



Research Skills for All: Introducing Statistical Research Methodology to Early Undergraduates at Carnegie Mellon

Peter E. Freeman (Department of Statistics & Data Science, CMU)



Sophomore Malaika Handa (May 2017)

Early Undergraduate Research at CMU

P. E. Freeman - May 2018

Teaching Statistics Group: <http://www.stat.cmu.edu/teachstat/>





In Envisioning the Statistics & Data Science Curriculum of Tomorrow There Is (As There Should Be!) Much Focus on Data Science...



Curriculum Guidelines for Undergraduate Programs in Data Science*

- De Veaux et al. 2017

Data Science in Statistics Curricula: Preparing Students to “Think with Data”

- Hardin et al. 2015

DATA SCIENCE FOR UNDERGRADUATES: OPPORTUNITIES AND OPTIONS

- Haas & Hero et al. 2018

(plus many other recent papers!)

Early Undergraduate Research at CMU

P. E. Freeman - May 2018

Teaching Statistics Group: <http://www.stat.cmu.edu/teachstat/>





...But, I Would Argue, Less Focus on Statistical Practice (And, In Particular, *Repeated* Practice That Starts Early)

American Statistical Association
Undergraduate Guidelines Workgroup

Curriculum Guidelines for Undergraduate Programs in Statistical Science

- ASA 14 (Horton Report)

“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.** Such scaffolded exposure helps students connect statistical concepts and theory to practice.”
[bolding mine]

Mere Renovation is Too Little Too Late: We Need to Rethink our Undergraduate Curriculum from the Ground Up

- Cobb 2015

“[W]e should put priority on two goals, to make ‘fundamental concepts accessible’ and to ‘minimize prerequisites to research.’”

Early Undergraduate Research at CMU

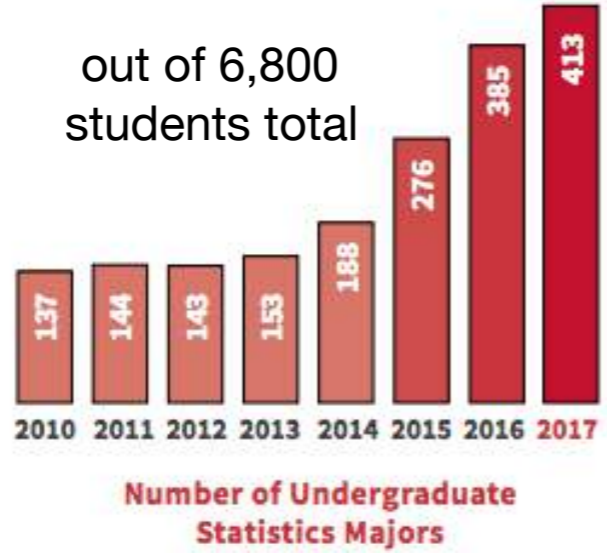
P. E. Freeman - May 2018

Teaching Statistics Group: <http://www.stat.cmu.edu/teachstat/>





Early Research Experience at CMU: the Context



(now: 500+)

Most of our students aim for jobs in industry.

On Teaching Statistical Practice: From Novice to Expert

- Greenhouse & Seltman 2017

But: “[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.” One can argue that these students lacked sufficient practice in statistics.

“There has been a tendency in statistics to have students first understand, then do” (Brown & Kass 2009): early students are generally denied opportunities.

Axioms: experience in statistical practice helps early undergraduates gain internships (and eventual long-term employment), and helps them in future classes.

→ I thus began advising early undergraduate research in 2014.

Early Undergraduate Research at CMU

P. E. Freeman - May 2018

Teaching Statistics Group: <http://www.stat.cmu.edu/teachstat/>





An Important Contextual Note...

I make no claim that we at CMU were the first to provide a mechanism by which early undergraduates (or more broadly, those who have previously taken a minimal number [or no] statistics courses, regardless of year) could gain experience in statistical research!

(See, e.g., Legler et al. 2010, Nolan & Temple Lang 2015, Wagaman 2016.)

Rather, I am here to speak of our experiences at CMU, and to sell the listener on the idea that a properly calibrated, conceptual introduction to statistical learning and its application to research has great benefits for students. (In short: you should do this!)

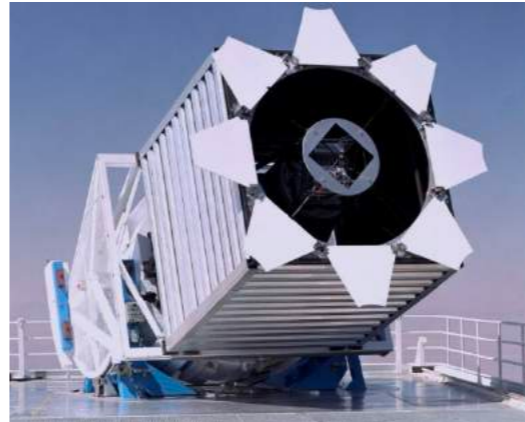


Dietrich College Research Training Program

The Dietrich College freshman-sophomore research training program is open to second semester freshmen and sophomores with a 3.0 QPA or by petition.



+



=

u	g	r	i	z	redshift	
17.8313	16.9077	16.4431	16.2099	16.0613	15.8732	0.038356
19.0731	17.7448	16.9789	16.5288	16.2551	15.9531	0.058309
21.638	21.0106	20.8286	20.6283	20.6552	20.528	0.063701
20.5474	19.5542	19.2387	19.0568	19.0887	18.9865	0.059006
21.2378	20.6876	20.5661	20.4371	20.4799	20.4503	0.063202
22.4627	21.4597	21.0484	20.8274	20.7639	20.6385	0.057773
23.8221	22.895	22.5779	22.3543	22.3225	22.2038	0.061548
23.0491	22.1536	21.8791	21.6889	21.7044	21.6381	0.063769
23.6742	23.0346	22.7857	22.6116	22.5813	22.5462	0.061427
23.5684	22.6635	22.3825	22.1811	22.1842	22.1003	0.066216
22.8421	21.9752	21.6568	21.4507	21.3821	21.2972	0.062465
22.27	20.98	20.3657	20.0249	19.8772	19.665	0.059334
22.9572	22.1043	21.8311	21.6471	21.6476	21.6021	0.064125
22.6374	21.7366	21.3624	21.0996	21.0509	20.8953	0.062879

My background is in astronomy. Thus I construct research questions with the following in mind: can the student answer a question that would help me, in the role of astronomer, better understand the processes underlying the data?

(cf. Cobb 2015: “research’ should be understood...broadly, namely, using data to study an unanswered real-world question that matters...”)

Early Undergraduate Research at CMU

P. E. Freeman - May 2018

Teaching Statistics Group: <http://www.stat.cmu.edu/teachstat/>





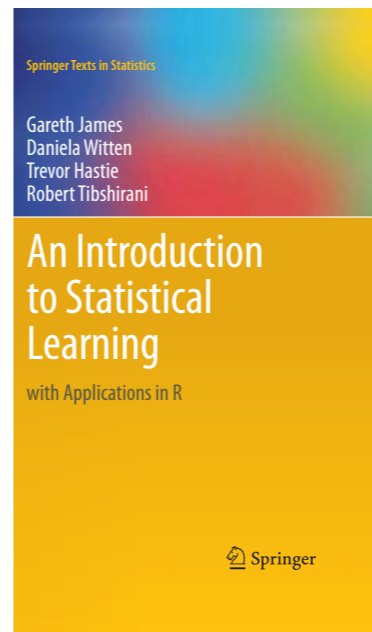
Dietrich College Research Training Program

The Dietrich College freshman-sophomore research training program is open to second semester freshmen and sophomores with a 3.0 QPA or by petition.

The Recipe:

```
u g r i z y redshift
17.8313 16.9077 16.4431 16.2099 16.0613 15.8732 0.038356
19.0731 17.7448 16.9789 16.5288 16.2551 15.9531 0.058309
21.638 21.0106 20.8286 20.6283 20.6552 20.528 0.063701
20.5474 19.5542 19.2387 19.0568 19.0887 18.9865 0.059006
21.2378 20.6876 20.5661 20.4371 20.4799 20.4503 0.063202
22.4627 21.4597 21.0484 20.8274 20.7639 20.6385 0.057773
23.8221 22.895 22.5779 22.3543 22.3225 22.2038 0.061548
23.0491 22.1536 21.8791 21.6889 21.7044 21.6381 0.063769
23.6742 23.0346 22.7857 22.6116 22.5813 22.5462 0.061427
23.5684 22.6635 22.3825 22.1811 22.1842 22.1003 0.066216
22.8421 21.9752 21.6568 21.4507 21.3821 21.2972 0.062465
22.27 20.98 20.3657 20.0249 19.8772 19.665 0.059334
22.9572 22.1043 21.8311 21.6471 21.6476 21.6021 0.064125
22.6374 21.7366 21.3624 21.0996 21.0509 20.8953 0.062879
```

+



+

R Markdown =
from R Studio

Determining the Proportion of Multiple-Star Systems using APOGEE Radial Velocity Data

Shannon Lu (Advisor: Peter Freeman)



Introduction

In the Milky Way galaxy, stars exist either in isolation or as part of a multiple-star system. Figure 4 at the lower right illustrates a binary-star system where both stars orbit the system's center of mass. The true proportion of stellar systems that contain multiple stars is unknown. However, the APOGEE data from the Sloan Digital Sky Survey provides information on radial velocities, or velocities relative to the Earth, for over 100,000 stars, which we can use to attempt to predict whether a given star orbits on its own or as part of a system. The radial velocities are useful because they can indicate whether or not a star wobbles back and forth relative to the system's overall center of mass, and typically the brightest star in a multiple-star system may be seen to wobble, whereas there is no wobble for a single star.

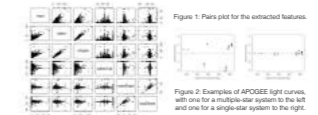
In this project, we seek to answer two questions. First, using methods of unsupervised learning, can we identify the populations of single and multiple-star systems? Second, if someday we acquired data with known single and multiple-star systems, how well would machine learning algorithms perform in classifying the data?

Data

The Sloan Digital Sky Survey's APOGEE experiment (Majewski et al. 2017) has collected multiple instances of high-resolution spectra for over 100,000 stars in the Milky Way. Each spectrum contains Doppler-shifted atomic lines, the magnitude of the shifts are proportional to the star's radial velocity. The dataset for this project, from SDSS Data Release 13, contains light curves (radial velocities as a function of time) for 75,276 stars, with the number of observations per star ranging from 2 to 28. The range of values for the radial velocities is -155 to 147 km/s, with an expected radial velocity error of approximately 0.1 km/s.

The following features were extracted from each light curve after its values were adjusted so as to have zero mean:

Feature	Description
median	absolute value of the median of the radial velocities
range	standard deviation of the radial velocities
width	range between minimum and maximum radial velocity
width2	percentage of radial velocities within 1 standard deviation from the mean
width3	maximum change in radial velocity between two observations
width4	average change in radial velocity between two observations



Analysis

We apply K-means clustering to the extracted features in order to see whether we can visually identify the populations of multiple-star and single-star systems. We assume two clusters because there are two populations of interest.

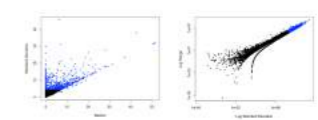


Figure 3: Examples of results from a K-means clustering analysis, assuming two clusters.

In the figure above, we observe how the population of statistics stretches continuously from the regime of single-star systems (values near zero) to the regime of multiple-star systems (values far from zero). However, there is no easy way to draw a boundary between the two populations. This is not surprising because multiple-star systems can rotate perpendicularly to the line of sight and thus look like a single star because the radial velocities are measured with respect to the Earth and may not be changing. There is an overlap between the two populations of stellar systems, so we cannot clearly separate them.

Supervised Learning

Here we address the question of whether, given a labeled dataset, commonly used machine learning algorithms would be helpful in classifying star systems. To generate labeled data, we use Python-based simulation software created by Carlos Badenes (PI), Eric Alpert (CMU), and Peter Freeman. Without loss of generality, we assume that for these labeled data, the fraction of multiple-star systems is 0.5

Algorithm	Misclassification Rate
Decision Forest	0.139
Naive Bayes	0.152
Ada Boost	0.287
Ada Boost	0.240

We see that supervised learning algorithms can classify multiple-star systems relatively well. Compared to the trivial misclassification rate of 50%, the random forest classifier performs better with a misclassification rate of 13.9%. However, the classifiers can never be perfect because of the fact that radial velocities of some multiple-star systems will be similar to those of single stars because their stars move (nearly) perpendicularly to our line of sight.

Looking at the results from random forest, we see that the median, range, and standard deviation are the most important features in classifying the stars as single, double, or multiple-star systems. Range and standard deviation are highly correlated, so it is not surprising that their importances are similar.

Conclusions

Given radial velocity data for 75,276 stars, we extracted features and performed both supervised and unsupervised learning analyses in order to try to classify stellar systems as single stars or multiple-star systems. Using unsupervised learning methods, we find that it is difficult to classify the stars into two distinct populations. With supervised learning, in a simulation where the proportion of multiple-star systems is known, we see that the machine learning algorithms are able to separate the data, however, there is still a misclassification rate of about 15% using these methods.

References

- Majewski S. R., et al. 2017, The Astronomical Journal, vol. 154, p. 94
- James, G., et al. 2013, An Introduction to Statistical Learning, Springer
- Bishop, C. M. 2006, Pattern Recognition and Machine Learning, Springer

Center of mass

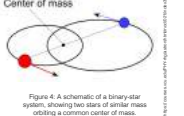


Figure 4: A schematic of a binary star system, showing two stars of similar mass orbiting a common center of mass.

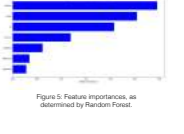


Figure 5: Feature importances, as determined by Random Forest.

Through this program, my students have carried out 18 individualized projects since Fall 2014, all presented as posters at CMU's Meeting of the Minds.

Early Undergraduate Research at CMU

P. E. Freeman - May 2018

Teaching Statistics Group: <http://www.stat.cmu.edu/teachstat/>

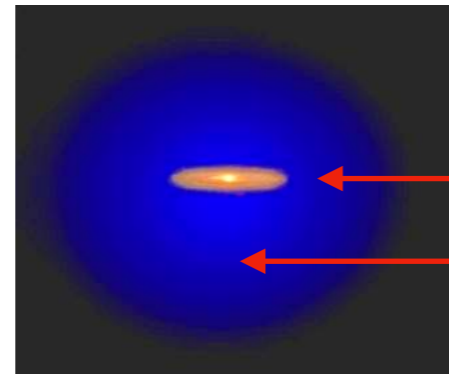


eCOTS 2018



Demonstrations

1) Stellar Mass - Dark-Matter Halo Relationship



galaxy (shines!)

dark-matter halo (does not shine!)

2) Clustering Applied to Draco Dwarf Galaxy Data



which stars belong to the dwarf galaxy and which do not?

(yep, there's a galaxy in this picture)

Early Undergraduate Research at CMU

P. E. Freeman - May 2018

Teaching Statistics Group: <http://www.stat.cmu.edu/teachstat/>

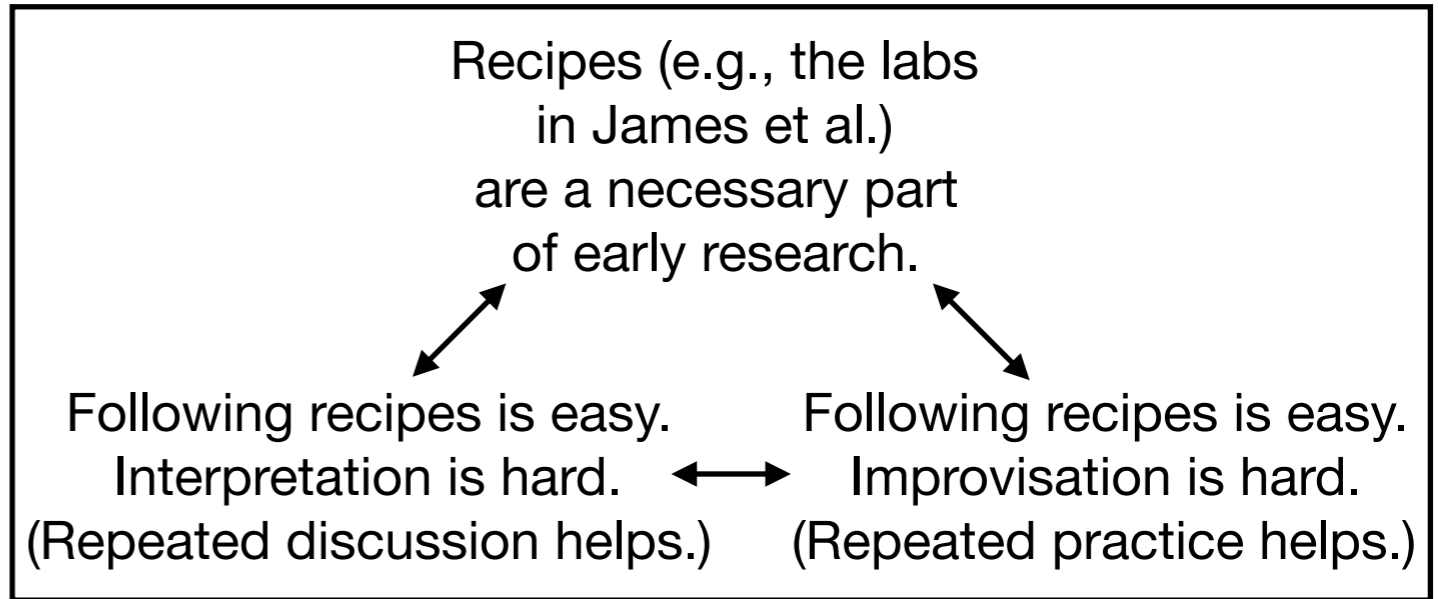


Reflections...

Students are not held back by lack of domain knowledge.

Tutorials for R (e.g., the *swirl* package) may be necessary to get students up to speed. Providing curated data and code for populating a data frame with those data helps!

Weekly meetings provide ample opportunity for feedback and for repeated discussion.



Students get a good qualitative feel for models.

“As a goal, we should seek a way to summarize profound concepts simply and succinctly, in words only.”
- Cobb (2015)

```

graph LR
    A[Students get a good qualitative feel for models.] <--> B["As a goal, we should seek a way to summarize profound concepts simply and succinctly, in words only." - Cobb (2015)]
  
```

Students gain confidence as they construct a contextual framework:
“so this is what research is”
“I can do this”

Students are invariably proud of their final poster!

Early Undergraduate Research at CMU

P. E. Freeman - May 2018

Teaching Statistics Group: <http://www.stat.cmu.edu/teachstat/>





Did statistical research help you in subsequent classes?

“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, which were all covered in-depth in...Intro[duction] to Machine Learning. It 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].”

Did statistical research help you attain internships?

“I want to say that every interview I've had has brought up this research...the two places I ended up working (Ascend Public Charter schools last summer, and Uber this coming summer) both specifically asked me about it and what skills I learned, and what methods I used. It was very nice to have an explicitly academic and quantitative experience to talk about, especially since I had minimal experience.”

“Yes, almost all the interviewers asked about the research projects that I wrote on my resume and wanted me to go into detail.”

“Absolutely. 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... It wasn't that the project had dominated the conversation...but it definitely left an impression.”



New Class: Introduction to Statistical Research Methodology

Preliminary Schedule (Subject to Change)

Week	What's Happening
1-2	Examine Breiman's "The Two Cultures" and introduce/review terminology. Download/install R + Python as necessary; review R basics.
3	Basics of data reading and exploratory data analysis. Pick research groups and select semester term projects.
4	Unsupervised learning: K-means, hierarchical clustering, and PCA.
5	Data splitting, illustrated with linear regression. Fit metrics: mean-squared error, prediction plots.
6	Dimension reduction: subset selection/lasso.
7	Generalization of linear regression. Discuss how to communicate analysis results. First DA due for semester term project.
8	Regression trees, random forest, and boosting.
9	k-nearest neighbors, data splitting with tuning parameters.
10	Support Vector Machines and Neural Networks.
11	Classification illustrated with random forest and knn. Fit metrics: confusion matrix, misclassification rate, ROC curves.
12	More classification (logistic regression, etc.) Second DA due for semester term project.
13	NO CLASS
14	Hackathon. (Details TBD.) Revised, final DA due for semester term project.
15	Group oral presentations and submission of final poster.

Dual track: an in-class track involving lectures, labs, and homework; and an out-of-class track involving the application of class concepts to the analysis of an external dataset, over the course of the semester.

Class Motto: Embrace the Ambiguity!

Early Undergraduate Research at CMU

P. E. Freeman - May 2018

Teaching Statistics Group: <http://www.stat.cmu.edu/teachstat/>

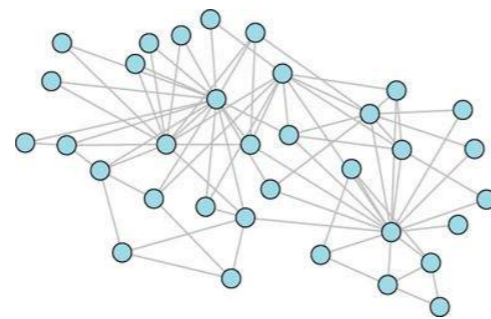
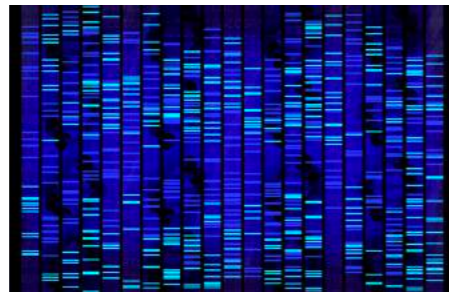




The Future



Does early undergraduate research have quantifiable benefits?



Expanding the scope and reach of early undergraduate research (i.e., getting more faculty involved, each operating in a different domain; this is preferable to making any single class larger!).



The Take-Home Message

Early exposure to research is an important aspect of an undergraduate's (*) professional development!
You should offer such exposure at your institution!

(*) The research detailed here was targeted toward, and primarily carried out by, statistics majors. However, properly calibrated, it could be targeted to early non-majors, to high school students, etc. Early research is an important aspect of Data Science for All!



Questions or Comments?

pfreeman@cmu.edu

Looking for (Curated) Data for Your Course?

See github.com/pefreeman/36-290/EXAMPLE_DATASETS

Figure Credits:

<https://www.cmu.edu/dietrich/docs/current/2017-year-in-review.pdf>

http://www.crossing-technologies.com/wp-content/uploads/2015/04/Big_data_image.jpg

https://en.wikipedia.org/wiki/Arp_271

<http://scienceblogs.com/startswithabang/2009/11/27/colliding-galaxies-for-fun-and/>

<https://chandra.as.utexas.edu/dm-halo-pic.html>

<http://www.greggsastronomy.com/ugc10822.html>

<https://www.alsde.edu/sec/sa/Pages/home.aspx>

<http://www.govtech.com/opinion/Precision-Medicine-Initiative-Why-You-Should-Worry-About-the-Privatization-of-Genetic-Data.html>

<https://deeplearning4j.org/graphanalytics>

<https://www.meritalk.com/articles/census-bureau-struggles-with-managing-new-it-for-2020-count/>