CZ4041/CE4041: Machine Learning

Lecture 1a: Introduction

Sinno Jialin PAN

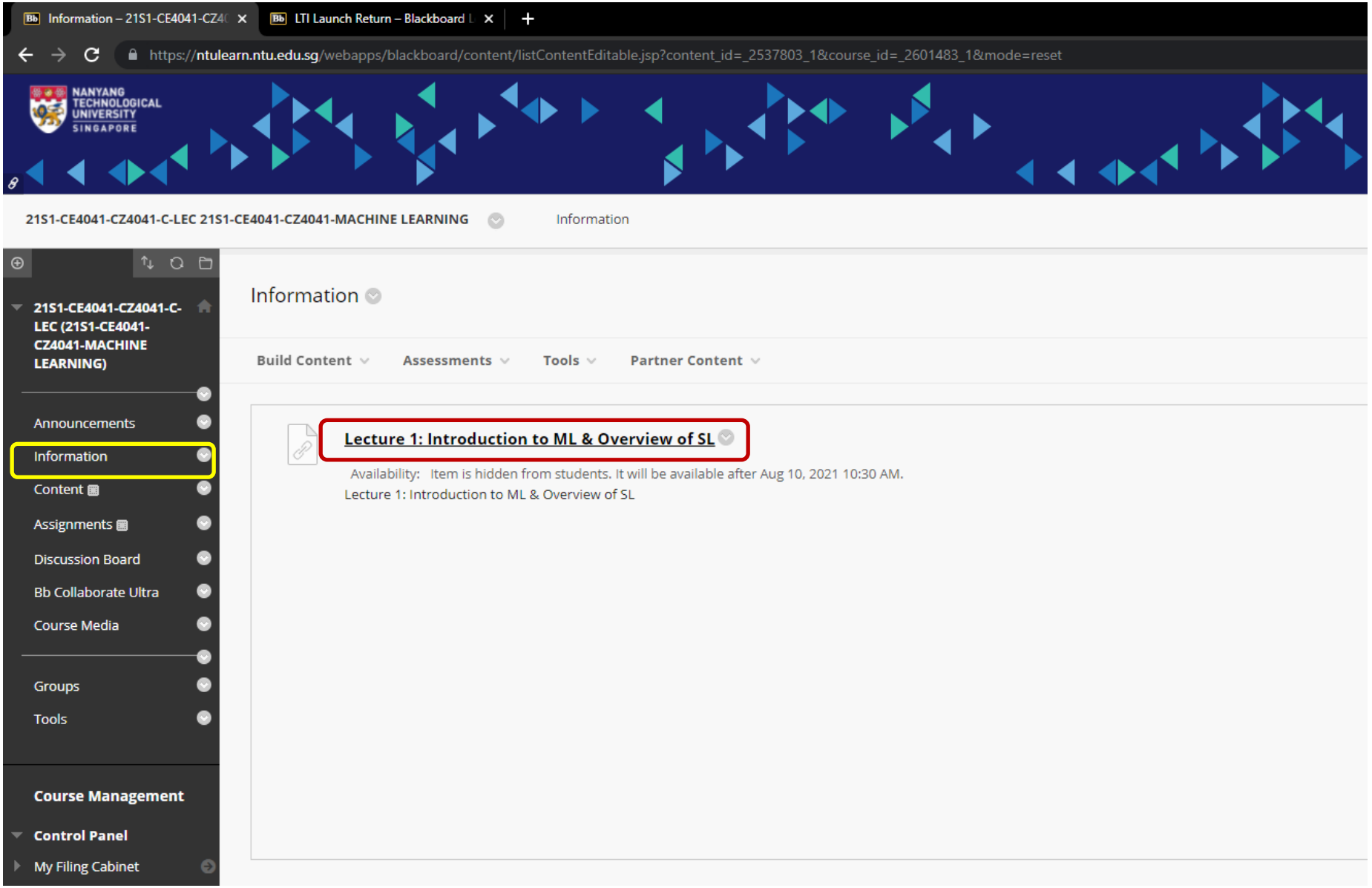
School of Computer Science and Engineering

NTU, Singapore

Homepage: https://personal.ntu.edu.sg/sinnopan/

General Information

- Instructors: Dr. Sinno Jialin PAN (First half: Week 1-6) and Dr. Kelly Yiping KE (Second half: Week 7-12)
- Lecture time/venue (first half)
  - Week 1-6, Tuesdays 11:30am – 1:30pm
  - Online via MS Teams
    - CZ/CE4041 in NTULearn → Information → Teams link (a NEW link will be available at 10:30am on every Tuesday)
- Tutorial time/venue (first half)
  - Starting from Week 2 (Week 2, 4, 6), Thursdays 1:30 – 2:30pm
  - Online via MS Teams
    - CZ/CE4041 in NTULearn → Information → Teams link (a NEW link will be available at 12:30pm on every Thursday)


Information – 21S1-CE4041-CZ4
LTI Launch Return – Blackboard
https://ntulearn.ntu.edu.sg/webapps/blackboard/content/listContentEditable.jsp?content_id=_2537803_1&course_id=_2601483_1&mode=reset
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
21S1-CE4041-CZ4041-C-LEC 21S1-CE4041-CZ4041-MACHINE LEARNING
Information
21S1-CE4041-CZ4041-C-LEC (21S1-CE4041-CZ4041-MACHINE LEARNING)
Announcements
Information
Content
Assignments
Discussion Board
Bb Collaborate Ultra
Course Media
Groups
Tools
Course Management
Control Panel
My Filing Cabinet
Information
Build Content
Assessments
Tools
Partner Content
Lecture 1: Introduction to ML & Overview of SL
Availability: Item is hidden from students. It will be available after Aug 10, 2021 10:30 AM.
Lecture 1: Introduction to ML & Overview of SL

General Information (cont.)

- Q&A (regarding Week 1-6 teaching content & course project)
  - Send questions via email sinnopan@ntu.edu.sg
  - Make an appointment via email (regarding first-half)
- Course Webpage
  - CZ/CE 4041 @ NTULearn (official course webpage)
  - https://personal.ntu.edu.sg/sinnopan/cz4041.html (check information when NTULearn is down, Week 1-6 only)

CZ/CE 4041: Machine Learning

Note: the official course webpage of CZ/CE 4041 is in NTULearn (21S1-CE4041-CZ4041-C-LEC 21S1-CE4041-CZ4041-MACHINE LEARNING). This webpage is only used when NTULearn is down and during the add/drop period.

To attend the weekly live class on Tuesdays (lectures) and Thursdays (tutorials), please login the course webpage in NTULearn, go to "Information", and click the link. A new link will be available after 10:30am on every Tuesday for lectures and after 12:30pm on every Thursday for tutorials.

If NTULearn is down or not stable, you can click the link shown below to join the online class

MS Teams link for the class on Aug. 10, 2021

Sinno J. Pan@NTU, Singapore

Evaluation

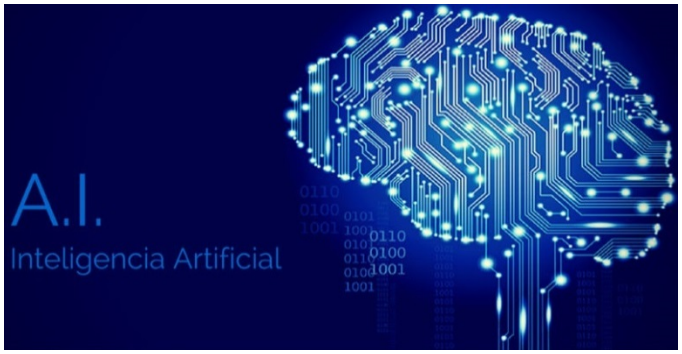
- Course project (40%)
  - Group-based (maximal size: 4 students)
  - Course report (30%) + presentation video (10%)
  - Either a Kaggle competition or a research topic. Detailed information including assessment criteria about the course project will be released in the tutorial session in Week 2
- Open book final exam (60%)
  - 2 hours
  - Note: in the case that the final exam is cancelled, an open book quiz in Week 13 or 14 will be a replacement (details will be released if necessary)

What Is New in AY 2021

- Starting form AY 2021, the Machine Learning module will be offered twice per academic year
  - In the past, it was only offered in Sem 2 for full-time/part-time students
  - Past exam papers set for Sem 1 (AY 2020 and before) do not have any reference value
- The teaching materials (Week 7-12) would be slightly different from pervious years
  - In the past 3 years, I was the solo instructor for the Machine Learning module
  - While for this semester, Prof. Ke will teach for the 2nd half

Hot Keywords in the IT Sector

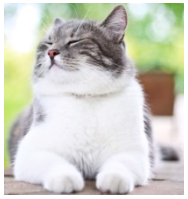
Machine Learning



What is Machine Learning?

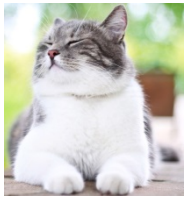
- Motivated by how human beings learn from examples/experience/exercise





- Focuses on the development of computer programs that can teach themselves to grow from data and change when exposed to new data

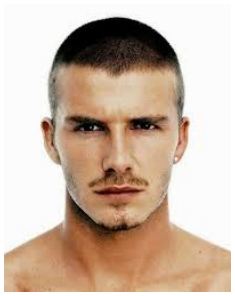




A Motivating Example

- Given a face image, to classify the face gender: 






- Once upon a time, to develop an AI system to solve such a task, developers or domain experts need to provide rules and implement them in the system

If the face has long hair and does not have moustache, then this is a "female" face;
If the face has short hair and moustache, then this is a "male" face.

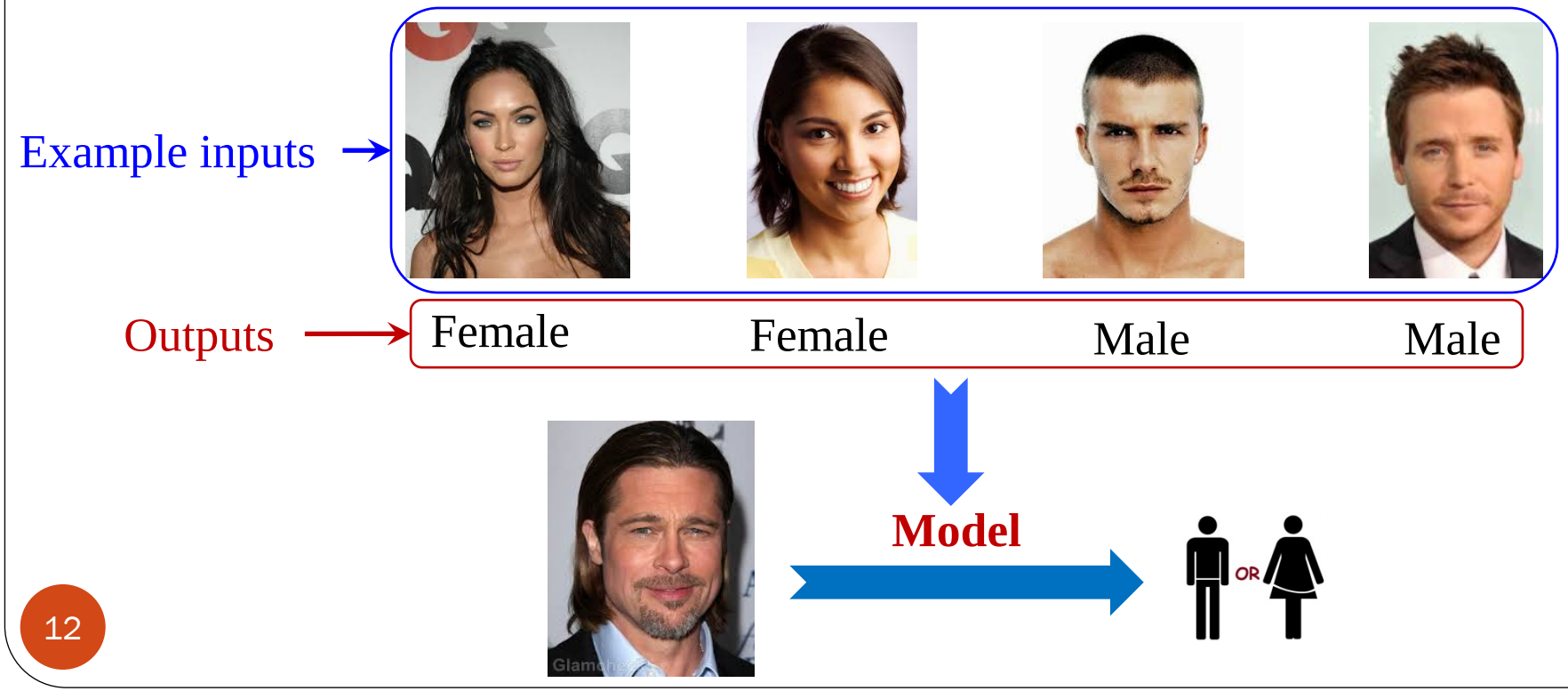
A Motivating Example (cont.)

- Limitations:
  - Time consuming
  - The defined rules may not be completed
  - Not able to handle uncertainty

If the face has long hair and does not have moustache, then this is a “female” face;
If the face has short hair and moustache, then this is a “male” face.

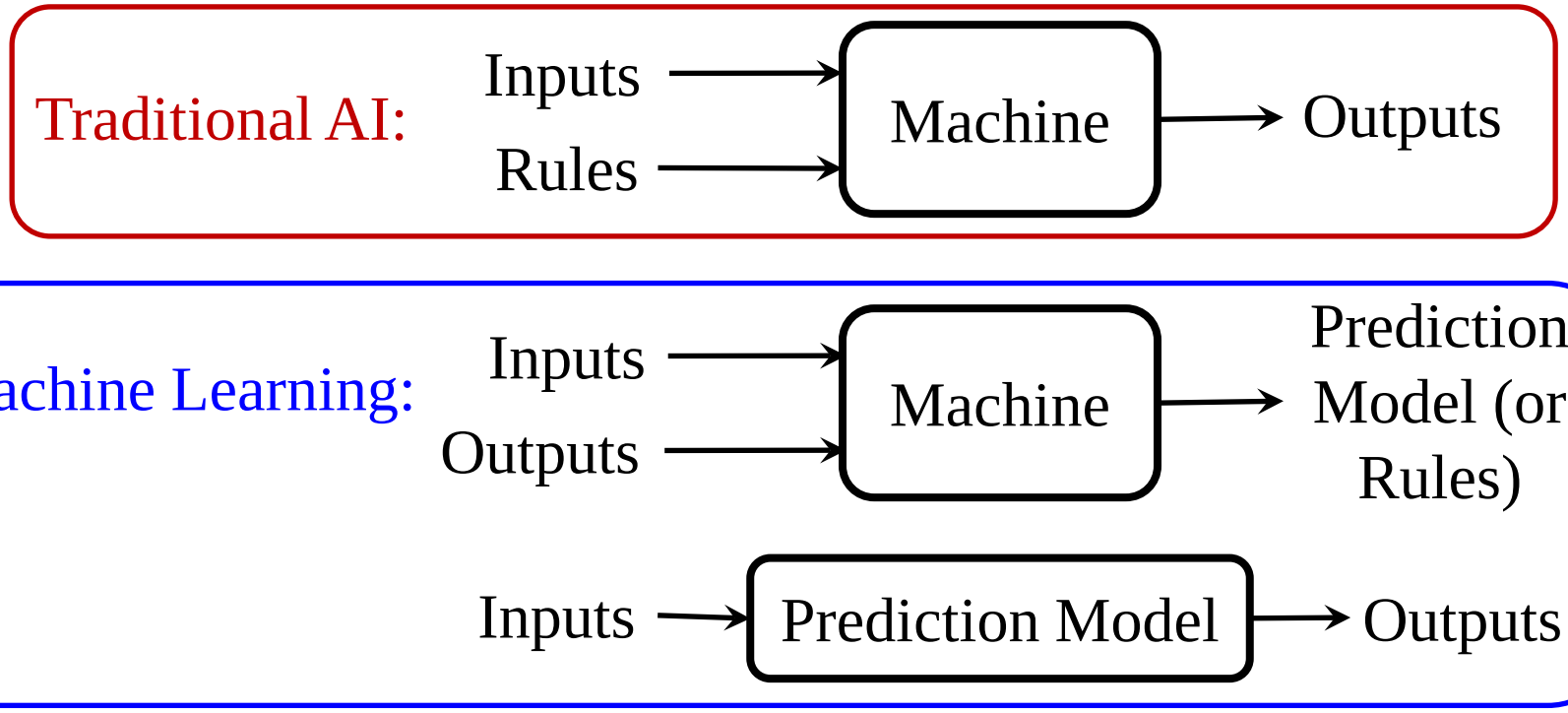
A Motivating Example (cont.)

- How about letting the machine learn the rules by itself?
  - The computer is presented with example inputs and their desired outputs, and the goal is to "learn" a set of general rules or "model" that maps inputs to outputs



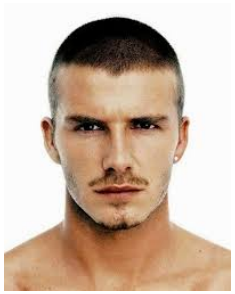
Machine Learning Definition

- Machine learning is a type of artificial intelligence that provides computers with the ability to learn from examples/experience without being explicitly programmed



How to Represent an Example?





- Feature engineering (not machine learning focus)
- Representation learning (one of the crucial research topics in machine learning)
  - Deep learning is the current most effective approach to representation learning

Machine Learning $\stackrel{?}{=}$ Deep Learning $\stackrel{?}{=}$ AI

- Machine learning is a field of AI – many other fields
- Deep learning is a type of methodologies of machine learning – many other methodologies in machine learning
- Machine learning has become a primary mechanism for data analytics (key in data science)
- Nowadays, machine learning is more and more interdisciplinary:
  - Distributed/parallel computing + machine learning → Distributed/parallel machine learning
  - Machine learning + hardware → AI chips

Different Paradigms/Settings

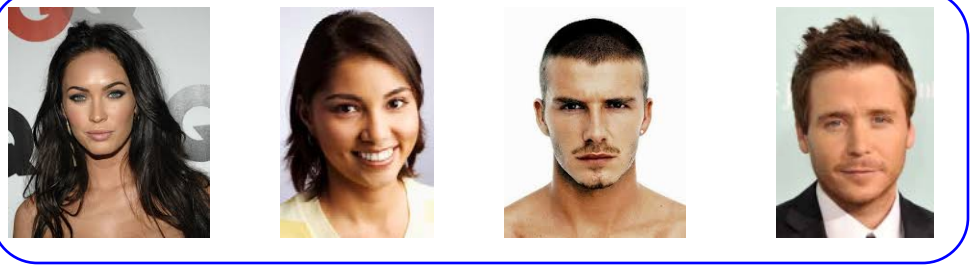
- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Advanced paradigms:
  - Semi-supervised learning
  - Active learning
  - Transfer learning

Supervised Learning

- The examples presented to computers are pairs of inputs and the corresponding outputs, the goal is to "learn" a mapping or model from inputs to labels

Labeled training data

Inputs: Face images



Outputs: Female or Male

Female Female Male Male

f: **label** = f(**input**)

Outputs are discrete (i.e., categorical) values → classification

Labels are continuous values → regression

Supervised Learning – Regression I

Supervised Learning – Regression II

Stock price prediction

Different Paradigms/Settings

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Advanced paradigms:
  - Semi-supervised learning
  - Active learning
  - Transfer learning

Unsupervised Learning

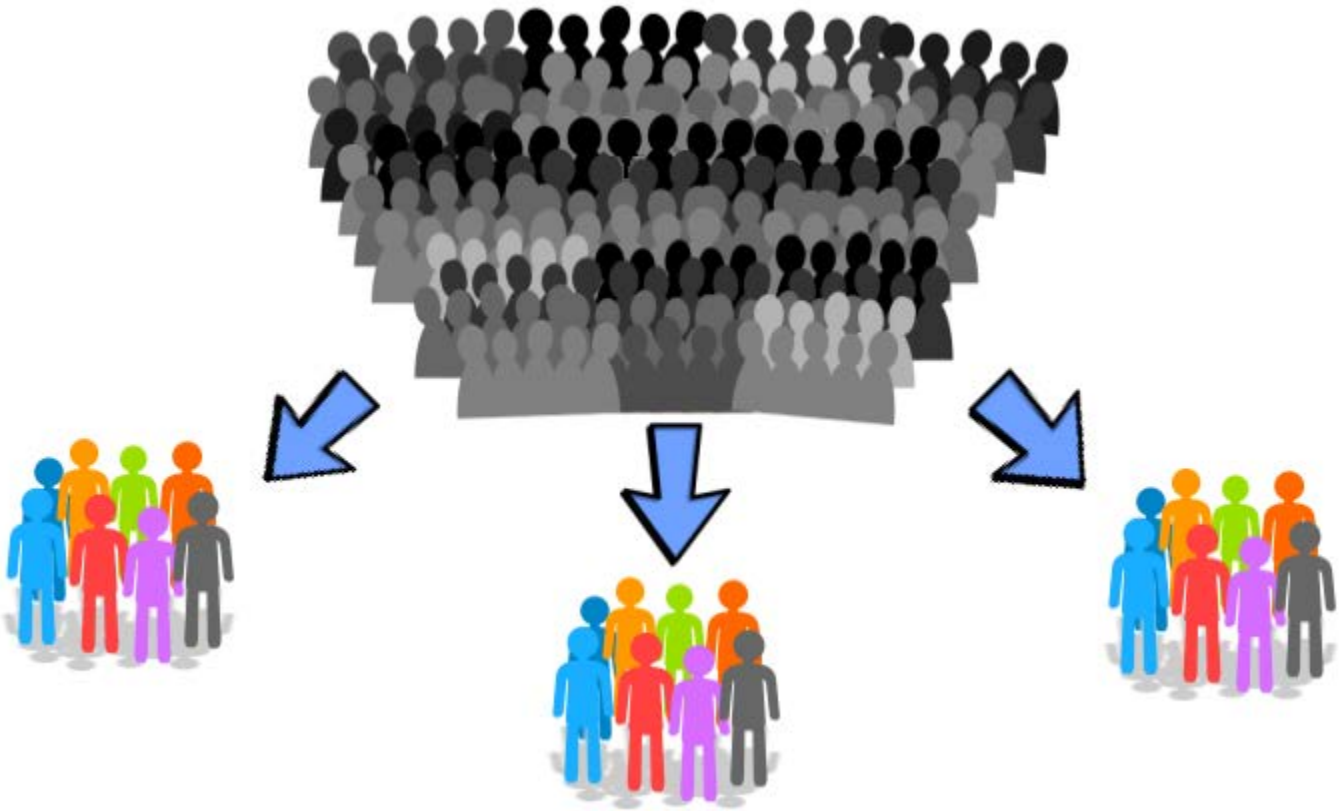
- The examples presented to computers are a set of inputs without any outputs, the goal is to "learn" an intrinsic structure of the examples, e.g., clusters of examples, density of the examples

Unlabeled training data

Inputs: Face images

Groups of similar faces

Unsupervised Learning – Clustering



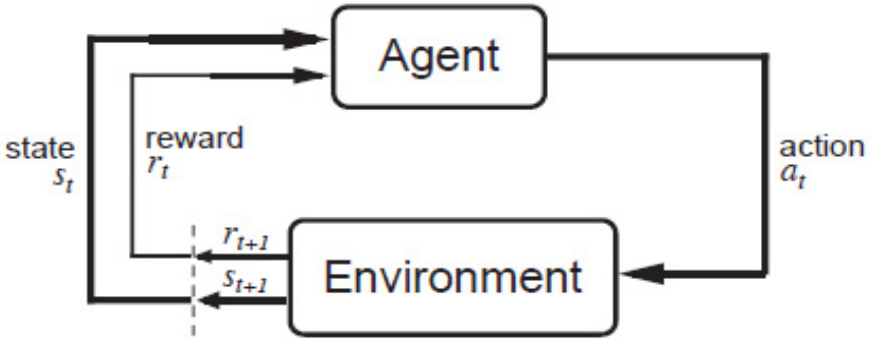
User Segmentation

Different Paradigms/Settings

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Advanced paradigms:
  - Semi-supervised learning
  - Active learning
  - Transfer learning

Reinforcement Learning

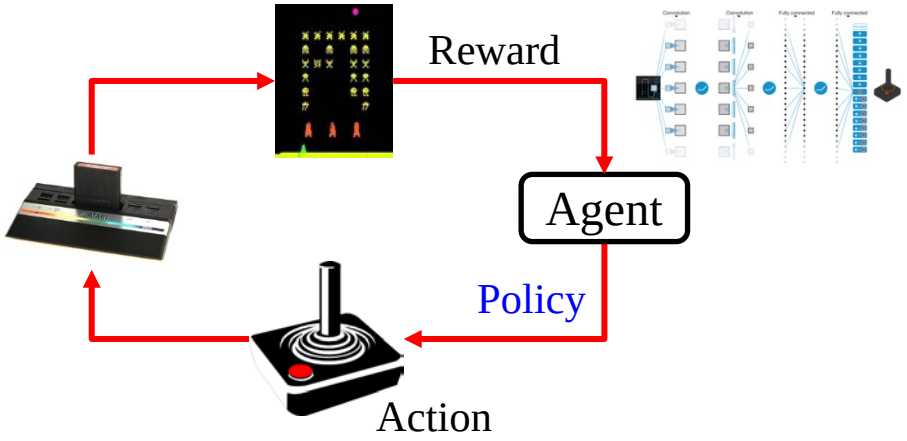
- Learning by interacting with an environment to achieve a goal
- Objective: to learn an optimal policy mapping states to actions

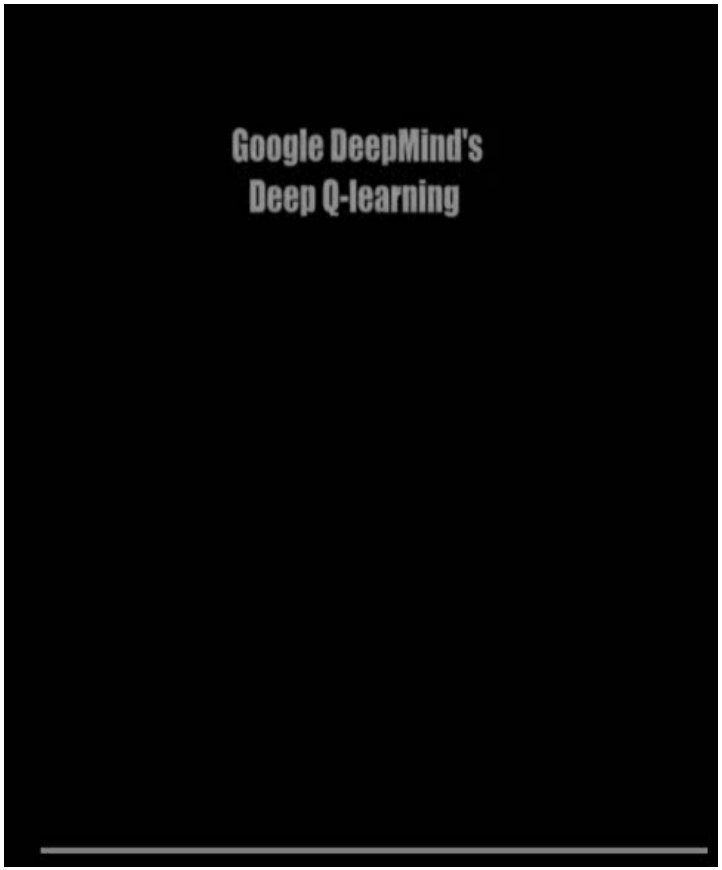


$$s_0 \xrightarrow[r_0]{a_0} s_1 \xrightarrow[r_1]{a_1} s_2 \xrightarrow[r_2]{a_2} s_3$$

Reinforcement Learning (cont.)

- Deep Q-Network (DQN) [1]
  - Play Atari 2600 Games





Google DeepMind

[1] Mnih et al, Human-level control through deep reinforcement learning. *Nature*, 2015

Different Paradigms/Settings

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Advanced paradigms:
  - Semi-supervised learning
  - Active learning
  - Transfer learning

Limitation of Supervised Learning

- Require sufficient labeled data to train a precise model (i.e., a model with good prediction performance)
  - Sufficiency of labeled data is context-aware, depending on different kinds of applications and specific datasets
- When there is insufficient labeled data, can we still train a precise model?
  - Advanced machine learning paradigms

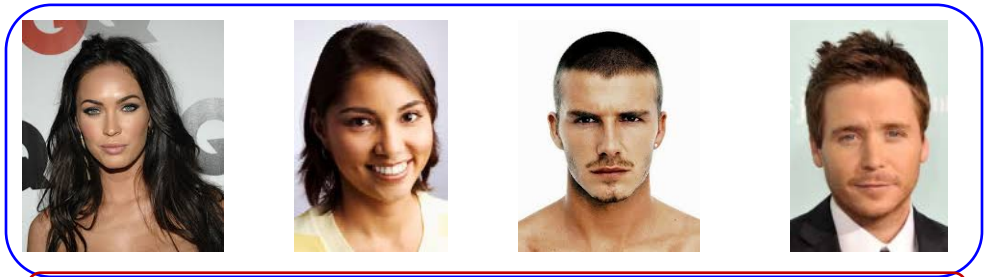
Different Paradigms/Settings

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Advanced paradigms:
  - Semi-supervised learning
  - Active learning
  - Transfer learning

Semi-supervised Learning

- The examples presented to computers include both labeled data and unlabeled data, the goal is to utilize unlabeled data to help supervised learning

Labeled training data



Female Female Male Male

Unlabeled training data

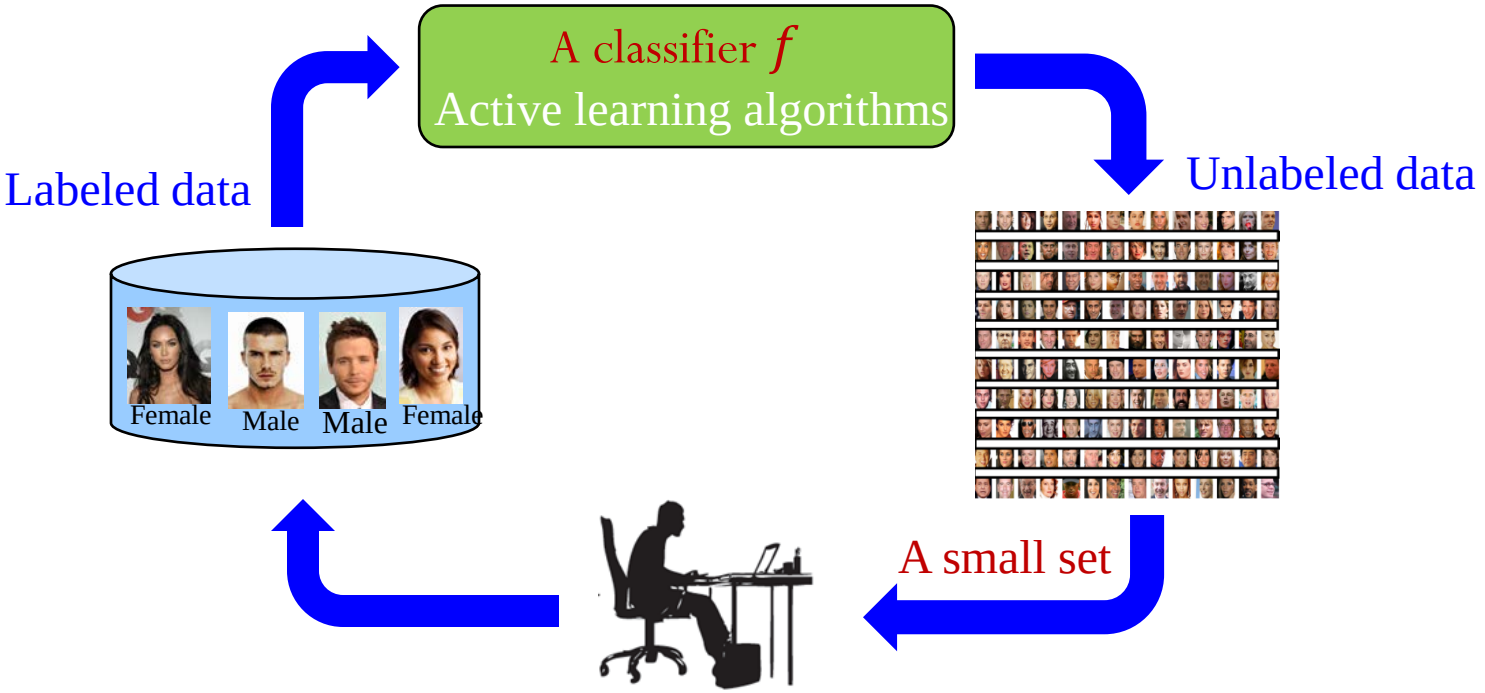
A more precise f

Different Paradigms/Settings

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Advanced paradigms:
  - Semi-supervised learning
  - Active learning
  - Transfer learning

Active Learning

- The examples presented to computers are a small set of labeled data and a pool of unlabeled data. An active learner (computer) can selectively choose some unlabeled data to inquire their ground-truth labels from an oracle (e.g., a human annotator) with some **cost**



Different Paradigms/Settings

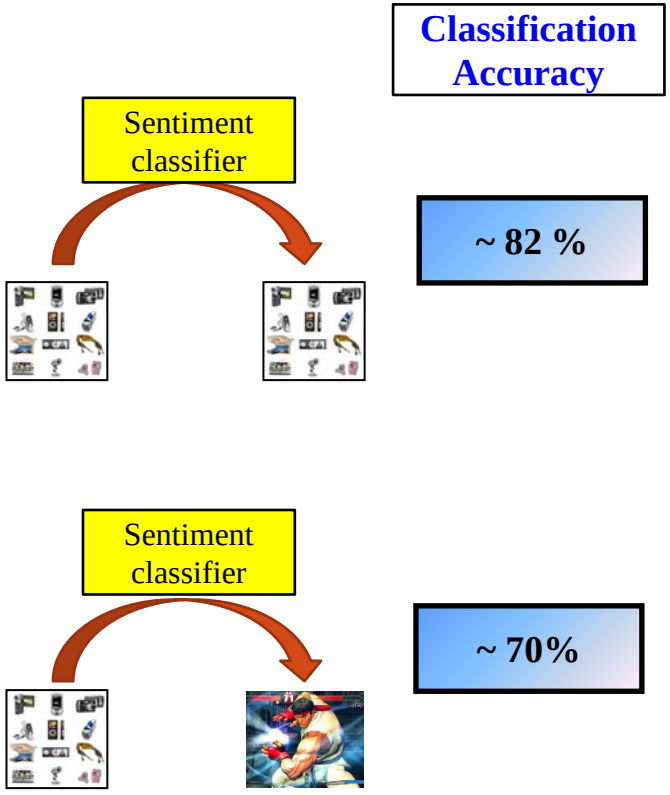
- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Advanced paradigms:
  - Semi-supervised learning
  - Active learning
  - Transfer learning

Motivating Example I:

Sentiment Analysis

Electronics	Video Games
(1) **Compact**; easy to operate; very good picture quality; looks **sharp**!	(2) A very good game! It is action packed and full of excitement. I am very much **hooked** on this game.
(3) I purchased this unit from Circuit City and I was very excited about the quality of the picture. It is really nice and **sharp**.	(4) Very **realistic** shooting action and good plots. We played this and were **hooked**.
(5) It is also quite **blurry** in very dark settings. I will never buy HP again.	(6) The game is so **boring**. I am extremely unhappy and will probably never buy UbiSoft again.

Product reviews on different domains



Motivating Example II:

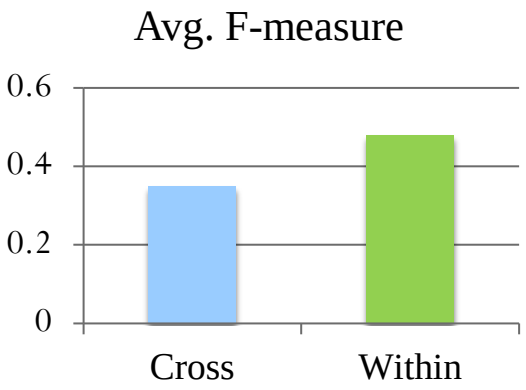
Defect Prediction

For a particular project:

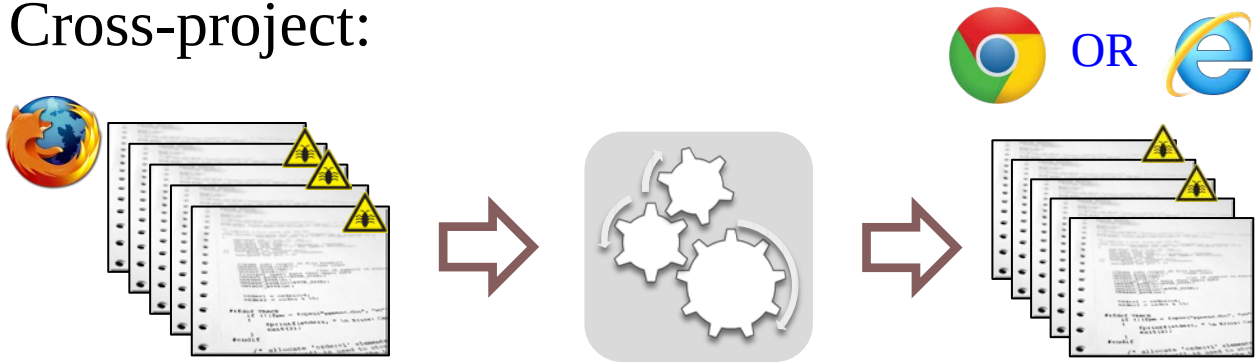
Program with defect information

Predictive Model

Future defects



Cross-project:

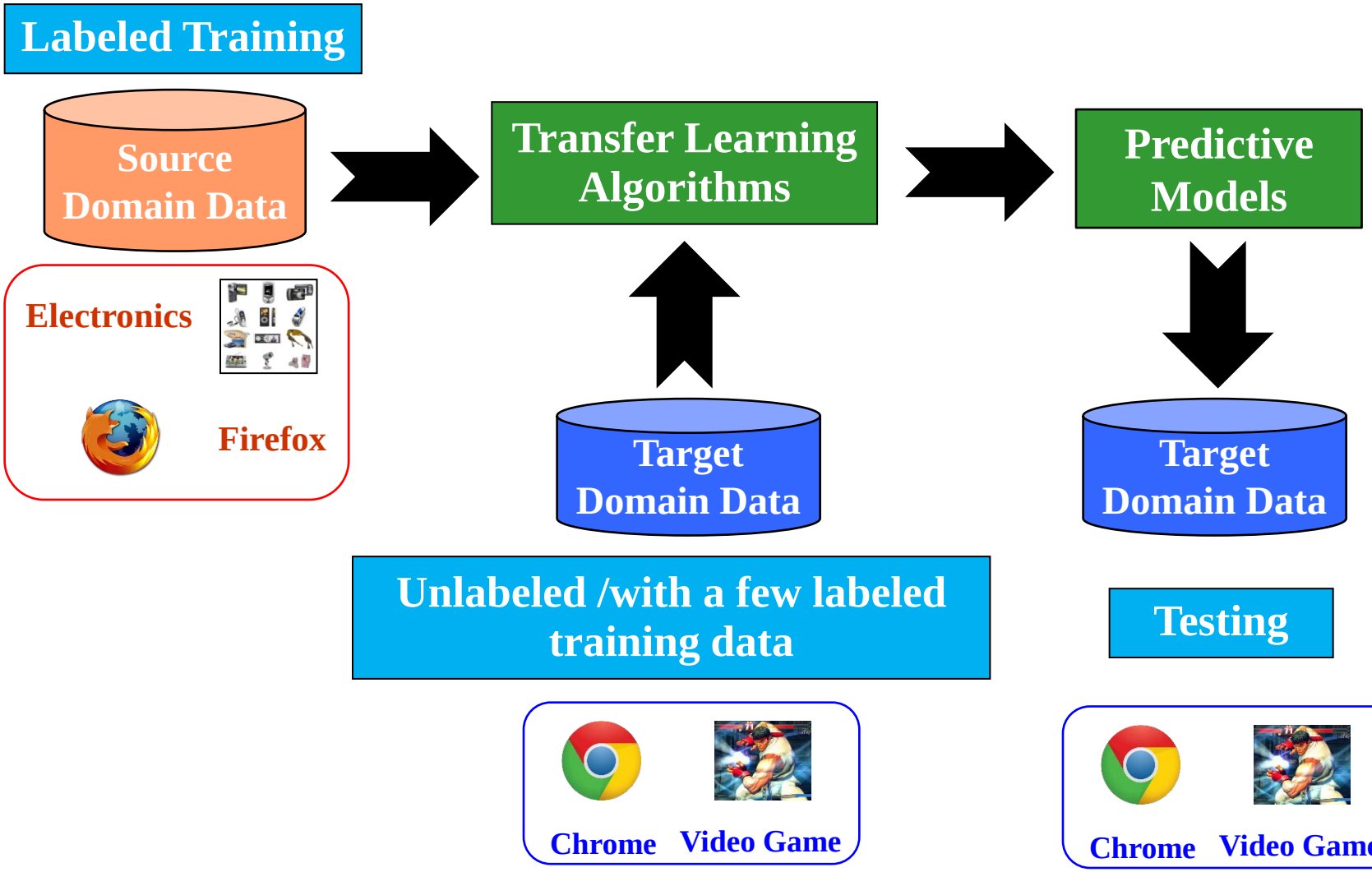


Program with defect information

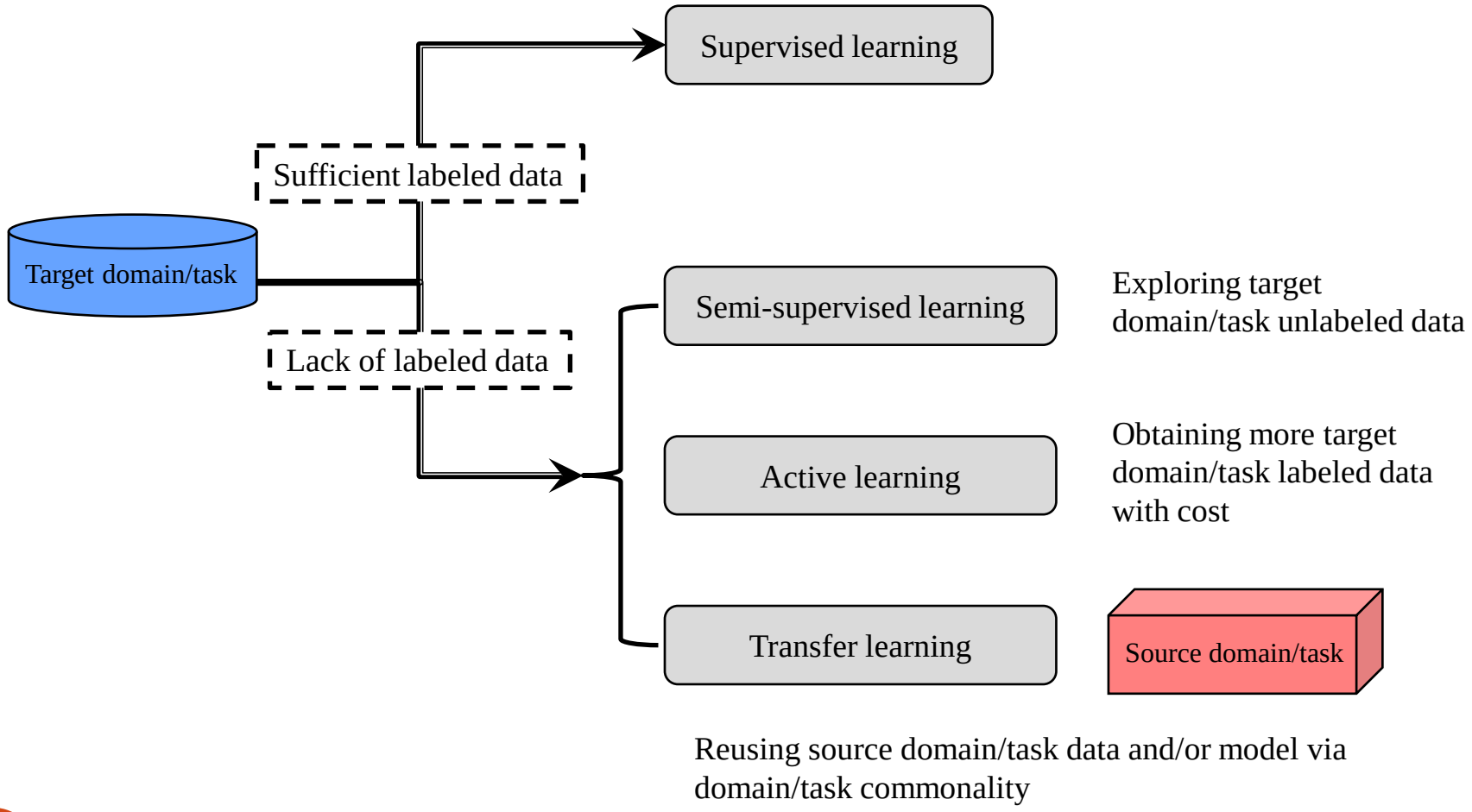
Predictive Model

Program in another Project

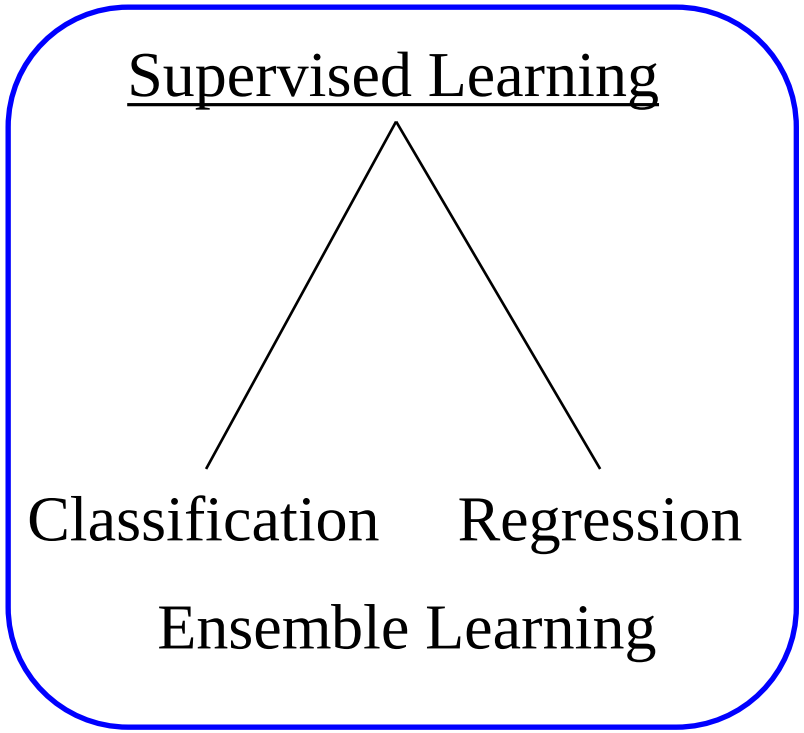
The Goal of Transfer Learning

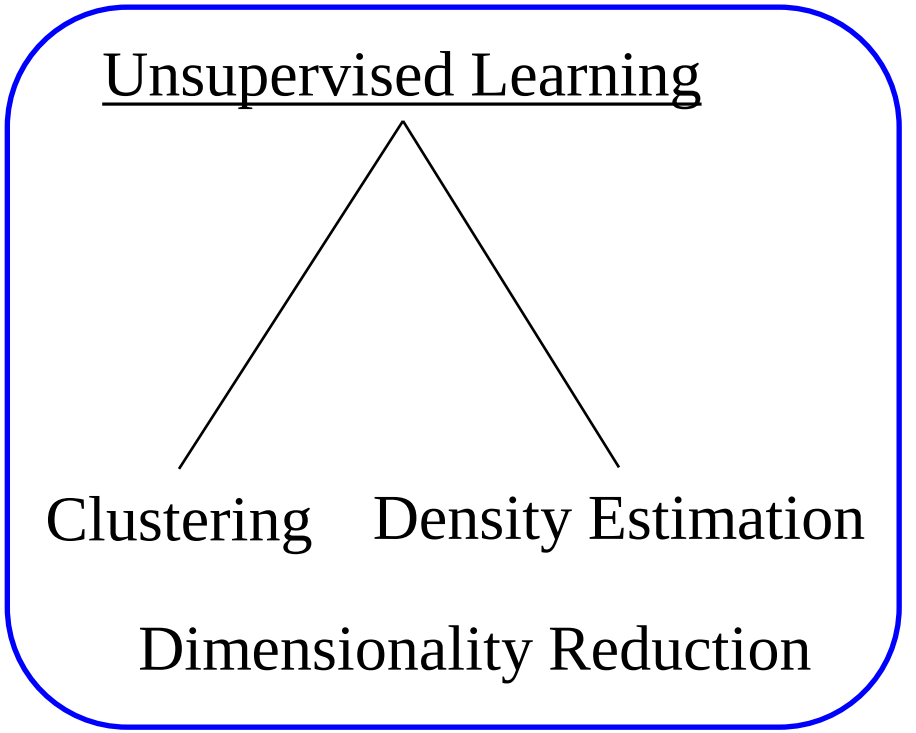


Relationships between Supervised Learning and Advanced Paradigms



Course Scope



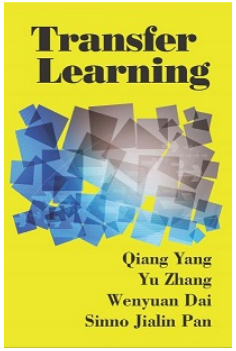


1st-half Course Schedule (Tentative)

Date		Topics
Week 1	10th Aug.	Introduction and Overview of Supervised Learning,
Week 2	17th Aug.	Bayesian Classifiers & Bayesian Decision Theory
Week 3	24th Aug.	Naïve Bayes Classifier
Week 4	31st Aug.	Bayesian Brief Networks
Week 5	7th Sept.	Decision Trees
Week 6	14th Sept.	Generalization & Nearest-Neighbor Classifier

Advanced Readings

- Reinforcement Learning:
  - *Reinforcement learning: a survey*
- Semi-supervised Learning:
  - *Semi-supervised learning literature survey*
- Active Learning:
  - *Active learning literature survey*
- Transfer Learning:
  - *A survey on transfer learning*
  - *Transfer Learning*,
    by Cambridge University Press



Course Objective

- To provide students with essential concepts and principles of machine learning algorithms
- To enable students to understand how to revise or design (beyond how to use) various machine learning algorithms to solve supervised learning and unsupervised learning problems

Breadth and Depth

- Through lectures:
  - Supervised learning techniques
    - Classic classification and regression algorithms, ensemble learning methods
  - Unsupervised learning techniques
    - Classic clustering, density estimation and dimensionality reduction algorithms
- Self-learning through doing a course project:
  - Real-world Applications or Research Topics

Breadth and Depth (cont.)

- Focus on introducing well-known concepts and fundamental methodologies of machine learning
  - Motivations
  - Induction of the mathematical models (mathematics)
- For those who want to learn more, some up-to-date techniques and advanced issues will be mentioned
  - Details cannot be covered in lecture, some additional materials for reading will be suggested (*optional*)

Relationships to Other Modules

CZ4041/CE4041: Machine Learning

Modern AI approaches:

- Classification:
  - Bayesian Decision Theory
  - Bayesian Classifiers (Naïve Bayes & Bayesian Networks)
  - Decision Trees
  - Artificial Neural Networks
  - Support Vector Machines & Kernelization
  - Nearest-Neighbor Classifier
- Regression:
  - Linear Regression & Kernelization
- Clustering:
  - K-means and its variants
  - Hierarchical clustering
- Density Estimation
- Ensemble Learning
- Dimensionality Reduction

Objective: Deeply understand principles

- CZ3005: Artificial Intelligence

Classic AI approaches:

- Search
- First Order Logic

Reinforcement learning

- CZ4042/CE4042: Neural Networks and Deep Learning

Various Architectures of Neural Networks

- CZ4032/CE4032: Data Analytics and Mining

Objective: Understand how to use

Mathematics Background

Various machine learning applications:
Face recognition, object recognition, text mining, activity recognition, stock price prediction, etc.

Various learning paradigms:
supervised learning, unsupervised learning, reinforcement learning, other advanced learning.

Various types of methodologies:
graphical models, deep learning, empirical risk minimization, entropy-based models, kernel methods, etc.

Various mathematical techniques:
Probability theory, linear algebra, calculus, optimization, information theory, functional analysis, etc.

NTU Confessions

January 23, 2017 ·

"There are a lot of year 4 CS modules that require a very solid math foundation to the extent that I think if math majors try taking them, most of them will score better than actual CS students themselves. I believe NTU math graduates will also perform better if they are to take CS graduate courses than actual NTU CS graduates too. This is because we're not exposed to linear algebra / statistics / calculus / number theory / functional analysis / optimization as deeply, if at all. We mostly are only taught about coding and how to software project management in year 2-3. The only math we do in year 1 is way too basic. I dont see how most of us have the foundation necessary to learn more advanced topics in CS and survive pursuing Masters / PhD in many interesting specializations in CS. It's like we are limited to only those areas that require little to no math at all despite us having an actual bachelors degree in CS.

Then again, most CS majors dont care about more specialized topics in CS and have no interest in pursuing further education in CS, because most of us are qualified to become software engineers once we receive our bachelors degree already which allow us to earn quite a lot already. But I think this issue shouldnt be neglected. We need more math in our CS course, whether you like it or not."

#NTUConfessions20807

Textbook and Reference

- Textbook:
  - **Introduction to Machine Learning (2nd Ed.)**, by Ethem Alpaydin, The MIT Press, 2010.
- Reference:
  - **Pattern Classification (2nd Ed.)**, by Richard Duda, Peter Hart, and David Stork, Wiley-Interscience, 2000.
  - **Introduction to Data Mining**, by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Addison Wesley, 2005.
  - **Pattern Recognition and Machine Learning**, by Christopher M. Bishop, Springer, 2006.
- Regarding Mathematics:
  - Part I of the MIT Press book "*Deep Learning*" http://www.deeplearningbook.org/

Useful Resources: Datasets

- UCI Repository:
  - http://www.ics.uci.edu/~mlearn/MLRepository.html
- Kaggle:
  - http://www.kaggle.com/

Useful Resources: Libraries

- scikit-learn (Python) – recommended:
  - http://scikit-learn.org/stable/
- MALLET (Java)
  - http://mallet.cs.umass.edu/
- Weka (Java)
  - http://www.cs.waikato.ac.nz/ml/weka/
- Tensorflow:
  - https://www.tensorflow.org/
- Pytorch:
  - https://pytorch.org/
- Many other libraries on deep learning
  - http://deeplearning.net/software_links/

scikit-learn: machine learning in

scikit-learn.org/stable/

Install User Guide API Examples More

Go

scikit-learn

Machine Learning in Python

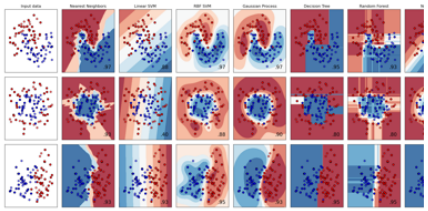
Getting Started | Release Highlights for 0.24 | GitHub

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.
Algorithms: SVM, nearest neighbors, random forest, and more...

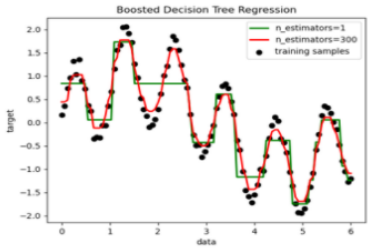


Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, nearest neighbors, random forest, and more...

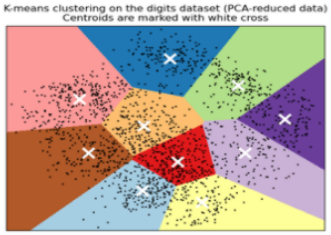


Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, and more...



Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: k-Means, feature selection, non-negative matrix factorization, and more...

Model selection

Comparing, validating and choosing parameters and models.

Applications: Improved accuracy via parameter tuning
Algorithms: grid search, cross validation, metrics...

Preprocessing

Feature extraction and normalization.

Applications: Transforming input data such as text for use with machine learning algorithms.
Algorithms: preprocessing, feature extraction, and more...

Useful Resources: Conferences

- International Conference on Machine Learning (ICML)
- Neural Information Processing Systems (NIPS)
- Conference on Learning Theory (COLT)
- Uncertainty in Artificial Intelligence (UAI)
- International Conference on AI & Statistics (AISTATS)
- International Joint Conference on Artificial Intelligence (IJCAI)
- AAAI Conference on Artificial Intelligence (AAAI)
- International Conference on Learning Representations (ICLR)

Useful Resources: Journals

- Journal of Machine Learning Research (JMLR)
- Machine Learning (MLJ)
- IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)
- IEEE Transactions on Neural Networks and Learning Systems (TNNLS)
- Artificial Intelligence (AIJ)
- Journal of Artificial Intelligence Research (JAIR)

Thank you!