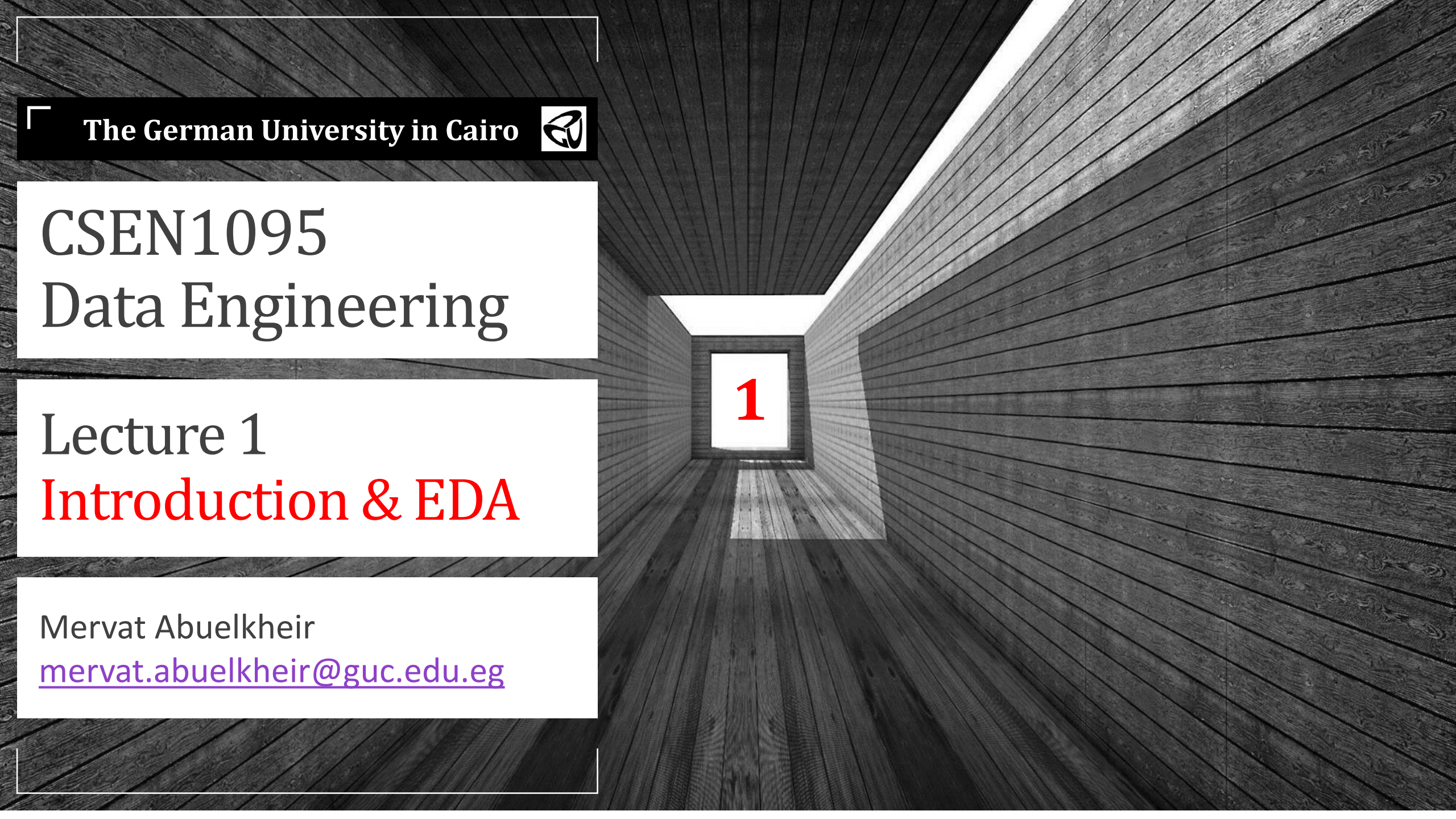
The German University in Cairo

CSEN1095
Data Engineering



Lecture 1
Introduction & EDA

Mervat Abuelkheir
mervat.abuelkheir@guc.edu.eg

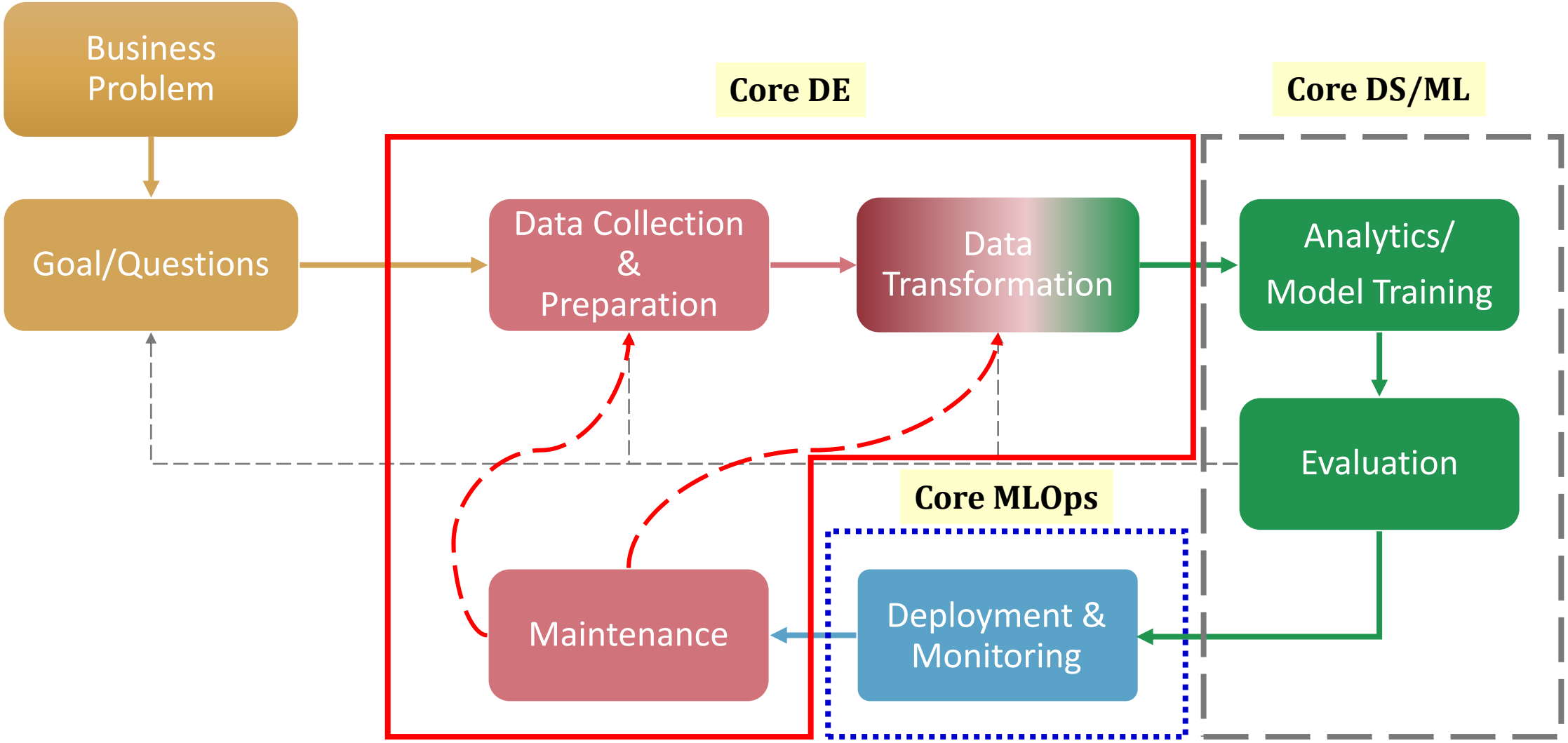
"...and if anyone here suspects that the algorithm that put these two together might be flawed, speak now..."

Data Engineering

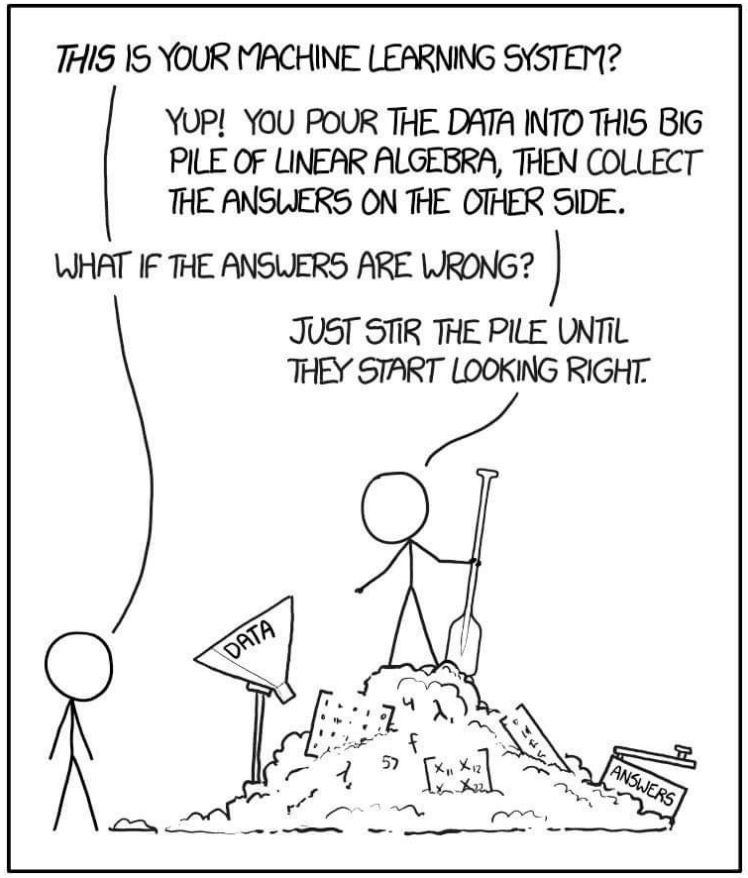
The processes that collect and integrate raw data from multiple diverse resources into a unified and accessible data repository that can be used for analytics and other applications (that may incorporate ML models)

- Raw data pose challenges of quality – incompleteness, inaccuracy, inconsistency, …
- Applications may need data to be processed in batches or as streams
- Analytics need continuous serving for continuous – and improved – insights

This Course is about building Data Pipelines and Infrastructures

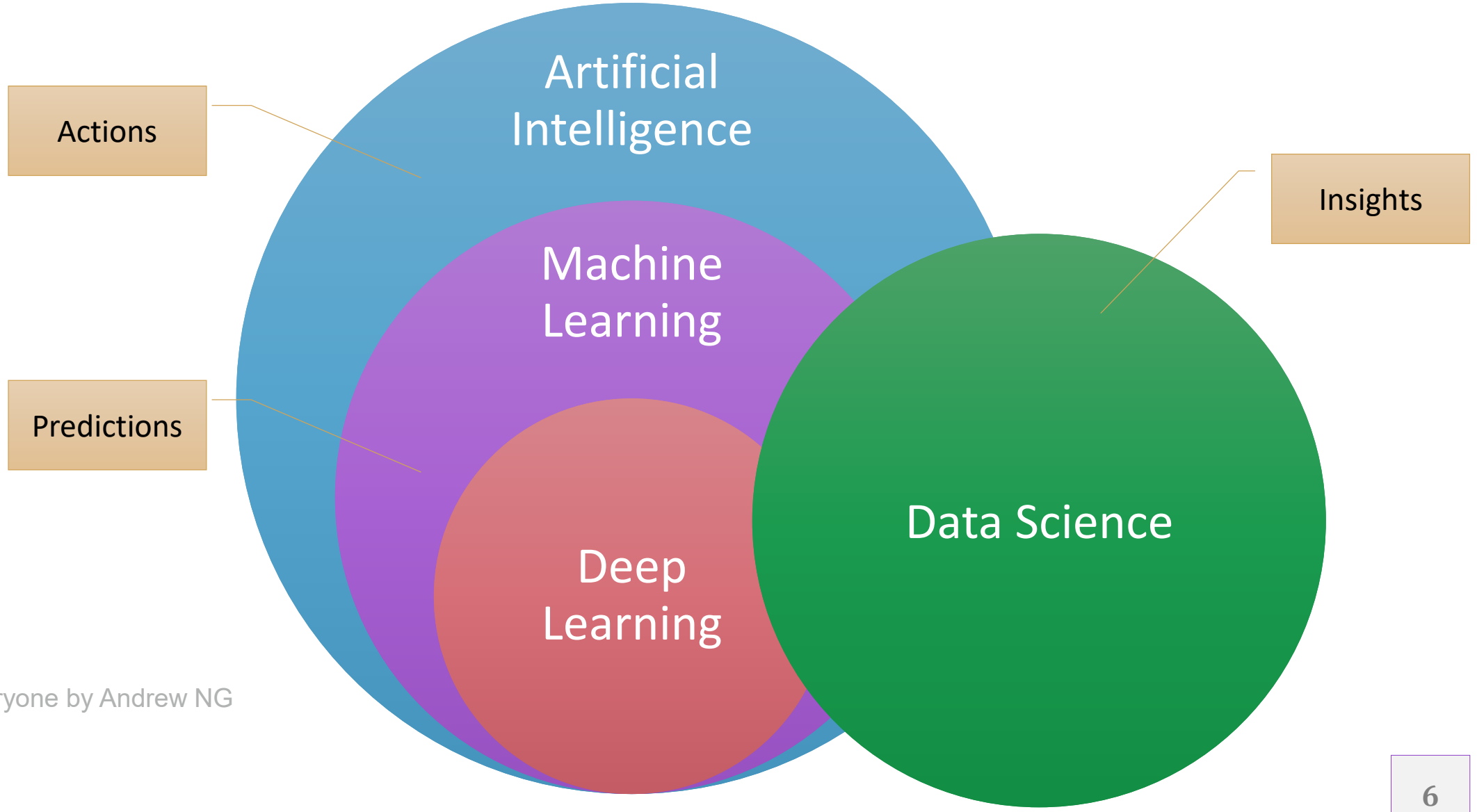


… with Data Quality as an Integral Design Concept



Analytics/ML is only as good as the data you provide to it as input

Decomposing The Jargon

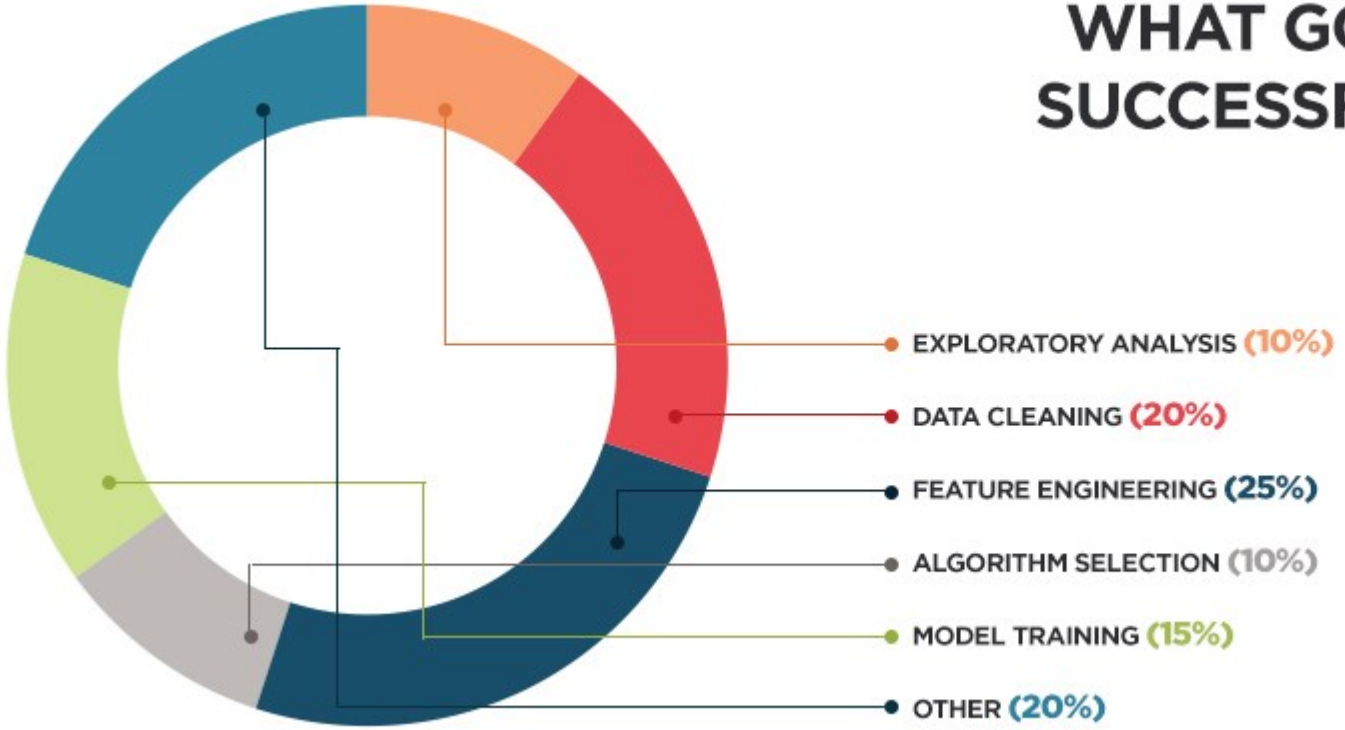


Source: AI for Everyone by Andrew NG

Jobs and whatnot

- Language Requirement: **Python**
- Solid Knowledge of **Operating Systems**
- Heavy, In-Depth Database Knowledge – **SQL** and **NoSQL**
- Data Warehousing and Data Processing at Scale – **Hadoop**, **MapReduce**, **Spark**, **Kafka**
- Basic **Machine Learning** Familiarity

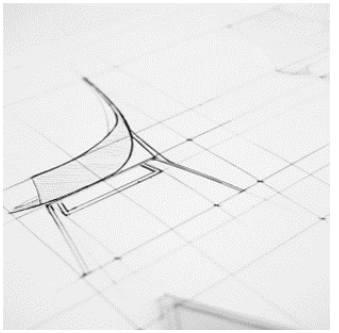
https://www.analyticsvidhya.com/blog/2018/11/data-engineer-comprehensive-list-resources-get-started/

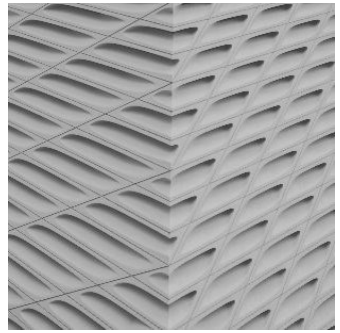


- Data workers spend more than 40% of their time searching for and preparing data
- On average, data workers leverage more than six data sources, 40m rows of data and seven different outputs along their analytic journey

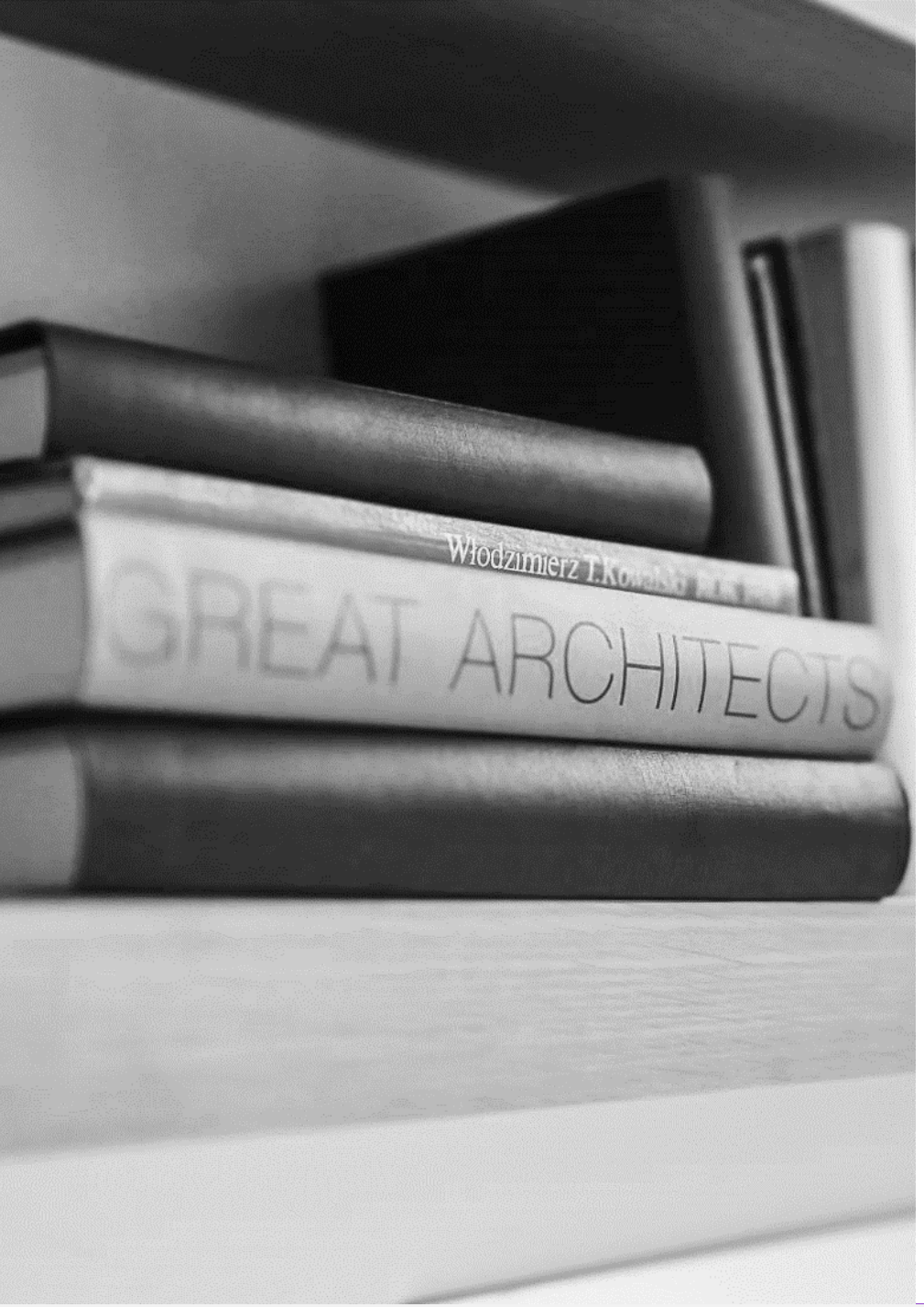
https://www.information-age.com/productivity-in-data-science-123482699/
https://www.information-age.com/data-engineer-sexiest-job-21st-century-123480578/

Week	Saturday Lecture	Practical elements to cover in the lab
1	Introduction Data exploration • Understand data • Use statistics	
2	Data cleaning and imputation • Missing values • Noise and Outliers	Python, Anaconda, Pandas, and EDA
3	Data transformation • Normalization • Encoding and Featurization	Cleaning I
4	ETL and Pipelining	Cleaning II and Data Transformation
5	Data Storage and Repositories • Data Warehouses • Data Lakes	Docker Images
6	Data integration	Docker Compose
7	Traditional and Modern Data Models • SQL and NoSQL	Docker II
8	Big Data Ecosystem I	PySpark I
9	Big Data Ecosystem II	PySpark II
10	Stream Processing	PySpark III
11	Building data Applications I	Airflow I
12	Building data Applications II	Airflow II + Dashboards

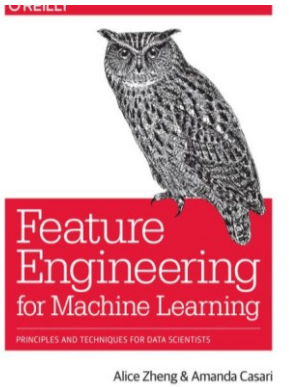


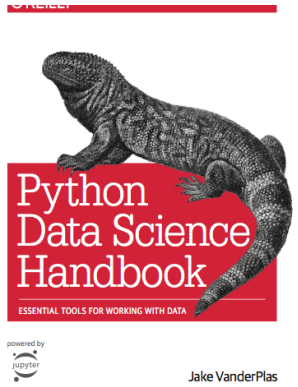


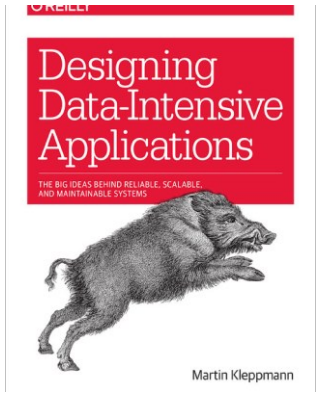
Course Outline

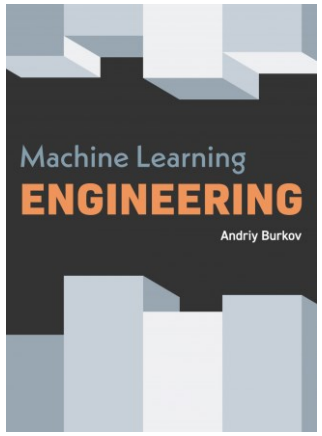


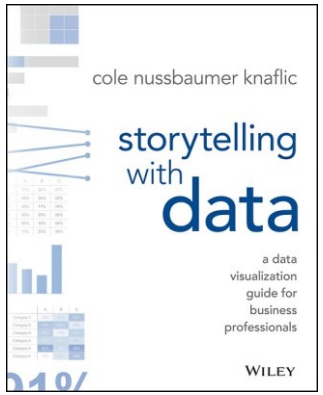
Resources





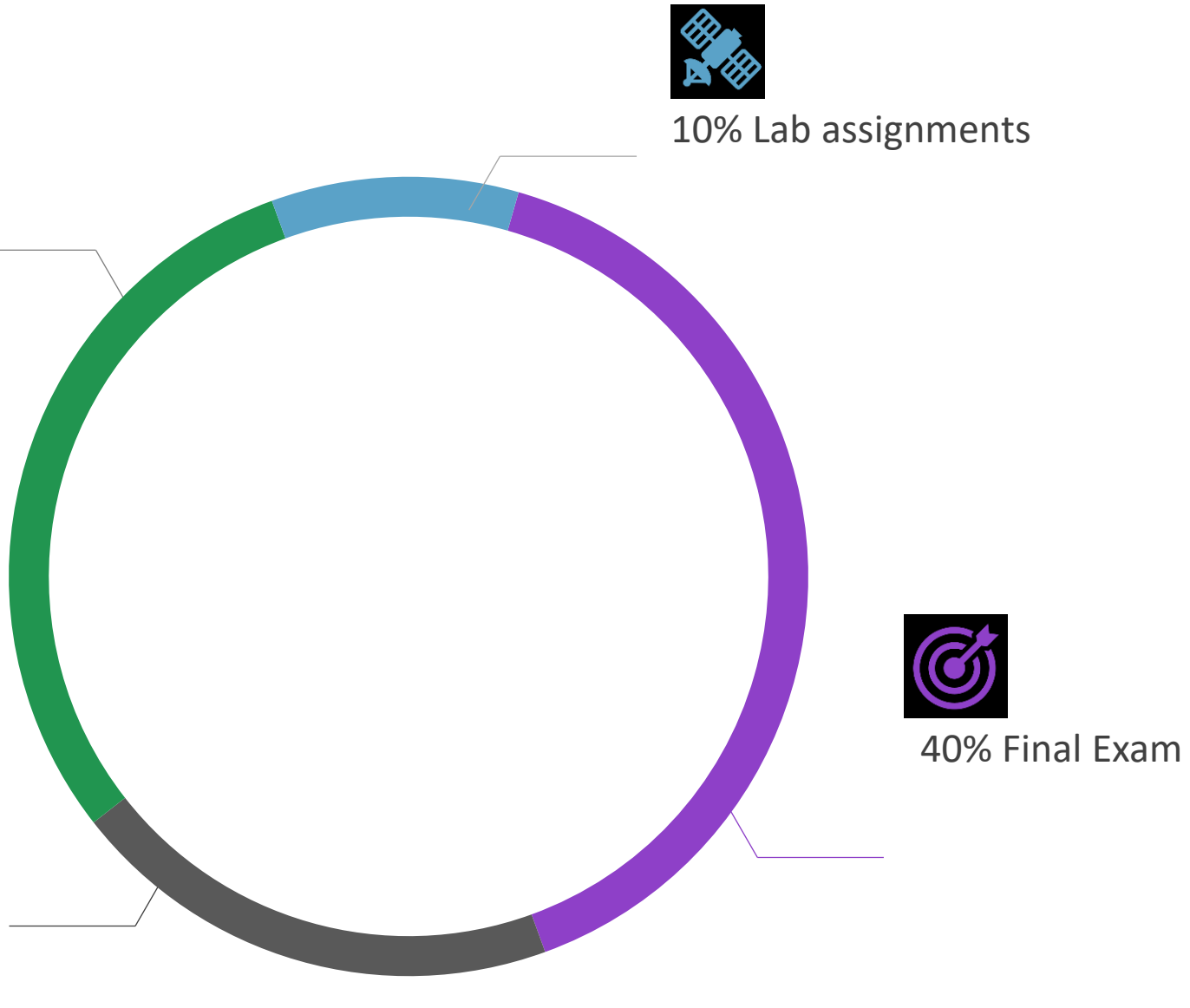






Click on picture for link to book chapters

Grade Distribution



Course Team

- Mervat Abuelkheir
- Badr El-Agoury

Project Workflow

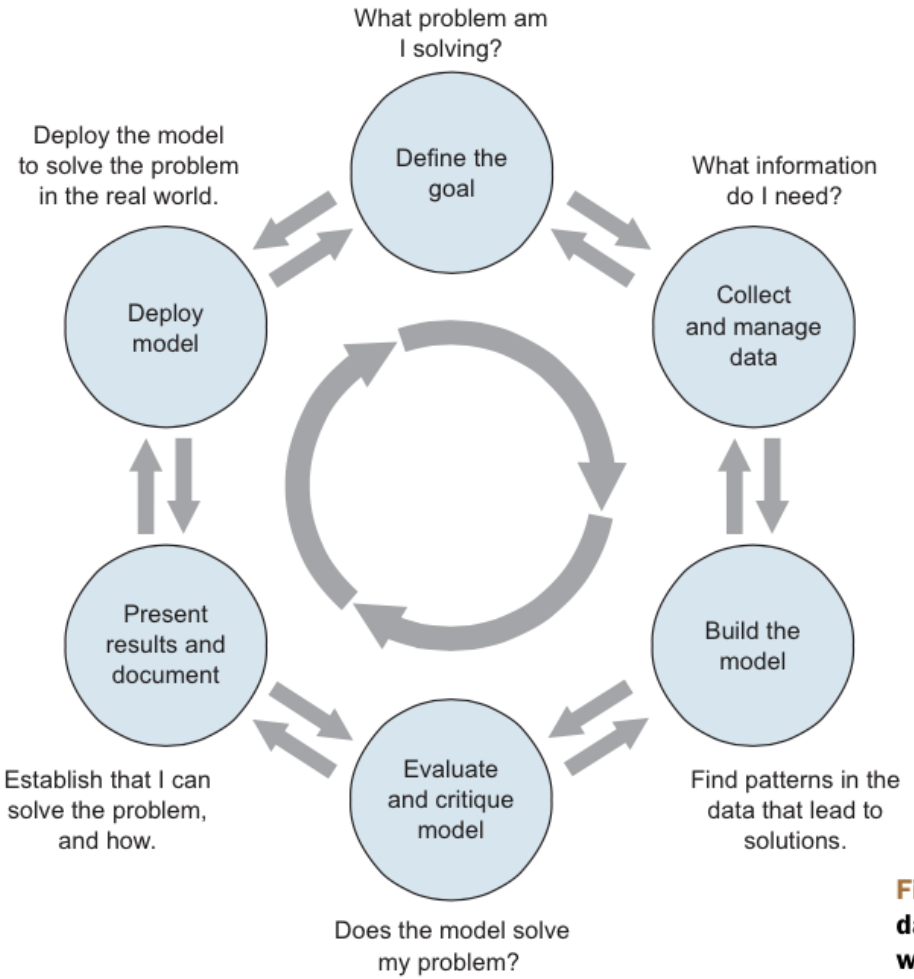


Figure 1.1 **The lifecycle of a data science project: loops within loops**

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?

Source: CMSC320 Data Science Course

Project Workflow

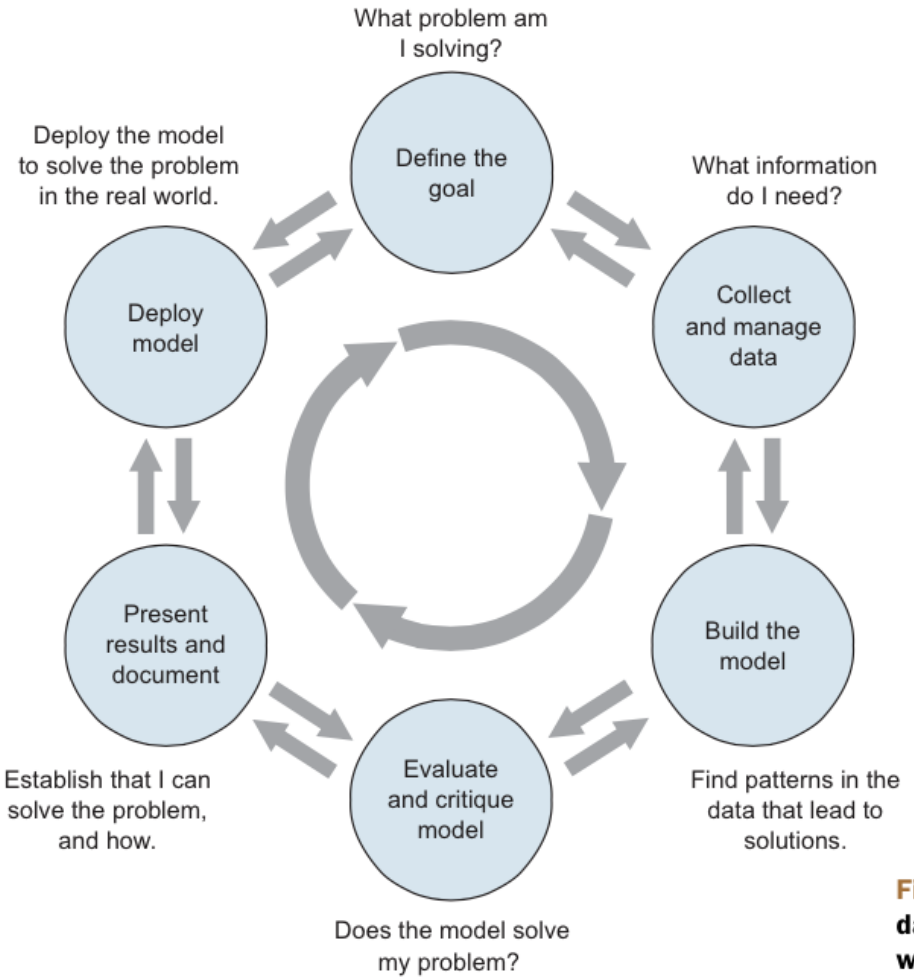


Figure 1.1 **The lifecycle of a data science project: loops within loops**

Data Collection and Management

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

Source: CMSC320 Data Science Course

Project Workflow

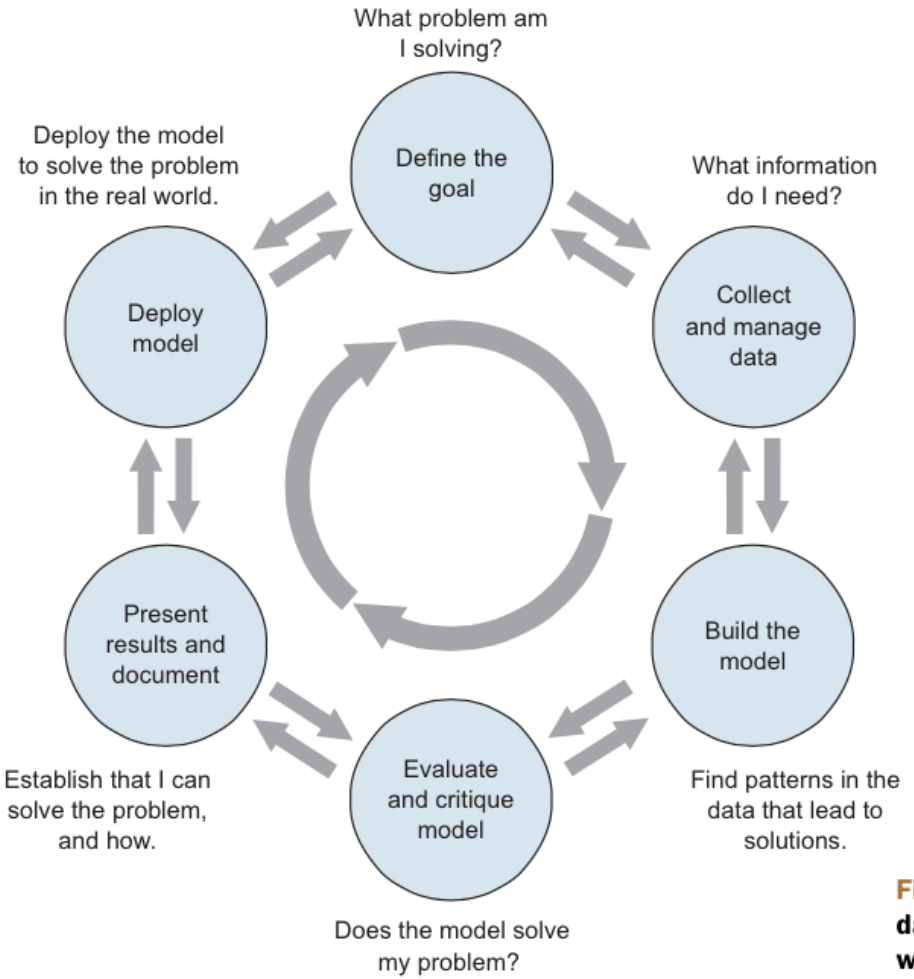


Figure 1.1 **The lifecycle of a data science project: loops within loops**

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?

Source: CMSC320 Data Science Course

Project Workflow

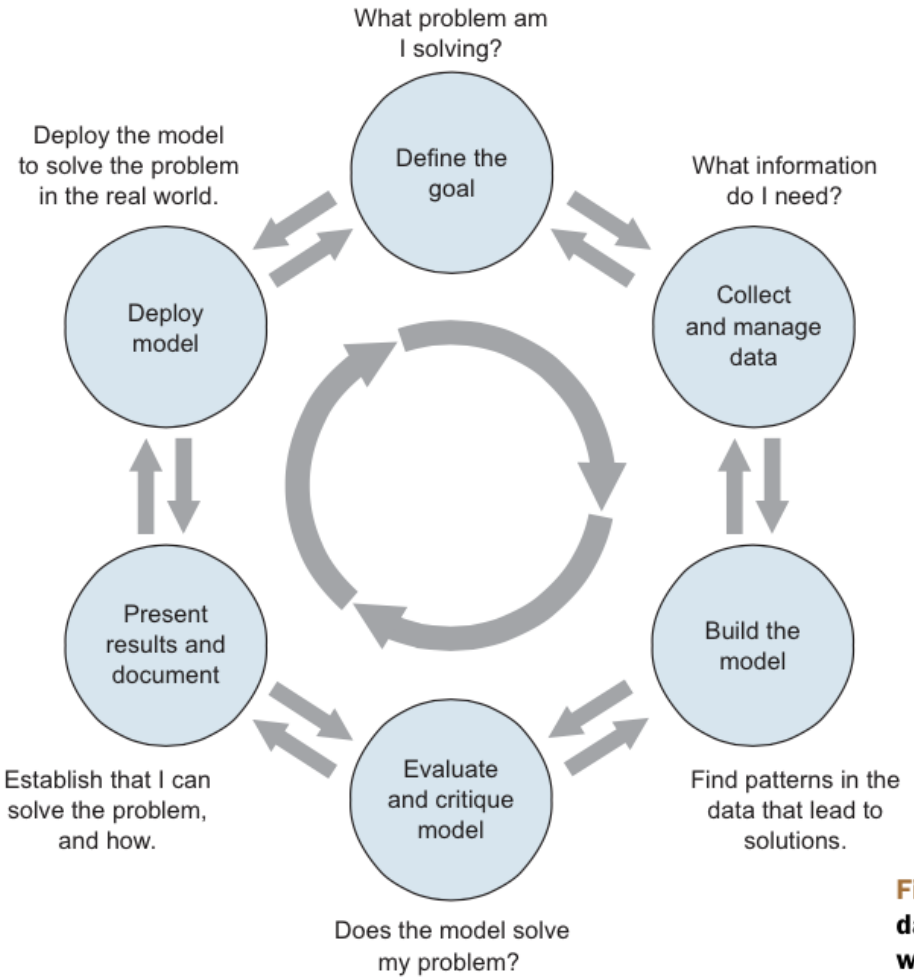


Figure 1.1 **The lifecycle of a data science project: loops within loops**

Model Evaluation

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

Source: CMSC320 Data Science Course

Project Workflow

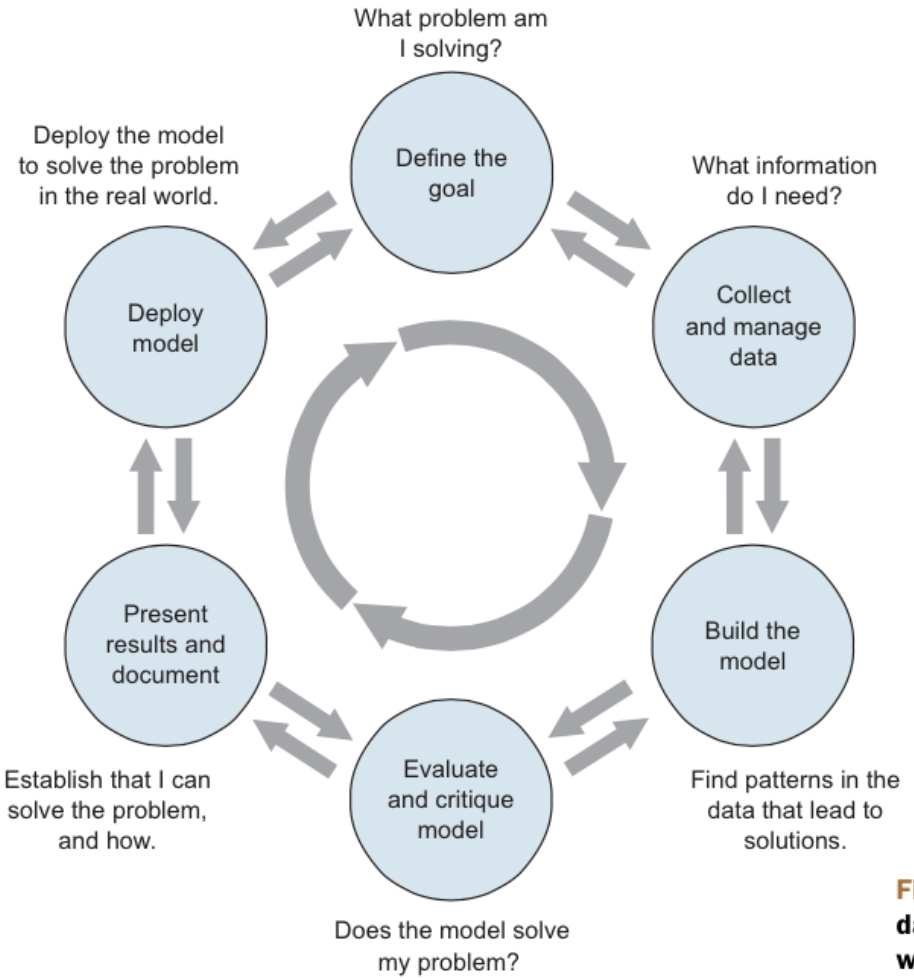


Figure 1.1 **The lifecycle of a data science project: loops within loops**

Presentation

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

Source: CMSC320 Data Science Course

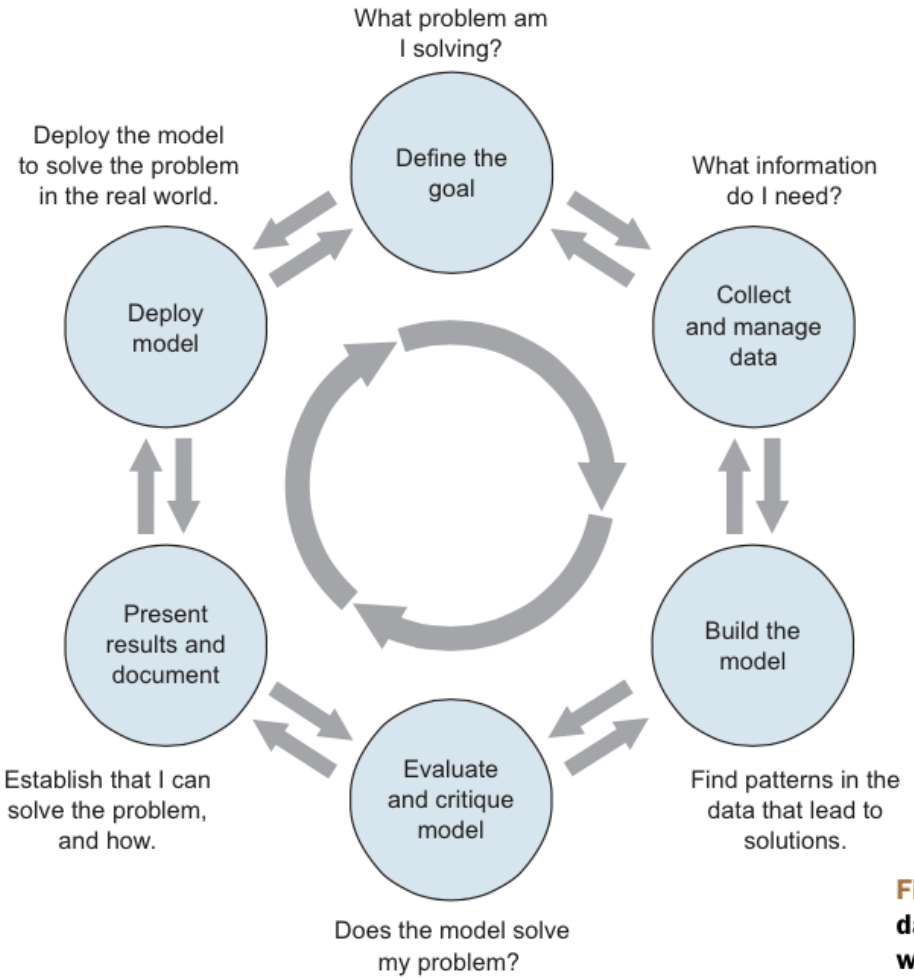


Figure 1.1 **The lifecycle of a data science project: loops within loops**

Deployment

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

Source: CMSC320 Data Science Course

Boston's Hubway Data Challenge

trip duration	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

Feature

Feature vector

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

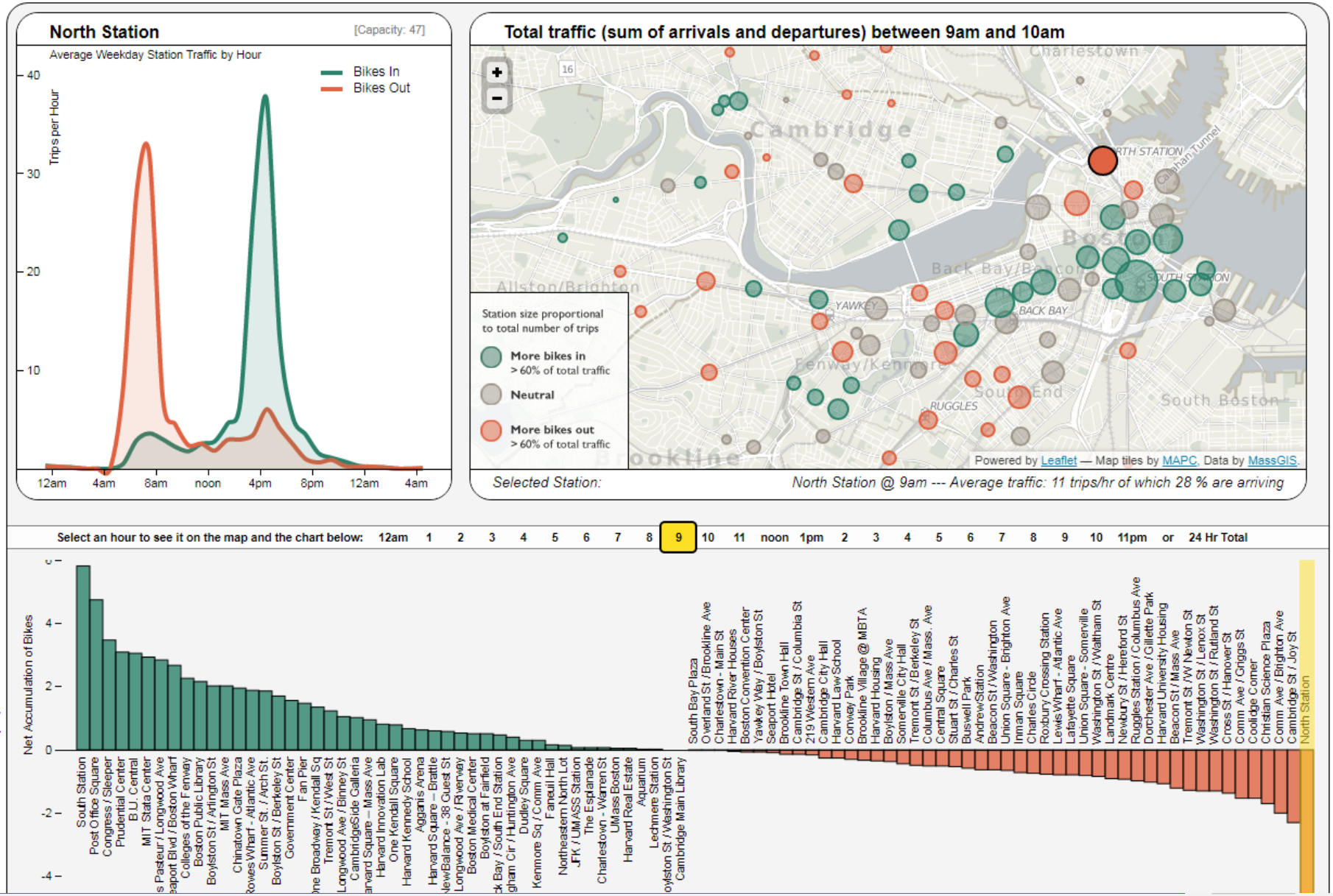
Boston's Hubway Data Challenge

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

Hubway Bikes In / Bikes Out

Explore this interactive to discover the flow of Hubway Bikes around Boston on a typical week day:

- **Select different hours** to see the total average of bikes in and out of stations on the map, and the net gains and losses in the bars below.
- **Select circles on the map or bars below** to see the character of the various stations in the hourly usage curves in the left panel. Note the commuter usage spikes at many stations, and more distributed afternoon usage in more touristy areas. More insights below | How to use this chart



Before We Process: Profiling & Stats

GUC

Building a Data Profile to Assess Data Quality

- Understanding the data in a given system – data values and their frequencies
- Helps with the quality assessment of your data
  - Should do to identify issues and decide on mitigation techniques to improve quality
    - Completeness – missing values, nulls
    - Conformance – formats, value domains
    - Correctness – business logic
- How to measure data quality?
  - Statistics
  - Thresholds
- Data Quality Indicators and Sanity Checks
  - Accepted values in the data model
  - Density – # entities available both temporal (within defined time window) and a temporal
  - % messingness
  - Uncertain values

Building a Data Profile – Task 1

- Use WhiteRabbit (https://github.com/OHDSI/WhiteRabbit) to profile the following datasets:
  - https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
  - https://www.kaggle.com/hugomathien/soccer
  - Or a dataset of your choosing
- Generate a report and submit to my email (DE – Task 1 – Data Profile)
- Think of more complex data quality features and metrics and how to design a tool that includes them

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*

Attribute Types

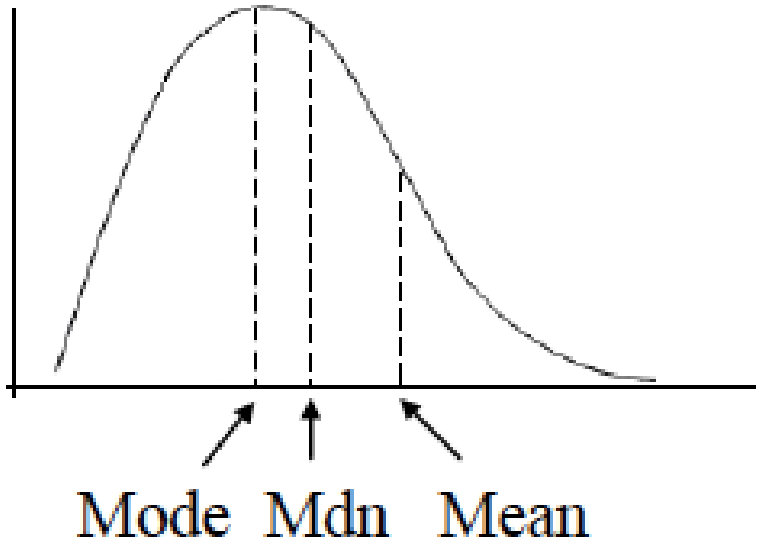
- **Interval-scaled**
  - Measured on a *scale of equal-size units*
  - e.g. *temperature, year*
  - Do not have a true zero point
  - Not possible to be expressed as multiples
- **Ratio-scaled**
  - Have a true zero point
  - A value can be expressed as a *multiple* of another
  - e.g. *years of experience,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

Quantitative Attributes

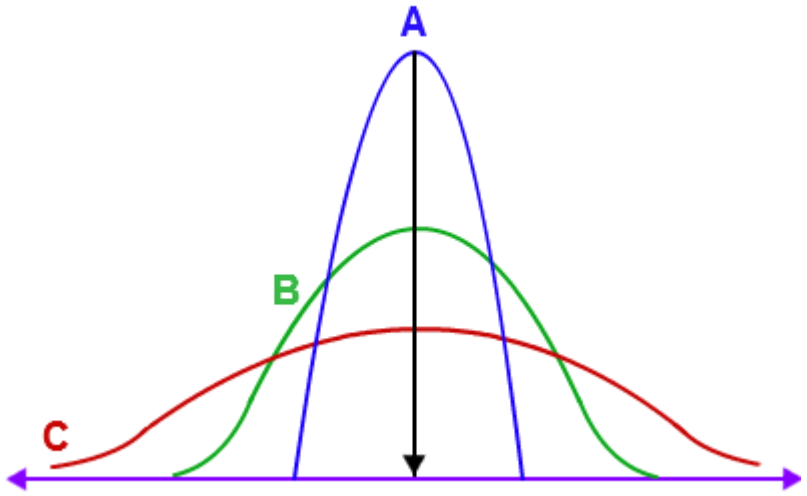
- Sometimes we need to normalize quantitative data
- Sometimes we need to discretize quantitative data – Back to categorical!

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

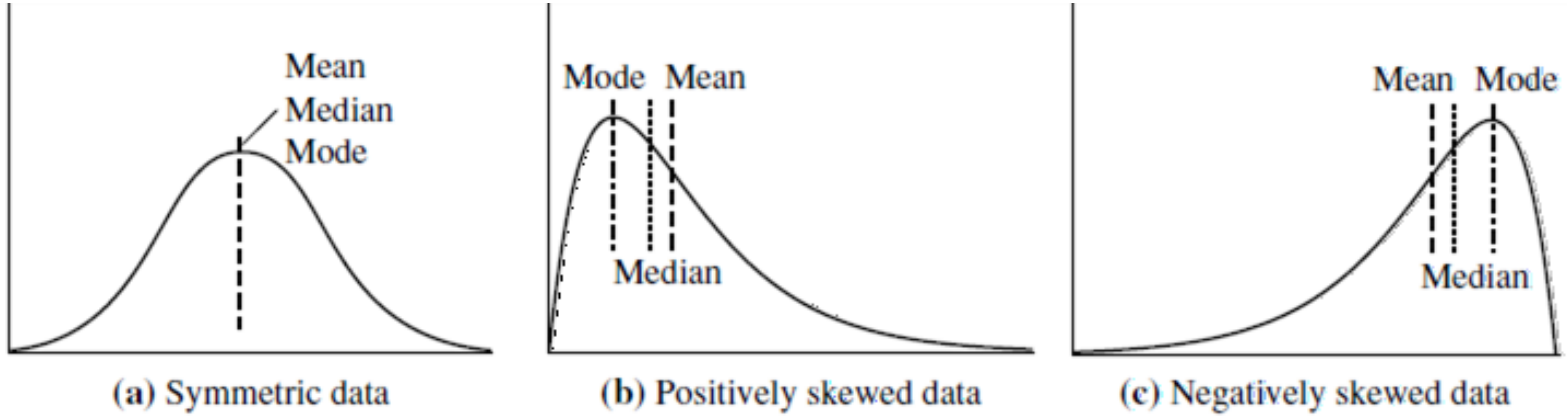
Measuring Central Tendency

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?



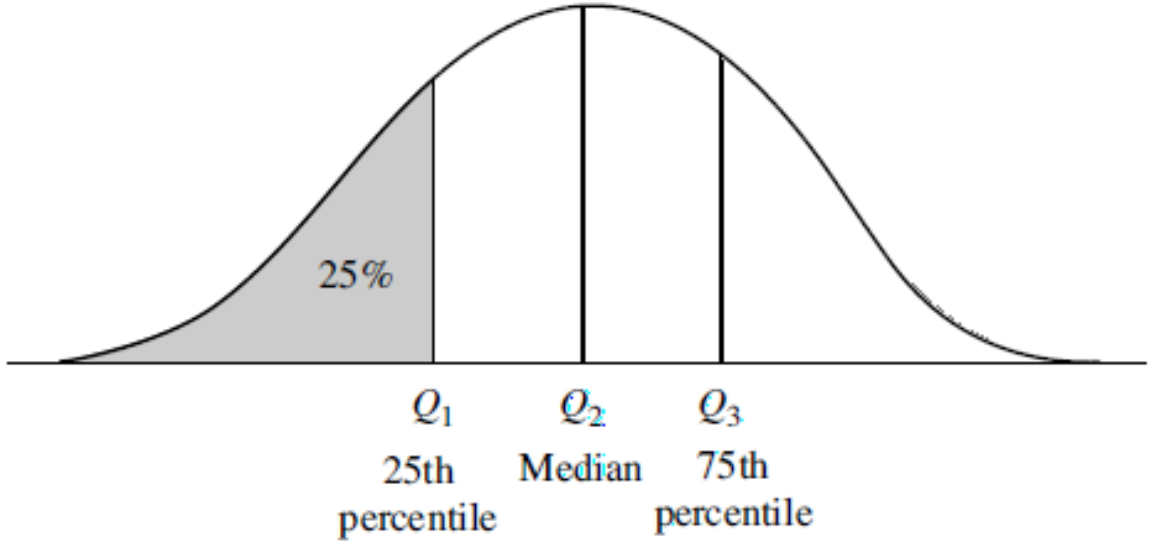
(a) Symmetric data (b) Positively skewed data (c) Negatively skewed data

Measuring Dispersion of Data

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

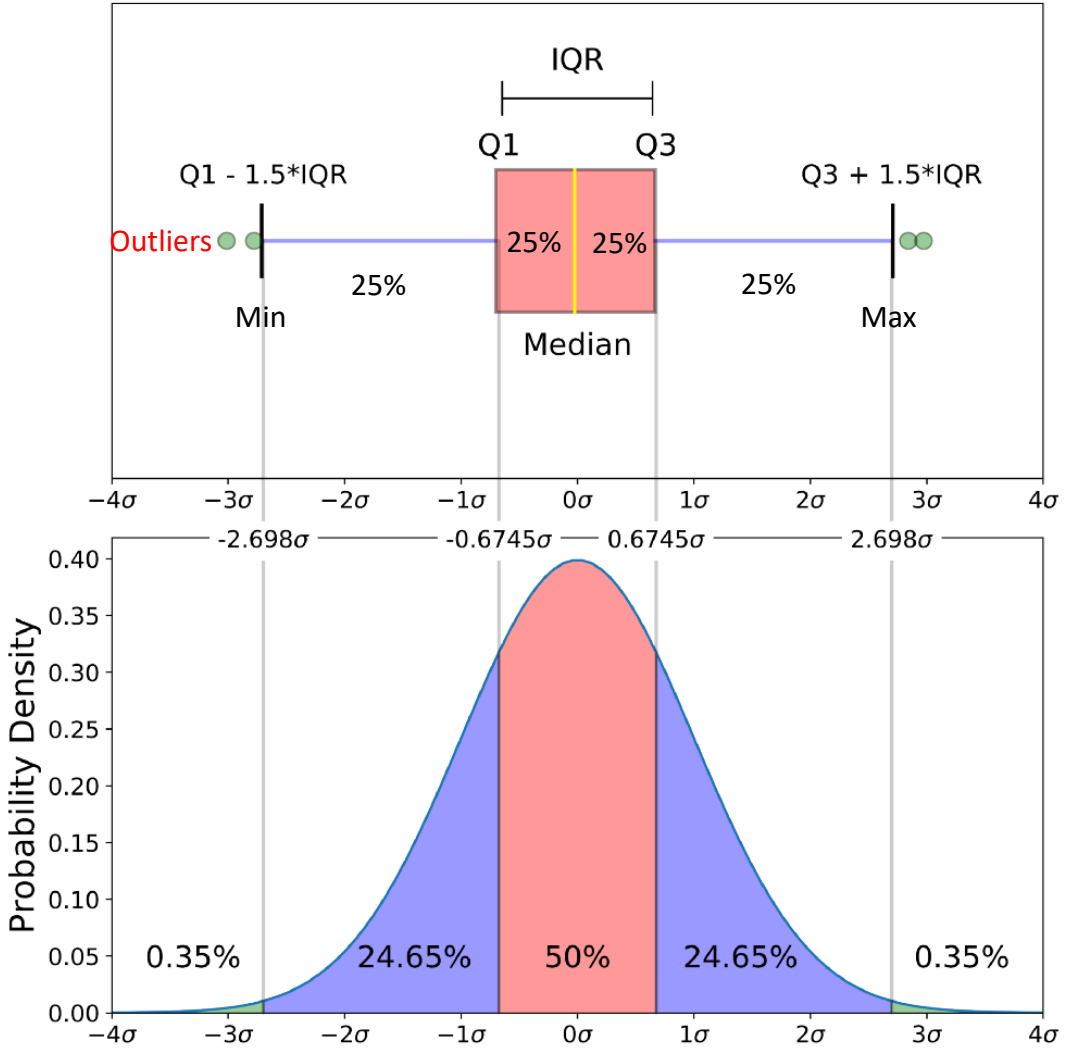
For *N* observations of numerical variable X: $x_1, x_2, \ldots, x_N$

- **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:** = *Q*3 - *Q*1



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!)**



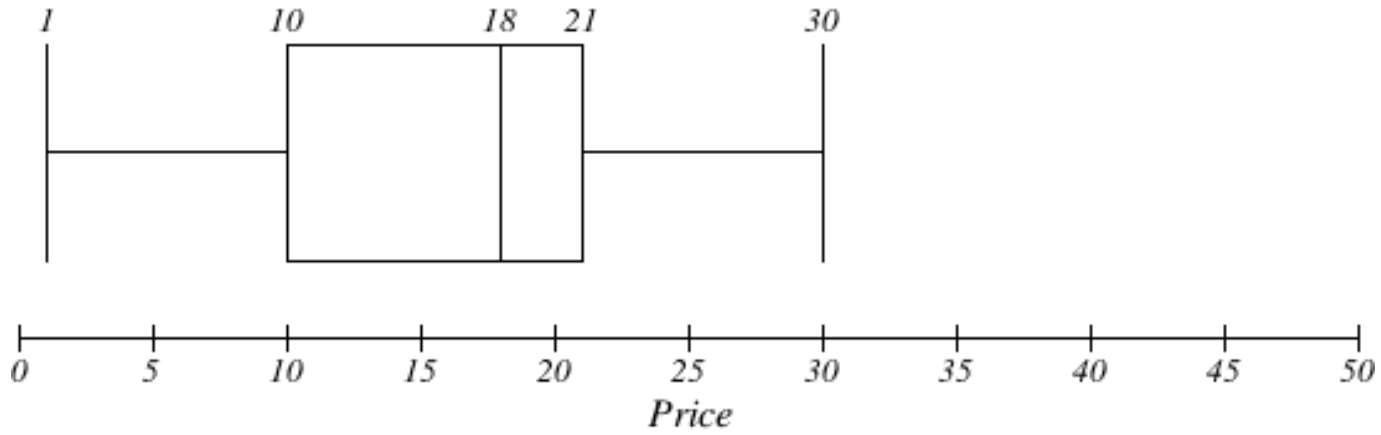
Working Example: https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/box-plot-review

Pop Quiz

Example: Item prices at a store are: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30

Total # observations is 52

1. Identify Q1, Q2, Q3
2. Draw the boxplot

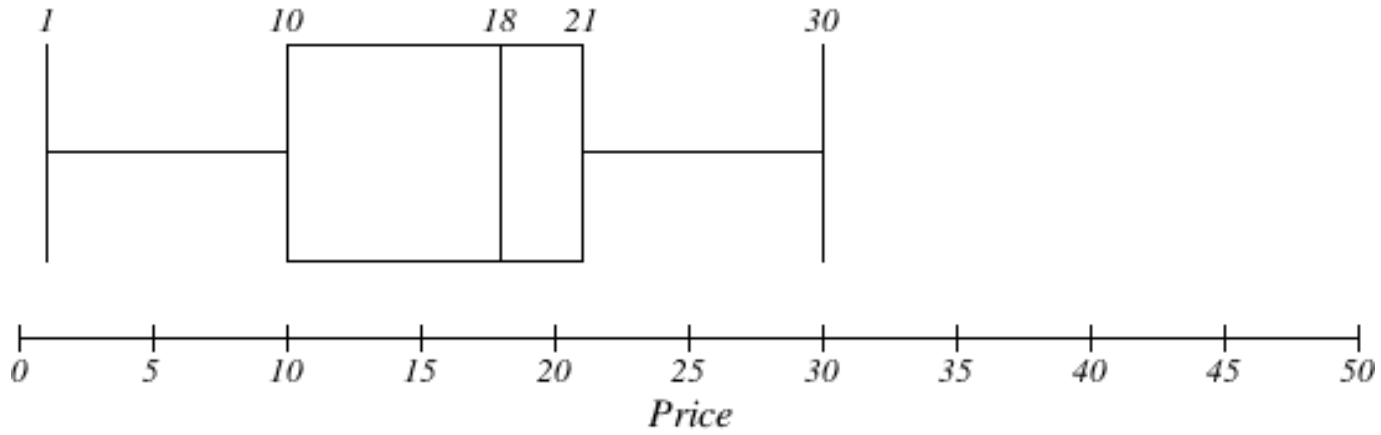


Pop Quiz

Example: Item prices at a store are: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, ~~30~~ 40?

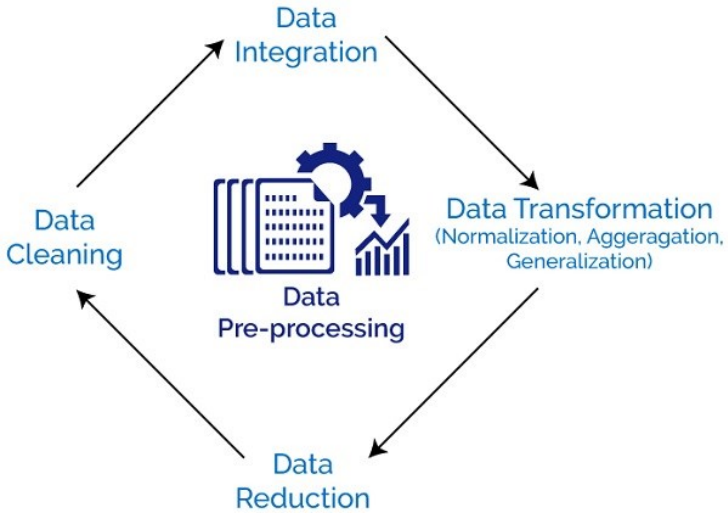
Total # observations is 52

1. Identify Q1, Q2, Q3
2. Draw the boxplot



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
  - $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} {x_i}^2\right) - \bar{x}^2$
  - $SD = \sigma$



Why Prepare/ Preprocess Data?

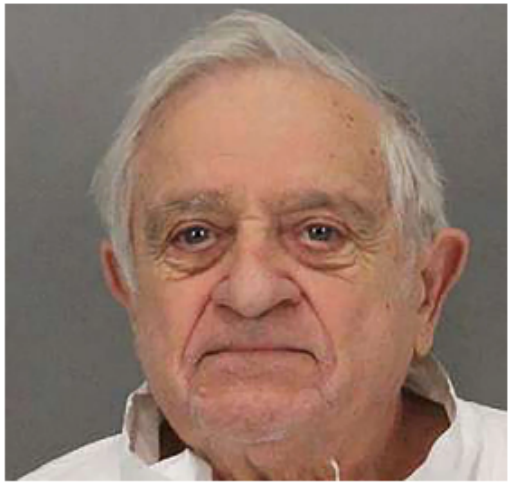
GUC

Police Use Fitbit Data to Charge 90-Year-Old Man in Stepdaughter's Killing

By Christine Hauser

Oct. 3, 2018

https://www.nytimes.com/2018/10/03/us/fitbit-murder-arrest.html



Anthony Aiello, 90, has been charged with murder in the death of his stepdaughter in San Jose, Calif., the police said.
San Jose Police Department

The last time Anthony Aiello spoke to his stepdaughter, he took homemade pizza and biscotti to her house in San Jose, Calif., for a brief visit. Mr. Aiello, 90, told investigators that she then walked him to the door and handed him two roses in gratitude.

But an unnoticed observer in the house later revealed that their encounter ended in murder, a police report said.

When Ms. Navarra's Fitbit data was compared with video surveillance from her home, the police report said, the police discovered that the car Mr. Aiello had driven was still there when her heart rate stopped being recorded by her Fitbit.

Bloodstained clothes were later found in Mr. Aiello's home, the document said. He was arrested on Sept. 25.

Mr. Aiello was "confronted" with the Fitbit information during questioning, said Brian Meeker, a San Jose police detective. "After explaining the abilities of the Fitbit to record time, physical movement, and heart rate data, he was informed that the victim was deceased prior to his leaving the house," Detective Meeker said in the report.

Mr. Aiello said that could not be true, insisting Ms. Navarra had walked him to the door, and he suggested that someone else could have been in the home, the report said.

"I explained that both systems were on internet time, and there was no deviation," Detective Meeker said.

After they finished their questions, detectives left Mr. Aiello alone in the interview room. He began talking to himself, the report said, saying repeatedly, "I'm done."

year worldwide.

As more people used the devices, it was inevitable that they would be worn by victims or suspects in crimes and potentially hold tantalizing clues or even plausible answers: Does a suspect's alibi of being at home asleep hold up? Does a victim's steady heart rate at the time of an alleged attack suggest the charge was fabricated?

Using trackers this way, of course, assumes that the devices are accurate—and not just accurate on average, but at very specific moments in time, a sort of black box for the body that reveals physiological truths that its wearer might prefer to conceal. Research on fitness trackers, however, shows they don't always perfectly mirror reality. An analysis of 67 studies on Fitbit's movement tracking concluded that the device worked best on able-bodied adults walking at typical speeds. Even then, the devices weren't perfect—they got within 10 percent of the actual number of steps a person took half of the time—and became even less accurate in counting steps when someone was resting their wrist on a walker or stroller, for example. "It's not measuring actual behavior," says Lynne Feehan, a clinical associate professor at the University of British Columbia and the lead researcher on the paper. "It's interpreting motion."

Many fitness tracker users experience moments of misinterpretation: the piano playing session that was categorized as cycling; the times during sweaty exercise when it stops picking up a heart rate. Even Fitbit's own terms of service point out that it is a consumer product with accuracy that is "not intended to match that of medical devices or scientific measurement devices."

Smartwatches decipher heart rate using green LEDs that beam hundreds of times per second into capillaries through the skin. Those capillaries allow in more of the light when full of blood, and less between beats, and the device measures how much light is absorbed. That measurement is then siphoned through a proprietary algorithm to generate a heart rate figure. University of Wisconsin researchers looked at how well wrist-worn fitness trackers measured heart rate, comparing it to an electrocardiograph, the gold standard for heart monitoring. They found that the fitness trackers' heart rate deviated more from the actual rate when a subject exercised on a treadmill than when at rest. (Fitbit won't talk specifics about its accuracy, saying in a statement, "We are confident in the performance of all our devices" and that the company continues to test them.)

Tony's defense lawyers signaled that they would attack the reliability of the Fitbit data. They assembled a grab bag of disqualifications: They said Karen wore the device for only two weeks or less, and it hadn't yet normalized to her signal; they said that Fitbit, which assigns a confidence score of 0 to 3 to its data collection, at times assigns zero confidence to the data on Karen's device on the day the prosecution says she was murdered; and Edward Caden, one of the defense attorneys, said that what the prosecution calls a "spike" in Karen's heart rate is more like "a pimple." Caden even asserted that there were moments after 3:28 pm when Karen's Fitbit seems to still report heartbeat data.

Angela Bernhard, the chief trial deputy for Santa Clara County, told me in August that she expected that the defense would "be fighting to keep out a lot of the evidence that we want in" and that she intended to present the Fitbit evidence at trial. "Ultimately it's up to the judge what evidence gets brought in and what doesn't," she said. At a grand jury hearing in August, Bonham, the Fitbit executive, testified that Fitbit had turned over a voluminous Excel spreadsheet of Karen's raw heartbeat and step data. He also clarified that a confidence rating of zero means the device isn't registering a heartbeat at all, and detectives say that Karen's device showed no heartbeat and zero confidence at 3:28 pm and after. Detective Meeker testified to the reliability of Karen's device specifically: At two times in early September that Karen was visible on surveillance footage walking in stores, her Fitbit recorded movement. (Fitbit declined to comment on Aiello's case.)

https://www.wired.com/story/telltale-heart-fitbit-murder/

Data Challenges

- Massive data (500k users, 20k movies, 100m ratings)
- Curse of dimensionality (high-dimensional problem in terms of features)
- 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

Source: CS109 Stanford's Data Science Course

Why Prepare/Preprocess Data?

To increase the quality of data (*usability* and *reliability*) and make it suitable for the requirements of the intended use

- Factors of data quality
  - *Accuracy* → lack of is due to faulty instruments, errors caused by human/computer/transmission, deliberate errors …
  - *Completeness* → lack of is due to data acquired over different design phases, optional attributes
  - *Consistency* → lack of is due to semantics, data types, field formats …
  - *Timeliness* → data should be current, and available when it's needed by user
  - *Integrity* →lack of due to poor definitions of data relationships
  - *Interpretability* → how easy the data is understood

Example 1

Two records from a pipe-delimited file:

T.Das|97336o8327|24.95|Y| – |0.0|1000

TedJ.|973 – 360 – 8779|2000|N|M|NY|1000

- Interpretability?
- Accuracy?
- Integrity?
- Completeness?
- Consistency?

Example 1

Two records from a pipe-delimited file:

T.Das|97336o8327|24.95|Y| – |0.0|1000

TedJ.|973 – 360 – 8779|2000|N|M|NY|1000

- Interpretability? name, phone number, revenue, indicator, gender, state, usage

Example 1

Two records from a pipe-delimited file:

T.Das|97336o8327|24.95|Y| – |0.0|1000

TedJ.|973 – 360 – 8779|2000|N|M|NY|1000

- Interpretability? name, phone number, revenue, indicator, gender, state, usage
- Accuracy?

Example 1

Two records from a pipe-delimited file:

T.Das|97336o8327|24.95|Y| – |0.0|1000

TedJ.|973 – 360 – 8779|2000|N|M|NY|1000

- Interpretability? name, phone number, revenue, indicator, gender, state, usage
- Accuracy?
- Integrity?

Example 1

Two records from a pipe-delimited file:

T.Das|97336o8327|24.95|Y| – |0.0|1000

TedJ.|973 – 360 – 8779|2000|N|M|NY|1000

- Interpretability? name, phone number, revenue, indicator, gender, state, usage
- Accuracy?
- Integrity?
- Completeness?

Example 1

Two records from a pipe-delimited file:

T.Das|97336o8327|24.95|Y| – |0.0|1000

TedJ.|973 – 360 – 8779|2000|N|M|NY|1000

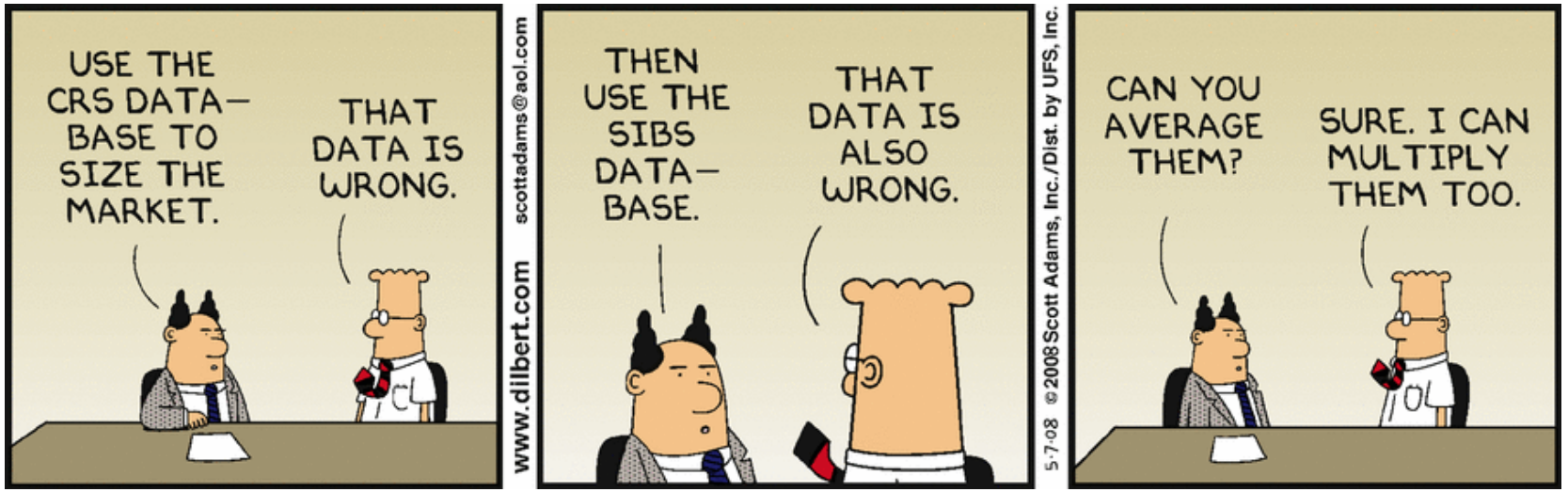
- Interpretability? name, phone number, revenue, indicator, gender, state, usage
- Accuracy?
- Integrity?
- Completeness?
- Consistency?

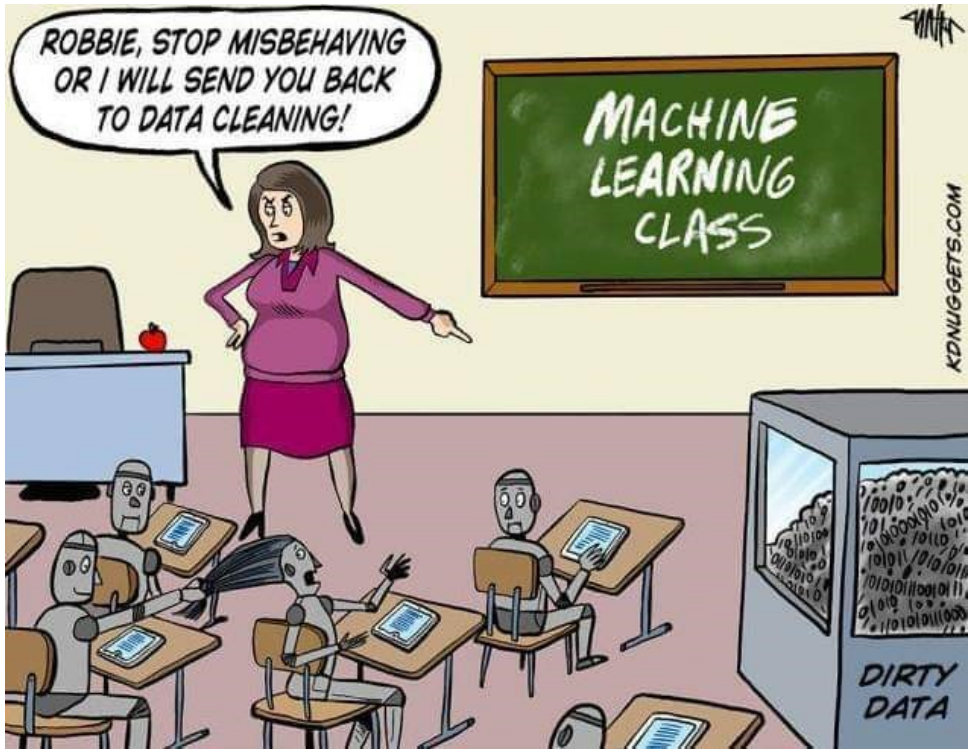
Data Preparation is itself a Challenge!

- Data quality problems are highly complex and context dependent
  - Extensive domain knowledge is needed
  - Solutions need to be chosen case by case
- No single tool can solve a majority of data quality problems!
- Do not apply data preprocessing methods manually – **AUTOMATE**
  - For large datasets
  - Reuse of code for similar issues

Major Preparation Tasks
That Improve Quality of Data

- ***Data cleaning*** → filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies
- ***Data transformation*** → normalization, discretization
- ***Data reduction*** → obtain a reduced representation of the data set that is much smaller in volume, while producing almost the same analytical results
- ***Data integration*** → include data from multiple sources in analysis, map semantic concepts, infer attributes …


USE THE CRS DATA-BASE TO SIZE THE MARKET.
THAT DATA IS WRONG.
THEN USE THE SIBS DATA-BASE.
THAT DATA IS ALSO WRONG.
CAN YOU AVERAGE THEM?
SURE. I CAN MULTIPLY THEM TOO.
www.dilbert.com scottadams@aol.com
5-7-08 © 2008 Scott Adams, Inc./Dist. by UFS, Inc.


ROBBIE, STOP MISBEHAVING OR I WILL SEND YOU BACK TO DATA CLEANING!
MACHINE LEARNING CLASS
DIRTY DATA
KDNUGGETS.COM

GUC

Thank You