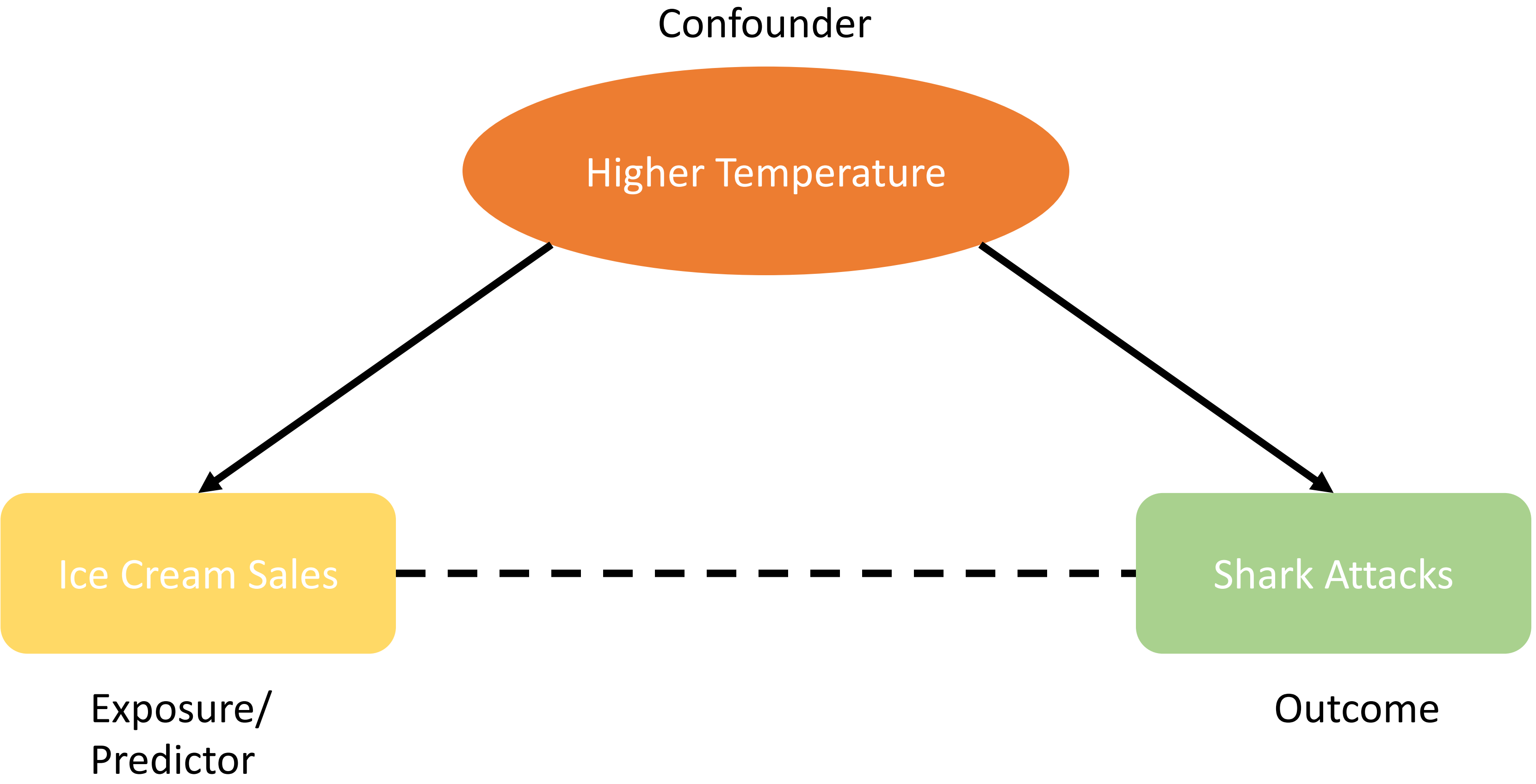
Special Topics in Biostatistics and Bioinformatics
Week V

Ege Ülgen, M.D.

31 March 2022

ACIBADEM
MEHMET ALİ AYDINLAR
ÜNİVERSİTESİ

What is confounding?



Confounder

Higher Temperature

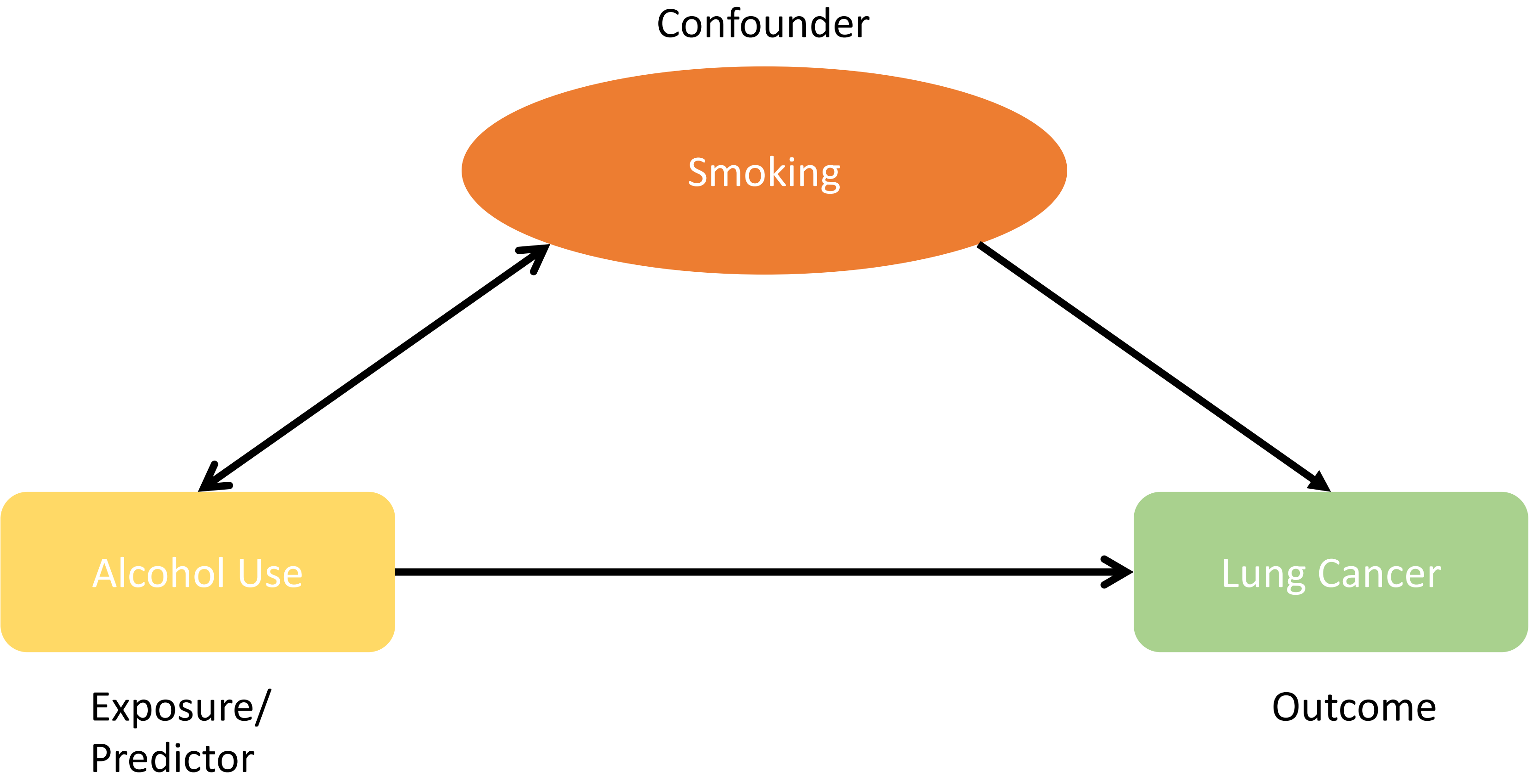
Ice Cream Sales

Shark Attacks

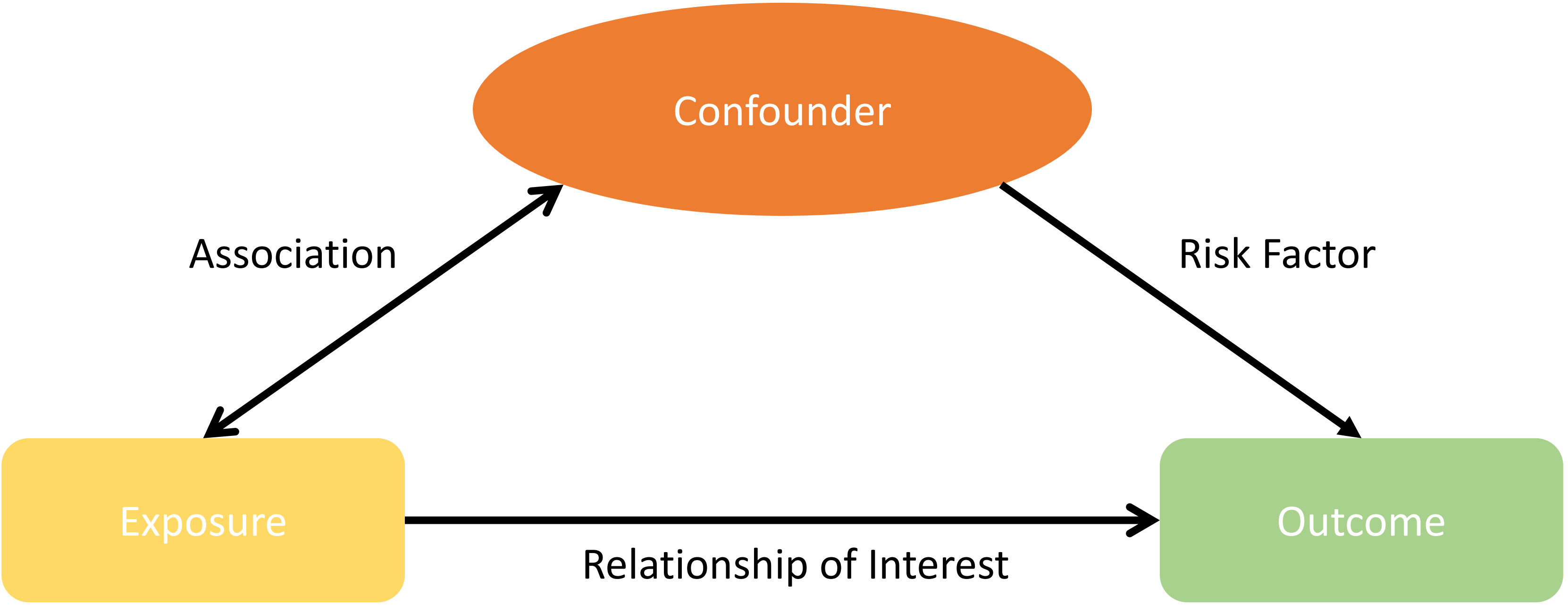
Exposure/
Predictor

Outcome

What is confounding?



What is confounding?


Confounder

Association

Risk Factor

Exposure

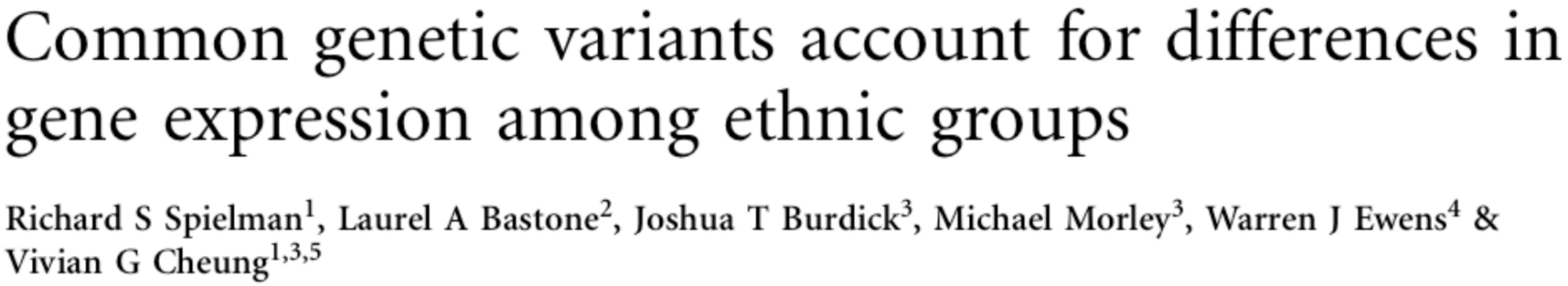
Outcome

Relationship of Interest

https://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture1.pdf

The Most Common Confounder

- Batch Effects



Common genetic variants account for differences in gene expression among ethnic groups

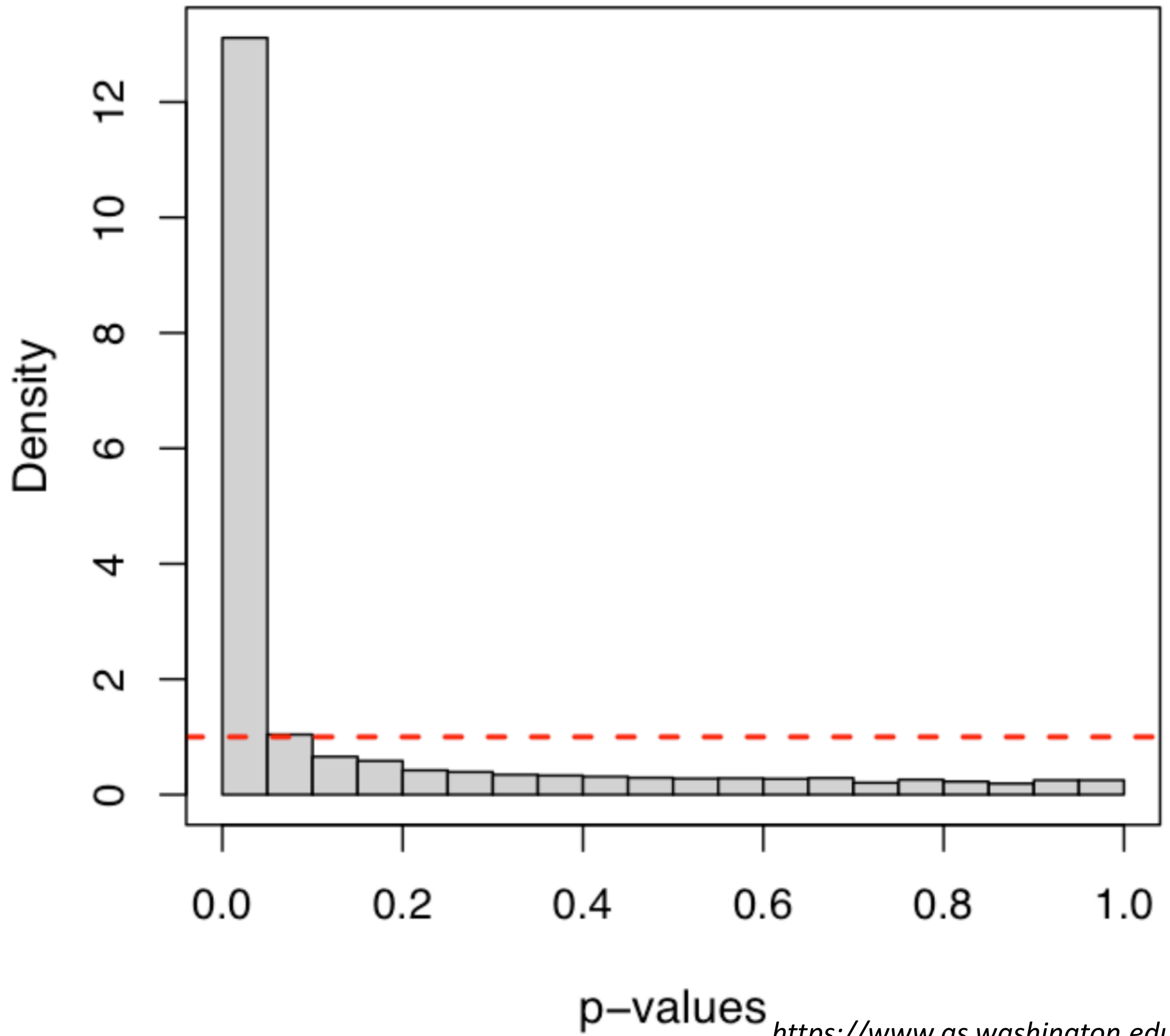
Richard S Spielman[1], Laurel A Bastone[2], Joshua T Burdick[3], Michael Morley[3], Warren J Ewens[4] & Vivian G Cheung[1,3,5]

- Compared gene expression levels between 60 CEU and 82 ASN HapMap individuals

- Tests of differential expression performed by ***parametric t-tests*** and adjustment for ***multiple testing*** through Sidak corrections

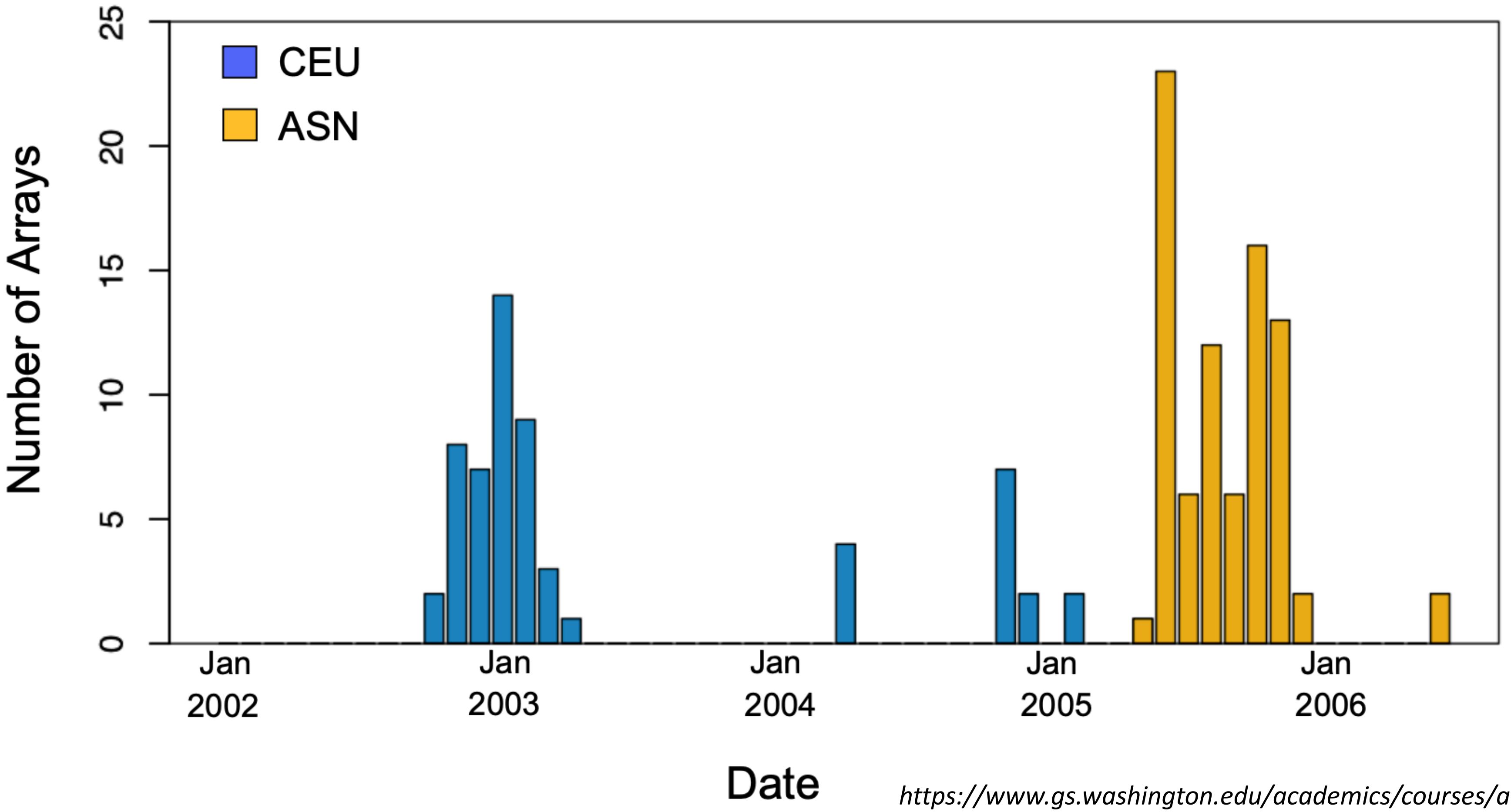
- Estimate ~26% of genes to be differentially expressed

78% of Genes Are Estimated To Be Differentially Expressed



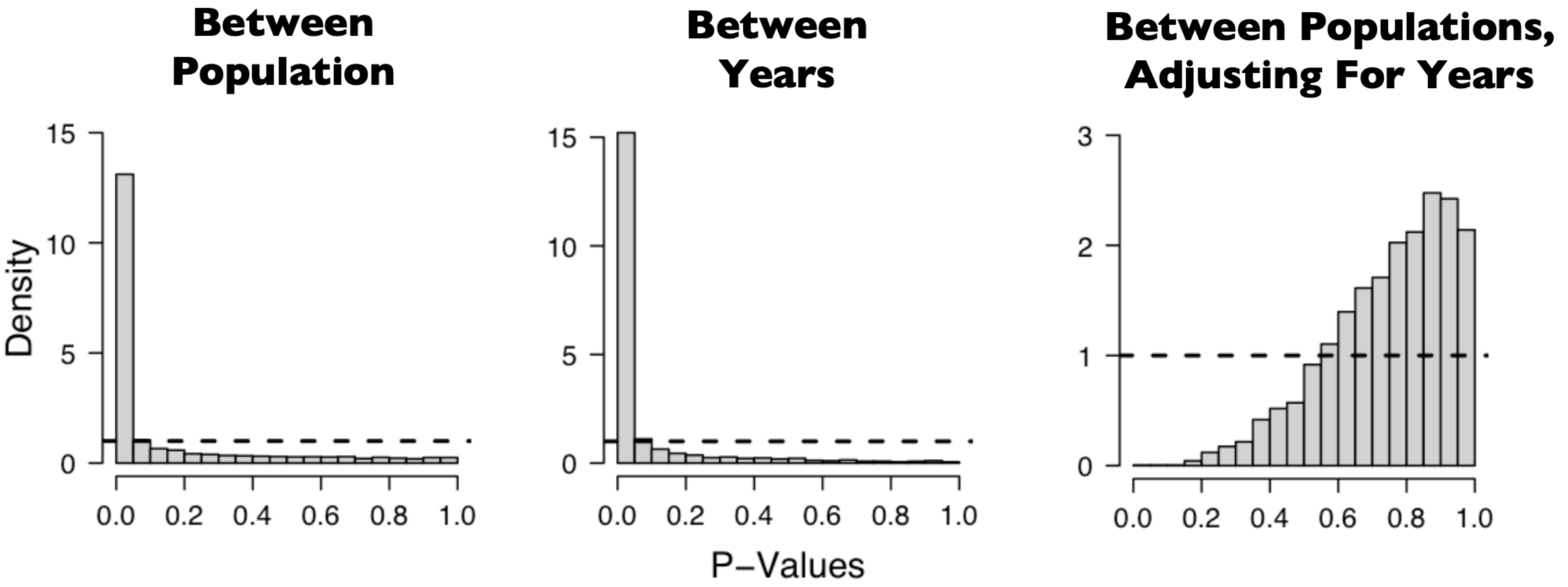
https://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture1.pdfcc

Population and Time of Processing Are Confounded



https://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture1.pdfcc

Batch Effects Can Completely Account For Differential Expression



Between Population

Between Years

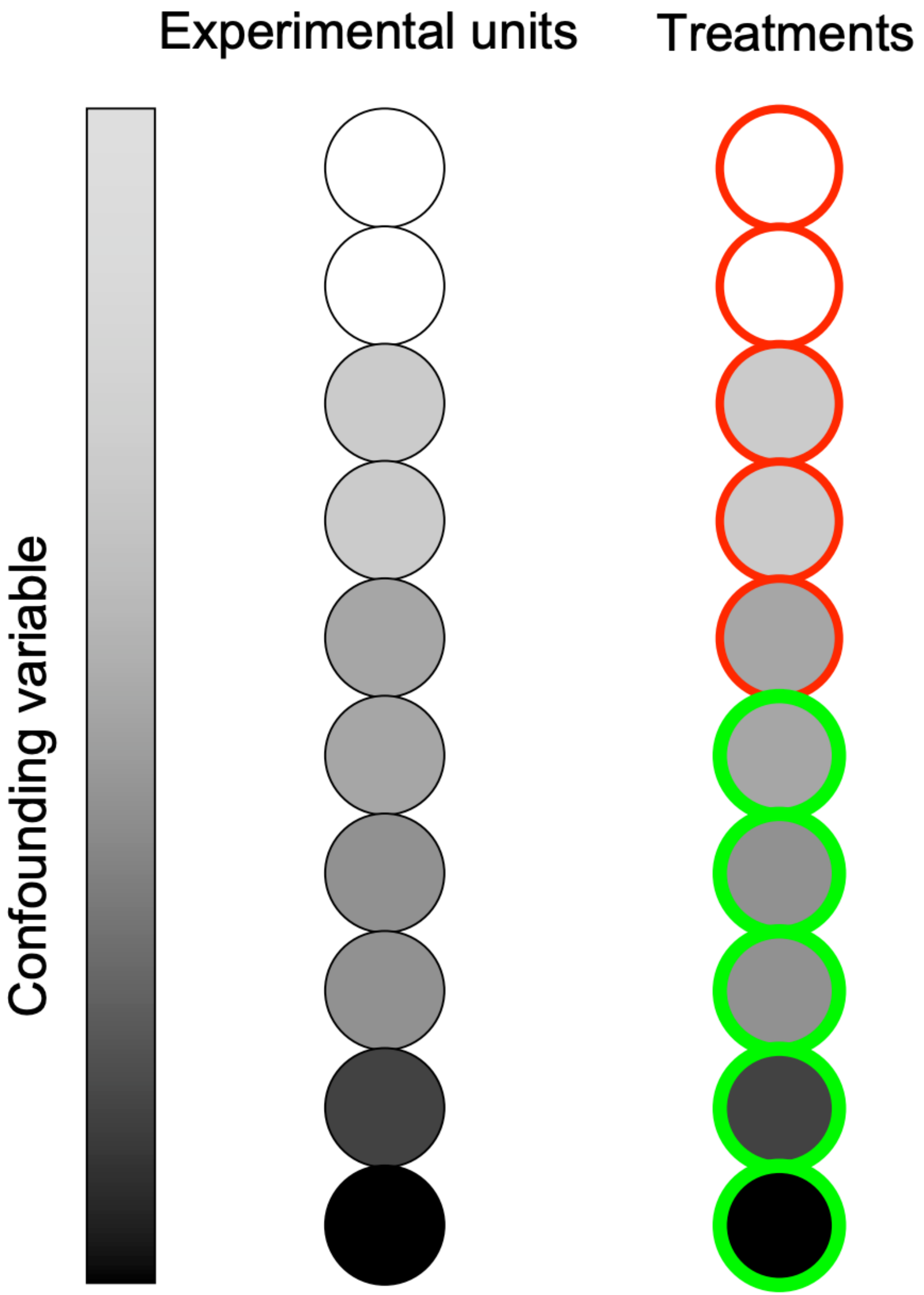
Between Populations, Adjusting For Years

78% of genes estimated to be differentially

96% of genes estimated to be differentially

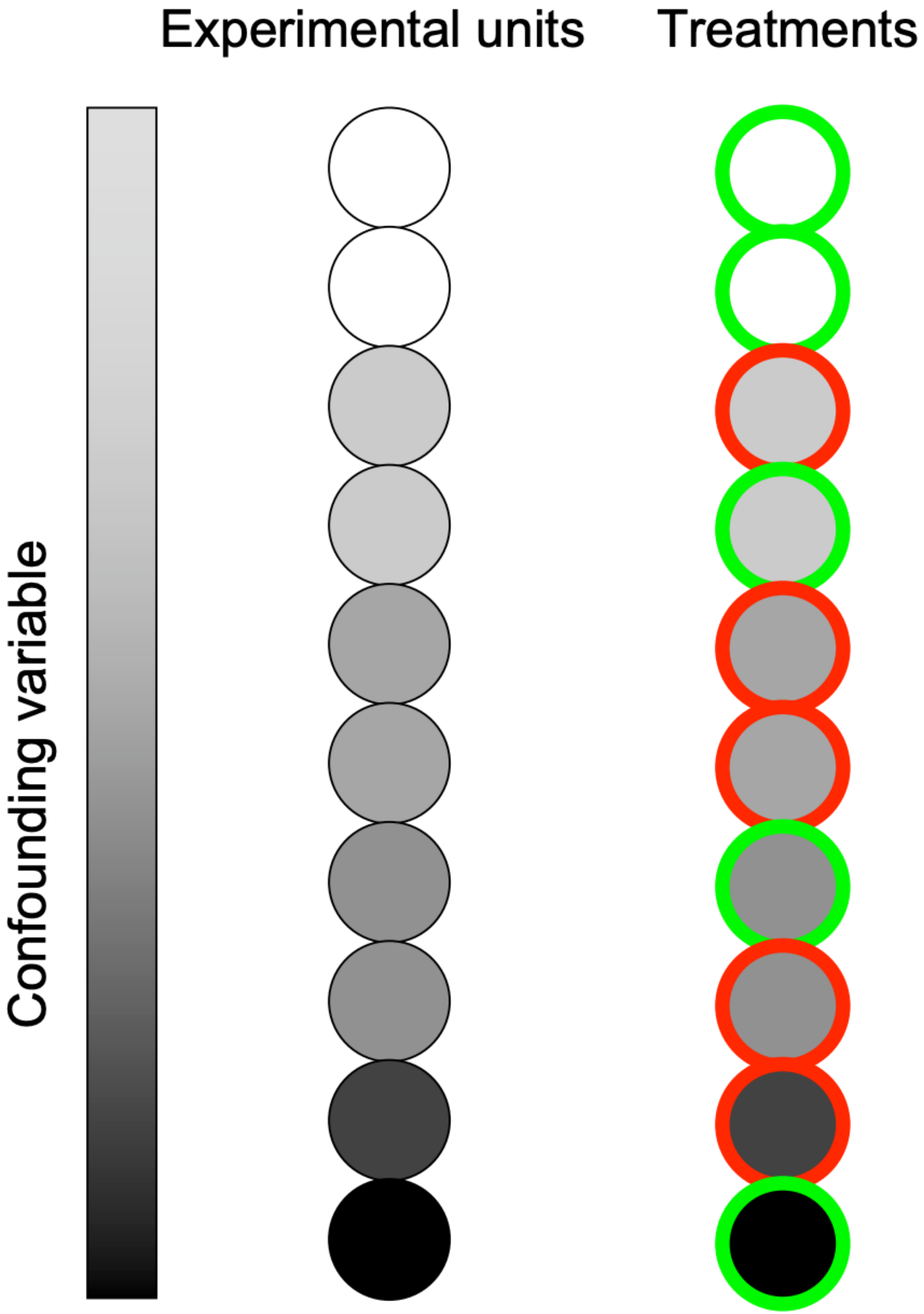
0% of genes estimated to be differentially

https://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture1.pdfcc



Without randomization, the confounding variable differs among treatments

https://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture1.pdfcc

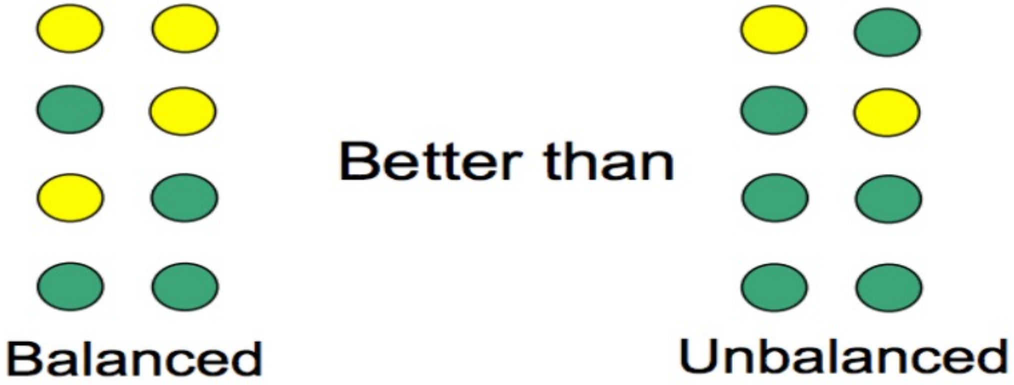


With randomization, the confounding variable does not differ among treatments

https://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture1.pdfcc

More good study design characteristics

- Balanced
- Replicated
- Has controls



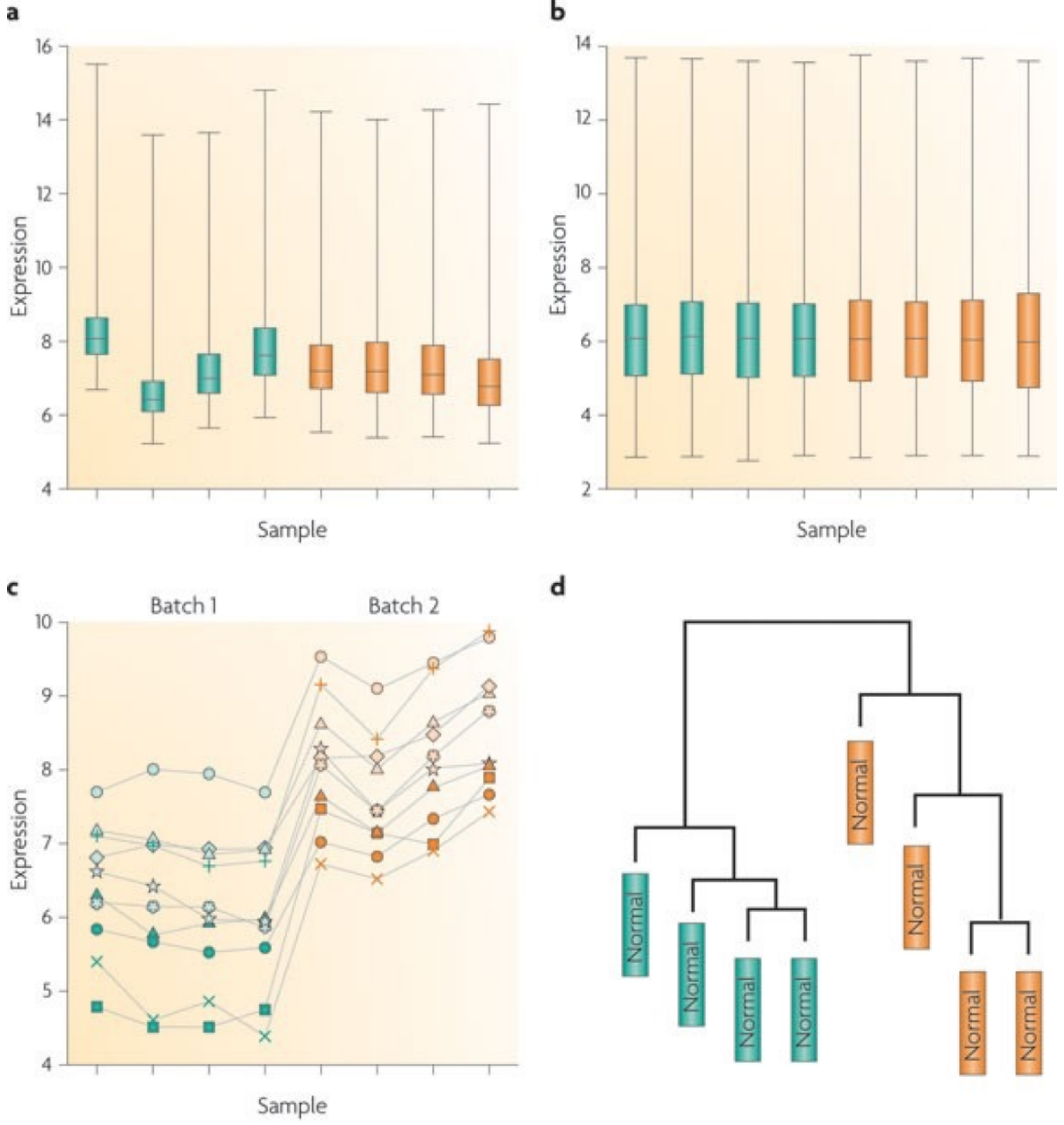
Batch Effects

- Batch effects are **unwanted sources of variation** caused by different processing date, handling personnel, reagent lots, equipment/machines, etc.

- A common method for visualizing the existence of batch effects is PCA
- The first two principal components are plotted with each sample colored by the suspected batch, and **separation of colors** is taken as evidence of a batch effect

Green and orange represent two different processing dates

a. Box plot of raw gene expression data (log base 2)
b. Box plot of data processed with quantile normalization
c. Example of ten genes that are susceptible to batch effects even after normalization (Hundreds of genes show similar behavior but, for clarity, are not shown)
d. Clustering of samples after normalization. Note that the **samples perfectly cluster by processing date**

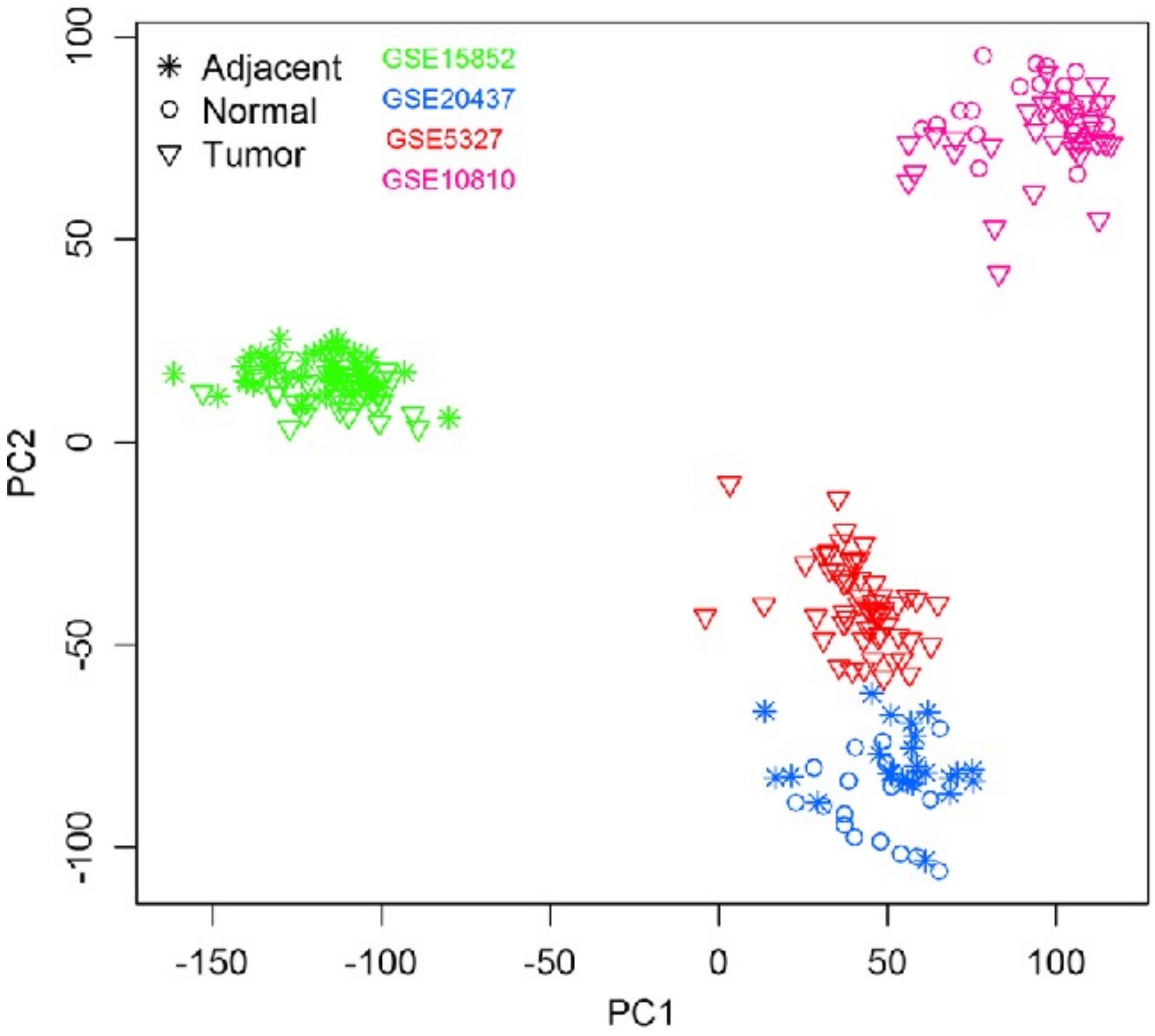


Nature Reviews | Genetics

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics. 2010 Oct;11(10):733–9.

Batch Effect Example

Thillaiyampalam G, Liberante F, Murray L, Cardwell C, Mills K, Zhang S-D. An integrated meta-analysis approach to identifying medications with potential to alter breast cancer risk through connectivity mapping. BMC Bioinformatics. 2017 Dec 21;18(1):581.



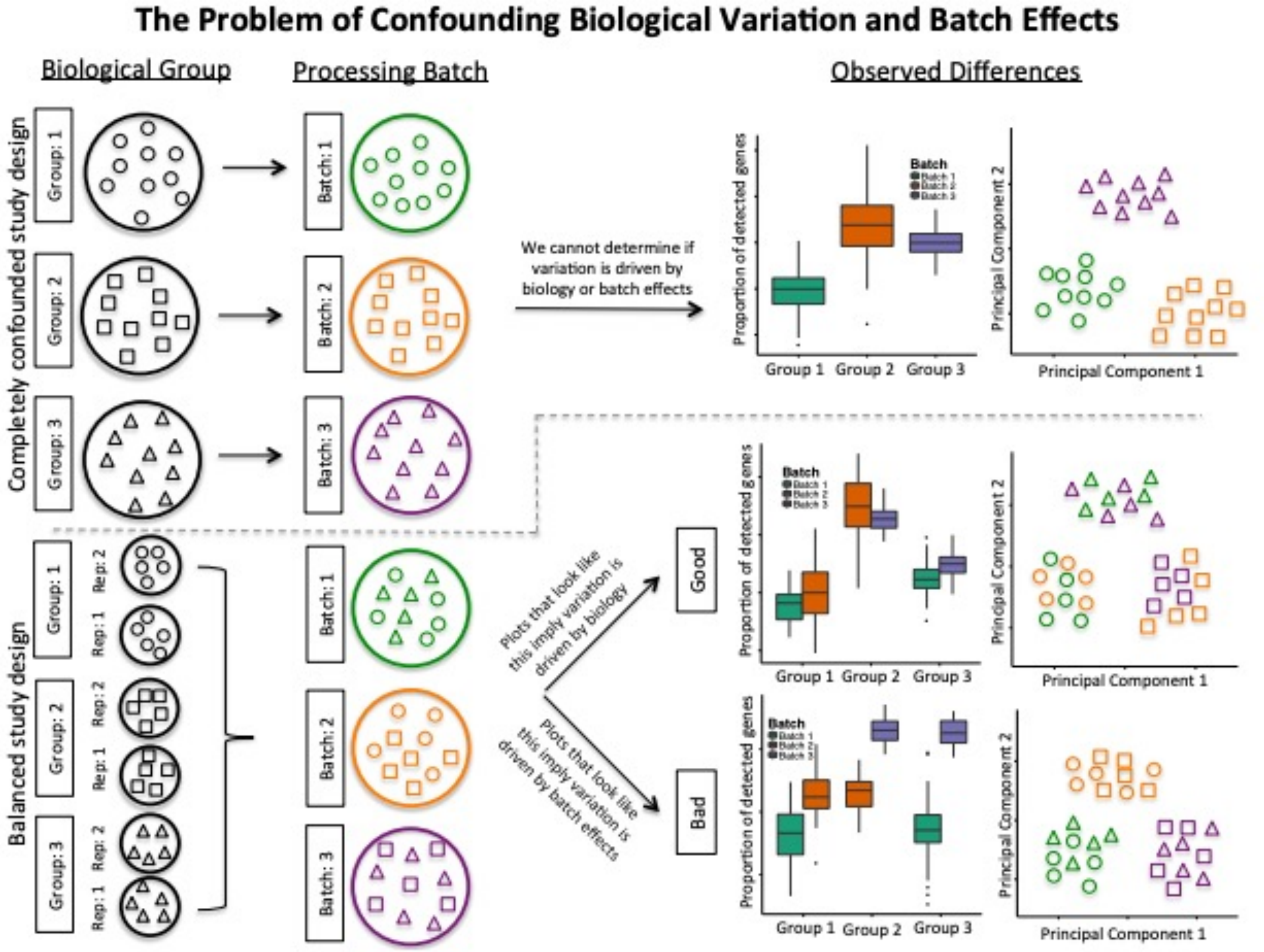
Sources of “Batch” Effects

- External Factors (e.g., environment)
- Genetics/Epigenetics
- Technical Factors

https://www.biorxiv.org/content/biorxiv/early/2015/08/25/025528.full.pdf

When can you remove batch effects?

- When they don’t perfectly overlap with what you care about



The Problem of Confounding Biological Variation and Batch Effects

Biological Group
Processing Batch
Observed Differences
Completely confounded study design
Group: 1
Group: 2
Group: 3
Batch: 1
Batch: 2
Batch: 3
We cannot determine if variation is driven by biology or batch effects
Proportion of detected genes
Batch
Batch 1
Batch 2
Batch 3
Group 1
Group 2
Group 3
Principal Component 2
Principal Component 1
Balanced study design
Rep: 1
Rep: 2
Plots that look like this imply variation is driven by biology
Plots that look like this imply variation is driven by batch effects
Good
Bad

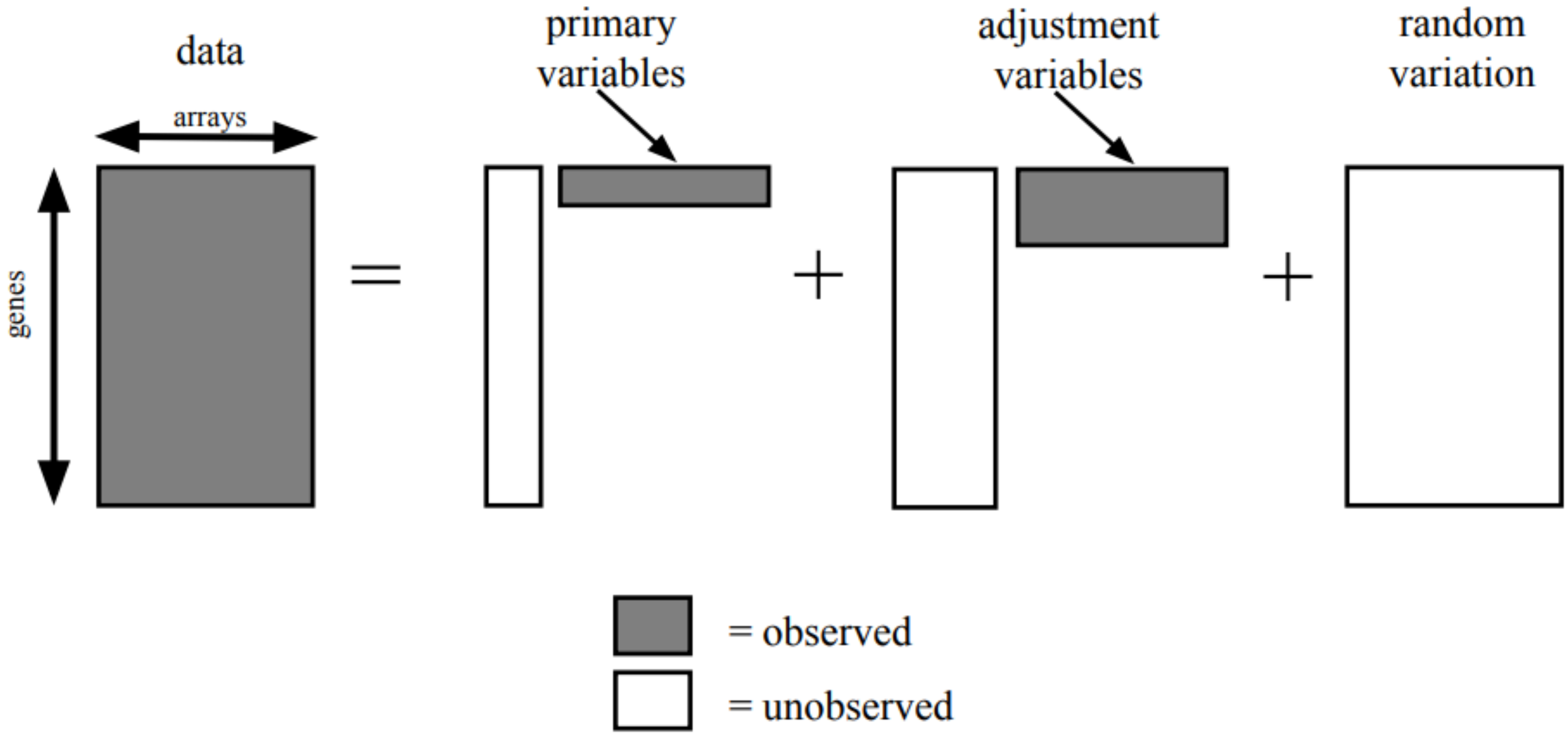
When batch is known

$$Y = \beta_0 + \beta_1 P + \beta_2 B + \varepsilon$$

P: phenotype you're investigating

B: Batch

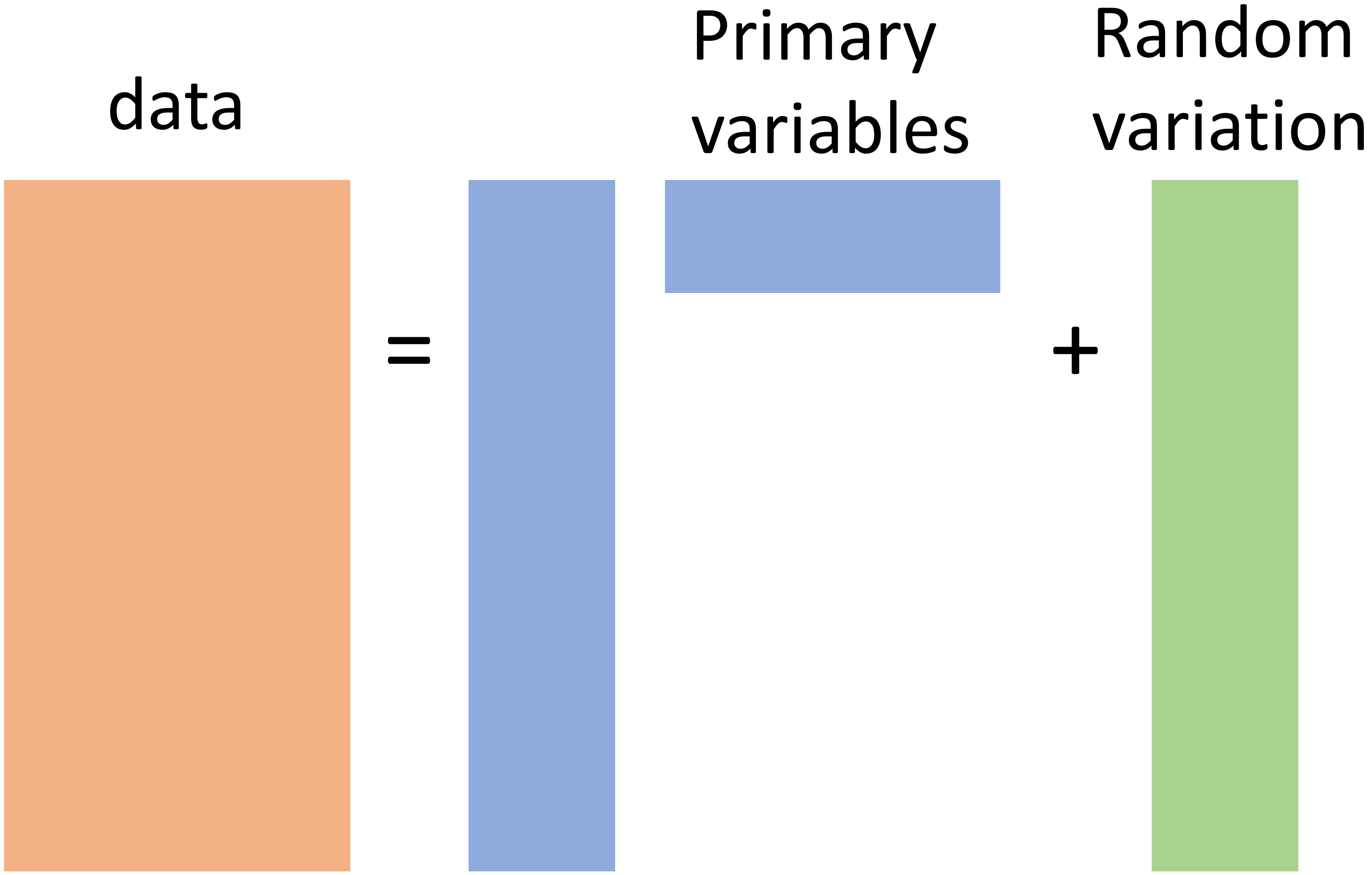
http://jtleek.com/genstats_site/lecture_notes/02_12_Batch_Effects_and_Confounders.pdf



https://rdrr.io/bioc/sva/man/ComBat.html

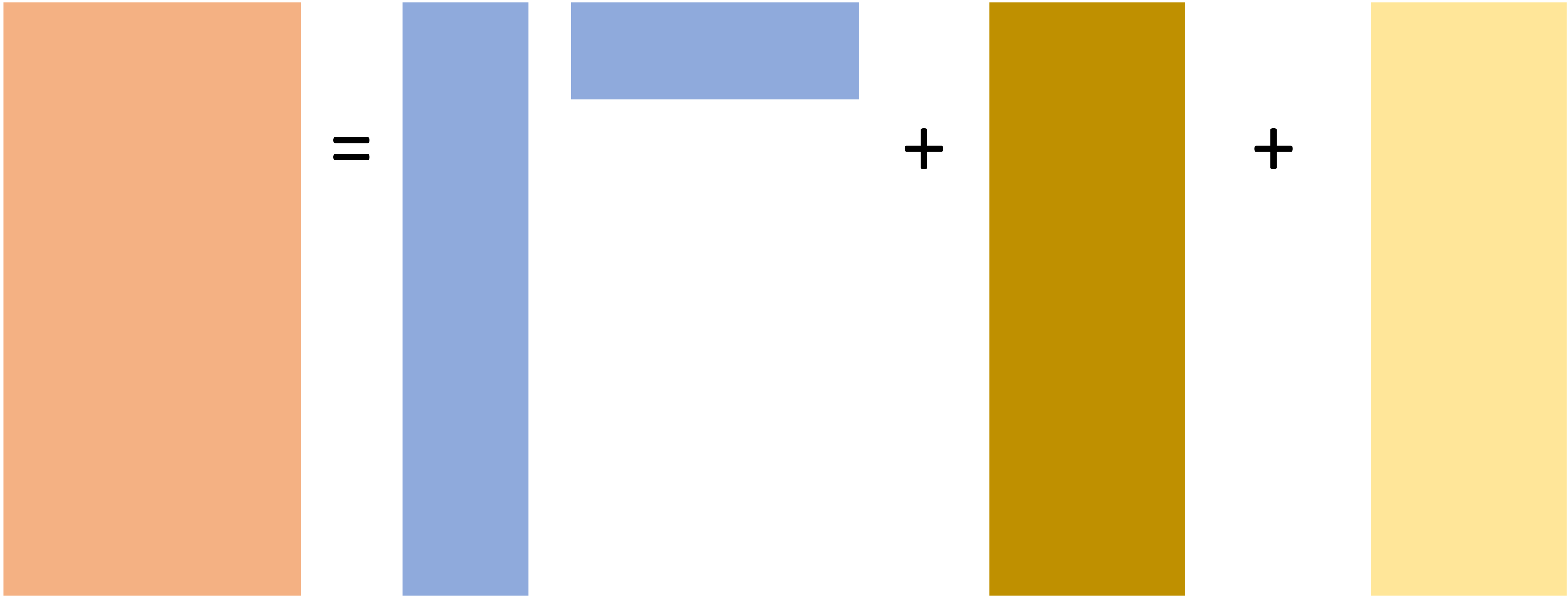
When batch is unknown

Adapted from
Jeff Leek



$$X = B\;S + E$$

Adapted from
Jeff Leek



data

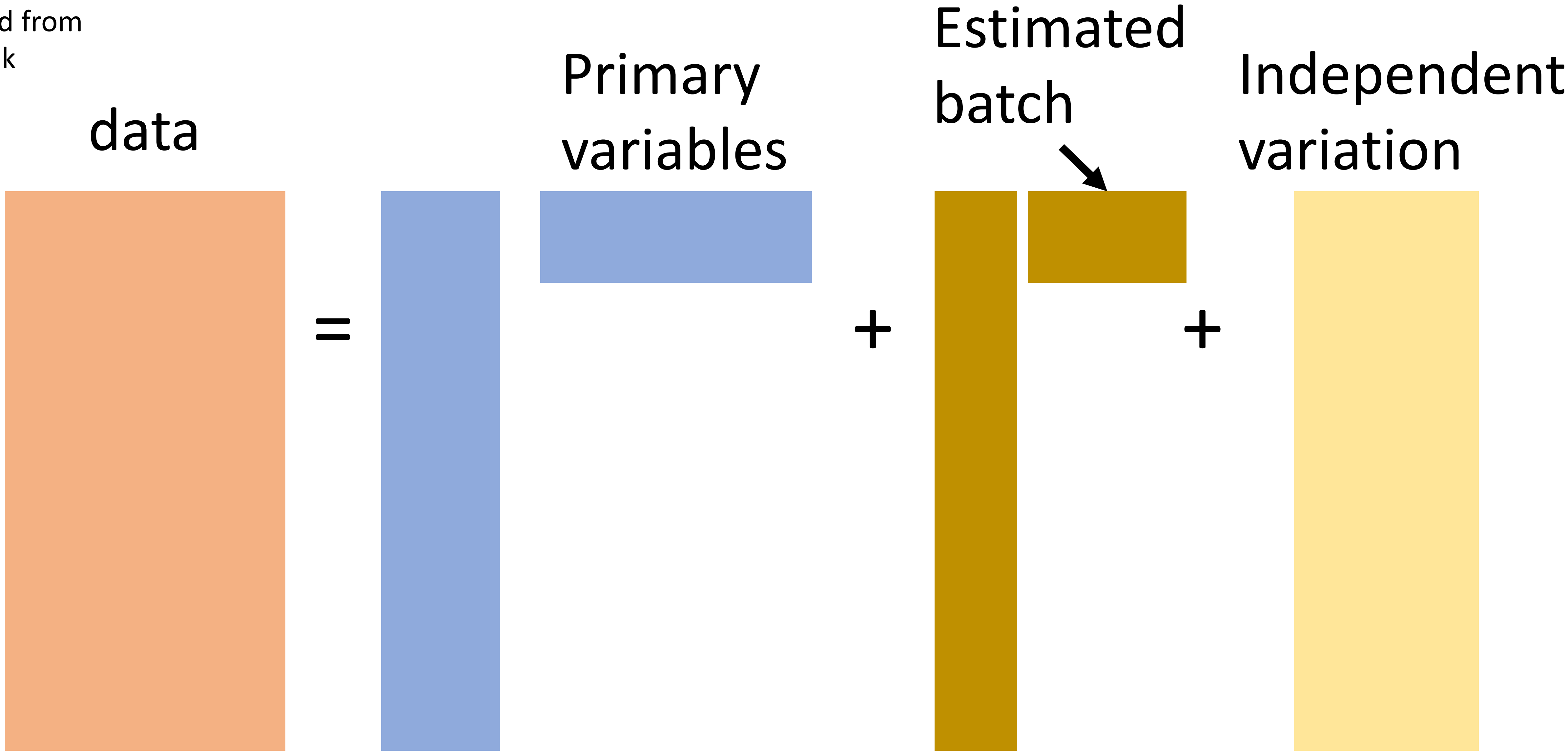
Primary variables

Dependent variation

Independent variation

$$X = B\;S + H + U$$

Adapted from
Jeff Leek



data

Primary variables

Estimated batch

Independent variation

$$X = B\,S + \Gamma\,G + U$$

Surrogate Variable Analysis

- https://www.pnas.org/content/105/48/18718.full
- https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0030161