What's beyond

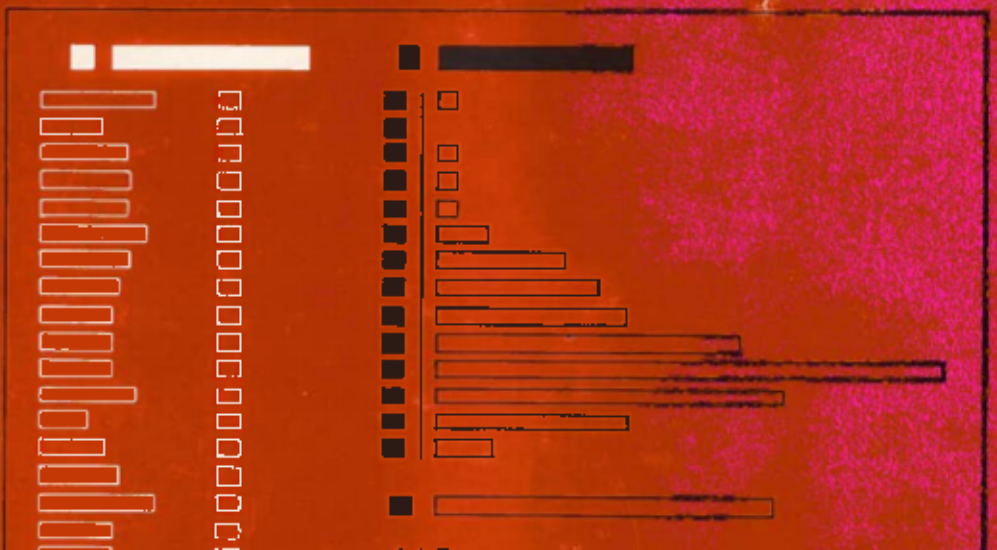
H_0 vs H_1 ?

The hardest thing of all is to find a black cat in a dark room, especially if there is no cat.

~ Confucius

John W. Tukey

EXPLORATORY DATA ANALYSIS



Preface

This book is based on an important principle:

It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it.

(Very) High Dimensional Data

Lung cancer : 260 observations

47 000 probes

Curse of dimensionality

- sparsity
- distance

H_0 vs H_1 solution

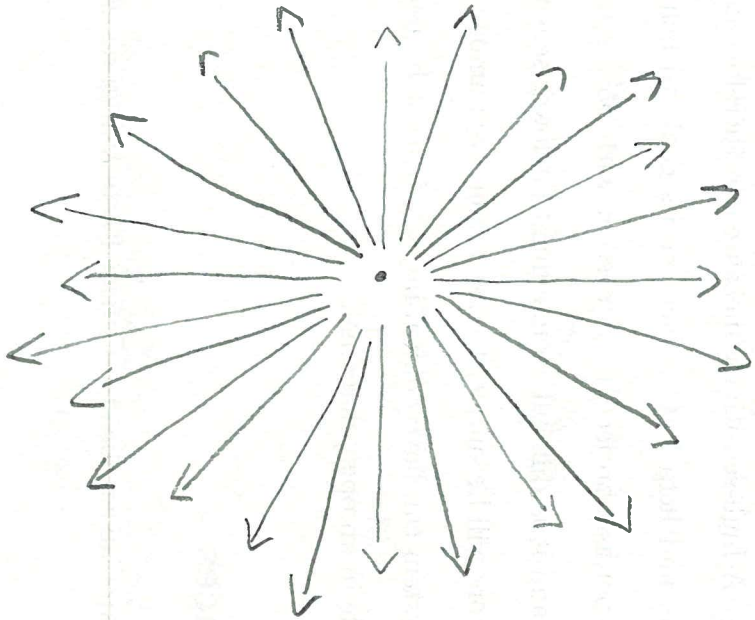
- 47,000 1-dim data
  adjust for multiple testing

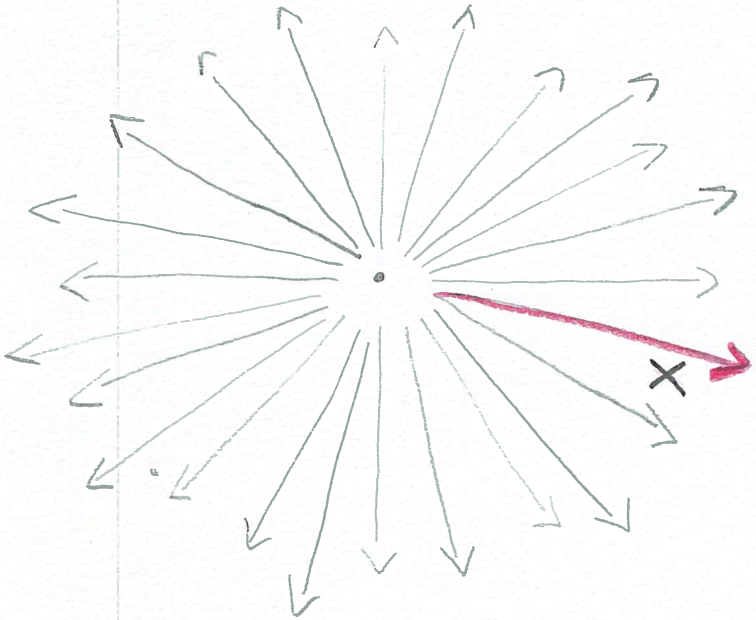
- reduce 47,000 → 10,000
  by selection of probes

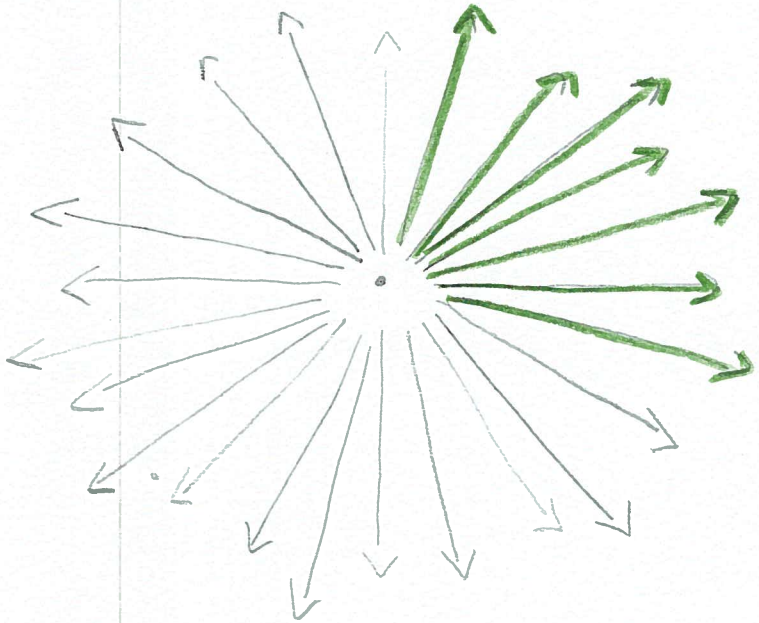
Which assumptions did you make?

- the software
- the function
- the parameter

ALWAYS VIOLATED




x



Are any of the 47,000 probe values correlated to lung cancer?

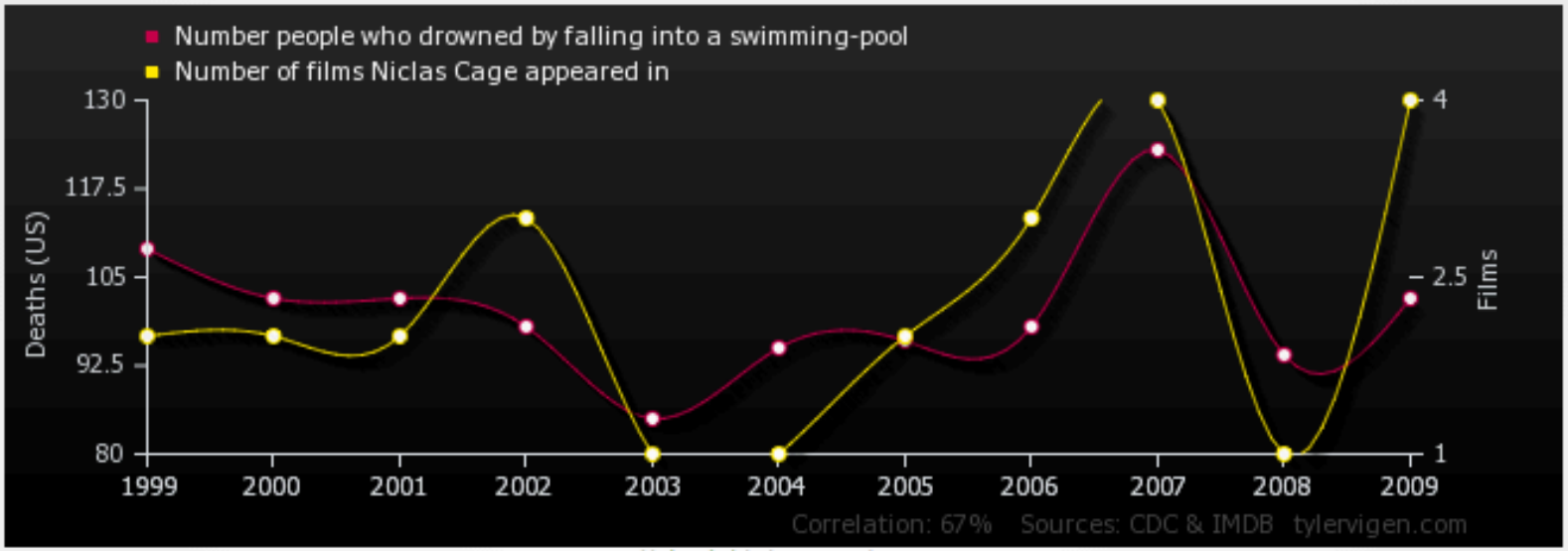
Which of the 47,000 probe values are most likely correlated to lung cancer?

How can the data best be described?

Number people who drowned by falling into a swimming-pool

correlates with

Number of films Nicolas Cage appeared in



Upload this image to imgur

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Number people who drowned by falling into a swimming-pool Deaths (US) (CDC)	109	102	102	98	85	95	96	98	123	94	102
Number of films Nicolas Cage appeared in Films (IMDB)	2	2	2	3	1	1	2	3	4	1	4

Correlation: 0.666004

Permalink - Mark as interesting (24,541) - Not interesting (11,190)

View all correlations - Discover a new correlation

Re-Chart