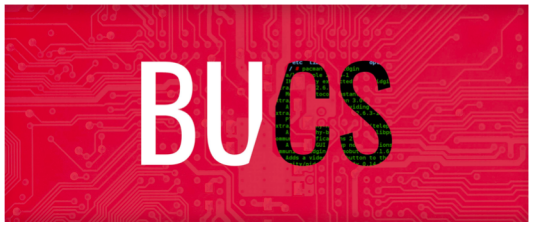
Lecture 18
Interpreting Results: Pitfalls
CS 506 & EC 500



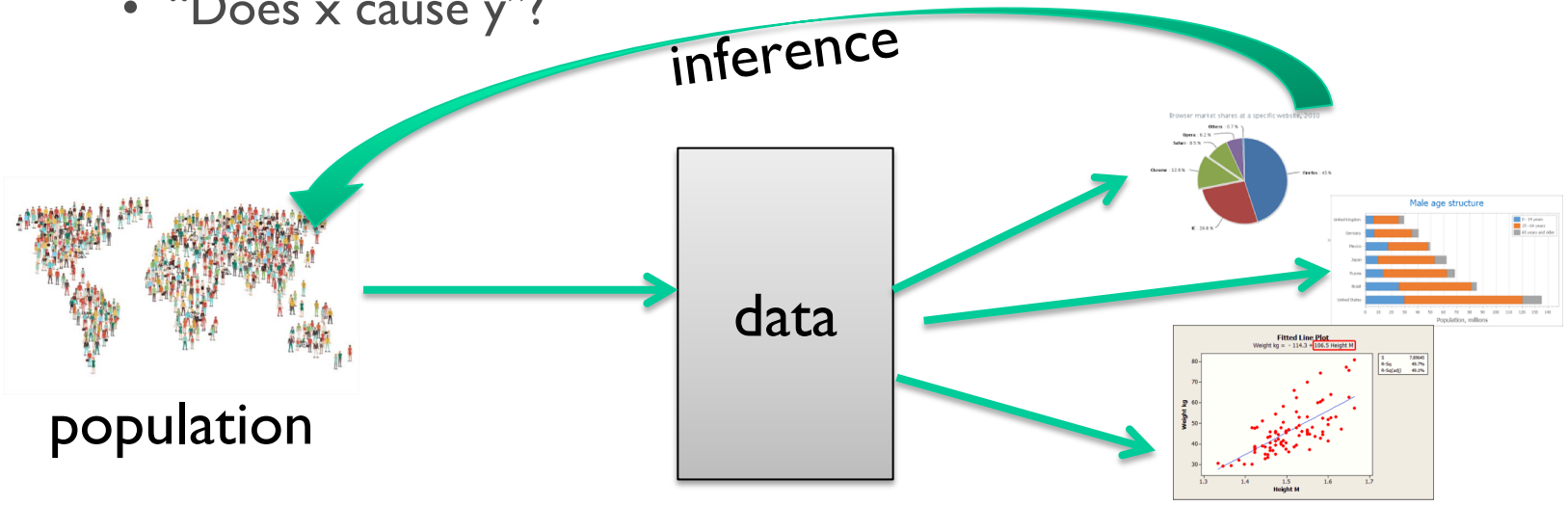
Adam Smith

BU Computer Science

Data Science is about…

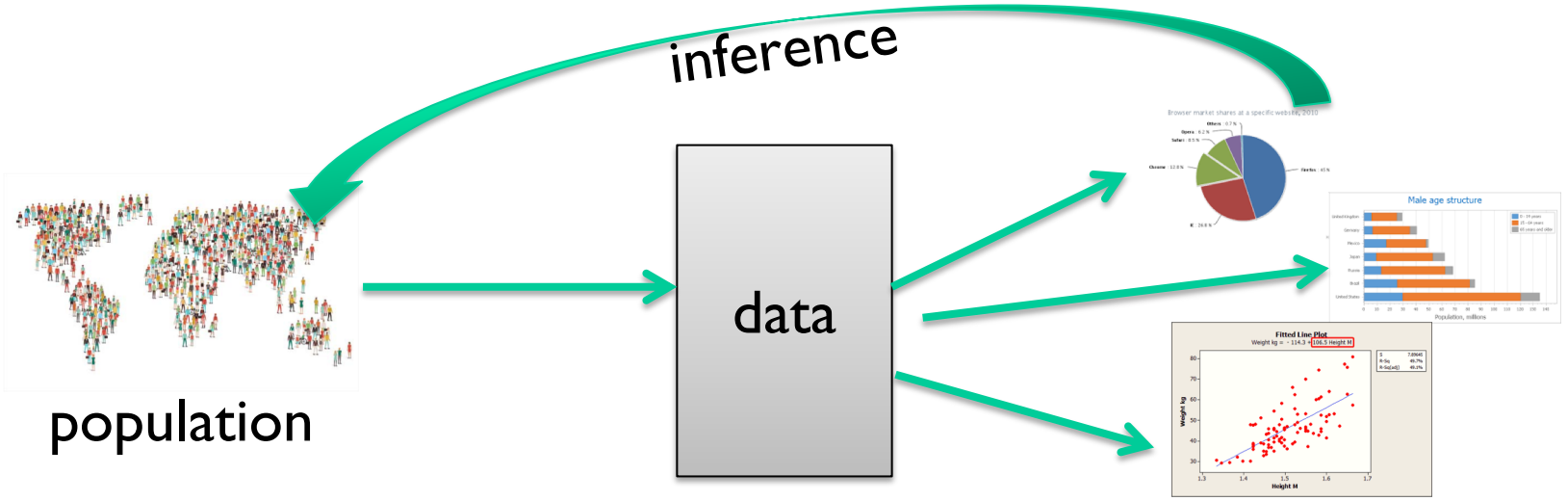
… formulating and answering questions based on data

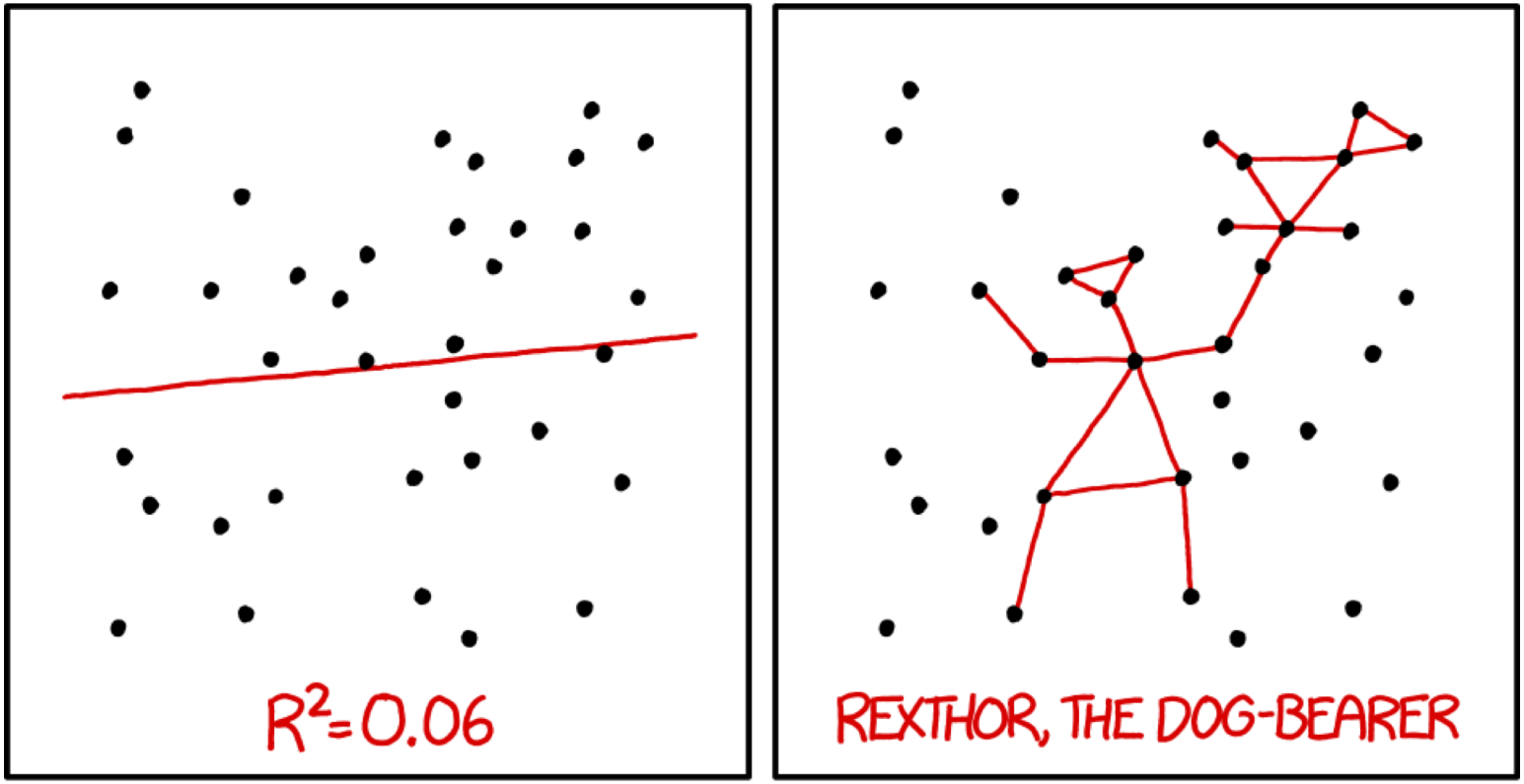
- Many different types of questions
  - Summarization of this data, e.g.
    - histograms, maps, visualizations
  - Inference about (unknown) population or future, e.g.
    - Prediction on unseen examples
    - "Is x significantly correlated to y"?
    - "Does x cause y"?



Interpreting results

- Most subtle aspect of data science
  - Focus of most of statistics!
- This course: "interpreting" = "(how) does this answer my questions"?
  - We'll see some tools, but formulas are not enough
  - Common sense and judgment are necessary



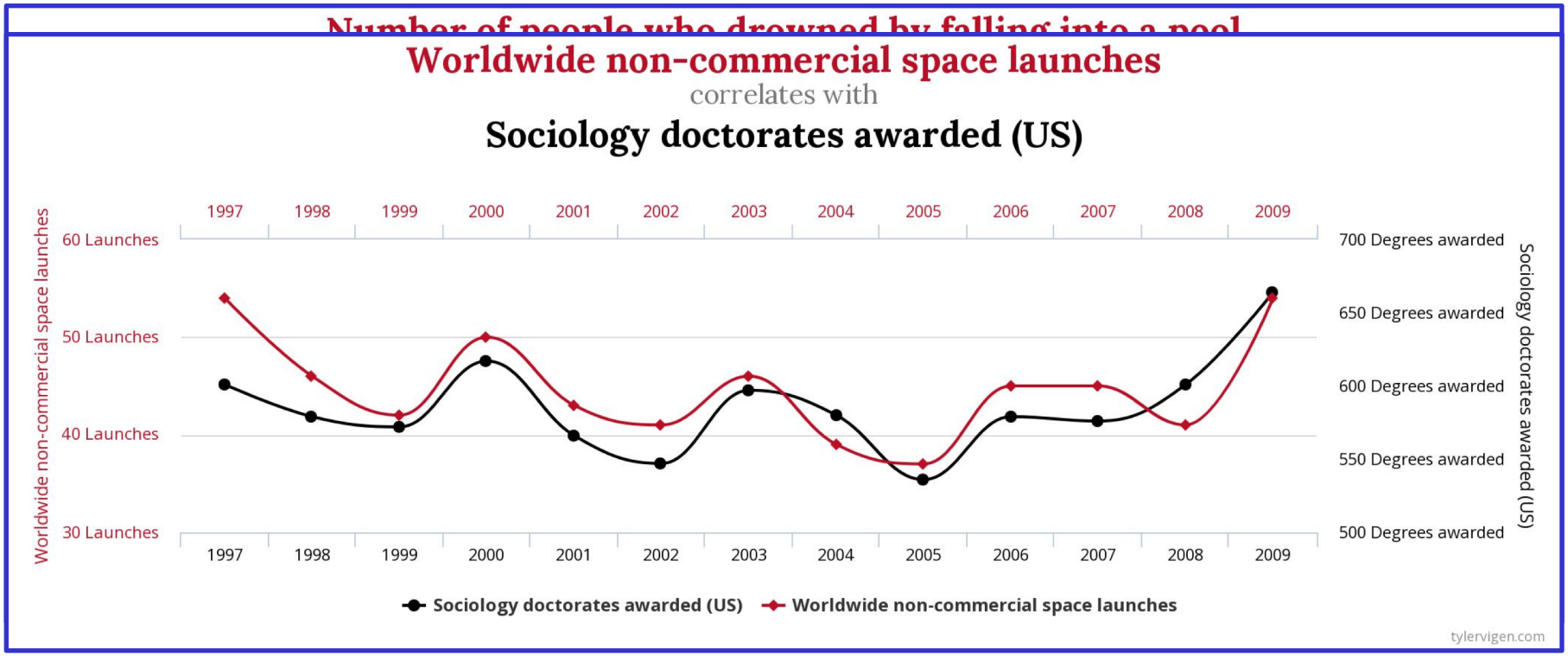


I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

https://xkcd.com/1725/

Signal versus noise

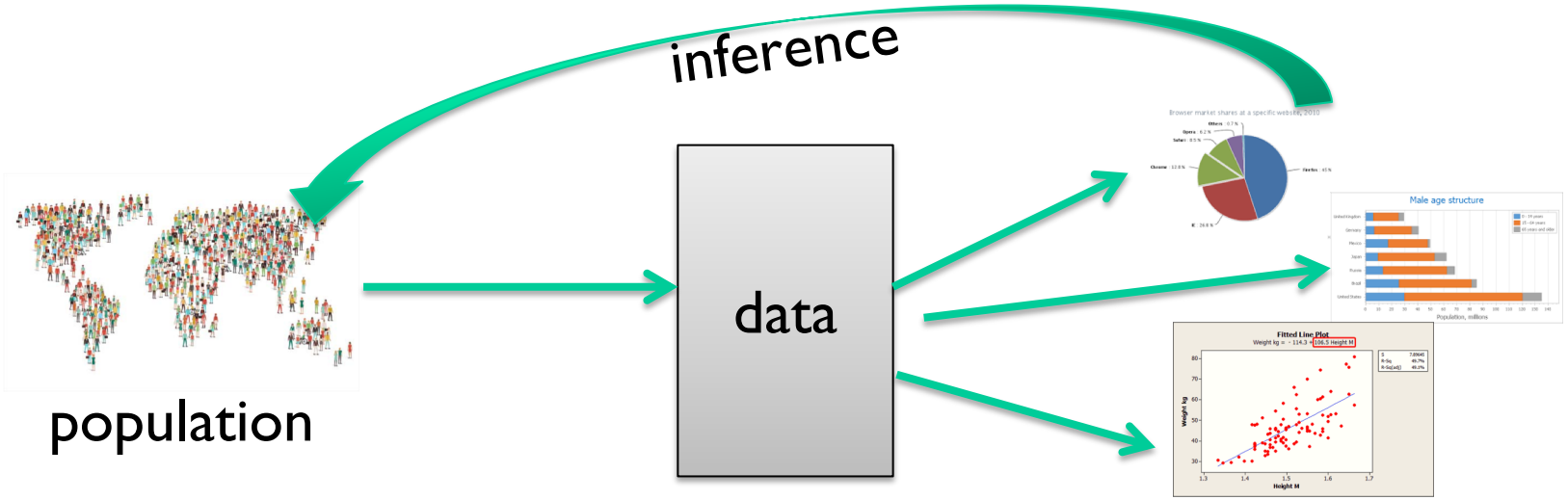
- Basic problem in inference
- How can we tell apart real effects from random occurrences?



http://www.tylervigen.com/spurious-correlations

Important tools

- Holdouts and cross-validation
  - Crucial for evaluating prediction
- Goodness of fit measures
  - E.g. R^2, clustering scores (adjusted Rand index, silhouette),…
- Confidence intervals
- Hypothesis tests and p-values



Confidence intervals

- Suppose we want to estimate a parameter of the population distribution
  - Running example: success rate of a classifier
- We might make some assumptions
  - E.g. data points are drawn independently from the population
  - (Not true, e.g., for time series)
- Based on these assumptions, for each value of the parameter, there is a distribution on the observed data
  - Example: fix classifier, draw $n = 1000$ fresh data points
  - Observe misclassification rate
  - Distributions is _____?
- Top hat question: if the true misclassification rate is $q \in [0,1]$, what is the expectation of the observed number X of mistakes?

a) $E(X) = nq$
b) $E(X) = n(1-q)$
c) $E(X) = nq(1-q)$
d) $E(X) = nq^2$
e) None of the above

Confidence intervals

- Given q, what is the distribution of X?

a) Binomial (n,q)

b) Normal(μ, σ^2) with $\mu = nq$, $\sigma = \sqrt{nq(1-q)}$

c) Poisson(nq)

d) None of the above

- Given X, what is the distribution of q?

a) Normal

b) Binomial

c) Poisson

d) Beta

e) None of the above

Confidence intervals

- The distribution of q given X is undefined since we have not assumed a prior distribution on q
  - We are missing terms in Bayes rule
- How do we estimate our uncertainty about q?
- A confidence interval is a function that maps data to pairs of numbers $a(data), b(data)$
  - If assumptions are satisfied, then for all q,
    $$\Pr_{X \sim Bin(n,q)}\left(\; q \in [a(X), b(X)] \; \right) \geq 0.95$$
  - Here 0.95 is called the "coverage"
- A Bayesian credible interval has a more straightforward interpretation
  - Assume a prior, compute an interval with posterior probability at least 0.95
  - Have to be careful about choosing a good prior

How can we compute confidence intervals?

- Say we have code for the probability mass function and cumulative distribution function for binomial
- How do we compute a confidence interval with 95% coverage?

- We did this on the board. Given X, the idea was to compute $\hat{q} = \frac{X}{n}$ and then find a, b such that

$$CDF_{Bin(n,a)}(\hat{q}) = 0.975$$
$$CDF_{Bin(n,b)}(\hat{q}) = 0.025$$

  - We said that we can get very good approximations to a, b using binary search, and the CDF computed, for example, by scipy.stats.binom.cdf(…) or scipy.stats.binom.ppf(…)