BCA - THIRD YEAR SIXTH SEMESTER

DATA SCIENCE

CSA 7109T

Detailed & Descriptive Notes

HandNotes created by Kamal Kishor

UNIT-1: Introduction to Data Science

1. What is Data?

Data is a collection of raw facts and figures which do not have meaning by themselves.

- Marks of students
- Temperature readings
- Customer purchase records

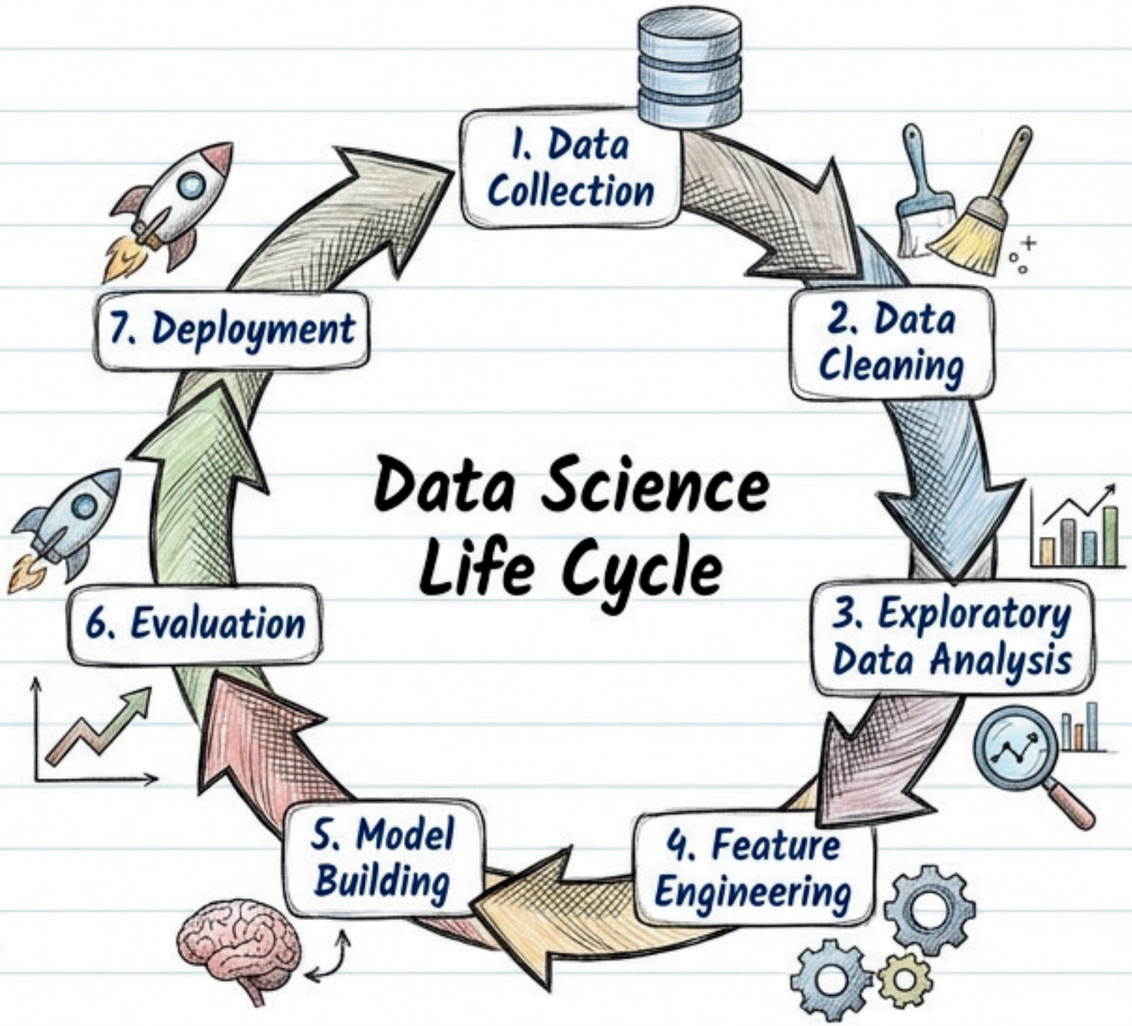
2. What is Data Science?

An interdisciplinary field that uses Statistics, Mathematics, Computer Science, and Machine Learning to extract meaningful insights, patterns, and knowledge from data.

Goal: To help organizations make data-driven decisions.

Need: Huge amount of data (Social media, IoT), Manual analysis is impossible, Need automation/intelligent analysis.

Applications: Netflix recommendations, Fraud detection, Weather prediction, Healthcare diagnosis.



The World of Big Data

Extremely large and complex datasets that traditional databases cannot handle efficiently.

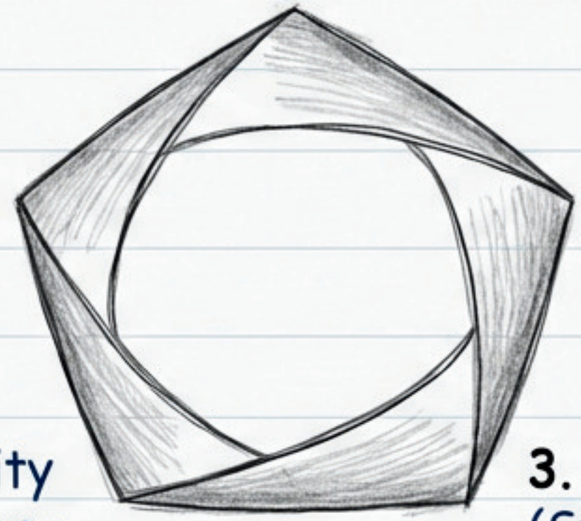
1. Volume - Huge amount of data (TB, PB)

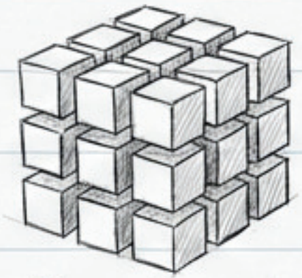
2. Velocity - Speed of data generation

3. Variety - Types of data
(Structured, semi-structured, unstructured)

4. Veracity - Data quality and accuracy

5. Value - Useful insights from data





Structured
(Databases)

Semi-Structured
(XML/JSON)

Unstructured
(Audio/Video/Text)

Big Data in Business

Role of Big Data

- ☑ Improves decision making
- ☑ Helps in predictive analysis
- ☑ Enhances customer experience
- ☑ Cost reduction

Big Data Driven Business Models

- Data-as-a-Service (DaaS)
- Analytics-as-a-Service
- Recommendation-based models
- Personalized marketing
- Targeted advertising
- Predictive maintenance
- Fraud detection

Challenges in Data Science

- ⚠ Data quality issues
- ⚠ Privacy & security
- ⚠ Scalability
- ⚠ Handling unstructured data
- ⚠ Skill shortage

UNIT-2: Descriptive Statistics

Formula Sheet

Measures of Central Tendency

Mean (Average value):

$$\text{Mean} = \frac{\text{Sum of } x_i}{n}$$

Median (Middle value):

If n is odd: $\left(\frac{n+1}{2}\right)^{th}$ value

If n is even: $\frac{x_{n/2} + x_{(n/2)+1}}{2}$

Mode: Most frequently occurring value.

Measures of Dispersion

Range:

$$\text{Range} = X_{max} - X_{min}$$

Variance:

Population: $\sigma^2 = \frac{\sum(x_i - \text{mean})^2}{N}$

Sample: $s^2 = \frac{\sum(x_i - \text{mean})^2}{n - 1}$

Standard Deviation:

$$\sigma = \sqrt{\text{Variance}}$$

Inferential Statistics & Probability

Used to draw conclusions about a population from a sample.

Hypothesis Testing

Null Hypothesis (H_0): No effect / no difference

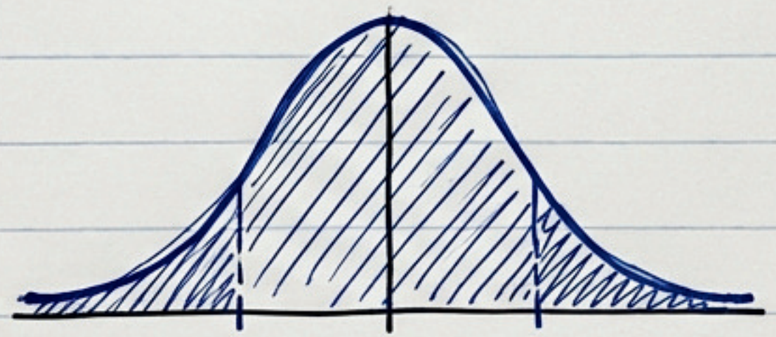
Alternative Hypothesis (H_1): There is an effect / difference

Level of Significance (alpha): Probability of rejecting true H_0.
Common values: 0.05 or 0.01

Random Variables

Discrete: $P(X=x)$

Continuous: Integral $f(x)\,dx$



$$f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-0.5\left(\frac{x-\mu}{\sigma}\right)^2}$$

Key Formulas

Z-Test: $z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$

t-Test: $t = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}}$

Probability:

$$P(A) = \frac{\text{Favorable outcomes}}{\text{Total outcomes}}$$

UNIT-3: Pattern Recognition Basics

1. **Problem Definition** - Understand objective,
   Define inputs and outputs.

↓

2. **Representation of Input** - Raw data converted to structured format
   (e.g., Image → pixel values).

$$\text{Feature Vector } X = (x_1, x_2, x_3, \cdots, x_n)$$

↓

3. **Feature Engineering** - Process of selecting, extracting, and transforming variables.

Good features = Better model

Examples : Height, weight → BMI
Date → day, month, year

Dataset Management

$$D = D_{train} + D_{validation} + D_{test}$$

Learn patterns | Tune model | Final evaluation

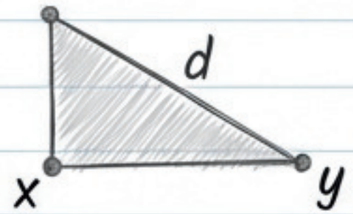
Accuracy Formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

UNIT-4: Distance Metrics

1. Euclidiean Distance - Straight-line distance. Used in kNN, Clustering.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$



2. Manhattan Distance - Sum of absolute differences. Grid-based distance.

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

3. Hamming Distance - Number of mismatched positions.

$$d = \sum (x_i \neq y_i)$$

Example:

1 0 1 1 1 0

1 0 1 0 0

Mismatch at pos 3 and 5 → Hamming distance = 2

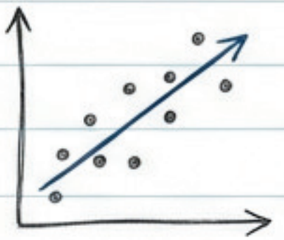
Similarity Measures & Correlation

Correlation

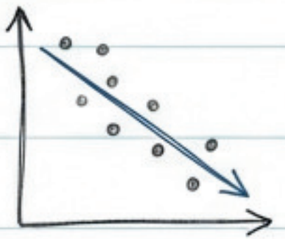
Measures relationship between variables: Positive, Negative, Zero.

Pearson Correlation Coefficient Formula

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



Positive



Negative

Higher-Order Measures

Cosine Similarity $\cos(\theta) = \frac{\dot{X} \cdot Y}{\|X\| \|Y\|}$

- Jaccard Index

UNIT-5: Supervised Learning Techniques

1. Regression (Predicts continuous values, e.g., Salary)

Simple Linear Regression Formula: $y = mx + c$

Slope $m = \frac{\text{Sum}((x-\bar{x}\_bar)(y-y\_bar))}{\text{Sum}((x-\bar{x}\_bar)^2}$

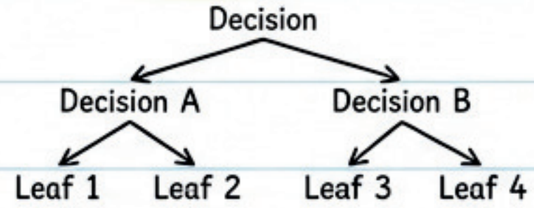
2. Classification (Predicts categories, e.g., Spam)

Logistic Regression Formula: $P(y=1|x) = \frac{1}{1+e^{-z}}$

3. Decision Trees - Tree-like structure. Nodes=decisions, Leaves=output.

Entropy Formula: $H(S) = -\sum (p_i \cdot \log_2(p_i))$

Information Gain: $IG = H(S) - \sum (|S_i|/|S| \cdot H(S_i))$



4. Support Vector Machine (SVM) - Finds optimal hyperplane.

Formula: $\dot{w} \cdot \vec{x} + b = 0$

5. Random Forest - Collection of decision trees. Reduces overfitting.

Formula: $\text{Prediction} = (1/N) \cdot \text{Sum}(T_i(x))$

Unsupervised Learning & Algorithms

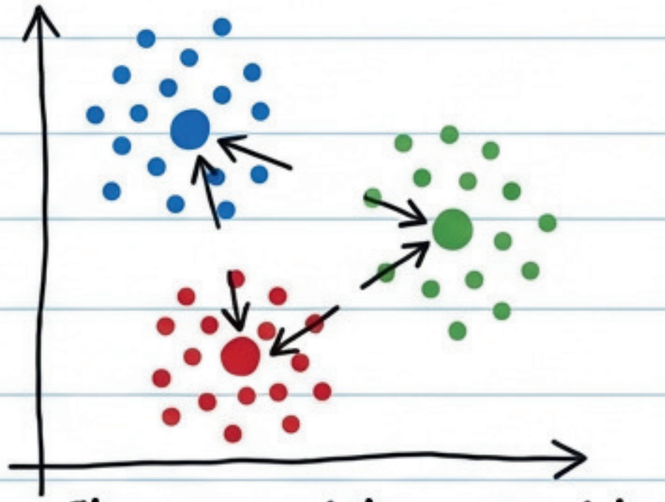
1. Clustering

Groups similar data points.

2. K-Means Algorithm

1. Choose K
2. Assign clusters
3. Update centroids
4. Repeat until convergence

Objective Function Formula: $J = \sum(\sum(\|x - \mu_i\|^2)$



Clusters with centroids

3. Association Rule Mining

Finds relationships (e.g., Bread -> Butter)

Key Metrics

Support(A) = (Transactions containing A) / (Total transactions)

Confidence(A->B) = Support(A intersect B) / Support(A)

Lift = Confidence(A->B) / Support(B)

Challenges in Machine Learning

Overfitting: Model learns noise instead of pattern (performs well on train, bad on test).

Underfitting: Model is too simple to capture the pattern.

Bias: Error due to overly simplistic assumptions.

High Dimensionality: Dataset has too many features, making it hard to find patterns (Curse of Dimensionality).

UNIT-6: Data Visualization

Graphical representation of data.

Graph Types: Bar chart, Line graph, Pie chart, Histogram.

Graph Types: Bar chart, Line graph, Pie chart, Histogram.

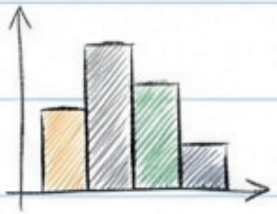
Data Summaries: Mean, Median, Charts, Tables.

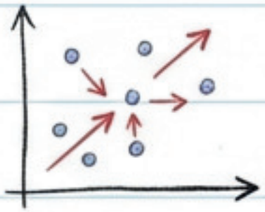
Visualization Dimensions:

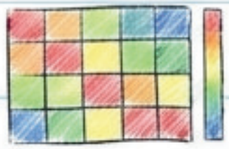
One-dimensional -> Histogram

Two-dimensional -> Scatter plot

Multidimensional -> Heatmaps



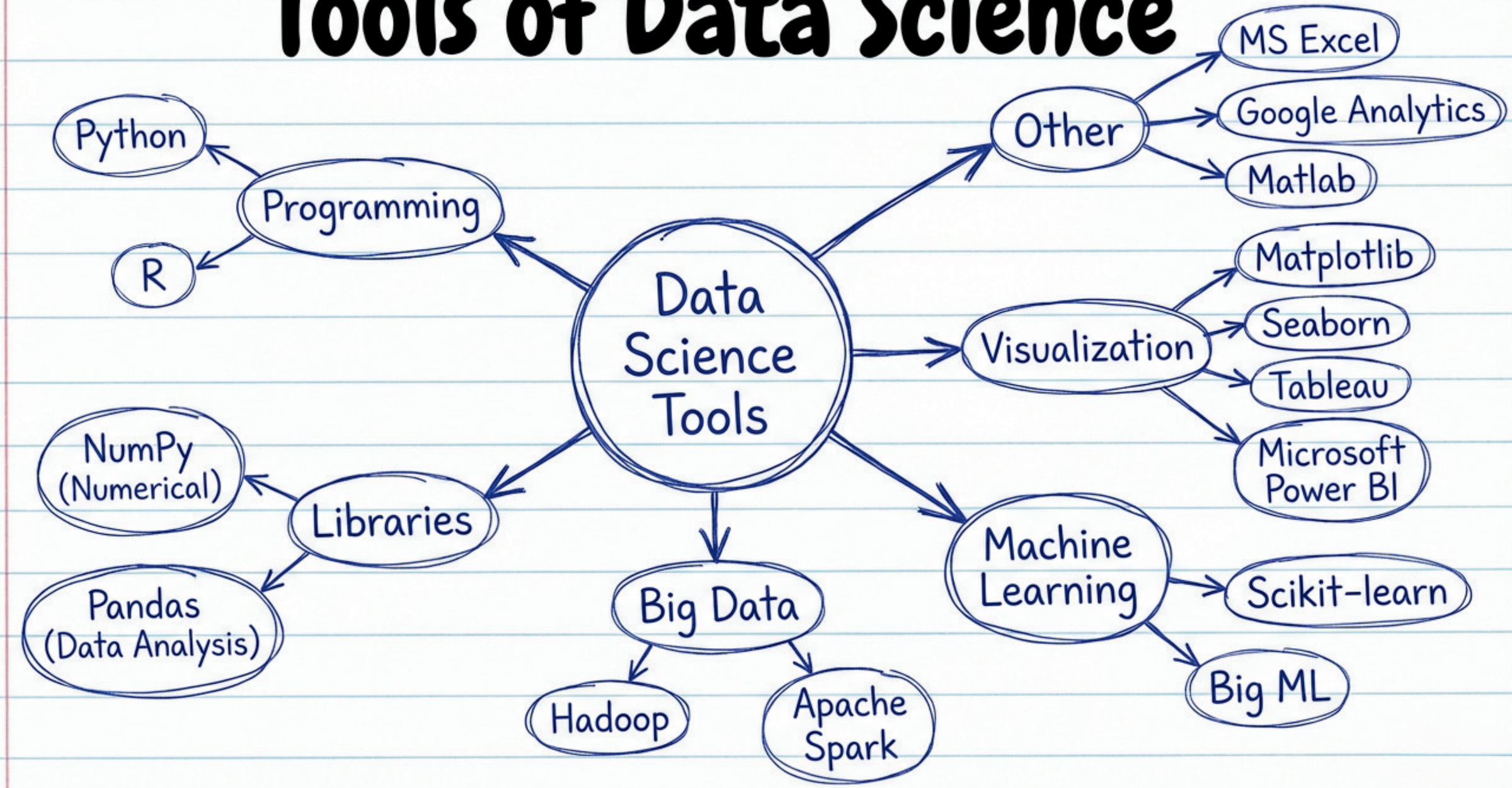




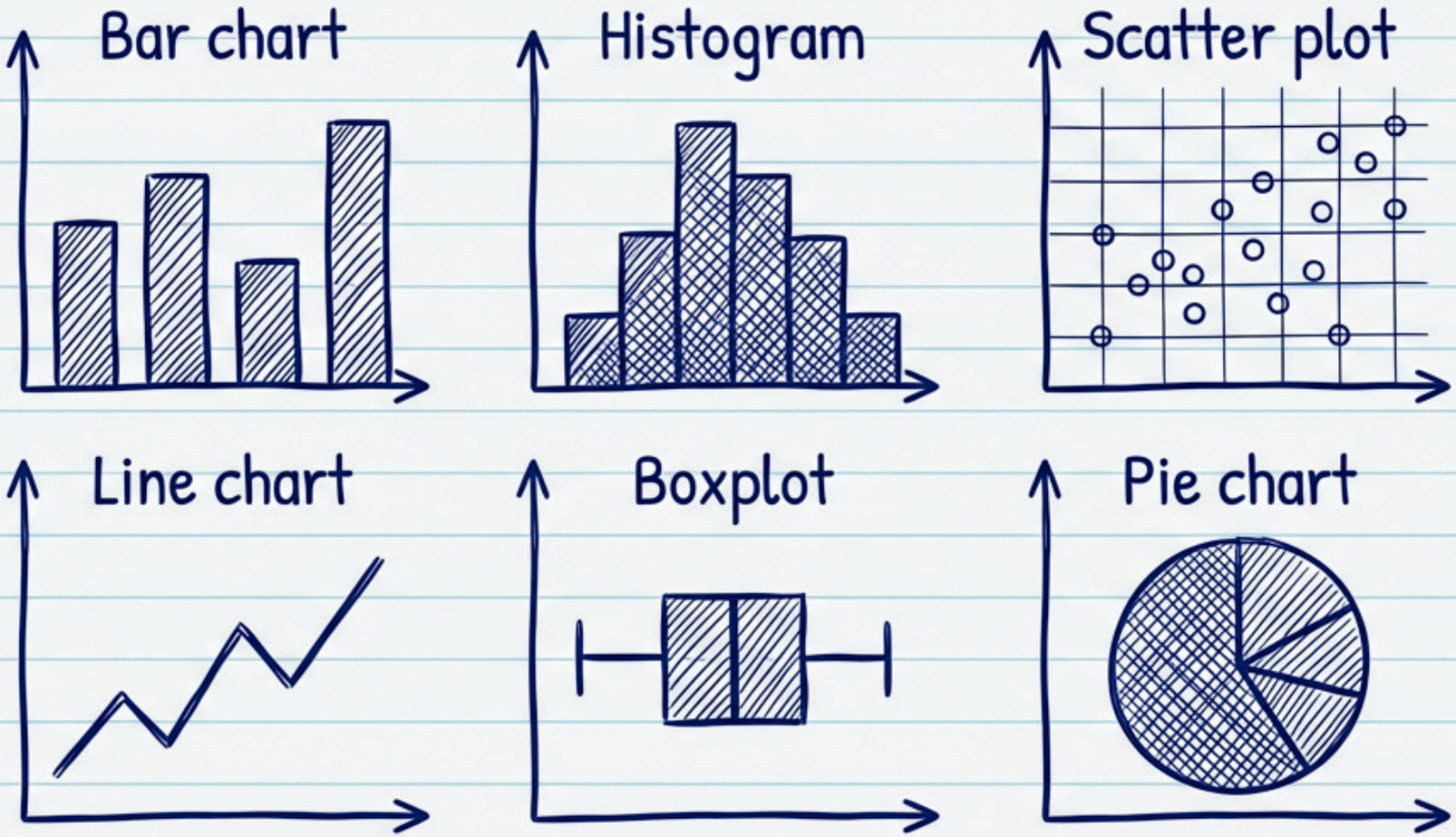
Histogram Formula:

$$\text{Frequency density} = \frac{\text{Class frequency}}{\text{Class width}}$$

Tools of Data Science



Visual Summary



End of Notes - CSA 7109T