Introduction to machine learning

Tao Tan

Who is teaching this

People

- Tao Tan

- Generalized electric medicine (GEM) research group – in faulty of applied science
- Image and video processing / analysis
- Three application domains:

Medical Imaging

Video monitoring

Nature language processing

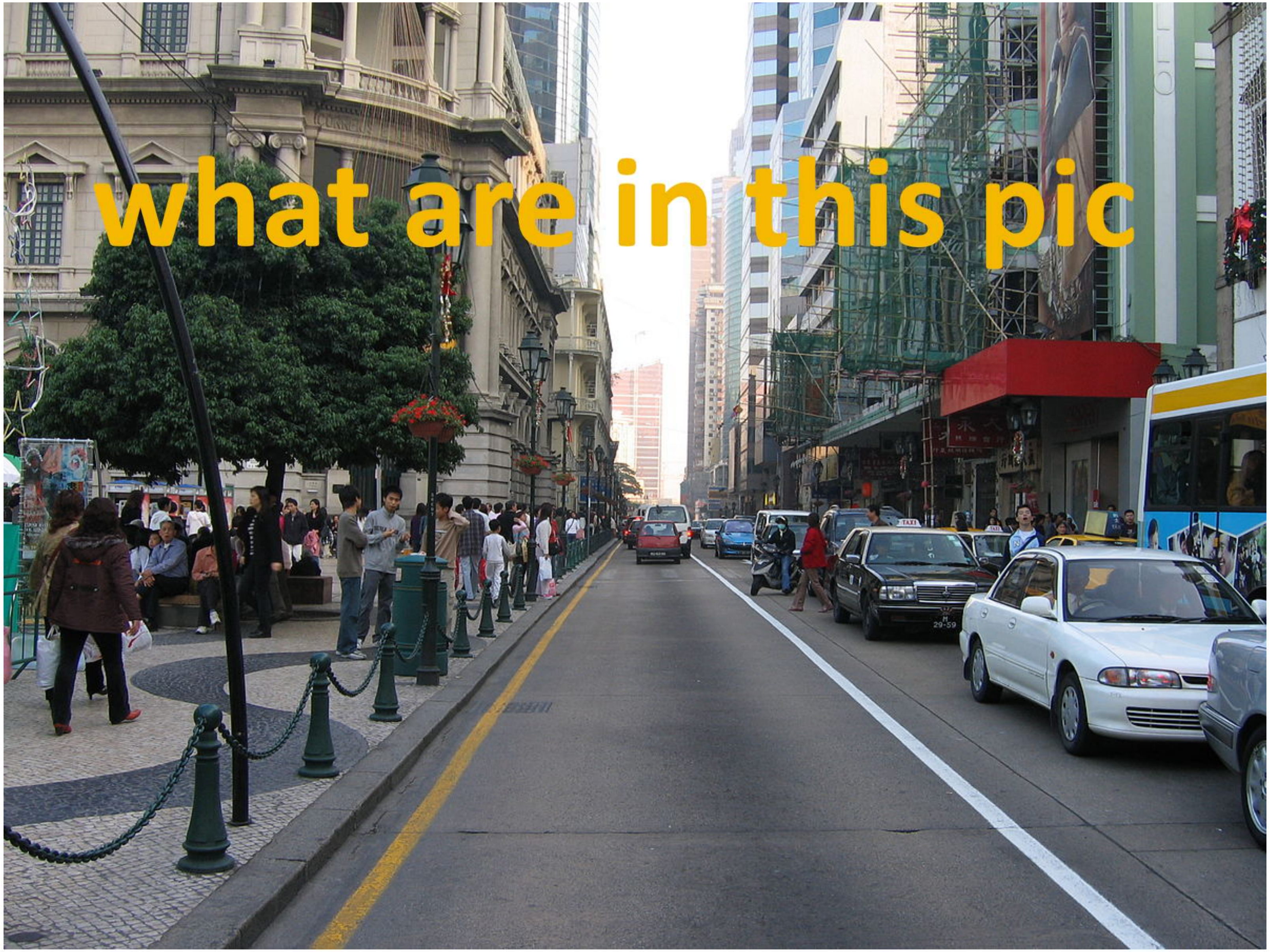
Introduce each other

个人简历

2002年10月至2007年7月	浙江大学，生物医学工程系本科 优秀毕业生
2007年10月至2009年9月	荷兰艾因霍恩理工大学，生物医学工程硕士
2010年1月至 2014年2月	荷兰 Radboud 大学，计算机系/放射科博士 （USNews 放射科世界排名第四）
2017年6月至2020年12月	荷兰艾因霍恩理工大学， 客座助理教授
2014年1月至2014年10月	荷兰Radboud 大学医学院博士后研究员
2014年10月至2018年3月	荷兰 ScreenPoint Medical 高级科学家，钼靶乳腺AI产品负责人
2014年10月至2019年5月	美国 Qview Medical，算法顾问，三维乳腺AI产品负责人
2019年6月至2022年7月	**通用医疗资深科学家，医疗人工智能项目经理**
2022年8月至今	澳门理工大学副教授

Contents

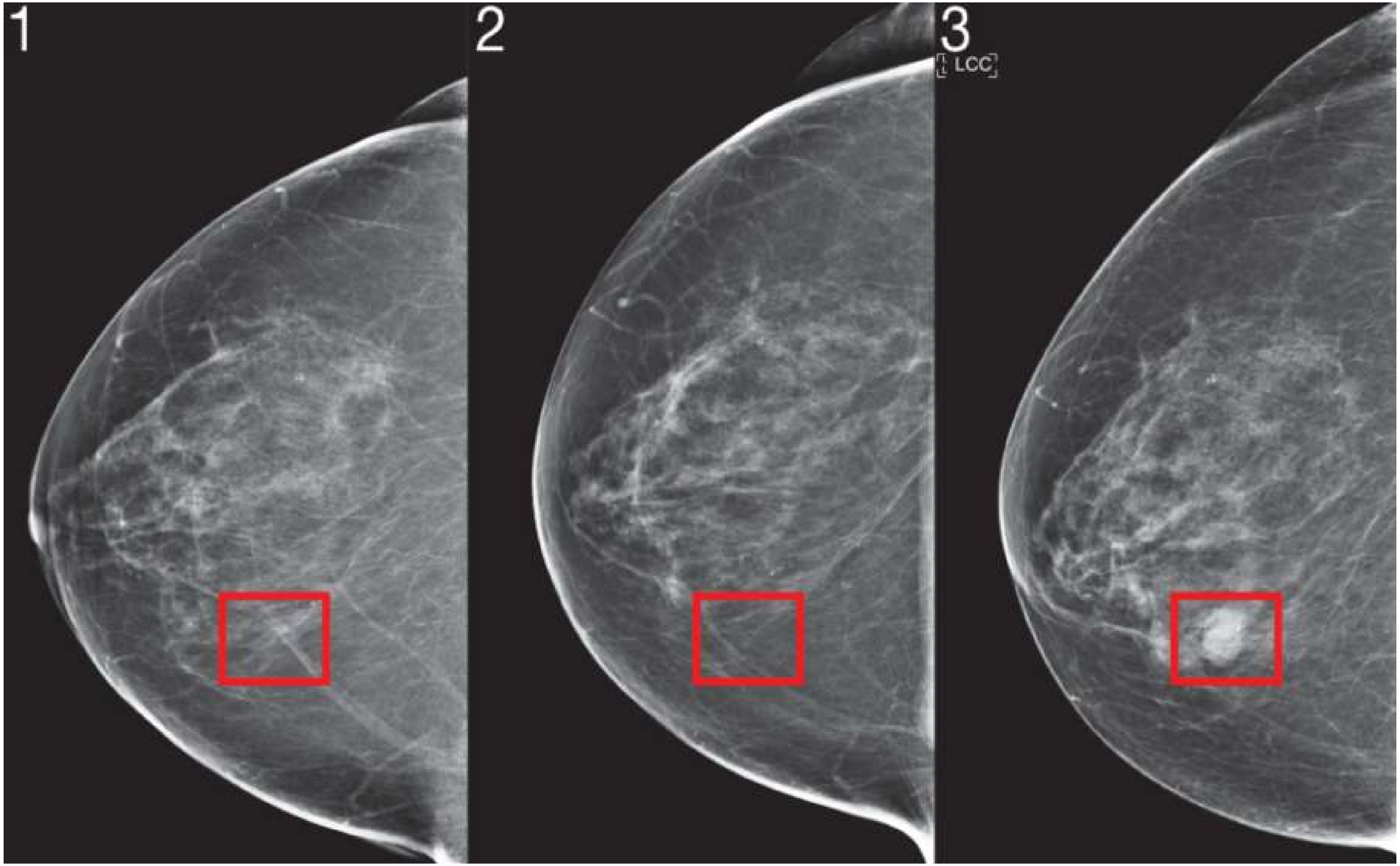
- 1.Machine learning (ML) applications
- 1.2What constitutes an ML algorithm?
- 1.3Supervised learning & unsupervised learning



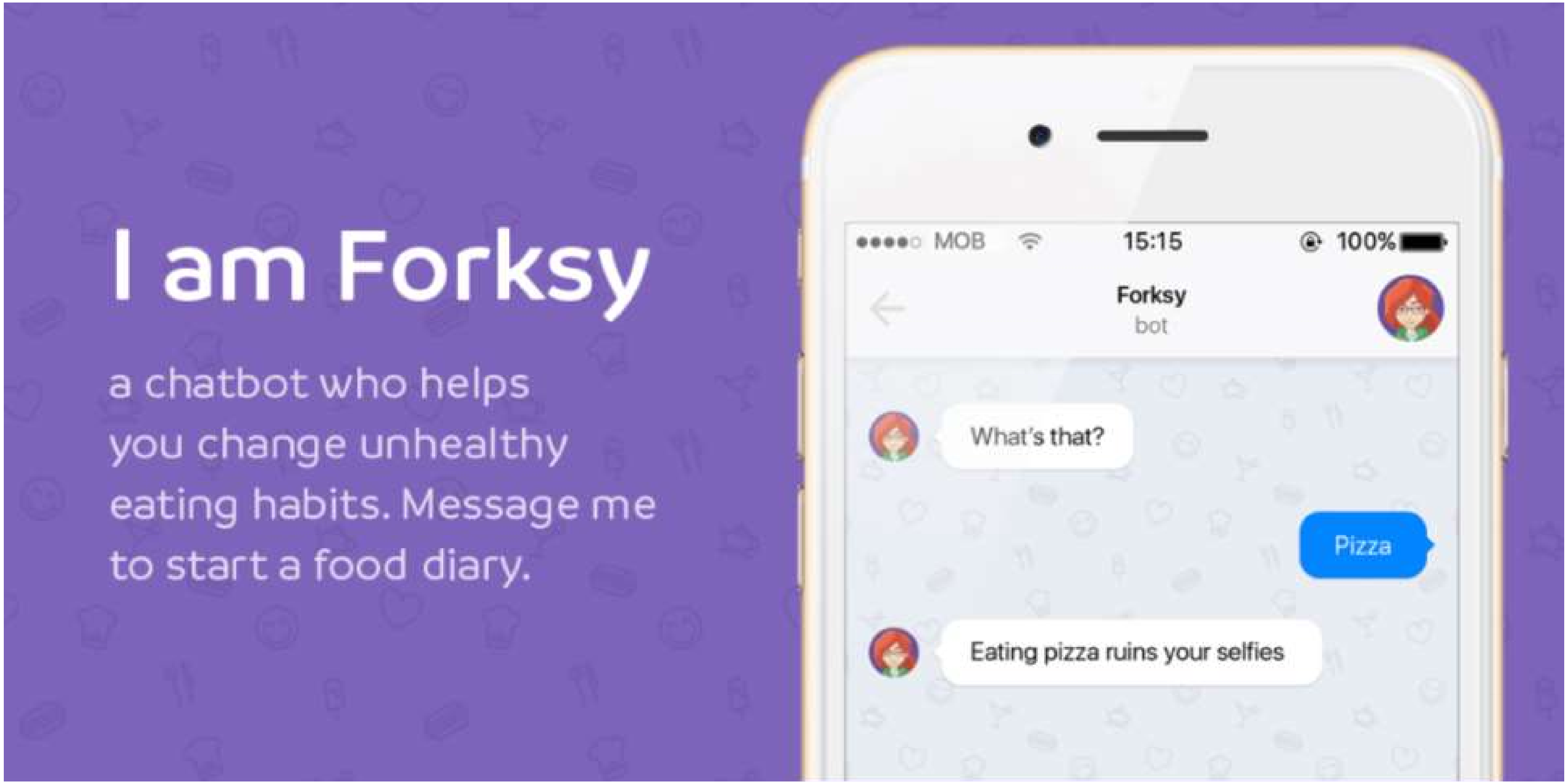
what are in this pic

Wikipedia

WWW.pexels.com



www.wired.com



I am Forksy

a chatbot who helps you change unhealthy eating habits. Message me to start a food diary.

●●●●○ MOB 15:15 100%

Forksy
bot

What's that?

Pizza

Eating pizza ruins your selfies

Forksy's Facebook page

What is this course about?

- Machine learning
  - Image/video analysis

- Why is this hard for a computer?
  - Huge changes in the input
  - We cannot define a general set of rules for each identification
  - The forms of output vary considerably
  - Nature/human language is very flexible

What is this course about?

- Machine learning

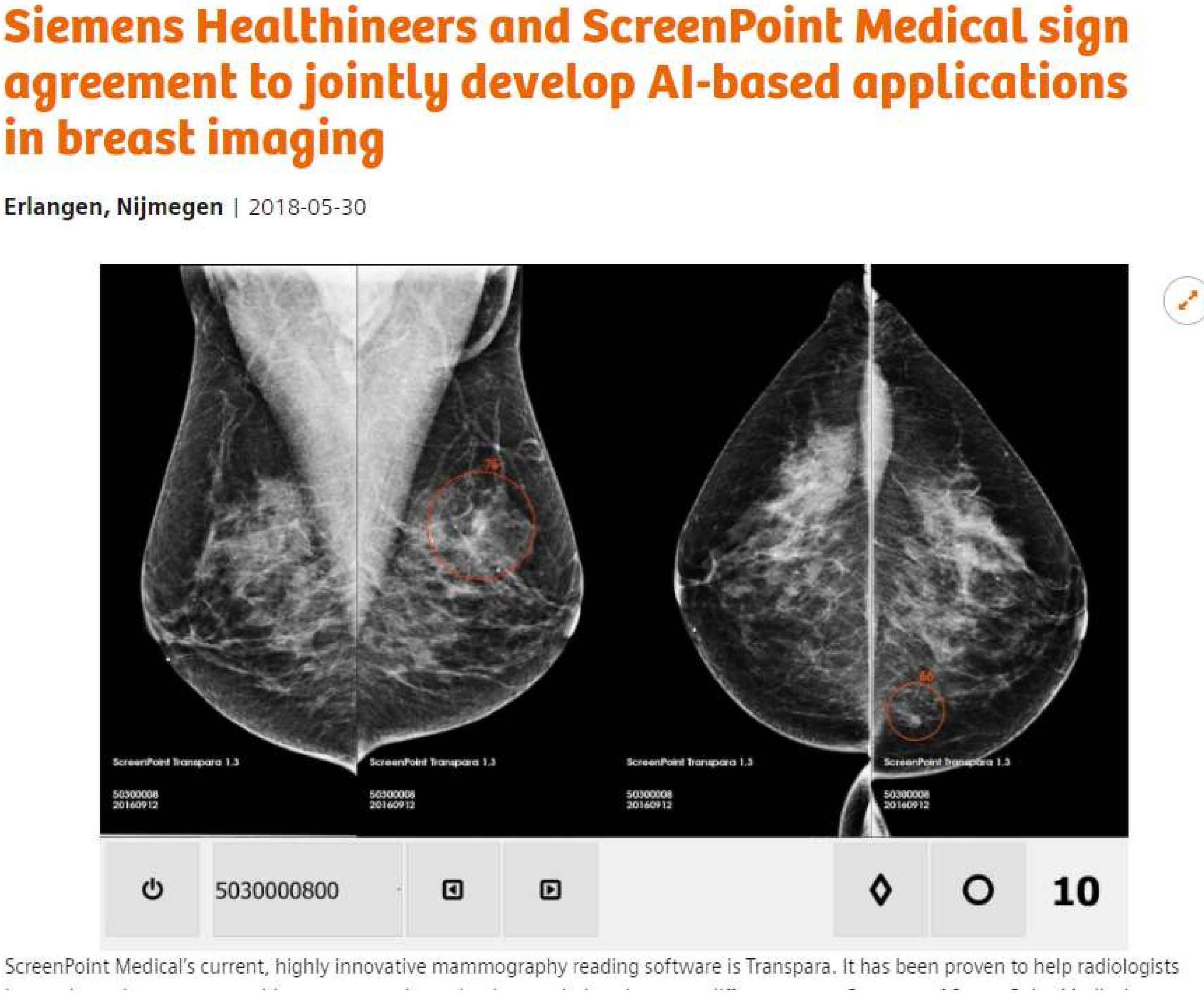
- Deep learning
  - Convolutional Neural Network (CNN)

- Computer vision
  - Image processing
  - Time-sequence processing
  - NLP…

Medical applications

乳腺癌的钼靶精准检测

- **先进的算法**：深度学习+乳腺癌独有计算机检测算法(i.e., 量子噪声抑制)，识别块状和微小钙化点乳腺癌
- **庞大的数据库**：收集整理荷兰以及英国**二十年体检数据**
- **灵活的读片方式**：自主开发移动工作平台
- **科研落地**：在ScreenPoint 落地：获得CE、FDA认证，全球60家临床机构使用, 同时融入西门子工作站



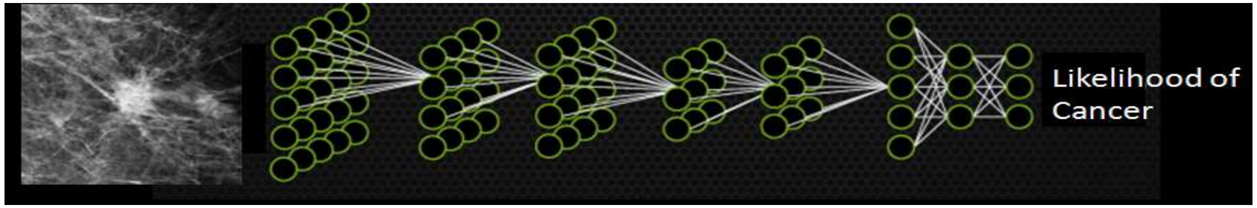
Siemens Healthineers and ScreenPoint Medical sign agreement to jointly develop AI-based applications in breast imaging

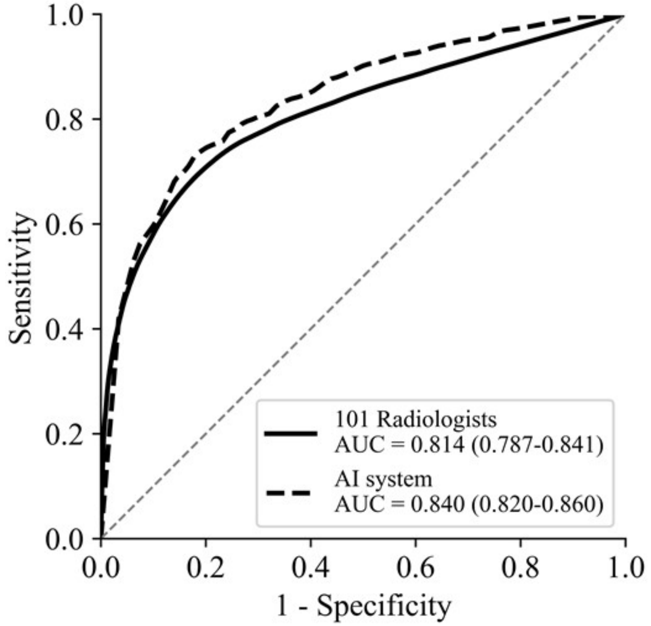
Erlangen, Nijmegen | 2018-05-30

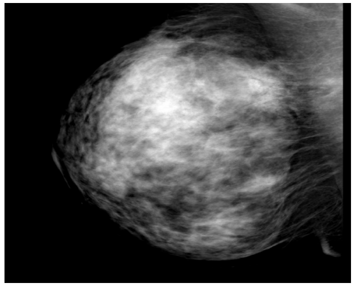
ScreenPoint Medical's current, highly innovative mammography reading software is Transpara. It has been proven to help radiologists

乳腺癌的钼靶精准检测

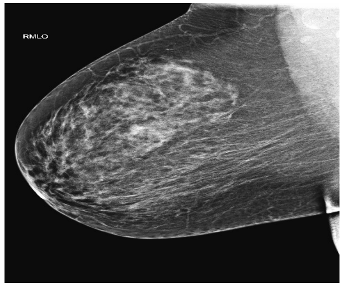
- 计算机水平（灵敏度和特异度）**接近顶尖放射科医生水平**
- 创新的图像归一化算法，消除几百万张不同厂商图像的差异性，解决了**大数据深度学习使用问题**
- **横向比较**，和行业金标准美国**Hologic 的R2系统**相比，**癌症检出率高出8%**



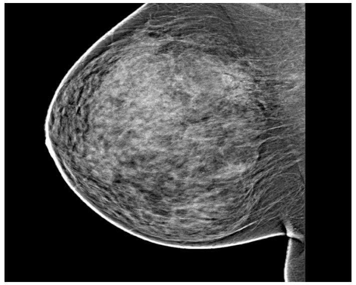




归一化前



参照图像



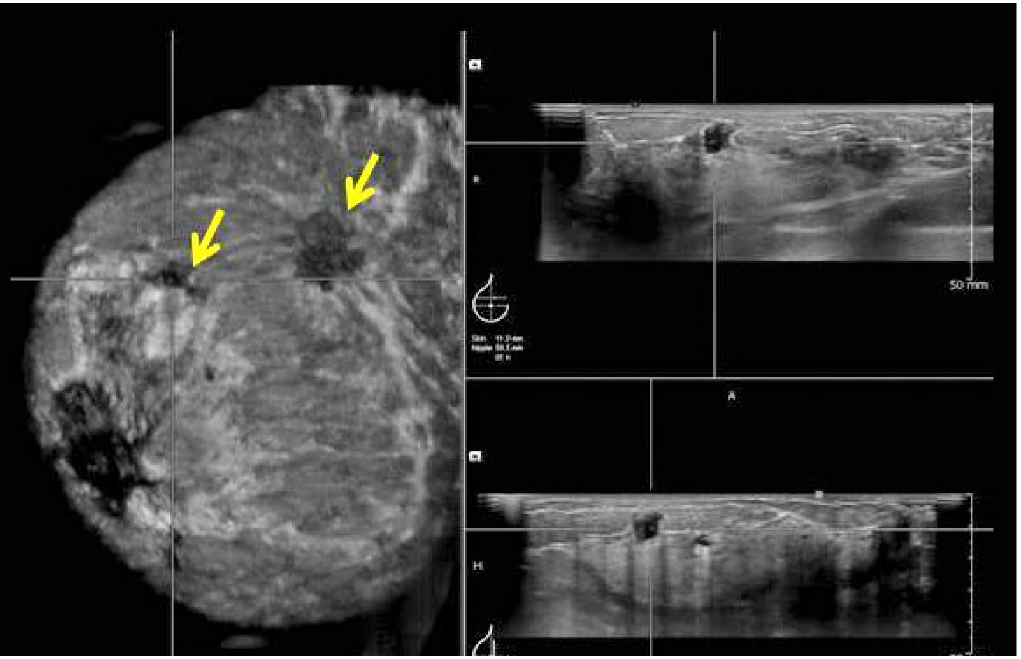
归一化后

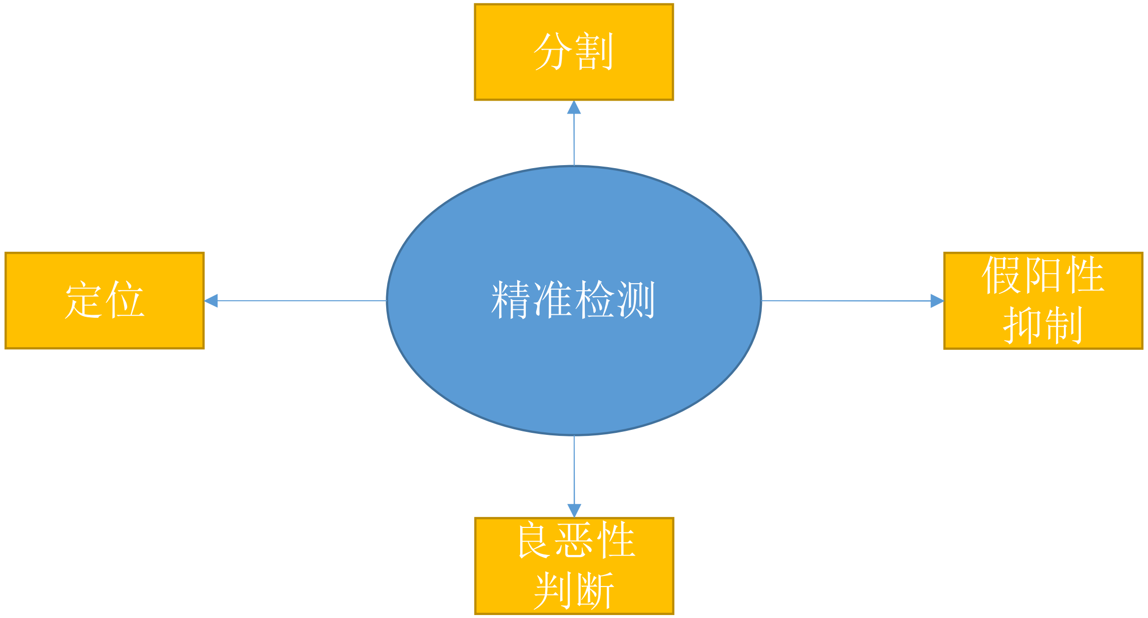
中科院一区 JNCI (IF=11.4)发表

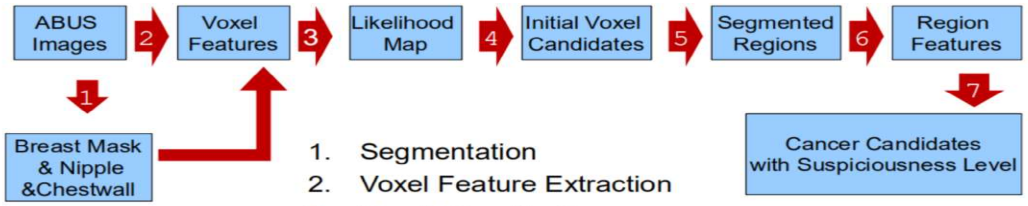
乳腺癌在三维乳腺超声精准检测

现有乳腺筛查痛点

1. 乳腺X射线不适合亚洲女性的致密性乳房
2. 可替代的三维乳腺超声数据量大，无法快速读片，容易漏诊



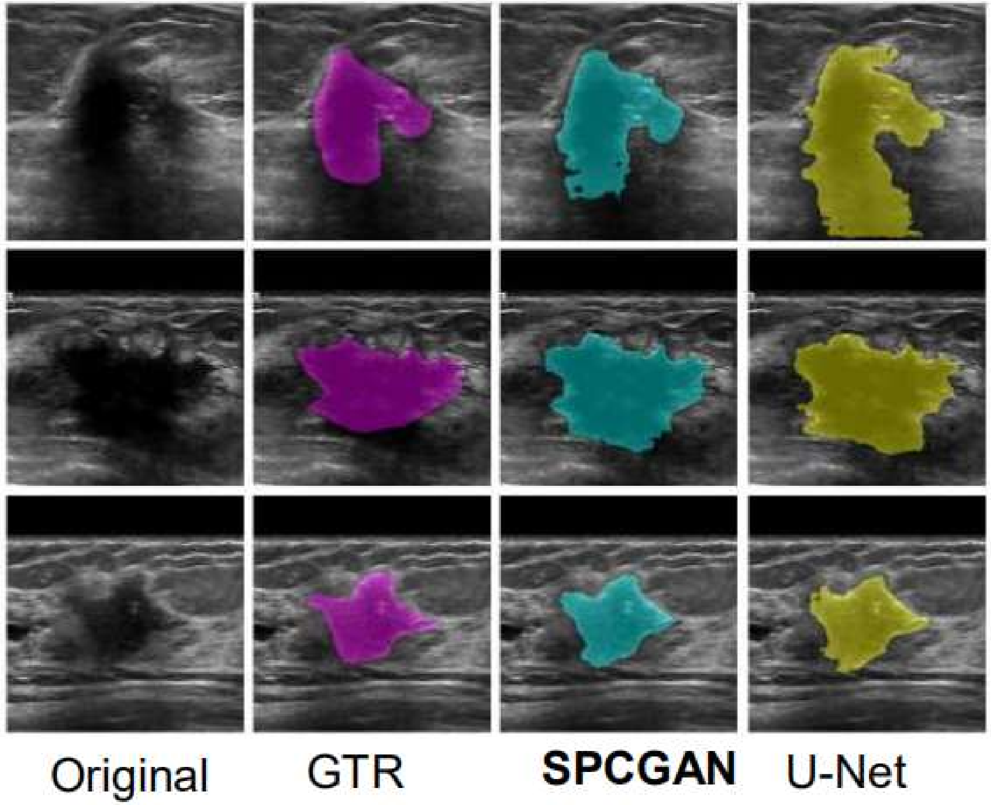


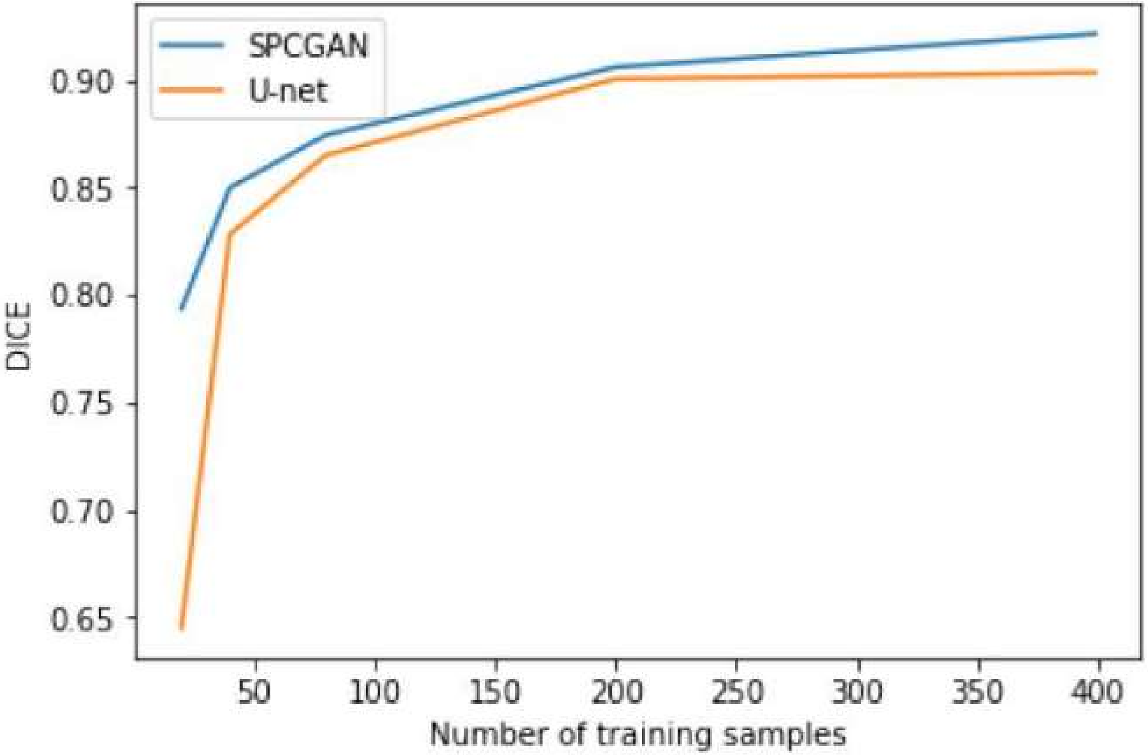


1. Segmentation
2. Voxel Feature Extraction
3. Voxel Classification
4. Local Maxima Detection
5. Region Segmentation
6. Region Feature Extraction
7. Region Classification

肿瘤分割-融入先验知识

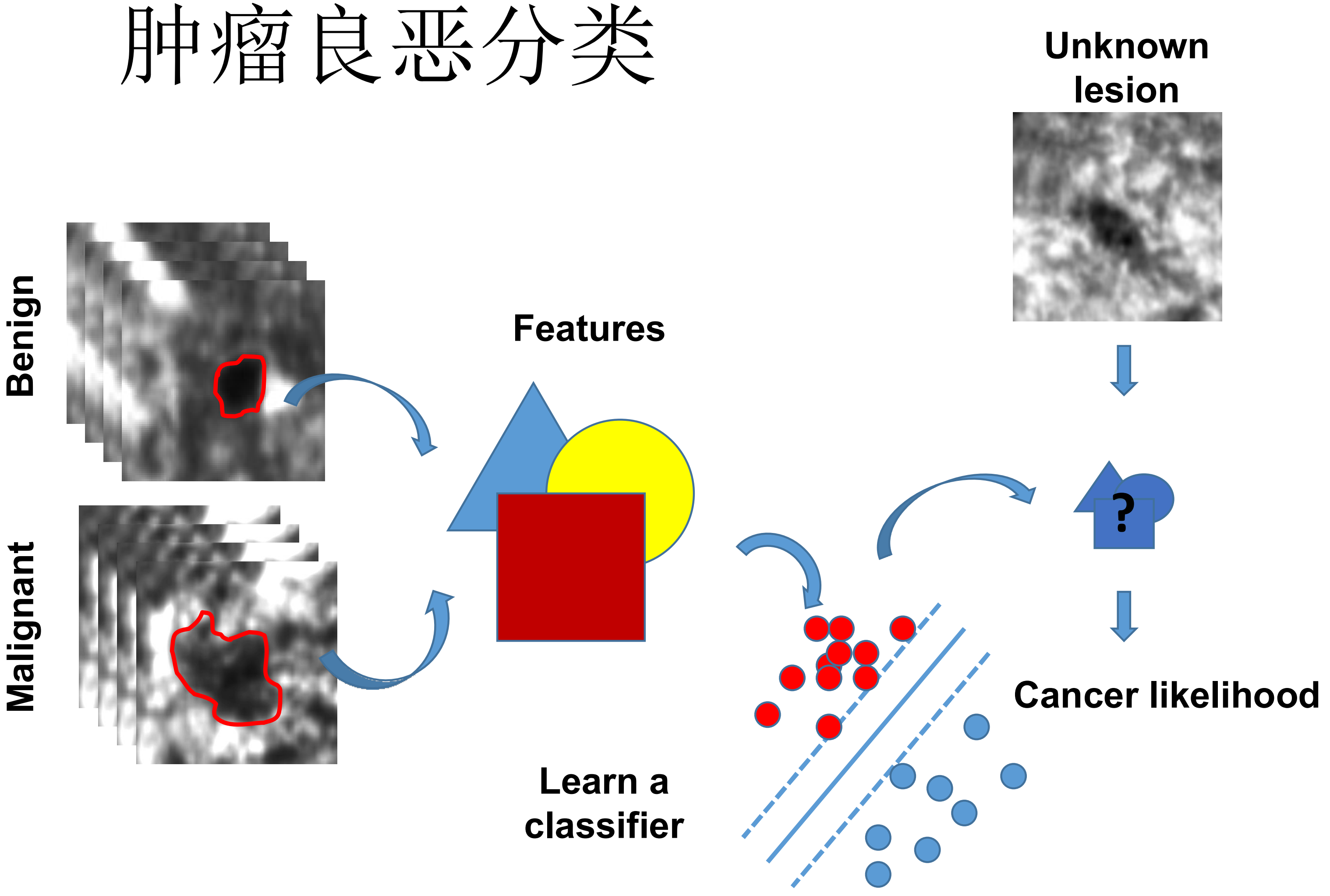
神经网络融入先验知识，更适应于少量数据训练， 优于行业标杆U-Net

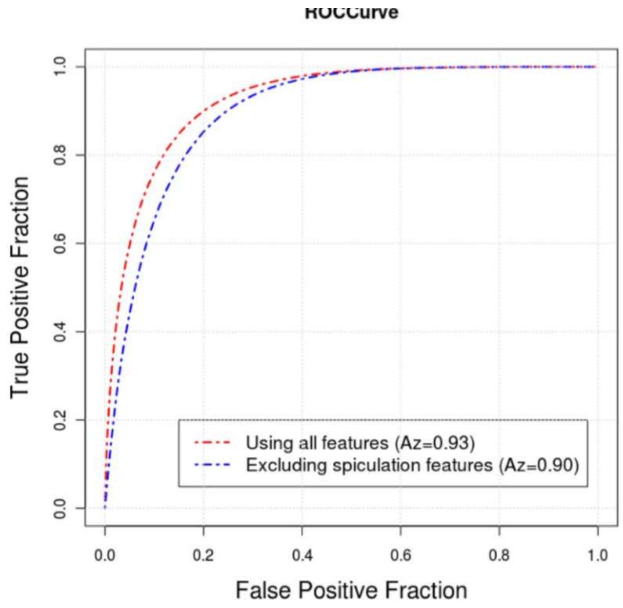


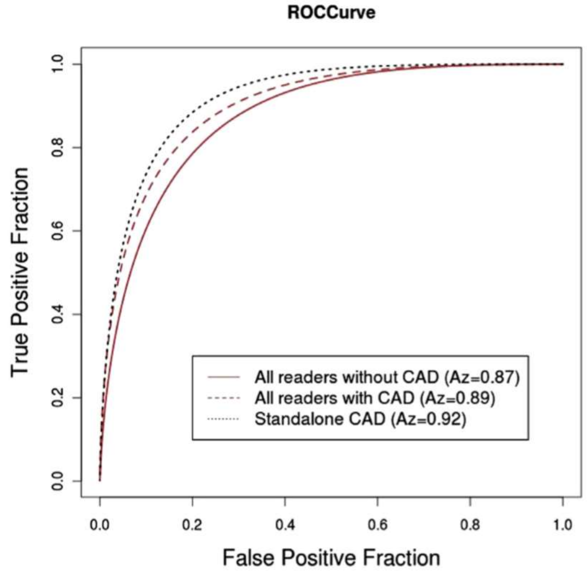


J.Xing et al. ***T. Tan et al.** IEEE/ACM Transactions on Computational Biology and Bioinformatics 2020

肿瘤良恶分类







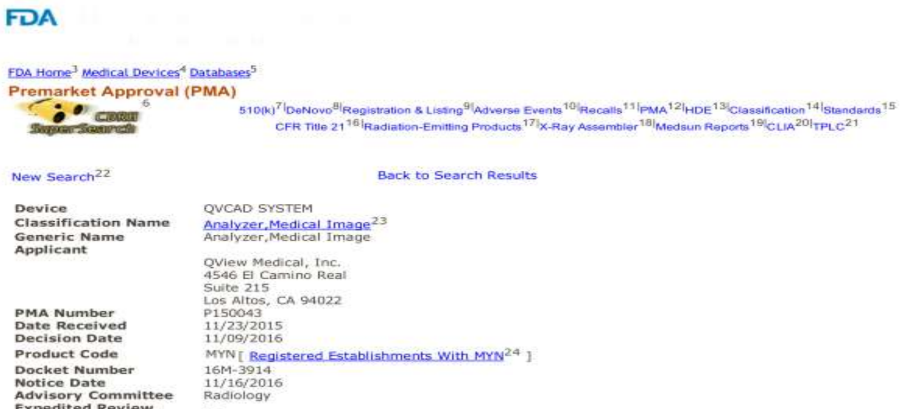
AI 远超医生水平

T. Tan et al. TMI(IF=9.7), 2012

T. Tan et al, Academic radiology

乳腺癌在三维乳腺超声精准检测

- 实现在硅谷科研成果转化
- 芝加哥大学FDA临床试验，荷兰CE临床试验证明由于导航辅助，**节约读片时间30%，提高癌症检出率10%以上，解决漏诊问题**
- **首个也是唯一的**美国食品药监局FDA批准的三维乳腺超声计算机辅助系统



FDA

FDA Home[3] Medical Devices[4] Databases[5]

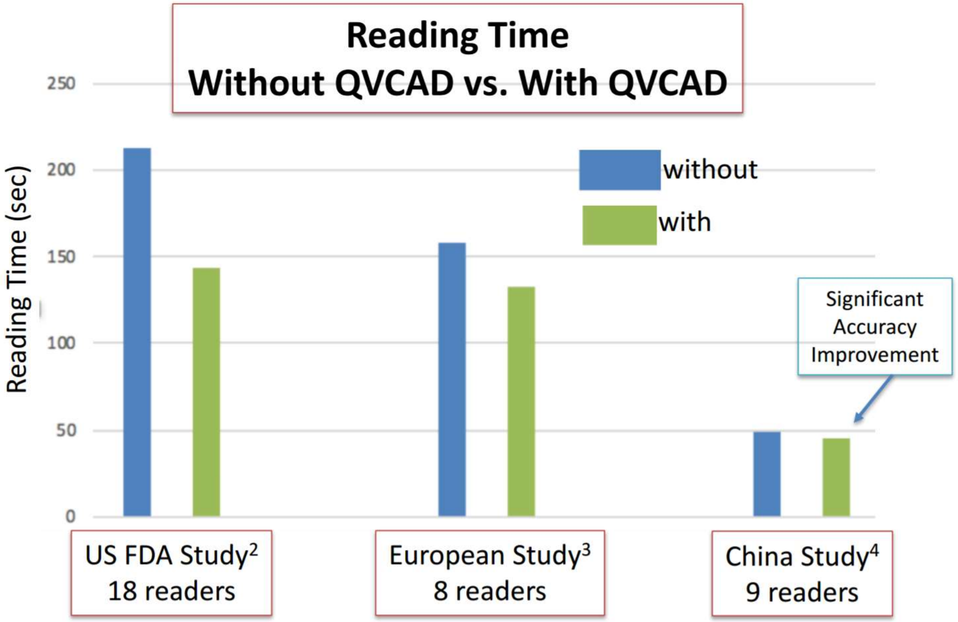
Premarket Approval (PMA)

CDRH SuperSearch[6]

510(k)[7]|DeNovo[8]|Registration & Listing[9]|Adverse Events[10]|Recalls[11]|PMA[12]|HDE[13]|Classification[14]|Standards[15]
CFR Title 21[16]|Radiation-Emitting Products[17]|X-Ray Assembler[18]|Medsun Reports[19]|CLIA[20]|TPLC[21]

New Search[22]   Back to Search Results

Device	QVCAD SYSTEM
Classification Name	Analyzer,Medical Image[23]
Generic Name	Analyzer,Medical Image
Applicant	QView Medical, Inc. 4546 El Camino Real Suite 215 Los Altos, CA 94022
PMA Number	P150043
Date Received	11/23/2015
Decision Date	11/09/2016
Product Code	MYN [Registered Establishments With MYN[24]]
Docket Number	16M-3914
Notice Date	11/16/2016
Advisory Committee	Radiology
Expedited Review	



Reading Time
Without QVCAD vs. With QVCAD

Significant Accuracy Improvement

Personal Industry Experience

ScreenPoint Medical

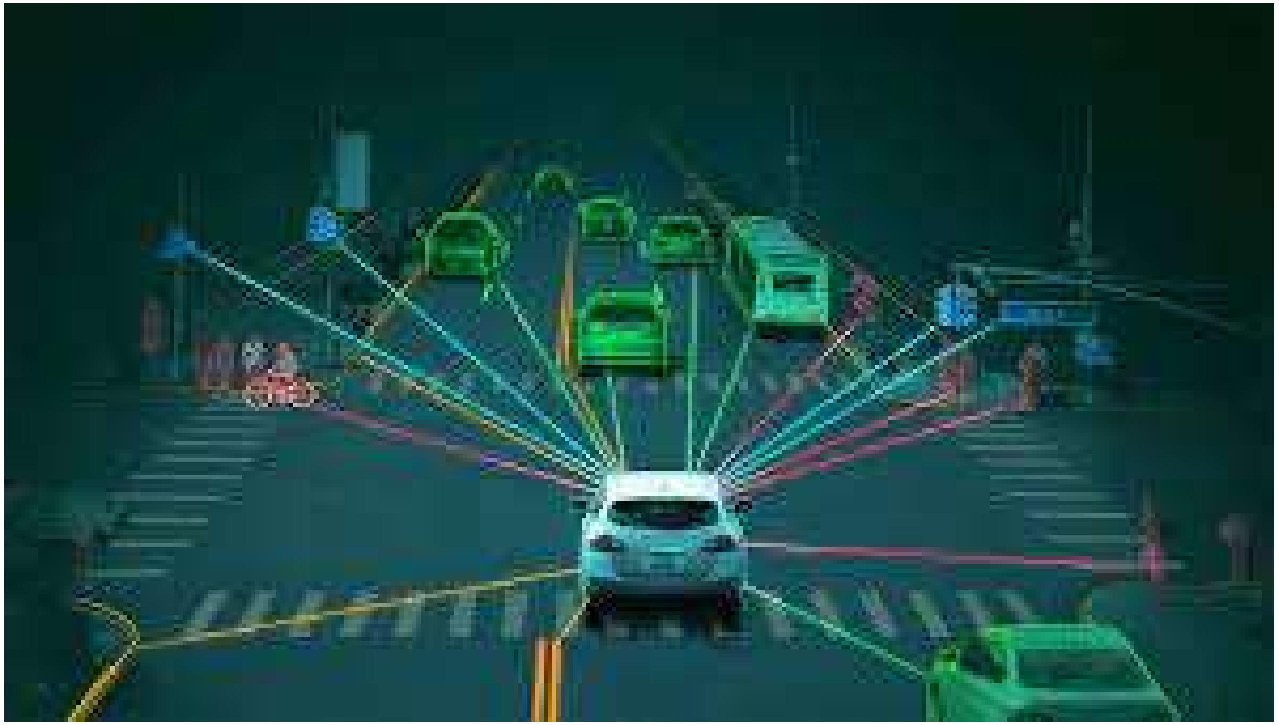
QView MEDICAL

General Electric

Spam filtering

- With a tiny investment, a spammer can send over 100,000 bulk emails per hour.
- Content based filters
    - ➢ Rule based (blacklists)
    - ➢ Bayesian filters
- Cost of false positive/false negative

Automated driving

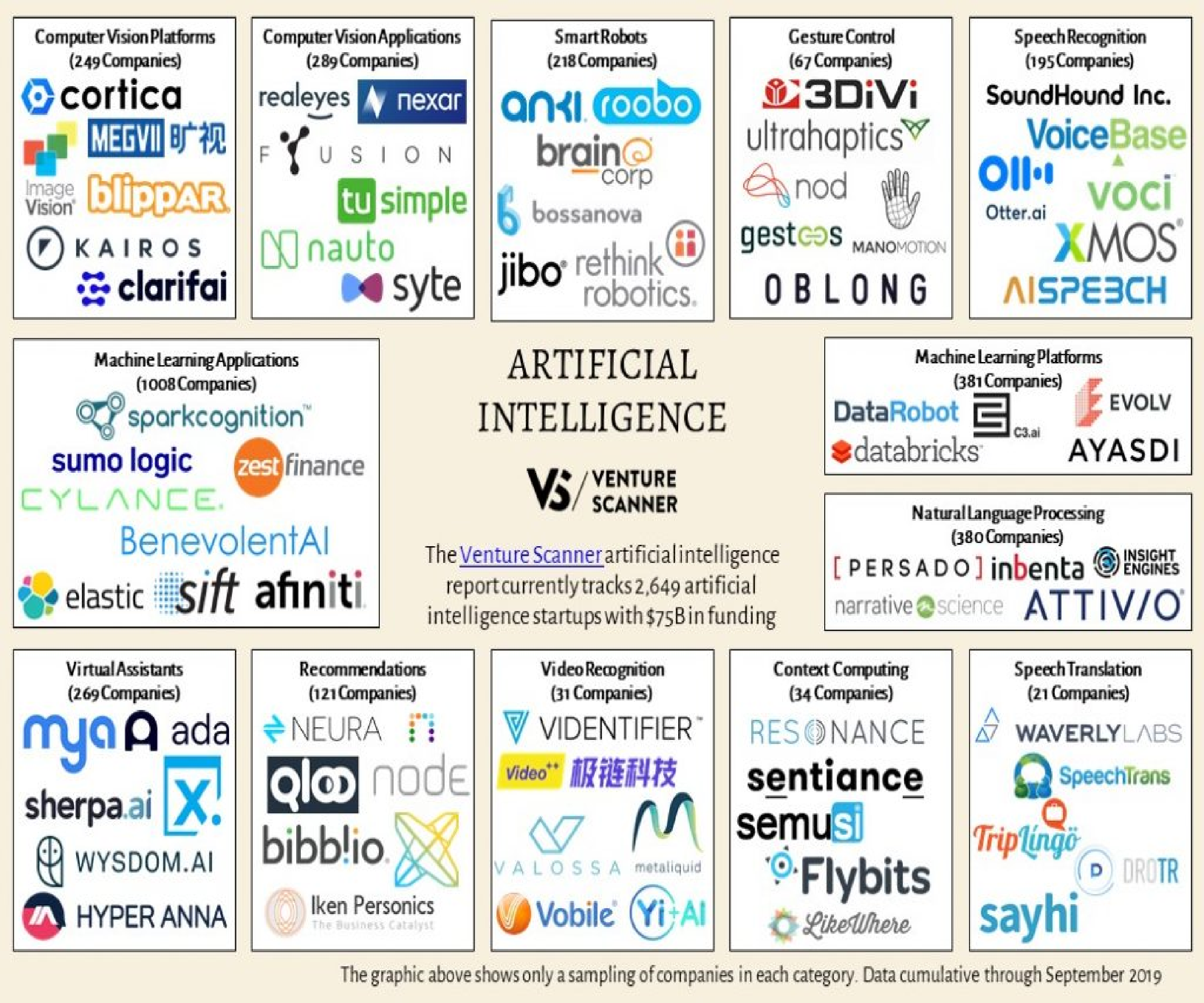


it is legal for autonomous cars to drive on roads in June 2011 in Nevada

https://researchleap.com/research-in-autonomous-driving-a-historic-bibliometric-view-of-the-research-development-in-autonomous-driving/

What can you do with it?

- Amount of data is exploding
  - 22.5 MRI exams per day
  - 8 hours of video uploaded to YouTube every second
  - 2.5 quintillion bytes of data created each day



The Venture Scanner artificial intelligence report currently tracks 2,649 artificial intelligence startups with $75B in funding

The graphic above shows only a sampling of companies in each category. Data cumulative through September 2019

Web search

Computational biology

Finance

Space exploration

Healthcare

Robotics

Information extraction

Social networks

Debugging software

E commerce

www.venturescanner.com/2019

How hard is this course

- What was the difficulty level
  - 30% fundamental and 70% applied
  - 50% conceptual and 50% concrete

What you should know

- Basics
  - Calculus
  - Linear algebra
  - Probability theory
  - Some programming

+ Signal processing experience

1D, 2D, 3D, 4D data

Course details & planning

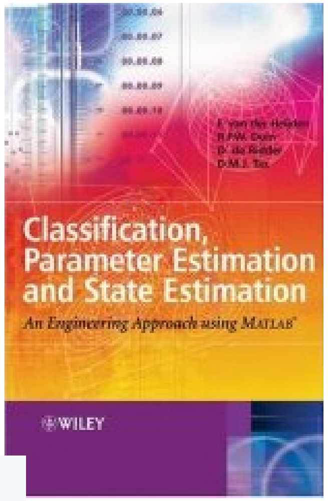
- Lectures
- Effectively 15 weeks
- Lectures on Tuesday 10am - 1pm

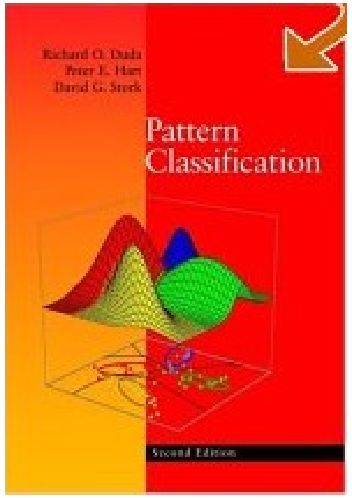
- Important dates
- Test: week 8
- Final project presentation: week 15

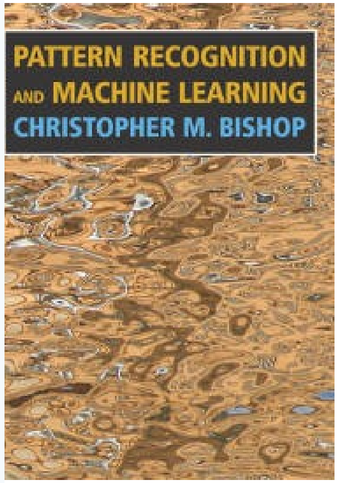
Course details & planning

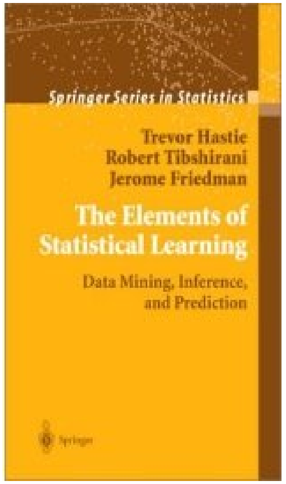
- Slides are sufficient to study for the exam

- Optional book on machine learning fundamentals


F. van der Heijden
R.P.W. Duin
D. de Ridder
D.M.J. Tax
Classification, Parameter Estimation and State Estimation
An Engineering Approach using MATLAB
WILEY






Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Springer

- Optional book on deep learning:
  - Goodfellow, Bengio & Courville, “Deep Learning” (2015)
  - Free online! (https://www.deeplearningbook.org/)

Chris Bishop

FRS FRSE FREng

Chris Bishop in 2015

Born	Christopher Michael Bishop 7 April 1959 (age 63)[1] Norwich[1]

Machine learning

- Two definitions of Machine Learning are offered.
  - Arthur Samuel described it as: “the field of study that gives computers the ability to learn without being explicitly programmed.” This is an older, informal definition.
  - Tom Mitchell provides a more modern definition: “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.”

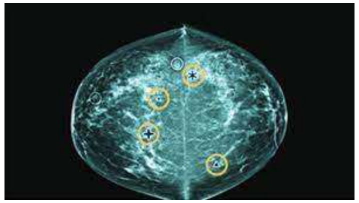
Machine learning

- Imagine you have some sets of the pair of numbers. Then you put only 1 number of the pair into a machine to predict the other half of the pair.
- (2,4),(3,6),(4,9) . The computer program has to predict the second number for (5,?)
- The program first needs to find the logic between the pairs and then apply the same logic to predict the number. To find that logic is called “machine learning”. So that after finding the logic it can apply the same logic to predict each number.

What is machine learning

Machine learning is sued when

- Human performance varies(CAD)



- We want reduce spent time (workflow AI, face recognition)

- Humans can not perform (deep ocean project)

- Personalized model (oncology outcome)

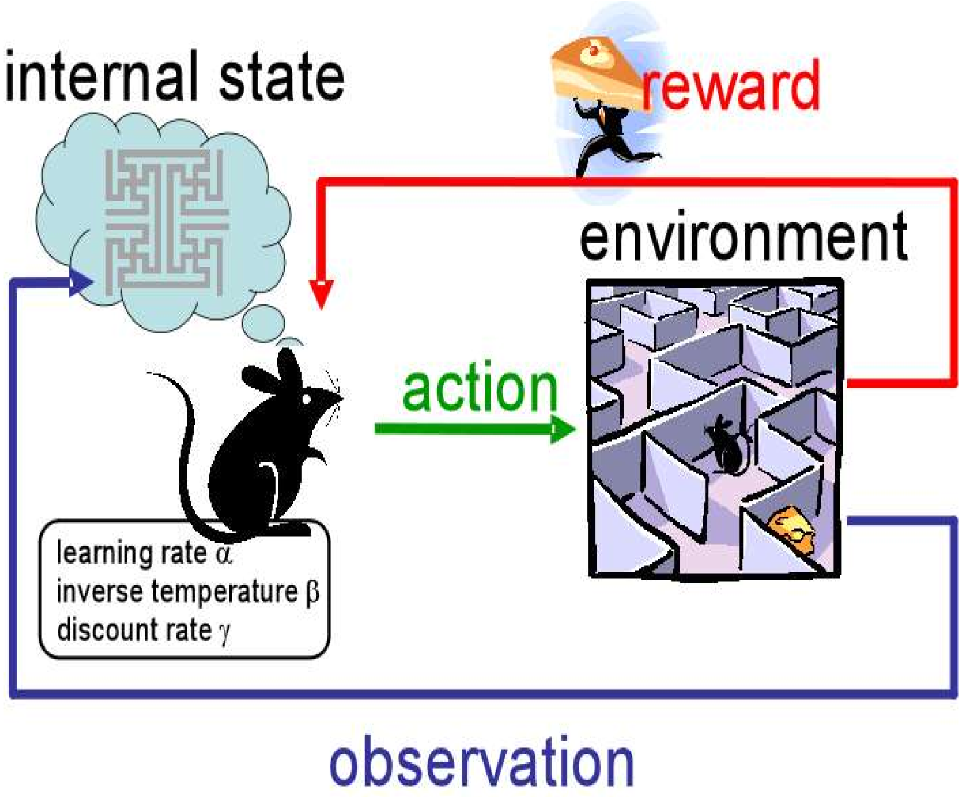
- Huge amounts of data (genomics)

What is machine learning

- Build general models from a data of examples
- Build a model that is a good and useful approximation to the data
- Accuracy vs robustness

Machine learning?

- Supervised learning
- Unsupervised learning
- Reinforcement learning



https://becominghuman.ai/the-very-basics-of-reinforcement-learning-154f28a79071?gi=70c715fb76ce

Supervised learning

- In supervised learning, we are given a data set and already know what our correct output should look like, having the idea that there is a relationship between the input and the output.
- Supervised learning problems are categorized into “regression” and “classification” problems. In a regression problem, we are trying to predict results within a continuous output, meaning that we are trying to map input variables to some continuous function. In a classification problem, we are instead trying to predict results in a discrete output. In other words, we are trying to map input variables into discrete categories.

A practical example

Housing price

- Faced during buying a house
- Buying a house is stressful thing.
- Buyers are generally not aware of factors that influence the house prices.
- Many problems are faced during buying a house.
- Hence real estate agents are trusted with the communication between buyers and sellers as well as laying down a legal contact for the transfer. This just create a middle man and increases the cost of houses

Housing price

Common factors

- City
- Area
- Neighbourhood
- No. of bedrooms
- Apartment or house
- Area outside of the house
- Which floor

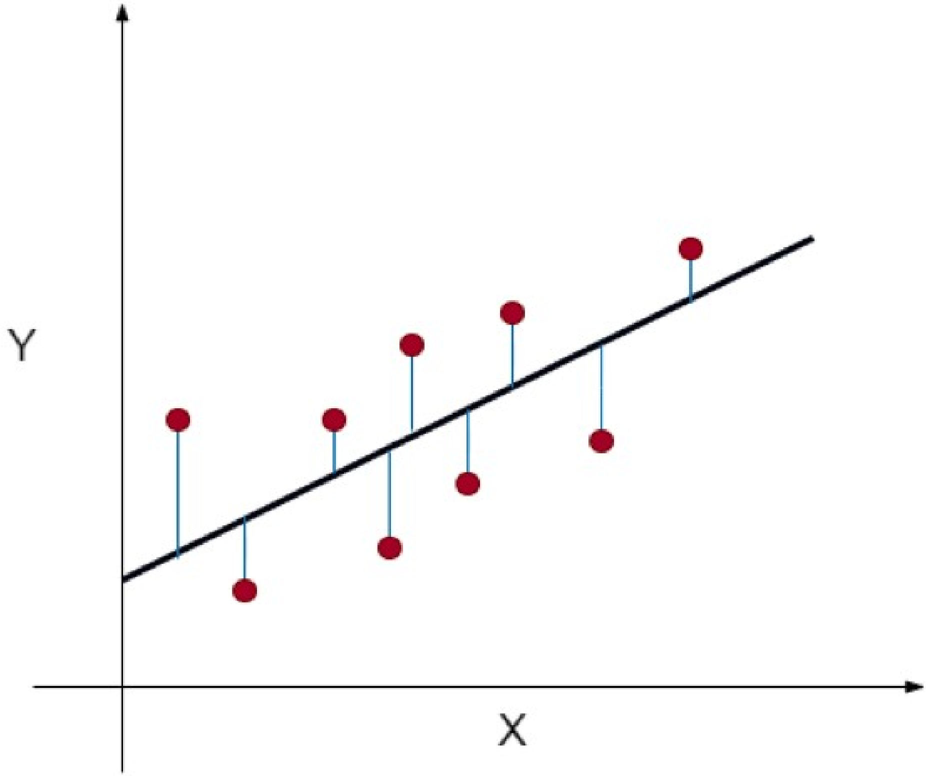
Housing price

- We want to leverage machine learning app, based on specifications of future home and try to guess the most accurate prices.

- Information such as city, area, shops.

Housing price

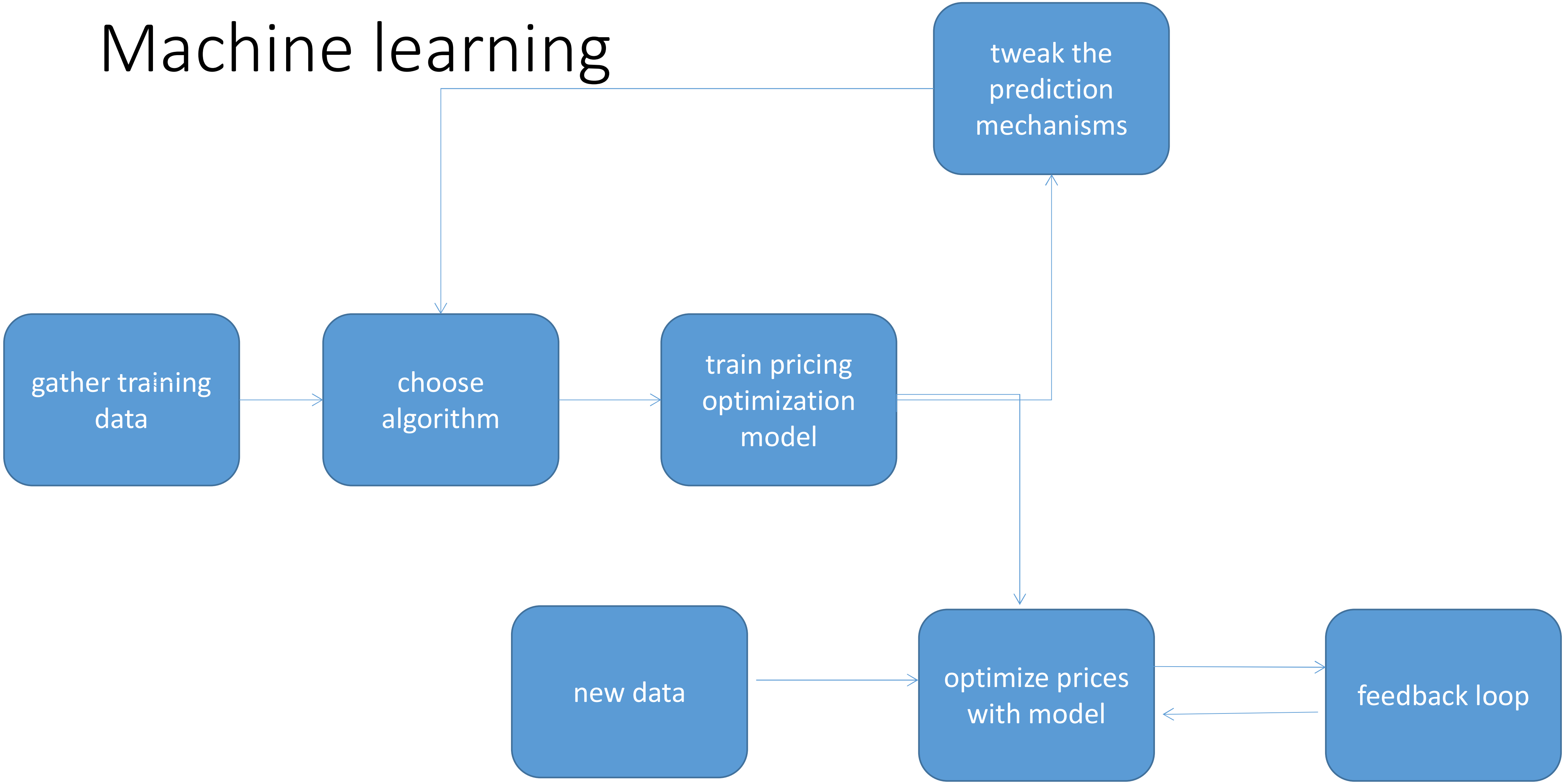
- Price as a function of size is a continuous output, so this is a regression problem.
- We could turn this example into a classification problem by instead making our output about whether the house “sells for more or less than the asking price.” Here we are classifying the houses based on price into two discrete categories.



Regression vs Classification

- (a) Regression — Given a picture of a person, we have to predict their age on the basis of the given picture
- (b) Classification — Given a patient with a tumor, we have to predict whether the tumor is malignant or benign.

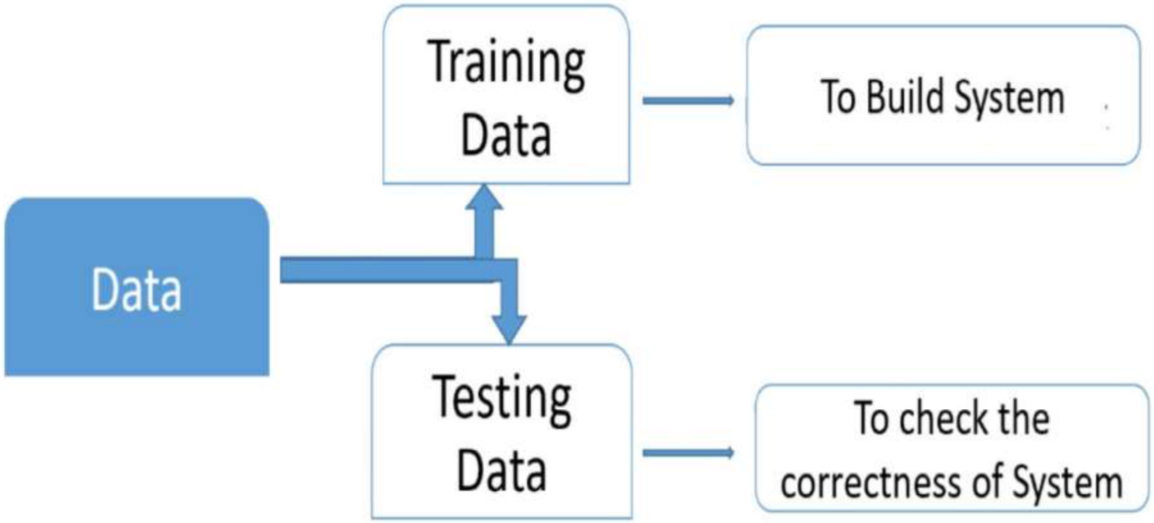
Machine learning



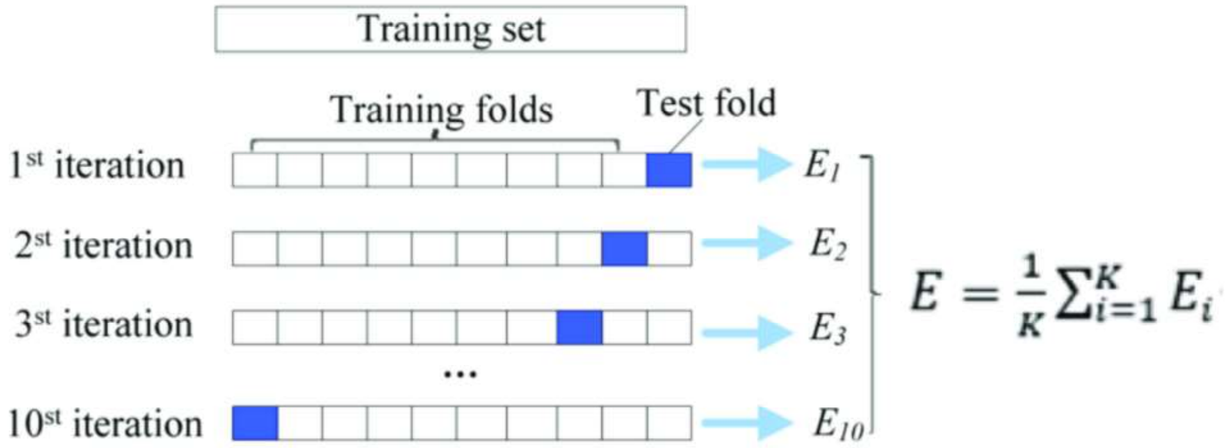
tweak the prediction mechanisms
gather training data
choose algorithm
train pricing optimization model
new data
optimize prices with model
feedback loop

How to perform evaluation

- Accuracy
- Sensitivity and specificity
- Confusion matrix
- MSE
- AUC
- F1 score



Cross validation



Ten-fold cross validation diagram. The dataset was divided into ten parts, and nine of them were taken as training data in turn, and one was used as test data for testing. The average value E of the ten-groups test results is calculated as an estimate of the model accuracy and is used as a performance indicator for the current K-fold cross-validation model. Where E i represents the cross-validation error of the ith group.3.4. The RFAmyloid Online Prediction Server.

https://www.researchgate.net/figure/Ten-fold-cross-validation-diagram-The-dataset-was-divided-into-ten-parts-and-nine-of_fig1_326465007

Machine learning components

- Data digitization
- Data normalization
  - Minmax
  - Zscore
- Modeling
- Optimization
- Evaluation

Break

History of Machine Learning

- https://roboticsbiz.com/machine-learning-the-complete-history-in-a-timeline/

- https://www.techtarget.com/whatis/A-Timeline-of-Machine-Learning-History

Resources: coding

- w3 schools: https://www.w3schools.com/python/default.asp

- NumPy

NumPy is a popular Python library for multi-dimensional array and matrix processing because it can be used to perform a great variety of mathematical operations.

Resources: coding

- Scikit-learn

Scikit-learn is a very popular machine learning library that is built on NumPy and SciPy. It supports most of the classic supervised and unsupervised learning algorithms, and it can also be used for data mining, modeling, and analysis. Scikit-learn’s simple design offers a user-friendly library for those new to machine learning.

Resources: coding

Pandas

- Pandas is another Python library that is built on top of NumPy, responsible for preparing high-level data sets for machine learning and training.

Resources: data

- Kaggle challenge: https://www.kaggle.com/

- Grand challenge: https://grand-challenge.org/

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

Resource: computation power

- https://colab.research.google.com/

- https://www.youtube.com/watch?v=RLYoEyIHL6A

Academic resources

conferences

International Conference on Computer Vision (ICCV）

Conference on Computer Vision and Pattern Recognition (CVPR)

International Conference on Machine Learning (ICML)

Neural Information Processing Systems (NIPS)

International Conference on Pattern Recognition (ICPR)

Journals

International Journal of Computer Vision

IEEE Trans on Medcial Imaging

Neural Networks

IEEE Trans on Neural Networks and Learning Systems

IEEE Trans on Pattern Analysis and Machine

What we will cover in this course

- See COMP407-2223-1-yyy.docx