Big Data : A Practitioner's Perspective

叶邦宇
中科院信工所信息检索组
蚂蚁金融服务集团支付宝基础数据部

What is Big Data

Google Web Search: 1999 vs 2010

- # 文档数: 数千万 to 数万亿

---100000X

- 更新频率: 几个月to 数十秒

---50000X

- 查询所需时间: <1s to <0.2s

---5X

What is Big Data

- 众说纷纭！
- 个人认为，大数据中的“大”，不仅仅是涉及数据规模，而且包含“价值”这个层面。

task1

- 某父亲怒气冲冲地进到一家商店斥责经理给他女儿邮箱发婴幼儿用品的优惠券，而他女儿还是个高中生而已。经理后来打电话跟他表示歉意时，这位父亲却反过来向他道歉，说女儿瞒着他发生了一些事，确实怀孕了。

task2

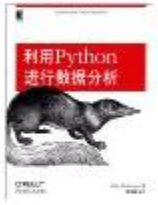
task3

购买此商品的顾客也同时购买

统计学习方法
李航
97
平装
￥28.00

计算机科学丛书:机器学习
米歇尔 (Mitc ...
107
平装
￥24.20

Python自然语言处理
伯德 (Steve ...
11
平装
￥72.00



利用Python进行数据分析
麦金尼 (Wes ...
30
平装
￥62.30

task4

支付峰值：每分钟285万笔

订单创建峰值：每秒钟8万笔

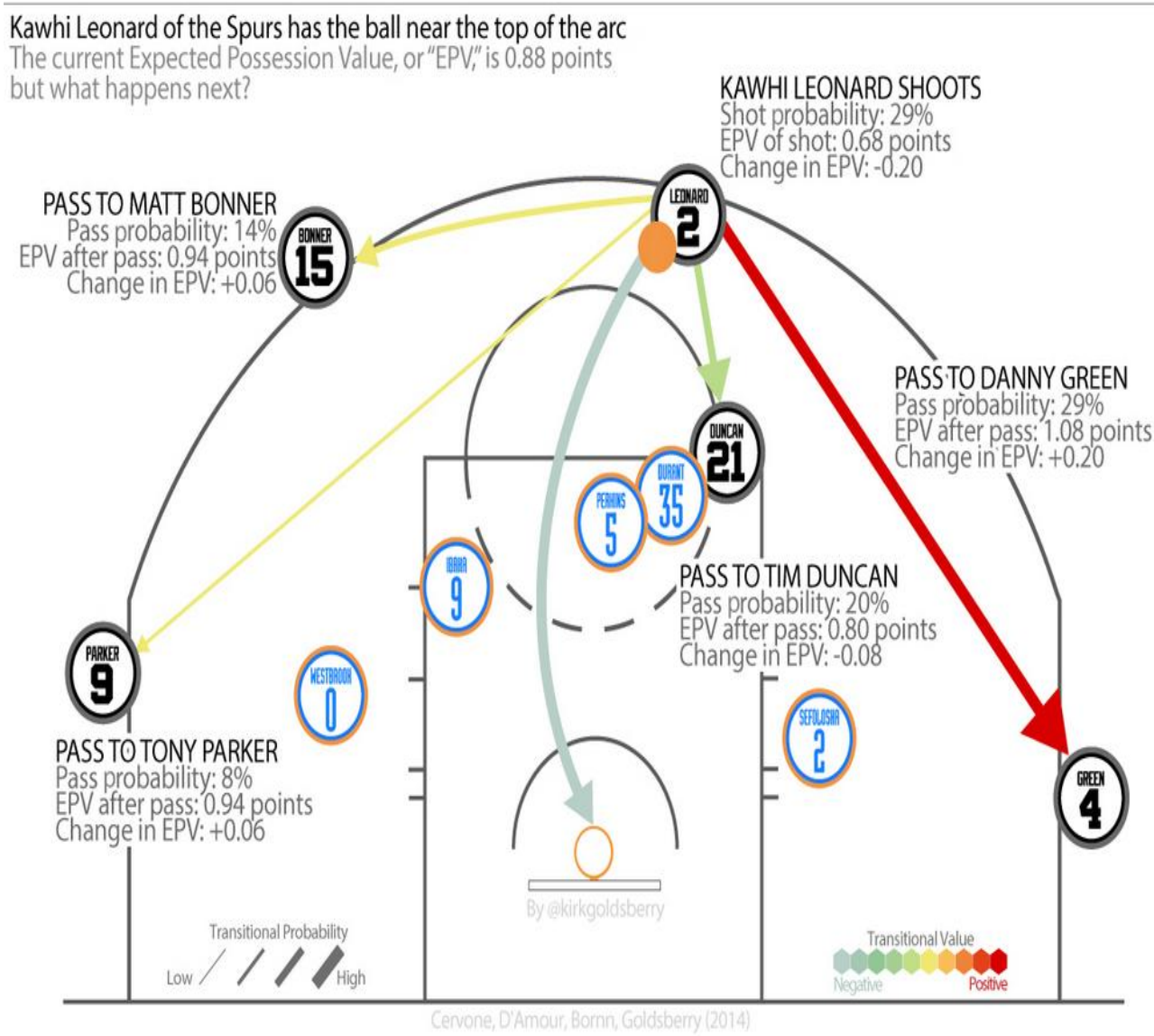
task5

- 淘宝数据平台显示，购买最多的文胸尺码为B罩杯。B罩杯占比达41.45%，其中又以75B的销量最好。
- 其次是A罩杯，购买占比达25.26%.
- C罩杯只有8.96%。
- 在文胸颜色中，黑色最为畅销。
- 以省市排名，胸部最大的是新疆妹子。

task6

task7

- 93 GB.
- 500 parallel processors
- 2TB memory

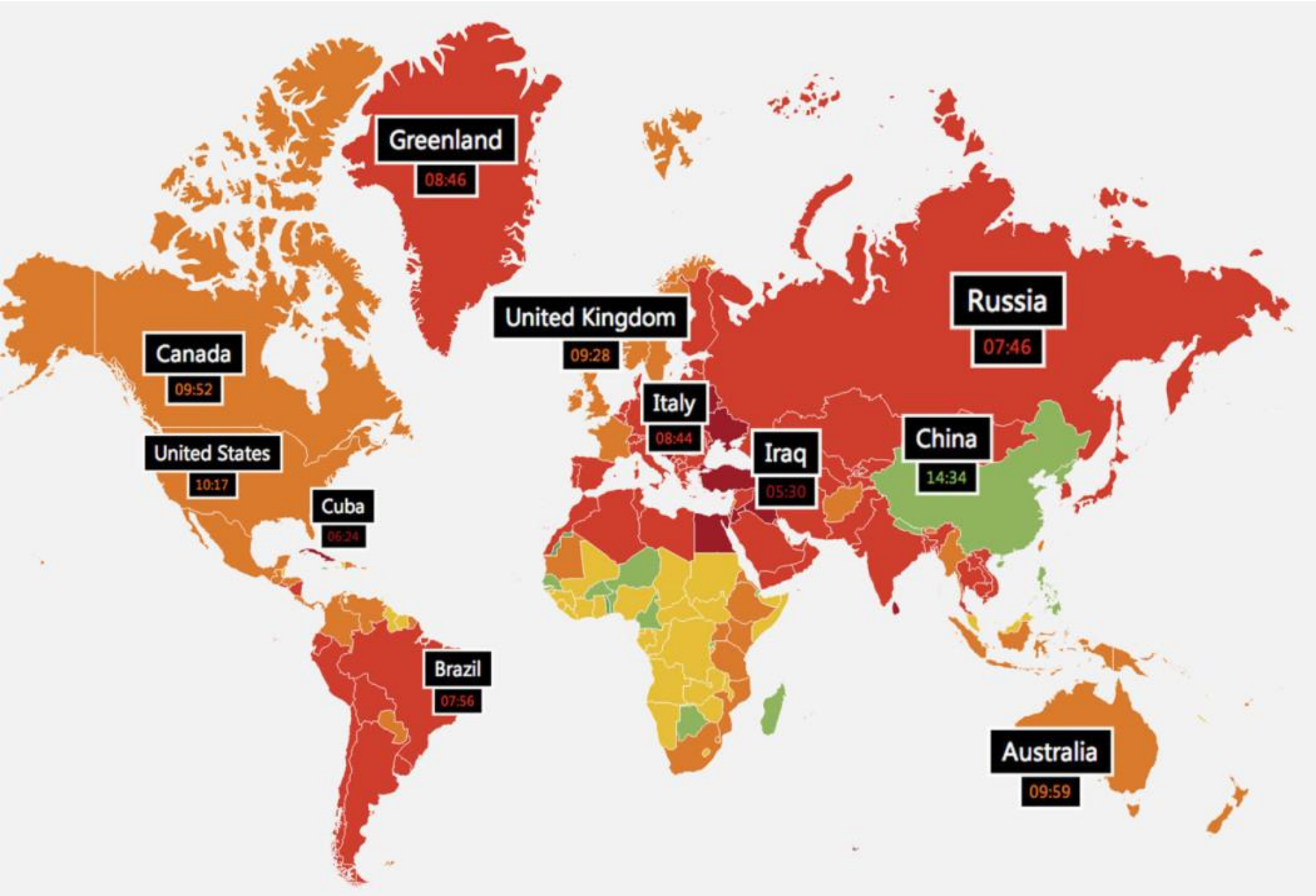


task8

- 美国印第安纳大学的约翰·博伦通过跟踪Twitter社交网站上股民的发言情绪，可以对3天后道琼斯工业平均指数进行预测，预测精度高达86.7%

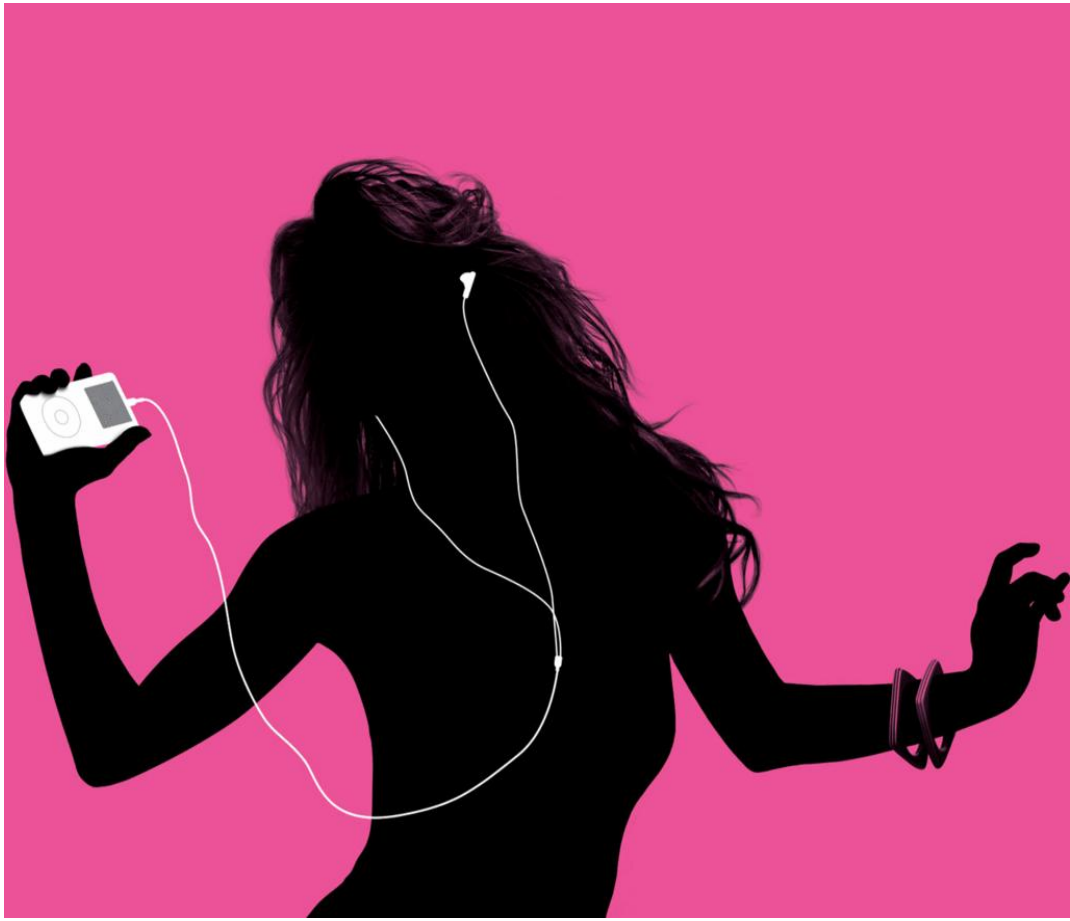
task9

- 世界最大的色情影片分享网站“Pornhub”根据他们访客的资料，统计全球各国和地区网友观看色情影片的时常。大部分的国家或地区观看色情影片的时间落在7分钟～10分钟之间，非洲国家的时间偏高，大多超过11分钟。统计数据显示大陆人平均观看色情影片时间最长，平均为14分钟左右。


Greenland
08:46
Canada
09:52
United States
10:17
Cuba
06:24
Brazil
07:56
United Kingdom
09:28
Italy
08:44
Iraq
05:30
Russia
07:46
China
14:34
Australia
09:59

task10

- 美国某疾病治疗中心跟踪观察病人的心跳等，来给病人推荐适当风格的音乐，来辅助治疗疾病



task11

首页 > 科技 > IT业界 > 正文

谷歌可预测电影票房 准确率达94%

摘自：虎嗅网 | 2013-06-07 13:31:40 | 我来说两句

下载新闻客户端

（原文来自BusinessInsider，虎嗅编译） 在一份名为"用谷歌搜索量化电影魔术（Quantifying Movie Magic with Google Search）"的研究报告当中，谷歌写到，从一部电影上映之前的月搜索量能够预测首映周末的票房。

[查看全文]

task12

耐克锁定大数据营销研发NIKE+新产品

2012-09-14 09:13:15　中国鞋网 www.cnxz.cn　来源：中国鞋网

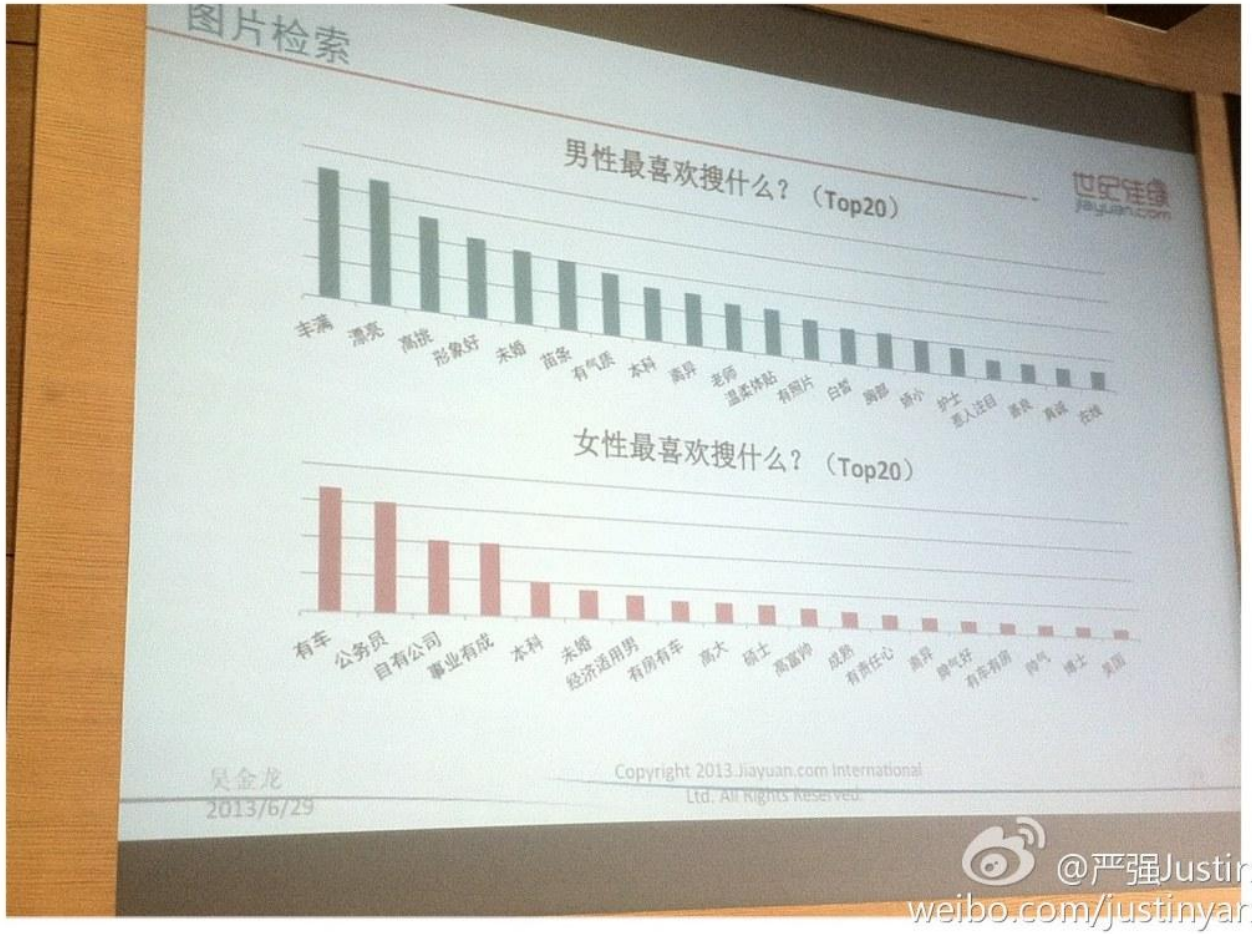
鞋子品牌大全 女鞋品牌大全 男鞋品牌大全 童鞋品牌 女鞋加盟 童鞋加盟

【中国鞋网-品牌动态】大数据时代正在改变、颠覆传统商业规则和人们的消费习惯。

运动鞋品牌耐克最近正在凭借一种名为NIKE+的新产品变身为大数据营销的创新公司。

所谓NIKE+，是一种以“Nike跑鞋或腕带+传感器”的产品，通过无线Nike+iPod运动组件与网络实现信息互通。只要运动者穿着NIKE+的跑鞋运动，iPod就可以存储并显示运动日期，时间、距

task13


图片检索
男性最喜欢搜什么？（Top20）
女性最喜欢搜什么？（Top20）
吴金龙
2013/6/29
Copyright 2013 Jiayuan.com International Ltd. All Rights Reserved.
@严强Justin
weibo.com/justinyan

task14

- 林彪从红军带兵时起，身上就有个小本子，上面记载着每次战斗的缴获、歼敌数量。某次战役后，林彪发现三个问题：

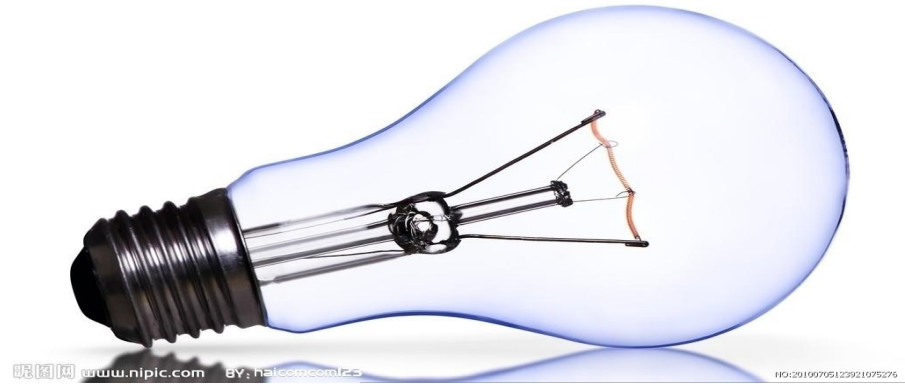
1，“为什么那里缴获的短枪与长枪的比例比其它战斗略高”？

2，“为什么那里缴获和击毁的小车与大车的比例比其它战斗略高”？

3，“为什么在那里俘虏和击毙的军官与士兵的比例比其它战斗略高”？

task15

- Bidgely利用对家用电器具有识别能力的机器学习算法，来分析用户家庭内能源消耗的方式，在网页端和移动端实时显示直观的数据，并给出“省钱妙招”，为用户节电省钱。



task16

- 莱克斯易网盾木马监控系统通过数据挖掘，检测各种木马通信痕迹、识别木马特征和行为。弥补了防火墙、入侵检测系统、防毒墙等在网络层对木马检测的技术空白

task17

- 专家通过数据挖掘方法得出，对城镇就业人员整体而言，最优退休年龄为64.14岁

task18

- 沃尔玛全球移动部门通过数据分析得到：
  ❶现在Walmart.com近1/3的流量来自移动端
  ❷安装了app的用户以更高的频率光临沃尔玛，停留时间比普通顾客高40%。

task19


情侣开房最多
北京
最甜蜜


经济型最多
“省下的钱可以
吃大餐了！”
“oh yeah”
西安
最居家


旅行最多
武汉
最行走


单人房最多
深圳
最寂寞

task20

- 微信公众号相似文章发现

conclusion

- 1，家庭生活
- 2，医疗
- 3，电子商务
- 4，游戏/娱乐
- 5，科学研究
- 6，国民政策制定/总统选举
- ……

Four categories of Big Data

- Fetching 采集/爬虫/抓取
- Storage 数据存储/分布式系统/数据库
- Mining 数据挖掘/机器学习/推荐系统
- Processing Hadoop/Spark/VW/Mahout

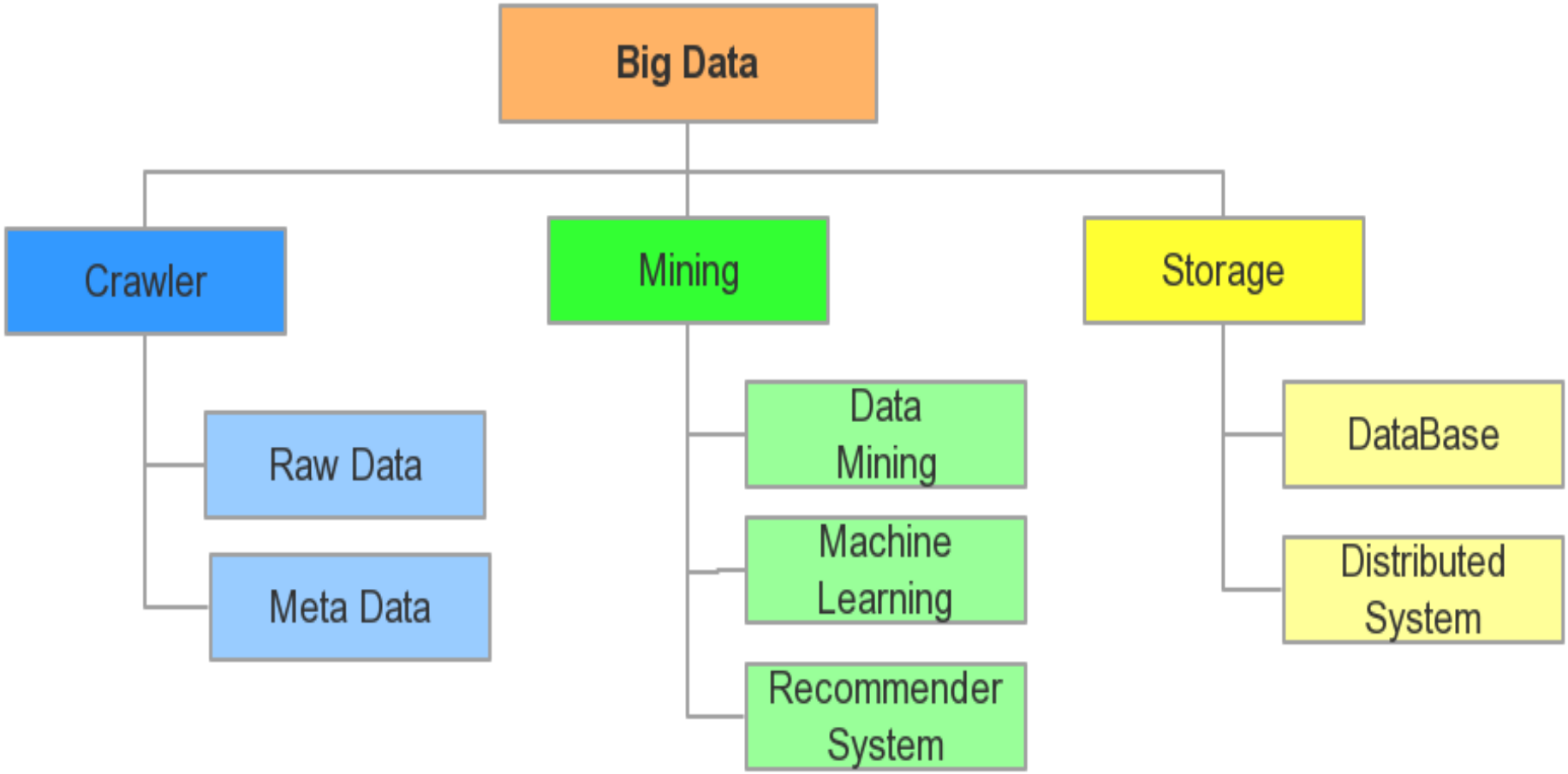
Four categories of Big Data

- Fetching 采集/爬虫/抓取
- Storage 数据存储/分布式系统/数据库
- Mining 数据挖掘/机器学习/推荐系统

因为时间关系，今天我们只介绍前面三类

- Processing Hadoop/Spark/VW/Mahout

Four categories of Big Data



Four categories of Big Data

- Fetching 采集/爬虫/抓取
- Storage 数据存储/分布式系统/数据库
- Mining 数据挖掘/机器学习/推荐系统
- Processing Hadoop/Spark/VW/Mahout

案例&实战

- 任务：一共有1000个查询，每个查询都是大学名称。我们需要定期地周期性地把每个查询扔到google、bing、百度、搜狗、搜搜、有道、360等7大搜索引擎中，将得到的url进行整理。

案例&实战

```
int main()
{
        while(true)
        {
                string url=generateOneUrl();
                string content=getContentViaUrl();
                MetaInfo info=extract(content);
                process & save info();
        }
}
```

案例&实战

优化和改进1：单线程下载，太慢。单线程==》多线程

- while(true)

{

string url=generateOneUrl();

Thread t=new Thread(new Worker(url));

t.start();

}

Worker():

string content=getContentViaUrl(url);

MetaInfo info=extract(content);process & save info();

案例&实战

优化和改进2：创建线程需要开销。多线程==》线程池

- while(true)

```
 {
            string url=generateOneUrl();
             AddworkToThreadPool(worker,url);
 }
```

Worker():

string content=getContentViaUrl(url);

MetaInfo info=extract(content);process & save info();

案例&实战

优化和改进3：消费者和生产者模型

Producer

int main(){while (true){produce url;}} //别忘了，多线程

Worker

int main(){while (true){cosume url;}} //别忘了，多线程

问题来了：怎么通信？怎么交互？

一种简单方法：消费者和生产者同一进程不同线程

案例&实战

两个任务，A和B。既可以多进程，也可以多线程实现。

多线程：

1，便于变量和数据共享。（全局变量 & lock)

2，利用线程池，可以更方便的在主线程里控制。

多进程：

1，稳定和健壮。当A crash了，对B没有影响。

所以，本例中，我们更倾向于多进程

案例&实战

优化和改进4：队列的引入。队列中的每个任务：消息

- Producer将url放入队列
- Worker从队列中取出url

问题：

1，如何防止队列容量爆满内存不够用？

2，如何持久化队列中的消息？

3，如何实现负载均衡？

4，成熟的开源工具？Rabbitmq、zeromq、redis

案例&实战

如何实现getContentViaUrl(url)？

Httpclient & URLConnection (Java)

urllib2(Python)

socket & accept & bind(C in linux)

wget curl (linux os)

1，如何解决被封的问题？代理ip+cookie+友好访问

2，含js/flash的网页如何抓取？Selenium htmlunit

案例&实战

如何实现getMetaInfo(content)？也就是说，如何从网页里抽取出所需的信息？

1，正则表达式

2，Jsoup/Tika

3，模板

Four categories of Big Data

- Fetching 采集/爬虫/抓取
- Storage 数据存储/分布式系统/数据库
- Mining 数据挖掘/机器学习/推荐系统
- Processing Hadoop/Spark/VW/Mahout

Subjects…

- Data Mining，数据挖掘
- Machine Learning，机器学习
- Natural Language Processing 自然语言处理
- Recommender System，推荐系统
- Social Network，社交网络
- Search Engine，搜索引擎

Social Network

- 微博 Twitter Facebook
- 有趣的应用：

  a，社团发现

  b，僵尸粉/水军

  c，情感分析/性别判断/学历……

僵尸粉

判断一个粉丝是不是僵尸粉

特征

- 转发多，原创少
- 关注多，粉丝少
- 内容基本上都是广告
- 头像基本上都是美女图片
- .....

判断一个粉丝是不是黑粉

- If(粉丝数<=10&&关注数>=1000)

  if(原创数<=100&&转发数>=1000)

  if(广告所占比例>=90%)

  ……

- 基于规则的方法，有什么缺点？？？

Why we need machine learning

- Drawbacks of Rule-based (if else if else....)

  Not elegant

  Hard to maintain

  Lack of explanations of probability

Machine Learning

- 预测：分类 & 回归

Machine learning

- 1，特征抽取(feature extraction)

[转发数，粉丝数，关注数….]

- 2，标记数据(labelling training data)

ID	转发数	粉丝数	关注数	y/n
1	10	20	21	0
2	2300	10	3000	1

……

Machine learning

- 3，模型选择(model selection)

Y=a1*转发数+a2*粉丝数+a3*关注数

- 4，模型训练(training)

=>确定参数a1，a2，a3的值

- 5，得到模型

Y=0.0052*转发数+0.019*粉丝数+0.0743*关注数+0.99902

Why we need machine learning

- WHY

人工智能思潮

- 基于规则的专家系统
- 基于数据的机器学习

Data Mining

- 目的: 发现数据的规律和模式
- 应用：疾病诊断

NLP

- 女人如果没有了男人就恐慌了
- 明日逢春好不晦气，来年倒运少有余财
- 此屋安能居住，其人好不悲伤

NLP

@全球震惊创意

冬天：能穿多少穿多少；夏天：能穿多少穿多少。「转」

收起 | 查看大图 | 向左转 | 向右转

6月18日 12:18 来自新浪微博

(49) | 转发(590) | 评论(131)

Recommender System

- Amazon：

“查看此商品的顾客也查看了”

“购买此商品的顾客也购买了”

Recommender System

- 推荐系统严格依赖于场景和业务
- 考虑淘宝推荐衣服 & 美团网推荐酒店

Recommender System

- 喜欢还是不喜欢？喜欢是多喜欢？

Recommender System

- rating
- rating=1代表非常讨厌
- rating=2代表讨厌
- rating=3表示一般般
- rating=4表示喜欢
- rating=5表示非常喜欢

Recommender System

	战狼	速度与激情7	左耳
A：	1	2	5
B:	5	5	4
C:	5	5	1
D:	1	1	5
E：	1	2	4

Recommender System

	战狼	速度与激情7	左耳
A：	1	2	?
B:	5	5	4
C:	5	5	1
D:	1	1	5
E：	1	2	4

User-based

- Rating(D,左耳)=5
- Rating(E,左耳)=4

User-based

- Rating(D,左耳)=5
- Rating(E,左耳)=4
- Prediction(A，左耳)
- =(5+4)/2=4.5

User-based

- Rating(D,左耳)=5
- Rating(E,左耳)=4
- Prediction(A，左耳)
- =(5+4)/2=4.5
- =5*0.5+4*0.5=4.5

Different weights

- Rating(D,左耳)=5
- Rating(E,左耳)=4
- Prediction(A，左耳)
- =5*0.8+4*0.6=6.4

Normalization

- Rating(D,左耳)=5
- Rating(E,左耳)=4
- Prediction(A，左耳)
- =(5*0.8+4*0.6)/(0.8+0.6)=6.4/1.4=4.57

Recommender System

- 2， Content-based

User profile

Item description

Recommender System

- 3， Hybrid approach

Recommender System

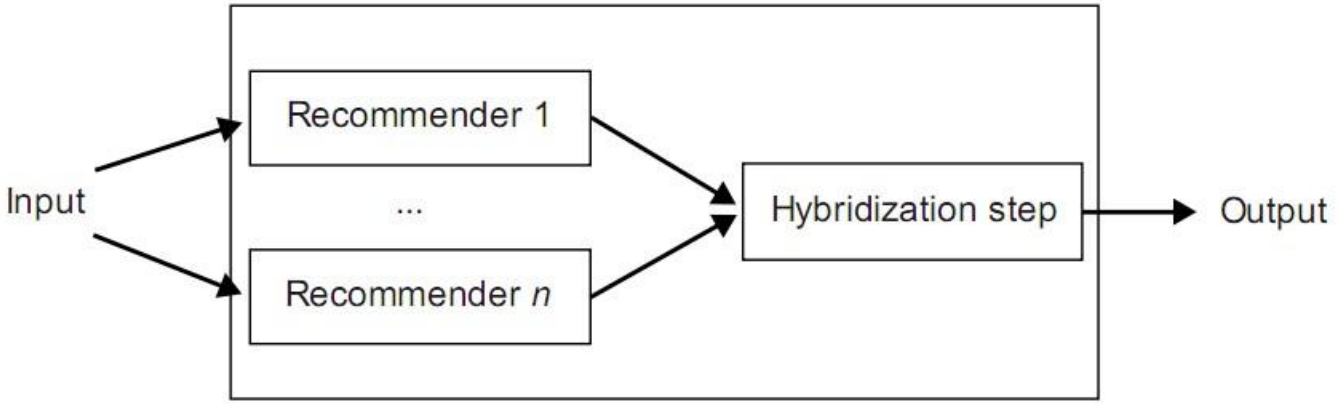


Figure 5.3. Parallelized hybridization design.

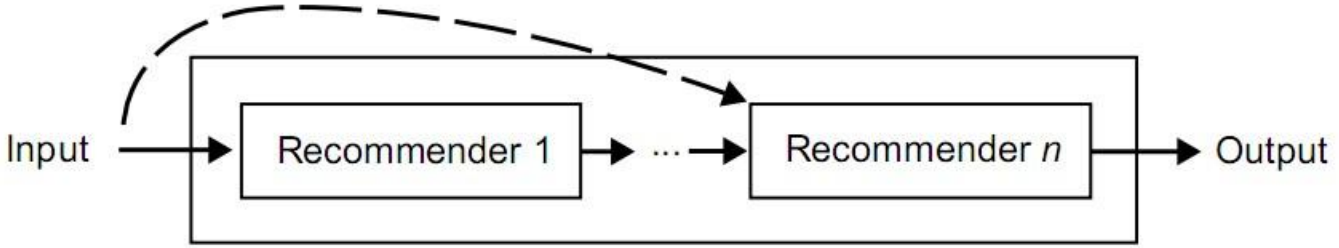


Figure 5.4. Pipelined hybridization design.

Recommender System

- What to do right now and next?
  1，Cold start /Data sparsity problem
  2，Scalability of model
  3，Online-learning
  4，Explanations in recommender systems
  5，Attacks and protections
  ……

反思

- 相关性 PK 因果性

Four categories of Big Data

- Fetching 采集/爬虫/抓取
- Storage 数据存储/分布式系统/数据库
- Mining 数据挖掘/机器学习/推荐系统
- Processing Hadoop/Spark/VW/Mahout

NoSQL

- NoSQL means Not Only SQL
- Why we need Nosql now?
- Some drawbacks in Mysql?

Mysql的缺点

- Schema-based

  不好水平拆分/不大适合互联网环境
- transactions

  很多时候没有必要 /分布式环境难以实现
- Traditional applications

  并发能力较差 （lock）
- Disk-based

  还能更快点吗？

案例&实战

- 场景：在淘宝，买家可以收藏宝贝，将宝贝添加到收藏夹。
- 需求：列出某个买家收藏的所有宝贝的信息（包括最新的人气、价格等）

案例&实战

- 背景：

1.热门宝贝可能被数十万买家收藏

2.每个买家可能收藏上千个宝贝

3.宝贝的价格和人气随时可能变化

案例&实战

- 方案一：
- 两张表：info表和item表

UserId ItemId

1 10

1 45

2 60

ItemId price popularity

10 ￥46 0.98

45 ￥122 0.99

It takes: 4000*5ms=20000ms=20s SO BAD！！！

案例&实战

- 方案二：
- 一张表：info表

UserId	ItemId	price	popularity
1	10	￥46	0.98
1	45	￥122	0.99
2	10	￥46	0.98

It takes: 300000*1000=300000000 =3亿次 SO BAD！！！

NoSQL

- 键值(key-value)存储系统 (eg: redis)
- 无模式文档型存储系统 (eg: mongodb)

Redis

- Key-Value memory based database
- Value：

  String

  List

  Set

  HashMap

案例& 实战

- 给你一堆学生的学号 & 姓名
- 查询：输入学号，快速返回姓名
- 用Mysql怎么做？
- 用Redis: set & get

  set (“110”, ”张帅哥”)

  get (“110”) // “张帅哥”

案例& 实战

- 给你一堆新浪微博账号信息。每个账号的信息类型不全相同
- 查询：快速返回指定的相关信息
- 用Mysql怎么做？
- 用Redis: hset hget

Redis---hashmap

hset(“张帅哥”, //key

{

‘年龄’:’23’,

’电话’: ’1388888888’

‘身高’:’173’

‘年龄’:’25’,

’籍贯’: ’福建福州’

}

Redis---hashmap

hset(“李美女”, //key

{

‘出生年月 ’ : ’1990-09-22’,

’电话’ : ’13999999999’

‘身高’ : ’181’

‘单身’ : ’1’,

’政治面貌’ : ’共产党员’

}

Redis---hashmap

- hget(“张帅哥”,”年龄”) // ”23”
- hget(“李美女”,”单身”) // ”1”

案例&实战

- 求同时选了数据结构和编译原理这两门课的学生
- 用Mysql如何实现？
- 用Redis: sadd sinter

Redis---set

- sadd(“数据结构”，{110,112,119……})

  key　　value

- sadd(“编译原理”，{110,113,140……})

  key　　value

- 共同？

Redis---set

- sadd(“数据结构”，{110,112,119……})

  key　　value

- sadd(“编译原理”，{110,113,140……})

  key　　value

- 共同？
- sinter

Redis

- 效率高，每秒10W次读写
- 数据结构非常丰富
- 支持Java、C++、Python等语言
- disk persistence
- Cluster

But...

- 内存是便宜的，但是不是免费的。
- 内存是很大的，但是不是无限的。
- Redis不（显式）支持复杂的查询

- 我想存很多东西，redis无法全部存下
- 我想无模式
- 试试SSDB？

What should we do?

- 怎么更快处理数据？
- 怎么利用数据？

冷静思考

1，真的需要大数据吗？

很多时候简单的规则和统计即可

2，真的需要很牛逼的算法吗？

很多时候，重要的不是算法牛，而是数据多

研究现状

- 医疗，教育，金融，互联网.....
- 相关牛人：（计算领域）

@明风

@吴甘沙

@Andrew-Xia

@尹绪森

研究现状

- 医疗，教育，金融，互联网.....
- 相关牛人：（存储领域）

@阿里正祥

@阿里日照

@平民架构

@joehan100

@CrazyJvm

@ ideawu

研究现状

- 相关牛人：（挖掘领域）
  @余凯_西二旗民工
  @老师木
  @南大周志华
  @王斌_IIEIR
  @phunter_lau
  @黄萱菁

就业形势

[内推] 算法工程师

待遇水平：	年薪30~50万
公司部门：	爱奇艺 / 系统架构部
所在城市：	北京
详细地址：	北京市海淀区海淀北一街2号鸿城拓展大厦11层
发布信息：	bullud于2013-11-21

机会吸引力：

爱奇艺系统架构部X-Team寻找优秀工程师，协同解决音视频处理、信息检索、模式识别、自然语言处理、机器学习、数据挖掘等技术在实际应用中遇到的各种挑战性问题，研发改变行业和生活的系统级产品。

机会详情：

我们希望将现代信息科技的成果恰到好处的注入到产品中，拥有良好的基本素质和相关专业知识，将有助您这份工作上获得成功。因此，

我们期望优秀的您：

1. 拥有计算机、电子工程、自动化、数学、物理等相关方向硕士或博士学位
2. 已熟练掌握 C/C++ ，Java，Python语言中的任意一种，并熟悉基础性数据结构和相关算法
3. 熟练掌握了机器学习、机器视觉、信号处理、模式识别等任一方面的基础知识（Plus）。
4. 拥有音频识别、音视频处理、人脸识别、文本挖掘、语音识别、场景检测等任一方面的项目经验（Plus）。
5. 具备良好的英语文献阅读能力，学习能力、团队协作能力。

就业形势

[内推] 数据挖掘工程师

待遇水平：	年薪20~40万
公司部门：	百度 / 百度国际化事业部
所在城市：	北京
发布信息：	BigHuge于2013-12-20

机会详情：

工作职责：

1.建设和挖掘结构化数据，用以支撑国际化事业部推荐及其他产品

2.有大规模数据挖掘与机器学习系统经验，能实现数据采集、分析和挖掘，产出对数据分布规律、变化趋势、关联关系的知识

3.结合具体产品，设计合理的策略和算法对用户数据和内容数据进行分析，提升产品效果

职位要求：

1.热爱互联网，对推荐技术、数据挖掘、探索解决问题有浓厚的兴趣

2.良好的逻辑能力，良好的学习能力，有不拘一格的灵活思路

3.扎实的机器学习/数据挖掘理论和技术基础，有2年以上的相关研究或工程经验尤佳

4.精通Java,PHP或者Python等程序设计语言，对数据结构有深刻的理解和掌握

5.了解Map-Reduce, MPI等分布式计算框架，具备相关开发能力的尤佳

6.注重团队协作，有良好的沟通能力

就业形势

GZ老杨同志V：广州鑫亚集团股份有限公司董事长张长德在老杨读书汇分享大数据在实际中的应用

收起 | 查看大图 | 向左转 | 向右转

今天16:40 来自三星android智能手机 | 转发(1) | 收藏 | 评论(1)

How to be a data scientist

- 1，数学 （基础）
  高等数学 线性代数 概率论
- 2，编程语言 （基础）
  首选：Python！ Python！
  其他：C++ /Java /R
- 3，算法和模型 （理论）
  机器学习 数据挖掘 推荐系统 自然语言处理
- 4，开源库 （工具）
  单机：Scikit Learn
  分布式：Spark
- 5，实操 （实践）
  阿里巴巴大数据竞赛 /Kaggle竞赛等

References
数据挖掘（入门）

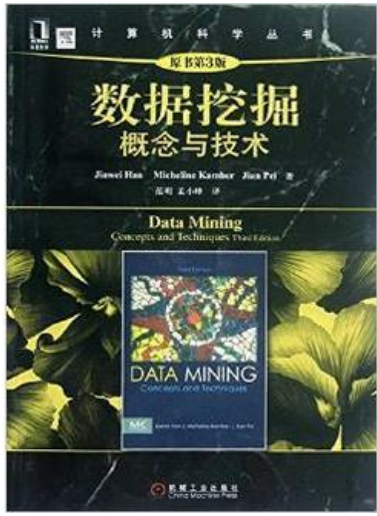
- Jiawei Han

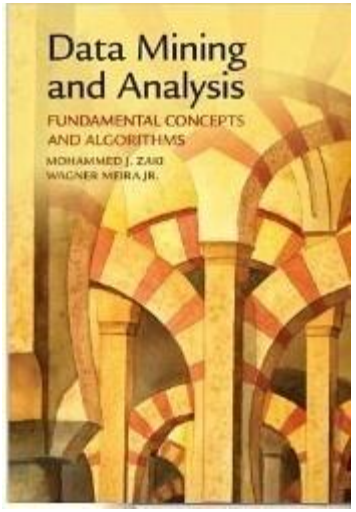
《Data Mining：Concepts and Techniques》

《数据挖掘：概念与技术》，机械工业出版社

- 《Data Mining and Analysis:

Fundamental Concepts and Algorithms》





References
数据挖掘（延伸）

- Anand Rajaraman

  Jure Leskovec

  Jeffrey D. Ullman

  《Mining of Massive Datasets》

《大数据：互联网大规模数据挖掘与分布式处理》

人民邮电出版社

References
机器学习（入门）

- 强烈推荐Andrew Ng的课程

  中文：

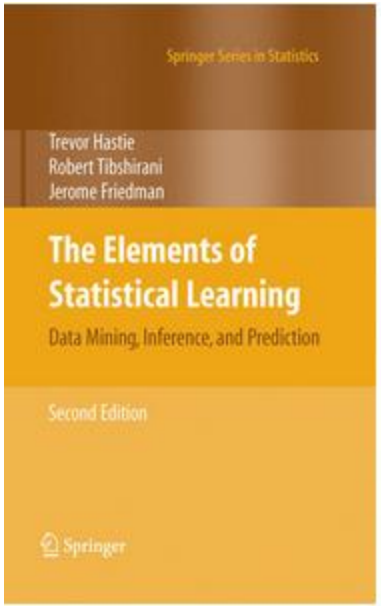
http://v.163.com/special/opencourse/machinelearning.html

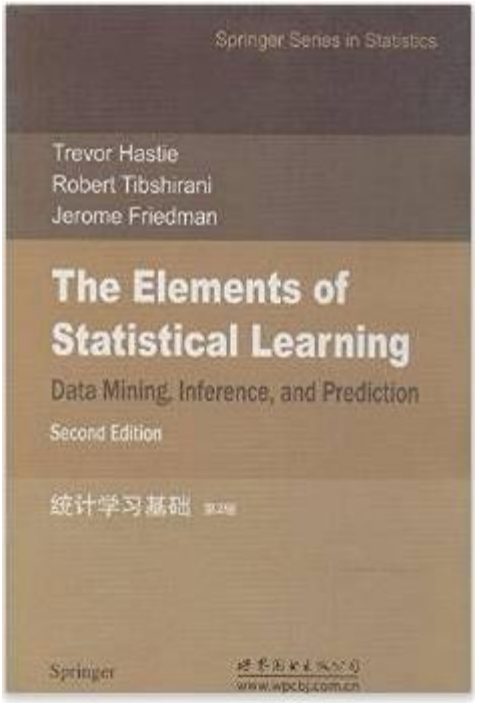
  英文：

http://cs229.stanford.edu/

References
机器学习（延伸）

- ***The Elements of Statistical Learning***
- 统计学习基础





References

推荐系统（入门）

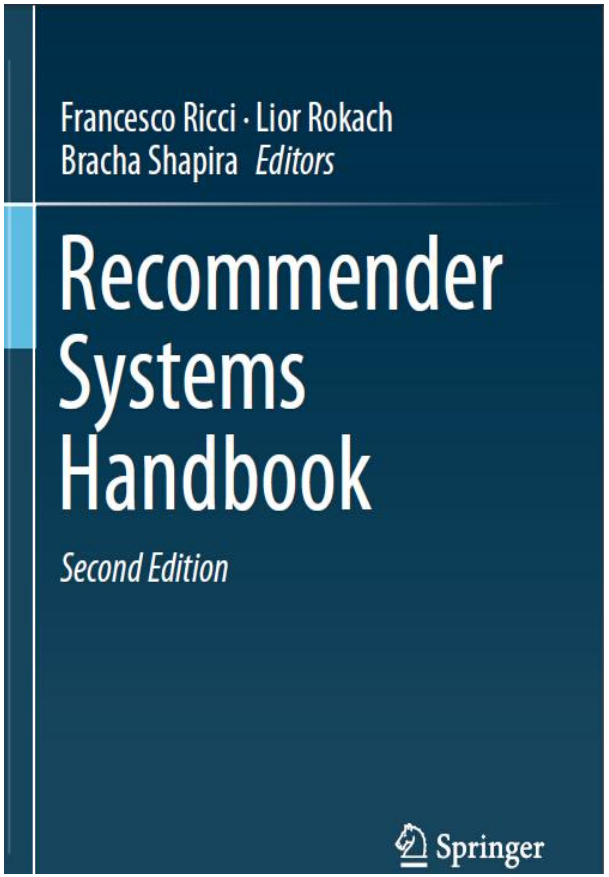
《Recommender Systems: An Introduction》

《推荐系统》，人民邮电出版社

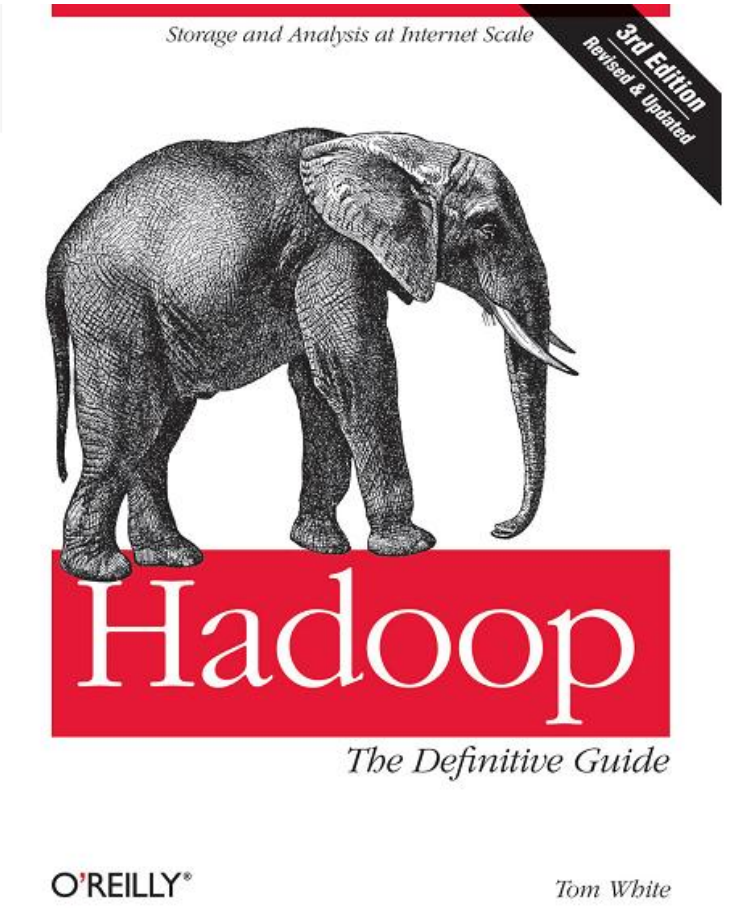
References

推荐系统（延伸）

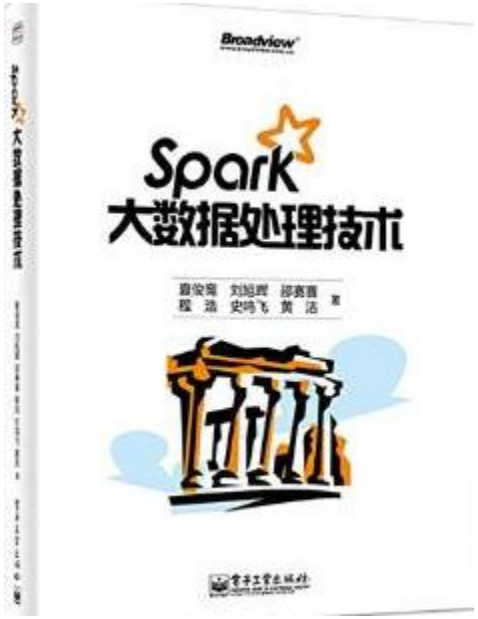
《Recommender Systems Handbook》



References
Hadoop



References
Spark



References
Scikit Learn

休闲读物


BIG
DATA
A REVOLUTION
THAT WILL TRANSFORM HOW
WE LIVE, WORK, AND THINK
大数据时代
生活、工作与思维的大变革

其他

- 以下几张ppt，专门送给对计算机本科生

计算机专业学生的出路

- I 技术类
- II 非技术类

计算机专业学生的出路

- I 技术类
- 程序猿/软件工程师/码农/算法工程师/研发工程师

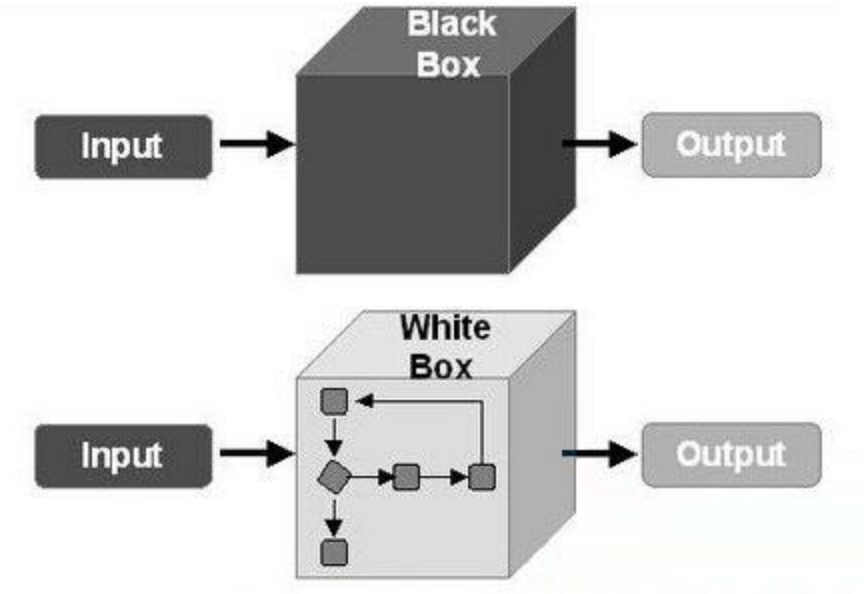
- 前端/Android & iOS/服务器/数据库/数据挖掘/机器学习/推荐系统/

计算机专业学生的出路

- I 技术类
- 产品经理/产品狗/PM/Product Manager/

计算机专业学生的出路

- I 技术类
- 测试



计算机专业学生的出路

- II 非技术类
- 考研（本专业 || 跨专业）
- 老师
- 公务员
- 事业单位
- 转行
- 打字员
- 赢娶白富美/嫁给高富帅，走向人生巅峰

成为技术帝？？？

- I 认真学习，每门课100分。奖学金拿到手软。
- II 编程语言（高级语言+辅助语言）

  C/C++/Java　　Python/Perl/Ruby/R
- III 算法

  ACM竞赛　OJ刷题
- IV 项目
- V 大数据
- VI 准备考研
- VII 比赛

  数学建模 蓝桥杯 中科杯

Q&A

- Thanks A Lot
- Any questions?