



Jhelum Chakravorty

Published Sep 29

Fork of Large-Scale Inference and Experimental Design by Kris Sankaran



# Large-Scale Inference and Experimental Design

IFT6758, Fall 2020

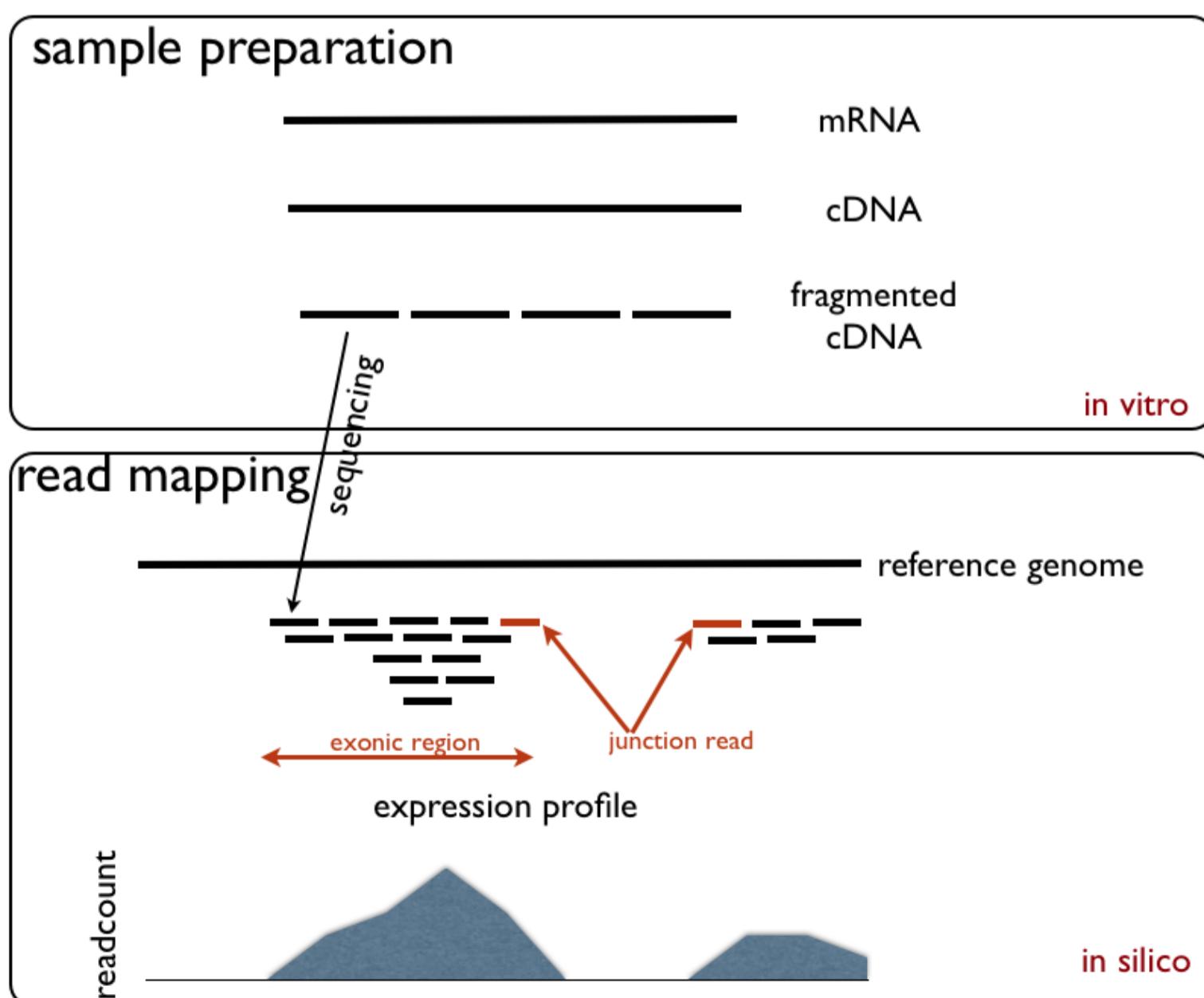
Reading: [MSMB 6.7 - 6.11, 13.1 - 13.4.](#)Optional reading: [CASI 15.1 - 15.3](#)

...

## The Modern Regime

- We have high-throughput data collection schemes, and it's now routine to collect many measurements per sample.
- These sensors collect many aspects of each sample, and it's no longer possible (or desirable) to specify the relevant features to collect in advance

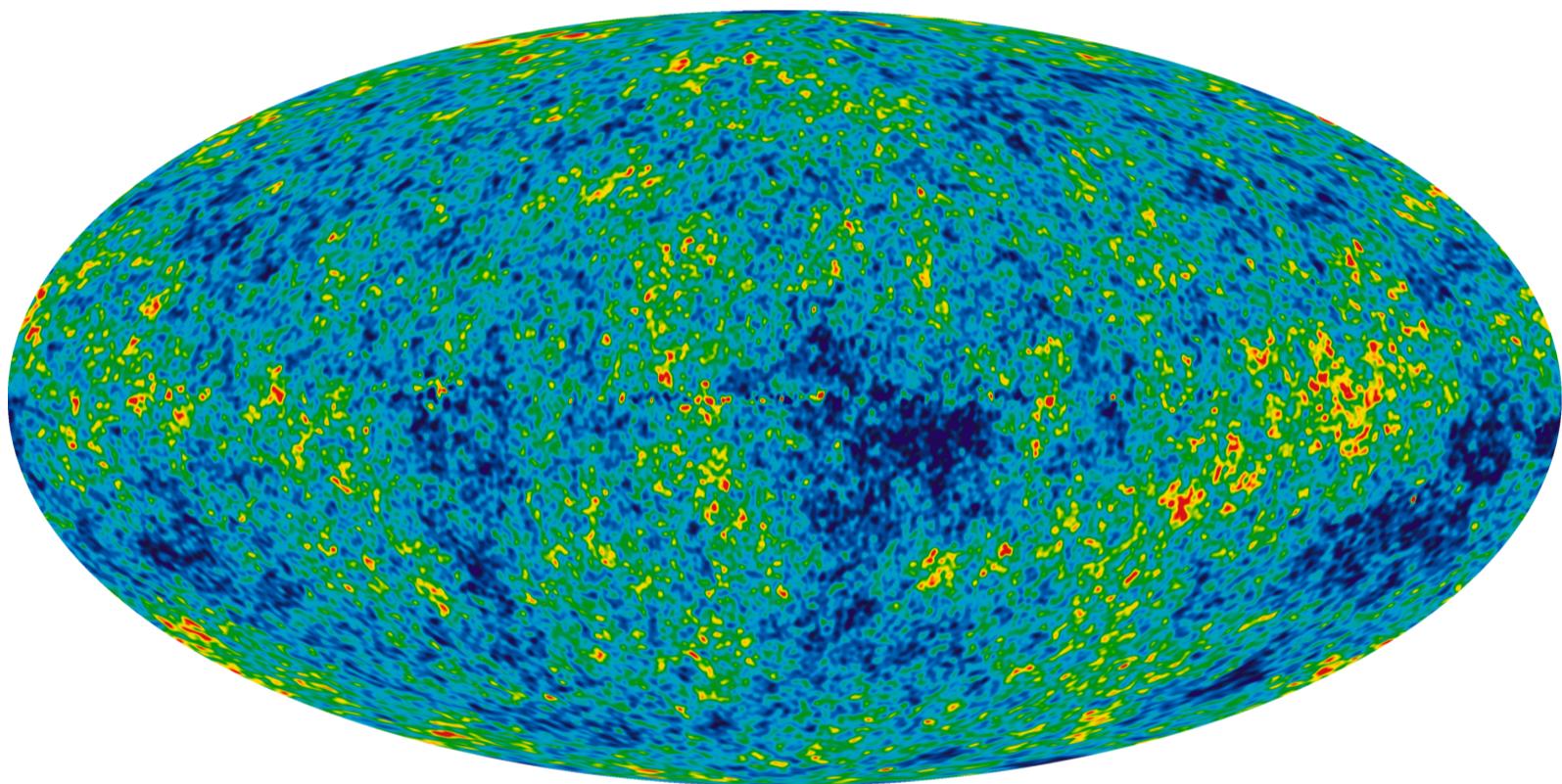
?



Are any of these genes associated with a disease of interest?

## The Modern Regime

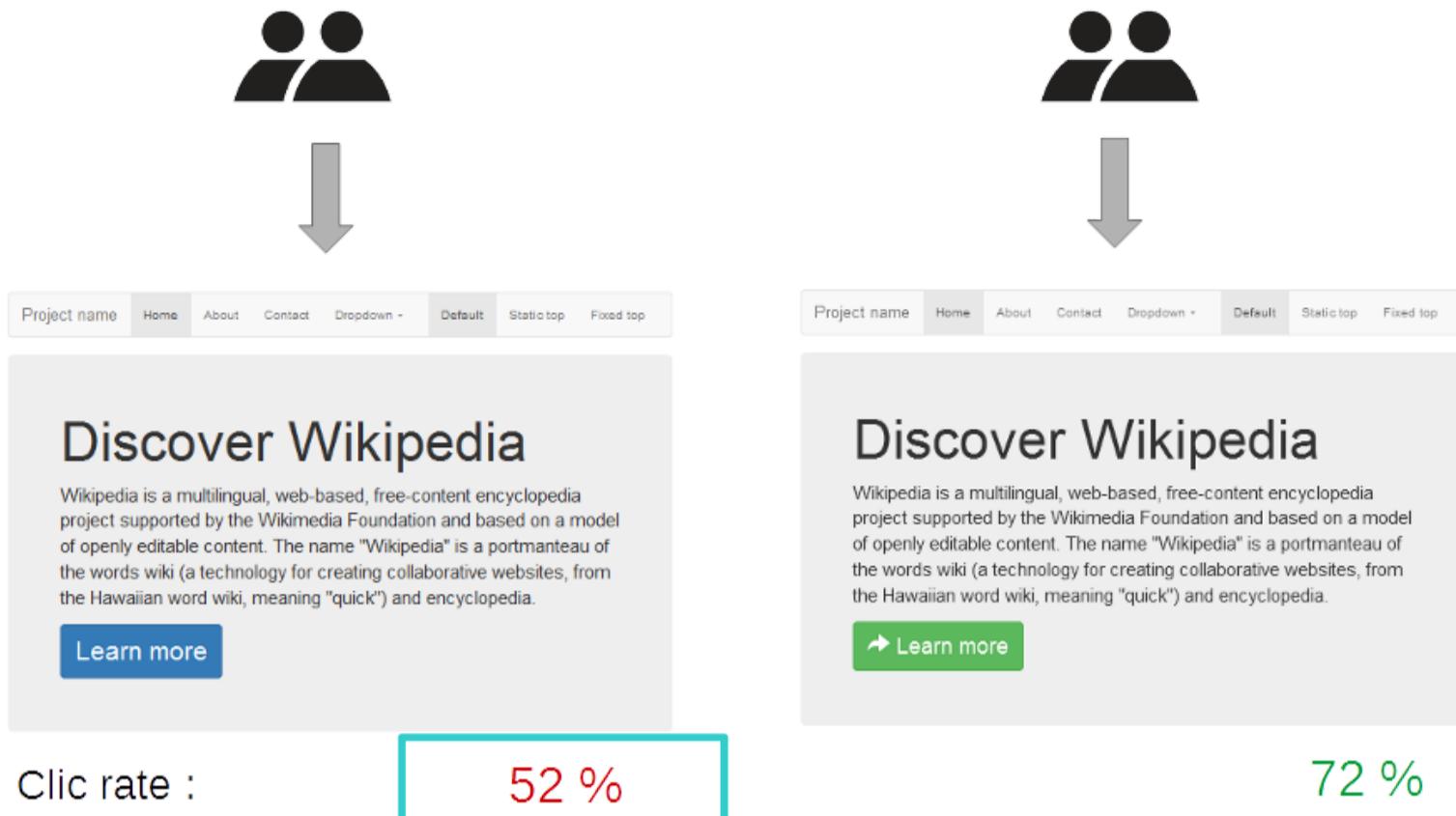
- We have high-throughput data collection schemes, and it's now routine to collect many measurements per sample.
- These sensors collect many aspects of each sample, and it's no longer possible (or desirable) to specify the relevant features to collect in advance



Is anywhere in the cosmic microwave background nongaussian?

## The Modern Regime

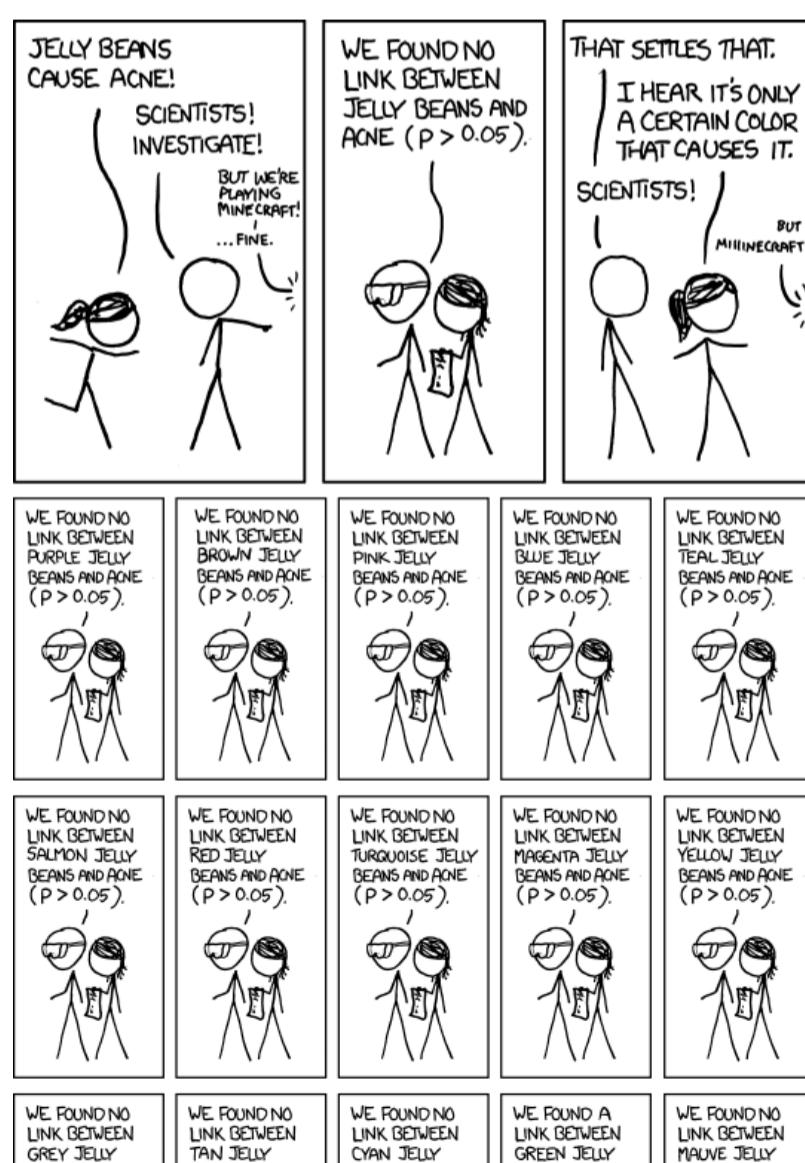
- We have high-throughput data collection schemes, and it's now routine to collect many measurements per sample.
- These sensors collect many aspects of each sample, and it's no longer possible (or desirable) to specify the relevant features to collect in advance

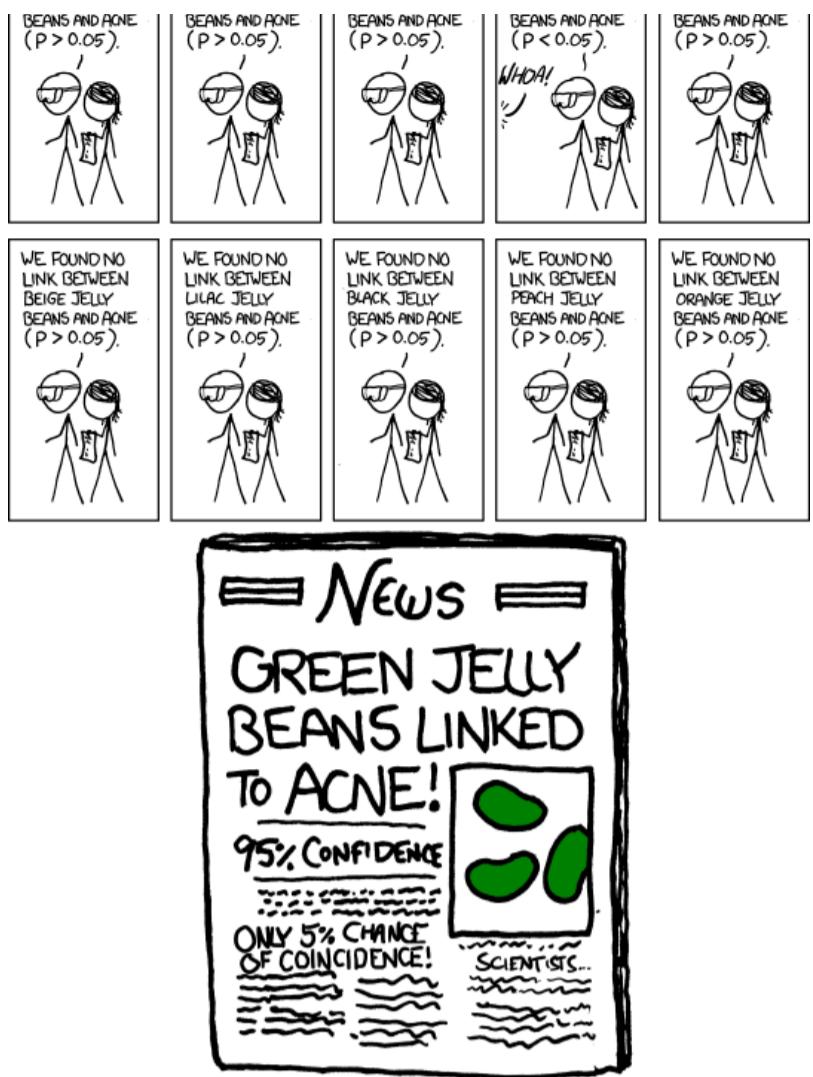


Among many possible changes to interface design, do any actually affect user behavior?

## Let's take a fun example

### Source





## Abstraction

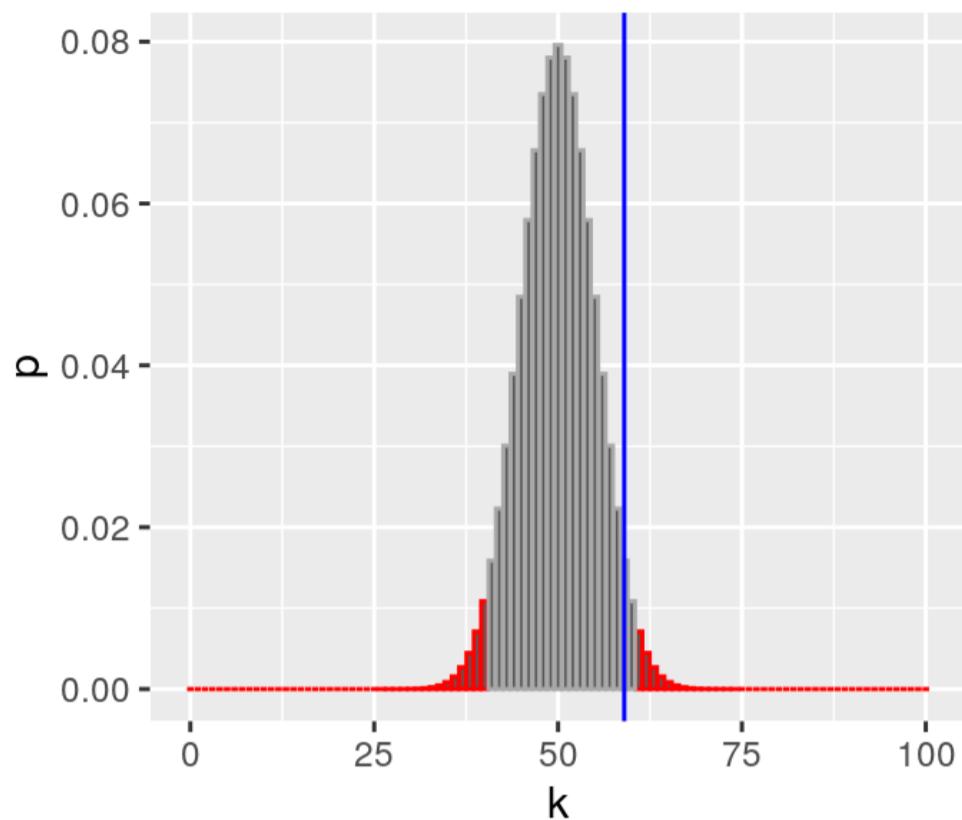
- You need to scan through a large number of features to extract those that are most relevant
- The ones that are interesting deviate from some pre-existing model of the system
- Let's associate feature  $i$  with a hypothesis test,

$H_{i0}$  : Feature follows expected behavior

$H_{i1}$  : Feature is of interest

## False Discoveries

- By design, even if everything really is null, some fraction of tests will result in rejections



By chance, you will end up in the red area  $\approx \alpha\%$  of the time. Recall: jelly beans!

?

## Family-Wise Error Rate

- How can we protect against this?
- One idea is to control the probability of any falsely rejected hypothesis, when they are all in fact null

$$\mathbb{P}_{\cap H_{i0}} [\text{any } H_{i0} \text{ is rejected}] \leq \alpha$$

## Notation

- How can we protect against this?
- One idea is to control the probability of any falsely rejected hypothesis, when they are all in fact null

$$\mathbb{P}_{\cap_m H_{i0}} [\text{any } H_{i0} \text{ is rejected}] \leq \alpha$$

?

## Notation

We need some new notation, suited to this point of view.

	Null is true	Null is false	Total
Rejected	$V$	$S$	$R$
Not Rejected	$U$	$T$	$m - R$
Total	$m_0$	$m - m_0$	$m$

## Notation

Some interpretations so it doesn't become too confusing

some interpretations, so it doesn't become too confusing.

	Null is true	Null is false	Total
Rejected	False alarm	Correct detection	Number alarms
Not Rejected	Proper silence	Missed signal	Number silent
Total	Actually uninteresting	Actually relevant	Number scanned

like, confusion matrix

## Family-Wise Error Rate

With this notation, the goal of FWER is to find a way to evaluating hypothesis so that

$$\mathbb{P}[V > 0] \leq \alpha$$

?

## Probability of false positives

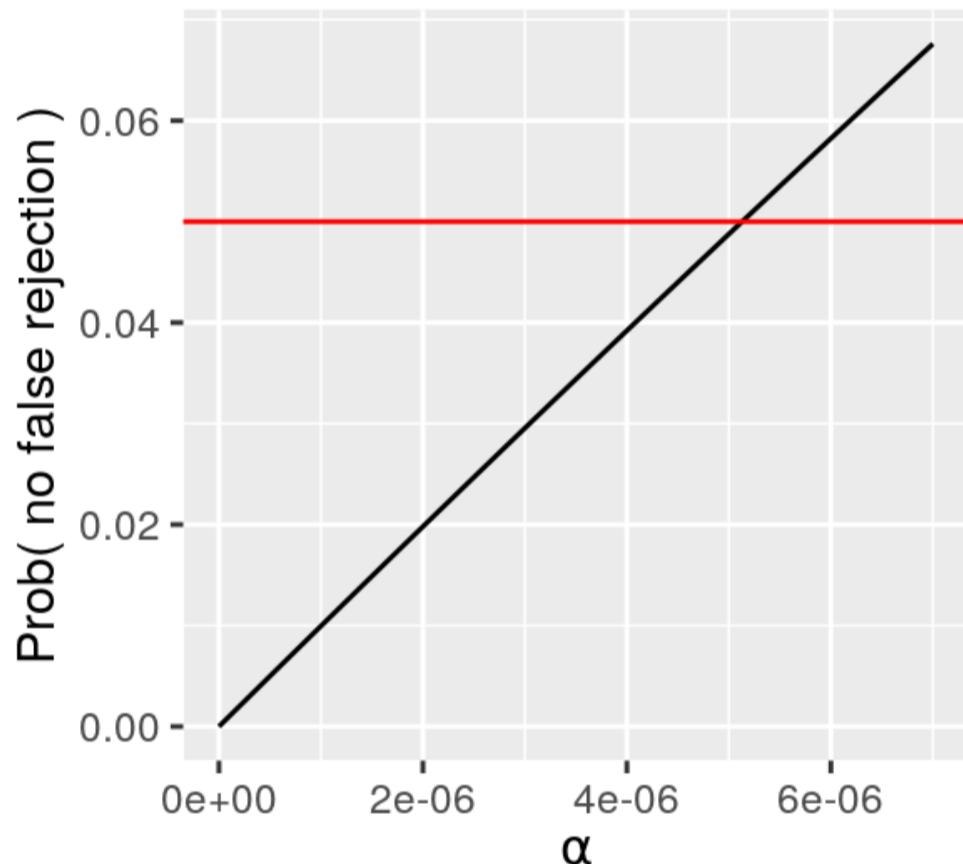
With many hypotheses, you'll almost definitely have at least one false alarm.  $\alpha$ : probability of false alarm in one sample

$$\begin{aligned}\mathbb{P}[V > 0] &= 1 - \mathbb{P}[V = 0] \\ &= 1 - (1 - \alpha)^{m_0} \\ &\xrightarrow{m_0 \rightarrow \infty} 1,\end{aligned}$$

except when  $\alpha = 0$  (and we never reject anything).

## Bonferroni Correction

- One way to do this is to make the original tests more stringent.
- What rejection threshold  $\alpha$  would we need, to control the FWER below 0.05? Here's a picture when  $m_0 = 10,000$ .



The black line is the function  $1 - (1 - \alpha)^{m_0}$ . The crossing happens at  $\frac{0.05}{10,000}$ . The idea of using  $\frac{\alpha}{m_0}$  as a rejection threshold is called the "Bonferroni correction."

## Justification

Why does this work?

$$\mathbb{P}[V > 0] = \mathbb{P}[\cup_{i=1}^{m_0} \{H_{i0} \text{ is rejected}\}]$$

$$\begin{aligned}
 &\leq \sum_{i=0}^m \mathbb{P}[H_{i0} \text{ is rejected}] \\
 &= \sum_{i=0}^m \frac{\alpha}{m} = \alpha
 \end{aligned}$$

## Issues with FWER

- This approach becomes very conservative with even a moderate number of tests
- Even if there are *few* interesting hypotheses, you'll never reject them with such a stringent threshold

## False Discovery Rate

Instead of trying to prevent the occurrence of any false positive, just try to control their frequency.

Formally, let's provide a guarantee on

$$FDR := \mathbb{E} \left[ \frac{V}{R + 1} \right],$$

the expected proportion of false positives, among all the

rejected hypothesis.

*The number of false alarms is allowed to grow with the number of hypotheses tested.*

- We want to reject those hypotheses for which  $\text{FDR}(p)$  is less than  $\alpha$ .

## Estimating FDR

Consider the histogram of the  $p$ -values across all hypothesis, both those that are true and those that are not.

- If a test statistic is drawn from the null, the associated  $p$ -value will be uniform
- Otherwise, you will expect it to be small

Ref: [How to interpret a p-value histogram](#)

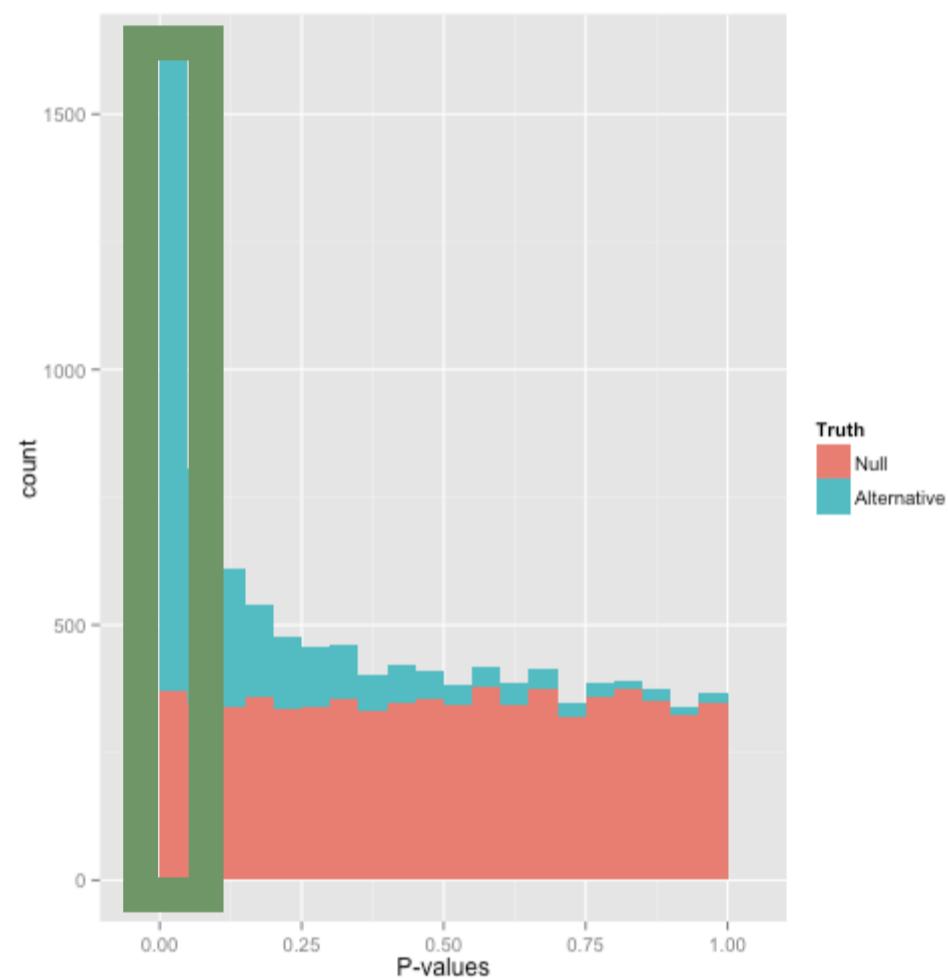
## Digression: Uniformity under null

- Under the null, a  $p$ -value has 5% chance of being  $\leq 0.05$ 
  - It gives the probability of a false alarm under the null
- More generally, it has  $\alpha$  chance of being smaller than  $\alpha$ .
- Therefore, under the null,  $p$ -values are uniform

Ref: [How to interpret a p-value histogram](#)

## Estimating FDR

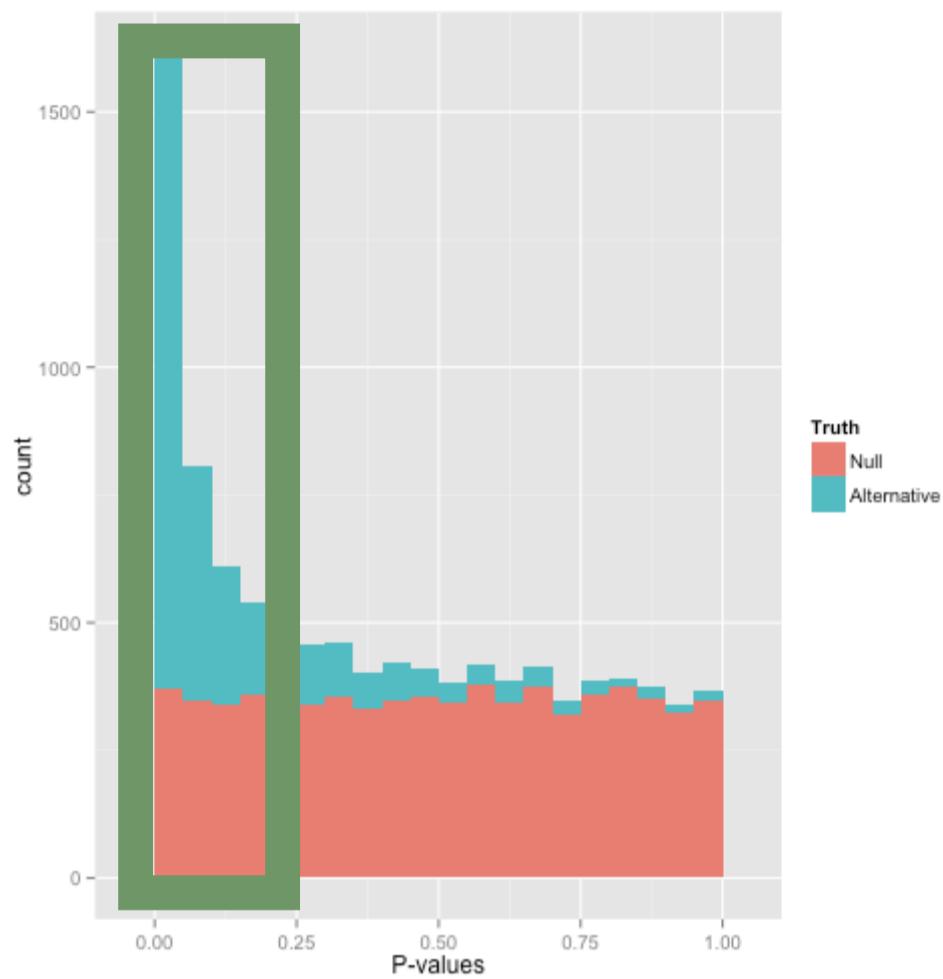
- You will reject all hypotheses to the left of the vertical bar
- The fraction of the rejection area that seems to come from the uniform part is an estimate of the *FDR*



Most of the rejection (green) region is blue, so the FDR will be small.

## Estimating FDR

- You will reject all hypotheses to the left of the vertical bar
- The fraction of the rejection area that seems to come from the uniform part is an estimate of the *FDR*

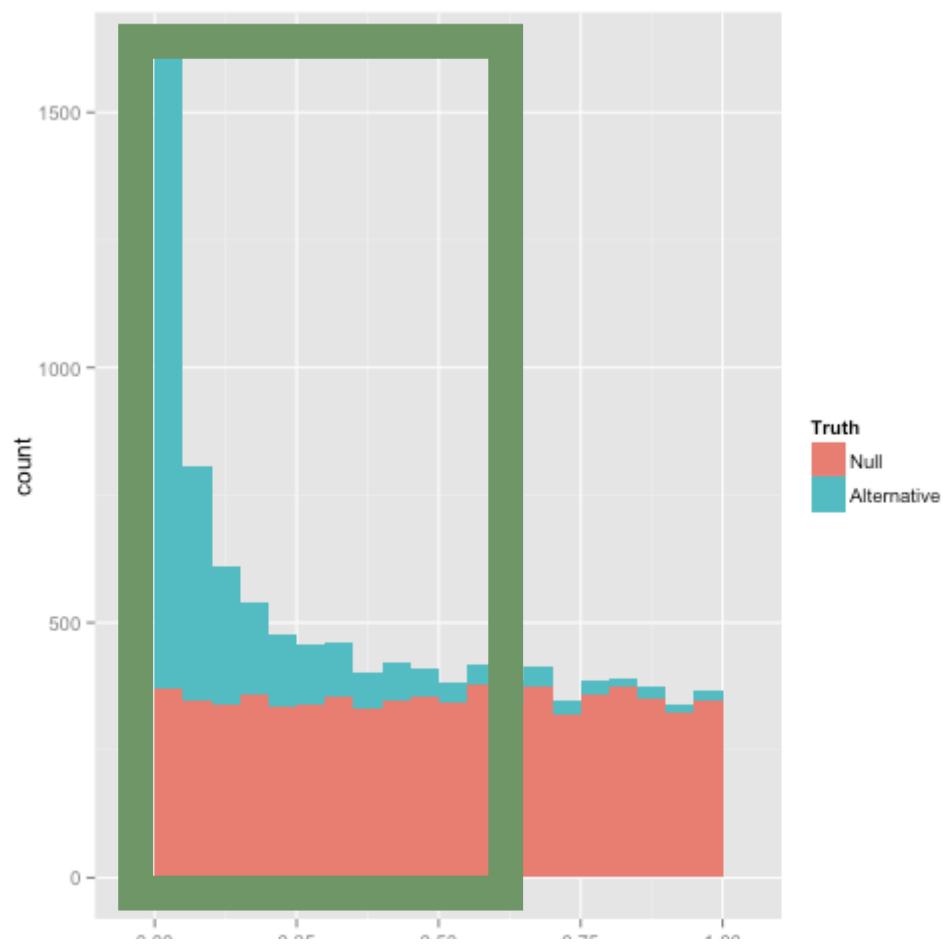


Now the FDR is larger.

?

## Estimating FDR

- You will reject all hypotheses to the left of the vertical bar
- The fraction of the rejection area that seems to come from the uniform part is an estimate of the *FDR*



And now even larger.

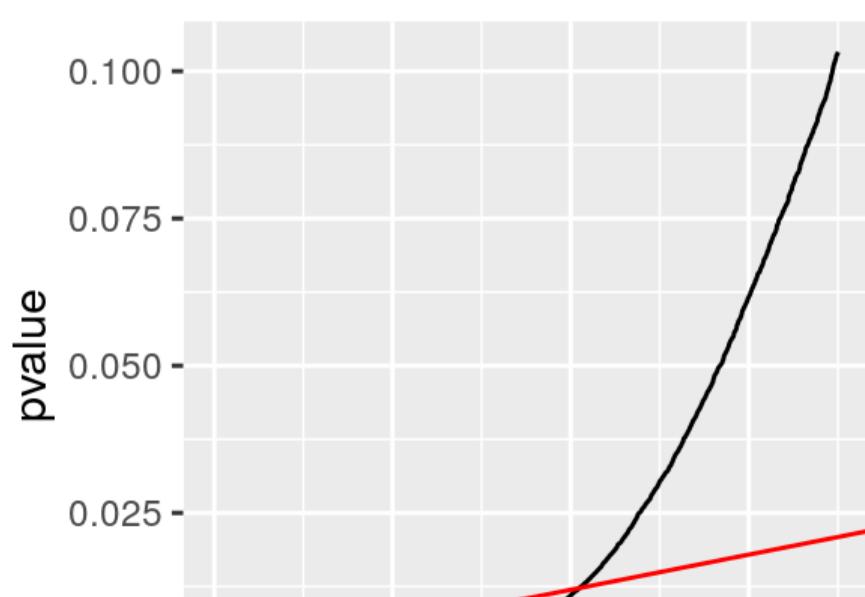
## Benjamini-Hochberg (BH) procedure

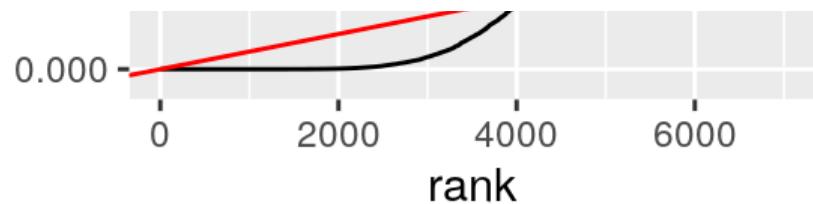
- When the  $p$ -value histogram is almost uniform, at most a small percentage of hypotheses are non-null
- The BH procedure to identify that percentage. It lets us control the FDR
- Intuition: Make the rectangle as large as you can (make the most rejections), while limiting the estimated FDR

?

## Formal Definition

- Sort the  $p$ -values,  $p_{(1)}, \dots, p_{(m)}$
- Find the last of the  $p_{(k)}$  below the line  $\frac{\alpha k}{m}$
- Reject all hypotheses with  $p$ -values smaller than that





## Justification

- This procedure is not as arbitrary as it might seem at first
- We need to quantify the intuition from the histogram

Since the histogram has both a null (uniform) and an alternative component, we can think of it as a mixture density,

$$f(p) = \pi_0 \text{Unif}[0, 1] + (1 - \pi_0) f_{alt}(p)$$



## Justification

Let's represent the "instantaneous" false discovery rate associated with any point on the histogram using

$$fdr(p) = \frac{\text{null part}}{\text{rejection total}} = \frac{\pi_0}{f(p)},$$

where  $f(p)$  is the density associated with the mixture histogram.



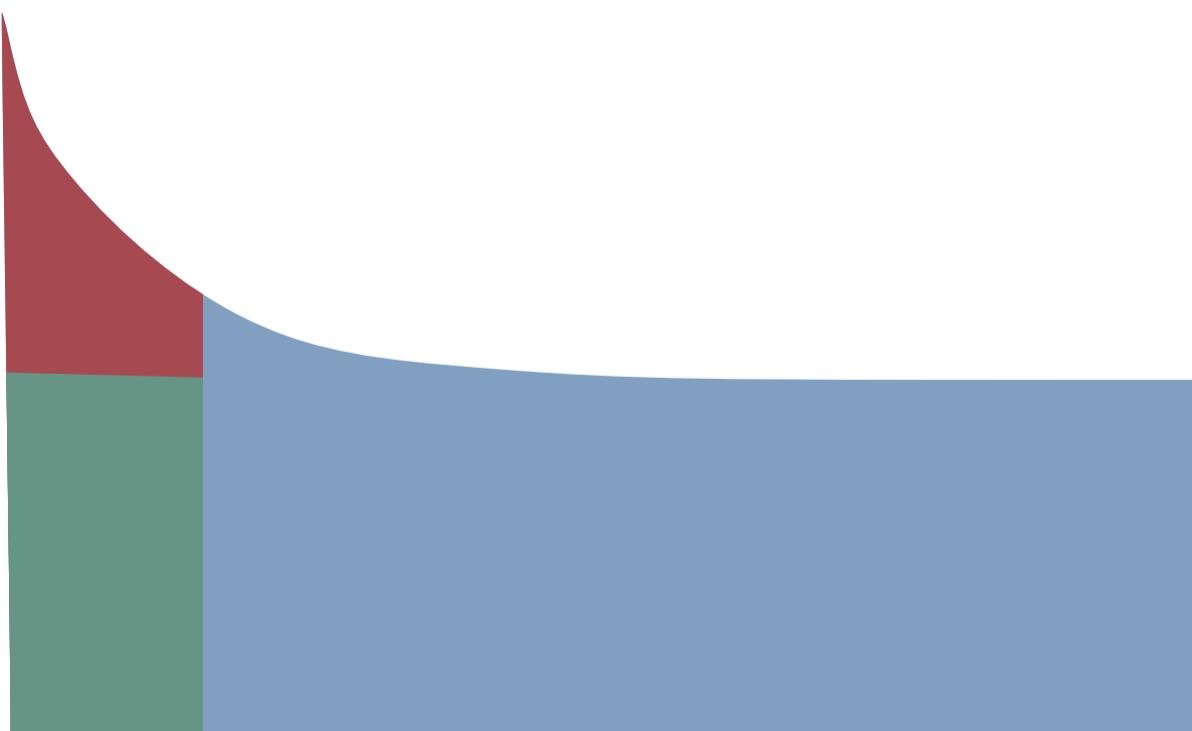
...

## Justification

Let  $F$  represent the CDF associated with this mixture. Then we can define a false discovery rate in terms of the ratios of the areas in the histograms,

$$Fdr(p) = \frac{\text{null part to left}}{\text{rejection total to left}} = \frac{\pi_0 p}{F(p)},$$

where  $F(p)$  is the CDF associated with the mixture histogram.



?

## Justification

We'll estimate the FDR by plugging in the empirical  $p$ -value distribution in for  $F$ ,

$$\widehat{Fdr}(p) = \frac{\pi_0 p}{\hat{F}(p)}$$

We want the largest  $p$  which still satisfies  $\widehat{Fdr}(p) \leq \alpha$ .

## Justification

Think about the smallest  $p$ -value,  $p_{(1)}$ . Notice that,

$$\widehat{Fdr}(p_{(1)}) = \frac{\pi_0 p_{(1)}}{1/m},$$

so that this estimate is smaller than  $\alpha$  if and only if

$$p_{(1)} \leq \frac{\alpha}{m\pi_0}.$$

So if  $p_{(1)} \leq \frac{\alpha}{m}$ , that's good enough (because  $\pi_0 \leq 1$ ).

## Justification

Similarly, for the  $k^{th}$ -largest  $p$ -value,

$$\widehat{Fdr} (p_{(k)}) = \frac{\pi_0 p_{(k)}}{k/m} \leq \alpha$$

if and only if,

$$p_{(k)} \leq \frac{k\alpha}{m\pi_0},$$

and it would be good enough if  $p_{(k)} \leq \frac{k\alpha}{m}$ .

## Justification

Therefore, the BH procedure is just trying to use the observed histogram of  $p$ -values to select the largest  $p_{(k)}$  so that the estimated  $\widehat{Fdr}$  is less than  $\alpha$ .

## The Multiple Testing problem opportunity

- Usually, people think of the fact that we have to screen many hypotheses as a problem
  - (because you can get false positives)
- However, all these extra (mostly null) hypotheses give us

extra information

- We were able to use them to estimate the FDR, and it's in fact possible to use to estimate reference distributions
  - No need for theory-driven references

## Experimental Design

?

### Why Experimental Design?

- The effects you observe can be driven by many factors
- To draw effective inferences, need to minimize potential for alternative explanations
- The best data scientists have an understanding of (or some control over) the collection process

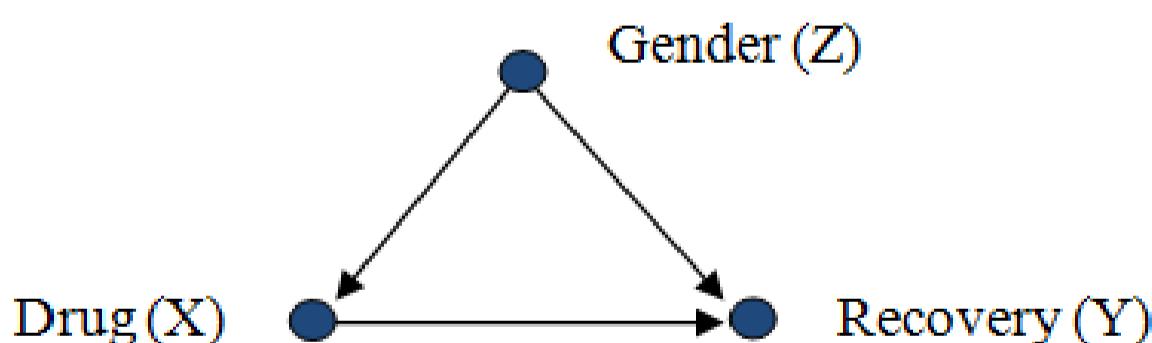
## Basic Principles

- Confounding
- Blocking
- Power analysis

?

## Confounding

$X$  and  $Y$  are *confounded* by  $Z$  whenever  $Z$  causally influences both  $X$  and  $Y$ .



So, effect of  $Z$  should be accounted for while computing the effect of  $X$  on  $Y$ .

## Blocking

Blocking: arranging of experimental units in groups (blocks) that are similar to one another

Aim: reduce unexplained variability

Principle: variability which cannot be overcome is confounded

- Large, systematic sources of variation should be "blocked"
- This makes it possible to see the residual variation

Example: An experiment is designed to test a new drug on patients. There are two levels of the treatment, drug, and placebo.

**Blocking factor:** gender of the patients

?

## Blocking

You are asked to evaluate whether a new type of shoe sole will last longer than the previous design. How should you design your experiment?

- Predictors (e.g. raw materials)? Response (e.g. longevity)?
- What are possible blocking factors (e.g., location, climate)?



## Power Analysis

The power of any study is driven by

- 1) The sample size used to estimate effects of interest
- 2) The variability in the outcome of interest, within otherwise comparable samples
- 3) The underlying **effect size** on the outcome of interest
  - a way of **quantifying the size of the difference between two groups**: *standardized mean difference* between *experimental group and control group*
  - Effect size and *Z-score*
  - Effect size and its confidence interval

?

## Power Analysis

Unfortunately, (3) is very hard to know before starting the study. There are a few common ways around this,

- Simulate: Generate data that look like what you expect to see, under a few different effect sizes. See how data collection and analysis choices influence the downstream power, across effect sizes.
- Pilot study: Get a small preliminary study, for the explicit purpose of getting a ballpark effect size estimate.
- Sequential design: Conduct your study in stages, getting better estimates of the effect over time.

## Types of Experiments (informal)

- Controlled Experiment: You have control over all factors that may influence the outcome
- Study: You do not have control over some factors
- Meta-Analysis: You combine the results from many experiments

## Variants: Observational Study

- In these studies, you control neither the samples that are collected nor the treatments they were assigned
- Be wary that "correlation is not causation"
  - Examples of illogically inferring causation from correlation
    1. The faster that windmills are observed to rotate, the more wind is observed. Therefore, wind is caused by the rotation of windmills.
    2. Children that watch a lot of TV are the most violent. Clearly, TV makes children more violent.
  - Ref: Correlation does not imply causation

?

## Randomized Control Trial

- You have no control over most factors, but you **can** randomize the assignment of the treatment
- By randomizing the assignment, other influences will cancel out, on average
- Here, correlation *does* imply causation (assuming that the model is correct...)

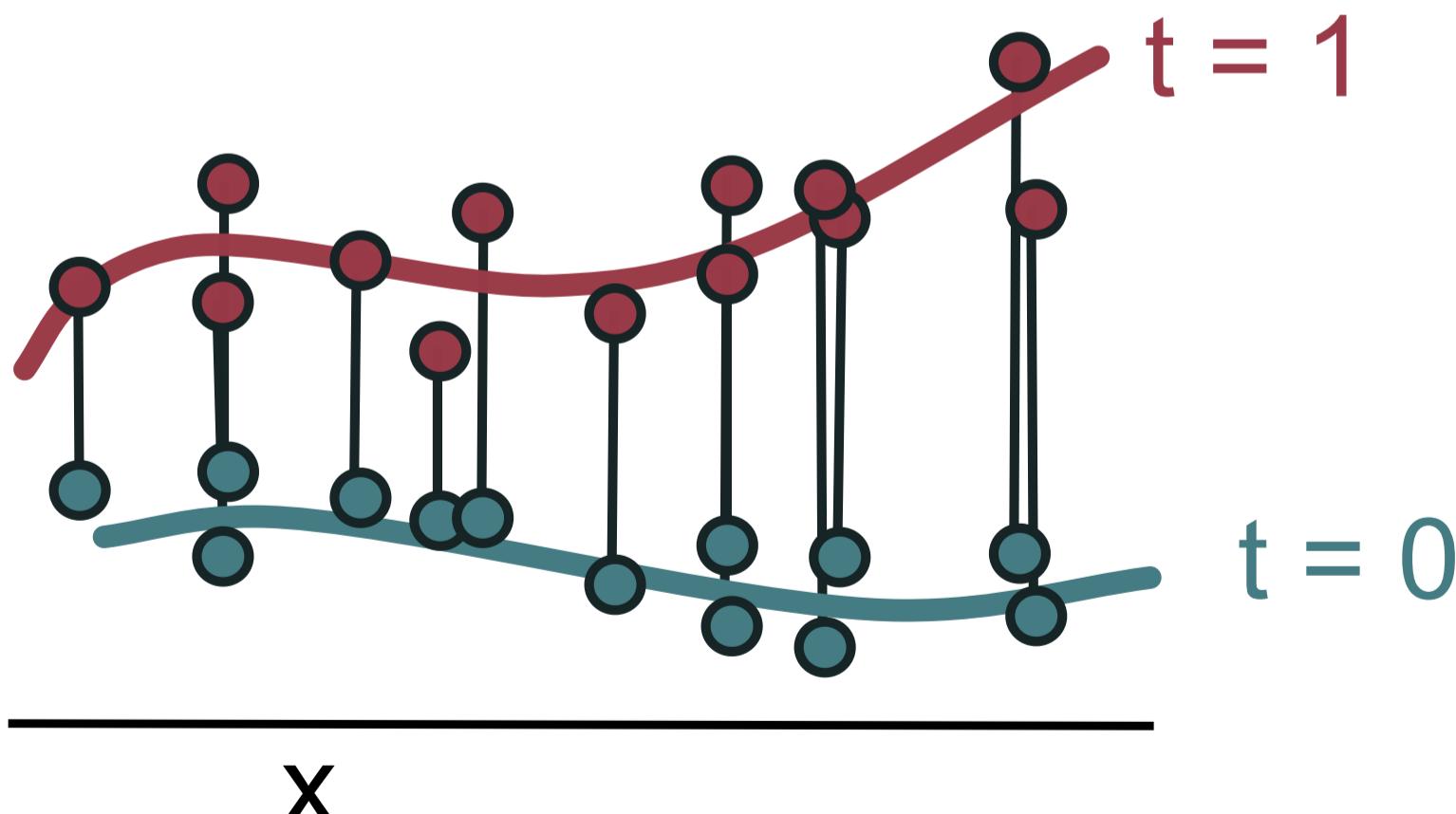
## Types of Experiments (formal)

Formally, there is a process influencing the outcome,

$$y_i = f(x_i, t_i) + \epsilon_i$$

?

- $t_i$  is the treatment,
- $x_i$  are measured influencing factors
- $\epsilon_i$  are i.i.d. with mean 0 (from unmeasured factors)



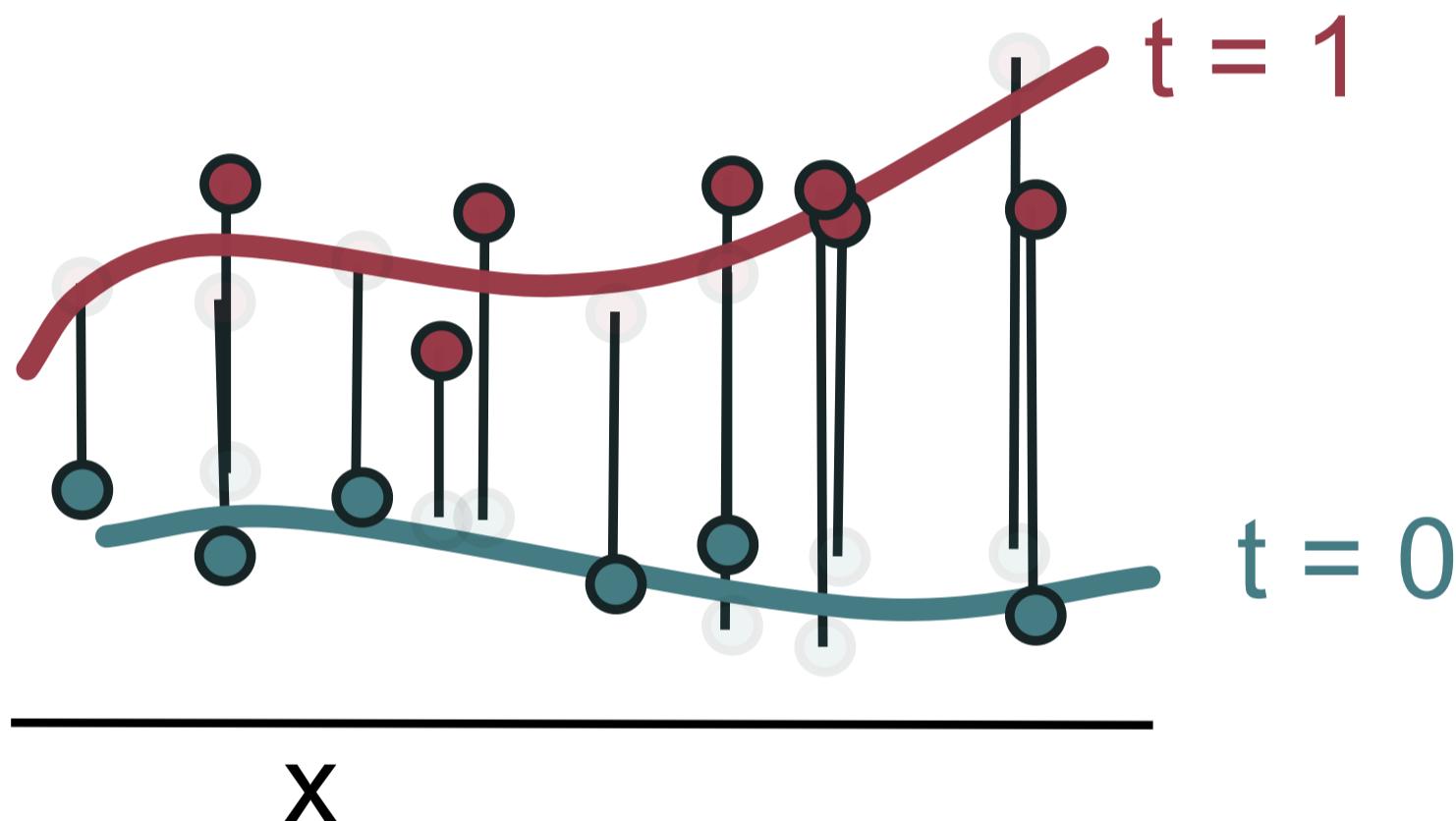
In theory, each sample could have two outcomes, depending on whether they are given the treatment or not.

## Types of Experiments (formal)

Formally, there is a process influencing the outcome,

$$y_i = f(x_i, t_i) + \epsilon_i$$

- $t_i$  is the treatment,
- $x_i$  are measured influencing factors
- $\epsilon_i$  are i.i.d. with mean 0 (from unmeasured factors)



In reality, you can observe each sample under either the treatment or the control (not both).

## Types of Experiments (formal)

Formally, there is a process influencing the outcome,

$$y_i = f(x_i, t_i) + \epsilon_i$$

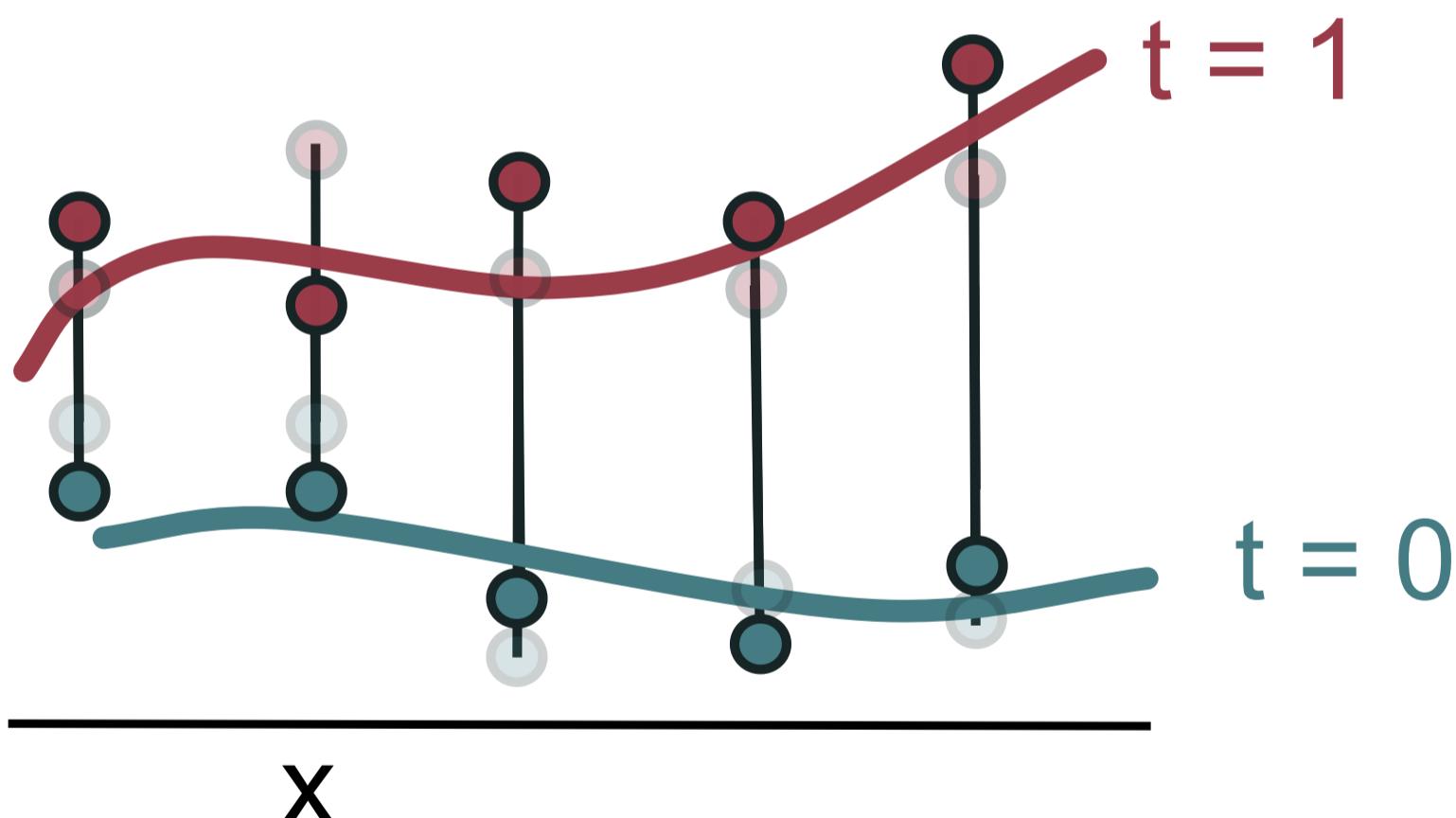
- Controlled experiment: Can manipulate all  $x_i, t_i$ , and  $\epsilon_i \equiv 0$ .

- Study: Able to manipulate some of the  $x_i$  and / or  $t_i$
- Observational study: Cannot manipulate any of  $x_i$  or  $t_i$ <sup>\*</sup>
- Randomized controlled study: Randomize over  $t_i$ , can't manipulate  $x_i$

## Controlled Experiment

$$y_i = f(x_i, t_i) + \epsilon_i$$

Can manipulate all  $x_i$ ,  $t_i$ , and  $\epsilon_i \equiv 0$ . In the example here, we get both a treatment and a control sample on an even grid of  $x$ 's.

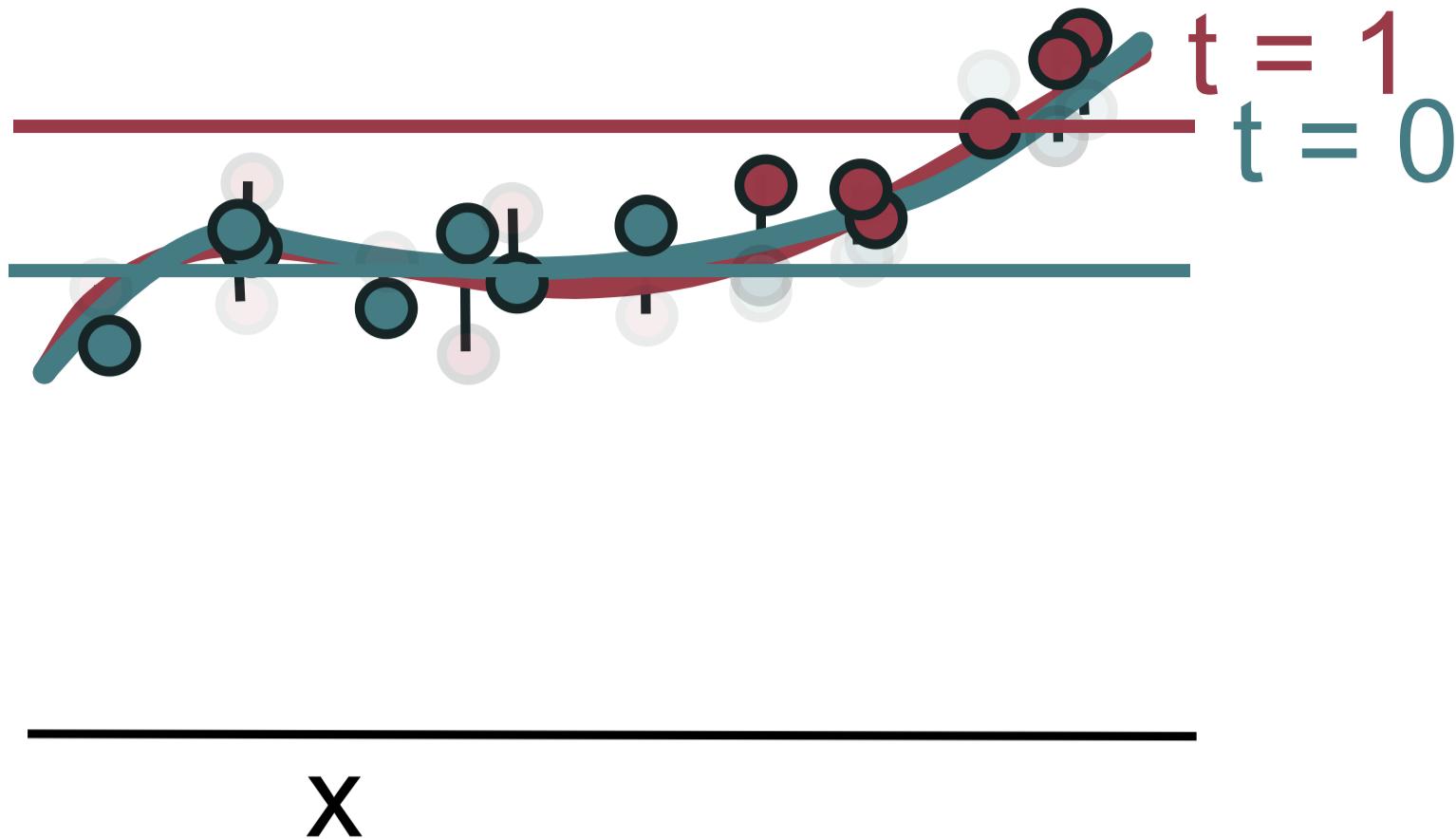


## Observational Study

$$y_i = f(x_i, t_i) + \epsilon_i$$

Have to guard against confounders

have to guard against confounds.



## Randomized Controlled Trial

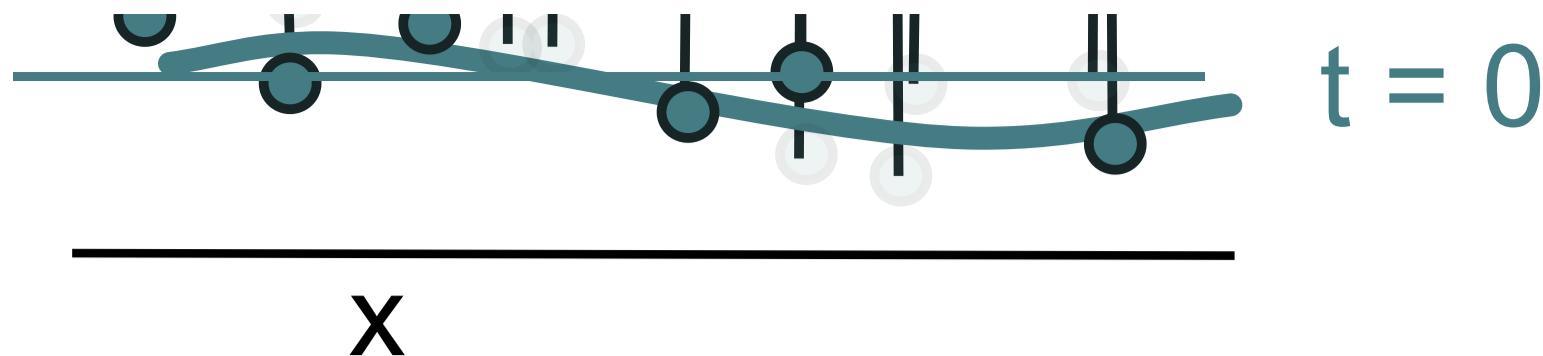
$$y_i = f(x_i, t_i) + \epsilon_i$$

Randomize over  $t_i$ , can't manipulate  $x_i$ . When the treatment effect is uniform over  $x$ , this randomization causes any effect of  $t_i$  to get averaged out, leaving only the overall effect of  $x$ .

Chain rule:  $P(y|x) = \sum_t P(y|x, t)P(t|x)$

When it's not constant over  $x$  (like the figure below), there is some approximation error.





## Data Quality

Data quality is a vague term.

Instead, consider your data's *fitness for purpose*.

What you want to do will guide your data collection and analysis strategies.

?

## Data Quality

Section of plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of the plane	1.8

A famous **story** about Abraham Wald (founder of decision theory). Where should you put the armor?

# Data Science in context



Figure from [this post](#). Some other useful references are,

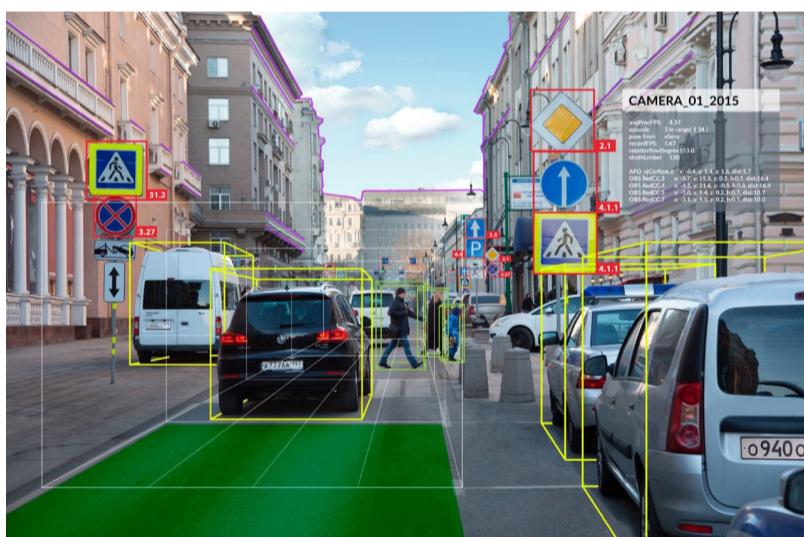
- How to share data with a statistician
  - e.g., raw data, clean data, metadata
- Design Study Methodology
  - 9 stages of design
  - common pitfalls

Where to next?

- We've reviewed the roles of prediction, unsupervised learning, and statistical inference in data science
- We assumed the following
  - Samples drawn  $x_i \stackrel{i.i.d.}{\sim} F$
  - Generic structure between coordinates of  $x_i$

## Where to next?

- What if there are features that might be useful, but which weren't directly included?
- What if your samples have relationships with one another?
- What if the correlation has very specific types of structure?



```
import {slide} from '@mbostock/slide

<style>

mteq_block = f()

mteq = f()
```



© 2020 Observable, Inc.

[Home](#)   [Teams](#)   [About](#)   [Jobs](#)   [Contact](#)   [Terms](#)

?