Fall 2024

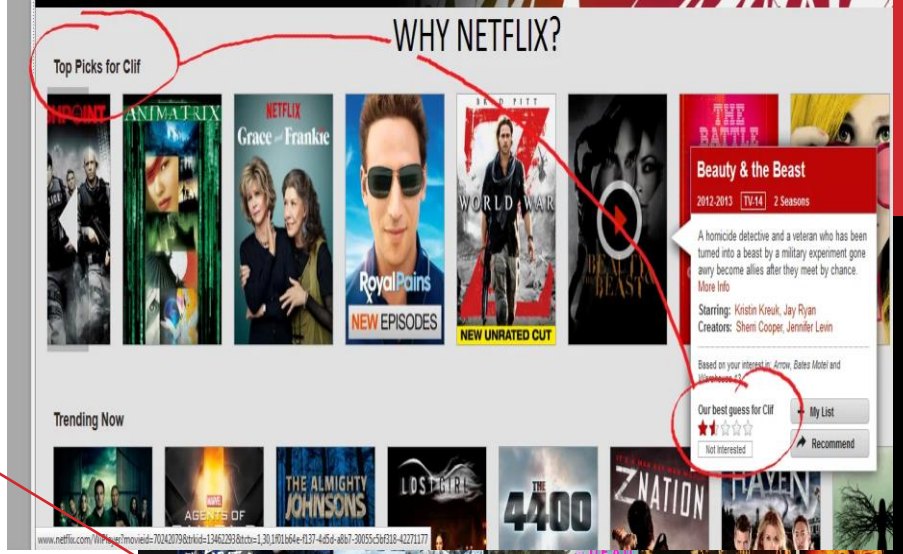
DATA, MSML, BIOI 602

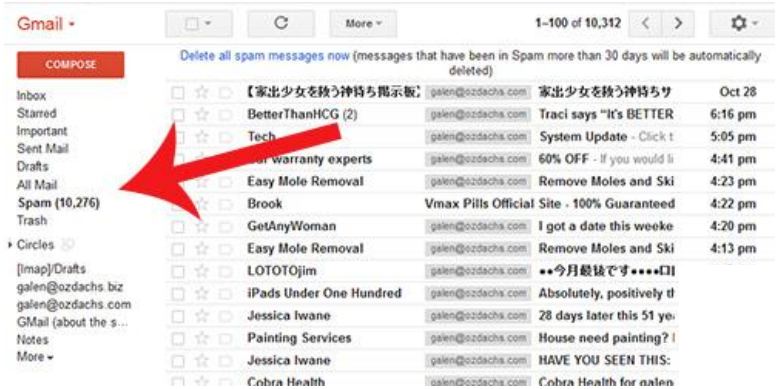
Principles of Data Science

LECTURE: 01

[Slides from Fardina Fathmiul Alam]

Let’s start by understanding what is data science?





Need availability of DATA!!!

DATA SCIENCE

Data science is all about using data to solve problems.

- **Decision making**
  - Which email is spam and which is not?
- **Product recommendation**
  - Which movie to watch?
- **Predicting the outcome**
  - Who will be the next President of the USA?
  - Many more .....

In simple terms: Using data to draw an inference or predict an outcome. Such information can help us to make better decisions.

WHAT IS DATA SCIENCE?

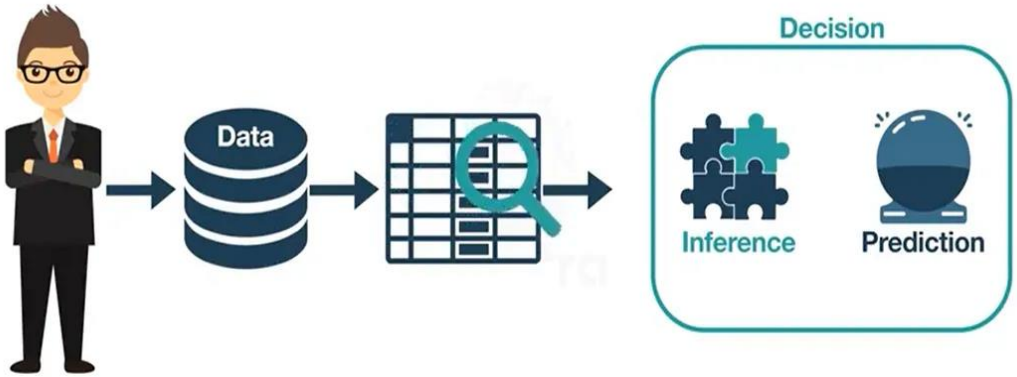
Data science is an interdisciplinary field focused on discovering patterns and describing relationships using data.

Data science is the application of computational and statistical techniques to address or gain [managerial or scientific] insight into some problem in the real world.

Zico Kolter
Machine Learning Prof, CMU
Chief Scientist of AI Research, Bosch

WHAT IS THIS COURSE ABOUT?

This semester we will learn to take raw data and turn it into insights about the world or predictions about the future.



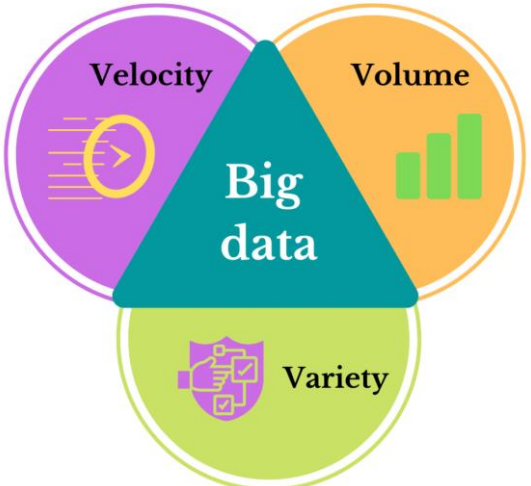
BIG DATA AND DATA SCIENCE

The rise of data science over the last 20 years is partially a result of big data.

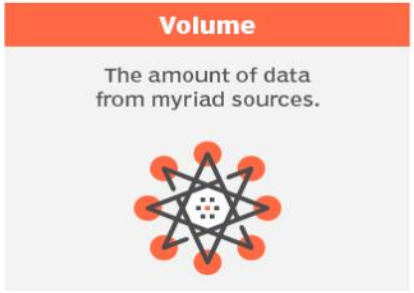
Big data describes datasets with large volume, created and updated with high velocity, that have variety in structure and format. As more companies and organizations collect and use big data, the demand for people with data science skills grows.

The 3 V's of big data

Big data is a collection of data from various sources, often characterized by what's become known as the 3 V's: *volume*, *velocity* and *variety*.



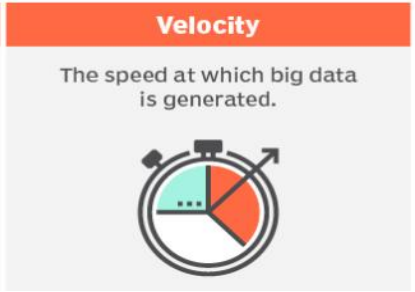
BIG DATA AND DATA SCIENCE



3 V's:

- **Volume:** vast amount of data generated or collected. the volume of global data has increased exponentially.
    - Data scientists use specialized tools and software to work with big datasets, such as: Apache Spark, Hadoop, Cloud-based storage, like Amazon Web Services, Google Cloud, or Microsoft Azure for storing and analyzing large amounts of data.

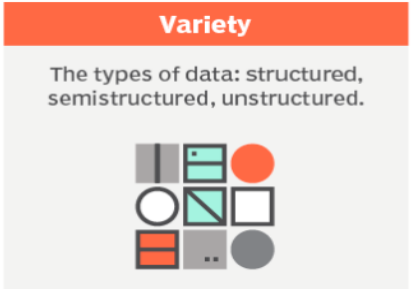
BIG DATA AND DATA SCIENCE



3 V's:

- **Velocity:** Represents the speed at which data is generated, processed, and analyzed, often in real-time or near real-time.
  - Think about everyday **900** million photos are uploaded on Facebook, **500** million tweets are posted on Twitter, **0.4** million hours of video are uploaded on Youtube and **3.5** billion searches are performed in Google.
  - This is like a nuclear data explosion

BIG DATA AND DATA SCIENCE



3 V's:

- **Variety:** Big data comes in a variety of forms: tables, spreadsheets, images, videos, sound, text, etc.
  - Data scientists deal with variety in data by *using techniques from* statistics, computer science, machine learning, and artificial intelligence.
  - The ability to deal with data variety helps set data science apart from other computational and analytical fields.

EXAMPLE: ANALYZE PATIENT'S MEDICAL RECORD, PREDICT DISEASE OUTBREAK

Applies Techniques and Models: Analyze patient data to identify risk factors, predict disease progression, and recommend personalized treatments.

Raw Data: Medical records, patient demographics, lab results.

Insights and Predictions: Identifies patterns in patient data to

- predict disease outbreaks,
- optimize treatment plans,
- provide insights for medical research.

SOME MORE EXAMPLES

- Given the results of a drug trial, determine if the drug is effective.
- Given a dataset of movies with ratings, predict what movies someone will like.
- Given a set of labeled images, identify what is in a given picture.
- Given a dataset describing people who have and haven't paid off their loans, predict if a new person will repay a loan.

Visual Object Categorization

We are given categories for these images:

What are these?

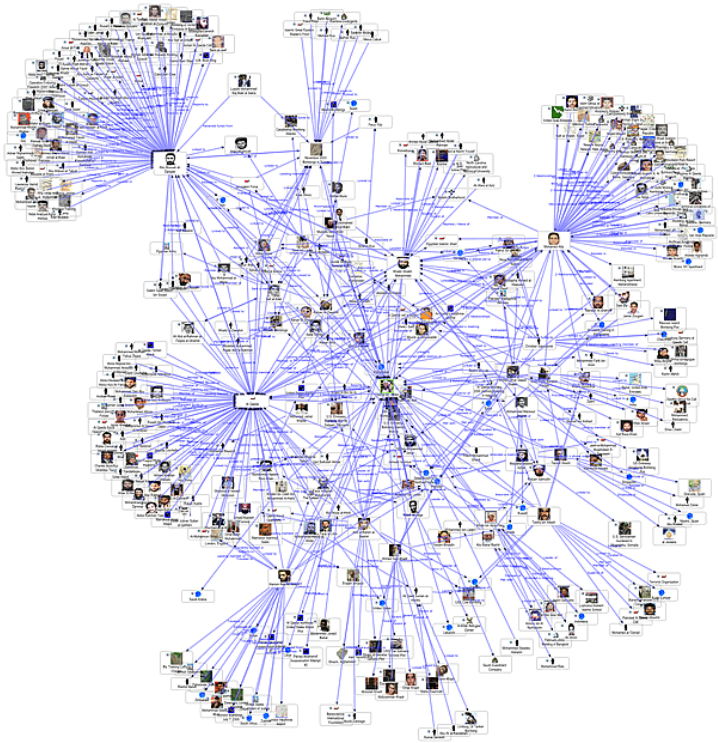
A classification problem: predict category y based on image x.
Little chance to “hand-craft” a solution, without learning.
Applications: robotics, HCI, web search (a real image Google...)

Object Detection

Example training images
for each orientation
(Prof. H. Schneiderman)

TV - Person of Interest

Social Network and Sentiment Analysis



"What people think?"

What others think has always been an important piece of information

"Which car should I buy?"

"Which schools should I apply to?"

"Which Professor to work for?"

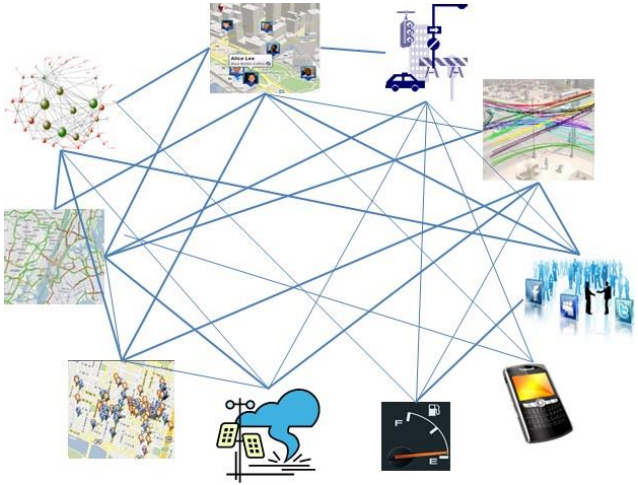
"Whom should I vote for?"

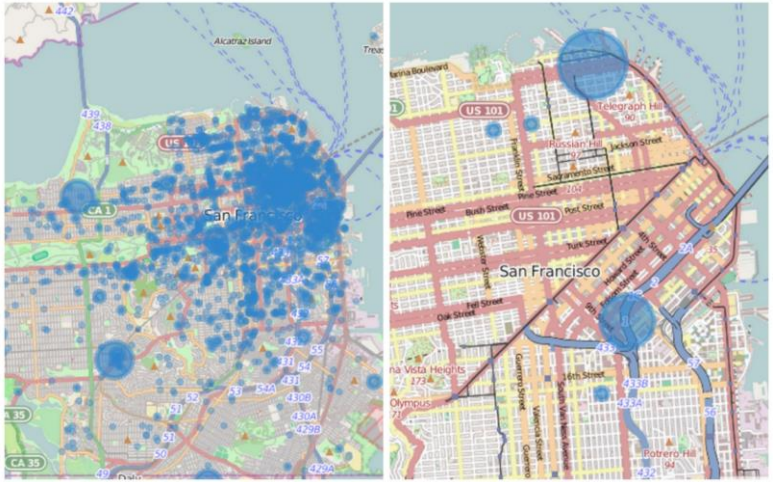
Data Science in Urban Computing

Uber, Lyft, DiDi, etc.

Cyber physical systems

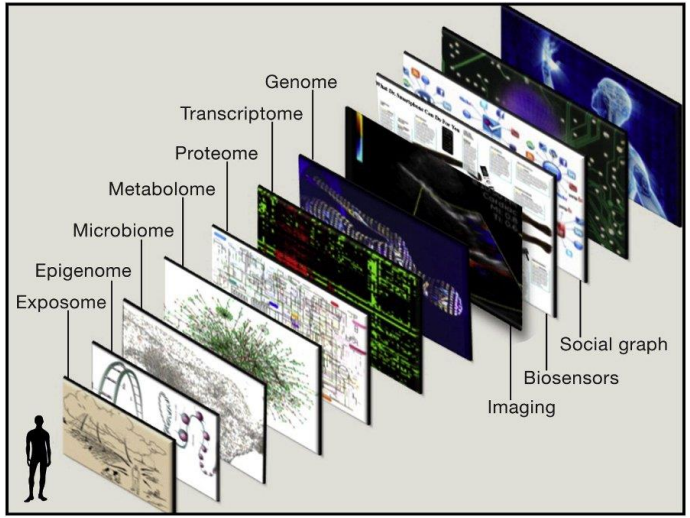
Internet of Things

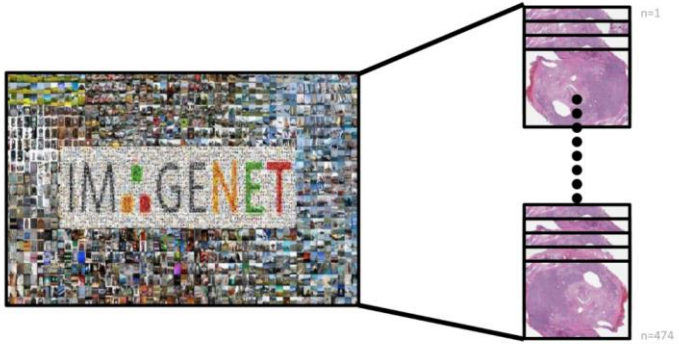




Financial Prediction

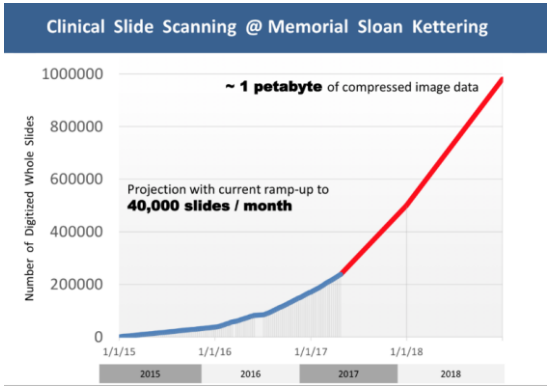
Biomedical Data Science





All of ImageNet
482 x 415 * 14,197,122
= 2.8 trillion pixels

474 Whole Slides
100,000 x 60,000 *474
= 2.8 trillion pixels

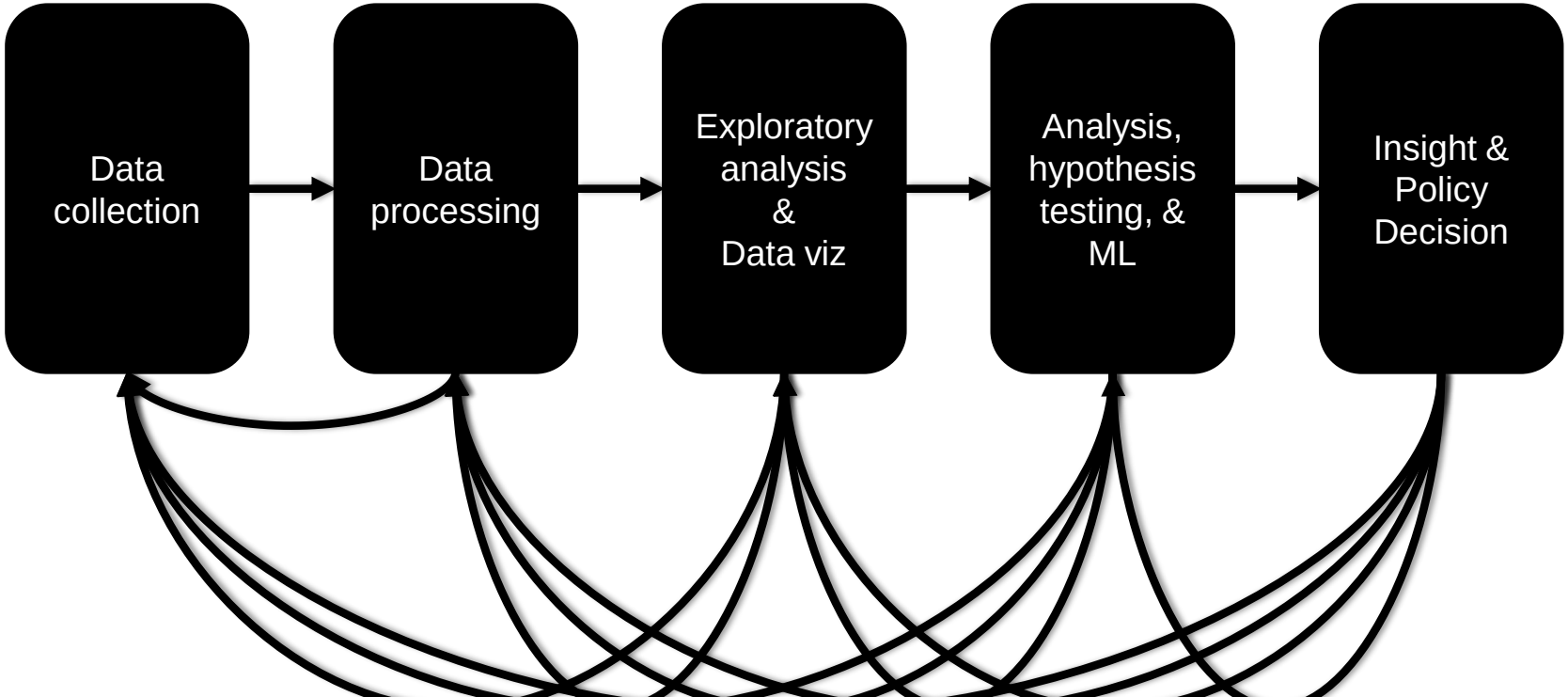


Recommendation Systems

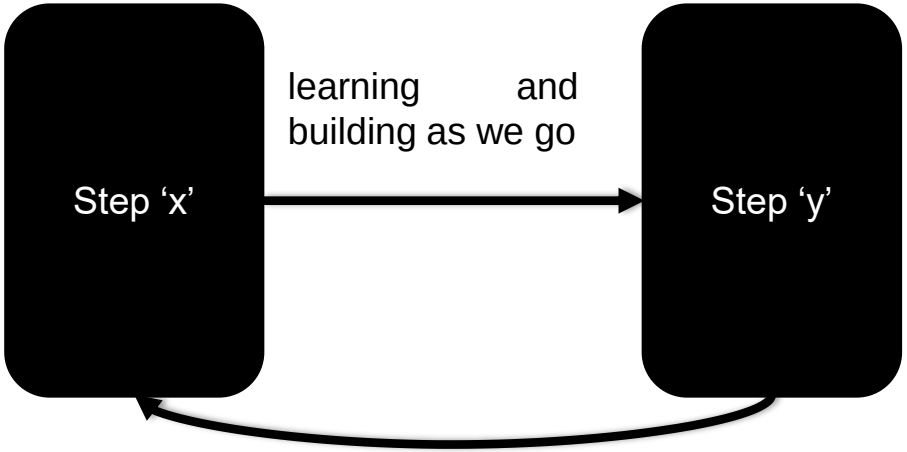
Challenge: to improve the accuracy of movie preference predictions

Netflix $1m Prize. Competition started Oct 2, 2006 and awarded 2009.

THE DATA LIFECYCLE



REMEMBER: DATA SCIENCE IS NOT A STRICTLY ONE-WAY LINEAR PROCESS; IT'S DYNAMIC, ITERATIVE, AND ADAPTIVE



If need to revisit previous steps due to new insights or challenges that arise during later stage. **Example:** Later realize you need to collect more data. This allows for *constant refinement* and *improvement* as new information emerges and insights evolve.

BEFORE THAT: DEFINE PROBLEM STATEMENT

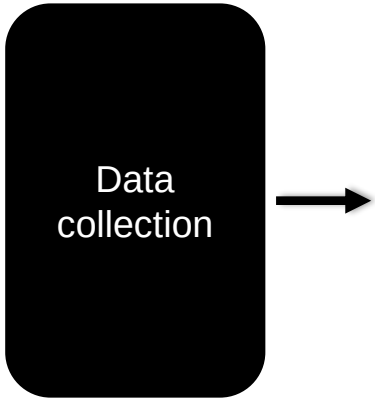
What problem are you going to solve?

- Why do we need a well-defined problem statement?

Example: “I want to increase the profit” - is it a well defined problem statement?

How much to increase the average profit/ revenue such as 20% or 30% ?
What is the average time frame to increase the revenue?

THE DATA LIFECYCLE



1. COLLECT DATA

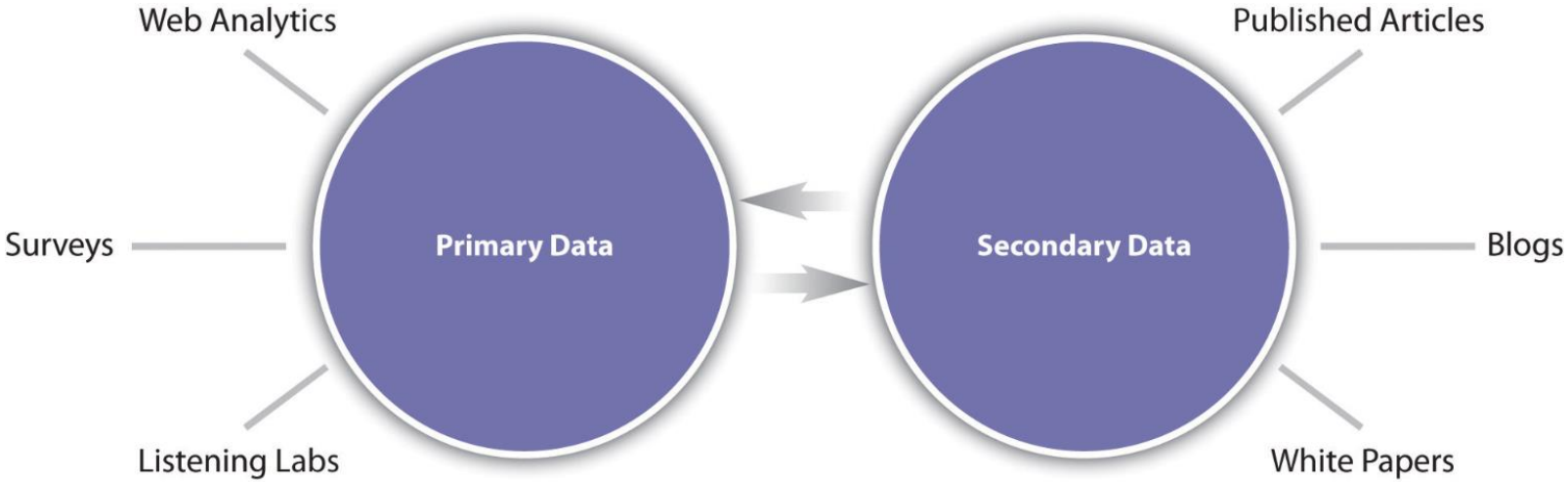
Data collection is a systematic approach to gather relevant information from a variety of sources.

- Gathered from external sources
- Gathered from existing company databases
- Gathered by tools created by you

DATA COLLECTIONS METHODS:

2 types of data collections methods:

1. Primary Data Collection
2. Secondary Data Collections



DATA COLLECTIONS METHODS: PRIMARY

Situation: Some unique problem and no related research is done on the subject.
Solution: Collect new data → Primary data collection.

Example: Average time that employees spend during lunch break across companies.

- ***Problem*** → No public data available of these.
- ***Solution:*** Collect the data through various methods.
- ***Different Methods:*** Surveys, Interviews of employees and by Monitoring the time spent by employees in cafeteria.
- **This methods are time consuming**

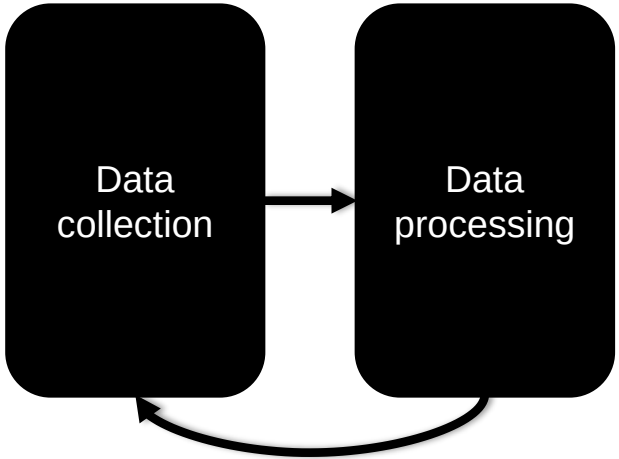
DATA COLLECTIONS METHODS: SECONDARY

Situation: Some problem and the data is readily available or collected by someone else.

Solution: Use the data → Secondary data collection.

- ***Different Methods:*** The internet, news articles, government census, magazines and so on.
- **This methods are less time consuming than the primary method.**

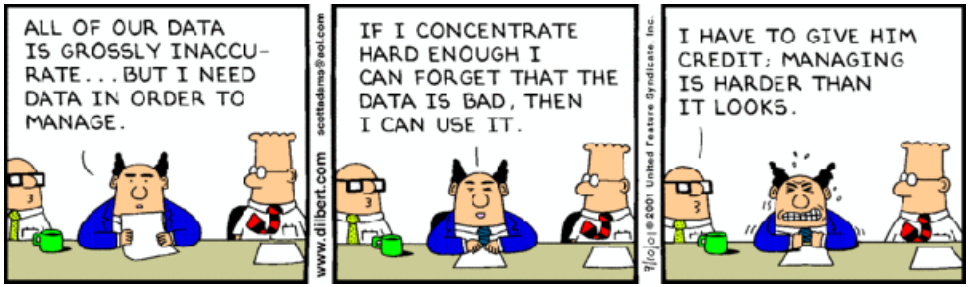
THE DATA LIFECYCLE

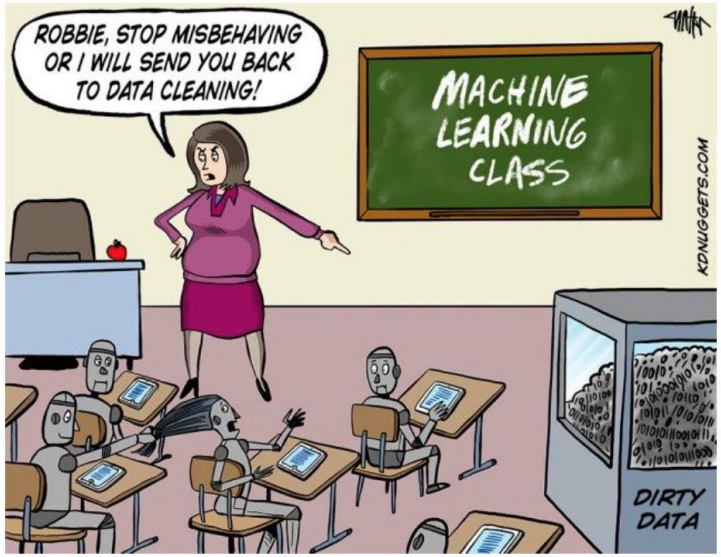


2. DATA PROCESSING

Clean or scrub data to ensure the data quality

- Important: Do sanity check on data.
- **Why ?** Bad quality may lead to unexpected results or misleading information.
  - Deal with duplicates
  - Formatting
  - Weird outliers
  - Mistakes



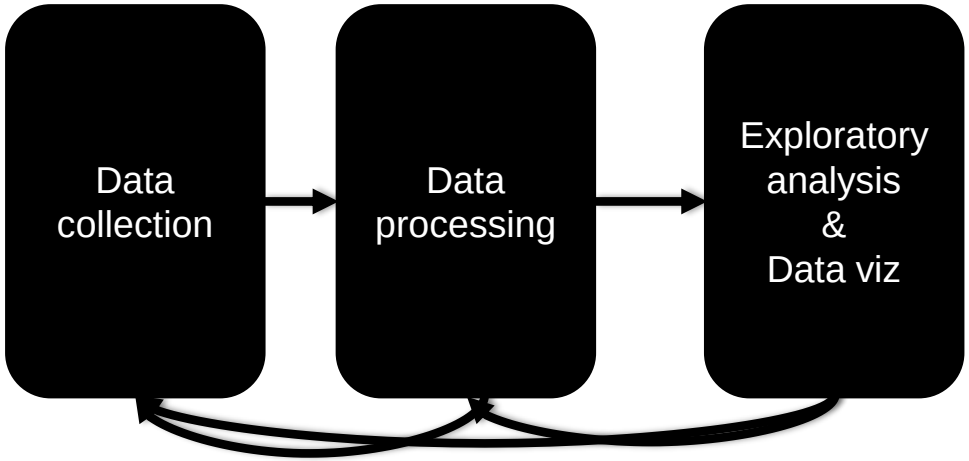


2. DATA PROCESSING: EXAMPLE

You Collect data about students' test scores.

- **But →** Some left SCORE scores blank, and others wrote "N/A" (Missing Values).
- **Process Data**
  - Replace "N/A" entries with a neutral value like 0 and
  - Fill in the missing scores with the correct value

THE DATA LIFECYCLE



3. EXPLORE DATA: FIGURE OUT WHAT YOU HAVE

You'll be sitting on like a terrabyte of raw data

- What is there?
- Are there any interesting correlations?
- Do you have everything you need?

3. EXPLORE DATA CONT.

- **Extract useful insights** from the data, understanding patterns, and setting the stage for effective model building and decision-making
- Important to **analyse the data and build familiarity** with the data
- **Skipping this step may lead to inaccurate models** as well as insignificant variables in your models.

"It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it." – *As quoted by John Tukey, developer of Exploratory Data Analysis*

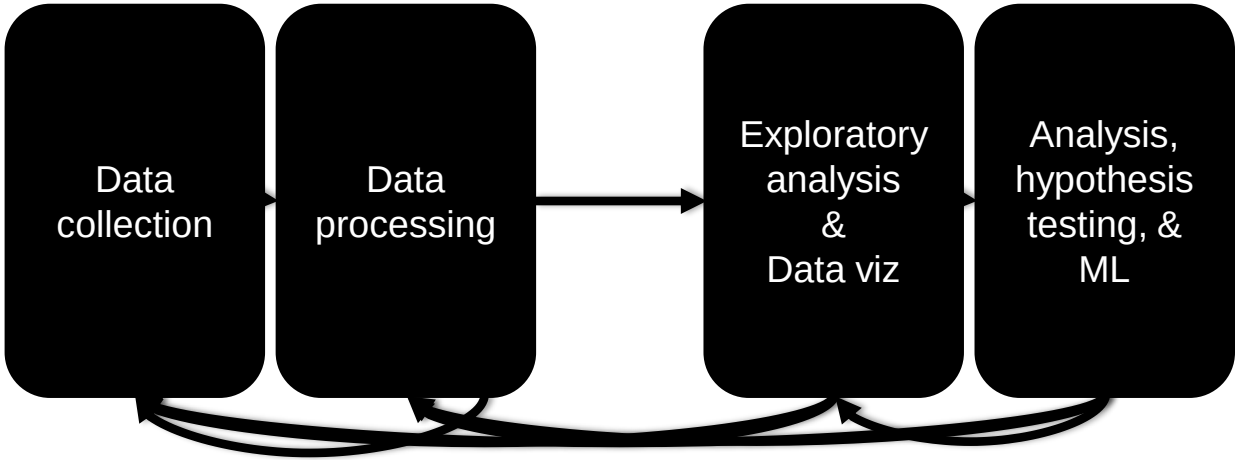
3. EXPLORE DATA: EXAMPLE

You want to understand more about the scores and how they relate to other factors like study hours.

- **Exploratory Analysis:** Find Connection between study hours and scores by calculating the average score, find the range of scores, and notice that some students scored exceptionally well.

- **Data Visualization:** Visually show the relationship between study hours and average scores for different groups of students in order to reveal that students who study more tend to have higher average scores.

THE DATA LIFECYCLE

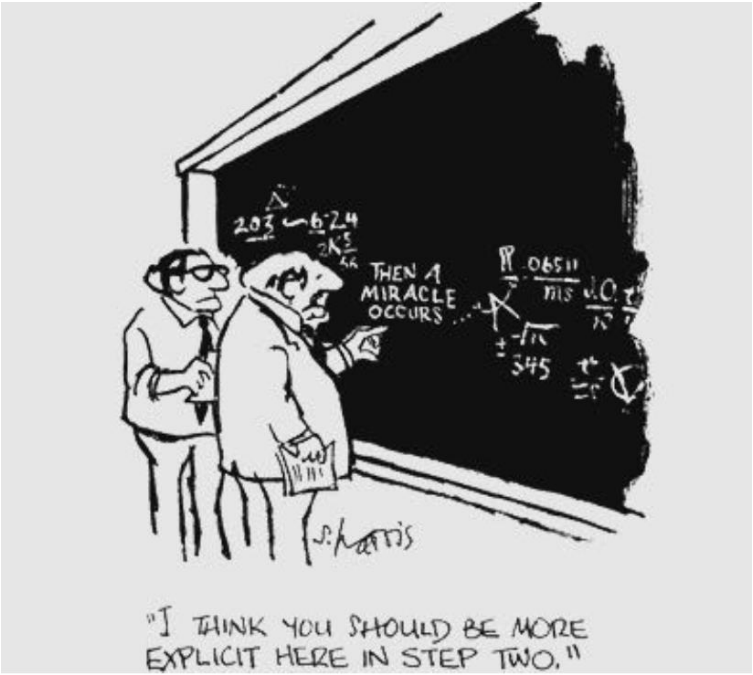


4. BUILD A MODEL

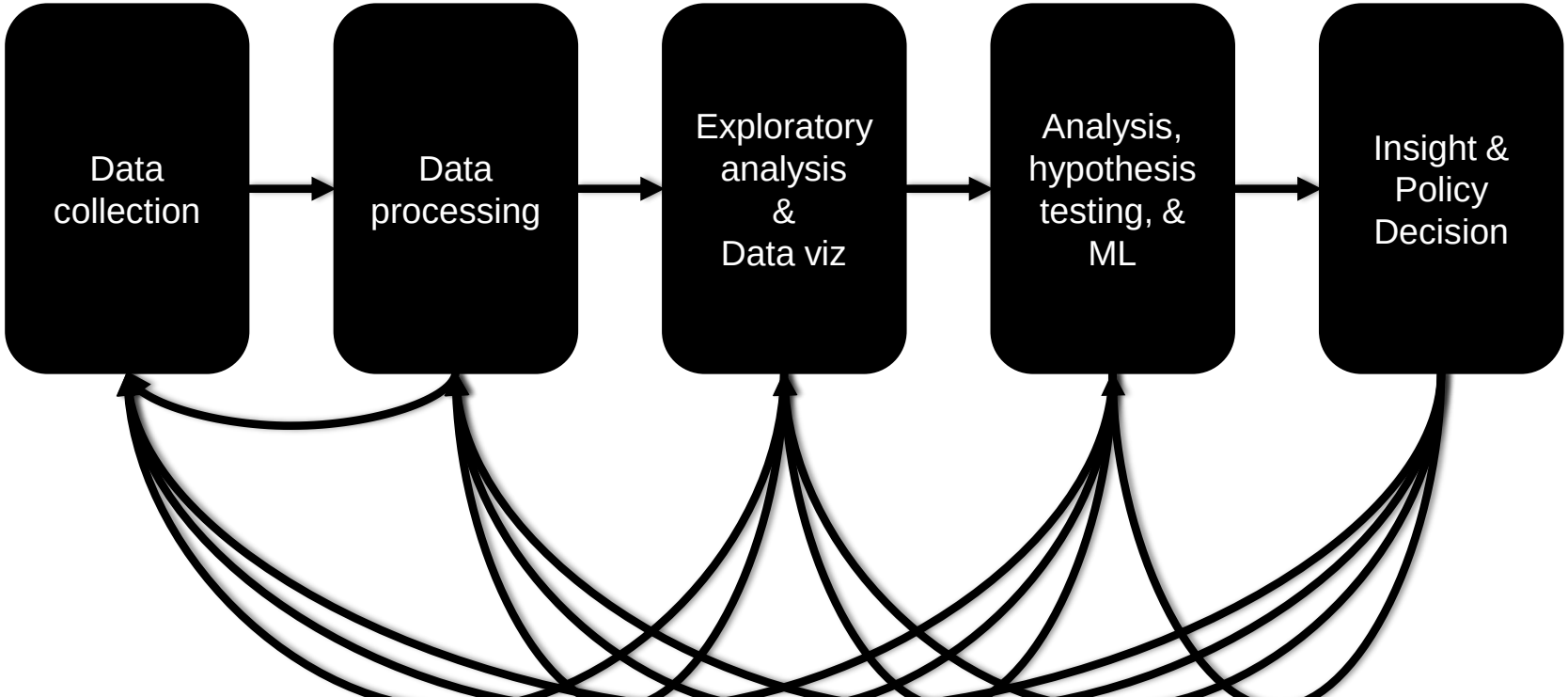
Steps to the solution.

Example: build a machine learning model that **predicts a student's test score based on their study hours**.

- Train the model → using the historical data,
- Once it's trained, **you can input a student's study hours to get a predicted test score**.



THE DATA LIFECYCLE

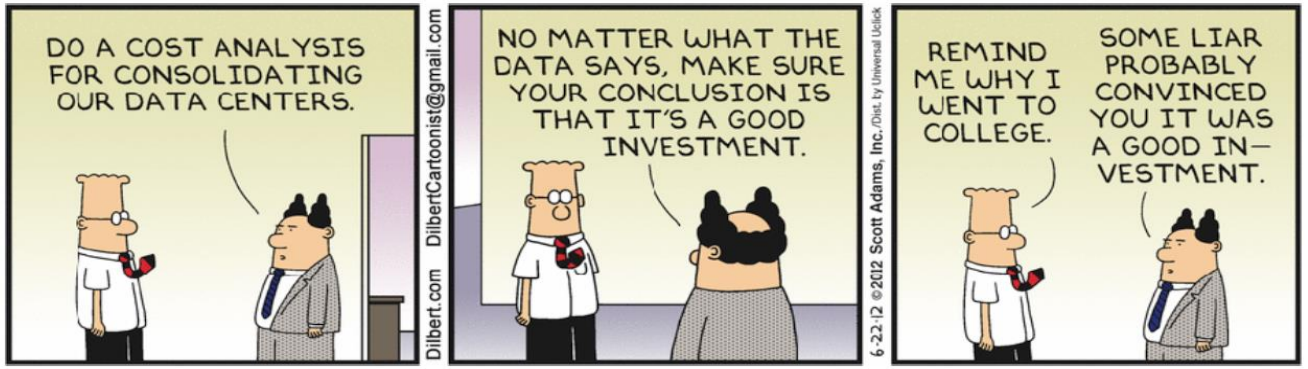


5. INTERPRETATION

Translate these findings into actionable insights.

- Deriving Insights and make policy decisions if needed
- Present the results from your analysis to the stakeholders.
- **Convince People:** Explain the specific conclusion and critical findings, probably in understandable manner.

The last step is getting a bunch of non-technical people to understand what your magical model is doing and why it's right and they should listen to you.



5. INTERPRETATION

Example: Let's say your analysis confirmed that there's a strong positive correlation between study hours and test scores. This insight is crucial for both students and educators to understand the importance of consistent study habits.

- Policy Decisions: You can lead to policy decisions that encourage regular study time or improved teaching methods.
- Convincing people to do what you want:
  - Often, you must share findings with non-technical groups like marketing or executives. Your goal is clear communication, allowing stakeholders to create actionable plans based on the results.

CAREER IN DATA SCIENCE

Title	Description
Data engineers	Data engineers specialize in data gathering and storage. Data engineers extract, transform, and load datasets for later analysis.
Data scientists	Data scientists gather data, transform data, and use models and algorithms to extract meaningful insights from datasets.
Data analysts	Data analysts work with industry experts to analyze datasets and create visualizations. Data analysts use some data science models, but tend to use data visualization and summary more than modeling.
Business intelligence analysts	Business intelligence analysts specialize in data related to financial and market transactions. Data analysts and business intelligence analysts are similar roles, but the term business intelligence is more common in business and finance.
Machine learning engineers	Machine learning engineers specialize in machine learning models instead of statistical models. Machine learning engineers often focus on the implementation and development of a model rather than selection and interpretation.

CAREER IN DATA SCIENCE

Data Analyst	Data Scientist	Machine Learning Engineer
Focus: ● Primarily deals with analyzing and interpreting data to provide actionable insights, answering specific business questions ● Emphasizes data visualization, statistical analysis, and creating reports.	**Focus:** ● Involves both analysis and interpretation of complex data. ● Applies statistical models and machine learning algorithms to extract insights. ● Emphasizes pattern recognition and predictive **Probable Skills:** modeling.	**Focus:** ● Specializes in deploying and operationalizing machine learning models. ● Focuses on the technical implementation, scaling, and optimization of models for production.
Probable Skills: Proficient in statistical analysis, data cleaning, and visualization tools. Strong Excel skills and familiarity with databases. May have knowledge of basic programming for data manipulation.	**Probable Skills:** Strong statistical and mathematical background. Proficient in programming languages (e.g., Python, R). Expertise in machine learning and data modeling.	**Probable Skills:** Strong programming and software engineering skills. Expertise in machine learning frameworks and libraries. Knowledge of model deployment and optimization.

“The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill in the next decades …”

Hal Varian
Chief Economist at Google

THIS COURSE

You'll learn to take data:

- Process it
- Visualize it
- Understand it
- Communicate it
- Extract value from it

All course information will be posted at ELMS.

COURSE INFORMATION

Instructor: Dr. Heng Huang

- **Office location:** IRB 3238
- **Office hours:** Tuesdays 4-5:50pm (Other time is ok, please make appointment)
- **Email:** heng@umd.edu
  - Best way to contact: Email, office hours
    - Do not send me an email through ELMs
  - If it is urgent, add [URGENT] to the subject line
  - Email instructor with **[DATA602]** in the email subject line.
- Slides will be posted at ELMS
  - Memory tool, not a substitute for lectures
  - Take your own notes
  - Do the reading
- Assignments and Project
  - Will be posted on ELMS

PREREQUISITE KNOWLEDGE

Aimed at entry-level data science students – but likely accessible to others with programming experience and mathematical maturity.

We do not assume:

- Experience with Python (not expert but basic use), NumPy, pandas, scikit-learn, PyTorch, matplotlib, etc …
- Deep statistics or any ML knowledge
- Database or distributed systems knowledge

We do assume:

- You want to be here!

How To Succeed in this Course

This class is organized with the idea that students will:

- **Attend Class**
- **Participate / Take notes**
- **Do the readings**
- **Be amenable to very open ended assignments and questions**
- **Check ELMS daily**

"Go to class! College isn't that hard if you actually, you know, show up!"
– Andrew, Financial Analyst, University of Wisconsin Madison, Class of 1993, from The Advice I'd Give My College Freshman Self, PBS News Hour

Students are also expected to check the course at ELMS in regular basis for any updates.

Course Setup

Components	Percentages
Programming Homework	**30%**
Weekly Quizzes (at least 10-12, one lowest would be dropped automatically)	**15%**
1 Final Project	**30%**
1 Midterm Exam	**20%**
Participation & Engagement	**5%**

Course Setup

QUIZZES:

- **There will be multiple very short quizzes weekly (duration: 5-15 minutes).**
  - Topics are based on the weekly topics covered.
  - The quizzes will be available starting from Tuesday any time after class (an announcement will be given).
  - You will have until Monday 11:59 pm to complete them.
  - You will have just one chance to submit each quiz; you cannot take any quiz twice.

Late Policy

- **Turn in your homework by the due date**
  - There will be a 10% penalty for late submissions of **homework** within 24 hours after the deadline.
  - After this 24-hour period, **no submissions will be accepted (you will receive 0 automatically)**.
  - In ELMS (as instructed in the assignment or project), you can submit multiple times, and only the last submission will be graded.
  - This policy applies to both homework and projects, **except for the final project**.
    - Late submission is not allowed for final project submissions.

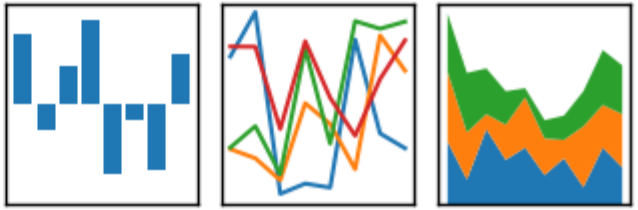
Late Policy

- **Extensions:**
  - Extensions will be handled on a case-by-case basis
  - Things that will probably get you an extension:
    - Documented illness or mental health issues
    - Deaths in the family
  - Things that are less convincing:
    - Your partner is in town
    - You forgot you were taking this course
  - Ask anyway! You never know.

SOME TECHNOLOGIES WE MIGHT USE (MOSTLY)

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



IMPORTANT WALLS OF TEXT

ANTI-HARASSMENT

Common Sense!

The open exchange of ideas and the freedom of thought and expression are central to our aims and goals. These require an environment that recognizes the inherent worth of every person and group, that fosters dignity, understanding, and mutual respect, and that embraces diversity. For these reasons, we are dedicated to providing a harassment-free experience for participants in (and out) of this class.

Harassment is unwelcome or hostile behavior, including speech that intimidates, creates discomfort, or interferes with a person's participation or opportunity for participation, in a conference, event or program.

(Adapted from ACM SIGCOMM's policies)

ACADEMIC INTEGRITY

Common Sense!

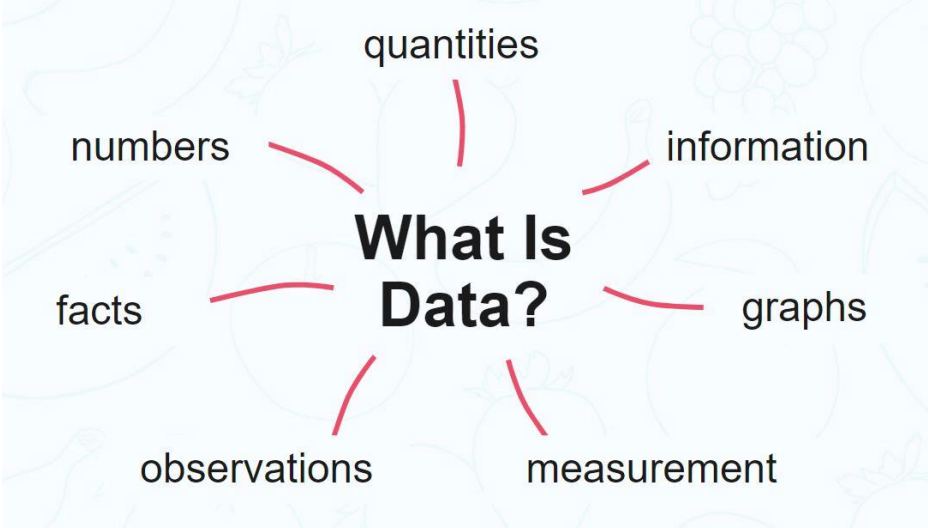
Any assignment or exam that is handed in must be your own work (unless otherwise stated). However, talking with one another to understand the material better is strongly encouraged. Recognizing the distinction between cheating and cooperation is very important. If you copy someone else's solution, you are cheating. If you let someone else copy your solution, you are cheating (this includes *posting solutions online in a public place)*. If someone dictates a solution to you, you are cheating.

Everything you hand in must be in your own words, and based on your own understanding of the solution. If someone helps you understand the problem during a high-level discussion, you are not cheating. We strongly encourage students to help one another understand the material presented in class, in the book, and general issues relevant to the assignments. When taking an exam, you must work independently. Any collaboration during an exam will be considered cheating. Any student who is caught cheating will be given an F in the course and referred to the University Office of Student Conduct. Please don't take that chance – if you're having trouble understanding the material, please let me know and I will be more than happy to help.

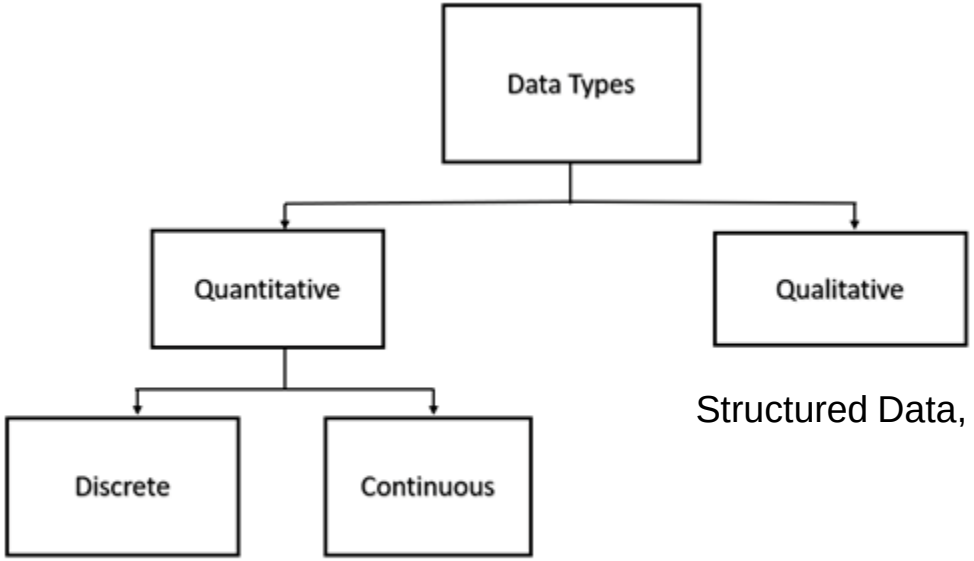
What is Data?

Data

Data is raw information, facts, or statistics that can be in various forms such as numbers, text, images, or more.



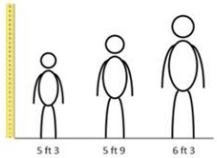
Broad Category of Data



Structured Data, Unstructured Data

No. of Laptops

No. of Cars



Height

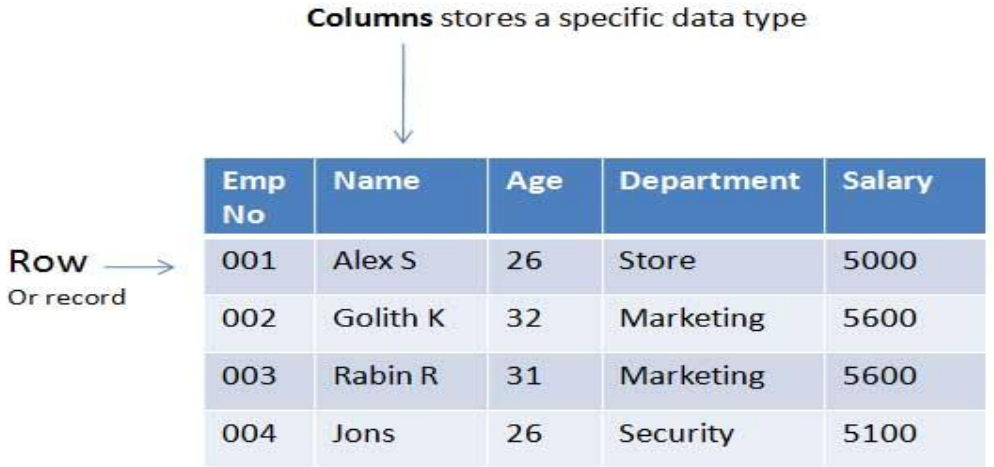
Time

Types of Data

Tabular Data–(Things that are in tables): Structured data organized into rows and columns, often resembling a spreadsheet or database table.

Example:

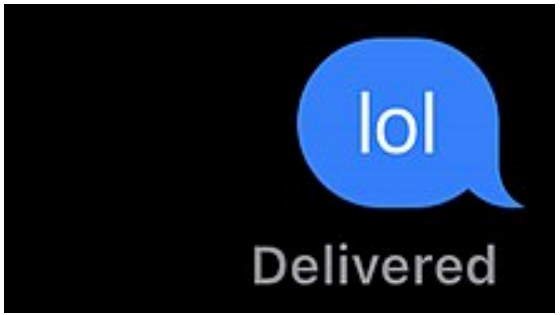
- Demographic info
- Grades
- Many more….



Emp No	Name	Age	Department	Salary
001	Alex S	26	Store	5000
002	Golith K	32	Marketing	5600
003	Rabin R	31	Marketing	5600
004	Jons	26	Security	5100

Types of Data cont.

Text: human-readable text



Examples:

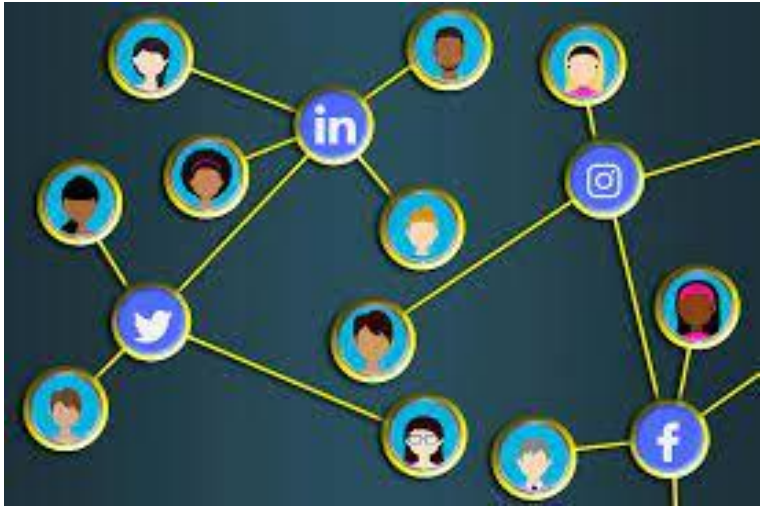
- **Reviews, Books, Articles, Emails**
- **Translation**
- **ChatGPT → generate human-like text responses in a conversational manner.**
- **Social Media Post**

Types of Data cont.

Graph: Represents relationships between entities using nodes (vertices) and edges.

Examples:

- **Social connections**
- **Websites**
- **Network traffic**
- **Roads**



Types of Data cont.

Unstructured Data: lacks a predefined structure or format, challenging to analyze and process.

- **Videos**
  - Tik Tok
- **Images**
  - James Webb
  - Faces
  - Handwriting
  - Road signs
- **Audio**
  - Alexa
  - Real-time translation
  - Music
- **Biometrics**
  - Fingerprints
  - Facial recognition
- **Haptics**
  - Phone vibrates to notify you of a message,

Some More Examples of Different Types of Data

- **Tabular Data**
  - Text document of the heights of everyone in this class in inches
  - IRS Data for taxpayers
  - Netflix show data
- **Graph Data**
  - Social networks
  - Course prereqs
  - Highway
- **Geo Data**
  - Flight paths
  - Weather patterns
  - All phones on verizon

Some More Examples cont.

- **Raw Data**
  - Image
  - Video
  - Audio
  - Telemetry
- **Hierarchies (Graphy)**
  - Taxonomy for something
  - Family tree
  - File directory

- **Text**
  - All tweets
  - Chat logs
  - Search history
  - Shakespeare
- **Time Series**
  - People in store
  - Stock prices

Data Formats

- **CSV / TSV**
- **Image**
  - .jpg
  - .png
- **Audio**
  - .wav
  - .mpg
- **SQL Database**
  - mySQL
  - Postgres
  - etc…

- **No SQL Database**
  - Bigtable
  - Accumulo

- **JSON**
- **XML / HTML**

Data Format: TSV / CSV

- **CSV (Comma-Separated Values)** **A plain-text format where data values are separated by commas; commonly used for tabular data.**
- **TSV (Tab-Separated Values)**
  - Both are common file formats used to store structured data in a simple tabular format, with rows and columns.
  - Easy to import and export across various data analysis tools and programming languages.

Any CSV reader worth anything can parse files with any delimiter, not just a comma (e.g., “TSV” for tab-separated values)

Delimiter: The separator character : the comm (,), the tab (`\t`), colon (`:`) and semi-colon (`;`) characters.

Tabular Data: Example

classic_rock_playlist.csv (39.93 kB)

Detail | Compact | Column

10 of 13 columns

Artist	Music	Album	Year	Genre	2022	2021
The Black Crowes	Remedy	The Southern Harmony and Musical Companion	1992	Southern Rock	500	
Asia	Only Time Will Tell	Asia	1982	Progressive Rock	499	
Collective Soul	Shine	Hints Allegations and Things Left Unsaid	1993	Alternative Rock	498	
Billy Idol	Sweet Sixteen	Whiplash Smile	1986	Rock	497	
Collective Soul	December	Collective Soul	1995	Alternative Rock	496	
Duran Duran	Save a prayer	Rio	1982	Synthpop	495	466
Men at Work	Down Under	Business as Usual	1981	New Wave	494	
Brian Setzer	Summertime Blues	La Bamba soundtrack	1987	Rock and Roll	493	
Simple Minds	Dont You Forget About Me	The Breakfast Club Original Motion Picture Soundtrack	1985	Pop Rock	492	

Representation of tabular data: how data might be structured in CSV Files

Artist,Music,Album,Year,Genre,2022,2021,2020,2019,2018,2017,2016,2015

The Black Crowes,Remedy,The Southern Harmony and Musical Companion,1992,Southern Rock,500,,324,290,132,64,36,

Asia,Only Time Will Tell,Asia,1982,Progressive Rock,499,,,,,,,

Collective Soul,Shine,Hints Allegations and Things Left Unsaid,1993,Alternative Rock,498,,,419,485,403,,

- CSV (Comma-Separated Values) files are a common way to store tabular data.
- Python's pandas library makes it easy to load (df = pd.**read_csv**('data.csv')) and manipulate CSV data.

CSV Files in Python

ID	Date	Topic	Reading	Slides	Lecturer
1	26-Jan	Introduction	—	"pdf, pptx"	Fardina
2	31-Jan	Scraping Dat	Anaconda's Test Drive.		Fardina
3	2-Feb	"Vectors, Ma	Introduction to pandas		Fardina
4	7-Feb	Jupyter notebook lab			"Denis, Anant, & Neil"
5	9-Feb	Best Practices for Data Science Projects			Fardina

Input file: schedule.csv

Not necessary write your own CSV or JSON parser

```
import csv
with open("schedule.csv", "r") as f:
    reader = csv.reader(f, delimiter= ",", quotechar='"')
    next(reader)
    for row in reader:
        print(row)
```

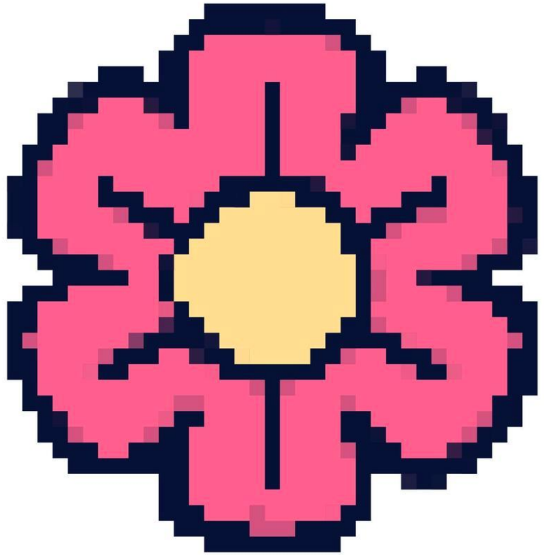
Output:

```
['1', '26-Jan', 'Introduction', '—', '"pdf, pptx"', 'Fardina']
['2', '31-Jan', 'Scraping Data with Python', "Anaconda's Test Drive.", '', 'Fardina']
['3', '2-Feb', '"Vectors, Matrices, and Dataframes"', 'Introduction to pandas', '', 'Fardina']
['4', '7-Feb', 'Jupyter notebook lab', '', '', '"Denis, Anant, & Neil"']
['5', '9-Feb', 'Best Practices for Data Science Projects', '', '', 'Fardina']
```

(We'll use pandas to do this much more easily and efficiently)

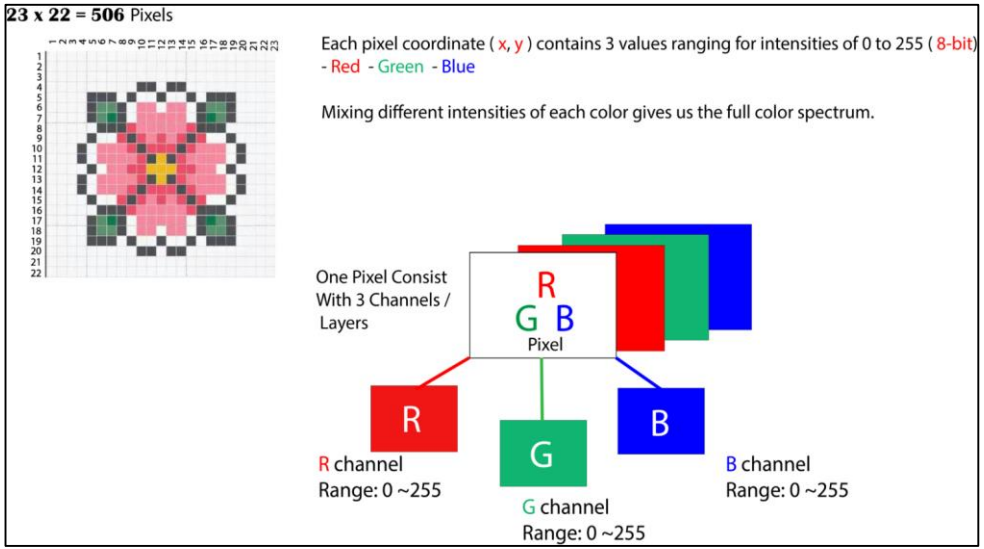
Data Format: Images

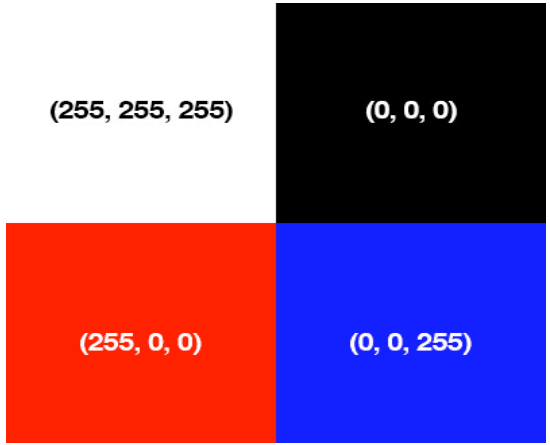
Image data encompasses **the visual content and properties of an image** (visual representation), including details such as **colors, shapes, patterns, and pixel values**.



Data Format: Images

- **Pixel Structure:** Images are organized as grids of pixels (tiny building block).
- **Pixel Data:** Each pixel holds color information.
- **Color Channels:** In color images, pixels have red, green, and blue (RGB) channels.
  - **Channel→ is like a separate layer of information about color.**
- **Channel Values:** Each channel ranges from 0 (no intensity) to 255 (maximum intensity).
- **Intensity Values:** Values between 0 and 255 create varying shades of the color





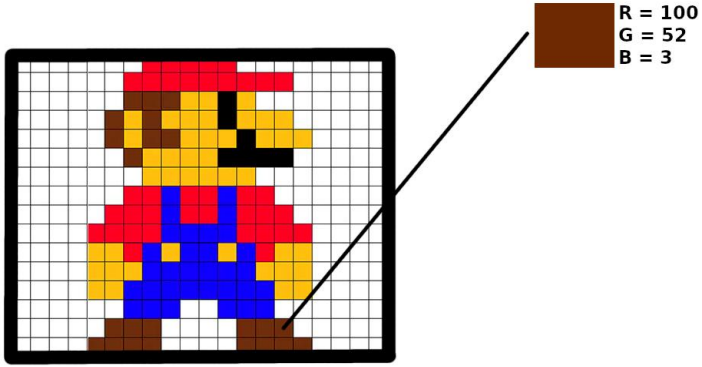
Data Format: Images

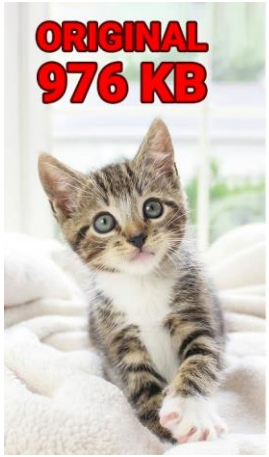
Images can be compressed to save storage space.

- Lossy vs. Lossless

- **Lossy Compression:** Sacrifices some data to reduce file size.
  - Suitable for photos where minor quality loss is acceptable (e.g., JPEG).
- **Lossless Compression:** Reduces file size without any loss of quality.
  - Ideal for preserving image quality in medical imaging, digital art, or graphics (e.g., PNG, GIF).

→Varying Underlying Resolutions (level of detail and clarity)





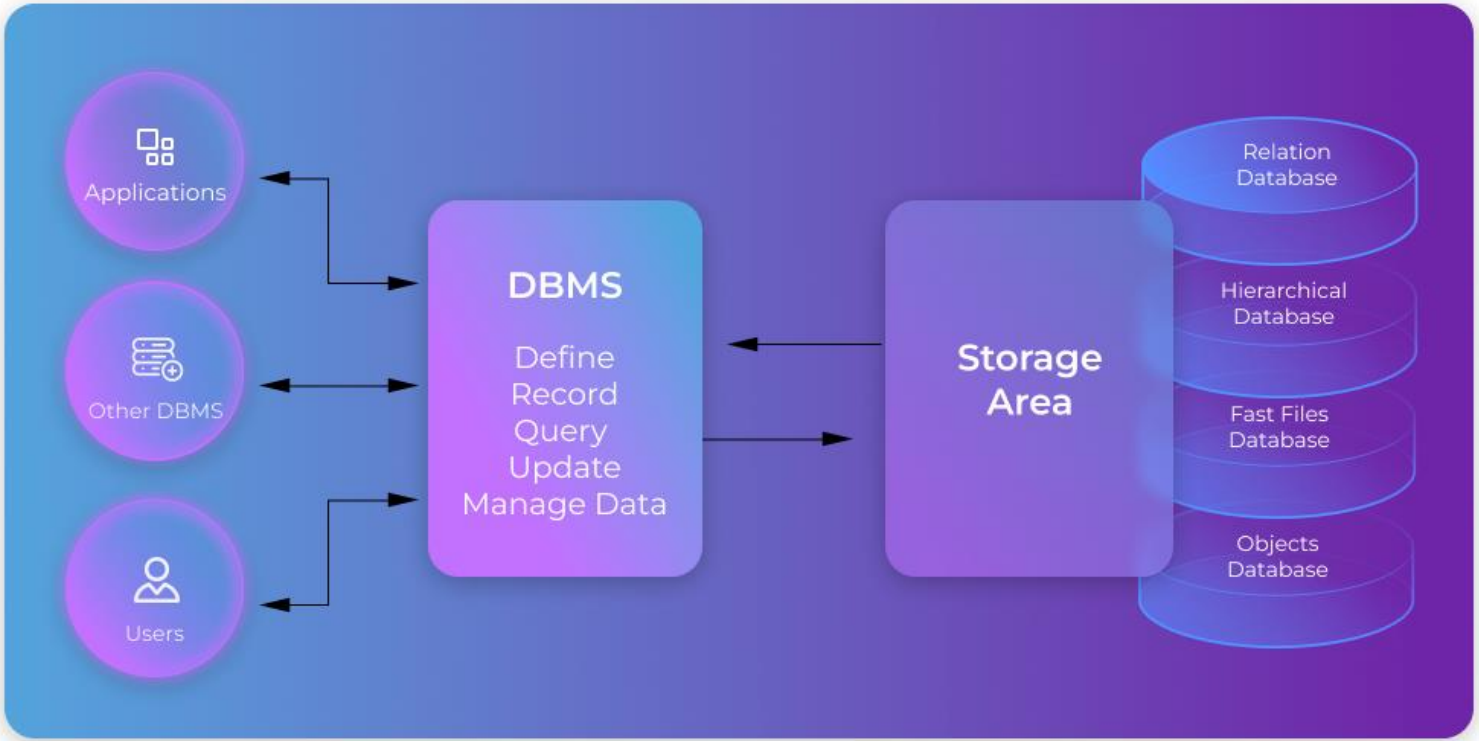
Databases

A database is an organized collection of structured information, or data, typically stored electronically in a computer system.

- Handle more complex data relationships, often organized in tables.
- It **efficiently manages and allows retrieval, updating, and manipulation** of information for various applications.

```
dvdrental=# select title, release_year, length, replacement_cost from film
dvdrental-#   where length > 120 and replacement_cost > 29.50
dvdrental-#   order by title desc;
          title          | release_year | length | replacement_cost
-------------------------+--------------+--------+------------------
 West Lion               |         2006 |    159 |            29.99
 Virgin Daisy            |         2006 |    179 |            29.99
 Uncut Suicides          |         2006 |    172 |            29.99
 Tracy Cider             |         2006 |    142 |            29.99
 Song Hedwig             |         2006 |    165 |            29.99
 Slacker Liaisons        |         2006 |    179 |            29.99
 Sassy Packer            |         2006 |    154 |            29.99
 River Outlaw            |         2006 |    149 |            29.99
 Right Cranes            |         2006 |    153 |            29.99
 Quest Mussolini         |         2006 |    177 |            29.99
 Poseidon Forever        |         2006 |    159 |            29.99
 Loathing Legally        |         2006 |    140 |            29.99
 Lawless Vision          |         2006 |    181 |            29.99
 Jingle Sagebrush        |         2006 |    124 |            29.99
 Jericho Mulan           |         2006 |    171 |            29.99
 Japanese Run            |         2006 |    135 |            29.99
 Gilmore Boiled          |         2006 |    163 |            29.99
 Floats Garden           |         2006 |    145 |            29.99
 Fantasia Park           |         2006 |    131 |            29.99
 Extraordinary Conquerer |         2006 |    122 |            29.99
 Everyone Craft          |         2006 |    163 |            29.99
 Dirty Ace               |         2006 |    147 |            29.99
 Clyde Theory            |         2006 |    139 |            29.99
 Clockwork Paradise      |         2006 |    143 |            29.99
 Ballroom Mockingbird    |         2006 |    173 |            29.99
(25 rows)
```

Databases


Applications
Other DBMS
Users
DBMS
Define
Record
Query
Update
Manage Data
Storage Area
Relation Database
Hierarchical Database
Fast Files Database
Objects Database

JSON

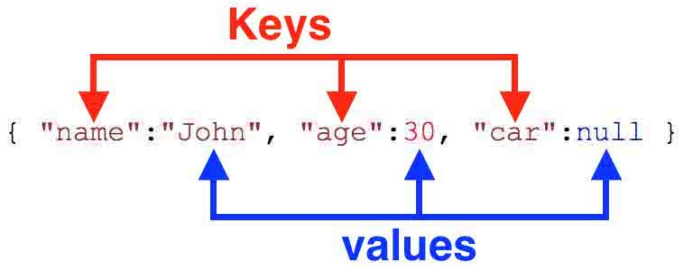
- **JSON stands for JavaScript Object Notation (a lightweight data interchange format).**
- **A text format for storing and transporting data; ex: Web APIs**
- **JSON is "self-describing" and easy to understand**
  - It is easy for humans to read and write and easy for machines to parse and generate.
  - popular choice for representing data in web APIs, mobile app development, configuration files, and client-server communication

```
'{"name":"John", "age":30, "car":null}'
```

In this Example:
It defines an object with 3 properties:

- name
- age
- car

Each property has a value.



JSON Structure

- **JSON represents data as key-value pairs; organized in a hierarchical structure using objects and arrays**
  - Objects (collections of *key-value pairs*),
  - Arrays (ordered lists of values)

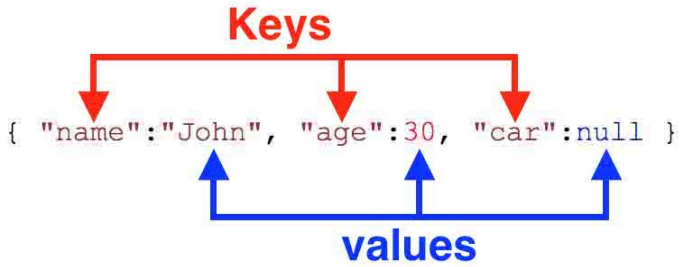
- **JSON accommodates various data types, such as strings, numbers, arrays, objects, booleans, and null, allowing for nested structures.**

```
'{"name":"John", "age":30, "car":null}'
```

In this Example:
It defines an object with 3 properties:

- name
- age
- car

Each property has a value.



JSON: Exampl

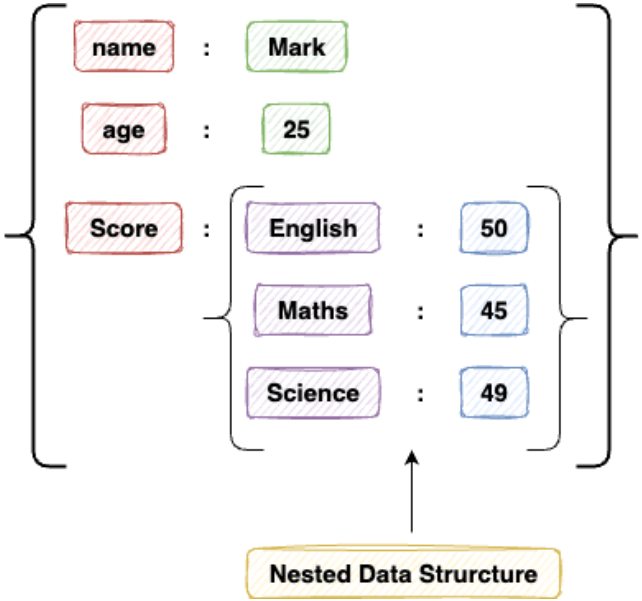
Object

```
{
   creationDate : 2015-11-20T23:19:43.701Z,
   modifiedDate : 2015-11-20T23:19:43.701Z,
   name : demo,
   members : [
      {name : first},
      {name : second},
      {name : third}
   ]
}
```

Name-value pair
comma separated

Array
containing objects with a name-value pair

```
{
  "volume": "blaring",
  "current" : {
              "band": "rednex",
              "song": "cotton eye joe",
              "members":[
                       {"firstname":"Kent","lastname":"Olander"},
                       {"firstname":"Urban","lastname":"Landgren"},
                       {"firstname":"Jonas","lastname":"Lundstrom"},
                       {"firstname":"Tor","lastname":"Nilsson"}
              ]
             },
  "next" : {
          "band": "the dubliners",
          "song": "finnegan's wake",
          "members":[
                   {"firstname":"Ronnie","lastname":"Drew"},
                   {"firstname":"Luke","lastname":"Kelly"},
                   {"firstname":"Ciaran","lastname":"Bourke"},
                   {"firstname":"Barney","lastname":"McKenna"}
          ]
        }
}
```



Json in Python

In Python, the json module is commonly used to work with JSON data.

- **Use json.dumps() to convert Python objects to JSON format**
- **Use json.loads() to convert JSON data back to Python objects.**

```
# Python object
data = {
    'name': 'John Doe',
    'age': 30,
    'city': 'New York',
    'skills': ['JavaScript', 'Python', 'SQL']
}

# Convert Python object to JSON
json_data = json.dumps(data, indent=2)
print("JSON Data:")
print(json_data)

# Convert JSON data to Python object
parsed_data = json.loads(json_data)
print("\nParsed Data:")
print(parsed_data)
```

XML / HTML

- HTML →**H**yper**t**ext **M**arkup **L**anguage
- XML → e**X**tensible **M**arkup **L**anguage

Both are markup languages (define the text document within tag which defines the structure of web pages).

Used for structuring and organizing content on the web and other digital documents, but they serve different purposes:

HTML

```
<!DOCTYPE html>
<html>
    <head>
        <title>Example</title>
        <link rel="stylesheet" href="st
    </head>
    <body>
        <h1>
            <a href="/">Header</a>
        </h1>
        <nav>
            <a href="one/">One</a>
            <a href="two/">Two</a>
            <a href="three/">Three</a>
        </nav>
```

XML

```
<person>
    <address>
        <first_name>Peter</first_name>
        <last_name>Miller</last_name>
        <street>Hauptstrasse</street>
        <number>20</number>
        <city>Zurich</city>
    </address>
</person>
```

HTML

- HTML is primarily used for creating web pages and presenting content on the internet.
- HTML has a fixed set of predefined tags for structuring content, emphasizing elements like headings, paragraphs, links, etc.

```
<!DOCTYPE html>
<html>
    <head>
        <title>Example</title>
        <link rel="stylesheet" href="st
    </head>
    <body>
        <h1>
            <a href="/">Header</a>
        </h1>
        <nav>
            <a href="one/">One</a>
            <a href="two/">Two</a>
            <a href="three/">Three</a>
        </nav>
```

XML

XML is designed to transport and store data, with a focus on describing the structure of the data and is not concerned with presentation.

- **XML allows users to define custom tags, allowing flexibility to create custom structures**

```
<person>
    <address>
        <first_name>Peter</first_name>
        <last_name>Miller</last_name>
        <street>Hauptstrasse</street>
        <number>20</number>
        <city>Zurich</city>
    </address>
</person>
```

How to Get Data?

- Given to you by your company
- Gathered from databases
- From the internet
- From a restful API

Beautiful Soup and Parsing HTML

Beautiful Soup is a Python library for parsing HTML and XML, making web scraping (extracting data directly from web pages) and data extraction easier.

```
soup = BeautifulSoup(page.content, 'html.parser')
soup.find_all('p')
```

```
[<p>Here is some simple content for this page.</p>]
```

Note that `find_all` returns a list, so we'll have to loop through, or use list indexing, it to extract text:

```
soup.find_all('p')[0].get_text()
```

```
'Here is some simple content for this page.'
```

Notes: Don't write own parser. Install Beautiful soup and Use Beautiful Soup to parse HTML content by creating a Beautiful Soup object.

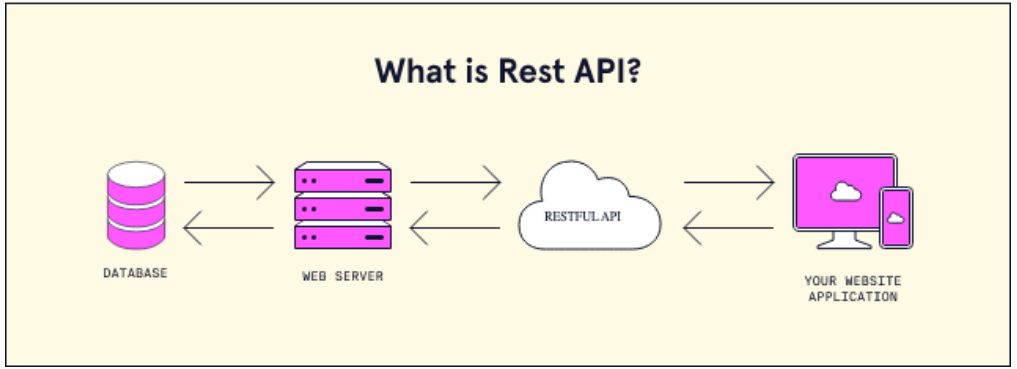
Restful APIs (Application Programming Interface)

RESTful API (Representational State Transfer API) provide **a structured and documented way** to access data from websites or services that offer them.

- APIs are a reliable way to access web data.
- APIs create a service agreement between applications.
- They enable communication via requests and responses.
- API documentation guides effective usage and data interpretation.effectively, interpret the received data correctly etc.

```
import requests
response = requests.get("http://api.open-notify.org/astros.json")
print(response)
```

"If you send me a specific request, I will return some information in a structured and documented format."



Summary

- **Data Types: Various data types (tabular, text, graph, unstructured) impact data preparation in the data science lifecycle.**
- **File Formats: Knowledge of file formats (e.g., CSV, JSON etc.) aids data ingestion and transformation during data preparation.**
- **Databases: are structured repositories for storing and retrieving data, playing a central role in data management during the data science process.**
- **Data Acquisition: RESTful APIs and web scraping are some key for gathering data at the beginning of the data science lifecycle.**

Understanding these elements is crucial for progressing through the data science lifecycle, leading to data-driven insights and solutions.

Next: Additional Reading Slides

These slides are created for your understanding, use them as a helping slides. These additional slides will not be included in quiz / exams

JSON Files & Strings

- **Easy for humans to read (and sanity check, edit)**
- **Example: The JSON object represents a person's information:**

```
{
"name": "John",
"age": 25,
"city": "New York"
}
```


“Key”:“value” pair

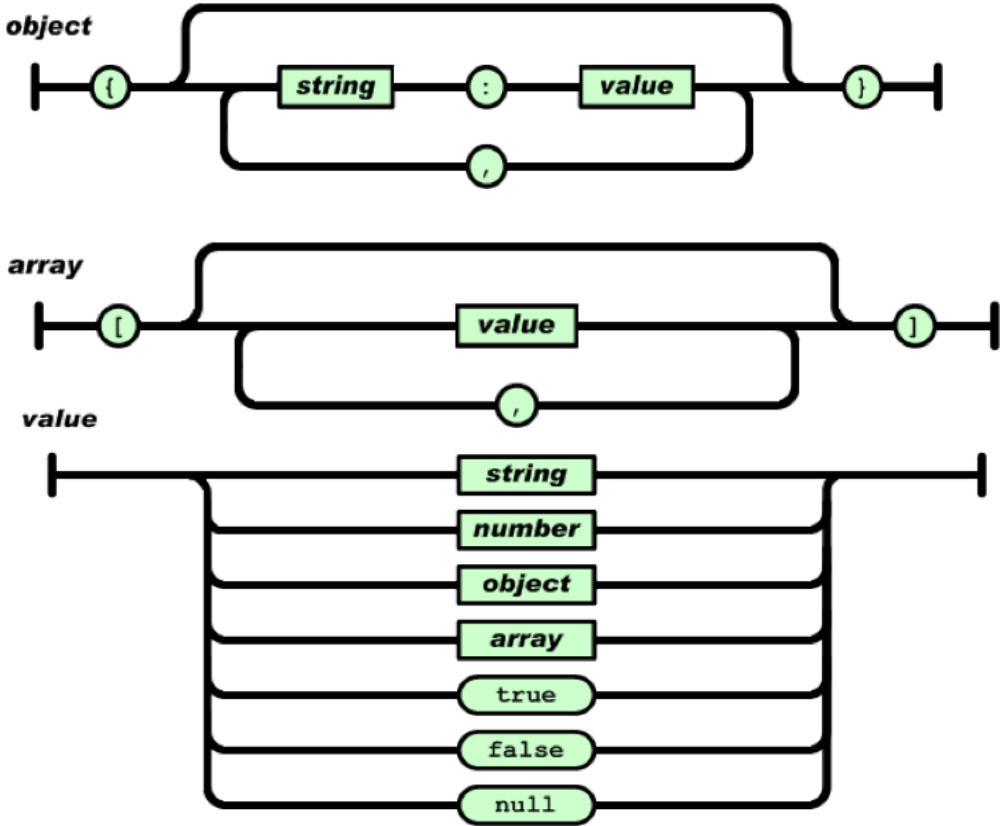
Popular for data interchange and storage due to its simplicity and broad compatibility: Serialization & Deserialization

- **JSON is a method for serializing objects:**
  - Convert an object into a string
  - Deserialization converts a string back to an object

JSON Files & Strings

Valid JSON data type for "value": objects, arrays, strings, numbers, booleans, and null

Defined by three universal data structures



Python dictionary, Java Map, hash table, etc …

```
{
"name": "John",
"age": 25,
}
```

Python list, Java array, vector, etc …

```
["apple", "banana", "orange"]
```

Python string, float, int, boolean, JSON object, JSON array, …

```
"Hello, world!"
```

Images from: http://www.json.org/

JSON In Python

Some built-in types: “Strings”, 1.0, True, False, None

Lists: [“Goodbye”, “Cruel”, “World”]

Dictionaries: {“hello”: “bonjour”, “goodbye”: “au revoir”}

Dictionaries within lists within dictionaries within lists (Nested Structures):

```
[1, 2, {“Help”:[
                “I’m”, {“trapped”: “in”},
                “CMSC320”
                ]}]
```