



中山大學
SUN YAT-SEN UNIVERSITY



2012级 《多元统计分析与数据挖掘》

第1周

2015.3.5

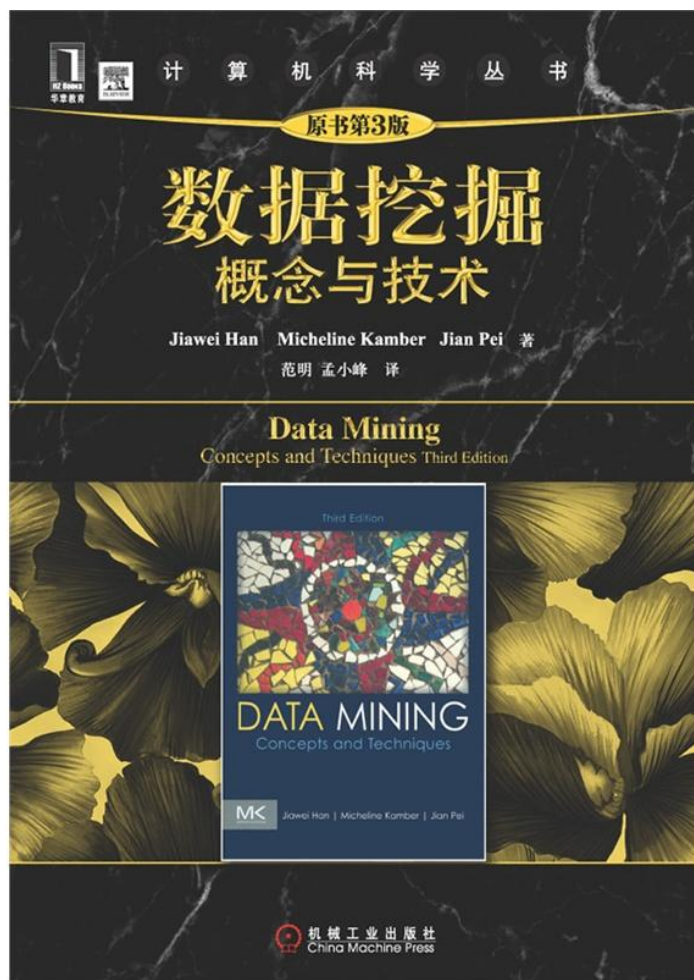
课程介绍



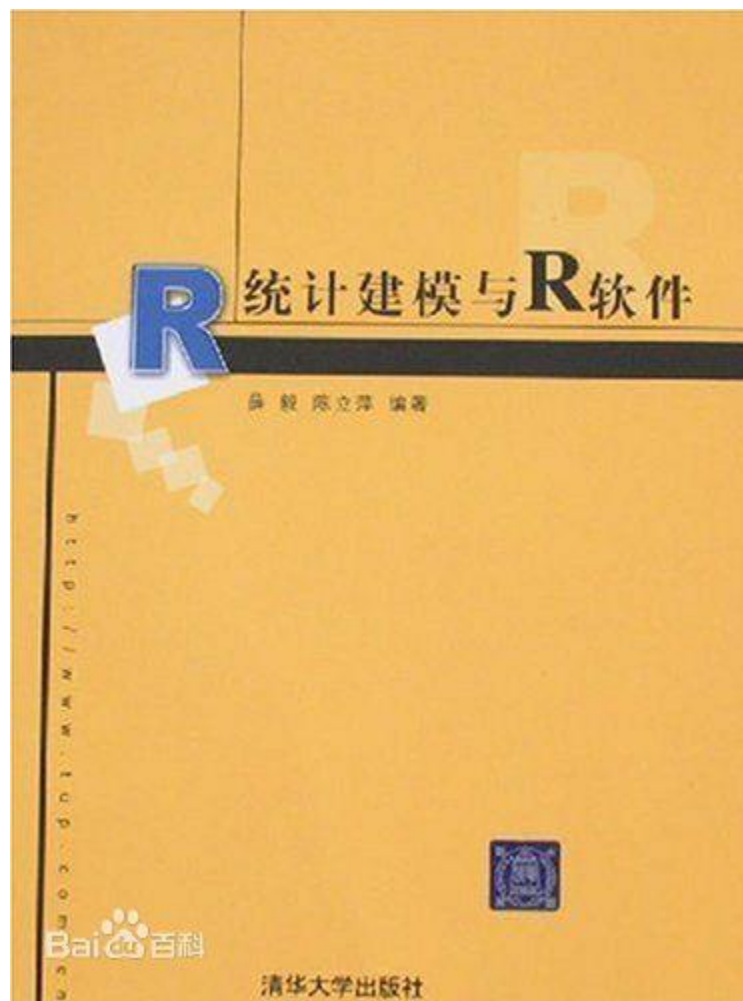
中山大學
SUN YAT-SEN UNIVERSITY

- 上课时间，地点：周二，周四上午1-3节，1308。上课12周，至5月底结束
- 课程需要：计算机，上网环境，英语阅读能力
- 电子作业：书面，互动，大作业。作业，比赛、互动优异者可以加分
- 课程资源和作业平台（加入口令为“sysu@2015”）：
http://www.dataguru.cn/myclass.php?mod=new_basicforlesson&op=basic&lessonid=345
- 考试：3-3.5-3.5，期中考与期末考
- 老师的联系方式：手机13802502960，[邮件stswzh@sysu.edu.cn](mailto:stswzh@sysu.edu.cn)，QQ1829118
- 课程交流qq群：414907025。加入时请注明自己的学号，专业，姓名以便审核

- 1 数据挖掘与R语言概述
- 2 使用R进行回归分析
- 3 使用R进行数据归约：主成分分析和因子分析
- 4 MINE算法，在大数据集中提取规律
- 5 高级回归话题，LASSO，岭回归等
- 6 plyr包与多维数据处理
- 7 R的统计图数据展现
- 8 关联规则挖掘与Apriori算法
- 9 贝叶斯分类器：朴素贝叶斯与贝叶斯信念网络
- 10 决策树
- 11 提升分类器准确率，随机森林，adaboost
- 12 神经网络
- 13 支持向量机，更多的分类器
- 14 聚类
- 15 时间序列
- 16 数据挖掘的扩展话题，序列挖掘，图挖掘



2015.3.5



2015.3.5

数据仓库，数据分析，数据挖掘



中山大學
SUN YAT-SEN UNIVERSITY

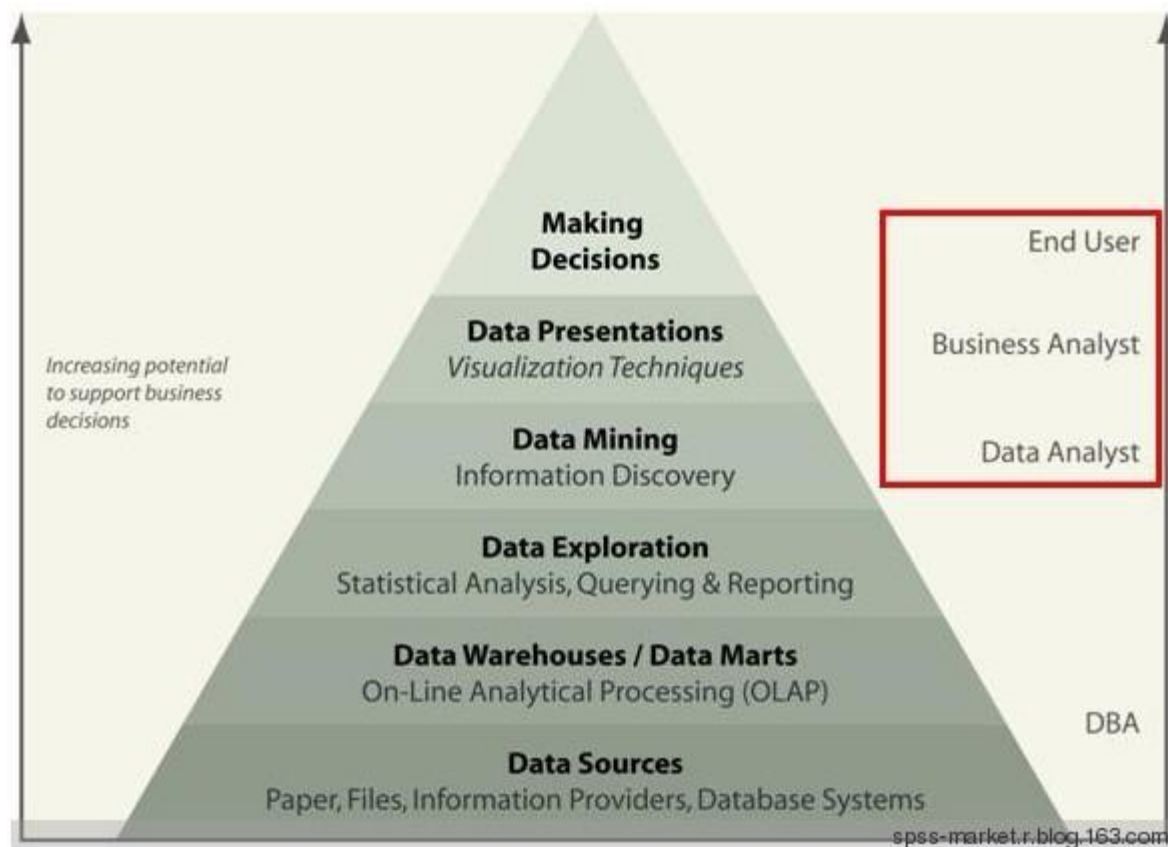


- 数据仓库与数据集市
- ETL是什么？
- OLAP，ROLAP，MOLAP，HOLAP
- 数据分析和数据挖掘有什么区别？

多层模型



中山大學
SUN YAT-SEN UNIVERSITY



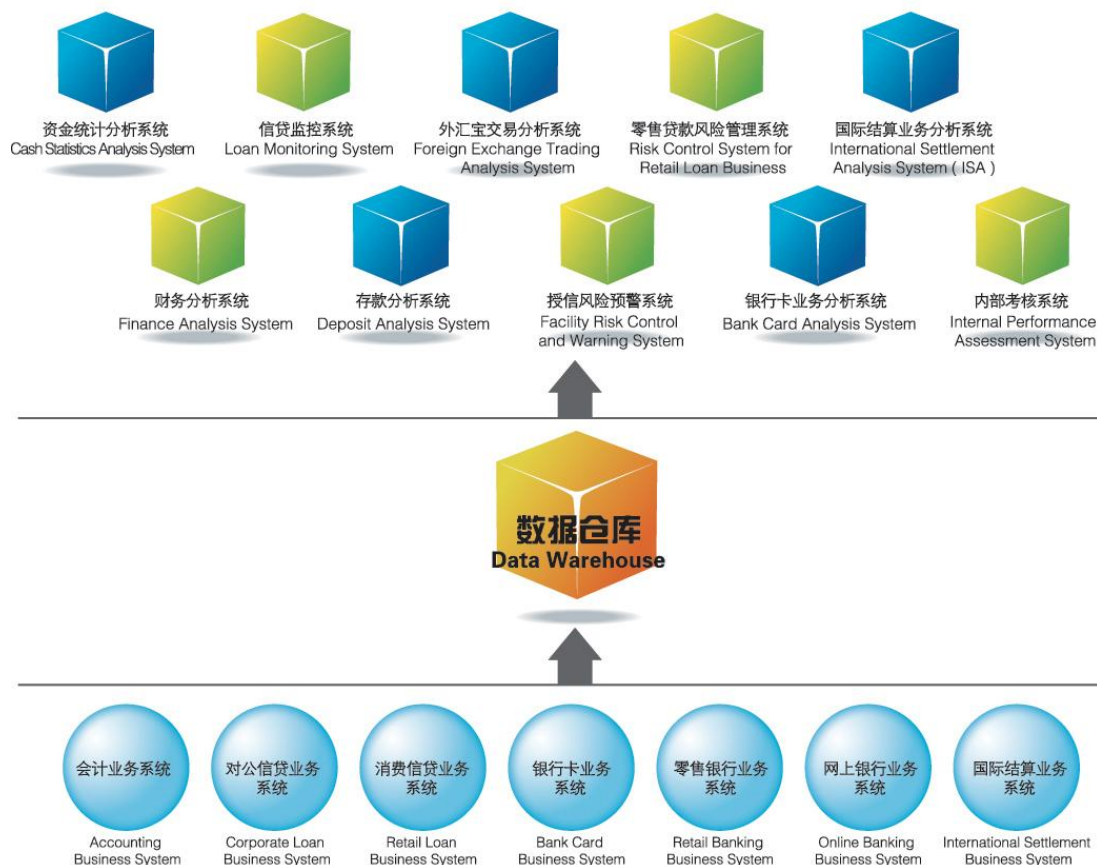
2015.3.5

银行数据仓库



中山大學
SUN YAT-SEN UNIVERSITY

公司 BI产品组图 Business Intelligence Product Grouping

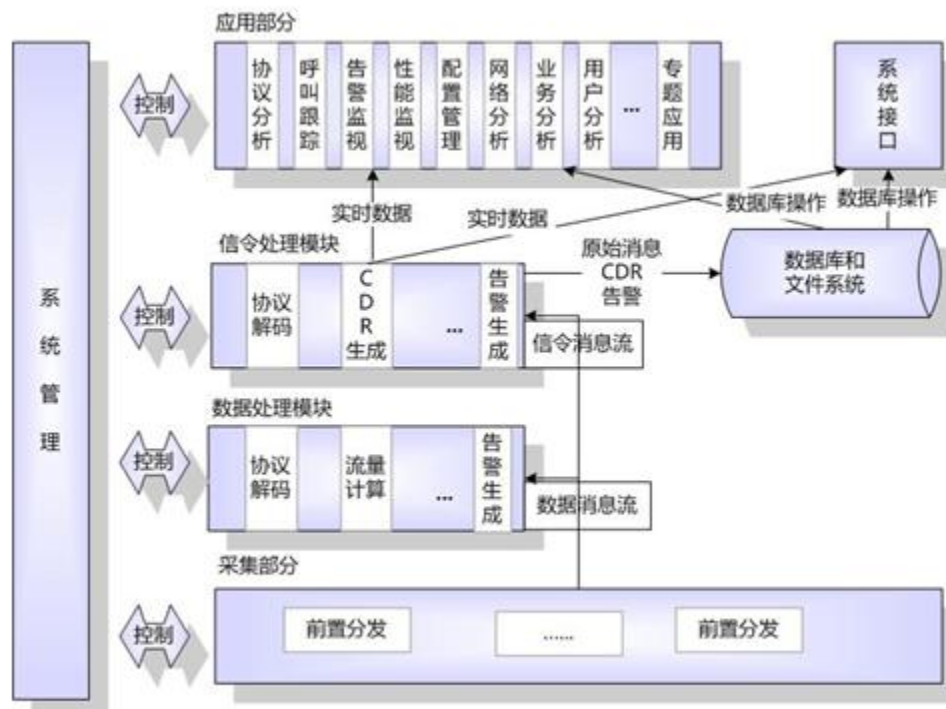


2015.3.5

电信运营商信令分析监测系统



中山大學
SUN YAT-SEN UNIVERSITY



2015.3.5

- Extraction-Transformation-Loading
- 常见的ETL动作：摘取，删除错误数据，合并重复，数据替换，预计算
- 元数据管理，ETL规则规划
- ETL主要的知识基础：业务，IT
- ETL工具：专用的商业产品，通用的商业产品，自写程序

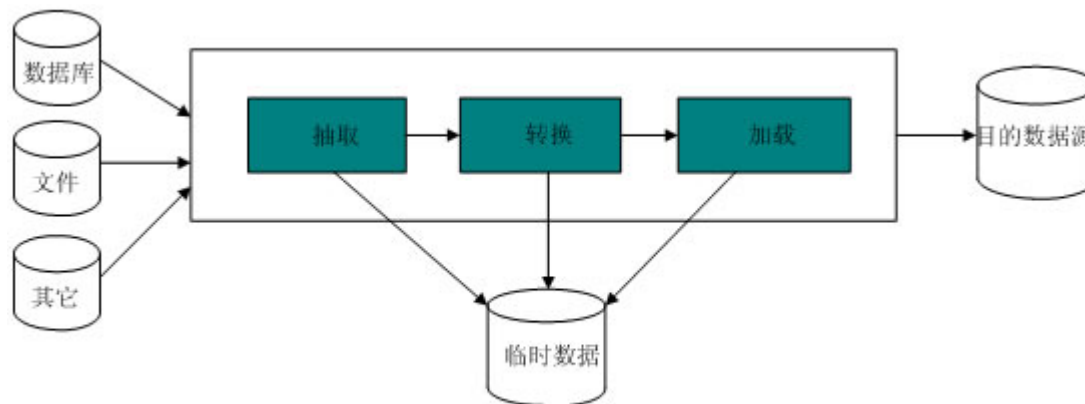
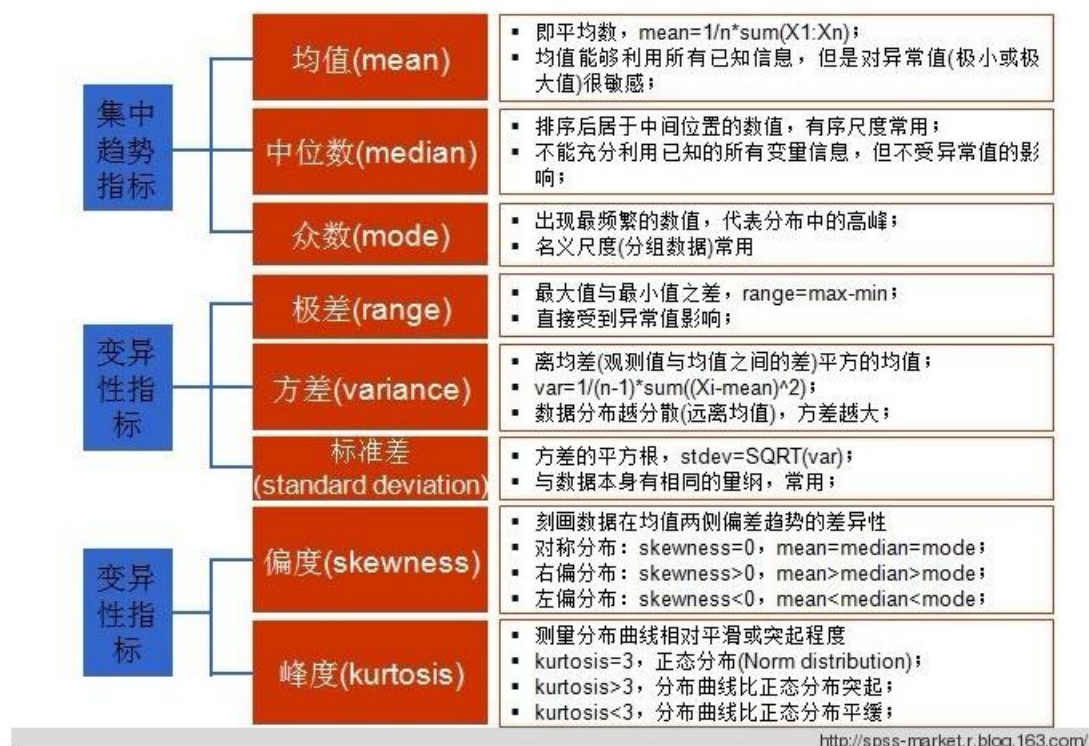


图1.1 ETL 体系结构

- 数据仓库 (Data Warehouse) 是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合，用于支持管理决策。【Inmon , 1991】
- OLAP：联机分析处理，关系型OLAP，多维OLAP，混合OLAP，OLAP的实现
- $DW = ETL + OLAP$
- 数据仓库所需具备的知识：业务+数据建模+IT
- 数据仓库平台及产品
- 数据仓库职位：DBA，数据治理，数据仓库工程师

- 使用统计方法，有目的地对收集到的数据进行分析处理，并且解读分析结果



■ 常用算法



■ 数据分析工具



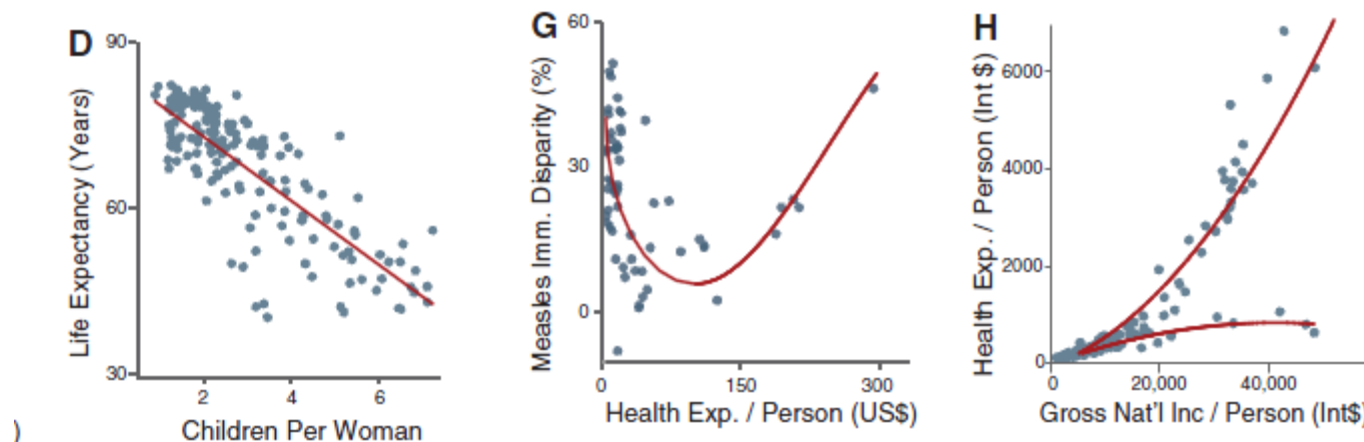
- 数据解读。参考《统计数字会撒谎》达莱尔.哈夫1954年的统计学经典作品
- 实例：毕业生收入统计问题，很多报告会写某某著名大学的毕业生年收入多少，其实样本是偏了的，因为很多混得不好的毕业生就不会回答这类问题，而有些毕业生为了虚荣心又会夸大，如果采用这些样本就会存在问题。
- 实例：房地产公司公布附近居民的平均年收入有10万元，假设有10个居民，9个居民加起来的年收入只有18万，也就是年均收入2万，但是有一个居民是富翁，年收入却有82万，这样一平均下来，平均每个居民的年收入就是10万，显然这个平均数是不能说明大多数居民的收入情况的。
- 实例：抛硬币问题，假如抛5次硬币，那么很有可能4次朝上，1次朝下，很多时候这些数据就会被某些人用来成为某些报告的数据来源。
- 实例：抽烟者的大学成绩比不抽烟者的成绩差，进而推断抽烟使人的头脑变笨，这个谬论的模式是如果B紧跟着A出现，那么A形成了B。

其实跟大的可能是两个因素不互为因果，而同为第三个因素的产物。有可能是性格的原因，例如有些学生喜欢社交，那么他们就喜欢抽烟，时间都放在社交上，学习的时间少了那么自然成绩上不去等。真正的相关往往能够通过相关系数这个令人信服的精确数值来证明事物之间存在关联关系。

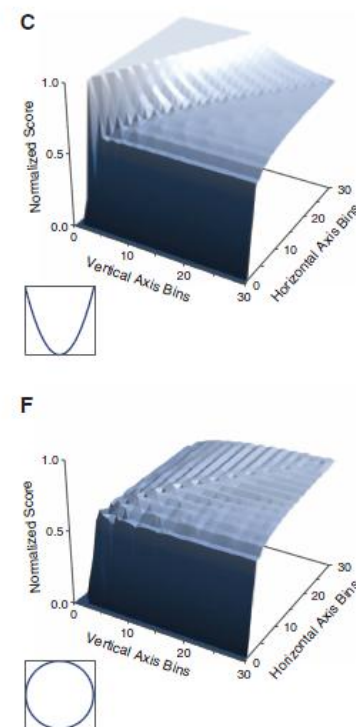
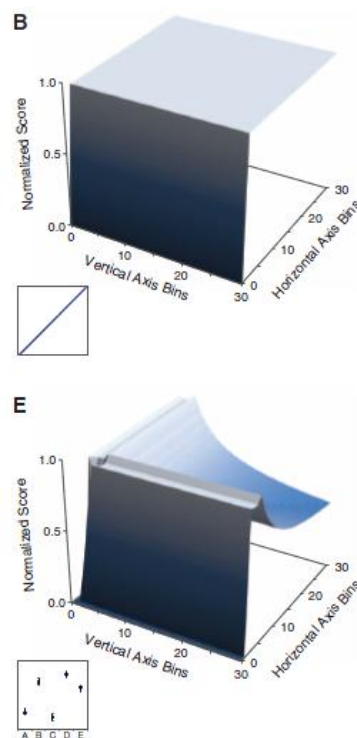
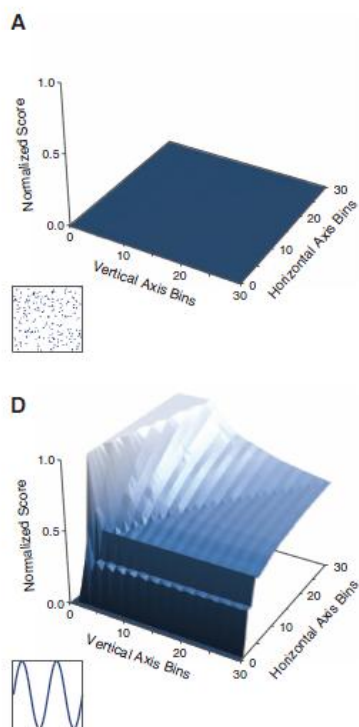
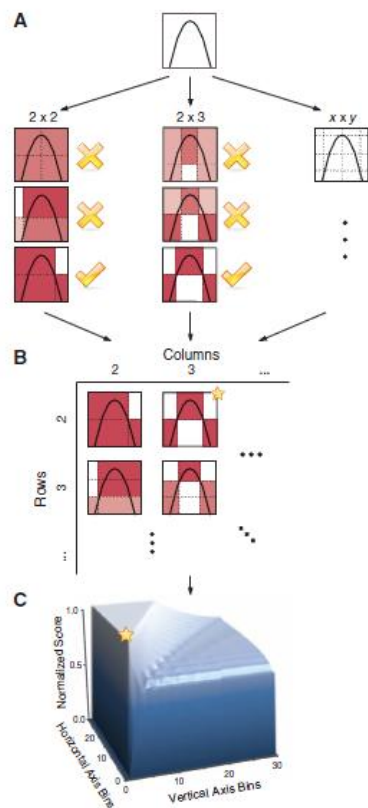


- 数据分析的主要知识技能：业务+数学（统计、数学建模、算法）+IT（分析软件）
- 数据分析职位：数据分析师——金融建模分析师，保险精算师，网站分析师，电子商务分析师，生物信息分析师等等

- 数据挖掘是以查找隐藏在数据中的信息为目标的技术，是应用算法从大型数据库中提取知识的过程，这些算法确定信息项之间的隐性关联，并且向用户显示这些关联
- 数据挖掘思想来源：假设检验，模式识别，人工智能，机器学习
- 常见数据挖掘任务：关联分析，聚类分析，孤立点分析等等
- 例：啤酒与尿布的故事
- 例：《Science》的文章《[科学家摸索出大型数据集内的趋势](#)》



2015.3.5





- 数据挖掘有关知识：业务+数学+IT
- 数据挖掘职位：数据分析师，高级数据分析师



概念澄清

- 数据分析和数据挖掘有什么区别？
- 数据仓库和数据挖掘是什么关系？

什么是机器学习

- 机器学习是指一套工具、方法或程式，从现实世界的海量数据里提炼出有价值的知识，规则和模式
- 把规则应用到前台系统，辅助业务的进行，使其达到更好的效果，例如推荐，辅助决策（沙盘推演，博弈，预测结果），精准辨别，参与服务等，使到业务能产生更大的效益
- 给用户带来“机器具备人类般高智能”的震撼性体验
- 人力成本又越来越高，机器学习能降低企业成本，提高投入产出比
- 第二次机器革命——以具备人类智能为核心价值的机器占主导地位（第一次机器革命——动力系统革命）
- 与人工智能，模式识别，数据挖掘等概念的区别，同一座山峰在不同视角下的侧影，技术内涵几乎一样



- 当当网的图书推荐
- 汽车之家同类汽车推荐
- 淘宝的同类商品推荐
- 新浪的视频推荐
- 百度知道的问题推荐
- 社交推荐
- 职位推荐

推荐系统：当当网



中山大學
SUN YAT-SEN UNIVERSITY

图书 > 风水/占卜 > 运程/风水 > 商品详情

看过本商品的还看了



¥29.00

一次完全读懂运程 (图解民间传统文化百科1000问)

★★★★★ (126条评论)



¥3.50

《好运来》第十期/年 内刊
包含 董易林2014年十

★★★★★ (1条评论)



畅销



分享到: 查看大图

[批量购买入口>>](#)

最佳拍档

明大师2014年马年好运笔记本 (最受欢迎的新年礼物; 十二生肖每月运势; 全年不求人; 幸运号码大揭秘; 随书赠送金箔开运牌一枚)

当当价 **¥19.90** (5折)

定价 ¥39.80

评论 ★★★★★ 96.8%推荐 31条

配送至 广东广州市海珠区, 有货 运费说明 本商品提供礼品包装服务

明天(1月18日)可送达, 请在23小时40分钟内下单并选择“普通快递送货上门”

作者 明大师 编著

出版社

出版时间 2014-1-1

I S B N 23399576

所属分类 图书 > 风水/占卜 > 运程/风水

我要买 件

加入购物车

收藏商品



2015.3.5

、氯等常用消毒药都很敏感。

6、若有发热及**呼吸道**症状，应戴上口罩，尽快就诊，并切记告诉医生发病前有无外游或与禽类接触史。

7、一旦患病，应在医生指导下治疗和用药，多休息、多饮水，注意**个人卫生**。

评论(9)

787


34




sunny闪电雷霆 | 二级 采纳率50%

擅长：暂未定制


其他类似问题

H7N9禽流感有哪些症状 [百度经验]  23 2013-04-09

h7n9禽流感早期症状是什么样的?  217 2013-04-16

H7N9禽流感的症状是什么?  75 2013-04-17

H7N9禽流感症状是什么?  31 2013-04-03

h7n9禽流感什么症状?  14 2013-04-20

[更多关于H7N9的问题>>](#)

问题分类

手机提问 **NEW**

电脑/网络 >

硬件 常见软件 互联网

生活 >

服装/首饰 美容/塑身 购物

医疗健康 >

内科 妇产科 人体常识

体育/运动 >

足球 篮球 健身

电子数码 >

手机/通讯 照相机/摄像机

商业/理财 >

股票 财务税务 创业投资

教育/科学 >

理工学科 外语学习

社会民生 >

法律 求职就业 时事政治

文化/艺术 >

等待您来回答

更多提问 >

我关注的关键词	我关注的分类	为我推荐的问题
春暖花开....性吧		0回答
10 铁观音的茶叶梗子能泡茶喝吗? 对身体好吗?		0回答
穿越火线获得英雄武器黑龙的办法了!!!! [已失效]		0回答
5 给一个可以测定输入的float类型数据小数位数的多少的...		0回答
在常州市老人机哪卖得好?		0回答
跪求小漠国服第一系列泽拉斯三分钟的时候背景音乐		0回答
手拿包什么牌子好呢? 请问		0回答
100 品牌折扣店		0回答
想参加云南14年法检考试, 但基础有些差, 想报个培训班, ...		0回答

贝叶斯分类：判定垃圾邮件



中山大學
SUN YAT-SEN UNIVERSITY

收取 发送 撰写 回复 全部回复 转发 删除 邮件提醒 地址簿 远程管理 中转站

Foxmail

huangzh@139.com

收件箱

反垃圾邮件设置

常规 规则过滤 贝叶斯过滤 黑名单 白名单

在学习邮件前需要整理您的邮件夹，以避免把垃圾邮件作为非垃圾邮件学习或把非垃圾邮件作为垃圾邮件学习。

☐ 使用贝叶斯概率模型判定接收的邮件是否垃圾邮件(U)

已学习信息

非垃圾邮件:	2620	垃圾邮件:	4104
非垃圾词:	1106333	垃圾词:	786541
更新时间:	2014-1-16 下午 11:43:34		

学习(L)... 高级(A)...

过滤强度

移动下面的标记设定过滤的强度。

低 中 高

过滤强度设定越高，邮件被判定为垃圾邮件的可能性越大

☒ 自动删除垃圾邮件箱中以下天数之前的旧邮件

30 天之前

☐ “设定为非垃圾邮件”时不显示提示窗(D)

导入... 导出...

确定 取消

发件人: qingbianji88

主题

来自 qingbianji88 的邮件	2013年12月12日
这儿有件事要说，最近请关注一下	2013年12月12日
自然会议安排	2013年12月11日
全国1800家分店,星级优眠大床房77元即可入...	2013年12月11日
老师，您好	2013年12月7日
韩编辑	2013年12月6日
论文翻译: stswzh@mail.sysu.edu.cn	2013年12月6日
2013 研究生优秀论文展示-Emerald	2013年12月5日
Reference Form Submitted to UBC Graduate Stu...	2013年12月5日
特色专业建设项目研讨会	2013年12月5日
可以 799390	2013年12月4日
三亚旅游国际学术会议邀请函	2013年12月1日
您有1篇论文成果。确认成果，提高工作效率和...	2013年11月29日

收获明显吗？
发，货比三家，价格战满天飞！
“询盘质量低”，“成交价格低”，“客户忠

解决这件难题唯有：主动出击，抢先同行联系客户，实现一对一交流！双喜外贸客户搜索与开发系统帮助您主动式24小时搜遍你们产品行业的上万上游目标客户资源，模拟手工一对一智能群发，24小时让目标客户知道贵司及产品。具有搜索速度快，搜索质量高，信息准确率强，开发信到达率高，投入成本低等特点。让你一天联系100个客户变为一天联系上万个高质量目标潜在客户。询盘订单不断！！避开

2015.3.5

- 分词
- 贝叶斯公式与贝叶斯分类器

若 B_1, B_2, \dots 为一系列互不相容的事件，且

$$\bigcup_{i=1}^{\infty} B_i = \Omega, \quad P(B_i) > 0, i = 1, 2, \dots$$

则对任一事件 A ，有

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{k=1}^{\infty} P(B_k)P(A|B_k)}, \quad i = 1, 2, \dots$$



网页自动分类

- 自动化门户系统（百度新闻，谷歌新闻等）
- 搜索引擎根据用户标签类型推送不同类别的搜索结果

焦点新闻

日本砸22亿元应对钓鱼岛局势 将建专属部队

人活每一天 马英九外甥吓退绑匪 王岳伦死磕到底

张安薇大哥张大公谈救妹过程 特别感谢余靖

- 中国军方高度评价AK-47之父 美媒：人类悲剧 08:11
- 美政界盘点奥巴马政绩：内外交困 被中俄夺主动权 08:33
- 澳媒：中国人即将给亲日的澳领导人一个教训 08:27
- 罗德曼访问朝鲜 金正恩竟为他安排色情服务 12-24 08:17
- 朝鲜第一夫人为张成泽提供性服务被朝鲜证实 12-12 11:30
- 盘点2013中国军队出国十大事件 东海识别区上榜 10:20
- 甲骨文记载：巨人帮助古代中国人大战外星人 12-12 11:41
- 嫦娥之父：美国人去过月球 中国人也一定要去 08:13
- 外媒：中国连射洲际导弹意义重大 令美国不安 08:35
- 安理会通过向南苏丹大规模增派维和部队决议 16:16
- 空军上将：围绕强军目标学习研究毛泽东军事思想 08:59
- 叙利亚称化武储藏点遭反对派袭击 11:21
- 俄媒：直20先用直10发动机 量产型动力舍外国技术 15:57
- 解放军四总部党的群众路线教育实践活动取得成效 15:04
- 共和国“第一号烈士”段德昌：被冤杀的未来元帅 13:10
- 一专多能的女兵台长 05:52
- 航空军工行业：大军工时代的到来 15:32



媒体称中国十天连射两洲际导弹 核打击能力增强？



南苏丹冲突至少8万人流离失所 联合国关切(图)



2013，这些事件峰回路转！

军事评论

- 毛泽东军事思想的伟大建树
- 美驱日制华“鹰犬战略”很危险
- 华报：应重视俄罗斯对中日关系
- “AK-47之父”曾称其枪支发明
- 陈政雄：认识自我核心能力
- 中国罕见海战利器！“潜水战
- 解放军接连展示两大战略神器
- 面对朝鲜变局，韩国有必要紧

图片报道



国际晚班车：《时代》称奥巴马成2013



中英美印四国航空母舰“正脸”照大比拼

2015.3.5

评论自动分析



中山大學
SUN YAT-SEN UNIVERSITY

酒店详情

酒店点评(3027)

立即预订



luya****
2013-12-23

总评: 5.0 卫生: 5 服务: 5 设施: 5 位置: 5

价格便宜 性价比高 交通便捷 靠近市区 服务不错。[详情]

豪华房

有用(0)



luya****
2013-12-23

总评: 5.0 卫生: 5 服务: 5 设施: 5 位置: 5

价格公道 性价比高 交通便捷 酒店餐厅很好吃 服务也很到位。[详情]

高级房

有用(0)



luya****
2013-12-23

总评: 5.0 卫生: 5 服务: 5 设施: 5 位置: 5

五星级酒店而言 价格便宜 性价比高 交通便捷 服务到位。[详情]

豪华房

有用(0)



1100****
2013-12-23

总评: 5.0 卫生: 5 服务: 5 设施: 5 位置: 5

价格合理, 出行方便[详情]

高级房

有用(0)

酒店回复: 2013-12-24

尊敬的顾客您好, 感谢您入住上海明悦大酒店并对我们酒店做出的肯定, 期待您的下次光临!



300720****
2013-12-23

总评: 3.8 卫生: 5 服务: 5 设施: 3 位置: 2

在携程订购的话给的房间都是最小的。别的还行[详情]

高级单人房

有用(0)

来自: 手机用户

酒店回复: 2013-12-24

尊敬的顾客您好, 感谢您入住上海明悦大酒店并对我们酒店做出的肯定, 期待您的下次光临!



109216****
2013-12-23

总评: 5.0 卫生: 5 服务: 5 设施: 5 位置: 5

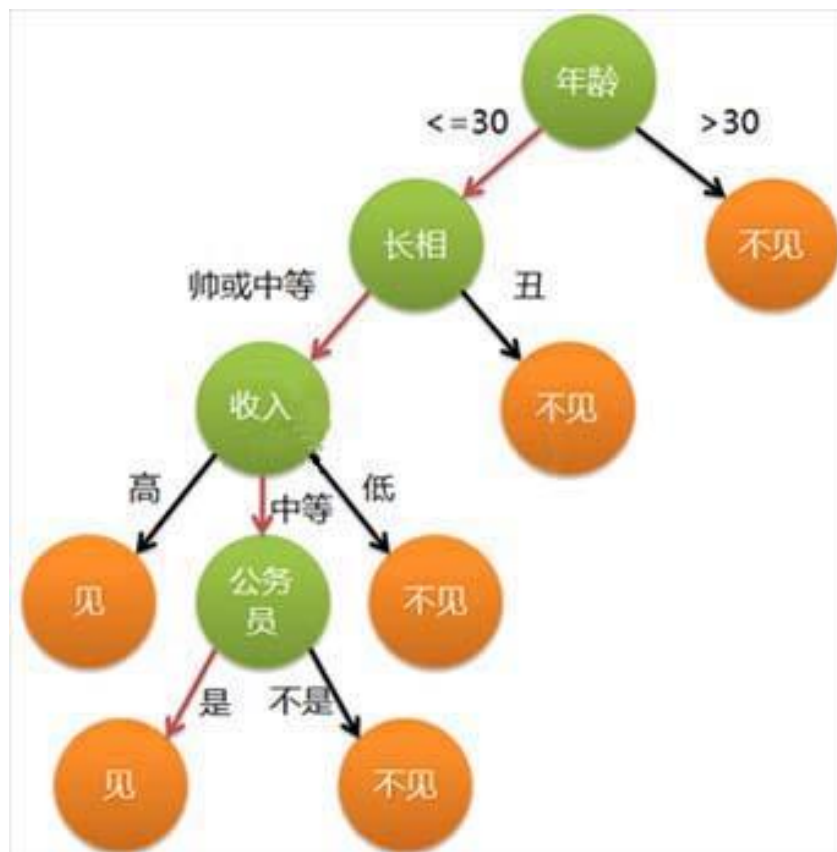
还不错。[详情]

高级单人房

有用(0)

2015.3.5

- 给出样本集，学习后输出的产物是一颗决策树

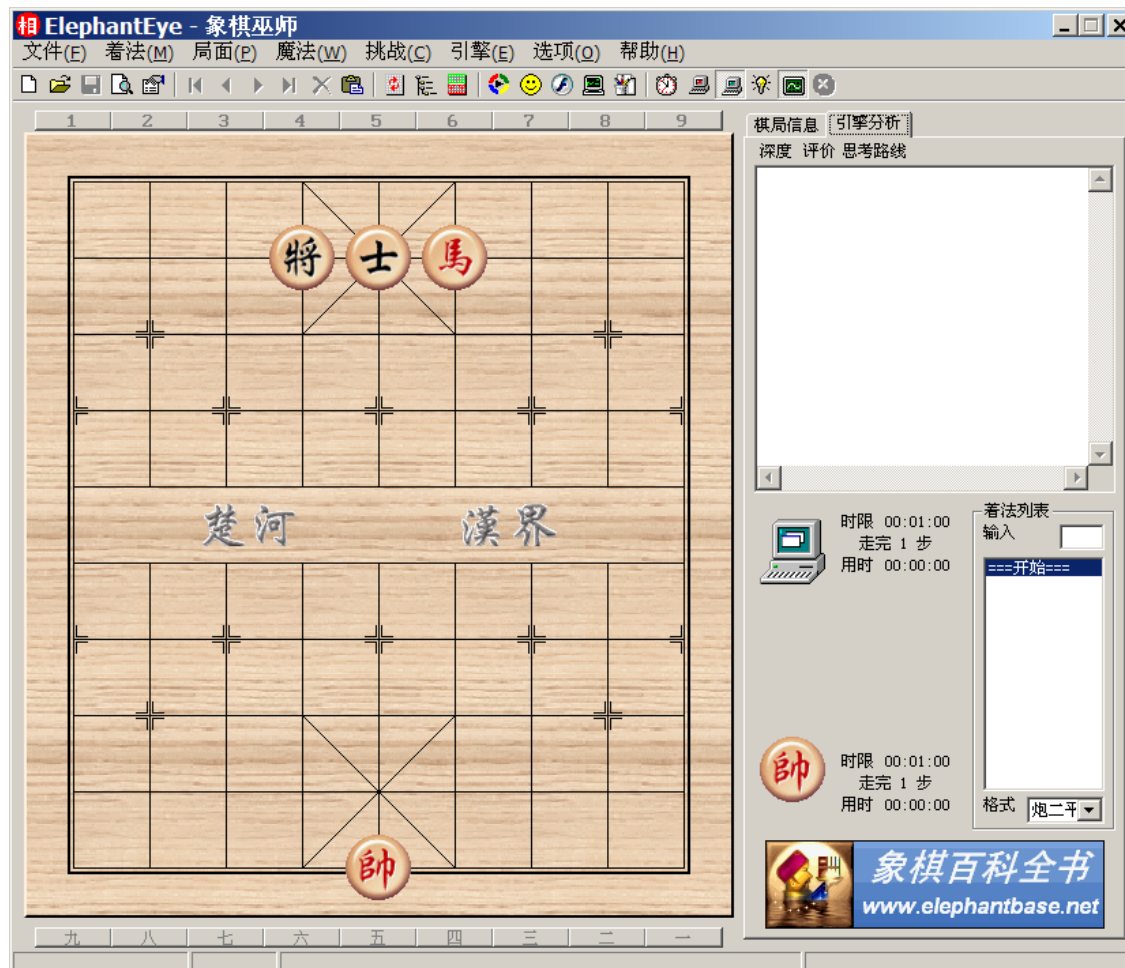


智能博弈：中国象棋云构想



中山大學
SUN YAT-SEN UNIVERSITY

- 局面标准化
- 局面评估函数
- 棋谱学习



2015.3.5



语音识别



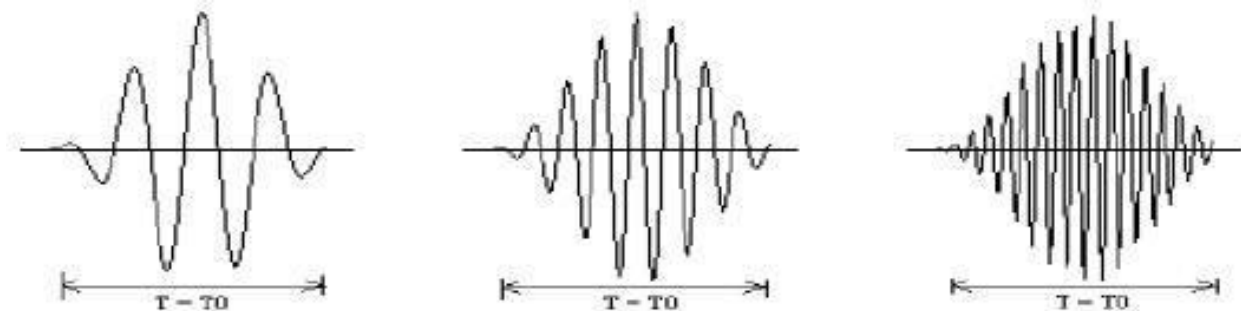
中山大學
SUN YAT-SEN UNIVERSITY

- 语音输入
- 规范化语音：滴滴打车
- 语音属主鉴别

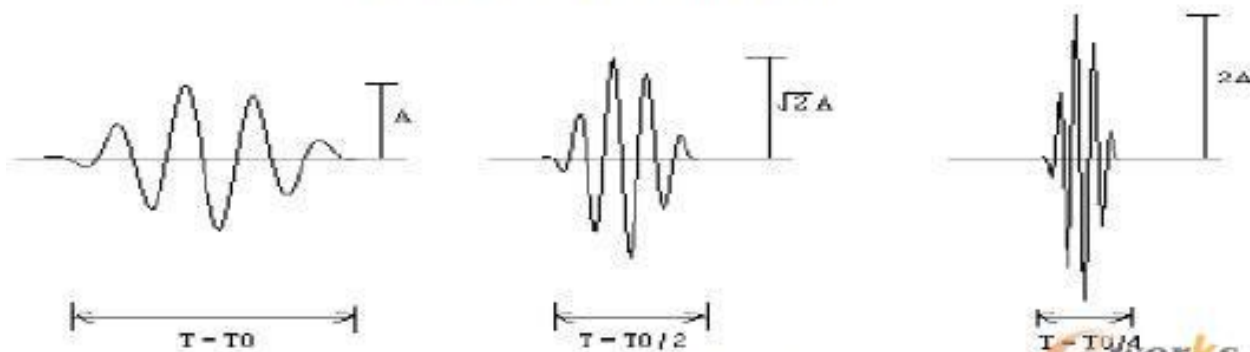


2015.3.5

- 指纹、虹膜纹识别
- 脸像识别
- 车牌识别
- 动态图像识别
- **小波分析**



B: 短时傅里叶变换基函数示意图



C: 小波变换基函数示意图



- R
- Weka
- Matlab
- Python
- 参考：<http://blog.csdn.net/hzxhan/article/details/8548801>

展现层：报表与图形






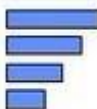







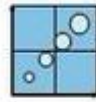


中山大学
SUN YAT-SEN UNIVERSITY

■ 老土的报表

2006年资金预算收支执行情况表																
																单位：万元
月份	收 入								支 出							
	预算情况				实际情况				预算情况				实际情况			
	经营活动	投资活动	筹资活动	合 计	经营活动	投资活动	筹资活动	合 计	经营活动	投资活动	筹资活动	合 计	经营活动	投资活动	筹资活动	合 计
1月份	2100			2100	3610		0.17	3610.17	5476	2082	50	7608	4961	1175	35	6171
2月份	3800			3800	2420		10.2	2430.2	3809	1244	50	5103	2887	108	54	3049
3月份	4274			4274	5474		11	5485	4526	1496	50	6072	4529	6088	30	10647
4月份	12396			12396	11121	68	2097	13286	5586	1514	50	7150	4246	1230	33	5509
5月份	5311	152		5463	5784	98	94	5976	5841	2431	440	8712	4785	792	432	6009
6月份	3801			3801	1217	15	103	1335	4332	2904	87	7323	4067	1903	33	6003
7月份	5951			5951	4427	65	3593	8085	4085	2591	331	7007	5218	2187	332	7737
8月份	5388			5388	1883		2021	3904	3375	3830	2120	9325	3133	3472	2120	8725
9月份	2830			2830	2459	2	914	3375	3955	2905	93	6953	2800	1469	85	4354
10月份	3250			3250	2855		49	2904	4283	2209	40	6534	3526	1591	39	5156
11月份	3870		700	4570	647		134	781	5873	6036	540	12449	810	3861	540	5211
12月份	4105		2150	6255	7723		2576	10299	7631	3551	88	11270	7065	1838	86	8989
合 计	57676	152	2850	60678	53620	248	11602.37	65470.37	58774	32793	3939	95506	48027	25714	3819	77560

2015.3.5

■ 常见的报表

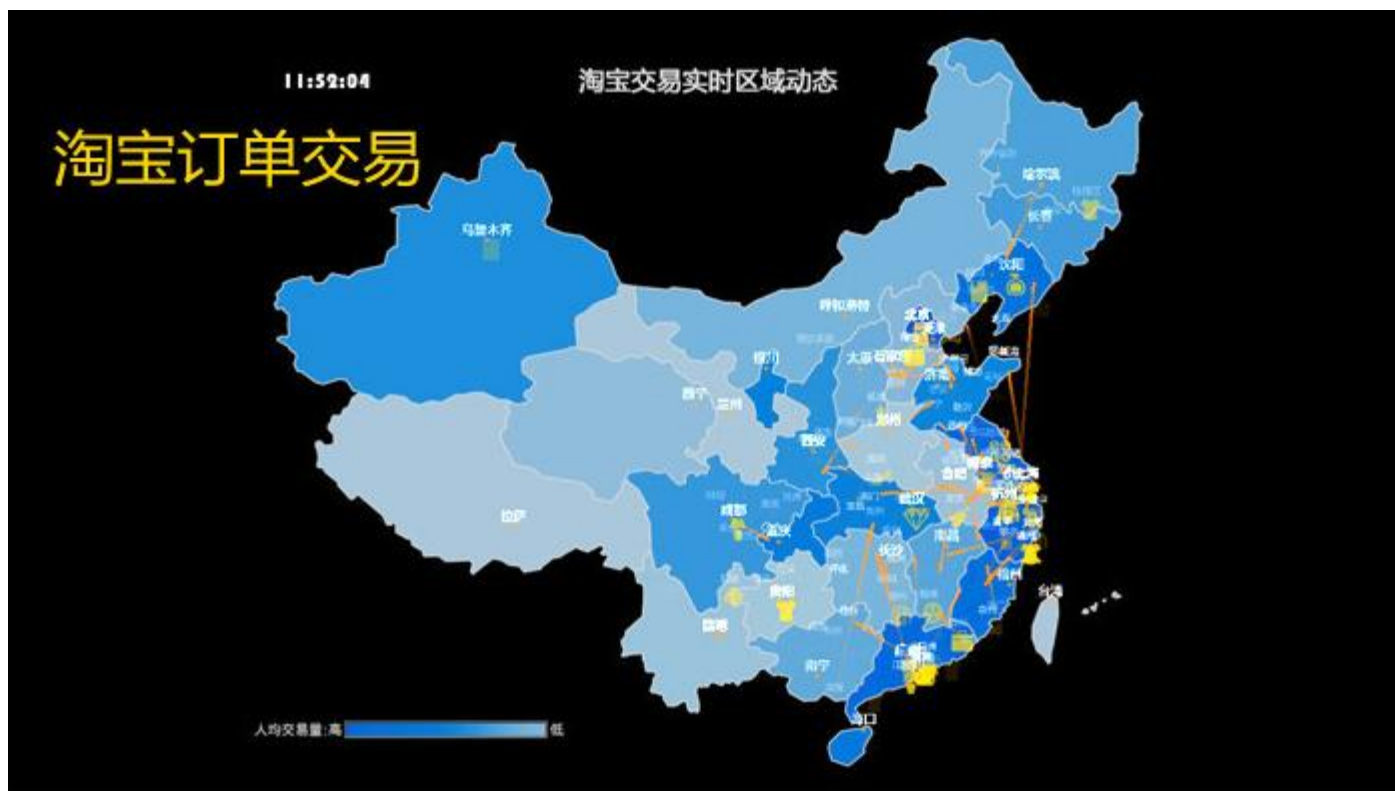
要表达的数据和信息	建议采用图形					
	饼图	垂直柱	水平柱	线图	水泡	其他
整体的一部分						
不同数据的比较						
时间序列						
频率						
两组数据的相关性						
和多重数据、标准相比较						

spss-market.r.blog.163.com

■ 仪表盘

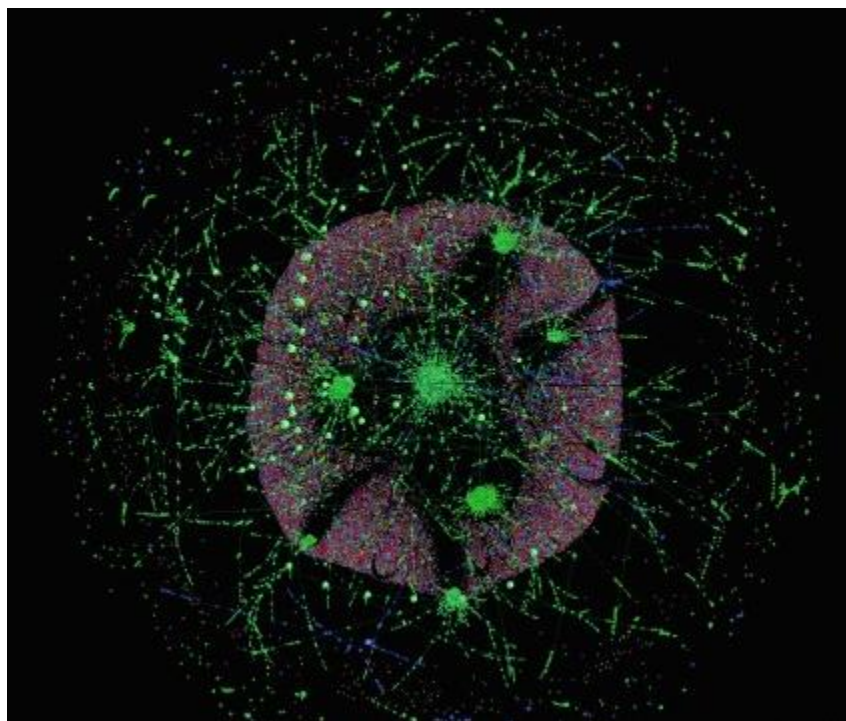


■ 一些有趣的图表



2015.3.5

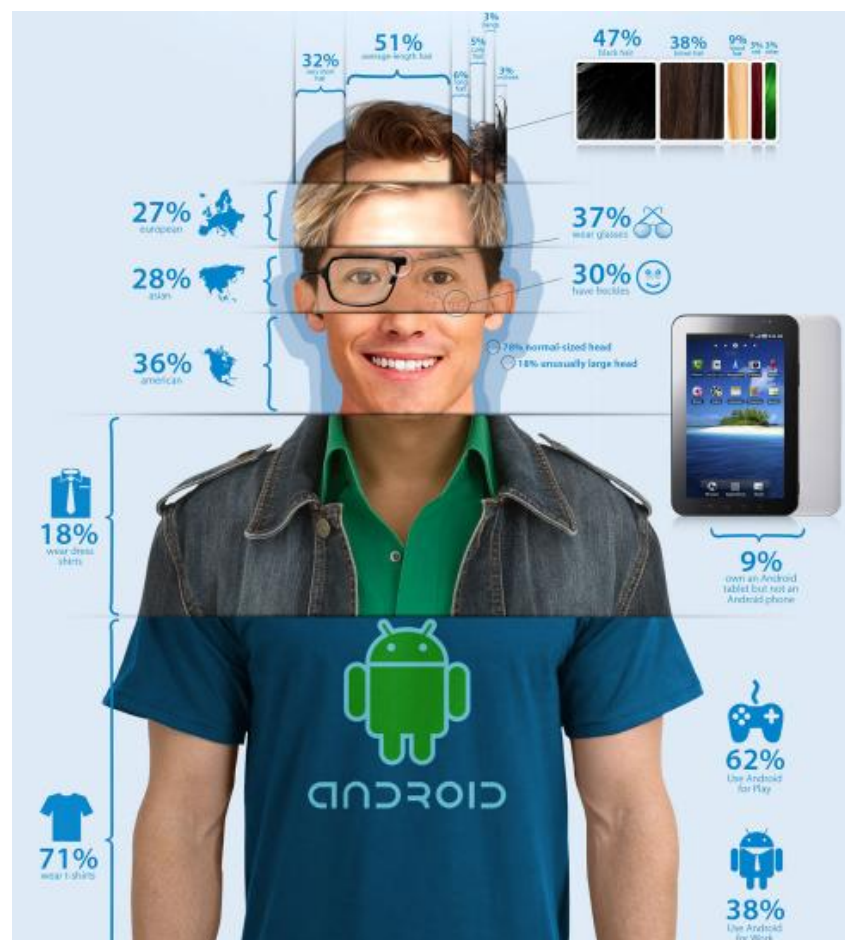
■ 某条微博的扩散路径





■ Mr Android

根据信息图显示，Android先生的头发有47%的可能是黑色的，戴眼镜的几率为37%，有36%的可能是北美人，30%的可能脸上长雀斑。71%的时间会穿T恤，下身穿牛仔裤的时间占了62%。工作只占了38%，玩游戏却占了62%，平均每个月会用掉582MB的数据流量。

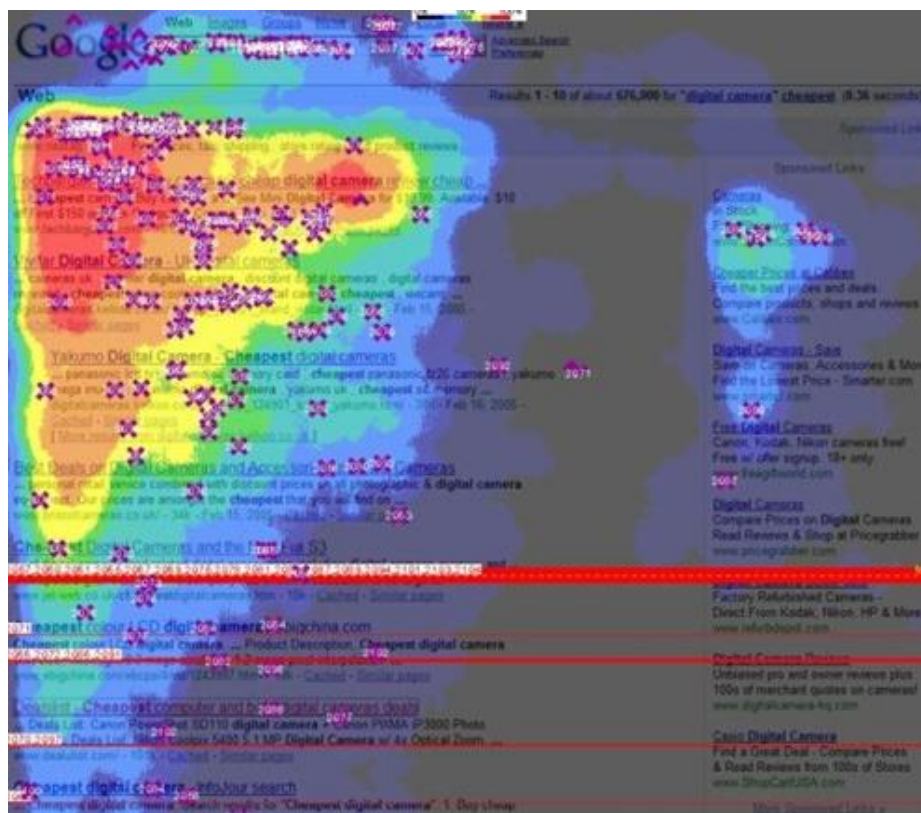


■ Mr Android



2015.3.5

■ 网站点击“热力图”

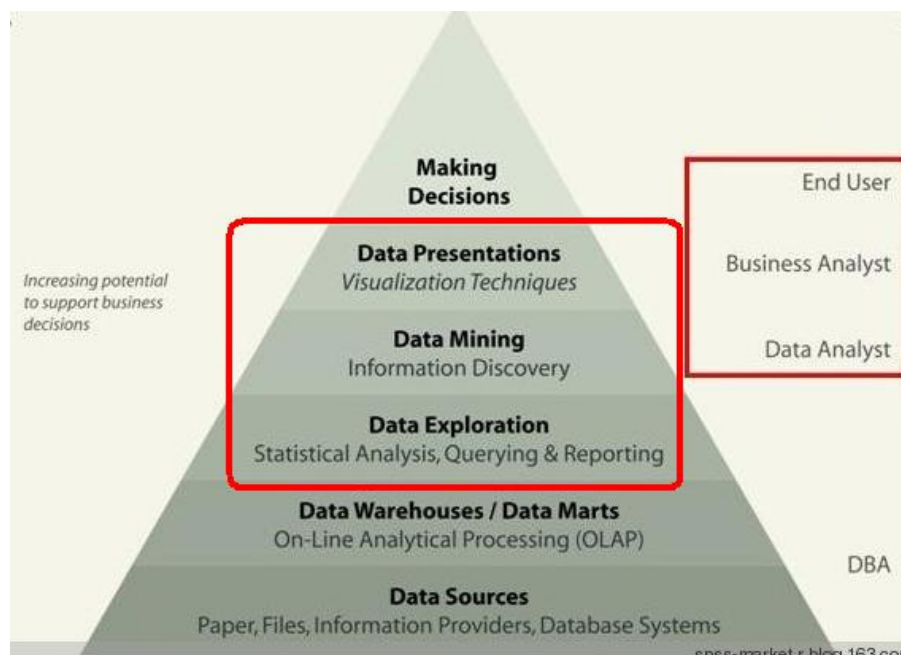


2015.3.5



- 报表展现所需知识：IT（报表工具与绘图工具）+美术素养+心理学
- 相关职位：报表工程师，图表设计师，界面设计师，视觉效果设计师等等

- Business Intelligence , 简写为BI
- BI=数据仓库（存储层）+数据分析和数据挖掘（分析层）+报表（展现层）
- 我们课程的位置



■ R的源起

R是S语言的一种实现。S语言是由 AT&T贝尔实验室开发的一种用来进行数据探索、统计分析、作图的解释型语言。最初S语言的实现版本主要是S-PLUS。S-PLUS是一个商业软件，它基于S语言，并由 MathSoft公司的统计科学部进一步完善。后来Auckland大学的 Robert Gentleman 和 Ross Ihaka 及其他志愿人员开发了一个R系统。R的使用与S-PLUS有很多类似之处，两个软件有一定的兼容性。

■ R is free

R是用于统计分析、绘图的语言和操作环境。R是属于GNU系统的一个自由、免费、源代码开放的软件，它是一个用于统计计算和统计制图的优秀工具。

R是一套完整的数据处理、计算和制图软件系统。其功能包括：数据存储和处理系统；数组运算工具（其向量、矩阵运算方面功能尤其强大）；完整连贯的统计分析工具；优秀的统计制图功能；简便而强大的编程语言：可操纵数据的输入和输出，可实现分支、循环，用户可自定义功能。

R是一个免费的自由软件，它有UNIX、LINUX、MacOS和WINDOWS版本，都是可以免费下载和使用的，在那儿可以下载到R的安装程序、各种外挂程序和文档。在R的安装程序中只包含了8个基础模块，其他外在模块可以通过CRAN获得。

R官方网站地址：<http://www.r-project.org>

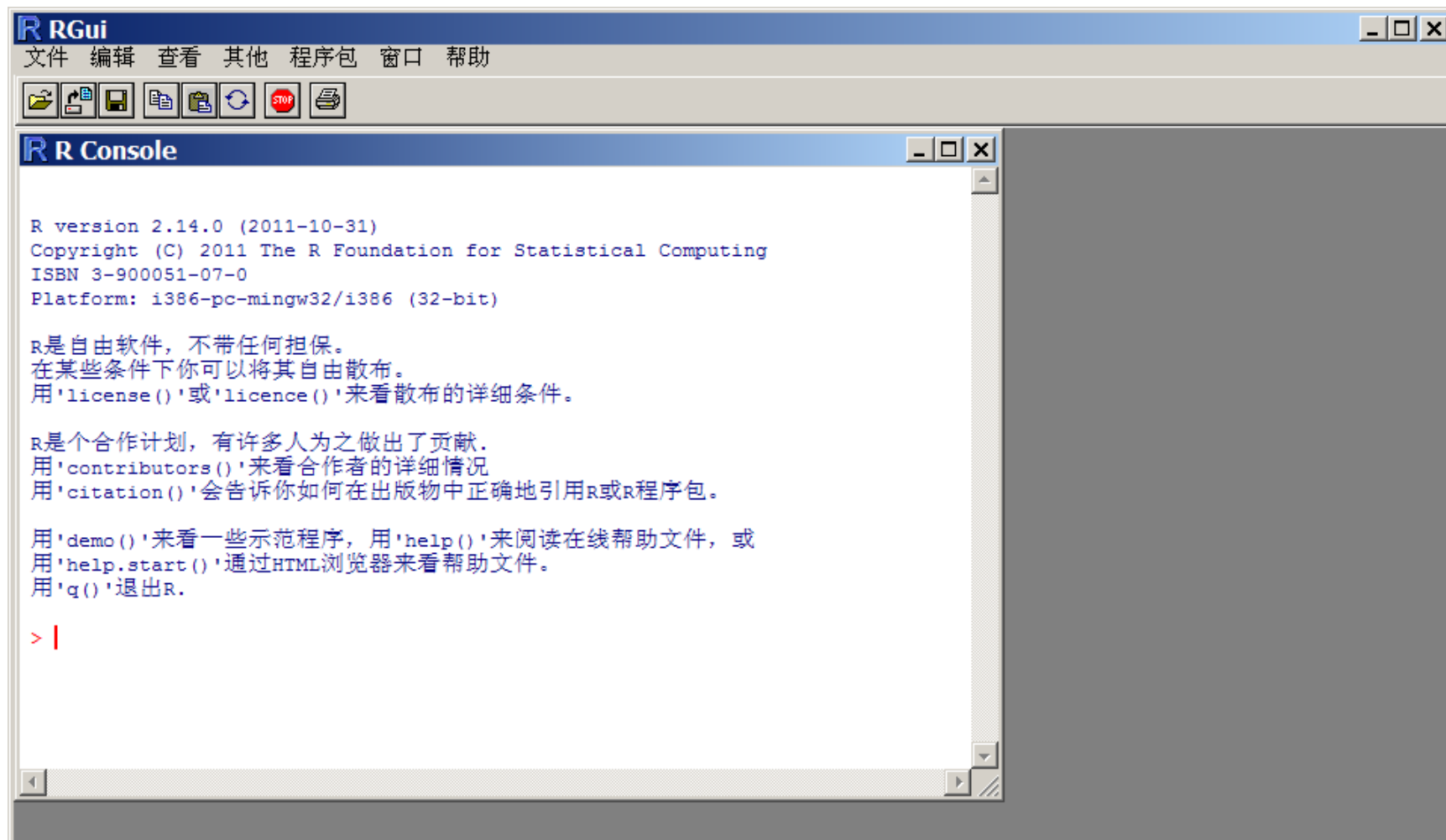
■ R的特点

1. 有效的数据处理和保存机制。
2. 拥有一整套数组和矩阵的操作运算符。
3. 一系列连贯而又完整的数据分析中间工具。
4. 图形统计可以对数据直接进行分析和显示，可用于多种图形设备。
5. 一种相当完善、简洁和高效的程序设计语言。它包括条件语句、循环语句、用户自定义的递归函数以及输入输出接口。
6. R语言是彻底面向对象的统计编程语言。
7. R语言和其它编程语言、数据库之间有很好的接口。
8. R语言是自由软件，可以放心大胆地使用，但其功能却不比任何其它同类软件差。
9. R语言具有丰富的网上资源

■ 商业版本的R

Revolution R (官网 : <http://www.revolutionanalytics.com/>)

很多大型厂商也在开始推出自己的R或兼容R的产品 , 例如Oracle、IBM、Sybase



创建向量和矩阵



- 函数 `c()`, `length()`, `mode()`, `rbind()`, `cbind()`

```
> x1=c(2,4,6,8,0)
> x2=c(1,3,5,7,9)
> length(x1)
[1] 5
> mode(x1)
[1] "numeric"
> |
```

```
> x1
[1] 2 4 6 8 0
> x1[3]
[1] 6
> |
```

```
> a1=c(1:100)
> length(a1)
[1] 100
> |
```

```
> rbind(x1,x2)
      [,1] [,2] [,3] [,4] [,5]
x1      2    4    6    8    0
x2      1    3    5    7    9
> m1=rbind(x1,x2)
> m1
      [,1] [,2] [,3] [,4] [,5]
x1      2    4    6    8    0
x2      1    3    5    7    9
> |
```

```
> cbind(x1,x2)
      x1 x2
[1,]  2  1
[2,]  4  3
[3,]  6  5
[4,]  8  7
[5,]  0  9
> |
```



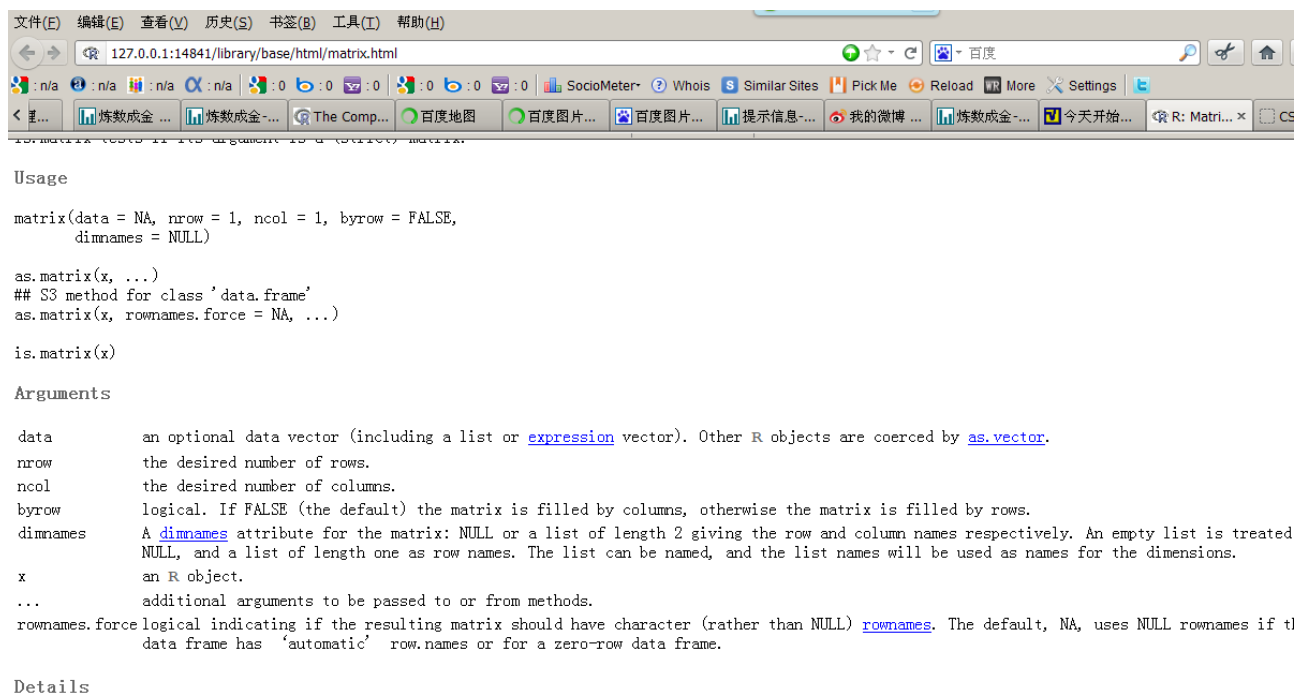

求平均值，和，连乘，最值，方差，标准差

- 函数mean(), sum(), min(), max(), var(), sd(), prod()

```
> x=c(1:100)
> mean(x)
[1] 50.5
> sum(x)
[1] 5050
> max(x)
[1] 100
> min(x)
[1] 1
> var(X)
错误于is.data.frame(x) : 找不到对象'x'
> var(x)
[1] 841.6667
> prod(x)
[1] 9.332622e+157
> sd(x)
[1] 29.01149
```

■ 函数help()

```
> help(matrix)
starting httpd help server ... done
> |
```



■ 函数matrix()

```
> a1=c(1:12)
> matrix(a1,nrow=3,ncol=4)
      [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
> matrix(a1,nrow=4,ncol=3)
      [,1] [,2] [,3]
[1,]    1    5    9
[2,]    2    6   10
[3,]    3    7   11
[4,]    4    8   12
> |
```

```
> matrix(a1,nrow=4,ncol=3,byrow=T)
      [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
[4,]   10   11   12
> |
```

■ 函数t(), 矩阵加减

```
> a=matrix(1:12,nrow=3,ncol=4)
```

```
> a
```

	[,1]	[,2]	[,3]	[,4]
[1,]	1	4	7	10
[2,]	2	5	8	11
[3,]	3	6	9	12

```
> t(a)
```

	[,1]	[,2]	[,3]
[1,]	1	2	3
[2,]	4	5	6
[3,]	7	8	9
[4,]	10	11	12

```
> |
```

```
> a=b=matrix(1:12,nrow=3,ncol=4)
```

```
> a+b
```

	[,1]	[,2]	[,3]	[,4]
[1,]	2	8	14	20
[2,]	4	10	16	22
[3,]	6	12	18	24

```
> a-b
```

	[,1]	[,2]	[,3]	[,4]
[1,]	0	0	0	0
[2,]	0	0	0	0
[3,]	0	0	0	0

```
> |
```

■ 矩阵相乘，函数diag()

```
> a=matrix(1:12,nrow=3,ncol=4)
> b=matrix(1:12,nrow=4,ncol=3)
> a%*%b
      [,1] [,2] [,3]
[1,]    70   158   246
[2,]    80   184   288
[3,]    90   210   330
> |
```

```
> a=matrix(1:16,nrow=4,ncol=4)
> a
      [,1] [,2] [,3] [,4]
[1,]     1     5     9    13
[2,]     2     6    10    14
[3,]     3     7    11    15
[4,]     4     8    12    16
> diag(a)
[1]  1  6 11 16
> diag(diag(a))
      [,1] [,2] [,3] [,4]
[1,]     1     0     0     0
[2,]     0     6     0     0
[3,]     0     0    11     0
[4,]     0     0     0    16
> diag(4)
      [,1] [,2] [,3] [,4]
[1,]     1     0     0     0
[2,]     0     1     0     0
[3,]     0     0     1     0
[4,]     0     0     0     1
> |
```

■ 矩阵求逆，函数 `rnorm()`, `solve()`

```
> a=matrix(rnorm(16),4,4)
> a
      [,1]      [,2]      [,3]      [,4]
[1,] 0.60714591 0.9354156 0.6471921 1.7788818
[2,] 0.03972303 -0.4784529 0.1773237 0.1755301
[3,] -1.59620992 -0.4553338 2.1706594 1.3569393
[4,] -0.56335648 -1.2811563 1.7136756 -1.2032154
> solve(a)
      [,1]      [,2]      [,3]      [,4]
[1,] 0.58375607 0.6422022 -0.51939339 0.37098393
[2,] 0.23718393 -1.8782607 0.01174602 0.08990033
[3,] 0.42730886 -0.5205396 -0.03332176 0.51823304
[4,] 0.08272524 0.9578674 0.18321945 -0.36243660
> |
```

■ 函数solve(a,b)

```
> a=matrix(rnorm(16),4,4)
> a
      [,1]      [,2]      [,3]      [,4]
[1,]  0.09502486 -0.2002975 -0.9340249  1.067134
[2,]  0.91382126 -0.8181392  0.8628442 -2.094286
[3,] -1.32881330 -0.5173477 -0.9182241 -1.635026
[4,]  0.28637823  0.6505220 -0.0399500 -1.469619
> b=c(1:4)
> b
[1] 1 2 3 4
> solve(a,b)
[1]  1.8470707  0.9692533 -3.1994782 -1.8458519
..
```



矩阵的特征值与特征向量

■ 函数eigen()

```
> a=diag(4)+1
> a
```

	[,1]	[,2]	[,3]	[,4]
[1,]	2	1	1	1
[2,]	1	2	1	1
[3,]	1	1	2	1
[4,]	1	1	1	2

```
> a.e=eigen(a,symmetric=T)
> a.e
```

	[,1]	[,2]	[,3]	[,4]
[1,]	2	1	1	1
[2,]	1	2	1	1
[3,]	1	1	2	1
[4,]	1	1	1	2

```
$values
```

[1]	5	1	1	1
-----	---	---	---	---

```
$vectors
```

	[,1]	[,2]	[,3]	[,4]
[1,]	-0.5	0.8660254	0.0000000e+00	0.0000000
[2,]	-0.5	-0.2886751	-6.408849e-17	0.8164966
[3,]	-0.5	-0.2886751	-7.071068e-01	-0.4082483
[4,]	-0.5	-0.2886751	7.071068e-01	-0.4082483



正定矩阵的Choleskey分解

■ $A=P' P$ 函数chol()

```
> a
      [,1] [,2] [,3] [,4]
[1,]     2     1     1     1
[2,]     1     2     1     1
[3,]     1     1     2     1
[4,]     1     1     1     2
> a.e=chol(a)
> a.e
      [,1]      [,2]      [,3]      [,4]
[1,] 1.414214 0.7071068 0.7071068 0.7071068
[2,] 0.000000 1.2247449 0.4082483 0.4082483
[3,] 0.000000 0.0000000 1.1547005 0.2886751
[4,] 0.000000 0.0000000 0.0000000 1.1180340
> |
```



数据的R语言表示——数据框

- 矩阵形式，但列可以不同数据类型
- 每列是一个变量，每行是一个观测值

```
> x1=c(10,13,45,26,23,12,24,78,23,43,31,56)
> x2=c(20,65,32,32,27,87,60,13,42,51,77,35)
> x=data.frame(x1,x2)
> x
```

```
  x1 x2
1  10 20
2  13 65
3  45 32
4  26 32
5  23 27
6  12 87
7  24 60
8  78 13
9  23 42
10 43 51
11 31 77
12 56 35
> |
```

```
> (x=data.frame('重量'=x1,'运费'=x2))
  重量 运费
1    10   20
2    13   65
3    45   32
4    26   32
5    23   27
6    12   87
7    24   60
8    78   13
9    23   42
10   43   51
11   31   77
12   56   35
> |
```

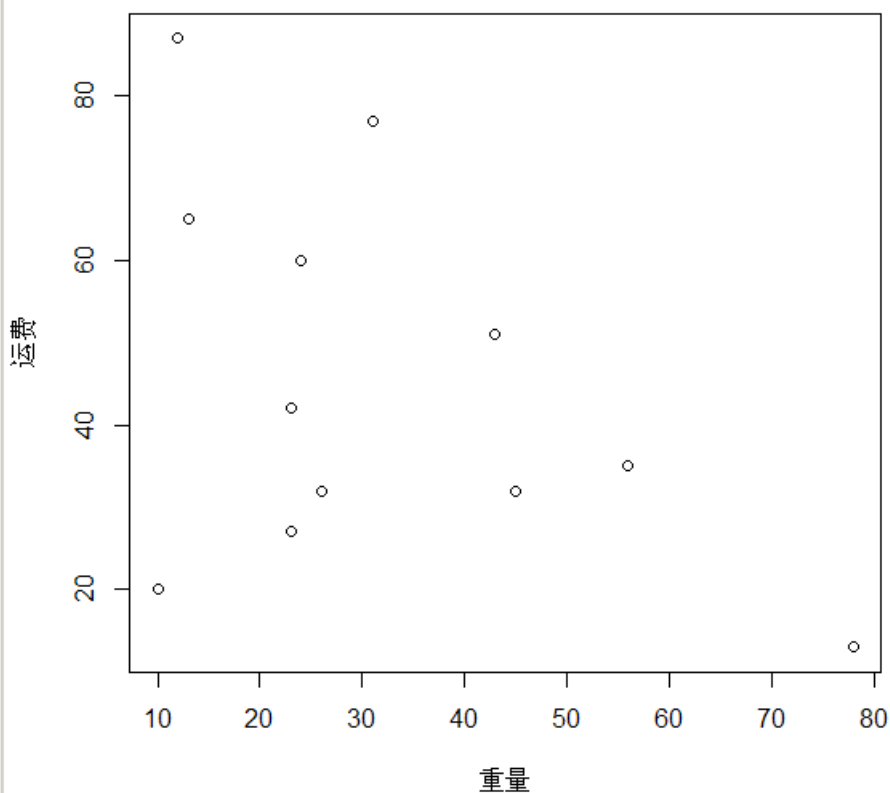
画散点图



中山大學
SUN YAT-SEN UNIVERSITY

■ 函数plot()

```
> plot(x)  
> |
```



2015.3.5

- 先设置工作目录，把文本文件放于该目录下

```
> (x=read.table("abc.txt"))  
      V1 V2  
1    175 67  
2    183 75  
3    165 56  
4    145 45  
5    178 67  
6    187 90  
7    156 43  
8    176 58  
9    173 60  
10   170 56
```

- 文本或excel的数据均可通过剪贴板操作

```
> y<-read.table("clipboard",header=F)
```

```
> y
```

```
      V1 V2  
1    175 67  
2    183 75  
3    165 56  
4    145 45  
5    178 67  
6    187 90  
7    156 43  
8    176 58  
9    173 60  
10   170 56
```

```
> |
```

```
> z<-read.table("clipboard",header=T)
```

```
> z
```

```
 商品  价格  
1    A    2  
2    B    3  
3    C    5  
4    D    5
```

```
> |
```

- 方法1：先把excel另存为空格分隔的prn文本格式再读

```
> w<-read.table("test.prn",header=T)
> w
  商品  价格
1    A    2
2    B    3
3    C    5
4    D    5
> |
```



读Excel文件数据

- 方法2：安装RODBC包，再通过ODBC读

```
> local({pkg <- select.list(sort(.packages(all.available = TRUE)), graphics=T$  
+ if(nchar(pkg)) library(pkg, character.only=TRUE))})
```

警告信息：

程辑包‘RODBC’是用R版本2.14.1 来建造的

```
> library(RODBC)
```

```
> z<-odbcConnectExcel("test.xls")
```

```
> (w<-sqlFetch(z, "Sheet1"))
```

	商品	价格
--	----	----

1	A	2
---	---	---

2	B	3
---	---	---

3	C	5
---	---	---

4	D	5
---	---	---

```
> |
```




中山大學
SUN YAT-SEN UNIVERSITY

Thanks

FAQ时间