Data Science: Course Overview

Dong-Kyu Chae

PI of the Data Intelligence Lab @HYU
Department of Computer Science
Hanyang University

Data
Intelligence
LAB

Information

- **Instructor: Dong-Kyu Chae**
  - Contact information
    - Email: dongkyu@hanyang.ac.kr
    - Office: Room 616, ITBT
    - Tell: 02-2220-2896

- **Textbook**
  - Jiawei Han, Micheline Kamber, and Jian Pei, **Data mining: concepts and techniques**, Morgan Kaufmann
  - It is optional to purchase the book.
  - All the exam questions will be originated from PPT slides

Grading Scheme

- **Weights on graded parts**
  - Midterm exam: **40%**
  - Final exam: **40%**
  - Programming: **15%**
    - **3 programing assignments**
    - **Only Java and Python must be used**
  - Attendance: **5%**
- **The students who take this class again (재수강) will be able to get at most A0**
- **Grade 'F' will be given if you**
  - copy somebody else's program (i.e., from classmate or from the Internet) or allow others to copy yours
  - do not take both the midterm and final exams
  - get 0 scores for both midterm and final exams

Important Notice

- **This course will be an online lecture**
  - Recorded lecture videos will be uploaded (not a real-time Zoom class)

- **Lecture upload schedule**
  - 1~2 lecture videos for a week. (I will try my best to upload videos before Thursday morning.
  - Next week (3/9): no upload (will be uploaded later)

Assignments: General Information

- **3 programming assignments**
  - Frequent pattern mining: **Apriori**
  - Classification: **Decision tree**
  - Clustering: **DBSCAN**

- **Assignment submission: through an e-mail**
  - The email address will be noticed later

Assignments: General Information

- **Late submission policy**
  - *20% penalty* : after a week
  - *50% penalty* : within 1~2 weeks
  - *Will not be accepted*, after two weeks

- **Requirements unsatisfied**
  - Up to 100% penalty will be given depending on how the requirements are not well-satisfied

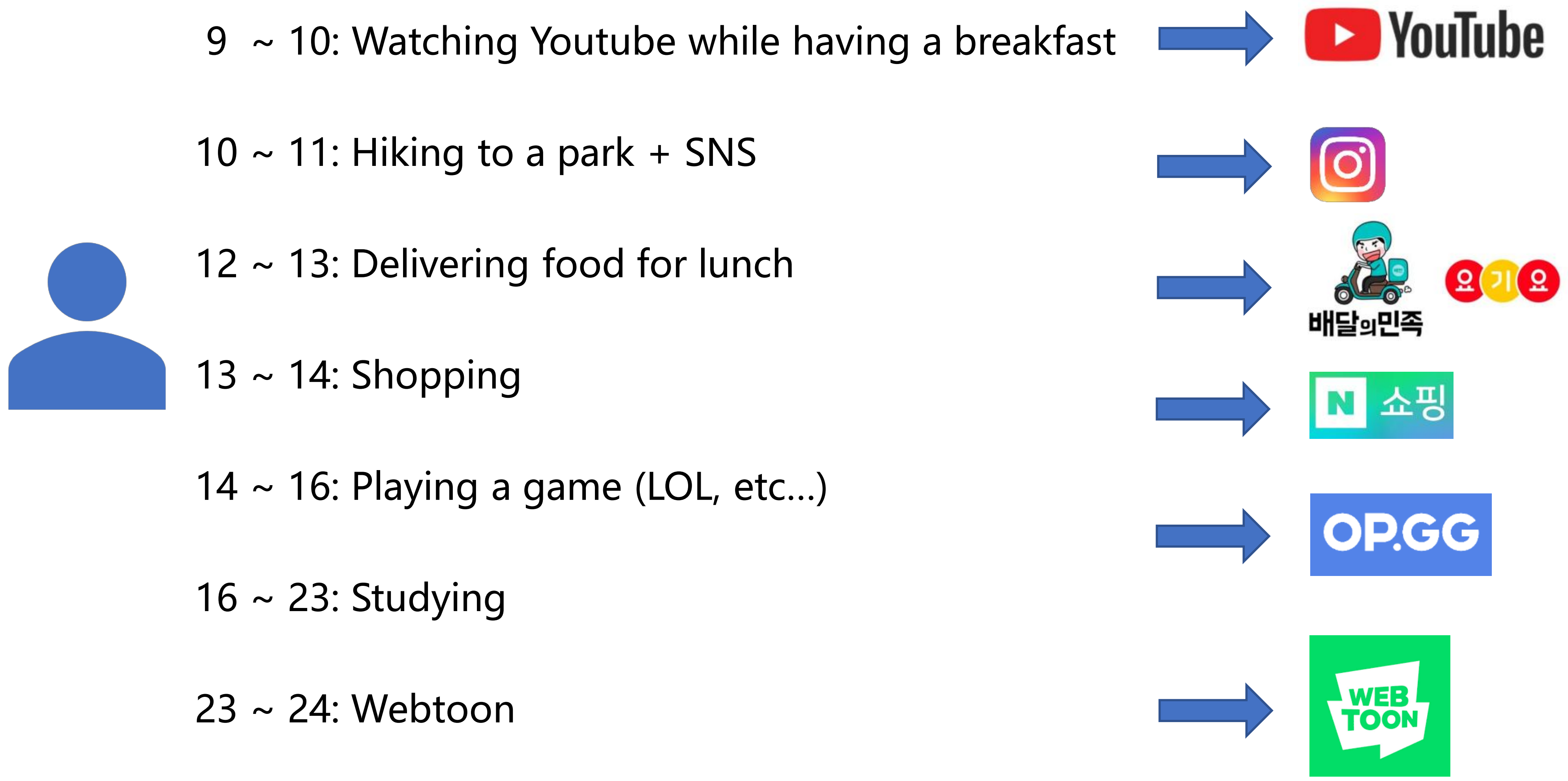
Exam schedule

Date and time

- The exam schedule will be confirmed on one of the following days, based on your votings
  - Midterm exam: 4/20(Thu.) ~ 4/26(Wed.) 19:30 (except weekend)
  - Final exam: 6/15(Thu.) ~ 6/21(Wed.) 19:30 (except weekend)

Overview

❑ The Explosive Growth of Data



9 ~ 10: Watching Youtube while having a breakfast

10 ~ 11: Hiking to a park + SNS

12 ~ 13: Delivering food for lunch

13 ~ 14: Shopping

14 ~ 16: Playing a game (LOL, etc…)

16 ~ 23: Studying

23 ~ 24: Webtoon

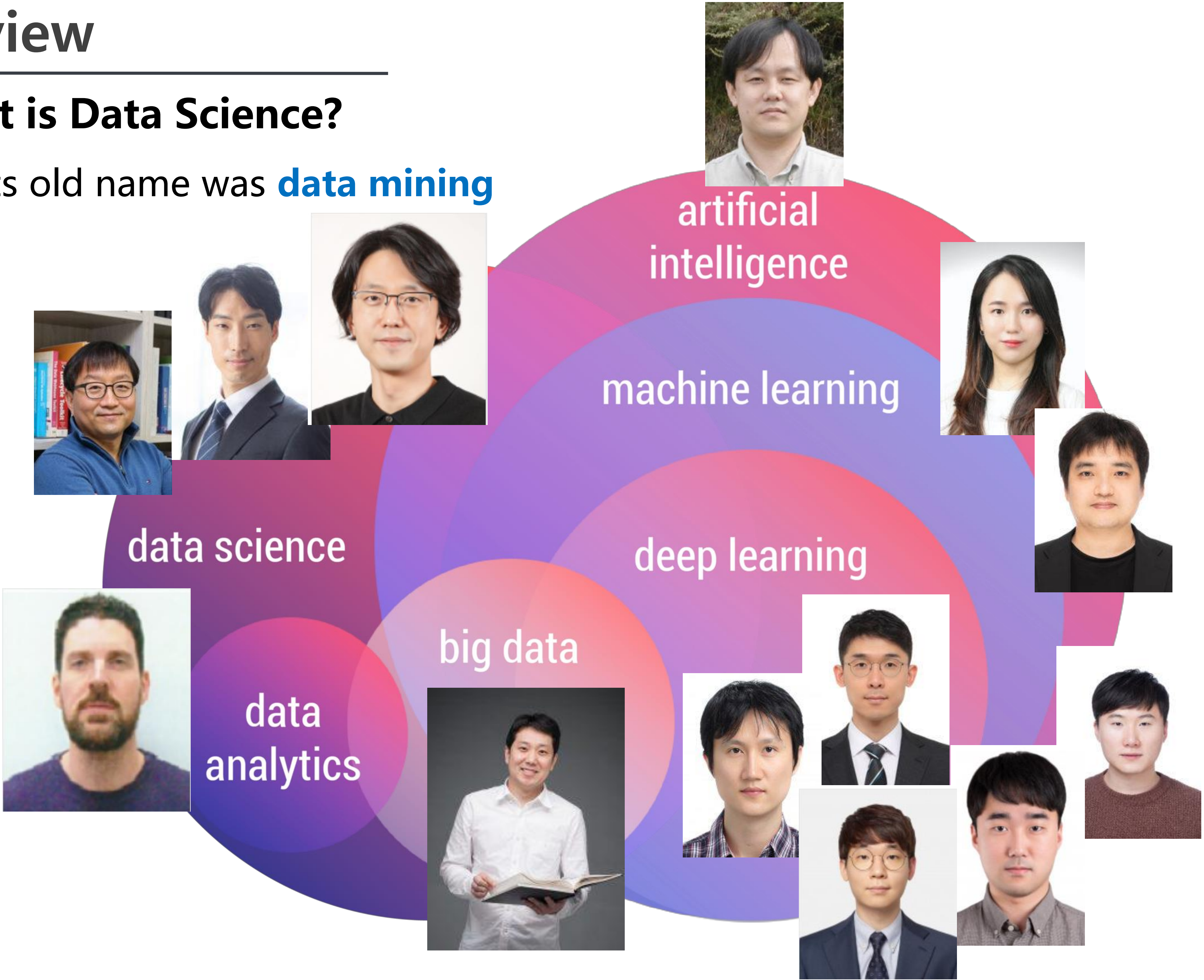
<Our daily life>

Overview

- **The Explosive Growth of Data**
  - Major sources of abundant data
    - **Business**: e-commerce, transactions, stocks, …
    - **Science**: IoT, **bioinformatics**, scientific simulations, …
    - **Society and everyone**: news, digital cameras, YouTube, …
  - Data collection and data availability
    - Automated data collection tools, database systems, computerized society
- **We are drowning in data, but starving for knowledge!**

Overview

- ❑ **What is Data Science?**
  - ❑ Its old name was **data mining**



History of Data Science

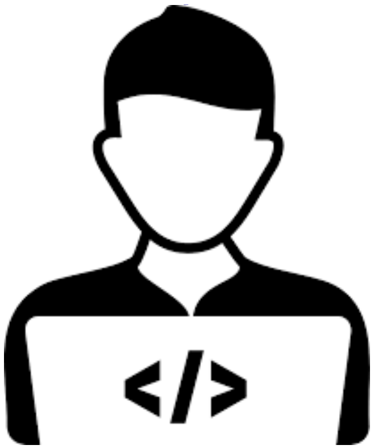
- ❑ **Association rules: Beer and diapers**

History of Data Science

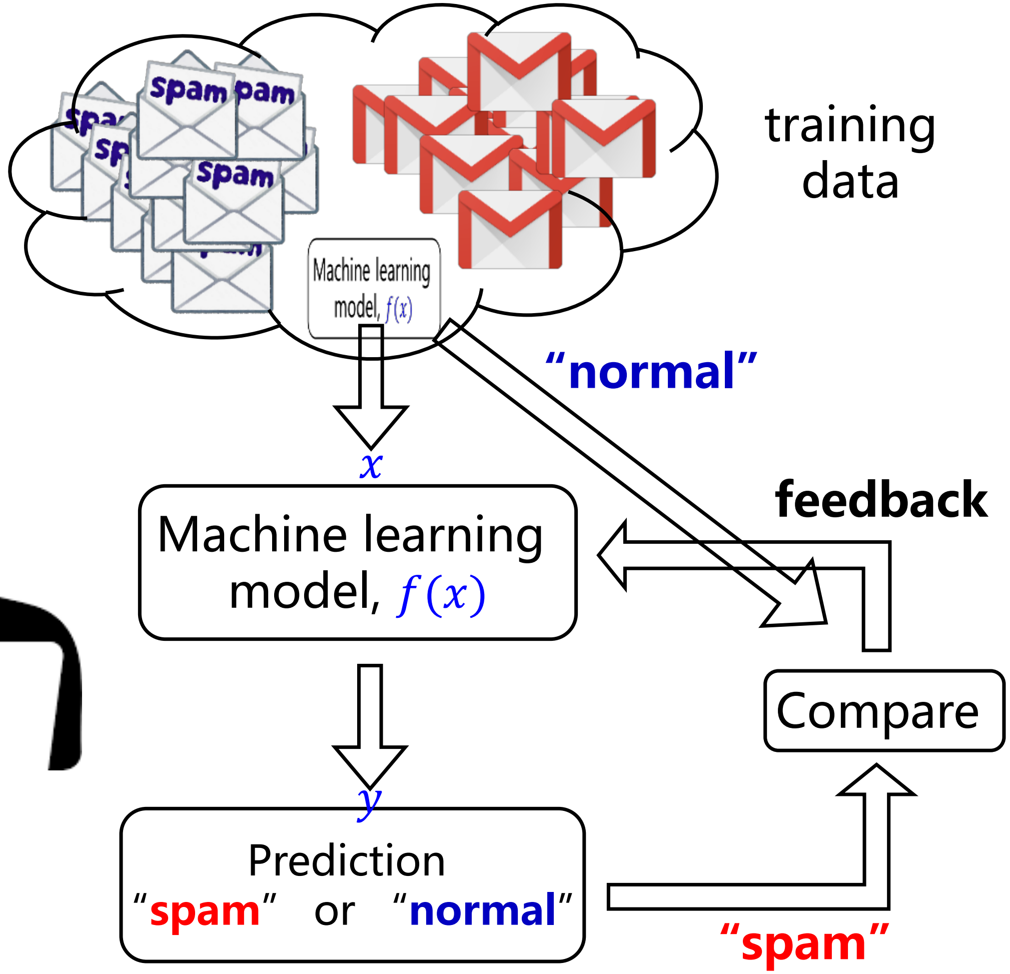
❑ Machine learning: SVM, Random Forest, etc…

- ❑ Spam mail detector example

```
if email contains "bitcoin" and "send" :
   then mark spam;
if email contains "FREE" and "click" :
   then mark spam;
if email contains …
   then …
if email contains …
   then …
if email contains …
   then …
…..
```



Traditional programming



Machine learning

History of Data Science

- Deep learning (deep neural networks): CNN, RNN, LSTM, …

A cat sitting on a suitcase on the floor

A cat is sitting on a tree branch

A dog is running in the grass with a frisbee

A white teddy bear sitting in the grass

Two people walking on the beach with surfboards

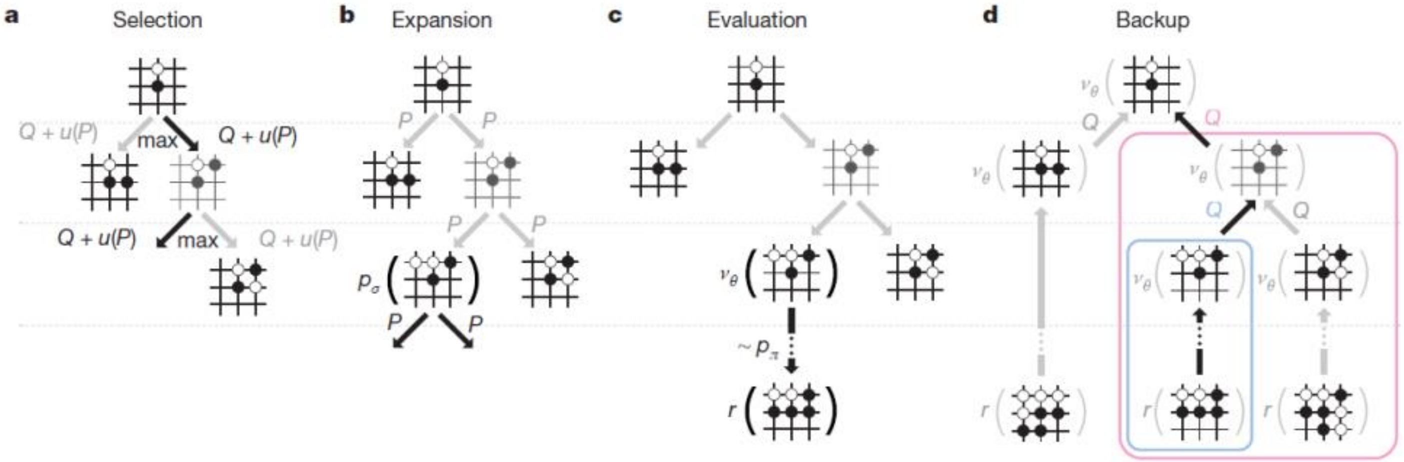
A tennis player in action on the court

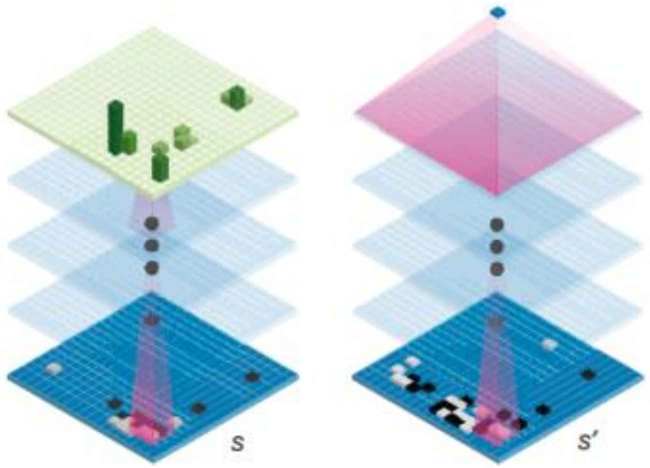
Two giraffes standing in a grassy field

A man riding a dirt bike on a dirt track

History of Data Science

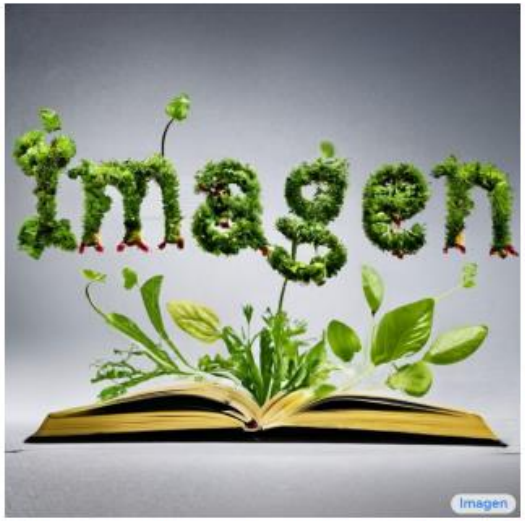
- Deep reinforcement learning: AI gamers beat human experts



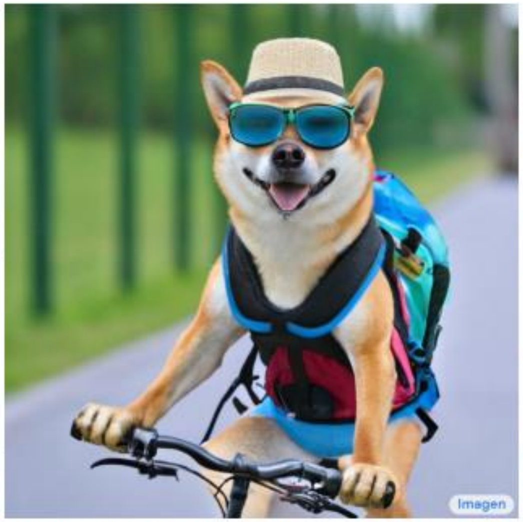


History of Data Science

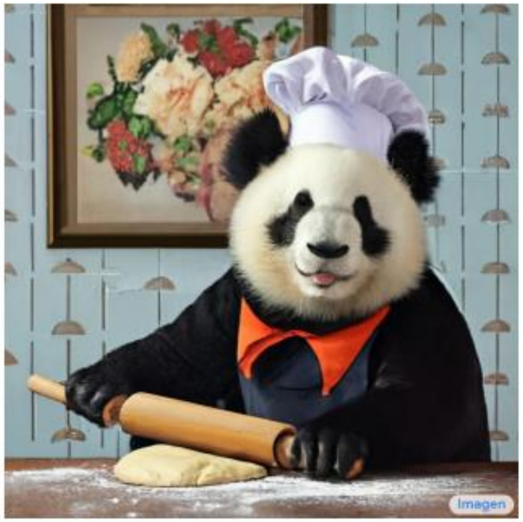
- **Generative AI (Diffusion model, ChatGPT, …): human-level AI**



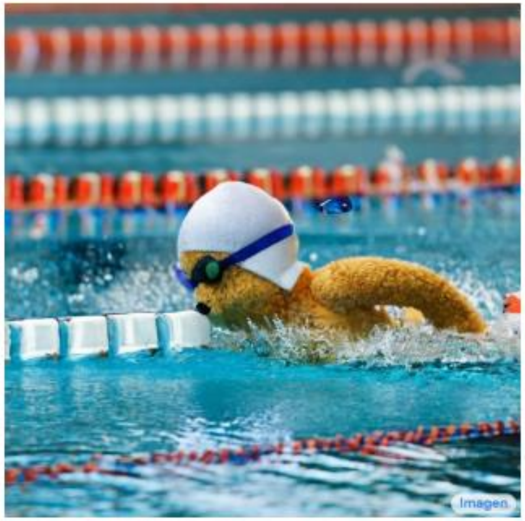
Sprouts in the shape of text 'Imagen' coming out of a fairytale book.



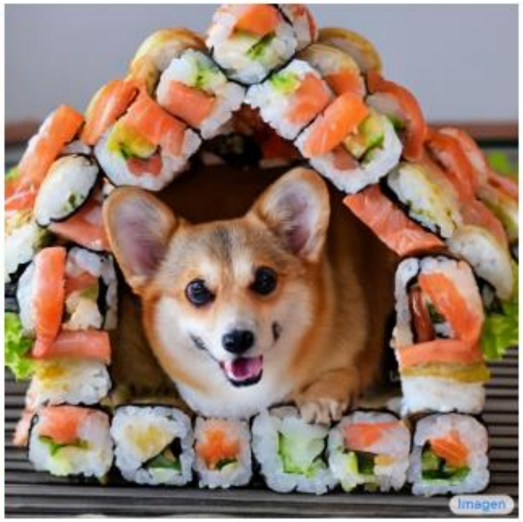
A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.



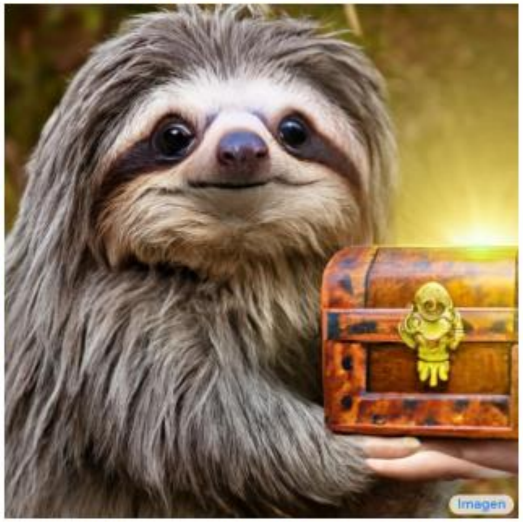
A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.



Teddy bears swimming at the Olympics 400m Butterfly event.



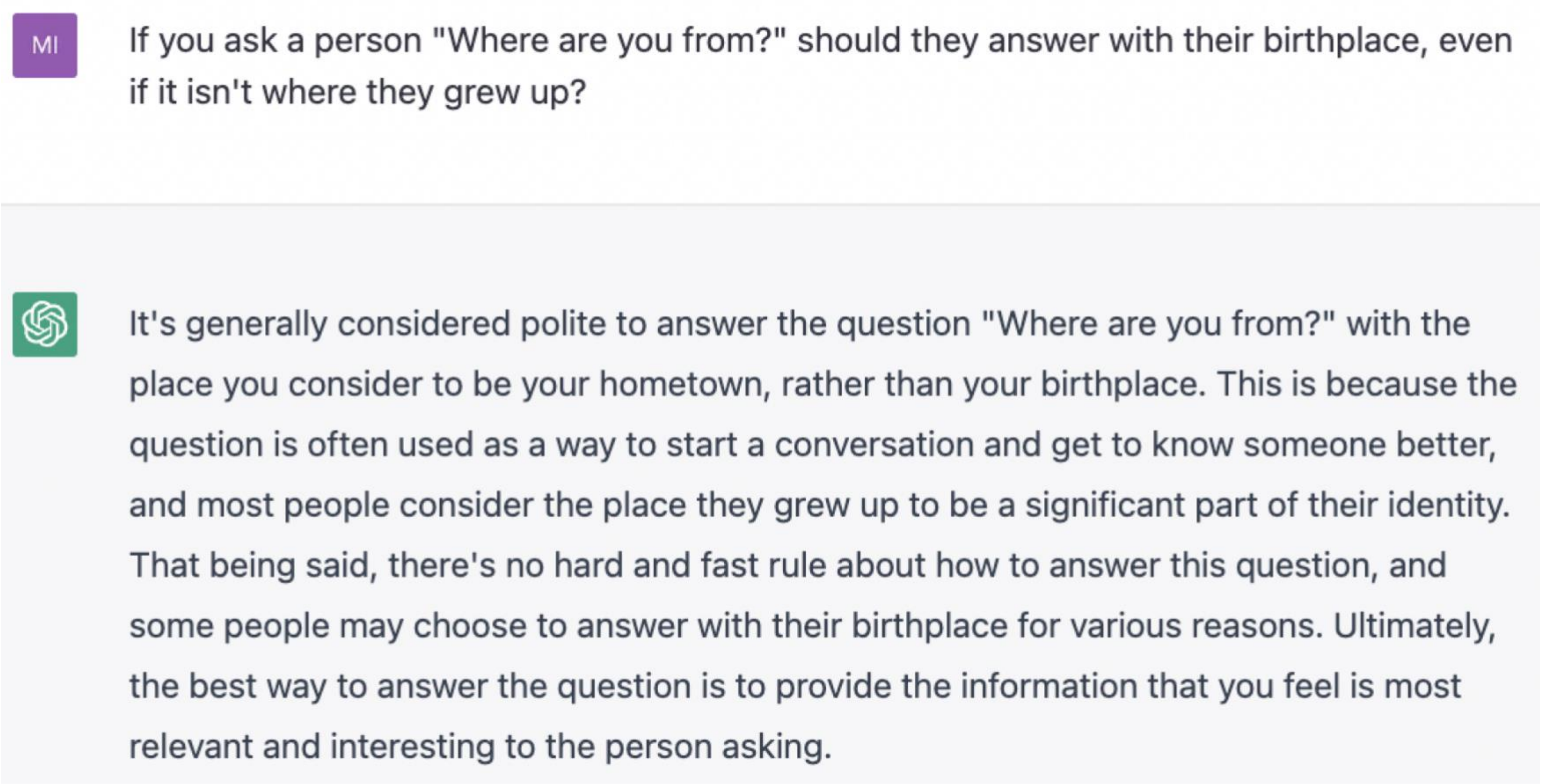
A cute corgi lives in a house made out of sushi.



A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.

History of Data Science

- ❑ **Generative AI (Diffusion model, ChatGPT, …): human-level AI**



MI

If you ask a person "Where are you from?" should they answer with their birthplace, even if it isn't where they grew up?

It's generally considered polite to answer the question "Where are you from?" with the place you consider to be your hometown, rather than your birthplace. This is because the question is often used as a way to start a conversation and get to know someone better, and most people consider the place they grew up to be a significant part of their identity. That being said, there's no hard and fast rule about how to answer this question, and some people may choose to answer with their birthplace for various reasons. Ultimately, the best way to answer the question is to provide the information that you feel is most relevant and interesting to the person asking.

Topics to be Addressed: popular in 1980~2000

- ❑ **Frequent Pattern Mining**
- ❑ **Association Rule Mining**
- ❑ **Getting to Know Your Data**
  - ❑ Data Generalization
  - ❑ Data Preprocessing
  - ❑ Outlier Analysis
- ❑ **Basic Machine Learning Techniques**
  - ❑ Classification (decision tree, etc…)
  - ❑ Clustering (K-means, DBSCAN, etc…)
- ❑ **Other Issues in Data Science**
  - ❑ Graph Analysis
  - ❑ Recommendation

Topics that will NOT be Addressed

- **Very recent machine learning & AI models such as SVM, neural nets, deep learning, etc.…**
- **Complex and high-level problems such as computer vision, natural language processing, etc…**
  - Will be addressed in other lectures

Goals

- ❑ **To learn techniques and applications of data mining in large databases**
  - ❑ To understand the concepts of data mining
  - ❑ To study a variety of data mining techniques
  - ❑ To understand the applications of data mining
  - ❑ To analyze real-world data by using data mining tools
  - ❑ To improve programming skills by developing data mining techniques and applications

Thank You