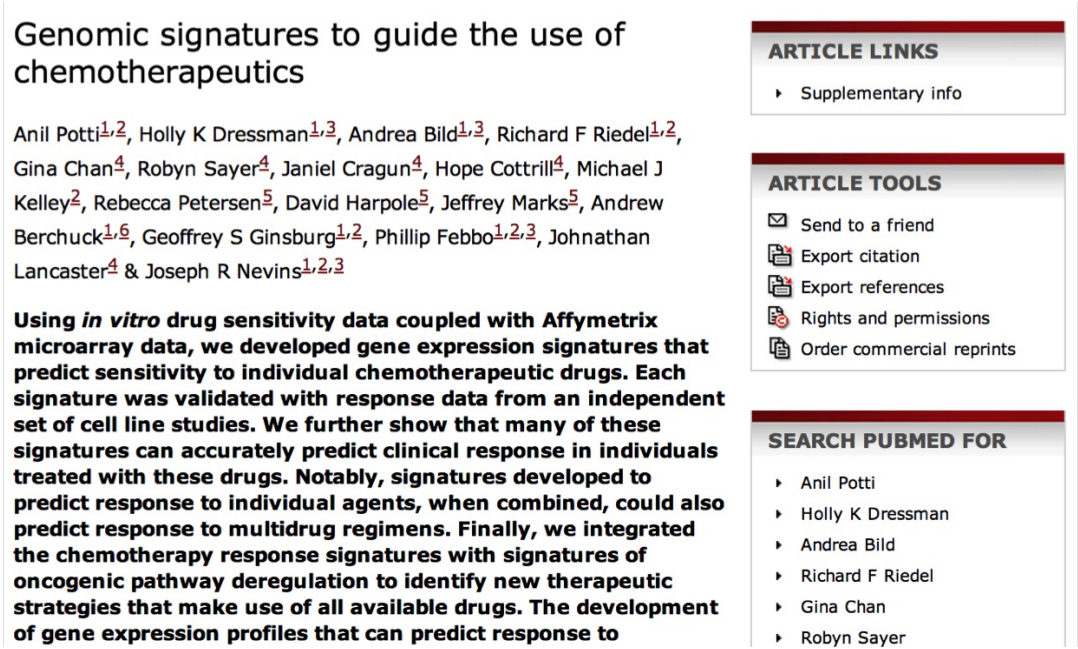
Experimental design

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

Why you should care - an exciting result!



Genomic signatures to guide the use of chemotherapeutics

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

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

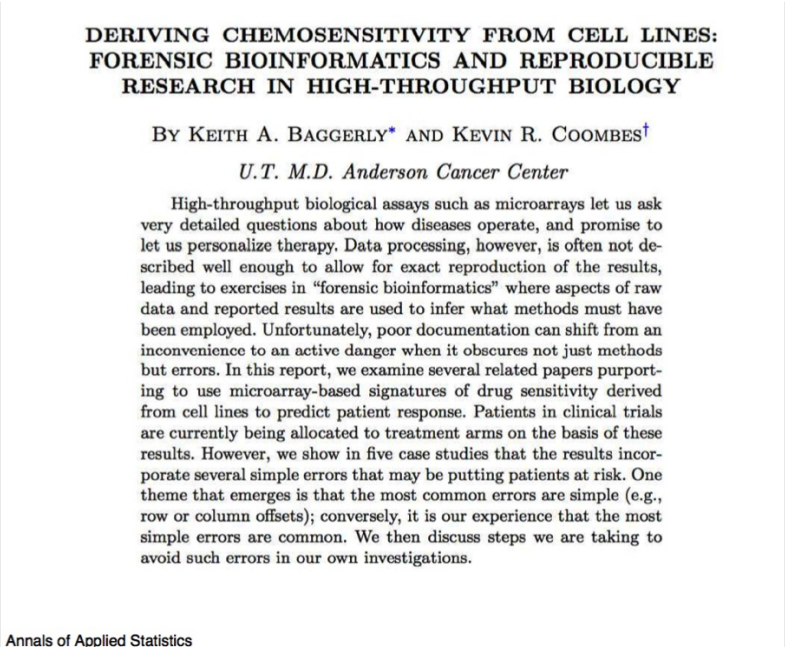
- Send to a friend
- Export citation
- Export references
- Rights and permissions
- Order commercial reprints

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†

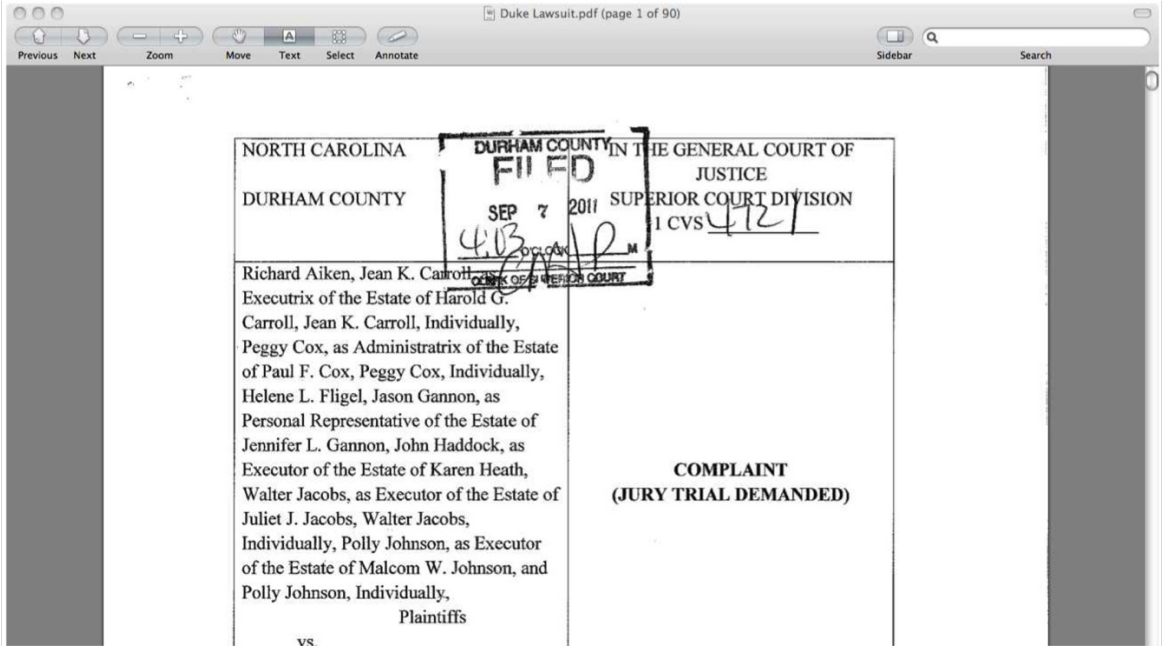
U.T. M.D. Anderson Cancer Center

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.

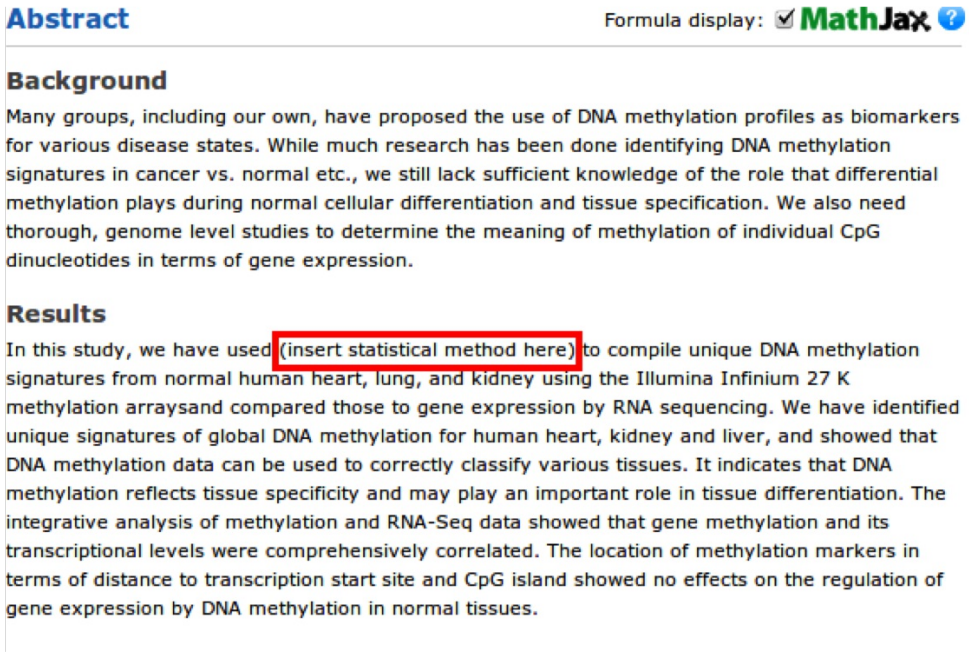
Annals of Applied Statistics

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

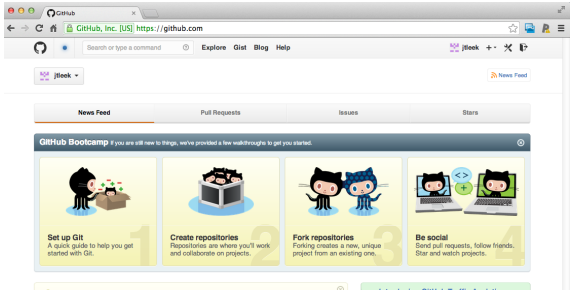
Many groups, including our own, have proposed the use of DNA methylation profiles as biomarkers for various disease states. While much research has been done identifying DNA methylation signatures in cancer vs. normal etc., we still lack sufficient knowledge of the role that differential methylation 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

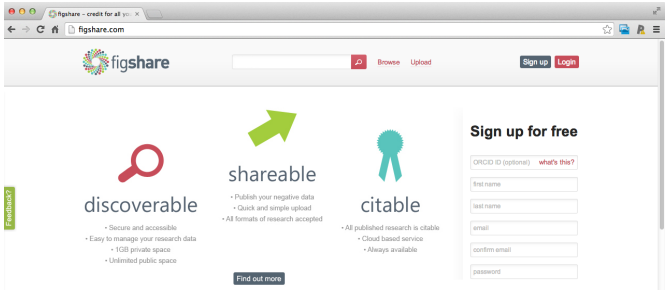
In this study, we have used (insert statistical method here) to compile unique DNA methylation signatures from normal human heart, lung, and kidney using the Illumina Infinium 27 K methylation arraysand compared those to gene expression by RNA sequencing. We have identified unique signatures of global 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/

Have a plan for data and code sharing

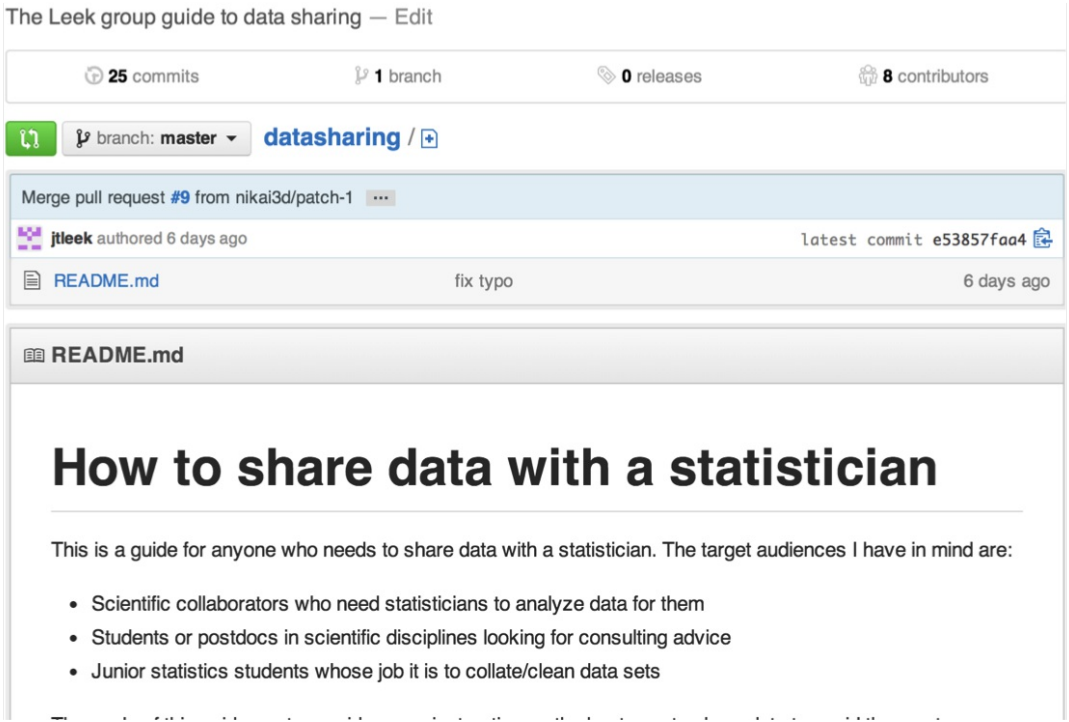


https://github.com/



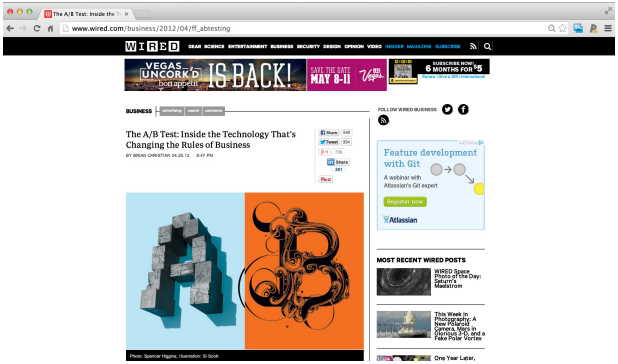
http://figshare.com/

May I recommend?



https://github.com/jtleek/datasharing

Formulate your question in advance



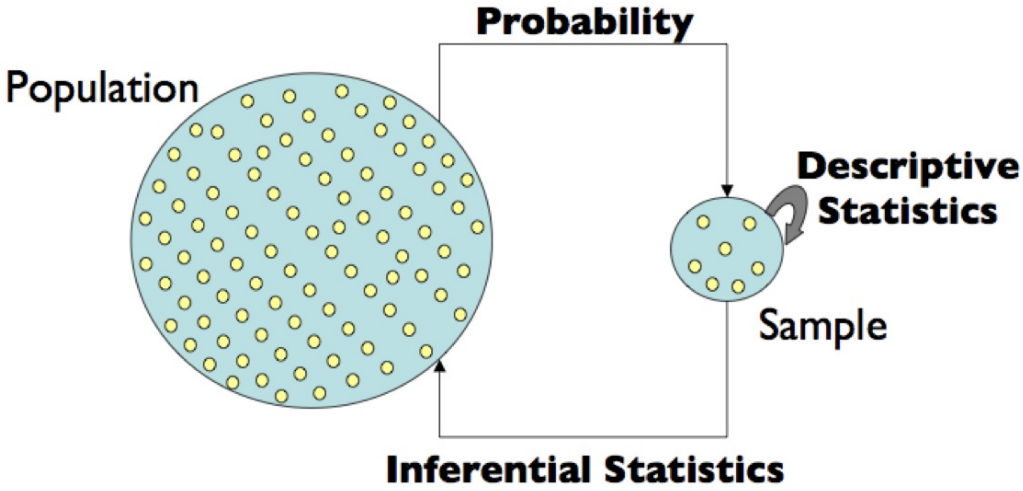
Question: Does changing the text on your website improve donations?

Experiment:

1. Randomly show visitors one version or the other
2. Measure how much they donate
3. Determine which is better

http://www.wired.com/business/2012/04/ff_abtesting

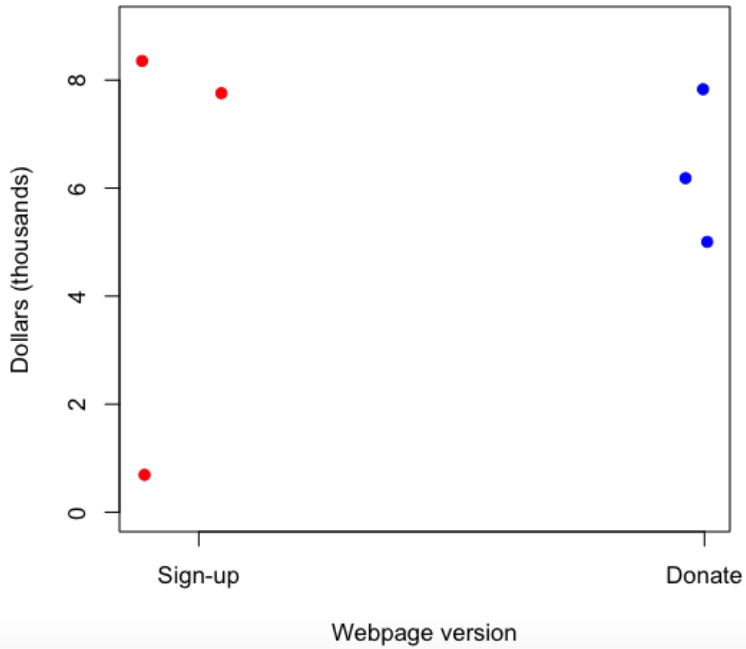
Statistical inference



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

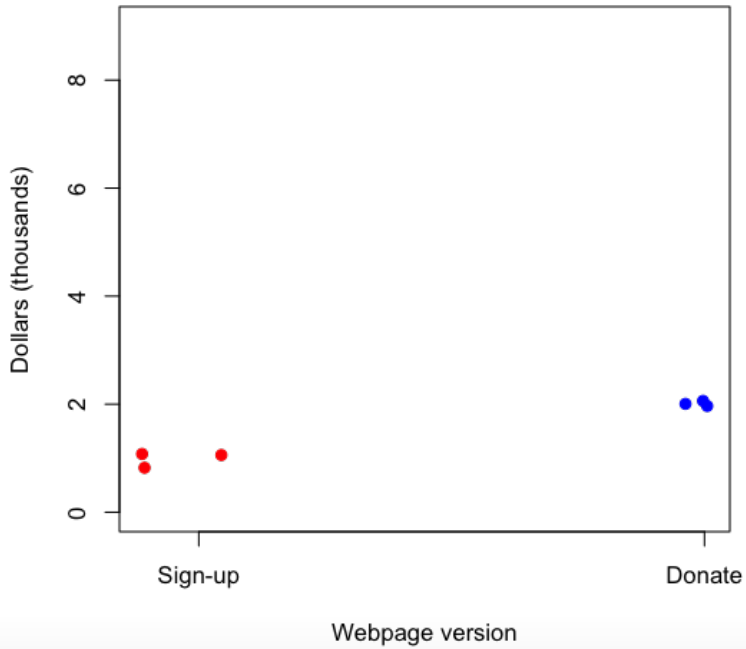
Variability - Scenario 1

兩個版本對捐款的差異不大



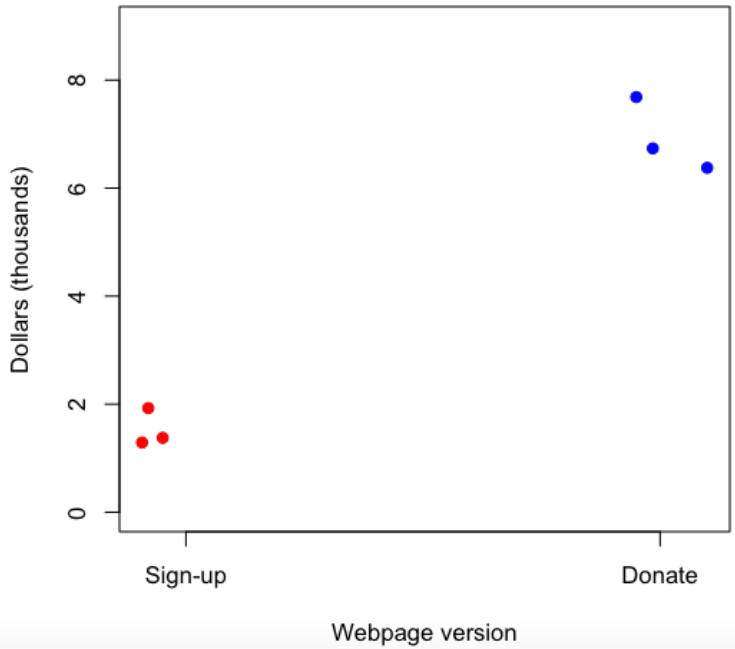
Variability - Scenario 2

兩個版本對捐款的差異不大（都不捐，上面是都捐）

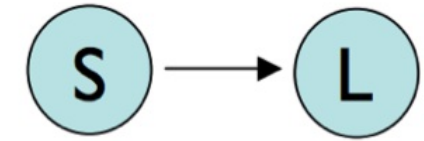


Variability - Scenario 3

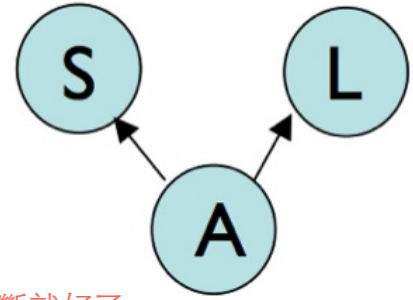
差異很大，sign-up 版本的捐款低，donate 版本捐款高



Confounding



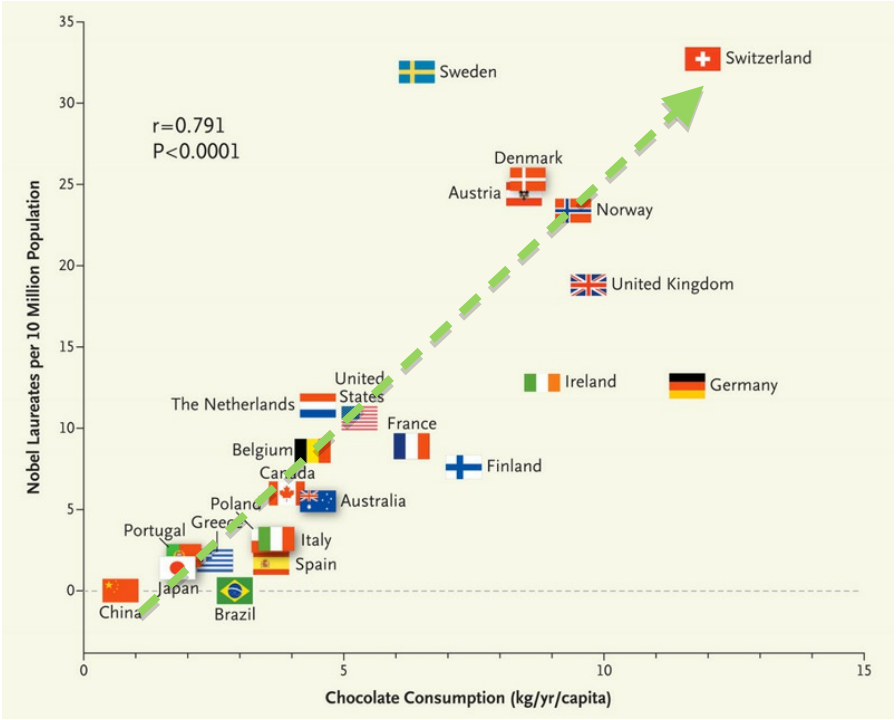
Shoes's size　　Literacy



Age

我們是要找 S 跟 L 之間的 correlation ，應該由第一個圖來判斷就好了
但我們把 age 加進來後，反而導致 S & L 之間的 correlation 不明顯，
此時我們把這種情況稱為 confounding， Age 稱為 confounder

Correlation is not causation*



從圖表中我們可以看出明顯的趨勢，告訴我們似乎巧克力消耗量越多的國家，諾貝爾獎得主會越多；但實際上巧克力消耗量或許不是直接原因，因為有可能這些國家本身經濟條件較好，所以花在教育上的預算可以較多，
同時也因為經濟條件佳而消耗較多的巧克力，因此雖然兩筆數據有關係，但巧克力消耗量並非諾貝爾獎人數的直接原因．

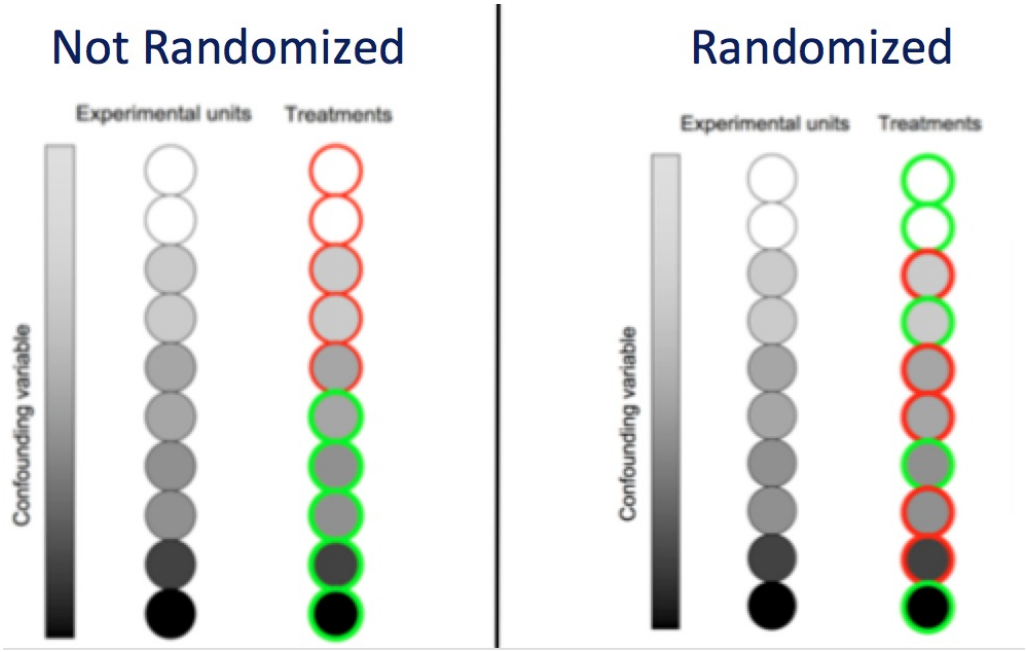
http://www.nejm.org/doi/full/10.1056/NEJMon1211064

*Sometimes called spurious correlation**

Randomization and blocking

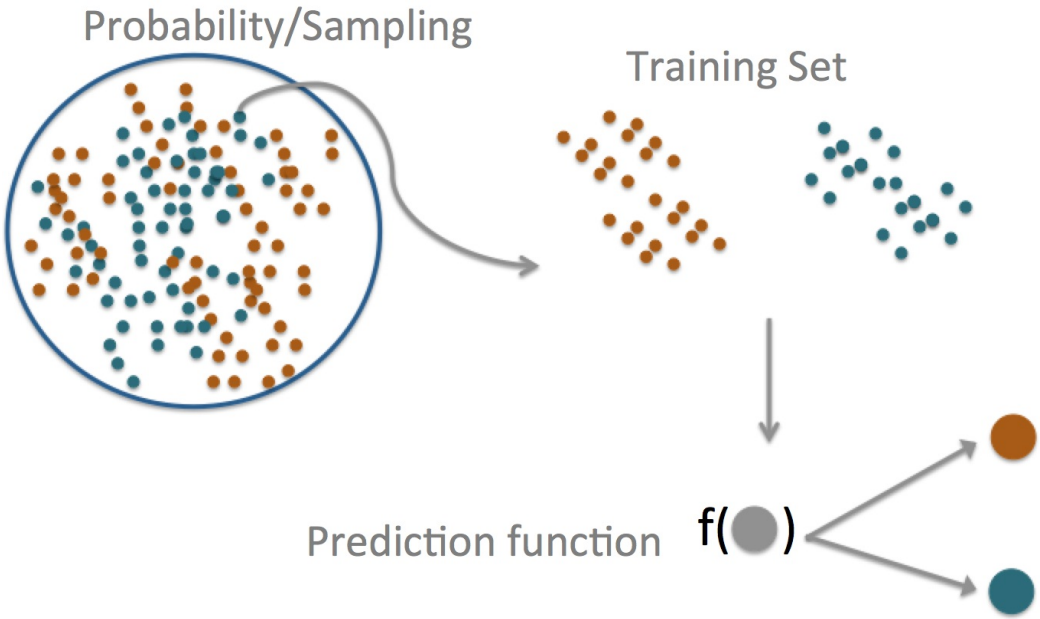
- If you can (and want to) fix a variable
  - Website always says Obama 2014 on it
- If you don't fix a variable, stratify it
  - If you are testing sign up phrases and have two website colors, use both phrases equally on both.
- If you can't fix a variable, randomize it

Why does randomization help?

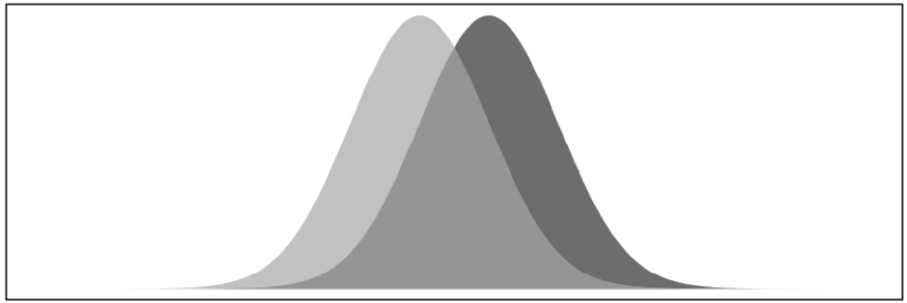


http://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture1.pdf

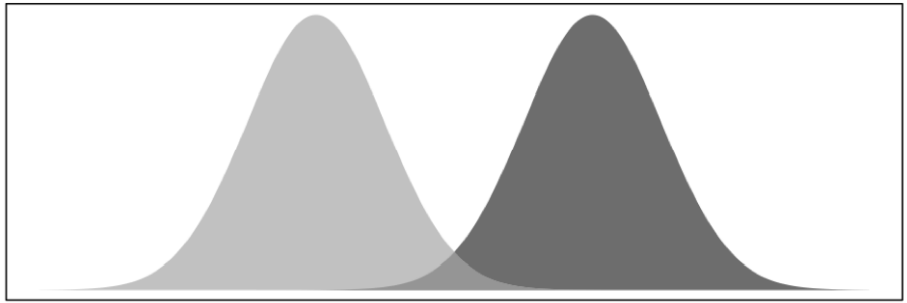
Prediction



Prediction versus inference



圖（一）

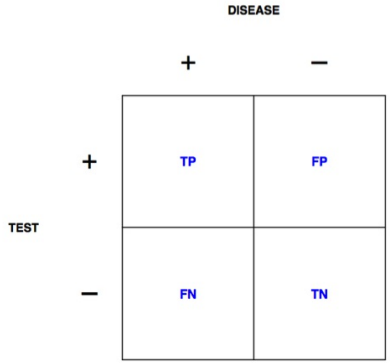


圖（二）

prediction 比起 inference 更加困難，如下圖（一），因為我們兩個平均值太過相近，我們無法去做預測，如果需要做預測，我們會需要像圖（二）這樣兩個平均值至少相差一定距離才行．

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

Prediction key quantities



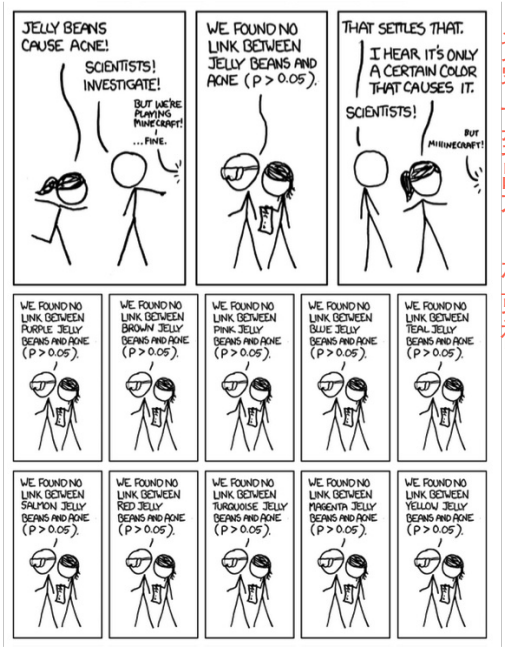
Sensitivity	→	Pr (positive test | disease)
Specificity	→	Pr (negative test | no disease)
Positive Predictive Value	→	Pr (disease | positive test)
Negative Predictive Value	→	Pr (no disease | negative test)
Accuracy	→	Pr (correct outcome)

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

Beware data dredging

http://xkcd.com/882/

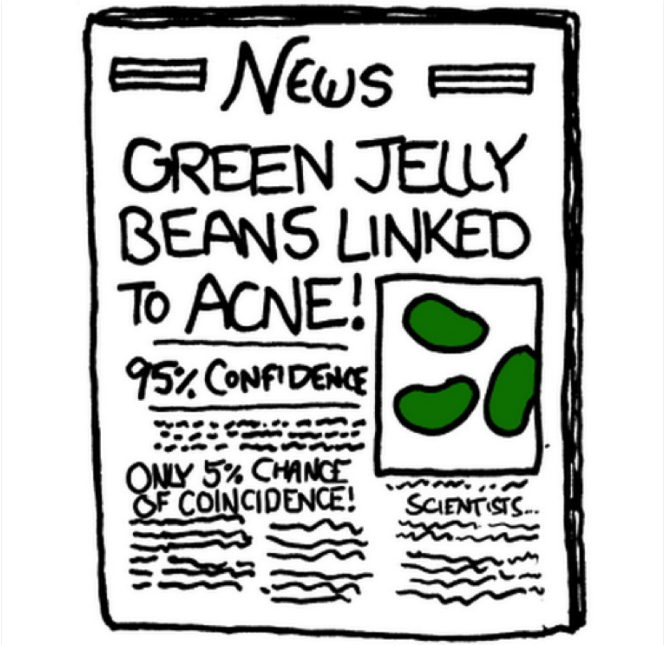
Beware data dredging



資料挖泥，一個在資料探勘過程中容易犯的錯誤，如圖畫中所說，在一堆其實沒有關聯的數據之中，找到一個相對較大的關聯，有時還會出現十分可笑的應用，例如發掘出不存在但看起來振奮人心的模式（特別的因果關係），這些根本不相關的、甚至引人誤入歧途的、或是毫無價值的關聯，在統計學文獻裡通常被戲稱為「資料挖泥」（Data dredging, data fishing, or data snooping）。

http://xkcd.com/882/

Beware data dredging



http://xkcd.com/882/

Summary

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