

431 Class 01

Thomas E. Love

2019-08-27

This is PQHS 431 / CRSP 431 / MPH 431

I USED TO THINK
CORRELATION IMPLIED
CAUSATION.



THEN I TOOK A
STATISTICS CLASS.
NOW I DON'T.



SOUNDS LIKE THE
CLASS HELPED.
WELL, MAYBE.



Please take a copy of (1) the **survey** and (2) an **index card** from the teaching assistants.

Wait for further instructions before writing anything down.

Instructions for the Survey

Please read these instructions carefully before writing anything down.

- ① Introduce yourself to someone that you don't know.
- ② Record the survey answers **for that other person**, while they record your responses.
- ③ Be sure to complete all 15 questions (both sides of the paper).
- ④ Also, write **YOUR** answer to question #4 on the index card, and keep that, for now. You'll need it later.
- ⑤ When you are finished, thank your partner and raise your hand.
Someone will come to collect your survey.

Regarding Question 4, Professor Love is the large fellow standing in the front of the room.

Course Details

Instructor: Thomas E. Love, Ph.D.

Email (best way to reach me): Thomas.Love@case.edu

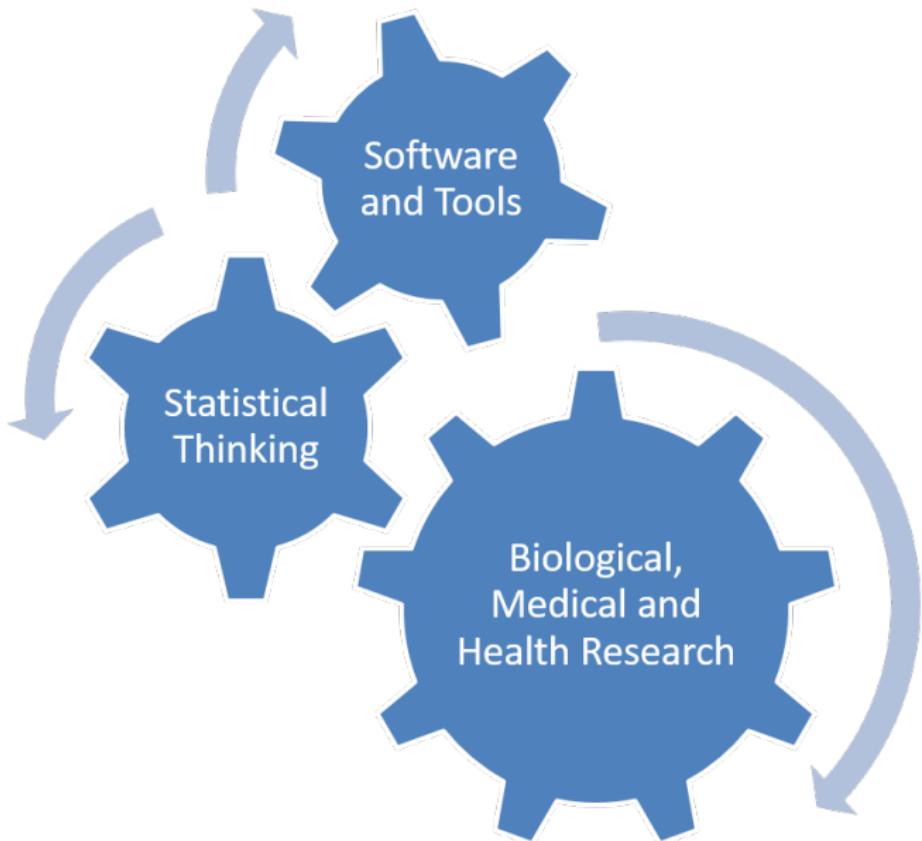
Our web site: <https://github.com/THOMASELOVE/2019-431>

Links there to:

- Course Syllabus
- Course Calendar
- Course Notes (essentially a textbook)
- Slides Page (in-class materials, slides as PDF and R Markdown)
- Software Details (R and R Studio, installation and R Basics)
- Data and Code
- Deliverables, Quizzes and Project (after Labor Day)

How to Get Help: 431-help@case.edu

What is this course about?



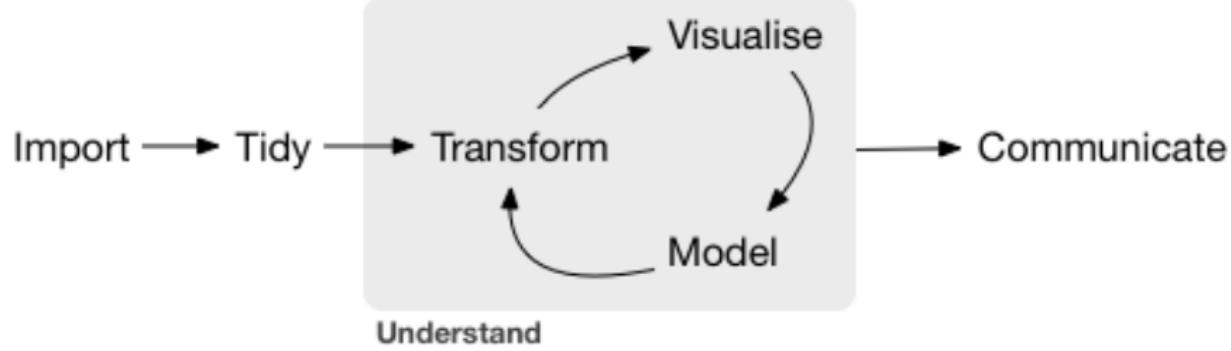
What is this course about?

- a. Exploratory Data Analysis, Visualization
- b. Statistical Inference, Making Comparisons
- c. Linear Regression and related Models

The course is about biostatistics, replicable research, using state-of-the-art tools (R, R Studio, R Markdown), and thinking about how science is most effectively done.

- It is more a course in **how** to do things (highly applied) rather than a theoretical/mathematical justification for **why** we do them. We focus here on practical work.
- It's mostly about getting you doing data science projects for biological, medical and health applications.

What is Data Science about?



Program

Source: <http://r4ds.had.co.nz/introduction.html>

Teaching Assistants ([email 431-help@case.edu](mailto:431-help@case.edu))

431: 2019 Teaching Assistants

431-help@case.edu



Amin
Saad
MD



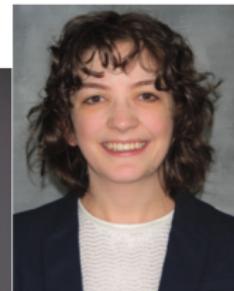
Amr
Mahran
MD MS



Harry
Persaud
BS



Joseph
Hnath
BA



Julijana
Conic
MD



Noah
Lorincz-Comi
BA MSc

- All TAs work with PQHS/CRSP/MPHP students and both Sections.
- TA **office hours** start Tuesday, details on Course Calendar.

To get help at any time starting now, email 431-help@case.edu

What will we be reading?

Dr .Love's Course Notes

Data Science for Biological, Medical and Health Research: Notes for PQHS 431

Thomas E. Love, Ph.D.

Version: 2019-08-26 21:35 29

Introduction

These Notes provide a series of examples using R to work through issues that are likely to come up in PQHS/CRSP/MPH 431.

Silver

new york times bestseller
noise and the noise
the signal and the noise
and the noise and the noise
and the noise and the noise
why so many noisy predictions fail—but some don't
and the noise and the noise and the noise
data silver the noise

"Could turn out to be one of the more momentous books of the decade." —The New York Times Book Review

Leek

The Elements of Data Analytic Style



Jeff Leek

R4DS

O'REILLY



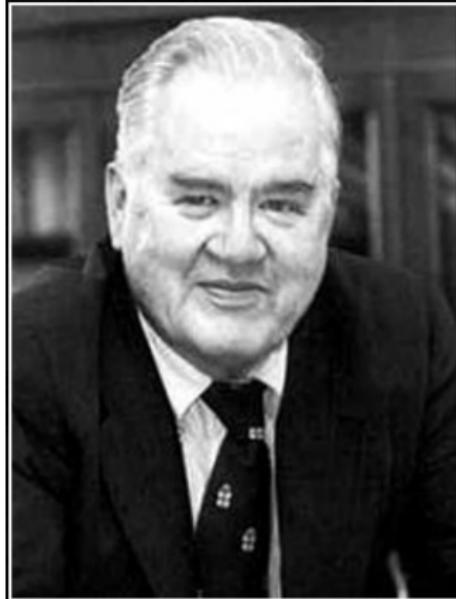
R for Data Science

VISUALIZE, MODEL, TRANSFORM, Tidy, AND IMPORT DATA

Hadley Wickham & Garrett Grolemund

Our web site: <https://github.com/THOMASELOVE/2019-431>

Great Statisticians in History



The greatest value of a picture is
when it forces us to notice what we
never expected to see.

— *John Tukey* —

AZ QUOTES

Photo Source: http://www.azquotes.com/author/14847-John_Tukey

John Tukey (1915-2000)

Your First Task (Do Today, Please)

Visit <http://bit.ly/431-2019-welcome-survey> if you haven't already.

You'll need to **log in to Google via CWRU** to see the form.

Gathering Some Data: Age Guessing Activity

- You will join one of ten groups, with 6-7 students in each group.
- Your group will receive a sheet to keep track of your guesses (estimated ages.)
- Your group will then receive one of a series of cards, with a photo of a person on it.
- For each card your group receives...
 - estimate the age of the person on the card
 - write your (group) guess in the table on the sheet in the row corresponding to that numbered card
- Later, you will be told the true ages and will be able to compute errors.

Scientists Gather Their Own Data

If you have a little time between cards, make sure everyone in your group . . .

- ① knows the name and field of everyone else in the group, and knows your group's letter.
- ② writes down a new guess as to my age on their index card, now that you know me better.

So if your initial guess was that I was 18, but now you think I'm 19, your card should read 18/19.

Age Guessing Robots?

Well, yes, of course, there's a tool online to do this. More than one, in fact.

Visit <https://how-old.net/>

<https://how-old.net>



The AI's guess was...

7 years too high

6 years too low

Do you think you did that well?

Card 1



Card 1

Eric Chong
Master Chef Canada winner
Photo date: April 2014

Age 21

Card 2



Card 2

Katherine Archuleta
Former U.S. OPM Director
Photo date: 2013

Age 64

Card 3



Card 3

Elise Mayfield
Chef, Actor, Baker
Photo date: 2014

Age 28

Card 4



Card 4

Kevin Love
(then) High School Student
Photo date: June 2014

Age 14

No, not THAT Kevin Love



THIS Kevin Love, on the right (January 2019)



Card 5



Card 5

Rosemary McGinn

Photo date: July 2013

Age 54

Card 6



Card 6

John Chaney
Basketball Coach
Photo date: 2006

Age 74

Card 7



Card 7

David Storm

Photo date: August 2014

Age 44

Card 8



Card 8

Margo Glantz
Writer
Photo date: 2013

Age 83

Card 9



Card 9

Quade Ross Honey
Fugitive
Photo date: 2012

Age 24

Card 10



Card 10

Bianca Lawson
Actress
Photo date: 2013

Age 34

So, How did we do?



#1 Age 21



#2 Age 64



#3 Age 28



#4 Age 14



#5 Age 54



#6 Age 74



#7 Age 44



#8 Age 83



#9 Age 24



#10 Age 34

Collecting the Results

We'll collect some key results in a Google sheet, that you should be able to reach when logged into CWRU for Google.

The sheet is at <http://bit.ly/431-2019-day1-ageguess>

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	group	group size	females	card 1	card 2	card 3	card 4	card 5	card 6	card 7	card 8	card 9	card 10
2	A												
3	B												
4	C												
5	D												
6	E												
7	F												
8	G												
9	H												
10	I												
11	J												

And how did the AI at <https://how-old.net> do?



#1 Age 21
AI guess 27



#2 Age 64
AI 44



#3 Age 28
AI 22



#4 Age 14
AI 19



#5 Age 54
AI 36



#6 Age 74
AI 63



#7 Age 44
AI 55



#8 Age 83
AI 79



#9 Age 24
AI 35

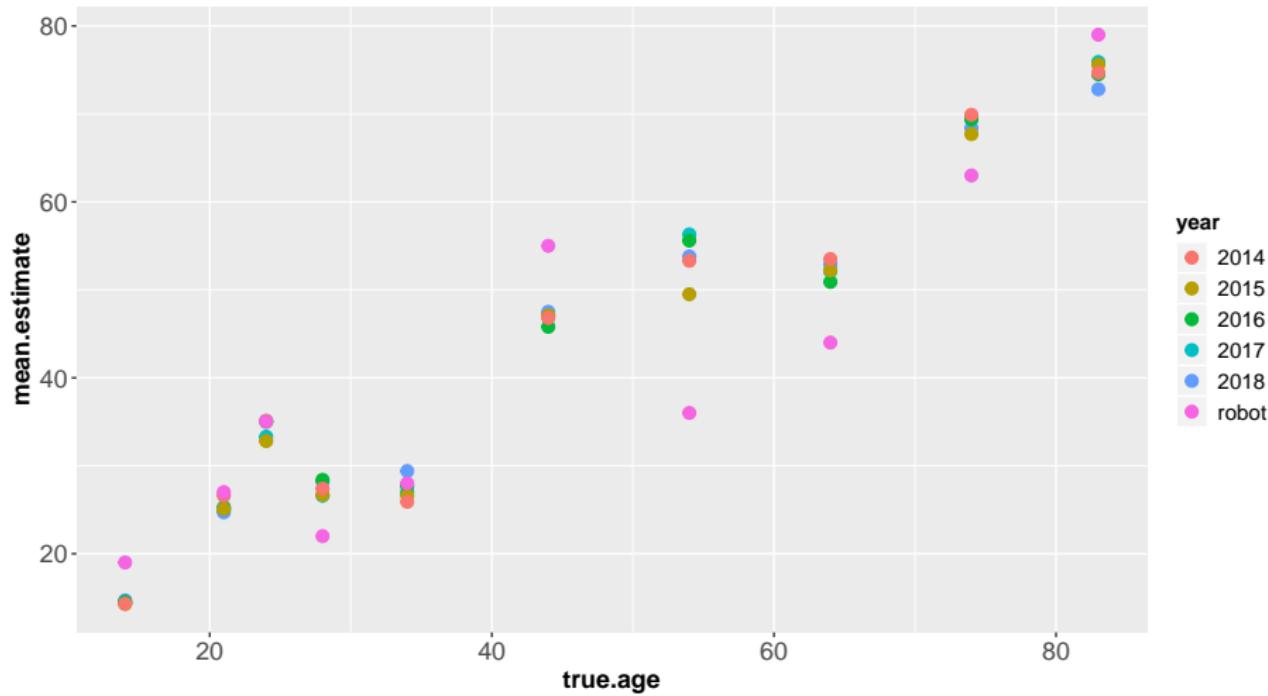


#10 Age 34
AI 28

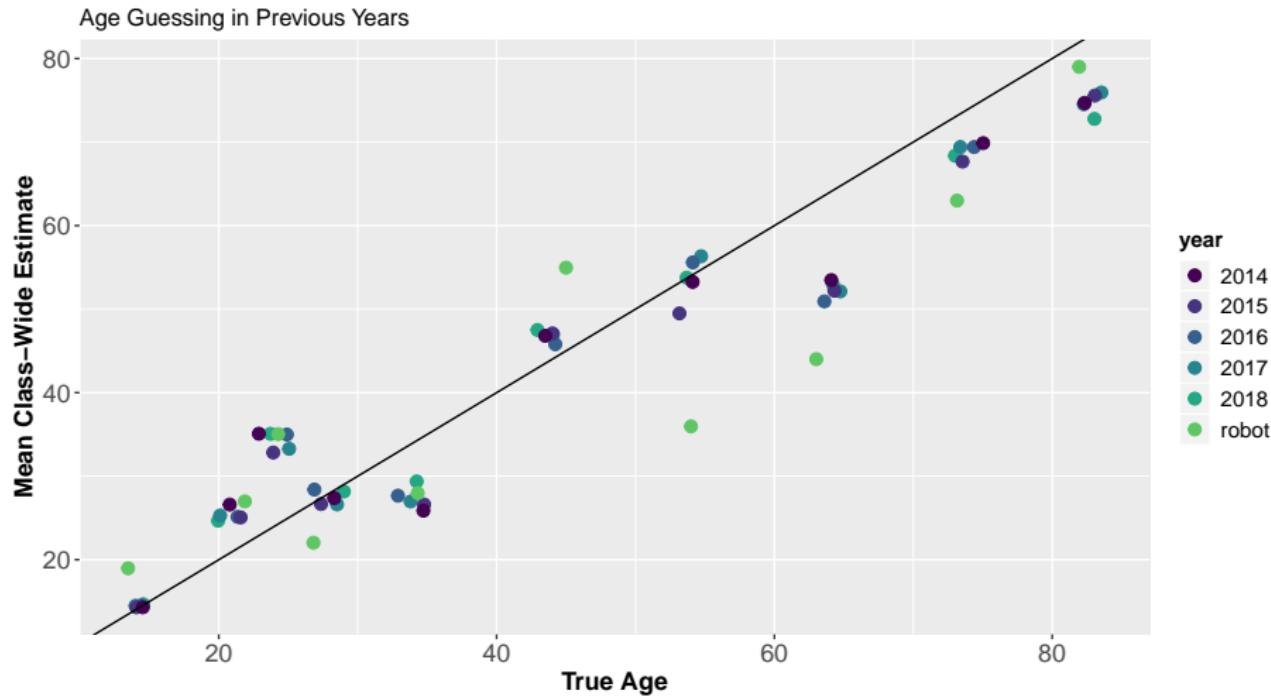
photo-age-history-2018.csv Data Set (excerpt)

card	label	true.age	sex	facing	year	mean.estimate	error
1	Chong	21	M	R	2018	24.7	3.7
2	Archuleta	64	F	L	2018	52.9	-11.1
3	Mayfield	28	F	L	2018	28.2	0.2
4	Love	14	M	L	2018	14.7	0.7
5	McGinn	54	F	R	2018	53.8	-0.2
6	Chaney	74	M	L	2018	68.4	-5.6
7	Storm	44	M	R	2018	47.5	3.5
8	Glantz	83	F	L	2018	72.8	-10.2
9	Honey	24	M	L	2018	35.1	11.1
10	Lawson	34	F	R	2018	29.4	-4.6
1	Chong	21	M	R	2017	25.3	4.3
2	Archuleta	64	F	L	2017	52.1	-11.9

Scatterplot of Prior Results, 1



Scatterplot of Prior Results, 2



Mean Class-Wide Guesses (2014-17 combined)



#1 Age 21 2014-17 Mean Guesses	25.5	#2 Age 64 52.2	#3 Age 28 27.3	#4 Age 14 14.4	#5 Age 54 53.7
#6 Age 74	69.1	#7 Age 44	46.7	#8 Age 83 75.2	#10 Age 34 26.8



Mean Class-Wide Errors (2014-17 combined)



#1 Age 21
2014-17
Errors +4.5
 -4.9

#2 Age 64
 -11.8
 +2.7

#3 Age 28
 -0.7
 -7.8

#4 Age 14
 +0.4
 +10.1
 -7.2

#6 Age 74

#7 Age 44

#8 Age 83

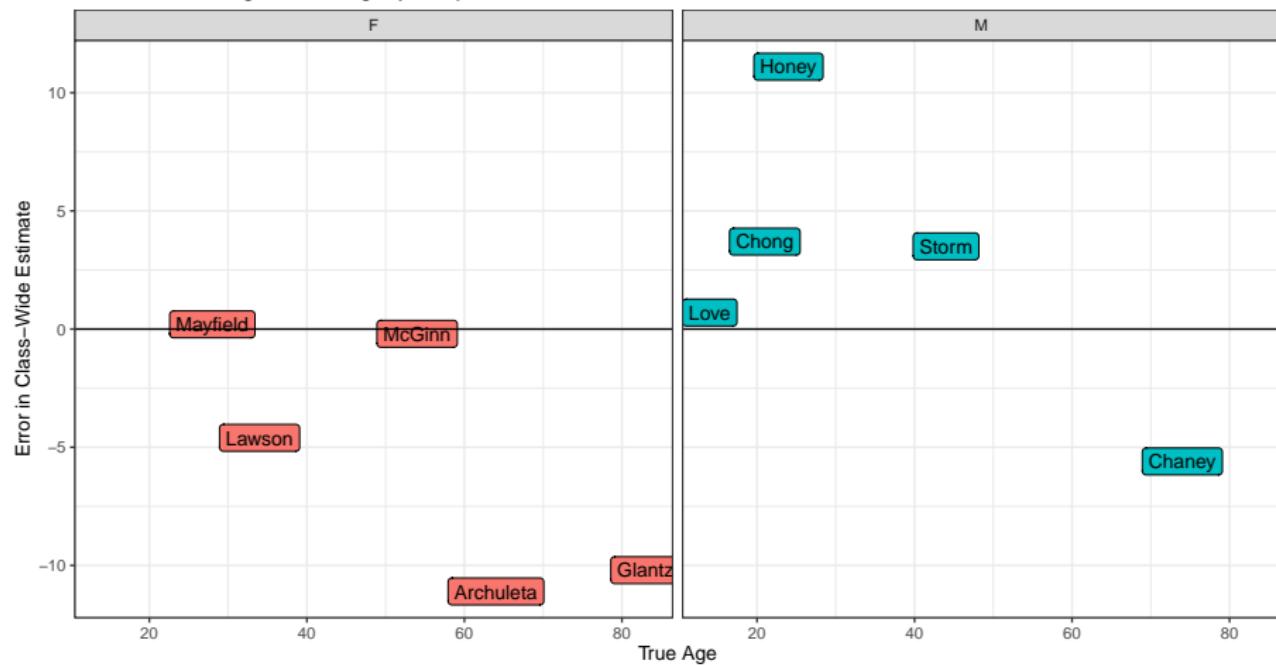
#9 Age 24

#10 Age 34



Scatterplot of 2018 Results with Labels

Errors in 2018 Age Guessing, by Subject's Sex



Hans Rosling and “The Joy of Stats”

200 countries over 200 years using 120,000 numbers, in about 4 minutes.

<http://bit.ly/431-rosling>

And if you liked that ...

- The 20 minute version (from 2007):
<https://www.youtube.com/watch?v=RUwS1uAdUcl>
- The full documentary from the BBC:
<https://www.gapminder.org/videos/the-joy-of-stats/>
- Video playlist from Gapminder: <https://www.gapminder.org/videos/>

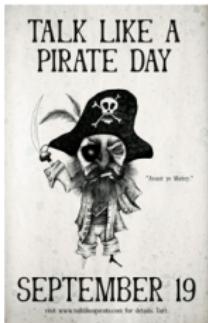
What's next?



RStudio makes R easier to use. It includes a code editor, debugging & visualization tools.



R Packages



R Markdown

from R Studio

Analyze. Share. Reproduce.

Your data tells a story. Tell it with R Markdown.
Turn your analyses into high quality documents, reports, presentations and dashboards.

What's next?

- ① Turn in your index card with your two guesses of my age, please.
 - ① Visit <http://bit.ly/431-2019-welcome-survey> if not already done.
 - ② Follow the software instructions to get R, R Studio, R Packages and 431 Data on your computer.
 - ③ Obtain Jeff Leek's [The Elements of Data Analytic Style](#).
 - ④ Obtain Nate Silver's [The Signal and the Noise](#).
 - ⑤ Read [the syllabus](#) and look at the rest of [the website](#). Make sure you view the [Course Notes](#).
 - ⑥ Ask us questions. TA office hours start next week, but email is available now.
- Course Web Site: <https://github.com/THOMASELOVE/2019-431>
 - Want help? Email 431-help@case.edu