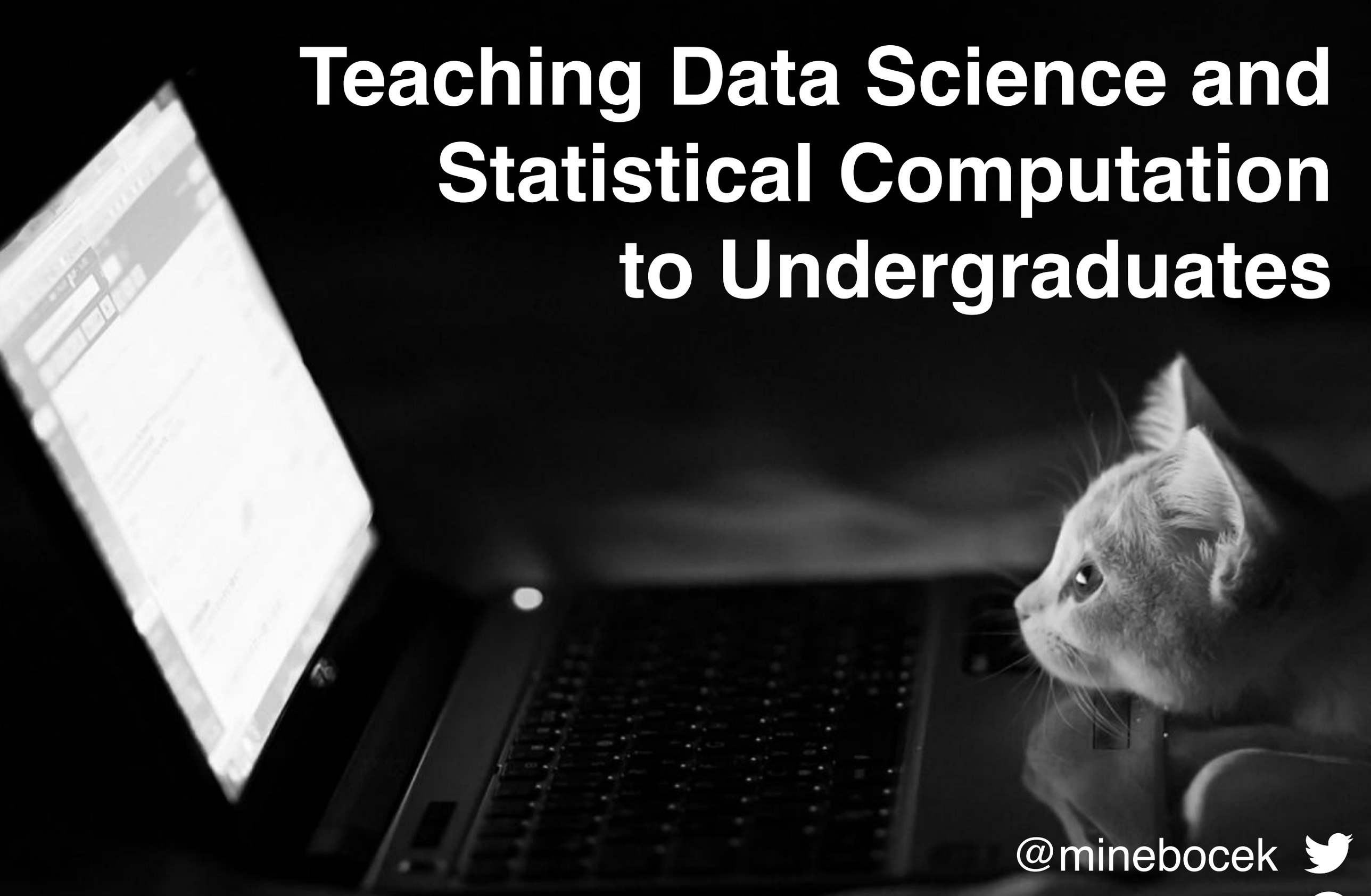
Teaching Data Science and Statistical Computation to Undergraduates



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 (think, code, write, present) collaboratively

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 progress, peer evaluations and contribution diagnostics

motivation

computation

interest & impact

course overview

data analysis examples

curricular considera-tions

computation

core:

R + RStudio

toolkit:

(mostly) `tidyverse`

reproducibility:

R Markdown + Git/GitHub

R + RStudio

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 → easier grading

pedagogy:

code + output + prose together

syntax highlighting FTW!

key to success:

knit early, and often

Git + GitHub :: why?

version control:

lots of mistakes along the way, need ability to revert

accountability:

transparent commit history

collaboration:

platform that removes barriers to well documented collaboration

early intro:

mastery takes time, earlier start the better

marketability

Git + GitHub :: how?

organization:

one organization per course

one repo per student/team per assignment

teams:

for collaboration

for assigning individual students to repos

for graders

interface:

via RStudio

no local git install required since using RStudio Server

assessment:

check reproducibility via clone + compile

feedback through issues

Git + GitHub :: day one

Working with GitHub

- Create a GitHub account at https://github.com/
    - This will be a public account associated with your name
    - Choose a username wisely for future use
    - Don't worry about details, you can fill them in later
- Create a repository called `intro_demo`
    - Give a brief and informative description
    - Choose "Public"
    - Check the box for "Initialize this repository with a README"
    - Click "Create Repository"

Cloning the repository

- Go to RStudio
- File -> New Project
    - Version Control: Checkout a project from a version control repository
    - Git: Clone a project from a repository
    - Fill in the info:
        - URL: use HTTPS address
        - Create as a subdirectory of: Browse and create a new folder call sta112
- Note for the future: Each course component you work on (an application exercise, a homework assignment, project, exam, etc.) should be its own repository, and should be fully contained in a folder inside the folder `sta112`.

Merge conflicts

- On GitHub (on the web) edit the README document and `Commit` it with a message describing what you did.
- Then, in RStudio also edit the README document with a different change.
    - Commit your changes
    - Try to push – you'll get an error!
    - Try pulling
    - Resolve the merge conflict and then commit and push
- As you work in teams you will run into merge conflicts, learning how to resolve them properly will be very important.

Git + GitHub :: lessons learned

if you plan on using git in class, start on day one, don't wait until the "right time"

first assignment should be individual, not team based to avoid merge conflicts

remind students on that future projects should go on GitHub with instructor / PI approval

students need to remember to pull before starting work

impossible (?) to avoid shell intervention every once in a while

motivation

computation

interest & impact

course overview

data analysis examples

curricular considera-tions

sample exercises

scraping data off the web + interactive visualization

scrape	clean	visualize
scrape data with **`rvest`** from goduke.statsgeek.com	clean the data with (mostly) **`dplyr`**	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

modeling paris paintings

data

auctions 1764 - 1780
[3,393 x 57]

seller / buyer, painter,

clean

clean the data with (mostly) `dplyr`

model

model log(price) and do model selection

	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	author standard	artistliving	author style	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	Lebrur
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
2521	R1777-89b	R1777	89	R	1777	D/FL	D/FL	D/FL	0	5,300.0	1	Déchargement d'un chariot, estran	Wouwerman, Philips	0	n/a	Philippe Wouwerman	Langlie

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

to do:
update 2015 data

pipeline for stats:

2014: 19% declared
2015: 38% expressed interest

diversity:

% female
2014: 44%
2015: 50%

compared to ~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

???

this section still needs work

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