

大数据时代的到来和蓬勃发展，让越来越多的人们开始觊觎或对大数据分析和大数据挖掘的爱好兴趣越来越浓厚，开始蠢蠢欲动想进去大数据分析与应用行列。这很正常，毕竟大数据时代下的数据分析师或数据挖掘师的岗位需求是不断被增加或持续增长的。但大数据的火热也终究会带来僧多粥少的局面。也许当初你一股脑钻进大数据分析培训的学习行列，想在数据分析培训认证的帮助让自己的数据分析师能力得到认可和提升，但是我们必须首先要搞明白的是我们是否真的了解大数据，自己是否真的对大数据非常感兴趣。很多人口口声声说自己的数据分析能力不错，或者说自己可以算得上是一名数据分析师了，但被人问起何为大数据时，可能还是一头雾水。所以啊，在我们决定进行数据分析培训或数据挖掘培训之前，在我们决心进入大数据分析或大数据挖掘行列前，我们要真正搞懂什么是真正的大数据。下面小编就和大家分享一篇文章，用白话大数据告诉你什么才是真正的大数据。

**一、大数据的四大误区**

**误区一 大数据就是数据大**

“光大是不够的!”当我掷地有声用这句话开场时，正好一个妹子推门而入，听到这句话，微微一怔，低头坐下。

如今在很多场合，一提起大数据，基本都会说“日处理数据量XXGB，上传图片XXGB，并发数XXX”“Hadoop集群拥有XXXX节点，总存储XXPB”……诸如此类的技术语言。但是不是数据大了，就可以达到大数据的境界，可以一起达到人生的大和谐了?

数据再大，不去使用，呆在机房里偏安一隅，那绝不是大数据，而是败家子。在这个问题上，传统的门户网站基本上就属于坐拥金山却无钱上青楼。无论搜狐新浪网易，每天的用户数以亿计，但是除了简单的广告呈现，并没有通过对数据的分析产生更多的价值。之所以没有提企鹅，是因为他们手里有最庞大的QQ用户，没推广但不代表没有做，例如当年推出QQ圈子的那种惊艳感，现在想起来还触目惊心。

而对于中小网站来说，不用盲目追求先进技术牛逼架构。首先需考虑的是商业运作和推广，只有用户量飙升见红的情况下，再去考虑技术升级。给君一个双选题，A.日登陆用户1000人，架构完全参照美国亚马逊从不宕机，B.日登陆用户10万人，每天因为高并发不得不宕机三次。你会选什么?

你不觉得“我那个网站因为用户太多太热情又宕机了”这句话又骄傲又装逼么?

**误区二 懂大数据就必须要懂技术**

“我不懂技术，我能学会大数据吗?”

大数据更多的是一种能力，而非技能，这种能力能从无尽的数据里看出商机看出价值。诸葛亮懂兵法，他知道该在哪里伏兵该在哪里生火，他不必知道关羽是如何耍大刀，也不必知道张飞杀人时，蛇矛是扎还是砍。

**误区三 所有公司都必须懂大数据**

我承认如果卖煎饼果子的大妈能开发一个APP，拿到客户对煎饼里薄脆是否爽口，辣椒酱是否够味的直接意见，那么很有利于产品的改进。但是你会为一个3元钱的煎饼去下载一个8M的APP吗?

而在移动互联网大会上，Evernote的CEO Phil Libin明确说自己产品的商业模式就是向用户收费，让他们甘心为产品体验付费，而不玩当下流行的大数据那一套。

知道自己能玩什么很常见，而知道自己不能玩不必玩什么更可贵。

**误区四 数据越多越好**

从陈冠希到李宗瑞，无不透露出浓浓的收藏癖。那些被大家藏在硬盘深处的姑娘们，真的有重见天日的一天吗?无论东京热还是加勒比，由来只有新人亚买碟，有谁听得旧人COME ON?

而搜一搜“存款 贬值”的关键词，从“五十年前的百万变13块”，到“一万元存一年赔19元”都可以看出，钱必须要用才有价值，数据也是一样。

只有不停的使用数据，挖掘数据背后的关系和价值，才能如滚雪球一般，使数据之间的相互关系更丰富更完善。



**二、大数据的核心思想**

**1.有什么不重要，怎么用才重要**

在会上，无奈的发现每当举男女之间的例子，众人理解能力就飙升，而讲偏产品技术方面时，下面就一脸木然昏昏欲睡。(诸位!!!你们是来听色情讲座呢?)

Netflix公司利用大数据来拍《纸牌屋》的励志故事大家都听过无数遍。Netflix是美国最大的DVD与网络视频租赁网站，他们在国内有2700万用户，全球有3300万。他们从流媒体视频用户处收集到的数据是异常惊人的，每一次搜索，每一次暂停，每一次积极或者消极的评价，还有自己的位置数据设备数据社交媒体数据，在分析完这些数据之后，发现自己的这些观众都喜欢演员凯文-史派西，还喜欢导演大卫-芬奇，并且喜欢1990年的英国同名电视剧。三项综合在一起，Netflix就下定决心拍摄《纸牌屋》，并将数据分析运用得淋漓尽致。在Netflix的观影页面上，提供暂停后截图的功能，他们就依靠这种数据来判断观众更喜欢哪种布景和画面。

当然，这三个因素其实都有不稳定性，例如凯文-史派西也曾拍过只被评了4.2分的《弗雷德圣诞老人》，大卫-芬奇首次执导的《异形3》也成系列中口碑最差的一部。但是从概率上来说，这三个成功因素都包括的剧集更靠谱。

另外国政通推出的“同名同姓”应用，虽然看起来比较偏娱乐化实用性不强，但是这背后的那些身份数据是可以进行各种二次利用。每次看到类似应用，我总会把前同事“操瑞士”的奇葩名字输进去，现在只有这款应用准确的找到了全国只有1个操瑞士，并且是安徽人。呼叫瑞士大使馆迅速锁定他的位置吧!

所以大数据的核心并不是拥有数据，而是拿数据去做了什么。也就是说你不能仅仅占有人家的身子，也要占据人家的心灵。

**2.对数据的容错性更强，来源更加多样化**

一个500M的用户数据excel表算不算大数据?曾经在做某款产品的用户筛选时，这么一个excel成功拖死我电脑三次，我当时愤怒的说：“KAO，这SB大数据!”现在想起来，我冤枉了大数据君，您受苦了。

真正的大数据，应该是从不同维度，不同途径过来的各种格式数据碎片，并不限于文字/视频/声音/位置/图片等。只有将不同维度的数据放在一起判断，得出来的趋势才有可能更真实。雷同的数据积累的再多，超过某一个限度之后，我们再从新样本上获得的有用信息就越来越少，就如同经济学上的边际效应递减一样。而用简单点儿的话解释就是如果你已经交往过五个IT码农，那么第六个其实也不会新鲜到哪里去，倒不如重新调整方向，换个高富帅找点儿不同的刺激。

多样化的来源同样可以避免陷在一个死胡同里出不来。例如以前经常说的“三年自然灾害”，如果仅仅去查询这三年的天气状况和死亡人数的关系。那么可能最后得出结论是“晴朗天气比阴天更容易致人死亡”。但实际上如果去结合《人民日报》亩产十万斤的相关翔实公正报道再加上那些年中国粮食对外进出口情况，那么就会得出更有建设性的结论。在一90后妹子的追问下，只说了四个数字。1958年，中国出口粮食288.34万吨，进口22.35万吨。困难时期第一年，1959年，中国出口415.75万吨，进口0.20万吨，那年，听说大家都好饿。别不多说，免得喝茶。

再举个例子，这两天淘宝首页给我推送的个性化广告除了情趣用品就是“花花公子”服饰大优惠，我不就是为了找素材和配图搜了一下“情趣用品”么?你们没完没了的给我推送这个，节操何在?如果你们能拿到我小学老师的评语，初中老师的评语和操行评定，高中老师的评语和操行评定，通过对多个来源不同的数据源分析，那么你们给我推送的一定会是《钢铁是怎么炼成的》《雷锋日记》等好书。(大学老师的评语就算了，那厮打CS被我蹂躏了太久容易怀恨在心。)

**3.拥有大数据的身，也要有大数据的心**

以往的数据分析，更多的是精确的样本/深度的数据挖掘，“精确”就是其代名词。不符合规格的样本过滤掉，然后再深度挖掘数据字段间的关系，得出几个精确无比数字去做PPT，或者从一系列数据里精准的找到某个正一脸猥琐偷拍姑娘的人。

但是大数据更多的是通过对各种数据分析得出某种趋势，这种趋势不必过于精确，但是能让相关决策人有底气去做某项决定。大数据不重要，重要的是使用大数据的人。

因为哪怕面向完全相同的数据源，不同的人得出的结论或者决定也可能是截然不同的。三国赤壁大战，当庞统献计“若以大船小船各皆配搭，或三十为一排，或五十为一排，首尾用铁环连锁，上铺阔板，休言人可渡，马亦可走矣”时，同时听到这话的两个人，曹操下席而谢，“非先生良谋，安能破东吴耶!”，而徐庶却私下里扯住庞统“你好大胆，只恐烧不尽绝。”由此可见，“以人为本”是多么精辟的一句废话啊!

无论你NB还是SB，数据总在那里，不离不弃。

**4.强调趋势和未来**

大数据，更多的应该是分析过去，提醒现在，展望未来。无法用到实践中去的大数据都是耍流氓，无论这个结果是造福了全人类，还是帮助网站提高1%转化率，这都是有用的。

上图这充满着暴力气息的玩意是当年冷战期间美国和苏联研发的高速列车，他们将轰炸机上的涡轮直接装在了火车顶上。其中美国的M-497号在1966年俄亥俄州的铁轨上跑出了295.54公里/小时。虽然几十年后，才有了现代高速铁路，但是如果没有最初的这种野蛮实验，恐怕现在自主研发的高铁也没有那么充足的底气。

谈趋势谈大数据，怎么能不提谷歌的流感趋势呢?谷歌分析自己上亿次的搜索查询，近乎实时的提供全球许多国家和地区的流感疫情评估，从截图可以看出，谷歌的趋势曲线跟美国官方公布的数据重合度非常高，但是后者在时间和效率上完全无法比拟谷歌趋势。

如果你在准备出差的时候，发现你要去的地方正爆发鼠疫，我想大部分人肯定就哭着喊着辞职不去了。

再例如2012年美国某著名避孕套品牌发布校园性健康报告，报告表明在美国大学生中，25%得了性病，其中排名第一的是尖锐湿疣，得病率最高的学校是位于科罗拉多州的美国空军学院(U.S. AIR FORCE ACADEMY)。然后紧接着，无聊的我又从学院官网中看到学校男生占比78.1%，女生占比21.9%。

从这些数据至少我们可以分析出一个结果：跟海归发生关系之前，请查一下学位证。



**三、大数据应用瞎想**

**1.医疗服务**

通过一系列的病历等数据，得出某一类人的健康趋势，例如“做IT8年”“每天加班6小时”“病中还坚持工作”“每天吃盒饭”“每天一包烟提神”“每天一杯咖啡提神”将这些标签具体到某一群人，然后个性化的推送一条如上图的信息。

还有人要加班吗?

**2.罪案预警**

通过对全市的历史罪案以及出警纪录得出城市各种罪案高发区域，例如剁手区/掏包区/性侵区……等等，每天发布提醒，推送到市民的手机/平板等所有终端。我相信性侵区的流氓早晚会不堪重负而投案自首的。

**3.大龄青年相亲**

此想法如果搭配上类似Google Glass的可穿戴设备就更棒了，在上百人的男女相亲大会上，你带着功能齐全的眼镜四处看妞，系统自动分析各种数据，帮你找到最搭配的另一半，按照匹配度自动走在一起。当然，如果那时红外线透视功能很发达，能加入匹配度最高可以付费看裸体的VIP功能就更棒了。