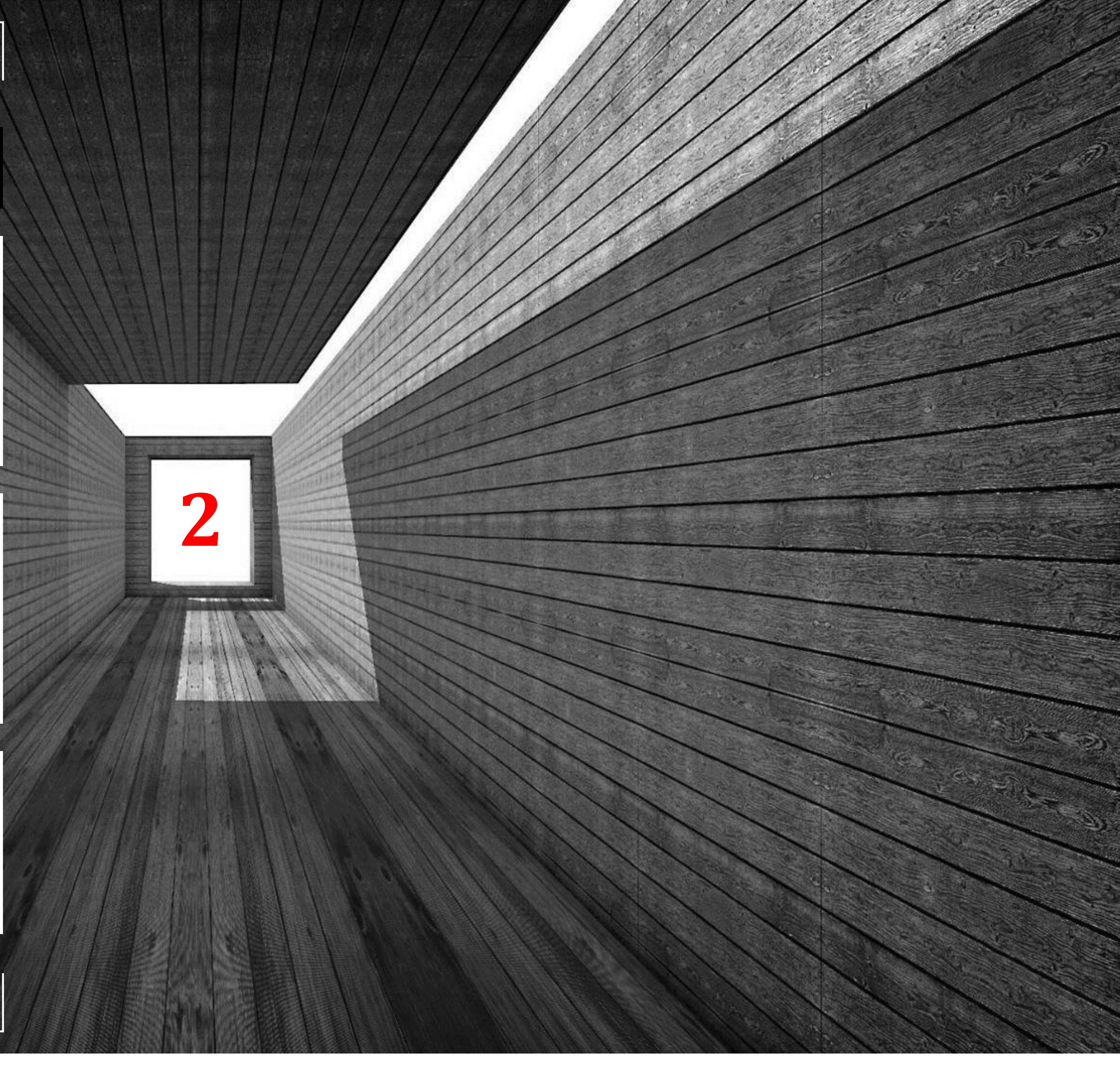
The German University in Cairo

CSEN1095 Data Engineering

Lecture 2 Explore Your Data I

Mervat Abuelkheir
mervat.abuelkheir@guc.edu.eg



Project Workflow

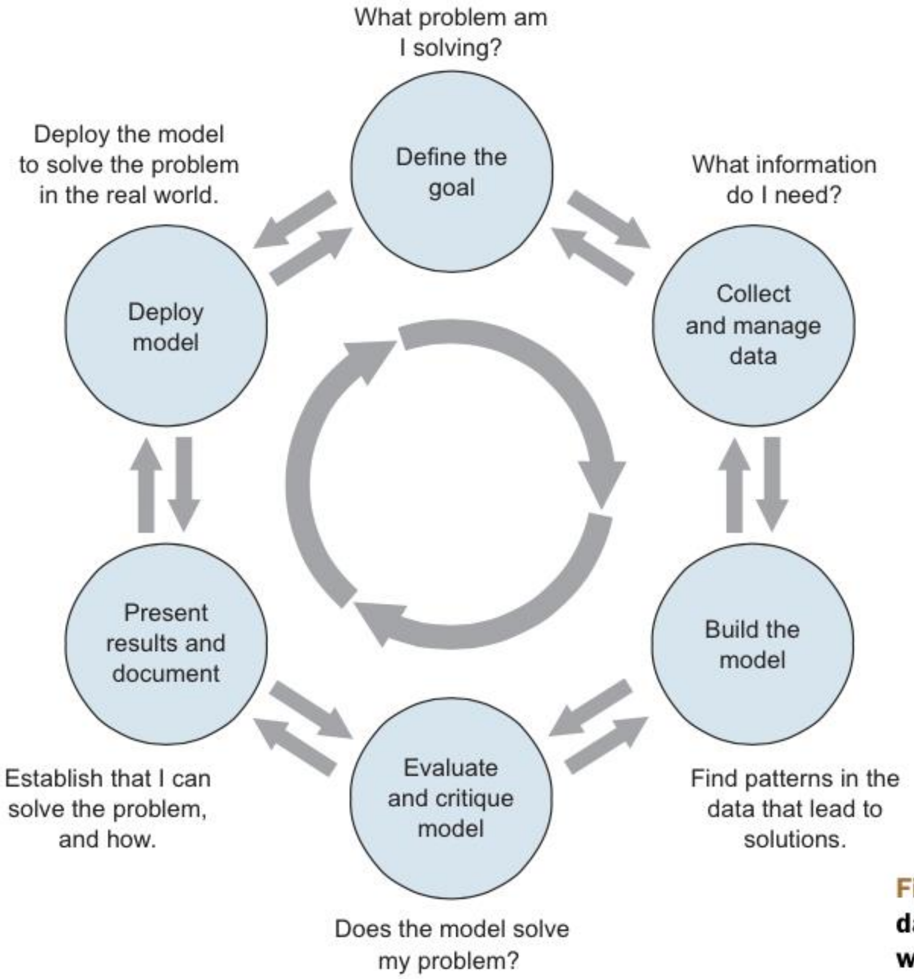

What problem am I solving?
Define the goal
What information do I need?
Collect and manage data
Build the model
Find patterns in the data that lead to solutions.
Evaluate and critique model
Does the model solve my problem?
Present results and document
Establish that I can solve the problem, and how.
Deploy model
Deploy the model to solve the problem in the real world.

Figure 1.1 The lifecycle of a data science project: loops within loops

Source: CMSC320 Data Science Course

Define The Goal

- What is the question/problem?
- Who wants to answer/solve it?
- What do they know/do now?
- How well can we expect to answer/solve it?
- How well do they want us to answer/solve it?

Project Workflow

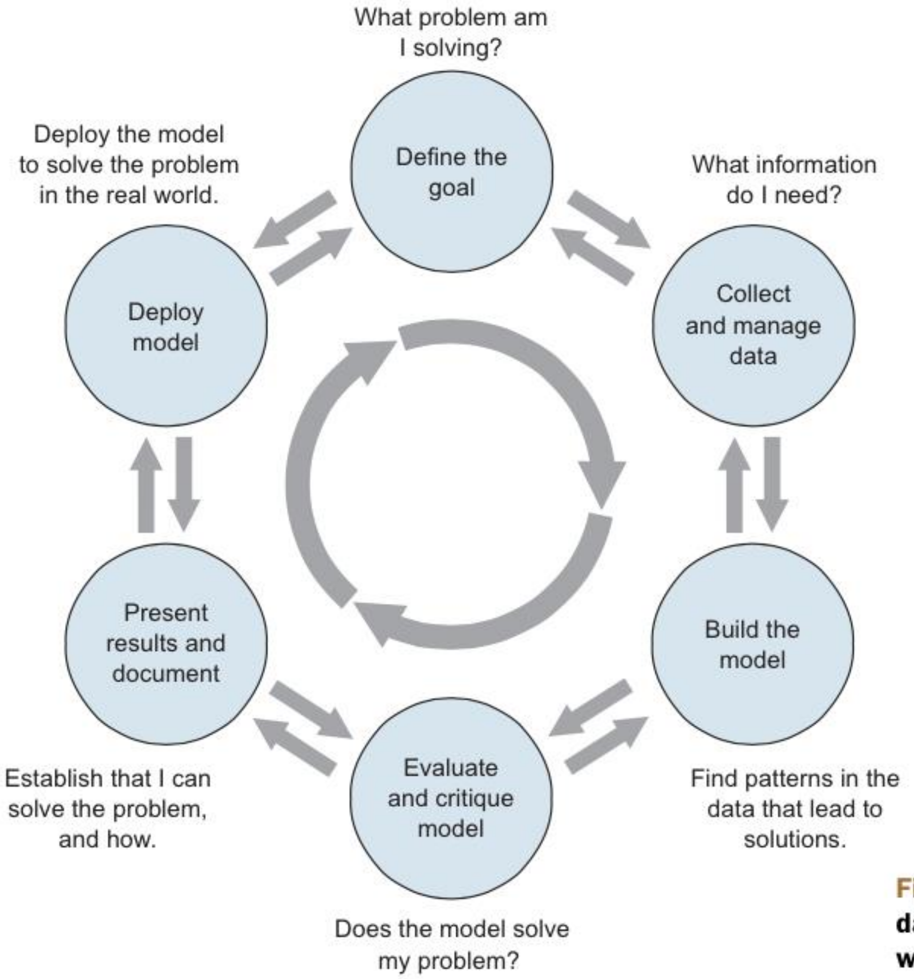

What problem am I solving?
Define the goal
What information do I need?
Collect and manage data
Build the model
Find patterns in the data that lead to solutions.
Evaluate and critique model
Does the model solve my problem?
Present results and document
Establish that I can solve the problem, and how.
Deploy model
Deploy the model to solve the problem in the real world.

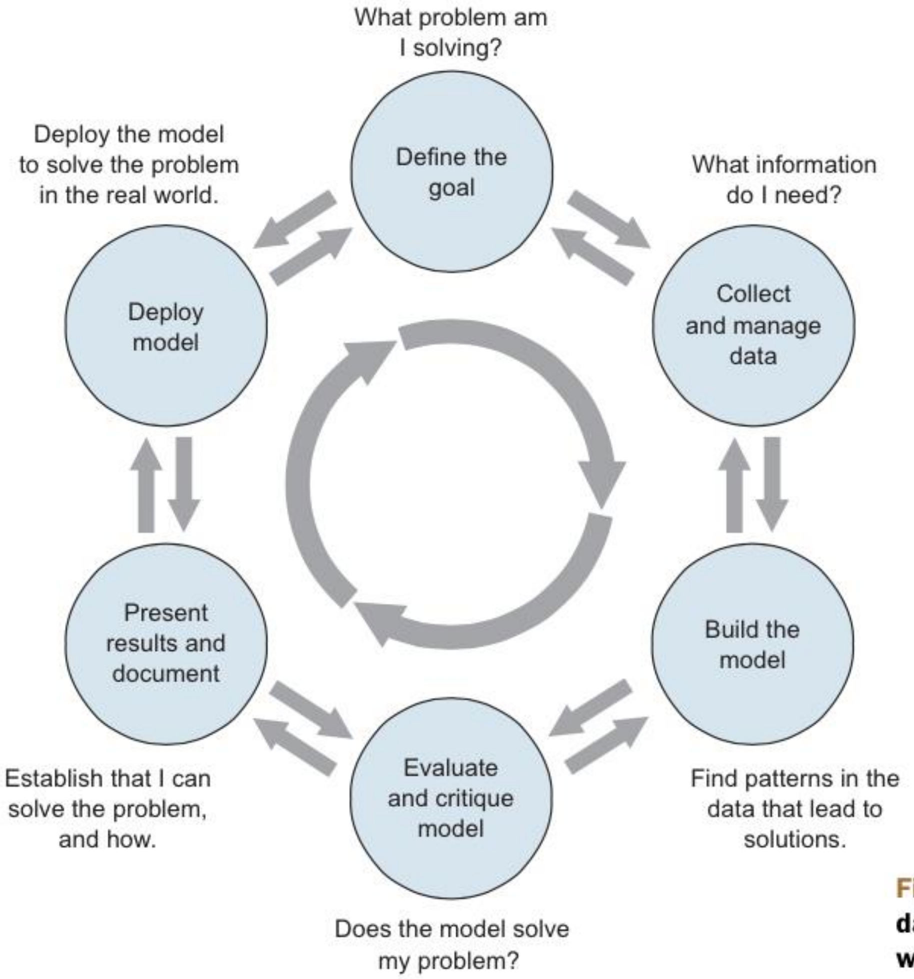
Figure 1.1 The lifecycle of a data science project: loops within loops

Source: CMSC320 Data Science Course

Data Collection and Management

- What data is available?
- Is it good enough?
- Is it enough?
- What are sensible measurements to derive from this data? Units, transformations, rates, ratios, etc.

Project Workflow



What problem am I solving?
Define the goal
What information do I need?
Collect and manage data
Build the model
Find patterns in the data that lead to solutions.
Evaluate and critique model
Does the model solve my problem?
Present results and document
Establish that I can solve the problem, and how.
Deploy model
Deploy the model to solve the problem in the real world.

Figure 1.1 The lifecycle of a data science project: loops within loops

Source: CMSC320 Data Science Course

Modeling

- What kind of problem is it? e.g., classification, clustering, regression, etc.
- What kind of model should I use?
- Do I have enough data for it?
- Does it really answer the question?

Project Workflow

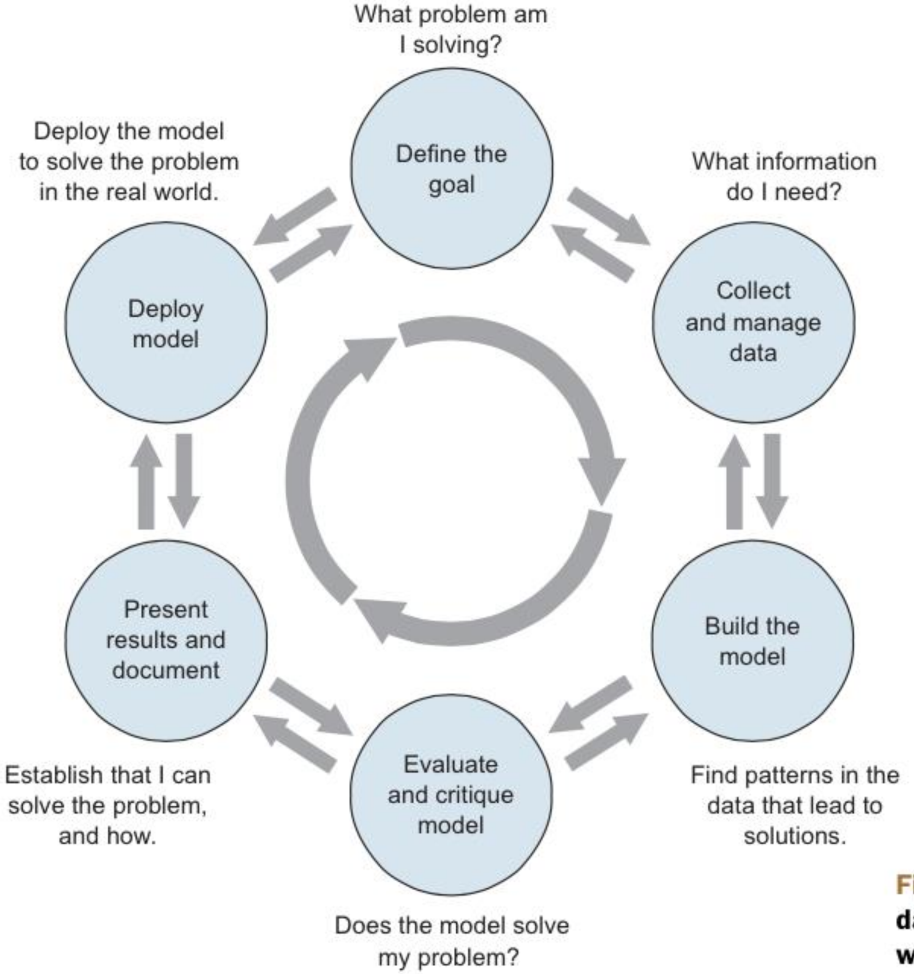

What problem am I solving?
Define the goal
What information do I need?
Collect and manage data
Build the model
Find patterns in the data that lead to solutions.
Evaluate and critique model
Does the model solve my problem?
Present results and document
Establish that I can solve the problem, and how.
Deploy model
Deploy the model to solve the problem in the real world.

Figure 1.1 The lifecycle of a data science project: loops within loops

Source: CMSC320 Data Science Course

Model Evaluation

- Did it work? How well?
- Can I interpret the model?
- What have I learned?

Project Workflow

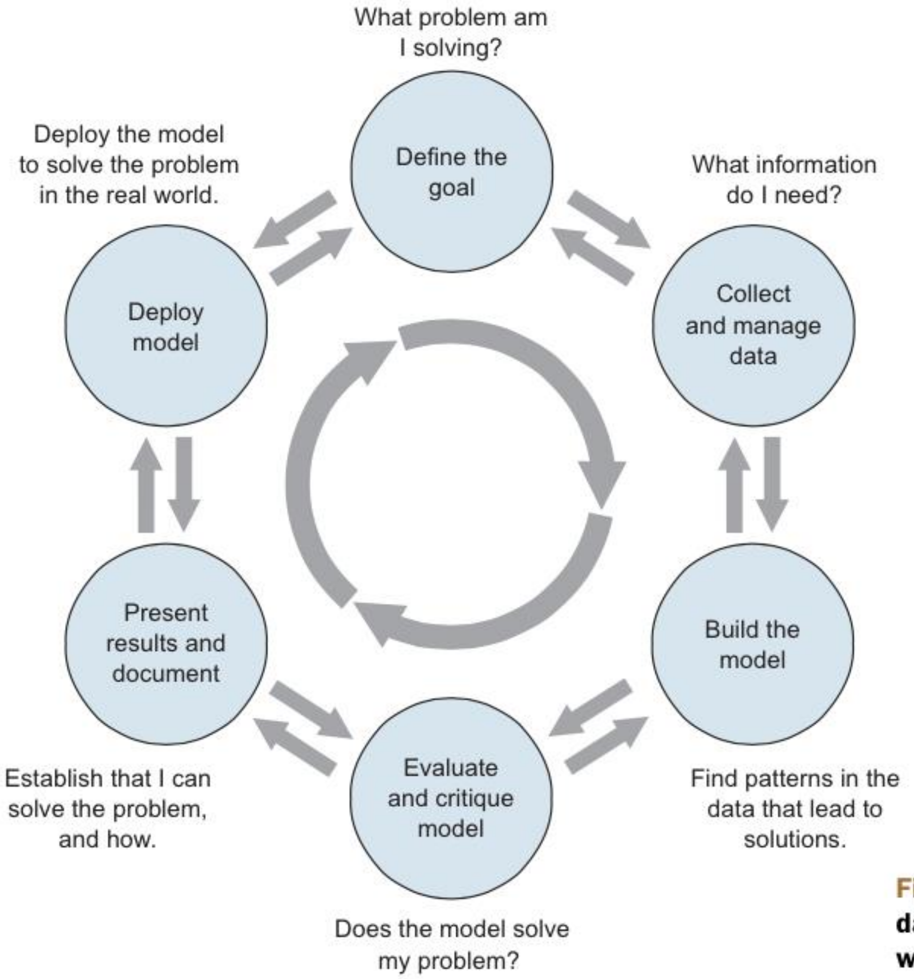

What problem am I solving?
Define the goal
What information do I need?
Collect and manage data
Build the model
Find patterns in the data that lead to solutions.
Evaluate and critique model
Does the model solve my problem?
Present results and document
Establish that I can solve the problem, and how.
Deploy model
Deploy the model to solve the problem in the real world.

Figure 1.1 The lifecycle of a data science project: loops within loops

Presentation

- Again, what are the measurements that tell the real story?
- How can I describe and visualize them effectively?

Source: CMSC320 Data Science Course

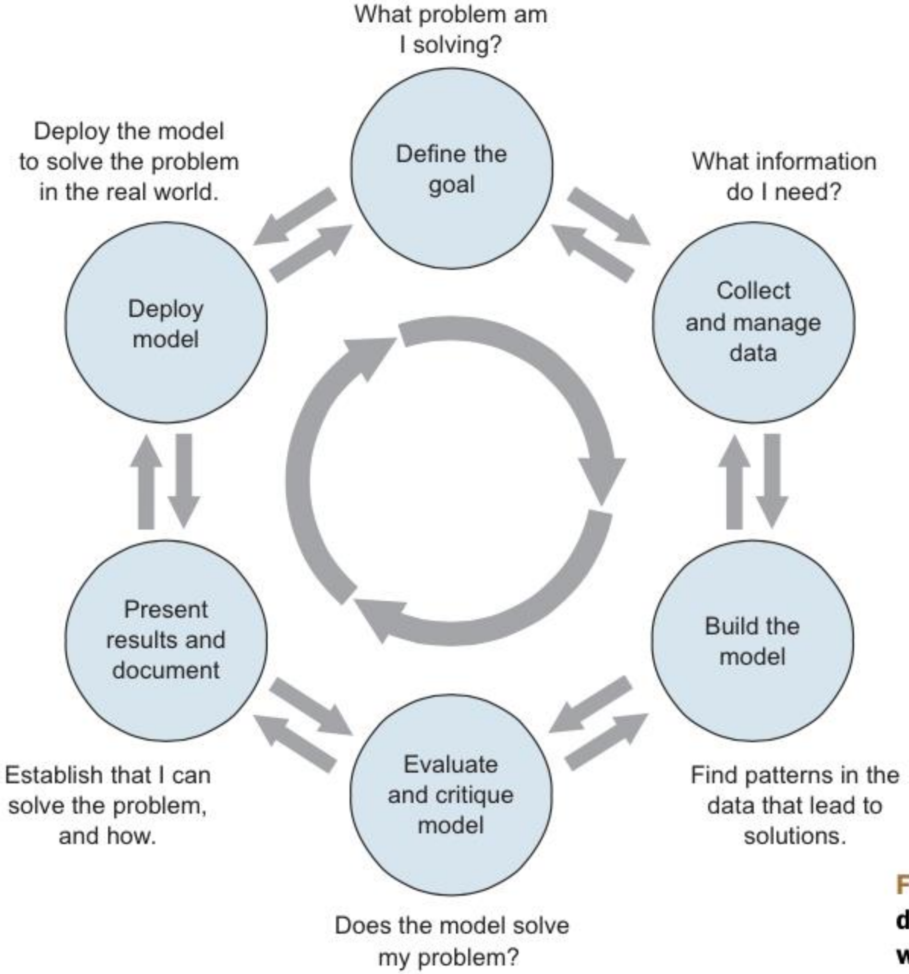

What problem am I solving?
Define the goal
What information do I need?
Collect and manage data
Build the model
Find patterns in the data that lead to solutions.
Evaluate and critique model
Does the model solve my problem?
Present results and document
Establish that I can solve the problem, and how.
Deploy model
Deploy the model to solve the problem in the real world.

Figure 1.1 The lifecycle of a data science project: loops within loops

Deployment

- Where will it be hosted?
- Who will use it?
- Who will maintain it?

Source: CMSC320 Data Science Course

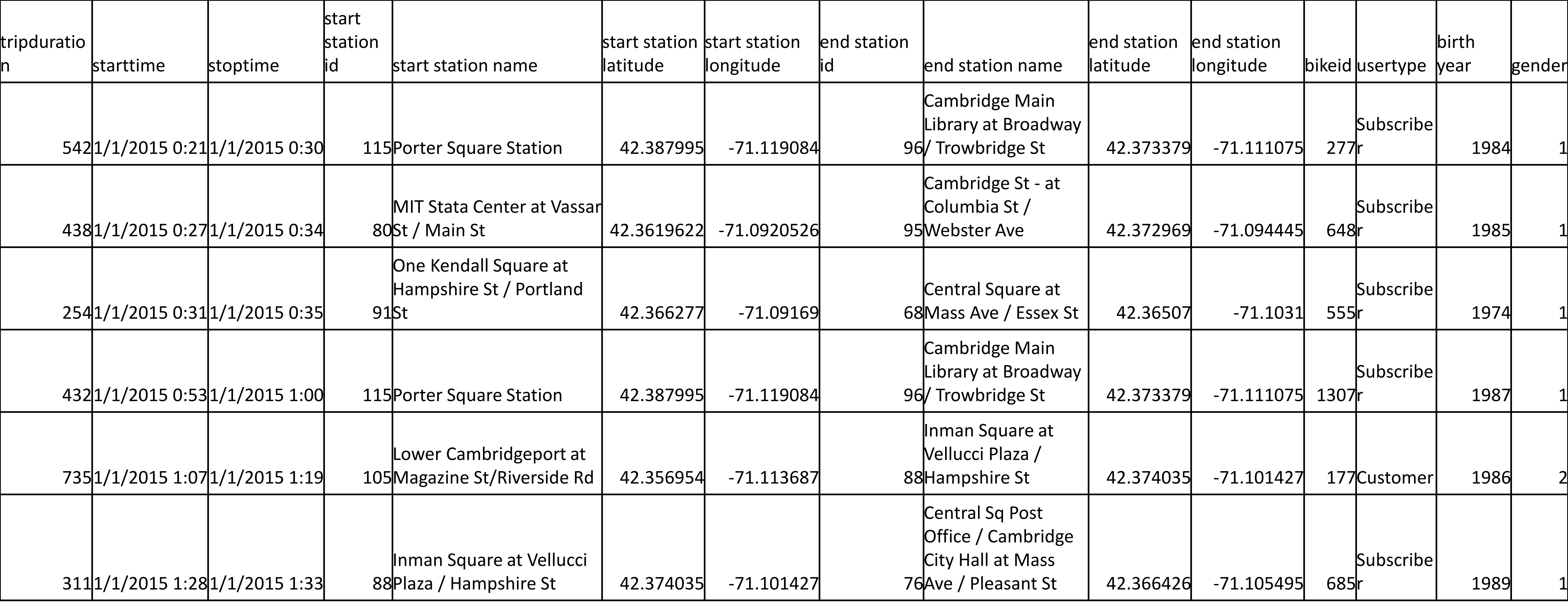
Data Challenges

- Massive data (500k users, 20k movies, 100m ratings)
- Curse of dimensionality (very high-dimensional problem)
- Missing data values (sometimes not missing at random)
- Wrong data values (needs detection and correction)
- Sometimes data is not factual (yet not technically wrong!) and we have a complicated set of factors that affect user-provided data values
- Need to avoid overfitting (test data vs. training data)

Source: CS109 Stanford's Data Science Course

Boston's Hubway Data Challenge

Winner:
http://zsobhani.github.io/hubway-team-viz/



tripduration	starttime	stoptime	start station id	start station name	start station latitude	start station longitude	end station id	end station name	end station latitude	end station longitude	bikeid	usertype	birth year	gender
542	1/1/2015 0:21	1/1/2015 0:30	115	Porter Square Station	42.387995	-71.119084	96	Cambridge Main Library at Broadway / Trowbridge St	42.373379	-71.111075	277	Subscriber	1984	1
438	1/1/2015 0:27	1/1/2015 0:34	80	MIT Stata Center at Vassar St / Main St	42.3619622	-71.0920526	95	Cambridge St - at Columbia St / Webster Ave	42.372969	-71.094445	648	Subscriber	1985	1
254	1/1/2015 0:31	1/1/2015 0:35	91	One Kendall Square at Hampshire St / Portland St	42.366277	-71.09169	68	Central Square at Mass Ave / Essex St	42.36507	-71.1031	555	Subscriber	1974	1
432	1/1/2015 0:53	1/1/2015 1:00	115	Porter Square Station	42.387995	-71.119084	96	Cambridge Main Library at Broadway / Trowbridge St	42.373379	-71.111075	1307	Subscriber	1987	1
735	1/1/2015 1:07	1/1/2015 1:19	105	Lower Cambridgeport at Magazine St/Riverside Rd	42.356954	-71.113687	88	Inman Square at Vellucci Plaza / Hampshire St	42.374035	-71.101427	177	Customer	1986	2
311	1/1/2015 1:28	1/1/2015 1:33	88	Inman Square at Vellucci Plaza / Hampshire St	42.374035	-71.101427	76	Central Sq Post Office / Cambridge City Hall at Mass Ave / Pleasant St	42.366426	-71.105495	685	Subscriber	1989	1

Half a million Hubway rides from 2011 to 2013!

‘What does the data tell us about Boston’s ride share program?’

Data Exploration/Question Refinement

- **Who?** Who's using the bikes?
  - More men or more women?
  - Older or younger people?
  - Subscribers or one time users?

- **Where?** Where are bikes being checked out?
  - More in Boston than Cambridge?
  - More in commercial or residential?
  - More around tourist attractions?

Source: CS109 Stanford's Data Science Course

Data Exploration/Question Refinement

- **When?** When are the bikes being checked out?
  - More during the weekend than on the weekdays?
  - More during rush hour?
  - More during the summer than the fall?

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

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

Source: CS109 Stanford's Data Science Course

Data Exploration/Question Refinement

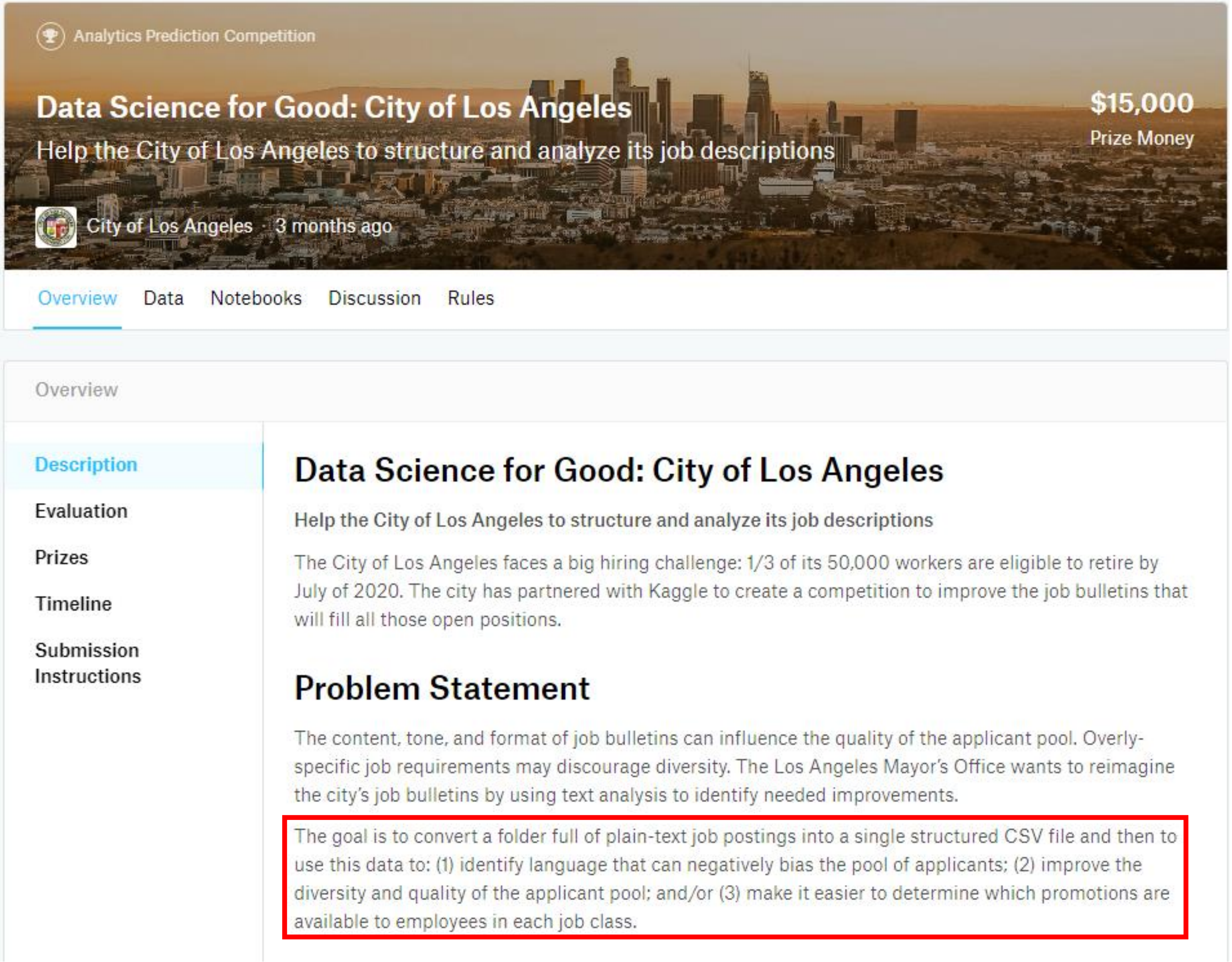
- **How?** Questions that investigate/model relationships between variables
  - How does user demographics impact the duration the bikes are being used? Or where they are being checked out?
  - How does weather or traffic conditions impact bike usage?
  - How do the characteristics of the station location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

- *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?*
- *Sometimes the feature you want to explore doesn't exist in the data, and must be **engineered**!*
- *Sometimes the data is given to you in pieces and must be **merged**!*

Source: CS109 Stanford's Data Science Course

Data Representations

- Tabular – Ideal for ML!
- Structured – XML, JSON, …
- Semi-structured – graph, DNA, …
- Unstructured – images, text, video, …



Data Science for Good: City of Los Angeles

Help the City of Los Angeles to structure and analyze its job descriptions

$15,000

Prize Money

City of Los Angeles · 3 months ago

Data Science for Good: City of Los Angeles

Help the City of Los Angeles to structure and analyze its job descriptions

The City of Los Angeles faces a big hiring challenge: 1/3 of its 50,000 workers are eligible to retire by July of 2020. The city has partnered with Kaggle to create a competition to improve the job bulletins that will fill all those open positions.

Problem Statement

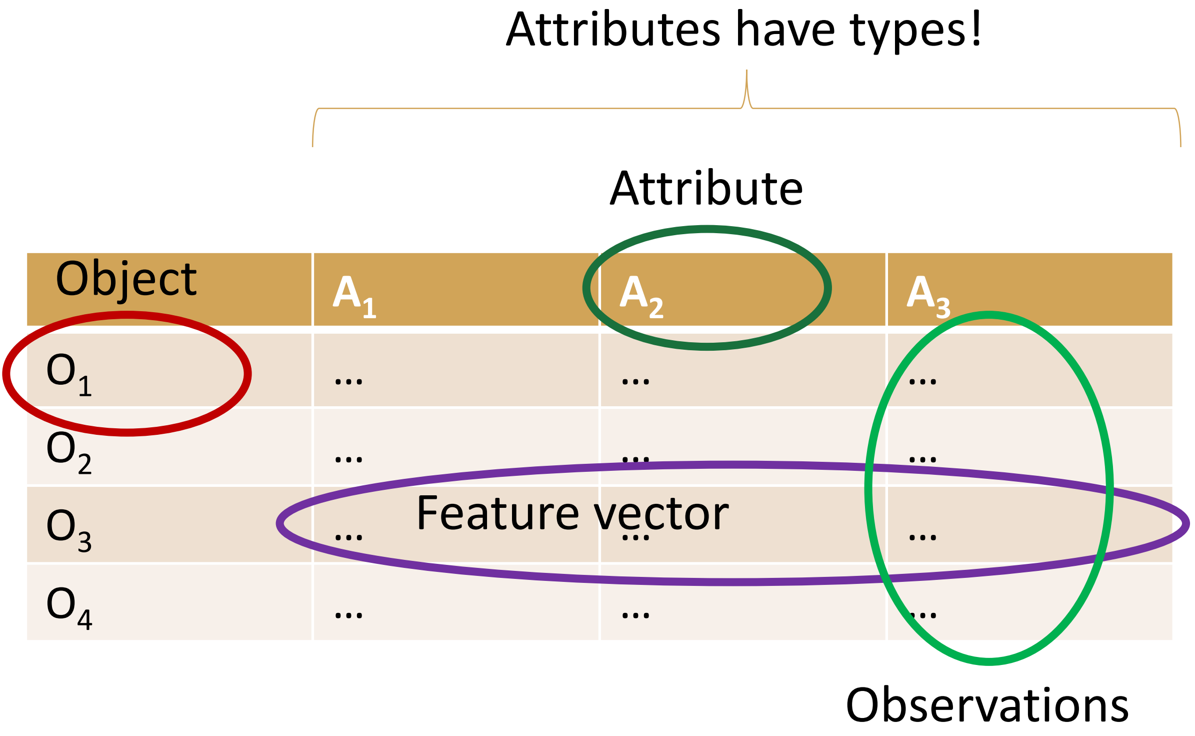
The content, tone, and format of job bulletins can influence the quality of the applicant pool. Overly-specific job requirements may discourage diversity. The Los Angeles Mayor's Office wants to reimagine the city's job bulletins by using text analysis to identify needed improvements.

The goal is to convert a folder full of plain-text job postings into a single structured CSV file and then to use this data to: (1) identify language that can negatively bias the pool of applicants; (2) improve the diversity and quality of the applicant pool; and/or (3) make it easier to determine which promotions are available to employees in each job class.

https://www.kaggle.com/c/data-science-for-good-city-of-los-angeles

Objects and Attributes

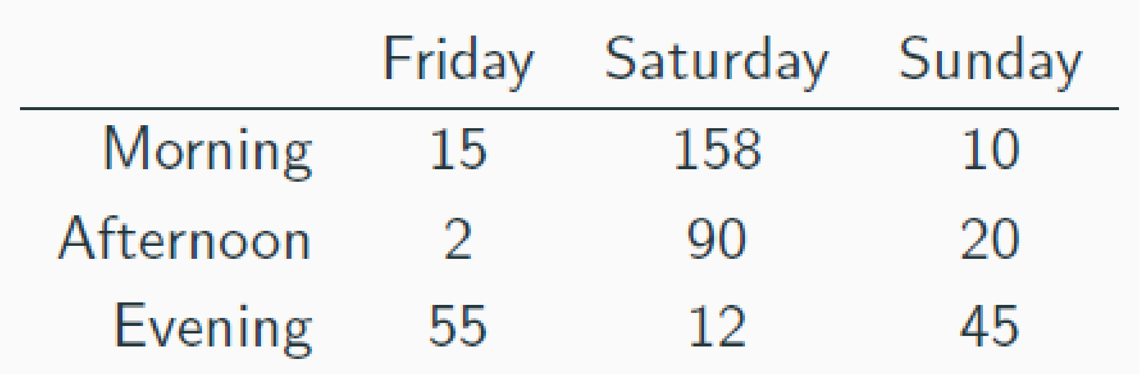
- A **data object** represents an entity
  - Also sample, example, **instance**, data point
  - e.g. *customers, students, patients, books*
- An **attribute** is a data field, representing a characteristic or feature of a data object
  - Also **dimension**, **feature**, and **variable**
  - e.g. *name, age, salary, gender, grade, …*
  - Attribute/feature vector → A set of attributes that describe an object
  - Observed values for an attribute → **observations**



Object	A_1	A_2	A_3
O_1	…	…	…
O_2	…	…	…
O_3	…	…	…
O_4	…	…	…

Interjection – Correct Tabular Data

The following is a table for the number of produce deliveries over a weekend



	Friday	Saturday	Sunday
Morning	15	158	10
Afternoon	2	90	20
Evening	55	12	45

- What are the variables in this dataset?

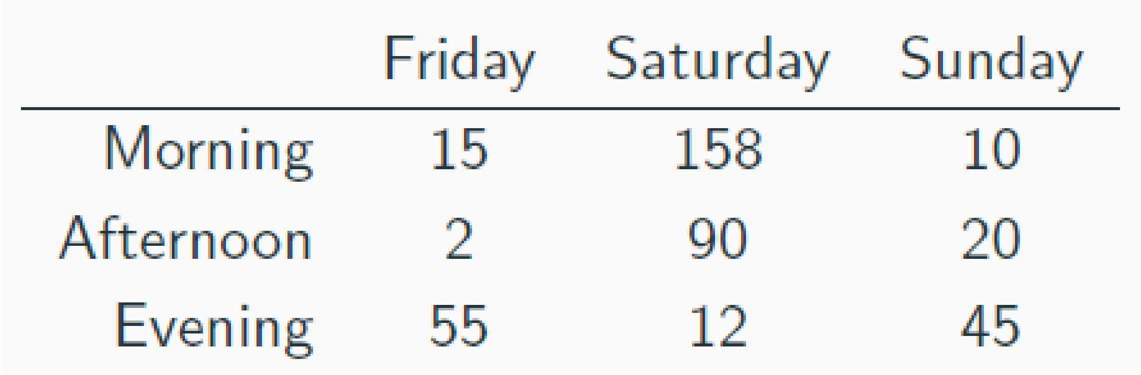
Variables should be: Time, Day, # Produce Deliveries

- What object or event are we measuring?
- What's the issue? How do we fix it?

Source: CS109 Stanford's Data Science Course

Interjection – Correct Tabular Data

The following is a table for the number of produce deliveries over a weekend



	Friday	Saturday	Sunday
Morning	15	158	10
Afternoon	2	90	20
Evening	55	12	45

- What are the variables in this dataset?
- What object or event are we measuring?
- What's the issue? How do we fix it?

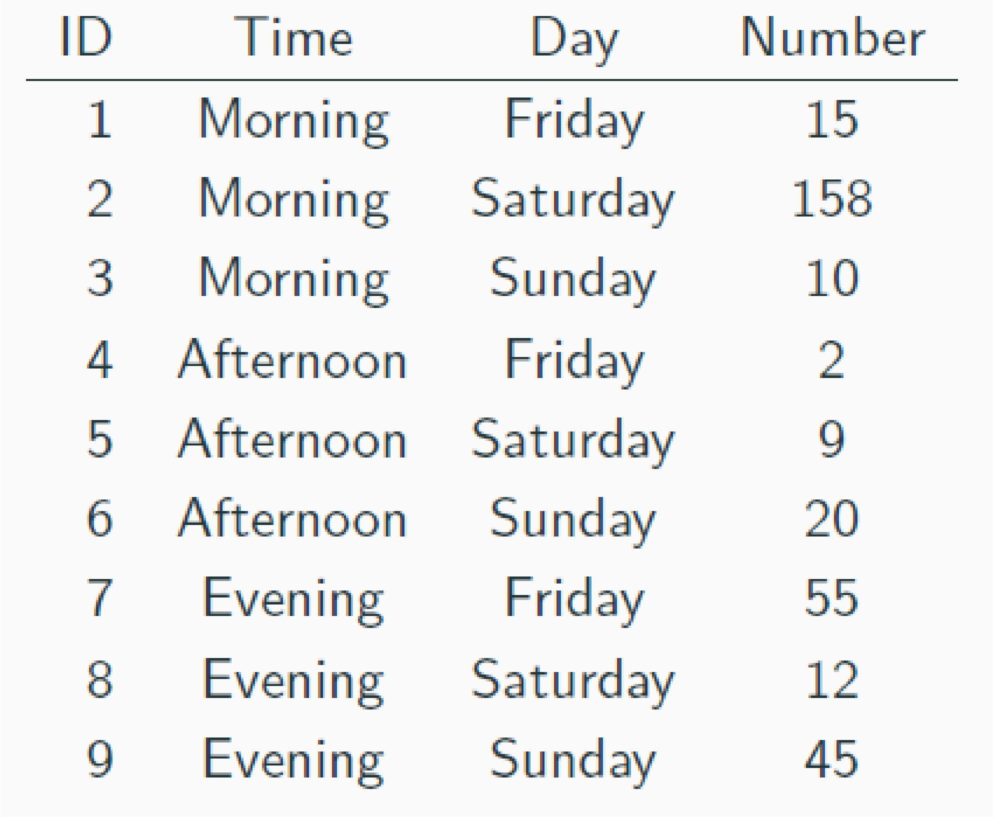
Variables should be: Time, Day, # Produce Deliveries

- Each column header represents a value, not a variable
- The values of the variable “# Produce Deliveries” are not recorded in a single column

Source: CS109 Stanford’s Data Science Course

Interjection – Correct Tabular Data

Reorganize the data to make explicit the event we're observing and the variables associated to this event



ID	Time	Day	Number
1	Morning	Friday	15
2	Morning	Saturday	158
3	Morning	Sunday	10
4	Afternoon	Friday	2
5	Afternoon	Saturday	9
6	Afternoon	Sunday	20
7	Evening	Friday	55
8	Evening	Saturday	12
9	Evening	Sunday	45

Source: CS109 Stanford's Data Science Course

Interjection – Things to Consider ..

- Are column headers values and not variable names?
- Are variables stored in both rows and columns?
- Are multiple variables stored in one column?
- Are multiple types of experimental units stored in the same table?

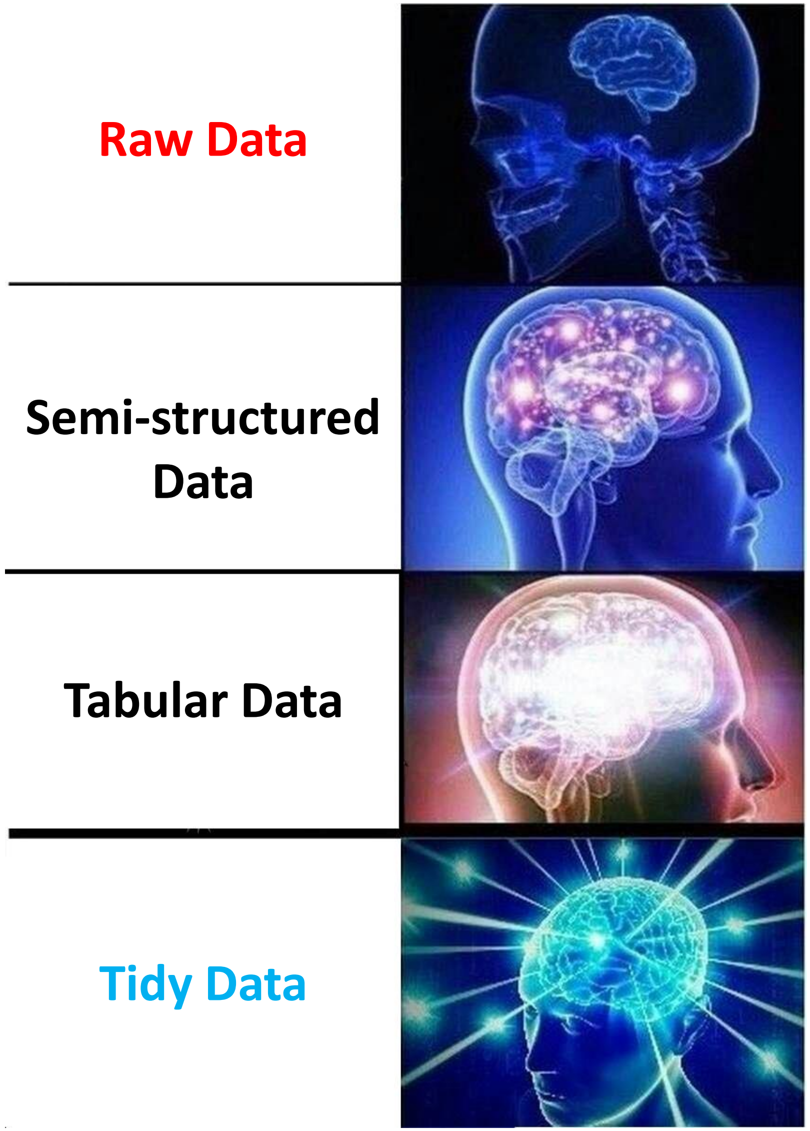
- In general, ***we want each file to correspond to a dataset, each column to represent a single variable and each row to represent a single observation***
- We want to tabularize the data. This makes Python happy!

Source: CS109 Stanford's Data Science Course

Exploratory Data Analysis – How To

- Each row describes a single object
- Each column describes a property of that object
- Columns are numeric whenever appropriate
- Column values have same measurement unit
- Columns contain atomic properties that cannot be further decomposed

This is Tidy Data



Raw Data

Semi-structured Data

Tabular Data

Tidy Data

Source: CS109 Stanford's Data Science Course

Attribute Types

Qualitative Attributes

- Categorical/Nominal
- Binary
- Ordinal

Quantitative Attributes

- Numeric

Attribute Types

Qualitative Attributes

- Most algorithms are designed to work with numbers!
- *Qualitative attributes may need to be encoded into numbers*

Categorical/Nominal

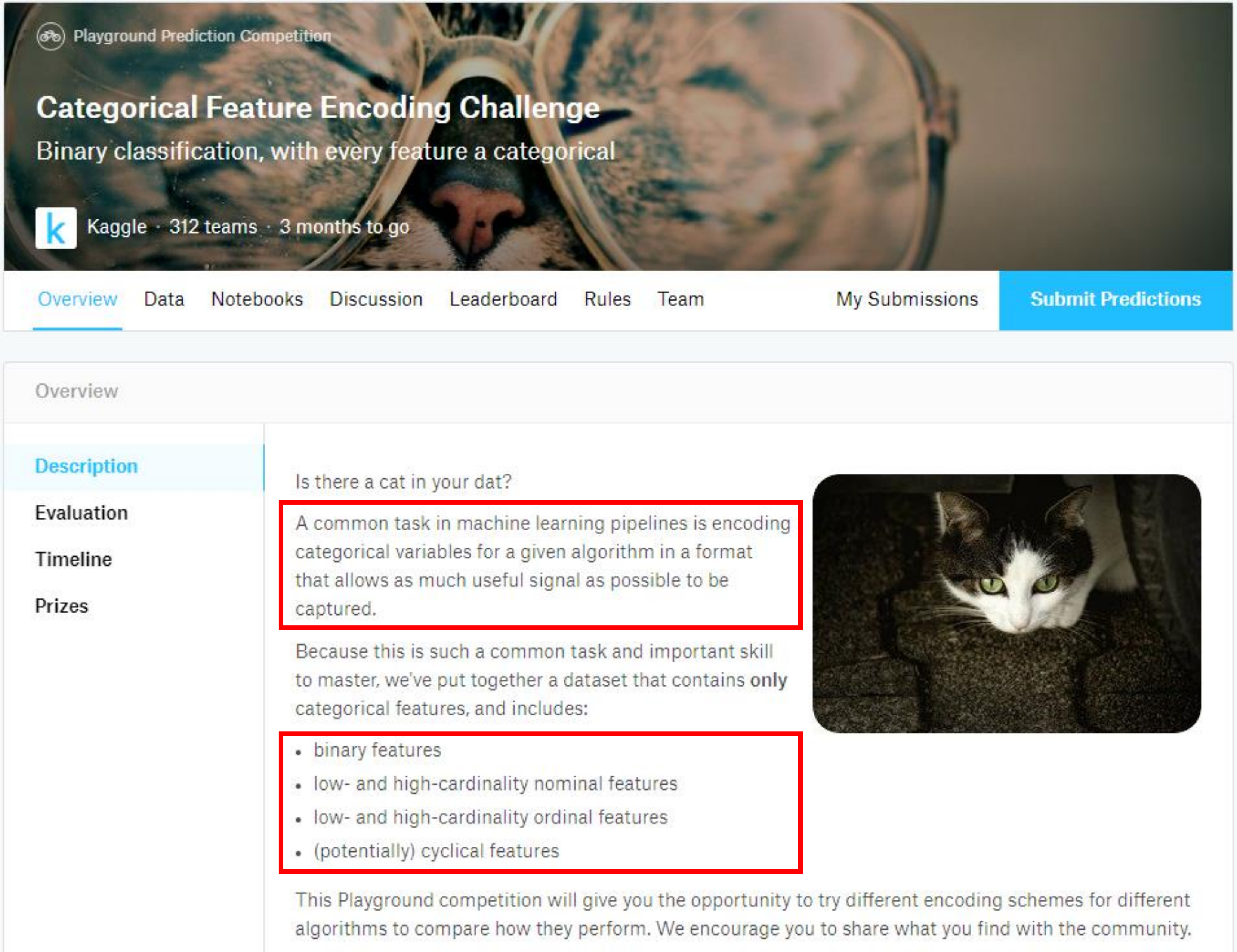
- Each value represents *category*, *code*, or *state*
- e.g. *hair color, marital status, customer ID*
- Possible to be represented as numbers (*coding*)

Binary

- Nominal with only two values; *two states* or *categories*: 0 or 1 (absent or present, true or false)
- Symmetric: both states are equally valuable and have the same weight
  - e.g. *gender*
- Asymmetric: states are not equally important
  - e.g. *medical test outcomes – +ve or -ve* (Which outcome should take 1?)

Ordinal

- Values have a meaningful order or *ranking*, magnitude between successive values is not known
- e.g. *professional rank, grade, size, customer satisfaction*



Categorical Feature Encoding Challenge

Binary classification, with every feature a categorical

Kaggle · 312 teams · 3 months to go

Is there a cat in your dat?

A common task in machine learning pipelines is encoding categorical variables for a given algorithm in a format that allows as much useful signal as possible to be captured.

Because this is such a common task and important skill to master, we've put together a dataset that contains **only** categorical features, and includes:

- binary features
- low- and high-cardinality nominal features
- low- and high-cardinality ordinal features
- (potentially) cyclical features

This Playground competition will give you the opportunity to try different encoding schemes for different algorithms to compare how they perform. We encourage you to share what you find with the community.

https://www.kaggle.com/c/cat-in-the-dat/overview

Attribute Types

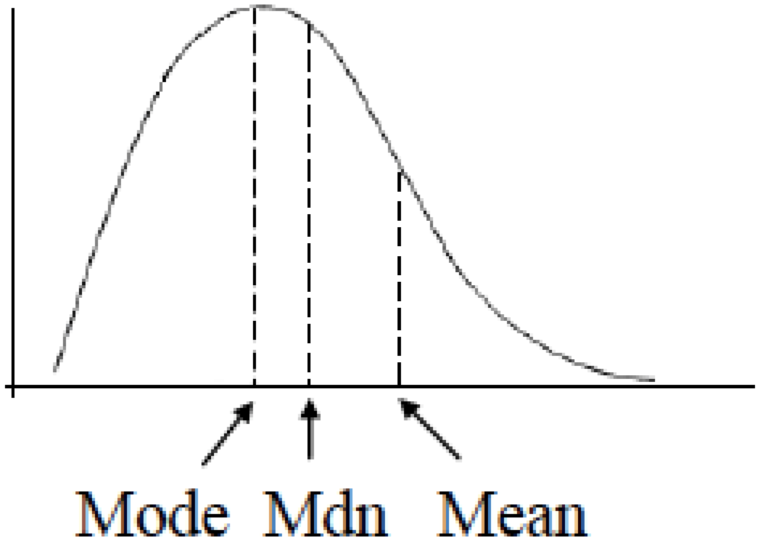
- **Interval-scaled**
  - Measured on a *scale of equal-size units*
  - e.g. *temperature, year*
  - Do not have a true zero point
  - Not possible to be expressed as multiples
- **Ratio-scaled**
  - Have a true zero point
  - A value can be expressed as a *multiple* of another
  - e.g. *years of experience, weight, salary*

Quantitative Attributes

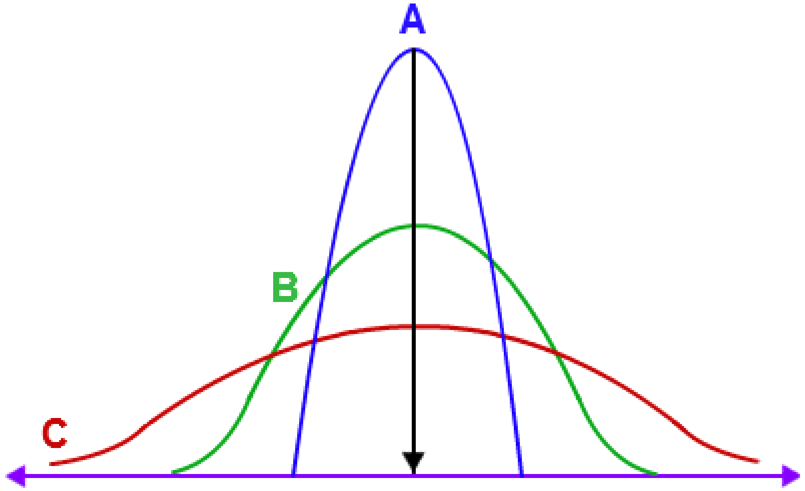
- Sometimes we need to normalize quantitative data
- Sometimes we need to discretize quantitative data – Back to categorical!

Basic Statistical Descriptions of Data

Measuring Central Tendency



Measuring dispersion of Data



Measuring Central Tendency

Population versus sample:

- A **population** is the entire set of objects or events under study. Population can be hypothetical "all students" or all students in this class
- A **sample** is a "**representative**" subset of the objects or events under study. Needed because it's sometimes impossible or intractable to obtain or compute with population data

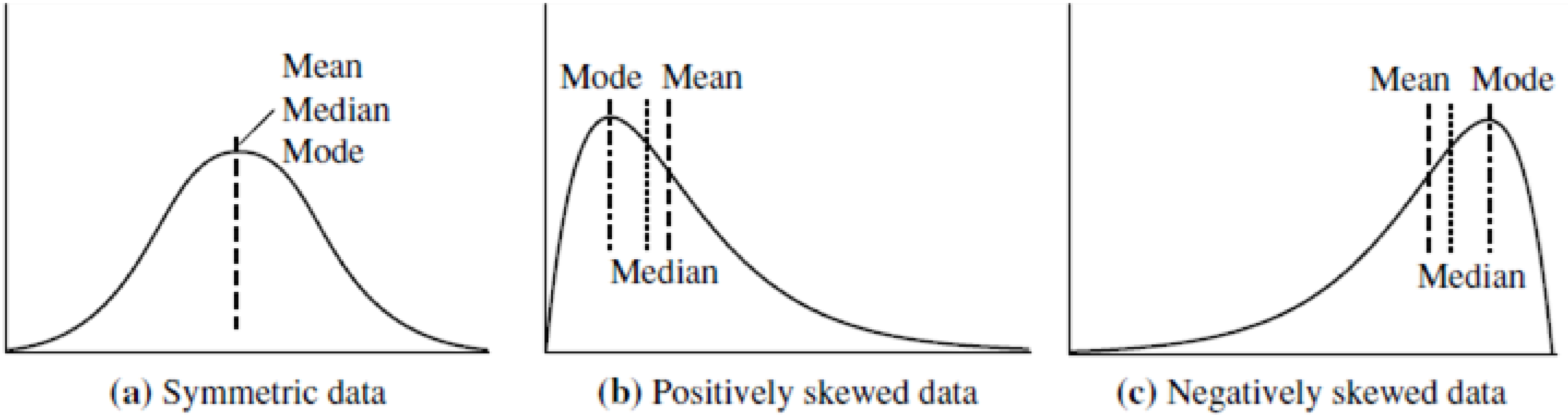
Measuring Central Tendency

For *N* observations of numerical variable X: $x_1, x_2, \ldots, x_N$

- **Mean**: or *average* of values
  - $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \ldots + x_N}{N}$
- **Weighted Average**: a *weight* is associated with each value
  - $\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{N} = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_N x_N}{N}$
- Problem: sensitivity to outlier values
  - e.g. *mean salary, mean student score*
  - *Trimmed mean* → chop off extreme values at both ends
- There is always uncertainty involved when calculating a sample mean to estimate a population mean

Measuring Central Tendency

- **Median**: *middle value* in set of ordered values
  - *N* is odd → median is middle value of ordered set
  - *N* is even → median is not unique → average of two middlemost values
  - Expensive to compute for large # of observations
- **Mode**: value that occurs *most frequently* in the attribute values
  - Works for both **qualitative** and quantitative attributes
  - Data can be *unimodal*, *bimodal*, or *trimodal*
    - ➢ No mode?

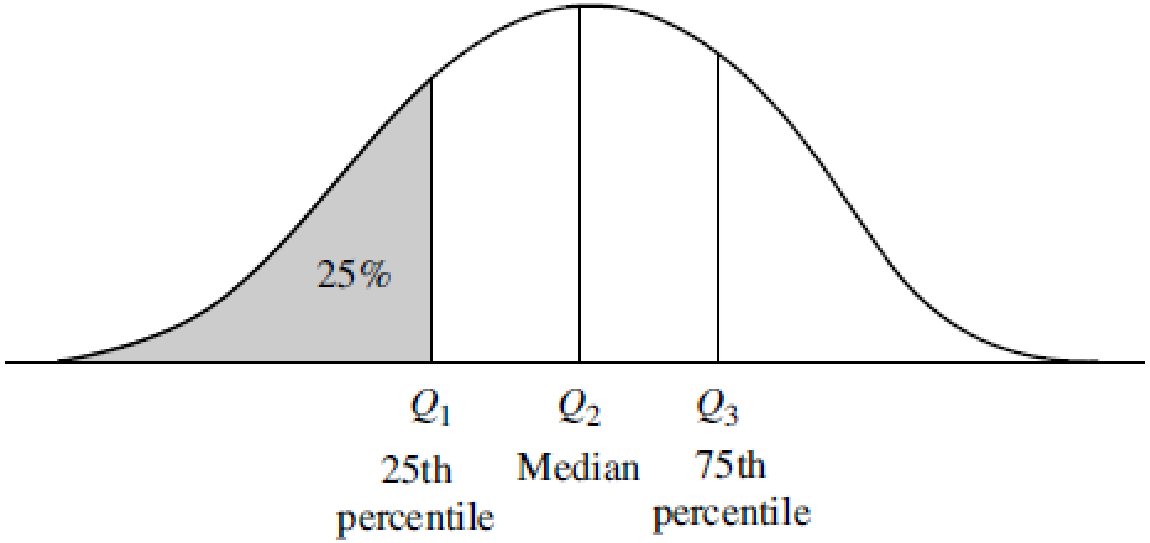


Measuring Dispersion of Data

The spread of a sample of observations measures how well the mean or median describes the sample

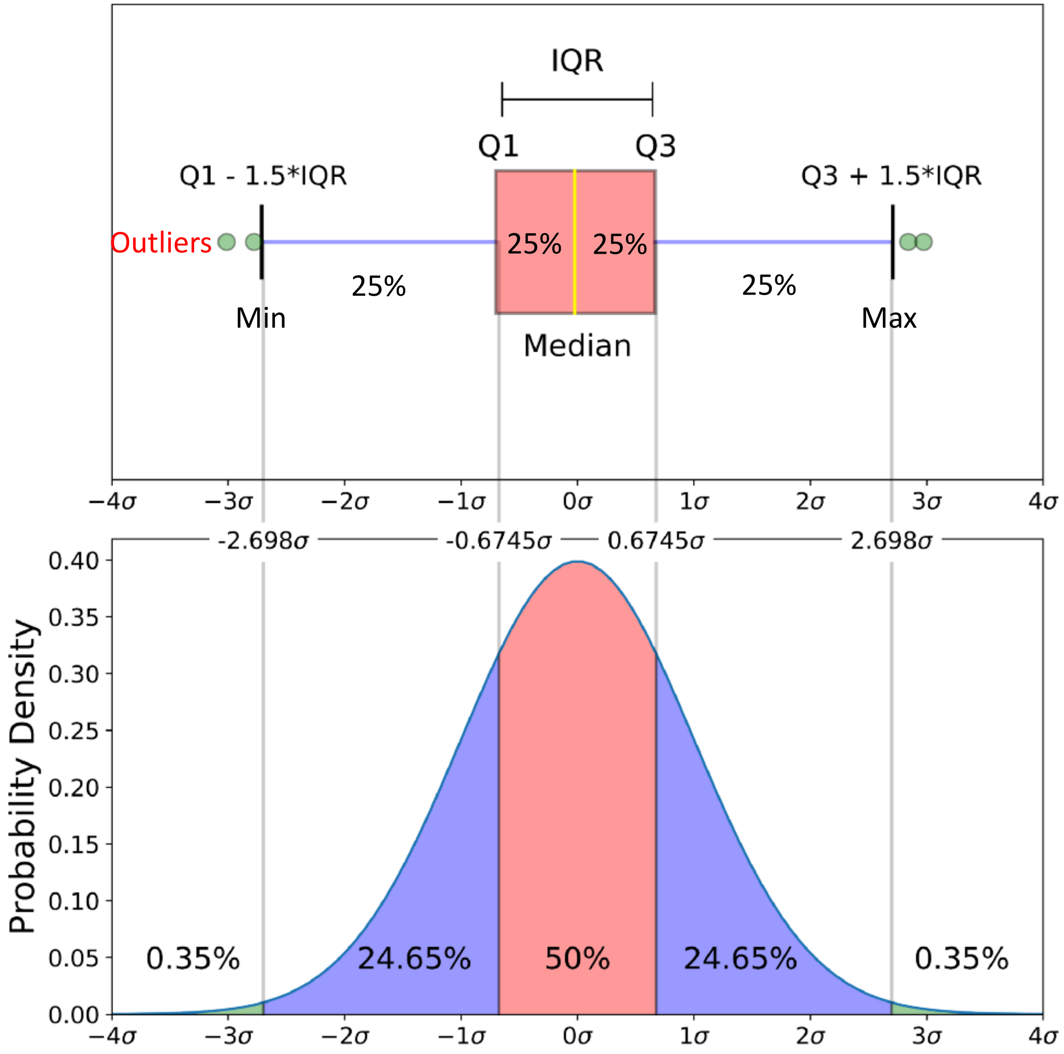
For *N* observations of numerical variable *X:* $x_1, x_2, \ldots, x_N$

- **First, we order the observations! Then, we can compute …**
- **Range**: *difference* between the largest and smallest values
- **Quantiles**: points taken at *regular intervals* of a data distribution, dividing it into (almost) equal-size consecutive sets
  - Most famous → *percentile*
    - 100 equal-sized sets
  - Quartiles → 4 Quantiles
- **Interquartile Range:** = *Q3 - Q1*



Measuring Dispersion of Data

- **Five-Number Summary**:
  - Min, Q1, Median (Q2), Q3, Max
- **Boxplots**: *visualization* for the five-number summary
  - *Whiskers* terminate at *min* & *max* **OR** the most extreme observations within
    - 1.5 × *IQR* of the quartiles →
      - Lower whisker: Min OR Q1 – (1.5 × *IQR*)
      - Upper whisker: Max OR Q3 + (1.5 × *IQR*)
    - **Remaining points are plotted individually (outliers!)**



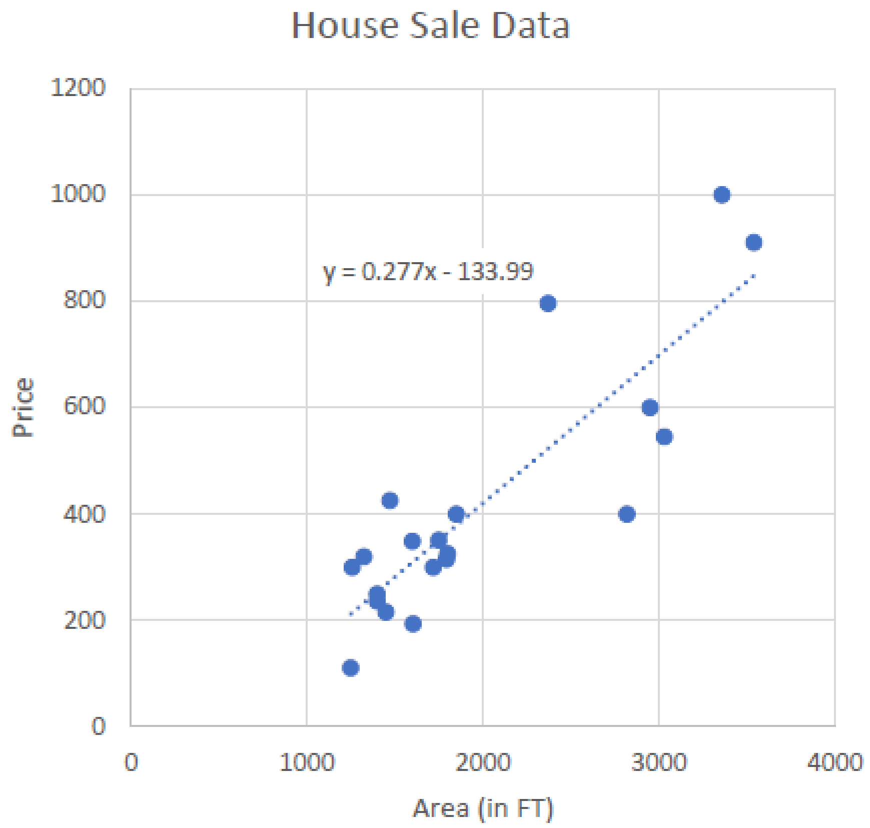
Working Example: https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/box-plot-review

Measuring Dispersion of Data

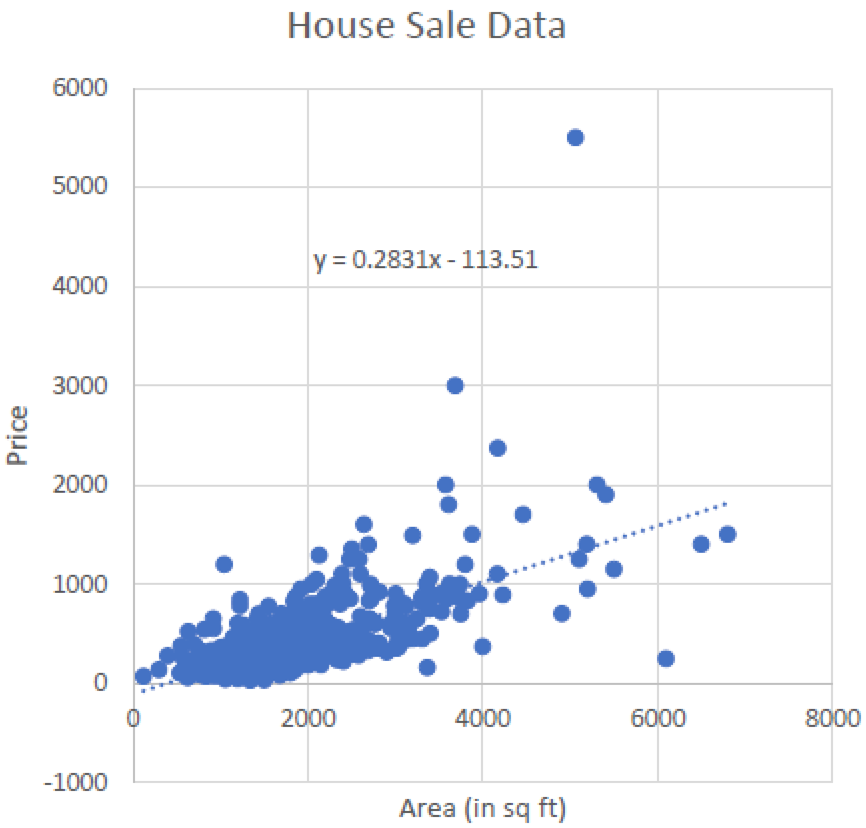
- **Variance & SD**: indicate *how spread out* a data distribution is
  - *Low SD* → data observations tend to be very close to the mean
  - *High SD* → data is spread out over a large range of values
  - $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} {x_i}^2\right) - \bar{x}^2$
  - $SD = \sigma$

Careful with Estimations of Centrality and dispersion Parameters!

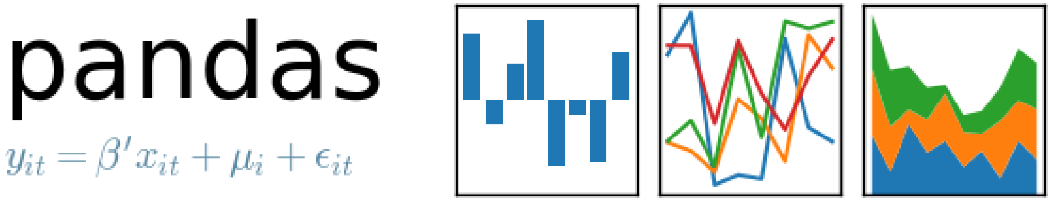
The first 20 points

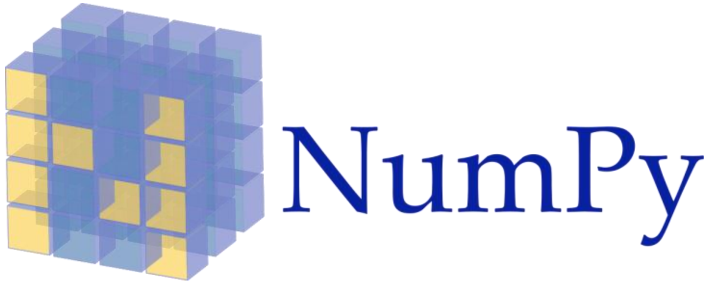


All the points in the dataset



Preparing for Next Week’s Practice Sessions





GUC

Thank You