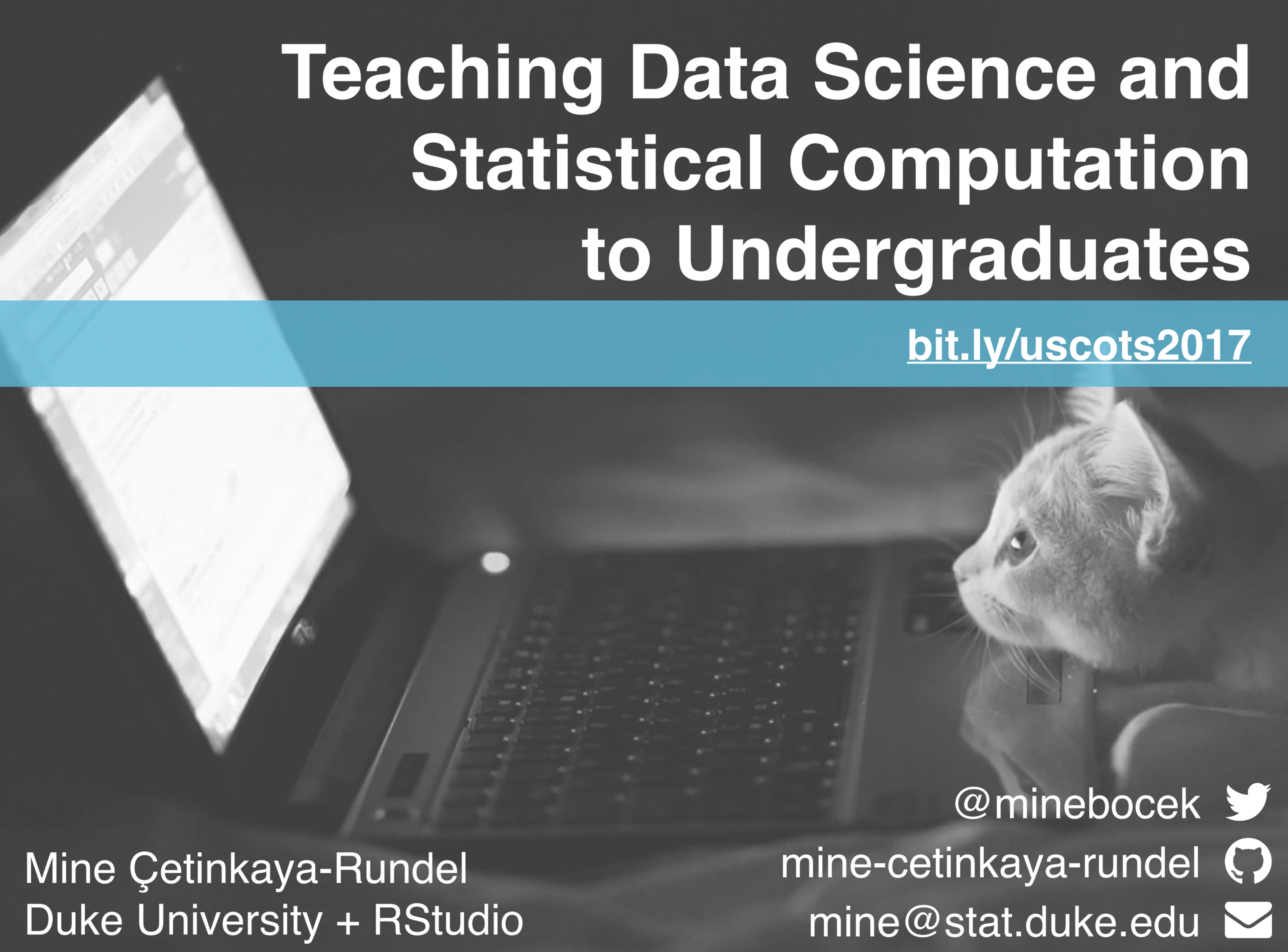
# Teaching Data Science and Statistical Computation to Undergraduates

bit.ly/uscots2017




Mine Çetinkaya-Rundel
Duke University + RStudio

@minebocek
mine-cetinkaya-rundel
mine@stat.duke.edu


I'm interested in stats, but I have no idea where to get started!

I'm interested in stats, but I have no idea where to get started!

Sta 101

I'm interested in stats, but I have no idea where to get started!

Calculus

I'm interested in stats, but I have no idea where to get started!

Probability

I'm interested in stats, but I have no idea where to get started!

Regression


I'm interested in stats, but I have no idea where to get started!
Intro to data science and statistics

motivation

computation

interest & impact

course overview

data analysis examples

curricular considera-tions

motivation

computation

interest & impact

course overview

data analysis examples

curricular considera-tions

**goal:**

a course that provides a common (gateway) experience to students wanting to get started with stats, and that is

1. modern
2. places data front and center
3. quantitative (but not mathematical)
4. different than any HS stats course
5. challenging (but not intimidating)

# 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

motivation

computation

interest & impact

course overview

data analysis examples

curricular considera-tions

*this course doesn't yet exist, but…*

# Better Living Through Data Science:

## Exploring / Modeling / Predicting / Understanding

Combines techniques from statistics, math, computer science, and social sciences, to learn how to use data to understand natural phenomena, explore patterns, model outcomes, and make predictions. Case studies include examples from election forecasts, movie reviews, and online dating match algorithms. Discussions around reproducibility, data sharing, data privacy will accompany these case studies. Gain experience in data wrangling and munging, exploratory data analysis, predictive modeling, and data visualization, and effective communication of results. Course will focus on R statistical computing language. No computing background necessary. For students in the FOCUS Program.

Part of the What If? Explaining the Past/Predicting the Future cluster.

first-year seminar for undergrads interested in quantitative fields

# course overview

## curriculum:

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

## structure:

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

## applications:

movie reviews, airline delays, paris paintings, basketball, professor evals, etc.

## assessment:

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

motivation

computation

interest & impact

course overview

data analysis examples

curricular considera-tions

# computation

**core:**

R +
RStudio Server

**toolkit:**

(mostly) `tidyverse`

**reproducibility:**

R Markdown +
Git/GitHub

# R + RStudio Server

**goal:**

get started "like a knife through butter" to minimize time to first data visualization

**how:**

avoid local installation with RStudio Server (Pro)

**at the end:**

provide instructions for + help with local install

# R Markdown

**reproducibility:**

train new analysts whose only workflow is a reproducible one

**efficiency:**

consistent formatting + built in “show your work”
= easier grading

**pedagogy:**

code + output + prose together

syntax highlighting + notebooks FTW!

**key to success:**

iterative development: knit early, and often

# Git + GitHub

**version control:**

lots of mistakes along the way, need ability keep track of history (revert)

**accountability:**

transparent commit history

**collaboration:**

platform and interface designed to enable collaboration

**early intro:**

mastery
takes time,
start early (day one)

marketability + discoverability

motivation

computation

interest & impact

course overview

data analysis examples

curricular considera-tions

# data expeditions

pairs of grad students, work with course instructor to formulate a question, and a pathway through a dataset to explore the question

element of an undergraduate course that introduces students to exploratory data analysis

graduate student participants receive a travel grant

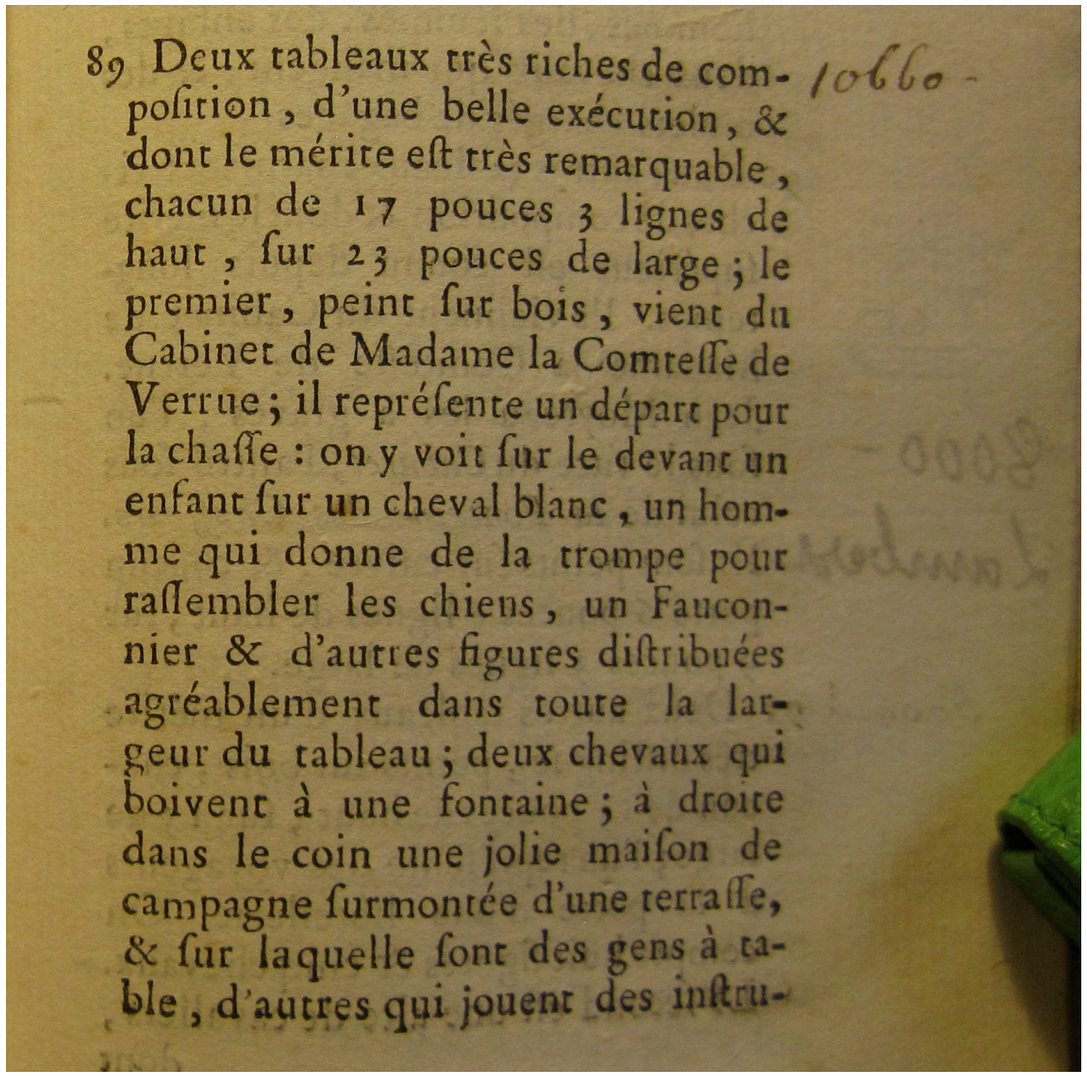
Source: http://bigdata.duke.edu/data-expeditions

# meet the experts

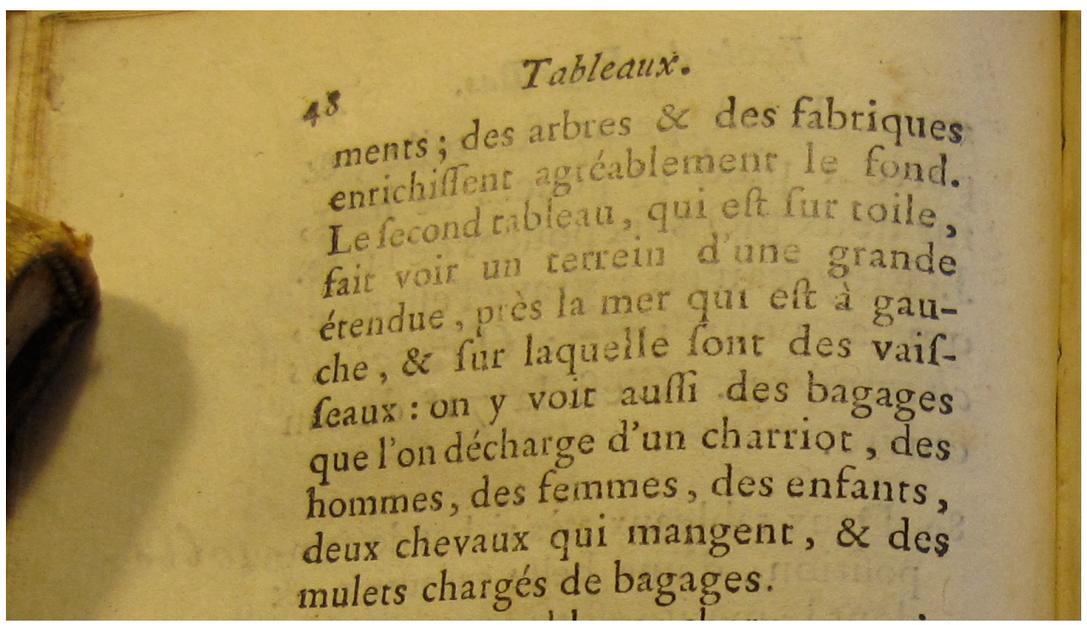

Sandra Van Ginhoven
PhD, Art History

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 enfant ſur un cheval blanc, un homme qui donne de la trompe pour raſſembler les chiens, un Fauconnier & d'autres figures diſtribuées agréablement dans toute la largeur du tableau; deux chevaux qui boivent à une fontaine; à droite dans le coin une jolie maiſon de campagne ſurmontée d'une terraſſe, & ſur laquelle ſont des gens à table, d'autres qui jouent des inſtru-



*Tableaux.*

48

ments; des arbres & des fabriques enrichiſſent agréablement le fond. 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_Rnd | 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 (mostly) with `ggplot2`

**clean:**

data cleaning (mostly) with `dplyr`

**model:**

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

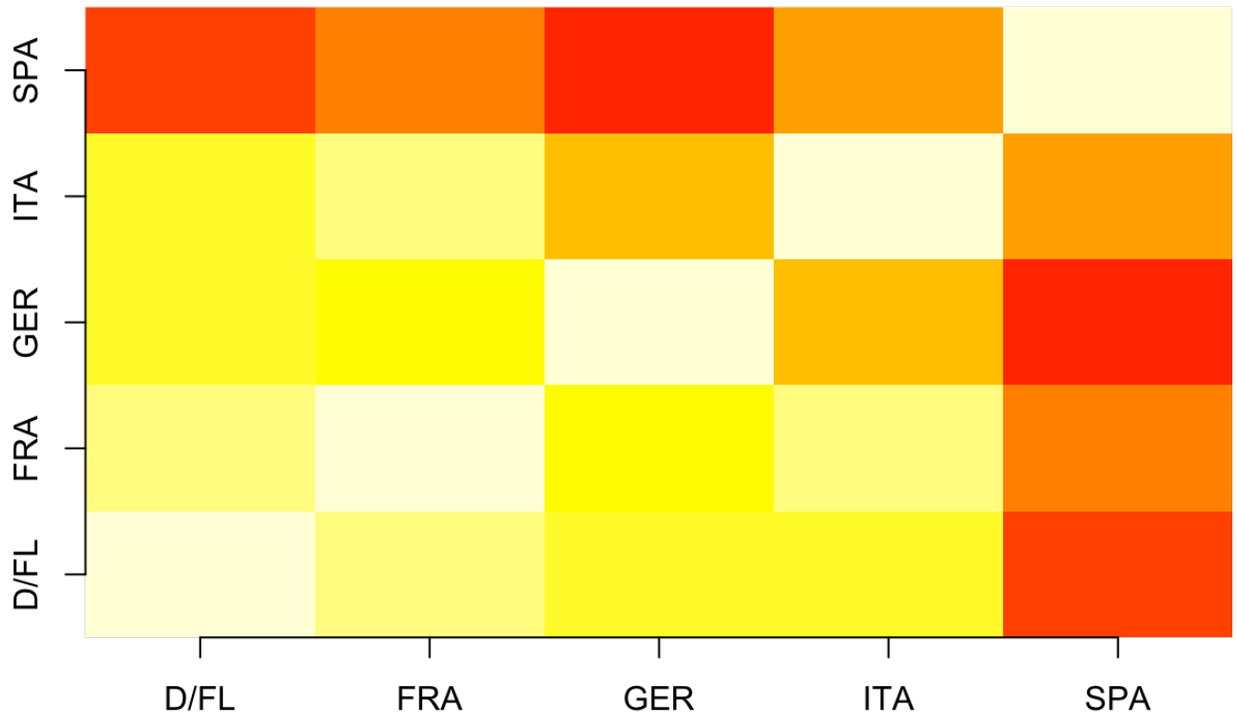
# sample exploration #1

## 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.

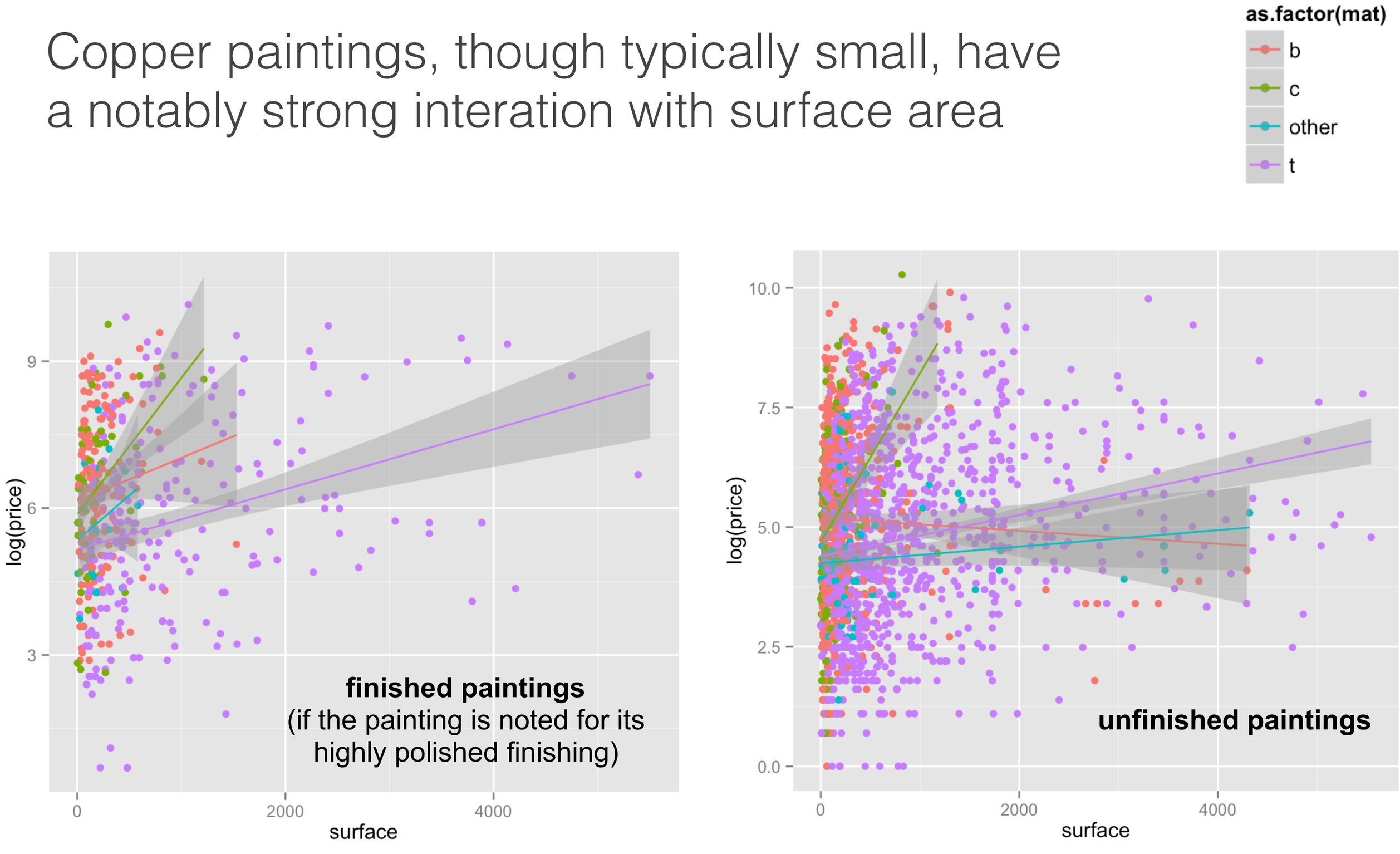
```
similarity = function (vec1, vec2) {
  mag1 = sqrt(vec1 %*% vec1)
  mag2 = sqrt(vec2 %*% vec2)
  return(vec1 %*% vec2 / mag1 / mag2)
}
```

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



# material and price

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



# experience

non-standard application piqued student interest

“massive” data overwhelming but expert input refreshing

unfamiliar variables made narrative challenging

novel application pushed creativity

# basketball

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 |

# basketball

**gather:**

scrape data with `rvest`

**clean:**

clean the data with (mostly) `dplyr`

**visualize:**

visualize the data with `ggplot2` and `shiny`

**Mine CetinkayaRundel**
@minebocek

Students upset b/c website they need to scrape data from for hw assignment is down. Bad assignment or good lesson in working w/ real data?

RETWEET
1

LIKES
10

9:48 AM - 26 Nov 2015

motivation

computation

interest & impact

course overview

data analysis examples

curricular considera-tions

# 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 previous years”

## 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!

motivation

computation

interest & impact

course overview

data analysis examples

curricular considera-tions

# 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 and often!


@minebocek
mine-cetinkaya-rundel
mine@stat.duke.edu

bit.ly/uscots2017