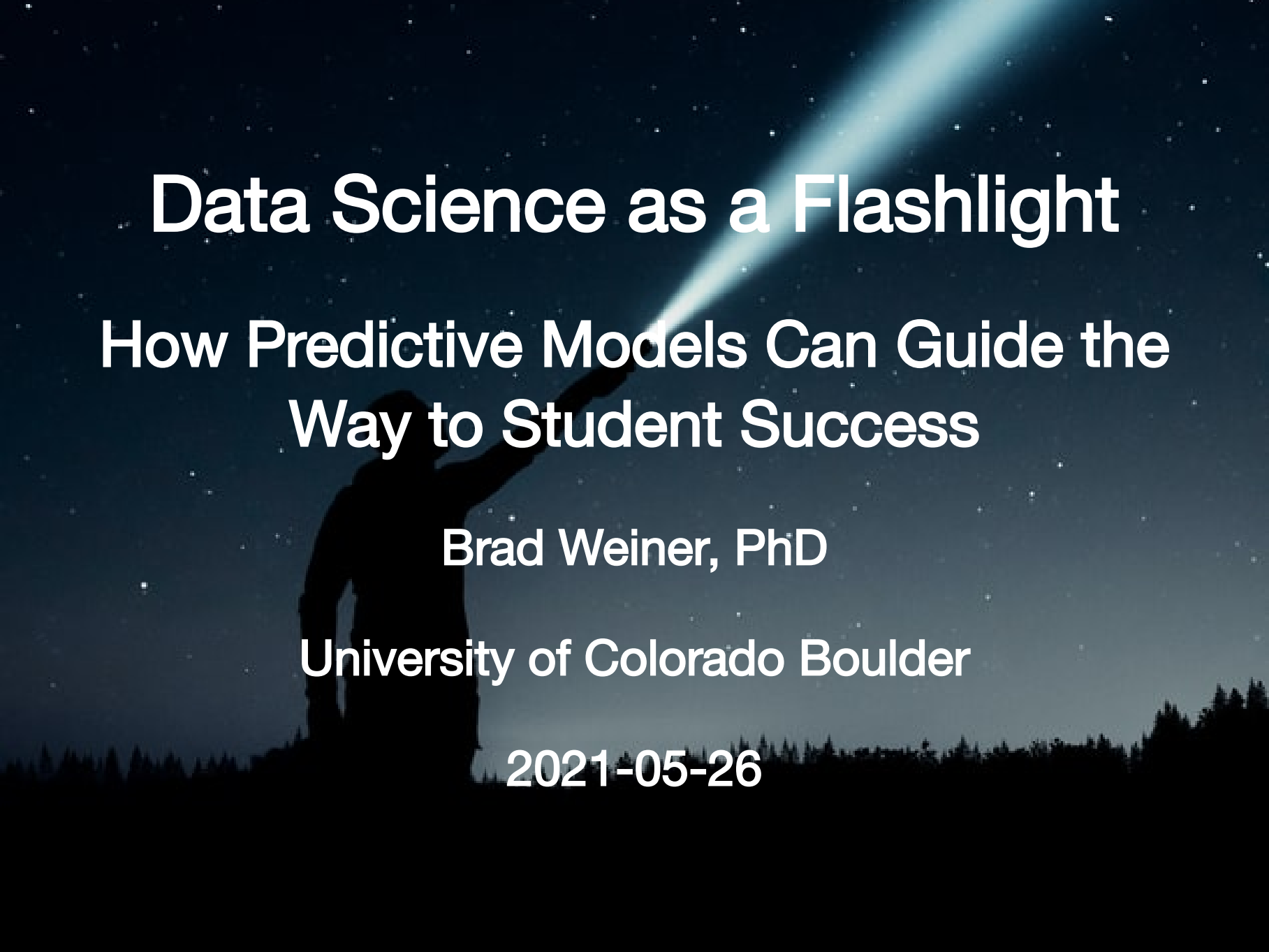


Data Science as a Flashlight

How Predictive Models Can Guide the Way to Student Success

Brad Weiner, PhD

University of Colorado Boulder

2021-05-26

About Me

Director of Data Science, University of Colorado Boulder

19 years experience in higher education

14 years on campus (Kansas, Vanderbilt, Minnesota, Colorado)

5 years in Ed-Tech/Consultancy

11 years Higher Ed Analytics/Data Science

Contact

brad.weiner@colorado.edu

brad_weiner

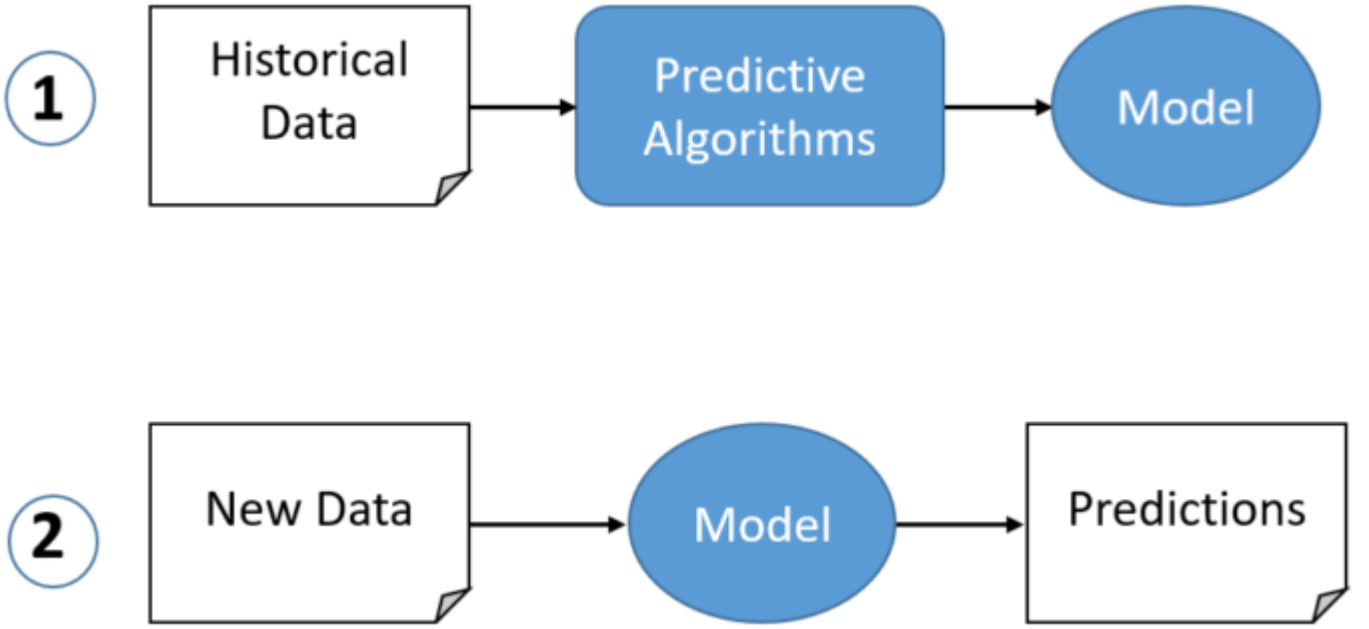
bradweiner.info

What is a predictive model?

It's an equation

Built on historical data (Beware!)

Estimates the likelihood of a future outcome



Why Use Predictive models in higher education?

Colleges and universities are in the public trust

We don't (currently) have the capacity to provide customized interventions for each individual student

Our research, teaching, and public engagement missions depend on us to fairly and efficiently allocate resources

If we don't use data or existing research, we're guessing

If we're guessing, we're biased toward the *status quo*

How do we use predictive models effectively *and* fairly?

✅ Understand the use case. Are you predicting success or failure?

✅ Assess risk. Don't deploy if the model could deny opportunities

✅ Specify the model transparently and test for accuracy

✅ Train users on intended uses. Avoid "off-label" efforts

✅ Align incentives and organizational structures with the outcome

✅ Learn to point the outcome toward populations of interest. If we want to enroll and retain more BIPOC Engineers, then use the model as a flashlight for those students


Let's Predict Retention
Reminder: This is an example. Be Careful.

Explore the Data (this is not real student data)

student_id	1	2	3	4	5	6
retained	0	1	1	1	0	0
income_group	Pell Eligible	No Aid	Pell Eligible	No Aid	Pell Eligible	Pell Eligible
sex	male	female	female	female	male	male
age	22	38	26	35	35	NA
siblings_enrolled	1	1	0	1	0	0
peers_from_hs	0	0	0	0	0	0
net_tuition	283	2783	309	2073	314	330
residency	Resident	Non-Resident	Resident	Resident	Resident	International
total_peer_group	1	1	0	1	0	0

Pre-Process the Data

student_id	1	2	3	4	5	6
retained	0	1	1	1	0	0
income_group	Pell Eligible	No Aid	Pell Eligible	No Aid	Pell Eligible	Pell Eligible
sex	male	female	female	female	male	male
age	-0.5300051	0.5714304	-0.2546462	0.3649113	0.3649113	NA
siblings_enrolled	0.4325504	0.4325504	-0.4742788	0.4325504	-0.4742788	-0.4742788
peers_from_hs	-0.4734077	-0.4734077	-0.4734077	-0.4734077	-0.4734077	-0.4734077
net_tuition	-0.5021568	0.7865640	-0.4887541	0.4205673	-0.4861766	-0.4779288
residency	Resident	Non-Resident	Resident	Resident	Resident	International
total_peer_group	1	1	0	1	0	0
income_group_no_aid	0	1	0	1	0	0
income_group_pell_eligible	1	0	1	0	1	1
income_group_state_grant_eligible	0	0	0	0	0	0
sex_female	0	1	1	1	0	0
sex_male	1	0	0	0	1	1
residency_international	0	0	0	0	0	1
residency_non_resident	0	1	0	0	0	0
residency_resident	1	0	1	1	1	0
residency_na	0	0	0	0	0	0

Split Into Training/Test Sets

retn_train$retained	n	percent
0	412	0.6158445
1	257	0.3841555

retn_test$retained	n	percent
0	137	0.6171171
1	85	0.3828829

Build Basic Regression Model

(reminder, I only have 8 minutes)

```
mod.1 <- glm(retained ~
               total_peer_group +
               net_tuition +
               sex_female +
               income_group_no_aid,
             data = retn_train,
             family = "binomial")
```

Review and Interpret the Results

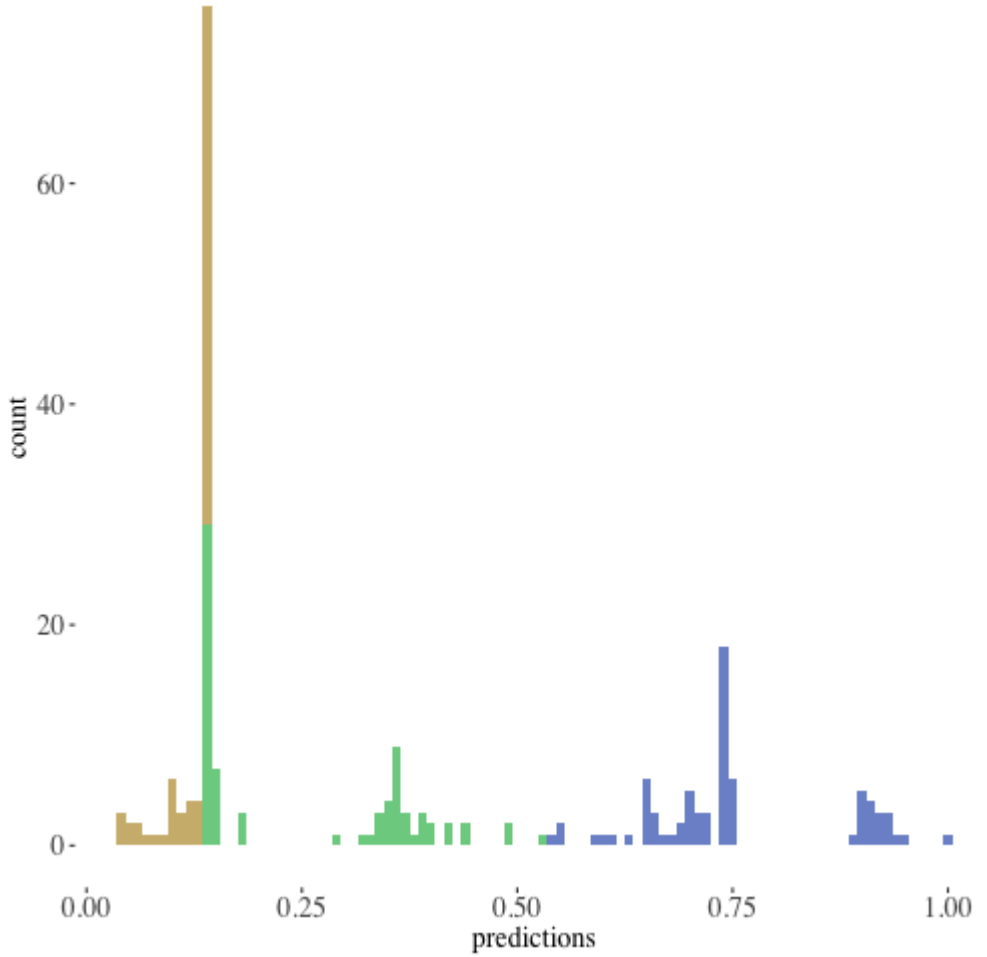
term	estimate	std.error	statistic	p.value
(Intercept)	0.190	0.179	-9.255	0.000
total_peer_group	**0.786**	**0.074**	**-3.245**	**0.001**
net_tuition	1.381	0.168	1.923	0.054
sex_female	**17.582**	**0.225**	**12.744**	**0.000**
income_group_no_aid	**2.992**	**0.297**	**3.691**	**0.000**

Interpretation

Students in the Income No Aid Group are 2.992 times more likely to retain than those in the baseline group when controlling for other features

Female Students 17.5 times more likely to retain than those in the baseline group when controlling for other features

Make New Predictions

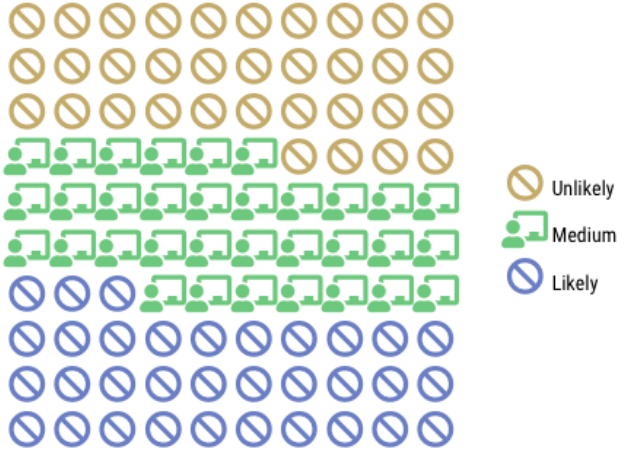


Allocate Scarce Resources With Interventions/Programming

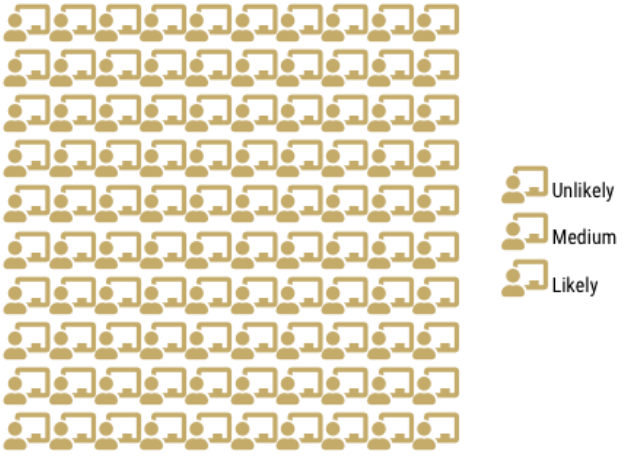
THIS is where your model goes from an _equation_ to an _intervention_

Option #1: All Students

Option #2: All Students



Option #3: All Engineers



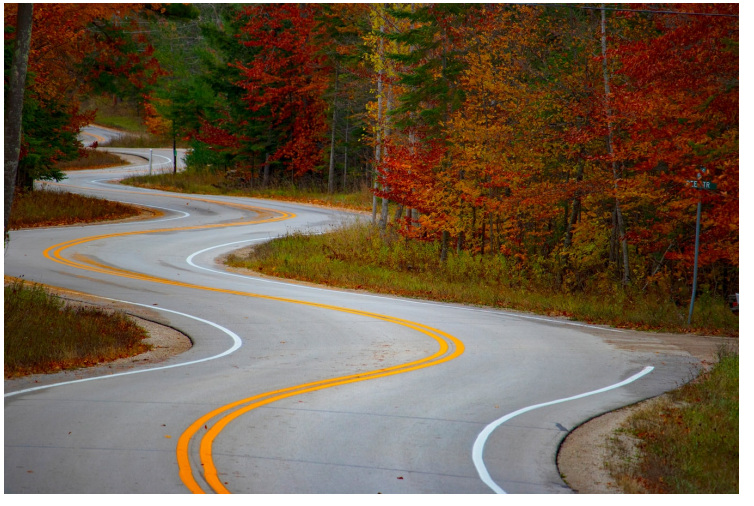
Option #4: BIPOC Engineers

Which option will you choose?

Who "succeeds" in higher education is...

Structural

Cultural



As educators we should *appropriately* use data to allocate limited resources

That means we can...

Reproduce the Past

Create More Opportunities in the Future

Thanks

Code and Slides available at

bradweiner.info/talk