LECTURE 09:
RESAMPLING WITH CROSS-VALIDATION AND BOOTSTRAP

October 11, 2017

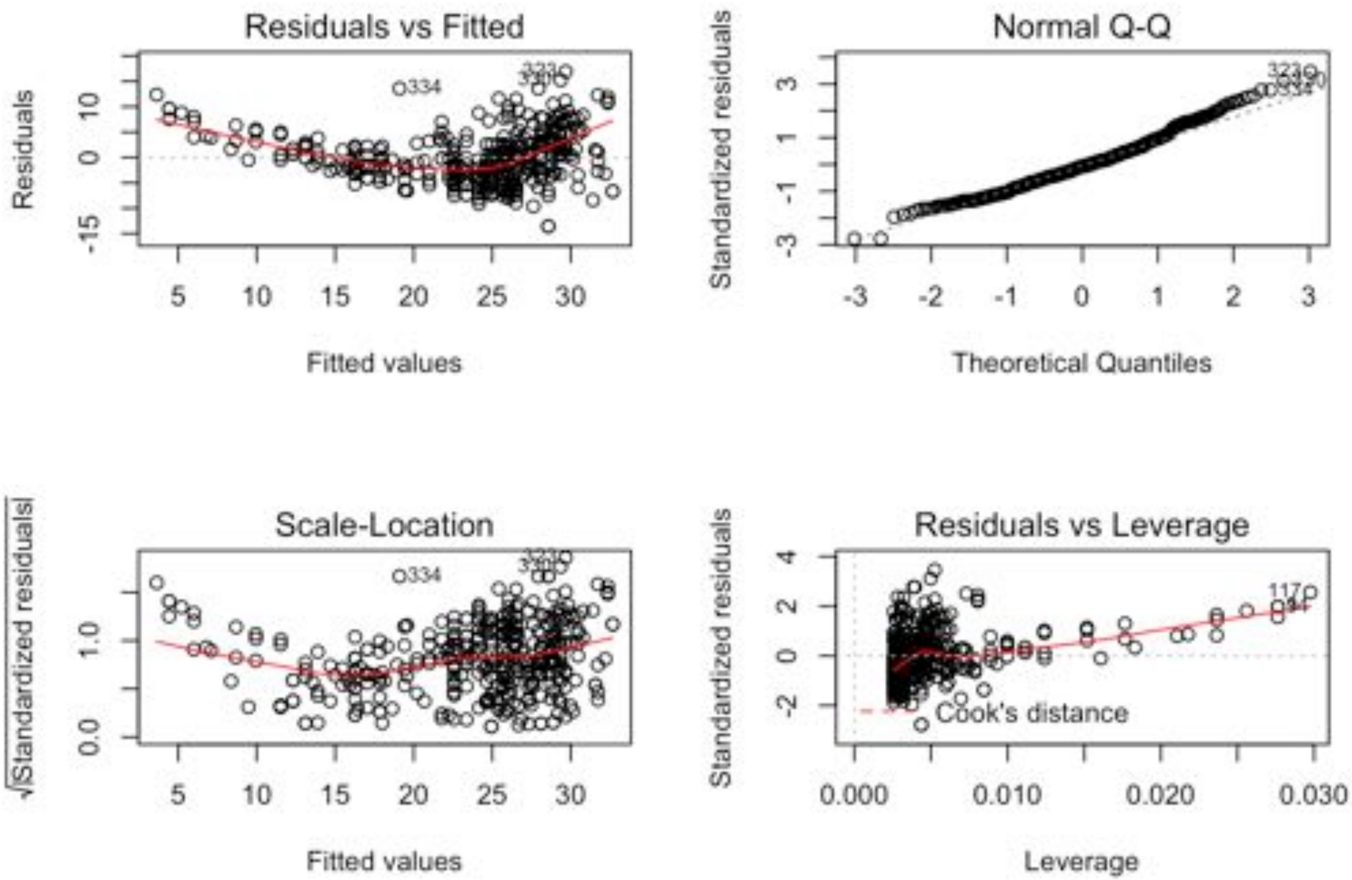
SDS 293: Machine Learning

Announcements / reminders

- Stats TAs available **every weeknight** 7-9 in Burton 301
- Labs due 24 hours after class
  - No late labs without prior arrangement
  - If you miss a deadline, post anyway to get participation credit
- Homework:
  - Applied problems ➔ .Rmd or .ipynb (not PDF)
  - No need to submit knitted version
  - If you work with a group, remember to **attribute them**
  - Late submissions take a **10% hit per day** (starting at 12:00am)
  - **Extension**? Request 48+ hours in advance, or talk to your Dean

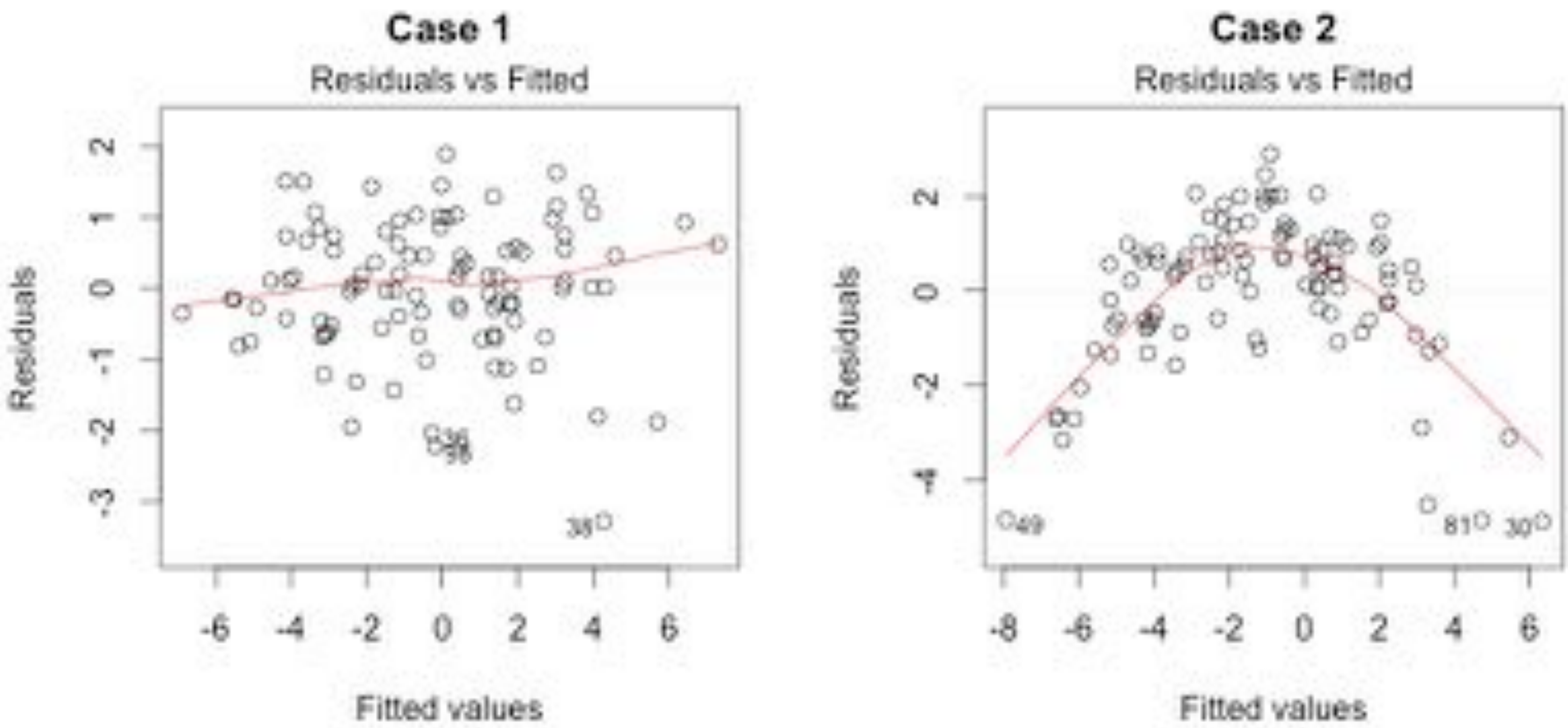
Lingering questions

- **Question:** how do I interpret the results of plot(model)?



Lingering questions

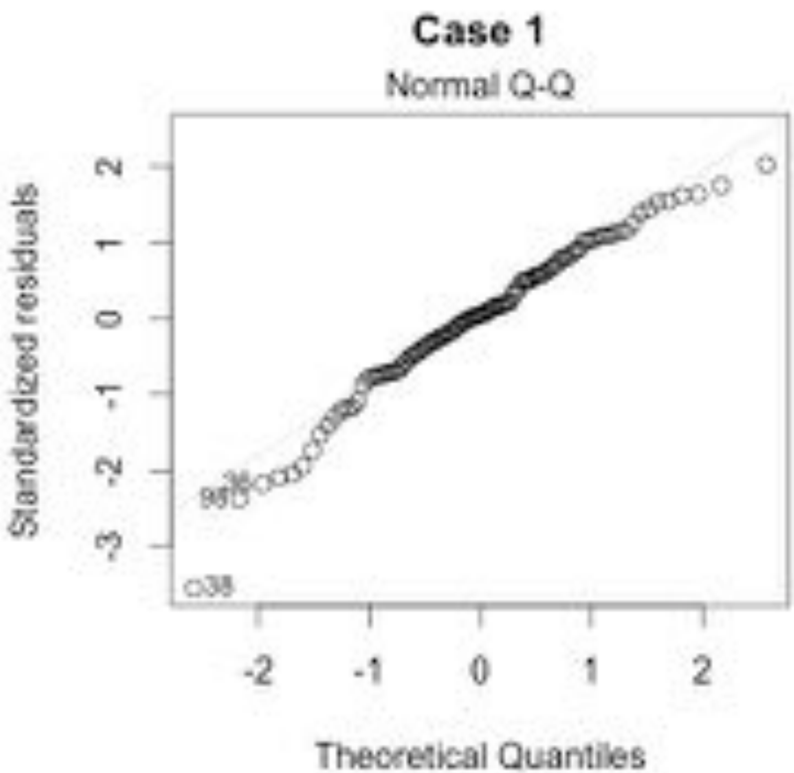
- **Answer (upper left)**: residuals vs. fitted

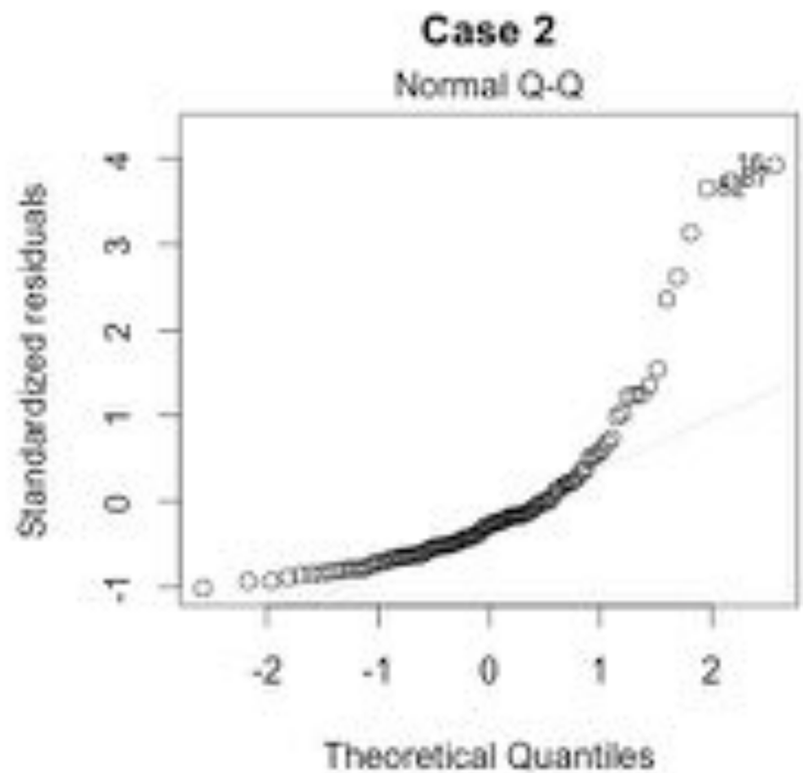


http://data.library.virginia.edu/diagnostic-plots/

Lingering questions

- **Answer (upper right)**: Normal Q-Q

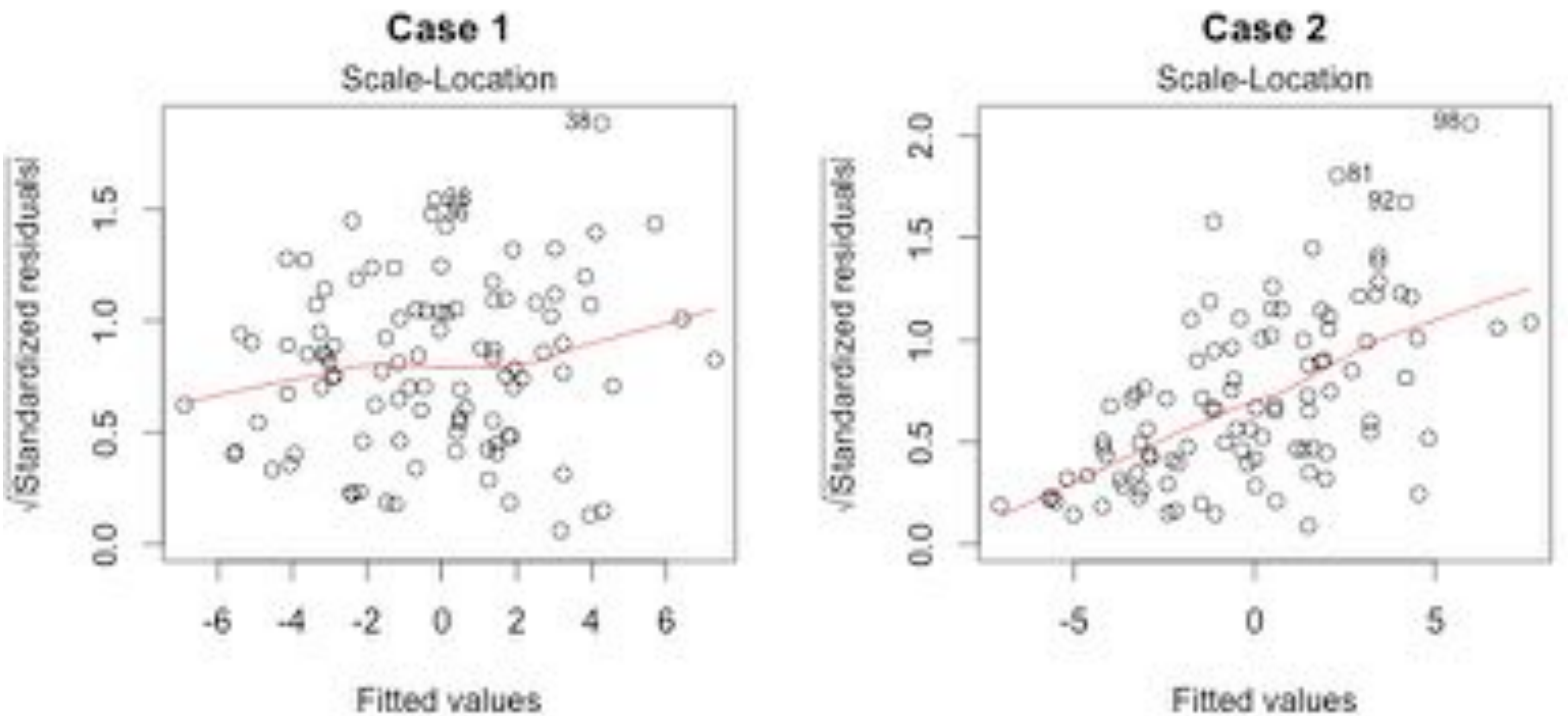




http://data.library.virginia.edu/diagnostic-plots/

Lingering questions

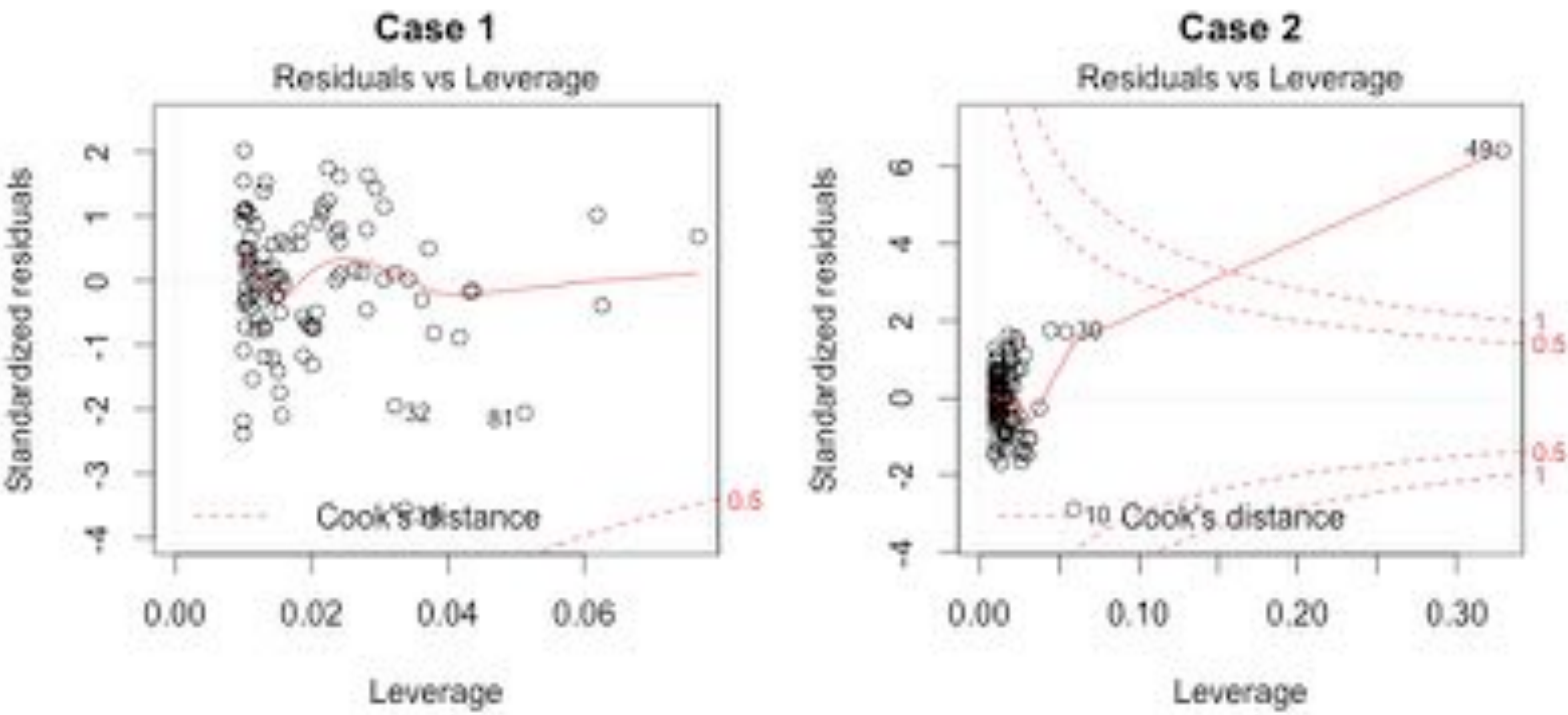
- **Answer (bottom left)**: Scale-Location



http://data.library.virginia.edu/diagnostic-plots/

Lingering questions

- **Answer (bottom right)**: Residuals vs. Leverage



http://data.library.virginia.edu/diagnostic-plots/

Lingering questions

- **Question:** how do I pick a good subset of predictors?
- **Answer:** tune in on Monday!

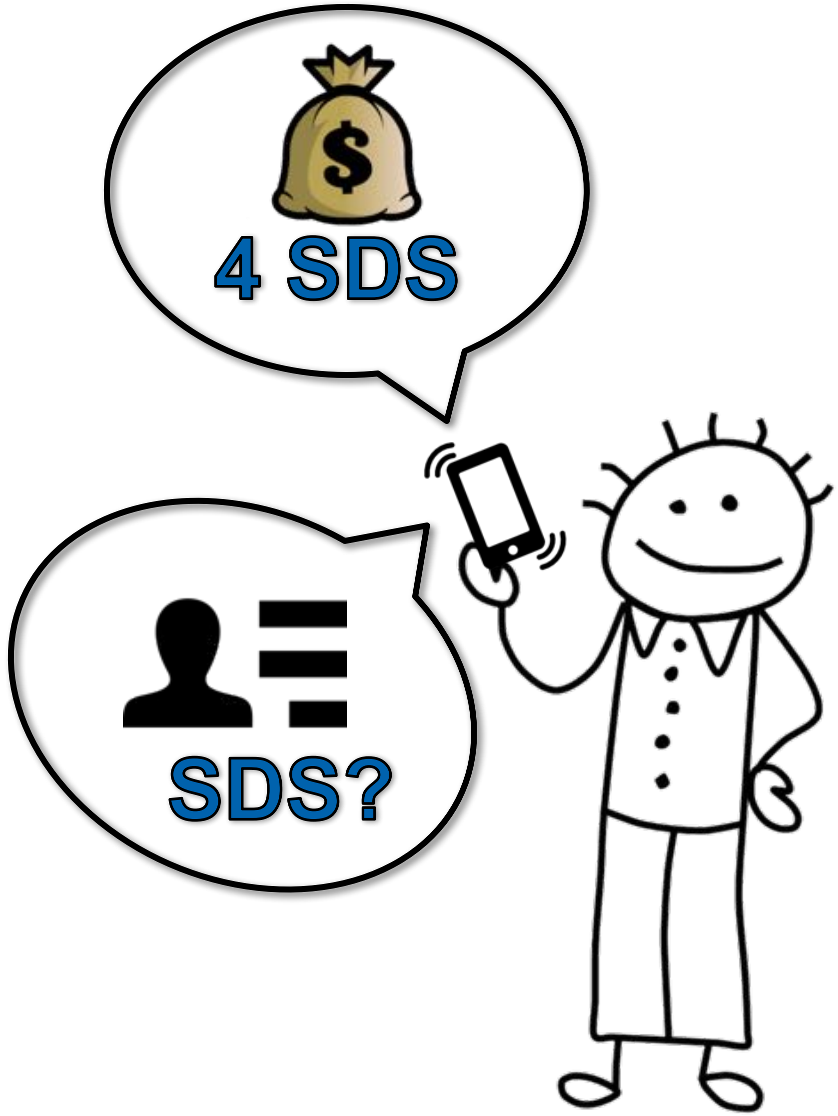
Lingering questions

- **Question:** it seems counterproductive to reserve data for testing. Isn’t there a better way?
- **Answer:** Why yes, yes there is ➔ today’s class ☺

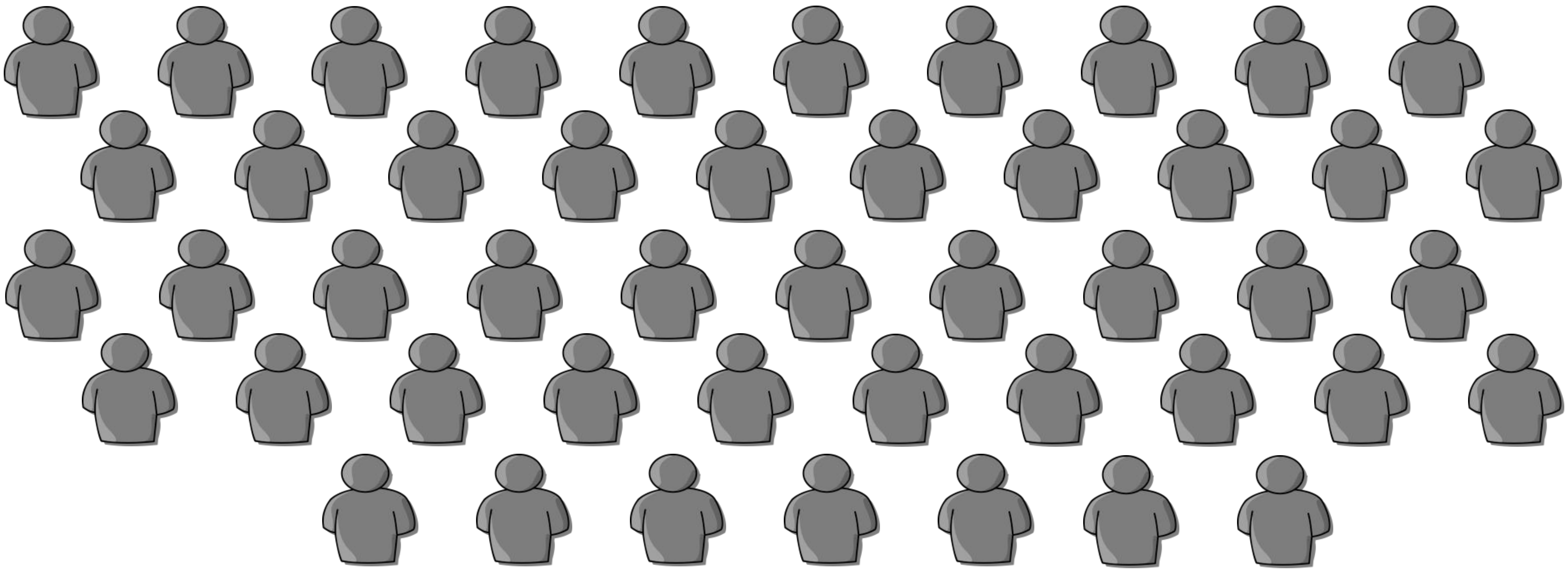
Outline

- Evaluating models using **resampling**
  - Running example
  - Cross-validation
  - Bootstrap
- Bootstrap activity
- Lab

Running example

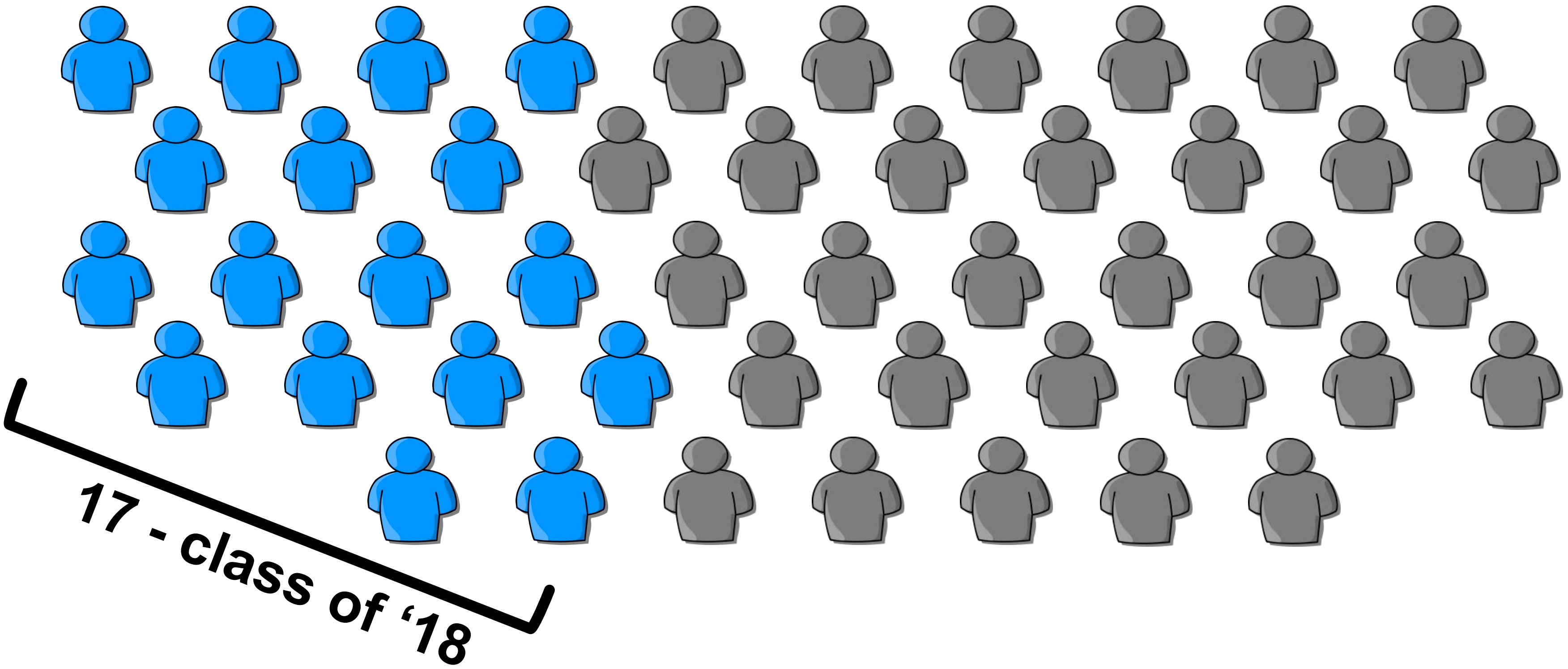


Running example

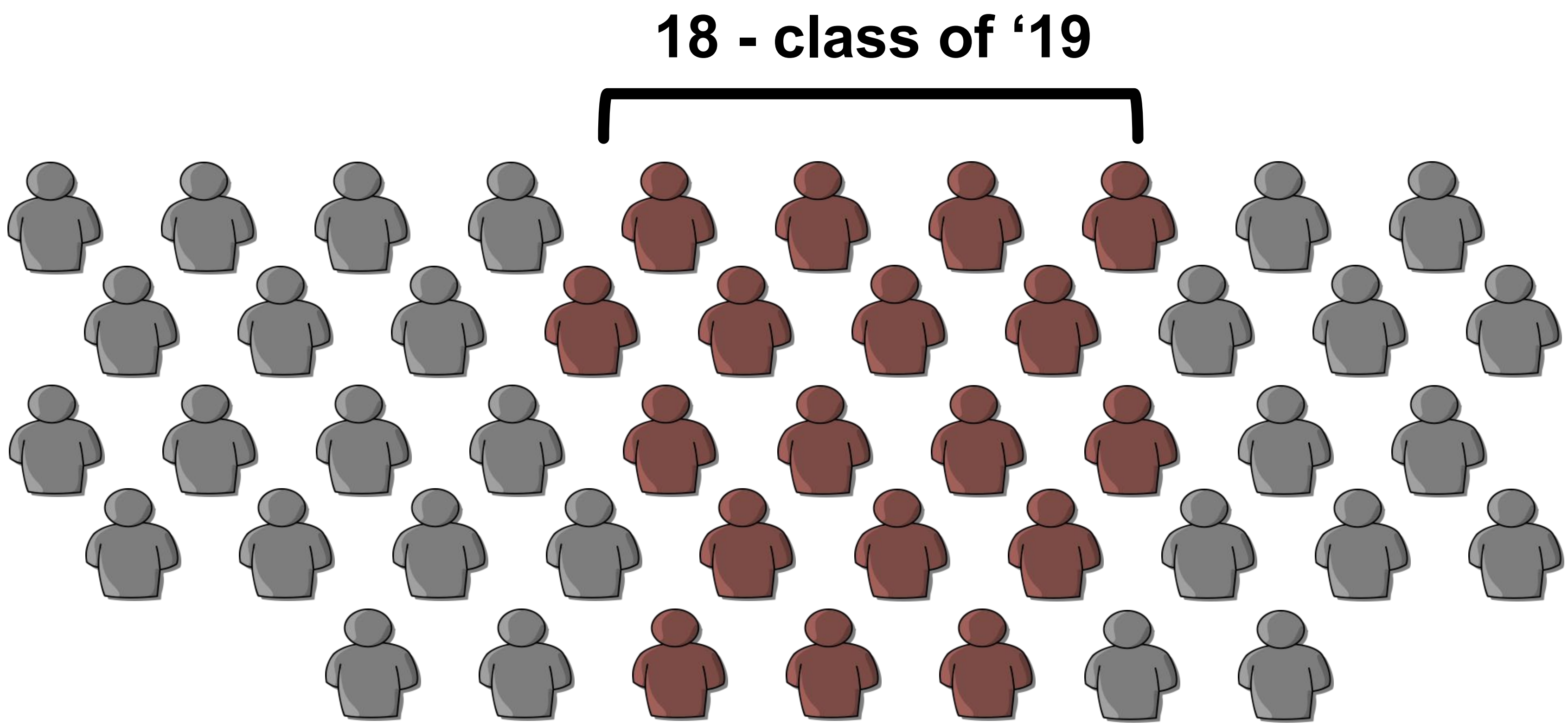


47 students in SDS 293

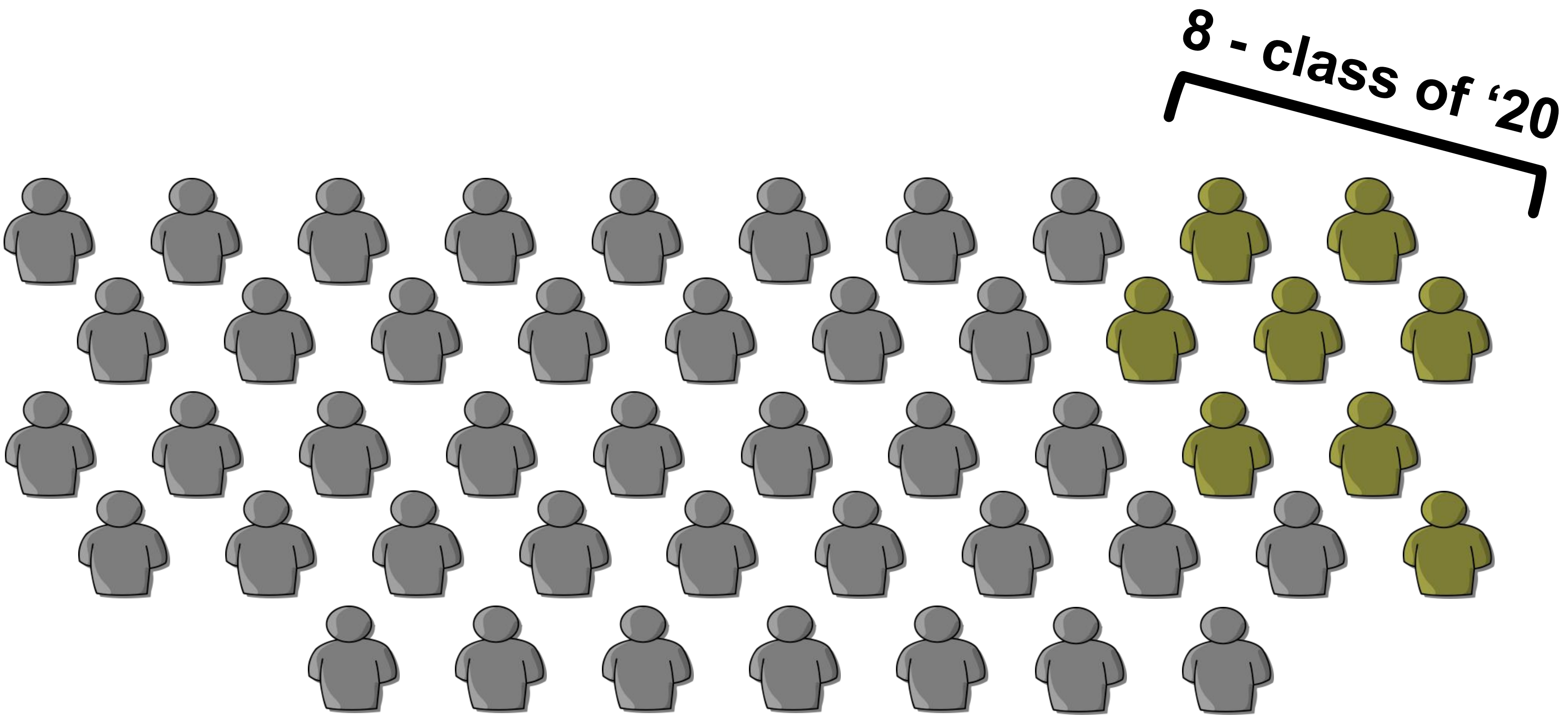
Running example



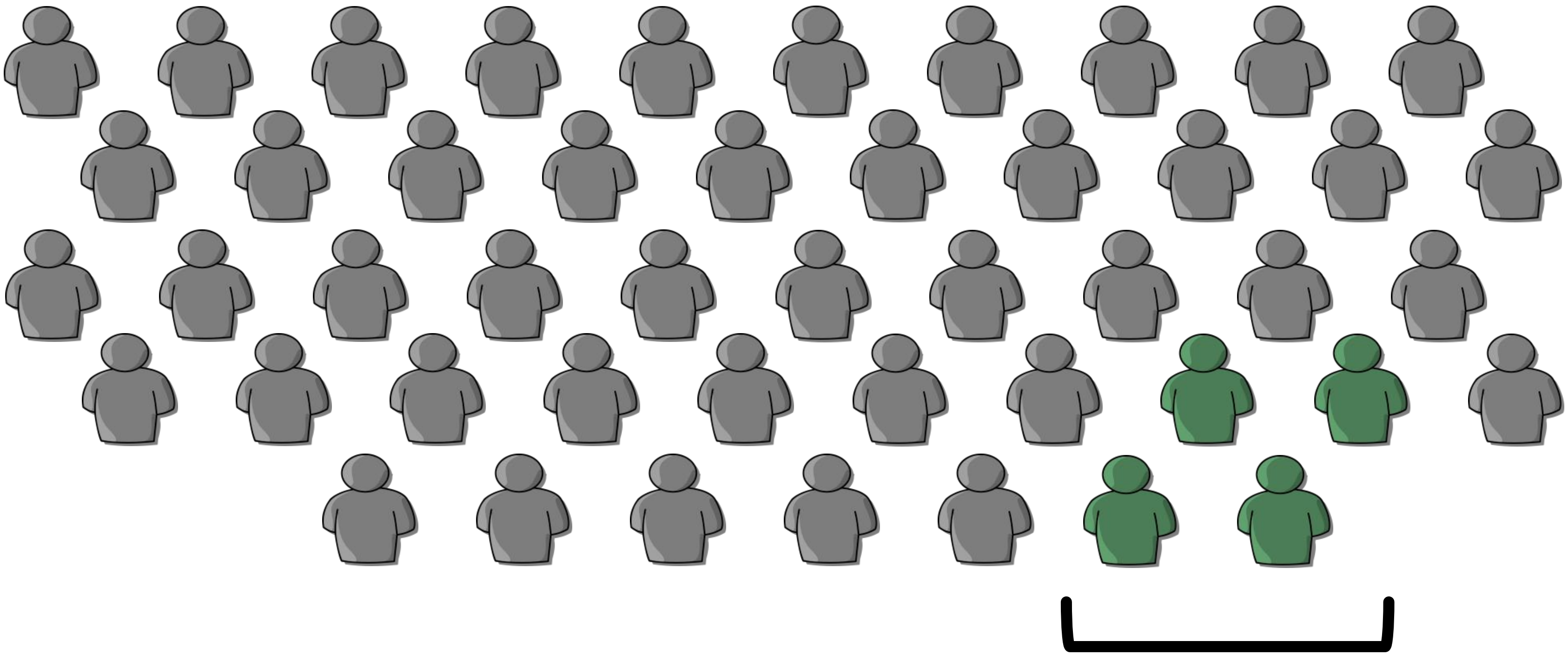
Running example



Running example

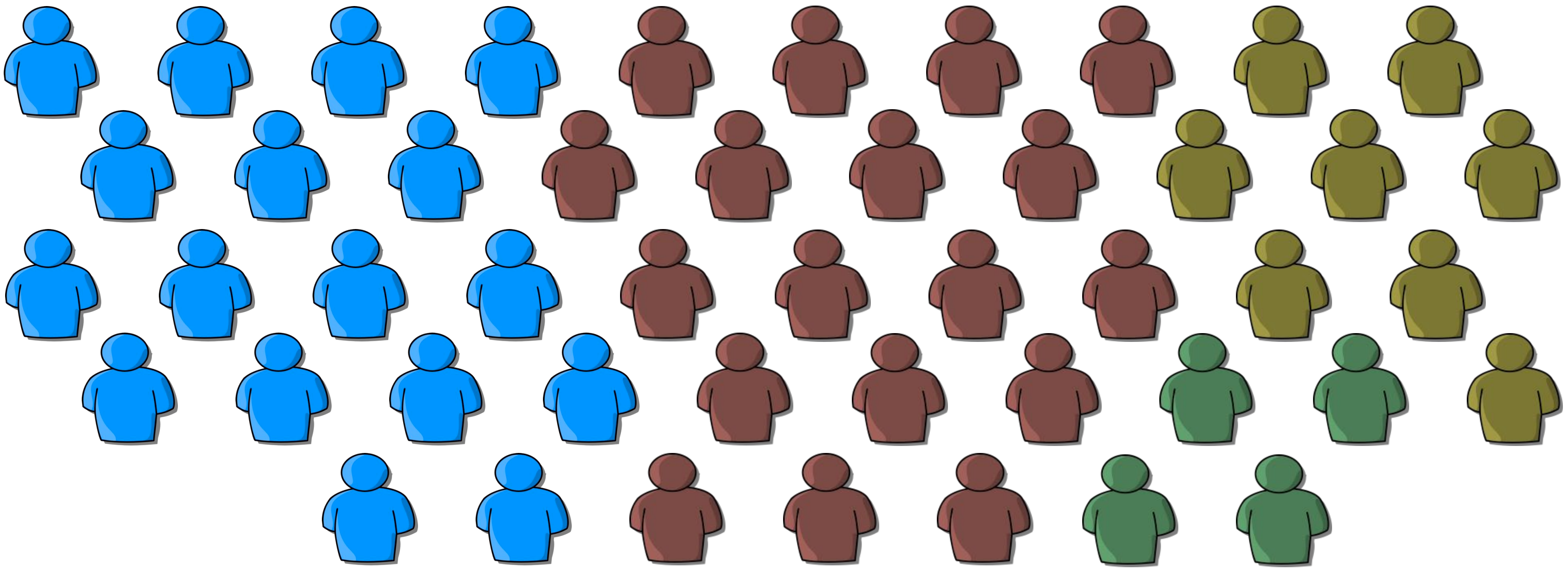


Running example



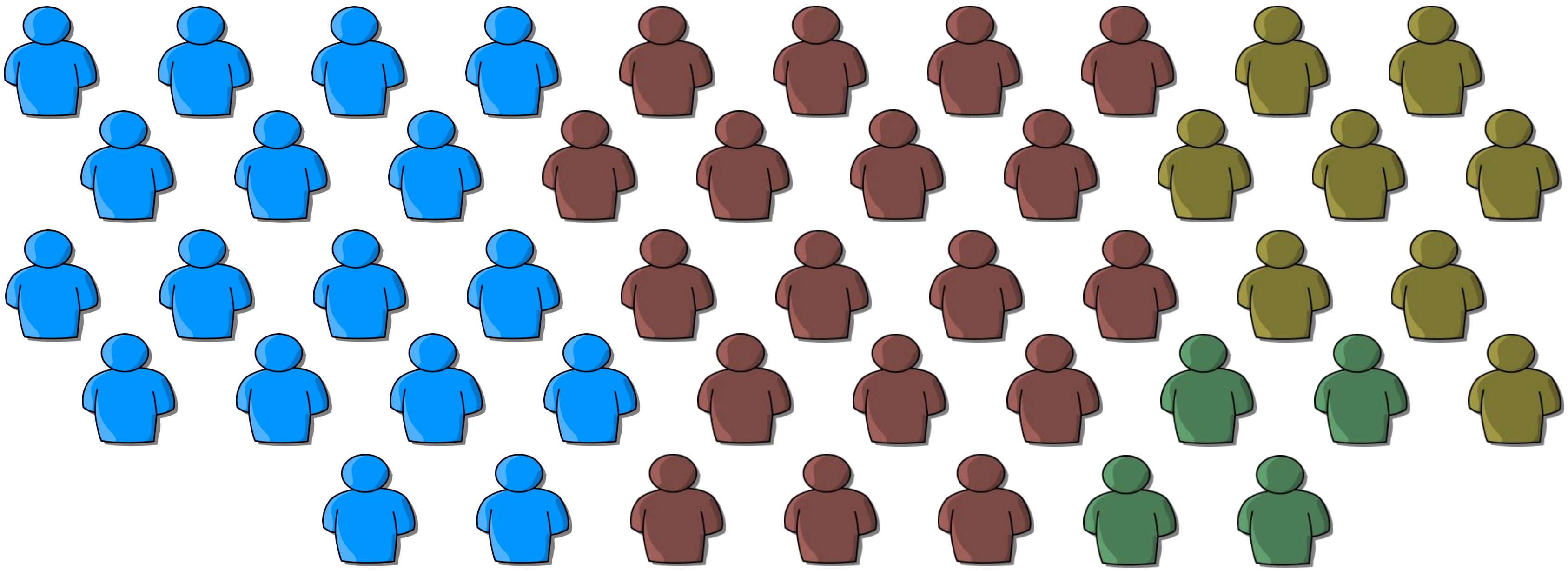
4 – ’21, PB, HS, etc.

Running example



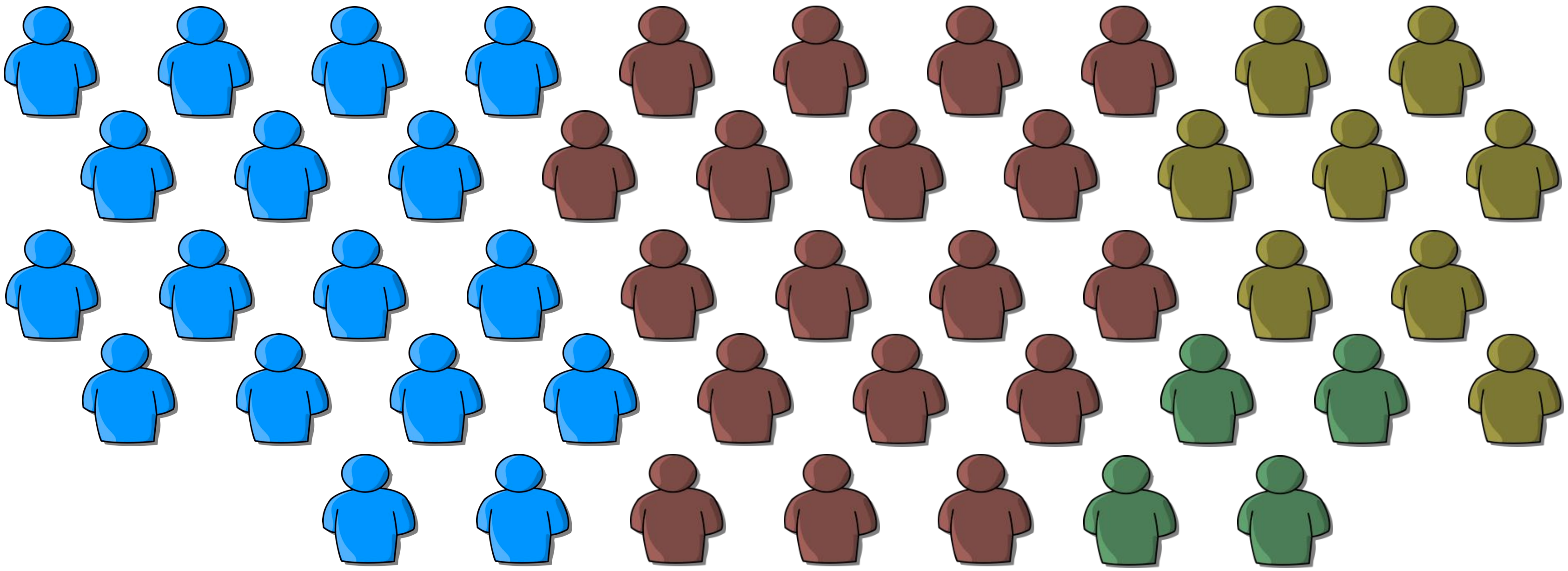
$$MODEL(\texttt{SDS293}) \sim \texttt{SDS Program}$$

Running example



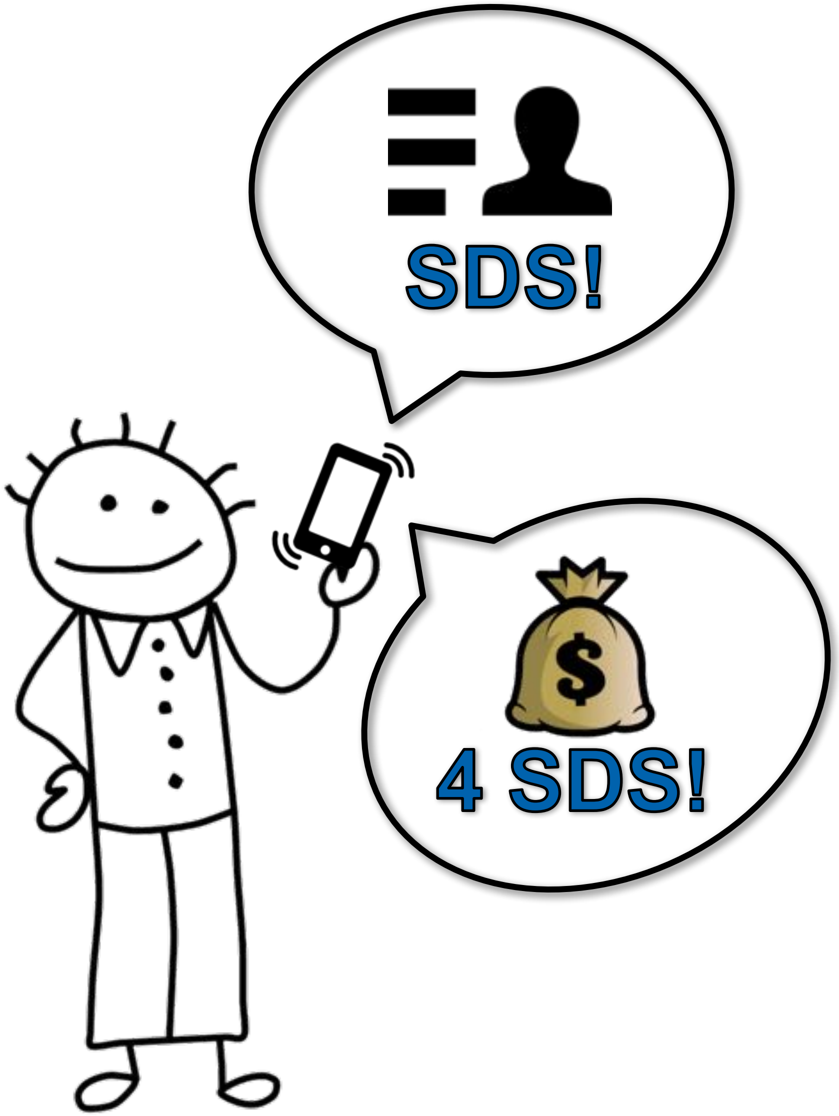
training set

Running example



test (“validation”) set

Running example



Discussion

- There are several issues with “validation set” approach
- Two big ones are:
  1. The **test error rate** depends on which observations we used for training vs. testing
  2. We’re only training on a **subset** of the data

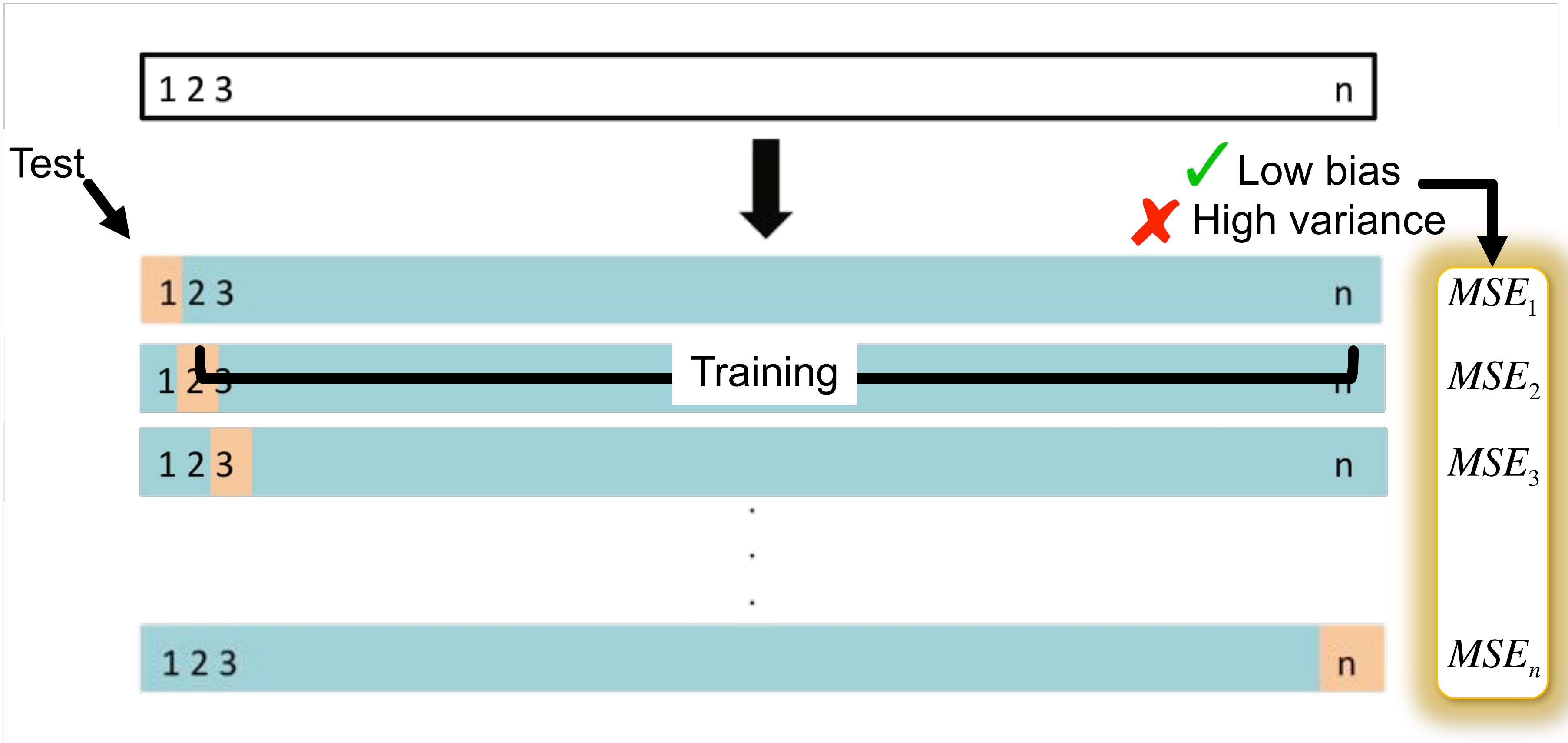
We need a new method…

Cross-validation

- **Goal 1:** avoid sensitivity to test set selection
- **Goal 2:** train on as much data as possible

Big idea:

Leave-one-out cross-validation (LOOCV)



$$CV_{(n)} = avg\left(MSE_i\right)$$

✓ Low bias
✓ No variance

Discussion

- LOOCV is extremely general, and can be used with **any** kind of predictive modeling
- **Question:** what's the catch?
- **Answer:** fitting n models could be awfully expensive…

Cheap LOOCV for least-squares regression

- **Good news:** there's a special trick when we're working with least-squares regression models
- **Fun fact:** remember when we talked about *leverage*?

how much an observation **influences its own fit**

$$h_i = \frac{1}{n} + \frac{\boxed{\left(x_i - \overline{x}\right)^2}}{\boxed{\sum_{j=1}^{n}\left(x_j - \overline{x}\right)^2}}$$

big values = "outliers in the predictors"

- Can use h_i along with MSE to calculate what the LOOCV error *would* be **without ever actually performing it**

Cheap LOOCV for least-squares regression

- Fit a least-squares regression model on the **full dataset**
- Calculate the MSE of the model, but divide each residual by 1 minus the point's **leverage**:

$$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \rightarrow \boxed{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2 = CV_{(n)}}$$

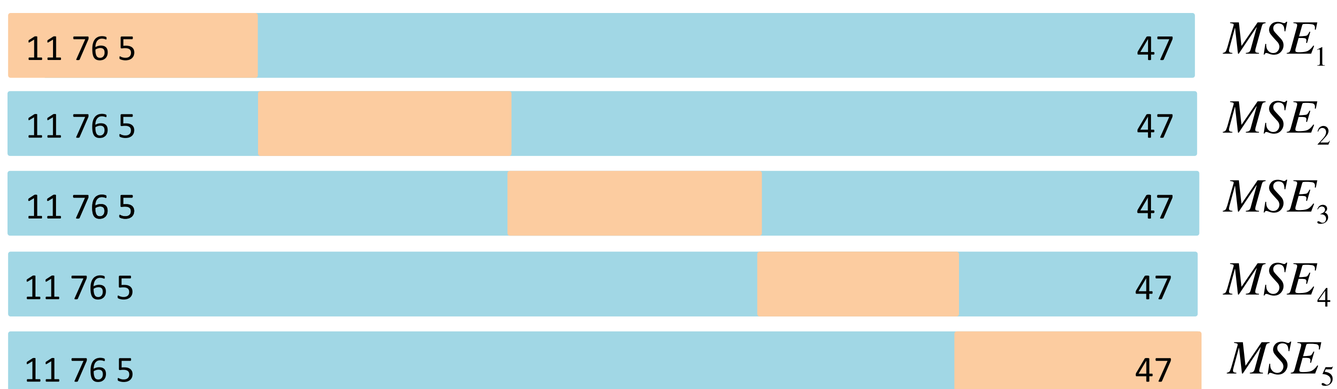
- This normalization "inflates" high leverage points by *just* the right amount…

Note: this sadly only holds for least-squares regression

K-fold cross-validation

- LOOCV is often too expensive on large datasets, but the same idea works even if we can't build n separate models
- Start by randomly dividing the data into k non-overlapping groups (or **folds**)*

- And then round-robin:



$$CV_{(k)} = avg\left(MSE_i\right)$$

*Empirical evidence indicates that k = 5 or 10 usually works well

Cross-validation for choosing variables

- Recall that in regular ol' regression, adding parameters never increases our error even if they're useless (why?)
- **Question**: will a cross-validated model have the same problem?

Cross-validation for choosing variables

- **Answer**: generally not – cross-validated error will tend to:
  - decrease with the addition of useful predictors
  - increase with the addition of junk predictors

Cross-validation for classification

- So far we've only talked about regression
- **Question**: what do we need to do to make this work for classification?

Cross-validation for classification

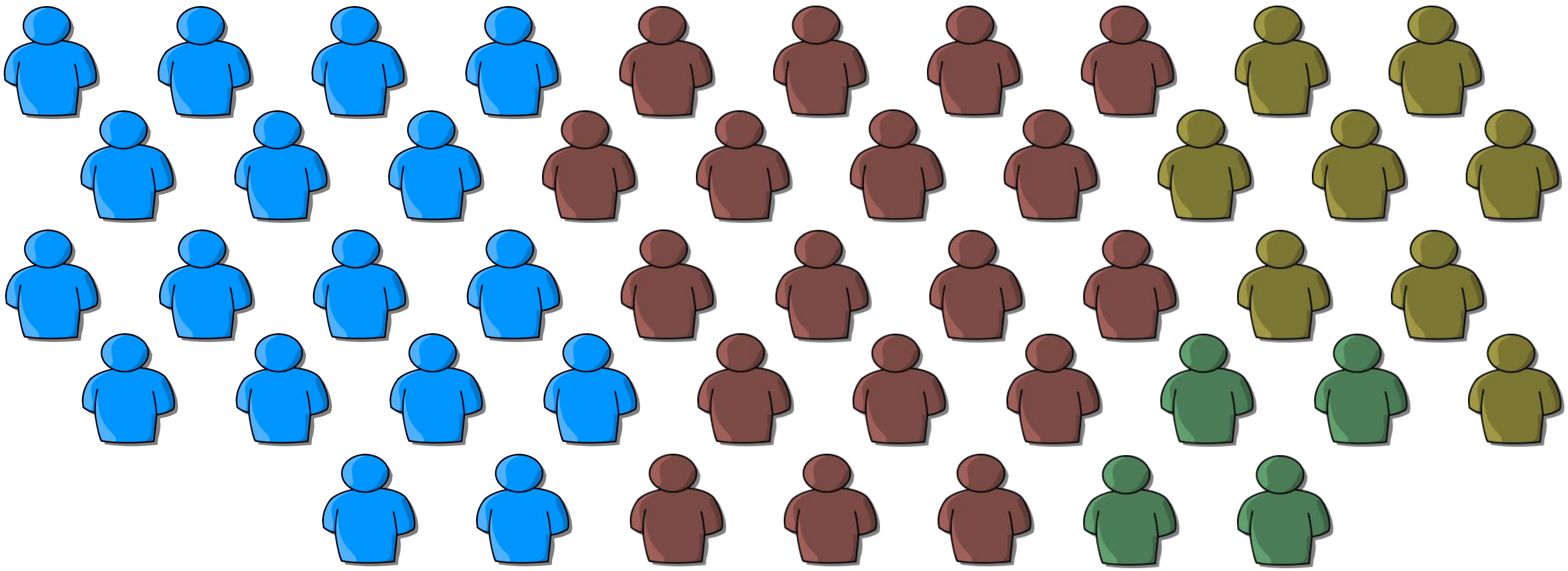
- **Answer:** Good news! Not much needs to change
- We just need to tweak our measure of error

$$CV_{(k)} = avg\left(MSE_i\right) \rightarrow avg\left(TE_i\right)$$

where:

$$TE = avg\left(I\left(y \neq \hat{y}\right)\right)$$

Back to our example

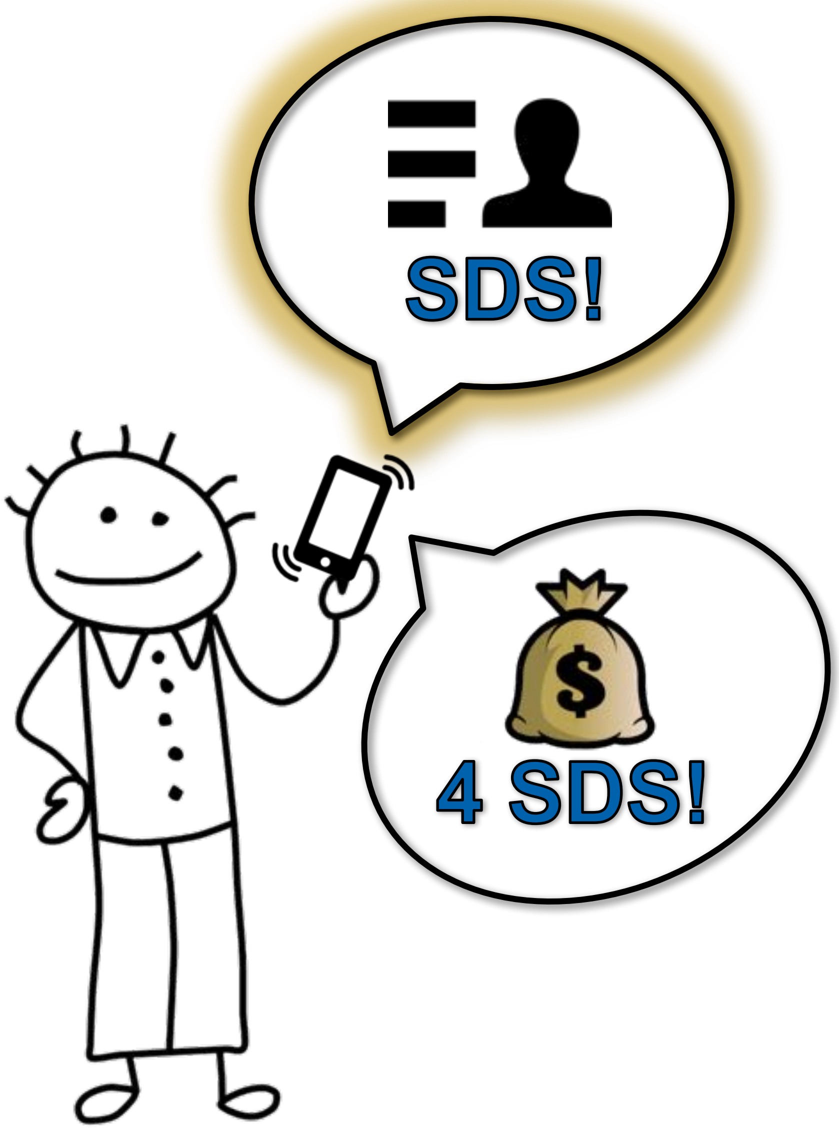


$CV_{47}(\texttt{SDS293}) \sim \texttt{SDS Program}$

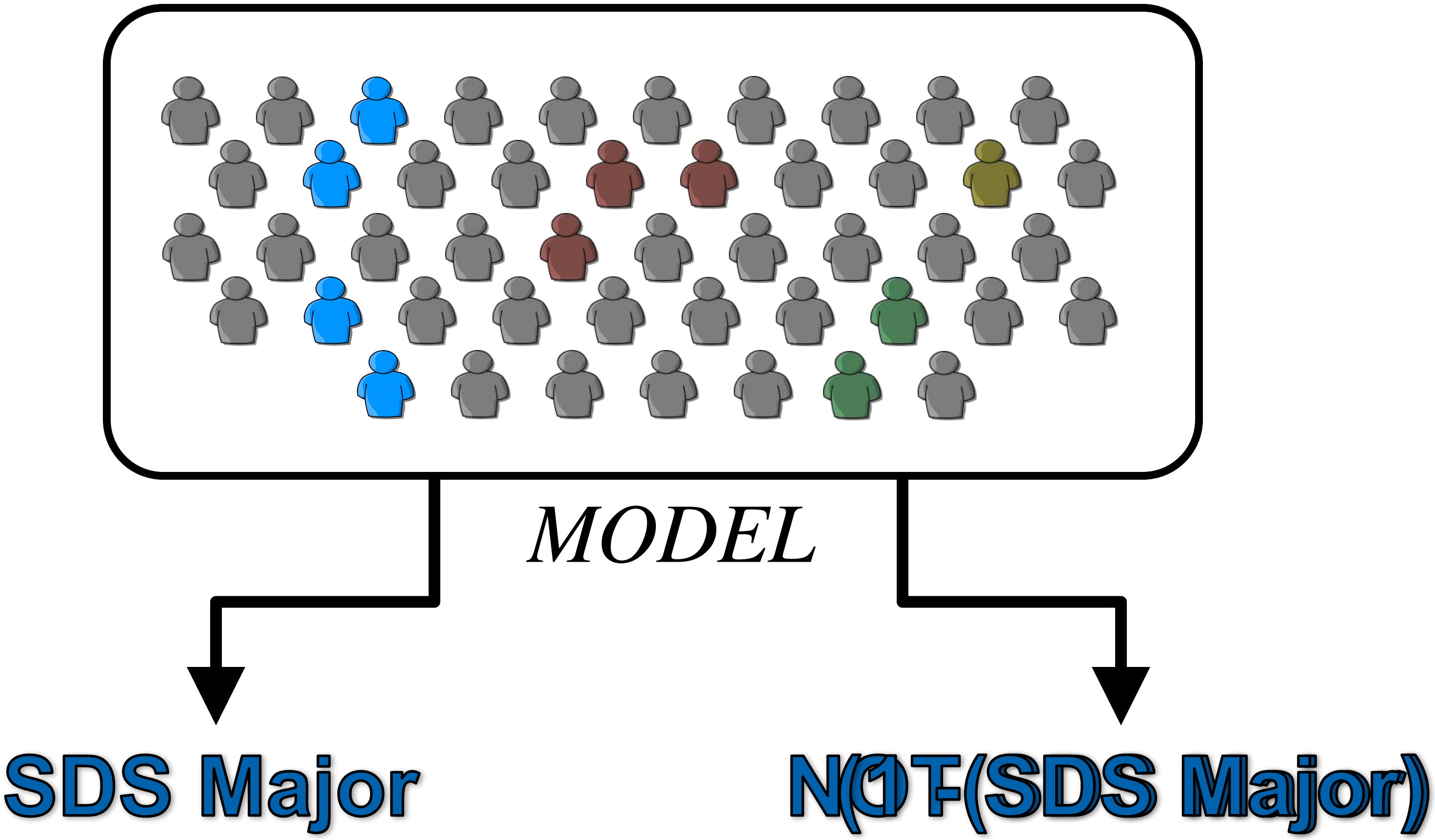
Discussion

- **Question:** assuming nothing strange shows up during cross-validation, what do I know about my model?
- **Answer:** none of the observations in my sample have *undue influence* on the model

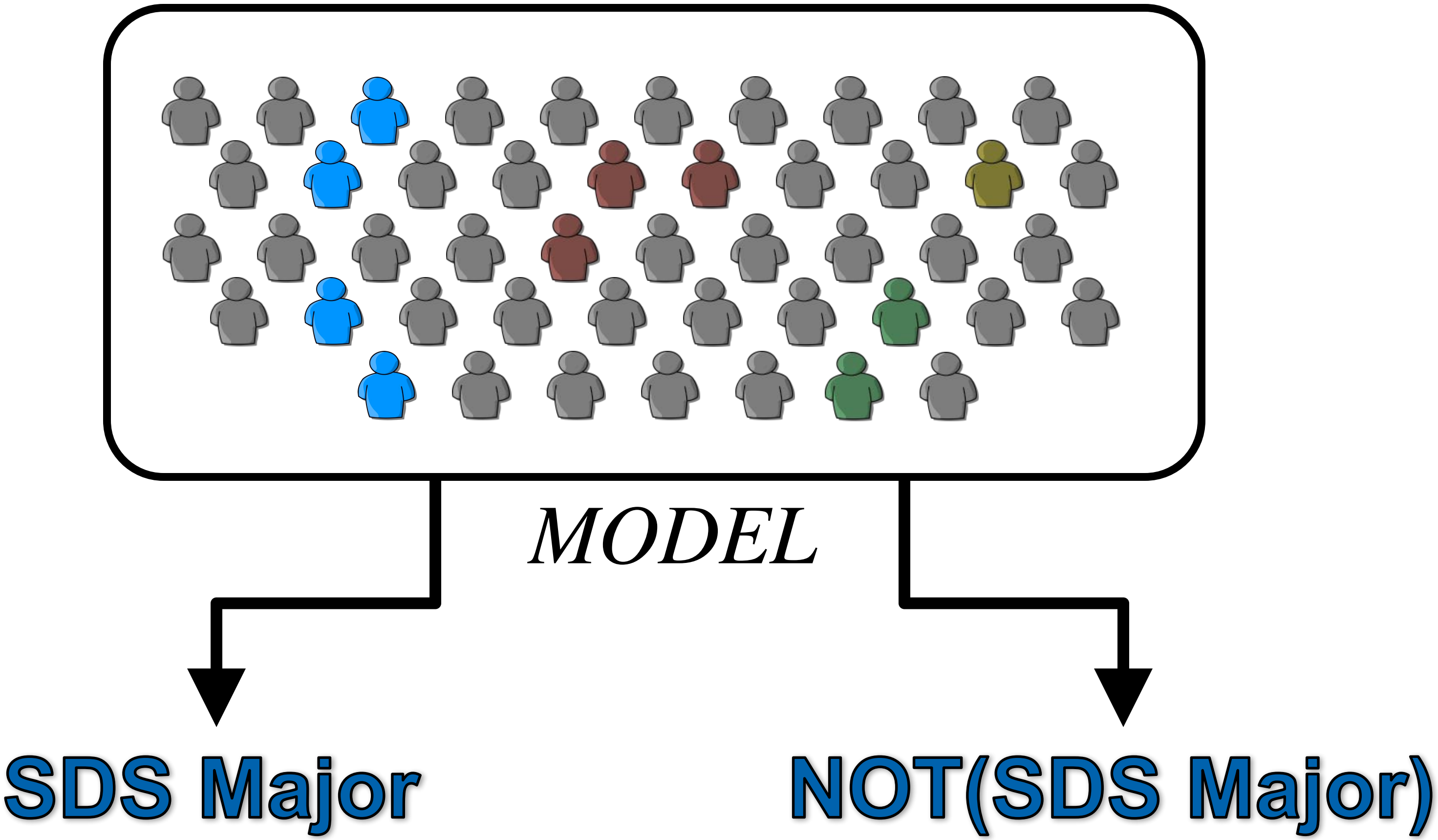
Progress!



Estimating SDS Majors



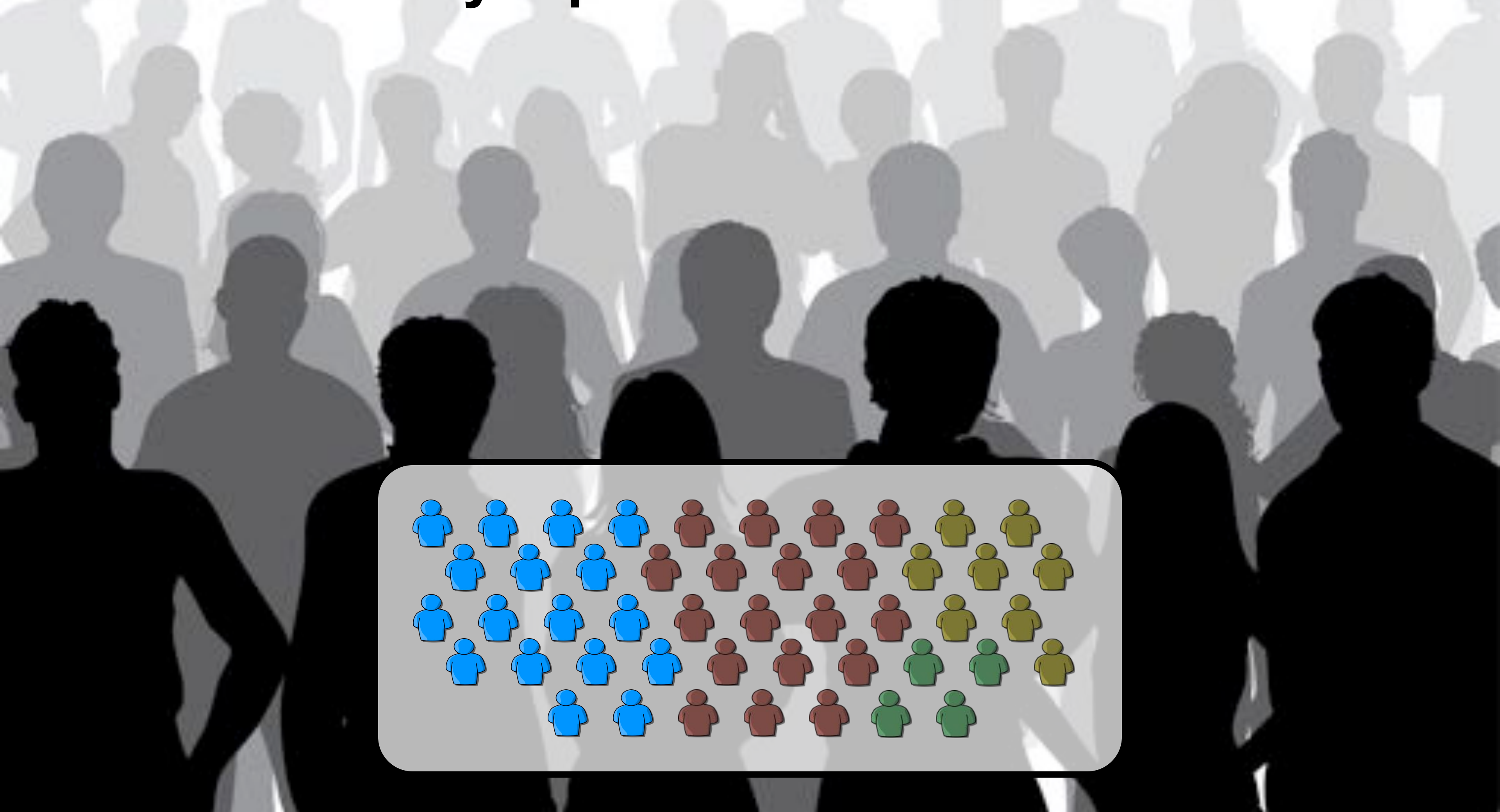
Estimating SDS Majors



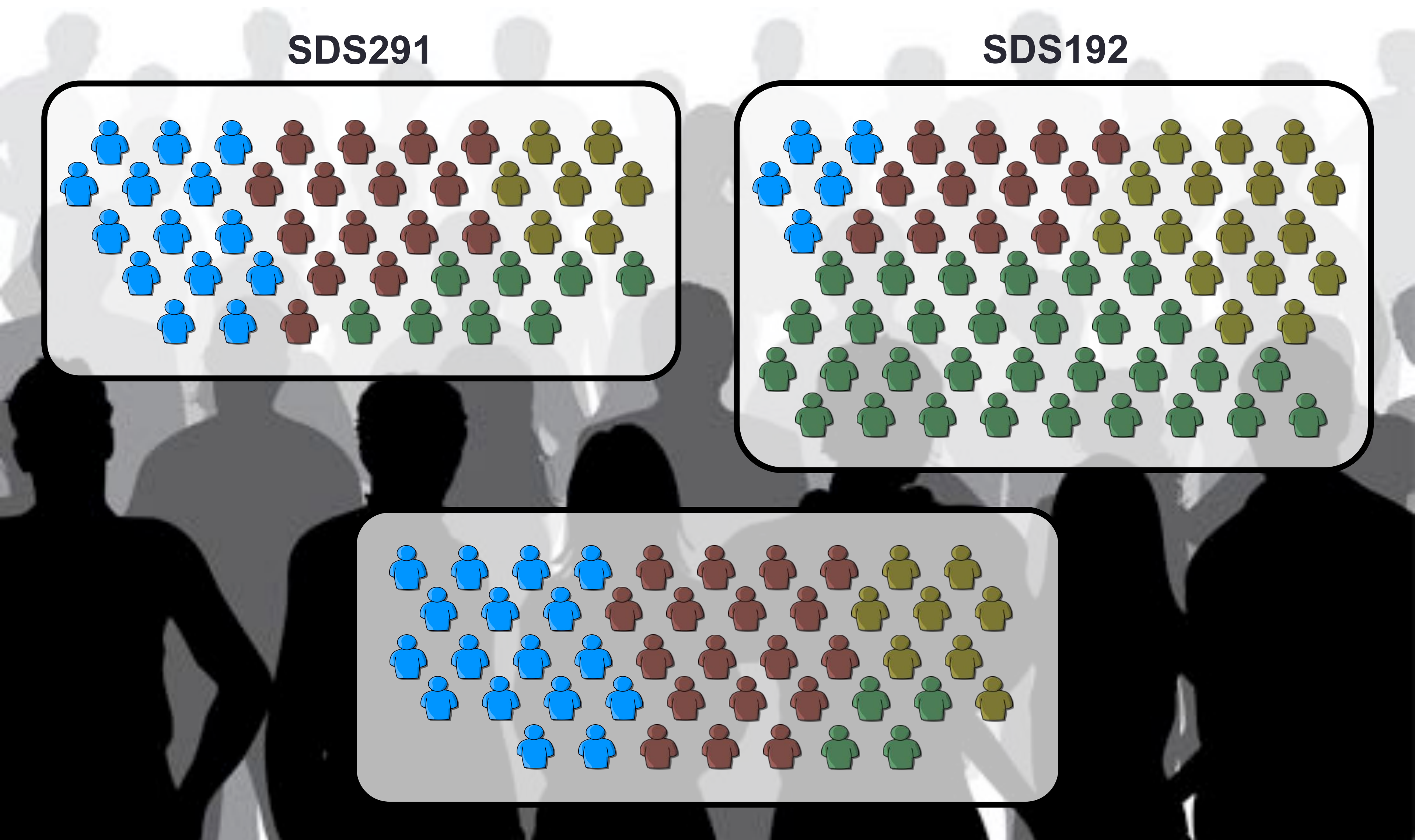
$$\Pr\left(\textbf{SDS Major}\right) \in \left[\min, \max\right]$$

One problem…

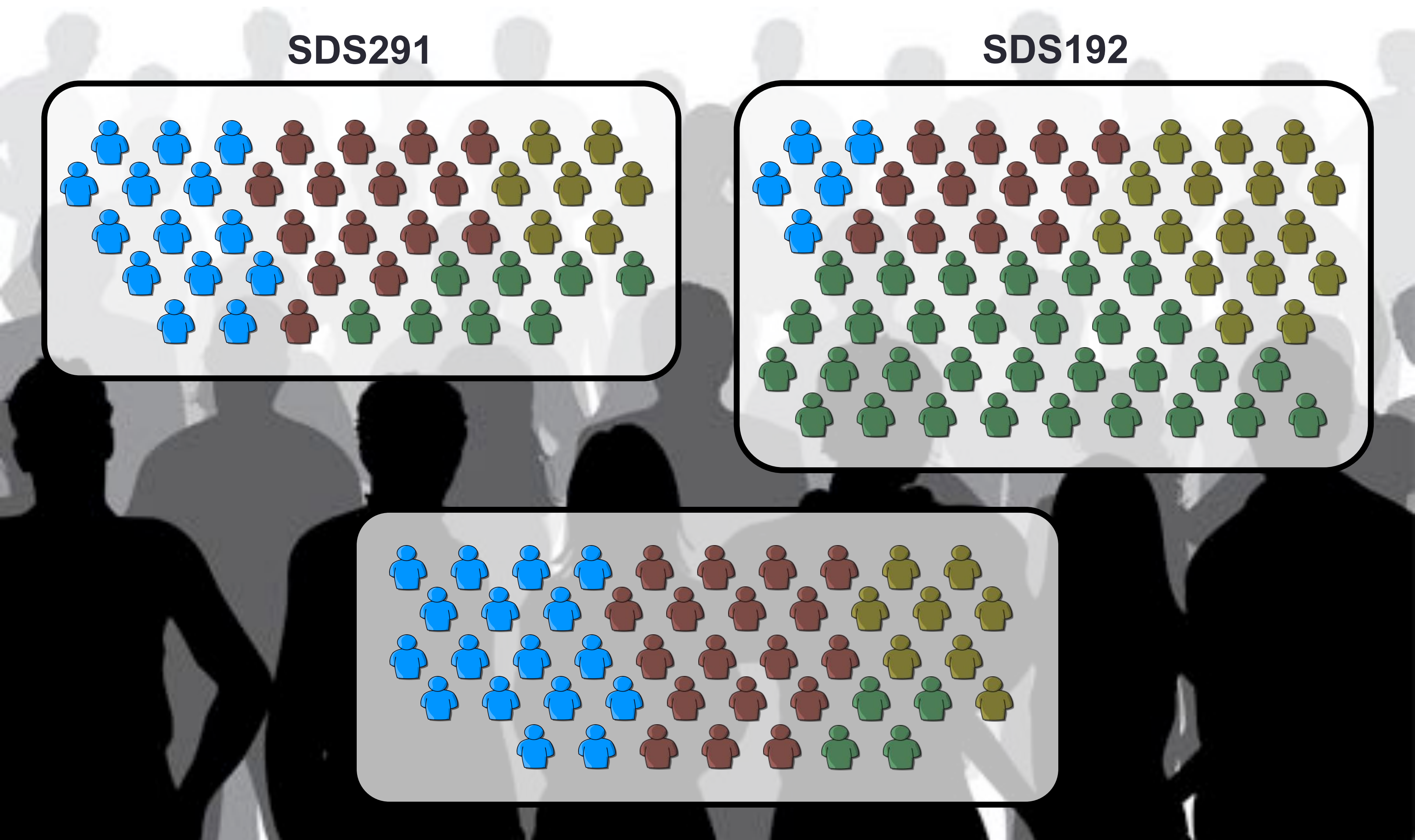
How can we be sure the model we built on SDS293 **accurately represents SDS *as a whole*?**



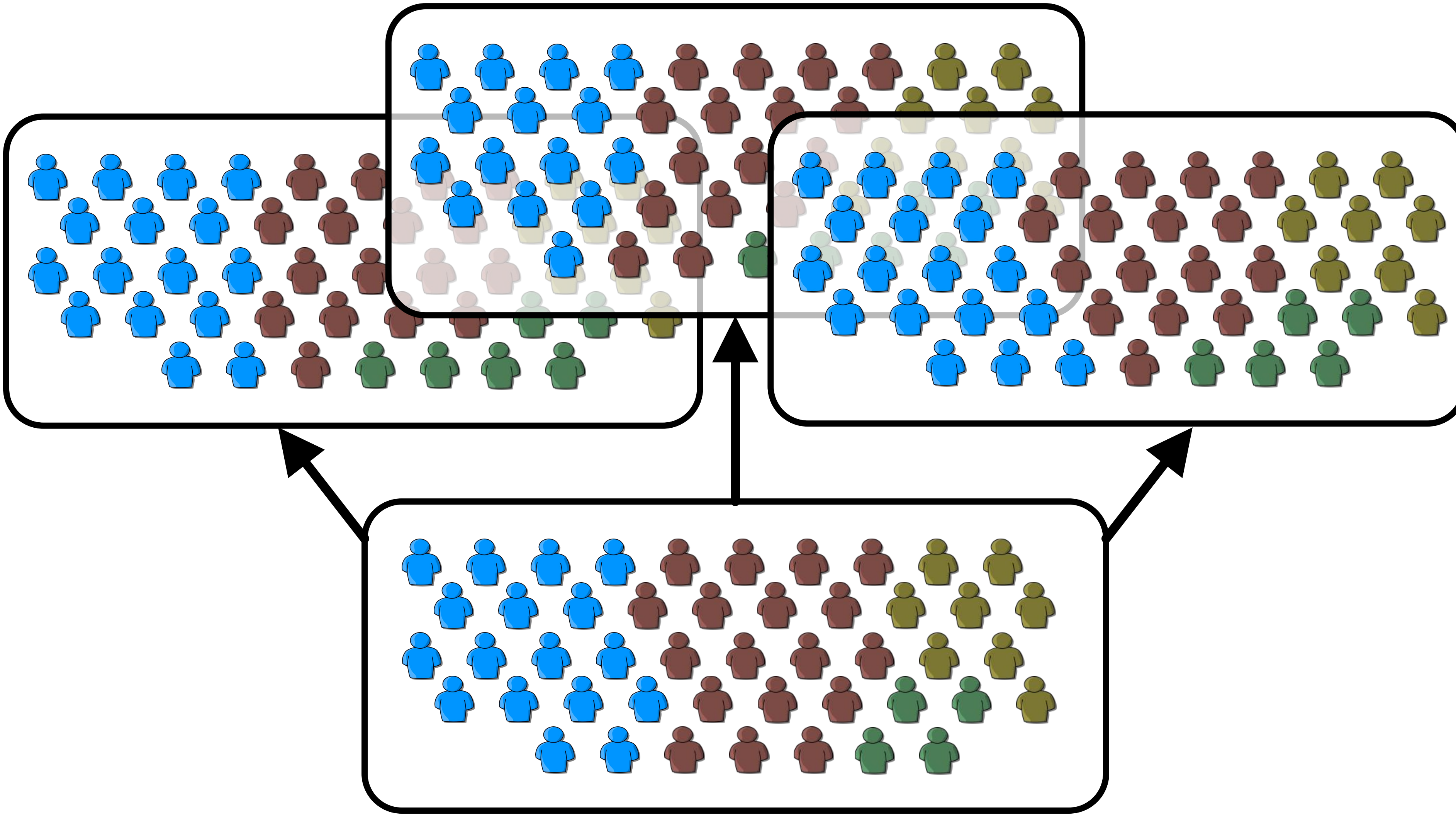
In a perfect world…



The real world



Can we fake it?

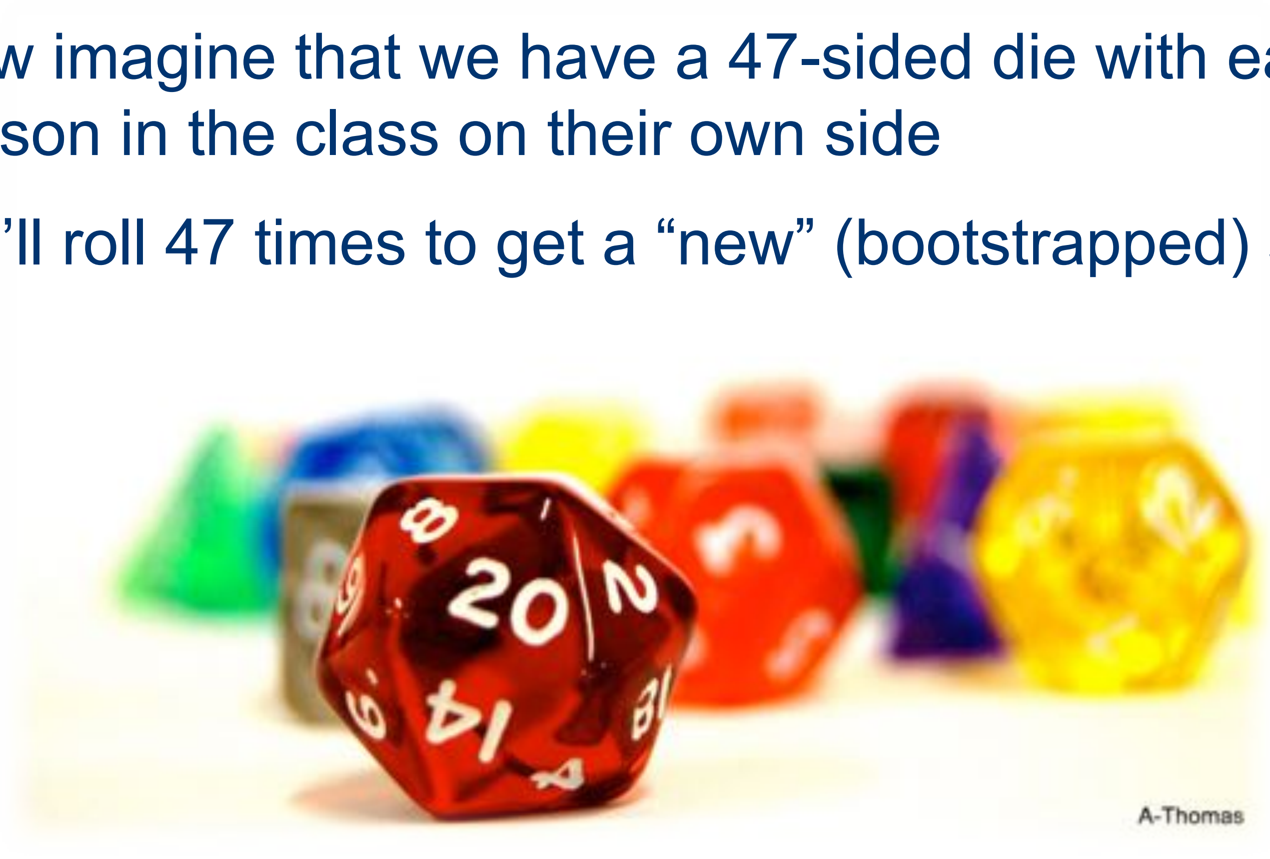


"Bootstrapping"

- Resample the original data to get a "new" dataset
- **Assumption**: original data was uniformly sampled
  - Each obs. is equally likely to appear in the resampled dataset
- Perform resampling **with replacement**
  - Each obs. may appear more than once in the resampled dataset

Bootstrap activity

1. Imagine everyone has an envelope containing several copies of their "predictors" (i.e. class year)
2. Now imagine that we have a 47-sided die with each person in the class on their own side
3. We'll roll 47 times to get a "new" (bootstrapped) sample



Bootstrap estimates

- Let’s say we’ve generated some large number of bootstrapped datasets
- Big idea: generate estimates and calculate the standard error **across all of them** (*bootstrap error*)
- This should help us to better capture the variation in the population (why?)

Discussion

- **Question**: when is bootstrapping useful?
- **Answer:** Two cases:
    1. When the **sample size is too small** for straightforward statistical inference. If we know the underlying distribution, bootstrapping lets us account for any distortions caused by the specific sample.
    2. When the underlying **distribution is complicated or unknown**. Bootstrapping is an *indirect* method to assess the properties of the distribution and estimate parameters that are derived from it.

Lab: cross-validation and bootstrap

- To do today's lab in R: **`boot`**
- To do today's lab in python`: <nothing new>
- Instructions and code:

  [course website]/labs/lab7-r.html

  [course website]/labs/lab7-py.html

- Full version can be found beginning on p. 190 of ISLR

Coming up

- Monday: **linear model selection**
- A1 and A2 have been graded and returned
- A2 solution posted
- A3 due **tonight by 11:59pm**