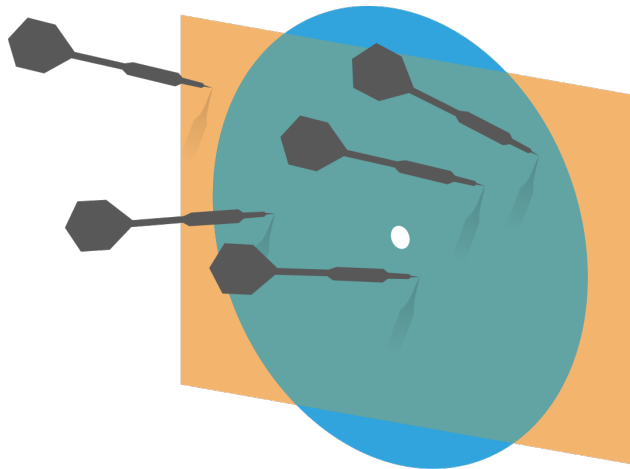


Probability and Statistics for Computer Science



“The eternal mystery of the world is its comprehensibility ... The fact that it is comprehensible is a miracle.”
– Albert Einstein

Credit: wikipedia

How to Zoom in the lectures

- ✱ Students' Video and Audio will be both muted during the lecture unless permitted by the instructor for questions.
- ✱ You can use the chatbox to ask questions or write comments.
- ✱ Questions will be collected by the assistant for answers or summary.

Test Poll₁

✱ Have you read the syllabus on the course website?

A. Yes. B. No.

Test Poll2

✱ Have you done the survey on the course
Compass website?

A. Yes. B. No.

Test Poll3

✱ Have you watched the welcome video in the Orientation module?

A. Yes. B. No.

Test Poll4

- ✱ How many ways are there to watch a lecture video of CS361 Sp21?
 - A. 1.
 - B. 2.
 - C. 3.
 - D. None of the above.

Objectives

- ✱ Welcome/Orientation
- ✱ Big picture of the contents
- ✱ Lecture 1 - Data Visualization & Summary (I)

Vision (PCA)

- ✱ **P**assion for learning
- ✱ **C**ompassion for each other
- ✱ **A**uthentic understanding

How to succeed in this course?

- ✱ Factors that will hinder you from success
- ✱ Factors that will help you succeed

Avoid these that could cause failure

- ✱ Academic integrity infraction – by all means!
- ✱ Missing homeworks, project or quizzes
- ✱ Late/Poor homeworks or project
- ✱ Insufficient viewing of the contents
- ✱ Poor time management
- ✱ Too many challenging classes at the same time
- ✱ Not motivated/not interested in the topic

Factors that will help you succeed

- * **Be engaged/motivated,**
- * **Do not hesitate to ask** for help.
- * Be **Active** in class participation
- * Do as much practice as possible, not just the homework and project.
- * Read the textbook and other recommended books.
- * Clear your doubts/misconceptions **asap (every lecture/discussion is important)**

Interactions are important!

- ✱ Try to go to office hours as much as possible
- ✱ Try to meet or talk to the instructor as least once personally
- ✱ You are encouraged to join the team work (extra points opportunities)
- ✱ Show compassion via community service

Graded Team work



Extra Points



Quizzes



Course materials

- ✱ Compass Course Site

Find it through Compass for CS361 Spring 2021 AL1

- ✱ Public Website

- ✱ <https://courses.grainger.illinois.edu/CS361/sp2021/>

Lecture videos and ClassTranscribe

- ✱ Lecture and discussion will be recorded and accessible at <https://mediaspace.illinois.edu/>
- ✱ ClassTranscribe provides transcripts for these videos
<https://classtranscribe.illinois.edu/home>
- ✱ The Zoom recording links and the specific links of the above two channels are all on Compass

Our Staff

Instructor: Hongye Liu

Teaching Assistants:

Weikai Xu (ADA & ADB),

Sneha Krishna Kumara (ADC & ADD),

Aditya Karan (ADE & ADF),

Yiren Wang (ADG & ADH).

Office hours are listed under Zoom Meetings.

Our Staff (II)

Course Assistants: Ajay Fewell, Christina Hu, Chenhui Zhang, Lilac Lai, Matthew Chen, Shirley Mao, and Vishesh Gupta.

Let's dive into the content

- ✱ Probability and Statistics in action

- ✱ What does this course teach?

Textbook: Forsyth, D. A. "Probability and Statistics for Computer Science," Springer (2018)

- ✱ Why are there 4 sections? How are they related?

This field really started with gaming

- ✱ We are familiar with flipping a coin or throwing a dice, the result is uncertain!



Head
Or Tail?



Which side
is front?

Life is uncertain so aim for long-term average

- ✱ We repeat a lot of experiments and see if there is regularity



Head
Or Tail?



Which side
is front?

Throwing a lot of “coins” for many times in one touch

✱ Galton board, the Bead Machine

<https://www.youtube.com/watch?v=Kq7e6cj2nDw>

Probability and Statistics

Experiment in action



Simulation of random draw of a picture on computer



✱ It's the same as
throwing a 4-sided die.



What does this course teach?

- ✱ Describing Datasets

 - Summary & visualization

- ✱ Probability

- ✱ Inference – Statistical Inference

- ✱ Tools – Machine Learning tools

Describing datasets (Summary & visualization)

Descriptive & Graphical

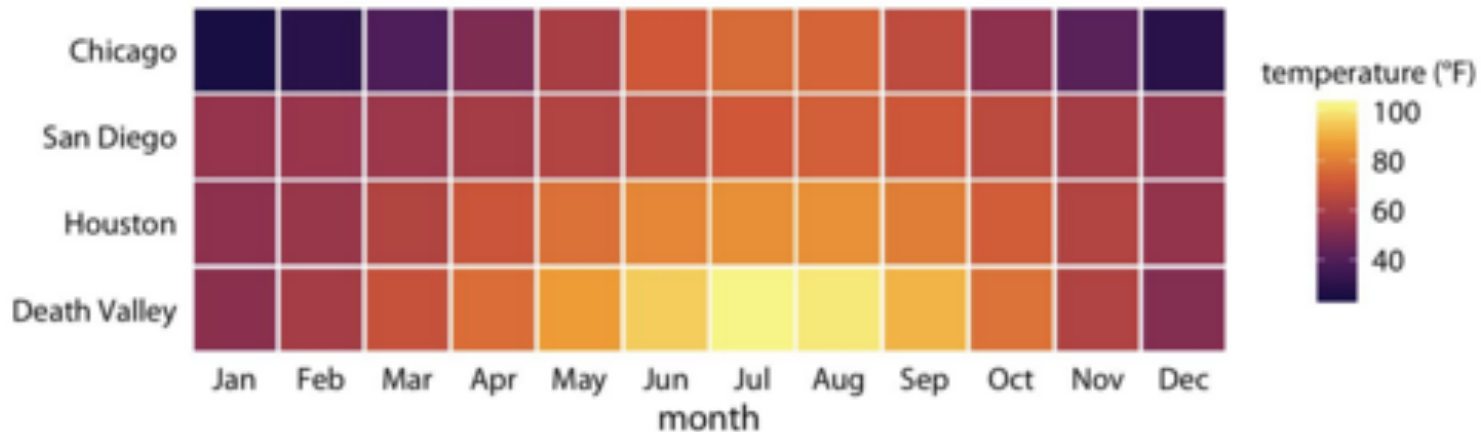


Figure 2-4. Monthly normal mean temperatures for four locations in the US. Data source: NOAA.

Summarization of 4 locations' annual mean temperature by month

Probability

✻ Mathematical

Romeo and Juliet have a date

Each arrives with a delay btw 0 and 1 hour. The first to arrive leaves after $1/4$ hour. All pairs of delays are equally likely.

What's the probability that they will meet?

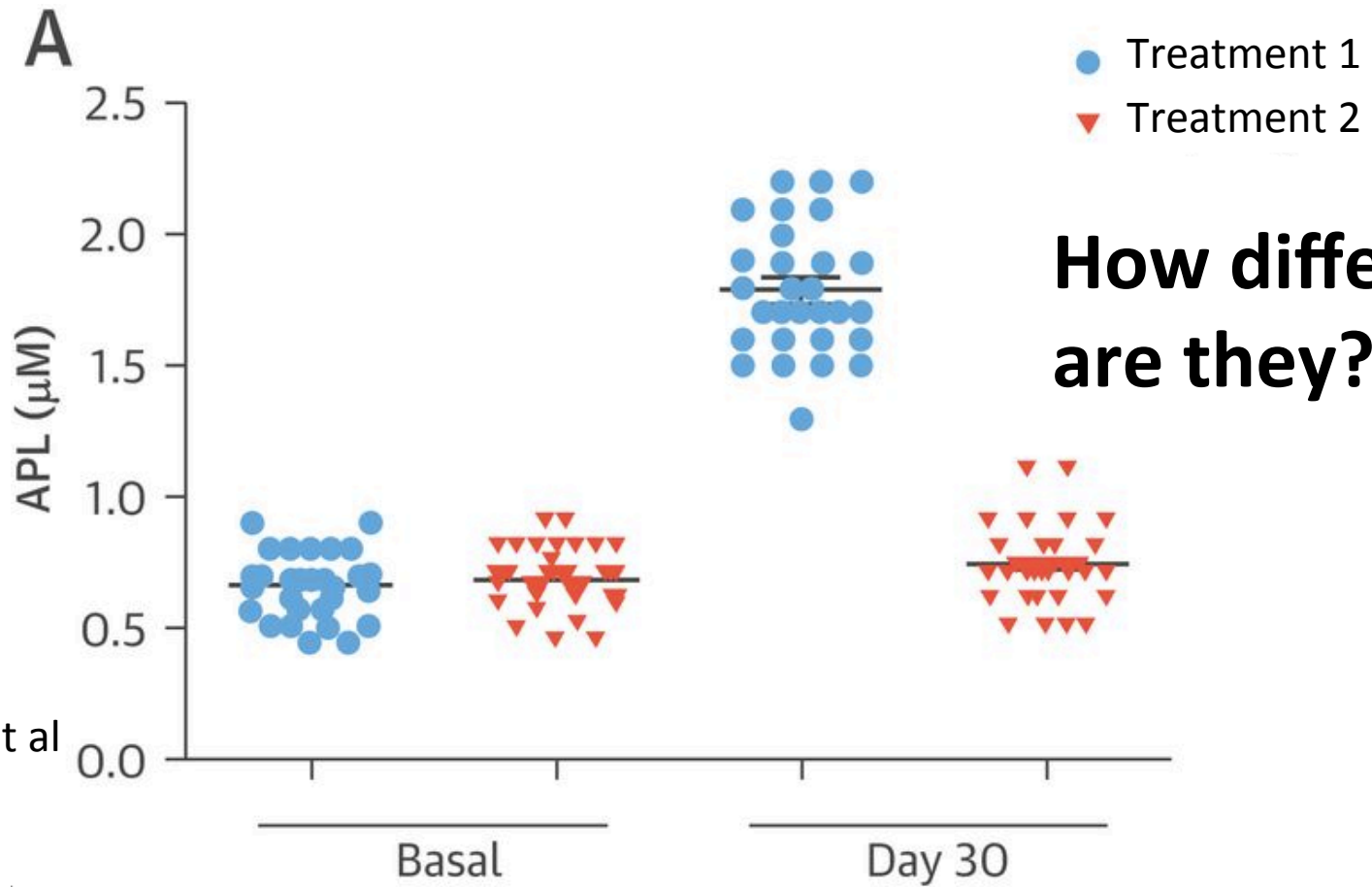
Probability

✱ Mathematical

How many slots are empty on average for a simple hashing table?

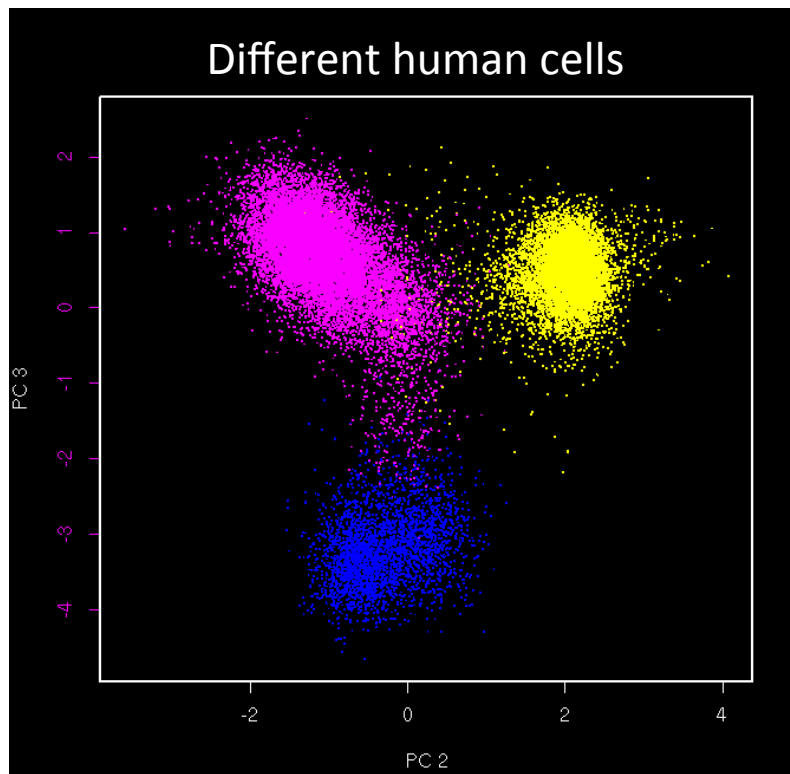
Inference

✱ Analytical



Tools (Machine learning)

✪ Algorithmical



High-dimensional or complex shaped data sets need tools! Humans are limited in 2-3D.

Machine learning is Highly desired!
Often depends on Statistics.

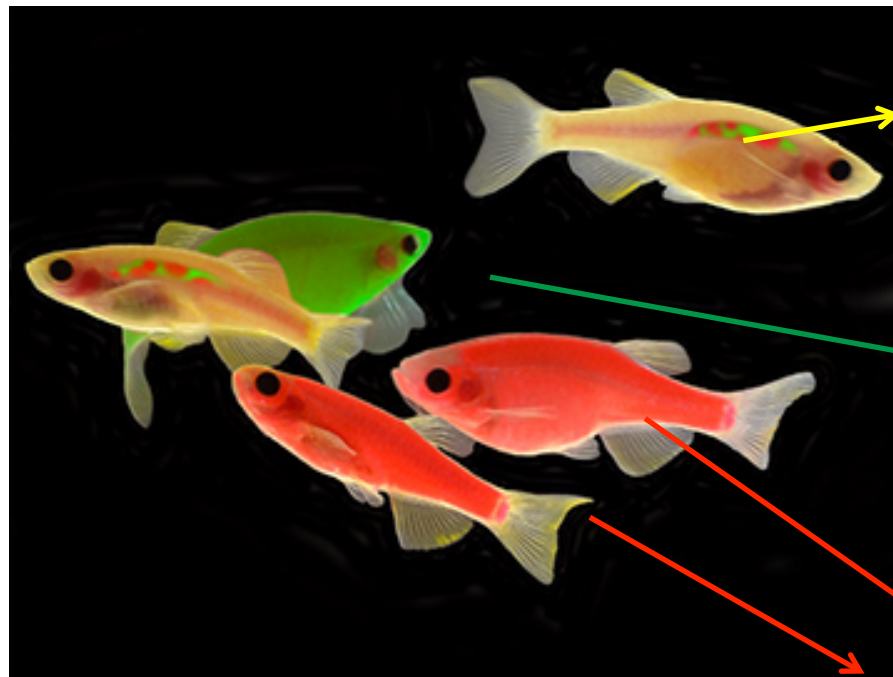
Why these 4 sections?

- ✱ Summary & visualization
Graphical
- ✱ Probability
Mathematical
- ✱ Inference – Statistical Inference
Analytical
- ✱ Tools – Machine Learning tools
Algorithmical

Why these 4 sections?

- ✱ The common thread is **Data**.
- ✱ We are doing computer science and so

are like
these
yellow
fish



Data Science +
Comp. Science

Statistics

Mathematics

What is special of Data? For Data?



Why these 4 sections?

- ✱ Real world data is often high dimensional and complex
- ✱ These 4 parts of knowledge or techniques are inseparably/organically connected in many real world applications.

What do we emphasize?

- ✱ Mathematical principle
- ✱ Critical thinking
- ✱ Working with real world data

LECTURE 1

Q. What do you feel about it when we speak of data visualization?

Example 1: Black hole

Constructed image
using data collected
from many different
telescopes' view of the
same object

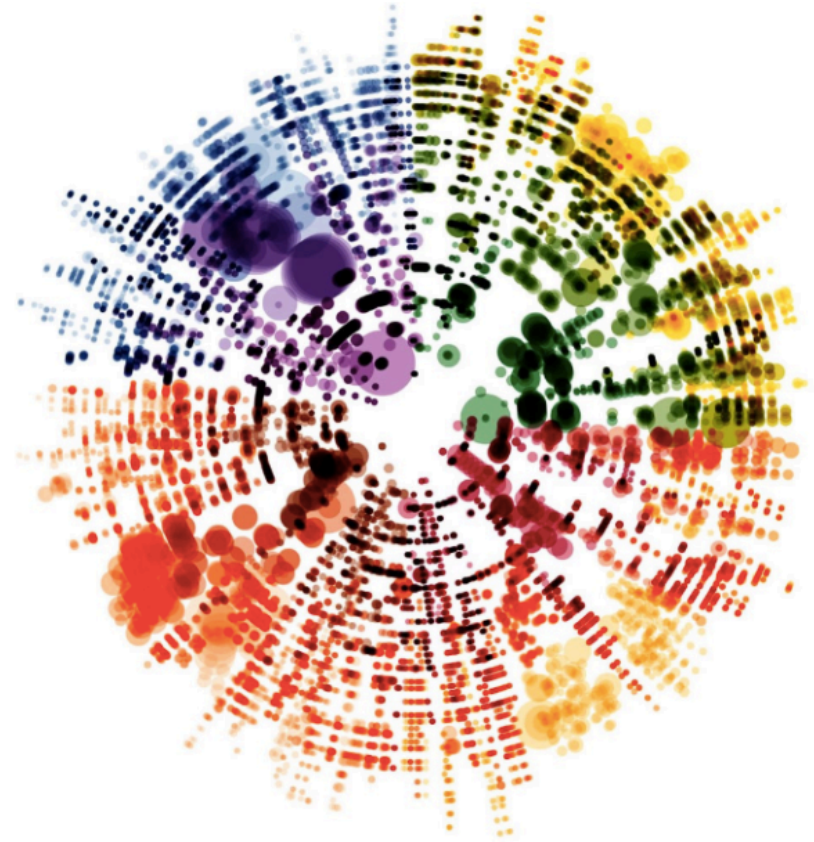
This project received a
3million-dollar award



Credit: NASA

Example 2: Four seasons by Vivaldi

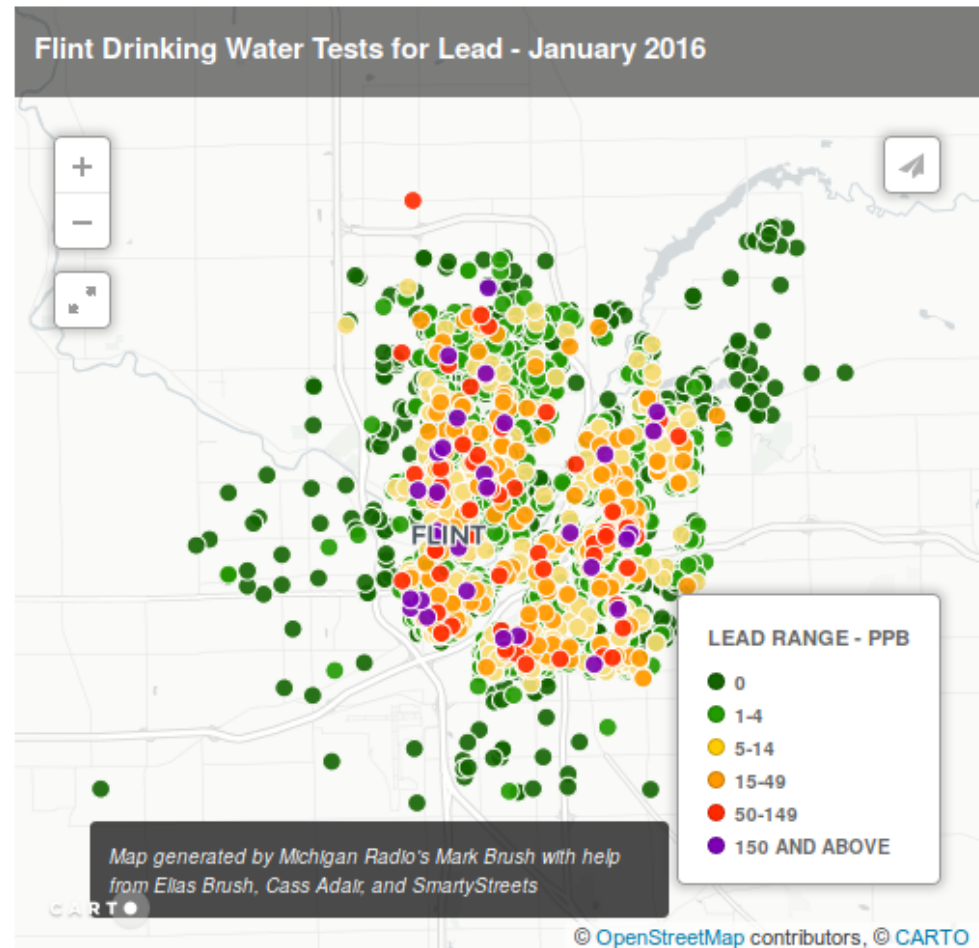
Pitch is shown by the distance from center;
Length of the note is the size of dot
Instrument is shown by the color



<https://medium.com/future-today/off-the-staff-an-experiment-in-visualizing-notes-from-music-scores-58f6ee9f0cef>

Example 4: GIS map

Color scaled dots show the lead level in water in an area in Michigan



Lecture I: Data Visualization & Summary

- ☼ Datasets $\{x\}$ – a set of N items x_i , $i=1\dots N$, each of which is a tuple

Proteins \longrightarrow

Cells
 \downarrow

Cell ID	CD45	CD3e	CD19	CD11b	Ki67
1	7.10543765	1.99490875	2.13073358	7.82894178	2.57289058
2	6.5957055	4.65342077	1.62918585	0.88137359	0.88137359
3	6.81991147	1.76259579	4.63429706	2.74452653	0.88137359
4	6.90112651	1.41502227	4.54593607	0.88137359	0.88137359
5	6.75571436	2.87597714	2.18671075	6.72464322	0.91192661
6	7.39538689	2.55285118	4.55845203	1.57273629	0.88137359
7	6.50181654	0.9030504	0.88137359	6.55459538	1.61883699
8	6.60986569	2.1753298	1.52779681	6.44086205	1.5347653
9	6.97651408	2.38246511	1.90249637	3.41580053	1.85303806
10	7.14397512	3.36924119	9.23325502	4.79035059	0.88137359

Each row is a tuple

Lecture I: Data Visualization & Summary

- ☼ Convention: columns are the *features*; the number of features is *dimension*.

Proteins →

Cells ↓

Cell ID	CD45	CD3e	CD19	CD11b	Ki67
1	7.10543765	1.99490875	2.13073358	7.82894178	2.57289058
2	6.5957055	4.65342077	1.62918585	0.88137359	0.88137359
3	6.81991147	1.76259579	4.63429706	2.74452653	0.88137359
4	6.90112651	1.41502227	4.54593607	0.88137359	0.88137359
5	6.75571436	2.87597714	2.18671075	6.72464322	0.91192661
6	7.39538689	2.55285118	4.55845203	1.57273629	0.88137359
7	6.50181654	0.9030504	0.88137359	6.55459538	1.61883699
8	6.60986569	2.1753298	1.52779681	6.44086205	1.5347653
9	6.97651408	2.38246511	1.90249637	3.41580053	1.85303806
10	7.14397512	3.36924119	9.23325502	4.79035059	0.88137359

Each row is a tuple with dimension =5

Data types

✱ Categorical

✱ Ordinal

✱ Continuous

Data types

- ✱ Categorical

Smoker or non-Smoker, Female or Male etc.

- ✱ Ordinal

Not satisfied, satisfied, very satisfied

- ✱ Continuous (any real number within a range)

Temperature

Q. Which of the following data is not categorical?

- A. Number of enrolled students in a class
- B. Weight of apples in a grocery store
- C. Instruments played by an orchestra
- D. Type of chemical reagents in a lab
- E. A & B

Simple Visualization of Data

- ✱ General principles
- ✱ Bar chart
- ✱ Histogram
- ✱ Conditional histogram

Simple Visualization of Data

✱ Tables

In Python or R, there is table format and there is `data.frame` which is a very versatile table for storing all kinds of data type.

Simple Visualization of Data

✱ General principles

Must not mislead or distort;

Aesthetically pleasing;

Clear, Attractive, Convincing;

Show message/significance.

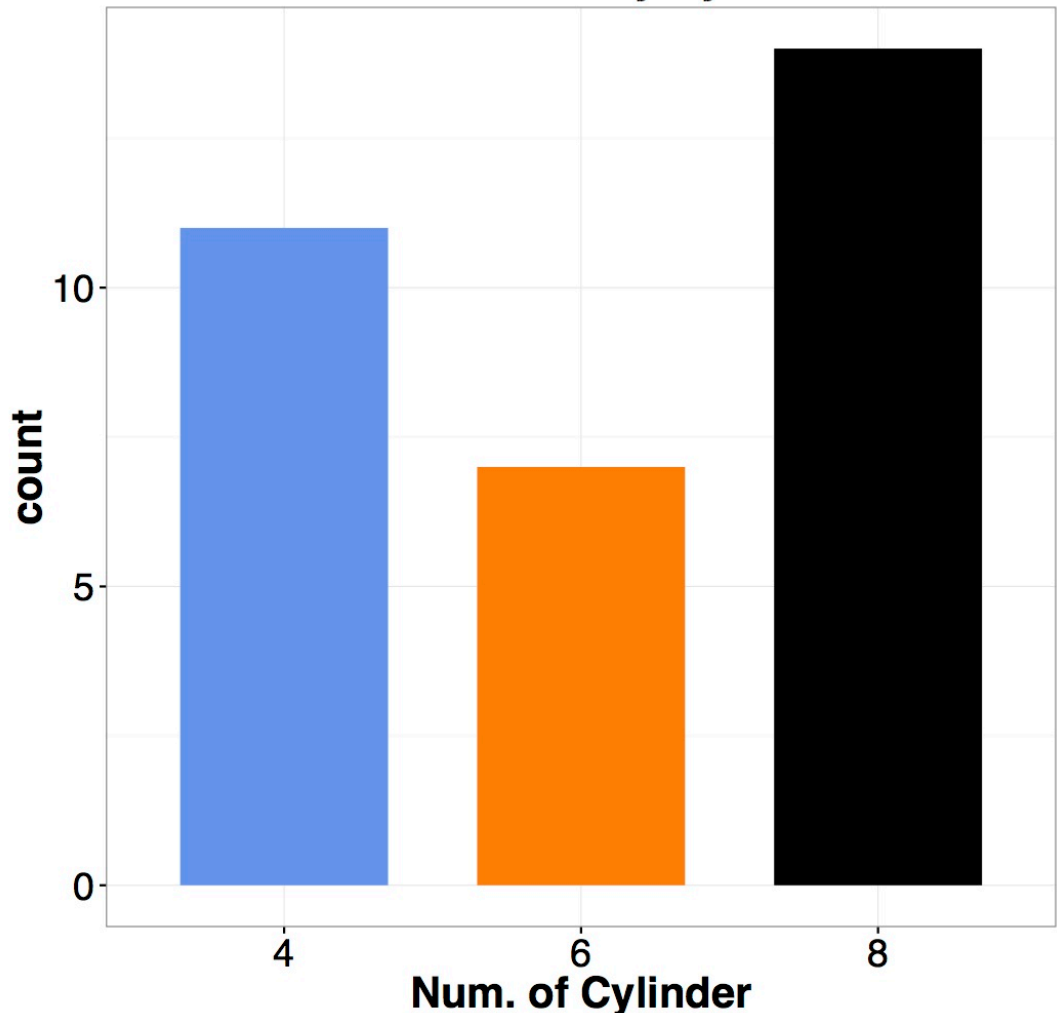
Simple Visualization of Data

✱ Bar chart

A set of bars that are organized by categorical or ordinal feature

Data: "mtcars"

Count of cars by Cylinder



An example of good, ugly, bad, wrong

Dr. Wilke illustrated the difference between *good*, *ugly*, *bad* and *wrong* visualization

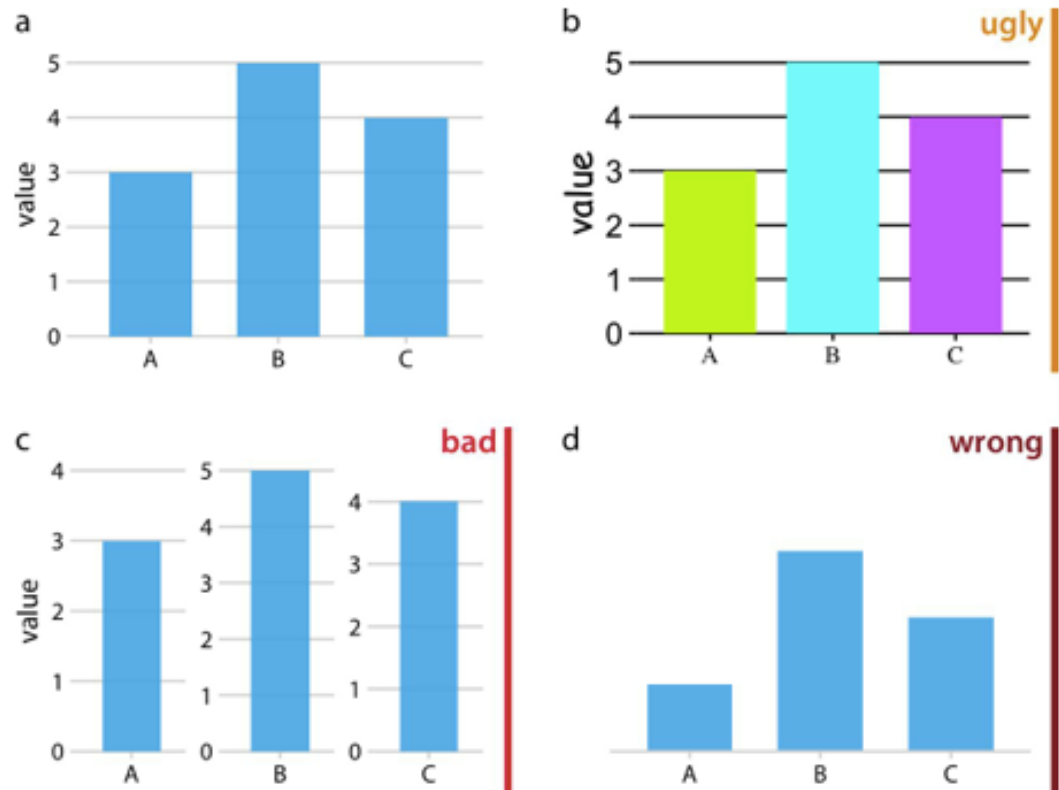


Figure 1-1. Examples of ugly, bad, and wrong

Q: Is this a good bar chart?

Q1 (by day)

Chart Type ▾

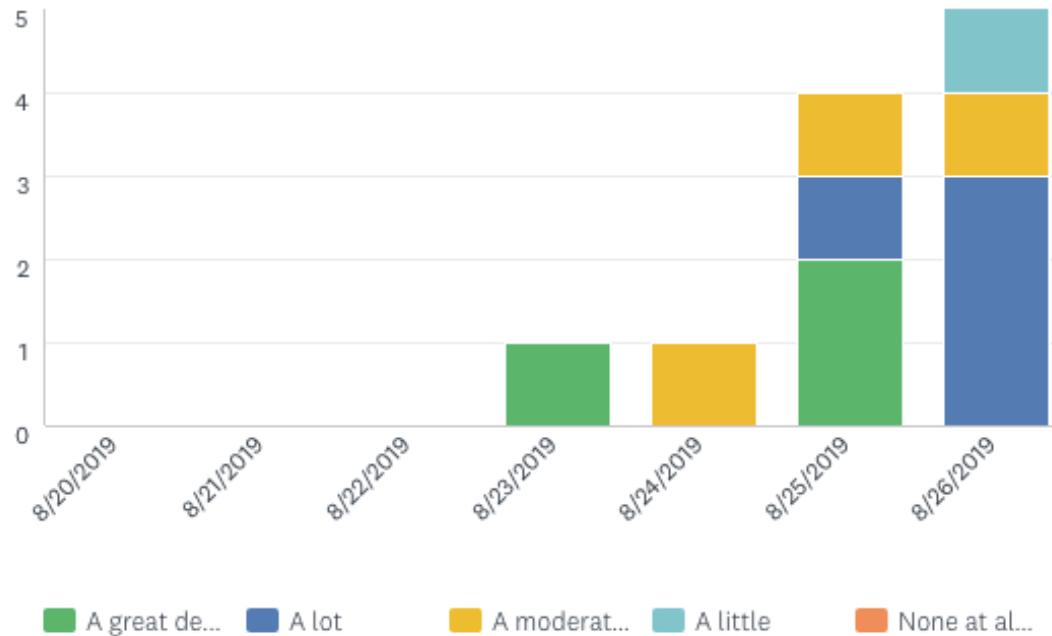
Display Options ▾

Trend by... ▾

Zoom ▾

How much do you expect this course to relate to your future career?

Answered: 11 Skipped: 0 First: 8/23/2019 Zoom: 8/20/2019 to 8/26/2019



A. Yes

B. No

How about using a color scale

Q1 (by day)

Chart Type ▾

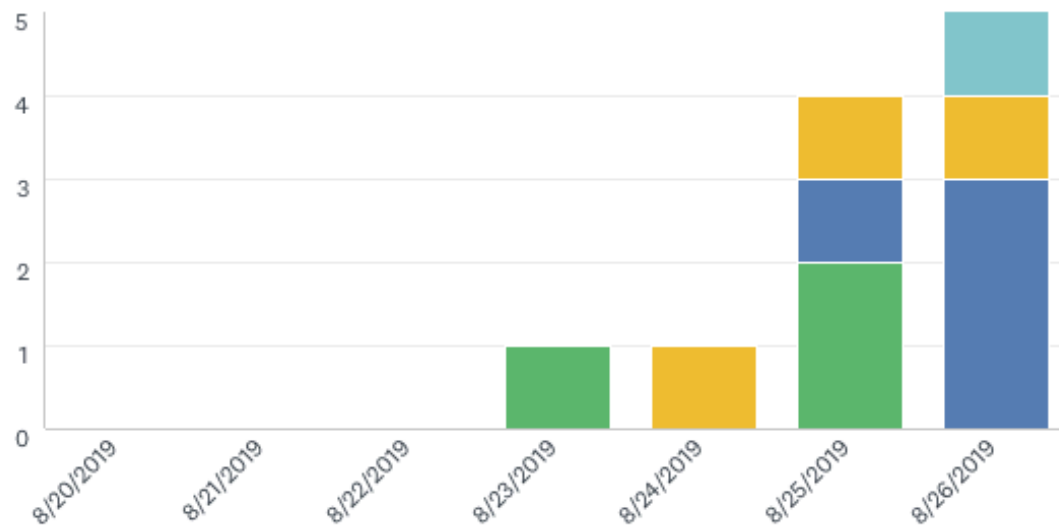
Display Options ▾

Trend by... ▾

Zoom ▾

How much do you expect this course to relate to your future career?

Answered: 11 Skipped: 0 First: 8/23/2019 Zoom: 8/20/2019 to 8/26/2019



■ A great de... ■ A lot ■ A moderat... ■ A little ■ None at al...

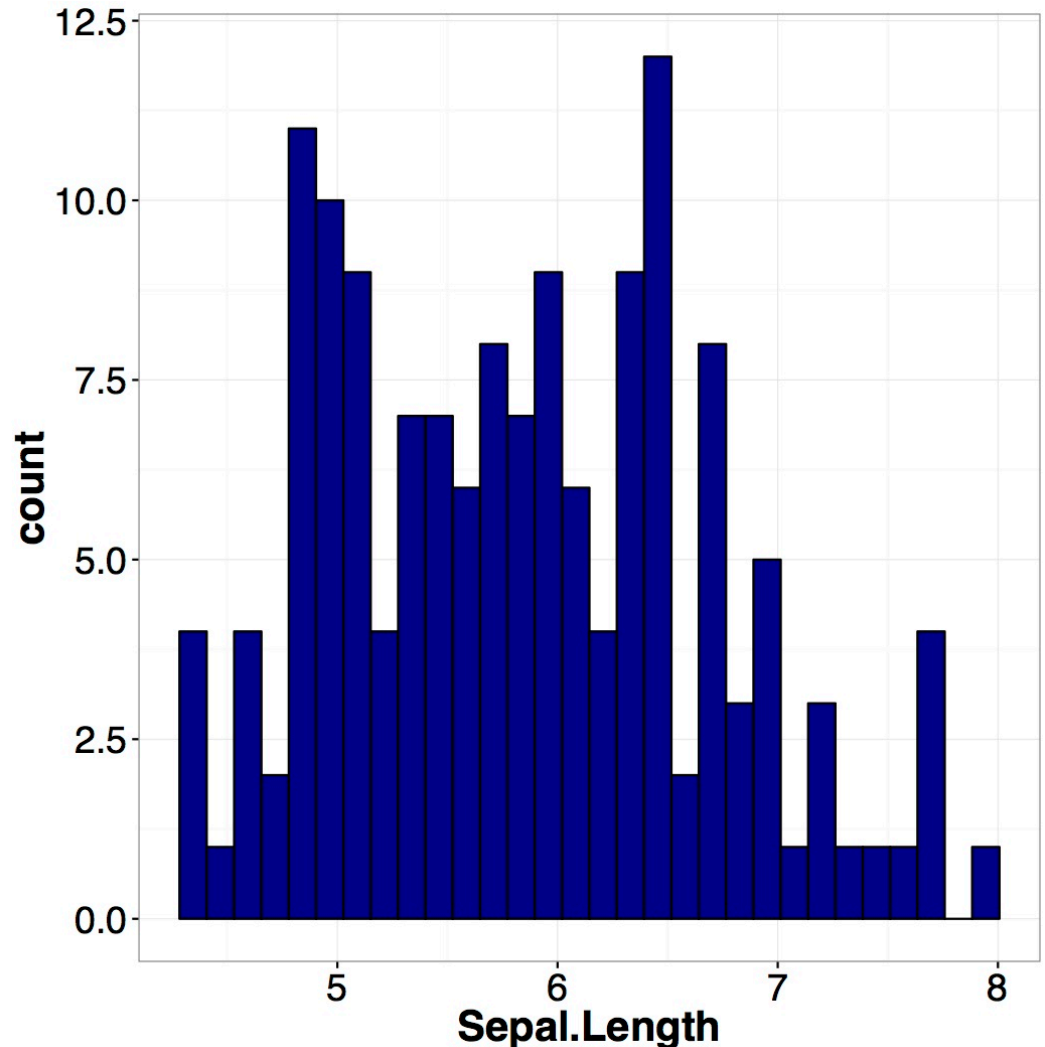


Visualizing Data with Histogram

✧ Histogram

A set of bars that are organized by bins that contains numerical data (discrete or continuous)

Data: "iris"

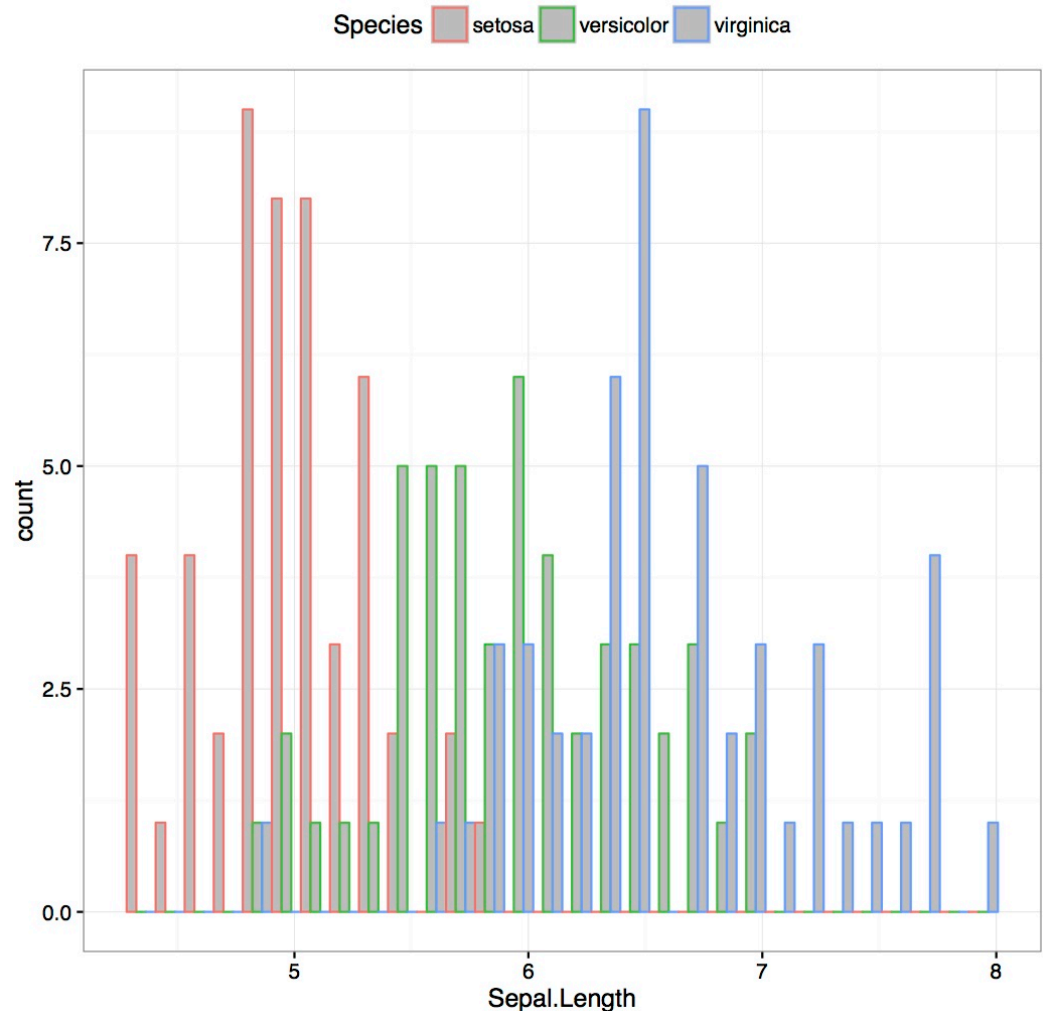


Visualizing Data with Histogram (II)

☼ Conditional histogram

Histogram generated by subsets of the data

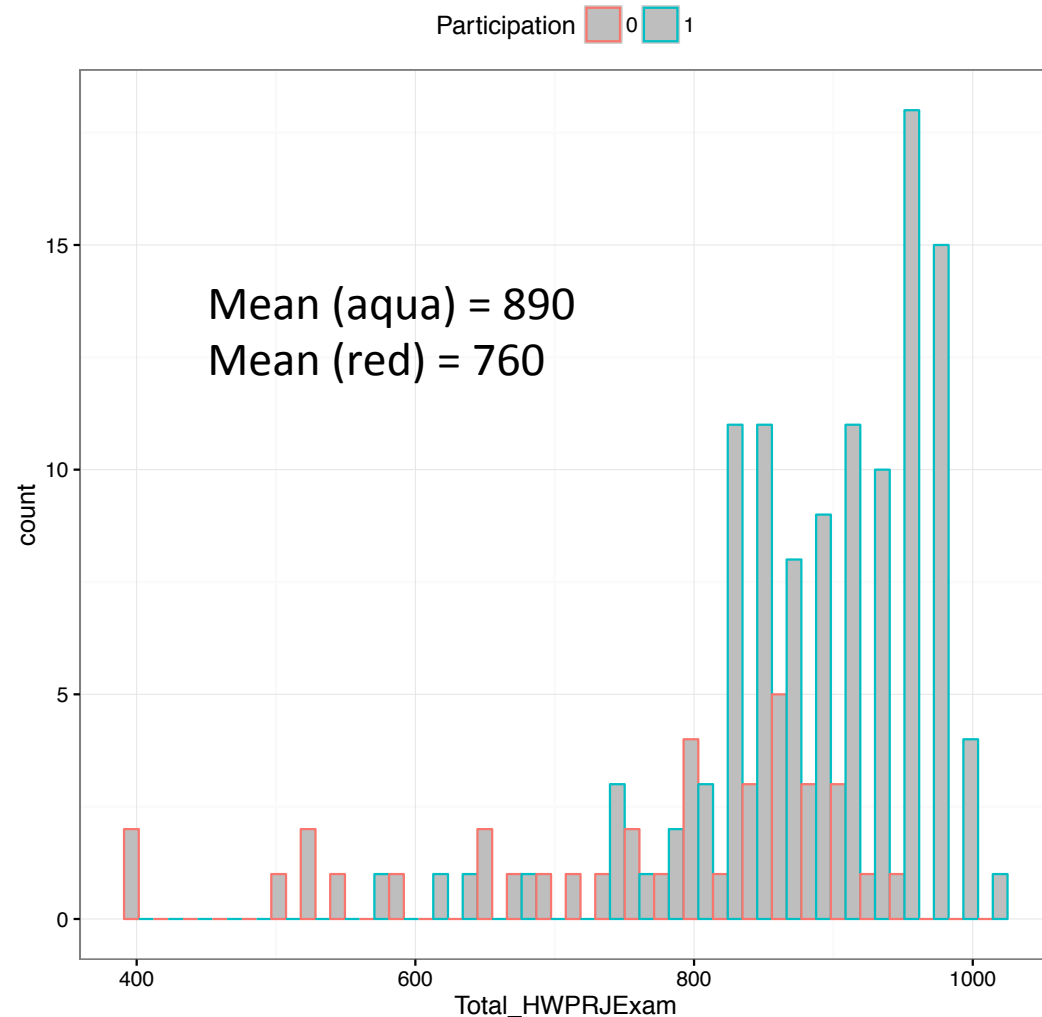
Data: "iris"



Visualizing Data with Histogram (III)

☼ Conditional histogram

Data: Combined Score (HWs, Prj and Exams) grouped by students with **full** participation or **not full** in CS361 fall 2019



Assignments

- ✱ Finish the orientation module on Compass
- ✱ Submit HW0 to Gradescope to test it
- ✱ Start week1 module on Compass

Additional References

- ✱ Charles M. Grinstead and J. Laurie Snell
"Introduction to Probability"
- ✱ Morris H. Degroot and Mark J. Schervish
"Probability and Statistics"

See you next time

See you!

