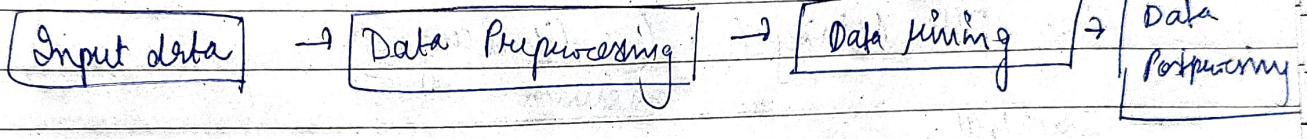
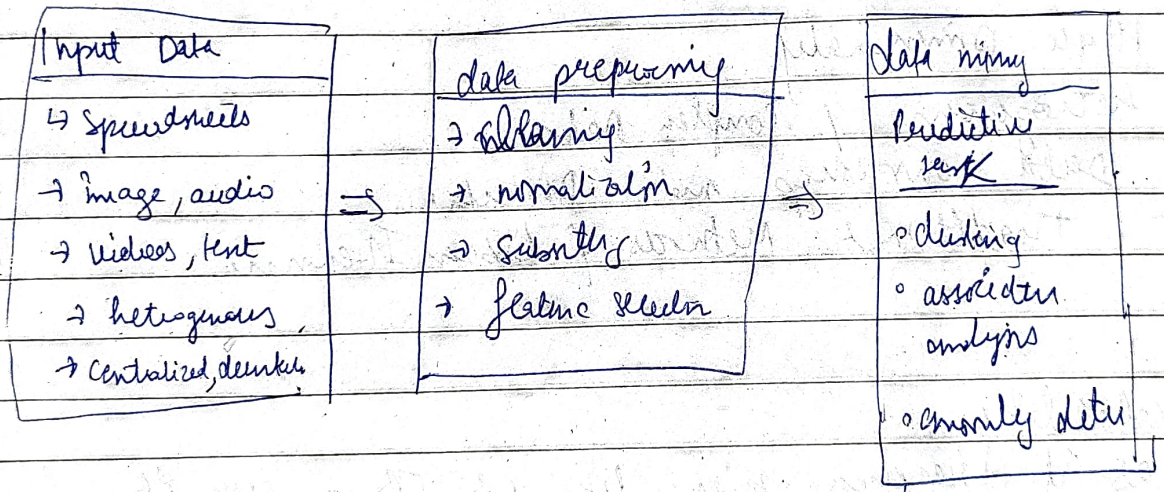
DT. ☐☐ ☐☐ ☐☐☐☐

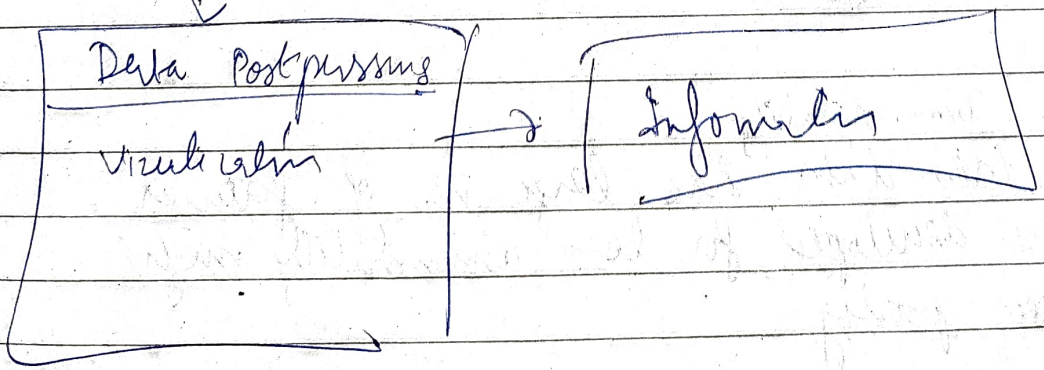
1-1

What is Data Mining?

- process of automatically discovering useful info from large data repositories.
- it is an integral part of knowledge Discovery in databases (KDD) which is overall the process of converting raw information into useful info.

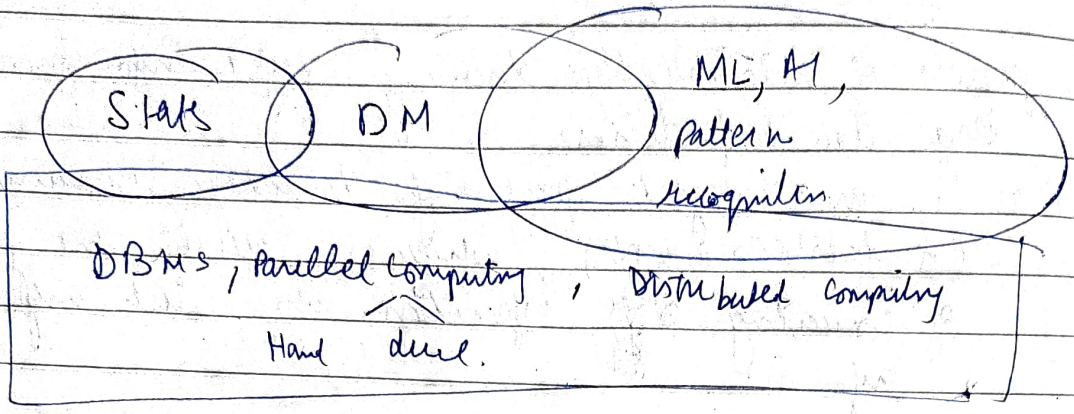






DT. ☐☐ ☐☐ ☐☐☐☐

L3 Origin



L2 Why Data Mining? (Challenges)

- Scalability
- High Dimensionality
- Heterogenous / Complex Data
- Data ownership and Distribution
- Traditional methods / Lasson Interning

① Scalability
as it reaches sizes like GB, TB or even PB, so algo's must be scalable enough to handle these volumes.

② High Dimensionality
as new data has large no. of features. Algo is developed for low dimensionality might perform poorly.

③ Heterogenous and complex Data
Traditional data deals with large no. of features. ~~Algo developed for low dimensionality~~

Spiral

DT. ☐☐ ☐☐ ☐☐☐☐

/ attributes of same types (continuous/categorical) so techniques must be able to handle heterogenous attributes and complex objects.

(4) Data Ownership and distribution
necessary data may not be stored centrally / owned by one organization, requiring development of distributed DM techniques.

(5) Traditional Methods.
Instead of starting with one hypothesis we let data suggest thousands of pattern (rules, clusters, correlations)

1.4 Data Mining Tasks

Predictive
↓
To predict the value of a particular attribute (target / dependent value) based on values of other attributes

Descriptive tasks
↓
Objective is to derive pattern that summarize underlying relationship in data.

attribute to be predicted
↓
target / dependent variable

attributes used for making prediction
↓
explanatory / independent variable

Spiral

DT.

4 core data mining tasks.

- Predictive modeling
  - Classification
  - Regression
- anomaly detection
- association analysis
- Clustering

Predictive modelling

building a mathematical / statistical model that can predict an unknown (target) variable based on known values (explanatory variables)

target variable → outcome we want to predict (Yes / No)

explanatory variables → the factors / inputs that influence (predictors, features) the target.

It is of 2 types.

- Classification : target variable is discrete
  for eg → spam / not spam
  species of a flower.
- Regression : target variable is continuous (numerical values)
  for eg → predicting house price

DT. ☐☐ ☐☐ ☐☐☐☐

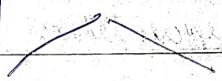
Association analysis
discovers co-occurrence patterns among items or features

expressed $A \Rightarrow B$

Cluster analysis
groups observations into clusters so that points within a cluster are more similar to each other than to points in other clusters

Anomaly (Outlier) detection
identifies observations that are significantly diff from the majority (rare / abnormal patterns)
ex: credit card fraud detection

anomaly detection



Supervised
↓
when we do have labels (fraud vs legit, attack vs normal)

Unsupervised
↓
when we don't have reliable labels for anomalies

goal / data required	supervised learning predict a known label labeled data	unsupervised learning discover hidden structure unlabeled data
task	classification, regression	clustering, association analysis, dimensionality red, density estimation

Spiral

DT. ☐☐ ☐☐ ☐☐☐☐

1) diff b/w DM & KDD.

2) what is unsupervised learning, explain it with the help of an example

3) [illegible]

---

2) (1) Data

- Quantitative (numeric)
- Qualitative (categorical)

spatial data — Time series (sequence) — Text, Image.

(2) Data Quality (how accurate, complete, consistent, representative the data is)

Common issues

Noise (random error)
Outliers (rare/extreme observations)
Missing data (blank / placeholders like 0 / "missing")
Inconsistency/duplicates (conflicting / repeated rows)
Bias/unrepresentative (sample doesn't reflect the population)

(3) Preprocessing (to make data suitable for mining)

↓

Transforming (reshaping data to fit a technique and improve results

Spiral

DT. ☐☐ ☐☐ ☐☐☐☐

Scaling (z-score)

$$x' = \frac{(x - \bar{x})}{s_x}$$

Min max normalization

$[0, 1]$.

Discretization
↓
turn numeric into bins (like low / med / high)

Binarization / encoding (categorical into binary)

Aggregation / sampling (fewer rows for speed, higher level summaries)

Dimensionality reduction

4) Analyze via relationships
↓
replace raw objects with their pairwise proximity and analyze those.

DT. □□ □□ □□□□

2.1 Types of Data.

Data object : a single record / case / row. / instance / vector / sample.

attribute = column / feature / dimension.

- Attribute.
  ~~characteristic~~ characteristic of an object that can vary across objects over time.
  Ex: eye color
  temperature.

- Measurement scale
  a rule that maps the true property to a numeric / symbolic value.

Attribute type

- Nominal (categorical) → values are names / labels
- Ordinal (categorical) → values have order / rank
- Interval (numeric) → differences play an imp role
- Ratio (numeric) → diff and ratios are imp.

DT. ☐☐ ☐☐ ☐☐☐

Discrete vs Continuous Vs Binary.

Discrete → finite / countable values
Binary → Special case of discrete with two values
Continuous → any real value

Asymmetric Attributes and Sparsity

Asymmetric attribute
only presence (non-zero) is informative; absence (0) is less meaningful.

Sparse data.
most values are 0.
attributes are of same type and asymmetric.
non zero are considered imp

Resolution
Data can be collected at multiple levels (fine vs coarse)
↙ ↘
more detail — Smoothing but
but more noise — may hide patterns.

If resolution is too fine then pattern may not be visible as may be ignored into noise whereas. If pattern is too coarse then it may disappear.

Spiral

DT. ☐☐ ☐☐ ☐☐☐☐

Types of Datasets

General characteristics of Dataset that has significant impact on DM.

1) Dimensionality

↓

Curse of Dimensionality

It happens when we have too many features (dimensions) in our data.

for eg 1D graph: data is pretty close

2D graph: same no. of points feel more spread out

3D: they are farther apart.

100 dimens: everything feels equally far from everything else.

2) sparsity It saves computational time & storage resources

3) Resolution.

DT. ☐☐ ☐☐ ☐☐☐

Types of Dataset

Record Data	Graph based	Order Data
• Market Basket • Data Matrix • Sparse Data Matrix.	• Relationship among objects • Objects that are graph	• Sequential / Temporal Data • Sequence. • Time series • Spatial

Record Data

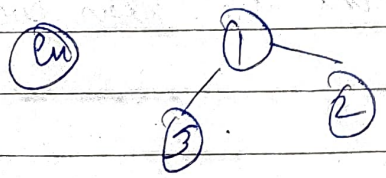
- fixed scheme; each record has same attributes
- stored in flat files

- Market Based
  each record is a set of items bought together

- Data Matrix / Pattern Matrix
  Rows = objects
  Columns = numeric attributes

here same fixed number of numeric attribute

- Sparse Data Matrix
  attributes are of same type



data objects are mapped to the nodes of the graph.

Relationship among the objects are captured by the links b/w objects and their properties such as directions and weights

Spiral

DT. ☐☐ ☐☐ ☐☐☐☐

Graph based :

relationship among objects
nodes = objects, edges = links.

Objects that are graphs
each object has internal structure (atoms/bonds)

Ordered Data :

Sequential / Temporal data

- there is no timestamp but there are positions in an ordered sequence
- ordered symbols without sequences (DNA sequences)

(ex)
- sequence of words or letters.
- sequence of molecules in a genetic seq data. could infer if a person is healthy or not.

time series :

- measurements over time (stock prices, monthly temp)
- here a series of measurment are taken over time.

(ex) temp during diff years.

DT. □□ □□ □□□□

Temporal autocorrelation

If two measurements are close in time then values are often similar to each other.

Spatial

- measurement by location (and time)
- objects have spatial attributes such as point/ areas

eg: weather data collected for diff geographical locations

Spatial autocorrelation
objects that are physically closed tend to be similar in other ways as well.

handling non-record data

non record data is when order or relationship is inherent

General characteristics after analysis

- dimensionality : no. of attributes
  - curse of dimensionality → motivates dimensionality reduction
- distribution : frequency / concentration of values
- resolution : level of detail, affects visibility of pattern and noise.

Spiral

DT. ☐☐ ☐☐ ☐☐☐☐

2.2 Data Quality

DM uses data collected for other purposes
We must detect / correct issues

measurement and data collection issues
↳ as values might differ from true values

- Measurment Error
  recorded value ≠ true value.
- Data collection error
  missing / extra objects / inappropriately including some data objects

Spurious

Occurs in Spatial (by location) or Temporal (timestamp) data

Artifact

deterministic distortion of the data.
(ex) 8 traces in the same place in a set of pics.

Noise

random error (sensor fluctuations)
Hard to fully remove.

DT. ☐☐ ☐☐ ☐☐☐☐

Precision

- closeness of repeated measurements (to each other) of same qty.
- it is measured by Standard deviation

$$s = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n-1}}$$

Bias

- it is diff b/w the mean of the set of values and known value of qty being measured.
- systematic deviation from the true value

Accuracy

- closeness of measurement to the true value by using significant digit.
- Goal is to use as many digits to represent the result as are justified by precision of data.

Mass of 1 g in weighing 5 times.

{ 1.015 , 0.9901 , 0.001 , 0.996 , 0.131 }

Bias 0.399

Mean = 0.601.

Outliers / Anomalous objects

- objects / values that are v different from the bulk.
- values that are unusual / considerably different from the typical values of that attribute
- they are not the same as noise; they can be legitimate and interesting.

Spiral

DT. □□ □□ □□□□

Missing values

- Info is not collected because people decline to share
- that attribute may not be applicable to all the cases.

Handling missing values. (Viva) ?

- Eliminate the data object / attribute.
- estimate missing values
- ignore the missing values

Inconsistent values

conflicting entries

ex: - zip code of a city is correct but it is not appropriate for a particular location.
- height of a person cannot be negative.

⇒ correction of inconsistency requires additional info from a master database.

Duplicate data.

↳ same object appearing multiple times

- Care must be taken while merging data from heterogeneous sources
- these could be two objects that represent a single object

Spiral

DT. ☐☐ ☐☐ ☐☐☐☐

- avoid accidentally combining data objects that are similar but not duplicates

eg. deduplication

↓

It requires solving attribute discrepancies while avoiding merging distinct entities with similar info.

Issues related to the application

- timeliness
  data becomes stale, refresh required.
- relevance.
  data must contain relevant variables needed for the goal.
- sampling Bias
  sample doesn't represent the whole population
- documentation / knowledge about data.
  a good data dictionary prevents serious mistakes

↓

(types, units, encoding, missing-value codes, known relationships, provenance)

the quality of data

Spiral

DT. ☐☐ ☐☐ ☐☐☐☐

2.3 Data Preprocessing

- aggregation
- sampling
- dimensionality reduction
- feature subset reduction
- feature creation
- discretion and binarization
- variable transformation

Aggregation

Combining two or more objects into single attribute (object)

purpose

- data reduction : reduction in no. of attributes/objects
- change of scale : monthly view of data instead of a daily / per item view.
- more stable data : aggregated data tends to have less variability.

advantages

- Smaller datasets after data reduction requires less memory & processing time and hence aggregation may permit the use of more expensive data mining algorithms
- aggregation can act as change of scope/scale by providing high level view instead of low level view.

Spiral

DT. ☐☐ ☐☐ ☐☐

disadvantage
- potential loss of interesting details

Sampling
- analyze a representative subset instead of whole data.
- using sample can work almost as well as using the entire data if the sample is representative
- it is used for DM because processing the entire data is too expensive and time consuming

Sampling
- simple random
  - with replacement
  - without replacement
- stratified
- progressive.

Simple random sampling (equal probability of selecting any particular item)
- sampling without replacement
  - as each item is selected it is removed from the set of all objects that constitute population.
- sampling with replacement
  - objects aren't removed from population as they are selected from sample
  - same object can be picked more than once
  - simpler to analyze

Spiral

DT. ☐☐ ☐☐ ☐☐☐☐

Disadvantage of sampling (simple random)

when population consist of different types of objects with widely different no. of objects, simple random sampling can fail to adequately represent those types of objects that are less frequent.

Solution → stratified sampling

- starts with pre specified group of objects.
- equal no of objects are drawn from each group even though groups are of diff sizes
- the number of objects drawn from each group is proportional to the size of that group.

Progressive / adaptive sampling

- proper sample size can be difficult to determine
- these approaches start with a small sample and the increase the sample size until a sample of sufficient size has been obtained
- while this advantage eliminates the need to determine the correct sample size initially it requires that there be a way to evaluate the sample to judge if it is large enough.

DT. ☐☐ ☐☐ ☐☐☐☐

Dimensionality reduction

↓

make data fewer features by projecting data into a lower-dimensional space

advantage

- avoids curse of dimensionality
- eliminate irrelevant features and reduces noise
- reduce amount of time and memory required by data mining
- allows data to get visualized easily
- allow data to be more understandable

Techniques

↓

- by creating new attributes that are a combination of old attributes
- by selecting new attributes that are a subset of old is known as feature subset selection / feature selection

Spiral

DT. □□ □□ □□□□

Feature subset selection

Keep only useful original features
(drop irrelevant ones)

Techniques

- Some irrelevant & redundant attributes can be removed by using domain knowledge
- Try all possible features as input to data mining algorithm.

Standard approaches

- embedded approaches
  - feature selection occurs naturally as a part of the data mining algorithm
  - algo itself decides which attribute to use and which to ignore

- filter approaches

  features are selected before data mining algo is run.

  use select general criteria

- Wrapper

  search subset using final model

DT. ☐☐ ☐☐ ☐☐☐☐

Feature creation

↓

Creating a new attribute from existing ones

Why? raw inputs are mostly poor shaped

Methods

1) feature extraction (domain transforms)
2) map data to a new space
3) feature constructs

feature extraction : summarizing messy raw data into useful signals

[in en] → we turn pixel into images
words into sentences → text
Raw sound to audio

map data to a new space

↓

transform data to reveal hidden patterns.

fourier transform (time → frequency) wavelet transform (patterns over time)

feature construction

↓

use simple formulas on existing features to make better ones

DT. □□ □□ □□□□

Discretization and Binarization

Discretization

transforming continuous attributes to categorical attributes.

Steps
1) Choose split points
2) map any value to its interval label.

unsupervised (no labels)	supervised (use labels)
◦ class info is not used	◦ class info is used
◦ equal width → $w = \frac{(max - min)}{k \text{ (bins)}}$	◦ pure intervals, entropy based
◦ equal frequency → $\frac{n}{k}$ (bin no.s)	
◦ clustering	

Data → 0, 4, 12, 16, 16, 18, 24, 26, 28.

Width = $\frac{max - min}{k}$.

$\frac{28 - 0}{3} = 9.333 \equiv (10)$

here no of bins is 3 [0-10]
[10,20]
[20,30]

$\frac{n}{k}$ eq freq = $\frac{9}{3} = (3)$

Spiral

DT. □□ □□ □□□□

Data : 2, 4, 6, 7, 10, 12, 15, 18, 20, 28

$$\frac{25-2}{3} = \frac{23}{3} = 7.66$$

Binarization

- Conversion into binary
- binarize categorical attributes

Awful	0	0	0	0
Poor	1	0	0	1
OK	2	0	1	0
Good	3	0	1	1
Great	4	1	0	0

Standardization

mean = 0 SD = 1 (σ^2)

If diff variables are to be continued in some way then it's necessary to avoid having a variable with large values.

Min – Max Normalization

$$v' = \frac{v - min}{max - min} (new\ max - new\ min) + new\ min$$

Spiral

DT. ☐☐ ☐☐ ☐☐☐☐

marks	
8	0
10	$\frac{10-8}{20-8} = \frac{2}{12} = 0.16$
15	$\frac{7}{12} = 0.38$
20	$\frac{20-8}{20-8} = 1$

Max = 20 Min = 8

2 Scale normalization

$$\frac{\sum v_i}{n}$$

$$v' = \left(\frac{v-\mu}{\sigma}\right) = \frac{\sum v_i^2}{n}$$

$$\rightarrow \frac{\sqrt{\sum (v_i - \mu)^2}}{n}$$

DT. ☐☐ ☐☐ ☐☐☐☐

Measures of similarity and Dissimilarity

→ Similarity b/w two objects is the numerical measure and is usually non negative. It is often b/w 0 and 1 where

0 → no similarity
1 → high similarity

→ Dissimilarity is the numerical measure of how the objects are diff from each other
- It is lower for more similar objects.
- Distance is used to find it.

Minkowski Distance.

$$d(x, y) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{\frac{1}{r}}$$

n = no. of dimensions
r = no. of parameters

if r = 1,

L1 (Manhattan) = sum $|x_k - y_k|$

if r = 2

L2 Euclidean = $\sqrt{\text{sum}} (x_k - y_k)^2$

if r = ∞,

L∞ Supremum = max $|x_k - y_k|$

Spiral

DT. □□ □□ □□□□

	x	y
P_1	0	2
P_2	2	0
P_3	3	1
P_4	5	1

$(5-2) + (1-0)$

L_2 $\sqrt{(5-2)^2 + (1-0)^2}$ $\sqrt{10}$