TOPIC 2. BASIC CONCEPTS OF DATA ANALYSIS

Outline

- Definition of data
- What is database
- What is data analysis
- Concepts of some data analysis methods

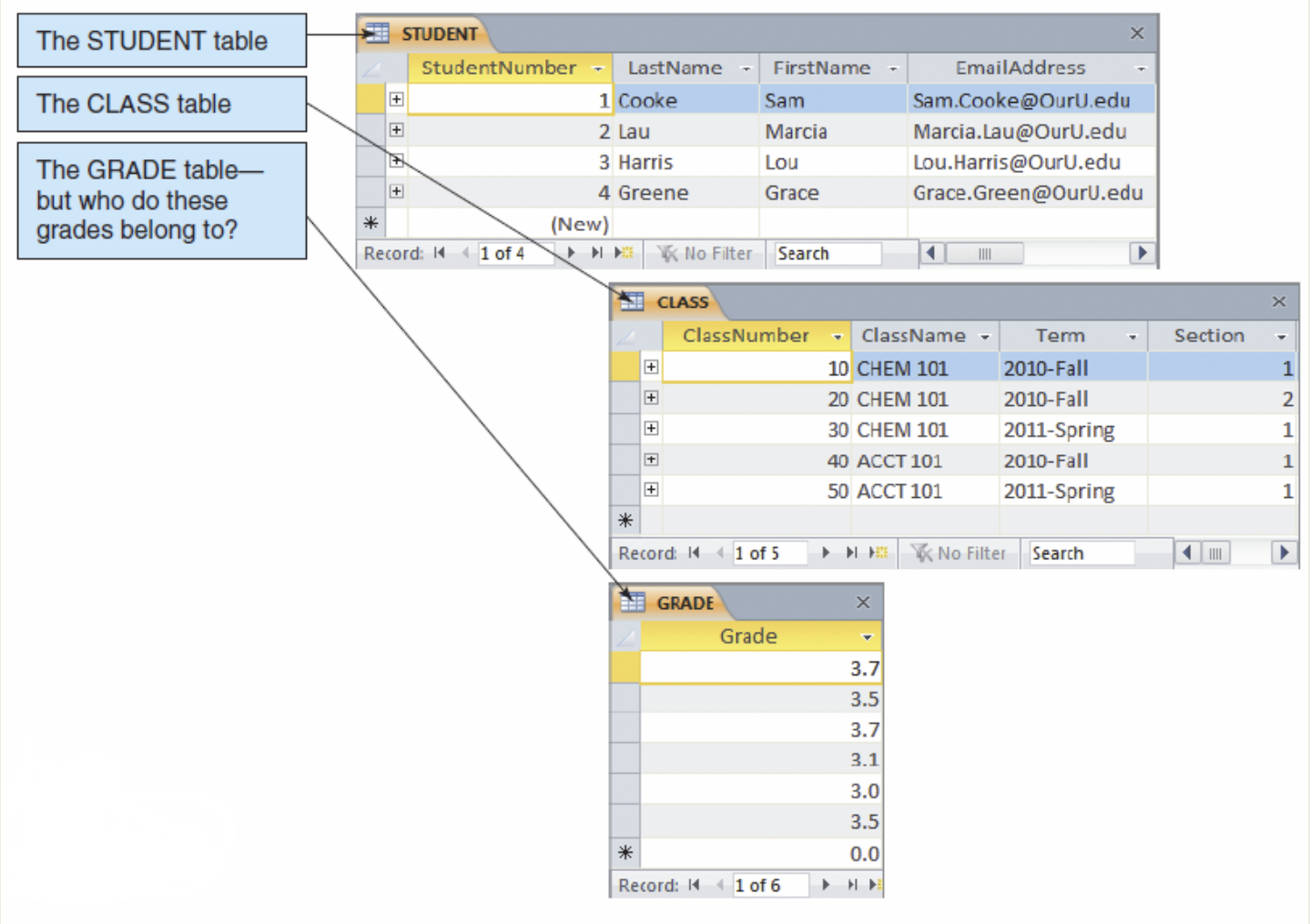
Data and Information

- Data = observations and measurements
- Information = knowledge derived from data
- Databases record data, but they do so in such a way that we can produce information from the data.
  - *The data on STUDENTs, CLASSes, and GRADEs could produce information about each student's GPA.*

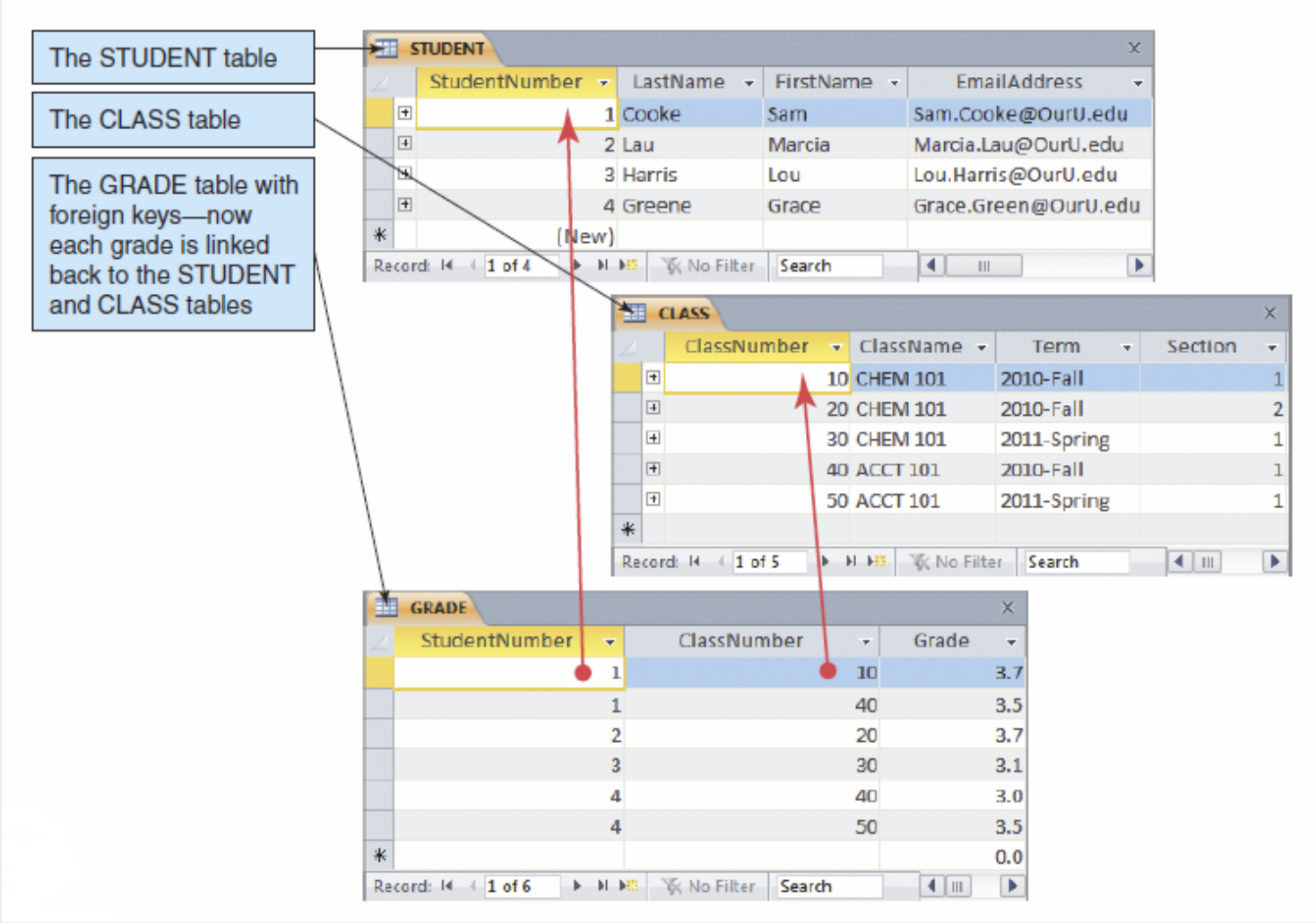
Data and Database

- The purpose of a **database** is to help people track things of interest to them.
- Data is stored in **tables**, which have rows and columns like a spreadsheet. A database may have multiple tables, where each table stores data about a different thing.
- Each row in a table stores data about an occurrence or **instance** of the thing of interest.
- A database stores **data** and **relationships**.

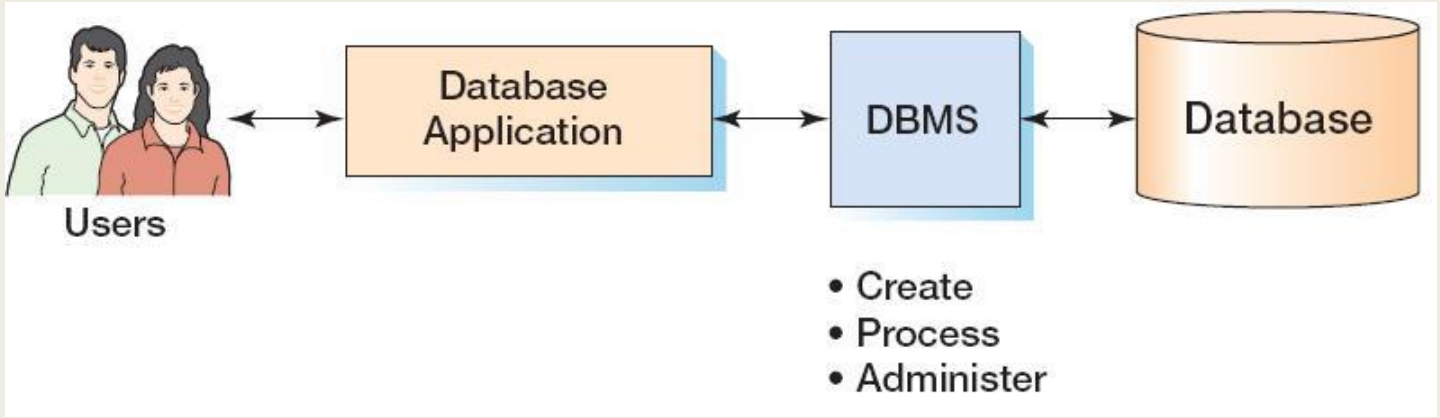
Data in Tables



Database: Related Tables



Components of A Database System



Database

- A **database** is a self-describing collection of integrated tables.
- The tables are called **integrated** because they store data about the relationships between the rows of data.
- A database is called **self-describing** because it stores a description of itself.
- The self-describing data is called **metadata**, which is data about data.

Examples of meta data

USER_TABLES Table

TableName	NumberColumns	PrimaryKey
STUDENT	4	StudentNumber
CLASS	4	ClassNumber
GRADE	3	(StudentNumber, ClassNumber)

USER_COLUMNS Table

ColumnName	TableName	DataType	Length (bytes)
StudentNumber	STUDENT	Integer	4
LastName	STUDENT	Text	25
FirstName	STUDENT	Text	25
EmailAddress	STUDENT	Text	100
ClassNumber	CLASS	Integer	4
Name	CLASS	Text	25
Term	CLASS	Text	12
Section	CLASS	Integer	4
StudentNumber	GRADE	Integer	4
ClassNumber	GRADE	Integer	4
Grade	GRADE	Decimal	(2, 1)

Process of Data Analysis

- Think about analysis EARLY
- Start with a plan
- Code, enter, clean
- Analyze
- Interpret
- Reflect
  - *What did we learn?*
  - *What conclusions can we draw?*
  - *What are our recommendations?*
  - *What are the limitations of our analysis?*

Why Do We Need An Analysis Plan?

- To make sure the questions and your data collection instrument will get the information you want.
- To align your desired “report” with the results of analysis and interpretation.
- To improve reliability--consistent measures over time.

Key Components of A Data Analysis Plan

- Purpose of the evaluation
- Questions
- What you hope to learn from the question
- Analysis technique
- How data will be presented

Analyzing and Interpreting Quantitative Data

- Quantitative Data is
  - *Presented in a numerical format, numbers*
  - *Collected in a standardized manner*
    - *e.g. surveys, closed-ended interviews, tests, automated sensors*
  - *Analyzed using statistical and data mining techniques*

Common descriptive statistics

- Count (frequencies)
- Percentage
- Mean
- Mode
- Median
- Range
- Standard deviation
- Variance
- Ranking

Are Your Data Ready?

- Assign a unique identifier
- Organize and keep all forms (questionnaires, interviews, testimonials)
- Check for completeness and accuracy
- Remove those that are incomplete or do not make sense

Discussing Limitations

Written reports:

- Be explicit about your limitations

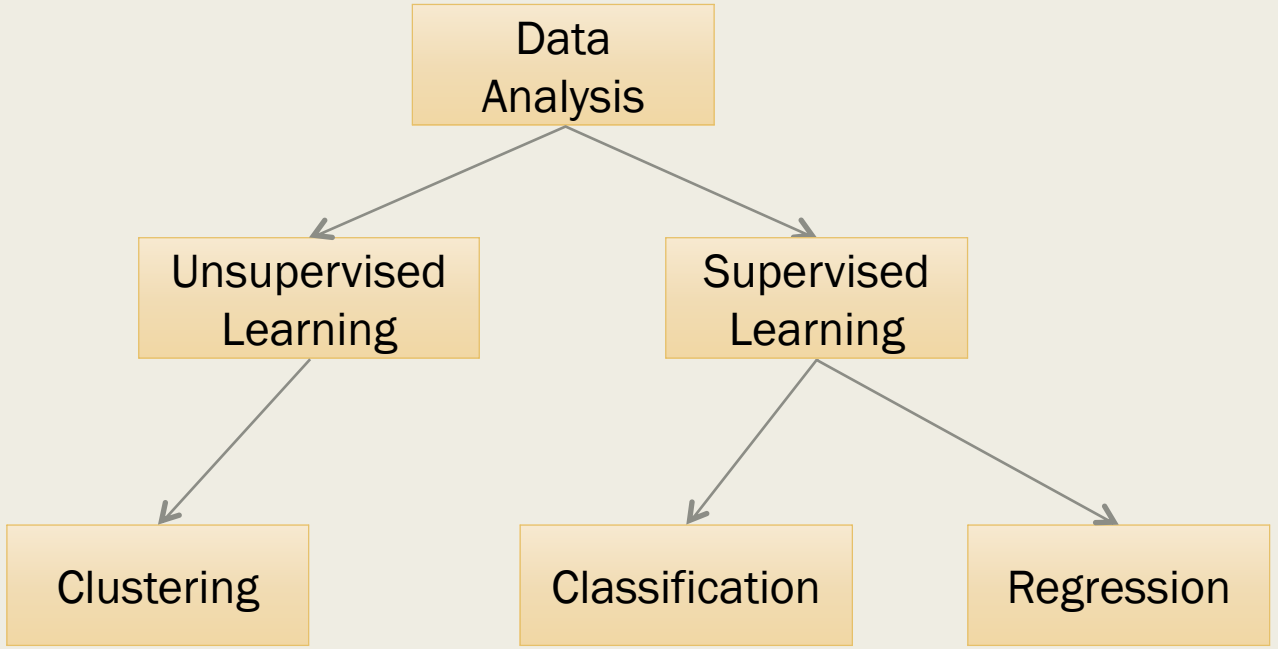
Oral reports:

- Be prepared to discuss limitations
- Be honest about limitations
- Know the claims you cannot make
  - *Do not claim causation without a true experimental design*
  - *Do not generalize to the population without random sample and quality administration (e.g., <60% response rate on a survey)*

Analyzing and Interpreting Qualitative Data

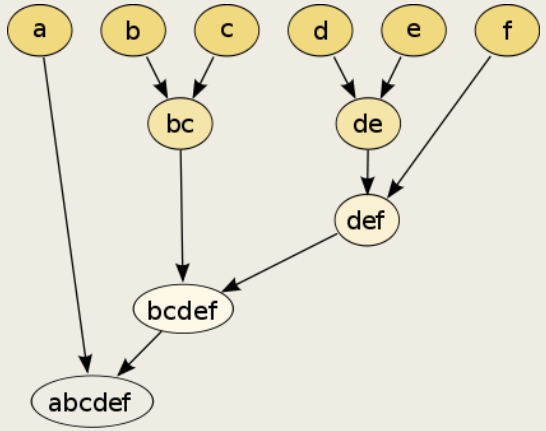
- Qualitative data is thick in detail and description.
- Data often in a narrative format
- Data often collected by observation, open-ended interviewing, document review
- Analysis often emphasizes understanding phenomena as they exist, not following pre-determined hypotheses

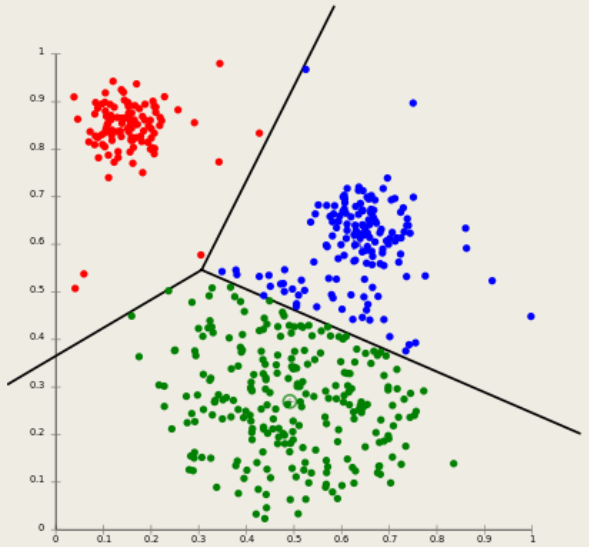
Data Analysis Techniques



Unsupervised Learning

- Hierarchical Clustering
- K-means Clustering
- Expectation Maximization (EM) Clustering



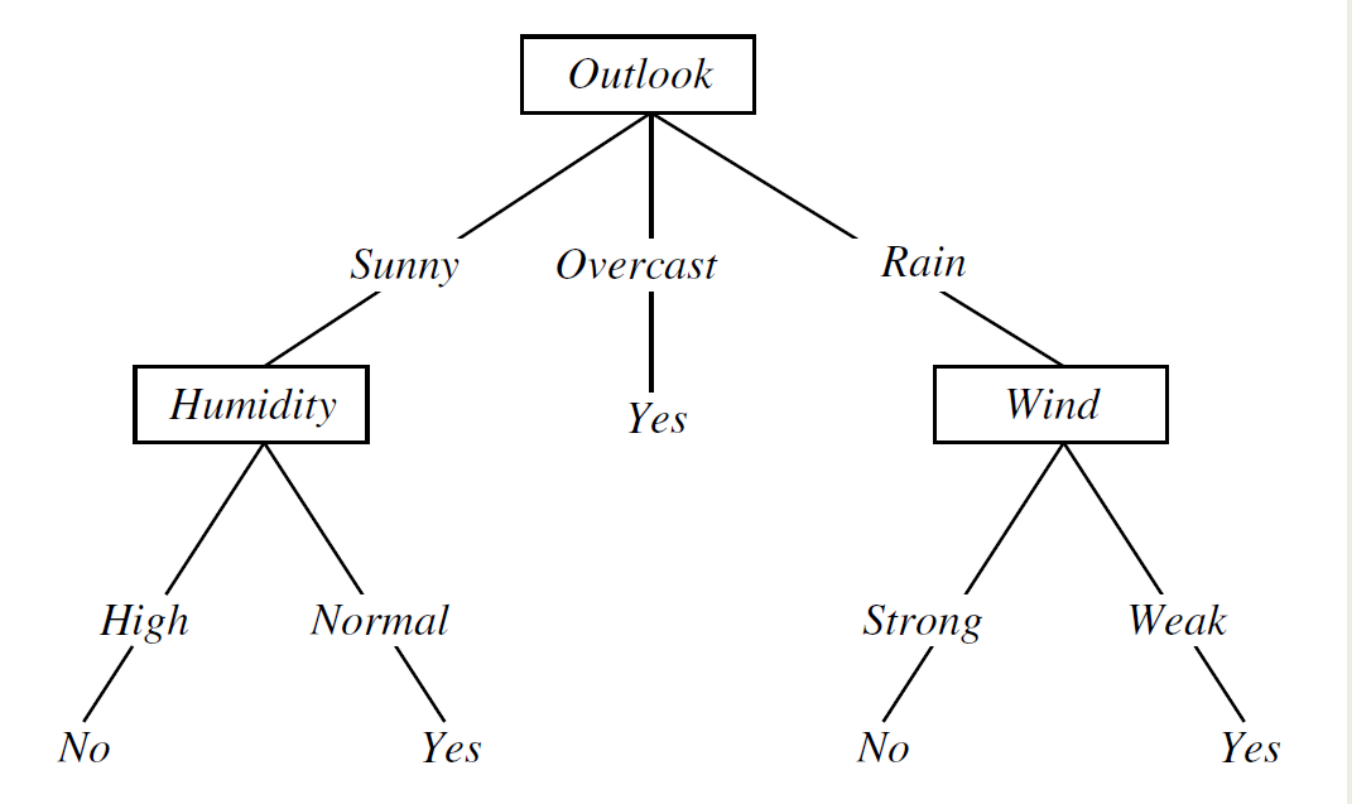


Supervised Learning

- Well-known supervised learning algorithms:
- Decision Tree Based:
  - *Classification and Regression Tree*
  - *Boosting Tree*
  - *Random Forests*
- K nearest neighbors
- Linear regression and time-series methods
- Support vector machine
- Neural Networks

Classification Tree

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



K nearest neighbors

k nearest neighbors – Classification

Voting scheme!