a first-year undergraduate data science course

@minebocek

mine@stat.duke.edu

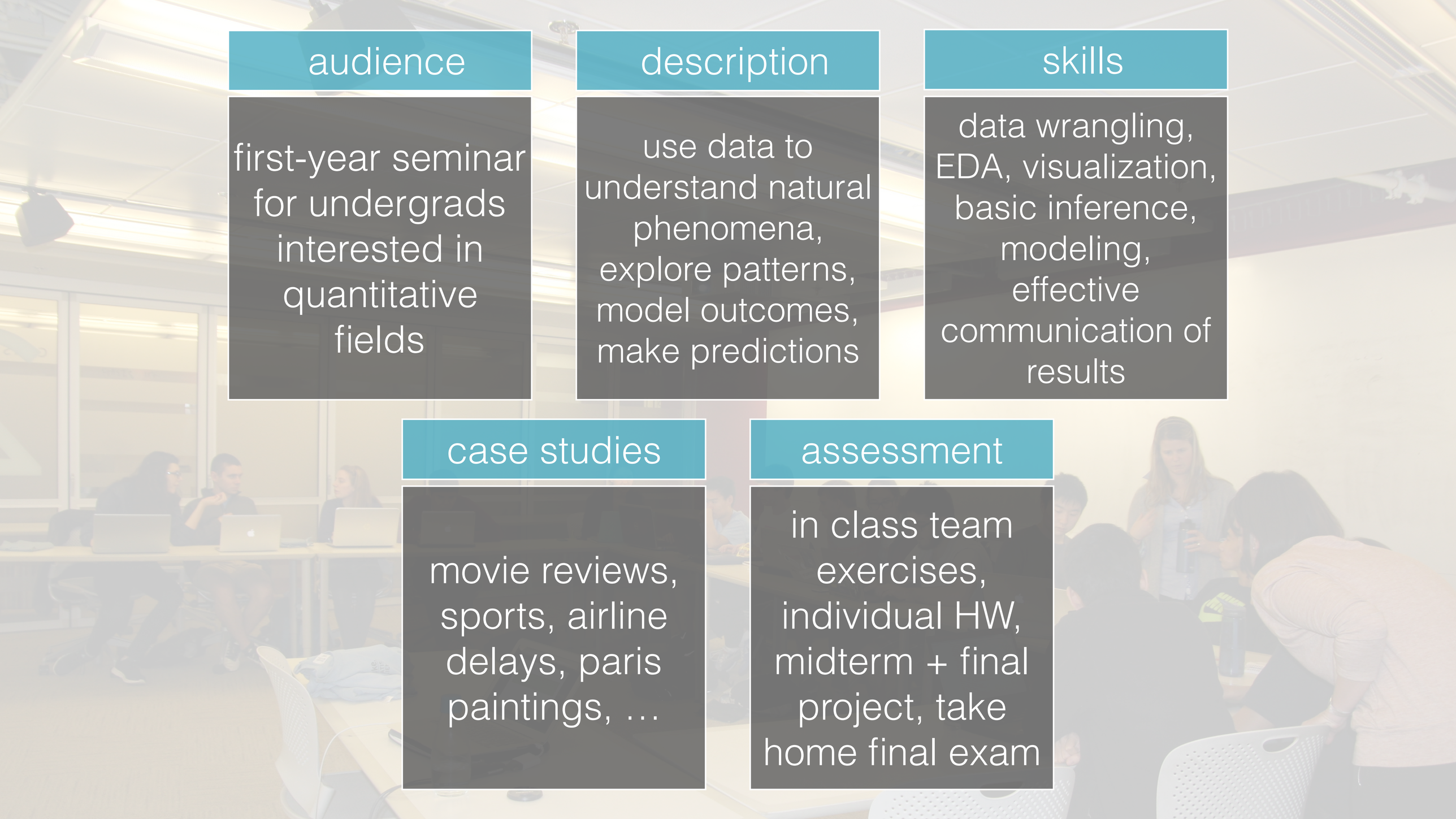
mine-cetinkaya-rundel

mine çetinkaya-rundel

duke university

course info

sta 112fs: better living with data science



audience

first-year seminar for undergrads interested in quantitative fields

description

use data to understand natural phenomena, explore patterns, model outcomes, make predictions

skills

data wrangling, EDA, visualization, basic inference, modeling, effective communication of results

case studies

movie reviews, sports, airline delays, paris paintings, …

assessment

in class team exercises, individual HW, midterm + final project, take home final exam

computation

R + RStudio + R Markdown + git + GitHub

R + RStudio

goal	why	how	at the end
get started “like a knife through butter”	avoid local installation to minimize time to first data visualization	RStudio Server Pro individual accounts on dept. server	provide instructions on / help with local install

R Markdown

reproducibility

train new analysts whose only workflow is a reproducible one

pedagogy

code + output + prose together

syntax highlighting FTW!

efficiency

consistent formatting → easier grading

key to success

knit early and often

Sta112FS-Fall2015

git + GitHub

why

version control	collaboration	accountability	early introduction
lots of mistakes along the way, need ability to revert	platform that removes barriers to well documented collaboration	transparent commit history	mastery takes time, earlier start the better marketablity

git + GitHub

how

Sta112FS-Fall2015

organization	interface	teams	assessment
one organization per course one repo per student/team per assignment	via RStudio no local git install required since using RStudio Server	for collaboration for assigning individual students to repos for graders	check reproducibility via clone + compile feedback through issues

Working with GitHub

git + GitHub

day one

- 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 PI approval

students need to remember to pull before starting work

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

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]

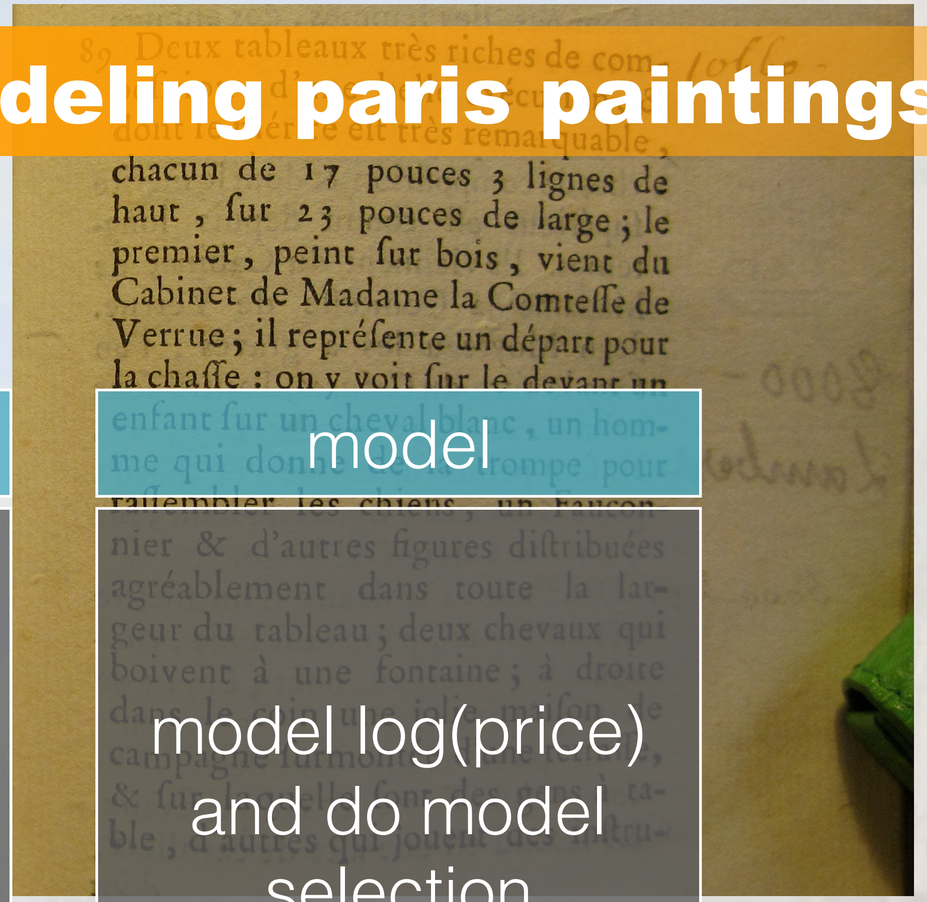
seller / buyer, painter, painting classification, painting attributes

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	authorstandard	artistliving	authorstyle	author	winnir
2517	R17'7-86	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

interest & impact

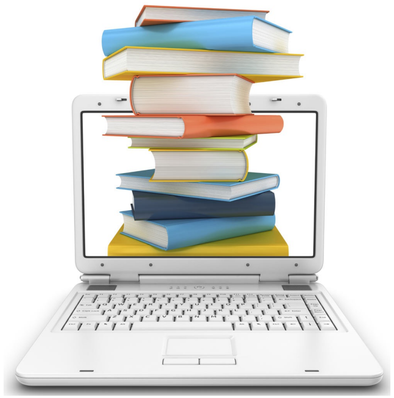
interest

duke focus	interest in What if?	interest in DS
first -year undergrads modeling theme cluster: “What if? Explaining the Past, Predicting the Future”	no hard data available, but “definitely significant increase in applications the last two years than previous years”	% of What If applicants interested in DS 2015: 76% 2016: 83%

impact

pipeline to stats	curriculum	gender distribution
% to StatSci 2014: 19% declared 2015: 38% expressed interest	plan to offer as Statistical Science gateway course starting in 2017 / 2018	% female 2014: 44% 2015: 50% compared to ~25% in current gateway (Probability)

thank you!



bit.ly/user_data_sci

@minebocek

mine@stat.duke.edu

mine-cetinkaya-rundel