



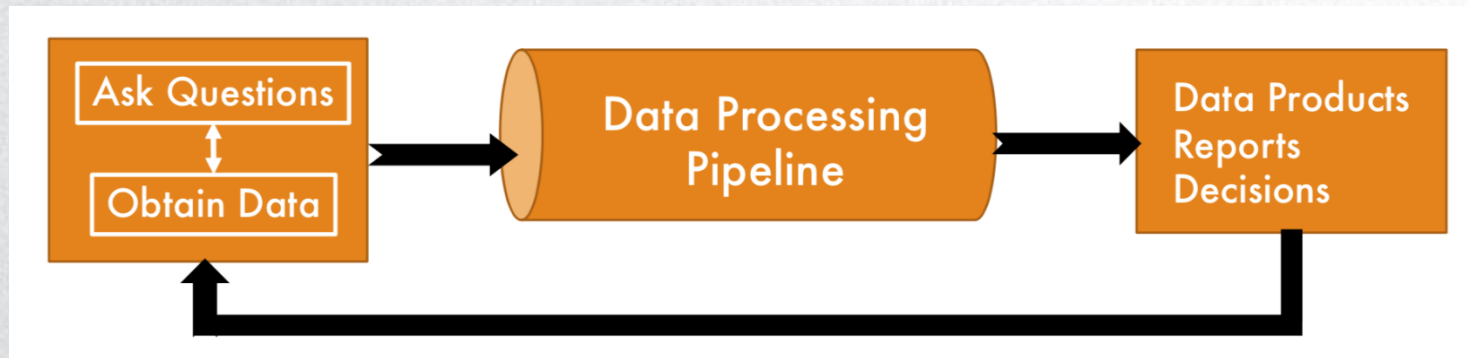
数据模型与数据采集



覃雄派

数据模型与数据采集

- 数据科学的工作流程



循环迭代式的工作流程

- 先提出问题，再收集与分析相关的数据
- 先收集数据，再分析可以回答哪些问题

数据模型与数据采集

- 数据科学的工作流程
- 三个基本任务
 - 获取原始数据
 - 准备待分析数据
 - 针对特定问题进行数据分析

数据采集
数据准备
数据分析

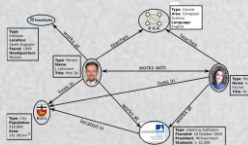
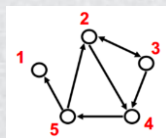
数据采集

	item ₁	item ₂	item ₃	...	item _m
user ₁	3	5	2		1
user ₂	1		3		
...					
user _n	5		4		2
user _m		4			3

Primary Key	Partition Key	Sort Key
Product ID	Type	
1	Book D	
2	Album D	
2	Track D	
3	Movie D	

Attributes
Schema is defined per Item
Odyssey Homer 1871
Parties Bach
Parties Bach
The Kid Drama, Comedy
Chaplin

ID	Name	Contact
NA01	Hello World Tech.	334-55-7476
NA02	ABC Technologies	283-92-8311



来源：科技日报

据《科学》网站最新发布的消息，超过40%的昆虫物种可能在未来二十年内灭绝。其中蝴蝶、蜜蜂和蛾类等受到的影响最大。主要原因是栖息地的丧失，以及对杀虫剂使用造成的负面影响。

“这种影响对全球生态系统将是灾难性的，因为昆虫是地球上许多生态系统的基石。”

研究人员说，昆虫减少的最大原因是栖息地丧失。其次，杀虫剂和农药的过度使用。例如，杀虫剂的使用导致蜜蜂种群数量的下降。最后，气候变化和全球变暖，导致昆虫的生存环境变得更加恶劣。研究人员已经提出全球变暖将有助于下降。



数据准备

数据分析

特征				标签
...	1
...	0

待分析数据



数据模型与数据采集

- 数据科学的工作流程
 - 数据科学与烹饪



买菜

数据采集



洗菜



备菜

数据准备



炒菜

数据分析

本讲重点：数据、数据采集 + 数据准备 → 给分析算法准备更**优质**的数据



提纲

- 数据模型
- 数据采集

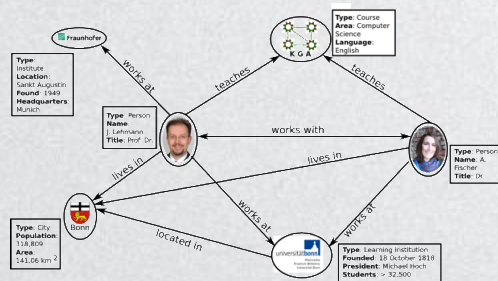
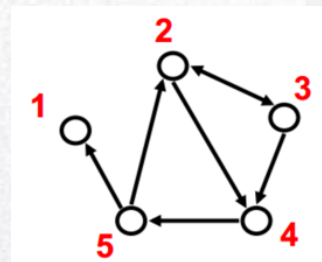
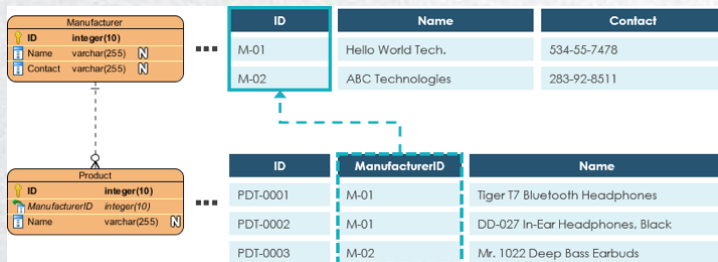
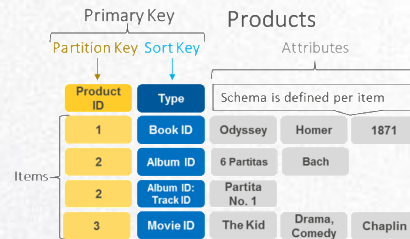


数据模型与数据采集

数据模型与数据采集

- 数据的种类繁多
- Variety: 数据的种类繁多
 - 数组、矩阵
 - 键值对
 - 实体-关系表
 - 时序数据、流数据
 - 图数据
 - 文本数据
 - 多媒体数据
 - ...

	$item_1$	$item_2$	$item_3$...	$item_n$
$user_1$		5	2		1
$user_2$	3				
$user_3$	1		3		
...					
$user_{m-1}$	5		4		2
$user_m$		4			3



来源: 科技日报

据《新科学家》网站最新发布的信息, 超过40%的昆虫物种可能在未来几十年内灭绝, 其中蝴蝶、蜜蜂和苍蝇受到的影响最大, 主要原因是栖息地的丧失。这是对过去40年来所有昆虫长期调查得出的令人震惊的结论。

“这种影响对地球生态系统将是灾难性的, 因为昆虫是世界上许多生态系统的基石。”论文作者说, 他们来自澳大利亚悉尼大学和中国农业科学院。

研究发现, 昆虫减少的最大原因是栖息地丧失; 其次, 寄生虫和疾病也起着重要作用, 例如, 瓦螨的蔓延导致蜜蜂种群的衰退; 最后, 气候变化似乎也有影响, 热带地区的昆虫可能对温度变化的耐受性较差, 其数量可能已经因全球变暖而有所下降。



数据模型与数据采集

- 数组与矩阵
- 数据项同类型，可以利用下标访问
 - 例子：NumPy的多维数组 (ndarray)
 - 例子：推荐系统中的user-item矩阵



两个用户对三个商品打分：

- $u_1 \rightarrow 1(5); 3(2)$
- $u_2 \rightarrow 2(3); 3(5)$

请用**NumPy**构造矩阵

A. `mat = np.array([[5,0,2],[0,3,5]])`

B. `mat = np.array([[5,np.nan,2],[np.nan,3,5]])`

π
@
R
U
C

商品

用户

	<i>item₁</i>	<i>item₂</i>	<i>item₃</i>	...	<i>item_n</i>
<i>user₁</i>		5	2		1
<i>user₂</i>	3				
<i>user₃</i>	1		3		
.					
.					
.					
<i>user_{m-1}</i>	5		4		2
<i>user_m</i>		4			3

评分

数据模型与数据采集

- 数组与矩阵
- 数据项同类型，可以利用下标访问
 - 例子：NumPy的多维数组 (ndarray)
 - 例子：推荐系统中的user-item矩阵



两个用户对三个商品打分：

- $u_1 \rightarrow 1(5); 3(2)$
- $u_2 \rightarrow 2(3); 3(5)$

请用**NumPy**构造矩阵

A. `mat = np.array([[5,0,2],[0,3,5]])`

B. `mat = np.array([[5,np.nan,2],[np.nan,3,5]])`

```
import numpy as np
mat = np.array([[5,np.nan,2],[np.nan,3,5]])
mat
array([[ 5., nan,  2.],
       [nan,  3.,  5.]])
```

商品

用户

	<i>item₁</i>	<i>item₂</i>	<i>item₃</i>	...	<i>item_n</i>
<i>user₁</i>		5	2		1
<i>user₂</i>	3				
<i>user₃</i>	1		3		
.					
.					
.					
<i>user_{m-1}</i>	5		4		2
<i>user_m</i>		4			3

评分

数据模型与数据采集

- 关系数据 (Relational Data)
- 简单的关系数据：单表数据
 - 行：表示一条记录 (Record)
 - 列：表示一个属性 (Attribute)

Team	Win	Loss	Win%
Houston Rockets	20	4	0.83
Golden State Warriors	21	6	0.78
San Antonio Spurs	19	8	0.7
Minnesota Timberwolves	16	11	0.59
Denver Nuggets	14	12	0.54
Portland Trail Blazers	13	12	0.52
New Orleans Pelicans	14	13	0.52
Utah Jazz	13	14	0.48

使用pandas表示单表数据

```
nba_df = pd.DataFrame ({'Team': team_col,
                        'Win': win_col,
                        'Loss': loss_col})
print (nba_df)
```

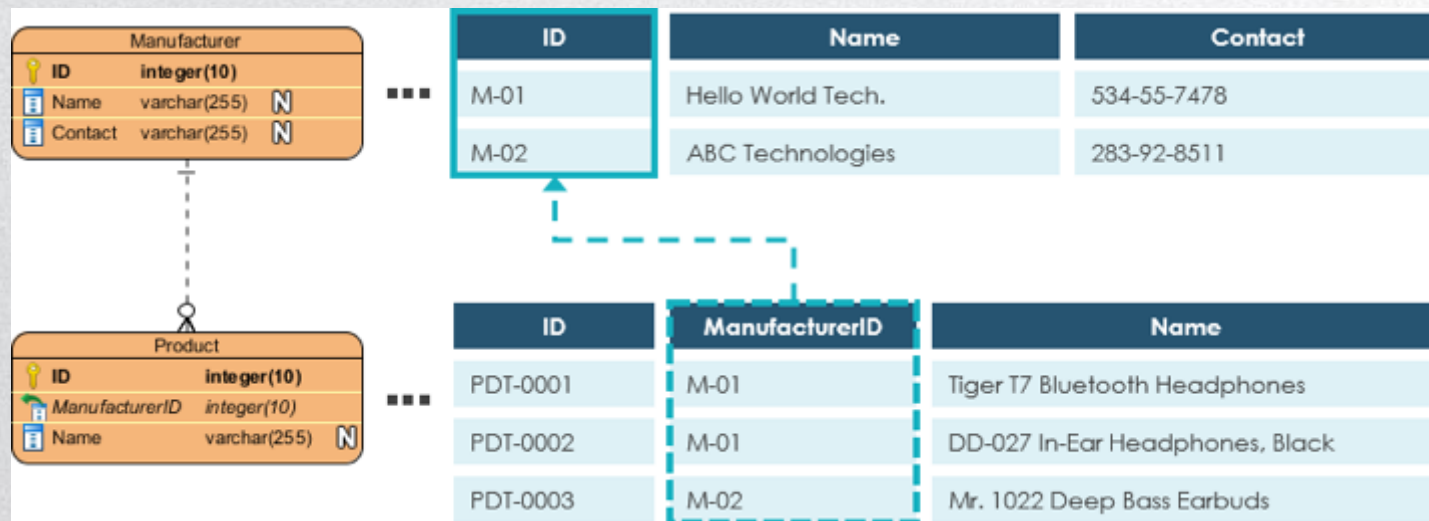
列标签

	Team	Win	Loss
0	Houston Rockets	20	4
1	Golden State Warriors	21	6
2	San Antonio Spurs	19	8
3	Minnesota Timberwolves	16	11
4	Denver Nuggets	14	12
5	Portland Trail Blazers	13	12
6	New Orleans Pelicans	14	13
7	Utah Jazz	13	14

行标签

数据模型与数据采集

- 关系数据 (Relational Data)
- 关系数据库：将数据表示为多个彼此可关联的表格
 - ER模型组织数据
 - 表格、属性、主外键



数据模型与数据采集

- 文本数据
- 自然语言是人们交流信息最为自然的表达方式
 - 互联网网页、论坛评论等
 - 企业文档
 - 聊天记录

来源：科技日报

据《新科学家》网站最新发布的消息，超过40%的昆虫物种可能在未来几十年内灭绝，其中蝴蝶、蜜蜂和螳螂受到的影响最大，主要原因是栖息地的丧失。这是对过去40年来所有昆虫长期调查得出的令人震惊的结论。

“这种影响对地球生态系统将是灾难性的，因为昆虫是世界上许多生态系统的基础。”论文作者说，他们来自澳大利亚悉尼大学和中国农业科学院。

研究发现，昆虫减少的最大原因是栖息地丧失；其次，寄生虫和疾病也起着重要作用，例如，瓦螨的蔓延导致蜜蜂种群的衰退；最后，气候变化似乎也有影响，热带地区的昆虫可能对温度变化的耐受性较差，其数量可能已经因全球变暖而有所下降。

- 缺少结构支持，给文本分析处理带来巨大挑战
- 理解词语、实体、句子、关系等
- 自然语言的语义鸿沟

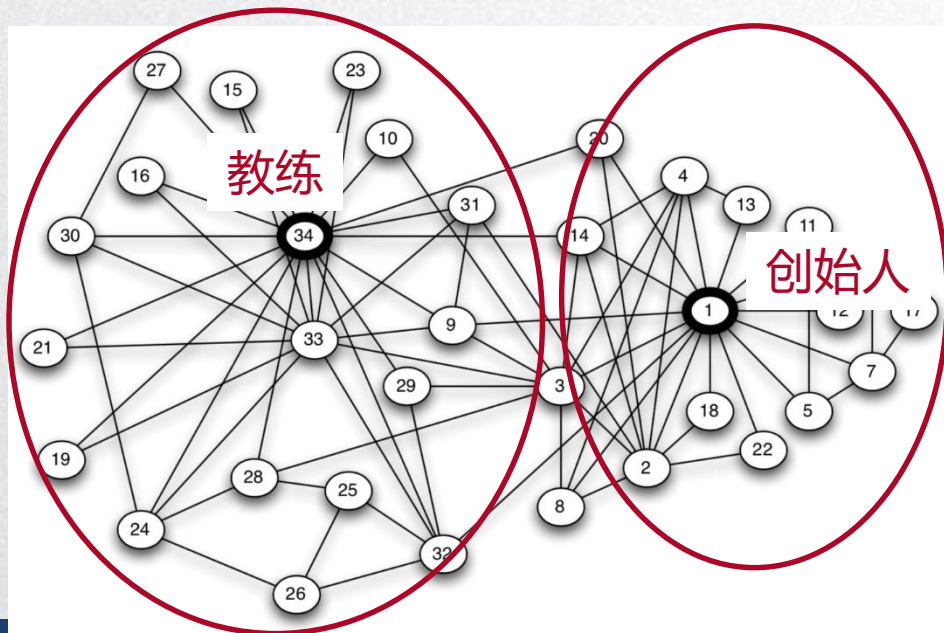
数据模型与数据采集

- 图数据
- 顶点一般表示实体或者属性值
 - 顶点之间的边，表示被连接的两个顶点间的关系
 - 实例
 - 社交网络
 - 知识图谱



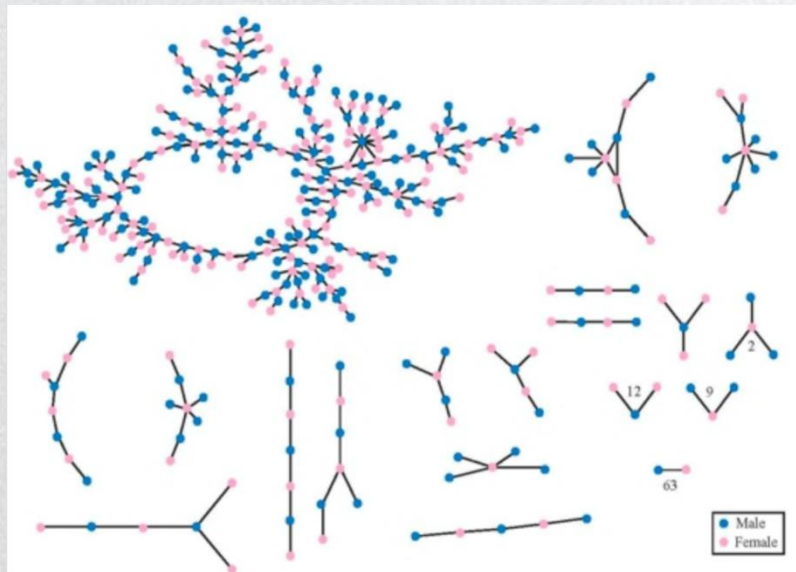
请你预言该俱乐部在不久的将来会：

- A.** 分裂为两个俱乐部
- B.** 团结在创始人的周围



数据模型与数据采集

- 图数据：直观地理解群体的行为
 - 例子：美国高中生恋爱关系图（边代表二人在18个月内恋爱过）



Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks

Peter S. Bearman
Columbia University

James Moody
Ohio State University

Katherine Stovel
University of Washington

July 2004 · *American Journal of Sociology*. 110(1)

DOI:[10.1086/386272](https://doi.org/10.1086/386272)

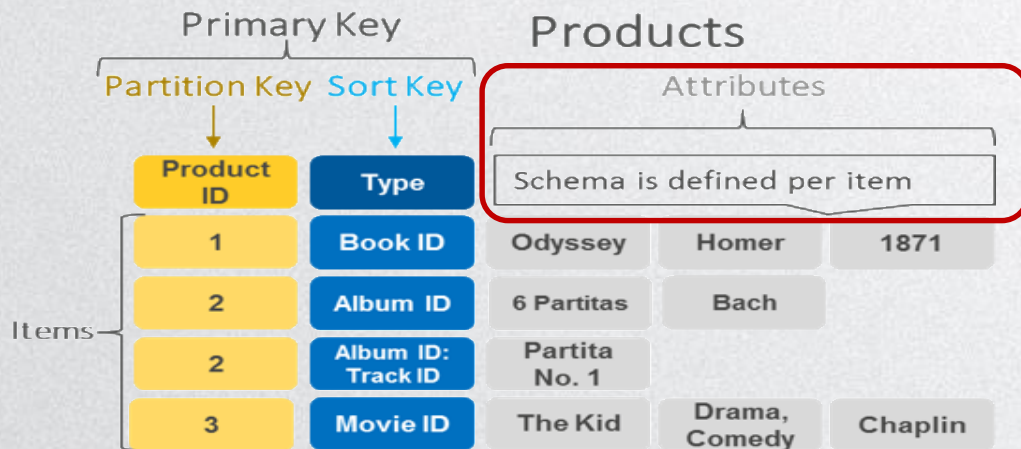
数据模型与数据采集

- 时序数据
- 随时间不断变化或累积的数据
 - 每个数据项有时间戳
 - 关注一段时间内的数据值变化、关注异常值
 - 新的数据价值更高
 - 多用于监控传感等场景



数据模型与数据采集

- 键值对
- 键值对灵活定义属性，每行可以有多个不同的属性
 - 例子：用户画像
 - 通过键直接访问值
 - 简单的如Hash table, Map等数据结构



数据模型与数据采集

- 多媒体数据
- 图像、视频、音频等
 - 多种媒体类型的混合
 - 更关注语义
 - 处理复杂，计算代价高
 - 数据量相对更大
 - 在自媒体应用中普遍存在



【简介】比尔及梅琳达·盖茨基金会联席主席比尔·盖茨12日在通过新华社独家发布的视频里说，过去一年里中国在促进全球发展方面继续作出重要贡献。具体聊了哪些贡献？快戳视频看看吧！



ISSN: 1077-3142

Computer Vision and Image Understanding

Editor-in-Chief: [N. Paragios](#)

> [View Editorial Board](#)

> [CiteScore: 8.7](#) ^① [Impact Factor: 3.121](#) ^②



数据模型与数据采集

- 大数据时代：多模态数据并存
 - 以关系数据为代表的结构化数据
 - 数据量占比低于20%
 - 数据价值相对高
 - 以文本、图数据为代表的非结构化数据
 - 数据量占比高于80%
 - 数据价值相对低
 - 需要融合结构化数据和非结构化数据
 - 信息抽取
 - 实体链接与数据融合

数据模型与数据采集



数据模型与数据采集

- 数据的采集：数据采集案例
- 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？



数据模型与数据采集

- 数据的采集：Where to Collect
- 你要采集哪些数据来支撑你的分析？
- 内部数据
 - 产品数据库（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）

数据模型与数据采集

- 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：

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

华为依然是中国市场的老大，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%。

数据模型与数据采集

- 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 interface. It features a sidebar with navigation links and a main content area with two tables of API endpoints.

微博		
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	第三方分享到微博 

评论		
	comments/show	获取某条微博的评论列表
	comments/by_me	我发出的评论列表

数据模型与数据采集

- 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)?

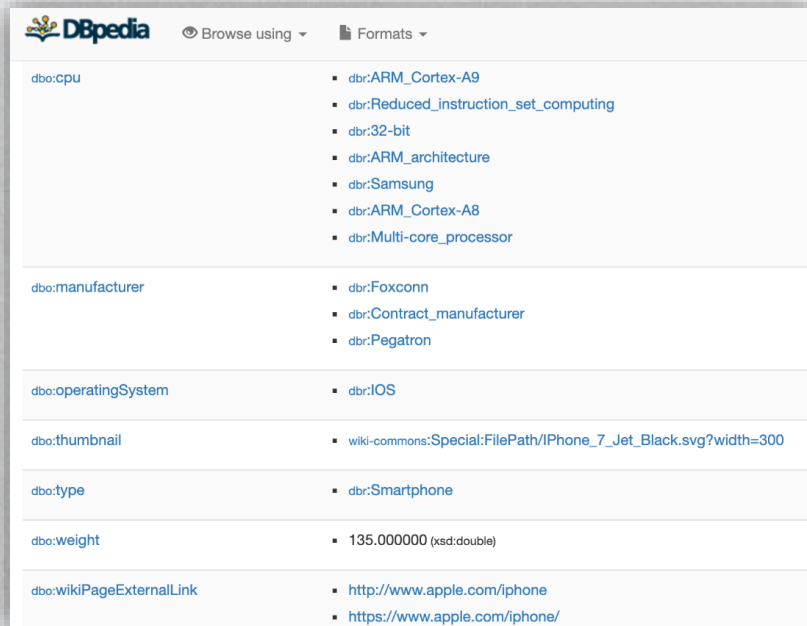
国内资源不完全统计：

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/

数据模型与数据采集

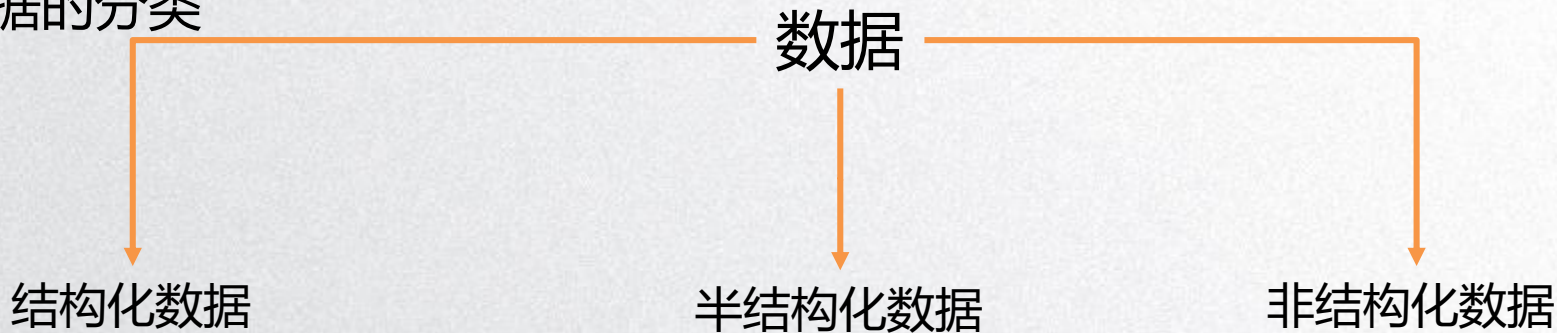
- 数据采集：Where to Collect
- 你要采集哪些数据来支撑你的分析？
- 外部数据
 - 网页数据
 - Web API
 - 开放数据 (Open Data)
 - 知识图谱 (DBpedia)



dbo:cpu	<ul style="list-style-type: none">dbr:ARM_Cortex-A9dbr:Reduced_instruction_set_computingdbr:32-bitdbr:ARM_architecturedbr:Samsungdbr:ARM_Cortex-A8dbr:Multi-core_processor
dbo:manufacturer	<ul style="list-style-type: none">dbr:Foxconndbr:Contract_manufacturerdbr:Pegatron
dbo:operatingSystem	<ul style="list-style-type: none">dbr:iOS
dbo:thumbnail	<ul style="list-style-type: none">wiki-commons:Special:FilePath/IPhone_7_Jet_Black.svg?width=300
dbo:type	<ul style="list-style-type: none">dbr:Smartphone
dbo:weight	<ul style="list-style-type: none">135.000000 (xsd:double)
dbo:wikiPageExternalLink	<ul style="list-style-type: none">http://www.apple.com/iphonehttps://www.apple.com/iphone/

数据模型与数据采集

数据的分类



结构化数据

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

Counterpoint
Strategy Analytics

半结构化数据

6***m PLUS会员	★★★★☆ 手机还行，信号不怎么好。	银色 公开版 256GB 2018-12-03 10:43
阳***哥	★★★★☆ 用起来还好，还是很相信京东的！	深空灰色 公开版 256GB 2018-12-04 17:18
O***b	★★★★☆ 商品很好。信号很差	深空灰色 公开版 64GB 2018-12-14 18:45
h***8 PLUS会员	★★★★☆ 物流速度快，信号有点问题！	深空灰色 公开版 256GB 2018-10-03 07:00

非结构化数据

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。”

数据模型与数据采集

- 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



可选掌握



了解

数据模型与数据采集



数据模型与数据采集

- 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
- 使用Python **Pandas**读取CSV文件

```
import pandas as pd

df = pd.read_csv("./employee.csv", delimiter=',')
df.head()
```

	EMPID	FirstName	LastName	Salary
0	1001	Amal	Jose	100000
1	1002	Edward	Joe	100001
2	1003	Sabitha	Sunny	210000
3	1004	John	P	50000
4	1005	Mohammad	S	75000

数据模型与数据采集

- 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）

```
df2 = pd.read_json("./employee.json")  
df2.head()
```

	EMPID	FirstName	LastName	Salary
0	1001	Amal	Jose	100000
1	1002	Edward	Joe	100001
2	1003	Sabitha	Sunny	210000
3	1004	John	P	50000
4	1005	Mohammad	S	75000

```
1 [{"EMPID":1001,"FirstName":"Amal","LastName":"Jose","Salary":100000},  
2 {"EMPID":1002,"FirstName":"Edward","LastName":"Joe","Salary":100001},  
3 {"EMPID":1003,"FirstName":"Sabitha","LastName":"Sunny","Salary":210000},  
4 {"EMPID":1004,"FirstName":"John","LastName":"P","Salary":50000},  
5 {"EMPID":1005,"FirstName":"Mohammad","LastName":"S","Salary":75000}]
```

employee.json的内容



数据模型与数据采集

- 读取网页数据
 - requests库

```
import requests
```

```
r = requests.get('https://www.baidu.com')
r.encoding=requests.utils.get_encodings_from_content(r.text)
# 注意get_encodings_from_content的参数是字符串，所以要用r.text而不是r.content
print(r.text)
```

```
<!DOCTYPE html>
<!--STATUS OK--><html> <head><meta http-equiv=content-type content=text/html;charset=utf-8><meta http-equiv=X-UA-Compatible content=IE=Ed
ge><meta content=always name=referrer><link rel=stylesheet type=text/css href=https://ssl.bdstatic.com/5eN1bjq8AAUYm2zgoY3K/r/www/cache/b
dor/baidu.min.css><title>百度一下，你就知道</title></head> <body link=#0000cc> <div id=wrapper> <div id=head> <div class=head_wrapper> <
div class=s_form> <div class=s_form_wrapper> <div id=lg> <img hidefocus=true src=/www.baidu.com/img/bd_logol.png width=270 height=129>
</div> <form id=form name=f action=/www.baidu.com/s class=fm> <input type=hidden name=bdorz_come value=1> <input type=hidden name=ie val
ue=utf-8> <input type=hidden name=f value=8> <input type=hidden name=rsv_bp value=1> <input type=hidden name=rsv_idx value=1> <input type
=hidden name=tn value=baidu><span class="bg s_ipt_wr"><input id=kwd name=wd class=s_ipt value maxlength=255 autocomplete=off autofocus=aut
ofocus></span><span class="bg s_btn_wr"><input type=submit id=su value=百度一下 class="bg s_btn" autofocus></span> </form> </div> </div>
<div id=ul> <a href=http://news.baidu.com name=tj_trnews class=mnav>新闻</a> <a href=https://www.hao123.com name=tj_trhao123 class=mnav>h
ao123</a> <a href=http://map.baidu.com name=tj_trmap class=mnav>地图</a> <a href=http://v.baidu.com name=tj_trvideo class=mnav>视频</a> <
a href=http://tieba.baidu.com name=tj_trtieba class=mnav>贴吧</a> <noscript> <a href=http://www.baidu.com/bdor/login.gif?login&tpl=mn&u=http%3A%2F%2Fwww.baidu.com%2F%3fbdorz_come%3d1 name=tj_login class=lb>登录</a> </noscript> <script>document.write(' <a href="htt
p://www.baidu.com/bdor/login.gif?login&tpl=mn&u=' + encodeURIComponent(window.location.href+ (window.location.search == "" ? "?" : "&")+
"bdorz_come=1")+ ' ' name="tj_login" class="lb">登录</a>');
</script> <a href=/www.baidu.com/more/ name=tj_briicon class=bri style="display: block;">更多产品</a> </div> </div> </di
v> <div id=ftCon> <div id=ftConw> <p id=lh> <a href=http://home.baidu.com>关于百度</a> <a href=http://ir.baidu.com>About Baidu</a> </p> <
p id=cp>&copy;2017&nbsp;Baidu&nbsp;<a href=http://www.baidu.com/duty/>使用百度前必读</a>&nbsp;<a href=http://jianyi.baidu.com/ class=cp-
feedback>意见反馈</a>&nbsp;<a href=http://www.baidu.com/img/g.gif></p> </div> </div> </div> </div> </body> </html>
```




数据模型与数据采集

- 读取网页数据：中文网页
 - requests库

```
import requests
```

```
url = "https://new.qq.com/omn/20211111/20211111A0AQ7700.html"  
r = requests.get(url)  
r.encoding='gb2312' # 根据网页编码设置
```

```
print(r.text)  
mytext = r.text
```

```
<!DOCTYPE html>  
<html lang="zh-CN" dir="ltr">  
  <head>  
    <title>北京谱仪精确测量中子电磁结构 揭开光子-核子相互作用之谜_腾讯新闻</title>  
    <meta name="keywords" content="北京谱仪精确测量中子电磁结构 揭开光子-核子相互作用之谜,中科院高能所,光子,质子,中子,北京谱仪">  
    <meta name="description" content="《自然·物理》以封面文章形式发表成果论文。 中科院高能所 供图中新网北京11月11日电 (记者 孙自法)记者11日从中国科学院高能物理研究所(中科院高能所)获悉,北京谱仪III(BESIII)作为北……">  
    <meta name="apub:time" content="11/11/2021, 8:51:41 PM">  
    <meta name="apub:from" content="default">  
    <meta http-equiv="X-UA-Compatible" content="IE=Edge" />  
    <link rel="stylesheet" href="//mat1.gtimg.com/qqcdn/qqindex2021/qqdc/css/index.css" />  
    <!--[if lte IE 8]><meta http-equiv="refresh" content="0; url=/upgrade.htm"><![endif]-->  
    <!-- <meta name="sogou_site_verification" content="SYWy6ahy7s"/> -->  
    <meta name="baidu-site-verification" content="jJeIJ5X7pP" />  
    <link rel="shortcut icon" href="//mat1.gtimg.com/www/icon/favicon2.ico" />  
    <link rel="stylesheet" href="//vm.gtimg.com/tencentvideo/txp/style/txp_desktop.css" />  
    <script src="//is.qq.com/is/qq_common.js"></script>
```

数据模型与数据采集

- 读取网页数据：中文网页
 - String的split
 - 切割新闻主体内容

```
temp = mytext.split("<div class=\"content-article\">")[1]
temp = temp.split("<div id=\"Status\"></div>")[0]
print(temp)
```

```
<!--导语-->
<p class="one-p">
</p>
<p class="one-p">《自然·物理》以封面文章形式发表成果论文。 中科院高能所 供图</p>
<p class="one-p">中新网北京11月11日电 (记者 孙自法) 记者11日从中国科学院高能物理研究所(中科院高能所)获悉,北京谱仪III(BESIII)
作为北京正负电子对撞机核心科研装置之一,其国际合作组最近已实现对中子电磁结构精确测量,从而揭开困扰学界20多年的光子-核子相互作用之谜。</p>
<p class="one-p">北京谱仪III国际合作组最新完成的对中子的类时电磁形状因子进行精确测量,实验结果不仅解决了长期存在的光子-核子耦合反常的问题,还观测到中子电磁形状因子随质心能量变化的周期性振荡结构。这项重要物理实验结果论文,近日以封面文章形式在国际学术期刊《自然·物理》发表。</p>
<p class="one-p">据中科院高能所实验物理中心介绍,中子和质子统称为核子,它们是构成可见物质世界的主要成分。迄今为止核子的内部结构仍有许多未解之谜,长达20余年的光子-核子相互作用之谜即是其中之一。1998年意大利芬尼斯(FENICE)实验首次测量了中子的类时电磁形状因子,其结果表明光子-中子相互作用强于光子-质子相互作用,与夸克模型预期不符。</p>
<p class="one-p">
</p>
```

切割以后的结果

数据模型与数据采集

- 读取网页数据：中文网页
 - String的replace
 - 去除无关html标记<p> </p>

```
temp = temp.replace("<p class=\"one-p\">", "")  
print(temp)
```

```
temp = temp.replace("</p>", "")  
print(temp)
```

<!--导语-->

《自然·物理》以封面文章形式发表成果论文。 中科院高能所 供图

中新网北京11月11日电 (记者 孙自法) 记者11日从中国科学院高能物理研究所(中科院高能所)获悉,北京谱仪III(BESIII)作为北京正负电子对撞机核心科研装置之一,其国际合作组最近已实现对中子电磁结构精确测量,从而揭开困扰学界20多年的光子-核子相互作用之谜。

北京谱仪III国际合作组最新完成的对中子的类时电磁形状因子进行精确测量,实验结果不仅解决了长期存在的光子-核子耦合反常的问题,还观测到中子电磁形状因子随质心能量变化的周期性振荡结构。这项重要物理实验结果论文,近日以封面文章形式在国际学术期刊《自然·物理》发表。

据中科院高能所实验物理中心介绍,中子和质子统称为核子,它们是构成可见物质世界的主要成分。迄今为止核子的内部结构仍有许多未解之谜,长达20余年的光子-核子相互作用之谜即是其中之一。1998年意大利芬尼斯(FENICE)实验首次测量了中子的类时电磁形状因子,其结果表明光子-中子相互作用强于光子-质子相互作用,与夸克模型预期不符。

数据模型与数据采集

- 读取网页数据：中文网页
 - 正则表达式
 - 去除无关html标记

```
import re
s = temp
replaced = re.sub('<img .*>', '', s)
print (replaced )
```

<!--导语-->

《自然·物理》以封面文章形式发表成果论文。 中科院高能所 供图

中新网北京11月11日电（记者 孙自法）记者11日从中国科学院高能物理研究所（中科院高能所）获悉，北京谱仪III（BESIII）作为北京正负电子对撞机核心科研装置之一，其国际合作组最近已实现对中子电磁结构精确测量，从而揭开困扰学界20多年的光子-核子相互作用之谜。

北京谱仪III国际合作组最新完成的对中子的类时电磁形状因子进行精确测量，实验结果不仅解决了长期存在的光子-核子耦合反常的问题，还观测到中子电磁形状因子随质心能量变化的周期性振荡结构。这项重要物理实验结果论文，近日以封面文章形式在国际学术期刊《自然·物理》发表。

据中科院高能所实验物理中心介绍，中子和质子统称为核子，它们是构成可见物质世界的主要成分。迄今为止核子的内部结构仍有许多未解之谜，长达20余年的光子-核子相互作用之谜即是其中之一。1998年意大利芬尼斯（FENICE）实验首次测量了中子的类时电磁形状因子，其结果表明光子-中子相互作用强于光子-质子相互作用，与夸克模型预期不符。



数据模型与数据采集

- 从html网页解析 (Parsing) 内容
 - 可以选用如下python库
 - 正则表达式解析 re
 - BeautifulSoup (<https://www.crummy.com/software/BeautifulSoup/>)
 - lxml (<http://lxml.de/>)

此处不展开讨论

数据模型与数据采集

- 获取html网页以后
 - 可以从文本数据中抽取结构化信息

For years, [Microsoft Corporation](#) [CEO](#) [Bill Gates](#) was against open source. But today he appears to have changed his mind. "We can be open source. We love the concept of shared source," said [Bill Veghte](#), a [Microsoft VP](#). "That's a super-important shift for us in terms of code access. "

[Richard Stallman](#), [founder](#) of the [Free Software Foundation](#), countered saying...

PEOPLE

<u>Name</u>	<u>Title</u>	<u>Organization</u>
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	Founder	Free Soft..

Select Name
From PEOPLE
Where Organization = 'Microsoft'

Bill Gates
Bill Veghte

数据模型与数据采集

- 从关系数据库获取数据

- 以MySQL数据库为例

- 创建连接

- 写SQL语句

- 执行SQL语句

- 解析结果

- 关闭连接

```
import pymysql

# Open database connection
con = pymysql.connect(host='localhost',
                      user='root',
                      password='rootroot',
                      database='test1',
                      cursorclass=pymysql.cursors.DictCursor)

# prepare a cursor object using cursor() method
cursor = con.cursor()
sql = "select * from namelist"
# Execute the SQL command
cursor.execute(sql)

# Fetch all the rows in a list of lists.
results = cursor.fetchall()
for row in results:
    id = row['id']
    name = row['name']
    # Now print fetched result
    print ("id=%s,name=%s" % (id, name))

# disconnect from server
con.close()
```

```
id=1, name=徐君
id=2, name=陈跃国
id=3, name=覃雄派
```


数据模型与数据采集

- 数据模型小结
 - 不同类型的数据与数据模型
 - 人们如何理解与表达数据
 - 计算机如何存储与处理数据

关系数据库里使用的
数据模型三要素

数据结构：描述数据有由什么元素构成，是什么类型，有什么关系等

数据操作：可以施加于数据对象的操作以及相关规则

数据完整性约束条件：指在给定的数据模型中，数据及其联系所遵守的一组通用的规则，以保证数据的正确性和一致性



数据模型与数据采集

- 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”

文本模块的数据采集部分，展开讲述

数据模型与数据采集

