# a first-year undergraduate data science course

Slides at http://bit.ly/first-ds


@minebocek

mine-cetinkaya-rundel

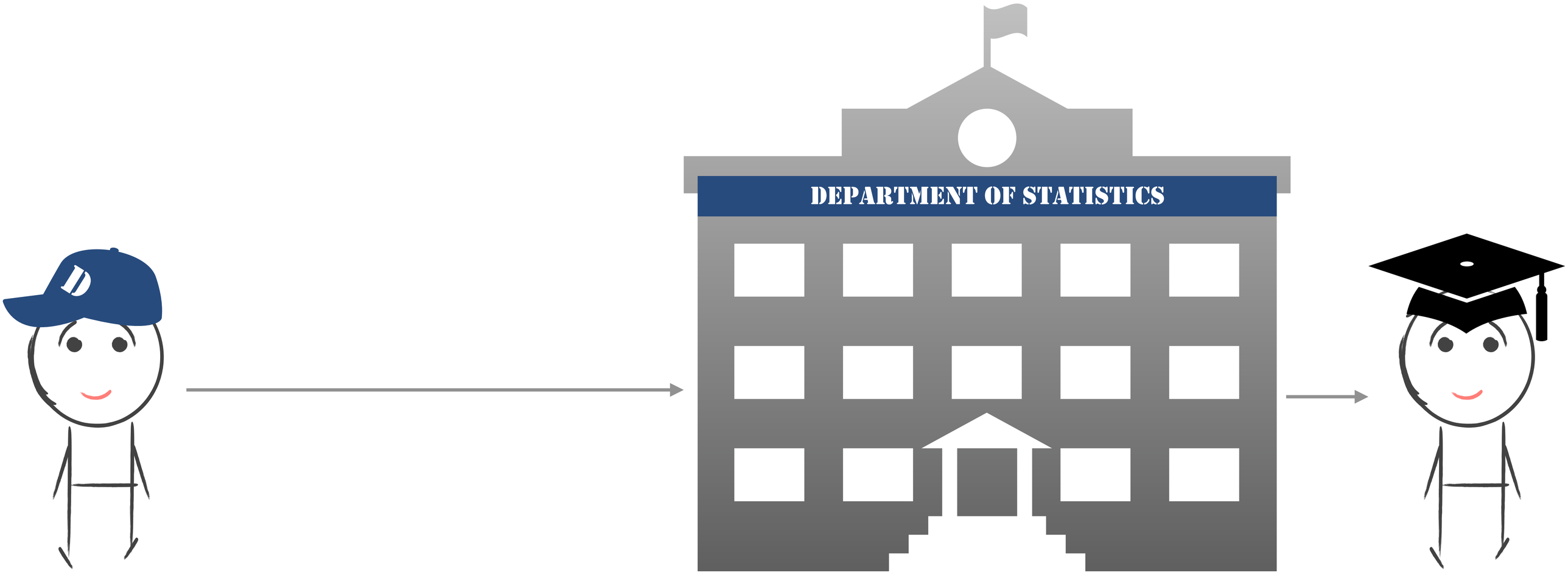
mine@stat.duke.edu

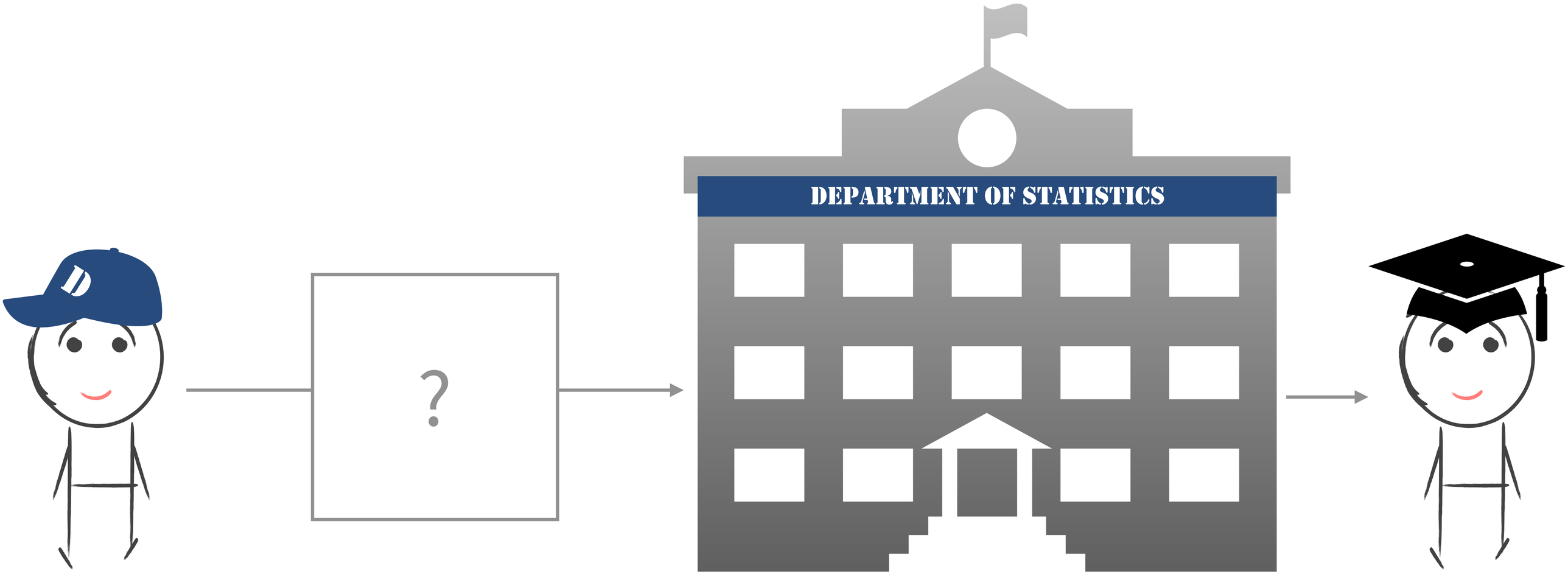
mine çetinkaya-rundel

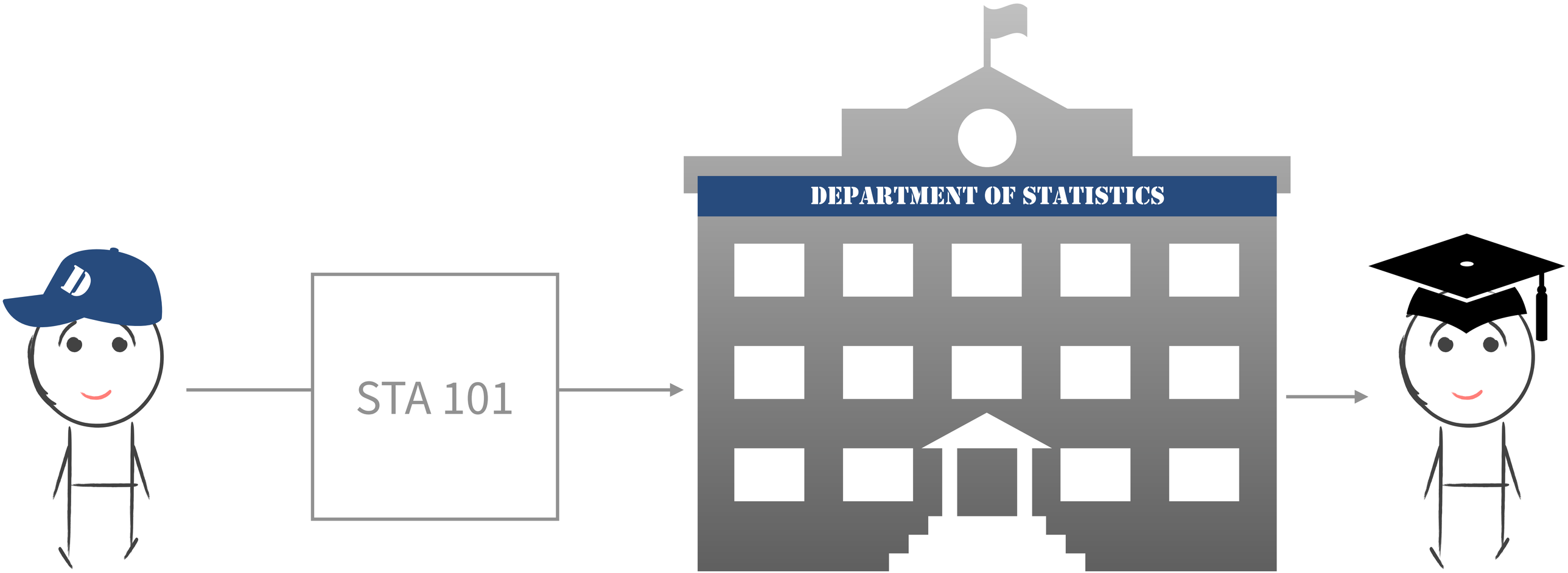
Duke UNIVERSITY

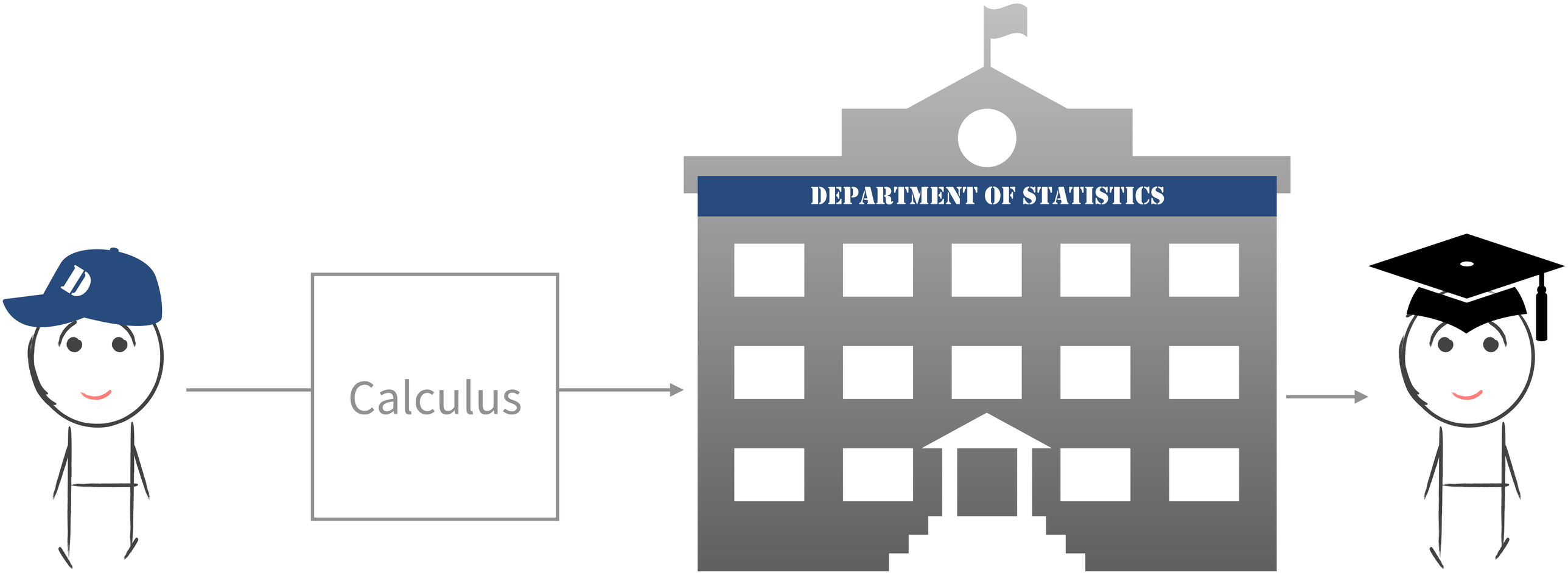
RStudio

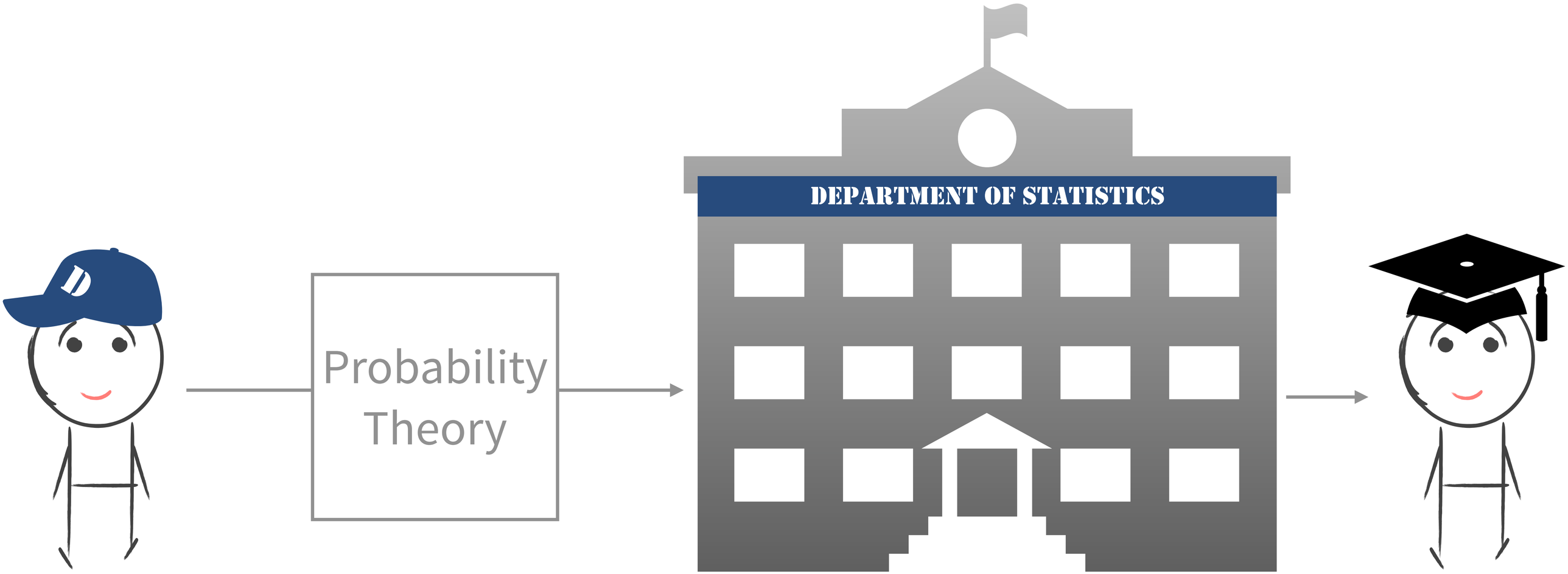
R Ladies

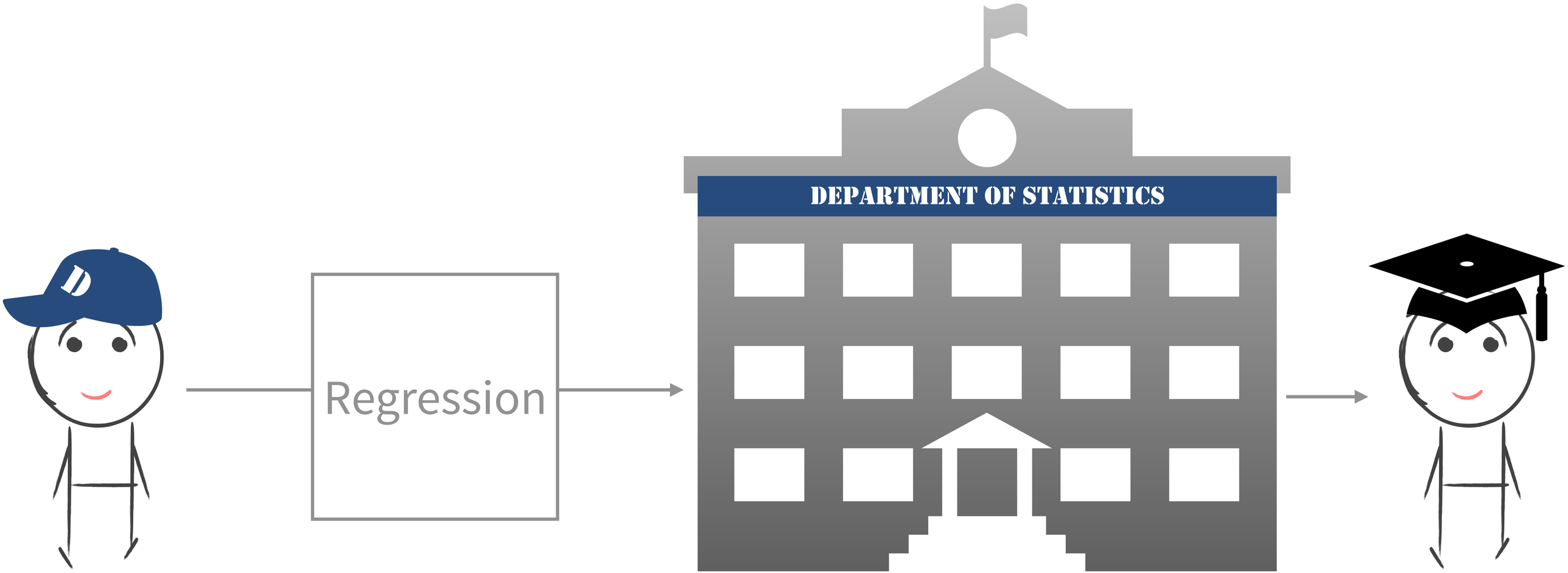

DEPARTMENT OF STATISTICS

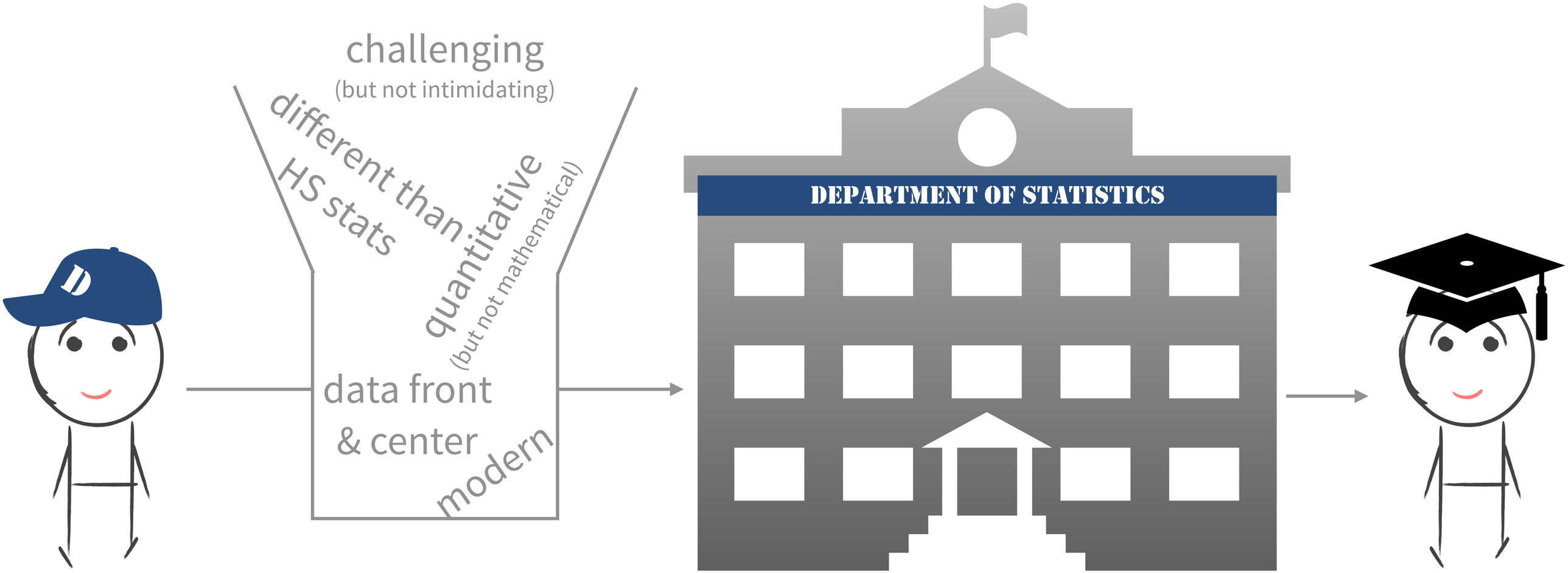

DEPARTMENT OF STATISTICS
?


STA 101
DEPARTMENT OF STATISTICS


DEPARTMENT OF STATISTICS
Calculus


Probability
Theory
DEPARTMENT OF STATISTICS


Regression
DEPARTMENT OF STATISTICS


challenging
(but not intimidating)
different than HS stats
quantitative
(but not mathematical)
data front & center
modern
DEPARTMENT OF STATISTICS

# this course should...

emphasize modern and multivariate EDA + data visualization

start at the beginning of data analysis cycle with data collection and cleaning

encourage + enforce working collaboratively (think, code, write, present)

teach (not just expect) reproducible computation

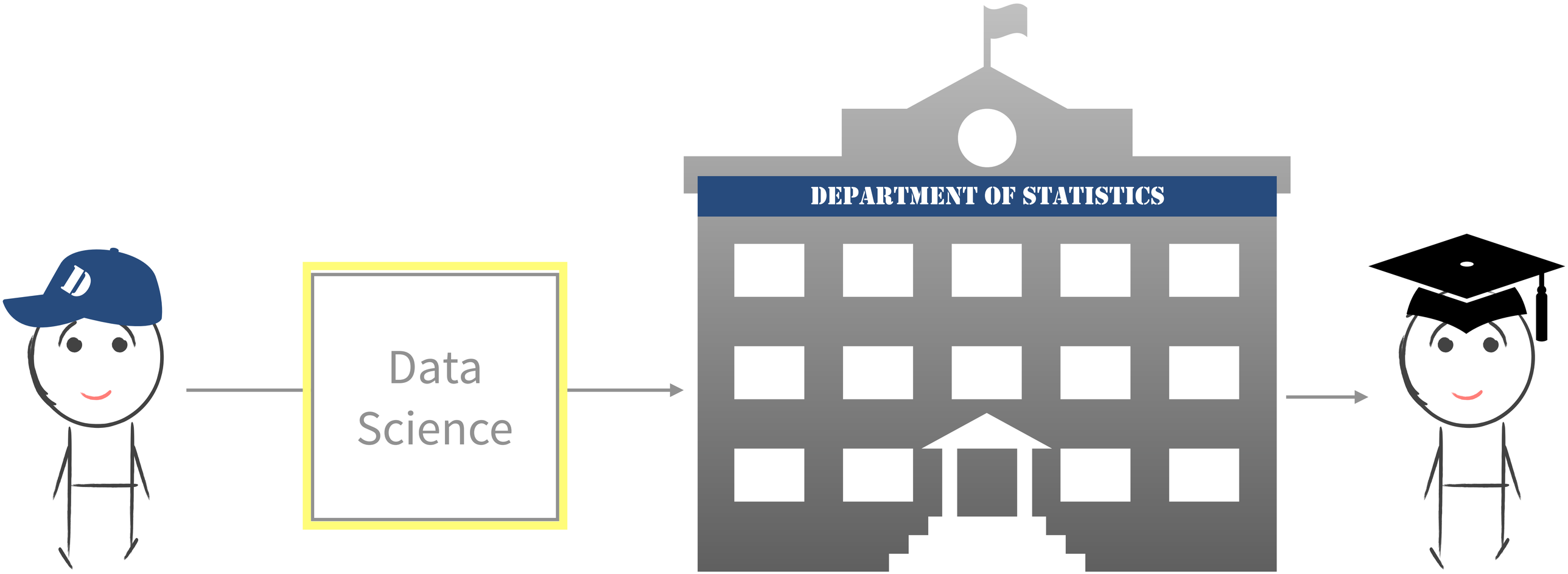
approach statistics from a model based perspective

underscore effective communication of findings

## and maybe more importantly...

ask questions that students want to answer

equip students with the tools to answer questions of their own choosing


DEPARTMENT OF STATISTICS
Data
Science

outline

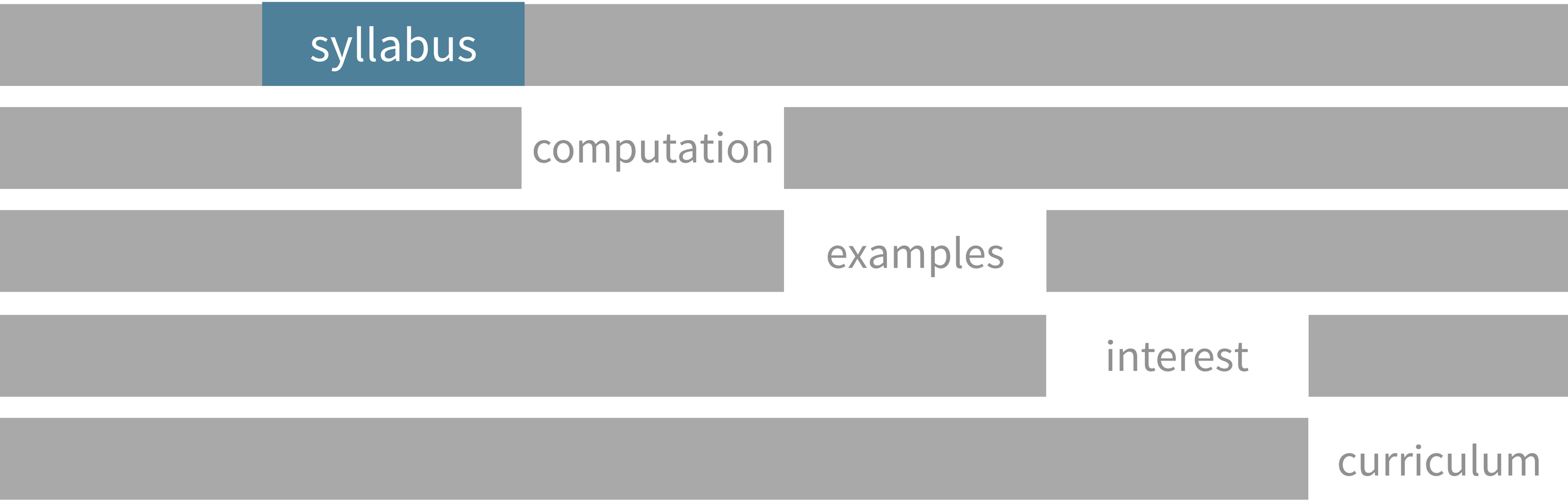
syllabus

computation

examples

interest

curriculum


syllabus
computation
examples
interest
curriculum

# course overview

**curriculum:**

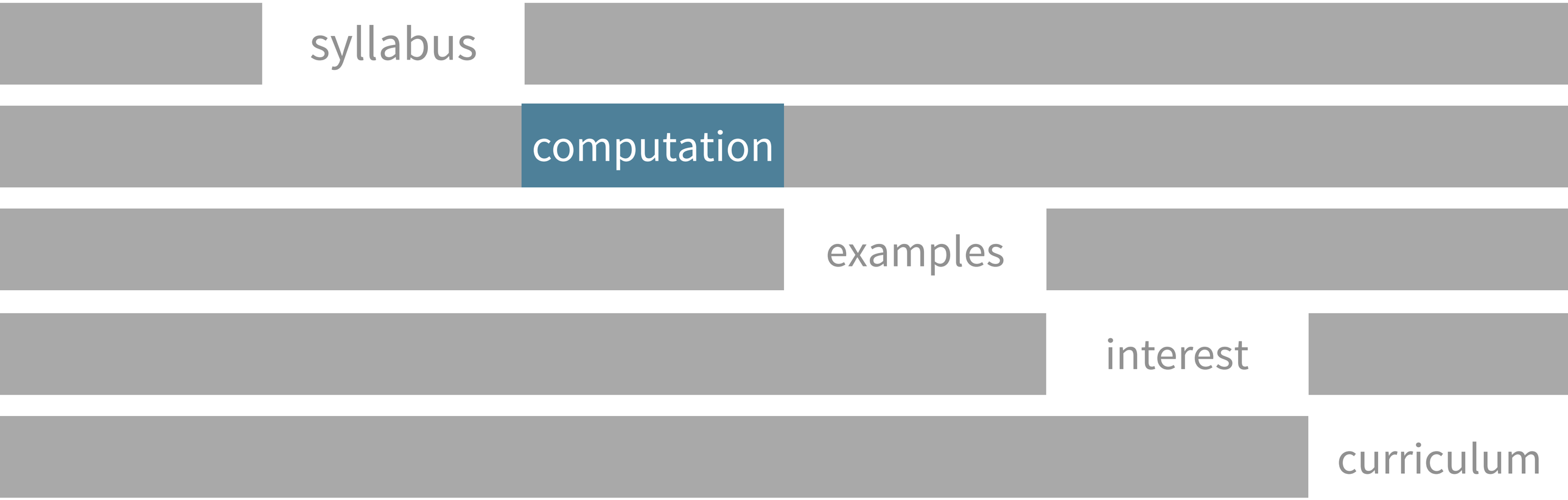
data gathering + wrangling, EDA + viz, multivariate modeling, basic inference, communication

**structure:**

*teams:* in class exercises + projects
*indivudual:* HW + take home midterm and final

**assessment:**

not just final work but also the process, peer evaluations and contribution diagnostics


syllabus
computation
examples
interest
curriculum

core

**what:**

R + RStudio server

**why:**

minimize onboarding friction and time to 1st data viz

ex: start up instrutsions

## # Local install

- Install R: https://cran.r-project.org/
- Install RStudio: https://www.rstudio.com/products/rstudio/
- Install the following packages:
  - rmarkdown
  - knitr
  - tidyverse
  - …
- Load these packages

vs.

## # RStudio Server

- Go to smith.stat.duke.edu:8787
- Log in with your Net ID & pass

# data analysis toolkit

**what:**

(mostly)
`tidyverse`

**why:**

(closer to) human readable, consistent syntax, easy multivariate visualizations

# ex: recoding a binary variable

## # base R

```
mtcars$transmission <-
  ifelse(mtcars$am == 0,
         "automatic",
         "manual")
```

vs.

## # tidyverse

```
mtcars <- mtcars %>%
  mutate(
    transmission =
      case_when(
        am == 0 ~ "automatic",
        am == 1 ~ "manual"
      )
  )
```

# ex: recoding a multi-lelvel variable

## # base R

```
mtcars$gear_char <-
  ifelse(mtcars$gear == 3,
         "three",
         ifelse(mtcars$gear == 4,
                "four",
                "five")
  )
```

vs.

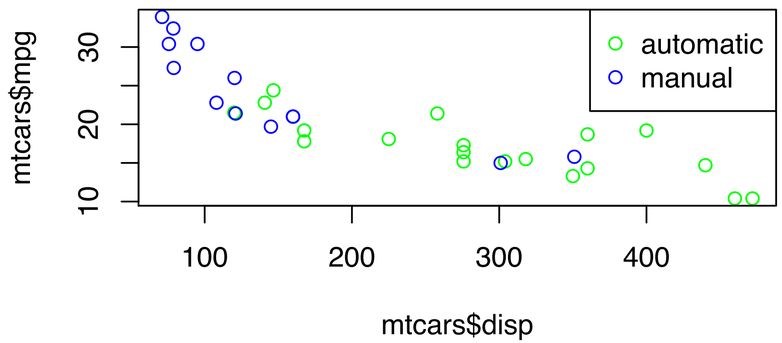
## # tidyverse

```
mtcars <- mtcars %>%
  mutate(
    gear_char =
      case_when(
        gear == 3 ~ "three",
        gear == 4 ~ "four",
        gear == 5 ~ "five"
      )
  )
```

# ex: visualizing multiple variables

## # base R

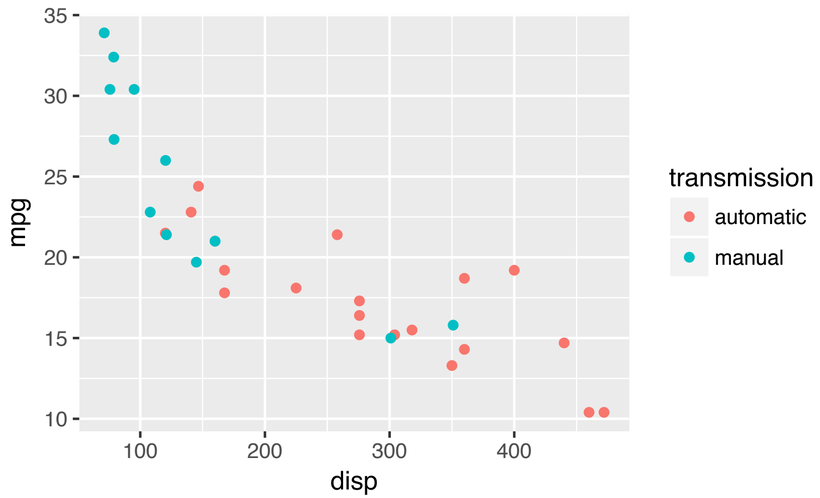
```
mtcars$trans_color <-
  ifelse(mtcars$transmission == "automatic",
         "green",
         "blue")

plot(mtcars$mpg ~ mtcars$disp,
     col = mtcars$trans_color)
legend("topright",
       legend = c("automatic", "manual"),
       pch = 1, col = c("green", "blue"))
```




vs.

## # tidyverse

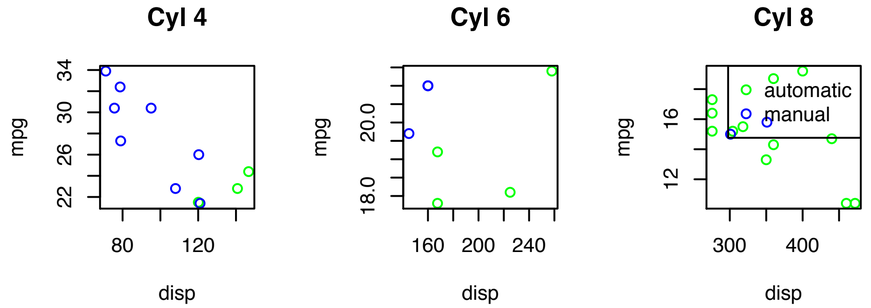
```
ggplot(mtcars,
       aes(x = disp, y = mpg,
           color = transmission)) +
  geom_point()
```



# ex: visualizing even more variables

## # base R

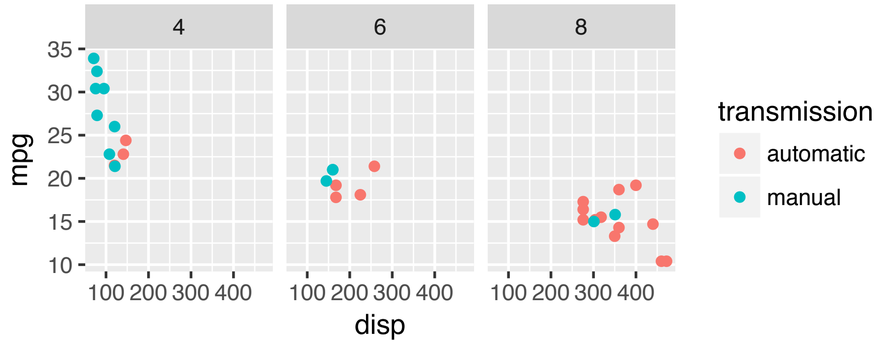
```
mtcars_cyl4 = mtcars[mtcars$cyl == 4, ]
mtcars_cyl6 = mtcars[mtcars$cyl == 6, ]
mtcars_cyl8 = mtcars[mtcars$cyl == 8, ]

par(mfrow = c(1, 3))
plot(mpg ~ disp, data = mtcars_cyl4,
     col = trans_color, main = "Cyl 4")
plot(mpg ~ disp, data = mtcars_cyl6,
     col = trans_color, main = "Cyl 6")
plot(mpg ~ disp, data = mtcars_cyl8,
     col = trans_color, main = "Cyl 8")
legend("topright",
       legend = c("automatic", "manual"),
       pch = 1, col = c("green", "blue"))
```




vs.

## # tidyverse

```
ggplot(mtcars,
       aes(x = disp, y = mpg,
           color = transmission)) +
  geom_point() +
  facet_wrap(~ cyl)
```



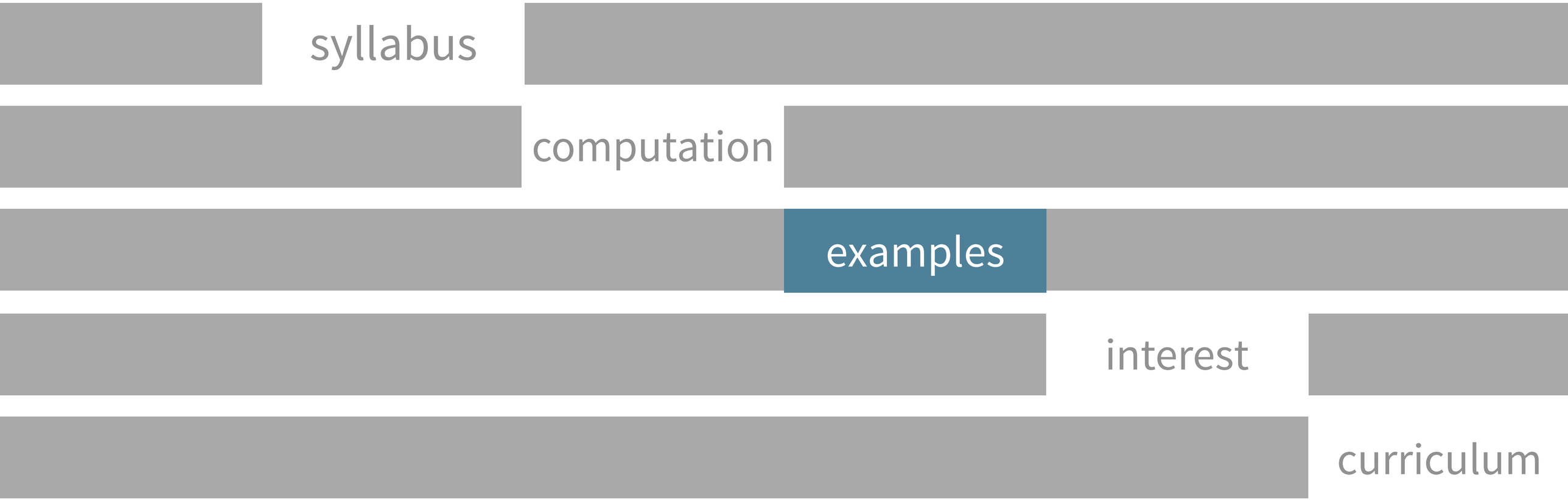
# reproducibility

**what:**

R Markdown +
Git / GitHub

**why:**

train new analysts whose only workflow is a reproducible one


syllabus
computation
examples
interest
curriculum

# #1 paris paintings


Sandra Van Ginhoven
PhD, Art History


Hilary Coe Cronheim
PhD, Art History

data source: auction catalogs

89 Deux tableaux très riches de compoſition, d'une belle exécution, & dont le mérite eſt très remarquable, chacun de 17 pouces 3 lignes de haut, ſur 23 pouces de large; le premier, peint ſur bois, vient du Cabinet de Madame la Comteſſe de Verrue; il repréſente un départ pour la chaſſe : on y voit ſur le devant un

1066 -

enrich

Le ſecond tableau, qui eſt ſur toile, fait voir un terrein d'une grande étendue, près la mer qui eſt à gauche, & ſur laquelle ſont des vaiſſeaux : on y voit auſſi des bagages que l'on décharge d'un charriot, des hommes, des femmes, des enfants, deux chevaux qui mangent, & des mulets chargés de bagages.

*Two paintings very rich in composition, of a beautiful execution, and whose merit is very remarkable, each 17 inches 3 lines high, 23 inches wide; the first, painted on wood, comes from the Cabinet of Madame la Comtesse de Verrue; it represents a departure for the hunt: it shows in the front a child on a white horse, a man who gives the horn to gather the dogs, a falconer and other figures nicely distributed across the width of the painting; two horses drinking from a fountain; on the right in the corner a lovely country house topped by a terrace, on which people are at the table, others who play instruments; trees and fabriques pleasantly enrich the background.*

data transcription

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | name | sale | lot | dealer | year | origin_ author | origin_ cat | school_ pntg | diff_origin | price | count | subject | authorstandard | artistliving | authorstyle | author | winnir |
| 2517 | R1777-86 | R1777 | 86 | R | 1777 | D/FL | D/FL | D/FL | 0 | 620.0 | 1 | 2 femmes, enfants, paysage vu à travers une arcade | Bega, Cornelis Pieterszoon | 0 | n/a | Corneille Bega | Lebrun |
| 2518 | R1777-87 | R1777 | 87 | R | 1777 | D/FL | D/FL | D/FL | 0 | 12,000.0 | 1 | Course du hareng | Wouwerman, Philips | 0 | n/a | Philippe Wouwerman | Donjeu |
| 2519 | R1777-88 | R1777 | 88 | R | 1777 | D/FL | D/FL | D/FL | 0 | 8,000.0 | 1 | Paysage sablonneux | Wouwerman, Philips | 0 | n/a | Philippe Wouwerman | Lambe |
| 2520 | R1777-89a | R1777 | 89 | R | 1777 | D/FL | D/FL | D/FL | 0 | 5,300.0 | 1 | Départ pour la chasse | Wouwerman, Philips | 0 | n/a | Philippe Wouwerman | Langlie |

| | R | S | T | U | V | W | X | Y | Z | AA | AB | AC | AD | AE | AF | AG | AH |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | winningbidder | winningbiddertype | endbuyer | Interm | type_interm ed | Height_in | Width_in | Surface_Rect | Diam_in | Surface_Rn d | Shape | Surface | material | mat | quantity | nfigures | engraved |
| 2516 | Feuillet | D | D | 0 | | 16 | 20 | 320 | | | squ_rect | 320 | toile | t | 1 | 0 | 0 |
| 2517 | Lebrun, Jean-Baptiste-Pierre | D | D | 0 | | 13.25 | 11 | 145.75 | | | squ_rect | 145.75 | bois | b | 1 | 0 | 0 |
| 2518 | Donjeux, Vincent | D | D | 0 | | 23 | 29.25 | 672.75 | | | squ_rect | 672.75 | toile | t | 1 | 50 | 0 |
| 2519 | Lambert, John (Chevalier Lambert) | C | C | 0 | | 23 | 30 | 690 | | | squ_rect | 690 | toile | t | 1 | 0 | 1 |
| 2520 | Langlier, Jacques for Poullain, Antoine | DC | C | 1 | D | 17.25 | 23 | 396.75 | | | squ_rect | 396.75 | bois | b | 1 | 0 | 0 |

| | AG | AH | AI | AJ | AK | AL | AM | AN | AO | AP | AQ | AR | AS | AT | AU | AV | AW | AX | AY | AZ | BA | BB | BC | BD | BE | BF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | nfigures | engraved | original | prevcoll | othartist | paired | figures | finished | lrgfont | relig | landsALL | lands_ sc | lands_ figs | lands_ment | arch | mytho | peasant | othgenre | singlefig | portrait | still_life | discauth | history | allegory | pastorale | other |
| 2516 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2517 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2518 | 50 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2519 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2520 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

paris paintings

**data:**

painting
auction data
1764 - 1780
[3,393 x 57]

**visualize:**

data visualization to explore patterns and possible interactions

**clean:**

data cleaning and wrangling

**model:**

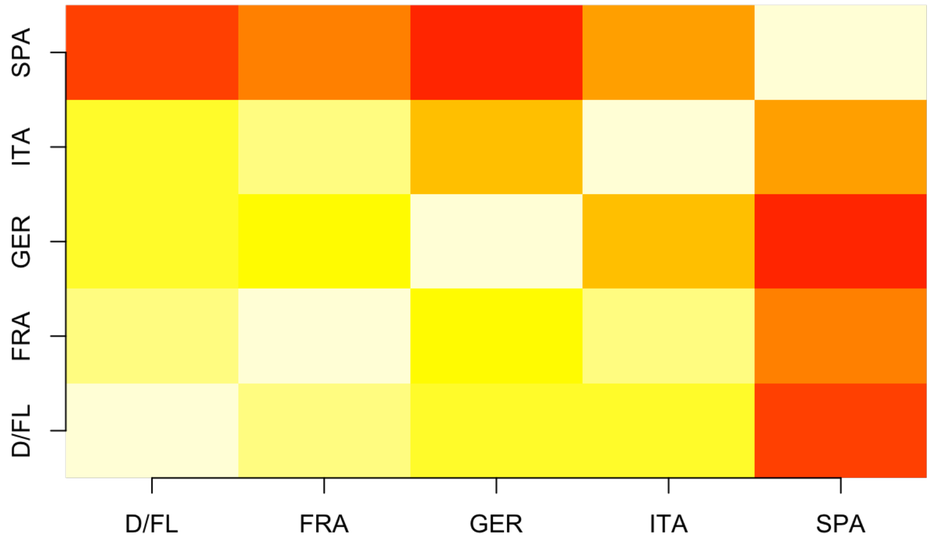
model log(price) and perform procedural and expert opinion based model selection

# sample: similarity of schools

Calculate a similarity score between different classes of art - score between 0 and 1, higher scores reflect a greater degree of similarity among features; i.e. a score of 1 would indicate identical vectors while a score of 0 would indicate vectors with no features in common.

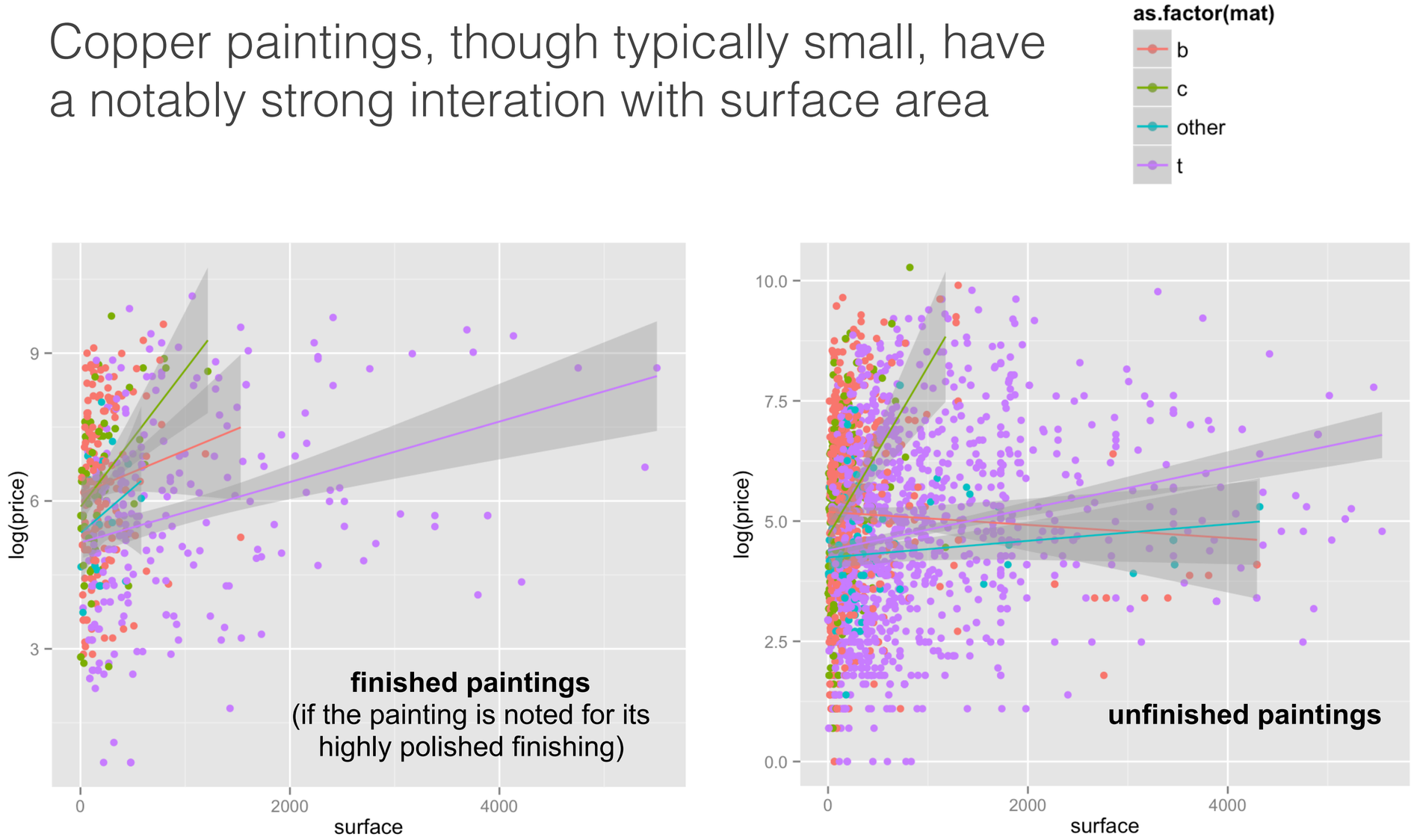
```
similarity = function (vec1, vec2) {
  mag1 = sqrt(vec1 %*% vec1)
  mag2 = sqrt(vec2 %*% vec2)
  return(vec1 %*% vec2 / mag1 / mag2)
}
```

Spanish art is most notably different from the other schools (Lighter colors indicate similarities, while deep red indicates large differences).



# sample: material and price

Copper paintings, though typically small, have a notably strong interation with surface area



(2) bbasketball

data

goduke.statsgeek.com/basketball-m/seasons/schedule.php?season=2014-15

## 2014-15 Schedule & Results

| Date | | [Rk] Opponent | Duke Rank | Location (Venue) | Score (OT) | Att. | Tip Time | TV |
|---|---|---|---|---|---|---|---|---|
| 11/14 | ~ | **Presbyterian** | 4 | Durham, N.C. (Cameron Indoor Stadium) | W 113-44 | 9,314 | 6 p.m. | ESPNU |
| 11/15 | ~ | **Fairfield** | 4 | Durham, N.C. (Cameron Indoor Stadium) | W 109-59 | 9,314 | 8 p.m. | ESPN3 |
| 11/18 | ¶¶ | vs. [19] Michigan State | 4 | Indianapolis, Ind. (Bankers Life Fieldhouse) | W 81-71 | 19,306 | 7 p.m. | ESPN |
| 11/21 | ~ | vs. Temple | 4 | Brooklyn, N.Y. (Barclays Center) | W 74-54 | 10,135 | 9:30 p.m. | TruTV |
| 11/22 | ~ | vs. Stanford | 4 | Brooklyn, N.Y. (Barclays Center) | W 70-59 | 10,046 | 9:30 p.m. | TruTV |
| 11/26 | | **Furman** | 4 | Durham, N.C. (Cameron Indoor Stadium) | W 93-54 | 9,314 | 5 p.m. | ESPNU |
| 11/30 | | **Army** | 4 | Durham, N.C. (Cameron Indoor Stadium) | W 93-73 | 9,314 | 12 p.m. | ESPNU |
| 12/3 | # | at [2] Wisconsin | 4 | Madison, Wisc. (Kohl Center) | W 80-70 | 17,279 | 9:30 p.m. | ESPN |
| 12/15 | | **Elon** | 2 | Durham, N.C. (Cameron Indoor Stadium) | W 75-62 | 9,314 | 7 p.m. | ESPNU |
| 12/18 | | vs. Connecticut | 2 | East Rutherford, N.J. (Izod Center) | W 66-56 | 16,541 | 8 p.m. | ESPN |
| 12/29 | | **Toledo** | 2 | Durham, N.C. (Cameron Indoor Stadium) | W 86-69 | 9,314 | 7 p.m. | ESPN2 |
| 12/31 | | **Wofford** | 2 | Durham, N.C. (Cameron Indoor Stadium) | W 84-55 | 9,314 | 3 p.m. | RSN |
| 1/3 | * | **Boston College** | 2 | Durham, N.C. (Cameron Indoor Stadium) | W 85-62 | 9,314 | 4 p.m. | RSN |
| 1/7 | * | at Wake Forest | 2 | Winston-Salem, N.C. (Joel Coliseum) | W 73-65 | 12,651 | 9 p.m. | ACCN |
| 1/11 | * | at N.C. State | 2 | Raleigh, N.C. (PNC Arena) | L 75-87 | 19,500 | 1:30 p.m. | CBS |
| 1/13 | * | **Miami** | 4 | Durham, N.C. (Cameron Indoor Stadium) | L 74-90 | 9,314 | 9 p.m. | ESPNU |
| 1/17 | * | at [6] Louisville | 4 | Louisville, Ky. (KFC Yum! Center) | W 63-52 | 22,791 | 12 p.m. | ESPN |
| 1/19 | * | **Pittsburgh** | 5 | Durham, N.C. (Cameron Indoor Stadium) | W 79-65 | 9,314 | 7 p.m. | ESPN |
| 1/25 | | at St. Johns | 5 | New York, N.Y. (Madison Square Garden) | W 77-68 | 19,812 | 2 p.m. | FOX |
| 1/28 | * | at [8] Notre Dame | 4 | Notre Dame, Ind. (Joyce Center) | L 73-77 | 9,149 | 7:30 p.m. | ESPN2 |
| 1/31 | * | at [2] Virginia | 4 | Charlottesville, Va. (John Paul Jones Arena) | W 69-63 | 14,593 | 7 p.m. | ESPN |
| 2/4 | * | **Georgia Tech** | 4 | Durham, N.C. (Cameron Indoor Stadium) | W 72-66 | 9,314 | 7 p.m. | ESPN2 |
| 2/7 | * | [10] **Notre Dame** | 4 | Durham, N.C. (Cameron Indoor Stadium) | W 90-60 | 9,314 | 1 p.m. | CBS |
| 2/9 | * | at Florida State | 4 | Tallahassee, Fla. (Donald L. Tucker Center) | W 73-70 | 11,498 | 7 p.m. | ESPN |
| 2/14 | * | at Syracuse | 4 | Syracuse, N.Y. (Carrier Dome) | W 80-72 | 35,446 | 6 p.m. | ESPN |
| 2/18 | * | [15] **North Carolina** | 4 | Durham, N.C. (Cameron Indoor Stadium) | W 92-90 • | 9,314 | 9 p.m. | ESPN/ACCN |
| 2/21 | * | **Clemson** | 4 | Durham, N.C. (Cameron Indoor Stadium) | W 78-56 | 9,314 | 4 p.m. | ESPN |
| 2/25 | * | at Virginia Tech | 4 | Blacksburg, Va. (Cassell Coliseum) | W 91-86 • | 9,847 | 9 p.m. | ESPN2 |
| 2/28 | * | **Syracuse** | 4 | Durham, N.C. (Cameron Indoor Stadium) | W 73-54 | 9,314 | 7 p.m. | ESPN |
| 3/4 | * | **Wake Forest** | 3 | Durham, N.C. (Cameron Indoor Stadium) | W 94-51 | 9,314 | 8 p.m. | ACCN |
| 3/7 | * | at [19] North Carolina | 3 | Chapel Hill, N.C. (Dean Dome) | W 84-77 | 21,750 | 9 p.m. | ESPN |
| 3/12 | $$$ | vs. N.C. State | 2 | Greensboro, N.C. (Greensboro Coliseum) | W 77-53 | 22,026 | 7 p.m. | ESPN |
| 3/13 | $$$$ | vs. [11] Notre Dame | 2 | Greensboro, N.C. (Greensboro Coliseum) | L 64-74 | 22,026 | 9 p.m. | ESPN |
| 3/20 | !! | vs. Robert Morris | 4 | Charlotte, N.C. (Time Warner Cable Arena) | W 85-56 | 16,945 | 7 p.m. | CBS |
| 3/22 | !!! | vs. San Diego State | 4 | Charlotte, N.C. (Time Warner Cable Arena) | W 68-49 | 18,482 | 2 p.m. | CBS |
| 3/27 | !!!! | vs. [19] Utah | 4 | Houston, Texas (NRG Stadium) | W 63-57 | 21,168 | 7:45 p.m. | CBS |
| 3/29 | !!!!! | vs. [7] Gonzaga | 4 | Houston, Texas (NRG Stadium) | W 66-52 | 20,744 | 4 p.m. | CBS |
| 4/4 | !!!!!! | vs. [23] Michigan State | 4 | Indianapolis, Ind. (Lucas Oil Stadium) | W 81-61 | 72,238 | 6 p.m. | TBS/TNT |
| 4/6 | !!!!!!! | vs. [3] Wisconsin | 4 | Indianapolis, Ind. (Lucas Oil Stadium) | W 68-63 | 71,149 | 9:15 p.m. | CBS |

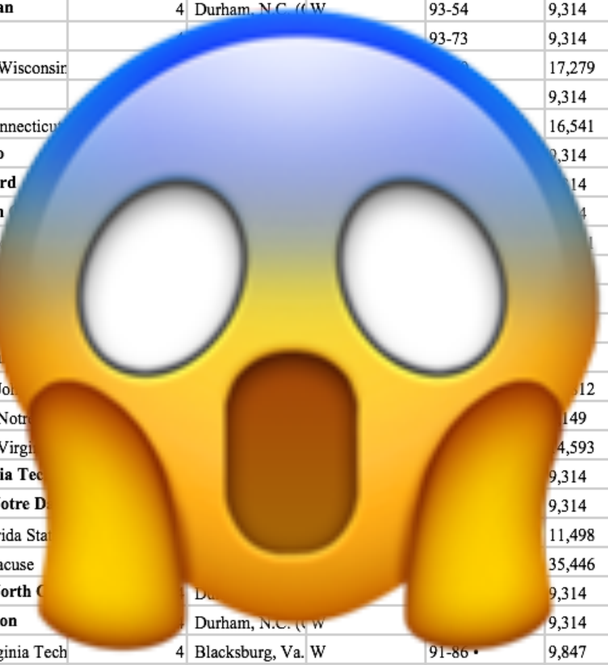
copy

# 2014-15 Schedule & Results

| Date | | [Rk] Opponent | Duke Rank | Location (Venue) | Score (OT) | Att. | Tip Time | TV |
|---|---|---|---|---|---|---|---|---|
| 11/14 | ~ | **Presbyterian** | 4 | Durham, N.C. (Cameron Indoor Stadium) | W 113-44 | 9,314 | 6 p.m. | ESPNU |
| 11/15 | ~ | **Fairfield** | 4 | Durham, N.C. (Cameron Indoor Stadium) | W 109-59 | 9,314 | 8 p.m. | ESPN3 |
| 11/18 | ¶¶ | vs. [19] Michigan State | 4 | Indianapolis, Ind. (Bankers Life Fieldhouse) | W 81-71 | 19,306 | 7 p.m. | ESPN |
| 11/21 | ~ | vs. Temple | 4 | Brooklyn, N.Y. (Barclays Center) | W 74-54 | 10,135 | 9:30 p.m. | TruTV |
| 11/22 | ~ | vs. Stanford | 4 | Brooklyn, N.Y. (Barclays Center) | W 70-59 | 10,046 | 9:30 p.m. | TruTV |
| 11/26 | | **Furman** | 4 | Durham, N.C. (Cameron Indoor Stadium) | W 93-54 | 9,314 | 5 p.m. | ESPNU |
| 11/30 | | **Army** | 4 | Durham, N.C. (Cameron Indoor Stadium) | W 93-73 | 9,314 | 12 p.m. | ESPNU |
| 12/3 | # | at [2] Wisconsin | 4 | Madison, Wisc. (Kohl Center) | W 80-70 | 17,279 | 9:30 p.m. | ESPN |
| 12/15 | | **Elon** | 2 | Durham, N.C. (Cameron Indoor Stadium) | W 75-62 | 9,314 | 7 p.m. | ESPNU |
| 12/18 | | vs. Connecticut | 2 | East Rutherford, N.J. (Izod Center) | W 66-56 | 16,541 | 8 p.m. | ESPN |
| 12/29 | | **Toledo** | 2 | Durham, N.C. (Cameron Indoor Stadium) | W 86-69 | 9,314 | 7 p.m. | ESPN2 |
| 12/31 | | **Wofford** | 2 | Durham, N.C. (Cameron Indoor Stadium) | W 84-55 | 9,314 | 3 p.m. | RSN |
| 1/3 | * | **Boston College** | 2 | Durham, N.C. (Cameron Indoor Stadium) | W 85-62 | 9,314 | 4 p.m. | RSN |
| 1/7 | * | at Wake Forest | 2 | Winston-Salem, N.C. (Joel Coliseum) | W 73-65 | 12,651 | 9 p.m. | ACCN |
| 1/11 | * | at N.C. State | 2 | Raleigh, N.C. (PNC Arena) | L 75-87 | 19,500 | 1:30 p.m. | CBS |
| 1/13 | * | **Miami** | 4 | Durham, N.C. (Cameron Indoor Stadium) | L 74-90 | 9,314 | 9 p.m. | ESPNU |
| 1/17 | * | at [6] Louisville | 4 | Louisville, Ky. (KFC Yum! Center) | W 63-52 | 22,791 | 12 p.m. | ESPN |
| 1/19 | * | **Pittsburgh** | 5 | Durham, N.C. (Cameron Indoor Stadium) | W 79-65 | 9,314 | 7 p.m. | ESPN |
| 1/25 | | at St. Johns | 5 | New York, N.Y. (Madison Square Garden) | W 77-68 | 19,812 | 2 p.m. | FOX |
| 1/28 | * | at [8] Notre Dame | 4 | Notre Dame, Ind. (Joyce Center) | L 73-77 | 9,149 | 7:30 p.m. | ESPN2 |
| 1/31 | * | at [2] Virginia | 4 | Charlottesville, Va. (John Paul Jones Arena) | W 69-63 | 14,593 | 7 p.m. | ESPN |
| 2/4 | * | **Georgia Tech** | 4 | Durham, N.C. (Cameron Indoor Stadium) | W 72-66 | 9,314 | 7 p.m. | ESPN2 |
| 2/7 | * | [10] **Notre Dame** | 4 | Durham, N.C. (Cameron Indoor Stadium) | W 90-60 | 9,314 | 1 p.m. | CBS |
| 2/9 | * | at Florida State | 4 | Tallahassee, Fla. (Donald L. Tucker Center) | W 73-70 | 11,498 | 7 p.m. | ESPN |
| 2/14 | * | at Syracuse | 4 | Syracuse, N.Y. (Carrier Dome) | W 80-72 | 35,446 | 6 p.m. | ESPN |
| 2/18 | * | [15] **North Carolina** | 4 | Durham, N.C. (Cameron Indoor Stadium) | W 92-90 • | 9,314 | 9 p.m. | ESPN/ACCN |
| 2/21 | * | **Clemson** | 4 | Durham, N.C. (Cameron Indoor Stadium) | W 78-56 | 9,314 | 4 p.m. | ESPN |
| 2/25 | * | at Virginia Tech | 4 | Blacksburg, Va. (Cassell Coliseum) | W 91-86 • | 9,847 | 9 p.m. | ESPN2 |
| 2/28 | * | **Syracuse** | 4 | Durham, N.C. (Cameron Indoor Stadium) | W 73-54 | 9,314 | 7 p.m. | ESPN |
| 3/4 | * | **Wake Forest** | 3 | Durham, N.C. (Cameron Indoor Stadium) | W 94-51 | 9,314 | 8 p.m. | ACCN |
| 3/7 | * | at [19] North Carolina | 3 | Chapel Hill, N.C. (Dean Dome) | W 84-77 | 21,750 | 9 p.m. | ESPN |
| 3/12 | $$$ | vs. N.C. State | 2 | Greensboro, N.C. (Greensboro Coliseum) | W 77-53 | 22,026 | 7 p.m. | ESPN |
| 3/13 | $$$$ | vs. [11] Notre Dame | 2 | Greensboro, N.C. (Greensboro Coliseum) | L 64-74 | 22,026 | 9 p.m. | ESPN |
| 3/20 | !! | vs. Robert Morris | 4 | Charlotte, N.C. (Time Warner Cable Arena) | W 85-56 | 16,945 | 7 p.m. | CBS |
| 3/22 | !!! | vs. San Diego State | 4 | Charlotte, N.C. (Time Warner Cable Arena) | W 68-49 | 18,482 | 2 p.m. | CBS |
| 3/27 | !!!! | vs. [19] Utah | 4 | Houston, Texas (NRG Stadium) | W 63-57 | 21,168 | 7:45 p.m. | CBS |
| 3/29 | !!!!! | vs. [7] Gonzaga | 4 | Houston, Texas (NRG Stadium) | W 66-52 | 20,744 | 4 p.m. | CBS |
| 4/4 | !!!!!! | vs. [23] Michigan State | 4 | Indianapolis, Ind. (Lucas Oil Stadium) | W 81-61 | 72,238 | 6 p.m. | TBS/TNT |
| 4/6 | !!!!!!! | vs. [3] Wisconsin | 4 | Indianapolis, Ind. (Lucas Oil Stadium) | W 68-63 | 71,149 | 9:15 p.m. | CBS |

paste

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | Date | | Old Opponent | Rank | Location (Attendance) | Score (OT) | Att. | Time | TV | |
| 4 | 14-Nov | ~ | **Presbyterian** | 4 | Durham, N.C. ( | W | 113-44 | 9,314 | 6 p.m. | ESPNU |
| 5 | 15-Nov | ~ | **Fairfield** | 4 | Durham, N.C. ( | W | 109-59 | 9,314 | 8 p.m. | ESPN3 |
| 6 | 18-Nov | ¶¶ | vs. [19] Michiga | 4 | Indianapolis, In | W | 81-71 | 19,306 | 7 p.m. | ESPN |
| 7 | 21-Nov | ~ | vs. Temple | 4 | Brooklyn, N.Y. | W | 74-54 | 10,135 | 9:30 p.m. | TruTV |
| 8 | 22-Nov | ~ | vs. Stanford | 4 | Brooklyn, N.Y. | W | 70-59 | 10,046 | 9:30 p.m. | TruTV |
| 9 | 26-Nov | | **Furman** | 4 | Durham, N.C. ( | W | 93-54 | 9,314 | 5 p.m. | ESPNU |
| 10 | 30-Nov | | **Army** | | | | 93-73 | 9,314 | 12 p.m. | ESPNU |
| 11 | 3-Dec | # | at [2] Wisconsin | | | | | 17,279 | 9:30 p.m. | ESPN |
| 12 | 15-Dec | | **Elon** | | | | | 9,314 | 7 p.m. | ESPNU |
| 13 | 18-Dec | | vs. Connecticu | | | | | 16,541 | 8 p.m. | ESPN |
| 14 | 29-Dec | | **Toledo** | | | | | ,314 | 7 p.m. | ESPN2 |
| 15 | 31-Dec | | **Wofford** | | | | | 14 | 3 p.m. | RSN |
| 16 | 3-Jan | * | **Boston** | | | | | 4 | 4 p.m. | RSN |
| 17 | 7-Jan | * | at Wal | | | | | 1 | 9 p.m. | ACCN |
| 18 | 11-Jan | * | at N.C | | | | | | 1:30 p.m. | CBS |
| 19 | 13-Jan | * | **Mian** | | | | | | 9 p.m. | ESPNU |
| 20 | 17-Jan | * | at [6] | | | | | | 12 p.m. | ESPN |
| 21 | 19-Jan | * | **Pittsbu** | | | | | | 7 p.m. | ESPN |
| 22 | 25-Jan | | at St. Jo | | | | | 12 | 2 p.m. | FOX |
| 23 | 28-Jan | * | at [8] Notr | | | | | 149 | 7:30 p.m. | ESPN2 |
| 24 | 31-Jan | * | at [2] Virgi | | | | | 4,593 | 7 p.m. | ESPN |
| 25 | 4-Feb | * | **Georgia Tec** | | | | | 9,314 | 7 p.m. | ESPN2 |
| 26 | 7-Feb | * | [10] **Notre D** | | | | | 9,314 | 1 p.m. | CBS |
| 27 | 9-Feb | * | at Florida Sta | | | | | 11,498 | 7 p.m. | ESPN |
| 28 | 14-Feb | * | at Syracuse | | | | | 35,446 | 6 p.m. | ESPN |
| 29 | 18-Feb | * | [15] **North C** | | D | | | 9,314 | 9 p.m. | ESPN/ACCN |
| 30 | 21-Feb | * | **Clemson** | | Durham, N.C. ( | W | | 9,314 | 4 p.m. | ESPN |
| 31 | 25-Feb | * | at Virginia Tech | 4 | Blacksburg, Va. | W | 91-86 • | 9,847 | 9 p.m. | ESPN2 |
| 32 | 28-Feb | * | **Syracuse** | 4 | Durham, N.C. ( | W | 73-54 | 9,314 | 7 p.m. | ESPN |
| 33 | 4-Mar | * | **Wake Forest** | 3 | Durham, N.C. ( | W | 94-51 | 9,314 | 8 p.m. | ACCN |
| 34 | 7-Mar | * | at [19] North Ca | 3 | Chapel Hill, N.C | W | 84-77 | 21,750 | 9 p.m. | ESPN |
| 35 | 12-Mar | $$$ | vs. N.C. State | 2 | Greensboro, N.C | W | 77-53 | 22,026 | 7 p.m. | ESPN |
| 36 | 13-Mar | $$$$ | vs. [11] Notre D | 2 | Greensboro, N.C | L | 64-74 | 22,026 | 9 p.m. | ESPN |
| 37 | 20-Mar | !! | vs. Robert Morr | 4 | Charlotte, N.C. ( | W | 85-56 | 16,945 | 7 p.m. | CBS |
| 38 | 22-Mar | !!! | vs. San Diego S | 4 | Charlotte, N.C. ( | W | 68-49 | 18,482 | 2 p.m. | CBS |
| 39 | 27-Mar | !!!! | vs. [19] Utah | 4 | Houston, Texas | W | 63-57 | 21,168 | 7:45 p.m. | CBS |
| 40 | 29-Mar | !!!!! | vs. [7] Gonzaga | 4 | Houston, Texas | W | 66-52 | 20,744 | 4 p.m. | CBS |
| 41 | 4-Apr | !!!!!! | vs. [23] Michiga | 4 | Indianapolis, In | W | 81-61 | 72,238 | 6 p.m. | TBS/TNT |
| 42 | 6-Apr | !!!!!!! | vs. [3] Wisconsi | 4 | Indianapolis, In | W | 68-63 | 71,149 | 9:15 p.m. | CBS |



scrape

```
# Load packages ----------------------------------------------------
library(rvest)
library(stringr)
library(dplyr)

# Read page with season data ---------------------------------------
page <- read_html("http://goduke.statsgeek.com/basketball-m/seasons/schedule.php?season=2014-15")

# Harvest fields ---------------------------------------------------
date <- page %>%
  html_nodes(".stattextline b") %>%
  html_text()

opponent <- page %>%
  html_nodes(".stattextltgray2:nth-child(3)") %>%
  html_text() %>%
  str_trim()

venue <- page %>%
  html_nodes(".stattextltgray2:nth-child(5)") %>%
  html_text() %>%
  str_trim()

# Put fields into a tibble -----------------------------------------
blue_devils_1415 <- data_frame(date, opponent, venue)
```

voila!

blue_devils_1415

Filter

| | date | opponent | venue |
|---|---|---|---|
| 1 | 11/14 | Presbyterian | Durham, N.C. (Cameron Indoor Stadium) |
| 2 | 11/15 | Fairfield | Durham, N.C. (Cameron Indoor Stadium) |
| 3 | 11/18 | vs. [19] Michigan State | Indianapolis, Ind. (Bankers Life Fieldhouse) |
| 4 | 11/21 | vs. Temple | Brooklyn, N.Y. (Barclays Center) |
| 5 | 11/22 | vs. Stanford | Brooklyn, N.Y. (Barclays Center) |
| 6 | 11/26 | Furman | Durham, N.C. (Cameron Indoor Stadium) |
| 7 | 11/30 | Army | Durham, N.C. (Cameron Indoor Stadium) |
| 8 | 12/3 | at [2] Wisconsin | Madison, Wisc. (Kohl Center) |
| 9 | 12/15 | Elon | Durham, N.C. (Cameron Indoor Stadium) |
| 10 | 12/18 | vs. Connecticut | East Rutherford, N.J. (Izod Center) |
| 11 | 12/29 | Toledo | Durham, N.C. (Cameron Indoor Stadium) |
| 12 | 12/31 | Wofford | Durham, N.C. (Cameron Indoor Stadium) |

Showing 1 to 13 of 39 entries

# (3) interactivity

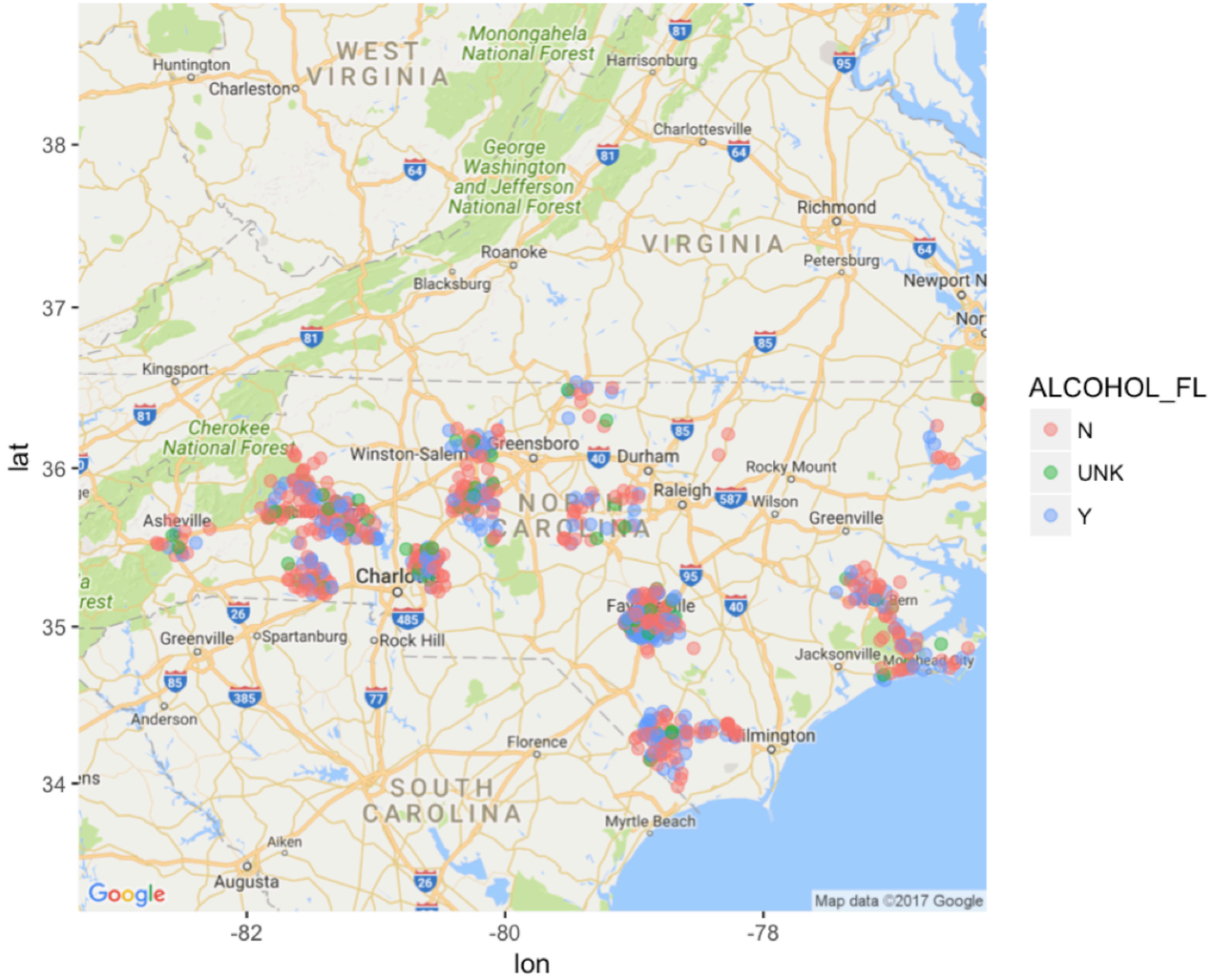
```
> library(shiny)
```

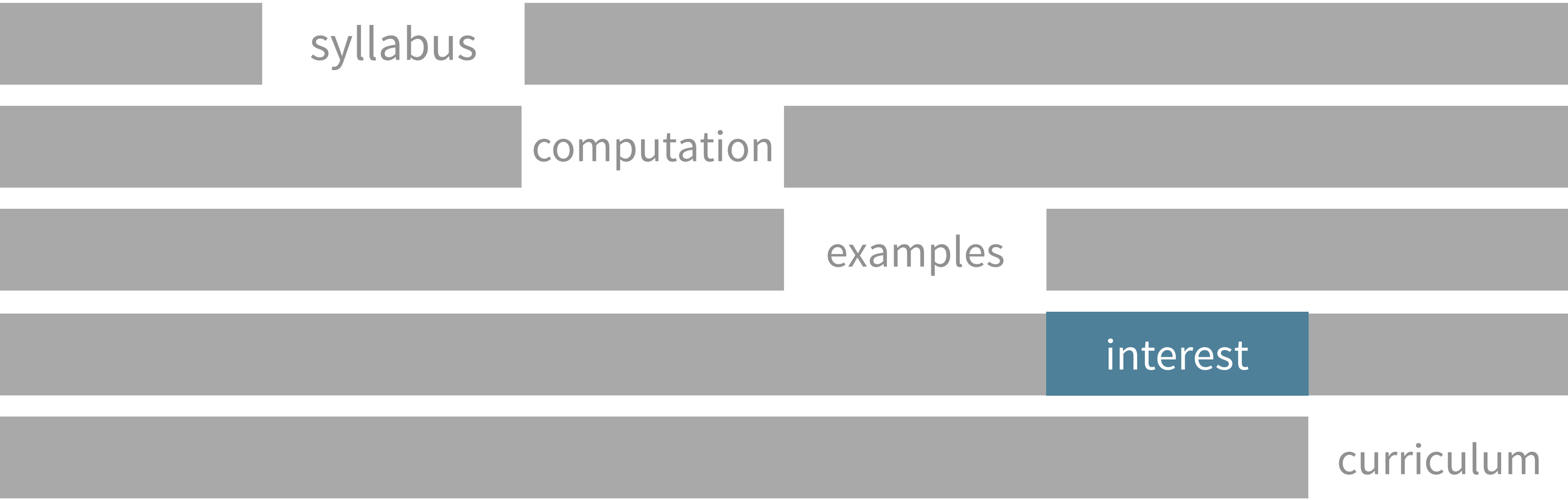
# Modeling the Distributions of Fatal Car Crashes

**Select categorical variable**

alcohol involved

- crash type
- alcohol involved
- teen driver
- older driver
- direction




syllabus
computation
examples
interest
curriculum

# interest

**duke focus:**

first-year undergrads modeling cluster: “What if? Explaining the Past, Predicting the Future”

**interest in What If:**

no hard data, but “definitely significant increase in applications the last two years than

**interest in DS:**

% of What If applicants interested in DS

2015: 76%

2016: 83%

impact

**pipeline for stats:**

2014: 19% declared

2015: 31% declared

2016: ~40%

expressed interest

**diversity:**

% female

2014: 44%

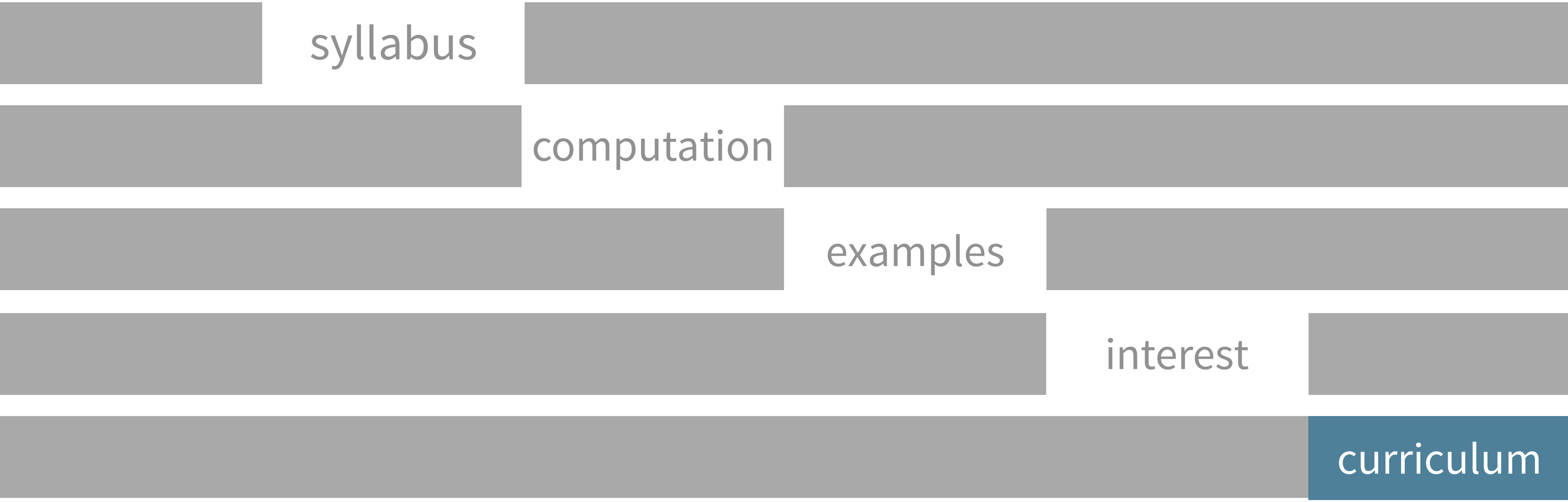
2015: 50%

2016: 35%

~25% in Probability

**curricular:**

basis for
gateway to stats
major course
to be offered in
Spring 2018!


syllabus
computation
examples
interest
curriculum

# curricular considerations

move away from ad-hoc computing education and/or expecting students to pick it up along the way

uniformity of tools is important: choose a toolkit that works for you and stick to it throughout the curriculum

teach computing early (without any prereqs) and often!

# a first-year undergraduate data science course

Slides at http://bit.ly/first-ds


@minebocek

mine-cetinkaya-rundel

mine@stat.duke.edu

mine çetinkaya-rundel

Duke UNIVERSITY

RStudio

R Ladies