

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

goal

get started
“like a knife
through butter”

why

avoid local
installation to
minimize time to
first data
visualization

how

RStudio Server
Pro

individual
accounts on
dept. server

at the end

provide
instructions on /
help with local
install

The screenshot shows the RStudio interface with an R Markdown file open. The title bar reads "rmarkdown.Rmd". The left pane displays the R Markdown code, and the right pane shows the generated HTML output.

R Markdown

Console

Files Plots Packages Help Viewer

RMarkdown

Mine Cetinkaya-Rundel

June 27, 2016

reproducibility

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

train new analysts whose only workflow is a reproducible one

pedagogy

code + output + prose together

syntax highlighting FTW!

efficiency

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both the source code and the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
## cars
```

key to success

consistent formatting → easier grading

#	speed	dist
## Min.	: 4.0	Min. : 12.00
## 1st Qu.	: 12.0	1st Qu.: 26.00
## Median	: 15.0	Median : 36.00
## Mean	: 15.4	Mean : 42.98
## 3rd Qu.	: 19.0	3rd Qu.: 56.00
## Max.	: 25.0	Max. : 120.00

Including Plots

You can also embed plots, for example:

2:1 R Markdown R Markdown



Sta112FS-Fall2015

git + GitHub

why

version control

lots of mistakes
along the way,
need ability to
revert

collaboration

platform that
removes barriers
to well
documented
collaboration

accountability

transparent
commit history

early introduction

mastery
takes time,
earlier start
the better

marketability

opEx_DasCrew PRIVATE

updated on Nov 17, 2015

opEx_TheStatian PRIVATE

HTML ★ 0 ⚡ 0

R ★ 0 ⚡ 0



Sta112FS-Fall2015

git + GitHub

how

organization

one organization
per course

one repo per
student/team per
assignment

interface

via RStudio
no local git install
required since
using RStudio
Server

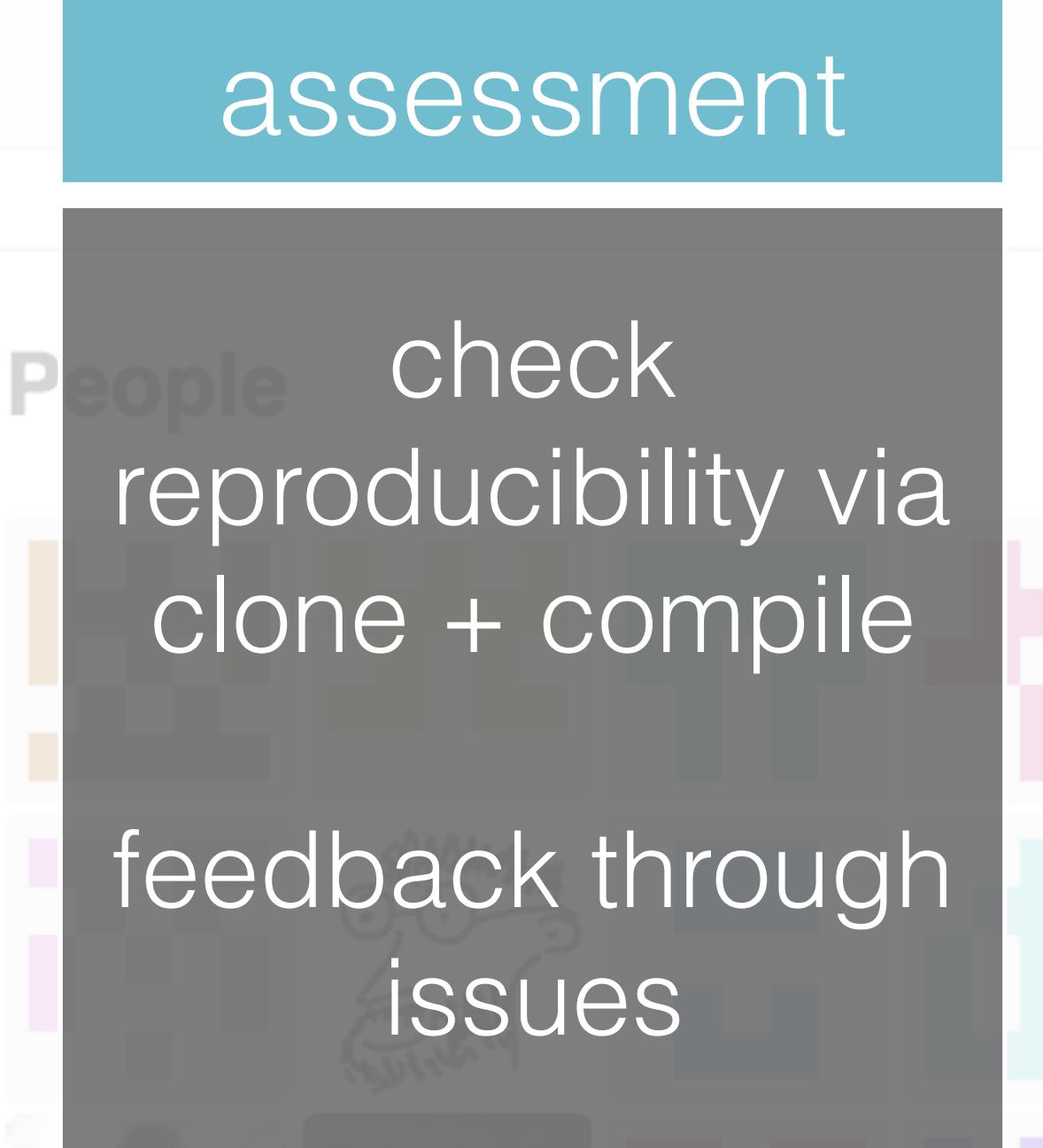
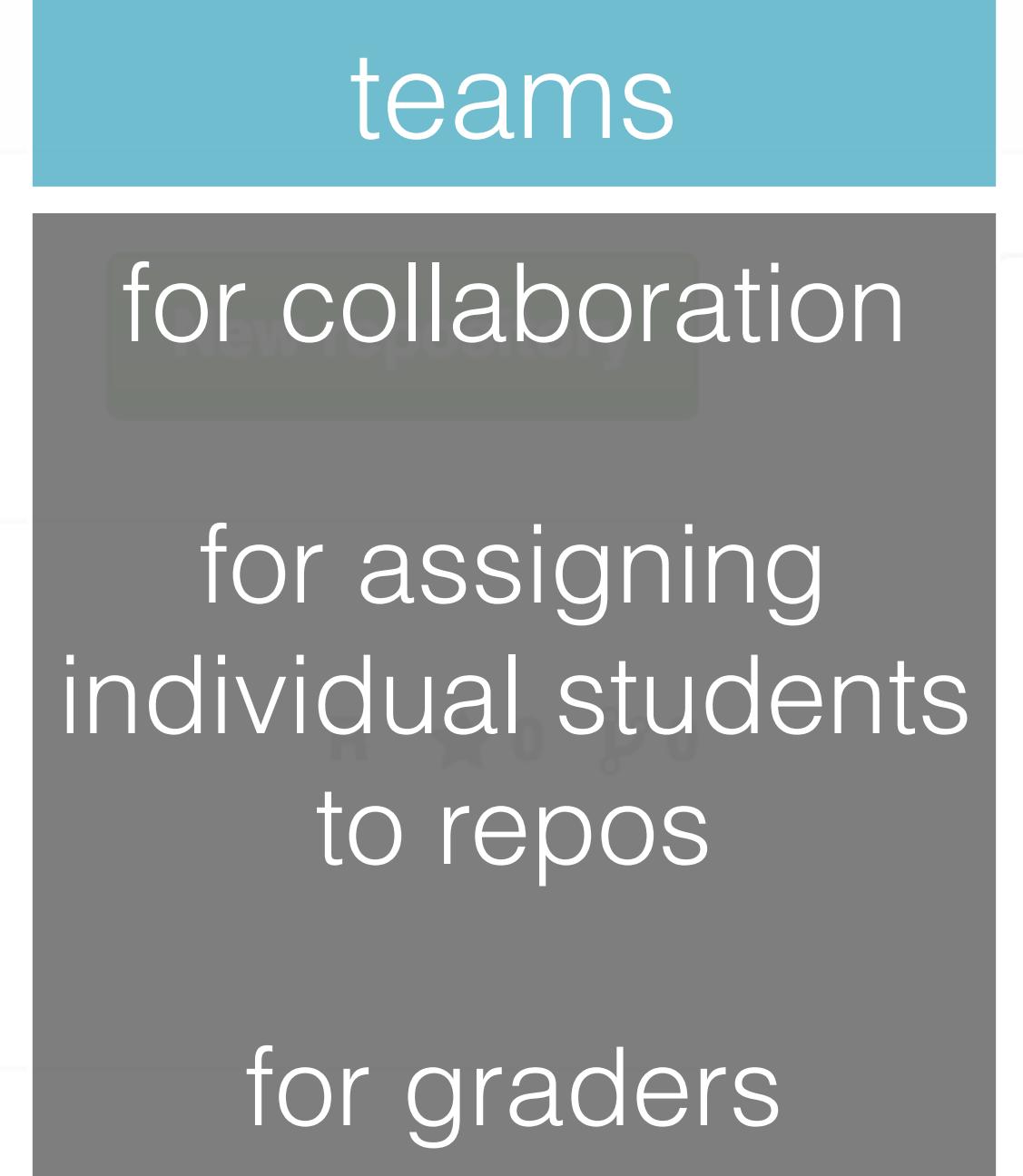
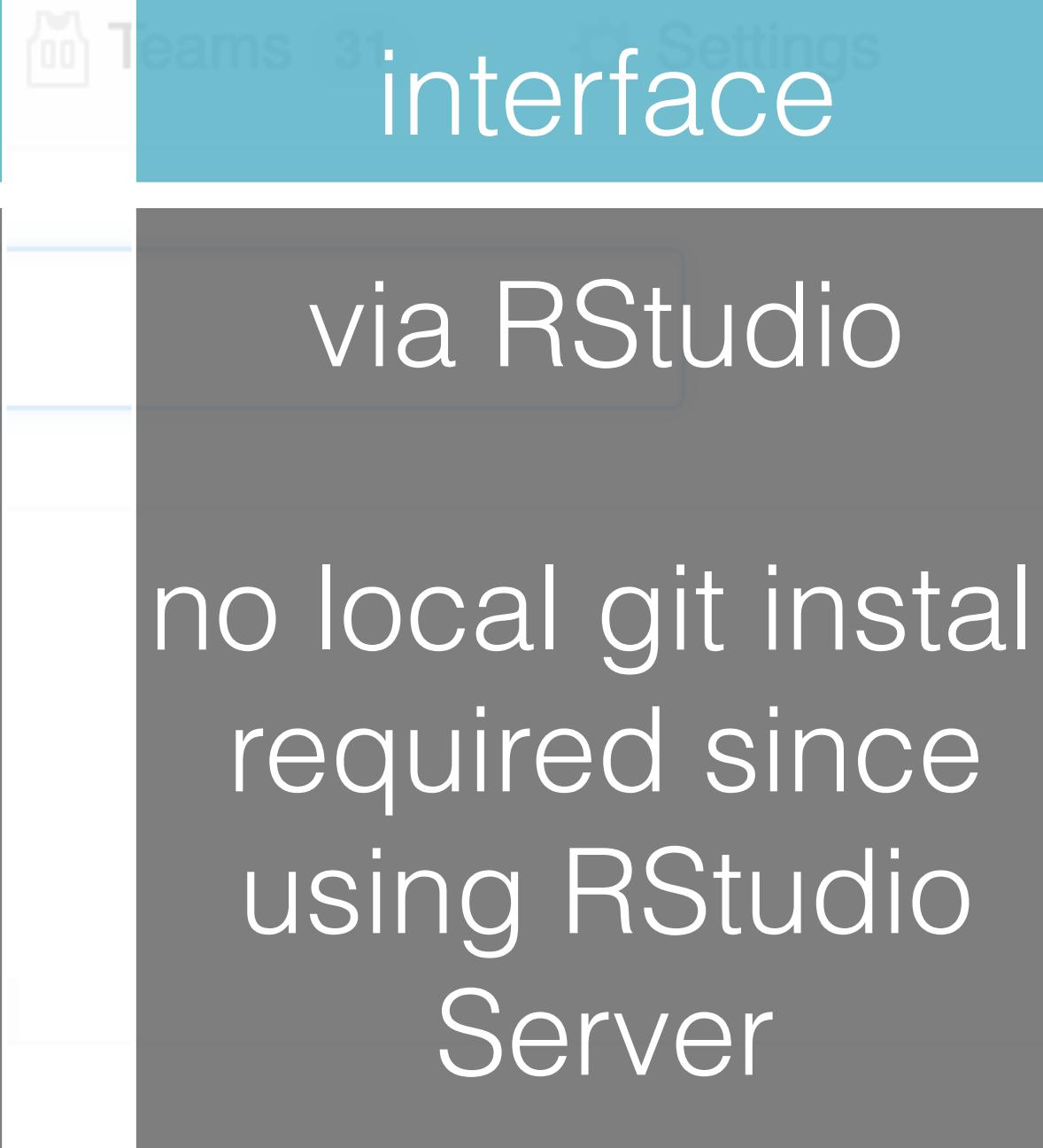
teams

for collaboration
for assigning
individual students
to repos
for graders

assessment

check
reproducibility via
clone + compile

feedback through
issues



R ★ 0 ⚡ 0

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"

git + GitHub

day one

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.

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

students need to remember to pull before starting work

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

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

sample exercises

scraping data off the web + interactive visualization

scrape

scrape data with
rvest from
goduke.statsgeek.com

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

modeling paris paintings

data

auctions 1764 - 1780
[3,393 × 57]

seller / buyer,
painter, painting
classification,
painting attributes

	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	author	style	winner
2517	R1777-86	R1777	86	R	1777	D/FL	D/FL	D/FL	0	620.0	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
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

clean

clean the data
with (mostly)
dplyr

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 l'armont le long de la table , & sur l'autre font des gens à table , d'autres qui jouent des instruments

model log(price),
and do model
selection



interest & impact

interest

duke focus

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

interest in What if?

no hard data
available, 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%

pipeline to stats

% to StatSci

2014: 19%
declared

2015: 38%
expressed interest

curriculum

plan to offer as
Statistical Science
gateway course
starting in
2017 / 2018

gender distribution

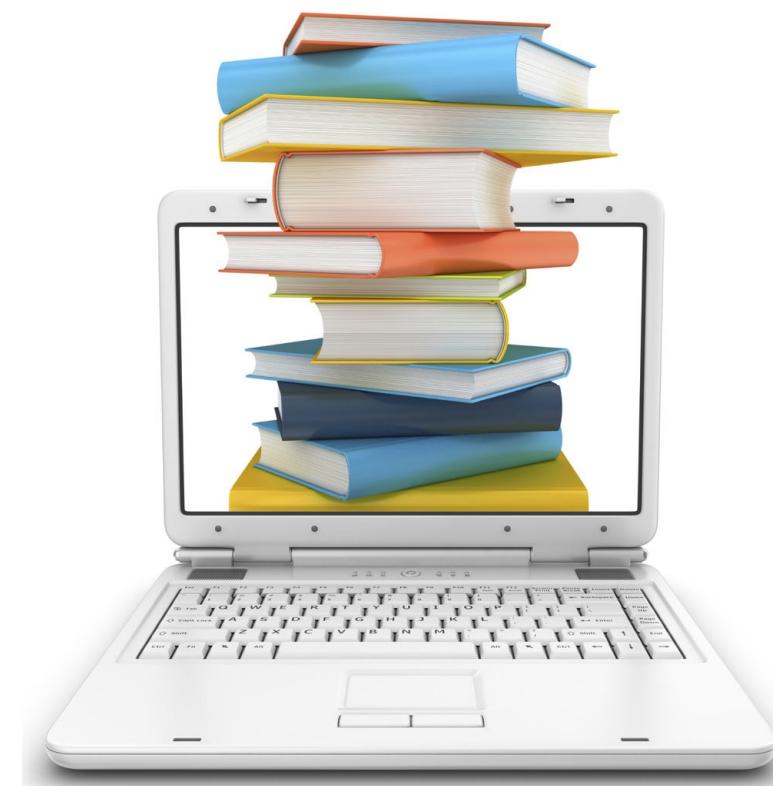
% female

2014: 44%

2015: 50%

compared to ~25%
in current gateway
(Probability)

thank you!



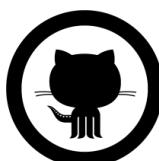
bit.ly/user_data_sci



@minebocek



mine@stat.duke.edu



[mine-cetinkaya-rundel](https://github.com/mine-cetinkaya-rundel)