

AI5213 / EC4213 / AI4021 / CT5303 / ET5402 / FE5402

Machine Learning and Deep Learning



Introduction

Mon/Wed 13:00-14:30

Sundong Kim (sundong@gist.ac.kr)

Syllabus

- Date/Time & Location
 - Class: College Bld. C 104
 - Mon/Wed 13:00-14:30

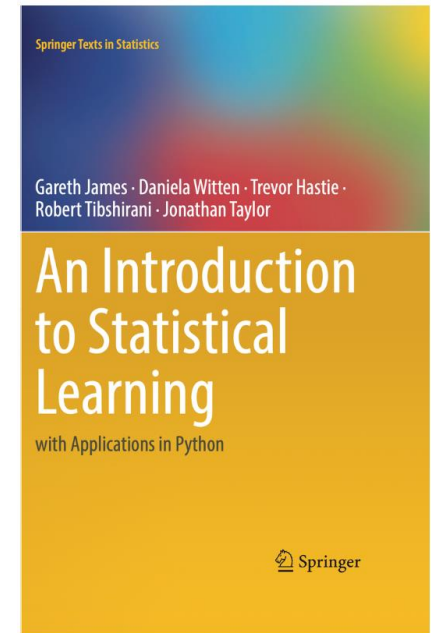
- Class website
 - <https://sundong.kim/courses/mldl23f>

- Online forum
 - Ed discussion
 - Gradescope for assignments

→ More information on class website.

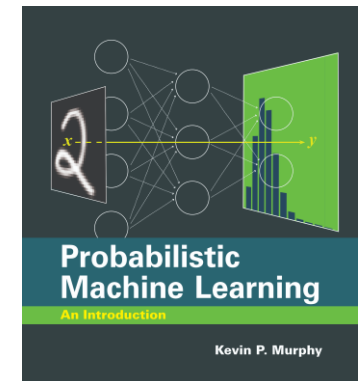
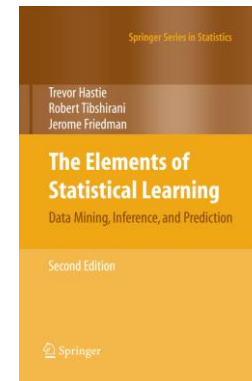
Textbook

- **An Introduction to Statistical Learning with Application in Python (2023)**
(PDF available at: <https://www.statlearning.com/>)



- **References**

- The Elements of Statistical Learning, (Hastie et al, 2017)
<https://hastie.su.domains/ElemStatLearn/>
- Probabilistic Machine Learning – An Introduction (Murphy, 2023)
<https://probml.github.io/pml-book/book1.html>



Good News

- Machine Learning and Deep Learning (MLDL) course will also be held every semester!
- I encourage students who are not yet prepared to attend the course next semester.

We assume that you know...

- Good understanding about probability theory
- Calculus and some Linear Algebra
- English writing
- Handy in python



So, what is machine learning?

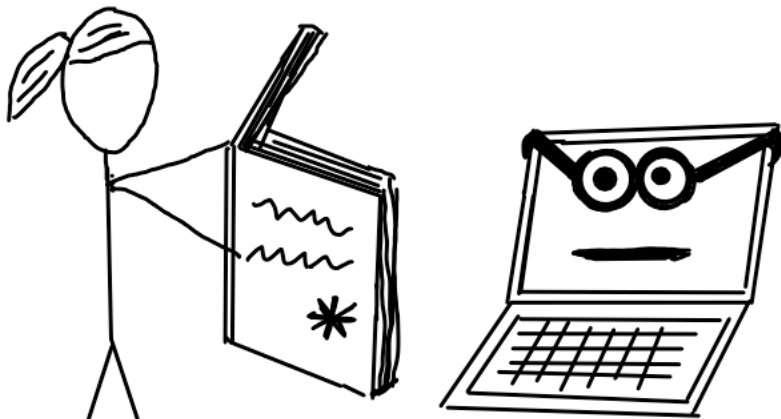


Machine Learning

- Automating automation
- Getting computers to program themselves
- Writing software is the bottleneck
- Let the data do the work instead!

Traditional Programming vs. Machine Learning

Without Machine Learning



* VERY SPECIFIC INSTRUCTIONS

With Machine Learning

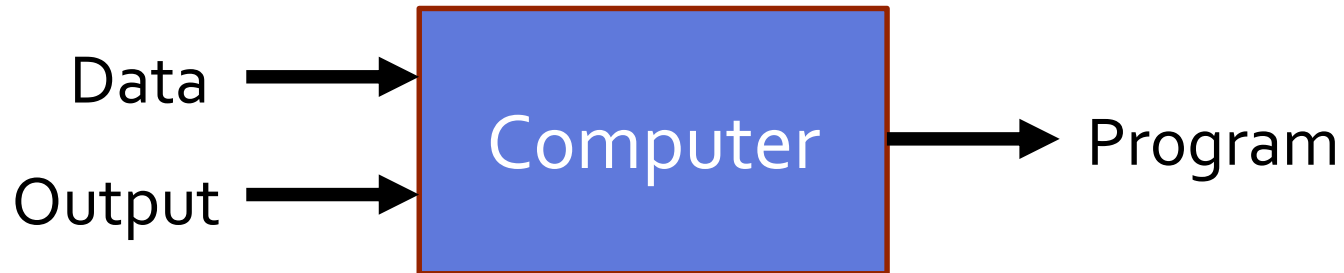


Traditional Programming vs. Machine Learning

Traditional Programming



Machine Learning

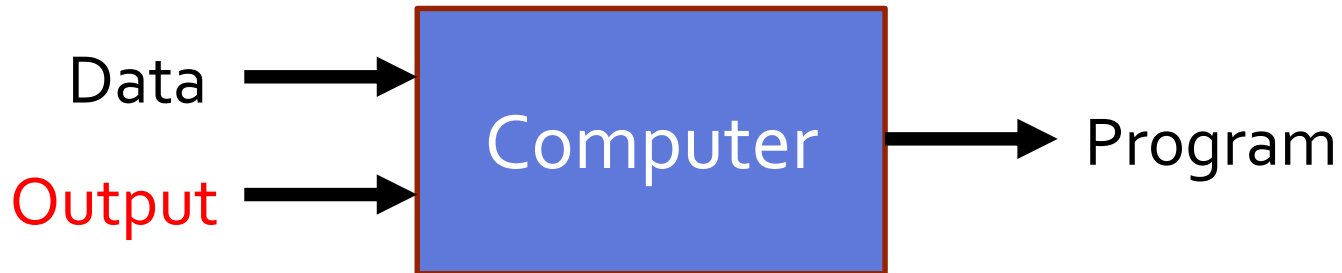


Traditional Programming vs. Machine Learning

Traditional Programming



Machine Learning

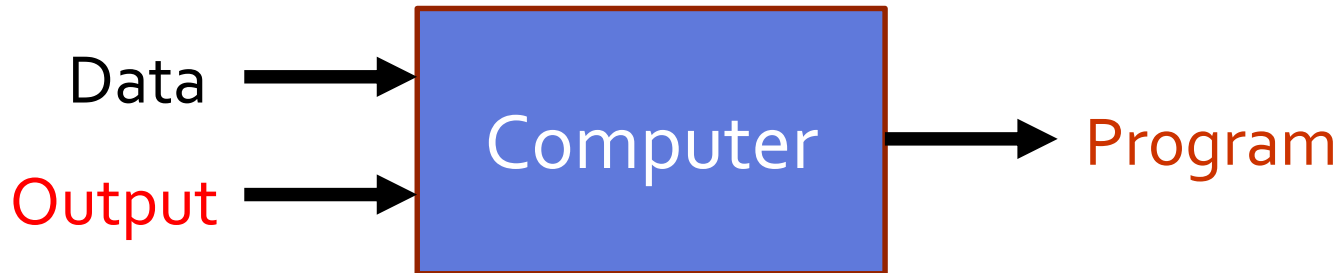


Traditional Programming vs. Machine Learning

Traditional Programming



Machine Learning



OK, but more concrete?

Machine



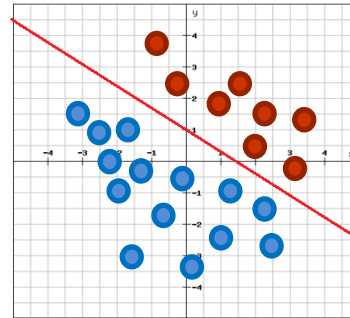
Function

Learning



Computing coefficients

$$ax+by+c = 0$$



How to compute the coefficient? = Machine Learning

The goal of machine learning

- From data to predict the output for never seen input

Generalization

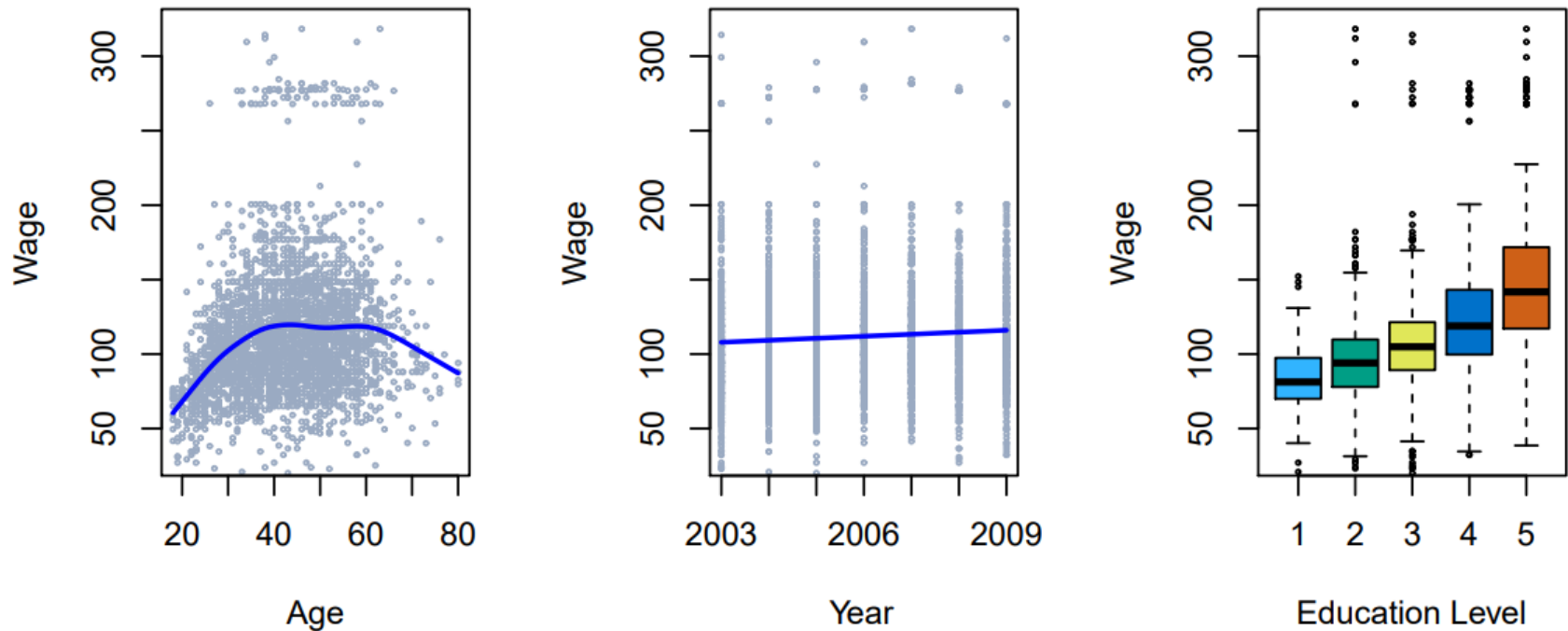
Three axes of machine learning

- Data
- Tasks What knowledge we seek from data?
- Models (Algorithms)

Data

- Fully supervised
- Partially supervised
 - Some variables missing sometimes
 - Using a combination of labeled and unlabeled data
- Actively supervise/collect/sense data
 - Having the learning system decide which examples to ask an oracle to label

Datasets in this book - Wage Data

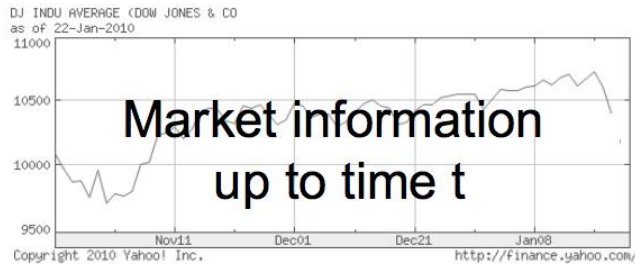


Income survey data for males from the central Atlantic region of the USA in 2009.

Tasks

- Prediction Problems
 - Estimate **output** given **input**

Feature Space \mathcal{X}



Label Space \mathcal{Y}



“Sports”
“News”
“Science”
...



Share Price
“\$ 24.50”

Task: Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$.

Models that we will cover during classes

- Linear Regression
- Classification
- Resampling Methods
- Model Selection & Regularization
- Moving beyond Linearity
- Tree-Based Methods
- Support Vector Machines
- Survival Analysis and Censored Data
- Unsupervised Learning
- Multiple Testing
- Deep Learning

Machine Learning Problem Types

- Based on output:
 - Regression, Classification, Clustering, Embedding, ...
- Based on data:
 - Supervised, Unsupervised, Semi-supervised, Reinforcement Learning, ...
- Based on models:
 - Discriminative, Generative,

Machine learning problems

- Classification
 - Data to discrete class label
 - Predicting a class label
- Regression
 - Predicting a numeric value
- Similarity
 - Finding similar/dissimilar data
- Clustering
 - Discovering structure in data
- Embedding
 - Data to a vector
- Reinforcement Learning
 - Training by feedback

Machine learning problems

- Classification
 - Data to discrete class label
 - Predicting a class label
- Regression
 - Predicting a numeric value
- Similarity
 - Finding similar/dissimilar data
- Clustering
 - Discovering structure in data
- Embedding
 - Data to a vector
- Reinforcement Learning
 - Training by feedback



Usually, supervised



Usually, unsupervised

Literally, RL

Categorizing machine learning methods

- Supervised learning
 - Train a model with data with label
- Unsupervised learning
 - Train a model with data without label
- Semi-supervised learning
 - Train a model with data with label for some
- Active learning
 - Train a model with selected data with label
- Reinforcement learning
 - Train a model with indirect label

Supervised Learning Problem

Starting point:

- Outcome measurement Y (also called dependent variable, response, target).
- Vector of p predictor measurements X (also called inputs, regressors, covariates, features, independent variables).
- In the *regression problem*, Y is quantitative (e.g price, blood pressure).
- In the *classification problem*, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).
- We have training data $(x_1, y_1), \dots, (x_N, y_N)$. These are observations (examples, instances) of these measurements.

Objectives

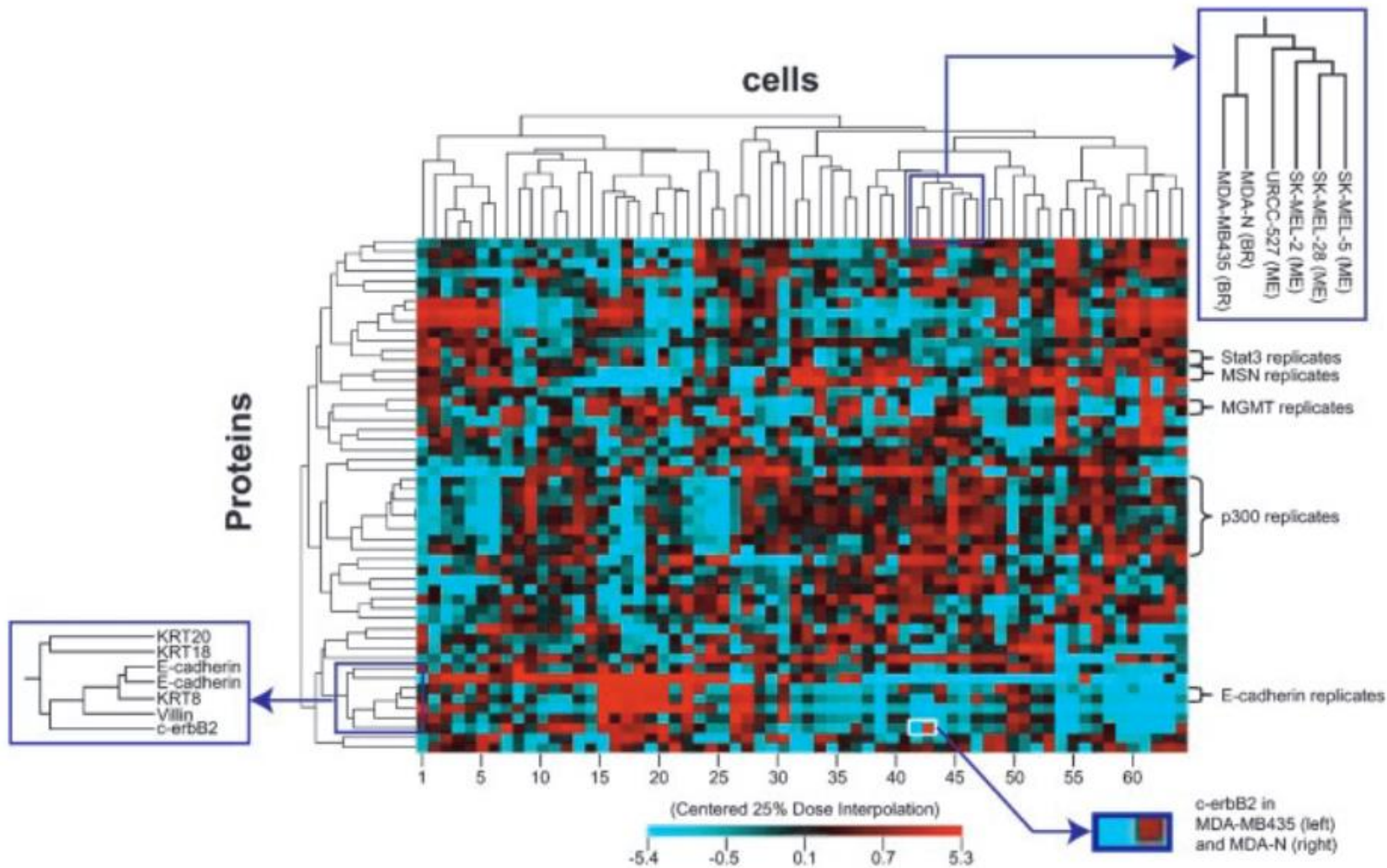
On the basis of the training data we would like to:

- Accurately predict unseen test cases.
- Understand which inputs affect the outcome, and how.
- Assess the quality of our predictions and inferences.

Unsupervised Learning

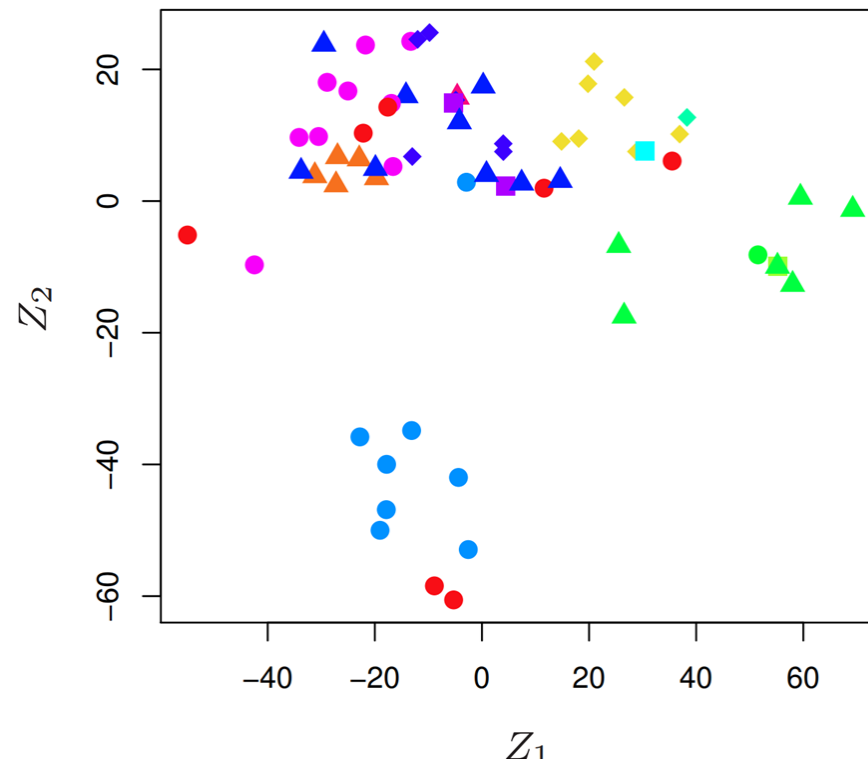
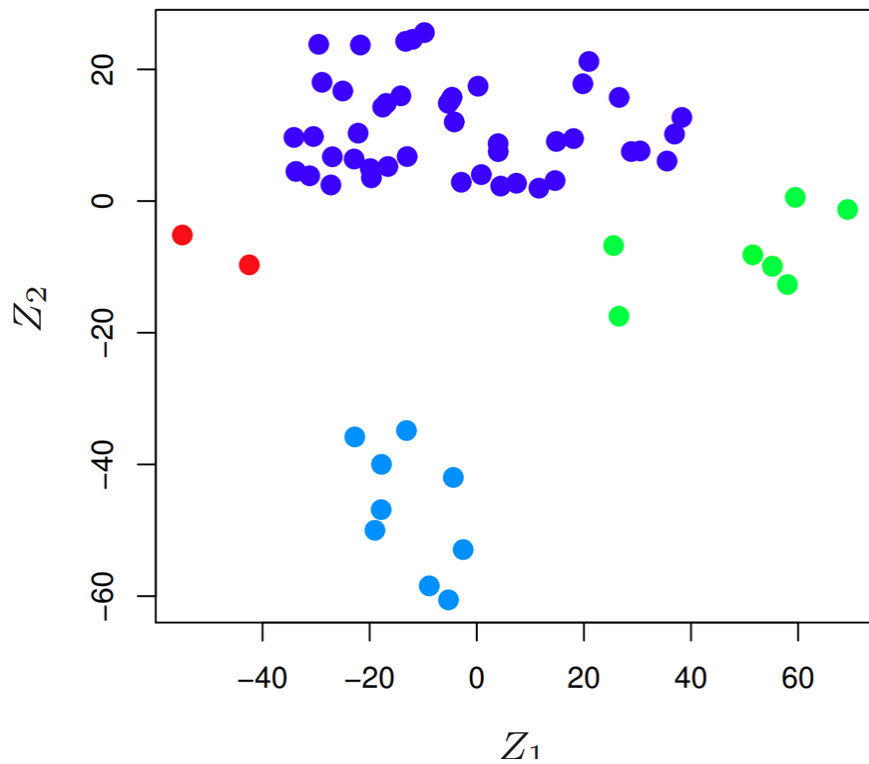
- No outcome variable, just a set of predictors (features) measured on a set of samples.
- objective is more fuzzy — find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- difficult to know how well you are doing.
- different from supervised learning, but can be useful as a pre-processing step for supervised learning.

NCI60 gene expression dataset – Hierarchical Clustering



<https://www.pnas.org/doi/epdf/10.1073/pnas.2331323100>

NCI60 gene expression dataset – in 2D space



Philosophy

- It is important to understand the ideas behind the various techniques, in order to know how and when to use them.
- One has to understand the simpler methods first, in order to grasp the more sophisticated ones.
- It is important to accurately assess the performance of a method, to know how well or how badly it is working [simpler methods often perform as well as fancier ones!]
- This is an exciting research area, having important applications in science, industry and finance.
- Statistical learning is a fundamental ingredient in the training of a modern *data scientist*.



Applications

Recommender system



Real Python

Spam Detection

- data from 4601 emails sent to an individual (named George, at HP labs, before 2000). Each is labeled as *spam* or *email*.
- goal: build a customized spam filter.
- input features: relative frequencies of 57 of the most commonly occurring words and punctuation marks in these email messages.

	george	you	hp	free	!	edu	remove
spam	0.00	2.26	0.02	0.52	0.51	0.01	0.28
email	1.27	1.27	0.90	0.07	0.11	0.29	0.01

*Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between **spam** and **email**.*

Virtual Assistant



Natural Language Processing

The screenshot shows the ChatGPT web interface in a browser. The address bar displays the URL: `chat.openai.com/chat/fe88e6cc-2c9d-431f-aac6-198f82b5af3d`. The browser's taskbar includes icons for Gmail, Google Calendar, My Drive, Toggl Track, Bookmarks, Rick Davies (@Man...), LinkedIn, ParEvo - Aptivate, and Other bookmarks.

Chat History Sidebar (Left):

- + New chat
- ChatGPT Language Mo [edit] [delete]
- International Tensions Over SA
- Most Significant Change (MSC)
- Rick Davies Unknown Consulta
- NAM coalition vs Western nati
- Clear conversations
- Dark mode
- OpenAI Discord
- Updates & FAQ
- Log out

Main Chat Window:

Prompt: Describe ChatGPT in 200 words, for a general, non-specialist reader

Response: ChatGPT is a language model developed by OpenAI. It is based on a deep learning technique called transformer, which allows it to generate human-like text. It has been trained on a large dataset of text from the internet, and can be fine-tuned for specific tasks such as language translation, text summarization, and text generation. ChatGPT is capable of understanding and generating text in multiple languages, and can respond to a wide range of questions, prompts, and commands. It can also generate creative content such as stories and poetry. The model can be used for various applications like virtual assistants, chatbots, automated writing and more. However, it is important to note that while ChatGPT is highly advanced, it is not a human and may not always understand context or nuances in language and may produce errors or biases present in the training data.

Buttons: Regenerate response

Footer: [ChatGPT Jan 9 Version](#). Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

Self-driving car – Computer Vision

