

BT5153

Applied Machine Learning for Business Analytics

Zhao Rui
diszr@nus.edu.sg

Agenda

1. Course Overview
2. What is Machine Learning?
3. Machine Learning is everywhere
4. Overview of machine learning
5. Group Project

What BT5153 covers

Goals of this Course

Learn and improve upon the applications of machine learning

- Understand conceptually the mechanism of machine learning algorithms
- Implement the whole data science pipeline
- Select appropriate machine learning tools/techniques for business applications.

Course Background and Overview

- After DSC5106 Foundation in Data Analytics I
- Together with BT5151 Foundation in Data Analytics II
- Basic Machine Learning/Data Mining models have been covered in these two above modules
- In BT5153:
 - “**Advanced**” topics
 - **Hands-on** Experiences
 - In each lecture, roughly 75% Slides and 25% IPython notebooks.
 - More **Practical** Assignments/Exams

Models

- Representation Learning
 - Autoencoder
 - Word Embeddings
- Bayesian Learning
 - Naive Bayes
- Deep Learning
 - Neural Networks
 - Convolutional Neural Network
 - Recurrent Neural Network
 - Variational Autoencoder
 - Generative Adversarial Networks
- Model Ensembles

Applications

- Spam Detection
- Document Classification
- Recommendation
- Machine Translation
- Image Categorization
- Sentiment Analysis
- Image/Text Generation
- Portfolio Optimization
- Name Entity Recognition
- Part-of-Speech Tagging
- Etc

Hands-on Experience

- **Understanding domain, prior knowledge**
- **Date integration, selection, clearing, pre-processing, etc**
- Learning models (little math, more intuitive ideas)
- Compare models
- **Model Interpretability**
- Consolidating and deploying discovered knowledge
- **Apply discovered knowledge to practical problems**
- Python programming is not the teaching focus.

Course Assessment

- Assignments:
 - One week individual assignment (10%):
 - A two-week mini kaggle competition (10%)
- Midterm Exam (30%)
- Group Project (50%)
 - Project Proposal (10%)
 - Final Presentation (20%)
 - Final Report (20%)

Course Schedule

ML Practices

Date	Topic	Content	Assignment
Fri 01/17	Introduction to Machine Learning	LINK	N.A.
Fri 01/24	Machine Learning Practice	LINK	N.A.
Fri 01/31	Explainability-Accuracy Tradeoff	LINK	Form your team
Fri 02/07	Bayesian Learning: Navie Bayes	LINK	N.A.

Two guest speakers from Alibaba Damo Academy and DBS FX Quant Team will be invited.

Fri 02/14	From Logistic Regression to Deep Learning	LINK	Group Project Proposal Due
Fri 02/21	Representation Learning: Autoencoder	LINK	Individual Assignment Due
Fri 02/28	Recess Week	N.A.	N.A.
Fri 03/06	Representation learning: Word2Vec	LINK	Kaggle Start
Fri 03/13	Convolutional Neural Networks: Why It Works	LINK	In-class Midterm Exam
Fri 03/20	Generative Deep Learning	LINK	Kaggle Due
Fri 03/27	Responsible Machine Learning	LINK	N.A.

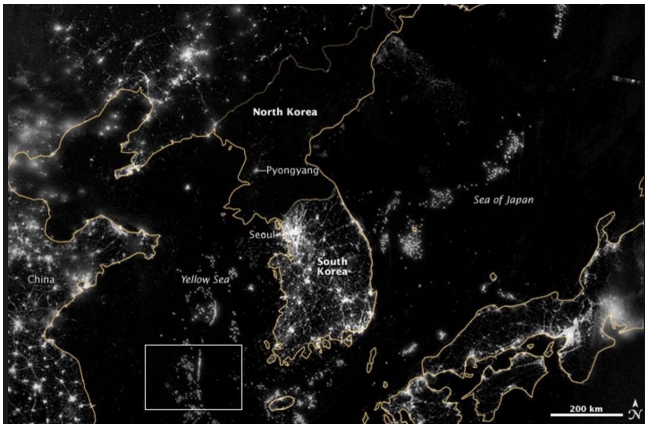
Deep Learning and Representation Learning

Fri 04/03	Group Presentation 1	N.A.	N.A.
Fri 04/10	Good Friday	N.A.	N.A.
Fri 04/17	Group Presentation 2	N.A.	N.A.

<https://nusmsba.github.io/#schedule>



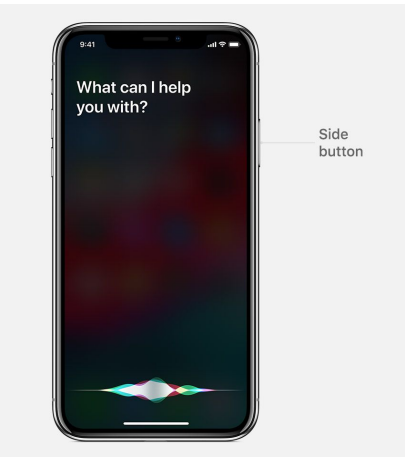
Face Recognition



GDP Prediction



Alphago



Siri



Hedge fund use ML for trading



Special Effects in Tiktok



Amazon Recommendation



Machine Translation



Self-driving Car

What is Machine Learning



Mat Velloso

@matvelloso

Follow



Difference between machine learning
and AI:

If it is written in Python, it's probably
machine learning

If it is written in PowerPoint, it's
probably AI

5:25 PM - 22 Nov 2018

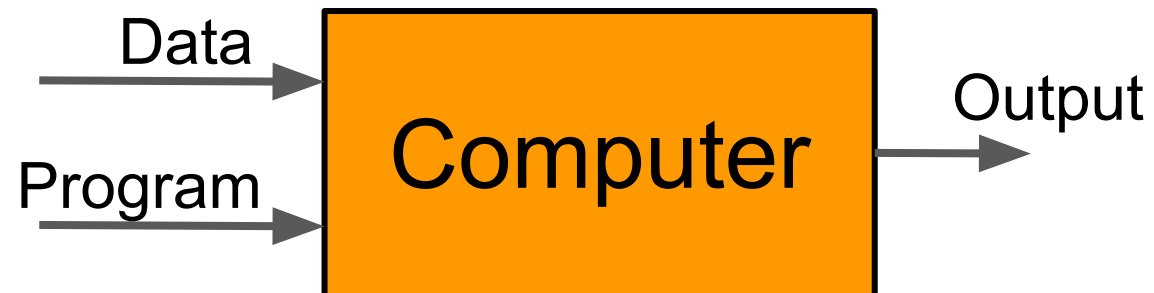
8,541 Retweets 23,778 Likes



Python Programming

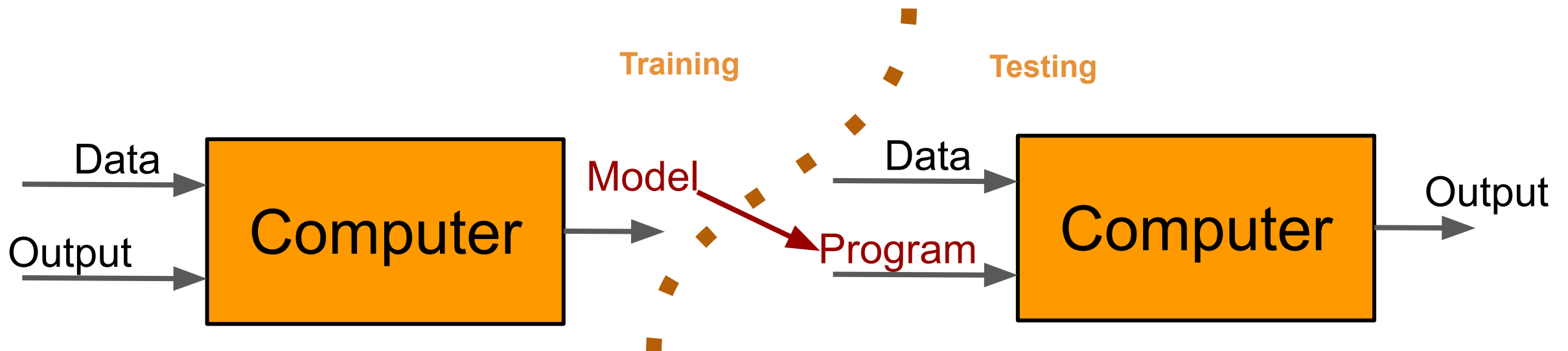
```
In [1]: a = 3  
b = 1  
q = 3*a + 2*b  
print('result is {}'.format(a + b))
```

result is 4



Machine Learning

```
] : from sklearn.neighbors import KNeighborsClassifier
    from sklearn.metrics import accuracy_score
    #create an object of KNN
    neigh = KNeighborsClassifier(n_neighbors=3)
    #train the algorithm on training data and predict using the testing data
    pred = neigh.fit(data_train, target_train).predict(data_test)
```



Definition of Machine Learning

“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**”



Tom Mitchell

T, **P**, **E** are three basic elements to define a complete machine learning tasks

AlphaGo

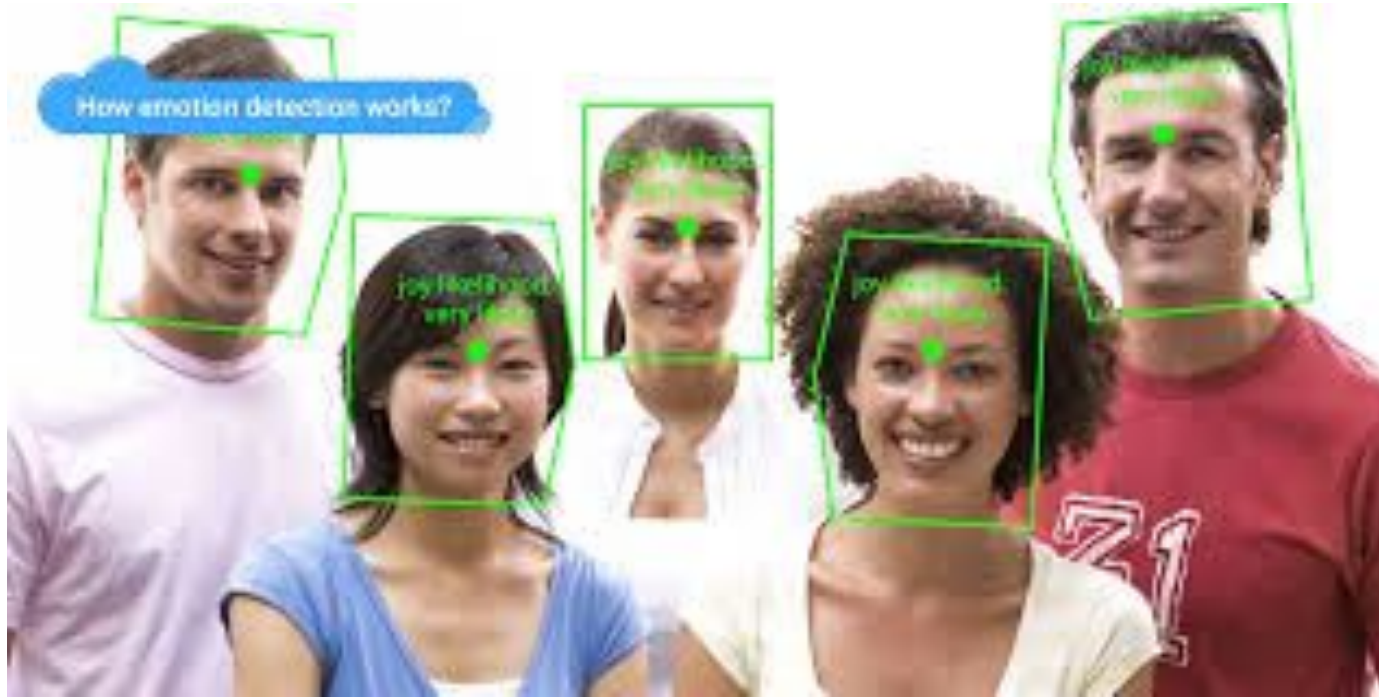


T: Play Go Games

P: Win rates of all matches

E: Match Experiences with many go players or itself

Face Recognition



T: Identify or verify human faces

P: Accuracy that human faces are detected

E: Dataset of labelled human faces

Machine Translation



T: Translate source language into target language

P: Accuracy that human faces are detected

E: Corpus of source-target language pairs

More about E

- For machine learning algorithms, E is **data**.

- Data types:

- Unstructured vs Structured
- Raw vs Processed



Computers Processable and Understandable

The very import step for Machine Learning Project is that how to **preprocess these unstructured/raw data**.

Structured

- Structured: Table (Matrix) or Tensor

Player	Height (inches)	Weight (pounds)	Position
Player 1	76	225	C
Player 2	75	195	PG
Player 3	72	180	SF
Player 4	82	231	PF

Features (points to Height column)

Feature Values (points to 225)

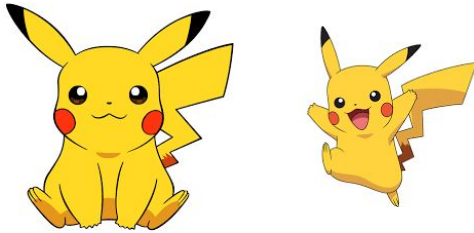
Labels (points to Position column)

Data Sample (points to Player 4 row)

Unstructured

- The original data can not be stored in an “table”
- More abstract, more fuzzy, and more high-dimensionality

Images



Audio



Video

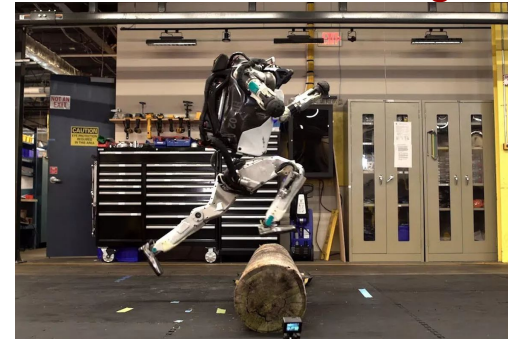


Text

Content

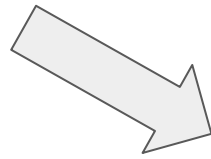
This module provides students a deep overview of various advanced machine learning techniques applied to business analytics tasks. The focus of this course will be the key and intuitive idea behind machine learning models and hands-on examples instead of theoretical analysis. The tentative topics include machine learning pipeline, unsupervised learning, structure learning, Bayesian learning, deep learning and generative models. The programming languages used will be Python.

Environment around agent



Raw Data

- Unstructured Data such as images and text
- Some structured data like categorical data



Labels

Position

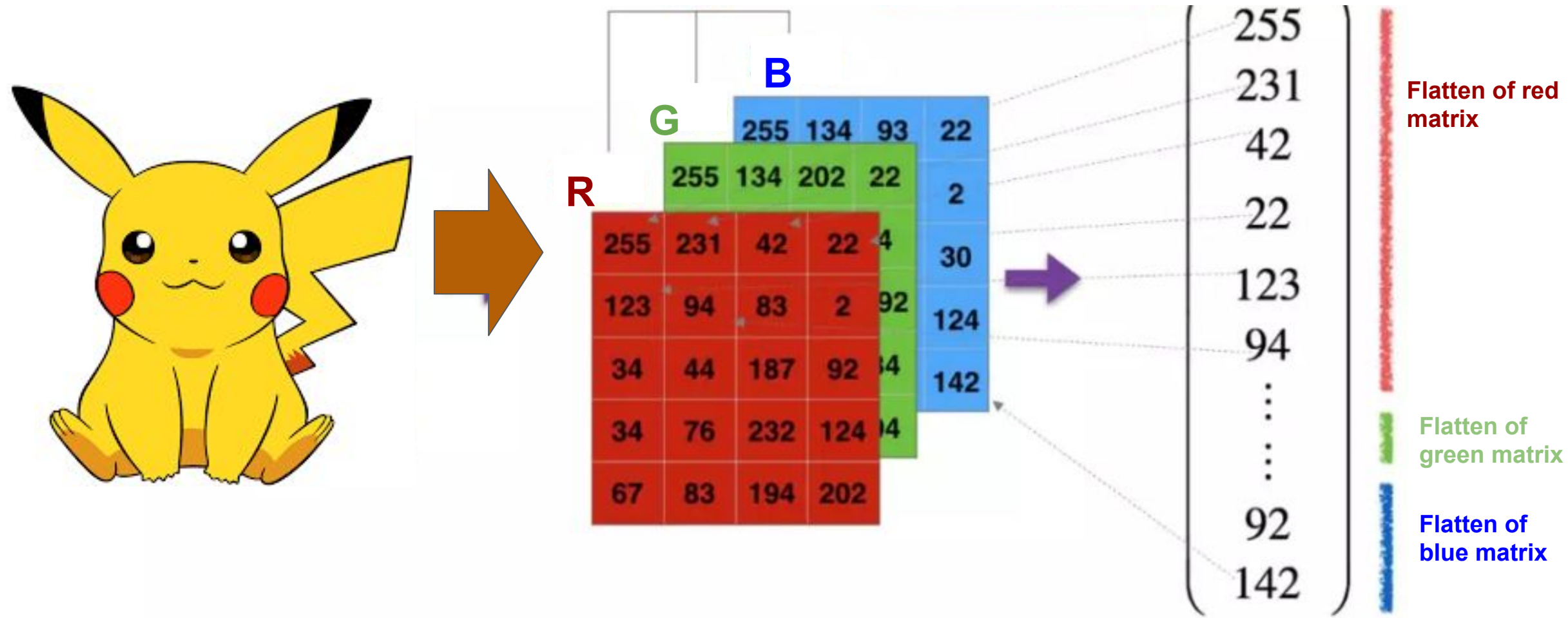
C

PG

SF

PF

Processed Data (from Raw)



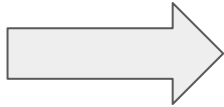
Processed Data (from Raw)

the dog is on the table

0	0	1	1	0	1	1	1
are	cat	dog	is	now	on	table	the

Processed Data (from Raw)

Position	Label
C	0
PG	1
SF	2
PG	1
PF	1



Position	Label
C	1 0 0 0
PG	0 1 0 0
SF	0 0 1 0
PG	0 0 1 0
PF	0 0 0 1



Which one is better?

Terms

- Artificial Intelligence: **Intelligence** exhibited by machines to mimic a human mind
- Machine Learning: Computers being able to learn without hand-coding each step
- Deep Learning: **Multi-layered** algorithms for learning from data
- Data Science: Methods, processes, and systems to extract **insights** from data
- Analytics: Discovery of meaningful patterns in data

What is what

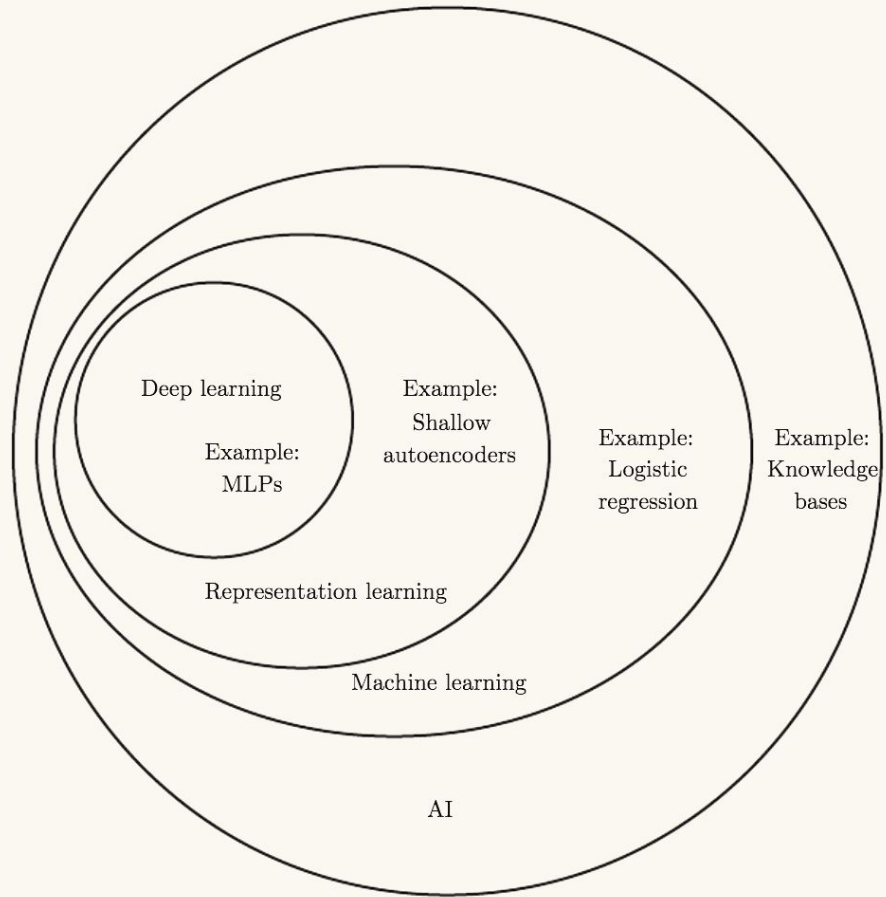


Figure 1.4: A Venn diagram showing how deep learning is a kind of representation learning, which is in turn a kind of machine learning, which is used for many but not all approaches to AI. Each section of the Venn diagram includes an example of an AI technology.

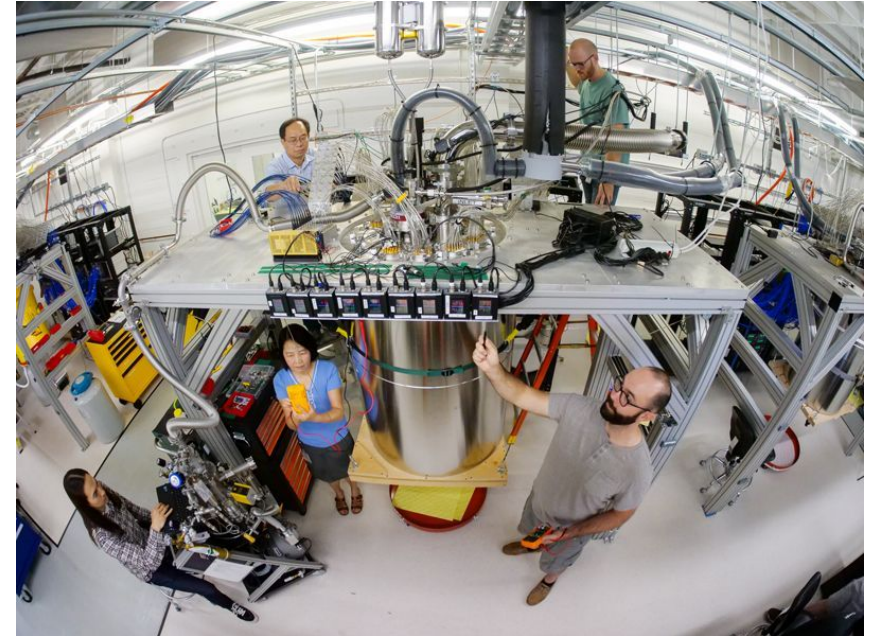
**Machine Learning is
everywhere**

Why Study Machine Learning?

- Reduce Human Efforts:
 - Allow the computers learn automatically
 - Writing rule-based program is too **challenging**
 - Let the data **SPEAK**
- Machine learning brings better career opportunity
 - The trend will be **MLaas** (Machine learning as a service)
 - Machine learning engineer, Data scientist, Data Product Manager, Cloud Engineer, etc

Why Machine Learning is Powerful?

- Recent progress in algorithms and theory
- Big data era
 - Flow of online data and mobile data
 - 5G is developing
 - Cloud computing
- Computational power is available
 - TPU, GPU, **Quantum Computing**,



Google Quantum Computer

Three Niches for Machine Learning

- **Data mining**
 - Use historical data to improve decisions.
- **Software applications that are hard to be programmed by hand**
 - Speech Recognition
 - Autonomous Driving
 - Etc
- **User modeling**
 - Recommendation System
 - Micro-credit Loan
 - Etc

Credit Risk Analysis

Customer103: (time=t0)

Years of credit: 9

Loan balance: \$2,400

Income: \$52k

Own House: Yes

Other delinquent accts: 2

Max billing cycles late: 3

Profitable customer?: ?

Customer103: (time=t1)

Years of credit: 9

Loan balance: \$3,250

Income: ?

Own House: Yes

Other delinquent accts: 2

Max billing cycles late: 4

Profitable customer?: ?

Customer103: (time=tn)

Years of credit: 9

Loan balance: \$4,500

Income: ?

Own House: Yes

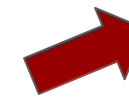
Other delinquent accts: 3

Max billing cycles late: 6

Profitable customer?: No

- Learned Rules

- Other-Delinquent Account > 2 and Number Delinquent billing cycles > 1
- Other-Delinquent Account = 0 and (Income > \$30k or Years-of-Credit > 3)

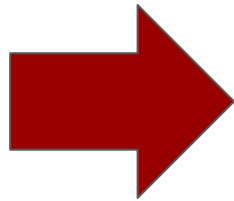
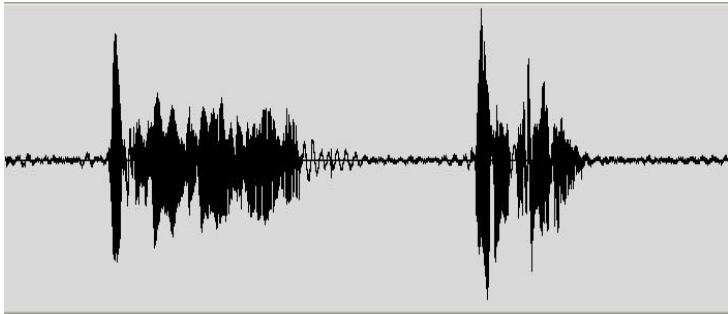


No Profitable



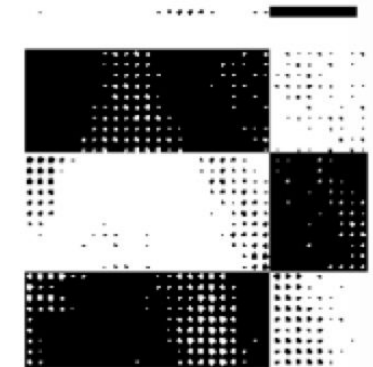
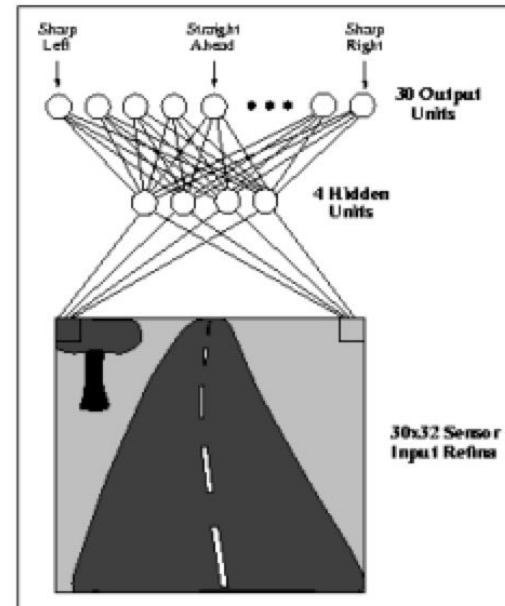
Profitable

Speech Recognition



This is Friday

Self-Driving Car



ALVINN: an autonomous land vehicle in a neural network (1989)

Overview of Machine Learning Concepts

Basic Paradigm

- Define the **T**asks (what should be learned)
- Find the training **E**xperiences(datasets)
- Quantify the **P**erformance via a measure
- Choose a machine learning model to complete the **T** and improve the **P** via the **E**

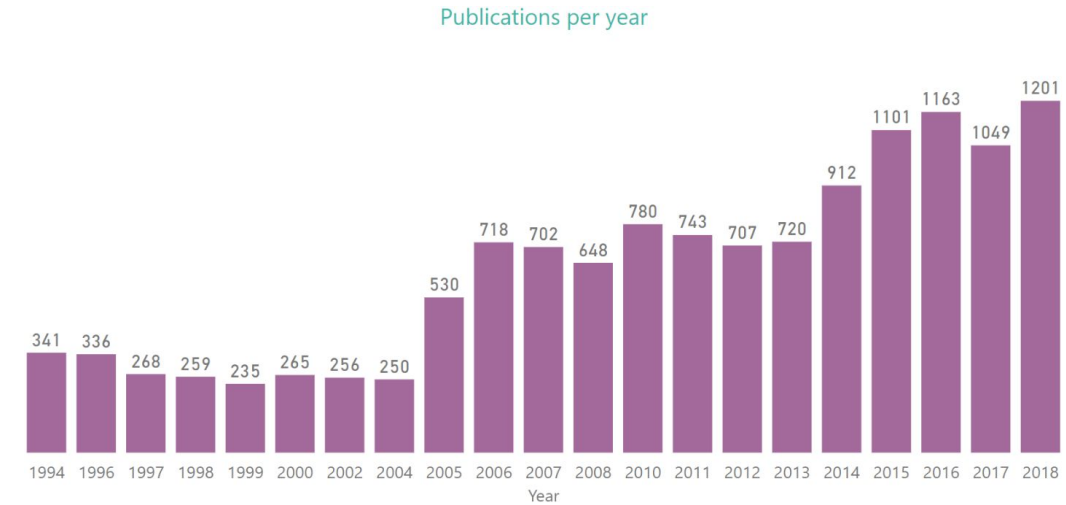
Machine Learning in a nutshell

- Ten of thousands of machine learning algorithms

- Hundreds New per month

- Each ML algorithm can be decomposed into:

- **Representation**
- **Evaluation**
- **Optimization**



<https://www.microsoft.com/en-us/research/project/academic/articles/aaai-conference-analytics/> AAI Conference

Representation

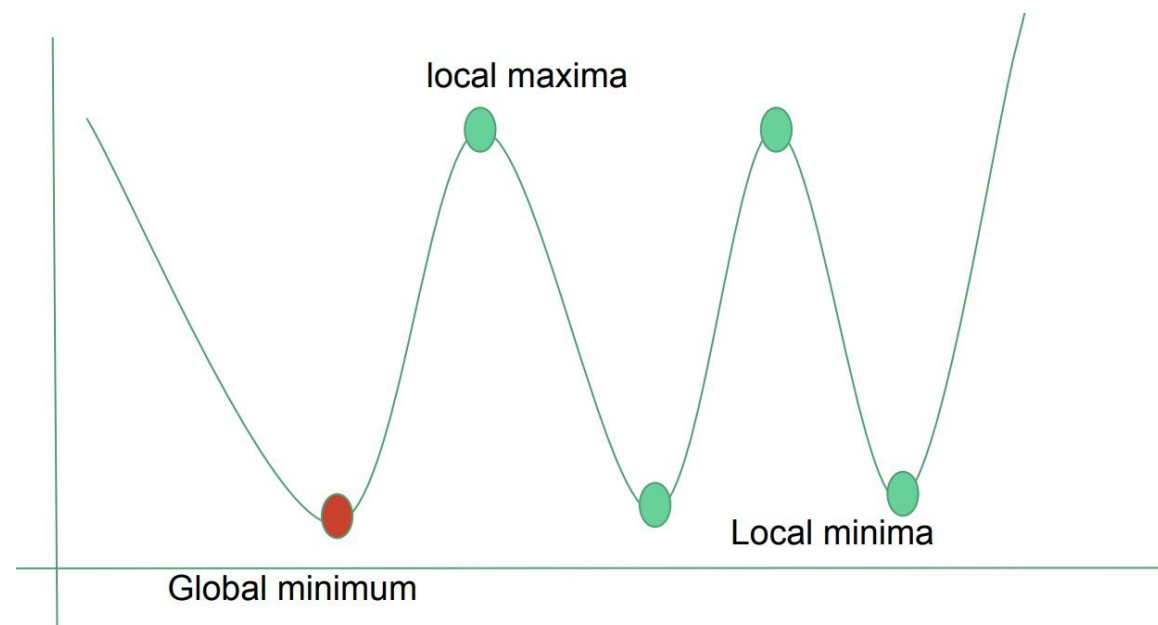
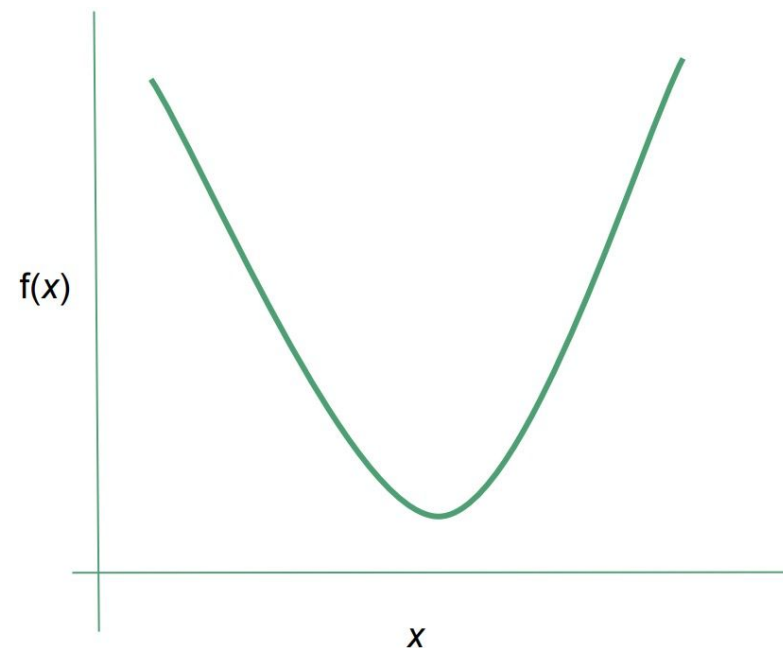
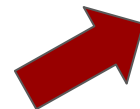
- Decision Trees
- Support Vector Machine
- Set of rules
- Instances-based Learning (K Nearest Neighbor)
- Probabilistic Graph Models (Naive Bayes and Hidden Markov Models)
- Neural Networks and Deep Learning
- Ensemble
- Others

Evaluation

- Confusion matrix
- Accuracy
- Precision and recall
- Mean Squared Error
- Likelihood
- Posterior probability
- Margin
- Entropy
- K-L Divergence
- Etc

Optimization

- Combinatorial Optimization
 - Grid search
- Convex Optimization
 - Least Squares
 - Linear Programming (with constraints)
 - Semidefinite Programming
 - Etc
- Non-convex Optimization
 - Gradient descent algorithm
 - Bayesian Optimization
 - Etc



Types of Machine Learning Models

- Supervised Learning
 - Training data contain the desired outputs (labels)
- Unsupervised Learning
 - Training data do not contain labels
- Reinforcement Learning
 - Rewards from **sequence actions**
- Semi-supervised Learning
 - Training data include a few labels

Supervised Learning

- Given (training data x , training label y), predict (new data, ?)

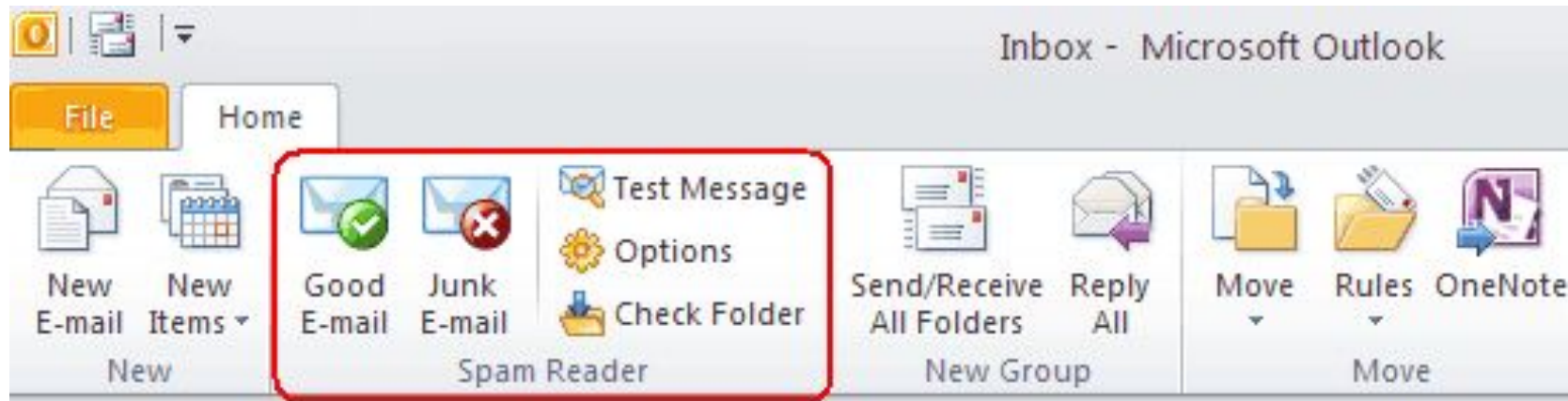
$$h(x) \approx f(x)$$



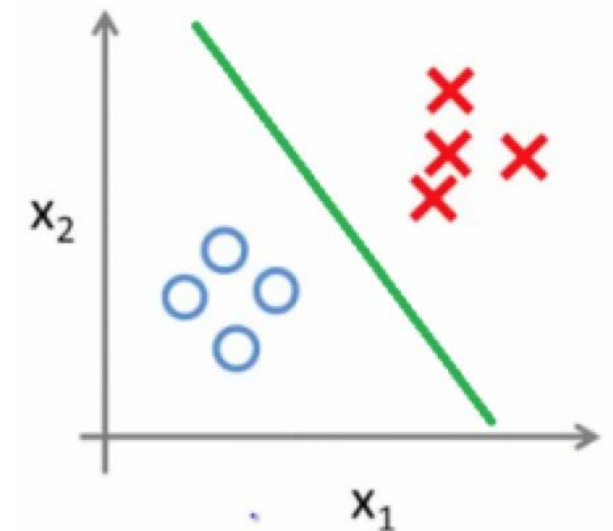
- Classification: **discrete** label
 - Binary Classification: label y in $\{0, 1\}$
 - Multi-class Classification: label y in $\{0, 1, 2, 3, \dots, k-1\}$
- Regression: **continuous** label
 - y is real-valued space

Binary Classification

- Junk Email Filter

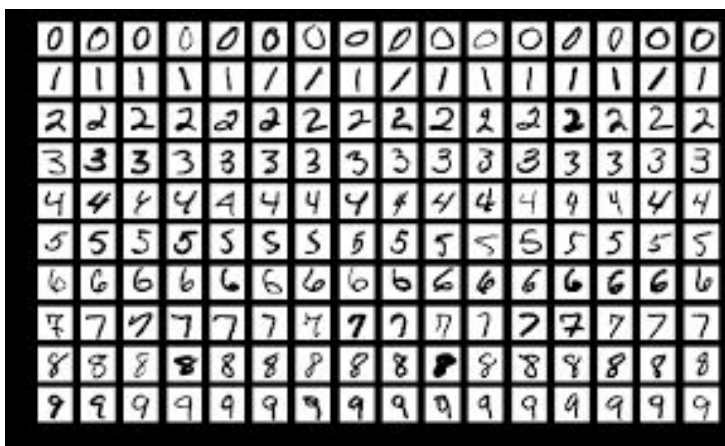


- Two classes
 - 1 Normal emails
 - 0 Spam emails

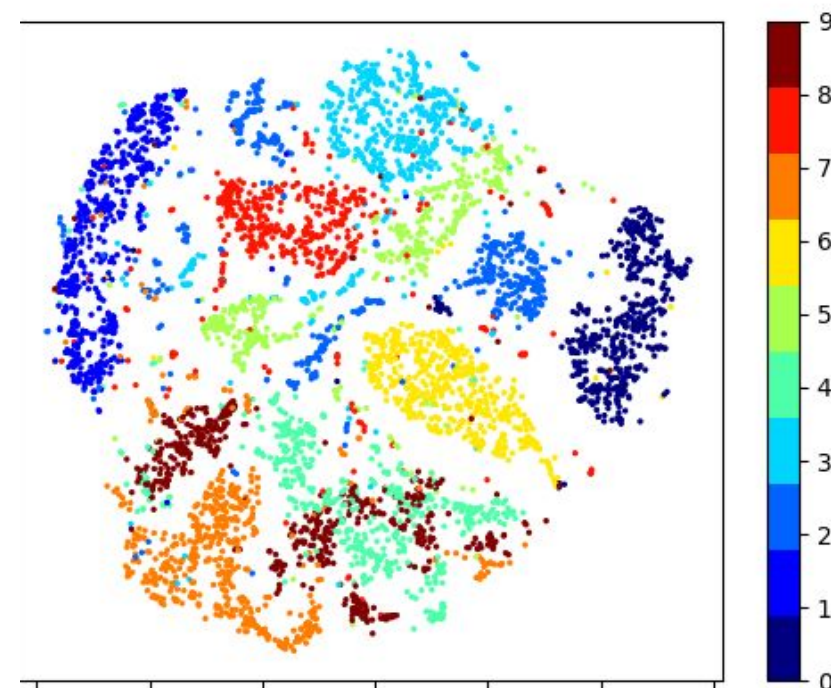


Multi-class Classification

- Handwritten digits recognition

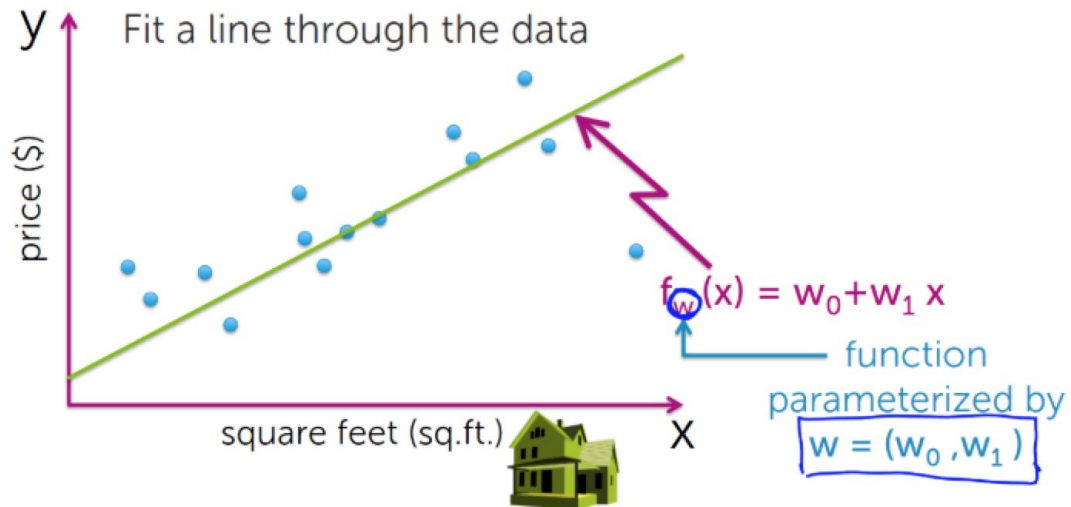


- Ten classes
 - Each number is one class from 0 to 9



Regression

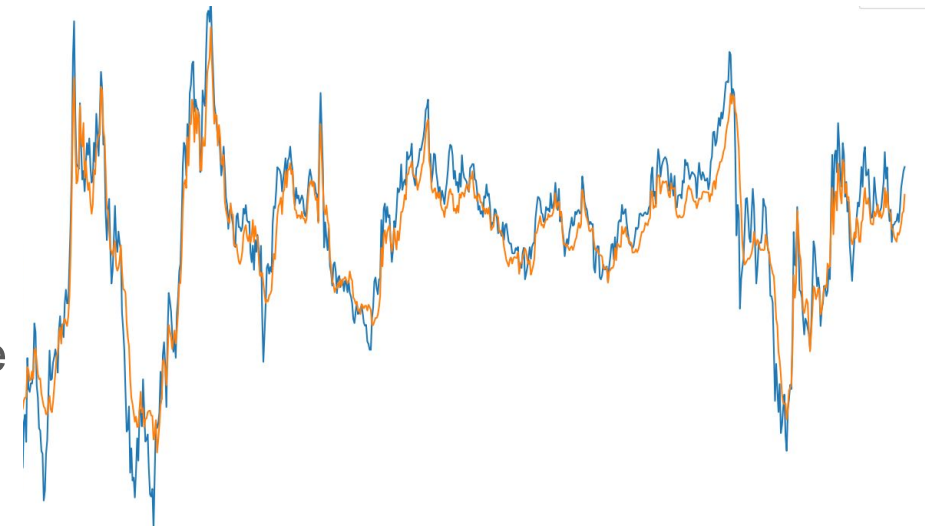
- Linear Regression



- Non-linear Regression

- Time series prediction
- Based on previous k time steps, predict the current value

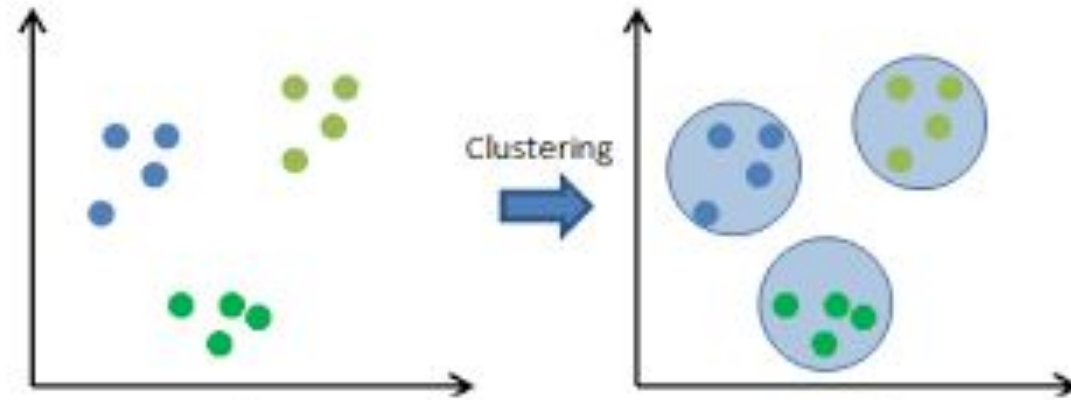
— True Value
— Prediction



Unsupervised Learning

- No training labels
- Clustering
 - Divide data into a number of groups
- Dimension Reduction
 - Find a sub feature space for the data representation
- Novelty/Anomaly detection
 - Find the odd one out
- Etc

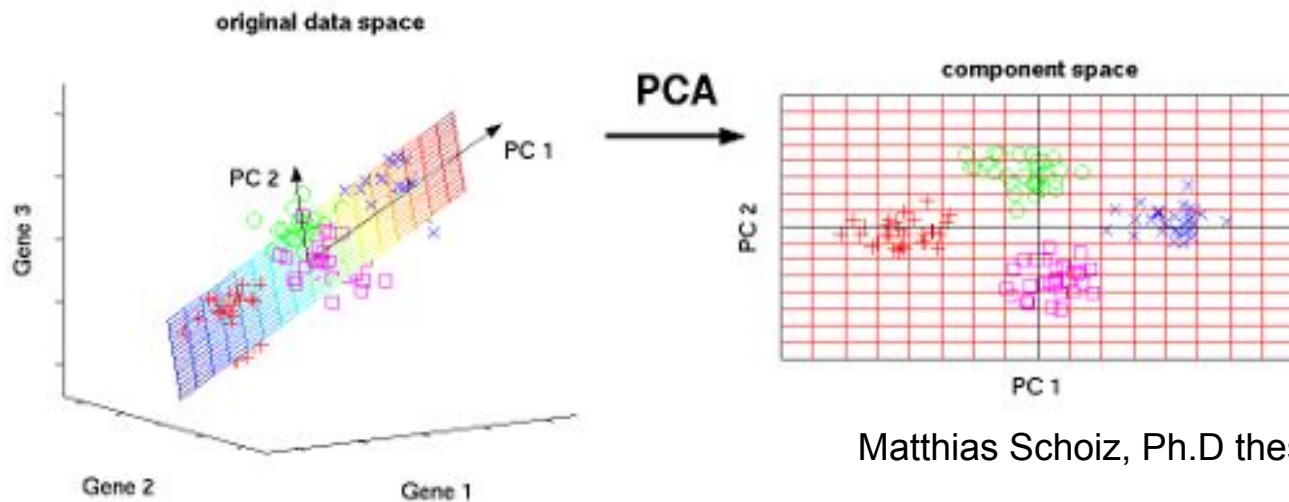
Clustering



- Customer/Marketing segmentation
- Clustering of news, documents, pictures, ..

Dimension Reduction

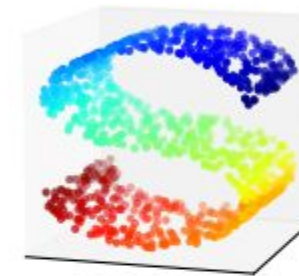
- Principal Component Analysis



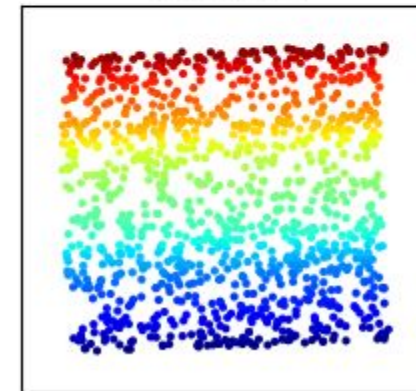
Matthias Schoiz, Ph.D thesis

- Manifold Learning

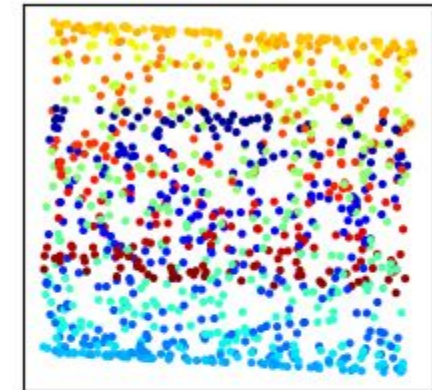
- Find nonlinear subspace/embedding



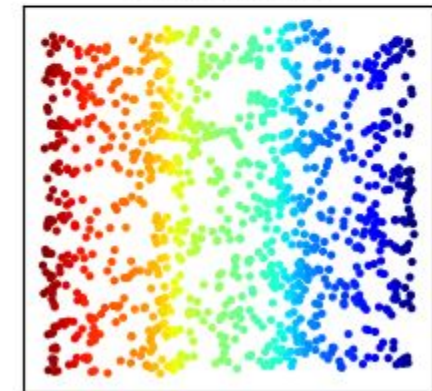
LLE projection



PCA projection



IsoMap projection



Novelty / Anomaly Detection

- Identify new/unknown patterns



Interacting with Environment

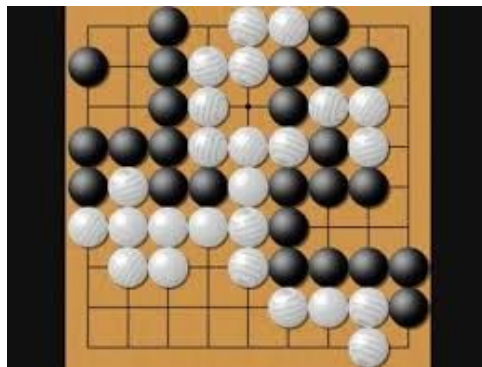
- How to get experience from Environment
- Batch Learning
 - Observe a **batch** of training data $(x_1, y_1), \dots (x_k, y_k)$, then train and used for prediction
- Online Learning
 - **Sequential**: Observe x_1 , predict $f(x_1)$, train with (x_1, y_1) , observe x_2
- Active Learning
 - **interactively** query the user/database to obtain the desired outputs
 - Reduce the amount of labelled data that is needed
- Reinforcement Learning
 - Take action, environment responds, take new actions
 - Play Go, Autonomous Driving

Reinforcement Learning

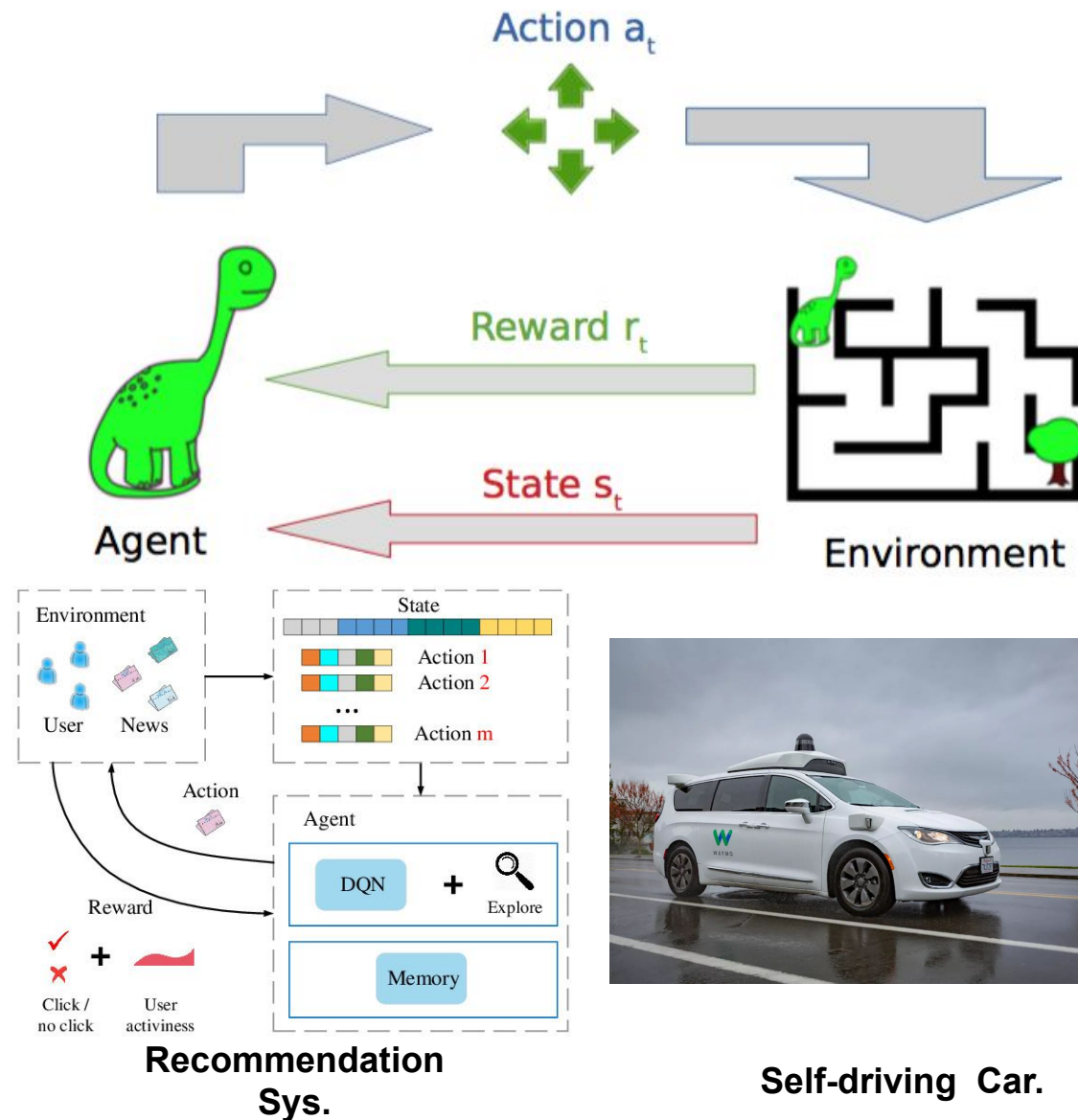
- Do not have known-correct answers
- Repeat
 - Take action
 - Environment reacts
 - Reward/Penalty
 - Update model
- Application



Play Dota2



Play Go



Key Issues in Machine Learning

- Obtaining experience

- How to obtain training data?
 - Supervised or Unsupervised
- How many examples are enough?
 - PAC learning theory

- Learning algorithms

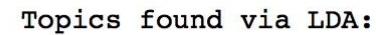
- Which kind of machine learning model can approximate function well, when?
- How does the complexity of learning algorithms impact the learning accuracy?
- Whether the objective function is learnable?

- Representing inputs

- How to represent the inputs?
- How to remove the irrelevant information from the input representation?
- How to reduce the redundancy of the input representation?

Group Projects

1. <https://nusmsba.github.io/project/2019fyp.html>
2. https://nusmsba.github.io/project/pyp_analyze/PDF_Extraction.html
3. https://nusmsba.github.io/project/pyp_analyze/Reprots_Analysis.html



Topic #1:
feature model review text data using

Topic #3:
comment feature model character article editor

Topic #5:
word model data wine tweet question

Project Hint 1

- Find a new problem which can be solved by machine learning technique
 - **Visualize the impact of climate change**

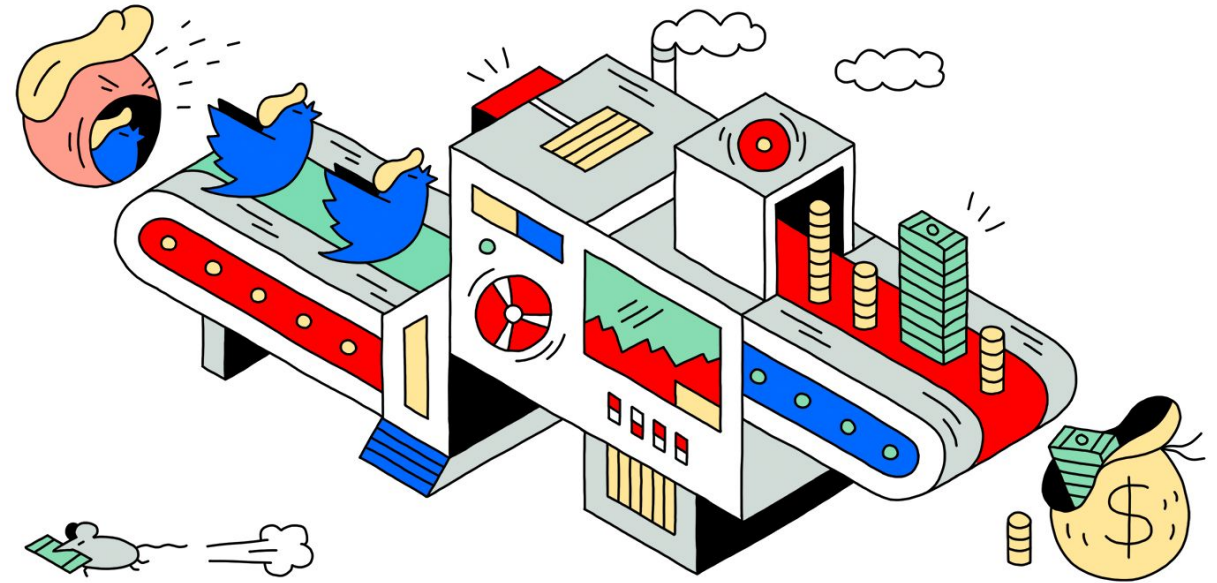
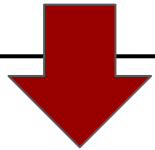


Applying generative models to create personalized images of an extreme climate event, flooding

Project Hint 2

- **Trump2Cash**

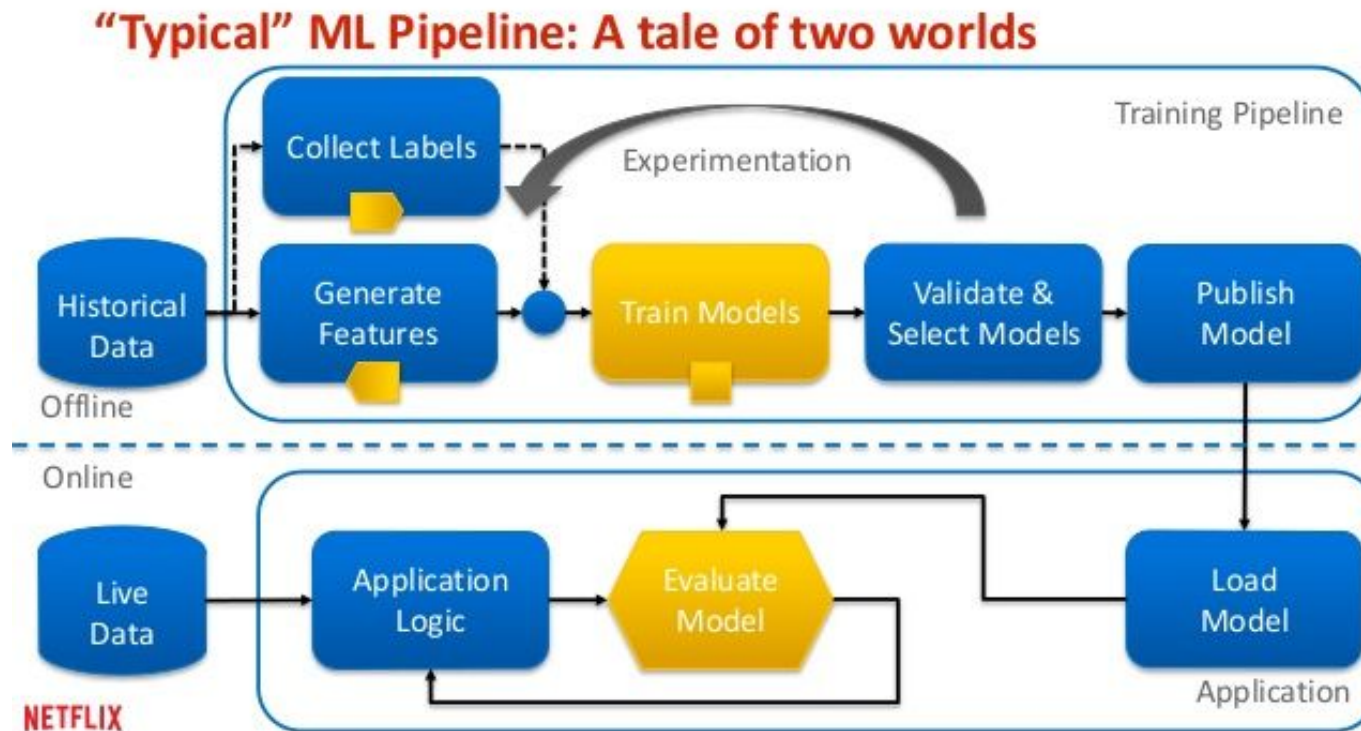
1. Monitor Trump's twitter feed
2. Analyze the twitter
If it mentions of any publicly traded stocks
and compute its sentiment
 - a. Long it if the sentiment is positive
 - b. Short it if the sentiment is negative



Source: <https://github.com/maxbbraun/trump2cash>

Project Hint 3

- Build a whole pipeline ML system (real-life one)



Project Hints 4

- In-depth analysis of machine learning algorithms on one specific application
- **Try to explain the findings**

Model	MR	SST-1	SST-2	Subj	TREC	CR	MPQA
CNN-rand	76.1	45.0	82.7	89.6	91.2	79.8	83.4
CNN-static	81.0	45.5	86.8	93.0	92.8	84.7	89.6
CNN-non-static	81.5	48.0	87.2	93.4	93.6	84.3	89.5
CNN-multichannel	81.1	47.4	88.1	93.2	92.2	85.0	89.4
RAE (Socher et al., 2011)	77.7	43.2	82.4	—	—	—	86.4
MV-RNN (Socher et al., 2012)	79.0	44.4	82.9	—	—	—	—
RNTN (Socher et al., 2013)	—	45.7	85.4	—	—	—	—
DCNN (Kalchbrenner et al., 2014)	—	48.5	86.8	—	93.0	—	—
Paragraph-Vec (Le and Mikolov, 2014)	—	48.7	87.8	—	—	—	—
CCAE (Hermann and Blunsom, 2013)	77.8	—	—	—	—	—	87.2
Sent-Parser (Dong et al., 2014)	79.5	—	—	—	—	—	86.3
NBSVM (Wang and Manning, 2012)	79.4	—	—	93.2	—	81.8	86.3
MNB (Wang and Manning, 2012)	79.0	—	—	93.6	—	80.0	86.3
G-Dropout (Wang and Manning, 2013)	79.0	—	—	93.4	—	82.1	86.1
F-Dropout (Wang and Manning, 2013)	79.1	—	—	93.6	—	81.9	86.3
Tree-CRF (Nakagawa et al., 2010)	77.3	—	—	—	—	81.4	86.1
CRF-PR (Yang and Cardie, 2014)	—	—	—	—	—	82.7	—
SVM _S (Silva et al., 2011)	—	—	—	—	95.0	—	—

From Yoon Kim

Take Action

- Form your group
- Find a cool team name

2020BT5153 Project Groups ☆

File Edit View Insert Format Data Tools Add-ons Help [Last edit was yesterday at 4:36 PM](#)

100% \$ % .0 .00 123 Default (Ari... 10 B I S A

	Name				
	A	B	C	D	E
1	Name	Group Name	Student Number	Email	Group Number
2					1
3					1
4					1
5					1
6					1
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					