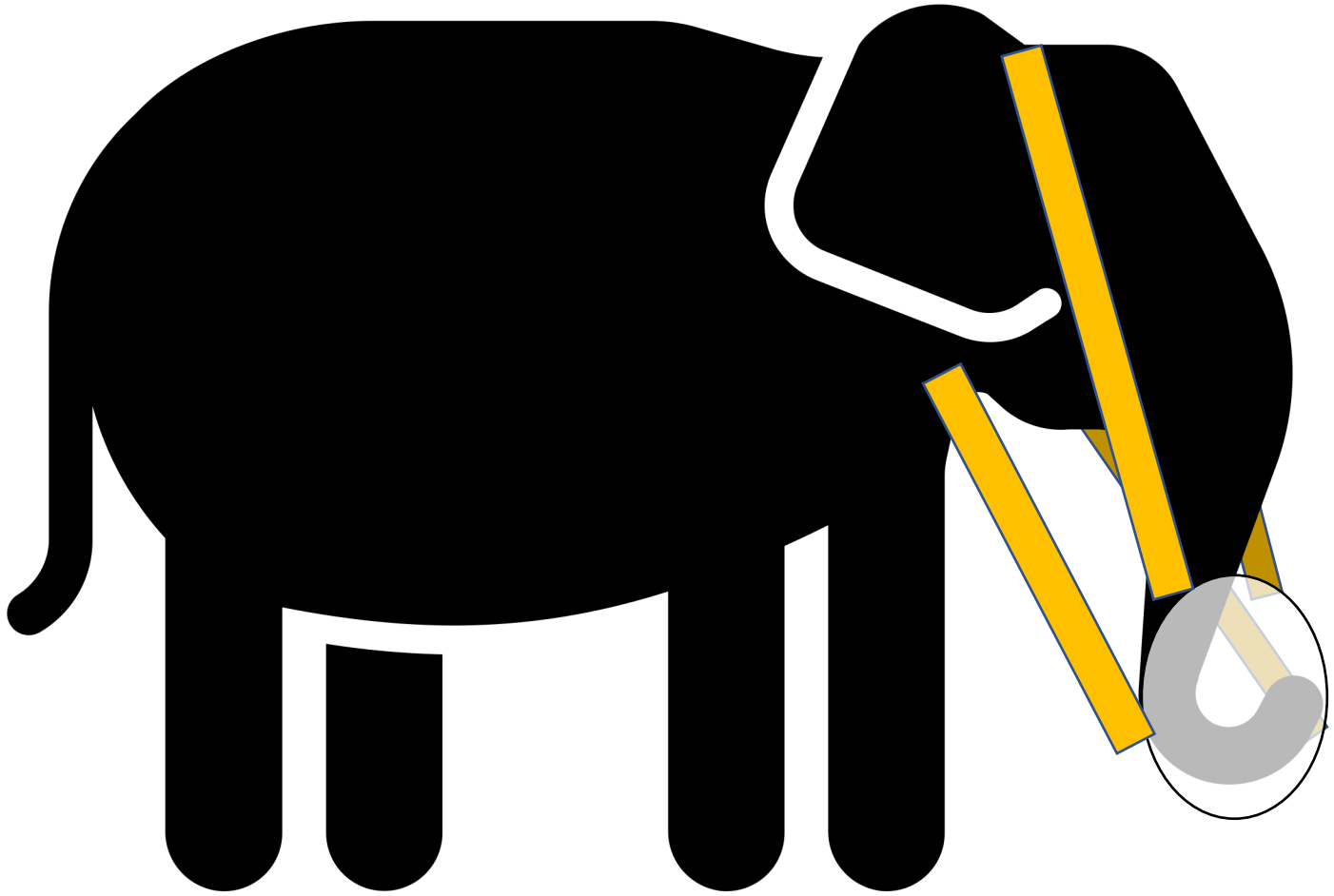
Welcome to CME 250 Introduction to Machine Learning!

Spring 2020 – Online version

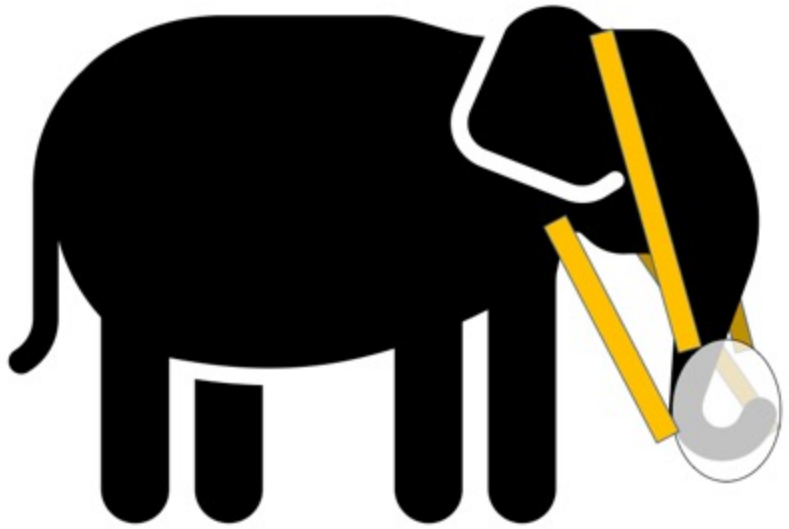
Today's schedule

- Getting to know each other
- What is Machine Learning?
- Class expectations
- Course Logistics
- Exploratory Data Analysis

First, the elephant in the room…



First, the elephant in the room…



COVID-19 emergency

- Synchronous vs. Asynchronous
- Time Zones
- Responsibilities and conditions at home

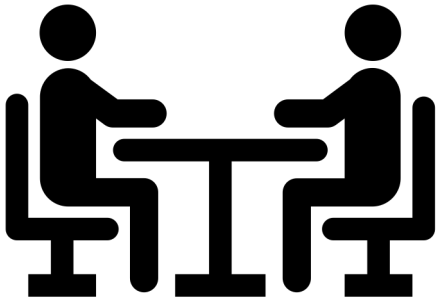
Thank you!

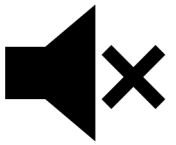
New experiment together

Zoom etiquette

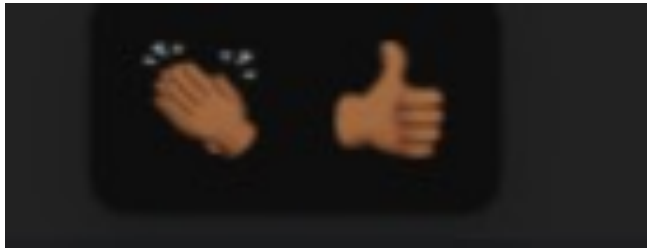
Zoom Reactions

Chat



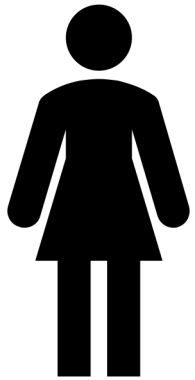
Sound 

Video 

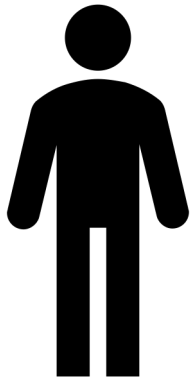


Let's get to know each other…

Breakout room



You



Fellow student

- Name
- Location
- Department
- Year
- Why Intro to ML
- Nickname + Story

3 mins

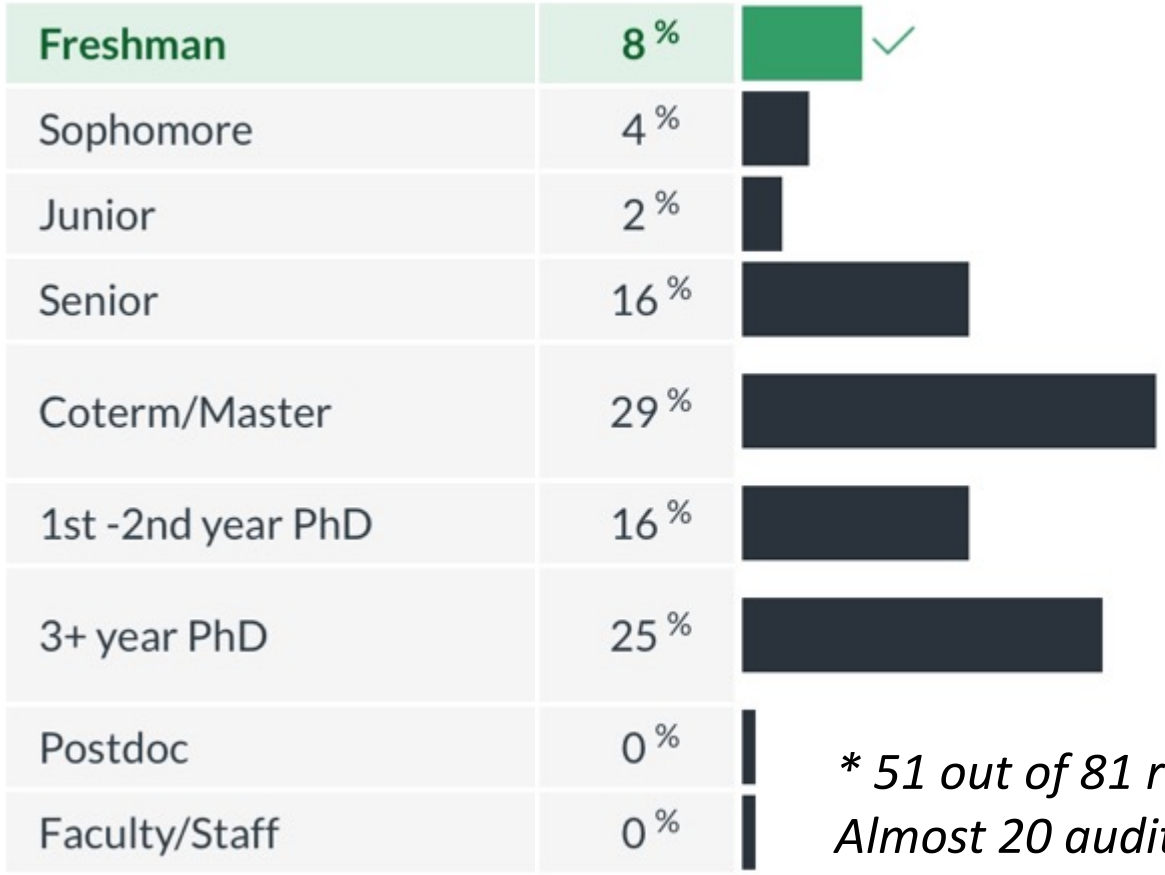
Chat/Audio/Video

About the survey

What is your experience with machine learning, linear algebra and statistics?

I do not feel comfortable with linear algebra and statistics	**4 %**
I know basic linear algebra and statistics but I have never worked on machine learning	71 %
I know some algorithms used on machine learning, but I do not know how to apply them	16 %
I have use standard machine learning tools to solve small projects, and I want to learn more about technical details and math	8 %
I have used machine learning on the past and I want a refresh/update in the methods used nowadays	2 %

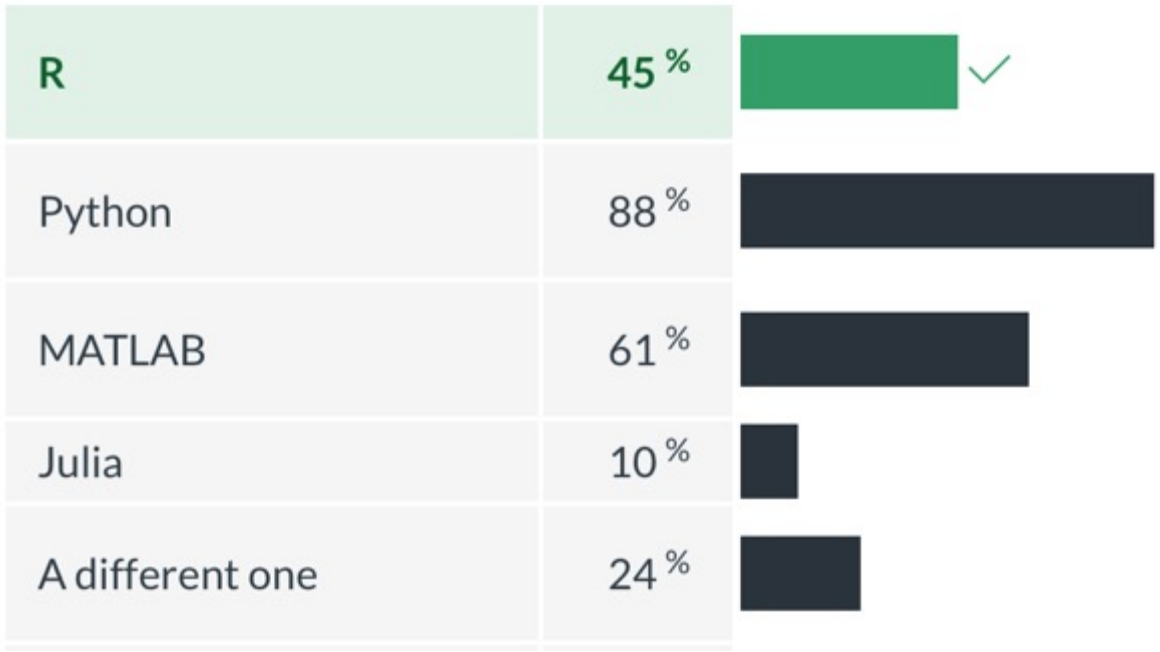
About the survey



** 51 out of 81 registered students*
Almost 20 auditing students, between senior PhDs, Postdocs and Staff

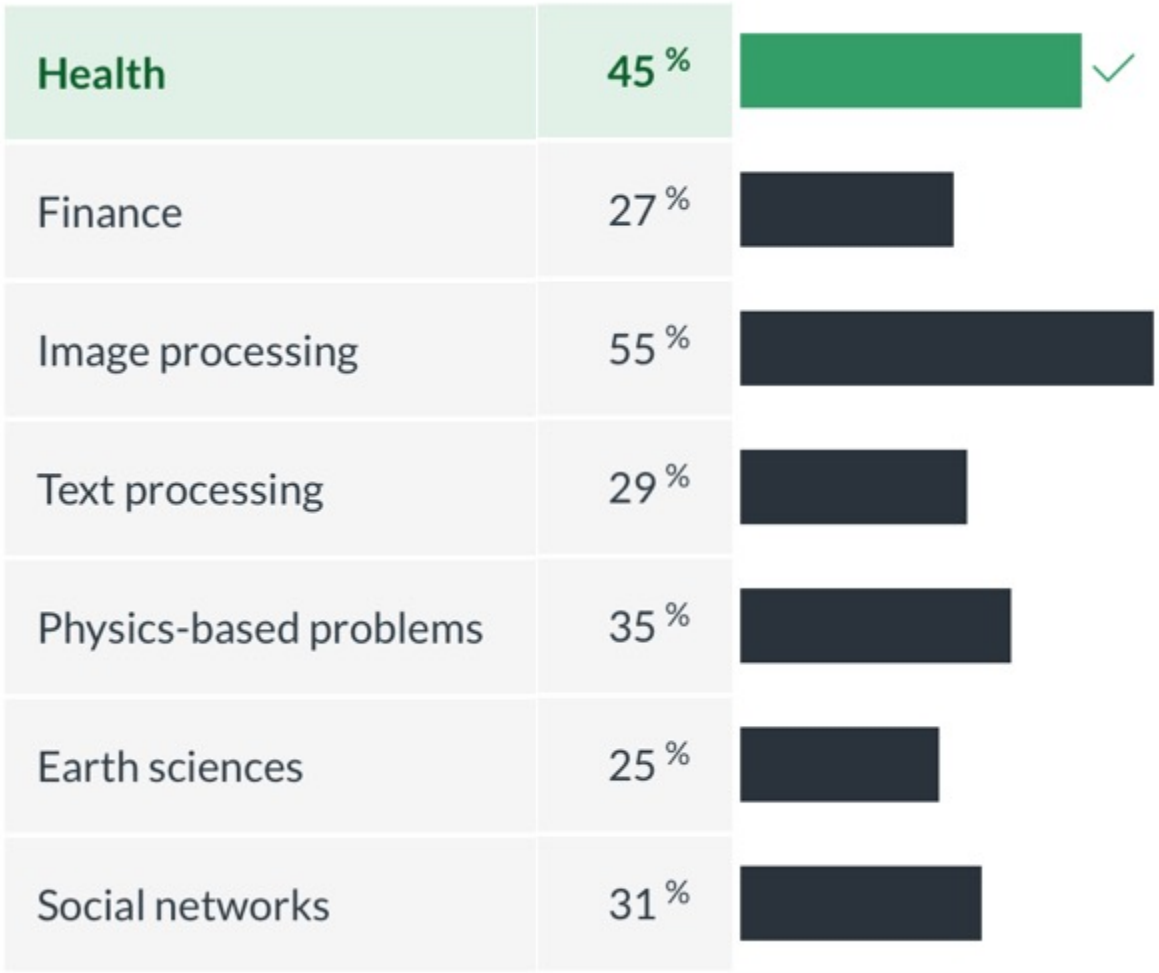
About the survey

Programing language



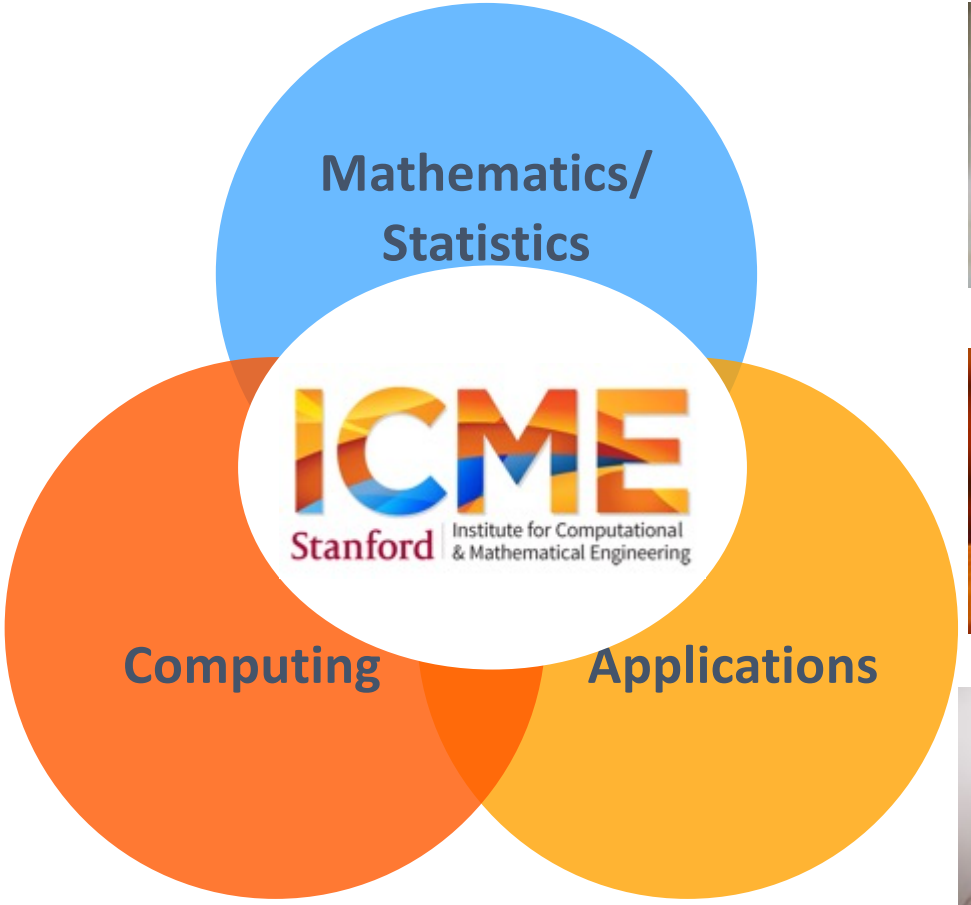
About the survey

Application

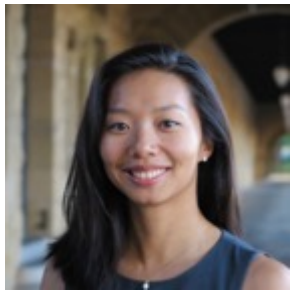


About myself …

Cindy Orozco (orozcocc@stanford.edu)



https://icme.stanford.edu/



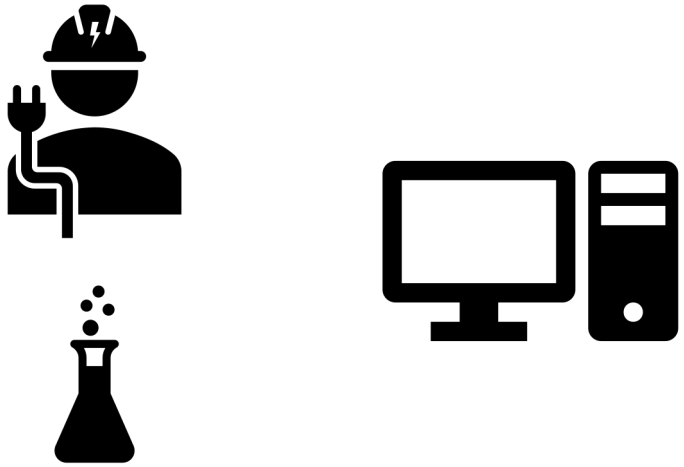
Sherrie Wang

Alex Ioannidis

Karianne Bergen

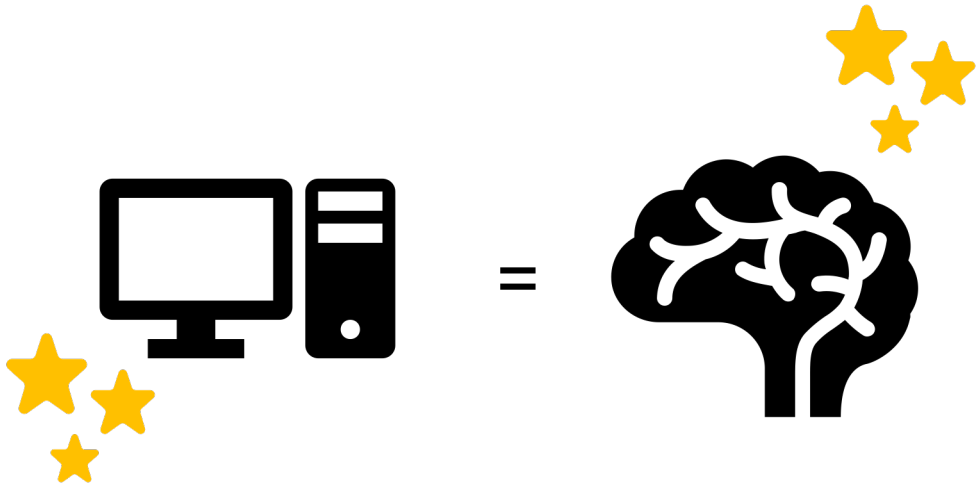
In 2015 …

Numerical Analysis



Theory -> Algorithms

Machine Learning



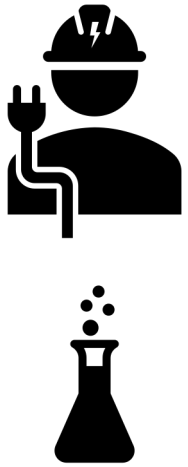
Algorithms -> Human Behavior

In 2020 …

Theory

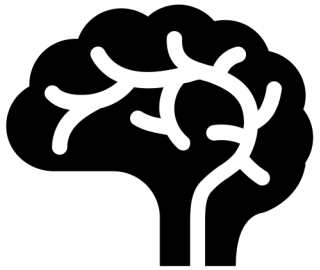
Numerical Analysis

Machine Learning



My Research

=



Theory -> Algorithms

Modeling

Algorithms -> Human Behavior

What is machine learning?

"Field of study that gives computers the ability to learn without being explicitly programmed"

Arthur Samuel (1959)

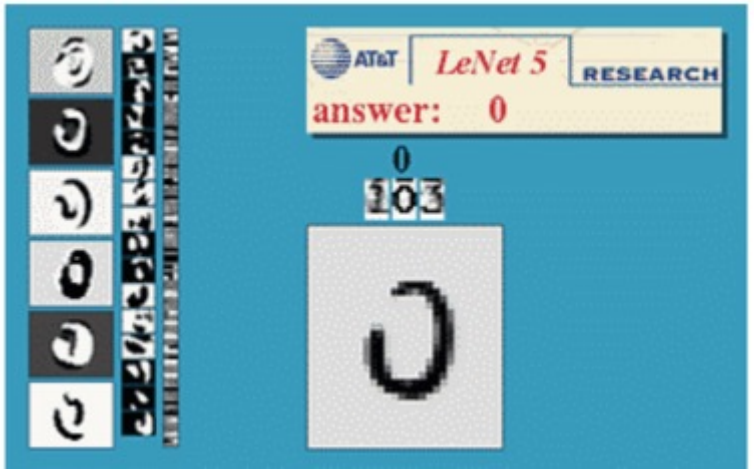
How is machine learning used?

Anomaly Detection

How is machine learning used?

Recommendation Systems

How is machine learning used?



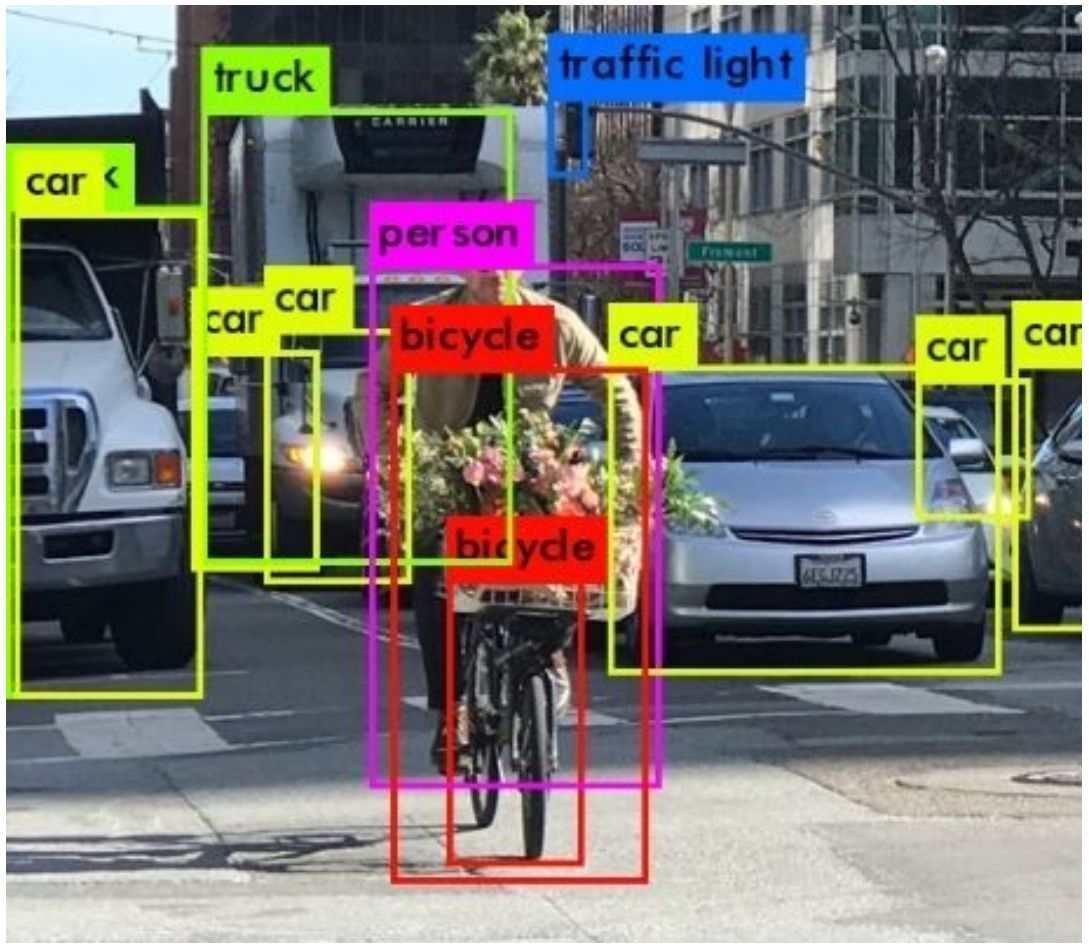
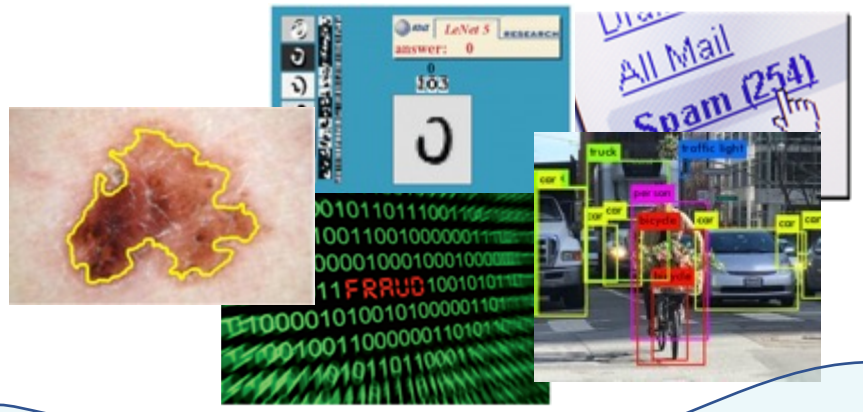


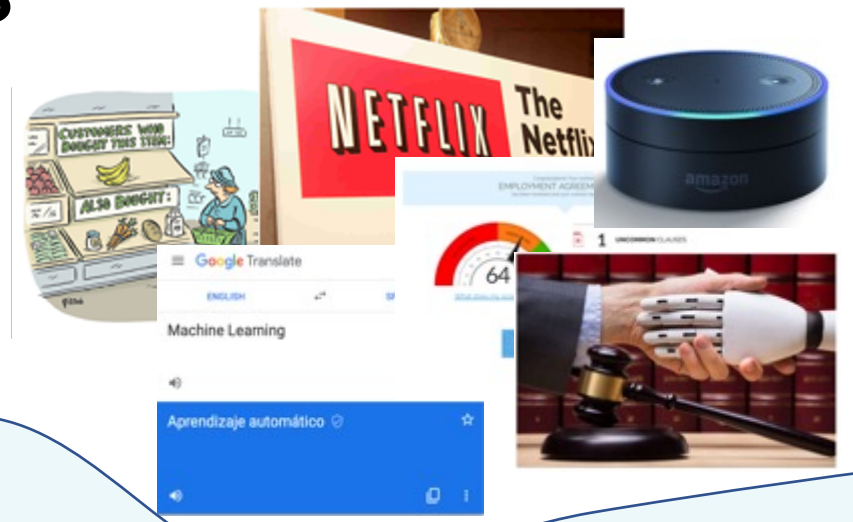
Image recognition

How is machine learning used?

Natural Language Processing

This is just the tip of the iceberg …





What is machine learning?

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Tom M. Mitchell (1997)

Experience

Computer Program

Performance Measure

Task

Stage 0: Finding (useful) Data

Experience

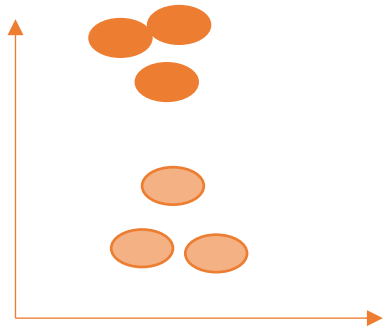
Data

Task

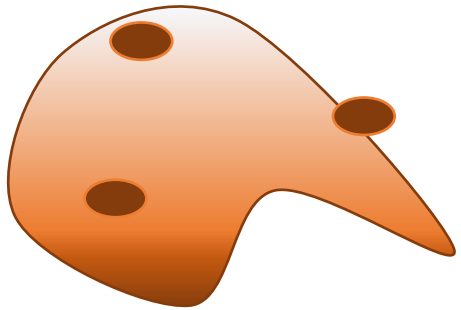
Problem

Stage 1: Data Exploration

Data

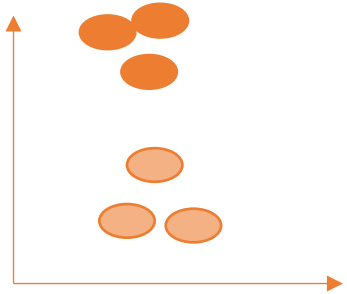


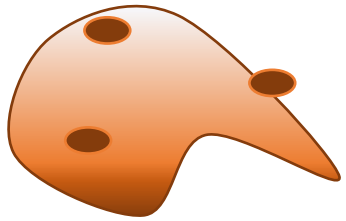
Problem



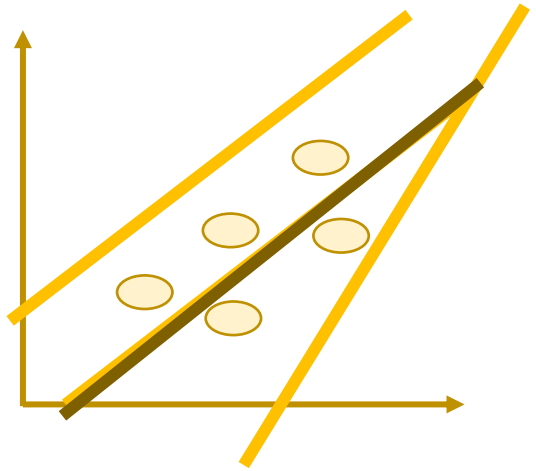
What are the properties of the data?

Stage 2: Prediction Models



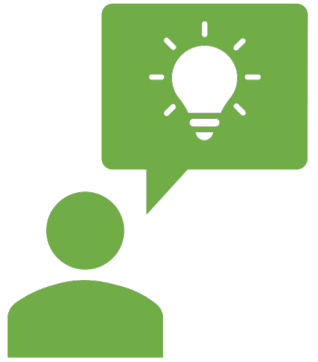


Data Exploration



$$y = a\,x + b$$

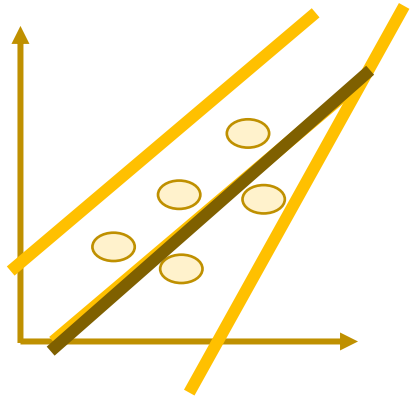
Find optimal parameters



Problem

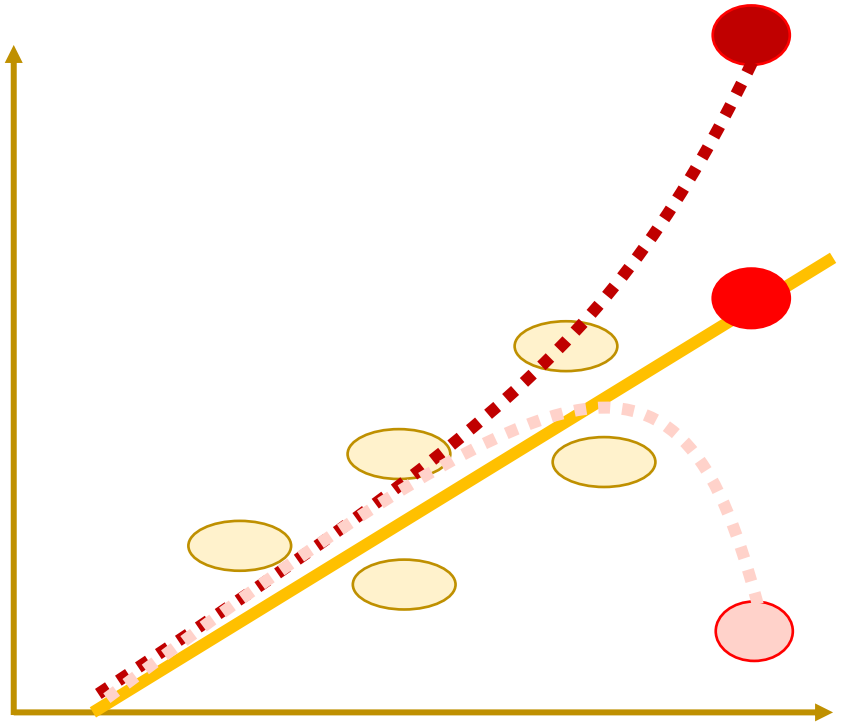
Given x … y?

Stage 3: Performance Analysis



$y = a\,x + b$

Model Creation



Model selection for unseen data

Problem

Given x … y?

What is machine learning?

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Tom M. Mitchell (1997)

Experience

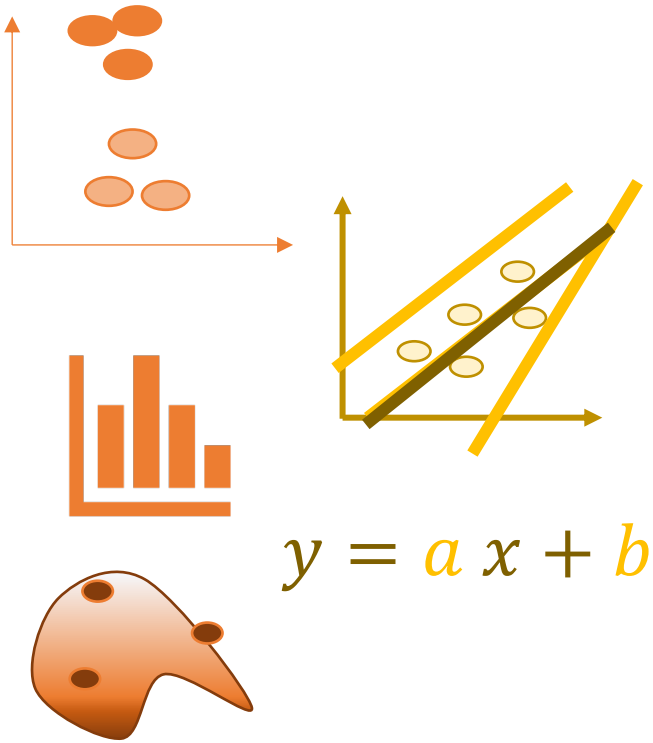
Computer Program

Performance Measure

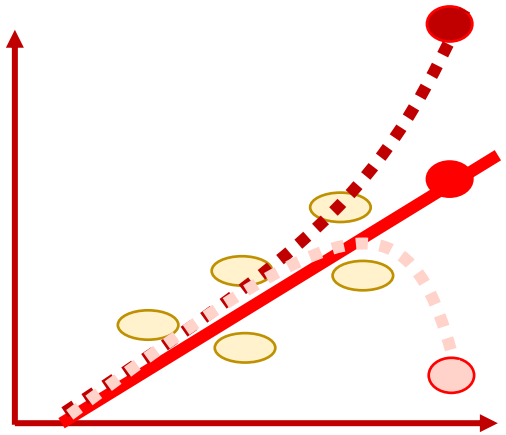
Task

What is machine learning?

Experience



Computer Program

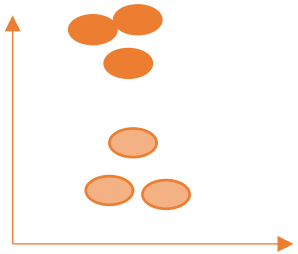


Performance Measure

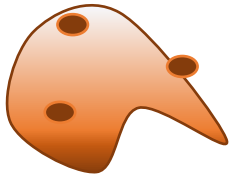
Task

Class Schedule

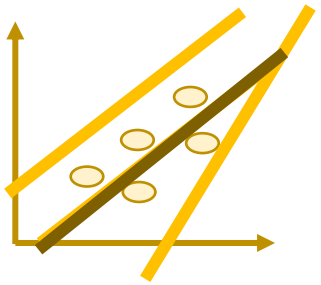
April 9: Exploratory Data Analysis



April 14: Clustering



April 16: Dimensionality Reduction

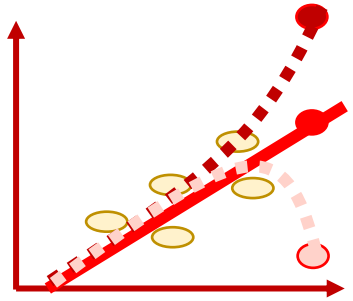


April 21: Regression Models

April 28: Support Vector Machines

April 30: Trees and Random Forest

May 5: Neural Networks



April 23: Model Evaluation, Regularization

Class expectations

You

Machine Learning

Class expectations

Terminology, Models
Best Practices.

Mathematical proofs.
Implementation tricks.

Class expectations

Introduction	Foundations	Theory	Applications	
CME 250: Introduction to Machine Learning	CS 229: Machine Learning	CS 229T: Statistical Learning Theory	CS 224N: Natural Language Processing with Deep Learning	CS 325B: Data for Sustainable Development
CS 229A: Applied Machine Learning	CS 221: Artificial Intelligence	STATS 315A/B: Modern Applied Statistics	CS 231N: Convolutional Neural Networks for Visual Recognition	CS 273B: Deep Learning in Genomics and Biomedicine
	CS 230: Deep Learning	CS 234: Reinforcement Learning	CS 246: Mining Massive Data Sets	…and much more

Mathematical proofs.
Implementation tricks.

Class expectations

Terminology, Models
Best Practices.

+

Working with a project

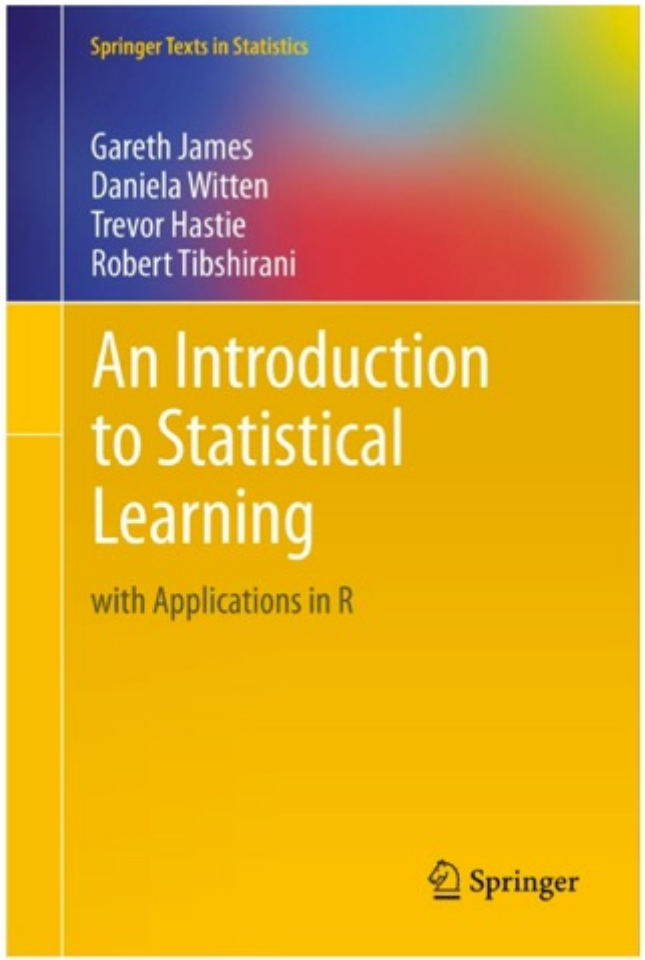
Prerequisites

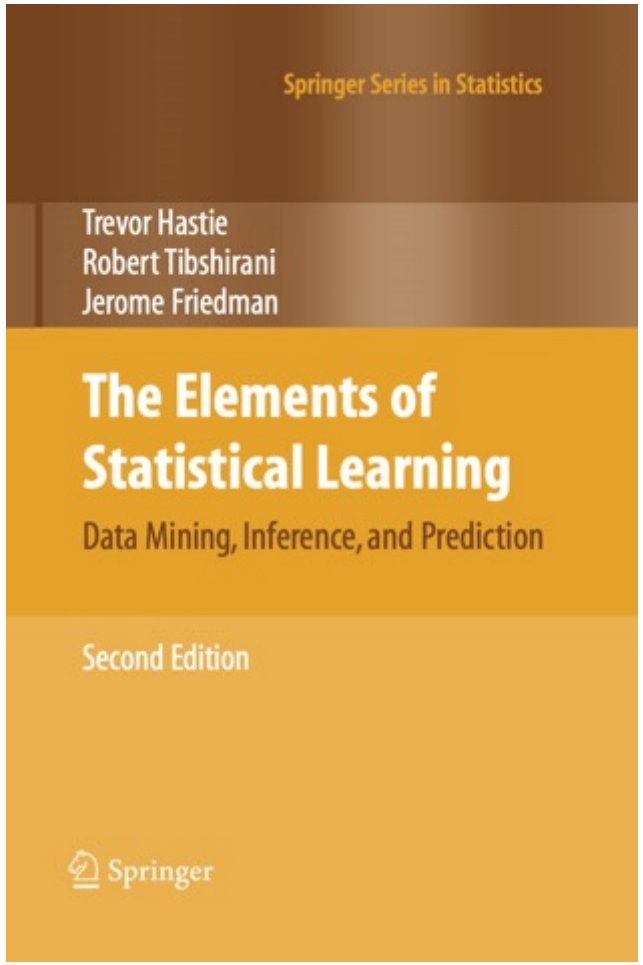
Statistics

Linear Algebra

Programming: R, Python, MATLAB

Tools

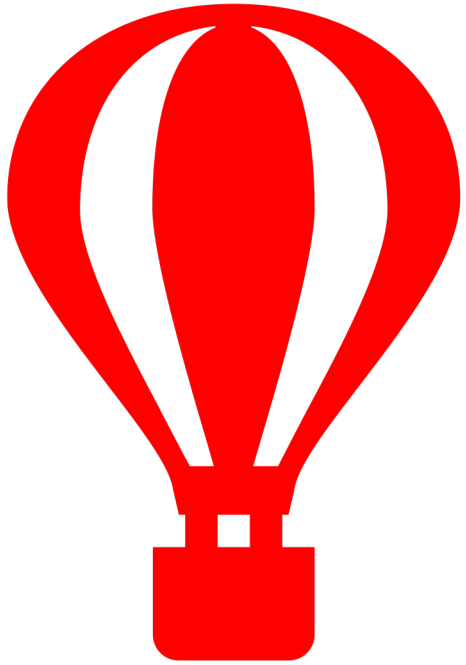

Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

Tools

	Python	R	MATLAB
Desktop	ANACONDA.	RStudio	
Browser	Google Colaboratory	RStudio Cloud	
Server	Sherlock / Farmshare		

Class Logistics



Live lectures on Zoom T-TH 9-10:20 am

Recordings available after class

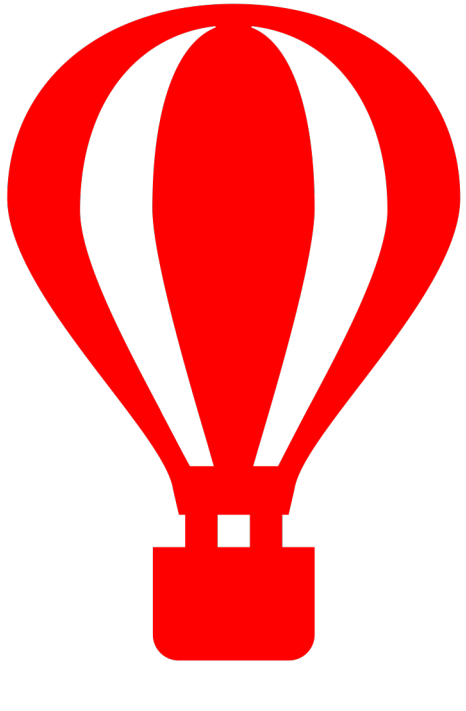
Slides + Notes on Canvas.

Questions? Discussion boards

More questions? Office hours

Feedback: We will be adapting during the quarter

Office Hours



	M	T	W	TH	F	S	S
9							
10							
11							
12							
1							
2							
3							
4							
5							
6							
7							
8							

Place a stamp where at the time that work best for you

Project

Part I: Data Exploration. April 26

Formulate Question

Find the data

EDA + Unsupervised Learning

Part I: Peer Review. May 3

Part II: Prediction & Selection. May 17

Refine Question

Construct 2 models (or more)

Model Selection

Part II: Peer Review. May 25

Project

Groups
1 – 2 People (Use Canvas)

Anonymous Peer Review

Deliverables
Report explaining process / conclusions
Code/Notebook reporting calculations

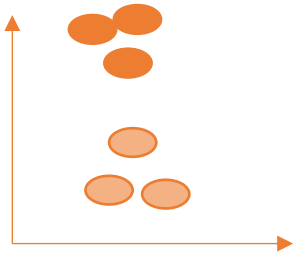
Complete Part I & Part II + Peer Review to get credit for the course

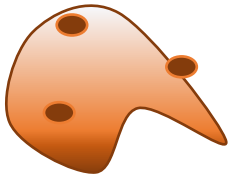
Consider the scope of your question, and the timeline.

Other short courses offered by ICME in Spring

- CME 193: Introduction to Python
- CME 195: Introduction to R
- CME 197: Human-Centered Design methods for Data Science

Keep an eye on Data Science Summer Workshops





$$y = a\,x + b$$

Experience

Data Exploration

Prediction Models

Performance Analysis

Task

Exploratory Data Analysis

based on

Modern Statistics for Modern Biology

Chapter 3 High Quality Graphics in R

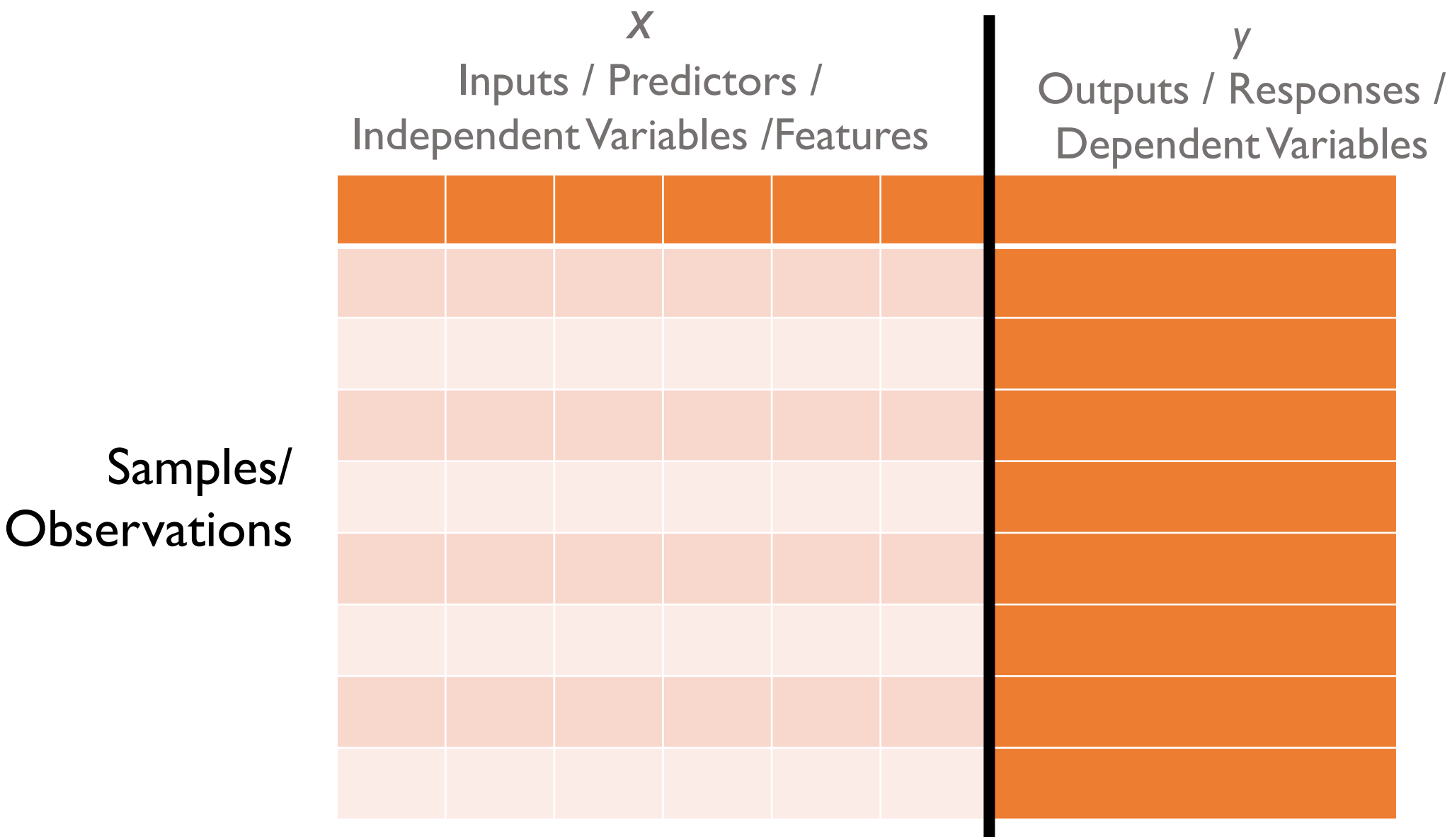
Susan Holmes, Wolfgang Huber

http://web.stanford.edu/class/bios221/book/Chap-Graphics.html

Data Exploration

How is the *typical* data?

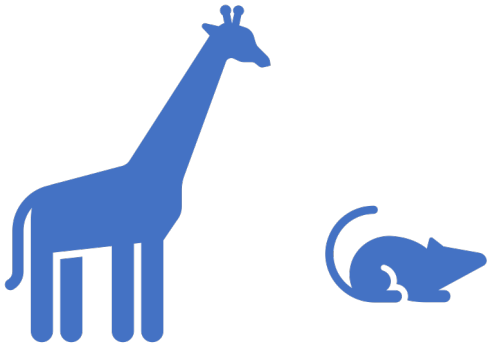
Attributes / Variables

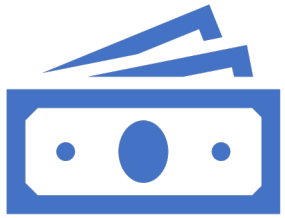


How is the *typical* data?

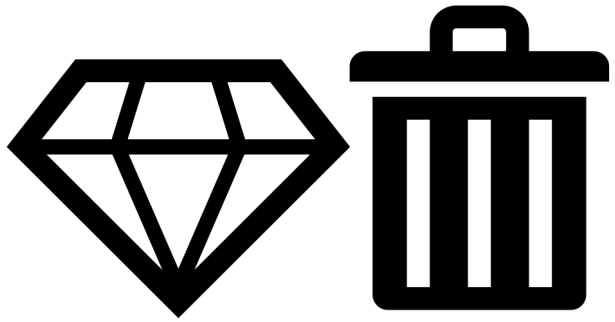
Variables

Quantitative	Qualitative	Ordered Categorical





Exploratory Data Analysis



Data Quality

Quick Insights

Summaries

Pandas - Python

tidyverse -R

Visualization

Seaborn, Plotly … - Python

ggplot -R

Summaries Pandas - Python tidyverse -R

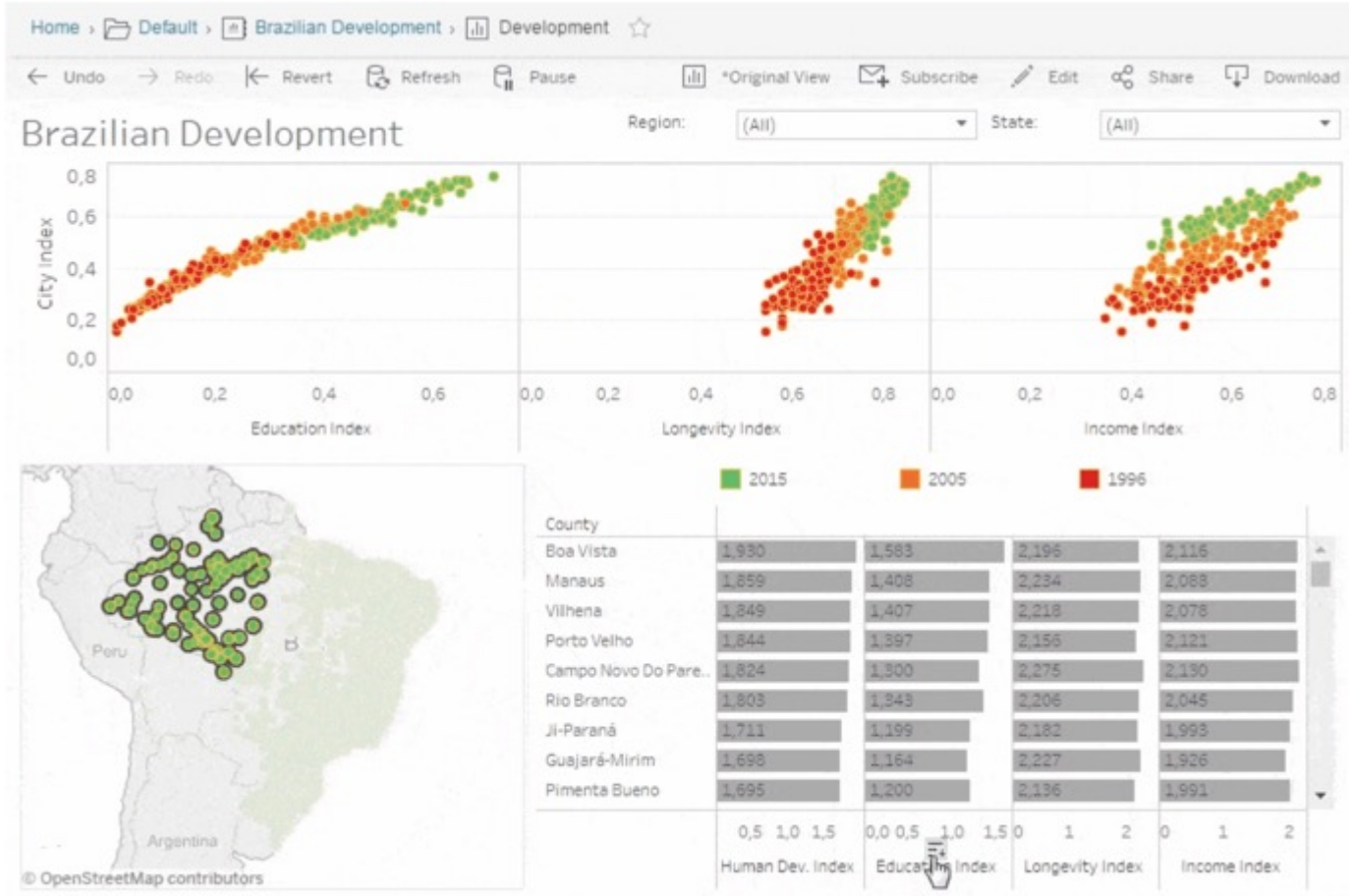
	mtcars2 (N = 32)	cyl_factor: 6 cylinders (N = 7)	cyl_factor: 4 cylinders (N = 11)	cyl_factor: 8 cylinders (N = 14)
Miles Per Gallon				
min	10.4	17.8	21.4	10.4
max	33.9	21.4	33.9	19.2
mean (sd)	20.09 ± 6.03	19.74 ± 1.45	26.66 ± 4.51	15.10 ± 2.56
Displacement				
min	71.1	145.0	71.1	275.8
median	196.3	167.6	108.0	350.5
max	472	258.0	146.7	472.0
mean (sd)	230.72 ± 123.94	183.31 ± 41.56	105.14 ± 26.87	353.10 ± 67.77
Weight (1000 lbs)				
min	1.513	2.620	1.513	3.170
max	5.424	3.460	3.190	5.424
mean (sd)	3.22 ± 0.98	3.12 ± 0.36	2.29 ± 0.57	4.00 ± 0.76
Forward Gears				
Three	15 (47)	2 (29)	1 (9)	12 (86)
Four	12 (38)	4 (57)	8 (73)	0 (0)
Five	5 (16)	1 (14)	2 (18)	2 (14)

https://cran.r-project.org/web/packages/qwraps2/vignettes/summary-statistics.html

Visualization

The most important skill of a data scientist

Storytelling



Visualization

a) Plots involving
- 1 variable
- 2 variables

b) Tips to keep in mind

Visualization: 1 variable

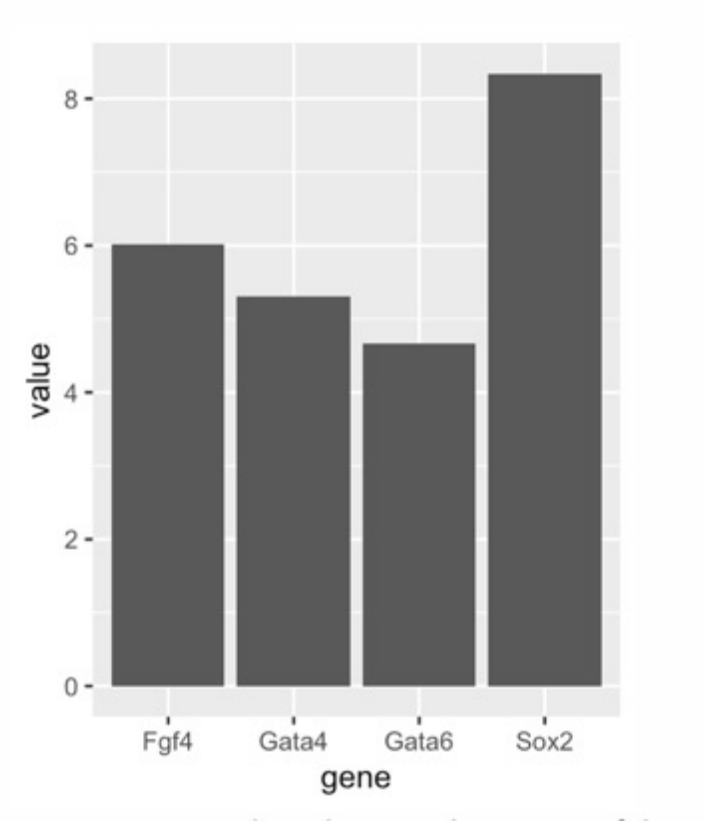
Intensity measures of 4 genes: Fgf4, Gata4, Gata6 and Sox2

Fgf4	Gata4	Gata6	Sox2
	Intensities in [0,12.5]		

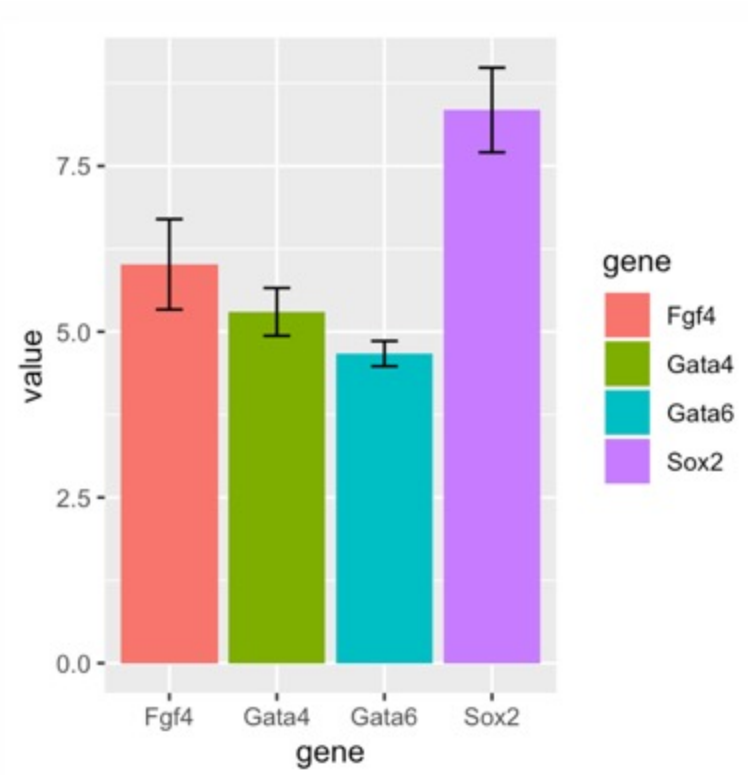
based on Holmes, Huber, *Modern Statistics for Modern Biology*

Visualization: 1 variable

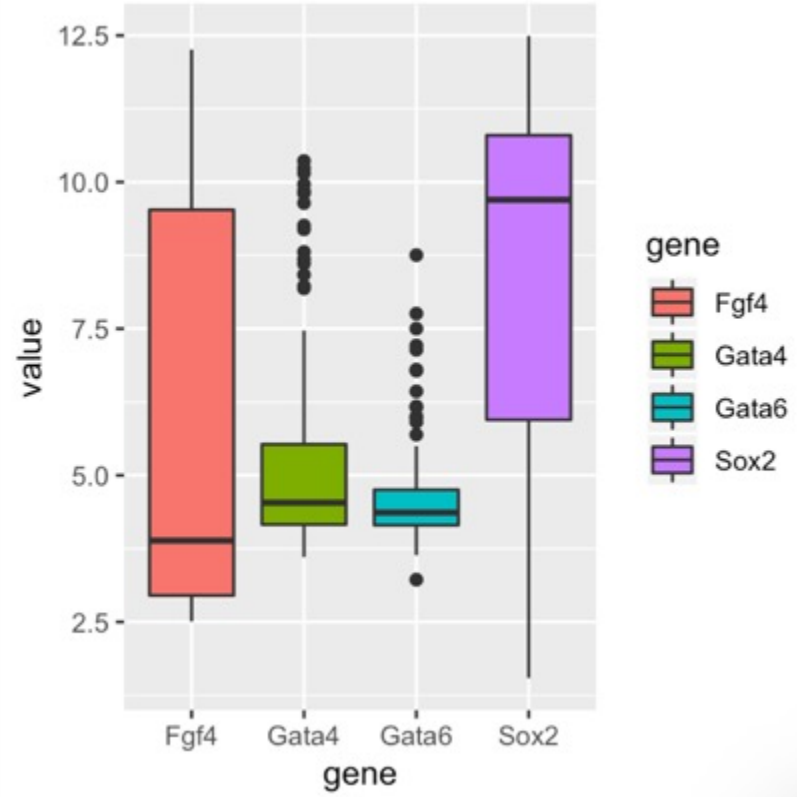
Intensity measures of 4 genes: Fgf4, Gata4, Gata6 and Sox2



Mean



Mean + Error bar

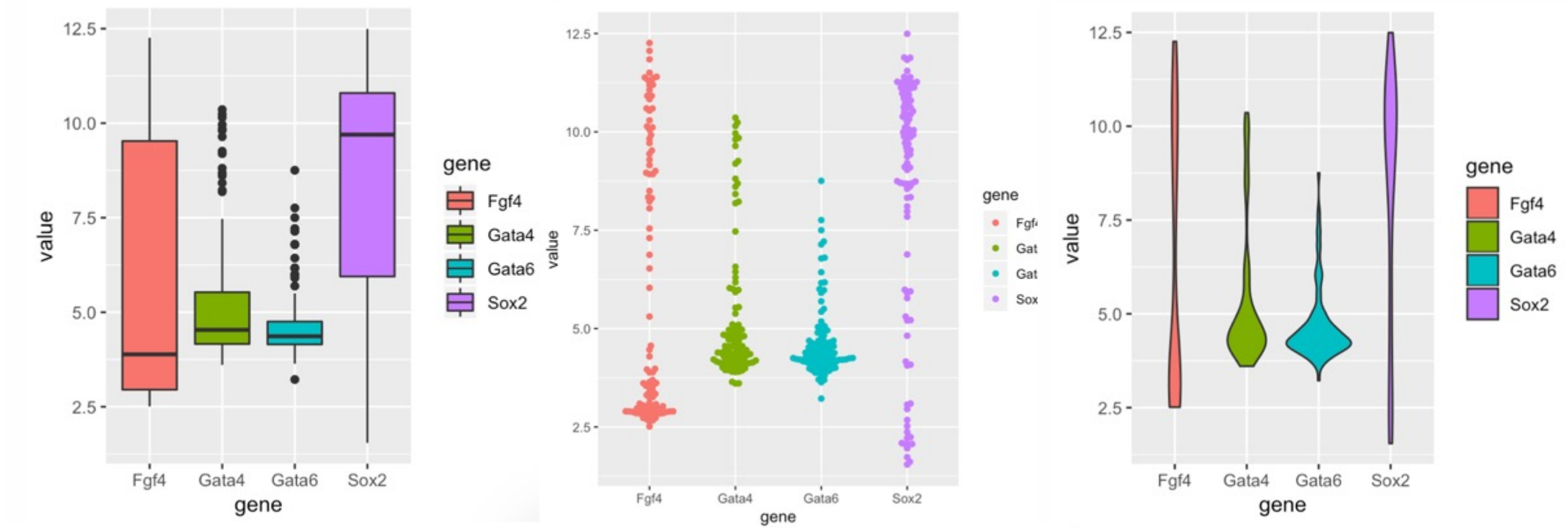


Boxplot

based on Holmes, Huber, *Modern Statistics for Modern Biology*

Visualization: 1 variable

Intensity measures of 4 genes: Fgf4, Gata4, Gata6 and Sox2



Boxplot

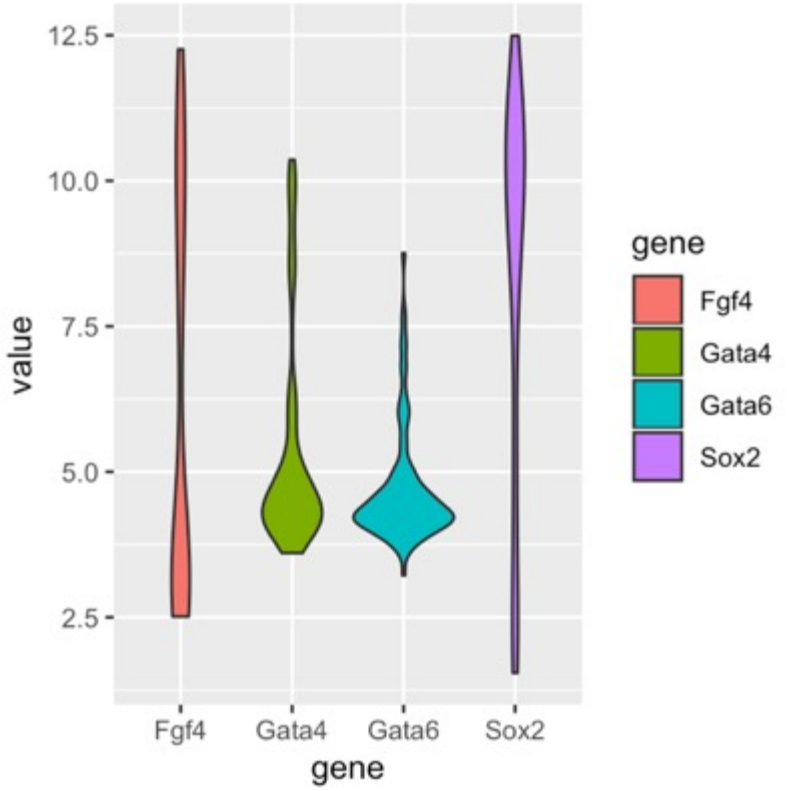
Beeswarm plot

Violin plot

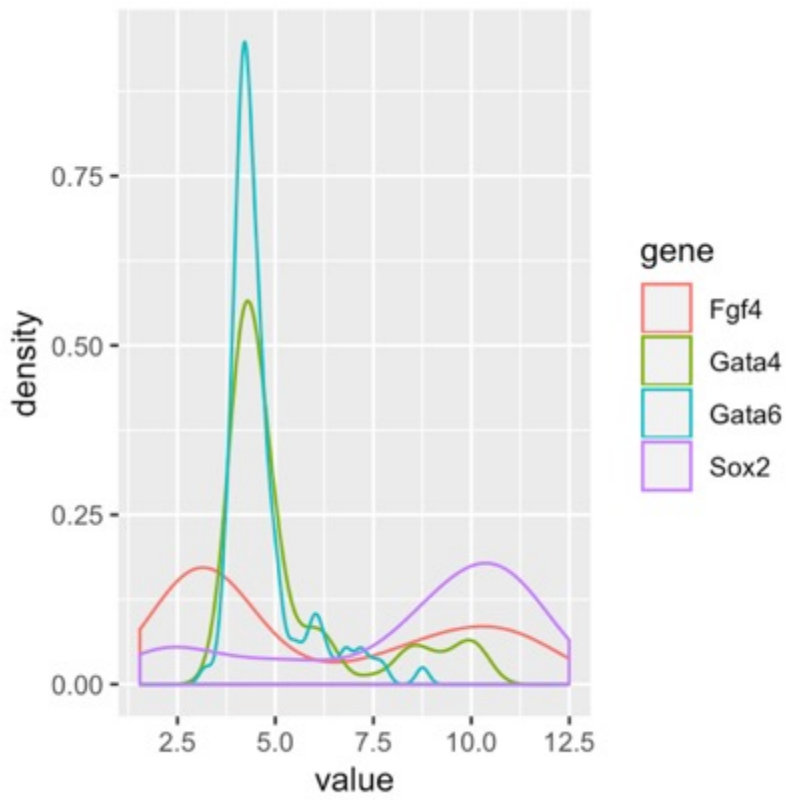
based on Holmes, Huber, *Modern Statistics for Modern Biology*

Visualization: 1 variable

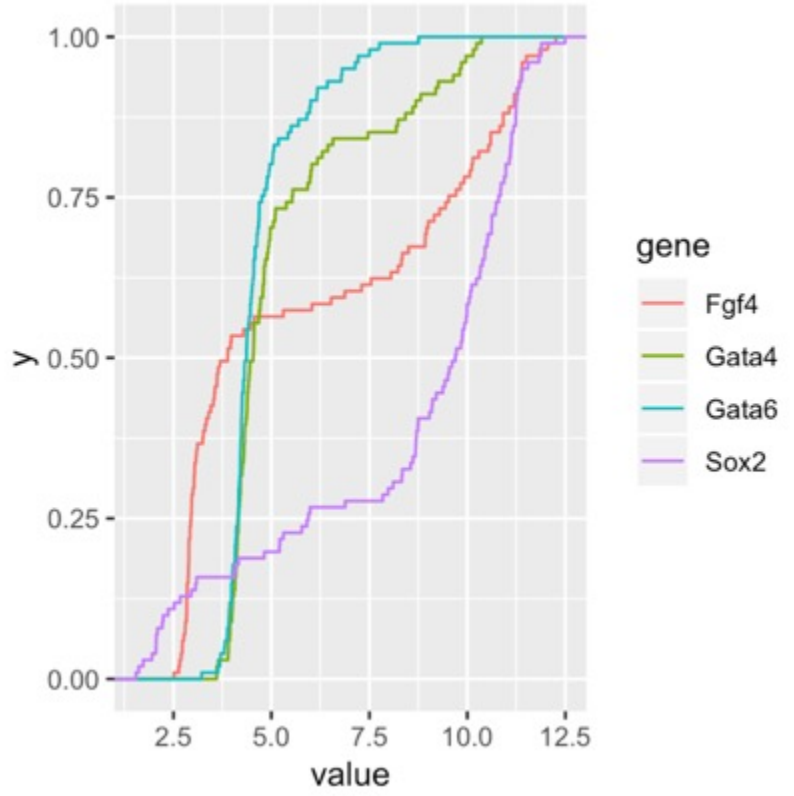
Intensity measures of 4 genes: Fgf4, Gata4, Gata6 and Sox2



Violin plot



Density plot

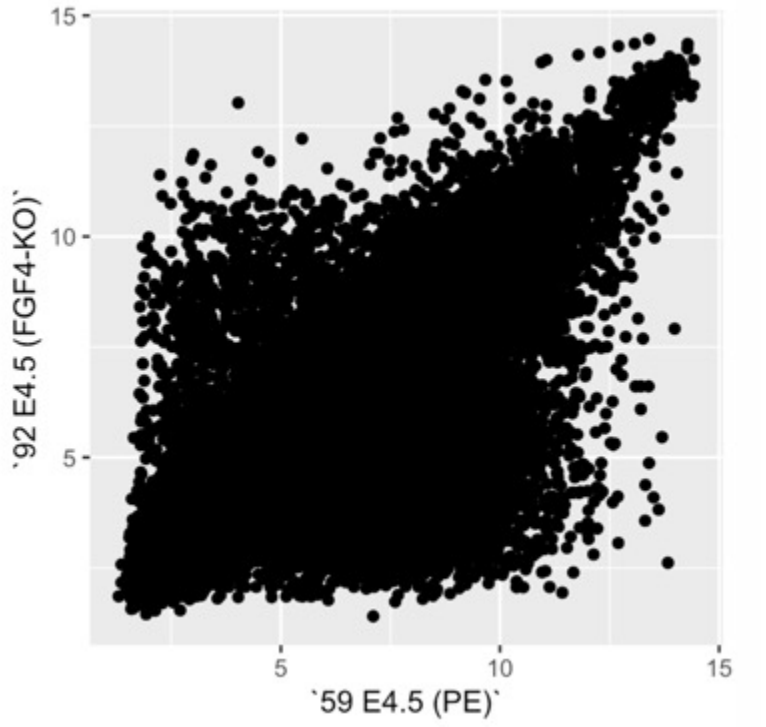


Empirical Cumulative Distribution

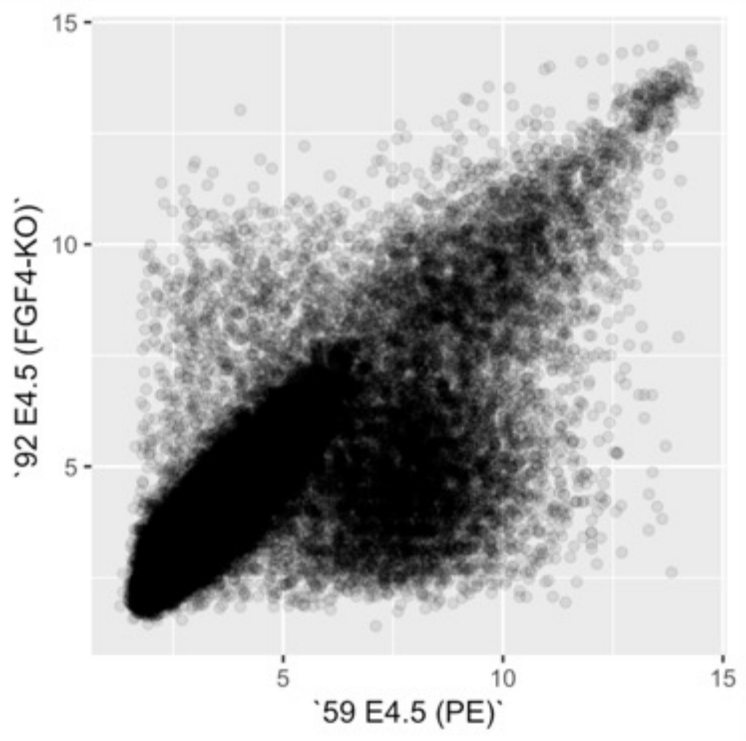
based on Holmes, Huber, *Modern Statistics for Modern Biology*

Visualization: 2 variables

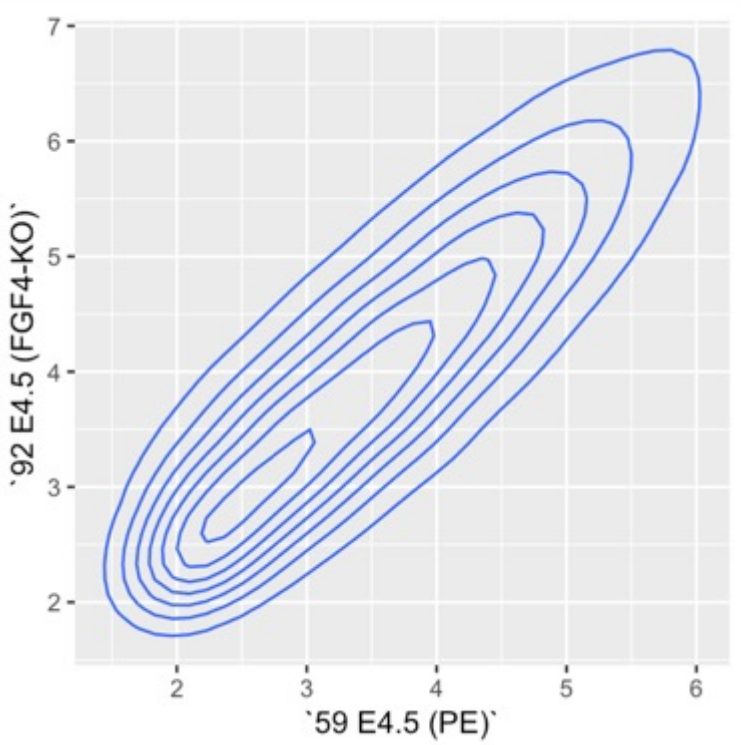
Measure with/without treatment



Scatterplot



Scatterplot + transparent

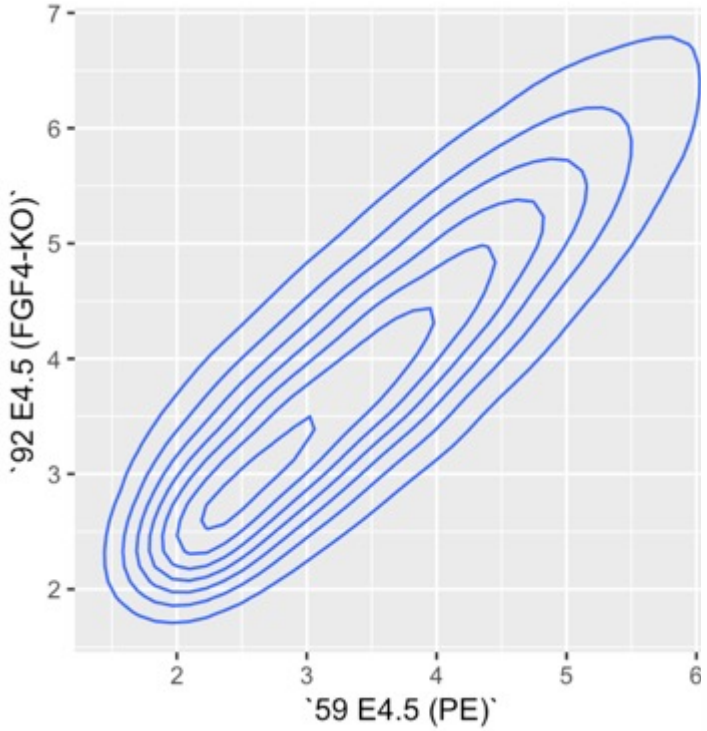


Contour plot

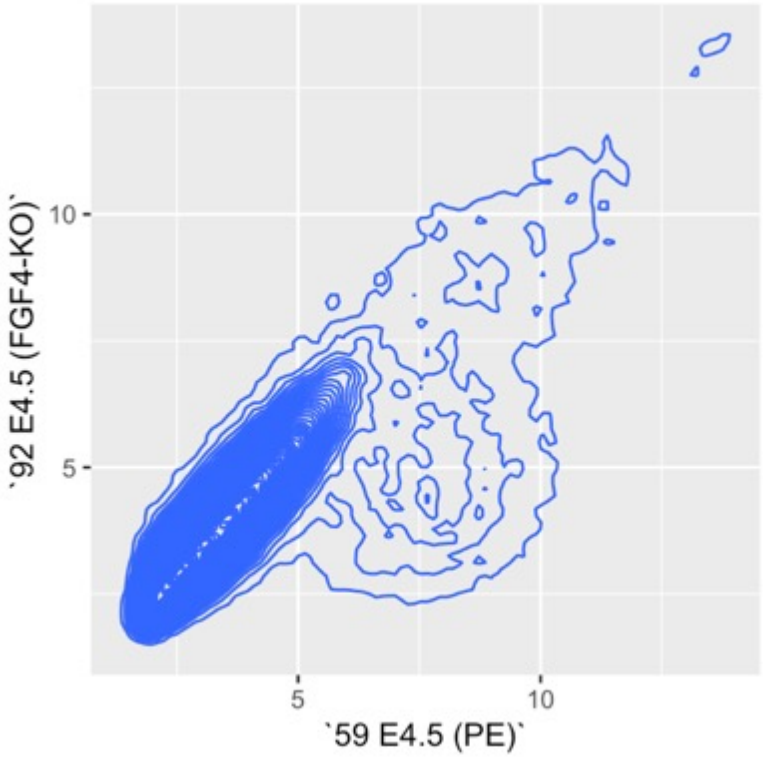
based on Holmes, Huber, *Modern Statistics for Modern Biology*

Visualization: 2 variables

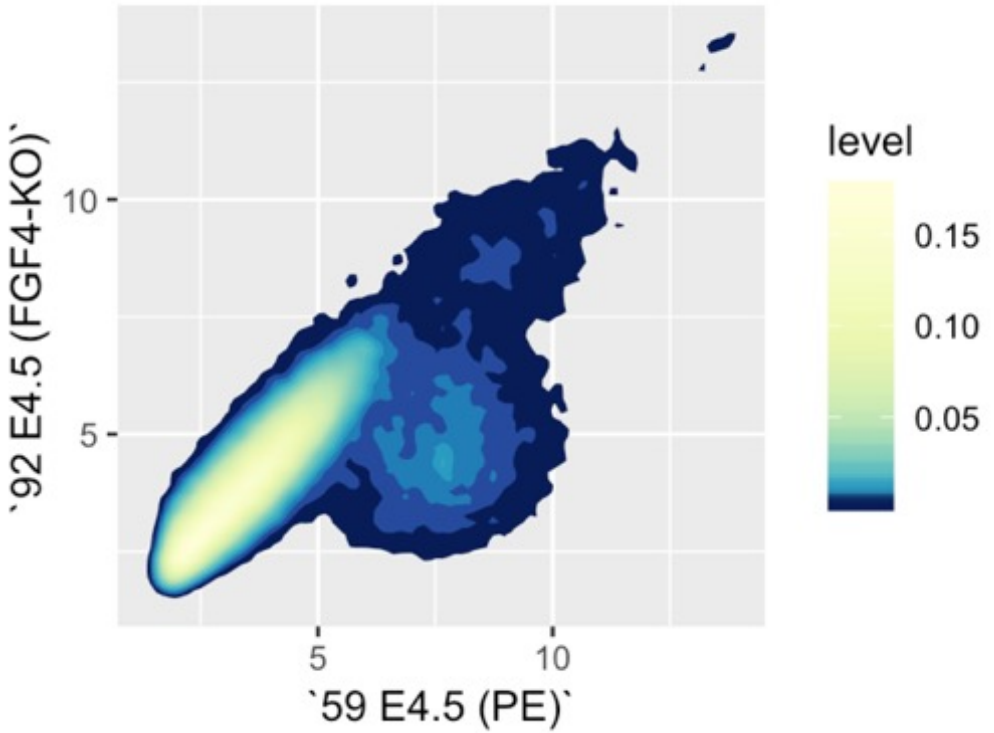
Measure with/without treatment



Contour plot



Contour plot +
more bins

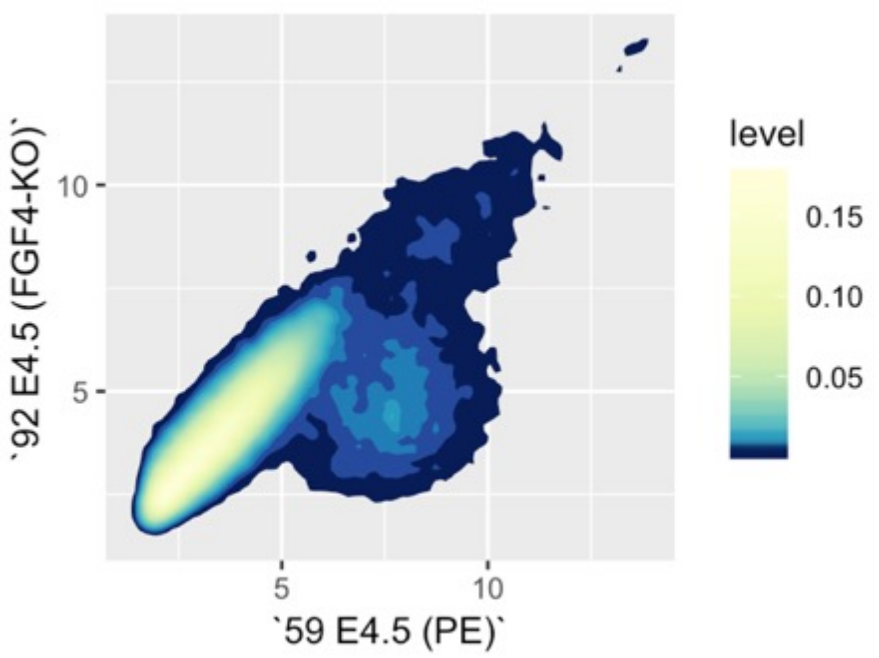


Contour plot +
color scheme

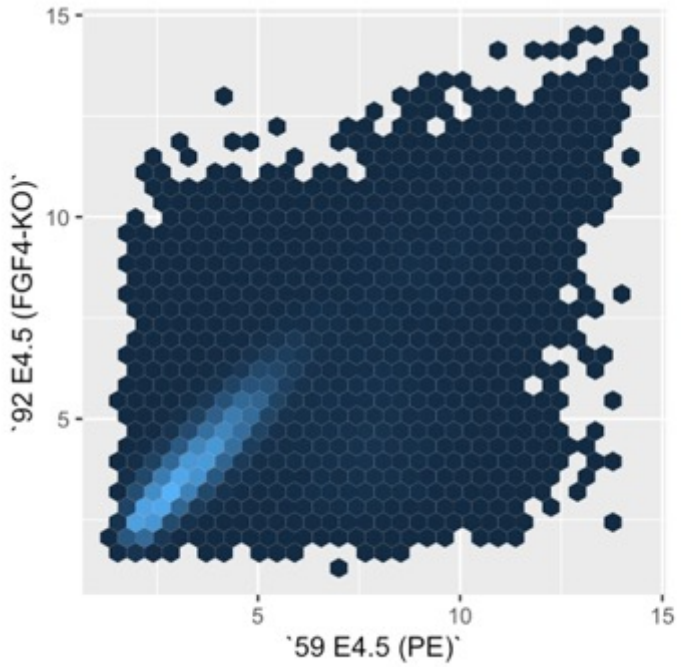
based on Holmes, Huber, *Modern Statistics for Modern Biology*

Visualization: 2 variables

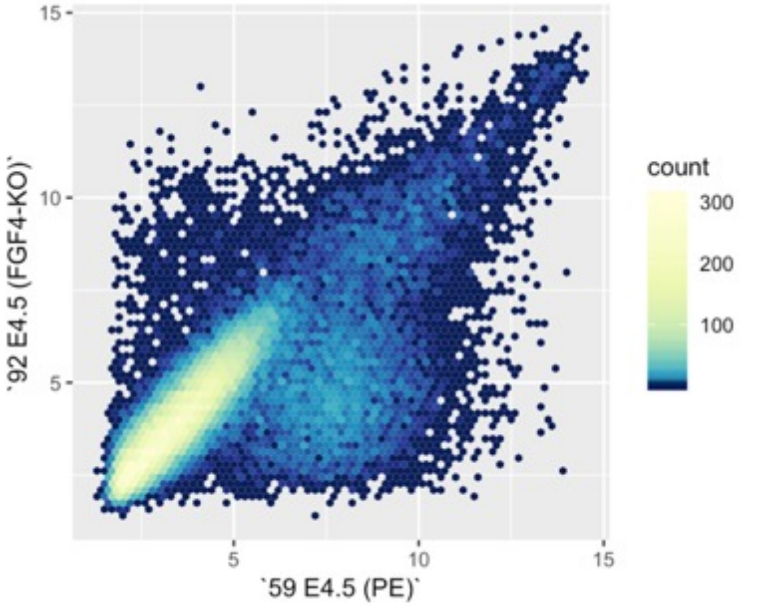
Measure with/without treatment



Contour plot +
color scheme



Hexagonal Binning

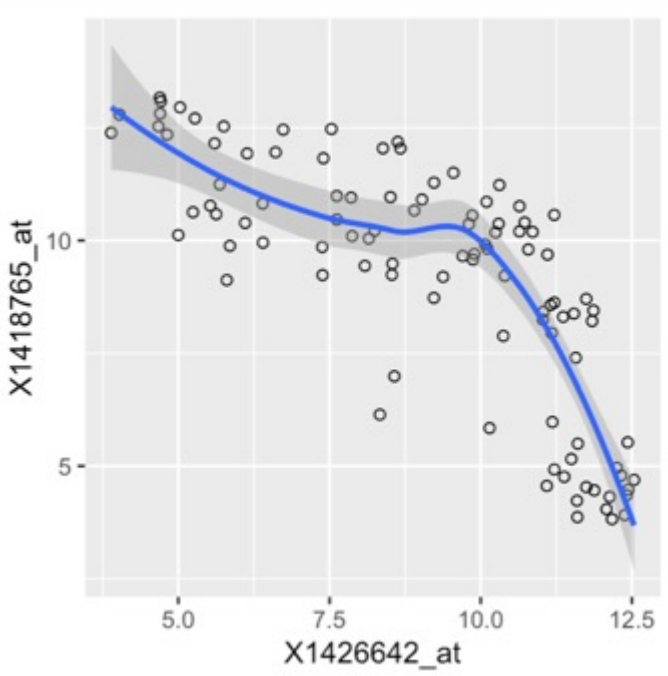


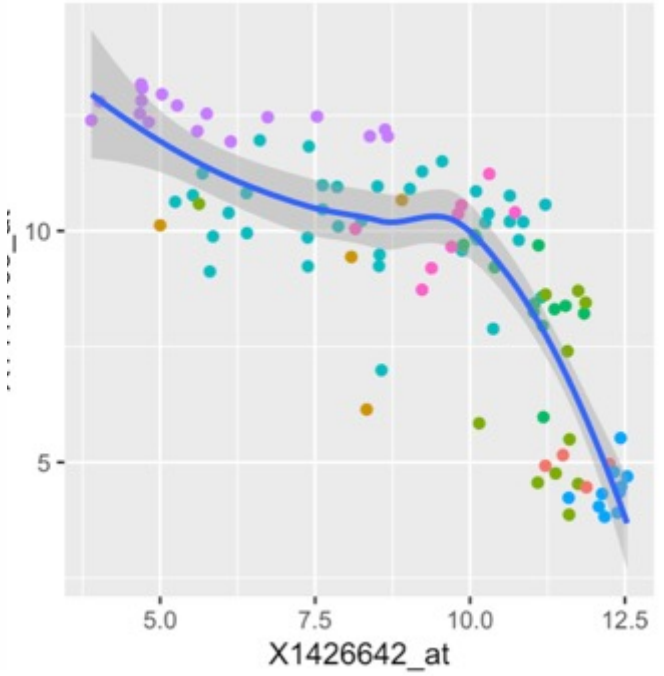
Hexagonal Binning +
Better color scheme

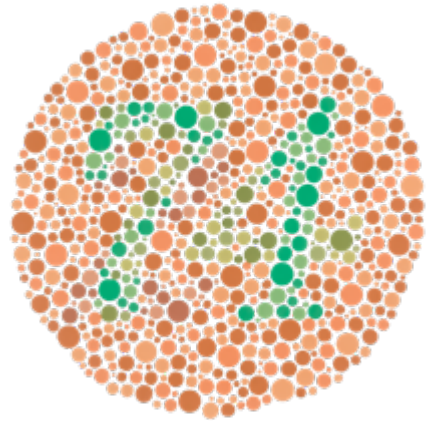
based on Holmes, Huber, *Modern Statistics for Modern Biology*

Tips to remember

Color



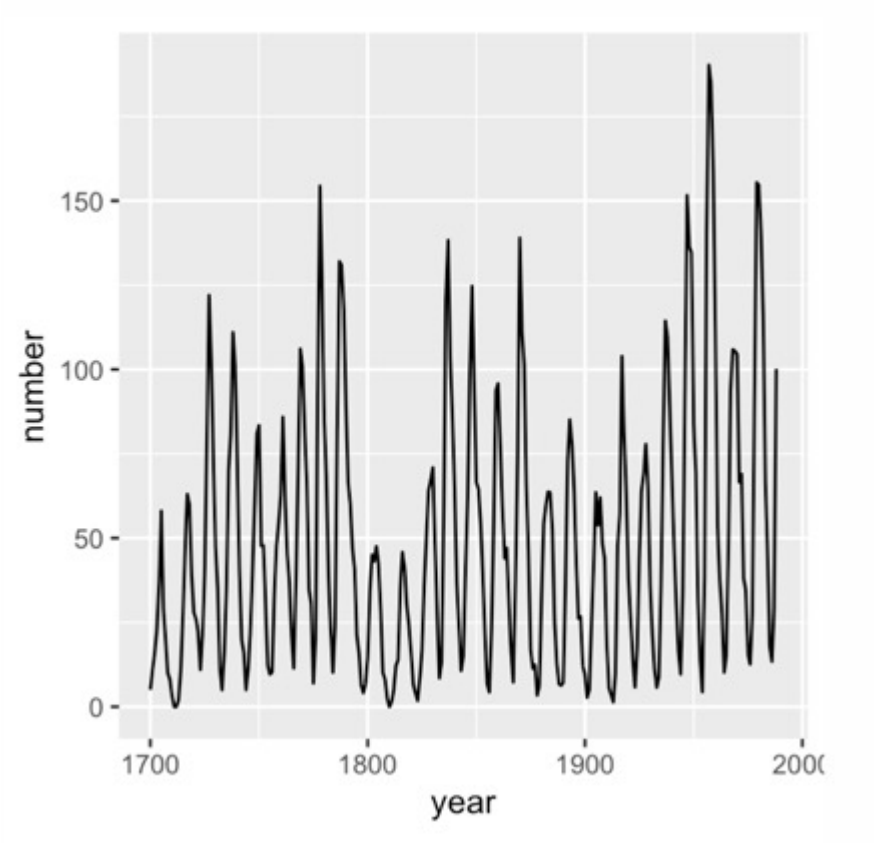


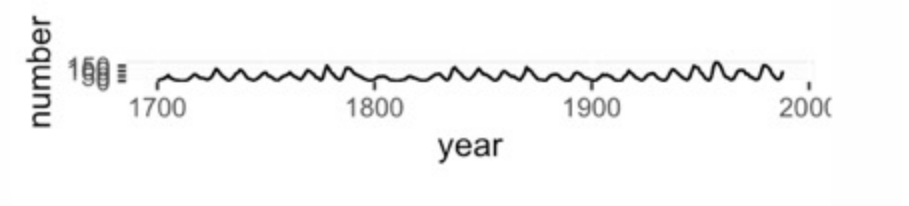


based on Holmes, Huber, *Modern Statistics for Modern Biology*

Tips to remember

Axes

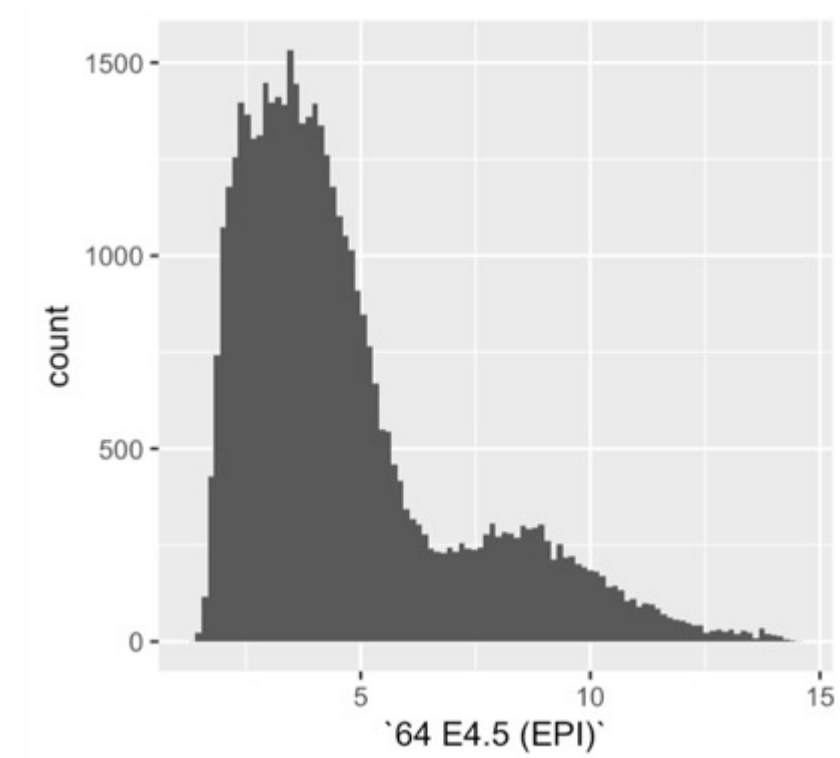




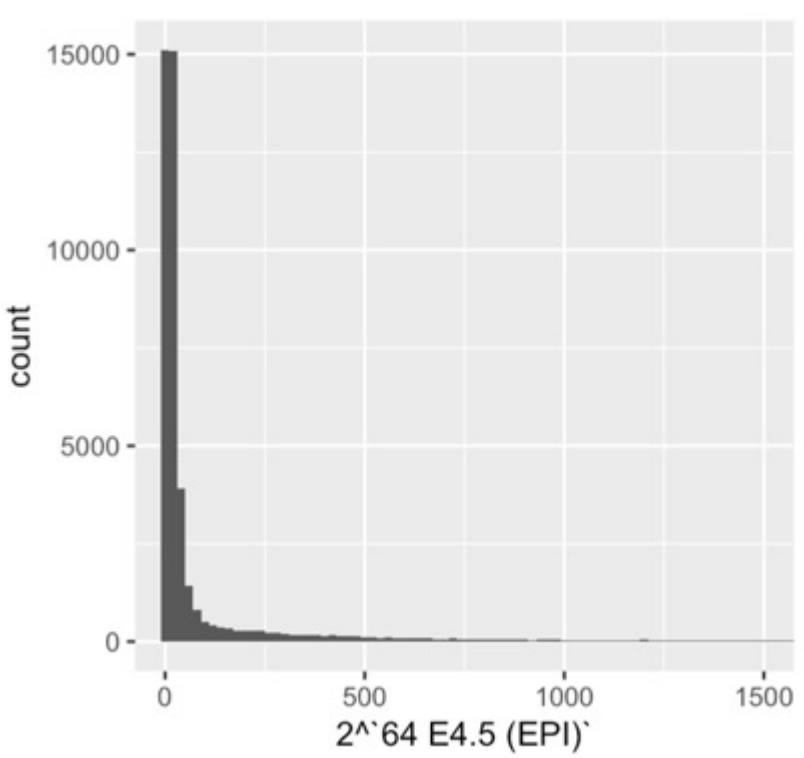
based on Holmes, Huber, *Modern Statistics for Modern Biology*

Tips to remember

Axes



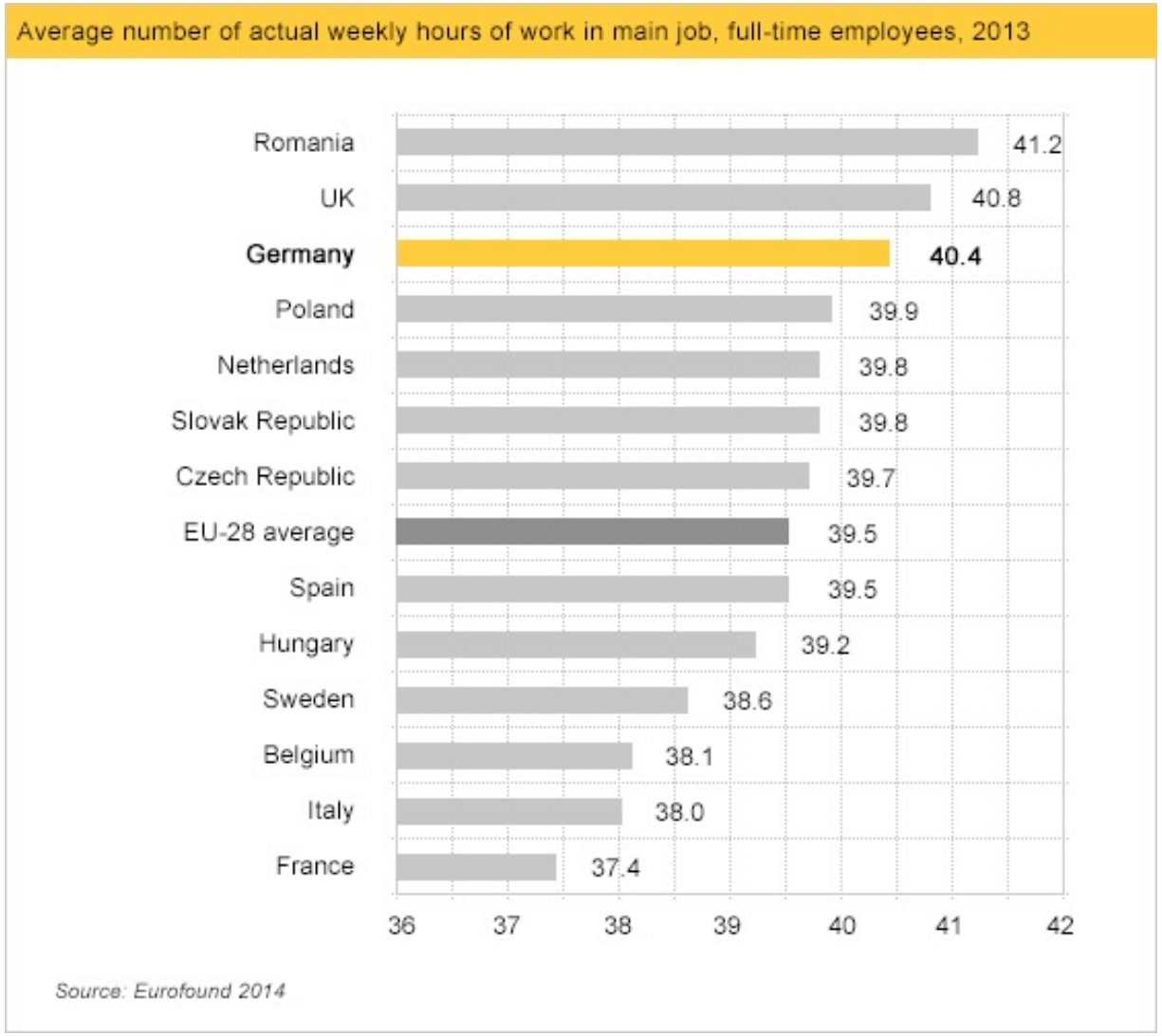
Logarithmic x-axis

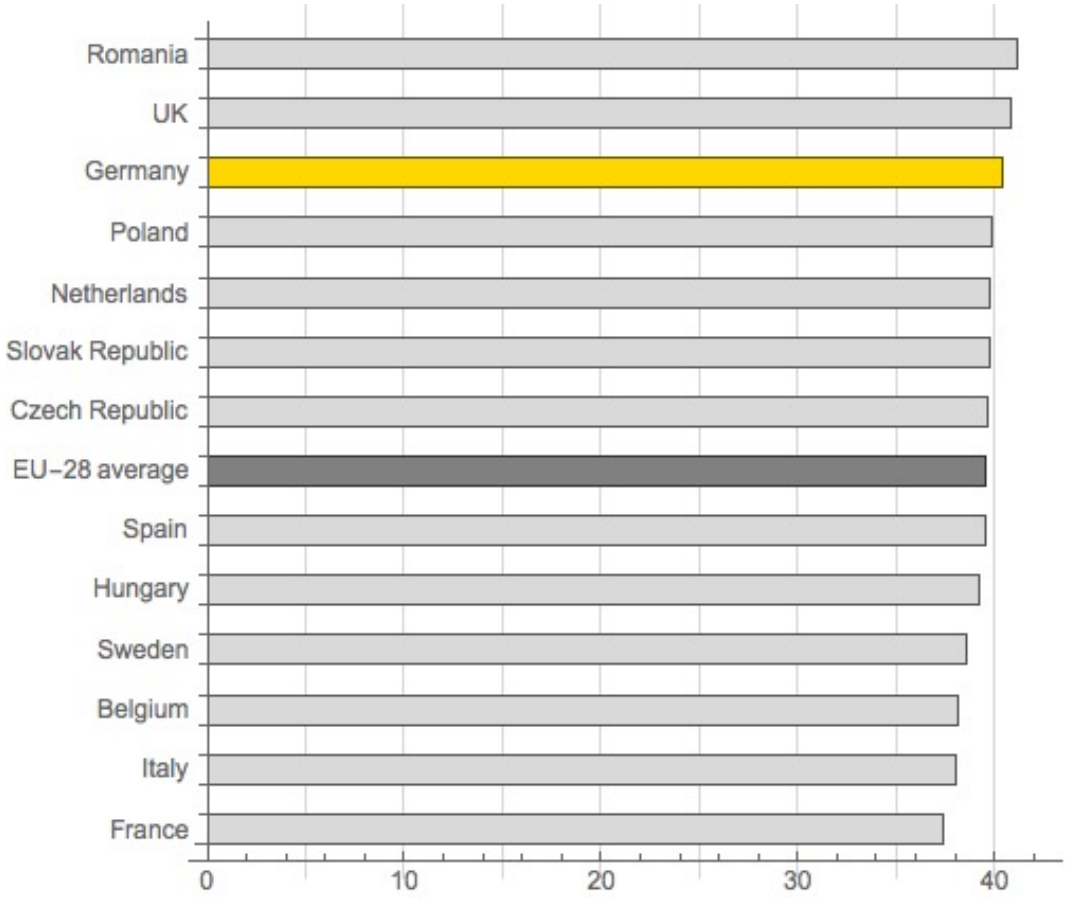


Linear x-axis

Tips to remember

Axes

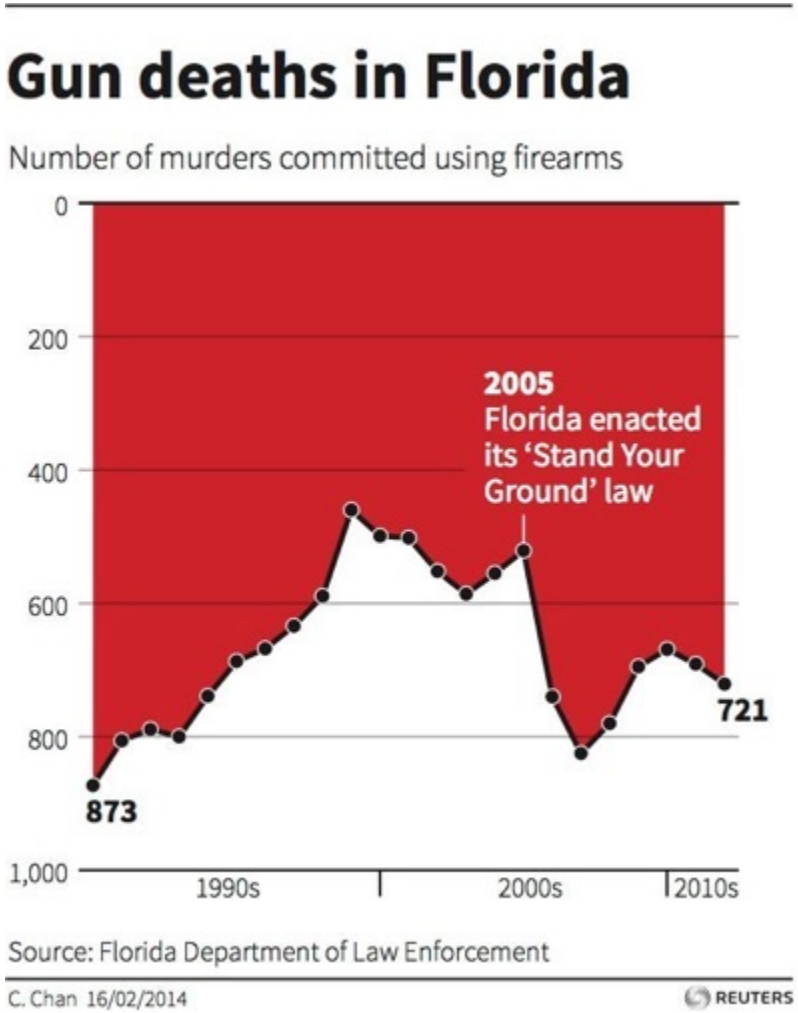


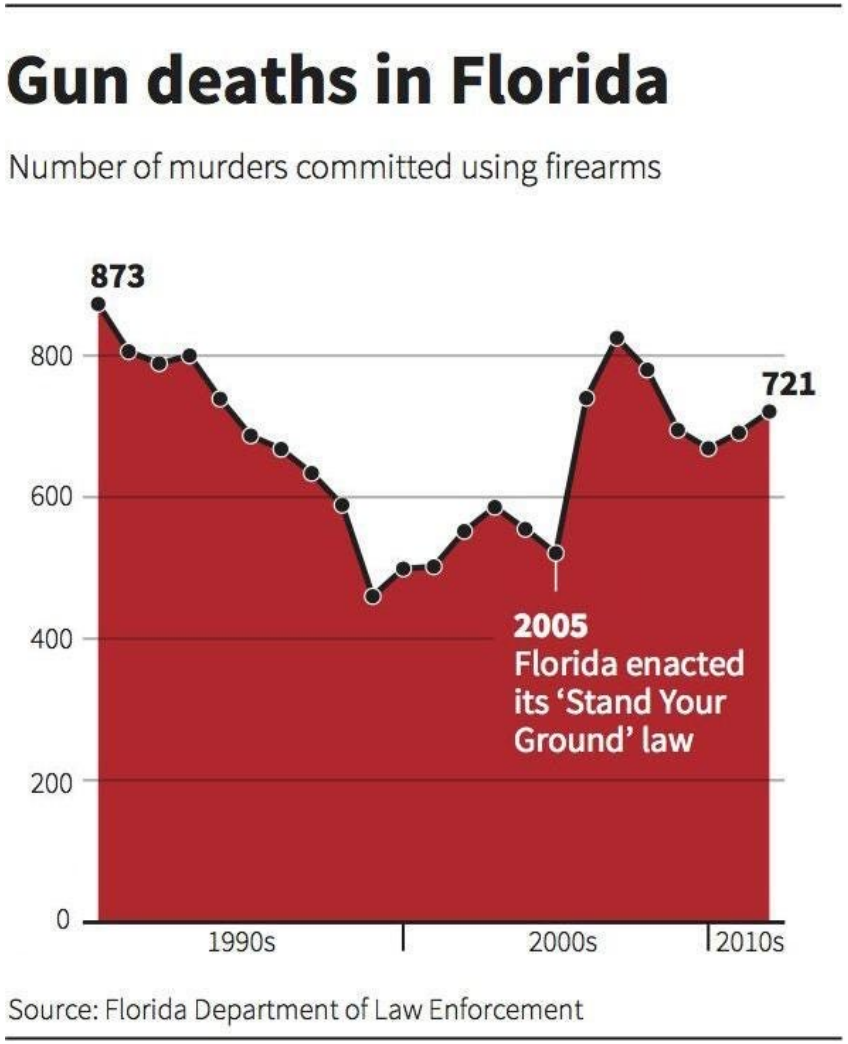


https://callingbullshit.org/tools/tools_misleading_axes.html

Tips to remember

Axes





Have a healthy weekend

Office Hours announcement
Project Description
Class recording + Slides