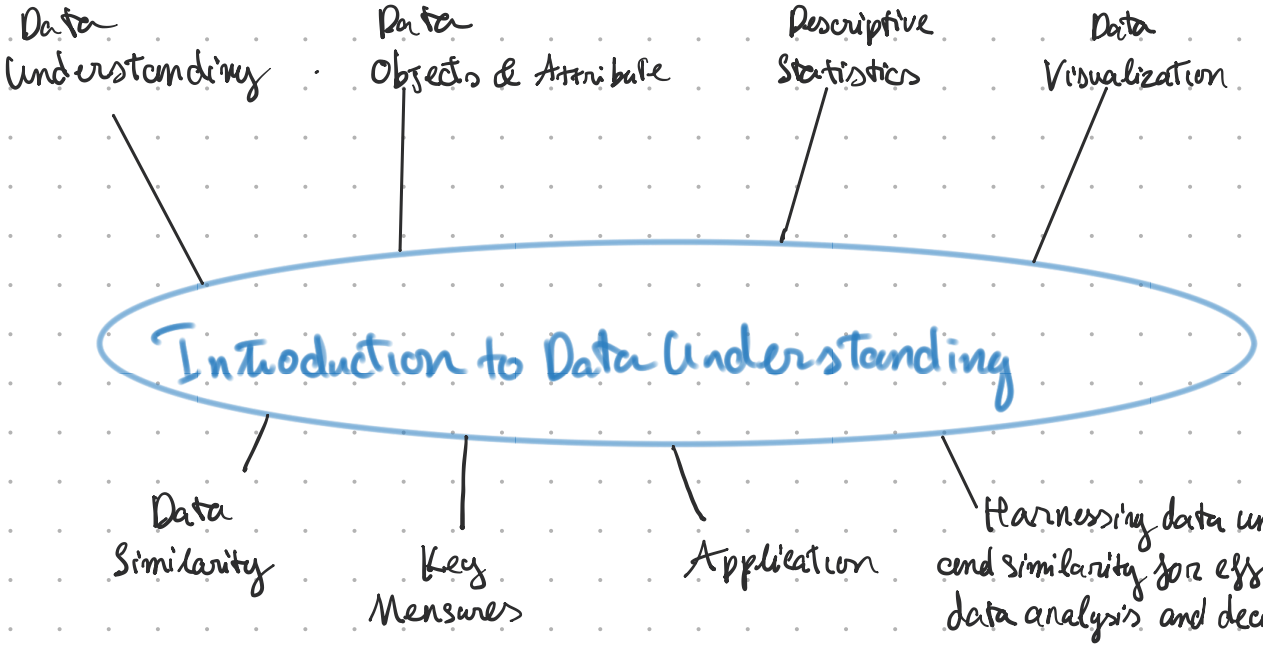
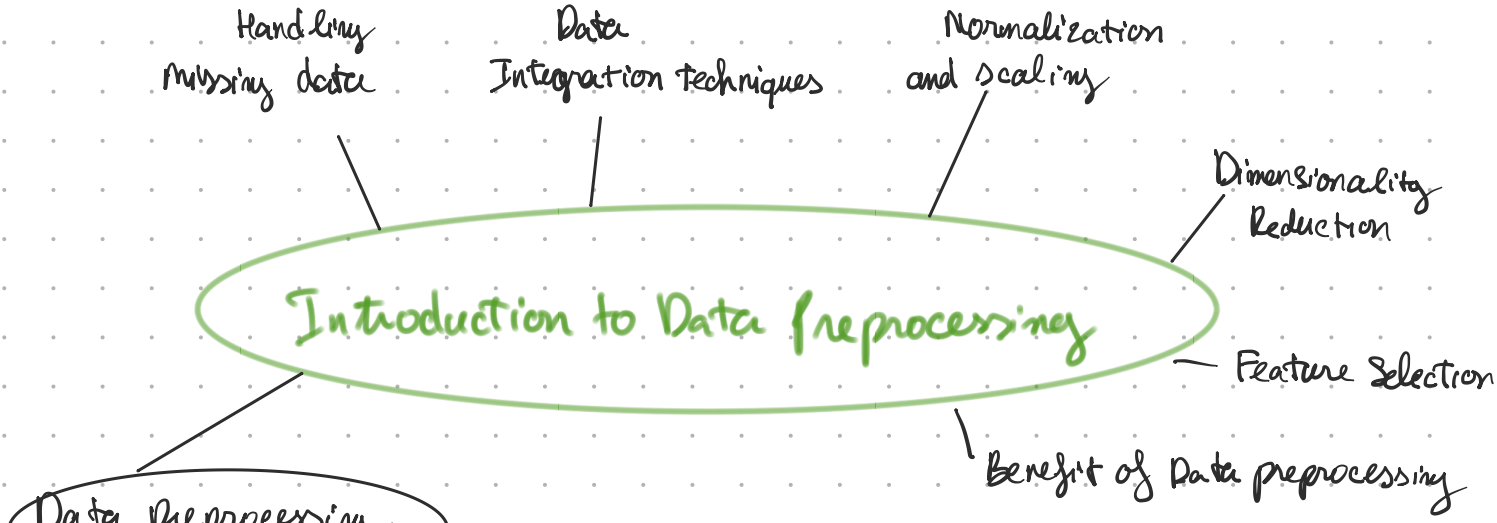
Introduction to Data Mining

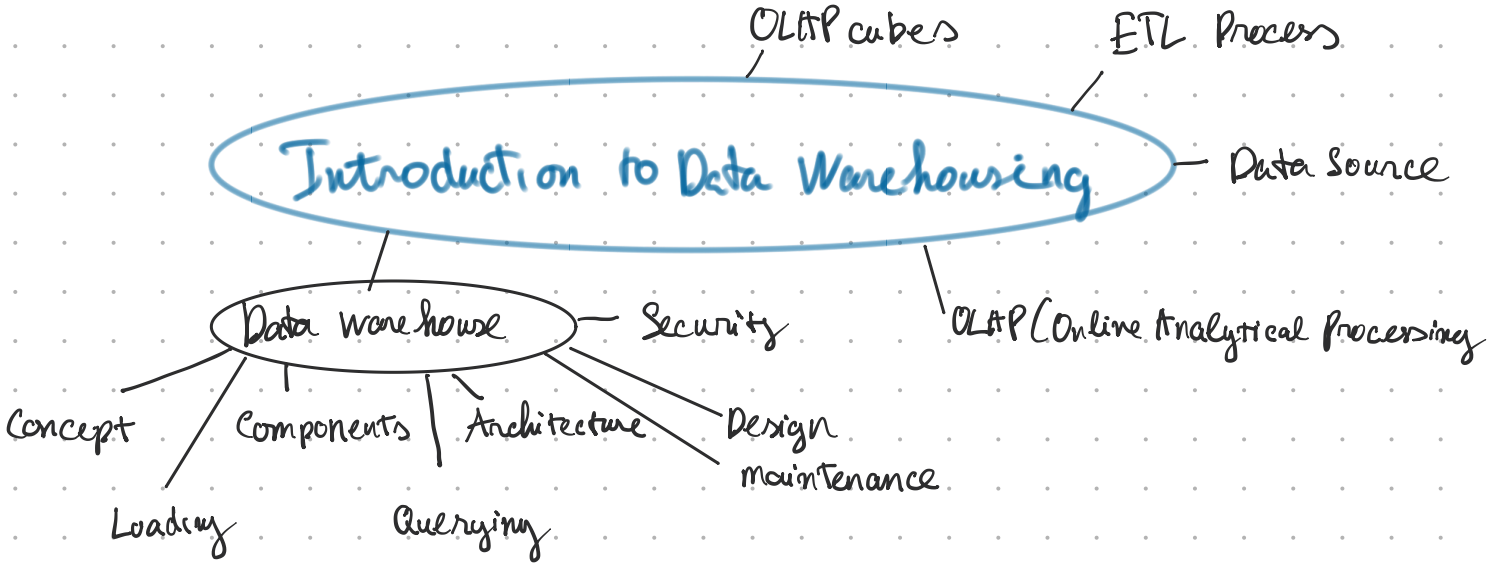
- Key components
  - Collection
  - Preprocessing
  - Mining
  - Evaluation
  - Presentation
- What?
  - Extraction
  - Interesting
  - Huge
- Why?
  - explosive data growth
  - Drowning in data
  - need automative
- 4 views
  - Data
    - The (3, 4, 5) Vs
    - Relational, transaction data
    - Sequential, Temporal, streaming data
  - Knowledge
    - Freq pattern, association, correlation
    - categorization
    - Anomaly, outliers
  - Application
    - Market Analysis, target advertisement
    - Healthcare, medical research
    - Security
  - Technical
    - Freq pattern analysis
      - Freq
        - Itemset
        - Structure
        - Seq
      - Association
      - correlation analysis
    - Classification
      - pre-defined classes
      - need training data
      - Build model to distinguish classes
    - prediction — numerical
    - clustering
      - No predefined classes
      - Intra-cluster similarity
      - Inter-cluster dissimilarity
    - Trend and evolution analysis — changes over time
    - Anomaly detection — Anomaly / outlier





Data preprocessing

- cleaning
  - handeling
    - missing value
    - duplicate records
  - correcting inaccuracies
  - Effective data cleaning
- Integration
  - challenge
    - Data format
    - Data redundancy
  - Benefits
    - Comprehensive analysis
    - Improved decision-making
    - Data integration Streaming data utilization
- Transformation
  - Normalization
  - Attribute Construction
- Reduction
  - Dimensionality Reduction
  - Numerosity Reduction.


Introduction to Data Warehousing
OLAP cubes
ETL Proces
Data Source
OLAP (Online Analytical Processing
Data warehouse
Security
Concept
Components
Architecture
Design
maintenance
Loading
Querying