COMS BC1016
Introduction to Computational Thinking and Data Science

Lecture 20: Regression Inference and Classification

Dec 01, 2025

The rest of the semester

- **Monday, Dec 1**: Regression Inference and Classification ← HW 8 due
- **Wednesday, Dec 3**: Computing Fellows Workshop
  - Final Project Consultations during Lab
- **Friday, Dec 5**: Progress Report due

- **Monday, Dec 8**: Special Topics (Data Privacy) ← HW 9 due

  Last day of class :(
- **Friday, Dec 12**: Final Projects Due

Final Project Progress Reports

- Good news: Due date moved from Tuesday to Friday

- Bad news: You will not be able to get your grades back on this portion before your final reports are due!

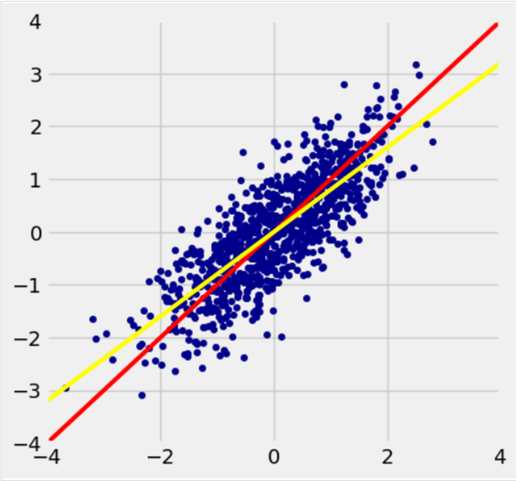
    - Please find your TAs or go to office hours if you would like specific feedback to incorporate into your reports

- Semi-related news: I’m traveling and not having office hours this week

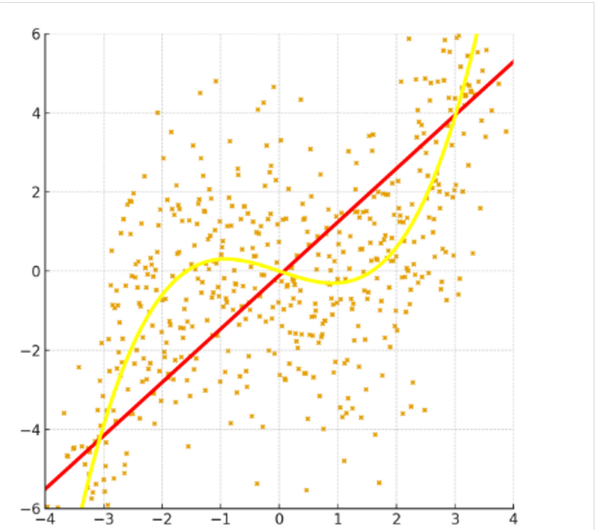
Last time: Least Squares and Residuals

Least Squares and Residuals

How can we know we've created the best line to fit through our data (i.e., that we've minimized error)?



How can we check whether a line is appropriate (versus a non-linear model)?



Least Squares Line

- **Minimizes** the **root mean squared error** (**rmse**) among all possible lines
  - Equivalently, minimizes the **mean square error** (**mse**) among all lines
- Other names for this line include:
  - “Best fit” line
  - Least squares line
  - Regression line

Computing the RMSE:

1. Compute the errors between the regression line and actual value and square them
2. Compute the mean of the squared errors
3. Compute the square root

Using `minimize` with mse to minimize errors

Suppose we have a dataset with x, y pairs.

If we define a function **`mse(a,b)`** to compute the mean square error of

$$\text{estimate of } y = a \times x + b$$

then `minimize(mse)` returns an array $[a_0, b_0]$

where a_0 is the **slope** and b_0 is the **intercept** that minimizes mse

Residuals

Residual: The error for *individual* regression estimates

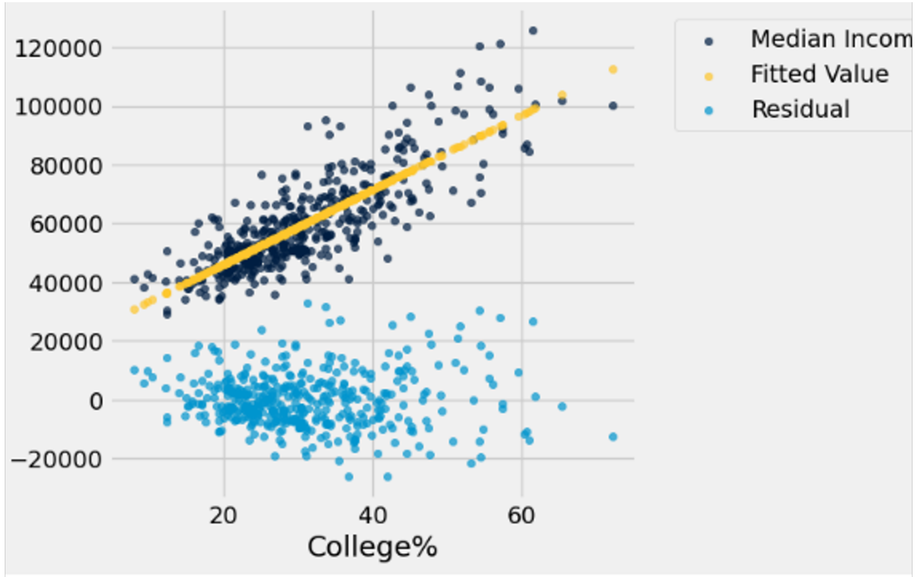
$$\text{residual} = \text{observed } y - \text{regression estimate of } y$$

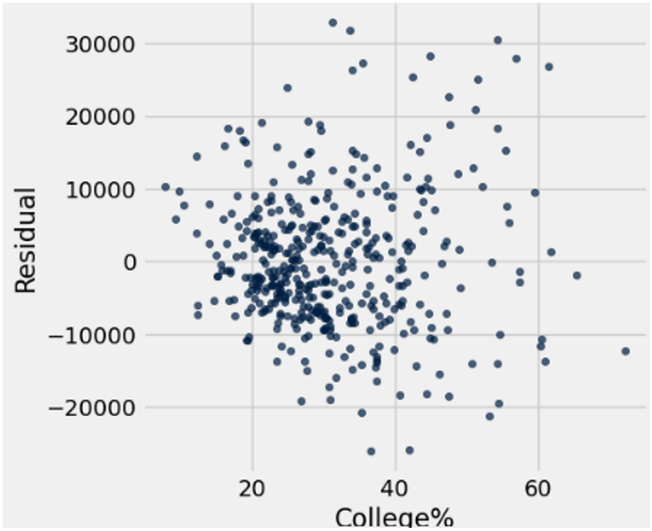
- Can calculate the residual for each individual (x, y) point
- It's the **vertical distance** between the point and the line of best fit

Residual Plots

Scatter diagrams of residuals:

- Should look like unassociated blobs for linear relationships
- Will show patterns for non-linear relationships
- Look for curves, trends, changes in spread, outliers, etc. as examples of non-linear patterns





Regression

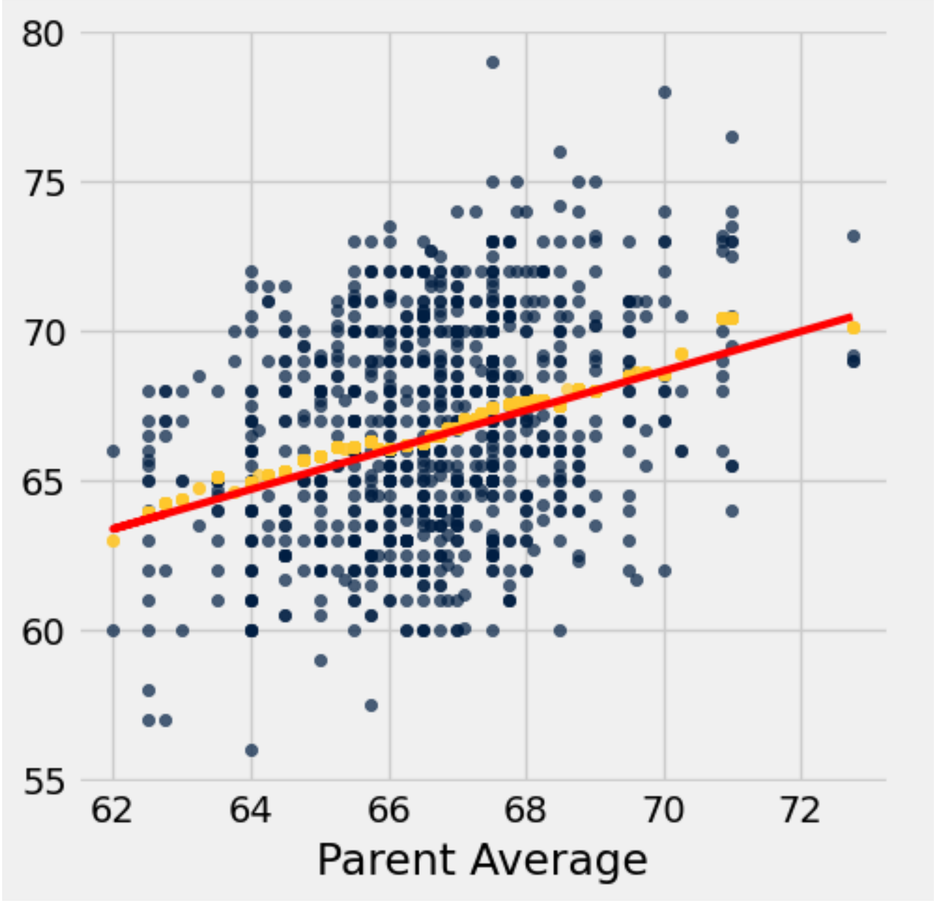
- **Regression** quantifies the **linear relationship** between two variables

- Process for checking linearity:

  - Plot our data in a scatter plot and observe whether it seems linear

  - If so, calculate a regression line

  - Look at **residual plot** to confirm whether linearity makes sense



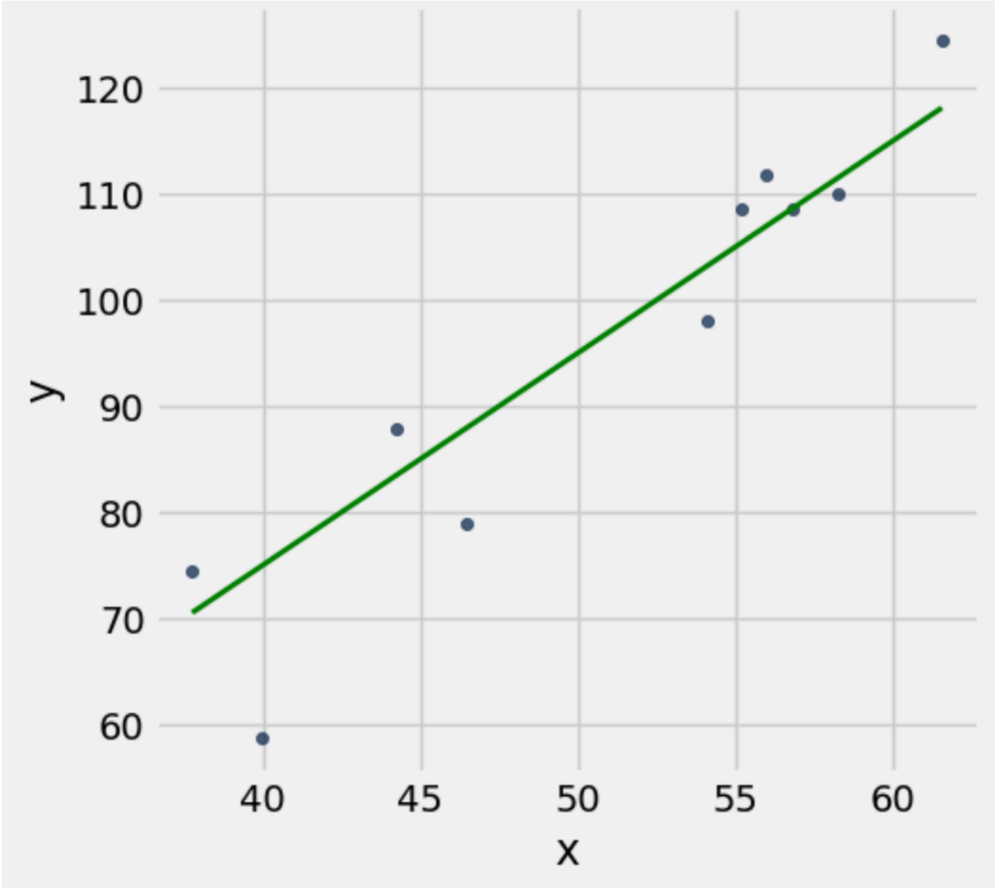
Regression Inference

Regression Inference: Premise

- Our data represents a sample of a larger population
    - The linear relationship (regression line) we determined is dependent on our sample

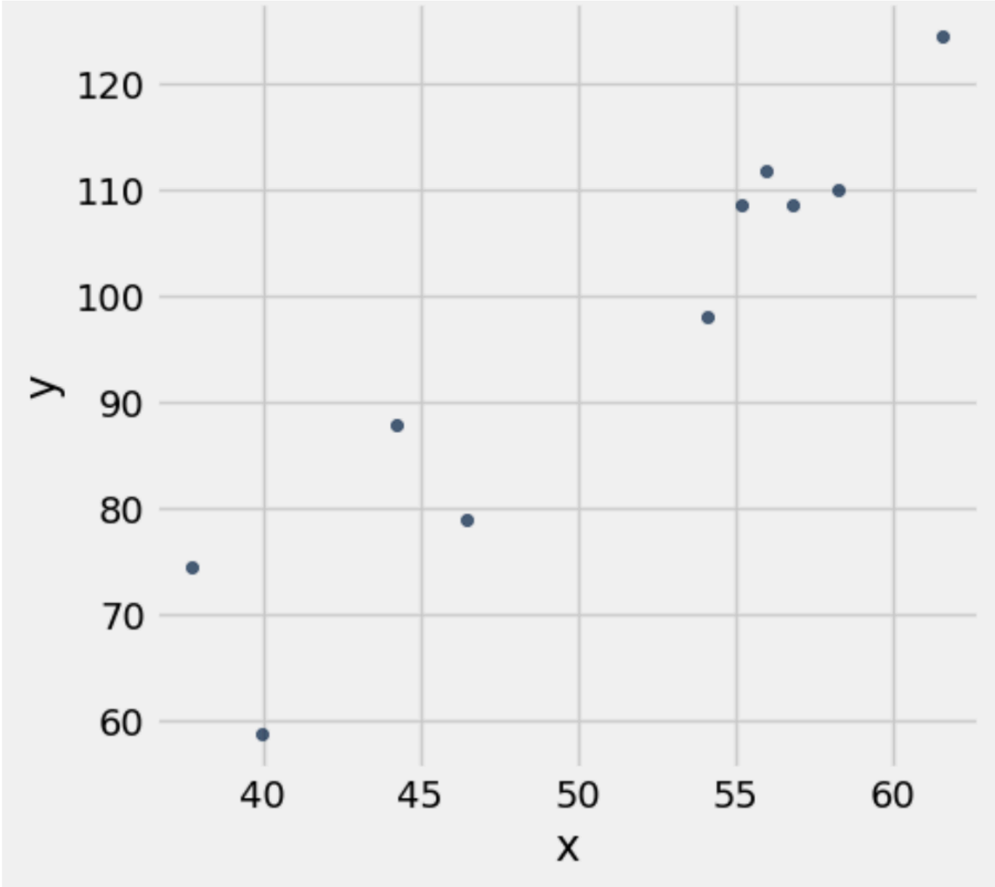
Regression Prediction

True line and 10 samples



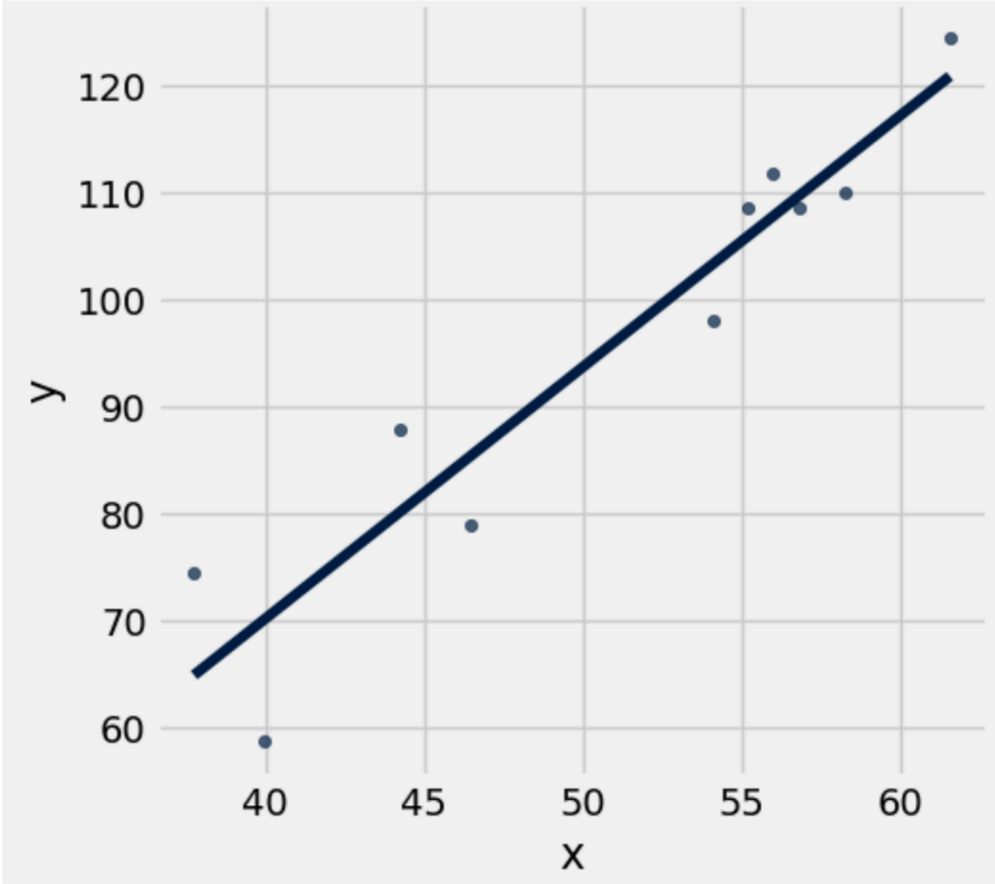
Regression Prediction

What we get to see



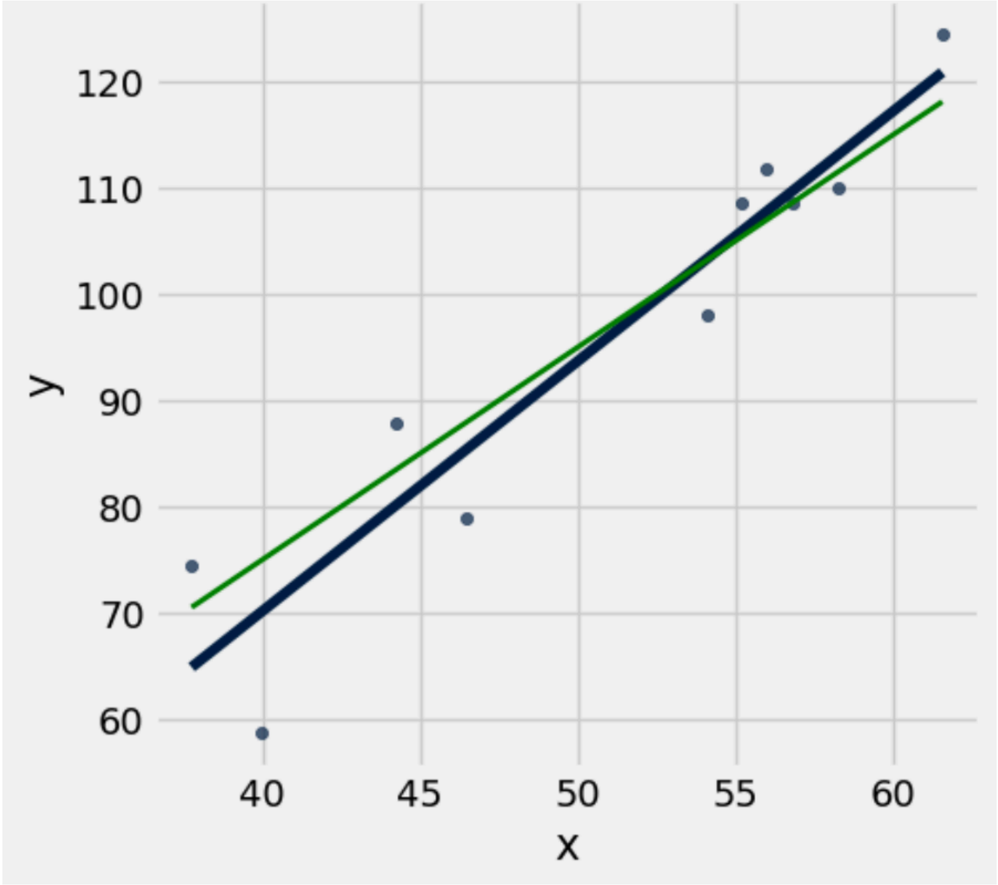
Regression Prediction

Regression Line: Estimate of the True Line



Regression Prediction

Regression Line and True Line



Regression Inference: Premise

- Our data represents a sample of a larger population
    - The linear relationship (regression line) we determined is dependent on our sample
- How confident are we in the regression line that we found?
    - **Estimate uncertainty** with a confidence interval for our **regression prediction**
    - We'll do this using our familiar bootstrap method

Regression Inference: Premise

- **Estimate uncertainty** with a confidence interval for our **regression prediction**

    1. Uncertainty around our **predicted value** for a given x-value

    2. Uncertainty around our regression line **slope**

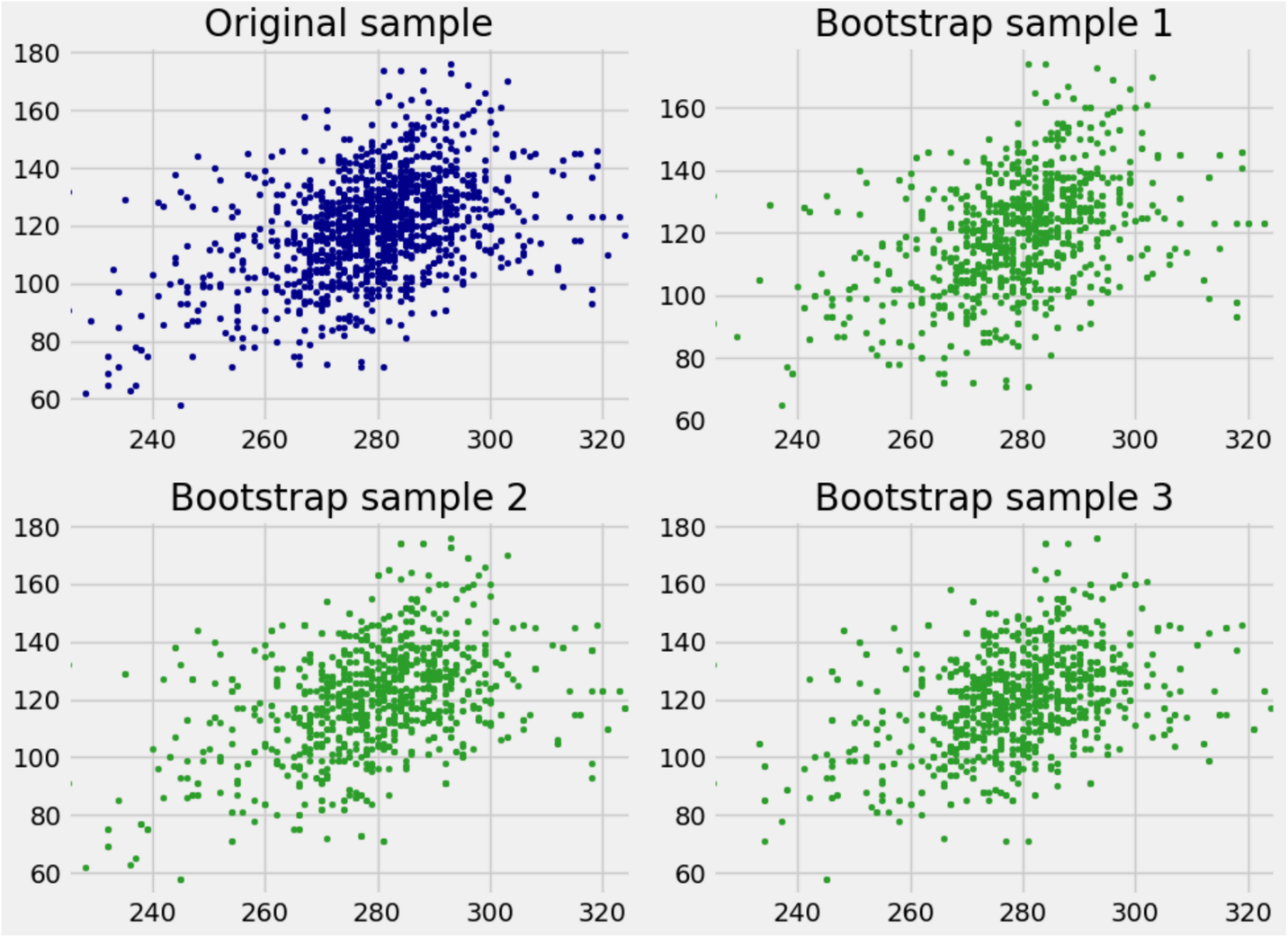
        - Do we think that the variables are linearly related?

Regression Prediction

- If the data in our sample comes from the regression model, the *true value* of the response y at a given value of x is the height of the *true line* at x
  - We don't know the true line because we don't know the entire population!
- The regression line is most likely close to the true line
  - Given a value of x, predict y by finding the point of the regression line at that x

Regression Inference: Bootstrapping

- Recall that bootstrapping is resampling *with replacement* from our original sample many times
  - Create scatter plot for resampled scatter plot
  - Determine regression line through each



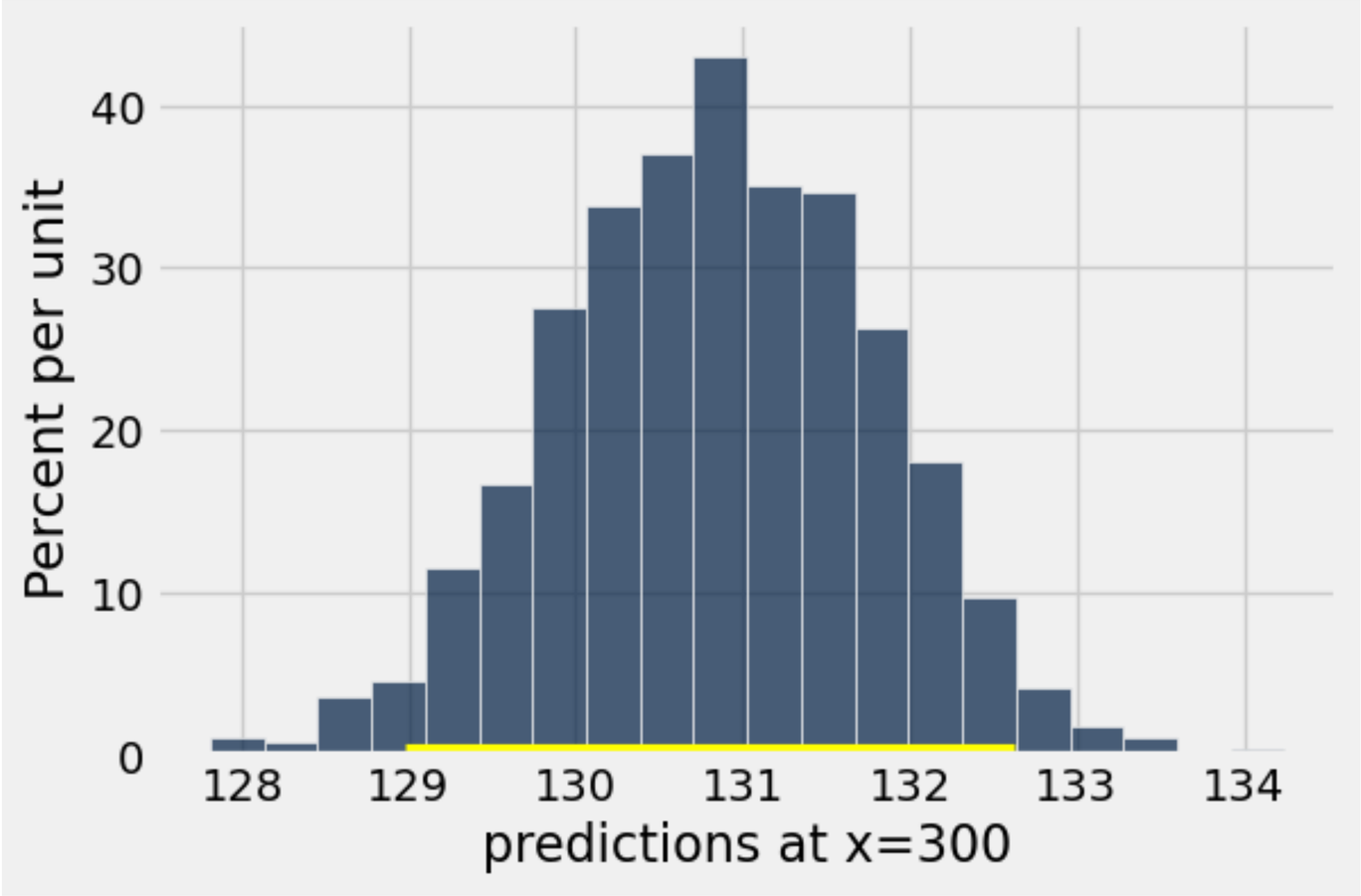
Regression Inference: Bootstrapping

To predict a value y at a given value of x,

we calculate the **predicted value** for each regression line

then draw a **histogram of these predicted values**

and construct our confidence interval



Determining the Prediction Interval

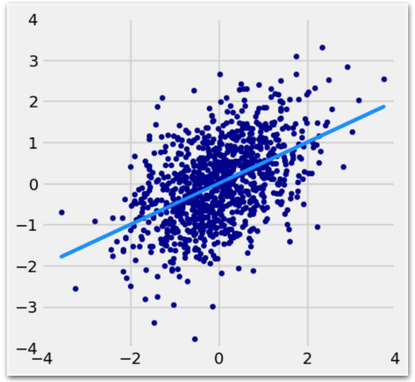
- Bootstrap the scatter plot
- Calculate the regression line for that scatter plot
- Get a prediction for y using the regression line
  - Repeat many times to get a histogram of values for y
- Get the middle 95% interval for a 95% confidence interval
- Interpretation:
  - **95% confidence interval for the height of the true line at x**

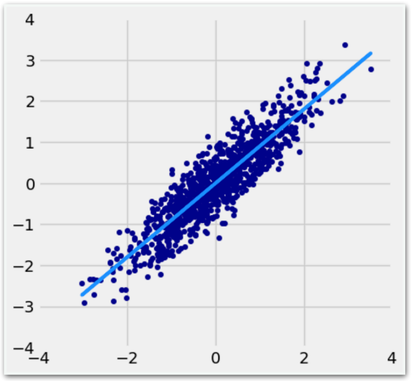
Confidence Interval for Slope

Same process as we just did for confidence interval of prediction of y at value of x, but instead create a **histogram of slopes**

Confidence Interval for Slope

Same process as we just did for confidence interval of prediction of y at value of x, but instead create a **histogram of slopes**



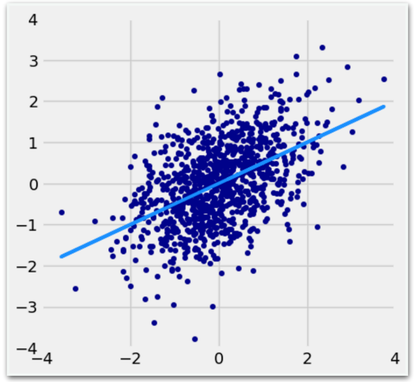


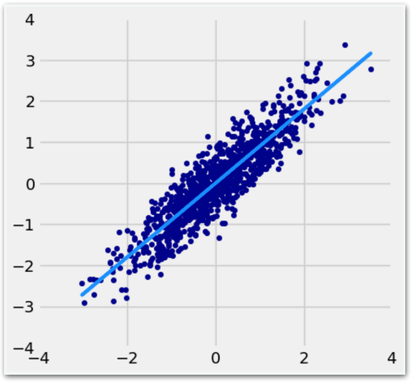
Recall what we know about slope:

- Metric for strength of the linear relationship!

Confidence Interval for Slope

Same process as we just did for confidence interval of prediction of y at value of x, but instead create a **histogram of slopes**

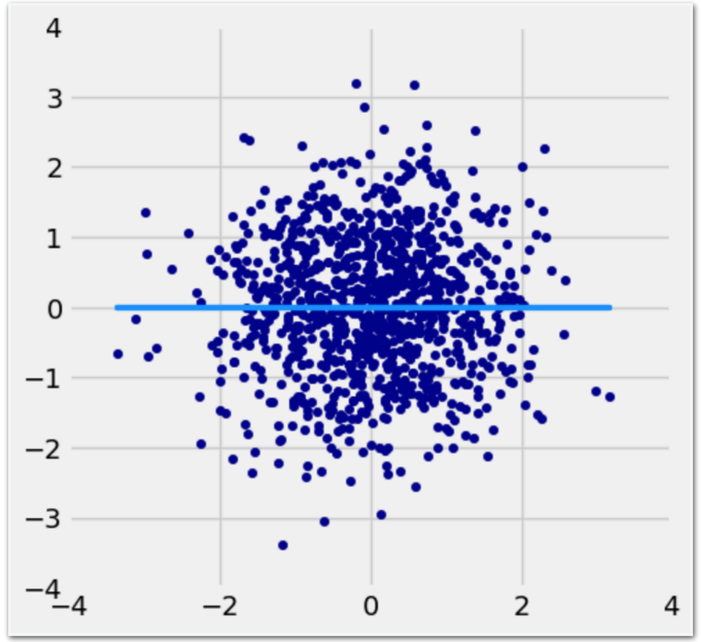




Recall what we know about slope:

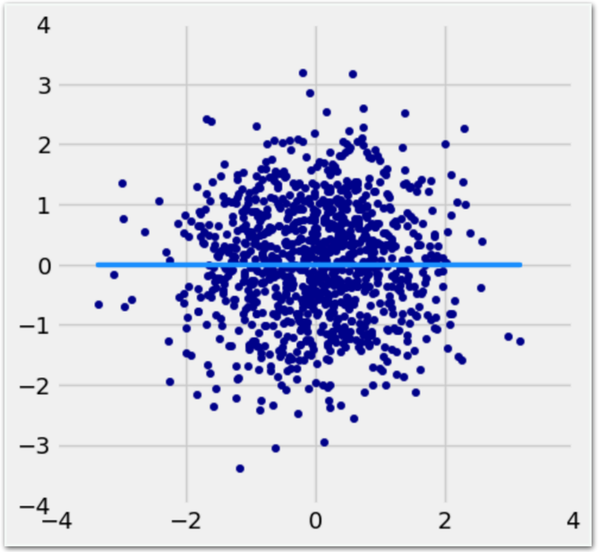
- Metric for strength of the linear relationship!

- *When the slope is zero, there's no linear association*

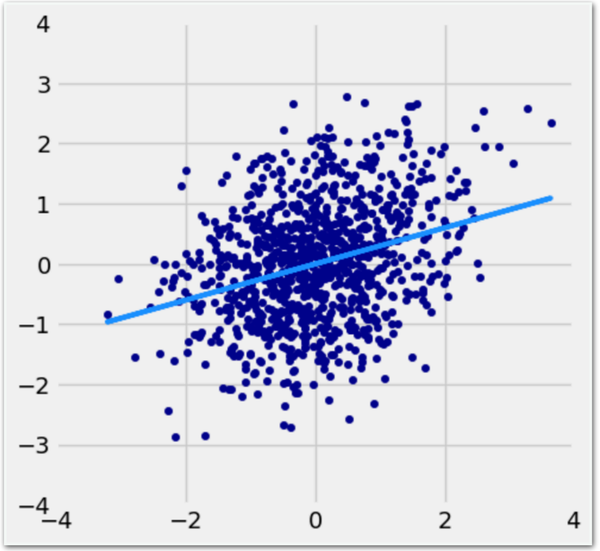


Hypothesis Testing Slope

Null Hypothesis:

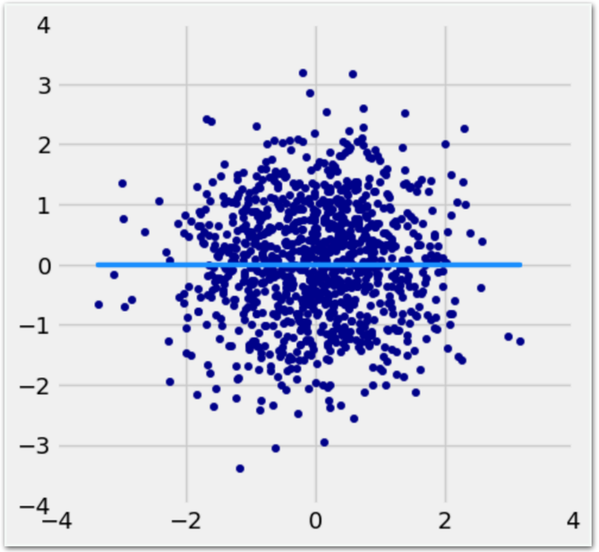


Alternative Hypothesis:

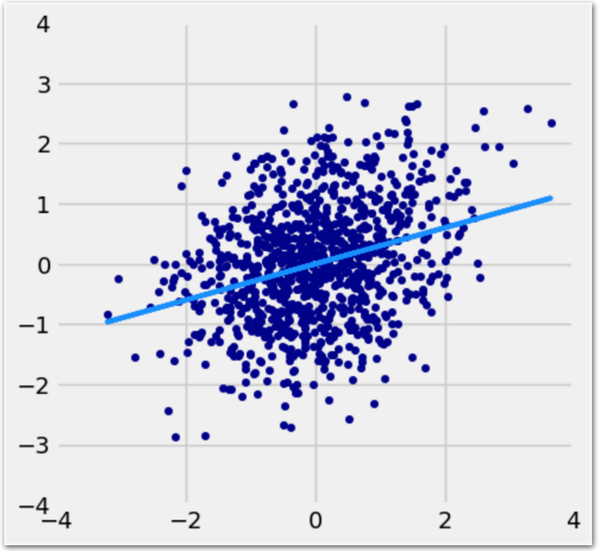


Hypothesis Testing Slope

Null Hypothesis: The slope of the true line is 0 (the variables are *not* linearly correlated)



Alternative Hypothesis: The slope of the true line is non-zero



Hypothesis Testing Slope

Null Hypothesis: The slope of the true line is 0 (the variables are not linearly correlated)

Alternative Hypothesis: The slope of the true line is non-zero

- Process:
  - Construct a bootstrap confidence interval for the true slope
  - If the interval doesn't contain 0, **reject null**
  - If the interval does contain 0, **there isn't enough evidence to reject the null**

Classification

Machine Learning

- **Machine learning** is a class of techniques for automatically finding patterns in data and using it to draw inferences or make predictions

- **Machine learning algorithms** are mathematical models that are calculated based on sample data (“training data”) to make predictions / decisions without being explicitly programmed to perform the task

Does this sound familiar?

Predicting Values

- Based on incomplete information

- One way to make predictions:

  - To predict an outcome for an individual

    Find others who are like that individual
    and whose outcomes you know

    Use those outcomes as the basis of
    your prediction

Predicting Values

- Based on incomplete information
- One way to make predictions:
  - To predict an outcome for an individual

    Find others who are like that individual and whose outcomes you know

    Use those outcomes as the basis of your prediction

<u>Two types of predictions:</u>

- **Regression**: Numeric
- **Classification**: Categorical

Classification

- Our data has different attributes (e.g., columns in our tables) and we refer to the attribute we want to predict as the **class**

Spam or not?

Angell Animal Medic. Invoice 10/16/2024 - This is an automated notification. Please do not reply to this message.

Your Free $4,000 B. **>>eysa.lee!! - eysa.lee -Bonus Code: Use BETMGM**

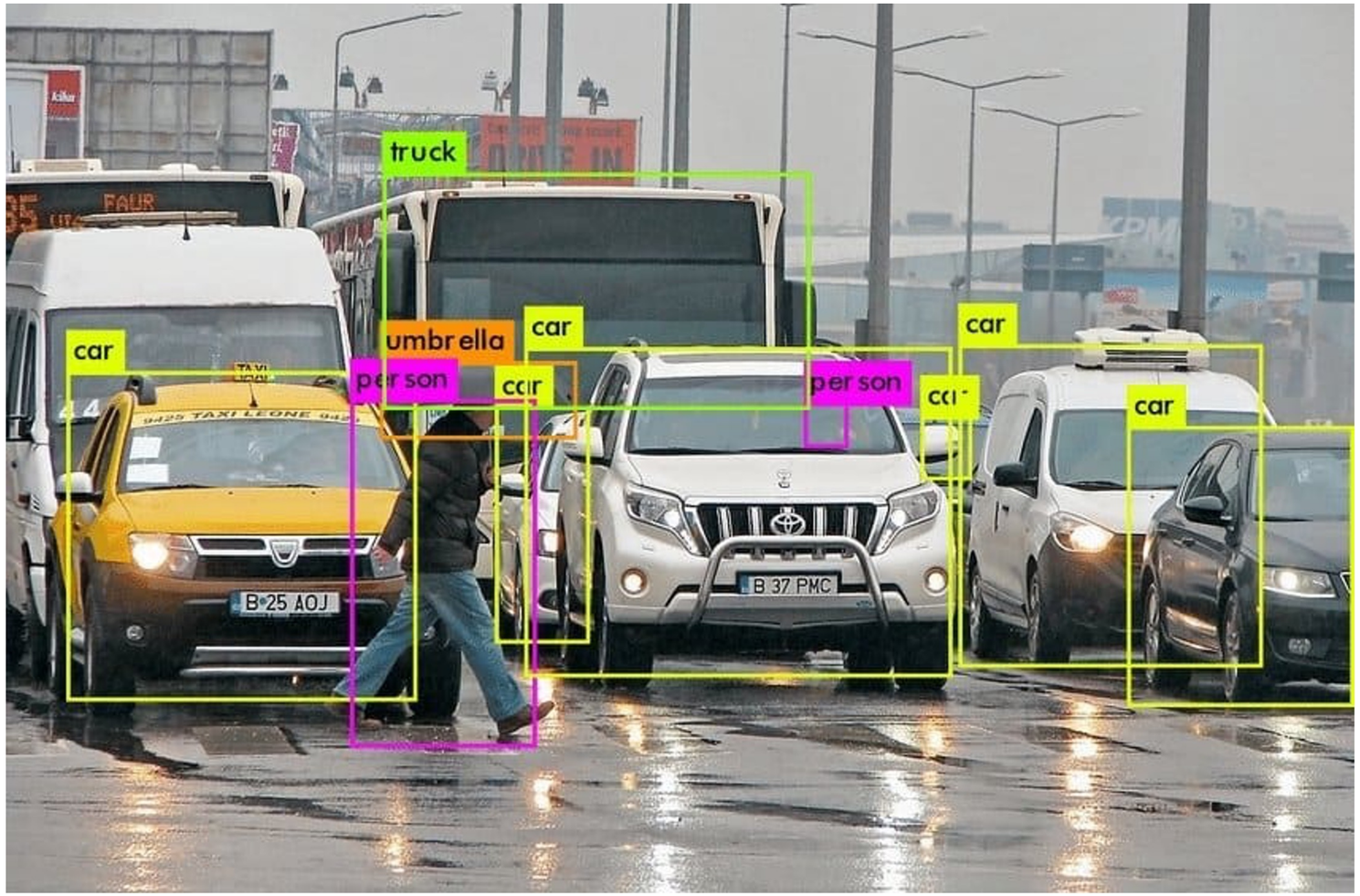
me, Riverside 3 Cat is constipated? - Good Morni...

no-reply Administrator has responded to your request for 'Internal Reports - Classified' - Good news. You now have access to 'Intern...

mailings [IACR] 2025 Election - Dear Eysa Lee, Dear IACR Member, The 2025 Election for Board positions is now open.

idontknowyou...@gmail.c... (no subject) - Hello, Please could you drop a contact to text you on, Thank you. Laura Rosenbury President Barnard College

Object Classification



https://viso.ai/computer-vision/image-recognition/

Classification

- Our data has different attributes (e.g., columns in our tables) and we refer to the attribute we want to predict as the **class**

- To predict, we look for patterns in the existing data

  - The data we're using to look for patterns we call **training data**

- To check accuracy of our predictions, we'll use a separate set of data we call **testing data**

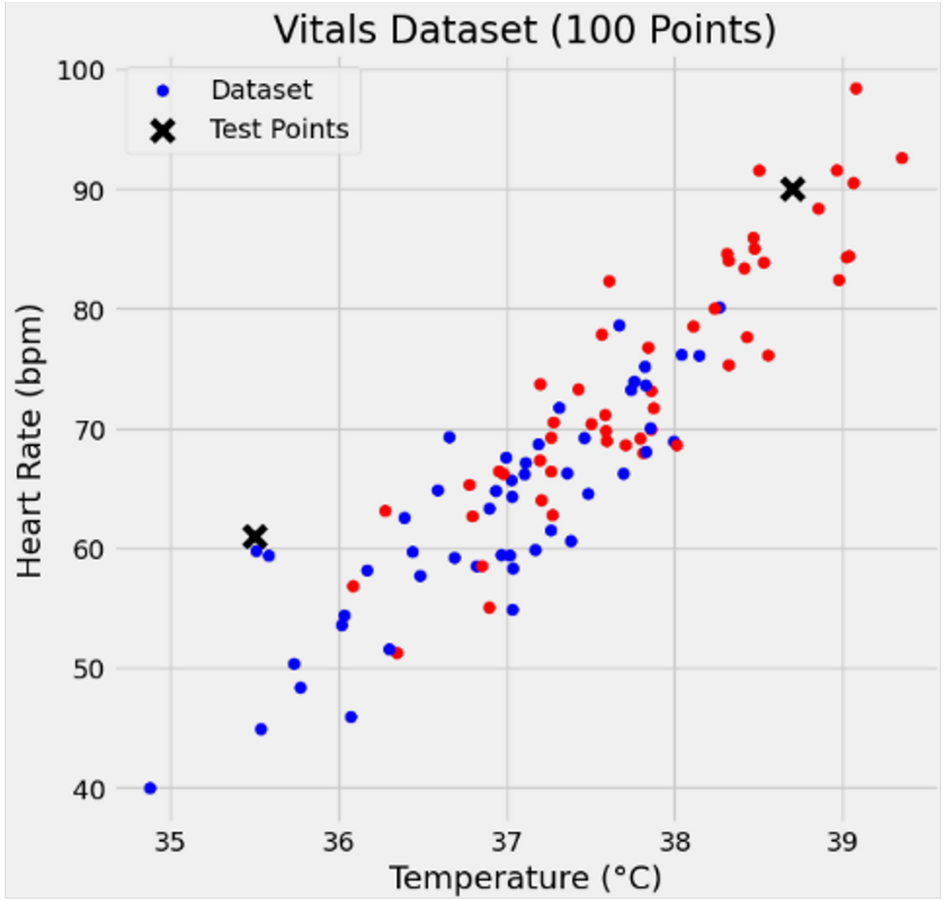
  - Though we'd like our predictions to always be correct, it can still be useful even if it doesn't predict correctly 100% of the time

Classifier Example: Disease or not

Temperature	Heart Rate	Disease
39.5	120	1
38.5	110	1
36.8	78	0
37	150	0
38	92	0
36.5	70	0
39	115	1
37.2	80	0

Classifier Example: Disease or not

Temperature	Heart Rate	Disease
39.5	120	1
38.5	110	1
36.8	78	0
37	150	0
38	92	0
36.5	70	0
39	115	1
37.2	80	0



Distance between points

- Distance between two points with attributes x and y:

$$D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}$$

- Distance between points with attributes x, y, z:

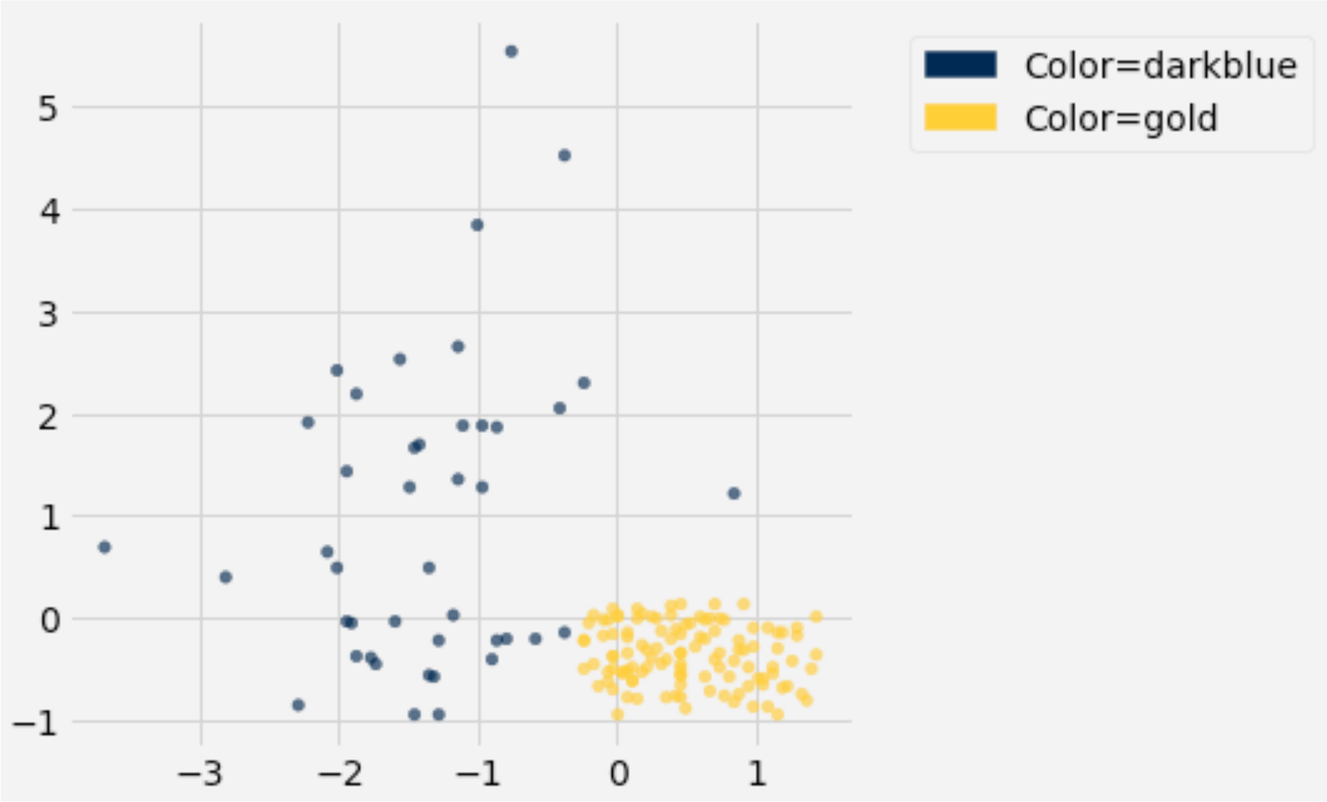
$$D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2 + (z_0 - z_1)^2}$$

Nearest Neighbor Classification

To classify a point:

Find the *nearest* point to it (i.e., shortest distance)

Assign the label of the nearest point



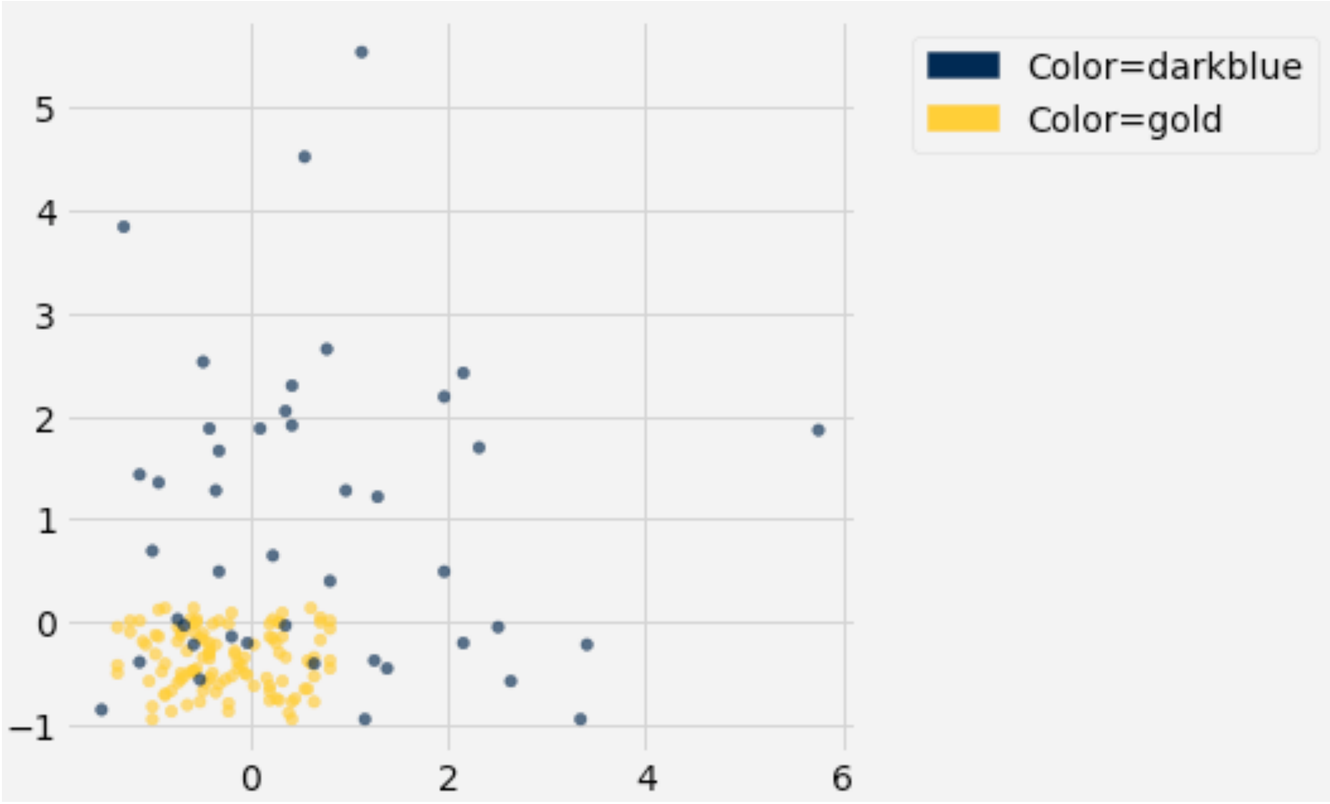
k-Nearest Neighbor Classification

To classify a point:

Find the closest k neighbors

Find the *majority* label among those neighbors

Assign the majority label to the new point



Accuracy of a classifier

- Accuracy of a classifier on a labeled data set is the proportion of examples that are labeled correctly

- Need to compare classifier predictions to true labels

- If the labeled data set is sampled at random from the population, then we can infer accuracy on that population

Evaluation Metrics

- Suppose a model performs 95% accuracy on a test set. Is this a good performance?

- Depends!

  - If 90% of the population is Group A, then 95% accuracy isn't that much better than labeling everyone Group A

  - If 50% of the population is Group A, then 95% accuracy is pretty good

- Evaluation depends on contextualizing the performance with the baseline

Evaluation: Confusion Matrix

	Actually Positive (1)	**Actually Negative (0)**
Predicted Positive (1)	True Positive 😌	False Positive ☹️
Predicted Negative (0)	False Negative ☹️	True Negative 😌

Extra Topic:
Common Errors with Data

Errors with Data Overview

- Up to this point in the course, datasets have been provided and we have assumed the data is “good” / representative
    - But in reality, there are many cases in which there are issues and errors with data
    - Knowing how to look out for them is critical to responsible data science!
- Side note: A large part of data science is also involves “cleaning” your data (as some of you have experienced with the final project)
    - Apart from errors, real-world data isn’t always as pretty as what we’ve mostly worked with in class!

Errors with Data Overview

1. Errors with data collection

2. Errors with interpreting results

Errors with Data Collection

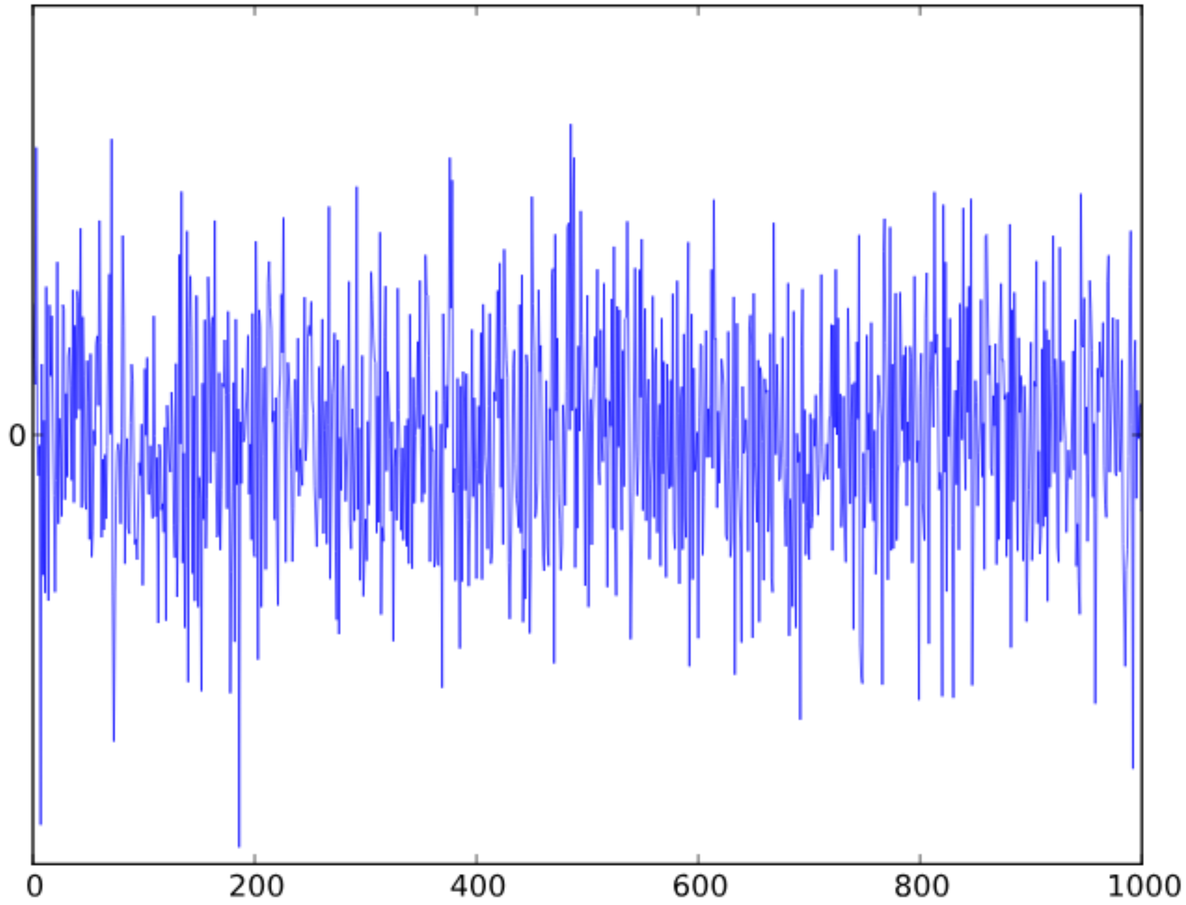
Random Errors

Most forms of data collection have random errors that in theory should cancel themselves out with large enough datasets

- Manual data entry errors

- Noise with data collection probes

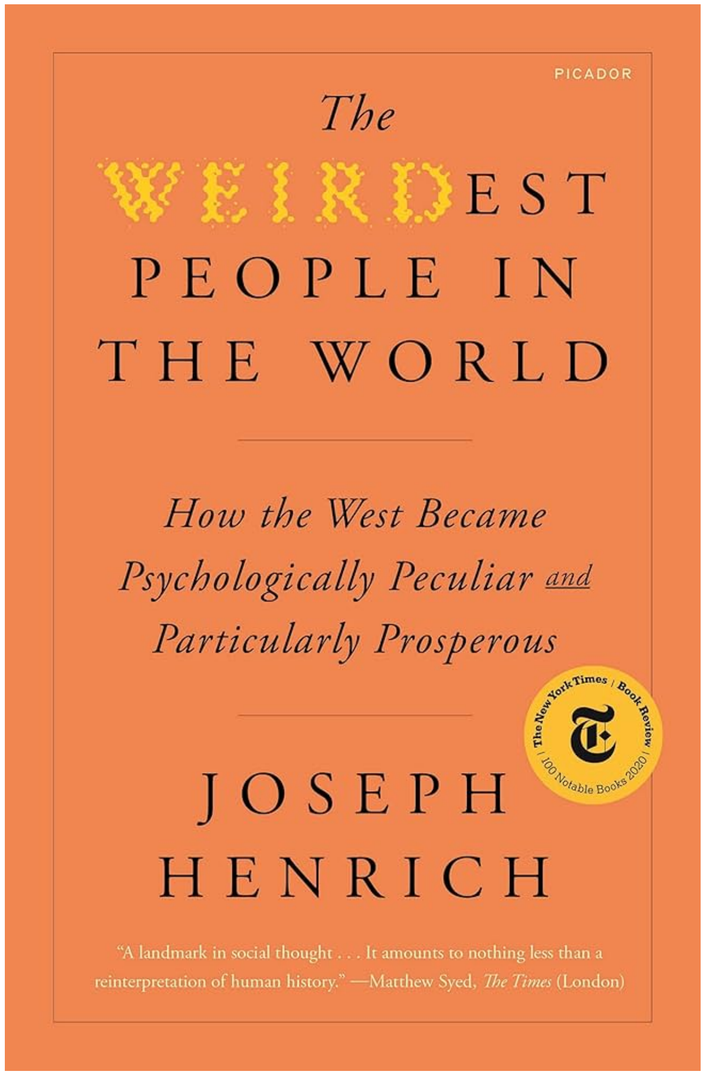
However, we will be discussing non-random errors



Selection Bias

Errors due to non-random / unrepresentative selection

Example: WEIRD (**W**estern, **E**ducated, **I**ndustrialized, **R**ich, and **D**emocratic) studies on Western, college-age students



Selection Bias

- Meta-analysis [1] revealed that between 2003-2007, undergrads made up 80% of study subjects in 6 top psychology journals

- 96% of psychology samples come from countries making up only 12% of the world's population

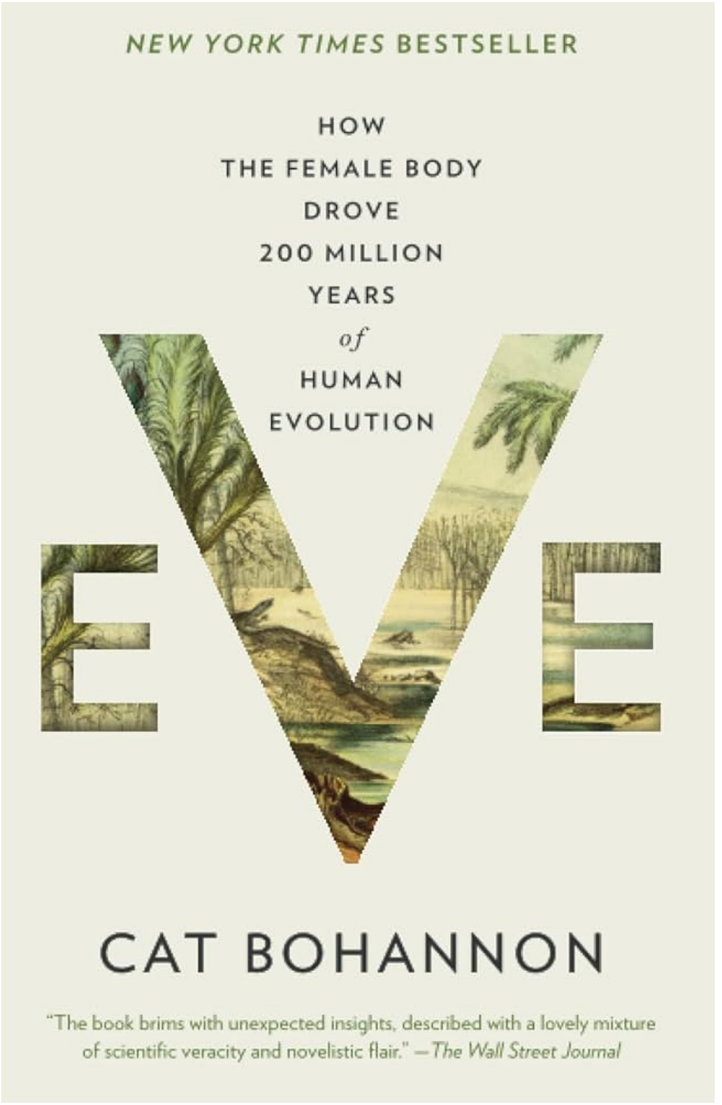
- Issue because **WEIRD** societies tend to have certain characteristics that may not be representative such as:

  - More individualistic, more overconfident, etc.

[1] Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. Behavioral and brain sciences, 33(2-3), 61-83.

Selection Bias

- Clinical drug trials have traditionally not included women of child-bearing age
    - 1977: FDA created a policy to exclude women from phase 1 and phase 2 clinical trials
    - Only in 1986 and 1993 they revisited this and congress passed a law to require inclusion of women in clinical research
    - Example implications: woman bodies were found to handle anesthesia differently than men (tend to come out of medication *faster*)



Selection Bias: Right-Censored Data

- **Right-Censored Data**: When data at the upper bound is excluded because it hasn’t “expired” yet

Selection Bias: Right-Censored Data

- **Right-Censored Data**: When data at the upper bound is excluded because it hasn't "expired" yet

- Example: Average age of death

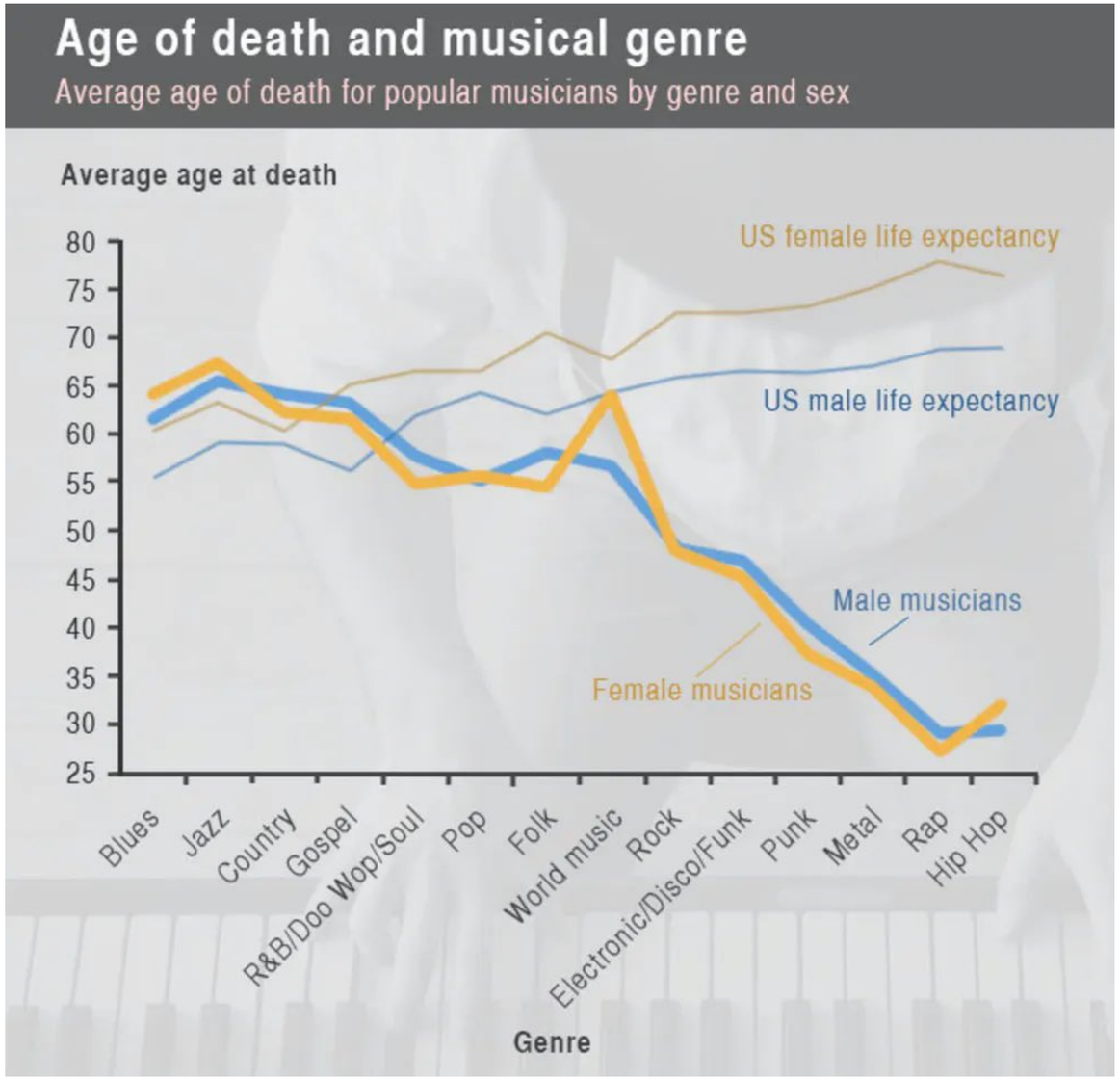
  - Data that deals with age of death includes those who die young but *not* those who are still alive

  - May be skewed towards the lower end (left)

Selection Bias: Right-Censored Data

Age of death and musical genre:

- Rap and Hip-Hop are younger genres (~40 years old)
    - The only “data” we have is on people who died young (other artists are still living)
- Since we are only displaying those who have died, the average is not representative (right-censored)



Life expectancy data from: http://demog.berkeley.edu/~andrew/1918/figure2.html

theconversation.com Source: Author

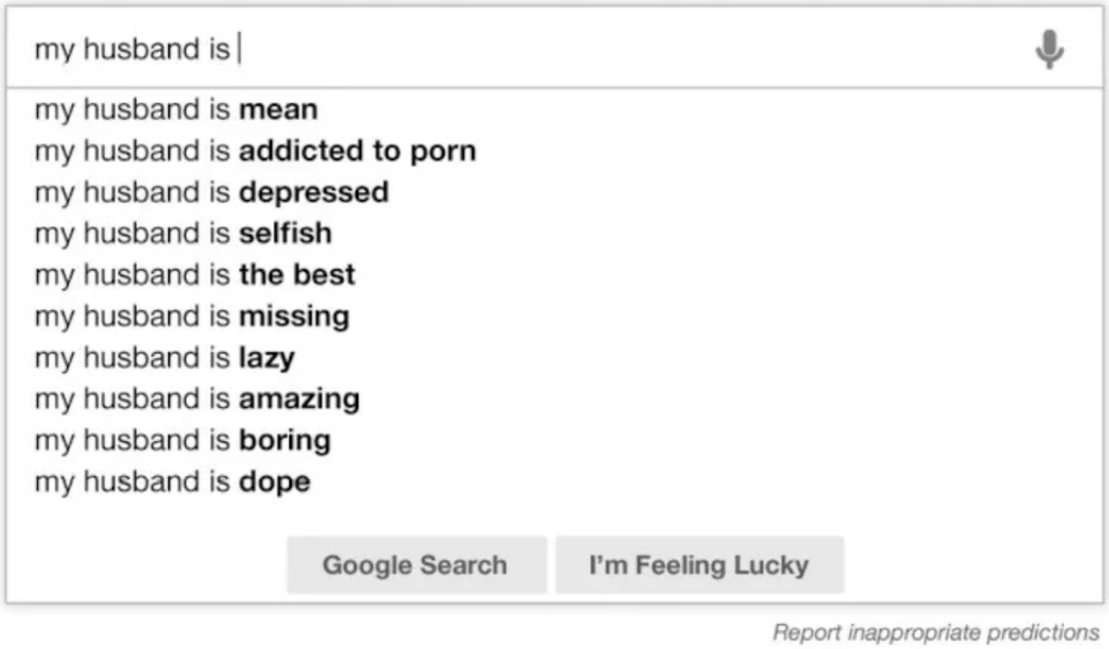
Selection Bias: Where Data is Collected

Facebook Search

Selection Bias: Where Data is Collected

Facebook Search

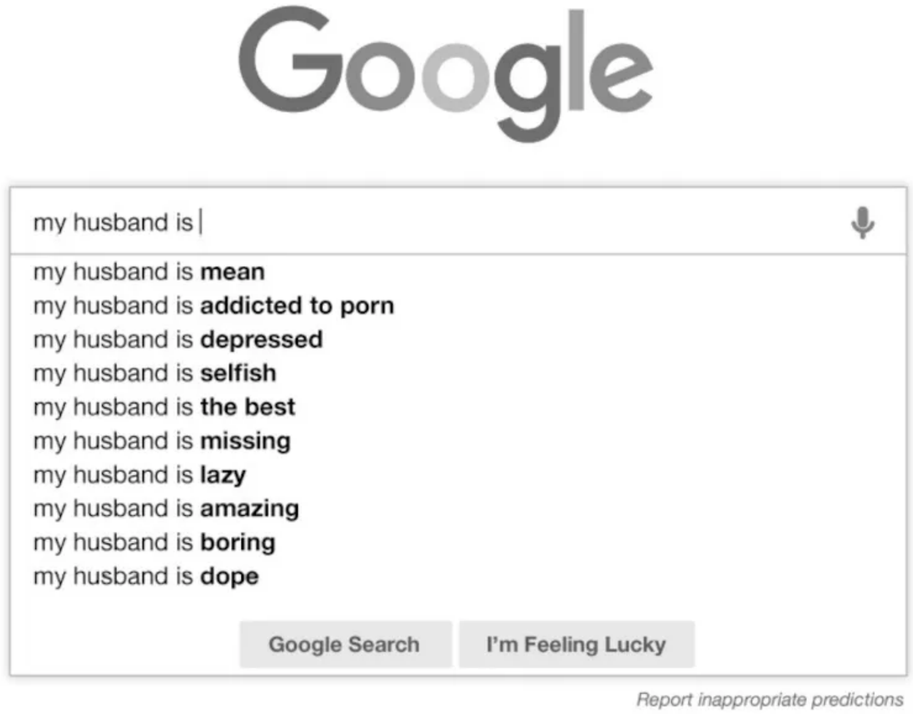
Google Search



Selection Bias: Where Data is Collected

Search queries are dependent on what platform the search is carried out on

- Social media: people share positive content
- Search engines: people look for help

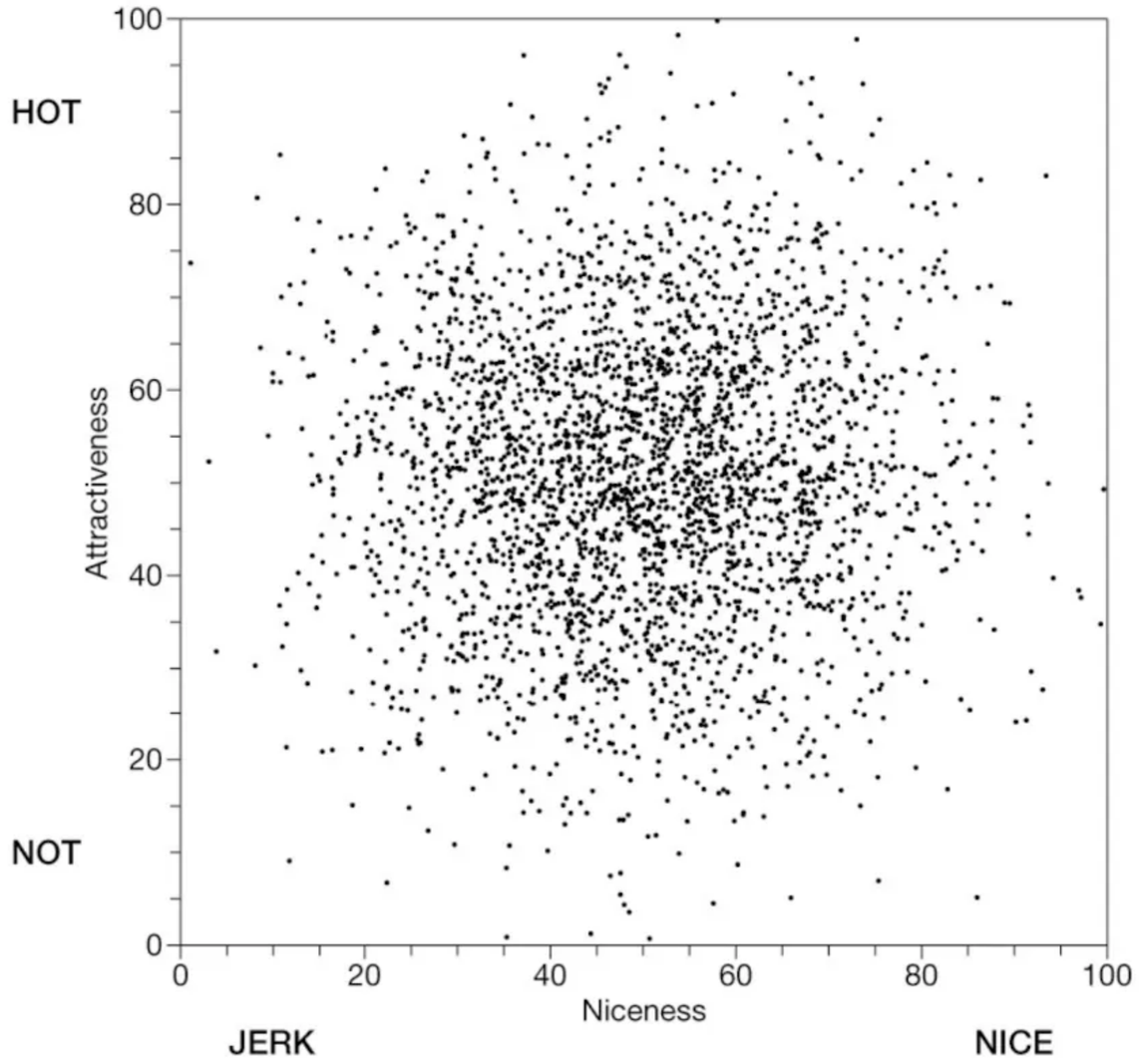


Bergstrom, C. T., & West, J. D. (2021). Calling bullshit: The art of skepticism in a data-driven world. Random House.

Selection Bias: Berkson's Paradox

- Two values can seem negatively correlated even when they're uncorrelated in the population
    - We systematically observe some events more than others
- Let's consider as an example attractiveness and niceness
    - What if you were to judge people based on your dating pool?

Example: Attractiveness vs Niceness

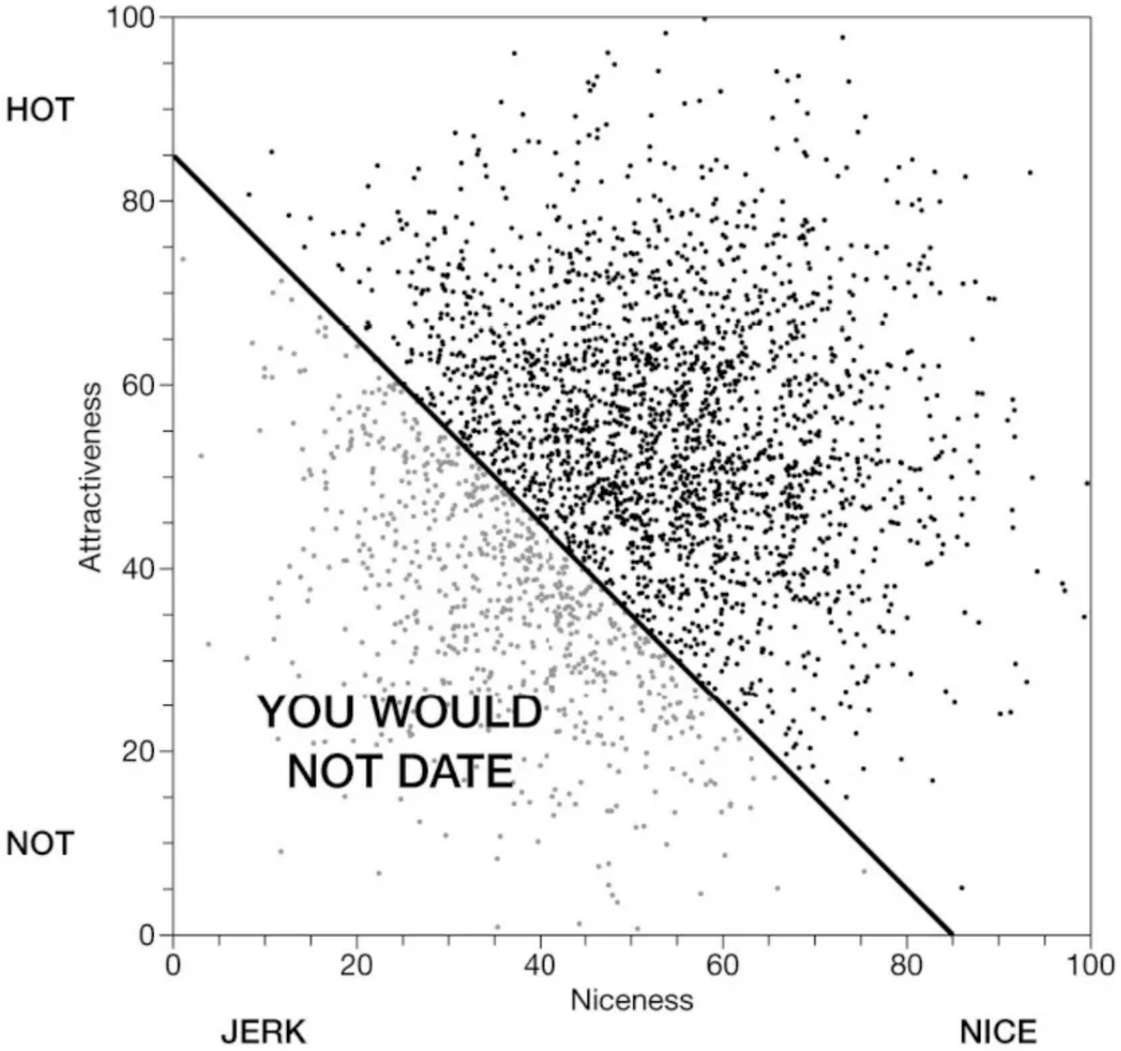


$r = 0.03$

Ellenberg, J. (2015). How not to be wrong: The power of mathematical thinking. Penguin. Chicago

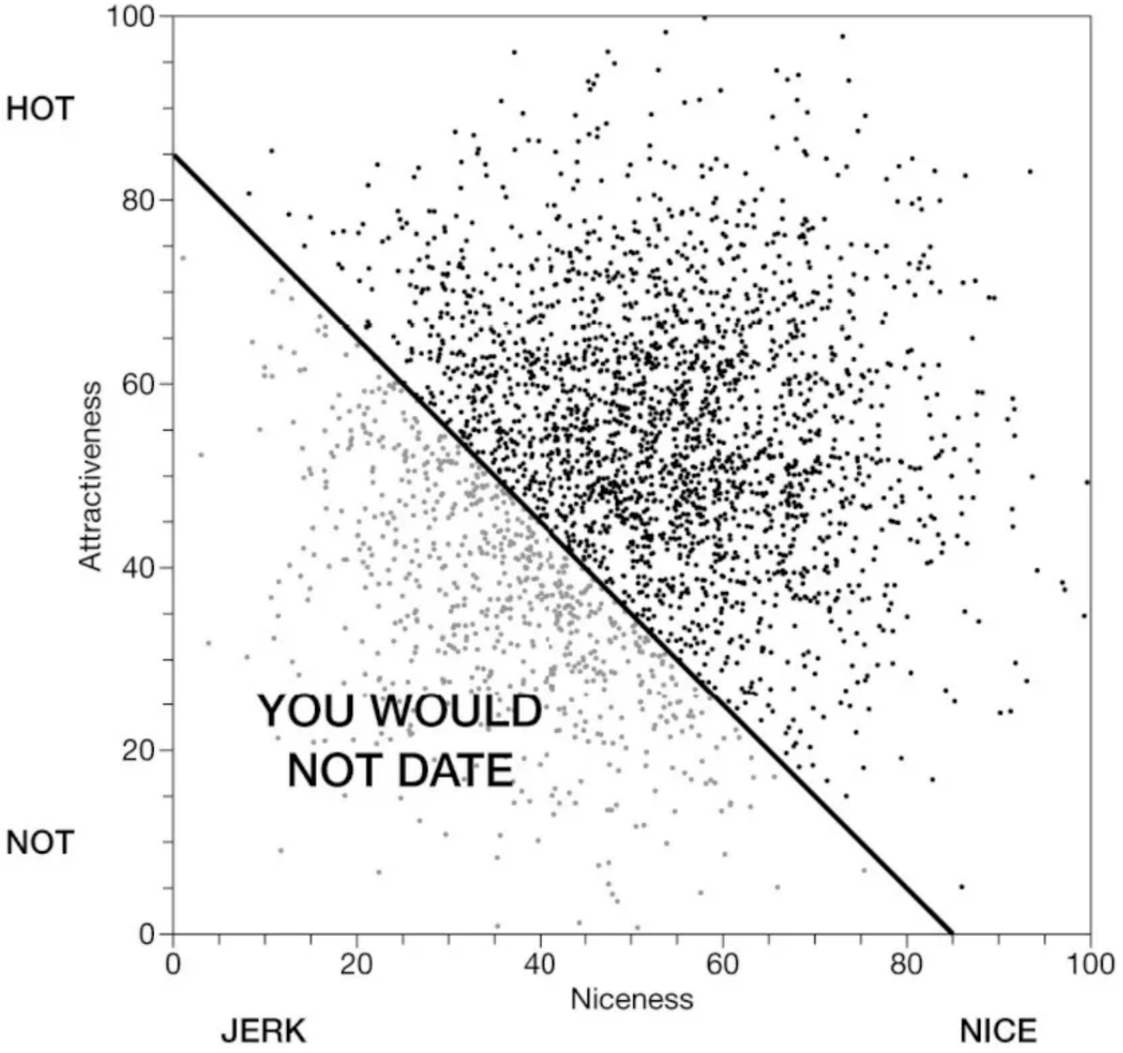
Example: Attractiveness vs Niceness

$r = 0.26$

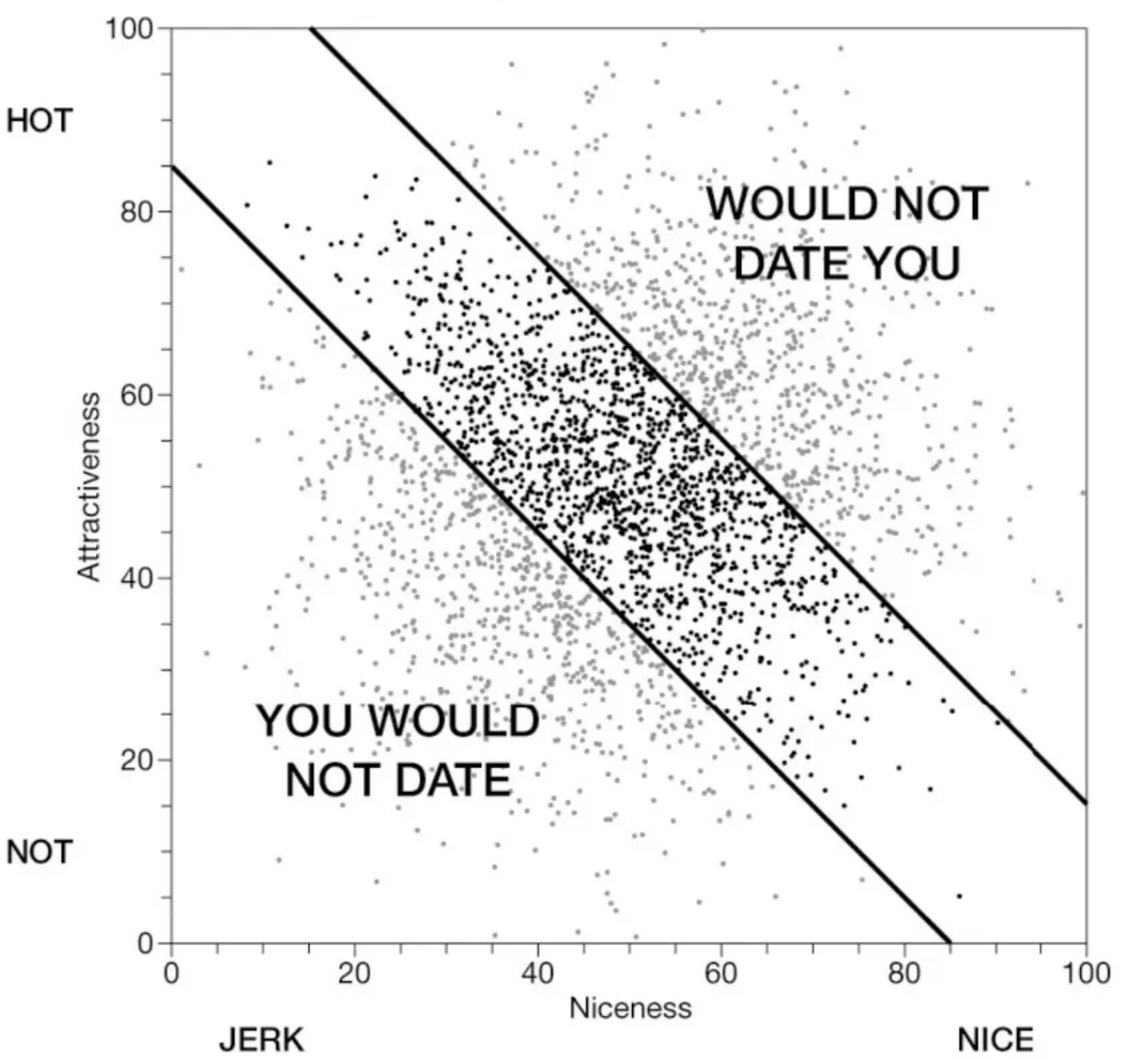


Example: Attractiveness vs Niceness

$r = 0.26$



$r = 0.75$



Selection Bias: Social Desirability Bias

- When respondents' answers align with what they *think* the researcher wants to hear
    - Example: Dentist asking you how many times you floss a week, your professor asks when you started on your homework

Errors with Interpreting Results

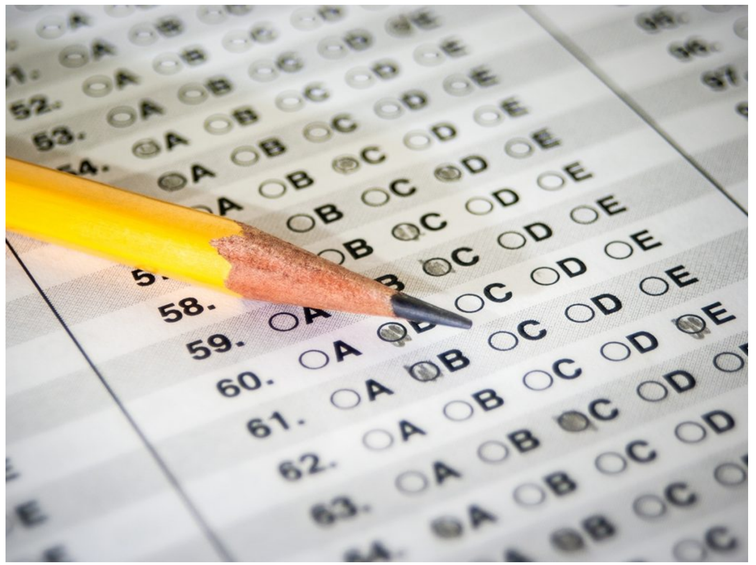
Validity

When we fail to measure what we think we're measuring

Example: Standardized Tests

- SAT scores tend to be correlated with academic achievement at post-secondary level

- But are we measuring students' abilities, or other underlying factors (e.g., access to test prep materials, etc.)?



Validity

Example: Marshmallow Test & Delayed Gratification

- Original Study: Stanford 1960s, 32 children
- Delayed gratification associated with better life outcomes (SAT scores, educational attainment, BMI)
- However… Many of the original study participants were students of Stanford students or professors

https://www.yalemedicine.org/news/why-you-shouldnt-rely-on-bmi-alone

Validity

Replicated by NYU study in 2018 with 900 children (diverse backgrounds)

- Outcomes were weaker than previously found

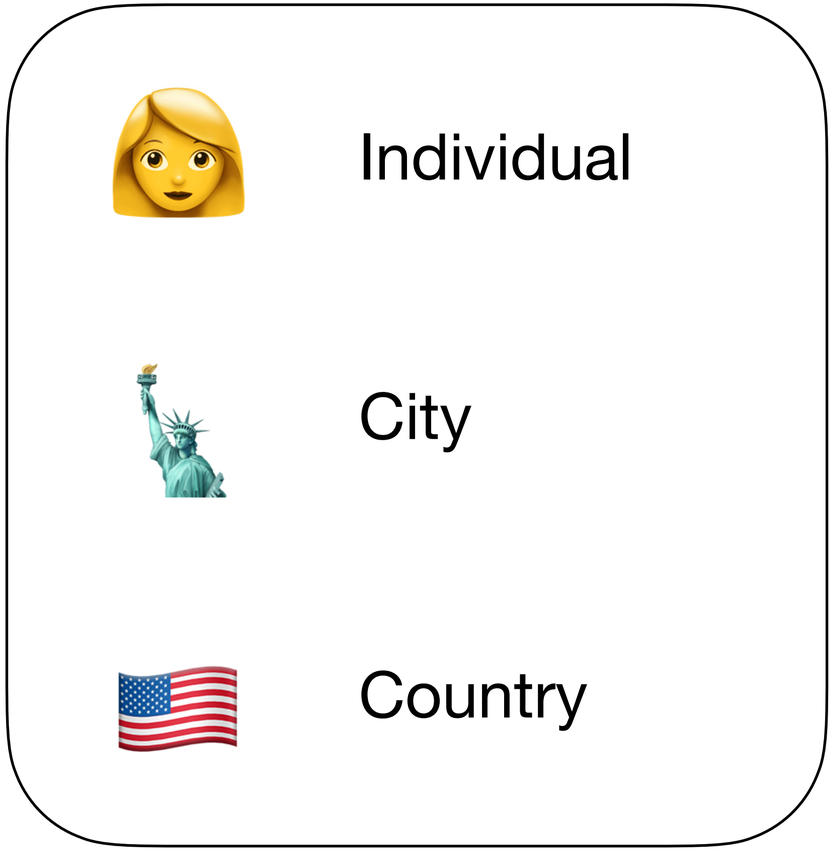
- … and disappeared when accounting for family's socioeconomic status and home environment

  - Children from more affluent families more likely to display delayed gratification

Watts, T. W., Duncan, G. J., & Quan, H. (2018). Revisiting the marshmallow test: A conceptual replication investigating links between early delay of gratification and later outcomes. Psychological science, 29(7), 1159-1177.

Ecological Fallacy

Data collected at a given level (e.g., average per country) is interpreted at a different level (e.g., individuals within a country)

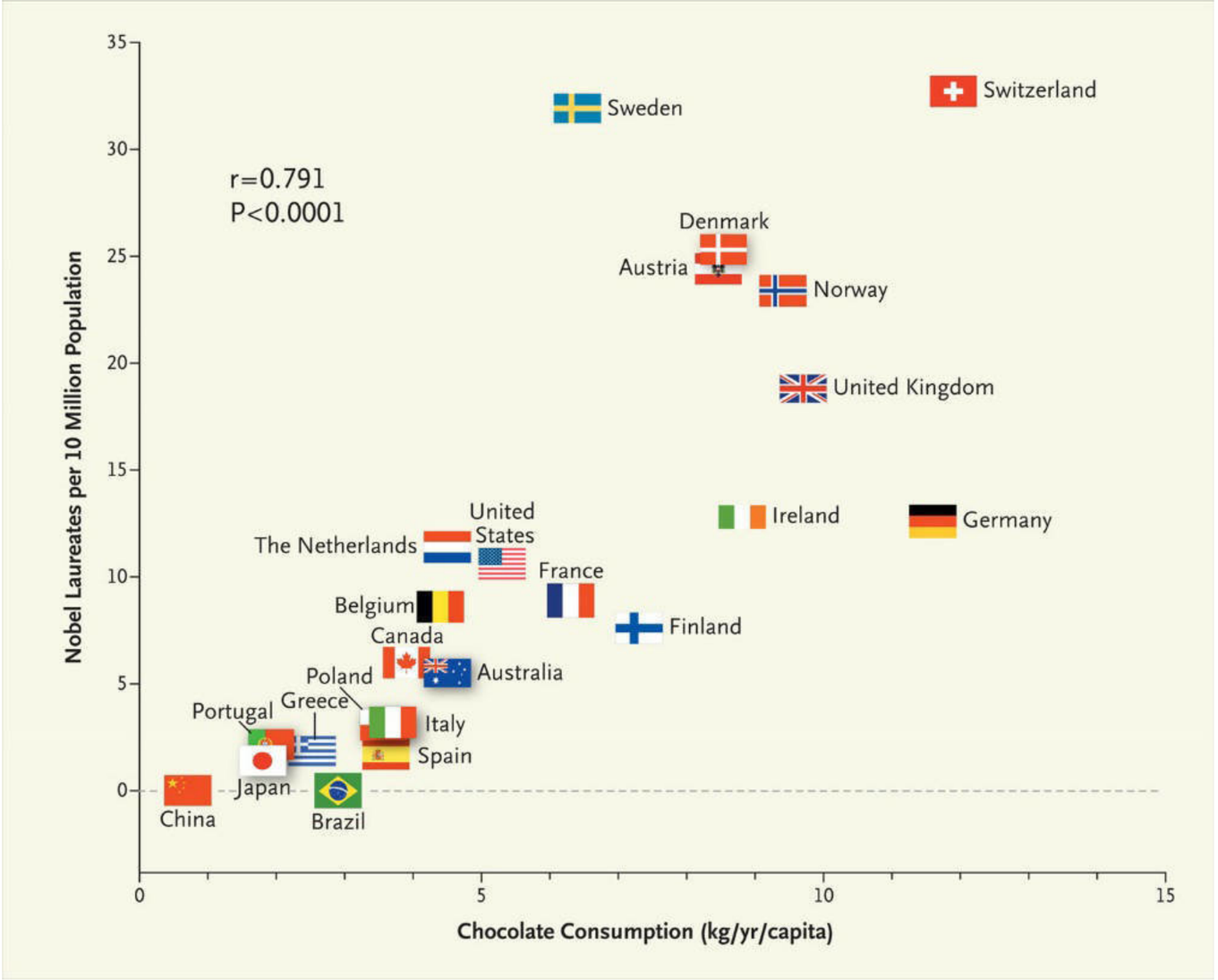


Ecological Fallacy

Example: Chocolate consumption and Nobel Laureates

- Chocolate consumption is average consumption within a country

- We know nothing about Nobel laureates and their individual consumption



Messerli, F. H. (2012). Chocolate consumption, cognitive function, and Nobel laureates. N Engl J Med, 367(16), 1562-4.

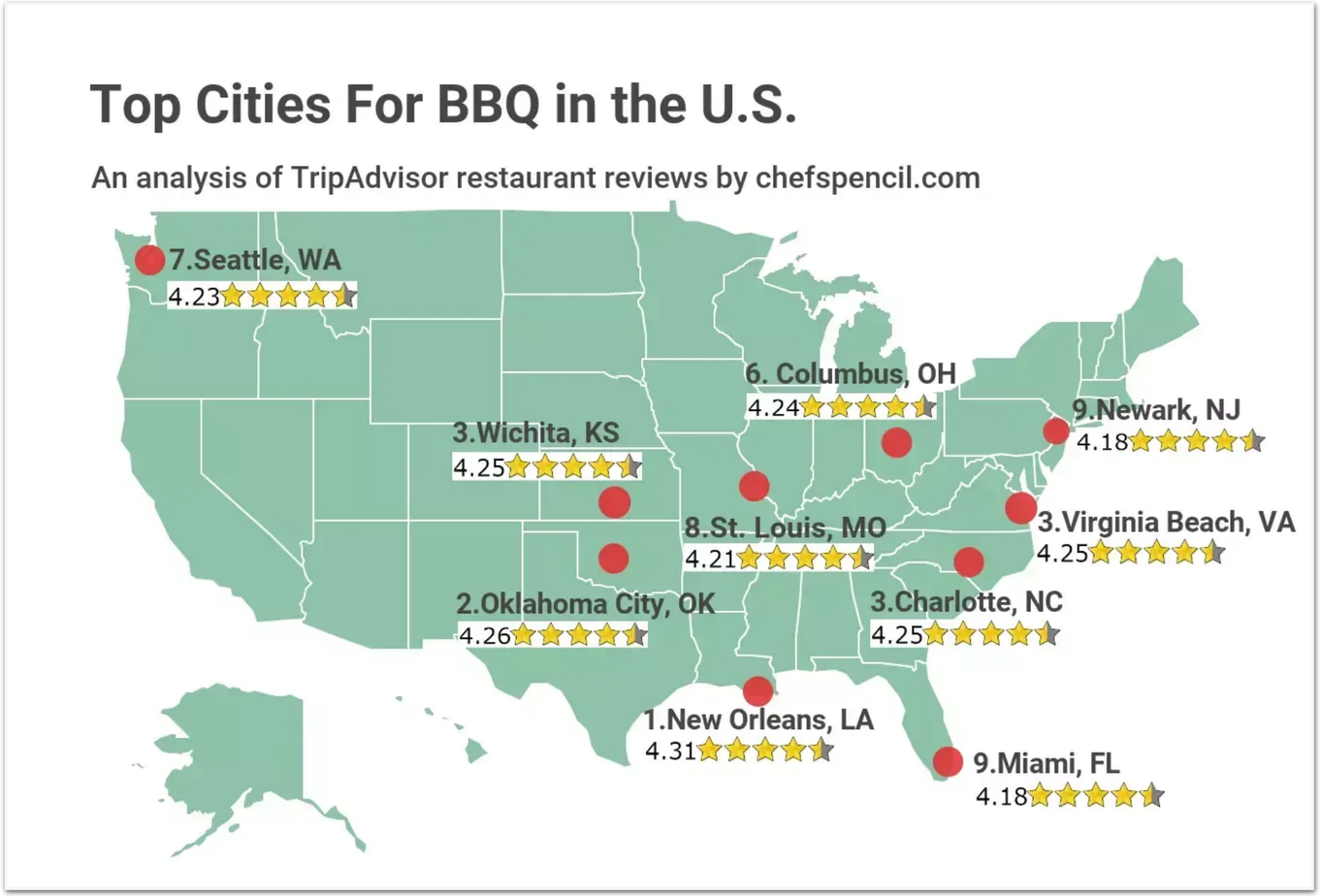
Ecological Fallacy

Example: America's Best BBQ (Chef's Pencil)

- Used Trip Advisor ratings for BBQ restaurants in 75 different US Cities to come up with a top 10 List

Anything surprising?

What might be wrong with this approach?



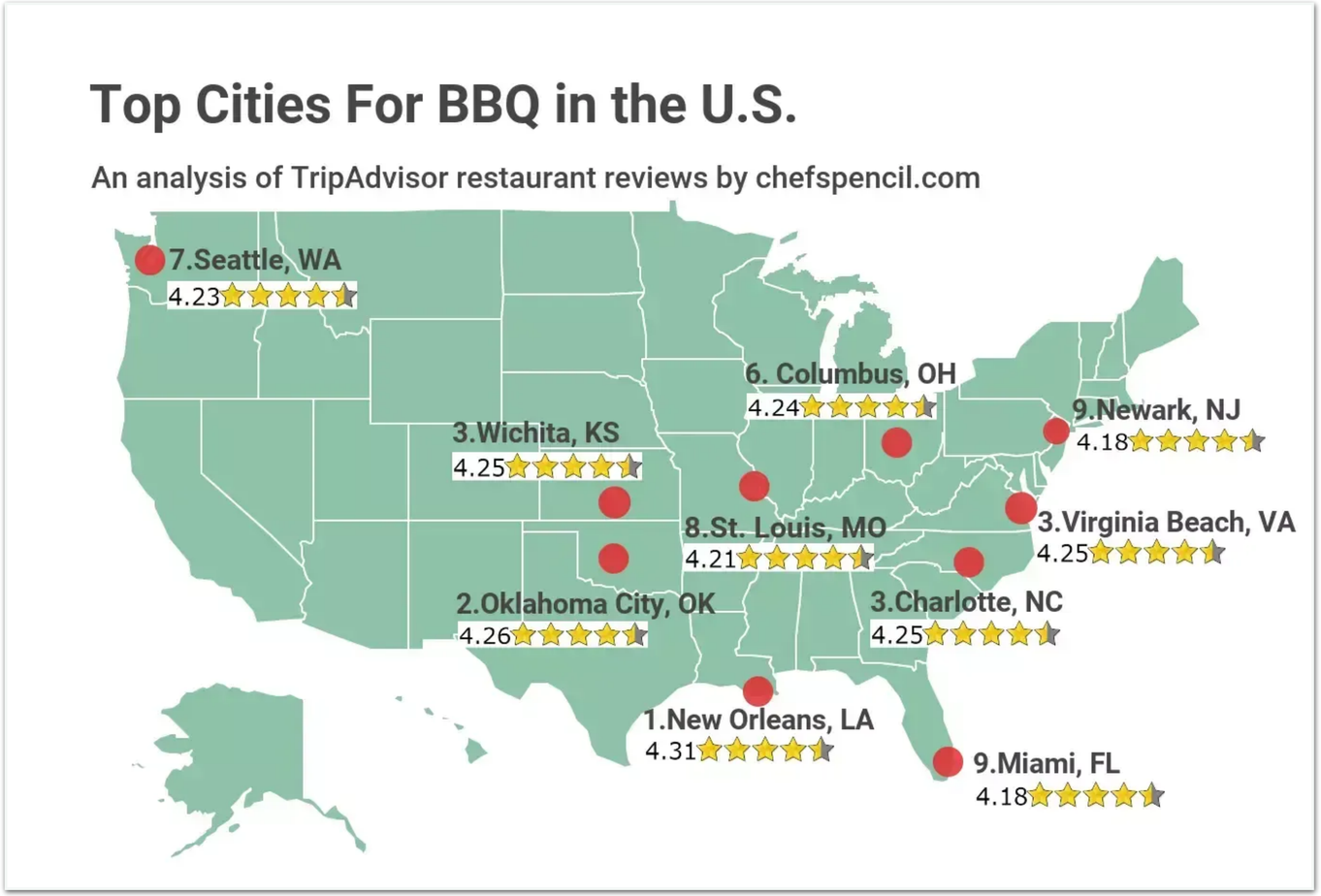
https://callingbullshit.org/case_studies/case_study_barbecue.html

Ecological Fallacy

Reviews *within* a city are being used to compare *between* cities

- Cities with fewer options for BBQ may be judged differently than those with lots of choices

- Average user review may be higher in places that don't have many BBQ places



https://callingbullshit.org/case_studies/case_study_barbecue.html

Errors with Data Collection

- Selection Bias (overrepresentation, exclusion)
- Right-censored data
- Berkson’s Paradox
- Social Desirability Bias

Errors with Interpreting Data

- Validity
- Ecological Fallacy

Conclusion: Think Critically of Data!

- Be aware of these pitfalls!

- Look at how data was collected and apply domain knowledge / understanding to think through other interpretations

- Have healthy skepticism 😅

Errors in Data References

- NYU Data Science for Everyone
- Calling Bullshit: The Art of Skepticism in a Data-Driven World - Carl Bergstrom and Kevin West
- The WEIRDest People in the World - Joseph Henrich
- Eve: How the Female Body Drove 200 Million Years of Human Evolution - Cat Bohannon

Our remaining time together

- **Monday, Dec 1**: Regression Inference and Classification ← HW 8 due
- **Wednesday, Dec 3**: Computing Fellows Workshop
  - Final Project Consultations during Lab
- **Friday, Dec 5**: Progress Report due

- **Monday, Dec 8**: Special Topics (Data Privacy) ← HW 9 due

  Last day of class :(
- **Friday, Dec 12**: Final Projects Due