# Data Retrieval & Cleaning

## Retrieving Data

- SQL Databases → relational, fixed schema.  
- NoSQL Databases → non-relational, flexible, often JSON-based.  
- APIs.  
- Cloud Data Sources.  
- Flat Files → CSV (comma-separated), TSV (tab-separated), or other delimiters.

## Reading in Database Files (SQLite + Pandas)

1. Create a path variable: path to the database.  
2. Create a connection variable: connect to the database.  
3. Create a query variable: SQL query to read the table.  
4. observations = pd.read\_sql(query, connection) → read data.  
5. tables = pd.read\_sql("SELECT \* FROM sqlite\_master", connection) → read metadata tables.

## JSON & NoSQL

- JSON: standard cross-platform format, similar to Python dictionaries.  
- NoSQL: non-relational, flexible structures, often store data in JSON format.

## Data Cleaning

Important because messy data leads to unreliable results.  
  
Common issues:  
- Duplicate or unnecessary data.  
- Inconsistent data, typos.  
- Missing data.  
- Outliers.  
- Data source issues.  
  
Handling Missing Data:  
- Remove rows with missing values.  
- Impute missing values.  
- Mask → create a special category for missing.  
  
Handling Outliers:  
- Detection: plots, statistics, residuals.  
- Policies: remove, impute, transform, or use robust models.