机器学习方法与应用

中科院大学电子电气与通信工程学院

叶齐祥

雁栖湖园区，学园二，457房间，qxye@ucas.ac.cn

助教:张天亮、刘畅

雁栖湖园区，学园二，330房间

zhangtianliang13@mails.ucas.ac.cn liuchang615@mails.ucas.ac.cn

提纲

- 机器学习的定义
- 机器学习的目的
- 机器学习的一般步骤
- 机器学习的输入
- 机器学习：新一代计算技术的浪潮
- 机器学习 vs. 数据挖掘vs.统计学习
- 机器学习的相关期刊会议

提纲

- 机器学习的定义
- 机器学习的目的
- 机器学习的一般步骤
- 机器学习的输入
- 机器学习：新一代计算技术的浪潮
- 机器学习 vs. 数据挖掘vs.统计学习
- 机器学习的相关期刊会议

机器学习的定义

- ✓ 通过机器学习的算法研究及其与具体问题的恰当结合，获得合适的模型；
- ✓ 对于于一些工程应用任务，依赖于数学与经验，设计学习模型，提高算法的性能
- ✓ Studying models from existing information or from observation.
  依赖于对**现存数据**的学习或者观察获取新的推理模型的过程。但是，最新的强化学习超出了现存数据的范畴
- ✓ ML是具体的实现方法：回归分析、SVM，NeuralNetwork, 概率方法, 聚类方法…

提纲

- 机器学习的定义
- 机器学习的目的
- 机器学习的一般步骤
- 机器学习的输入
- 机器学习：新一代计算技术的浪潮
- 机器学习 vs. 数据挖掘vs.统计学习
- 机器学习的相关期刊会议

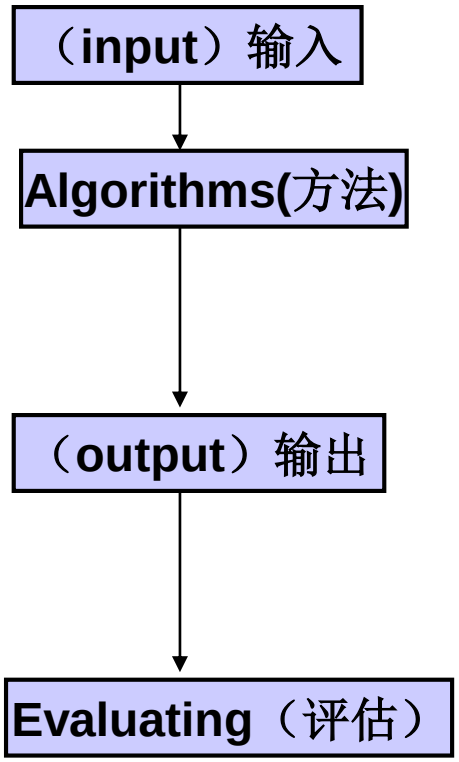
机器学习的目的

- ✓ 分类（Classification、Clustering）
  - ✓ 身高1.15m，体重60kg的儿童健康么？
  - ✓ 如何将教室里的学生按爱好、身高划分为5类？
- ✓ 预测（Regression、Prediction）
  - ✓ 如何预测中关村周边的房价？
  - ✓ 天气预报
  - ✓ 预测下一步围棋的最优落点
  - ✓ 预测机器人下一步应该做的动作

提纲

- 机器学习的定义
- 机器学习的目的
- 机器学习的一般步骤
- 机器学习的输入
- 机器学习：新一代计算机技术的浪潮
- 机器学习 vs. 数据挖掘vs.统计学习
- 机器学习的相关期刊会议

机器学习的一般步骤*



Concepts, instances, and features

Statistical modeling, decision trees, linear models, instance-based learning, clustering, …

Decision tables, classification rules, association rules, instance-based representation, clusters…

Training and testing, predicting performance, cross-validation, leave-one-out, boostrap, comparing data mining methods, counting the cost

提纲

- 机器学习的定义
- 机器学习的目的
- 机器学习的一般步骤
- 机器学习的输入
- 机器学习：新一代计算技术的浪潮
- 机器学习 vs. 数据挖掘vs.统计学习
- 机器学习的相关期刊会议

样本

- **输入的基本单元**
  - 机器学习的输入是一系列样本（带表号/无标号），机器学习是要将这些样本分类、回归、关联或者聚类。
  - 每一个样本都是样本（样例）都有一系列特性
  - 多个样本及其特性构成一个矩阵，或者一张表，构成ML的基本输入

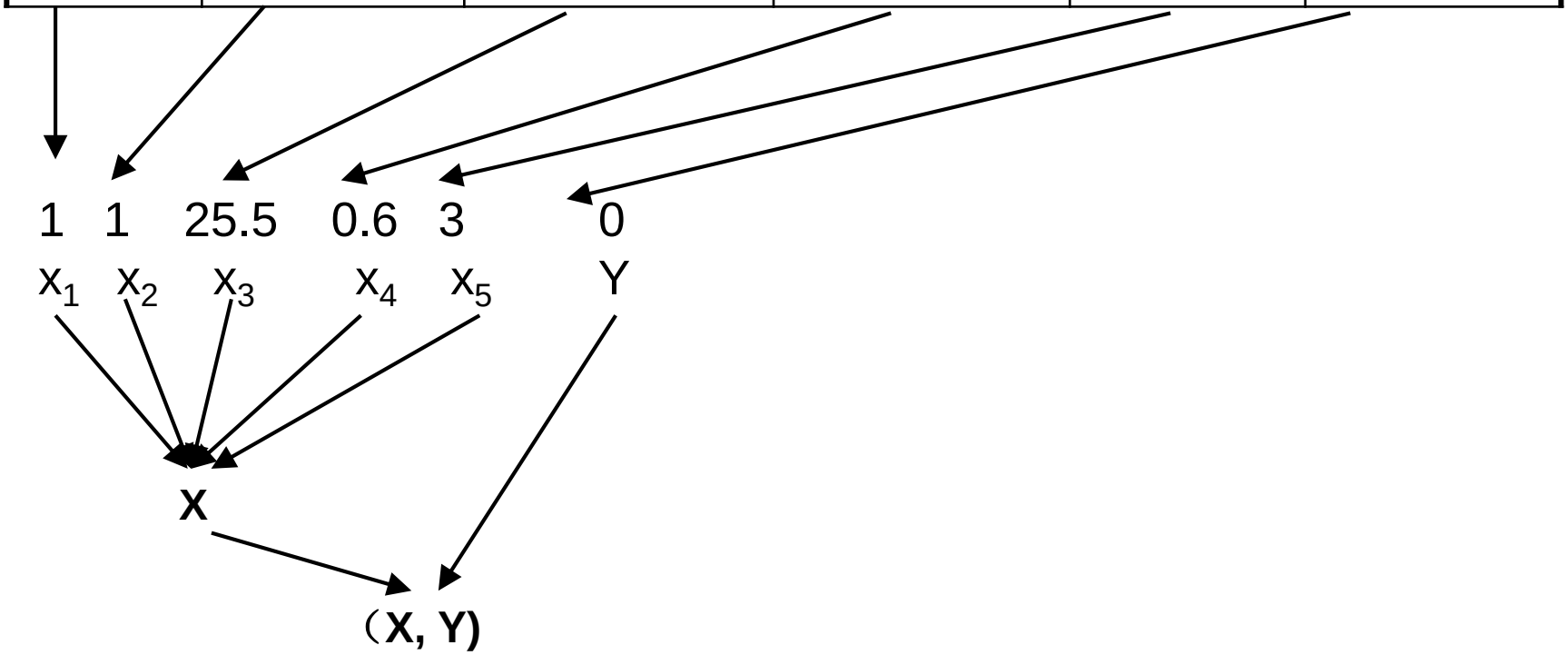
数据特征化

- **每一个样本都有一系列固定的，事先确定的特征**
  - **Each individual, independent instance that provide the input to machine learning is characterized by its values on a fixed, predefined set of features or attributes.**

编号	天气	温度	湿度	风	是否去打球
1	晴天	炎热	高	弱	不去
2	晴天	炎热	高	强	不去
3	阴天	炎热	高	弱	去
4	下雨	适中	高	弱	去
5	下雨	寒冷	正常	弱	去
6	下雨	寒冷	正常	强	不去
7	阴天	寒冷	正常	强	去
8	晴天	适中	高	弱	不去
9	晴天	寒冷	正常	弱	去
14	下雨	适中	高	强	不去

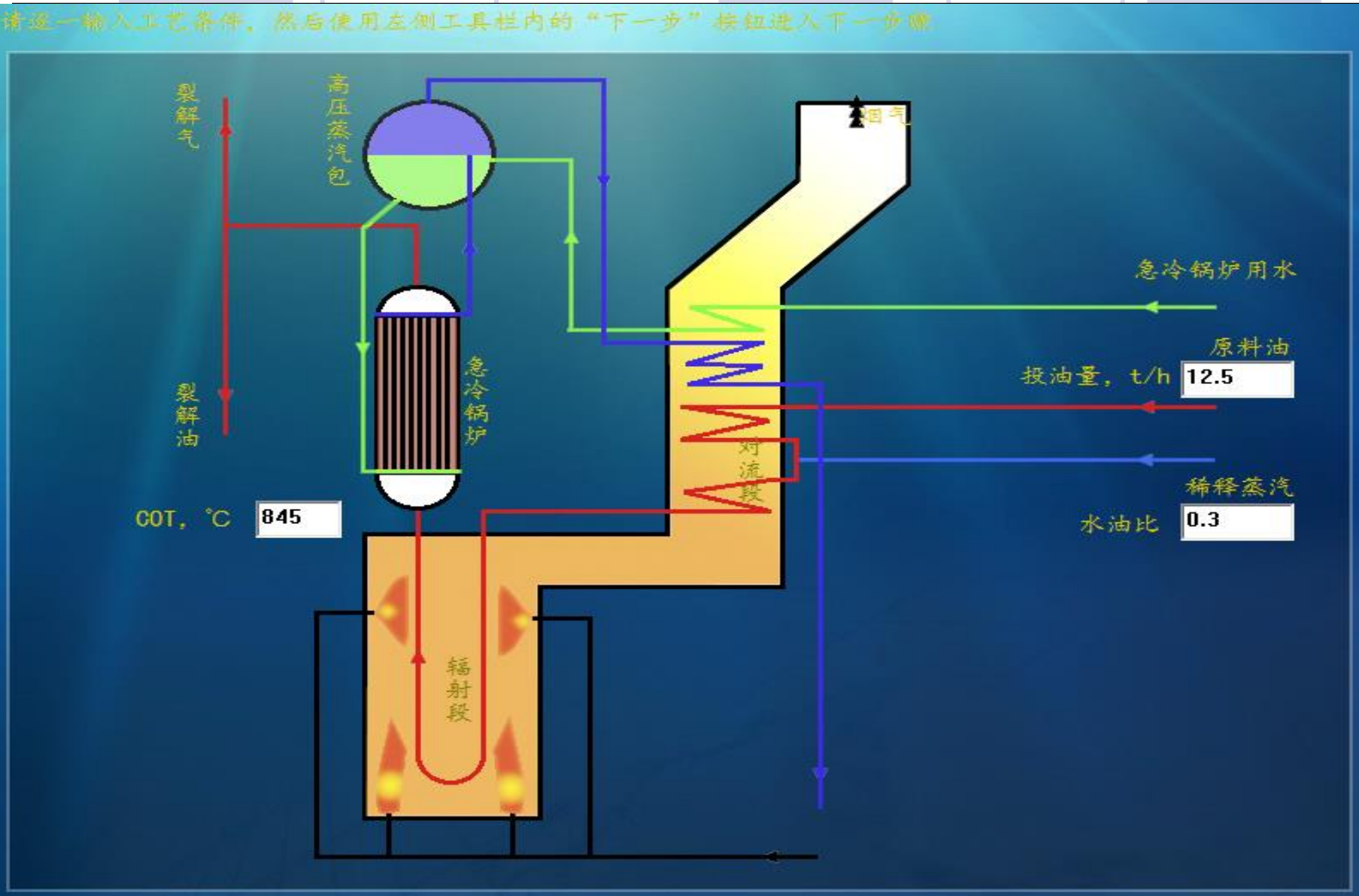
数据特征化

编号	天气	温度	湿度	风	是否去打球
1	晴天	25.5	60%	3	不去

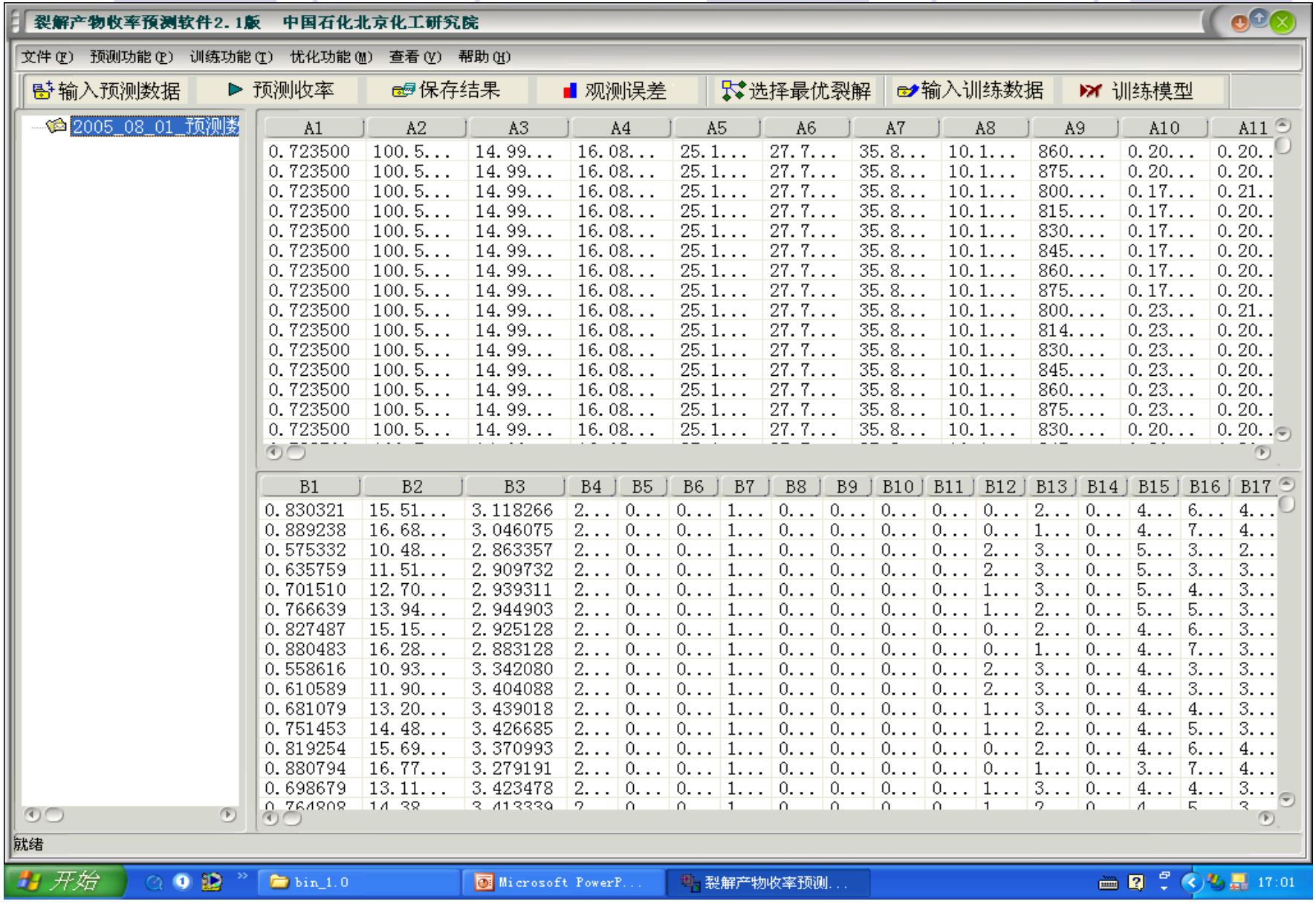


数据特征化

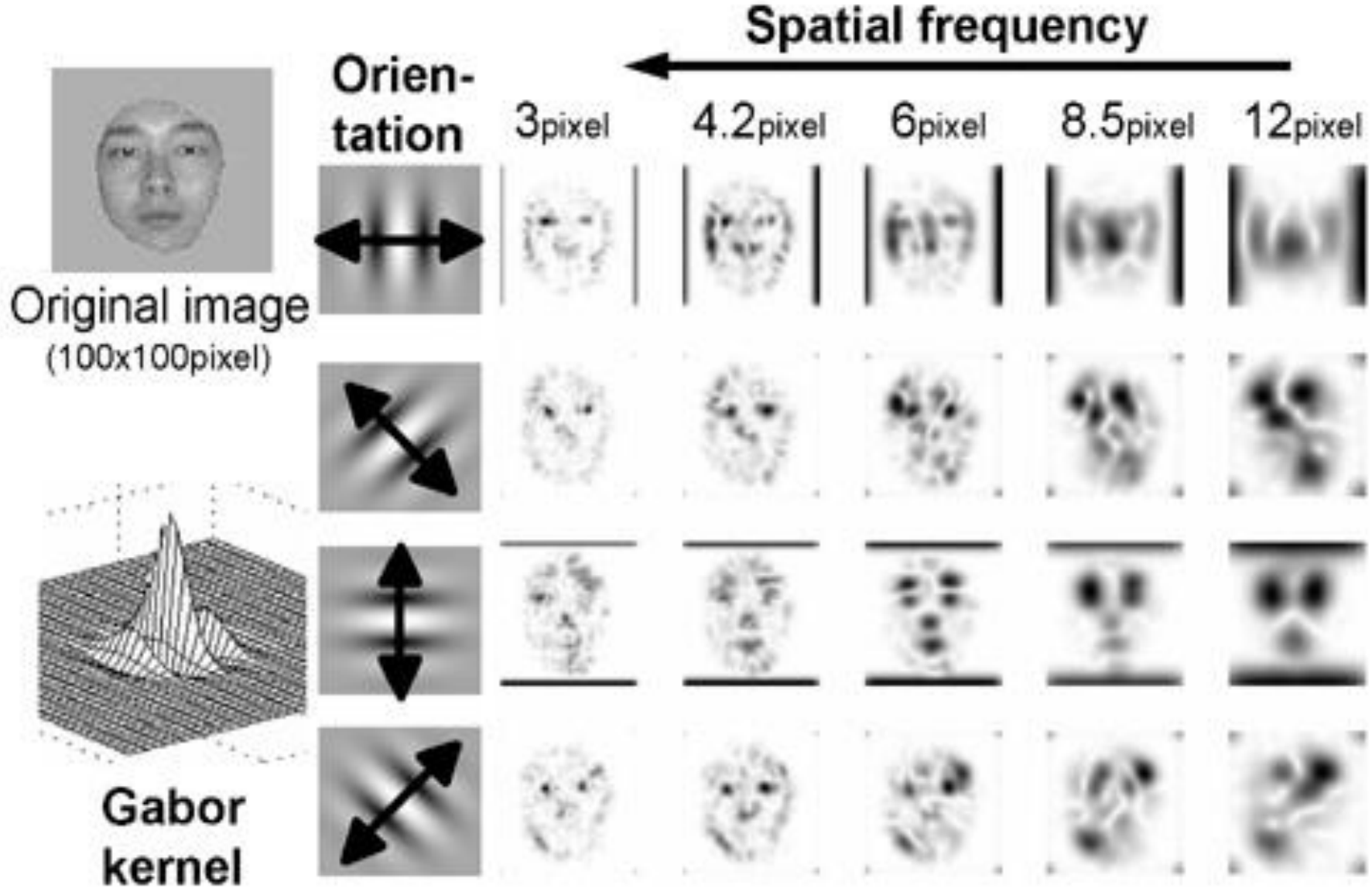
数据特征化


请逐一输入工艺条件，然后使用左侧工具栏内的“下一步”按钮进入下一步骤
裂解气
高压蒸汽包
烟气
急冷锅炉用水
原料油
投油量，t/h
12.5
裂解油
急冷锅炉
对流段
稀释蒸汽
水油比
0.3
COT，℃
845
辐射段

数据特征化


裂解产物收率预测软件2.1版 中国石化北京化工研究院
文件(F) 预测功能(P) 训练功能(T) 优化功能(M) 查看(V) 帮助(H)
输入预测数据
预测收率
保存结果
观测误差
选择最优裂解
输入训练数据
训练模型
2005_08_01_预测类
A1 A2 A3 A4 A5 A6 A7 A8 A9 A10 A11
0.723500 100.5... 14.99... 16.08... 25.1... 27.7... 35.8... 10.1... 860.... 0.20... 0.20..
0.723500 100.5... 14.99... 16.08... 25.1... 27.7... 35.8... 10.1... 875.... 0.20... 0.20..
0.723500 100.5... 14.99... 16.08... 25.1... 27.7... 35.8... 10.1... 800.... 0.17... 0.21..
0.723500 100.5... 14.99... 16.08... 25.1... 27.7... 35.8... 10.1... 815.... 0.17... 0.20..
0.723500 100.5... 14.99... 16.08... 25.1... 27.7... 35.8... 10.1... 830.... 0.17... 0.20..
0.723500 100.5... 14.99... 16.08... 25.1... 27.7... 35.8... 10.1... 845.... 0.17... 0.20..
0.723500 100.5... 14.99... 16.08... 25.1... 27.7... 35.8... 10.1... 860.... 0.17... 0.20..
0.723500 100.5... 14.99... 16.08... 25.1... 27.7... 35.8... 10.1... 875.... 0.17... 0.20..
0.723500 100.5... 14.99... 16.08... 25.1... 27.7... 35.8... 10.1... 800.... 0.23... 0.21..
0.723500 100.5... 14.99... 16.08... 25.1... 27.7... 35.8... 10.1... 814.... 0.23... 0.20..
0.723500 100.5... 14.99... 16.08... 25.1... 27.7... 35.8... 10.1... 830.... 0.23... 0.20..
0.723500 100.5... 14.99... 16.08... 25.1... 27.7... 35.8... 10.1... 845.... 0.23... 0.20..
0.723500 100.5... 14.99... 16.08... 25.1... 27.7... 35.8... 10.1... 860.... 0.23... 0.20..
0.723500 100.5... 14.99... 16.08... 25.1... 27.7... 35.8... 10.1... 875.... 0.23... 0.20..
0.723500 100.5... 14.99... 16.08... 25.1... 27.7... 35.8... 10.1... 830.... 0.20... 0.20..
B1 B2 B3 B4 B5 B6 B7 B8 B9 B10 B11 B12 B13 B14 B15 B16 B17
0.830321 15.51... 3.118266 2... 0... 0... 1... 0... 0... 0... 0... 0... 2... 0... 4... 6... 4...
0.889238 16.68... 3.046075 2... 0... 0... 1... 0... 0... 0... 0... 0... 1... 0... 4... 7... 4...
0.575332 10.48... 2.863357 2... 0... 0... 1... 0... 0... 0... 0... 2... 3... 0... 5... 3... 2...
0.635759 11.51... 2.909732 2... 0... 0... 1... 0... 0... 0... 0... 2... 3... 0... 5... 3... 3...
0.701510 12.70... 2.939311 2... 0... 0... 1... 0... 0... 0... 0... 1... 3... 0... 5... 4... 3...
0.766639 13.94... 2.944903 2... 0... 0... 1... 0... 0... 0... 0... 1... 2... 0... 5... 5... 3...
0.827487 15.15... 2.925128 2... 0... 0... 1... 0... 0... 0... 0... 0... 2... 0... 4... 6... 3...
0.880483 16.28... 2.883128 2... 0... 0... 1... 0... 0... 0... 0... 0... 1... 0... 4... 7... 3...
0.558616 10.93... 3.342080 2... 0... 0... 1... 0... 0... 0... 0... 2... 3... 0... 4... 3... 3...
0.610589 11.90... 3.404088 2... 0... 0... 1... 0... 0... 0... 0... 2... 3... 0... 4... 3... 3...
0.681079 13.20... 3.439018 2... 0... 0... 1... 0... 0... 0... 0... 1... 3... 0... 4... 4... 3...
0.751453 14.48... 3.426685 2... 0... 0... 1... 0... 0... 0... 0... 1... 2... 0... 4... 5... 3...
0.819254 15.69... 3.370993 2... 0... 0... 1... 0... 0... 0... 0... 0... 2... 0... 4... 6... 4...
0.880794 16.77... 3.279191 2... 0... 0... 1... 0... 0... 0... 0... 0... 1... 0... 3... 7... 4...
0.698679 13.11... 3.423478 2... 0... 0... 1... 0... 0... 0... 0... 1... 3... 0... 4... 4... 3...
就绪
开始
bin_1.0
Microsoft PowerP...
裂解产物收率预测...
17:01

数据特征化


Spatial frequency
Orien-
tation
3pixel
4.2pixel
6pixel
8.5pixel
12pixel
Original image
(100x100pixel)
Gabor
kernel

数据特征化

$$\begin{bmatrix} x_1^{(1)}, x_2^{(1)}, \ldots, x_K^{(1)} \\ x_1^{(2)}, x_2^{(2)}, \ldots, x_K^{(2)} \\ \ldots \\ x_1^{(n)}, x_2^{(n)}, \ldots, x_K^{(n)} \end{bmatrix} \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \ldots \\ y^{(n)} \end{bmatrix}$$

（X (1), y(1)）

（X(2), y(2)）

（X(3), y(3)）

…

（X(n), y(n)） ⟶ （X, Y）

数据特征化

●数据标准化

- 观察数据分布: 集中趋势, 差别 和 分布
- 计算数据统计特性：median, max, min, outliers, variance, 等.

$$X = \{x_1, x_2 ..., x_k, ... | y\}$$

$$x_k = \frac{x_k - min(x_k)}{max(x_k) - min(x_k)}$$

$$x_k = \frac{x_k - \overline{x_k}}{\sigma(x_k)}$$

提纲

- 机器学习的定义
- 机器学习的目的
- 机器学习的一般步骤
- 机器学习的输入
- 机器学习：新一代计算机技术的浪潮
- 机器学习 vs. 数据挖掘vs.统计学习
- 机器学习的相关期刊会议

机器学习：新一代计算机技术的浪潮

- **Mainframe**
- **Minicomputer**
- **Workstations**
- **Personal computer**
- **Smartphones+Clouds**
- **Ubiquitous computing+Machine Learning**

案例1:基于机器学习的化工生产数据预测

模型指标

平均误差

平均误差控制在8%以内或者更低

标准差在8%以内

单数据预测时间

普通 PC 机， 预测时间小于0.001s

时间复杂度要求

非在线学习，复杂度要求低

训练耗时，但是可以离线更新模型

案例2:基于机器学习的文档分类

是公司主页？

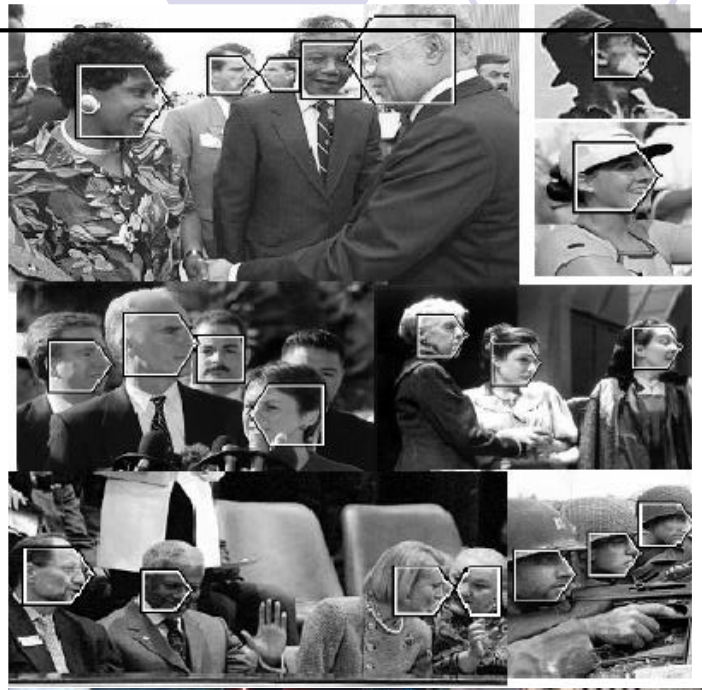
是个人主页？

还是学校网站？

邮件过滤、 搜索引擎、

案例3:基于机器学习的目标识别

训练样本（机器学习的输入）



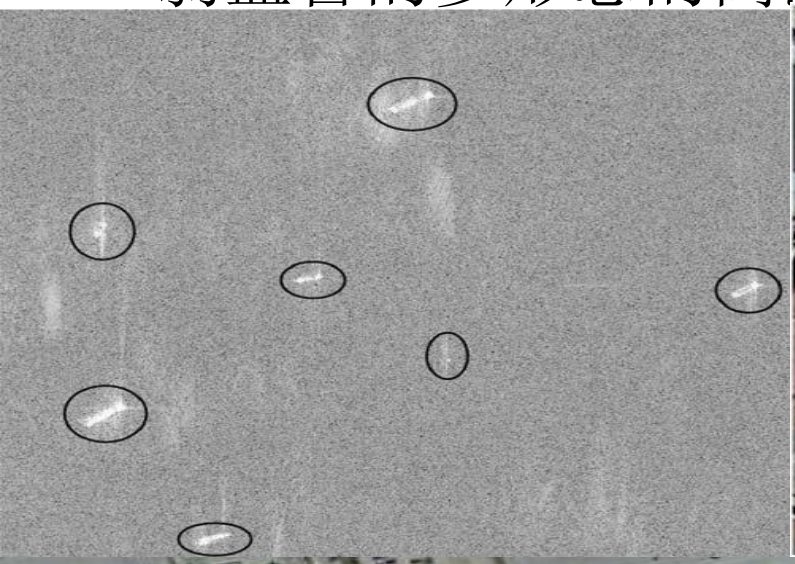
测试样本

训练样本（机器学习的输入）

测试样本

案例3:基于机器学习的目标识别

- 弱监督的多形态的高清航拍图像目标检测识别



案例4:智能交互

Kinect交互Demo

GoogleGlasses

KINECT
for XBOX 360.

一些与机器学习有关的演示

DemoLPR

http://yann.lecun.com/exdb/lenet/a35.html

ObjectDetection

ImageNet

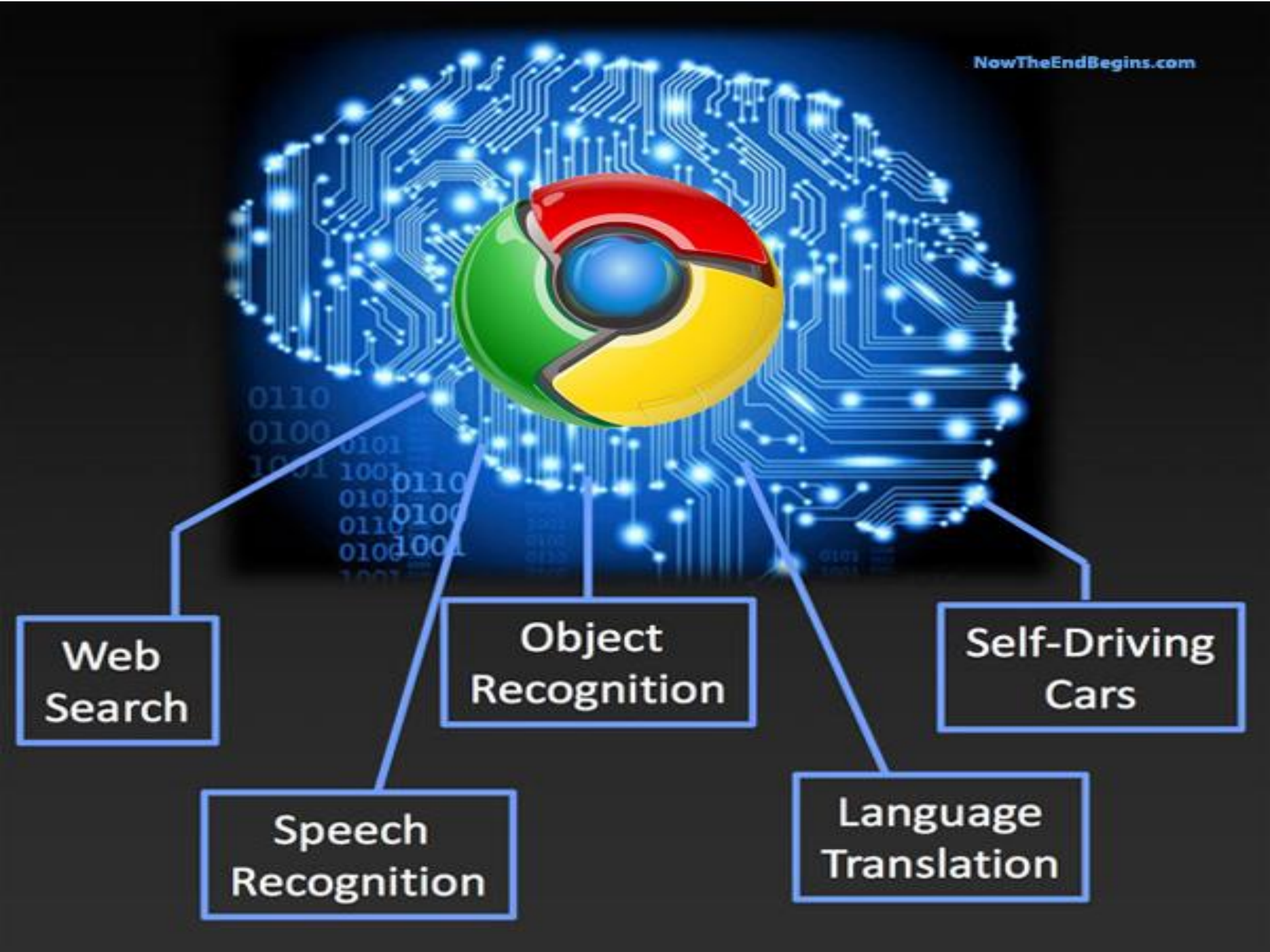
http://shitu.baidu.com/

语音识别

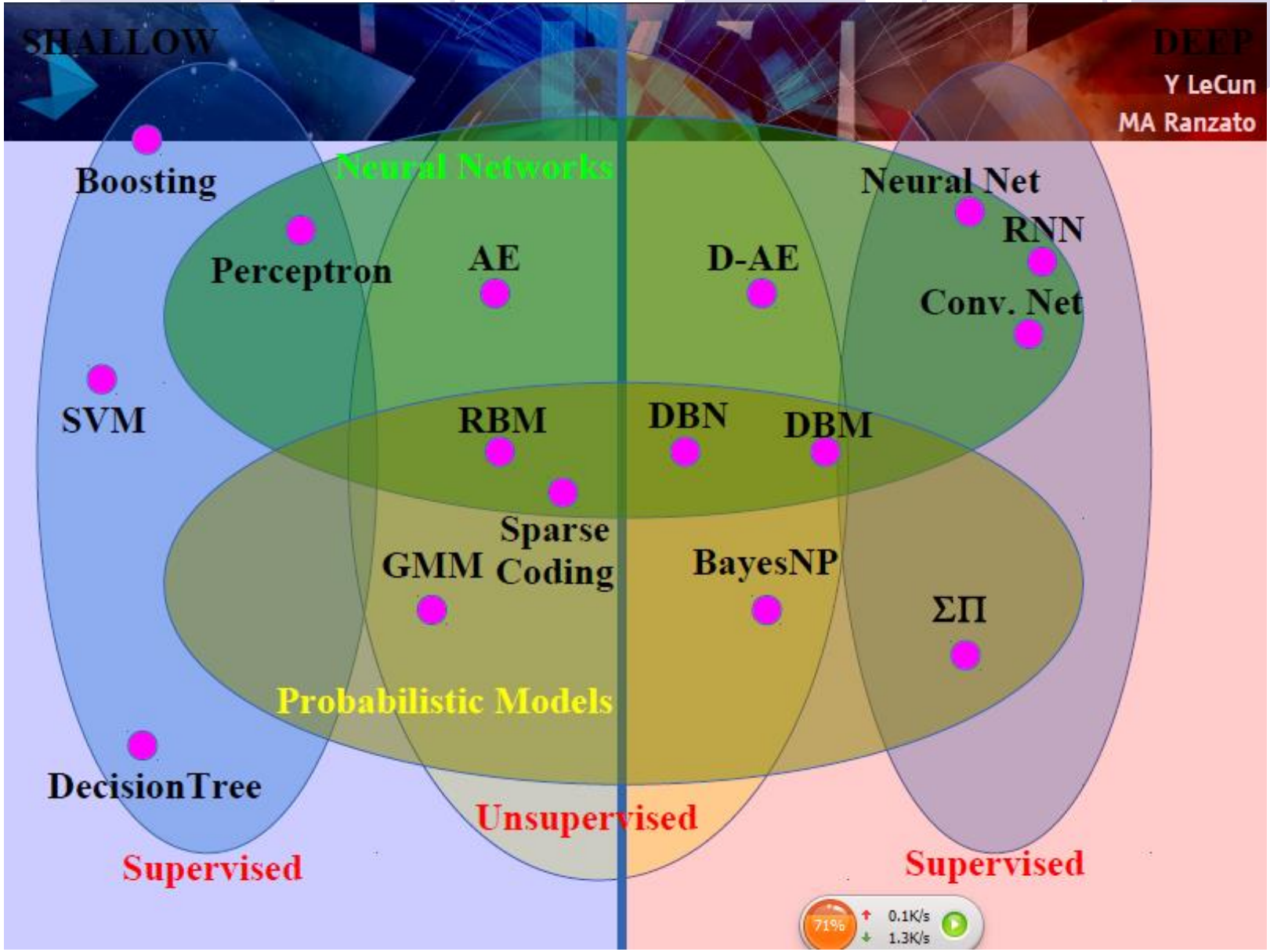
机器学习、强化学习与智能决策（AlphaGo）

机器学习：新一代计算技术的浪潮

深度学习带来了机器学习的新浪潮，推动“**大数据＋深度模型**”时代的来临，以及人工智能和人机交互大踏步前进。**如果我们能在理论、建模和工程方面突破深度学习面临的一系列难题，人工智能的梦想不再遥远。**


NowTheEndBegins.com
Web Search
Speech Recognition
Object Recognition
Language Translation
Self-Driving Cars

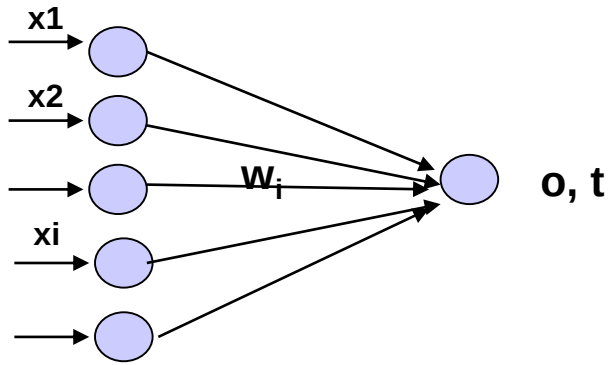
从浅层学习到深度学习-Lecun

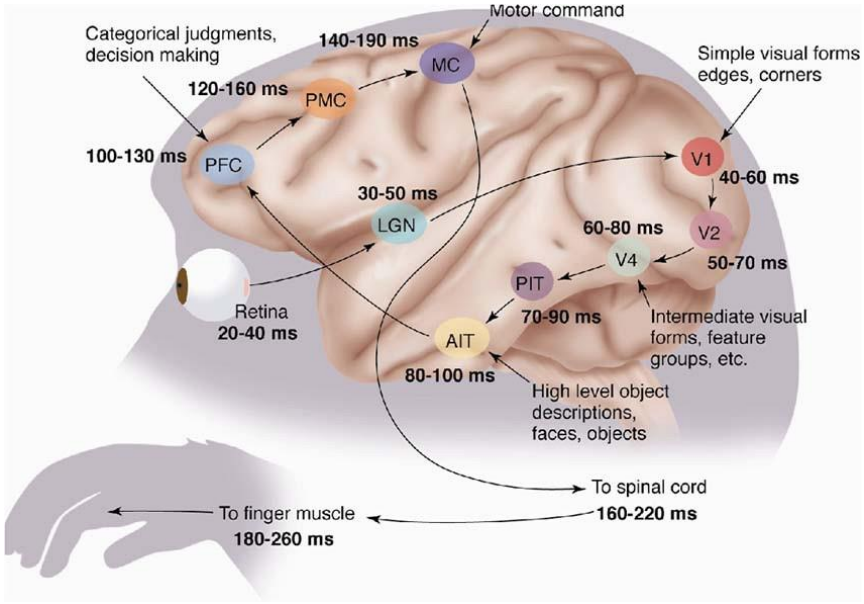


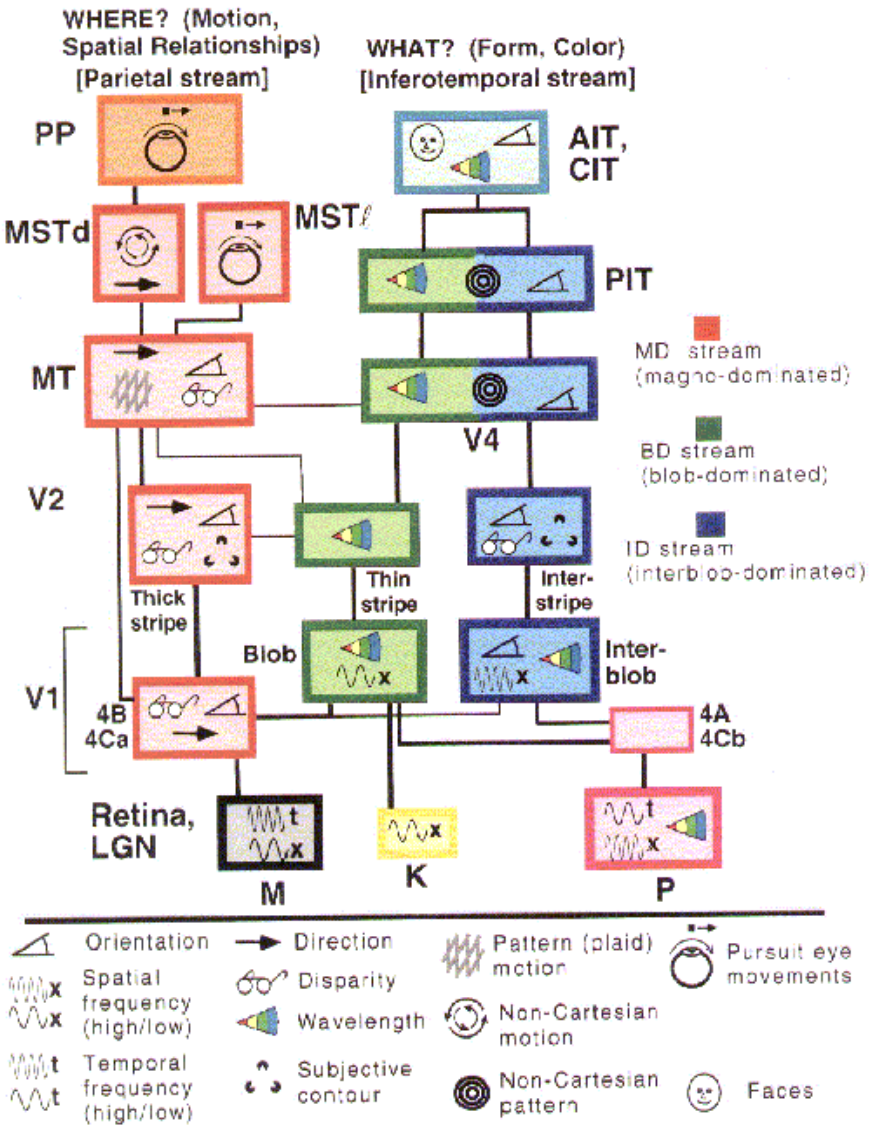
机器学习：新一代计算机技术的浪潮

从浅层学习到深度学习-Lecun

$$H = \{\vec{w} \mid \vec{w} \in R^{n+1}\}$$

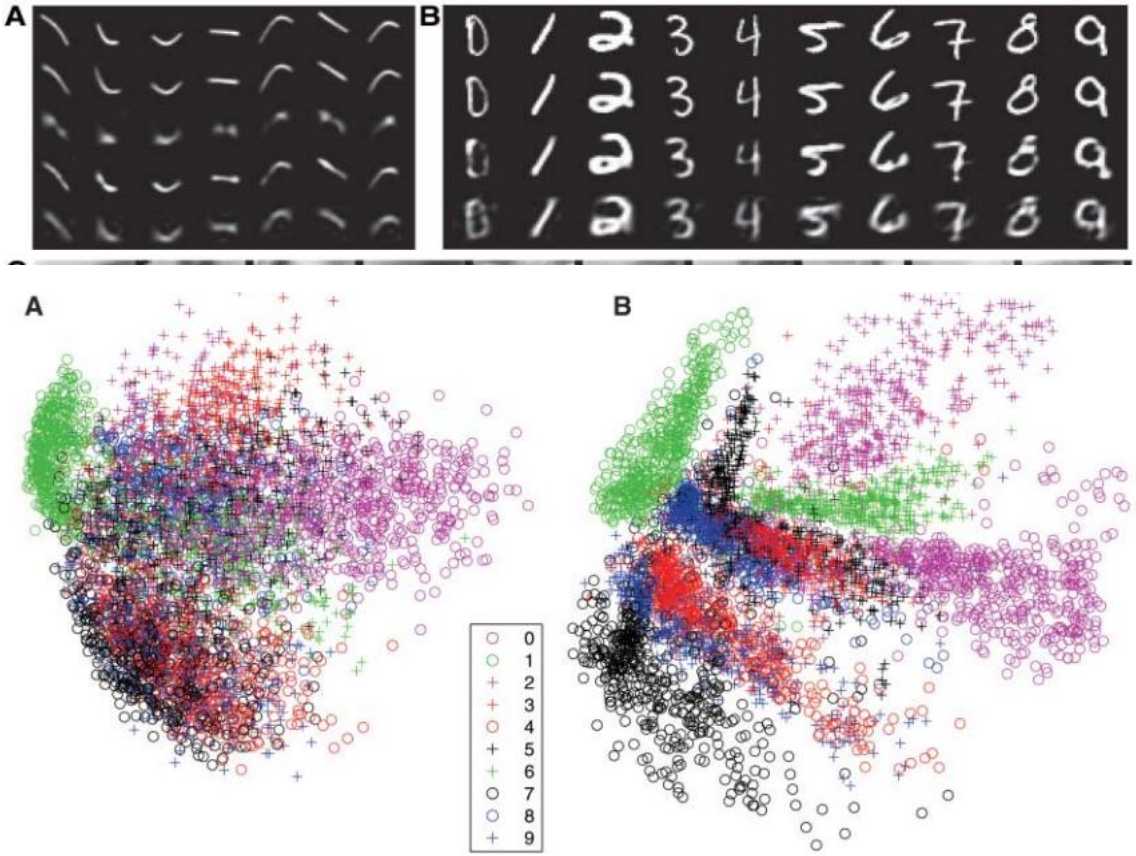




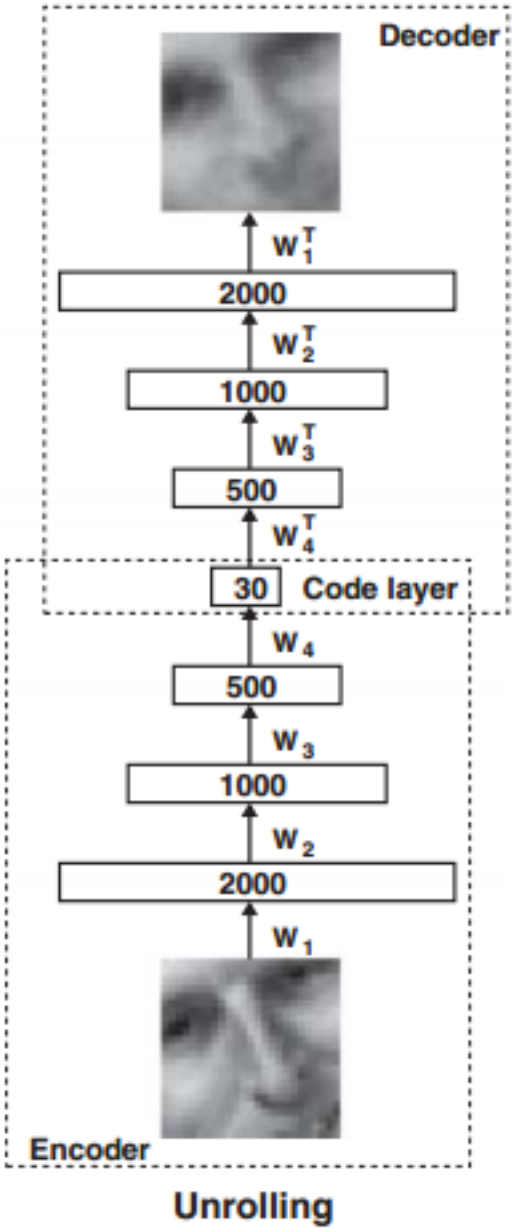


机器学习：新一代计算机技术的浪潮

从浅层学习到深度学习-Lecun



G. Hinton：(A) The twodimensional codes for 500 digits of each class produced by taking the first two principal components of all 60,000 training images. (B) The two-dimensional codes found by a 784-1000-500-250-2 autoencoder. For an alternative visualization, see (8).



深度学习的应用

- 人工智能
- 机器人与自动化
- 智能交通
- 模式识别
- 智慧城市与家居

提纲

- 机器学习的定义
- 机器学习的目的
- 机器学习的一般步骤
- 机器学习的输入
- 机器学习：新一代计算机技术的浪潮
- 机器学习 vs. 数据挖掘vs.统计学习
- 机器学习的相关期刊会议

Machine learning vs Statistics

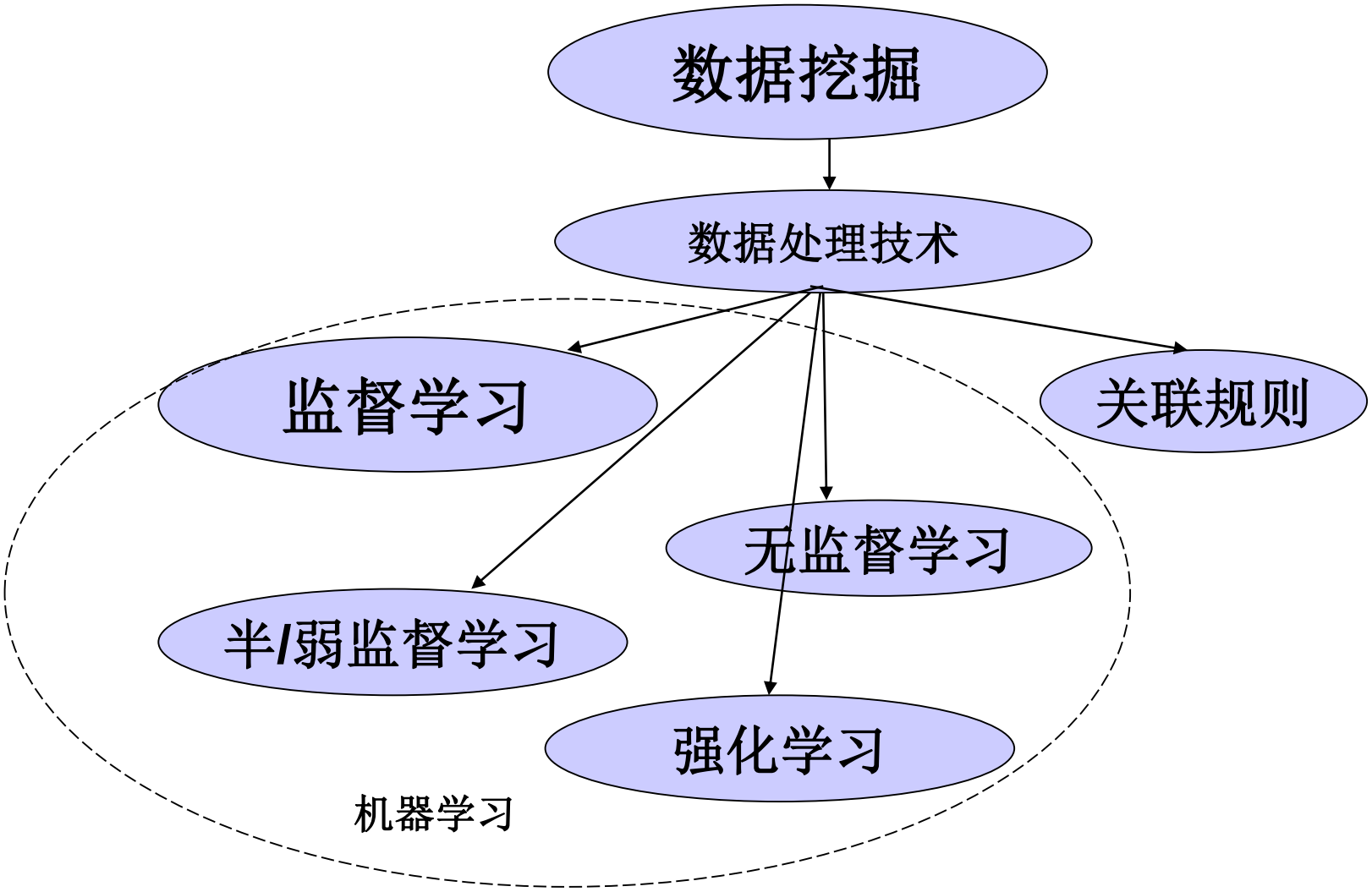
机器学习vs统计理论

If forced to point out a single difference of emphasis, it might be that statistics has been more concerned with testing hypotheses, whereas machine learning has been more concerned with formulating the process of generalization as search through possible hypotheses.

But, this is a gross oversimplification: statistics is far more than hypothesis testing, and many machine learning techniques do mot involve any searching at all.

传统的统计理论侧重于假设检验，机器学习侧重寻找可泛化的模型

机器学习vs数据挖掘



提纲

- 机器学习的定义
- 机器学习的目的
- 机器学习的一般步骤
- 机器学习的输入
- 机器学习：新一代计算技术的浪潮
- 机器学习 vs. 数据挖掘vs.统计学习
- 机器学习的相关期刊会议

机器学习的参考书、期刊与会议

参考书：

1. Machine Learning: a Probabilistic Perspective" by Kevin Patrick Murphy, 2012
2. 《机器学习》，周志华，著，清华大学出版社，2015
3. 斯坦福大学，Andrew N.G. 机器学习开放课程视频.

参考资料：

《计算机学报》《软件学报》《自动化学报》《电子学报》

IEEE International conference on Machine Learning

Machine Learning

Journal of Machine Learning Research

IEEE, Trans. Pattern Analysis and Machine Intelligence

IEEE Trans. Neural Network

机器学习领域的著名学者

Tom Mitchell

CMU
经典机器学习理论

Eric Xing

CMU
Graphic model

Michael Jordan

UCB
Graphic model

机器学习领域的著名学者

朱军

清华大学
非参贝叶斯模型

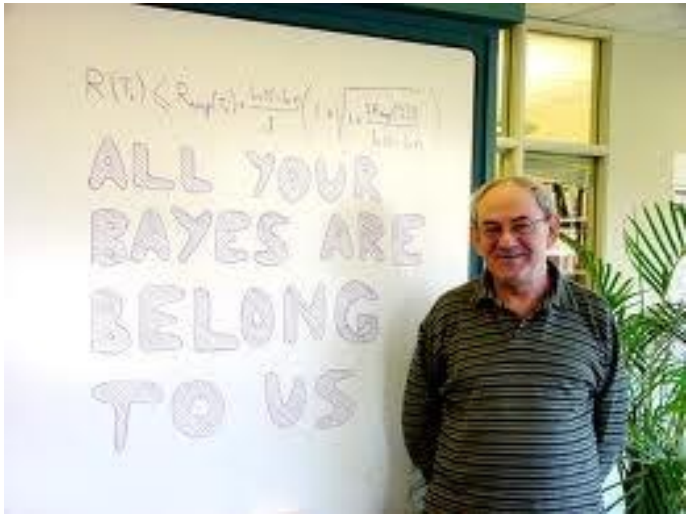
周志华

南京大学
半监督、多标记学习等

于凯、李航

百度、华为
深度学习、排序

机器学习领域的著名学者



Vapnik

支撑向量机

Thorsten Joachims

Cornell
结构化支撑向量机

林智仁

台湾大学
LibSVM, LibLinear

机器学习领域的著名学者

Geoffrey E. Hinton

Toronto University
深度学习

Andrew N.G

Stanford, Baidu
无监督学习

Lecun Yann

NCU, Facebook
深度学习

机器学习应用，10年后我们的梦想

机器“人”尚不具备真正的智能

机器学习是解决这个问题的途径之一

http://v.ku6.com/show/v0tkdFOMAEhJxWmKdXgjzA...html

机器学习应用，10年后我们的梦想

让摄像机具有自主学习功能，长上眼睛（Self-learning Camera)

机器学习应用，10年后我的梦想

让机械手臂具有自主学习功能（Self-learning Mechanical Arms)

Thanks