ID430B: Data Analytics for Designers 디자인 특강V <디자이너를 위한 데이터 분석>

# Lecture 1
# Course Introduction

Tak Yeon Lee <takyeonlee@kaist.ac.kr> (takyeonlee.com)
AI-Experience-Lab (reflect9.github.io/ael)

In this course students will

1. Learn basic concepts of data analytics and their applications in the context of designing products / services

2. Get their hands dirty to collect, clean, transform, analyze, and visualize data

3. Understand common limitations of data and how to fight against them.

# Draft Outline

Please note that this schedule is not fixed, and subject to change at any point

| Week | Topic |
|---|---|
| 1 | Introduction<br>Data Literacy<br>Rationales and Examples |
| 2 | Handling Data Files (CSV and JSON) |
| 3 | Data Cleaning and Transformation |
| 4 | Data Types<br>Descriptive Statistics 1 |
| 5 | Descriptive Statistics 2 |
| 6 | Data Visualization 1 |
| 7 | Data Visualization 2 |
| 8 | Mid-term exam |

| Week | Topic |
|---|---|
| 9 | Probability<br>Bayes' Theorem<br>Root cause analysis on UX |
| 10 | Information Theory |
| 11 | Basic Machine Learning Models |
| 12 | Graph data<br>Journey Analysis |
| 13 | Natural Language Processing<br>Bag-of-words model; Vector Embedding<br>Word Cloud |
| 14 | Limitations of Data<br>How to fight against the limitations |
| 15 | Review of the course |
| 16 | Final Exam |

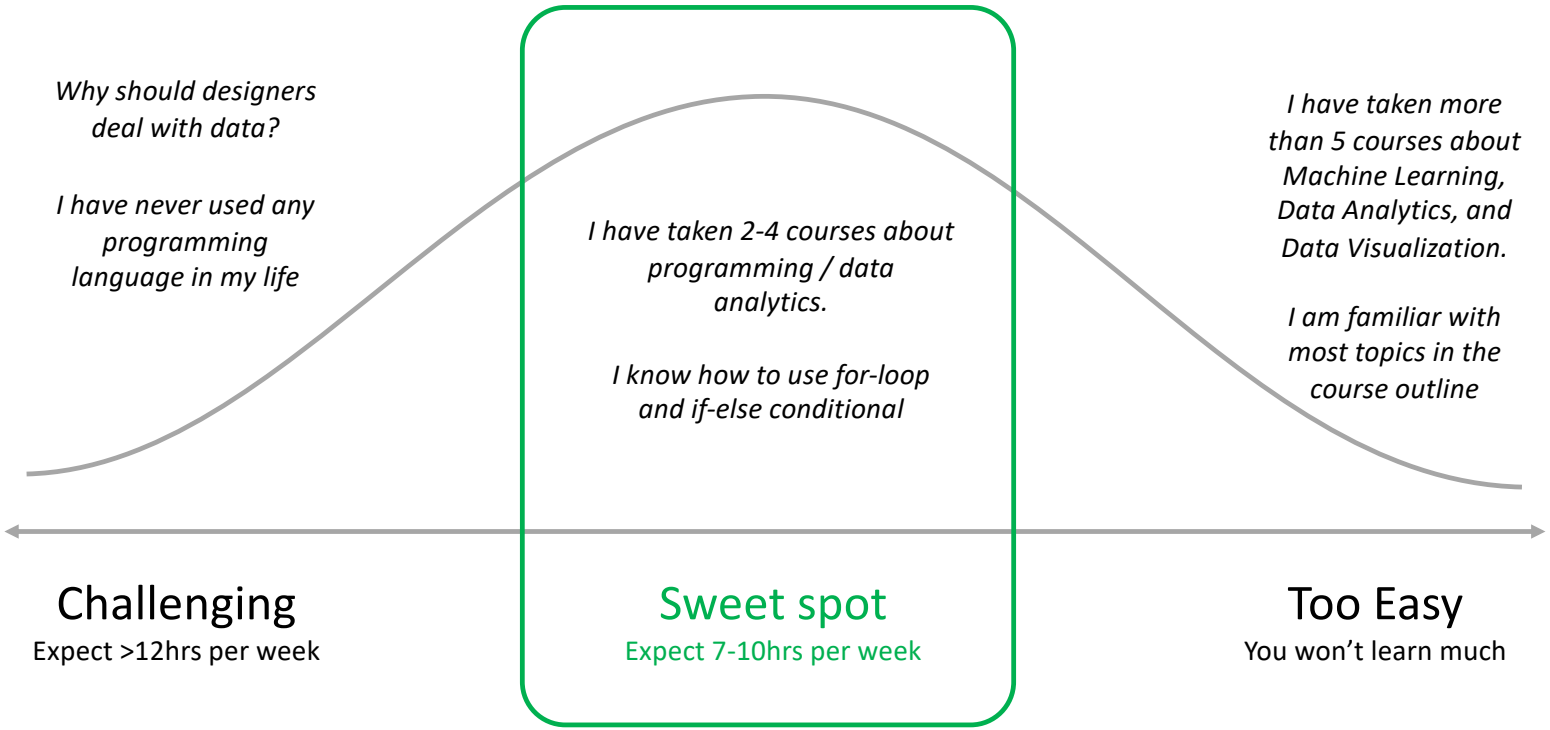
# Educational Philosophy

1. Introducing a wide range of topics with little depth

2. Minimal use of libraries / More hands-on practices

3. Repetitions and repetitions

# Never Too Late To Drop



# Gradings

- Assignments: 42% (+14% extra)
  - There will be 14 assignments (1 per week except mid-term / final exam weeks).
  - You may expect approximately 3% per each assignment, but exact percentages will be decided when they are out.
  - For each assignment you can earn 1% extra point if you solve challenge questions.
- Mid-term exam: 28% (+5% extra)
- Final exam: 28% (+5% extra)
- This course is not curved. We will follow the standard grading system as below (as much as possible).
  - A-(90<=score<93%), A0(93%<=score<97%), A+ (97%<=score<114%)
  - B-(80<=score<83%), B0(83%<=score<87%), B+ (87%<=score<90%)
  - Ranges of C and D are 10% and 20% below the ranges of B, respectively
  - F is below 60%

# Sample Assignments

Topic: Data Transformation

## Regular Question (20-50 per week)

Create a method that count positive numbers in the input. For instance, given [5,3,-2,-1,1,6] the method should return 4.

**Solution:**

```
def count_pos(INPUT):
  count = 0
  for n in INPUT:
     if n >= 0:
        count = count + 1
  return count
```

## Challenge Question (1-3 per week)

Our log data contains multiple user segments (e.g. S1, S2, …, Sn). Each segment contains a list of numbers that indicate how many times an individual user opened our app. Our business team told us that an active user would have opened the app more than *m* times (*m* is not specified yet). Your job is to create a method to find out segments that have at least *p%* of the total users in the segment opened our app at least m times.

In short, for a given list-of-list, count how many lists contain at least *p%* numbers above *m*.

Solution:

```
def find_active_segment(log, k, p):
    active_segments=[]
    for s,l in log.items():
        aun = sum([1 if v>k else 0 for v in l])
        aup = aun / len(l)
        if aup*100 > p:
            active_segments.append(s)
    return active_segments

log = {
    's1': [1, 3, 4, 1],
    's2': [5, 6, 2, 1, 2],
    's3': [1, 1, 1, 1, 1, 4, 4, 4, 4, 4]
    }

print(find_active_segment(log, 2, 40))
```

Output: `['s1', 's3']`

# Sample Assignments

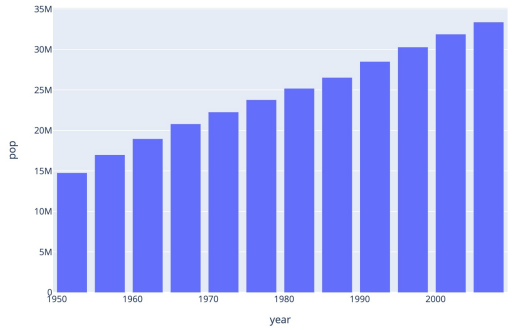
Topic: Data Visualization

Regular Question (20-50 per week)

Draw a simple bar chart

**Solution:**

```
import plotly.express as px
fig = px.bar(data, x='year', y='pop')
fig.show()
```

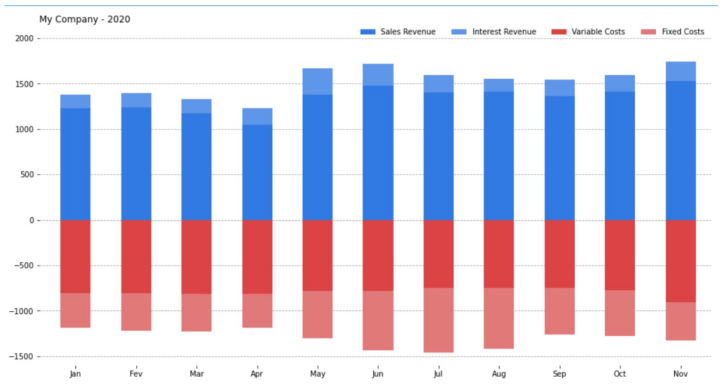



Challenge Question (1-3 per week)

Draw a stacked bar chart aligned at zero.

**Solution:**

Last example on this page



# Logistics

- KLMS is the hub for all the materials
- Materials of each week will be out on each Monday (exceptions may happen)
- Late submissions will be deduced by 10% per day. However, homework submitted after 3 days (72hrs) from the deadline will get zero point
- Academic Integrity and Collaboration
  - You may discuss assignments with others, but you should always give credit and be intellectually honest. For all individual assignments, you should write the solution entirely on your own. Sharing or seeing other students' solutions (written material or code) is not allowed.
  - You may not use any third-party libraries (unless explicitly mentioned in the instruction).
  - Failure to adhere to these policies may lead to serious penalties, including an F in the course and reference to the departmental and university committee.Every homework must be done by yourself. Students may discuss about the topic but not the answer.
- Office hours are 19:00 - 20:30 on Tuesday and Thursday
  - Zoom Link. https://kaist.zoom.us/j/8827121106
  - TA: Seon Gyeom Kim <ksg_0320@kaist.ac.kr>
  - Instructor: Tak Yeon Lee <takyeonlee@kaist.ac.kr>

**Tak Yeon, Lee**
Assistant Professor
takyeonlee@kaist.ac.kr

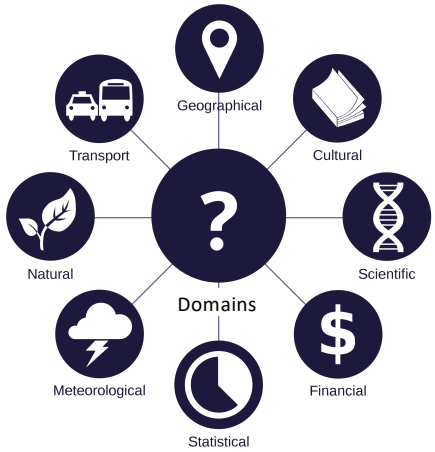
**Seon Gyeom, Kim**
Master Student
ksg_0320@kaist.ac.kr

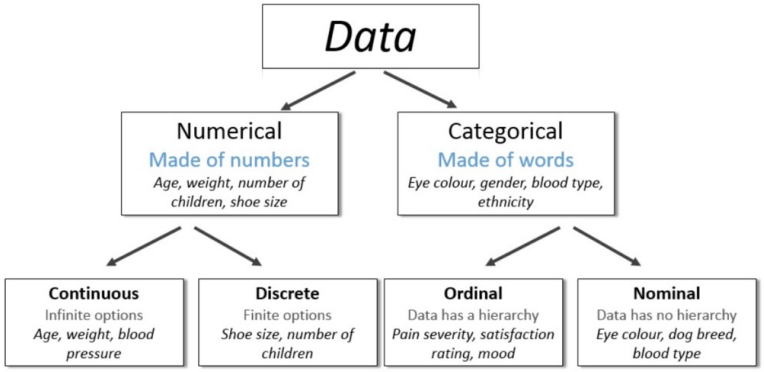
# Data Literacy

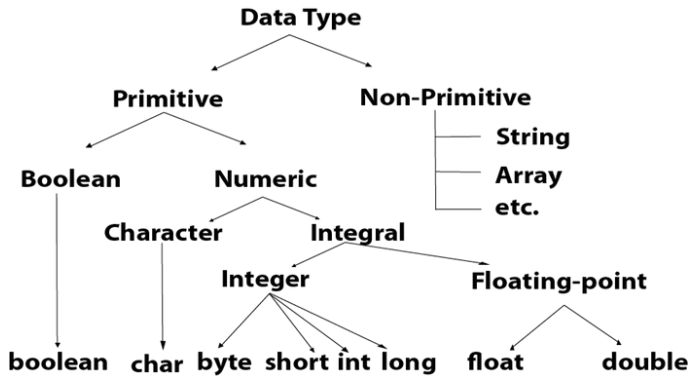
…is the ability to read, understand, create, and communicate data as information - *Wikipedia*

# What is Data?

- "**Factual information** (e.g. measurements, statistics) used as a basis for reasoning, discussion, or calculation" - *Merriam Webster dictionary*
- "Data are individual facts, statistics, or items of information, often numeric, that are **collected through observation**. More technically, data is a set of value of **qualitative / quantitative variables**" - *OECD Glossary of Statistical Terms*
- "Data are **measured, collected, reported**, and **analyzed**, and used to create data visualizations such as graphs, tables or images. Data as a general concept refers to the fact that some existing information or knowledge is represented or coded in some form suitable **for better usage or processing**." - *Wikipedia*



# Quality of Data

These are not hard rules but potential limitations of the dataset that analysts should keep in mind

Is it **factual**?

Does the dataset contain any non-factual values such as theories or interpretations?

Is it **complete**?

Does the dataset include values for all the fields required by our analysis?

Is it **reliable**?

Will it be consistent if we measure tomorrow (or under other consistent conditions)?

Is it **unbiased**?

Is it collected from the (entire/randomly/evenly sampled) population of our interest?

Is it **non-redundant**?

Does it contain any duplicates or dummy records?

Is it **accurate**?

If no, what are types of noise? How can we adjust them?

Is it **relevant**?

Does it help us get wisdom at the top of the DIKW model?

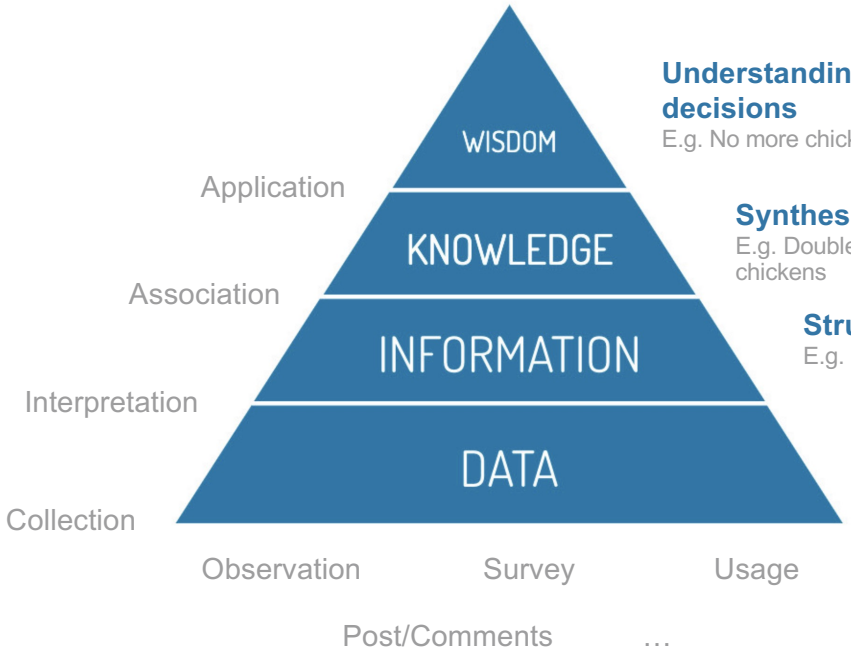
# DIKW Model

*Where is the **Life** we have lost in **living**?*

*Where is the **wisdom** we have lost in **knowledge**?*

*Where is the **knowledge** we have lost in **information**?*

from "The Rock" by T.S.Eliott, 1934




**Understanding, Integrated, and Actionable decisions**
E.g. No more chicken delivery on night

**Synthesized to answer "why"**
E.g. Doubled spending in the last month on chickens

- Knowledge is information connected with other pieces.
- How can we apply the information to achieve our goal?
- What comparisons can we make to help us better grasp the relevance of the data?

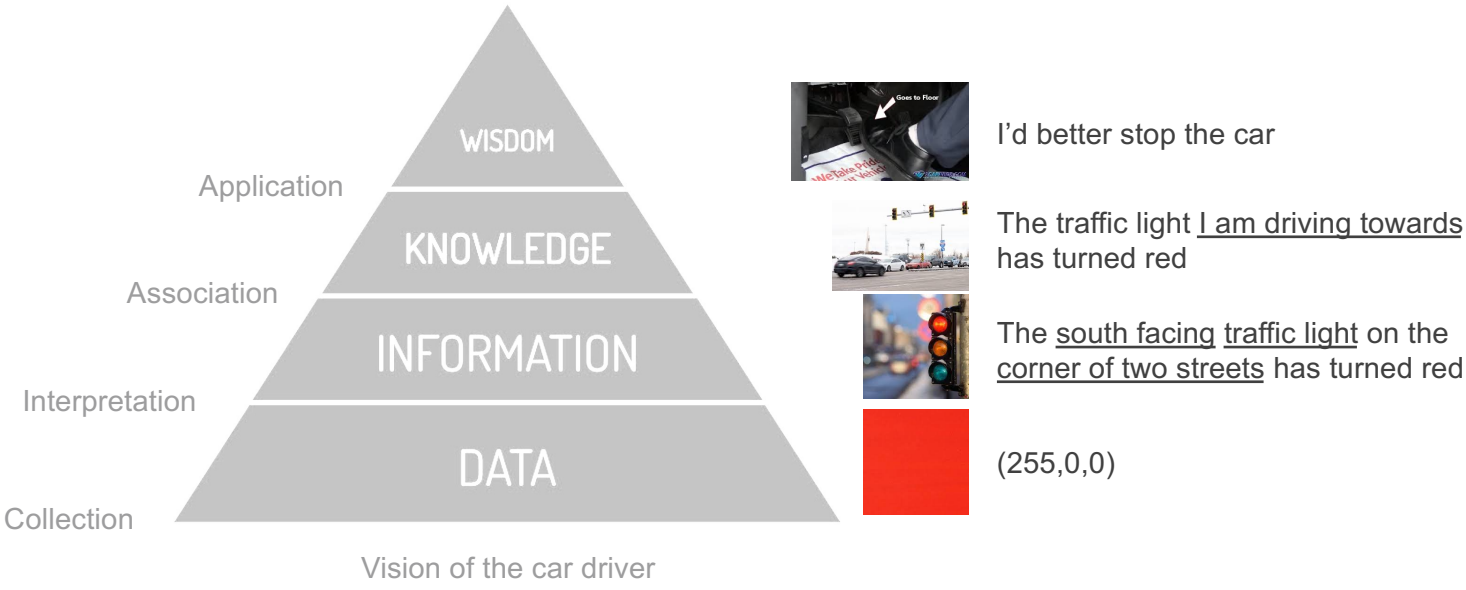
**Structured and contextualized**
E.g. Balance: $320, $220, $20
Spending: $100, $200

Who, What, When, Where? This is data that has been cleaned or errors and processed to make it easier to measure, visualize, and analyze for a specific purpose.

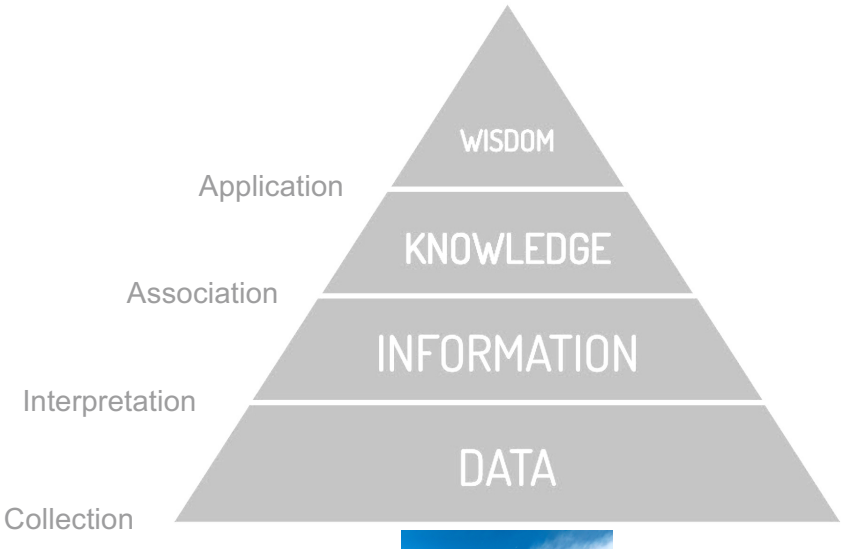
**Discrete, objective facts about a phenomenon**
E.g. 320, 220, 20

Data is a collection of facts in a raw or unorganized form such as numbers or characters. Check if the dataset is factual, complete, reliable, unbiased, non-redundant, accurate, and relevant

# E.g. Stopping car at the red light



# E.g. Mount Everest




Mount Everest

The practical climbing of Mount Everest's peak
E.g. "*We must retreat or die*"

A climber's guidebook containing practical information on the best way to reach Mount Everest's peak
*E.g. Be careful at the top of the mountain in September*

Geographical characteristics of Mount Everest
*E.g. In September the top of the mountain is highly likely to get storms*

Measured facts of Mt. Everest
*E.g. Elevation, Rates of seasonal storm, etc*

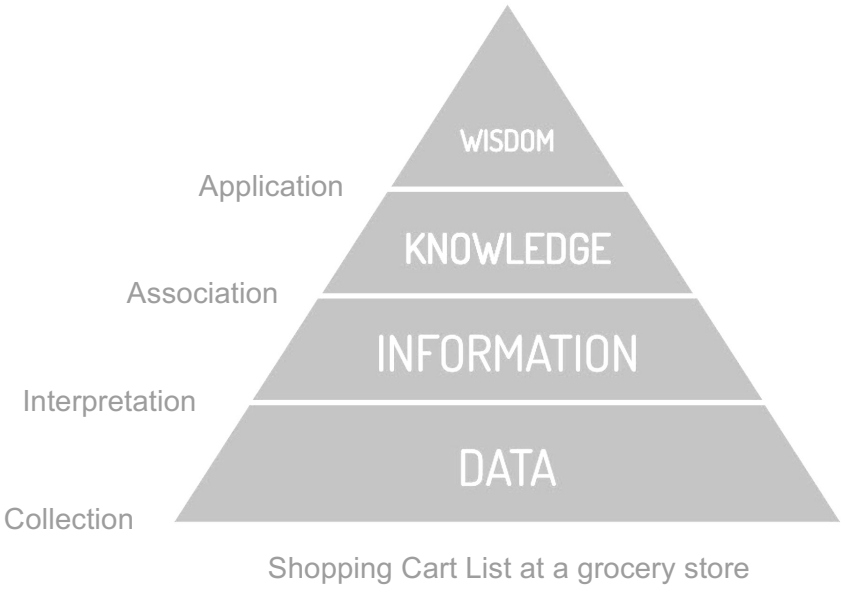
*This example was from "data" in Wikipedia*

Abstract, Integrated



Concrete, Individual

# E.g. Sales Manager's Job


To increase our sales, we can build a special section for parents who want to relax from child-caring.

The co-purchase pattern was surprisingly high. A follow-up survey also tells that majority of the co-purchasing was done by 30s-40s males during evening hours.

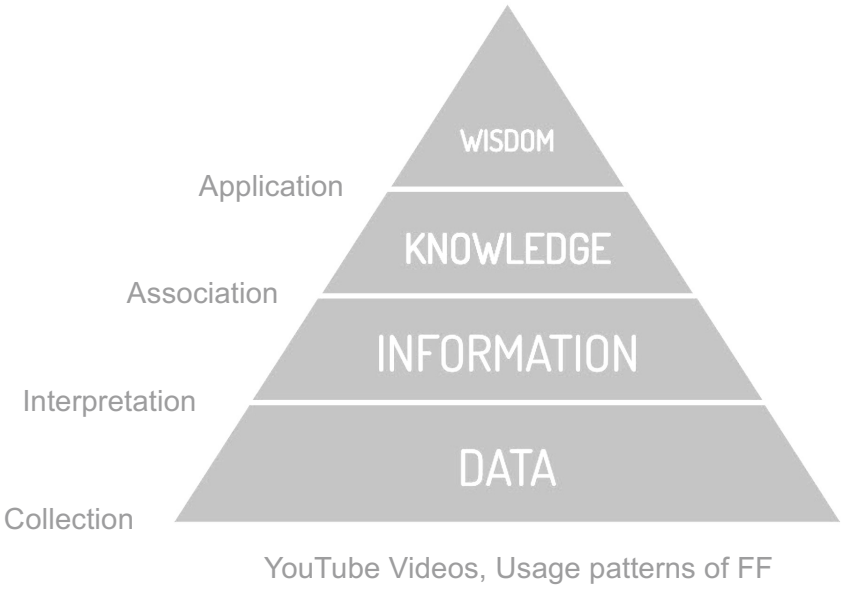
42% of the customers who bought beer also bought diapers

CUSTOMER 1  ...

CUSTOMER 2  ...

⋮

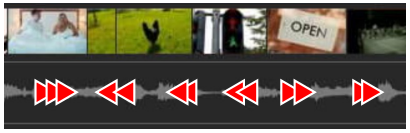
# E.g. YouTube Usage Pattern




Let's redesign the video player to jump to the next / prev scene when user presses right / left button respectively

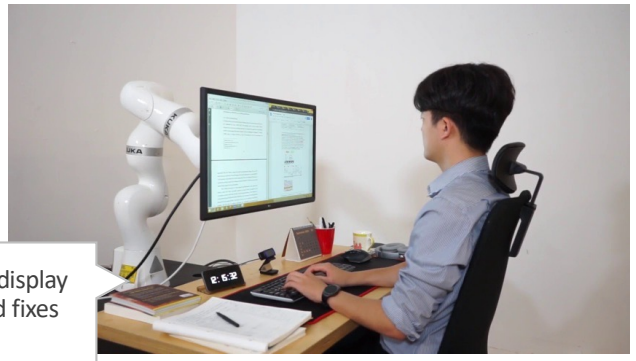
People tend to press buttons to jump forward / backward when they want to move to the next scene.


(1) A sequence of scenes that are detected from the data
(2) Frames that users restart watching the video after jumping back and forth



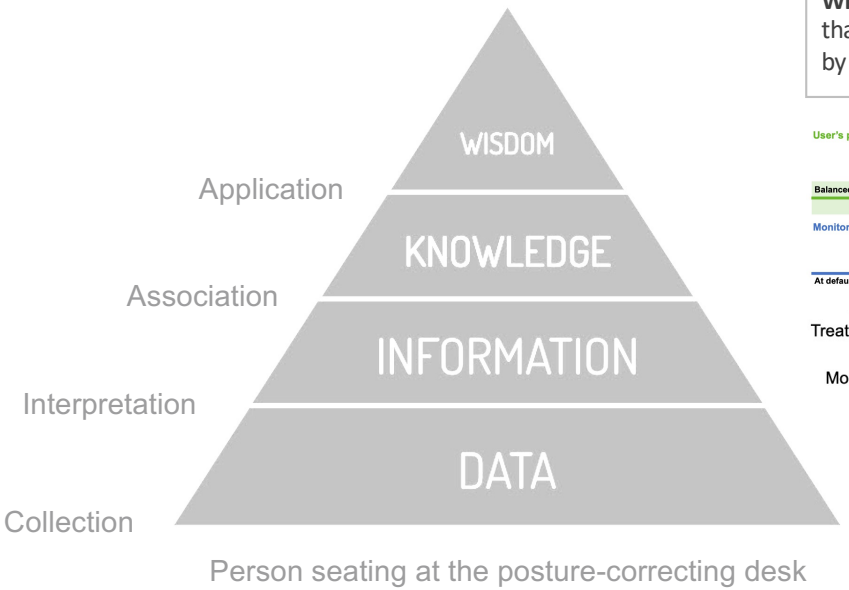

(1) Video, audio, and subtitle information of each frame
(2) When users press left/right button to jump 10s forward / backward

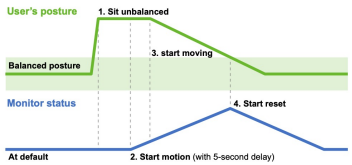
# E.g. Correcting Unbalanced Posture



**WISDOM:** "Let's make an intelligent display that detects unbalanced posture, and fixes by moving its position."

Slow Robots for Unobtrusive Posture Correction. Joon Gi Shin, Eiji Onchi, Maria Jose Reyes, Junbong Song, Uichin Lee, Seung-Hee Lee and Daniel Saakes.



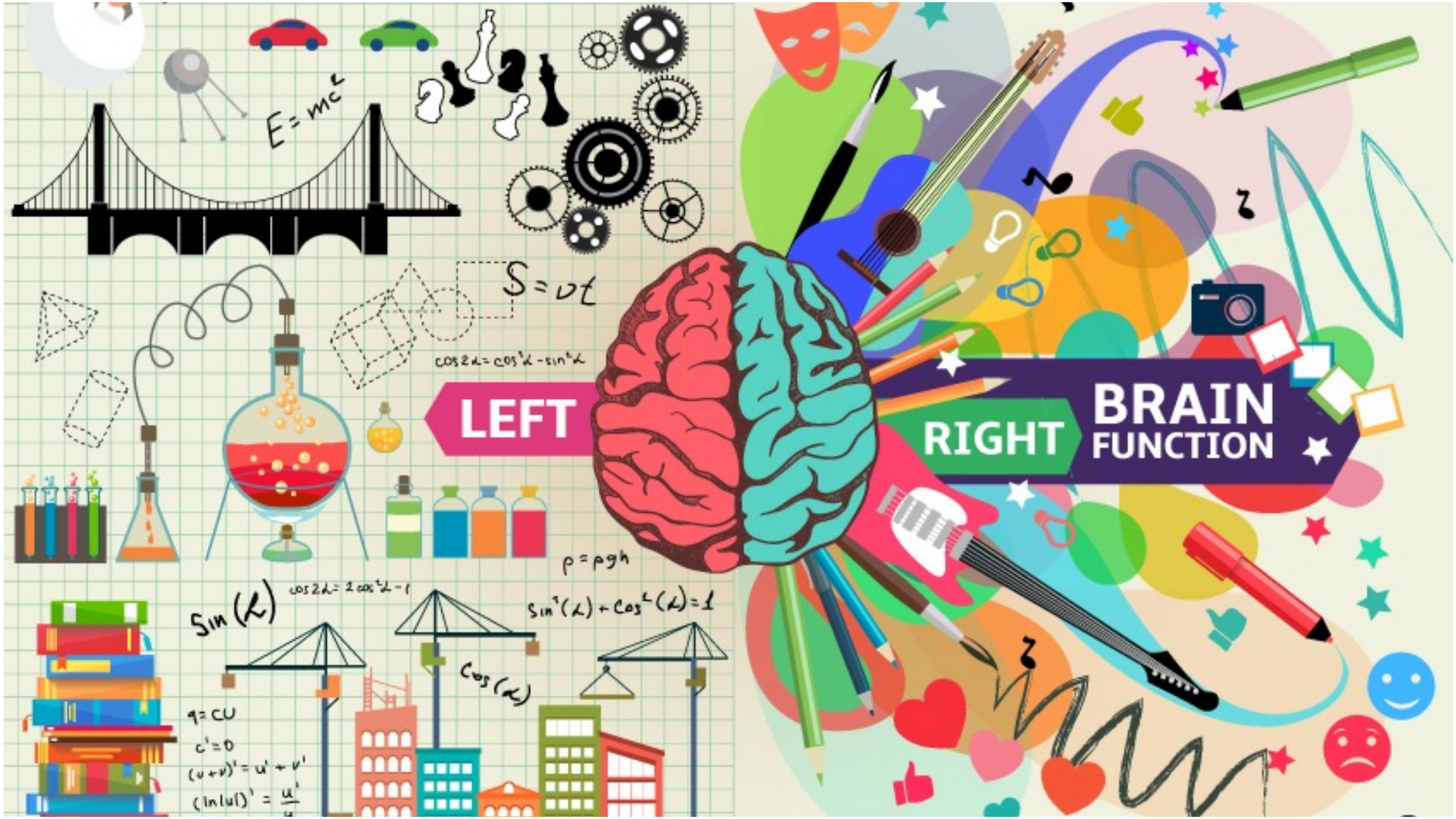


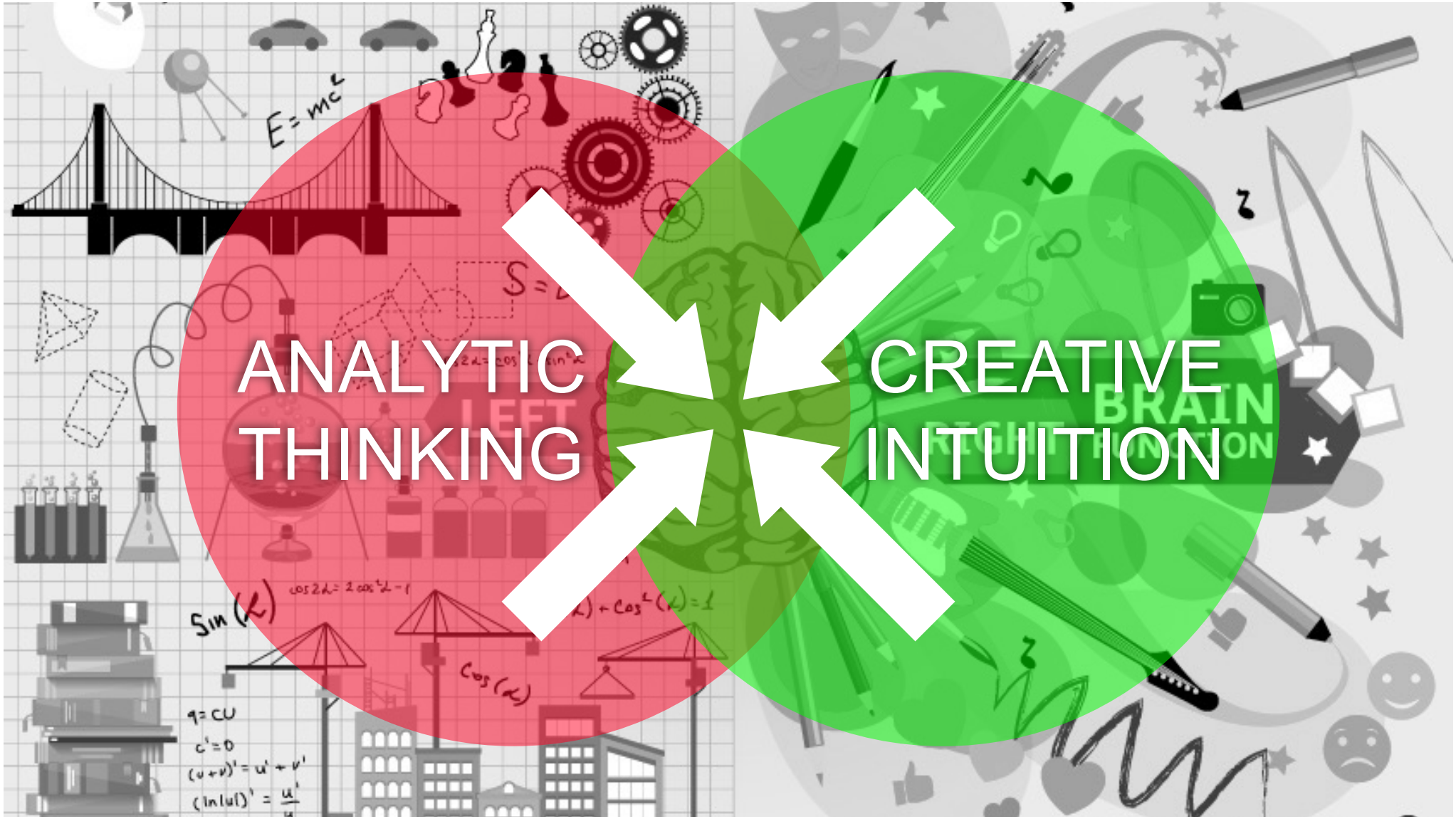
"Slow motion of display can fix user's unbalanced postures"

Timeline of the two dataset (display motion and person's behavior after treatment)

(1) Motion data of the moving display
(2) Person's behavioral data captured by cameras from different angles

# How Analytics and Design Help Each Other?


LEFT
RIGHT
BRAIN
FUNCTION


ANALYTIC
THINKING
CREATIVE
INTUITION

# Why Analytic Thinking is important for designers

## 1. Analytic thinking gives designers new business opportunities

Digital Transformation has created a lot data-centric businesses. Analytic Thinking is the best way for designers to find new opportunities. For instance, designers equipped with AT skills can design and build their own conversational agents, recommendation systems, and other AI-powered services. In addition, data literacy (a basic AT skill) is the key to communicate with engineers and business experts

## 2. Analytic thinking help designers improve creative intuition

While being a strong asset of designers, creative intuition could be biased or fixated based on prior knowledge. Analytic thinking is one of the best ways to test and fix biases in designer's creative intuition.

# Why Creative Intuition is important for data analytics

1. Knowing what the data is telling / not telling us
2. Knowing where the look next
3. Knowing when to stop looking and take action
4. Knowing who needs to hear and how to get through to them
5. Knowing why any of it matters in the first place

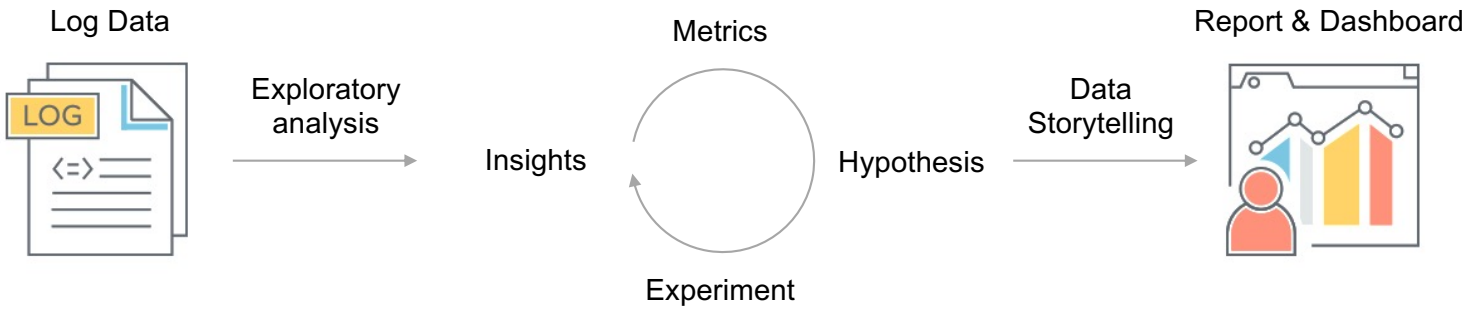
**While analytic thinking is hands-on tactical skills,**
**Intuition helps us make high-level strategic decisions**

# Common Patterns of Design + Analytics

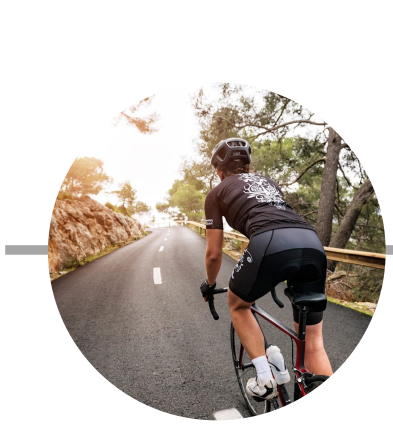
# Log Analysis beyond off-the-shelf tools

Log analysis is a common method to understand user behavior from . While off-the-shelf analytic tools (e.g. Google Analytics) allows to get basic insights, designers want much more: (1) exploratory analysis, (2) Gain insights via custom metrics and experiments at scale, and (4) data storytelling.

As results of log analysis, designers usually create reports or dashboard to be shared within the organizations. Going further, insights gained from log analysis can guide designers toward intelligent data-driven services and features.



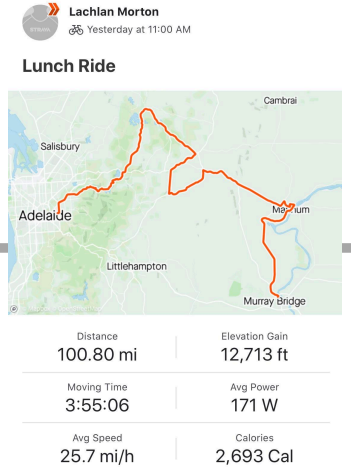
# Designing the entire data pipeline



PERSON
- Age, Sex, Occupation, …
- BMS (Body Mass Index), …

RIDE
- Location: Start, End
- Time: Start, End
- Elevation, Speed, Coord, …
- Heart rate, Pedal angle, …
- Wind, Temperature, …

…




**Real World** | **Data collection** | **Collected Data** | **Data-driven Features**
(e.g. Report, Dashboard, Recommendation, …)

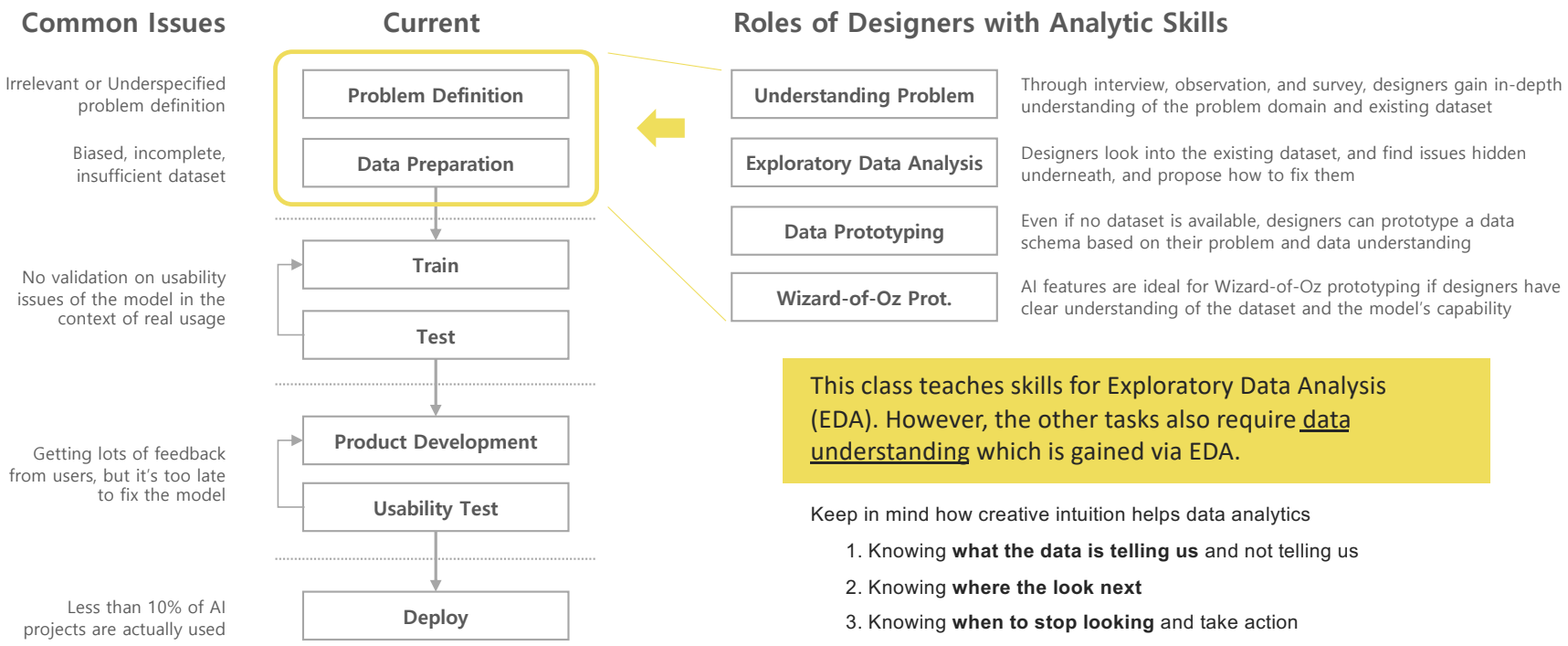
Notice that the entire pipeline can be designed.

CASE 1. If the activity does not exist, designers must envision every step from scratch.

CASE 2. If there's a collected dataset, analysts would first examine collected data to develop data-driven features. If the dataset is not satisfactory, they would try different collection methods. If cyclists' behavior does not have signal of interest, they can even redesign cycling.

Both cases require analytic thinking and creative intuition at the same time.

# Before ~~and After~~ AI Modeling

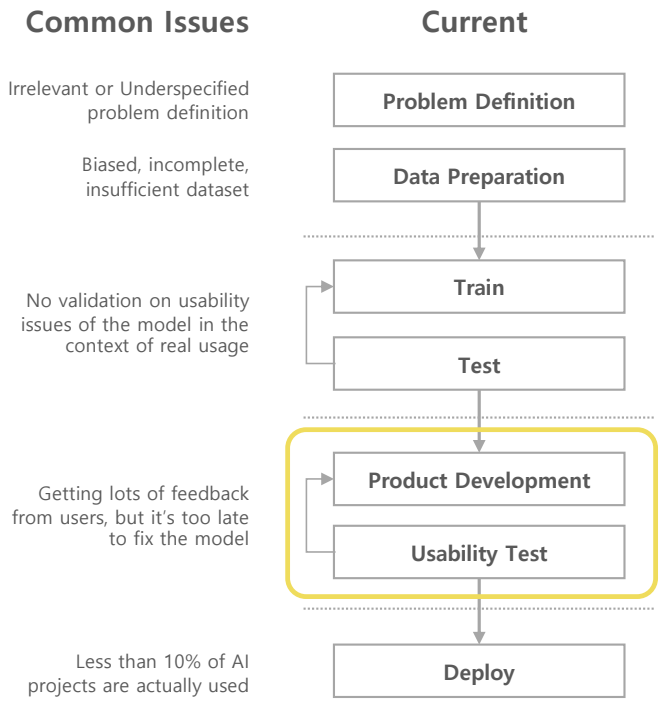



This class teaches skills for Exploratory Data Analysis (EDA). However, the other tasks also require data understanding which is gained via EDA.

Keep in mind how creative intuition helps data analytics

1. Knowing **what the data is telling us** and not telling us
2. Knowing **where the look next**
3. Knowing **when to stop looking** and take action
4. Knowing **who needs to hear** and how to get through to them
5. Knowing **why any of it matters** in the first place

# ~~Before and~~ After AI Modeling




## Roles of Designers with Analytic Skills

| | |
|---|---|
| **Understanding Model Performance** | Most AI models make mistakes, designers must have in-depth understanding of when / how AI models fail. |
| **Make the worst case scenario tangible** | Not only the best but also the worst-case scenarios (i.e. model returning wrong (or even no) results) should be designed |
| **Gatekeeping data-driven features** | If a data-driven feature does not have significant benefits for users, designers would drop / redesign it. |
| **Envision Iterative Cycles** | In usual, data-driven AI-infused systems are not complete at their first versions. Designers propose a roadmap (i.e. iterative cycles of development) rather than finalized systems. |

# Summary

## A. Course overview

A. Goals, Outline, Philosophy, Grading, Sample Assignments, Logistics

## B. Data Literacy

A. Definition of Data

B. DIKW (Data → Information → Knowledge → Wisdom) model and examples

## C. How analytics and design help each other

A. Analytic thinking skills provide new business opportunities, and improve creative intuition

B. Creative intuition helps high-level strategic decisions

## D. Common Patterns of Design + Analytics

A. Advanced Log Analysis

B. Designing the entire data pipeline

C. Contribute before / after AI Modeling