

FROM A BLUE  
M ● N



---

# 자기 소개

---

- ❖ 이름
- ❖ 자신을 기억하도록 만드는 특징?
- ❖ 쉴 때 뭐해요? (뭐하고 쉬어요?)

# 시간표

모든 주차							
시간/요일	월	화	수	목	금		
10:00 ~ 10:30	월요일 그룹 활동	데일리 스크럼					
10:30 ~ 11:00	주간 수업	개발 활동	개발 활동	개발 활동	주간 수업		
11:00 ~ 11:30							
11:30 ~ 12:30					주간피드백		
12:30 ~ 14:00		점심 시간					
14:00 ~ 14:30	주간 수업	그룹세션	그룹세션	그룹세션	개발 활동		
14:30 ~ 15:00							
15:00 ~ 16:00	개발 활동	개발 활동	개발 활동	개발 활동			
16:00 ~ 17:00				스쿼드 세션			
17:00 ~ 18:00							
18:00 ~ 18:30	PR				주간회고		
18:30 ~ 19:00	성장노트 작성 + 그룹 회고						

# Honor Code

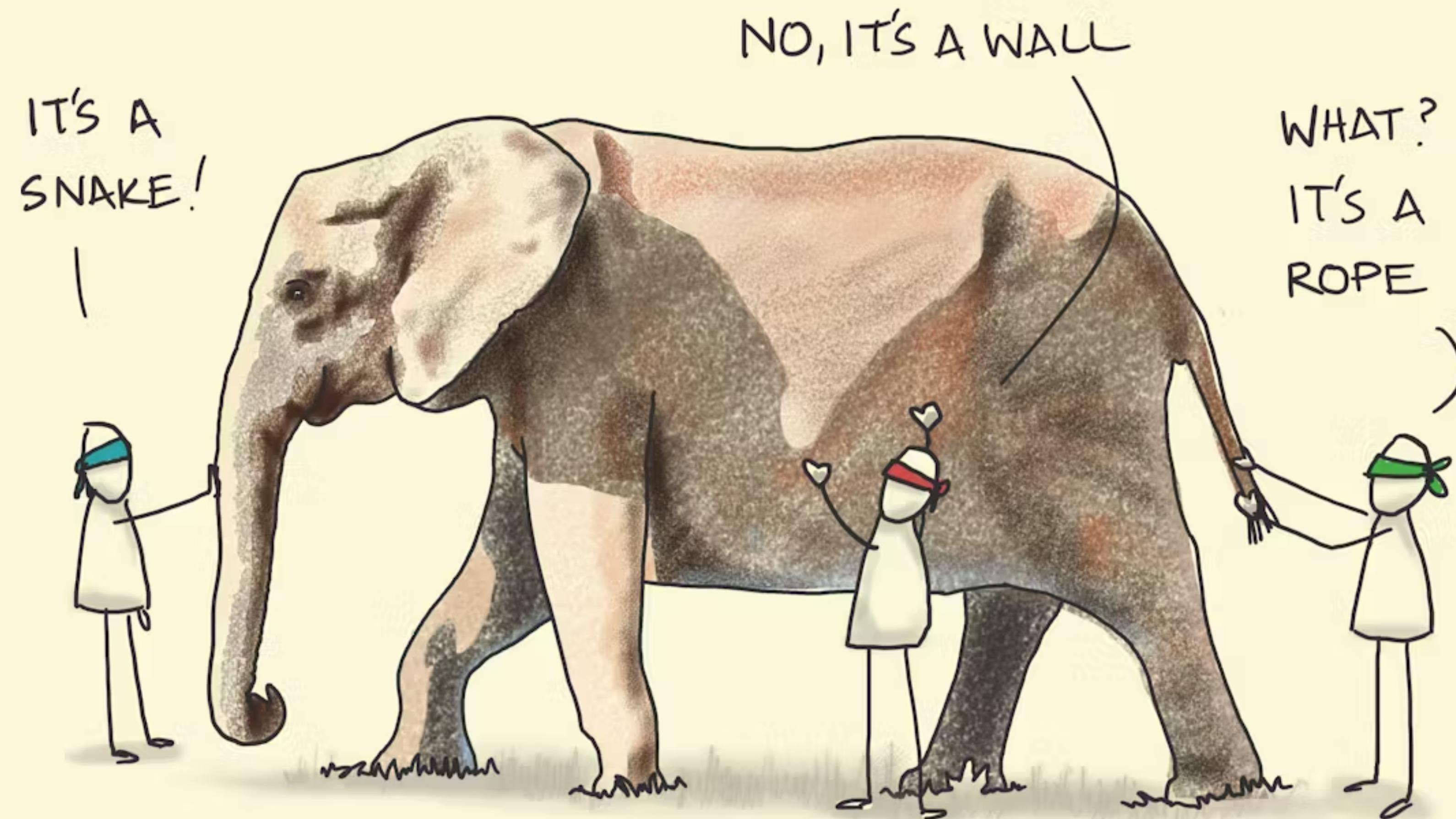
- 어떤 경우에도 동료를 존중하고, 함부로 대하지 않는다.
- 혼자가 아닌 함께 성장하려고 노력한다.
- 지식을 적극적으로 나누고, 서로를 돋는다.



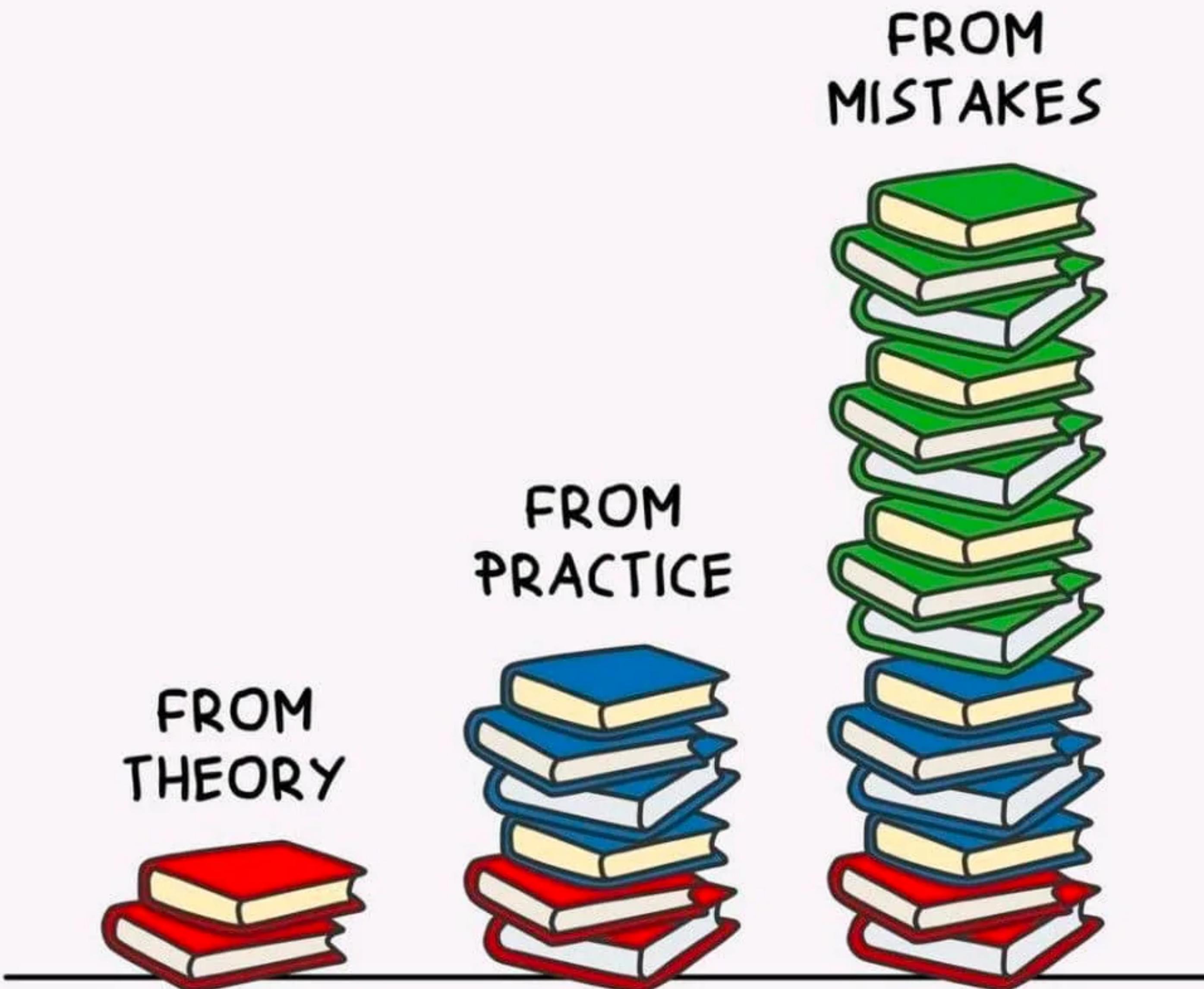
What If You're Learning  
in a Highly Uncertain Environment?

# THE BLIND AND THE ELEPHANT

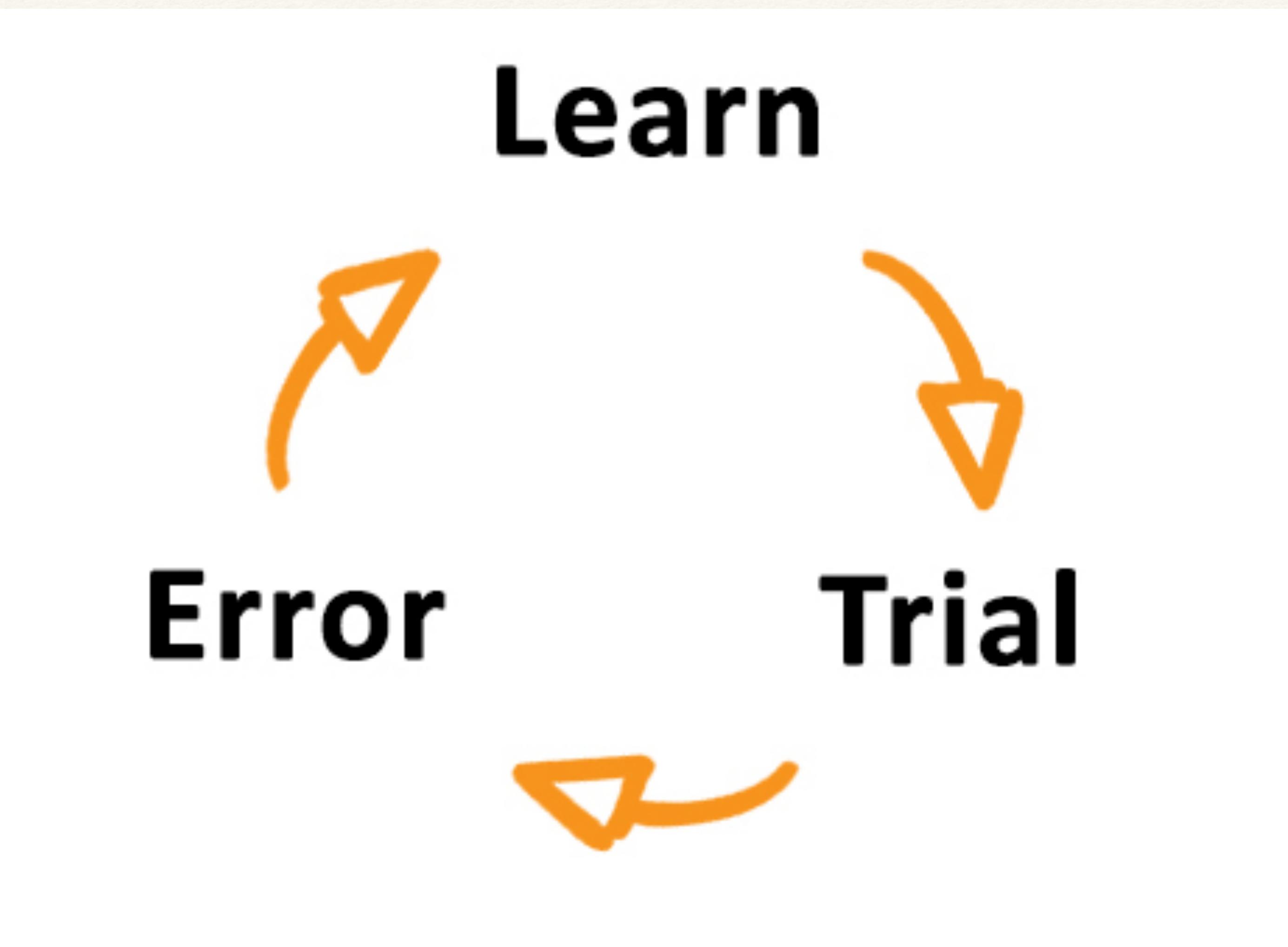
OUR OWN EXPERIENCE IS RARELY THE WHOLE TRUTH



# HOW MUCH YOU LEARN



# Trial and Error (시행착오)





HMG 소프티어 부트캠프 7기

# Introduction to Data Engineering

Dano Lee

# Modern Data Ecosystem



**DATA**

Why use Data?

# Monetary Value

- ❖ Data를 금전적 가치가 있는 정보로 바꾼다.
- ❖ 처리비용이 비싸다는 건 ‘그비용보다 훨씬 더 가치있는 정보가 나와야 한다’는 것.



# Problem - Solution

Monetary Value

Data + Technology

“If I were given one hour to save the planet,  
I would spend 59 minutes defining the problem  
and one minute resolving it,”

# AMD CEO 리사 수 일하는 비법 알려줌



출처 : Bloomberg Originals

<https://www.youtube.com/watch?v=HS4tBHpaWVo>

---

# Why Data Matters

---

We often rely on **heuristic** decision-making — fast, intuitive, but prone to bias.

But we need to shift toward **evidence-based** decisions — grounded in facts, not just instincts.

Data provides the evidence we need to make informed, reliable, and scalable decisions.

---

# Actionable Data

---

information that can be acted upon or information that gives enough insight into the future that the actions that should be taken become clear for decision makers.

---

# Question

---

- CPM
- CPC
- CPI - Acquisition Funnel

---

# Data Professionals

---

- Data Engineers
- Data Analysts
- Data Scientists
- Business Analysts
- Business Intelligence Analysts

# Data Professionals

---

- Data Engineering converts raw data into usable data
- Data Analytics uses data to generate insights
- Data Scientists use Data Analytics and Data Engineering to predict the future using data from the past
- Business Analysts and Business Intelligence Analysts use these insight and predictions to drive decisions that benefit and grow their business

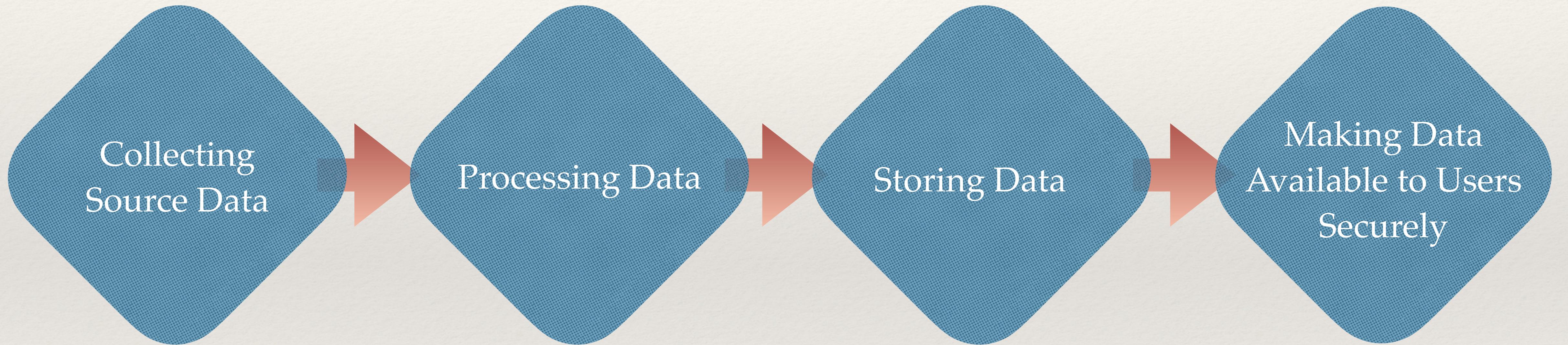
---

# Data Engineering

---

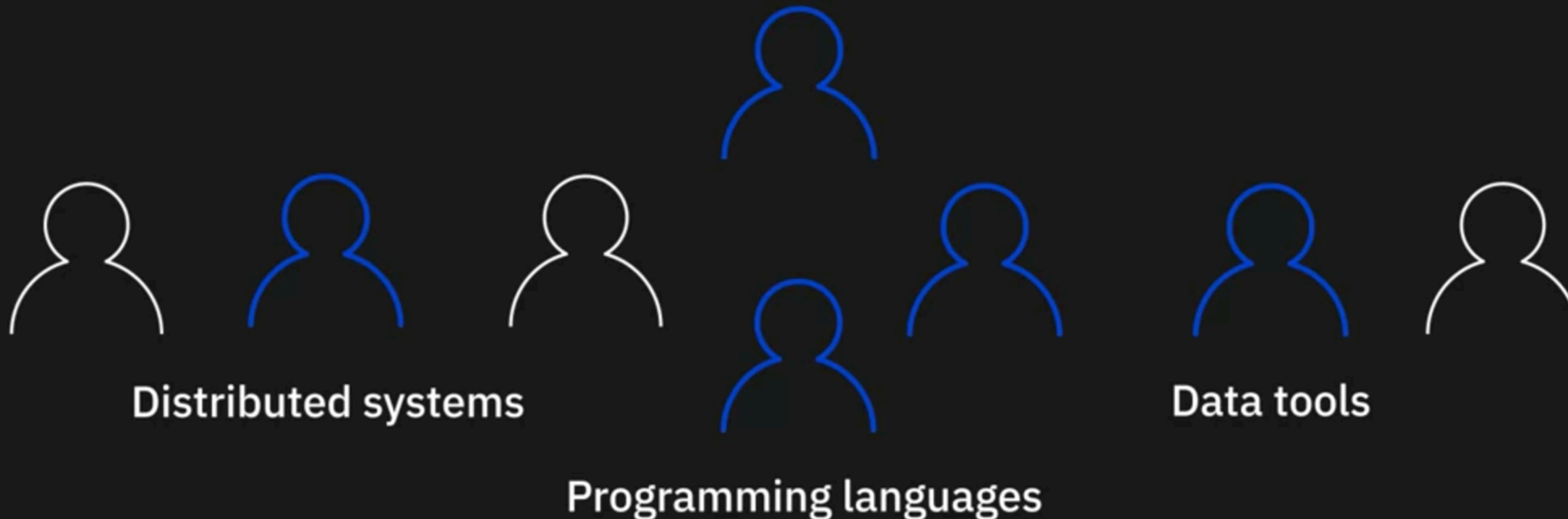
The goal of Data Engineering is to make quality data available for analytics and decision-making. And it does this by collecting raw source data, processing data so it becomes usable, storing data, and making quality data available to users securely.

# Data Engineering



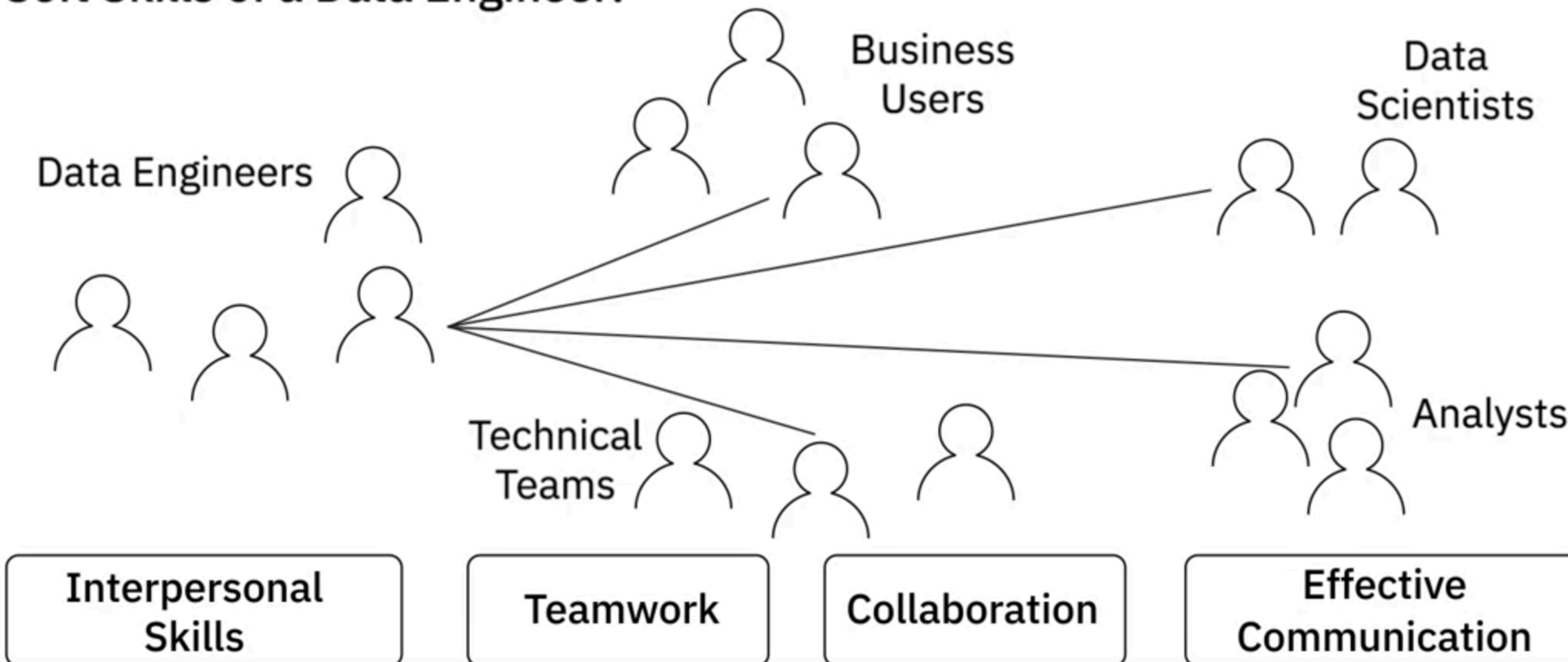
# Data Engineering is a team sport

Optimize data stores for high availability



# Soft Skills

## Soft Skills of a Data Engineer:





What's the moral of the story?

# Trial and Error

Learn



Error

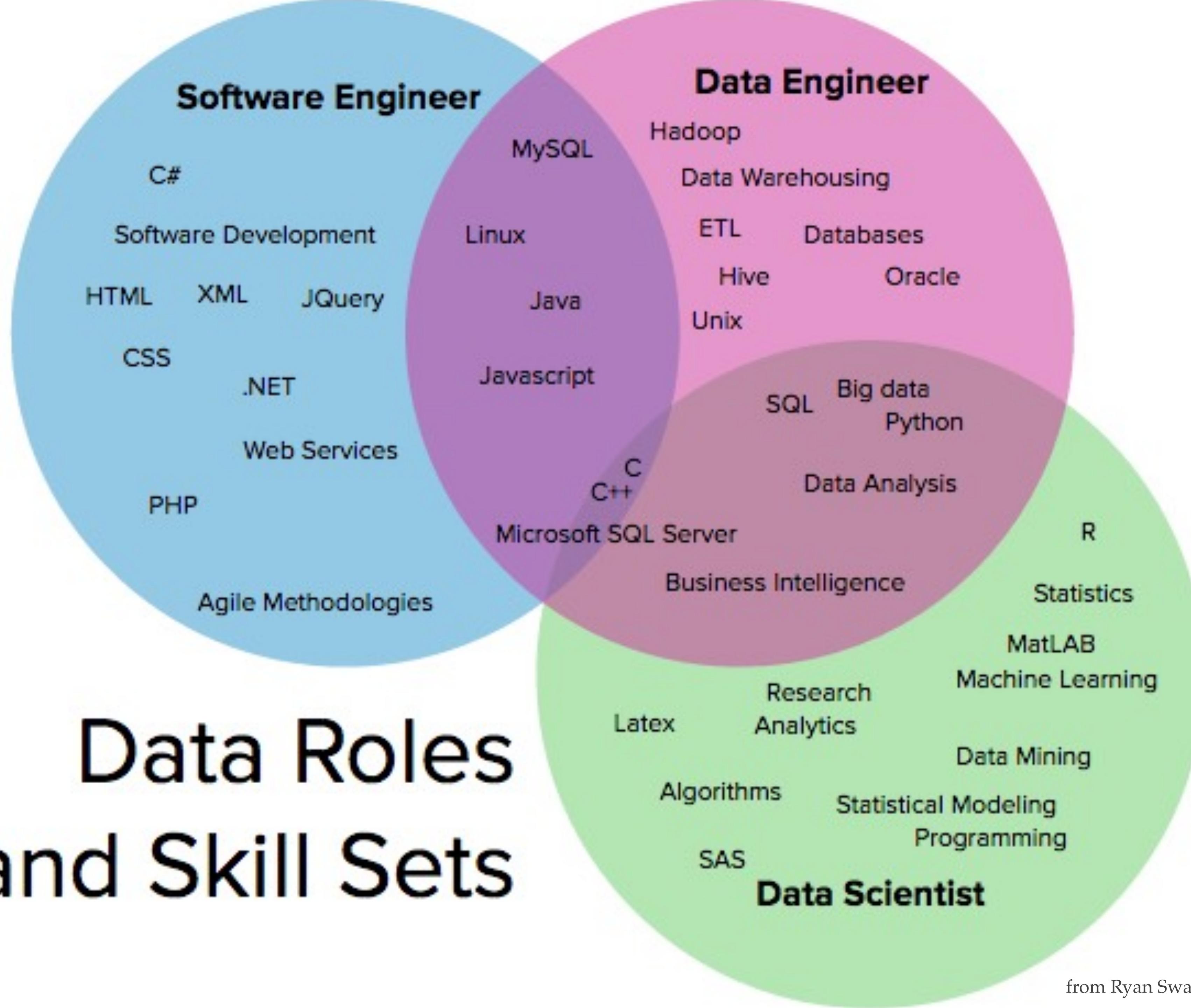


Trial



What if you are the daughter?

# Data Roles and Skill Sets



---

# Technical Skills

---

- ❖ Working with different operating systems and infrastructure components such as virtual machines, networks, and application services.
- ❖ It also includes working with databases and data warehouses, data pipelines, ETL tools, big data processing tools, and languages for querying, manipulating, and processing data.

---

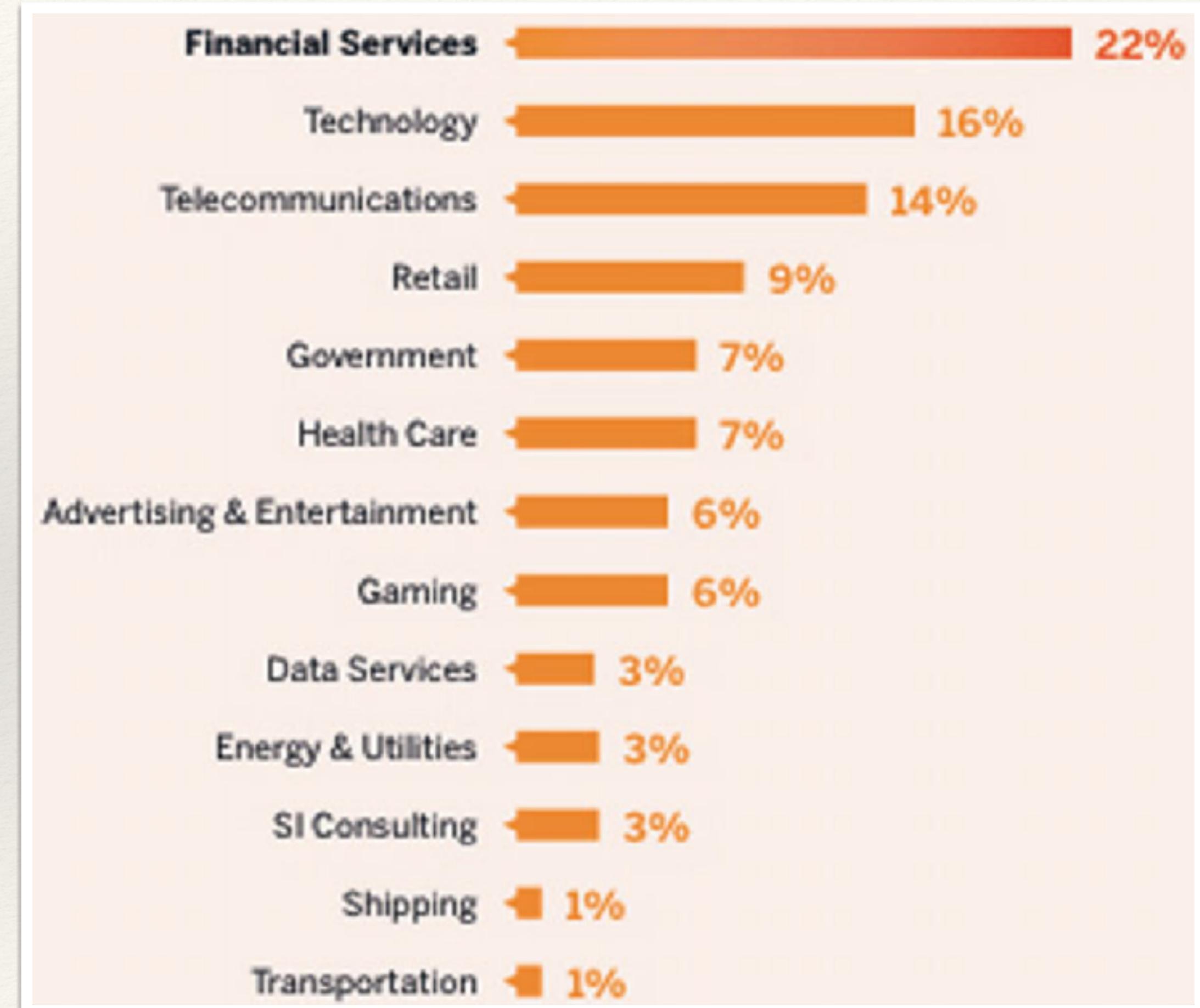
# Functional Skills

---

- ❖ Convert business requirements into technical specifications
- ❖ Work with the complete software development lifecycle
  - ❖ Ideation -> Architecture -> Design -> Prototyping -> Testing -> Deployment -> Monitoring -> Optimization
- ❖ Understand data's potential application in business
- ❖ Understand risks of poor data management
  - ❖ Data quality, Data privacy, Security, and Compliance

# Big Data Usage by Industry

- ❖ Where do you want to go?



source: Utilizing Big Data for Health Care Automation: Obligations, Fitness and Challenges

# Data Types

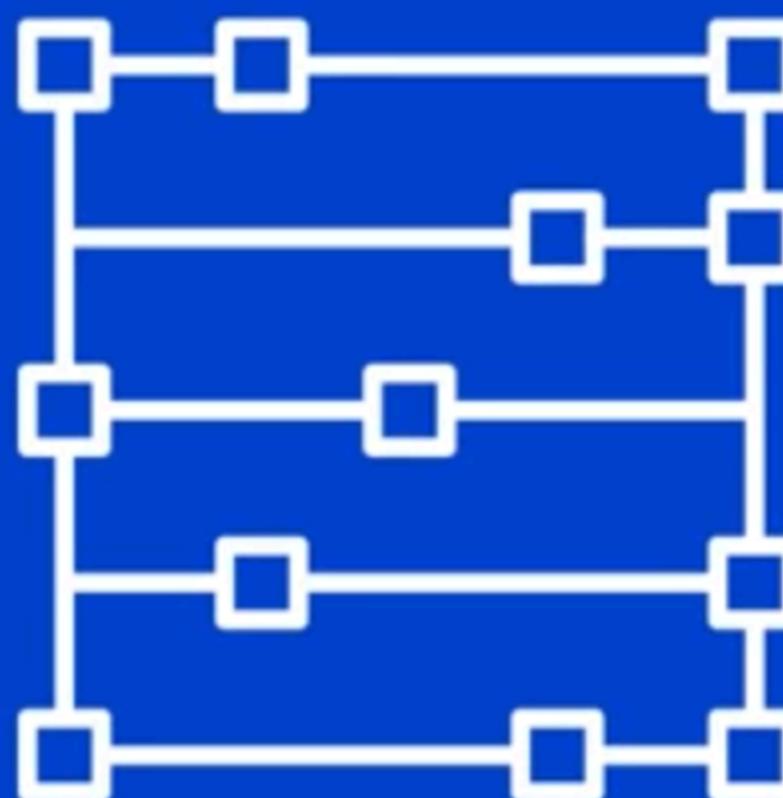
---

# Types of Data

---

- Structured Data
- Semi-structured Data
- Unstructured Data

# Structured Data



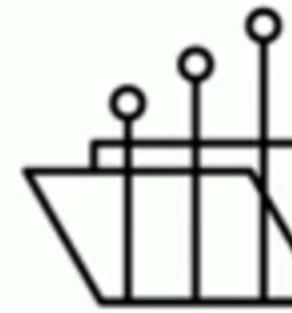
SQL Databases



Online Transaction Processing



Spreadsheets



Online forms



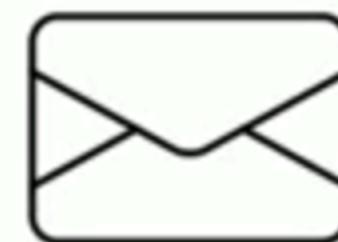
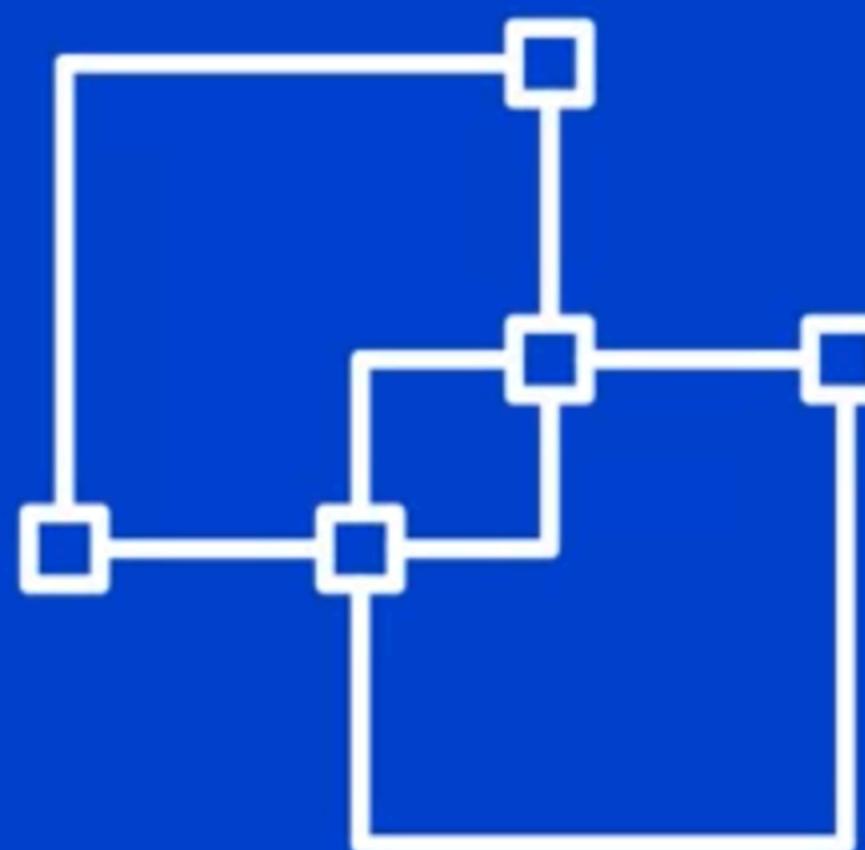
Sensors GPS and RFID



Network and Web server logs



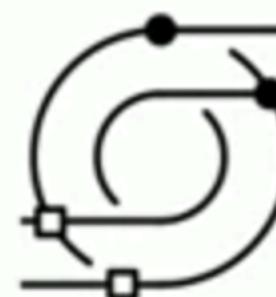
# Semi- Structured Data



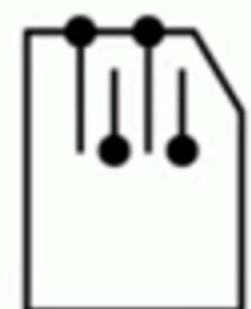
E-mails



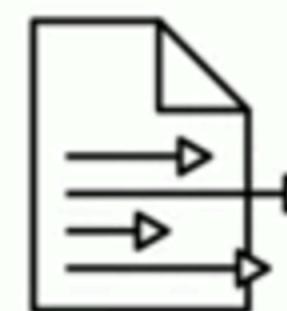
XML and other markup languages



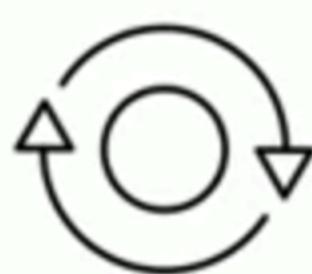
Binary executables



TCP/IP packets



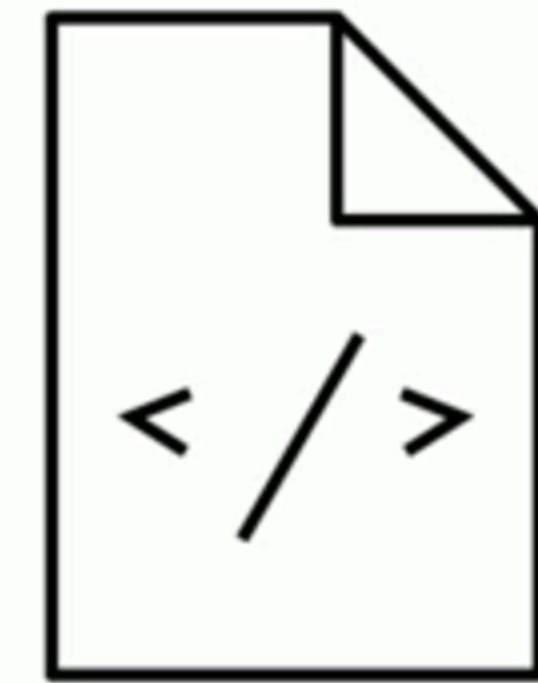
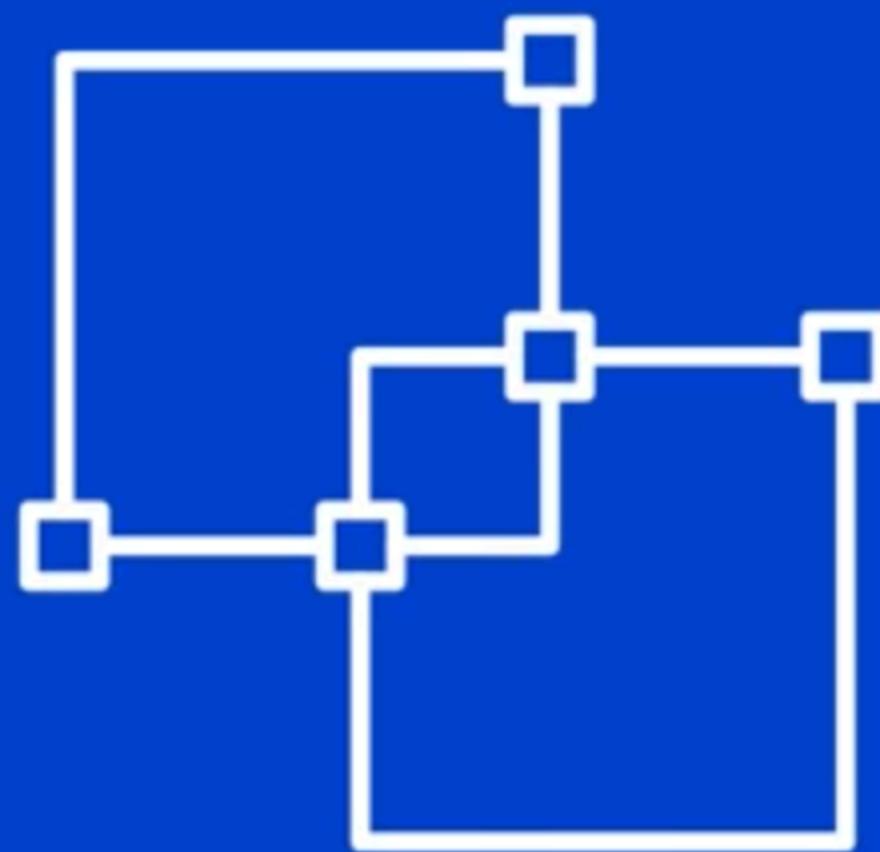
Zipped files



Integration of data



# Semi- Structured Data

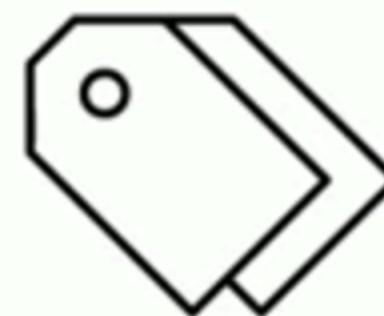


XML



JSON

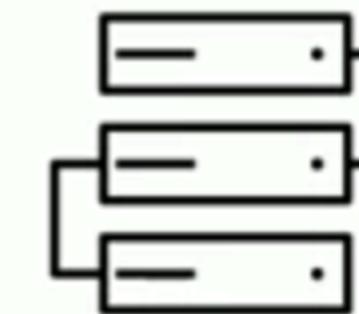
Allow users to



Define Tags



Attributes



To store data



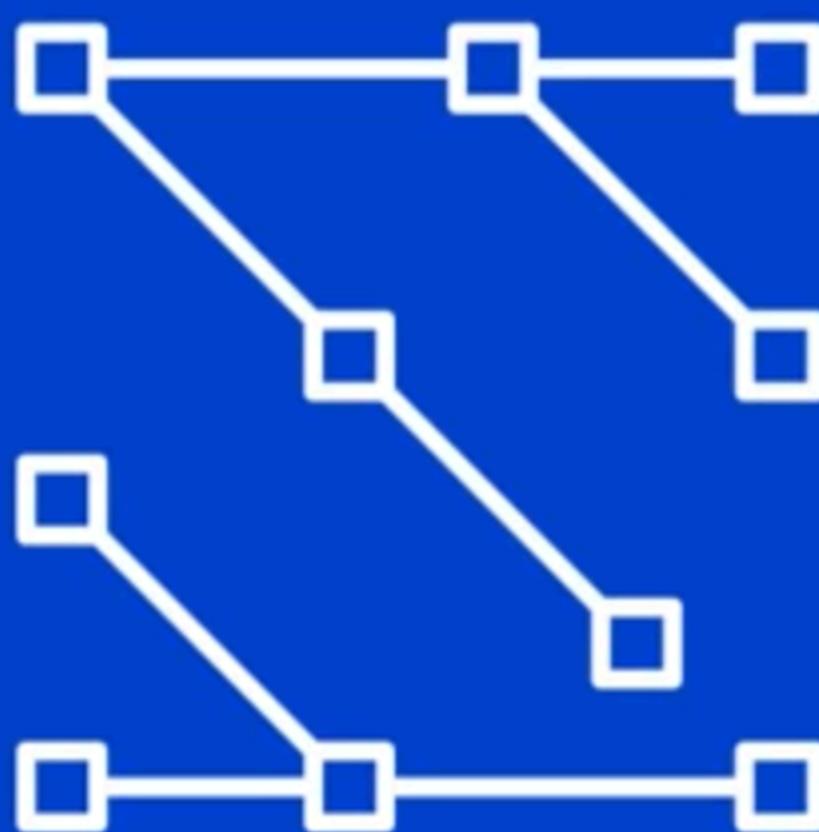
# Unstructured Data



- Web pages
- Social media feeds
- Images in varied file formats
- Video and Audio files
- Documents and PDF files
- PowerPoint presentations



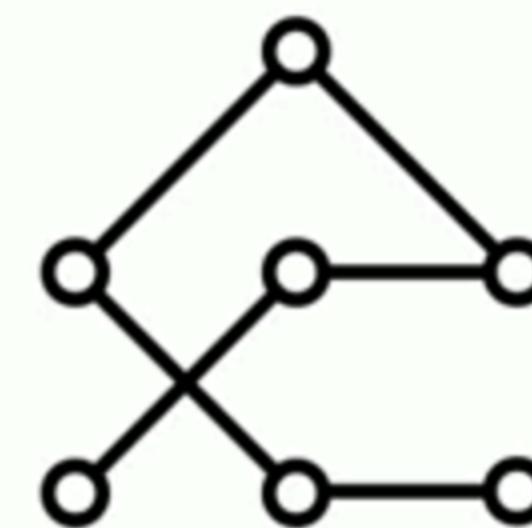
# Unstructured Data



Files and Docs



Manual Analysis



NoSQL



Analysis Tools



---

# Home Address?

---

- String?
- Structured data?
- Semi-structured data?

# 결국 중요한 정보가 어떤 형태의 데이터냐?

- 동영상 파일: 동영상의 내용 vs 동영상의 metadata
- Email: Email의 body vs Email의 header
- 테이블 (표): cell의 값 vs ????

---

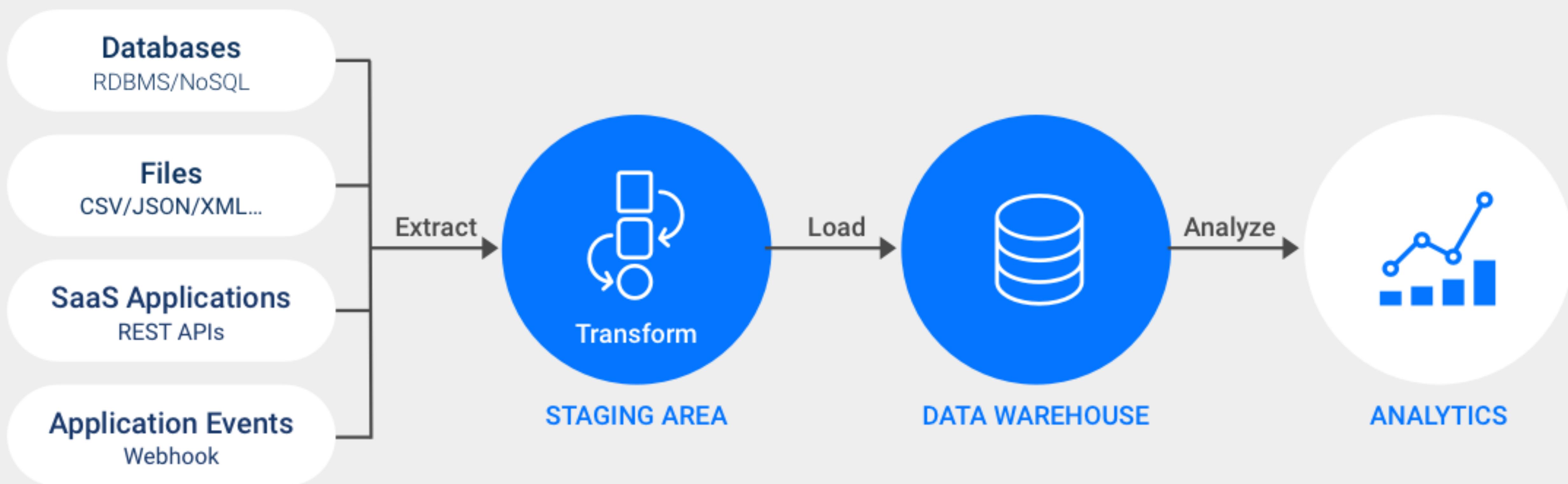
# Question

---

- Table: Person
- columns: last name, first name, identification number, address, etc
- address?
- etc column에 json 형식의 데이터가 들어있다.

# Extract, Transform & Load

# ETL PROCESS



# Python Programming

# PEP 8 – Style Guide for Python Code

---

- <https://peps.python.org/pep-0008/>
  - Indentation: Use 4 spaces per indentation level
  - Tabs or Spaces?
    - Python disallows mixing tabs and spaces for indentation.
- Pylint: a static code analyser for Python. Pylint analyses your code without actually running it. It checks for errors, enforces a coding standard, looks for code smells, and can make suggestions about how the code could be refactored.

---

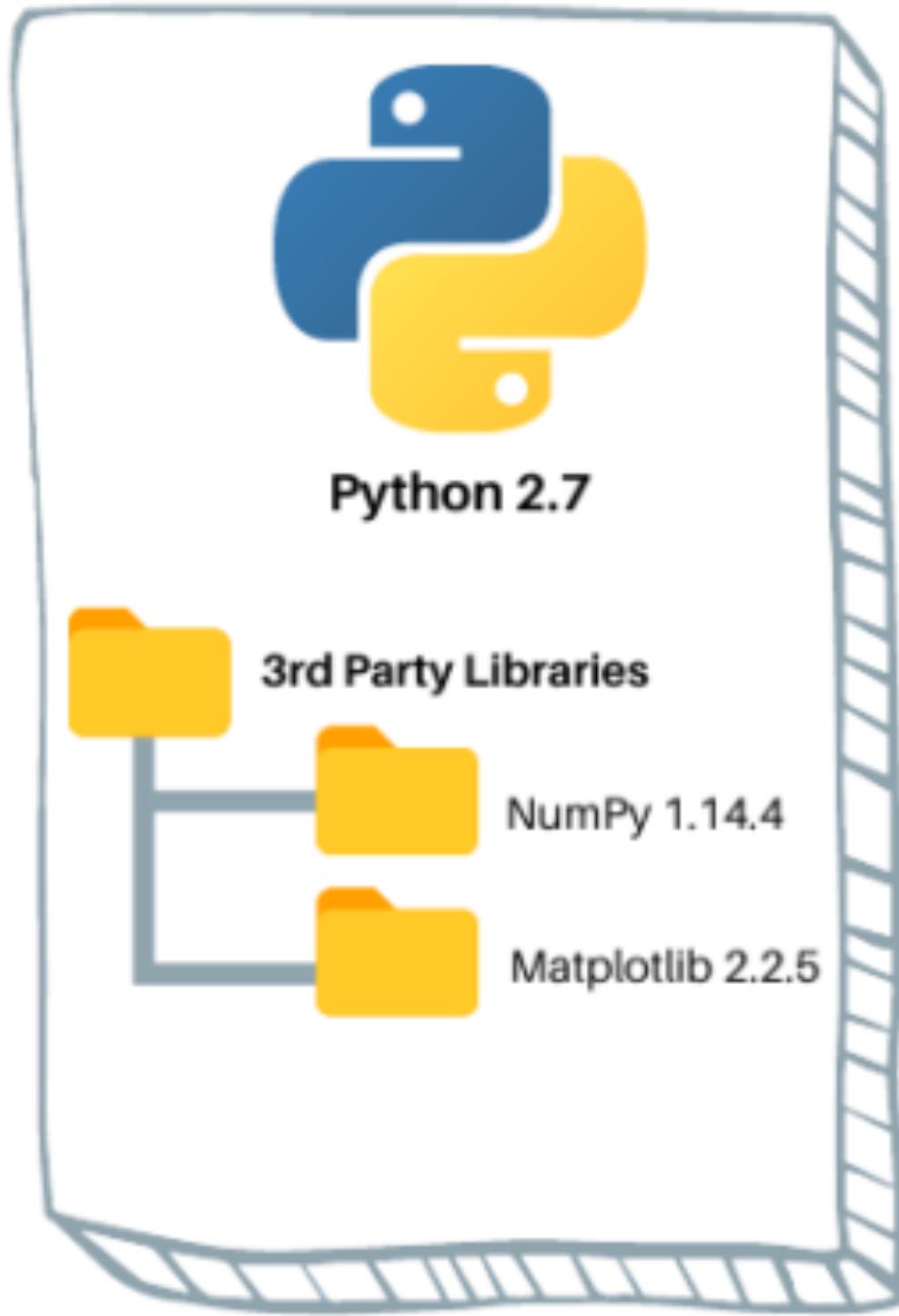
# pyenv

---

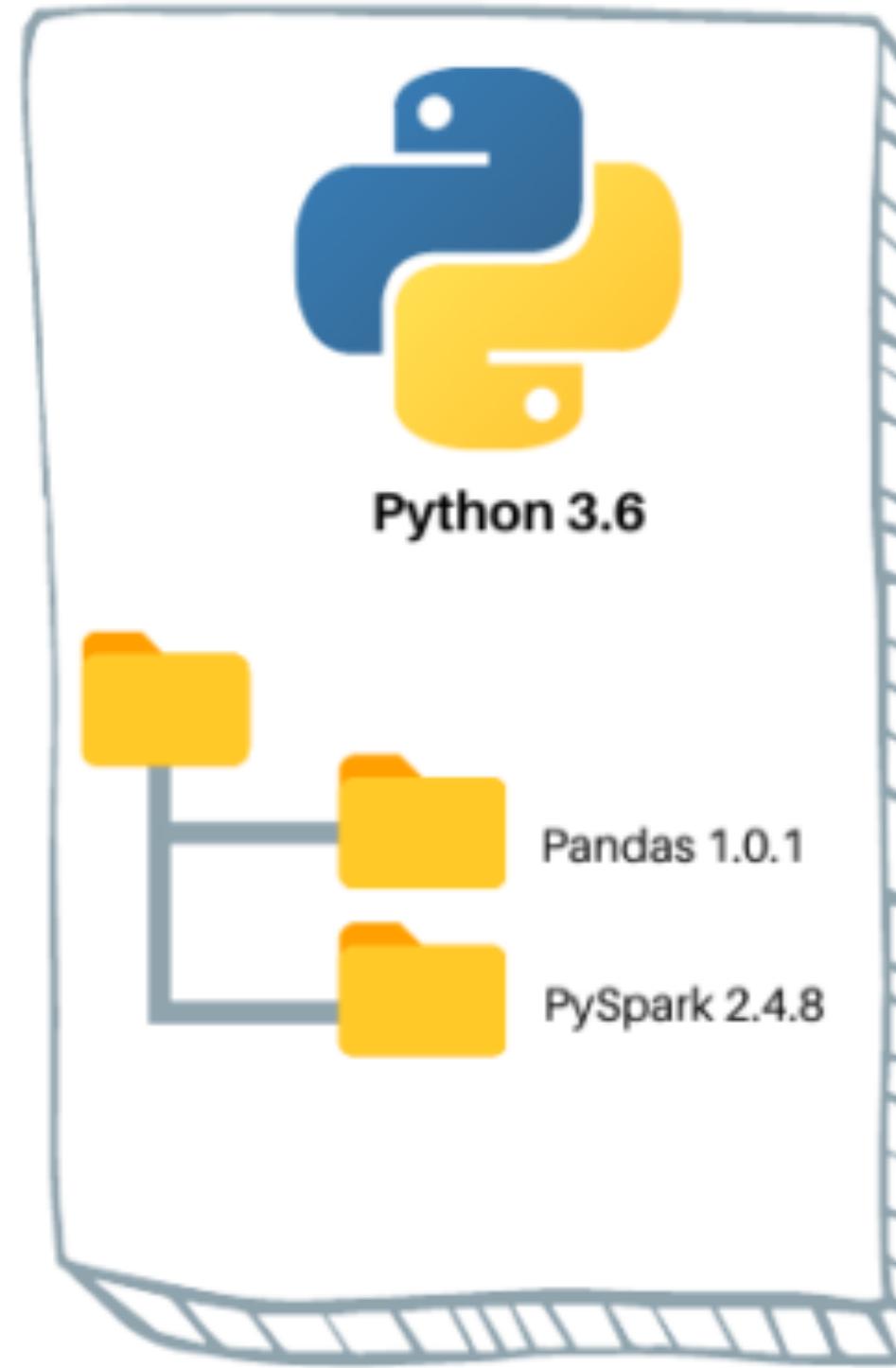
- Simple Python Version Management: pyenv
- <https://github.com/pyenv/pyenv>

# venv

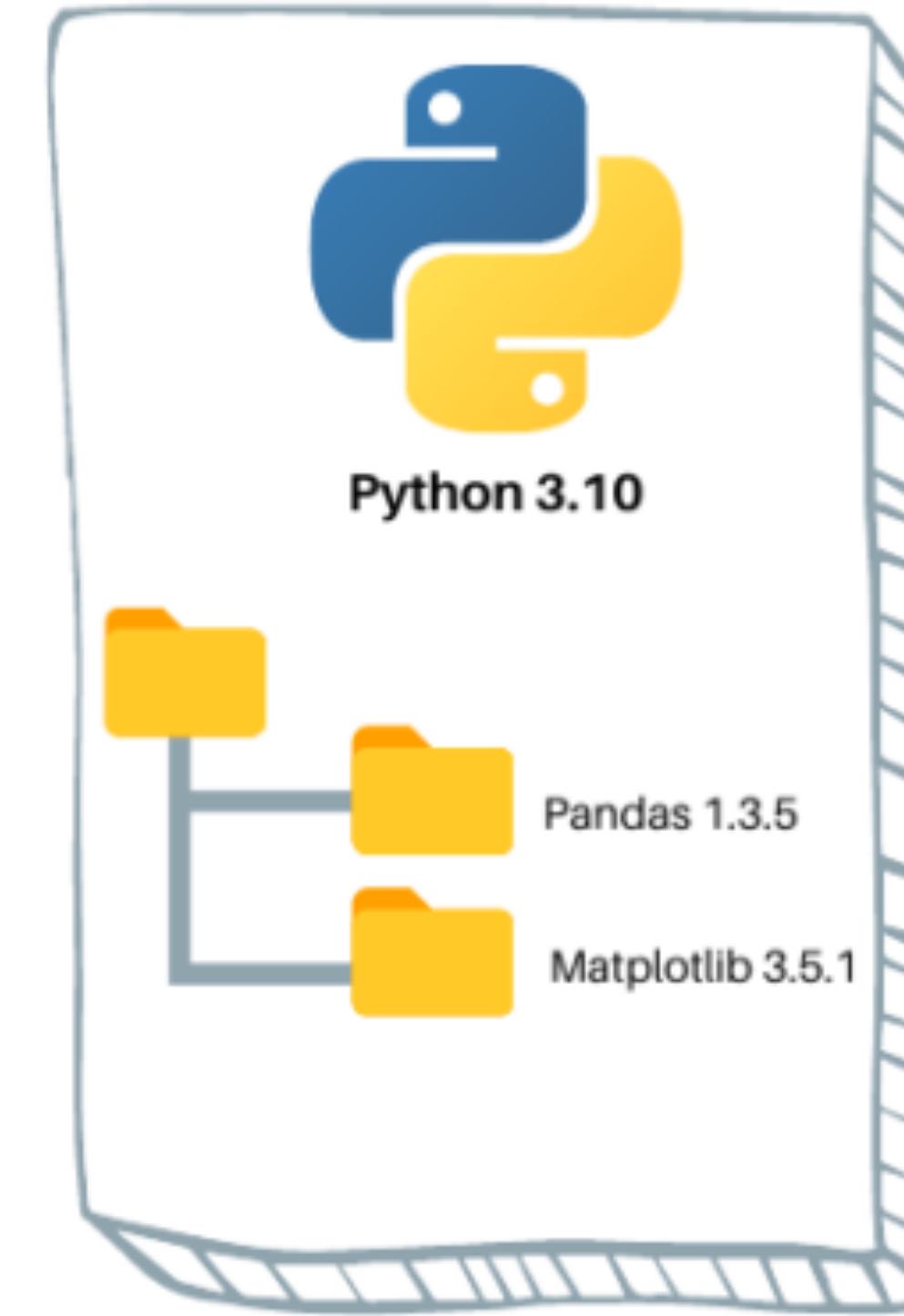
Virtual Environment 1



Virtual Environment 2



Virtual Environment 3



---

# GitHub

---

- GitHub
  - personal repositories: wiki에 모든 리서치한 내용들을 담으세요.
    - missions 폴더 아래에 주차별 폴더를 만들고 과제 파일을 담으세요. ex)W1, 1주차
- git
  - .gitignore: code로 관리하지 않아도 되는 것들. ex) .venv
- GitHub desktop (optional)

# Tools for Data Analysis

---

# JupyterLab and Jupyter Notebook

---

- <https://jupyter.org/>
- **JupyterLab** is the latest web-based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extensions to expand and enrich functionality.
- The **Jupyter Notebook** is the original web application for creating and sharing computational documents. It offers a simple, streamlined, document-centric experience.

---

# Pandas

---

- pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.
- refer to [API reference](#)

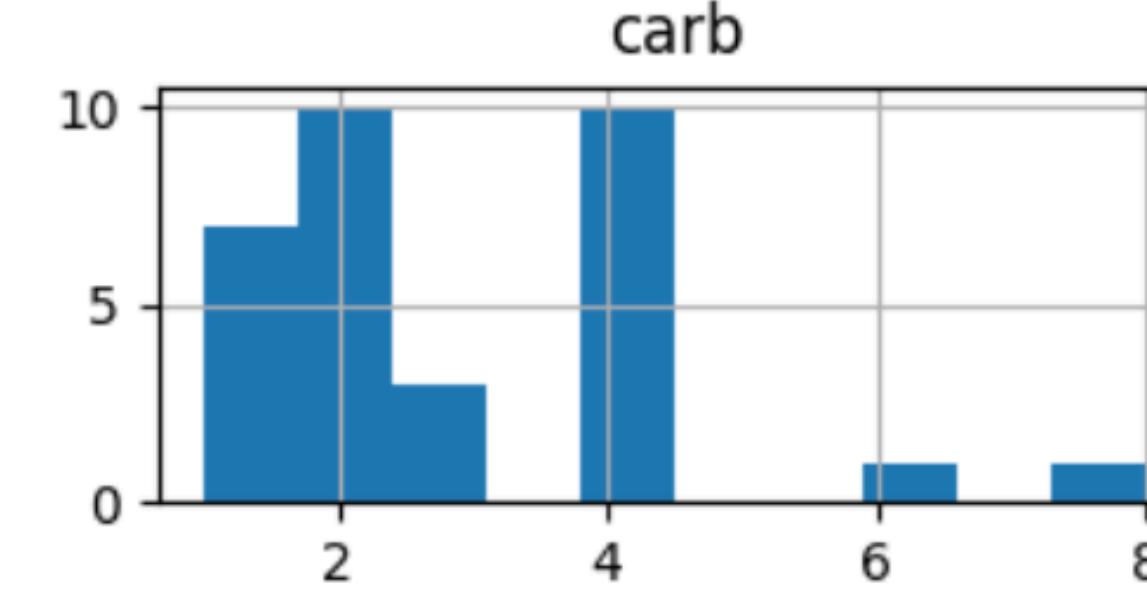
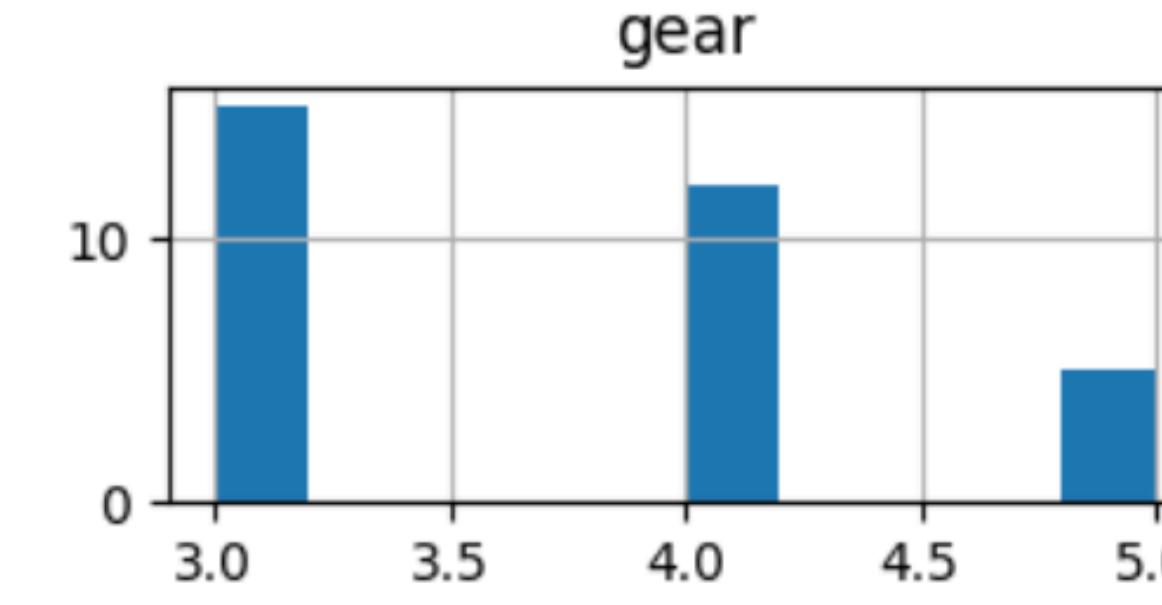
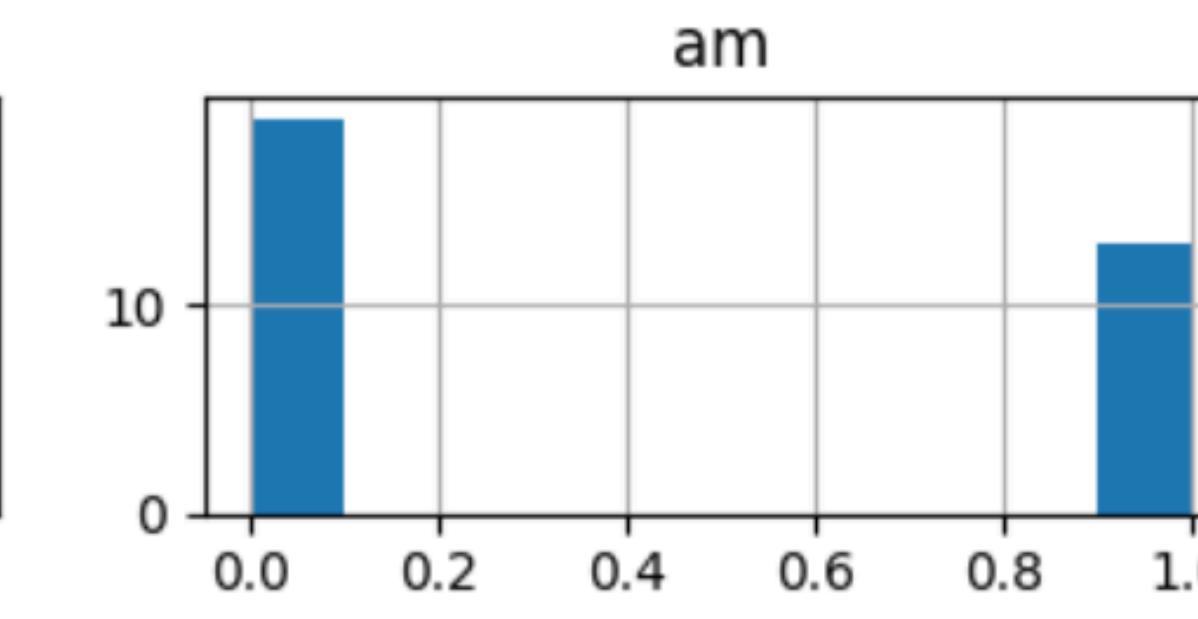
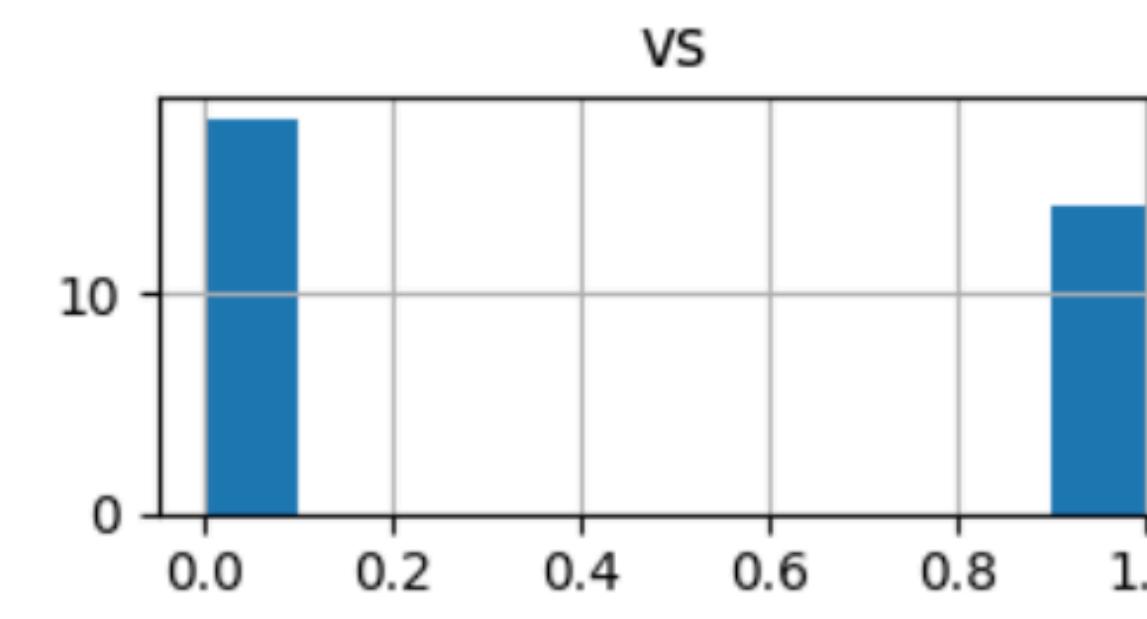
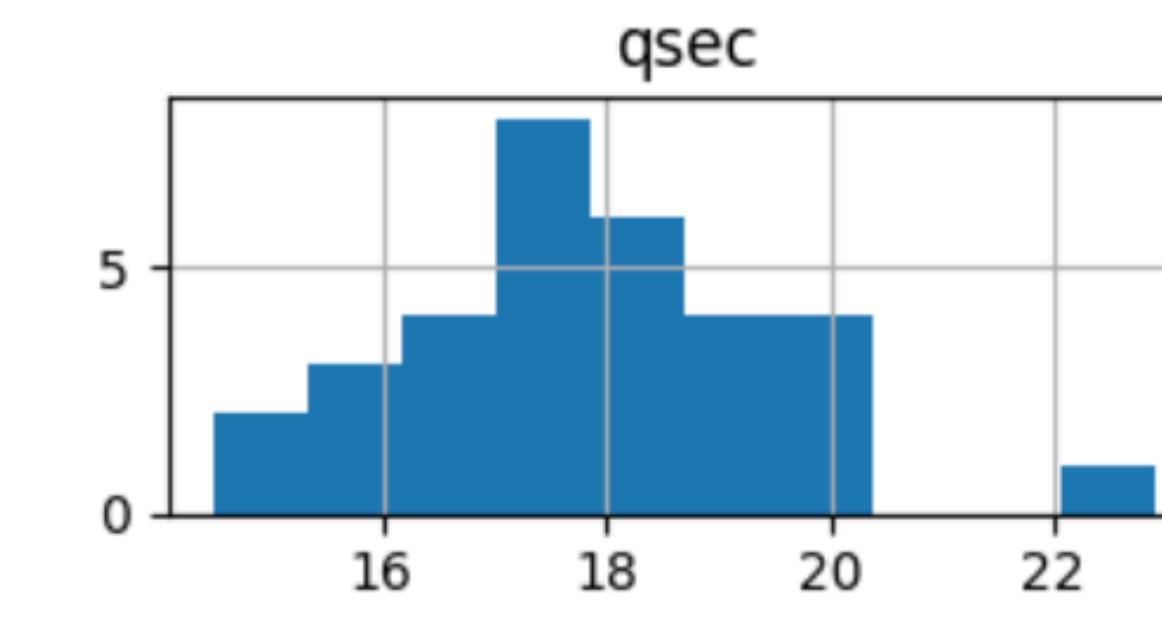
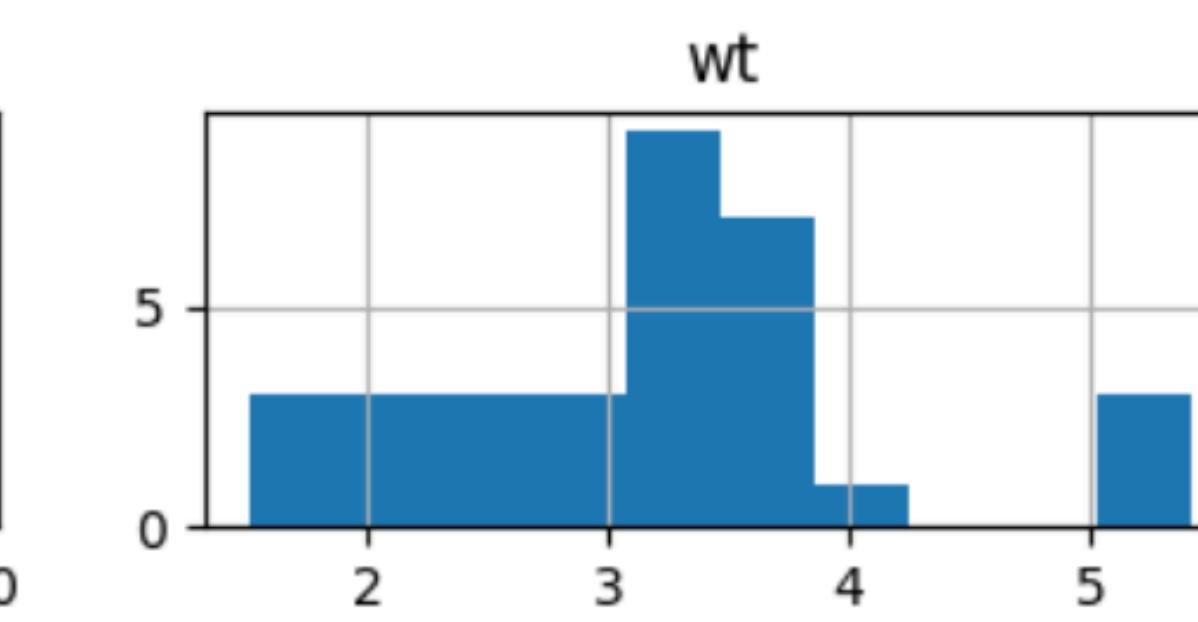
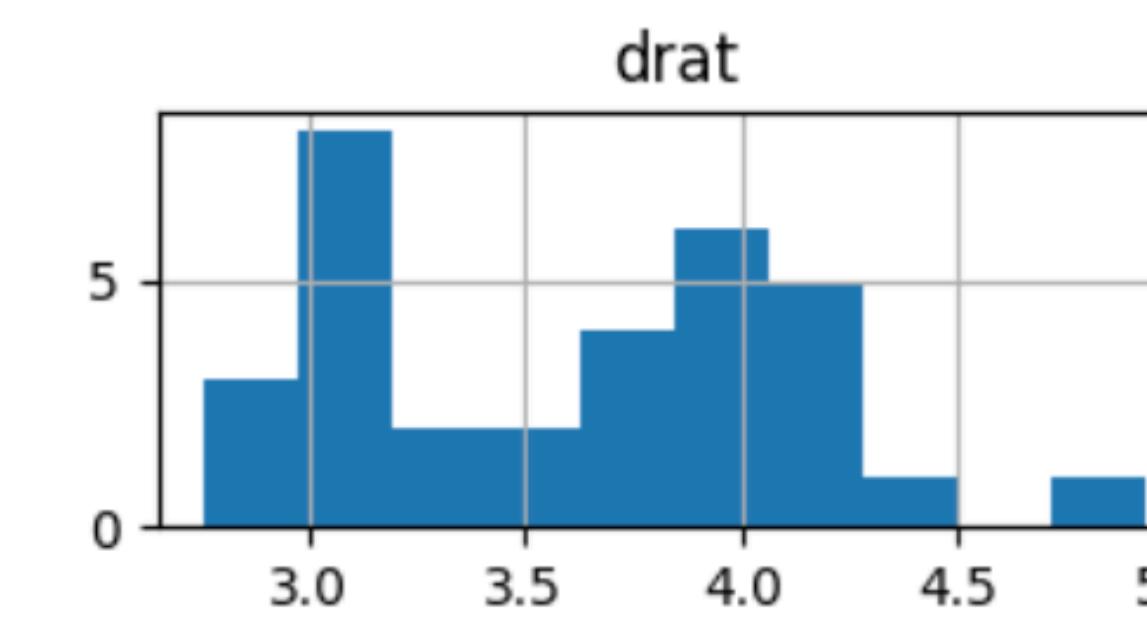
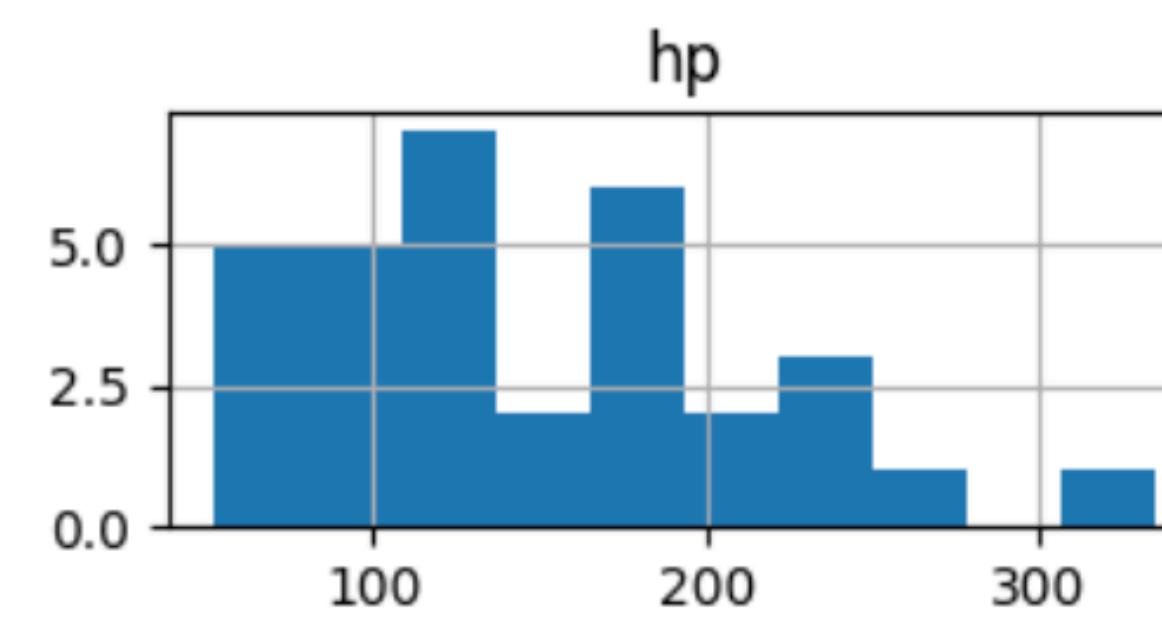
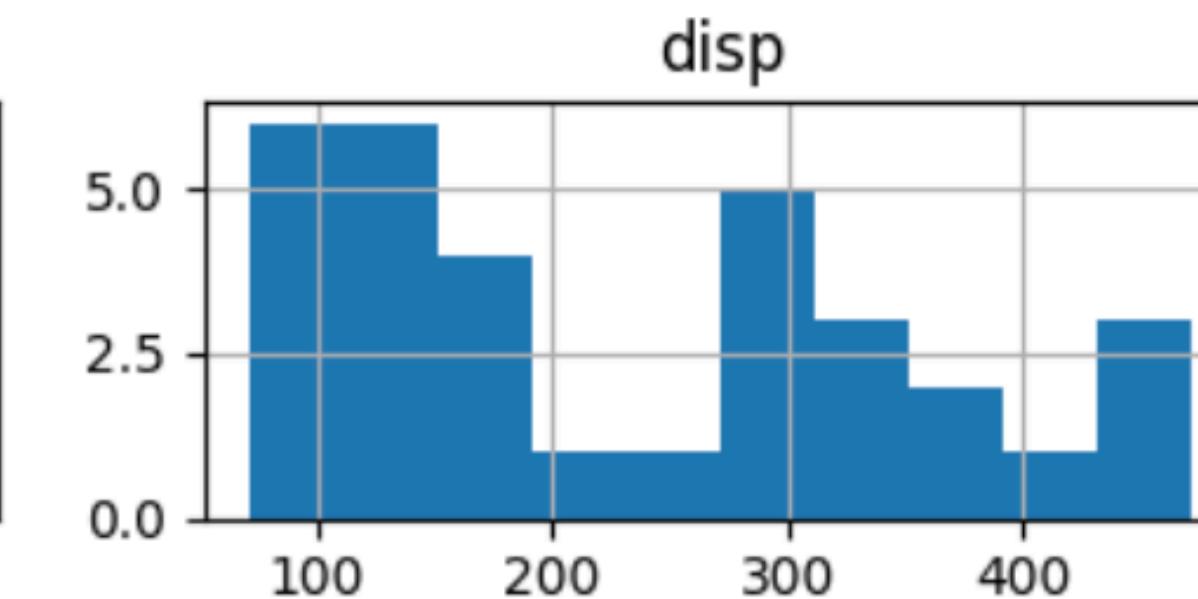
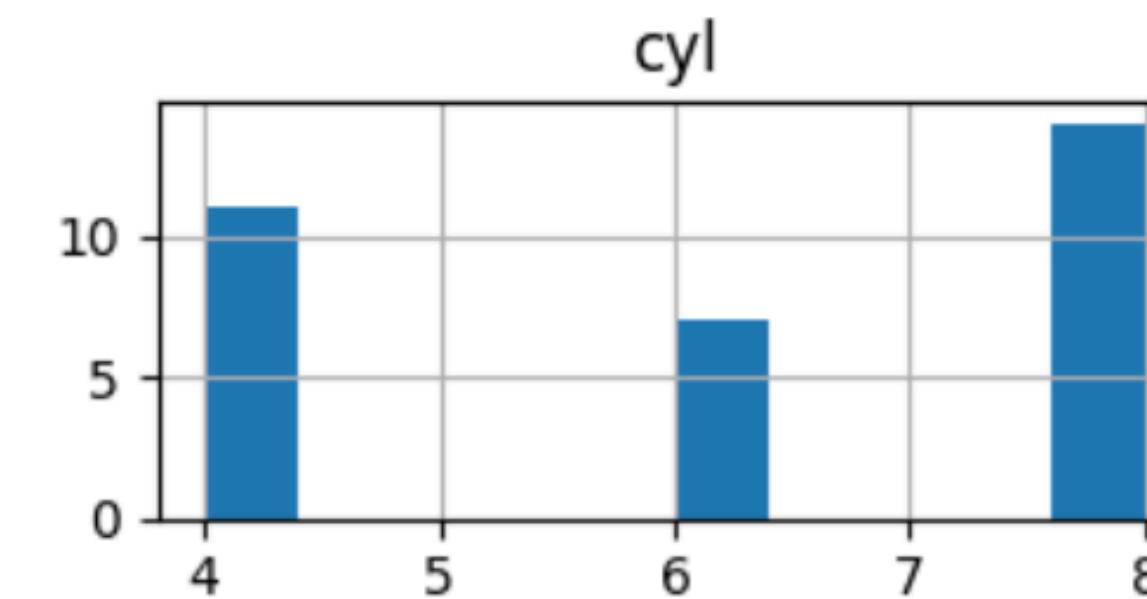
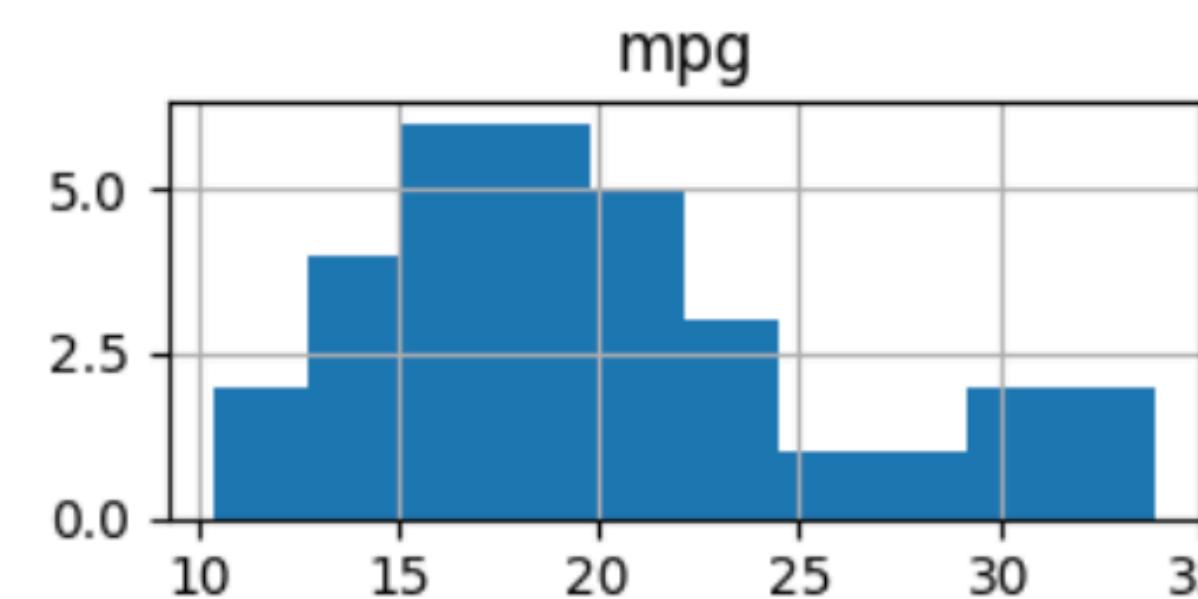
---

# Matplotlib

---

- Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.
- refer to [API reference](#)

# Histograms of Variables



Ref: matplotlib