Statistical Methods for Quantitative Proteomics

Design of Experiments and Interpretation of Results

Meena Choi
Statistics, Purdue University
choi67@purdue.edu

Brendan MacLean
MacCoss lab, Genome Sciences, U. Washington
brendanx@proteinms.net

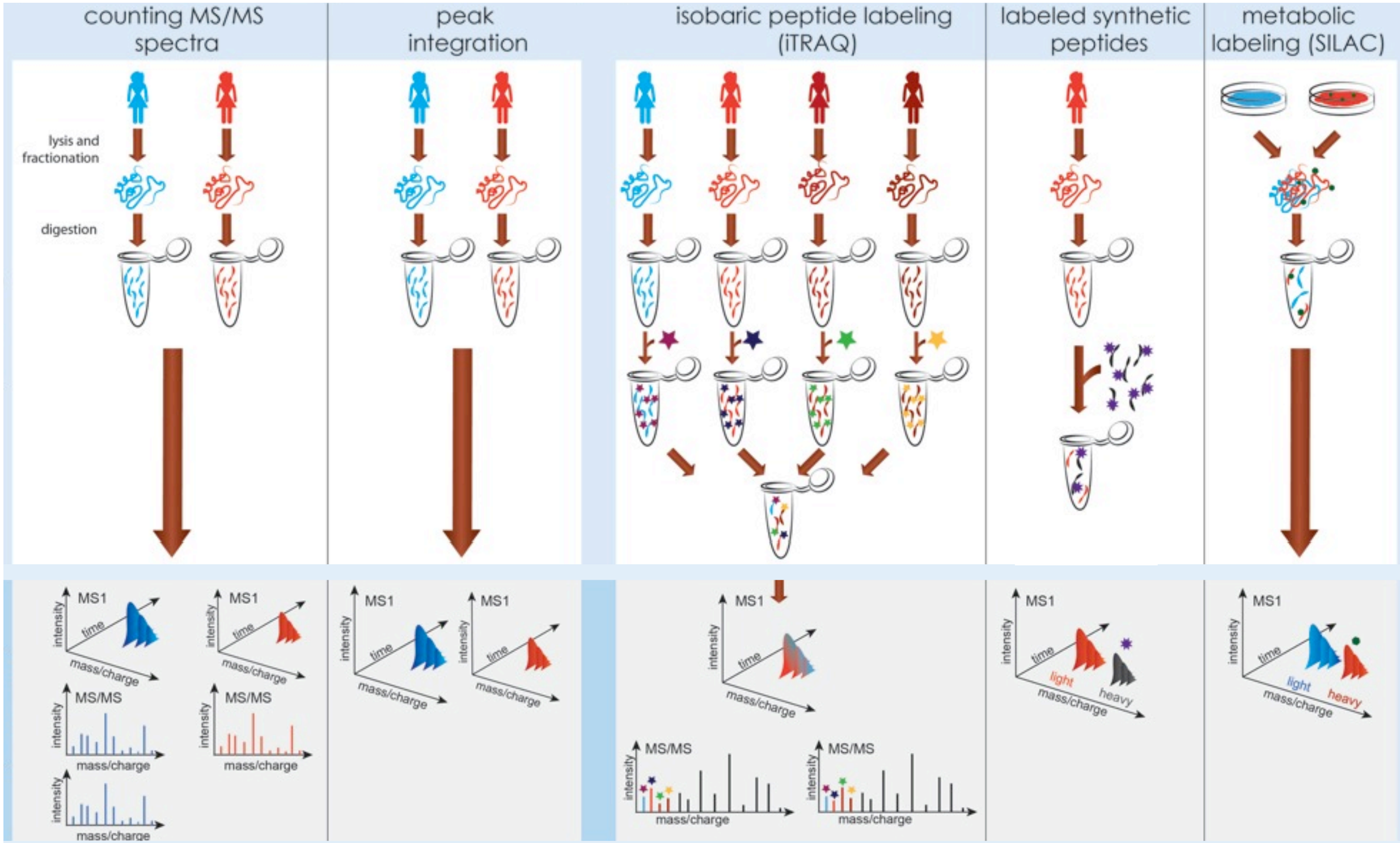
Olga Vitek
Statistics and Computer Science, Purdue University
ovitek@purdue.edu

Quantitative proteomic workflows: global (unbiased)

Label-free *Label-based*

Sample preparation

Global LC-MS/MS



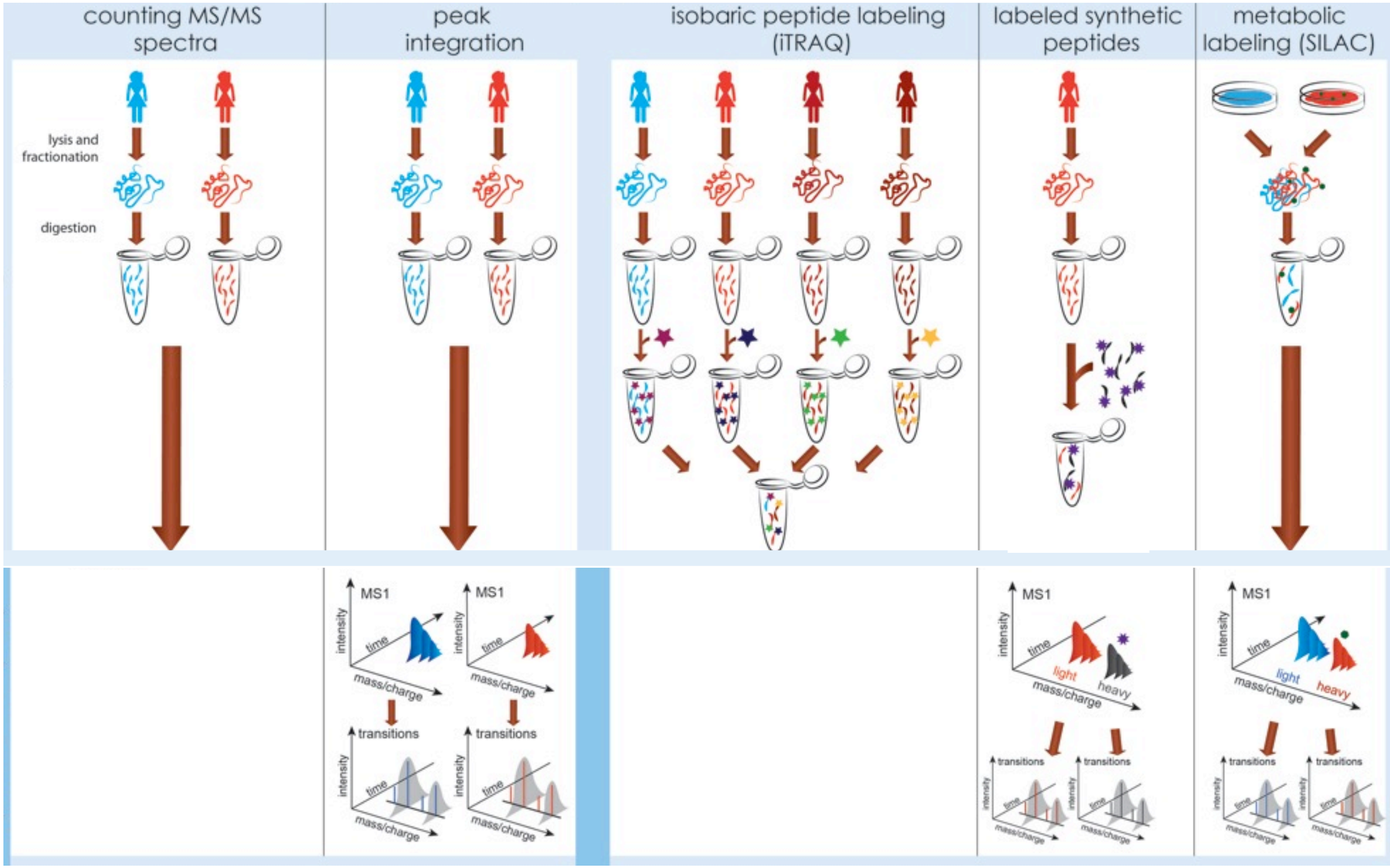
Käll and Vitek, *PLoS Computational Biology*, 7, 2011

Quantitative proteomic workflows: targeted

Label-free *Label-based*

Sample preparation

Targeted SRM



Käll and Vitek, *PLoS Computational Biology*, 7, 2011

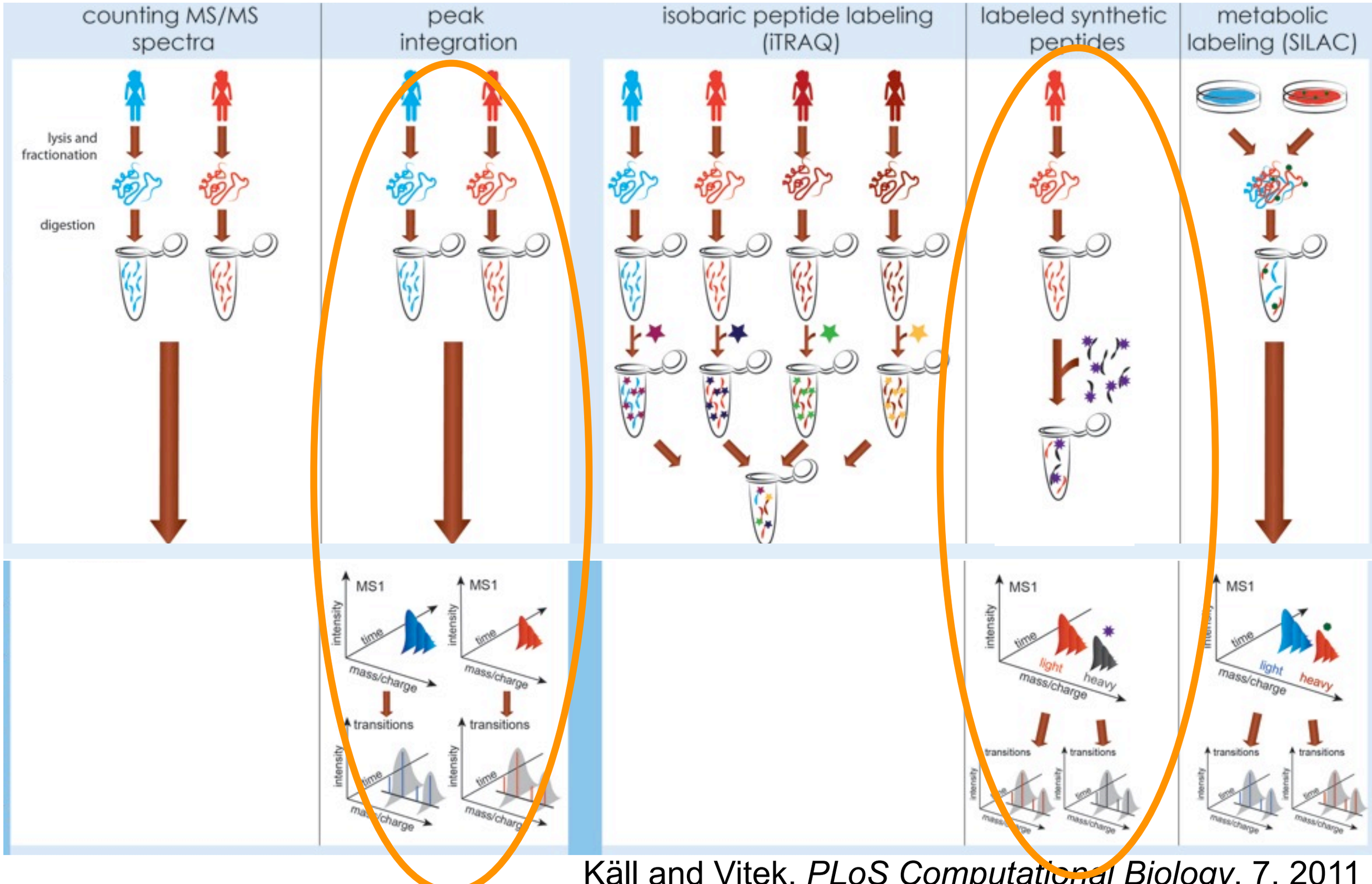
Today: label-based and label-free SRM

But most of the discussion generally applies

Label-free ***Label-based***

Sample preparation

Targeted SRM



Käll and Vitek, *PLoS Computational Biology*, 7, 2011

Scope of discussion: finding differentially abundant proteins

Experimental design, signal processing, significance analysis

- Stochastic variation and uncertainty are unavoidable
  - *Biological variation:* natural variation in protein abundance
  - *Technical variation:* sampling handling, storage, processing
  - *Mass spectrometric variation:* elution time, ion suppression
  - *Signal processing:* ambiguous peak boundaries, identity, intensity

- Statistical reasoning enables efficient, reproducible research
  - *Experimental design:* unbiased and resource-efficient experiments
  - *Data analysis:* objective conclusions in presence of uncertainty
  - *Statistical tools:* re-analysis, peer review, reproducibility

Plan for the day

- **Morning**
  - 9:00am-10:00am Olga: Statistical experimental design
  - 10:00am-10:30am Brendan: Data processing with Skyline
  - 10:30am-11:00am *Refreshments*
  - 11:00am-12:00pm Brendan: Data processing with Skyline
- **Afternoon**
  - 1:00pm-2:00pm Olga: Statistical significance analysis
  - 2:00pm-2:30pm Meena: Statistical analysis case studies
  - 2:30pm-3:00pm *Refreshments*
  - 3:00pm-4:00pm Meena: Statistical analysis case studies

Steps of statistical experimental design

- Define the problem
  - Populations of interest
  - Comparisons of interest
  - Scope of conclusions
- Utilize 3 principles of experimental design
  - Replication
  - Randomization
  - Blocking: known biological and technical variation
  - Blocking: MS run

Motivating example: a case study of coronary artery disease

- Collection of plasma samples of 3290 disease subjects and controls
  - treated at the Munich Heart Center between 2005 and 2006
  - collected at single time point at diagnosis
  - recorded clinical characteristics
- Focus on 5 disease groups
  - *STEMI*, *NSTEMI*, *unstable angina*, *stable angina*, *controls*
- General goal: an initial quantitative LC-MS screening
  - select a subset of plasma samples
  - examine protein profiles
  - a follow-up study will focus on a subset of proteins and disease groups

  - Clough et al. *Methods in Molecular Biology*, 2011

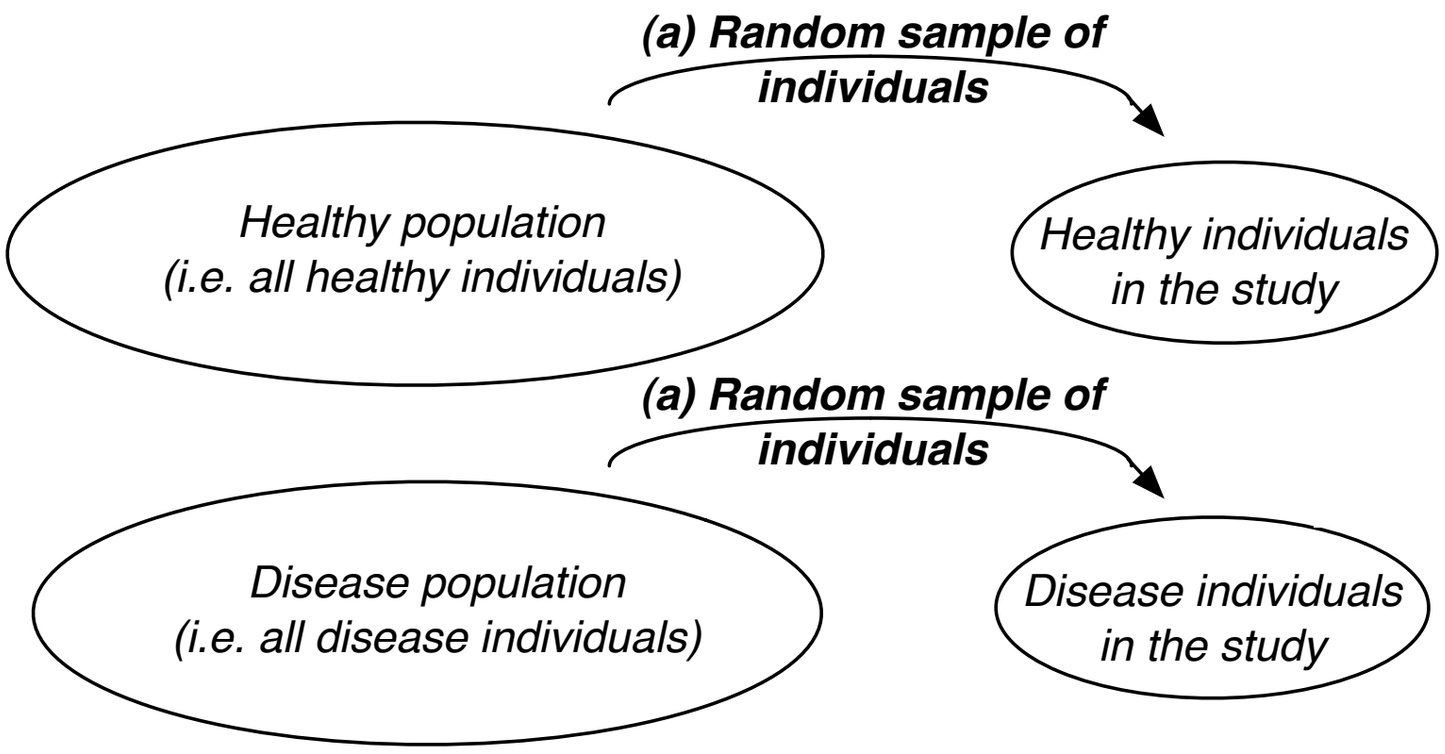
Here is how a statistician views this experiment

Healthy population
(i.e. all healthy individuals)

Disease population
(i.e. all disease individuals)

Large populations of individuals

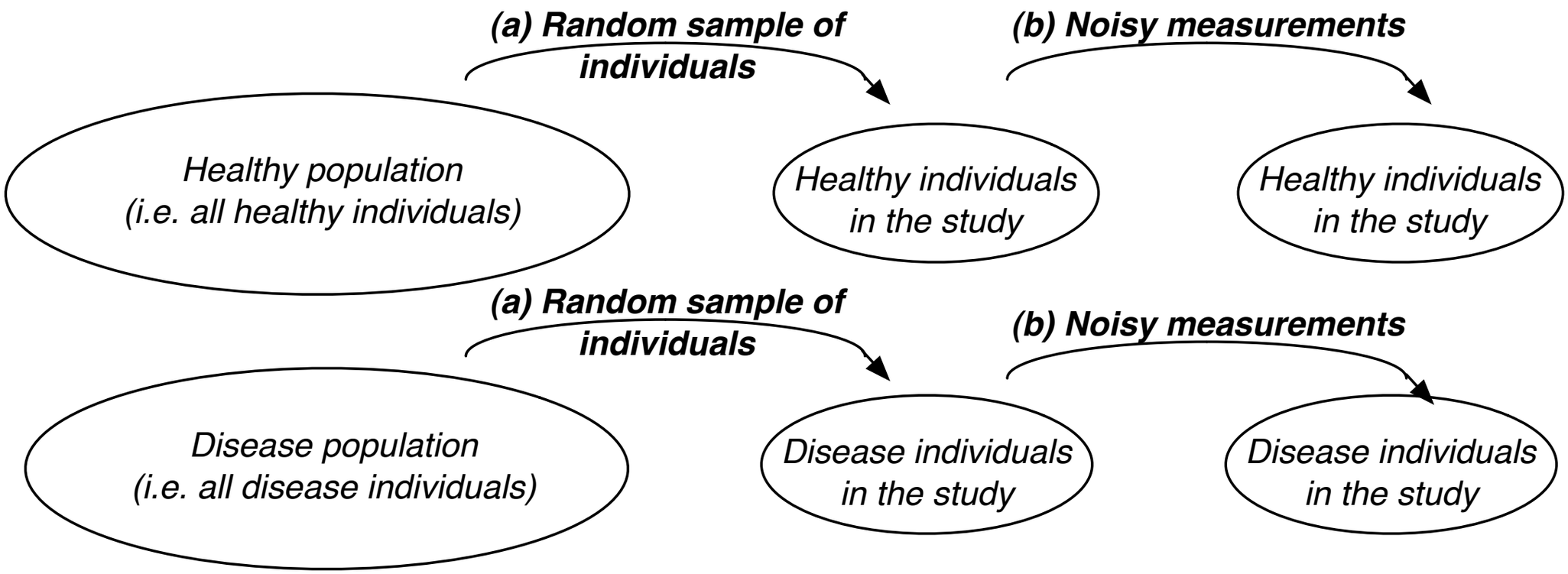
Here is how a statistician views this experiment



Large populations of individuals

Randomly selected individuals

Here is how a statistician views this experiment



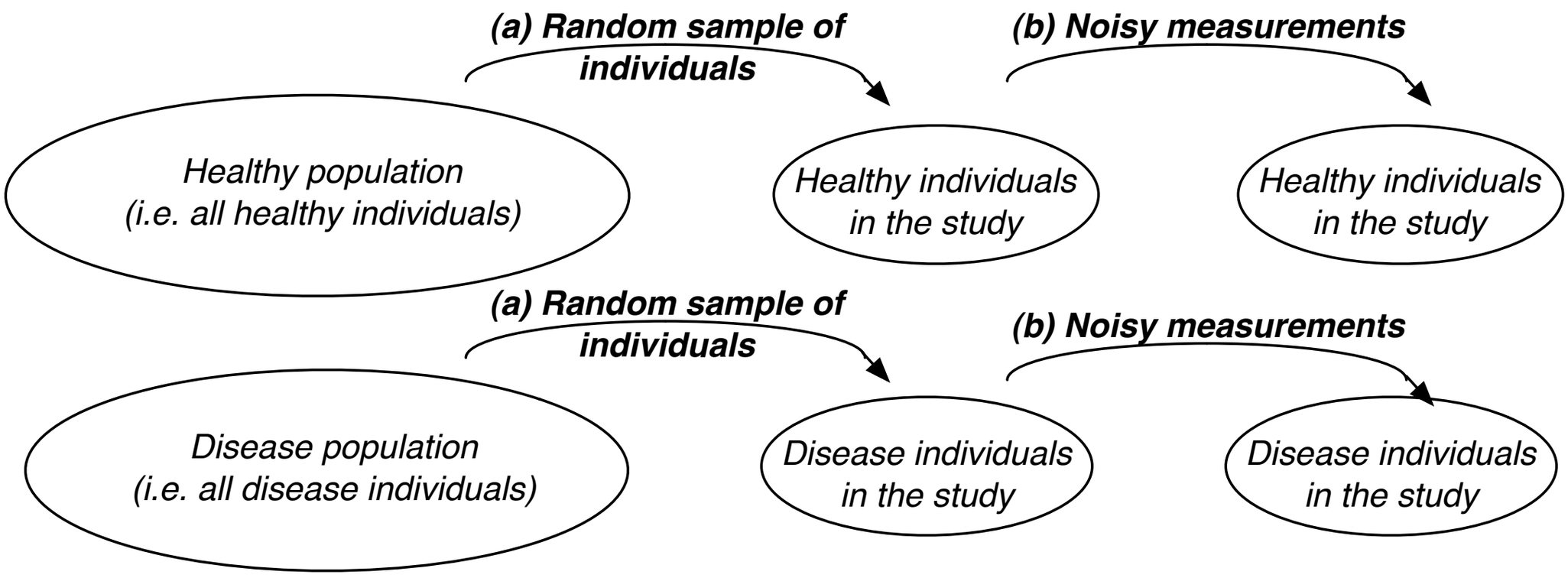
Large populations of individuals

Randomly selected individuals

Noisy measurements on selected individuals

Here is how a statistician views this experiment

Define the problem: *Which conditions to compare?*
Which subjects to compare?



Test:
mean of all disease patients
=
mean of all control patients

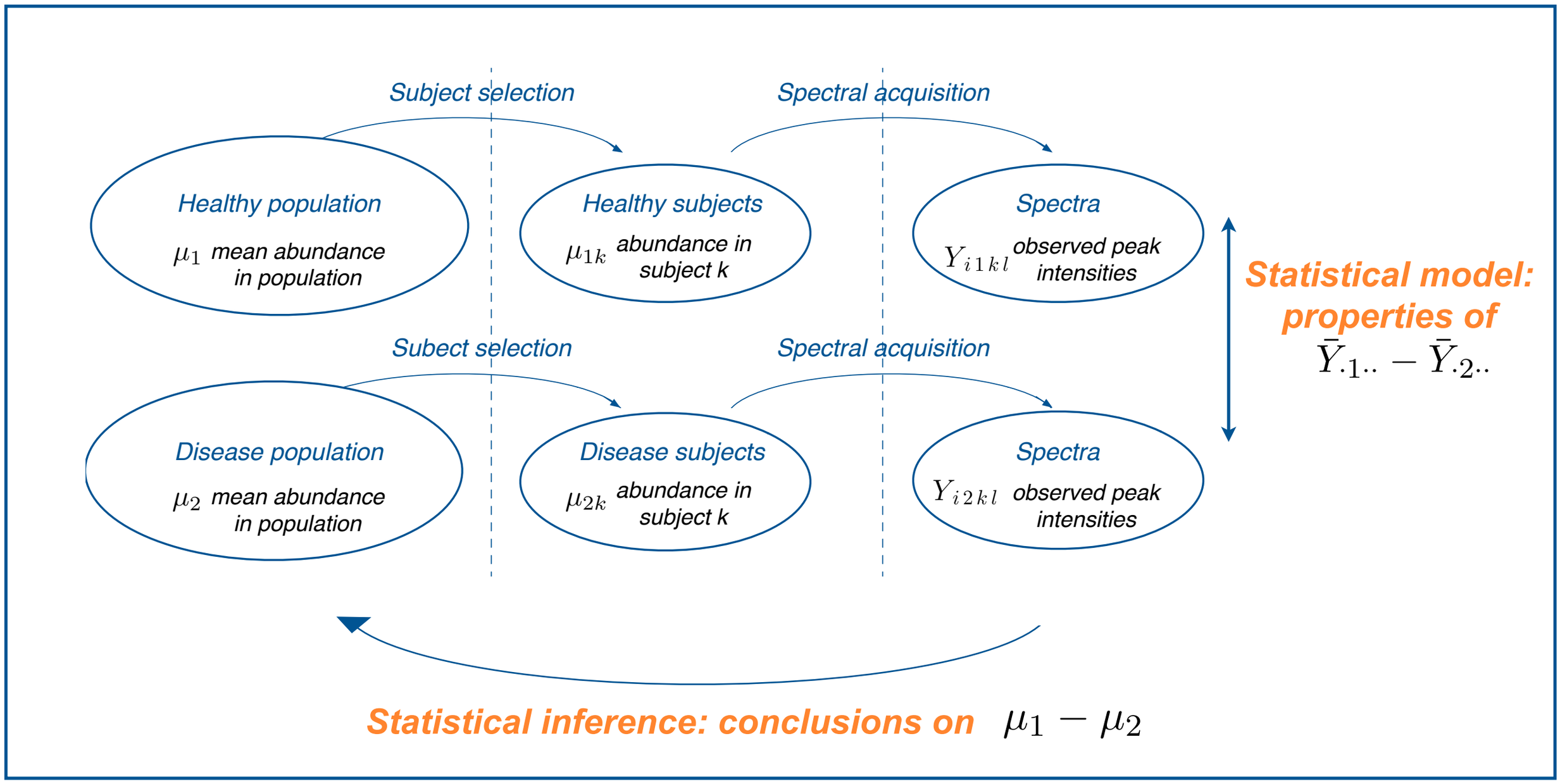
Useful for validation experiments

Test:
mean of selected disease patients
=
mean of selected control patients

Useful for screening experiments

← *Scope of conclusions*

Here is how a statistician would use the data to perform the comparisons

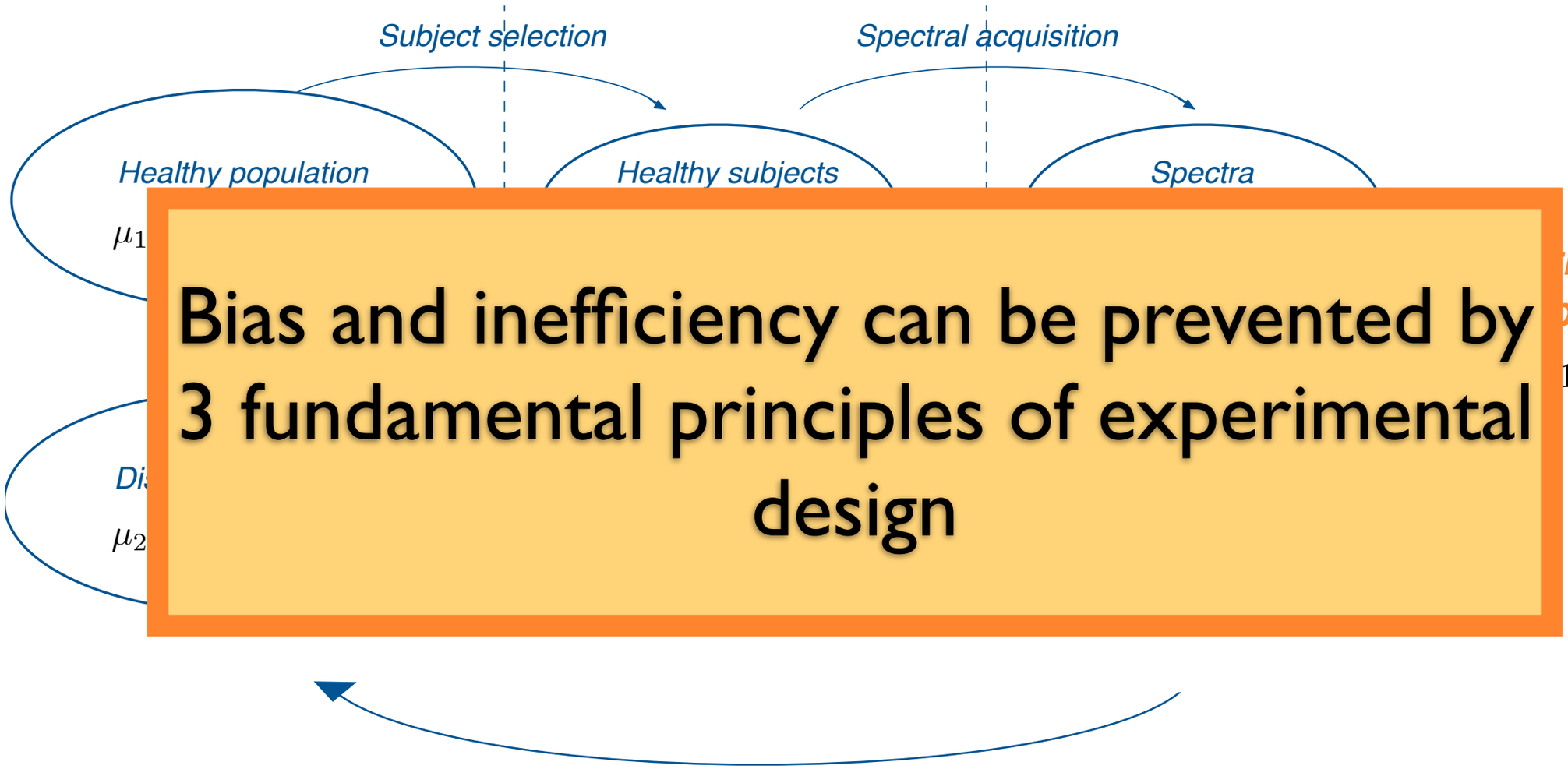


Potential dangers:

Bias: $\bar{Y}_{.1..} - \bar{Y}_{.2..}$ *systematically different from* $\mu_{1k} - \mu_{2k}$

Inefficiency: *Large* $Var(\bar{Y}_{.1..} - \bar{Y}_{.2..})$

Here is how a statistician would use the data to perform the comparisons



ical model:
perties of
$\bar{Y}_{1..} - \bar{Y}_{2..}$

Potential dangers:

Bias: $\bar{Y}_{.1..} - \bar{Y}_{.2..}$ *systematically different from* $\mu_{1k} - \mu_{2k}$

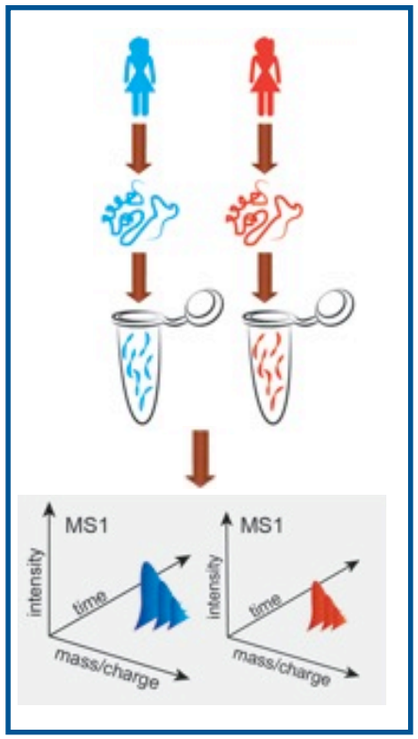
Inefficiency: *Large* $Var(\bar{Y}_{.1..} - \bar{Y}_{.2..})$

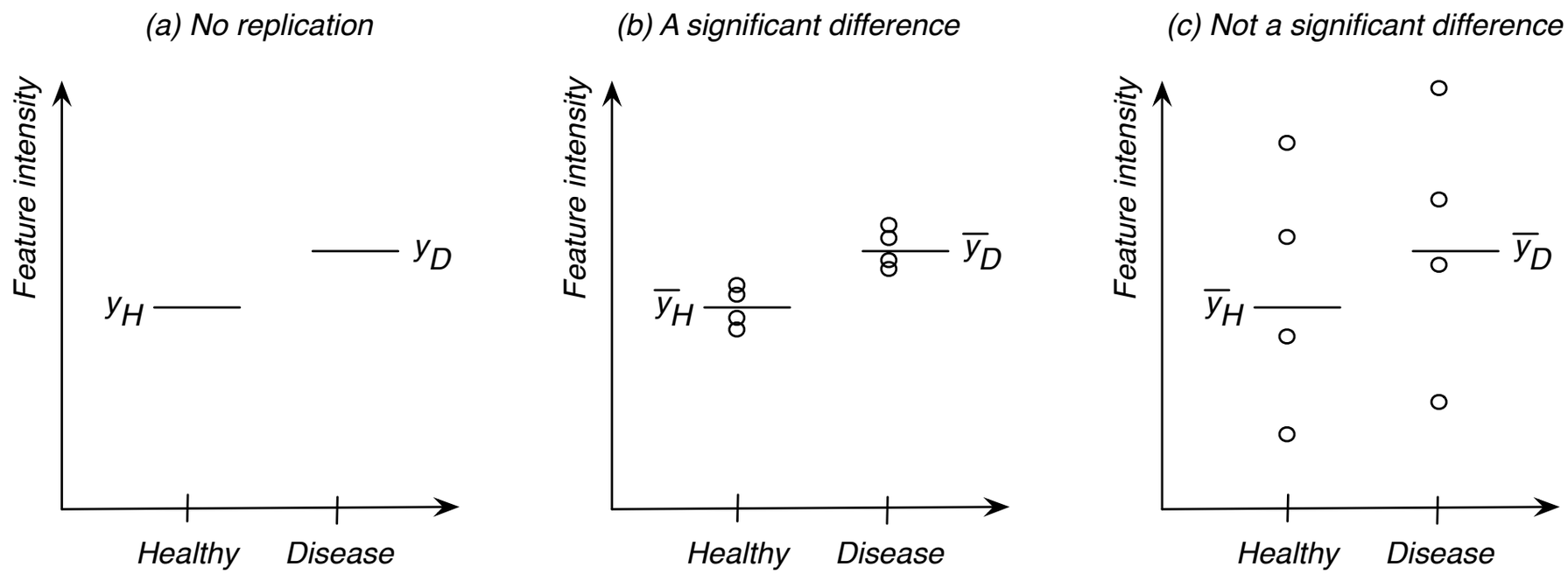
Steps of statistical experimental design

- Define the problem
  - Populations of interest
  - Comparisons of interest
  - Scope of conclusions
- Utilize 3 principles of experimental design
  - Replication
  - Randomization
  - Blocking: known biological and technical variation
  - Blocking: MS run

Fundamental principle 1: replication

Required to (1) carry out the inference and (2) minimize the variance





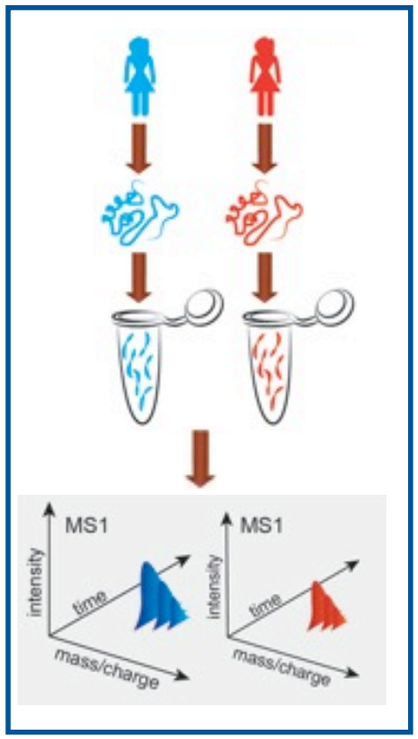
Two levels of randomness imply two types of replication:

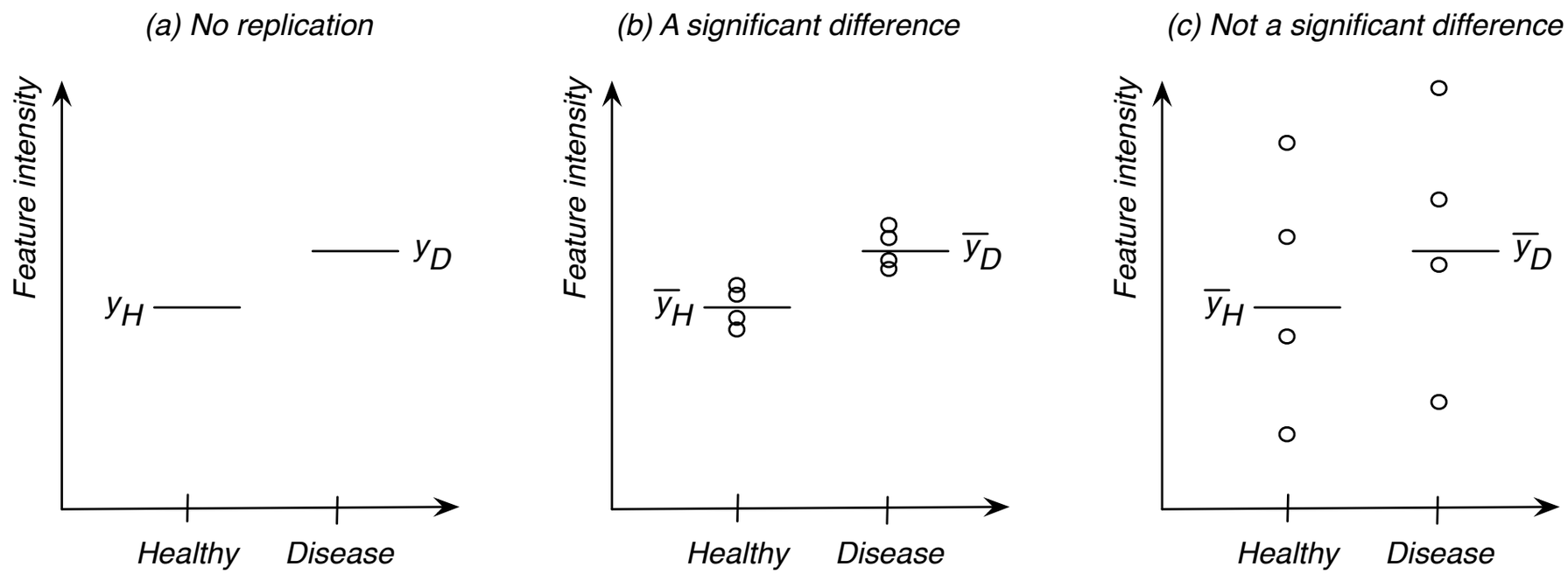
- *Biological replicates:* selecting multiple subjects from the population
- *Technical replicates:* multiple runs per subject

Oberg and Vitek, *J. Proteome Research*, 8, 2009

Fundamental principle 1: replication

Required to (1) carry out the inference and (2) minimize the variance



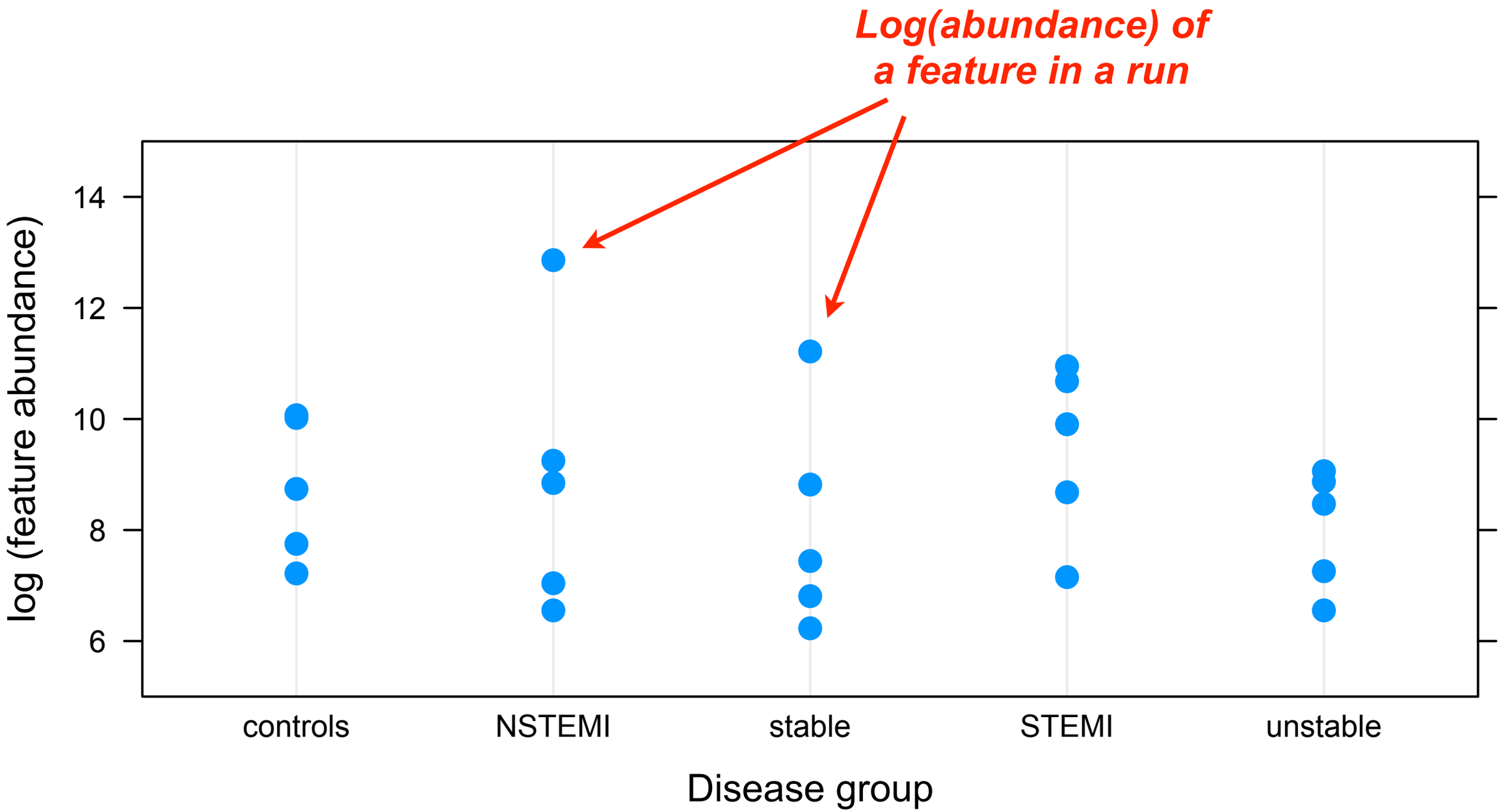


Coronary artery disease experiment:

- *Biological replicates:* 50 subjects per disease group from the population
- *Technical replicates:* no technical replication in this case

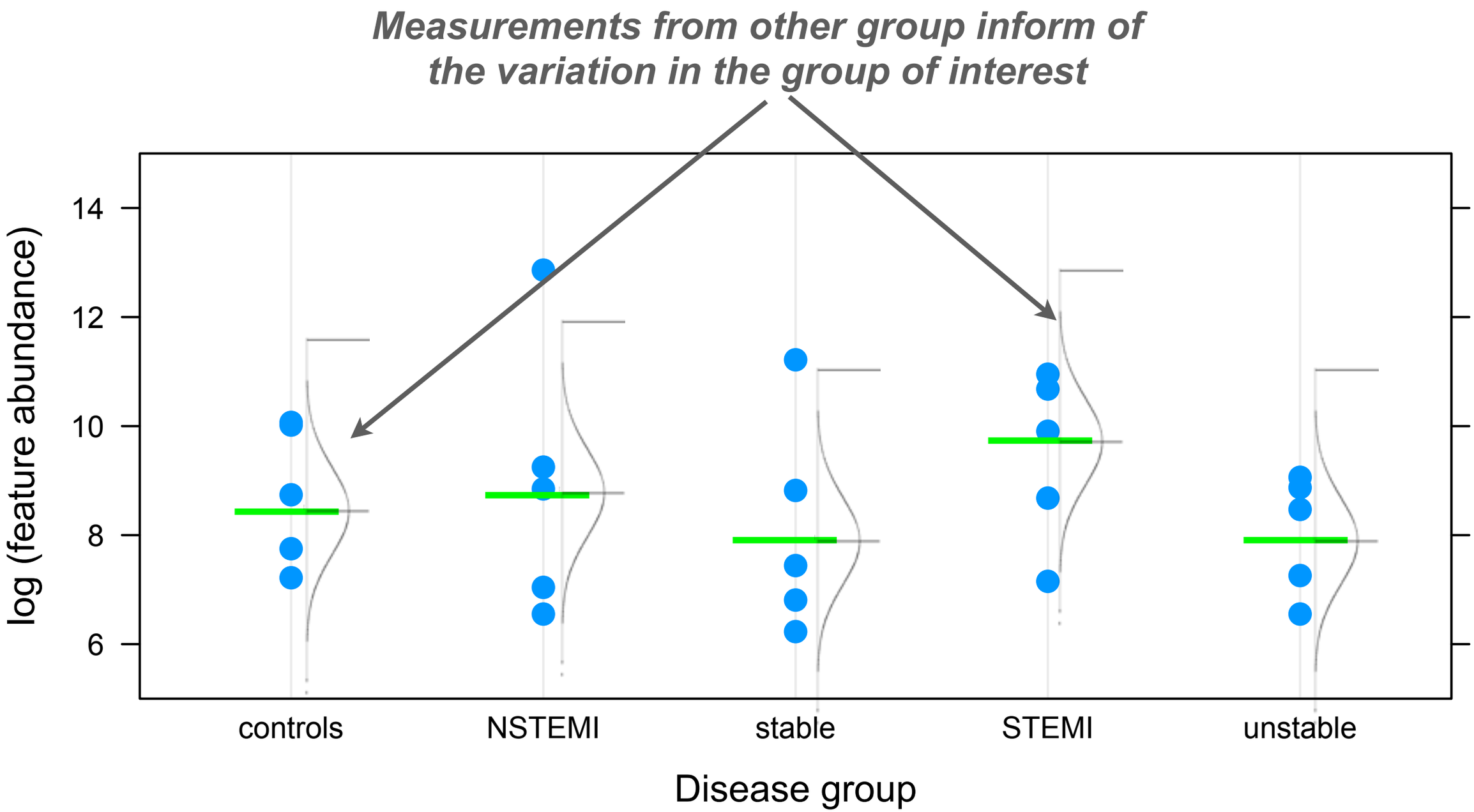
Oberg and Vitek, *J. Proteome Research*, 8, 2009

Jointly analyzing multiple conditions effectively increases the number of replicates



Often can assume that the variation is same across groups

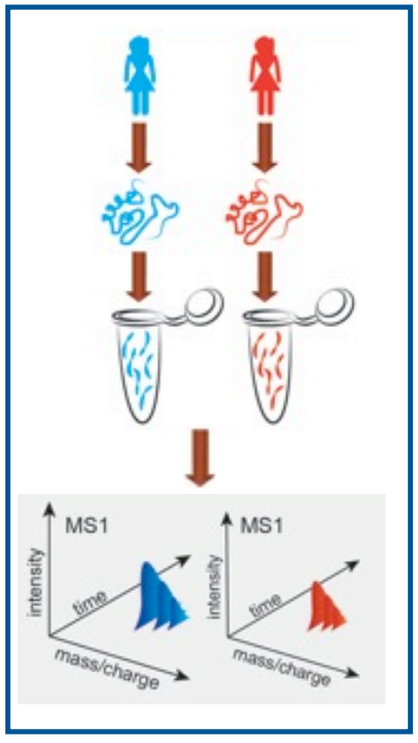
Does not need to be constant (e.g. function of intensity)

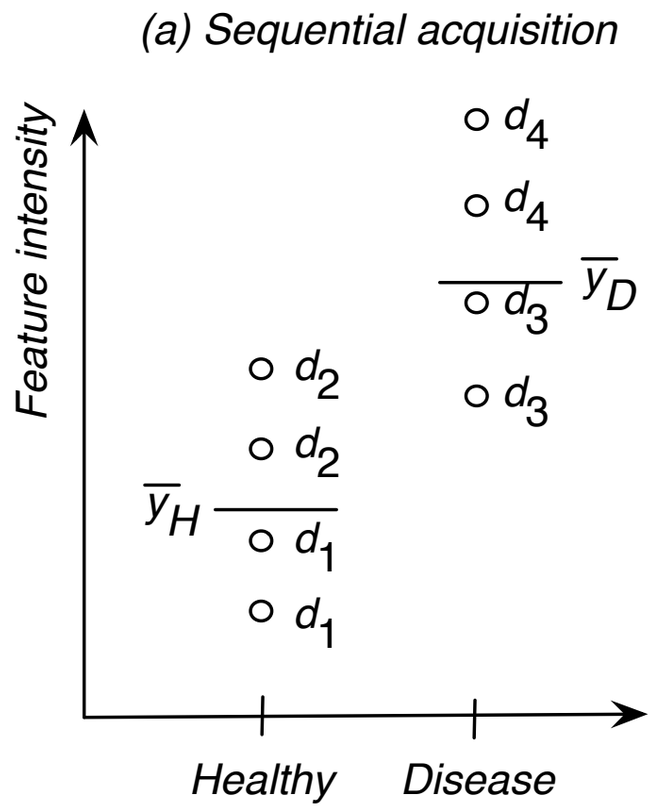


Same when jointly analyzing all features of a protein

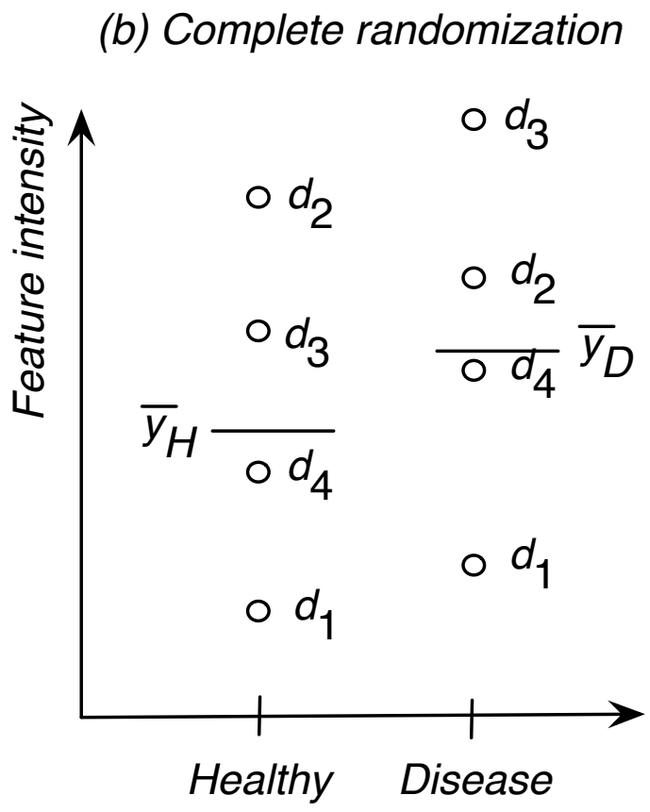
Fundamental principle 2: randomization

Required to prevent bias





No randomization
= confounding
= bias



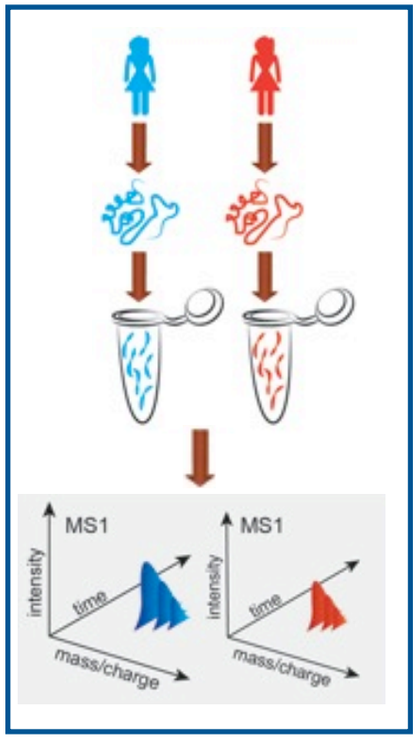
Complete randomization
= no bias

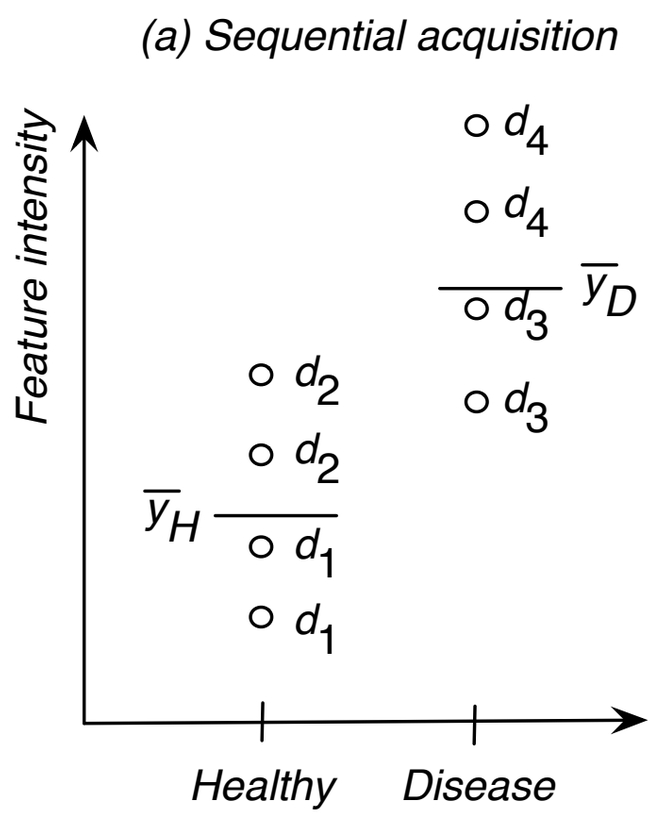
Two levels of randomness imply two types of randomization:

- *Biological replicates:* random selection of subjects from the population
- *Technical replicates:* random allocation of samples to all processing steps

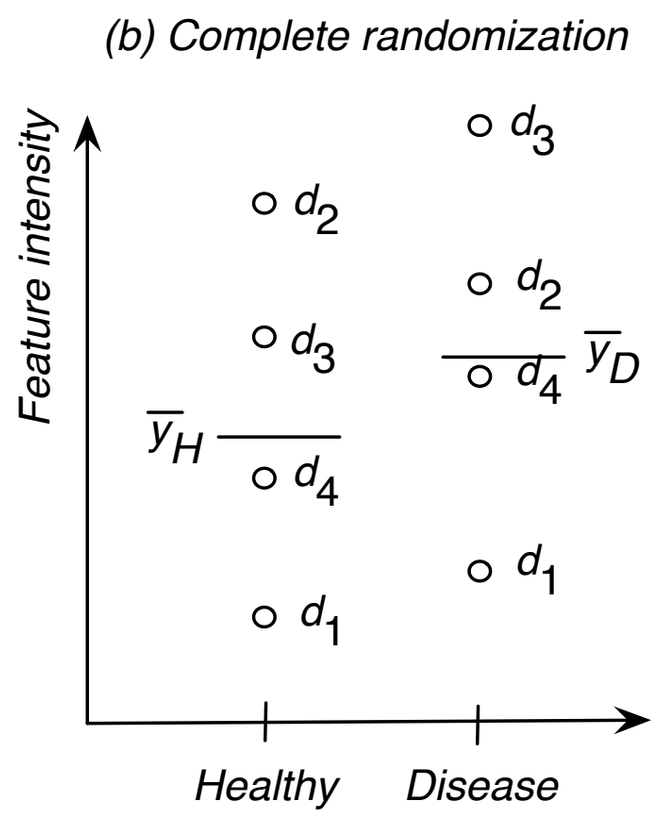
Fundamental principle 2: randomization

Required to prevent bias





No randomization
= confounding
= bias



Complete randomization
= no bias

Coronary artery disease experiment:

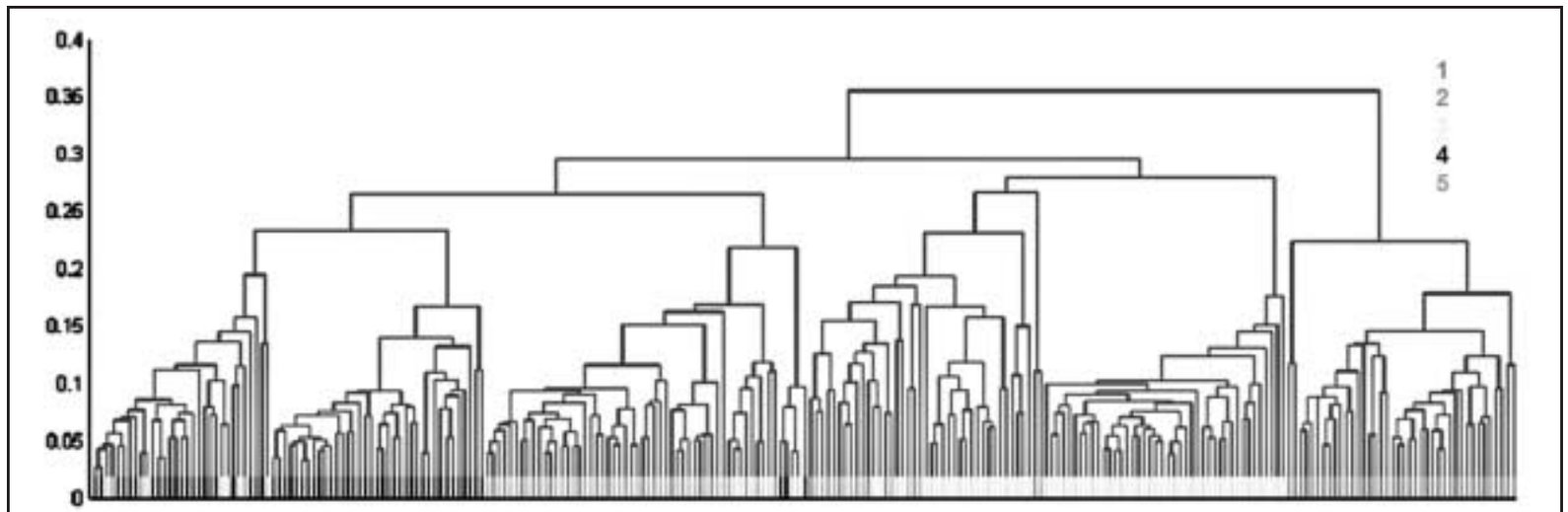
- *Biological replicates:* randomized selection from the repository
- *Technical replicates:* random order of samples

Example: technical replication and randomization

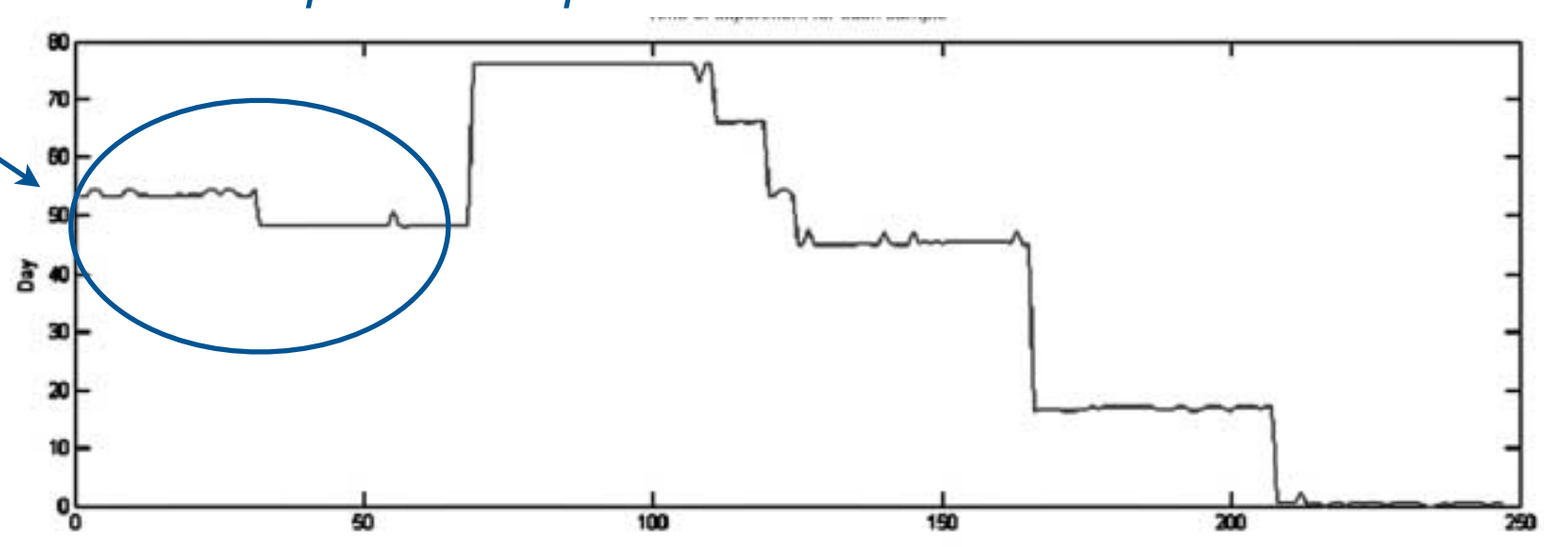
Hu, Coombes, Morris, Baggerly, Briefings in Functional Genomics, 2005

- Serum samples with five types of cancer
- SELDI-TOF MS
  - normalized, peak picked

Hierarchical clustering of samples



Time of spectral acquisition



Cancer subtype confounded with time

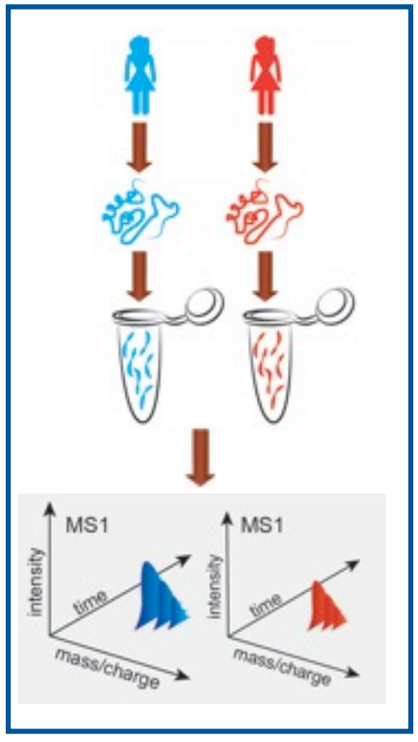
Same time-based clustering on the QC samples!

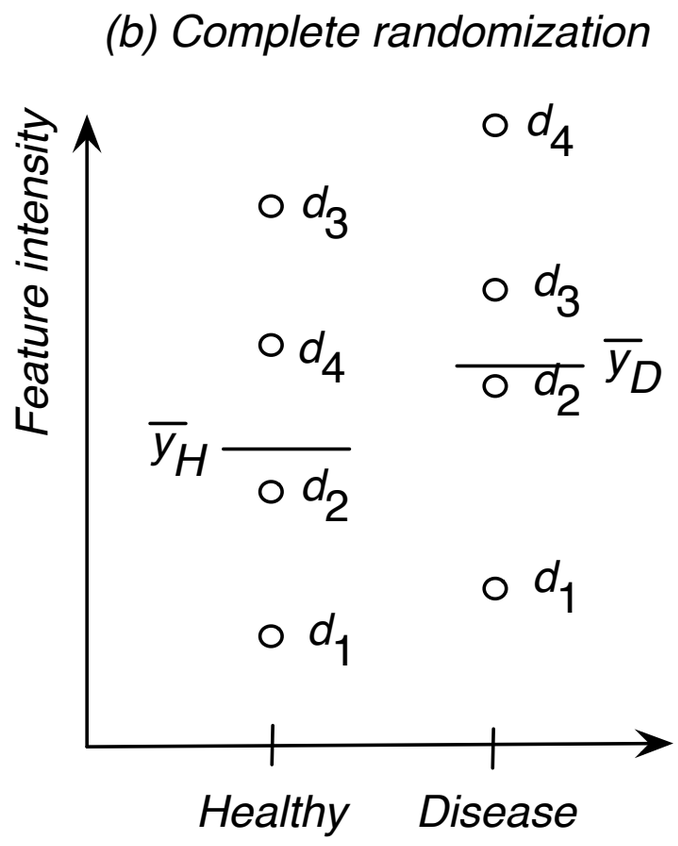
Steps of statistical experimental design

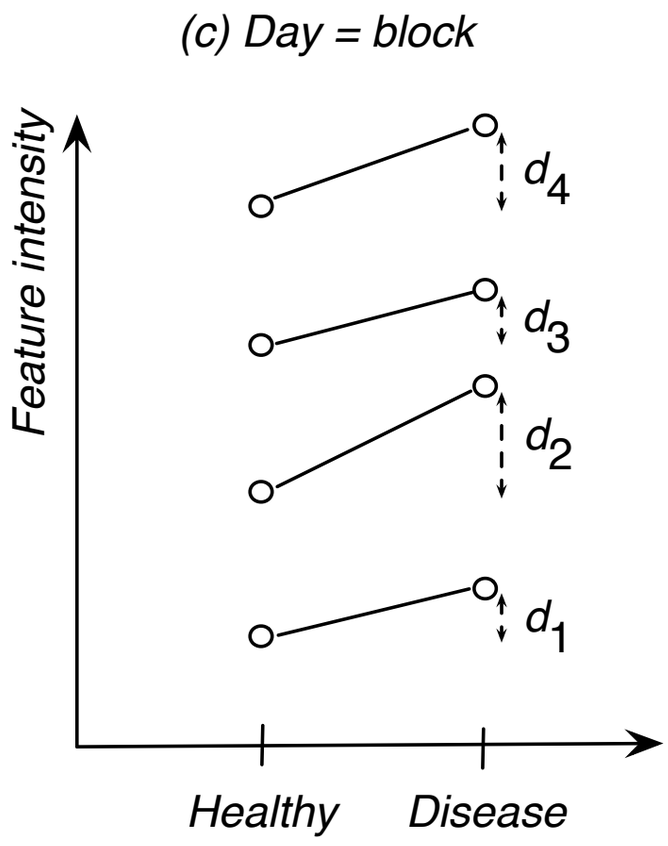
- Define the problem
  - Populations of interest
  - Comparisons of interest
  - Scope of conclusions
- Utilize 3 principles of experimental design
  - Replication
  - Randomization
  - Blocking: known biological and technical variation
  - Blocking: MS run

Fundamental principle 3: blocking

Helps reduce both bias and variance







Complete randomization
= inflated variance

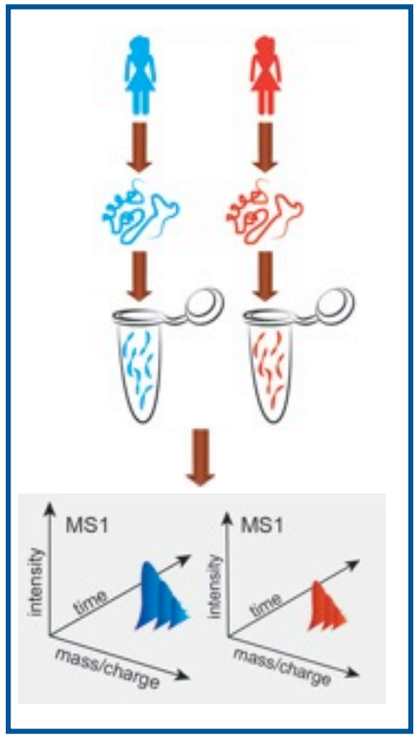
Block-randomization
= restriction on randomization
= systematic allocation

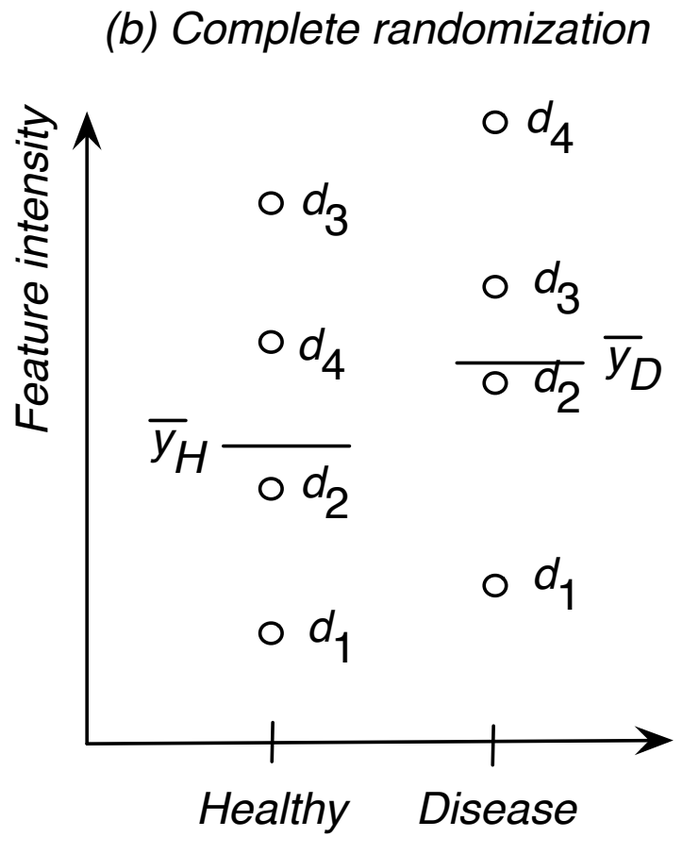
Two levels of randomness imply two types of blocks:

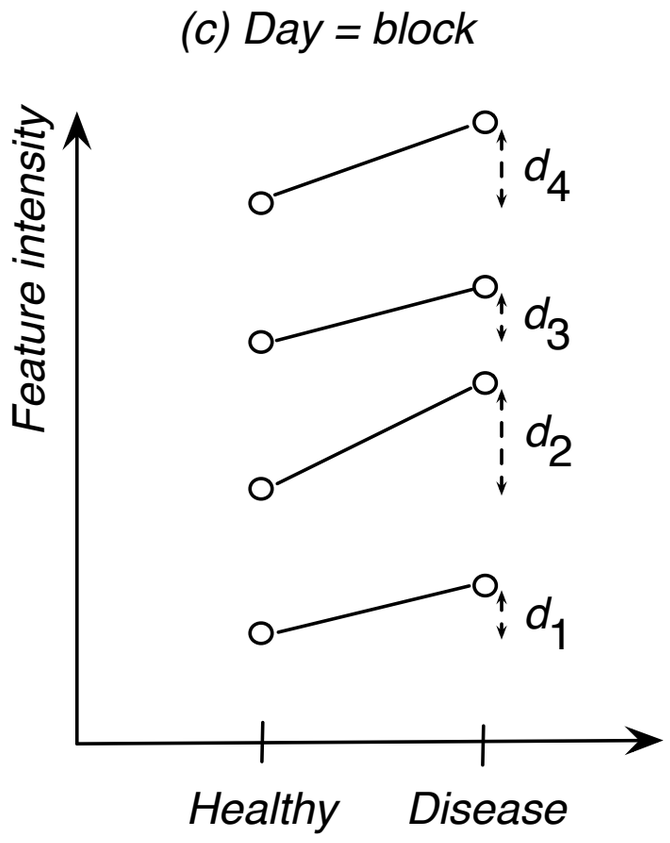
- *Biological replicates:* subjects having similar characteristics (e.g. age)
- *Technical replicates:* samples processed together (e.g. in a same day)

Fundamental principle 3: blocking

Helps reduce both bias and variance







Complete randomization
= inflated variance

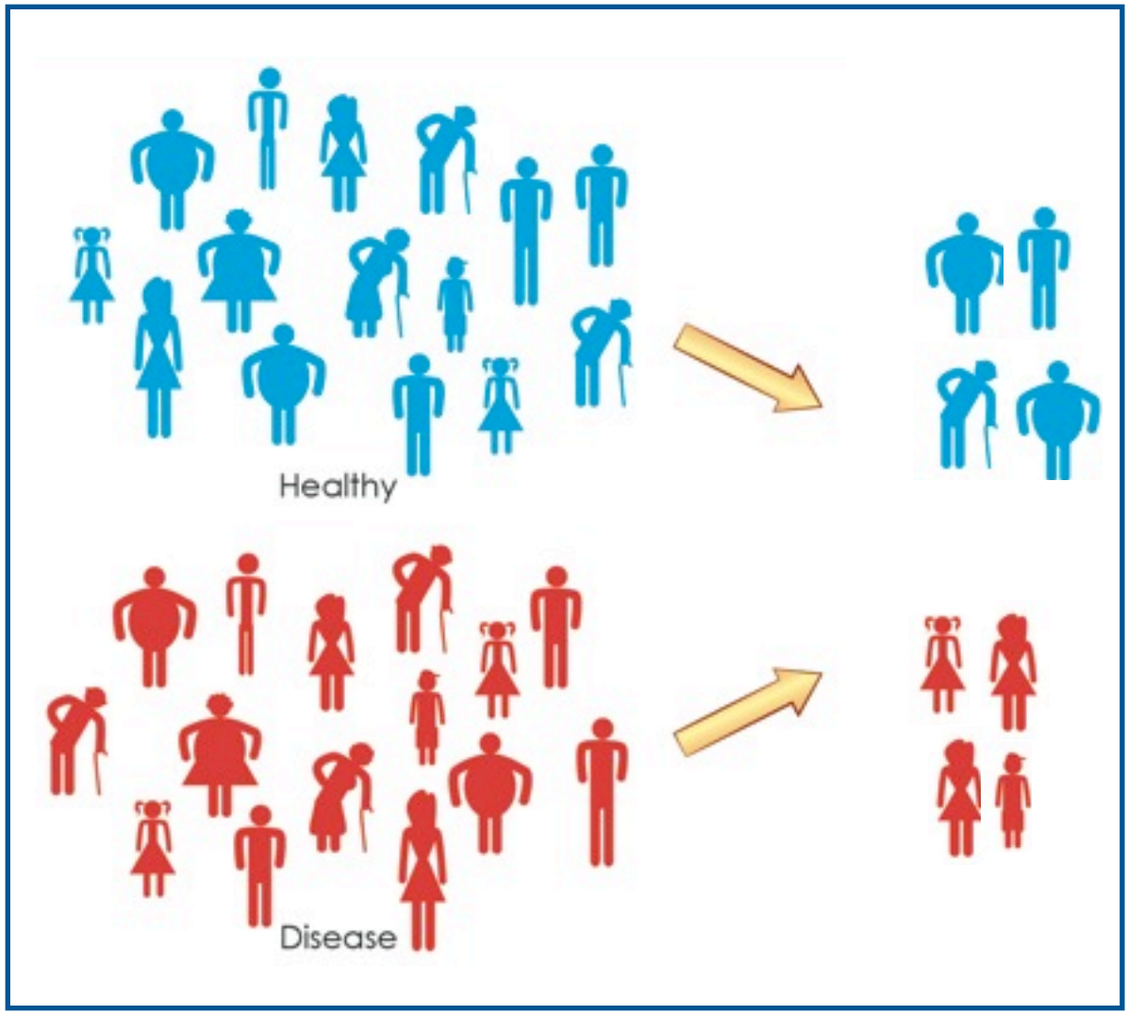
Block-randomization
= restriction on randomization
= systematic allocation

Coronary artery disease experiment:

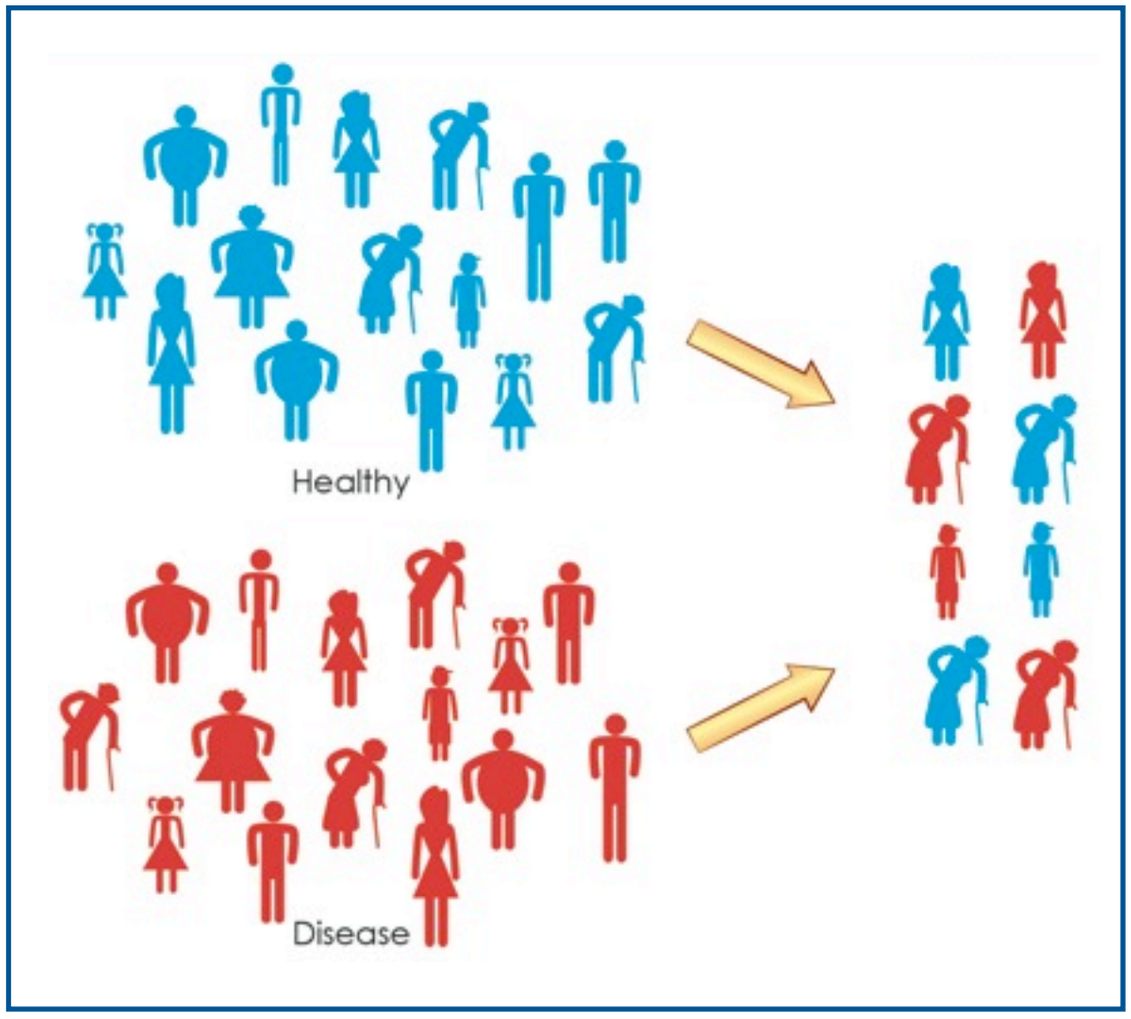
- *Biological replicates:* block-randomized sample selection
- *Technical replicates:* no important blocking factors were anticipated

Blocking with respect to biological factors (= matching)

Time course experiments are also instances of blocking (subject=block)



Complete randomization
= inflated variance



Block-randomization
= restriction on randomization
= systematic allocation

Käll and Vitek, *PLoS Computational Biology*, 7, 2011

Case study: an illustration of block-randomized selection of subjects from the repository

		Disease group				
		Control	Stable angina	Unstable angina	NSTEMI	STEMI
Stratification	$\geq$ 58 y.o; Female	354	300	49	39	29
	$\geq$ 58 y.o; Male	701	843	143	86	54
	$<$ 58 y.o; Female	80	56	5	5	8
	$<$ 58 y.o; Male	264	190	34	23	27

Counts in the initial repository of samples

		Disease group				
		Control	Stable angina	Unstable angina	NSTEMI	STEMI
Stratification	$\geq$ 58 y.o; Female	3	3	3	3	3
	$\geq$ 58 y.o; Male	3	3	3	3	3
	$<$ 58 y.o; Female	2	2	2	2	2
	$<$ 58 y.o; Male	2	2	2	2	2

Counts of subjects included in the study

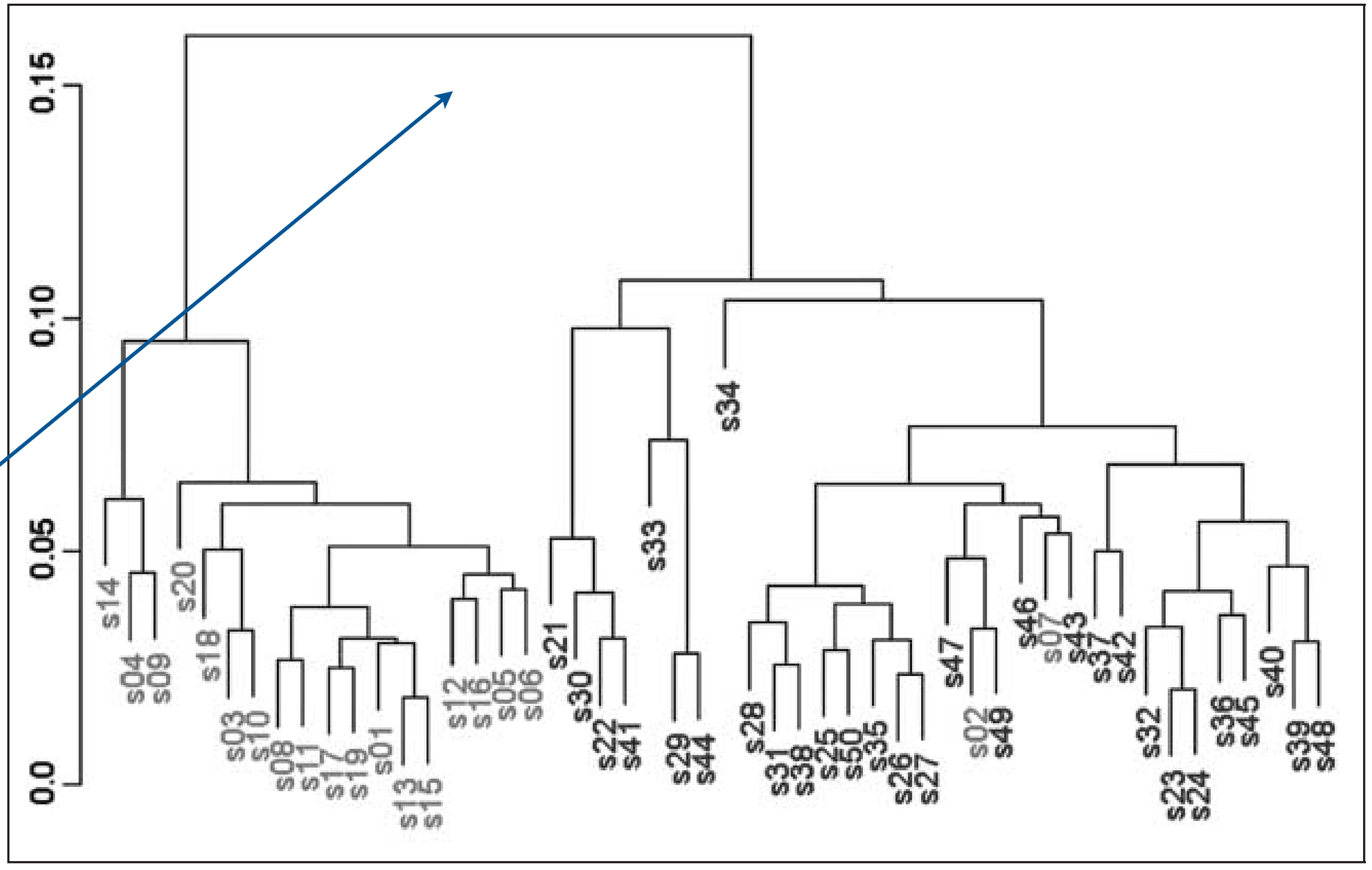
Example: blocking with respect to technical factors

Hu, Coombes, Morris, Baggerly, Briefings in Functional Genomics, 2005

- Serum samples with two types of cancer
- SELDI-TOF MS, 3 fractions
  - normalized, peak picked

Hierarchical clustering of samples

Protocol change



Summary of the experimental design of the coronary artery disease case study

- Define the problem
  - Populations: Munich Heart Center patients in 2005-2006
  - Comparisons of interest: 5 well-defined disease groups
  - Scope of conclusions: selected subjects (screening experiment)
- Utilize 3 principles of experimental design
  - Replication: 50 subjects per group, no technical replicates
  - Randomization & blocking
    - patients randomly selected from the population
    - matched by age and gender
    - random order of sample processing and spectral acquisition
    - label-free LC-MS

Alternative: block-randomized spectral acquisition

(5 subjects, one from each group, in random order),

...,

(5 subjects, one from each group, in random order),

Example in this tutorial

Differentially abundant proteins in a Dahl Salt sensitive rat model

- Define the problem
  - Populations: Dahl salt sensitive rats
  - Comparisons of interest: high vs low salt diet
  - Scope of conclusions: selected subjects (screening experiment)
- Utilize 3 principles of experimental design
  - Replication: 7 rats per group, 3 technical replicates
  - Randomization & blocking
    - rats randomly selected from the population
    - rats randomly assigned to treatment
    - random order of sample processing and spectral acquisition
    - label-free SRM

Alternative: block-randomized spectral acquisition

(2 rats, one from each group, in random order),

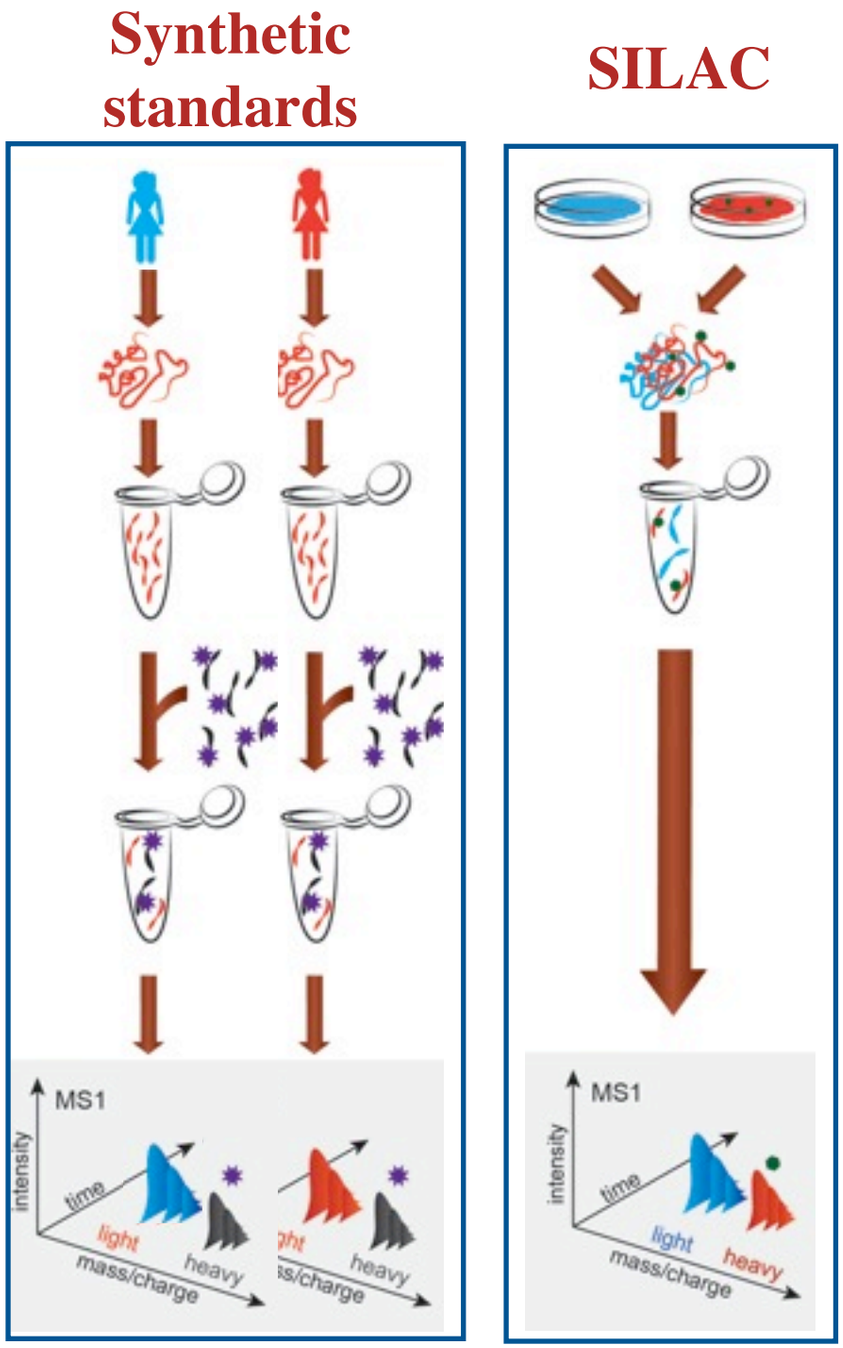
...,

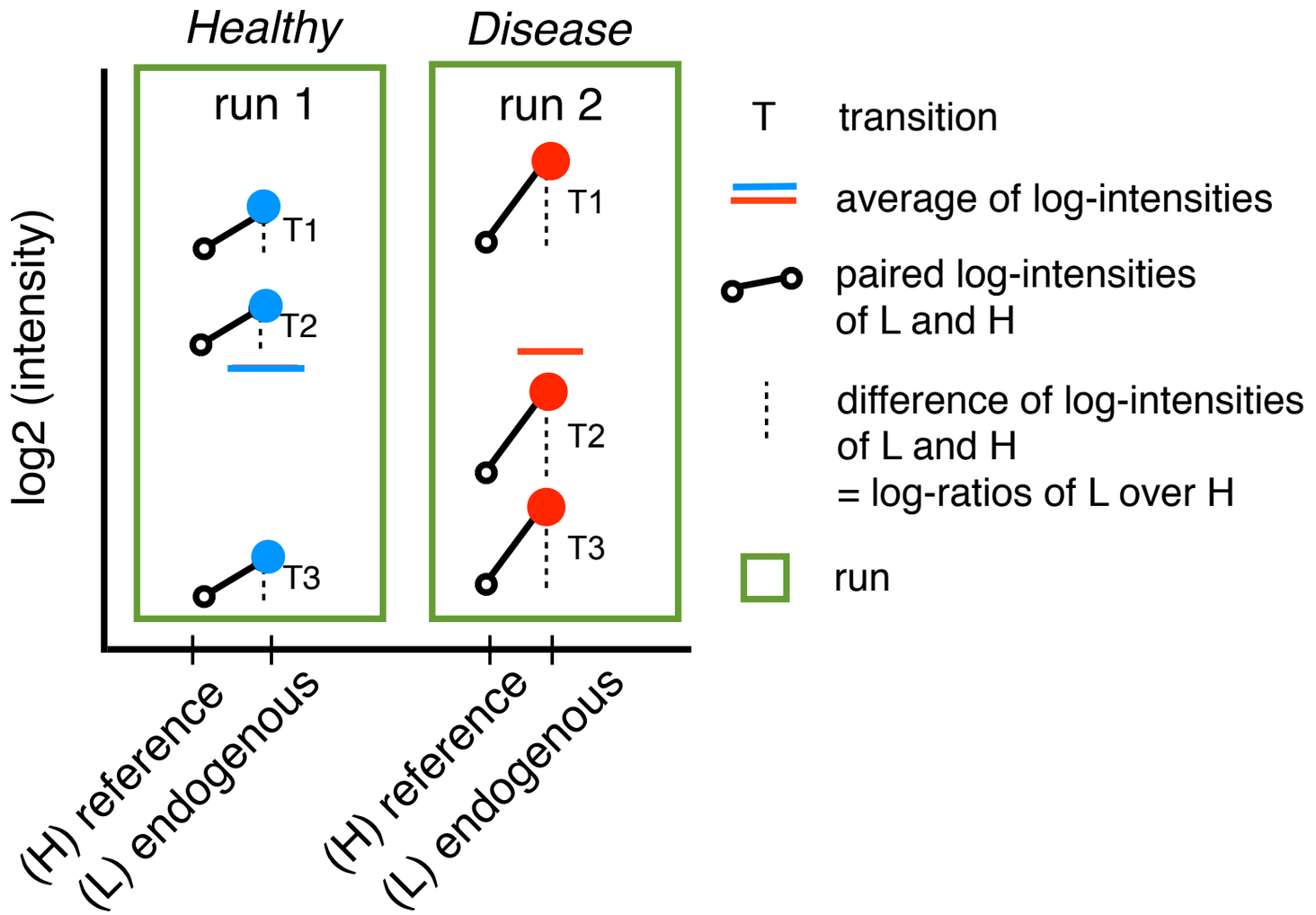
(2 rats, one from each group, in random order),

Steps of statistical experimental design

- Define the problem
  - Populations of interest
  - Comparisons of interest
  - Scope of conclusions
- Utilize 3 principles of experimental design
  - Replication
  - Randomization
  - Blocking: known biological and technical variation
  - Blocking: MS run

Labeling (multiplexing) is also an instance of blocking





Multiplexing reduces both bias and variance

(assuming that extra sample handling does not introduce extra variation)

Example in this tutorial

Differentially abundant proteins in ovarian cancer patients

- Define the problem
  - Populations: Patients at University Hospital Zürich with no previous history of disease
  - Comparisons of interest: disease vs controls
  - Scope of conclusions: selected subjects (screening experiment)

- Utilize 3 principles of experimental design
  - Replication: 6 disease and 10 control patients, no technical reps
  - Randomization & blocking
    - random order of sample processing and spectral acquisition
    - label-based SRM

Alternative: block-randomized spectral acquisition

(2 subjects, one from each group, in random order),

...,

(2 subjects, one from each group, in random order),

How to allocate samples to runs?
Allocation of resources in a 2-label workflow > 2 groups

(a) Balanced Incomplete Block

Disease group	Replicate set 1 Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8	Block 9	Block 10	⋯
D_1	X_{L_1}	X_{L_2}	X_{L_1}	X_{L_2}							⋯
D_2	X_{L_2}				X_{L_1}	X_{L_2}	X_{L_1}				⋯
D_3		X_{L_1}			X_{L_2}			X_{L_1}	X_{L_2}		⋯
D_4			X_{L_2}			X_{L_1}		X_{L_2}		X_{L_1}	⋯
D_5				X_{L_1}			X_{L_2}		X_{L_1}	X_{L_2}	⋯

(b) Reference

Disease group	Replicate set 1 Block 1	Block 2	Block 3	Block 4	Block 5	⋯
R	R_{L_1}	R_{L_1}	R_{L_1}	R_{L_1}	R_{L_1}	⋯
D_1	X_{L_2}					⋯
D_2		X_{L_2}				⋯
D_3			X_{L_2}			⋯
D_4				X_{L_2}		⋯
D_5					X_{L_2}	⋯

(c) Loop

Disease group	Replicate set 1 Block 1	Block 2	Block 3	Block 4	Block 5	⋯
D_1	X_{L_1}				X_{L_2}	⋯
D_2	X_{L_2}	X_{L_1}				⋯
D_3		X_{L_2}	X_{L_1}			⋯
D_4			X_{L_2}	X_{L_1}		⋯
D_5				X_{L_2}	X_{L_1}	⋯

- Reference design
  - allocate a same control subject in every run
  - keep same channels across groups
- BIB and loop designs
  - systematically rotate group allocation to runs
  - randomize or systematically rotate channels across groups

Calculate model-based variances of comparisons for each allocation to determine the best design given resource constraints

How to allocate samples to runs?
Allocation of resources in a 4-label workflow

(a) Randomized Complete Block

Disease group	Replicate set 1 Block 1	Replicate set 2 Block 2	⋯ ⋯
D_1	X	X	⋯
D_2	X	X	⋯
D_3	X	X	⋯
D_4	X	X	⋯

(b) Balanced Incomplete Block

Disease group	Replicate set 1 Block 1	Block 2	Block 3	Block 4	Block 5	⋯
D_1	X	X	X	X		⋯
D_2	X	X	X		X	⋯
D_3	X	X		X	X	⋯
D_4	X		X	X	X	⋯
D_5		X	X	X	X	⋯

- 4 groups or less
  - allocate a subject from each group to a run
  - randomize or systematically rotate channels across groups
- 5 groups or more
  - systematically rotate group allocation to runs
  - randomize or systematically rotate channels across groups

Calculate model-based variances of comparisons for each allocation to determine the best design given resource constraints

Concluding thoughts

- Clearly define the problem before starting the experiment
  - Do not change the comparisons of interest and the scope of conclusions after seeing the data
- Experimental design is critical
  - Randomization, replication and blocking
  - Statistical analysis will not correct the faults of design
- Need a statistical model to finalize the design
  - Jointly analyzing all conditions & all features gains sensitivity
  - Compare designs in terms of expected variation
- Involve a statistician in all steps of experiment planning!

References

- Experimental design
  - A. Oberg, O. Vitek. *J. Proteome Research*, 8, p.2144, 2009.
  - D. Ransohoff . *Nature Reviews Cancer,* 5, p.142, 2005.
- Case studies
  - *Cardiovascular disease:*
    T. Clough et al. *Methods in Molecular Biology*, 728, 2011.
  - *Ovarian cancer:*
    C.-Y. Chang et al. *Molecular & Cellular Proteomics*, 2012.
  - *More examples:*
    Hu et al. *Briefings in Functional Genomics & Proteomics*, 3, p.322, 2005.
- Reproducible computational research
  - R. Peng. *Science*, 334, p.6060, 2011