



中國人民大學
RENMIN UNIVERSITY OF CHINA

计算传播理论与实务

2019-2020秋季学期

第二讲

统计思维与实用机器学习

授课教师：范举副教授、塔娜讲师

时间：2019年9月30日

第2.2节

实用机器学习

机器学习 – Machine Learning

- 人们不需要理解过去，更需要预测未来

理解过去 (数据描述)	预测未来 (机器学习)
苹果：今年卖了多少部手机？	苹果：明年会卖多少部手机？
百度：你是否点击了广告？	百度：你会不会点击广告？
保险公司：哪些保户已经挂了？	保险公司：这个新保户会挂吗？
本课程：哪些同学已经得了4.0？	本课程：你会不会得4.0？
.....

机器学习 – Machine Learning

- 人们不需要手动分类，更需要自动识别



狐狸



狗



狐狸 or 狗？



华为Mate30很好！



华为Mate30不太好！



Mate30不能再好了！



？ ? ?

回顾：机器学习能回答什么问题

- 问题 I:

Is This A or B?

- 分类问题 (Classification)
 - 这张照片中是猫还是狗?
 - 这条评论体现了积极还是消极情绪?
 - 这个用户是否会点击给他推送的广告?

回顾：机器学习能回答什么问题

- 问题2：

How much or How Many?

- 回归问题 (Regression)
 - 我发完这条微博会涨多少粉？
 - 下礼拜一的气温会是多少？
 - 中国第四季度的GDP增长会是多少？

回顾：机器学习能回答什么问题

- 问题3：

Is This Weird?

- 异常检测 (Outlier Detection)
 - 那个群体的情绪变化与大众明显不同？
 - 那个用户最近买的商品与以往显著不同？
 - 哪条借贷记录属于欺诈行为？

回顾：机器学习能回答什么问题

- 问题4:

How Is This Organized?

- 聚类问题 (Clustering)
 - 哪些评论在话题上非常类似?
 - 哪些顾客在购物行为上非常类似?
 - 哪些用户喜欢同样类型的电影?

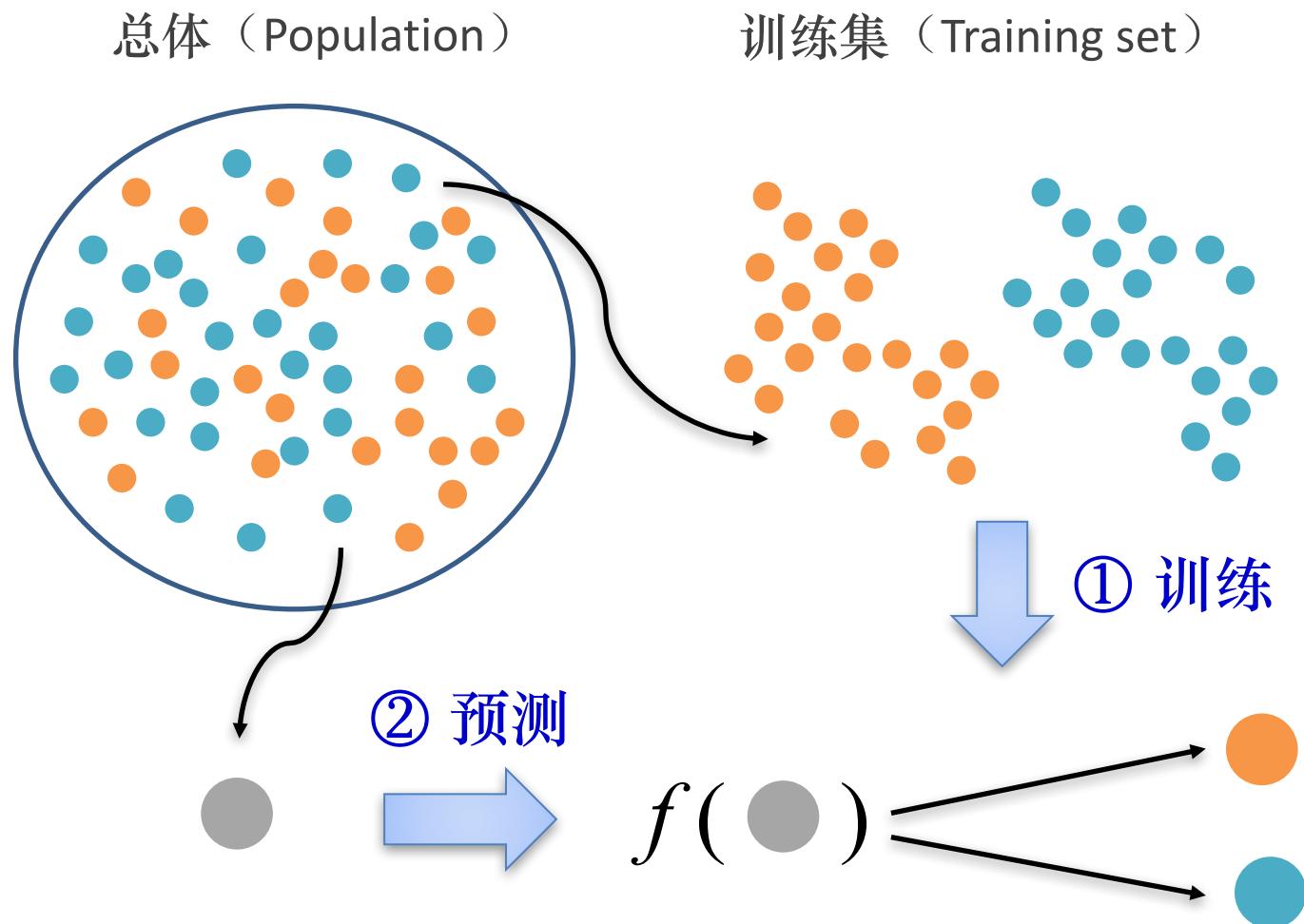
本讲主要内容

- 主要侧重二分类问题:

Is This A or B?

- 主要内容
 - 机器学习如何预测
 - 如何设计机器学习方法
 - 机器学习算法原理
 - 数据采集与数据准备

机器学习如何预测



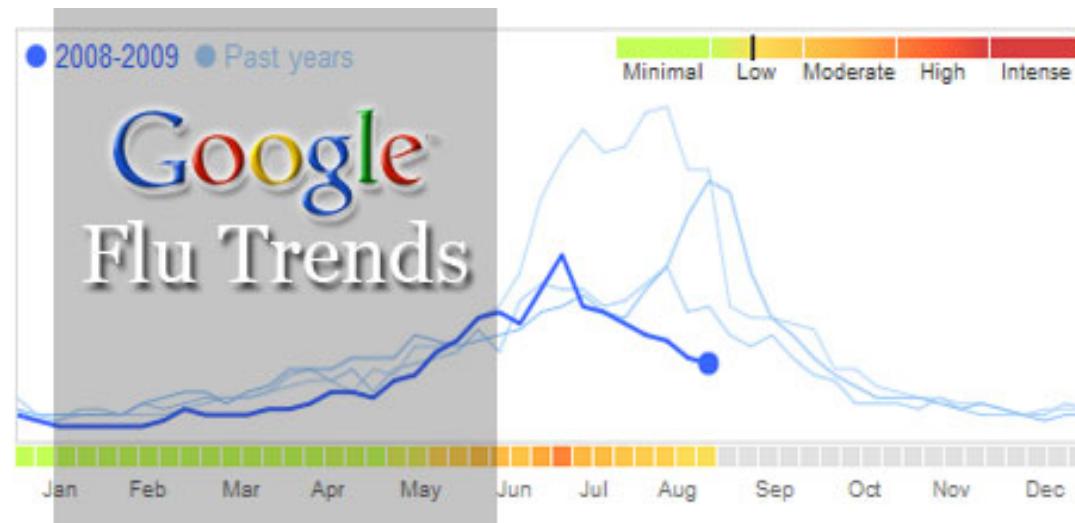
案例分析：Google Flu



Aah Choo! I got the flu...

案例分析：Google Flu

- 流行病学调查
 - Center for Disease Control and Prevention
- 如何使用搜索关键词预测流感？



Google Flu Trends (GFT)

Detecting influenza epidemics using search engine query data: Nature 2009

案例分析：Google Flu

- 基本思想

- 如果一个用户在Google搜索引擎上搜索与“流感”相关的词，则预测他可能得了流感
- 观察以下查询关键词，你预测谁得了流感.....

user	query	month
1477	head flu symptoms	2006-03
2178	pennsylvania college savings plan	2006-03
2178	fuel additives check engine light	2006-03
9852	a cure for the flu ivy blossom	2006-04
2178	bare minerals make up	2006-04
3587	bird flu usa	2006-04
1849	drugs that shorten flu symptoms	2006-04
1326	konig wheels	2006-04
1326	jet blue airlines	2006-04
3587	bird flu symptoms	2006-05

案例分析：Google Flu

- 良好的预测效果.....

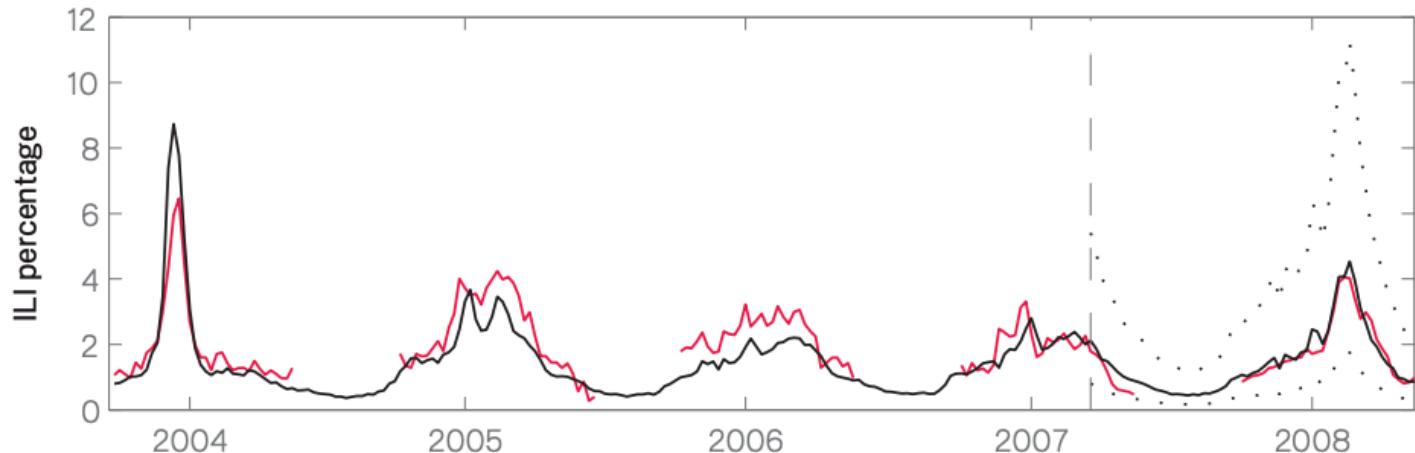
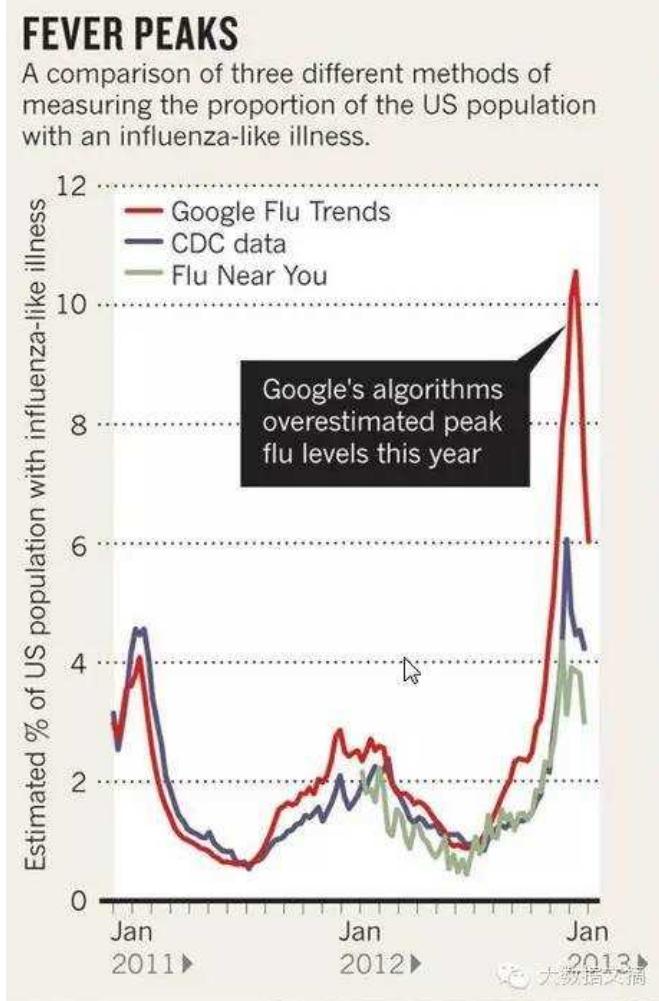


Figure 2: A comparison of model estimates for the Mid-Atlantic Region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, while a correlation of 0.96 was obtained over 42 validation points. 95% prediction intervals are indicated.

Source:

<https://static.googleusercontent.com/media/research.google.com/en//archive/papers/detecting-influenza-epidemics.pdf>

案例分析：Google Flu



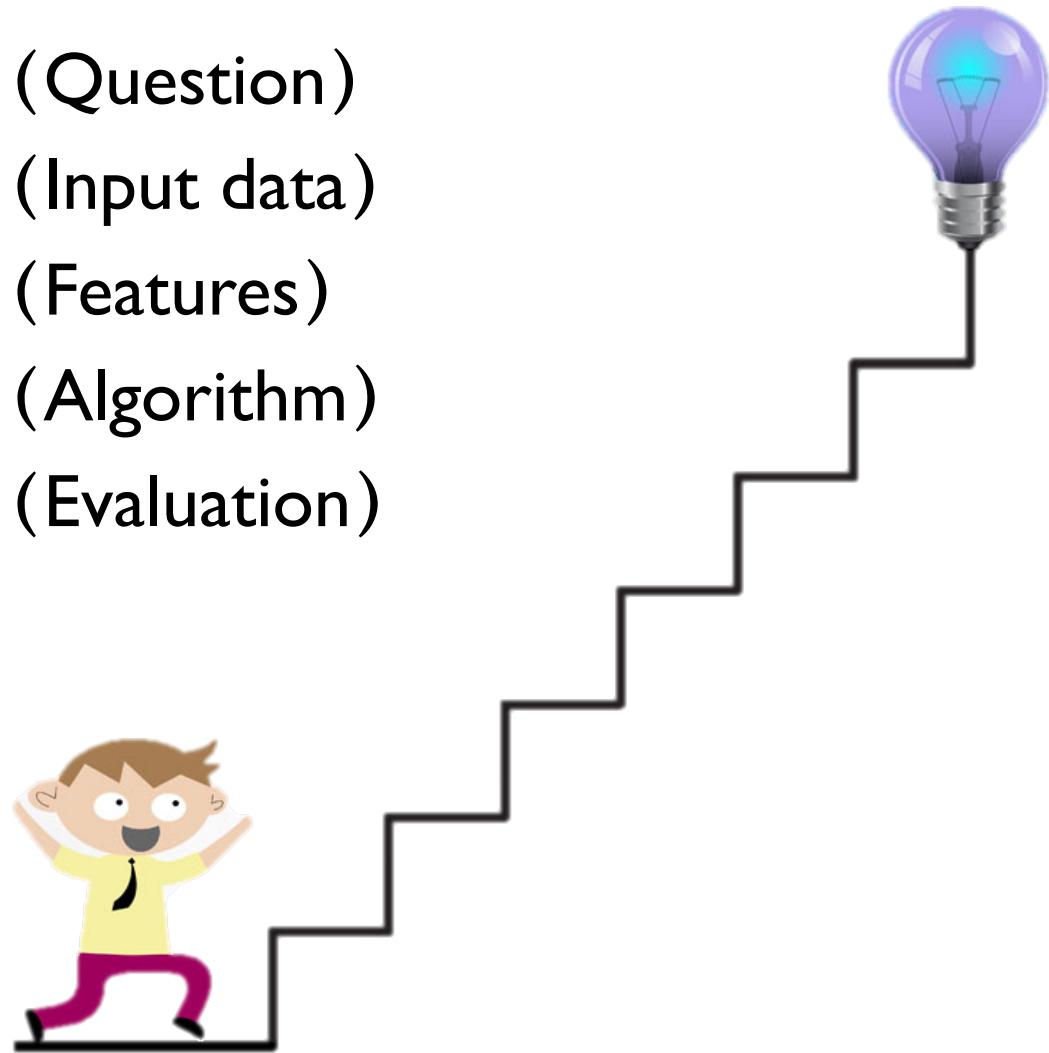
- 然而，2012年12月的一次流感爆发中，GFT预测显示某次的流感爆发非常严重然而疾控中心CDC在汇总各地数据以后发现谷歌的预测结果比实际夸大了几乎一倍
- 原因分析
 - 模型本身：搜索查询词的复杂性
 - 模型之外：搜索流感不代表得了流感.....

第2.2.1节

○
基于机器学习的预测方法
应当如何设计？

方法设计的基本步骤

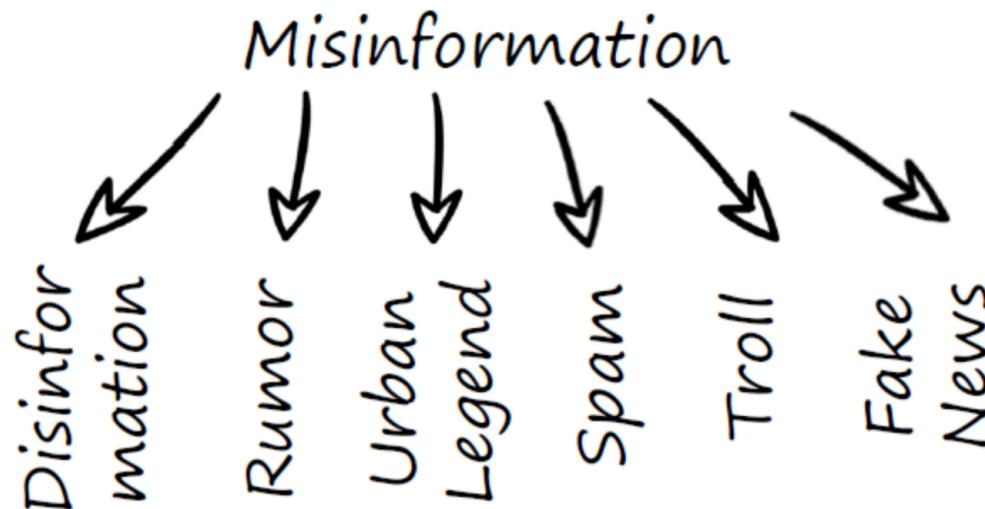
- 提出问题 (Question)
- 准备数据 (Input data)
- 选择特征 (Features)
- 学习算法 (Algorithm)
- 评价模型 (Evaluation)



案例分析：虚假信息检测

Question → Data → Features → Algorithm → Evaluation

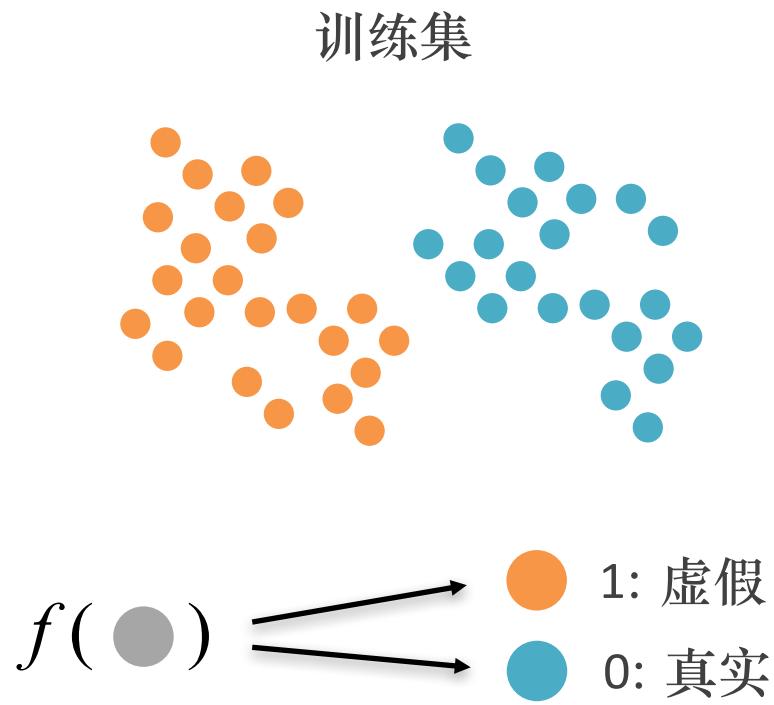
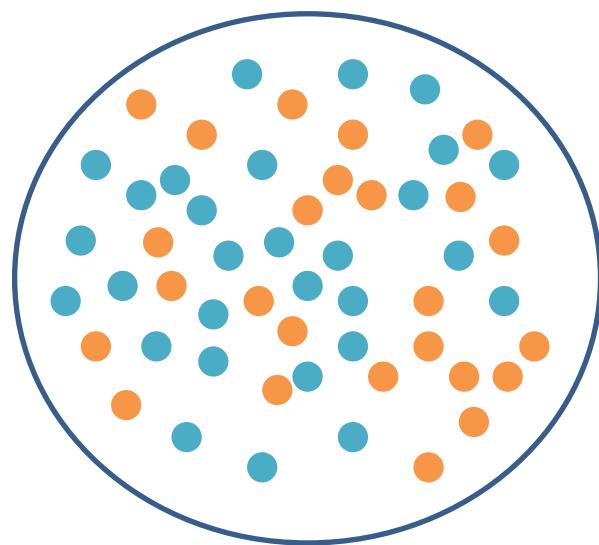
- 虚假信息 (Misinformation) 检测
 - 在社交媒体上自发传播的错误（假）或者不准确（虚）的信息
- 与虚假信息 (Misinformation) 相近的概念



案例分析：虚假信息检测

Question → Data → Features → Algorithm → Evaluation

- 虚假信息（Misinformation）检测
 - 问题定义应明确而具体



案例分析：虚假信息检测

Question → Data → Features → Algorithm → Evaluation

- 准备虚假信息检测的数据



- 原始数据获取
 - 根据Twitter API获取数据
 - 后面会介绍数据准备的具体步骤与技术
- 数据标注
 - 通过人工标注，为收集到的每条tweet打标签：1为虚假信息；0为真实信息

案例分析：虚假信息检测

Question → Data → Features → Algorithm → Evaluation

- 一条信息包含哪些数据？



虚假信息散布者

虚假信息内容

虚假信息时空属性

虚假信息传播者

- 文本
 - 文字
 - 标签
 - 链接
 - 表情
- 图片
- 视频 (GIF)

- 日期、时间
- 地点

- 点赞、回复
- 转发

案例分析：虚假信息检测

Question → Data → **Features** → Algorithm → Evaluation

- 什么是特征？



你如何区分
汪星人
喵星人



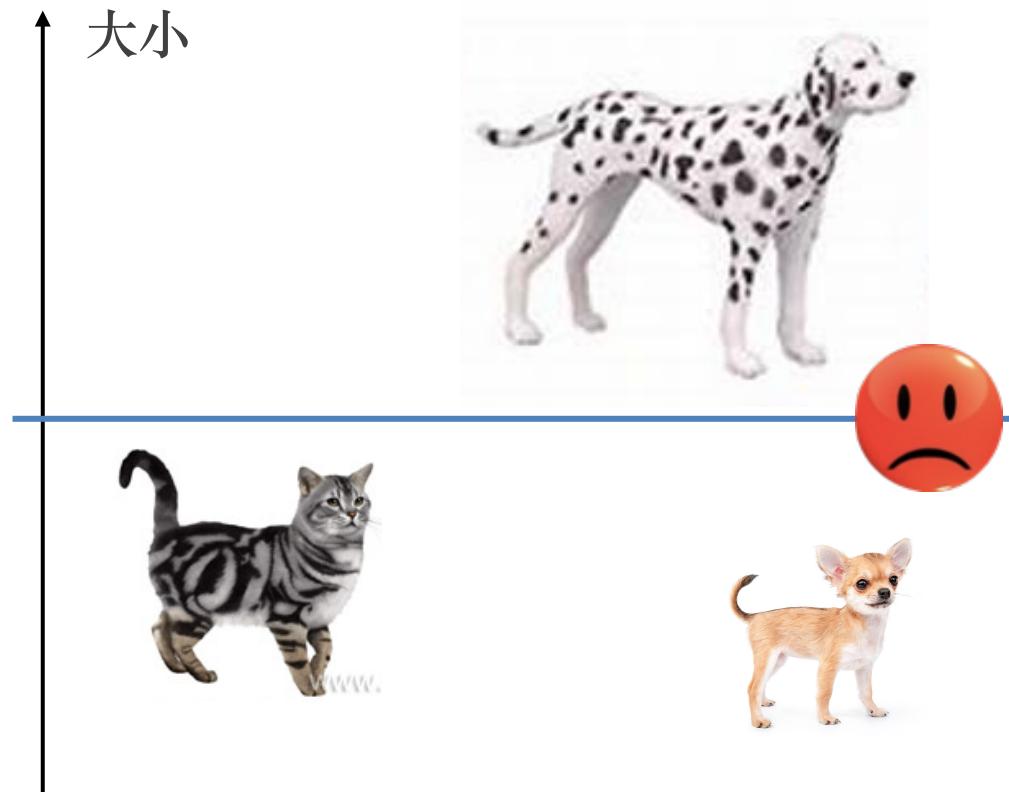
案例分析：虚假信息检测

Question → Data → **Features** → Algorithm → Evaluation

- 什么是特征？



你如何区分
汪星人
喵星人



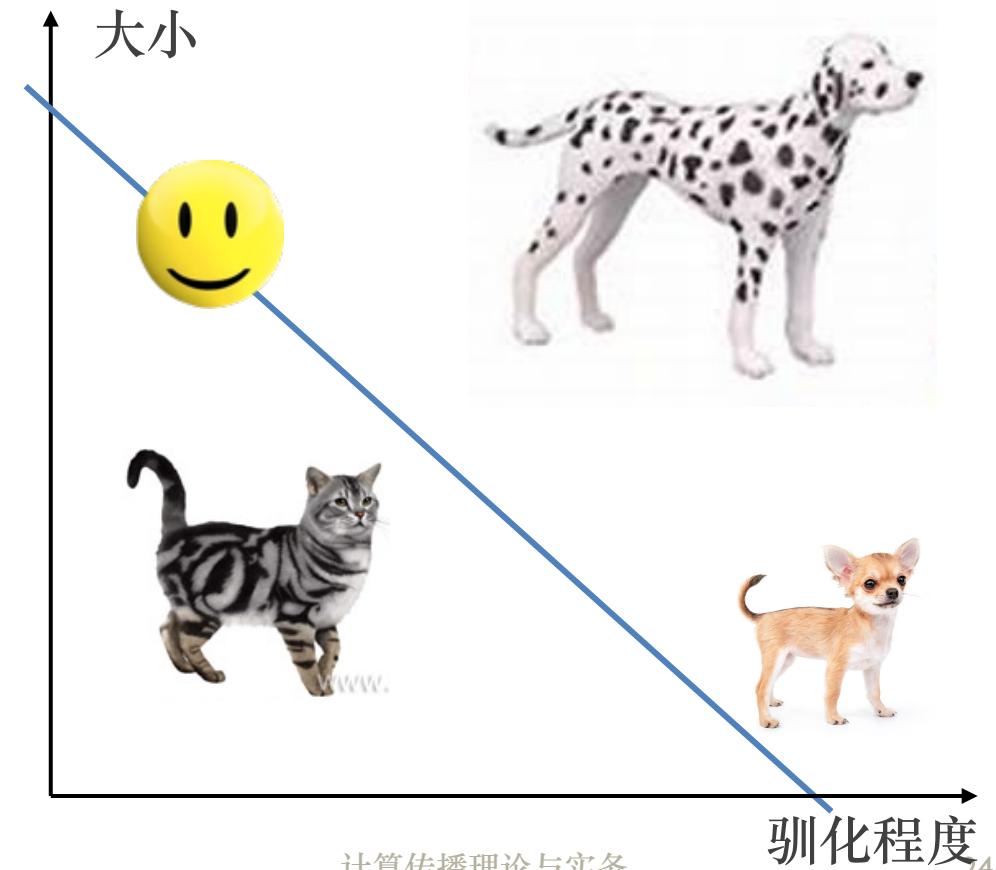
案例分析：虚假信息检测

Question → Data → **Features** → Algorithm → Evaluation

- 什么是特征？



你如何区分
汪星人
喵星人



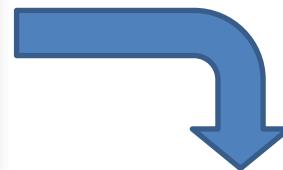
案例分析：虚假信息检测

Question → Data → **Features** → Algorithm → Evaluation

- 如何给一条Tweet信息提取特征？

#hurricaneharvey2017

Believe it or not, this is a shark on the freeway
in #Houston, #Texas. #HurricaneHarvy
#HarveyStorm #houstonflood



Text Feature	Example
Length of post	#words, #characters
Punctuation marks	Question mark ? Exclamation!
Emojis/Emoticons	Angry face ;-L 😡
Sentiment	Sentiment/swear/curse words
Pronoun (1 st , 2 nd , 3 rd)	I, me, myself, my, mine
URL, PageRank of domain	@AndrewGirdwood Have you heard Google was hiring people to work from home? pretty cool i thought
Mention (@)	http://dwarfurl.com/1f291
Hashtag (#)	Free President #Gbagbo #8demarzo #stopkony #SOLARSTORM #iwvd #ftmedia12 #BarackObama #cnn #breakingNews #Syria #sarkozy

案例分析：虚假信息检测

Question → Data → **Features** → Algorithm → Evaluation

- 案例分析：真实的中文虚假信息

娱乐类（标题）

曾毅退出凤凰传奇，与庞龙合唱，原来他唱功这么好！

唐嫣学生装现身机场，宽松胖大疑似怀孕？罗晋要升级当爸爸！

《七仙女》开拍，赵丽颖和郑爽谁更适合演七仙女和鹿晗搭档

今天鹿晗和关晓彤分手了吗？

佟丽娅被爆离婚，网友为何大喊：痛快人心！

社会类（标题）

家长注意了！初三有这6个现象的孩子，成绩会越来越差！

摊贩城管一家亲！10月1日起，农户进城摆摊不用愁了，可放心买卖

电瓶车用久了电池就得换？别傻了，教你1招，再多骑10年都没问题！

医院交停车费，找回来一枚硬币，价值3000元，真不敢相信！

还用别人的驾驶证代扣分？现在已经行不通了！

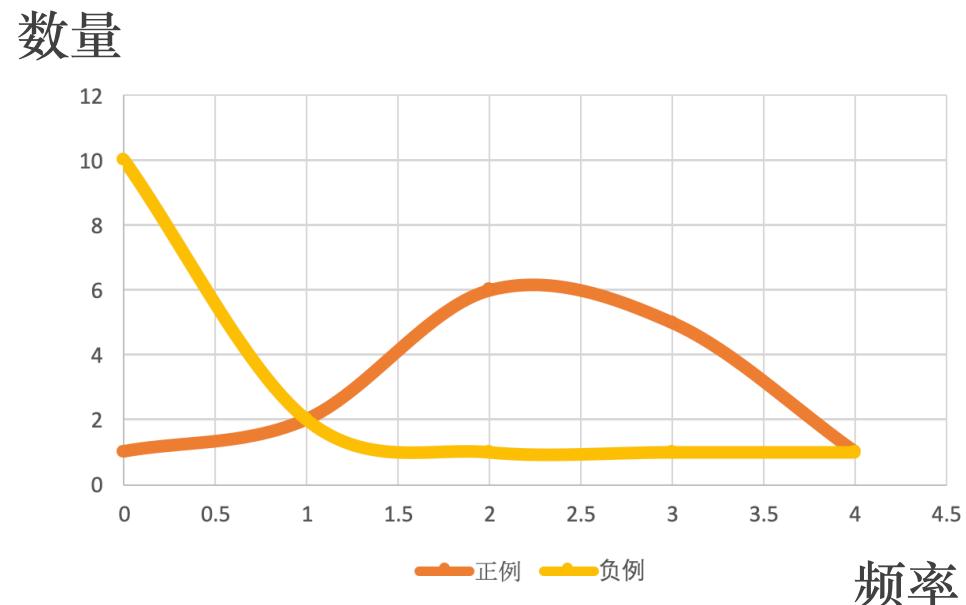
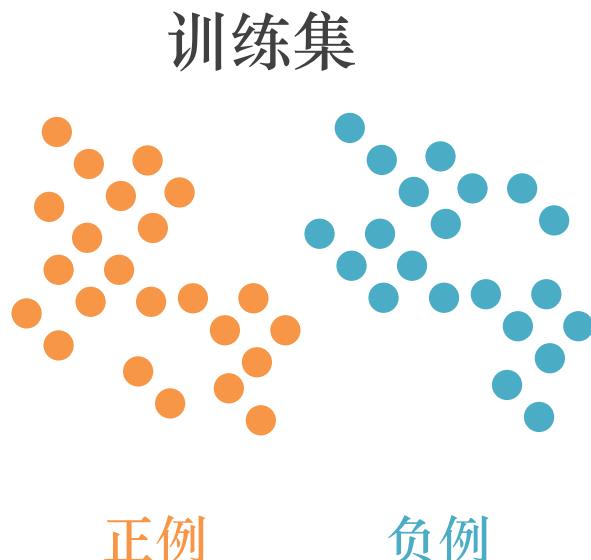
你会
提取什么
特征

？

案例分析：虚假信息检测

Question → Data → **Features** → Algorithm → Evaluation

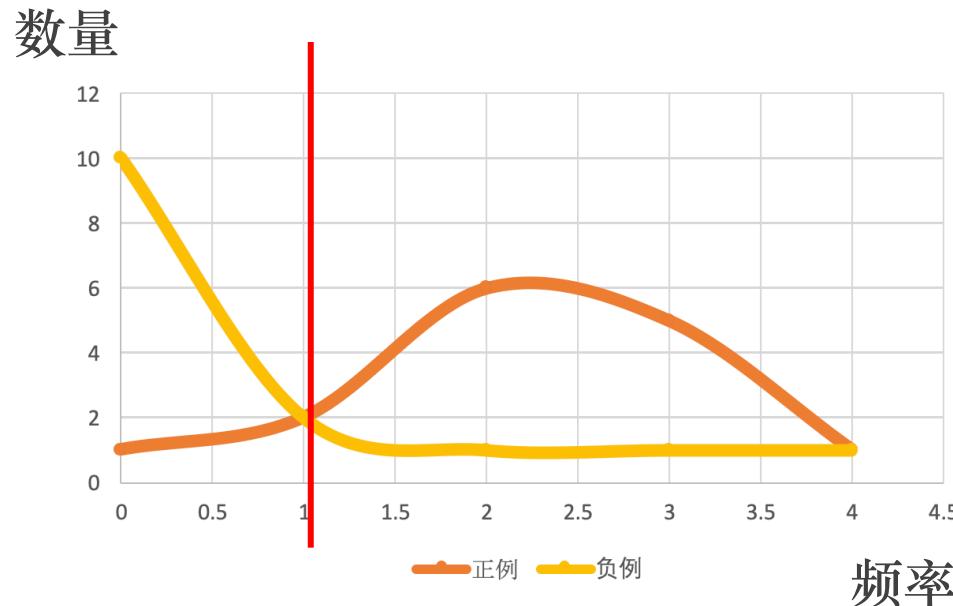
- 最简单的特征选取策略
 - 问号加感叹号出现在第 i 篇消息中的频率：记为 F_i



案例分析：虚假信息检测

Question → Data → Features → **Algorithm** → Evaluation

- 最简单的分类算法
 - 设置一个阈值C，如C=1



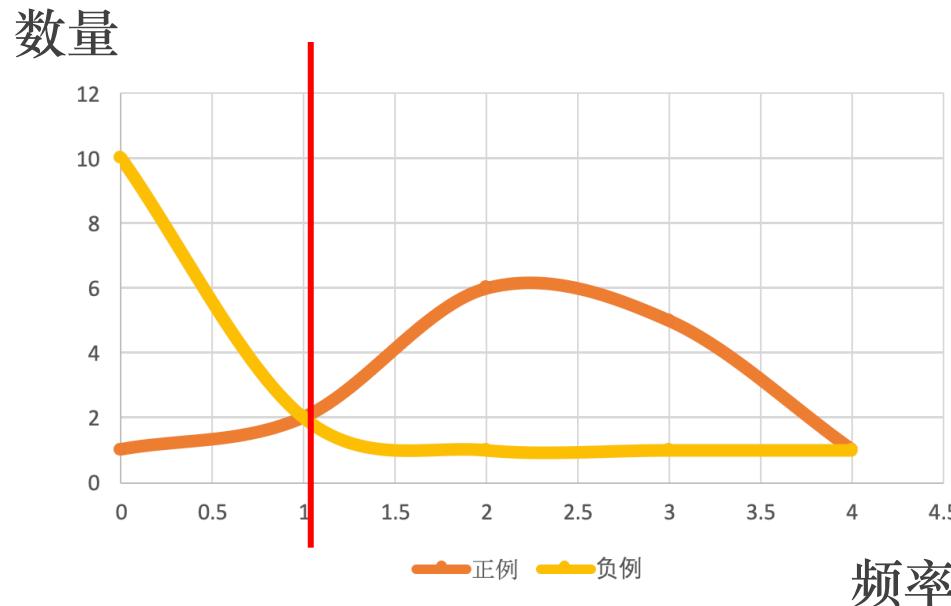
```
if F_i > C:  
    return 1  
else:  
    return -1
```

你的第一个
分类器

案例分析：虚假信息检测

Question → Data → Features → Algorithm → **Evaluation**

- 如何评价你分类器的优劣?
 - In Sample Error: 直接在训练集评测



预测结果

标准答案	预测结果	
	正例	负例
正例	12	3
负例	3	12

$$\text{Accuracy} = \frac{24}{30} = 80\%$$

案例分析：虚假信息检测

Question → Data → Features → Algorithm → **Evaluation**

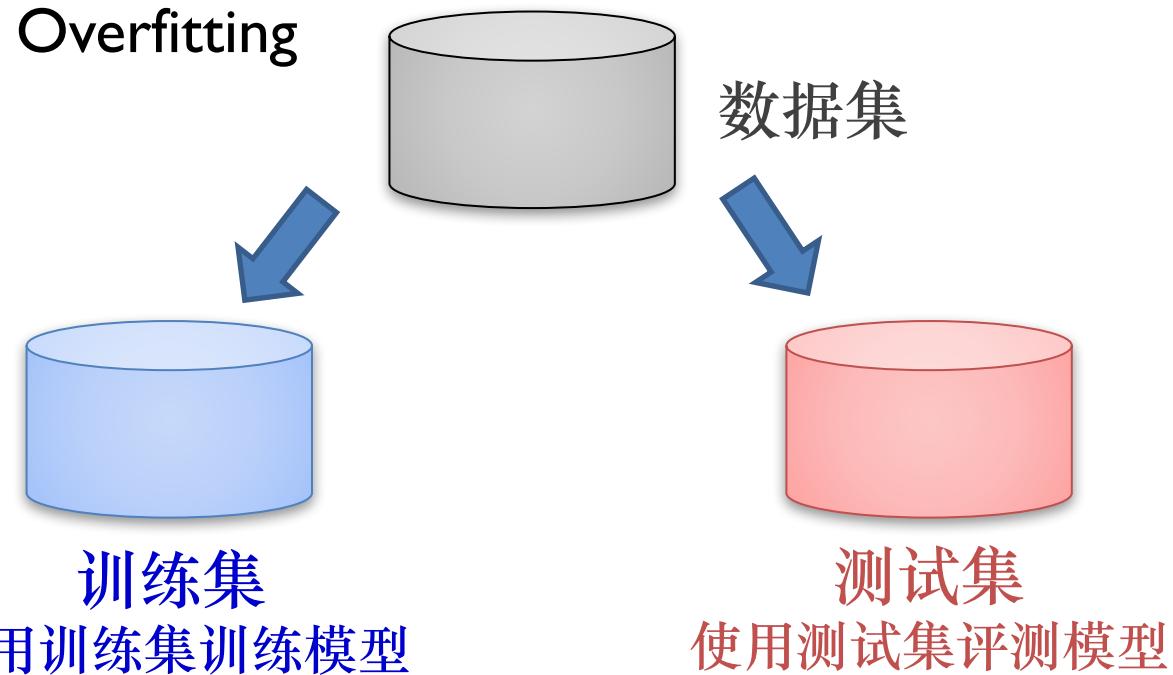
- 只在训练集中评测是否合理？
- 回顾上节课的例子
 - 从Web上收集了1000张图片，训练出分类模型，其分类的准确率为95%
 - 总体：所有在Web上的图片
 - 样本：从Web上收集到的1000张照片



案例分析：虚假信息检测

Question → Data → Features → Algorithm → **Evaluation**

- 如何评价你分类器的优劣?
 - Out of Sample Error: 划分为训练集和测试集
 - 了解: Overfitting



案例分析：虚假信息检测

Question → Data → Features → Algorithm → **Evaluation**

- 使用测试集评测分类算法，结果如下：

预测结果		
	正例	负例
正例	8	12
负例	8	72

$$\text{Accuracy} = \frac{80}{100} = 80\%$$



看指标很不错

但总感觉哪里不对

- 几个关键概念
 - True Positive (TP): 正确识别
 - False Positive (FP): 错误识别
 - True Negative (TN): 正确拒绝
 - False Negative (FN): 错误拒绝
- 练习：将 TP/FP/TN/FN 填到下表

预测结果		
	正例	负例
标准答案	TP	FN
负例	FP	TN

案例分析：虚假信息检测

Question → Data → Features → Algorithm → **Evaluation**

- 使用测试集评测分类算法，结果如下：

预测结果		
	正例	
正例	8	12
负例	8	72

$$\text{Accuracy} = \frac{80}{100} = 80\%$$



看指标很不错
但总感觉哪里不对

- 引入新的指标

- Precision – 准确率

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall – 召回率

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F1 Score - F值

$$F1 = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

请计算左表

案例分析：虚假信息检测

Question → Data → Features → Algorithm → **Evaluation**

- 使用测试集评测分类算法，结果如下：

预测结果		标准答案
正例	负例	
正例	8	12
负例	8	72

$$\text{Accuracy} = \frac{80}{100} = 80\%$$



看指标很不错

但总感觉哪里不对

- 引入新的指标
 - Precision = 50%
 - Recall = 40%
 - F1 Score = 44%



虚假信息识别效果很差！

案例分析：虚假信息检测

Question → Data → Features → Algorithm → **Evaluation**

- 使用测试集评测分类算法，结果如下：

预测结果		
	正例	负例
正例	8	12
负例	8	72

$$\text{Accuracy} = \frac{80}{100} = 80\%$$



看指标很不错
但总感觉哪里不对

- 在医疗检测或社会科学领域也使用评测指标：

- Sensitivity – 灵敏度**

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

- Specificity – 特异度**

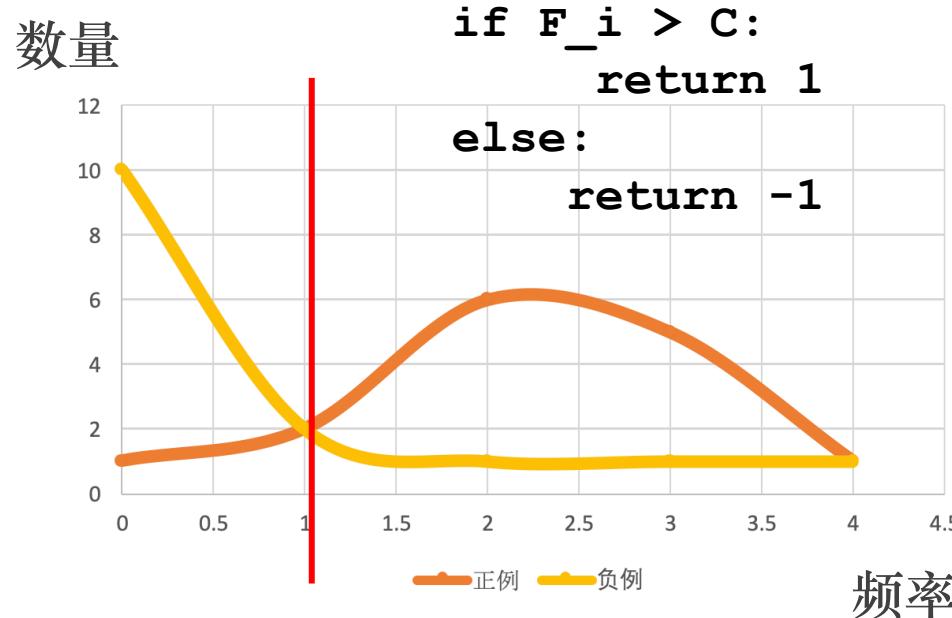
$$\text{Specificity} = \frac{TN}{TN + FP}$$

请计算左表

案例分析：虚假信息检测

Question → Data → Features → Algorithm → **Evaluation**

- 如果变化阈值C的取值，请计算并观察趋势

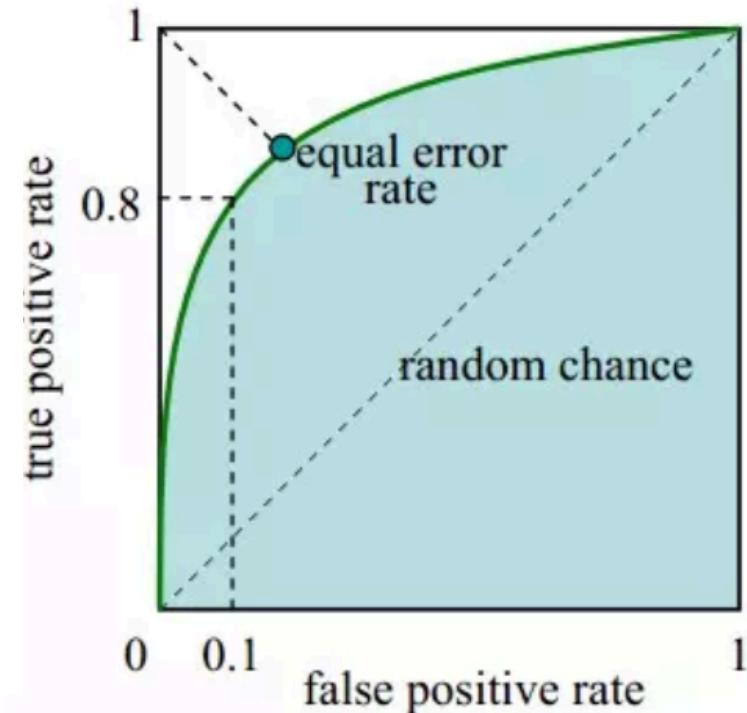
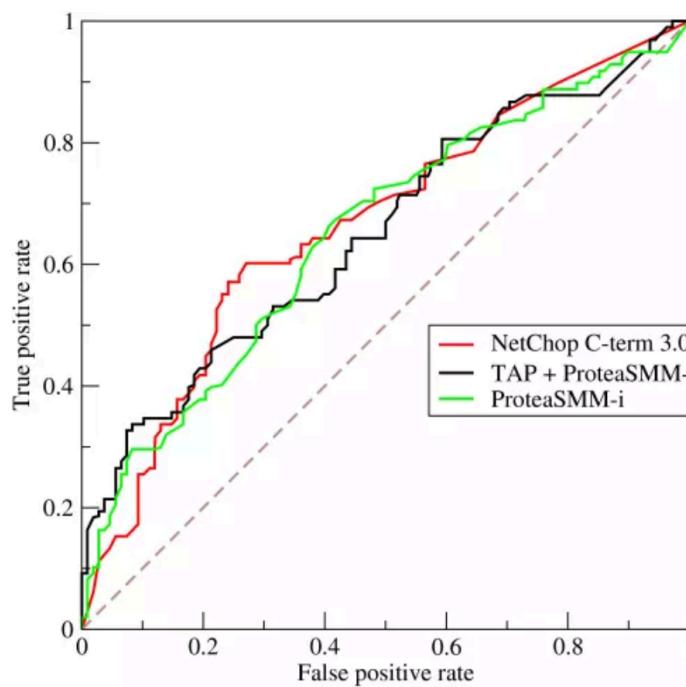


C	Sensitivity	1-Specificity
0		
1		
2		
3		
4		

案例分析：虚假信息检测

Question → Data → Features → Algorithm → **Evaluation**

- 新的指标：ROC曲线与AUC (Area Under Curve)



总结：哪步

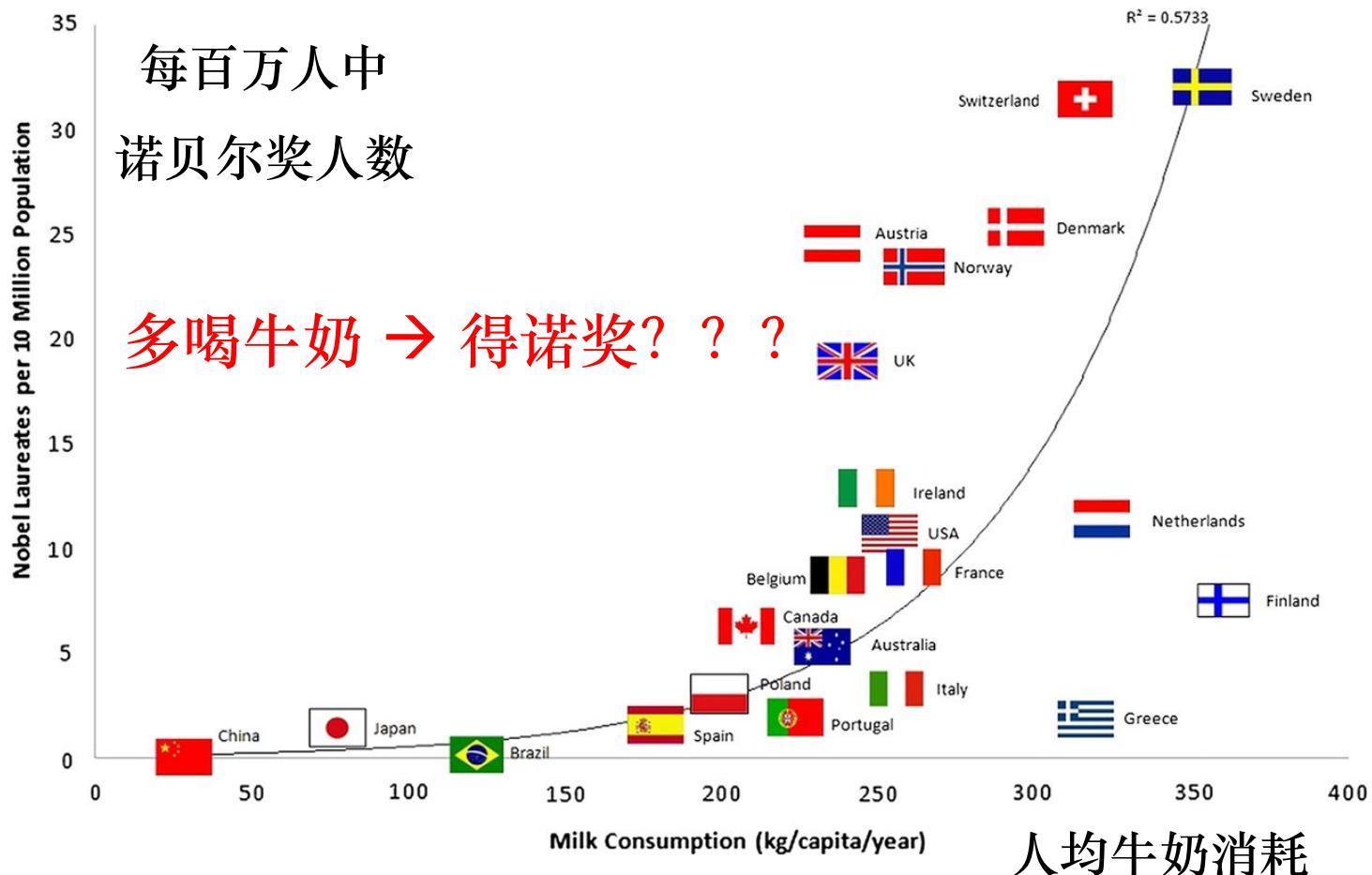
- 提出问题 (**Q**uestion)
- 准备数据 (**I**nput **D**ata)
- 选择特征 (**F**eatures)
- 学习算法 (**A**lgorithm)

$Q > D > F > A$



为什么数据比算法重要？

- GIGO: Garbage In, Garbage Out

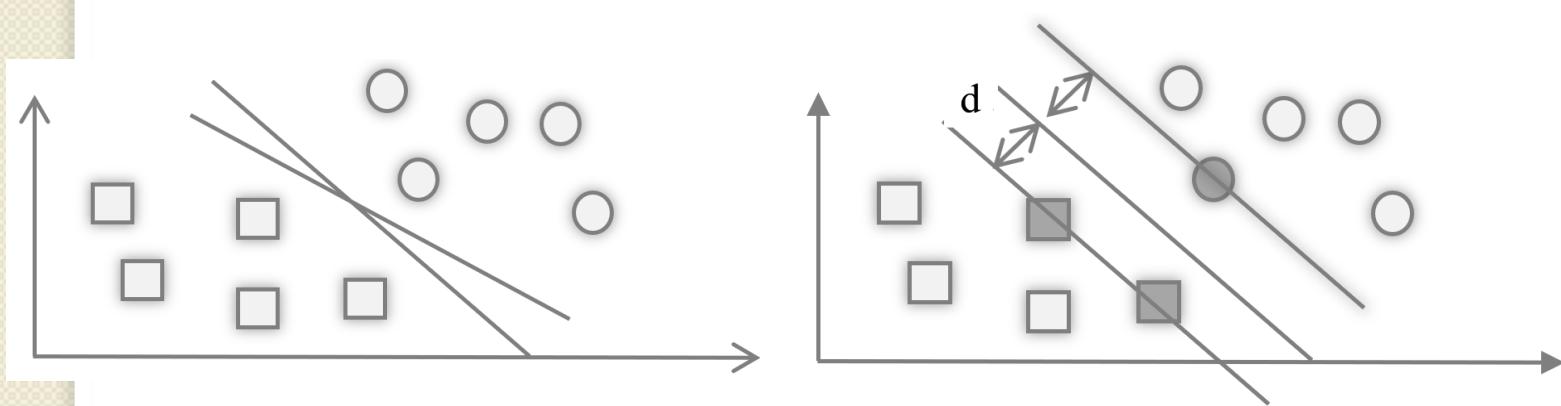


第2.2.2节

机器学习典型算法

SVM模型

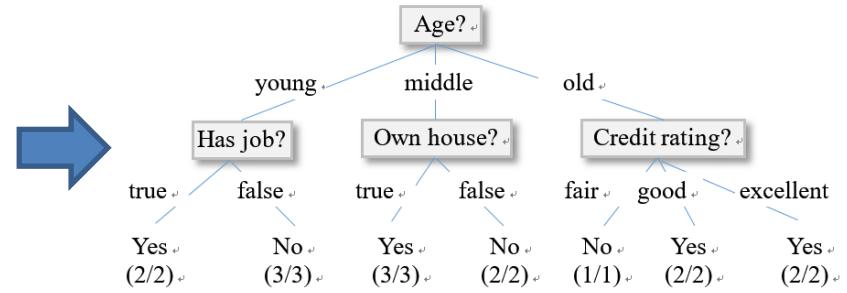
- 机器学习中的分类模型
- 分类模型：以SVM为例
 - 二维平面上把两类点分开，可用的直线有多条
 - 最优的那条，到两类数据点的距离都最大
 - 确定这条直线的少量数据点，称为支持向量



决策树模型

- 机器学习中的分类模型
- 分类模型：以决策树模型为例
 - 决策树中的非叶子节点，表示对象属性的判断条件，其分支表示符合节点条件的所有对象，树的叶子节点表示对象所属的预测结果。

ID	Age	Has Job	Own House	Credit Rating	Class
1	Young	false	false	fair	No
2	Young	false	false	Good	No
3	Young	true	false	Good	Yes
4	Young	true	true	fair	Yes
5	Young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	True	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No



第2.2.3节

数据采集

数据采集：Where to Collect

- 考虑一个场景：请你基于数据分析原因

中国iPhone销量下滑速度是整个市场的两倍

2019年02月11日 23:03 3558 次阅读 稿源：威锋网 4 条评论

苹果在其假日季度财报电话会议上透露，iPhone 在中国的糟糕销售是导致该公司季度收入达不到预期的主要原因。市场分析公司 IDC 本周对 iPhone 在中国市场的糟糕程度进行了估计，在中国，iPhone 销量的下滑速度是智能手机市场整体下滑速度的两倍。



- 你要采集哪些数据来支撑你的分析？

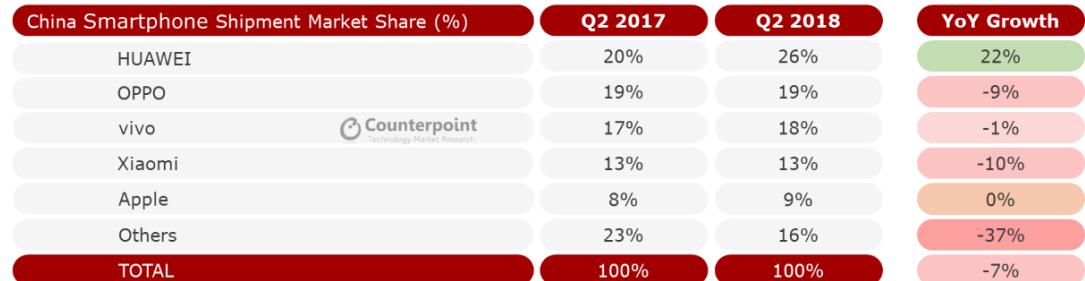
数据采集：Where to Collect

- 你要采集哪些数据来支撑你的分析？
- 内部数据
 - 应用场景数据库（关系数据）
 - 例如：iPhone不同型号，及在不同销售地的定价
 - 系统日志（文本数据）
 - 例如：用户在苹果官网搜索、购买iPhone及其周边的历史
 - 文档数据（Word, Excel, PDF, CSV）
 - 例如：销售渠道汇总来的表格数据
 - 多媒体数据（视频、音频、图片）

数据采集：Where to Collect

- 你要采集哪些数据来支撑你的分析？
- 外部数据
 - 网页数据

2018Q2中国市场手机市场份额：



华为依然在中国市场的老大，主要得益于子品牌荣耀多渠道分销策略带来的快速增长，而且华为是唯一一家能够实现同比增长的制造商，出货量暴涨了 22%，其余均不同程度下降，小米出货量跌幅达到 10%，“其他”类别暴跌 37%，说明小厂商几乎已无法生存。就出货量占比而言，华为出货量达到 26% 的份额，其次是 OPPO 的 19%，vivo 的 18%，小米的 13% 和苹果的 9%。

数据采集：Where to Collect

- 你要采集哪些数据来支撑你的分析？
- 外部数据
 - 网页数据
 - Web API

The screenshot shows the Weibo Open Platform API documentation. The top navigation bar includes links for 微连接, 微服务, 文档, 支持, 推广, and 我的应用. On the left, there's a sidebar with sections for 微博API (including 微博API, API更新日志, 接口访问频次权限, and 新版接口迁移指南), 资源下载, 常见问题, and 联系我们. The main content area is titled '微博' and contains two tables: '读取接口' and '评论'. The '读取接口' table lists various API endpoints:

读取接口	statuses/home_timeline	获取当前登录用户及其所关注用户的最新微博
	statuses/user_timeline	获取用户发布的微博
	statuses/repost_timeline	返回一条原创微博的最新转发微博
	statuses/mentions	获取@当前用户的最新微博
	statuses/show	根据ID获取单条微博信息
	statuses/count	批量获取指定微博的转发数评论数
	statuses/go	根据ID跳转到单条微博页
	emotions	获取官方表情
	写入接口	statuses/share

The '评论' section is partially visible at the bottom.

数据采集：Where to Collect

- 你要采集哪些数据来支撑你的分析？
- 外部数据
 - 网页数据
 - Web API
 - **开放数据
(Open Data)**

哪一些网站提供中国的开放数据(open data)?

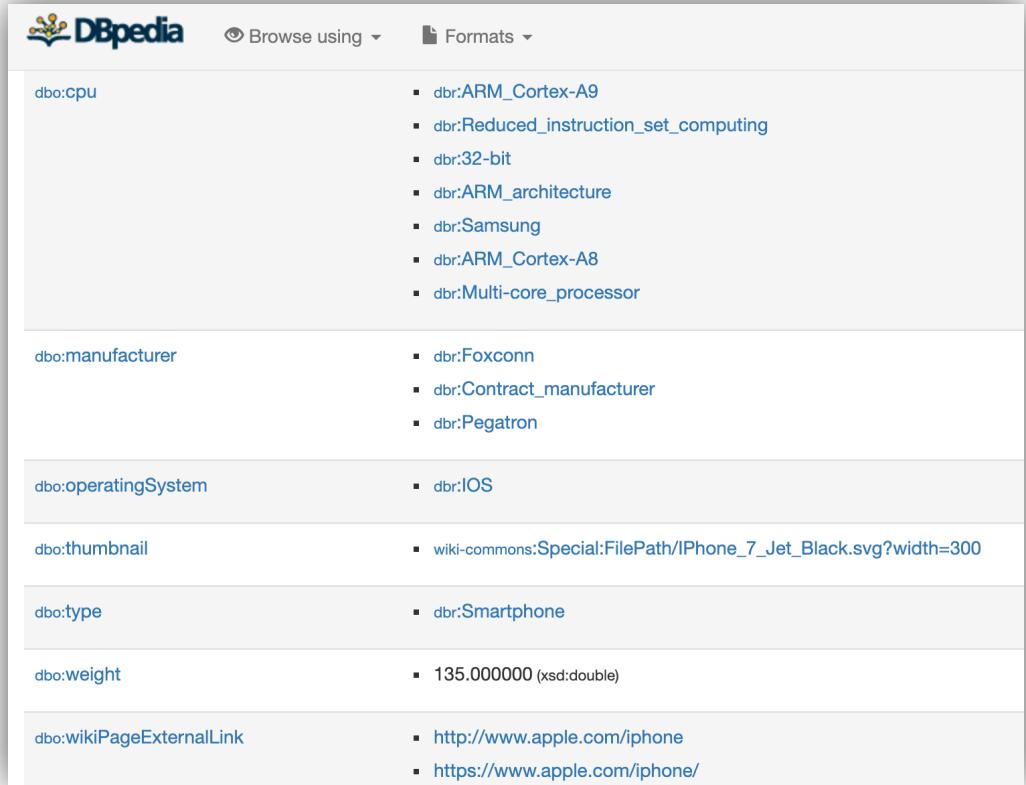
国内资源不完全统计：

北京 bjdata.gov.cn/
上海 datashanghai.gov.cn/
浙江省 data.zjzwfw.gov.cn/
武汉 <http://wuhandata.gov.cn>
青岛 data.qingdao.gov.cn/
杭州 114.215.249.58/
贵阳 datagy.cn/
无锡 opendata.wuxi.gov.cn/
湛江 data.zhanjiang.gov.cn/
宁波海曙 data.haishu.gov.cn/hs_m...
佛山南海 data.nanhai.gov.cn/
深圳罗湖 szlh.gov.cn/opendata/
深圳质量监管 szscjg.gov.cn/fz/openda...
深圳住建 szjs.gov.cn/fzlm/openda...

中国气象开放服务平台 openweather.weather.com.cn...
中国专利数据 patdata.sipo.gov.cn/
国家数据 data.stats.gov.cn/

数据采集：Where to Collect

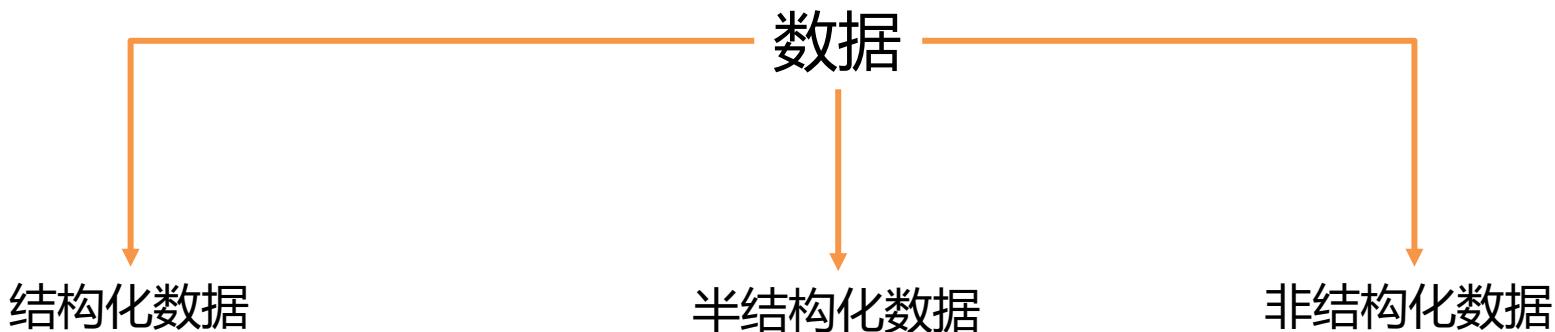
- 你要采集哪些数据来支撑你的分析？
- 外部数据
 - 网页数据
 - Web API
 - 开放数据
(Open Data)
 - 知识图谱
(DBpedia)



The screenshot shows a list of properties from the DBpedia ontology. The properties are listed on the left, and their corresponding values are listed on the right. The properties include dbo:cpu, dbo:manufacturer, dbo:operatingSystem, dbo:thumbnail, dbo:type, dbo:weight, and dbo:wikiPageExternalLink. The values for each property are listed as a bulleted list.

dbo:cpu	dbr:ARM_Cortex-A9 dbr:Reduced_instruction_set_computing dbr:32-bit dbr:ARM_architecture dbr:Samsung dbr:ARM_Cortex-A8 dbr:Multi-core_processor
dbo:manufacturer	dbr:Foxconn dbr:Contract_manufacturer dbr:Pegatron
dbo:operatingSystem	dbr:iOS
dbo:thumbnail	wiki-commons:Special:FilePath/IPhone_7_Jet_Black.svg?width=300
dbo:type	dbr:Smartphone
dbo:weight	135.000000 (xsd:double)
dbo:wikiPageExternalLink	http://www.apple.com/iphone https://www.apple.com/iphone/

数据的分类



China Smartphone Shipment Market Share (%)		Q2 2017	Q2 2018	YoY Growth
HUAWEI		20%	26%	22%
OPPO		19%	19%	-9%
vivo	Counterpoint	17%	18%	-1%
Xiaomi		13%	13%	-10%
Apple		8%	9%	0%
Others		23%	16%	-37%
TOTAL		100%	100%	-7%

A screenshot of a mobile application displaying reviews for the iPhone Xs and iPhone Xs Max. The reviews are presented in a card-based format with user profiles, star ratings, and short comments. The data is semi-structured, containing both structured user information and unstructured text reviews.

User	Rating	Comment	Date
6***m PLUS会员	★★★★☆	手机还行，信号不怎么好。。	2018-12-03 10:43
阳***替 PLUS会员	★★★★☆	用起来还好，还是很相信京东的！	2018-12-04 17:18
O***b PLUS会员	★★★★☆	商品很好。信号很差	2018-12-14 18:45
h***8 PLUS会员	★★★★☆	物流速度快，信号是有点问题！	2018-10-03 07:00

全球各地的评论媒体对 iPhone Xs 和 iPhone Xs Max 进行了测试。下面是他们做出的一些评论：

Mashable

“再度改进的摄像头硬件结合了新的‘智能 HDR’自动技术，由神经网络引擎和 A12 仿生的图像信号处理器再添动力，意味着你可以充分享用先进的摄像头光学技术和计算摄影技术带来的益处。”

TechCrunch

“谈到中央处理器性能，这款开创性的规模化 7 纳米架构已带来显著成效。iPhone Xs 拥有可媲美笔记本电脑的运行速度和远超 iPhone X 的处理性能，其架构的成效由此可见一斑。”

Daring Fireball

“iPhone 镜头和感光元件的品质无法与体积更大的专业相机相比，甚至相差较远。这是由于物理定律的限制。但是，传统的相机企业在定制化芯片和软件方面却逊色于 Apple，他们的相机无法像 iPhone 一样便于随身携带，也无法随时连接互联网进行分享。从长期考虑，明智的投资应当用于芯片和软件。”

数据采集：How to Collect

- 按数据源类型进行分类
 - 来自CSV文件
 - 来自JSON文件
 - 来自网页Web Pages
 - 来自关系数据库（如MySQL）
- 来自HDFS
- 来自Web API
- 来自Open Data网站



掌握



了解

从CSV文件读取数据

- CSV的全称是Comma-separated values，是一种用逗号分隔的方式来表示与存储表格数据的文件格式

	Team	Win	Loss	Win%
1	Houston Rockets	20	4	0.833
2	Golden State Warriors	21	6	0.778
3	San Antonio Spurs	19	8	0.704
4	Minnesota Timberwolves	16	11	0.593
5	Denver Nuggets	14	12	0.538
6	Portland Trail Blazers	13	12	0.52
7	New Orleans Pelicans	14	13	0.519
8	Utah Jazz	13	14	0.481
9				
10				

- 技能包
 - 使用Python Pandas读取CSV文件

```
import pandas as pd  
  
df = pd.read_csv('./samples/nba.csv')  
print(df['Team'])  
print(df['Win'].max())
```

扩展阅读：<https://pythonspot.com/pandas-read-csv/>

从JSON文件读取数据

- JSON是一种存储嵌套数据的文件格式（类似Python中的List, Dict）

```
players.json
1 {
2   "Kobe Bryant": {
3     "Born": "08/23/1978",
4     "Number": ["8", "24"],
5     "Team": ["Los Angeles Lakers"]
6   },
7   "Michael Jordan": {
8     "Born": "02/17/1963",
9     "Number": ["23"],
10    "Team": ["Chicago Bulls", "Washington Wizards"]
11  }
12 }
```

- 技能包
 - 使用Python **Pandas**读取JSON文件

```
import pandas as pd
df = pd.read_json('./samples/nba.json')
```

阅读：http://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_json.html

从网页获取数据

- 访问网页

- urllib2 (<https://docs.python.org/2/library/urllib2.html>)
- request (<http://docs.python-requests.org/en/master/>)

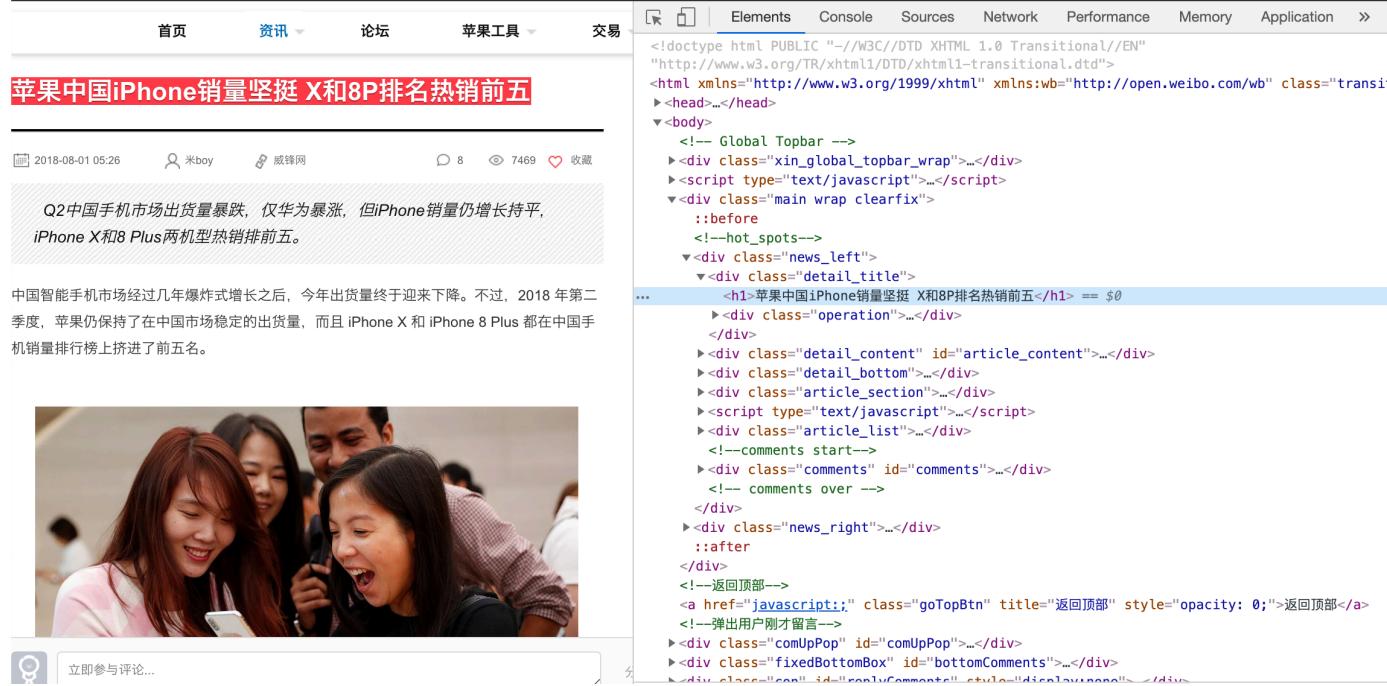
```
import urllib2
response = urllib2.urlopen('http://python.org/')
print ("Response:", response)

# Get the URL. This gets the real URL.
print ("The URL is: ", response.geturl())

# Getting the code
print ("This gets the code: ", response.code)
```

从网页获取数据

- 解析网页 (Parsing)
 - 正则表达式解析 re
 - BeautifulSoup
(<https://www.crummy.com/software/BeautifulSoup/>)
 - lxml (<http://lxml.de/>)



苹果中国iPhone销量坚挺 X和8P排名热销前五

2018-08-01 05:26 米boy 威锋网 8 7469 收藏

Q2中国手机市场出货量暴跌，仅华为暴涨，但iPhone销量仍增长持平，iPhone X和8 Plus两机型热销排前五。

中国智能手机市场经过几年爆炸式增长之后，今年出货量终于迎来下降。不过，2018年第二季度，苹果仍保持了在中国市场稳定的出货量，而且iPhone X和iPhone 8 Plus都在中国手机销量排行榜上挤进了前五名。



立即参与评论...

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xmlns:wb="http://open.weibo.com/wb" class="transi
><head>...</head>
<body>
    <!-- Global Topbar -->
    <div class="xin_global_topbar_wrap">...</div>
    <script type="text/javascript">...</script>
    <div class="main_wrap clearfix">
        :before
        <!--hot_spots-->
        <div class="news_left">
            <div class="detail_title">
                ...<h1>苹果中国iPhone销量坚挺 X和8P排名热销前五</h1> == $0
                <div class="operation">...</div>
            </div>
            <div class="detail_content" id="article_content">...</div>
            <div class="detail_bottom">...</div>
            <div class="article_section">...</div>
            <script type="text/javascript">...</script>
            <div class="article_list">...</div>
            <!--comments start-->
            <div class="comments" id="comments">...</div>
            <!-- comments over -->
        </div>
        <div class="news_right">...</div>
        ::after
    </div>
    <!--返回顶部-->
    <a href="#" class="goTopBtn" title="返回顶部" style="opacity: 0;">返回顶部</a>
    <!--弹出用户刚才留言-->
    <div class="comUpPop" id="comUpPop">...</div>
    <div class="fixedBottomBox" id="bottomComments">...</div>
    <div class="comment_id replyComment" id="comment_id" style="display:none">...</div>
    <div class="comment_id replyComment" id="comment_id" style="display:none">...</div>
```

从网页获取数据

- 解析网页 (Parsing)
 - 正则表达式解析 re
 - Beautiful Soup
(<https://www.crummy.com/software/BeautifulSoup/>)
 - lxml (<http://lxml.de/>)

```
import urllib2
from bs4 import BeautifulSoup

optionsUrl = 'http://finance.yahoo.com/q/op?s=AAPL+Options'
optionsPage = urllib2.urlopen(optionsUrl)
soup = BeautifulSoup(optionsPage)
soup.findAll(text='AAPL130328C00350000')
# [u'AAPL130328C00350000']
soup.findAll(text='AAPL130328C00350000')[0].parent.parent
# <td><a href="/q?s=AAPL130328C00350000">AAPL130328C00350000</a></td>
```

阅读 <https://www.pythongcentral.io/python-beautiful-soup-example-yahoo-finance-scraper/>

从网页获取数据

- 网页数据获取套装
 - Scrapy (<https://scrapy.org/>)
- 网页数据获取经验谈
 - 劳动力密集型：网页“千站千面”
 - 横看成岭侧成峰，远近高低各不同
 - 不识庐山真面目，边吐老血边coding

阅读 <https://www.analyticsvidhya.com/blog/2017/07/web-scraping-in-python-using-scrapy/>

从关系数据库获取数据

- 以MySQL为例

- 创建连接

```
import datetime
import mysql.connector

cnx = mysql.connector.connect(user='scott', database='employees')
cursor = cnx.cursor()
```

- 写SQL语句

```
query = ("SELECT first_name, last_name, hire_date FROM employees "
          "WHERE hire_date BETWEEN %s AND %s")
```

```
hire_start = datetime.date(1999, 1, 1)
hire_end = datetime.date(1999, 12, 31)
```

- 执行SQL语句

```
cursor.execute(query, (hire_start, hire_end))

for (first_name, last_name, hire_date) in cursor:
    print("{} {}, {} was hired on {}".format(
        last_name, first_name, hire_date))
```

- 解析结果

```
cursor.close()
cnx.close()
```

<https://dev.mysql.com/doc/connector-python/en/connector-python-examples.html>