

大数据解析与应用导论

Introduction to Big Data Analytics and Application

授课教师: 赵春晖

联系方式:

Email: chhzhao@zju.edu.cn

Phone: 13588312064

Room: 工控新楼308室



第一部分：课程综述

- **课程地位** 控制学院专业选修课（大三本科生）
- **预修课程**
 - 概率论与数理统计（大二秋冬）
 - 线性代数（大一秋冬）
- **学分2，共8周，课时4/周**



第一部分：课程综述

▣ 考核方法

➤ 平时（50%），个人为主体

讨论发言（课堂表现、交流）、3次课外作业、考勤（随机3次）：

- （1）3次随机点名，每次3分，缺席一次扣3分，满分10分
- （2）3次课外作业，每次10分，满分30分
- （3）课堂表现，10分

➤ 期末大作业（50%），不苟求个体作战

一个具体案例的技术报告和demo演示，包括期末成果展示效果、工作量。



第一部分：课程综述

□ 教学形式（互动）

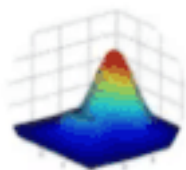
- **教师讲授**（讲授方法的核心思想、答疑、确定讨论主题、布置作业、总结点评等）；**强调互动**
- **课后组队**（按照课堂讲述和讨论主题推荐参考文献、提供案例数据，分小组讨论解决问题，并作demo展示）
- **课堂表现**（由课堂讨论、作业展示和质疑-应答等若干环节组成，学生在讨论中如能提出创新思想，则会在其绩效记录中有所体现）

第一部分：课程综述

□ 教学目标

概括认识 实际应用

模型



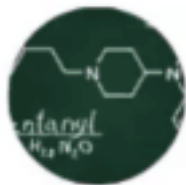
根据不同的问题，如何选择合适的模型

策略



根据具体的模型，选择什么样的评判准则进行评优

算法



针对合适的模型，如何用合适的计算方法求最优解



第一部分：课程综述

□ 为什么在本科生阶段开设这门课？

“大数据”这个概念几乎应用到了所有人类智力与发展的领域中。

直接就业：各大企业实际生产经营决策过程中都产生了丰富的数据，对数据产生价值的利益驱动转化为对数据分析数据挖掘人才的迫切需求。


继续深造：对于数据挖掘理论方法的深入研究并用于解决实际问题已成为学术界研究热点。可以为对数据分析数据挖掘方法进一步的深入研究和学习带下坚实的基础。

实际需要：生活中到处都有数据挖掘，数据挖掘无处不在



我们可能每天都在用 数据挖掘

类别 (Class)	目的 (Purpose)
数据 (Data)	Know-Nothing (一无所知)
信息 (Information)	Know-What (知道是什么)
知识 (Knowledge)	Know-How (知道是怎样)
智慧 (Wisdom)	Know-Why (知道是为何)



人类从依靠自身判断做决定到依靠数据做决定的转变，也是大数据作出的最大贡献之一。——《大数据时代》



第一部分：课程综述

□ 与研究生课程《实用多元统计分析》的区别

- **针对研究生：**学习的理论深度更强，侧重于理论方法的推导以及理论深度的提高，学习的面不一定宽，但更深入（结合自己的研究课题进行针对性的研究，如何结合具体研究对象及其特性对理论方法进行提升）；（矩阵论等知识）
- **针对本科生：**弱化理论，侧重于开阔视野和知识面(**大数据背景**)，建立一种数据分析的思维体系，以及运用基本方法去解决各种实际问题的能力（**具体操作实践能力**）。



第一部分：课程综述

□ 参考书目

1. 《大数据分析挖掘》（适合作为课程教材）
<https://item.jd.com/32856815714.html>
2. 《Python数据分析与挖掘实战》（适合课后实验参考）
<https://item.jd.com/31373896774.html>

其他参考资料：

1. 《机器学习》 周志华著 清华大学出版社，2016年1月。
2. 《工业过程运行状态智能监控：数据驱动方法》 赵春晖，王福利，化工出版社，2019年2月出版。
3. 《间歇过程统计监测与质量分析》 赵春晖，陆宁云，科学出版社，2014年10月出版。
4. 《大数据挖掘与统计机器学习》(大数据分析统计应用丛书)吕晓玲、宋捷著中国人民大学出版社，2016年7月。
5. 《实用多元统计分析》，王学仁、王松桂著，上海科技出版社，1990年6月
6. 《多元统计分析-写于大数据.云计算时代》李庆来著，上海交通大学出版社，2015年9月。
7. 朱道元等编，《多元统计分析与软件SAS》，东南大学出版社，1999年8月。。。



第二部分：网上实战案例

纸上得来终觉浅，绝知此事要躬行



大数据比赛推荐

天池:

<https://tianchi.aliyun.com/>

Ai challenger:

<https://challenger.ai/>

Kaggle:

<https://www.kaggle.com/>

知名数据竞赛



<https://www.kaggle.com/>

2010年成立于墨西哥

2017年被谷歌收购



<https://tianchi.aliyun.com/>

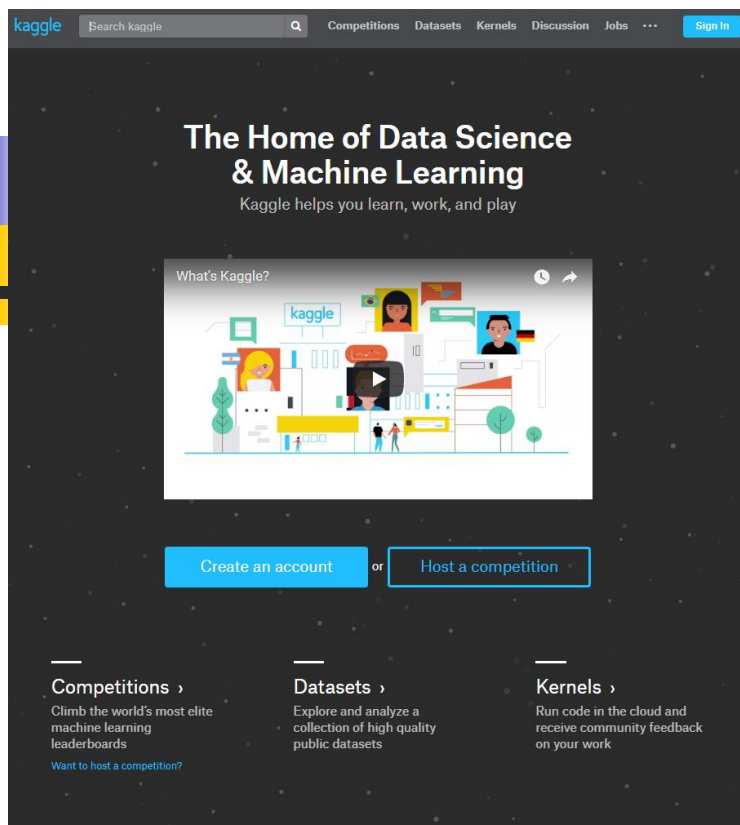
2014年由马云发起

140000+ 数据开发者

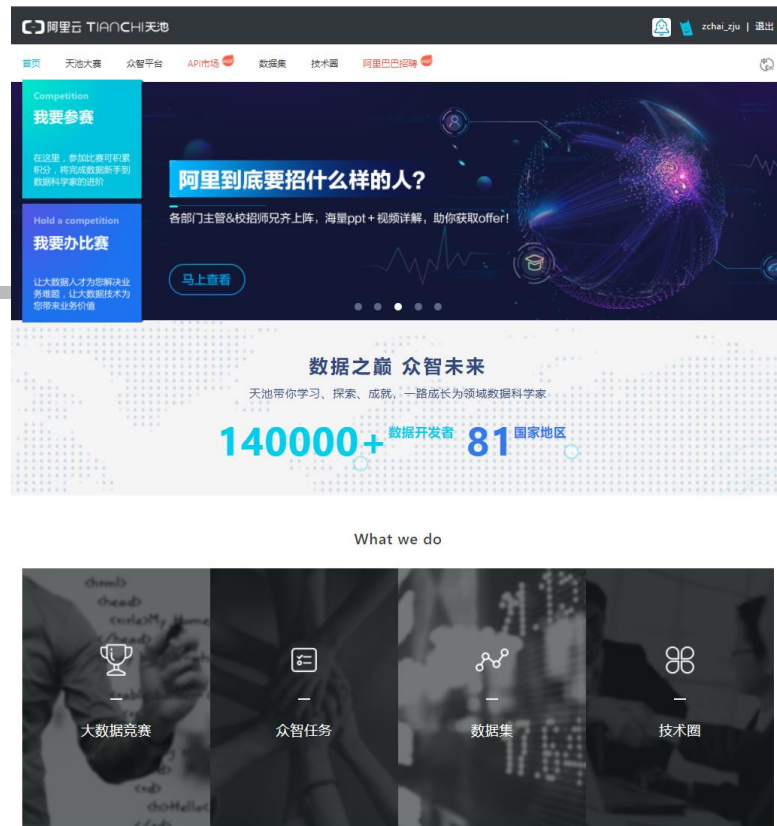
81 国家地区

知名数据竞赛

<https://www.kaggle.com/>



<https://tianchi.aliyun.com/>



阿里云天池大数据竞赛平台

<p>TIANCHI 天池</p> <p>探索智能天体分类 开启奇幻天文之旅</p> <p>天文数据挖掘大赛</p> <p>2018-1-19</p>	<p>TIANCHI 天池</p> <p>2017印象盐城 数创未来大数据竞赛</p> <p>印象盐城 数创未来</p> <p>2017-12-18</p>	<p>TIANCHI 天池</p> <p>聚焦光电制造质量智能预测</p> <p>天池工业AI大赛</p> <p>2017-12-07</p>	<p>TIANCHI 天池</p> <p>气象数据领航无人飞行器线路优化</p> <p>英国气象局天池大赛</p> <p>2017-10-30</p>
<p>TIANCHI 天池</p> <p>限量天池好礼共分享</p> <p>天池数据英雄征集令</p> <p>2017-9-13</p>	<p>TIANCHI 天池</p> <p>"数据引领 飞粤云端"</p> <p>广东政务数据创新大赛</p> <p>2017-9-5</p>	<p>TIANCHI 天池</p> <p>当最强算法遇上Hacker</p> <p>第二届阿里云安全算法挑战赛</p> <p>2017-8-12</p>	<p>TIANCHI 天池</p> <p>"数聚华夏 创享未来" 中国数据创新行</p> <p>智慧交通预测挑战赛</p> <p>2017-8-12</p>
<p>TIANCHI 天池</p> <p>挑战双十一万亿级消息引擎</p> <p>第三届阿里中间件性能挑战赛</p> <p>2017-6-13</p>	<p>TIANCHI 天池</p> <p>电改创造有为时代</p> <p>"智造扬中" 电力AI大赛</p> <p>2017-7-10</p>	<p>TIANCHI 天池</p> <p>天池医疗AI大赛</p> <p>第一季：肺部结节智能诊断</p> <p>2016-10-23</p>	<p>TIANCHI 天池</p> <p>数据引领，飞粤云端</p> <p>广东航空大数据创新大赛</p> <p>2016-10-23</p>
<p>TIANCHI 天池</p> <p>当最强算法遇上Hacker</p> <p>阿里云安全算法挑战赛</p> <p>2016-10-15</p>	<p>TIANCHI 天池</p> <p>云端智慧物流</p> <p>菜鸟网络全球算法大赛</p> <p>2016-9-18</p>	<p>TIANCHI 天池</p> <p>首届基因组云计算技术开发者大会</p> <p>华大基因风云挑战赛</p> <p>2016-08-05</p>	<p>TIANCHI 天池</p> <p>公益云图</p> <p>可视化暑期学校报名啦！</p> <p>2016-07-22</p>
<p>TIANCHI 天池</p> <p>公益云图</p> <p>数据可视化创新大赛</p> <p>2016-06-02</p>	<p>TIANCHI 天池</p> <p>全国巡回技术沙龙启动</p> <p>中间件性能挑战大赛</p> <p>2016-05-17</p>	<p>TIANCHI 天池</p> <p>你敢来挑战吗？</p> <p>天池竞赛挑战baseline</p> <p>2016-03-05</p>	<p>TIANCHI 天池</p> <p>邀请同学打比赛有礼了</p> <p>"全民星探" 计划</p> <p>2016-03-05</p>

合作方：

国家天文台

英国国家气象

广东省政府

厦门航空

华大基因

菜鸟、口碑

More...

重要通知

最终数据评分规则说明

在初赛阶段，为方便参赛选手训练、评估模型，组委会提供了风机的全部数据。通过比较选手上传的数据结果，组委会发现部分选手在“齿形带断裂故障预测”问题所提交的结果中出现了部分非正常高分...

>>

大赛介绍 >

- 工业和信息化部指导、中国信息通信研究院主办的全国性的工业大数据应用实践赛事
- 美国自然科学基金会智能维护系统中心特别支持
- 基于企业真实数据源，以竞赛模式求解工业大数据实际应用问题

本届主题：

- 风力发电机故障预测性维护

比赛项目：

- 风机叶片结冰预测大赛
- 风机齿形带故障分类大赛

[查看更多>](#)

我要报名

赛程安排 >



比赛题目一：风机叶片结冰预测大赛



赛事简要：低温环境所导致的叶片结冰问题是风力发电设备维护面临的一个全球范围难题。叶片附着较大质量的冰层，会改变风机叶片的共振频率，进而改变其动态响应行为，造成叶片断裂的风险。风机的SCADA系统普遍会具备结冰探测和除冰系统，当实际功率与理论值严重不符时触发报警。但是多数情况下当系统触发报警时结冰现象已很严重，风机的运行已经存在很大风险。风机的叶片结冰故障预测期望选手可以提出更好的算法，提高预测准确度，从而提高除冰系统的效率，降低风机的效率损失和风机运行的风险。

比赛题目二：风机齿形带故障分类大赛



赛事简要：变桨系统是保障风机按照设计的功率曲线运行的核心系统，在使用电机驱动的变桨系统中，齿形带是连接电机和叶片的核心传动零部件。由于变桨次数频繁造成材料疲劳的累计速度快，其断裂频次远远高于其他配件。目前在齿形带发生断裂之前很难从SCADA的相关参数中发现明显的迹象，还没有有效手段在齿形带断裂前精确监测和停机，因此提前预测齿形带断裂对于保障风机安全运行具有十分重要的意义。风机的齿形带断裂故障预测期望选手可以提出更好的算法，在故障发生前精确预测，从而指导SCADA系统进行保护性停机和预测性维护。

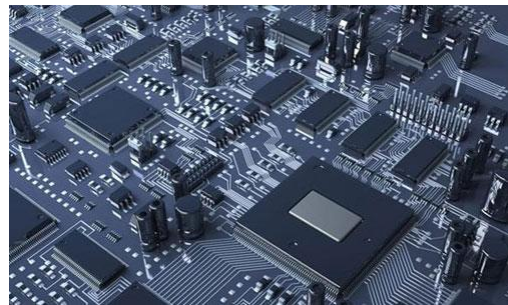
天池工业AI大赛

Tianchi industry AI competition

智能制造质量预测

数据背景:

半导体产业是一个信息化程度高的产业，高度的信息化给数据分析创造了可能性。基于数据的分析可以帮助半导体产业更好的利用生产信息，提高产品质量。在机器学习，人工智能快速发展的今天，我们希望能由机器生产参数去预测产品的质量，来达到生产结果即时性以及全面性。更进一步的，可基于预先知道的结果，去做对应的决策及应变，对客户负责，也对制造生产更加敏感。



官方提示的难点:

- 1) TFT-LCD（薄膜晶体管液晶显示器）的生产过程较为复杂，包含几百道以上的工序。每道工序都有可能对产品的品质产生影响，故算法模型需要考虑的过程变量较多。
- 2) 变量的取值可能会存在异常（如测点仪表的波动导致、设备工况漂移等现象），模型需要足够稳定性和鲁棒性。
- 3) 生产线每天加工的玻璃基板数以万计，模型需要在满足较高的精准度前提下尽可能实时得到预测结果，这样才能给在实际生产中进行使用。

比赛数据介绍

本次比赛为参赛者提供了**500**个半导体生产的样本作为训练数据，每个样本**8028**个属性，一个标签为样本的质量评估值。测试样本**100**个

数据展示：

ID	TOOL_ID	210X1	210X2	210X3	210X4	210X5	210X6	210X7	210X8	210X9	210X10	210X11	210X12	210X13	210X14	210X15	210X16	210X17	210X18	210X19
ID001	N	102.05	0.465	0.27	1.43	67.45	4.62	-0.54	-1.05	-0.13	26.3	27.95	0.532	0.077	0.079	0.078	0.079	750	0.4	0.398
ID002	M	100.95	0.805	0.22	3.477	62.08	3.412	-2.12	1.02	0.08	28.2	24.27	2.653	0.072	0.065	0.073	0.067	750	0.4	0.398
ID003	L	98.56	0.555	0.24	1.172	56.7	3.08	-2.25	0.88	0.17	26.6	24.51	0.523	0.076	0.072	0.077	0.073	750	0.398	0.398
ID004	M	100.35	0.901	0.22	3.631	62.25	3.949	-1.98	0.82	0.08	25.2	24.38	2.582	0.074	0.068	0.074	0.067	750	0.4	0.398
ID005	M	100.25	0.854	0.23	3.429	61.42	3.63	-1.89	1.02	0.08	27.3	24.36	2.535	0.073	0.067	0.074	0.067	750	0.4	0.398
ID006	M	100.55	0.882	0.22	3.462	61.85	3.747	-2.1	0.93	0.08	25.3	24.34	2.434	0.073	0.067	0.073	0.067	750	0.4	0.398
ID007	N	102.2	0.44	0.27	1.429	67.38	4.628	-0.2	-0.44	-0.13	26.5	27.79	0.496	0.077	0.079	0.078	0.079	750	0.4	0.398
ID010	N	102.25	0.42	0.27	1.312	67.47	4.614	-0.99	-0.99	-0.13	27.1	27.9	0.485	0.076	0.08	0.078	0.079	750	0.4	0.398
ID011	N	102.25	0.421	0.27	1.311	67.44	4.631	0.32	-1.19	-0.14	25.9	27.88	0.487	0.078	0.079	0.078	0.079	750	0.4	0.398
ID012	M	100.85	0.958	0.22	3.705	62.88	4.184	-2.14	0.99	0.08	27.7	24.37	2.474	0.073	0.067	0.074	0.067	750	0.4	0.398
ID013	L	98.95	0.525	0.25	1.132	57.14	2.992	-2.04	0.3	0.15	27.3	24.88	0.487	0.077	0.072	0.077	0.072	750	0.398	0.398
ID015	J	101.5	0.312	0.4	0.554	71.63	3.697	0.2	-0.44	-0.16	24.3	41.75	0.349	0.079	0.085	0.079	0.085	750	0.4	0.398
ID018	J	101.45	0.311	0.4	0.56	71.61	3.712	0	-0.61	-0.18	24.9	41.74	0.336	0.079	0.085	0.079	0.085	750	0.4	0.398
ID019	J	101.45	0.314	0.4	0.571	71.67	3.715	0.03	-0.58	-0.16	25	41.73	0.335	0.079	0.085	0.079	0.085	750	0.4	0.398

比赛的真正难点在于，**500**个训练样本和**8028**个样本属性，这是一个典型的小样本高纬度的预测问题，在数据挖掘的过程中特征的提取和选择会成为至关重要的问题



本次大赛旨在通过糖尿病人的**临床数据**和**体检指标**来预测人群的糖尿病程度，以**是否患病**为预测指标。参赛选手需要设计高精度，高效，且解释性强的算法来挑战糖尿病精准预测这一科学难题。

大数据辅助糖尿病精准医疗



回归问题（根据若干体检指标预测病人血糖值）：

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
id	性别	年龄	体检日期	*天门冬氨酸氨基转换酶	*丙氨酸氨基转换酶	*碱性磷酸酶	*r-谷氨酰基转换酶	*总蛋白	白蛋白	*球蛋白	白球比例	甘油三酯	总胆固醇	高密度脂蛋白胆固醇	低密度脂蛋白胆固醇
1	男	41	19/10/2017	24.86	22.1	60.58	20.22	76.98	48.6	33.38	1.92	1.21	4.43	1.27	2.5
3	2男	41	19/10/2017	24.57	36.25	67.21	79	79.43	47.76	31.67	1.51	2.81	4.06	0.93	2.63
4	3男	46	26/10/2017	20.82	15.23	63.69	38.17	86.23	48	38.23	1.26	0.99	4.13	1.64	2.01
5	4女	22	25/10/2017	14.99	10.59	74.08	20.22	70.98	44.02	26.96	1.63	1.06	6.89	1.43	5.04
6	5女	48	26/10/2017	20.07	14.78	75.79	22.72	78.05	41.83	36.22	1.15	0.97	5.37	1.27	3.65
7	6女	74	18/10/2017	23.72	22.59	81.23	23.35	76.46	45.85	30.61	1.5	2.45	6.65	1.81	4.28

→ 血糖值？

分类问题（根据基因表达丰度对妊娠女性是否患糖分类）：

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	
id	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	SNP11	SNP12	SNP13	SNP14	SNP15	RBP4	年龄	孕次	产次	
2	1	2	1	2	3	1	3	3	2	2	2	1	1	1	2	10.92	20	1	1	
3	2	3	2	3	2	2	2	3	3	2	2	3	2	1	1	2	33.84	36	1	1
4	3	1	1	3	1	3	2	3	3	2	2	2	1		3	3	12.4	36	2	2
5	4	1	3	3	1		1	3	1	1	1	2	2	2	2	2	12.87	27	3	1
6	5	1	2	3	2	2	1	2	2	1	2	3	1	1	1	1	26.43	33	4	1
7	6	2	2	3	2	1	2	3	2	1	2	3	2	1	1	2	53.69	38	1	1
8	7	2	2	3	1	3	1	3	3	2	2	1	2	2	3	3	16.14	27	1	1
9	8	3	2	3	2	2	2	3	2	1	2	1	1	1	1	2	20.37	31	5	1
10	9	2	2	3	1	1	1	3	3	2	2	3	3	1	1	2	20.23	37	1	1

→ 是否患糖尿病？

IJCAI
2018

阿里妈妈国际广告算法大赛

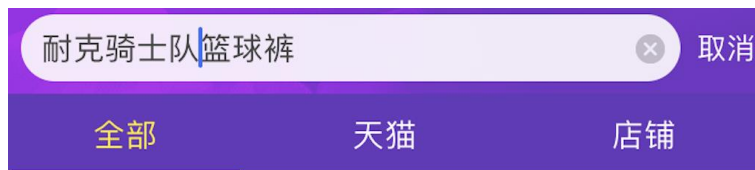
IJCAI-18 ALIMAMA INTERNATIONAL ADVERTISING
ALGORITHM COMPETITION

数据背景：

搜索广告是一种常见的互联网营销方式，商家（广告主）根据商品特点自主购买特定的关键词，当用户输入这些关键词时相应的广告商品就会展示在用户看到的页面中。

例如：

1.在淘宝搜索 “耐克骑士队篮球裤”



2.当你再次打开淘宝时，你的首页会出现相关的产品



搜索广告的转化率，即广告商品被用户点击后产生购买行为的概率，是衡量广告转化效果的指标。

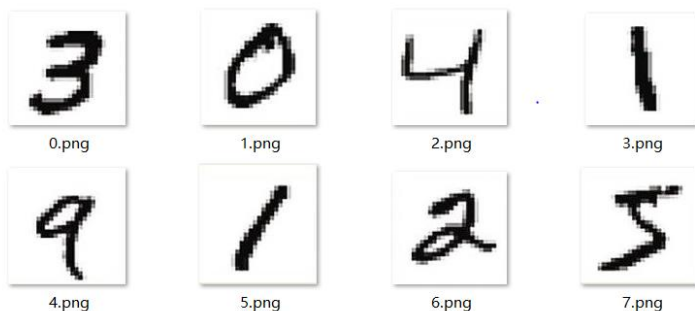
这里讨论推荐的目标是为用户展示之后用户会去点击，但是比赛要求的是预测用户是否去购买）。

如何更好地利用海量的交易数据来高效准确地预测用户的购买意向，是人工智能和大数据在电子商务场景中需要解决的技术难题。

手写字体识别



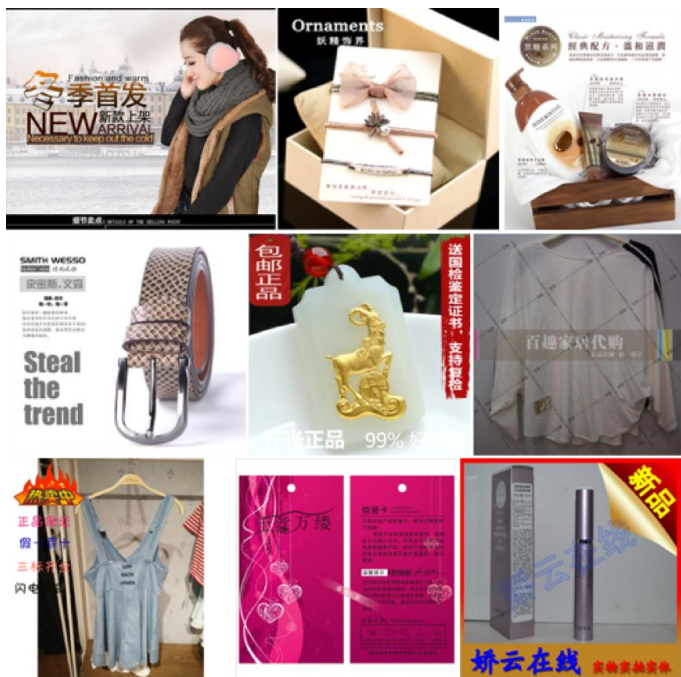
网上截图数据识别预测



```
conv2d_1 (Conv2D)          (None, 64, 26, 26)    640
max_pooling2d_1 (MaxPooling2 (None, 64, 13, 13)    0
dropout_1 (Dropout)         (None, 64, 13, 13)    0
conv2d_2 (Conv2D)          (None, 64, 11, 11)    36928
max_pooling2d_2 (MaxPooling2 (None, 64, 5, 5)    0
dropout_2 (Dropout)         (None, 64, 5, 5)    0
flatten_1 (Flatten)         (None, 1600)          0
dense_1 (Dense)             (None, 128)           204928
dropout_3 (Dropout)         (None, 128)           0
dense_2 (Dense)             (None, 10)            1290
=====
Total params: 243,786
Trainable params: 243,786
Non-trainable params: 0
=====
./num/0.png
Glib-GIO-Message: Using the 'memory' GSettings backend. Your settings will not be saved or shared with other applications.
./num/1.png
./num/2.png
./num/3.png
./num/4.png
./num/5.png
./num/6.png
./num/7.png
The results of the prediction are as follows:
[3 0 4 1 9 1 2 5]
```

```
./num/3.png
./num/6.png
./num/7.png
The results of the prediction are as follows:
[3 0 4 1 9 1 2 5]
```


网络图像的端到端文本检测和识别



在互联网世界中，图片是传递信息的重要媒介。特别是电子商务，社交，搜索等领域，每天都有数以亿兆级别的图像在传播。**图片文字识别（OCR）**在商业领域有重要的应用价值，是数据信息化和线上线下打通的基础，也是学术界的研究热点。本竞赛基于网络图片的中英混合数据集，该数据集数据量充分，涵盖几十种字体，几个到几百像素字号，多种版式，较多干扰背景。

网络图像的端到端文本检测和识别



494.91,36.36,494.91,81.45,596.0,81.45,595.27,32.73,三星
614.91,34.91,614.91,77.82,783.64,77.82,783.64,34.91,N7100
524.73,93.82,524.73,146.18,784.36,146.18,784.36,93.82,钢化玻璃膜
164.0,143.27,164.0,157.09,251.27,157.09,251.27,143.27,SNMSUNG
316.73,174.55,316.73,189.09,353.82,189.09,353.82,174.55,17:39
117.45,236.36,117.45,284.36,298.55,284.36,298.55,236.36,17:39
142.91,292.36,142.91,309.82,268.73,309.82,268.73,292.36,9月12日星期三
262.91,354.91,262.91,367.27,345.82,367.27,345.82,354.91,小到中雨转阵雨
321.82,338.91,321.82,350.55,345.82,350.55,345.82,338.91,北京
70.57,378.86,70.57,384.57,87.14,384.57,87.14,378.86,###
88.86,376.57,88.86,385.71,118.57,385.71,118.57,376.57,新浪天气
206.0,371.43,206.0,381.71,324.29,381.71,324.29,371.43,已更新2012/09/1116:59
76.86,545.71,76.86,556.0,106.0,556.0,104.86,546.29,沃·3G
150.57,545.71,150.57,557.71,183.71,557.71,183.71,545.71,沃商店
226.57,546.86,226.57,557.71,264.29,557.71,264.29,546.86,116114
294.57,545.14,294.57,557.71,350.0,557.71,350.0,545.14,手机营业厅
67.71,682.86,67.71,693.14,100.29,693.14,100.29,682.86,联系人
135.71,682.29,134.57,694.29,156.29,694.29,156.29,681.14,手机
195.14,682.86,195.14,693.71,218.57,693.71,218.57,682.86,信息
251.14,682.86,251.14,692.57,282.57,692.57,282.57,682.86,互联网
307.14,681.71,307.14,693.14,349.43,693.14,349.43,681.71,应用程序
232.29,513.14,232.29,522.86,261.43,522.86,261.43,513.14,116114
150.0,514.29,150.0,522.86,172.86,522.86,172.86,514.29,W0
153.43,523.43,153.43,532.0,180.86,532.0,180.86,523.43,###
76.29,522.29,76.29,529.14,105.43,529.14,105.43,522.29,###
76.29,504.0,76.29,514.86,107.14,514.86,107.14,504.0,W0
69.43,341.14,69.43,366.86,125.43,366.86,125.43,341.14,29° C
141.43,340.0,141.43,366.86,195.71,366.86,195.71,340.0,17° C
547.71,701.14,547.71,746.86,784.29,746.86,784.29,701.14,送贴膜工具
568.29,554.86,568.29,608.0,792.29,534.29,775.14,481.71,防爆防刮
551.14,449.14,577.43,510.86,795.71,434.29,785.43,372.57,智能贴合
542.0,344.0,564.29,396.57,788.86,323.43,771.71,265.71,防指纹油
659.14,543.43,659.14,544.0,659.71,544.0,659.71,543.43,###

任务：

网络图像的端到端文本检测和识别，同时评估文本检测和识别性能，不可以使用自己生成的数据，不可以使用预训练模型，只能用提供的训练集训练模型。

训练集（10000张图片）：
只考虑中文、英文和符号。

测试集（10000张图片）：
输入：一张图片

输出：对于每一个检测到的文本框，按行将其顶点坐标和文本内容输出到对应的[图像文件名].txt中。

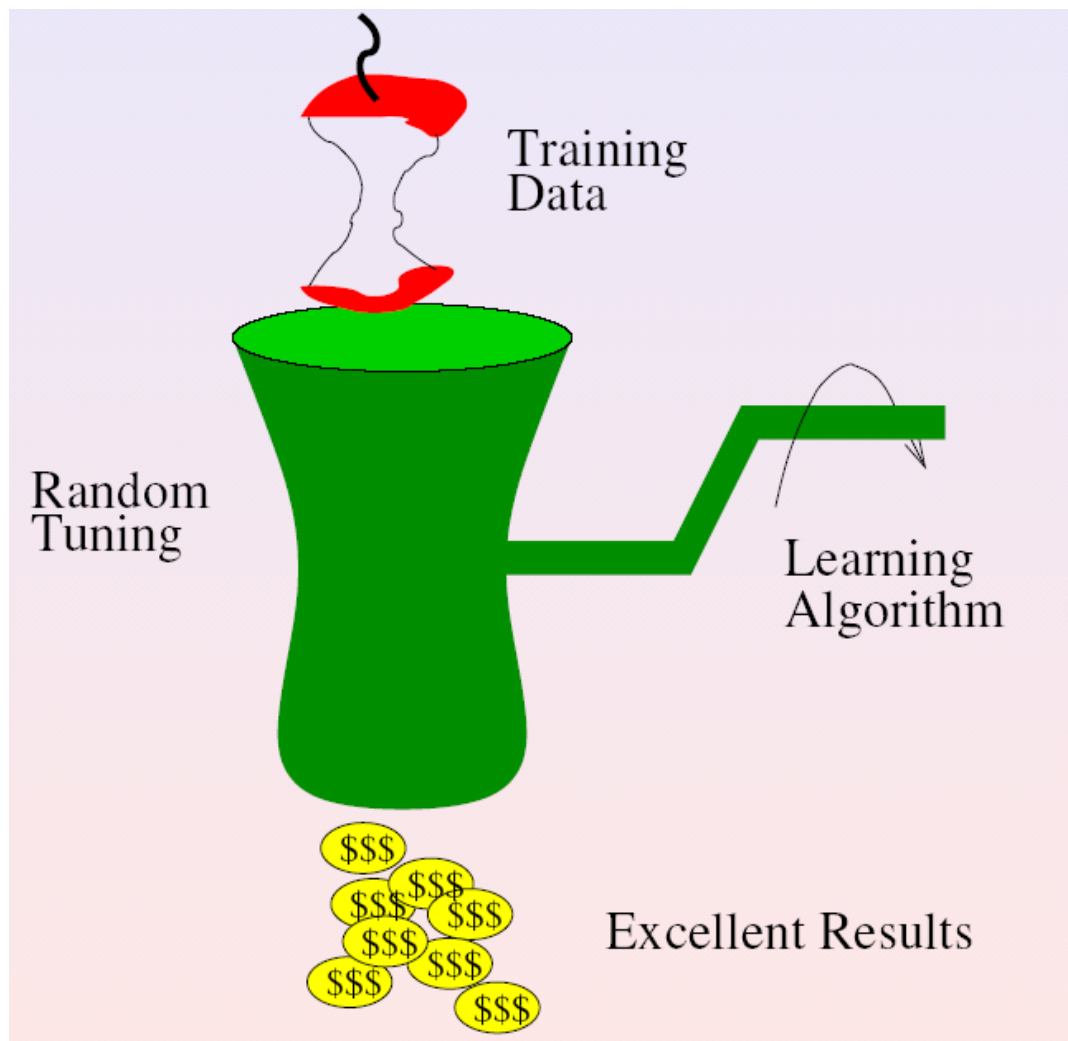


小结

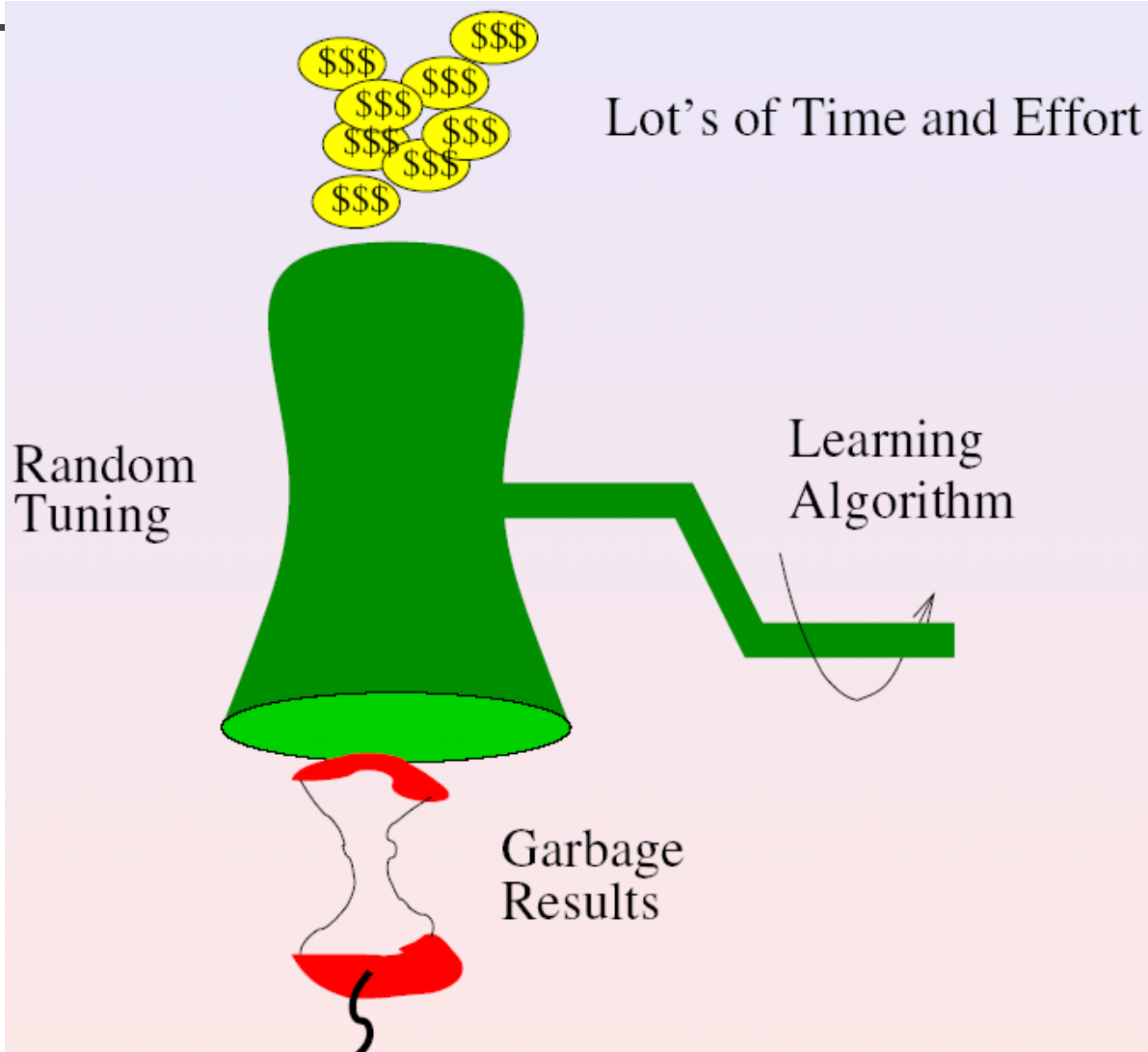
□ 我想要告诉大家的心里话

- ✓ **数据分析是一门实战艺术：**纸上得来终觉浅，绝知此事要躬行
- ✓ **锤子和钉子理论：**研究实际问题，精准定位碰到的具体场景，**抓准具体对象本身的特点、特性和问题，以问题驱动，而非以技术为导向，不要哪个技术热，追逐哪个。切忌脱离问题空谈花哨的方法**
- ✓ **活用数据，不要迷信数据以及被数据绑架！**

数据挖掘的魔力——数据的力量与陷阱



数据挖掘的魔力——数据的力量与陷阱





开始组队啦！

期末大作业组队要求：

≤ 3 人

每个人分工明确，承担一定工作量

明确说明队伍成员分工与贡献，要有区分度

确实组员贡献没有区分度的要写明原因



谢谢大家

推荐阅读：

深度好文：大数据，小数据，哪道才是你的菜

http://www.cbdio.com/BigData/2015-08/14/content_3697325_all.htm