Introduction to Data Science With Probability and Statistics

Lecture 1: Introduction

CSCI 3022 - Summer 2020
Sourav Chakraborty
Dept. of Computer Science
University of Colorado Boulder

Logistics

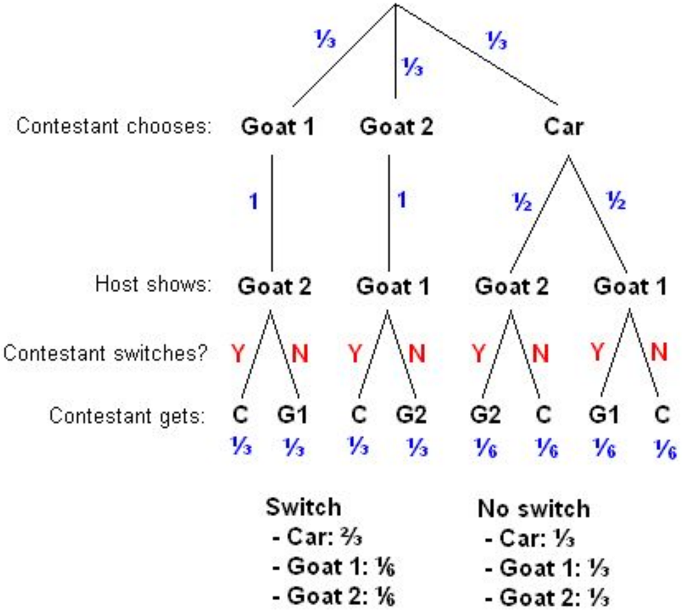
- **Instructor**
  - Name: Sourav Chakraborty (Pronouns: He/Him/His)
  - Email: sourav.chakraborty@colorado.edu
  - Office Hour: Every Thursday 3.30pm-4.30
  - Zoom link: https://cuboulder.zoom.us/j/94424300585
- **GSS**
  - Name: Amit Baran Roy (Pronouns: He/Him/His)
  - Email: amit.roy@colorado.edu
  - Office Hour: Every Friday 3.30pm-4.30pm
  - Zoom link: TBD
- **Course Assistant**
  - Name: Yibo Yang (Pronouns: He/Him/His)
  - Email: yibo.yang@colorado.edu
  - Office Hour: 10am-11am Monday to Thursday.
  - Zoom link: TBD

Logistics

- **Course Web Page**
  - Canvas: https://canvas.colorado.edu/courses/62409
  - Slides, Video recordings, Coding assignments, Grades, Online quizlets and resources.
- **Piazza Page**
  - Signup Link: piazza.com/colorado/summer2020/csci3022
  - Access Code: **SUM20**
  - Class Discussion Forum. Do not post direct solutions/codes.
- **Gradescope**
  - Access Code: **9GBBZ2**

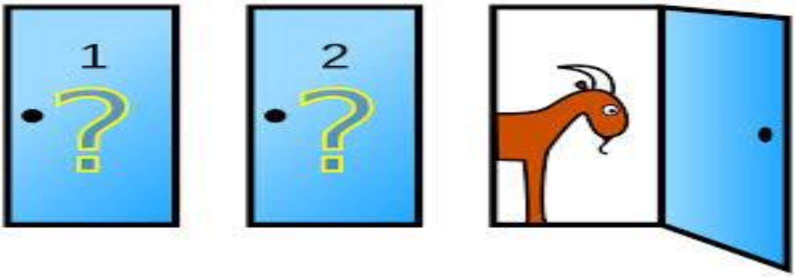
The Monty Hall Show

The Monty Hall problem is a brain teaser, in the form of a probability puzzle, loosely based on the American television game show Let's Make a Deal and named after its original host, Monty Hall. There are 2 goats and 1 car behind the 3 doors. You are to choose a door first, and then Monty will open another door with a goat behind it, now you will be given a choice to switch to the other remaining door or to stick your original choice. Will you switch?



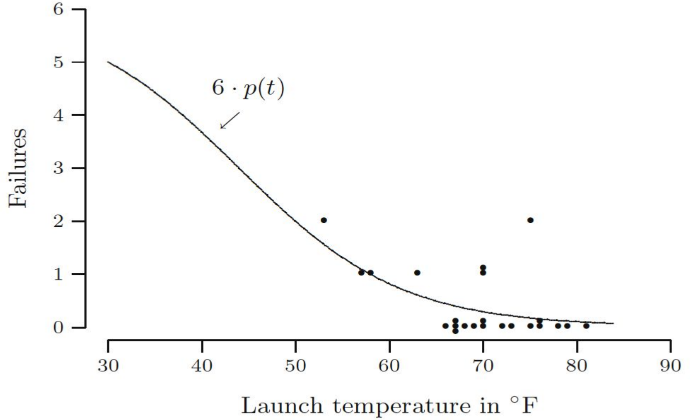
Switch
- Car: ⅔
- Goat 1: ⅙
- Goat 2: ⅙

No switch
- Car: ⅓
- Goat 1: ⅓
- Goat 2: ⅓



Challenger Disaster 1986 [video link]

The Space Shuttle Challenger disaster was a fatal incident in the United States space program that occurred on Tuesday, January 28, 1986, when the Space Shuttle Challenger broke apart 73 seconds into its flight, killing all seven crew members aboard. An investigating committee was set up, which included Richard Feynman. They used statistics to find the root cause of the problem and had a press conference revealing it.



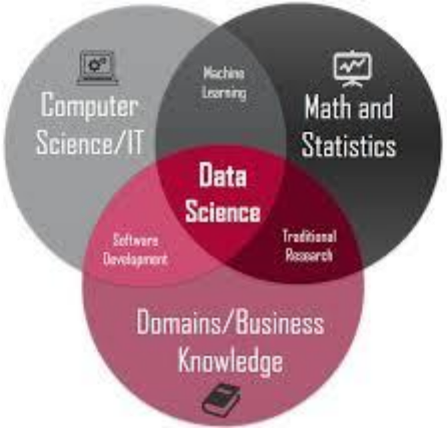
Source: based on data from Volume VI of the Report of the Presidential Commission on the space shuttle Challenger accident, Washington, DC, 1986.

World War 2 - The case of the missing bullets

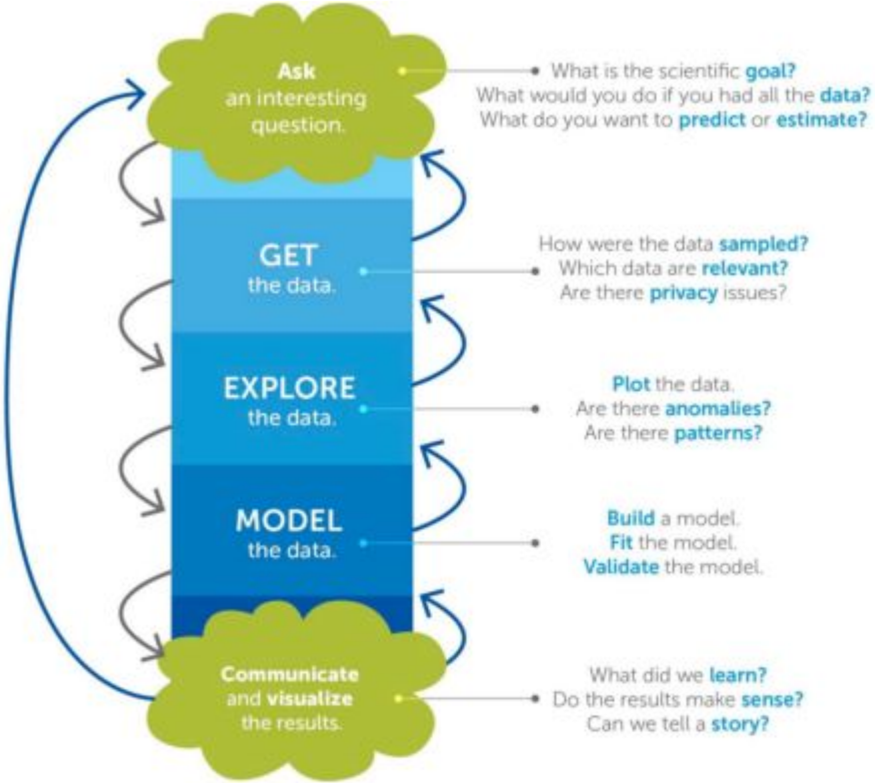
The statistical research group(SRG) at Columbia was a classified program that yoked the assembled might of American statisticians to the war effort. They analysed the bullet holes in the battle planes returning from Europe to design optimal quantities of protective shielding to the planes to not weigh them down and decrease their agility, but just enough to save them from the barrage of bullets.

What is Data Science?

- Recovering insights & trends hiding within the data
- Using data to answer interesting questions
- Using data to understand the world around us
- Probability & Statistics



What is Data Science?



Ask an interesting question.

What is the scientific **goal?**
What would you do if you had all the **data?**
What do you want to **predict** or **estimate?**

GET the data.

How were the data **sampled?**
Which data are **relevant?**
Are there **privacy** issues?

EXPLORE the data.

Plot the data.
Are there **anomalies?**
Are there **patterns?**

MODEL the data.

Build a model.
Fit the model.
Validate the model.

Communicate and visualize the results.

What did we **learn?**
Do the results make **sense?**
Can we tell a **story?**

Hypothesis

↓

Observations

↓

Analysis

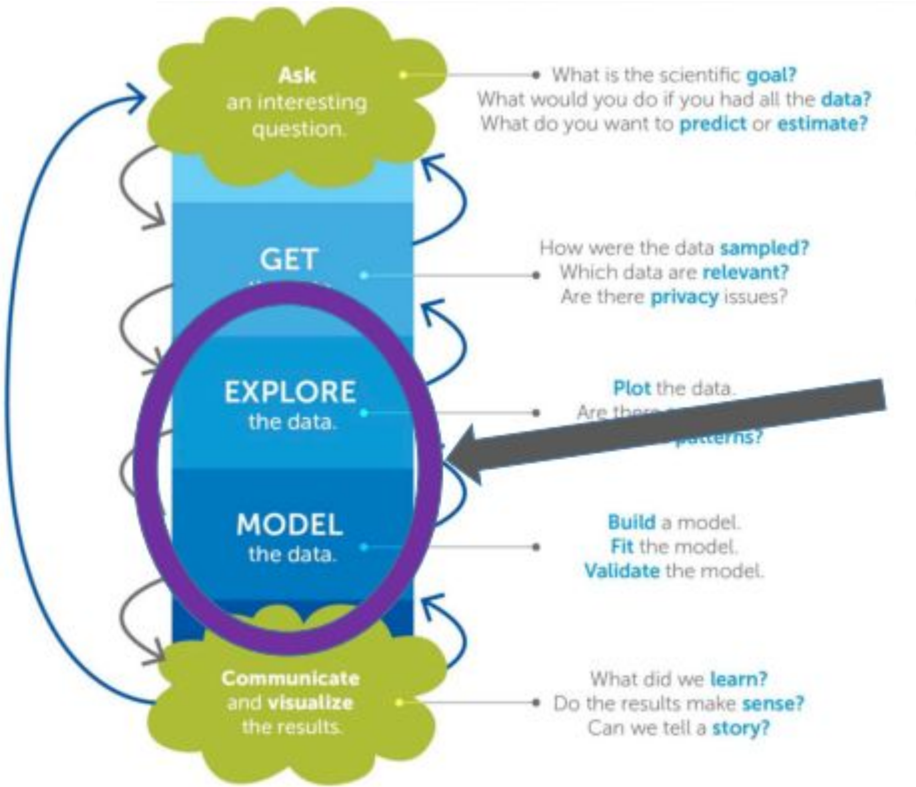
↓

Conclusions

↓

Refinement (Do it all over again?)

What is Data Science?



Ask an interesting question.

What is the scientific **goal?**
What would you do if you had all the **data?**
What do you want to **predict** or **estimate?**

GET

How were the data **sampled?**
Which data are **relevant?**
Are there **privacy** issues?

EXPLORE the data.

Plot the data.
Are there
patterns?

MODEL the data.

Build a model.
Fit the model.
Validate the model.

Communicate and **visualize** the results.

What did we **learn?**
Do the results make **sense?**
Can we tell a **story?**

We will focus largely on this part

- Exploration
  - Data Mining (**discover**)
- Modeling
  - Statistical analysis (**understand**)
  - Rudimentary machine learning (**predict**)

Course Topics

- **Exploratory Data Analysis**
  - Data wrangling, cleaning, packages, visualization.
- **Probability Theory and Simulation**
  - Basic axioms, laws, conditioning, Bayes' theorem.
- **Random Variables and Distribution**
  - Expectation, Variance, Law of large numbers, Central Limit Theorem.
- **Basic Statistical Modelling**
  - Bootstrap, Estimators, Mean Squared Errors, Maximum Likelihood etc.
- **Hypothesis Testing**
  - What is the data really telling us and how confident should we be in our conclusions?
  - T-Test, Z-test, Confidence Intervals
- **Special Topics**
  - Linear Regression, Logistic Regression

Grading

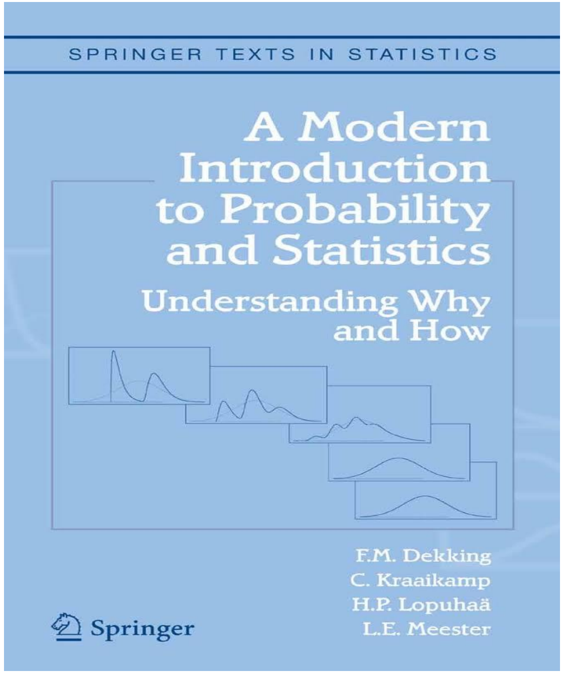
- **40% - Homework assignments.**
  - Once in ~2 weeks. Total 3-4.
- **30% - Midterm exam.**
  - ~ After 3 weeks of classes. date TBD.
- **25% - Final Project/Exam**
  - End on July. date TBD.
- **5% - Quizlets on Canvas**
  - Small in-class quizzes with 2-3 questions each.
  - Once in every 2-3 lectures, only covering the recent-most content of the time.
  - Questions will be simple and just to freshen up the content of the previous 2 lectures.

Text Book

A Modern Introduction to Probability and Statistics

(MIPS)

By F.M Dekking et al.


SPRINGER TEXTS IN STATISTICS
A Modern Introduction to Probability and Statistics
Understanding Why and How
F.M. Dekking
C. Kraaikamp
H.P. Lopuhaä
L.E. Meester
Springer

Acquiring Data

- Not always do we get the data in the preferred format to start analysis.
- Data is acquired from various sources
    - Internet Website (Govt websites, Social Media, Personal Websites)
    - Data Storage Systems (Big Data from Cloud systems)
    - Some Relational Database (SQL Tables)
- Data might be messy and incomplete.
- Data Wrangling/Munging to the rescue.

Data Wrangling

- The process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.

Photo : Margaret Hamilton with her code for the Apollo Launch, circa 1969.
Source: Google

Data Wrangling

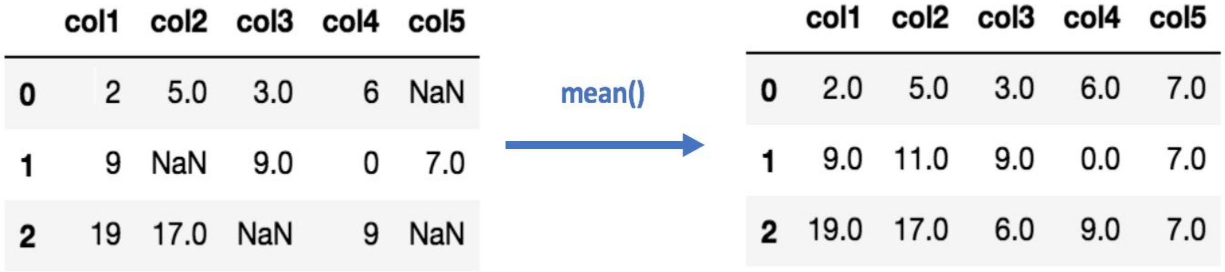
- Acquiring data from internet may involve scraping websites, downloading their data and convert them into various convenient data structures like Python dictionaries, Numpy Arrays, Pandas Dataframes.
- Few sources have their own Application Programmable Interfaces (APIs) available, which makes the whole process easier to acquire data using libraries like python **requests**.
- Some common data formats from internet websites are CSV, JSON, XML.
- These forms are then generally converted to pandas dataframes.

Data Wrangling

- To get data from data storage systems like the cloud, we use various big data frameworks like Apache Hadoop and Spark to work with huge volumes of data.
- Relational Databases are still very relevant and we require SQL query language to get data.

Data Wrangling

- Most of the time, the data is incomplete and one has to deal with missing data using techniques like imputation and linear interpolation.



	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

Practice Notebooks

- NB0: Python & *Numpy*
- NB1: Titanic Dataset with *pandas* Dataframe.

Continuous Feedback!

- Email anytime.
- Google form attached to the homework assignments (Anonymous)