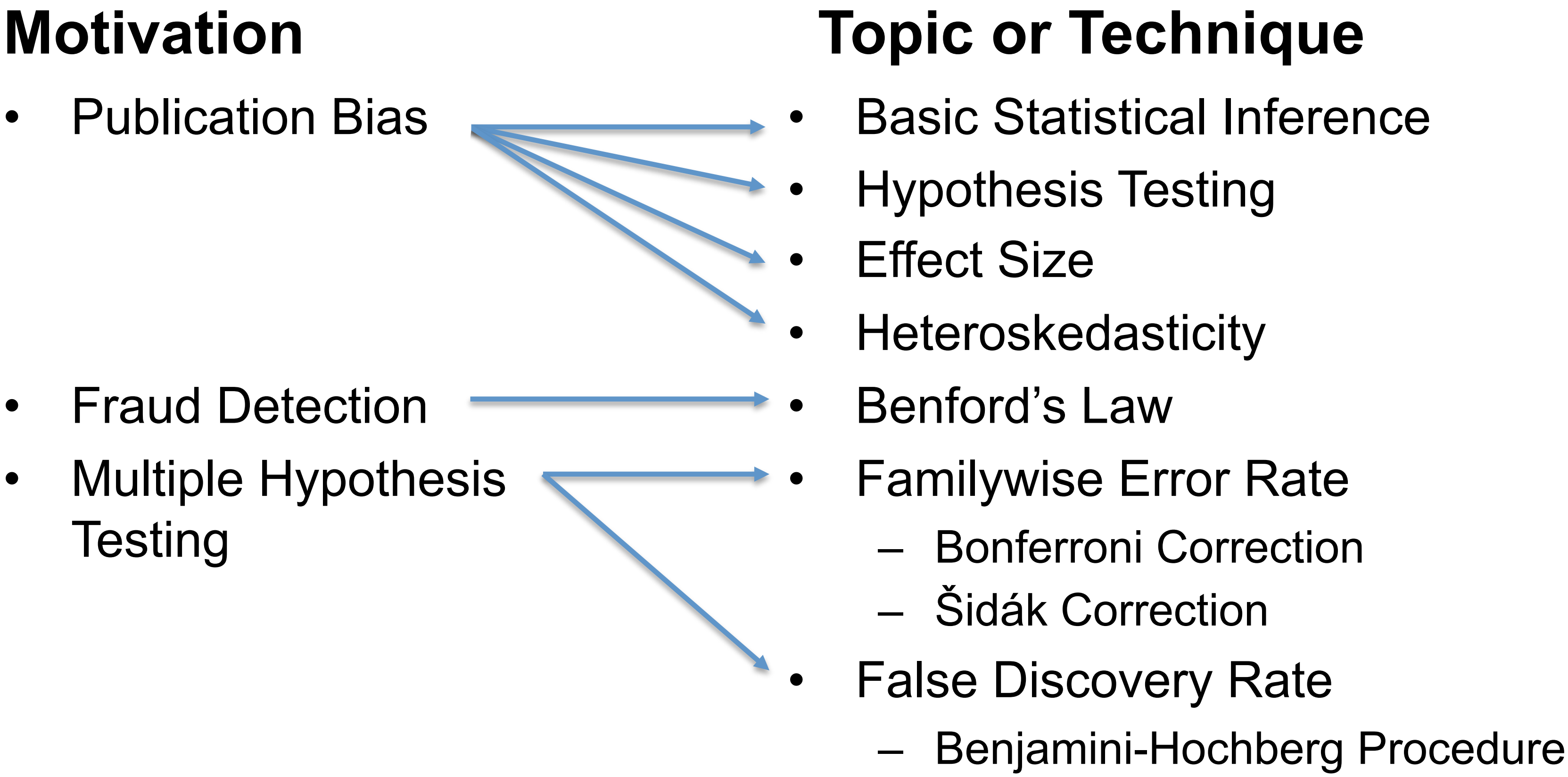
Where we are



Motivation

- Publication Bias
- Fraud Detection
- Multiple Hypothesis Testing

Topic or Technique

- Basic Statistical Inference
- Hypothesis Testing
- Effect Size
- Heteroskedasticity
- Benford’s Law
- Familywise Error Rate
  - Bonferroni Correction
  - Šidák Correction
- False Discovery Rate
  - Benjamini-Hochberg Procedure

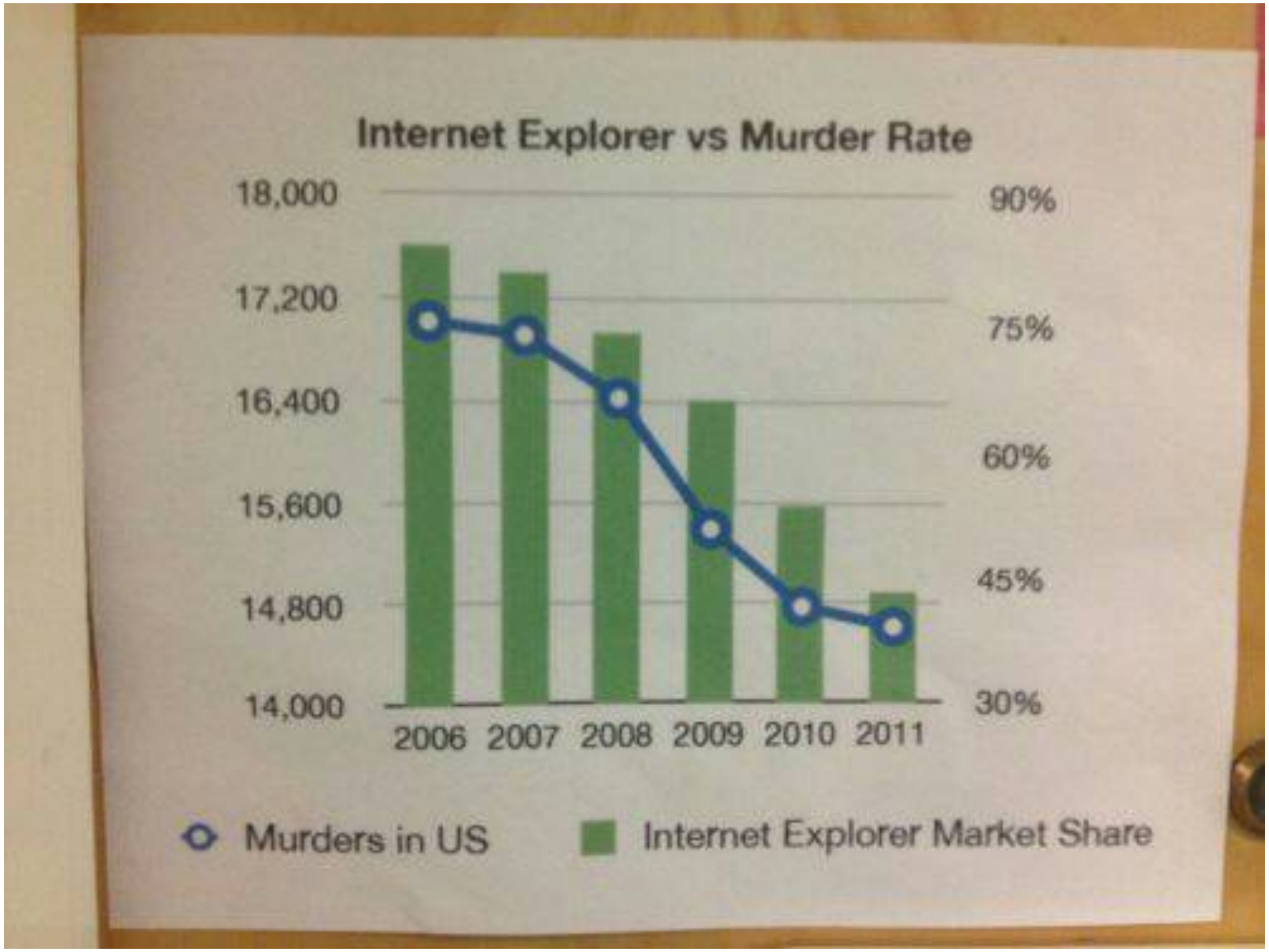
What about Big Data?

"Classical statistics was fashioned for small problems, a few hundred data points at most, a few parameters."

"The bottom line is that we have entered an era of massive scientific data collection, with a demand for answers to large-scale inference problems that lie beyond the scope of classical statistics."

Bradley Efron,
Bayesians, Frequentists, and Scientists

http://www-stat.stanford.edu/~ckirby/brad/papers/2005BayesFreqSci.pdf


Internet Explorer vs Murder Rate
18,000
17,200
16,400
15,600
14,800
14,000
90%
75%
60%
45%
30%
2006
2007
2008
2009
2010
2011
Murders in US
Internet Explorer Market Share

Positive Correlations

- Number of police officers and number of crimes (Glass & Hopkins, 1996)
- Amount of ice cream sold and deaths by drownings (Moore, 1993)
- Stork sightings and population increase

(Box, Hunter, Hunter, 1978)

The “curse” of Big Data?

“…the curse of big data is the fact that when you search for patterns in very, very large data sets with billions or trillions of data points and thousands of metrics, you are bound to identify coincidences that have no predictive power.”

Vincent Granville

http://www.analyticbridge.com/profiles/blogs/the-curse-of-big-data

Vincent Granville's Example

- Consider stock prices for 500 companies over a 1-month period
- Check for correlations in all pairs

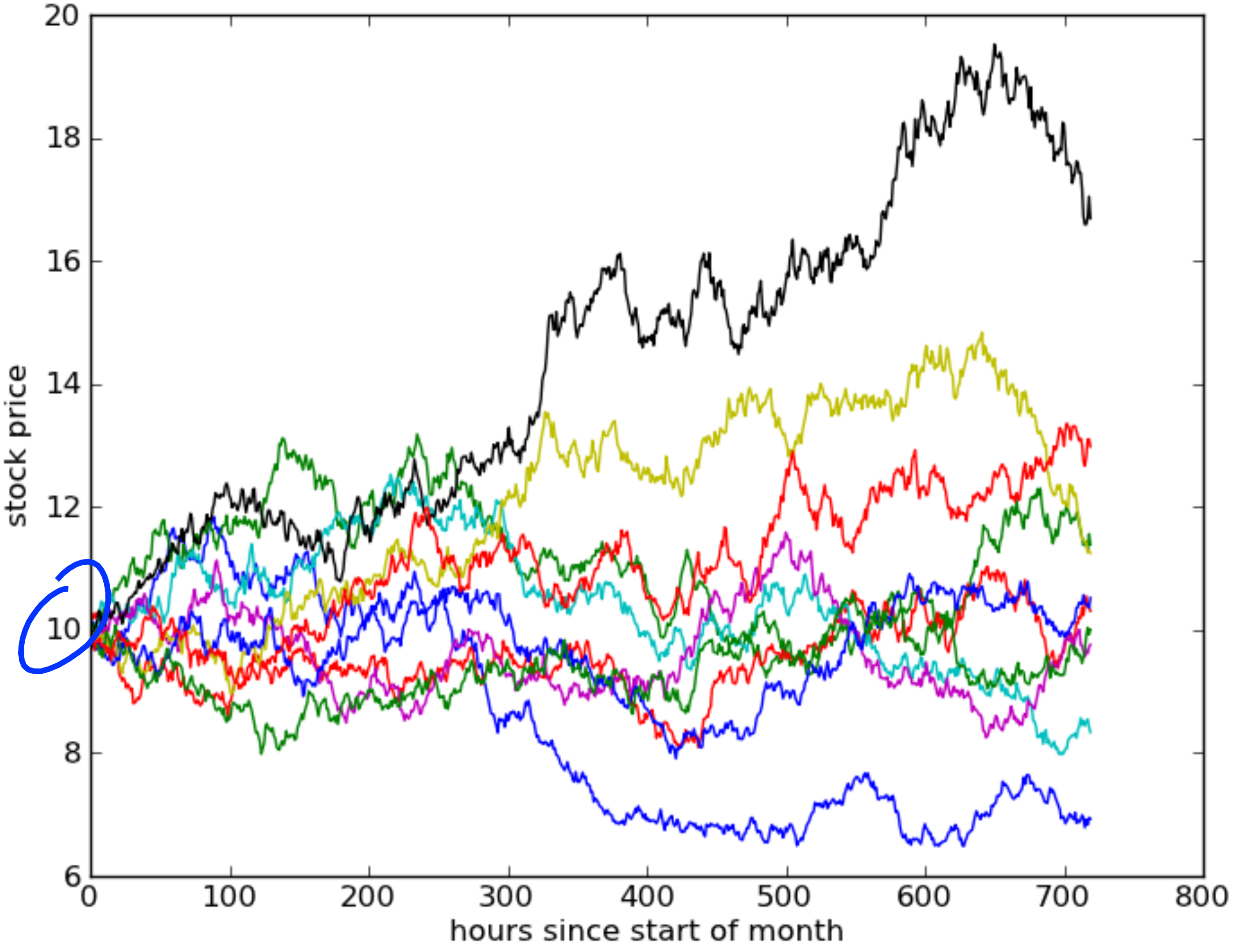
Aside: Very Basic Timeseries Analysis (1)

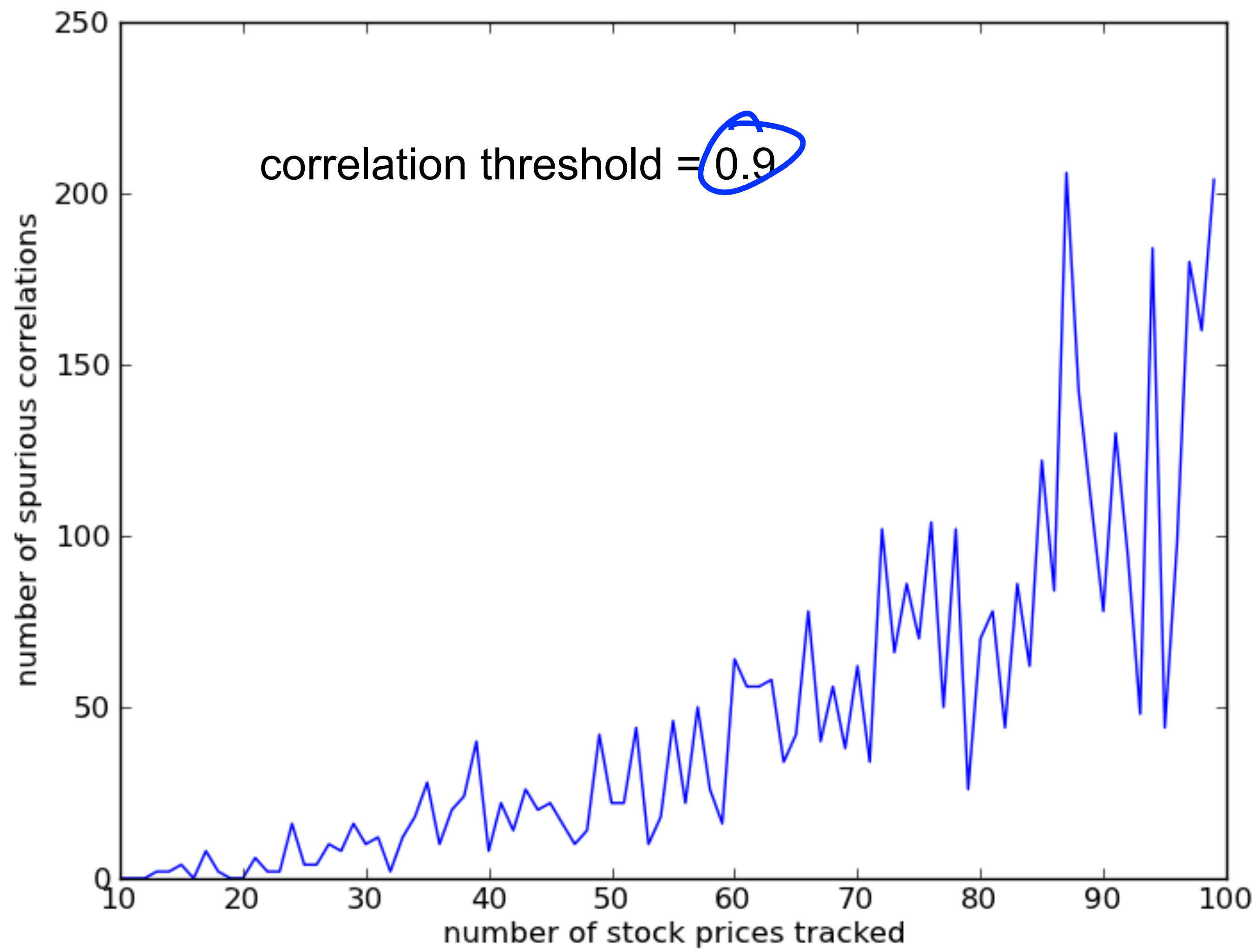
$$cov(x, y) = \sum_i (\underline{x_i} - \underline{u_x})(\underline{y_i} - \underline{u_y})$$

$$corr(x, y) = \frac{cov(x, y)}{\sqrt{\sum_i (x_i - u_x)^2}\sqrt{\sum_i (y_i - u_y)^2}}$$

standard deviation

random walk
each step is normally distributed @ 1% of current price




correlation threshold = 0.9
number of spurious correlations
0
50
100
150
200
250
10
20
30
40
50
60
70
80
90
100
number of stock prices tracked

Is Big Data different?

- Big P vs. Big N
  - P = number of variables (columns)
  - N = number of records
- Marginal cost of increasing N is essentially zero!
- But while >N decreases variance, it amplifies bias
  - Ex: You log all clicks to your website to model user behavior, but this only samples current users, not the users you want to attract.
  - Ex: Using mobile data to infer buying behavior
- Beware multiple hypothesis tests
  - “Green jelly beans cause acne”
- Taleb’s “Black Swan” events
  - The turkey’s model of human behavior