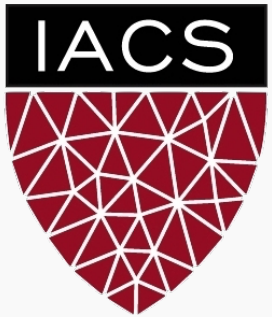
Lecture #0: Introduction to CS109A

CS 109A, STAT 121A, AC 209A: Data Science

Pavlos Protopapas Kevin Rader



Lecture Outline

What is Data Science

What is This Class?

Practicing Data Science

What is Data Science

Slide Title

What is This Class?

Slide Title

Practicing Data Science

The Data Science Process

The Data Science Process is similar to the scientific process - one of observation, model building, analysis and conclusion:

- Ask questions
- Data Collection
- Data Exploration
- Data Modeling
- Data Analysis
- Visualization and Presentation of Results

Note: This process is by no means linear!

Analyzing Hubway Data

Introduction: Hubway is metro-Boston's public bike share program, with more than 1600 bikes at 160+ stations across the Greater Boston area. Hubway is owned by four municipalities in the area.

By 2016, Hubway operated 185 stations and 1750 bicycles, with 5 million ride since launching in 2011.

The Data: In April 2017, Hubway held a Data Visualization Challenge at the Microsoft NERD Center in Cambridge, releasing 5 years of trip data.

The Question: What does the data tell us about the ride share program?

The Data Exploration/Question Refinement Cycle

Our original question:

'What does the data tell us about the ride share program?'

is a reasonable slogan to promote a hackathon. It is not good for guiding scientific investigation.

Before we can refine the question, we have to look at the data!

	seq_id	hubway_id	status	duration	start_date	strt_statn	end_date	end_statn	bike_nr	subsc_type	zip_code	birth_date	gender
0	1	8	Closed	9	7/28/2011 10:12:00	23.0	7/28/2011 10:12:00	23.0	B00468	Registered	'97217	1976.0	Male
1	2	9	Closed	220	7/28/2011 10:21:00	23.0	7/28/2011 10:25:00	23.0	B00554	Registered	'02215	1966.0	Male
2	3	10	Closed	56	7/28/2011 10:33:00	23.0	7/28/2011 10:34:00	23.0	B00456	Registered	'02108	1943.0	Male
3	4	11	Closed	64	7/28/2011 10:35:00	23.0	7/28/2011 10:36:00	23.0	B00554	Registered	'02116	1981.0	Female
4	5	12	Closed	12	7/28/2011 10:37:00	23.0	7/28/2011 10:37:00	23.0	B00554	Registered	'97214	1983.0	Female

Based on the data, what kind of questions can we ask?

The Data Exploration/Question Refinement Cycle

- **Who?** Who's using the bikes?

  Refine into specific hypotheses:

The Data Exploration/Question Refinement Cycle

- **Who?** Who's using the bikes?

  Refine into specific hypotheses:

  - More men or more women?

The Data Exploration/Question Refinement Cycle

- **Who?** Who's using the bikes?

  Refine into specific hypotheses:

  - More men or more women?
  - Older or younger people?

The Data Exploration/Question Refinement Cycle

- **Who?** Who's using the bikes?

  Refine into specific hypotheses:

  - More men or more women?
  - Older or younger people?
  - Subscribers or one time users?

The Data Exploration/Question Refinement Cycle

- **Where?** Where are bikes being checked out?

  Refine into specific hypotheses:

The Data Exploration/Question Refinement Cycle

- **Where?** Where are bikes being checked out?

  Refine into specific hypotheses:

  - More in Boston than Cambridge?

The Data Exploration/Question Refinement Cycle

- **Where?** Where are bikes being checked out?

  Refine into specific hypotheses:

  - More in Boston than Cambridge?
  - More in commercial or residential?

The Data Exploration/Question Refinement Cycle

- **Where?** Where are bikes being checked out?

  Refine into specific hypotheses:

  - More in Boston than Cambridge?
  - More in commercial or residential?
  - More around tourist attractions?

  Sometimes the data is given to you in pieces and must be merged!

The Data Exploration/Question Refinement Cycle

- **When?** When are the bikes being checked out?

  Refine into specific hypotheses:

The Data Exploration/Question Refinement Cycle

- **When?** When are the bikes being checked out?

  Refine into specific hypotheses:

  - More during the weekend than on the weekdays?

The Data Exploration/Question Refinement Cycle

- **When?** When are the bikes being checked out?

  Refine into specific hypotheses:

  - More during the weekend than on the weekdays?
  - More during rush hour?

The Data Exploration/Question Refinement Cycle

- **When?** When are the bikes being checked out?

  Refine into specific hypotheses:

  - More during the weekend than on the weekdays?
  - More during rush hour?
  - More during the summer than the fall?

  Sometimes the feature you want to explore doesn't exist in the data, and must be engineered!

The Data Exploration/Question Refinement Cycle

- **Why?** For what reasons/activities are people checking out bikes?

  Refine into specific hypotheses:

  - More bikes are used for recreation than commute?
  - More bikes are used for touristic purposes?
  - Bikes are use to bypass traffic?

  Do we have the data to answer these questions with reasonable certainty?

  What data do we need to collect in order to answer these questions?

The Data Exploration/Question Refinement Cycle

- **How?** Questions that combine variables.
  - How does user demographics impact the duration the bikes are being used? Or where they are being checked out?
  - How does weather or traffic conditions impact bike usage?
  - How do the characteristics of the station location affect the number of bikes being checked out?

How questions are about modeling relationships between different variables.

Inspirations for Data Viz/Exploration

So how well did we do in formulating creative hypotheses and manipulating the data for answers?

Check out the winners of the Hubway Challenge:

http://hubwaydatachallenge.org