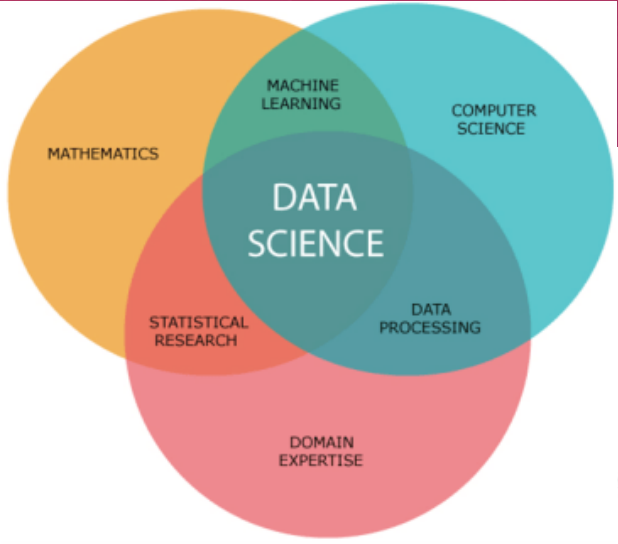
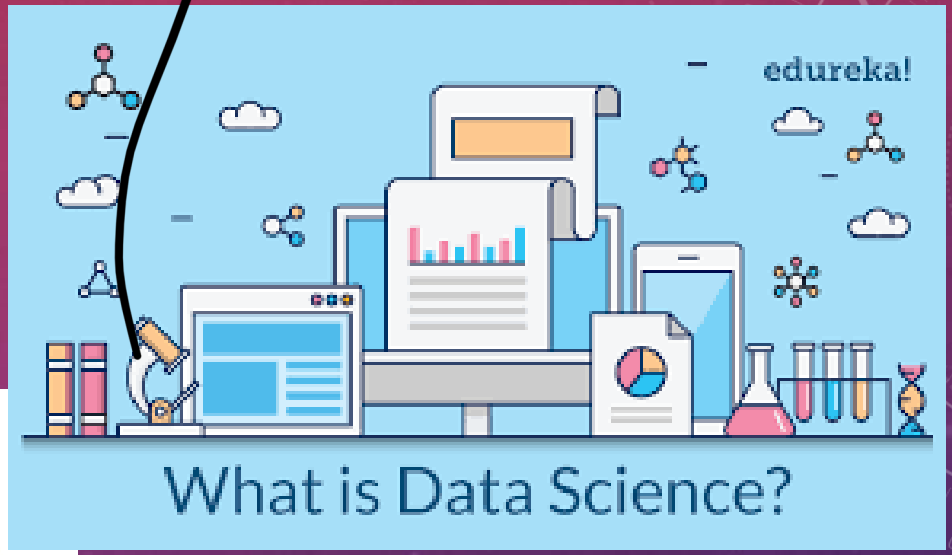
INTRODUCTION TO DATA SCIENCE

WHAT IS DATA SCIENCE?



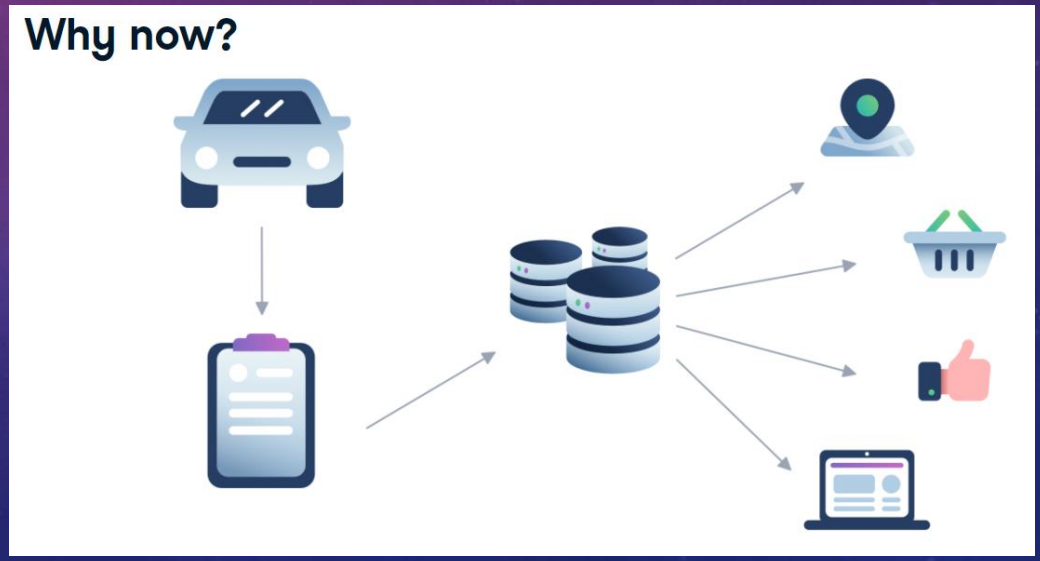


WHAT IS DATA SCIENCE AND WHAT CAN THEY DO?

- Making data work for you - It's a set of methodologies for taking in thousands of forms of data that are available to us today and using them to draw meaningful conclusions.
- So what can data do?
  - Data can describe our current state, like our energy consumption. This can be accomplished with dashboards or alerts, simplifying time-intensive reporting processes.
  - It can help detect anomalous events, such as fraudulent purchases. If we have data on what has happened previously, we can increase efficiency by automatically detecting a new event that is unexpected or abnormal.
  - Data can also diagnose the causes of observed events and behaviors, for instance your activity on Spotify or Netflix.

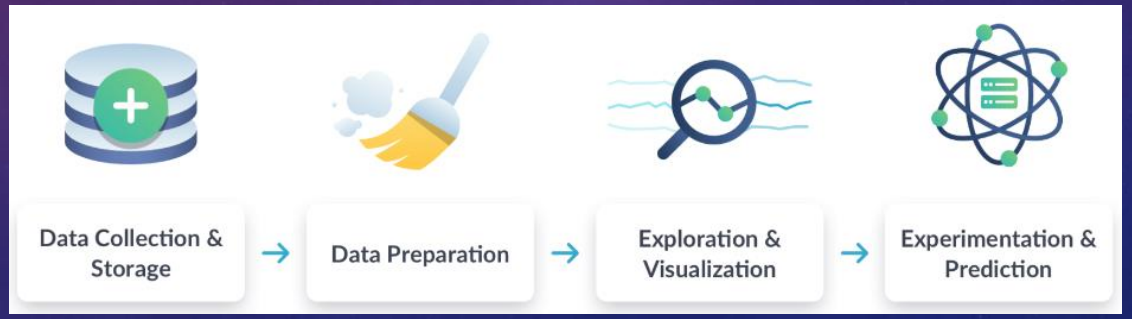
WHY NOW? WHY IS IT SO POPULAR?

- Sample Scenario - Suppose that you visit a car dealership and fill out some information.
- Once we have that data, it's easy to use the email address that you provided when you bought that car to tie your car purchase data with your data from social media or web browsing. Suddenly, we have a very complete picture about everyone who purchased a car in the last year: their ages, their likes and dislikes, their families, their friends. This additional data can be used to predict what price you can pay for your car, and what other purchases you're likely to make, or how best to sell you insurance for that new car.



DATA SCIENCE WORKFLOW

- Data Collection and Storage - First, we collect data from many sources, such as surveys, web traffic results, geo-tagged social media posts, and financial transactions. Once collected, we store that data in a safe and accessible way.
- Data Preparation – This includes "cleaning data", for instance finding missing or duplicate values, and converting data into a more organized format.
- Exploration & Visualization – Then, we explore and visualize the cleaned data. This could involve building dashboards to track how the data changes over time or performing comparisons between two sets of data.
- Experimentation & Prediction – Finally, we run experiments and predictions on the data. For example, this could involve building a system that forecasts temperature changes or performing a test to find which web page acquires more customers. Another example is to create a model to predict fraudulent activities.



LET’S ORDER THEM

Divide the users into different groups and predict the churn for each group.

Reformat the delivery rate on all entries to be in the same time zone.

Download the data.

Create a line chart that shows decay in subscriptions by group of customers.

BUILDING A CUSTOMER SERVICE CHATBOT

- Data Collection and Storage
- Exploration and Visualization
- Experiment and Prediction

a. Gather customer information for each conversation.
b. Use a machine learning model to predict possible responses for each question.
c. Create a bar chart of the number of conversations in each type.
d. Plot the number conversations vs. the time of day.
e. Load the transcripts into the data teams' database.
f. Collect the timestamps for each transcript.
g. Create an algorithm that classifies the initial customer question.

EXAMPLES OF APPLICATIONS OF DATA SCIENCE

- Traditional machine learning
- Internet of Things (IoT)
- Deep Learning

CASE STUDY: FRAUD DETECTION

Amount	Date	Location	...
149.62	2019-05-23	London	...
2.69	2018-10-03	Birmingham	...
378.66	2019-06-15	Liverpool	...
123.5	2019-01-12	London	...
69.99	2018-06-16	São Paolo	...
3.67	2019-03-06	Brussels	...
...	...	...	...

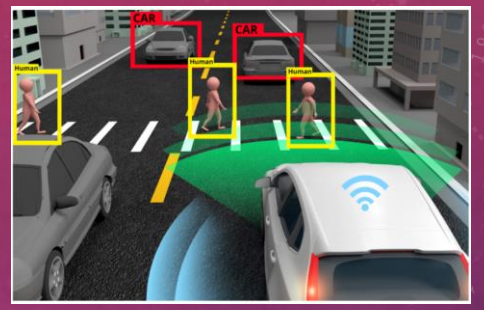
- Suppose you work in fraud detection at a large bank. You'd like to use data to determine the probability that the transaction is fake.
- To answer this question, you might start by gathering information about each purchase, such as the amount, date, location, purchase type, and card-holder's address. You'll need many examples of transactions, including this information, as well as a label that tells you whether each transaction is valid or fraudulent. Luckily, you probably have this information in a database. These records are called "training data" and are used to build an algorithm (a model). Each time a new transaction occurs, you'll give your algorithm information, like amount and date, and it will answer the original question: What is the probability that this transaction is fraudulent?

- What do we need for machine learning?
  - A well-defined question – What is the probability that this transaction is fraudulent?
  - A set of example data – Old transactions labeled as fraudulent or valid.
  - A new set of data to use our algorithm on – New credit card transactions.

CASE STUDY: SMART WATCH

- Now, suppose you're trying to build a smart watch to monitor physical activity. You want to be able to auto-detect different activities, such as walking or running. Your smart watch is equipped with a special sensor, called an "accelerometer", that monitors motion in three dimensions. The data generated by this sensor is the basis of your machine learning problem.
  - You could ask several volunteers to wear your watch and record when they are running or walking. You could then develop an algorithm that recognizes accelerometer data as representing one of those two states: walking or running.
- Your smart watch is part of a fast-growing field called "the Internet of Things", also known as IoT, which is often combined with Data Science.
  - IoT refers to gadgets that are not standard computers, but still can transmit data. This includes
    - smart watches,
    - internet-connected home security systems,
    - electronic toll collection systems,
    - building energy management systems, and much, much more.
- IoT data is a great resource for data science projects!

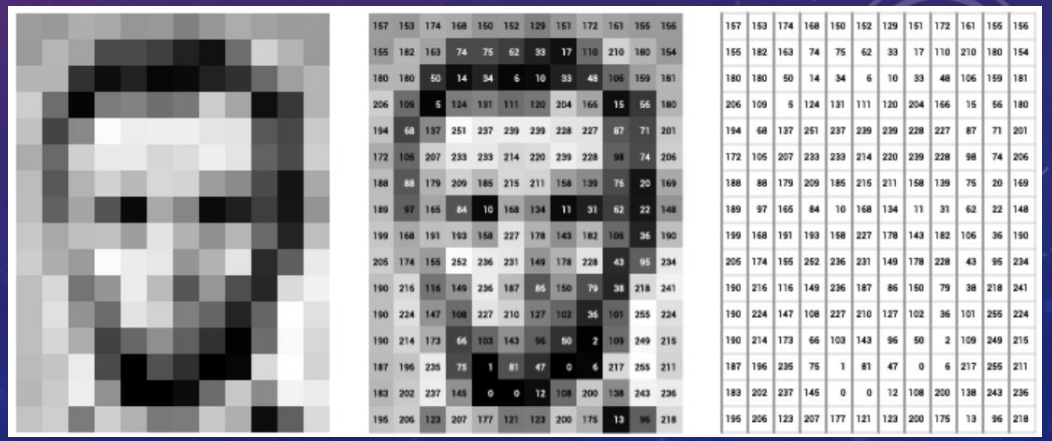
CASE STUDY: IMAGE RECOGNITION



- We could express the picture as a matrix of numbers where each number represents a pixel. However, this approach would probably fail if we fed the matrix into a traditional machine learning model. There's simply too much input data!
- We need more advanced algorithms from a subfield of machine learning called deep learning. In deep learning, multiple layers of mini-algorithms, called "neurons", work together to draw complex conclusions. Deep learning takes much, much more training data than a traditional machine learning model, but is also able to learn relationships that traditional models cannot. Deep learning is used to solve data-intensive problems, such as image classification or language understanding.

Deep Learning

1. Many neurons work together
2. Requires much more training data.
3. Used in complex problems
   a. Image Classification
   b. Language Learning/Understanding



EXERCISE

- Laura manages an analytics team and has a few tasks that she's hoping to achieve this quarter. The tasks are centered around the following domains:
  - **traditional machine learning**
  - **deep learning**
  - **Internet of Things (IoT)**
- The knowledge to build traditional machine learning and deep learning applications is present within her team. There is another team in the company that is specialized in working with IoT data. Laura wants to know for which tasks she'll need their help.

Traditional Machine Learning

Internet of Things

Deep Learning

Predict ride-sharing prices at a certain time and location based on previous prices TML

Cluster patients by symptoms to help doctors select a treatment TML

Automatically summarize text from news articles DL

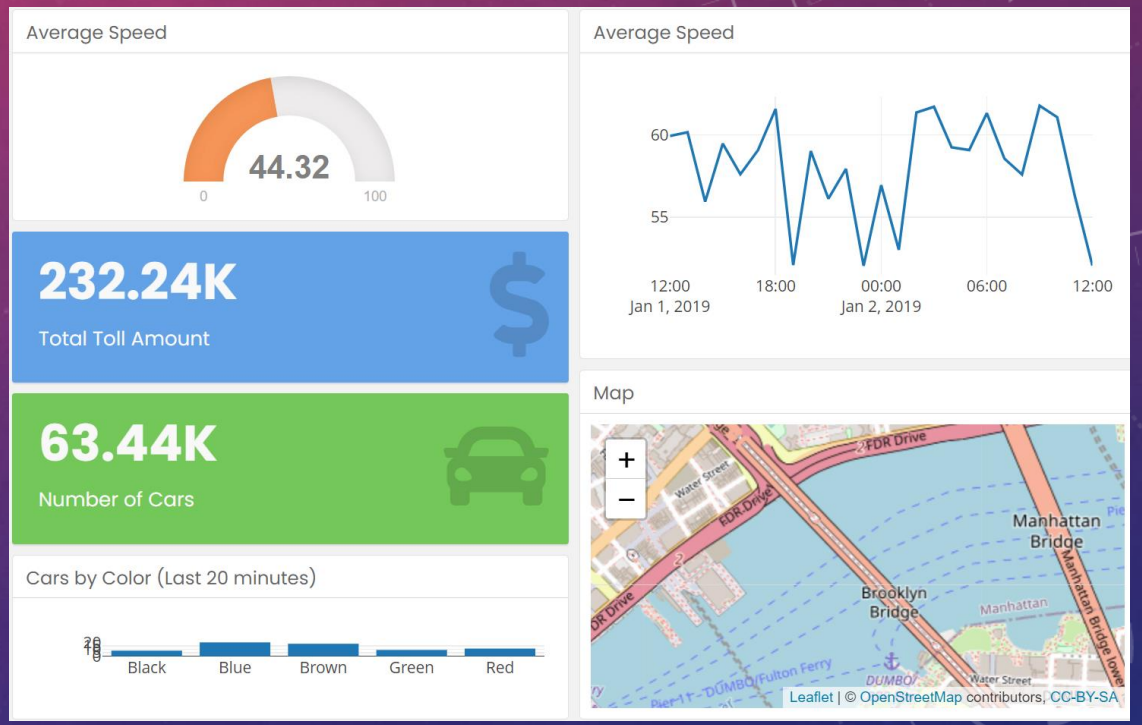
Detect machinery failure with vibration detectors

Automate building cooling using temperature sensors IoT

Detect machinery failure with vibration detectors IoT

EXERCISE

- **Investment research**
- Greg is an investment analyst. He's been asked to review a new start-up that uses easy-to-install vibration sensors to measure the amount of traffic on a bridge or highway. His boss has asked him to do some background research and decide which category this start-up belongs in.
- Take a look at the startup's dashboard and help Greg answer this question. Some of the plots in the dashboard are interactive and may change as you interact with them. You can always reset any plot to its original state by double-clicking it.
- Which category best fits this start-up?



1. Traditional machine learning
2. Deep learning
3. Internet of Things.
4. Natural language processing

ROLES IN DATA SCIENCE

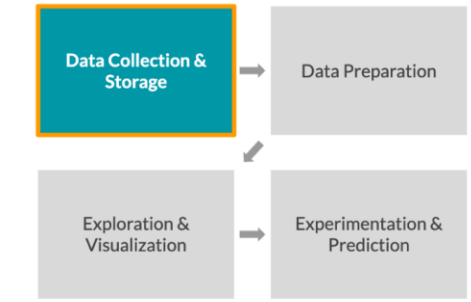
- Data Engineer
- Data Analyst
- Data Scientist
- Machine Learning Scientist/Engineer

DATA ENGINEER

- Data engineers control the flow of data: they build custom data pipelines and storage systems. They design infrastructure so that data is not only collected, but easy to obtain and process. Within the data science workflow, they focus on the first stage: data collection and storage.

Data engineer

- Information architects
- Build data pipelines and storage solutions
- Maintain data access



Data engineering tools

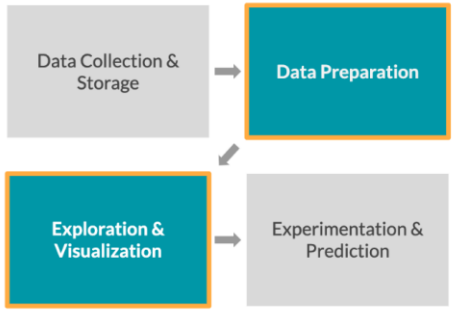
- **SQL**
  - To store and organize data
- **Java**, **Scala**, or **Python**
  - Programming languages to process data
- **Shell**
  - Command line to automate and run tasks
- **Cloud computing**
  - AWS, Azure, Google Cloud Platform

DATA ANALYST

- Data analysts describe the present via data. They do this by exploring the data and creating visualizations and dashboards. To do these tasks, they often have to clean data first. Analysts have less programming and stats experience than the other roles. Within the workflow, they focus on the middle two stages: data preparation and exploration and visualization.

Data analyst

- Perform simpler analyses that describe data
- Create reports and dashboards to summarize data
- Clean data for analysis



Data analyst tools

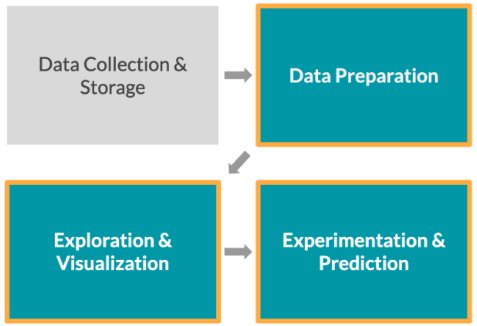
- **SQL**
  - Retrieve and aggregate data
- **Spreadsheets (Excel or Google Sheets)**
  - Simple analysis
- **BI tools (Tableau, Power BI, Looker)**
  - Dashboards and visualizations
- *May have:* Python or R
  - Clean and analyze data

DATA SCIENTIST

- Data Scientists have a strong background in statistics, enabling them to find new insights from data, rather than solely describing data. They also use traditional machine learning for prediction and forecasting. Within the workflow, they focus on the last three stages: data preparation and exploration and visualization, and experimentation and prediction.

Data scientist

- Versed in statistical methods
- Run experiments and analyses for insights
- Traditional machine learning



Data scientist tools

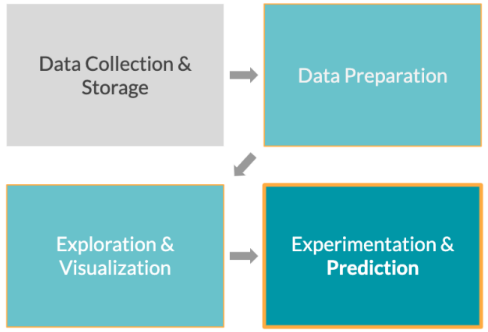
- **SQL**
  - Retrieve and aggregate data
- **Python and/or R**
  - Data science libraries, e.g., `pandas` (Python) and `tidyverse` (R)

MACHINE LEARNING SCIENTIST

- Machine learning scientists are similar to data scientists, but with a machine learning specialization. Machine learning is perhaps the buzziest part of Data Science; it's used to extrapolate what's likely to be true from what we already know. These scientists use training data to classify larger, unrulier data, whether its to classify images that contain a car, or create a chatbot. They go beyond traditional machine learning with deep learning. Within the workflow, they do the last three stages with a strong focus on prediction.

Machine learning scientist

- Predictions and extrapolations
- Classification
- Deep learning
  - Image processing
  - Natural language processing



Machine learning tools

- **Python and/or R**
  - Machine learning libraries, e.g., TensorFlow or Spark

Data Engineer	Data Analyst	Data Scientist	Machine Learning Scientist
Store and maintain data	Visualize and describe data	Gain insights from data	Predict with data
SQL + Java/Scala/Python	SQL + BI Tools + Spreadsheets	Python/R	Python/R

Matching skills to jobs

Johan manages a Data Science team and is looking to post some new job listings for a **Data Engineer** and a **Machine Learning Scientist**. Help Johan decide which skill requirements belong with each job.

Data Engineer

Data Analyst

Data Scientist

Give new team members data base access

Update Excel spreadsheet with new graphs

Train an anomaly detection algorithm

Run a correlation analysis between weather and ice cream sales

Create a dashboard for the marketing team

Create a new table in the SQL database

Data Engineer	Data Analyst	Data Scientist	Machine Learning Scientist
Store and maintain data	Visualize and describe data	Gain insights from data	Predict with data
SQL + Java/Scala/Python	SQL + BI Tools + Spreadsheets	Python/R	Python/R

Matching skills to jobs

Johan manages a Data Science team and is looking to post some new job listings for a **Data Engineer** and a **Machine Learning Scientist**. Help Johan decide which skill requirements belong with each job.

Data Engineer

Data Scientist

- *Expert at building and maintaining SQL databases*
- *Some higher education in statistics*
- *Proficient in using Python for prediction and modeling*
- *Strong Java skills*
- *Experience in using TensorFlow for implementing deep learning architecture*

DATA COLLECTION

DATA SOURCES

Sources of data

Company data

- Collected by companies
- Helps them make data-driven decisions

Open data

- Free, open data sources
- Can be used, shared, and built-on by anyone

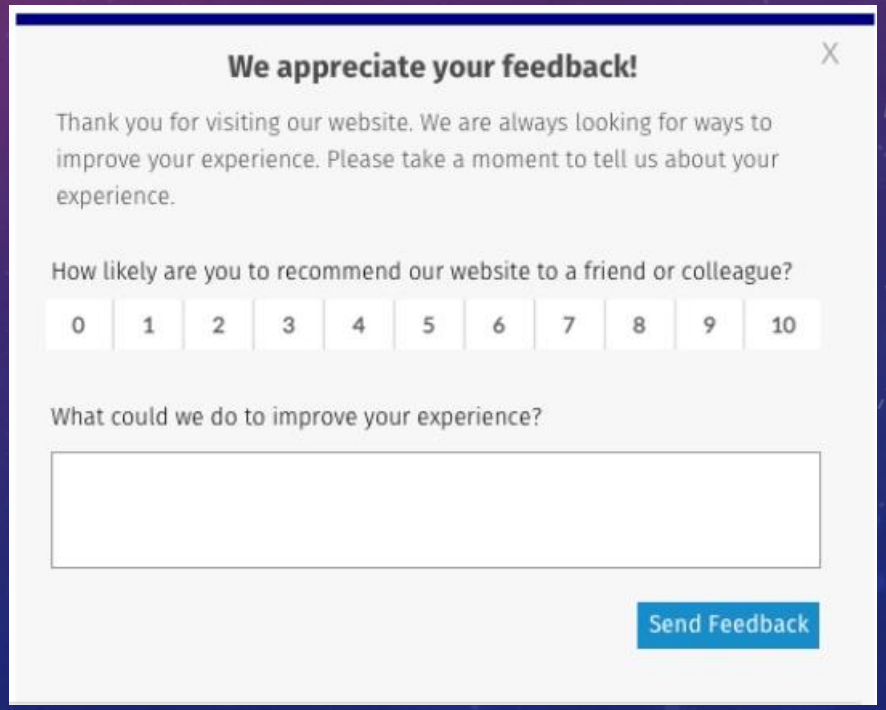
COMPANY DATA

- Some of the most common company sources of data are web events, survey data, customer data, logistics data, and financial transactions.
- Web Data - When you visit a web page or click on a link, usually this information is tracked by companies in order to calculate conversion rates or monitor the popularity of different pieces of content. The following information is captured: the name of the event, which could mean the URL of the page visited or an identifier for the element that was clicked, the timestamp of the event, and an identifier for the user that performed the action.
- Survey Data - Data can also be collected by asking people for their opinions in surveys. This can be, for example, in the form of a face-to-face interview, online questionnaire, or a focus group.

Web data

event_name	timestamp	user_id
homepage_visit	2019-01-01 12:01:01	1234

- Events
- Timestamps
- User information



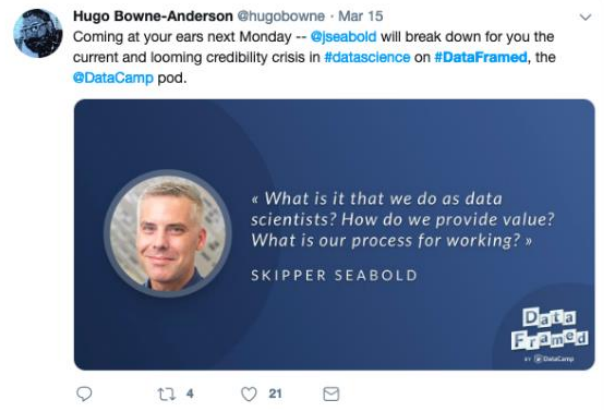
OPEN DATA

- There are multiple ways to access open data. Two of them are APIs and public records.
- Public data APIs - API stands for Application Programming Interface. It's an easy way of requesting data from a third party over the internet.
- Many companies have public APIs to let anyone access their data. Some notable APIs include Twitter, Wikipedia, Yahoo! Finance, and Google Maps, but there are many, many more.

- Sentimental analysis
- See if positive tweets are correlated with more downloads?

Tracking a hashtag

- All tweets with `#DataFramed` (DataCamp's podcast!)
- Use Twitter API



OPEN DATA

- Public records are another great way of gathering data. They can be collected and shared by international organizations like
  - the World Bank, the UN, or the WTO, national statistical offices, who use census and survey data, or
  - Government agencies, who make information about for example the weather, environment or population publicly available.
- For example, in the US, data.gov has health, education, and commerce data available for free download. In the EU, data.Europa.eu has similar data.

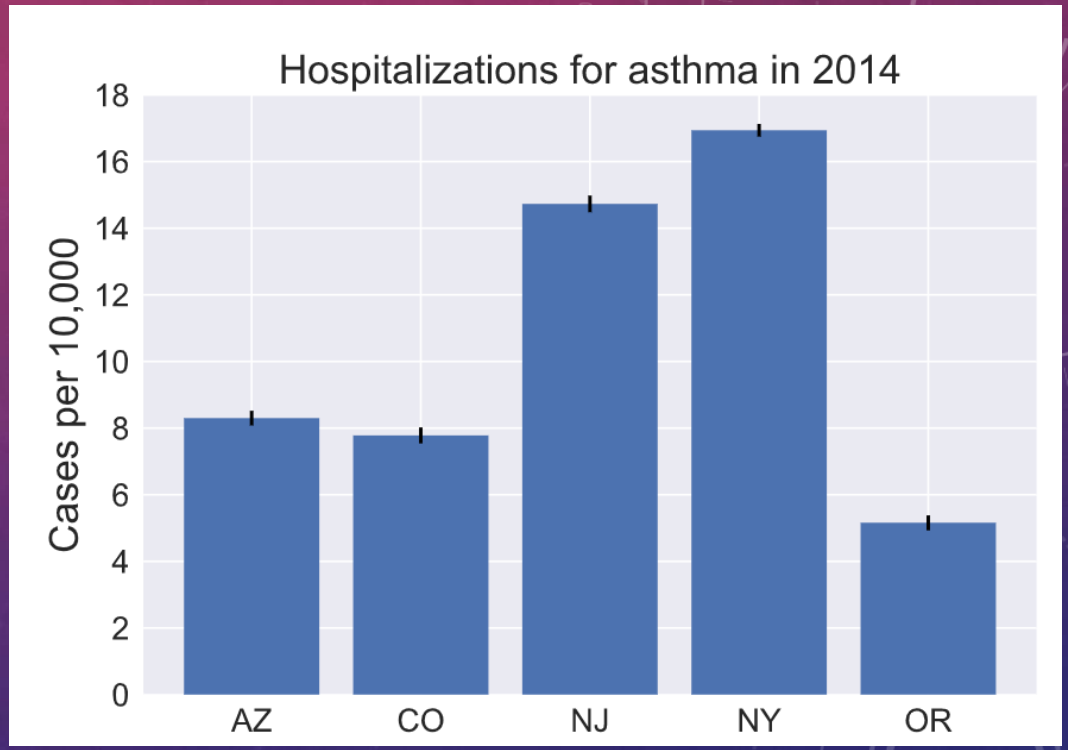
EXERCISE

- Company Data
- Open Data

- Average household income in Miami-Dade County
- Stock prices of different beauty brands over time
- Number of people clicking on the Coming Soon link on a company website.
- The number of customers that bought an iPhone in Bangkok in December last year
- Number of women between the ages of 30 – 45 living in Orange County
- How likely an Airbnb user is to recommend the website to a friend or colleague

EXERCISE

- **Asthma frequencies**
- You've realized by now that data can come from various sources, and not all of them are publicly available. A data science report contains the following visualization. With your new knowledge of data sources, you are able to identify where the data that's behind it originated.
- What source has the data scientist most likely used to collect this data?

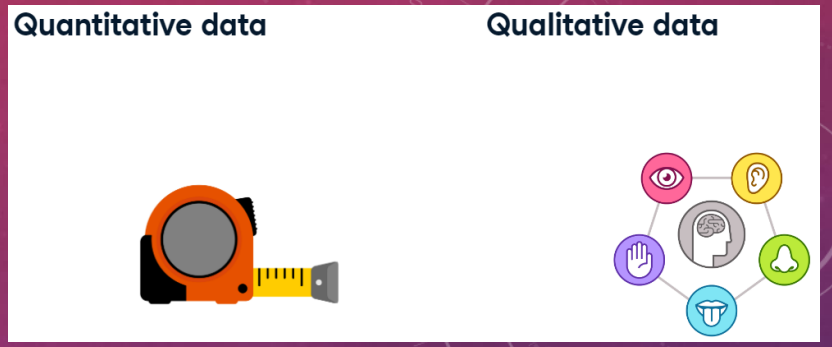


- **APIs**
- **Public records**
- **Web events**

DATA TYPES

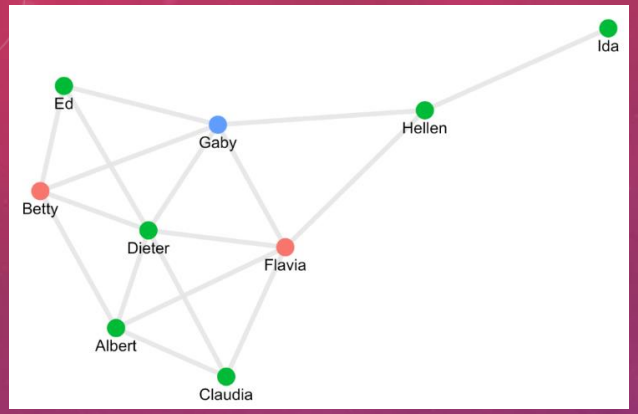
- Why care about data types?
  - Important when
    - Storing the data
    - Visualizing/analyzing the data

QUANTITATIVE VS. QUALITATIVE DATA

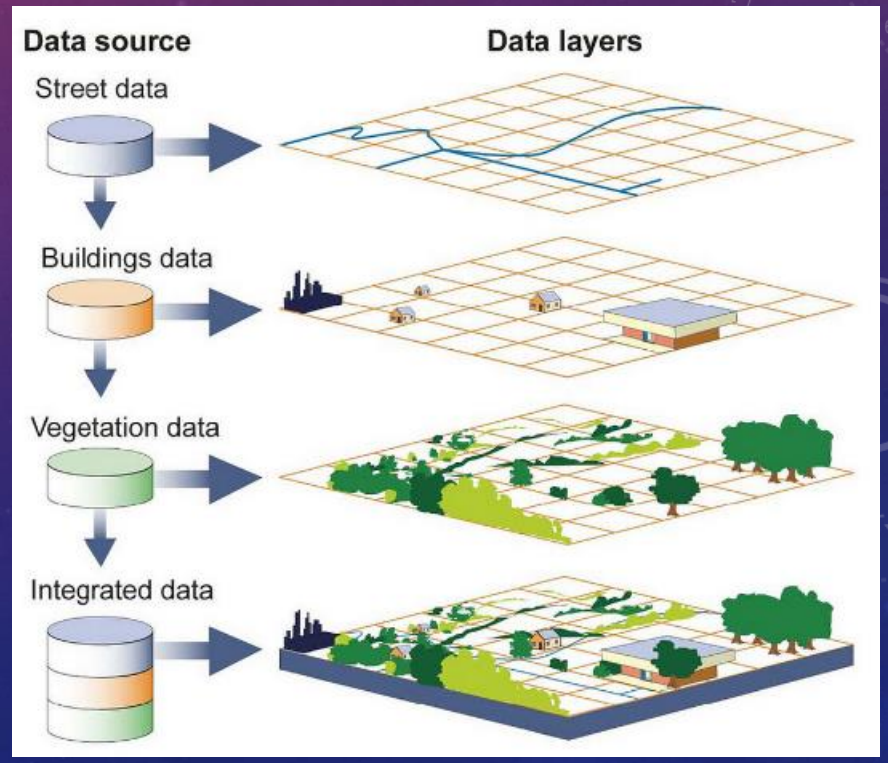


- There are two general types of data: qualitative and quantitative data. It's important to understand the key differences between both.
- Quantitative data can be counted, measured, and expressed using numbers. Qualitative data is descriptive and conceptual.
- Qualitative data can be observed but not measured.
- For example, **quantitative** data can be expressed in numbers. For example, the fridge is 60 inches tall, has two apples in it, and costs 1000 dollars. **Qualitative data**, on the other hand, are things that can be observed but not measured like the fridge is red, was built in Italy, and might need to be cleaned out because it smells like fish.

OTHER DATA TYPES?



- Other than the traditional quantitative and qualitative data, there are many other data types that are becoming more and more important.
  - Image data
  - Text data
  - Geospatial data – data with location information
  - Network data – e.g., Social network
  - and many more



EXERCISE – CLASSIFYING DATA TYPES

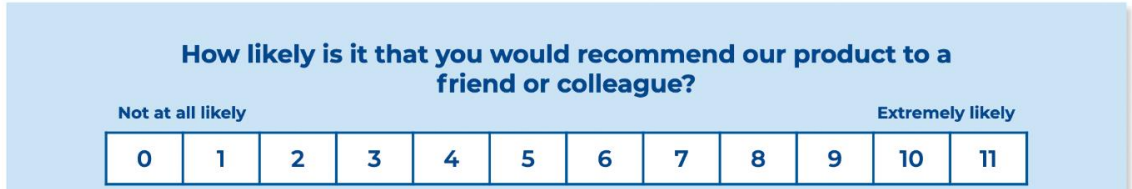
(number)

- Quantitative Data
- Qualitative Data

- The reviews for a property on Airbnb
- The price of a cup of coffee in StarLord Coffee
- The maximum speed of formula one cars during 2022 season
- The individual weight of all students in the class
- The daily average temperature in Bangkok during summer this year
- Images of several cats
- The eye color of people participating in a study
- The gender of students in Data Science fields

How likely is it that you would recommend [insert brand/website/service/product] to a friend or colleague?

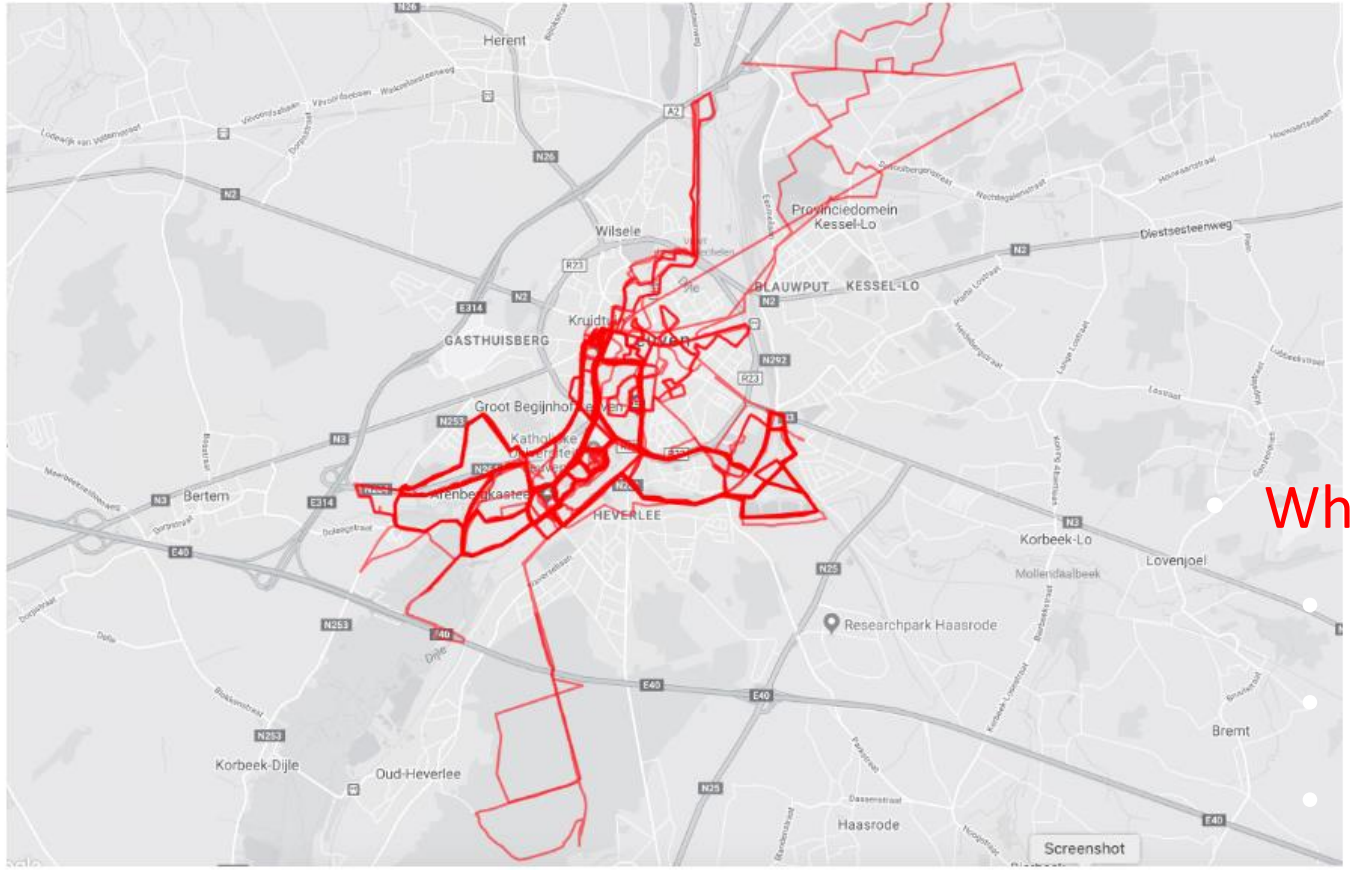
Users respond on a scale of 0 - 10 with 0 being not at all likely to recommend and 10 being extremely likely to recommend.



Activity tracker

Jane's New Year's resolution this year was to get into the best shape of her life. To help her achieve this goal she decided to invest in an activity tracker. After some months of tracking her activity, there is quite some data available.

The company that manufactured the activity tracker has a public API that allows access to your personal data. Jane is specifically interested in the GPS data of her runs because she wants to make a heatmap showing her most common running routes.



- Which one?
- Image data
- Text data
- Geospatial data
- Network data

What type of data will she be extracting from the API?

DATA STORAGE AND RETRIEVAL

Data Type	Query Language
Document Database	NoSQL
Relational Database	SQL

- When storing data there are multiple things to take into consideration.
  - Data storage location – single PC, a cluster of servers, cloud storage.
  - Data type
    - Unstructured – Email, text, video and audio files, web pages, social media, etc. typically stored in Document database
    - Structured – Relational database such as MySQL, PostgreSQL
  - Retrieval
    - At a basic level, we'll want to be able to request a specific piece of data, such as "All of the images that were created on March 3rd" or "All of the customer addresses in Montana".
    - NoSQL, SQL

Cloud platforms

Jerome has collected a lot of data for a data science project he's working on. His goal is to build a face recognition algorithm and to do that he has collected thousands of images. He needs to decide on which Cloud provider to choose for storing the data.

Which of the following is NOT an example of a Cloud provider?

(1) Google Cloud (2) Oracle Database (3) AWS S3 (4) Azure Storage (5) Tencent Cloud (6) IBM SQL Server

Querying a database

In this exercise, you'll see a command for selecting some data from the database. You can use the drop-down menus to modify the command. Whenever you want to see the results of the query, click "run query".

Based on your investigation of this database, which of the following statements is most accurate? Complete the following sentence:

This is a ____, and the query is written in ____.

- Relational Database, NoSQL
- Relational Database, SQL
- Document Database, NoSQL
- Document Database, SQL

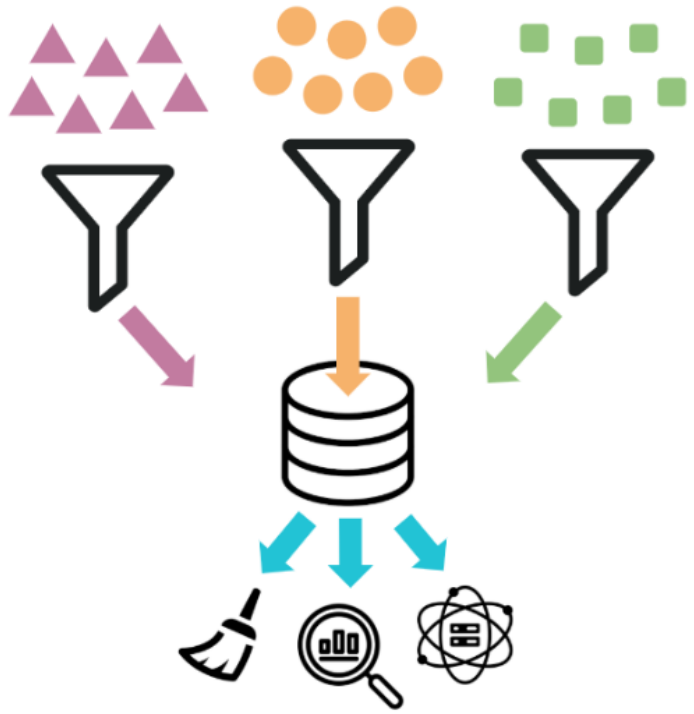
Relational Database

- Text from various emails sent and received by you
- The dates, times, subjects and recipient addresses for all emails you ever sent.
- Customer information for all students of VMS such as name, phone number and location.

Document Database

- Images of different traffic events, including metadata about the image’s contents

HOW DO WE SCALE DATA COLLECTION – DATA PIPELINES



How do we scale?

More than one data source:

- Public records
- APIs
- Databases

Different data types:

- Unstructured data
- Tabular data
- Real-time streaming data *e.g., tweets*

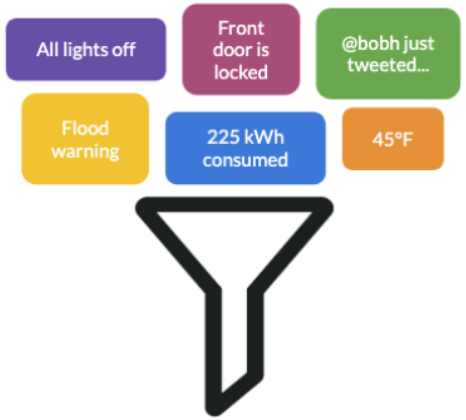
WHAT IS DATA PIPELINE? For Data Engineer

- Moves data into defined stages – for example, from data ingestion through an API to loading data into a database. A key feature is that pipelines automate this movement.
- Automated collection and storage - it would be tedious to ask a data engineer to manually run programs to collect and store data.
  - Scheduled hourly, daily, weekly, etc.
  - Triggered by an event
- Monitored with generated alerted – for example, when 95% of storage capacity has been reach or if an API is responding with an error.
- Necessary for big data projects – when working with a lot of data from different sources.
- Data engineers work to customize solutions.
- Extract Transform Load (ETL)

Case study: smart home

Data	Source	Frequency
Weather conditions	National Weather Service API	Every 30 minutes
Tweets in your area	Twitter API	Real-time stream
Indoor temperature	Smart home thermostat	Every 5 minutes
Status of lights	Smart light bulbs	Every minute
Status of locks	Smart door locks	Every 15 seconds
Energy consumption	Smart meter	Weekly

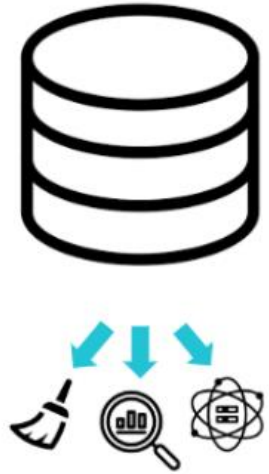
Extract



Extract

Source	Frequency
National Weather API	Every 30 minutes
Twitter API	Real-time stream
Smart home thermostat	Every 5 minutes
Smart light bulbs	Every minute
Smart door locks	Every 15 seconds
Smart meter	Weekly

Transform



Load

DATA PIPELINES – TRANSFORM & LOAD

- With all the data coming in, how de we keep it organized and easy to use? Example transformations:
  - Joining data sources into one data set.
  - Converting data structures to fit database schemas.
  - Removing irrelevant data - For example, the Twitter API, not only gives you a tweet but other details, like the number of followers the author has, which is not useful in this scenario.
- Note: Data preparation and exploration does not occur at this stage.
- Finally, we load the data into storage so that it can be used for visualization and analysis.
- Once we've set up all those steps, we automate. For example, we can say every time we get a tweet, we transform it in a certain way and store it in a specific table in our database.

DATA PIPELINE CHARACTERISTICS

- Which of the following statements is true?
  - A data pipeline is essential for every data science project. X
  - In the transform phase of ETL, data analysts perform exploratory data analysis. X For data engineer
  - Data engineers design and build custom data pipelines for projects. ✓
  - Data pipelines do not require automation. X

- Extract
- Transform
- Load

- Use the Apple Store API to get the latest download and rating information.
- Combine data on different users into one data set for all users.
- Collect the latest songs listened to by users on the mobile app.
- Group and arrange user listening data by musician.
- Store listening data into a database used by machine learning scientists to generate personalized playlists.

DATA PREPARATION

WHY PREPARE DATA?

Let's start cleaning

	Sara	Lis	Hadrien	Lis
Age	"27"	"30"		"30"
Size	1.77	5.58	1.80	5.58
Country	"Belgium"	"USA"	"FR"	"USA"

- Real-life data is messy. Data rarely comes in ready for analysis.
- Data preparation is done to prevent:
  - errors,
  - incorrect results,
  - biasing algorithms.

KEYS

- Tidiness
- Remove duplicates
- Unique ID

Tidy data output

Before

	Sara	Lis	Hadrien	Lis
Age	"27"	"30"		"30"
Size	1.77	5.58	1.80	5.58
Country	"Belgium"	"USA"	"FR"	"USA"

After

Name	Age	Size	Country
Sara	"26"	1.78	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"
Lis	"30"	5.58	"USA"

Remove duplicates | output

Before

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"
Lis	"30"	5.58	"USA"

After

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"

Unique ID | output

Before

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"

After

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	5.58	"USA"
2	Hadrien		1.80	"FR"

KEYS

- Homogeneity
  - Different units used in different countries
  - Abbreviations
  - Data types
    - str, int
  - Missing values

Homogeneity | output

Before

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	5.58	"USA"
2	Hadrien		1.80	"FR"

After

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	1.70	"USA"
2	Hadrien		1.80	"FR"

Homogeneity, again | output

Before

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	1.70	"US"
2	Hadrien		1.80	"FR"

After

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"BE"
1	Lis	"30"	1.70	"US"
2	Hadrien		1.80	"FR"

Data types | output

Before

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"BE"
1	Lis	"30"	1.70	"US"
2	Hadrien		1.80	"FR"

After

ID	Name	Age	Size	Country
0	Sara	27	1.77	"BE"
1	Lis	30	1.70	"US"
2	Hadrien		1.80	"FR"

Missing values | output

Before

ID	Name	Age	Size	Country
0	Sara	27	1.77	"BE"
1	Lis	30	1.70	"USA"
2	Hadrien		1.80	"FR"

After

ID	Name	Age	Size	Country
0	Sara	27	1.77	"BE"
1	Lis	30	1.70	"US"
2	Hadrien	28	1.80	"FR"

Last night, you were stargazing with a friend, and you saw something strange in the sky. You decide to look into UFO (Unidentified Flying Object) reports to see if anyone has seen something similar in your area. Luckily, you find a UFO reports dataset, but before analyzing it, you need to prepare the data! Here is a snapshot of the first 5 rows.

	summary	city	state	date_time	shape	duration	stats	report_link	text	posted	city_latitude
0	My wife was driving southeast on a fairly popu...	Chester	VA	2019-12-12T18:43:00	light	5 seconds	Occurred : 12/12/2019 18:43 (Entered as : 12/...	http://www.nuforc.org/webreports/151/S151739.html	My wife was driving southeast on a fairly popu...	2019-12-22T00:00:00	37.343152
1	I think that I may caught a UFO on the NBC Nig...	Rocky Hill	CT	2019-03-22T18:30:00	circle	3-5 seconds	Occurred : 3/22/2019 18:30 (Entered as : 03/2...	http://www.nuforc.org/webreports/145/S145297.html	I think that I may caught a UFO on the NBC Nig...	2019-03-29T00:00:00	41.664800
2	I woke up late in the afternoon 3:30-4pm. I we...	NaN	NaN	NaN	NaN	NaN	Occurred : 4/1/2019 15:45 (Entered as : April...	http://www.nuforc.org/webreports/145/S145556.html	I woke up late in the afternoon 3:30-4pm. I w...	NaN	NaN
3	I was driving towards the intersection of fall...	Ottawa	ON	2019-04-17T02:00:00	teardrop	10 seconds	Occurred : 4/17/2019 02:00 (Entered as : 04-1...	http://www.nuforc.org/webreports/145/S145697.html	I was driving towards the intersection of fall...	2019-04-18T00:00:00	45.381383
4	In Peoria Arizona, I saw a cigar shaped craft ...	Peoria	NY	2009-03-15T18:00:00	cigar	2 minutes	Occurred : 3/15/2009 18:00 (Entered as : 03/1...	http://www.nuforc.org/webreports/145/S145723.html	In Peoria, Arizona, I saw a cigar shaped craft...	2019-04-18T00:00:00	NaN

Based on the data shown in the previous slide, what should we do?

The last row is missing a value for `city_latitude`, you should drop it.

The third row is full of `NaN` values. A UFO sighting event has to have a place, a date, and a description (the shape of the object). This row does not have these information, you should drop it.

The third row is full of `NaN` values, which is unusual. You should try to understand why that is the case before taking the decision to drop or keep it.

These five observations report seeing aliens, so what you and your friend saw was probably aliens too.

Removing duplicates, verifying data types, handling missing values, ensuring values use the same measurement system are part of data preparation.

Data often comes in a ready to analyze state.

Missing values are tricky, but there are several ways of dealing with them.

Tidy data means having the features as rows and the observations as columns.

Country names should always be abbreviated.

Non-prepared data can lead to errors, biased algorithms and incorrect results.

Which of the above is TRUE or FALSE?

EXPLORATORY DATA ANALYSIS (EDA)

WHAT IS EDA?

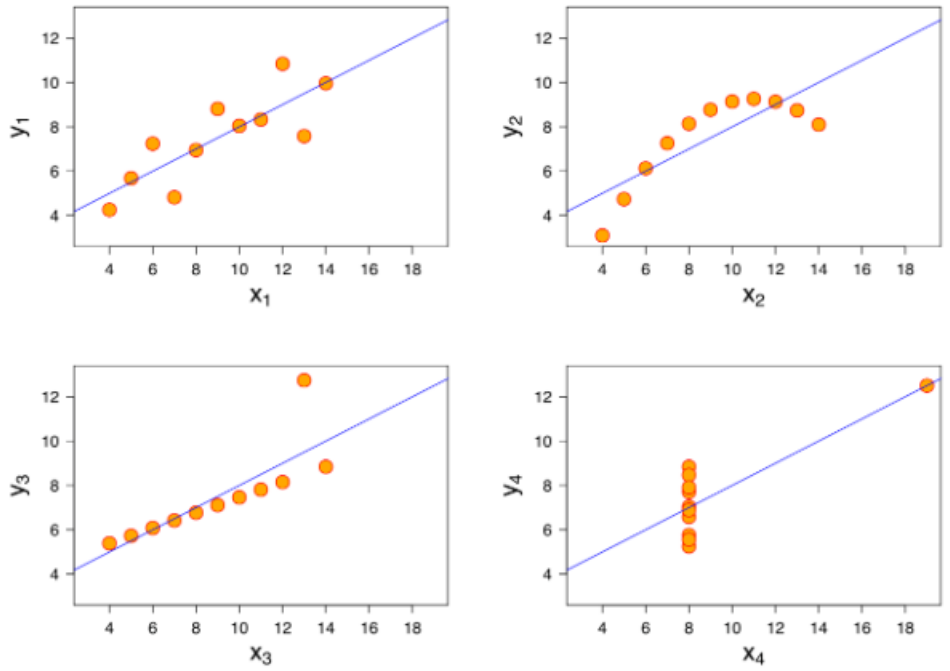
- It consists of
  - exploring the data and
  - formulating hypotheses about it, and
  - assessing its main characteristics, with a strong emphasis on visualization.

- All for data sets display:
  - Identical mean and variance for x
  - Identical mean and variance for y
  - Identical correlation coefficient
  - Identical linear regression equation
- In short: they look quite similar
- Note: Data till 3rd month of 2018

Dataset 1		Dataset 2		Dataset 3		Dataset 4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Dataset 1		Dataset 2		Dataset 3		Dataset 4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Anscombe's quartet



ANOTHER EXAMPLE – SPACEX LAUNCHES

- Knowing your data

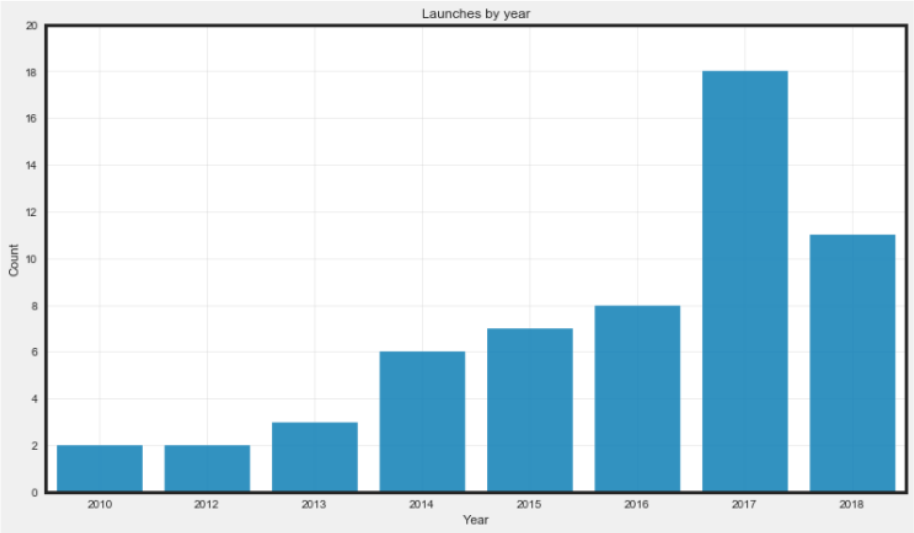
- Previewing your data

- Flight Number (number)
- Date (datetime)
- Time (UTC) (datetime)
- Booster Version (text)
- Launch Site (text)
- Payload (text)
- Payload Mass (kg) (number)
- Orbit (text)
- Customer (text)
- Mission Outcome (text)
- Landing Outcome (text)

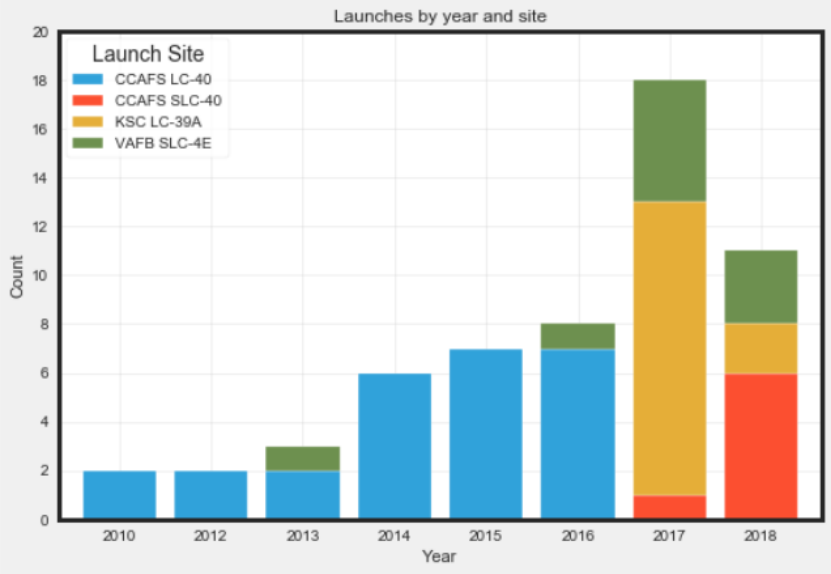
Flight	Date	Time (UTC)	Booster Version	Launch Site	Payload
1	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit
2	2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats...
3	2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2+
4	2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1
5	2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2

Payload Mass (kg)	Orbit	Customer	Mission Outcome	Landing Outcome
NaN	LEO	SpaceX	Success	Failure (parachute)
NaN	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
525	LEO (ISS)	NASA (COTS)	Success	No attempt
500	LEO (ISS)	NASA (CRS)	Success	No attempt
677	LEO (ISS)	NASA (CRS)	Success	No attempt

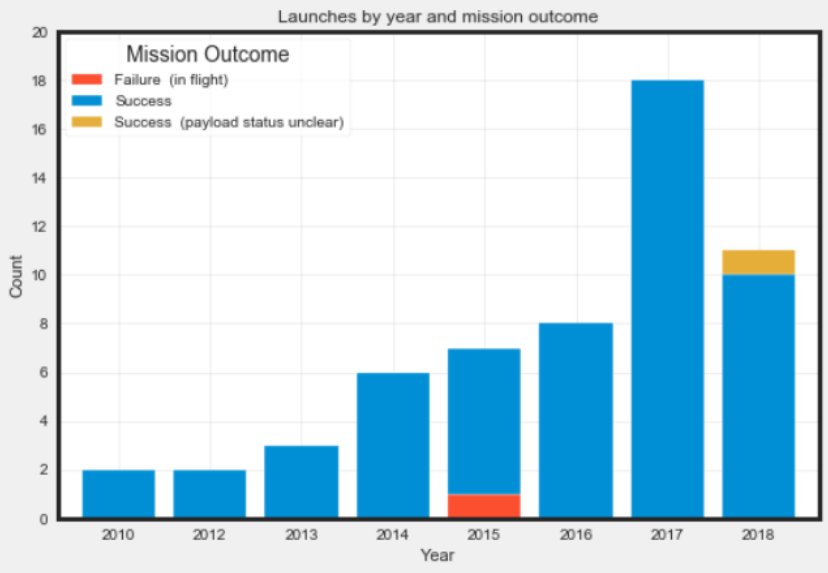
Visualize!



Ask more questions!

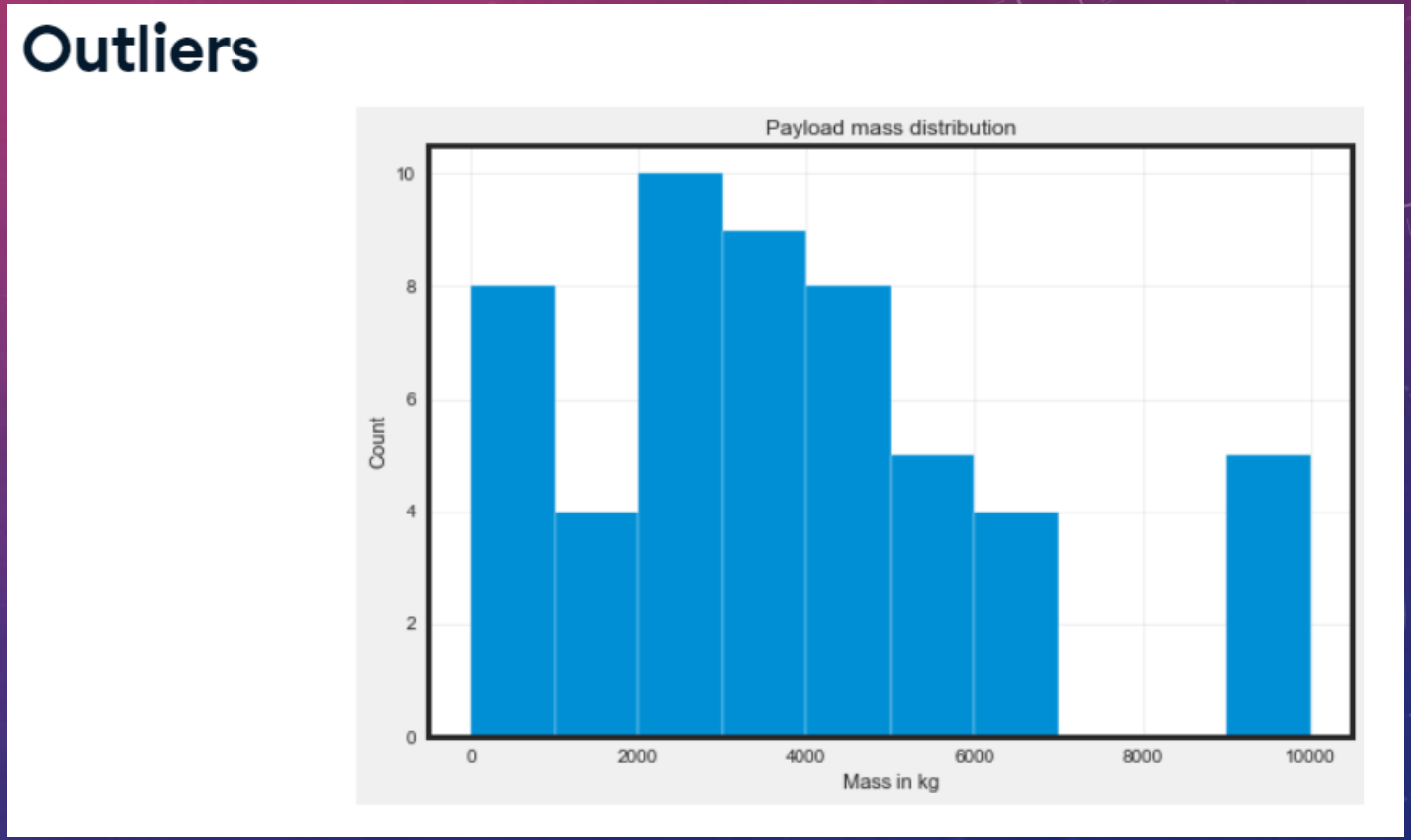


Ask more questions!



OUTLINERS – UNUSUAL VALUE

- Another thing you do during EDA is look for outliers, that is, unusual values. Whether they are errors or valid, it's nice to know about them, as they can throw your results off. Here, we can see we have only 5 launches with a weight greater than 7,000 kg, when the average mass is closer 3800 kg.



A few years ago, KIC 8462852, also known as Tabby's star, started decreasing strongly in luminosity. We're not sure why yet, but scientists are investigating by recording observations at different times and locations. Below are some descriptive statistics on the observations. Unlike the qualitative dataset we saw in the video, the Tabby's star dataset is made of quantitative data only, so we get statistics, like the mean and median.

Luminosity Observations	
count	52
mean	1.169356
std	1.230953
min	0.974110
25%	0.995637
50%	1.002465
75%	1.008685
max	9.874800

How many separate luminosity observations make up the dataset?

Visual EDA

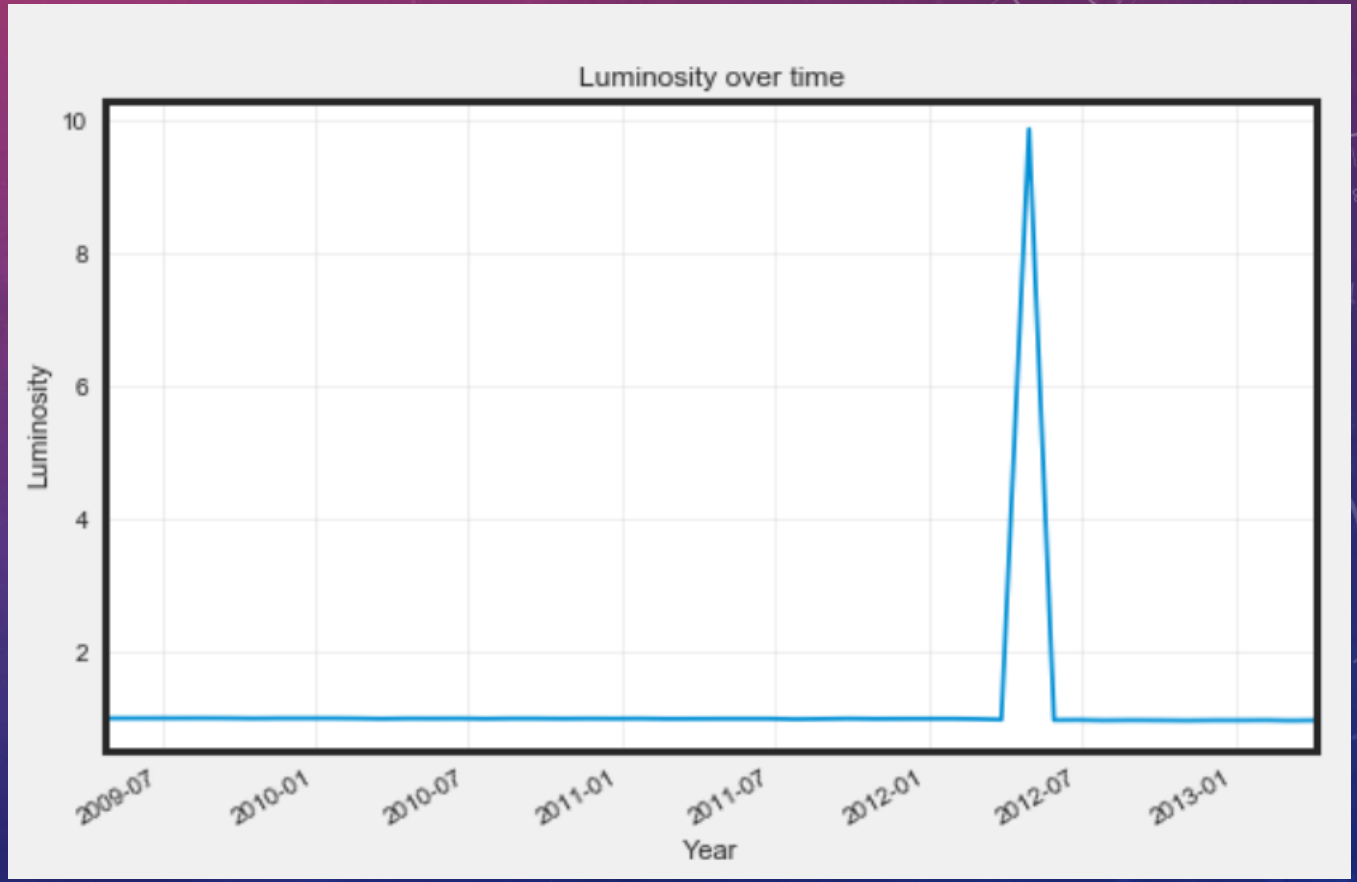
Your friend graphed the star's luminosity using data he copied from public sources. He shows you the graph below. He concludes that scientists have it completely backwards. Scientists say the star is dimming, as the graph shows the luminosity is actually pretty flat, except one moment when it increases a lot.

Which of the following statements is true?

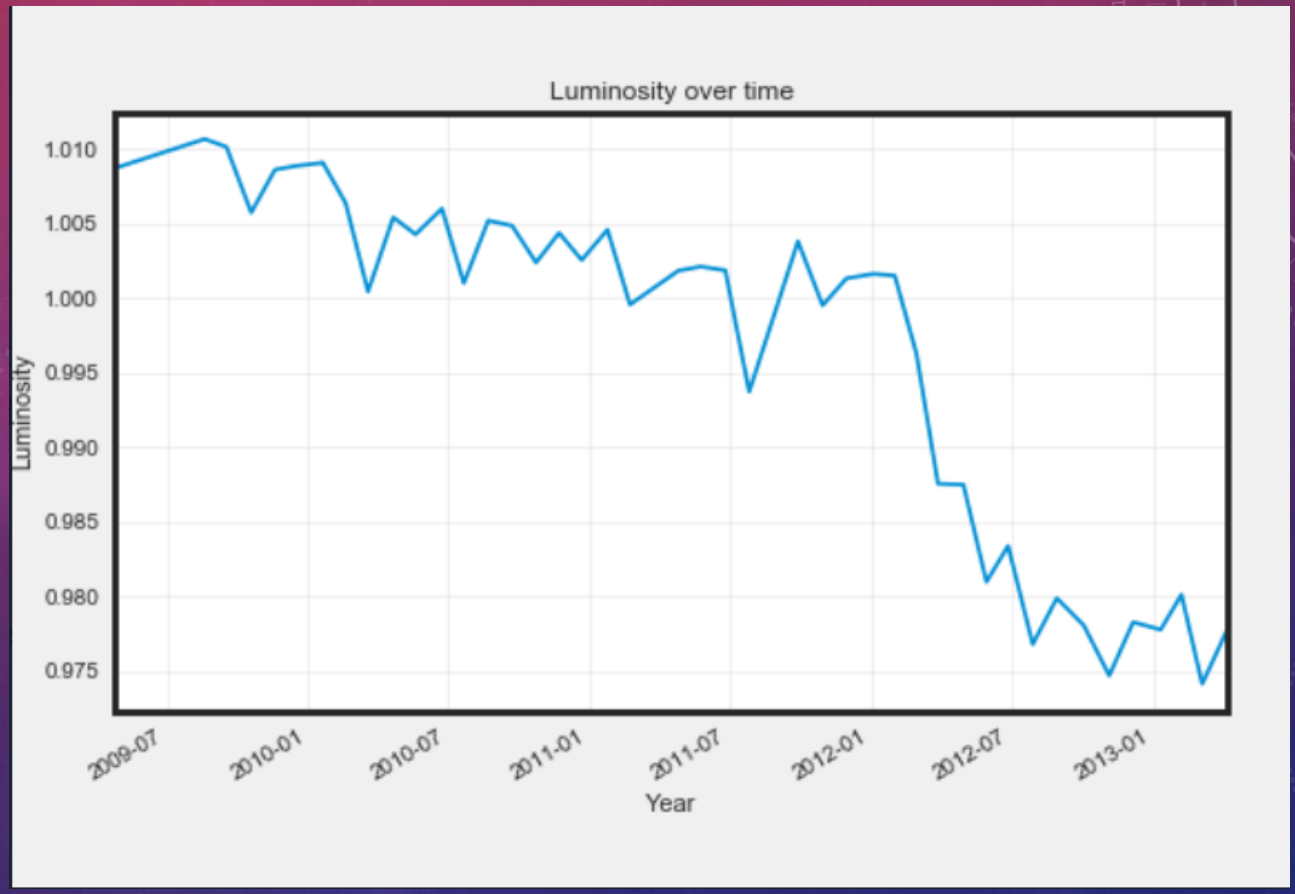
1. The data and graph has to be correct. As the graph shows, the star must have been stable, except for a brief luminous spike in mid-2012. Scientists should focus on what made that star spike in luminosity.

2. Something weird is going on. That jump here is an outlier and the data should be reviewed.

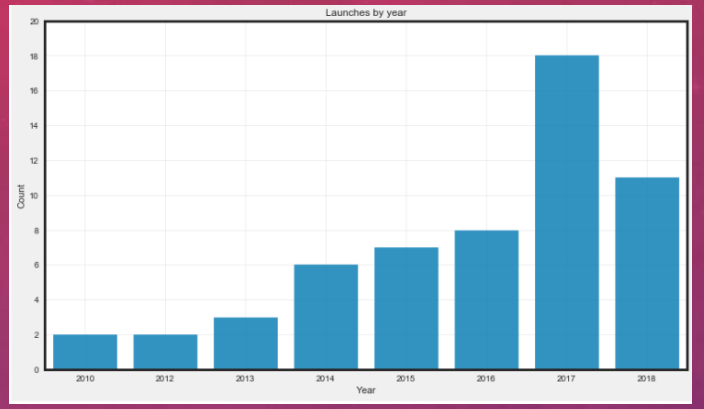
3. If scientists and journalists say the star is dimming, then the star is dimming. It does not matter that the data states otherwise.



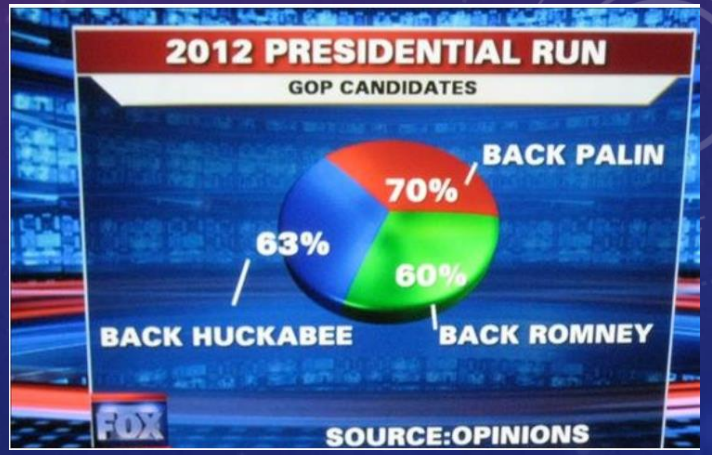
All values are centered around 1, except one occurrence at 9.8748. Turns out the correct value is 0.98748. Your friend misplaced the period when copying the values. After correcting, the graph looks much better, and the star is indeed dimming.



VISUALIZATION TIPS



- Use color purposefully
  - Colorblindness
- Readable fonts
- Label, Label, Label
- Title, x-Axis, y-Axis
- Easy-to-understand charts



Dashboards

BI tools

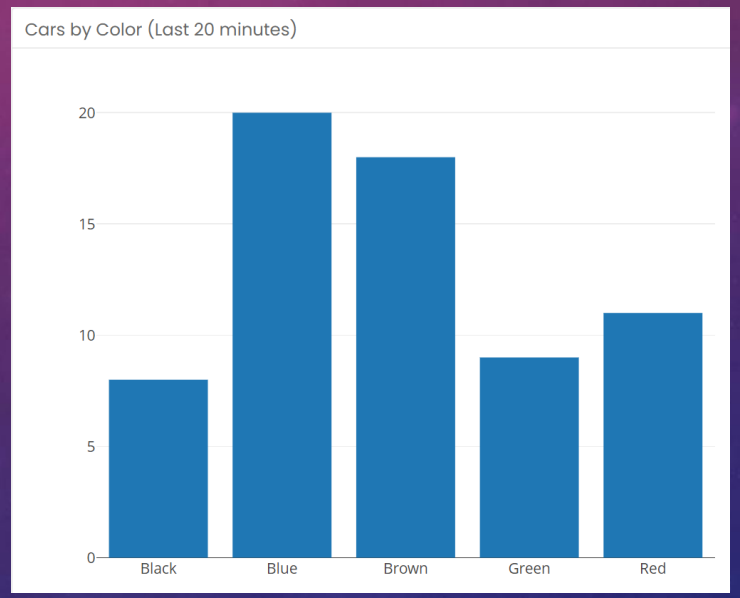
tableau

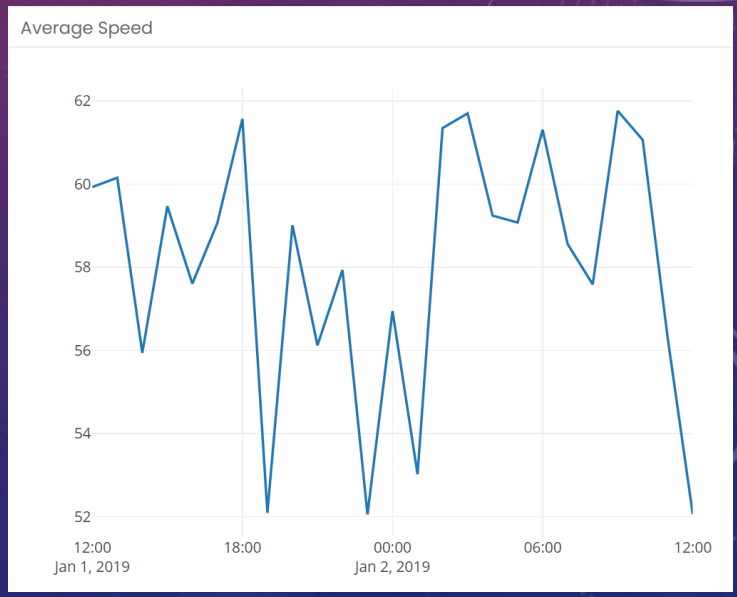
looker

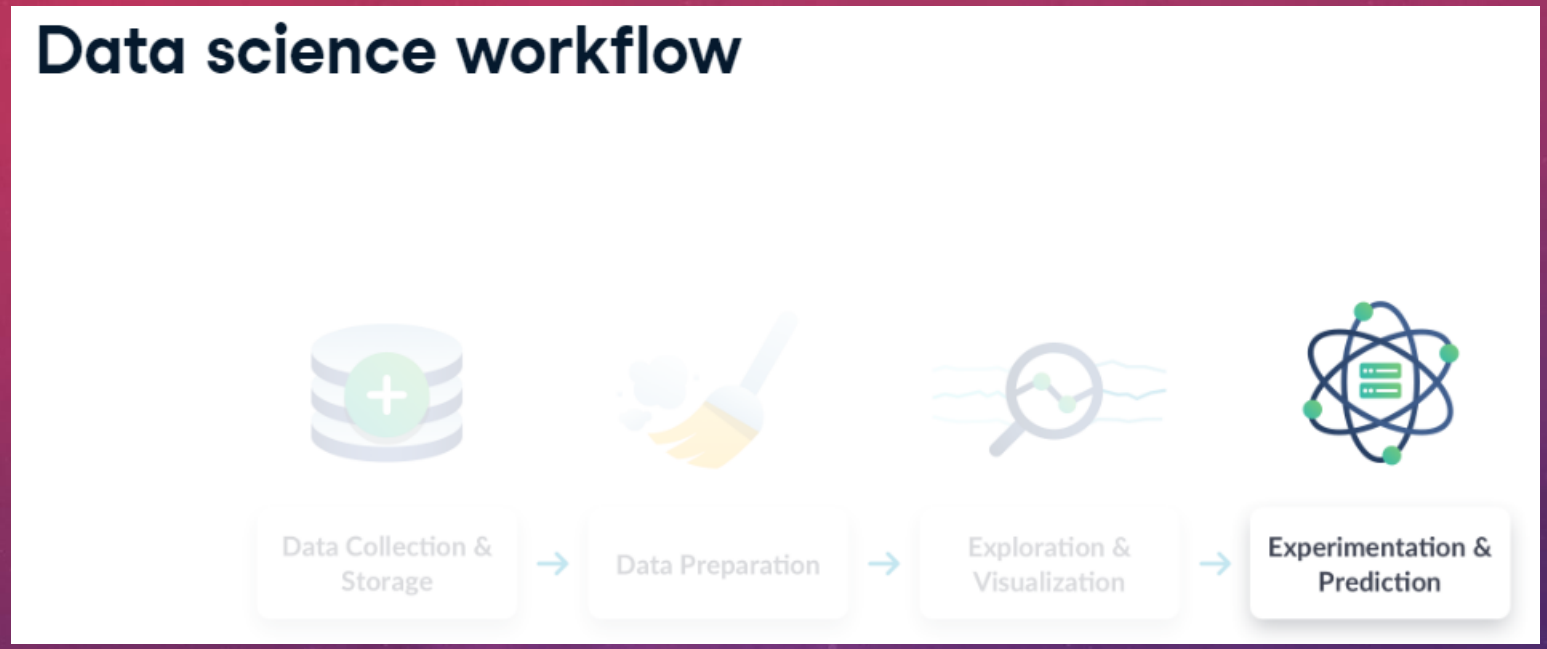
Power BI

Which suggestion would you not give?

1. The bars in the "Cars by Color" graph should match the car colors they represent.
2. MPH (miles per hours) should be added to the title and axis labels referring to average speed.
3. The "Total Toll Amount" and "Number of Cars" components need a date associated with them.
4. The y-axis in the top right "Average Speed" graph needs to start at 0, as opposed to 52.







EXPERIMENTAL AND PREDICTION

WHAT ARE EXPERIMENTS IN DATA SCIENCE

- Experiments help drive decisions and draw conclusions.
  1. Form a question
  2. Form a hypothesis
  3. Collect data
  4. Test the hypothesis with statistical test
  5. Interpret results

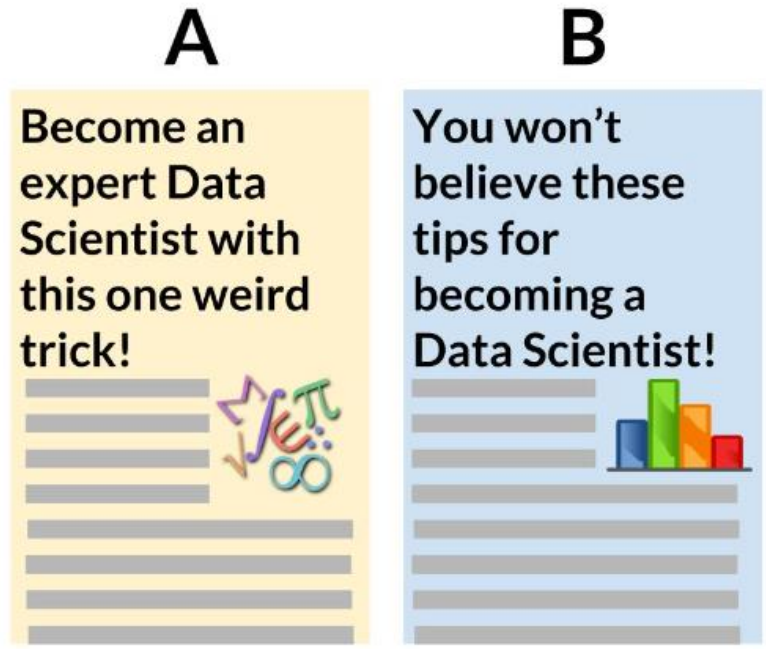
Case study: which is the better blog post title?

Form a question: Does blog title A or blog title B result in more clicks?

Form a hypothesis: Blog title A and blog title B result in the same amount of clicks.

Collect data:

- 50% users will see blog title A
- 50% users will see blog title B
- Track click-through rate until sample size reached

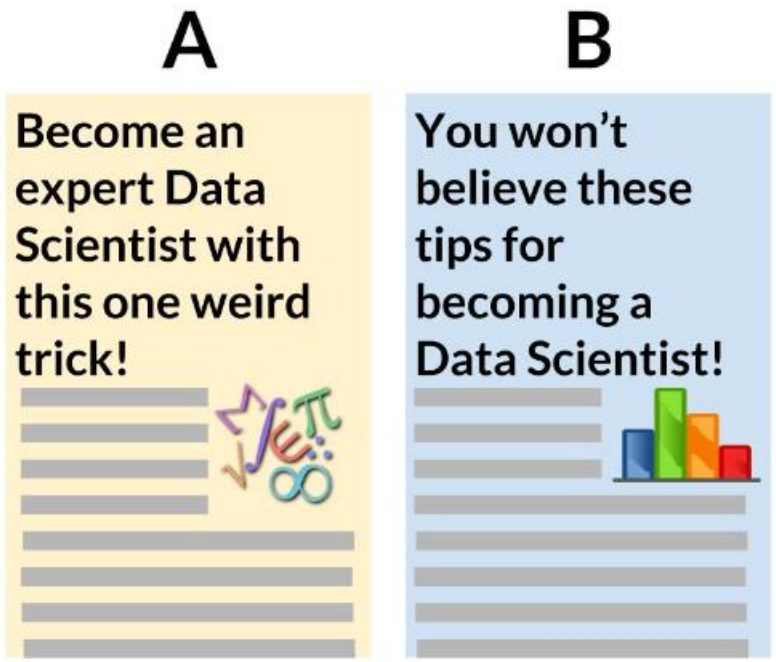


Case study: which is the better blog post title?

Test the hypothesis with a statistical test: Is the difference in titles' click-through rates significant?

Interpret results:

- Choose a title
- Or ask more questions and design another experiment!

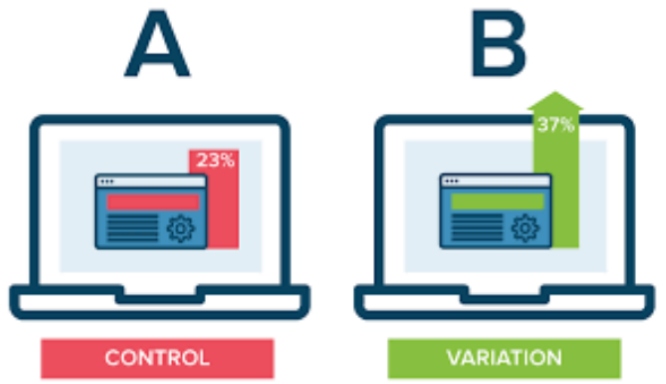


A/B Testing Steps

- Picking a metric to track
- Calculating sample size
- Running the experiment
- Checking for significance

What is A/B Testing?

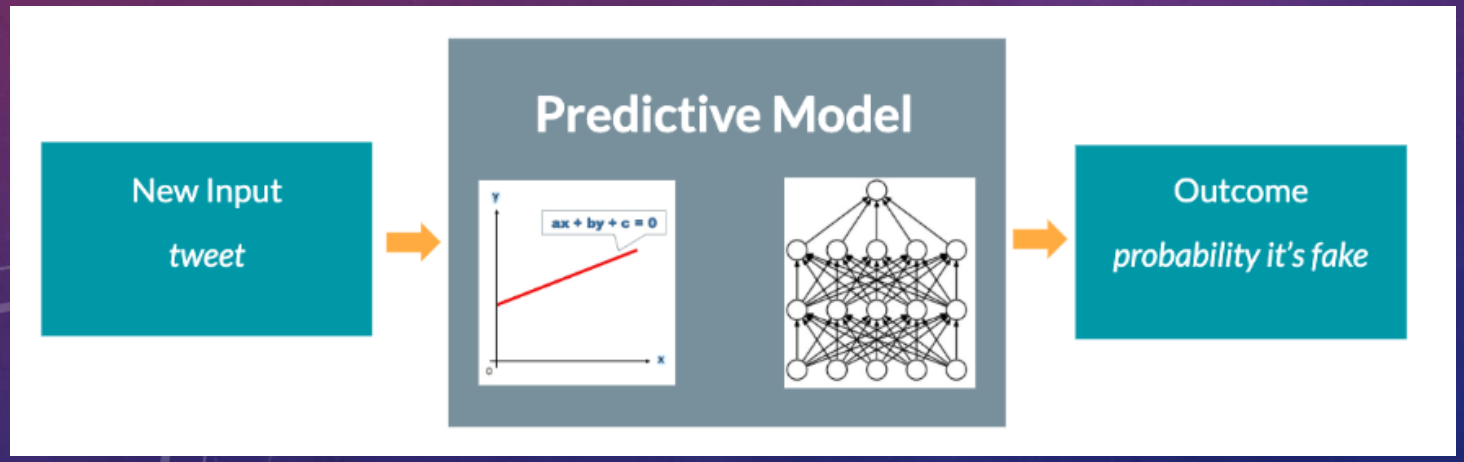
AKA Champion/Challenger Testing

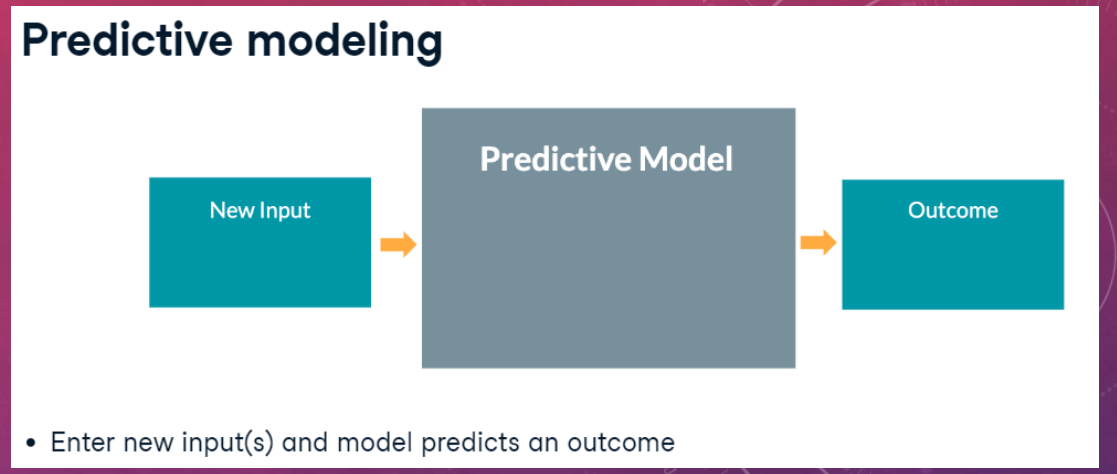


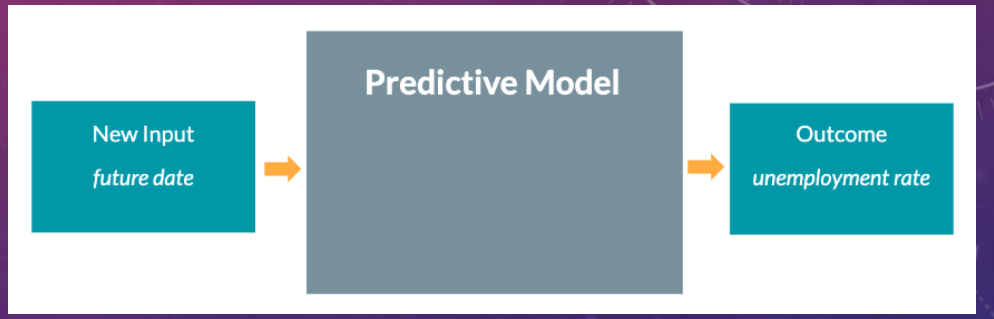
TIME SERIES FORECASTING

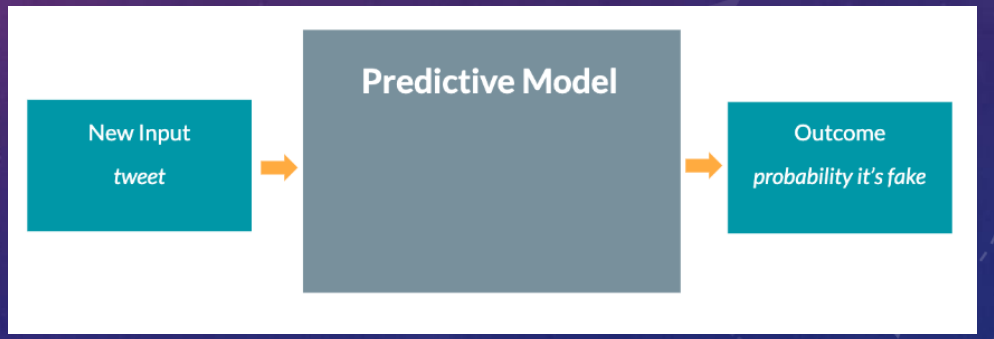
Modeling in DS

- Represent a real-world process with statistics
- Mathematical relationships between variables, including random variables.
- Based on statistical assumptions and historical data



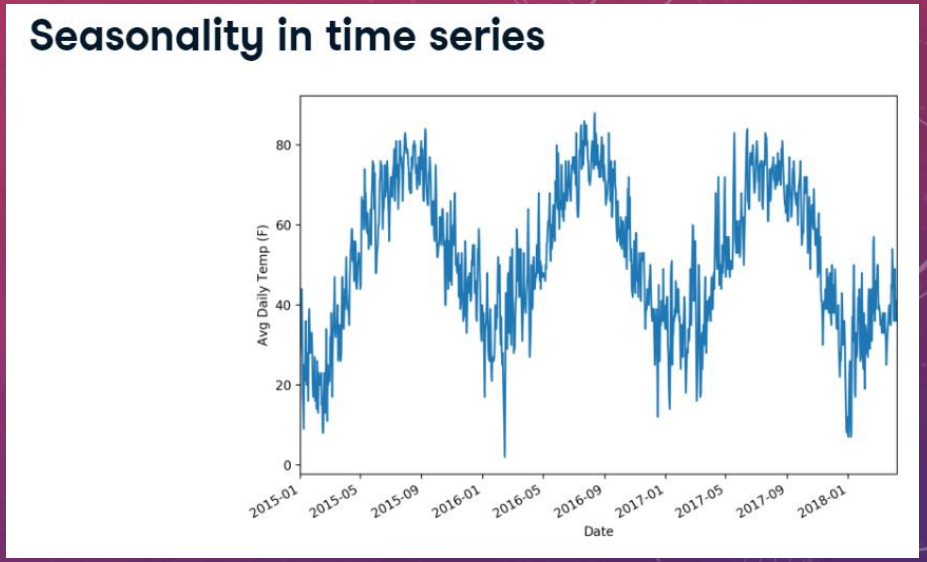




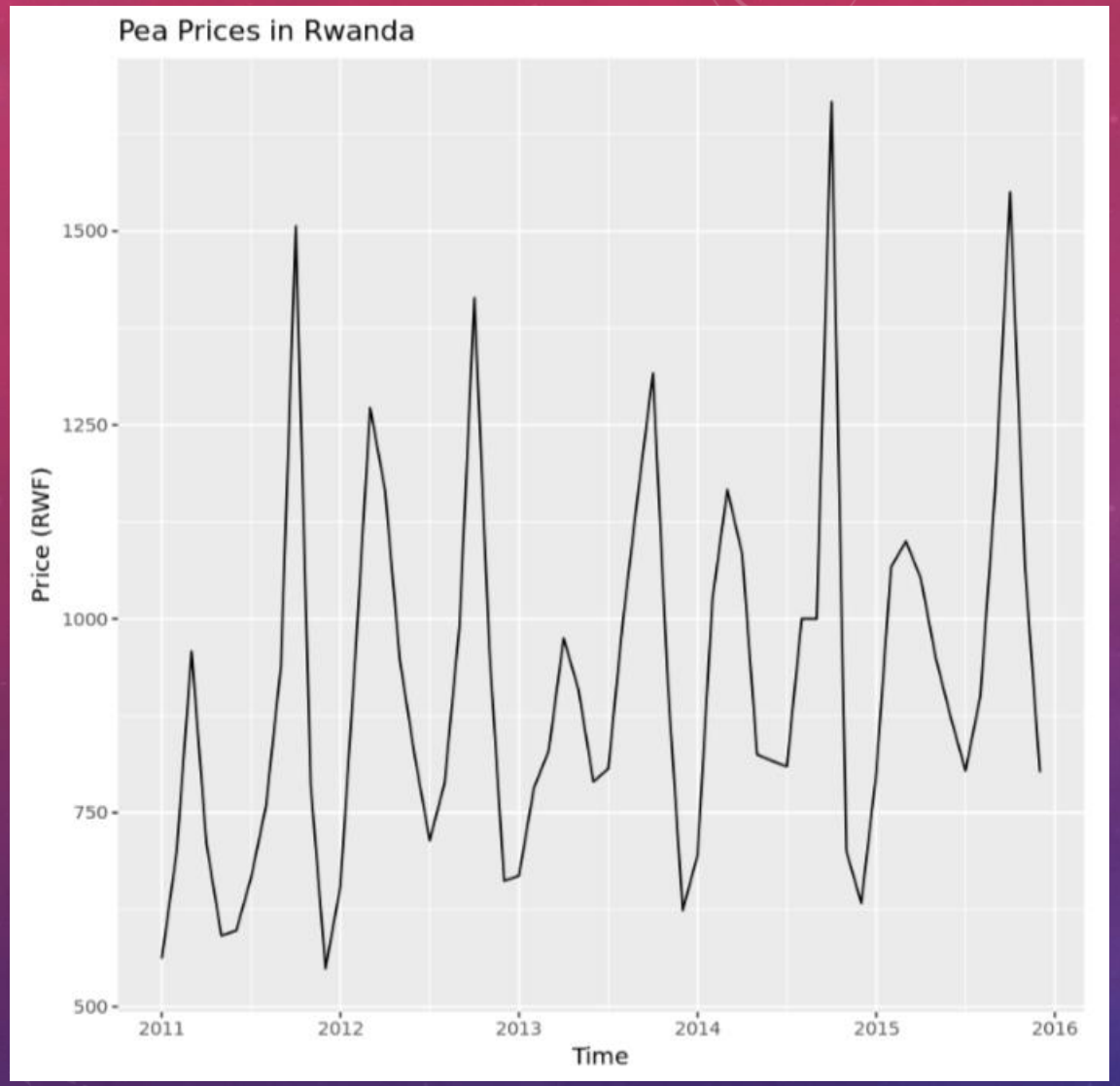


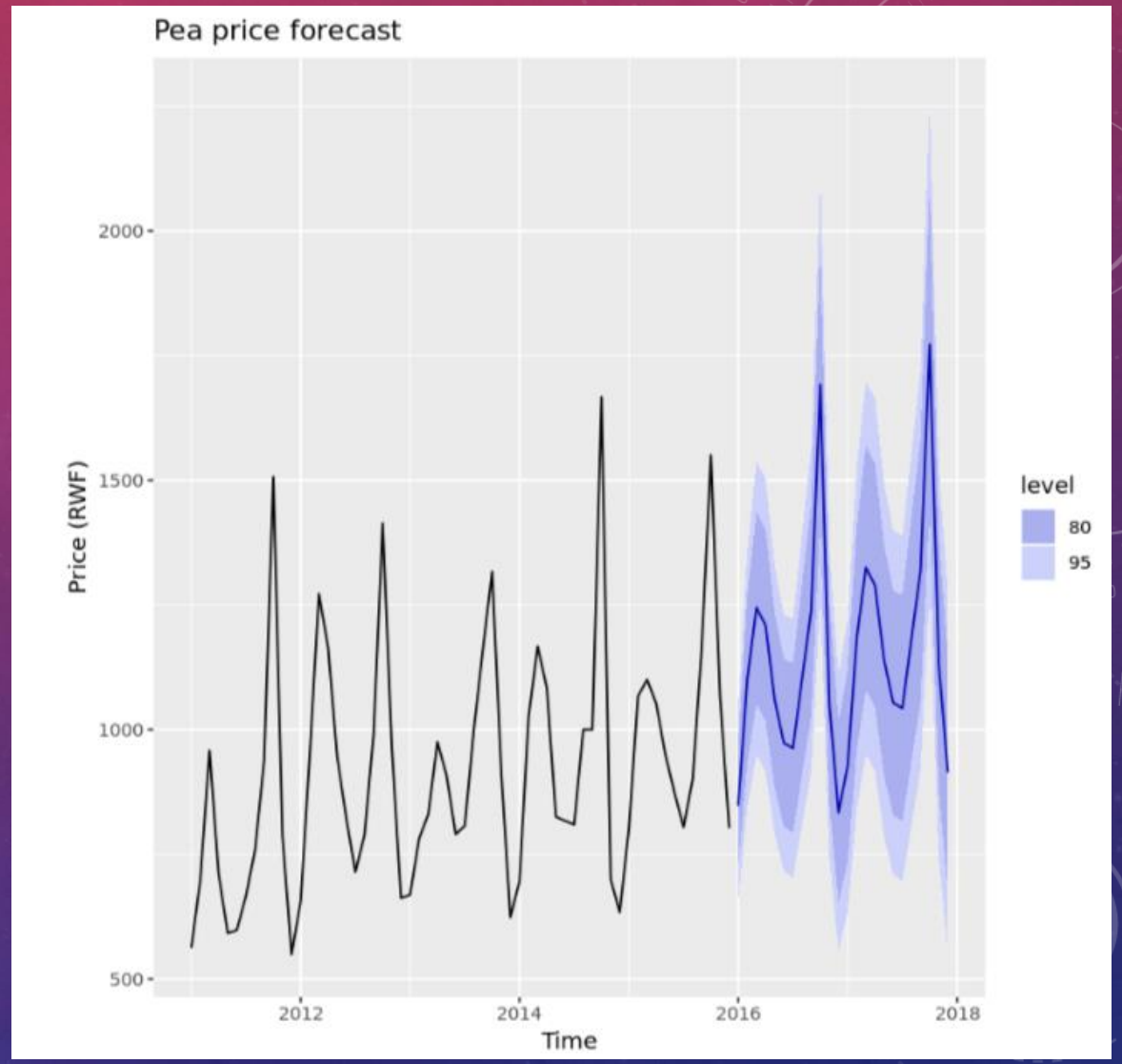
Ranges in complexity, from linear equation to a deep learning algorithm.

- Time series data
- A series of data points sequenced by time.
  - Stock prices
  - Gas prices
  - Unemployment rates
  - Heart rate
  - Inflation rate
  - CO2 levels
  - World temperature
  - Height of ocean tides
  - Fuel consumption during winter



- Forecasting time series
  - How much rainfall will we get next month?
  - Will traffic ease up in the next half hour?
  - How will the stock market move in the next six hours?
  - What will be earth's population in 20 years?
  - What will be % of EV cars in 5 years?
- Derive a model from historical data to generate predictions.
- Modeling methods use a combination of statistical and machine learning methods.





There are also two blue areas shown along the forecast. These are confidence intervals, which are extremely useful for evaluating predictions. We see two confidence intervals: 80% and 95%. The model is 80% sure that the true value will be in the area labeled as 80. Same goes with the area labeled as 95. If we're using this forecast to make big decisions, confidence intervals can help us buffer for the unexpected.

Which one is and is not Time series data?

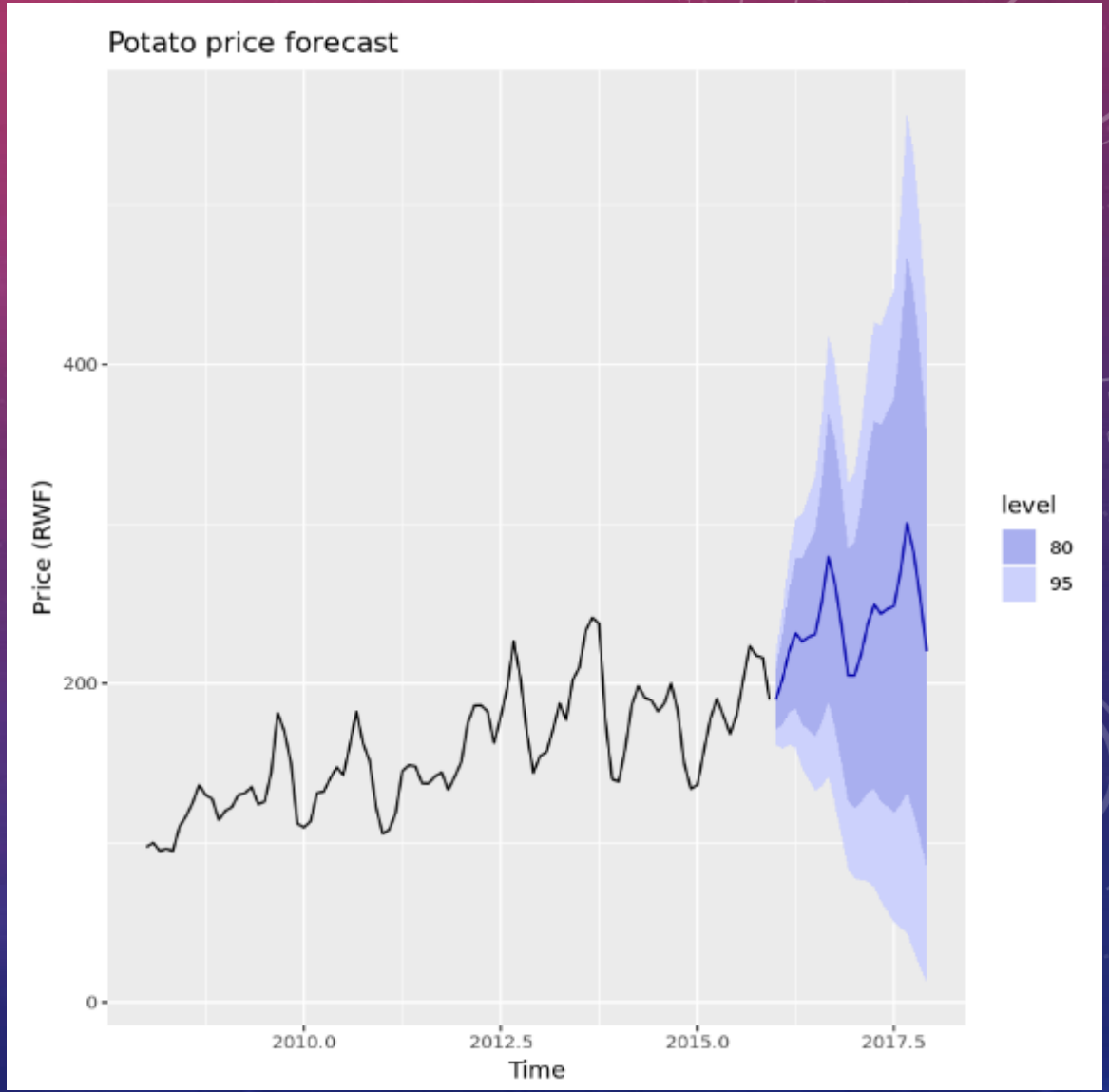
1. The price of an airline ticket during different seasons.
2. Weekly energy consumption.
3. Name and address of users.
4. Number of steps walked every day for the past year.
5. Geo-spatial of all hiking-trial in Yosemite park.
6. Food nutrients in Thai cuisine.
7. VMS intakes and retention rate for the last 5 years.
8. Engine RPM of different car models.

Interpret a time series plot

Below is a forecast of potato prices in Rwanda for the years 2016 and 2017.

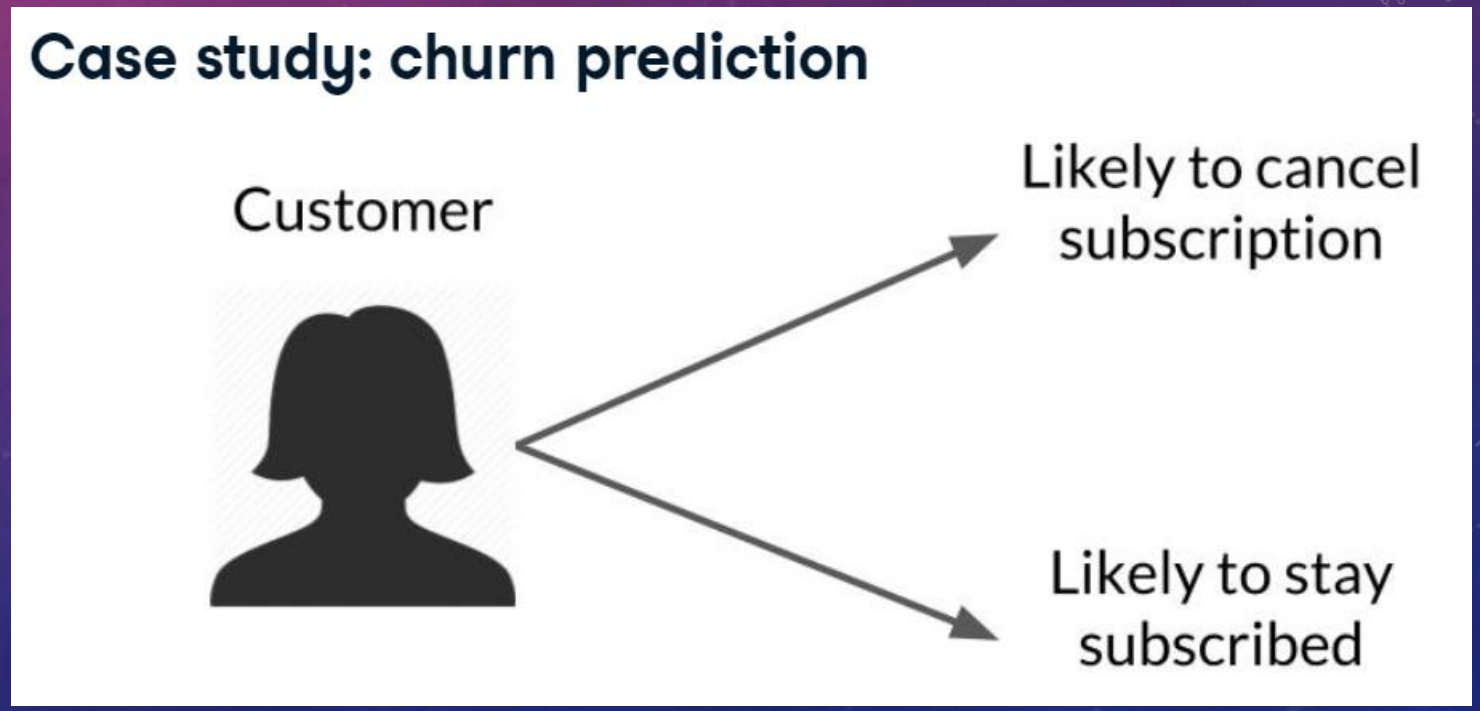
Which of the following statements is true about the model's prediction?

- The model is 95% confident that the highest the price will peak is 410 Rwandan francs.
- The model is 80% confident that the highest the price will peak is 450 Rwandan francs.
- The model is 80% confident that the lowest the price will go is 25 Rwandan francs.



WHAT IS SUPERVISED MACHINE LEARNING?

- Machine learning: Predictions from data
- Supervised machine learning: Predictions from data with labels and features.
  - Recommendation systems
  - Diagnosing biomedical images
  - Recognizing hand-written digits
  - Predicting customer churn



Training Data:
Customers

Labels
Customer outcomes

churn
subscribe
subscribe
churn
subscribe
churn

Age	Gender	Date of last purchase?	Date of last visit?	Likes cats?	Household $$	Location	Number of Kids	Profession	
									churn
									subscribe
									subscribe
									churn
									subscribe
									churn

Features:
Collected customer data

#	#	#	#	#	#	#	#	#

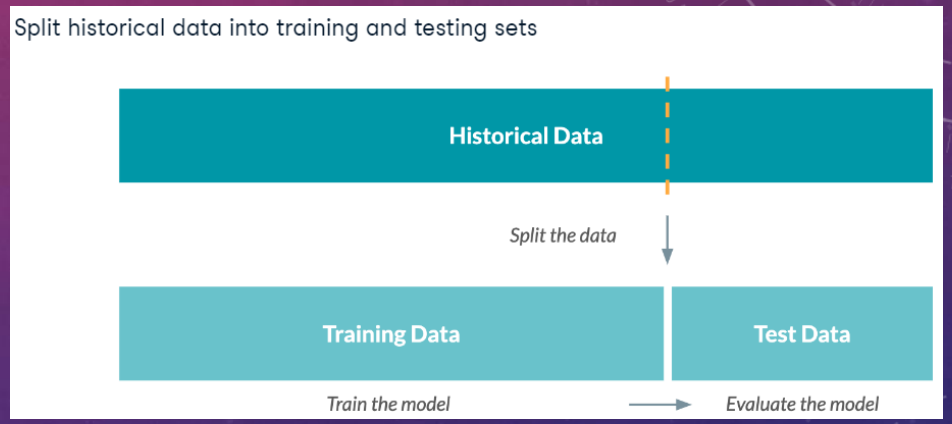
Trained Model

Prediction: Subscribe

HOW DO WE KNOW THE MODEL IS GOOD?

- Supervised machine learning recap
  - Make a prediction based on data
  - Data has features and labels
    - Label is what we want to predict.
    - Features are data that might predict the label.
  - Trained model can make predictions

- We need model evaluation.



Model Evaluation

Possible Labels	True Labels	Model Prediction	Model Accuracy
Customer remains	970	1000	# of correct predictions / # of predictions = 970/1000 = **97%**
Customer churns	30	0	

Which of the followings is A/B Testing, Supervised Machine Learning, and Dashboards.

1. **Select the menu format that leads to more online orders.**
2. **Test which shoes result in a faster marathon time.**
3. **Visually track hourly energy consumption to monitor for high usage.**
4. **Use statistics on past game performance to decide which team is most likely to win the world cup.**
5. **Based on achieved tweet, predict whether a tweet was tweeted by a bot.**
6. **Study VMS statistics to forecast next year retention rate.**
7. **Decide if Windows laptop or Macbook should be given to VMS freshmen.**
8. **VMS website showing statistical results and graphs of student's satisfaction survey.**

Features and labels

So far, you learned that features and labels are key elements in supervised machine learning. Furthermore, it's important to understand the difference between them when building a model.

Table

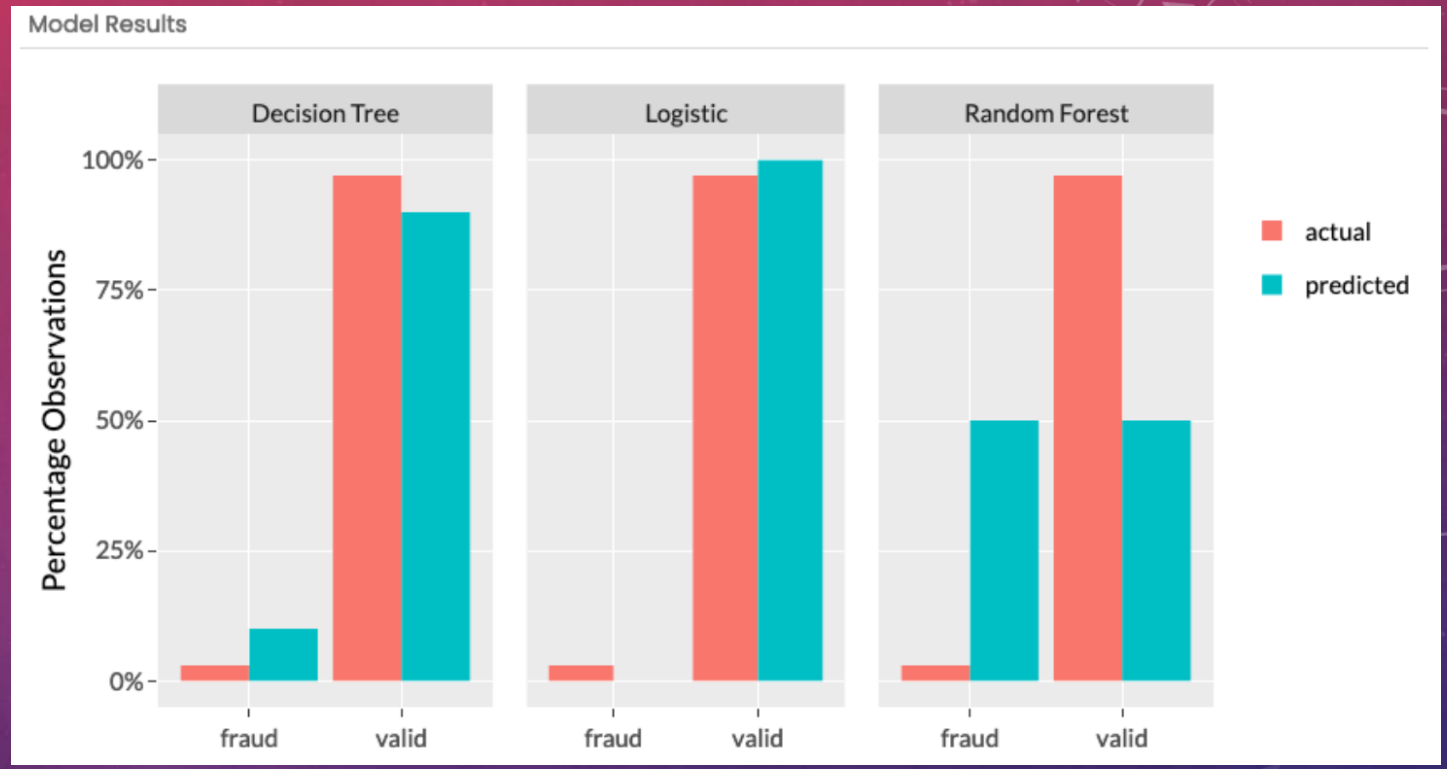
City	Marital Status	Income	Industry	Age	Loan Type
○ Feature ○ Label	○ Feature ○ Label	○ Feature ○ Label	○ Feature ○ Label	○ Feature ○ Label	○ Feature ○ Label
Cardiff	Single	$35,000 to $49,999	Broadcasting	35-44	College
Lauw	Divorced	$200,000 or more	Other Manufacturing	25-34	Boat
Les Bons Villers	Single	$75,000 to $99,999	Mining	35-44	Vacation
San Massimo	Common-Law	$75,000 to $99,999	Homemaker	45-54	Small Business
Dessau	Common-Law	$35,000 to $49,999	Retail	35-44	College

If you want to build a recommendation system that predicts the industry someone works in based on other personal information. You have a dataset that you want to use to train his model. Try to identify which column contains labels.

Model Evaluation

You are evaluating three different supervised machine learning models for detecting fraudulent bank transactions. You knows that one of the weaknesses of the training data was that there are many more examples of valid transactions compared to fraudulent transactions.

On the right, you can see how each model predicted that a transaction was fraudulent or valid compared to how often the transaction actually was fraudulent or valid:



According to this data, which model is the best at detecting both fraudulent and valid transactions, and why?

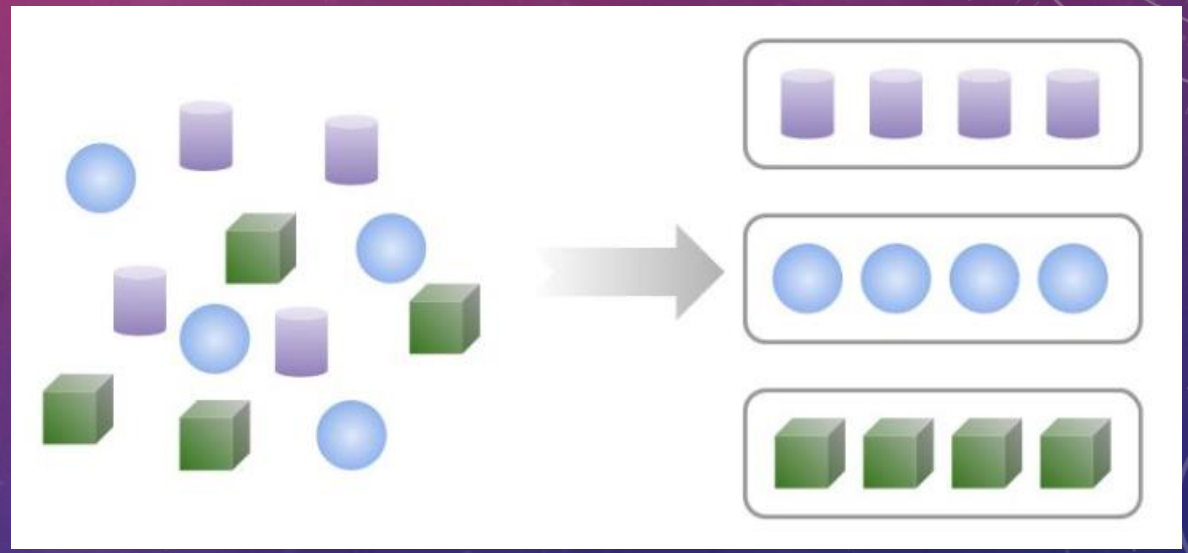
Decision tree - it has high rates of accuracy for both valid transactions and fraudulent transactions.

Logistic classifier - overall, it is accurate in >90% of cases.

Random forest - it has equal accuracy for both fraudulent and valid transactions.

UNSUPERVISED MACHINE LEARNING - CLUSTERING

- Clustering is a set of machine learning algorithms that divide data into categories, called clusters. Clustering can help us see patterns in messy datasets. Machine Learning Scientists use clustering to divide customers into segments, images into categories, or behaviors into typical and anomalous.

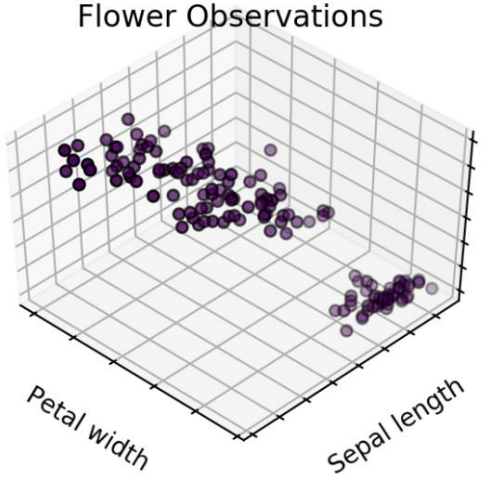


Supervised Machine Learning

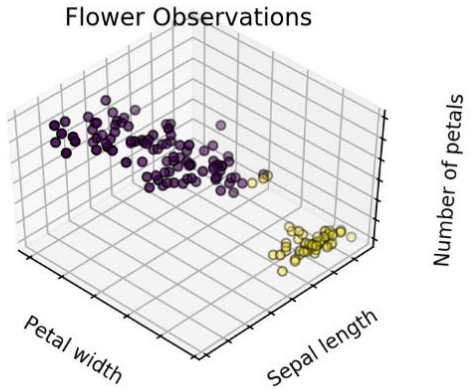
Unsupervised Machine Learning

Defining features

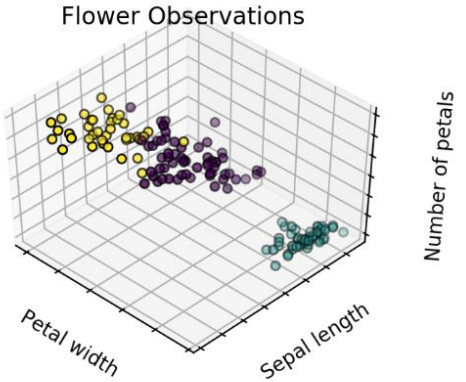
- Flower colors
- Petal length and width
- Sepal length and width
- Number of petals



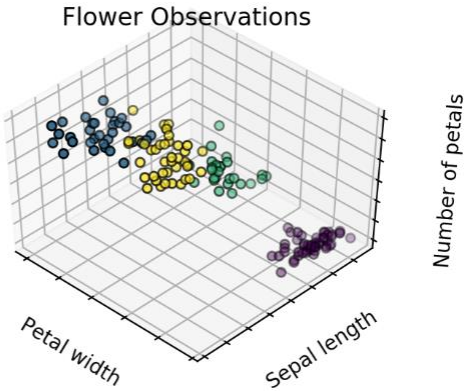
Two clusters:



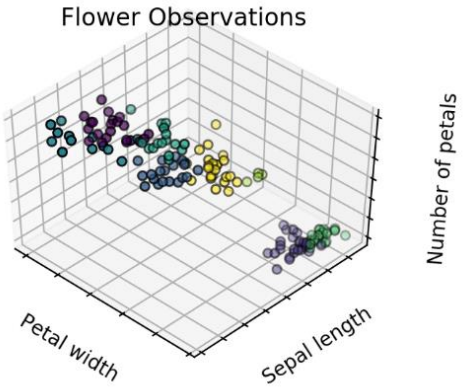
Three clusters:



Four clusters:



Eight clusters:



SUPERVISED OR UNSUPERVISED?

- To know whether a new clothing style will be successful based on previous season’s trends.
- To create groupings of clothing items that have similar features.
- To see if a customer will purchase an item based on what previous customers purchased.
- To divide customers into different categories based on their shopping habits.
- To know a group of problems/issues based on VMS satisfaction survey.

Assume you manage operations for a meal delivery service in Houston, Texas. You have latitude and longitude data for meals requests in the past month.

You'd like to create several new "delivery hubs" where demand has been greatest. In order to do this, you've asked a machine learning scientist to use a clustering algorithm to divide the orders into groups based on their location.

The machine learning scientist isn't sure how many clusters to make. Take a look at the clustering algorithm's results using different number of clusters.

What is the most suitable number of clusters?