Kris Sankaran

Published Oct 2, 2019

Large-Scale Inference and Experimental Design

IFT6758, Fall 2019

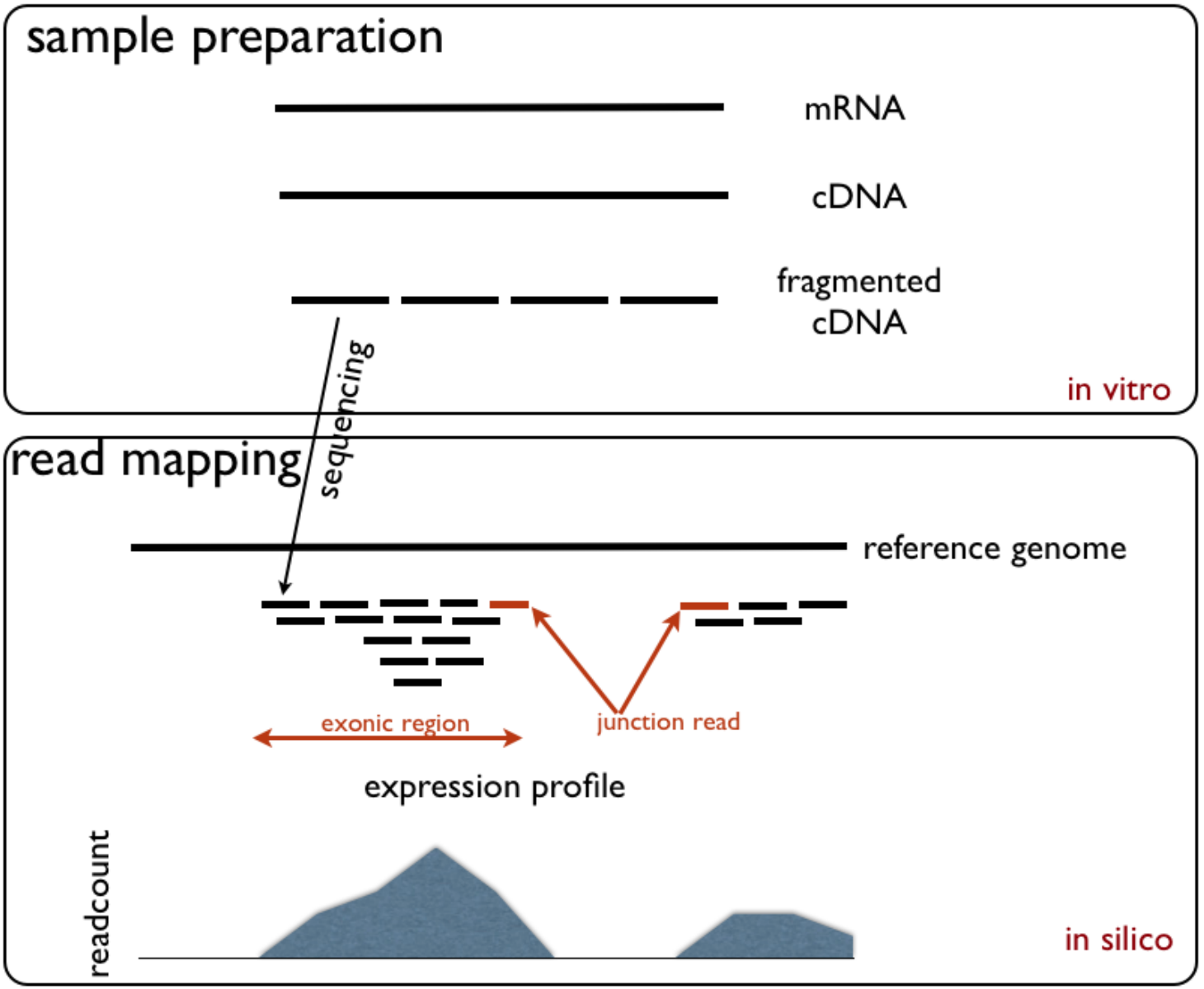
Reading: MSMB 6.7 - 6.11, 13.1 - 13.4.

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

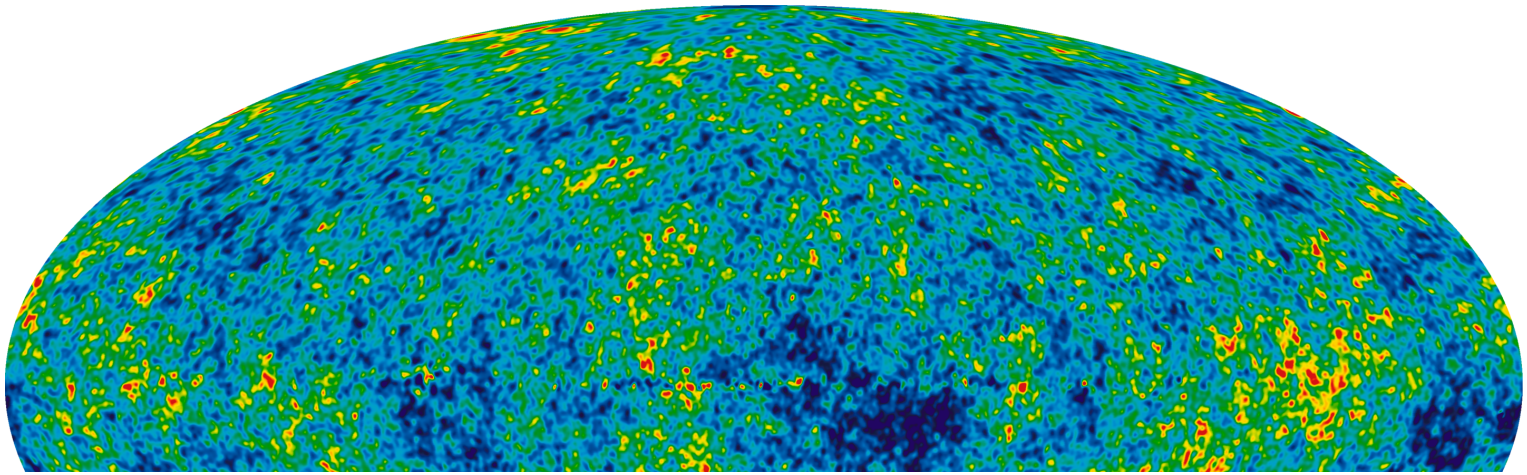
It's no longer possible (or desirable) to specify the relevant features to collect in advance

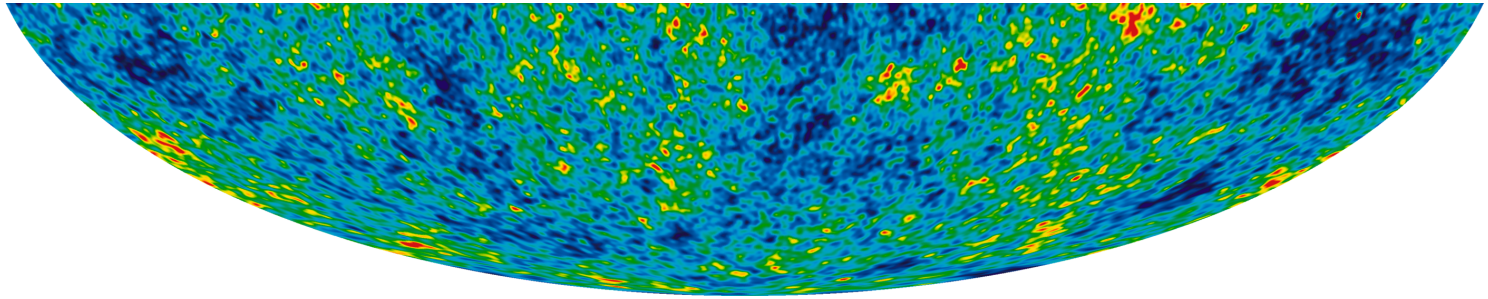


Are any of these genes associated with a disease of interest?

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

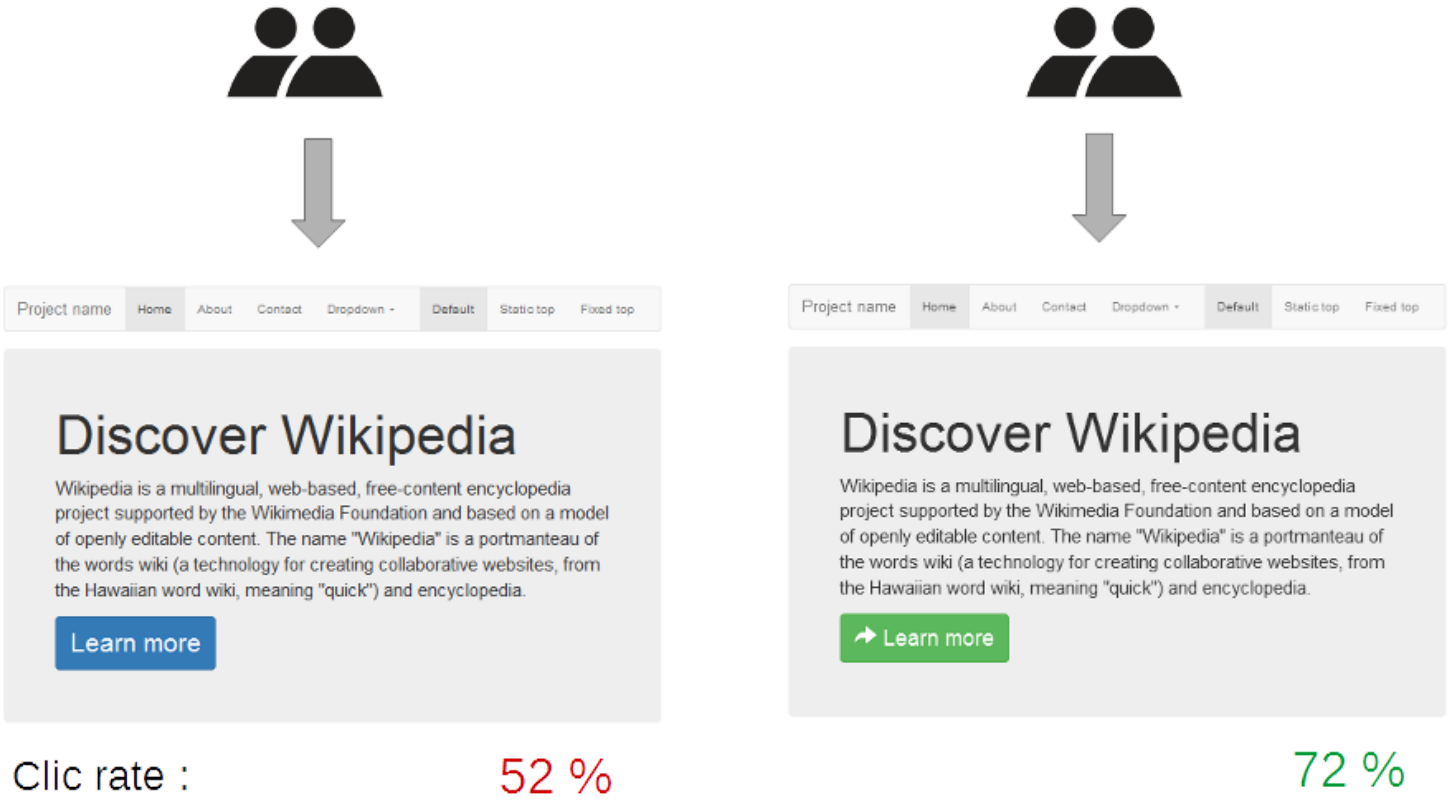




Is anywhere in the cosmic microwave background nongaussian?

The Modern Regime

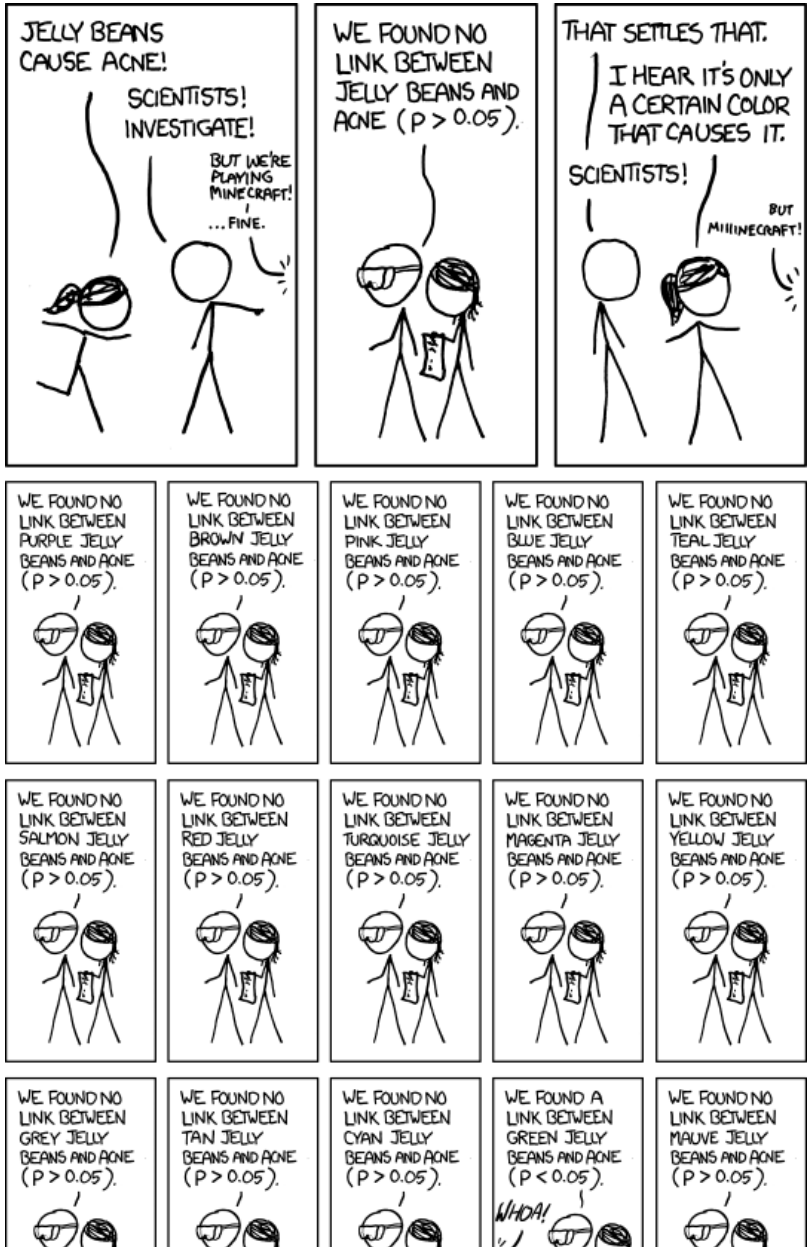
- We have high-throughput data collection schemes, and it's now routine to collect many measurements per sample.
- These sensors collect many aspects of each sample, and it's no longer possible (or desirable) to specify the relevant features to collect in advance

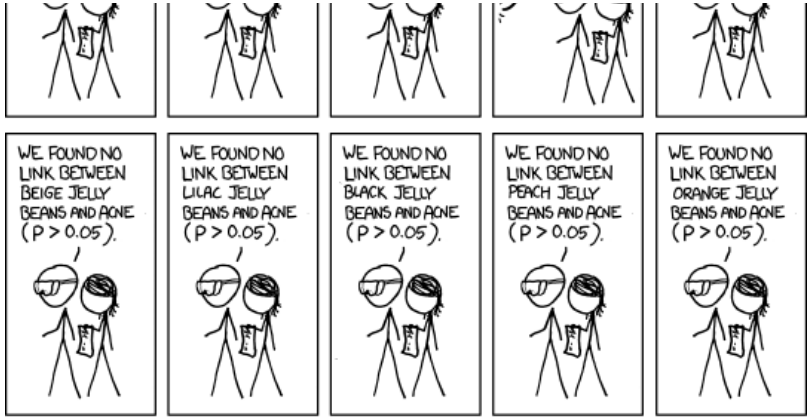


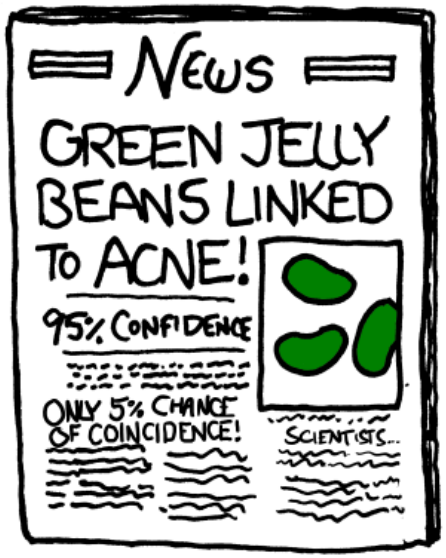
Among many possible changes to interface design, do any actually affect user behavior?

Alternatively

If these examples are too serious for you




WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Abstraction

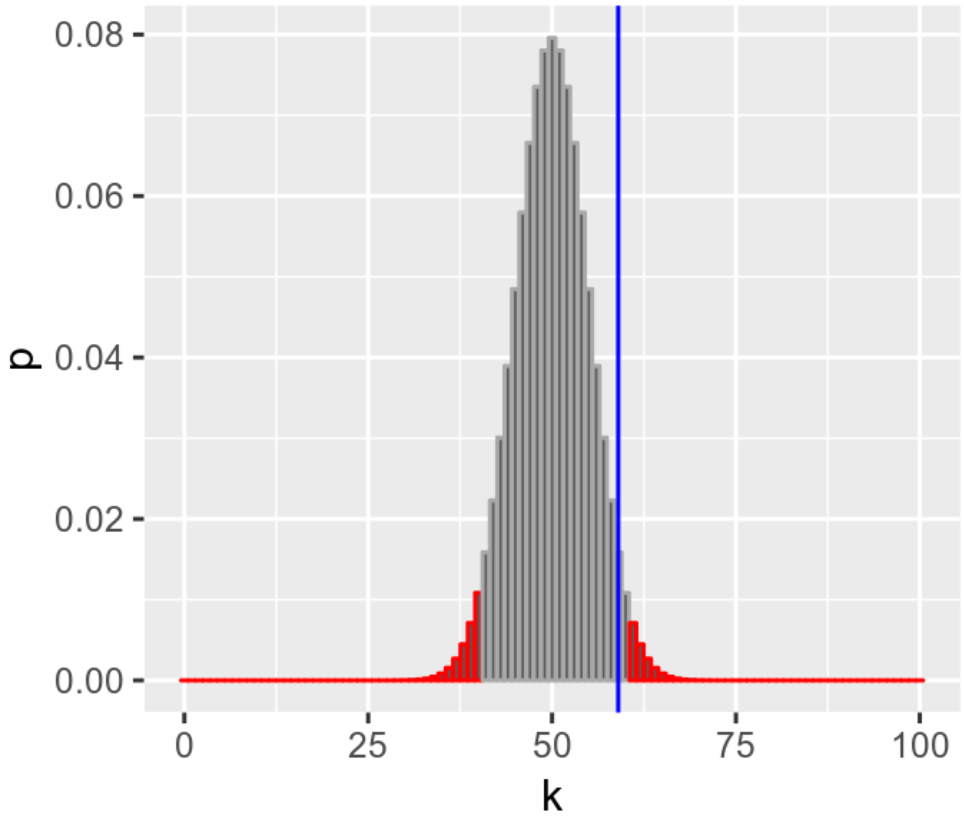
- You need to scan through a large number of features to extract those that are most relevant
- The ones that are interesting deviate from some pre-existing model of the system
- Let's associate feature i with a hypothesis test,

$$H_{i0} : \text{Feature follows expected behavior}$$

$$H_{i1} : \text{Feature is of interest}$$

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. This is the point of the XKCD cartoon.

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

$$\mathbb{P}_{\cap H_{0i}}\left[\text{any } H_{i0} \text{ is rejected }\right] \leq \alpha$$

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

$$\mathbb{P}_{\cap H_{0i}}\left[\text{any } H_{i0} \text{ is rejected }\right] \leq \alpha$$

Notation

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

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.

	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

Total Actually uninteresting Actually relevant Number declared

Family-Wise Error Rate

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

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

Probability of no false positives

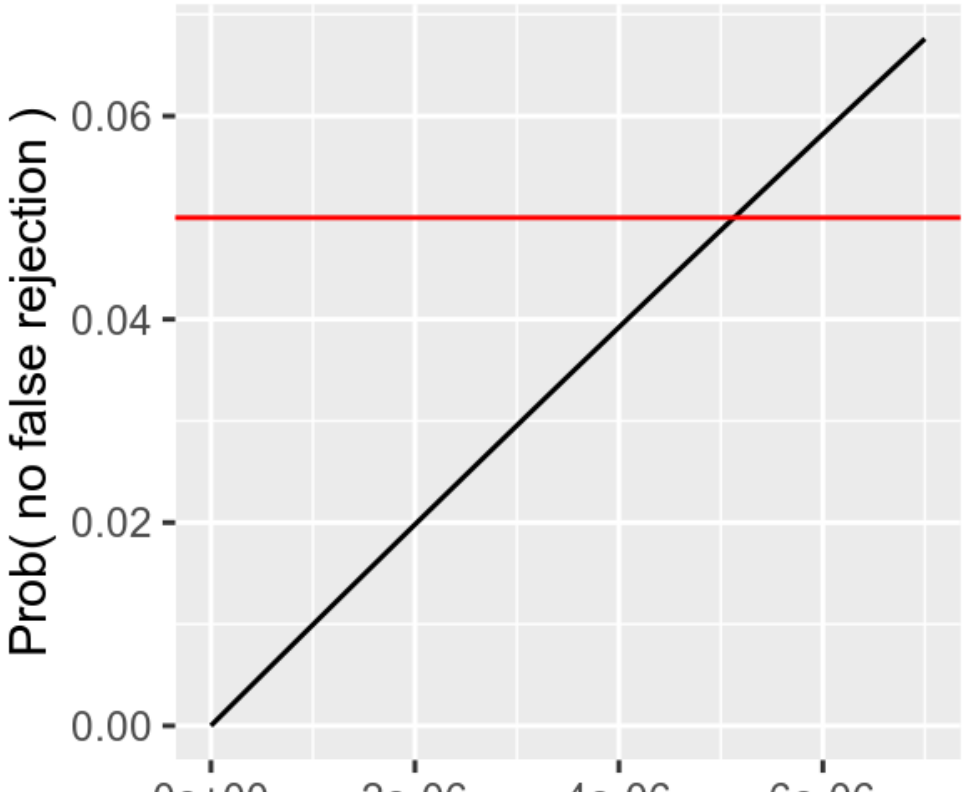
With many hypotheses, you'll almost definitely have at least one false alarm.

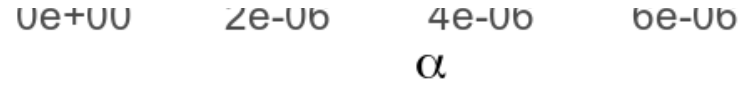
$$\begin{aligned}\mathbb{P}\left[V > 0\right] &= 1 - \mathbb{P}\left[V = 0\right] \\ &= 1 - \left(1 - \alpha\right)^{m_0} \\ &\xrightarrow{m_0 \to \infty} 1,\end{aligned}$$

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

Justification

Why does this work?

$$\begin{aligned}\mathbb{P}\left[V > 0\right] &= \mathbb{P}\left[\cup_{i=1}^{m}\{H_{0i} \text{ is rejected}\}\right] \\ &\leq \sum_{i=1}^{m} \mathbb{P}\left[H_{0i} \text{ is rejected}\right] \\ &\sum^{m} \alpha\end{aligned}$$

$$= \sum_{i=1} \frac{}{m} = \alpha$$

Issues with FWER

- This approach becomes very conservative with even a moderate number of tests
- Even if there *are* a 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

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

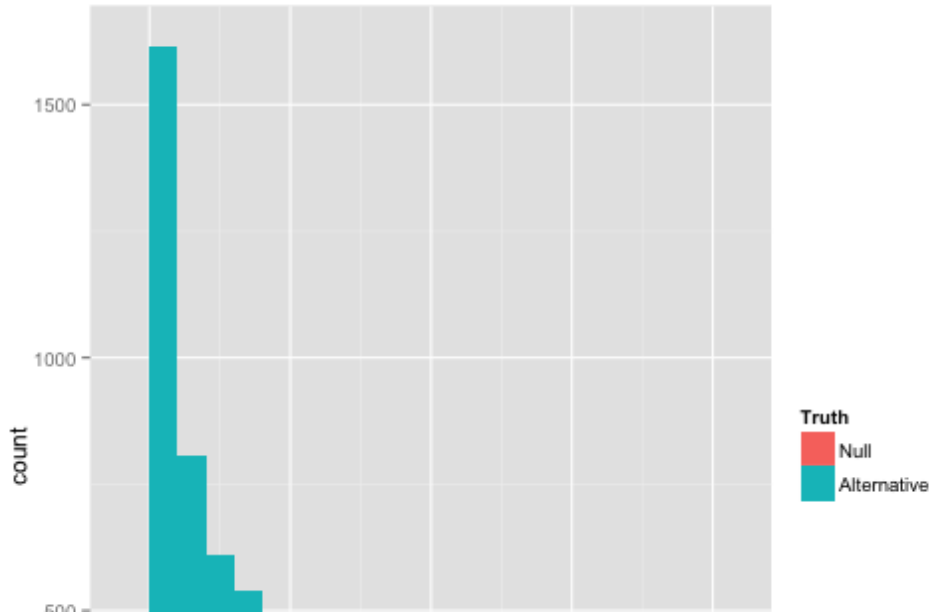
the expected proportion of false positives, among all the rejected hypothesis.

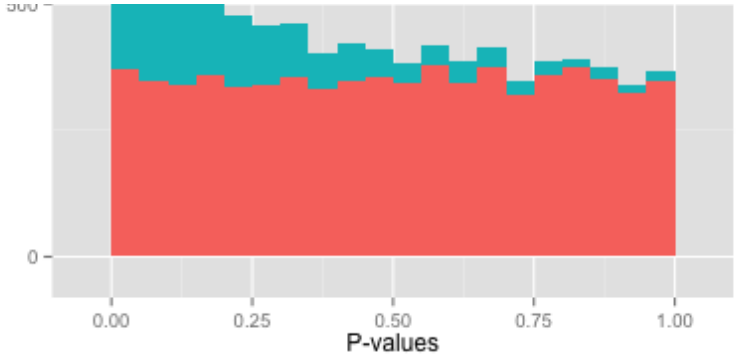
The number of false alarms is allowed to grow with the number of hypotheses tested.

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

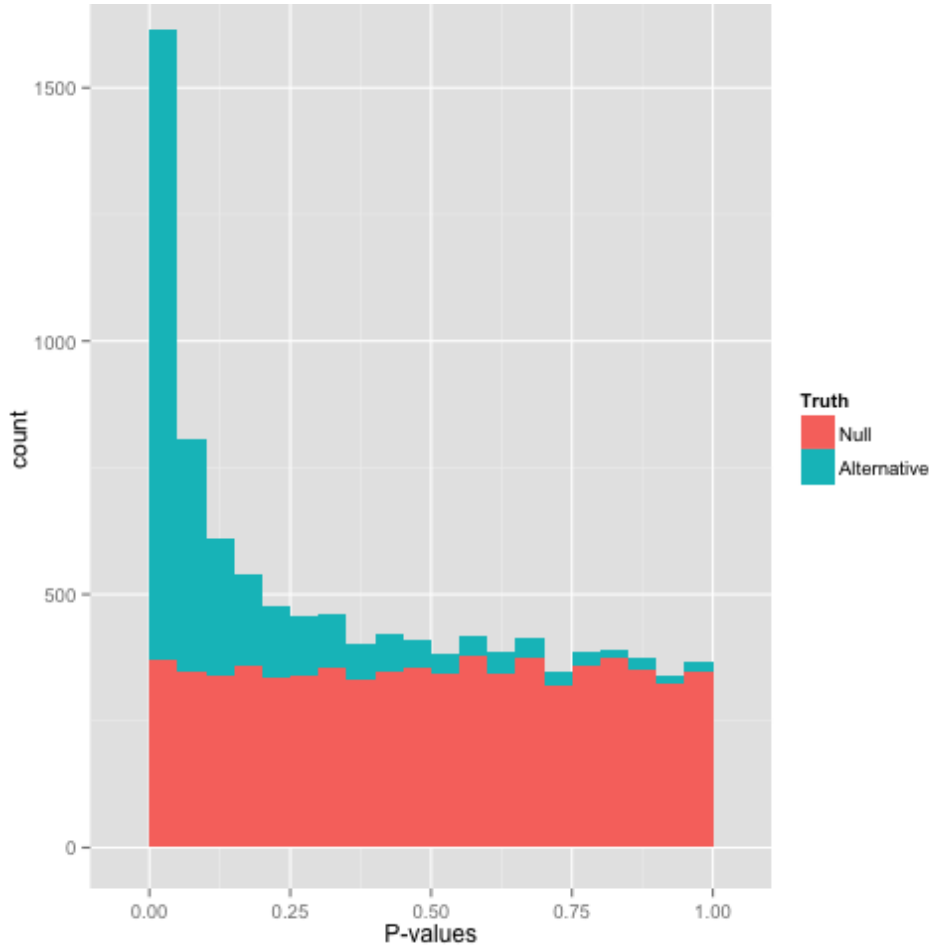




From this post

Digression: Uniformity under null

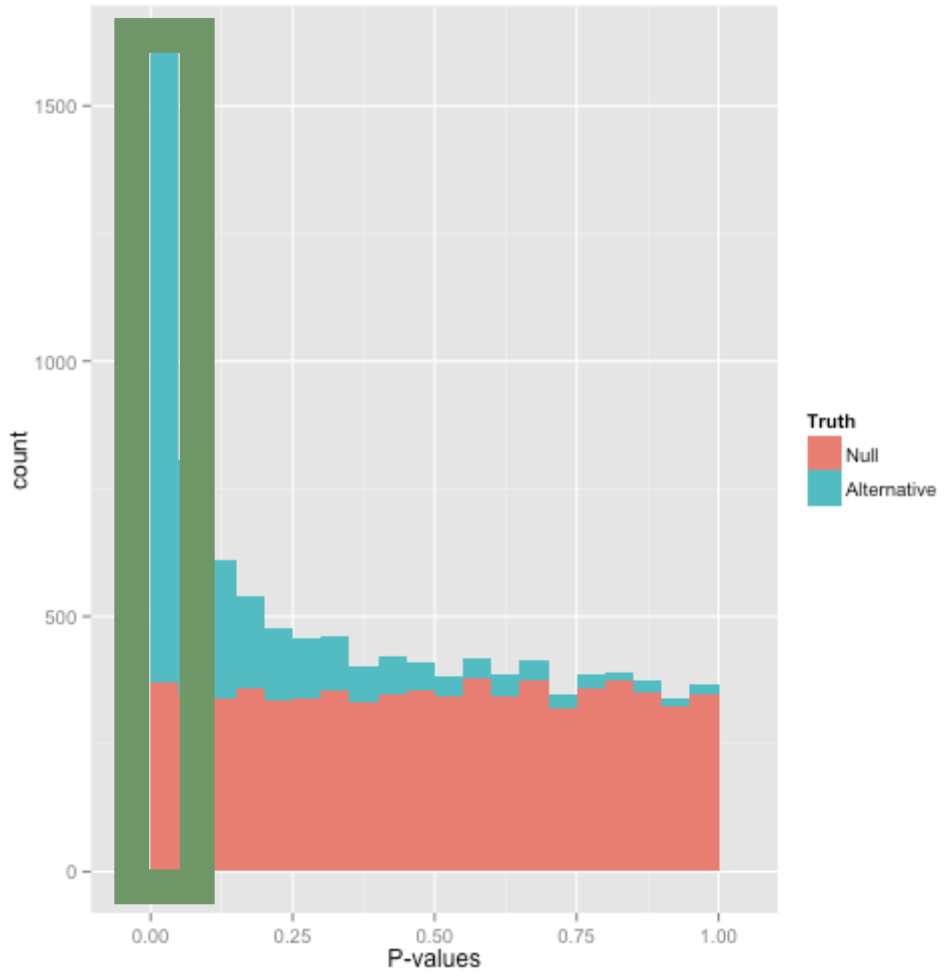
- Under the null, a p-value has 5% chance of being ≤ 0.05
  - It gives the probability of a false alarm under the null
- More generally, it has α chance of being smaller than α.
- Therefore, under the null, p-values are uniform



From this post

Estimating FDR

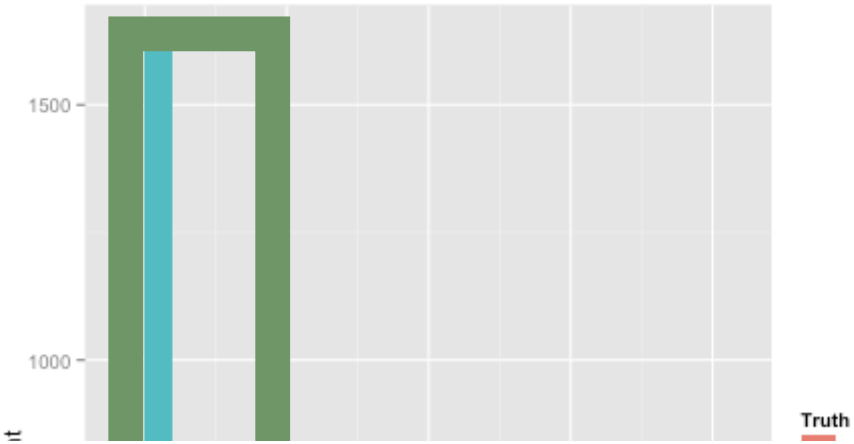
- You will reject all hypotheses to the left of the vertical bar
- The fraction of the rejection area that seems to come from the uniform part is an setimate of the FDR

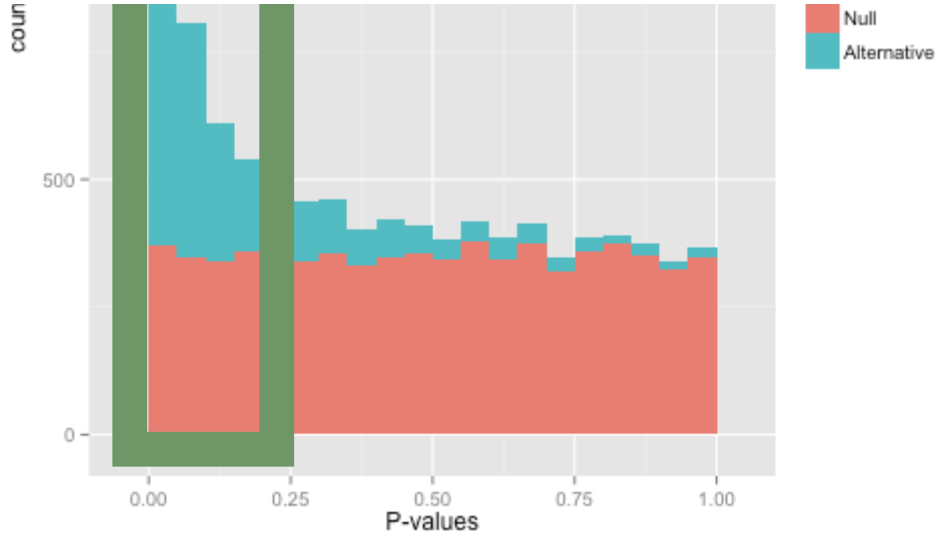


Most of the rejection (green) region is blue, so the FDR will be small.

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 setimate of the FDR

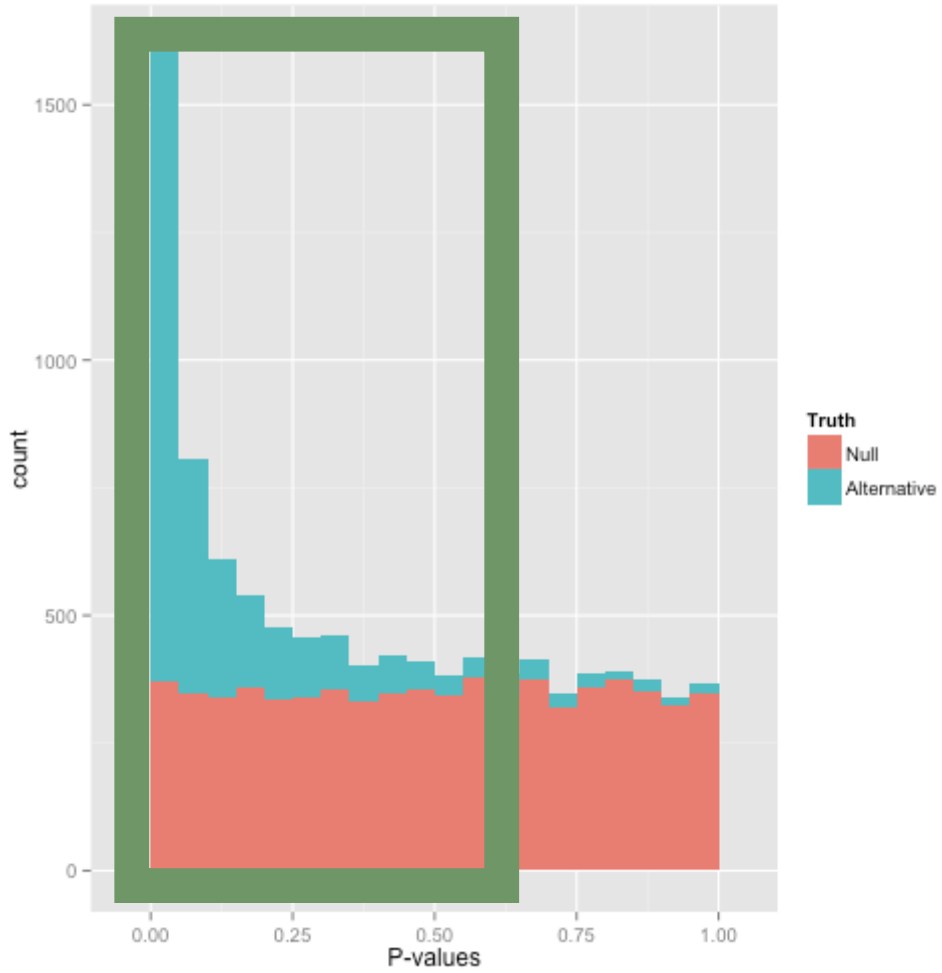




Now the FDR is larger.

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 setimate of the FDR



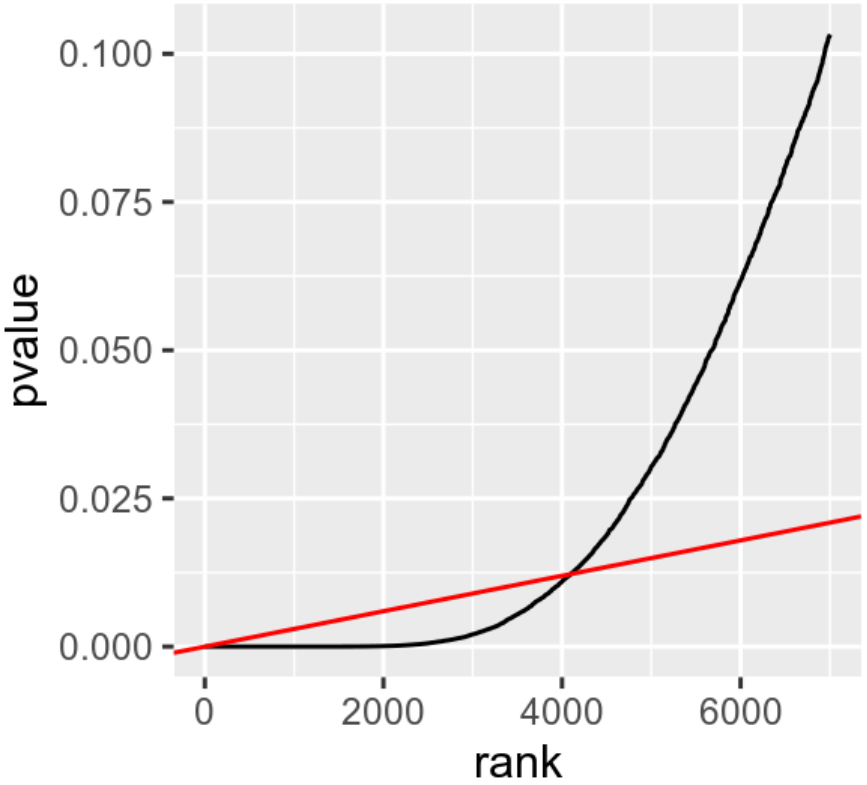
And now even larger.

Benjamini-Hochberg (BH) procedure

- The BH procedure 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, $p_{(1)}, \ldots, p_{(m)}$
- 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 is has both a null (uniform) and an alternative component, we can think of it as a mixture density,

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

Justification

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

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

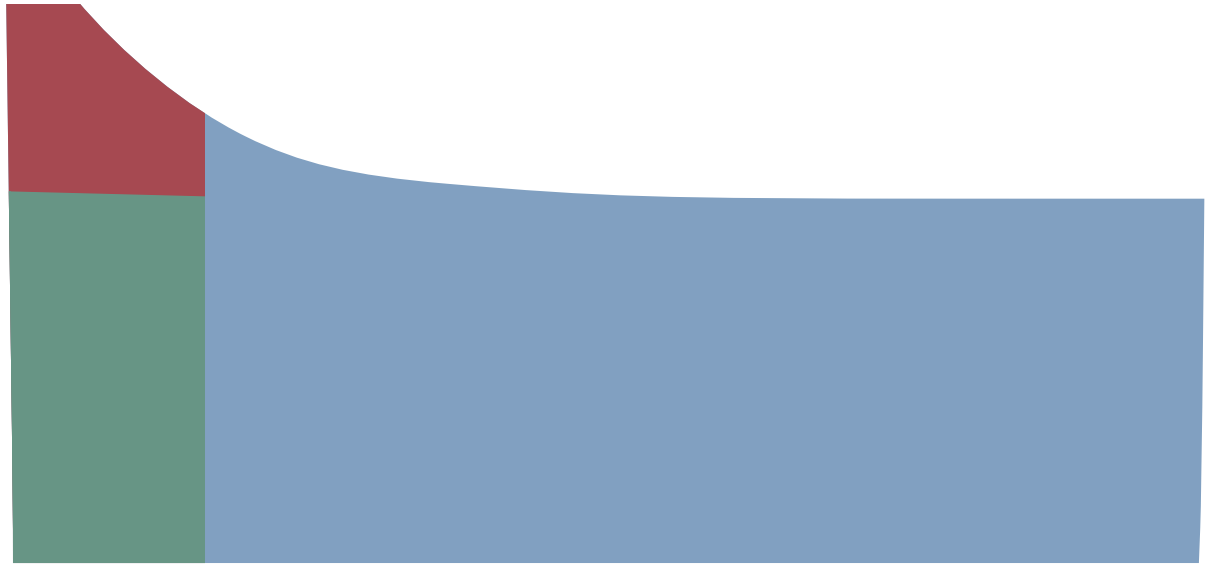
where $f\left(p\right)$ is the density associated with the mixture histogram.

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,

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

where $F\left(p\right)$ 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,

$$\widehat{FDR}(p) = \frac{\pi_0 p}{}$$

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

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

Justification

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

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

so that this estimate is smaller than α 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,

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

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 α.

The Multiple Testing ~~problem~~ opportunity

- Usually, people think of the 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

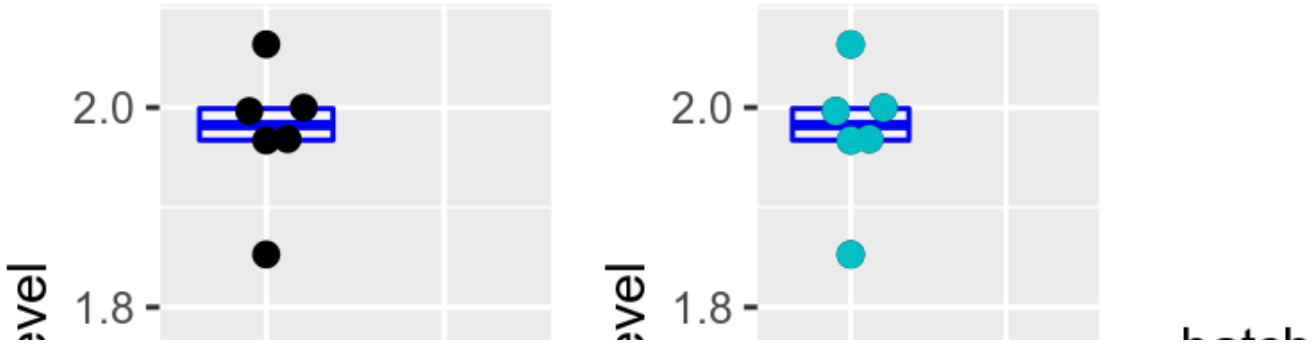
Basic Principles

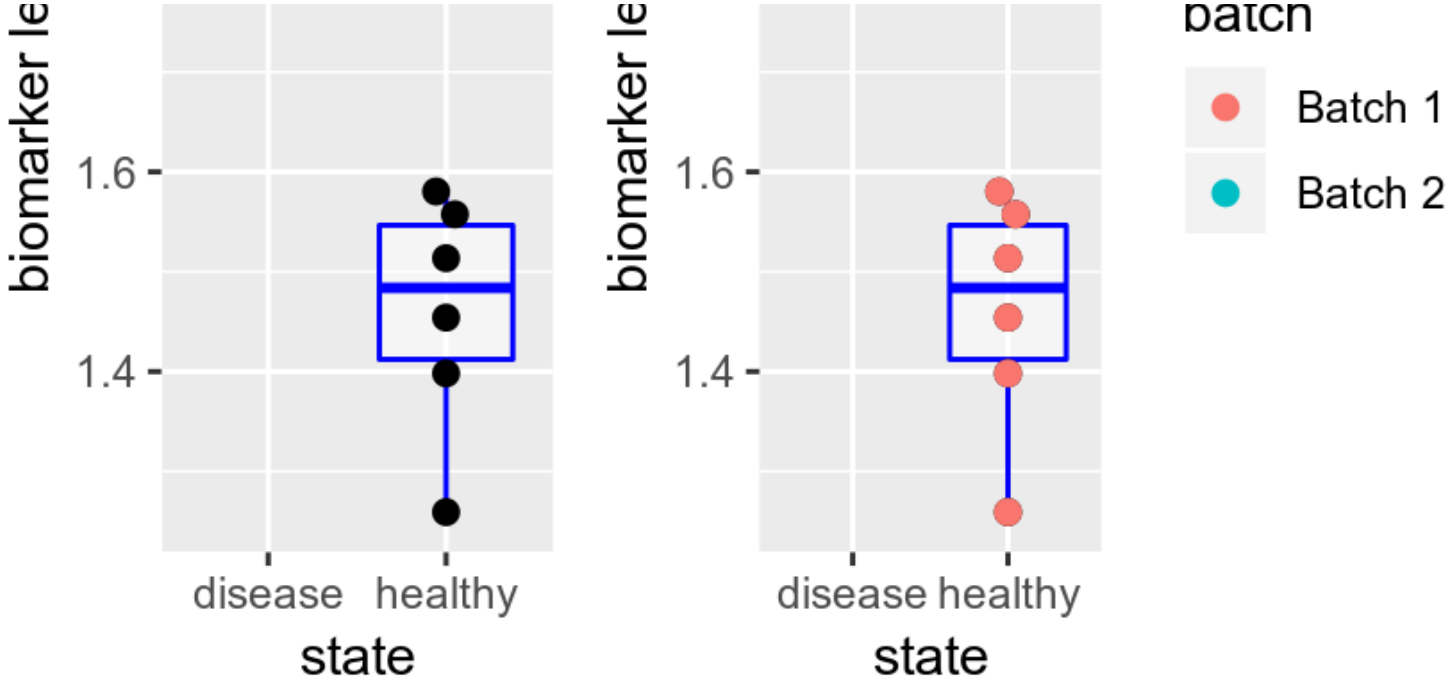
- Confounding
- Blocking and pairing

- Blocking and pairing
- Power analysis

Confounding

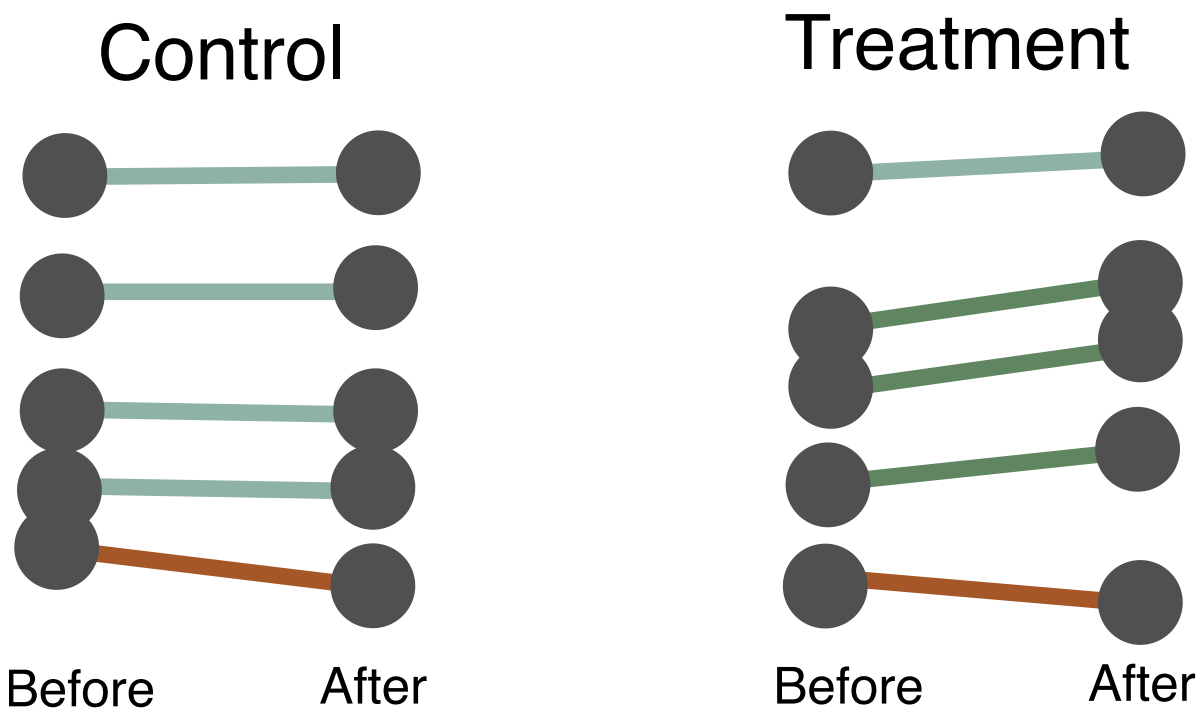
If there are factors that may influence both the treatment assignment and the observed outcome, then they can't be ruled out as potential explanations for the outcome.





Blocking and Pairing

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



Here, the effect of the treatment would get drowned out by inter-individual variation, unless you explicitly looked at the differences on a per-person level.

Blocking

Blocking

Story: 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?

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

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"

Randomized Control Trial

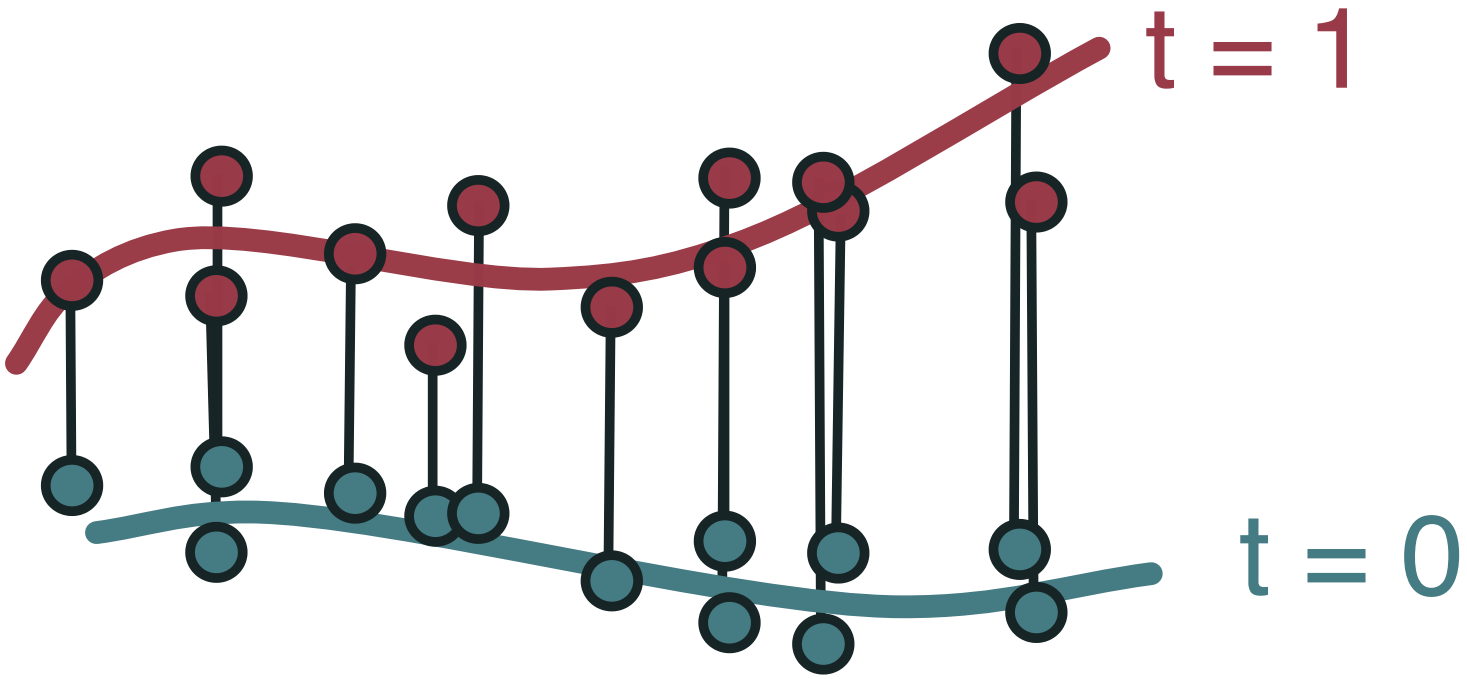
- You have no control over most factors, but you **can** randomize the assignment of the treatment
- By randomizing the assignment, other influences will cancel out, on average
- Here, correlation *does* imply causation (assuming that the model is correct...)

Types of Experiments (formal)

Formally, there is a process influencing the outcome,

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

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



X

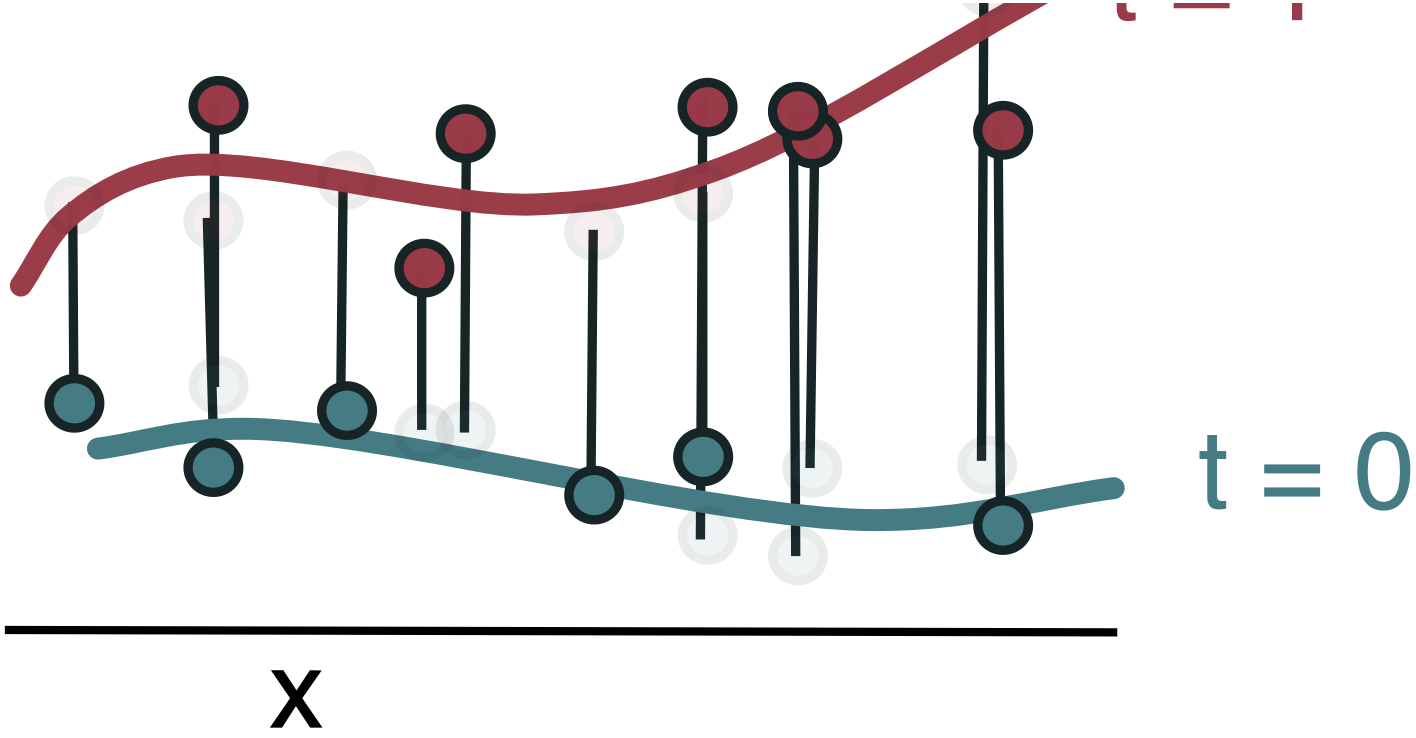
In theory, each sample could have two outcomes, depending on whether they are given the treatment or not.

Types of Experiments (formal)

Formally, there is a process influencing the outcome,

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

- t_i is the treatment,
- x_i are measured influencing factors
- ϵ_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,

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

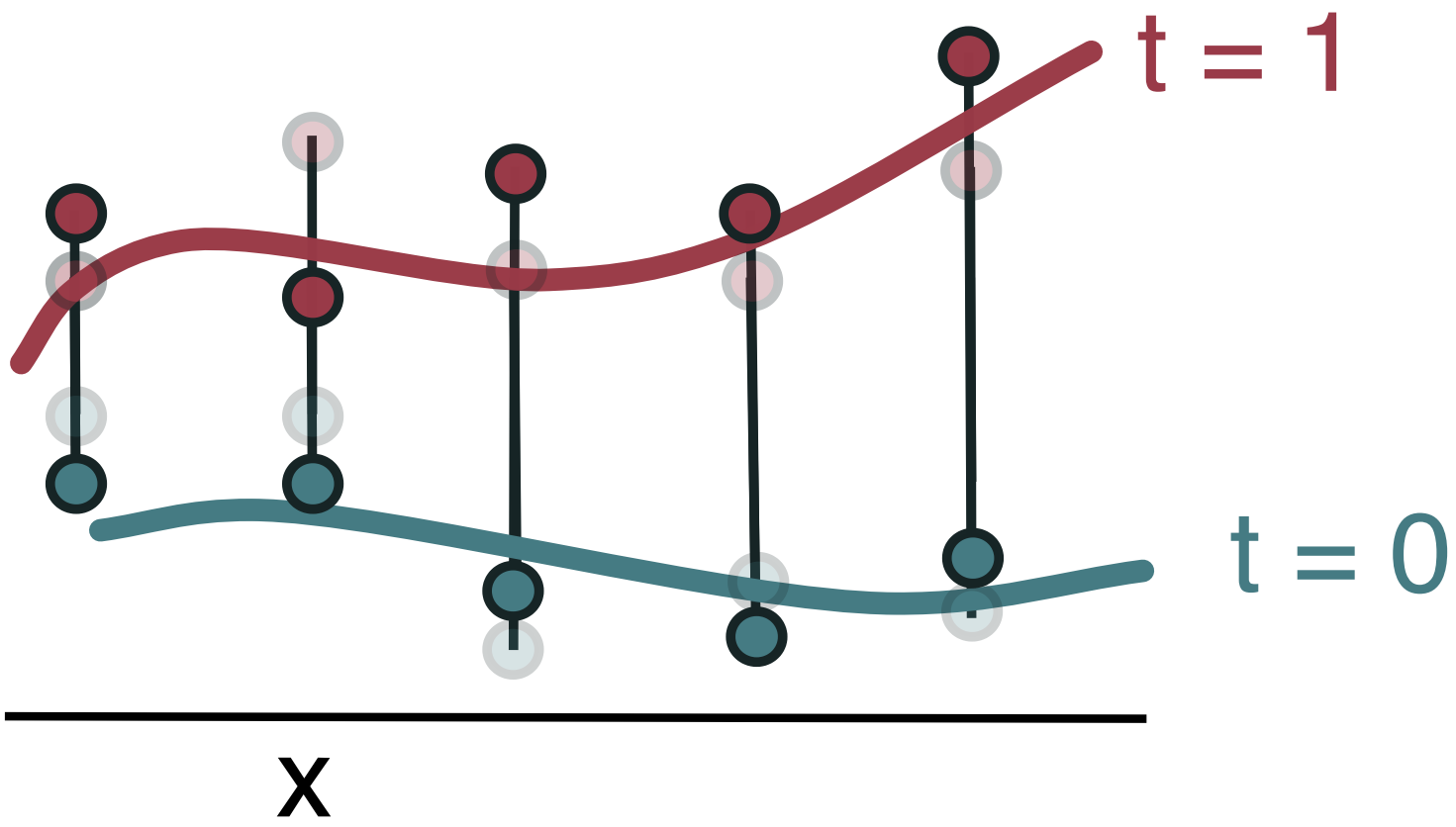
- Controlled experiment: Can manipulate all x_i, t_i, and $\epsilon_i \equiv 0$.
- Study: Able to manipulate some of the x_i and / or t_i
- Observational study: Cannot manipulate any of x_i or t_i* Randomized controlled study: Randomize over t_i, can't manipulate x_i

Controlled Experiment

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

Can manipulate all x_i, t_i, and $\epsilon_i = 0$. In the example here

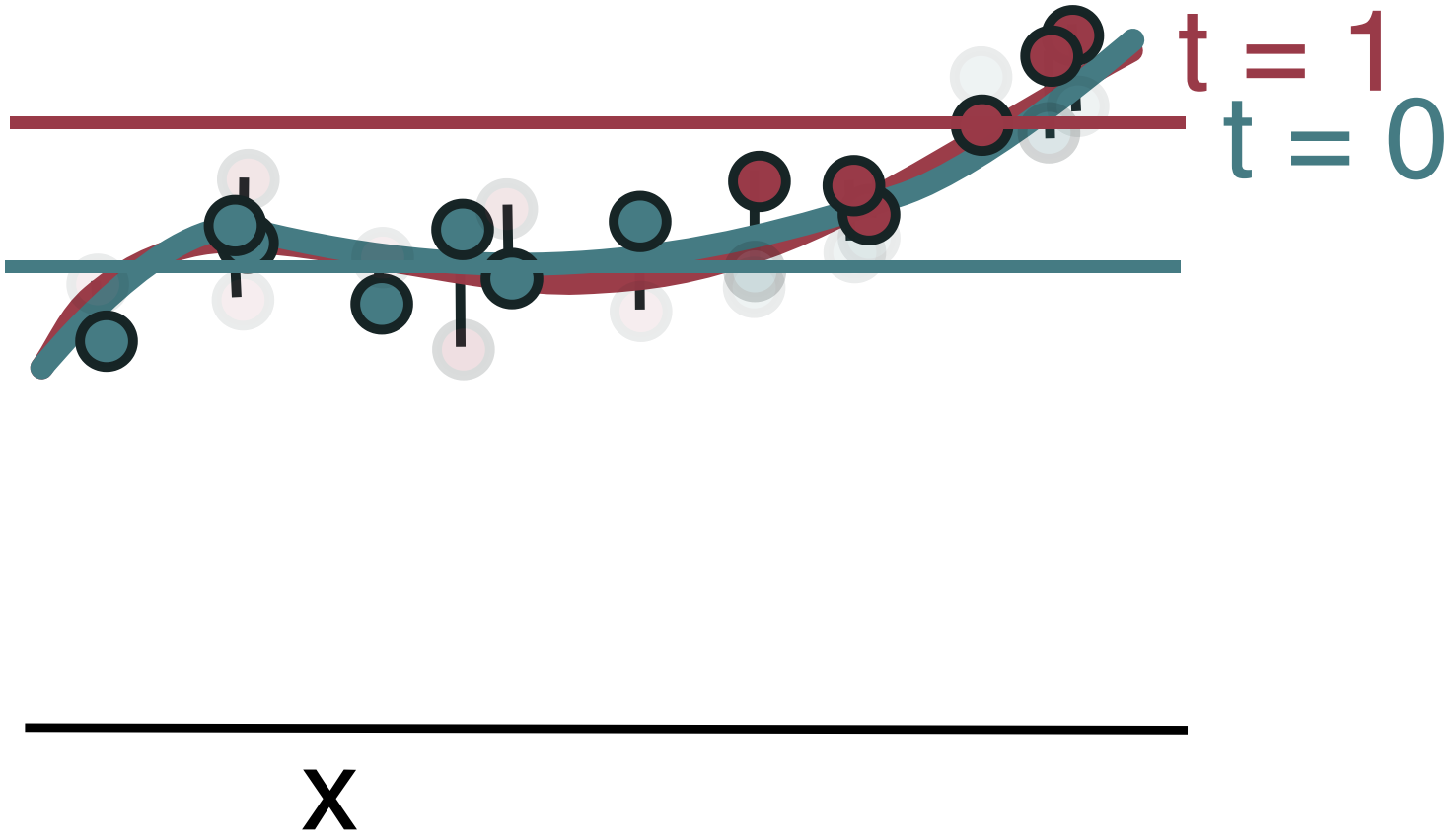
Can manipulate all x_i, t_i, and $\epsilon_i = 0$. In the example here, we get both a treatment and a control sample on an even grid of x's.



Observational Study

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

Have to guard against confounders.

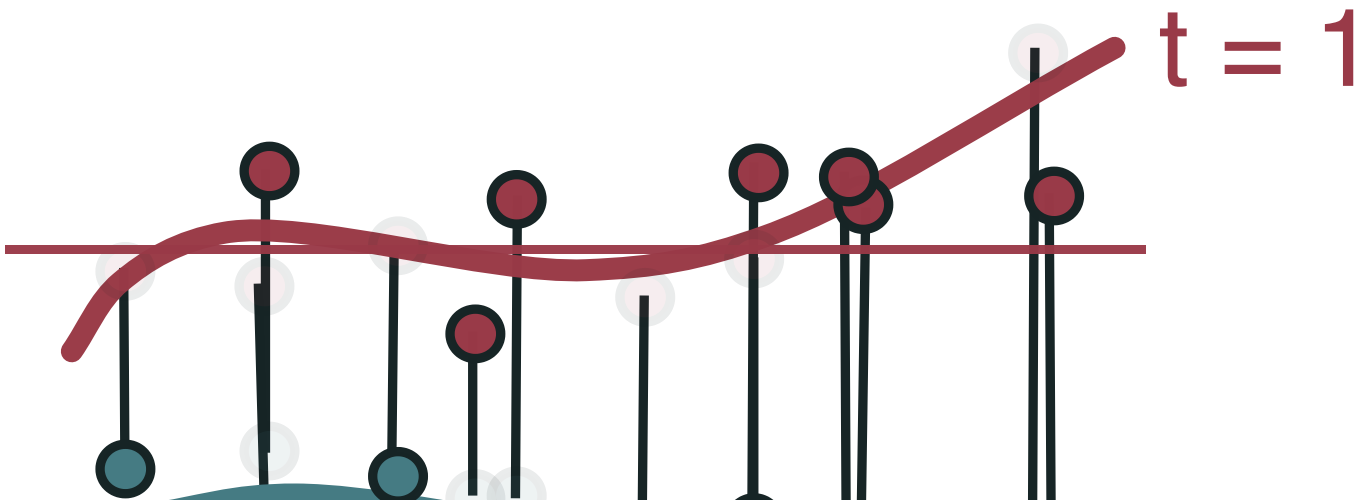


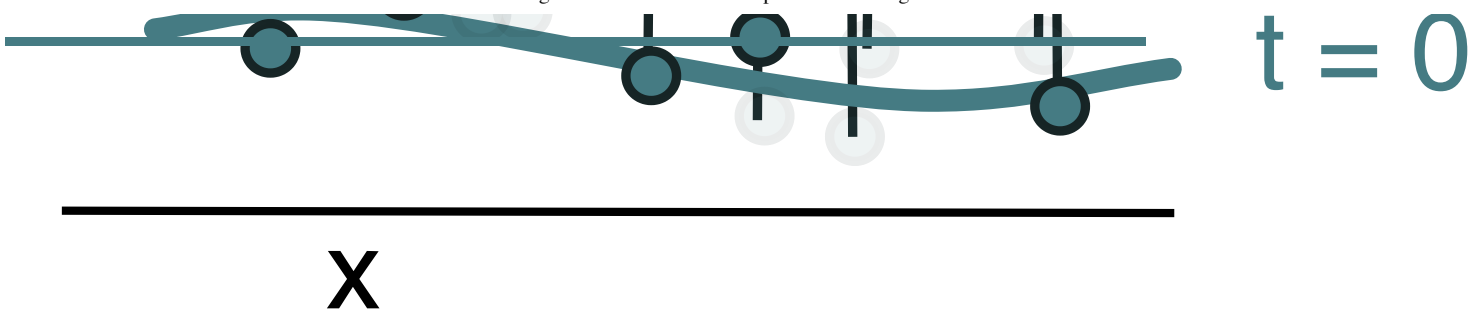
Randomized Controlled Trial

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

Randomize over t_i, can't manipulate x_i. When the treatment effect is constant over x, any effect caused by x_i gets averaged out, leaving only the overall treatment effect.

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





Data Quality

Data quality is a fuzzy 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

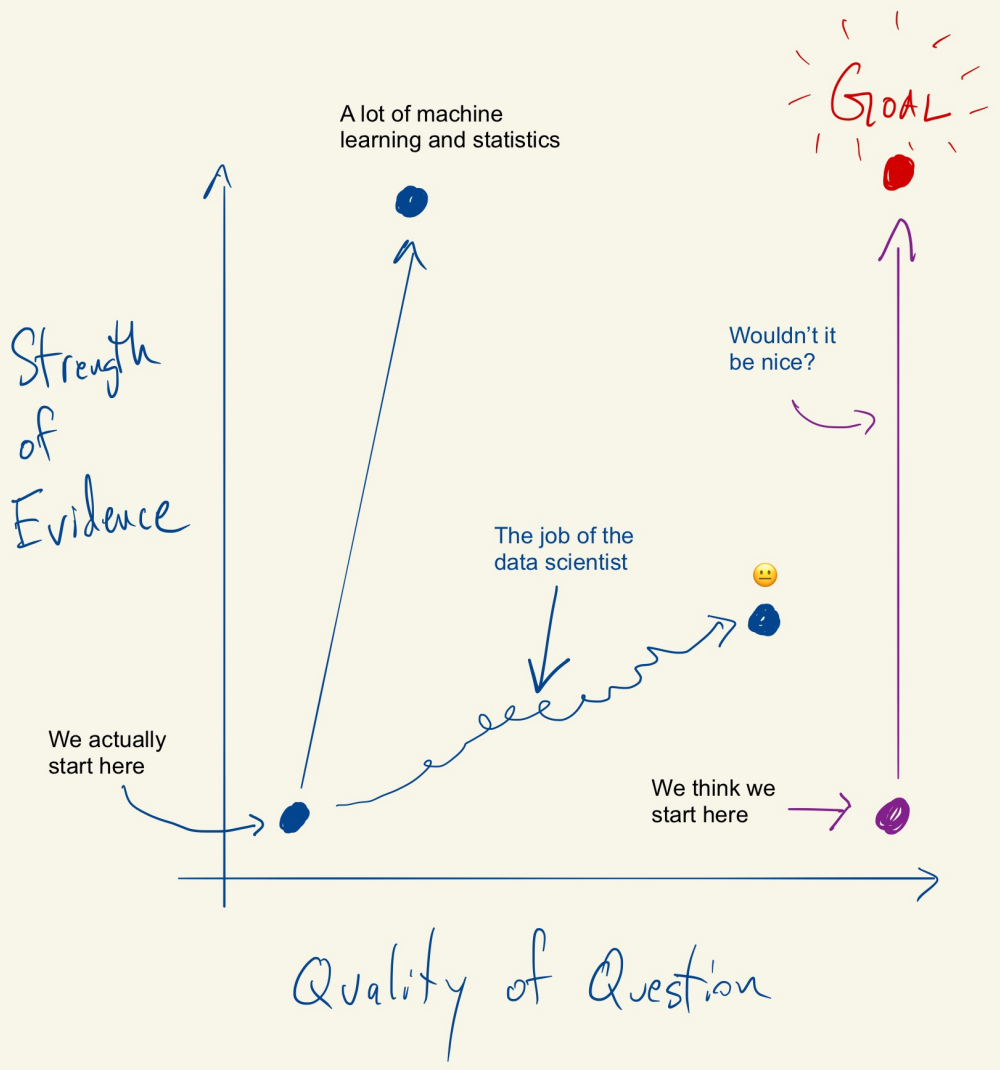


Figure from this post. Some other useful references are,

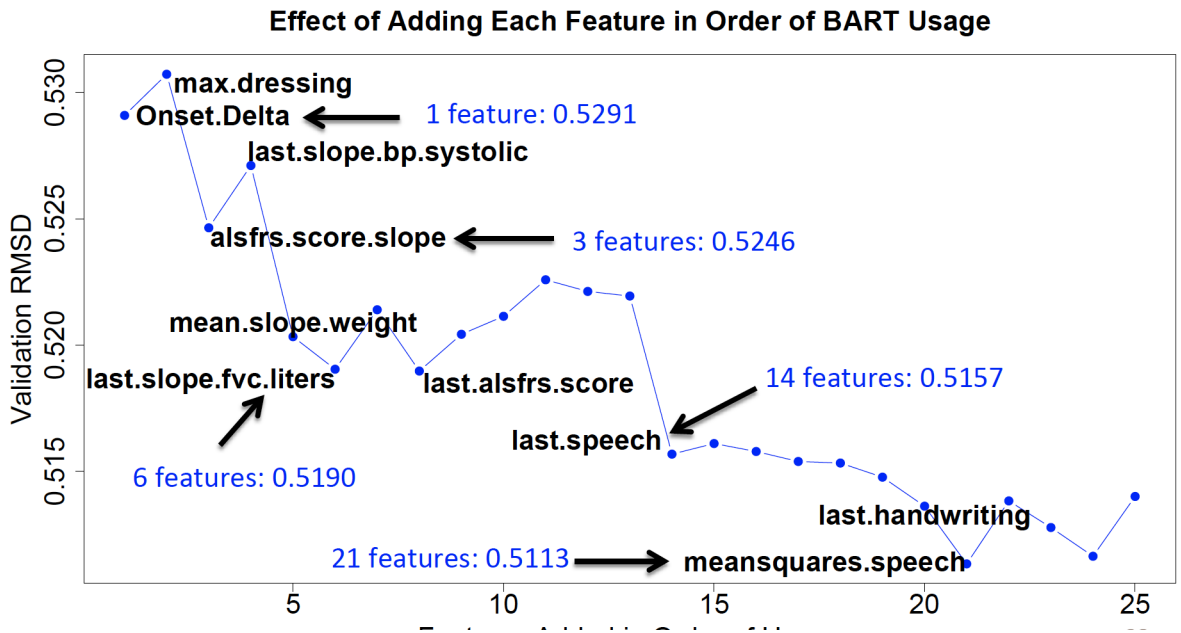
- How to share data with a statistician
- Design Study Methodology

Where to next?

- We've reviewed the roles of prediction, unsupervised learning, and statistical inference in data science
- We always assumed that our data are tabular
  - Samples drawn $x_i \overset{i.i.d.}{\sim} F$
  - 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?

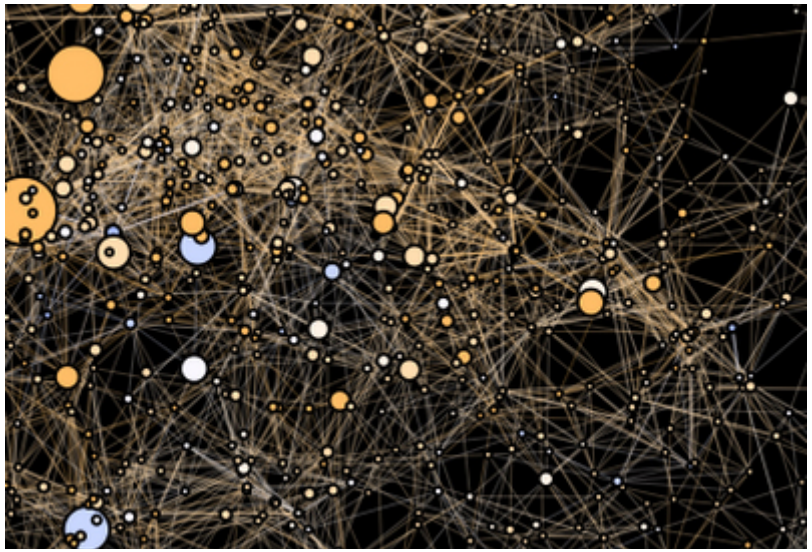


Features Added in Order of Usage 38

From the winning solution described here.

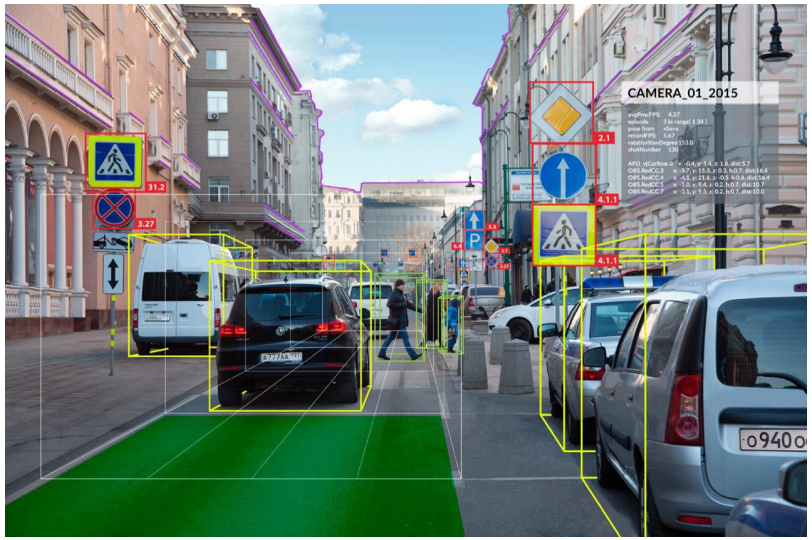
Where to next?

- What if your samples have relationships with one another?



Where to next?

- What if the correlation has very specific types of structure?



Where to next?

(to be continued...)

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

```
<style>

mtex_block = f()

mtex = f()
```