Aug 20

Data Mining Pipeline

- Get ~~eff~~ data efficiently, lower latency, higher throughput
- Quality of data.
- System Design – to get data efficiently.
- Algorithms
- Interdisciplinary Collaboration.

Four Views

Data Mining – Knowledge discovery from data.

Huge amount of data – scalability, efficiency

Data
Application
Knowledge
Technique.

Data mining tasks.
Integrated.

Data View – 3Vs, 4Vs, 5Vs,

3V – Volume, Variety, Velocity (dynamics of data)

4V – Veracity (quality, trustworthy)

5V – Value

Data View – Relational, transactional data.
eg – student records, bank accounts, store purchases.

Sequential, temporal, streaming data.
eg – gene sequences, stock prices, sensor reading

* Spatial, spatial-temporal data.
  eg- land use, bird migration, traffic condition

* Text, multimedia, web data.
  eg- News article, audio/video/image, hypertext

* Graph, Network data.
  eg. social network, power grid, co-authorship.

So, Data is Single or Mixture of multiple data types

* Application View

* Market Analysis, target advertisement
  eg- customer profiling, product recommendation

* Healthcare, Medical Research
  eg- disease diagonosis, patient care, drug discovery

* Science and engineering.
  eg. air pollution, marine life, electric vehicles.

* Security.
  eg. surveillance, intrusion/crime, fraud, cyberattack.

* Govt. non profit.
  eg. urban planning, traffic control, education.

Knowledge View

- Frequent pattern, association, correlations
  eg. songs listened together or in certain sequence
  eg. A is (more/less) likely to happen given B.

- Categorization.
  eg. similarity among users within certain purchases
  eg. differences between two patient groups.

- Anomaly, Outliers
  eg. sensor errors, fraud activities, extreme events

- changes over time
  eg. emerging new patterns, shift of user interest

Descriptive - describe, how it looks like
Predictive - make predictions.
prescriptive - recommendation, prescribe strategy.

Technique View

- Frequent pattern Analysis
- Classification, prediction
- Clustering
- Anomaly detection
- Trend and evolution analysis.

Frequent ~~Item~~ Pattern Analysis

- ~~Fe~~ Frequent Itemset

- Frequent sequence
- Frequent structure
- Association rules
- Correlation analysis.

Classification

- predefined classes.
- Need training data
- Build model to distinguish

Prediction.

- Numerical prediction. e.g. weather, stock price. traffic.
  (continuous value)

clustering.

- ~~Pre def~~ No predefined classes
- Intra-cluster similarity
- Inter-cluster dissimilarity.

Anomaly Detection

- differ from the Norm
- error, noise
- fraud.
- extreme events.

Trend and Evolution Analysis

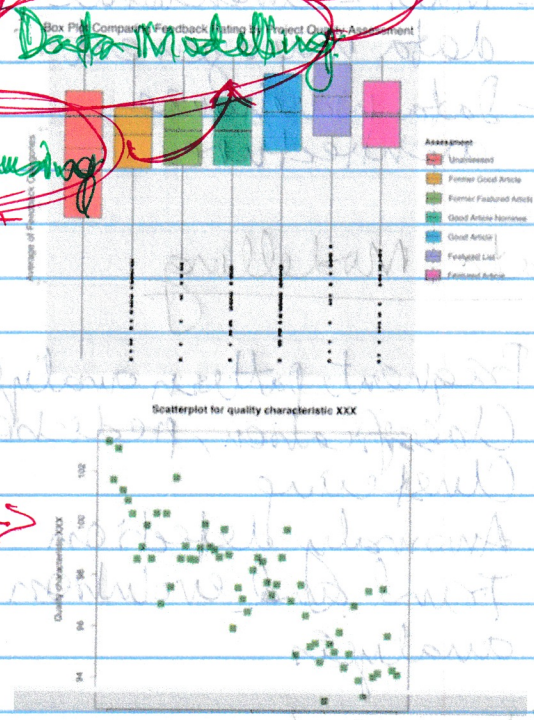
- Changes over time
- overall trend
- periodical patterns
- Anomalies.

Data Mining Pipeline

Key Components of Data Mining pipeline.

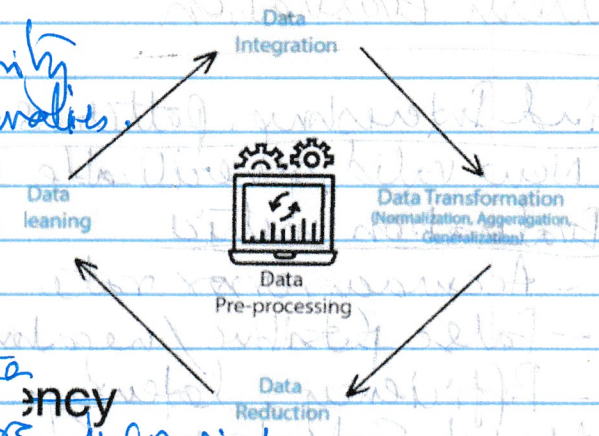
4 - Data modeling
1 - Data understanding
2 - Data preprocessing
3 - Data warehousing - data managing.
5 - Pattern evaluation.

Not in order



Data Understanding

- What type of data?
- What do they look like?
- Statistics & Visualization
- Similarity and dissimilarity
- General patterns vs. anomalies.



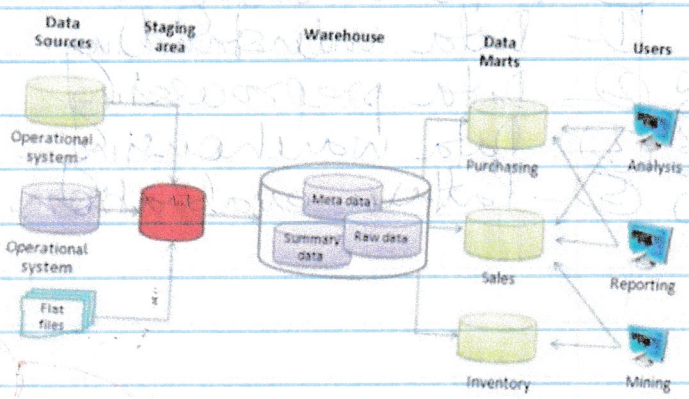
Data preprocessing.

- Potential Issues with data
  eg. missing data, errors, inconsistancy
- Preparing data for the mining process.
  - cleaning, integration, transformation, reduction
    combine data sources

→ "No good data, no good data mining"

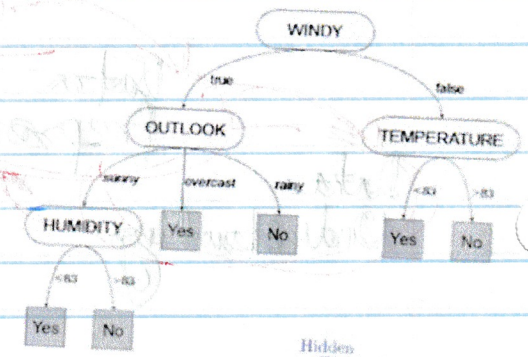
~~Data preprocessing~~.

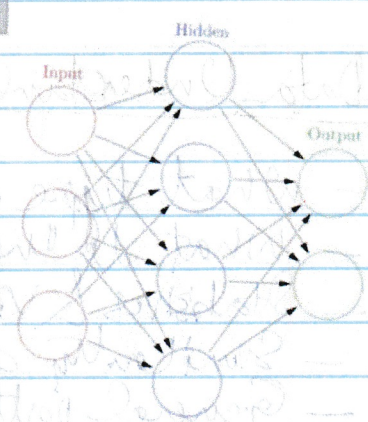
Data Warehousing



- Data warehousing vs operational data ~~missing~~
- Data Cube & OLAP
  - Multi-dimensional data management
  - Data warehouse architecture

Data Modelling

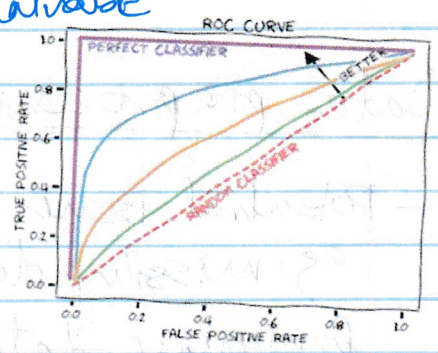




- Frequent pattern analysis
- Classification, prediction
- Clustering
- Anomaly detection
- Trend and evolution analysis

Pattern Evaluation

- Find Interesting patterns from data
- New valid, generalizable useful, explainable
- Evaluation metrics
  - Accuracy, error rate
  - False positive / negative rate
  - Efficiency, latency
- Model Selection



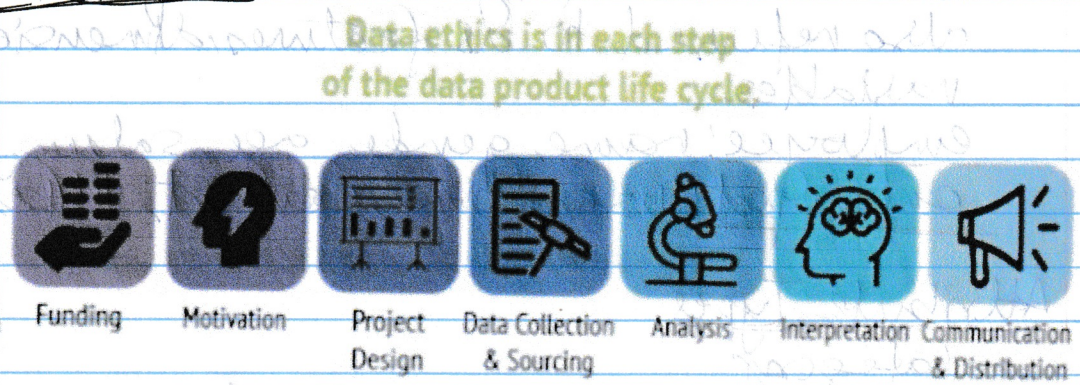
Data Mining Examples

- Business Intelligence
  - customers, products, logistics, promotion, fraud
- Cyberspace
  - service providers, online social media, security

Data Mining: Major Issues, Ethics, Resources.

Veracity –
- → Diverse data ⇒ diverse knowledge
- → Data Quality Issues.
- → Supervised vs unsupervised learning
- → Performance evaluation
- → Effectiveness vs efficiency
- → Incremental, interactive mining.
- → Integration of domain knowledge.
- → Visual analytics.
- → Privacy-preserving mining.

Data Ethics



- → Data Ownership
- → Privacy, anonymity
- → Data and model validity.
- → Data and model bias (algorithmic fairness)

→ Interpretation, application, societal consequence

Data Mining Resources

- https://www.kdd.org/

MODULE 2 – DATA UNDERSTANDING

Data Objects, Attributes, Statistics

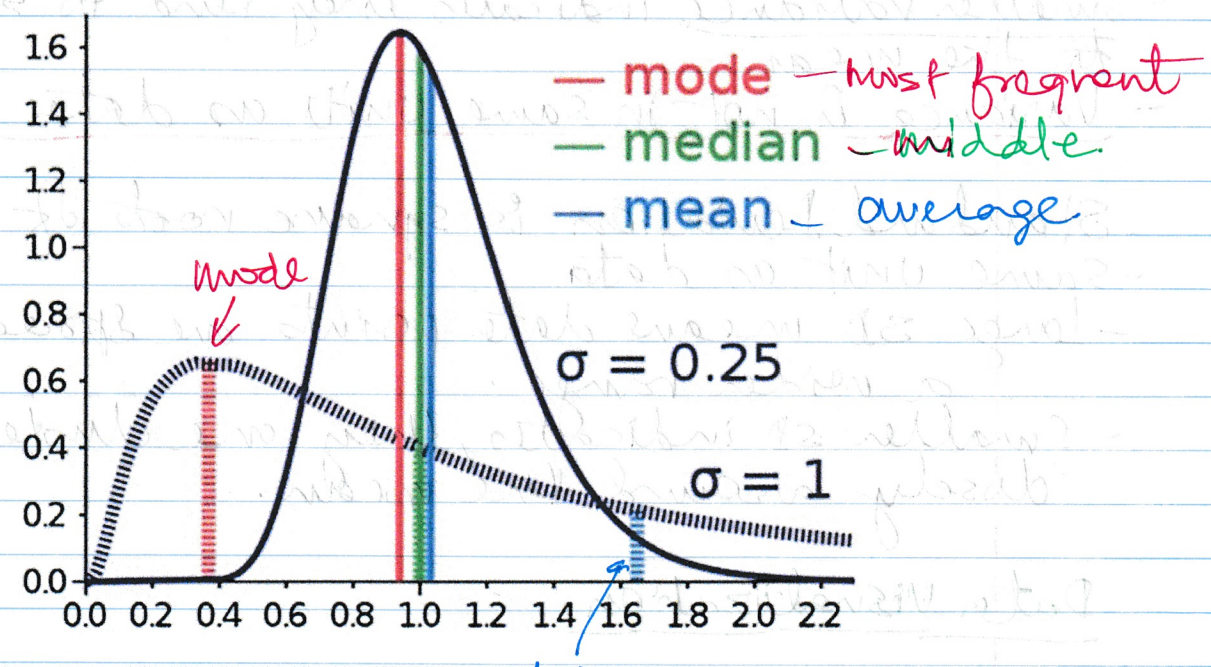
Data Understanding
- Data objects and attributes
- Data statistics
- Data visualization
- Data similarity

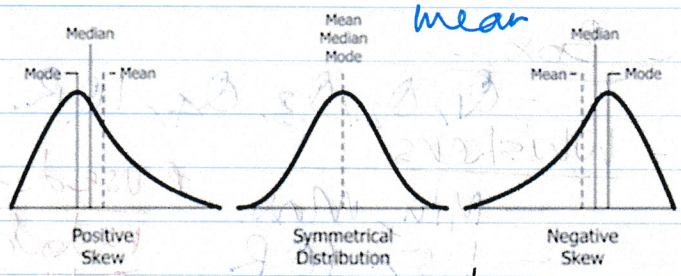
Dataset
- Collection of data objects.
  e.g. employee records, product catalog, online post
- each described by a number of attributes
  also referred to as features, dimensions
  variables.
  employee: name, gender, age, salary, job title
  online post: user, time, content, #likes, responses

Attribute Types
- Category
  - Nominal: binary, ordinal.
    - major, academic ~~mon~~ grades.
- Numeric: discrete or continuous.
  - Interval scale or ratio-scaled
    year 2000, number of users, annual income.

Data statistics

- # of objects, # of attributes.
- Distribution of each attributes values.
  - Categorical: % of each value.
  - Numerical: central tendency, dispersion.
- Comparision across attributes & datasets.
- Central tendency
  - Mean
  - Median
  - Mode
  - Midrange. - (Max − Min)/2



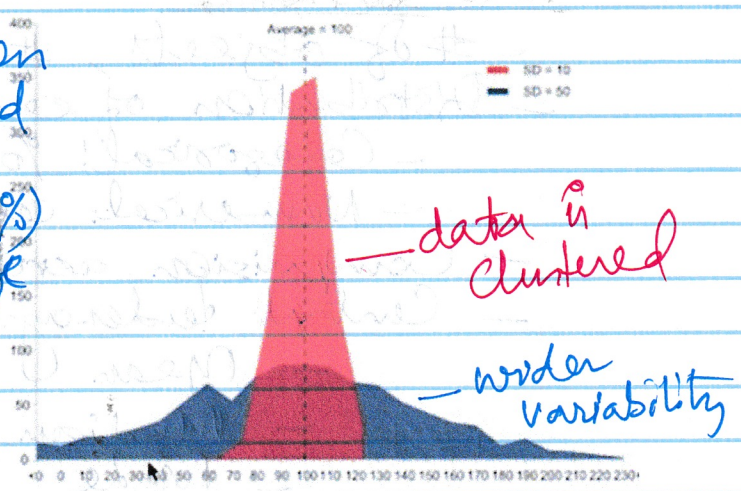


Right skewed
Mean is right of Median

left skewed
Mean is left of Median

Dispersion

- How much a distribution is stretched or squeezed
- Range: max - min
- Quartiles: Q_1 (25%), Q_3 (75%)
- IQR (interquartile range
- Variance
- Standard Deviation.



← data spread →

- Larger Variance means data points are farther from the mean.
- Smaller variance indicates they tend to be closer to the mean.
- Variance is not in same units as data

- Standard Deviation is square root of variance.
- Same unit as data
- large SD means data points are spread over a wider range.
- Smaller SD indicates, they are clustered closely around the mean.

Data Visualization

Boxplot
- Box
  - Q_1, Q_2, Q_3, Q_4, IQR.
- Whiskers
  - Min, Max,
  - $1.5 \times$ IQR
- Outliers.

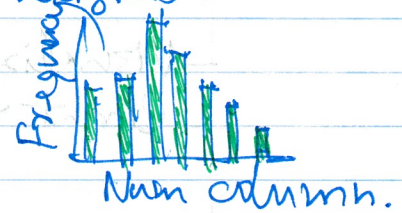
* Used to find spread of Num variable

* used to ~~comp~~ find relation between Cat and Num variable.

* Outliers or extreme values.

Histogram - Continuous variable

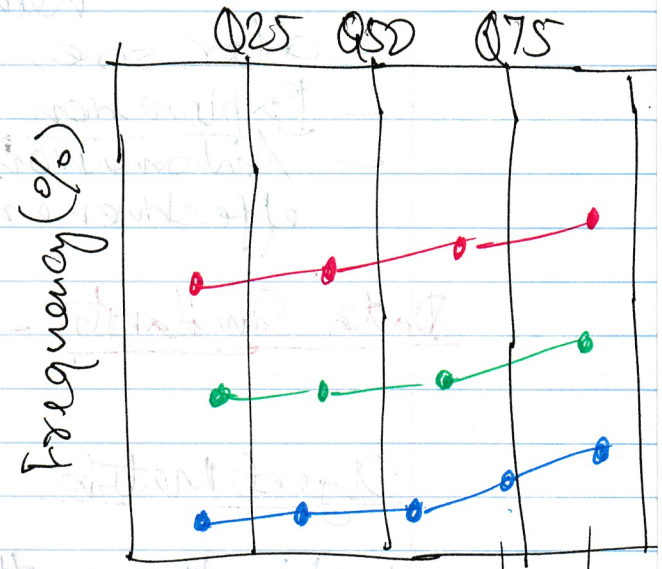
- tells the distribution, bars heights.



Quantile Plot

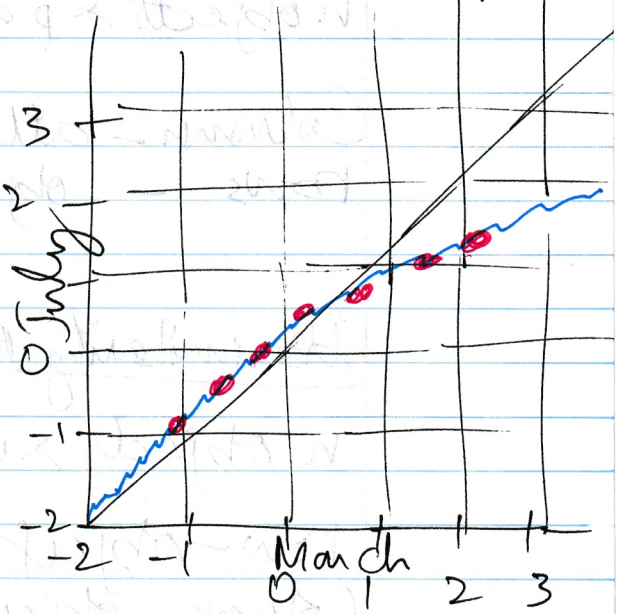
- Quantile: percent of points below the given value.
  X: percent
  Y: quantile

  seems similar to boxplot
  Cat var + num var.



Quantile-Quantile (Q-Q) Plot

- Comparision of two quantiles
- 45° reference line.



Scatter Plot.

- between two attributes X and Y to find relationship.

- Pairwise plot - pairwise comparision across multiple attributes.

Data Visualization Methods

- Visualizing complex data and relationships
- Chart type: line, pie, stacked bar, bubble, area, heatmap, word cloud, network.
- color, size, layout hierarchy.
- Exploration vs explanation
- Information, interaction, efficiency & effectiveness.

Data Similarity – Understand how objects are similar or dissimilar.

Object Matrix

n objects × p attributes

Column – attributes
rows – objects

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Dissimilarity Matrix

n objects × n objects

row – object
Column – object

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ & & & 0 & \\ d(n,1) & d(n,2) & \cdots & & 0 \end{bmatrix}$$

Nominal Attributes - are category labels No order

	Car1	Car2
Type	SUV	Sedan
Make	Honda	Honda
Model	CRV	Civic
Color	White	Silver.
Condition	Good	Good

Similarity

$s=1$, if $x=y$; otherwise $s=0$

Dissimilarity.

$d=0$ if $x=y$; otherwise $d=1$

Customized.

Eg. Color: White is more similar to silver than red.

Binary Attributes - takes two possible values.

	Item1	Item2	Item3	Gender	Cough	Fever.
UserA	Y	Y	N	F	N	N
UserB	Y	N	N	M	Y	N.

- Symmetric: ie equal chance of Y or N.
  - Same calculation as nominal attributes
- Asymmetric: ie. Y is less likely than N.
  - eg - items purchased or medical symptoms

Matrix with 4 possible scenarios

Users ⇒ A and B

	B=Y	B=N	Sum
A=Y	q	r	q+r
A=N	s	t	s+t
Sum	q+s	r+t	q+r+s+t.

q, r, s, t are scenarios.

→ For symmetric variables.

dissimilarity ⇒ $d(i,j) = \frac{r+s}{q+r+s+t}$ ⇒ cases they are different ⇒ total cases

→ For Asymmetric variables.

$$d(i,j) = \frac{r+s}{q+r+s}$$ t is ignored

* Jaccard Coefficient – is widely used by similarity measure.

$$\boxed{Sim(i,j) = \frac{q}{q+r+s} = 1 - d(i,j)}$$

Similarity for Ordinal, Numeric, Mixed Attributes

Ordinal Attributes – have the order.
eg – Gold, Silver, Bronze

→ Map them to the order/ranks.

$$r_{ij} \in \{1, \ldots M_f\}$$

⇒ (1, 2, 3)

⇒ Normalize the rank.

Map to [0, 1]. ⇒ (0, 0.5, 1.0) $z_{if} = \frac{r_{if}-1}{M_f-1}$

Ordinal Attributes are categorical variables, they have meaningful order, but the difference between them isn't necessarily uniform.

example:

Education Level: High school < Bachelor's < Master's < PhD.

Customer Satisfaction: Very Unsatisfied < Unsatisfied < Neutral < Satisfied < Very Satisfied.

~~Python~~! Things to avoid:

- Avoid one-hot encoding. That will destroy the order. Use label encoding or rank mapping instead.

How to Map to Ranks

```
ordinal_map = {
    'Very Unsatisfied': 1,
    'Unsatisfied': 2,
    'Neutral': 3,
    'Satisfied': 4,
    'Very Satisfied': 5
}

df['satisfaction_rank'] = df['satisfaction'].map(ordinal_map)
```

Use OrdinalEncoder from scikit-learn

```
from sklearn.preprocessing import OrdinalEncoder

encoder = OrdinalEncoder(categories=[['Very Unsatisfied',
'Unsatisfied', 'Neutral', 'Satisfied', 'Very Satisfied']])
df['satisfaction_encoded'] =
encoder.fit_transform(df[['satisfaction']])
```

⇒ Dissimilarity between the mapped values.

Numeric Object Dissimilarity

⇒ Usually measured by distance
⇒ Minkowski distance (l_p norm)

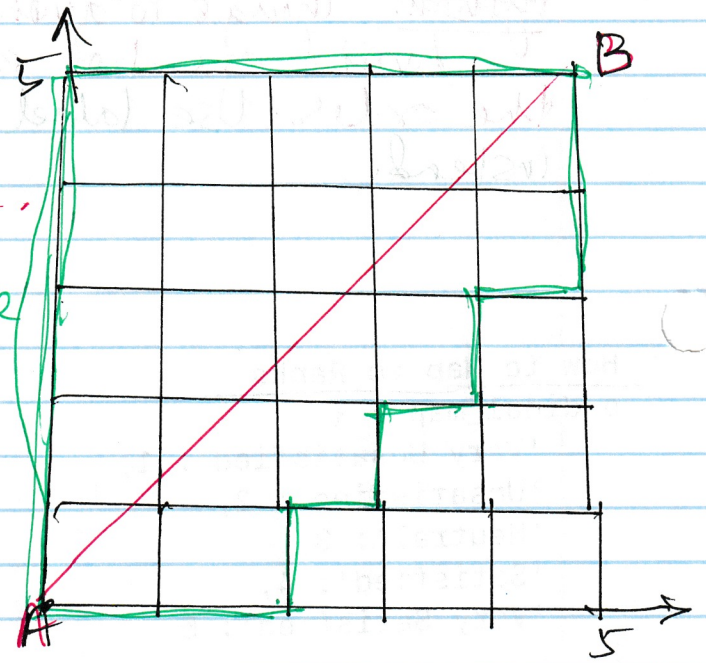
$$d(i,j) = \left(|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \dots + |x_{in} - x_{jn}|^p\right)^{1/p}$$

$p = 1$: Manhattan distance
$p = 2$: Euclidean distance

Distance Measure

— Euclidean Distance.

= Manhattan Distance



Cosine Similarity

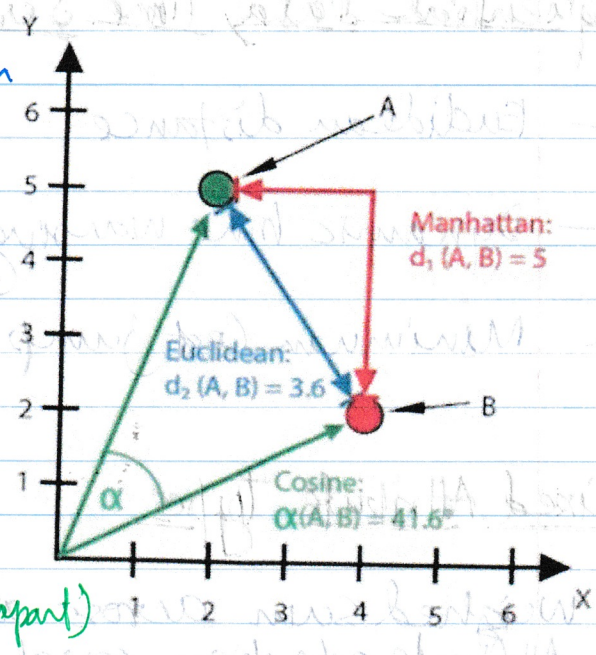
eg. text documents:
- Frequency of word occurrence
- High dimensional, sparse

- Angular Similarity of vectors: Inner product.

$$\cos\theta = \cos(A, B) = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum A_i B_i}{\sqrt{\sum A_i^2} \sqrt{\sum B_i^2}}$$

Cosine Similarity is a metric used to measure how similar two vectors are, based on the angle between them, rather than the magnitude.

Very useful in high dimensional spaces like text data, where direction of the data matters more.



if they point in same direction. $\cos 0 = 1$

if they are orthogonal (90° apart) $\cos 90 = 0$.

if they are in opposite direction $\cos 180 = -1$. (completely dissimilar)

Cosine Similarity Example

	Game	Basketball	Player	Injury	Won
Document 1	3	5	2	1	2
Document 2	1	3	4	2	1

Step 1. Recall the formula.

$$\cos\theta = \cos(A, B) = \frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\sum A_i B_i}{\sqrt{\sum A_i^2}\sqrt{\sum B_i^2}}$$

$$D_1 \cdot D_2 = 3\times1 + 5\times3 + 2\times4 + 1\times2 + 2\times1 = 30$$

$$\|D_1\| = \sqrt{\sum D_1^2} = \sqrt{3^2+5^2+2^2+1^2+2^2} = \sqrt{9+25+4+1+4} = \sqrt{43} = 6.55$$

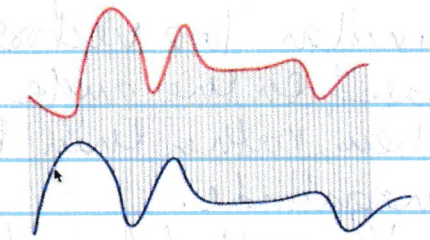
$$\|D_2\| = \sqrt{\sum D_2^2} = \sqrt{1^2+3^2+4^2+2^2+1^2} = \sqrt{1+9+16+4+1} = \sqrt{31} = 5.56$$

$$\cos\theta = \frac{30}{6.55 \times 5.56} = 0.822$$ (measure of similarity)

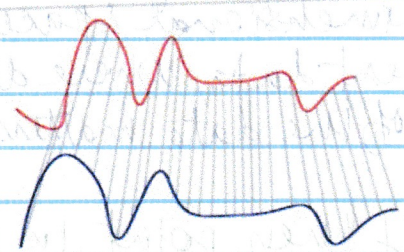
Quite high similarity

Sequential Data, Time Series

- Euclidean distance
- Dynamic time warping.
- Minimum Cost jump.



Euclidean Matching



Dynamic Time Warping Matching

Mixed Attribute Types

- weighed sum across attributes.
- Attribute selection, correlated attributes, normalization weights, missing values.

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

Data Similarity/Dissimilarity

How to choose the right measure?
What kind of data one we talking about
→ Dense, continuous data:
Euclidean, Manhattan Distance.
→ Asymmetric attributes: ignore the null cases.
→ Sparce data: cosine similarity, Jaccard similarity.
→ Subset: seasonal patterns, subgroups.
→ Domain knowledge: types of pattern to learn

Case Study: Online Social Media

* Types of data:
  - user, post, timestamp, reactions, network.
* Statistics, visualization.
  #objects #attributes, distributions, outliers
* Similarity/dissimilarity.
  - Nominal, binary, numeric, text, temporal.

Air Quality Sensing Data

- Size: # of sensors, # users, duration of study, area.
- statistics:
- visualization
- similarity/distance.
- location, temperature, humidity.
- activity, biking, riding bus, running, sleeping.

3. Data Quality Issues and Causes

Data Preprocessing - 3.

* Identify potential issues in the data and apply specific techniques to preprocess your data.
  - missing data, errors, inconsistency availability.
* Preparing data for the mining process

- Data cleaning, integration (multiple data sources), transformation, reduction.

Data Quality

* Get Relevant data
* Accessibility of data
* Interpretability - good metadata
* Reliability - Data sources, quality control process
* Timeliness - how quickly we can get new information.
* Accuracy - any error in data.
* Consistency - for multiple data sources and multiple attributes.
* Precision -
* Granularity - spacial data - data at country city level.
  temporal - minute, hourly. monthly
* Completeness -

Real World Data Issues

* Incomplete - missing values/attributes
* Noisy - Imprecision, errors, outliers age = -10.
* Inconsistant - age vs birthday, rating scale as 1 to 5 with 1 as high, sometimes 5 is high.

Causes of Data Issues

* Data Collection/transmission/processing

* Human, hardware and software
  - Limitations, errors, multiple sources
* Changes over time
  - Updated survey, new sensing capabilities.

Data Cleaning

* Incomplete: Remove or fill in missing values
* Noisy: Smooth noisy data, identify/remove errors/outliers.
* Inconsistent: Resolve

Incomplete Data

- Remove object/attributes
- Manually fill in missing values
- Automated methods - global constant, attribute mean, class mean.
- Estimated value: regression, KNN, probabilistic

Noisy Data

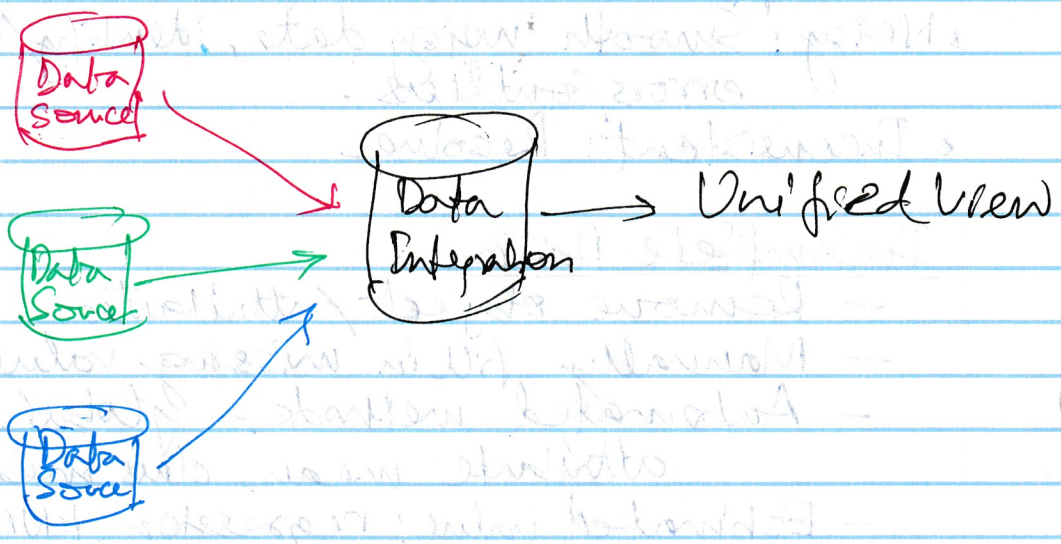
* Regression -
  - Fit data with regression functions
    eg. linear, polynomial, logistic
* Clustering -
  - Group data into clusters
  - Detect and remove outliers

Inconsistent Data

* Semantic-based checking
  - Metadata, attribute relationships
    eg. age vs DOB
* Data understanding
  - statistical analysis, visualization
    eg. scatter plot

Data Integration.

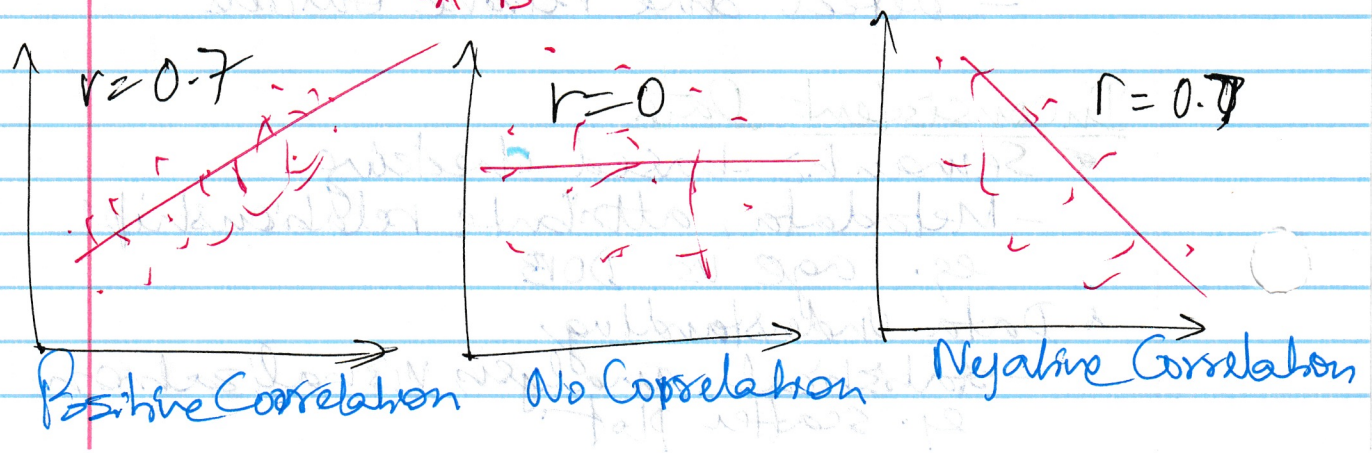
* Combine data from multiple sources.
  - Entity identification
    e.g. users, items
* Redundant data
  e.g. correlation analysis



Correlation Analysis

Numerical attributes: correlation Coefficient

$$r_{ab} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i - N\bar{A}\bar{B})}{N\sigma_A\sigma_B}$$



Nominal Attributes: chi-square test

	$B=b1$	$B=b2$
$A=a1$	$a1b1$	$a1b2$
$a=a2$	$a2b1$	$a2b2$

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

$O \rightarrow$ Observed value.

$e \rightarrow$ expected value.

$$e_{ij} = \frac{Count(A = a_i) \times Count(B = b_i)}{N}$$

Does Correlation imply Causality

example: The more fireman fighting a damage, the more damage there is going to be.

Data Normalization

Data Transformation

- Smoothing - noise removal / reduction
- Aggregation - eg cities $\Rightarrow$ states (n to 1)
- Generalization - eg city $\Rightarrow$ state (1 to 1)
- Normalization - feature scaling
- Discretization - continuous $\Rightarrow$ intervals.
- Attribute Construction - from existing one

Normalization

$\rightarrow$ Rescaling (Min-max normalization)

- take min and max and convert into new min' and max'
- example:

Income Range [50k, 200k], map to [0, 1.0]

V = 100k. (say)

$$V' = \frac{V - min}{max - min}(max' - min') + min' \rightarrow \underline{\text{Formula}}.$$

$$V' = \left(\frac{100-50}{200-50}\right)(1.0-0) + 0 = 0.33$$

so 100k is mapped to 0.33.

Mean Normalization

Formula $V' = \dfrac{V - mean}{max - \cancel{mean}\ min}$

example: Income Range [50k, 200k], mean = 120k.
V = 100k

$$V' = \frac{100 - 120}{200 - 50} = -0.13$$

Standardization (z-score normalization) symetric

$$V = \frac{V - mean}{stdev.}$$

Eg. Income mean = 120k
stdev = 30k V = 100k.

$$V' = \frac{(100 - 120)}{30} = -0.67$$

Discretization

- Continuous ⇒ intervals
  example income ⇒ 10k increments
- By split or merge
  - Recursive splitting or merging.
- Supervised or unsupervised: class labels

Unsupervised Discretization

- Binning and histogram analysis.
  - equal width, equal frequency
- Clustering analysis
- Intuitive partitioning.

Supervised Discretization

- Predefined class labels
- Entropy-based interval splitting.
  - lower entropy means 'purer' class distribution
- χ^2 analysis-based interval merging.
  - lower χ^2 value means class is independent of interval

Data Reduction

→ Large data takes a long time to mine.
  - Reduce data to make it more efficient.
→ - Should still find similar patterns.
→ Dimensional Reduction — # of attributes redn.
→ Numerosity Reduction — # of objects redn.

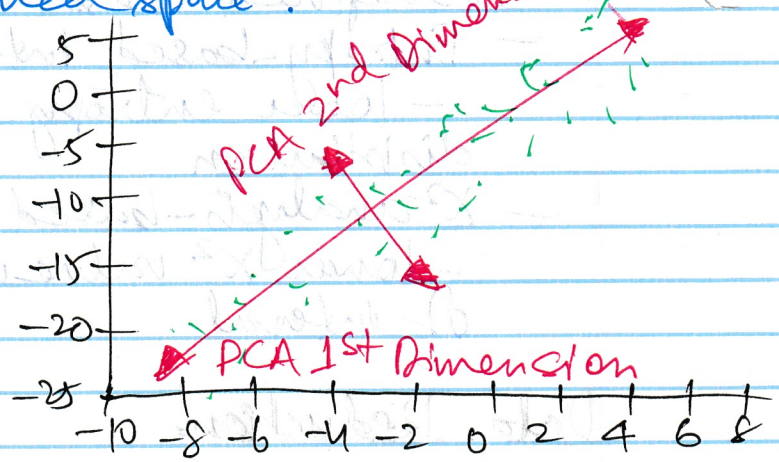
Attribute Selection

- Forward selection – keep adding (more informative) attributes
- Backward elimination – keep removing (least informative) attributes.
- Feature Engineering
  - Domain knowledge, decision tree induction.

Principle Component Analysis (PCA)

* n-dimensional data.
* Use first few orthogonal vectors (principle components)

"We take the original space, then find the transformed space, such that orthogonal vectors capture majority of the information in the transformed space".



Wavelet Transformation

- Linear signal processing, multi-resolution
- Store a small fraction of the strongest wavelet coefficients.
- Used in image processing.

Numerosity Reduction

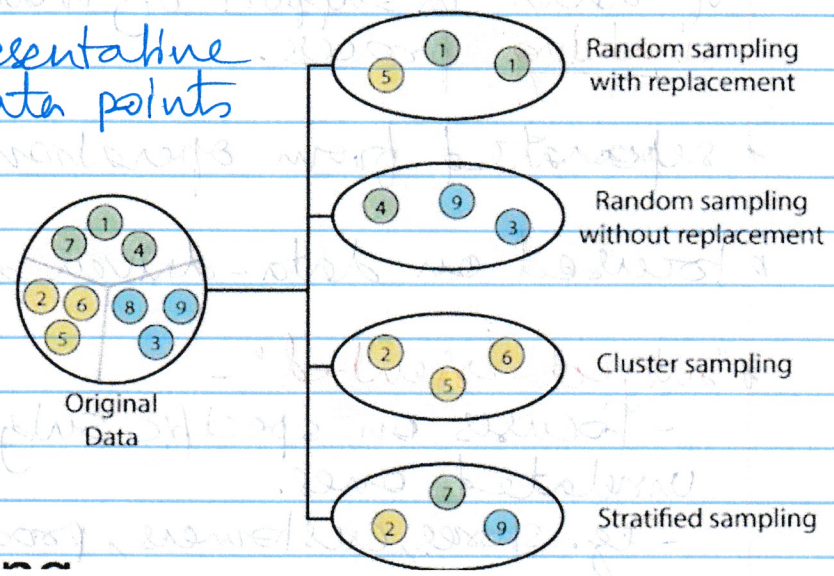
- Parametric method.
  - Assume the data fits a certain model.
  - Estimate model parameters.
  - eg. linear/multi-linear/log-linear regression

- Non-parametric methods
  - Do not assume a certain model
  - Use fewer/smaller data representations.

Sampling

- select a representative subset of data points

- Different sampling methods for different scenarios.



Aggregation -

- daily sales ⇒ monthly sales
- vehicles by city ⇒ by state.

Histogram -

- Store bucket intervals & frequencies.

Clustering -

- Store cluster representations

4. Data Warehousing, OLTP vs OLAP

* Data Warehouse vs operational data
* Data Cube and OLAP
  - Multi-dimensional data management
* Data warehouse architecture - key components to be included and how they are linked

* Why Data warehousing

* A data warehous is "subject-oriented", "integrated", "time-variant", and 'nonvolatile' collection of data in support of management's "decision making" process.

* separated from operation data.

* Focused on data-driven decision support.

* "subject-oriented" -
  - Focuses on specific subjects, ignore unrelated ones.
  - eg. stores, customers, products, students etc.

* "Integrated"
  - Heterogeneous sources, data cleaning and integration
  - eg. New store location: customers, businesses

* "Time-variant"
  - Historical data, longer time span. timestamped data.
  - eg. traffic information in the past 5-10 yrs.

- "Non volatile"
  - separate from operational data, not updated inplace.
  - typical data operations. initial loading, append, read.

OLTP vs OLAP

Online Transactional Processing (OLTP)

- Transaction-oriented tasks: bank transfer, purchase.
- Daily operations: Insert, Update, delete

Online Analytical Processing (OLAP)

- Complex queries on historical data.
- Data analysis for insights and decision making.

OLTP vs. OLAP (2)

	OLTP System Online Transaction Processing (Operational System)	OLAP System Online Analytical Processing (Data Warehouse)
Source of data	Operational data; OLTP's are the original source of the data	Consolidation data; OLAP data comes from the various OLAP Databases
Purpose of data	To control and run fundamental business tasks	To help with planning, problem solving, and decision support
What the data Reveals	A snapshot of ongoing business processes	Multi-dimensional views of various kinds of business activities
Inserts and Updates	Short and fast inserts and updates initiated by end users	Periodic long-running batch jobs refresh the data
Queries	Relatively standardized and simple queries Returning relatively few records	Often complex queries involving aggregations
Processing Speed	Typically very fast	Depends on the amount of data involved; batch data refreshes and complex queries may take many hours; query speed can be improved by creating indexes
Space Requirements	Can be relatively small if historical data is achieved	Larger due to the existence of aggregation structures and history data; requires more indexes than OLTP
Database Design	Highly normalized with many tables	Typically de-normalized with fewer tables; use of star and/or snowflake schemas
Backup and Recovery	Backup religiously; operational data is critical to run the business, data loss is likely to entail significant monetary loss and legal liability	Instead of regular backups, some environments may consider simply reloading the OLTP data as a recovery method

Data Warehouse Model. – organizes data into
- Fact (eg sales) vs dimension (eg. item)

Fact tables: Quantitative data (Sales, revenue)
Dimension tables: Contexual data (time, customer, product).

Common schemas

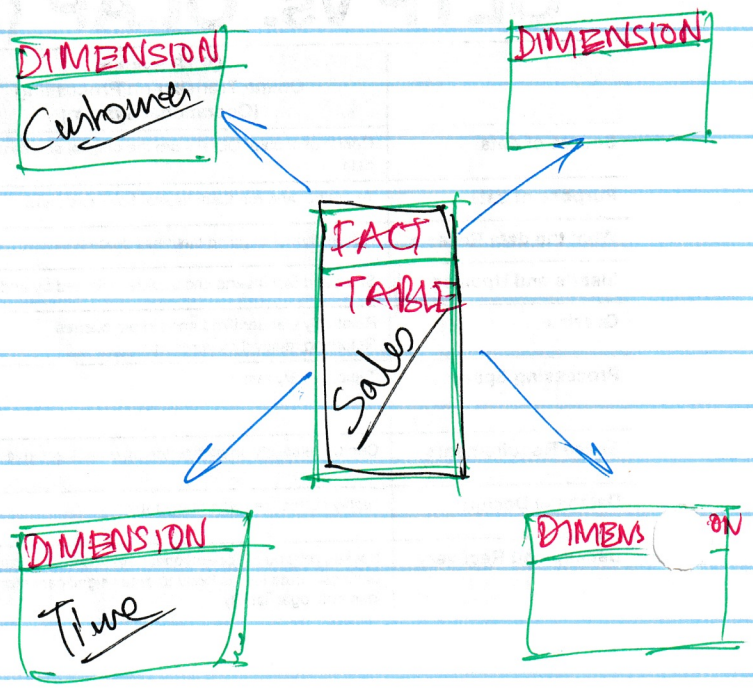
- Star schema: Fact table at center and dimension tables radiating out
- Snowflake schema: Normalized dimensions for reduced redundancy, one fact table.
- Fact constellation schema:
  - multiple fact tables, shared dimension tables. (Galaxy schema)

Example: Star schema

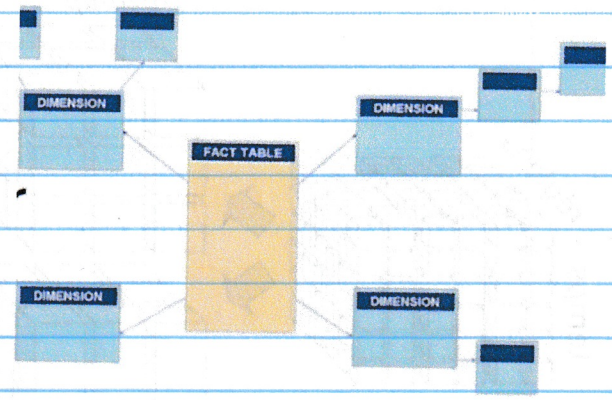
Fact: Sales.
- Customer, item, time

Dimension: Customer
- Name, address, DOB

Dimension: Time
- Year, month, date



Snowflake schema

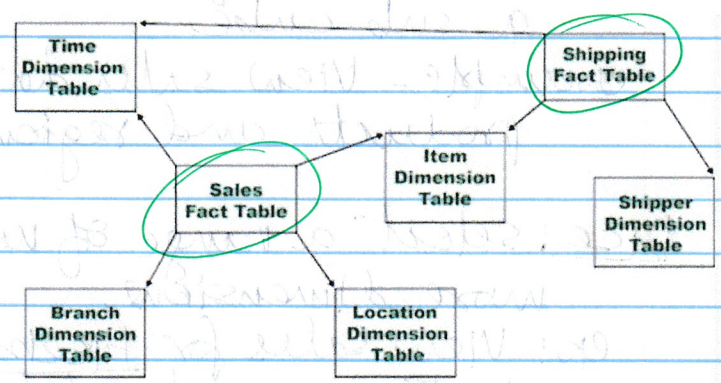


Fact: Sales
· Customer, item, time.

Dimension: Customer
· Name, Address card.

Dimension: Card
· Name. exp date. CVC

Fact Constellation schema



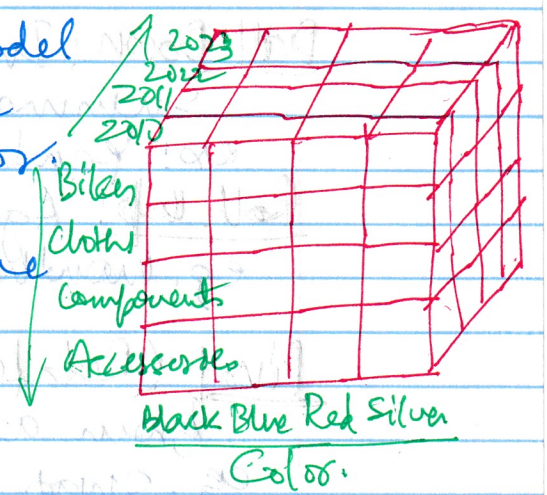
Fact: Sales

Dimension: Item

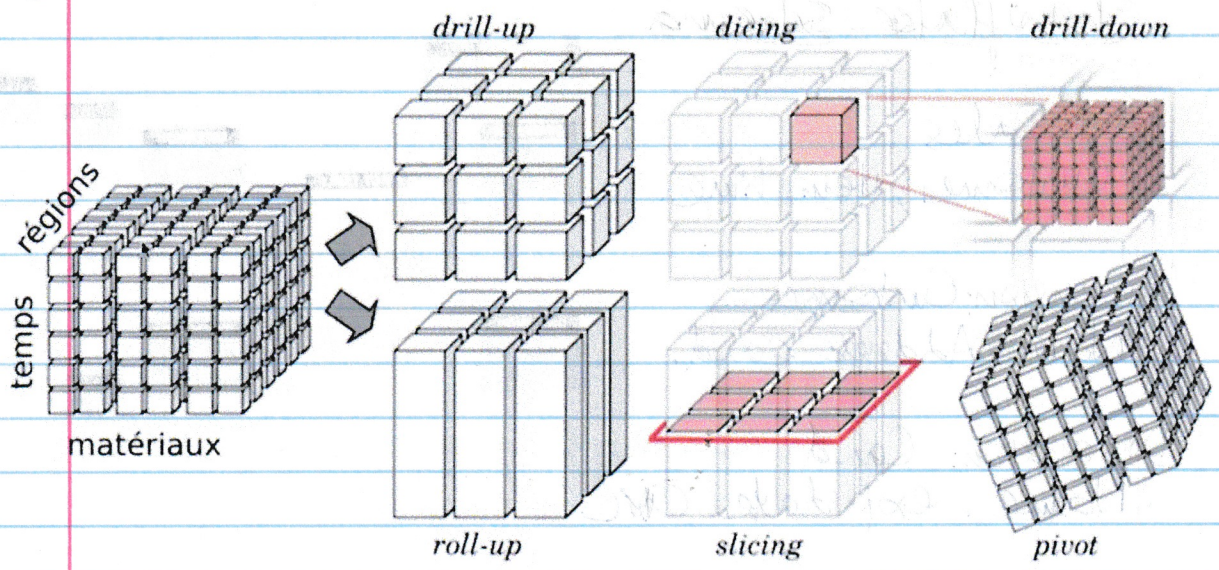
fact: Shipping.

Data Cube

· Multi-dimensional data Model
Dimensions: Cube attribute
eg. year, product, color.
Facts: numeric measure
eg. Sales volume/value



Data Cube Operations



slice: Select a single dimension value to create a sub-cube
example - View sales data for 2024 across all products and regions.

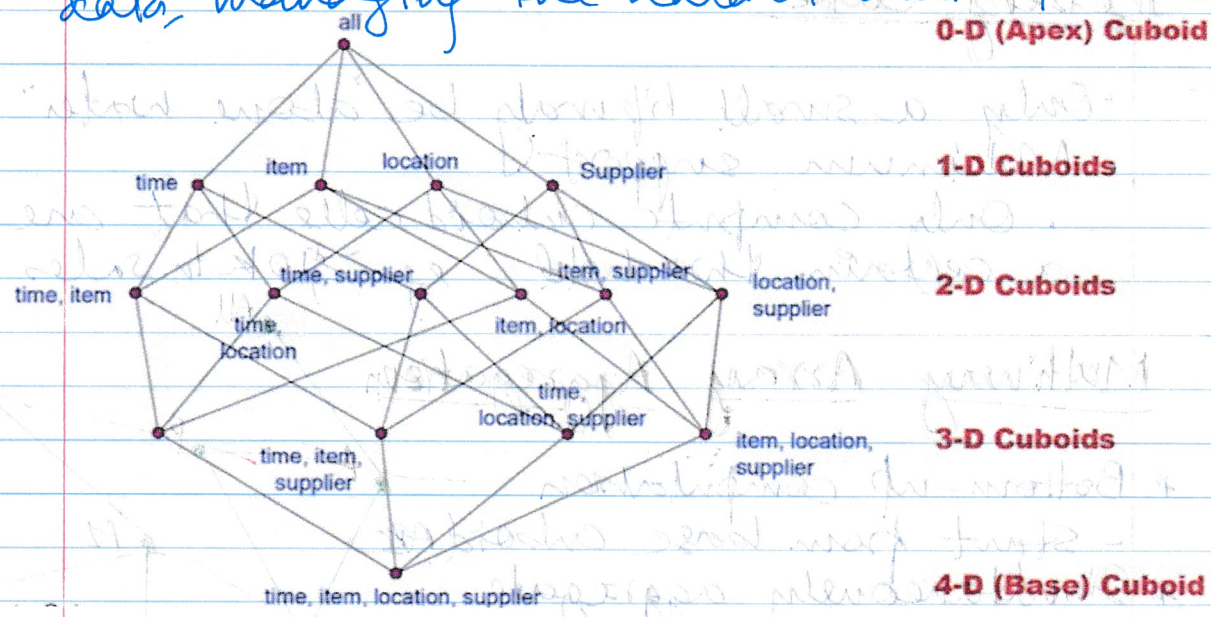
Dice: select a range of values across two or more dimensions.
ex: View sales for Electronics and Furniture in Texas and California during Q1

Drill Down: Increase granularity by moving from summary to detail.
ex: Yearly Sales → Monthly → weekly sales

Roll Up: Aggregate data to a higher level.
eg. monthly Sales → Quarterly → Yearly sales

Pivot (Rotate): Reorient the cube to view data from a different perspective.
eg. swap rows and columns to view Products by Region instead of Regions by Product

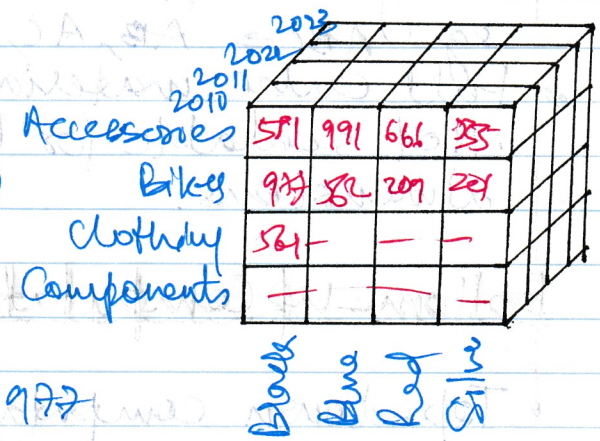
Data Cube: Lattice of Cuboids - Organizing the data, managing the data at certain granularity/dimensions



Cuboid Cells

- Base Cuboids.
- Aggregation (Roll-up)

eg. (year, item, color)



a(2010, bikes, red) = 977

b(2010, bikes, *) = 977+582+209+201 = ---.

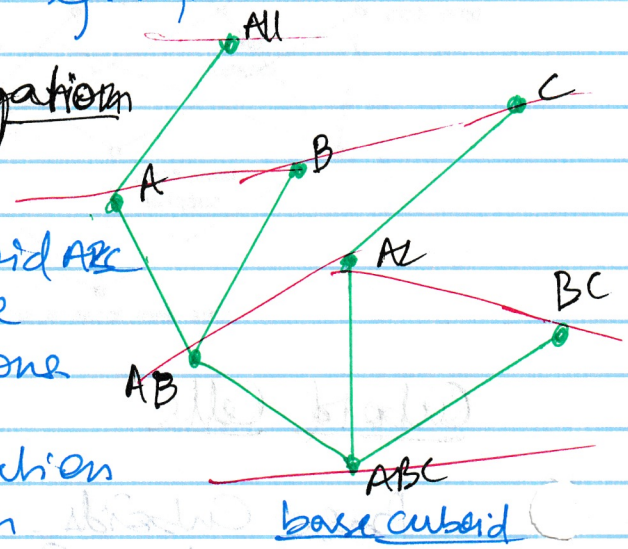
Data Cube Materialization

* Full materialization
  - Pre-compute all cuboids and cells
* No materialization.
  - No precomputation, on-demand.
* Partial Materialization.
  - (heuristically) pre-compute some cuboids and cells

Iceberg Cube

- Only a small tip may be "above water"
- Minimum support
  - Only compute cuboid cells that are above a certain threshold e.g. >10k in sales

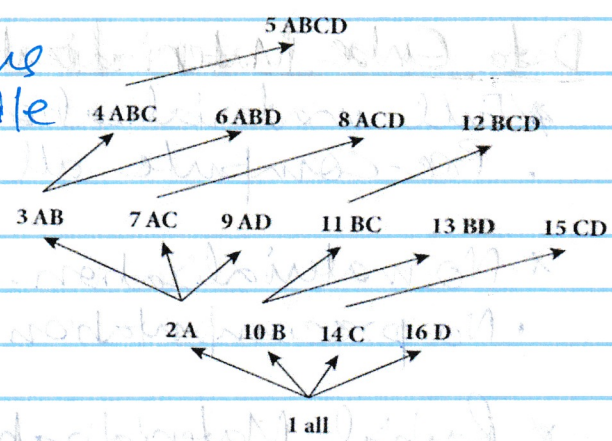
Multiway Array Aggregation



- Bottom-up computation
  - Start from base cuboid ABC
- Simultaneously aggregate along multiple dimensions
  e.g. ABC ⇒ AB, AC, BC
- Full cube materialization
  - Not scalable for high dimensions.

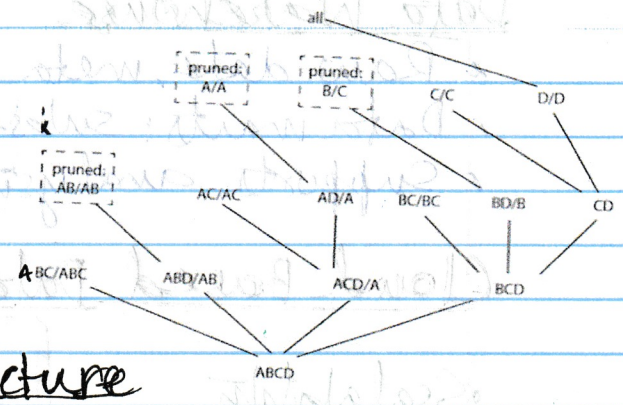
Bottom-Up Computation (BUC)

- Top-down computation, starting from "1 all".
- Iceberg pruning
  - Divide dimensions into partitions while > threshold.

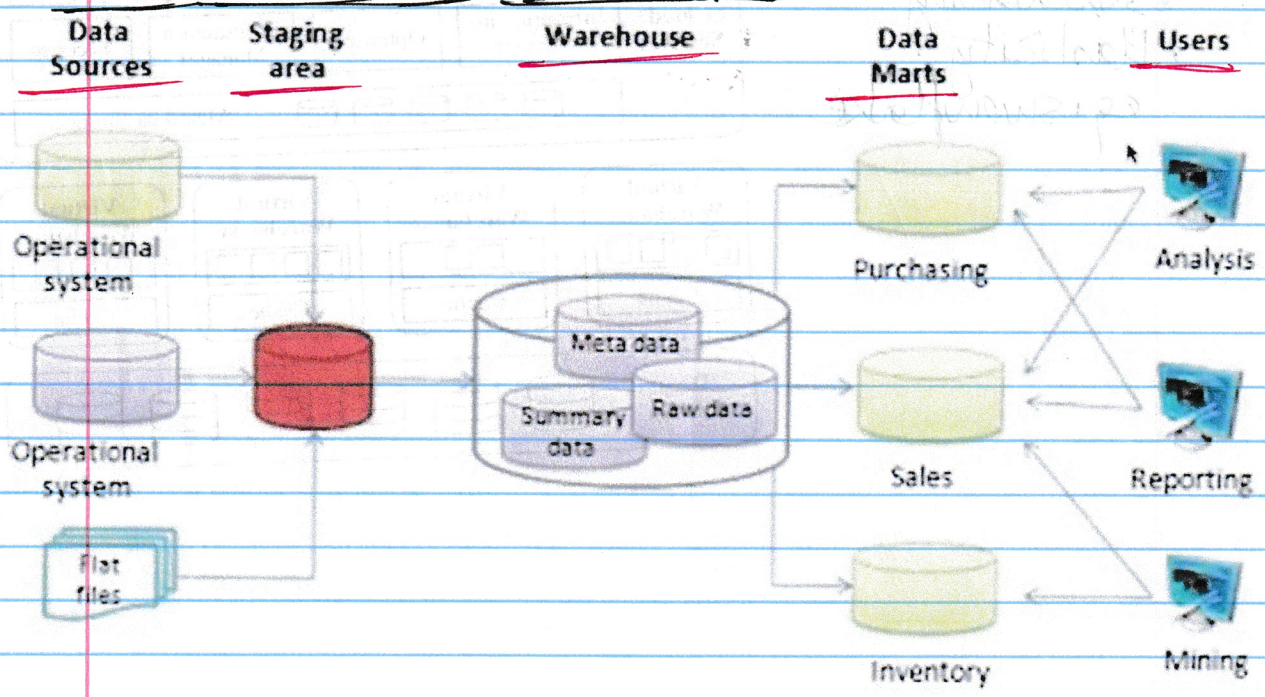


Star-Cubing

- Bottom-up computation & top-down expansion of shared dimensions



Data Warehouse Architecture



Data Sources

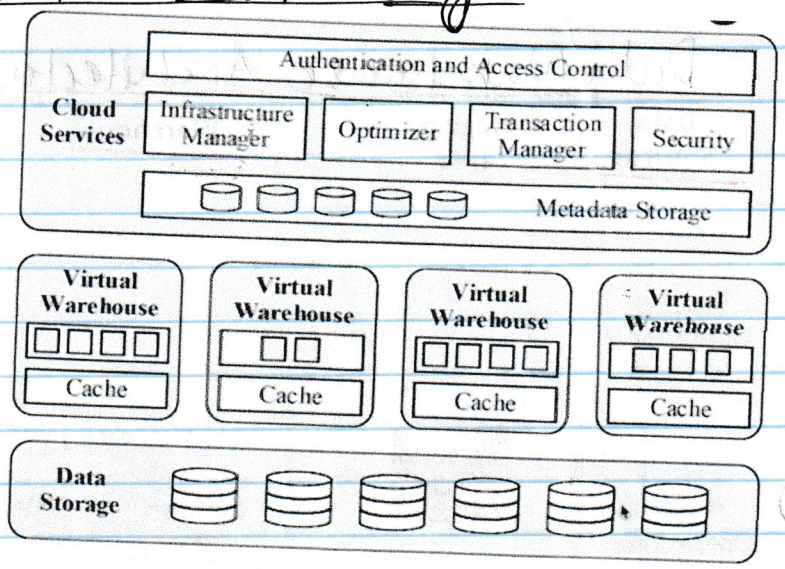
- Flat files eg. csv.
- operational systems
- CRM (Customer Relationship Management)
- ERP (Enterprise Resource Planning)
- structured, semi-structured, unstructured.

Staging

Extract data from various data sources
Transform data
Load data into the data warehouse

Data Warehouse

- Raw data, meta data, summary data.
- Data marts: subsets with specific focuses
- Supports analysis, reports, data mining.

Cloud Based Data Warehousing

- Scalability
- Elasticity
  eg: snowflake