Broadview
www.broadview.com.cn

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

中国工信出版集团
電子工業出版社

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。

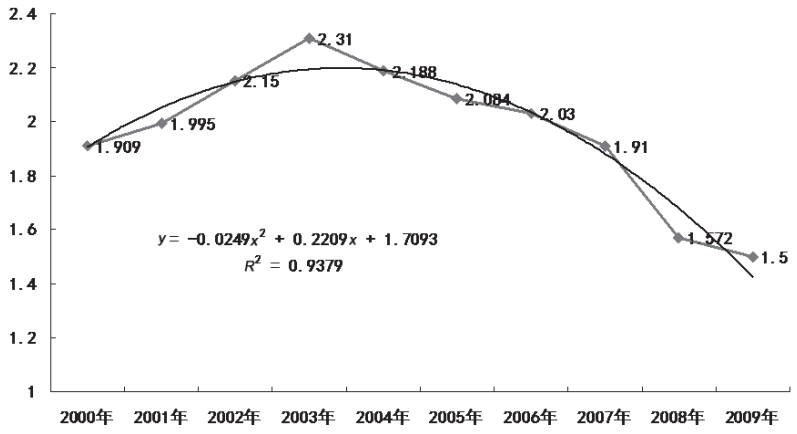
这个时候，我们就会用到所谓的“MECE”分析法，也就是麦肯锡方法里面重点的分析思维。

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。

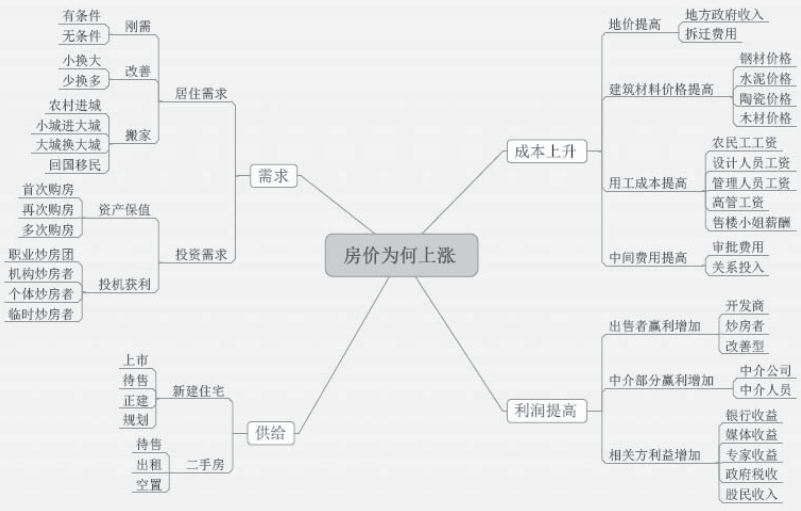
5*X*	2*Y*	1*Z*
2	5	0
	4	2
	3	4
	2	6
	1	8
	0	10
1	5	5
	4	7
	3	9
0	5	10

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。

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。

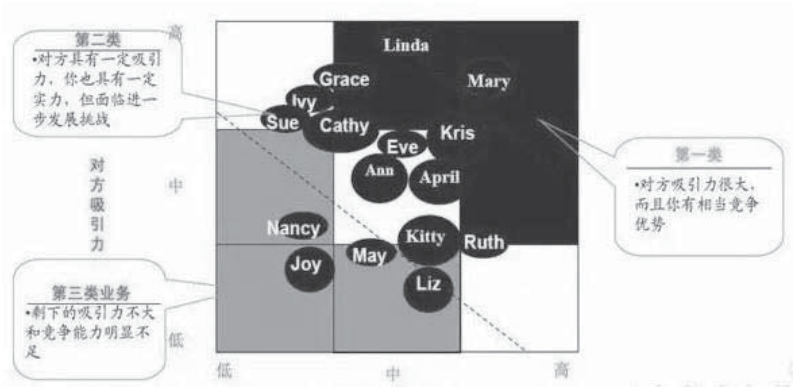
相亲是感性的还是理性的

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：

$$P(k)=x\int_{x}^{1}\frac{1}{t}dt=-x\ln x$$

$$\frac{d}{dx}(-x\ln x)=-1-\ln x=0\Rightarrow \qquad x=1/e$$

$1/e$ 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