Berkeley Engineering | BerkeleyHaas

PROFESSIONAL CERTIFICATE IN MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Module 4

Fundamentals of Data Analysis

Office Hours with Viviana Márquez
September 19, 2024

Let's give everyone a couple of minutes to join…

AGENDA

- Required activities for Module 4
- Content review Module 4: Fundamentals of Data Analytics
- Questions

Required Activities for Module 4

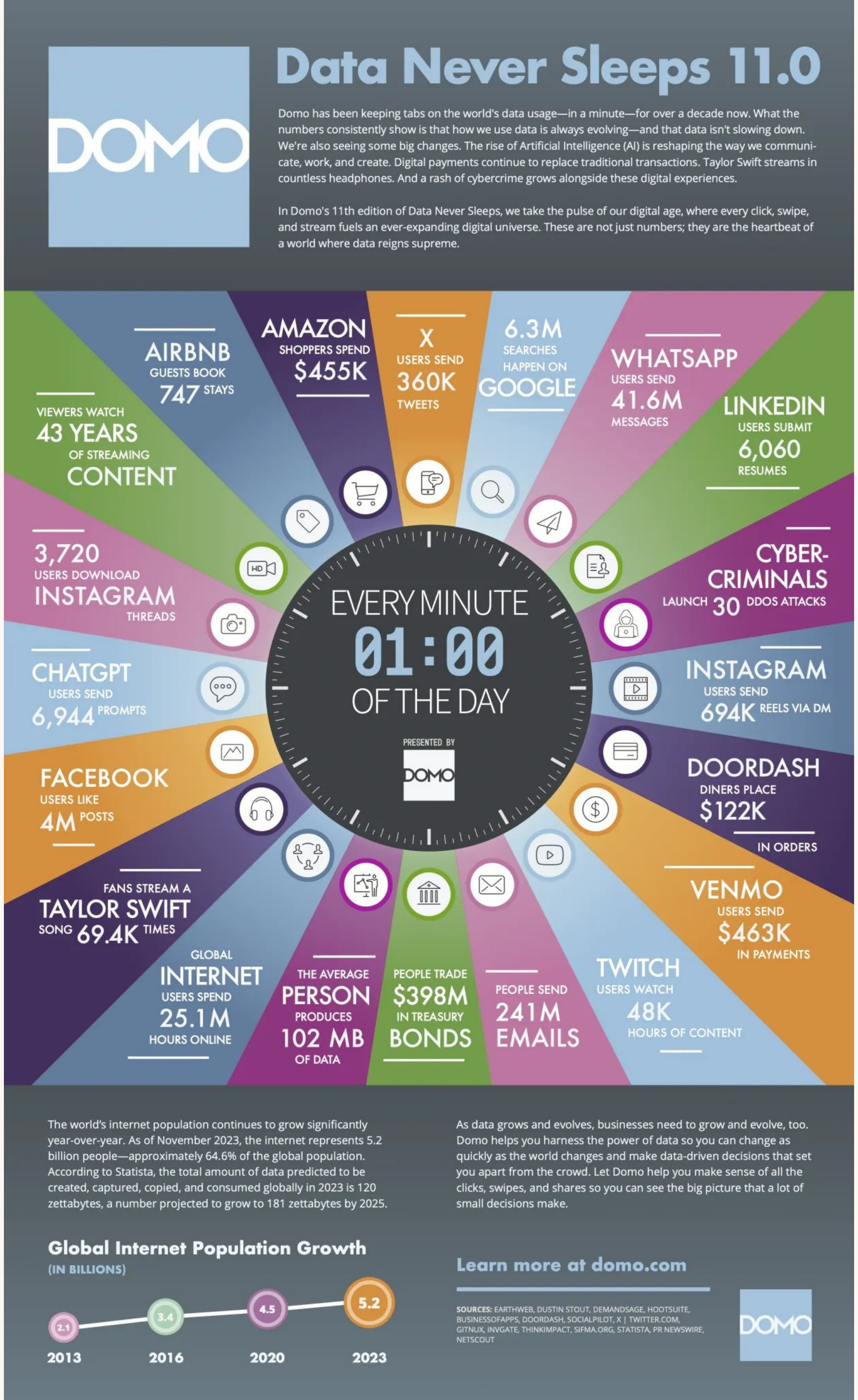
- Required Knowledge Check 4.1: Basic Joins [20:00]
- Required Knowledge Check 4.2: Joining by Multiple Fields [20:00]
- Required Codio Assignment 4.1: Complex Joins on Datasets [60:00]
- Required Knowledge Check 4.3: Creating Plots [20:00]
- Required Knowledge Check 4.4: More Plots [20:00]
- Required Knowledge Check 4.5: String Operations [20:00]
- Required Codio Assignment 4.2: String Operations [60:00]
- Required Codio Assignment 4.3: Data Cleaning [02:00:00]

Content review Module 4: Fundamentals Data Analysis

- Panoramic view: Why are we doing what we are doing?
- Joining tables
- Data cleaning
- Code
    - Pandas merge
    - Data cleaning
        - String operations

Data Never Sleeps 11.0

DOMO



Data Never Sleeps 11.0

Domo has been keeping tabs on the world's data usage—in a minute—for over a decade now. What the numbers consistently show is that how we use data is always evolving—and that data isn't slowing down. We're also seeing some big changes. The rise of Artificial Intelligence (AI) is reshaping the way we communicate, work, and create. Digital payments continue to replace traditional transactions. Taylor Swift streams in countless headphones. And a rash of cybercrime grows alongside these digital experiences.

In Domo's 11th edition of Data Never Sleeps, we take the pulse of our digital age, where every click, swipe, and stream fuels an ever-expanding digital universe. These are not just numbers; they are the heartbeat of a world where data reigns supreme.

EVERY MINUTE 01:00 OF THE DAY

PRESENTED BY DOMO

AMAZON SHOPPERS SPEND $455K

X USERS SEND 360K TWEETS

6.3M SEARCHES HAPPEN ON GOOGLE

WHATSAPP USERS SEND 41.6M MESSAGES

LINKEDIN USERS SUBMIT 6,060 RESUMES

CYBER-CRIMINALS LAUNCH 30 DDOS ATTACKS

INSTAGRAM USERS SEND 694K REELS VIA DM

DOORDASH DINERS PLACE $122K IN ORDERS

VENMO USERS SEND $463K IN PAYMENTS

TWITCH USERS WATCH 48K HOURS OF CONTENT

PEOPLE SEND 241M EMAILS

PEOPLE TRADE $398M IN TREASURY BONDS

THE AVERAGE PERSON PRODUCES 102 MB OF DATA

GLOBAL INTERNET USERS SPEND 25.1M HOURS ONLINE

FANS STREAM A TAYLOR SWIFT SONG 69.4K TIMES

FACEBOOK USERS LIKE 4M POSTS

CHATGPT USERS SEND 6,944 PROMPTS

3,720 USERS DOWNLOAD INSTAGRAM THREADS

VIEWERS WATCH 43 YEARS OF STREAMING CONTENT

AIRBNB GUESTS BOOK 747 STAYS

The world's internet population continues to grow significantly year-over-year. As of November 2023, the internet represents 5.2 billion people—approximately 64.6% of the global population. According to Statista, the total amount of data predicted to be created, captured, copied, and consumed globally in 2023 is 120 zettabytes, a number projected to grow to 181 zettabytes by 2025.

As data grows and evolves, businesses need to grow and evolve, too. Domo helps you harness the power of data so you can change as quickly as the world changes and make data-driven decisions that set you apart from the crowd. Let Domo help you make sense of all the clicks, swipes, and shares so you can see the big picture that a lot of small decisions make.

Global Internet Population Growth (IN BILLIONS)

Learn more at domo.com

SOURCES: EARTHWEB, DUSTIN STOUT, DEMANDSAGE, HOOTSUITE, BUSINESSOFAPPS, DOORDASH, SOCIALPILOT, X | TWITTER.COM, GITNUX, INVGATE, THINKIMPACT, SIFMA.ORG, STATISTA, PR NEWSWIRE, NETSCOUT

Data is the new oil!

What kind of data can be gathered from something as simple as buying a cup of coffee?

Instead of relying on guesses or intuition, we can harness the power of data to make informed, accurate decisions.

Data Professionals – What are the different job families?

Data Analysts & BI

They look at past data to figure out what happened and why.

What happened?

Data Science

They use data to build models that predict what will happen in the future.

What will happen?

Data Engineering

They design and maintain the systems that store and move data to make it usable.

How can we make data accessible and usable?

Data Governance

They ensure data is accurate, secure, and used according to rules and regulations.

Is the data accurate, secure, and compliant?

Machine Learning Engineer

They take predictive models and turn them into practical tools and systems that work in real life.

How can we make it happen?"

Data Professionals – What tools do they use?

Data Analysts & BI

- Tableau
  PowerBI
  LookerStudio
- Excel
- SQL
- Some R
  Some Python

Data Science

- Python/R
- Jupyter notebooks
- Scikit-Learn
- PyTorch
  TensorFlow
- SQL

Data Engineering

- Hadoop
- Kafka
- SQL
- MongoDB
- Java

Data Governance

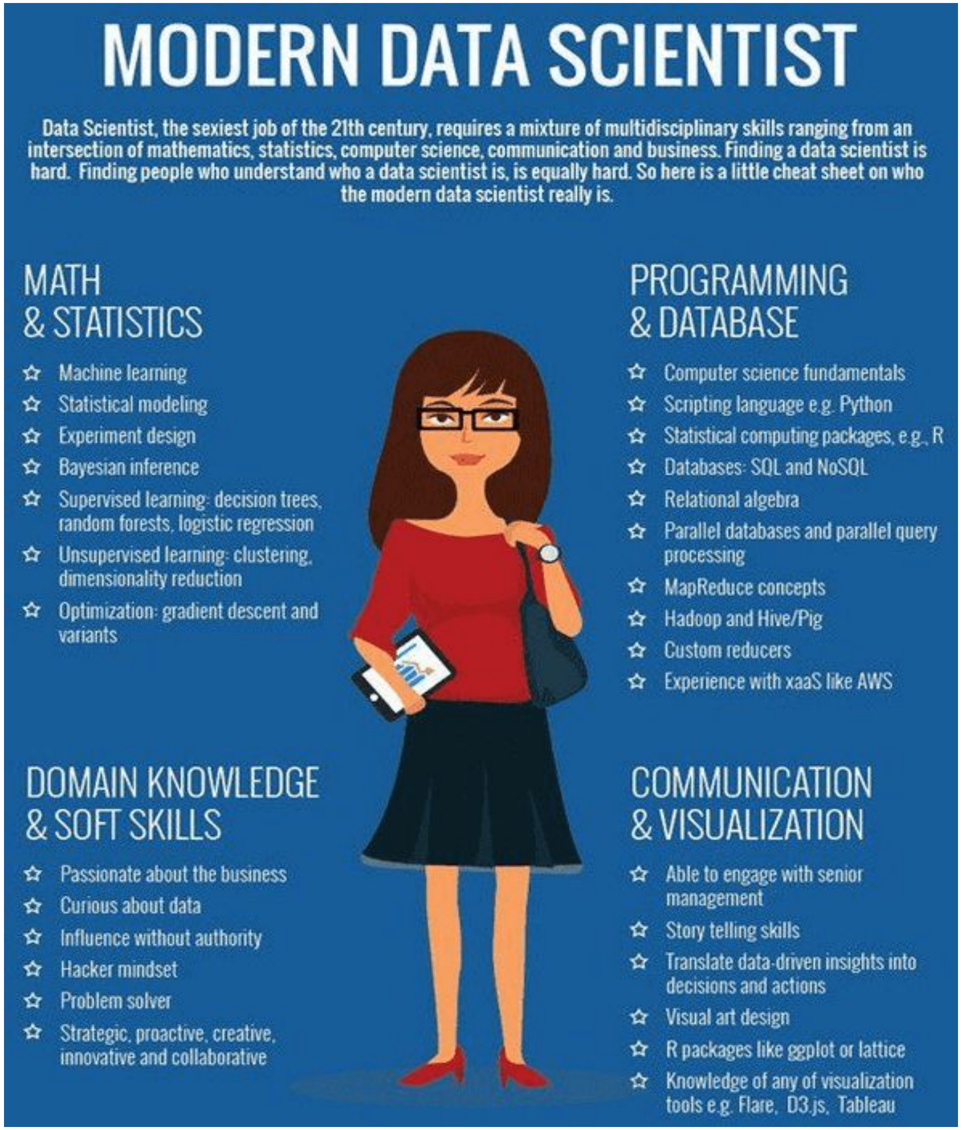
- Erwin
- others…

Machine Learning Engineer

- Python/R
- Scikit-Learn
- PyTorch
  TensorFlow
- SQL
- Spark
- Docker
- Git
- AWS
  GCP
  Azure

Data Scientists – What skills do they have?

A data scientist will use anything that helps—whether it's **AI**, stats, or something new—to better understand and make sense of data.



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

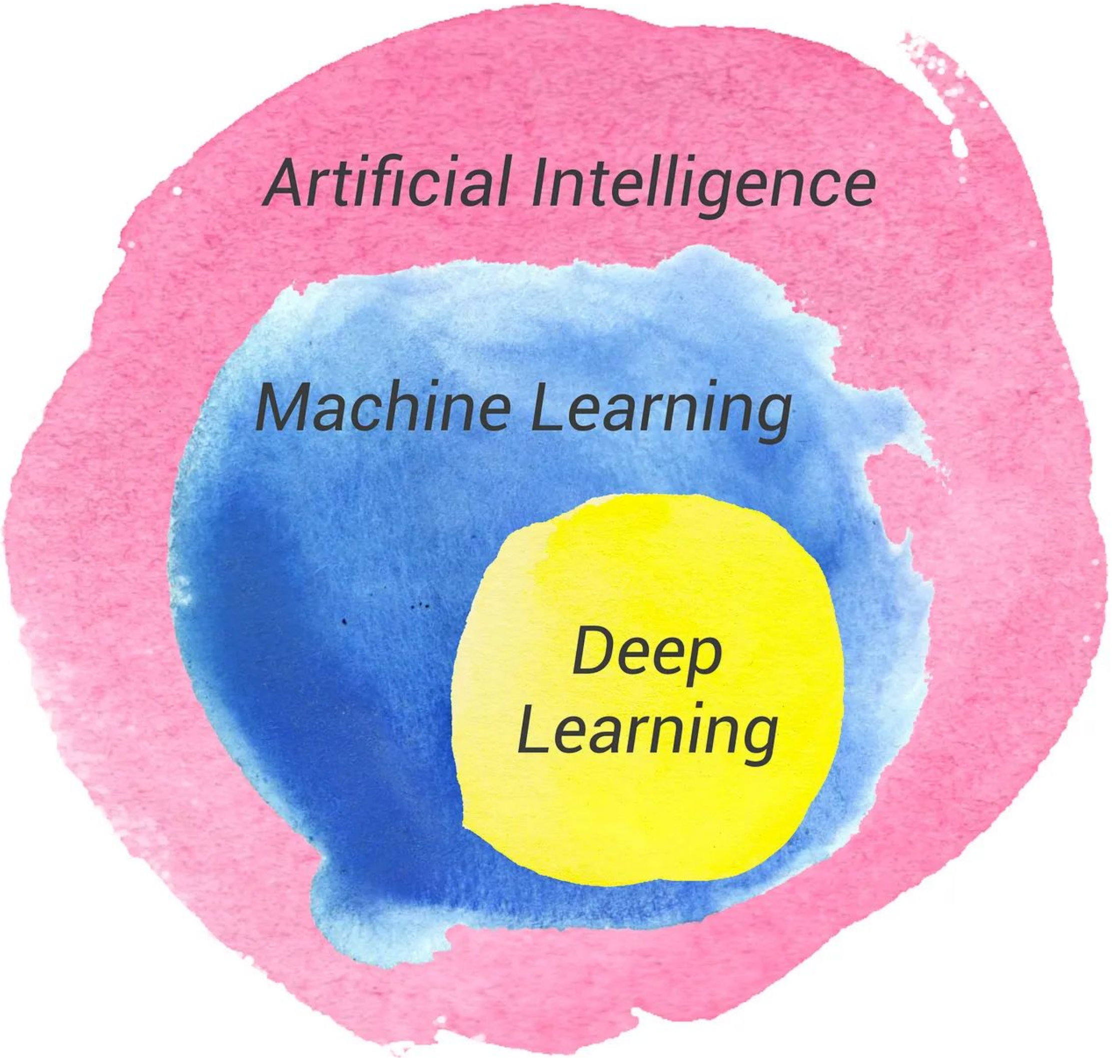
- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, e.g., R
- ☆ Databases: SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

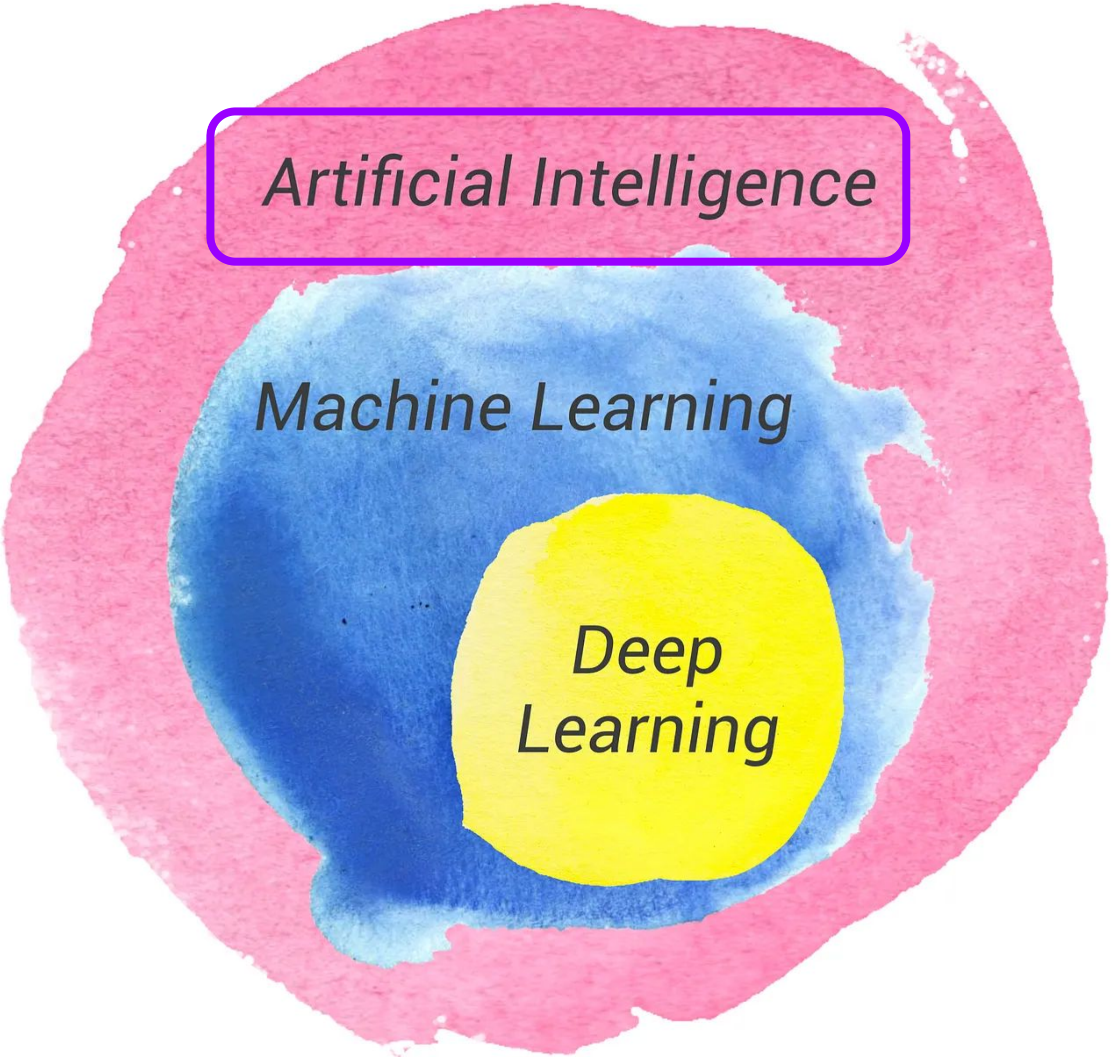
- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau


Artificial Intelligence
Machine Learning
Deep Learning

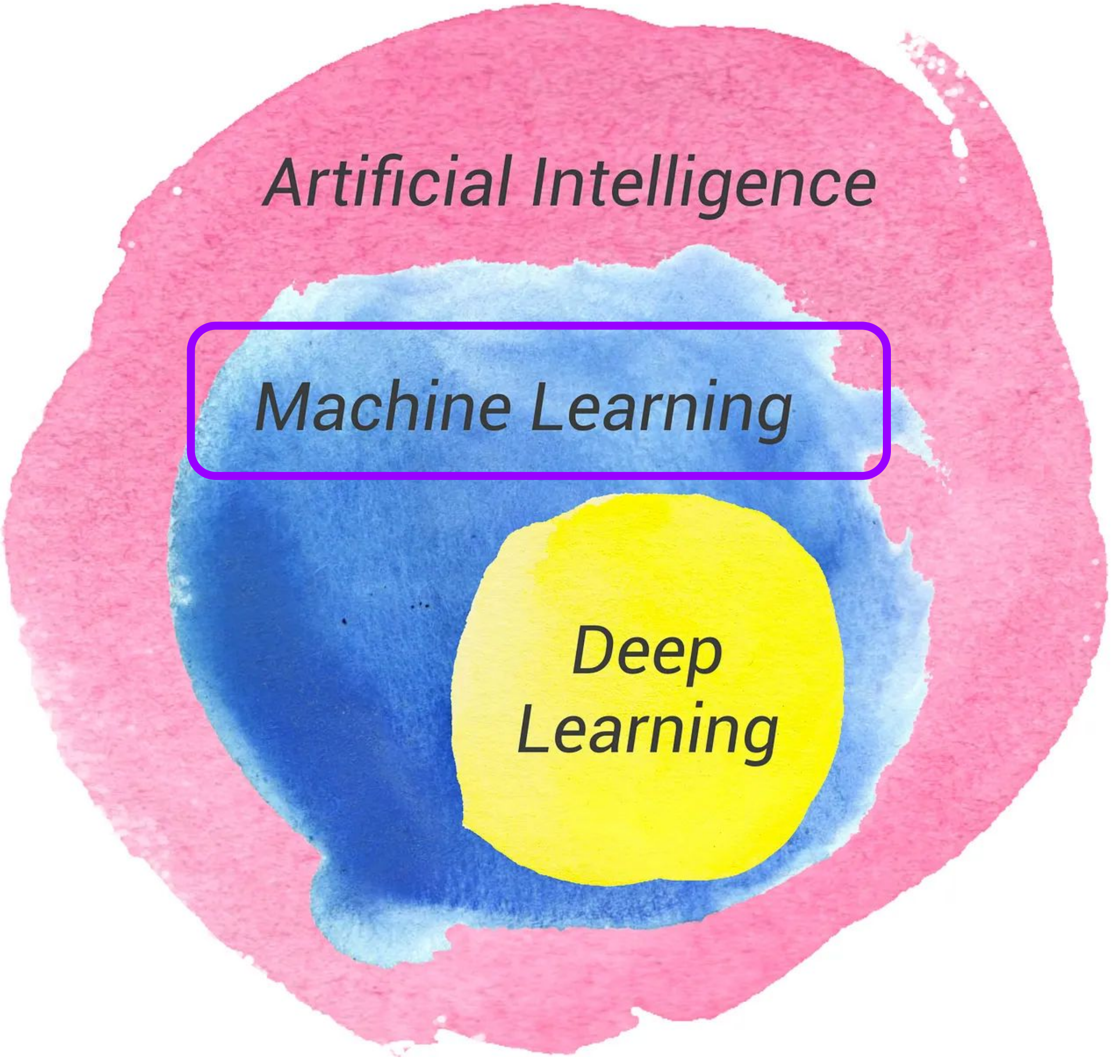
AI
AI everywhere



Artificial Intelligence

AI is the broad field where machines are designed to mimic human intelligence, enabling them to perform tasks like decision-making, language understanding, and problem-solving.

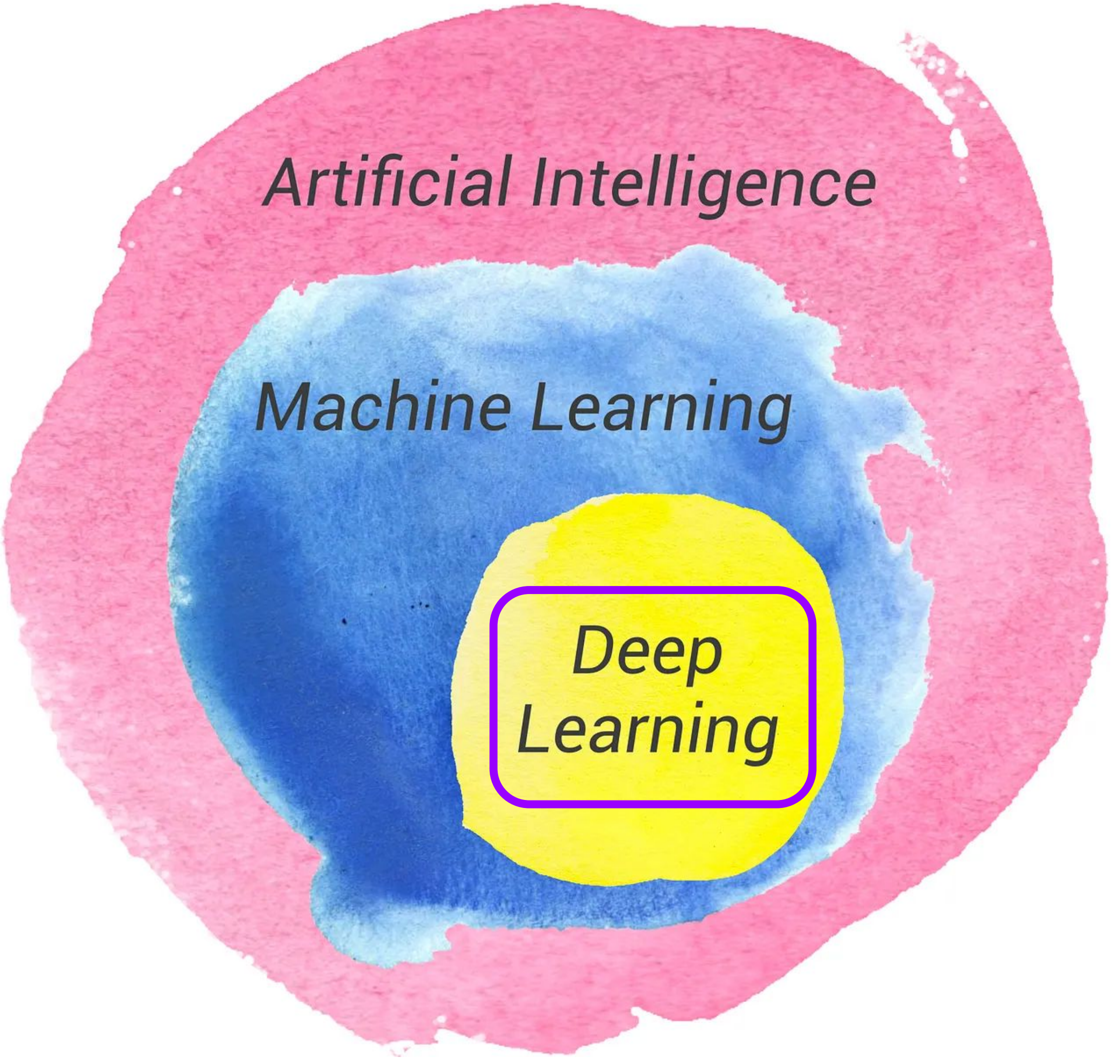
Scope: Broad. AI encompasses everything that allows computers to imitate human intelligence, including robotics, natural language processing, and problem-solving.



Machine Learning

ML is a subset of AI that focuses on creating systems that can learn and improve from experience or data, without being explicitly programmed for every task.

Scope: Moderate. It includes various techniques such as regression, classification, clustering, and ensemble models.



Deep Learning

DL is a specialized type of machine learning that uses neural networks with many layers to analyze vast amounts of data and learn complex patterns, often achieving results comparable to human performance in areas like image recognition or language translation.

Scope: Narrow. DL is a specific, yet powerful, form of machine learning.

Obtaining data for a Data Science project

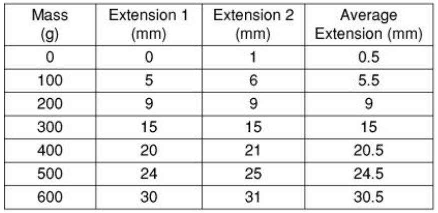
- Without data, there is no project, as the entire foundation of machine learning relies on having the right data to train, test, and validate models

- Data acquisition is the process of identifying, collecting, and extracting useful information from various sources for use in data science

- The quality, relevance, and variety of the data obtained directly influence the model's effectiveness, accuracy, and performance, making data an essential component for the success of the project

GARBAGE DATA
GREAT MODEL
GARBAGE RESULTS

Structured data vs unstructured data



Mass (g)	Extension 1 (mm)	Extension 2 (mm)	Average Extension (mm)
0	0	1	0.5
100	5	6	5.5
200	9	9	9
300	15	15	15
400	20	21	20.5
500	24	25	24.5
600	30	31	30.5

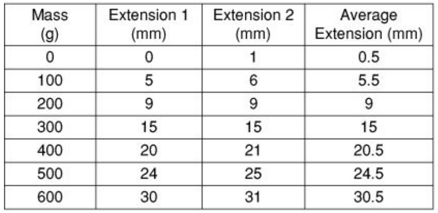
- **Structured data** is highly organized and easily readable by machines. It is typically stored in tabular formats, such as spreadsheets (CSV, Excel) or relational databases (SQL).

  Each observation is in a row, and its features are in predefined columns, making it easier to process and analyze.

- **Unstructured data** does not follow a specific format or structure, making it more challenging to organize and analyze. This type of data includes free text, images, videos, audio, and other multimedia formats.

  Due to its nature, unstructured data often requires advanced techniques such as Natural Language Processing (NLP) or Convolutional Neural Networks (CNN).

Structured data vs unstructured data



Mass (g)	Extension 1 (mm)	Extension 2 (mm)	Average Extension (mm)
0	0	1	0.5
100	5	6	5.5
200	9	9	9
300	15	15	15
400	20	21	20.5
500	24	25	24.5
600	30	31	30.5

Usually Machine Learning

- **Structured data** is highly organized and easily readable by machines. It is typically stored in tabular formats, such as spreadsheets (CSV, Excel) or relational databases (SQL).

  Each observation is in a row, and its features are in predefined columns, making it easier to process and analyze.

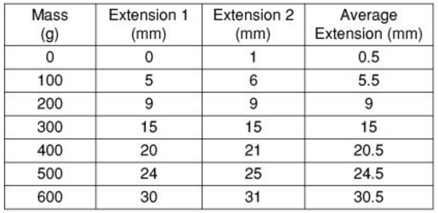
Usually Deep Learning

- **Unstructured data** does not follow a specific format or structure, making it more challenging to organize and analyze. This type of data includes free text, images, videos, audio, and other multimedia formats.

  Due to its nature, unstructured data often requires advanced techniques such as Natural Language Processing (NLP) or Convolutional Neural Networks (CNN).

Structured data vs unstructured data

Usually Machine Learning



Mass (g)	Extension 1 (mm)	Extension 2 (mm)	Average Extension (mm)
0	0	1	0.5
100	5	6	5.5
200	9	9	9
300	15	15	15
400	20	21	20.5
500	24	25	24.5
600	30	31	30.5

- **Structured data** is highly organized and easily readable by machines. It is typically stored in tabular formats, such as spreadsheets (CSV, Excel) or relational databases (SQL).

  Each observation is in a row, and its features are in predefined columns, making it easier to process and analyze.

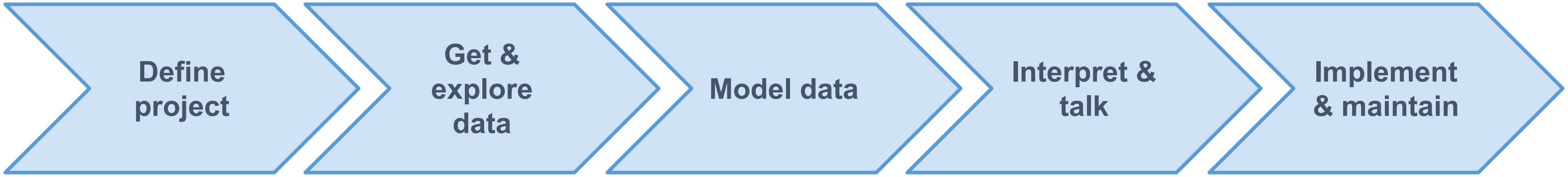
We'll focus on this kind of data today

Usually Deep Learning

- **Unstructured data** does not follow a specific format or structure, making it more challenging to organize and analyze. This type of data includes free text, images, videos, audio, and other multimedia formats.

  Due to its nature, unstructured data often requires advanced techniques such as Natural Language Processing (NLP) or Convolutional Neural Networks (CNN).

The Data Science Lifecycle



Define project

- Specify business problem
- Acquire domain knowledge

Get and explore data

- Find appropriate data
- Exploratory Data Analysis
- Clean and pre-process data
- Feature engineering

Model data

- Determine ML task
- Build candidate models
- Select model based on performance metrics

Interpret & talk

- Interpret model
- Communicate model insights

Implement & maintain

- Set up function to predict on new data
- Document process
- Monitor and maintain model

Exploratory Data Analysis (EDA)

- It’s an initial process in data analysis, with the main goal of gaining a deep understanding of the dataset. It is used to explore, summarize, and visualize the data before applying any machine learning model.

Importance of EDA in machine learning:

- **Identifying patterns and relationships:** It allows for the discovery of important patterns in the data that could influence the model's outcomes.

- **Detecting errors and anomalies:** EDA helps identify outliers, measurement errors, or missing data that could negatively impact model performance.

- **Understanding the problem context:** It aids in understanding the data's characteristics, distribution, relationships between variables, and data structure, which improves decision-making.

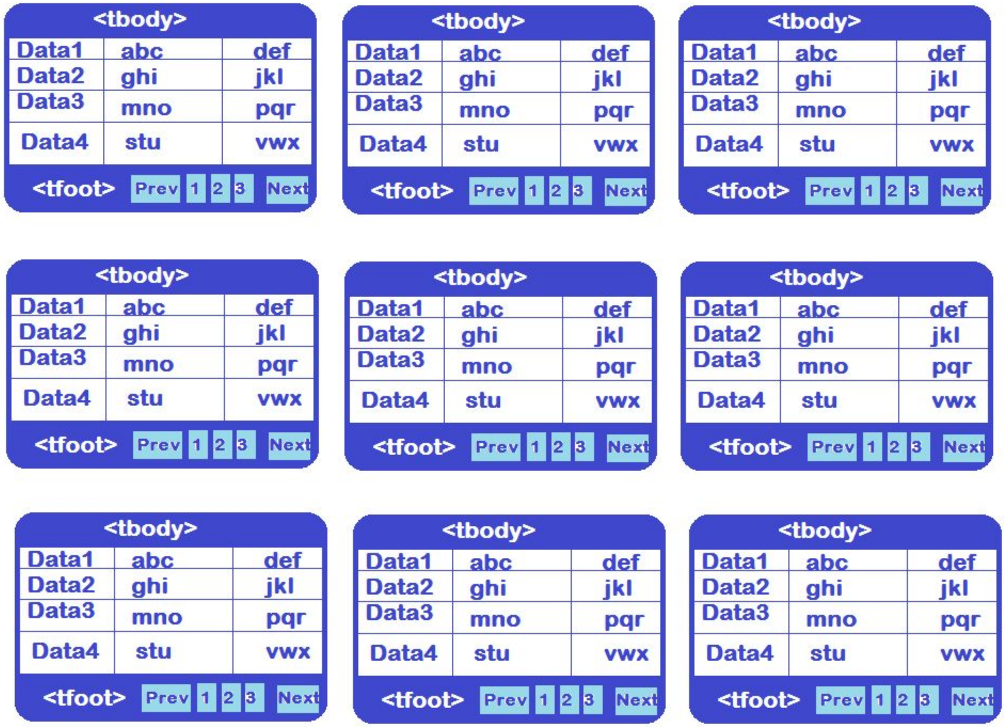
- **Improving data quality:** A well-conducted EDA can lead to more effective data cleaning, enhancing the final model's accuracy.

Let’s code!

-

Joining tables

- In the real-world, often data will be split across multiple tables
- You will have to combine records from those tables based on related columns



<tbody>		
Data1	abc	def
Data2	ghi	jkl
Data3	mno	pqr
Data4	stu	vwx
<tfoot>	Prev 1 2 3	Next

<tbody>		
Data1	abc	def
Data2	ghi	jkl
Data3	mno	pqr
Data4	stu	vwx
<tfoot>	Prev 1 2 3	Next

<tbody>		
Data1	abc	def
Data2	ghi	jkl
Data3	mno	pqr
Data4	stu	vwx
<tfoot>	Prev 1 2 3	Next

<tbody>		
Data1	abc	def
Data2	ghi	jkl
Data3	mno	pqr
Data4	stu	vwx
<tfoot>	Prev 1 2 3	Next

<tbody>		
Data1	abc	def
Data2	ghi	jkl
Data3	mno	pqr
Data4	stu	vwx
<tfoot>	Prev 1 2 3	Next

<tbody>		
Data1	abc	def
Data2	ghi	jkl
Data3	mno	pqr
Data4	stu	vwx
<tfoot>	Prev 1 2 3	Next

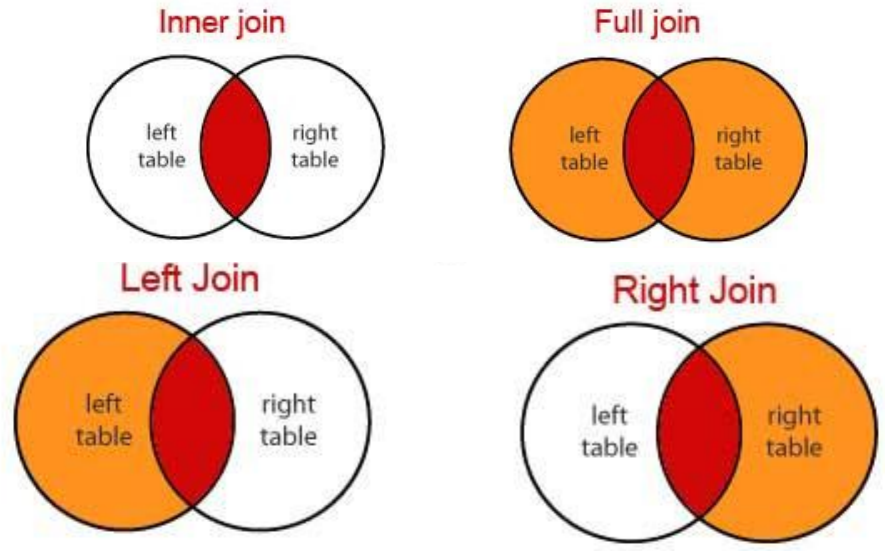
<tbody>		
Data1	abc	def
Data2	ghi	jkl
Data3	mno	pqr
Data4	stu	vwx
<tfoot>	Prev 1 2 3	Next

<tbody>		
Data1	abc	def
Data2	ghi	jkl
Data3	mno	pqr
Data4	stu	vwx
<tfoot>	Prev 1 2 3	Next

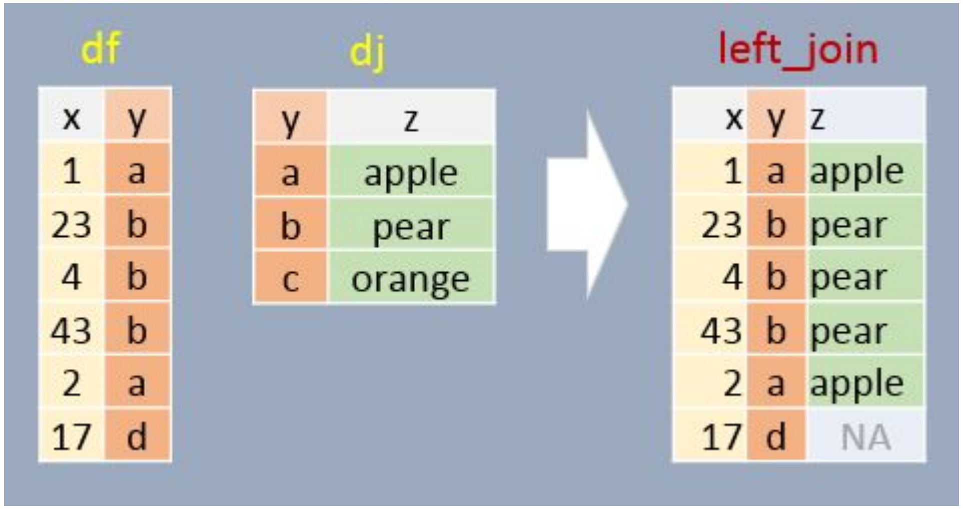
<tbody>		
Data1	abc	def
Data2	ghi	jkl
Data3	mno	pqr
Data4	stu	vwx
<tfoot>	Prev 1 2 3	Next

Types of joins

- **INNER JOIN**: Returns records that have matching values in both tables.
- **LEFT (OUTER) JOIN**: Returns all records from the left table, and the matched records from the right table.
- **RIGHT (OUTER) JOIN**: Returns all records from the right table, and the matched records from the left table.
- **FULL (OUTER) JOIN**: Returns all records when there is a match in either the left or the right table.
- **CROSS JOIN**: Returns the Cartesian product of the two tables.
- **SELF JOIN**: Joining a table with itself.



Left join



df

x	y
1	a
23	b
4	b
43	b
2	a
17	d

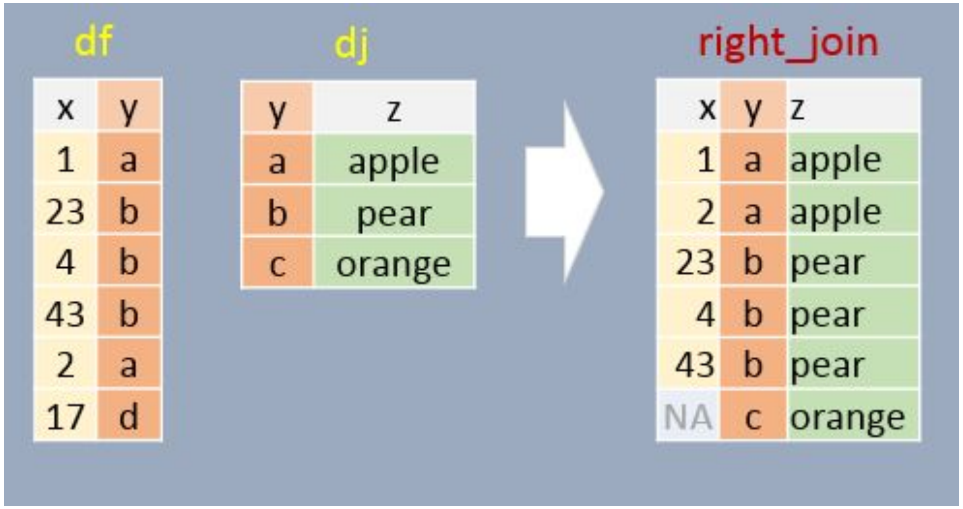
dj

y	z
a	apple
b	pear
c	orange

left_join

x	y	z
1	a	apple
23	b	pear
4	b	pear
43	b	pear
2	a	apple
17	d	NA

Right join



df

x	y
1	a
23	b
4	b
43	b
2	a
17	d

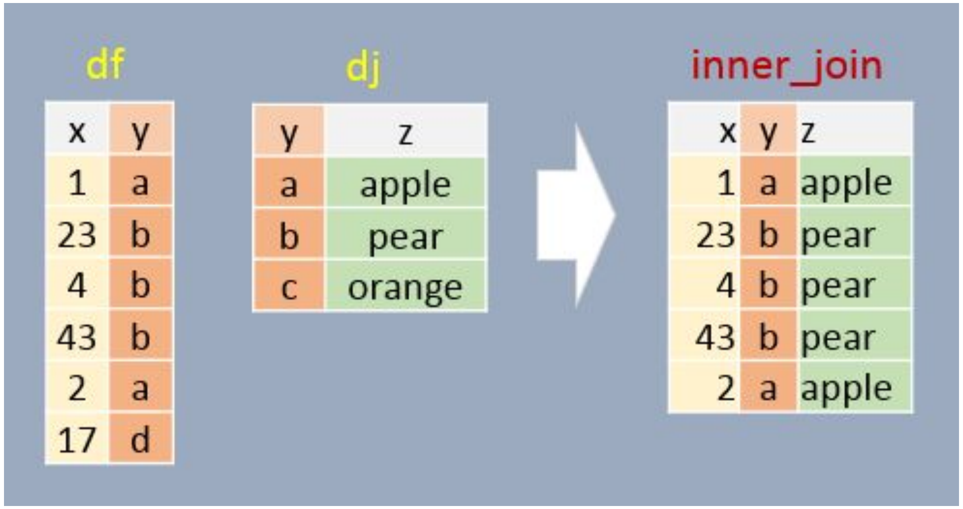
dj

y	z
a	apple
b	pear
c	orange

right_join

x	y	z
1	a	apple
2	a	apple
23	b	pear
4	b	pear
43	b	pear
NA	c	orange

Inner join



df

x	y
1	a
23	b
4	b
43	b
2	a
17	d

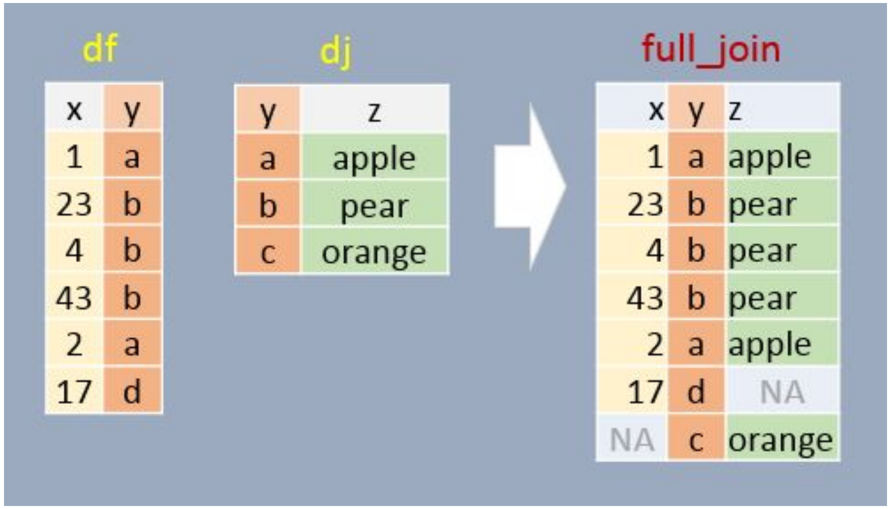
dj

y	z
a	apple
b	pear
c	orange

inner_join

x	y	z
1	a	apple
23	b	pear
4	b	pear
43	b	pear
2	a	apple

Full join



df

x	y
1	a
23	b
4	b
43	b
2	a
17	d

dj

y	z
a	apple
b	pear
c	orange

full_join

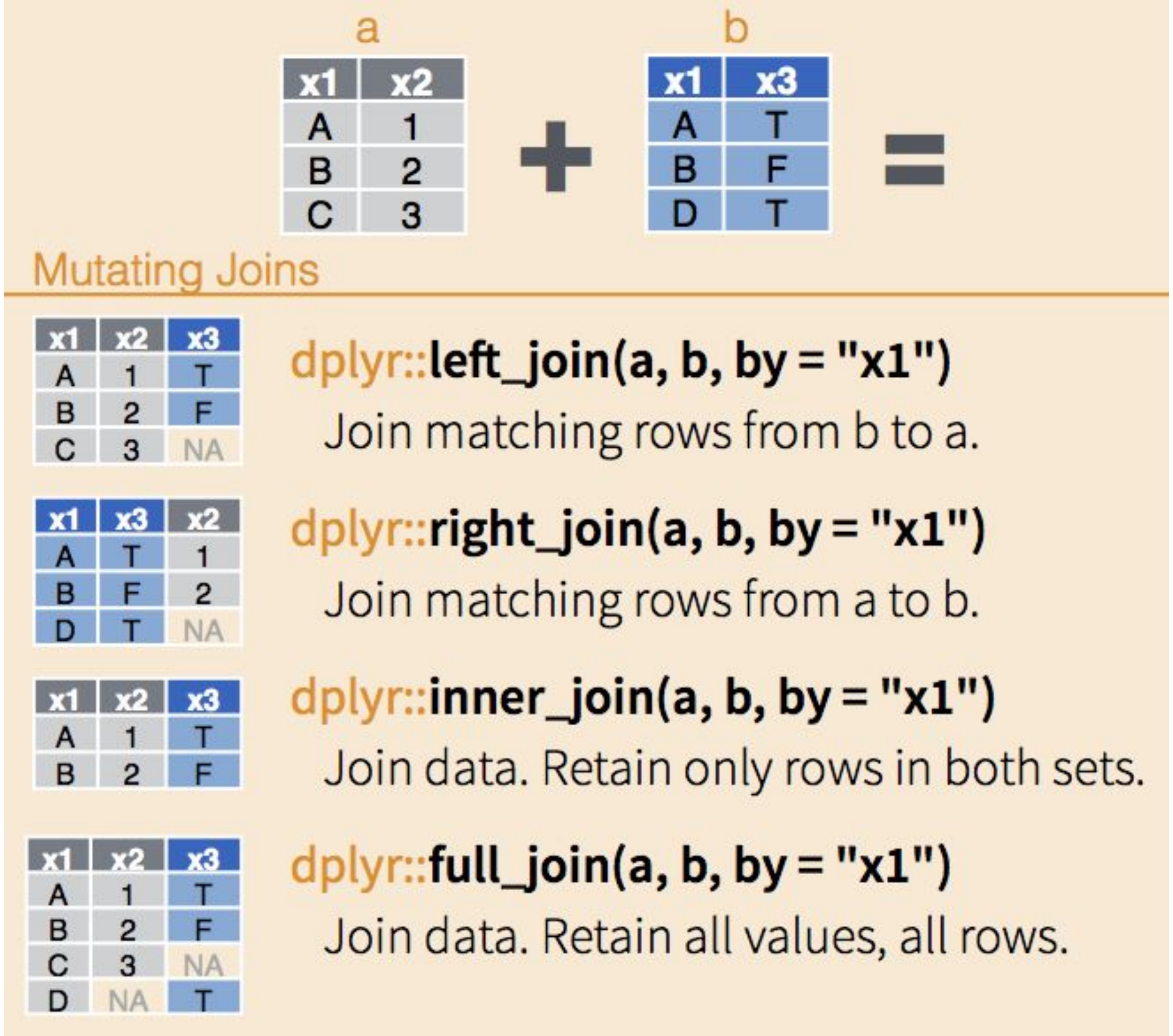
x	y	z
1	a	apple
23	b	pear
4	b	pear
43	b	pear
2	a	apple
17	d	NA
NA	c	orange

Data cleaning

- Included but not limited to:
  - Handling missing values: drop them or account for them
  - Handling outliers: drop them or account for them or keep them
  - Remove duplicates
  - Handling incorrect data types
  - Handling inconsistent data (example: age shouldn't be negative)

QUESTIONS?

R Joins



a

x1	x2
A	1
B	2
C	3

+

b

x1	x3
A	T
B	F
D	T

=

Mutating Joins

x1	x2	x3
A	1	T
B	2	F
C	3	NA

dplyr::**left_join(a, b, by = "x1")**

Join matching rows from b to a.

x1	x3	x2
A	T	1
B	F	2
D	T	NA

dplyr::**right_join(a, b, by = "x1")**

Join matching rows from a to b.

x1	x2	x3
A	1	T
B	2	F

dplyr::**inner_join(a, b, by = "x1")**

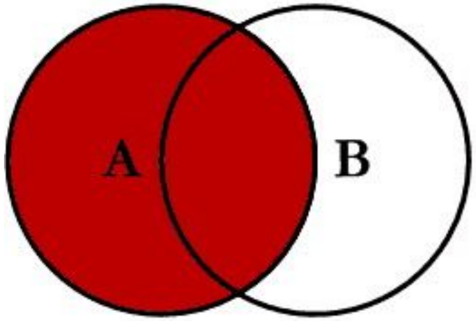
Join data. Retain only rows in both sets.

x1	x2	x3
A	1	T
B	2	F
C	3	NA
D	NA	T

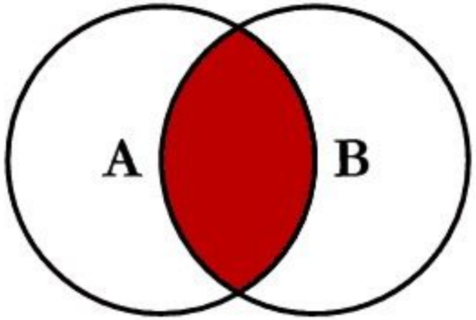
dplyr::**full_join(a, b, by = "x1")**

Join data. Retain all values, all rows.

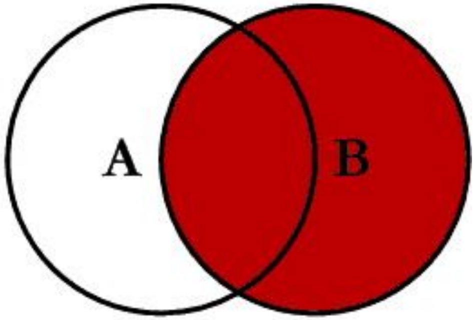
SQL JOINS



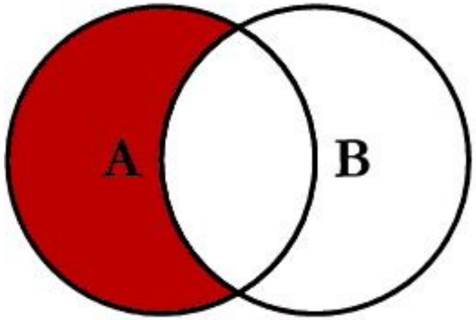
```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
```



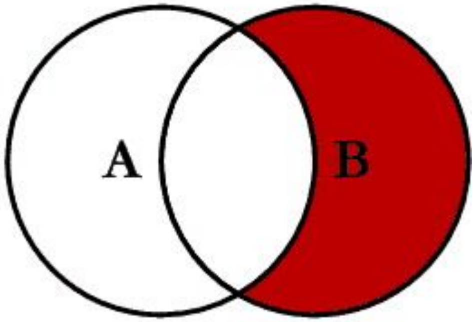
```
SELECT <select_list>
FROM TableA A
INNER JOIN TableB B
ON A.Key = B.Key
```



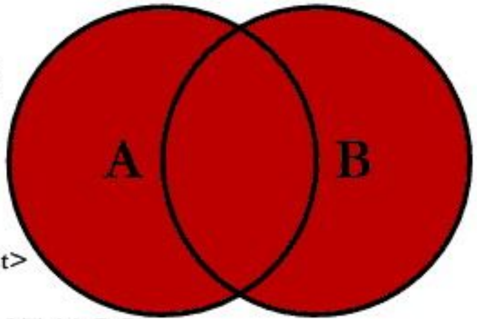
```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
```



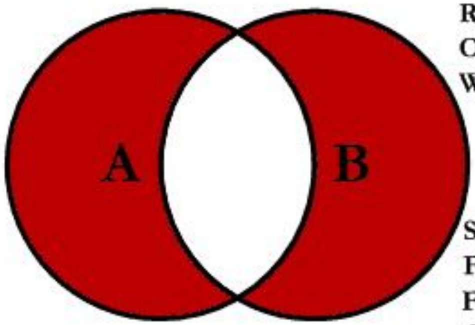
```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
WHERE B.Key IS NULL
```



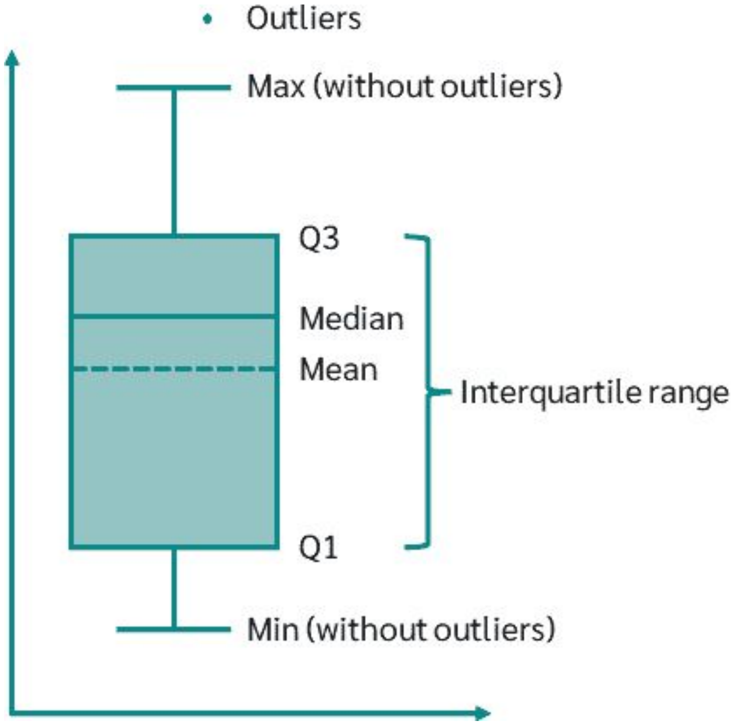
```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
```



```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
```



```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
OR B.Key IS NULL
```



The box indicates the range in which the middle 50% of all data lies

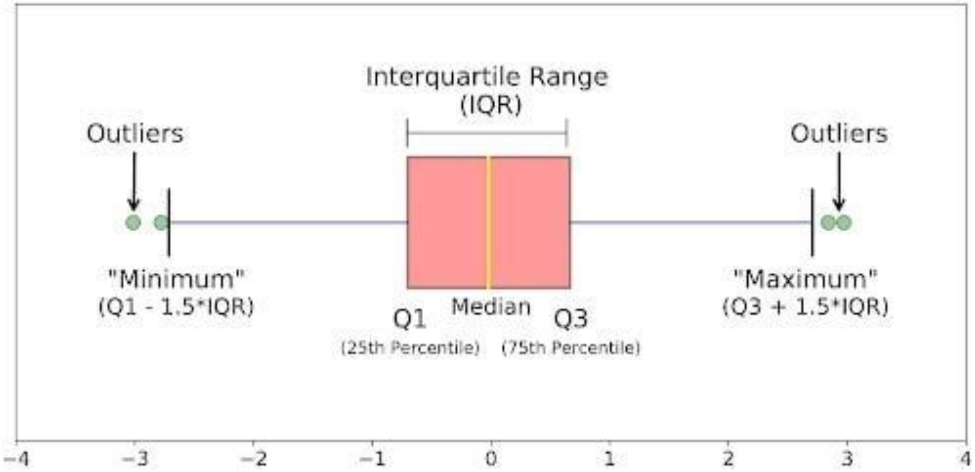
Thus, the lower end of the box is the 1st quartile and the upper end is the 3rd quartile

Between Q1 and Q3, is the interquartile range

In the boxplot, the solid line indicates the median and the dashed line indicates the mean.

The T-shaped whiskers go to the last point, which is still within 1.5 times the interquartile range.

Points that are further away are considered extreme values (outliers).



Descriptive statistics

Method to summarize and describe the main features of a dataset

Measures of central tendency

- **Mean**
  Arithmetic average of the data.
  Calculated by adding all the values and dividing them by the number of values
- **Median**
  Middle value when the data points are arranged in order
- **Mode**
  The most frequently occurring value in the dataset

Measures of variability (dispersion)

- **Range**
  The difference between the maximum and minimum values
- **Variance**
  A measure of how much the data points differ from the mean on average
- **Standard deviation**
  Square root of the variance, giving a sense of how much data points typically deviate from the mean in the same units as the data

Correlation

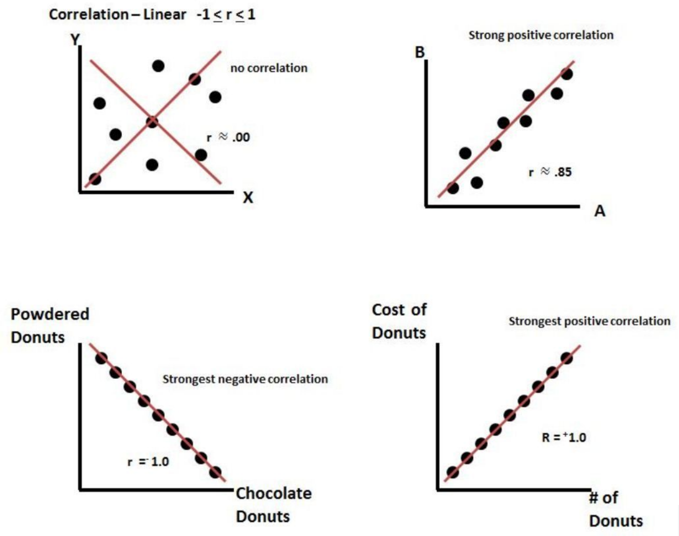
Correlation is a statistical measure that quantifies the **strength** and **direction** of the linear relationship between two variables. It ranges from -1 to 1, where:

- A value of 1 indicates a perfect **positive correlation**, meaning that as one variable increases, the other variable increases proportionally.

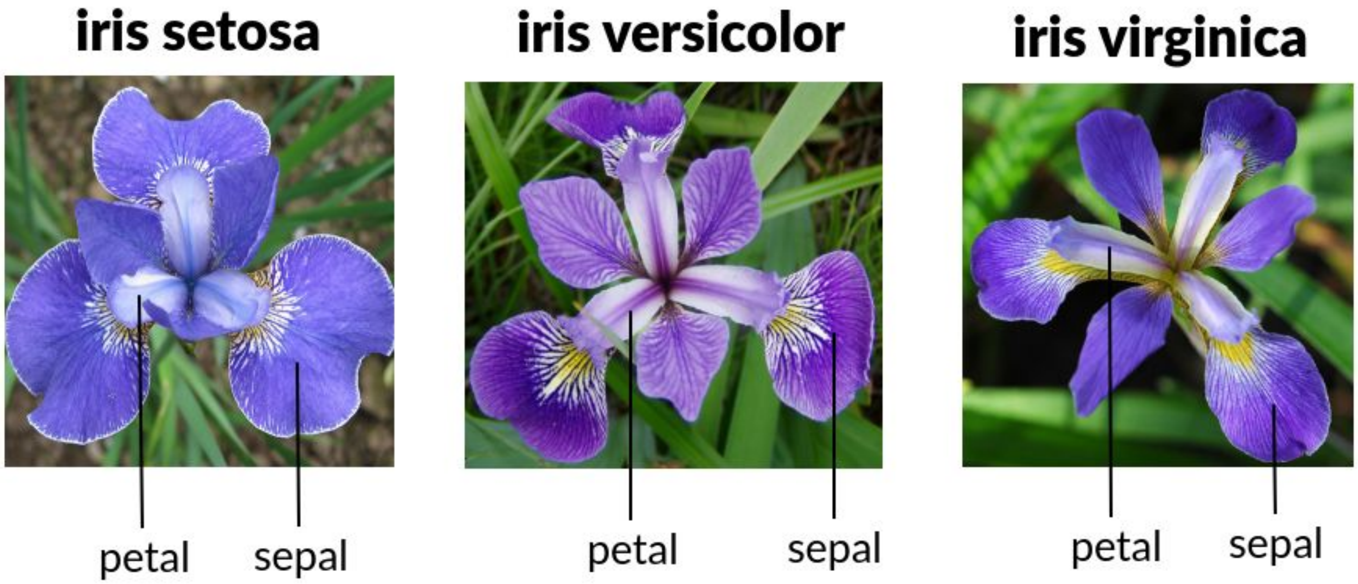
- A value of -1 indicates a perfect **negative correlation**, implying that as one variable increases, the other variable decreases proportionally.

- A value of 0 indicates **no linear correlation** between the variables.

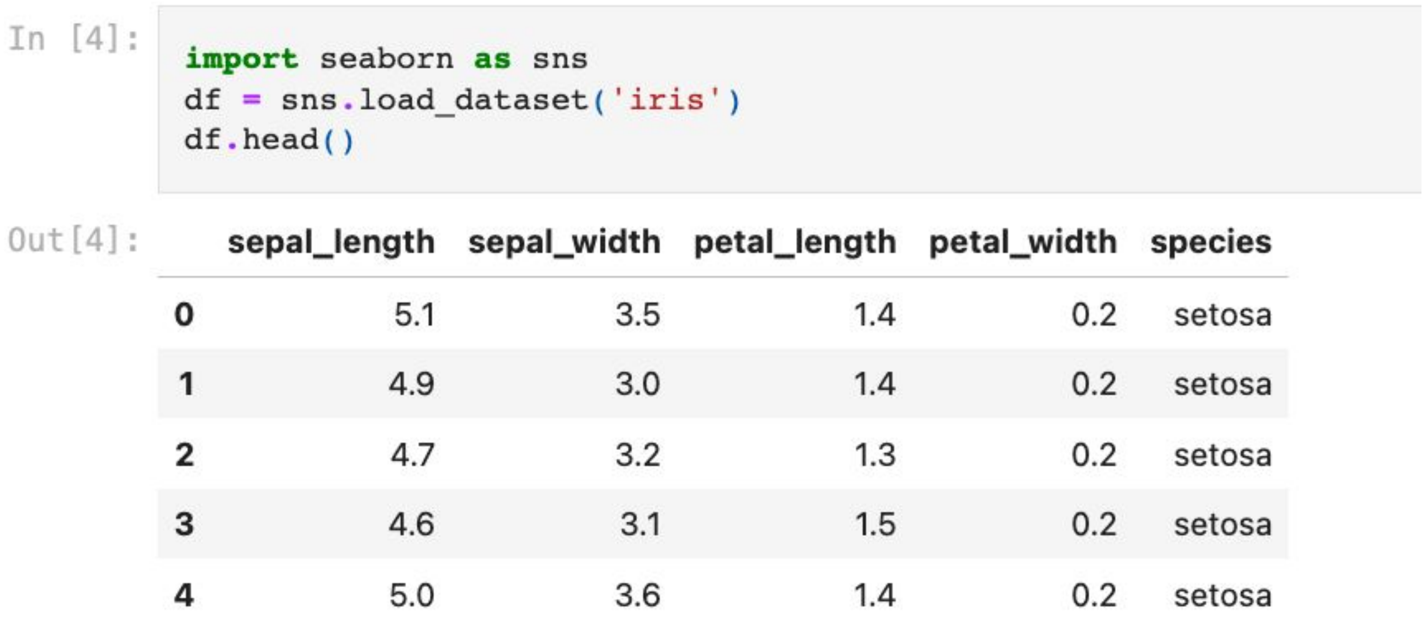
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$



Parts of a Machine Learning model



Parts of a Machine Learning model



In [4]:

```
import seaborn as sns
df = sns.load_dataset('iris')
df.head()
```

Out[4]:

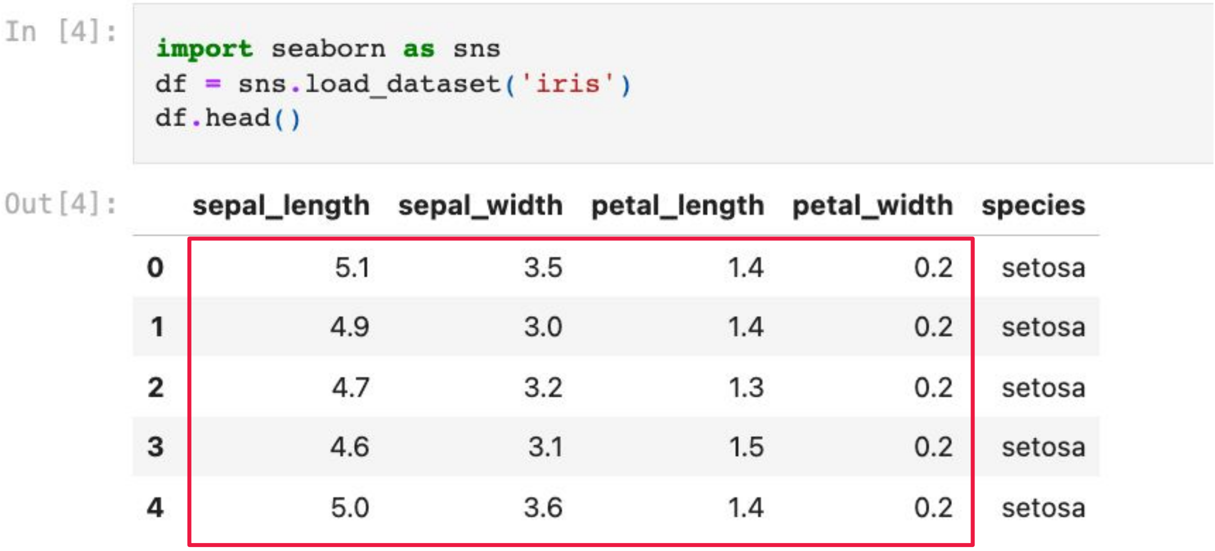
	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

Parts of a Machine Learning model

Model Inputs

Also known as:

- Features
- Attributes
- Predictors
- Inputs
- **Independent Variables**
- Dimensions
- X
- Probably more…



In [4]:

```
import seaborn as sns
df = sns.load_dataset('iris')
df.head()
```

Out[4]:

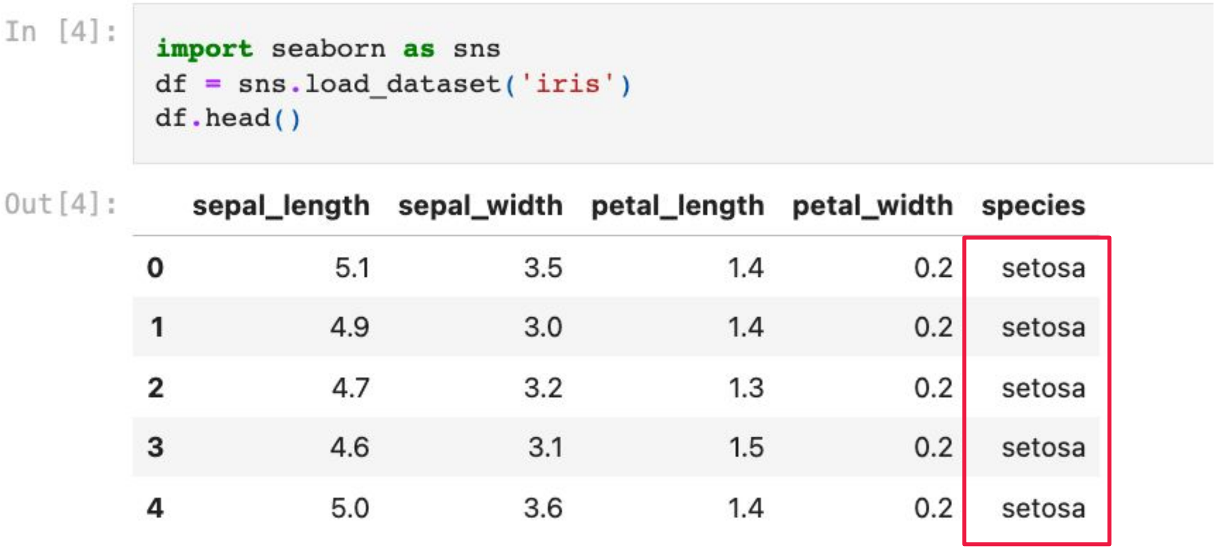
	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

Parts of a Machine Learning model

Model Outputs (What you're trying to predict)

Also known as:

- Target
- Response
- Output
- **Dependent Variable**
- Labels
- Y
- Probably more…



In [4]:

```
import seaborn as sns
df = sns.load_dataset('iris')
df.head()
```

Out[4]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa