Machine Learning Introduction
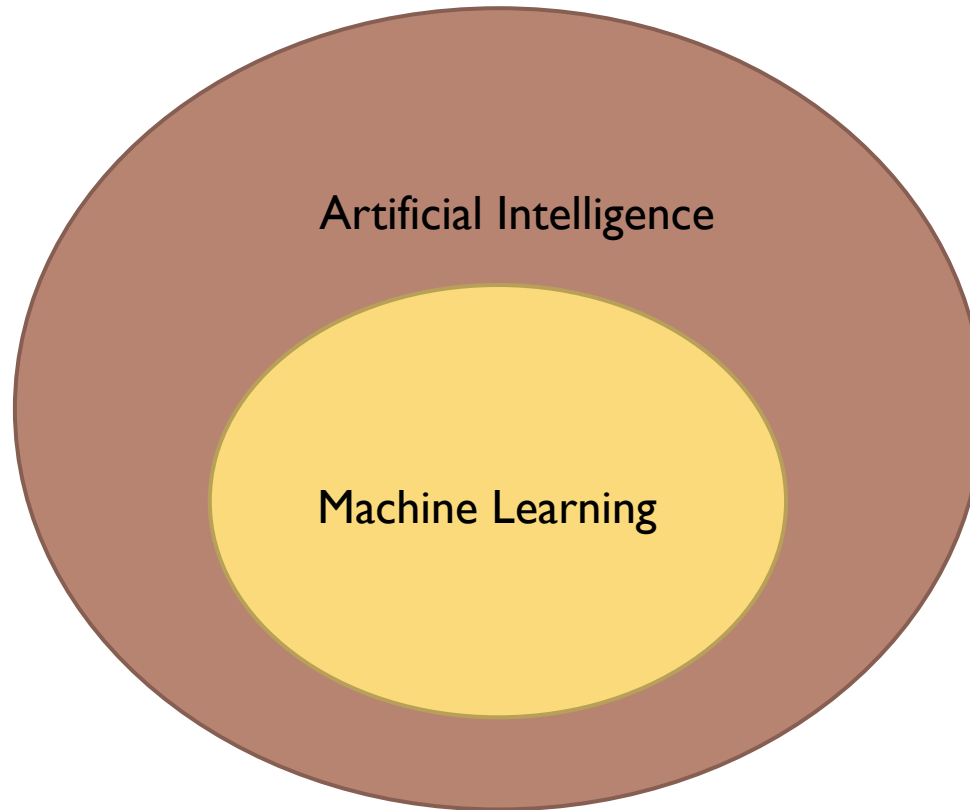
Outline

▶ Machine Learning concept

▶ Applications in Machine Learning
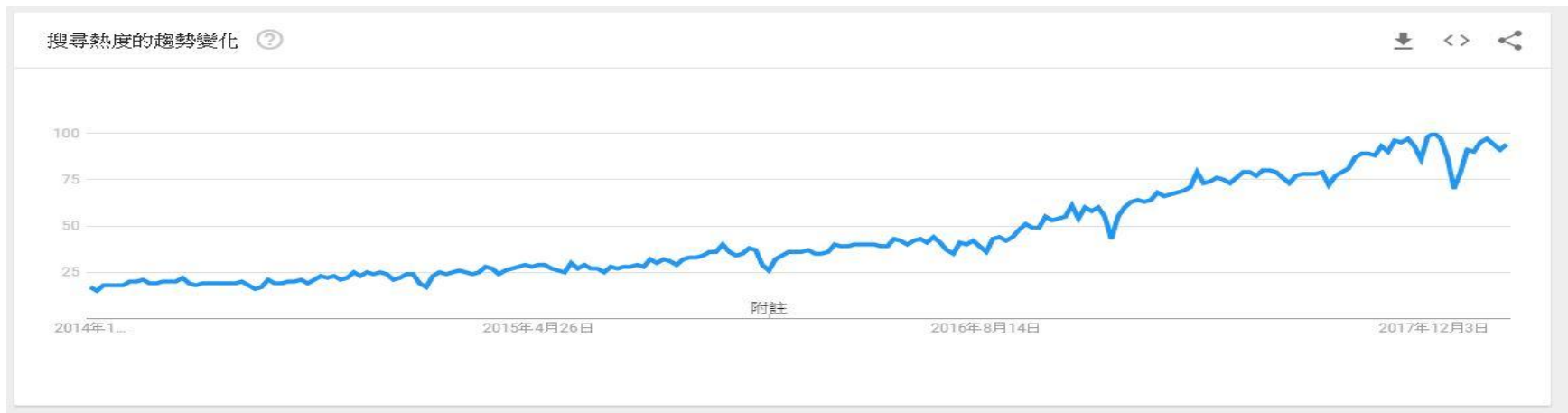
Machine Learning concept

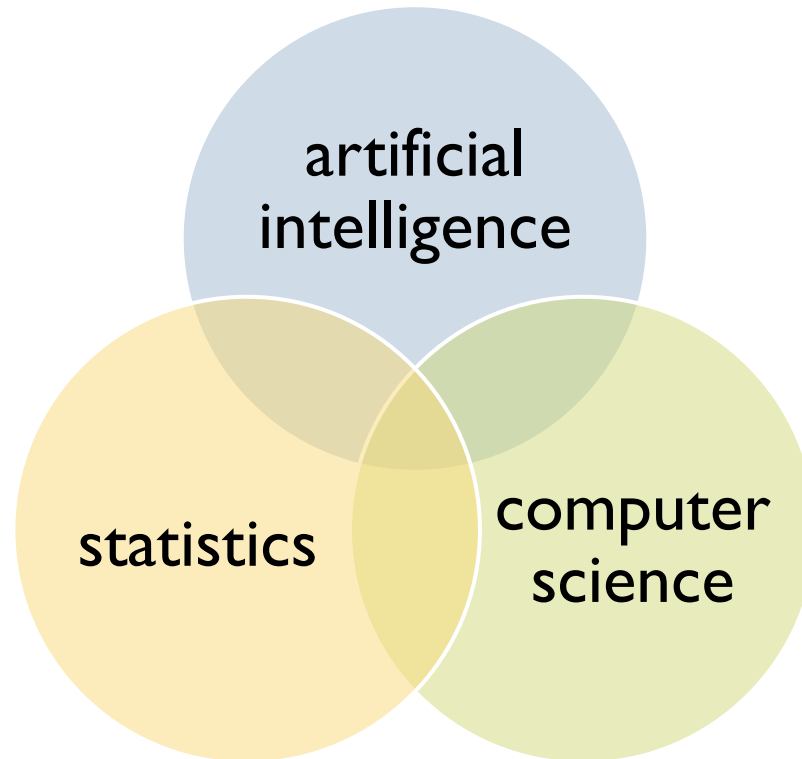# AI V.S ML

Artificial Intelligence

Machine Learning

Popularity of AI
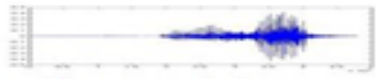

https://trends.google.com.tw/trends/explore?date=today%205-y&q=machine%20learning

Machine Learning knowledge

artificial intelligence

statistics

computer science

What is Learning?

▸ We want to find a function such that……

"Stock price today"

$$f(\theta)$$

"hello"

"2"

"Stock price tomorrow"

Different kind of Learning

supervised	unsupervised
semi-superviesed	reinforcement

Different Machine Learning style
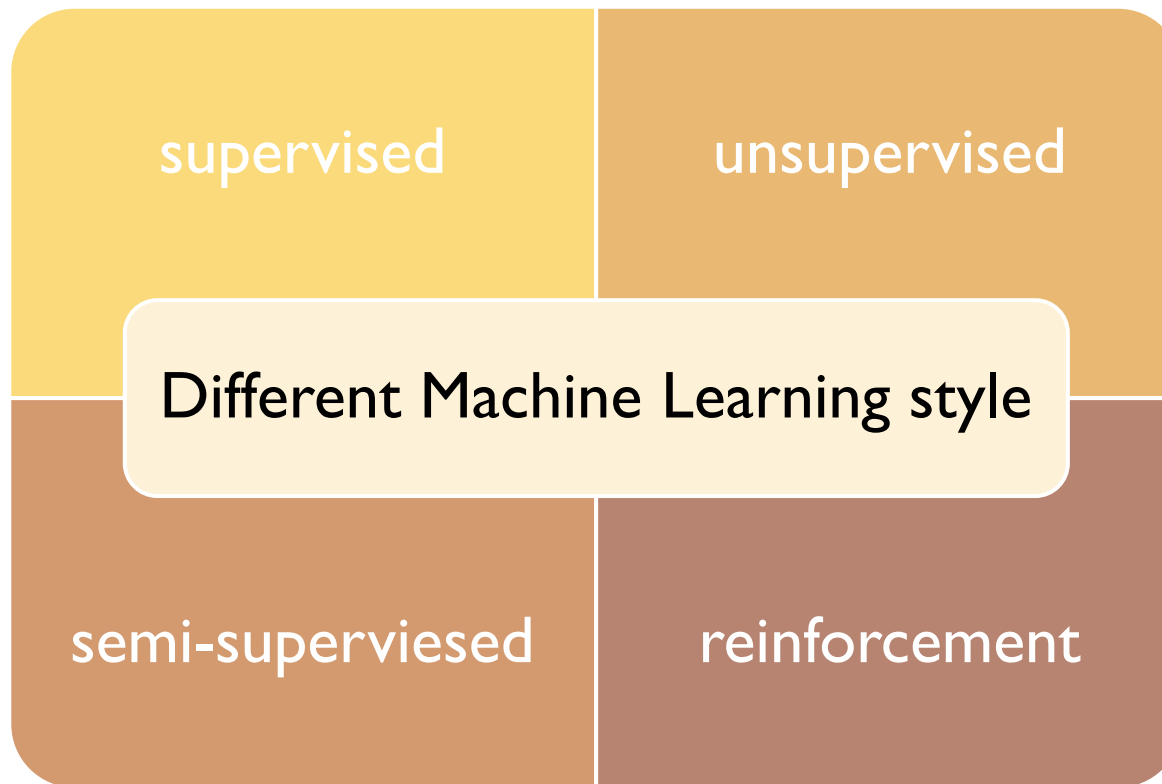
Different kind of Learning

▸ ## Supervised Learning

  ▸ Give machine data and corresponding label

 This is zero

 This is one

 This is one

 This is zero

Different kind of Learning

▸ Unsupervised Learning

  ▸ Cluster unlabelled data

Different kind of Learning

▶ Semi-supervised Learning

  ▶ some data are labelled but some are not

  ▶ supervised + unsupervised learning
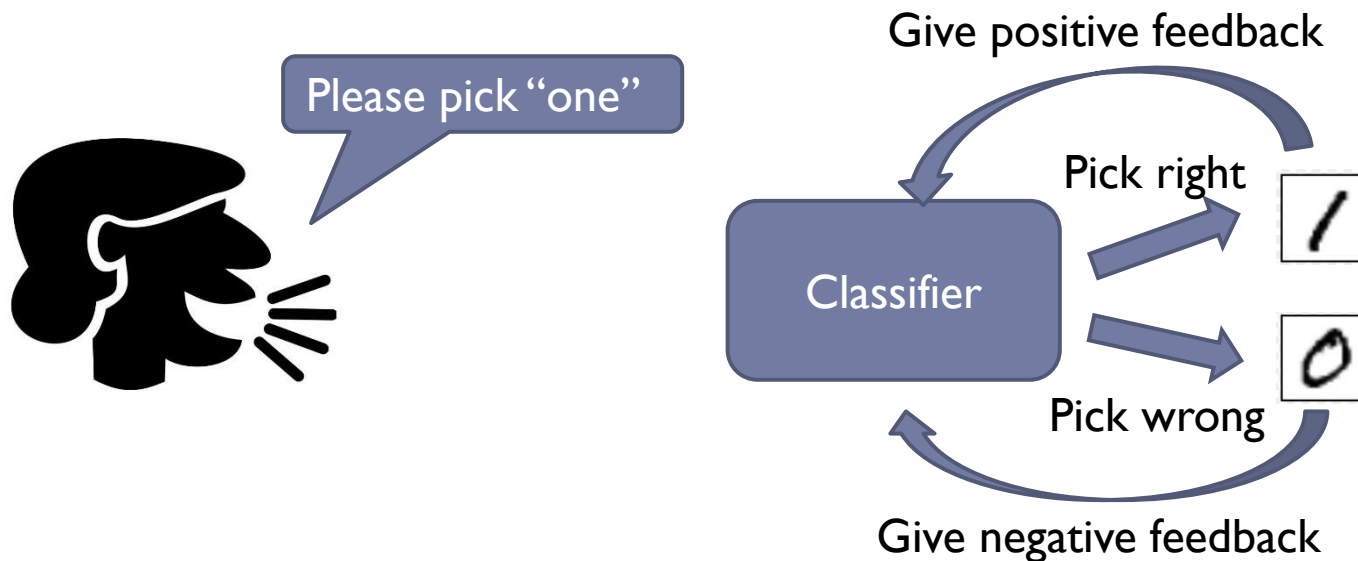
This is zero

This is one

???

???

Different kind of Learning

▸ Reinforcement Learning

    ▸ Given unlabelled data, make machine self-learning

Regression V.S. Classification
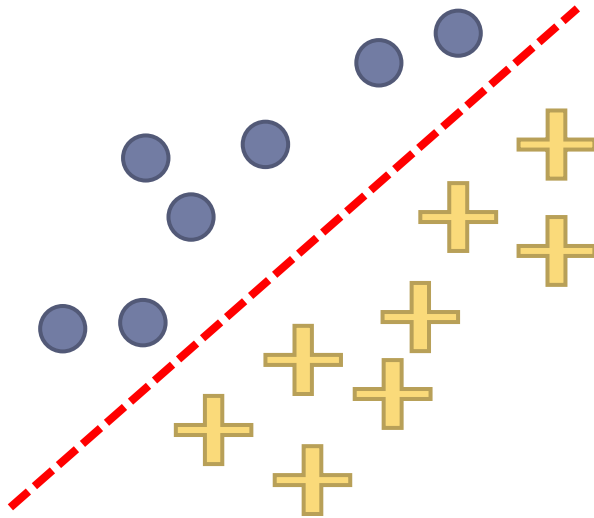
Regression: y is a real value, $y \in R$

$$f: R^d \rightarrow R \ \ f \ is \ called \ a \ regressor$$

Classification: y is discrete, $y \in \{-1, +1\}$
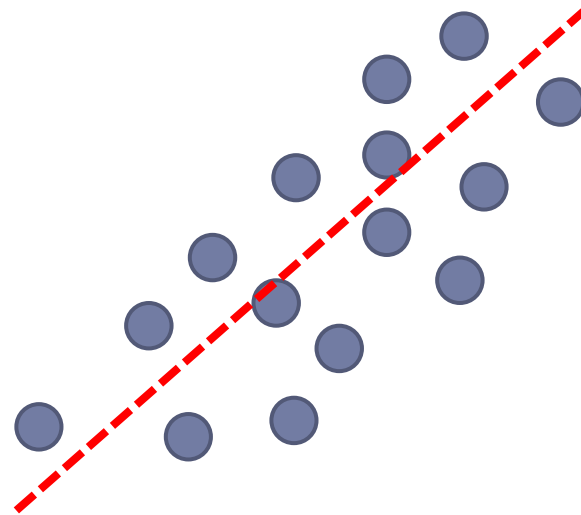
$$f: R^d \rightarrow \{-1, +1\} \ \ f \ is \ called \ a \ classifier$$

Regression V.S. Classification

Classification

Regression

Regression example

Size (feet²)	Number of bedrooms	Number of floors	Age of home (years)	Price ($1000)
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
...	...	...	...	...

house price prediction

Classification example

? cat or dog

Machine Learning Workflow

Input → Feature Engineering → ML Model → Output
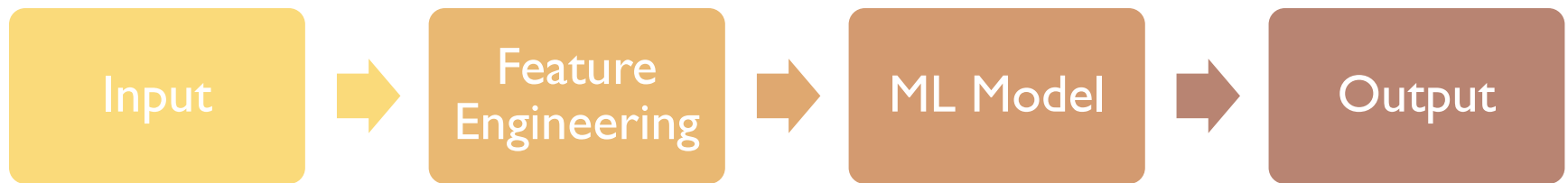
Feature engineer

- Fill missing value
  - Numerical
    - drop data that have missing value
    - Fill median/mean as missing value
  - Categorical
    - Fill mode as missing category
    - Fill "other" as missing category
- Outliers detection
- ……
- more reference
  - https://vinta.ws/code/feature-engineering.html

Split data

- Split total house price data into two part
  - 80% as training data
  - 20% as testing data
- Training data is used to make machine learn
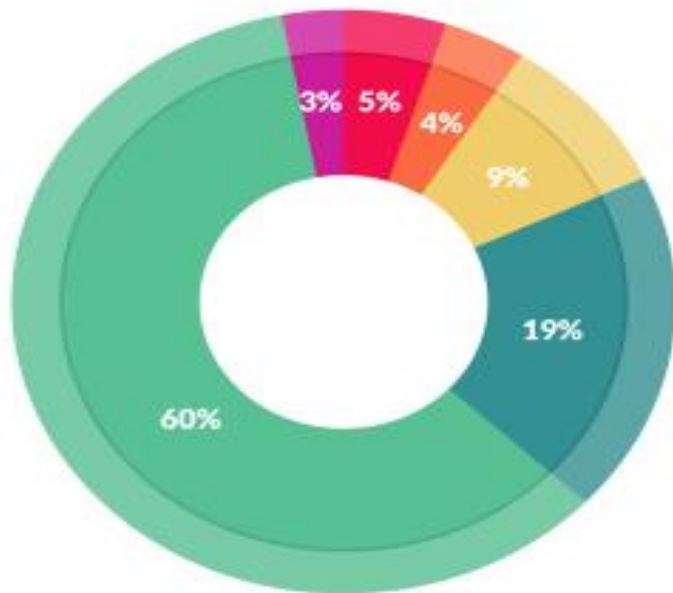- Testing data is used to validate if machine actually learn well

Training Data	Testing Data
80%	20%
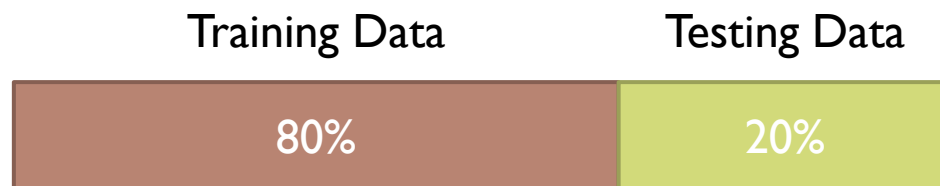
Data Normalization

▸ Try to scale all features into [0, 1] or [-1, 1]

▸ Some common normalization method

▸ $\dfrac{x_i - x_{min}}{x_{max} - x_{min}}$

▸ $\dfrac{x_i - \mu}{\sigma}$

original data → zero-centered data → normalized data
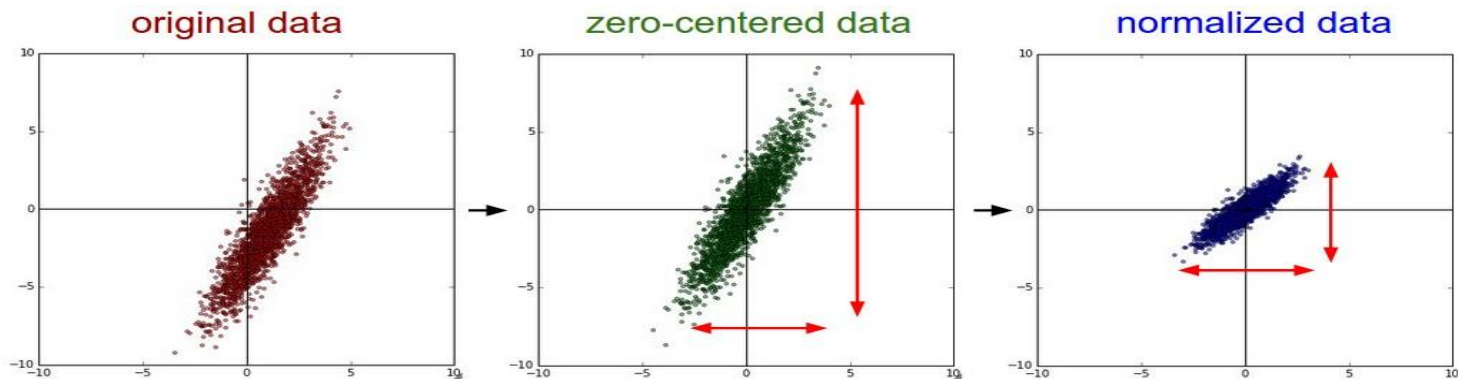
Feature engineer

What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Select model

	Unsupervised	Supervised

Continuous	• Clustering & Dimensionality Reduction
	    ○ SVD
	    ○ PCA
	    ○ K-means
		• Regression
		    ○ Linear
		    ○ Polynomial
		• Decision Trees
		• Random Forests

Categorical	• Association Analysis
	    ○ Apriori
	    ○ FP-Growth
	• Hidden Markov Model
		• Classification
		    ○ KNN
		    ○ Trees
		    ○ Logistic Regression
		    ○ Naive-Bayes
		    ○ SVM

what we will focus in this course

validate trained model - regression

▸ MSE

   ▸ average of the squares of the errors/deviations(difference between the estimator and what is estimated)

▸ R squared(coefficient of determination)

   ▸ proportion of the variance in the dependent variable that is predictable from the independent variable

   ▸ R squared is close to 1 mean that the model is fitting better

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (f_i - \hat{y}_i)^2 \qquad\qquad R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

validate trained model – classification

▸ The most common way to measure of classification performance

  ▸ Note that please use testing data to evaluate

$$accuracy = \frac{\# \; of \; correct \; prediction}{\# \; of \; data}$$

validate trained model – classification

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

accuracy paradox

Assume we have a classifier to identify spams and it shows following results

	classified positive	classified negative
actual positive	10	15
actual negative	25	100

$$accuracy = \frac{10 + 100}{10 + 15 + 25 + 100}$$
$$= 73.3\%$$

What if the result change like the following?

	classified positive	classified negative
actual positive	0	25
actual negative	0	125

$$accuracy = \frac{0 + 125}{0 + 25 + 0 + 125}$$
$$= 83.3\%$$

accuracy paradox

validate trained model – classification

TP	TN

Correctly classified

FP	FN

Misclassified

Classified positive   Classified negative
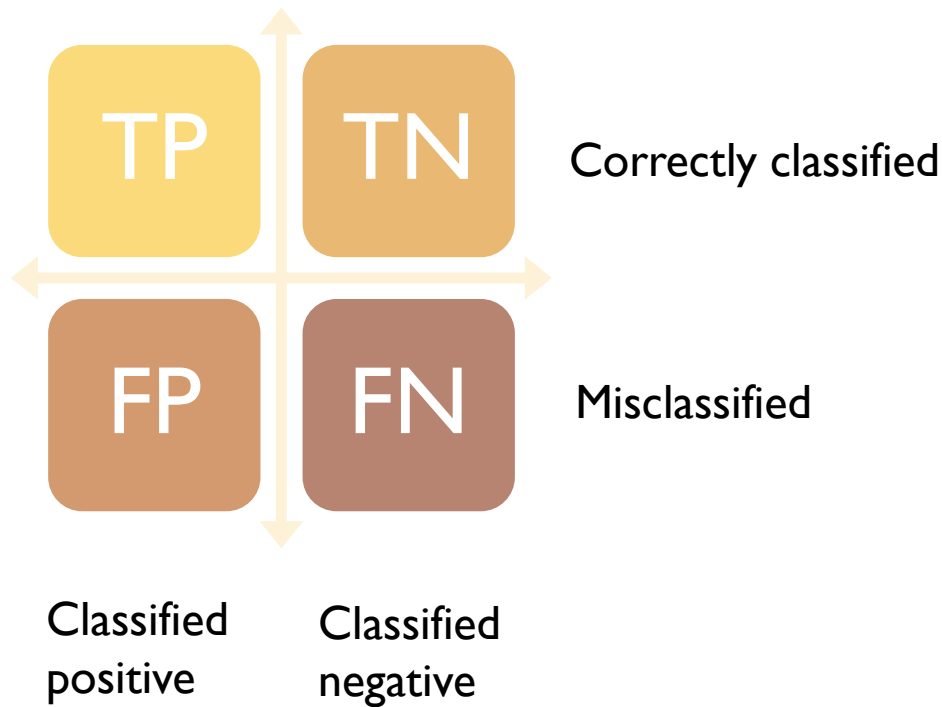
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$
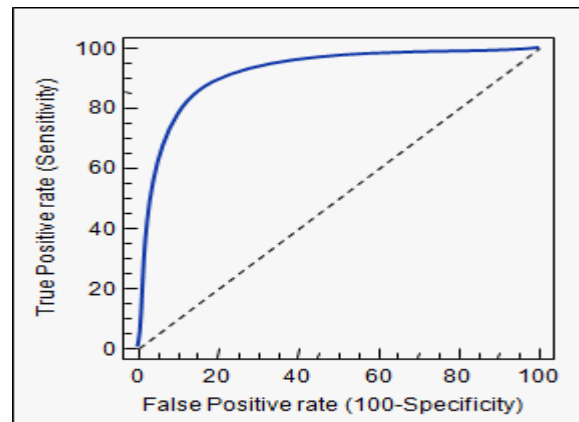
validate trained model – classification

▶ **F1-score**

   ▸ harmonic mean of precision & recall

   ▸ best value at 1 (perfect precision and recall) and worst at 0

$$F_1 = 2 * \frac{1}{\dfrac{1}{recall} + \dfrac{1}{precision}} = 2 * \frac{precision * recall}{precision + recall}$$

validate trained model – classification

- ROC curve(receiver operating characteristic curve) illustrates the diagnostic ability of a binary classifier system under different threshold
  - created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings
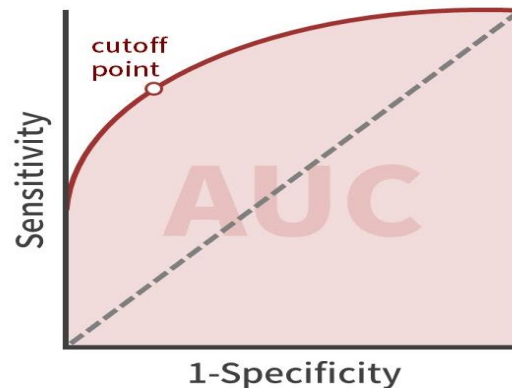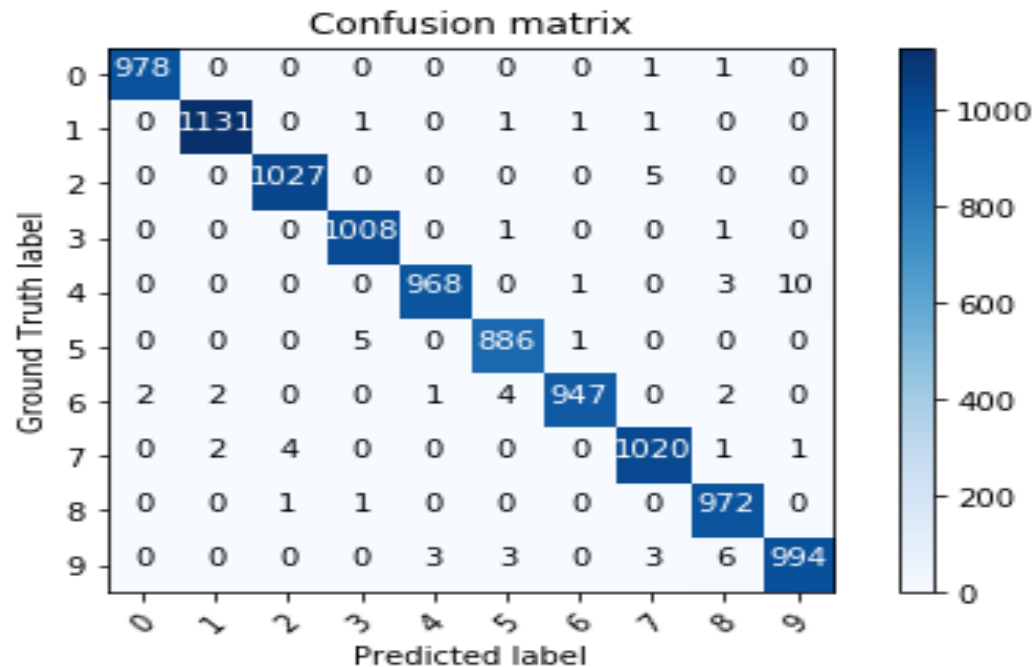
validate trained model – classification
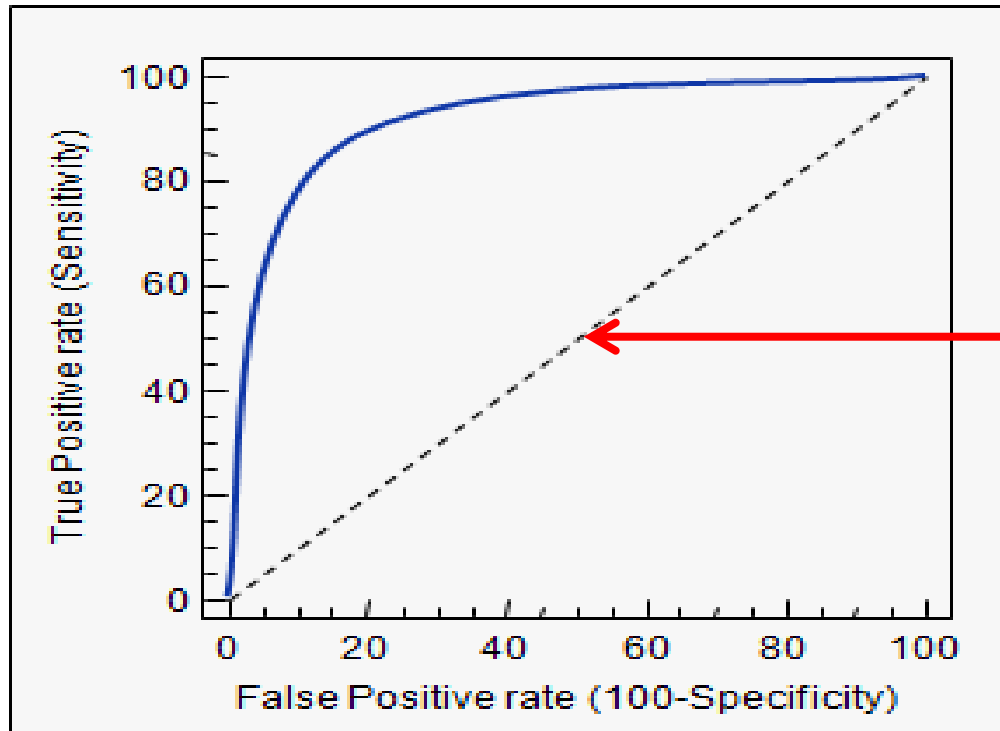
random choose

validate trained model – classification

▶ AUC(area under curve) value is the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example

  ▶ Usually range in [0.5, 1]
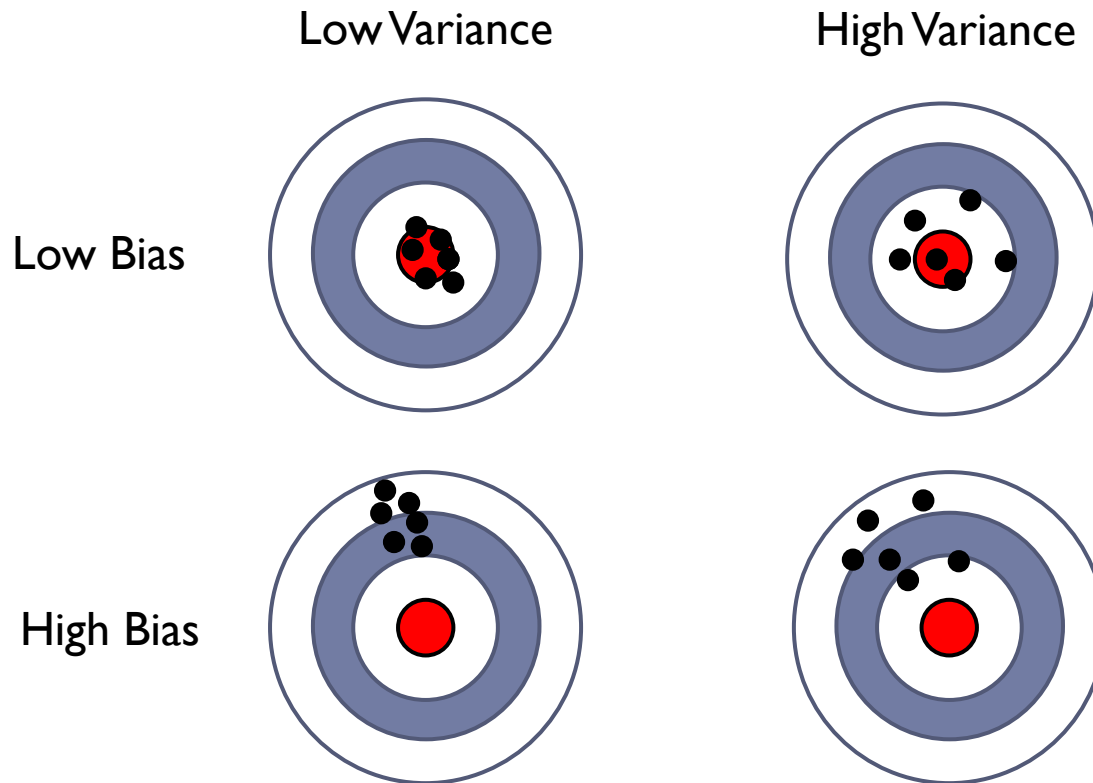
  ▶ AUC values bigger mean the classifier perform better
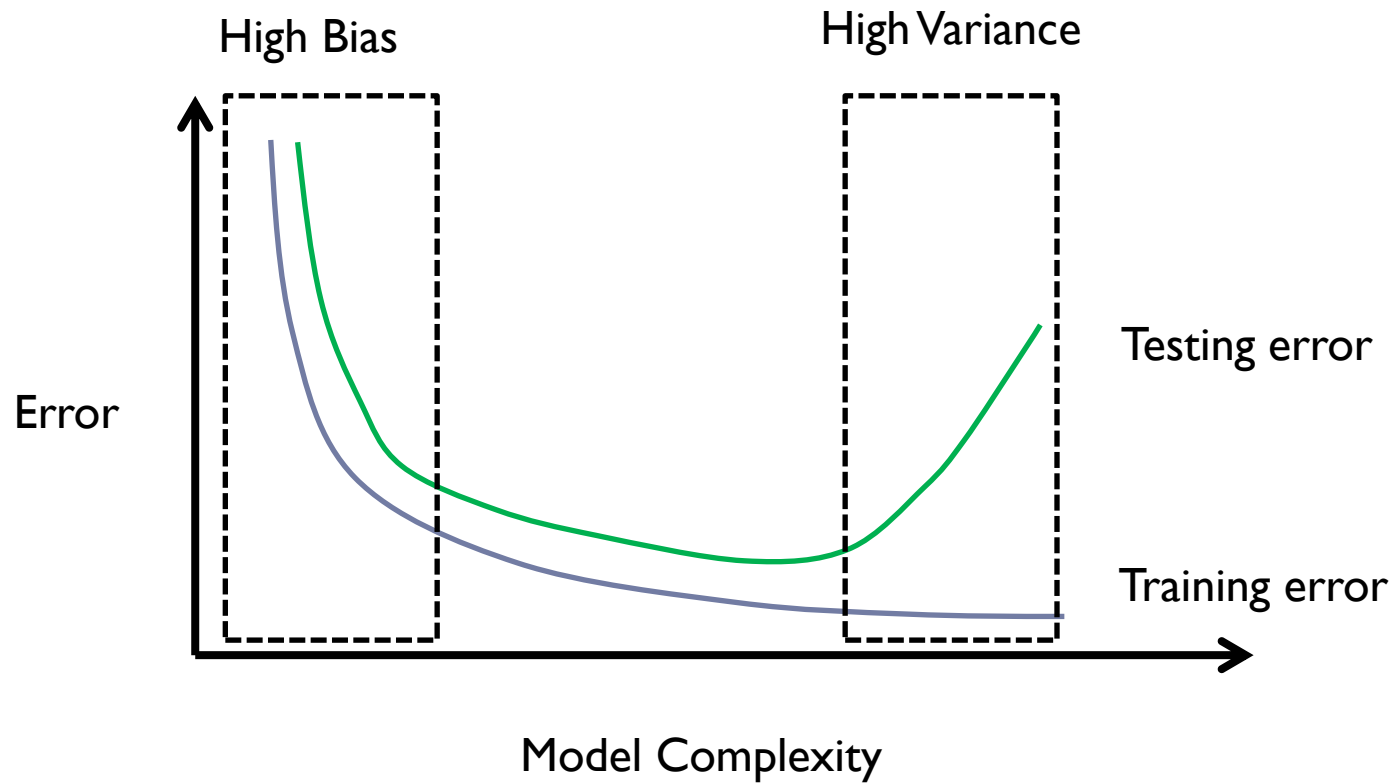
validate trained model – classification
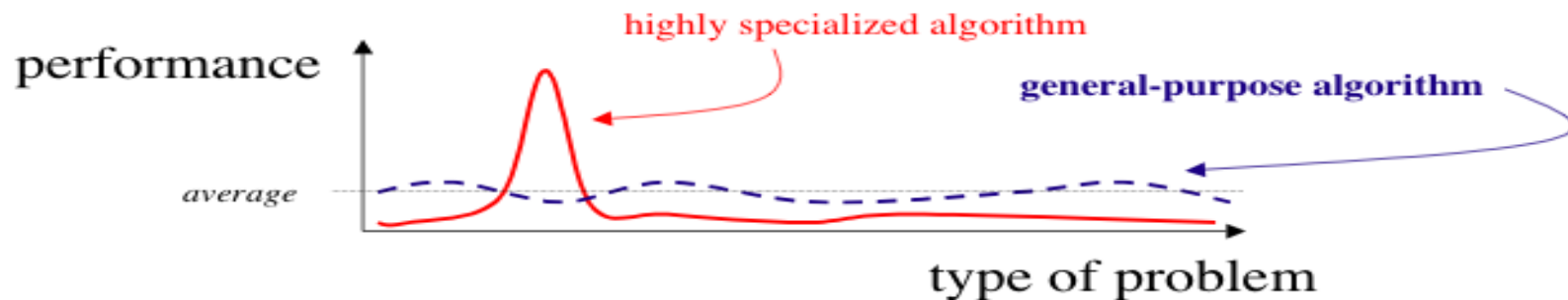

confusion matrix

Bias-Variance Tradeoff

Low Variance   High Variance

Low Bias

High Bias

Bias-Variance Tradeoff

High Bias

High Variance

Error

Model Complexity

Testing error

Training error

No free lunch theory

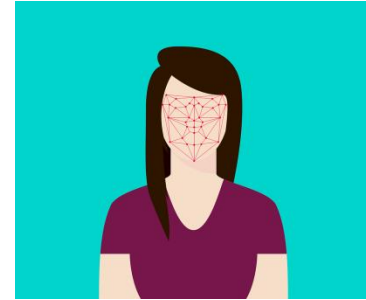▸ No free lunch theory

▸ http://www.no-free-lunch.org/

applications in machine learning

Machine Learning is Everywhere

auto vehicle

payment

medical treatment

robot advisor

Machine Learning is Everywhere

drone

precision marketing

smart factory
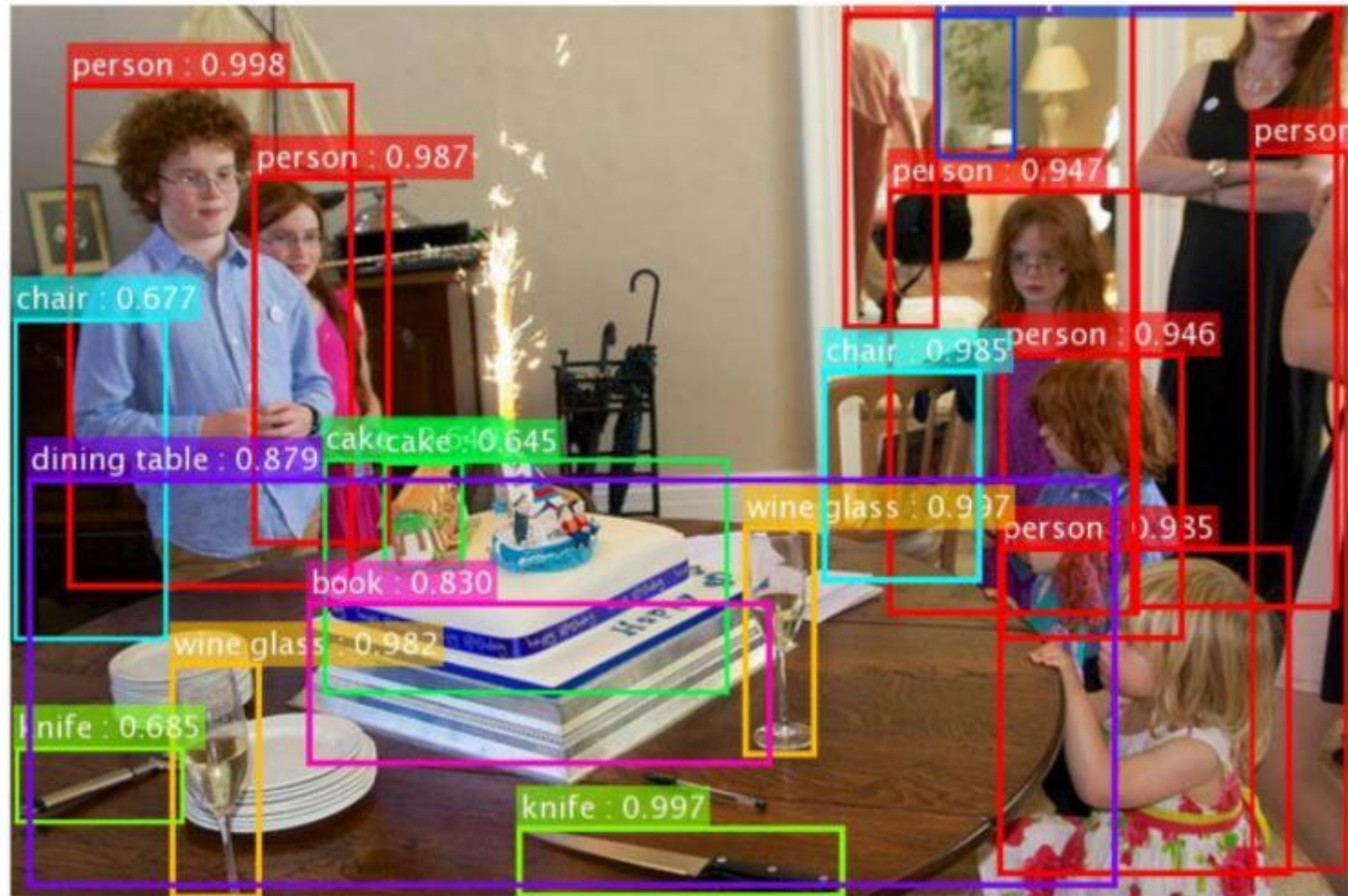

voice assistant

Object Detection

Image Caption

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.
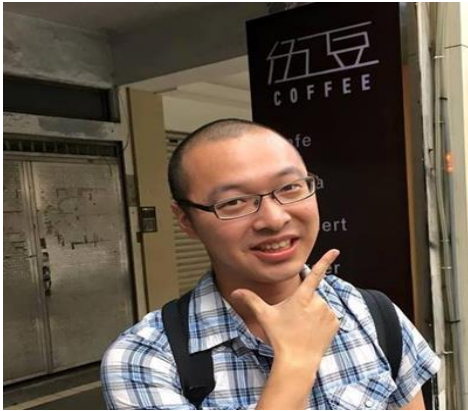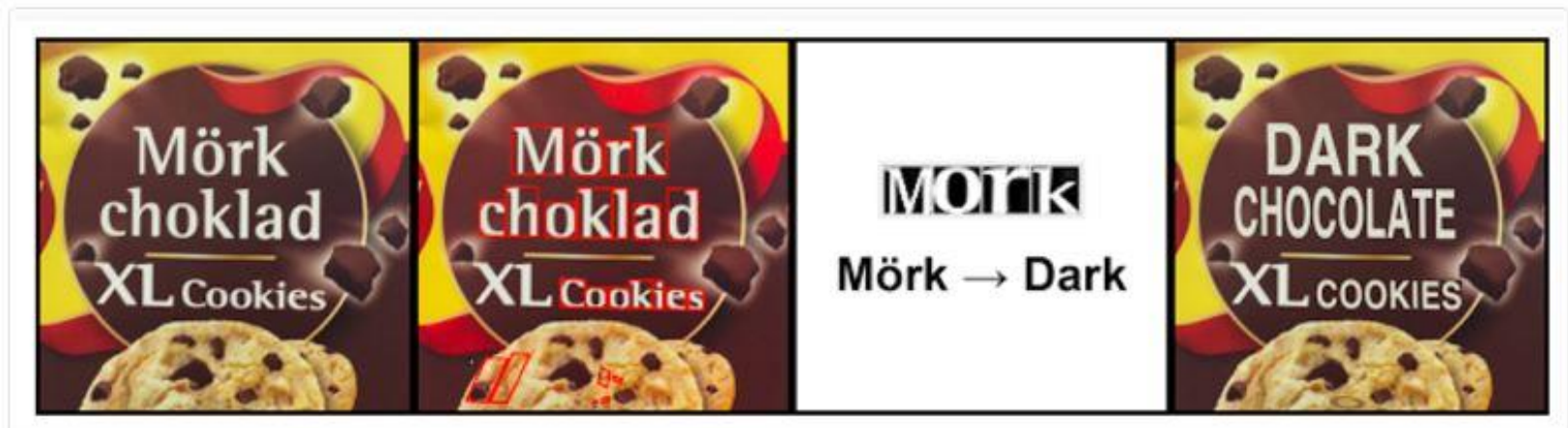
Text to image

Image Style transfer

Image Style transfer

Translation

Mörk → Dark

Text Generator

```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.
```

Text Generator

```
/*
 * Increment the size file of the new incorrect UI_FILTER group information
 * of the size generatively.
 */
static int indicate_policy(void)
{
  int error;
  if (fd == MARN_EPT) {
    /*
     * The kernel blank will coeld it to userspace.
     */
    if (ss->segment < mem_total)
      unblock_graph_and_set_blocked();
    else
      ret = 1;
    goto bail;
  }
  segaddr = in_SB(in.addr);
  selector = seg / 16;
  setup_works = true;
  for (i = 0; i < blocks; i++) {
    seq = buf[i++];
    bpf = bd->bd.next + i * search;
    if (fd) {
      current = blocked;
    }
  }
  rw->name = "Getjbbregs";
  bprm_self_clearl(&iv->version);
  regs->new = blocks[(BPF_STATS << info->historidac)] | PFMR_CLOBATHINC_SECONDS << 12;
  return segtable;
}
```

Generated text

Text Generator

楊是“”過是道“是不，”：”是道“：“過一了。是“是：”：”“是不是，過是是，”“”“：：是：”“不一不，”：：“：過：”過一，“：：是：“”“不不，過一一一，“”過是，過是不，過道是一一，“”過不的不一，”“”是道過一。“是：是是不，“是不是一，”是：：過一，“”是是是，“：“”過一了的，“是：：“”是：“”是：是道“：“是是，“”“：”“”是是是一。是不一一。過：“”道過，”過是是一。”：：過道”“：是“：“是一不，是“是道：是一，“是是道是是不，“是一，是不不，“是道”道是是一。是

Epoch = 1

Text Generator

楊過不住，只不是他們，但不能不能是他。" 楊過道："這位你這一句話們，不知是我不能，你也要不知不會，但不知你們的。" 郭襄道："我們一聲不是你？" 楊過道："那麼！"楊過笑道："好啊，你是你這一人，我不是他的的。" 楊過道："我……這是你，我是這一個人的一個，我便是不能。" 小龍女道："你不能再不知你，我們們們的不是你。" 楊過笑想："那是他一句話來，我不是不能好，你不是我？我不肯跟他說，我也不肯跟你。" 黃蓉道："你們這般好好。你說我不是？你不知不會。"郭芙道："你說到我？"黃藥師道："你說到我們

Epoch = 5

Text Generator

楊過一燈大師，也不知這兩個字，就算他一起去。" 楊過聽得"啊"的一聲，臉色大亂。楊過笑嘻的道："我不知道，那也不是不錯？"楊過微微一笑，說道："你師父和我相見，那也不敢說話。"黃蓉微微一笑，叫道："你們在那里去了？"黃蓉道："我不肯說的，我要去瞧瞧，你也要去了！我說甚麼？"郭靖笑了起來，道："這是你的孩子，不是你的，你便去啦。" 楊過心想："我不知是我姑姑，我自然不知。"說著又道："你們不肯再說，他就要去。" 楊過一驚，問道："我是誰？"楊過道："你不知道。" 楊過道："我不知道，我也就算得了。"

Epoch = 20

Text Generator

楊 "哥，他越自言還。" 黃蓉一怔，一時不暇細想。郭靖與黃蓉夫婦說起郭靖、黃蓉及孩，是以心存一縱的敬意而已。郭芙道："過兒，你也知道。我和小龍女重赴大事，小妹已在古墓中住不了幾年，這時再來助告。"公孫止一生心情，心中一動："終南山一時尚未明白，但我一生孤苦，若是出言誘他治愈致功夫的情由，何以我一刻不死？而我自己也是難以挽到，又說你聰明過約，我也不想活了？" 郭芙大叫："姑姑！" 楊過伸手接住，笑道："你瞧，我也不用聽見麼？"郭芙道："你說我有甚麼不好？" 小龍女一呆，道："我媽媽這般說，我也決意要

Epoch = 100

AI poetry

看那星，闪烁的几颗星
西山上的太阳
青蛙儿正在远远的浅水
她嫁了人间许多的颜色

Microsoft 小冰

AI poetry

https://webapps.msxiaobing.com/V3/Portal?task=poem
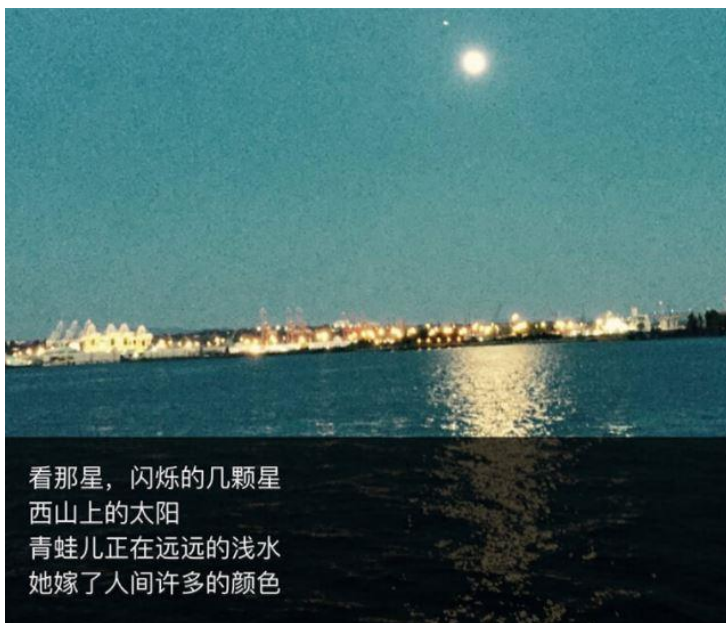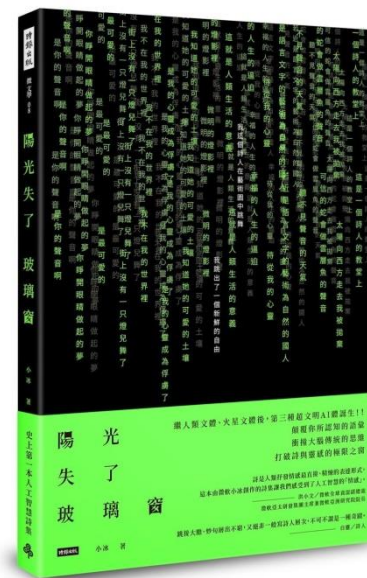
Char level model

- Fake Latex document
  - https://cs.stanford.edu/people/jcjohns/fake-math/4.pdf
- Generate baby name
  - https://cs.stanford.edu/people/karpathy/namesGenUnique.txt
- Jin-Yong novel generator
  - https://github.com/isaac60103/TensorFlow-LSTM
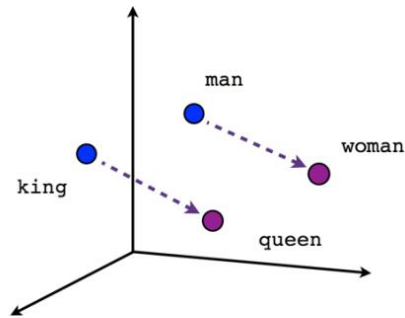
Word to vector
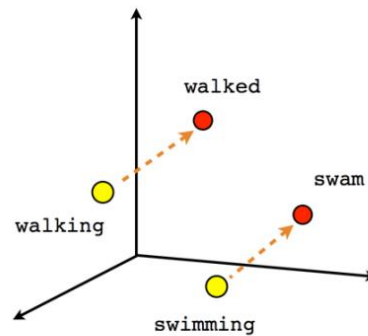
Male-Female   Verb tense   Country-Capital
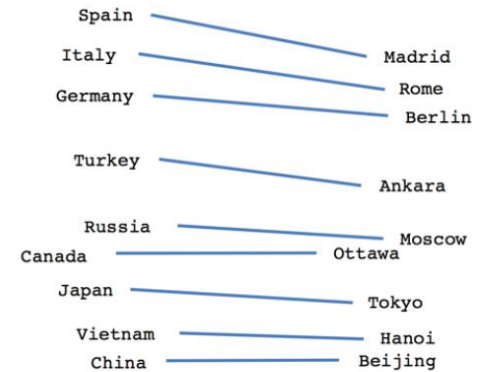
vector(king) − vector(queen) = vector(man) − vector(woman)

vector(walking) − vector(walked) = vector(swimming) − vector(swam)

vector(Spain) − vector(Italy) = vector(Madrid) − vector(Rome)

Word to vector
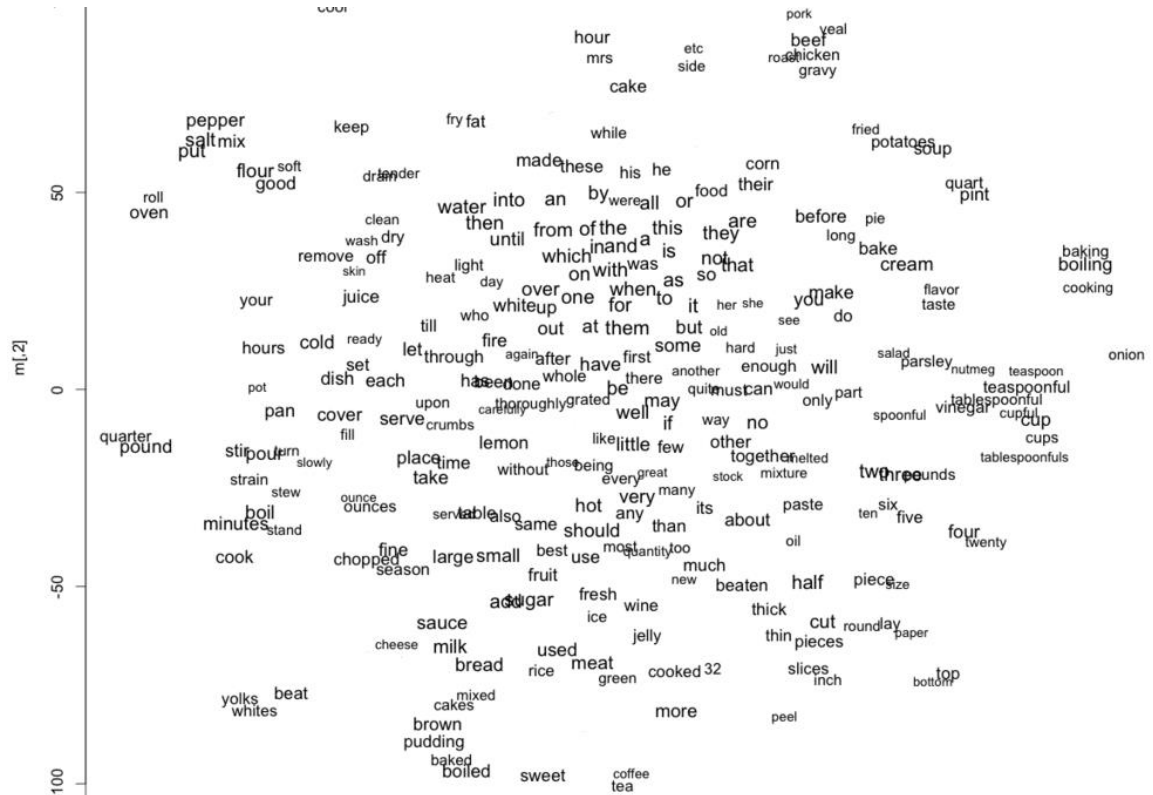
AlphaGo

AlphaGo Zero

▶ AlphaGo zero was released at 2017/10/18

40 days

AlphaGo Zero surpasses all other versions of AlphaGo and, arguably, becomes the best Go player in the world. It does this entirely from self-play, with no human intervention and using no historical data.
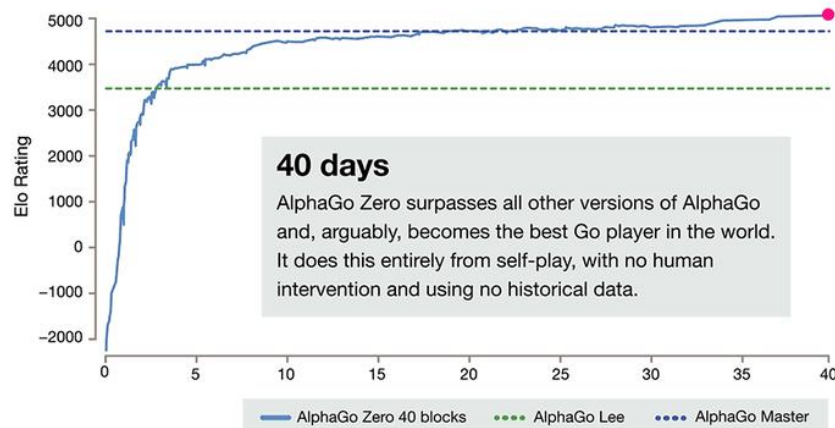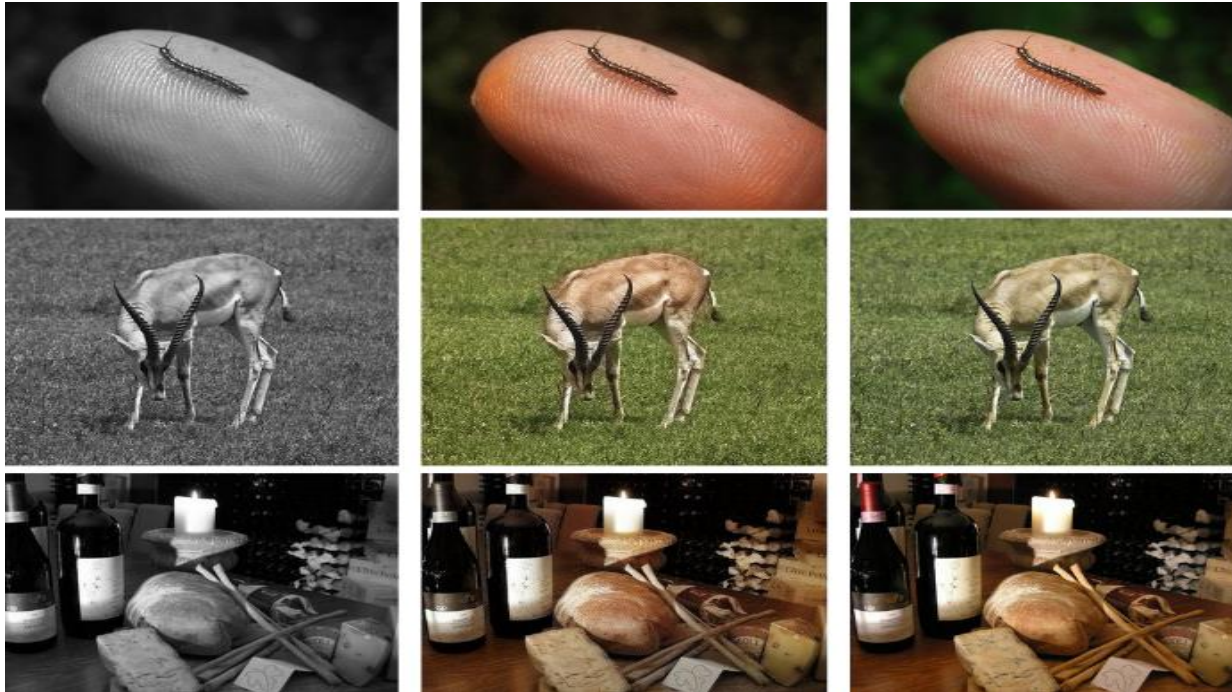
It only took 40 days to self-study to beat all previous version Alphago

AlphaGo Zero
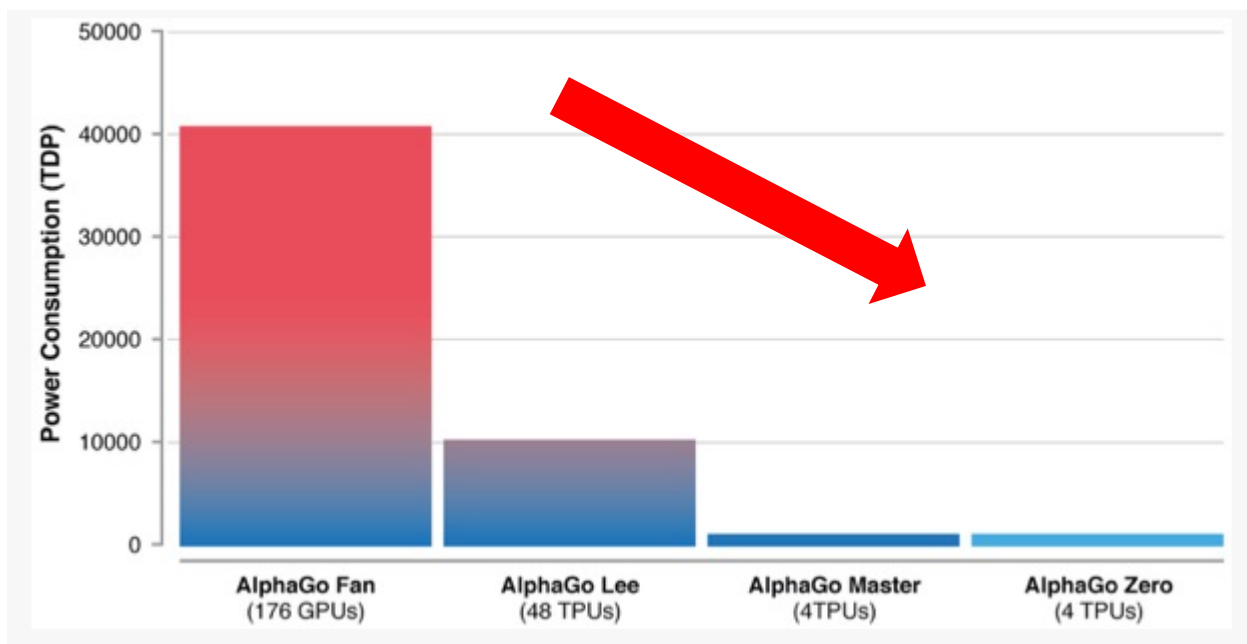
樊麾(2015/10)　　李世乭(2016/03)　柯潔(2017/05)　　　(2017/10)
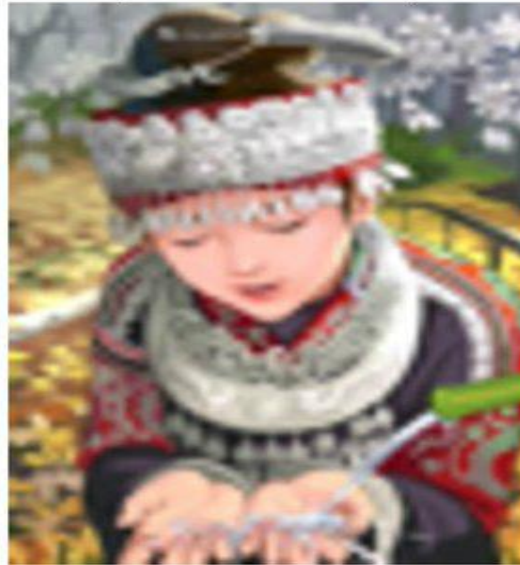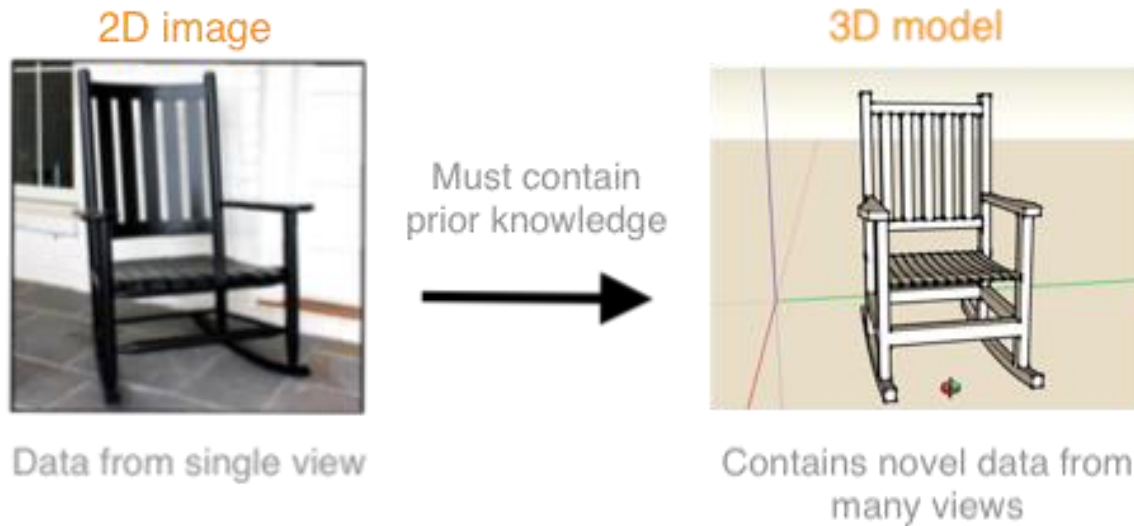
Automatic Colorization

Fake image

smiling woman − neutral woman + neutral man = smiling man

man with glasses − man without glasses + woman without glasses = woman with glasses
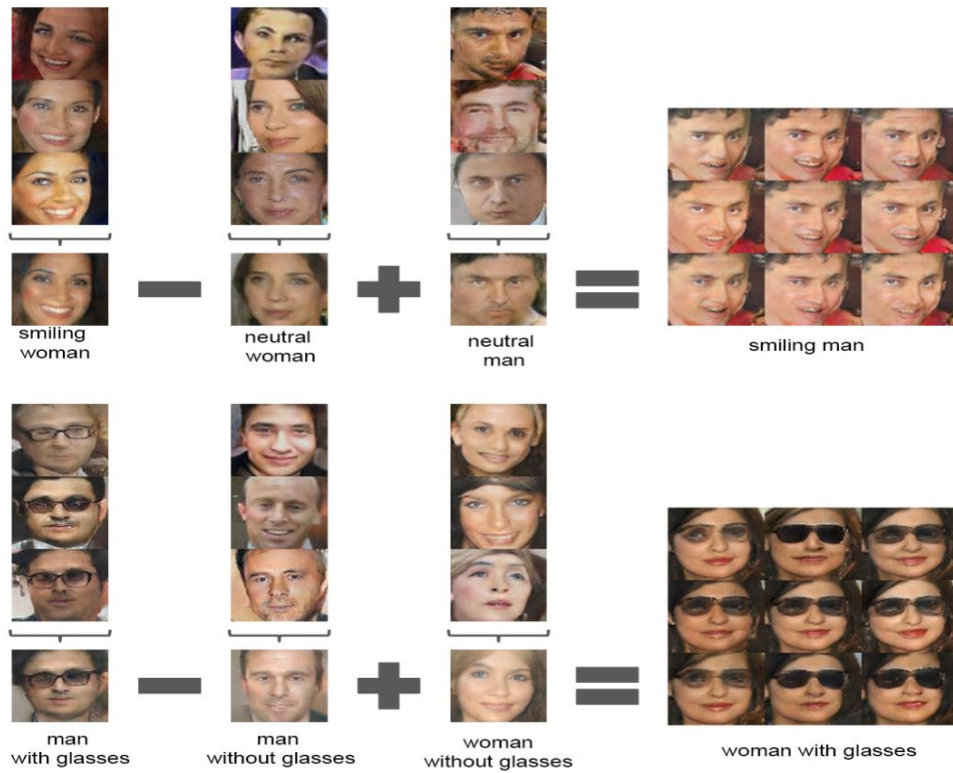
Super resolution

original

bicubic
(21.59dB/0.6423)

SRGAN
(20.34dB/0.6562)

2D to 3D image

2D image — Data from single view

Must contain prior knowledge

3D model — Contains novel data from many views

AI trading

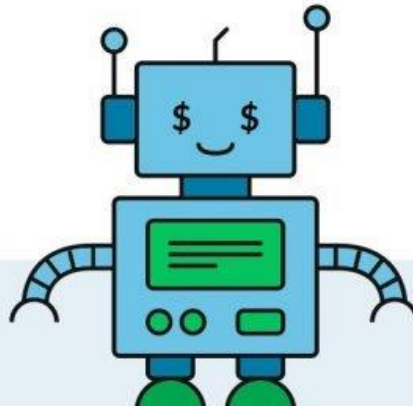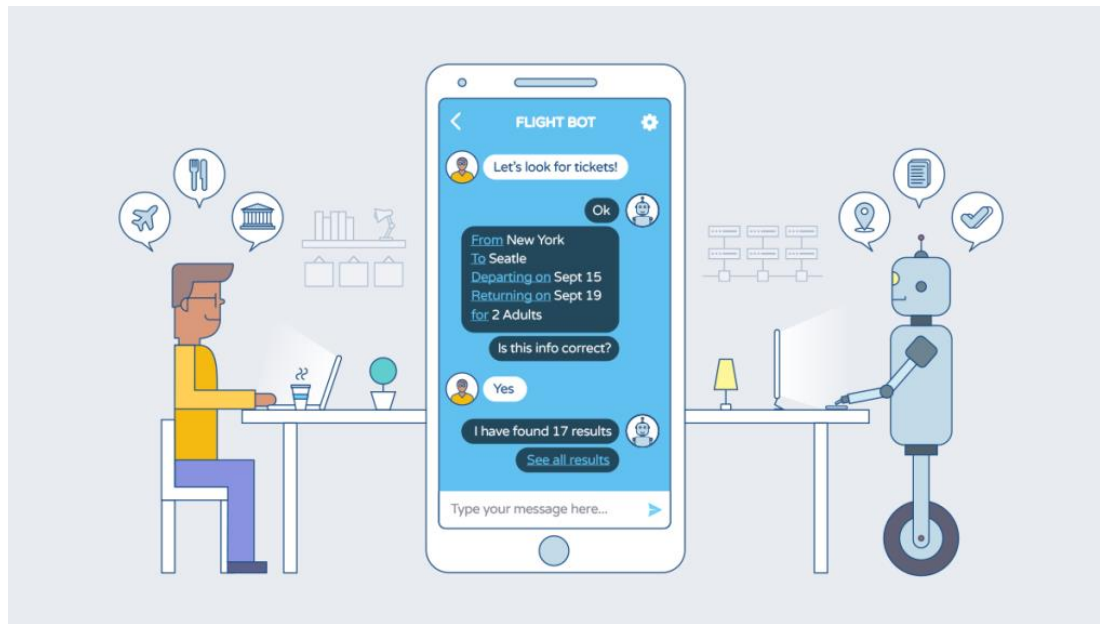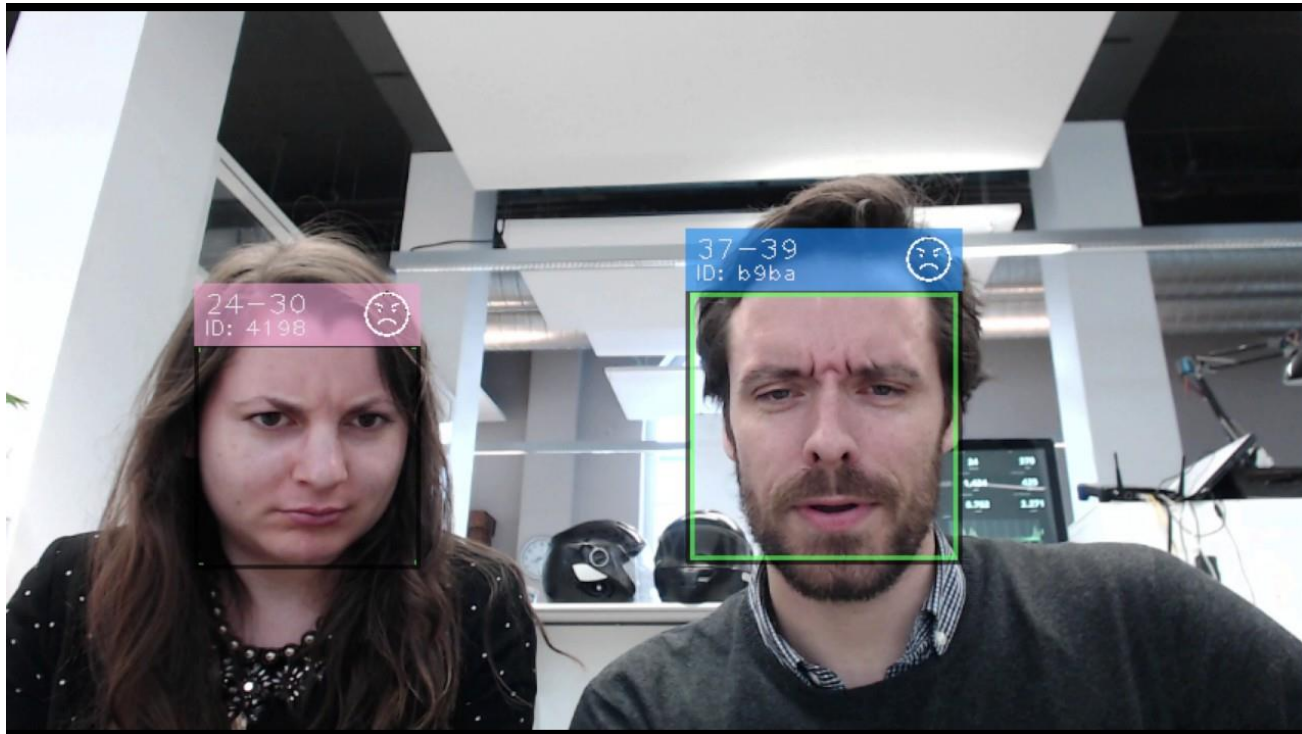# Robot Advisor

Do-It-Yourself

Robo Advisor

Financial Advisor

Chatbot

Age/Gender Detection

Face recognition

Optical Character Recognition(OCR)

AI in Design

"鲁班" system by alibaba

AI in Education
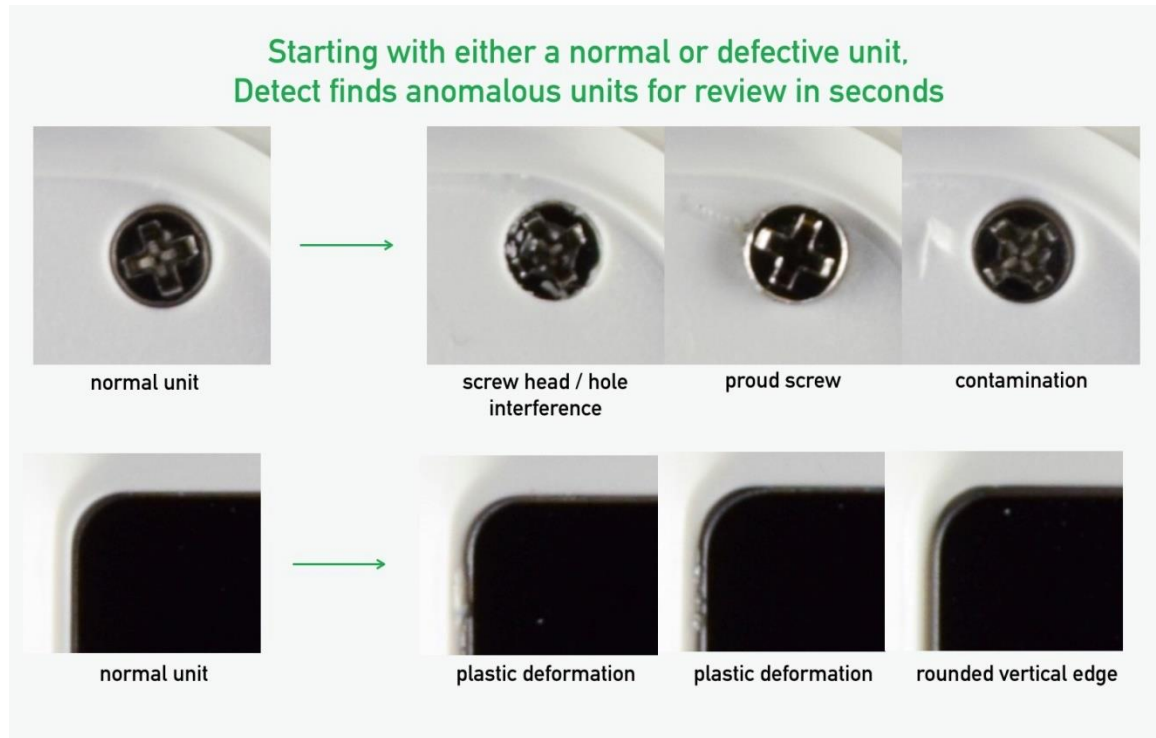

Qubena system by COMPASS company

AI in Recruiting

AI recruiting by leoforce company

AI in Entertainment

AI virtual celebrity by star.ai

AI in manufacturing
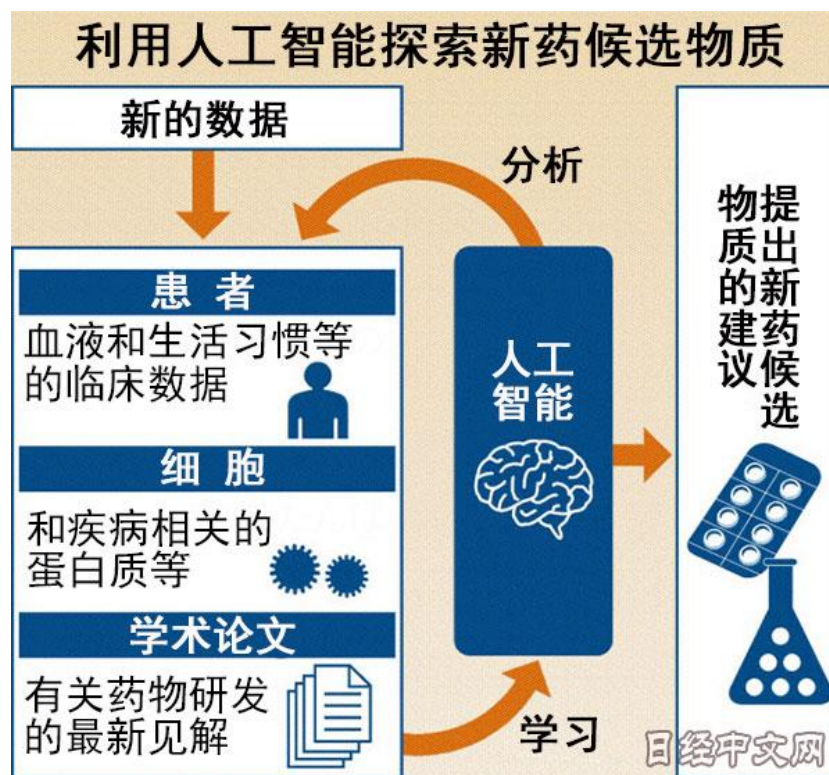

Starting with either a normal or defective unit,
Detect finds anomalous units for review in seconds

normal unit → screw head / hole interference | proud screw | contamination

normal unit → plastic deformation | plastic deformation | rounded vertical edge

Unmanned store

AI in Agriculture

AI in Agriculture

AI in medical discovery

利用人工智能探索新药候选物质

新的数据

分析

患者
血液和生活习惯等的临床数据

细胞
和疾病相关的蛋白质等

学术论文
有关药物研发的最新见解

人工智能

学习

提出的新药候选物质的建议

日经中文网

AI in security

AI in biometric