Review Session 4

API-202 2022
Sophie Hill

Agenda

- Review the quiz
  - Overall, great job
  - Discuss some of the harder questions
- Review important concepts that weren't on the quiz
- Review notation
- Data viz with ggplot

Quiz #1

Feedback: great job!!

Tricky questions:

- Q5: association vs correlation
- Q7: prediction vs causal

Q5: association vs correlation

"We have learned that it is possible for two variables to be associated even if their association is not causal. Is it also possible for two variables to be associated even if their correlation is zero? Why or why not?"

Q5: association vs correlation

“We have learned that it is possible for two variables to be associated even if their association is not causal. Is it also possible for two variables to be associated even if their correlation is zero? Why or why not?”

Yes!

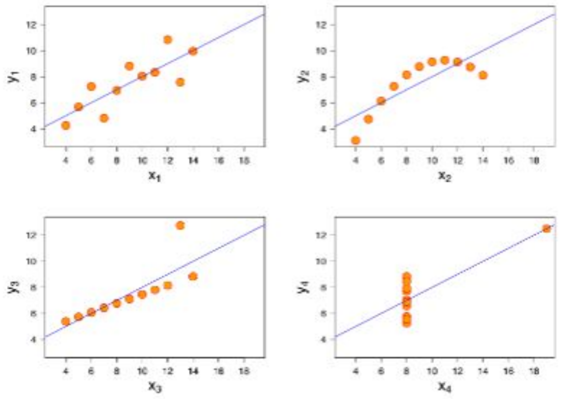
It is possible for two variables to have a correlation of 0 but still be associated. For example, it could be that the relationship is *non-linear* or that the relationship only holds within a subset of the dta.

When did we cover this in class?

Section B / Bloome / Class 2 Slide #15

You must plot your data!

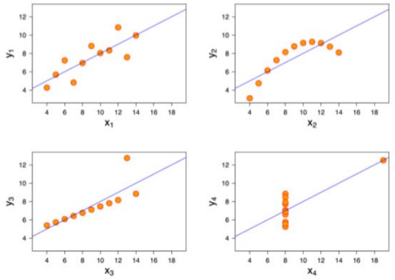
Different patterns can result in the same correlation. Be aware of nonlinearities and outliers!



Question: Which pattern generates the most misleading correlation (that is, the association worst summarized by a straight line)?

Section C / Schneer / Class 1 Slide #36

Different patterns can yield the same correlation

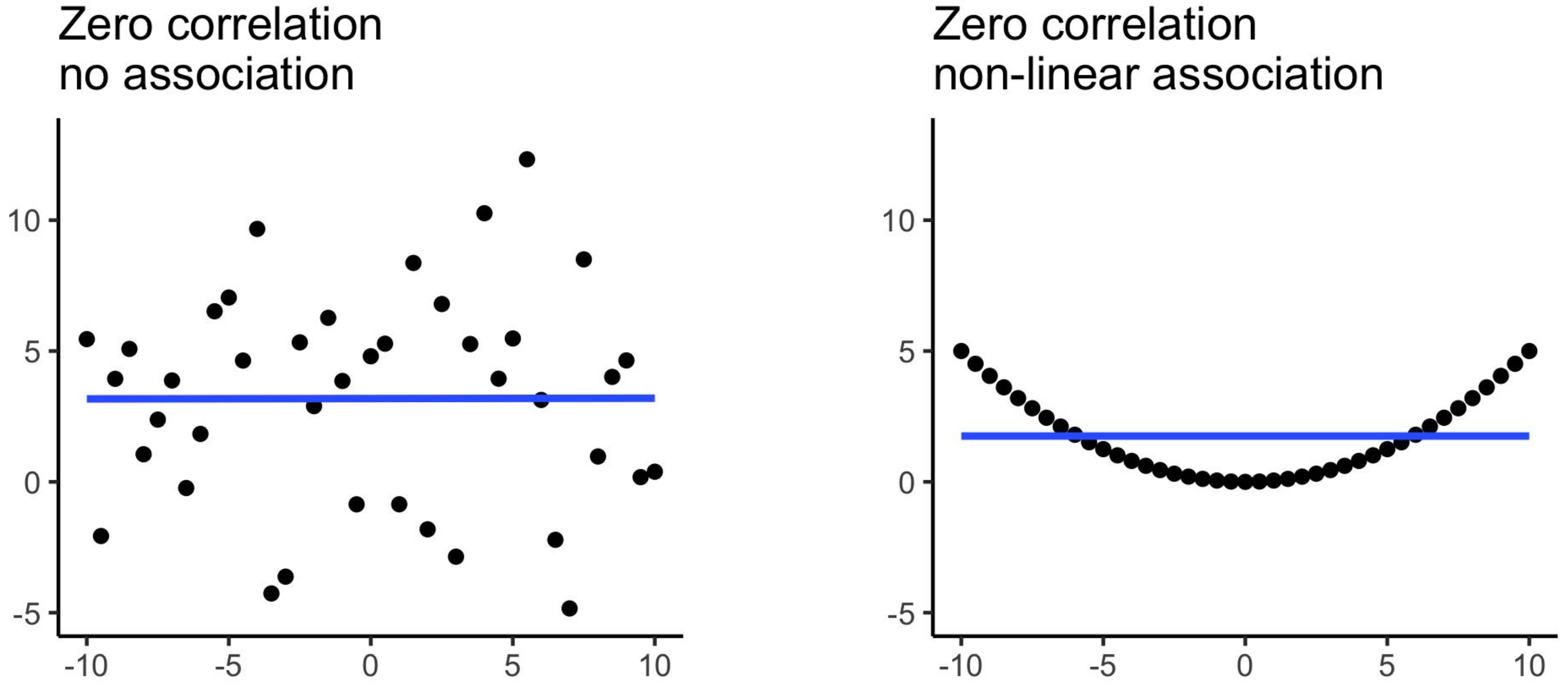


← This is why it's good to look at your data

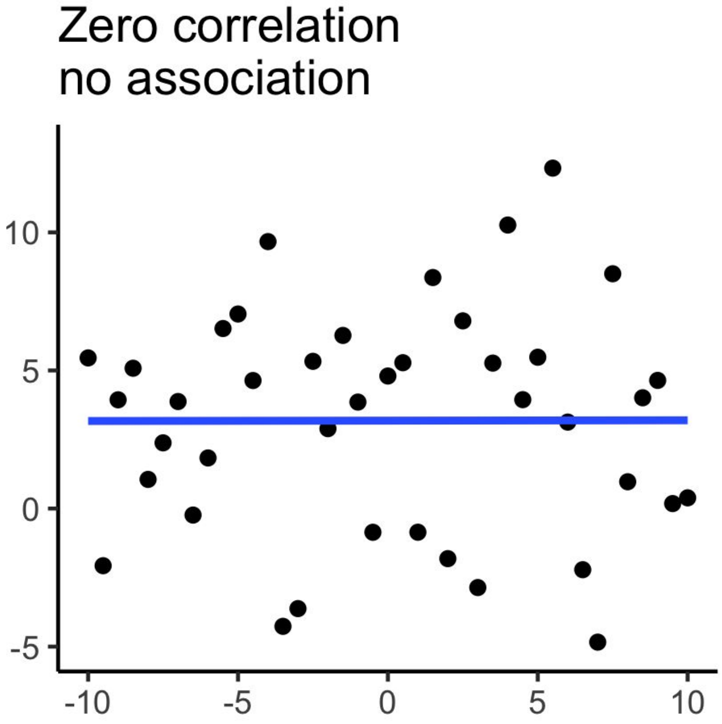
Which model do you trust most?

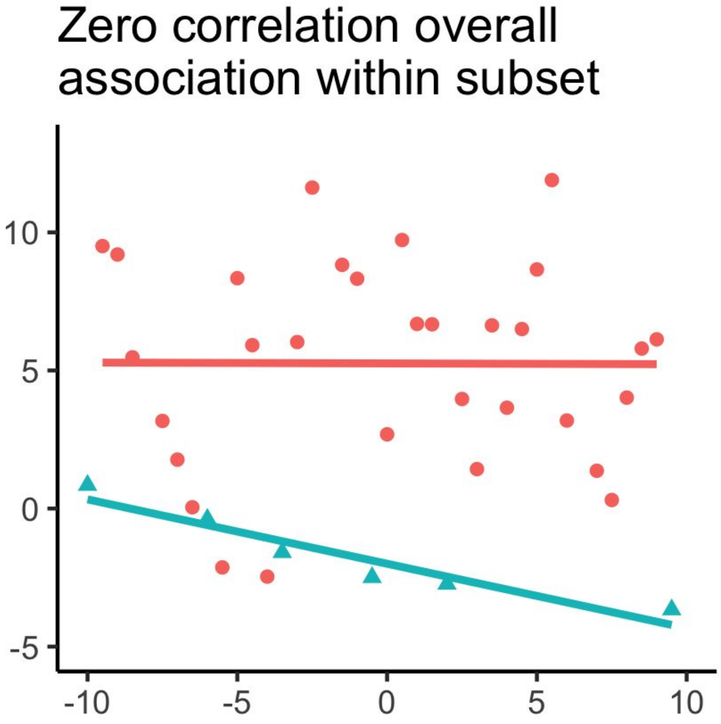
- Top left
- Top right
- Bottom left
- Bottom right

Q5: association vs correlation



Q5: association vs correlation





Q7*: prediction vs causal

*If you are in section C this question may have appeared as Q8, since an additional question was included by mistake.

Your partner (from Question 7) says that in their predictive model, they included a bunch of predictor variables that they know are not causally associated with housing cost burden.Is that OK? Explain why or why not in one sentence.

Q7*: prediction vs causal

*If you are in section C this question may have appeared as Q8, since an additional question was included by mistake.

Your partner (from Question 7) says that in their predictive model, they included a bunch of predictor variables that they know are not causally associated with housing cost burden.Is that OK? Explain why or why not in one sentence.

Yes, because the goal of the analysis is **predictive** (to learn about future cost burdens), not **causal** (to learn about how changing some input of interest might change cost burdens).

When did we cover this in class?

Section B / Bloome / Class 2 Slide #45

Prediction

We would like to predict what life expectancy will be in Pakistan in 2023. Should we use the number of TVs per capita to help us make this prediction?

(A) Yes

(B) No

Section C / Schneer / Class 2 Slide #6-7

Prediction: What *predicts* the outcome of interest?

Suppose **losing a job** makes one more likely to:

- Receive SNAP (food stamps)

  – and –
- **Experience homelessness**

In **prediction**:

losing a job → **experience homelessness**

– or –

receive SNAP → **experience homelessness**

Causality: What *causes* the outcome of interest?

Suppose **losing a job** makes one more likely to:

- Receive SNAP

  – and –
- **Experience homelessness**

In **causal inference**, we want to isolate:

losing a job → **experience homelessness**

– or –

~~receive SNAP → **experience homelessness**~~

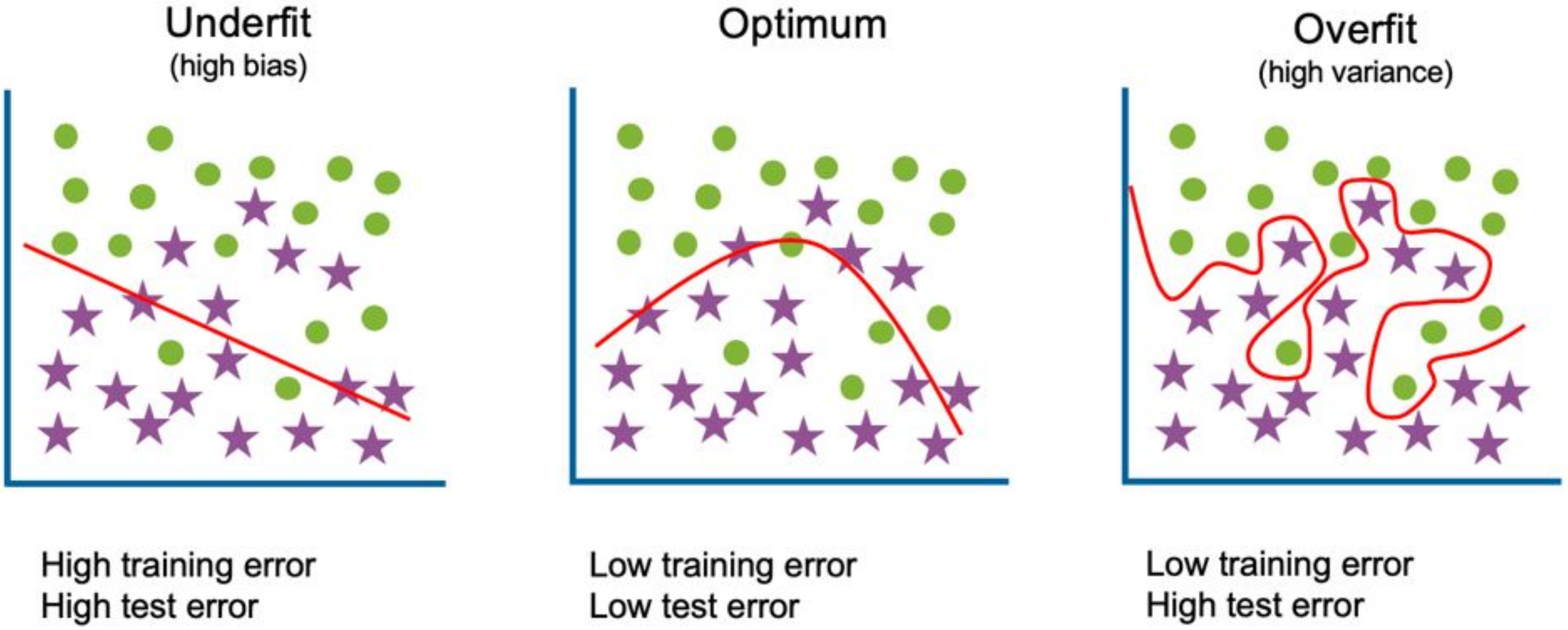
Important concepts that weren't on the quiz

Overfitting

Bias-variance trade-off

Prediction accuracy: recall and precision

Overfitting



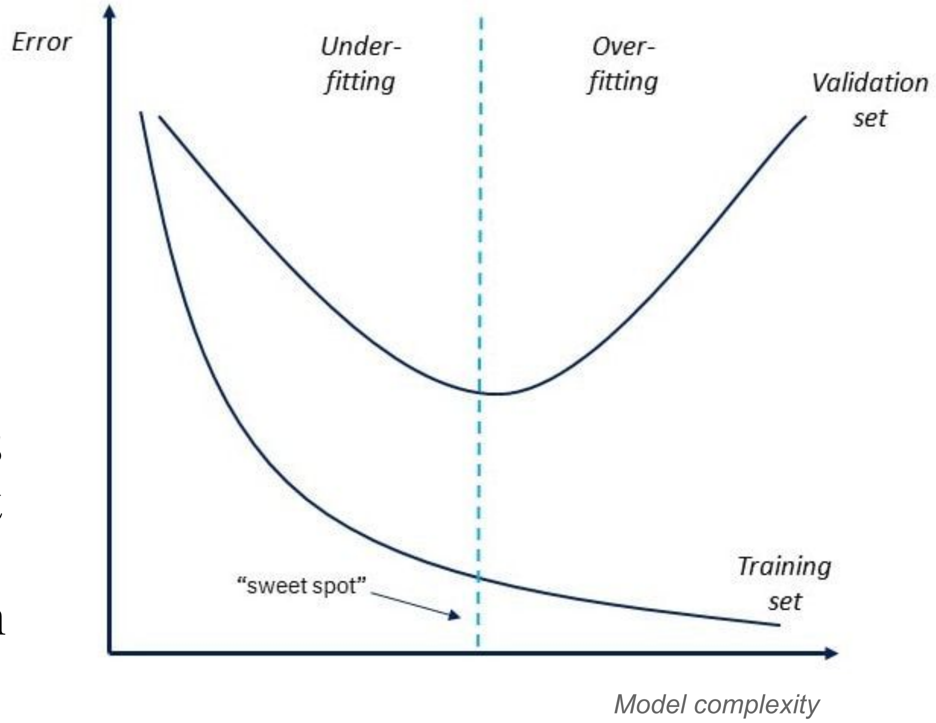
Source: [IBM "What is overfitting?"]

Overfitting

As we fit our model more and more closely to our training set, we can reduce our prediction errors on both the training set and the test set (or "validation set").

But beyond a certain point, our model is just "learning the noise", meaning that it is getting better at fitting the training data but *worse* at predicting outcomes in the test set!

This is overfitting.



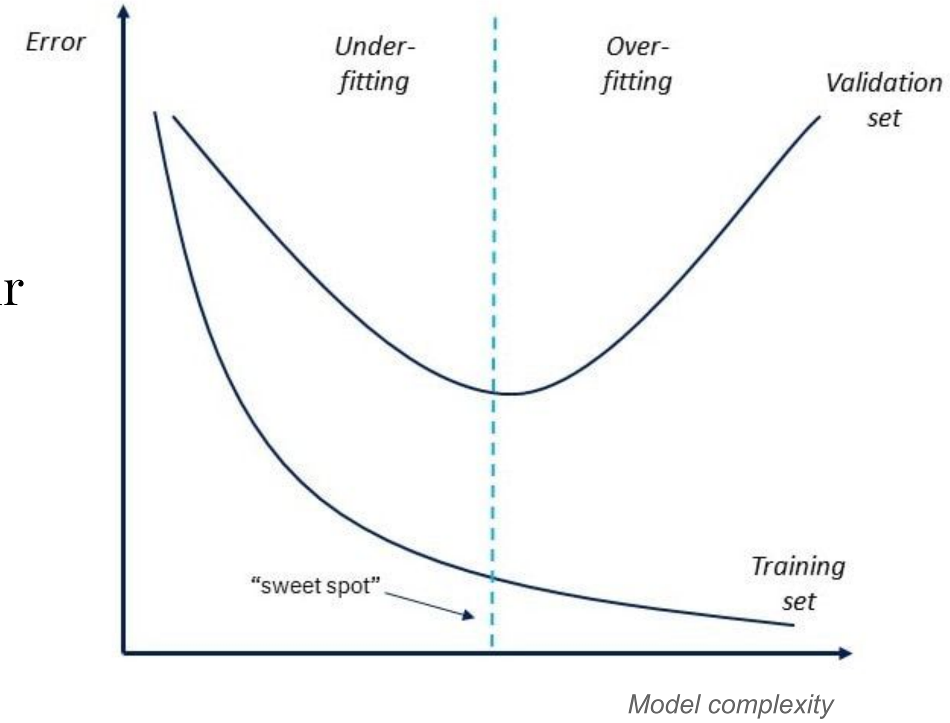
Source: [IBM "What is overfitting?"]

Overfitting

There is no *a priori* way to know where this “sweet spot” is.

Instead, we focus on comparing how our model performs on both sets of data.

If it performs well on the training data but badly on the test data, that is a sign of overfitting.

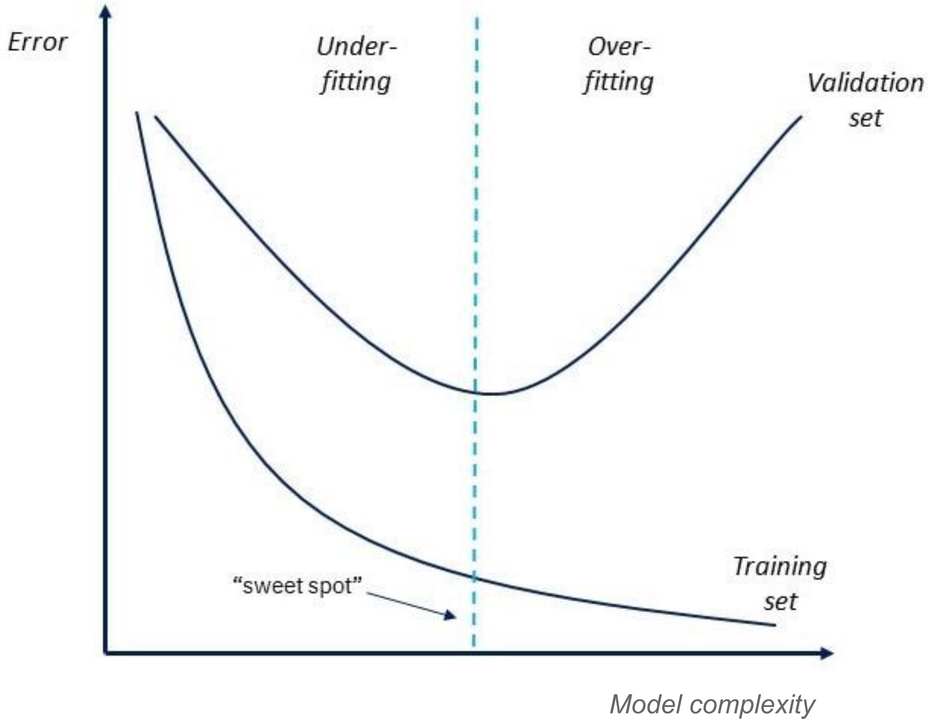


Source: IBM “What is overfitting?”

Bias-variance trade-off

The problem of overfitting stems from the bias-variance trade-off!

Beyond the "sweet spot", if we continue reducing the bias (i.e. reducing the prediction errors on the training set), we will inevitably increase the variance (i.e. increase the sensitivity of our model to idiosyncracies in the training set).



Source: IBM "What is overfitting?"

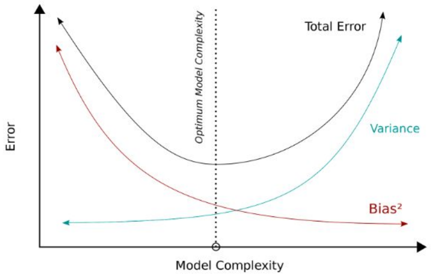
When did we cover this in class?

Section B / Bloome / Class 2 Slide #31

Bias/variance contributions to total error

$$\text{Total Error} = \text{Bias}^2 + \text{Variance}$$

- To minimize the total error, ideally we would minimize both bias and variance... but sadly, we cannot!
- There is a **bias/variance tradeoff**
  - This tradeoff appears all over statistical analyses
  - For prediction, the terms have specific meanings



Section C / Schneer / Class 1 Slide #53

Bias/Variance tradeoff is key to model performance

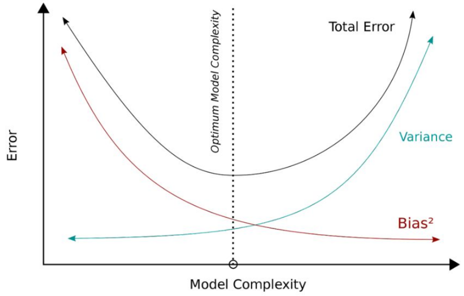
Bias:

- How well your model fits your training data.
  - Strong training fit = low bias, because model is ***not making assumptions that differ*** from the relationships seen in your data

Variance:

- How consistent your training and test fit are. Similar error = low variance, because the model's performance is ***not sensitive to the exact data*** it's given

Optimize bias and variance to get low error



Prediction accuracy: recall and precision

We often think about accuracy as a 1-dimensional concept: more vs less accurate.

But as we know from API201, there are several quantities we can look at.

For example:

Sensitivity = P(+ | COVID)

Specificity = P(- | no COVID)

PPV = P(COVID | +)

NPV = P(no COVID | -)

Prediction accuracy: recall and precision

Sensitivity = P(+ | COVID)

Specificity = P(- | no COVID)

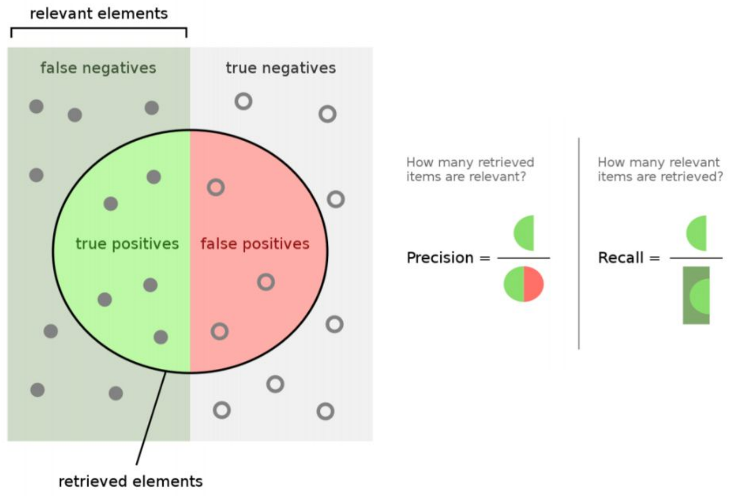
PPV = P(COVID | +)

NPV = P(no COVID | -)

Precision + Recall are common error summaries

Precision considers ***all predicted positives*** and finds what share is correctly classified

Recall considers ***all real positives*** and finds what share is correctly classified



Question:

Can you “match up” precision and recall with 2 of the concepts you learned in API201?

Prediction accuracy: recall and precision

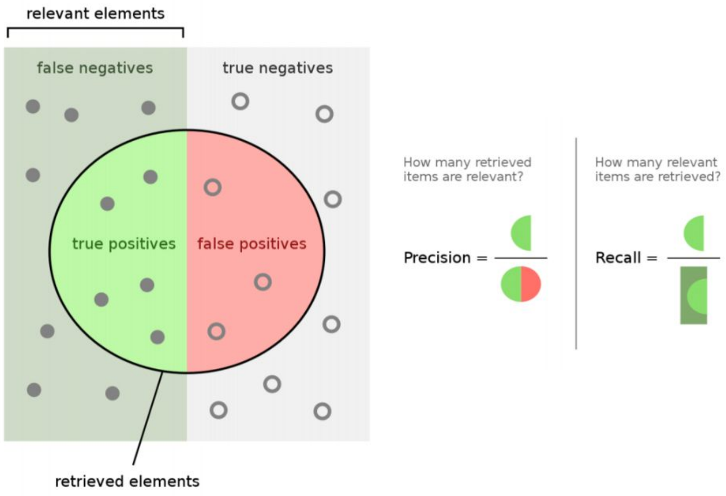
Sensitivity = P(+ | COVID)   PPV = P(COVID | +)

Specificity = P(- | no COVID)   NPV = P(no COVID | -)

Precision + Recall are common error summaries

Precision considers ***all predicted positives*** and finds what share is correctly classified

Recall considers ***all real positives*** and finds what share is correctly classified



Question:

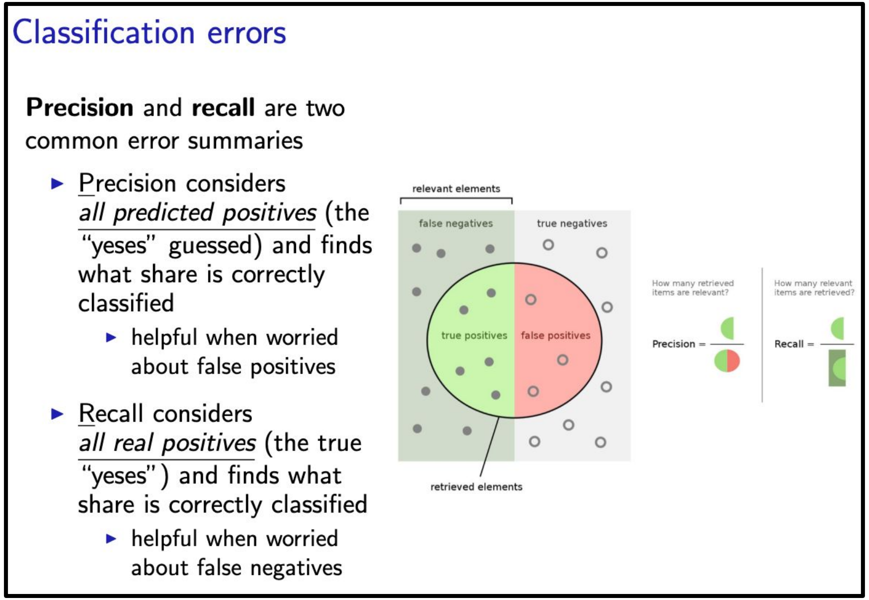
Can you "match up" precision and recall with 2 of the concepts you learned in API201?

Answer:

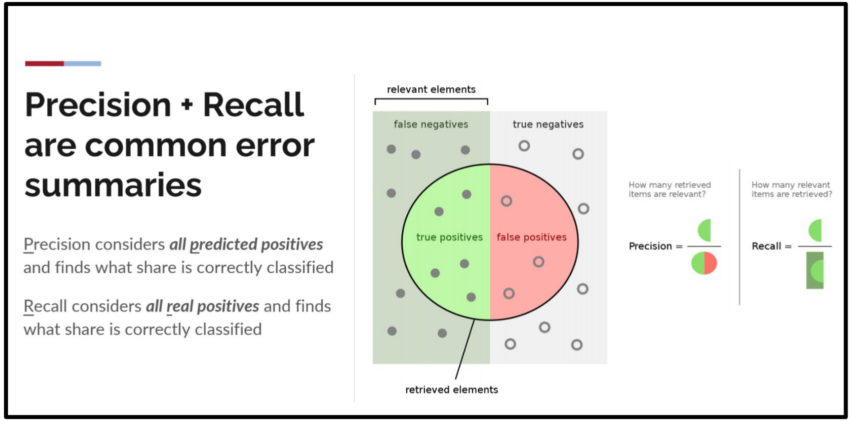
Precision = PPV
Recall = Sensitivity

When did we cover this in class?

Section B / Bloome / Class 2 Slide #22



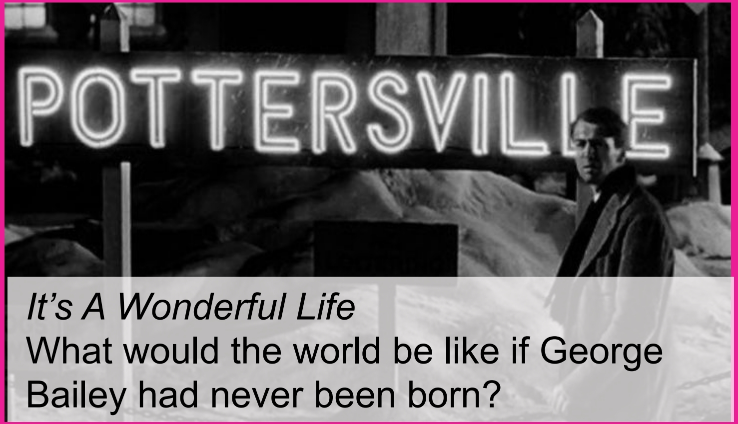
Section C / Schneer / Class 1 Slide #47



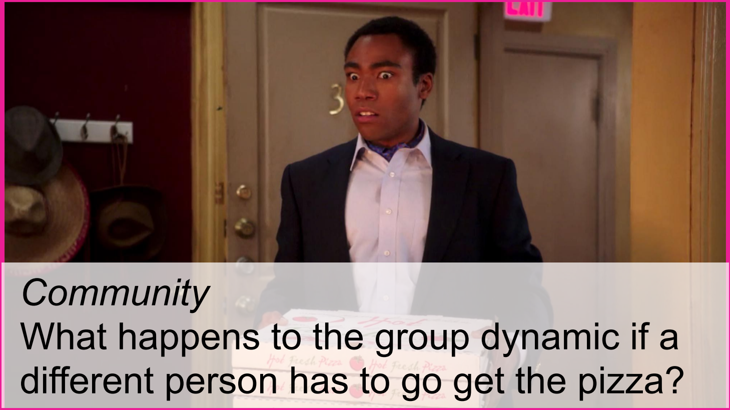
POP (CULTURE) QUIZ!

What is the common link?

Counterfactuals!



It's A Wonderful Life
What would the world be like if George Bailey had never been born?



Community
What happens to the group dynamic if a different person has to go get the pizza?

Sliding Doors
What happens to Gwyneth Paltrow’s life if she does / does not catch the tube?

The Man in the High Castle
What is the U.S. like if the Axis Powers won WWII?

Unlike in the movies, in real life we don’t get to observe the counterfactual.

The framework of “potential outcomes” helps us think about these unobserved counterfactuals in a rigorous way.

Review notation

Notation	Words
Y_i	Observed outcome for individual i
$Y_i(1)$	Potential outcome for individual i under treatment
$Y_i(0)$	Potential outcome for individual i under control
$Y_i(1) - Y_i(0)$	Difference in potential outcomes for individual i under treatment vs control
$E[Y_i(1) - Y_i(0)]$	Average difference in potential outcomes under treatment vs control for individuals $i = 1, 2, ..., n$

Review notation

Let's put this in context:
Y = did this person turn out to vote in 2020?
T = did this person receive a GOTV text message?

Notation	Words
Y_i	Observed outcome for individual i
$Y_i(1)$	Potential outcome for individual i under treatment
$Y_i(0)$	Potential outcome for individual i under control
$Y_i(1) - Y_i(0)$	Difference in potential outcomes for individual i under treatment vs control
$E[Y_i(1) - Y_i(0)]$	Average difference in potential outcomes under treatment vs control for individuals $i = 1, 2, \ldots, n$

Review notation

Let's put this in context:
Y = did this person turn out to vote in 2020?
T = did this person receive a GOTV text message?

Notation	**Words**
Y_i	Observed turnout for individual i
$Y_i(1)$	Whether individual i *would have* turned out *if* they were sent a GOTV text
$Y_i(0)$	Whether individual i *would have* turned out *if* they were not sent a GOTV text
$Y_i(1) - Y_i(0)$	Difference in turnout for individual i, with vs without a GOTV text
$E[Y_i(1) - Y_i(0)]$	Average difference in turnout with vs without a GOTV text for individuals $i = 1, 2, ..., n$

Dataviz with ggplot()

See `rs4_dataviz.html` on Canvas (Files ↣ Review Sessions ↣ RS4) for a discussion of some key principles of dataviz and how to implement them with ggplot().

You are not required to memorize this stuff, I am just providing it as a resource for you! :)