

# Experimental design

Jeffrey Leek

May 17, 2016

# Why you should care - an exciting result!

## Genomic signatures to guide the use of chemotherapeutics

Anil Potti<sup>1,2</sup>, Holly K Dressman<sup>1,3</sup>, Andrea Bild<sup>1,3</sup>, Richard F Riedel<sup>1,2</sup>, Gina Chan<sup>4</sup>, Robyn Sayer<sup>4</sup>, Janel Cragun<sup>4</sup>, Hope Cottrill<sup>4</sup>, Michael J Kelley<sup>2</sup>, Rebecca Petersen<sup>5</sup>, David Harpole<sup>5</sup>, Jeffrey Marks<sup>5</sup>, Andrew Berchuck<sup>1,6</sup>, Geoffrey S Ginsburg<sup>1,2</sup>, Phillip Febbo<sup>1,2,3</sup>, Johnathan Lancaster<sup>4</sup> & Joseph R Nevins<sup>1,2,3</sup>

**Using *in vitro* drug sensitivity data coupled with Affymetrix microarray data, we developed gene expression signatures that predict sensitivity to individual chemotherapeutic drugs. Each signature was validated with response data from an independent set of cell line studies. We further show that many of these signatures can accurately predict clinical response in individuals treated with these drugs. Notably, signatures developed to predict response to individual agents, when combined, could also predict response to multidrug regimens. Finally, we integrated the chemotherapy response signatures with signatures of oncogenic pathway deregulation to identify new therapeutic strategies that make use of all available drugs. The development of gene expression profiles that can predict response to**

### ARTICLE LINKS

- ▶ Supplementary info

### ARTICLE TOOLS

- ✉ Send to a friend
- ✉ Export citation
- ✉ Export references
- ✉ Rights and permissions
- ✉ Order commercial reprints

### SEARCH PUBMED FOR

- ▶ Anil Potti
- ▶ Holly K Dressman
- ▶ Andrea Bild
- ▶ Richard F Riedel
- ▶ Gina Chan
- ▶ Robyn Sayer

http:  
//www.nature.com/nm/journal/v12/n11/full/nm1491.html

# Why you should care - uh oh!

## DERIVING CHEMOSENSITIVITY FROM CELL LINES: FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

BY KEITH A. BAGGERLY\* AND KEVIN R. COOMBES<sup>†</sup>

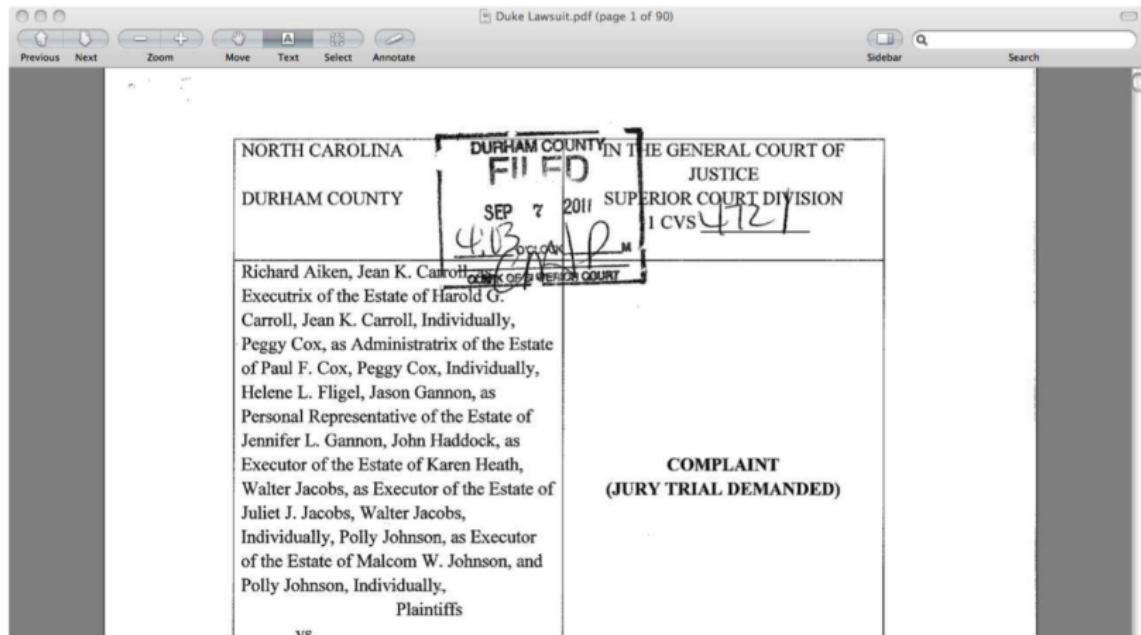
*U.T. M.D. Anderson Cancer Center*

High-throughput biological assays such as microarrays let us ask very detailed questions about how diseases operate, and promise to let us personalize therapy. Data processing, however, is often not described well enough to allow for exact reproduction of the results, leading to exercises in "forensic bioinformatics" where aspects of raw data and reported results are used to infer what methods must have been employed. Unfortunately, poor documentation can shift from an inconvenience to an active danger when it obscures not just methods but errors. In this report, we examine several related papers purporting to use microarray-based signatures of drug sensitivity derived from cell lines to predict patient response. Patients in clinical trials are currently being allocated to treatment arms on the basis of these results. However, we show in five case studies that the results incorporate several simple errors that may be putting patients at risk. One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common. We then discuss steps we are taking to avoid such errors in our own investigations.

Annals of Applied Statistics

<http://arxiv.org/pdf/1010.1092.pdf>

# Why you should care - serious trouble



# Know and care about the analysis plan!

## Abstract

Formula display:  **MathJax** [?](#)

## Background

Many groups, including our own, have proposed the use of DNA methylation profiles as biomarkers for various disease states. While much research has been done identifying DNA methylation signatures in cancer vs. normal etc., we still lack sufficient knowledge of the role that differential methylation plays during normal cellular differentiation and tissue specification. We also need thorough, genome level studies to determine the meaning of methylation of individual CpG dinucleotides in terms of gene expression.

## Results

In this study, we have used (insert statistical method here) to compile unique DNA methylation signatures from normal human heart, lung, and kidney using the Illumina Infinium 27 K methylation arrays and compared those to gene expression by RNA sequencing. We have identified unique signatures of global DNA methylation for human heart, kidney and liver, and showed that DNA methylation data can be used to correctly classify various tissues. It indicates that DNA methylation reflects tissue specificity and may play an important role in tissue differentiation. The integrative analysis of methylation and RNA-Seq data showed that gene methylation and its transcriptional levels were comprehensively correlated. The location of methylation markers in terms of distance to transcription start site and CpG island showed no effects on the regulation of gene expression by DNA methylation in normal tissues.

[http://nsaunders.wordpress.com/2012/07/23/  
we-really-dont-care-what-statistical-method-you-used/](http://nsaunders.wordpress.com/2012/07/23/we-really-dont-care-what-statistical-method-you-used/)

# Have a plan for data and code sharing

The screenshot shows the GitHub homepage with a user profile for 'jtleek'. The main navigation bar includes 'Explore', 'Gist', 'Blog', and 'Help'. Below the navigation is a 'News Feed' section with tabs for 'News Feed', 'Pull Requests', 'Issues', and 'Stars'. A prominent feature is the 'GitHub Bootcamp' section, which provides four numbered steps for new users:

- 1 Set up Git**: A quick guide to help you get started with Git.
- 2 Create repositories**: Repositories are where you'll work and collaborate on projects.
- 3 Fork repositories**: Forking creates a new, unique project from an existing one.
- 4 Be social**: Send pull requests, follow friends. Star and watch projects.

<https://github.com/>

The screenshot shows the figshare homepage. The top navigation bar includes 'Browse' and 'Upload' buttons, along with 'Sign up' and 'Login' links. The main header features the figshare logo and a search bar. A large green arrow points upwards towards the 'Sign up for free' button, which is highlighted in red.

Sign up for free

# May I recommend?

The Leek group guide to data sharing — Edit

A screenshot of a GitHub repository page for 'datasharing'. The repository has 25 commits, 1 branch, 0 releases, and 8 contributors. The branch dropdown shows 'branch: master'. The repository was last updated 6 days ago by jtleek, with a commit titled 'fix typo'.

Merge pull request #9 from nikal3d/patch-1 ...

**jtleek** authored 6 days ago latest commit e53857faa4

**README.md** fix typo 6 days ago

**README.md**

## How to share data with a statistician

This is a guide for anyone who needs to share data with a statistician. The target audiences I have in mind are:

- Scientific collaborators who need statisticians to analyze data for them
- Students or postdocs in scientific disciplines looking for consulting advice
- Junior statistics students whose job it is to collate/clean data sets

<https://github.com/jtleek/datasharing>

# Formulate your question in advance

The screenshot shows a WIRED website page. At the top, there's a banner for a Vegas Uncork'd event. Below it, a navigation bar includes links for GEAR, SCIENCE, ENTERTAINMENT, BUSINESS, SECURITY, DESIGN, OPINION, VIDEO, INSIDER, MAGAZINE, and SUBSCRIBE. A search bar is also present.

The main headline is "The A/B Test: Inside the Technology That's Changing the Rules of Business". Below the headline, it says "BY BRIAN CHRISTIAN 04.25.12 8:47 PM". To the right of the headline are social sharing buttons for Share (542), Tweet (334), StumbleUpon (795), LinkedIn (261), and Print (261).

The central image is split into two halves: the left half shows a large, 3D-printed letter 'A' made of grey blocks, and the right half shows a stylized, ornate letter 'B' in black on an orange background. Below the image, the caption reads "Photo: Spencer Higgins; Illustration: Si Scott".

To the right of the main content, there's a sidebar with a section titled "Feature development with Git" featuring an Atlassian logo. It includes a small diagram of two circles connected by arrows. Below this, there's a "Register now" button and the Atlassian logo again.

Further down the page, there's a section titled "MOST RECENT WIRED POSTS" with two visible thumbnails:

- "WIRED Space Photo of the Day: Saturn's Maedstrom" (Thumbnail shows a swirling space image)
- "This Week in Photography: A New Paradise, Cities Under Siege, and a Fake Polar Vortex" (Thumbnail shows a bridge over water)

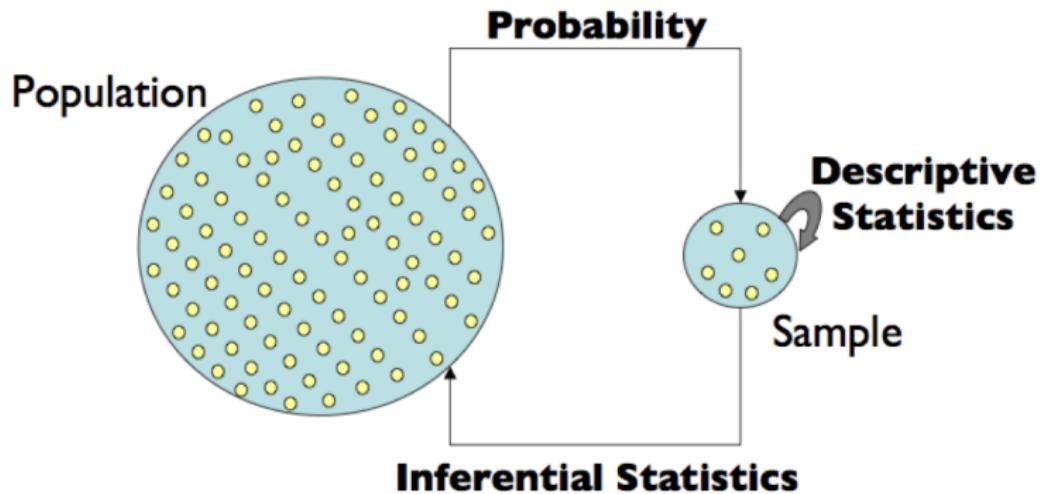
At the bottom of the page, there's a "One Year Later" section with a thumbnail showing a colorful bar chart.

**Question:** Does changing the text on your website improve donations?

**Experiment:**

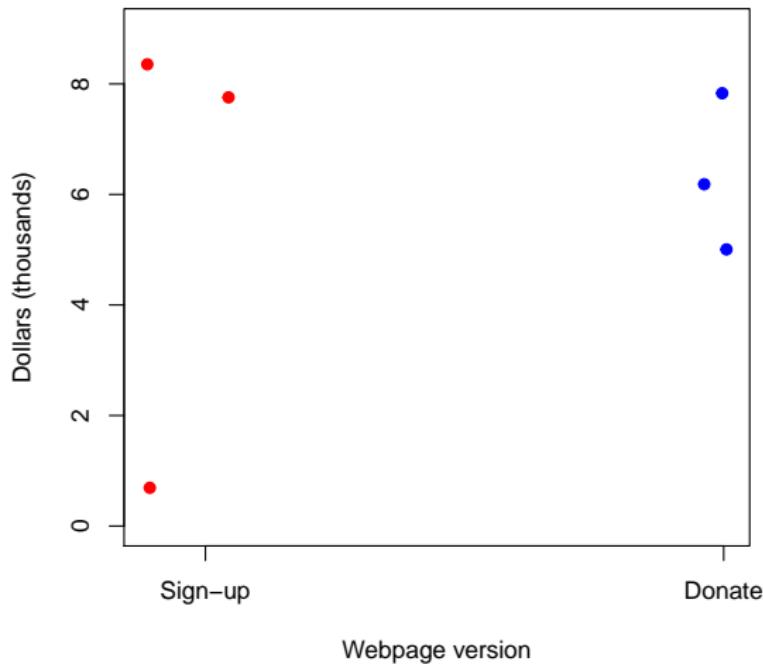
- 1 Randomly show visitors one version or the other

# Statistical inference

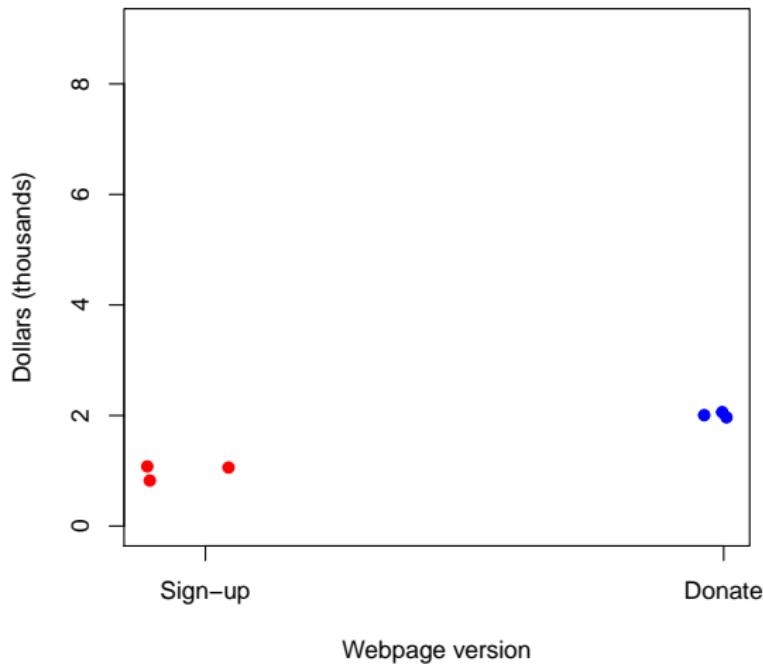


<http://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture2.pdf>

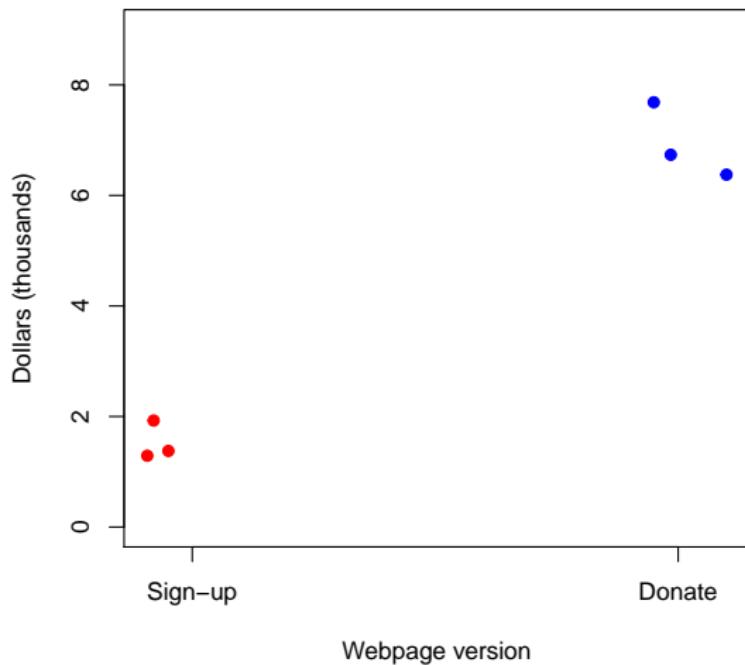
## Variability - Scenario 1



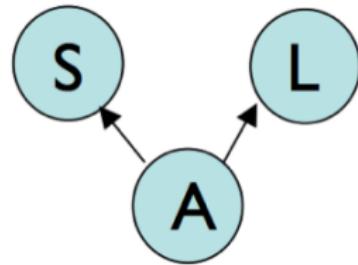
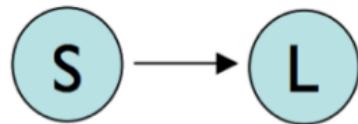
## Variability - Scenario 2



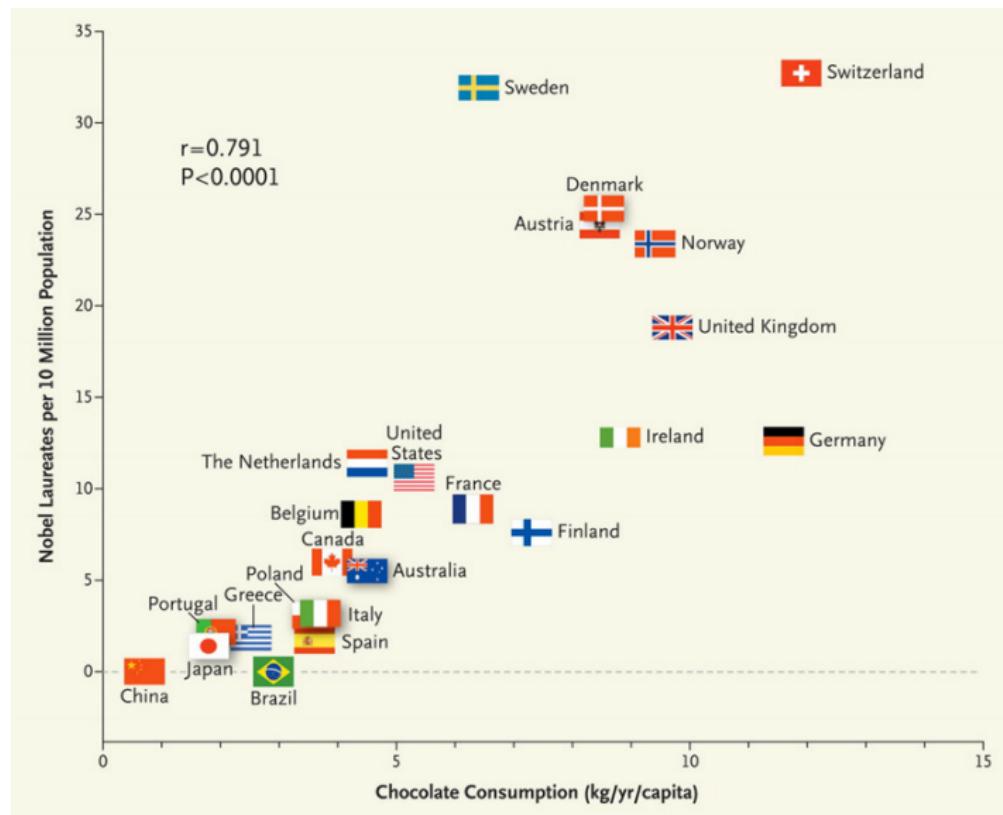
## Variability - Scenario 3



# Confounding



# Correlation is not causation\*



<http://www.nejm.org/doi/full/10.1056/NEJMoa1211064>

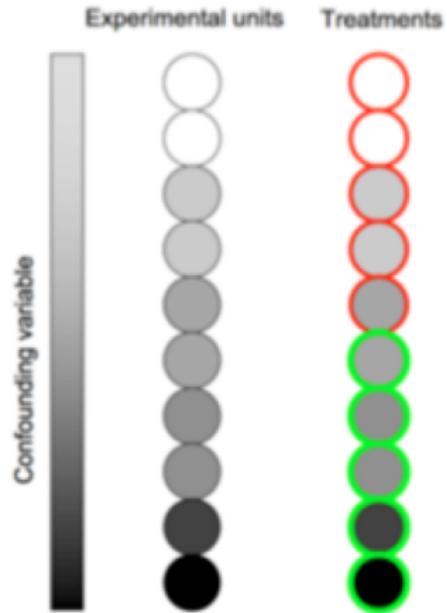
Sometimes called spurious correlation\*

## Randomization and blocking

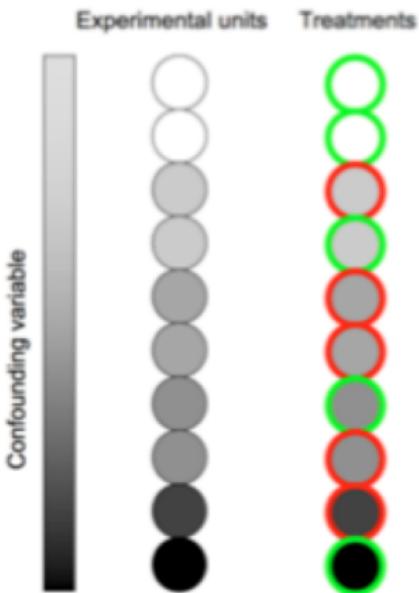
- ▶ If you can (and want to) fix a variable
- ▶ Website always says Obama 2014 on it
- ▶ If you don't fix a variable, stratify it
- ▶ If you are testing sign up phrases and have two website colors, use both phrases equally on both.
- ▶ If you can't fix a variable, randomize it

# Why does randomization help?

## Not Randomized

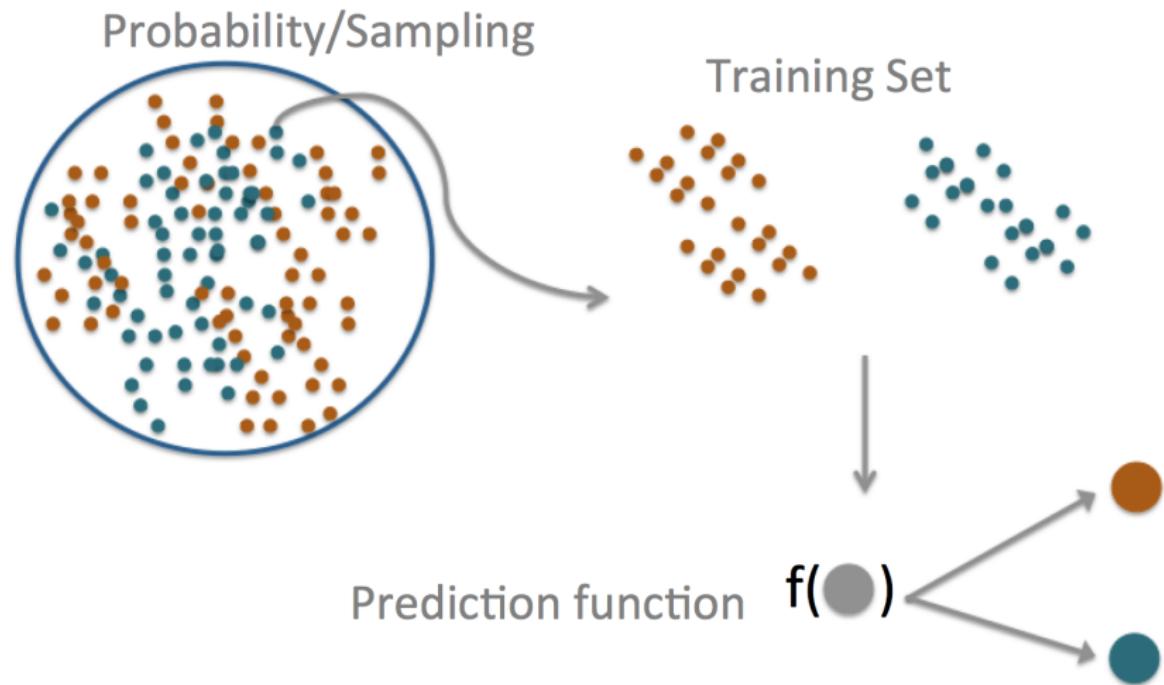


## Randomized

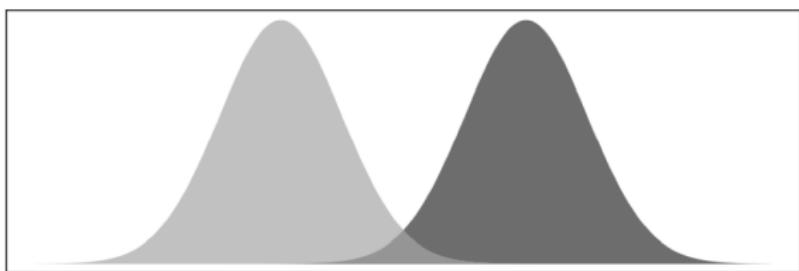
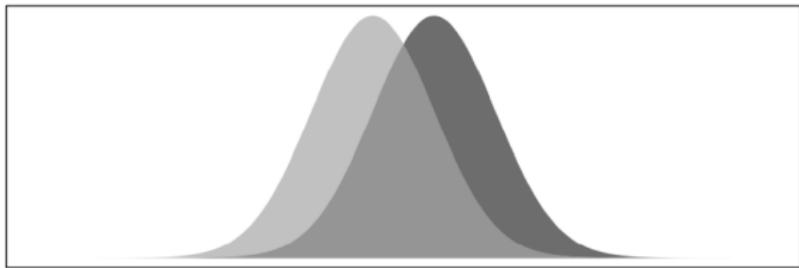


<http://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture1.pdf>

# Prediction



## Prediction versus inference



<http://www.biostat.jhsph.edu/~iruczins/teaching/140.615/>

# Prediction key quantities

		DISEASE	
		+	-
TEST	+	TP	FP
	-	FN	TN

Sensitivity

→  $\Pr(\text{positive test} \mid \text{disease})$

Specificity

→  $\Pr(\text{negative test} \mid \text{no disease})$

Positive Predictive Value

→  $\Pr(\text{disease} \mid \text{positive test})$

Negative Predictive Value

→  $\Pr(\text{no disease} \mid \text{negative test})$

Accuracy

→  $\Pr(\text{correct outcome})$

http:

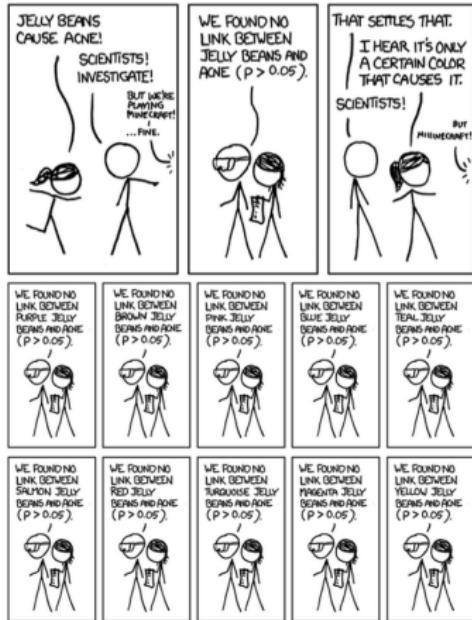
//www.biostat.jhsph.edu/~iruczins/teaching/140.615/

## Beware data dredging



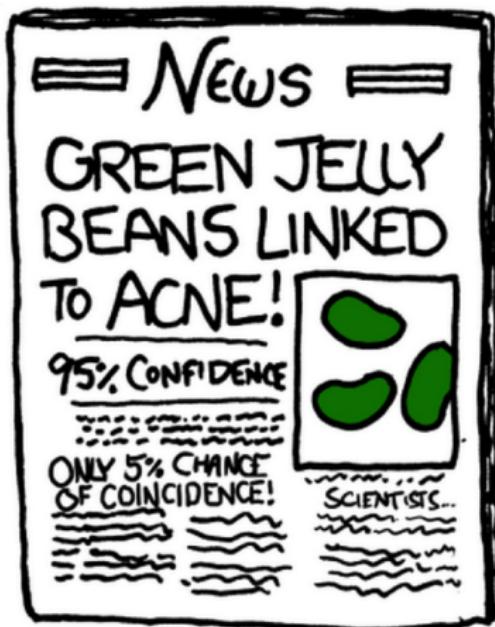
<http://xkcd.com/882/>

# Beware data dredging



<http://xkcd.com/882/>

## Beware data dredging



<http://xkcd.com/882/>

# Summary

- ▶ Good experiments
- ▶ Have replication
- ▶ Measure variability
- ▶ Generalize to the problem you care about
- ▶ Are transparent
- ▶ Prediction is not inference
- ▶ Both can be important
- ▶ Beware data dredging