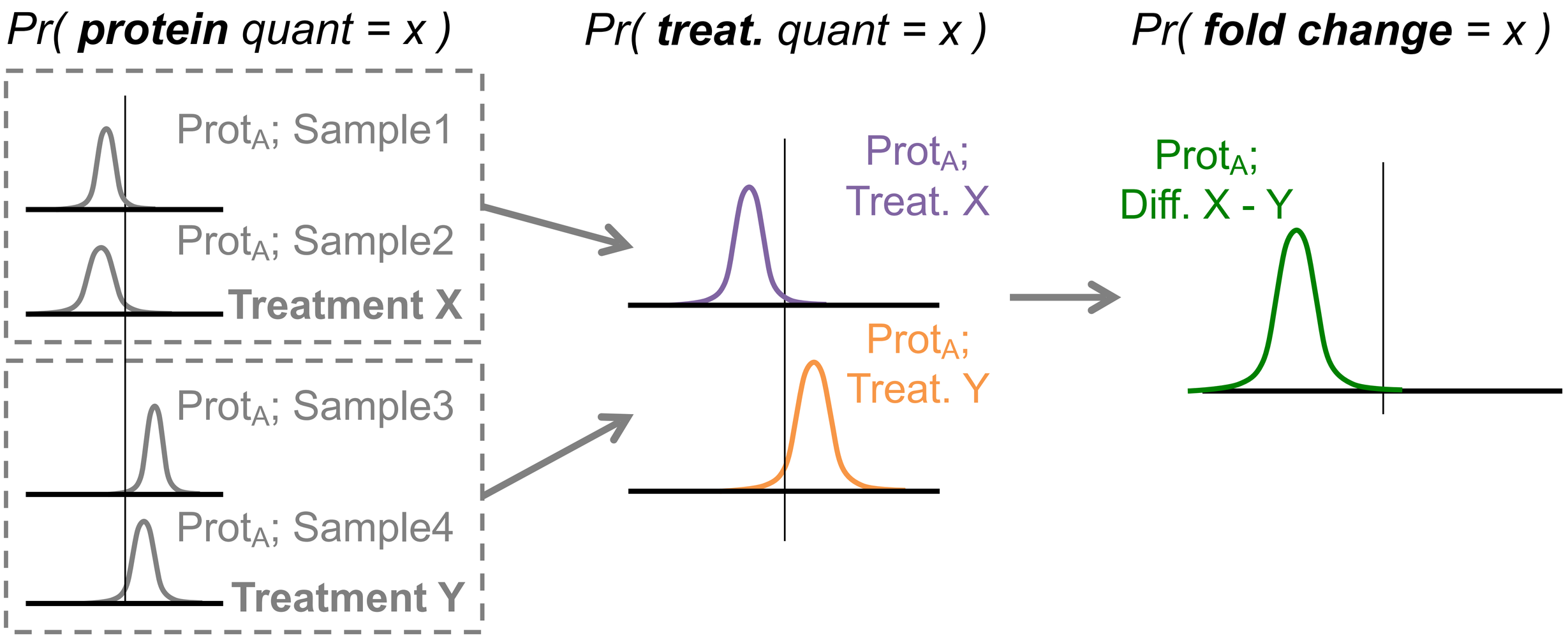
# Posterior distributions reflect our uncertainty



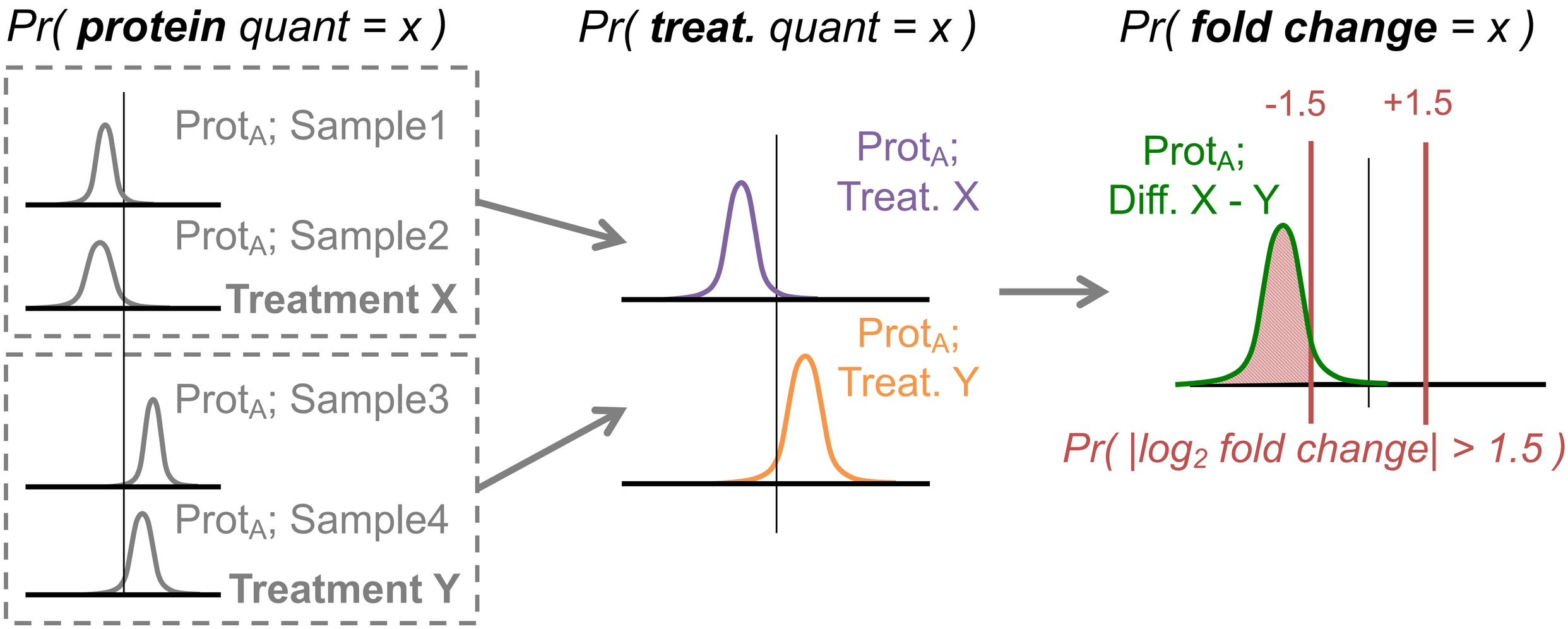
# Posterior distributions reflect our uncertainty



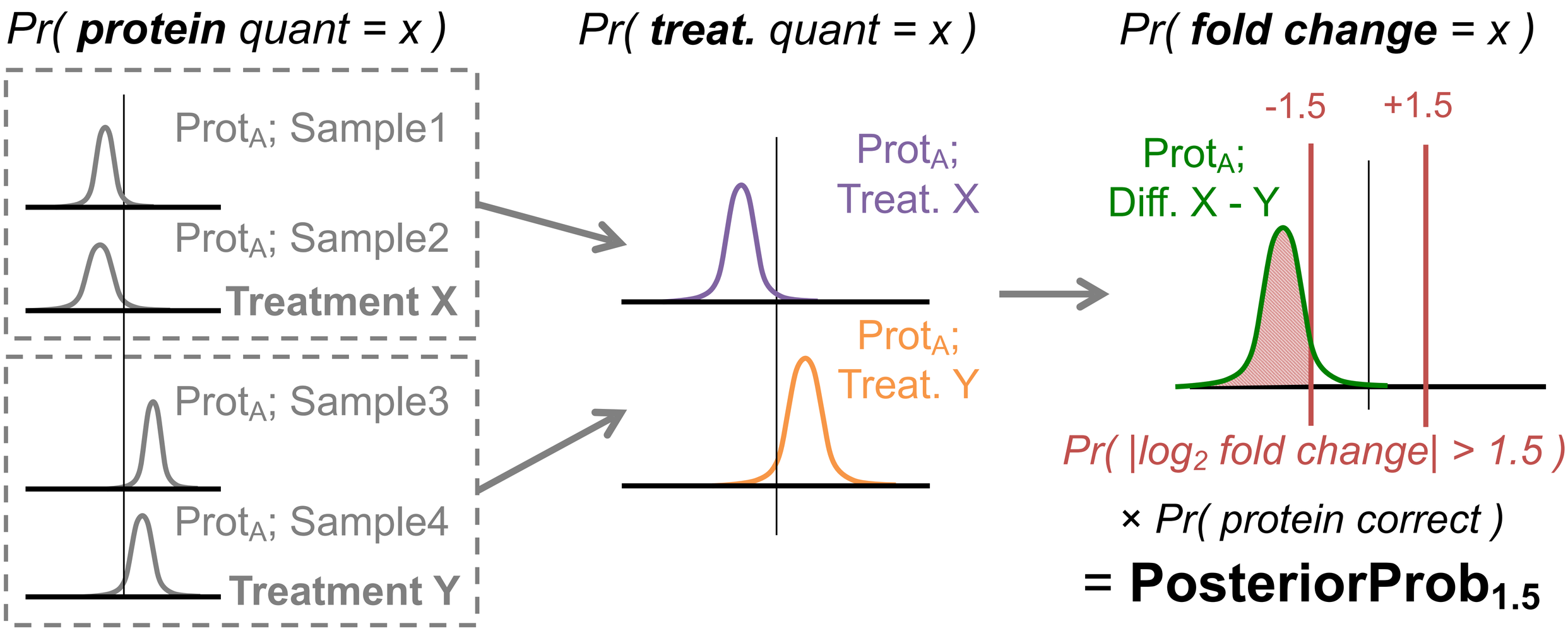
# Forget p-values, use posterior probabilities



# Forget p-values, use posterior probabilities



# Forget p-values, use posterior probabilities



# Forget p-values, use posterior probabilities

p-value < 0.05

- **NOT** >95% chance that the protein is diff. exp.

$PosteriorProb_{1.5} = 0.95$

- 95% chance that the protein has a $|\log_2$ fold change$| > 1.5$

# Forget p-values, use posterior probabilities

p-value < 0.05

- **NOT** >~~95% chance that the protein is diff. exp.~~

$PosteriorProb_{1.5} = 0.95$

- 95% chance that the protein has a $|\log_2$ fold change$| > 1.5$

# Forget p-values, use posterior probabilities

p-value < 0.05

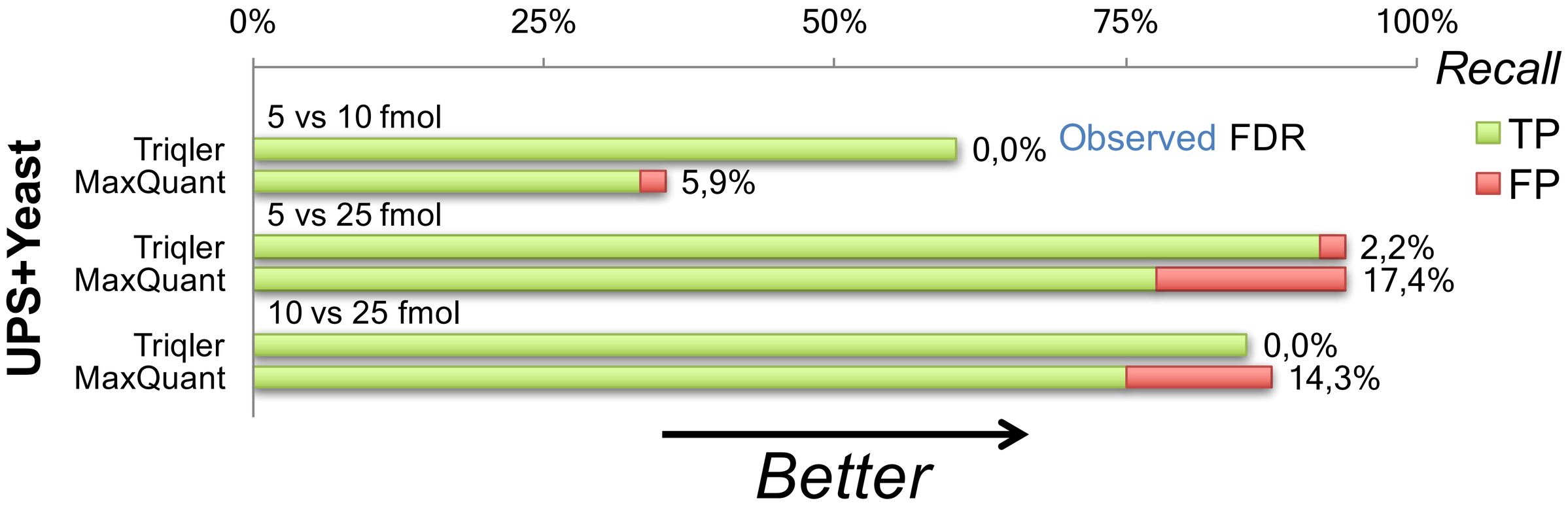
- **NOT** >~~95% chance that the protein is diff. exp.~~

$PosteriorProb_{1.5} = 0.95$

- 95% chance that the protein has a $|\log_2$ fold change$| > 1.5$ ✔

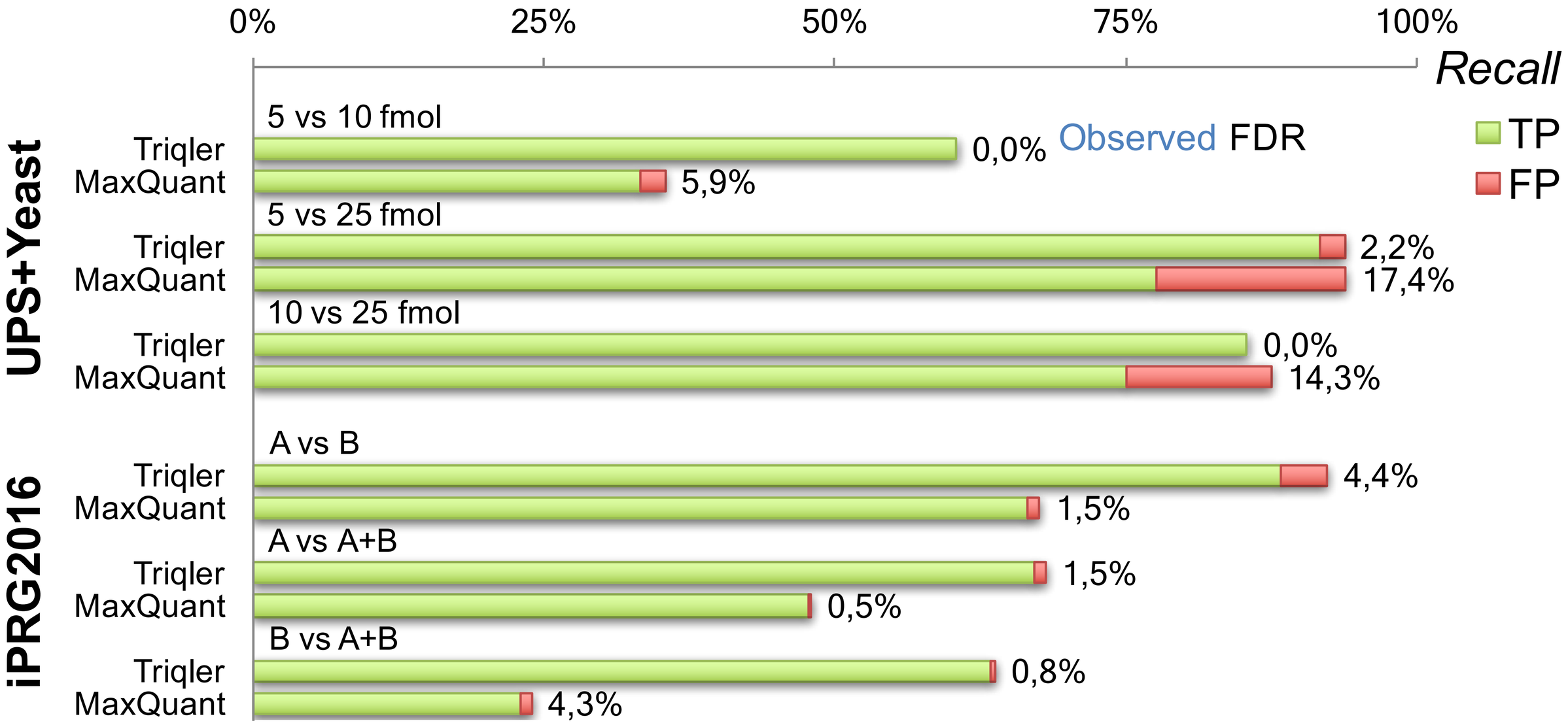
# Results!

# Triqler controls FDR at high recall




5% Reported FDR

# Triqler controls FDR at high recall



# Take-homes

- Triqler combines identification and quantification errors.

# Take-homes

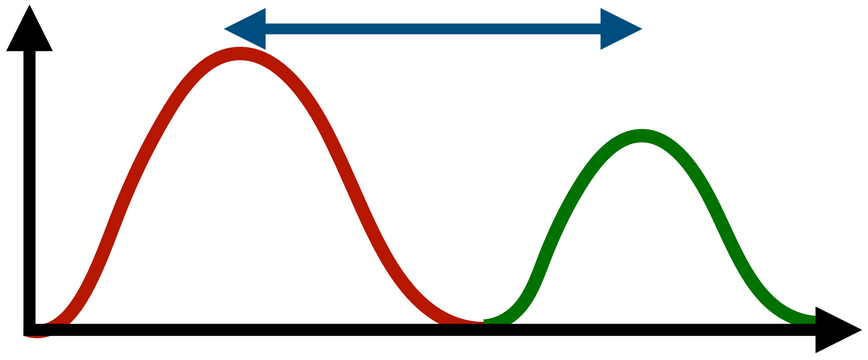
- Triqler combines identification and quantification errors.
- Triqler uses posterior probabilities instead of thresholds, imputation and p-values to control the FDR of quantitative experiments.

# Take-homes

- Triqler combines identification and quantification errors.
- Triqler uses posterior probabilities instead of thresholds, imputation and p-values to control the FDR of quantitative experiments.
- Controlling for errors in complex processing pipelines leads to higher yields for experiments.

# Outline

1. Background on label-free quantification
2. Combined identification and quantification error rates — Triqler
3. Clustering and Quantification MS/MS data — Quandenser



# Clustering and Quantification — Quandenser



New Results

Comment on this paper

Posted June 13, 2019.

**Focus on the spectra that matter by clustering of quantification data in shotgun proteomics**

Matthew The, Lukas Käll

**doi:** https://doi.org/10.1101/488015





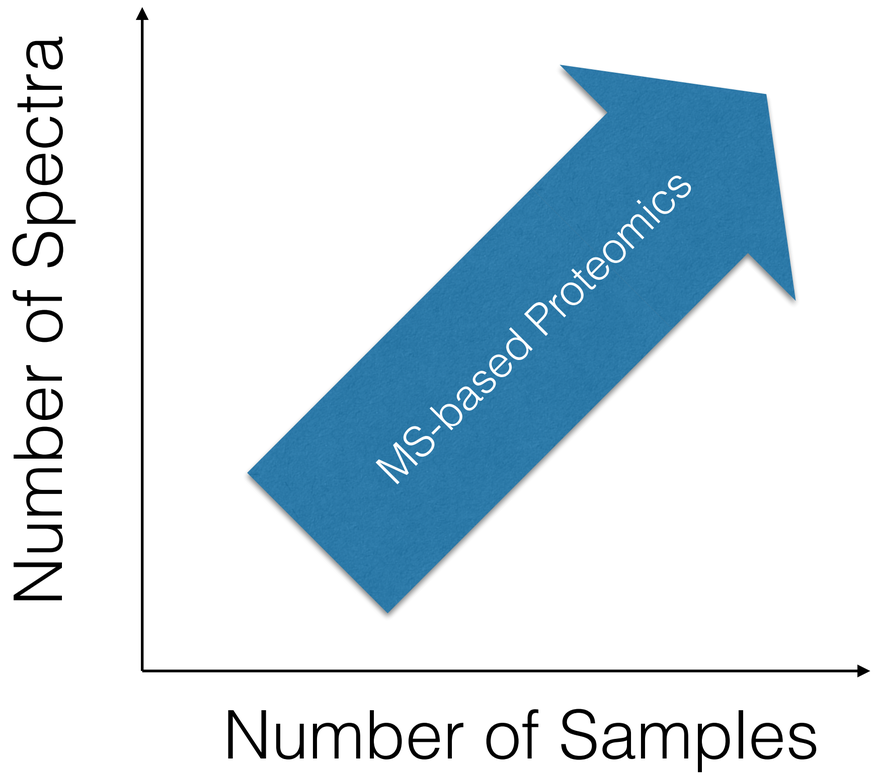
**Abstract**

In shotgun proteomics, the information extractable from label-free quantification experiments is typically limited by the identification rate and the noise level in the quantitative data. This generally causes a low sensitivity in differential expression analysis on protein level. Here, we propose a quantification-first approach that reverses the classical identification-first workflow. This prevents valuable information from being discarded prematurely in the identification stage and allows us to spend more effort on the identification process. Specifically, we introduce a method, Quandenser, that applies unsupervised clustering on both MS1 and MS2 level to summarize all analytes of interest without assigning identities. Not only does this eliminate the need for redoing the quantification for each new set of search parameters and engines, it also reduces search time due to the data reduction by MS2 clustering. For a dataset of partially known composition, we could now employ open modification and de novo searches to identify analytes of interest that would have gone unnoticed in traditional pipelines. Moreover, Quandenser reports error rates on feature level, which we integrated into our probabilistic protein quantification method, Triqler. This propagates error probabilities from feature to protein



Matthew The
(again)

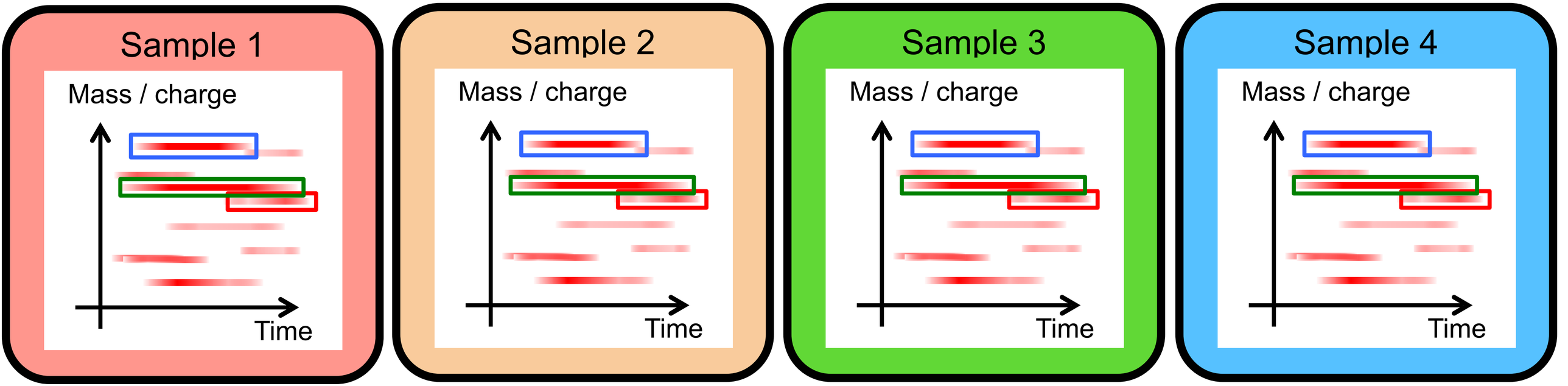
# Proteomics generates an increasing amount of data



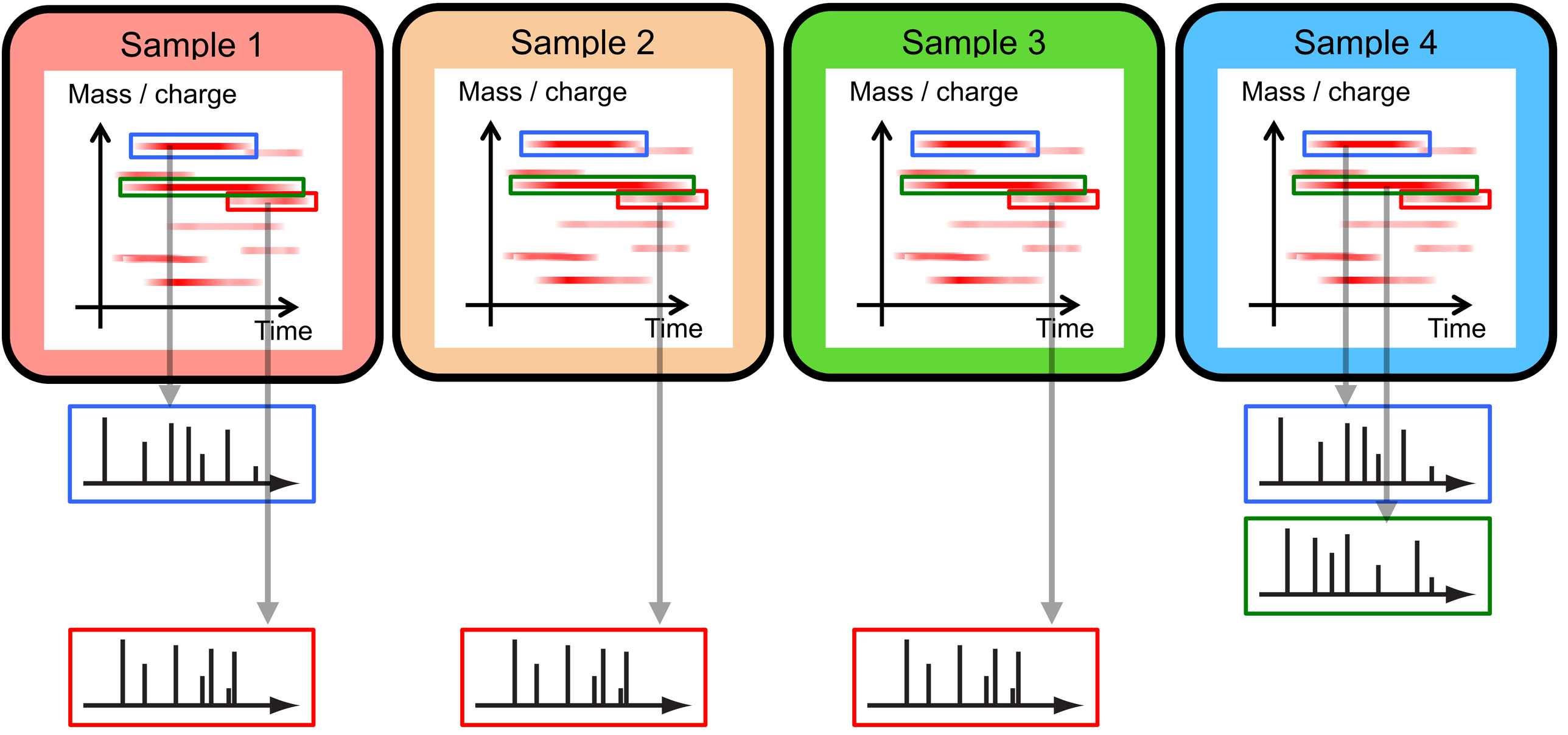

MS-based proteomics needs methods that can interpret large numbers of:

- Spectra
- Samples

# MS1 features can often be retrieved across samples…

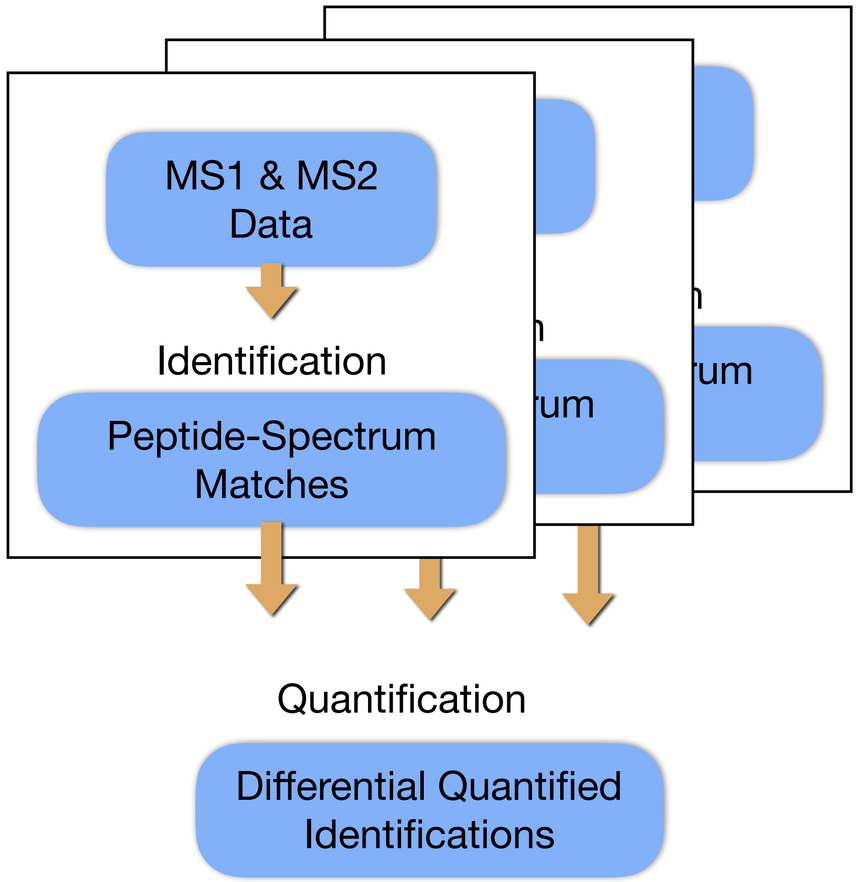


# … but MS2 spectra are sparse



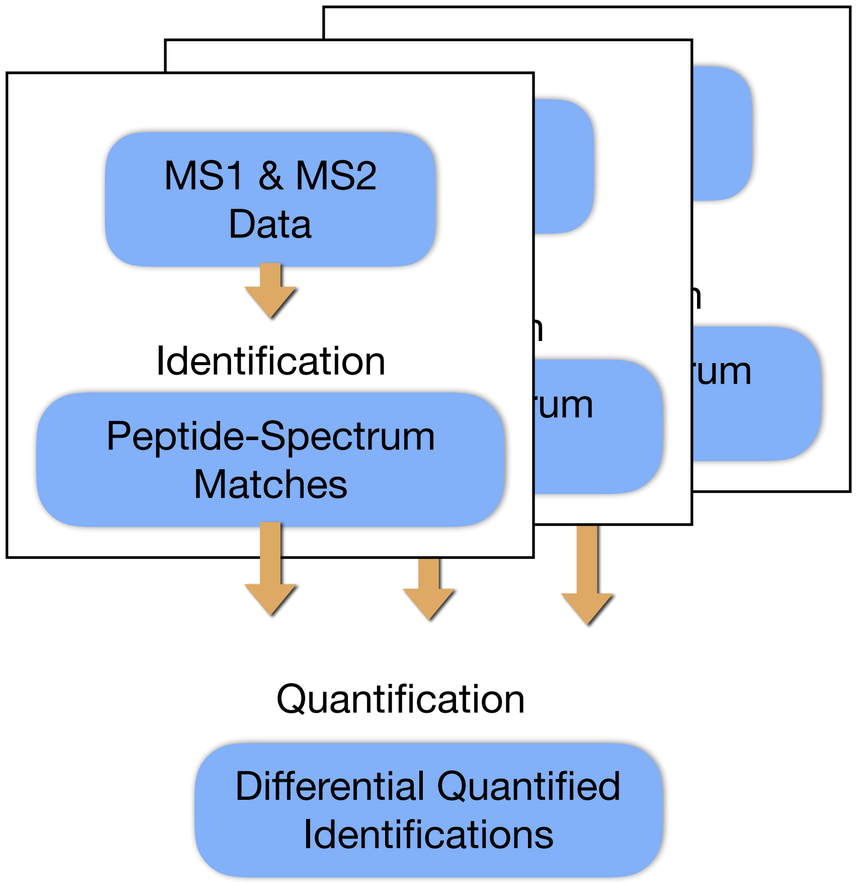
# Flipping the pipeline

*Identification-first*

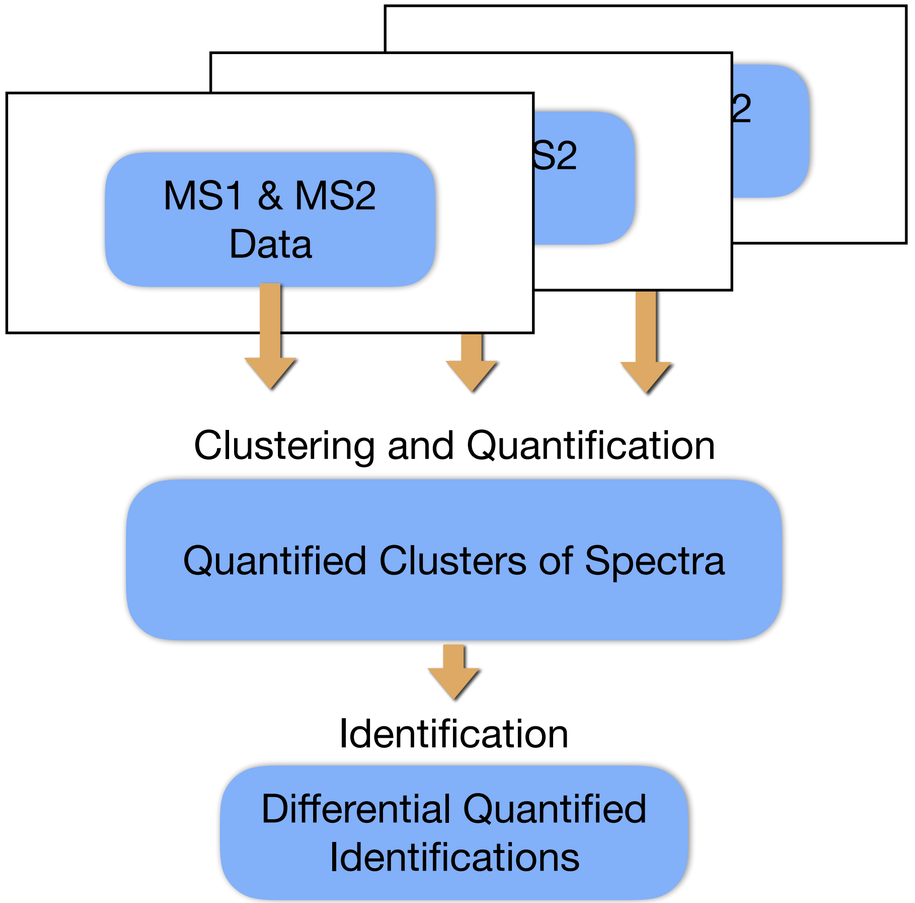


# Flipping the pipeline

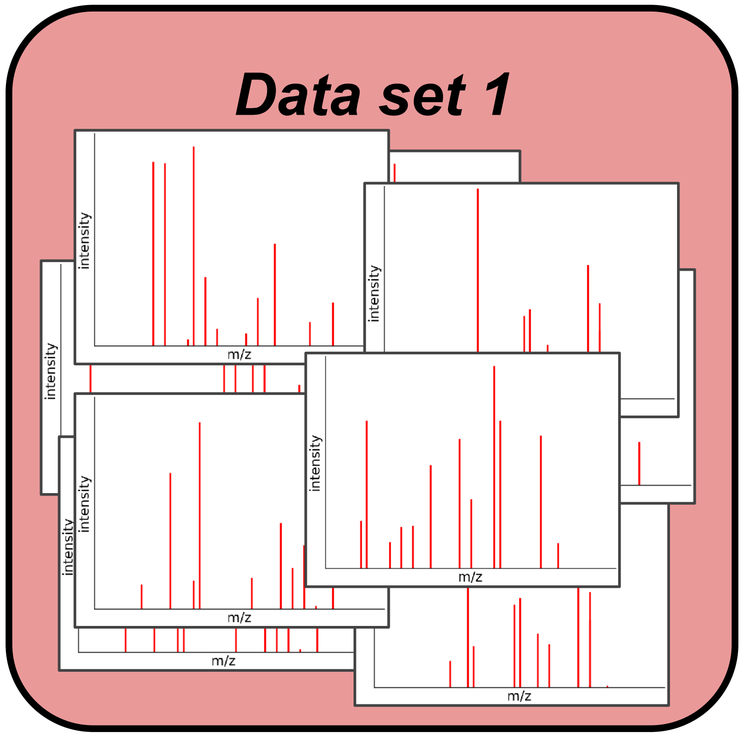
*Identification-first*



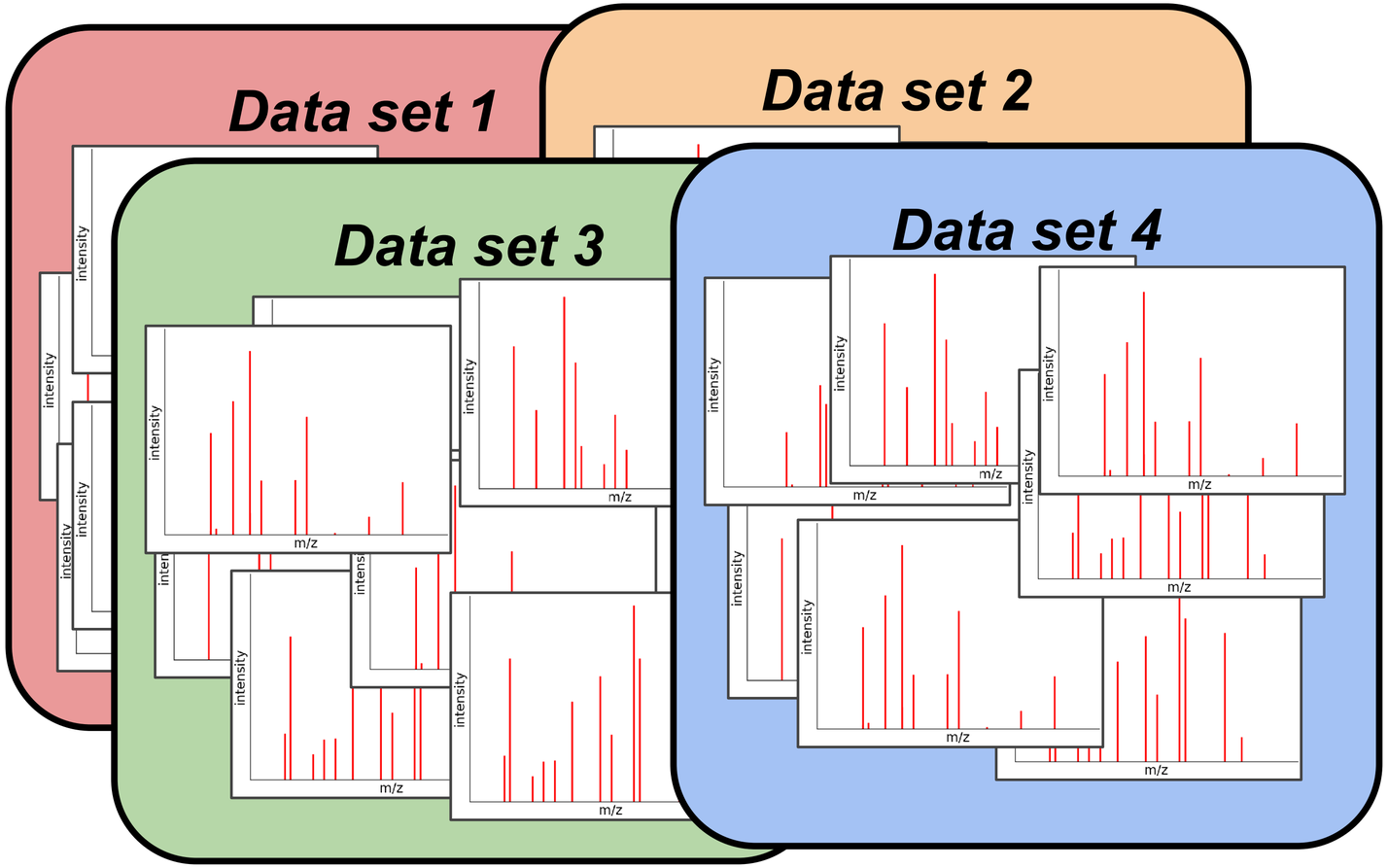

*Quantification-first*



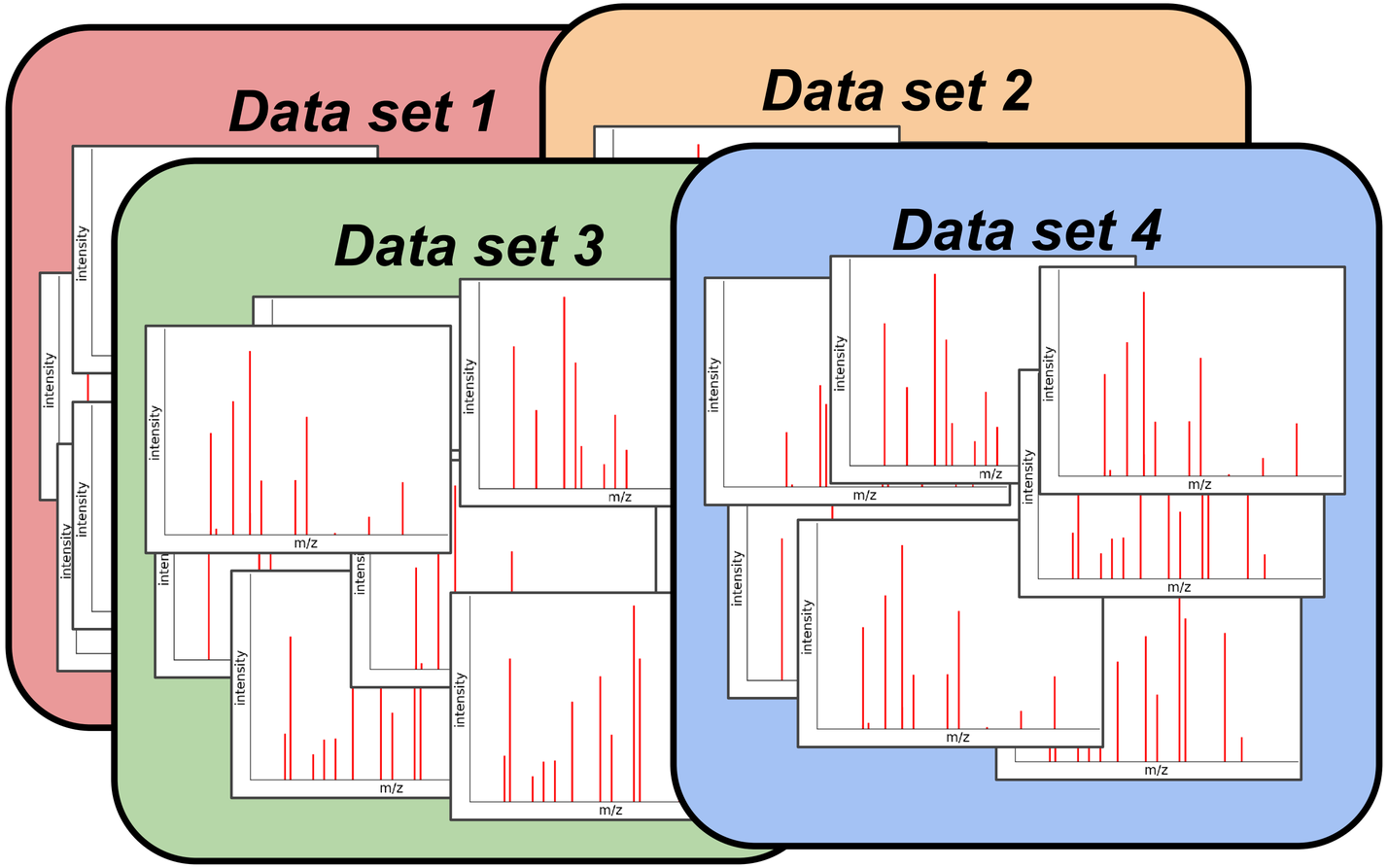
# Finding structure in data by clustering



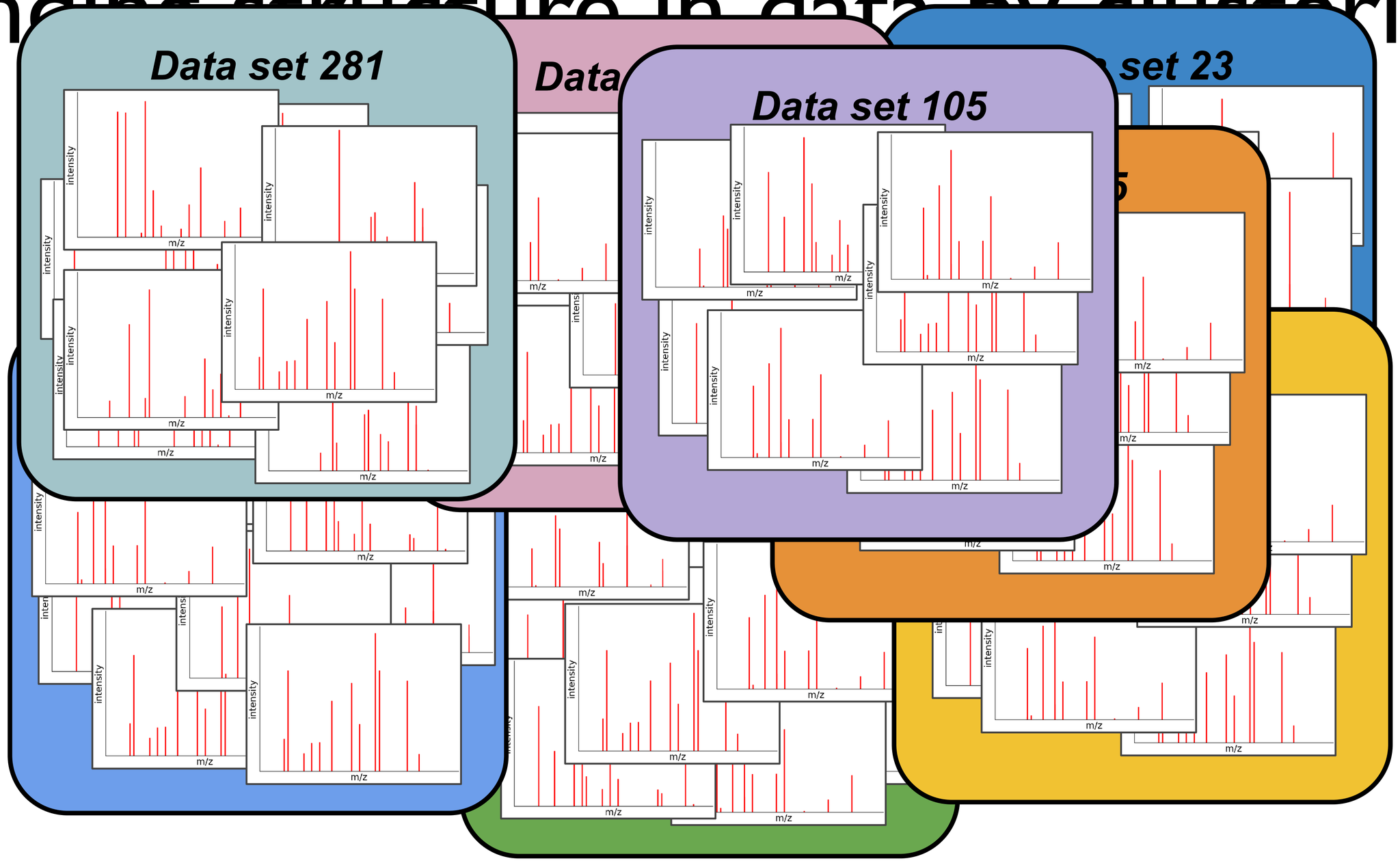
# Finding structure in data by clustering



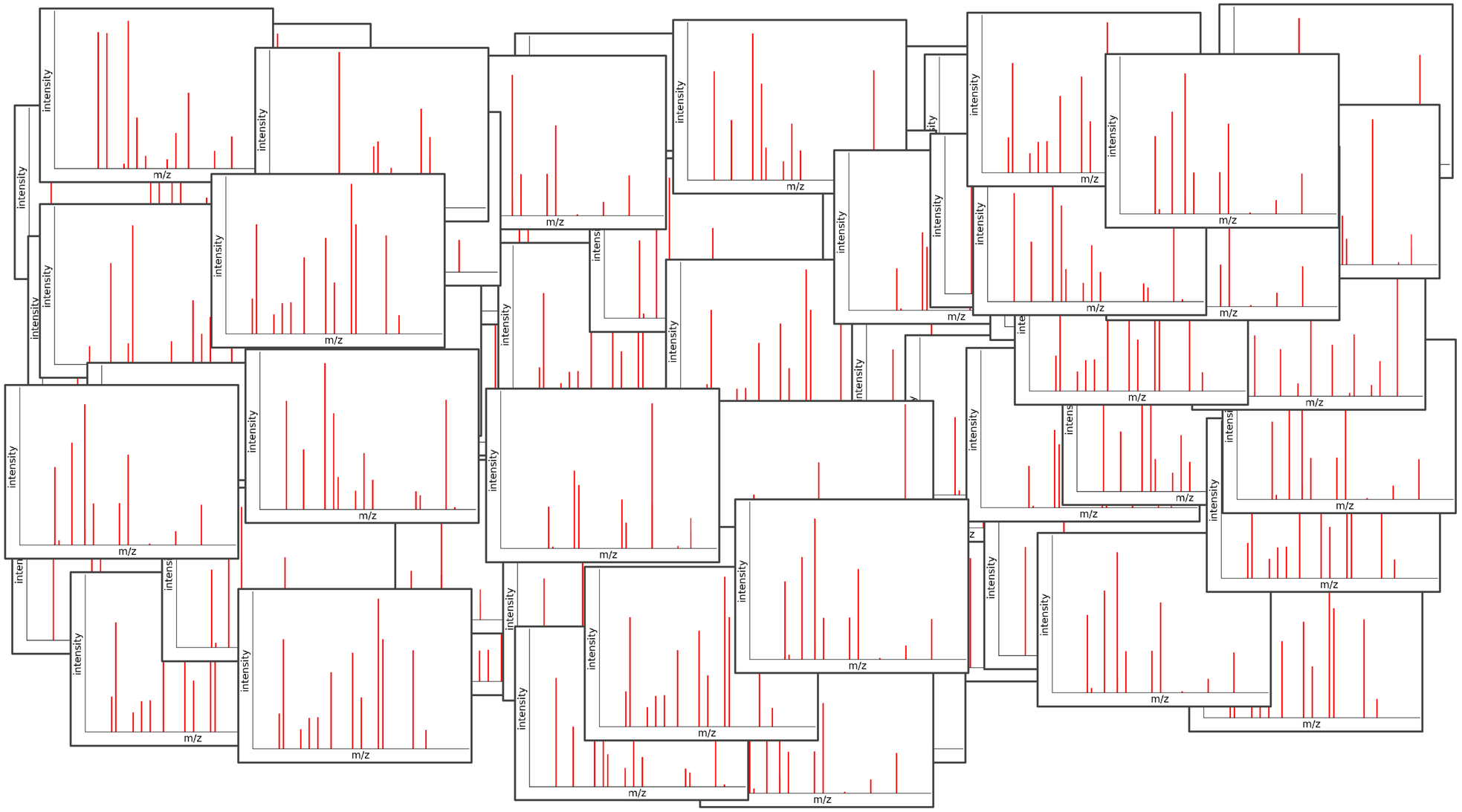
# Finding structure in data by clustering



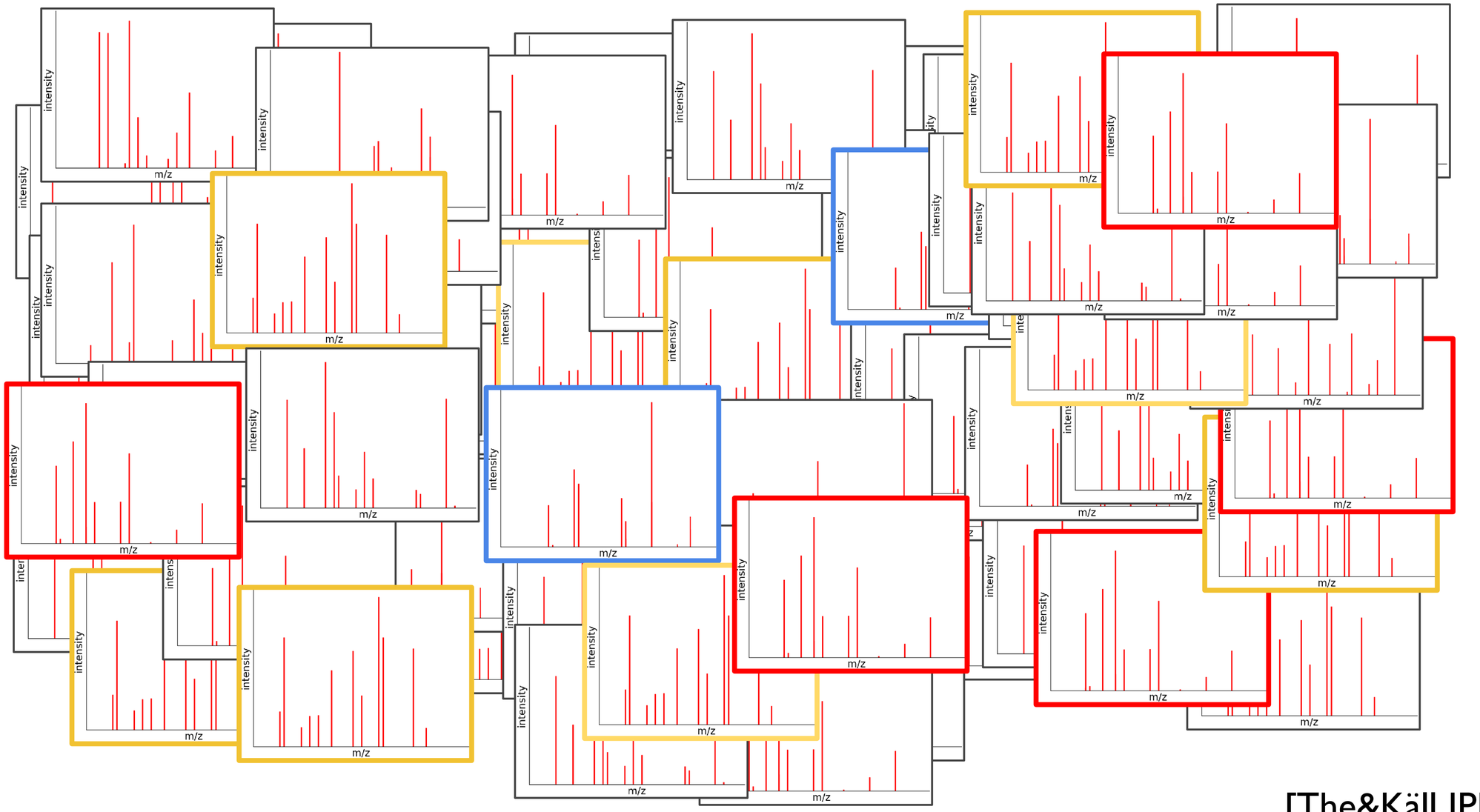
# Finding structure in data by clustering



# Finding structure in data by clustering



# Finding structure in data by clustering



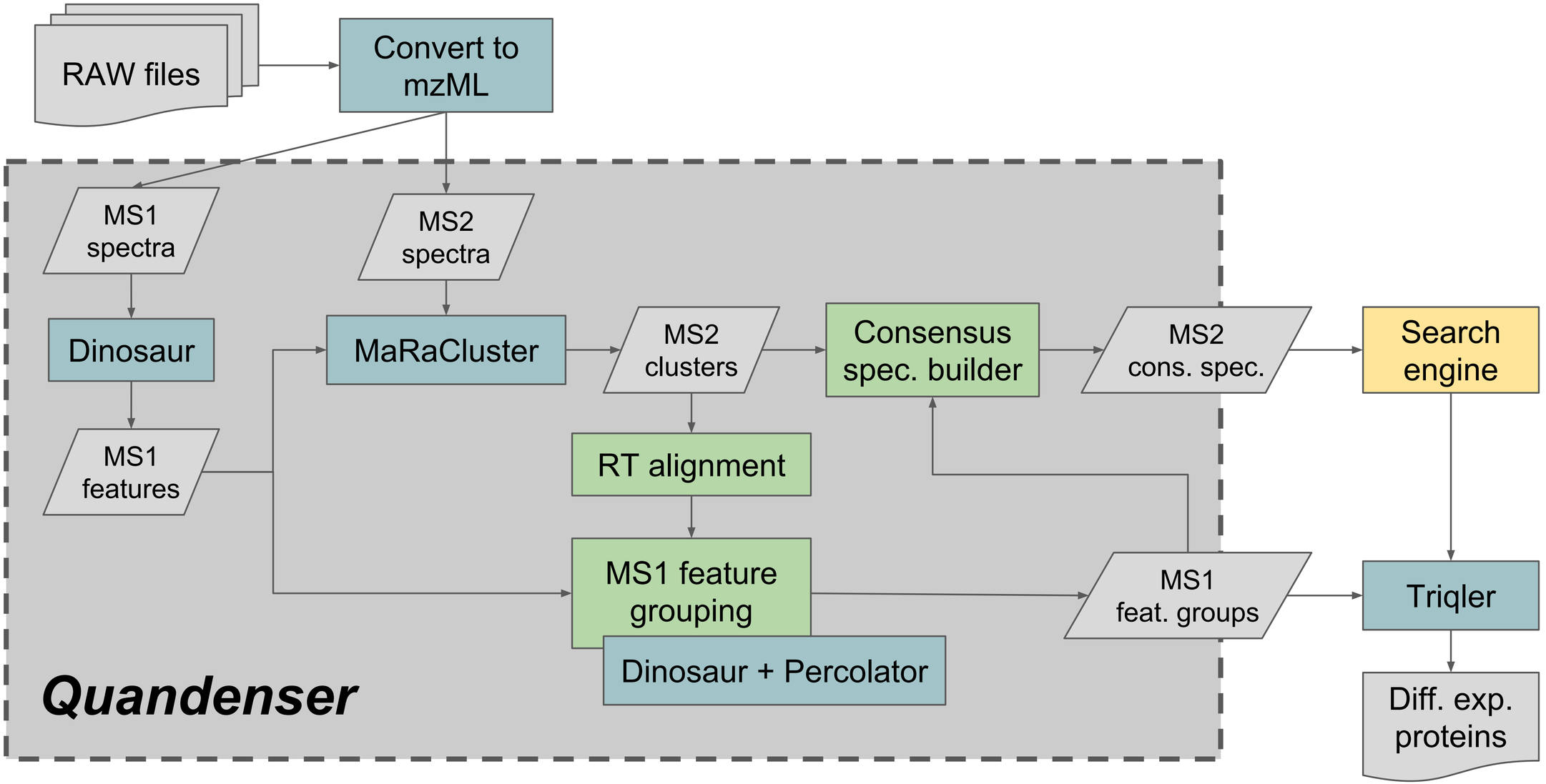

[The&Käll JPR 2015]

# Benefits of quantification-first

- No need to rediscover the same peptides for each run
- Lowering the number of spectra
  - Enables more advanced identification strategies
  - Faster identification
  - Fewer hypothesis tests

Focus on the spectra that matters!

# The Quandenser Workflow



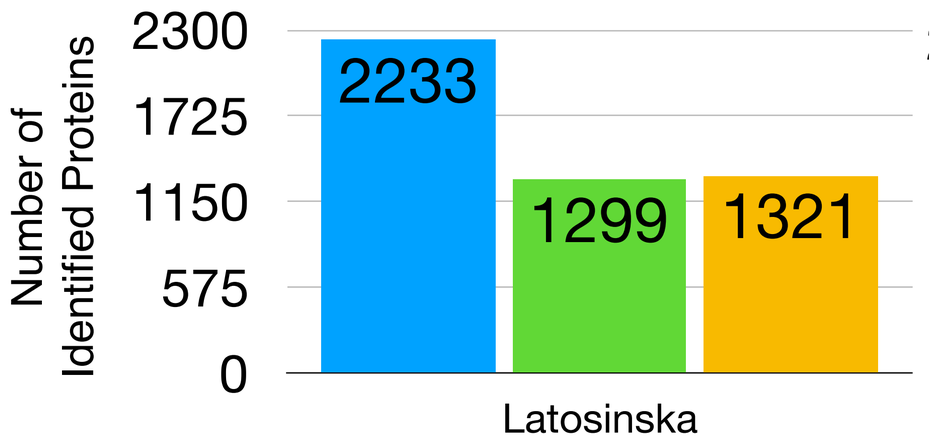
# Results!

# Benchmarking sets

- Latosinska: 4 muscle-invasive vs. 4 non-invasive Bladder Cancers. Latosinska *et al.* (PLoS One 2015) reported 77 proteins as differential with $p<0.05$, 0 after multiple testing correction, (Pride: PXD002170)

- Bracht: 27 severe vs non-severe Hepatitis C Virus-associated hepatic fibrosis. Bracht *et al.* (JPR 2015) reported 70 proteins as differential with $p< 0.05$, 0 after multiple testing correction.  (Pride: PXD001474)

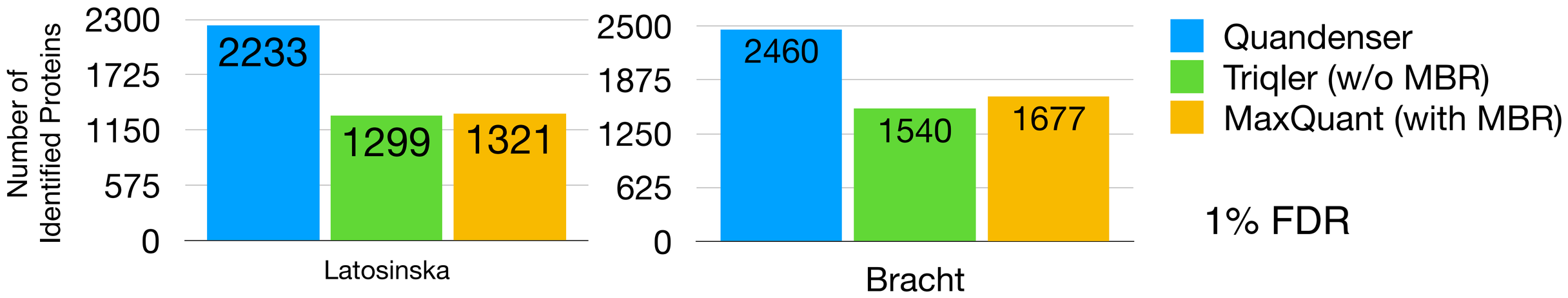
# Quandenser identifies more proteins



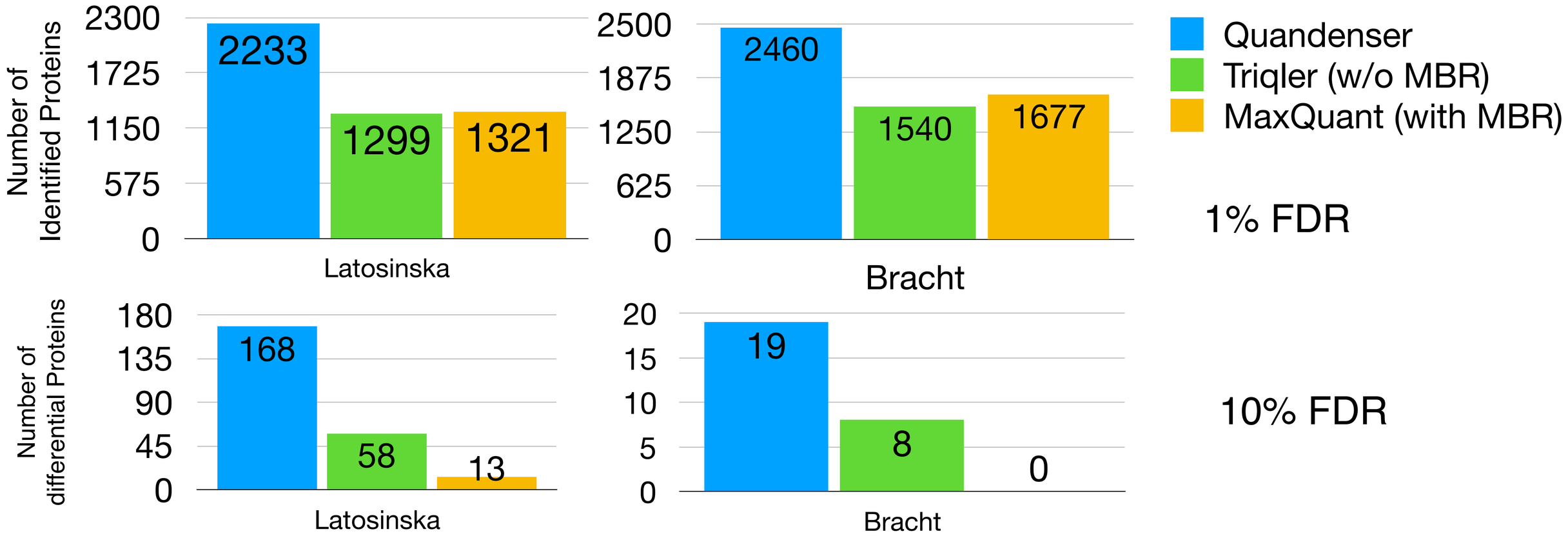

Quandenser
Triqler (w/o MBR)
MaxQuant (with MBR)

1% FDR

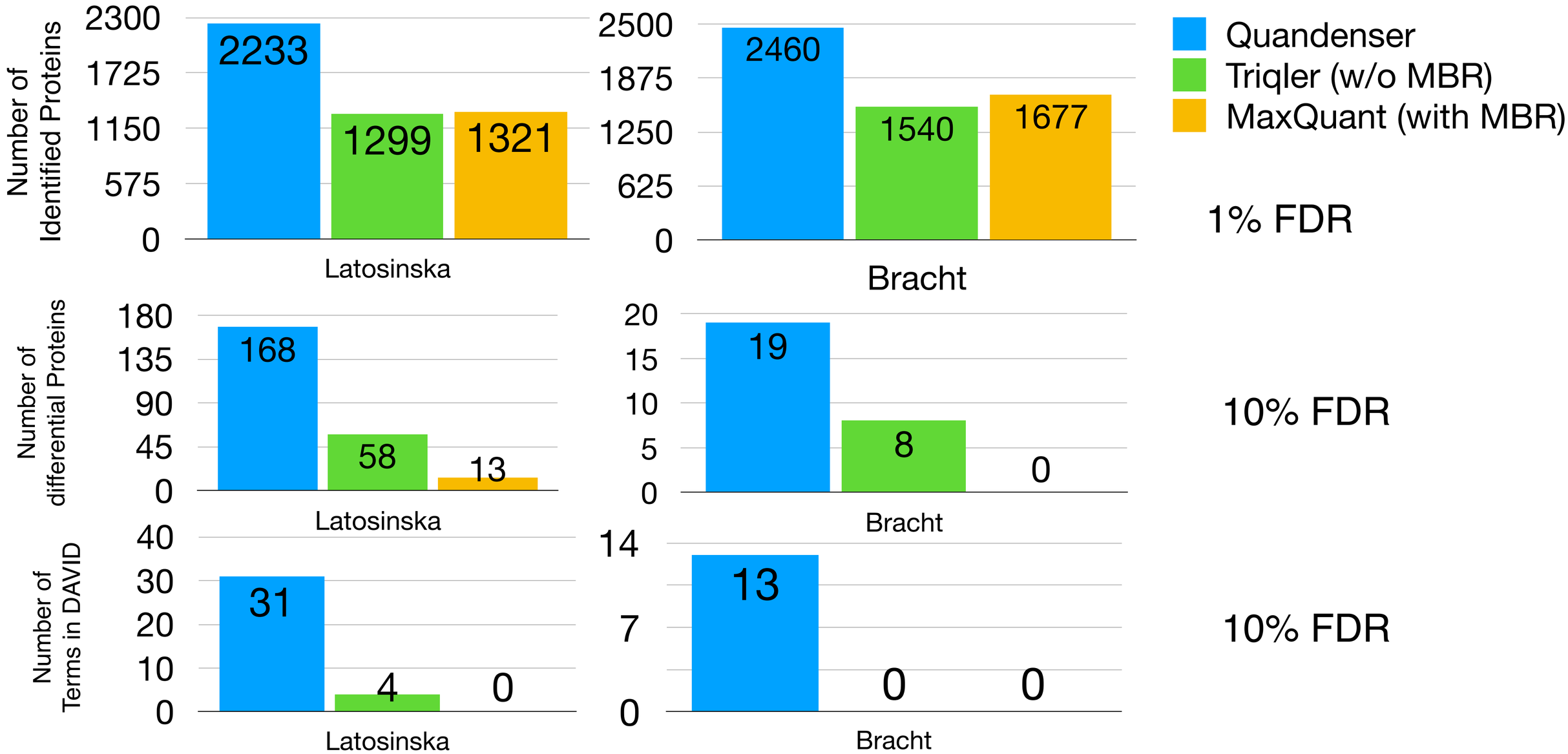
# Quandenser identifies more proteins



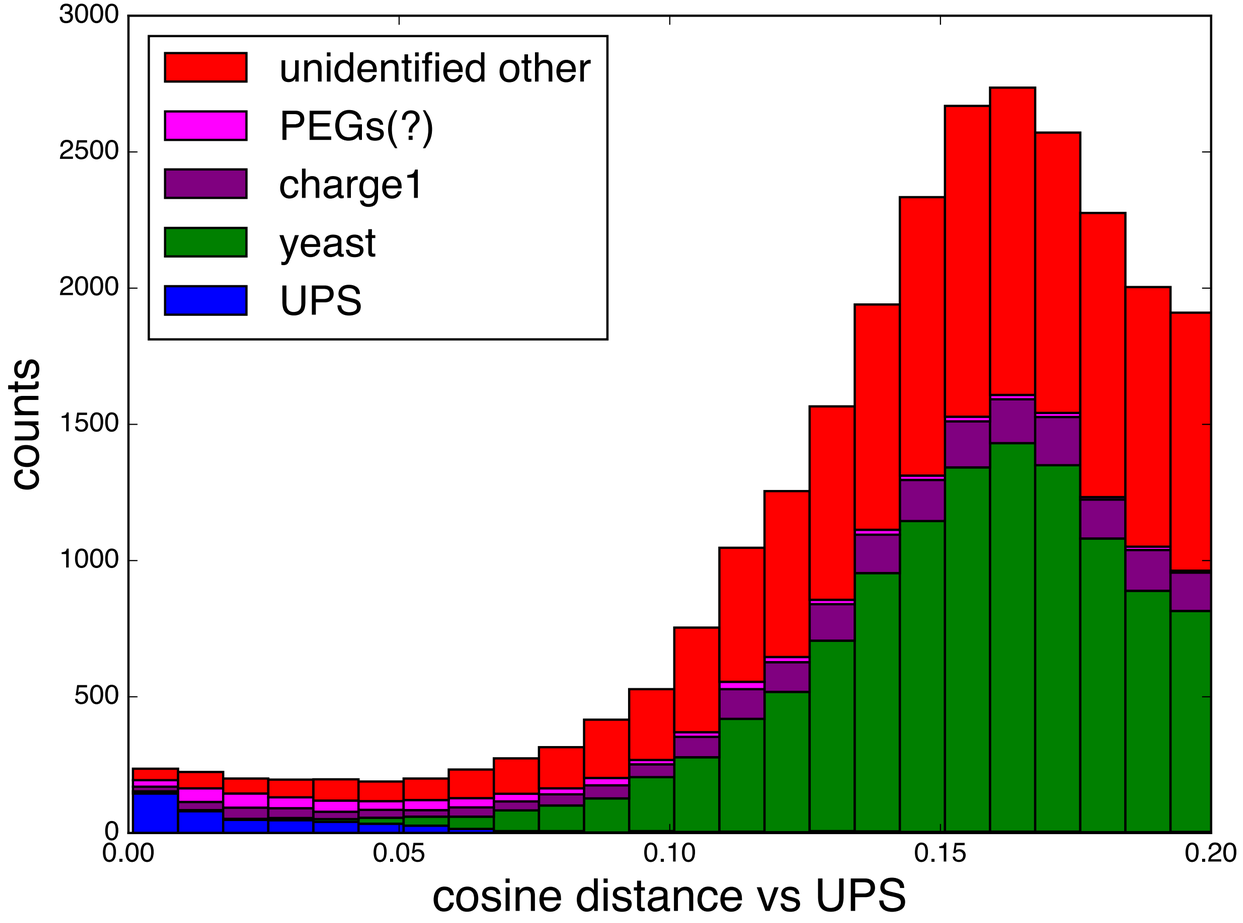
# Quandenser is more sensitive for differential expression

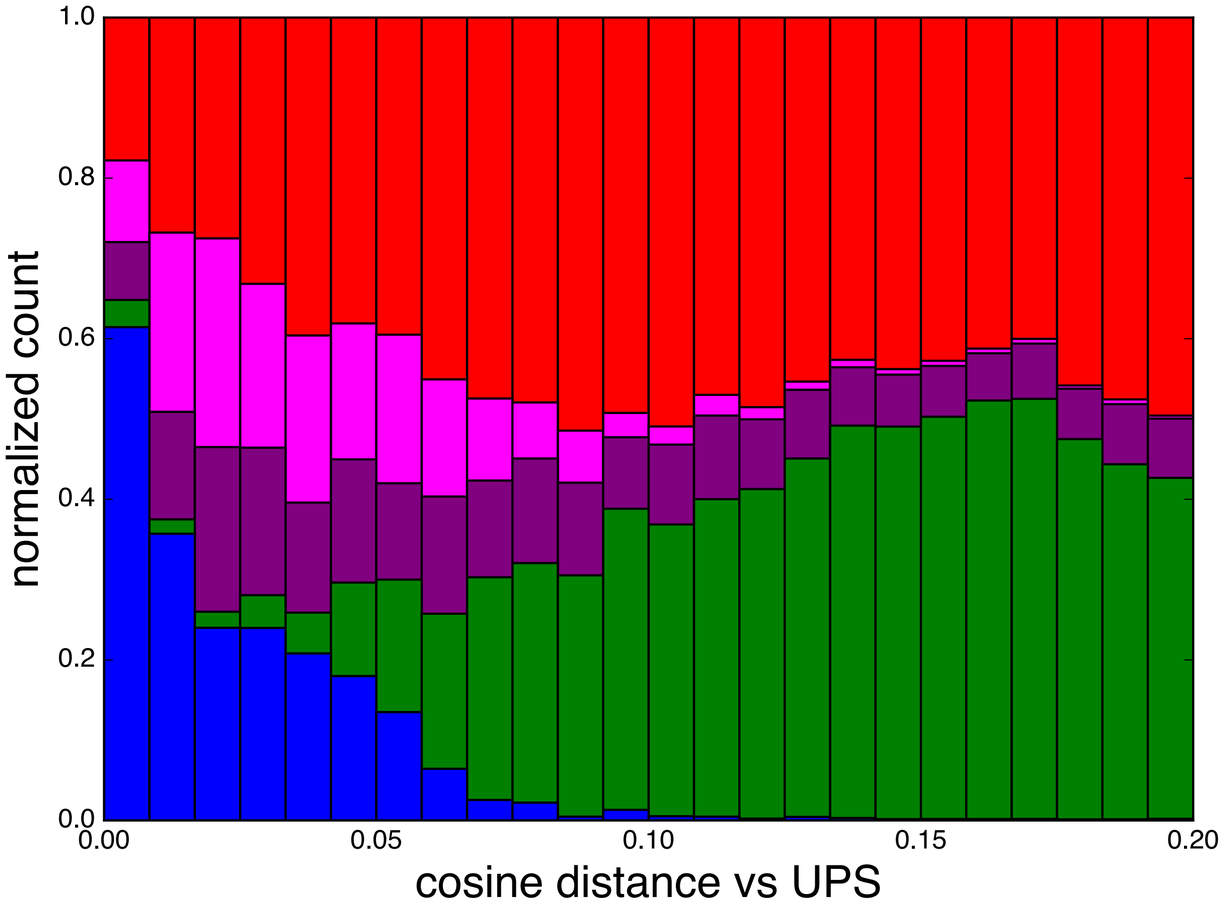


# Quandenser elucidates more biology

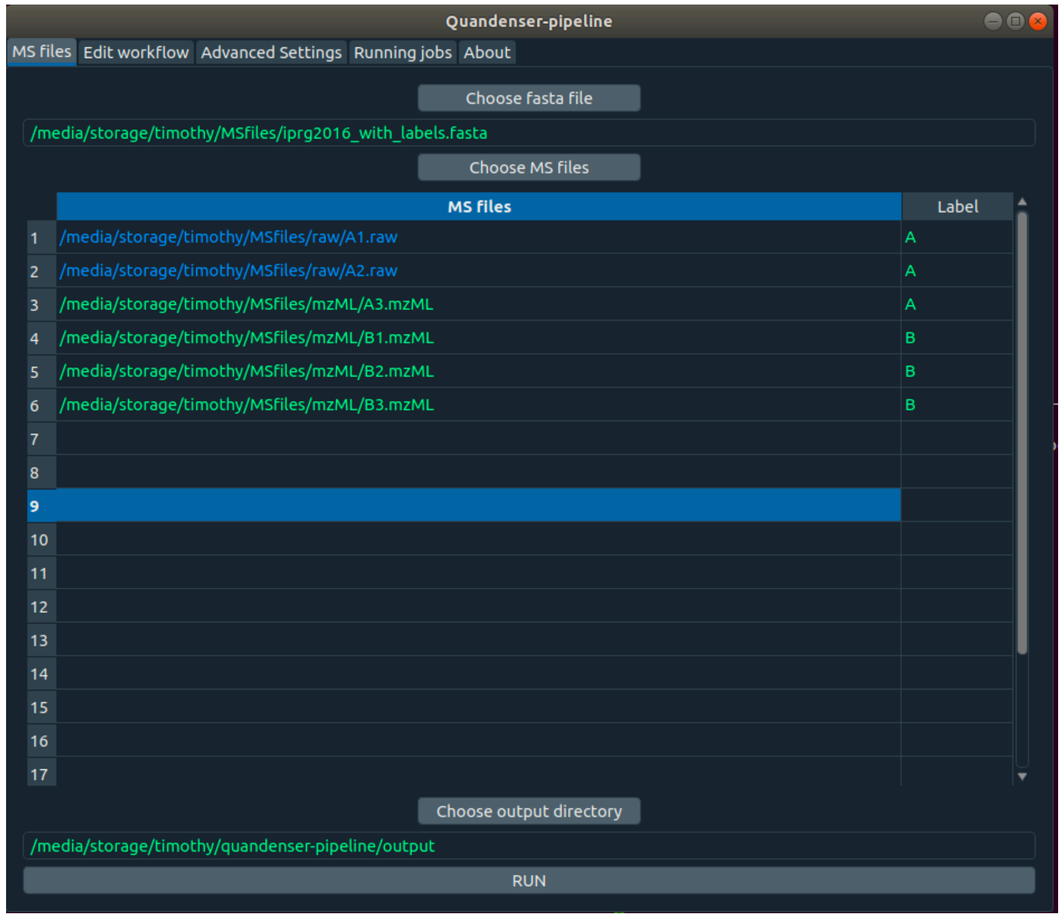


# Identification of unidentified covariate spectra



# Quandenser-pipeline




Input: raw-files or mzML-files

Output: proteins with *Pr(diff. exp.)*

Runs a next flow pipeline of Quandenser through a Singularity container.

Extremely easy to install.

Support for SLURM queue parallelisation.

https://github.com/statisticalbiotechnology/quandenser-pipeline

**Timothy Bergström**

nextflow

# Take-homes

- Methods that are designed for small datasets are not necessarily the best methods for processing larger datasets

# Take-homes

- Methods that are designed for small datasets are not necessarily the best methods for processing larger datasets
- Quandenser is a quantification-first method that outperforms other label-free quantification methods.

# Take-homes

- Methods that are designed for small datasets are not necessarily the best methods for processing larger datasets
- Quandenser is a quantification-first method that outperforms other label-free quantification methods.
- The method is easy to install and available under Apache 2.0 license from:
  `https://github.com/statisticalbiotechnology/quandenser-pipeline`

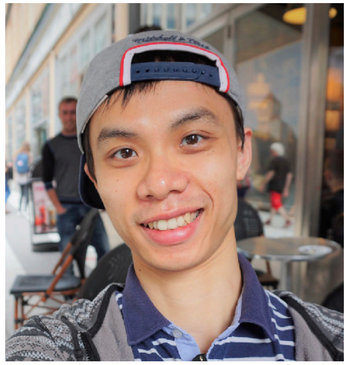
# Thank you!

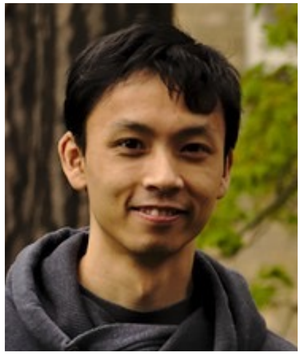

**Statistical Biotechnology, KTH**
Matthew The
Gustavo Jeuken
Patrick Truong
Timothy Bergström
Miriam Bovelett

**Zubarev Lab, Karolinska Institutet**
Roman Zubarev
Bo Zhang

**Calico-labs**
Jonathon O'Brien

**University of British Columbia**
Andrew Roth




Swedish Foundation for Strategic Research

Vetenskapsrådet

SciLifeLab