CDA数据分析师系列丛书

大数据思维

从掷骰子到纸牌屋

马继华 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

数据分析不在于你掌握了多少先进的软件工具，也不在于你拥有多么高智商的头脑，而是要靠更大视野、更宽角度和更具有逻辑性的思维。本书不是一本介绍大数据概念的流行读物，也不是开讲编程工具高深理论的专业教材，而是立足于大数据之上的思维模式的普及。读者不需要任何统计学知识，也没必要掌握复杂的公式与算法，在最通俗易懂的案例介绍和娓娓道来中就可以轻松理解大数据分析的基本模式与方法。

作为读者，你可以是大中专院校的数据分析专业学生，也可以是企事业单位的经营分析人员，或者是任何行业任何职业中喜欢“头头是道”的分析爱好者。开卷有益，即便你从来不需要大数据，也可以从本书中领悟到思维魔力，因此让工作与生活更充满智慧与乐趣。

图书在版编目（CIP）数据

大数据思维：从掷骰子到纸牌屋 / 马继华著. —北京：电子工业出版社，2016.7
（CDA 数据分析师系列丛书）
ISBN 978-7-121-29407-5

Ⅰ. ①大… Ⅱ. ①马… Ⅲ. ①数据处理 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2016)第 163950 号

策划编辑：石　倩
责任编辑：石　倩
印　　刷：北京季蜂印刷厂
装　　订：北京季蜂印刷厂
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：720×1000　1/16　印张：17.5　字数：281 千字
版　　次：2016 年 7 月第 1 版
印　　次：2016 年 7 月第 1 次印刷
定　　价：55.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819　faq@phei.com.cn。

前言

早就想写一本关于数据分析的书，最主要的原因就是，自己是统计专业毕业，又从事过多年数据分析的工作。工作几经变迁，现在已经很少用软件重操旧业，但却越来越感觉到数据分析的重要性。

经常看网络、电视和报纸上的很多分析，在信誓旦旦的说教与言之凿凿的数字之外，很多却是惨不忍睹的分析过程，甚至说是误人子弟也不为过。因为自媒体的流行，很多人根本没有基本的分析方法和技巧，在违背常理的情况下做出了很多奇异的解释，将大家引导到错误的方向。

最为可笑的，曾经有一次看到某知名报纸上的文章，分析的是中国信息分类领域的两家互联网巨头：58 同城与赶集网（这两家公司在 2015 年宣布合并）。当时，58 同城刚刚上市，这家报纸的专栏作者发表了一篇针对性的分析文章，文中称，他查阅了 ALEX 网站，58 同城的流量排名在世界网站的第 300 名，而赶集网排名是第 900 名。于是，这位作者就果断地下结论说，以上数据足以证明 58 同城的网络流量是赶集网的 3 倍。呜呼，如此分析竟然逃过了多少编辑的眼睛，甚至还

被众多读者接受，是多么可悲！

在实际工作中，一些人虽然科班毕业，通晓各种分析工具，甚至对各种各样的软件如数家珍，编程造模轻车熟路，但却对具体的分析套路与方法形同陌路，只能机械刻板地对数字结论进行解读。实际上，这样的数据分析还不如不做，错误的分析和错误的解读同样都是害人不浅。

当然，由于分析能力不到位，让自己吃亏上当丢人的案例更是不胜枚举。中国足协就是典型案例。2013 年，人所共知的原因，中国足球终于迎来了出人头地的机会，中国足协更是喜出望外。为了配合隆重的节日气氛，也是要彰显一下中国足球有雄起的能力，中国足协费尽心思地组织了一场国际足球友谊赛。

中国足协应该在邀请友谊赛的对手方面煞费苦心。邀请德国队？肯定不行，严谨的德国人不明就里的职业精神会破坏比赛气氛。邀请西班牙队？鼎盛时期的西班牙与中国队比赛也必须让自己有一个可以接受的成绩，否则被人笑掉大牙。于是，中国足球邀请了我们的近邻，泰国队，可怕的比赛开始了。估计包括中国足协官员在内的中国球迷都没有想到，一场友谊赛进了 6 个球，更重要的是，我们只进了一个，泰国队进了 5 个。

如果中国足协进行了充分的数据分析，也许就会避免这场悲剧的发生。历史数据证明，中国队此前已经多年没有胜过泰国队。如今的中国队不再是以前的那支“中国头球队”，依靠身高与体重就可以战胜东南亚球队，几年来学西班牙控制脚下球的中国队既没有学到技术，也忘记了本分，对付泰国这样的小老虎已经心有余而力不足。或者，这场比赛还不如邀请韩国，场面也不会失控。

如果我们非要挖苦一下数学水平奇差的中国足协，那也是可以的。

因为，某年某月某日的世界杯外围赛亚洲区预选赛，中国与黎巴嫩同组，在最后一轮比净胜球决定出线的关键时刻，中国足协竟然鬼使神差地算错了账。当全场球迷因为中国队 7∶0 战胜中国香港而成功惊险获得出线权而欢呼的时候，足协才明白过来，8∶0 才出线，我们已经被淘汰出局。这样的数据分析能力怎有能力让中国足球拿下大力神杯？

从历史上看，中国一直不是一个靠数据化进行管理的国家，我们太多的中庸之道和模糊分辨，“好好好”、“是是是”、“差不多”，贯穿着经济和社会管理的始终，这个模式也对中国的国家统计局产生着潜移默化的影响，也直接造成了人们对国家统计机构数字的不信任。

数据分析是每个人生活与工作的基本功，小时候对父母的察言观色也是在分析，长大以后的相亲娶妻也要分析，工作中的汇报决策更需要分析，炒股理财也离不开分析。数据分析无处不在，数据分析无时不在，数据分析伴随我们生命的始终。

我们生活的世界变化是如此之快。电力引入美国 46 年后，才覆盖 1/4 国民；电话花了 35 年；电视机 26 年；宽带呢？只用了 6 年。2007 年，数码世界容纳了 2810 亿 GB 的数据，全球平均每人 45GB，数码资料首次超越保存空间总量，目前，互联网每小时处理的数据量已经超过 1EB。

要给美国国会图书馆填满逾 5700 万份手稿、2900 万册书籍和期刊、1200 万张照片及其他，需时 2 个世纪，现在全球每日生成的数码资料几乎是这些的 100 倍。人类 5000 年的文字记载总共是 5EB，今后每年将产生的数字内容超过 1000EB。

我们所拥有的数据量在海量暴增，我们认识世界的水平也在不断提高。大数据时代来了，我们的思维是不是也应该有所改变？

目录

第 1 章

大数据与人脑的较量

BAT 为何如此了解我们

开篇，我们来讲一个简单的问题，你知道腾讯的 QQ 与微信的重要区别是什么吗？

现在的中国人，如果有人问你，你用 QQ 或者微信吗？估计很少有人会回答“否”。因为，QQ 或者微信已经深入到我们生活的各个方面，成为工作与生活的必需品。

可是，如果问你，QQ 与微信有什么区别？估计很多人答不上来。或者有人会说，QQ 有空间，微信有朋友圈；还有人会说，QQ 能穿衣服，微信没有。这些也是差别，但却没看到本质。

通过大数据的分析，我们也许能得到更为靠谱的答案。我们试着再提示一下，你在使用 QQ 的时候，使用频率最高的词是什么？这个问题如果问腾讯，腾讯可以通过系统地查询很容易地得到答案。我们普通用户实际上也能说得出来。一些人说，QQ 上使用频率最高的词是“呵呵”或者“哈哈”，还有“哦”，但更多人会联想到一个词，那就是“在吗？”

是的，我们需要的答案就是“在吗”。因为，我们可以对比一下，你在使用微信的时候，还会经常使用“在吗”吗？答案是，不会。

以上的分析，我们就是使用了最简单的词频分析，以最简单的数数的方式获得了最佳的分析路径，因为一句“在吗”就能充分地展示 QQ 与微信的本质差别。

我们通过进一步分析可知，因为 QQ 是互联网时代的产物，后来与移动互联网相结合，因此，QQ 有电脑客户端，也有手机客户端。大家使用 QQ 的时候之所以经常第一句说“在吗”，是因为我们无法判断

对方是否在线（或者没在电脑前或者在隐身），即便有人在电脑前，我们也无法断定是否本人正好坐在电脑前，所以，先问“在吗”可以确认身份，以便开启下一步的对话聊天。而微信是移动互联网的产品，其主要使用环境是在手机端，手机是绝大多数人形影不离的用品，而且是个人用品，移动互联网又是实时在线，我们与人用微信联系的时候根本无需先问“在吗”，因为，只要这个人还在，他就一定在。你这个时候问对方“在吗”，实际的含义是“你还活着吗？”

一个简单的“在吗”就形象地刻画出了腾讯的两个产品 QQ 与微信的代差，也找到了互联网与移动互联网产品分析的钥匙，这是多么神奇？

接下来，如果你是中国移动的员工，或者是通信行业的分析师，如果要分析中国移动的飞信产品，那与之进行对比分析的产品应该是 QQ 还是微信？很简单，应该是 QQ，而不是同样有一个“信”字的微信，因为，飞信与 QQ 同样都是互联网时代的产品，都拥有电脑客户端和手机客户端，而且都可以同时在线。

分析就是如此，只要你找到了窍门，四两拨千斤，简单的方法可以解释大道理，何必非要扎在数据堆里当无头苍蝇呢？

对用户的使用行为研究最充分的，无疑是阿里巴巴。很多人都发现，只要你打开淘宝，首页上的推荐就让你欲罢不能，特别是网页中间那张跳动的大图，怎么看都是自己想要的商品。是的，淘宝说要实现千人千面，每个人看到的网页都是不一样的，因为那个页面就是根据你最近的搜索、下单等历史行为结合你的各种资料进行“定制”的。

有这样一个小故事：一个连锁商店，专门有一个铺子卖婴幼儿产品。因为客户信息很多，就发现当人怀孕之后，行为会出现改变。比如会更多选择没有香味的洗发水，买营养品的时候口味也和怀孕前有不同。商店便可以根据客人购买行为的变化，预测是否可能怀孕了，然后给可能怀孕的客人寄婴幼儿产品广告，说买我的尿布吧，买我的奶粉吧。一天，一个父亲很愤怒地过来说，“我女儿还在高中，你们现在天天给她寄婴儿尿布、奶粉的广告，什么意思？你鼓励未婚怀孕啊？”然后商场说，“对不起，我们搞错了！”过了一个星期，这个爸爸又回来，说：“对不起，我搞错了，我女儿已经向我坦白了，她真的怀孕了。”

在现代企业经营中，电子商务都非常重视针对性的产品推荐，比如淘宝，更具有大数据应用意义的就是信用评价，比如芝麻信用分。芝麻信用公布了基本的计算模型，综合考虑了个人用户的信用历史、行为偏好、履约能力、身份特质、人脉关系五个维度的信息，没有任何一个单项信息能够直接或完全决定个人的芝麻分，其五个维度包含

的内容举例如下：

（1）信用历史：过往信用账户还款记录及信用账户历史；

（2）行为偏好：在购物、缴费、转账、理财等活动中的偏好及稳定性；

（3）履约能力：享用各类信用服务并确保及时履约；

（4）身份特质：在使用相关服务过程中留下的足够丰富和可靠的个人基本信息；

（5）人脉关系：好友的身份特征，以及跟好友互动的程度。

根据这个计算模型，我们大概可以总结出一些规律，能够帮助个人提高自己的信用得分。

（1）你要至少办一张信用卡，并经常在网上进行消费，特别重要的是要记得按时还款，如果你是使用支付宝进行按时还款，那么肯定会增加信用分。

（2）即便你有钱，也要使用下“花呗”、小额信用贷款等，并设置自动还款，保证你的账户里有这笔钱到时候准时还上，如果你不设置自动还款却能按时手动还款，那信用的分数肯定会提高。

（3）使用支付宝进行慈善捐款，如果是每年每月都坚持下来，即便数额不大，也会对信用分数帮助不小，因为理论认为做慈善的人信用比较好。

（4）发发红包，不管是定向发还是抢红包，都表明你乐善好施并且不差钱，信用不会差。

（5）多交几个有钱的朋友，并经常在网络上互动，如果发现谁经常信用卡不还，赶快绝交，至少也要在网络上不要来往。

（6）在网上买东西，要记得收到货物之后尽早地主动支付而不是

等系统默认付款，最好要给买家进行评价，如果能不厌其烦地多写几句话，就更好了。

（7）网购时的收货地址要力争保持稳定，如果你是租房或经常变换居住地，或者是房子太多经常换地方住，那也要选最稳定的地址来收货，比如办公室的地址，或者直接是一个居住稳定的朋友代收。经常换地方收网购商品对信用影响很大。

（8）如果可能，就把自己的网购账户的信息多填点，那些多人或家人公用一个账号的自然在个人信用评分上会受到影响。

（9）如果你有钱，在各互联网公司的理财产品里放些闲钱，既能保障收益，也可以让自己看起来是个有钱人。

怎么样？数据分析很有用吧，不仅可以帮助企业了解客户需求，还可以帮助客户找到针对性地提升自己社会信用的方法。掌握简单的科学的数据分析方法，对所有人都是必要的。

战争是各种矛盾最为激烈的表达，而数据分析更是战场指挥员不可缺少的工具。最为著名的案例就是，林彪靠战利品分析意外地快速结束了辽沈战役。

据资料记载，在中国革命战争年代的十大元帅中，林彪非常有特点，从白山黑水到天涯海角，战功卓著。据说，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，其实这就是在积累大数据。1948 年的辽沈战役，是决定国共命运的大决战开端。每天深夜，林彪都在东北野战军前线指挥所里听取军情汇报，由值班参谋读出下属各个纵队、师、团用电台报告的当日战况和缴获情况，而林彪则认真细致地记录着他的大数据：每支部队歼敌多少、俘虏多少；缴获的火炮多少、车辆多少、枪支多少、物资多少……作为

司令员，林彪的要求很细，俘虏要分清军官和士兵，缴获的枪支，要统计出机枪、长枪、短枪，击毁和缴获尚能使用的汽车，也要分出大小和类别。

一天深夜，值班参谋正在读着下面某师上报下属部队的战报，说他们的部队碰到了一个难度不大的胡家窝棚遭遇战，歼敌部分，其余逃走。与其他之前所读的战报看上去并无明显异样，值班参谋就这样读着读着，林彪突然叫了一声“停!”。林彪接连问了三句：“为什么那里缴获的短枪与长枪的比例比其他战斗略高？”“为什么那里缴获和击毁的小车与大车的比例比其他战斗略高？”“为什么在那里俘虏和击毙的军官与士兵的比例比其他战斗略高？”林彪不等别人回答，指着地图上的那个点说：“我猜想，不，我断定！敌人的指挥所就在这里!”结果，部队集中兵力攻击，很快抓获了廖耀湘。从大批杂乱无序的数据中将信息集中、提炼，分析出研究对象的内在规律，找到蛛丝马迹的异常变动，从而为决策提供最强支撑。

神奇的不仅仅是林彪，还有柳传志，更是擅长根据蛛丝马迹的数字做出自己的判断。据说，柳传志的创业起因非常具有传奇色彩，只是因为看了一张再普通不过的报纸。有时候，借助敏锐的数据分析能力就可以发现别人不易察觉的变化，从而让自己的人生大不相同。

1978 年 11 月 27 日，中国科学院计算所 34 岁的工程技术员柳传志按时上班，走进办公室前他先到传达室拎了一个热水瓶，跟老保安开了几句玩笑，然后从写着自己名字的信格里取出了当日的《人民日报》，一般来说他整个上午都将在读报中度过。20 多年后，他回忆说：

“记得 1978 年，我第一次在《人民日报》上看到一篇关于如何养牛的文章，让我激动不已。自打‘文化大革命’以来，报纸一登就全

是革命，全是斗争，全是社论。在当时养鸡、种菜全被看成是资本主义尾巴，是要被割掉的，而《人民日报》竟然登载养牛的文章，气候真是要变了！”

从现在查阅的资料看，日后创办了赫赫有名的联想集团的柳传志可能有点记忆上的差失。因为在已经泛黄的 1978 年的《人民日报》中，并没有如何养牛的文章，而有一篇科学养猪的新闻。在这天报纸的第三版上，有一篇长篇报道是“群众创造了加快养猪事业的经验”，上面细致地介绍了广西和北京通县如何提高养猪效益的新办法，如“交售一头可自宰一头”、“实行公有分养的新办法”，等等。柳传志看到的应该是这一篇新闻稿。

不过，是养牛还是养猪似乎并不重要，重要的是，举国之内，确有一批像柳传志这样的人，“春江水暖鸭先知”，他们在这个寒意料峭的早冬，感觉到了季节和时代的变迁（节选自《激荡三十年》）。

还有更神奇的大数据应用，即便是很多美女最喜欢玩的自拍，也有可能成为大数据应用的先驱，因为网络上忽悠你做明星脸对比的，往往都是一些人脸识别的程序在收集素材训练“机器人”。

媒体报道，史上最昂贵的自拍照应该是诞生于 2007 年。两名美国大兵在伊拉克的军营中玩自拍传到了社交网络上，结果几天之后，这个秘密的驻扎地就遭到了恐怖分子火箭弹的袭击。四架“阿帕奇”直升机惨遭击落，两亿美金灰飞烟灭。美军情报部门百思不得其解，最后才发现：原来是大兵的自拍照中附带了经纬度信息，让“好友”轻易掌握了他们的位置。但是就在 2015 年，某 ISIS 成员在其“总部大楼”自拍，并且在社交网络上大肆吹嘘这里的指挥能力有多么“炸裂”。结果一语成谶，22 个小时之后，这幢大楼就被美军三枚导弹“强拆”了，

“炸裂”得粉身碎骨。

其实，每个人生活的痕迹就是大数据。如果有一种技术可以轻易地记下你的脚印，那么你的爱好、习惯、职业、经济状况、婚姻状况都可以通过你去的地方精确展现出来。只不过问题在于，脚印这种数据非常难以记录。

在央视 2015 年 10 月 25 日晚播出的《挑战不可能》总决赛中，董艳珍通过观察 15 个孩子的行走步态，顺利将其中来自同一家庭的四胞胎全部选出，并将四个孩子分别与其光脚脚印一一对应，最终获得“年度挑战王”桂冠。连节目评委、有华人神探之称的李昌钰也为之折服，称要拜她为师。

董艳珍从小就继承了祖传的“足迹追踪学”，16 岁的时候就曾经协助过警察破案，在 18 年的时间里，有很多地方的公安局会聘请她担任刑侦技术员，而她主要擅长的也是足迹追踪和鉴别，依靠董艳珍的“足迹追踪术”破获的大小刑事案件超过了一千余件，而董艳珍也因为自己特殊的才能成为了家喻户晓的“民间女神探”。

还有，猫眼电影整合了 2015 年上半年的售票数据，报告根据用户购买电影票的习惯，结合用户在美团上的相关消费行为，发现了有意思的现象。数据显示，用户在购买电影票的同时，有 79%会进行餐饮消费，10%会选择唱 K、桌游、足疗等休闲活动，还有 11%会选择酒店消费，其中有 81%选择的是经济型酒店……

大数据预测世界杯真的很准吗

在 2014 年的巴西世界杯上，卫冕冠军西班牙连续两场失利，小组赛即遭淘汰，不仅让西班牙球迷伤心欲绝，让彩民损失不小，还顺便连累了众多预测世界杯的高人欲哭无泪。

这届世界杯在大数据火爆之后，不管是民间还是官方，都把大数据的概念运用到了世界杯预测上，但这些预测真的准吗？下面选取国内外主要的世界杯预测机构，对他们的预测方法进行简要的分析，看看谁的更准一些。

百度分析最传统

据验证，2014 年全国高考作文题目 18 卷中 12 卷的作文方向被百度大数据预测命中，被戏称“神预测”。因此，这次百度收集网上的综合数据，然后进行整理、分析，最终通过大规模机器学习等人工智能技术，开始预测世界杯。

百度预测世界杯的主要数据来源包括：百度搜索数据、球队基础数据、球员基础数据、赔率市场数据。百度大数据通过分析过去 5 年 987 支球队的 3.7 万场比赛数据，共涉及 29610 名球员，112,285,543 条相关数据，构建了足球赛事预测模型。据说为了验证模型是否准确，百度用 2010 年南非世界杯的淘汰赛数据进行了准确性验证，输入 2010 年世界杯期间的比赛、球队、球员等相关数据，由预测模型计算出淘汰赛比赛结果，与当时的比赛结果进行对比，准确率为 75%。

评：百度用的是传统统计分析，注重近期球队和球员表现，这种预测是至今为止在技术上最稳定的方法，但受意外因素（如天气、伤病、裁判等）影响较大。

德银推算最胡闹

德银根据各个球队的 FIFA 排名、历史战绩、球员构成和赌场赔率等因素，建立了量化分析模型，并根据复杂计算得到一份夺冠概率表格，从夺冠概率表格中挑选出了前 10 强，依据“轮流转周期”，由此排除了 2014 年巴西、意大利和西班牙夺冠的可能性，然后根据另一个

假设：强队会回来，即夺取过世界杯的强队，未来必然还会夺取世界杯或至少打入一次决赛。最后，本届英格兰队有 6 名队员来自利物浦，而正是在利物浦的球员最多的 1966 年，英格兰获得了历史上唯一一次世界杯冠军。同时，德银报告的主笔人承认自己是利物浦队的铁杆球迷，因此，最后确定英格兰将获得世界杯的冠军。

评：还好，德银报告主笔不是中国队的球迷！

高盛模型最神秘

高盛对世界杯决赛周 32 支国家队的胜算，有它自己的一套评估方法（命名为 Elo），在所有因素中分量最重。Elo 是高盛自设的动态系统，不断根据球队近绩更新评分和排名。

为此，分析师要收集多项数据，包括：世界各个国家足球队历史成绩数据库给出的各队排名得分；比赛中双方球队过去 10 场和 5 场比赛的进球数；比赛双方是不是巴西主场；比赛球队是不是美洲球队；还有以往各队在世界杯的进球数优于平时多少个。最后，他们把这几项数据按照一定的权重相加到一起，可以得出每一个球队在对阵另外某一个球队时平均会进多少个球。按照这样的方式，从小组赛一路到最后决赛，每一场比赛双方的进球数都可以期望一番，最后获得一个“最平均”的世界杯全程模拟结果。

评：投行一贯用神秘模型来忽悠投资者，Elo 模型就是高深黑洞，关键环节恕不奉告，至于准确与否，只有神知道。

严格地讲，以上几家世界杯预测都不能算“大数据分析”，只是传统的统计分析，虽然数据“大”，但并未融合多种因素综合考虑，可见在专业领域还是相信经典理论。

以下这些不靠谱的预测才是大数据：

霍金想法最娱乐

霍金收集了大量的数据，包括历史记录、温度、球场的海拔高度等，把所有数据都集中起来，分析你事先不知道的事情，或许能发现一些规律。它的原理不是传统的分析，更多是基于关系的一种预测。霍金 19 页的分析结果是关于如何提高英格兰队的夺冠几率的，但最后却抛出一个让英格兰球迷伤心的终极结论：个人更看好巴西队夺冠。霍金认为英格兰队首先需要在海拔 500 米以下的球场比赛，气温的提升会降低赢球可能，在巴西当地时间 15 时是最好的比赛时间。从球队自身来说，433 阵形无疑是夺冠的节奏，而且必须穿上红色战袍。提到点球大战，霍金认为助跑必须不少于三步，如果速度上不去，进球几率只有 58%。瞄准上角的点球有 84%的命中率，金发和秃头的球员射中的概率达到更高的 84%，前锋的进球概率超过 80%，中场与后卫递减。

评：霍金老爷爷最近几年很喜欢预测，还预测过世界将在两百年之后灭亡，这次娱乐世界杯一下，也是比黑洞更沾地气。当然，事后的结果证实，霍金老先生看好的巴西队早早出局，德国队获得了冠军。

科隆体育最繁琐

德国科隆体育学院根据复杂的计算机模拟测算得出的本届世界杯预测结果：科隆体育学院的格罗尔教授领导研究小组以自己设计的计算机模拟算式一共进行了 10 万次测算，综合考虑各队的世界排名、足彩赔率、市值、预选赛表现，还包括可能的伤病、战术、气候条件、主场优势因素。他们预测，巴西队与阿根廷队将争冠，卫冕冠军西班牙有可能止步小组赛，从西荷大战那个惊悚的 5 比 1 赛果，看来德国

人的模拟测算还是靠谱的。

评：德国人的严谨是出了名的，而且竟然没有预测德国队夺冠，对于西班牙却一语中的，最后德国队的夺冠让这个预测显得很不靠谱。

熊猫预测夭折了

世界杯开幕前，据媒体报道，中国保护大熊猫研究中心称将派出一到两岁的熊猫宝宝来预测世界杯。小组赛阶段，主办方会拿出三个竹筐代表主队的胜平负，熊猫宝宝则通过选择哪个筐里的食物来预测比赛结果。等到了淘汰赛，熊猫宝宝们还会通过爬树和赛跑来预测结果。前者是让熊猫爬上挂有一方球队国旗的树木来预测，后者则是两个熊猫宝宝分别穿上两队球衣，通过谁先跑到目的地来预测比赛结果。就在世界杯开赛之后，“熊猫预测世界杯”活动已经被取消。

评：本来要顶替章鱼保罗的国宝没了用武之地，国人还是缺乏点娱乐精神，借此机会宣传下大熊猫，有何不可，万一要是预测对了，那大熊猫基地岂不成了大师圣地，还愁旅游不火？

微软相信 Excel

微软必应大数据之前曾多次成功预测奥斯卡奖项、投票大选。微软的预测考虑过往比赛历史、主场客场、地理位置、草坪状况、天气及“群众智慧”等多种因素，还使用大量的公开数据——博彩市场、民意调查、社交媒体及其他在线数据，利用大数据分析来判断每场比赛的结果。据说这一切都是用 Excel 来完成的，我们权当其是软件推广策划吧。

微软：相信 Excel 是万能的，但预测足球估计是万万不能的，不过，人家说奥斯卡、大选都预测对了。

雅虎相信网络流言

雅虎用轻博客网站 Tumblr 的数据来估计每支国家队的优势，最终计算出最可能获胜的是巴西队。雅虎研究小组分析的前提是，Tumblr 上所有有关世界杯的讨论都具有一定价值。为了查明哪些国家将相互较量，小组会根据之前比赛的结果为每支队伍赋予优势值。针对每一次比赛，雅虎会利用名为泊松分布的不同参数的概率论来估计每一支队伍可能的进球数量。

评：雅虎相信的是目前最火的社交网络数据，据说可以预测传染病和犯罪现场。

当然，虽然很多人相信大数据能够帮助我们预测世界杯，也有不可预测派。美国的洛斯·阿拉莫斯国家实验室的三位统计物理学家曾经对大型体育比赛的赛况进行数据化分析，发现在棒球、曲棍球、篮球、橄榄球及足球五大项目中，足球比赛是其中最具悬念，赛果最具不确定性的，弱旅战胜强队的概率居高不下，即使使用科学方法也未能得到准确的预测。

说实话，作为统计专业人士，对足球预测不敢太相信，体育比赛确实可以预测，足球也不例外，但足球项目影响因素太多，特别是世界杯足球比赛，相对场次不多、间隔周期太长，致使数据量很小，比赛中又有太多的主观因素（比如裁判），有时候这种比赛的预测和算命没什么差别。

如果要问为何总有人预测正确？正如一家报纸所说，每届世界杯都会有无数的“保罗”，大部分都在前几次猜测失败后从媒体视线中消失。贝利也不是真正的乌鸦嘴，只不过他预测成功的时候没有后续报道。预测大师都是这样炼成的！

数据分析的五个基础

数据分析这种事情，每个人都可以做，并不分高低贵贱和专业学识，只是，不同的人分析出来的结果会有不同。婴儿在咿呀学语的时候就已经在分析父母的表情变化，以此来决定自己应该怎样撒娇啼哭才能获得最大的好处。在每一家公司里，老板、中层和基层的员工，也包括门口的保安、打扫卫生的阿姨，都在进行着自己的分析，只是，每个人的目标会有差异，每个人分析的角度也会不同，至于分析能力，老板并不一定比保安要高明多少。

一般的分析，主要可以分为描述性分析、探测性分析和因果性分析三种，三种分析有时候是独立的，有时候是密切结合在一起，但大多数企业的分析都会是逐步展开的。我们一般要先进行描述性的分析，然后根据描述的结果进行探测性分析，探测完成以后会开展因果分析，三家共同构成了完整的经营分析。

我们可以把描述性分析比喻为考古。我们首先要做的是企业的经营行为的描述，比如用户数量多少、业绩如何、客户的评价怎样，也包括报告里的增长或减少程度，还可能要有公司产品销售的结构比例及成本费用情况等。总之，描述就是在进行古墓考古，要把古墓中的一切说清楚，到底挖掘出了多少宝贝，是否有盗洞等。当然，这工作只是考古的第一步，我们接下来需要弄明白的就是这个墓到底是什么年代的，墓主人是谁？等等。那我们就要用到探测性分析。

探测性分析往往是建立在描述性分析之上的，没有清晰的描述，就很难去探测。探测就是要发现问题，比如，通过对公司情况的描述，我们进行对比及评估，可以发现公司在经营中存在哪些问题，主要问

题是什么，公司用户数增长或下降是否严重，等等。探测性分析类似电脑游戏“扫雷”，我们知道这块地方里有地雷，但不知道到底有多少地雷，也不知道地雷到底在哪，通过我们的分析，可以找出一定的规律，发现地雷的大概位置，并逐步地排除掉。总之，探测性分析是用来发现问题和指向问题的，而寻找问题和症结所在正是企业经营管理中非常重要的工作内容。

发现问题之后就需要弄明白问题是怎么产生的，到底因何而来，目的是要解决问题。我们始终要清楚，分析的目标是解决问题，不是为了分析而分析，不能发现问题解决问题的分析是劳民伤财和自欺欺人。因此，探测性分析之后往往就是因果分析。

因果分析是要找到问题的成因，并经过严密的推理和实证确定，此后就是针对原因想出解决方案，把影响经营的因素解决掉，让企业的经营回归正常轨道，或者更上一层楼。

假设，我们是公安部门的侦查员，有人报警在某宾馆内发现凶杀案，当我们出警去到现场时，就需要对现场进行详细描述和认真探测，目的就是发现蛛丝马迹以便锁定凶手进而破案。在破案过程中，我们要针对遇害人可能接触到的对象进行一一排查，从逻辑上去查寻凶手的痕迹，并分析凶手的杀人动机，即便抓获了凶手，也需要弄清楚凶手的杀人缘由和过程，形成证据链，才能将其移送法院量刑定罪。

但是，我们普通人往往在分析的过程中会犯下错误，简单主观地根据自己的常识进行推断，比如一定是某某某做的，因为其有前科，或者是某某某无疑，因为其和被害人曾有纠纷，这种判断有一定的合理性，但对分析却是有害的。分析不能简单粗暴，更不可偏听偏信，需要全面深入，公平公正。

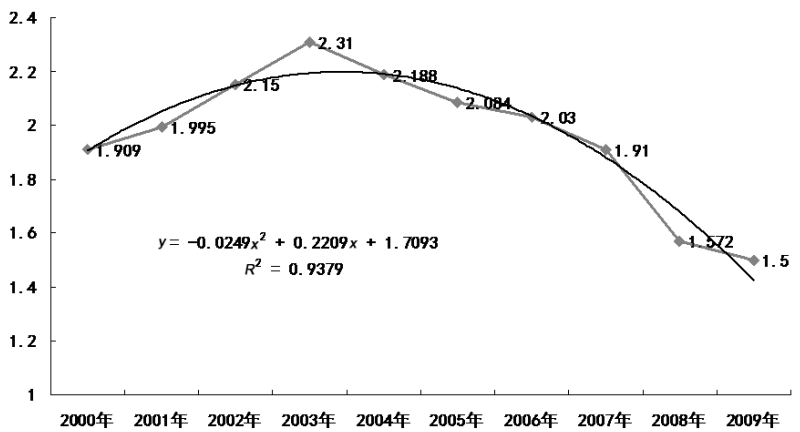
这个时候，我们就会用到所谓的“MECE”分析法，也就是麦肯锡方法里面重点的分析思维。

带着思维去分析

很多人认为，数据分析就是数学，通过数学计算就可以找出数据之间的关系，从而发现数据背后的真相，如此就是完美的数据分析。

其实，这是完美主义的思维，也是不切实际的想法。数学计算只是数据分析的工具，也是医生手里的血压计、战士手中的枪，最多只是实现分析目标的必备手段而已。任何的分析，主体都是进行分析的人，数学上的分析结果只能作为进行判断的辅助，归根结底是要靠人脑结合具体场景来得到结论。

有一个案例，那就是中国电信业的增加值占国家 GDP 的比重，这个数据在 2005 年之后就一直在下降，有的年份下降的程度还非常大。如下图所示。



在这张图上，我们使用了最简单的趋势线，来描述电信业增加值占 GPD 比重的趋势，从数学角度看，这条趋势线很好地描述了发展趋

势轨迹，其 R^2 几乎等于 1，非常完美的分析。

我们画出一条趋势线，目的显然不只是为了描述趋势，更是为了进行预测。就图上所示，我们可以用这个方程预测 2010 年的数据，显然，按照这个趋势发展，2010 年的数字会很惨，大概只能维持在 1.2 左右。这样的情况会发生吗？

即便不站在目前这个时点向后看，就是站在 2009 年向前看，这样的结果也不会发生。因为，正是在 2008 年，中国政府发放了 3G 牌照，几家电信运营商在 2008 ~ 2009 年掀起了一轮建设高潮，而设备商等相关领域都是大大的受益者，电信业的增加值不会再快速下滑了。由此证明，图上的分析是错误的，虽然在数学上几乎完美无瑕，可实际上却在误导我们的结论。

其实，这张图我们可以看出更多信息，比如，政府为何要在 2008 年发放 3G 牌照？

所有的数据分析人士都应该切记，任何的数据分析都不是在做纯粹的数学题，而是结合具体场景和一定的背景条件而产生的应用题。很多在数学上看似正确的分析结论却经不起实践的检验。当然，这里并不是说数学不重要，而是说，数学是从社会现象中抽象出来的普遍规律，是我们学习和工作的参考，而不是分析的全部。

除此之外，我们还必须了解一些必要的假设，离开了假设，公理或定理往往都会变成错误，而分析更不例外，特别是在一些预测的环节。

很多人以为，短期预测比长期预测更简单，因为数据变量更容易控制，可应用的分析方法也更多。但在实际工作中，我们却会发现，短期的预测最难做，而长期的预测却可以有很多发挥的空间。

究其原因，短期预测容易被验证。比如，你预测下个月或者明年的公司业绩，这些结果将在不久的将来被验证，对错可循，作为分析人士，压力可想而知。但是，如果预测公司十年后的发展情况，分析的压力会小很多，毕竟，十年之后的情况变化会更大，公司的组织结构和领导都会发生很大的变化，甚至战略都彻底变更了，到那个时候就没有人去较真预测的成败。比如，霍金预测地球将在两百年之后灭亡，谁又有能力去验证是否真实可信呢？

还有，我们进行的很多预测分析，都是建立在一定的假设基础上，比如公司业务不变化、公司不会被重组、技术革新不会突然出现颠覆性结果，甚至还需要考虑很多可能突然出现的政治干预。即便在分析技术上，也会有很多假设前提，否则，很多模型是没有办法得到应用的。离开了这些假设，任何的分析预测都毫无意义。

诸葛亮曾经在《隆中对》中对当时的中国格局进行了准确的分析，并且预测到了三分天下的格局，并由此提出了三国鼎力最终统一的路线图，可是这样的宏图大略也遭遇了意外因素，由于关羽的大意失荆州，导致整体战略无法继续实施，夷陵之战更是让蜀汉彻底伤了元气，最终导致诸葛孔明出师未捷身先死。孔明尚且如此，谁又能保证我们普通人的预测都会变成现实呢？

最为可笑的是，某中国非常有影响力的证券机构曾经出版了一本对未来 50 年的中国房地产市场的预测报告，论述逻辑严密，使用方法高级，涉及的因素全面，可是就是这样一份报告，如果我们深入分析，却有着难以自圆其说的漏洞。因为，这本报告在序言的最下面列有一行不为人注意的小字，小到一般人看不到，那串文字清楚地告诉读者：因为中国房地产市场发展还不成熟，数据积累太少，因此，我们只选

取了 2009 年和 2010 年两年的数据进行了计算。天啊！要知道，这份报告预测的可是未来 50 年的情况，使用的数据竟然只有两年。任何对数学稍有研究的人都会知道，两个点的连线，怎么可能确定出它符合什么模型曲线呢？

先把研究对象的方方面面吃透

很多为人父母的人大概都有与我一样的感受。记得我家宝宝刚出生不久的时候，都是妈妈在家带孩子。作为经常出差的我，难得有时间在家陪孩子玩耍。在孩子几个月大的时候，我就发现有很多奇怪的现象。

比如，我们一家经常利用孩子睡觉的时候吃饭，而小孩子往往会在出人意料的很短时间醒来，妈妈便只能中断吃饭去带孩子。有很多次，我们在餐厅吃饭，孩子在卧室睡觉。突然，妈妈就放下碗筷，嘴里说“宝宝醒了”，迅速跑到卧室，结果，孩子正好刚刚醒过来，还没来得及哭出声来，看看妈妈甜甜的微笑，宝宝也就会安静地继续睡下去。

这样的事情发生了多次，而我却从来没有任何的感觉，所以就觉得妈妈有特异功能。但宝宝妈妈却说，自己也只是感觉孩子醒了，并没有听到孩子的哭声或其他动静。也许，所谓的母子连心，正是如此吧！

我们从数据分析的角度来看待这件事情，就会发现所谓的第六感也是有道理的。因为妈妈天天和宝宝在一起，对宝宝极端关注，而且，天长日久母子连心，宝宝的作息习惯和行为方式已经刻画在妈妈脑子里，这些基本要素就促成了妈妈拥有其他人所不可能有的分析特质，

由此才能准确地感应到孩子的行为。

作为数据分析人士，我们首要做的并非是掌握太多高级的分析方法，而是要对分析的对象充分了解，十分关注。对于任何毫不知晓的领域，发表任何看法都是草率的，即便是博士院士，在其无知的领域也只能是无知，其能力不会比普通人多哪怕一点点。做企业分析的人，就必须花力气最大限度地了解企业的过去和现在、业务和产品、销售与服务、人员和管理，当你烂熟于心的时候，就可能具备了“第六感”，也许拍拍脑门都可以做出恰当的决策。反之，对企业的情况一知半解，即便掌握再高超的分析方法，也不会有用武之地，强行使用，可能得到南辕北辙的结论。

我有一位朋友，科班出身的数据分析专业高手，毕业后就进入到一家大公司工作。她发现公司内部的分析工作非常原始，基本都是柱状图或者摊大饼，分析的方法更是只有简单的对比与排序。于是，在工作刚刚两个多月的时候就废寝忘食地根据能拿到的公司信息数据进行了高科技分析，认真加工之后将十几页 PPT 发送到了部门领导的邮箱中。

可是，邮件发出便石沉大海，两周后，实在忍耐不住好奇的这位朋友利用一个机会向领导打听“观后感”，得到的回答却是“毫无用处，没有价值”。这位朋友痛哭之余向我求助。我问她，你对公司了解吗？她认真地说，自己入职培训非常认真，工作这两个月也是多多用心，对公司情况应该掌握得差不多。后来，我又问她，你拿到的数据都是真实的吗？她告诉我，那些数据都来自公司原有的 PPT 材料，有些不全的也是向同事问来的，会不准确吗？

我和她进行了分析：入职培训再认真，两个月的工作再用心，对

于一个员工数十万人年收入数千亿的巨型企业来说都是完全不够的，甚至连很多皮毛都没摸到，这个时候进行涉及全局性的分析注定不会有价值。至于收集到的数据，企业往往会根据不同的需求产生不同的口径数字，将这些数字（特别是已经在 PPT 中简单化的数字）拼凑到一起做分析，更是犯了分析的大忌，何况这些数据资料也存在被无意“造假”的嫌疑。

数据分析需要扎实的基本功，也需要扎实的分析行功，把可能涉及的方方面面都研究明白，把事情的本来面目都搞清楚，甚至还要清晰业务的来龙去脉与隐情内幕，否则，分析结果的价值就毫无保障。

快，才能解决现实问题

天下武功，唯快不破。很多人信奉这样的武林秘诀，小李飞刀所向无敌。在现实的市场竞争中，行动快的企业也会比行动慢的企业有更强的生命力，所谓快鱼吃慢鱼。

数据分析也是一样，如果分析的过程太过缓慢，不管你分析的结果有多正确，都可能因为时效性的问题而变得一文不值。

在大学或者研究所，为了深入研究一个课题，往往需要做严密的规划、详细的论证，仅仅取样或者组建数据库就要花费数月或数年的功夫，加上各种认真的分析过程和验收程序，一个项目下来动辄两三年，甚至要十数年。这种情况对于研究机构无可厚非，因为，对于这些深度研究，往往时间不是问题，结果才是问题。

可是，企业的经营分析却等不了。一个企业遇到了经营困难，或者客户在流失，或者产品销售不畅，或者客户服务评价在降低，我们要找到原因所在，就必须严格遵守时间限制，用最快的速度将分析完

成，拖延几个月甚至几天都可能变得毫无价值。对于一些激烈的商家竞争，胜负甚至只在一念之间，这时候的分析更是要分秒必争，甚至要转瞬完成。这不是苛刻，而是现实的挑战，也是必须要完成的任务。

地震预报一直是世界难题，至今都没有大的进展，虽然很多机构或者组织也曾经有过成功预报地震的先例，比如已经载入地震预报辉煌历程的海城地震，可最后总是成为“偶然”碰到的好运，不久之后的唐山大地震彻底让沉醉在地震预测成功喜悦中的人们惊醒。

有些人认为是我们还没有掌握到足够多的数据信息，有些人觉得是我们使用的分析方法走错了方向，还有人认为人类根本就无法做到对地震的预报。但不管怎样，地震预报肯定是越快越好，如果能提前 30 秒，都可能帮助到很多人，如果能提前 30 分钟，那几乎可以将灾害降低到最小。

但是，至今，对于地震信息的分析依然处在初级阶段，我们甚至都无法确定我们真的已经在开始收集与地震预报有关的数据。也就是说，数据分析的结果确实要快，还必须保证质量，“萝卜快了不洗泥”，也是不可以的。

地震的预报有一个前提条件，那就是可靠性。如果我们对地震信息的分析有足够把握，而且十次预报会有九次或八次是正确的，我们就可以有充分的信心进行预报，从而实现减灾的梦想。但是，如果我们的分析结果只可能有一两次是正确的，那谁敢去轻易地预报呢？要知道，预报错误所带来的损失不亚于一次小型地震，更不要说多次预报错误会带来“狼来了”的灾难性后果。

客户的流失预警也是一样。对于集中出现的客户流失，公司需要做出最快的分析，找到原因和解决方案，否则很可能带来灾难性的经

营后果，这种分析刻不容缓。但是，客户流失的监测分析却也和地震预报一样，必须保证一定的准确度。如果辛辛苦苦建立起来的客户流失预警模型，提取出 100 个据说有离网倾向的客户，结果，在实际流失挽留的过程中发现其中只有十几个是真正有想法的，其他根本就是忠诚客户。这样的情况发生一两次，就不会再有人愿意使用这样的分析工具了。

理解管理者的意图

如果一个分析者从事的是自己爱好的科学研究，那无论怎样去分析或得出怎样的结论，都没有关系，只要对自己负责就可以。可是，如果这个人是一家企业的市场分析人员，或者要对某位领导者与委托人提供意见，那就需要认真考虑，三思而后行。

企业都是有经营目标的，企业的管理者也有自己的想法，任何的企业分析都需要充分结合这样的必然前提。理解管理者的意图，为特定的目标服务，每位分析人员都会面临这样的境况，只有理论联系实际做出合情合理的分析，才会让分析变得有价值。

我们可以有一个可能并不很严谨的比喻，如果你是红军长征队伍里的一员，也可能只是一个团里的参谋。最高决策层已经制定了北上方针，准备去陕北开辟根据地，让作为团参谋的你出谋划策，制定下一步的行动方案。这个时候，你要做的只能是认真分析敌我形势和我军面临的战场状况，做好飞夺泸定桥或者穿越草地雪山的应对策略。如果你却非要去分析队伍南下的好处或者应该去攻取上海，那一定是不正常的。不是你不可以有这样的想法，而是你作为团参谋不能有这样的想法，离开目标和管理需要的分析都会变得毫无价值。

结构化思维与分析的类别

一般来说，要做好数据分析，需要从思维、方法、模型和解读四个方面来行动。思维是最高阶，也是做好数据分析的基础。和很多人的想法并不一致，做好数据分析并不是首先要强化自己的 Excel 或者 SPSS 操作能力，甚至也不是什么统计学知识，而是在于锻炼自己的思维能力。

方法比思维低一个层次。所谓的方法，主要是将我们已经具备的思维能力转化为具体的行为，通过适当的方式方法来解决具体的问题。思维是解决问题的灵魂，而方法是解决问题的肉身，是执行者，是行动派。

什么是结构化的思维（MECE）

学杂费应该怎么交？有这样一道题目，是小学二年级的寒假作业题。不要小看这样的题目，孩子的题往往是用来考家长的。

题目是这样的：某老师要向孩子收 20 元的学杂费，当小孩子告诉家里人后，孩子的母亲只在家里找到了 2 张 5 元、5 张 2 元和 10 张 1 元的纸币（注：2 元的纸币已经退出了人民币的序列）。问，这家有多少种交钱的方法？

不服的人们可以在空白的地方试着演算下，看看你用了多长时间得到正确答案。

先给出结果：10 种！

怎么得出来的呢？其实很简单，千万不要用什么高级的算法，更不要用什么排列组合，要知道这只是二年级的题目啊！

我们可以用最原始的凑数方法，列出一张表格，既然是 *XYZ* 三个变量，那就一定要固定住一个变量，然后让另外的两个变量构成唯一的组合，最后累加到一起，就是我们要的结果。

比如我们先考虑 5 元，一共最多只有三种情况可选，一种是用 2 张，一种是用 1 张，一种是用 0 张，接下来，我们可以绘制出这样的表格。

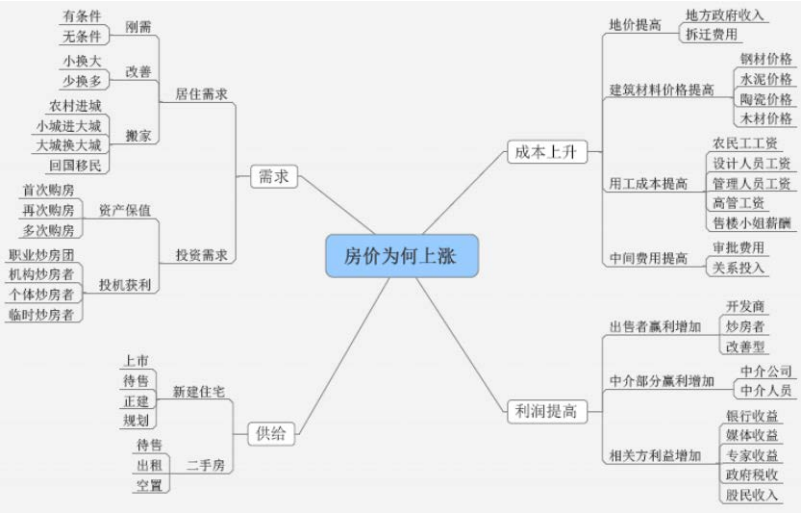
5*X*	2*Y*	1*Z*
2	5	0
	4	2
	3	4
	2	6
	1	8
	0	10
1	5	5
	4	7
	3	9
0	5	10

当结果出来的时候，我们就知道这道题目做对了，因为我们遵循了 MECE 的法则，也就是做到了“不遗漏、不重复”。

我们在分析问题的时候一定要坚持这样的原则，不要遗漏掉任何因素，也要将因素之间的相互重复或影响去除，比如回归分析时的多重共线性。

利用这样的方法，我们可以借助思维导图这样的软件来清晰地描绘出问题的分析思路，比如房价为什么上涨？有人算了一笔账，有人 2005 年以总价 115 万元买了世纪公园的房子，至 2015 年，成交价 1050 万元，涨价 935 万元，一年涨 93.5 万元，一个月涨 77916 元，每天涨 2597 元；每小时涨 108 元，平均每分钟涨 1.8 元。所以就这样每分钟

一块八、一块八、一块八……日夜不停数了十年。



以下是结合网络上的讨论，列出的可能的中国房地产市场火爆的原因，有兴趣的读者可以试着用结构化的思维进行分类归总，通过思维导图的方式梳理清晰。

（1）福利房制度取消，将全体国民推向市场化的房地产市场

（2）分税制的推行出现了富中央、穷地方现象，导致地方出现土地财政。

（3）学习香港的超级地租模式。政府通过房产超级地租来支持国内基建。

（4）加入 WTO 后，外贸出口发展吸引更多农民工进城，导致住房需求暴涨。

（5）人民币在加入 WTO 后外汇储备增多，推动人民币升值，国外资金通过投资人民币资产获利。

（6）国内许多资产项目不开放，导致国内私人资金可投资项目不

多（理财、股票、房产），大都涌入房地产行业。

（7）贪腐，“房叔、房奶、房爷爷”，大量的人将资产封存在正在价格上涨的房子上，特别是贪官污吏。

（8）20 世纪 70 年代末开始的独生子女一代结婚，往往可以获得家庭的倾囊购房支持，对高房价有承受能力。

（9）全社会集体的看涨预期导致的心理推动。

（10）中介出现后的推波助澜，不断忽悠市场不停地转买转卖，价格不断被炒高翻倍。

（11）取消单位宿舍，单位廉租房房改后，导致廉租房减少，住房需求增加。

（12）中国老太太和美国老太太不同的成功故事。

（13）土地流转受到限制，凡是农业用地一律不许进入商业市场流通，虽然守住耕地红线，实际上却卡住土地供应，推高房价。

（14）中国广义货币 M2 从 2002 年前的 22 万亿元到现在的 120 万亿元。

（15）央企入市，国务院整合央企、银行联手哄抬地价，多块地王被央企拿下。

（16）过去十年对房价的放纵，地方政府从未真正调控，越调越高，所有号称调控都是演戏。

（17）国家资源分配不均，一线城市占据太多资源，导致平民只能尽可能地涌进去。城市比农村资源更多，导致有钱的农民希望进城买房。

（18）钉子户的狮子大开口，拆迁费用上涨，既得利益者和体制相勾结，瓜分拆迁费。

（19）国企改制，中国人的社会等级弱化，农村人想进入城市成为市民不再麻烦。

（20）大学扩招，使得平民子女享受到更多高等教育的机会，这些人不愿意回到乡村，扩大了购买力。

（21）农村并村、并校，使得大量的农村人口不得不跟随孩子进城买房。

人脑在大数据时代并没有过时

2015 年 3 月，一条谷歌制造的机器狗，在韩国大战围棋高手李世石，这个只有东方人才玩的棋类游戏，怎么就成了西方科技代表的谷歌的“眼中钉”，恐怕不仅仅是围棋代表了人类智慧的高峰那么简单。

谷歌的机器狗开头连胜三场，路数大胆可是都取得了胜利，第四盘被李世石完胜，在最后一盘双方几乎打成平手，最后是机器险胜。虽然机器赢了，可很多人将这场围棋游戏比赛看成了机器和人的对决，上升到了人类生存还是毁灭的高度，估计让谷歌自己都哭笑不得。

机器下棋获胜有很多棋外因素，并非完全是智能的力量

游戏就是游戏，再高级也是游戏，即便是机器狗全胜，也不代表什么，因为机器狗本来就是人造的，其程序也是程序员们写进去的，至于说什么机器自己学习，也只是很多人根据“深度学习”这个翻译的词汇臆想出来的。

计算机机器狗在与人对弈的时候，确实有很多优势，最大的优势就是，计算机没有感情，不会紧张，也没有胜负压力，更不会疲劳，

除非断点或者死机。人就不同，连续几个小时的高体力思考是对人脑极限的考验。一般来说，一场高强度的围棋比赛之后，很多棋手都会体重减轻几斤甚至十几斤，机器估计仅仅是稍许发热而已。

虽然我们可以说谷歌的机器人程序并非针对围棋而来，更不是为战胜李世石开发的，但这套程序适合围棋是显而易见的，而在对阵李世石之前在围棋上做过很多磨合，针对李世石的特点做了大量的研究也应该是事实。

从这个意义上讲，计算机（实际上是控制计算的那些程序员们）对李世石可以说是了如指掌，而李世石对这台电脑的了解却知之甚少，甚至谷歌之前的对局都是保密的，计算机已经知己知彼，可李世石却是盲目仓促上阵。当年，卡斯帕罗夫也是遇到了这个问题。当然，看李世石的布局，也是想到了这一层，所以出招有点怪，直接赢得了中盘优势，可后来还是因为准备不足而吃了大亏。

从机器所显示出来的水平看，已经达到了一流高手的水准，可距离超一流还不到位，至少是人与机器对抗时完全有取胜的机会。机器并不能和任何人下，而是针对性地做了开发，这也是要找在棋坛征战多年的老将李世石而非其他刚刚出道不久的高手的原因。

计算机的人工智能没有其名字所呈现得那么神奇

现在计算机最大的问题是三个：一是计算机的理论模型从来没有变化，这种结构永远不会超越人类；二是暴力算法的解题逻辑一直没有改变过；三是计算机是数字化的而不是连续世界，不可避免地出现断点，对于绝大多数问题采用的都是最大近似的方法，比如，圆周率就是用 3.14159265358979323846264338327950288419716939937510582

09749445923078164062862089986280 34825 34211 7067982148 08651 32823 06647 09384……

对于围棋来说，现在的计算机下棋也是基于对围棋棋谱的学习，人类曾经达到的高度就是围棋所能达到的最高高度，因为计算机自己还不会创造，也没有自我意识，数据所能表达出来的东西最多只能和数据质量一样好，不可能超越。如果哪一天，计算机可以去创造性地下棋，完全不顾及以往的棋谱，那才是真正的智能。

至少这些问题得不到解决，即便是在围棋这样的游戏层面，计算机也不敢说一定会赢，因为围棋的变化太多，是现在的计算机无法穷尽的。只有到了计算机可以穷尽围棋变化的时候，就可以完全控制局面，不管怎样下，棋手都不再有胜利的可能。

与围棋的对阵确实可以换来高的知名度，特别是在东亚地区，可这种即便能战胜围棋的谷歌人工智会还会在哪些领域有这样的超能力展示，我们还很难说。按照李开复的说法，未来像保姆、记者、中介等助理性质的职业都会被机器人替代，可这些到底与下棋获胜有多大关联呢？计算机仍然没有解决自我意识的问题，只能依靠固有设计来做事，离真正的人工智能还相差太远，距离造福人类的更好地应用也有很长的路要走。

我是歌手让阿里云“小 Ai”初试锋芒

在 2016 年的 4 月，“我是歌手”总决赛现场，阿里云人工智能程序“小 Ai”对这场比赛的结果跟踪做出预测，通过大量数据的收集处理推断，成功预测了李玟获得冠军。

据称，“小 Ai”主要基于神经网络、社会计算（Socialcomputing）、

情绪感知，善于洞察本质和实时预测，并能理解人类情感，还具有强大的计算和机器学习能力，能够不断自我进化。首先，小 Ai 需要积累一首歌曲的下载量、点评量这些可以判断歌曲受欢迎程度的数据，以及歌曲本身音频特征和谱曲音乐的关联因素。接下来，运行在阿里云大数据平台上的爬虫系统、情绪分析系统和现场效果采集系统协同工作，预判最终结果。爬虫系统是通过一定的规则，自动抓取互联网上的评论变化，其数据来源主要是新浪微博等，并以此形成大量的数据供给第二个系统。情绪分析系统会根据抓取回来的评论进行实时文本分析，以便分析出现场 500 位听众评审对歌手的评价。然后对现场音频数据和舞台效果进行实时采集，并做出判断，以此调节判断歌手夺冠的几率算法的权重。

从结果来看，第一轮，小 Ai 的判断依据较少，所以对选手获得冠军的概率预测与结果相差较大，但成功预测出淘汰选手；第二轮，小 Ai 成功预测了对决名单，不过出场次序略有错误；第三轮，小 Ai 顺利预测出了冠军李玟，但亚军和季军的顺序预测与结果相反。

人工智能“小Ai”的预测因子分析

从分析的角度来看，在方法分类上，一般会分成定性分析和定量分析。简单地说，定性研究主要是回答“为什么”的问题，我们应用定性研究进行“认识、发现、判断、了解”，而不能使用它进行“测量、监控、估计、预测”，这方面的问题应当用定量研究的方法去解决。定性研究的方法一般包括焦点小组座谈会和深度访谈。

定性分析就是对研究对象进行“质”的方面的分析，运用归纳和演绎、分析与综合及抽象与概括等方法，对获得的各种材料进行思维加工，从而能去粗取精、去伪存真、由此及彼、由表及里，达到认识事物本质、揭示内在规律的作用。定量分析是对社会现象的数量特征、数量关系与数量变化的分析，功能在于揭示和描述社会现象的相互作用和发展趋势。

从分析的内容看，定性分析与定量分析应该是统一的，相互补充的；定性分析是定量分析的基本前提，没有定性的定量是一种盲目的、毫无价值的定量；定量分析使定性分析更加科学、准确，它可以促使定性分析得出广泛而深入的结论。

定量分析是依据统计数据，建立数学模型，并用数学模型计算出分析对象的各项指标及其数值的一种方法。定性分析则是主要凭分析者的直觉、经验，凭分析对象过去和现在的延续状况及最新的信息资料，对分析对象的性质、特点、发展变化规律做出判断的一种方法。

相比而言，前一种方法更加科学，但需要较高深的数学知识，而后一种方法虽然较为粗糙，但在数据资料不够充分或分析者数学基础较为薄弱时比较适用，更适合于一般的投资者与经济工作者。但是必须指出，两种分析方法对数学知识的要求虽然有高有低，但并不能就此把定性分析与定量分析截然划分开来。

事实上，现代定性分析方法同样要采用数学工具进行计算，而定量分析则必须建立在定性预测基础上，二者相辅相成，定性是定量的依据，定量是定性的具体化，二者结合起来灵活运用才能取得最佳效果。不同的分析方法各有其不同的特点与性能，但是都具有一个共同之处，即它们一般都是通过比较对照来分析问题和说明问题的，正是通过对各种指标的比较或不同时期同一指标的对照才反映出数量的多少、质量的优劣、效率的高低、消耗的大小、发展速度的快慢等，才能为鉴别、判断提供确凿有据的信息。

在大数据时代，很多人觉得定性分析已经无用，我们依靠强大的计算机技术可以通过数量解决一切问题，但计算机至今还不是人脑，大数据信息再全面也很难有足够的“智慧”，更无法参透各国文字之间的玄妙。

以下文字是来自网络上对某国外交语言的解释，大数据懂吗？

（1）亲切友好的交谈——字面意思；

（2）坦率交谈——分歧很大，无法沟通；

（3）交换了意见——会谈各说各的，没有达成协议；

（4）充分交换了意见——双方无法达成协议，吵得厉害；

（5）增进了双方的了解——双方分歧很大；

（6）会谈是有益的——双方目标暂时相距甚远，能坐下来谈就很好；

（7）我们持保留态度——我们拒绝同意；

（8）尊重——不完全同意；

（9）赞赏——不尽同意；

（10）遗憾——不满；

（11）不愉快——激烈的冲突；

（12）表示极大的愤慨——现在我拿你没办法；

（13）严重关切——可能要干预；

（14）不能置之不理——即将干涉；

（15）保留做出进一步反应的权利——我们将报复；

（16）我们将重新考虑这一问题的立场——我们已经改变了原来的（友好）政策；

（17）拭目以待——最后警告；

（18）请于×月×日前予以答复——×月×日后我们两国可能处于非和平状态；

（19）由此引起的后果将由××负责——可能的话我国将诉诸武力（这也可能是虚张声势的俗语）；

（20）这是我们万万不能容忍的——战争在即；

（21）这是不友好的行动——这是敌视我们的行动；可能引起战争的行动；

（22）是可忍孰不可忍——不打算忍了，要动手了。

（23）悬崖勒马——想被××吗？

（24）勿谓言之不预也——准备棺材吧。

如果以上这个对于我们普通人意义不大，那么下面这些图形要记好，一旦在你家附近的墙壁上发现，你可要小心了，因为据说这就是小偷行动的暗号。

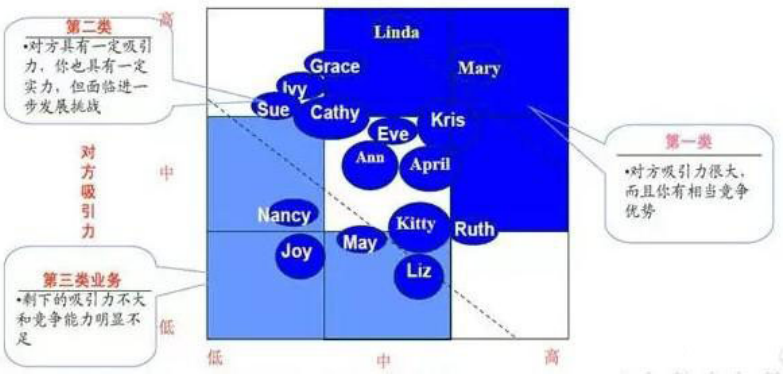
相亲是感性的还是理性的

据说，有位在金融行业从事分析师工作的美女多年没有找到对象，原因之一就是太理性。这位美女自己研发了一个模型，一共设计了 16 个维度，每次见到相亲的男性，总会将注意力放在收集相关的数据之上，然后根据得到的信息进行计算，比较评估，最终确定交往的可能性与策略，最终的结果就是从来没成功过一个。

按照一位高人的数据分析思路，选择对象是这样的模型：

1．选择谁

回答这个问题既要考虑对方的吸引力，也要考虑自身的竞争实力。因此 GE 矩阵模型是不二的选择。下图是某人的 GE 矩阵分析结果，从中可以看出，Mary 和 Linda 是需争取的主要对象。



注：以上人名为虚拟，圆圈大小表示投入成本（如时间成本和物质成本等）。

2．为什么选择他/她

换句话说，从哪些方面评价对方的吸引力和自身的竞争实力？可以考虑 7S 模型。

（1）Sharedvalue（共同的价值观）：体现在对生活、金钱、后代、亲人等重要问题的看法上。例如享乐者和节约者如果结合，则常常会因为钱该花不该花的问题而争吵不休。

（2）Structure（结构）：也就是对方是如何平衡家庭、工作、生活、亲人、朋友等多种关系的，是否能实现多种关系结构的和谐。

（3）Satisfaction（满意）：不同的人选择标准不同，比如外表、性格、家庭出身等，当对方的条件达到甚至超过你的标准时，你就会觉得这是自己的“菜”，感到满意。

（4）Sense（感觉）：也就是我们常说的“来电”。

（5）Style（风格）：体现在饮食、兴趣、爱好、习性等方面。

（6）Sex（性）：你懂的。

（7）Skill（技能）：引起对方注意的独特能力，比如沟通的技能、

生存的技能等。

在利用 GE 矩阵模型进行选择时，可以从这 7 个维度考虑，根据自身偏好，为这 7 个维度设置权重，并为自己和对方打分，从而得到吸引力和竞争实力的具体得分。

大数据主义者认为，所有决策，都应当逐渐摒弃经验与直觉，并且加大对数据分析的倚重。相对于全人工决策，科学的决策能给人们提供可预见的事物发展规律，不仅让结果变得更加科学、客观，在一定程度上也减轻了决策者所承受的巨大精神压力。实际上，并不一定如此！

在百科里，我们找到一段非常精彩的关于感性与理性的分析，全文复制在这里，给大家共享。

> “感性”：是人对大自然最本能的直觉，也是大自然给予人的生存能力中最强大的武器。这种直觉往往高于理性，但是在理性渐渐占据人类的思维之时，感性的敏锐精准度却被慢慢地消钝，对“直觉”使用的荒废，使得人们逐渐地忽略了其重要性而使其更趋退化。人深层的感性是超越时空的能量，是对时刻改变着的世界最本能、最直接且最精确的反应，它没有容量限制而又无限延展。只有真正深度的感性能够打破常规和局限，带领理性跃入局限之外的世界，而理性是在感性的认识之后逐渐形成，由“准确而却模糊”到“从属而却相对清晰”的过程，也就是感性到理性的过程。理性因其清晰而让人更易于把握，感性因其不确定而令常人无法控制，感性一时被蒙昧着的人们下降为低于理性的东西。然而人们却不知道在理性的过程中始终是“无意识里的感性”在推动着理性的运行使其日趋完善。理性就像感性无意间造出来

> 的玩具（工具），是感性用来探索自然的最好工具，而每次的突破和超越却依然必须由感性来完成——灵感。人们总认为灵感是基于理性，其实灵感是生命最本源的东西，在生命之初是仅有灵性而无理性的，由灵性而至理性，再在已发掘出的理性之上获得更高的灵性，从而获得新一层次的理性，循环往复，螺旋上升，但到了如今，人们却忘记了最本源的东西，并将人本身最宝贵的东西——直觉给抛弃不用。而真正能改变这个世界的那些人都是感性思维异常发达的人，灵感和精力异常充沛的人，举目望去，古往今来大凡如此。“理性”是“感性”有益部分的产物，是对“感性”的合理归纳和总结。而无益的部分则成了人们弃之的东西，如冲动、失控、盲目和执迷等。如果将“感性”比作“探索”，则“理性”是探索出来的规律；如果“感性”是只笔，则“理性”是这只笔画出来的作品。“感性”不能被“理性”取代正如人脑不可能被电脑取代。情之不定，皆因浮于表面未达深层，而刻骨铭心的情则已无力可变。“感性”始终充满了人类饱满的激情，而“理性”则是过滤了激情之后的冷静和思索方式。“感性”激发出艺术的人生，而“理性”则摈弃了情绪的干扰。理性往往借由高智慧的人而出，因为高智慧的人的感性思维往往比常人来得丰沛而旺盛，也由此创造出了更辉煌的理性产物，科学家、哲学家、艺术家……无不如此。

就相亲来说，国防科学技术大学吴孟达教授《数学建模》中也有关于真正理性的分析：

假设一生总共相亲 n 个对象，不选择前 k 个对象，从第 k+1 个开始，一旦发现有比前面优秀的对象马上出手。我们求解出 k，找出我们

取 k 为某值时能够选到最优秀的对象的概率最大。

解：设 i 为选到最优秀对象时的位置，其中 $i>=k$，那么你选到最优秀对象的概率 P（k）为：

用 x 来表示 k/n 的值，并且假设 n 充分大则上述公式可以近似表示为积分形式：

$$P(k)=x\int_{x}^{1}\frac{1}{t}dt=-x1\mathrm{n}\,x$$

$$\frac{d}{dx}(-x1\mathrm{n}\,x)=-1-1\mathrm{n}\,x=0\Rightarrow \qquad x=1/e$$

$1/e$ 大约等于 37%，即 k/n=37%——37%法则！按此策略，找到最中意男生的概率也是 37%!

也就是说，如果你目标相亲 100 个对象，你找到最优秀的对象应该从第 38 位开始选择，从第 38 位开始，只要发现比前 37 位优秀的对象即马上接受，这时你能够达到模型里的选择到最优秀的对象的概率最大，最大的概率结果 37%，当然按照此模型，你也有 37%的概率没有机会选择到理想的对象，因为最优秀的那位已经在前面的 37 位中出现，而你没有选择，因此你成为了剩男或者剩女。

在大数据时代，有婚恋网站设计了大数据系统，在一定程度上是模仿红娘的做法，搜集用户的个性化信息，为用户提供建议，以实现更加有效和精准的推荐。新系统将根据用户的浏览轨迹和填写恋爱问卷的数据等信息，将适合的双方进行匹配，从而实现个性化、高效率的速配。这些网站的数据显示，通过红娘一对一服务，用户在线下门店相亲的成功率是线上的 3 倍。

婚恋网站还有其他的大数据分析发现，男性和女性相处之道十分微妙。男女在等待对方回信息的耐心程度上，男性的平均时间是 8.5 小

时，女性则是 8.7 小时。而在恋情关系中，女性对于财产的重视程度远远高于男性。就异地恋接受程度来说，男性希望伴侣不要远离，而女性的心理较为复杂，她们偏向于同城，但是当距离特别远时，却认为远距离不是问题。再来说颜值问题，结果显示男性通常是视觉动物，相比之下，女性对于颜值不是那么看重。

世界上就是有不信邪的数据科学家。美国波士顿数学家克里斯·麦金利（Chris McKinlay）自己写程序，只花了不到 90 天时间就在茫茫人海中找到了心仪的对象。

这位克里斯开设了 12 个账户，利用计算机程序随意做答网站的配对问卷，从 2 万名用户中收集到 600 万条问题的答案，然后利用演算程序筛选出 5000 名住在美国的活跃用户，从中按性格分类又选出最符合择偶条件的 2 组女子。之后克里斯又创建了两个账号，诚实地回答这两类姑娘们最关注的 500 个问题。回答完问题后，他发现和自己匹配度在 90%以上的超过 10000 人，最高匹配度达到了 99%。为了获得这些姑娘们的关注。克里斯又编写了一个新程序，自动访问与他匹配度高的对象，对方回访他的页面时，就会给他留言。在经过不少尝试后，克里斯终于约到一名亚裔女孩。他见面时主动披露破解网站的秘诀，对方极为欣赏，二人开始恋爱关系。并在恋爱一周年后克里斯求婚成功，二人终成眷属。

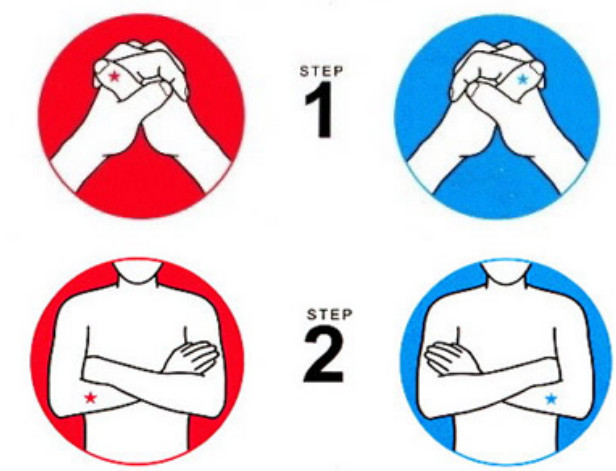
据说，某国为自己的士兵都配备了数字化的头盔，单兵计算机和综合头盔子系统能定时、定位与导航，进行信息采集、处理与记录，进行数据传递，便于指挥员正确实施、调整和制订作战计划，使战场真正成为完整高效的、数字化的一体战场。也就是说，这个头盔不仅能保护脑袋，还能够实时与后方指挥系统相连，通过数据链，后方的

指挥部会将战场情况传输过来，告诉士兵自己目视耳闻做不到的一切。

可就是如此高级的智慧设备，很多士兵上了战场之后却第一个将其抛弃。后来，军方研究后才发现，很多士兵认为，面对战场上瞬息万变的形势，保持头脑冷静，用最快的时间来反映才是最重要的，太多的信息让自己无所适从。对于身处战场前沿的士兵来说，首先干掉正在向自己瞄准的敌人最重要，而这个敌人长什么样子、身高多少、体重多少，甚至拿的什么枪，都无所谓。

由此，我们知道，在复杂的形势下，需要快速做出决定的时候，感性思维往往比理性思维更好用。如果竞争对手已经采取了急风暴雨式的营销活动，我们却还在那里收集数据、磨合模型、研究方案，三个月之后方案出来了，对方的营销效果已经达到，这个时候再出来多好的方案也毫无用处。

如果你想知道自己是个感性的人还是个理性的人，最简单的方法就是伸出你的双手，如图所示。



理性　　感性

日本的（USAUSA~UNO SANO URANA）性格诊断，利用人类左右脑各司其职的特性，设计了简单的两个惯性动作，分辨出这个人是习惯以左脑（主理性、语言、计算、分析）还是右脑（主感性、直觉、想象、创造），来作为解读讯息用的“接收脑”，还是决定怎么说，怎么行动的“传达脑”？进而了解一个人的潜在性格与行为模式。

第 2 章

大数据看起来是无所不能

从三只麻雀之死看大数据的起源

不知道从什么时候开始，“大数据”作为一个概念就火热了起来，很多连数据都没搞清楚是怎么回事的人也开始张口闭口“大数据”。就如同段子里说的，连算命的都改称大数据了。

要更清晰地了解大数据，让我们先从几只被毒死的麻雀说起：

【请看新闻】

“很多麻雀抢食大米后死了。”6 月 29 日上午，一位居民报警称，一艘货船停靠在夜明珠码头处装运大米，其间有不少大米散落在了地上，很多麻雀都飞来抢食地上的大米，可不久后却开始相继死亡，不知什么原因。接到报警后，宜昌市政府高度重视，立即安排工作人员和市公安局、食品药品监督管理局稽查分局工作人员一起赶至夜明珠码头。工作人员赶到现场时，货船正在装运大米，地上确实散落有不少米粒，现场四周还有 20 余只死亡麻雀的尸体。技术人员经过连夜抽检化验，4 批大米均无任何质量问题，货船于 30 日上午离开夜明珠码头驶往重庆。而对于麻雀的死亡，技术人员分析说，它们可能是抢食大米过多导致撑死，也可能是在其他地区食入不健康食品后，恰好在此抢食大米时出现死亡。

一篇“码头散落大米麻雀抢食成批死亡，官方：吃撑死的”的微博近日引发网民热议。湖北省宜昌市政府 7 月 3 日发布消息称，死亡麻雀已被送检，初步发现麻雀体内含有杀虫剂呋喃丹成分，并称从未有技术人员说过“麻雀被撑死”。宜昌市政府部门表示，事发后，有关方面迅速将死亡麻雀送检，公安部门对死亡麻

雀进行检验，发现麻雀胃内有杀虫剂呋喃丹成分。7 月 2 日下午，三峡食品药品检验检测中心、宜昌市土肥站对码头装卸现场散落物及土壤继续取样检验分析。为得出更加权威的结论，宜昌市已将部分死亡麻雀送湖北省级权威机构进行化验，并邀请相关专家赶赴宜昌，对麻雀死亡原因做进一步的分析。

这件事后来变成了罗生门，似乎，毒大米和死麻雀之间的大数据逻辑是这样的：

据说，20 只麻雀吃了散落的大米，死了。

有人传言说，麻雀是吃了有毒大米，中毒死的。

后来，有人说：专家认为麻雀可能是吃多了，撑死的。

再后来，有人又辟谣说：没有人说过麻雀是撑死的。

再再后来，有人又辟谣说，大米没有毒。

再再再后来，有人又辟谣说，大米还没有卖出去，卖出去的都追回来了。

再再再再后来，有人又辟谣说，还有一部分大米没追回来，但大米确实没检测出有毒。

我们不再说后来了，因为这个故事还没有结束。而且，即便被人为地确定终结，好事者也不会就此认为事情结束了。

这次的毒大米与死麻雀的事件，看似传言绕来绕去，实际上却是一次典型的大数据分析的实践，从中可以看出，盲目的所谓大数据分析是多么容易误导公众。

（1）我们找到了所有的麻雀了吗？

我们不知道谁在现场数数了，可以肯定当时贪吃了大米的就是 20 只，如果是很多很多只，那些麻雀去哪里了，为何那些麻雀没有死？

我们做大数据分析，往往号称拿到了所有的数据，但实际上仅仅是能够拿到的那部分而已，也许恰恰是那些我们没有能力拿到或者没准备却拿到的部分，将大大影响我们最终的分析结论。当年，美国总统大选，那么有名的《文学文摘》拿到了 240 万份的读者投票意向，最终却预测失败，相反，盖洛普凭借 5000 个很小的样本就预测成功，也是这个道理。

（2）这 20 只麻雀就是那吃了大米的麻雀吗？

麻雀是否吃了大米，应该比较好检验，但是否正好是吃了这一堆大米，却有点难度。当然，如果是时间比较短，检验起来也应该可以确认。总之，我们要确认大米与麻雀之间的相关性。

大数据分析首先要确认事物之间的相关性，而且要密切相关，一对一的直接相关，如果我们仅仅是把毫不相关的或者可能有一点关联的事物放到一起分析，最终的结论可能很无聊。比如，有人连续看到中央电视台的《新闻联播》结束的时候太阳就落山了，由此得出结论，太阳落山与新闻联播结束相关。

（3）麻雀之死是因为吃了大米导致的吗？

麻雀死了，这是事实；麻雀死之前吃了大米，也是事实。那我们是否就可以说，麻雀之死与大米有关联呢？也不能下结论。我们需要在麻雀的死亡与吃大米之间构建确切的因果关系，也就是说，我们需要找到麻雀之死的死因，而且这个死因是大米之毒。

大数据分析非常关注相关性，甚至对因果关系不予理睬，但这种相关性却往往需要因果关系的支撑。只要是关联密切的直接相关，一定会找到某种因果关系，或者排除某种因果关系。我们做大数据分析，不能仅仅就凭借简单的相关性来下结论，必须通过严谨的因果论证，

才能被严肃地使用。

（4）麻雀之死是因为吃了毒大米导致的吗？

严格来说，麻雀确实有可能是吃大米太多而“撑死”的，我们并不能完全排除这种可能性，所以，专家的话实际上说得在理。即便认定麻雀之死是毒大米造成的，还要分析这毒是如何来的，是大米生产过程中还是有人投毒？当然，这就是公安部门的职责了。

我们只有发现了大米有毒，且大米之毒足以致死麻雀，而麻雀也确实是吃进了这些大米，这样才可以下结论“大米毒死了麻雀”，可事实上舆论早已经抛开了这些逻辑，自顾自地开始从中国的食品安全惯性来考虑。

大数据分析中可能发现很多关联，这些看似可贵的发现却多数都可能是无用的，而且，有些可能是毫无意义的。我们需要对其进行深入分析，特别是要建立起一系列的可证逻辑，由此才可能发现对于我们非常重要的线索，但是，我们却往往不愿意采用“MECE”方法，不想把所有的可能性都考虑到，更愿意先入为主地自以为是，而这往往是误判的主要来源。

对于大数据（Big Data），研究机构 Gartner 给出了这样的定义：大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

麦肯锡全球研究所给出的定义是：一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。

有一个经典的大数据应用案例。来自微软纽约研究院的一名经济

学家，利用大数据分析，成功预言了 2012 年美国大选选举结果和 2013 年奥斯卡颁奖礼奖项归属，准确性高于 98%。

2014 年 3 月 2 日，第 86 届奥斯卡颁奖典礼如约在杜比剧院举行。提名入围者谁能最终捧得小金人，是各界热议的焦点，也成为各大博彩公司的热门盘口。然而早在 2013 年，第 86 届奥斯卡颁奖礼的悬念已被提前揭晓了，做到这一点的就是大数据的分析。大卫·罗斯柴尔德是微软纽约研究院的一名经济学家，他率领的团队通过对入围影片的相关数据进行分析，成功预测出第 86 届奥斯卡颁奖礼 13 项大奖的结果。而且早在 2012 年美国总统选举中，大卫·罗斯柴尔德就曾经使用一个通用的数据驱动型模型，准确预测了美国 50 个州和哥伦比亚特区共计 51 个选区中 50 个地区的选举结果，准确性高于 98%。

大卫说，“我预测奥斯卡金像奖得主的方法与预测其他事情的方法完全相同，其中包括政治。科学是相同的，但证明哪些数据最有用却存在千差万别。”大卫团队的工作方法是，首先关注最有效的数据，然后创建不受任何特别年份结果干扰的统计模型，在建模时要非常谨慎，确保模型能够正确预测将来的样本结果，而不仅仅是过去发生的结果。投票数据、预测市场数据、基本数据和用户产生的数据，这四种不同类型的数据是关注的重点。大卫表示，在预测奥斯卡时，“我更关注的是预测市场数据，这是主要因素，同时采用部分用户产生的数据，这有助于理解电影内部和不同类别之间的相关度。”大卫团队的实践充分证明了大数据分析成为“预测帝”的能力。人们可以通过较为完善的建模，进行快速地数据处理和分析，并让这一分析结果用于商业用途。

大数据技术的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化处理。换言之，如果把大数据比作一种

产业，那么这种产业实现盈利的关键，在于提高对数据的“加工能力”，通过“加工”实现数据的“增值”。有人把数据比喻为蕴藏能量的煤矿。煤炭按照性质有焦煤、无烟煤、肥煤、贫煤等分类，而露天煤矿、深山煤矿的挖掘成本又不一样。与此类似，大数据并不在“大”，而在于“有用”。价值含量、挖掘成本比数量更为重要。

大数据会让我们失去做梦的权力吗

按照百科的解释，大数据（Big Data），或称巨量资料，指的是所涉及的资料量规模巨大到无法通过目前的主流软件工具，在合理时间内达到撷取、管理、处理，并整理成为帮助确认企业经营决策等更积极目的的资讯。

即便如此，有关大数据，也仍然没有大家都能普遍接受的统一定义。可以说，数据量大并非大数据，再大量的数据如果不能被利用也不能被称为大数据，而单一领域的大量数据的集合更不是真正意义上的大数据。根据一般的理解，大数据应该是围绕特定的主题而将看起来毫不相干的数据集成在一起构成统一视图，然后寻找到期间合理的关联因素，从而超越简单的统计分析而得出意想不到的结论。

阿莱克斯·彭特兰教授指出了大数据应用比较成功的几个领域，包括营销场景的预测、城市管理、疾病预测、金融预测等，这些方面都要依靠海量的数据积累和不同的客户应用场景，互联网搜索引擎具有先天优势。

对于大数据的采集，微软算是先驱之一。当初微软每年卖掉几亿

份复制的 Windows，却无法知道用户在家究竟是怎么使用这个系统的。于是他们便对用户的鼠标点击数据进行收集，给 Windows 升级提供依据。这就是最早的“用户体验改善计划”。而同样是收集用户的点击数据，谷歌却做到了知晓用户的性格和爱好，从而实现精准的广告投放，产生了远大于微软的商业价值。

“大数据”之“大”，更多的意义在于：人类可以“分析和使用”的数据在大量增加，通过这些数据的交换、整合和分析，人类可以发现新的知识，创造新的价值，并让很多常态化的认知、判断、思维定势、产品形态、服务模式，形成全新的面貌和演进方向。

横空出世的小米手机、特斯拉的电动车、乐视的超级电视、海尔的空气盒子、引发热潮的微软小冰、热播的《纸牌屋》之类的产品，它们和传统的创新型产品似乎并无很大差异，但背后其实都有大数据应用的影子。以大悦城为例，当消费者想去某一个商家时，百度会通过大数据存储和分析告诉他，这个商家在几层，里面有多少人；消费者想离店，百度地图将指引具体路线，怎么去停车场，更准确地找到自驾车辆。

大数据的价值要通过相应的产品体现出来，比如，智能可穿戴设备就离不开大数据的应用，否则将变成死气沉沉的玩具。在大数据的利用上，国内比较成熟的领域包括互联网金融方面的风险控制、网购领域的智能推荐及物联网交通管理等，比较成功的产品有阿里巴巴的余额宝、咕咚智能手环、百度的百度指数等。

有这样一个流传很广的案例。据说，在细菌还没有被发现的时候，就有一个医生发现医生从停尸房回来后直接做接生手术，产妇的死亡率会明显提高。因此他建议医生从停尸房回来后用肥皂洗手。虽然我

们现在觉得这很正常，但当时的人们没有细菌的概念。洗手跟死亡率有什么关系呢？那位医生就说“我也不知道有什么关系，反正听我的就行，洗完手之后再去接生。”

在一个赌场，你去赌博的时候要在门口先办一张电子磁卡，其实你在办这张电子磁卡的时候，相关的信息已经被赌场获取了。比如说第几次来、大概年龄、种族、职业等。赌场有一个庞大的数据库，拿到数据后就做预测。人跟人的确是不一样的，有的人到赌场输了 10 元钱就心疼得睡不着觉；有的人输几百万元也面不改色心不跳。但不管是谁，都会有一个痛苦点。当在这个赌场里输的钱超过了痛苦点之后，这个人会从此再也不踏进这家赌场一步。因为已经输得恶心了，会觉得这个地方太背，以后也不再来了。从赌场的角度，最好的选择是当赌客快要达到痛苦点时，让赌客住手。赌场里面有很多摄像头，可以看到客人大概现在输了多少。比如你一进去，赌场根据它的数字预测，像你这样的中国人、男性、35 岁、土豪，大概痛苦点比如说是 1 万美金。当你输到 9800 美元的时候，奇迹发生了，你旁边会突然出现一个年轻貌美的公关经理，说：“先生玩得很累了吧，我们的赌场刚请了一个法国名厨，会做世界一流的法国大餐。恭喜你，你被选为幸运顾客。要不带着家人去享受法国大餐，休息一下吧。”为什么服务这么好？因为你的最后一分钱已经被它榨完了。为什么赌场能够精准预测你的行为？因为你的行为和别人的行为不一样。

活在当下，是互联网的正常思维，我们更关注发生了什么，而不再用心思于为何会发生。因为大数据的出现，这种趋势越来越明显。

从本质上说，大数据之于商家，就是通过采集的大量用户行为数据寻找“众数”，发现共同的兴趣点或痛点，然后投其所好地进行产品

设计和针对营销。对于商业机构，甚至社会学研究，大数据都是极好的工具，是传统的市场研究的升级。

《纸牌屋》火了，如果你不看，那就会被人觉得过时了。因为，这部剧是站在了大数据的基础上，根据你的喜好进行设计，你爱看什么就演什么，你爱怎么看就给你怎么演。实际上，这理念一点不陌生，就是我们原来说的“群众的呼声”。央视的元宵晚会，把网络上北京台春晚评价很高的相声安排进来，这也是大数据的一种体现。

当然，一样的理念，不一样的操作。美国人的数据分析能力和手段更先进，数据处理过程更系统化更规范，而我们的应用还比较碎片化，没有形成可持续的挖掘能力。

不过，大数据对于影视的应用、产品的开发，有很大的局限性，不能被过度神话和滥用。互联网公司对于大数据的过分吹捧，有其自身掌握的资源变现的潜在目的。

大数据对于用户需求的前瞻性有限，事实早已不断证明，战略性的发现和规划的真理往往掌握在少数人的手里，即便是靠大数据获得的预测，也往往不是机器和计算的大众来使用。

大数据的预测需要很强的相关性，对于突破创新的影响比较小，过分相信和使用，会失去对市场的敏感预测与突破性产品的诞生。可以这样说，如果乔布斯用所谓的大数据设计手机，一定不会有苹果手机的改变世界。

大数据对于影视作品的影响也非常深远，用大数据来迎合客户，会产生很火爆的流行作品，这非常符合商业利益，但会越来越缺乏伟大的作品。那些领先时代能传颂千古的画作、文学作品等很多都会被扼杀于摇篮。我们不排除很多伟大的作品在当时也很流行，但更多孤

高和寡的作品却是时代前进的更大动力。

有没有人比你自己还了解你的购物需求？ Weather Co 是美国一家能够基于对人们查看天气情况的时间、地点和频次的分析预测消费者行为的机构。该公司积累了超过 75 年的气象信息，覆盖北美等地区的天气、云量等方面的数据。基于这些大数据，Weather Co 不仅能为用户提供单纯的天气信息，而且可以通过数据挖掘，分析天气会对用户消费产生什么影响。比如，某位消费者有在下雨天购买零食的习惯，那么，当他下次查询到天气预报可能有雨时，系统会自动推送一些优惠的零食商品信息给他。这种对用户消费行为的预判，不仅能让用户感受到一种全新的购物体验，而且还可以吸引那些对广告投放精准度要求较高的广告主。例如，Weather Co 发现，在达拉斯，杀虫剂在春天露点（湿度指标）低于平均水平的时候会非常热销；但在波士顿，杀虫剂则是在春天露点高于平均水平的时候畅销。宝洁的营销总监 Kevin Crociata 表示，根据 Weather Co 的特定数据，结合女性消费者所处的准确位置和天气，可帮助投放高度精准的广告。他指出，对于在高温湿热地区查看天气的女士，就应该向她推送柔顺产品的广告；而如果处于低湿度的地区，她的头发没有弹性，那就应该向她投放富弹性配方的洗发水广告。

大数据是商业创新的利器，也是改变人类文化的双刃剑，是个别人的大财富工具，也是让伟大更加落寞的厚壁。那些超越时代的人和作品会更孤独。大数据，很好用，关键看谁来用，怎么用？

数字时代已经让我们失去了对哥德巴赫猜想的兴趣，大数据还会让我们失去做梦的渴望吗？

运营商的大数据为何抱着金碗要饭吃

大数据概念已经火了多年，逐渐从原理阐述过渡到了数据应用阶段，越来越多的公司声称自己将转型为数据公司，依靠大数据来赚钱。

在互联网公司中，阿里巴巴很早就把自己定位为数据公司，看似是做电商业务的淘宝、天猫等都是数据平台，各种交易行为、用户态度和商品信息都成为了宝贵的数据资源，依靠这些数据阿里系衍生出来的网络信贷、芝麻信用甚至也包括余额宝、招财宝等，现在更是向着影视、旅游等各场景进军。

互联网公司其实个个都是收集数据的高手，百度拥有几乎所有中国网民的搜索数据，你想干什么曾经干过什么都逃脱不了百度的眼睛，腾讯平台上数以亿计的用户在交友、聊天和游戏，你在腾讯平台做的一切腾讯都知道。一度很多人把互联网公司的这种经营行为看成是争抢客户，然后是争抢入口，现在看来，它们都是在争夺数据。

电信运营商最早就是搭建平台的，电话线两头伸，所有的通信行为都要通过运营商提供的网络，但运营商们以前过得太悠闲了，也受到隐私保护的严格限制，在语音时代几乎不收藏任何客户的数据信息，有的也只是与收费相关的通话时间和通话时长等有限资料。据说，以前手动交换的时代，话务员是很容易就会偷听到用户之间的谈话的，如果将这些谈话内容收集起来，也许就是最初的大数据。从这个角度看，战争时期的电话或无线电监听，算是最早大数据应用的案例之一。

运营商掌握了相当强的大数据资源，这个没有任何争议，国内外行业里的专家都是这样认为的，但就大数据的应用成果来看，运营商还都停留在枝节，没有形成规模化和系统化，至于大数据的价值变现，

更是处在摸着石头过河的初级阶段。

从理论上来说，运营商的大数据资源得天独厚，数据维度多且数量大。按照中国联通相关部门的说法，对比其他数据方，运营商的数据更全面、丰富、真实。

运营商的用户随时都在产生大量信息，积累了用户身份、套餐消费、语音通信、短信通信、位置信令、手机上网等庞大的数据量，而且数据的真实度和质量都很高，且具有较高的含金量。其中，仅移动用户每天就会产生 800 ~ 1000 亿条访问记录，约 30TB 的数据量。

此外，互联网公司可能具有用户在社交、搜索、电商等某一领域的数据，而运营商则能采集到用户通过移动互联网所产生的全网内容。同时，实名制的号码与每位用户的真实行为紧密相关，并且联通可以通过用户的通信关系验证用户行为特征，这种基于社交网络的信用评估方式，比单一分析每个用户个体行为具有更高的可信度，可有效避免信息伪造。更重要的是，运营商的网络特性使其天然掌握用户的位置数据，这些位置数据实时、连续且唯一，成为大数据应用的宝贵资源。

也正是因为此，运营商可以将 IT 系统所包含的结构化数据与用户上网日志、访问记录、位置信息、终端信息等信息进行整合，更加清晰地感知用户的实际需求，知道用户是谁、在什么地方、用什么样的终端、需要什么、访问什么……也就是说，运营商可以通过数据，发现和感知客户需求，从而为用户提供更好的产品和服务。这不就是大数据应用的最基本样式吗？

媒体报道，2015 年 11 月 27 日，在中国企业大数据联盟峰会上，中国电信正式发布了大数据开放平台和“天翼大数据”品牌，

并推出精准营销、风险防控、区域洞察、咨询报告四类数据型产品及大数据云平台型产品，重点服务于旅游、金融、广告、政府、交通等行业。中国电信与浪潮集团、全联房地产商会、东方国信、中诚信征、中智诚征、华为、中兴、神州泰岳等十余家合作伙伴签署了战略合作协议。中国电信将与战略合作伙伴在大数据产品和解决方案等领域持续开展深度合作。中国电信集团公司副总经理高同庆表示，大数据联盟由中国电信牵头推出了大数据共同成长计划。该计划以大数据应用为主线，旨在发挥各成员单位的技术、行业和市场优势，带动上下游大数据产业服务链条的发展，共同促进大数据产业生态圈的建立。

此前，据中国通信网 2015 年 5 月报道，中国联通与西班牙电信达成合作协议成立合资公司。合资公司总资本为 1 亿元，其中联通出资 6000 万元人民币，西班牙电信出资 4000 万元。由联通宽带公司旗下专事运营位置数据的北京新时讯无限传媒广告有限公司牵头筹备，该合资公司将在中国市场开展基于位置的大数据业务，比如精准营销等。中国联通同时还在与澳大利亚电讯进行谈判，有望成立一家车联网及相关大数据服务公司。业内分析认为，这是中国联通进一步加大大数据业务的决心，未来会有计划进一步整合其他类型数据，以形成结构化的数据分析，提高数据含金量。

更早的 2014 年，在移动互联网国际研讨会上，中国移动原董事长奚国华提出了大数据时代全新的移动互联网战略，即：构筑“智能管道”、搭建“开放平台”、打造“特色业务”与提供“友好界面”。这 16 字方针，体现了中国移动在移动互联时代全面开启

之际的全新战略定位。一年多来，中国移动的各省市公司陆续成立以大数据应用为核心业务的独立公司，进军大数据的布局已经基本完成。

运营商们之所以纷纷成立大数据联盟或者加紧产业合作，都是苦于找不到大数据价值变现的场景，缺乏像阿里巴巴那样完整的内部大数据应用场景成为了运营商抱着金碗要饭的根本原因。但是，这种联盟的方式太松散，将大数据成果变成咨询报告一样对外销售，也绝对不会是大数据的主流。

但是，有数据却不一定就有大数据，有大数据也不一定就能进行数据经营。有数据是一回事，用好数据是另一回事，而数据经营就是完全不同的另外一件事。

很多人关注到了京东的一个动作，作为大数据使用的一个方向，京东成立了京东调研这样的一个独立部门，甚至还为此召来了调查业同行的指责。京东号称拥有快速建立样本库、快速进行分析的能力，如果真的将数以亿计的用户调动起来，那真的可能改变调查业（至少是零售调研）的格局，但现实会如此吗？

与京东类似，作为拥有数据量更为可观且时间积累时间更长的阿里巴巴却没有做类似的产品出来，其实并非是想法不同。此前，马云已经宣布，阿里巴巴的数据将逐渐封闭，只会给合作商家来使用，这几乎是与京东的策略南辕北辙。

商业产生数据，数据推动商业，阿里形成了数据产生与使用的内部循环，而且这个内部循环是内生的、可持续的、价值性的，与此相比，不管是银行还是运营商，甚至电商中的 B2C 平台，数据的价值都只能更多地体现在自身业务管理和营销方面，而很难进行变现，或者

是充分地可以赖以经营的变现。

所以，京东才会成立京东调研，而阿里巴巴根本不需要这样的机构。对于阿里巴巴来说，不管是直通车，还是信用分，所有的数据产品都来自需求各方的直接推动，而这些数据产品也成为需求方必须的生产要素，数据经营和变现是一体化的。由这些商业数据出发，阿里巴巴和蚂蚁金服不仅可以支撑公司平台上的商家业务发展，还可以在信用贷款、基金管理、金融产品开发甚至影视制作放映等领域拥有独一无二的优势。

也正是秉承了这样的思维，阿里巴巴旗下的淘宝才从来不收取商家或卖家的费用，天猫才希望与淘宝差异化的互补发展，支付宝才开放给越来越多的商家使用，看似是在用“钱”连接一切，实际上却是用数据连接所有，纷繁复杂的业务的中心在于数据，这是在地地道道地做着数据经营。

真正的数据经营一定是用数据来经营赚钱，而非通过销售数据产品做咨询公司，数据经营应该是使用数据底层来支持公司的行为由此获得巨大的商业价值，甚至可以仅仅凭借数据资源就可以构筑商业帝国。从这里看，利用自身掌握的数据开发一些具有社会热点的小排行榜发布，或者使用数据做到自己的业务精确化的营销，这些都还只是数据经营的摸索阶段。除了阿里巴巴开发的那些数据产品，也只有百度的大数据股票、疾病预测等算是数据经营的个别案例。

数据经营能否实现主要看数据应用生态是否存在，运营商虽然是大数据的拥有者，但并不是大数据经营的好企业，缺乏应用场景是致命伤。从现在开始，运营商要想实现数据经营，如果不能自己掌握周边服务场景的一切，也需要整合一切周边的资源，让自己的数据有地

地道道的用武之地，而不是变身成为可获得蝇头小利的数据咨询公司。

大数据方法真能解决交通拥堵吗

连续几年的“五・一”劳动节，京藏高速 55 公里的大堵车震惊了世界，中国的交通拥堵问题被空前关注起来。那这种拥堵真的没有解决之道了吗?

如今大数据被赋予了神一样的能量，好像只要是大数据当道就可以解决一切难题。这种想法显然不对，即便大数据可以帮助我们了解得更多，也不能预测到我们想象中的程度。

智能手机已经很普及，大多数的人们都拿着具有定位功能的手机，而 4G 网络又是这样的覆盖广泛，以至于我们每个人的行动时时刻刻都被运营商、互联网应用提供商所“监控”，这些数据被整合脱敏之后可以成为大数据分析的基本信息来源，从而为交通和出行提供管理上的帮助。

媒体报道，2006 年，斯德哥尔摩与 IBM 合作，在通往市区的 18 个路段安装了传感器和照相机。搭载了感应装置的汽车在通过该路段时，系统会自动识别该车辆，并对其征收通行费。没有搭载感应装置的汽车通过该路段时，系统会自动识别照相机拍摄的车头照片上的车牌号码，确认汽车所有者，并对其征收通行费。该系统实施后，斯德哥尔摩市区交通量降低了 25%，二氧化碳排放量减少了 14%。

资料显示，目前大数据在交通中的应用主要有以下几种方式：

（1）公共交通部门发行的一卡通大量使用，因此积累了乘客出行

的海量数据，这也是大数据的一种，由此，公交部门会计算出分时段、分路段、分人群的交通出行参数，甚至可以创建公共交通模型，有针对性地采取措施提前制定各种情况下的应对预案，科学的分配运力。

（2）交通管理部门在道路上预埋或预设物联网传感器，实时收集车流量、客流量信息，结合各种道路监控设施及交警指挥控制系统数据，由此形成智慧交通管理系统，有利于交通管理部门提高道路管理能力，制定疏散和管制措施预案，提前预警和疏导交通。

（3）通过卫星地图数据对城市道路的交通情况进行分析，得到道路交通的实时数据，这些数据可以供交通管理部门使用，也可以发布在各种数字终端供出行人员参考，来决定自己的行车路线和道路规划。

（4）出租车是城市道路的最多使用者，可以通过其车载终端或数据采集系统提供的实时数据，随时了解几乎全部主要道路的交通路况，而长期积累下的这类数据就形成了城市区域内交通的“热力图”，进而能够分析得出什么时段的哪些地段拥堵严重，为出行提供参考。

（5）智能手机已经很普及，多数智能手机都会使用地图应用，于是始终打开 GPS 或北斗定位系统，地图提供商将收集到的这些数据进行大数据分析，由此就可以分析出实时的道路交通拥堵状况、出行流动趋势或特定区域的人员聚集程度，这些数据公布之后会给出行提供参考。

以车联网为例，专家的理性分析告诉我们：一个城市，如果把车和车，车和道路充分链接到位的话，从理论上来说，可以提升这个城市道路通行能力的 270%。

以上这些都是大数据在交通管理方面的应用，会有助于提升道路交通信息的透明度，也对缓解交通拥堵有所帮助，但如果就此认为，

大数据可以解决交通拥堵问题，那就是文不对题了。

我们很多人都乐观地估计，只要信息足够，通过大数据分析来实现的智慧交通系统就会帮助我们做出理性的规划，从而路路畅通。

理想很美好，可现实却很残酷。即便是各部门的大数据应用都起到了作用，国庆节出行的道路却依然拥堵，且没有任何改善的迹象。很多人都体会了 2015 年 10 月 1 日各地道路上的堵车盛况，甚至有乘客下车在高速路上开始遛狗。在这一刻，大数据选择了失灵。

实际上，很多公司通过大数据已经对交通拥堵做出了预测。比如，全国最堵的京藏高速预计从 9 月 30 日到 10 月 1 日下午拥堵超过 24 小时，“十・一”的返程高峰会出现在长假结束前一天下午 3 点到长假最后一天的 23 点。但这些数据都没有能够帮到很多人，大多数人还是会一如既往地走上拥堵的道路。

交通拥堵的核心是通行能力与通行需求不匹配，可能是常态化的道路资源不够，也可能是瞬时车流高峰导致的不协调，但就一般情况而言，多数的城市或郊区道路拥堵都无法通过大数据的交通信息公开来缓解。

大数据肯定不是万能的，即便再强，也只是基于现实数据进行的一种分析，可以给我们提供参考，但这种参考的价值却不应该被无限放大。比如，我们可以提前通过大数据分析进行预警，哪条道路会拥堵，会拥堵到什么程度，可如果条条大路都是超负荷的，大数据的提前预警作用也就失效了。

大数据可以帮助我们提前规划路线，避开拥堵的道路，但一旦道路全在拥堵，我们就失去了选择的机会。在这种情况下，“理性的人”应该选择留在家里，这样就可以让自己不被堵在路上，也不会造成更

大的拥堵，这样选择的人多了，道路可能就通畅了。问题是，很多人都这样想，大家都觉得别人会不出行，结果，群体性理性的选择带来了更大的拥堵。还有一种情况是，大家只能在这个时候出行，再挤也要去，否则就没有别的机会可以选择。

2014 年以后，北京市把使用进京证的范围扩大，进京证的有效期缩短，由此造成了各进京路口办理进京证的排队状况盛况空前，据很多司机反映，正常情况下，办理一张进京证需要 2 ~ 3 个小时，也就是说，如果从天津到北京开车办事，路上只需一个小时，而办理进京证就需要 3 个小时。但大家都知道，办理进京证的排队时长是不固定的，也有司机幸运地半个小时就办完了。很多司机都在预测什么时间办理的人少，有司机选择在半夜 1 点去办，结果排队了 3 个多小时，因为与他有共同想法的人太多，结果造成了人员拥堵。

其他交通领域也一样，大数据的交通信息公开会带来交通流量的透明化，而大家同样的选择会导致下一个交通拥堵的出现，景点的热力图也只代表现在，如果大家都得到同样的信息，结果冷点就会很快变成热点。当然，饭店可能是个例外，如果你发布的某个饭店排队人数多，很可能导致的是这个饭店的排队人数更多。

有人说，单一个体的出行是随机和不可控的，而一旦每个交通参与者通过某种方式“连接”起来形成一个大的可实时分享交通信息的群体，那么这个群体就具备了某种“智能”，通过互相影响来达到自我调节和自我优化，而结果一定会朝着减轻拥堵的方向发展。确实，在现代移动互联网状态下，每位终端用户既是交通信息的生成者，又是交通信息的提供者，从而以互联网彼此连接、相互影响，但这种交通智能对交通拥堵的缓解起不到多大的作用。

因此，大数据的分析结果在群体性公共知识的面前，一定会变得毫无意义，甚至会起到负面作用。很多人认为，信息不对称是导致交通拥堵的重要原因，而在实践中，信息太对称，也一样会导致拥堵。

我们获得的大数据也并非全面，还有很多人并不使用智能手机的定位功能，一些大数据分析公司无法获得数据。斯德哥尔摩是通过在公共交通工具上安装传感器，分析这些传感器数据，来掌握道路的拥挤情况，这种方式对城市道路很实用，而对于高速公路来说，目前大数据分析普遍采用的用户个人的智能手机定位数据并不可靠。

大数据分析也是十分复杂科学的工作，任何的理论或操作上的微小失误都可能造成分析结果的被错误使用。即便获得了用户数据，在分析的方法和使用的策略上也存在不足，难以充分发挥大数据的价值，这也造成了分析上的偏差，错误的引导会带来局部更为严重的拥堵。

与此同时，大数据在偶发事件面前也无能为力。在国庆节这样的大车流的情况下，一起偶然发生的交通事故就可以造成蝴蝶效应，由此带来一个路段的拥堵，然后是整个路段的拥堵，接着会造成更多辐射的路段上的连环拥堵的发生。这种事故是不可预测的，其后果也很难提前预知，而节日道路的变通余地很小，一旦发生突发事件，交通拥堵的严重程度就会超出想象。

实事求是地说，大数据确实可以提升道路管理水平，但大数据却无法解决信息沟通中的群体错位决策，也无法解决超出负荷的刚性需求到来的道路绝对拥堵，更没有办法应对随时可能出现的随机性的事故影响。大数据对于节假日期间的交通拥堵问题，绝对是有心无力。

德国足球队中的“第十二人”

首先，请不要误会，这里的第十二人并非是指“黑哨”，而是悄然影响绿茵场强弱较量的“大数据”。

无论是作为教练员指挥比赛，作为球迷观看比赛，还是进行市场分析，经验和感觉往往不大可靠，尤其是我们还存在或多或少的迷信和不可靠的预感，凭这些来做判断显然不如用数据分析的方法来得科学。

我们在这里以足球为例，如果要进行数据分析，那一定是要先建立一个模型，将对比赛或对市场的影响因素列出来，给出他们的相互关系，然后我们就可以通过这个模型进行计算，得到我们想知道的结果。

根据一般的情况，在研究球赛的时候，我们可以把关系球队胜负的影响因素归纳为下面的八个方面：

（1）球队本身的综合实力；（2）教练员的水平和战术；（3）裁判是否公正或倾向性；（4）踢球的场地在何地；（5）观众和球迷的取向和文明程度；（6）气象条件；（7）后勤保障条件；（8）历史交战成绩。

在此基础上，我们可以继续将这八个因素进行细分，比如球队的综合实力可以包括如下方面：（1）球员个人技术；（2）体现教练战术意图的能力；（3）整体配合水平；（4）士气；（5）心理素质；（6）体能。

如果需要还可以继续细分，比如心理素质又可以分解为下面几个因素：（1）是否有恐惧对手的毛病；（2）在落后或失利的情况下是否有耐受力；（3）是否经受得住胜利；（4）在重大比赛面前是否心理负

担过重；（5）是否有因家庭或个人问题而影响情绪；（6）是否因队友或教练矛盾而影响情绪。

我们把每个大项、小项都列出以后，还需要去研究各因素之间量的关系，看看各个要素在其中的影响占多大比例，也就是我们常说的重要性系数，可以根据自己的或者专家的经验来确定，如果有历史数据，当然也可以通过统计计算得到。系数得到了，我们就基本完成了对一场比赛的分析模型。

对于球迷或者教练可以用这样的模型来进行定量的分析，初步预测比赛的胜负，但是这样的分析属于静态分析。在静态分析中虽说有些要素是在发展变化，但绝大部分的要素是静止不变的，两队还没有上场比赛，我们所做的分析只不过是一种预测，而预想可能与实际情况相差甚远。这就是为什么足球专家们在预测比赛的时候说得头头是道，可他们买彩票也很少中奖，更不要说被称为乌鸦嘴的“贝利”的预测了。

有些资料容易获得，但由于各种信息封锁或伪装，我们通常面对的是“黑箱”，虽然我们可以通过考察其外部特征来做基本的评估，但我们了解的对手情况仍然可能并不真实，或者我们的队员在真正比赛的时候受到对方的压制没有很好地发挥水平，我们可能考虑了许多因素但是对各因素之间的相互作用和牵制估计不足，比如出现了“黑色三分钟”或者是打疯了，导致我们的预测失败。在市场研究中也经常会出现这样的情况，你可能了解到对手正在研究一种会走路的机器人，可在发布会上你看到的是一个向你爬过来的家伙。

比赛越是激烈，场上局面越是变化万千，那我们的动态分析就越要及时，这种分析也被称为随机分析。坐在场边的教练时刻关注着场

上的变化，他的随机分析一直不间断地进行，我们看到他请求换人、面授机宜，甚至焦虑地在场外大喊大叫，实质就表明他对随机分析做出了决策，并采取了措施。球迷在观看球赛的时候也是时刻提心吊胆，根据场面的形势不断修正自己的看法。

在比赛的过程中，赛事转播方为了球迷更好地了解比赛，会经常及时地提供一些场上的比赛数据，比如双方的射门次数、越位次数、控球时间、犯规次数、红黄牌等，还会通过动画、图表等来形象地演示进球线路，甚至球员的动作。这些都可能帮助我们更好地理解比赛。

时间变化的频率决定我们随机分析的次数，如果变化过快就可能没有时间通过这样的动态分析影响比赛。在战争中，一颗导弹命中目标可能只要几分钟时间，而留给拦截或者躲避的时间更短，过了这个时间导弹将突破防御系统，动态分析也就失去了价值。

如果说上面这些分析还停留在纸面，那么大数据已经成为了德国队的“第十二人”。2014 年巴西世界杯，德国队再享冠军荣光，而帮助德国队获胜的“秘密武器”之一，正是在悄然影响绿茵场强弱较量的“大数据”。

在世界杯比赛开始前，德国足协与 SAP 公司合作了一款名为“Match Insights”的足球解决方案，用以迅速收集、处理、分析球员和球队的技术数据，基于“数字和事实”优化球队配置，提升球队做战能力，并通过分析对手技术数据，找到在世界杯比赛中的“制敌”方式。这款数据分析系统首先通过摄像头、传感器等工具捕捉到球员跑动速度、位置、控球时间、防御范围、动作细节等大量数据，并传入数据库，随后，基于 SAP HANA 平台运行的分析工具可迅速对这些数据进行后台分析处理。在短短 10 分钟内，10 名球员用 3 个球进行训练，

可产生超过 700 万个可供分析的数据点，而 SAP 数据分析平台完全可对这些数据实现实时处理。通过这一数据工具，德国队教练可以迅速评估比赛状况，每个球员的特点和表现，球员的防守范围，对方球队的空档区等信息。通过这些信息，教练可以更有效地对球员上场时间、位置、技战术等情况优化配置，以提升球队表现。

国家统计局经常面向社会公布一些调查和统计的数字，比如进出口额、物价指数、居民收入等，还会进行农业普查、工业普查、人口普查，企业也会自己或者请专业的公司来进行市场调查，主要就是为了得到动态分析的随机数据库，以便据此做出正确的分析和决策。时效性很重要，如果我们用过期的数据来进行研究，那很可能不能得到正确的答案。

在进行动态分析的时候，思维必须是连续的，因为行为是连续不间断的。无论是看球赛的球迷还是指挥作战的教练员，都必须始终关注场上的变化，根据自己掌握的知识和分析方法来进行判断。很多球迷都有体会，足球场上的变化很快，必须一直盯着屏幕，否则可能整场比赛中仅有的一个进球恰恰出现在你离开的那一刻。如果你是球队的主教练，你敢离开吗？在进行市场分析的时候也是如此，任何孤立的分割的研究都不会有太大的价值。

大数据之下，人而无信，不知其可也

中国是一个有信用的社会吗？显然不是，或者不太是。在中国，契约精神高度缺失，不守信已经成为中国社会商业发展的大敌。由此

引发出很多社会的不和谐行为，特别是一些游客在黄金周期间出国旅行的不文明行为更是被曝光太多。

不过，即便每个人都在骂中国是个缺乏商业信用的社会，但却很少有人去思考为何中国社会缺乏信用，甚至很多人只是简单地将其原罪归于几十年前的那场十年浩劫。

从深层上来看，中国的信用问题大爆发与所谓的社会变革关系并不大，而是来自信用无用处。如果一个人只要失去了信用就无处可藏，那么这个人根本就不敢也不会去让自己的信用受损。

古人说，人而无信，不知其可也，但在现实中却是，人而无信，做什么都可以。甚至于，一些不讲信用的人却获得了社会上的“好评”，成为了各方面的优秀杰出成功人士。

在这个社会上，我们到处看到的是，骗子发财致富，老赖生活富足，欠账的衣食无忧，应收账款让很多企业破产倒闭，三角债问题竟然多次成为中国经济最大的毒瘤。正是在这样的负能量的引导下，全民的信用水平开始下降，直到老太太自己摔倒却要赖上伸出援手的好心人。

中国有自己的征信系统，但这样的征信系统使用场景却极少，甚至甚少有人关注。之所以大家都不重视信用，因为国家的征信系统也只有在银行贷款买房的时候才被人用上。

因为个人信用的使用范围狭窄，几乎没有任何地方会考虑个人信用，在这样的社会中，我们谁还会把诚实守信当成天大的事情？

人类都是相似的，但我们却一致认为欧美国家的社会信用比我们好。从经济学的角度来看，都说美国、欧洲的商业信用好，那是因为，在那样的国家，如果你做了违背诚信的事情，你将寸步难行，这里不

守信用的代价是高昂的，人们就不敢也不值得为了蝇头小利而让自己的信用受损。如果一次逃票就可以让自己未来找工作无人敢要，你还愿意去逃票吗？

中国高高在上的国家征信系统非常严密，但却始终难以覆盖全体国民，更是在实践中的使用率极低，没有社会普及的征信系统就变成了银行发放贷款和信用卡的专属工具。在这样的情况下，整个社会的信用体系根本建立不起来。

随着互联网的快速发展，中国社会建立全民信用的机会到来了。到目前为止，蚂蚁金服的“芝麻信用分”、腾讯征信的信用评级、前海征信的“好信度”、中诚征信“万象分”、拉卡拉“考拉分”、华道征信的“猪猪分”开始运行，在我们每个人都离不开互联网生活的背景下，这些分数的高低已经可以得到准确的评估计算，并将深刻影响我们以后的生活。

在 2015 年开启的首个 6.6 信用日的测试过程中，据媒体的报道，在北京“无人超市”现场，有三位女性现场拿走了价值昂贵的货物，而没有付钱；还有人往返好几次，拿走数袋价值不菲的烟酒，并只支付了 10 元钱。这些人视信用为无物，贪恋小财，但也折射了中国信用建立的难度。

按照芝麻信用发布的首份社会信用调查报告，通过对 1.5 万名用户的调查，82%的中国消费者认为个人信用对自己非常重要，85%的消费者对个人征信体系的未来看好，能接受第三方公司提供的个人信用评估。知道个人信用记录的人当中，只有 44%的人前去查询过个人信用记录。在对个人信用的使用上，91%的消费者都是集中在“银行贷款”上，89%的被调查者希望个人信用的应用可以扩大范围。可见，中国并

不缺乏信用发展的社会基础，缺乏的只是社会监督与执行机制。

如今，仅仅依靠信用分，就可以享受到“信用签证”服务，芝麻分达到一定标准的用户，凭借芝麻分和芝麻信用报告，就可申请新加坡或者卢森堡签证，减少很多证明材料的准备和提交。此外，消费金融公司招联推出了贷款利率优惠活动，芝麻分达标的用户在信用日当天贷款可享受利率 6.6 折，并且还有 10 个免利息贷款名额，要知道这家公司正是招商银行与中国联通的合资公司。还有，依靠信用分还可以在出行、租车等方面享受到先用后付的优势。

此外，作为已经纳入国家信用体系的互联网信用分，未来还会影响到银行贷款等大额的个人消费的支出，甚至会影响到社会人际交往和各种商业合作。

当然，随着信用分的引入，围绕着电商网购、在线旅游、餐饮 O2O、P2P、消费信贷等都将产生新的商业模式，而像阿里巴巴与腾讯等也就拥有了新的竞争优势，随着信用分使用范围的扩大，也必将倒逼很多人开始更加珍视自己的信用积累，在现实生活中时时刻刻遵守信用。信用无价，但信用分将值千金。

2015 年 7 月 24 日，最高人民法院与芝麻信用签署对失信被执行人信用惩戒合作备忘录。同月，双方通过专线方式实现数据对接，共享失信被执行人信息。芝麻信用会同神州租车、趣分期、去啊、我爱我家公寓等各应用平台在消费金融、融资租赁、信用卡、P2P、酒店、租房、出行等场景全面限制失信被执行人，主要措施包括：（1）限制失信被执行人申请贷款、融资等金融行为；（2）限制失信被执行人预订机票、列车软卧，及非经营必须车辆、旅游、度假产品等；限制预定三星级以上宾馆、酒店；（3）限制失信被执行人进行奢侈品交易等其

他高消费行为。截至 2015 年年底，芝麻信用通过各应用平台限制失信被执行人购买机票、租车、贷款等已超过 13 万人次，5300 多名失信被执行人因此还清债务，其中 1500 多名失信被执行人是长达三、四年一直逃避执行的老赖。

随着围绕互联网大数据建立起来的信用评价体系的成熟，各种应用场景越来越多，信用分的多少已经开始关系到个人的教育、生活和工作。先是为了利而珍惜自己的信用积累，然后一定是为了弊而不敢去冒信用分缩水的危险，整个国家的信用体系就在这个过程中得到了最底层的建构。

我们每个人都暴露在互联网大数据之下，我们每个人都生活在互联网应用无处不在的场景之下，只要信用会关系到每个人的一生，谁还愿意为了一点蝇头小利而让自己的信用受到伤害呢?

继芝麻分之后，芝麻蚁盾反欺诈服务已经在金融、手机 3C 产品网络直销平台、O2O、出行、团购等行业开始广泛引用。芝麻信用反欺诈产品总监林述民曾表示，芝麻蚁盾反欺诈产品可以帮助金融机构和互联网商户发现哪些用户存在欺诈行为，帮助他们更好地进行风险管理。

目前信用卡贷款的包装、组团欺诈骗贷的情况屡见不鲜，尤其是在信用贷款领域，约有 60%来自于欺诈，这其中有一半以上是由于身份造假和资料包装。在数据维度不全面的情况下，银行等放贷机构缺乏充分和有效的交叉核验手段，容易被组团骗贷者钻空子，而且存在大量的人工审批工作。

芝麻信用等第三方征信机构由于数据来源丰富，可以进行信息的交叉验证，同时通过机器学习等技术手段，有效的验证和评估信息的有效性。芝麻信用曾表示，其数据来源包括电商数据、互联网金融数

据、公安部人口户籍、最高法老赖，工商注册等政府机构数据、合作伙伴数据，以及各种用户自主递交的信息等，涵盖购物、出行、住宿、转账支付、投资理财、生活、公益等数百种场景数据。

芝麻信用蚁盾反欺诈产品 IVS（Information Validation Service，信息验证服务）已经给金融行业风险管理带来了新成效。通过对高风险客户的自动筛选和低风险客户的审批流程简化，芝麻信用蚁盾反欺诈产品有效地帮助某银行信用卡中心优化了审批流程，节约了 10%的人工工作量。此外，通过信息验证等反欺诈服务，已有银行将虚假办卡的识别能力提高近 3 倍；通过芝麻信用行业关注名单识别不良用户占比达到平均的 4 倍以上。

热门手机销售，经常采用定期网上预约抢购的方式。然而大量的黄牛账户，通过技术手段比正常账户更快地抢购成功，导致正常优质用户很难买到手机。这种行为，一方面扰乱了市场价值秩序，另一方面也给手机品牌带来了负面影响。

为打击互联网黄牛倒卖，魅族上线了芝麻信用蚁盾反欺诈产品 RAIN 分（Risk of Activity，Identity and Network），双方数据与技术共同沟通、推进，黄牛订单比例大幅度下降，压制到 10%以内，有效地提升了客户的网购体验。通过芝麻信用的帮助，魅族也成为手机行业反欺诈领域的标杆。

以后，以芝麻信用评分为主打，以风险识别和评估、实名认证和反欺诈等为辅的芝麻信用全线产品体系已开始深度应用。与花田、世纪佳缘合作，打造诚信婚恋体系，提供的是实名身份认证，或者说是身份核实；专线连接最高人民法院，实时更新“老赖”名单，可以作为其反欺诈服务的重要数据来源之一；免押租车、免押租房、消费金

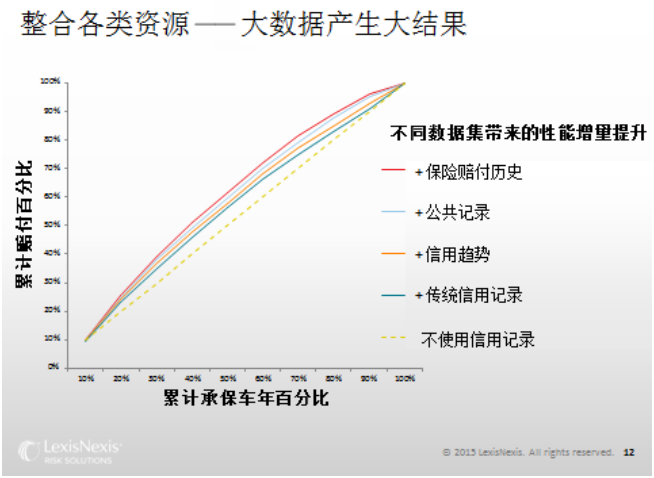
融等服务，更是风险识别和评估能力的体现，背后的服务不只是芝麻信用评分，兴业银行、华夏银行、北京银行等的合作，更是金融行业综合信用解决能力的体现。

近年来，核心的银行征信数据已经发生了变化，除了消费者行为的改变和数据明细程度的提升，还产生了一些全新的数据字段，为消费者风险评估带来了许多有价值的新洞见。面对不断演变的数据来源，律商联讯长期致力于扩展消费者风险分析维度，从全球超过 1 万 3 千多个数据源采集了 500 亿条消费者和企业记录，为保险和金融服务等行业积累了海量的数据资源，其中包括：历来的居住地址和住址稳定性、电话和水电煤气记录、职业证书、教育历史、，破产、抵押、判决和驱逐等数据。

除了丰富的公共记录和第三方数据资源，律商联讯通过建立保险行业共享型数据平台，为行业引入了一个全新的数据成分，完善了为以保险为中心的消费者金融视图。

律商联讯将这些非传统数据引入保险市场，生成独特的变量和行业风险评分，与传统征信数据一起用于风险定价和承保决策，帮助保险行业利用数据优化工作流程，更好地评估风险，从而提升从展业到理赔、覆盖客户完整保险生命周期的各个环节的工作效率。

如下图所示，掌握的数据越多，保险赔付风险模型的预测能力就越强。每增加一个数据集，我们都能看到模型的预测准确度获得显著提升——改良后的信用记录，加上公共记录，再加上保险赔付历史，可以在传统信用记录的基础之上带来 30%的模型效能提升。



实际上，蚂蚁金服的运营管理本身也是大数据的使用者和受益者。2015 年“双 11”期间，蚂蚁金服 95%的远程客户服务已经由大数据智能机器人完成，同时实现了 100%的自动语音识别，蚂蚁金服客户中心整体服务量超过 500 万人次，客服人员的精力可以更好地集中处理复杂类客户问题和工作。

当用户通过支付宝客户端进入“我的客服”后，人工智能开始发挥作用，“我的客服”会自动“猜”出用户可能会有疑问的几个点供选择，这里一部分是所有用户常见的问题，更精准的是基于用户使用的服务、时长、行为等变量抽取出的个性化疑问点。在交流中，则通过语义分析等方式获得关键信息再给予匹配。由于不断积累扩大的样本库及持续调优的算法模型，使得交流更加智能。这也将是未来客户服务领域大数据应用的典范。

大数据助传统银行涅槃重生

2015 年招商银行宣布取消一切转账手续费，随后宁波银行也跟进。我们可以预测，很快就会有更多的银行加入到免费大军。这是为什么呢？

其实，不仅仅是转账收费，很多人也疑惑自己的钱存在银行还有可能被收取小额账户管理费。从银行的角度讲，传统银行的 IT 系统在用户进行转账或是账户管理上都是有成本的，而且还不低。据测算，银行单笔交易成本以“角”计，单账户年成本按银行规模在 30~100 元不等。

不过，这一切都在被金融云彻底改变。喧嚣了多年的云计算终于在 2015 年前后开始落地，特别是微众银行、网商银行等云中银行的诞生，更是将云的作用充分发挥出来，而金融云已经是炙手可热。

传统金融企业要建设网点，要建设规模庞大的 IT 系统，这些投入都会分摊到金融业务的成本之中，由此造成了这些金融机构的经营模式局限，特别是服务中小企业或小额低价值用户的时候捉襟见肘。

不过，以上这些劣势恰恰是互联网公司的优势。比如，互联网公司为了发展各种各样的丰富互联网应用而建设起来的数据中心及云服务，同样可以为其金融业务服务，甚至可以为其他金融机构提供云服务支持。这些云能力已经经过严苛的互联网业务考验，安全性有充分的保障，低廉的成本更是让传统金融机构望尘莫及。

据公开资料显示，2011 年移动银行的用户数量是 3200 万个，到 2013 年这一用户数量已经超越 2 亿，到了 2017 年移动银行的用户数会将近到 5 亿，渗透率将近 1/3，1/3 的中国人口将会使用移动银行服务。

这样的用户使用习惯的变化催生了移动端的金融业务兴起，也造就了微信红包和支付宝钱包的火爆，更让银行也不得不去主动适应，否则传统银行一定会被淘汰。

根据估算，如果依然采取传统方式不做改变，到 2020 年，银行的 ROE（净资产收益率）可能会从今天的 19%降低到 5%，如果能够从业务模式及成本上面做出一些调整，有可能会维持在 10%～12%，这是一个市场普遍水平。这种改变必然来自云服务，而中小银行完全自建云服务基本不可能，也不经济，选择与互联网云服务商合作就成为了必由之路。

专业机构测算，如果采用云服务，作为银行主要运作成本的 IT 架构的运维成本会大大降低。据不完全统计，小型银行每个账户的 IT 成本 100 元，大型银行每个账户的 IT 成本 20～30 元。蚂蚁金服报告认为，金融云把单笔支付交易成本降到 1 分钱左右，单账户成本已经降到 1 元以下。微众银行数据也显示，利用海量服务分布式的架构，将成本下降了 80%。因此，我们可以说，使用了云服务，只需要小型银行的 5%单位成本，就可以服务好用户，这种竞争力没有一家银行可以视而不见。

因此，我们可以预计，当金融机构逐渐用金融云等云计算代替现有 IT 系统后，面向个人用户的转账收费、小额账户收管理费都将成为历史，而面对中小金融的服务也将提升到新的高度。

面对正在到来的互联网金融时代，很多中小金融却没有技术能力，也没有财力去搭建一套与之相应的系统。比如，很多地方性银行甚至没有能力搭建网银系统。这些中小金融机构，恰恰是服务于三四线城市、偏远地区的主力。通过金融云等金融基础设施公共服务，让中小

金融机构可以“拎包入住”，用较低成本快速开展互联网金融业务，大大提升金融普惠性。同时以往金融行业的 IT 系统多是以产品为中心的交易式结构，而金融云带来的将是以用户为中心的交互式结构，所以基于金融云生长出来的金融产品，在用户体验上会更好。

不仅是低价，金融云带来的大数据处理能力，还能够让金融机构利用大数据，低成本地实现信贷业务。金融机构可以在线判断用户信用水平，无需用户再当面提交各种证明材料，或是担保抵押，就能让那些小微企业、草根用户非常方便地通过网络贷款，而且贷款成本也会更低。

未来金融云普及后，金融机构可以根据平台上积累的数据，实时发现客户的贷款需求，在发现一个人或一家企业可能需要贷款后，主动第一时间找到客户。在理财方面，投资者也不用再担心没有理财知识，看不懂复杂的产品说明了，系统会自动根据数据，在平台上推荐合适的投资组合。

总之，云计算让我们从 IT 进入 DT 时代，它带来的大数据必将极大改变金融行业。在 IT 时代，数据是应用的产物；在云时代，应用是数据的表现形式，数据本身即是应用。金融云已经在改变金融业的传统力量，在一些大的互联网金融平台更加成熟之后，整个金融业都将发生根本性的变革，传统银行也将拥有更加低成本发展的能力，金融业的历史新时代正在到来。

用大数据方法保护大数据的安全

大数据时代，各种各样的信息充斥而透明，几乎没有任何人、任何组织可以脱身事外，但这种信息暴露的直接后果就是隐私保护更加艰难。特别是，如果你的银行账号、密码泄露怎么办？你的文件服务器中存储有重要的机密数据，该怎么进行保护？

既然是大数据，那就有可能也用大数据的方法来进行数据保护，很多公司都在进行这方面的科技攻关，包括网络层面的安全解决方案，也包括应用层面的用户信息保护机制。

网络安全要依赖网络管理上的大数据应用

在网络层面，作为全球领先的信息与通信解决方案供应商，华为在美国 RSA2014 安全峰会上阐释“用大数据分析铸就安全敏捷网络”的理念，并发布了下一代 Anti-DDoS 解决方案，提供 T 级 DDoS 防护性能，同时宣布 T 级高性能数据中心防火墙，成功通过了美国 NSS 实验室的测试，成为业界首款经过第三方认证的 T 级数据中心防火墙产品。这两款业界领先的高性能产品，引领安全进入 T 级防护时代，为“大数据”的安全保驾护航。

在大数据的背景下，全网的视角看安全和单点看安全，是不一样的。华为使用基于 Controller 的技术方案可以看到全网的东西，用大数据分析的方法去发现一些潜在的威胁，由此建立更高的安全防范。

华为还提出了一个新的技术叫沙箱，就像一个病毒培养皿，它可以模拟一个 Linux 的环境，模拟一个 Windows 的环境，一个 Android 的环境，一个 iOS 的环境。如果发现可疑应用就会把这个应用放到里

面，让它在一个假环境里跑，监视它的各种行为，如果它去攻击 Office，几个送到假 Office 的响应就报警，最后分析出来的确是个潜在的攻击或者危险。网络用户把数据送到沙箱里面去观测，自动观测、自动分析，然后自动报警。这样可以把非常潜在的初级阶段的威胁抓出来，更好地保护网络。

用户在应用端的安全更需要大数据理念

另外的风险来自普通用户，用户的重要信息可能丢失，可能被盗，在极端情况下，涉及用户资金的账号密码、身份证件等都可能同时被其他人获取。如此，还能保护用户的信息及资金安全吗？

在这方面，阿里巴巴因为电子商务和互联网金融的原因会首当其冲遇到难题。根据相关人员的介绍，阿里巴巴也在利用大数据的方法进行信息保护的探索，即便在极端情况下也要保护用户的资金安全。

其实，我们经常在影视剧中看到战场上曾经出现过的声东击西的经典方法，一只主力部队准备偷袭战场，为了掩盖调动的信息，往往会仅留下总部的发报员，继续在原地进行伪装的收发电报和指挥，这种方法也确实在战争中成功应用，究其原因就是，每个发报员都会有自己独特的指法、速度，形成“指纹”，敌方的监听部门会根据收发特点来识别军队番号和行动路线，这实际上就是一种大数据的应用。

依据这样的原理，我们每个人在使用 PC 或手机等登录账号、输入密码、点击链接等也会形成自己的习惯动作，这些动作形成的大数据信息也会被记录和分析，如果哪一天哪一次系统突然发现这些动作都出现了异常，就会采取拦截措施，通过一系列的新增信息核对步骤来保证交易的安全，特殊条件下会中止交易并与资金所有人进行直接沟

通核实。

感觉安全对用户来说非常重要

多数人都知道在网络上进行资金的交易存在安全隐患，所以，往往使用很复杂的密码，或者经常更换密码，以此来提高安全水平。这种做法是正确的，使用含有数字、字母或者其他特殊符号的密码当然有利于提高安全等级，但这种做法在很大程度上也只是提高了用户自己对安全的感知。

安全是一种个人的感知，就如同我们离开家的时候都会锁上门，甚至会为了更强的安全感而选择安装最贵的防盗门或超 B 门锁。可事实上我们也都清楚，这些防盗门和门锁对于职业盗贼都是“小儿科”，并不能保证家庭财产的安全。不过，正因为在安全上的投入增加，我们的安全感也增加了。

同样的道理，账号和密码在网络上也是防君子不防小人的安全程序，并不能抵抗黑客或诈骗分子的各种攻击与圈套，我们要保障网络上的支付安全需要更为先进的理念或方式，其中最重要的安全依靠的是支付系统的后台安全机制。

在这方面，支付公司会在用户的支付环节上设置多种安全“印象”，比如，要求用户两次输入账号或密码，而且不能使用复制，这样可以很大限度地保证用户不会支付到错误的账户，还有，在用户登录账户或进行支付的时候要求输入验证码，包括随机数字、字母或文字、图片识别等，甚至，12306 网站现在都在要求用户输入需要经过“智力测验”一般的图形问题。正是因为这些的“麻烦”，用户会感觉到比较安全。

更重要的是，对于用户的安全感知来说，互联网金融公司在安全领域的投入越大，用户对安全的感知就会越好。这些投入包括资金方面的投入，也包括在科技研发、人力资源及系统建设上的持续加强等，让用户知道这些努力，会大大提高用户的安全感知。

当然，如果要让用户有更强的安全感，并不能完全依赖技术的提升，还必须通过保险设计来达到。按照现在的技术标准，支付宝已经达到了百万分之一的风险控制率，这样的标准在业内都是领先的，但也不能完全打消用户的担心。于是，支付宝推出了账户安全险，通过金融的方式解决金融的问题，0.88 元可以一年保 100 万元，出现支付安全问题可以全额得到赔偿，也就打消了用户的担心。

密码仍然是用户安全的第一道防线

虽然更高级的密码并不能更好地保障用户的支付安全，但密码仍是用户安全的第一道防线。对于非职业的网络攻击或者意外引发的安全隐患，密码还是具有很好的保护作用的，至少可以让用户躲过很多初级的安全风险。

根据支付宝的数据和安全防护经验，用户密码的被盗取或丢失有几种类型，占比最大的是扫号和社工。

所谓扫号，是指你在别的网站的账号密码被坏人知道了，然后坏人用这套密码来登录支付宝等，因为不少懒人在所有网站都是用一套密码，所以很多坏人会利用其他渠道得到的密码来试着打开你的支付账号。因此，要想保护密码，我们最好将重要的支付账号和密码设置成与其他普通的网络账号密码完全不同的名称或组合，这将大大提高你的支付安全性。

所谓社工，就是假冒各种公检法、熟人好友、假客服等，通过短信、聊天工具，把你的各类信息骗走，然后盗取或是更改你的密码，以此来使用你的支付工具进行转账或消费。这种方式最难以防范，属于典型的诈骗犯罪的受害者，最好的应对便是掌握根本原则，密码决不告诉任何人，打死都不说，因为合法的机构或者客服是绝对不会向用户索取密码的。

此外，钓鱼和木马也是盗取密码的重要方式。所谓钓鱼，就是搞个假网站，比如弄个 tiaobao.com，长得和淘宝很像，蒙骗你去输入，当你一输入，信息就泄露了。木马就是中毒，这些木马隐藏在你的电脑或手机中，记录下你的各种录入传送给黑客。面对这种威胁，最好的方式是多个心眼，不乱打开网络链接，不随意安装不明的应用程序，还要安装相关安全软件定期更新。

支付宝的数据显示，之前外界很担心的手机丢失导致的问题占比并不高，大概是 2%，可见大家的密码保护等还是有一定的帮助，特别是手机锁屏等。当然，一旦手机丢失或发现自己原来的手机号被二次放号，就应该快速地更改重要的账户密码信息，或者与运营商进行沟通处理。

有密码也取不走钱是支付安全所追求的重要目标

在这个互联网大发展的世界里，只要上网，每个人的信息都不可能绝对安全。事实上，安全只是相对概念，世上没有绝对的安全。对于用户来说，账号和密码的被盗始终存在可能性，再高级的密码设置也不能彻底保障用户的安全。

对于互联网金融企业而言，也不能将支付安全寄托于用户自己的安全意识和安全保护，系统建设和安全机制发挥作用才是保障用户支付安全的必须。

于是，支付企业会设置安全的几道防线。比如，支付宝会要求用户设置密码保护问题，还会要求与用户的手机进行捆绑，这样，当用户密码出现异常的时候，就会通过比较私密性的问题回答来验证是否本人，或者通过短信验证码来保证支付更为安全。当一个用户连续多次输入错误密码之后，还会暂时锁定以防机器破解的发生。

此外，很多互联网金融机构还会通过增加安全验证程序来进一步保障安全。比如，银行特别流行使用 U 盾，通过硬件与软件的结合提升安全系数，而支付宝等也会要求在电脑上安装支付证书。未来，随着生物技术的发展，指纹、虹膜、刷脸等都会被利用起来加强安全保证。

很多人遇到过，当你输入错误密码，或者刚刚到达一个从来没有去过的地区，或者使用了一个以前没有使用过的通信网络，支付宝也许会突然要求你在登录的时候输入图形验证码或者通过手机短信来进行验证。其实，这就是支付宝 8 年来致力于建设的 CTU 风控大脑正在发挥作用。

CTU 风控大脑是目前蚂蚁金服重点研发的安全系统的代号，实际上就是现在火热的人工智能在支付安全上的应用，目的就是要实现密码即便被盗也有能力保护用户的资金安全。

简单地说，这个风控大脑通过对用户资料和交易行为等大数据的积累，包括用户账户资料、设备、位置、行为、关系和偏好等方面，对用户进行了系统性的长期信息识别，形成了用户的支付行为画像。

如果用户在某次登录或支付的过程违背常理或者表现异常，系统就会自动识别出来，对风险进行评估打分，会要求用户提供更多的资料来审核，甚至会直接叫停支付行为，从而保护用户的资金安全。

风控大脑技术并非未来科技，早已经被应用。据国外实验室测算，这个技术能让判断风险的成功率提升 7 倍，用了这个技术后，支付宝风控大概提升了 5 倍。案例表明，2014 年 6 月 7 日，主人接收了伪基站 10086 的短信，主动输入了身份证信息和银行卡信息，并中手机木马。

当日深夜，骗子结合上述信息，成功获取校验码后修改登录密码，并在广州某小区登录，之后又修改支付密码。接着，得意洋洋下单一台 iPhone5，打算用别人的钱，给自己换手机。

没想到，风控大脑直接判定交易失败，并对账户进行了限制。第二天，支付宝客服给用户打电话，确认用户账户是被盗了，并引导其重置密码，成功杜绝了一次可能发生的安全事故。

安全永远是相对的概念，而现在的网络支付的安全相比线下的钱包安全早已经超出了何止万倍，但道高一尺，魔高一丈，来自各种场景的威胁始终不会消除，安全防护也将是永恒的话题。作为用户，要提高安全意识，减少信息泄露的风险，而支付企业更是要通过技术升级与系统建设来构筑更为安全的防洪堤，在新时代用大数据的方式来保护大数据的安全。我们相信，只要我们不断进步，安全便会一直伴随着我们，支付安全也就能说到做到。

大数据让运营商成为旅游业的智囊

随着移动互联网的发展，围绕着每个用户的信息数据正在形成海量存储，而大数据的各种应用也随之成熟起来。在 2014 年的世界杯期间，很多机构都通过大数据来预测比赛结果，而百度在 2014 年的春节期间就联合央视推出了春运迁徙图。

中国是一个人口大国，围绕着各种假日经济，旅游业日渐壮大，但每当黄金周就会遭遇各个景点的人满为患，相关道路的拥堵更是让出行的人焦头烂额。虽然各种机构都在利用自己掌握的数据资料进行分析，对旅游景点的客流疏导做出贡献，但始终难有实质性的效果。分析原因，主要是因为一般的机构掌握的数据并不全面，也无法实时动态地采集到所有游客的即时信息。

国内外运营商都在进行大数据应用探索

现实中，有一个行业在大数据应用中具有得天独厚的条件，那就是通信运营商们。运营商们数以亿计的通信用户基数保证了数据的海量和多元性，这些数据还具有可持续性，运营商可以通过对海量数据的有效分析，精准、高效地为广大用户和社会各界提供产品和服务。

比如，通信运营商多年来都在全面采集用户各方面的通信使用信息，包括用户的个人背景资料、实时的移动位置信息，如今还可以获得更多的移动互联网应用情况，只要是加以合理利用，完全可以准确清晰地分析出行走路线、旅游偏好等，成为大数据应用的样板。

在这方面，已经有国际上成功的应用案例。据媒体报道，美国运营商 Verizon 公司在举世闻名的超级碗比赛现场进行观众分析，短时间

内就能得到用户的行为轨迹并对散场后可能的交通情况进行预测，交通部门由此作为依据进行应对，取得了满意的使用效果。比如，信息表明，在超级碗体育场内，从巴尔的摩来的粉丝人数是来自旧金山的三倍，这样的数据通过其他渠道很难获得，可运营商却是手到擒来。

国内，一些运营商也已经开展了类似的大数据应用探索，并逐渐开始展现出巨大的应用前景。比如，某运营商通过对某省内的旅游景区所覆盖的网络信令数据提取，结合云计算分析引擎，站在大数据的视角上，为旅游主管部门和旅游相关从业者的行业决策和运营规划提供了第一手数据支持。

旅游业有可能是运营商大数据应用的先行者

作为旅游主管部门、旅行社或者景区管理方，最关注的无非是四个问题：游客从哪里来？游客怎么来？游客去哪儿玩？游客怎么玩？

要想解决这四个问题，就必须掌握游客的行动轨迹信息，但游客在各个机构填写的表格信息实在有限且不一定按计划执行，家庭及朋友一起自由行的游客更是行踪难觅。在这种时刻，只有几乎能做到人手一部的手机信息能够帮上忙。

在运营商所做的旅游大数据应用案例中，大数据项目组通过对旅游景点的检测与统计，从海量人群中准确提取出游客的相关特征，并对游客人群进行跟踪分析，最终在四个维度上给出大数据洞察报告。

1．游客从哪里来

对于景区来说，知道游客的主要来源地非常重要，因为这将是下一步制订营销策划和广告目的地的重要依据。每位游客都有一部手机，即使不是本运营商的用户，也能够通过网间通信数据分析获知，因此，

通过对游客号码归属地的调查，获取游客来源信息，包括省内、省外或国外等，可以清晰列出到此旅游的大多游客的归属地。

2．游客怎么来

对于旅行社和景区来说，如果能知道游客喜欢出行的交通方式，就可以针对性地设计旅游产品，并在相关的交通工具上进行宣传推广。运营商大数据项目组通过对到访游客的行动轨迹追踪，包括经过的交通枢纽、火车站、机场等记录，游客移动速度的分析等进行综合比对，可以还原出游客到达的方式，比如是通过公路、铁路还是航空。

3．游客去哪儿玩

游客到了一个城市，都去哪些景点关系到景点的收益，也对地方旅游主管部门的管理和服务非常重要。在大数据应用中，通过对景点进行实时的人流量统计，得出每日人流趋势图，并给出游客达到峰值时刻的统计，以便健全景区安全预警机制，可以提前行动做好各种保障措施。

4．游客怎么玩

运营商大数据团队在实践中通过对到达游客的持续跟踪，统计出在单一景区内的游玩时长，并结合游客的上下游出行地点、每日游玩作息、特点、活动区域来分析归纳游客的旅游轨迹，以便旅游主管部门及相关从业者为游客制订更个性化的旅游路线套餐，提供配套的餐饮、住宿、娱乐一条龙服务。

从实践中可以看到，以上这四个方面的大数据分析的实用价值明显，下一步，正在积极引入更多部门的数据，如景区门票数据、机票、

火车票数据等，一旦整合到一个平台，能够实现多维度全视角的数据集成，一定会发挥出更大的作用。

当然，大数据的应用需要完善信息保护，这样的大数据应用都是不针对用户个人的，而是根据大量的不记名的用户数据做出的信息产品，对营销和服务具有极强的指导作用却不会涉及用户个人的隐私。这样的大数据应用已经成为旅游行业发展的贴身智囊，成为了旅游从业机构的必备工具，不可或缺。

第 3 章

七种必备的大数据思维

从 1–0≠8–7 开始说起

请思考一个问题，1–0 在什么情况下不等于 8–7？

1–0 ≠ 8–7

史书记载，宋太祖是个非常有心计的皇帝，杯酒释兵权的故事就发生在他身上，但却不仅只有这一次。有一天早晨，文武大臣都一个个地汇报自己的工作，接着退到殿外。走在最后的是后周老宰相范质，他现在仍是宰相。当范质快要走出殿门时，宋太祖突然传话，说："范老爱卿，请稍稍留步，朕有一事与你相商。"听到传话，范质转过身走回到殿上，重新坐到自己的宰相之座。原来，在中国古代，宰相的地位是很高的，可以和皇帝坐着说话。人们常说宰相是一人之下，万民之上的官儿，就是皇帝对宰相也是很尊重，也得让礼三分。因此在上朝君臣议事的时候，宰相可以坐着跟皇帝说话，而其他官员只能够站着。范质坐下来以后，宋太祖递给他一份大臣汇报的奏折，范爱卿，你看这事如何解决才好？范质接过奏折仔细地看了起来。这时宋太祖从龙椅上站了起来，向后宫走去。宰相范质看完奏折后，心里已经想好解决的方法，可是，左等不见皇帝出来，右等也不见皇帝出来，范质实在等不住了，就起身去找皇帝。这时，宋太祖走了出来，范质连忙要坐下，可是回头一看，椅子没有了。原来，趁范质起身不注意时，身边的侍卫悄悄把椅子拿走了。范质不知道如何是好，只得站着和宋太祖说话。以后再上朝，宰相也和其他大臣一样只能站着和皇帝说话，这一制度后来被各朝所沿用，宰相站了千年。

事实上，在宋太祖之前，历朝历代的宰相都拥有决策权、议政权

和行政权，只是在逐渐地变小。从宋太祖之后，内阁就变成了只是皇帝的参谋，决策权在皇帝，行政权在六部。并且，这以后的宰相往往“身份”低下，阁臣通过票拟制度取得相当于前代丞相的权力。而“票拟”即对诸多臣民奏章提出处理意见，以供皇帝参考。

由此，我们可以形象地比喻成 1–0 和 8–7 的变化，前者的变化是质变，而后者仅仅是量变，两个结果看似都是 1，可实际上却有天壤之别。就像有人说的，如果你给领导做 PPT，前后做了 8 个才被接受，这等于是被枪毙了 7 个，当然与仅仅做了 1 个 PPT 就被接受不一样。

我们还可以看一个例子，假设有两位客户经理。在年初的时候，其中 A 负责重点维系一个大客户，B 负责重点维系另外 8 个大客户，等到年底比较绩效的时候，发现 A 维系的大客户没有变化，可 B 维系的 8 个大客户跑掉了 7 个，也剩下了一个。这个时候，如果仅仅看最终的结果，两位客户经理都只剩 1 个客户，结果是一样的，可如果我们看看过程，当然是不一样的，而且是天壤之别。

所以，我们在分析问题的时候，不仅要看最终的结果，也要看其中的过程，即便结果一样，如果过程不一样，也不能得到一样的结论。有些时候，即便结果有差异，但过程却非常类似或一样，那么两者可能差异并不大。

当然，以上这个案例，我们还可以从绝对数与相对数的角度来分析。A 客户经理保有客户 1 人，保有率达到了 100%，可另外那位 B 客户经理的保有率只有 12.8%，应该属于要被辞退的范围了。

在中国通信业的历史上，2008 年是个具有关键意义的年份，在这一年运营商重组且发放了 3G 牌照，中国电信拿着从中国移动“赠送”过来的 500 亿元人民币购买了中国联通手中的 CDMA 网络及配套运营

系统人员，由此形成了在移动通信市场上的三强争霸。在这种情况下，如果站在中国移动的视角上，其竞争对手的客户群体并没有发生数量上的变化，但竞争对手却由以前仅有的中国联通，变成了中国联通与中国电信两家，数未变但却发生了根本性的质变。

蚂蚁金服的分析师也介绍过一个关于客户信用分的案例。有三位消费者，他们都得到了相同的芝麻信用分数，比如差不多 750 分，这是某一个时间点的状态。从这个时间点来说，他们三位状态是一样的，但是把这个时间轴放长，看一下是否有一些变化。结果，可以看到，在两个月以前，第一个消费者获得的分数更高，比如说有 850 分，后来发现他有一些问题，比如说他会在还信用款账的时候有一些延迟，所以他的分数在降低。第二位消费者，可以看到他非常稳定，波动也不厉害，他总是准时付账。第三位得到了一个新工作，有了一个很好的职务，他现在将他之前的债务都还掉，在还信用卡账的时候每个月都还，而且他的趋势是向上的。有了这样一个新的知识以后，我们还会把这三个客户相同的看待和对待吗？ 所以，分析问题，既要看结果也要看过程。

我们也知道，量的积累达到阀值就会催生质变。“不积跬步，无以至千里，不积小流，无以成江海。” 任何事物的运动变化，总是先以微小的、不显著的变化开始，经过逐步积累而达到显著的、根本性质的变化。在哲学上，就把事物这种逐渐的、不显著的变化叫作量变；而把事物显著的、根本性质的变化叫作质变。

话说一栋十余层的旧大楼要拆除。一群工人忙活了半个月，用各种办法破坏大楼根基，下了很大工夫。但大楼虽旧，却无比坚固，兀自纹丝不动。工人们无法，便放弃行动，去跟楼主磨工钱。大楼闲置

数月。一日，一农民在楼旁不远处放牛，手闲无事，拿起一块石头朝大楼掷去。只听“砰”的一声小响，大楼上的一块玻璃应声而碎，随即，“隆，隆隆，隆隆隆隆”，一阵轰响，这庞然大物竟然顷刻散架，哗啦啦地塌了下来！那农民做梦也没想到，一块石头竟把一栋偌大楼房给打垮了。

做分析，就是要从量变看到未来的质变，或者于量变不显著的时候就看到内在的质变。

统计，一门与赌博密不可分的技术

要做分析，自然离不开统计学，而统计学是建立在概率论基础之上的学科，与大数据实际上“格格不入”。

我们现在谈的“大数据”，如果非要找一个相对的词汇，应该叫作“抽样数据”，也就是说，大数据并不是强调“大”，而是强调“全”。

在古代，人们实际上也是在研究大数据，因为人们的认识水平有限，不可能收集和整理出太多的信息，在有限的信息时代，我们完全可以将所有的信息都放到一起进行分析，也就是所谓的大数据。后来，当信息越来越多以后，我们收集信息的能力远远超过了分析工具和分析人的能力，所以，人类就产生了一种新的科学分支，叫作“统计学”。

统计学是建立在抽样和概率基础之上的，是科学家面对太多的海量数据不得不做出的聪明的选择，仅仅牺牲掉了一点点的可靠性，就让我们的分析变得可行与简单，统计学给人类的科学认知带来的贡献是巨大的。

统计学的诞生一直在等待概率论的成果，而概率论又是从机会性游戏中酝酿而来，世界上第一部概率论著作就是《论掷骰子游戏中的计算》。那是在遥远的 1657 年，荷兰著名的天文、物理兼数学家惠更斯从掷骰子中总结出的规律，据说他是一个嗜赌如命的人。

很多书中是这样记述的：早在公元前 1500 年，埃及人为了忘却饥饿的困扰，经常聚集在一起掷骰子和紫云英，这是一种叫作“猎犬与胡狼”的游戏，按照一定规则，根据掷出各种不同的紫云英而移动筹码。大约从公元前 1200 年起，人们把纯天然的骨骼（如脚上的距骨）改进成了立方体的骰子，方法是摩擦骨骼使其成为一个粗糙的立方体，骰子的六面就形成了，再在骰子面上刻上不同的数字。它是游戏中常用的随机发生器，可能因为当时没有表示数字的符号或简单标记，早期骰子各面的数字都被刻成浅浅的印迹。现在相对面的数字之和是 7 的骰子大约产生于公元前 1400 年的埃及。在玩骰子游戏的几千年的时间里，概率理论的某些思想可能早应该出现了，但是一直没有迹象表明人们观察到赌博与数学之间的直接关系，甚至没有发现有人意识到骰子点数下落的频率的计算是可能的、有效的，或每一面会以相同的频率出现等这些最简单的概率思想的萌芽。可能是由于缺少完美平衡和“诚实”的骰子，因而阻碍了人们发现任何可察觉的规律；或者由于缺少适当的数学概念和符号而阻碍了数学的探索；缺乏刺激概率思想研究的经济问题。直到文艺复兴时期，随着阿拉伯数字和计算技术的广泛传播，简单代数和组合数学的发展，并且哲学的思想开始转变、拓展时，随机事件的试验和计算在本质上才有所进展，概率的思想才开始逐渐浮出水面。现在有史可查的对于赌博问题最早加以研究的是意大利。

最初人们研究的重点是赌博输赢的各种可能性或次数。卡尔达诺（Girolamo Cardano，1501~1576）是意大利数学和医学教授，他天资聪明，常常不循规蹈矩，有着有趣而丰富的经历。在他一生中超过 40 年的时间里，卡尔达诺几乎每天都参与赌博，在一本名叫《机会性游戏手册》（*Liber de Ludo Aleae*）的书中，他公布了关于赌博实践的体会。有人认为卡尔达诺可以被称为是“概率论之父”。后来伟大的天文学家伽利略也开始对掷骰子的问题进行数学化的思考。

但概率论成为一门研究随机现象规律的数学分支，其起源于十七世纪中叶，来自赌博者的问题，即 1657 年，荷兰的另一数学家惠根斯（1629~1695）写成了《论掷骰子游戏中的计算》一书，由此奠定了古典概率论的基础。几乎与此同时，瑞士数学家雅各—伯努利（1654~1705）建立了概率论中的第一个极限定理，我们称为“伯努利大数定理”，即“在多次重复试验中，频率有越趋稳定的趋势”。到了 1730 年，法国数学家棣莫弗出版其著作《分析杂论》，当中包含了著名的“棣莫弗–拉普拉斯定理”。这就是概率论中第二个基本极限定理的原始雏形。而接着拉普拉斯在 1812 年出版的《概率的分析理论》中，首先明确地对概率做了古典的定义。另外，他又和其他数学家一起建立了关于“正态分布”及“最小二乘法”的理论。另一在概率论发展史上的代表人物是法国的泊松。他推广了伯努利形式下的大数定律，研究得出了一种新的分布，就是泊松分布。概率论继他们之后，其中心研究课题则集中在推广和改进伯努利大数定律及中心极限定理。

概率论发展到 1901 年，中心极限定理终于被严格地证明了，后来数学家正是利用这一定理第一次科学地解释了为什么实际中遇到的许多随机变量近似服从正态分布。到了 20 世纪的 30 年代，人们开始研

究随机过程，而著名的马尔可夫过程的理论在 1931 年才被奠定其地位。而苏联数学家柯尔莫哥洛夫在概率论发展史上亦做出了重大贡献，到了近代，出现了理论概率及应用概率的分支，即将概率论应用到不同范畴，从而开展了不同学科。因此，现代概率论已经成为一个非常庞大的数学分支。

如果你对统计学的深奥理论不感兴趣，大概可以掠过上面这么长的文字，继续向下看，应该不受影响。但是，如果你想真正做好数据分析工作，了解以上这些人和这些基础理论还是非常必要的，因为他们都是统计学的基础。

学好统计学，还有一位地地道道的高人，那就是二战中赫赫有名的偷袭珍珠港的日海军大将山本五十六。这位山本在二战之前最出名的不是海军指挥才能，而是其精湛的赌博技术。有记载说，在 1921 年，山本五十六欧美之行时特意到赌城摩纳哥去了一趟，以试自己的身手。在大赌院里，山本旁若无人，赌技高超，每战必胜，令赌院老板大伤脑筋，不得不禁止其入内。据说他是自蒙特·卡尔罗赌场开设以来第 2 个因赌技高超而被拒绝入场的人。连一向出言谨慎的山本对自己赌博的才能也不无得意，曾向朋友夸口说：“如果给我两年的时间游遍欧洲各地，我能赚到建造一艘战舰的费用。”

山本对赌博不是一般的兴趣，而是拥有一套理论。他认为，赌博虽然与自己的物质利益有关，但不能与物质利益纠缠不清，如果纠缠在一起，判断就容易产生错误。正确的态度应是出于内而超乎其外。他常说赌博虽靠运气，但还必须进行科学的计算，仅仅热衷于胜负是没有意义的。如果用高等数学进行冷静的计算，会清楚地计算出每场必胜的结果。当然在必胜的结果来到之前，需要长时间的等待和忍耐，

这是非常重要的。在许多绅士和小姐陷于狂热的时候，自己要忍受周围人的白眼，冷静地等待经过计算的获胜时间的到来。冷静沉着是最为重要的。山本从他赌博的切身体验中学到了不少可用于战争的东西，赌博对他的思维和行动方式产生了十分重大的影响。他常说：“战争就是赌博。”

谁能说山本总结的这些赌博技巧不是战争指挥的谋略，谁又能说现在做的企业经营分析和市场分析不需要同样的理论支持呢？做数据分析，就要站在中立的立场，不可考虑自己的利益在其中，超脱的分析才可靠。做数据分析也必须耐得住寂寞，长期的跟踪与冷静的判断，有时候也要忍受领导或者同事的白眼，当机会成熟的时候，一个分析的结果也许就能改变企业的命运。

虽然统计学与大数据有一定的差异，甚至在基础理论上风马牛不相及，可统计学的一些方法还是可以在大数据分析中使用的。记住，不是全部。基础的统计分析，包括汇总、排序、集中趋势、离散程度、分布形状，也包括相关分析、回归分析、聚类分析、因子分析等，也都有用武之地。

我们还需要明确的是，大数据分析并不一定比统计分析更准确，因为大数据分析经常会遇到异常或特异情况的干扰，即所谓的一颗老鼠屎坏一锅粥。

串联，一种简单实用的日常分析法

把串联当成一种分析方法，以前的任何一本统计学的书都没有介

绍过。在此前的书本时代，串联分析虽然经常在日常生活中被应用，但要成型起来，还是在互联网流行之后。借助网络的记忆和分类功能，串联分析变得简单实用。

物理学上，串联（Series Connection），是连接电路元件的基本方式之一，指电路中的元件或部件排列得使电流全部通过每一部件或元件而不分流的一种电路连接方式。

在大数据分析的时候，我们借助某条线索，可以是时间轴，也可以是人物、地点或者其他具有关联性的事物，把很多相关的事项连接起来，从而让人有一目了然的结论。

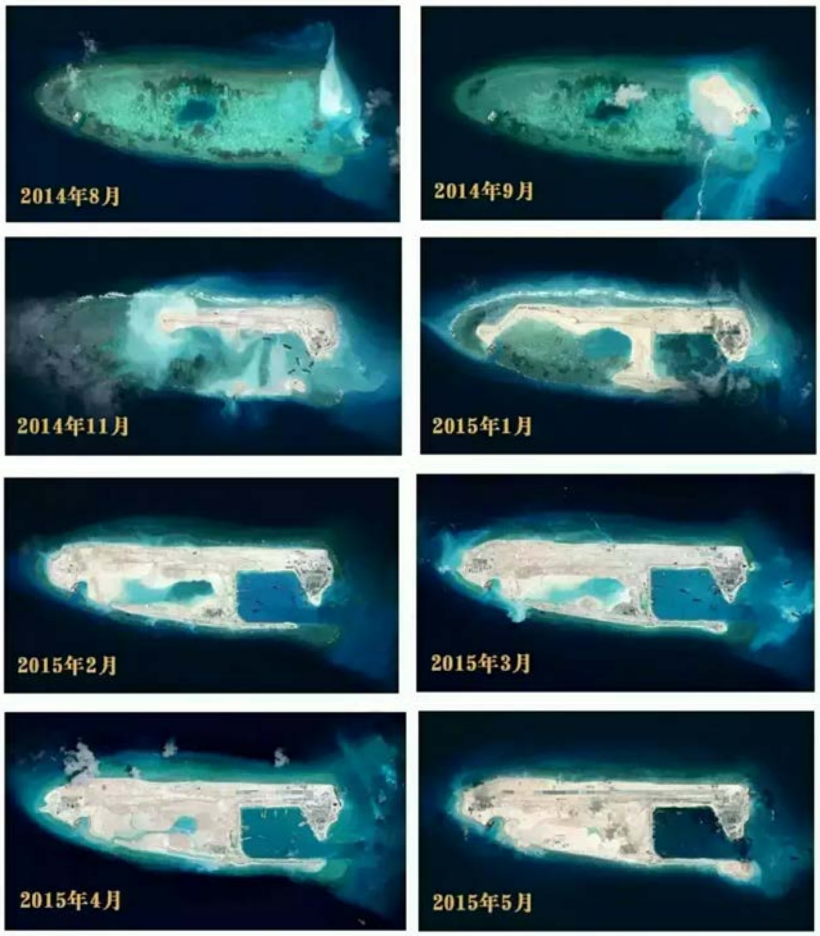
比如，中国的春运一票难求问题存在了多年，每到春节前后就是各种各样抢票排队的新闻，铁道部也做了各种努力，却一直也没得到根本性的解决。在网络上，有好事者将铁道部的新闻稿按照时间的先后排列在了一起，醒目地告诉了我们“真相”，铁道部从来说话都没有算数过。

串联的分析方法往往需要长期的资料积累，更需要专业性的知识储备，看似简单的逻辑使用起来却很难。

一位北京市民自 2013 年 1 月 27 日开始，坚持每天早晨在同一个地点拍摄一张照片，直观记录北京的天气情况，实际上也是一种串联分析。

不过，进行串联分析的时候，也需要遵循一定的规则，最好是有一条明确的串联线索，而不是多条交叉或者仅仅是“拉郎配”。

最后，我们再看看永暑礁是如何在短短九个月时间变身永暑岛的！永暑岛位于南沙群岛中部，距海南岛榆林港 560 海里，地理位置十分重要。如今，永暑岛已成为中国大陆实际控制的南沙最大岛屿，也是南沙群岛第一大岛。这仅仅是一个开始，它的面积还会继续扩大，并在机场等一系列基础设施上继续发展，将来可能成为新的度假圣地，因为它的面积早已经超过了马尔代夫首都马累。



对比，最常用也最实用的分析方法

对比是一种分析方法，也是最简单的分析。对于一个人来讲，生活中一直在进行对比，而工作中的分析，对比也是最常用的。古人有诗说，“横看成岭侧成峰，远近高低各不同”，实际上就是在不同角度的对比。

如果说到对比分析，一般是指，根据经济现象之间的内在联系，对相关指标进行对比，以分析其数量关系及形成原因的分析方法，是最基本的分析方法。在进行对比的时候，我们进行文字比较，也可以进行图片比较，还可以进行数字比较，当然也可以进行视频比较，方式方法略有差异。

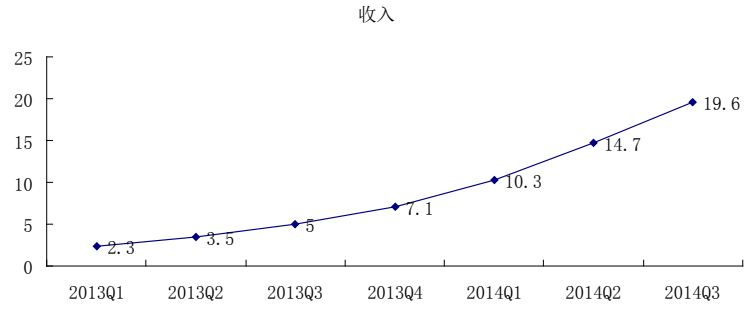
我们日常进行的比较分析，可以从与计划对比、与上期对比、与去年同期、与历史最好水平对比、与总体平均水平对比、与国际国内最好水平对比这些角度进行，也还要考虑绝对数与相对数的比较。

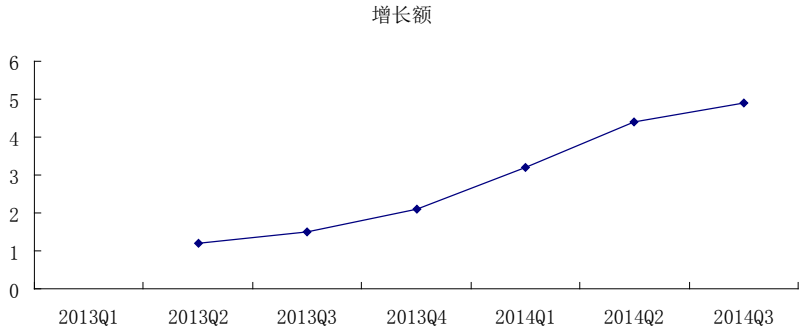
我们使用最多的比较是同比分析和环比分析。同比分析，一般是指本期水平与上年同期水平对比分析，如今年 12 月比去年 12 月。环比指报告期发展水平与前一时期水平之比，如计算一年内各月与前一个月对比，即 2 月比 1 月，3 月比 2 月，说明了逐月的发展程度。

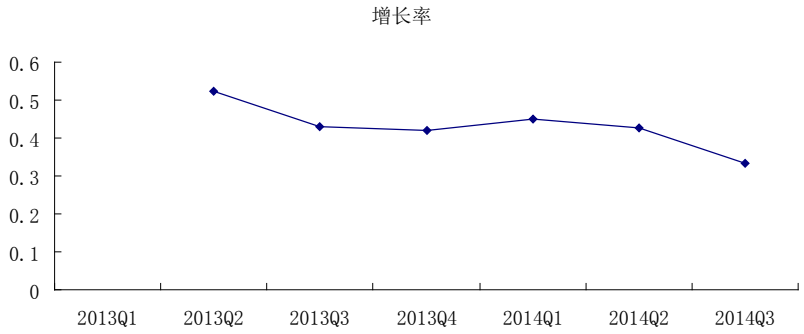
比如下面的例子，我们先看环比的分析，这里需要将收入量、增长量、增长率分别做出环比，然后综合起来做结论，否则容易走偏。

下面是一家公司的季度财报数据，进行一下环比分析：

	2013Q1	2013Q2	2013Q3	2013Q4	2014Q1	2014Q2	2014Q3
收入	2.3	3.5	5	7.1	10.3	14.7	19.6
增长额		1.2	1.5	2.1	3.2	4.4	4.9
增长率		52.17%	42.86%	42.00%	45.07%	42.72%	33.33%



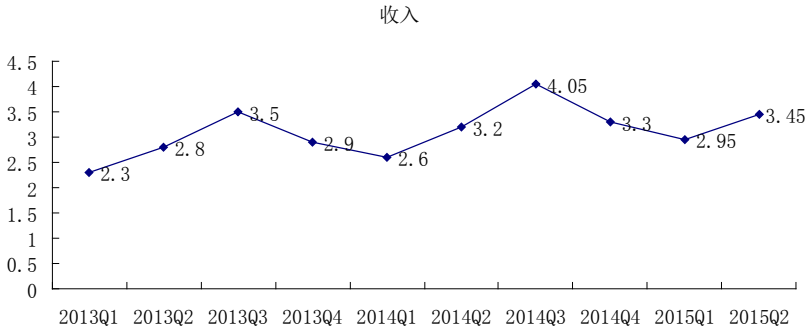


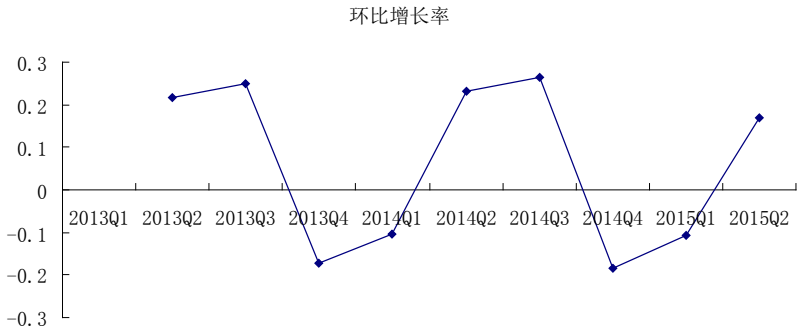


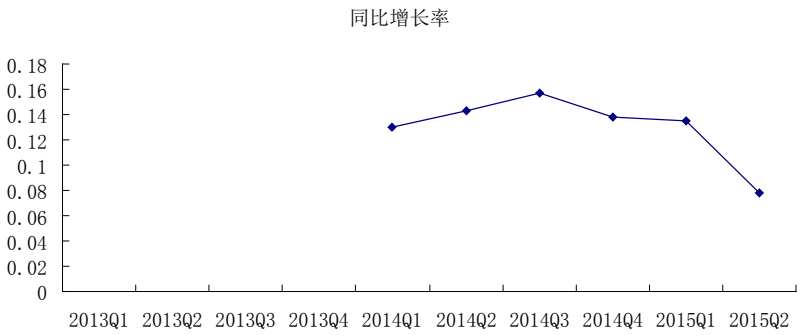
我们同时看了收入、增长额和增长率的环比数据，结论是，收入在稳步增长，增长额已经放缓，增长率在持续下滑，公司经济形势严峻。

接下来再看同比，同时也结合了环比的分析，结论更加可靠。

	2013Q1	2013Q2	2013Q3	2013Q4	2014Q1	2014Q2	2014Q3	2014Q4	2015Q1	2015Q2
收入	2.3	2.8	3.5	2.9	2.6	3.2	4.05	3.3	2.95	3.45
环比增长额		0.5	0.7	–0.6	-0.3	0.6	0.85	-0.75	-0.35	0.5
环比增长率		21.74%	25.00%	–17.14%	–10.34%	23.08%	26.56%	–18.52%	–10.61%	16.95%
同比增长额					0.3	0.4	0.55	0.4	0.35	0.25
同比增长率					13.04%	14.29%	15.71%	13.79%	13.46%	7.81%







综合起来看，这家公司的收入在两年半的时间内呈现波动向上的趋势，环比增长率波动非常大，但看同期的数据却是已经连续多个季度在下降，发展趋势并不好。

再来说说定基比。定基比发展速度，也简称总速度，一般是指报

告期水平与某一固定时期水平之比，表明这种现象在较长时期内总的发展速度。可以以历史最好水平做基数，也可以用历史上的平均水平做基数，还可以用你认为任何一个合适的数据做基数，进行对比。

接下来，在进行全面的对比分析之前，我们需要再多学习一点统计学常识，那就是数据的计量尺度。统计学依据数据的计量尺度将数据划分为四大类，即定距型数据（Interval Scale）、定序型数据（Ordinal Scale）、定类型数据（Nominal Scale）和定比型数据（Ratio Scale）。

定距型数据是数字型变量，可以求加减平均值等，但不存在基准 0 值，即当变量值为 0 时不是表示没有，如温度变量，当温度为 0 时，并不是表示没有温度，这样温度就为定距变量，而不是定比变量。

定序型数据具有内在固有大小或高低顺序，但它又不同于定距型数据，一般可以用数值或字符表示。如职称变量可以有低级、中级和高级三个取值，可以分别用 1、2、3 等表示，年龄段变量可以有老、中、青三个取值，分别用 A、B、C 表示等。这里无论是数值型的 1、2、3 还是字符型的 A、B、C，都是有大小或高低顺序的，但数据之间却是不等距的，因为低级和中级职称之间的差距与中级和高级职称之间的差距是不相等的，因此可以排序，但不能加减。

定类型数据是指没有内在固定大小或高低顺序，一般以数值、字符、文字表示的分类数据，如性别男和女。

定比型变量就是常说的数值变量，拥有零值及数据间的距离是相等被定义的，通常指诸如身高、体重、血压等连续性数据，也包括诸如人数、商品件数等离散型数据，也可以做基本运算。

从这四类数据出发，我们可以简单地理解，最好用的分析数据是定比数据，也就是那些连续性的数据变量，如收入、利润、用户数等，

而定类数据分析能力最差，一般只能进行类别之内的汇总，如果要跨类别进行统计，往往需要将不同的类别先综合成高一级别的大类，如统计男女合计就需要用“人”来当类别。一些定序数据分析能力也比较差，但我们可以人为进行赋值，让其具备一定的距离，比如客户满意度，我们把“非常好”定义为 5 分，“很好”定义为“4 分”，“一般”为“3 分”，“比较差”为“2 分”，“非常差”为“1 分”，这样我们就可以按照定距甚至定比数据来进行分析了。

我们来看一个全面的例子，这个是 2014 年能够进行当年世界 500 强的全球运营商的名单及相关信息。

排名	上年排名	公司名称（中英文）	营业收入（百万美元）	国家
34	34	美国电话电报公司（AT&T）	128752.0	美国
42	48	威瑞森电信（VERIZON COMMUNICATIONS）	120550.0	美国
53	32	日本电报电话公司（NIPPON TELEGRAPH & TELEPHONE）	109054.3	日本
55	71	中国移动通信集团公司（CHINA MOBILE COMMUNICATIONS）	107647.3	中国
99	105	德国电信（DEUTSCHE TELEKOM）	79829.0	德国
109	97	西班牙电话公司（TELEFÓNICA）	75752.0	西班牙
135	257	软银（SOFTBANK）	66546.0	日本
141	124	沃达丰集团（VODAFONE GROUP）	65986.5	英国
146	145	美国康卡斯特电信公司（COMCAST）	64657.0	美国
154	182	中国电信集团公司（CHINA TELECOMMUNICATIONS）	62046.8	中国
156	158	美洲电信（AMÉRICA MÓVIL）	61562.2	墨西哥
189	170	Orange公司（ORANGE）	54404.8	法国
210	258	中国联合网络通信股份有限公司（CHINA UNITED NETWORK COMMUNICATIONS）	49399.2	中国
249	233	日本KDDI电信公司（KDDI）	43258.0	日本
319	281	意大利电信（TELECOM ITALIA）	36493.4	意大利
325	289	法国维旺迪集团（VIVENDI）	35873.4	法国
379	386	DirecTV公司（DIRECTV）	31754.0	美国
421	394	英国电信集团（BT GROUP）	29051.1	英国
453	444	澳大利亚电信（TELSTRA）	26641.6	澳大利

在这张表格中，有多种类型的数据，比如排名是定序数据，国家是定类数据，营业收入是定比数据，我们如何针对这样一份综合数据通过简单的对比写出一份有分量的分析报告呢？

首先，我们在分析这样的综合案例的时候，需要找一个切入点，也就是分析问题的入口，以后我们在分析任何问题的时候都要这样做，不管数据有多少，第一步都是最重要的。

一般来说，分析问题可以坚持从大到小、从全局到局部的原则，第一步的分析可以采取的是大的全局性的概括。比如，分析中国的人口问题，首先要描述说明，中国人口的总数是多少。

在这里，我们首先要进行汇总，也就是说，在 2014 年，世界上有多少家电信运营商进入了世界 500 强。好的，我们数过了，是 19 家。

继续看，19 家电信运营商在世界 500 强里是多还是少呢？这个数据表太简单，不能告诉我们这个答案，但还是可以计算出来比例关系，电信运营商占到世界 500 强的 3.8%，如果能找到其他行业在世界 500 强中的占比，就可以看出电信运营商的实力了。

根据这个数据，我们也无法断定电信运营商是变好还是变坏，也就是发展趋势，假设我们能从其他渠道详细分析世界 500 强的榜单，找到 2013 年、2012 年等年度的世界 500 强的历史数据，计算一下前几年的电信运营商在世界 500 强中的比例，就可以看出世界电信运营的发展趋势。

所以，分析问题，不管多牛的分析师，都只能以做到手头的数据为最好，不可能超越其中去胡思乱想，但在很多情况下，需要补充外围的数据，这样分析出的结果才更有价值。

当然，如果精力足够，或者信息面知识面足够大，还可以借助其

他领域的信息对现有的分析进行弥补。

2015 年 2 月，奇虎 360 收购 360.com 的域名，收购价格为 1 亿元人民币。360.com 是奇虎 360 从沃达丰手中收购的。之前有消息称，小米 mi.com 域名的交易价格为 400 多万美元，折合人民币 2500 万元左右。到了 2015 年 12 月，中兴旗下的努比亚以 200 万美元买下 nubia.com 域名。看看 2014 年电信运营商的表现，沃达丰的排名从 124 降到了 141 位，连域名都卖掉了，其他不用多说了吧。

域名：360.COM
所属注册商：ENAME TECHNOLOGY CO., LTD.
Sponsoring Registrar IANA ID: 1331
域名服务器：whois.ename.com
相关网站：http://www.ename.net
DNS服务器：NS5.360WZB.COM
DNS服务器：NS6.360WZB.COM
域名状态：clientDeleteProhibited http://www.icann.org/epp#clientDeleteProhibited
域名状态：clientTransferProhibited http://www.icann.org/epp#clientTransferProhibited
更新日期：2015-02-03
注册日期：2000-03-07
过期日期：2016-03-07

这时候，我们已经知道了电信运营商的整体实力，接下来就应该从全局到局部，可以先看一个点，那就是这些入榜的电信运营商，哪一家的 500 强排名上升最多，简单地观察就可以知道是日本的软银，然后你就可以去收集软银的信息，看看他表现如此之好的原因是什么？

当然，做完整体的分析，我们还要对内部结构进行一些分析，对这些电信运营商的国别进行分类归总，美国 4 家入围，日本 3 家入围，中国 3 家入围，中美日作为电信运营的第一集团，三个国家一共有 10

家，对比总数的 19 家，已经占到了一半以上，剩下的多数是欧洲国家的电信运营商。这里，如果是电信业内人士，就可以分析下形成这种格局的原因了。

有人说，我们可不可以比较一下世界各国运营商的国家营业收入之间的实力对比呢？答案是，不可以。因为中国在 2014 年只有三家电信运营商，三家的收入之和等于中国的电信市场整体，可其他国家并不是，美国、日本、欧洲还有很多的小运营商没有进入世界 500 强，仅仅根据这个数据表计算不了总体，所以不能进行简单比较。

分析完世界范围之后，我们可以回落到中国范围内进行分析，也是一样的分析顺序。先汇总，三家电信运营商都进入了世界 500 强，然后排序，中国移动排名最高，中国电信次之，中国联通最后。

我们还可以发现，中国的三家电信运营商的排名全部都是上升，其中中国移动从 7 位提升到 55 位，提高了 16 名，中国电信提高了 28 名，中国联通提高的幅度最大，是 38 名。于是，我们可以说，中国联通比中国移动表现要好吗？当然不能，因为排名是定序数据，之间的距离并不一定是均等的，排名变化的多少一般来说是不能代表相互之间的差距的。我们都知道，考试时，从第 30 名提高到第 3 名，还是比较容易，但要是从第 3 名提高到第 1 名，恐怕就没那么简单了，除了努力，还需要天分和运气。但，三家电信运营商的排名都上升却是世界所有国家中唯一的，也还是从一定程度上说明了中国电信运营市场的一枝独秀。

如此，是不是所有的表上信息都分析完了？还没有，因为有一列非常重要的定比数据还没有用，运营商们的营业收入，虽然这个收入不能进行国家与国家之间的比较，但还是可以在中国电信运营商之间

进行比较。

我们可以看到，中国电信和中国联通的营业收入之和大概等于中国移动的收入，由此可以看出中国的电信运营市场的基本格局与不均衡。

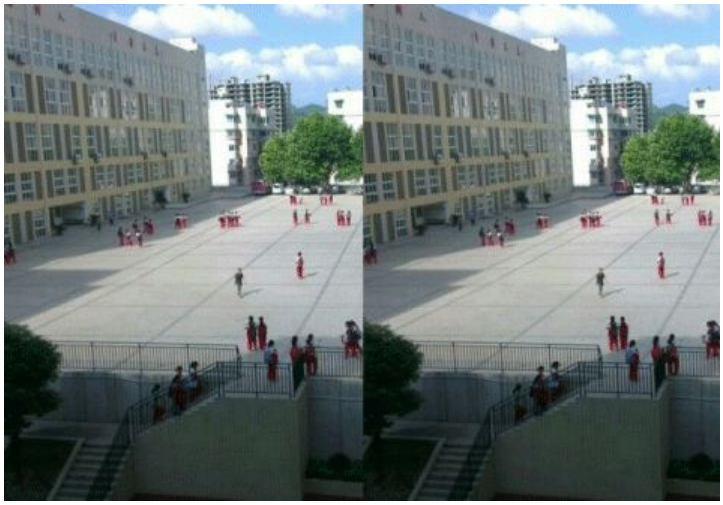
此外，三家中国电信运营商的营业收入之和大概是 2000 多亿美元，约等于 13000 亿元人民币，这个收入大概是什么水平呢？

中国互联网协会、工信部信息中心 2015 年 7 月 15 日联合发布 2015 年“中国互联网企业 100 强”排行榜。阿里巴巴、腾讯、百度、京东、奇虎 360、搜狐、网易、新浪、携程、搜房网位列百强榜前十位。百强企业 2014 年的互联网业务收入总额达 5735 亿元，占我国 2014 年信息消费总额的 20.5%。百强企业总体收入同比增速达 47%，带动信息消费增长 7.7%，贡献了 42.3%的信息消费增量。百强中有 71 家企业在全球各主要资本市场挂牌交易，4 家入围全球互联网前 10 名。

由此，可以清晰地看到，中国三家电信运营商的营业收入是中国互联网百强企业总收入的将近 3 倍，而互联网上所谓的 BAT（百度、阿里巴巴、腾讯）三家总收入也不过就是三家电信运营商的一个零头。虽然互联网行业营业收入增长迅猛，可这个差距在短期内都不会有多大的变化。

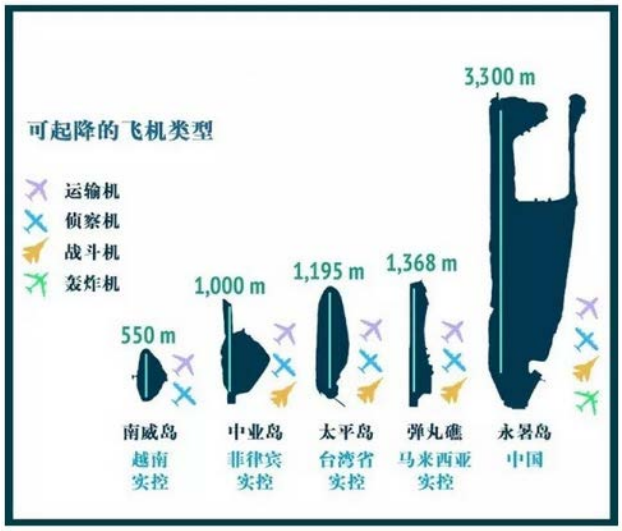
如果我们还继续去研究，也许从这张简单的表格中还会发现更多有价值的信息。在看到同样的数据信息的时候，每个人的分析能力不同，掌握的其他信息有多寡，直接会给分析的结果和数据的价值带来不可估量的影响。

至于图片比较，很多人小时候玩过找不同吧！好，先看图，不说话！



上面这张图有什么地方不同？

我们还可以把图做成这个样子直接比较。中国在南海建设了永暑岛机场，长度是 3300 米，这个长度有什么意义呢？



很多人即使看到永暑岛机场跑道如此长度，还是没有直观的感受，那就可以与首都机场的跑道比比看。

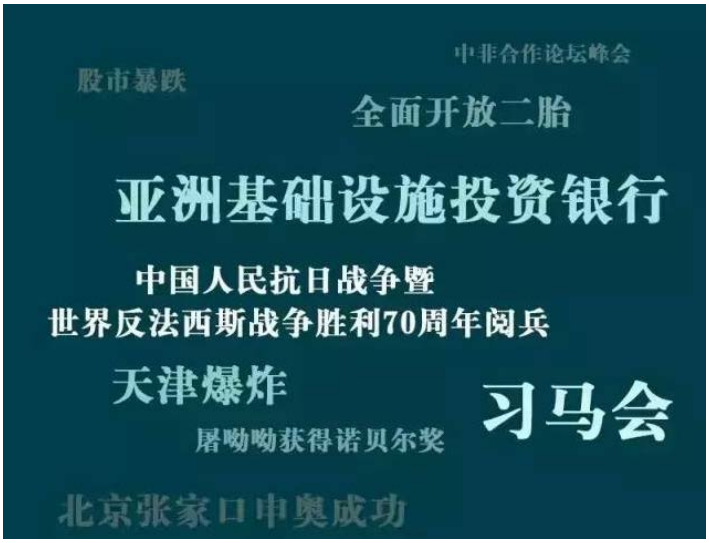
看完就明白了，永暑岛机场竟然比首都机场的一条跑道还要长，几乎可以起降任何飞机，当然，包括民航机也包括战斗机。

下面，最厉害的对比便是，12306 的验证码了，这一张是最简单的。

接下来还有难一点的。

文字也可以进行对比，比如进行文本分析。最简单的文本分析来自关键词统计，我们通过软件或者手工计算的方式进行词语频次的比较，往往能够发现很关键的信息。

中国人 2015 年最关注的词汇，请看下面。



当然，除了简单直接的计数，我们还要结合语境与文字的内涵来进行对比分析。下面给出一段文字，内容来自新华社的通稿，中国的电信运营商组建董事会制度的时候，董事长与总经理的分工如下：

> 中国联通董事长常小兵：（中国联合网络通信集团有限公司董事长、党组书记）负责公司全面工作，分管董事会办公室，人力资源部（高管人员部分）；中国移动董事长、党组书记王建宙主持公司全面工作。中国联通总裁陆益民（中国联合网络通信集团有限公司 总经理、副董事长、党组副书记）分管综合部、战略投资部，人力资源部（高管人员外的部分）、国际业务部、联通研究院、国家工程实验室；中国移动总裁、党组成员李跃主持公司生产经营管理工作，组织实施董事会决议。

有兴趣的读者可以分析一下，如果是用计算机，你能得到什么结论？不过，如果我们发挥人脑的作用，简短的一行文字就可以看出中国联通与中国移动在公司治理结构上的完全不同。

拆分，庖丁解牛之后的透视

先给大家讲一个挣钱的故事：

从前，有一个小城，小城里有一条街，街上住着一个叫“张三”的人。张三自己有10000元，通过街上“招财当铺”的搭桥，借给了同街的百年老字号“白家药铺”，白家承诺给他一年7.5%的回报，也就是说，一年之后张三可以从白家拿回10750元。不过，刚刚过了10天，张三的母亲生病急需支付医药费，张三找到白家药铺想把钱要回来，但白家说钱已被伙计拿着去外地买药，不能给付，要是现在要钱，不仅利息没有，还要扣30%的违约金，只能拿回7000元。张三觉得太亏，幸好有陶朱公路过，经高人指点，张三向街道附近其他闲杂人等借钱，承诺一年期的利息为6.0%，众人感觉高出银行太多，但质疑张三承诺的可信度，这个时候，张三拿出白家的借据，让“招财当铺”当场证实，于是众人你三百他五百正好给张三凑了10000元。转回来，张三找到“招财当铺”，和“招财当铺”商量好，到期的时候从白家和还款里直接支付给众乡邻，中间的7.5%–6.0%=1.5%利息差及张三投入的本金一并结算完成，而“招财当铺”也没白忙活，张三额外向“招财当铺”支付了0.1%的手续费。一年到期，众人在“招财当铺”亲手领回了“白家药铺”刚刚还回来的本金与6%的利息，大家皆大欢喜。

整个过程中，张三只用了几天的时间，就赚到了1.5%左右的收益，要知道，在银行这可是要一年的时间成本。从此，张三将拿回的钱再次投入，十天后再一个轮回，愉快地玩了起来，一年

后，他一共玩了 30 回，回报是（1+1.5%）的 30 次方，复利哦，

有兴趣的可以用科学计算器算算，高达 60%的收益啊！

很多人说，为啥众乡邻不直接投资给白家获取那 7.5%，非要让张三赚了一次差价呢？因为，对于白家药铺，向那么多人去直接借钱，太麻烦，也没有那么快捷，谁让张三钱多呢？同时，张三是白家药铺儿媳妇的哥哥的小舅子，人家是亲戚，或者比你消息灵通，或者人家比你胆子大。

整个事件，有风险吗？有，因为白家药铺有可能遇到强盗或者沉船导致血本无归，也就很可能没钱支付给张三，或者还账，而拿到了钱的张三是否会被众邻居起诉讨回欠款呢？还有，“招财当铺”要是卷款跑路呢，大家岂不都成了竹篮打水一场空。

最近一段时间，蚂蚁金融服务推出的招财宝平台受到了“小确幸们”的欢迎，因为其赚钱的方法就是让无数个“张三”实现了资金的高收益与流动性的兼得。特别是招财宝平台推出的万能险，一经面世就引发热捧，但由此也引来了各种各样的质疑，甚至有人夸大其词地将招财宝万能险描绘成“次贷危机”，事实真是这样吗？

万能险的投资风险并不高，比 P2P（网贷）风险要低得多

先要说明的是，只要你投资，就一定有风险。没有任何风险的投资品是不存在的。即便存在银行，也可能遭遇银行破产，万能险也是一样。

本质上，万能险是寿险的一种，至于万能险的投资风险高低，不在本文的详细讨论范围之内。根据一般规律，收益越高，风险越大，万能险一般只有 7% ~ 8%左右的收益，比那些动辄 15% ~ 30%的 P2P

收益要低得多，而万能险的发放公司都是财大气粗的大型保险公司，要通过投资获取 7%～8%这样的收益回报并不是太困难的事情。

再说一个基本事实吧！从 2000 年左右，万能险引入国内至今的情况来看，万能险大多通过银行、保险公司线下销售渠道来销售，相当于银行为万能险产品背书，所以银行不允许寿险公司的万能险产品出现未达计算收益率的情况。最近几年，寿险公司通过互联网渠道销售万能险产品，各大互联网金融平台都有销售，至今还没有出现过未达结算收益率的情况。如果还有质疑，可以看一下《保险法》的规定，“经营有人寿保险业务的保险公司，除分立、合并外，不得解散。”也就是说，经营寿险的公司想倒闭都不成。

当然，理论上讲，万能险的风险要高于货币基金，更高于银行定期存款，但收益也远远比这些要高。至于投资什么，就看个人的风险偏爱和资本情况进行选择了。

招财宝的万能险产品风险比直接在保险公司购买万能险的风险还要低得多

投资面临的最大损失，就是连本金都折在里面。如果招财宝的万能险投资也发生这样的情况，就意味着阿里巴巴关联公司蚂蚁金服旗下的招财宝跑路，售卖万能险的国内最大的几家保险公司也同时倒闭，投资人的钱就有可能颗粒无收，一去不返。实际上这种可能性根本不存在，原因根本不用解释。

首先，蚂蚁金融服务集团目前市场估值超过 500 亿美元，其旗下的支付宝占据中国第三方支付市场超过 70%的份额，关联公司的阿里巴巴市值 2300 多亿美元。在 2015 年，全国社保基金大手笔入股蚂蚁金服，中国邮政和中国邮储银行也入股。你觉得这样的公司会跑路吗？

其次，如果售卖万能险的保险公司连承诺利息都不能支付，投资人拿不回自己的最低 2.5%的回报。这种情况按照金融的逻辑是有可能发生的，但任何的金融投资都有风险，如果连 2.5%的投资风险都不能承担，就不要进入资本市场，而且，这样的承诺也是符合中国金融监管要求的做法。即便如此，招财宝的万能险还找了其他的财产保险公司或担保机构进行保障，保证用户的投资至少获得法定的承诺回报利息。

当然，这种 2.5%的收益甚至连本金都不能全额兑现的情况只会发生在“万能险公司超低的投资能力或投资失误”、“系统性金融风险（比如股市崩盘）”和“为此产品提供财产保险的那家保险公司或担保公司破产（众安保险或中投保等）”三者同时出现的时候。你觉得可能吗？

还有一种可能，如果售卖万能险的保险公司无力支付高额利息，投资人拿不回自己预期中的回报。比如，我们仍用上面的例子，有可能万能险最终的收益没有达到 7.5%，而只有 5%，那么，这个投资人其实并没有赔钱，只是赚得少了而已。

综合以上的分析，只要是投资了招财宝的万能险，2.5%的收益是有保障的，风险极低，比你在保险公司直接购买万能险或者通过银行购买万能险的风险还要低很多，因为增加了多重担保保障。当然，这只是在没有进行“变现”的情况下。

变现并没有增加风险，风险的高低与变现次数无关

招财宝的“变现”是一种具有划时代意义的创新，这种变现与金融领域的套利不同，在风险并不增加的情况下，实现了资金的快速流动，同时以市场化的方式实现了借贷双方的利率自由博弈。

我们知道，任何的定期产品，不管是银行的定期存款还是基金、保险、债券，一旦投入资金，必须要承诺在一定的期限内不能随意取回，因为这些钱被机构拿来投资，随意赎回会影响机构的投资行为和收益。所以说，定期产品是用损失流动性灵活性的方式来换取更好一点的回报。

简单地说，变现就是购买了理财产品的人通过招财宝平台拿回了自己已经投资出去的钱。投资到招财宝平台产品的人，购买了半年、一年或者两年期的理财产品之后，不用持有到期，如果需要把钱取回，可以使用“变现”的方式，向另外的投资人（在招财宝里预约或购买个人贷的人）借贷，等于是将贷来的钱转交给了自己要来投入资金的那家机构，换回自己的钱。

在很多人看来，这是多此一举，因为原投资人完全可以使用自己融资借贷来的资金，不用再去投资机构转一圈。但是，从事金融的人都知道，银行挣钱的奥秘就在里面，因为这里面有利差。

以上面的例子，我购买的万能险收益是 7.5%，但我“变现”的时候，也许市场的利率只有 6.0%，中间的差价是 7.5%–6.0%=1.5%，通过与投资机构、平台之间的“合作”，这个差价都成为了变现操作的收益。当然，还要支付一定的手续费。

通过前面张三的例子，我们也清楚，张三是有钱的，如果没有钱，是不能开始这个“游戏”的，也就是说，不管是谁接了张三的“盘”，都是有真金白银的货币作为“抵押”，而且，不仅有本金的抵押，连利息都是有抵押物的，因为张三的钱有白家药铺的投资回报支付利息，而且这个利息比张三去贷款的时候利息还多。不管张三向谁借款，最大的风险仍然来自白家药铺无法偿还，所以，不管怎么去变现，风险

都还是张三最初投资的那么多，并没有任何的增加。

实际中，在招财宝平台上，很多人买了万能险之后进行变现，而变现出来的个人贷被人购买之后可能再次选择利率低的时候进行变现，从而出现了下一个接盘者，以此类推。但是，不管多少次的变现，万能险的产品收益是 7.5%，这是所有人的收益总和，多次多人变现只是将这个总收益进行更细的拆分而已。

比如，甲购买的万能险收益是 7.5%，在 6.0%时变现，购买了 6.0%这个产品的乙在 5.0%的时候又进行了变现，丙购买之后持有到期。结果，甲拿走了差价 1.5%，乙拿走了差价 1.0%，丙在到期的时候拿到了剩下的 5.0%，其中，甲和乙各自支付了 0.1%的手续费，合计 1.5%+0.1%+1.0%+0.1%+5.0%仍然是 7.5%（这只是简单算法，因为变现金额上的差异会有误差，但总和不变）。

这个游戏的前提是不管多少次变现，投资的都是同时到期，也就是等于多人接力完成了万能险的长期持有。事实上，只要是仔细的关注自己的招财宝投资合同，投资时间长短差异显著，这也是招财宝的明示的投资期限上写“半年以内”、“一年以内”的原因，因为具体的投资期限要看变现的持有人的接力时段。

变现之后的重复理财实际上正是金融风险控制上的分散投资，等于是在降低风险

一些人认为，如果买入了万能险，不进行变现操作，风险是可控的，一旦进行变现，就等于是找人借款，而如果借来的款承诺出去的利息高于自己最终从万能险那里得到的回报，岂不就是亏， 如果你亏了，就没有钱给接盘侠，风险就会成倍的放大，甚至会发生“次贷危机”一样的多米诺骨牌效应。

理论上讲，虽然招财宝的风险只以最初投资者的风险为限，也就是说，招财宝提供的是保本保息的产品，但因为万能险只保证年化收益 2.5%或 3.5%（政策红线），高于此利率的变现，确实存在“入不敷出”的可能。比如说，你买入的是一年期历史年化收益为 7.5%的万能险，在“市场”利率为 6.0%时“变现”，那么 6.0%～2.5%中间的这 3.5%是有可能在极端情况下亏掉的，甚至还要赔上变现手续费。这种极端情况指的是，售卖万能险的公司真的无法提供已经承诺的所谓“历史年化收益率”（前文说了，十五年来从来没有出现过）。

即便这种情况真的出现了，那么“变现”也比不“变现”要合算。对于变现方来说，如果不变现，即便万能险真的只能给最低保证的 2.5%收益，也是赚的，可一旦变现就有可能出现“亏损”，但如果算上变现之后的再投资，而且是基于复利的再投资（你变现之后所拥有的下一次投资的金钱比上一次多了），只要接下来的投资获利，不仅可以抵偿前面的损失，还可能因为投入更高利率而略有盈余，比傻傻的放在那里只收到 2.5%还要好得多。即便一个万能险发生了亏损，另外的那些可以盈利，等于是多次分散投资，这正是投资上最好的抵御风险的方法。

对于接盘人来说，因为变现的人都是有真金白银的资金放在招财宝平台作为抵押，这个抵押物不会贬值（非货币贬值），并不会发生房屋贬值造成的所谓“次贷危机”。而且，招财宝为了保证每一次变现的完全履约，要求且强制变现用户在变现操作的同时去购买财产保险，为此要支付 0.1%的保险费，也就是说，万一有人真的因为“个人原因”而无法支付财产，保险公司会替其支付。这样，招财宝就把变现过程中的风险也采取了保障措施，接盘的人安全无虞。

招财宝实现了多方共赢，真正的输家只有银行

在招财宝的平台上，卖万能险的公司、购买万能险的人、变现接盘的人、再变现的人、再接盘的人、招财宝平台、财产保险公司这些参与者都是获益者。

卖万能险的人寿保险公司是最大的受益者，因为他们摆脱了以往长期主要依赖银行营业厅代销的困境，融资成本大幅降低，以 20 亿元的万能险为例，如果在银行的渠道卖，不仅佣金高，而且需要大概半年的时间，而在招财宝的平台上大概只要半个月就可以销售一空。这种融资能力和融资成本的大幅降低，让保险公司在 2015 年的股灾中得以频繁举牌抄底了一批最具有投资价值的大盘蓝筹。

购买万能险的人也是受益者，通过变现操作，在很短的时间之内就获得了相当高的收益，即便按照 1%的差价来变现，一年也可以有 30%的收益，即便是长期持有，也可以在有资金需求的时候随时变现出来，资金的流动性和收益都很好。

购买了变现后的个人贷的人也是受益者，也许是小额资金无法投资更高收益，也许是信息不对称，总之，这些人投资了比其他投资渠道更安全，收益却要高得多的理财产品，何乐而不为？

至于招财宝平台，每一次的用户变现，都要收一笔手续费，当然幸福。财产保险公司将最安全的一种保险卖掉了，而且费率不低。这两家是躺着挣钱的，当然是开心得很。

很多人有疑问，所有的参与者都在赚钱，到底是谁在赔钱呢？可以这样说，按照互联网的逻辑，参与者都成为了共赢者，谁没有参与，谁就被抛弃了，谁就是倒霉蛋。在这里，银行被排除在外了，而此前，万能险的销售主要依赖银行，不管是变现人还是接盘的人，中间多出

来的钱就是从银行嘴里虎口拔牙的结果。以前万能险的卖出价到用户拿到的收益之间的那部分全都是银行的收入，现在，招财宝联合大家给分了。

通过以上的案例剖析，我们终于明白，建设银行的行长和总理说“银行是弱势群体”，并不一定是玩笑话，因为，这样的互联网金融的创新确确实实是已经将银行变成了弱势群体。对于传统的银行来说，逆水行舟，不进则退。

最后，我们来回答最前面的问题，为什么不是所有人都去买那个高收益的 7.5%产品却去接别人的盘呢？

一是，更多的人不知道还可以买到 7.5%的产品，因为这些所谓的高收益产品并不是放在所有渠道的，最初的时候，这些 7.5%的高收益产品只会放到支付宝的 PC 页面，你想想自己有多久没打开过支付宝 PC 页面了，这就叫信息不对称。

二是，一些人风险意识很强，在招财宝里，那些收益相对低一些的个人贷，平台是通过第三方的担保机构进行“保本保息”的，而收益相对高的万能险、债券等标明的是“保证本金”，简单的差别就让很多人望而却步。

三是，一些人的钱实在是不多，即便是投资以 1000 元起步，也有人达不到，或者放到余额宝里的钱不是很多，买不了更高利息的理财产品。

综合起来看，大概也就是有这三种原因。还是金融赚钱的三件法宝，有信息赚没信息，风险高挣风险低，有钱的比没钱的赚得多。

在这个案例里，我们使用了拆分的分析方法，将一个产品的前前后后、里里外外，抽丝剥茧地进行了拆解，解剖一只麻雀一样地将事

物分析明白。

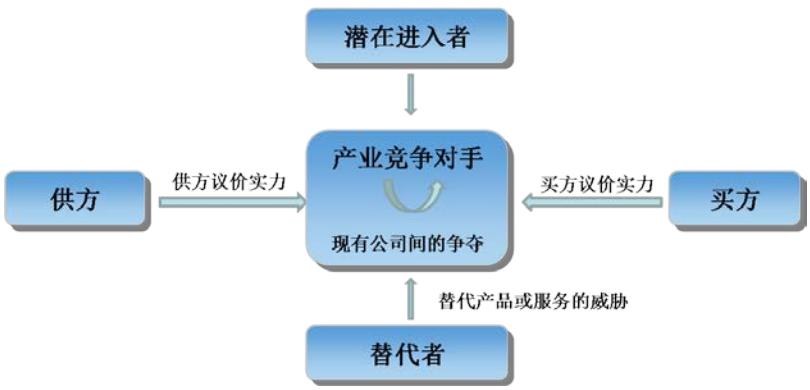
合成，组合起来的魅力

合成，汉语词典的解释，由几个部分合并成一个整体，或者，通过化学反应使成分比较简单的物质变成成分复杂的物质。

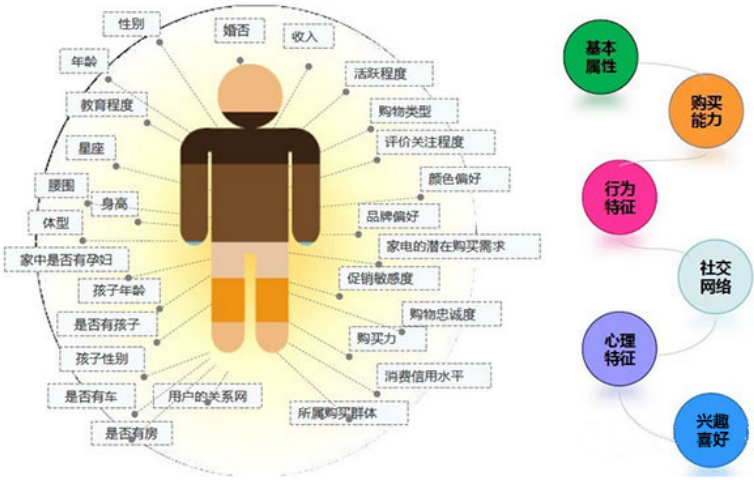
在数据分析上，也有很多合成的例子。任何一家公司和产品都面临多方面的竞争，在分析公司的竞争环境时就需要合成。

五力模型是由麦克尔·波特（Michael Porter）于 20 世纪 80 年代初提出的用于竞争战略分析的模型，可以有效地分析客户的竞争环境。

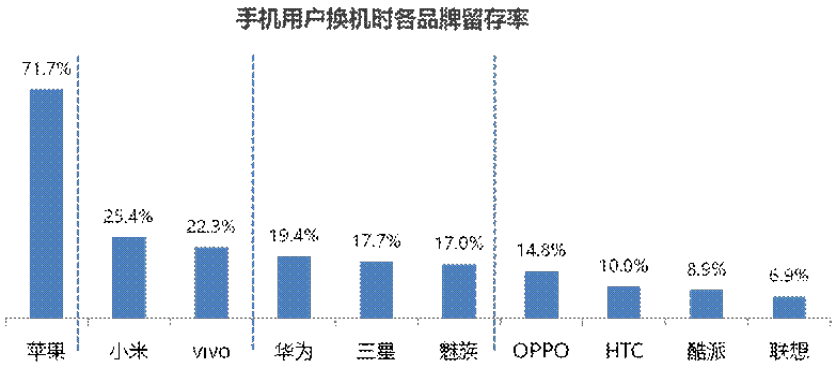
五种力量模型将大量不同的因素汇集在一个简便的模型中，以此分析一个行业的基本竞争态势。五种力量模型确定了竞争的五种主要来源，即供应商和购买者的讨价还价能力，潜在进入者的威胁，替代品的威胁，以及最后一点，来自目前在同一行业的公司间的竞争。一种可行战略的提出首先应该包括确认并评价这五种力量，不同力量的特性和重要性因行业和公司的不同而变化。

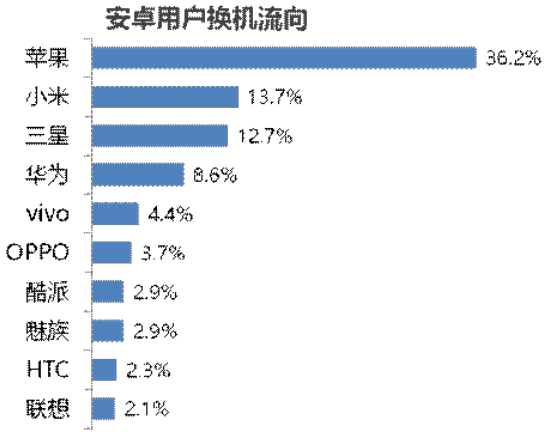


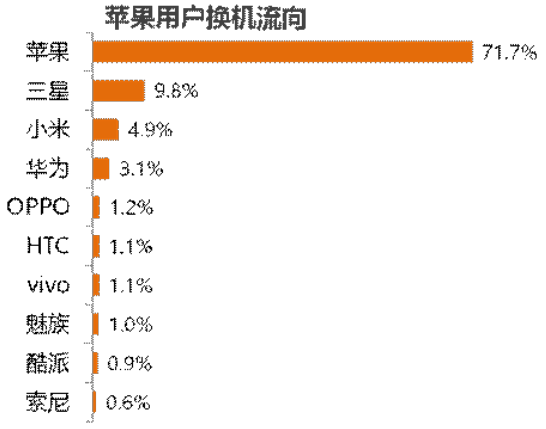
还有一种情况是需要合成的，比如在大数据分析上的数据全景视图。京东是一家大型全品类综合电商网站，海量商品和消费者产生了从网站前端浏览、搜索、评价、交易到网站后端支付、收货、客服等多维度全覆盖的数据体系，已经形成一个储量丰富、品位上乘且增量巨大的数据金矿。从用户画像分析来看，就是在解决把数据转化为商业价值的问题，就是从海量数据中来挖金炼银。这些以 TB 计的高质量多维数据记录着用户长期大量的网络行为，用户画像据此来还原用户的属性特征、社会背景、兴趣喜好，甚至还能揭示内心需求、性格特点、社交人群等潜在属性。了解了用户各种消费行为和需求，精准刻画人群特征，并针对特定业务场景进行用户特征不同维度的聚合，就可以把原本冷冰冰的数据复原成栩栩如生的用户形象，从而指导和驱动业务场景和运营，发现和把握蕴藏在细分海量用户中的巨大商机。



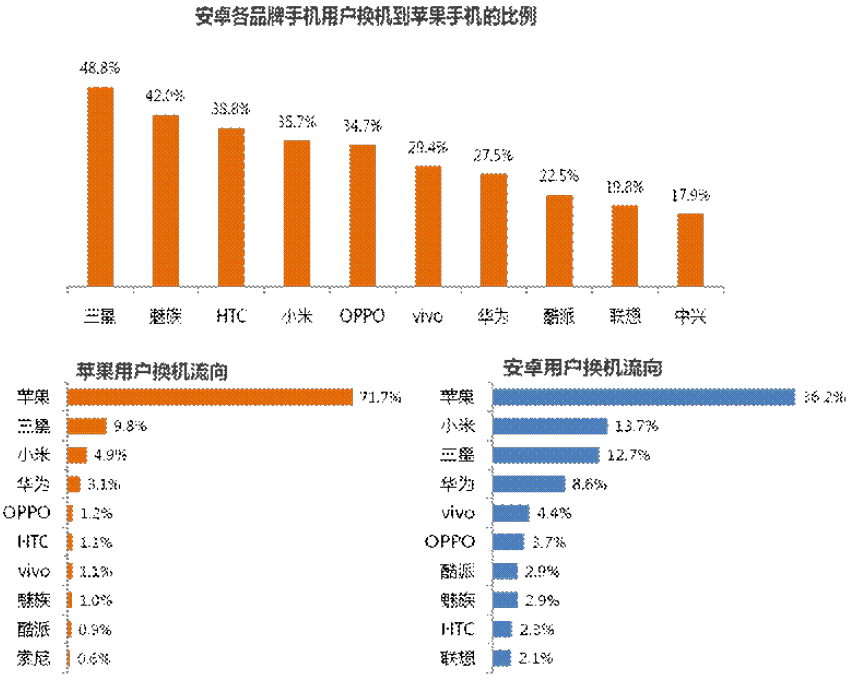
当然，一些简单的数据分析也需要通过合成的方式来进行。以下数据是新浪微博在 2015 年年初发布的智能手机微报告，三张图呈现的是手机用户换机的时候进行的品牌选择。


手机用户换机时各品牌留存率
71.7%
25.4%
22.3%
19.4%
17.7%
17.0%
14.8%
10.0%
8.9%
6.9%
苹果
小米
vivo
华为
三星
魅族
OPPO
HTC
酷派
联想


安卓用户换机流向
苹果 36.2%
小米 13.7%
三星 12.7%
华为 8.6%
vivo 4.4%
OPPO 3.7%
酷派 2.9%
魅族 2.9%
HTC 2.3%
联想 2.1%


苹果用户换机流向
苹果 71.7%
三星 9.6%
小米 4.9%
华为 3.1%
OPPO 1.2%
HTC 1.1%
vivo 1.1%
魅族 1.0%
酷派 0.9%
索尼 0.6%

分开看，我们也许只能得到简单的结论，如果我们把几张图放到一起，也许就能解释更复杂而全面的问题了。比如，我们可以看出三星为何衰落了。



逻辑与反证，大视野大转换下的推理

下面测试下，你的逻辑分析能力，看看你是否有炒股票的天份？这是最近非常火爆的一组漫画，作者是米洛斯拉夫·巴尔塔克（捷克），没有标题，你可以试着理解下作者要表达的意思。

分析问题，很多时候需要遵循基本的逻辑，还需要从更广阔的视角来入手，或者需要转换问题到另外的层面或领域，由此看清事物的本来面目。

据说，在移动互联网的江湖中，得入口者得天下，所以微信号称

获得了第一张门票，而在入口中又以支付为近水楼台，移动支付涉及众多应用场景，掌握着众多用户的支付数据，使用频率虽低却笔笔倾心，所以，移动支付也就成为了 2015 年微信与支付宝两强争夺的焦点。

当然，移动互联网时代躺着挣钱的机会并不多，既然移动支付这样重要，掌握网络的运营商、掌握终端的手机企业和那些仍有恃无恐的银行，都不会眼睁睁地看着 BAT 们垄断移动支付的江湖，并且有联合起来逆袭的可能。

媒体报道，苹果和三星几乎同时公布了各自的支付服务 Apple Pay 和 Samsung Pay 即将正式进入中国的消息，而且，中国银联与本土 15 家银行将与这些“踢门而入的野蛮人”达成合作联盟关系，这架势真有点“联盟拒曹”的感觉，也可以看作是银行系借助明星终端企业外力争夺市场的最后一战。

不过，从目前的形势来看，这种在移动支付上的合作冲击已经不占任何优势，很难撼动互联网企业的移动支付江山。

天时不再，错过了移动支付成长的黄金时期

毫无疑问，2014 ~ 2015 年是移动支付发展最为关键的两年，因为这两年是中国 4G 网络建设和大屏智能手机普及的最重要时期。数据显示，2014 年中国移动互联网市场规模为 2134.8 亿元，同比增长 115.5%，移动互联网接入流量消费达 20.62 亿 GB，同比增长 62.9%，2015 年这一数字会更惊人，4G 和智能手机的结合催熟了此前千呼万唤也难以普及的移动支付。

正是在这样的背景下，中国的大型互联网公司抓住机遇，采取抢红包、打折扣等方式将移动支付发展到了新高峰，2013 年、2014 年的

行业增速分别达到 800%和 500%。在 2015 年 2 月中国人民银行公布的数据显示，2014 年，全国共发生电子支付业务 333.33 亿笔，金额 1404.65 万亿元，其中，移动支付业务 45.24 亿笔，金额 22.59 万亿元，同比分别增长 170.25%和 134.30%。据易观智库发布的《中国第三方移动支付市场季度监测报告》显示，截至 2015 年第二季度，移动支付的交易规模达 34625 亿元，首次超过 PC 端的 32588 亿元。

与此同时，从 2015 年市场数据来看，中国的互联网巨头已经牢牢占据市场主导权，蚂蚁金服（阿里巴巴系）、支付宝和腾讯的财付通两家企业共占据了超过 90%的移动支付市场份额，支付宝一家甚至占到了 70%以上。反观中国银联，2013 年占据第三方支付平台份额 40%，现在却只有 9.2%。在一个用户使用率已经很高的市场，用户已经有了形成习惯的支付手段和品牌，想虎口夺食，难度可想而知。

中国银联当初是自己自绝于移动支付之外的，前有阿里巴巴主动上门寻求合作而顽固不化拒绝接受互联网，后有中国移动为代表的三家运营商合作联盟意向却逡巡不前达成双输，等到这个市场已经被互联网企业捷足先登之后寄希望于国外终端企业的实力来争食蛋糕，已经太晚。

在移动支付上，有三种类型，第一类是移动互联网远程支付，用 APP 实现手机端转账、消费等功能，第二类是 O2O 支付，基于移动互联网的交互技术，使用二维码、蓝牙、手机刷卡器、刷脸等支付技术实现支付功能，这两种都被互联网公司占尽优势，且已经形成了规模化。第三种类型是 NFC 近场支付，由银联、银行和移动运营商主导，虽号称技术先进却因为产业链复杂和内耗不断而市场惨淡，大概只占移动支付市场的 6%。

有一条互联网的发展规律，已经得到广泛认可。一种技术是不是先进，不是设计与生产厂商说了算，而是用户和市场说了算，即便是看起来在技术上落后的一方，只要先入为主形成规模，就会拥有成本优势，也会掌握主导权和话语权，后来的所谓先进技术也无法立足，最终往往被这些先导厂商在技术升级的过程中吸收消化掉。苹果三星用 NFC 连美国韩国本土的市场都没有形成优势，要靠什么来敲开已经在移动支付领先全球的中国互联网市场？

地利缺乏，移动支付场景和习惯已成

从金融发展的历史来看，支付从来都不是凭空产生的，缺乏场景的金融都是空中楼阁。在互联网金融的发展过程中，电子商务的发展促进了互联网金融的诞生，而移动端的业务增长才催生了移动支付的火爆。

互联网金融中，阿里巴巴的电子商务孕育和培养了支付宝，腾讯的微信和游戏让其支付站稳了脚跟，最近百度又借助智能设备和 O2O 机会杀入移动支付市场。这两年，人们通过打车软件、团购送餐、停车、酒店门票、便利店超市购物等熟悉了移动支付的操作，也享受到了移动支付的好处，使用习惯已经形成。

实际上，觊觎移动支付蛋糕的并非只有终端企业，拥有更大入口优势的运营商早在互联网公司行动之前就将其列入重点业务，经过数年奋斗却一无所获，原因也是仅仅有支付，缺乏与支付相关的场景土壤。

苹果和三星拥有众多的终端用户是事实，银联和银行也拥有比支付宝们更多更稳定的银联卡商户也是事实，但这并不等于拥有强势的

“地利”，也不等于有足够的支付场景。移动支付是从小额支付开始的，也是依托于现在快速发展的 O2O 业务，可这些正是银联的短板，即便最近银联也在改变自己，可效果并不明显。

这种合作依然是此前与运营商合作的翻版，造成支付两端用户分属不同公司，支付用户是苹果三星的，接受商户是银行和银联的，怎么与辛辛苦苦将两端彻底打通的互联网企业相提并论。可以说，这样的合作关系，只是会给用户提供一种新的支付选择而已，无论是营销效率还是使用效果，都会反差很大。

人和难望，国产终端强势政策支持难觅

苹果、三星选用类似的技术，联合中国银联、中国 15 家银行，这样的合作阵容强大，但在实际中却仍然只是 NFC 近场支付产业链的一部分，还缺乏 NFC 产业链中非常重要的移动运营商、应用开发商、系统集成商、商户及移动终端用户，庞大的产业链一直是 NFC 无法做大的根源，苹果和三星也解决不了。

苹果和三星自然是现在世界上炙手可热的手机企业，移动终端用户非常多，在中国更是有着强大的品牌影响力，可如今的国内智能手机市场已经是群雄逐鹿，华为、小米等企业已经分庭抗礼，用户数增长迅速。在这种情况下，苹果和三星号令商户和消费者的能量已经大不如前。在这种情况下，面对几乎可以覆盖全用户的支付宝、微信与只能覆盖一小部分用户的苹果三星，商户资源会更偏向谁呢？

金融是关系到国计民生的关键领域，有着各种各样复杂的进入限制，现在的终端企业又不是简单的售卖终端，而是会通过云计算掌控用户的各种行为数据，苹果和三星这样的国外终端企业注定没有占据

中国移动支付市场的可能。在合作中，中国银联和中国的这些银行如果让渡过多的敏感信息和资源，势必遭受监管，这些中国的金融机构也肯定不敢“引狼入室”，所以，苹果和三星的移动支付战略在中国一定是雷声大雨点小，如果能实现重度参与已经算是胜利。

中国的移动支付市场已经被 BAT 占据主导，中国银联和运营商已经错过了发展良机，苹果和三星的加入可以更进一步的催熟市场，也会给中国老百姓一种新的支付选择。虽然苹果和三星已经没有能力改变现有格局，但蛋糕会做得更大，所有的参与者都将是获益者。

京东净营收双降，危险真的降临了吗

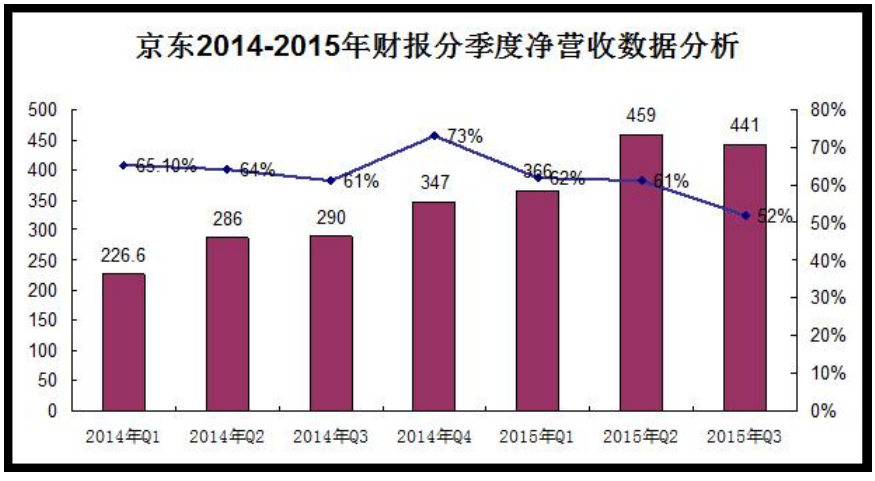
媒体报道，京东发布了 2015 年第三季度的财报，其中，2015 年第三季度交易总额（GMV）达到 1150 亿元人民币（约 181 亿美元），同比增长 71%。净收入为 441 亿元人民币（约 69 亿美元），同比增长 52%。第三季度归属于普通股股东的净亏损为 5.308 亿元人民币（约 8350 万美元），净利润率为–1.2%。第三季度完成订单量为 3.297 亿个，同比增长 85%。2015 年第三季度通过移动端渠道完成订单量约占总完成订单量的 52%，同比增长超过 210%。

正如京东官方所说，以上数据都很亮丽，“京东日益成为中国消费者无忧网购和快速履约的首选平台，2015 年第三季度，公司的业绩仍然保持强劲增长。京东在 2015 年三季度保持了健康的收入增长，用户数增长喜人，各品类业绩表现良好。”

净收入增幅增速首次双降，不可能只是因为“6・18”大促

但是，我们也从数据中看到了京东没有公开分析的另一面。数据显示，京东的收入出现了公司成立以来的首次负增长。京东第三季度净收入 441 亿元人民币，比第二季度 459 亿元人民币出现下滑，这是京东历史上第一次。按照京东的解释，主要由于 6 月份有“6・18”大促的因素，但是环比下滑体现出了京东增速正在放缓。

对比历史数据，京东在 2014 年的第三季度也曾经出现净收入增速的下降，从第二季度的同比 64%下降到 61%，降低了 3 个百分点，但是净收入仍然从 286 亿元提高到了 290 亿元。也就是说，同样有“6・18”的原因，去年的第三季度实现了净收入增长，2015 年却导致了收入下降，原因一定并非只是“6・18”那么简单。



平台数据首次超越自营，京东自营模式遭遇挑战

财报数据显示，2015 年第三季度线上自营与第三方平台核心交易总额分别为 613 亿元人民币与 497 亿元人民币，比 2014 年第三季度分别增长了 52%和 121%。2015 年第三季度线上自营业务的净收入同比增

长 48%，来自于服务项目与其他项目的净收入同比增长 111%，增长动力主要来自快速扩张的京东商城第三方开放平台业务，广告服务及向第三方商家提供的物流服务。

以上数据说明，京东平台部分的比重首次超过了自营。这几年来，京东自营部分比重持续下降，第三方的比例终于在 2015 年第三季度超过了 50%，达到了 55.23%。

更严重的是，京东赖以生存的数码 3C 销量也出现了负增长。数据显示，自营业务 GMV 第三季度为 613 亿元而第二季度 647 亿元，缩水 34 亿元，这也是京东历史上的首次。其中，第三季度 3C 销量 568 亿元，比第二季度销量的 590 亿元减少了 22 亿元。这应该受到了国内低迷的经济和 3C 整体下滑的不利影响，但由此给京东造成的整体营收伤害却不容小视。

如果这样的增长持续下去，京东以往所对外宣称的自营电商模式将遭遇前所未有的挑战。由此，我们就理解了刘强东为何在发布财报的时候没有提到引以为豪的自营，而是代之以“京东日益成为中国消费者无忧网购和快速履约的首选平台”。可以这样说，京东已经不是自营电商，而是与天猫、淘宝一样成为了平台型电商。

于是，我们就不难理解今年围绕“双 11”电商节前的商家合作是非，因为如今的京东再也难以对第三方商家的选择保持淡定。随之而来的挑战就是，京东以第三方商家为主的销售模式，还可以继续高举没有假货的大旗吗？

GMV 和现金流是京东的生命线，增速放缓确实有危险

京东是一家或者说曾经是一家以自营为主的电商企业，所以，GMV

就成了生存的命根。只有 GMV 不断刷新，才能维持新老账单的平衡，也才能有足够的收入去发展新业务和扩大投入，一旦 GMV 放缓，现金流就会出现困难。大概可以有个判断，这种企业经营模式的电子企业，必须保持不低于 50%的同比增速，否则就很难可持续增长。

财报数据显示，京东第三季度 GMV 同比增长 71%，上一季度同比增长 82%，增速下降 10%，虽然有所放缓，但仍处在良性发展区间。其中，自营 GMV 同比增速从 65%已经下降到 52%，增速下降 13 个百分点。同时，数码 3C 销量的增速从 70%下降到 59%，也下降 11 个百分点。由此，GMV 的下滑导致了京东的亏损进一步扩大到 5.3 亿元，而上季度的亏损只有 1.64 亿元，净现金流为–2.52 亿元。

当然，亏损并不是京东所害怕的，也不是京东最为关注的指标，只要有资本市场的输血和可持续增长的千亿级 GMV，京东这点亏损不会有大碍。不过，京东的账期已经从 2015 年第二季度的 40 天拉长到 51.08 天，应付款 253 亿元但现金储备只有 234 亿元，现金流遭遇的压力还是比较大。

京东财报同时也声明，公司的增长有可能继续放缓，京东 2015 年第四季度净收入介于 510 亿元至 525 亿元之间，同比增长约为 47 %至 51%之间，这个收入增幅与 2014 年第四季度的 73%已经不可同日而语。

任何一家公司，都不可能持续地保持高达 60%以上的高速增长，渡过初创期之后肯定会迎来稳定发展期，高速增长并非常态，也不是正常的企业经营状态。京东的季度交易总额（GMV）已经达到 1150 亿元人民币（约 181 亿美元）的水平，再保持急速增长将带来公司的严重管理问题，也会成为发展中最大的隐患。

我们可以下个结论，京东的发展速度降下来，对公司并非坏事，

关键是京东要找到不需要高速增长来维系企业发展的商业模式，与自营渐行渐远也许是走在了正确的路上。总之，以前京东嘲笑阿里巴巴的那些“电商坏东西”，最终都会落到自己的头上，以增幅和增速论英雄的时代在电商江湖可以休了，京东大船也要左满舵了。

大数据分析的关键在于有用

大数据比较热。传说中的牛公司都在做大数据。阿里巴巴甚至还专门收购了一家数据分析的公司，雅虎也入手一家中国社交数据分析的公司，由此可以看出大数据能力是多么被看重，即便是互联网巨头们也觉得自己囊中羞涩能力欠缺。

不过，与这股热闹相反的却是大数据应用的乏力，即便是号称数据能力超群的阿里巴巴也几乎没有摸到数据的门槛。可以毫不客气地讲，以目前淘宝上的通过数据分析展现出来的购物推荐看，这种大数据几乎是被糟蹋了。甚至有微博称，自己仅仅是好奇于朋友说淘宝上卖棺材是免费送货而打开了几个店家的页面看了看，结果，此后的一周自己的网页上便不厌其烦地开始被推荐各种各样的骨灰盒与棺材，真是让人汗颜。

不管你的数据来自哪里，是来自 COOKIE，还是用户的历史浏览和消费记录，线性的思考方式绝对不适合以人脑思维为基础的客户行为预测，即便号称神经网络等的非线性分析方法，也很难准确地进行预测评估，归根结底，人的行为是社会性的，人的决策也非完全理性的，更不要说毫无提前量的人云亦云的乱推荐。

大数据确实给分析人员提供了更好的基础，IT 技术的发展也让人们有了更方便的分析工具，但却导致了越来越多的分析过程被机械化的技术专业人士们主导，喜欢遨游在编程海洋中的技术天才们多数都是不食人间烟火的科技疯子，就数据论数据的方式严重制约了数据分析结果的使用价值。

因此，大数据分析的成功不仅仅在于应用，更在于能够有价值的应用，粗制滥造地去应用很可能导致彻底的失败。

做大数据分析，至少要做到以下几点：

（1）虽然你关注相关性，但这种相关性也应该在一定程度上被验证因果，毫无因果可言的相关也许是暗含与宇宙黑洞的秘密，至少现在对人类用处不大。

（2）先进的分析技术和高级的程序员都只是数据分析中的工具和操作手，都只能是作为决策的辅助，参谋不能带长，只会写报表的参谋永远不能当参谋长，更不能去当指挥战争的参谋总长。

（3）组织各个领域的专家委员会，让消费者研究专家们看那些似是而非的结论，然后根据发现的可能亮点再去有针对性地发掘，漫无目的或者成规制地出报表只适合绩效考核。

（4）让那些分析网络访问量、用户来自哪里、喜欢看什么网页等的传统互联网分析理念远离消费推荐领域，用那些为了网站运营而分析的套路去看待消费者行为是刻舟求剑，驴唇不对马嘴。

（5）忘掉那些高深的术语，被用专业门槛将公司里面的需求者阻挡在豪门之外，数据分析没那么神奇，即便多了一个大。

（6）数据也会说假话，片面相信数据的结果是彻底的教条主义，任何看似非常科学的结论都有可能是你自己的分析方法导致的。

（7）大数据当然有用，但要掌握在有用的人手里，更要掌握在会用的手里，更需要掌握在不乱用的人手里。

（8）让机器替代人脑，认为机器可以替代人脑，只有傻子才这样想，至少在人类文明的现阶段是傻子。

（9）根据客户行为做推荐的时候多设置几道识别好不好，特别不靠谱的要拦住，即便不能给客户提供帮助，也别给人家添堵。

（10）做好大数据，先从小数据开始吧，虽然你可以说大数据可不仅仅是个大，但任何的大都是从小来的，“不积小流无以成江河，不积跬步无以致千里”。

第 4 章

分析方法的全聚合

汇总与排序，你离不开的

求和，汉语的解释是，求得两个或两个以上数字相加的总数。如果仅仅是数字的求和，我们也可以将其称为“汇总”，统计上一般用∑表示。

我们做数据分析的时候，往往第一个步骤就是要做求和汇总，一共有多少用户，一共有多少收入和利润，等等。

2015 年 10 月，住房城乡建设部发布的《中国世界自然遗产发展公报（1985—2015）》显示，截至 2015 年 10 月，我国共有世界遗产 48 项，总数居全球第二，仅次于意大利，成为名副其实的世界遗产大国。

中国人兜里到底有多少钱？陆金所董事长首次公开提及陆金所过去几年研究的中国个人财富数据：中国整体个人的财产，加起来超 100 万亿元。如果把个人财产分成三段，高端客户可以投 50 万元以上，这个人口可能只有 1200 万人，但控制的整体财产有 40 万亿元。从可以投 1 万元到 50 万元间，这个人口接近 7000 万元，但控制的整体财产可能接近 10 万亿元。再下去，可以投 1 万元以下的大众客户，代表了 13 亿人口，控制财产大概是 45 万亿元。

在计算机编程理论上，将杂乱无章的数据元素通过一定的方法按关键字顺序排列的过程叫作排序。对数据来说，排序就是从小到大或者从大到小进行排列，然后就可以找出最大值、最小值，还可以借此找到四分位差等。简单的数字排序很容易实现，但复杂的数据资料进行排序就成了很艰巨的工作。

在美国斯坦福大学该校胡佛研究所档案馆所藏张嘉**璈**（曾任中国银行、中央银行总裁）档案中，有一份日本特务机关于 1939 年 10 月

17 日对国民党政府高级官员在上海外国银行存款情况所作做的秘密调查报告，名为《登集团特报丙第一号 政府要人上海外国银行预金（存款）调查表》。据该调查显示，不仅陈立夫有不少存款，蒋介石的存款数更是位居榜首。蒋介石与宋美龄夫妇的存款总数为 9733 万元（即 1186 万美元），约占当年国内存款总额的 1.6%，政府预算收入的 13%。外汇储备的 4.7%，高居于上述国民党政府官员私人存款额之首。

下面是 2015 年最贵的十个域名，看看吧。

	域名	成交价格	日期
1.	Porno.com	$8,888,888	2/4/15
2. tie	PX.com	$1,000,000	9/9/15
2. tie	588.com	$1,000,000	9/23/15
4.	989.com	$818,181.81	10/28/15
5.	899.com	$801,000	9/16/15
6.	345.com	$800,000	1/7/15
7.	GJ.com	$694,095	10/7/15
8.	NL.com	$575,000	3/4/15
9.	SX.com	$555,050	8/14/15
10.	QE.com	$554,000	7/22/15

汇总和排序只是说明全局，要想了解得更深刻，接下来一般就会进行结构分析，首先就是要算清楚其中的比例关系。

微信官方公布的数据显示，2015 年 9 月，平均每天有 5.7 亿人登录微信，根据国家统计局的数字，2014 年年末，中国内地总人口是 13.6782 亿人，这意味着每天登录微信的人数已经开始接近中国人口的一半。2015 年除夕当日，微信红包收发总量达 10.1 亿次；18 日 20:00～19 日 00:48，春晚微信摇一摇互动总量达 110 亿次。2015 年以来，“520 节”红包数量为 4 亿个，六一儿童节为 5 亿个，七夕为 14 亿个。2015 年元旦当天，微信红包收发总量达到 23.1 亿次，超过 2014 年除夕 2 倍

还多，微信红包的峰值出现在 1 月 1 日 00:05，在这一分钟内有 240 万个红包被发出，620 万个红包被拆开。

来自蚂蚁金融服务集团的数据显示，其旗下品牌支付宝在 2015 年“双十一”期间，共完成 7.1 亿笔支付。支付峰值出现在凌晨 0 点 05 分 01 秒，达到 8.59 万笔/秒，这一数值远远超出全球其他支付机构的处理能力。200 余家银行与蚂蚁金服共同打造了世界上交易处理能力最强的支付平台。

这个分析的最后，我们来看一下人们对唐诗的大数据词汇分析。大数据分析发现，李白与杜甫的用字习惯差别很大，在他们所有的五言诗中，李白撰字 6.8 万字，杜甫为 9.3 万字。但是，李诗中只有 3127 个字是互异的，杜诗中则有 3907 个互异字。他们同时都使用的字眼只有 2764 个字。换句话说，杜诗中 29%的字眼是李不用的。而李诗中，只有 363 个字未被杜使用。加在一起，他们共使用 4270 个互异的汉字。按照这个思路，有一位网友统计了汪峰在大陆发行的 9 张专辑共 117 首歌曲的歌词，同一个词语在一首歌中出现只算一次。形容词、名词和动词的前十名请看下图（词语后面的数字为出现的次数）。

	形容词		名词		动词
0	孤独：34	0	生命：50	0	爱：54
1	自由：17	1	路：37	1	碎：37
2	迷惘：16	2	夜：29	2	哭：35
3	坚强：13	3	天空：24	3	死：27
4	绝望：8	4	孩子：23	4	飞：26
5	青春：7	5	雨：21	5	梦想：14
6	迷茫：6	6	石头：9	6	祈祷：10
7	光明：6	7	鸟：9	7	离去：10
9	理想：6	8	瞬间：8	8	再见：9
9	荒谬：5	9	桥：5	9	埋：6

然后，我们就可以汪老师式创作了。你随便写一串数字，然后按数位，依次在形容词、名词和动词中取出名词，比如，圆周率 3.1415926，对应的词语就是：坚强，路，飞，自由，雨，埋，迷惘。然后，这首歌就成了这样："坚强的孩子，依然前行在路上，张开翅膀飞向自由，让雨水埋葬他的迷惘。"

谁说比例与频次不是分析

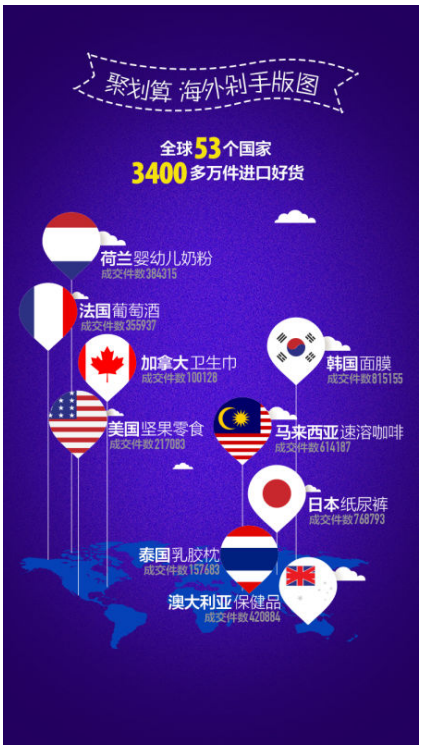
很多时候，简单的列举数量和比例就是最好的分析，而且，对大数据分析来说，也许 80%的工作就在分析比例（频数）上。

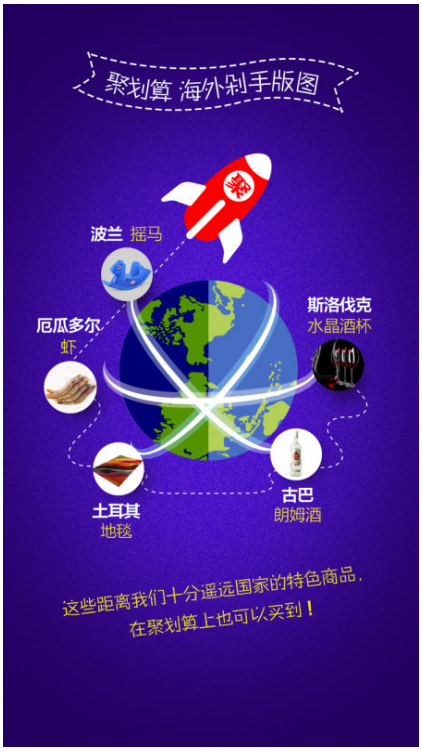
2015 年 12 月 30 日，聚划算公布 2015 年首份跨境报告。报告显示，女性购买进口商品占到 77%的比例，成为进口商品消费的绝对主力军，"进口女性"时代来临。

根据报告，2015 年 4 月 1 日至同年 12 月 12 日，共有超过千万的消费者通过聚划算全球精选购买了来自 53 个国家和地区的 3400 万件进口商品，德国牛奶、泰国乳胶枕、日本纸尿裤最畅销。同时，吃货吃遍全世界，占据进口商品消费近一半的市场份额。

日本、韩国、德国商品占据半壁江山

聚划算跨境报告数据显示，进口商品最受欢迎的 TOP 10 的国家和地区分别是：韩国、日本、德国、英国、法国、荷兰、新西兰、澳大利亚、中国台湾、泰国。其中日本、韩国、德国的进口商品成交占据了半壁江山，仅日韩两个国家的成交就占据 37%，中国消费者对日韩商品的喜爱程度可见一斑。





目前，聚划算全球精选的海外商品覆盖零食、生鲜、保健、母婴、玩具、个护、百货、美妆、家居、家电等多个类目，已经渗透到消费者需求的方方面面。

聚划算平台数据显示，在 2015 年的前 9 个月时间里，食品、母婴、保健品及生活用品成为国人心头好。期间，共卖出日本纸尿裤 768793 件，马来西亚速溶咖啡 614187 盒，德国牛奶 613399 盒，荷兰婴幼儿奶粉 384315 罐，法国葡萄酒 355937 瓶，澳大利亚保健品 420884 盒，美国坚果零食 217083 件，加拿大卫生巾 100128 包，泰国乳胶枕 157683 只。

其中，国人在奶制品的选择上更信赖进口奶源，德国位居第一。来自德国、新西兰和澳洲的进口牛奶共卖出 1100 多万件。

另外，来自韩国、日本、德国、美国、法国、荷兰、新西兰、澳大利亚、泰国、英国、意大利、瑞士和瑞典这 13 个国家的牙膏总共卖出 63 万件（平均每件在 1.5 支左右），等于售出百万支牙膏，进口面膜总共售出了 1100 万件，实现 1 亿元的销售额。

80 后、90 后仍是消费主力，天秤、天蝎、处女座成剁手党三甲

数据显示，聚划算“全球精选”的消费主力军以 80 后、90 后为主，占据了超七成比例，其中 80 后超五成。

同时，女性比男性更爱买进口商品，女性占到 77%的比例，是进口商品消费的绝对主力军。其中以 29 ~ 35 岁高学历、高收入的一线都市女性为主。她们在燕窝、高端厨房电器、空气净化器、胶原蛋白、鱼油、原汁机这些家庭型消费方面出手阔绰。

天秤座、天蝎座和处女座荣登全球精选“剁手榜”三甲。挑剔的处女座最爱在全球精选买扫地机器人，追求品质完美的金牛座最爱在全球精选买进口牛奶，霸气的狮子座最爱买厨房装备。而婴幼儿牛奶粉、膳食营养补充食品、纸尿裤则成为 12 星座的集体最爱。

上海人最爱吃，江苏人最爱丰胸，北京人爱减肥

在聚划算全球精选购买海外进口的全国地区 TOP 10 排行是:浙江、江苏、上海、广东、北京、山东、湖北、福建、四川、河南。前十榜单中江浙沪地区占据三席，其中浙江人最爱买杯子、水壶，江苏人最爱买蜂蜜，上海人最爱买吸尘器。

数据显示，在海外商品的选择上，食品和保健品最受宠，吃货尤其称霸，占整体海外商品消费的四成。最受吃货欢迎的是来自意大利的果仁巧克力，其中最著名的代表是费列罗，进口海苔紧随其后。来

自上海的吃货消费力最强，13.39%的巧克力被上海吃货搬回家。

上海人除了在吃的方面敢花钱，也更爱美、更注重享受。在购买风靡全球的酵素类减肥品排名前五的城市是：北京、上海、广州、深圳、杭州，可见北京人爱减肥。同时，来自 350 多座城市的消费者在全球精选购买了 50 多万只进口的避孕套和 6 万瓶进口玛卡，上海均位居第一，分别占 4.7%和 5.83%。

在丰胸霜和减肥产品的选择上，南北城市的女性均表现出了对好身材的高要求。南方妹子热衷丰胸，北京妹子减肥需求明显。在丰胸霜消费 TOP 10 城市的榜单中，江苏占据五席，包揽了进口丰胸霜 43%的销量，其中苏州居榜首。有趣的是丰胸霜还有 5%被都市男性买走。另外，在减肥类产品中，购买排名前五的城市是北京、上海、广州、深圳和杭州。

2015 年成聚划算跨境爆发年

2015 年，聚划算大力进军跨境电商，与天猫国际、淘宝全球购并肩作战，在“深化进口”的核心战略下，针对全球货品的供应链进行深度整合。

2015 年 5 月 11 日，聚划算联手国内杭州、宁波、广州、深圳、郑州、重庆六地保税区，开展“买空保税区”活动，重点品类集中在“进口母婴用品”、“进口食品保健品”、“进口百货”、“进口美妆”等四大品类，逾千种商品，短短 3 天时间共计成交 3135 万元，相当于深圳保税区日常出单 30 天的量。通过保税区的深化合作，有助于聚划算推出“正品保证”、“保税区直邮”等服务。

2015 年 6 月，阿里巴巴集团旗下聚划算平台和天猫国际联合开启“地球村”模式，加快与 20 国合作进程，更多的海外特色商品在聚划算实现首发。

此前，聚划算联合泰国商务部进行乳胶家纺销售。开卖不到 10 小时，6 款乳胶家纺接连售罄，紧急联系泰国工厂协调补货数次，最终在 3 天的时间里有 45000 多人次下单，使得泰国乳胶枕在这一年正式打开了中国市场。在高端进口家电上，如 2298 元的韩国惠人原汁机慢速榨汁机，三天就卖出 500 多台，实现千万级的销量。

据悉，聚划算在 2015 年 12 月 28 日 ~ 30 日推出的全球聚热榜单，泰国乳胶枕再次上榜。同时，HUO12FRM 韩国原装进口原汁机、日本花王纸尿裤、女性孕妇补铁补血营养品、韩国可莱丝 NMF 针剂水库面膜贴、绝世西餐厅牛排套餐、西班牙醉梦红酒、德国 Brita 碧然德进口净水壶过滤器、DYSON 戴森 DC45 无绳吸尘器都成为 2015 年的热销单品。

据商务部发布的全球贸易格局报告预测，2016 年，我国跨境电商进出口额将增长至 6.5 万亿元，年增速将超过 30%。全球最大管理咨询公司埃森哲于 2015 年 6 月发布的《2020 全球跨境电商趋势报告》显示，2020 年，中国有望成为全球最大的跨境 B2C 消费市场。

当然，以上的比例分析基本上都是简单的比例比较，在进行比例分析的时候还需要考虑比例之间的协调关系。比如，有一个分析认为，“30%的车祸是持驾照三年以下者所为，所以新驾驶员容易闯祸”，你觉得分析正确吗？

根据南京交警部门的分析，一天按 24 小时时段分析来看，6 点至 10 点和 17 点至 21 点为交通事故高发时段，53%的事故发生在这两个时间段内。其中，事故多发且损害后果严重的时段为早晨 6 点至 7 点，发生事故占事故总起数的 7.3%，同时死亡人数也在全天时段中最高，比例达到了 10.9%。其次，19 点至 20 点发生事故占事故总起数 7.1%，同时受伤人数也在全天时段最高，比例达到了 8.9%。针对交通事故大数据分析，交管部门专门分析了新手司机肇事的原因。从数据分析来看，驾龄 1 年以下的驾驶人引发的事故最多，而新手到了第 2 年后引发事故明显下降，到第 3 年和第 4 年开车的心态有所变化，自认为驾驶技能熟练了，一些不良驾驶习惯和违法行为会增多，比如抢绿灯尾或黄灯、超速、随意变道、加塞等，因而引发事故又有所上升。

事实上，“30%的车祸是持驾照三年以下者所为，所以新驾驶员容易闯祸”，这个结论可能并不正确。因为，我们需要先看一下“持驾照三年以下者”在总的驾驶员中的比例情况，如果“持驾照三年以下者”占总驾驶员的比例不到 30%却闯了 30%的车祸，那么这些新驾驶员就是容易闯祸，否则，如果“持驾照三年以下者”占总驾驶员的比例超

过 30%，可车祸只占到 30%，那就说明新驾驶员不容易闯祸。

比例分析就是如此，看起来多，并不是一定就多，还要看隔壁家是多还是少，或者在总体中的比例。在这里，真的是要患寡而患不均。

2015 年，由中国社会科学院新闻与传播研究所和社会科学文献出版社共同发布的《中国新媒体发展报告 No.6（2015）》蓝皮书指出，微博用户多是“三低人群”，即低学历、低年龄、低收入的人群。报告认为，从收入来看，微博用户平均收入水平依然较低。月收入 5000 元以上的微博用户约占 9.93%，5000 元以下的则占 90.07%，其中无收入群体最多，达到 8898.7 万人。但如果再看看中国社会的平均工资水平和月收入 5000 元以上的人口比例，也许就会得出相反的结论。

此外，比例还是会变化的，而不同比例的变化会带来总量的结构的变动。所以，分析比例，一定要站在动态的角度上，而不是静止地看待事物。

2014 年，全球航天经济持续增长，包括航天发射与地面服务、卫星制造、卫星电视与通信、政府研究、军事开销及其他领域在内的全球航天经济总量实现了 9%的增长，达到 3300 亿美元。其中，商业航天活动占 76%，额度较 2013 年增长了 9.7%。其余为政府投资，额度较 2013 年增长了 7.3%。美国政府的军事航天预算占全球航天预算的 54%。美国国家航空航天局 2014 年的预算较 2013 年提高了 4.6%，占全球政府航天投资的 22%。为了发展新能力和拓展现有能力，2014 年全球其他政府航天开支较 2013 年增长了 12.9%，增幅超过了美国。

本节最后再给大家提供一张图，不同行业里全球与中国的富豪人数比例和排名，看看你在里面吗？我们也可以看到，从全球来看，金融与投资领域的富豪总数比例遥遥领先，而在中国，富豪却主要集中

在房地产领域，由此足以看到中国发展模式存在的问题和隐患。

	全球富豪		中国富豪	
	比例	排名	比例	排名
金融与投资	19.2%	1	6.7%	3
资源	9.1%	2	6.5%	4
娱乐与文化	9.1%	3	1.4%	15
房地产	9.0%	4	23.5%	1
零售	8.9%	5	4.2%	10
IT	8.8%	6	5.8%	5
制造业	8.5%	7	19.1%	2
食品饮料	5.6%	8	2.9%	12
社会服务	4.1%	9	4.6%	9
医药	4.0%	10	5.5%	7
建筑	3.1%	11	2.5%	13
交通运输、仓储	2.7%	12	1.3%	16
服装纺织	2.5%	13	5.1%	8
农林牧渔	1.9%	14	2.0%	14
钢铁	1.8%	15	3.4%	11
新能源	1.7%	16	5.6%	6

平均数里隐藏的大秘密

计算平均数也许是所有数据分析方法里最简单也是最重要的一步。按照汉语词典里的说法：平均、一致、统一。平均如一，天下平均，合为一家。均匀；无轻重或多少之分，把总数按份儿均匀计算。

按照统计的定义，平均数是指在一组数据中所有数据之和再除以数据的个数。平均数是表示一组数据集中趋势的量数，它是反映数据集中趋势的一项指标。在分析数据中，平均数（均值）和标准差是描

述数据资料集中趋势和离散程度的两个最重要的测度值。用平均数表示一组数据的情况，有直观、简明的特点，所以在日常生活中经常用到，如平均速度、平均身高、平均产量、平均成绩等。一般来说，平均数可以分为简单平均数、加权平均数、调和平均数、几何平均数等，当然，还有一种叫作“截尾平均数”。

截尾平均数是指在一个数列中，去掉两端的极端值后所计算的算术平均数，也称为切尾均值。最常见的截尾平均数的例子是在一些比赛中（如跳水、体操等），计算选手的最终得分需要“去掉一个最高分，去掉一个最低分”，这种处理方法可以有效地去除“场外因素”的影响。

当然，除了平均数，还有两个指标也用来表示集中趋势，一个是众数，一个是中位数，这两个指标都与排序有关。我们把一组数据按从小到大的顺序排列，在中间的一个数字（或两个数字的平均值）叫作这组数据的中位数，而在一组数据中，出现次数最多的数就叫这组数据的众数。按统计学原理，只有在数据分布偏态（不对称）的情况下，才会出现均值、中位数和众数的明显区别。所以说，如果是正态的话，用哪个统计量都行。如果偏态的情况特别严重，可以用中位数。

实际上，平均数只是一个“虚拟”的数，而中位数并不完全是“虚”数，当一组数据有奇数个时，它就是该组数据顺序排列后中间的那个数据，是这组数据中真实存在的一个数据；平均数的大小与一组数据里的每个数据均有关系，其中任何数据的变动都会引起平均数的相应变动；众数着眼于对各数据出现的频数的考察，其大小只与这组数据中的部分数据有关；中位数则仅与数据的排列位置有关，某些数据的变动对中位数没有影响，当一组数据中的个别数据变动较大时，可用它来描述其集中趋势。

结合以上的分析，我们知道，简单平均数往往不能代表总体的真实水平。假设，我们公司里一群人在开会，这个时候李嘉诚或者马云进来了，我们计算一下在场人士的平均财富水平，那肯定让大家大吃一惊，因为我们都成了亿万富豪。有一个笑话说，老财有钱一千万，邻居九个穷光蛋，拉到一起算一算，家家户户是百万。

在 2015 韩国—四川省西部论坛上，西南财经大学中国家庭金融调查与研究中心家庭金融研究部首席研究员李凤副教授带来了中国家庭资产配置与变动趋势的最新调研成果。研究显示，2015 年中国家庭平均资产为 91.9 万元，比 2013 年增长两成。2015 年中国家庭总资产中，房产占比高达 69.2%，这比美国的两倍还多。

当然，平均数经常会忽悠人。比如，中国是个人均资源匮乏的国家，这是一个流传甚广的错误的“真理”。因为，中国确实是地大物博的，可是人口众多。

媒体报道，我国是一个干旱缺水严重的国家。淡水资源总量为 28000 亿立方米，占全球水资源的 6%，仅次于巴西、俄罗斯和加拿大，居世界第四位，但人均只有 2200 立方米，仅为世界平均水平的 1/4、美国的 1/5，在世界上名列 121 位，是全球 13 个人均水资源最贫乏的国家之一。这样的分析有一定的道理，但是却也有被简单平均数所忽悠的成分在里面。

比如，中国是世界矿产资源大国之一，探明矿产资源储量占全球 12%，仅次于美国和俄罗斯，全球 40 个主要矿产中，13 种有 3/4 集中在三个国家，23 种有 3/4 集中在 5 个国家。中国铁矿石储量占全球的 12.5%，人均却只有世界平均水平的 64%，但如果将全球人口分为三个群体，富铁五国人口 4.2 亿人，却拥有全球 63.4%的铁矿石储量，人均

276 吨，剩下的 50.8 亿人口只拥有 24%储量。中国的人均铁矿储量比全球 6.2%要少，但却比 74.3%的人要多。我们可以看到，这里是不是用了“截尾平均数”的方法，去掉了那些异常值的干扰，得到的结论更为科学。

此外，分组的平均数还可以解决大问题，我们在进行平均数分析的时候也可以分成几组来试试。

看一段比较老的媒体新闻报道，台湾消息，《红楼梦》这部中国经典文学巨著，最后四十回到底出自谁手一直是个谜，学界里有曹雪芹新撰、曹雪芹残篇、高鹗或程伟元补写等说法。台湾农委会林业试验所森林生物组研究员潘富俊从《红楼梦》中描述的植物入手，得出了红楼梦后四十回非曹雪芹亲撰的结论。

具体过程是这样的，根据台湾植物学家潘富俊在《红楼梦植物图鉴》一书序言所指出，《红楼梦》书中总计谈到 237 种植物，其中前四十回谈到 165 种，平均每回 11.2 种；中间四十回谈到 161 种，平均每回 10.7 种；最后四十回则只用到 66 种，平均每回 3.8 种。从引用植物的多寡，我们可以看出作者的植物素养，再一比较便略可知作者非同一人的可能性。潘富俊统计了大观园中的植物共 70 多种，包括热带、亚热带、温带及海岸、海中植物，所以大观园是南北各地园林的综合体，是作者塑造的理想花园，在现实中未必能找得到。

通过对喝茶的统计发现，在前八十回中，仅有六回没有提到茶，有茶的回数占 92.5%；而在后四十回中，却有十四回没有提到茶，有茶的回数仅为 65%，也是相差悬殊。

如此简单的分析，接近完美，也直接地解决了《红楼梦》作者的悬案，算是数据分析历史上的一段传奇。

方差，也许你不用关注，但还是要理解更好

在统计学中，有一个非常重要的概念，就是“方差”，很多高级的统计分析方法都是依托方差而逐渐发展起来的，大多数统计软件的计算中，方差都会被显示在结果之中，可是，在实际的工作中，我们却很少为了计算方差而去计算方差。

一般来解释，方差是各个数据与平均数之差的平方的平均数。在概率论和数理统计中，方差（Variance）用来度量随机变量和其数学期望（即均值）之间的偏离程度。

比如 1，2，3，4，5 这五个数的平均数是 3，方差就是 1/5[（1-3）² + （2-3）² +（3-3）² +（4-3）² +（5-3）²]=2，如果我们将方差的值进行开方，就得到了“标准差”。

计算方差做什么用呢？举例来说，两人的 5 次测验成绩如下。

A：50，100，100，60，50 E（A）=72；

B：73，70，75，72，70 E（B）=72。

两个人的平均成绩相同，但 A 不稳定，对平均值的偏离大。如此，我们就会说，A 的成绩起伏不定，可能是受到了外界因素的干扰，或者是上学贪玩造成的，而 B 的成绩很稳定，说明他就是这个水平，而 A 的提升空间很大。

用统计学的说法，当数据分布比较分散（即数据在平均数附近波动较大）时，各个数据与平均数的差的平方和较大，方差就较大；当数据分布比较集中时，各个数据与平均数的差的平方和较小。因此，方差越大，数据的波动越大；方差越小，数据的波动就越小。

当然，在统计学中还可以直接用方差分析，比如单因素方差分析，

通过方差比较确定组间差距是否显著，由此可以确认某因素是否会影响客户的选择。例如，年龄是否对手机的品牌选择构成影响，屏幕大小对用户使用手机流量是否影响很大，收入是否会对选择私家车的品牌构成影响等。

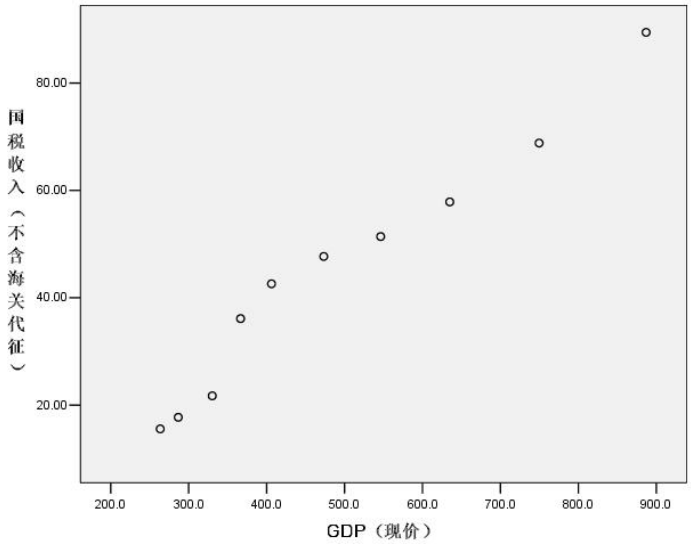
方差看似很容易计算，但使用起来却很麻烦，有兴趣的朋友可以在统计学的道路上好好用功钻研一下。

大数据时代的相关关系和因果关系

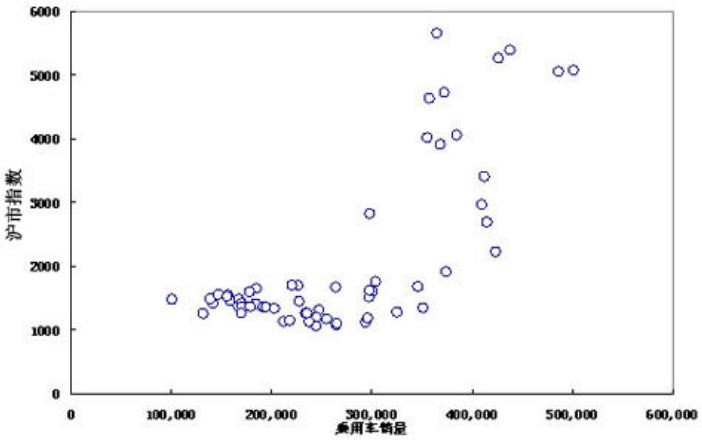
人类社会是普遍联系的，事物之间往往存在着冥冥之中的相关关系，就如同古人认为每一颗星星都对应着世上的一个人，当一颗流星划过，就代表一个人离开了。

相关分析（Correlation Analysis）是研究现象之间是否存在某种依存关系，并对具体有依存关系的现象探讨其相关方向及相关程度，是研究随机变量之间的相关关系的一种统计方法。相关关系是一种非确定性的关系，例如，以 X 和 Y 分别记一个人的身高和体重，或分别记录每天的页面浏览量与每天的网络销售量，则 X 与 Y 显然有关系，而又没有确切到可由其中一个去精确地决定另一个的程度，这就是相关关系。

下图形象地说明了 GDP 国税收入的强烈正相关关系。



此外，还有人做过研究，中国股市和车市有很强的相关性，乘用车销量和沪市表现相互有正面的影响，如下图所示。



乘用车销量

气象学家观测到，天气和太阳的活动也是有关系的，而天气气候的变动和国运的兴衰联系也很密切。研究发现，季风气候的不稳定性

使得中国气候灾害具有频率高、强度大的特点。发生的重大气候灾害往往更易引发社会危机，导致重大农民起义的爆发，甚至成为社会动荡乃至朝代更替的导火索。这也是造成中国历史上百年尺度的冷暖变化与社会经济的衰落呈现同期性，盛世往往悄随“流火”而去的重要因素。例如，17 世纪的小冰期寒冷阶段内崇祯大旱引发的李自成起义导致了明朝的灭亡，19 世纪小冰期的寒冷阶段内西南大旱引发的太平天国运动对清朝社会经济构成了重大打击。

据英国《每日邮报》网站于 2015 年 7 月 10 日报道，科学家警告称太阳将在 2030 年“休眠”，这将导致地球气温大幅度下降，使得地球步入“小冰河期”。这一发现是在英国皇家天文学会于威尔士兰迪德诺召开的国家天文会议上公布的。瓦伦蒂娜·扎尔科夫教授及其研究团队在会上介绍了他们研发的太阳活动周期新模型，该模型关注太阳的两个层面——一个靠近太阳表面，另一个深入太阳的对流区，预测到太阳活动将在 2030 年左右减少 60%，届时地球将很有可能进入“小冰河期”。

扎尔科夫的研究发现，在太阳活动的第 25 周期（该周期的太阳活动在 2022 年达到峰值），被列为观测对象的太阳两个层面的电磁波开始相互抵消；进入第 26 周期（2030 年至 2040 年）后，这两个层面的电磁波变得完全不同步，导致太阳活动剧烈减少。

“在第 26 周期，这两个层面的电磁波完全互为镜像——在相对的太阳半球同时达到峰值。”扎尔科夫说，“它们的相互作用是极具破坏性的，近乎相互彻底抵消。我们预测这将引发与‘蒙德极小期’相同的效应。”

公元 1645 年至 1715 年是蒙德极小期，在此期间太阳活动非常衰

微，持续时间长达不可思议的 70 年，此时也恰好是地球的小冰期，但两者是否有关联，仍然没有定论。当时，在寒冷的冬季，英国大部分河流都冻结了，当代油画显示人们甚至能够穿着旱冰鞋横穿泰晤士河。

如果地球真的要进入小冰河期，我们这个社会又将发生什么呢？2016 年 1 月 20 日开始，中国多地遭遇极寒天气，1 月 22 日，新华社记者在呼伦贝尔根河市的冷极村测出瞬间温度零下 60℃。广州也下了百年一遇的雪。

据历史记载，广州城区下雪非常罕见。其实，在有据可考的历史上，广州曾经下过 15 场大雪，2016 年年初这一次正是第 16 次。根据广州本地有关历史气候的研究：1500 年～1920 年广州地区共出现 41 次寒害，期间广州地区下过大雪有 15 次，广州下过连日大雪的记载有 8 次，分别是 1245 年、1618 年、1634 年、1787 年、1788 年、1832 年、1835 年、1892 年。

历史文献记载：南宋以降，广州下雪，见诸方志及时人题咏者，有“淳佑五年（1245 年）十二月广州大雪”。戴暟《广州通志稿》引《郡》云：“腊初，大雪三日，积盈尺余，炎方所未有也。时经略使方大琮躬出省视，贫民与诸营疲卒，均给缗钱，以恤其寒，阖郡大悦。”永乐十三年（1415 年）冬广州、南海、番禺“有雪，梅花枯死”。嘉靖十六年（1537 年）冬，番禺、南海大雪。万历四十六年（1618 年）十二月，从化大雪三日。“时，互阴寒甚，昼下如珠，次日如鹅毛，六日至八日乃已。山谷之中，峰尽壁立，林皆琼挺。老父俱言从来未有也。自是连岁大稔。”崇祯七年（1634 年）从化大雪。顺治十一年（1654 年）正月十八日，广州、龙门大雪。康熙五年（1666 年）十二月二十夜，广州、番禺、南海大雪。难道小冰河期，此言不虚？

沃尔玛是最早发现尿布和啤酒的销售有相关关系的。一开始不明白这两个东西为什么会有相关关系？后来发现当家里面有了小孩子之后，买尿布的任务往往是让新爸爸去干的。其实爸爸对孩子的出生贡献并不大，但是他觉得自己做出了很大的成绩。所以他买完尿布的时候，会想顺便买一瓶啤酒犒劳自己。所以后来沃尔玛就把啤酒和尿布放在一块儿，啤酒的销售量一下子就增加了，这是一个很经典的案例。

这样的分析非常经典，几乎所有的有关大数据的介绍里都会用到，但是，这个效应却没有在中国呈现，一些模仿沃尔玛的超市也收获了失望。为何？

究其原因，美国人的生活习惯与中国人差别很大。在中国的家庭里，给孩子买尿不湿的任务往往是母亲来完成，而且多是目标明确地去购买或者临时性随手购物，并不会同时带上笨重的啤酒箱。

大数据时代是一个注重相关关系的时代，人们变得不再对事物之间的因果进行深入细致的研究。这个偏好也并非是现在才有，数字化时代都存在同样的问题，当计算机出现之后，人们就不再对哥德巴赫猜想这样的话题感兴趣，也不再去探索公式化地求索圆周率的值，而是通过计算机无穷尽地接近这些真相的最终结果，但却不会去找到最终结果。

有时候，原因和结果之间真的很难找到答案。比如，曾经有一家很著名的体育机构，研究发现，大多数足球球星的生日都集中在上半年，所以他们认为，出生在上半年的人更容易成为球星。

1 月 23 日　罗本（1984）

1 月 28 日　吉安路易吉·布冯（1978）

1 月 29 日　罗马里奥（1966）

2 月 5 日 克里斯蒂亚诺·罗纳尔多（1985）

2 月 18 日 罗伯特·巴乔（1967）

4 月 16 日 永贝里（1977）

5 月 2 日 大卫·贝克汉姆（1975）

6 月 24 日 利昂内尔·安德雷斯·梅西（1987）

根据 CIES 对 2009～2010 赛季以来，欧洲 31 个国家顶级联赛的 28685 名球员研究表明，出生在每年前三个月的球员占比最多，达到 30.5%，而每年最后三个月出生的球员只有 19.3%。

这是为什么呢？CIES 表示，这可能和球员的选拔机制有关，在青少年时代，同一年出生的球员放在一起训练，出生在前几个月的球员因为身材发育更好，相对更容易脱颖而出。在十三四岁甚至更早的年龄阶段，早出生半年在身体上有着明显的差距。

实际上，这就是一个典型的相关却没有因果的例子，之所以会出现这种球员生日与成绩之间的关系，主要是因为制度催生的，因为足球比赛是分年龄段的，球队和教练会选择最合适的球员，小年龄段的球员在比赛中生日靠前在同年龄中占有明显优势，所以日积月累就形成了规律，只要制度改变，这种关系就会发生变化。比如，英格兰联赛比赛风格和制度与众不同，球员出生最多的反而是 9～12 月，占到 39.9%，其次是 5～8 月，占到 24.4%。

当然，这种足球球星生日的规律在中国足球并不生效，因为球员们的生日是“可调”的。西方一些球队和教练只知道使用生日占优势的球员，而中国的很多教练会让球员随心所欲地更改年龄。2004 年，中国足协曾对青少年球员虚报年龄的情况进行过一次检查，结果相当触目惊心：在对 1989、1990 年龄段的 1610 名少年球员进行了骨龄检

测之后，不合格率是 27.7%，不合格人数总计达 446 人。

为什么要更改年龄呢？因为足球比赛是分年龄段来进行的。比如，从 2011 年开始，中国足协正式启动了全国青少年足球联赛。根据全国青少年足球运动员情况，联赛设立了青年组（U-19、U-17）和少年组（U-15、U-13）四个组别，在赛制上设南北区省市联赛、大区联赛、大区决赛及全国总决赛等 4 个比赛阶段，国际上的相应比赛也以年龄为界限，更改年龄就可以参加相应年龄段的比赛，如果以大打小，自然容易出成绩。在中国足球史上曾经出现过中学生比赛，球员带着剃须刀参赛的奇观。1985 年在中国举行的首届柯达杯世少赛（U16），中国队闯进前八。但 2000 年中国足协注册办主任马成全曾打开疑团，这一届国少队是清一色的超龄球员。

当然，把年龄改小也是有的。比如某著名球员就曾经因为年龄问题闹得沸沸扬扬，原因竟然是注册年龄比实际年龄小了一岁，导致转会无法操作（球员转会有年龄上的限制）。事实上，这位球员变“小”一岁，实惠可谓巨大。因为 1984 年龄段并非奥运周期，不仅会错过 2005 年全运会，要参加奥运会也非常困难，而变成 1985 年出生，参加全运会和北京奥运会也就顺理成章了。

当年被公认为中国足球希望的健力宝队，多名主力都曾深陷年龄风波。其中最典型的当属郝伟。关于郝伟的出生年份，先后出现了从 1973 年到 1977 年的五个版本，整整横跨了一个奥运周期，从最小到最大竟然相差了五岁。而当时各方比较一致的看法是，当年在接受健力宝队挑选竞争去巴西留学的资格之前，郝伟将自己的出生年从 1975 年改为 1977 年。搞笑的是，后来为了提前登记结婚，郝伟又将出生年改为 1973 年。

总之，更改年龄参赛是中国足球长期低迷的重要原因之一。国外

的科学研究表明，球员虚报年龄造成的能力损耗有其客观规律：17 岁以下的球员年龄如果改小 1 岁，那他将失去 30%成功的机会；如果改小 3 岁，就将失去 90%成功的机会。

在现实中，我们也经常会说“出动消防员越多，火灾损失就越大”，或者“哪里有警察，哪里就堵车”，这是因果关系倒置的结果。实际情况应该是，因为火灾比较大，出动的消防员才比较多，因为太堵车，所以警察才经常出动。

2011 年 4 月 17 日，北京市交通委首次公布缓堵成绩。2011 年 1 月至 3 月，北京交通拥堵指数下降 16%，工作日平均拥堵持续时间减少 1 小时，为 1 小时 15 分钟。有媒体报道的时候说，“限购使北京交通拥堵状况下降了 16%”。这样的报道有问题吗？

目前，中国国内限购汽车的城市越来越多，已经有上海、北京、广州、杭州、石家庄、贵阳、深圳、天津等陆续开始限购汽车。“汽车限购令”是为解决城市交通拥堵问题，部分城市出台的限购汽车政策，意在缓解交通压力。

各地的政策不尽一致。1994 年开始，上海就对其车辆进行了限购政策，通过竞价进行车牌出售，每一次定的指标为 10 万个，也是第一个对私家车牌进行限制的城市。2010 年 12 月 23 日开始，北京就对其车辆进行了限购政策，同时还颁布了《北京市小客车数量调控暂行规定》，主要是通过摇号来出售车牌。2011 年到 2013 年期间，每期进行了 24 万个车牌的发放，但到了 2014 年就只有 15 万个车牌发放。2012 年 6 月 30 日开始对市民宣称要进行车辆配额管理，通过一半摇号和一半竞价来进行车牌出售。每年发放出来的车牌只有 12 万个，两种方式都是通过平均分散发放的。2014 年 12 月 29 日进行了车牌限购政策，

主要通过摇号、竞价、限行，每年进行 8 万个车牌发放。2014 年 3 月 25 日宣称，开始对全市进行汽车限购，但是在 3 月 26 日政策才开始执行。主要针对城市的小型客车通过总量控制和错峰限行来进行调控，也是通过摇号和竞价进行车牌出售。每年 8 万个车牌发放。

以北京为例，2010 年 12 月 23 日下午，北京正式公布《北京市小客车数量调控暂行规定》实施细则，2011 年度小客车总量额度指标为 24 万个（月均 2 万个），个人占 88%。每月 26 日实行无偿摇号方式分配车辆指标。外地人在北京购车需连续 5 年以上缴纳北京社保和个税的证明；港澳台居民、华侨及外籍人员需 1 年居住证明。更新指标无须摇号，直接申请更新即可。

如果你弄懂了这个限购令，就知道“限购使北京交通拥堵状况下降了 16%”这个结论完全站不住脚。因为，北京对“更新”是不限制的，也就是说，这个政策只是对新增用户进行了限制，存量并不调节，限购制度效果再好，也只能是延缓了道路拥堵继续恶化的进程，不会对现有的拥堵有任何帮助。原因和结果之间并没有必然的联系。

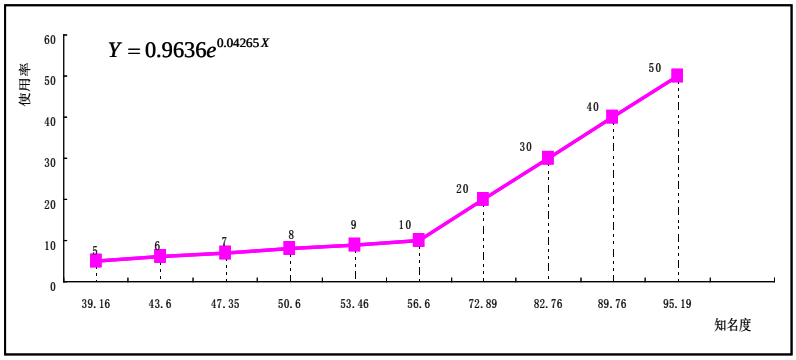
回归分析，你必须学会的分析方法

回归分析是应用最广的一种分析方法，也是统计学教学时的最重要内容。按照教科书里的解释，回归分析是根据已知的一个或一个以上变量（自变量）的值来估计另一个变量（因变量）的值，并且算出估计的误差，所建立的数学模型及所进行的统计分析。回归分析是希望得出一个有关各个变量之间联系的数字表达式，其中只有目标变量

因变量假设为随机变动，而自变量均为已知常数。

回归分析一般应用于度量影响程度的大小或者是对未来的预测等领域。比如，销售差异能够用广告支出、价格、分销水平上的差异解释，市场份额的差异能够用销售力量的强弱、广告支出和促销额来解释，消费者对产品质量判断由产品价格、品牌形象和属性决定。回归分析还被广泛应用于研究社会现象，如社会预测、人口预测、经济预测、政治预测、科技预测、军事预测、气象预测等。

这里有一个经典的例子，根据美国营销专家的研究模型，日用消费品的知名度与其使用率有着直接的关系，一般来说，只有当知名度达到一定程度后（如 60%），使用率才会急剧上升。



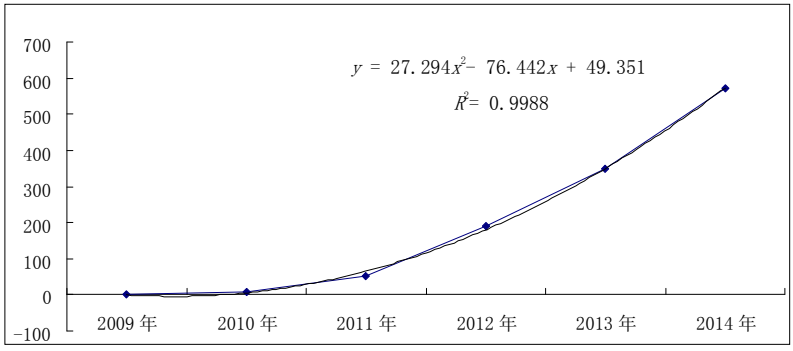
要理解回归的含义，也许需要从源头开始。一般认为，“回归”是由英国著名生物学家兼统计学家高尔顿（Francis Galton，1822—1911，生物学家达尔文的表弟）在研究人类遗传问题时提出来的。为了研究父代与子代身高的关系，高尔顿搜集了 1078 对父亲及其儿子的身高数据。他发现这些数据的散点图大致呈直线状态，也就是说，总的趋势是父亲的身高增加时，儿子的身高也倾向于增加。但是，高尔顿对试验数据进行了深入分析，发现了一个很有趣的现象——回归效应。因

为当父亲高于平均身高时，他的儿子身高比他更高的概率要小于比他更矮的概率；父亲矮于平均身高时，他们的儿子身高比他更矮的概率要小于比他更高的概率。它反映了一个规律，即这两种身高父亲的儿子的身高，有向他们父辈的平均身高回归的趋势。对于这个一般结论的解释是：大自然具有一种约束力，使人类身高的分布相对稳定而不产生两极分化，这就是所谓的回归效应。

回归分析是建立因变量 Y（或称依变量，反应变量）与自变量 X（或称独变量，解释变量）之间关系的模型。如果在回归分析中，只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，这种回归分析称为一元线性回归分析。如果回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。对具有相关关系的现象，择一适当的数学关系式，用以说明一个或一组变量变动时，另一变量或一组变量平均变动的情况，这种关系式称为回归方程。

比如，我们知道每年天猫“双 11”的销售数据，就可以进行简单的回归分析，对下一年的可能销售量进行预测。

通过下面的数据，我们可以预测 2015 年“双 11”的销售额，然后把预测销售额与实际销售额进行对比，以最终验证方程的可靠性。

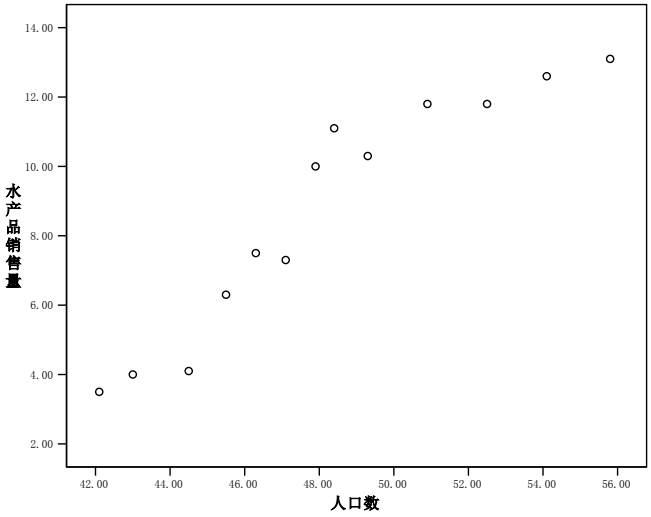


最后，我们认为，回归分析与相关分析关系密切，但差异也很明显。相关分析是用来度量变量与变量之间关系的紧密程度的一种方法，在本质上只是对客观存在的关系的测度。回归分析是根据所拟合的回归方程研究自变量与因变量一般关系值的方法，可由已给定的自变量数值来推算因变量的数值，它具有推理的性质。在进行相关分析时，不需要确定哪个是自变量，哪个是因变量，但回归分析的首要问题就是确定哪个是自变量，哪个是因变量。现象之间的相关分析只能计算一个相关系数；而回归分析时回归系数可能有两个，也就是两现象互为因果关系时，可以确定两个独立回归方程，从而就有两种不同的回归系数。

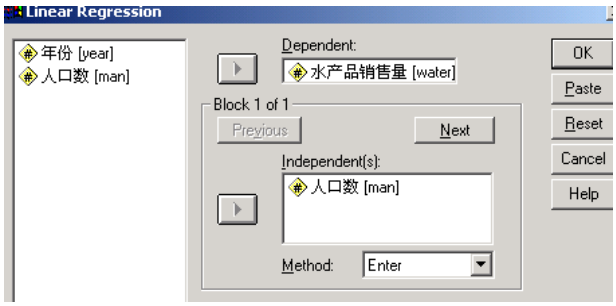
下面举一个一元回归预测的例子。某地区人口数与水产品销售量资料如下表所示，据此预测该地区 2005 年人口数达到 56.9 万人时，其水产品销售量会达到多少？

年份	人口数（万人）	水产品销售量
1992	42.1	3.5
1993	43	4
1994	44.5	4.1
1995	45.5	6.3
1996	46.3	7.5
1997	47.1	7.3
1998	47.9	10
1999	48.4	11.1
2000	49.3	10.3
2001	50.9	11.8
2002	52.5	11.8
2003	54.1	12.6
2004	55.8	13.1

首先用人口数和水产品销售量做出散点图，从趋势看，我们应用一元回归方法，设方程 $y=b+ax$。



我们使用 SPSS 中的 Regression 进行分析。



得到的结果如下：

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.943[a]	.890	.880	1.19920

a. Predictors: (Constant), 人口数

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	127.424	1	127.424	88.607	.000[a]
	Residual	15.819	11	1.438		
	Total	143.243	12			

a. Predictors: (Constant), 人口数

b. Dependent Variable: 水产品销售量

相关系数为 0.943，检验通过。

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	-28.899	4.011		-7.206	.000
	人口数	.780	.083	.943	9.413	.000

a. Dependent Variable: 水产品销售量

因此可以得到这样的方程　　$y= -28.899+0.78x$

我们得到的水产品预测值是　　$-28.899+0.78\times 56.9\approx 15.5$

通过计算标准差和查相关的 t 分布表（统计学术语，非专业人士可忽略），确定预测区间（12.3,18.6）。

以上的线性回归分析很简单，只是做一个示范。但是，在实际工作中，要特别注意线性回归的应用有 4 个前提条件：线性、独立性、正态性、等方差性。线性指因变量与自变量应大致呈一直线趋势。独立性指各观察值之间没有相关性。正态性指线性模型的残差要符合正态分布，实际中有时直接看因变量是否符合正态分布。等方差性指在自变量取值范围内，对于任意自变量取值，因变量都有相同的方差。

还有一种回归叫自回归，如果我们只知道依时间变化的一个变量，如股票价格、产值、工资水平等，就要用同一变量在不同时期中各个变量值之间的相关关系建立一元或多元回归方程，也就是说，用一个

变量的时间数列作为因变量数列，用同一变量向过去推移若干期的时间数列作为自变量数列，分析一个因变量数列和一个或几个自变量数列之间的相关关系，建立回归方程进行分析预测。

对于这样的预测，我们最好采用 SPSS 软件中 Time Series 程序来进行，但这个程序应用需要一定的技巧，在具体的预测实践中应该谨慎。下面我们介绍手动计算的方法。

例：某产品 1991～2003 年市场容量如下，据此预测 2004 年的市场容量。

年份	市场容量	后推一年	后退两年
1991	858.0	——	——
1992	929.2	858.0	——
1993	1023.3	929.2	858.0
1994	1106.7	1023.3	929.2
1995	1163.6	1106.7	1023.3
1996	1271.1	1163.6	1106.7
1997	1339.4	1271.1	1163.6
1998	1432.8	1339.4	1271.1
1999	1558.6	1432.8	1339.4
2000	1300.0	1558.6	1432.8
2001	2140.0	1300.0	1558.6
2002	2350.0	2140.0	1300.0
2003	2570.0	2350.0	2140.0

先取后推一年的自身回归模型，计算得到相关系数为 0.994，后推两年的自身回归模型相关系数为 0.982，因此，决定选用后推一年的自回归模型：

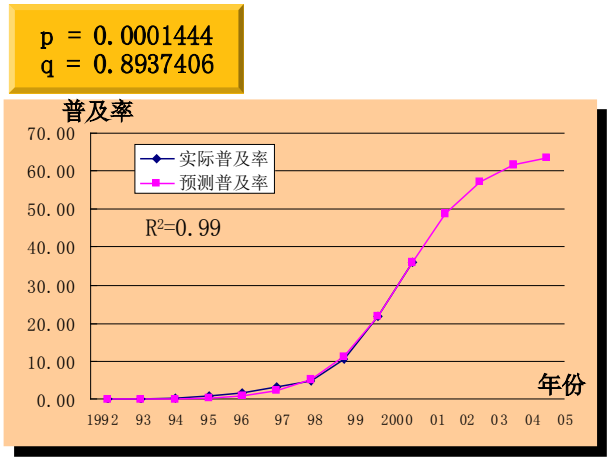
$$Y_t = b + a\ Y_{t-1}$$

我们可以通过回归分析方法计算，得到预测的方程：

$$Y_t = -57.5546 + 1.1416 Y_{t-1}$$

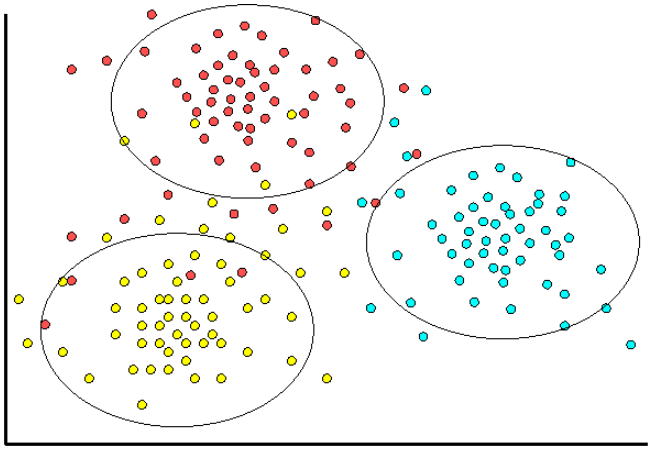
Excel 中制作的散点图、平滑曲线图等可以添加趋势线，也经常被用来进行市场预测，我们可以依据趋势线上面给出的公式，计算预测期的值。预测拟合的优劣可以通过 R^2 的值判断，越接近于 1，表明拟合程度越好，预测的结果越可信。

某通信公司根据 1992～2003 年的数据来预测 2004、2005 年的电话普及率，结果如下：



聚类、判别和因子分析

人以类聚，物以群分。聚类分析（Cluster Analysis）是根据事物（如消费者、产品、品牌、品牌属性）之间的相似性（Similarity）或同质性（Homogeneity）将它们归类分组的方法。聚类分析的结果寻求的是组内差异（Within-group Variation）最小，组间差异（Between-group Variation）最大。聚类分析也是数据分析中最经常使用的多元分析方法之一，它在有关市场细分研究中几乎是必不可少的分析工具。



在一次研究工作中，我们尝试通过聚类分析的方法，以全业务通信价值作为细分维度进行客户细分，从而进行客户分群、不同群描述、群识别等活动，然后通过制定精确的营销计划，达到用户购买产品的目的。其中，全业务通信价值结合了家庭客户价值（包括现有价值和潜在价值）和家庭客户业务取向（包括消费意愿和消费需求）两个方面。通过客户分群管理，在客户购买产品的过程中，通过产品设计和精确的营销，可以达到提升客户价值和提升客户消费需求、影响消费意愿（即业务取向）的目的。

从价值上看，影响城市家庭客户的现有价值因素主要包括消费心理、消费倾向等。如高端客户在消费决策时更多的考虑因素是休闲娱乐及自己生活和工作的需求，而且其对网络的应用和需求多样化。影响其潜在价值的因素主要包括收入、工作状态。如高收入家庭对移动电话和移动宽带的消费最高，而中低收入家庭的通信消费结构偏重于固话和固定宽带。

从业务取向上看，影响城市家庭客户的因素主要包括家庭收入、工作状态、消费心理等。如稳定高级工作人员（或中高收入的家庭）

对融合业务需求很大，对智能家庭需求明显；而中低收入家庭则更倾向于基础话音业务或亲情沟通方面的应用。

下表中列举了分群时所使用的变量的数量和这些变量的描述。家庭生命周期的维度并没有用到一开始的分群工作中，作为辅助分群纬度，并应用在客户群描述中（具体字段举例如下表所示）。

维　　度	变　　量	字段举例
场景	城市/农村	
客户价值	现实价值	ARPU 值
	潜在价值	学历
		收入
		家庭人口数
		家庭成员年龄
		家庭成员职业和职位
		忠诚度
业务取向	消费需求	本地区内通话费
		本地区间通话费
		传统国内长途通话费
		传统国际长途通话费
		传统港澳台长途通话费
		其他增值业务
		上网时长
		计费方式
		宽带入网时间
		宽带速率
	消费意愿	通信消费占总消费比
		消费占家庭收入比
		增值业务消费比例
		其他

通过以上分析，对各影响因素进行了一定的分组。再用客户价值

和业务取向两个维度形成一个二维表，从而对客户进行分群。家庭生命周期的维度并没有用到一开始的分群工作中，而是作为辅助分群纬度，并应用在客户群描述中。通过分别考虑中国城市家庭客户的消费能力、消费种类、家庭生活状态，以全业务运营下家庭客户的业务取向、现实及潜在价值为主要维度，将城市家庭客户分为六大类型。

1．高尚家庭（TOP）

高尚家庭用英文 TOP 表示，即高收入追求时尚的家庭，此类家庭的占比约为 2.4%。

此类家庭收入高，拥有名车豪宅，即居住在高档小区或别墅，开高档轿车；投资多样、消费高档，而且多元化，多为移动电话高端用户，对数字家庭有足够消费能力并具有消费倾向；同时，他们一般很少使用固话，对宽带上网速率的要求很高。很看重电信产品的品牌、品质和服务。追求高品质、高安全性的生活。一般由于工作等需要，社交很广泛，并且有一定的声望。

此类家庭是最好识别的家庭，主要通过家庭的住宅以及 ARPU 值消费即可识别。主要包括社会名流，如演艺文体明星和富二代；企业高管，包括企业家群体、专家学者和投资理财高手等。

2．潮流家庭（IN）

潮流家庭用英文 IN 表示，在现在的流行语中 IN=In Fashion，也就是流行中的、时尚的；它的反义词是 OUT，也就是过时的、落伍的。此列家庭的占比为 11.7%左右。

此类家庭拥有自有（按揭）住房及轿车，收入稳定、家庭核心成员在 25～35 岁之间，属于较年轻化的家庭。他们追求高品质的生活，

对时尚潮流敏感，喜欢各种各样的新奇事物和体验。学历偏高，收入颇丰，一般对品牌更加敏感。对电信公司来说，他们的通信消费高，对家庭信息化有需求，是数字家庭潜在消费者。

此类家庭主要包括一些单身贵族，他们对新鲜事物很敏感，注重消费体验方面的需求，还有商务白领或一般公务员、专业技术人员等。

3．传统家庭（ALL）

传统家庭并未有比较特殊的英文翻译，在中国传统家庭的占比较高，所以就用英文 ALL 表示，此类家庭占比约为 38.6%。

此类家庭主要是指中国传统的核心家庭，以三口之家为主，两口或三代同堂部分，主要强调家庭成员同房屋居住，同城市工作和学习。他们对固话、宽带和手机都有需求，是融合业务发展的主要目标客户。

主要包括有二人世界，也就是新婚夫妇或者结婚尚未有小孩的夫妇，他们追求健康的生活方式，看重消费品的品质；还有幼儿家庭，孩子在婴幼儿时期，家庭成员注重小孩的健康和保健，看重消费品品质和安全；学生家庭，家庭中孩子在学生时期，家庭注重小孩的教育和健康，看重消费品物美价廉；全职工家庭，家庭核心成员有稳定工作、稳定收入，有固定的消费需求，看重消费品品质。这些家庭都是在中国比较传统，占比很高的家庭。他们的消费行为可能存在一定的差异，但是总体上，这些消费行为都比较符合中国传统的消费习惯，所以将这些家庭归入此类。

4．“宅”家庭（SOHO）

“宅”一词是在网络兴起之后逐渐流行的一个词，针对这类家庭客户，我们也把他们划分为一类，称为 SOHO 一族，即“宅”家庭，此

类家庭占比 7.9%左右。

“宅”家庭即其家庭主要成员长期在家生活（工作），包括宅男宅女、SOHO 一族等，对网络有较强依赖，通话以手机为主，习惯使用电子商务、网络游戏、视频聊天、网络电话等。通过进一步细分，“宅”家庭又包括御宅族，即长期在家娱乐和工作，工作不稳定，但收入颇丰，对网络有依赖；还有一类是家庭办公型，他们长期在家工作，包括 SOHO 一族、自由职业者，其收入较稳定，他们追求时尚、潮流的生活方式。

5．离散型家庭（LONG）

在现在的中国家庭中，家庭成员已经并不一定长期住在一起，他们可能因为工作、学习等原因暂时异地分开。诸如近几年的出国热现象等，这些都导致了在中国离散型家庭的增加。在这里我们用距离上的 LONG 来代表离散家庭，此类家庭在中国占有将近 20.8%的比例。

顾名思义，此类家庭是指家庭成员（夫妻、子女等）长期异地分居生活，包括不同城市、城区与乡村、国内与国外等。包括城乡离散、城际离散和国际离散等。虽然家庭成员之间的距离不同，但是他们都有较强语音及视频沟通的需求，对长途优惠、网络电话等亲情沟通方面的服务和产品有很高的需求。

6．简单化家庭（EASY）

在中国的城市还存在这样一类家庭，由于各种原因，他们生活比较清贫，过着简单的生活，所以我们用 EASY 来表示此类家庭，这类家庭占比约 18.6%。

简单化家庭一般属于中低收入工薪阶层，家庭核心成员受教育程

度不高，年龄在 50 岁以上或幼儿处在小学以下，也包括特殊低保家庭（如空巢老人、残障人员）等，主要为语音通话方面的需求。

此类家庭还包括低收入劳动者，他们的家庭核心成员学历偏低、收入偏低、家庭消费追求价廉；普通空巢老人，即家庭核心成员为老人，收入偏低，消费需求简单；特殊家庭，即家庭成员有各种身体或智力方面的残障，生活负担很重，对各类消费需求低。此类家庭虽然可能电信价值较低，但是这些家庭客户也是公司发展中所不可忽视的。

我们还可以看到其他的案例。比如，淘宝通过 2015 年的年货指数对不同年龄的人进行了系统的画像分析。

00 后：当你还在想给宝贝准备最新发布的电子书时，他们可能对电子书配上的定制保护壳更感兴趣。淘宝年货数据显示，18 岁以下的淘宝用户喜欢不停地更换各种手机保护套，其排到了年货购买第四位。此外，创意礼品也进入该年龄层次的热门购买榜单。

90 后：当你觉得 90 后已经是个大人的时候，其实他们依然深藏着童心，淘宝年货数据显示，毛绒玩具类悄悄地占领了 90 后年货热购榜单。在淘宝上，1 ~ 3 米高的大型公仔玩具是 90 后最爱。除泰迪熊之外，河马、龙猫、超萌的哈士奇，这些卡通形象是 90 后女孩首选。

80 后：80 后纷纷进入家庭状态，他们囤得最起劲的年货就是尿不湿和奶粉，顺带买上一套益智玩具。初为人父人母，80 后同样孝心爆棚，在中老年服装等新年礼物的购买上，也很舍得花钱。对于自己的礼物，80 后对烘焙表现出浓厚兴趣，其中烘焙原料是他们的最爱。

70 后：70 后的生活相对枯燥平庸。年货指数中，40 ~ 49 岁的群体最爱买衣服、鞋包、化妆品及零食等。这个年龄的用户开始养起了宠物。狗狗主粮进入了年货热搜榜单。在一周之内，他们买走了 3 万多

份狗粮。

60 后：60 后才是真正意义上的爱俏一族。数据中，连衣裙排在他们的年货购买榜单第一位。面膜、香水、BB 霜、染发膏等也跻身热购榜单。在面膜的购买排行上，50 ~ 59 岁的购买排行甚至超过了 30 ~ 50 岁的用户群体。其次，老人喜欢滋补品。在淘宝上，铁棍山药、黄精、党参养胃茶、阿胶糕等位列药食同源的榜首。

50 后：50 后最爱手机，也会洋气地进行鲜花同城速递。数据显示，60 ~ 69 岁老人一周之内在淘宝上买走了 8.5 万多部手机。但别以为 200 多元的老人手机就能搞定花样爷爷，他们可是要买能实行亲情定位及申请儿女远程协助等功能的高端老人机。

判别分析又称分辨法，是在分类确定的条件下，根据某一研究对象的各种特征值判别其类型归属问题的一种多变量统计分析方法。其基本原理是按照一定的判别准则，建立一个或多个判别函数，用研究对象的大量资料确定判别函数中的待定系数，并计算判别指标。据此即可确定某一样本属于何类。当得到一个新的样品数据时，要确定该样品属于已知类型中的哪一类，这类问题属于判别分析问题。

判别分析的具体操作方法是，由若干个不同总体的样本来构造判别函数，以此决定新的未知类别的样品属于哪一类。例如，某医院已有 1000 个分别患有胃炎、肝炎、冠心病、糖尿病等的病人的资料，记录了他们每个人若干项症状指标数据。利用这些资料，在测得一个新病人若干项症状指标的数据时，能够判定他患的是哪种病；又如，在天气预报中，利用长时间的记录资料，判断是晴天或下雨天等。比如诸葛亮的神机妙算，很多应该是利用了判别分析。我们在实际工作中也会用判别分析来判断哪些是忠诚用户，哪些用户可能会流失。

子分析是指研究从变量群中提取共性因子的统计技术，最早由英国心理学家 C.E.斯皮尔曼提出。他发现学生的各科成绩之间存在着一定的相关性，一科成绩好的学生，往往其他各科成绩也比较好，从而推想是否存在某些潜在的共性因子，或称某些一般智力条件影响着学生的学习成绩。因子分析可在许多变量中找出隐藏的具有代表性的因子，将相同本质的变量归入一个因子，可减少变量的数目，还可检验变量间关系的假设。因子分析有着广泛的应用，可以用来进行消费者习惯和态度研究（U&A）、品牌形象和特性研究、市场划分识别、顾客、产品和行为分类等。

楼市命悬“一线”，“刚需”去哪里了

刚需，一个被中国楼市创造出来的词汇，终于落幕了。人们惊讶地发现，所谓的“刚需”根本不存在，需求仍然是要“有效需求”，脱离了购买能力的需求都是痴心妄想，而曾经看似无处不在的中国购房人的需求不再刚性。

这次的供给侧改革等于是彻底抛弃了“刚需”理论，世界上本来就没有什么刚性需求。大多数人都有住在北京王府井的需求，可一个月收入 3000 元的人有这种需求，如果不是祖上荫功，就只能是痴心妄想。这样所谓的“刚需”根本就不是需求。

在房地产高速发展的 20 年里，到处充斥着骗人的鬼话，什么丈母娘推高房价，什么买房就是爱国，什么有房才有青春，不管收入多少，不管条件怎样，不管年龄多大，人人都要去买房，把这样本来不切实

际的需求催化成了不考虑现实和未来的盲目冲动消费，后果可想而知。

经过多年的刚需忽悠之后，需求断档了，突然之间，开发商、中介、房地产专家们都找不到“刚需”去哪里了。

一线城市的刚需还存在吗？

现在，全中国的开发商都在紧盯着一线城市（也就是北上广深，即北京、上海、广州、深圳的简称）的房地产价格走势，因为只有这几个城市还被认为房价是坚挺的，未来也不会差，只要这几个城市有一个房价出现问题，面临的将是逃跑者的城门口踩踏。

数据表明，如今的北上广深房价不仅仅是坚挺，而且是“发烧”。2015 年 10 月，深圳房价同比上涨 39.9%，上海和北京房价分别同比上涨 10.9%和 6.5%。一线城市无一例外不是在大涨，而在全国 70 个大中城市中，房价环比上涨的城市只有 27 个。

事实上，在 2015 年，一线城市的房价都经历了轮番上涨，要是在往年，政府一定会采取各种限制措施，现在却基本都是偃旗息鼓的静观事态，只有北京市政府动迁造成通州房价异常之后采取了局部的限制措施。

现在在一线城市买房的，除了深圳那些根据一个虚无缥缈的直辖梦编造的故事而盲目投机的，就主要是前几年被大家定义的“刚需”，也就是此前因为各种限制无法买房，而现在终于等到解禁的人，可这些人之后呢？经过最后一轮的炒作，一线城市的所谓刚需也基本榨干了。

2015 年之前的 10 多年间，北上广深都有大量的人口涌入，每年新增常住人口都有 30 ~ 50 万人，大量的年轻劳动力涌入，还有的卖掉了

老家的房子，把资金和满腔的热爱撒向了一线城市，所以这些地方一年卖出 8 万套房子，也就在合情合理之中。那么我们要反过来研究了，这个持续涌入的情况会继续下去吗？事实上，北京在 2015 年仅仅新增了 19 万常住人口。

有媒体报道，北京新房总价迈入 464 万元时代，而全国上下，身价超过 600 万元的人大概只有不到千分之五，全国也就 100 ~ 200 万人，即使北京的比例高一倍，达到百分之一，2000 万常住人口中，也只有 20 万户能买得起，这里面九成的人不仅有房，还可能都有多套房，所以大约要买房的（包括换房）也就不到 2 万户。实际上，在北上广深买房早已经不是居住需求，而是资本投机的游戏。

中小城市的刚需还有吗？

中小城市的房地产已经严重失衡，一方面是楼房空置，另一方面是代售楼盘林立，更严重的是，这些城市以前创造刚需的手段不灵了。

以前，这些城市通过城市改造造就了大量的刚需购房者，这些人虽然分到了回迁房，但因为房子在涨价，而这些人手中的钱也没有其他的投资渠道，往往只能把钱再一次都投入到购房中，随着中国经济的新常态，很多城市已经没有能力继续造城，大拆大建很难继续，这些人的购房刚需也就消失了。

中小城市还有一个寄托，那就是进城打工的农民工，可因为房价太高，农民工的刚需并不是有效需求，很多老一点的农民工希望的是在外挣钱，然后回家造个大房子。更为严重的是，一些在一线城市打工的农民工用辛辛苦苦积累的钱在老家的城市买房，一年也不住一次，造成了事实上的空置，连出租都出租不出去。

东部的城市人口还在增加，主要是人口迁移过来的，而东北、西部的一些城市已经出现了人口萎缩现象，不仅是农村人口转移殆尽，连中小城市的人口也在向外流失，未来房屋出售的情况比新购的需求还多。

年轻人的刚需还存在吗？

以前，房地产行业经常把年轻人结婚买房当成最典型的“刚需”，还炒作出丈母娘的故事，后来，年轻人结婚就真的把买房当成了必需品。这种刚需支撑了中国房地产十年的发展。

中国房地产的快速发展，得益于人口红利，也得益于中国社会家庭结构的变动，一个一个大家庭的分开居住大大提高了房屋销售的套数。但是，随着人口增长放缓，中国的大家庭的分解已经结束，核心家庭早已经成为中国社会的主体，这部分强烈的购房需求也消失了。

年轻人结婚需要婚房，这个社会风气已经形成，短时间内难以改变，但是，随着老龄化的加速，正常的人口新陈代谢使得房屋已经不再稀缺，很多家庭有多套住房，年龄小一点的年轻人结婚的时候，一些房子已经空出来，即便这些年轻人不愿意住老房，可出售旧房产也会大大提升房屋的待售存量。和炒股一样，大量的人需要卖股炒新，势必造成股市价格的暴跌。

世上本无刚需，我们非要创造出这样一个词，这本来就是涸泽而渔，也是房地产市场的提前消费。在我们把大量的非有效需求忽悠或逼迫成刚性需求的时候，就注定了将来要承受泡沫破裂之苦。当然，我们还有最后的办法，用教育和户籍差异化措施催生新的刚需，可那就更是在走向灭亡之路了。

大数据分析可能用到的软件

Excel 作为电子表格软件，适合简单统计需求，缺点在于功能单一，且可处理数据规模小。SPSS（SPSS Statistics）和 SAS 作为商业统计软件，提供研究常用的经典统计分析（如回归、方差、因子、多变量分析等）处理。SPSS 轻量、易于使用，但功能相对较少，适合常规基本统计分析。SAS 功能丰富而强大，且支持编程扩展其分析能力，适合复杂与高要求的统计性分析。

数据挖掘作为大数据应用的重要领域，在传统统计分析基础上，更强调提供机器学习的方法，关注高维空间下的复杂数据关联关系和推演能力。代表是 SPSS Modeler（前身为 Clementine），

SPSS Modeler 的统计功能相对有限，主要是提供面向商业挖掘的机器学习算法（决策树、神经元网络、分类、聚类和预测等）的实现。

RapidMiner 是一个预测分析的神奇工具，功能强大，易于使用，而且背后有一个优秀的开源社区。你甚至可以把你自己的专用算法通过 API 集成到 RapidMiner 中。

另一个商业软件 Matlab 也能提供大量数据挖掘的算法，但其特性是更关注科学与工程计算领域。而著名的开源数据挖掘软件 Weka，功能较少，且数据预处理和结果分析也比较麻烦，更适合学术界或有数据预处理能力的使用者。

当然，我们不应该忘记 Oracle。Oracle 数据挖掘允许用户利用他们的 Oracle 数据发现洞察，做出预测。你可以构建模型发现客户行为，定位最好的客户和对他们画像。

近两年来出现了许多面向大数据、具备可视化能力的分析工具，

在商业研究领域，TableAU 无疑是卓越代表。TableAU 的优势主要在于支持多种大数据源/格式，众多的可视化图表类型，加上拖拽式的使用方式，能够涵盖大部分分析研究的场景。

Gephi 是免费软件，擅长解决网络分析的很多需求，其插件众多，功能强且易用。我们经常看到的各种社交关系/传播图谱，很多都是基于力导向图（Force Directed Graph）功能生成。

Qubole 能够简化、加速并扩展大数据分析工作负荷，数据存储在 AWS、Google 或 Azure 云上，省去了基础设施建设的麻烦。一旦 IT 策略到位，任意数量的数据分析师可以被解放出来，利用 Hive、Spark、Presto 和越来越多的其他数据处理引擎的力量协作进行“单击查询”。

基于自然语言处理（NLP）的文本分析，在非结构化内容（如互联网/社交媒体/电商评论）大数据的分析方面有重要用途。其应用处理涉及分词、特征抽取、情感分析、多主题模型等众多内容。

BigML 试图简化机器学习。它提供了强大的机器学习服务，易用的用户界面，可以导入你的数据并得到预测。你甚至可以使用它的模型进行预测分析。

前面介绍的各种大数据分析工具，可应对的数据都在亿级以下，也以结构化数据为主。当实际面临亿级以上/半实时性处理/非标准化复杂需求时，通常就需要借助编程（甚至借助于 Hadoop/Spark 等分布式计算框架）来完成相关的分析。

Hadoop 这个名字已经成为大数据的代名词。它是一个开源软件框架，用于非常大型数据集在计算机集群上的分布式存储。这意味着你可以向上和向下扩展你的数据而无需担心硬件故障。Hadoop 为所有类型的数据，极大的处理能力和几乎不受限制的并发任务/作业的控制能

力提供了海量存储。

MongdoDB 是一个现代的初创数据库方案。可以把它们作为关系数据库的替代，适用于管理频繁更改的数据或非结构化和半结构化的数据。常见使用案例包括存储移动应用数据、产品目录、实时个性化、内容管理和提供跨多系统单个视图的应用程序。MongoDB 也不适合新手。与其他任何数据库一样，你需要知道如何使用一种编程语言查询它。

当前适合大数据处理的编程语言，包括：

R 语言——最适合统计研究背景的人员学习,具有丰富的统计分析功能库及可视化绘图函数可以直接调用。通过 Hadoop-R 更可支持处理百亿级别的数据。相比 SAS，其计算能力更强，可解决更复杂更大数据规模的问题。

Python 语言——最大的优势是在文本处理及大数据量处理场景，且易于开发。在相关分析领域，Python 代替 R 的势头越来越明显。

Java 语言——通用性编程语言，能力最全面，拥有最多的开源大数据处理资源（统计、机器学习、NLP 等）直接使用。也得到所有分布式计算框架（Hadoop/Spark）的支持。

此外，OpenRefine（之前叫 GoogleRefine）是一个开源工具，专注于清理杂乱数据。你可以轻松快速探索大型数据集，即使数据不太结构化。DataCleaner 可以为你完成数据清理的工作，将杂乱的半结构化数据转换为干净可读的数据集，使得所有可视化工具能够使用。

Import.io 是最好的数据提取工具。使用一个非常简单的点击界面，打开一个网页并将其转换为易用的电子表格，之后可以用来分析、可视化和做出数据驱动的决策。

Tableau 是一款数据可视化工具，主要侧重商业智能。无需编程就可以创建地图、条形图、散点图和其他图形。他们最近的版本包含一个 Web 连接器，允许你连接一个数据库或 API，从而在数据可视化中获得实时数据。

CartoDB 是一个专门制作地图的数据可视化工具。它使任何人都可以轻松可视化位置数据，无需任何编程。CartoDB 可以处理大量数据文件和类型，甚至还包含你可以操作的样本数据集。

Chartio 允许合并数据源并执行浏览器内查询，通过几次点击创建强大的仪表板。Chartio 的可视化查询语言让任何人从任何地方获取数据，而不需要懂得 SQL 或其他复杂模型语言。它也可以生成 PDF 报告，从而能够导出仪表板，并用电子邮件发送给你希望的任何人。

FusionCharts Suite XT 不仅可以带给你漂亮的图表，还能帮你制作出生动的动画、巧妙的设计和丰富的交互性。它在 PC 端、Mac、iPad、iPhone 和 Android 平台都可兼容，具有很好的用户体验一致性，同时也适用于所有的网页和移动应用，甚至包括 IE6、7、8 这些绝大部分插件都不支持的主儿。在这个软件里，创建你的首幅图表也只需要 15 分钟。

Dygraphs 是一款快捷、灵活的开源 JavaScript 图表库，用户可以自由探索和编译密集型数据集。它具有极强的交互性，比如缩放、平移和鼠标悬停等都是默认动作。更棒的是，它还对误差线有很强的支持。Dygraphs 也是高度兼容的，所有的主流浏览器都可正常运行（包括不受待见的 IE8）。你甚至可以在手机和平板设备上使用双指缩放。

ZingChart 是一个强大的库，为用户提供了快速创造漂亮的图表、操作面板和信息图表的可能性。你可以在上百种图表类型中自由选择，

你的设计和个性化要求不会受到任何限制。你也可以使你的用户通过交互式图表特性参与到你的作品之中。

（以上软件介绍来自网络，相关软件部分可在网络上免费使用，但也有需要付费）

第5章

大数据，有时候很奇葩

看懂经济形势，奇葩大数据靠谱吗

网上流传一则阿里巴巴校园招聘的最新通知：“我们非常抱歉地通知：由于集团人才战略调整，阿里巴巴 2016 年校招名额确定将要缩减（名额从 3000 人减到了 400 人），各岗位将执行更加严格的“择优录取”标准。很多人评论说，这次阿里巴巴缩招降薪事件，预示着互联网冬天的到来，而最近一段时间确实有很多行业分析人士都在讨论互联网冬天的问题。

随着大数据的流行，人们越来越关注通过日常行为的观察来分析经济发展走势，希望一叶知秋。事实上，确实有很多人总结过一些颇具参考价值的社会信号，成为了经济分析的另辟蹊径。

发型长短

多年前，据日本最大日用品制造公司“花王”的“发型统计”调查显示，女人在蓄长发时显示经济在复苏中，反之则经济仍在恶化。例如：1997 年，留短发的比蓄长发的人多，那年为日本经济“最差”的一年，2008 年经济有所起色，超过八成受访女性头发都很长。

评：最近一些年，短发女星确实很受欢迎。

服务员颜值

还有更“靠谱”的，在如今刷脸时代，颜值处处都在。据纽约观察员的解读，当美艳的女服务员随店可见时，经济必陷困境，反之则显示经济兴旺，换句话说，当你到处碰见美女服务员，便可考虑抛售股票。观察家的解释是，当经济红火，颇有点“资本”的女性很容易找到工作环境舒适，即不属厌恶型行业的工作，诸如商品模特、推销

员等，此外，男性经济宽裕后更容易“金屋藏娇”。

评：豆腐西施、包子西施等已经成热门网络人物。

裙边理论

还有更广为人知的“裙边理论”，女性的裙子长短与经济发展密不可分。这个理论认为女人的裙子长度和社会经济情况成反比。“裙边理论”的提出者发现，20 世纪 20 年代和 60 年代繁盛时期的美国，妇女普遍选择短裙，裙边向上收，结果股市也随之上扬；相反三四十年代的经济危机时，她们选择穿长裙，市场也逐渐走低。也就是说，女人盛装打扮、着装性感是经济大好的兆头。

评：这两年突然发现小短裙不再被人关注了。

口红效应

“口红效应”是一种很有趣的经济现象，是说美国每逢遇到经济不景气时，口红反而会大卖。这是为什么呢？因为在经济萧条时，女性收入不多，不会再像以前那样随性地买一些时尚、赶潮流的衣服、化妆品等高端商品，尤其是奢侈品，而是趋向于购买那些性价比较高的用品。口红作为最便宜的奢侈品，既能满足女性的购物欲望，又能缓解经济低迷时的不好情绪，带来心理慰藉，最关键的是能够使女士们保持妩媚迷人的娇容，所以，口红才会有很大的市场。也有人因此而认为“口红效应”意味着经济不景气。

评：以卖化妆品为主的几家网站生意不错哦。

内裤指数

比如，美联储前主席格林斯潘（Alan Greenspan）曾提出过一个著

名的“男性内裤销量反映经济形势”的理论。即经济形势良好，内裤销量会平稳上升，反之则下降。原因很简单，在经济不景气时，男人不得不节省消费开支，不再经常换新内裤了；同时经济不景气导致离婚率上升，离婚的男人不再注重自身形象了，对新内裤需求会骤降。而随着经济复苏，可以稍加挥霍了，离婚的男人也要找新对象，于是内裤销量自然攀升。内裤这种内在的用品，不像口红、裙摆，不是那么容易就能让其他人看得见的，即使很破旧，也无关紧要。所以，内裤作为男人的必需品，其销量曲线一直以来都很平稳，没有太大的波动。但要是其销量曲线上出现少量的下滑，则表明经济开始走向萧条了。

评：问“淘宝君”吧。

票房指数

在宏观经济学上，很多人认为电影票房与整个经济环境的变化不无关系，严峻的经济环境反而有效促进了票房的走高。研究者认为，在经济不乐观时期，那种既能够满足消费者心理慰藉需求，价格又低到能消费得起的产品，往往能够获得更好的市场待遇，这表现为一种“低价产品偏爱趋势”，影院能够让人沉浸其中而获得短暂的心理慰藉。

评：今年好像是个电影都火爆，从来没有过的火爆。

克强指数

所谓“克强指数”，是英国著名政经杂志《经济学人》创造的用于评估中国 GDP 增长量的指标，以中国现任总理李克强的名字命名。“克强指数”是三种经济指标：耗电量、铁路运货量和银行贷款发放量的结合。不过，Gavekal Dragonomics 创始合伙人兼研究主管 Arthur

Kroeber 指出，“克强指数”正在被大量地滥用。该指数如今更多地反映了中国信贷和重工业发展状况，并不是一个了解中国经济全貌的好指标。近年来该指数可以说已经名誉扫地，现在更是“相当无用”。

评：大家很久没听到过拉闸限电的消息了。

4G 使用量

华尔街投资研究机构 SanfordC.Bernstein 亚太分析师 Michael Parker 已经构建了自己的等式，主要专注于中国日益增长的消费水平，并试图推动中国向服务型经济转型。他以电影票房收入和 4G 使用量等元素来制定反映中国实体经济的指标。

评：虽然 4G 使用量增长很快，但老百姓更吐槽费用难以忍受。

的士司机的谈吐

有这么一个指数，有些添堵——“读饱书的士司机指数”，每当大家在搭的士时经常碰上谈吐文绉绉的士司机时，不必查询 GDP 数据，便可断定经济已陷入或快将进入衰退。理由是：连有知识的“文化人”（本科以上学历）都来开的士了，那失业率可想而知。

评：看来，专车这么受欢迎，是这个原因啊！

我国航班正点率属国际中上水平

2010 年 8 月，民航局副局长夏兴华在回应飞机延误屡屡出现时称，中国民航航班的平均正常率一直保持在 80%左右，在国际上属于中上等水平。此言一出，社会哗然。真以为我们没坐过飞机吧！

经常坐飞机的人都有感觉，飞机延误是家常便饭，延误一个小时都可以说是正常航班，延误三四个小时的也不少见。资料显示，2013年，全国航空公司共执行航班278.0万班次，其中正常航班201.1万班次，不正常航班76.9万班次，平均航班正常率为72.34%。

从2013至2015年11月的统计看，航班大面积延误的高发期往往出现在旅客出行量较高的暑运期间。据飞常准统计，7、8月份的航班正点率为全年最低，仅为65%左右。

为什么7、8月份航班正点率最低？根据大数据的分析，从一年中航班不正常原因分类统计结果来看，航班不正常原因37.4%由航空公司引起，天气原因仅占21.8%，流量原因占27.60%。夏季雷雨大风较多，所以航班不正常情况比较严重。

是不是全球航空公司的正点率都不高呢？据资料显示，一些国外航空公司正点率高达90%左右。飞常准对全球主要的上百个航空公司的客运航班到港正点率进行排名，其中正点率最高的航空公司是芬兰航空，96.94%；第二、三、四名均为日本的航空公司。

有人说，航班正点率降低，与中国航班数量激增大有关系。中国作为新兴航空市场，2007年航班总量为167.2万班次，航班正点率达到了83.19%；而6年间，航班总量则以平均每年10%左右的速度递增。因航班增加较快，而机场有限和相关设施不足，到2012年航班正点率已逐步下降到了74.83%。

不过，以上的分析恐怕就很难站得住脚，正点率与航班数量之间的反比关系也不一定成立。航空数据服务商OAG报告称，2014年拉脱维亚的低成本航空公司波罗的海航空总体正点率最高，达到94.9%；2014年正点率表现最好的航空公司多集中在欧洲，前20名中有13家

位于欧洲，5 家位于亚太地区，2 家位于美国。可见，飞机起降的频次和正点之间，没有必然联系。

事实上，我们以上关于正点率的比较也不十分合适，因为你知道什么是“正点率”吗？

根据世界民航协会的有关规定，飞机在关闭舱门后允许有 15 分钟的时间。例如，你买的机票上写的飞机起飞时间是 12:00，飞机在 12:15 时间内起飞都算正点。原因一是机场飞机的起飞和降落时受航空管制部门指挥，二是考虑飞机的飞行安全性。

但是，一直以来，我国民航采用的正点起飞标准是关舱门，只要某趟航班按照机票上的时间关闭舱门，即只要旅客被按时关进飞机，这趟航班就算“正点”了。至于飞机何时才能起飞，旅客要在机舱里等多久，以及飞机何时到达目的地，这些则一概不论。由此来看，我国的所谓“正点率”比世界民航协会的标准要“宽松”得多，这样统计出来的数字依然如此之低，更令人扼腕叹息。

也正是因为“关舱门”这个标准被曝光，社会舆论哗然，民航局顺应民意修改了正点率计算的标准。从 2014 年起，中国民航局开始执行航班正点起飞统计新标准“撤轮挡”。撤轮挡是一个全球民航界通用的专业术语，如同地面的汽车一样，为避免汽车溜车，汽车停在车位里时，在汽车前面放置一个挡板。飞机撤掉轮挡后，即可启动发动机而滑行。

按照民航业内专家说，与原来关舱门即为航班正点的统计标准相比，新标准意味着，不仅航空公司做好了旅客全部上齐的准备，而且机场做好了行李装上飞机，航油做好了加油，空管发出飞机滑行到跑道上起飞的指令。

可以说，民航局从“关舱门”改为“撤轮挡”是一个伟大的进步，毕竟，相对于关舱门，撤轮挡距离起飞时间近一些，统计出来的“正点率”更真实一些，但这个进步是相当有限的，五十步笑百步而已。撤掉轮挡后飞机固然可以滑行，但常坐飞机的人都知道，有时候飞机滑啊滑啊就是不飞，还可能停下来等待很久。如此，飞机前进了一小步，正点率也提升了一大步。

随着新标准的实施，为了追求名义上的正点率，航空公司就让更多飞机先滑行再等待，也提高了航班的正点率。再后来，反正上有政策，下面一定还有对策。

这个案例告诉我们，做任何的分析，一定要先下功夫弄明白指标的定义，概念要一致，口径要统一，离开了这些，做的分析就会成为笑话。

最后，我们看看有没有什么办法能提高正点率呢？媒体报道，2012 年 11 月，华北地区流量管理及多机场放行协同系统（简称 CDM 系统）正式启用，就是一个较好的案例。据统计，运行一个月，首都机场 11 月的航班平均正常率提高了 3.21%，石家庄机场航班平均正常率提高了 7.7%，天津机场航班平均正常率提高了 3.88%，北京南苑机场航班平均正常率提高了 2.83%，四个机场航班正常率显著提高。另外，2013 年 12 月，民航局京昆航线将原来空中“单行”通道，升级为“双向”飞行，不仅正点率提高，飞行安全也提高了 50%，有 54 个机场 1100 个航班受益。看来，办法总是比困难多，就看想与不想了。

为什么互联网专车会造成城市拥堵

“大数据”概念很火，甚至连政府高层都在喊，这个有点像几年前的物联网。但概念再好，要落地还需要时日，物联网空喊了多年之后，到现在也只能是刚刚有点成形。大数据的概念很多，但大家讲来讲去都是那些事，很有点像传说中的朋友讲的朋友家的真事。

就现在的现实情况来看，大数据喊得最凶的是做计算机 IT 系统的，这些人往往把大数据给解释成通过 IT 系统的软件就可以包治百病，把大数据的分析给简单化为纯粹的编写程序和数学计算，还有一些做硬件的公司，把大数据给解读成数据的采集、存储，其实，这些都只是大数据的最初级阶段，与应用还差得很远。大数据和我们原来理解的数据一样，要想有价值，关键在于分析和使用，而大数据的应用与纯数学最大的区别就是那些数据不管大小多少都是有生命的，脱离了社会现实去做数学计算，毫无价值。

据说，阿里巴巴基于大数据构建了“RTB 广告交易平台”，名为 Tanx，能够实现让广告主从购买媒体变成直接购买用户。用户在购买了商品之后，你就被贴上了偏爱标签，卖商品的商家可以通过交易平台“买下”你，接着该平台会跟踪你的浏览行为，在你浏览其他网站的时候，恰到好处地把该商家的广告推送到你面前。而且，整个购买过程采用实时竞价的方式，即 RTB（Real Time Bidding），价高者得。这确实是大数据的一个应用，但也处在初级阶段，数据结果的有效性仍在探索。

一般的说法，大数据分析从原来统计分析看重的因果分析转为相关分析，只探究知道是什么，而不重点探究为什么。其实，在大数据的背景下，分析原因将变得更为重要，也更需要定性和直觉，因为大

数据分析经常会给出风马牛不相及的结论，只有后续进行深入细致的因果分析，才会更有价值。数据表明了用户的喜好有相关性，但因果关系却不一定，弄错因果，差异巨大，比如那个“尿裤与啤酒”的经典结论，却不是任何地方都可以照搬。

大数据分析重视对行为中的关联性研究进行预测，这种预测应该是具有预见性的，而不是说简单的联系。如果一个人在网上买了项链，然后这个人看视频的时候就弹出项链的广告，这种体验不仅不会增加购买，相反有时会让客户懊悔不已。我们需要找到的是看什么视频的人会买项链，买哪款项链。

还有一个最为眼前的案例，中央电视台在春运期间做了个“据说春运”的节目，用百度地图的位置数据来描述中国人春运的旅程，界面直观，刚刚轰动社会的南方扫黄事件中，百度地图再一次被研究者挖掘出事件前后发生地的人员流入和流出及目的地，算是娱乐民众，反映现实。这也算是大数据的应用，但这些数据的真实性还有偏差，特别是春运的分析，中国带有 LBS 功能的智能终端还没有普及到足以反映社会现实的代表性地位。

大数据的分析会大量收集用户的数据，虽然有一定的方法可以减小数据噪声的影响，但却也是不可能忽略的，“精确性不再重要”也只是适度而已，不能用不重视精确性的幌子来随便使用乱七八糟的大数据进行分析。可以肯定地说，不做任何清洗的大数据分析绝对不会有进行抽样统计得到的结果更好。

大数据分析需要连续的、真实的、少杂质的数据，而这些数据对于大多数中国企业而言简直是天方夜谭。在中国，也许银行、航空、电子商务公司好一些，其他的，即便是通信运营商也是支离破碎、断

断续续、真真假假的数据，这样的大数据分析就非常不靠谱了。

现在做大数据分析和应用，除了互联网公司 BAT 有相当的资本，其他主要都是徘徊在灰色地带，靠窃取盗用非法采集用户的数据来挣钱，更有无耻到直接售卖用户的个性化隐私信息，这不是大数据，这只是大数据的蛀虫，真正的大数据应该是用数据分析的结果来指导消费而非原始数据信息的本身价值。

我们应该好好利用自身掌握的珍藏的海量数据，充分利用大数据为企业发展和业务创新服务，但也不能太神话，在数据分析面前，智慧永远比算法和数量更重要，数据的多寡和技术的高低并不是决定结果是否有价值的核心标准。也许，在现阶段，大数据就像陪伴在纣王身边的妲己，看似妩媚动人，实际上却是狐狸精的化身，一旦盲目信从，后果不堪设想。

比如，某部一位官员说，互联网约租车（专车）的使用加剧了这些城市的交通拥堵，理由是约租车高峰期与城市道路拥堵的高峰期相吻合。

一时间，这样的言论被看成笑谈。更多的人认为，这种分析把因果关系弄反了，正是因为城市道路拥堵，出租车司机为避开拥堵时段而不上路，所以导致互联网叫车的增加。应该说，约租车的高峰是城市道路拥堵的高峰带来的结果，而不是原因。

这样的分析肯定是有道理的。大数据的分析和使用确实可以发现很多关联现象，比如啤酒和尿布，但这种分析的结果却不能反映因果关系，一些分析者会因为先入为主的偏见而将因果关系故意颠倒。我们经常会说，哪里有警察，哪里就堵车，和这个分析出错的原因是一样的。

但是，互联网专车的高峰与城市道路交通拥堵的关系并不是因果关系倒置这样简单，我们还需要从更多的层面去分析，因为，有关方面认为互联网专车的上路就是拥堵的原因之一。

一般来说，导致城市道路拥堵，我们可以从四个角度来看：

一是出行的必要性。也就是说，不必要的出行或者是开车出行会大大提高交通道路的承载量，人为地制造拥堵，如果非必要的出行减少，道路就会更通畅。

二是路权之争。有车的人和无车的人，步行的人与开车的人，肯定不是固定的，拥有汽车的人不能要求没有汽车的人就不要用车不要打车，让开自己私家车的人有通行的权力，也不要干涉没有私家车或正好赶上限行而不能开车的人的打车自由。

三是每个人占据的道路的面积大小。显然步行的最少，公交车仅次，自行车次之，开私家车会占用路面更多，如果一个人开车当然就会更多，所以，很多国家会规定中心城区不能一个人开车驶入。

四是出行者在路上停留的时间。这个与出行的距离有关，也与出行的速度有关，如果大家距离短（比如从家到单位距离短），行驶快（都开快车），那么交通就会减少拥堵，相反，如果出行的人数多，通行慢，那么停留的时间就会长，会导致更加严重的拥堵，而更加严重的拥堵就会导致通行更加慢，就会更堵，恶性循环。

基础理论讲完了，我们看看互联网专车是否会导致拥堵呢？

首先，我们看，专车的出行会导致不必要的行驶吗？简单来看，确实如此，因为这些所谓的专车大多是可以趴在停车场不出来的，只是因为有人约车才上路，等于是人为地增加了上路行驶的车辆，特别

是在拥堵高峰期，这几乎就是在造堵。但我们也要看到，因为大多数互联网专车并不进行道路巡航扫街，而是接到附近的订单之后才会通过最短的距离接到乘客，从而在最大程度上避免了空驶而带来的无价值的道路占用，所以，互联网提高道路利用率的作用还是非常大的。

其次，我们再看看关于路权的问题。当然，互联网专车司机的上路只是为了挣钱，那么这些车辆的路权在高峰期被剥夺是有理由的，但是，那些通过互联网叫车的打车人的路权却是应该被保障的。逼着不能开私家车的，有钱支付费用并想支付费用的人去坐公交，是拥有多辆车或交通便利的人的特权想法，也无助于交通的缓解。

一些城市有限行，一些城市有限购，导致很多人无法开着自己的私家车，或者有车也在某些时间不能开，这些人是打专车的主力。如果从政府制定政策的角度看，这些人就不应该使用车辆，而让这些人打专车出行就是导致了交通拥堵，也让这些城市的限行限购政策形同虚设。不过，反过来看，这些限行限购政策只应该是手段，而不是目的，只要道路通行效率高了，更多人会放弃无必要的购车与开车，即便使用了一些专车，也提高了车辆使用率，减少了道路上的车辆数量，两项抵消，专车对道路通行的正负贡献归零。

另外一些人，有临时的必需的出行需求，自己又没有开车，或者自己没有车却必须出行，那么，因为出租车都去躲高峰期而不出车，这些人在高峰期根本打不到车，只能使用公交，现在有了互联网约车，这些人就等于“开车”上路了。这些人的打车要求是完全合理的，只是因为出租车提供不了服务才去使用专车，没有抢夺本不属于自己的路权。

然后，专车上路之后在初期会有一段时间的独自驾驶，这个时间

的单位占用道路面积比较大，也确实从理论上增加了拥堵，但这段路程应该不长，否则这个专车司机也不合适，当接到客人之后，路权是来自乘客的，单车的占路面积没有变化，但即便去除司机的权力，乘客也仅仅占用了一个车位的面积，和行驶在路上的那些多数上班族的私家车占用的单位面积是一样的。乘客到站之后，专车如果空驶，也会增加道路拥堵，除非快速地接到乘客。

从这个角度上看，如果是一些上班的人兼职做专车或者顺路拼车，将大大降低道路占用面积，提高车辆的使用效率，而那些南辕北辙的专职专车确实会导致拥堵。所以，想要最大程度的减少专车带来的交通拥堵，就需要提高专车的使用效率，减少空驶率，让更多的人使用专车，才是减少交通拥堵的良方。

最后，我们看看道路上的停留时间。专车对比传统的巡航出租车，在路上行驶的时间要短很多，空驶的距离也更少，如果专车很流行，或者让更多的出租车通过互联网约车出行，道路上的空驶出租车会减少，道路状况有可能改善。提高专车司机的道路熟悉程度，改善专车司机的服务水平，提升专车司机的开车技术水准，将吸引更多的二把刀的司机放弃自己开车出行的想法，从而减少车辆数量和提高道路上车辆的行驶速度，最终减少交通拥堵。

以上的分析并不全面，但基本上从轮廓上分析了专车对道路行驶的影响。可以这样说，专车的使用在一些方面确实减少了交通拥堵，但也在另外的方面加剧了道路拥堵，功过分明，结果是正数还是负数则需要精确的计算，也对于不同的城市及城市出行的不同时刻有着不同的影响。

总之，城市拥堵的高峰时段，那些有充分理由不选择公交出行的

人们有足够的群体数量，选择互联网专车，并且专车司机道路熟悉开车熟练，对交通拥堵的影响应该微乎其微。就和以前有专家说，北京的雾霾是由家庭做饭引起的一样，说专车导致交通拥堵有些夸大其词，即便有影响，也是次要之次要的因素，何必拿出来秀呢？

坐飞机最危险的阶段是去机场的路上

对于经常乘坐飞机出行的人，最不愿意看到的新闻就是飞行事故，但对于统计出身的人来说，又有理性的数据证明，航空是目前地球上最为安全的交通方式。航空安全现状到底如何，大数据来告诉您。

按照国际航空运输协会的统计，只要一名普通乘客乘坐的是西方飞机制造商生产的飞机，那么他遭遇航空事故的几率低于五百三十万分之一。从事故发生的几率而言，就算是飞行时间最长的飞行员用一辈子的时间进行飞行，也很难超过两万架次。航空业事故发生几率非常低——即便是一个人天天坐飞机，也要一万四千年才有可能遇上一个航空事故。

在这个时候，网络和各种媒体上充斥各种各样的消息，人们的感性会战胜理性，统计学的知识也将让位给内心的感受。

关于航空安全，通过大数据的分析，至少可以告诉我们几个往往会误认的真理：

（1）数据统计的结论毫无疑问地告诉我们，飞机是目前地球上最安全的旅行交通工具，没有之一！飞机比汽车、火车、骑自行车、步行等的安全级别高太多。

飞机重大事故发生的频率如何？

重大事故绝少发生，造成多人伤亡的事故率约为三百万分之一。航空是远程交通最安全的方式，而且它变得越来越安全。30 年前，重大事故的发生率为每飞行一亿四千万英里一次。如今是 14 亿英里才发生一起重大事故，安全性提高了十倍。

坐飞机和坐汽车，哪个更安全？

据美国全国安委会对 1993～1995 年间所发生的伤亡事故的比较研究，坐飞机比坐汽车要安全 22 倍。事实上，在美国过去的 60 年里，飞机失事所造成的死亡人数比在有代表性的 3 个月里汽车事故所造成的死亡人数还要少。

（2）对于单个人来说，飞机、火车或者汽车，安全出行的概率其实差不多。

从行驶的距离和死亡人数的关系而言，乘飞机旅行是最安全的旅行方式。但要是按照死亡人数和单次旅行时间的关系来看，火车与飞机一样安全，而乘汽车旅行的危险几率是飞机的 4 倍；如果从死亡人数和旅行次数的关系来看，汽车要比飞机安全 3 倍，火车要比飞机安全 6 倍。

但人们必须注意到一种交通工具的可能性很难准确地与另一种交通工具的可能性相比较。飞机一次就有 250 名乘客和机组人员，而一辆汽车最多运载 5 名乘客。由此看来，飞机一次运载的人数是汽车的 50 倍，但安全性却是汽车的 60 倍（以行驶的距离为衡量依据）。对于单个乘客而言，飞机的安全性并不比汽车高出多少！

（3）飞机事故造成的社会影响却比其他事故更大，原因是事故少但严重程度高，受关注度大。

（4）美国的大数据专家通过对全球航空公司的运营数据的分析，揭示出，各国的航空安全指数实际上相差无几，并不是说发达国家的飞机就更加安全，当然，那些被制裁和处在混乱状态的非正常国家除外。

（5）国外专家确实也得到了数据的结论，国际航班往往比国内航班出事故的概率要低，所有的国家都一样，并不是发达国家的国际航班就更安全。

（6）各家航空公司的安全系数有差异，位于德国的航空事故数据评估中心（JACDEC）综合全球 60 家航空公司 30 年的飞行里程及事故数据，对各家航空公司的安全性进行了评估。

根据他们的数据，芬兰航空是目前全世界最安全的航空公司，已经有 50 年没有发生严重事故。紧随其后的是新西兰航空、国泰航空和阿联酋航空。中国的海南航空排名第 8，东航、国航和南航分别位列 36、43 和 48 名。最安全的 9 家航空公司在过去 30 年中没有损失一架飞机，也没有造成任何生命损失。但是这些公司中有多家成立时间较晚，且运营时间没有达到 30 年，例如阿联酋的阿提哈德航空是 2003 年才成立。

（7）飞机上的不同座位的安全程度相差无几，事故之后的安全主要取决于事故的严重程度和幸运，并不是前舱比后舱更安全，或者中间比两边更安全。

许多乘客喜欢选择飞机前排座位，除了觉得那里进出方便外，相当一部分人还认为前排的位置更安全。一位经验丰富的乘务员告诉记者，她的工作中也常遇见同样的问题，许多乘客会因为觉得安全系数高而选择靠前排的座位，更有人觉得头等舱、商务舱、经济舱的分布

就是根据安全系数排序的。对此，这位乘务员表示，飞机上其实并无所谓最安全的位置，安全位置只是相对的。“相对于选择座位，系紧安全带，仔细阅读安全须知卡更为重要。”一位资深安全员表示。2013 年，韩亚航空的波音 777 客机在旧金山机场着陆失事，当时飞机内有乘务员被甩到飞机尾部，飞机撞地再起飞，所有人又撞向天花板。这种情况的发生，很可能是由于乘务员和乘客以为航班下降即将着陆，将安全带解开了。其实，在飞机起飞和降落时，更要系好安全带。在乘坐飞机时，起飞和着陆占总飞行时间的 6%，事故率却高达 68.3%，所以有黑色 10 分钟之说。据介绍，如果遇到高空解体的情况发生，无论坐在飞机的哪一个位置，生还的希望都很渺茫。不过，即便如此，也必须在飞行过程中系好安全带，这其中的重要性，是不言而喻的。另外，乘飞机出行，对于天气自然也应谨慎考虑。雷、雨、雪、雾天气尽量避免出行，这都是飞行的安全隐患。这里还得说，飞机延误时，为了安全起见，起落城市或者航路上有不好的天气也是不能飞行的，因而导致的航班延误，则请您多谅解。

（8）飞行过程中的安全概率是不一样的，起飞和爬升到巡航高度，下降和着陆是飞行中最容易出问题的两个阶段。用极简单化的说法，起飞时在发动机推力和结构整体性方面对飞机的要求最高，而接近和着陆则对驾驶舱的机组人员要求最高。约有四分之三的严重事故都是在这两个短暂的飞行阶段中发生的。

（9）大飞机的安全系数并不比小飞机高多少，其执行的安全标准是基本一致的，人们对小飞机的安全性质疑多来自偏见。

真的是飞机越大，安全系数越高吗？是不是小飞机、私人飞机都尽量少坐呢？事实上，不同的机型，在起飞、爬升、下降过程中，操

作程序和外界环境影响是没有差别的，唯一不同的是在巡航阶段，也就是保持稳定平飞的阶段。中纬度地区的对流层是 10000 米 ~ 12000 米，通常大型飞机的巡航高度在 10000 米以上，而小型飞机的高度在 9000 米左右，所以乘坐小型飞机的乘客会时常感觉到颠簸，而乘坐大飞机则感觉更稳当。“实际上，飞机大小所影响的主要是舒适性，而非安全性”。一位资深飞行员作了这样一个比喻：其实飞机就像我们开汽车一样，家用 A 级车与 B 级车相比较而言，B 级车更加厚重和平稳，乘坐起来的舒适度更高。当你开车过沟坎时，轻薄紧凑型的 A 级车更敏感，舒适度低一些。大飞机受力面大，遇到气流时可以分散受力，乘客乘机时感觉不会太颠簸。如此看来，大飞机和小飞机并不存在哪一个更安全的问题，而出于舒适度考虑的话，建议您选择大型飞机。

（10）欧美公司宣称自己的飞机比其他飞机更安全。西方飞机制造公司生产的喷气式和螺旋桨式飞机的数量分别占到了世界航空市场的 95%和 80%。在 2012 年的 23 起导致乘坐人员死亡的事故之中，只有三起是使用了西方飞机制造公司所生产的飞机。但这个数据也有问题，对于飞行这样的大数据，这种比例的数据之间进行对比实际上意义不大，其实，造成这种差异的原因，主要应该是与飞机飞行的区域和国家的安全管理方面更为密切。

中医治未病，大数据四法助你看透 P2P 投资风险

银行的利率持续下调，老百姓更加关注收益更高的互联网金融，特别是各种类型的 P2P。但是，面对良莠不齐的 P2P，缺乏投资经验的

老百姓很容易被忽悠，在追求高收益的同时让自己的资金处在危险的境地。

在这里，结合各方面的研究成果和实际的操作经验，达睿咨询总结了一些简单的方法，帮助投资者更好的选择 P2P 平台，远离风险，保证资金安全。

望：横看成岭侧成峰，远近高低各不同

一望官网是否专业，粗制滥造的官网肯定不会有好的用心，但美轮美奂的官网也往往更让人不放心，那些专业度高、实实在在的 P2P 网站更可靠。看平台的网站，用户体验如果很差的话，基本就可以判定是某老板从某宝上买的某模板，眼不见为净。一家平台的真实性不是去看他网站的美观程度，更注重的应该看下服务器的地址及备案信息。如果服务器是在国外并且没有备案，那风险程度就要加分了。

二望内容是否全面，比如查看 P2P 平台经营团队的职业背景，如果投资者找不到这些资料，少投为妙。

三望团队历史背景，若这家 P2P 平台经营方此前就有杠杆融资炒股，或参与民间借贷的职业经历，投资者就要留心这个平台很有可能存在自融业务参与炒股或民间放贷。投资人不仅要关注运营团队的金融从业背景和网站运营背景，同时特别要关注股东背景，股东需要有真实的、多年的金融从业背景才符合正常的逻辑。一群非金融行业的股东是不太可能把 P2P 平台搞好的。

四望股东自融可能，如果观察到这家 P2P 机构的法人代表与股东出现在 P2P 平台借款人的名单里，风险会很大。

五望公司股东结构，多股东的比少股东的好，同时公司的注册资

本和实收资本尽量要在千万元以上，平台需要具备基本的抗风险能力，公司的实际办公地址在 CBD 的要比非 CBD 的好。

六望公司真实地址，如果条件允许，而且你又是大额投入，可以在得到公司地址的情况下，最好去实地考察，不方便的话就拜托所在城市的朋友去考察验证下是不是真实的公司。查一下注册法人和风控等主管的个人信息，留意下平台标的种类是信用标还是抵押标，平台的相关费率，债权可否转让，转让费用如何等。

闻：夜来风雨声，花落知多少

一闻公司品牌的味道，一般的互联网金融公司都会取一个相对时尚且具有吸引力的名字，但要是 P2P 公司的名称太多另类，或者是典型的山寨，都是很危险的，最危险的是那些带着中字头或知名企业名头的野鸡公司。

二闻产品的勾引味道，越是大额的、预期利率越高的 P2P 产品，越有可能是 P2P 机构为了募资发出的假 P2P 产品。

三闻资金的流向，关注 P2P 平台的投资者资金流向，如果大量投资者账户资金最终都流向同一个借款人账户，那么存在 P2P 平台自融的可能性就很大。一个 P2P 平台要进行自融，其虚构的 P2P 产品的相似度有时会非常高，感觉就是一个模板做出来的。往往仅在借款额、资金用途、借款人资料方面稍作修改。

四闻处理逾期能力，若一家 P2P 机构在出现 P2P 贷款逾期后，能用最短时间追回尽可能多的逾期贷款，就表示他们的坏账处理能力相当高，相应的跑路风险也会下降。

五闻网络媒体风声，如果网络上开始出现一些有关此 P2P 的负面

消息，就需要密切关注动态，在投资前也应该通过网络查看平台的口碑情况和此前业务的收益及风险情况。

问：问君何能尔，心远地自偏

一问资金的托管，关注这家 P2P 机构是否找第三方支付机构或银行进行资金托管。

二问是否有风险备付金，“风险保证金”是 P2P 企业以利润或者募集资金计提，在项目出现逾期或者风险的时候启用，用来先行垫付给投资者的一笔资金。但这其中又分为两种情况，专业银行人士告诉我们：“如果企业自己成立所谓的‘风险保证金’对于银行来说只是普通的存款，银行不会对这笔钱监管，网贷平台随时可以取走这笔钱；如果是 P2P 企业与银行合作建立的风险备付金，那么双方必须签约，约定缴存比例及启用条件，而每一次启用资金也必须是符合条件下经银行审批，且该笔资金网贷平台无法随意支配，更不要说直接转走。只有这种情况下，银行才起到监管的作用。”

三问客服是否在线，如果客服电话经常打不通，或者一段时间都打不通，问题就来了。

切：上寻鱼际，下寻尺泽，以求其终始

一切用户变化状态，新老客户的数量中往往代表平台的稳定性，假如新老投资者的数量维持在不高的幅度，代表平台最近发展相对平稳，假如突然出现较大的降幅，说明平台已经开始出现问题。新投资人和老投资人人数不变或者都有增加，说明近期该平台稳定、正常。如果都有较大的减少，说明平台可能出现问题。

二切产品购买热度，一个 P2P 平台的满标时间，说明了其投资者

的数量和对其的信任，如果一个 P2P 平台的满标时间不断地增长，说明投资者对其产生了疑虑，需谨慎投资。

三切产品周期，一般来说，借款周期短的投资风险指数要比借款周期长的小，但如果平台多是借款时间较长或全是借款时间较短的产品，都应选择谨慎。

四切平台资金流动，对借出金额与回收金额相比较，也就是一周内所有借款人成功借款的总额与一周内所有借款人需要还给投资人的还款总额相对比，假如回收金额为负，且持续较长时间，或许该平台已经悄悄出现了资金短缺的问题，甚至到后面会发展为提现困难等问题，风险系数为高。

五切平台现金流，如果现金流为负，并且此现象持续较长时间，那么平台近期可能还款压力较大，容易出现提现困难，甚至是资金链断裂，投资风险系数高。

六切产品集中度，如果前 10 名借款人占比 30%以上，风险系数很大。如果出现一个借款人逾期坏账，那么对平台的冲击是巨大的，严重的会造成资金链断裂。

以上只是一些简单的方法，可以通过各个侧面看到 P2P 平台的危险所在。如果要想控制风险，关键还是看自己，还是那句话，P2P 有风险，投资需谨慎。

你会叫个外卖给丈母娘拜年吗

2016 年正月初五，这个猴年春节也算过去了大半，春节假期也即

将结束。看着还在不断穿梭于亲戚家的拜年族，看着各个饭店兴旺的全家聚，在传统面前就会感受到互联网的渺小。

每年的春节都会引发人们对拜年方式的思考和讨论，而这些年老百姓的拜年方式也确实是在变化，就在很多人认为微信甚至红包拜年取代了短信的时候，我们真的应该回头看看，拜年的方式真的变了吗？

从前，拜年的方式只有一种，面对面的拜访，后来，人们通过书信、贺卡等方式和远方的亲人互致问候，直到现代化的通信工具出现，电报、电话等拜年都曾经红极一时，后来，手机流行，短信拜年成就了运营商黄金岁月，随后，互联网走进寻常百姓的生活，又出现了电子邮件拜年、飞信拜年、微信拜年，现在又有了视频拜年等。

不过，即便岁月变迁，拜年的手段顺应通信技术的发展而日新月异，可这些手段都是逞一时之勇。此前的短信拜年从盛到衰，只是历史发展的一个自然过程，微信拜年的流行也一样。

其实，都说是拜年，却存在亲疏远近的差异。一般来说，拜年包括给家里的长辈拜年、走亲访友、礼节性拜访、人情世故的拜年、串门拜访。

几十年前，甚至到了现在，春节给家里的长辈拜年，都是要磕头作揖的，长辈受拜以后，要将事先准备好的“压岁钱”分给晚辈。接下来，便是走亲访友，须带礼物，可以逗留吃饭、玩耍、座谈等。然后会有礼节性的拜访，如给朋友拜年，一般不久坐，寒暄两句客套话就要告辞，主人受拜后，往往择日回拜。还有一种是感谢性的拜访，凡一年来对人家欠情的，就要买些礼物送去，借拜年之机表示谢意。最后是串门式的拜访，对于左邻右舍的街坊，素日没有多大来往，但见面都能说得来，到了年禧，只是到院里，见面彼此一抱拳说：“恭禧

发财”、“一顺百顺”，在屋里坐一会儿而已，无甚过多礼节。

实际上，虽然短信拜年、微信拜年流行，也只是串门式拜访的一种形式变换，对于给家里长辈拜年、走亲访友、甚至也包括感谢性拜访，都没有冲击，更没有替代。

我们很多人仅仅从数据分析的角度出发，看到了短信群发量的暴增暴跌，看到了微信朋友圈的快速更新和各种红包的抢来抢去，就得出了拜访方式变更的结果，属于典型的“数据说假话”。量的变动与质的变动虽然往往有关系，量变达到一定程度往往会引发质变，可量的巨大变动并不一定代表质变。

手机通信和互联网的发展大大拓展了社会和商业的边界，也让人们拥有了更多更广泛的交往范围，也就是所谓的“朋友”越来越多，可这些朋友多数都仅仅是“朋友”，或者是某种意义上的“好朋友”，由于时空的距离和感情的网络化，主要通过现代化的群体通信方式来互致问候。可以这样说，微信拜年流行，是因为朋友太多而好朋友太少。

正是在社会交往的一般性的朋友的增加和社交空间的扩大化的基础上，这部分的拜年数量急速膨胀，导致了数据的倾斜，由此看起来是拜年方式发生了变化。实际上，给家里长辈的拜年、走亲访友等多年来极少甚至丝毫未发生变化，只是因为这些拜年需求并未大幅增加（甚至因为人口结构的原因而在减少），所以，让我们更多地看到了拜年方式变了。

中华民族几千年来的历史都表明，选择什么拜年方式，体现的是关系的亲疏远近。即便全球有 4.2 亿人在发微信红包，也不代表人们过年的拜年方式发生了任何变化，拜年的形式在不断创新和演化，可形

式虽然在变，但内核始终未变。

同事之间，发个微信红包，可以算成是拜年。合作伙伴之间，发个微信祝福，可以看作是拜年。远方的亲属或者师长，发个短信拜年也未尝不可。但是，给自己的爷爷奶奶爸爸妈妈，至少也要打个电话，让对方听到真实的声音吧！只要有可能，给家里的长辈、亲戚，只要有可能还是要登门拜年，你总不能自己窝在家里看电视，给同城的丈母娘叫一个饭店外卖送过去就当拜年吧！

第 6 章

善用数据，但别自作聪明

收集情报和信息的几种方法

> 明君贤将，所以动而胜人，成功出于众者，先知也。先知者，不可取于鬼神，不可象于事，不可验于度，必取于人知敌之情者也。
>
> ——《孙子兵法·用间篇》

信息和情报的收集是数据分析的基本功，也是要慢工出细活，更需要长久的积累，临时抱佛脚的方式是做分析工作的大忌。

平时多流汗，战时少流血。做数据分析也是如此。如果我们在生活和工作的时时刻刻都留心收集整理资料，等到需要用的时候，信手拈来，工作效率的提升可不只是一两倍。

好记性不如烂笔头，那是在以前，现在，有了计算机和手机，我们的存储和整理更容易更便捷。平时，在上网的时候看到什么有价值的信息，就设定清晰的目录或做成 PPT 等文件进行保存，也许不用什么复杂的文本美化，而且是越原始越好，分门别类地存放起来，未来一定有大用。

当然，这些存储的资料，一定要有原始的时间刻度，最好把出处也标记出来，以便未来引用的时候方便。不过，以现在的网络搜索水平，只要你记住了关键词和内容，实在不行，再上网搜索将内容补全也是可以的。

举个例子：国家广电总局陆续向七家企业颁发了互联网电视牌照，这张牌照非常有价值，因为这是广电企业向互联网视频进军的通行证，更是互联网视频企业必须与之合作才能开展相关业务的护身符。有了这张牌照，可以坐地收钱。但是这些牌照不是一天颁发的，也不是每

家媒体都会有相关的报道，这就需要用心的收集，发一张牌照就把相关的资料稍微加以整理，时间长了，收集齐全，价值巨大。

2010 年 3 月 24 日，国家广电总局向中央电视台旗下的 CNTV（中国网络电视台）颁出了第一张互联网电视牌照，并由未来电视具体运营。未来电视有三大股东：央视国际控股 60.2%；腾讯持股 19.9%；中数寰宇科技公司也持股 19.9%。

2010 年 3 月，以浙江电视台和杭州广播电视台为申请主体，华数传媒也获得了互联网电视牌照。2015 年 5 月 11 日，完成增发后，华数传媒的股权变为：华数集团占股 41.85%；云溪投资（由马云和史玉柱共同出资）占股 20%；浙江二轻集团占股 6.91%；湖南千禧龙投资占股 6.84%；上海源仓投资占股 4.97%，其他后五位股东分别为东方星空、浙江发展、北京光华贰陆柒企业管理公司、上海景贤投资、深圳孚威创业。

2010 年 7 月，上海广播电视台亦获得互联网电视牌照，由 SMG 旗下的新媒体公司百视通（600637.SH）运营。具体负责的上海视云网络科技有限公司是一家合资企业，由百视通控股 51%，联想持股 49%。

中国国际广播电台 2010 年 12 月取得互联网电视牌照，由“国广东方”运营。“国广东方”成立于 2006 年 11 月，股东包括华闻传媒（000793.SZ）、国广环球传媒公司。2014 年 7 月，优酷土豆的母公司合一信息技术（北京）有限公司向国广东方增资 5000 万元，持有国广东方 16.6667%的股权；同时，华闻传媒在国广东方的持股比例也调整为 36.8171%。

2011 年 3 月，以广东电视台为牌照的申请主体，南方传媒互

联电视“互联八方”集成服务平台、“云视听”节目内容服务平台通过国家广电总局验收，取得运营许可。这张牌照 2015 年 4 月 14 日之前，由南方传媒与优朋普乐合资的“广东南广影视互动技术有限公司”运营。整合后的广东广播电视台把牌照交给南方新媒体公司来运营。

中央人民广播电视台 2011 年 3 月也获得了互联网电视牌照，并交由旗下的银河互联网电视有限公司具体运营。“银河”成立于 2012 年 7 月，是由央广新媒体公司、江苏电视台、爱奇艺（百度旗下）合资。去年 9 月鹏博士电信传媒集团增资入股。

2011 年 7 月，湖南电视台正式获得互联网电视牌照，委托“快乐阳光”开展业务。2014 年 8 月 11 日之前，“快乐阳光”的股东为湖南广播电视台。此后，则变更为芒果传媒有限公司（业内称“芒果 TV”）。2015 年，芒果 TV 上市的传闻甚嚣尘上。

很多人的电脑里也存储了大量的资料，各种数据报表，各种 PTT 汇报材料，各种各样的文本，可是，书到用时方恨少，总是觉得找不到，或者收集的一些素材用不上，根本原因还是当初收集的时候没有用心，只是为了收集而收集，更没有使用一定的方法进行分析，所以信息的价值远远没有被开发出来。可以说，数据是分析的基础，分析也是收集数据的一种方式和能力。

如果我问你，是更相信电视台里的新闻内容，还是更相信网络上的新闻内容，很多人也许会不假思索地脱口而出是网络。可实际上，如果我们假设电视台与网络上传播的内容都是真实的，那么对我们的分析而言，信息也并不完全“真实”。

我们每个人都有视野障碍，往往只会看到自己关注的事物。比如，

如果你开了一辆宝马汽车，平时行使在路上，一定会经常看到宝马车的身影，甚至会奇怪，为何宝马汽车如此之多。同样，如果你是未婚的小青年，平日回家喜欢宅在家里，会感叹小区里冷冷清清缺乏人气，当你娶妻之后，特别是妻子怀孕，要陪着在小区里散散步，你这时候会发现小区里原来有这么多的孕妇，再后来，小孩子出生，你要推着婴儿车去晒晒太阳补补钙，你就会发现，小区里到处都是小孩子。实际上，小区没有任何变化，孕妇一直差不多，小孩也天天都在外边跑，变化的是你的关注点。

收集数据信息，随手可得，哪怕是被人总结为“三段论”的新闻联播，比如本书前面讲过的柳传志看报纸而萌发创业想法的案例。还有，自 2015 年 4 月 17 日起，招商证券金融工程分析师夏潇阳推出了一项新的股市“预测仪”，情绪指标，通过该指标捕捉沪深 300 指数异动来分析市场。自夏潇阳推出这项“神秘指标”起，至 2015 年 5 月 4 日为止，夏潇阳均维持“看多”，与市场走势一致。而 5 月 4 日晚，夏潇阳发出指标翻空预警，神奇的是，5 月 5 日，沪深主板市场迎来 3 个月以来的最大调整，沪深 300 指数重挫 3.99%；上证综指大跌 4.06%；深证成指跌 4.22%。

看电视可以收集素材，而走在大街上一样可以依靠细心观察获得新知。比如，电信运营商的门头是有大学问的，你天天路过，但如果能把他们放到一起串联起来进行对比，就会发现大不同。

观察分析的经典便是“油条老胡早知道”，有兴趣的可以去翻翻。当年，古巴领导人卡斯特罗在会见教皇保罗二世时，抱怨自己的关节炎，被别人偷听到了，这就成了的重要情报。有了互联网，我们还可以通过很多公开发布的照片或者声音来做出分析，从而完善自己的数

据。英国就有人通过中国国家领导人夫人的照片计算出衣服的尺寸，从而设计了一个大披肩作为国礼。在大庆油田发现之后，媒体刊登了相关的新闻照片，虽然没有提到地址信息，但日本情报机关根据照片上的人物着装、土壤情况、做饭等细节，几乎准确地确定了大庆油田的地理坐标，当然，令日本人非常懊恼的是，当年日本占领东北之后也曾经到处探油，而只是止步于大庆油田咫尺之遥的地方。

没有调查就没有发言权。是的，第一手的信息往往是最好的，这也是为何现代遥控技术如此先进，人类还是想亲自登上火星去看看的原因。也有一些基金经理会对自己关注的要准备投资的企业进行全方位考察，甚至会每天蹲守在企业大门附近，一边计算进出厂区的车辆，一边分析公司的销售量和收入等指标。

有些时候，我们无法接近要分析的事物，比如竞争的信息资料，这些信息对于分析来说却十分重要。这时候，我们可以根据一些外在的相关表现来得到近似的数据信息。

如果你是研究竞争对手情况，那就可以从其代理商、合作伙伴或者对手的用户那里得到消息，如果你是研究国家和社会，那也可以从公开的报道中找到蛛丝马迹。

我们都知道“克强指数”，主要是看发电量、铁路运输量和银行的贷款，这些数据往往都是权威部门发布之后才可以分析。但是，如果你经常看新闻，关注国家大事，完全可以从其他方面得到印证。

举一个例子：中国铁路总公司曾经推出过很“亲民”的满足市场需求的业务，把一些东部沿海城市的私家车放到火车上在黄金周假期的时候给运输到千里之遥的西部旅游地区。这样的新闻，你可以理解成是铁路公司的“市场化转型”，也同时可以看到，中国经济活跃程度

是多么低。要知道，铁路是中国经济的大动脉，在中国经济发展火热的时候，铁路车皮很难搞到，即便是求人托关系都需要排队，可铁路却有了精力帮助驴友们出去玩，可见运力的闲置状况。

收集信息和情报是如此重要，不允许有任何的差错，否则后果将非常严重。在红军长征过程中，遵义会议之后的土城之战，就因为情报错误，敌人的 4 个团变成了战斗力很强的 9 个团，差点让红军遭受灭顶之灾。

球探与中国足球的屡战屡败

知己知彼，百战不殆。即使在字字千金的《孙子兵法》中都有专门的章节来论述情报工作的重要。在现代信息社会中，情报工作更是渗透到市场竞争和社会生活的方方面面。

足球是很重视情报的。足球赛场上的情报战，甚至已经成为决定比赛胜负的一个关键因素。2004 年奥运会预选赛上，中国队再一次失利，其中一个因素就是韩国队对我们的队伍了解得非常详细，甚至对每个队员都建有详细的个人档案，而我们对对手的了解却明显不如对手对我们的了解。

有些情报可以在公开渠道很方便地获取，比如球队的网站、宣传品、教练和球员公开访谈，但是这些信息都非常有限，球队往往有很多的保密措施让我们只能得到只言片语。为了获得更多信息和真实的情报，就产生了一个特殊的职业：足球间谍，也可以称为球探。这些足球间谍在俱乐部老板或球队上层的支持下不择手段获取情报。他们

在搜集球星大腕资料的同时，又盯住了教练手中的布阵演兵图，甚至每一个细节。他们中的大多数人都可以获得丰厚报酬。

据说第一个职业球探是克莱德·迪肯，1949 年秋，克莱德·迪肯结束了在北非 10 年的英国职业间谍生涯，回伦敦做记者，他对前苏联、以色列、日本的情报工作都有过深入研究，是一名优秀的情报专家。作为球迷，迪肯开始专门从事体育报道。迪肯的主要活动是监视将在比赛中与球队相遇的客队。他亲临现场观察客队比赛，关注观众的评论和队员的竞技状态，另外在不暴露身份的情况下起草详尽的报告。每当赛前一两天，球队都会得到一份对手的材料，该队在足协杯赛中曾战胜多支劲旅。后来，他建立了一套培养球探的方法，同时，也为一些专业足球俱乐部输送了不少高水平的球探。

其实，足球领域的情报获取和商业竞争与社会生活中的情报获取都是类似的，就是指关于竞争环境、竞争对手、竞争态势和竞争策略的信息和研究。它既是一种过程（对竞争信息的收集和分析过程），也是一种产品（包括由此形成的情报或策略），一般简称为 CI（Competitive Intelligence）。情报的最大作用是帮助人们做出决策，减少在决策过程中对事物认识或判断的不确定性和信息的不对称性，是管理决策的基础。

获取情报的基础是合法性，情报尽管会涉及秘密情报这一内涵，但这并不意味着情报的获得要通过不正当的途径和方式。实际上，竞争情报在大多情况下是要通过正当的途径获得的，不能采用窃取、利诱、贿赂、胁迫等手段，非法获得情报。

美国中央情报局前身——战略服务办公室在二战期间发明的“四步情报周期”法。即：首先，提出问题并制定工作计划；第二，搜集

情报；第三，分析情报；最后，把搜集到的情报送给决策者。这一操作流程同样适用于足球或者商业情报的搜集、分析与运用。

首先要提出问题并制定相应的工作计划。如果是为了世界杯比赛的需要，那首先要确定在世界杯小组赛和可能的淘汰赛中遇到的是哪些对手，具体的比赛时间是怎样的，我们用来搜集情报的周期怎样确定，并做出详细的工作计划。

然后是搜集情报，确定采取怎样的方式通过什么途径来搜集。我们是派人去对手球队的基地考察还是雇佣内线提供信息？我们是在网上搜索还是到图书馆查阅资料？我们是到赛场观看有对手参加的比赛还是观看比赛的实况录像？我们通过对二手资料的整理和实地的调查可以得到自己希望得到的有用信息。科龙集团配备二十多位由博士、硕士等组成的团队每天从香港、日本等亚洲发达地区及北美、欧洲等重点市场的办事处获取信息，深入研究海外的个性化市场，为产品适销对路而出谋划策。曾经仅在一个月内，就提供颇具价值的建议信息 20 多条，做成空调、冰箱、冷柜等出口业务 12 宗，总额 2000 多万美元。平均 1 条信息，就产生 100 多万美元的价值。

第三步，分析情报。这一环节也是整个情报工作中最核心的部分。情报搜集回来后，关键是要对情报进行专业的分析。特别是从互联网上采集来的信息，很多都是未经过证实的。情报里面存在很多或夸大或缩小的成分。特别需要指出的是，来自竞争对手的情报很多都是不准确的，有利于自己的数据可能会被夸大，相反则会被缩小。1997 年戚务生带领的国家队虽然也注意了收集对手的情报，但是收集到的情报却是众所周知且没有价值，于是在金州与伊朗的比赛中，中国队注意了戴伊、巴盖里、阿齐兹，没有给他们在比赛中兴风作浪的机会，

但是却忽略了后来在德甲扬威的马达维奇亚，结果伊朗的“小马达”两记超远距离的射门令门将区楚良猝不及防，酿成大祸。

因此，我们收集和分析情报一定要客观、真实，虽然不要轻易放过任何一个信息，但也不要轻易地采信一个信息，以防被对手的假信息所惑。必须去粗取精、去伪存真，获取准确、有价值的情报，要注意从多个角度来获取信息进行分析和求证。如果是一个市场竞争对手的信息，我们可以从媒体报道、市场传闻、顾客反馈、卖场变化等方面综合的评价信息的真实性。

随着市场竞争的日趋激烈，经济间谍活动将更活跃。他们将像空气和水一样渗透到各个行业、各个部门、各个他们认为有油水可捞的地方，不择手段地窃取竞争对手的经济情报和商业秘密。因此，我们在重视获取对方情报的同时，必须非常注重自己的信息安全，建立更为严格有效的保密制度。

无论是获取情报还是分析情报，无论是球队比赛还是市场竞争，都有一个很重要的要素，就是要快，要领先对手一步。哪怕仅仅领先对手半步，可能历史的进程就会完全不同。

网络资料的鉴别与识别谣言

互联网上有海量数据，其中有很多优质的内容，当然也不可避免地存在糟粕，很多虚假信息或者不全面的内容会严重误导使用者。

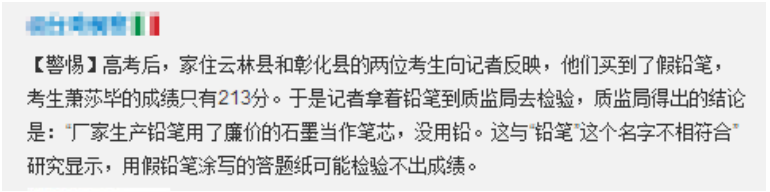
网络上的信息造假有三个特点：门槛低、传播快、影响大。所谓“门槛低”，就是网络上的任何一位网民，他都有相对比较丰富的手段，

提供一个假信息、假照片等。“传播快”是网络的自身特点，这样的假信息一旦进入到网络世界的平台上，能够迅速地传播，其传播速度，是传统媒体平台不可想象的。往往在人们还完全不能判断它的真假的时候，就已经获得了最大的传播效果。网络造假的第三个特点就是“影响大”：网络上的任何一点虚假信息都有可能造成比传统传播平台更大的负面影响。

如何鉴别网上的资料真伪呢？专家们给出的回答是，看有无客观物证，看是否直接来源，看提供者的利害关系，多方比较，交群众检验，交实践检验。

我们看一个例子。同样的一张图，如果剪辑的角度不同，给人的感觉会大不一样，所以，有时候，眼见也不一定是真的。

（1）凭空杜撰型的谣言，这个最好理解，多数谣言都属于这一类，没有任何事实依据的编造杜撰，不管其真实性是否被验证，因为是造谣者编写的，都是谣言。


【警惕】高考后，家住云林县和彰化县的两位考生向记者反映，他们买到了假铅笔，考生萧莎毕的成绩只有213分。于是记者拿着铅笔到质监局去检验，质监局得出的结论是："厂家生产铅笔用了廉价的石墨当作笔芯，没用铅。这与'铅笔'这个名字不相符合"研究显示，用假铅笔涂写的答题纸可能检验不出成绩。

（2）夸大其词性的谣言，这种谣言往往有基本的事实，但对事实进行了夸大其词和扩大化，比如本来受伤10人却在传播中被说成是100人，这种谣言迷惑性比较大，容易让人被基本事实蒙住眼睛和判断，比如经常被传播的三年自然灾害饿死数千万人，根本不符合人口学的基本常识却被很多人盲目相信。

（3）断章取义性的谣言，这种谣言是从某个大的内容中摘取的，通读整篇内容才可以理解其真实的含义，但如果被人从中间拉出一小段进行传播并不加以解释，就会造成完全不同甚至相反的理解，这种谣言只要看原来的整体就可以识别出来。

（4）拼凑剪接性或PS出来的谣言，这种谣言的基本组件都是真的，但这些组件是有其背景和条件的，脱离了具体的历史背景和场景被使用，就成为了谣言，比如一些领袖人物在特定的场合与特定的人开玩笑，或者是用一些典故与当时人人都可以理解的语言，删掉了当时的情景去传播，就成了谣言。

（5）半真半假性的谣言，这种谣言有真的成分也有假的内容，往往真的东西里面被掺入假的因素，真的东西是真实存在的，但假的却是被编写者杜撰添加上去的，比如说某个抗战老兵的历史遭遇的时候，有记者就加入了很多自己臆想中的情节用来感染人。

（6）假戏真做性的谣言，网络传播中有很多被人为设计的情节与场面，比如撑伞、喂饭等，这些都是推手们设计的舞台剧，但这种剧

作却被拿来当成偶发的社会真实去传播。

（7）刻意暗示性的谣言，传播谣言的人并没有直接针对某个事物进行编造，但所有的谣言内容却会给人以最直接的形象暗示，这会让人产生明显的联想和攻击效果。

（8）辟谣求证性的谣言，这也是大 V 们最喜欢用的传播谣言的方式，先以自己的小号或者马仔的号码发出，然后用紫禁城的大号进行求证式传播，既达到了传播谣言的目的也避免了自己的引火烧身。

（9）逻辑诡辩性的谣言，看似非常有道理的逻辑分析，其实是充满了狡辩，或者偷换概念，或者弄错前提，总之，这种是高级公知最擅长的谣言，最具有迷惑性，但其实是知识分子在耍流氓，只要用真正的逻辑来进行判断，谣言很容易拆穿。

（10）记忆偏差性的谣言，有些谣言的产生不是故意的，但却被有些人发布的时候出现了差错，后来又在传播中被误读，这种谣言本质恶性不大但有时候也会造成严重的社会影响。

此外，还有一些专家推荐的识别谣言的方法。一是可以看排版和语气。如果一篇文章是出自专业人士的手笔，语气会带着习以为常的平淡感。如果一篇文章里面充斥着各种叹号、语气词、形容词，那这个内容多半是传言。

文章的排版是和语气类似的另外一个标准，专业人士的文章也不会有各种新奇特的排版修饰。一篇花里胡哨的文章，发布者的专业性就是非常值得存疑的。

最后牢记一点：所有吓唬你的文章，默认都是谣言。结论越是绝对，越吓人，越可能是谣言。这个世界上很难突然从石头里面跳出来一条爆炸性的吓人新闻。

根据《中国青年报》的一项调查显示，50.5%的受访者坦言看到让人震惊、吸人眼球的消息时，第一反应是想办法确认消息的可信度。53.2%的受访者认为不动脑子就转发刷屏是幼稚的表现，但是，仅 22.9%的受访者看到社交平台上亲友熟人发的不实消息时会主动去辟谣。看来，谣言确实是止于智者。

网上的这些分析都是忽悠，你中招过吗

数据分析是一个很严肃的事情，但因为种种原因，很多数据分析者总是会犯下各种各样的错误，特别是在如今信息“快餐”时代，“萝卜快了不洗泥”，就造成了更多的分析错误。

在任何的数据分析中，都不能违反基本的逻辑，强词夺理或者巧言令色都会在基本的逻辑面前现出原形。下面，我们就举几个简单的例子，看看这些社会上流行的说法到底错在哪里？

1．虽然这样做是不对的，但是其带来的好处是……，所以这样做是对的。

网络上，经常会有人说，虽然……但是……，比如，在某个人或者某个组织做错了一件事情，甚至是恶意炒作造成了伤害之后，就会有人站出来说，虽然他这样是不对的，但是正因为他这样做，才带来了这样那样的好处，所以，他这样做是对的。在郭美美一案中，也在夏俊峰一案中，都有持有这样观点的人，甚至这些观点还大行其道。

下面这一段来自网络评论，一看就是典型的水军操作，在某篇文章分析某 B2C 购物网站也存在假货问题时，这位水军网友评价道：

真假暂且不说，完全可以尽量规避的。×××的隔日到真心的给力啊。说良心话，相对而言，×××售后是非常好的。请找出比×××售后还好的商家，你找得出来吗？国内相对好的，网商只有××××，实体也就是××。当然问题是有的，特别是第三方的买卖，因此有些商品我是必定只要××××自营的。

如此，这个问题变成了，虽然这家网站也有假货，但是他送货快售后好啊，所以这家网站没错误，这样的逻辑真是让人啼笑皆非。功过不能相抵，优劣更不能互充，对就是对，错就是错，即便这样做事是有意还是无意地造成了间接的进步，也不能说这种行为是正确的。结果的正义并不能表明过程的正义，何况结果也不一定正义。

2．因为 A 很流氓，虽然 B 这样的做法也是流氓，但比 A 还是要流氓少一点，所以 B 这样的做法是好的。

比比谁更流氓，这理论在互联网圈里非常流行，只要比同行或者同类流氓少一点，就可以认为自己不是流氓。实际上，五十步笑百步，不管是流氓多一点，还是流氓少一点，终归都是流氓。

比如，某手机企业或软件被对手揭发出来偷偷在用户的手机上安装用户不知情的或者根本不需要的软件，耗费了用户的流量，也增加了隐私信息泄露的风险，但该企业说，这样的做法太多了，我们做的是相对少的，其他家比我们更流氓。这样的结论也被各路媒体引用并认可，真是人心不古。

3．A 占市场份额 70%，B 占市场份额为 1%，但 B 最近增长了 50%，可 A 只增长了 30%，所以 B 对 A 构成了强烈冲击。

市场很大，但能够引起社会关注的市场很小，所以大家就容易将

目光聚焦到受人瞩目的领域或企业身上。与此同时，这些被聚焦企业的市场又不会短期内发生太多的变化，于是，为了分析的需要，很多人对数据采取了过于“敏感”的态度，一点点细微的变化也会给放大或者无限的放大来看。

如同上面的例子，A 企业的市场份额达到 70%，B 只是占到市场的 1%，可以说，A 占据了绝对的市场优势，而 B 可能仅仅是一家初创企业，因为基数和体量的差异，A 增长 30%已经算是高速之高速了，B 因为基数小，寻找市场缝隙的机会也多，增长 50%其实一点都不值得夸耀。但是，在很多分析中，有些人就得出 B 的发展态势比 A 好的结论，而且还会说成是对 A 造成了巨大的市场冲击。

我们假设整个市场原有的容量是 10000，A 就是 7000，增长了 30%，绝对值就是 2100，加到一起已经达到了 9100，而 B 占 1%，就是 100，即便增长了 50%，也只有 50，现在是 150，B 与 A 的绝对值差距从原来的 6900，已经扩大到了 8950，差距拉大了 2050，谈什么 B 冲击 A 呢？

4．A 的市场份额是 60%，B 的市场份额是 30%，C 的市场份额是 9%，D 的市场份额是 0.5%，其他公司占 0.5%，这个市场已经是四强鼎力。

电信市场和互联网市场有着天然的垄断特性，往往会形成巨头强势主导的格局，比如很多人讲的 7–2–1 分配。我们假设，在一个细分领域中，A 的市场份额是 60%，B 的市场份额是 30%，C 的市场份额是 9%，D 的市场份额是 0.5%，其他公司占 0.5%。这个时候，有些分析师就会说这个市场已经是四强鼎力的格局。

如果这样的分析不是 D 公司自己做的，也一定是 D 的朋友们做的，

从市场上看，整个市场可以看成是 AB 两强垄断才合适，至少也是 ABC 三强的天下，至于 D 只是第二集团的领军企业，要想进入鼎足而立，还需要付出巨大的努力。

5．A 是坏的，所以，A 这次拾金不昧，也是很恶心的举动。

这样的分析貌似符合逻辑学，但却真的是反逻辑，或者说是不尊重事实和哲学辩证法的。在这个世界上，好坏都是相对的，也几乎不存在完全的好和完全的坏，好人也会做坏事，坏人更是会有做好事的时候，哪怕是一瞬间的良心发现。

A 虽然是坏的，或者被人深恶痛绝，或者做过的业务都是不成功的，但却不能认为其所有的行为都是坏的，更不能认为其以后做的任何业务也都不会成功。比如，某运营商常年来被批判，话费高信号差，还经常强制用户使用付费应用，但这家运营商适应社会形势变化大幅度地对语音和流量进行降价，这怎么就不是好事了？如果非要分析出降价多么不对，就是存心找碴了。

由此推论，在很多人做分析的时候，总是依赖这样的自身逻辑，某公司很牛，所以某公司的产品都很牛，某人很牛，所以某人做的事情都很牛，也是不正确的。

6．一家公司有 A 和 B 两个产品，A 有用户 6 亿人，B 有用户 8 亿人，那么这家公司就有用户 14 亿人。

这家公司真的很牛，两个产品都有如此多的用户，所以很多人就把 AB 用户群相加得到了该公司的用户群数量。从数学上看，计算能力不错，但总觉得有点问题。

到底什么是用户呢？如果使用产品的叫用户，那么 A 的用户就应

该是指使用了 A 产品的人数，B 的用户就是使用了 B 产品的人数，而该公司的用户就应该是指使用了该公司产品的人数。如果有用户同时使用了 A 和 B 两种产品，这应该算一个用户还是两个用户呢？

由此延伸出来，很多网站和互联网应用都喜欢用“活跃用户”这个概念，但活跃用户的准确定义可能在不同的公司不同的领域不同的产品是不一样的。

简单从字面理解，“活跃”应该对应的是“不活跃”，到底什么算活跃，什么是不活跃呢？如果我一个月仅仅登录了一次，那算是活跃还是不活跃呢？恐怕，现在很多网站应用都是不使用活跃对比了，属于语义模糊故意造成认识偏差。

7．A 很牛，我和 A 合影了，所以我也很牛。

有图有真相总是容易被人认可和相信，所以，很多人喜欢秀合影来显示自己的身份。当然，能有这些合影的人也确实都是牛人，但频繁地秀合影，就有点多了，因为即使你与某位牛人合影，也并不代表你也是牛人，更不代表这位牛人和你是朋友，或者记住了你的名字。

这个社会上，频繁与牛人接触，确实会助长自己比较牛的感觉，也往往会让自己飘飘然，觉得自己也已经是牛人圈子里的，比如那位号称拥有数十位国家元首政府首脑好友的央视名记者，只有身陷囹圄之后才发现，一切都是浮云。

8．今天某牛人接受了大家的采访，我也参加了，所以我专访了某某牛人。

合影还好，至少表明这个人与某牛人零距离了，还有更不靠谱的，就是专访。不明就里的人看到某某专访了某某牛人，就以为是某某与

某某牛人像央视《面对面》一样的一对一访谈了两个小时。实际上，多数的所谓专访，都是一对多，这位号称专访了某某牛人的某某，只是众多参与者之一，甚至一个问题都没有提，或者离被采访者足有五十米远。至于专访的内容吗？其实和新闻稿或者公开谈话没啥区别。

9．某公司这一点有问题，所以某公司就是问题公司；某人有这个缺点被发现，所以这个人是坏的；我们调查了某某，他说没看到，所以这件事情就不存在。

其实，这样的逻辑和前面分析过的有些类似，但却并不完全相同。我们很多人往往会陷入两个极端，一是认定这个人好，这个人做得都是好的，二是认定这个人有一点不好，就认定这个人都不好。实际上，两种思路都不可取。

证明一个人是好的，也许任何的举例都不成立，即便案例再多，也无法代表全部，但要证明一个人是坏的，就仅仅需要一个点的证明就可以了。以点击面，是驳斥对方的好方法，但也容易导致走向以偏概全。

很多人，看似理性，非常善于抓住小细节和小辫子，总觉得自己聪明，可却只见树木不见森林。在分析中，典型案例的价值不可低估，但这个典型却应该是站在以面上的数据分析为基础之上，否则，就只是抬杠而已。

反过来，很多网络上的分析都采取的是以点击面的方式。比如，原来流传在飞机上能看到长城，后来，有人采访一个航天员，说你看到长城了吗，他说没有，所以，很多人就得出结论，飞机上能看到长城是谣言。有人说，西点军校挂过雷锋的照片，但最近有记者采访了一名西点军校的在校生，这位军校生说自己没见过雷锋的照片，所以，

很多人说西点军校挂雷锋照片是假的。也许结论是对的，但论证过程却绝对不正确，如是用穷尽法，需要找到所有人进行核实，全部都否认，这个结论才成立。很多人拿非充分条件来暗示或者直接得出结论，在逻辑上是错误的。

10．A 是好的，A 是好的，A 就是好的。

这种分析思路就不需要讨论了，有些人就是这样任性，不管你说什么，就是要好话说三遍，我认定的就是对的。同样，这些人还会坚持，B 是坏的，B 是坏的，B 就是坏的，你更没有办法反驳了。

数据分析需要掌握一定的基础知识，依照科学的方法来进行，任何错误的违背逻辑的分析都会带来有害的结果，更重要的是，任何的数据分析都需要站在客观公正的立场，不戴着有色眼镜看待事物，也不预设结论的进行分析，否则，数据分析就会成为欺世盗名的工具或帮凶。

为什么生儿子的司机车险出险率比生女儿的高

蚂蚁金服的首席技术架构师在分享会上透露了两个研究成果，一个是“生了儿子的司机的车险出险率要比生女儿的高”，另外一个是“朋友圈里朋友比较稳定的车险出险率比较低”。为了不曲解原意，将原文贴下：

> 通过大数据你可以更好地分析一个用户的行为，包括他的个人特征，包括金融资产、地理位置等。我们提供了强有力的平台，

而且经过蚂蚁金服这几年经验积累下的能力，我们有 3000 多个模型，能够帮助用户更好、快速地知道用户到底是什么样的，用户的画像是什么样的。大家对大数据没有太多感触，不是说你买一个尿不湿，旁边必须放一个啤酒这样老套大数据的产品。我给大家举一个我们自己亲身感受的例子，就是车险的例子，一个生了儿子的司机出险率大于有女儿的司机。大家可以感受一下，我不说结果。另外一个他的朋友圈经常变动，朋友圈的朋友经常变动，他的出险率大于一些固定朋友圈的这些人。

我们不禁要问，为什么生儿子之后的司机会经常出车险呢？经过在万能的朋友圈里众筹答案，终于恍然大悟，也许是因为一个人生了儿子以后，生活压力会比较大，要时不时地想着未来儿子的娶妻买房和就业，这样就会更着急地去工作挣钱，因此也就会比较容易发生交通事故。

如果这种分析是真的，那么，在中国生一个儿子该是多么负担沉重，而那些生了两个儿子的该怎么去面对生活压力呢？当然了，每个家庭都有自己的生活方式，如果不是去攀比，而是恰当地去追求和努力，也许生活就不会太过辛苦，也就不会有那么多急匆匆的司机，更会少很多的路怒族。

至于另外一个，朋友圈稳定的人交往的范围也比较稳定，日常生活方式可能比较固定，节奏也不会太快，生活压力也不大，也就是说活得比较从容，自然不容易发生交通事故。相反，朋友圈天天增长，变化非常大的，往往都是在外交际比较多的，也容易喝酒，发生交通事故的概率要大一些。

正是由于有了这样一些大数据的成果，现代的保险公司就可以更

准确地确定一个司机的车险费率，保险行业也就进入到了大数据时代。当然，保险行业对于数据的使用由来已久，是对概率统计最为重视的行业，也是完全依赖数据分析结论来挣钱的行业，只是在大数据时代，保险公司就有了更多维度和指标来对客户进行分析，或者研发更加具有前景的产品。

现在的互联网金融公司可以根据用户的一些日常行为，甚至是看起来和金融毫不搭界的信息，分析用户的信用情况，然后就可以以最快的速度处理客户的借贷申请。电子商务企业根据用户的日常浏览和购买行为进行分析，预测购买倾向和关心的商品，从而给用户进行有效的推荐。一些 APP 应用也可以结合用户的使用行为和终端系统的环境对用户的账户安全做出评价，在遇到账号信息泄露的时候有效地保护用户的资产安全。

不过，以上我们这些对司机和朋友圈的结论的因果分析也都是不靠谱的，没有任何数据和调查支撑我们的分析，所有的理由都是猜测，这也是大数据分析的普遍现象。重视相关，不重视因果，让大数据分析有现实的应用价值之外，也增添了更多的不可靠性。大数据给了我们一把刀的同时，也给了一块盾牌，就看我们用在哪里和怎么用了。

大数据营销不能自作聪明，别小瞧你的消费者

借助大数据，企业获得了很好的营销手段，可以精准地定位消费者，也可以根据消费者的喜好设计与生产适销对路的产品，还可以对消费趋势进行预测。于是，很多人认为，大数据时代，商业企业成为

了聪明人，背后的潜台词是，消费者就成为“傻子”。

事实可能不是这样。道高一尺，魔高一丈。在商业企业利用大数据提高自己营销能力的同时，消费者也在进步，甚至聪明进化的速度还可能超过商家，因此而形成新的买卖能力平衡。大数据只是工具，商家可以用，消费者也可以用，而且，很可能消费者的利用程度更高，能力也更强。

借助互联网，消费信息的获取更加容易，信息越来越透明

信息不对称是商业得以发展的本质。传统的商业时代，买卖之间的信息获取能力差异非常大，地域分隔、时空隔绝、传播困难，消费者获取信息的成本非常高，在这种情况下，只要稍有头脑的商家都可以借助信息差异获取利益。

美国经济型连锁酒店红屋顶酒店（Red Roof Inn）在 2013/2014 年业绩创纪录，原因是冬季航班取消率在 3%左右，这意味着每天有 90000 名乘客滞留，而这家酒店旗下的许多酒店毗邻各大机场。这家酒店的营销和分析团队协同工作，充分利用天气状况和航班取消方面的信息，知道大多数客人在移动设备上使用互联网搜索来查找附近的旅店后，启动了一项颇有针对性的营销活动，使得其在采用这项策略的地区的营业额增长了 10%以上。

在互联网时代，信息呈现爆炸性发展和蔓延，购物的时间差和区域差已经在互联网以秒计时的全球化传播冲击下不再成为壁垒。每个消费者都可以上网查询资料、比较价格甚至可以借助一些信息化技术进行虚拟试用，信息透明化让商家彻底曝露在消费者面前。

即便是大数据本身，不仅能服务商家企业，也能够服务消费者，

成为消费者购物时的重要参谋和助手。有经验的消费者，日积月累就会形成自己的大数据，还有一些专业机构综合汇总各种数据提供更加专业化的建议和参考服务，与一些商家孤立和割裂的数据相比，社会化的大数据更具有优势，也成为对抗商家大数据“忽悠”的利器。

商家销售行为的传播速度极快，很容易形成效仿和其他竞争者的及时应对，脱颖而出更难

与消费者博弈大数据的使用还只是一个侧面，更严重的威胁来自于与直接竞争对手或者潜在进入者的面对面对抗。

在世界各地拥有 1200 家酒店的喜达屋酒店及度假村集团系统分析当地及世界经济因素、活动和天气预报，以此优化房价。由于知道了北美核心客户群的本国天气如何影响那些客户在阳光灿烂的加勒比海度假一周愿意花的钱，他们知道了什么时候降低房价或开展营销促销活动最合适，其每间客房的收入增长了近 5%。这样的策略当然有效，可会是其一家独享的方案吗？

大数据时代，信息传播的速度极快，大数据也成为信息搜集和分析的重要方式。在这种情况下，一家企业开展的营销活动，很可能在发起的初期甚至还没有正式上线的时候就被对手获知，针对性的营销方案已经在制定中，商家已经很难建立起差异化的营销优势。

一种新产品上市，即便有专利的壁垒，但通过大数据的方式很可能被竞争对手反向工程，或者通过大数据分析出产品的优缺点与消费者的痛点，在竞争对手刻意的模仿与微创新之下，产品的优势也很难长期保持。

还有潜在的竞争对手在蠢蠢欲动，以往行业的门槛在大数据时代越来越低，一些跨界的巨头借助自身掌握的大数据能力切入新领域更

加容易，也给不同的行业带来了格局上的变化，新老企业都面临巨大的挑战和压力。

信息爆炸造成信息风暴，一招可以制胜，反过来，一个烂招就可能变得一败涂地

大数据也不会是百战百胜的。事实上已经有过很多大数据营销失败的教训，有平台预测过的某电影的票房会很高，可结果却以惨淡收场，至于那些号称用大数据预测球赛结果与竞选获胜的，更是屡屡失算。

在有些时候，大数据真有点像算命先生，即便很多次预测准确，但只要一次失手，就有可能前功尽弃，一世英名付于流水。2008 年，Google 第一次开始预测流感就取得了很好的效果，比美国疾病预防控制中心（Centers for Disease Control and Prevention）提前两礼拜预测到了流感的爆发。但是，几年之后，Google 的预测比实际情况（由防控中心根据全美就诊数据推算得出）高出了 50%。媒体过于渲染了 Google 的成功，出于好奇目的而搜索相关关键词的人越来越多，从而导致了数据的扭曲。

借助大数据的研究成果和大数据的手段，可以使用一个妙招或开发一个产品实现爆款，但也有很大的风险因为一个失误而马失前蹄，功败垂成。

如今，大数据的基础是信息大爆炸，同时伴随的也是信息的传播风暴，风暴口上站着，有可能被吹得飞将起来，也有可能被吹到大海里变得杳无音信。

消费者的信息太多，选择太多，大数据分析结果的适用性在下降

信息多是好的，但信息太多也有可能呈现负面结果。大数据需要大量的全面的数据资料，可越大的数据越全面的数据，就越容易受到噪声的影响，分析结论的可靠性反而会下降，错误地使用大数据，还不如没有大数据。

一家保险公司想了解日常习惯和购买生命保险意愿之间的关联性。由于随后觉得习惯太过于宽泛，该公司将调查范畴限定到是否吸烟上。但是，工作仍然没有实质进展。不到半年，他们就终止了整个项目，因为一直未能发现任何有价值的信息。

就消费者行为分析来说，商家借助各种手段来研究消费者，包括消费者的个人资料、家庭信息、收入情况、历史消费行为、爱好，甚至开什么车、吃什么饭、经常与怎样的异性约会等，但这些信息太多太杂以后，也会让分析者无所适从。

即便分析出来，因为现在的消费者追求个性化的程度很高，同时由于有跟风的习惯，其他人的消费行为也对每个人的决策构成巨大的影响，分析出来的结论在应用过程中时刻会发生场景变动，大数据也会表现得不如预期。不是大数据用错了，是这个世界变化太快。

大数据对于营销非常重要，信息的多寡甚至已经成为决定企业竞争力的核心要素，但大数据也不能盲目迷信，甚至都不能太过乐观。在大数据的应用上，商家与消费者是在同步提高的，自作聪明的商家肯定会聪明反被聪明误，诚实守信尊重顾客在任何时代都不会过时。

第 7 章

换个角度，让结论海阔天空

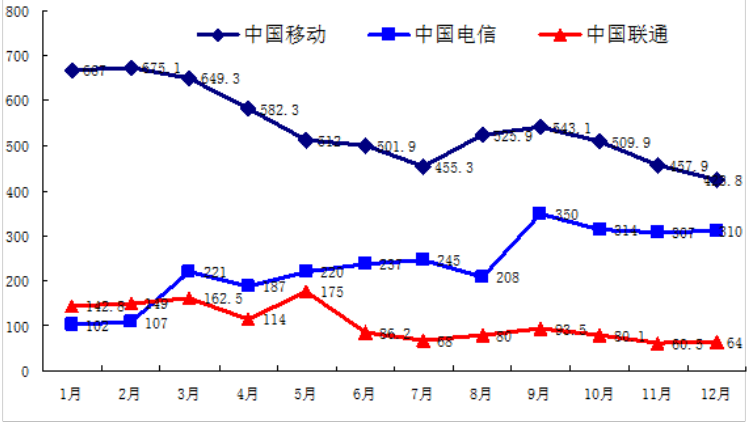
如何看不同的趋势图

看图识字是我们儿时学习文字之时的经历，而在数据分析中，看图更是基本功，同样的数据资料可以看不同的图表，同样的图表也可以有不同的解读，解读能力的不同带来的便是数据应用价值的高低。在一张普通的图表里发现更多更深刻的内容，是数据大师的能力。

图表中，最简单的莫过于条形图、柱状图和饼图，这些往往都比较容易解读，分析的时候主要考验制作的精美程度和图形的直观特征。

折线图是用直线段将各数据点连接起来而组成的图形，以折线方式显示数据的变化趋势。折线图可以显示随时间（根据常用比例设置）而变化的连续数据，因此非常适用于显示在相等时间间隔下数据的趋势。在折线图中，类别数据沿水平轴均匀分布，所有值数据沿垂直轴均匀分布。

下图是一张中国三家电信运营商在 2010 年的每个月新增用户数量的全年走势图，使用了折线图来展示。



先认真思考一下，我们能从这张图中发现哪些有用的信息呢？

（1）三家电信运营商这一年的年底与年初比较，新增用户数是增加了还是减少了？

（2）三家电信运营商的发展趋势是否一致？

（3）电信和联通的用户数发展为何走上了不同道路，中国移动呢？

（4）中国电信这条曲线是否是最有分析价值的？

一般来说，看折线图，可以按照七个步骤来推进：

- 一看整体趋势
- 二看结构变动
- 三看个体趋势
- 四看关键拐点
- 五看交叉断点
- 六看与众不同
- 七看匪夷所思

好，那我们来试一下：

一看整体趋势：从年初到年底，三家电信运营商的新增用户总数变化不大。

二看结构变动：虽然总数变化不大，但新增用户在不同运营商之间的差异变化非常大。

三看个体趋势：移动呈现波动下滑的趋势，联通稳中有降，电信应该算是节节攀升。

四看关键拐点：联通的 5 月，移动的 7 月，电信的 3 月和 9 月都是拐点，去分析原因吧。

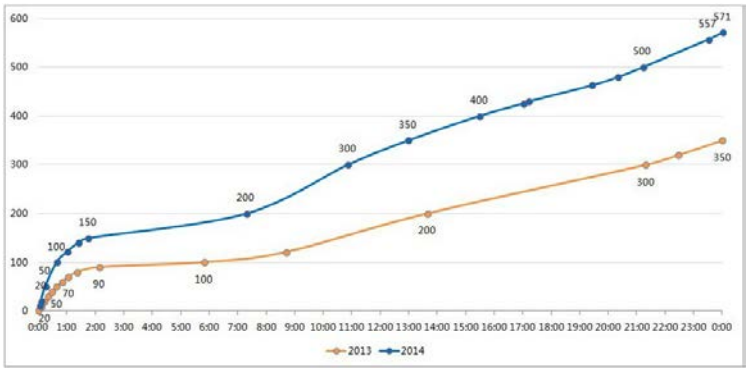
五看交叉断点：只有一个交叉点，这个交叉点意义重大，从此电信扬长而去。

六看与众不同：电信这条线与联通和移动的走势差异很大。

七看匪夷所思：一般来讲，用户数在营销活动前后变动很大，电信却是节节攀升。

怎样，如果你是一位电信分析师，把专业知识结合上面的描述，是不是这张图就分析得差不多了。普通的一张图，也变得很有内含的样子。

分析图表，不怕内容多，就怕内容少。下面这张图是阿里巴巴 2013 年和 2014 年的“双 11”全天销售走势图，使用了平滑曲线，记录了每个小时完成的销售量。



一张图，没有拐点、没有交叉点，也没有明显的整体趋势差异，两条曲线走势的差异也不大，这样的图表该怎么分析呢？

遇到这种情况，我们需要确定坐标，也就是要自己找到“参照物”，让不同的曲线能够在相对的位置上进行比较，从而发现有价值的信息。

这张图，只有时间轴和销售量的数量轴，所以，可以用时间为坐标，比如比较一下中午 12 点，或者其他具体时间点上的差距，也可以用数量为标杆坐标，比如完成 100 亿元、200 亿元这样的关键节点所使用的时间。在天猫的分析中，突出了完成 100 亿元交易的时间对比，

2013 年用了将近 6 个小时，2014 年用了不到 1 个小时的时间就完成了，当然，我们后来也知道，2015 年的双 11 交易额突破 100 亿元只用了 12 分钟，成长之迅速显而易见。更重要的是，我们可以去深挖，为何从 2014 年开始，第一个小时的交易额增长如此之快的原因了。

人均预期寿命提高，你真能多活一岁？

在十三五规划中，提出要将中国人均预期寿命再提高一岁，让很多人欢欣鼓舞，认为自己可以活得更长一些了。彭祖是中国传说中寿命最长的寿星，《国语》和《史记》上都有记载传说他活了八百年，我们什么时候可以有如此高寿呢？

提高一岁，到底多不多

据《中华人民共和国 2015 年国民经济和社会发展统计公报》显示，2015 年中国人的人均预期寿命 76.34 岁。而根据此前的资料，2010 年中国人均预期寿命是 74.83 岁，也就是说，此前五年，中国人的人均预期寿命提高了 1.51 岁。接下来的五年，如果仅仅是提高 1 岁，你觉得是有多高呢？

《世界卫生统计 2014》的报告指出，基于全球平均数据，2012 年出生的女孩预期寿命约为 73 岁，男孩预期寿命为 68 岁，均较 1990 年出生的人口预期寿命提高 6 岁。世界上，日本是公认的人均寿命最长的国家，在 2014 年就达到了创纪录的 87 岁。在性别差异上，日本女性和冰岛男性预期寿命分居全球之首。由此来看，我国的人均预期寿命还有很大的提升空间。

什么是人均预期寿命

按照人口统计学的解释，人口平均预期寿命（英文：Life Expectancy）是指假若当前的分年龄死亡率保持不变，同一时期出生的人预期能继续生存的平均年数。它以当前分年龄死亡率为基础计算，但实际上，死亡率是不断变化的，因此，平均预期寿命是一个假定的指标，它表明了新出生人口平均预期可存活的年数，是度量人口健康状况的一个重要指标。

人口平均预期寿命的计算并不容易，可以说是相当复杂。按照原理，应该是这样计算：我们对同时出生的一批人进行追踪调查，分别记下他们在各年龄段的死亡人数直至最后一个人的寿命结束，然后根据这一批人活到各种不同年龄的人数来计算人口的平均寿命。用这批人的平均寿命来假设一代人的平均寿命即为平均预期寿命。

但是，要跟踪同时出生的一批人的整个完整的生命过程几乎不可能，在实际计算时，往往可以利用同一年各年龄人口的死亡率水平，来代替同一代人在不同年龄的死亡率水平，然后计算出各年龄人口的平均生存人数，由此推算出这一年的人口平均预期寿命。因此，人口的平均预期寿命与同时代的死亡率水平有关，特别是与婴儿死亡率关系密切。事实上，这些年世界各国人均预期寿命的大幅提高也主要是得益于婴儿死亡率的快速下降。我国孕产妇死亡率由 2010 年的 30/10 万降至 2014 年的 21.7/10 万，婴儿死亡率由 2010 年的 13.1‰降至 2014 年的 8.9‰，直接推升了人均预期寿命。世界卫生组织数据，利比里亚人的均预期寿命在过去的 20 年里差不多提高了 20 岁，幅度居全球首位。

提高寿命，两方面很重要

一般来讲，寿命的长短受两方面的制约。一方面，社会经济条件、卫生医疗水平限制着人们的寿命，所以不同的社会，不同的时期，寿命的长短有着很大的差别。根据世界卫生组织的报告，2012 年出生在高收入国家的男孩预期寿命为 76 岁，比低收入国家高 16 岁；高收入国家女孩的预期寿命为 82 岁，比低收入国家高 19 岁。同样都是朝鲜民族，2015 年年底的数据，韩国国民的预期寿命为男 78.2 岁、女 85.0 岁，而朝鲜人的预期寿命仅为男 66.0 岁、女 72.7 岁。数据也显示，2015 年朝鲜婴儿的死亡率为 22/1000，达到韩国（2.9/1000）的 7.6 倍。

另一方面，由于体质、遗传因素、生活条件等个人差异，也使每个人的寿命长短相差悬殊。人均寿命主要与生活质量有关，生活质量越高，人均寿命就越高，还与种族因素有关，东亚农耕人口的人均天赋寿命最长，欧美白人的人均天赋寿命次之，黑人的人均天赋寿命最短。东、南亚大部为农耕人口，大多是性情温和，虽然与欧美的人相比，体质不算强壮，但是天赋寿命却可以很高。同等生活质量条件下，东亚人均寿命比欧美白人人均寿命要高两岁左右。

从长期看，寿命都是在提高的

从历史上来看，地球上的人均预期寿命都是在不断提高的，除了部分特殊年代，比如战乱、瘟疫或者饥荒。

资料显示，公元前欧洲人的平均预期寿命仅 20 岁左右，一直到 1850 年左右才达到 40 岁，也就是说，在漫长的近 2000 年的历史中终于延长了一倍，平均每百年不过增寿一岁。这种极其缓慢的增长速度，显然与古代社会生产力发展的缓慢有关。19 世纪工业化革命之后，社会

生产力的解放，人口的平均预期寿命迅速上升。自 1850 年以来的 100 多年，欧洲人的平均预期寿命大约增加了三十多岁，按 1977 年联合国人口年鉴所示，已达到平均 72 岁的水平。由于医疗技术的进步和卫生环境的改善，特别是抗生素的发现和免疫接种术的应用，欧洲的人口平均预期寿命实现了平均每十年增长 2.3 岁左右的速度。

我国也类似。古代中国人均预期寿命很低，东汉时期，我国人口的平均寿命只有 22~26 岁，唐朝为 27~29 岁，清朝为 30~33 岁，民国时的人均寿命还只有 35 岁。到 1957 年，我国人均寿命提高到 57 岁，1960 ~ 1970 年的 10 年间，人均寿命从 36 岁增长到 62 岁，并于 1975 年达到 65 岁，进入到了现代国家行列。1990 年，人均寿命增加到 68 岁，再过了近 15 年，已达到目前的组织公布的 76 岁。

当然，中国的人均预期寿命指标也是喜忧参半。按照中国卫生部门的数据，我国人均寿命在 2005 年就已经提高到 73 岁，但东西部省份人均预期寿命相差多达 15 岁，显示出严重的地区发展不平衡。

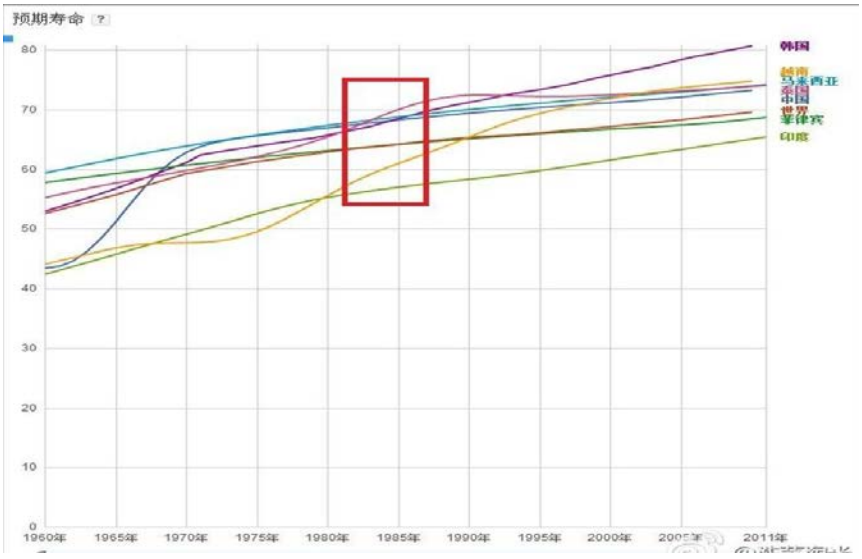
有关中国人均预期寿命存在的巨大争议

当然，中国人均预期寿命 76 岁这个指标也有争议。联合国开发计划署公布的 2011 年人类发展报告及人类发展指数排名中，人均寿命排在前几位的分别是日本、中国香港和瑞士。中国排名第 83 位，只有 73.5 岁。可根据 WHO 这个世界上最权威的医疗卫生组织提供的信息，2011 年中国人均预期寿命已经上升到 76 岁。两个国际机构的数据相差极大，也直接导致了是中国还是越南人均预期寿命更长的争论。

另外，更加引起争议的是中国南北方人均预期寿命的差异。根据以色列、中国和美国学者共同完成的研究发现，中国的空气污染已使

中国北方居民的预期寿命减少 5.5 年，后来被媒体形象地解读为中国南方人比北方人要多活五年，这也可以解释为何东北人口连年快速下降，而大量的人口都迁移到了广东、海南的原因。

在研究利用中国各地数十年期间的污染数据之后，由美国麻省理工学院、北京的清华大学和北京大学，以及耶路撒冷希伯来大学的多名教授联名发表报告称，在南方的平均污染水平之上，每增加每立方米 100 微克微粒物质，出生时的预期寿命就减少 3 年。据这项研究估算，华北的空气污染已在 20 世纪 90 年代致使淮河以北的 5 亿中国人总共减少 25 亿年的人类预期寿命。不过，北京是个例外，受到空气污染的北京 2013 年预期人均寿命超过 81 岁，空气清新的美国 2013 年人均预期寿命 77.9 岁，可见北京的经济发展水平已经超过美国，由此来解释北京房价可以买下美国也许比较合理。



最后，我们要说，人均预期寿命并非是每个人都会活到的年龄，因为有“人均”，而且，当年的人均预期寿命从本意上是说当年出生的人口在无意外的情况下的预计生存年数，可很多人在“当年”之前已经生活了很久，已经没有可能享受到所有此后的社会发展红利。当然，也并非说以后出生的人一定会比此前的人要活得久，还要看社会是否正常发展，不信看看叙利亚，在 2011 年还高于或者至少是与中国相当，可如今呢？我们祈祷和平，祝福安康，只有如此，才能活得更长。

跳楼？数据也会说假话

案例回放：

2010 年 1 月 23 日凌晨 4 时许，富士康 19 岁员工马向前死亡，公司悬赏 50 万元征集线索。

2010 年 3 月 11 日，富士康龙华基地生活区一李姓员工从宿舍楼 5 楼坠地身亡。

2010 年 3 月 29 日，龙华厂区，23 岁湖南籍男性员工从宿舍楼上坠下，当场死亡。

2010 年 4 月 6 日，观澜 C8 栋宿舍饶姓女工坠楼，仍在医院治疗，18 岁。

2010 年 4 月 7 日，富士康观澜厂区外宿舍，宁姓女员工坠楼身亡，18 岁，云南人。

2010 年 4 月 7 日，观澜樟阁村，富士康男员工孙丹勇身亡，死者 22 岁，湖北人。

在年初“跳楼事件”的声明中，富士康方面只是表面潦草解释员工自杀系自身心理因素所致，与企业无关。

2010 年 4 月 10 日，富士康集团媒体办主任刘坤、富士康集团卫生部部长芮新明以及富士康工会副主席陈宏方，接受采访时数度用“检讨”一词表态：“虽然富士康在深圳厂区有 40 多万人”，管理难度很大，“但是这不能成为我们推脱的借口”。

2010 年 5 月 6 日凌晨 4 时许，随着富士康员工卢新跳楼身亡，之前深陷“跳楼门”事件的富士康，再次被推到舆论的风暴中心。

据知情人透露，郭台铭因跳楼频发事件，委托富士康副总裁何友成，请来五台山高僧做法事，祈求公司能平静下来，为员工祈福。

2010 年 5 月 11 日晚间，又传来一名女员工跳楼的消息。这已经是自今年 1 月 23 日，富士康 19 岁员工马向前坠楼死亡后的第八起跳楼事件。

2010 年 5 月 14 日一名年仅 21 岁的安徽籍男员工，从宿舍楼 7 楼楼顶坠下，当场身亡。富士康“九连跳”事件引发了媒体和社会的关注。

2010 年 5 月 19 日，深圳市副市长、公安局长李铭来到深圳富士康科技集团，就富士康近期连续发生员工跳楼事件进行调查，并与该集团高层商讨防范措施。

2010 年 5 月 21 日，一名年仅 21 岁的男性员工南钢从 F4 栋楼跳下身亡。富士康“十连跳”事件引发公众关注。

深圳总工会、深圳公安局及宝安分局、富士康工会及富媒办等当日召开紧急联席会议，商议对策。最终采取的最有效措施：

建设安全防网，加强安保工作。富士康用巨资修建 300 万平方米的防护网，并增加安保人数，训练安保人员，防范再发生自杀事件。

富士康，专业从事电脑、通讯、消费电子、数位内容、汽车零组件、通路等 6C 产业的高新科技企业。富士康持续提升研发设计和工程技术服务能力，在中国大陆、中国台湾、日本、东南亚及美洲、欧洲等地拥有上百家子公司和派驻机构。2008 年富士康出口总额占中国大陆出口总额的 3.9%，连续 7 年雄居大陆出口 200 强榜首。2012 年进出口总额达 2446 亿美元，占中国大陆进出口总额的 4.1%，2012 年旗下 15 家公司入榜中国出口 200 强，综合排名第一。2013 年跃居《财富》全球 500 强第 30 位。我们很多人手里拿着的苹果手机，多数都来自这家工厂。

在富士康发生一系列的跳楼事件之后，“血汗”工厂、没世劳动者的权利、自杀者的个人因素、复杂的社会因素等对自杀的分析纷至沓来，可富士康却感到“委屈”。

来自中新网的报道，“中国是世界上自杀率最高的国家之一，总的自杀率为 10 万分之 23，而国际平均自杀率仅为 10 万分之 10。”我国自 2000 年以来每年自杀死亡人数约在 25 万人，自杀未遂人数至少在一百万人。在 15 岁至 35 岁之间，自杀成为死因之首。平均下来，我国每两分钟就有两人自杀，8 人自杀未遂。2008 年，中国自杀率大约是每 10 万人中有 12 名自杀者，而富士康的自杀率是每 10 万人大约有 2 名自杀者。

我们再计算一下富士康员工的自杀率。富士康员工仍然是中国打工者中幸福指数最高的一族。这是无法否认的事实。仅以富士康深圳

龙华工厂三十万以上员工来计算，按照概率，半年自杀九人，一年可能要为十八人，该工厂的自杀率为万分之零点六七，远远低于社会的平均水平。富士康公司自杀比率要比全国的自杀比率低得多!

当然，正如媒体在报道中所说的，年轻生命的非正常消逝，不能用冰冷的统计数字来遮盖。在这里，数据分析就让位给了感性的思考。数据向我们述说的并非我们感觉到的“真实”，如此是为什么呢?

数据有些时候会与我们的感受不一样，但也都是有原因的。对于富士康这个案例来讲，主要原因是员工自杀的方式非常集中，自杀的时间也相当集中，所以这个时候给人的感觉就会非常特别，所谓的理性的数据分析结论不会成为社会共识。数据在不同的时间以不同的方式呈现，会导致结果的差异。

网上有一位高人，他根据富士康员工自杀的日期做了一个回归分析，以下是具体的分析过程。在这里，我们最要向高手致敬：

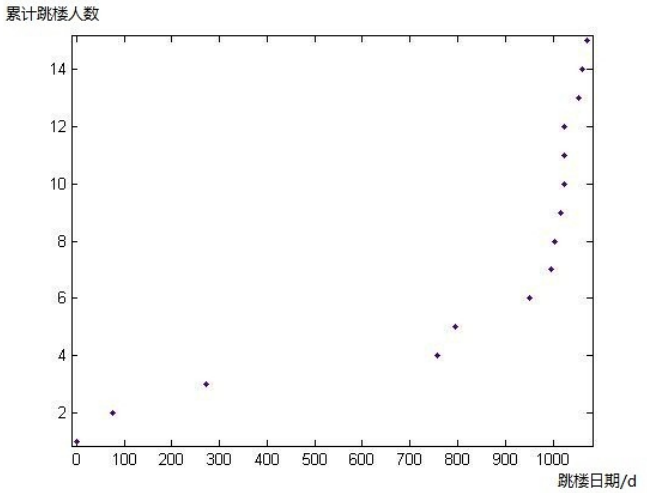
先列出如下数据：

以 2007 年 6 月 18 号，第一例自杀案例为原点，至今（2010 年 5 月 25 日）1072 天。

自杀时间 *x*/*d*：075 272 758 794 950 997 1003 1015 1023 1024 1024 1053 1061 1072

累计自杀人数 *y*：1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

在 Matlab 中容易做出散点图：

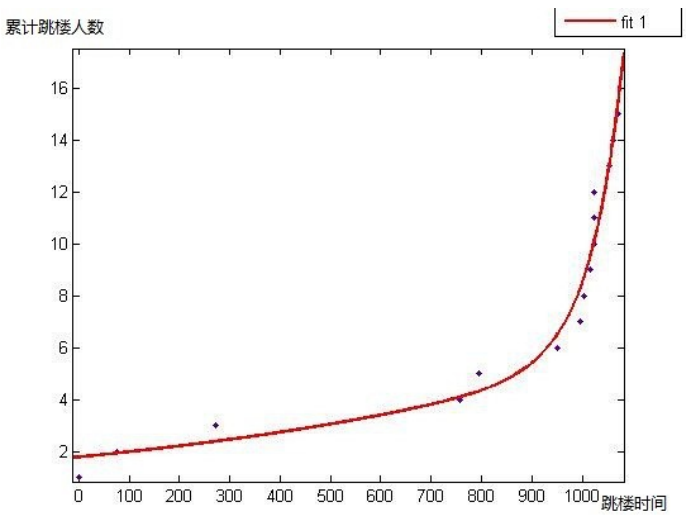


可见这是一个对数增长的曲线。

对此我认为自杀和流行病一样，自杀也是一种病，而且是一种可以传染的疾病。

因此其增长曲线与对数增长很接近。

对其做对数函数拟合：



```
General model Exp2:
       f（x） = a*exp（b*x） + c*exp（d*x）
Coefficients （with 95% confidence bounds）:
       a = 7.569e-007 （-6.561e-006, 8.075e-006）
       b = 0.01529 （0.006473, 0.0241）
       c = 1.782 （0.5788, 2.984）
       d = 0.001075 （2.37e-005, 0.002125）
Goodness of fit:
  SSE: 8.846
  R-square: 0.9684
  Adjusted R-square: 0.9598
  RMSE: 0.8968
```

可见相关度 0.96 也是非常高的。

然而和所有疾病一样，一旦其事件引起了人们的关注，则各方的反馈作用，将阻碍其继续上升。

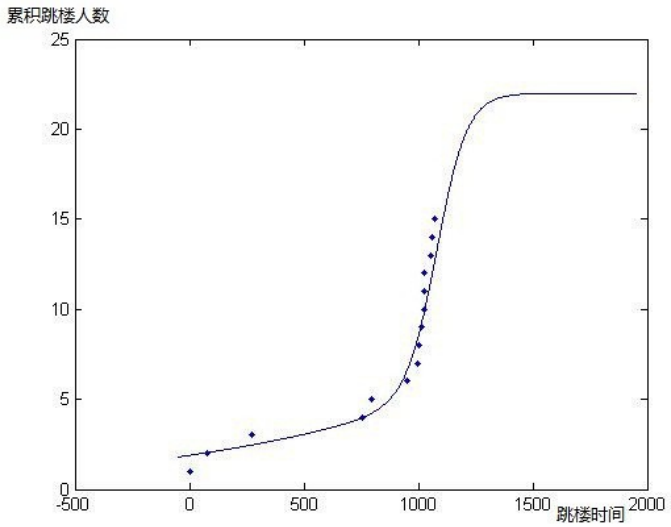
因此，和很多流行病分析一样，该曲线很有可能呈 S 型。对于该曲线的分析，使用 Logistic 回归。

首先我们假设 *Logis*（*B*,*x*）=*F*（*x*），之中 *B* 为参数数组，则由经验和可能的微分方程关系，回归曲线应该为

S（*x*）=*m*×*Logis*（*B*,*x*+*t*）/（*n*+*Logis*（*B*,*x*+*t*））格式

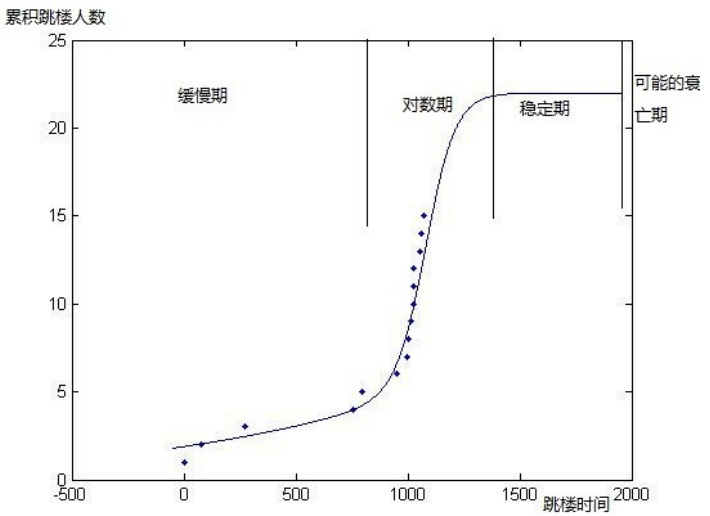
由于当 *Logis*（*B*,*x*）较小时 *S*（*x*）=*Logis*（*B*,*x*），则可以认为 *f*（*x*）的参数可以直接引入 *S*（*x*）作为一种近似，而对于 *m*，*n* 的确定，我以 1 为间隔，画出 *m*×*n*=40×20 的所有曲线，

选出其中最吻合的一条（*m*=22 *n*=20 *t*=50）：富士康跳楼曲线如下图所示。



由此可以分析出，富士康的跳楼人数最终会稳定在 22 人左右。因此仍然不会超过全国平均跳楼率。

对此曲线的分析，我们借鉴微生物生长曲线的方法，将其分为：缓慢期，对数期，稳定期，衰亡期，如图所示。



缓慢期，富士康员工虽然受到很大的工作压力，可是其自身的心理并没有崩溃，因此，跳楼这种事件发生频率很少，而且呈线性关系，说明没有跳楼者受到别的跳楼者的影响。

对数期，富士康员工由于受到工厂巨大的工作压力，以及来自社会各方的压力，甚至加上上级的欺压，心理防线渐渐崩溃，无处发泄。而一旦有想不开者跳楼，则为其提供了一个发泄的模板，这种情况下，很容易有相同经历的员工受到跳楼者的影响，从而一个接一个地跳楼自杀。

稳定期，由于社会、媒体各方面的关注，以及社会，广大人民对工厂的压力，工厂不得不做出改变，员工的心理压力渐渐得到释放，从而员工跳楼轻生频率会很快下降。

衰亡期，可能由于资料长期保存，不小心遗失；或者某机关的辟谣；或者所有人的健忘，导致跳楼人数被修正，被减少。

富士康的事情过去了，但整个事件带给全社会的反思永远也不会过去。对于数据分析工作者来说，冰冷的数据结果也许并不是真正的理性，数据也会说假话。片面相信数据的结果是彻底的教条主义，任何看似非常科学的结论都有可能不能符合“常识”。

一道被改过的阿里巴巴面试题

一般来讲，分析问题可以通过三步。

第一步是描述，也就是观察和分析数据，把数据展示出来的内容列举出来，也许不需要发现任何问题，但描述却是非常必要，如果描

述不清晰，以后的分析就会成为无源之水。

第二步是探测，对于描述出来的数据进行扫描，利用一些分析方法，找到数据中的价值所在，或者是找到数据中透露出来的关键信息。如果是做企业的经营分析，则主要是找到公司运营过程中存在的不足和出现问题的部分。

第三步是提出对策，要通过认真地核对和详细地探讨找到出现问题的原因，找到原因之后才可以对症下药，研究和提出应对方案。

据说，阿里巴巴在招聘数据分析师的时候，有过这样的一道题目。当然，下面的这个题，已经是经过了修改。

下面表格是一家 B2C 电子商务网站的 2012 年其中两周的销售数据，该网站主要用户群是办公室女性，销售额主要集中在 5 种产品上：

（1）从数据中，你看到了什么问题？你觉得背后的原因是什么？

（2）如果要求你提出一个运营改进计划，你会怎么做？

（3）如果可以进一步分析，你觉得应该从哪些方面入手？

	星期一	星期二	星期三	星期四	星期五	星期六	星期日
日期	11 月 4 日	11 月 5 日	11 月 6 日	11 月 7 日	11 月 8 日	11 月 9 日	11 月 10 日
销售额（元）	4789	4884	5001	4989	4950	2980	2768
日期	11 月 11 日	11 月 12 日	11 月 13 日	11 月 14 日	11 月 15 日	11 月 16 日	11 月 17 日
销售额（元）	2062	6045	6100	5998	6010	3420	3089

答案要点（读者可以在空白处试着回答）：

（1）数据中发现的问题

描述：

探测：

原因：

（2）改进

（3）进一步分析

以上，只要认真地分析，会发现更有价值的信息，不信你试试。

以下为参考答案。

（1）数据中发现的问题

① 周末销量少

② 周三销量最多

③ “双 11”销量极少

④ 第二周比第一周销量提高

⑤ “双 11”对本公司的影响仅限于当日流量

问题可能是办公室女性周末活动比较多或外出不方便购物，双 11 因为淘宝活动而被冲击了网站流量。

（2）改进

① 提高周末的销量，比如周末大促销，或者设计手机客户端

② 下次“双 11”活动的时候积极准备和参与，不错过机会

③ 深挖工作日的潜力，特别是周三的销量提升

④ 深入分析未来可能有效提升销量的时间点

（3）如果要想精准的设计营销方案，还需要深入研究每个时间点上的销量分布规律，热销产品的销售规律等。

楼市危急，农民工如何去救开发商

突然之间，农民工成了香饽饽，各地政府、开发商、专家们都把中国房地产拯救的希望寄托在了农民工身上，中国农民再一次要扛起中华民族振兴的大旗。要记得，在几年前，不止一个开发商说过，我们只给有钱人盖房子。

仿佛一夜之间，中国的老百姓也了解了中国房地产市场的真相，在拯救楼市政策出台的前一天，各路媒体还在报道中国房地产市场红红火火，房价一路上涨销售量屡创新高，第二天又都开始舆论转向，中国楼市到了崩溃的边缘，只有依赖农民工兄弟靠搬砖的钱来救市。

大家都有房了，很多人有几套房，刚需成了空想

一般的说法中，政府部门公布的数据显示，在 2007 年，中国城镇人均住宅建筑面积达到 23.7 平方米，农村人均住房面积达到 27.2 平方米，已经达到世界中高收入水平。2010 年，中国人均居住面积已经高达 31.6 平方米，按照 2010 年东欧国家人均住房面积多在 24 ~ 29 平方米之间，新加坡为 27 平方米，韩国也只有 33 平方米，中国已经高于很多发达国家。

在 2012 年，北京大学发布了《中国民生发展报告 2012》。调查显示，2011 年全国家庭现住房完全自有率为 84.7%。在住房类型上，42.2%的家庭现住房为平房，比前一年度下降 8.5 个百分点。全国家庭的平均住房面积为 116.4 平方米，人均住房面积为 36 平方米。

到了 2015 年，我国家庭的住房拥有率已达到 90.8%，其中城镇家庭住房拥有率为 87.%，农村家庭住房拥有率则为 95.8%。

2012 年 6 月，浙江大学不动产投资研究中心、清华大学媒介调查实验室与《小康》杂志联合发布《中国居住小康指数》报告，报告显示，有 21.4%的受访者表示没有买房，65.4%的受访者拥有一套住房，10.9%的受访者拥有两套住房，2.0%的受访者拥有三套住房，0.3%的受访者拥有四套住房。在 40 个城市排名中，居民拥有住房比率为 90.1%的长沙位居榜单榜首。而上海以 67.9%的比率排名 40 个城市末位，其余调查城市的居民拥有住房比率均在 70%至 80%之间。调查显示，自 2009 年以来，超过七成国人购买了住房，有报告数据表明，2013 年城镇家庭拥有多套房比例上升至 21.0%。

再看人口数据，中国 0~14 岁人口比重从 20 世纪六七十年代到现在一路下滑，到 2010 年已经降为 16.6%。随着新增人口的锐减和老龄化加快，住房越来越多会被“闲置”出来，购房动力也已经消失殆尽。

房子都被闲着，房子还在建着，供给严重过剩

当然，这种算法有问题，人均住房面积并不能准确反映中国人住房的现实，因为在投机盛行的情况下，平均数没有意义，太多的住房被小部分人购买拥有，这部分住房一部分用于出租，更多的是在闲置。

根据中国家庭金融调查与研究中心（CHFS）的报告，2013 年中国

城镇地区整体住房空置率为 22.4%，较 2011 年提高 1.8 个百分点，其中北京住房空置率 19.5%、上海 18.5%、天津 22.5%、重庆 25.6%。西南财经大学 2014 年年中发布的一份关于城镇住房空置率的调研报告提出，我国有大量城镇住房处于闲置状态。2013 年我国城镇住宅市场的整体空置率达 22.4%。据此估算，全国城镇空置房为 4898 万套。我国的住房空置率已高于美国、日本、欧盟等国家和地区。

即使美国楼市最差的 2007 年到 2008 年，租房空置率最高时候也就达到 10.7%，自有住房空置率最高只有 2.9%。欧洲国家的住房空置率也很低，荷兰、瑞典一般住房空置率只有 2%，法国为 6%左右，德国约为 8%。

2015 年 11 月末，全国商品房待售面积在 10 月末 68632 万平方米的基础上再创历史新高，达到了 69637 万平方米，折合约 680 万套。机构数据显示，目前 70 个大中城市中，去库存周期超过 12 个月的城市多达 27 个，其中北海去库存周期高达 30.1 个月，烟台 26.4 个月，荆门 25.5 个月，呼和浩特 24.9 个月，三亚也多达 23 个月。

农民工有心无力，根本就没有有效需求

国家统计局的数据显示，2014 年年底农民工的总量为 2.74 亿人，有人算了，这么多人里的十分之一去买房，就可以拯救楼市库存。但是，农民工真的要买房和能买房吗？

从 2010 年起，我国农民工的增速已连续 4 年出现下滑。观察者网查询《2014 年全国农民工监测调查报告》发现，根据抽样调查结果，2013 年全国农民工总量为 27395 万人，比上年增加 501 万人，增长 1.9%。2011 年、2012 年、2013 年和 2014 年农民工总量增速分别比上年回落

1、0.5、1.5 和 0.5 个百分点。

统计局数据显示，2014 年，16 ~ 20 岁年龄段的农民工较 6 年前减少了 1453 万人，减少的幅度超过 60%。与此对应的，2014 年一年的高龄农民工数量就增长了近 600 万人。根据第二次全国农业普查的数据，2006 年，51 岁以上的全国乡村户籍从业人员的数量为 13118 万人。据此，专家估计去掉中间不得不留守乡村（不从事非农工作）的比例剩余高龄农村劳动力只有不到 2400 万人。事实是，也就是说，未来 3 ~ 4 年，农村剩余的高龄劳动力也可能全部转移完毕。高龄农民工不会在城市买房，年轻农民工在减少，谁去买？

就是真的想买，农民工有这个钱吗？据国务院农民工工作领导小组办公室 2014 年 2 月发布的数据，外出农民工月平均收入 2864 元。如果平均一个农民工家庭有两个人工作，其年家庭收入约为 5 万元，以小城市的房价为例，房价收入比达到 10 倍，要在地市级城市购房，这些家庭不吃不喝也需要三年才付得起首付。但是，他们能不吃不喝吗？孩子上学、租房居住、日常消费、家里的老人养老、社会交往，恐怕 5 万元的年收入根本剩不下几个钱，还拿什么去买房救开发商？

有房地产中介老总提出来，以一线城市的收入去三、四线城市买房，大部分农民工支付首付款问题不大。这简直是信口开河胡思乱想，在一线城市的农民工拿着钱去三线城市买房，他接下来在哪里工作，如果去了三线城市，那房贷的按揭用什么来还？更重要的是，一线城市比二、三线城市还眼巴巴地指望这些农民工买房救市呢？

这些年，中国大城市，一边享受着农民工进城带来的人口红利和社会福利，一边使用各种手段排挤歧视农民工，用户口、学籍、社保等让农民工妻离子散造就了大量留守老人、儿童、妇女等社会问题，

可现在却要忽悠他们用微薄的收入去拯救纸醉金迷的地产大亨，这些所谓的优惠政策到底是在优惠谁？

可怜的中国农民工，勒紧裤腰带为中国人造房子，现在钱还没拿到，又要砸锅卖铁把自己亲手造的房子买回来，否则就可能拿不到工钱，只能是一声叹息。

模型都是靠不住的，挑战短板理论

在市场研究上我们一般把这样的缺陷称为“短木板”。要了解短木板还是让我们看看“木桶原理”是怎么描述的。一只水桶，由长短不同的木板条箍制。要想让这只水桶多盛一些水该怎么办？是把长木板加得更长，还是把短木板由短变长？水桶的容积为底面积与桶高的乘积。显而易见，水桶容积大小是由短木板的长度决定的。让长的更长，无助于水桶容积的增加；让短的变长才是办法，这就是管理学中的“水桶原理”。

在球队中出现了短木板，就会成为一颗不定时炸弹，随时可能成为比赛失败的罪魁祸首。如果是位置上的球员实力弱，就可能成为对手攻击的重点，导致其他队员的努力功亏一篑。在企业管理上出现了短木板，就会带来难以想象的灾难。日本“三泽房屋”建造的大量房屋屋顶，竟然在一次强台风中同时被刮走了，调查表明，公司的工程技术本身是过硬的。问题仅仅出现在一些螺丝拧得不紧。其实，公司经常教育员工房屋工程质量的重要性，每天都要举行员工研习会，职工们也有较强的质量意识，但这次被风刮走屋顶的事故偏偏是由一些

未受过这方面教育的工人所造成的，事件给公司造成了巨大的形象损失。因此，提升短木板就成为了当前市场研究和企业管理咨询的热门话题。

短木板真的应该提升吗？短木板真的能够提升吗？短木板真的值得提升吗？首先让我们对短木板进行一些分类研究。

短木板可以分为绝对的短木板和相对的短木板。在足球领域中，一只球队的某个方面处在绝对的弱势地位，和任何对手相比都是有差距的，比如南亚一些足球队的球员身高。还有一些，是因为和特定的对手相比存在劣势，在总体上仍然有一定的高度具备一定的竞争能力，比如中国队员的头球能力，如果和德国人相比自然是短木板，可要是和越南队比赛这个还是优势呢。

短木板也可以分为可提升的短木板和不可提升的短木板。就像 2014 年世界杯中的英格兰队，可能因为失去强力前锋鲁尼，在这个位置上埃里克松不可能寻找到达到或者超过目前的鲁尼水平的替代球员，这个短木板在世界杯看来是不可提升的。而德国队可能因为卡恩的不稳定使守门员这个位置有些问题，但是莱曼良好的竞技状态和最近出色的表现，能够让球迷和队员放心，这个短木板就得到了提升。

短木板可以分为容易提升的和不容易提升的。即使是可以提升的短木板，也存在提升容易不容易的问题，有些可能采取一些手段就提升幅度很大，有些可能费九牛二虎之力效果不显著。比如中国队的技术水平，被称为“糙哥”，但这些都是球员从小形成的，很难在短期内提高，只能通过系统的训练甚至几代球员以后得到提高。但是如果是一支新组建的球队，配合出现了问题，随着队友的相互熟悉和训练的磨合，很快就能得到大幅度地改进。

短木板还可以分为值得提升的和不值得提升的。在经济领域或者在进行市场研究的时候我们可以通过边际效益来衡量，如果提高短木板带来了多的收益，这样的提升就是值得的，相反就不值得。比如，中国队为了提升球队的防守能力，把有组织天才的中场核心郑智长期固定在中后位位置，导致球队进攻乏术，最终失去了世界杯出线的资格，这样的短木板提升就是不值得的。

我们要找出短木板，同时也要分清楚哪些短木板是可以提升的、是容易提升的，是值得提升的，然后采取相应的对策实现团队作战能力的整体提高。英格兰前锋出现问题，可能就会让拥有很强实力的中场球员来更多地参与进攻来弥补。2002 年的世界杯不被人看好的德国队获得了亚军，虽然德国队的技术是严重的短木板，可是德国人充分发挥了身体和配合的优势。同样，韩国队从实力上比欧洲劲旅相差很大，但是凭借主场优势、充沛的体能、快速的奔跑和不屈的意志先后战胜了对手进入四强。

因此，是充分发挥优势还是努力弥补劣势，是把自己的长木板加得更长还是把自己的短木板加长，应该根据客观形势来分析判断。特别是在激烈竞争的赛场或者市场中，有时候我们没有办法化腐朽为神奇，如果一味地追求整体的均衡，把有限的资源耗费在无用或者少用的地方，即使短木板加长了，竞争的结果也不一定是胜利。

大数据也有做不到的事

大数据是人类认识世界的一场革命，也为各个领域提供了新的手

段和方法，但是，大数据绝非万能。在很多地方，大数据也无能为力，甚至，大数据还可能会把我们引向歧途。

大数据也可能失之偏颇。虽然大数据强调的是“全”，抛弃了抽样的方法，但是，从认识世界的角度来看，“全”只是一个相对的概念，数据的完整性拥有都只是我们追求的目标，而不是实现的结果。比如，支付宝公布过一个数据，黄焖鸡米饭成为了中国人最爱点的美食，超越了兰州拉面和沙县小吃，这个数据就只能作为参考，因为就当时支付宝获得的数据来看，主要是来自使用其支付的用户群体，这就与用户群体的特征对数据的影响密不可分。

大数据看重量而非质量。腾讯可以分析出你有多少个微信好友，也可以知道你与每个好友聊天的次数和密度，甚至还可以知道你们聊天的具体内容，但是，次数多密度高可能代表社交关系的密切，但并不一定体现出人与人之间的关系强度，夫妻之间可能在微信上聊天很少，天天在那里聊的可能只是因为推销保险。

数据弄不懂幕后的背景。我们之所以是人，就是因为人的大脑还未被研究清晰，人类的决策过程是基于复杂到至今无法理解的连续性，大数据却都是会看成离散的事件。也许，依靠大数据，人们可以让某些机器通过深度学习而有一定的“智能”，但要讲替代人脑还并不现实。数据分析不懂思维的产生过程，也不能完成叙述之中的思路。

客户保有和流失预警对于任何一家公司都很重要，所以某银行聘请了一位数据专家，构建了评估用户是否即将流失的模型。该模型上线，银行也开始给那些被认为即将流失的客户发出信件加以挽留。结果，很快发现了令人惊奇的事情，有的客户的确即将流失，但并不是因为对银行的服务不满意。他们之所以转移财产（有时是悄无声息的），

是因为感情问题——正在为离婚做准备，这种对客户不合适的“挽留”让客户大为光火。

中国银联曾经针对 ATM 查询服务做过网络民意调查“每一笔查询收 1 毛 5 分钱，有多少人会放弃这项服务，采取其他的渠道查询银行卡的金额。结果，调查出来的结果是 85%的人会选择不再使用这种服务。可是后来收费开始以后，发现有 85%的人继续使用跨行查询业，只有 15%的下降。要知道，当时网上收集到的数据超过 1 万份。为什么得出来的结论和后面的真实情况会有如此大差异呢？实际上，这里面有很重要的数据没有被披露，这个栏目当天的点击率是十几万，但是参与调查的人只有一万多，也就是说，大概 90%的人并不关心，也不在乎这个查询收不收费。在真实的数据后面，也会得出很多并不客观的判断。

大数据会造出一些毫无意义的“大发现”。著名商业思想家 Nassim Taleb 提出，随着我们掌握的数据越来越多，可以发现的统计上显著的相关关系也就越来越多。这些相关关系中，有很多都是没有实际意义的，在真正解决问题时很可能将人引入歧途。这种欺骗性会随着数据的增多而指数级地增长。在这个庞大的“干草垛”里，我们要找的那根针被越埋越深。大数据时代的特征之一就是，“重大”发现的数量被数据扩张带来的噪音所淹没。

大数据也无法避免受到价值观的影响。在《“原始数据”只是一种修辞》一书中，作者认为，数据从来都不可能是“原始”的，数据总是依照某人的倾向和价值观念而被构建出来。数据分析的结果看似客观公正，但其实价值选择贯穿了从构建到解读的全过程。

大数据不是 IT 技术。一个好的数据分析体系，首先得有一个良好

的理论模型，用它去指导分析，然后通过数据不断修正它，任何把数据分析当数学和代码来搞的最后肯定会闹笑话。

支付宝的 2015 年全民账单数据显示，用户的餐饮消费平均每笔 36 元，吃货最多的是上海，其后是北京、杭州、武汉……各色餐饮中，黄焖鸡米饭力压兰州拉面和沙县小吃，成新一代最受欢迎的国民料理！这样的大数据有一定的参考价值，但也仅仅如此。

不管怎样，大数据时代已经到来，我们尽情地拥抱吧！