绪论

挖掘数据的金矿

我是王校长

- 清华大学物理系博士生
- 清华大学交叉信息研究院访问学生
- 原百度系统部实习研发工程师
- 原异构智能公司实习研发工程师

预备知识

课前

- Coursera吴恩达《机器学习》WEEK 1

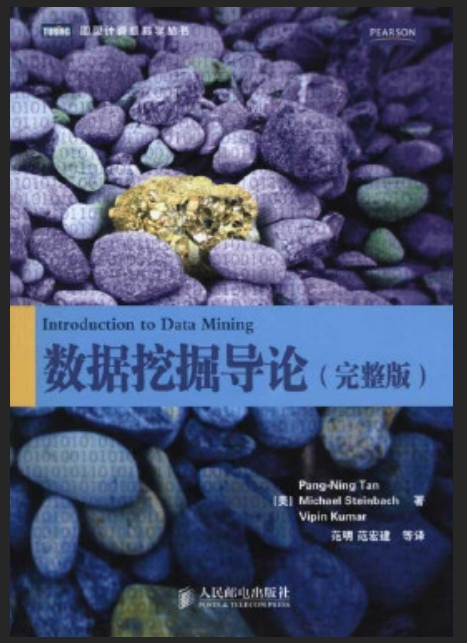
https://www.coursera.org/learn/machine-learning

- 《数据挖掘导论》第1、2章（选读第3章）

课后

- 学堂在线《数据挖掘：理论与算法》WEEK 1、2

http://www.xuetangx.com/courses/course-v1:TsinghuaX+80240372X+2016_T2/about



数据挖掘的起源

大数据技术的出现

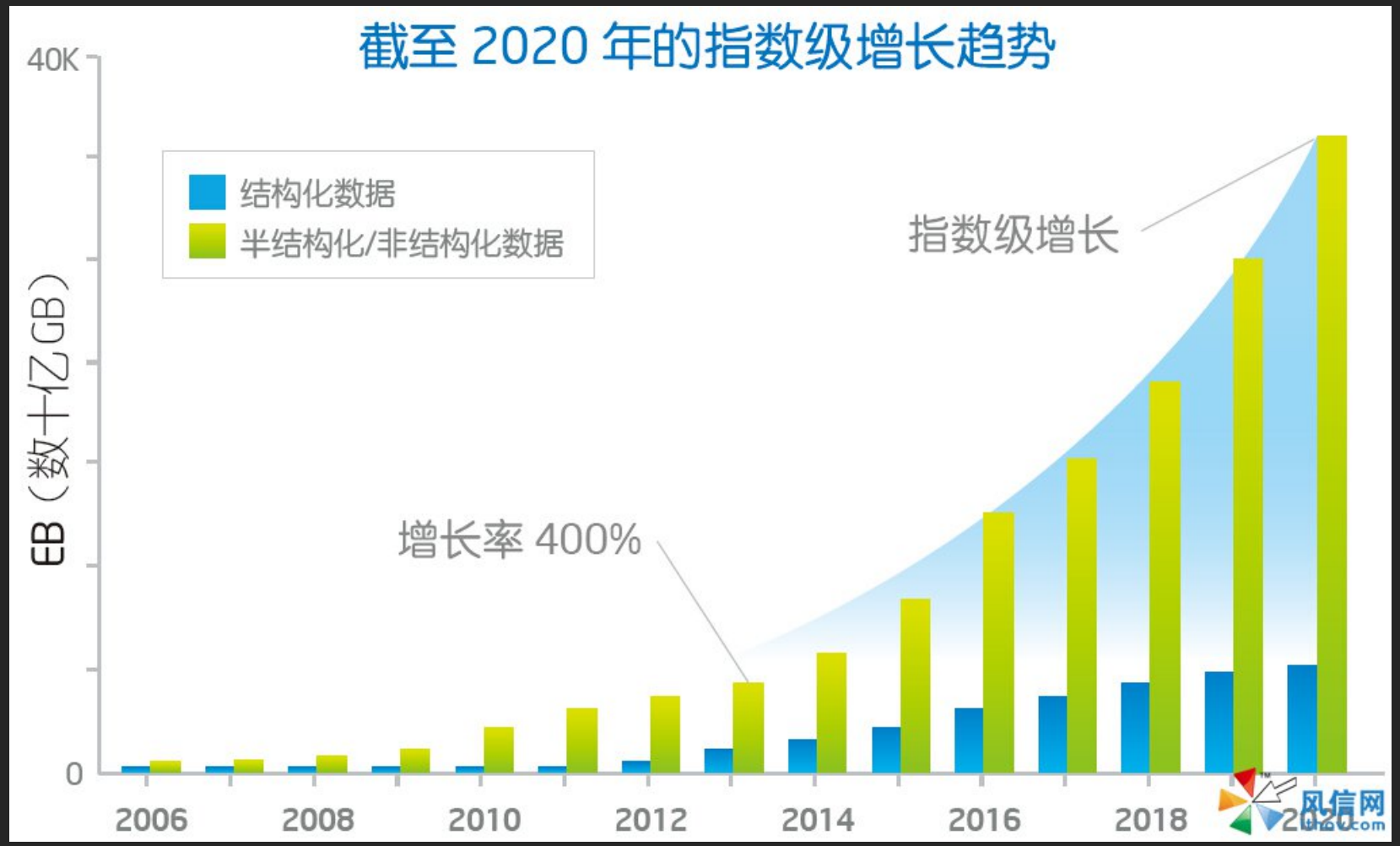
- 数据的涌现

4V – Volume（容量）; Variety（多样性）; Value（价值）; Velocity（速度）

- 计算能力的提升

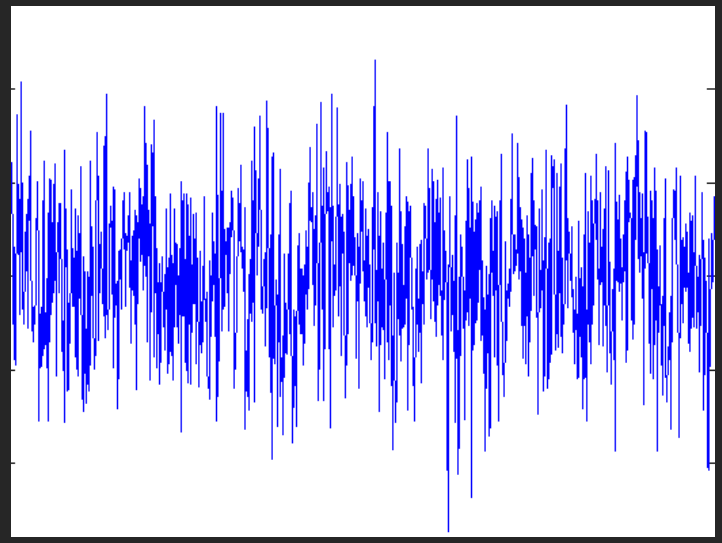
CPU; GPU; 异构计算; 分布式集群

Volume

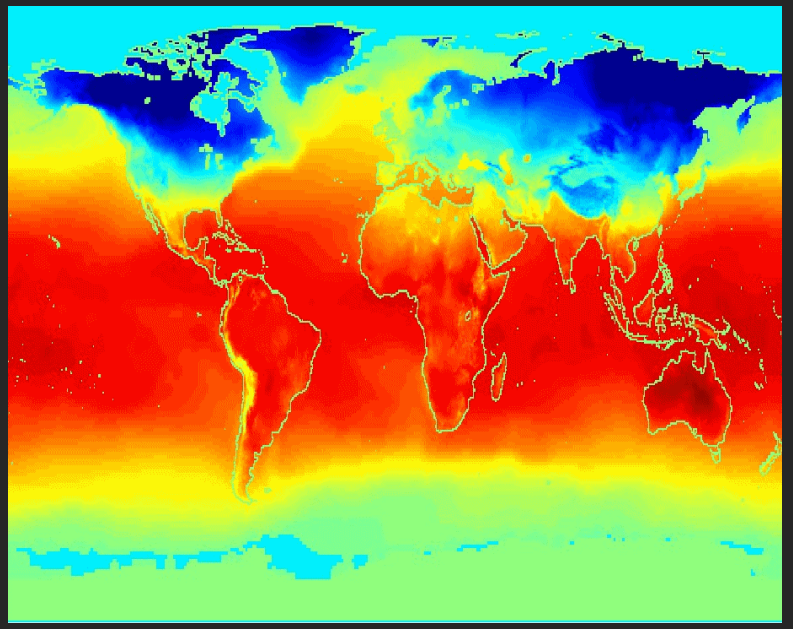

截至 2020 年的指数级增长趋势
40K
0
EB（数十亿GB）
结构化数据
半结构化/非结构化数据
指数级增长
增长率 400%
2006
2008
2010
2012
2014
2016
2018
2020

Variety

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

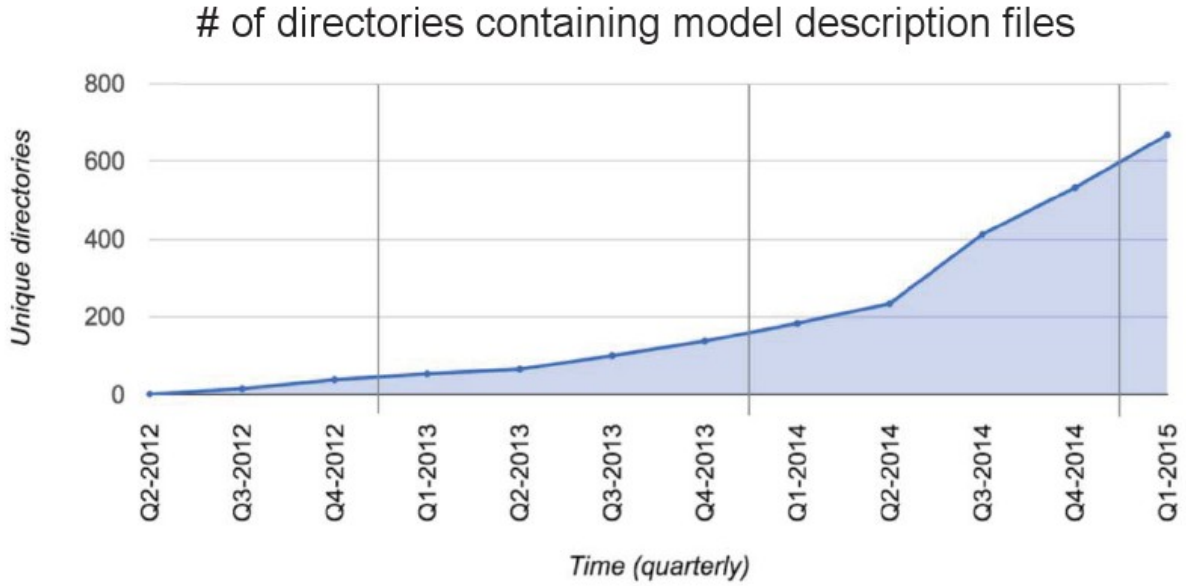


CGCAGGGCCCGCCCCGCGCCGT
CGAGAAGGGCCCGCCTGGCGG
GCGGGGGGAGGCGGGGCCGCC
CGAGCCCAACCGAGTCCGACCA
GGTGCCCCCTCTGCTCGGCCTAG
ACCTGAGCTCATTAGGCGGCAG
CGGACAGGCCAAGTAGAACAC
GCGAAGCGCTGGGCTGCCTGCT



Value

Growing Use of Deep Learning at Google

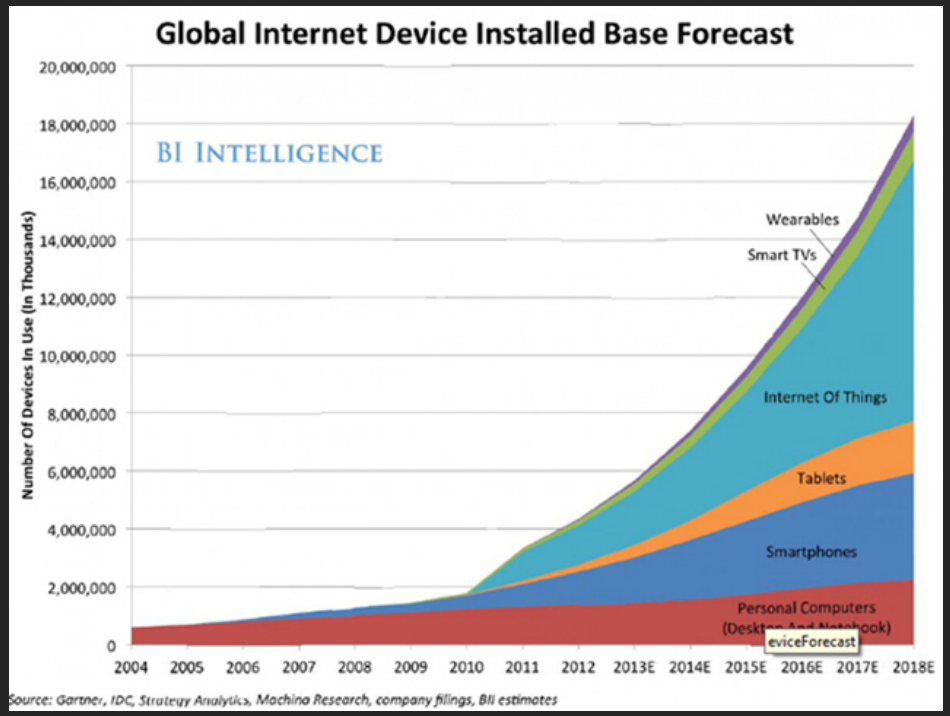


Across many products/areas:

Android
Apps
GMail
Image Understanding
Maps
NLP
Photos
Robotics
Speech
Translation
many research uses..
YouTube
… many others ...

Velocity

- 人 – Facebook; Twitter; 微信; 微博; ……
- 手机 – GPS; WIFI; 蓝牙; 短信; 电话; ……
- 仪器 – 天文望远镜; 测序仪; ……
- IoT – 各种传感器

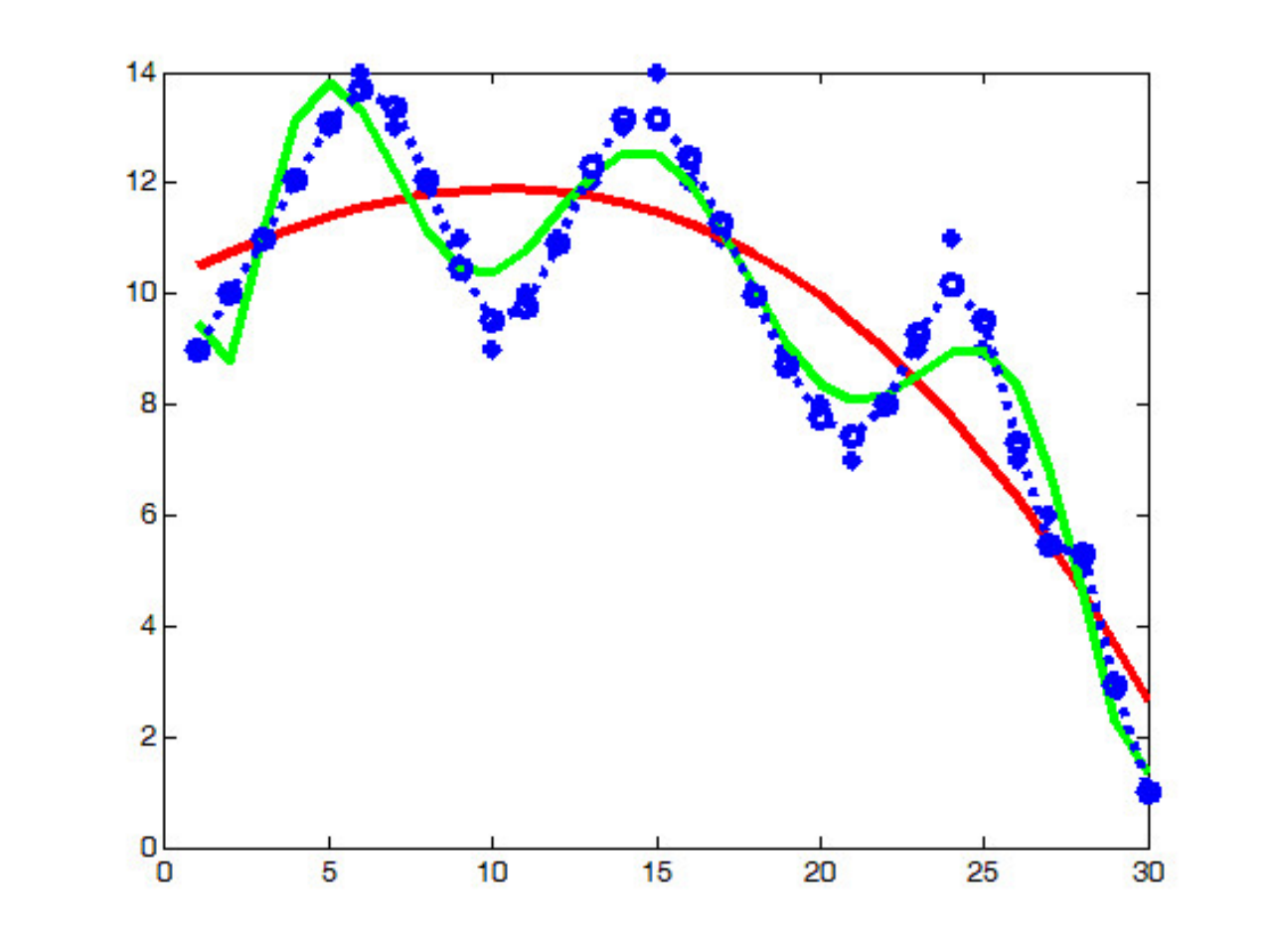


分布式集群 – Data Center As A Computer

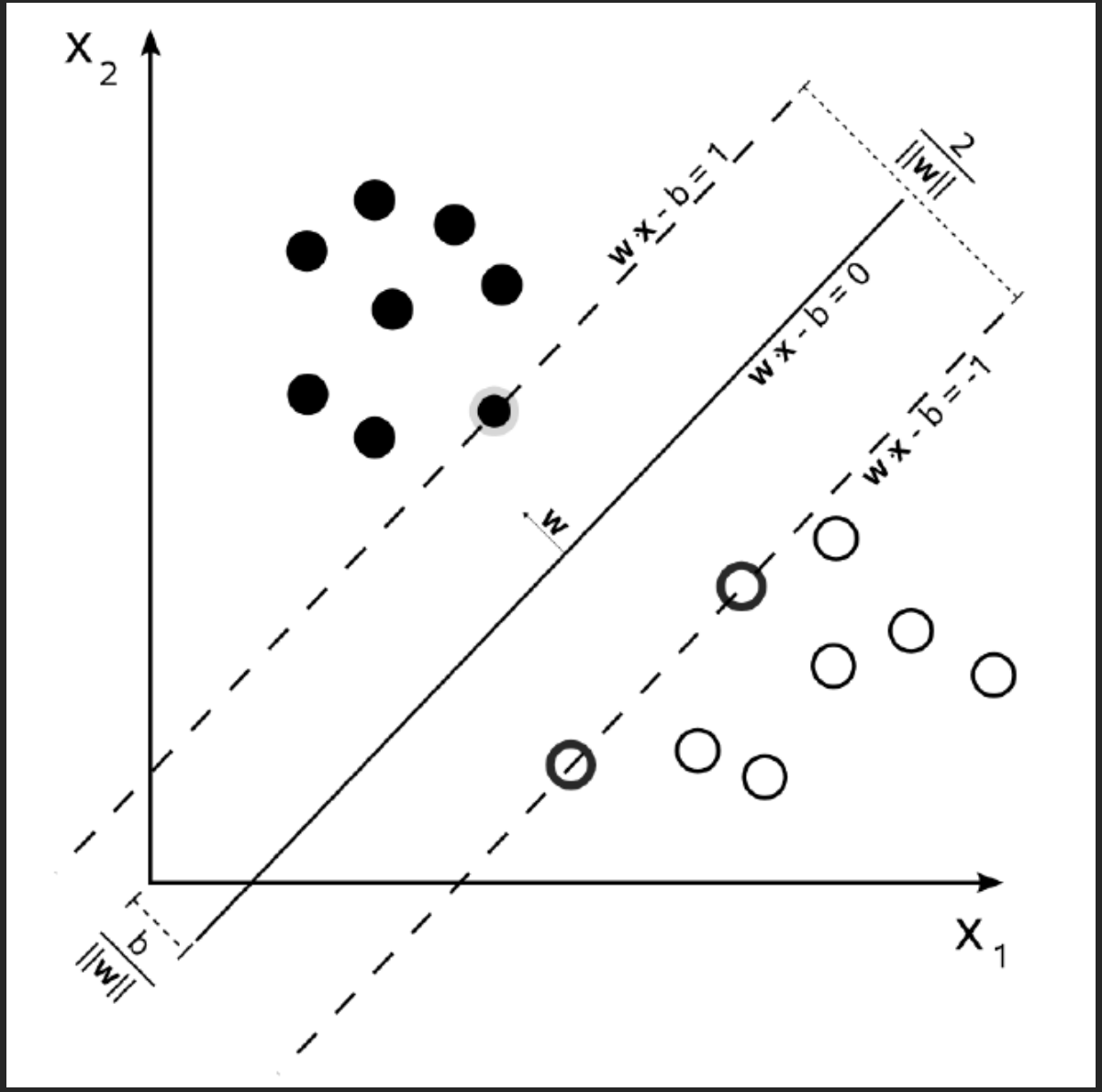
典型的数据挖掘问题

- 拟合 – Regression
- 分类 – Classification
- 聚类 – Clustering
- 异常检测 – Anomaly Detection

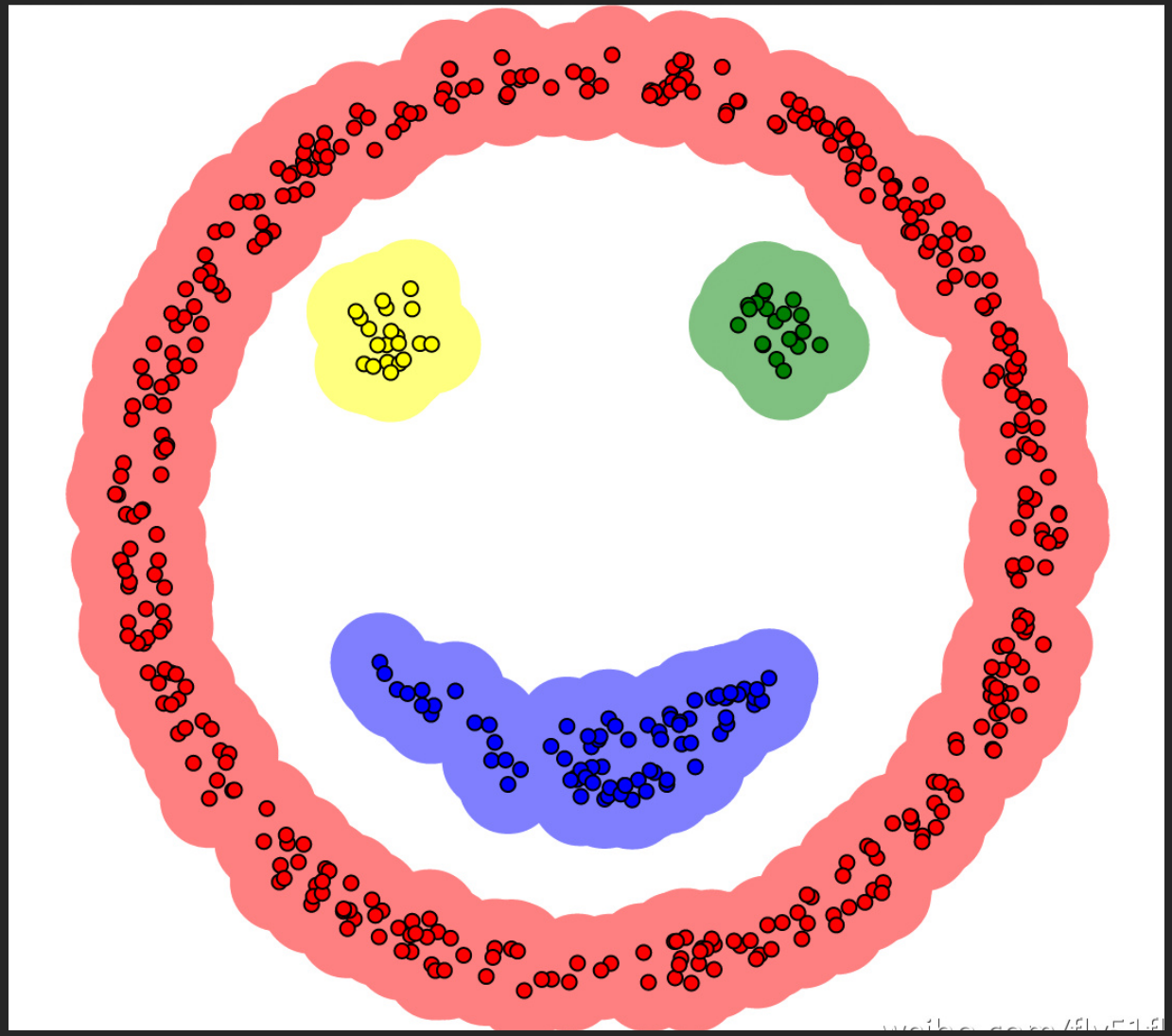
拟合



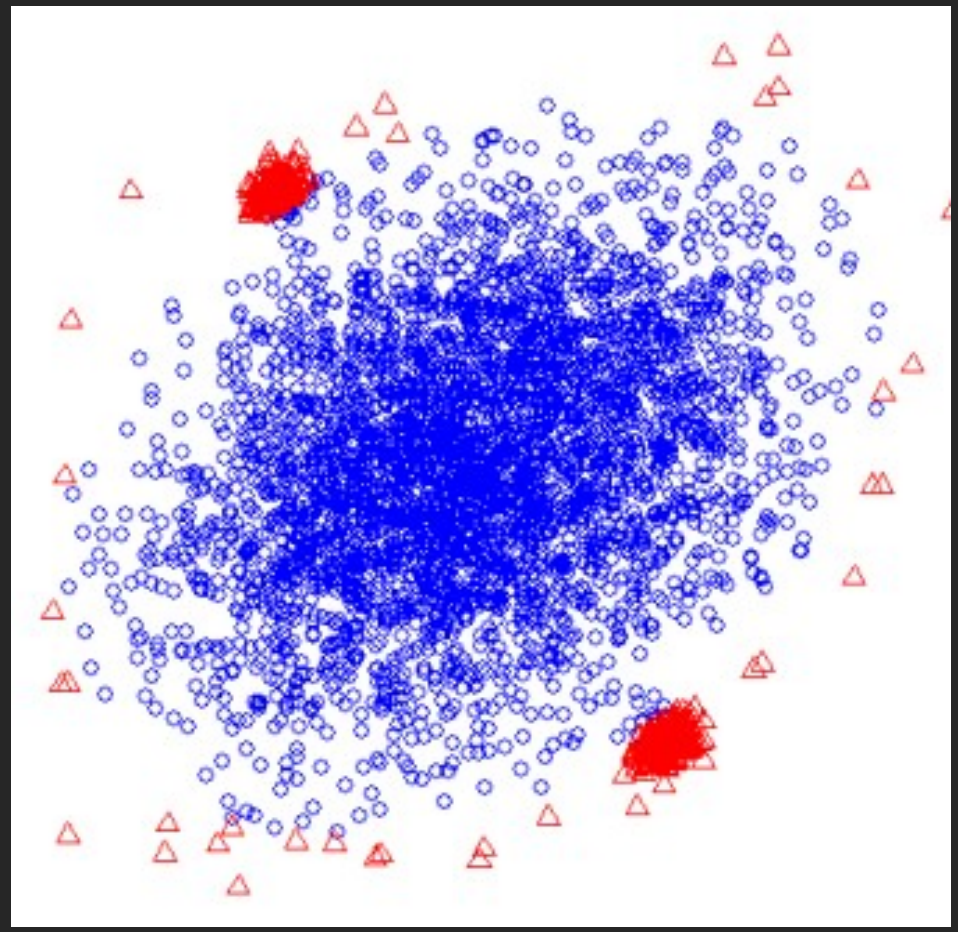
分类



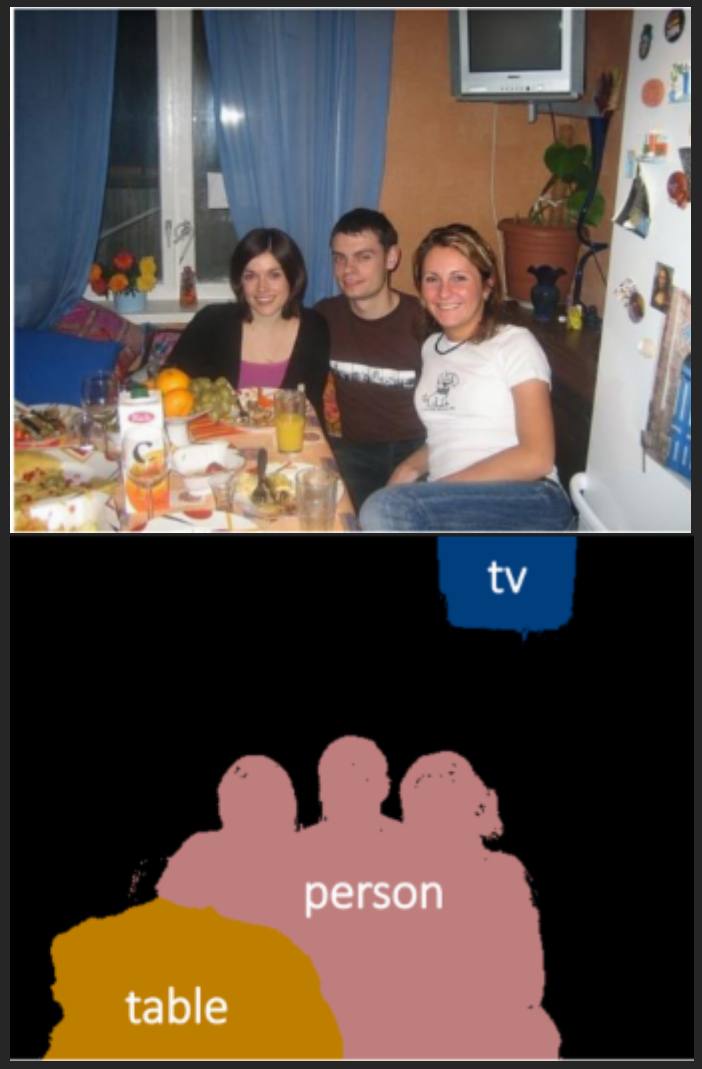
聚类

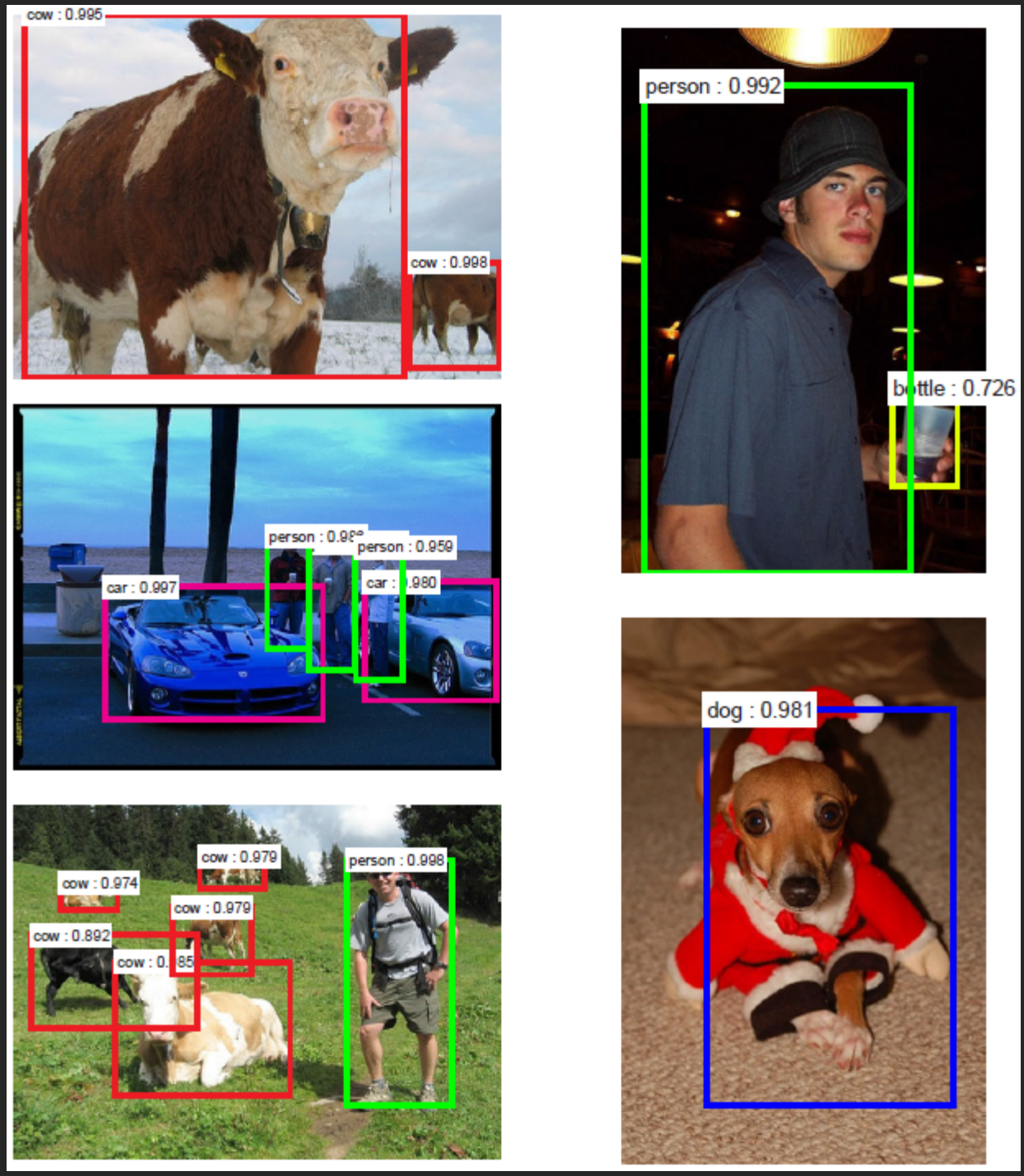


异常检测



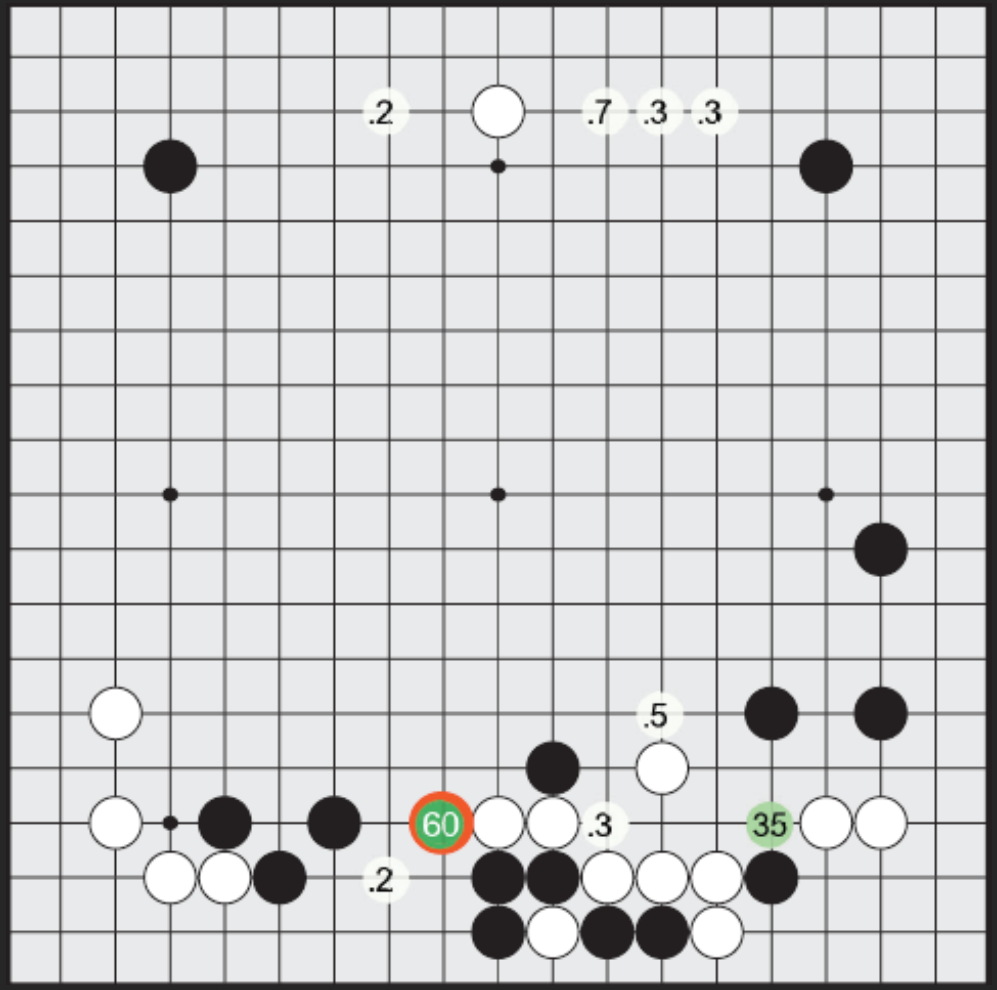
复杂任务


tv
person
table


cow : 0.995
cow : 0.998
person : 0.992
ttle : 0.726
car : 0.997
person : 0.959
car :
cow : 0.974
cow : 0.979
cow : 0.979
cow : 0.892
person : 0.998
dog : 0.981

数据挖掘的流程

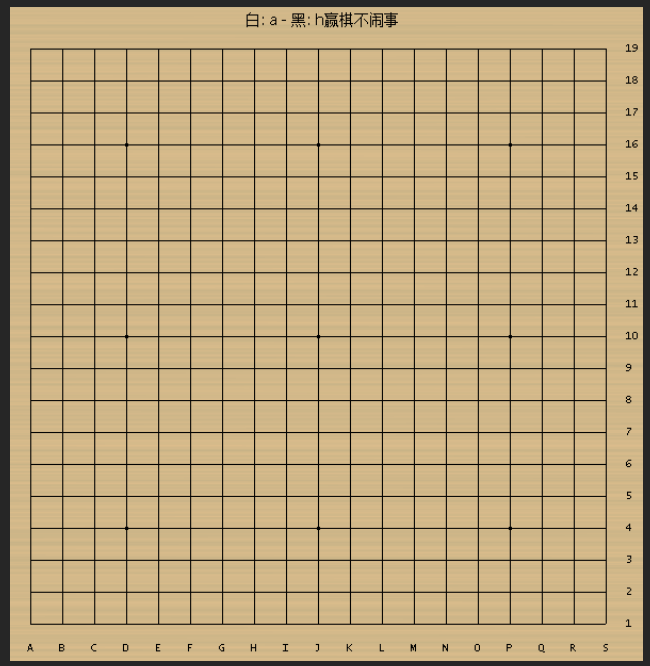
Alpha Go策略网络模型的训练



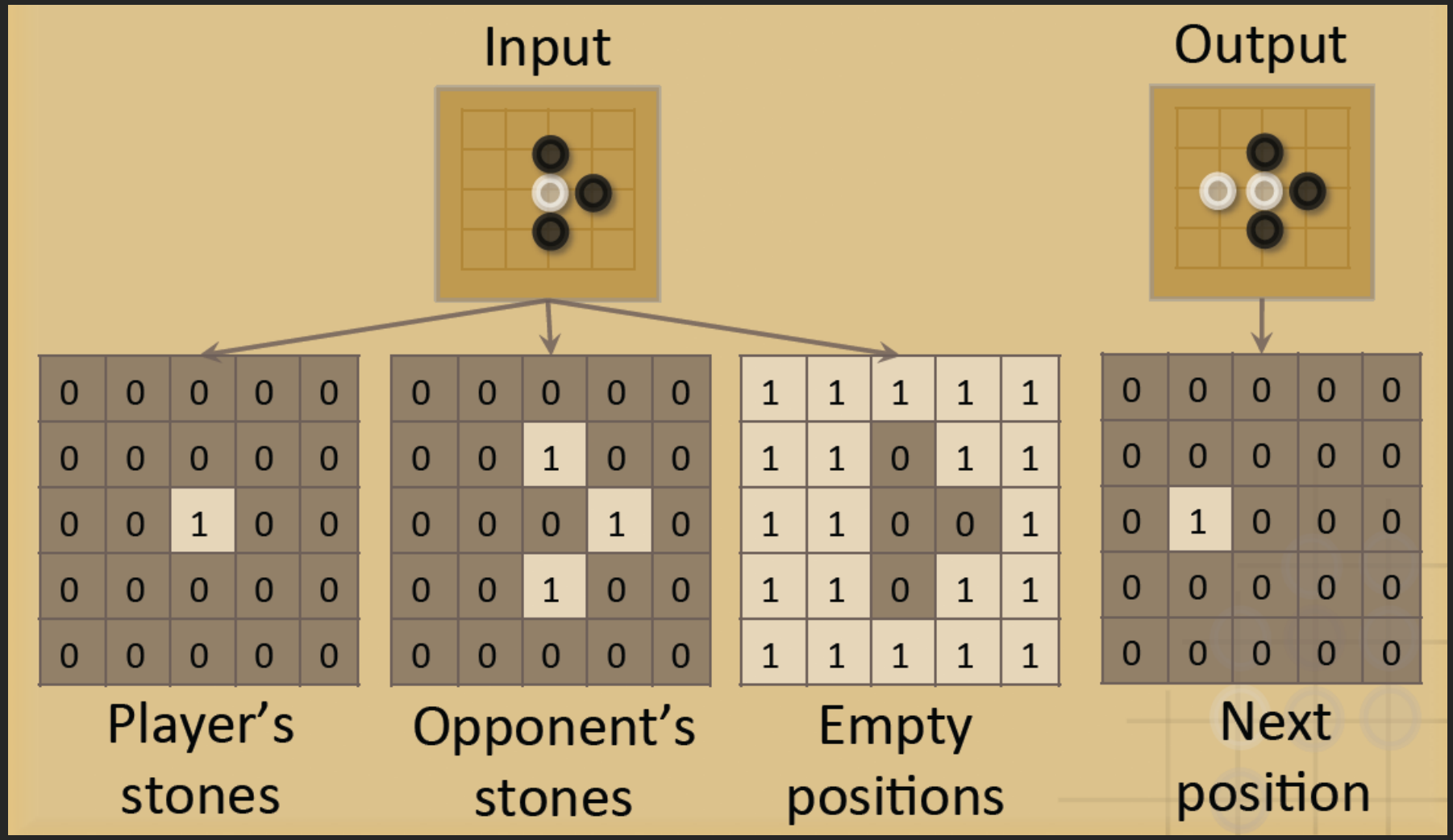
收集数据

- KGS对战平台 – 16W 6-9 dan的对战数据
- 数据格式 – SGF 文件

(;GM[1]FF[4]CA[UTF-8]AP[CGoban:3]ST[2]RU[Japanese]SZ[19]KM[0.50]TM[0]OT[5x10 byo-yomi]PW[bisushield]PB[cheater]WR[9d]BR[8d]DT[2016-10-04]PC[The KGS Go Server at http://www.gokgs.com/]C[cheater [8d\]: hibisushield [9d\]: hi]RE[W+Resign];B[pd]BL[10]OB[5];W[dp]WL[10]OW[5];B[qp]BL[10]OB[5];W[dd]WL[10]OW[5];B[nq]BL[10]OB[5];……)



数据清洗


Input
Output
0 0 0 0 0
0 0 0 0 0
0 0 1 0 0
0 0 0 0 0
0 0 0 0 0
0 0 0 0 0
0 0 1 0 0
0 0 0 1 0
0 0 1 0 0
0 0 0 0 0
1 1 1 1 1
1 1 0 1 1
1 1 0 0 1
1 1 0 1 1
1 1 1 1 1
0 0 0 0 0
0 0 0 0 0
0 1 0 0 0
0 0 0 0 0
0 0 0 0 0
Player's stones
Opponent's stones
Empty positions
Next position

领域知识

Input

Stone color:

3 planes

player, opponent, empty

0	0	0
0	1	0
0	0	0

0	1	0
0	0	1
0	1	0

1	0	1
1	0	0
1	0	1

Liberty:

8 planes

1~8 liberties

0	0	0
0	1	0
0	0	0

0	0	0
0	0	0
0	0	0

0	1	0
0	0	1
0	1	0

0	0	0
0	0	0
0	0	0

0	0	0
0	0	0
0	0	0

0	0	0
0	0	0
0	0	0

0	0	0
0	0	0
0	0	0

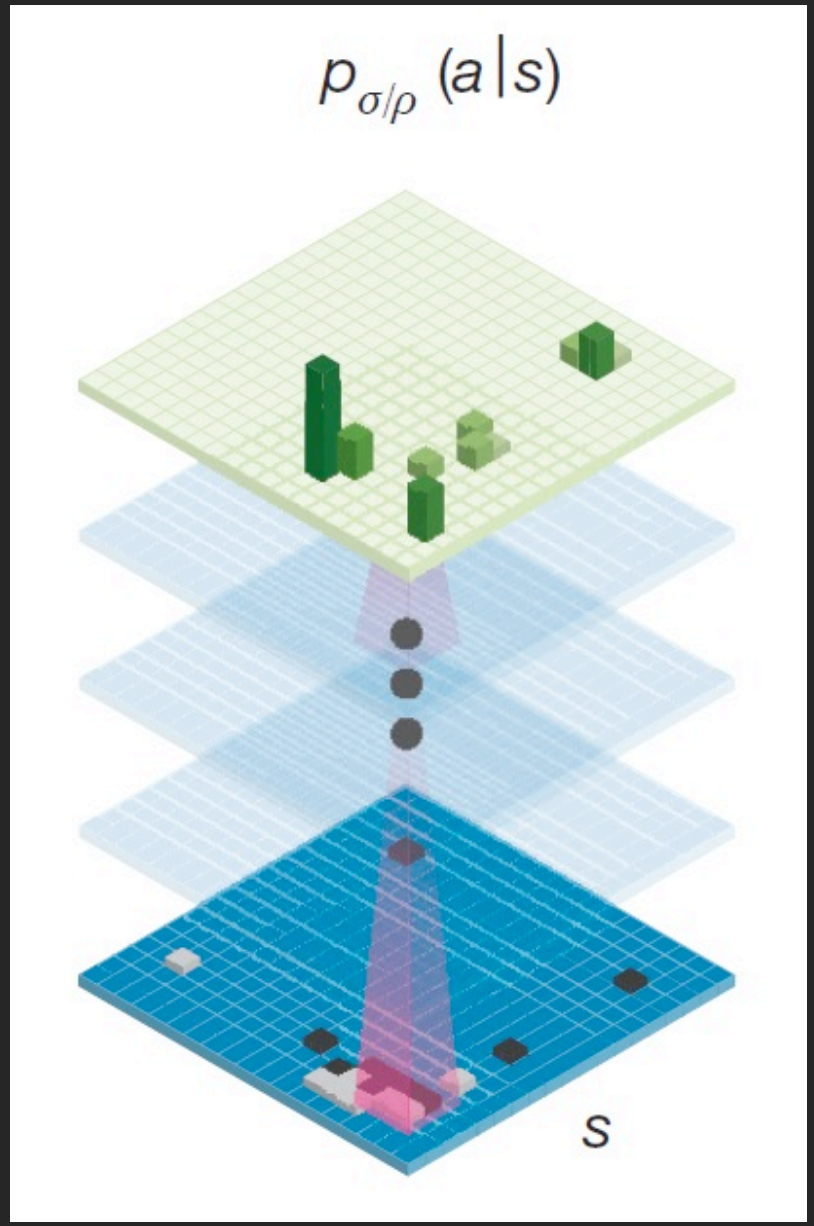
0	0	0
0	0	0
0	0	0

模型训练

训练集

验证集

测试集



模型调优

首先确认训练数据是没问题的！

- 训练误差大？

更大的模型; 训练更长时间; 新的模型结构; ……

- 验证误差大？

更多的数据; 正则化; 新的模型结构; ……

- 测试误差大？

更多的数据; ……

常见的问题

- 数据量不够怎么办？

尝试收集更多的数据; 使用数据放大的方法合成新数据; ……

- 不知道选择什么模型？

明确问题类型; 使用前人经验; 尝试最新的技术; ……

- 深度学习不能做什么？

人类思考时间大于1s的事情，仅靠目前的深度学习技术无法解决！

数据的诅咒

- 统计的前提 – 样本足够多

大数定理 - 样本数量很大的时候，样本均值和真实均值充分接近

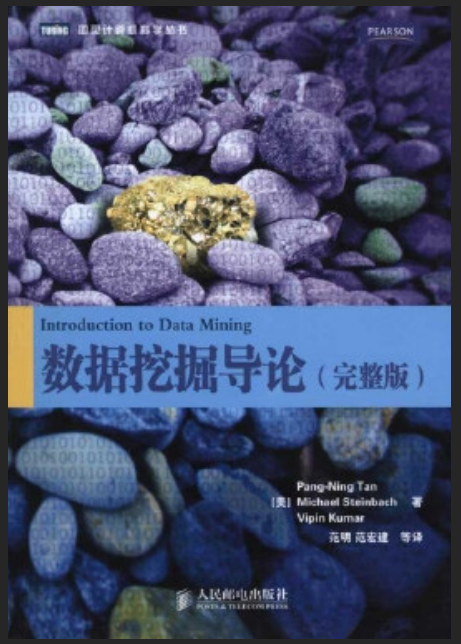
- 平均数与中位数

平均数往往具有欺骗性，柱状图最靠谱！

- 关联与因果关系

关联不等于因果关系，很多时候我们不用知道为什么！

Next Class – 拟合与优化



课前

- Coursera吴恩达《机器学习》WEEK 2、3

https://www.coursera.org/learn/machine-learning

- 《数据挖掘导论》附录D、E

课后

- 网易公开课《机器学习》第2、3课

http://open.163.com/special/opencourse/machinelearning.html