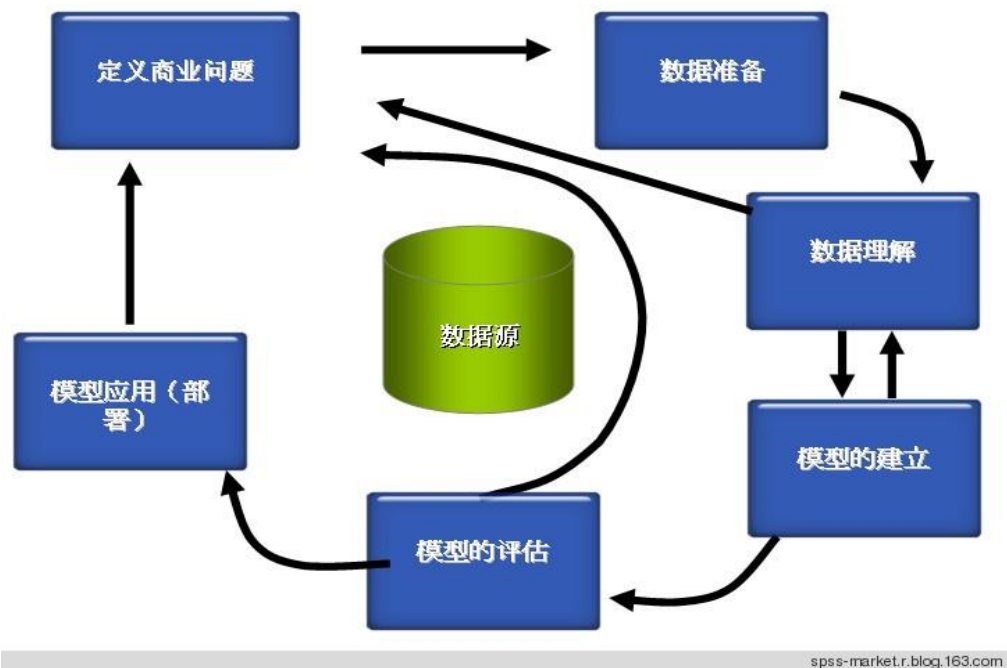


图说“什么是数据挖掘”

摘要: 1、数据挖掘需要‘神马样’的流程？ 2、哥，有没有详细点的，来个给力的！ 3、数据挖掘在商业上的理解是？ 4、数据在统计意义上有哪些类型？ 5、他们的含义是什么呢？ 6、基本的探测指标有哪些？ 7、数据挖掘的算法有哪些呢 ...

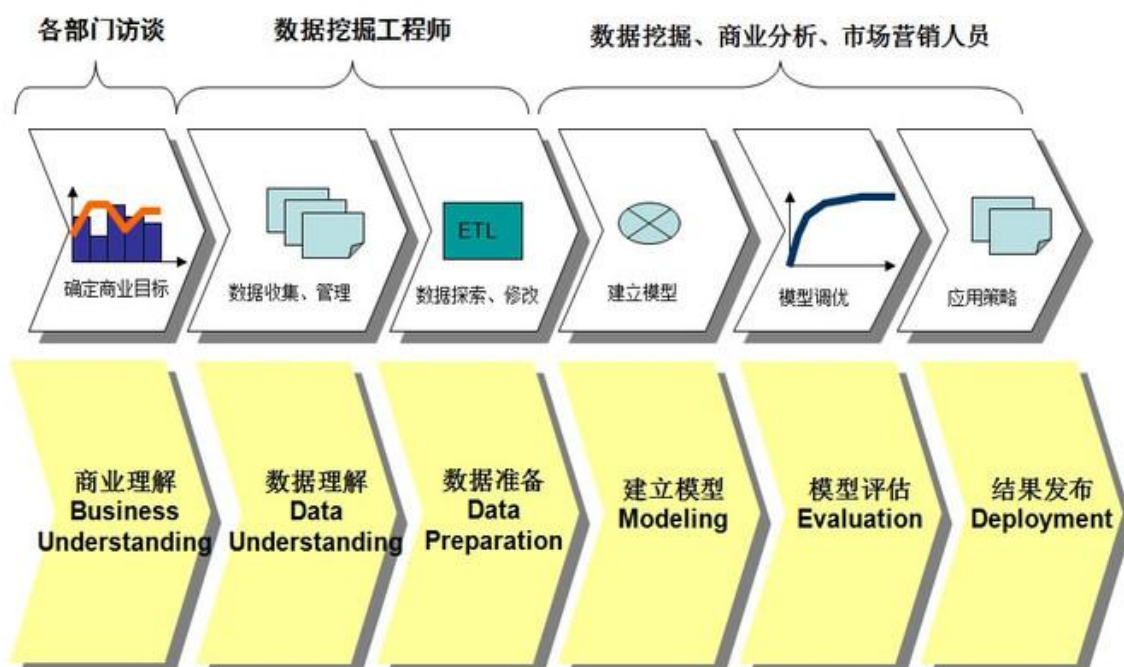
1、数据挖掘需要‘神马样’的流程？



2、哥，有没有详细点的，来个给力的！

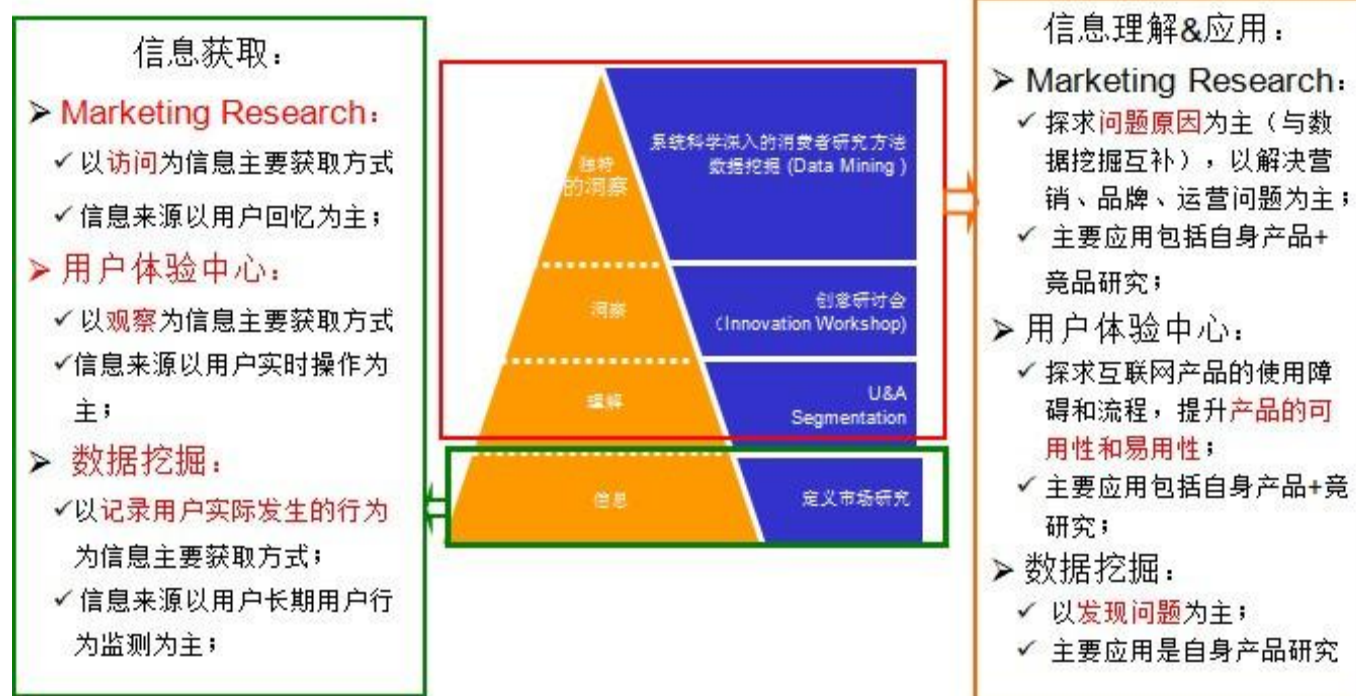


数据挖掘方法论—项目顺利实施的保证



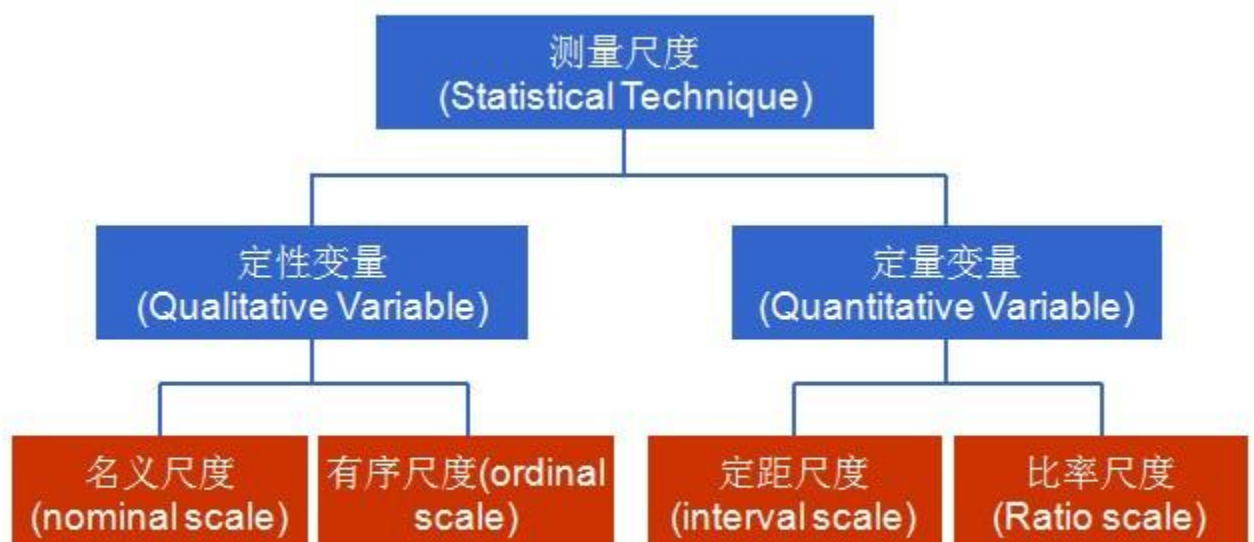
<http://spss-market.r.blog.163.com/>

3、数据挖掘在商业上的理解是？



<http://spss-market.r.blog.163.com>

4、数据在统计意义上有哪些类型？



<http://spss-market.r.blog.163.com>

5、他们的含义是什么呢？

名义尺度 (nominal scale)	<p>数字只用做对事物进行识别和分类的标志和标签</p> <ul style="list-style-type: none">▪ 例如：性别，婚姻状况，国籍/城市等；▪ 只允许计算有限的以频率计数为基础的统计指标，如百分比、众数等；
有序尺度 (ordinal scale)	<p>数字代表事物拥有某种属性的相对程度/位置，但没有指明差别的大小</p> <ul style="list-style-type: none">▪ 例如：偏好排序，市场/行业地位等；▪ 频率计数，以及基于分位点的统计指标(百分位数，中位数等)
定距尺度 (interval scale)	<p>尺度上数字相等的距离代表了被测特性的相等值，即可以比较事物之间差别的大小</p> <ul style="list-style-type: none">▪ 例如：偏好/态度量表(5-scale/7-scale)，重要性评分；▪ 零点位置不固定，即尺度可以变换；▪ 可以计算通常使用的统计量，但尺度值之间的比率及其它一些特殊统计量不适合计算；
比率尺度 (Ratio scale)	<p>可以依据尺度值对事物进行分类、比较等，以及计算相互之间的差值、比率等</p> <ul style="list-style-type: none">▪ 例如：年龄，收入，工作年数，花费等；▪ 有绝对零点，可以计算所有统计量；

<http://spss-market.r.blog.163.com/>

6、基本的探测指标有哪些？

集中趋势指标	均值(mean)	<ul style="list-style-type: none"> ▪ 即平均数，$mean = 1/n * \sum(X_1:X_n)$; ▪ 均值能够利用所有已知信息，但是对异常值(极小或极大值)很敏感;
	中位数(median)	<ul style="list-style-type: none"> ▪ 排序后居于中间位置的数值，有序尺度常用; ▪ 不能充分利用已知的所有变量信息，但不受异常值的影响;
	众数(mode)	<ul style="list-style-type: none"> ▪ 出现最频繁的数值，代表分布中的高峰; ▪ 名义尺度(分组数据)常用
变异性指标	极差(range)	<ul style="list-style-type: none"> ▪ 最大值与最小值之差，$range = max - min$; ▪ 直接受到异常值影响;
	方差(variance)	<ul style="list-style-type: none"> ▪ 离均差(观测值与均值之间的差)平方的均值; ▪ $var = 1/(n-1) * \sum((X_i - mean)^2)$; ▪ 数据分布越分散(远离均值)，方差越大;
	标准差 (standard deviation)	<ul style="list-style-type: none"> ▪ 方差的平方根，$stdev = \sqrt{var}$; ▪ 与数据本身有相同的量纲，常用;
变异性指标	偏度(skewness)	<ul style="list-style-type: none"> ▪ 刻画数据在均值两侧偏差趋势的差异性 ▪ 对称分布: $skewness = 0$, $mean = median = mode$; ▪ 右偏分布: $skewness > 0$, $mean > median > mode$; ▪ 左偏分布: $skewness < 0$, $mean < median < mode$;
	峰度(kurtosis)	<ul style="list-style-type: none"> ▪ 测量分布曲线相对平滑或突起程度 ▪ $kurtosis = 3$, 正态分布(Norm distribution); ▪ $kurtosis > 3$, 分布曲线比正态分布突起; ▪ $kurtosis < 3$, 分布曲线比正态分布平缓;

<http://spss-market.r.blog.163.com>

7、数据挖掘的算法有哪些呢？



8、需要掌握的工具有哪些？



9、知道这些工具不知道如何在工作中用呀？有没有‘浮云’般的角度？




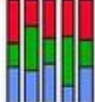

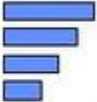






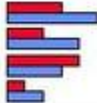






商业问题分析的角度



多维分析	从多个不同的角度及其组合去分析数据。
趋势分析	从时间序列分析随时间的变化趋势，找出其规律，如移动平均、同比、环比等。
意外分析	从大量历史数据中找出太高、太低、变化幅度过大等异常情况数据，支持预警显示、预警提醒；并可进一步进行相关影响原因的数据挖掘
排名分析	从大量数据中找出按某种分类方法的Top N或Bottom N数据，这些数据代表了需要特别关注的程度；
比较分析	从相同的角度去对不同数据集合进行对比，找出差异所在，并可进一步深入挖掘差异原因；
预测分析	利用决策树、回归分析、神经网络、时间序列等数据挖掘算法对客户响应、流失预警、收入预测、业务量预测、欠费预测等进行预测分析
.....

<http://spss-market.r.blog.163.com/>

10、结果如何可视化的展现？

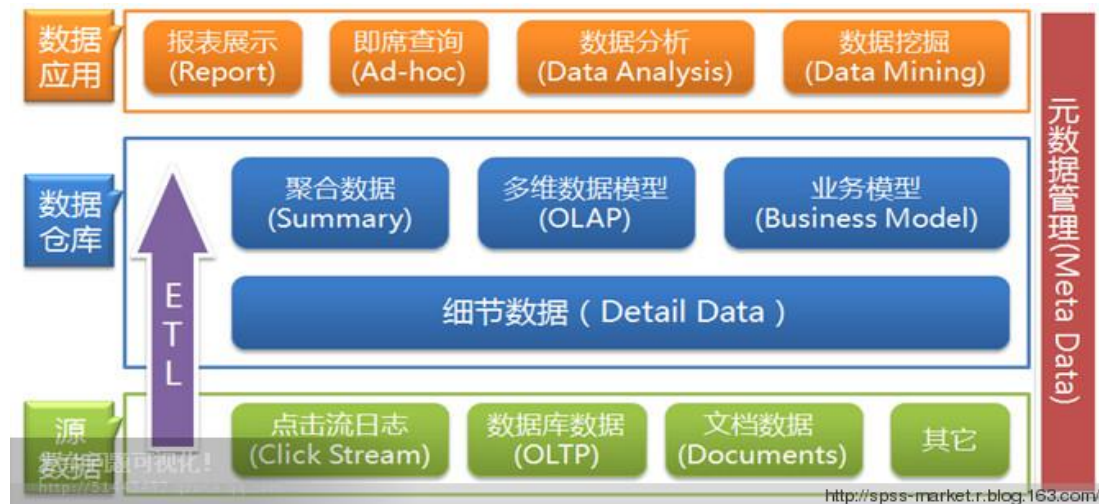
要表达的数据和信息	建议采用图形					
	饼图	垂直柱	水平柱	线图	水泡	其他
整体的一部分						
不同数据的比较						
时间序列						
频率						
两组数据的相关性						
和多重数据、标准相比较						   

spss-market.r.blog.163.com

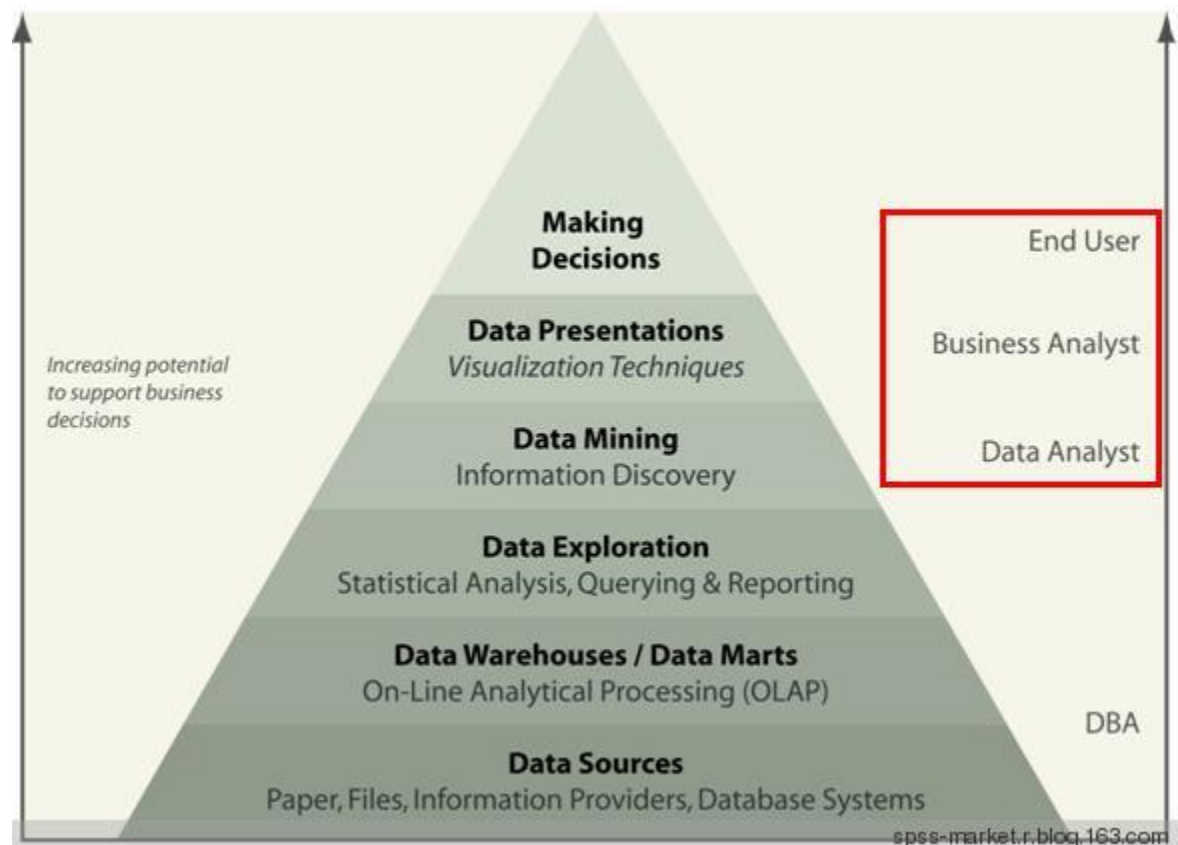
11、还有没有更人性化、智能化的展现？



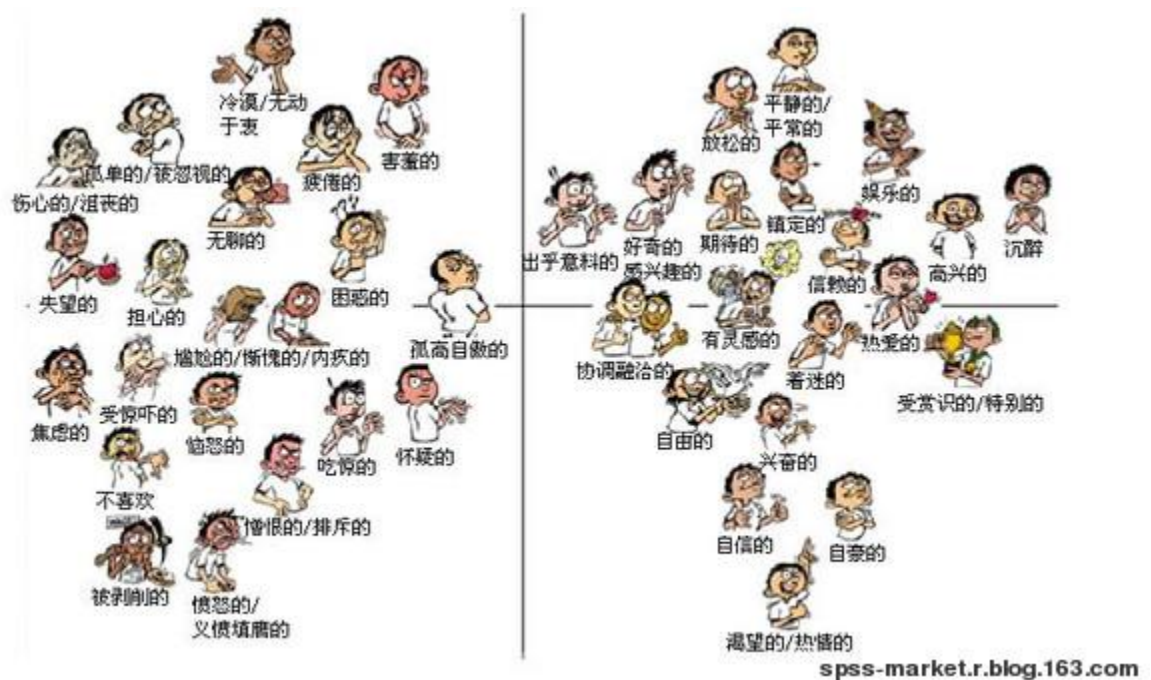
12、上面这图看起来很给力，背后很复杂吧？



13、职业的发展道路如何？



14、我的性格适合吗？（有志者，事竟成）



15、都说这行很累？NO！ 懂得生活。。。。



16、转载的留个来源，毕竟是我辛苦收集和想出来的，谢谢！

