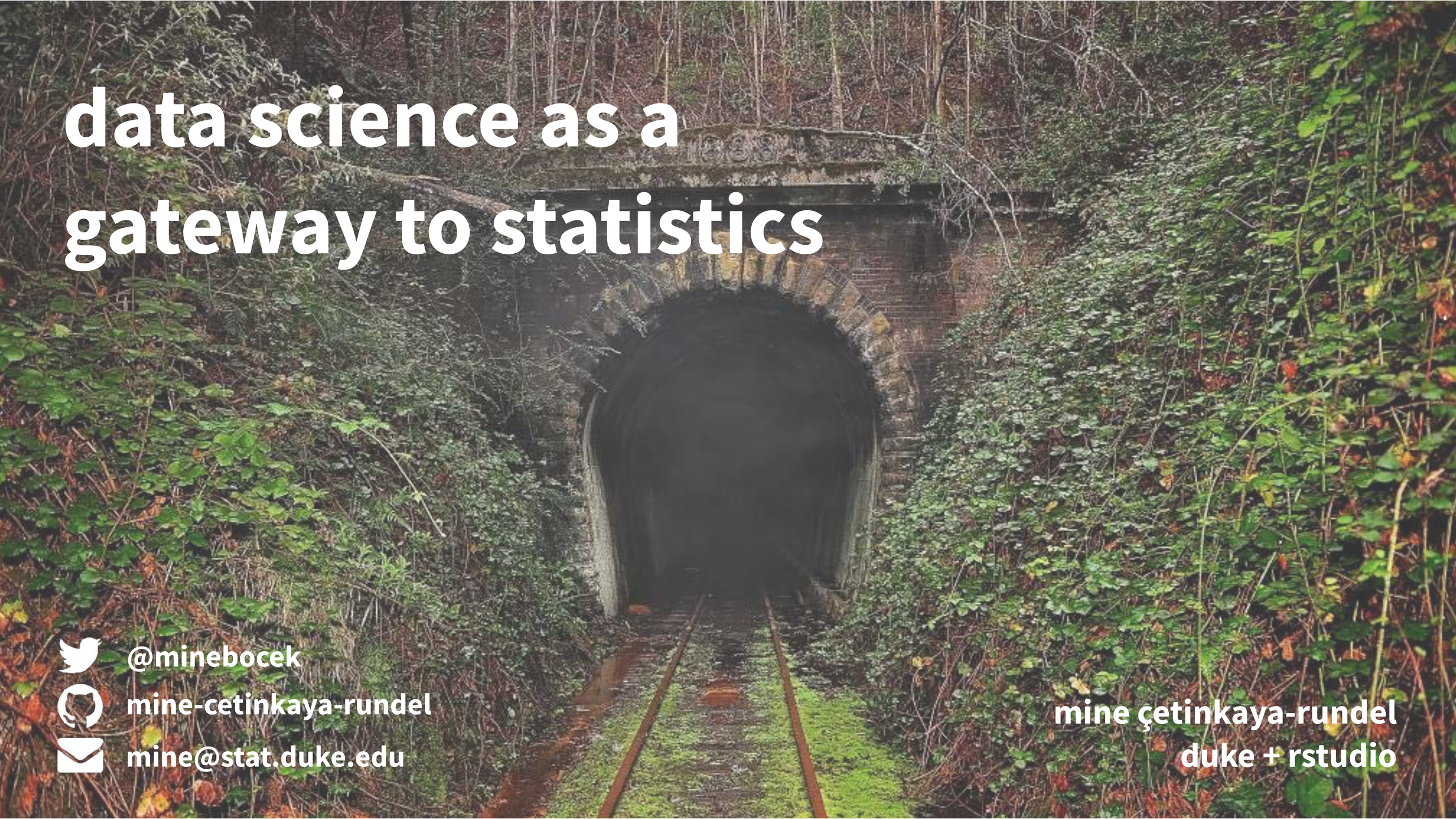
data science as a gateway to statistics



@minebocek

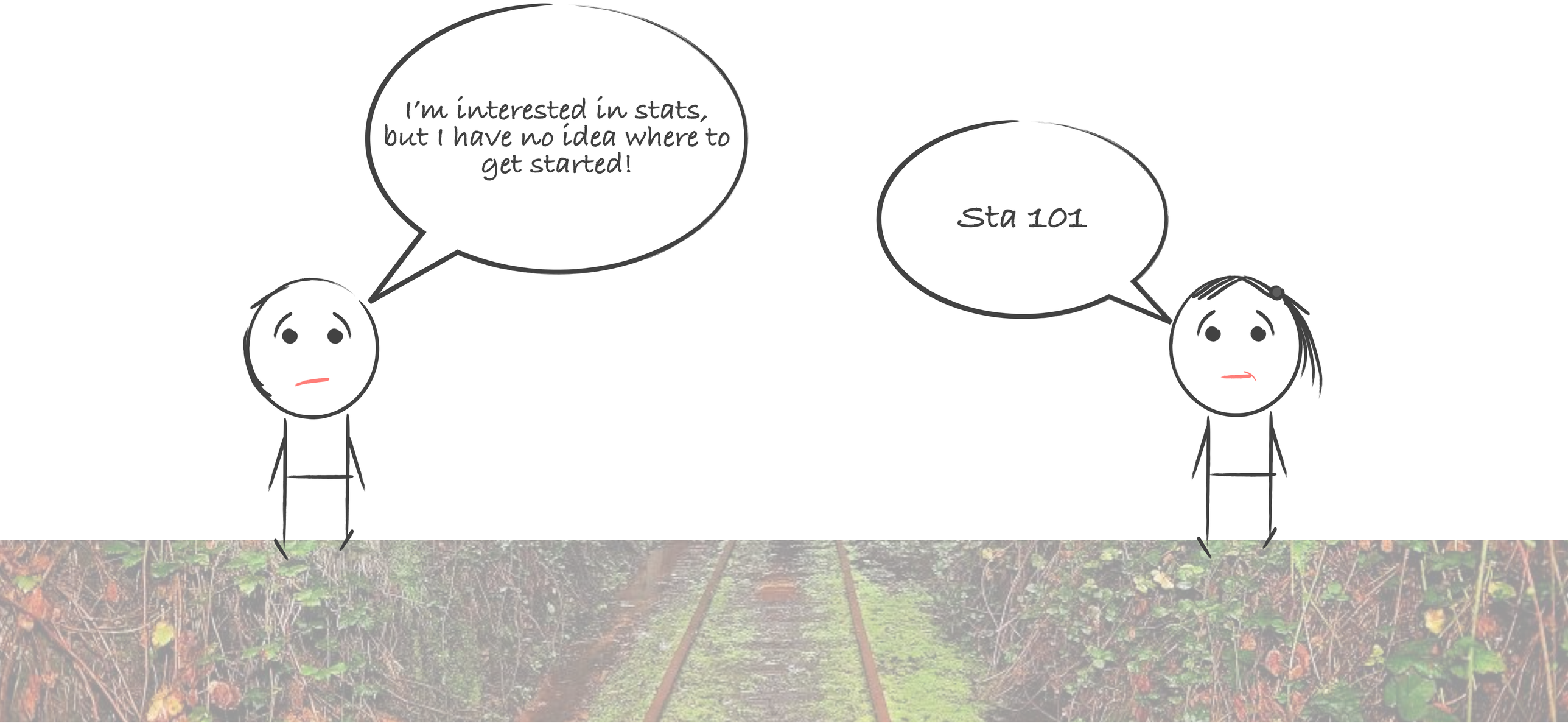
mine-cetinkaya-rundel

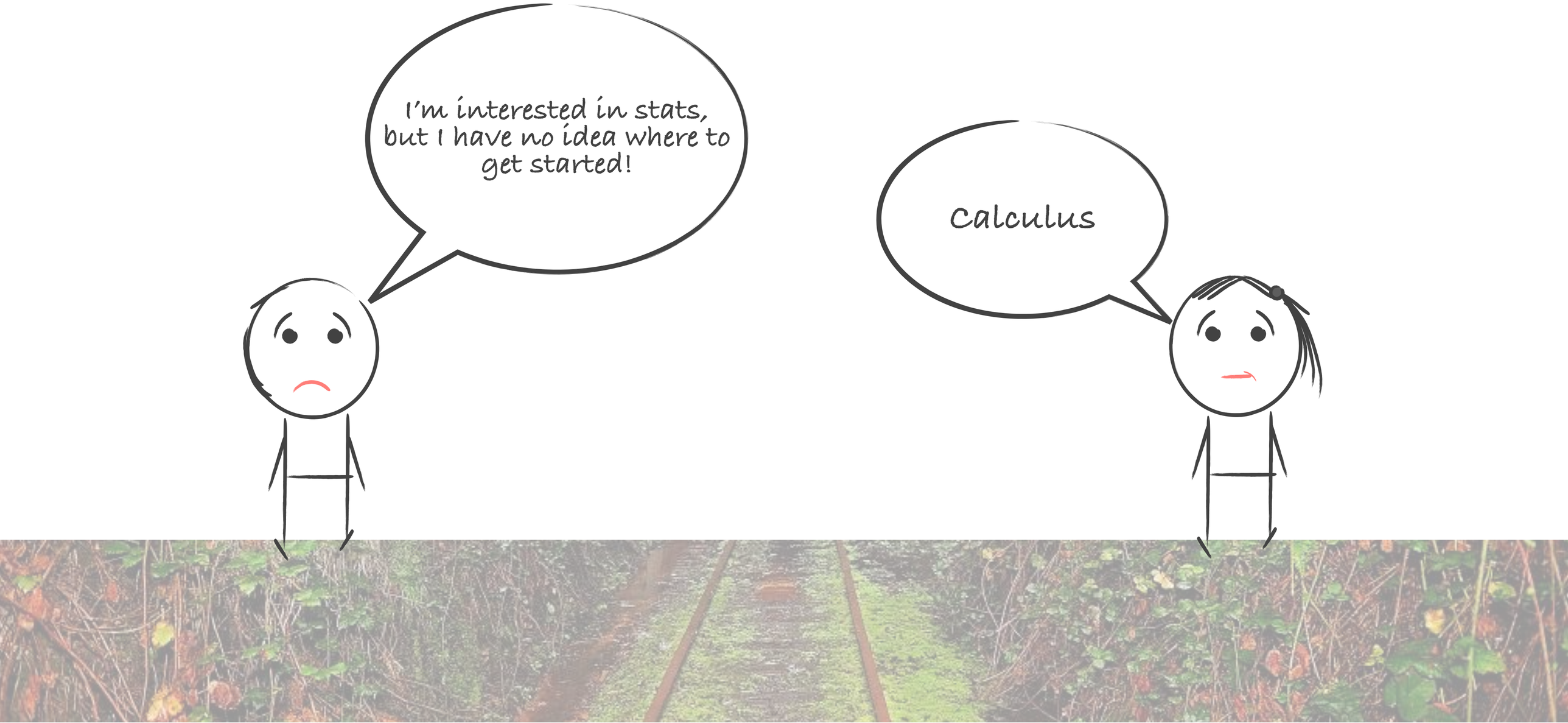
mine@stat.duke.edu

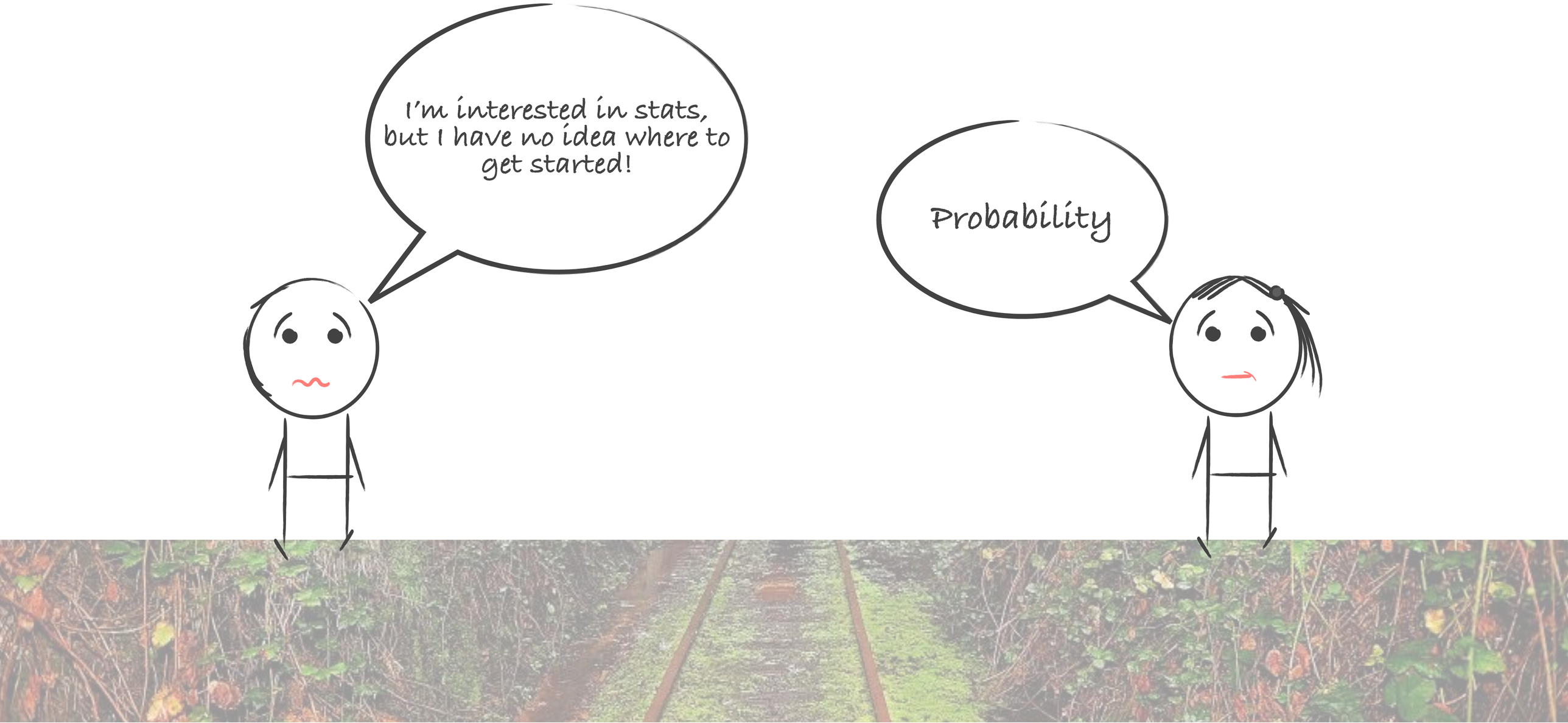
mine çetinkaya-rundel

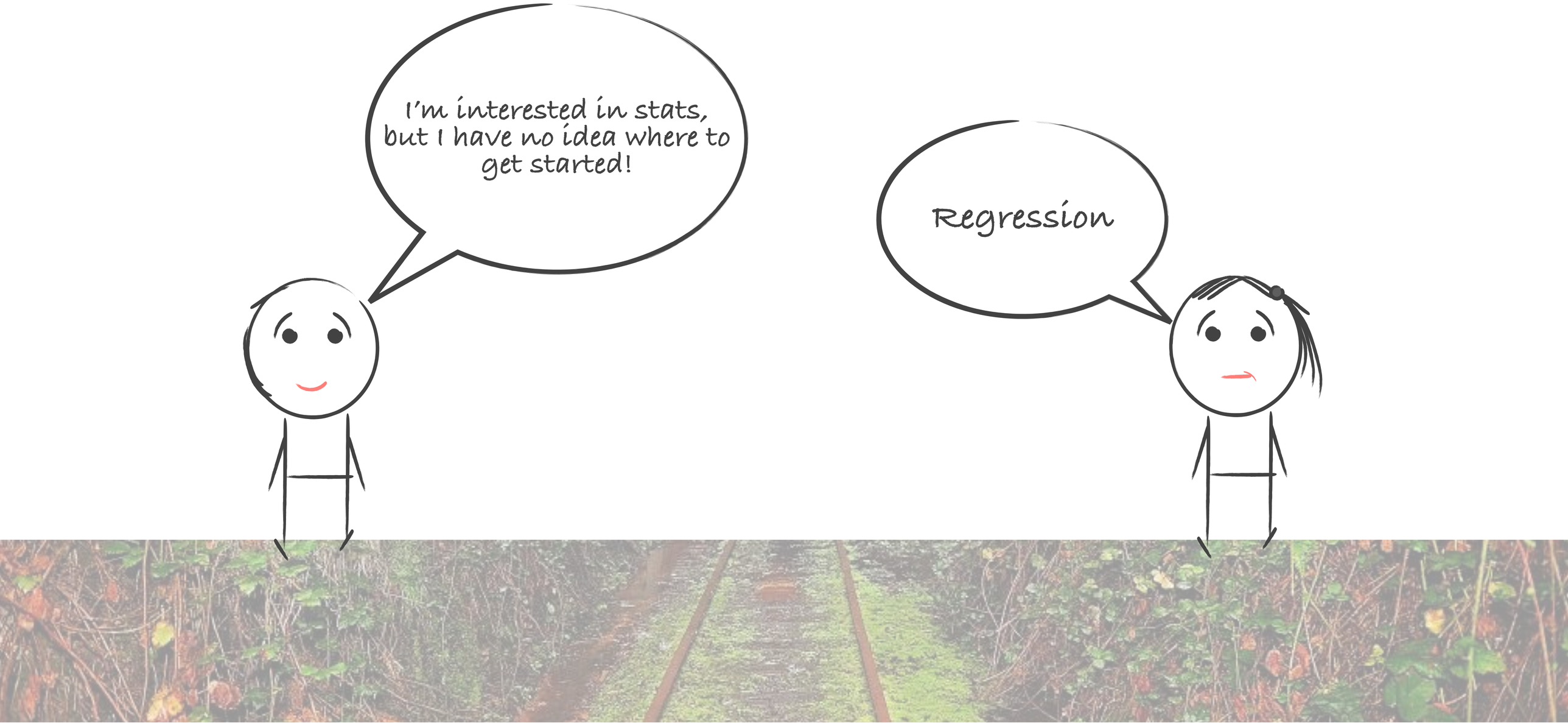
duke + rstudio


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

course overview

computing

data analysis examples

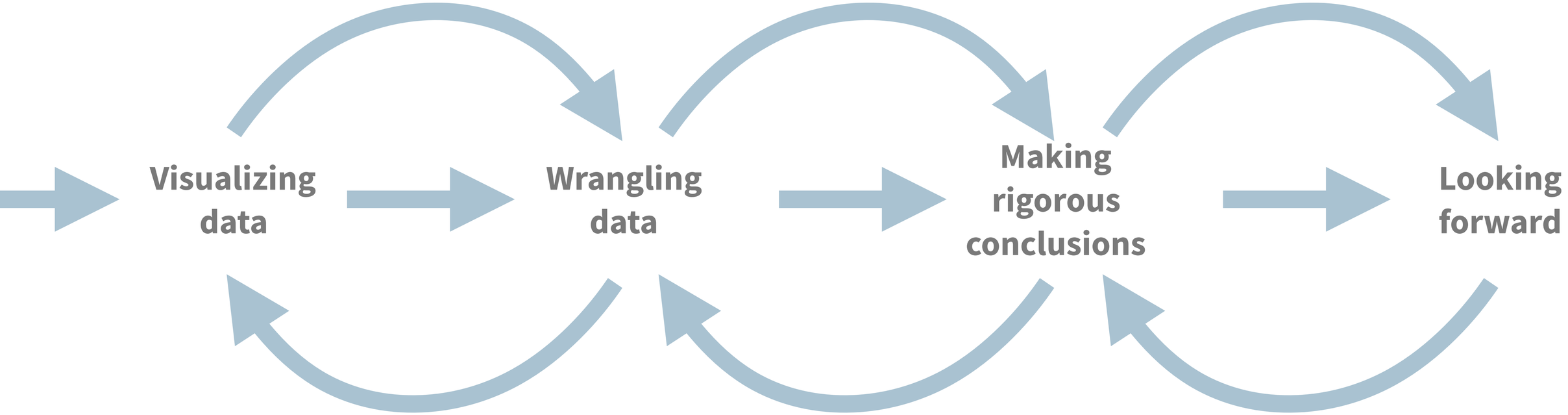
successes & challenges

goal

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

- modern
- places data front and center
- quantitative (but without math prereqs)
- different than HS stats
- challenging (but not intimidating)

curriculum



Fundamentals of data & data viz, revision exercises, confounding variables and Simpson's paradox (and git/GitHub)

Tidy data, data frames vs. summary tables, recoding and transforming variables (for fun, and for modelling)

Building and selecting models, visualizing interactions, prediction and model validation, inference via simulation & discussion of CLT

Web scraping, interactive visualization and reporting with Shiny, Bayesian inference, ???

course overview

opinionated computing

data analysis examples

successes & challenges

start up instructions

Local install

- Install R
- Install RStudio
- Install the following packages:
  - rmarkdown
  - tidyverse
  - …
- Load these packages
- Install git

vs.

RStudio Cloud

- Go to rstudio.cloud
- Log in with your Google ID & pass

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"
      )
  )
```

recoding a multi-level variable

base R

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

vs.

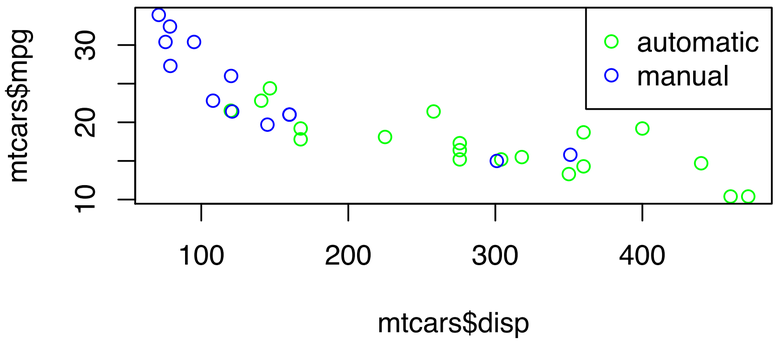
tidyverse

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

visualizing multiple variables

base R

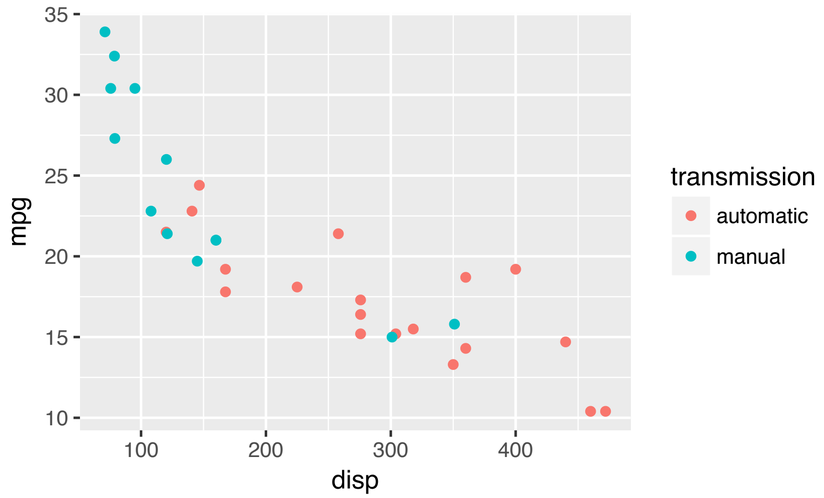
```
mtcars$trans_color <-
  ifelse(mtcars$transmission == "automatic",
         "green",
         "blue")

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



vs.

tidyverse

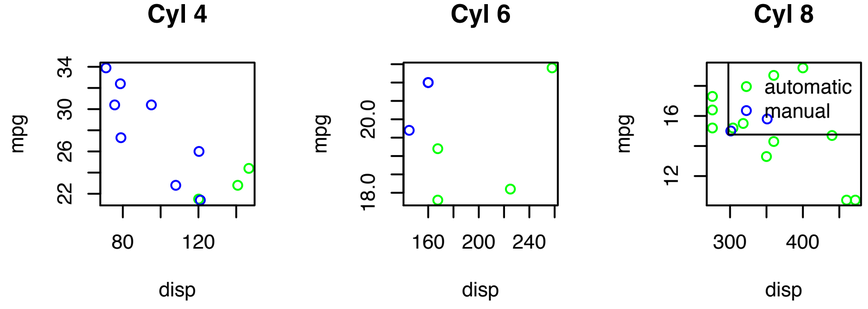
```
ggplot(mtcars,
       mapping = aes(
         x = disp, y = mpg,
         color = transmission
         )) +
  geom_point()
```



visualizing even more variables

base R

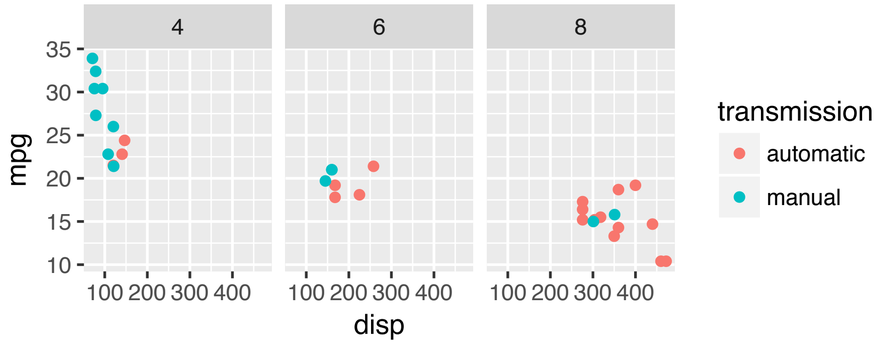
```
mtcars_cyl4 <- mtcars[mtcars$cyl == 4, ]
mtcars_cyl6 <- mtcars[mtcars$cyl == 6, ]
mtcars_cyl8 <- mtcars[mtcars$cyl == 8, ]

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



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 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 1)

marketability + discoverability

course overview

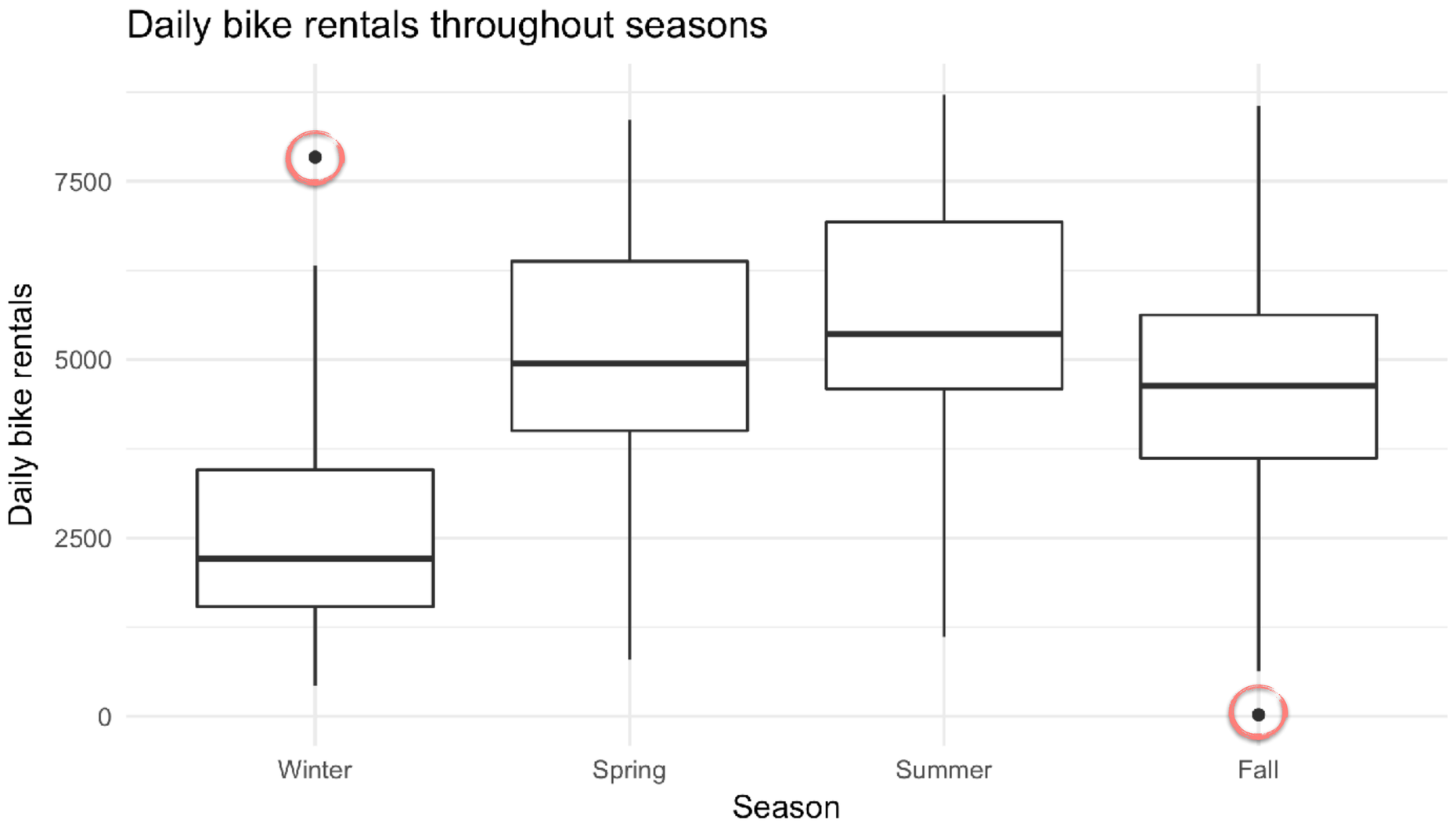
computing

data analysis examples

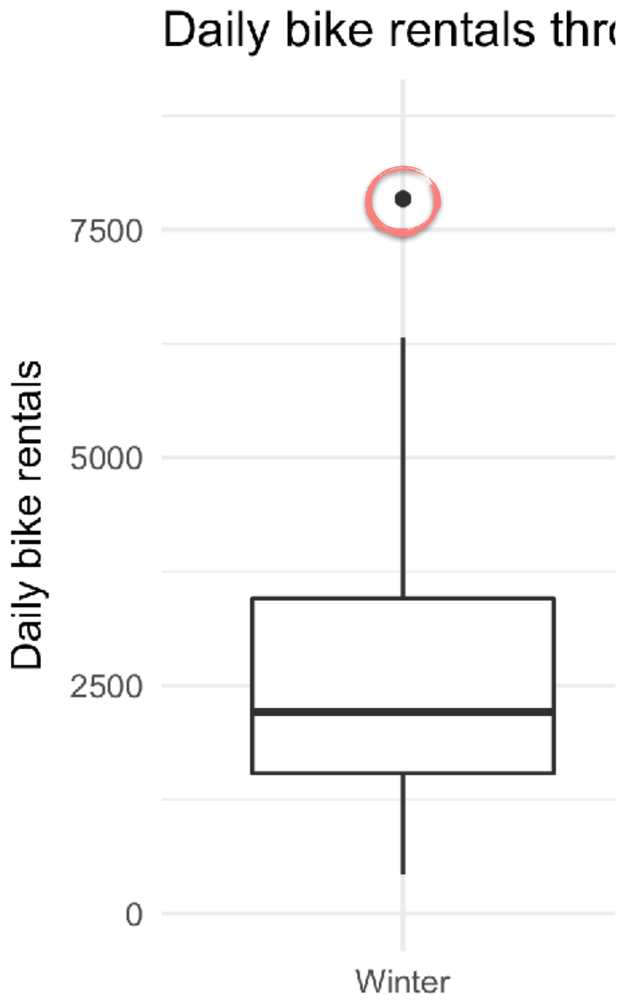
successes & challenges

#1 citybikes in DC

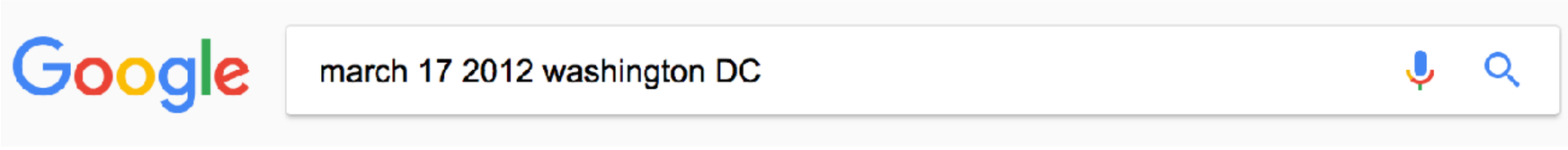
Question 8. Create a visualization displaying the relationship between bike rentals and season. Interpret the plot in context of the data.



Question 8. Create a visualization displaying the relationship between bike rentals and season. Interpret the plot in context of the data.

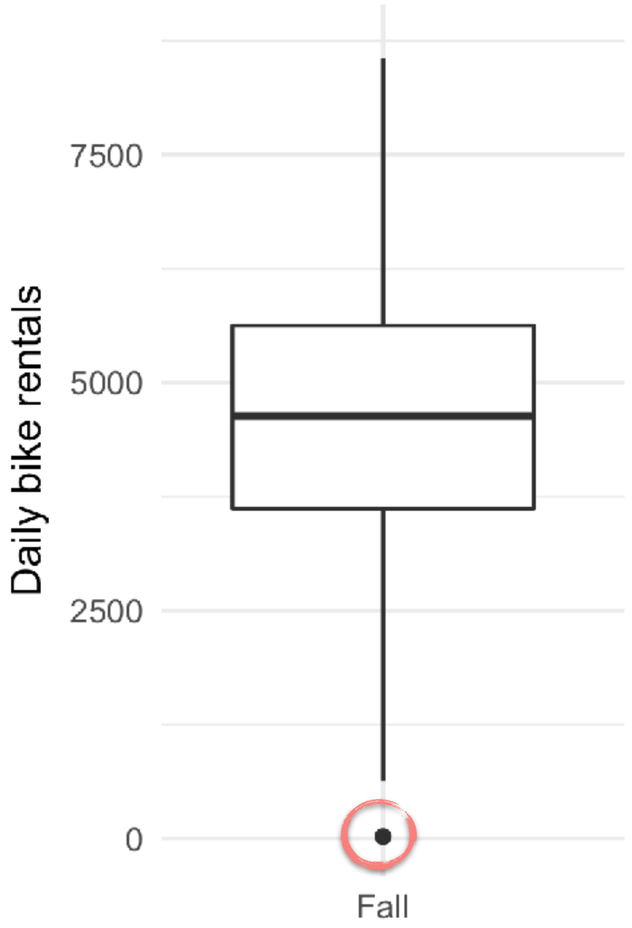



```
bike %>%
  filter(season == "Winter") %>%
  summarise(max = max(cnt), day_max = dteday[which.max(cnt)])
## # A tibble: 1 x 2
##     max day_max
##   <dbl> <date>
## 1  7836 2012-03-17
```

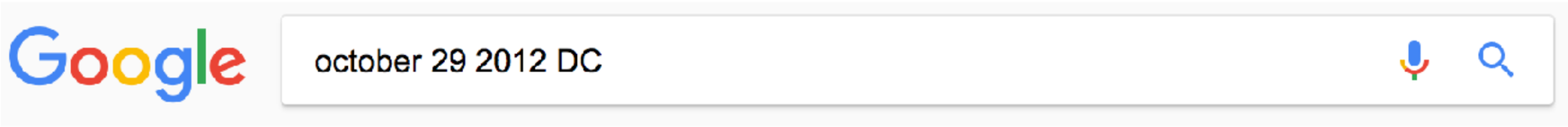


President Obama at the Dubliner on St. Patrick's Day | whitehouse.gov
https://obamawhitehouse.archives.gov/.../2012/.../17/president-obama-dubliner-st-patr...
17 Mar 2012 - President Barack Obama is reflected in a mirror at the Dubliner, an Irish pub in **Washington, D.C.**, with his Irish cousin, Henry Healy, and Ollie Hayes, a pub owner in Moneygall, Ireland, on St. Patrick's Day, **Saturday, March 17, 2012**. (Official White House Photo by Pete Souza). President Obama Greets the ...

Question 8. Create a visualization displaying the relationship between bike rentals and season. Interpret the plot in context of the data.




```
bike %>%
  filter(season == "Fall") %>%
  summarise(min = min(cnt), day_min = dteday[which.min(cnt)])
## # A tibble: 1 x 2
##     min day_min
##   <dbl> <date>
## 1  22.0 2012-10-29
```



Washington DC shuts down in preparation for hurricane Sandy | US ...
https://www.theguardian.com/.../2012/oct/29/washington-dc-shutdown-hurricane-sand...
28 Oct 2012 - Panic buying and fears over capability of power companies as first light rains reach city.

Hurricane Sandy: high winds and flooding hit US east coast – Monday ...
https://www.theguardian.com/.../us.../2012/oct/29/hurricane-sandy-new-york-live-blo...
30 Oct 2012 - Announces plans for a Storm Recovery Support Call Center operated by @Delaware_DHSS will open tomorrow at noon. Gov. Markell: The message for tonight is to hunker down. Weather Underground (@wunderground). 470,000 without power in 13 states and **DC**. **October 29**,

learning goals

main
prediction and
model selection

get for free
use of
outside data

#2 paris paintings

data source: auction catalogs

89 Deux tableaux très riches de com- *1666*
position, 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

enrichi...
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	Q
1	name	sale	lot	dealer	year	origin_ author	origin_ cat	school_ pntg	diff_origin	price	count	subject	authorstandard	artistliving	authorstyle	author	winner
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	Donjeux
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

- **mat** - category of material (a=silver, al=alabaster, ar=slate, b=wood, bc=wood and copper, br=bronze frames, bt=canvas on wood, c=copper, ca=cardboard, co=cloth, e=wax, g=grissaille technique, h=oil technique, m=marble, mi=miniature technique, o=other, p=paper, pa=pastel, t=canvas, ta=canvas?, v=glass, n/a=NA, (blanks)=NA)
- **Shape** - shape of painting

```
pp <- pp %>%
  mutate(
    Shape = fct_collapse(Shape, oval = c("oval", "ovale"),
                                round = c("round", "ronde"),
                                squ_rect = "squ_rect",
                                other = c("octogon", "octagon", "miniature")),
    mat = fct_collapse(mat, metal = c("a", "br", "c"),
                            canvas = c("co", "t", "ta"),
                            paper = c("p", "ca"),
                            wood = "b",
                            other = c("e", "g", "h", "mi", "o", "pa", "v", "al", "ar", "m"))
  )
```

learning goals

main

data provenance

modelling
diagnostic, log
transformation

get for free

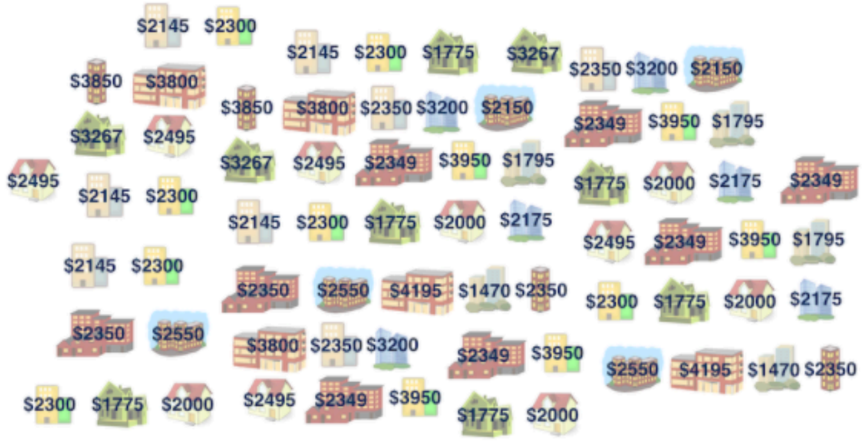
iterative
data
cleanup, i.e.
working with other
people's data

#3 manhattan apartments

observed sample

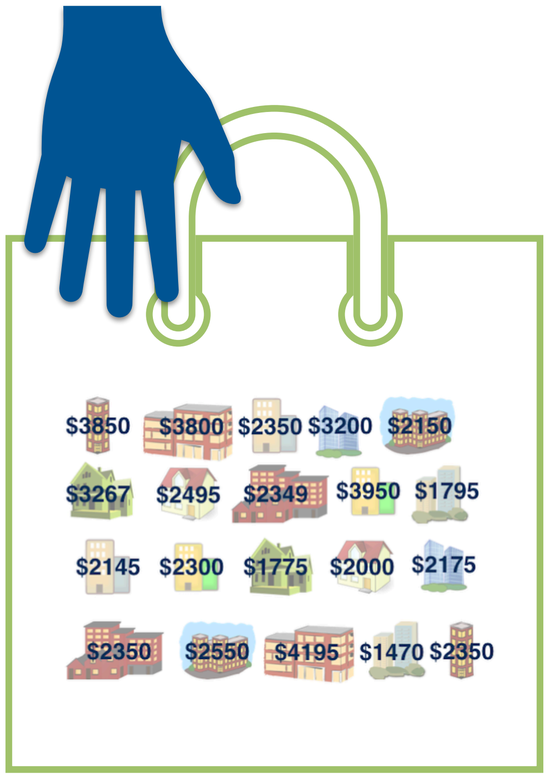
Sample median = $2350 😱

population



Population median = ❓

first, tactile simulation



Sample:

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20

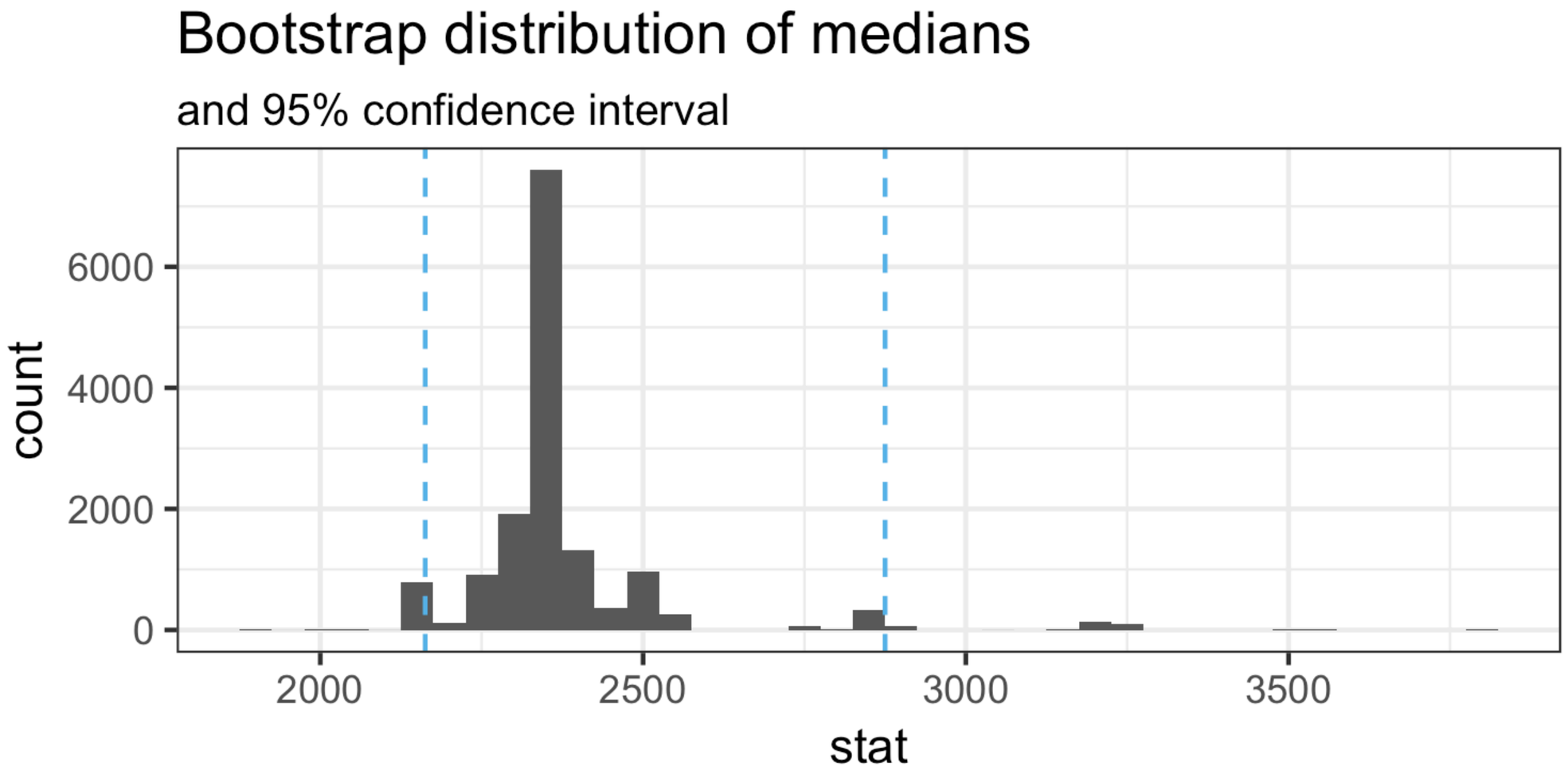
Ordered sample:

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20

Bootstrap median:

then, computational simulation

```
library(infer)

manhattan %>%

  # specify the variable of interest
  specify(response = rent) %>%

  # generate 15000 bootstrap samples
  generate(reps = 15000, type = "bootstrap") %>%

  # calculate the median of each bootstrap sample
  calculate(stat = "median")
```


Bootstrap distribution of medians
and 95% confidence interval
count
6000
4000
2000
0
2000
2500
3000
3500
stat

learning goals

main
estimation
via
bootstrapping

get for free
discussion on
representativeness
of samples

#4 basketball

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	1	Brooklyn, N.Y. (Barclays Center)	W 70-59	10,046	9:30 p.m.	TruTV
11/26		**Furman**	1	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	1	Syracuse, N.Y. (Carrier Dome)	W 80-72	35,446	6 p.m.	ESPN
2/18	*	[15] **North Carolina**	1	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	1	Indianapolis, Ind. (Lucas Oil Stadium)	W 81-61	72,238	6 p.m.	TBS/TNT
4/6	!!!!!!!	vs. [3] Wisconsin	1	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		[16] Opponent	Rank	Location (Qty.)	Result (OT)	Pts.	Att.	TV	
4	14-Nov	-	**Presbyterian**	4	Durham, N.C. (C	W	113-44	9,314	6 p.m.	ESPNU
5	15-Nov	~	**Fairfield**	4	Durham, N.C. (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. (C	W	93-54	9,314	5 p.m.	ESPNU
10	30-Nov		**Army**				93-72	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**		Du			9,314	9 p.m.	ESPN/ACCN
30	21-Feb	*	**Clemson**		Durham, N.C. (C	W		9,314	4 p.m.	ESPN
31	25-Feb	*	at Virginia Tech	4	Blacksburg, Va.	W	91-56 •	9,847	9 p.m.	ESPN2
32	28-Feb	*	**Syracuse**	4	Durham, N.C. (C	W	73-54	9,314	7 p.m.	ESPN
33	4-Mar	*	**Wake Forest**	3	Durham, N.C. (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	15,945	7 p.m.	CBS
38	22-Mar	!!!	vs. San Diego St	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

learning goals

main
data
harvesting

get for free
parsing
text
strings

course overview

computing

data analysis examples

successes & challenges

Better Living Through Data Science:
Exploring / Modeling / Predicting / Understanding

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

interest in DS:

% of What If applicants interested in DS

2015: 76%

2016: 83%

2017: 100%

pipeline for stats:

increasing interest in stats major from students in this course

Introduction to Data Science and Statistical thinking

offered for the first time in Spring 2018 as gateway to major and quantitative disciplines

enrollment > cap

lots of logistical updates to accommodate going from 18 to 80 students

growing pains

active “big group” time is difficult to manage

add “small group” lab sections

course open to all = students from all levels

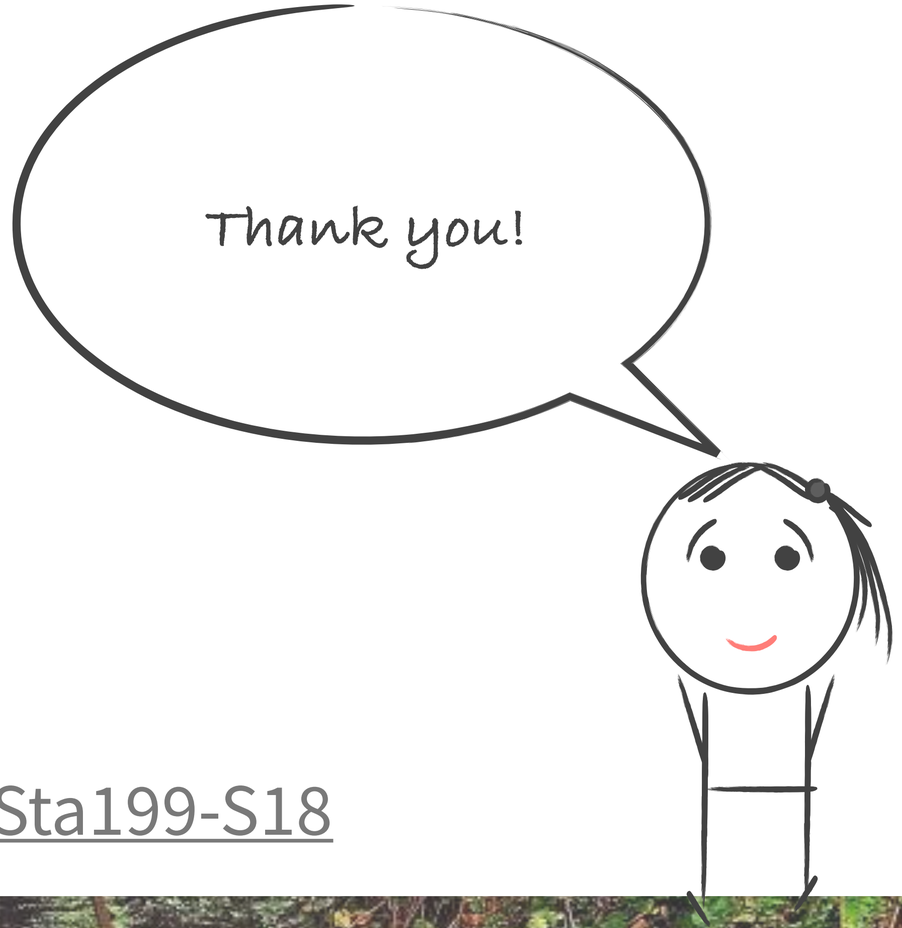
better placement guidelines

tl;dr

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!



course web: bit.ly/sta199-s18

course GitHub org: https://github.com/Sta199-S18

@minebocek

mine-cetinkaya-rundel

mine@stat.duke.edu