Introduction to Statistical Learning

Tianhang Zheng

https://tianzheng4.github.io

Course Website

https://tianzheng4.github.io/umkc-teaching/2023-fall-teaching-1/

What are on the course website:

- Lecture slides
- Lab material
- Contact Information

If you are interested in my research, feel free to contact me.

Assignments and Requirements

- Course + Lab: Theory and Implementation 30%
- Course Project: Maximum four people in one group 20%
- Mid-term and Final Exams 20% + 30%

Statistical Learning v.s. Machine Learning

Definition (Mitchell, 1998)

A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.

Machine learning and deep learning emphasize large scale applications and prediction accuracy.

Statistical learning emphasizes models and their interpretability, and precision and uncertainty.

Spam Detection

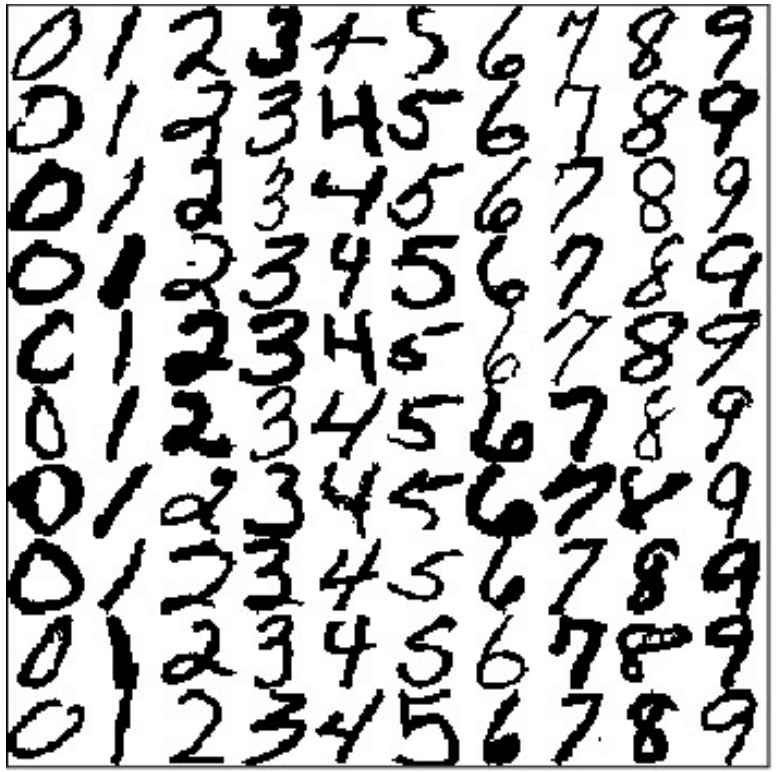
- Given features of emails, build a spam filter

	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.

Digit Recognition

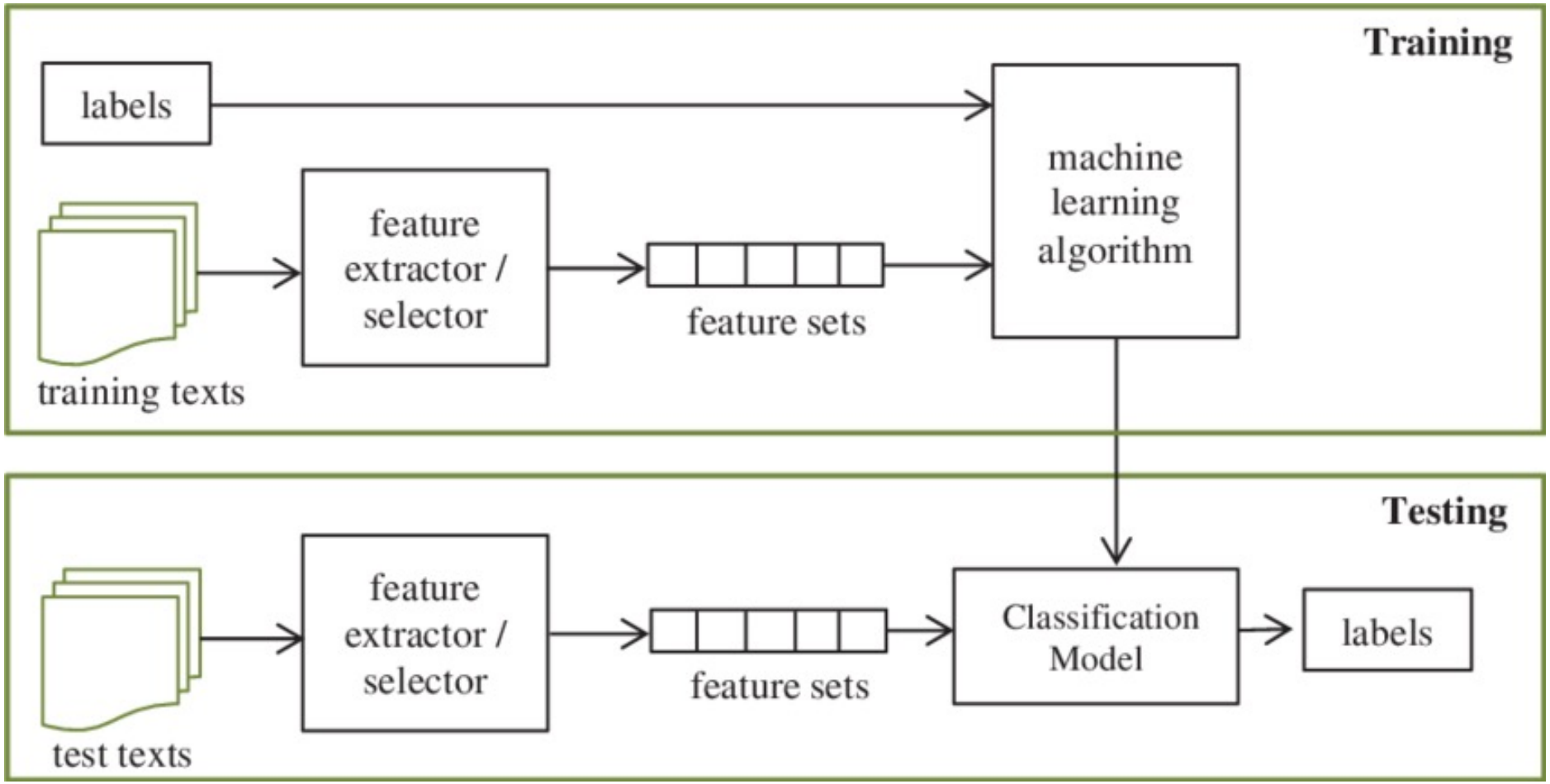
- Given the images of handwritten digits, recognize the digits



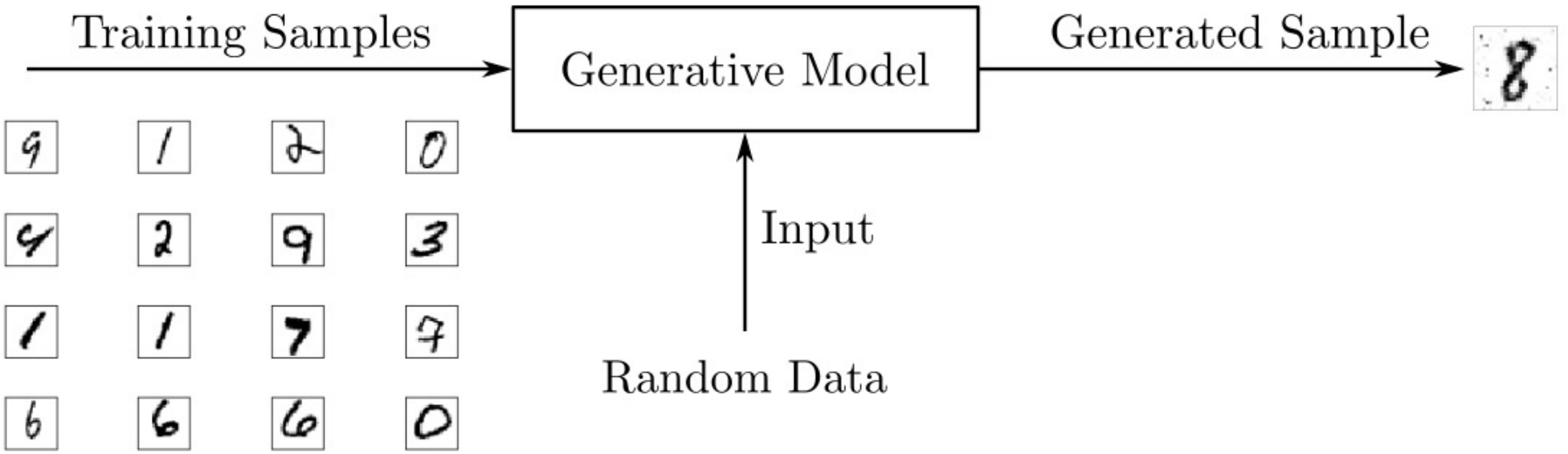
MNIST dataset

https://www.kaggle.com/datasets/hojjatk/mnist-dataset

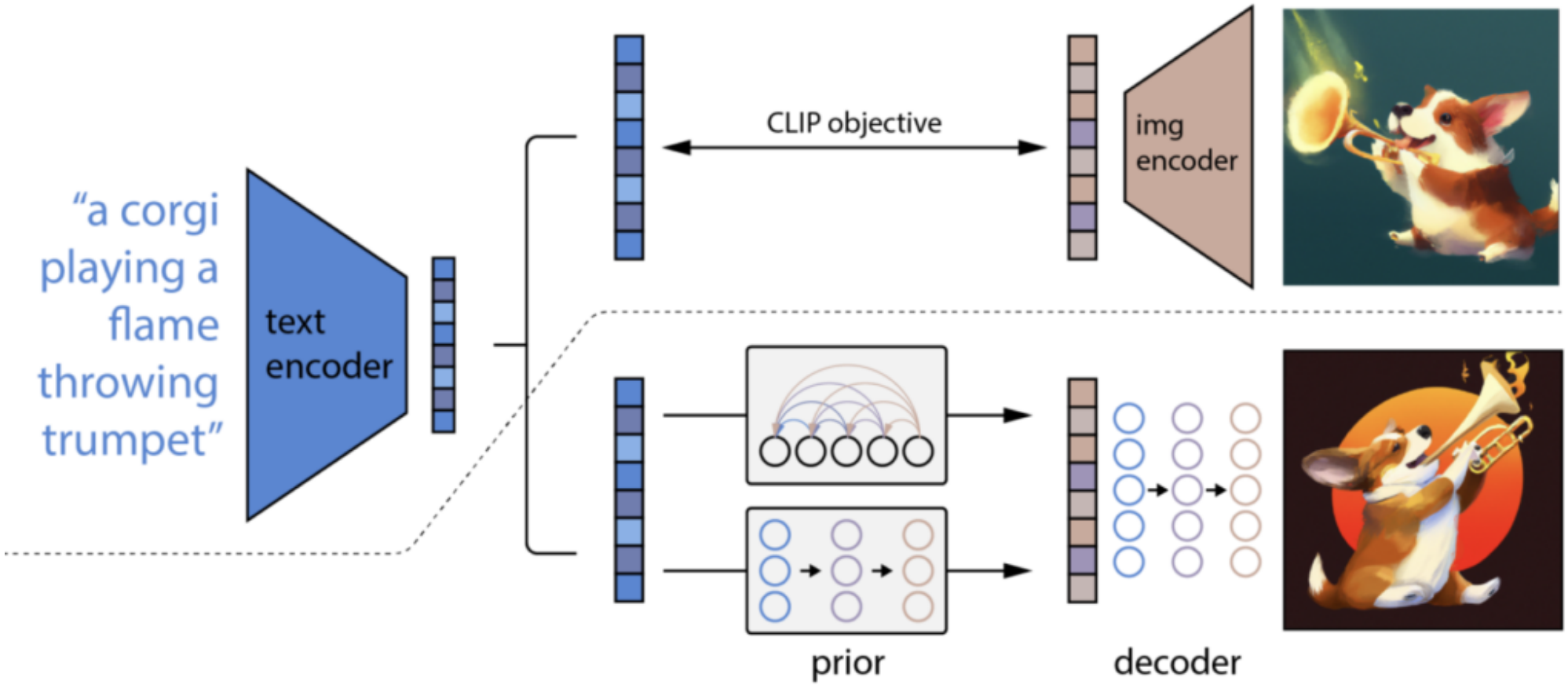
Text Classification



Generating New Samples (Generative Models)



Text-to-Image Generation



Applications of Statistical Learning

- bioinformatics
- detecting network intrusion
- neuroscience
- medical diagnosis
- stock market analysis
- social network analysis
- traffic and infrastructure planning

…

Textbooks

- An Introduction to Statistical Learning
- Machine Learning with Python Tutorial (Optional)

https://tianzheng4.github.io/umkc-teaching/2023-fall-teaching-1/

Supervised Learning

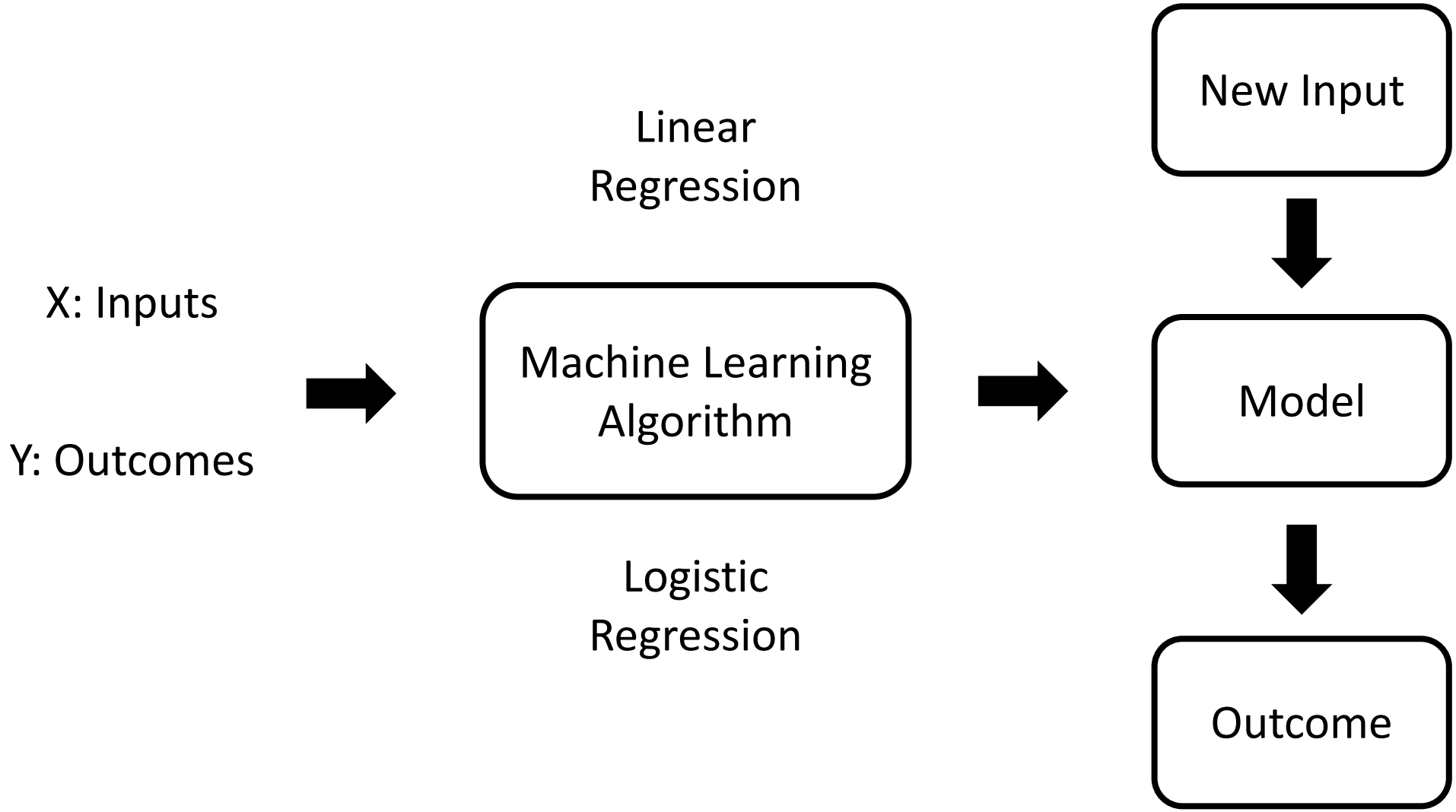
Vector/Matrix/Tensor predictor X (also called inputs, regressors, covariates, features, independent variables)

Outcome Y (also called dependent variable, response, target)

Objectives:

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

Supervised Learning

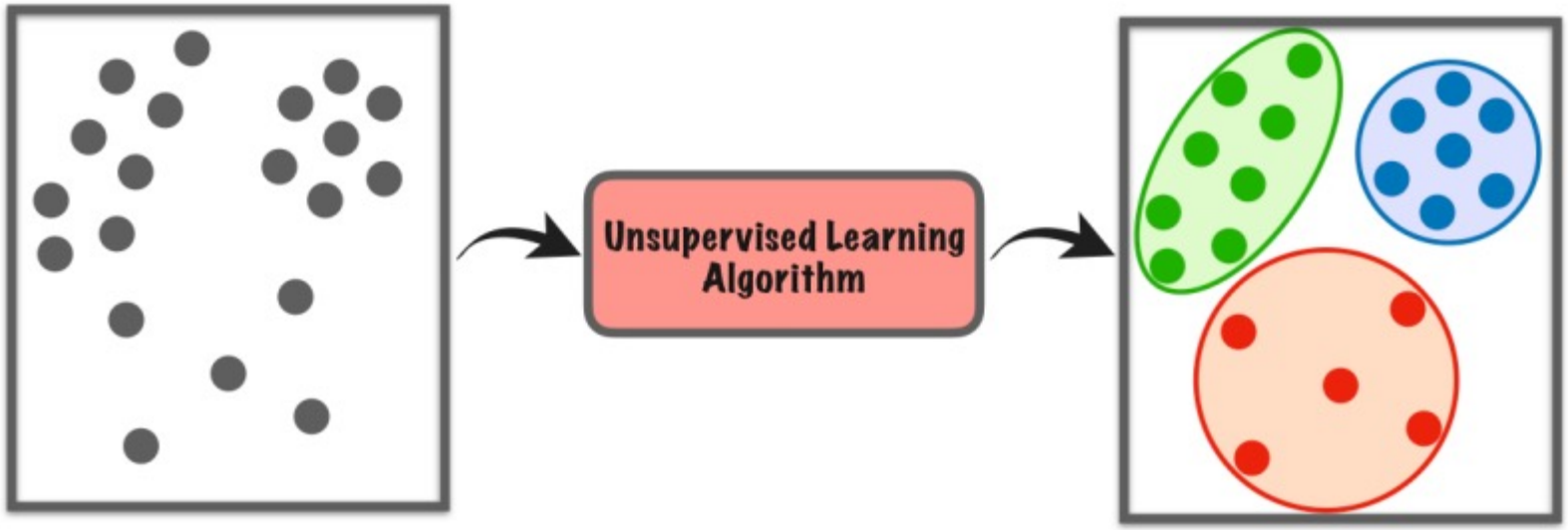


Unsupervised Learning

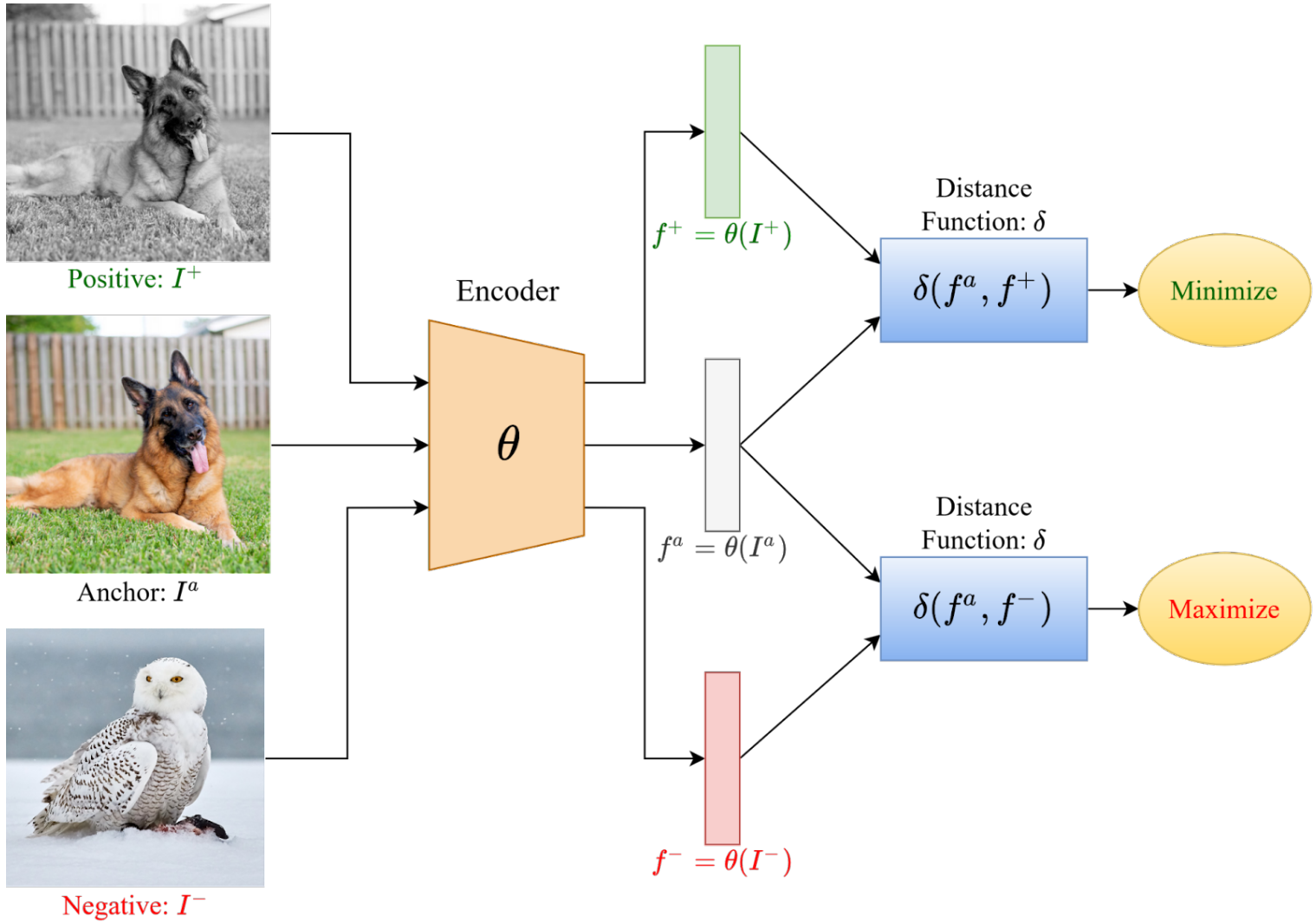
No outcome variable, just a set of predictors (features) measured on a set of samples.

Objectives:

1. Find groups of samples that behave similarly
2. Find the most important sets of features

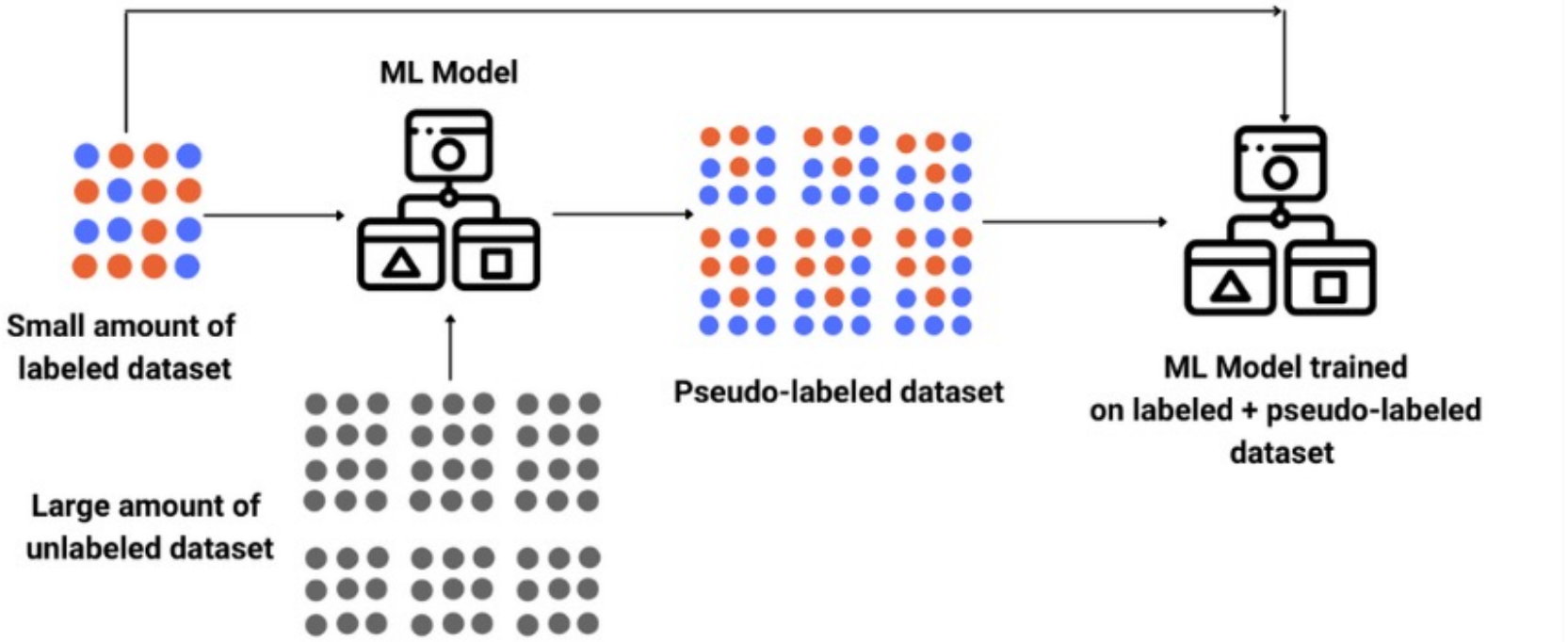


Unsupervised Learning

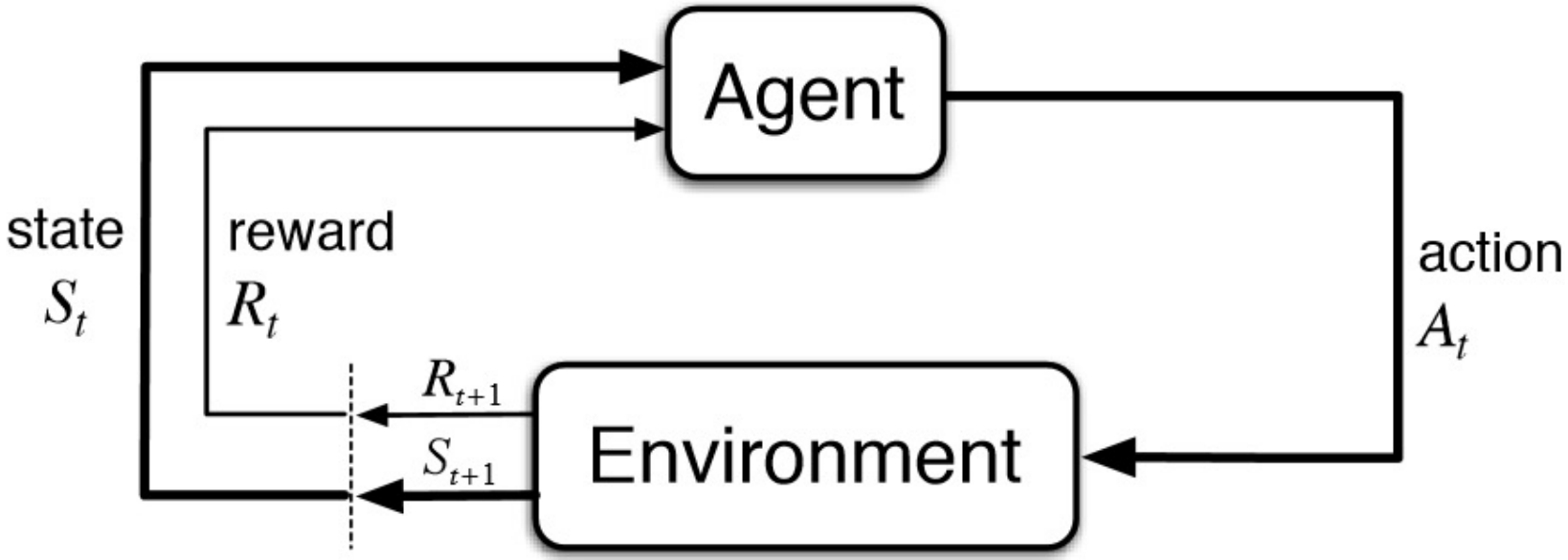


Semi-Supervised Learning

- Semi-supervised learning can be viewed as a mix of supervised and unsupervised learning.
- In semi-supervised learning tasks, some training examples have outcomes, but some do not.



Reinforcement Learning



Category of learning problems

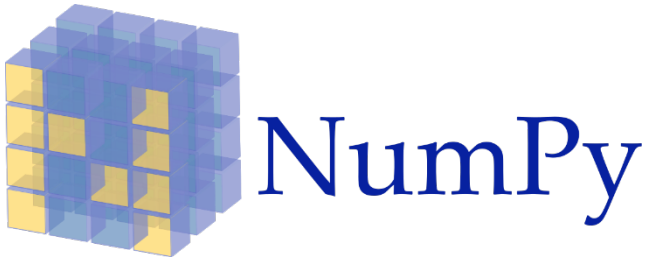
Spam Detection: Supervised or Unsupervised?

Digit Recognition: Supervised or Unsupervised?

Generative Models: Supervised or Unsupervised?

Text-to-Image Generation: Supervised or Unsupervised?

Implementation Tools



Python Resources

- Kaggle: https://www.kaggle.com/learn/python
- Learn Python: https://www.learnpython.org/
- Python for Beginners by Microsoft: https://learn.microsoft.com/en-us/shows/intro-to-python-development/
- Python for Everybody (up to 6 hours): https://www.youtube.com/watch?v=8DvywoWv6fI
- **Google it**

UMKC

Course Website

https://tianzheng4.github.io/umkc-teaching/2023-fall-teaching-1/

What are on the course website:

- Lecture slides
- Lab material
- Contact Information

If you are interested in my research, feel free to contact me.