Data Driven Sampling Methods

Lecture 1: An shadow overview of AI

Hongwei YUAN

University of Macau

Lecturer and TA

- Lecturer
  - Hongwei YUAN
  - Office: E11-2008E;
- TA
  - to be determined later
  - no tutorial classes this Wednesday and next Wednesday.
- UMMoodle
  - Teaching Material, Assignment
  - Notices such as Midterm Time, etc

Course Landscape

- A general but shadow introduction to the data science
- Review of Probability
- Sampling a random variable by computers
- Monte Carlo methods, Maximum likelihood estimator
- Expectation-maximum method
- Regression
- Resampling methods: Cross-validation, Bootstrap
- Bayes Inference, Markov chain Monte Carlo (MCMC), Thompson's Sampling
- Projects

Course Objectives

- To learn basic ideas and motivations for sampling
- To learn some basics sampling methods and statistical estimation methods
- To learn some advanced sampling methods such as resampling, MCMC, etc

Learning Outcomes

Upon completion of this course, you are expected to be clear with basic ideas and motivations for sampling, some basics sampling methods and statistical estimation methods, and understand advanced sampling methods. You will be able to use these sampling methods to analyze real data sets.

Student Assessment

- Assignments: 10 percent
- Projects : 20 percent
- Midterm examination: 20 percent
- Final examination: 50 percent
- No attendance

Artificial Intelligence

- Artificial Intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think, reason, learn, and act like humans.
- It encompasses a broad range of technologies aiming to enable computers to perform tasks that typically require human intelligence, such as understanding natural language, recognizing patterns, solving problems, and making decisions.
- The subject AI is about the methodology of designing machines to accomplish intelligence-based tasks.

Methodologies of AI

- Rule-based
  - Implemented by direct programing
  - Inspired by human heuristics
- Data-based
  - Expert systems: experts or statisticians create rules of predicting or decision making based on data
  - Machine learning
    - ⋆ Direct making prediction or decisions based on the data
    - ⋆ Data Science

What is Data Science

Data Science is an interdisciplinary field that uses scientific methods, algorithms, and systems to extract insights and knowledge from structured and unstructured data. It combines expertise in statistics, computer science, and domain expertise to analyze data and derive actionable conclusions.

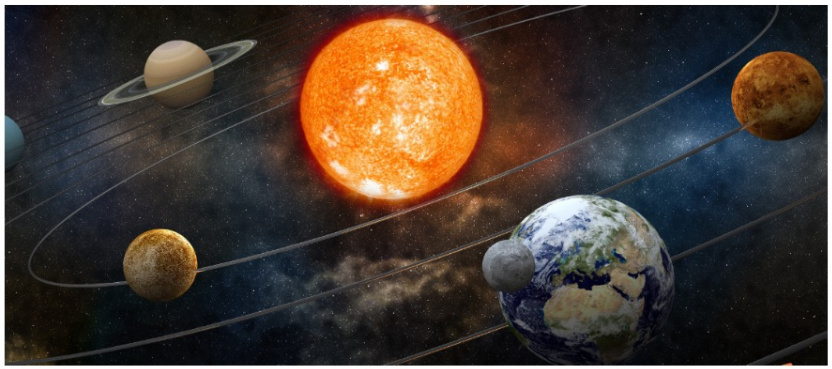
- Data Collection: Gathering data from various sources (e.g., databases, web scraping, APIs).
- Data Cleaning: Removing inconsistencies, duplications, and errors.
- Data Analysis: Using statistical methods to uncover trends and patterns.
- Data Visualization: Presenting findings through visual tools like charts and graphs.

Methodologies of Data Science

- Data Preprocessing: Handling missing values. Normalizing and scaling data for consistency.
- Exploratory Data Analysis (EDA): Summarizing data to identify patterns, trends, or anomalies. Tools include histograms, scatter plots, and box plots.
- Statistical Modeling: Hypothesis testing and regression analysis. Probability distributions and statistical inference.
- Machine Learning Models: Using algorithms to make predictions or classify data.
- Big Data Technologies: Leveraging tools like Hadoop, Spark, and cloud platforms to handle massive datasets.

Data Science

- Physics
  - Goal: discover the underlying principle of the world



  - Solution: build the model of the world from observations

$$F = G\frac{m_1 m_2}{r^2}$$

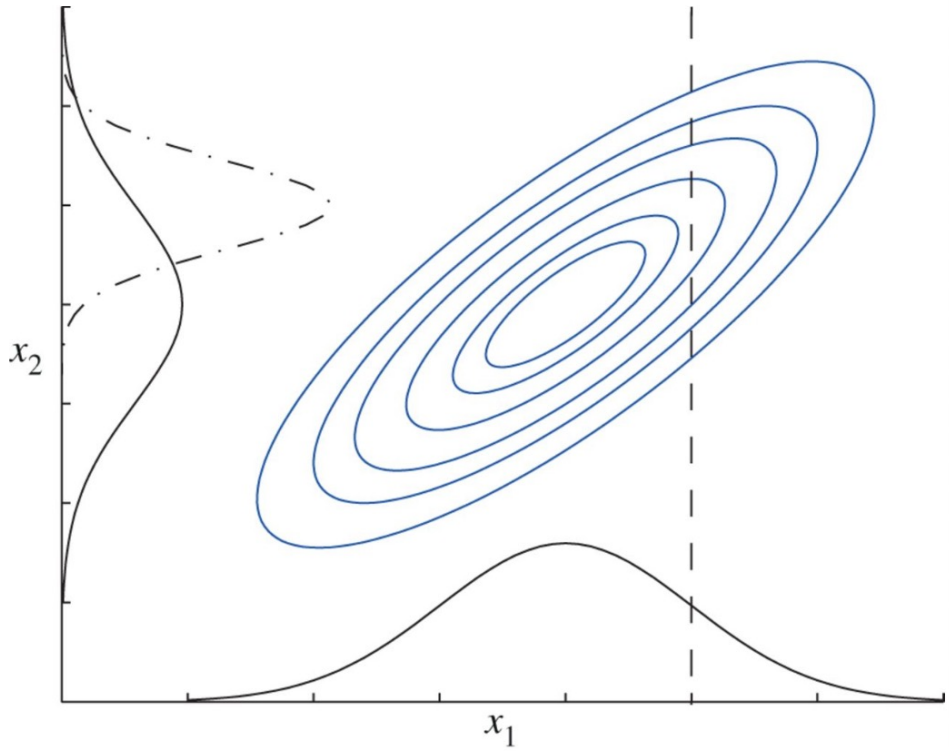
- Data Science
  - Goal: discover the underlying principle of the data

  - Solution: build the model of the data from observations

$$p(x) = \frac{e^{f(x)}}{\sum_{x'} e^{f(x')}}$$

- Mathematically
  - Find joint data distribution $p(x)$
  - Then the conditional distribution $p(x_2|x_1)$
- Gaussian distribution
  - Multivariate

$$p(x) = \frac{e^{-(x-\mu)^\top \Sigma^{-1}(x-\mu)}}{\sqrt{|2\pi\Sigma|}}$$



- Univariate

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

A simple example of User Behavior Modelling

Interest	Gender	Age	BBC Sports	PubMed	Bloomberg Business	Spotify
Finance	Male	29	Yes	No	Yes	No
Sports	Male	21	Yes	No	No	Yes
Medicine	Female	32	No	Yes	No	No
Music	Female	25	No	No	No	Yes
Medicine	Male	40	Yes	Yes	Yes	No

- Joint data distribution

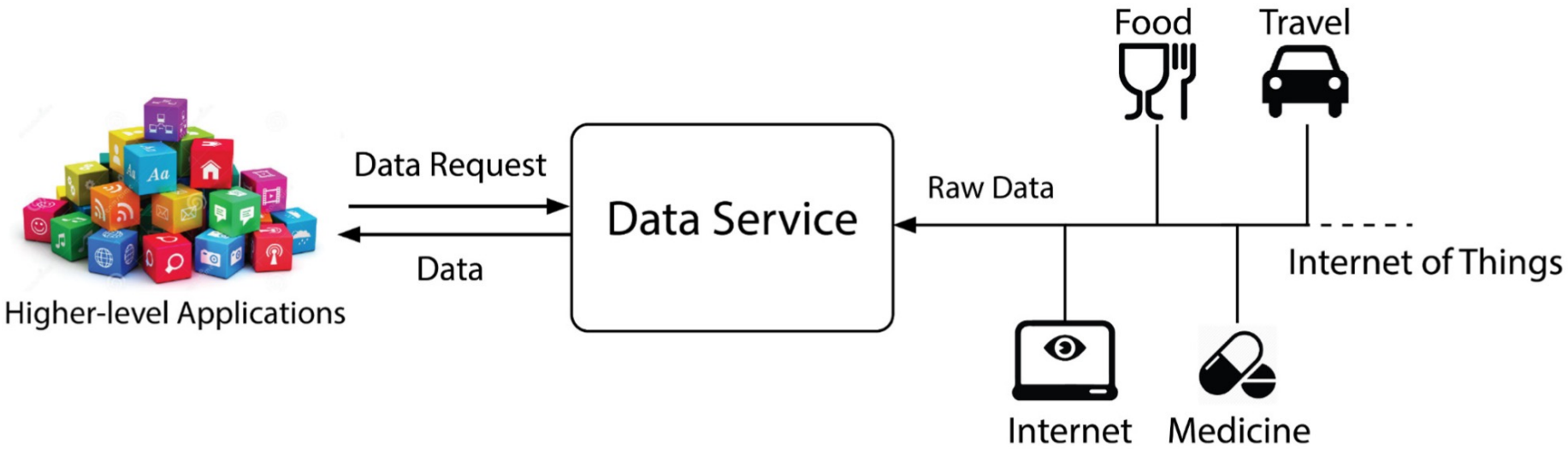
p(Interest=Finance, Gender=Male, Age=29, Browsing=BBC Sports,Bloomberg Business)

- Conditional data distribution

p(Interest=Finance | Browsing=BBC Sports,Bloomberg Business)

p(Gender=Male | Browsing=BBC Sports,Bloomberg Business)

Data Technology



Data itself is not valuable, data service is!

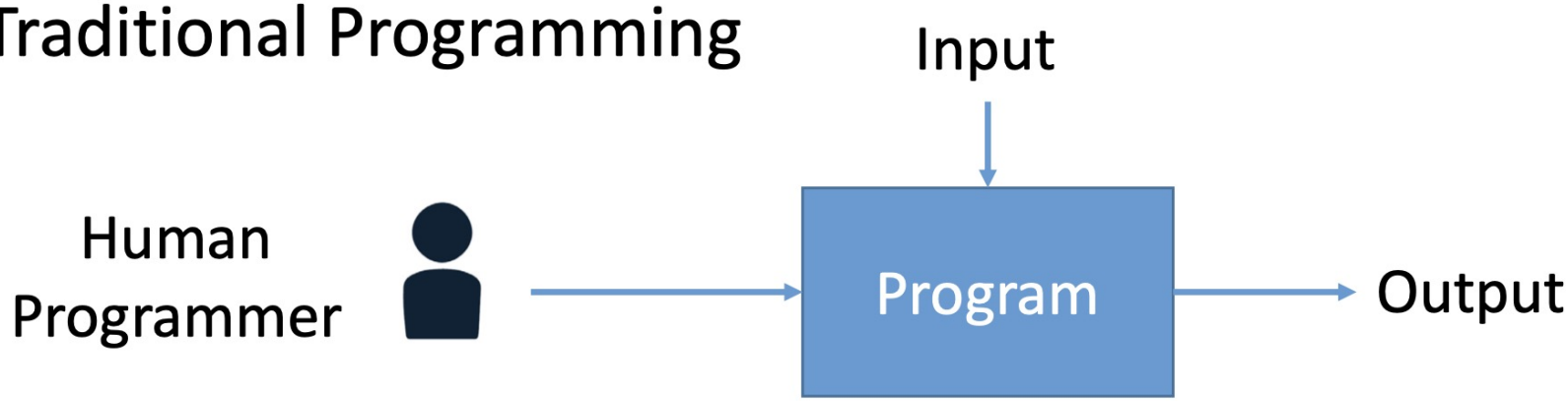
What is Machine Learning (ML)

- Machine learning is the study of algorithms that
  - improve their performance *P*
  - at some task *T*
  - based on experience *E*
  - with non-explicit programming

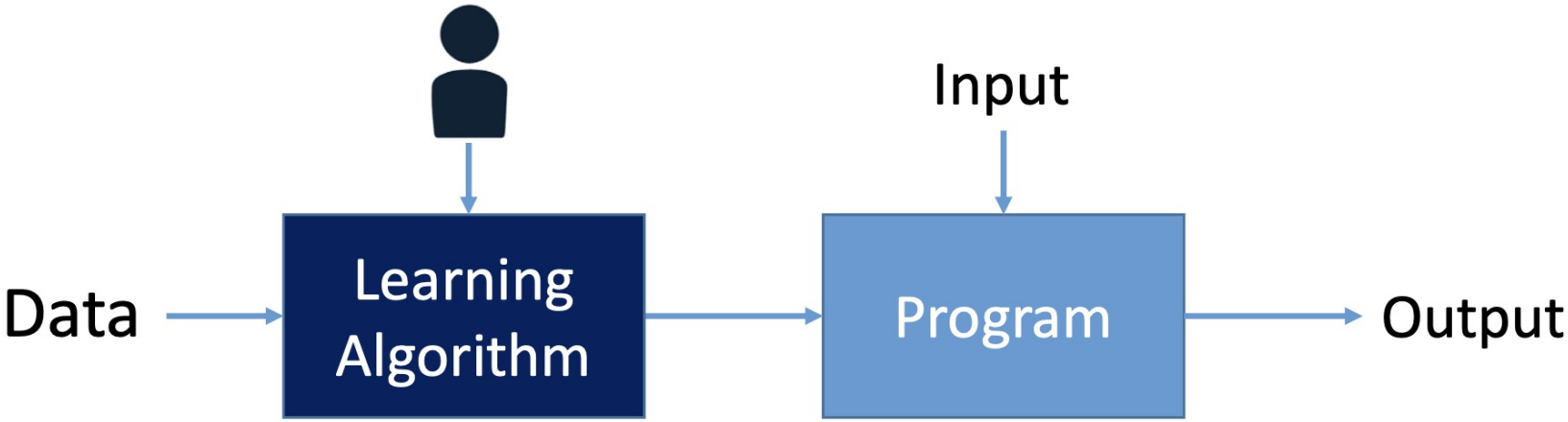
- A well-defined learning task is given by <*P*, *T*, *E*>

Programming v.s. Machine Learning

- Traditional Programming



- Machine Learning



The Advantages of Machine Learning

ML is used when

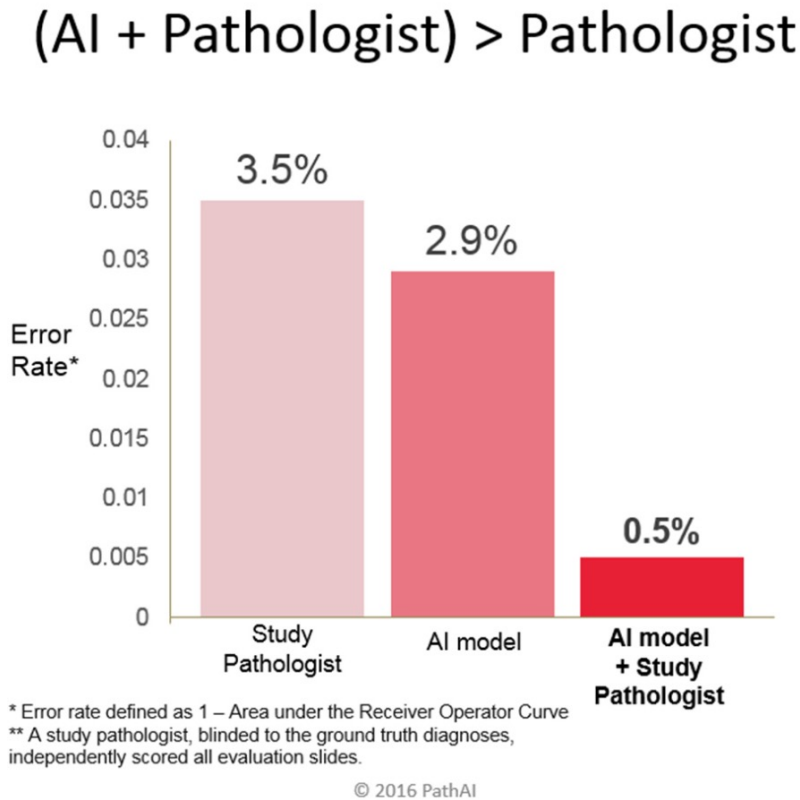
- Models are based on a huge amount of data
  - Examples: Google web search, Facebook news feed
- Output must be customized
  - Examples: News / item / ads recommendation
- Humans cannot explain the expertise
  - Examples: Speech / face recognition, game of Go
- Human expertise does not exist
  - Examples: Navigating on Mars

Two Types of Machine Learning

- Prediction
  - Predict the desired output given the data (supervised learning)
  - Generate data instances (unsupervised learning)
- Decision Making
  - Take actions in a dynamic environment (reinforcement learning)
    - to transit to new states
    - to receive immediate reward
    - to maximize the accumulative reward over time

Application Case of ML: Medical Image Analysis

- Breast Cancer Diagnoses

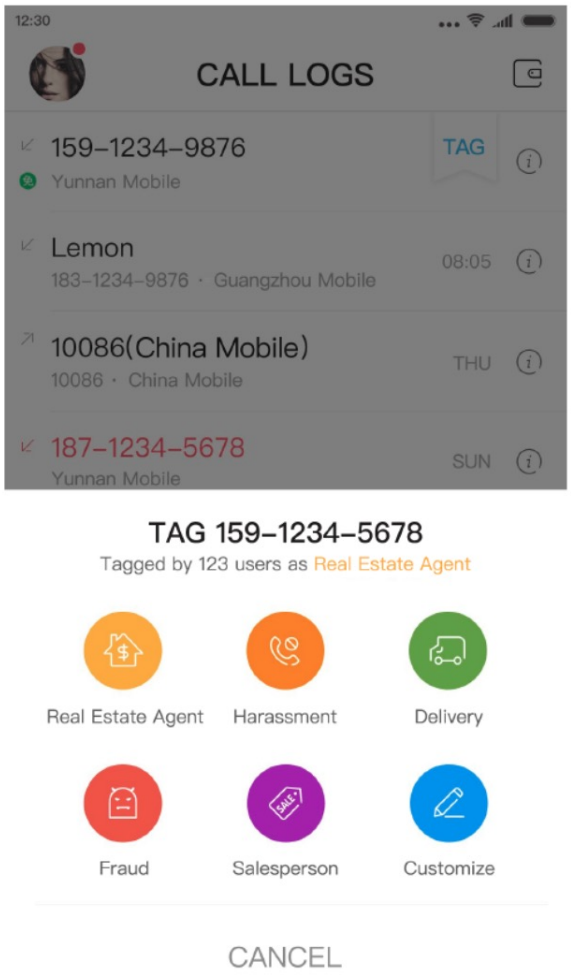


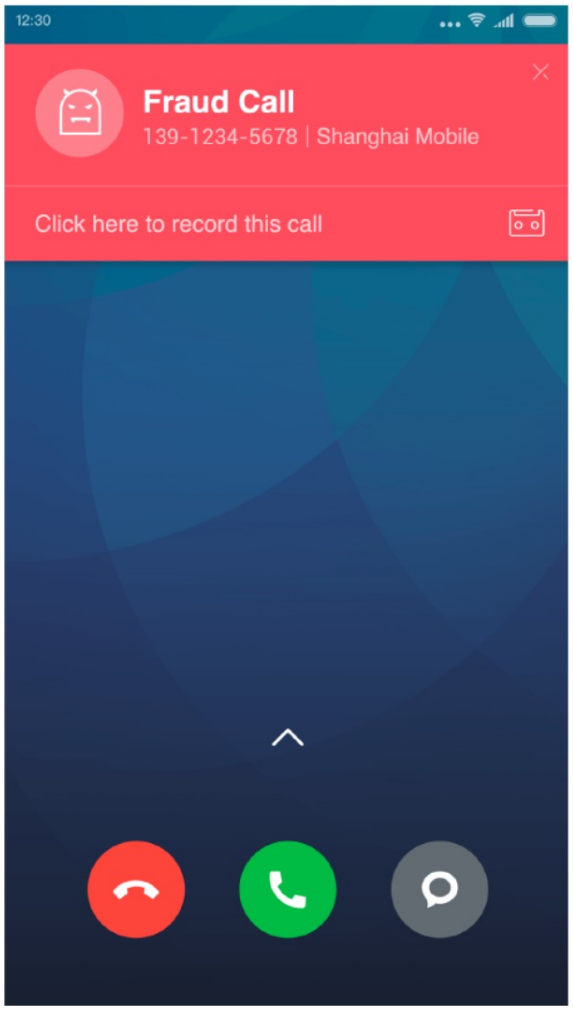
Application Case of ML: Social Networks

- Friends/Tweets/Job Candidates suggestion

Application Case of ML: Anomaly Detection

- Detect malicious calls





Application Case of ML: Robotics Control

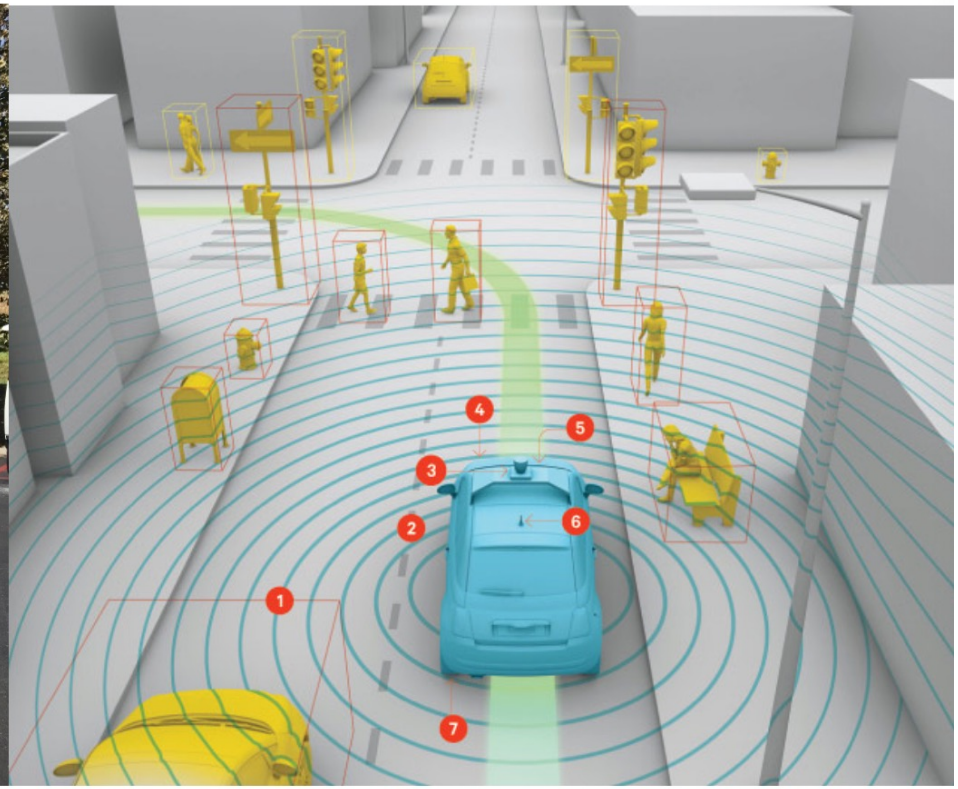
- Stanford Autonomous Helicopter
  - http://heli.stanford.edu/

Application Case of ML: Robotics Control

- Ping pong robot
  - https://www.youtube.com/watch?v=tIIJME8-au8

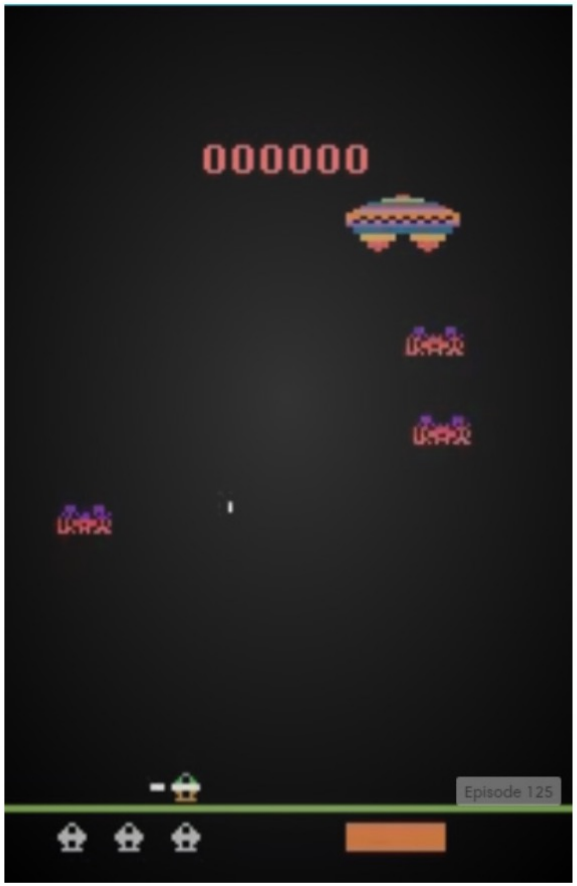
Application Case of ML: Self-Driving Cars

- Google Self-Driving Cars
  - https://www.google.com/selfdrivingcar/



Application Case of ML: Game Playing

- Take actions given screen pixels
  - https://gym.openai.com/envs#atari



Application Case of ML: AlphaGo

IBM Deep Blue (1996)

- 4-2 Garry Kasparov on Chess
- A large number of crafted rules
- Huge space search

Google AlphaGo (2016)

- 4-1 Lee Sedol on Go
- Deep machine learning on big data

Application Case of ML: Text Generation

- Making decision of selecting the next word/char
- Chinese poem example. Can you distinguish?

Human	Machine
南陌春风早，东邻去日斜。	山夜有雪寒，桂里逢客时。
紫陌追随日，青门相见时。	此时人且饮，酒愁一节梦。
胡风不开花，四气多作雪。	四面客归路，桂花开青竹。