《机器学习及其在化学中的应用》2025年课程

Sec. 1

机器学习引言

刘志荣（LiuZhiRong@pku.edu.cn)

北京大学化学学院

2025.9.8

身边的机器学习.....

一个新闻热点：ChatGPT

- ChatGPT：聊天机器人程序，能够通过理解和学习人类的语言来进行对话，甚至能完成翻译、文案、代码、撰写邮件、视频脚本、写论文等任务。

写代码、写情书、做题、写文章……
它啥都会！

一个新闻热点：换脸

迪士尼开发换脸技术，
用**deepfake**拍电影可以安排了？

技术源头：deepfakes（2017年）

另一个热门话题……

美国政府考虑封杀TikTok（2020年）

https://www.huxiu.com/article/376933.html

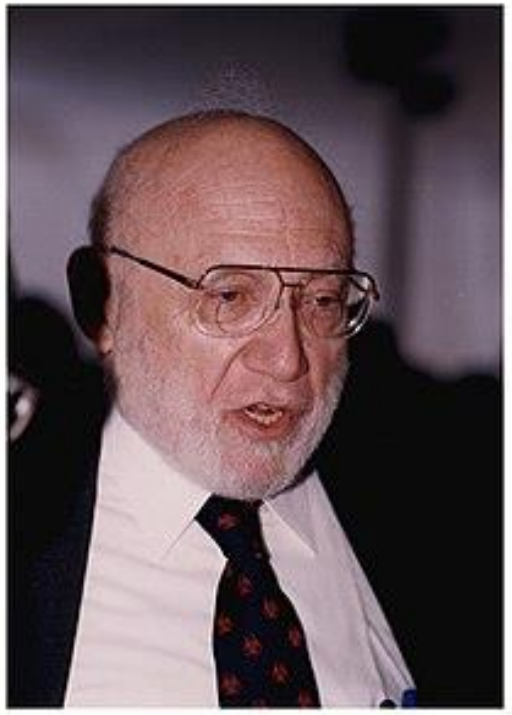
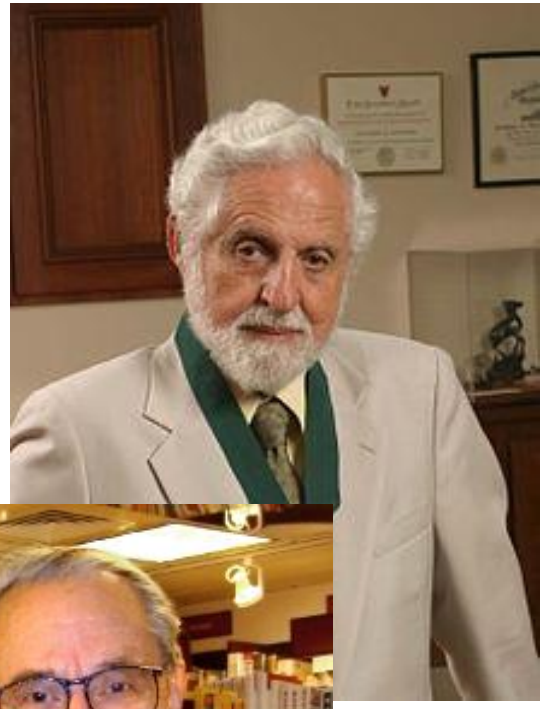
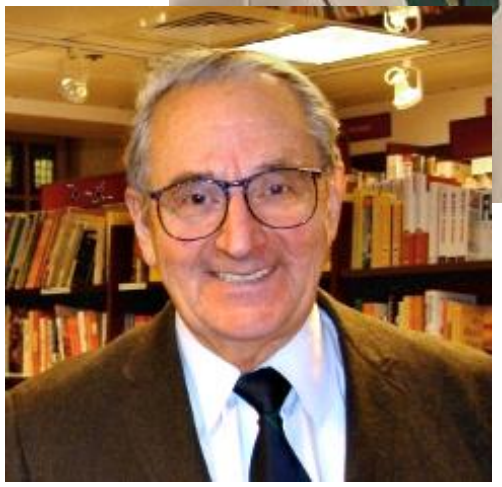
- 为什么TikTok能够横扫美国市场？

算法，能够克服文化壁垒！

在某些行业和产品类别中，基于机器学习的算法有着极高的响应性和准确性，能够穿透“文化无知之幕”。

机器学习在化学中的应用，很久很久以前……

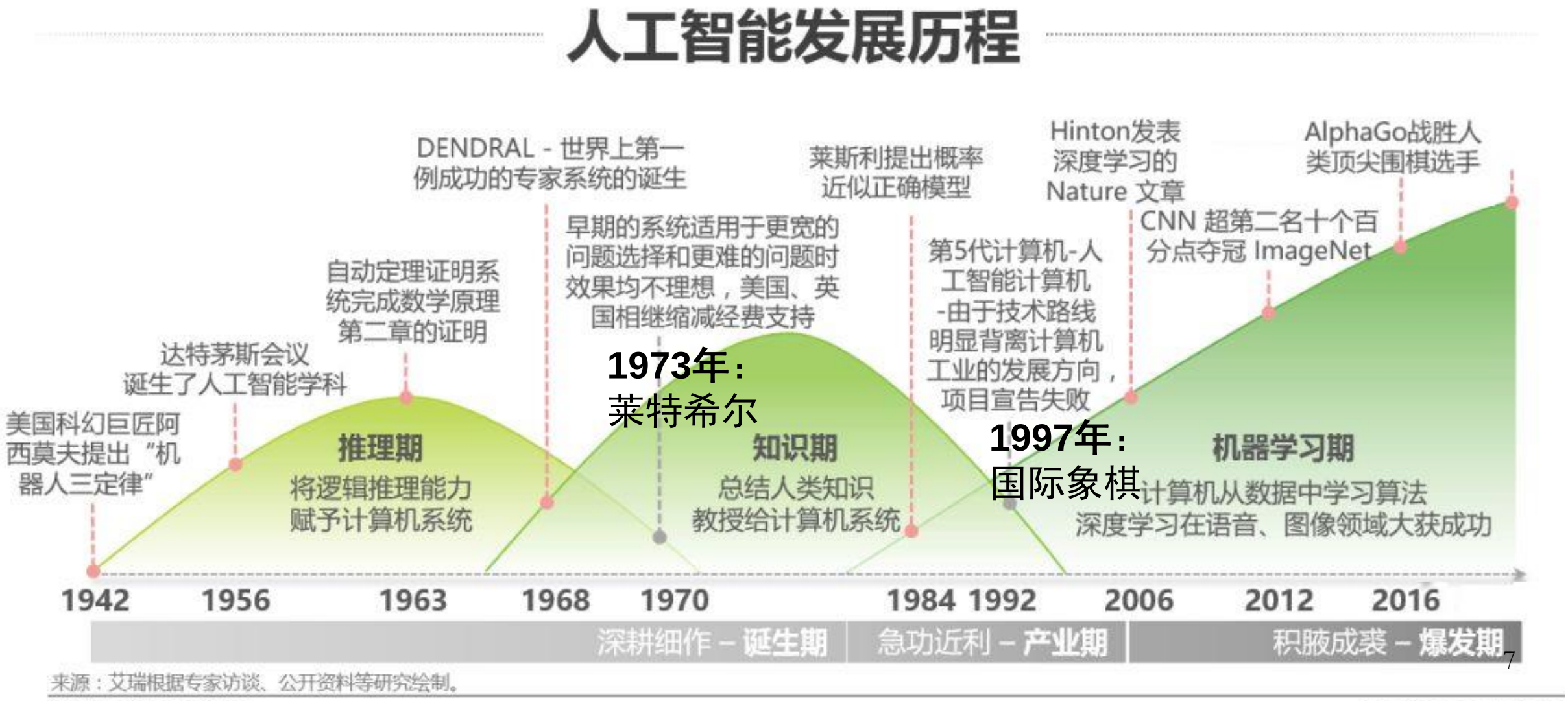
- **Dendral**：从质谱数据识别分子组成（1965-1968）
  - Edward Feigenbaum, Joshua Lederberg, and Carl Djerassi

- **Elias James Corey** 与逆合成分析
  - 1969，OCSS
  - 1976，LHASA

寒冬.....

人们总是高估一项技术所带来的短期效益，
却低估了它的长期影响。



最近几年.....

- 人工智能的崛起！

AlphaGo（阿尔法狗）!!

（2016年）

- Mark P. Waller et al., Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604 (2018).

- DeepMind又推AI杰作：**AlphaFold蛋白结构预测**

击败人类!!!（2018）

课程：

机器学习及其在化学中的应用

Machine Learning and its Applications in Chemistry

学分：2；　周学时：2；　　总学时：32

目的：介绍机器学习的相关基础知识，并分析机器学习在化学前沿领域的典型应用例子，　以便更好地面对机器学习给化学学科所带来的挑战与机会。

课程目的：

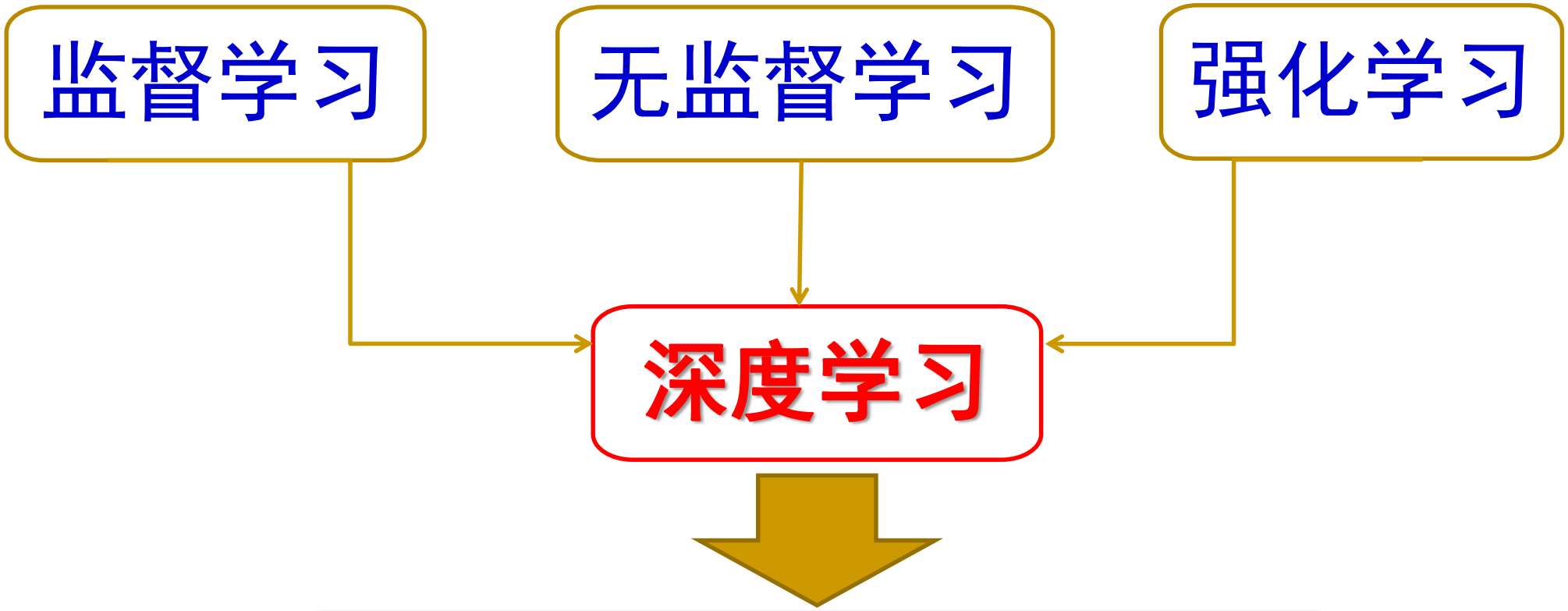
通过课程的学习，希望能够使学生

（1）掌握机器学习的基本知识；

（2）能够阅读、分析与评估相关领域的机器学习文章；

（3）能利用机器学习编写程序解决简单的问题。

课程内容



化学中的应用

重点例子：有机合成路线，AlphaFold，药物设计，力场

参考书：

[1] 刘志荣书稿草稿1-5章。

[2] Christopher M. Bishop, Pattern Recognition and Machine Learning (Springer, 2006). （有中文电子版）

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, The Elements of Statistical Learning: Data Mining, Inference, and prediction (Springer, 2016)

[4] Richard S. Sutton and Andrew G. Barto, Reinforcement Learning: An Introduction. 2nd edition. (MIT Press, 2018).（俞凯 等 译），强化学习（电子工业出版社，2019）

[5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, Deep Learning. (MIT Press, 2016). 伊恩.古德费洛、约书亚.本吉奥、亚伦.库维尔 著，赵申剑 等 译。深度学习（人民邮电出版社，2017）

[6] Peter Harrington 著，李锐，等 译。机器学习实战 [Machine learning in action]（人民邮电出版社，2013）。有代码与数据

网络资源

[1] 吴恩达（Andrew Ng），机器学习

网易云课堂：https://study.163.com/course/introduction/1210076550.htm

黄海广的笔记：https://github.com/fengdu78/Coursera-ML-AndrewNg-Notes

[2] 微软课程：Machine Learning for Beginners

https://github.com/microsoft/ML-For-Beginners

[3] Sklearn 与 TensorFlow 机器学习实用指南

https://hands1ml.apachecn.org/#/

[4] 伯克利课程：Introduction to Artificial Intelligence (人工智能导论)

https://inst.eecs.berkeley.edu/~cs188/fa18/index.html

[5] https://www.jiqizhixin.com/articles/123104

寒冬里的炭火：机器之心2018高分教程合集

课程进度安排

序号	授课内容	学时	参考书章节
1	机器学习引言	1	刘志荣1；Bishop 1.1；实战1；吴恩达1
2	线性回归：最简单的学习模型	5	刘志荣2；Bishop 1-3；实战8；吴恩达2,4,7
3	分类的线性方法：逻辑回归	2	刘志荣3；Bishop 4；实战5；吴恩达6
4	模型评估	2	刘志荣4；Bishop 1.3,1.5；Scikit-learn 3；实战7.7; 吴恩达10,11,15
5	人工神经网络	3	刘志荣5；Bishop 5；Goodfellow 9；吴恩达8,9
6	核方法：直方图、近邻法、高斯过程、支持向量机	3	刘志荣6；Bishop 6；实战2,6；吴恩达12
7	概率图模型：朴素贝叶斯分类器、贝叶斯网络、马尔科夫随机场	2	刘志荣7；Bishop 8；实战4
8	聚类：K-均值法、高斯混合模型	1	刘志荣8；Bishop 9；实战10；吴恩达13
9	降维：主成分分析	2	刘志荣9；Bishop12；实战13；吴恩达13
10	集成学习：Boosting、决策树、随机森林	2	刘志荣10；Bishop14；实战3,7,9
11	强化学习简介：多臂老虎机	1	刘志荣11；Sutton 1,2
12	强化学习的问题描述：有限马尔科夫决策过程	1	刘志荣12；Sutton 3
13	强化学习的三种方法：动态规划、蒙特卡罗、时序差分算法	2	刘志荣13；Sutton 4-6
14	强化学习：AlphaGo	1	刘志荣14；Sutton 8.10-8.11, 16.6
15	深度学习简介	2	刘志荣15； Goodfellow 1, 5, 8, 14, 20

监督学习

无监督学习

强化学习

深度学习

教学方式

- 以课堂讲授为主。（课件及学习资料下载：北大教学网 ）
  course.pku.edu.cn
  学生课后需完成书面及编程作业。（提供机房上机指导）
- 编程语言：Python

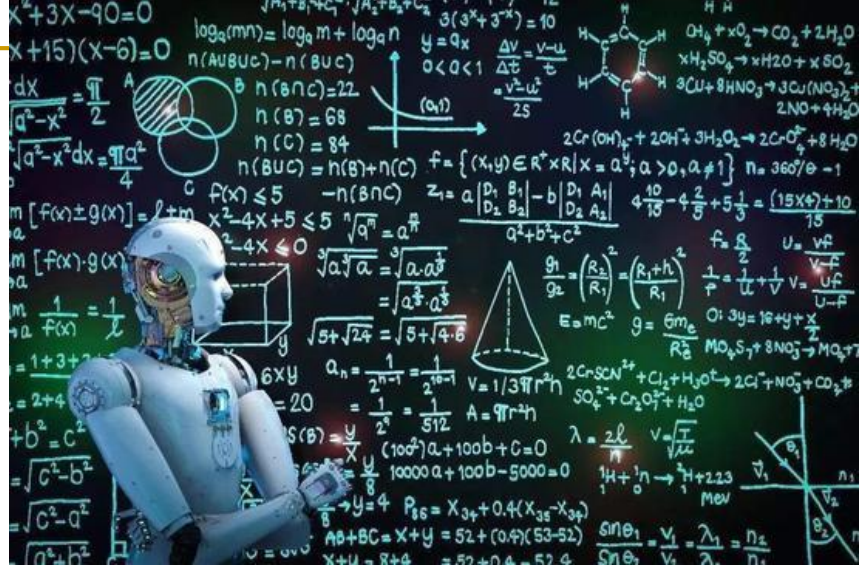
上机辅导：周一1-2节

- Python自学：
  - Eric Matthes（著），袁国忠（译），Python编程：从入门到实践。（人民邮电出版社，2016）
  - http://www.runoob.com/python/python-tutorial.html
    Python 基础教程，中文版的。网站上也有高级课程。
  - https://docs.python.org/zh-cn/3.7/
    Python官方文档，包括入门教程、标准库参考，等等。

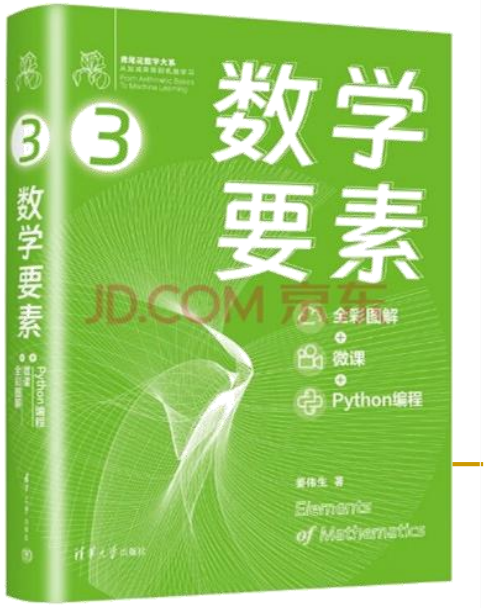
所用的机器学习的Python库：scikit-learn

- 官方网站：https://scikit-learn.org/
- 文档中文翻译： https://sklearn.apachecn.org/
  - sklearn_0.21.3_2019_12_13.epub
- scikit-learn（简称sklearn）是Python语言中专门针对机器学习应用而发展起来的一款开源框架。基本功能主要被分为六大部分：分类，回归，聚类，数据降维，模型选择和数据预处理。
- sklearn 中的模块都是高度抽象化的，所有的分类器基本都可以在3-5行内完成。这虽然限制了使用者的自由度，但增加了模型的效率，降低了批量化、标准化的的难度（通过使用pipeline）。
- sklearn主要适合中小型的、实用机器学习项目。

涉及数学基础知识



- 高等数学
- 线性代数
  - 自学：http://immersivemath.com/ila/index.html
  - 或自学：吴恩达 3
- 概率论
  - 自学：https://seeing-theory.brown.edu/cn.html
  - 或自学：Bishop 1.2-1.2.4
- 生姜(姜伟生)，鸢尾花书：从加减乘除到机器学习
  - https://github.com/Visualize-ML
  - 7册：编程不难；数学要素；矩阵力量；可视之美；统计至简；…



学生成绩评定方法

经申请可只按后两项计分！
33%：67%

- 平时书面与上机作业，占40%；
  - 2-3次书面作业，4-5次上机作业。
- 期末大作业-（组队）项目编程20%；

北大本科时已选过？直接继承！

  - 4-7人/队。
  - 根据自己搜集/整理的数据集，或感兴趣的实际问题，应用机器学习方法；
  - 或，基于已发表的文献及其提供的数据集，复现/改进文献结果，评价原文献的工作；
- 期末闭卷书面考试40%.

助教 （办公室：化学楼A502）

崔畅

- cuichang2022@stu.pku.edu.cn
- 手机：18851835626

阴钰骐

- yinyuqi0001@stu.pku.edu.cn
- 手机：15639917162

王思涵

- sh_wang@pku.edu.cn
- 手机：13021232237

北大相关课程：

~ 80 门/年

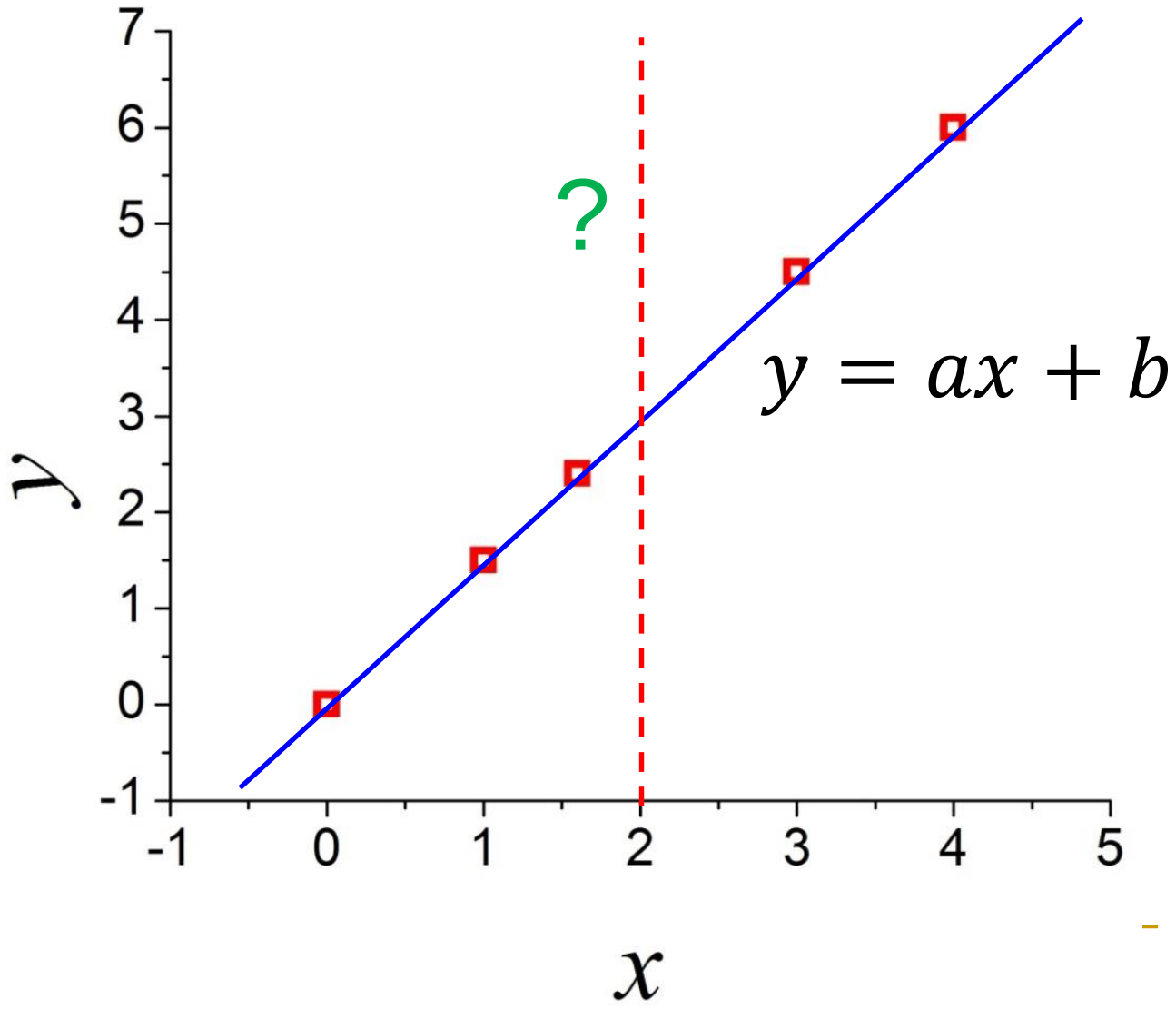
- 机器学习：
  - 21-22学年有28门；22-23学年有34门。
- 人工智能：
  - 21-22学年有41门；22-23学年有43门。
- 院系分布：
  - 信科48（其中智能学院8，微软学院5，人工智能研究院2，王选所2，集成电路学院1），光华9，深研院7，数院6，法学院4，教育学院4，医学3，物院2，工学院2，化院1，地空1，公共卫生学院1，交叉学院1，社会学系1，未来技术学院1，元培1，哲学系1，政管1

机器学习：引言

什么是机器学习（machine learning）？

- 会学习的机器！
- 机器在学习！！

输入	输出
0	0
1	1.5
1.6	2.4
3	4.5
4	6
2	?



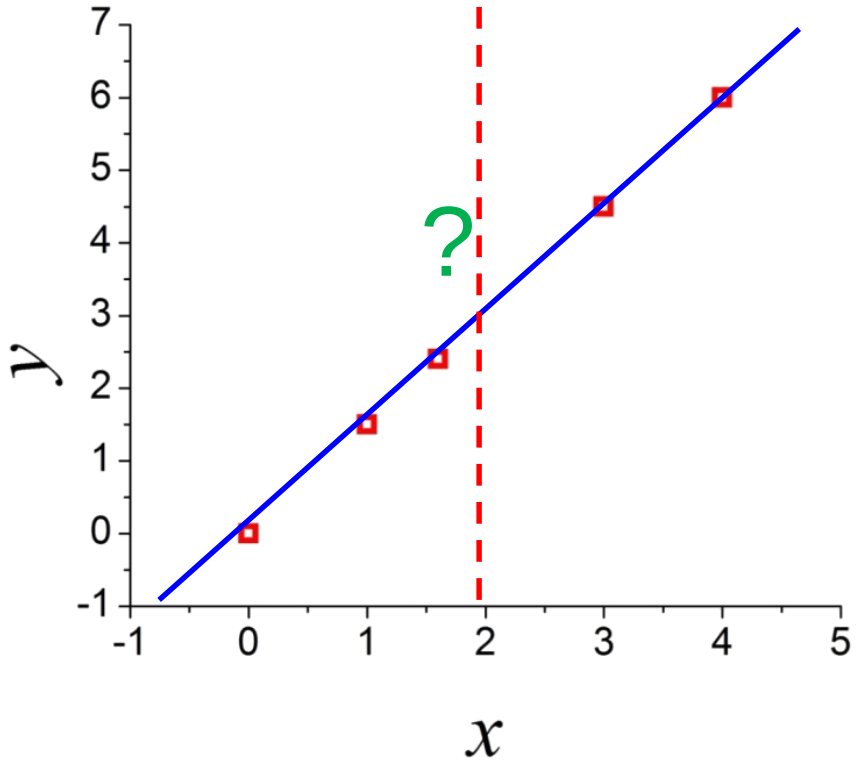
机器学习的定义……

- Herbert Simon（图灵奖、诺贝尔经济学奖获得者）：

  如果一个系统，能够通过执行某个过程，就此改进了它的性能，那么这个过程就是学习。

- Tom Mitchell (1998)：

A computer program is said to *learn* from experience **E** with respect to some task **T** and some performance measure **P**, if its performance on **T**, as measured by **P**, improves with experience **E**.



监督学习（Supervised Learning）

- **监督学习：**给学习算法一个包含了一系列问题与“正确”答案的数据集$\{\mathbf{x}_n, y_n\}$ $(n = 1,2, \dots, N)$，让算法据此预测某一问题$\mathbf{x}$的答案y。
  - $\mathbf{x}$小写粗体无斜体代表矢量，有D个分量（数据维度）。在前述例子中，$D = 1, N = 5$。大写粗体一般代表矩阵。

主要包含两类：

- **回归（Regression）：**y的可能取值范围是连续的。例如：曲线拟合。
- **分类（Classification）：** y的可能取值范围是分立的，例如$\{0,1\}$，代表答案“不是”与“是”。

x的例子.....

- 在前面的例子中，x只有一个分量：

$$\mathbf{x} = [x]$$

- 在图像识别的例子中，对于640*480分辨率的灰度图片，x具有640*480个分量：

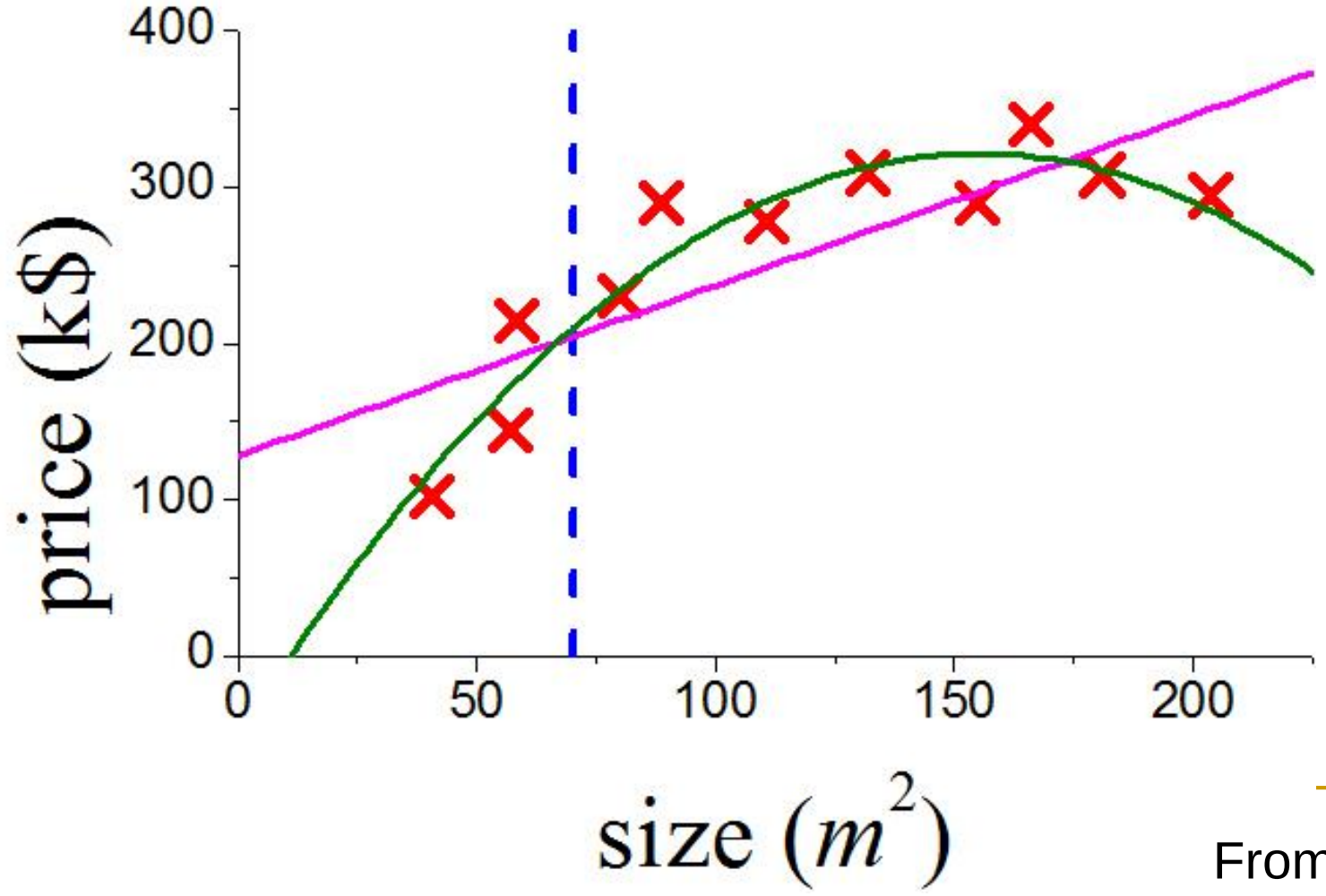
$$\mathbf{x} = \begin{bmatrix} s_{1,1} \\ s_{1,2} \\ \cdots \\ s_{1,480} \\ s_{2,1} \\ s_{2,2} \\ \cdots \\ s_{2,480} \\ s_{3,1} \\ \cdots \\ s_{640,480} \end{bmatrix}$$

对于640*480分辨率的彩色图片，x具有640*480*3个分量：

$$\mathbf{x} = \begin{bmatrix} r_{1,1} \\ g_{1,1} \\ b_{1,1} \\ r_{1,2} \\ \cdots \\ r_{1,480} \\ g_{1,480} \\ b_{1,480} \\ r_{2,1} \\ \cdots \\ b_{640,480} \end{bmatrix}$$

现实一些的回归问题.....

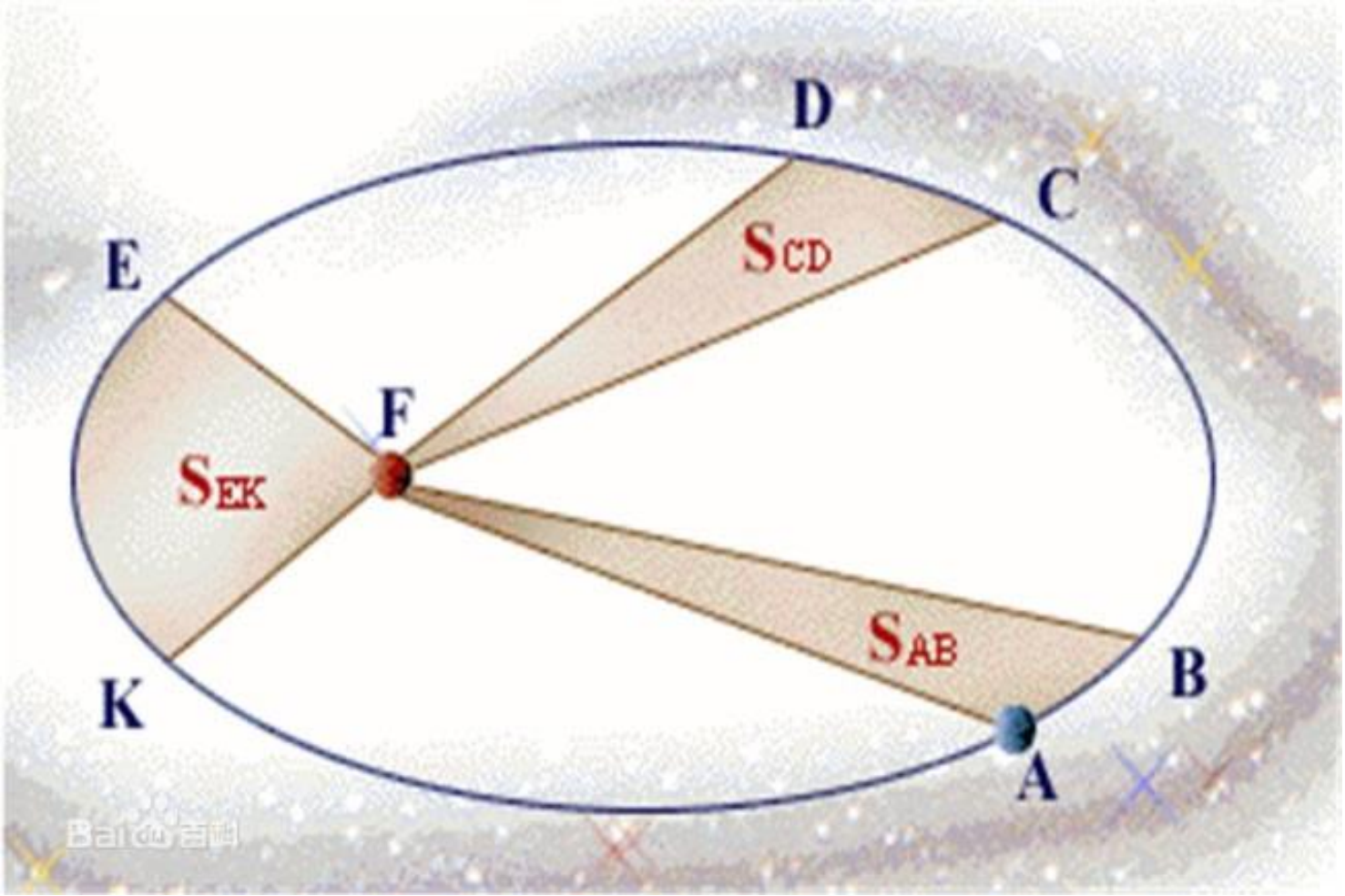
- 房子价格的预测



From A. Ng

一个重大的科学问题.....

- 行星的轨迹预测

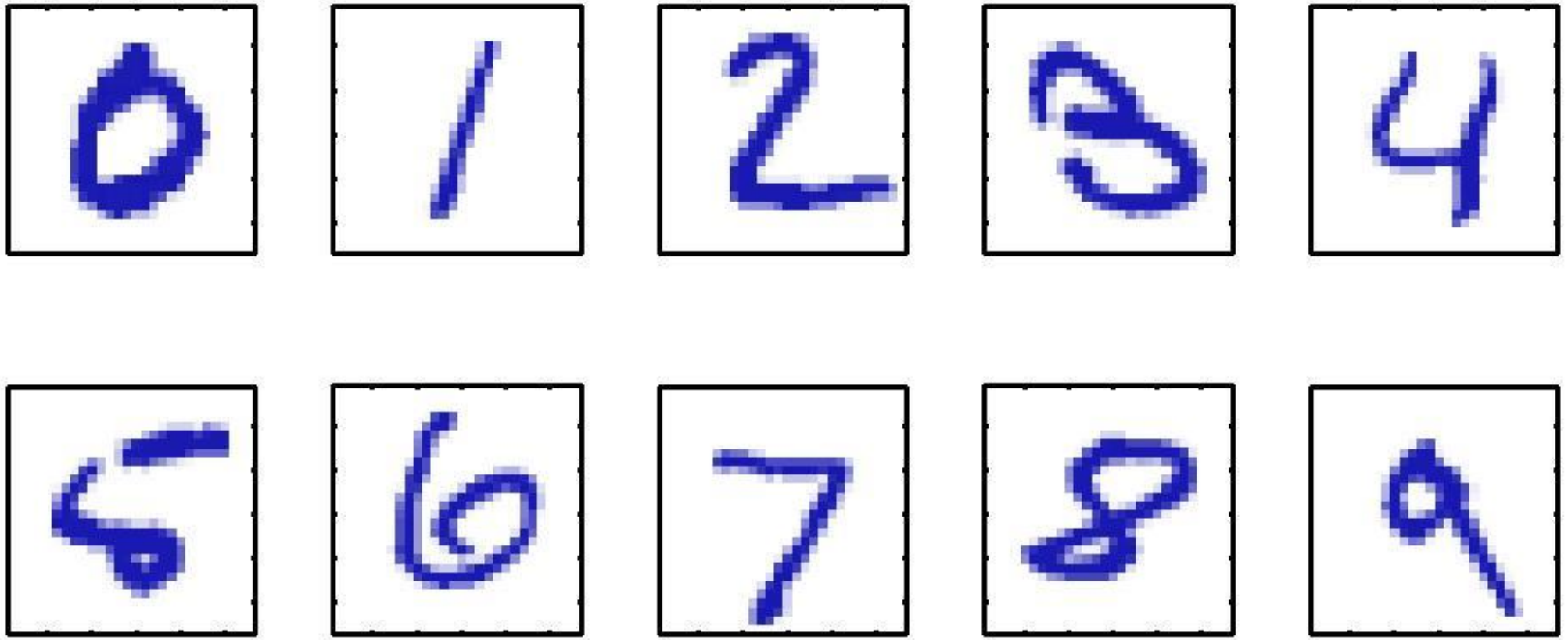


开普勒

分类问题.....

其它：人脸识别；垃圾邮件。

- 手写数字识别：邮政编码

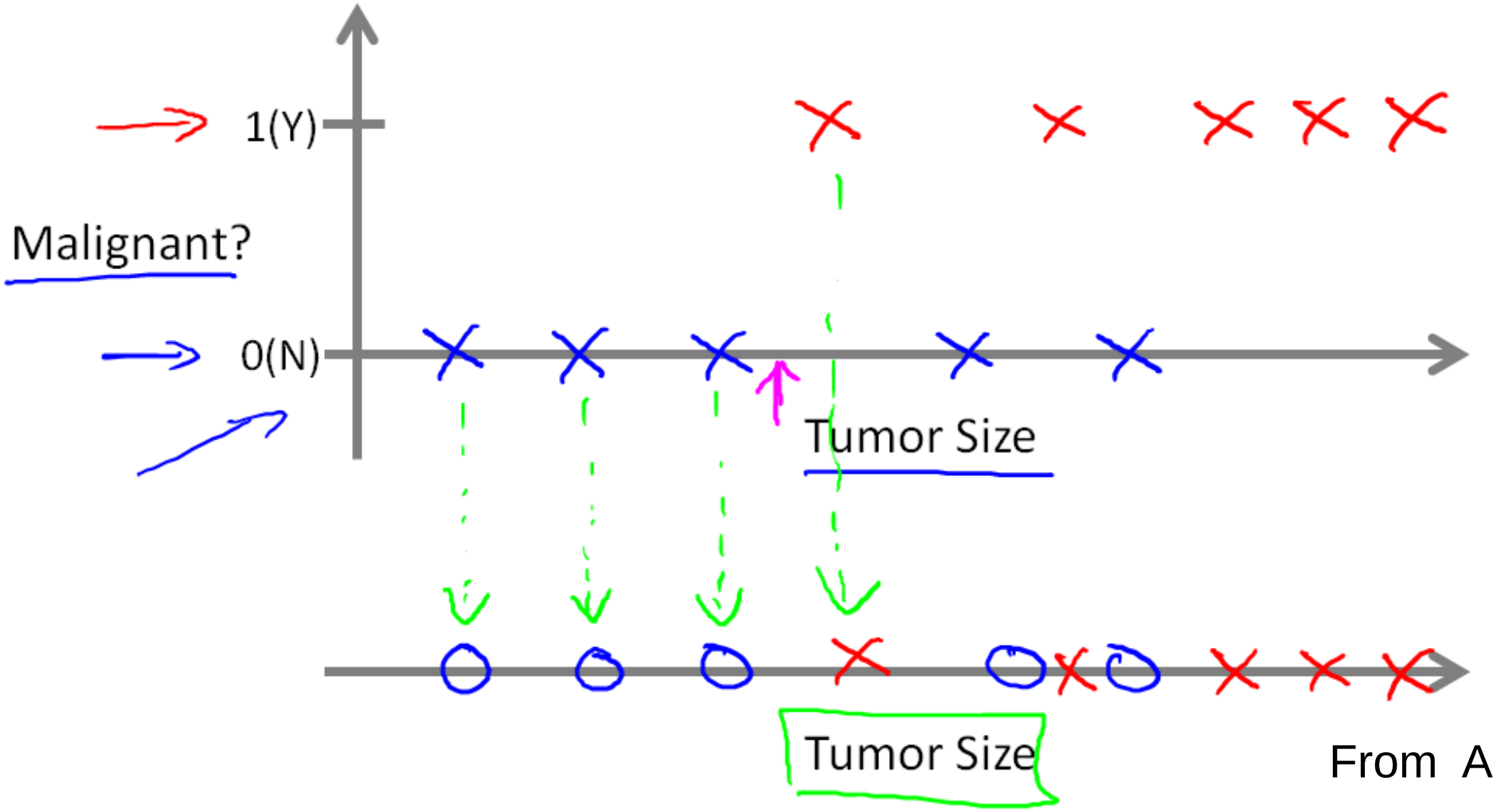


From: Bishop

1989年，Yann LeCun成功将反向传播神经网络应用于支票识别。

分类问题.....

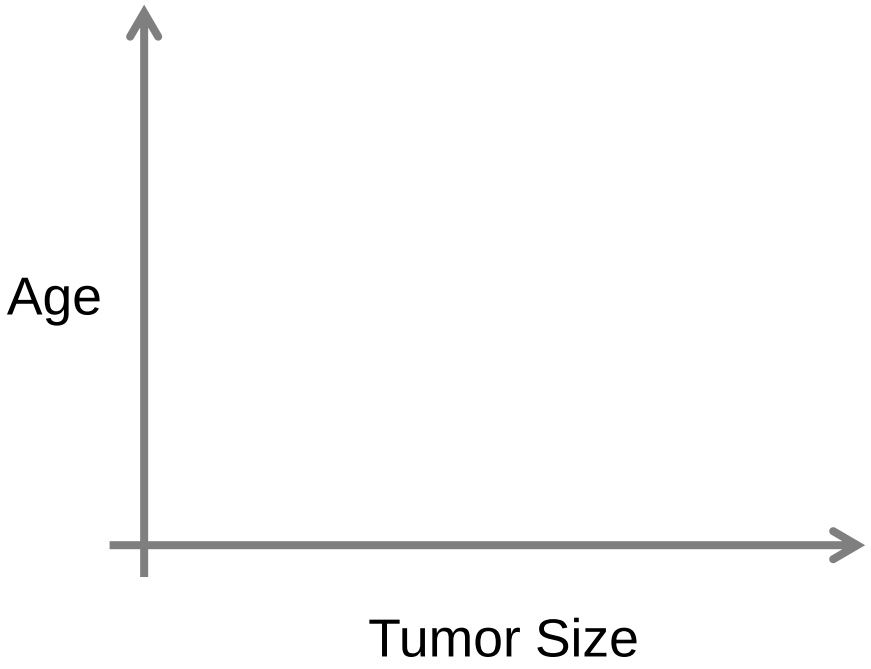
- 乳腺肿瘤是否恶性？



From A. Ng

特征（feature）/输入变量的选取。

- 特征：一事物异于其他事物的特点。
- 对原始数据进行选择和加工，将其变成特征/输入提供给机器学习模型。



- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape

…

From A. Ng

无监督学习（Unsupervised Learning）

- 给学习算法一个数据集$\{\mathbf{x}_n\}\,(n = 1,2,\ldots,N)$，里面没有答案或标签，让算法据此寻找其中的规律与结构。

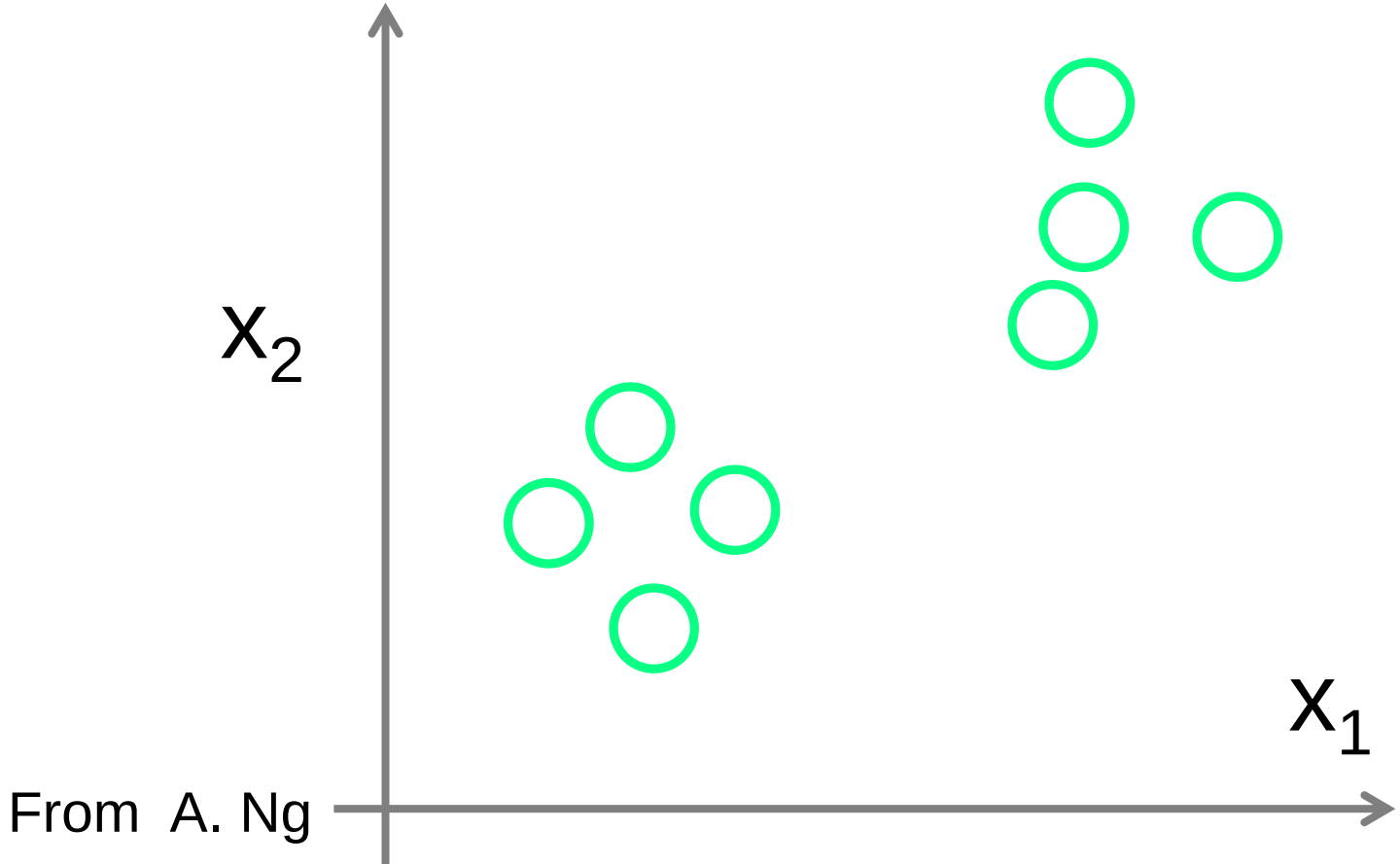
- 例子：

  微信群；

  手机套餐；

  新闻分组；

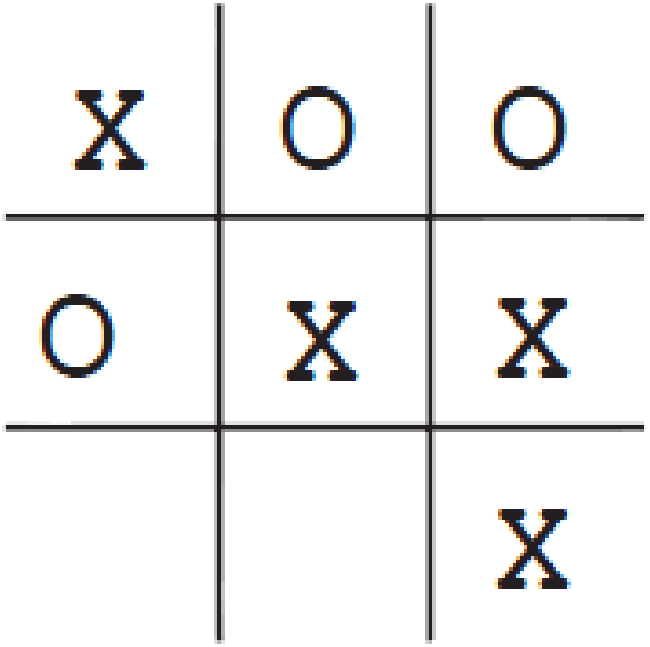
  购物网站分类；



From A. Ng

强化学习（Reinforcement Learning）

- 给学习算法一个规则和目标，让算法通过**主动**探索寻找达到目标的最佳策略。
- 又称评价学习或增强学习，用于描述和解决智能体（agent）在与环境的交互过程中通过学习策略以达成回报最大化或实现特定目标的问题 。
  - 没有现成数据。
  - 例子：围棋 。
  - 例子：三连棋游戏Tic-Tac-Toe
  - 例子：电子游戏



三连棋游戏
Tic-Tac-Toe

思考：与《物理化学》等课程的差别...

- 内容还在发展/变化中。
- 发展阶段：炼金术？炼丹术？
  - 炼金术是现代化学的古代先驱（雏形）。

- Frank Wilczek：我们从天文学史获得的重要教训是，大数据本身是解释不了自己的。构建简化的数学模型，再将其与真实的物理世界联系起来，这才是从数据这块原始矿石中提炼出“意义”这颗稀有宝石的可靠方法。
- Yann LeCun：在科学技术史上，工程学上的进步几乎总是先于理论认识。例如，蒸汽机先于热力学。

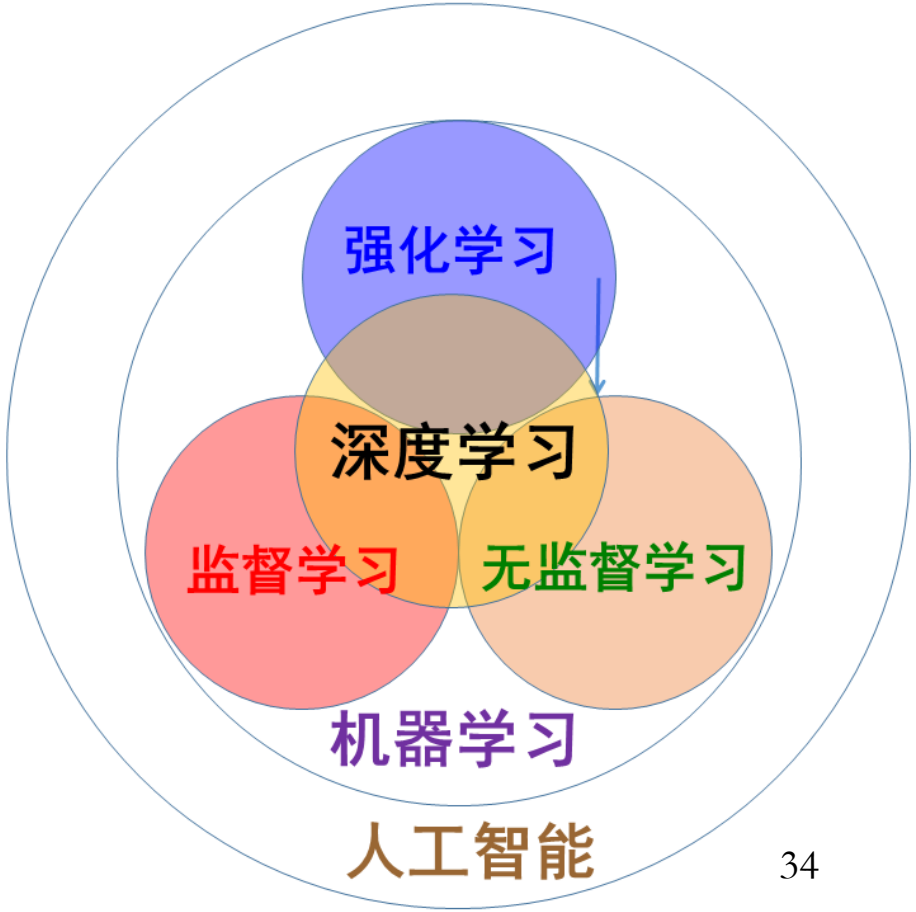
小结

监督学习

无监督学习

强化学习

- 回归
- 分类



思考题1.1.

假设你经营着一家公司，需处理这两个问题：

（1）你有一大批同样的货物，想预测接下来三个月能卖多少件？

（2）你有许多客户，你想写一个软件来检查并判断用户的账户是否曾经被盗过？

请问这两个问题，它们属于分类问题、还是回归问题?

- Reference:
  - 刘志荣1；
  - Bishop 1.1；
  - 实战 1；
  - 吴恩达 1。
  - 刘志荣《统计热力学》内容摘录
    - 概率论：贝叶斯学派的角度.pdf

扩展阅读：

- https://www.iyiou.com/p/38486

  从1308年到2016年，人工智能在这700年时间里发生了什么？.mht

- http://baijiahao.baidu.com/s?id=1646913007012147051

  大数据不等于科学规律.mht

- https://www.huxiu.com/article/376933.html

  为什么TikTok能够横扫美国市场？.mht

- https://wwww.huxiu.com/article/325277.html

  人工智能还是人工智障.mht

- https://baijiahao.baidu.com/s?id=1596520892125761222

  这是一份文科生都能看懂的线性代数简介.mht

- https://mp.weixin.qq.com/s?__biz=MzA3MzI4MjgzMw==&mid=2650731034&idx=1&sn=c700041fe10108ca1068c4d80aa0d05a&chksm=871b3664b06cbf726bd93d7b3db4f15125df77a6f69ddb0304b292b932071d2d78104ad5d4f0&scene=21#wechat_redirect

  从贝叶斯定理到概率分布：综述概率论基本定义.mht

- https://www.jiqizhixin.com/articles/2021-01-11-3

  人工智能十年回顾：CNN、AlphaGo、GAN……它们曾这样改变世界.mht

- https://www.huxiu.com/article/428329.html

  未来科学的进步，将取决于人类还是AI？.mht

- https://www.huxiu.com/article/384223.html

  当AI遇到网文.mht

  - 用AI技术来翻译网文，帮助中国文学出海。利用技术翻译一部网文，到最终全球上架，最快只需要48小时。

- https://www.huxiu.com/article/385803.html

  那些为AI研究做过贡献的人，29%在中国念的本科.mht

- https://www.jiqizhixin.com/articles/2019-01-25-13

  一文看懂机器学习3种类型的概念、根本差别及应用.mht

- https://www.bilibili.com/read/cv22407188/

一篇文章搞懂机器学习——文科生写给文科生的人工智能科普.mhtml

- 高文的帝国疆域刚好是一个规整的正方形。最近，他突然收到了一些汇报，声称帝国里陆续出现了大量一出生便觉醒超能力的“超凡者”。这本该是件好事，但因新生儿过于体弱，一旦觉醒超凡力量，反而会在几天内走向失控——除非派遣白骑士前往进行精神镇定。这是个大难题，毕竟全国的白骑士数量是有限的。

一个机器学习的笑话

谢谢大家！