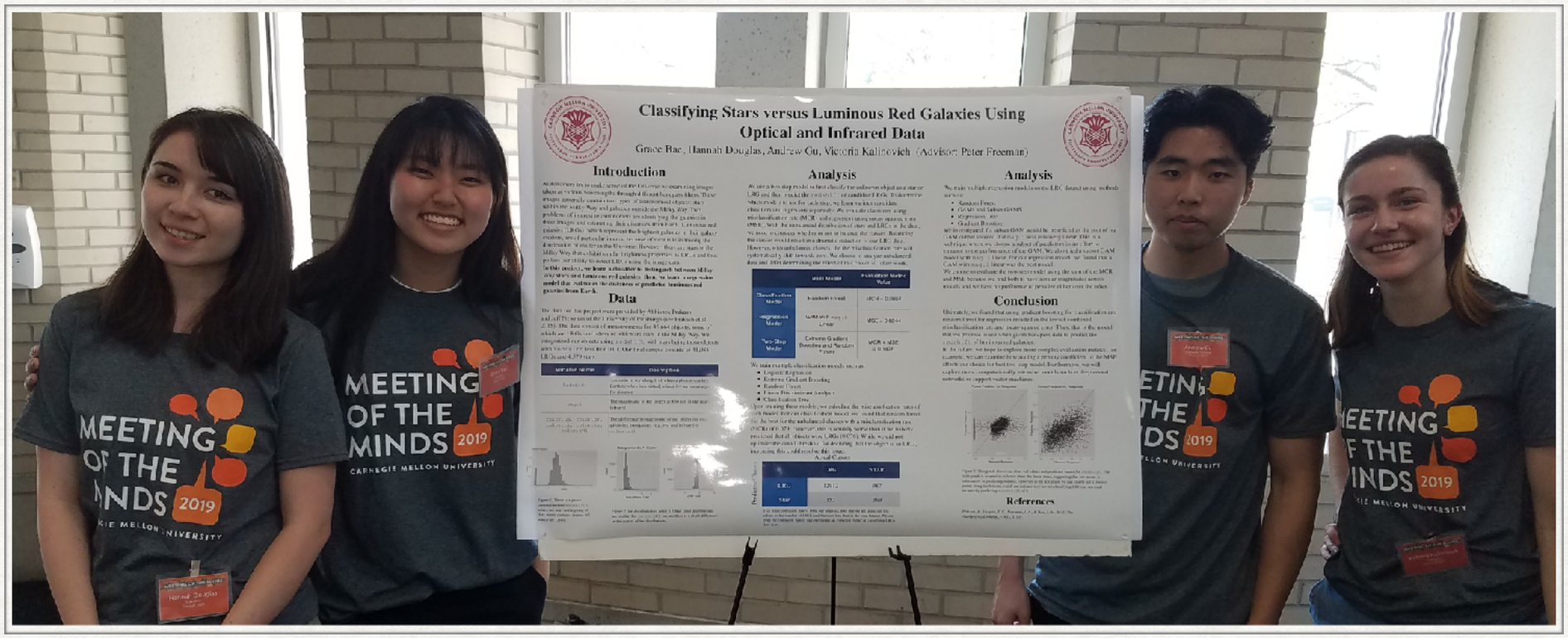
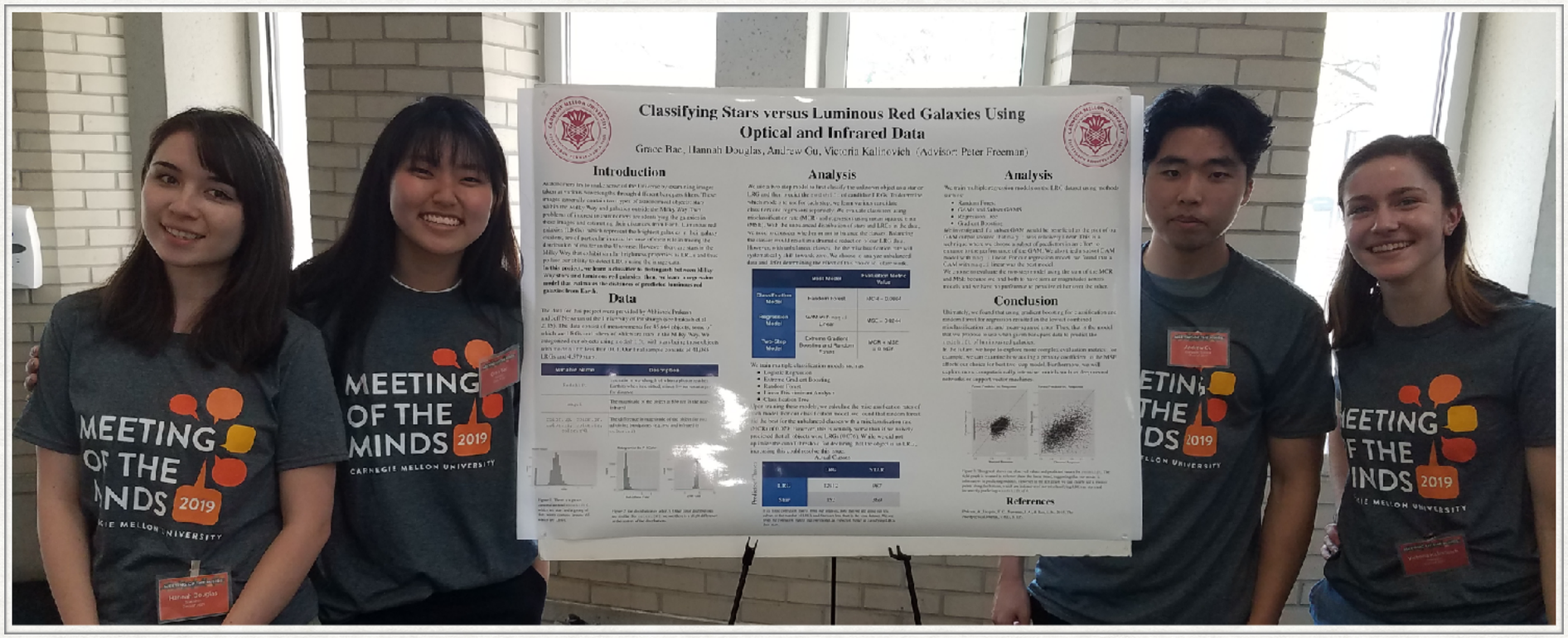
# Introducing Early Undergraduates to Statistical Practice: How You Can (and Why You Should) Provide Such Opportunities at Your Institution




A group of 36-290 students presenting their final project (May 2019)

Peter E. Freeman
Department of Statistics & Data Science
Carnegie Mellon University

# (Future) Practitioners Gotta Practice!




A group of 36-290 students presenting their final project (May 2019)

Peter E. Freeman
Department of Statistics & Data Science
Carnegie Mellon University

# The Proposal:

## We Should Introduce Freshman and Sophomore Statistics Majors to Statistical Practice


Early statistical practice, as practiced at CMU, involves introducing and contextualizing statistical learning (*) methods and focuses on application, not theory, and that is meant to complement later mathematical statistics classes (and…importantly…not replace any!).

(*) Statistical learning: determining associations between variables via linear or logistic regression, trees, random forest, etc. This exact focus is motivated by circumstance: CMU capstone experiences generally involve learning models, given observational data. In the end, the focus is not as important as the implementation: “all curriculum is local.”

# Why Should We Provide Early Statistical Practice?

*It fits naturally with current curricular recommendations…*

Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016

1. Teach statistical thinking.
   - Teach statistics as an investigative process of problem-solving and decision-making.
   - Give students experience with multivariable thinking.
2. Focus on conceptual understanding.
3. Integrate real data with a context and purpose.
4. Foster active learning.
5. Use technology to explore concepts and analyze data.
6. Use assessments to improve and evaluate student learning.

American Statistical Association
Undergraduate Guidelines Workgroup

Curriculum Guidelines for **Undergraduate Programs** in Statistical Science

Providing students with a strong foundation in statistical methods and theory is critically important for all undergraduate programs in statistics. These skills need to be introduced, supported, and reinforced throughout a student's academic program, beginning with introductory courses and augmented in later classes[25]. Such scaffolded exposure helps students connect statistical concepts and theory to practice.

*…and it acts to "minimize prerequisites to research" (Brown & Kass 2009)*

# Why Should We Provide Early Statistical Practice?

*It enhances student confidence.*

"[A] growing number of our own undergraduate students [at CMU], though well-trained, were reporting not feeling 'ready' to enter the job market with just a bachelor's degree."

- Greenhouse & Seltman (2018)

*It enhances future learning.*

"Hard yes. I learned about bootstrapping a year before I was 'supposed' to…and it made it wayyyy easier to do…assignments. I have explained bootstrapping probably three times to some stats friends…I also got a head start on SVMs, trees, and random forests…[i]t is a LOT easier to code a decision tree when I don't have to, at the same time, learn what a decision tree is."

"Though I wasn't able to fully understand every step of the methods when taking [the course] (like k-means clustering), the exposure helped me learn and understand faster in [later classes]."

- Former early practice students

*It is attractive to future employers.*

"In the phone call that gave me my offer for my internship this summer…they specifically mentioned that they liked the research I had done."

"[A]lmost all the interviewers asked about the research projects…and wanted me to go into detail."

- ibid.

*…and it helps students determine whether they "like" statistics.*

# How Should We Provide Early Statistical Practice?

*We should strive to **reduce cognitive overload** and keep the focus on practice!*

*So…start with a minimalist introduction to coding…e.g.,*

RStudio® + R Markdown

+ dplyr + ggplot2

*…that is reinforced (and expanded upon) in subsequent lab exercises.*

# How Should We Provide Early Statistical Practice?

*Provide fully curated data with specified research questions.*

Branch: master ▾ **36-290** / PROJECT_DATASETS / **KEPLER_OBJS_INTEREST** /

Create new file | Find file | History

**pefreeman** Tweaked README file. Latest commit 1b78e91 18 hours ago

..

| | | |
|---|---|---|
| README.md | Tweaked README file. | 18 hours ago |
| koi.Rdata | Added KOI dataset. | 18 hours ago |
| koi.csv | Added KOI dataset. | 18 hours ago |

The dataset in this directory contains 18 measurements for each of 9177 KOIs (or *Kepler* Objects of Interest). The measurements fall into four general groups:

| type | variables (all preceded with "koi_") |
|---|---|
| exoplanet orbit-related | period, eccen, sma, incl, dor |
| transit/eclipse-related | impact, duration, depth |
| exoplanet property-related | ror, prad, teq, insol |
| host star property-related | srho, steff, slogg, smet, srad, smass |

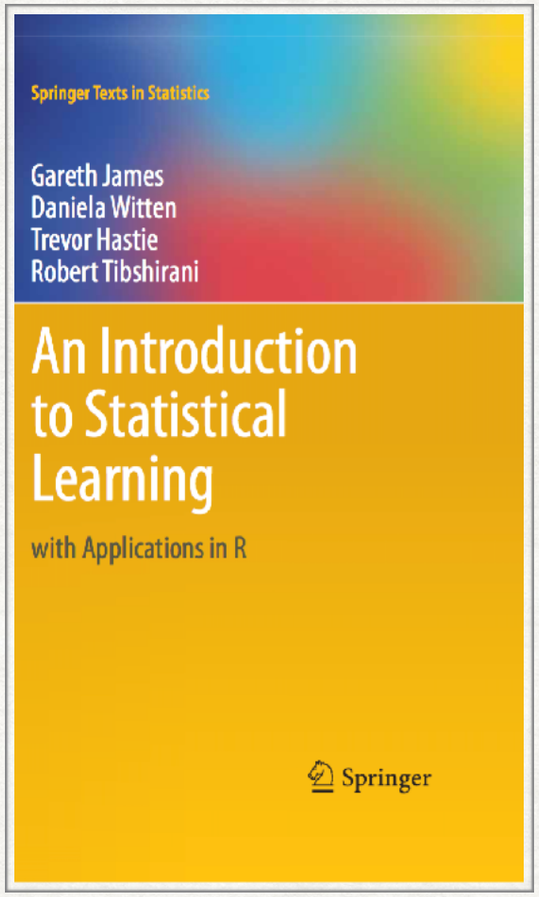
The definition of each measurement is given on this web page.

Research question: can you train a classifier that can effectively differentiate between `CONFIRMED` and `FALSE POSITIVE` exoplanetary candidates? Once you have done this, apply your model to the `CANDIDATE` data to inform NASA which candidates are most worthy of followup, independent observations.

(Note: the data I use are publicly available at github.com/pefreeman/36-290.)

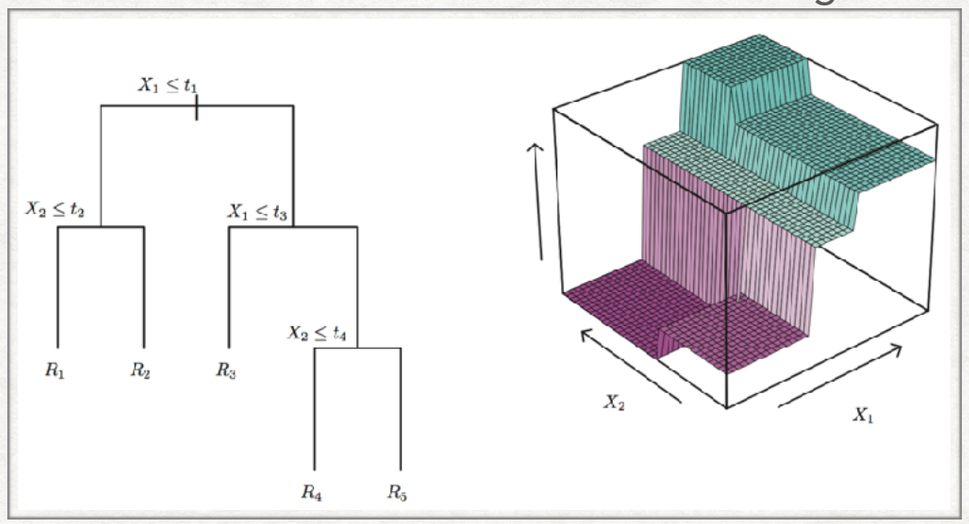
# How Should We Provide Early Statistical Practice?

*Use a textbook in which conceptual details are not obscured by mathematics.*




*ISLR* is an extremely good textbook around which to build an early statistical practice course: it provides qualitative contextual detail *and* lab exercises that students can use as a springboard.

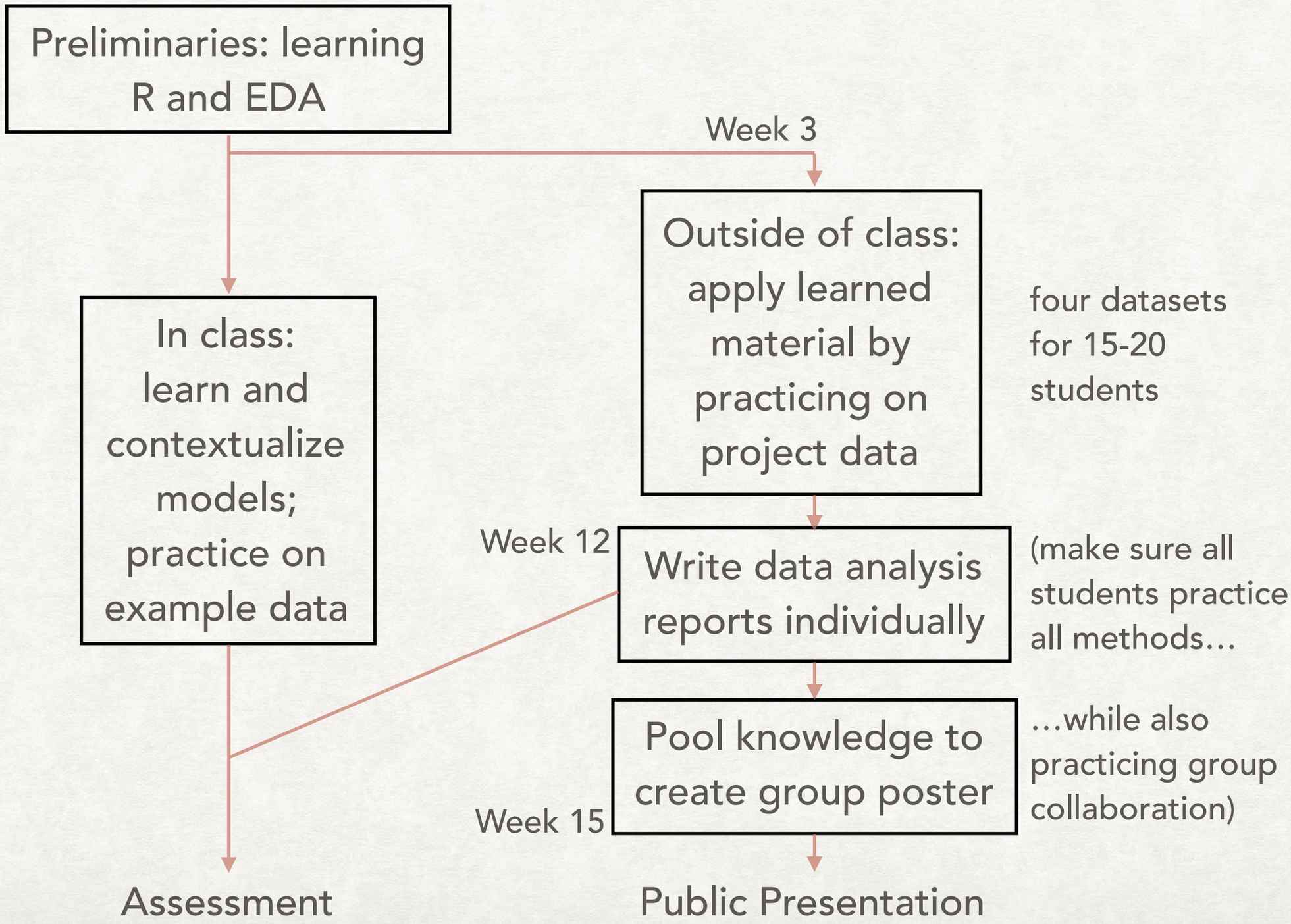
Figure 8.3




(e.g., we can teach students about how a tree is constructed while foregoing detail on how, algorithmically, split points are determined.)

# How Should We Provide Early Statistical Practice?

*Follow two paths: lecture + lab (in class) and semester project (outside of class).*



# How Should We Provide Early Statistical Practice?

In class: learn and contextualize models; practice on example data

**Preliminary Schedule** github.com/pefreeman/36-290

| Week | Day | Topic |
|---|---|---|
| 1 | Tu | introduction to R + pre-test assessment |
| | Th | R: vectors + lab |
| 2 | Tu | R: dplyr + ggplot + lab |
| | Th | exploratory data analysis + lab |
| 3 | Tu | statistical learning + K-means + hierarchical clustering + lab |
| | Th | PCA + lab |
| 4 | Tu | regression model assessment + lab |
| | Th | classification model assessment + lab |
| 5 | Tu | linear regression + lab |
| | Th | GLM + logistic regression + lab |
| 6 | Tu | best subset selection + lab |
| | Th | penalized regression + lab |
| 7 | Tu | ML + trees + lab |
| | Th | reserved for project dataset work |
| 8 | Tu | random forest + lab |
| | Th | cancelled: mid-semester break |

• • •

# How Should We Provide Early Statistical Practice?

*Use value-added (i.e., pre- and post-) assessment to quantify student growth.*

1. You intend to learn a statistical model. Assuming that computation time is not an issue, which of the following represents the best way to attack the problem?

   (a) You learn and assess the model using all samples.

   (b) You learn the model with all the samples except one, and assess the model using this one "held-out" sample. You repeat the process $n$ times until every sample has been held out exactly once.

   (c) You learn the model using one sample (and assess it using all other held-out samples), and repeat the process $n$ times until every sample has been fit exactly once.

   (d) You split the samples into four quarters, then learn the model using three quarters and assess the model using the remaining quarter.

   (e) You split the samples into four quarters, then learn the model using three quarters and assess the model using the remaining quarter. You repeat the process four times until every quarter of the sample has been held out exactly once.

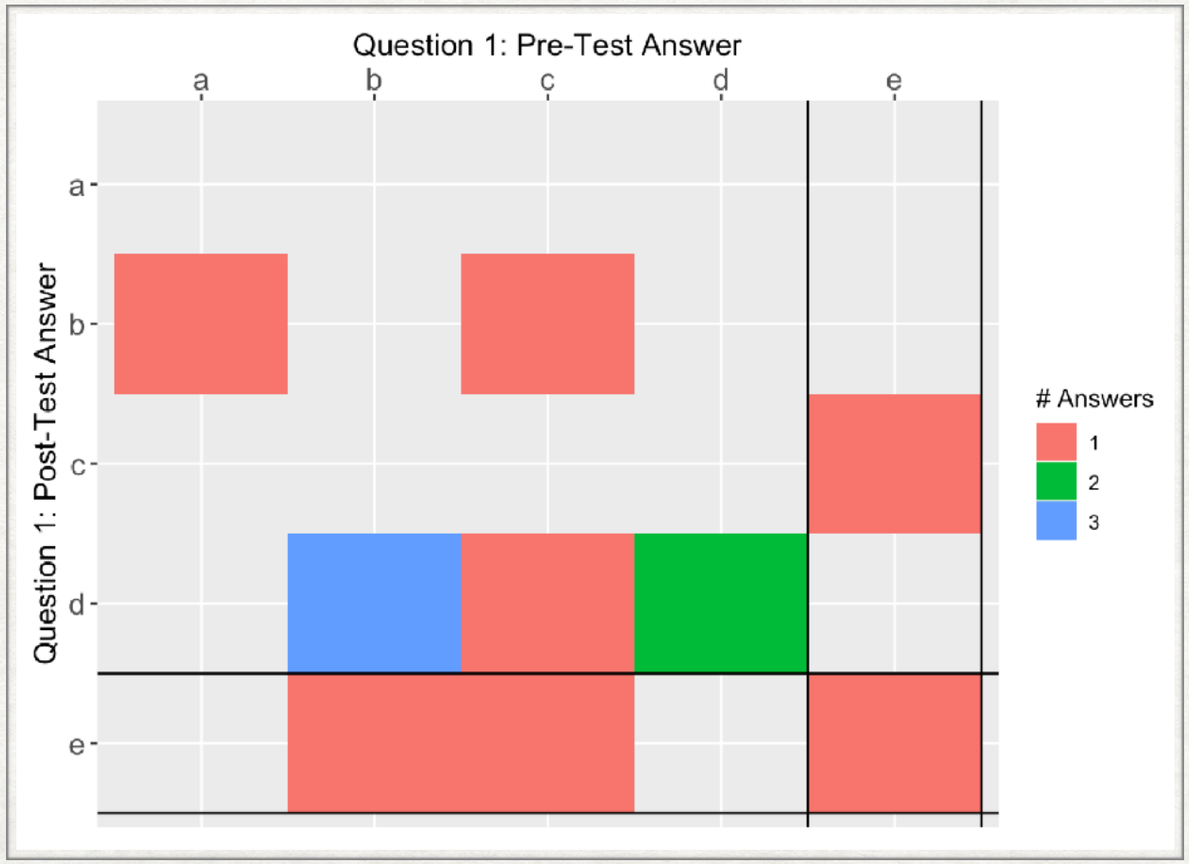
← no test dataset

← LOOCV

← anti-LOOCV

← 75/25 data splitting

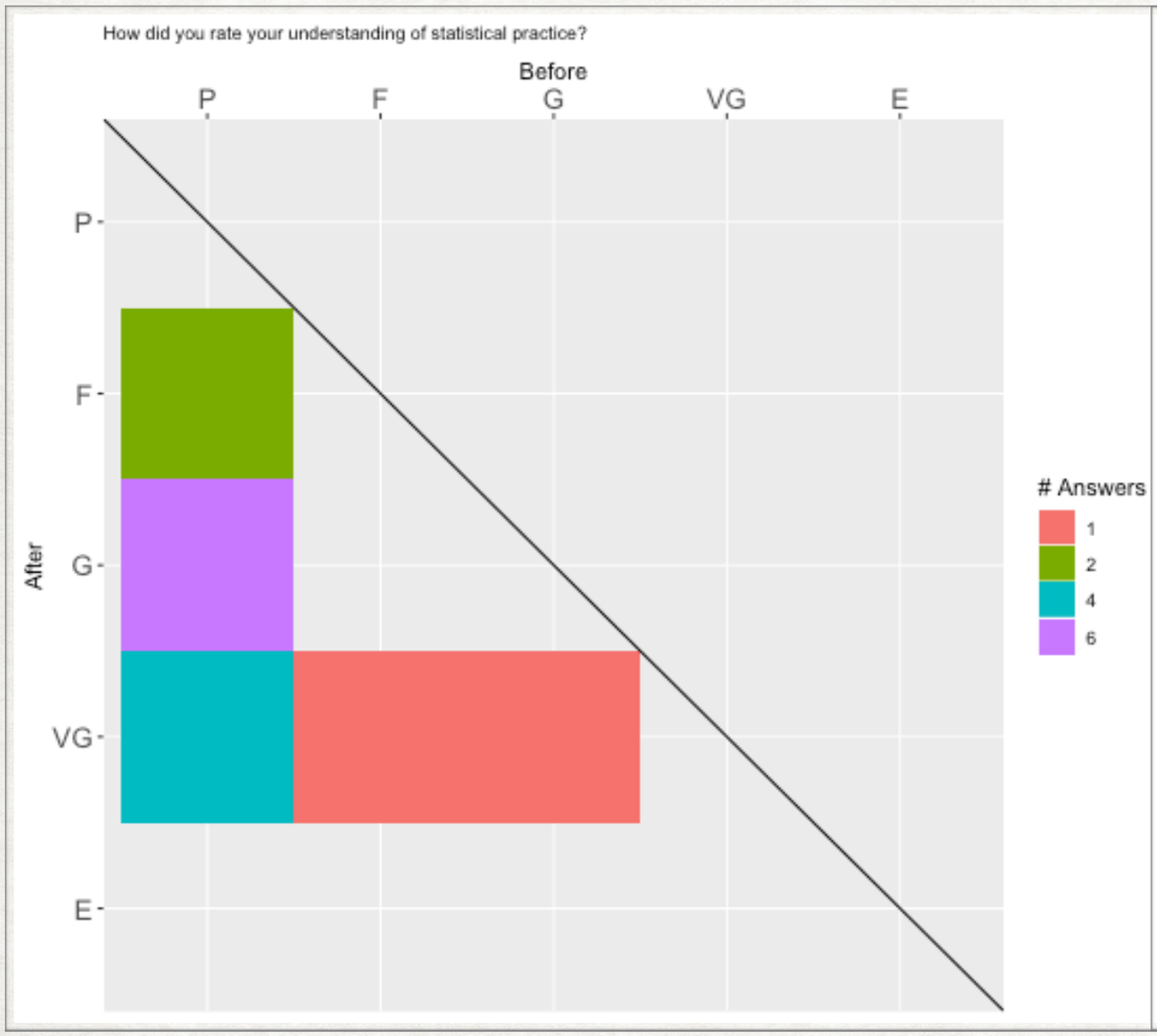
← 4-fold CV ✔



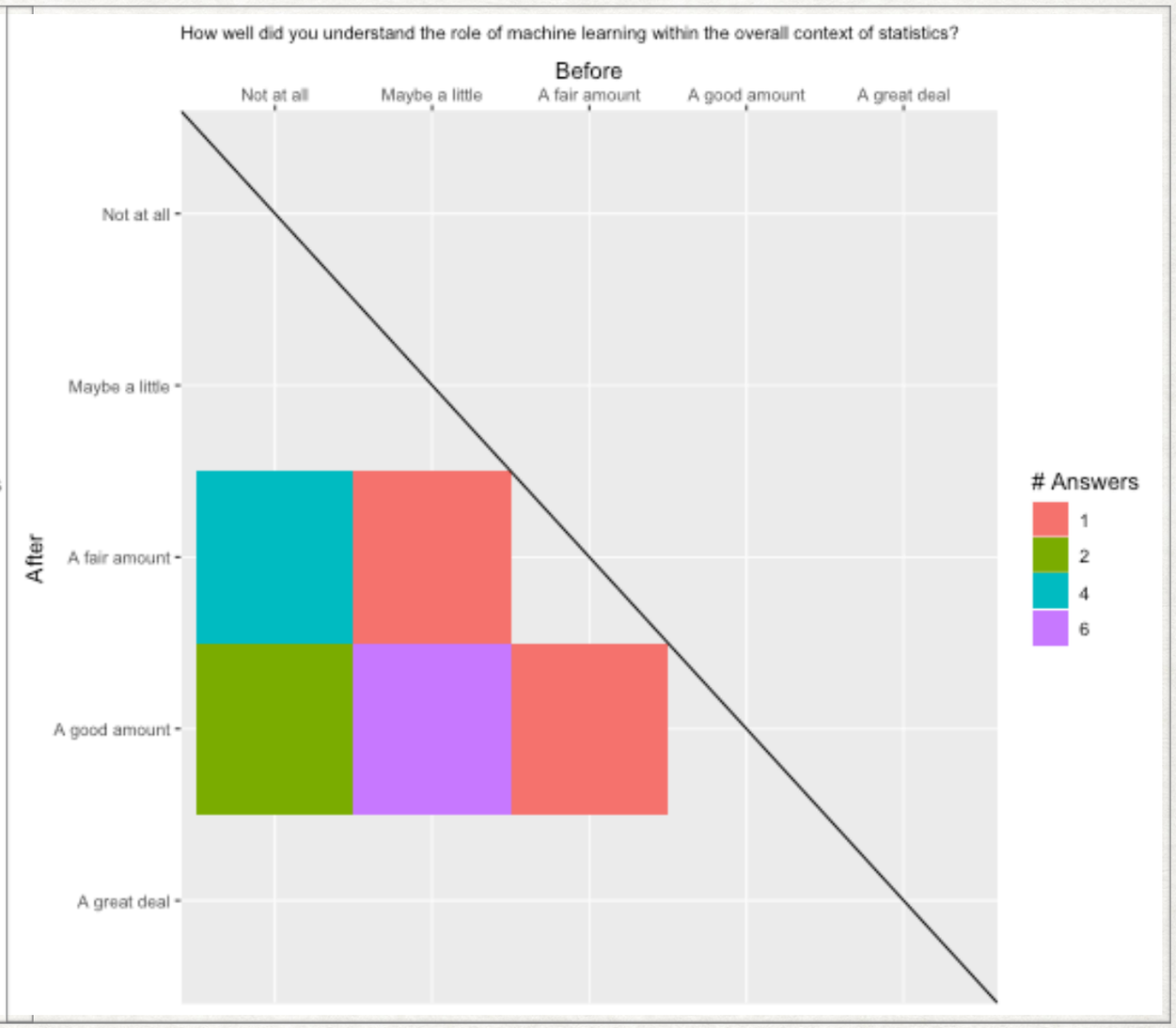

The majority of students come to recognize that data splitting/ k-fold CV is the best way to attack the problem (good!), although many do not recognize that k-fold CV is superior to data splitting (as it reduces the variance of test-set MSE).

# How Should We Provide Early Statistical Practice?

*Use retrospective surveys to quantify student attitudes.*



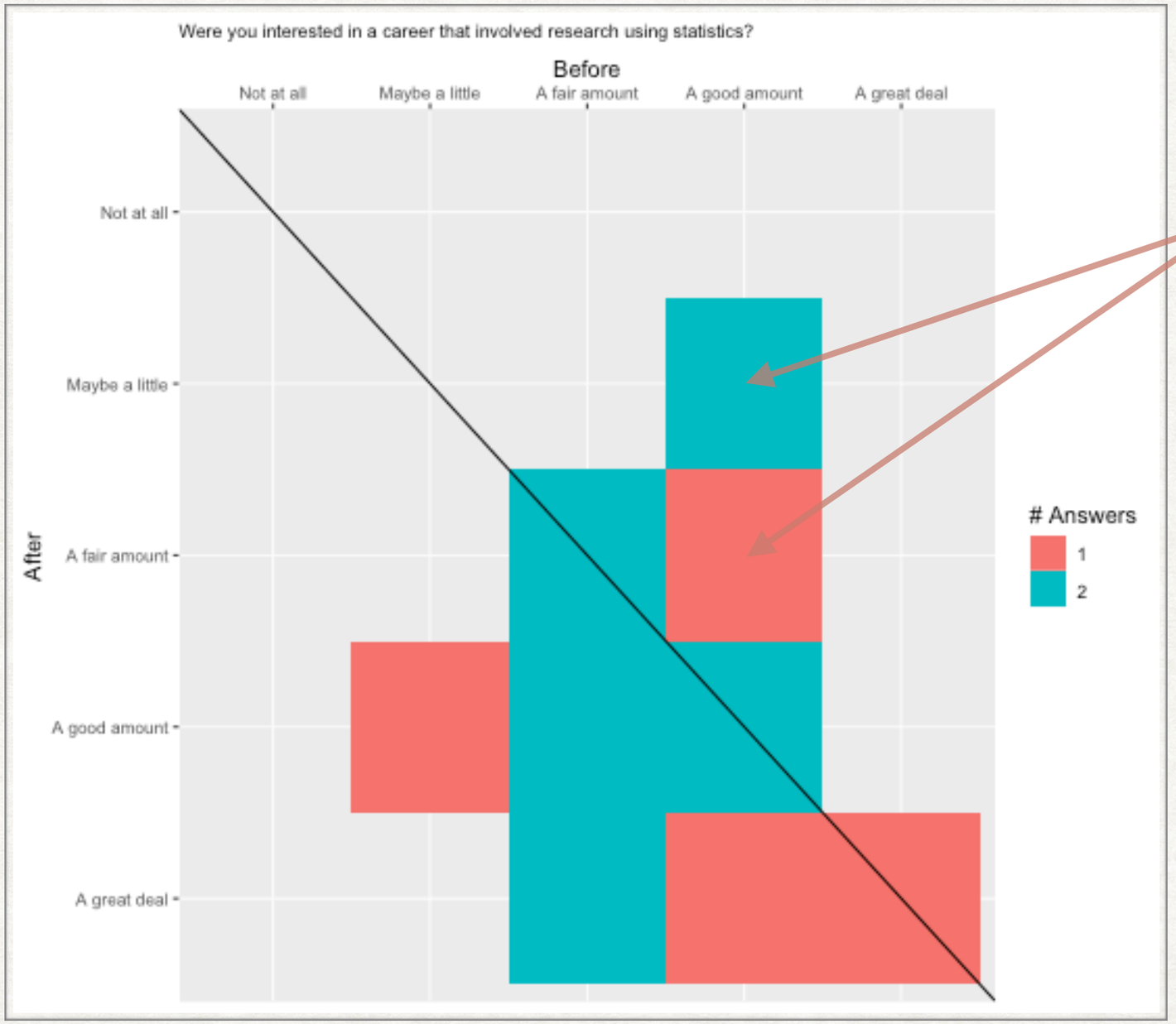

How did you rate your knowledge of statistical practice?




How well did you understand the role of machine learning within the overall context of statistics?

# Recall: Why Should We Provide Early Statistical Practice?

*It helps students determine whether they "like" statistics.*




Were you interested in a career that involved using statistics?

Less interested now than before.

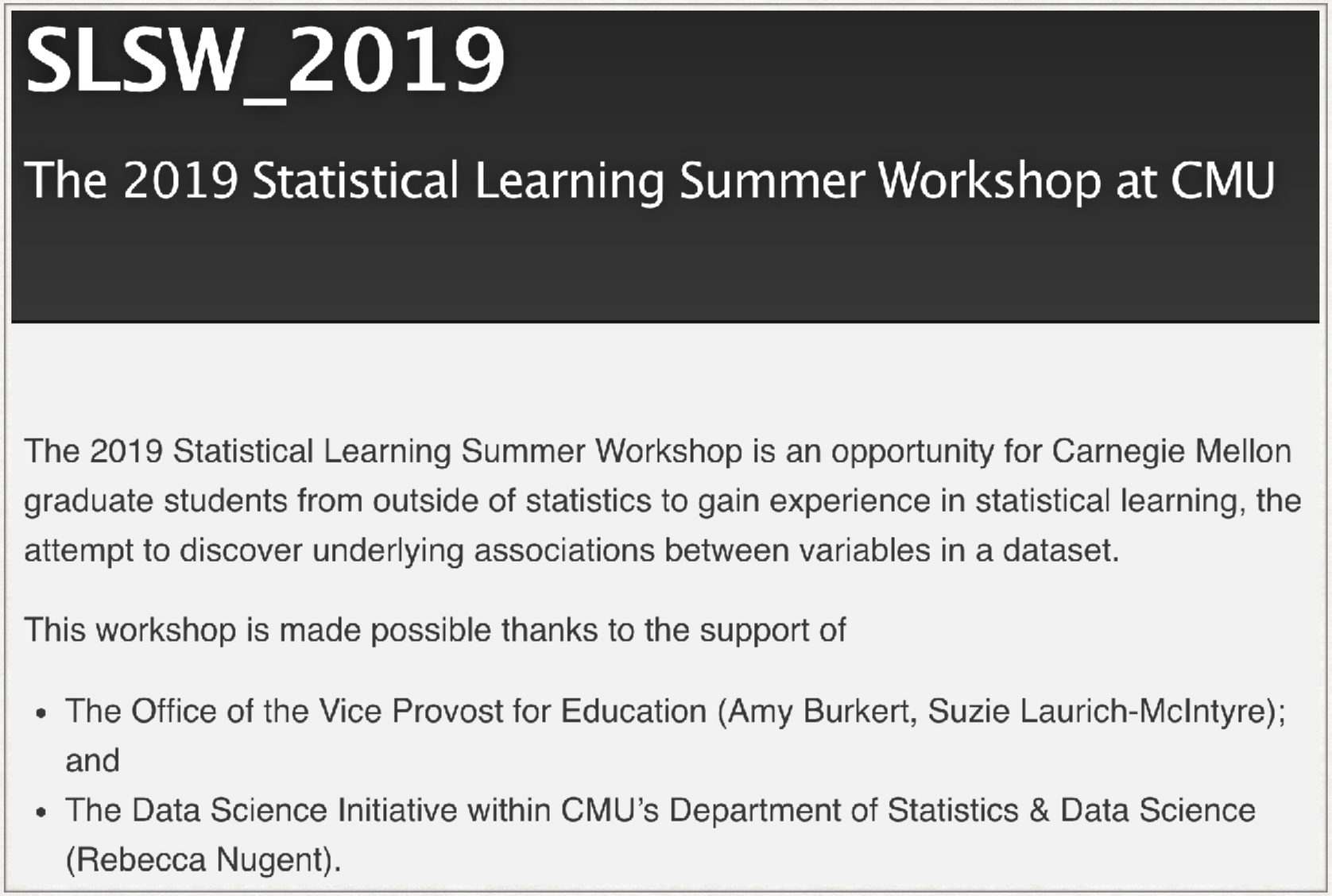
5: same interest

6: more interest

3: less interest

# An Extra Benefit to Providing Early Statistical Practice

*Course materials are easily extended, enabling outreach.*




www.stat.cmu.edu/slsw

This five-day workshop (organized with Joel Greenhouse and Rebecca Nugent) took 21 students from learning basic **R** to publicly presenting posters displaying analyses of project data.
(Over 100 students applied!…there's *definitely* an audience for such outreach.)

## To Sum Up…

*You should provide early undergraduate statistics majors with opportunities for "circumscribed" statistical practice that limits cognitive overload and thus maximizes the actual practice…*

1. Minimize coding overhead (e.g., via the use of tidyverse packages)
2. Use curated data (to reduce time lost to preprocessing)
3. Specify the research questions to answer
4. Treat models as "gray boxes" (focus on qualitative detail, not math)
5. Provide semester projects that encourage practice outside of class
6. Make your students publicly present their results

(7. Assess via pre- and post-course tests and retrospective surveys)

*…and then use your materials as a springboard to outreach (to external students, professors, corporate clients, executives desiring education, the public, etc.)*

# To Sum Up...

# Predicting Galaxy Mass and Star Formation Rate from Emission Line Spectra

Parvathi Meyyappan, Hal Rockwell, Filipp Shelobolin (Advisor: Peter Freeman)

## Introduction

In order to make statistical inferences about theories of galaxy formation and evolution, astronomers need to be able to estimate the masses and star formation rates of galaxies from the information present in their images. While conventional methods exist to estimate these quantities using physical models, these methods are computationally expensive, and it is desirable to determine simpler ways to generate the same estimates. Physical models undoubtedly must be used initially to provide training labels, but can we subsequently learn statistical models that adequately map observed properties of galaxies directly to masses and star-formation rates?

**In this poster, we specifically investigate whether we can utilize the strengths of the emission lines observed in galaxy spectra to predict masses and star-formation rates.**

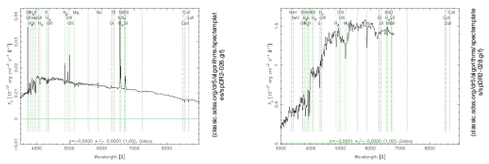


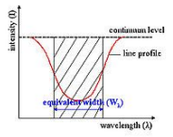
**Figure 1.** Example spectra from SDSS. (1 Angstrom is 10^-10 meters (or 0.1 nm))

| Predictors | Description |
|---|---|
| O II (372.9 nm) | an emission line strength of oxygen with no electrons missing |
| O III (495.9 nm) | an emission line strength of oxygen with one electron missing |
| O III (500.7 nm) | a different emission line strength of oxygen with one electron missing |
| N II (654.8 nm) | an emission line strength of nitrogen with one electron missing |
| N II (658.4 nm) | a different emission line strength of nitrogen with one electron missing |
| S II (671.7 nm) | an emission line strength of sulphur with one electron missing |
| S II (673.1 nm) | a different emission line strength of sulphur with one electron missing |
| H α | an emission line strength of hydrogen, the primary emission line of the Balmer series |
| H β | an emission line strength of hydrogen, the secondary emission line of the Balmer series |
| H ɣ | an emission line strength of hydrogen, the tertiary emission line of the Balmer series |
| **Responses** | **Description** |
| mass | the mass of the galaxy, in $\log_{10}$ solar masses |
| sfr | the star formation rate of the galaxy, in $\log_{10}$ solar masses per year |

Table 1: Descriptions of predictor and response variables.

## Data

Our predictors are measurements of spikes in spectra, called emission lines, obtained by spectroscopy from the Sloan Digital Sky Survey (SDSS). Emission lines are created when an atom in an excited state returns to a configuration of lower energy. The magnitudes of the lines are given in equivalent width, which is found by forming a rectangle with a height equal to the continuum emission and then finding a width such that the area of the rectangle is equal to the area of the spectral line. In Table 1 are all of our predictor variables. The roman numerals indicate the state of the atom: I means that the atom has all of its electrons and the reading is just an electron moving from an excited state to a less excited state. II means the atom is missing an entire electron, and III means it is missing two. Galaxy mass is just the mass of the galaxy and star formation rate is the the rate at which gas in galaxies is getting converted into new stars.



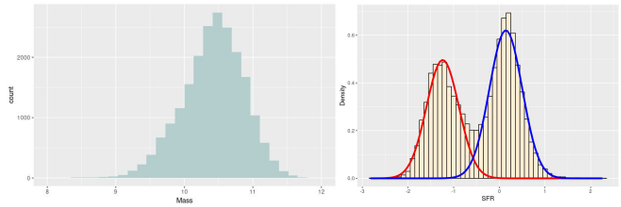
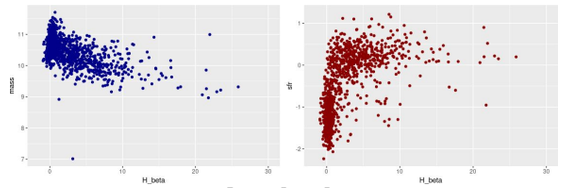




**Figure 2.** *Above left:* a histogram of the galaxy masses, showing a roughly normal distribution.
**Figure 3.** *Above right:* a mixture model of two normal distributions fit to the star formation rates, showing their bimodal distribution.
**Figure 4.** *Below left:* galaxy masses plotted against a sample predictor.
**Figure 5.** *Below right:* star formation rates plotted against a sample predictor.




## Analysis

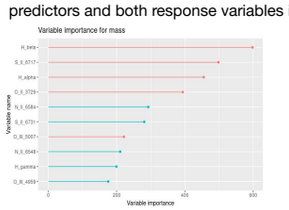
We try a variety of prediction algorithms to find which one yields the best predictions as measured testing our models on a test set, and comparing the mean of the squares of the differences between what our model predicts and the actual data (MSE). Several of these (linear, PC, lasso, ridge regression, and GAMs) are variants on the generalized additive model, in which the response is estimated as a linear combination of smooth functions of the predictors. The other methods we use are more nonlinear: support vector machines that project the data and then perform linear regression in the higher-dimensional space, boosted decision trees and random forest that construct trees that branch on learned values of the predictors, and K-nearest neighbors, predicting the mean of the nearest training data points.

## Results

| Model | Mass MSE | SFR MSE |
|---|---|---|
| Linear Regression | 0.117 | 0.334 |
| GAM | 0.089 | 0.176 |
| SVM | 0.083 | 0.171 |
| XGBoost | 0.080 | 0.170 |
| Random Forest | 0.073 | 0.160 |
| K Nearest Neighbors | 0.085 | 0.294 |
| PC Regression | 0.127 | 0.354 |
| Lasso Regression | 0.117 | 0.347 |
| Ridge Regression | 0.118 | 0.354 |

Table 2: Mean squared error of each method's predictions on the test set.

As shown above, Random Forest provides a significant reduction in MSE to other linear models such as ridge, lasso, or PC regression. However, the model's mean squared error value is still very similar to those from the SVM, GAM, and boosting models. This shows that the relationship between the predictors and both response variables is nonlinear by nature.



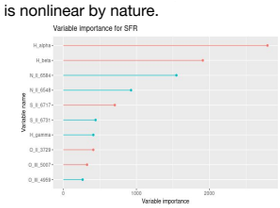
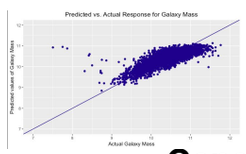
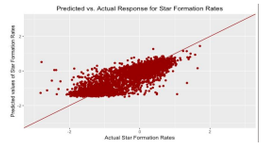




**Figure 6.** A plot showing the importance of the predictors in the random forest model for mass.
**Figure 8.** Predicted galaxy mass using Random Forest vs the observed galaxy mass.

**Figure 7.** A plot showing the importance of the predictors in the random forest model for star formation rate.
**Figure 9.** Predicted star formation rates using Random Forest vs the observed star formation rates.




## Conclusions

**We demonstrate in our analysis that galaxy mass and star formation rate can be reasonably well predicted from emission line spectra.** We also find that nonlinear methods, especially random forest, give much better predictions than linear ones. However, there may be room for improvement, as our best models still leave much of both responses' variance unexplained. We would like to see our models be run with more tuning and possible as a multivariate response, which could be possible with more computational resources.

### References:

James, Gareth et al. *An Introduction to Statistical Learning with Applications in R* : Springer; 1st ed. 2013
Sloan Digital Sky Survey Data Release 5: https://www.sdss.org/dr12/spectro/galaxy_mpajhu/

# Questions? Comments? Issues? Just Want to Talk?

## 4:00 PM Jumpstarting Early Undergraduate Research

Peter E. Freeman (Carnegie Mellon University)

The session is for those who are thinking of providing opportunities, or have provided opportunities, for statistical practice to early undergraduates: freshmen and sophomores. Basically we will discuss best practices: what works…and what doesn't.

- meet at the Statistics Education booth

# The Section on Statistics and Data Science Education invites you to contribute feedback on this presentation.

## You can use the JSM app or a paper form!

The feedback will be used to pick the
Ron Wasserstein Best Contributed Paper Award winner