Week 7 Inference for regression

2. Inference for prediction

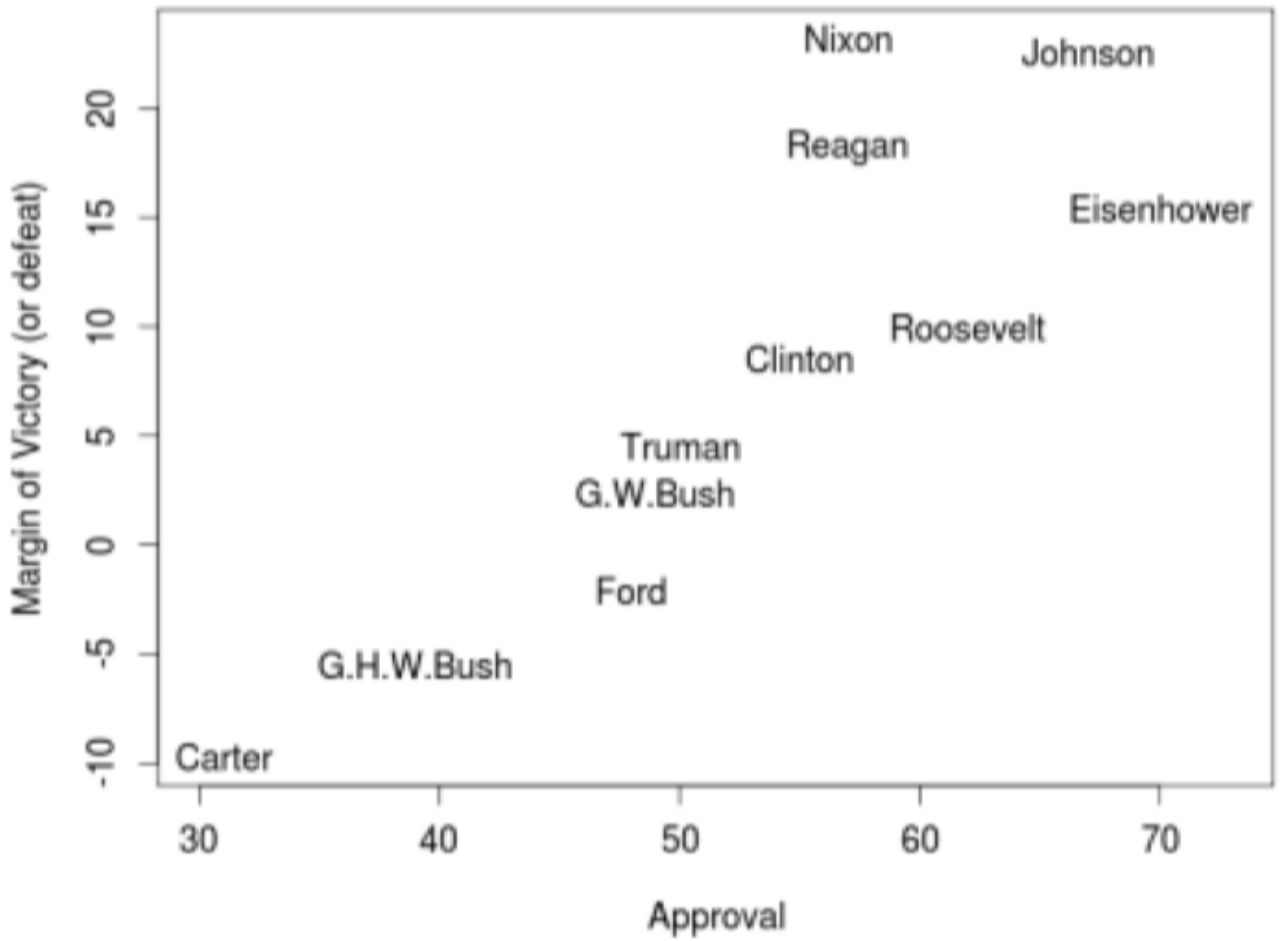
Stat 140 - 04

Mount Holyoke College

Dr. Shan Shan

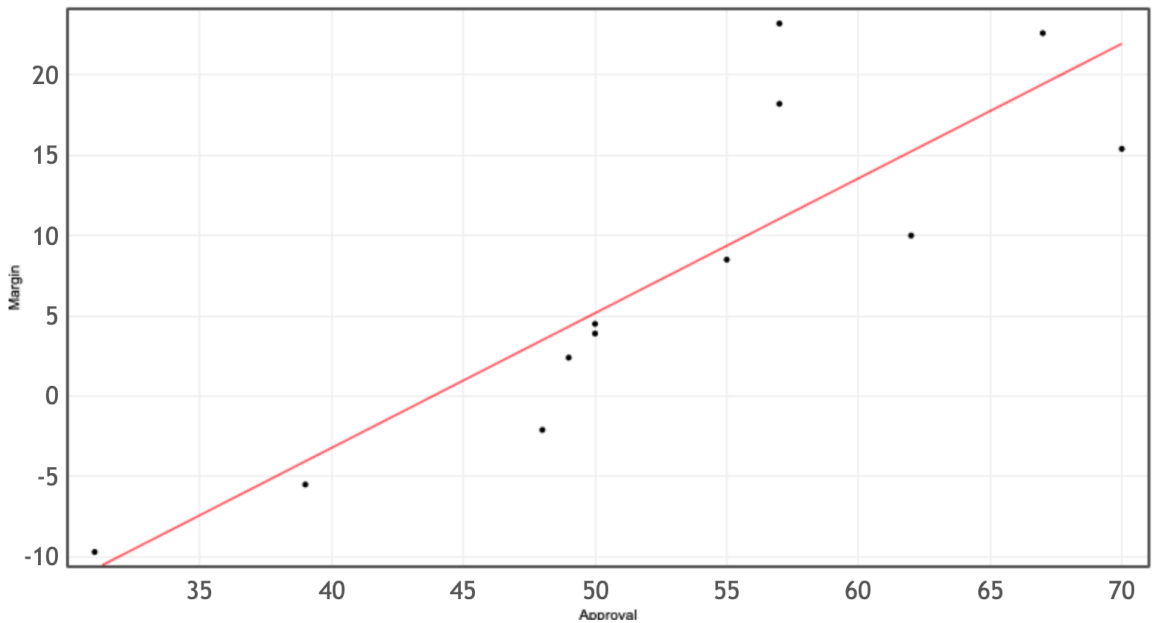
Slides posted at *http://sshanshans.github.io/stat140*

We can build a model using data from past elections to predict an incumbent's margin of victory based on approval rating



Example: Presidential Elections

Election Margin-2e (Margin vs Approval) | Show Data Table | Edit Data | Upload File | Change Column(s)



Summary Statistics Switch Variables

Statistic	Approval	Margin
Mean	52.917	7.617
Standard Deviation	11.041	10.717
Sample Size	12	
Correlation	0.864	
Slope	0.839	
Intercept	-36.76	

Scatterplot Controls

☑ Show Regression Line

What was Obama's predicted margin of victory, based on his approval rating on the day of the election (50%)?

- We would like to use the regression equation to predict y for a certain value of x
- For useful predictions, we also want **interval estimates**
- We will predict the value of y at $x = x^*$

The point estimate for the average y value at $x = x^*$ is simply the predicted value:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

The uncertainty in this point estimate comes from the uncertainty in the coefficients

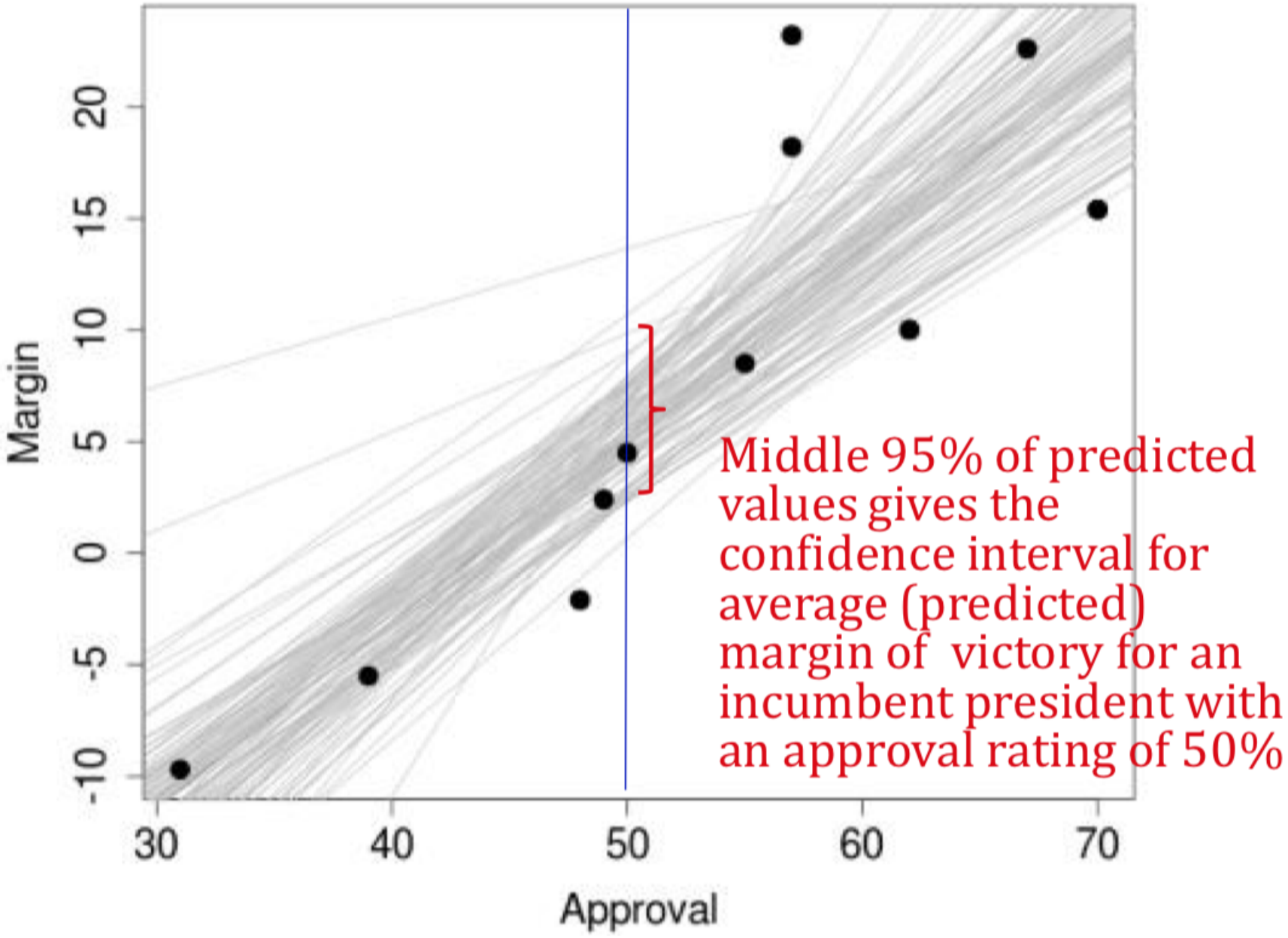
We can calculate a confidence interval for the average y value for a certain x value

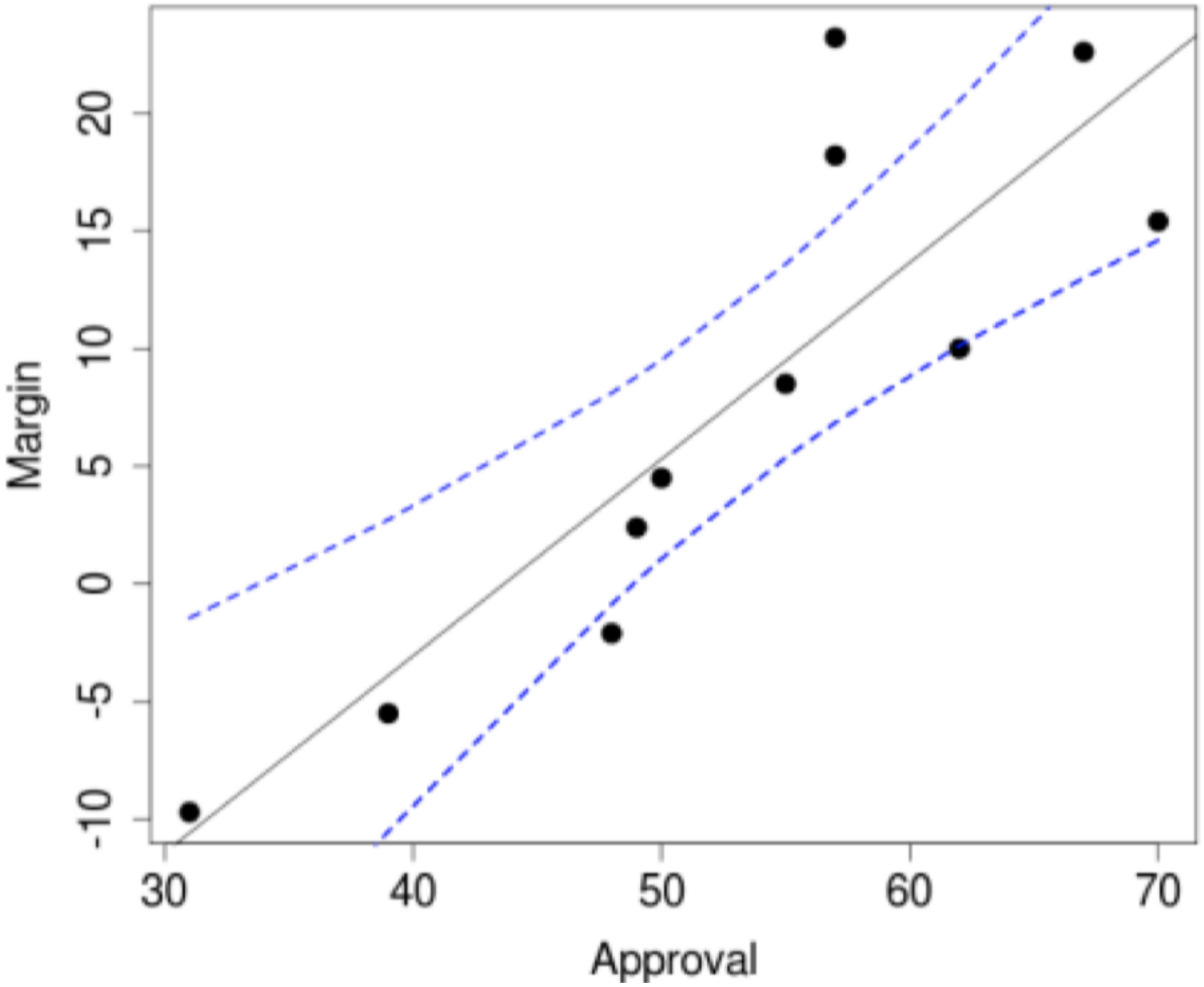
"We are 95% confident that the **average** y value for $x = x^*$ lies in this interval"

Equivalently, the confidence interval is for the point estimate, or the predicted value

This is the amount the line is free to "wiggle," and the width of the interval decreases as the sample size increases

- We need a way to assess the uncertainty in predicted y values for a certain x value... any ideas?
- Take repeated samples, with replacement, from the original sample data (bootstrap)
- Each sample gives a slightly different fitted line
- If we do this repeatedly, take the middle P% of predicted y values at x^* for a confidence interval of the predicted y value at x^*


Margin
20
15
10
5
0
-5
-10
30
40
50
60
70
Approval
Middle 95% of predicted values gives the confidence interval for average (predicted) margin of victory for an incumbent president with an approval rating of 50%


Margin
-10
-5
0
5
10
15
20
30
40
50
60
70
Approval

For $x^* = 50\%$, the confidence interval is $(1.07, 9.52)$

This means,

We are 95% confident that the average margin of victory for incumbent U.S. presidents with approval ratings of 50% is between 1.07 and 9.52 percentage points

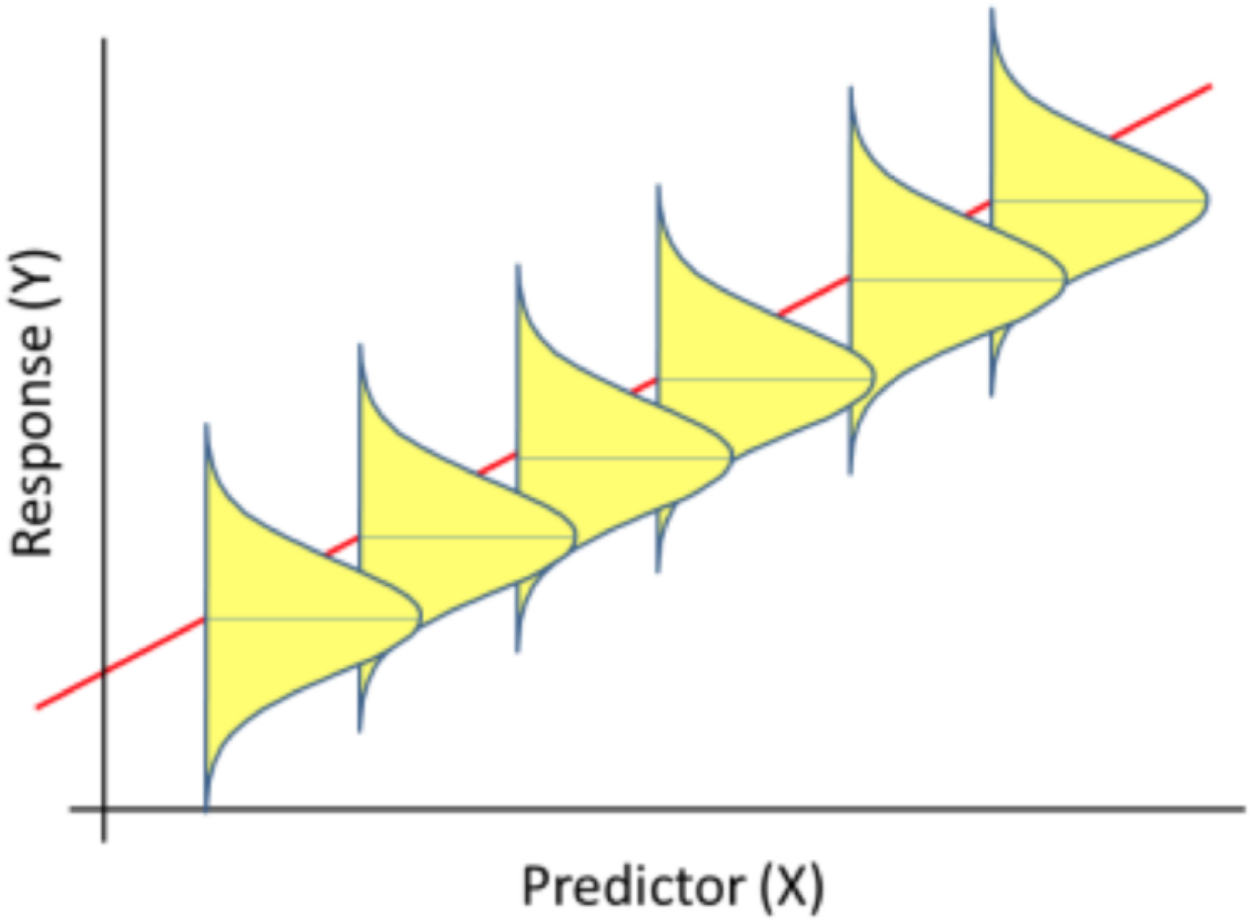
But wait, this still doesn't tell us about a particular incumbent! We don't care about the average, we care about an interval for one incumbent president with an approval rating of 50%!

We can also calculate a prediction interval for y values for a certain x value

"We are 95% confident that the y value for $x = x^*$ lies in this interval"

This takes into account the variability in the line (in the predicted value) AND the uncertainty around the line (the random errors)

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad \varepsilon \sim N(0, \sigma_\varepsilon)$$



- A confidence interval has a given chance of capturing the mean y value at a specified x value
- A prediction interval has a given chance of capturing the y value for a particular case at a specified x value

Poll question

For a given x value, which will be wider?

a. Confidence interval
b. Prediction interval

As the sample size increases:

- Confidence interval
  - the standard errors of the coefficients decrease
  - we are more sure of the equation of the line
  - the widths of the confidence intervals decrease
  - for a huge n, the width of the CI will be almost 0
- Prediction interval
  - The prediction interval may be wide, even for large n
  - The interval depends more on the correlation between x and y (how well y can be linearly predicted by x)

Based on the data and the simple linear model:

The predicted margin of victory for an incumbent with an approval rating of 50% is 5.3 percentage points

We are 95% confident that the margin of victory (or defeat) for an incumbent with an approval rating of 50% will be between -8.8 and 19.4 percentage points

NOTE: You will never need to use these formulas in this class – you will just have RStudio do it for you.

Confidence Interval:

$$\hat{y} \pm t^* \times s_e \sqrt{\frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{(n-1)s_x^2}}$$

Prediction Interval:

$$\hat{y} \pm t^* \times s_e \sqrt{1 + \frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{(n-1)s_x^2}}$$

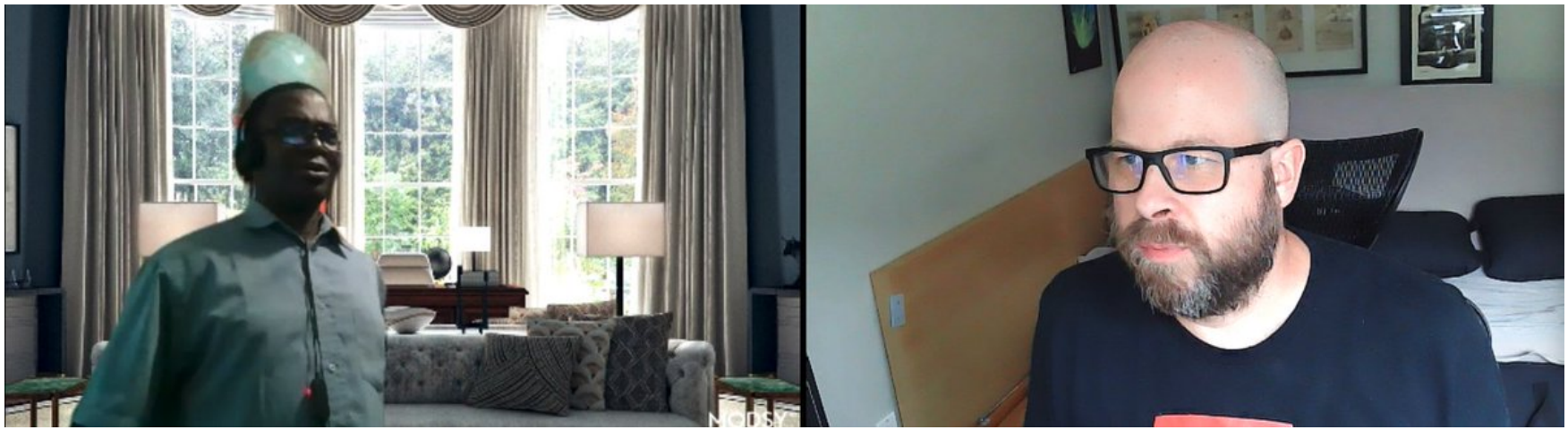
s_e : estimate for the standard deviation of the residuals

- No outliers (points that don't fit the trend)
- Straight enough?
- Does the plot thicken?
- Sample representative of population
- Independence
- Normally distributed residuals (or large enough sample size)

"All models are wrong, but some are useful"

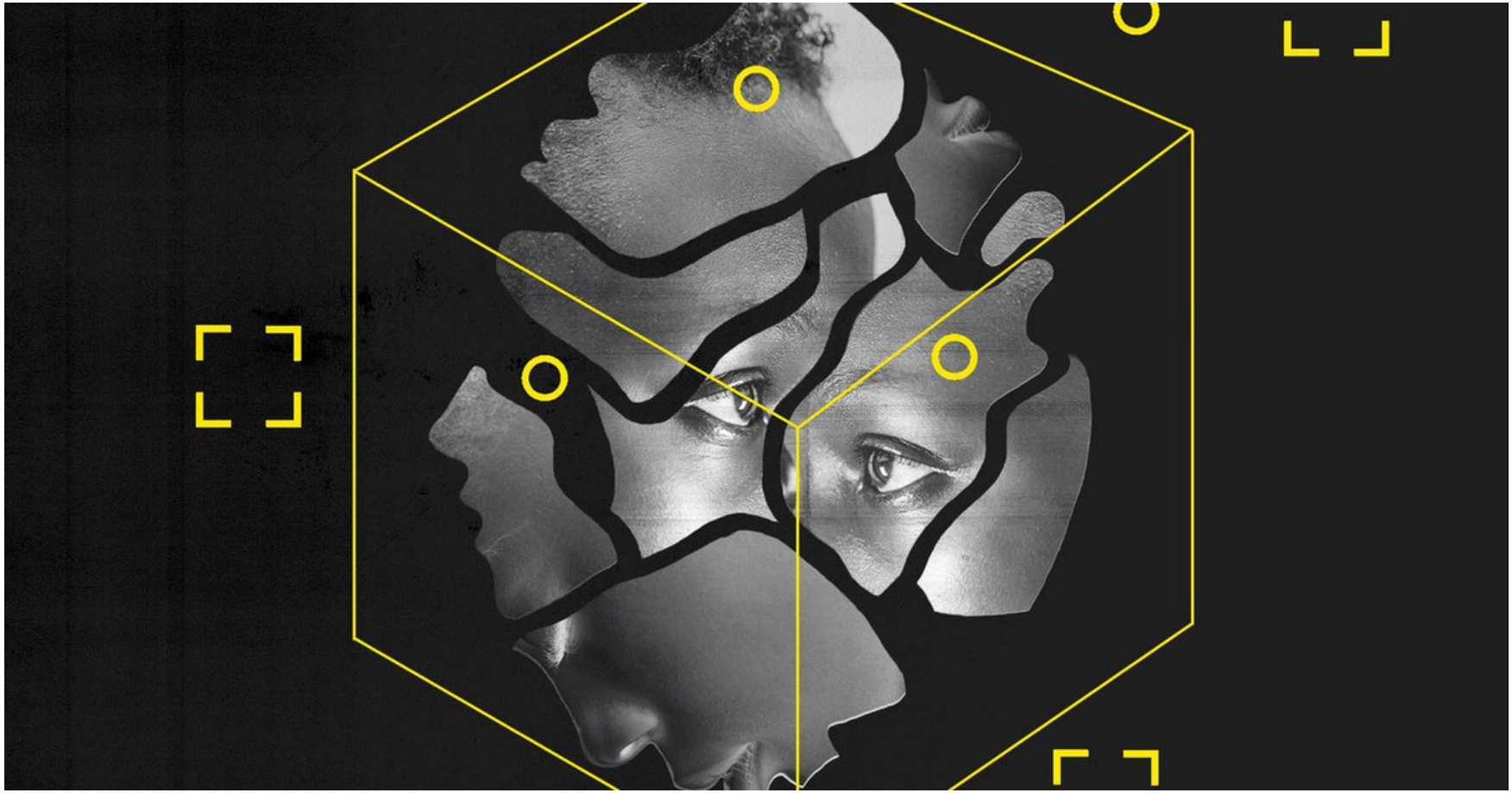
-George Box

Where is my face?

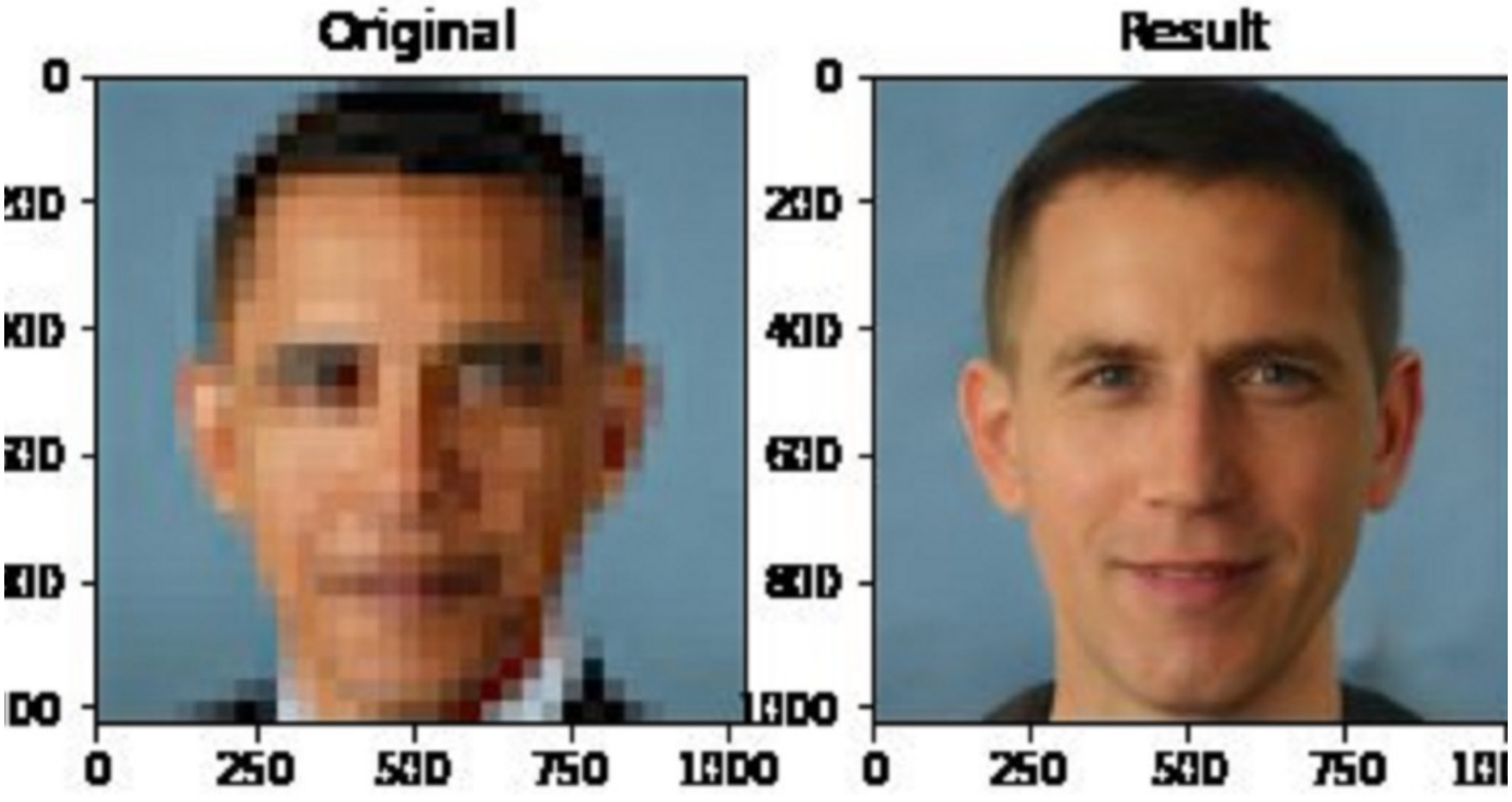


Turns out Zoom has a crappy face-detection algorithm that erases black faces...and determines that a nice pale globe in the background must be a better face than what should be obvious.

The best algorithm still struggles to detect black faces



https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/


Original
Result
0
250
500
750
1000

Similar idea

Bias in the AI system

- A training dataset that isn't representative
- A training dataset that has societal bias baked in
- A poorly chosen objective function in an ML model

What can you do?

- Defining and following a set of AI principles: `https://ai.google/responsibilities/responsible-ai-practices/`
- Investing in tools and technology approaches to support the operationalization of the principles, e.g, AI Fairness 360 `https://aif360.mybluemix.net`
- Diversify your team `https://arxiv.org/pdf/2002.11836.pdf`


WITH GREAT POWER
COMES GREAT
RESPONSIBILITY...