CS584 Machine Learning

Instructor: Dr. Binghui Wang
Lecture hours: MW 10:00AM – 11:15AM
Office: Stuart Building, 228F
Office hours: Wednesday 1:00 PM– 3:00PM
Email: bwang70@iit.edu
Communication: Piazza (Prefer), Blackboard, Email

- Signup Link: piazza.com/iit/spring2022/cs584
- Access code: 1729

Teaching Assistants

Ruo Yang (ryang23@hawk.iit.edu)

- Office hours: Tuesday 10:00AM – 1200PM

Gai Hao (ghao3@hawk.iit.edu),

- Office hours: Thursday 10:00AM -12:00PM

Course Description

Topics

- Supervised learning
  - Regression, classification
- Unsupervised learning
  - Clustering, dimen. reduction
- Semi-supervised learning

Advanced topics

- Generative adversarial nets/contrastive learning
- Trustworthy machine learning
- Federated learning
- Compressive sensing/sparse coding/dic. learning
- Causal inference (guest lecture)
- Machine learning for security (guest lecture)

Topics not covered: Reinforcement learning, Boosting/Ensemble learning, Bayesian learning, Active learning, Sequential models, Sampling methods, etc.

Textbooks (free E-Book)

- Pattern Recognition and Machine Learning, by Christopher M. Bishop
- Machine Learning: A Probabilistic Perspective, by Kevin P. Murphy

Grades & Assignments

Grades

- 5 assignments + 1 additional optional assignment (50%)
- Final project (5-min presentation & report) (40%)
- Paper reading (10%)

Assignments

- May come from extra readings. Optional assignment 6 is to replace the minimal grade in the first 5 required assignments
- Assignment submission: Single PDF for solution or/and separate source code files, NOT using zip
- Late submission: All assignments are due by Sunday at 11:59 PM. 50% of the grade will be deducted for that assignment if it is late within 1 day (24 hours). 0 grade if the assignment is late more than 1 day

Initial & Final Project

2-3 students form a group

An initial project topic (along with the group members) is due at the end of Spring Break

- Let me/TAs know as early as possible if you decide to change the topic after the due date

The final project includes a 5-min project representation (5%) and a final report (35%)

- The format of the final report should follow an academic paper (i.e., introduction, related work, problem, methods, results, conclusion)

- ALL group members should clearly clarify their contributions in the project

- You can use the NeurIPS template to organize the report

Project Lists

Deep learning theory

- Information bottleneck, Neural Tangent Kernel, ReduNet, Causality

Unsupervised representation learning

- Disentangled representation learning, causal representation learning, contrastive learning, self-supervised learning, variational methods

Semi-supervised learning

- Graph neural networks, graph embedding, mixup, self-training, co-training

Trustworthy machine learning (DNN, graph learning, federated learning, etc.)

- Security attacks (evasion, data/model poisoning, backdoor);
- Empirical defense (adversarial training, robust optimization); Provable defense (randomized smoothing, IBP)

Privacy-preserving machine learning (DNN, graph learning, federated learning, etc.)

- Privacy attacks (model stealing/inversion, property/attribute inference)
- Privacy preserving (differential privacy, Crypto, information theory)

Machine learning for security

- Blockchain security, Network security, Software/hardware security, Cyber-Physical System security

Project Lists (Continue)

Large-scale machine learning (High-dimensional/massive data)

- Randomized algorithms
- Streaming, sketching, compressive sampling, ...

Federated learning

- Communication efficient, computational-efficient, personalization, fairness

Meta learning

- Model-agnostic meta learning

Fair machine learning

Interpretable machine learning

Machine unlearning

(Deep) compressive sensing/sparse coding/dictionary learning

Other topics you may be interested in…

Academic Integrity & Course Recordings

Can discuss assignments with classmates, but all final work MUST be your own

Academic dishonesty of any kind may result in

- a 0 grade on the assignment
- a reduction in final grade
- and/or referral to the Dean

IIT code of Academic Honesty

- https://www.iit.edu/student-affairs/student-handbook/fine-print/code-academic-honesty

Class recordings are the intellectual property of the university or instructor and are reserved for use only by students in this class and only for educational purposes

Students SHOULD not post or otherwise share the recordings outside the class in any form

What is Machine Learning?

Arthur Samuel: Machine Learning is a field of study that gives computers the ability to learn without being explicitly programmed

Shapire: Machine learning studies how to automatically learn to make predictions based on past observations

Kevin Murphy: Machine learning as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data

Learning from Past Data/Observations

Future Data

Function f

Predict

"car"

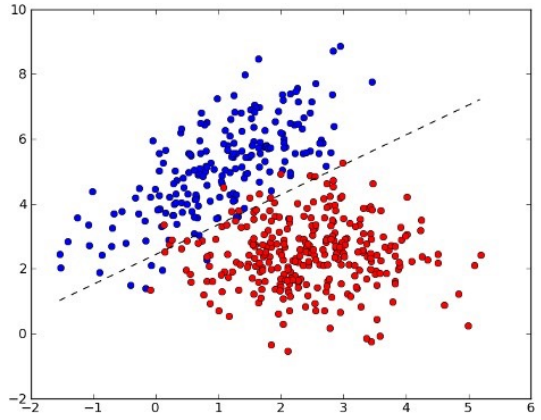
Train

Past Data
(perhaps with **labels**)

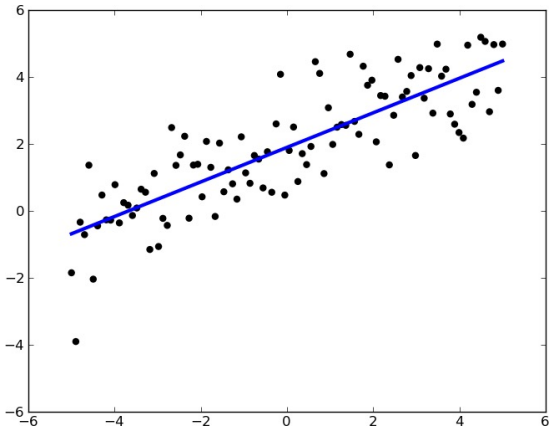
Machine Learning Draws Inspiration & Concepts from Many Scientific Fields

Statistics: Inference from data, probabilistic models, learning theory...
Mathematics: Optimization theory, numerical methods, tools for theory…
Engineering: Signal processing, robotics, control, information theory…
Economics: decision theory, operations research, econometrics...
Psychology/Cognitive science: Computational linguistics, learning, reinforcement learning, movement control...
Physics: Energy minimization principles, entropy…
Computational neuroscience: Neural networks, principles of neural information processing, ...

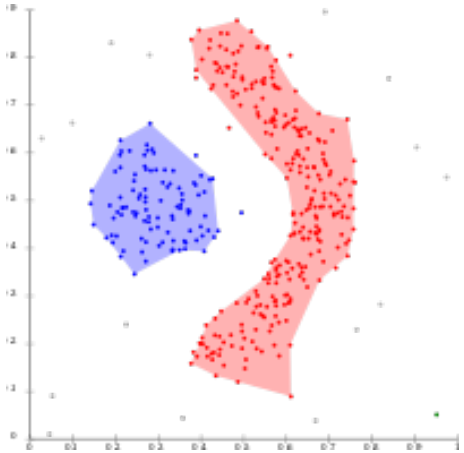
Machine Learning Tasks



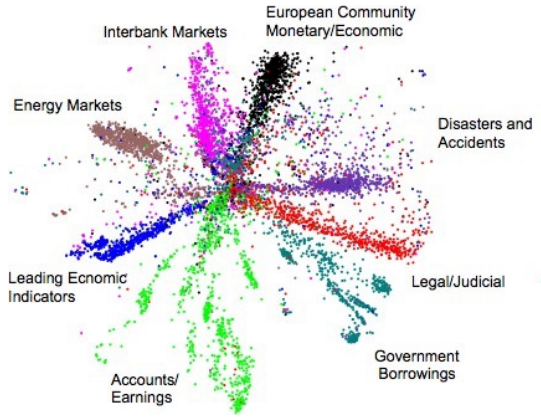
Classification



Regression

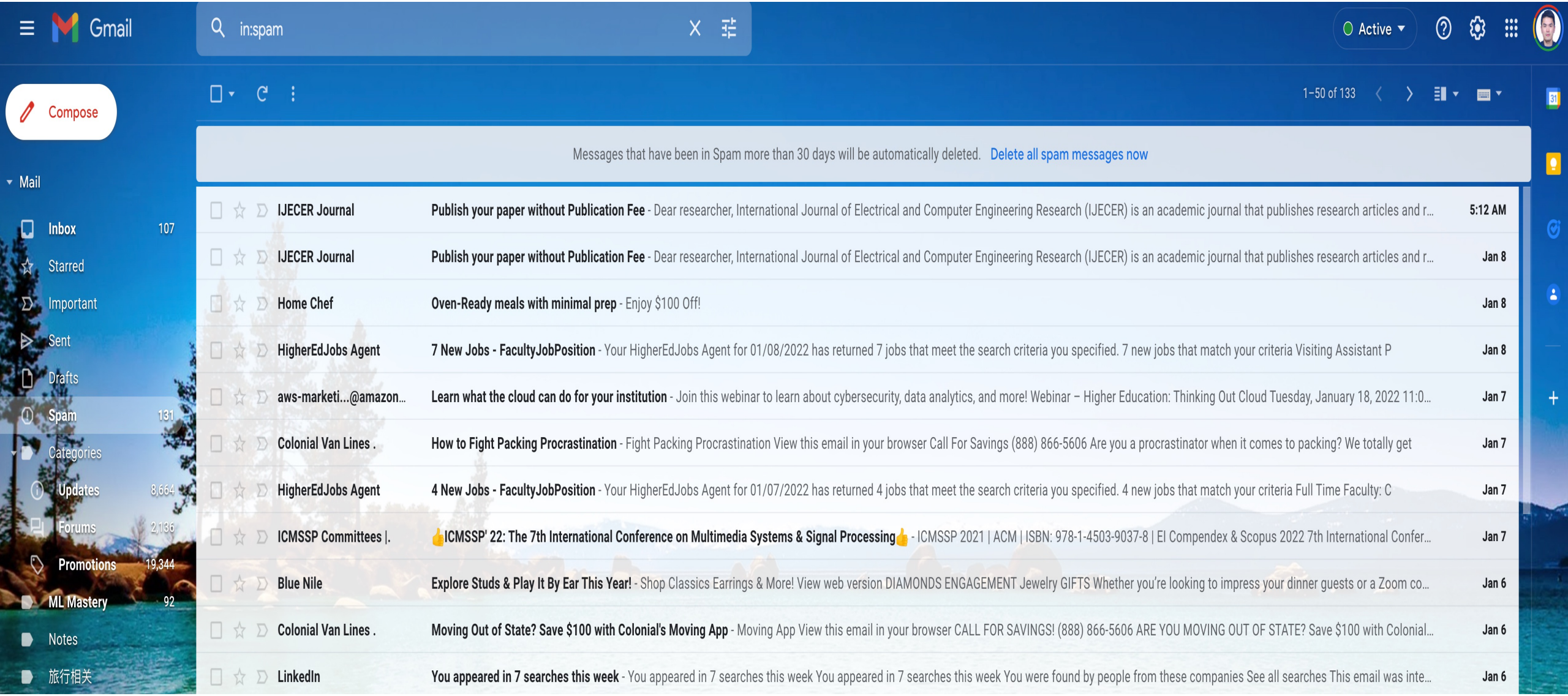


Clustering



Dimensionality Reduction

Classification: Spam Detection



Classification: Spam Detection

Sir / Madam,

We invite you to submit your manuscript(s) for publication. The journals include research papers, review articles, technical projects and short communications containing new insight into any aspect of the covered scope of the journal. Our objective is to inform authors of the decision on their manuscript(s) within weeks of submission. After acceptance, the paper will be published in the current issue immediately.
Keywords: English, Literature, Science, Economics, Engineering, Management, Agriculture, Horticulture, Environment .....

International Journal of Advanced Engineering Research and Science (IJAERS) ISSN: 2456-1908(O) | 2349-6495 (P)

DOI (CrossRef): 10.22161/ijaers

Thomson Reuters ResearcherID: P-3738-2015

Impact Factor: 4.192, SJIF: 4.072, IBI: 3.2, PIF: 2.465, ISRA-JIF: 1.317,

Website: http://www.ijaers.com

Kindly submit research articles to **http://ijaers.com/submit-paper/** or mail us at **editor.ijaers@gmail.com**

International Journal of English, Literature and Science (IJELS) ISSN: 2456-7620

New Email

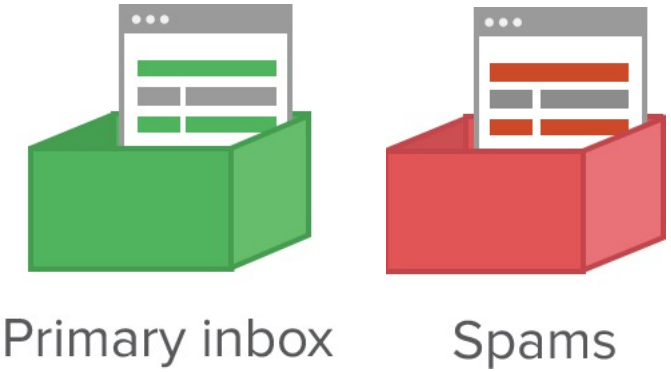
Classifier

Predict

Is Spam

Train

Binary classification



Primary inbox

Spams

Emails with Labels

Classification: Face Recognition

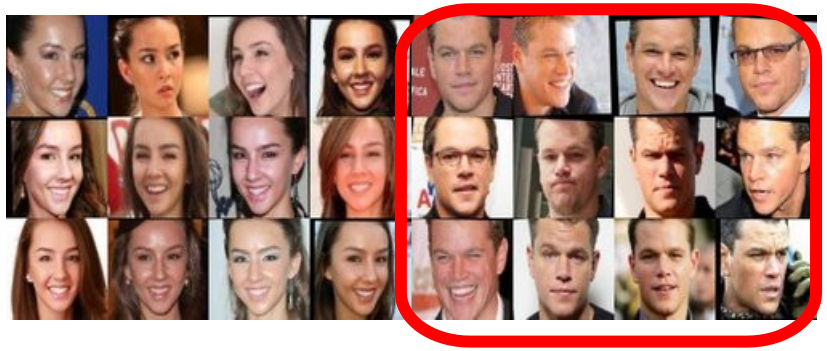
Classifier

Predict

"Matt Damon"

Train

multi-class



Faces & Names

Classification: Image Recognition

Classifier

Predict

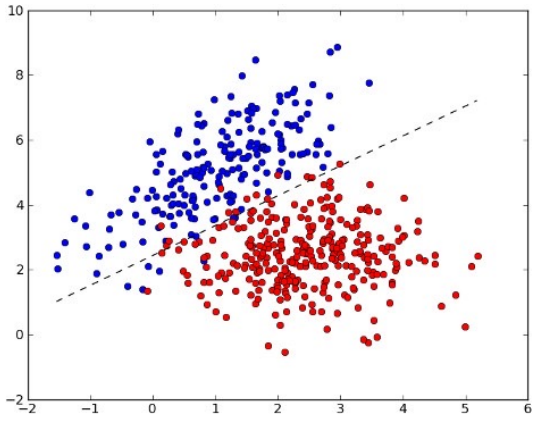
"Dog"
"Woman"

Train

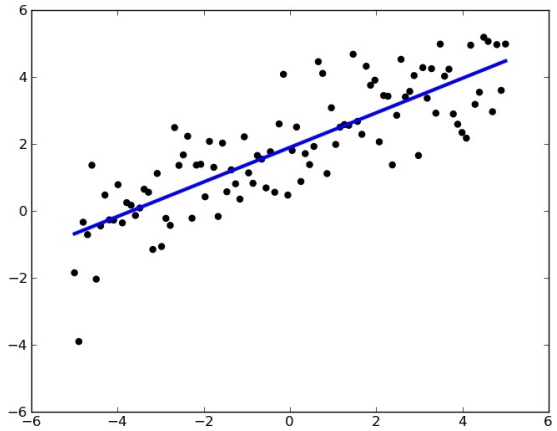
multi-class, multi-label

Images & Labels

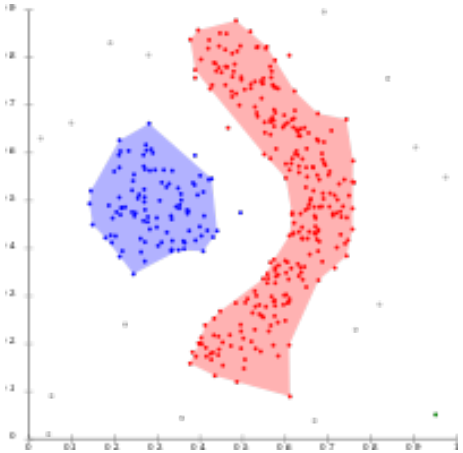
Machine Learning Tasks



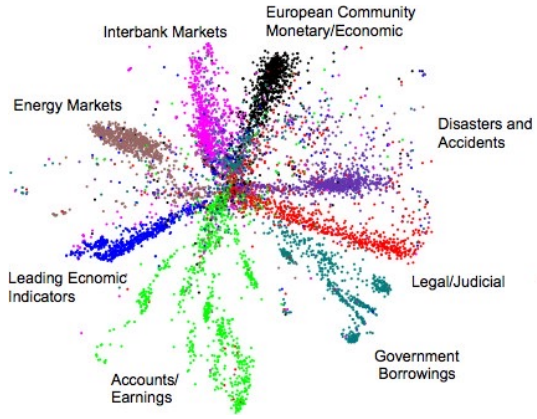
Classification



Regression



Clustering



Dimensionality Reduction

Regression: Housing Price

Features of a House

Regressor

Predict

Price = $0.5M

Train

label ← **Price** — Selling price of the property ($1,000)

features

Beds	Number of bedrooms in the house
Baths	Number of bathrooms in the house
Square Feet	Size of the house in square feet
Miles to Resort	Miles from the property to the downtown resort area
Miles to Base	Miles from the property to the base of the ski resort's mountain
Acres	Lot size in number of acres
Cars	Number of cars that will fit into the garage
Years Old	Age of the house, in years, at the time it was listed
DoM	Number of days the house was on the market before it sold

Regression VS. Classification

- **Regression: labels are continuous and ordered**

House Prices: $324K < $521K < $1.2M



Regression VS. Classification

- **Classification: labels are categorical**

Categories:

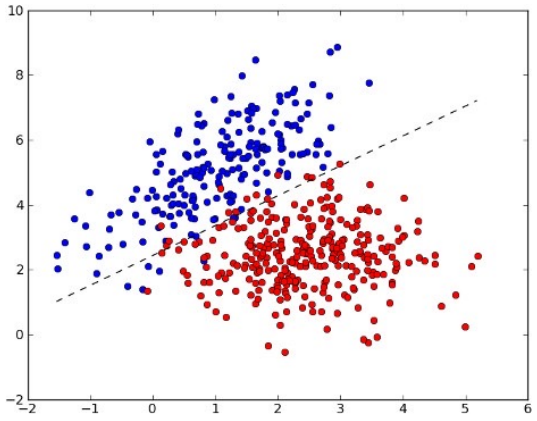
Class1

Class 2

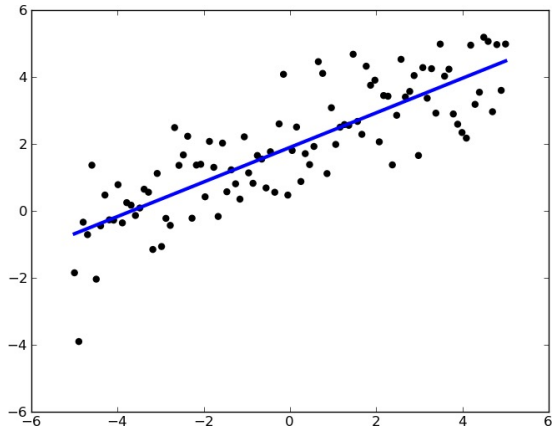
Class 3

"Class 1" and "Class 3" are not ordered and cannot be compared!

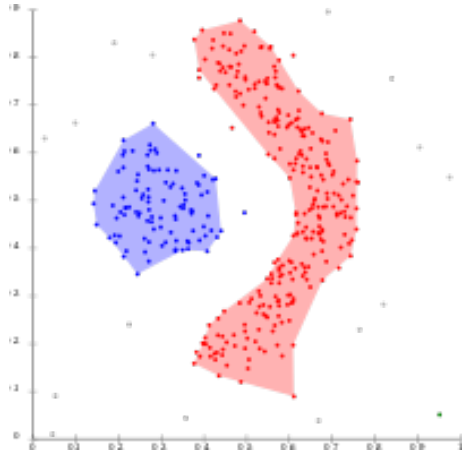
Machine Learning Tasks



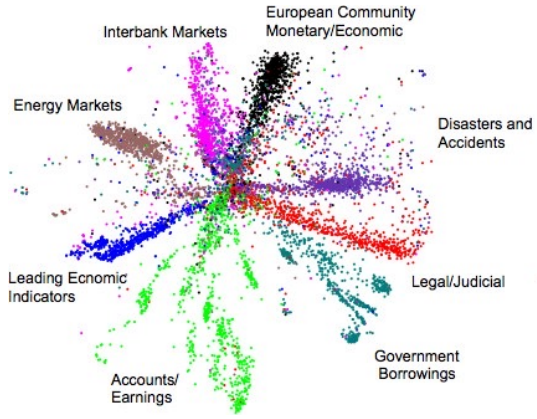
Classification



Regression



Clustering



Dimensionality Reduction

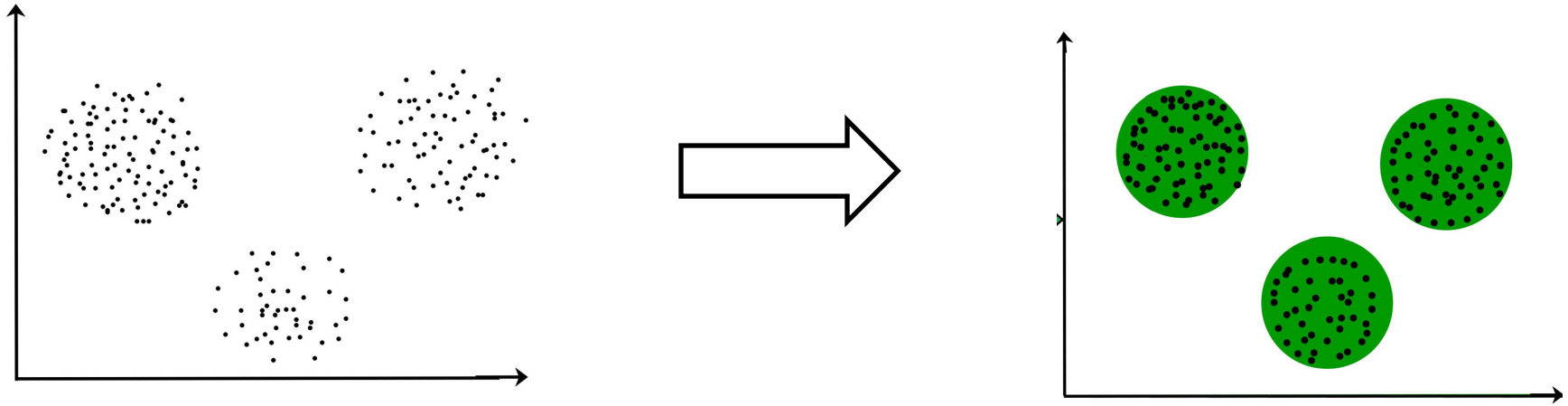
Clustering

Task: divide a set of data points into a number of groups such that

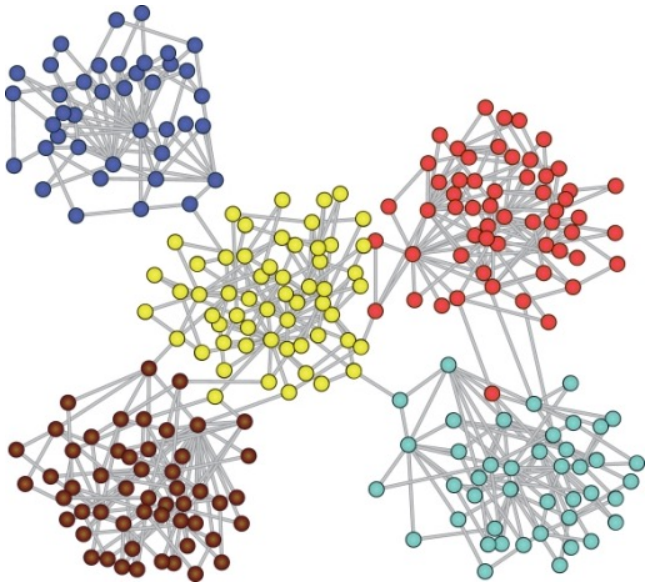
- data points in the same groups are similar to each other
- data points from different groups are dissimilar to each other

Input: data points $\mathbf{x}1, \mathbf{x}2, \cdots, \mathbf{x}n$

Output: predict clusters $y1, \cdots, yn \in \{1, \cdots, k\}$, k is #clusters



Clustering: Applications



Community detection

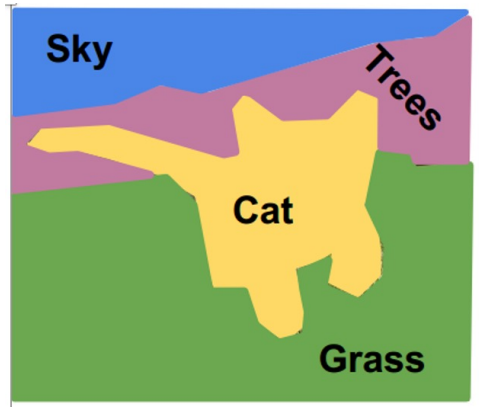
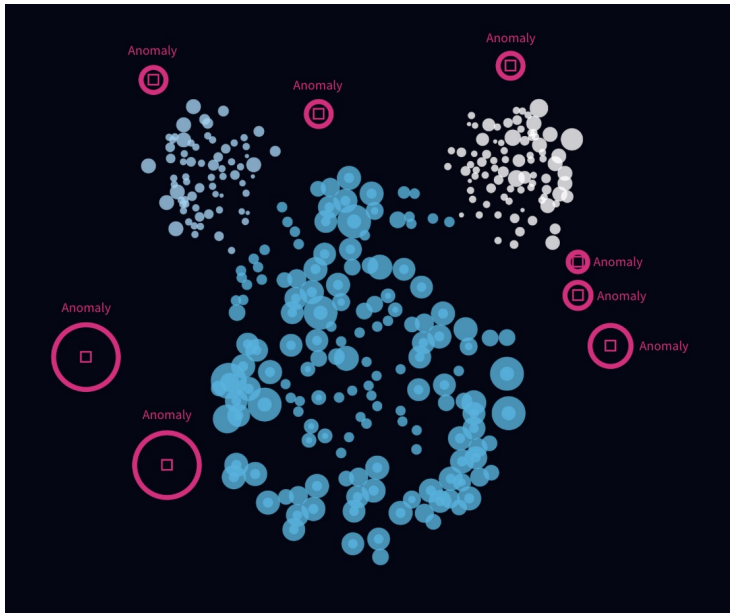
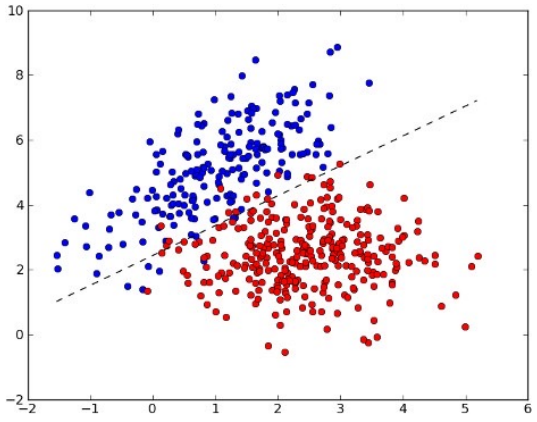


Image segmentation

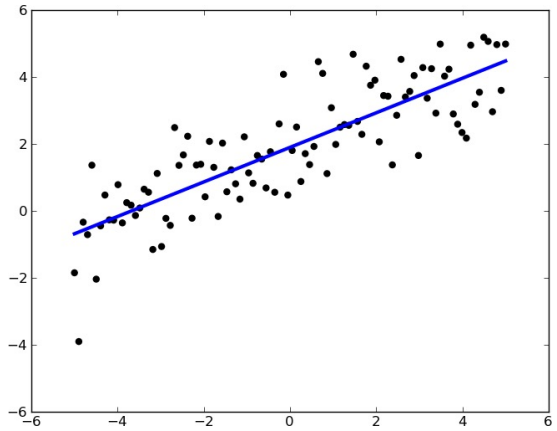


Anomaly detection

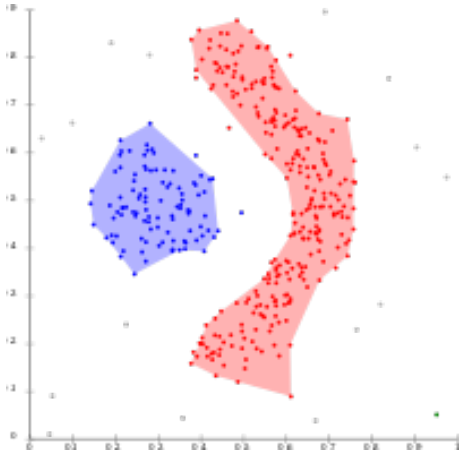
Machine Learning Tasks



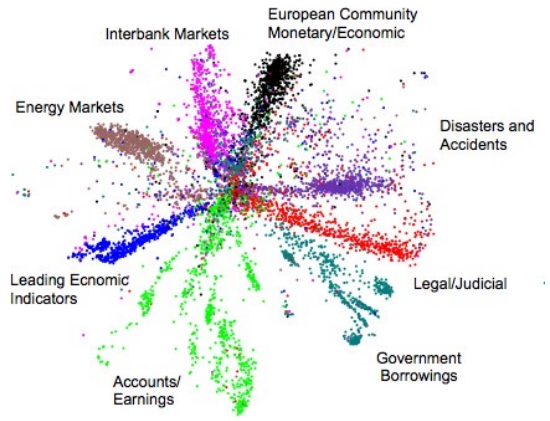
Classification



Regression



Clustering



Dimensionality Reduction

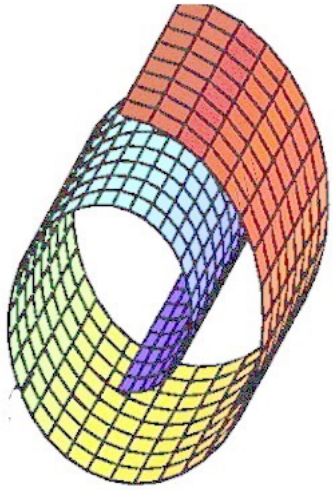
Dimensionality Reduction

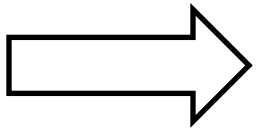
Task: map high-dim data points into a low-dim space such that

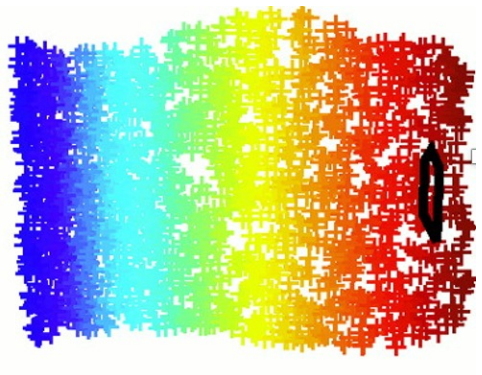
- Global information of data points (e.g., variance) can be kept
- Local information of data points (e.g., neighborhood) can be kept

Input: high-dim data points $\mathbf{x}1, \mathbf{x}2, \cdots, \mathbf{x}n$

Output: low-dim data representations $y1, y2, \cdots, yn$





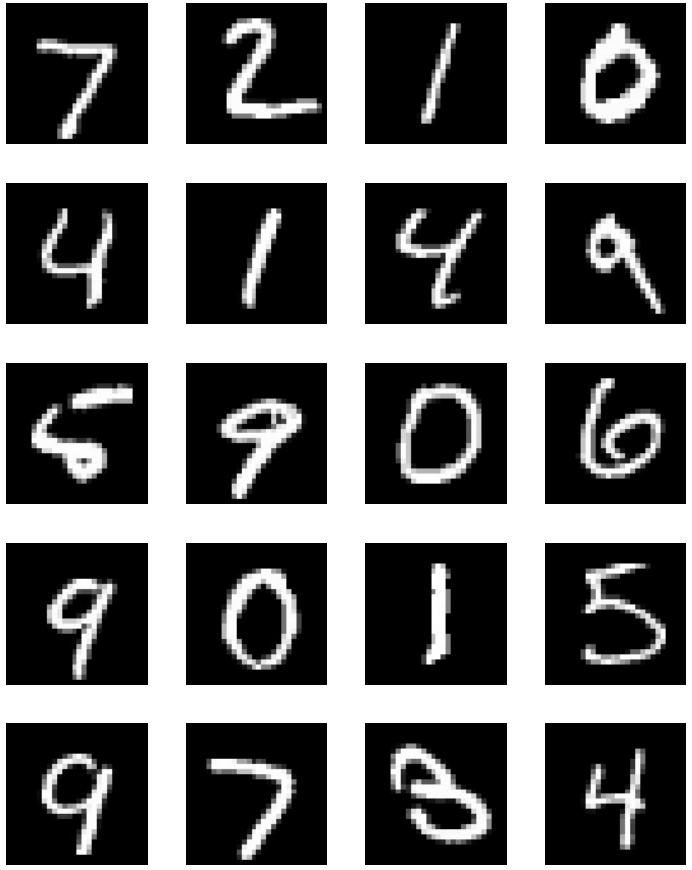


Visualization

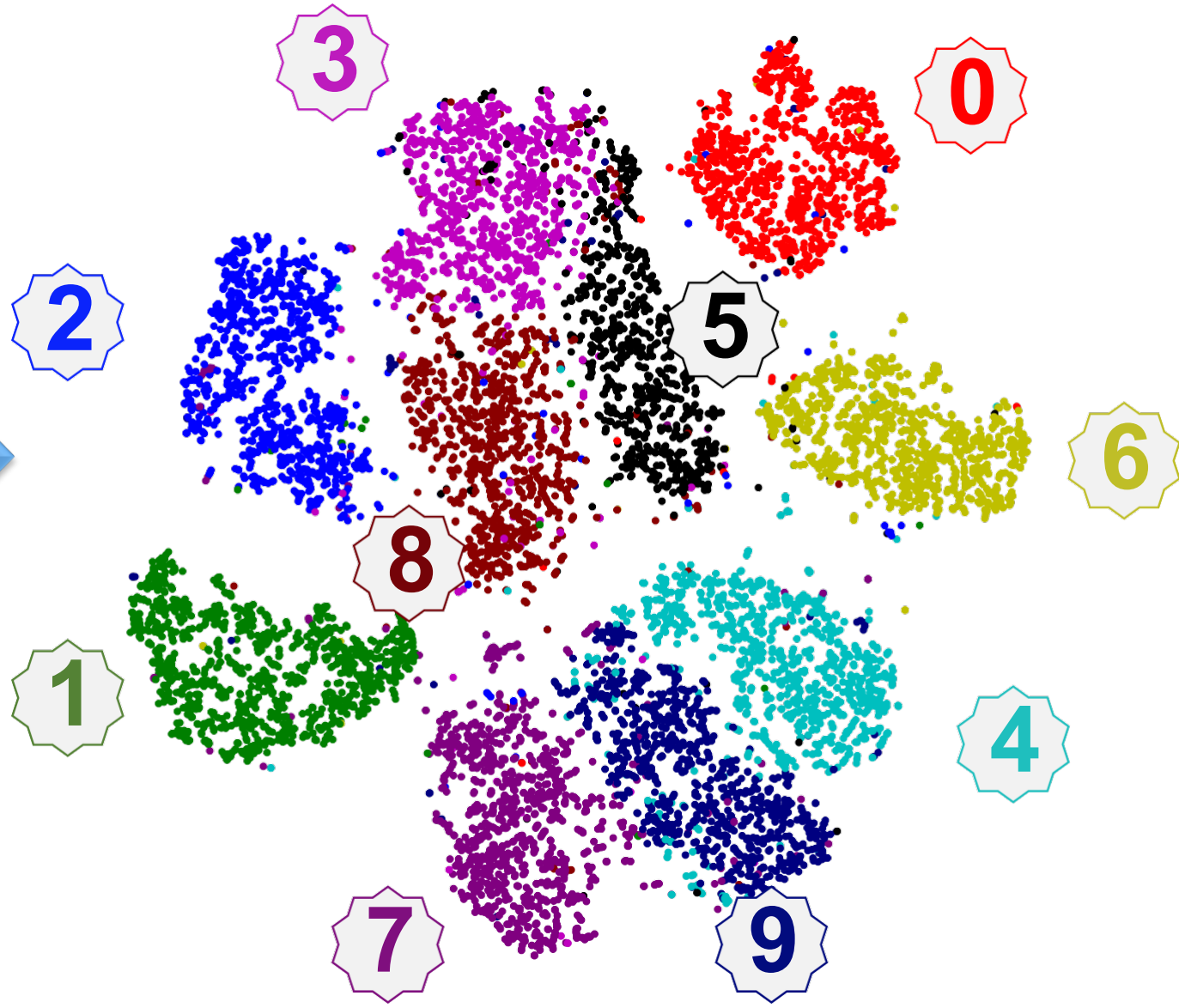
Data compression

Data preprocessing

Visualization: Handwritten Digit Data



60K images (size 28×28)



Supervised vs. Unsupervised Learning

Supervised learning : learn from labeled data

- Classification and regression are supervised learning tasks

Unsupervised learning : learn from unlabeled data

- Clustering and dimensionality reduction are unsupervised learning tasks

Machine Learning
Tasks & Methods

Tasks

Methods

Regression

Classification

Clustering

Dim Reduction

Tasks

Methods

Regression

Classification

Clustering

Dim Reduction

Tasks

- Regression
- Classification
- Clustering
- Dim Reduction

Methods

- Linear Regression ✓
- Polynomial Regression
- Ridge Regression ✓
- LASSO ✓
- Elastic Net

Tasks

Methods

Regression

Classification

Clustering

Dim Reduction

Tasks

Regression

Classification

Clustering

Dim Reduction

Methods

Logistic Regression

Naïve Bayes

Neural Networks

SVM

Nearest Neighbor

Decision Tree / Boosting

Tasks

Methods

Regression

Classification

Clustering

Dim Reduction

Tasks

Regression

Classification

Clustering

Dim Reduction

Methods

DBSCAN

K-means ✓

GMM ✓

Spectral Clustering ✓

Affinity Propagation

Sparse sub. Clustering ✓

Tasks

Regression

Classification

Clustering

Dim Reduction

Methods

Tasks

Regression

Classification

Clustering

Dim Reduction

Methods

PCA/SVD

Nonnegative MF

LLE/LE

Deep Belief Net

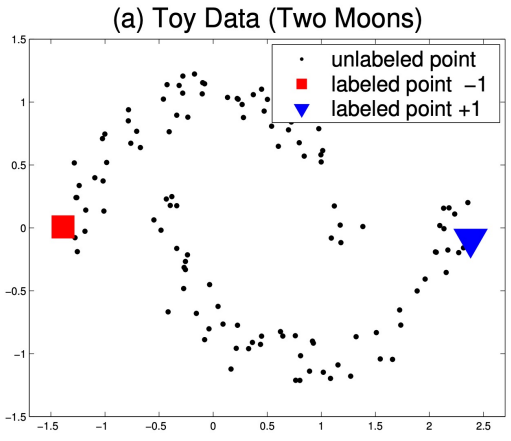
Random Projection

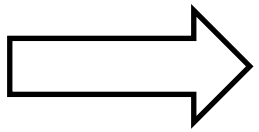
Semi-supervised Learning

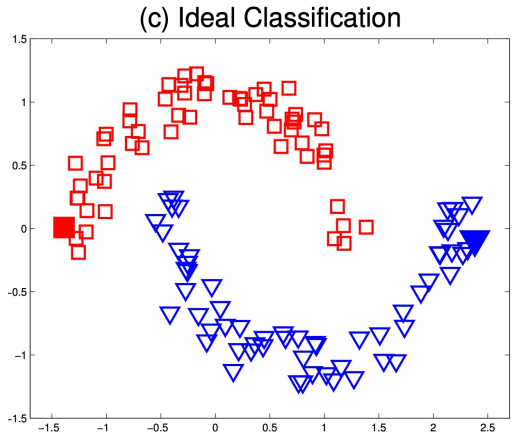
Supervised learning : learn from **labeled** data

Unsupervised learning : learn from **unlabeled** data

Semi-supervised learning : learn from both **labeled** and **unlabeled** data







Why/When Semi-supervised Learning?

Labeled data is insufficient/expensive, unlabeled data is sufficient/cheap

- Imagine the cost/time of tagging millions of images!

Speech analysis

- Switchboard dataset
- 400 hours annotation time for 1 hour of speech

Natural Language Processing

- Penn Chinese Treebank
- 2 years of 4,000 sentences

Semi-supervised Learning Methods

- Self-Training
- Co-Training
- Semi-supervised SVM
- Generative methods, mixture models
- Mixup & MixMatch
- Graph-based methods
  - Label propagation
  - Belief propagation
  - Graph neural networks

Semi-supervised Learning Methods

- Self-Training
- Co-Training
- Semi-supervised SVM
- Generative methods, mixture models
- Mixup & MixMatch
- Graph-based methods
  - Label propagation
  - Belief propagation
  - Graph neural networks

Acknowledgement

Some slides are from **Shusen Wang**
https://github.com/wangshusen/DeepLearning

Tasks

- Regression
- Classification
- Clustering
- Dim Reduction

Methods

- Linear Regression
- Polynomial Regression
- Ridge Regression
- LASSO
- Elastic Net