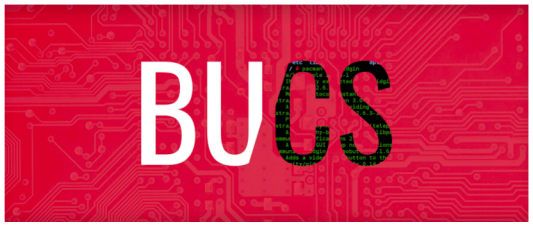
Lectures 18 & 19
Interpreting Results: Some Pitfalls
CS 506 & EC 500



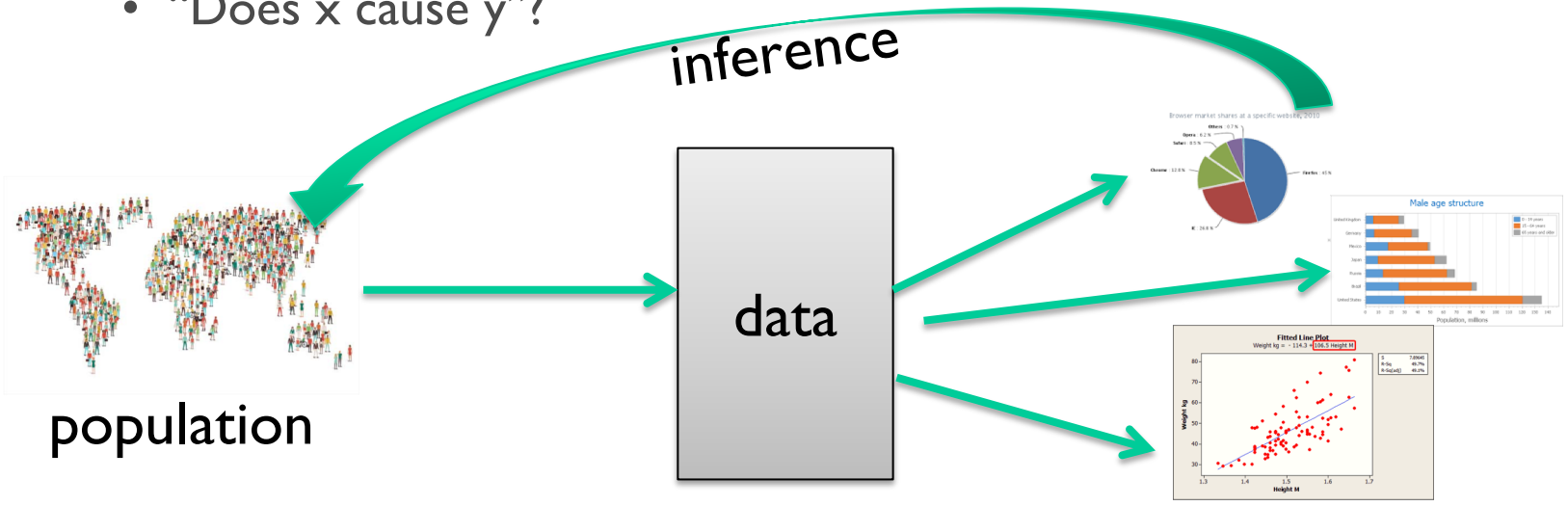
Adam Smith

BU Computer Science

Data Science is about…

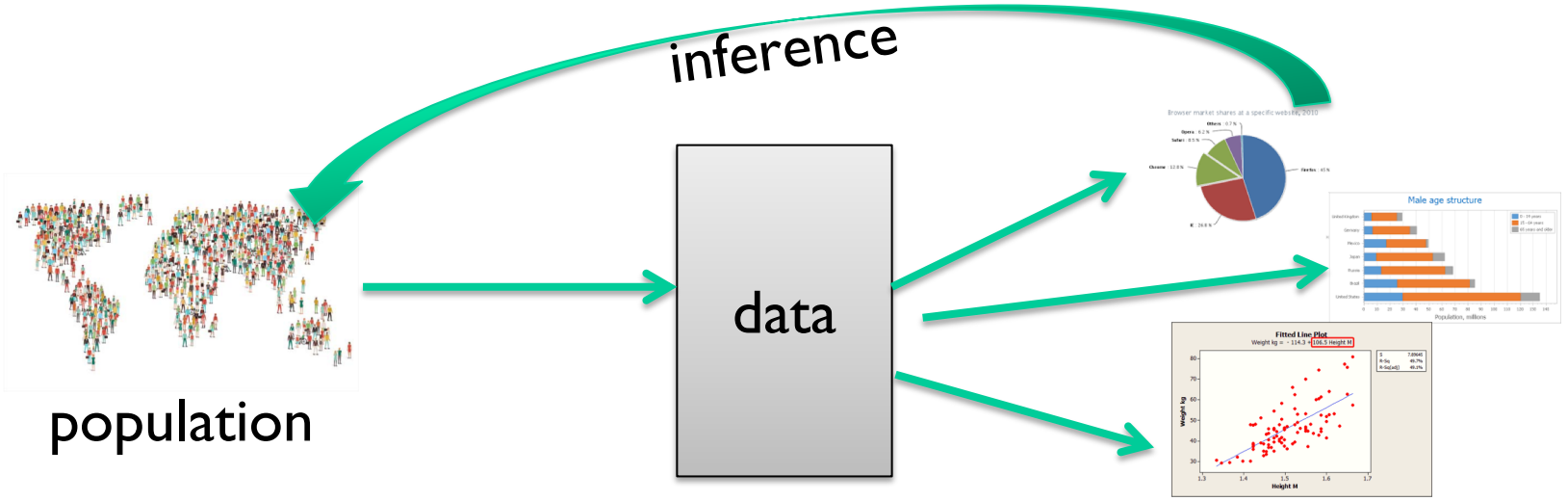
… formulating and answering questions based on data

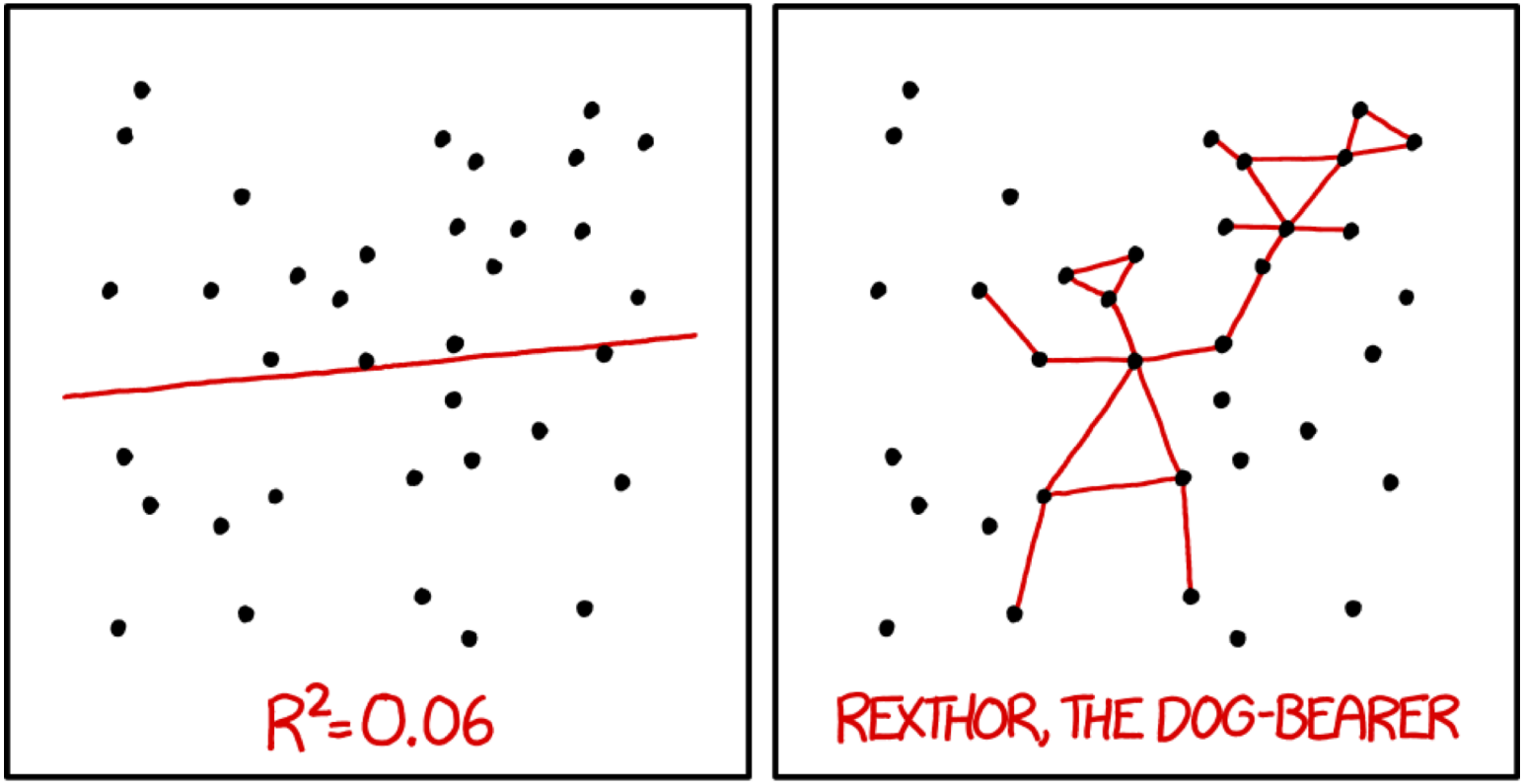
- Many different types of questions
  - Summarization of this data, e.g.
    - histograms, maps, visualizations
  - Inference about (unknown) population or future, e.g.
    - Prediction on unseen examples
    - "Is x significantly correlated to y"?
    - "Does x cause y"?



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



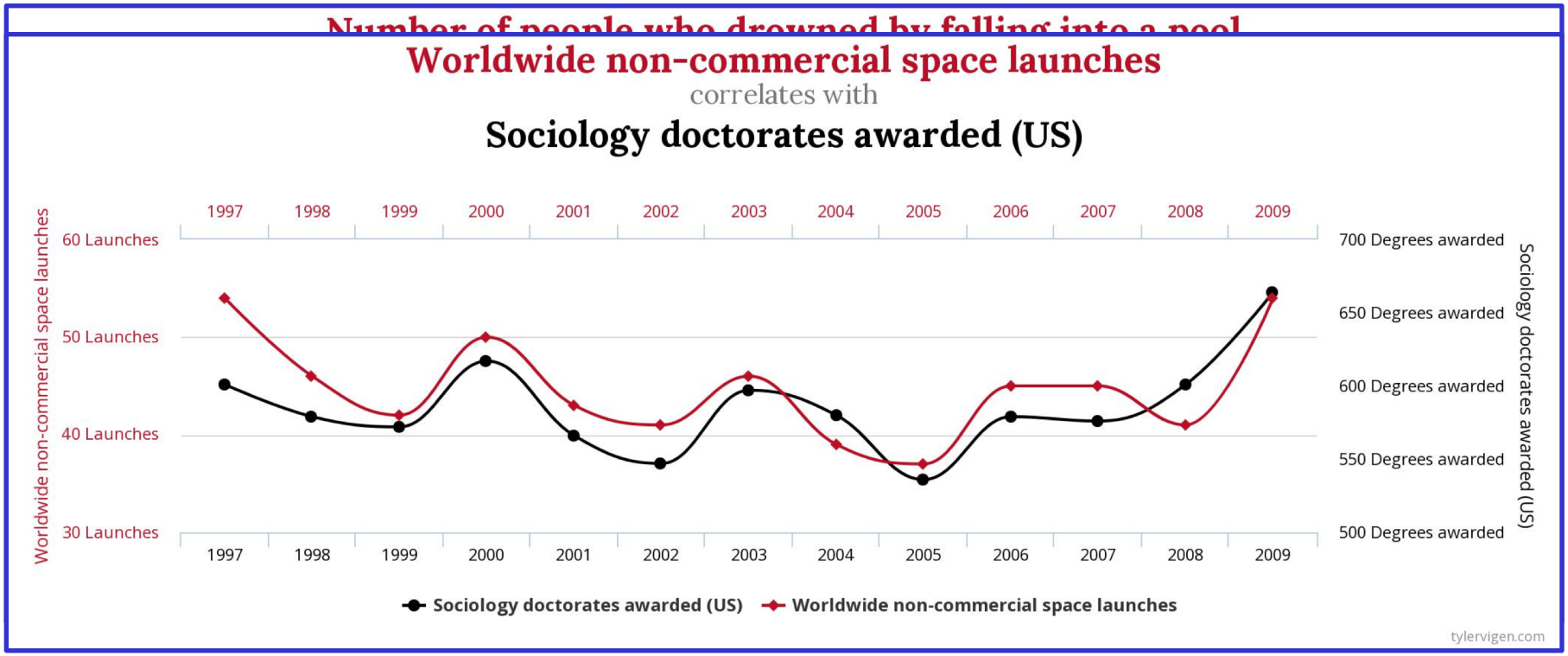


I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

https://xkcd.com/1725/

Signal versus noise

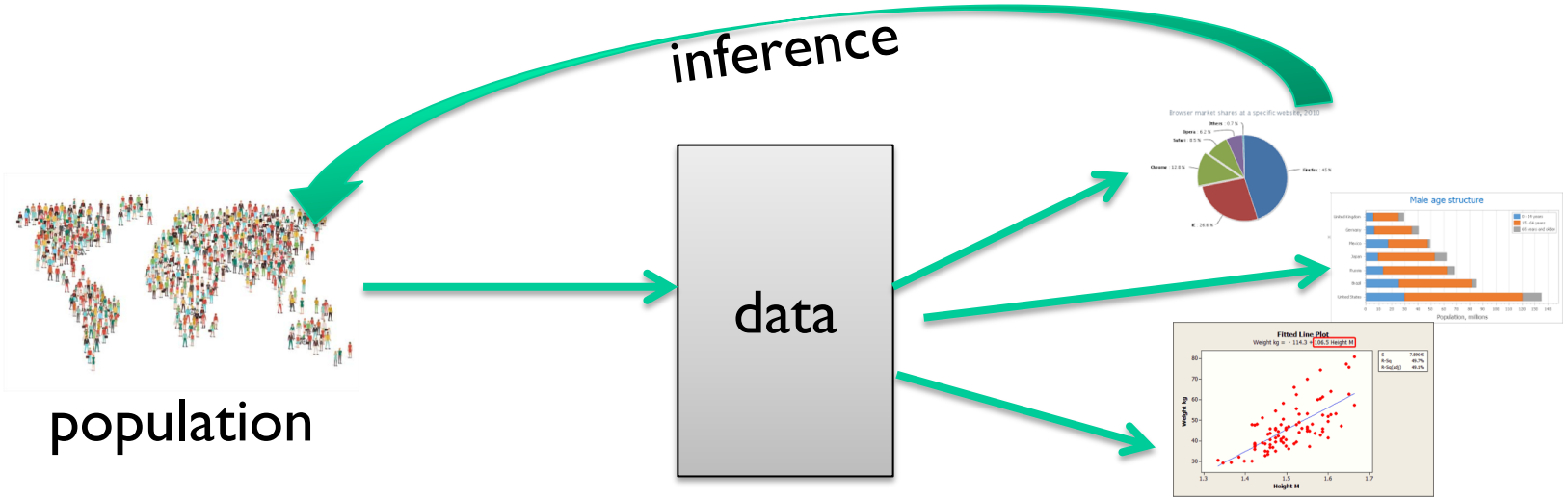
- Basic problem in inference
- How can we tell apart real effects from random occurrences?



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
  - But 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

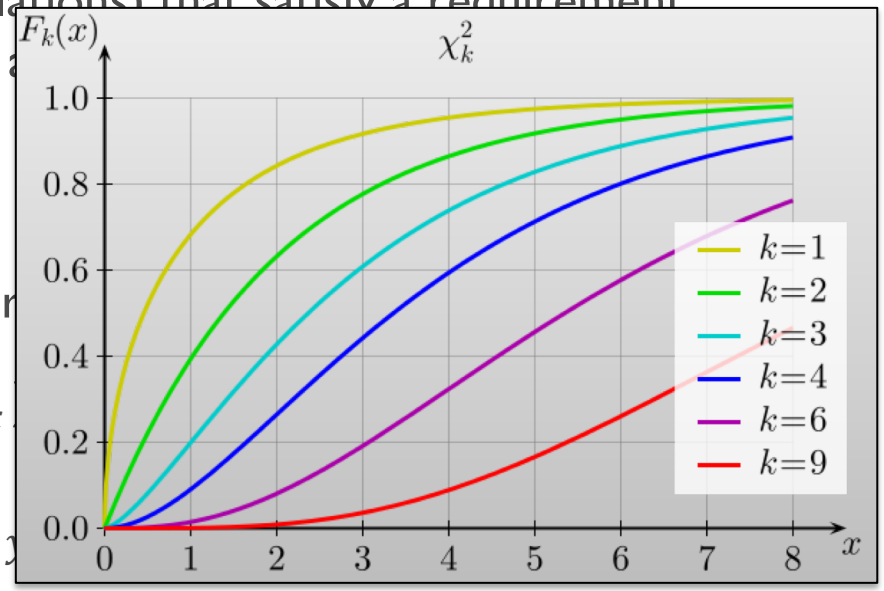
$$CDF_{Bin(n,a)}(\hat{q}) = 0.975$$
$$CDF_{Bin(n,b)}(\hat{q}) = 0.025$$

  - 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(…)

Hypothesis testing

Major tool for drawing qualitative "conclusions" about underlying population/distribution

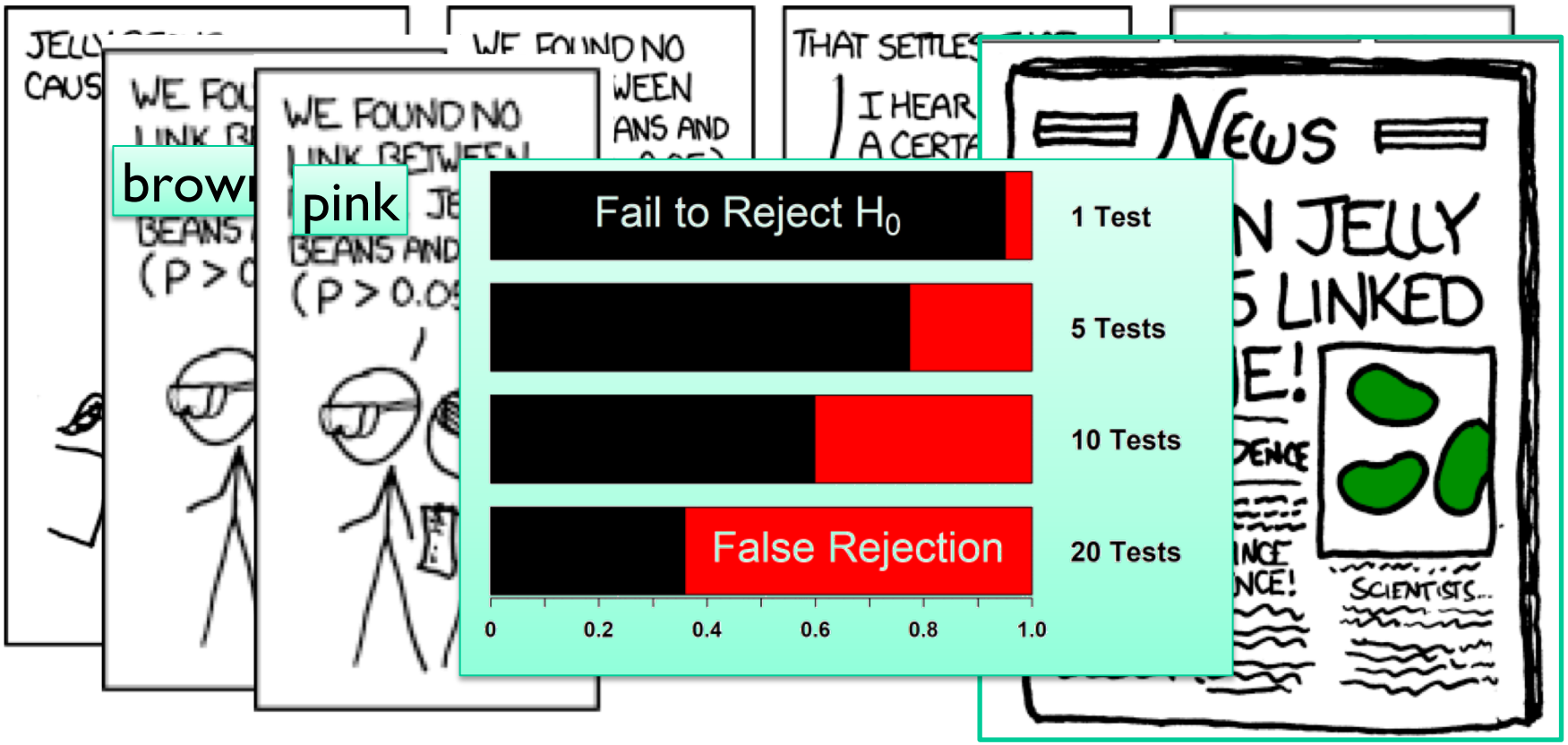
- Specify a "null hypothesis"
  - Set of possible distributions (= populations) that satisfy a requirement
  - Example: party affiliation (Rep/Dem) and opinion (for / against) are independent
- Compute a "test statistic"
  - It is a real-value function of the data
  - Should have the same distribution for
  - Example: chi-squared $T(data) = \sum_x$
    - Here $x \in \{R, D\}$ and $y \in \{F, A\}$
    - $O_{x,y} = count(Party = x, Opinion = y$
    - $E_{x,y} = \frac{count(Party=x) \times count(Opinion=y)}{count(all)}$
  - Theorem: (Under a few assumptions) If Party and opinion are independent, then $T(data)$ follows a chi-squared distribution with 1 degree of freedom
    - Works for (almost) all independent distributions. Magic!
- Compute a "p-value": Probability that test statistic would exceed its observed value assuming the null hypothesis



Hypothesis testing

- If the null hypothesis is "true"
  (that is, if the true distribution satisfies the hypothesis), then the p-value is uniformly distributed in $[0.1]$
  - So $\Pr(p - value \leq 0.05) = 0.05$
- General methodology:
  - Select the hypothesis, test statistic T and significance level α before looking at the data
  - "Reject the null hypothesis" if p-value is at most α
- NB: If p is large, we do not have evidence that the null hypothesis is true
  - For example, if

Pitfall 1: Testing many hypotheses



- P-values meaningful only if your first pick hypotheses, then look at data
- If you consider many hypotheses, some will appear significant

https://xkcd.com/882/

How many hypotheses?

- Top Hat Question: If we test 100 independent hypotheses, what is the probability that at least one will appear significant at $p = 0.01$?

a) 0.01

b) $100 \times 0.01 = 1$

c) 0.01^{100}

d) $(0.99)^{100} \approx 0.37$

e) $1 - (0.99)^{100} \approx 0.63$

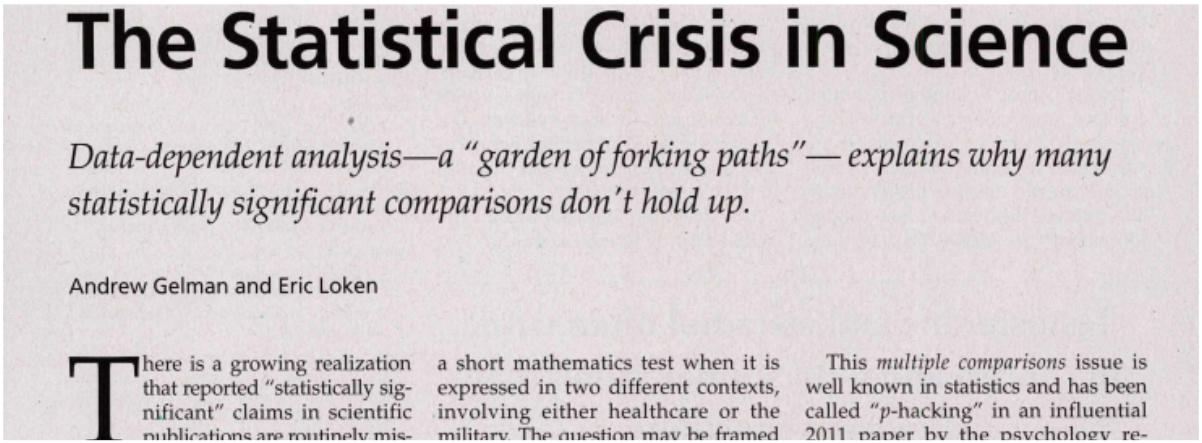
f) None of the above

Methodology for multiple hypothesis testing

- Adjust threshold so that p-values have the right interpretation:
  - If we test k hypotheses, Bonferoni correction sets each threshold at $p' = p/k$
  - Then the probability that any true hypothesis will be falsely rejected is at most $kp' = p$
- More sophisticated techniques exist
  - False discovery control sets thresholds so that fraction of rejected hypotheses is at most p
    - Requires extra ssumptions
- No matter what,
  - Track your hypotheses, use common sense
  - "If you torture the data enough, nature will always confess." Ronald Coase, 1981

Pitfall 2: Adapting hypothesis to data

- P-values meaningful only if your first pick hypotheses, then look at data
- If you choose the hypothesis based on the data, significance values are meaningless
- Example:
  - Test 1000 hypotheses
  - Pick most significant
  - Claim that is the question you were after



The Statistical Crisis in Science

Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up.

Andrew Gelman and Eric Loken

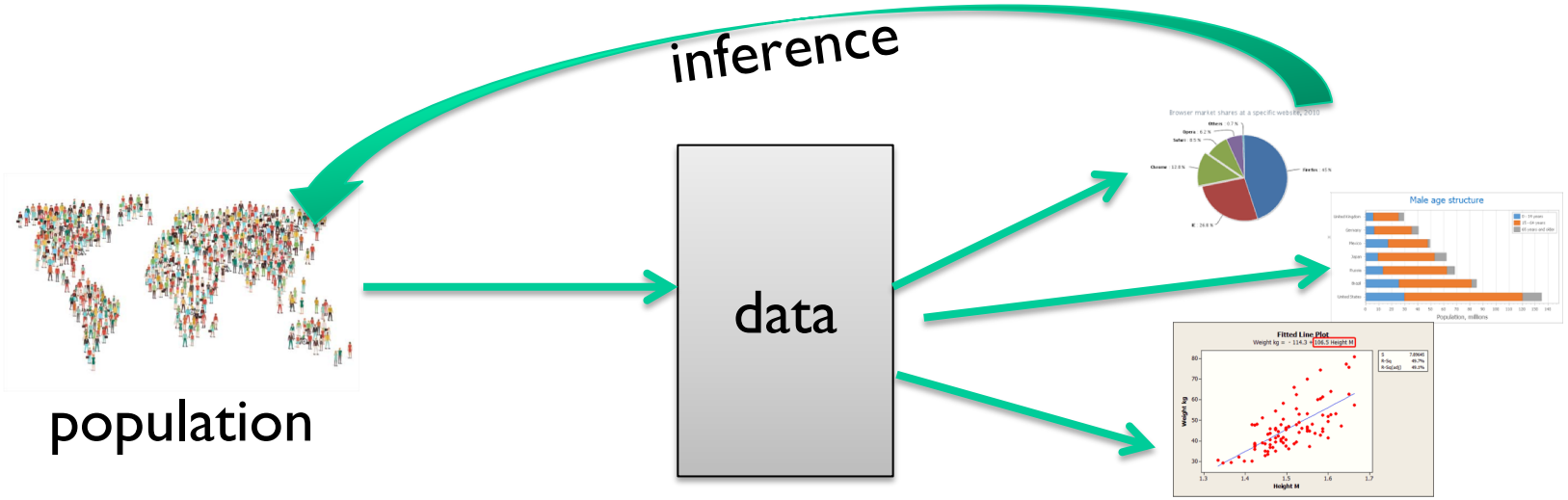
There is a growing realization that reported "statistically significant" claims in scientific

a short mathematics test when it is expressed in two different contexts, involving either healthcare or the military. The question may be framed

This *multiple comparisons* issue is well known in statistics and has been called "*p*-hacking" in an influential 2011 paper by the psychology re-

Important tools: reminder

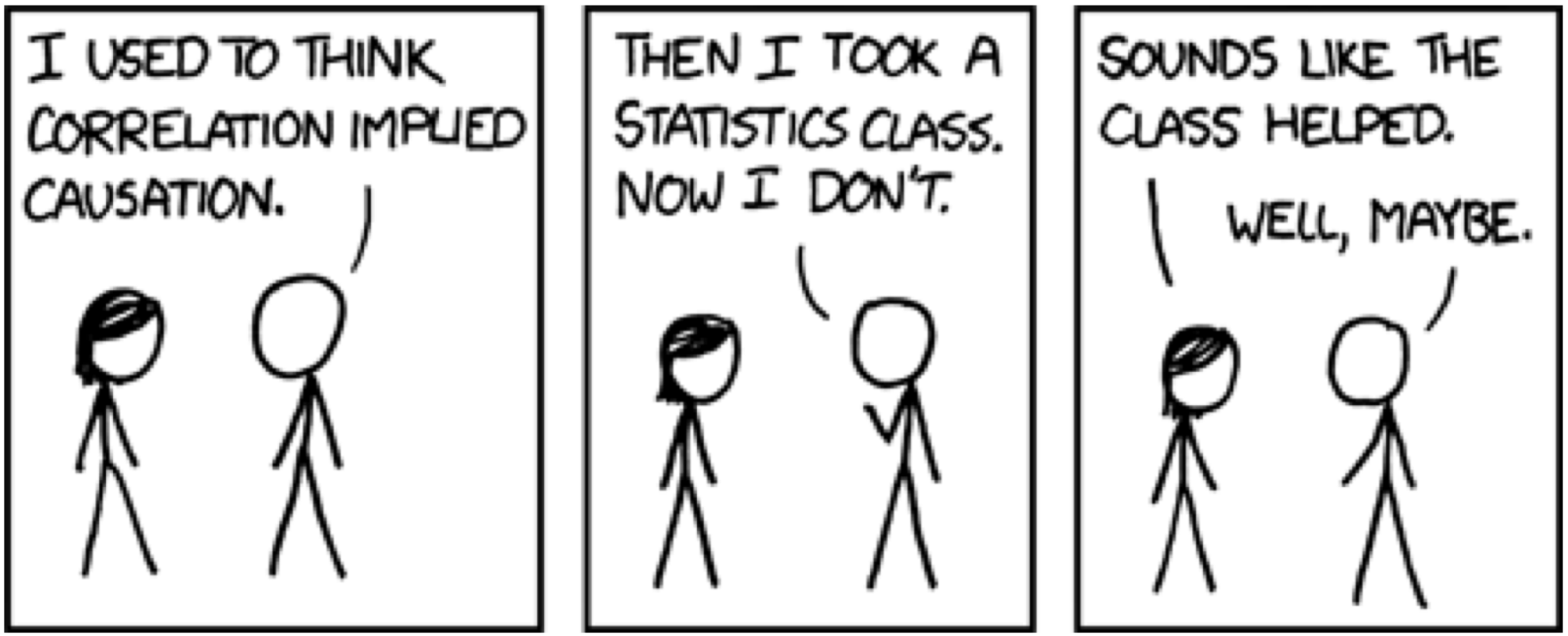
- Holdouts and cross-validation
  - Crucial for evaluating prediction
- Goodness of fit measures
  - E.g. R^2, clustering scores (adjusted Rand index, silhouette),…
- Confidence intervals
- Hypothesis tests and p-values



When statistical significance isn't enough

- Correlation v. causation
  - Wealthy people generally live in larger houses than poor people
  - But large houses don't (usually) cause wealth
- Unrepresentative data: asking the wrong people
  - Early person tracking in video worked only on subjects with white skin
- Unrepresentative data: correlated data

Correlation and causation



https://xkcd.com/552

Unrepresentative data

- Examples for the history books: 2016 US presidential election
  - Many polls placed Clinton in commanding lead
  - Media groupthink treated election as a sure thing
- Nice discussion:
  - http://fivethirtyeight.com/features/why-fivethirtyeight-gave-trump-a-better-chance-than-almost-anyone-else/

FiveThirtyEight

Politics Sports Science & Health Economics Culture

NOV. 11, 2016 AT 4:09 PM

Why FiveThirtyEight Gave Trump A Better Chance Than Almost Anyone Else

By Nate Silver

From the 538 article

- "Clinton was leading in the vast majority of national polls, […] So there wasn't any reasonable way to construct a polling-based model that showed Trump ahead.
  - Even the Trump campaign itself put their candidate's chances at 30 percent […]
- "But **people mistake having a large volume of polling data for eliminating uncertainty**.
  - Yes, having more polls helps to a degree, by reducing sampling error and by providing for a mix of reasonable methodologies. […]
  - Before long, however, you start to encounter diminishing returns.
- **"Polls tend to replicate one another's mistakes**:
  - If a particular type of demographic subgroup is hard to reach on the phone, … they're all likely to have problems of some kind or another.

Recommended reading

- "How to lie with statistics"
- "The Statistical Crisis in Science", Gelman and Loken

Other problems with interpretation

- Linear models are popular because they are intuitive
  - Coefficients tell you which factors "matter most"
- Modern models are harder to interpret
  - K-NN
  - Feed-forward neural nets
- What happens when we use these models for societally important decisions?
  - Credit scores
  - Recidivism scores (search for "compass recidivism controversy")
  - Deciding what news articles you see on Facebook
- Current controversy: what does it mean for models to be "transparent"?