Introduction

Communicating Results

Will Doyle

The Basics

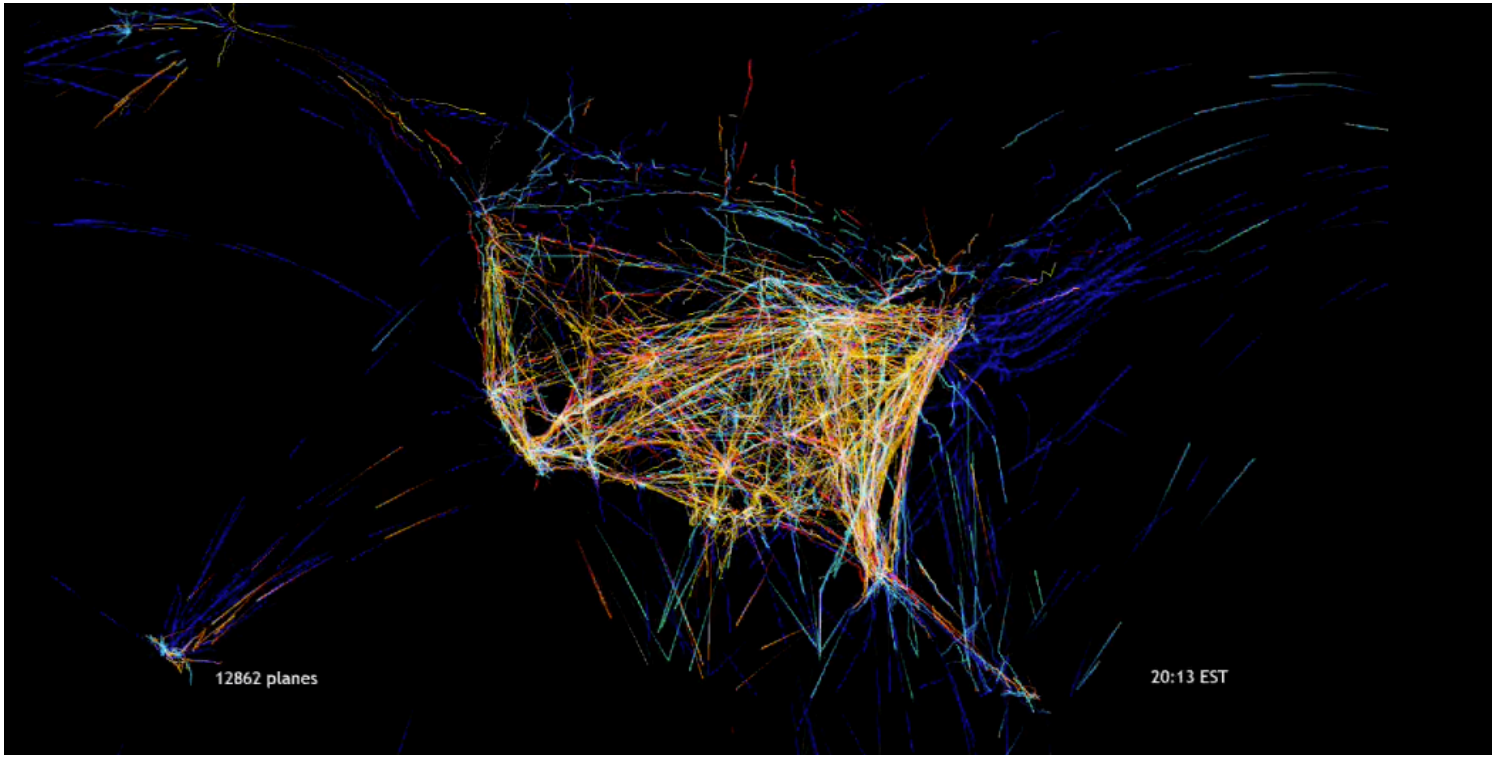
- Who is your audience?
- What do they need to know?
- What format will work?

Common Pitfalls

- Not understanding audience
  - Technical sophistication of most audiences is very low
  - Content knowledge among many audiences is very high
  - Data science presentations that emphasize technicality but miss basic context are very common

Common Pitfalls

- Pridefully obvious presentation



Source: https://vimeo.com/5368967

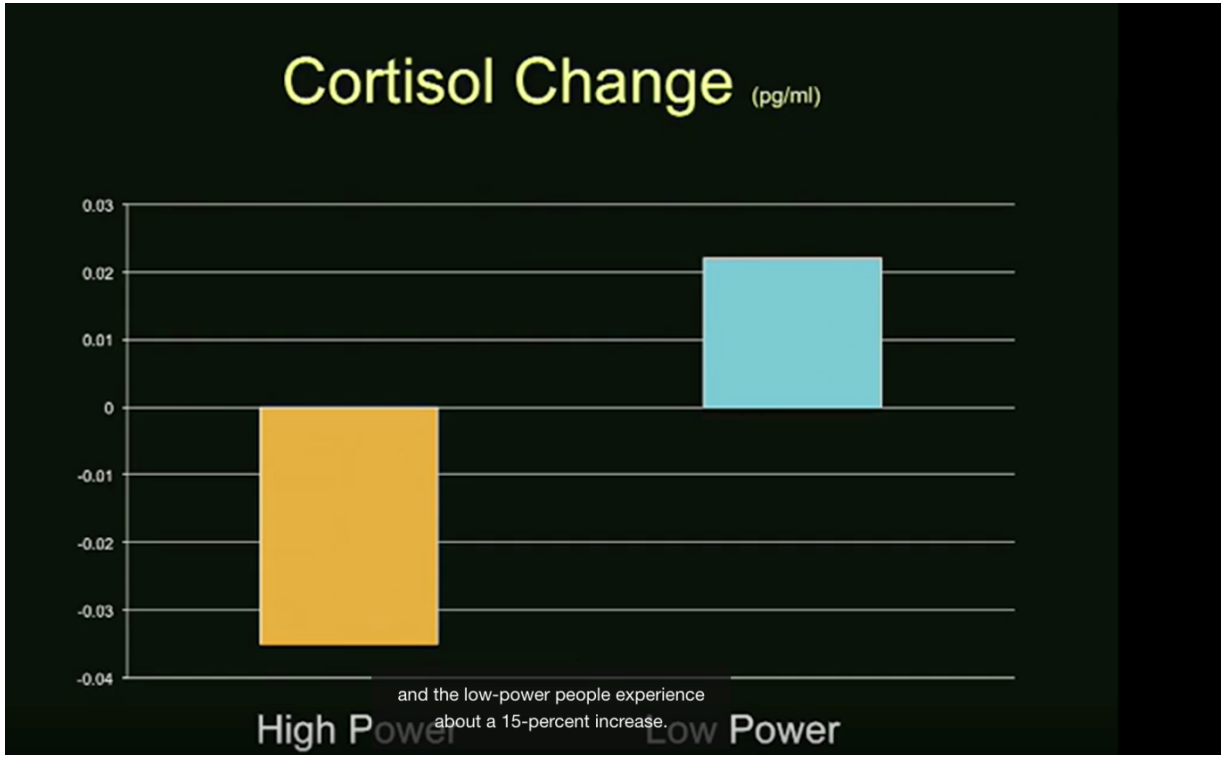
Common Pitfalls

- Failing to communicate uncertainty

Source: https://www.ted.com/talks/amy_cuddy_your_body_language_shapes_who_you_are?utm_campaign=tedspread&utm_medium=referral&utm_source=tedcomshare

Common Pitfalls

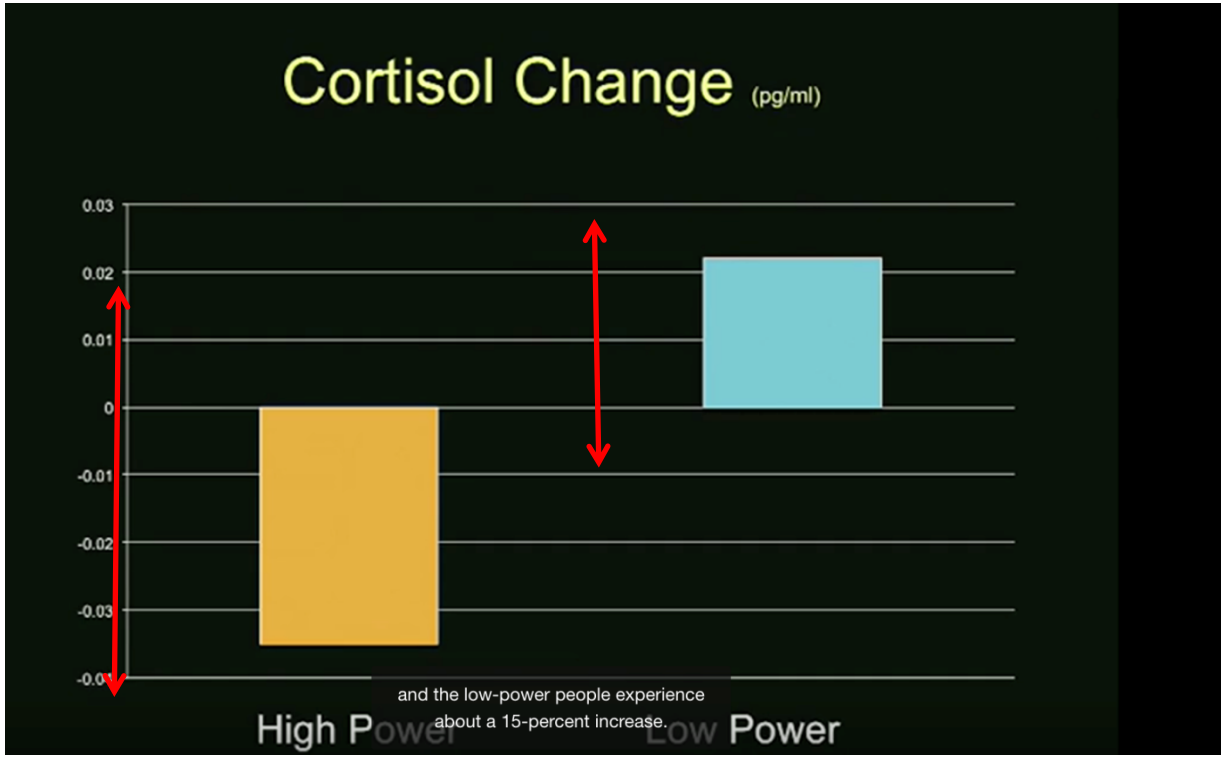
- Failing to communicating uncertainty



Source: https://www.ted.com/talks/amy_cuddy_your_body_language_shapes_who_you_are?utm_campaign=tedspread&utm_medium=referral&utm_source=tedcomshare

Common Pitfalls

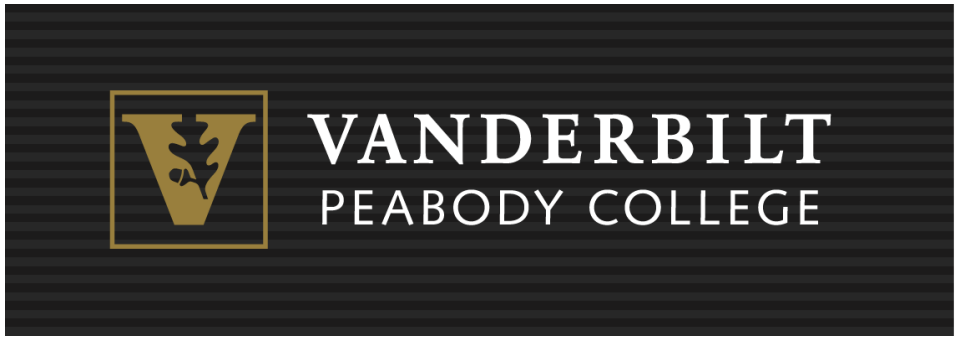
- Failing to communicating uncertainty



Source: https://www.ted.com/talks/amy_cuddy_your_body_language_shapes_who_you_are?utm_campaign=tedspread&utm_medium=referral&utm_source=tedcomshare

Where to Start

- Start with your conclusions
- Tell me; tell me how you know; tell me again
- Practice PGP
  - Particular
  - General
  - Particular


VANDERBILT
PEABODY COLLEGE

Who's Your Audience

Will Doyle

Three Dimensions for Audiences

1. Content knowledge
2. Statistical sophistication
3. Computing/data knowledge

Content Knowledge

- Most audiences have high content knowledge.
- It implies that the context is well known.
- Many times, it means that the audience knows the contours of the problem better than the analyst does.
- Questions to ask include:
  - Does this capture the nature of the case?
  - Will my audience trust the analysis?
  - Face validity: Did I miss something obvious?

Statistical Sophistication

- Audiences with statistical sophistication will have different concerns.
- Academic audiences tend to have the highest levels of statistical sophistication.
- Many times, audience will know of other (more arcane) ways to estimate models.
- Questions to ask include:
  - Are there other modeling procedures that I could have used?
  - Does my analysis satisfy (to some degree) the assumptions that underlie it?
  - Am I implying causality?

Computing/Data Knowledge

- Audiences who know computing/data analysis well will have yet another set of concerns.
- Programmers will want to know if the approach could have been made more general.
- Data analysts may likely be aware of better sources of data.
- Data analysts will also be acutely aware of data problems.
- Questions to ask include:
  - Did computing/programming limitations drive results?
  - What sources of data were considered?
    Why were some not used?
  - What level of confidence do we have in data quality?

Putting It Together

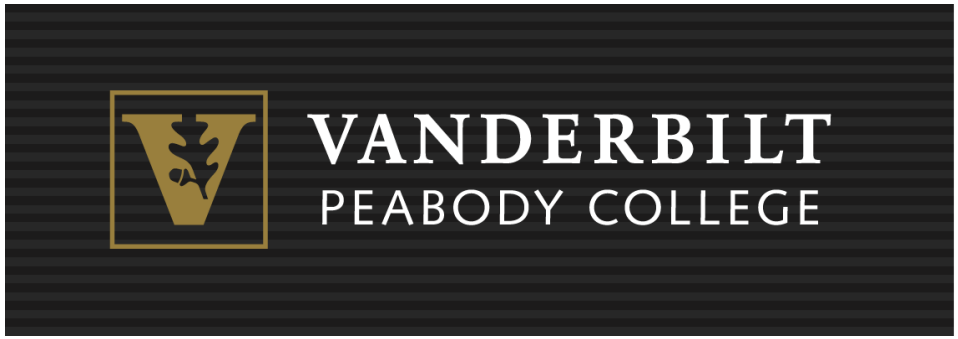
- The most common audience is one with high levels of content knowledge but low levels of statistical sophistication and computing/data knowledge.
- Be humble about the context—check for face validity.
- Communicate with an emphasis on clarity.
- Emphasize uncertainty when it exists.

Putting It Together

- The next most common audience has data/ computing knowledge and some content knowledge but, again, low levels of statistical sophistication.
- Emphasize the gain achieved by applying statistical models over much simpler approaches.
- Make sure you are on solid ground with data quality.
- Check face validity of assumptions.

Putting It Together

- Audiences with statistical sophistication are primarily academic.
- Describe content area clearly.
- Emphasize specifics of what needs to be answered.
- Avoid excessive confidence in any particular technique.
- Does this tell us (most) of what we need to know, as opposed to whether this is the perfect technique?


VANDERBILT
PEABODY COLLEGE

What Does the Audience Need to Know?

Will Doyle

Possible Goals for Reporting

- Provide context
- Change policy
- Planning

Providing Context

- Many times, the nature of the outcome is poorly understood.
- Answering basic questions about prevalence of the outcome (say employee attrition) among different groups is sufficient.
- Where is it most prevalent?
- Where is it least prevalent?
- Do we have any idea about the association between organization policies and prevalence of outcome?

Changing Policy

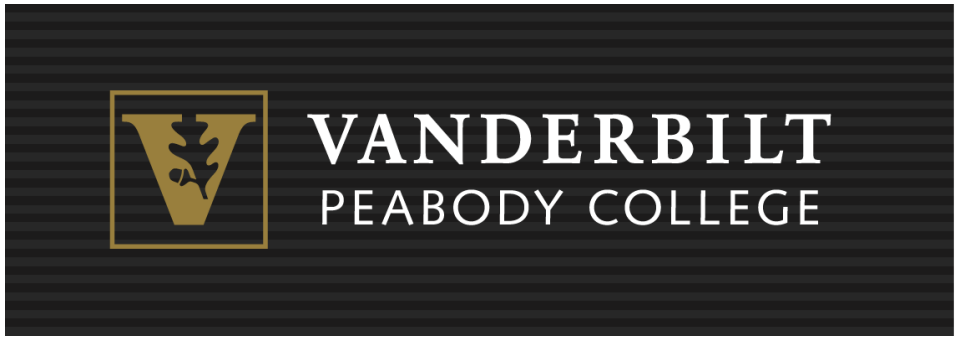
- This is also known as evaluation/policy analysis.
- The goal here is to say what would happen if a policy is changed.
- It implies causality, which requires experiments and A/B testing.
- What would be the impact of the policy change?
- Would it differ among key groups?

Planning

- Answering questions about what will happen in the future, based on current trends
- Not necessarily causal but based on past associations
- Provides decision makers with likely prevalence of the outcome at some future point, based on what we know today
- Example: Given the current characteristics of our employees, how many are likely to leave in the next year?

Goals: Concluding Thoughts

- Surprisingly “fuzzy” thinking around goals of many data analyses
- Ask decision makers for clarity
- Be clear what a given analysis can and can’t do (e.g., no evaluations without experiments)


VANDERBILT
PEABODY COLLEGE

What Format Should Be Used?

Will Doyle

Formats for Presenting Results

- Verbal reporting/informal methods
- Slide decks
- Briefing paper
- Full report
- Code

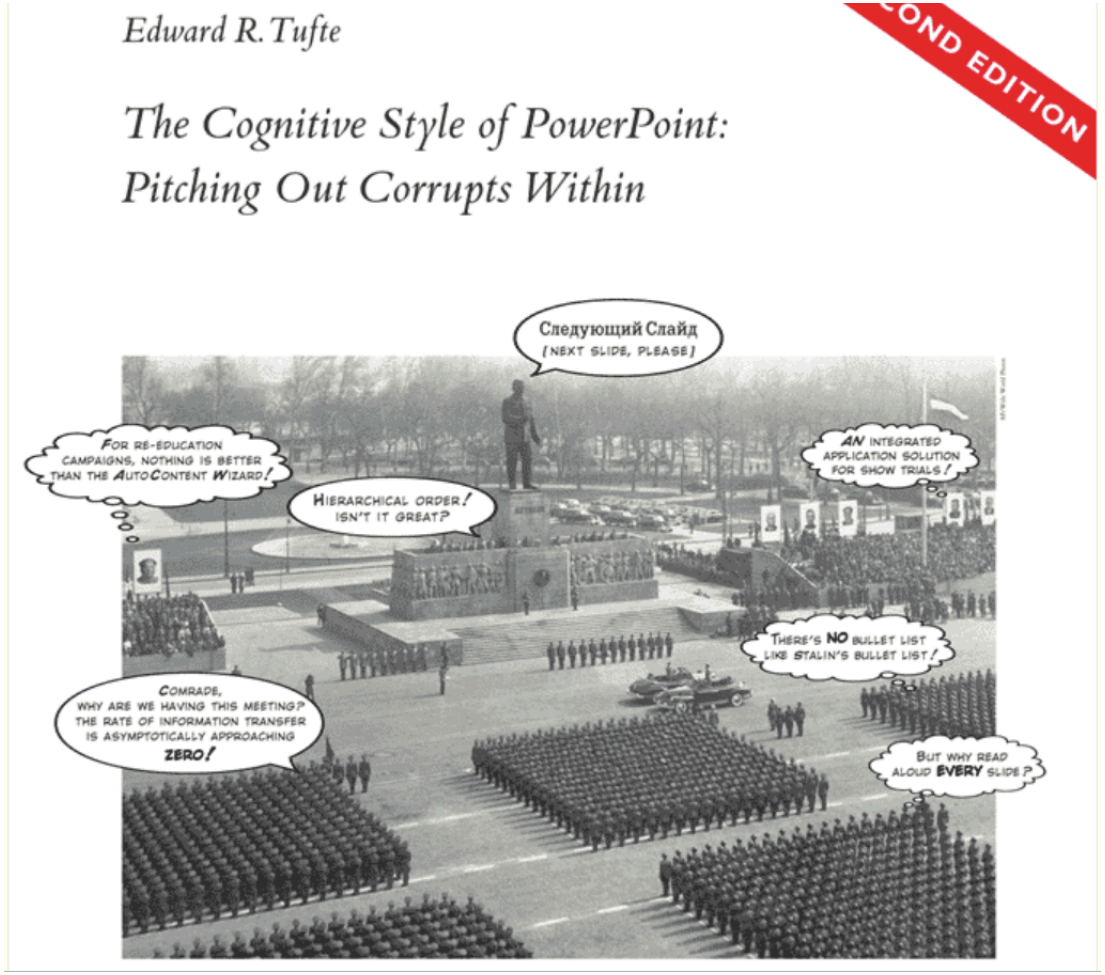
Verbal Reporting/Informal

- Quick questions in meetings, emails, Slack channels, and so on
- Quick answers are fine, but…
- **Document**
- Many times, these results propagate
- Make sure you can replicate the answer

Slide Decks

- By far, most common means of communication in organizations
- Simple structure
- Generally easy to follow
- But…

Slide Decks



Source: https://www.edwardtufte.com/tufte/powerpoint

On this one Columbia slide, a PowerPoint festival of bureaucratic hyper-rationalism, 6 different levels of hierarchy are used to display, classify, and arrange 11 phrases:

Level 1	Title of Slide
Level 2	● Very Big Bullet
Level 3	– big dash
Level 4	• medium-small diamond
Level 5	• tiny square bullet
Level 6	( ) parentheses ending level 5

The analysis begins with the dreaded Executive Summary, with a conclusion presented as a headline: "Test Data Indicates Conservatism for Tile Penetration." This turns out to be unmerited reassurance. Executives, at least those who don't want to get fooled, had better read far beyond the title.

The "conservatism" concerns the *choice of models* used to predict damage. But why, after 112 flights, are foam-debris models being calibrated during a crisis? How can "conservatism" be inferred from a loose comparison of a spreadsheet model and some thin data? Divergent evidence means divergent evidence, not inferential security. Claims of analytic "conservatism" should be viewed with skepticism by presentation consumers. Such claims are often a rhetorical tactic that substitutes verbal fudge factors for quantitative assessments.

As the bullet points march on, the seemingly reassuring headline fades away. Lower-level bullets at the end of the slide undermine the executive summary. This third-level point notes that "Flight condition [that is, the debris hit on the Columbia] is significantly outside of test database." How far outside? The final bullet will tell us.

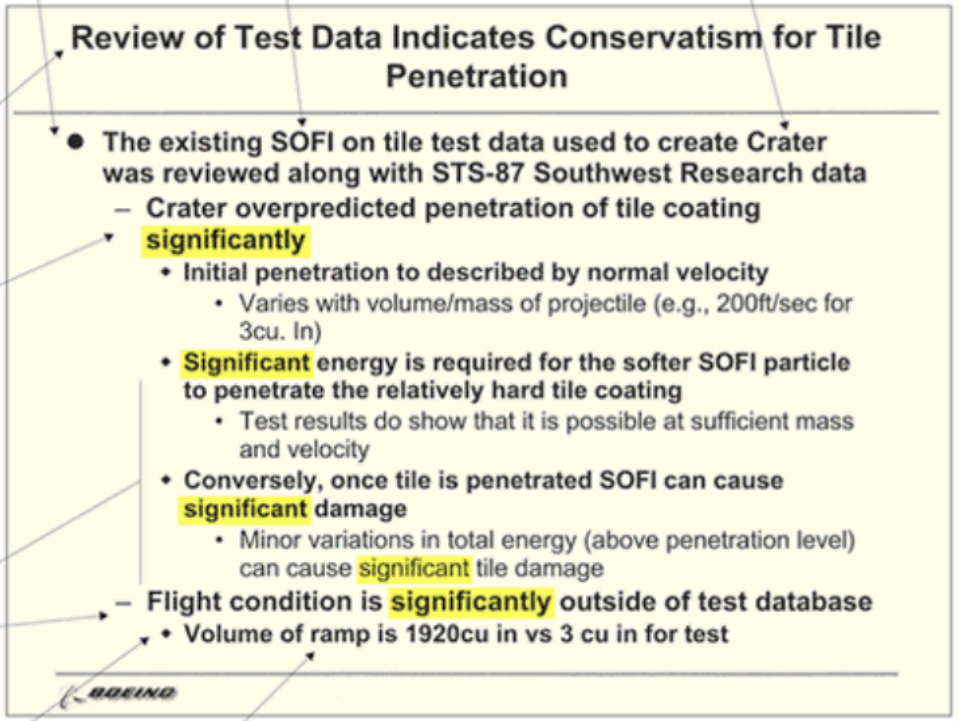
This fourth-level bullet concluding the slide reports that the debris hitting the Columbia is estimated to be 1920/3 = 640 times larger than data used in the tests of the model! The correct headline should be "Review of Test Data Indicates Irrelevance of Two Models." This is a powerful conclusion, indicating that pre-launch safety standards no longer hold. The original optimistic headline has been eviscerated by the lower-level bullets.

Note how close readings can help consumers of presentations

The Very-Big-Bullet phrase fragment does not seem to make sense. No other VBB's appear in the rest of the slide, so this VBB is not necessary.

Spray On Foam Insulation, a fragment of which caused the hole in the wing

A model to estimate damage to the tiles protecting flat surfaces of the wing



Here "ramp" refers to foam debris (from the bipod ramp) that hit Columbia. Instead of the cryptic "Volume of ramp," say "estimated volume of foam debris that hit the wing." Such clarifying phrases, which may help upper level executives understand what is going on, are too long to fit on low-resolution bullet outline formats. PP demands the shorthand of acronyms, phrase fragments, and clipped jargon in order to get at

Source: https://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0001yB:

Slide Decks

- Slide decks should be based on a more comprehensive report that can be supplied.
- If the slide deck must stand alone, then it can be more text-oriented.
- If the slide deck is purely to support verbal presentation, the orientation should be toward high-quality graphics.

Briefing

- Generally, best way to report results
- Set maximum length: Amazon uses six pages, single-spaced
- Length includes tables and graphics

Proposed structure

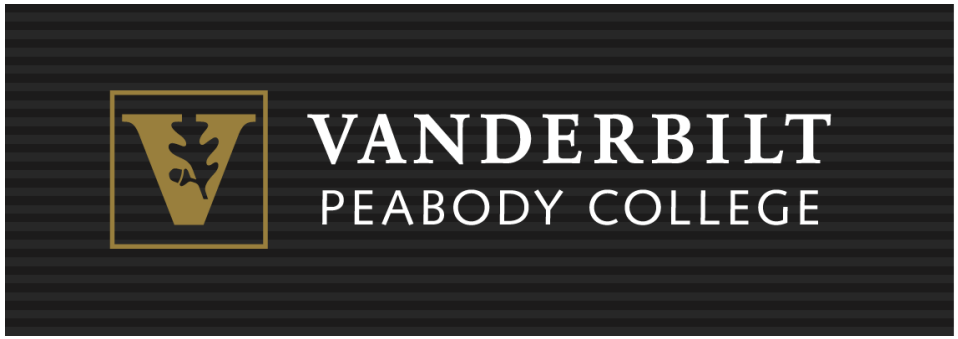
- What do we need to know?—problem statement
- Bottom line: the answer (by end of first page)
- Data: Where did we get it? Is it any good?
- Exploratory data analysis (conditional means)
- Predictive model: What was used? What assumptions are made?
- Accuracy of model (results from cross-validation)
- Predictions from model
- Recommendations

Full Report

- If the goal is to create a sustainable project, a full report should be created
- A briefing paper serves as an executive summary
- Sections are based on the briefing paper
- It includes much more detail and narrative, providing a roadmap for continued development

Code

- Your code is the best guide for replication.
- The code must be primarily readable by humans, secondarily by computers.
- It should be carefully documented.
- This class has emphasized literate programming: combining narrative and code in a single document (.Rmd file for us).
- Well-documented code should be part of the product.


VANDERBILT
PEABODY COLLEGE

Predicting College Enrollment

Setup

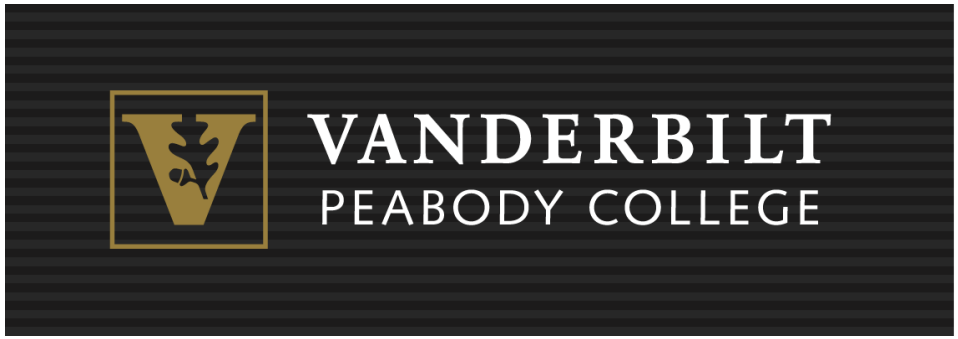
Will Doyle

Predicting College Enrollment

- Large, multi-campus public college system
- Given characteristics of applicants, how many will enroll next year?
  - Tuition
  - State appropriations
  - Staffing levels

What Data Will You Need

- Unit of analysis
- Coverage
- Dependent variable
- Independent variables


VANDERBILT
PEABODY COLLEGE

Predicting College Enrollment

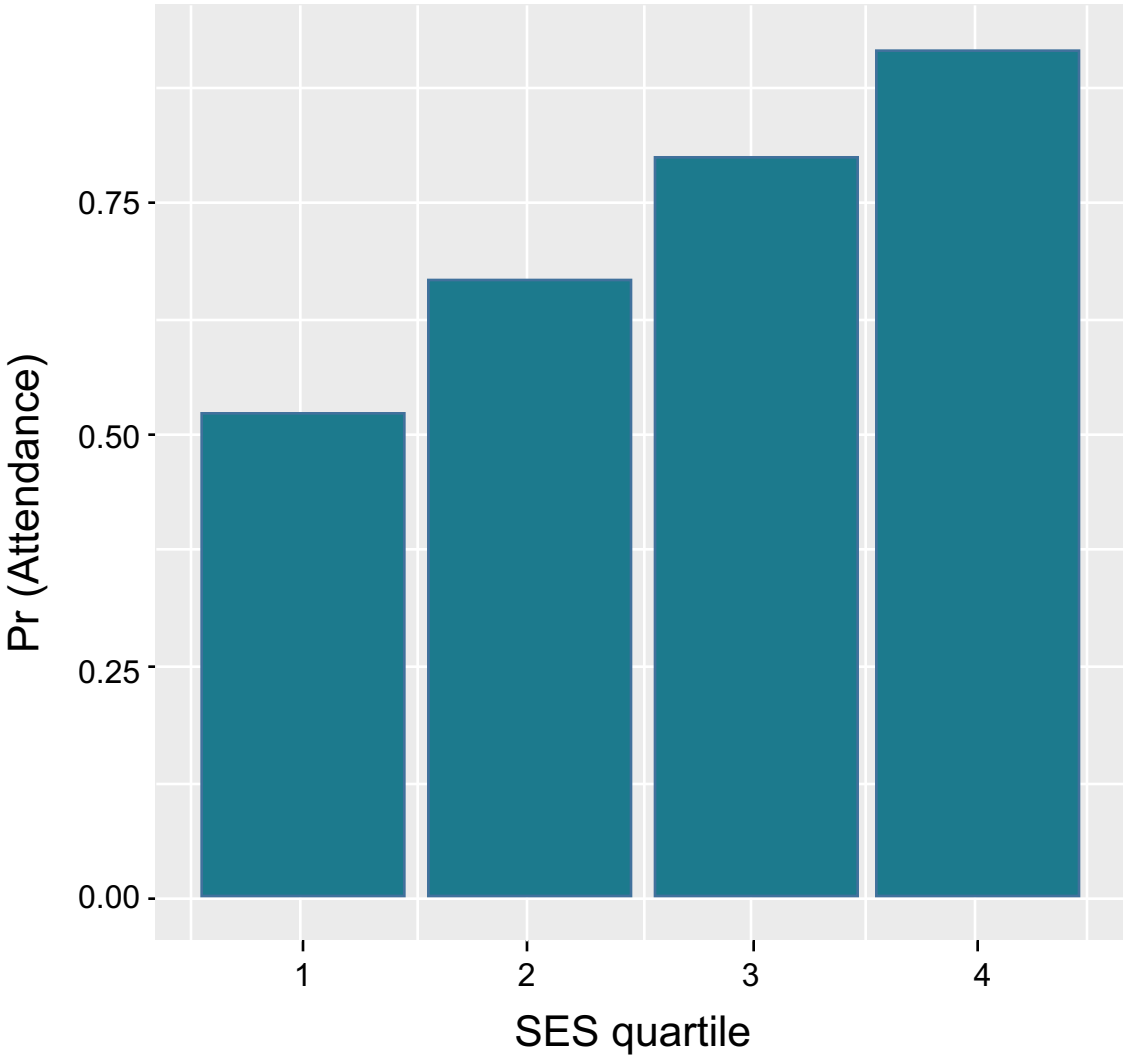
Data

Will Doyle

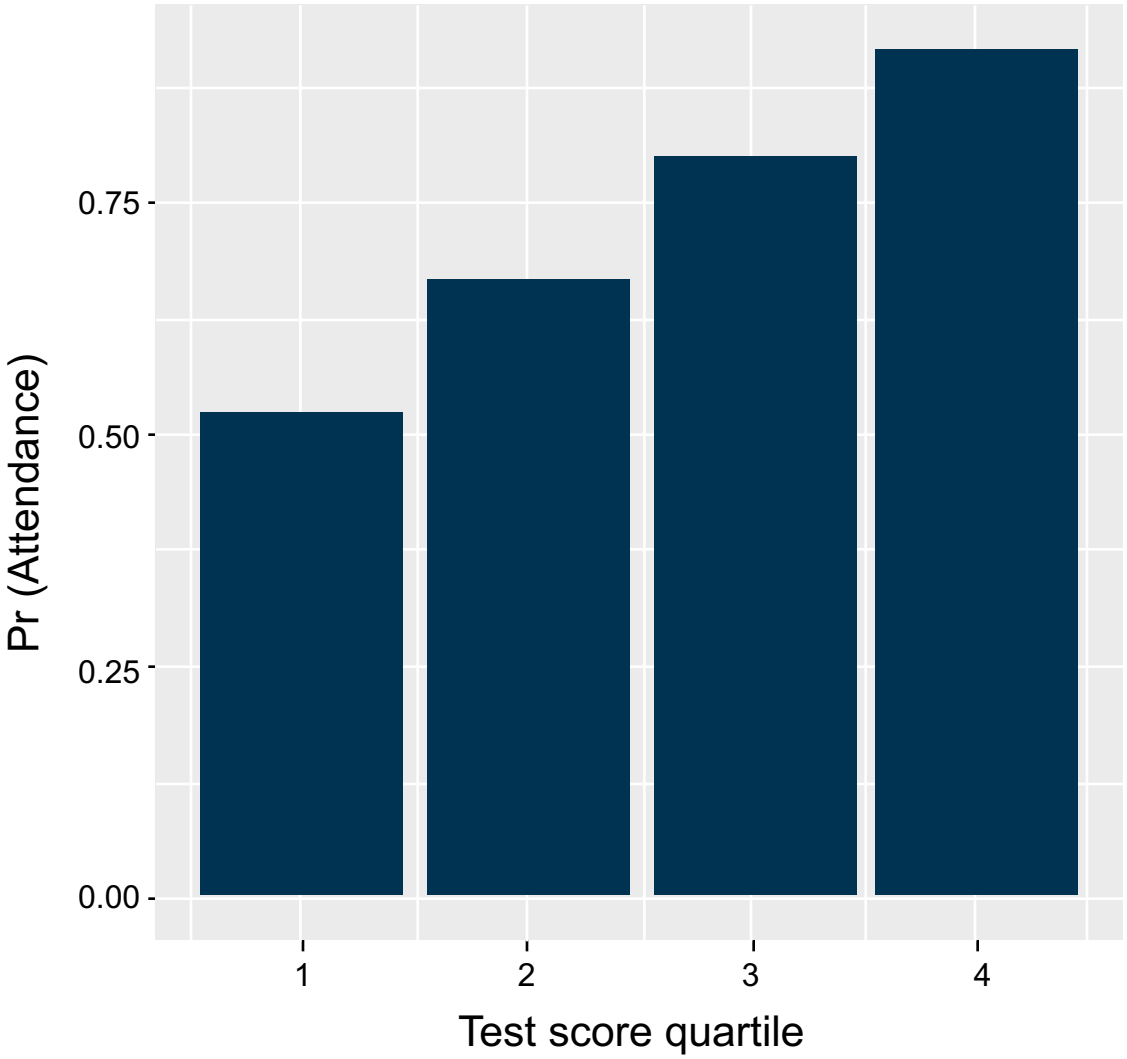
What Data Will You Need

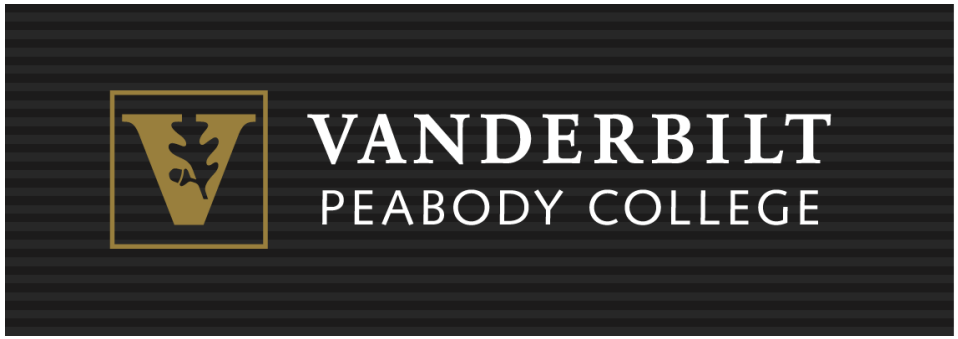
- Unit of analysis: **individual student**
- Coverage: **full sample of applicants**
- Dependent variable: **enrolled in system (0,1)**
- Independent variables
  - **SES**
  - **Race**
  - **Gender**
  - **Test scores**

Basic Conditional Means



Basic Conditional Means




VANDERBILT
PEABODY COLLEGE

Predicting College Enrollment

Analysis

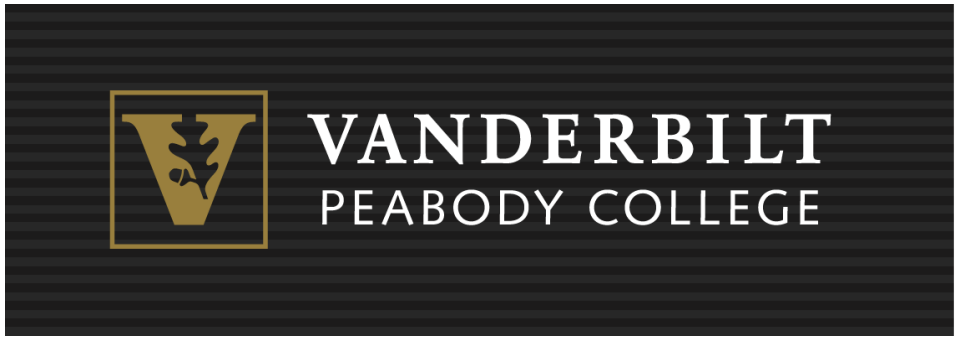
Will Doyle

What Analysis to Conduct

- Binary dependent variable
- Classification problem
- Logistic regression is the best approach
- Dependent variable: f2evratt, 0 = did not attend, 1 = attended
- Independent variable: race, sex, SES, test scores

Results

```
##
## Call:
## glm(formula = f2evratt ~ byses1 + bynels2m + bynels2r + amind +
##     asian + hispanic + multiracial + bysex, family = binomial(link = "logit"),
##     data = enr)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9287   0.1410   0.4181   0.6907   2.1422
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.390211   0.116786 -20.467  < 2e-16 ***
## byses1       0.873086   0.037152  23.500  < 2e-16 ***
## bynels2m     0.042924   0.002672  16.062  < 2e-16 ***
## bynels2r     0.029531   0.003787   7.797 6.33e-15 ***
## amind       -0.433130   0.218149  -1.985   0.0471 *
## asian        0.890845   0.097229   9.162  < 2e-16 ***
## hispanic     0.146637   0.063340   2.315   0.0206 *
## multiracial -0.403646   0.100146  -4.031 5.56e-05 ***
## bysex        0.578326   0.046907  12.329  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```


VANDERBILT
PEABODY COLLEGE

Predicting College Enrollment

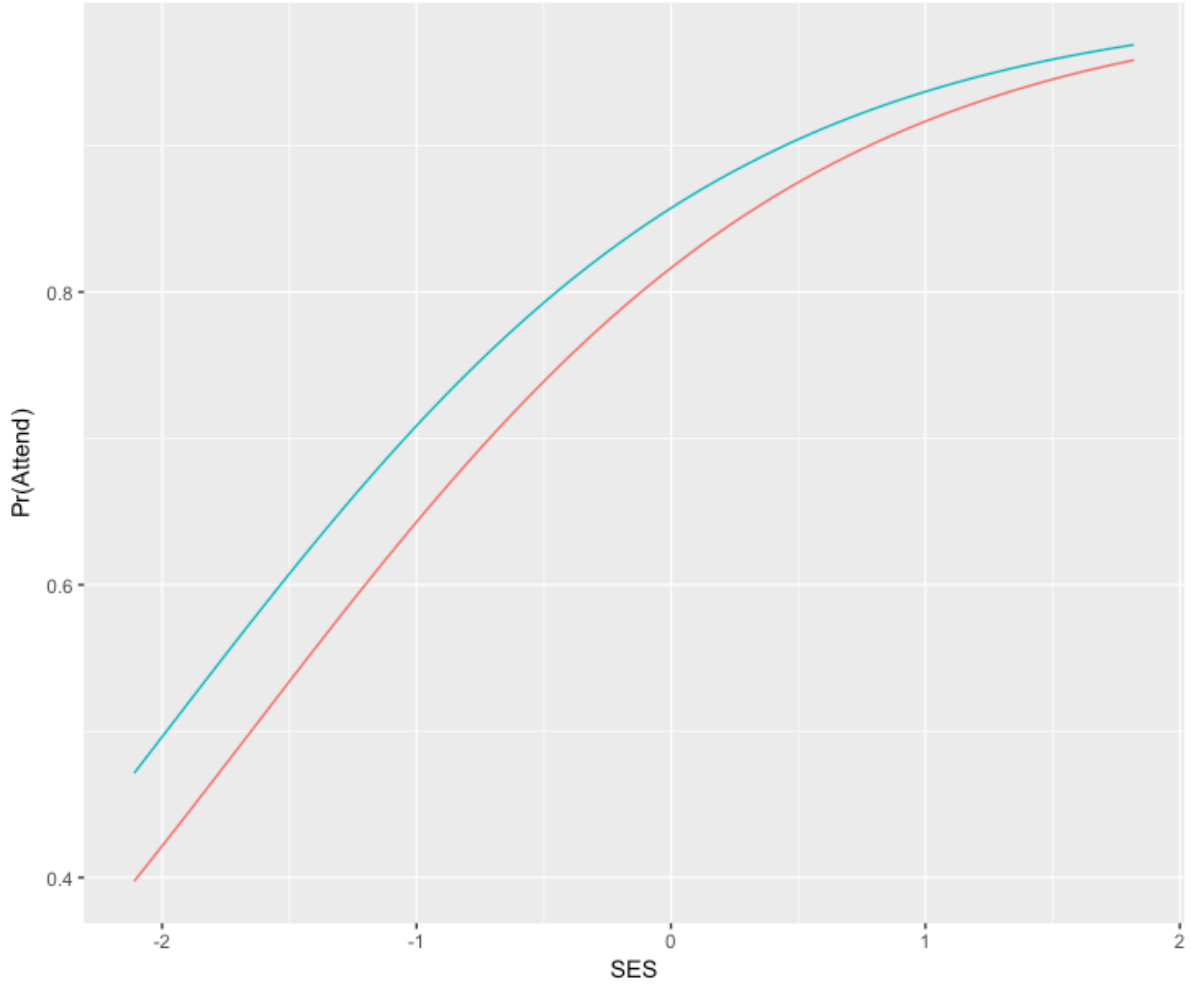
Results

Will Doyle

Basic Outline

- SES and test scores are strongly positively associated with enrollment.
- If we have more high-SES, high-test-score applicants, our enrollments will be higher.
- All else being equal, African American, Hispanic, and Asian students are more likely to enroll.
- All else being equal, female students are more likely to enroll.

Plotting Probabilities

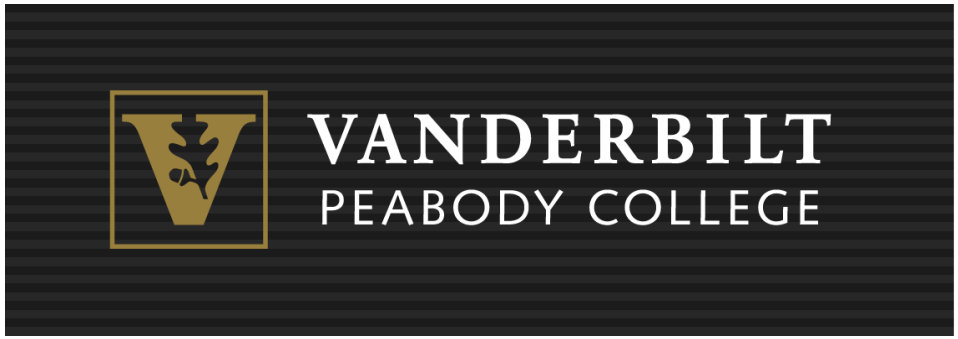


Evaluating Model Fit

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction    0    1
##          0 1175  719
##          1 2068 9322
##
##                Accuracy : 0.7902
##                  95% CI : (0.7832, 0.7971)
##     No Information Rate : 0.7559
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3384
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.36232
##             Specificity : 0.92839
```

Evaluating Model Fit

- Percent correctly predicted = 79%
- Specificity (who isn't going) is **much** better than sensitivity (who **is** going)
- AUC = .8—not great, but good
- Model is reasonably good at predicting nonenrollment—may need to adjust to get better at predicting enrollment


VANDERBILT
PEABODY COLLEGE