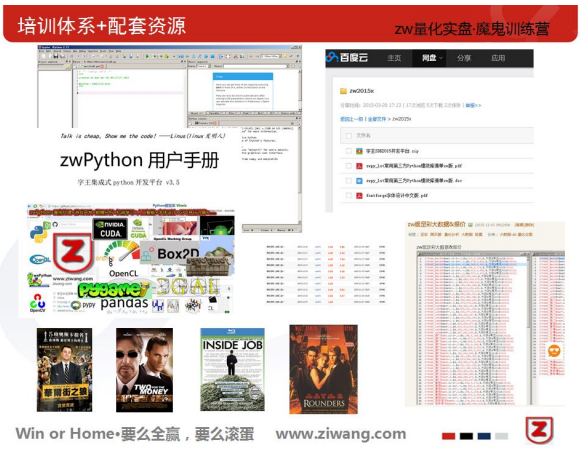
《AI 量化交易、足彩大数据·zw 文选 2015》

作者：zw=智王+字王

编辑：余勤 吴娜



- 字王 Git 项目总览：github.com/ziwang-com/，

  包括：字王 4k 云字库，zwPython、zwpy_lst
- QQ 群：124134140 （AI 量化，足彩大数据、云字库、zwPython）
- 技术 Blog：blog.sina.com.cn/zbrow （AI 量化、足彩大数据、字库）
- www.cnblogs.com/ziwang/ （机器视觉）
- 网盘下载：pan.baidu.com/s/1tY7Wq
- z w 论坛：http://www.ziwang.com

字王看：高频量化交易

凡是无法通过“足彩数据”进行实盘测试的方案、算法，都是在耍流氓。

对于高频量化交易，ZW 的观点主要分以下几点：

1. 采用 2-3 个维度作为数据分析坐标。主坐标，一般可用分（秒）钟的实时数据；第二、三坐标，可以采用关联金融产品，如外汇、贵重金属等参数，这个需要具体测试后再细化。
2. 另外，如果可能，与百度、新浪微博、微信、淘宝等机构，建立实时的 API 数据接口，进行元数据搜索，作为一个参照维度。
3. 数据源，不宜超过 3 个维度，便于数据的 2D、3D 可视化分析。数据维度过高，会带来几何级的数据量，无法保证实时运算和精度。实战测试，数据越多，反而会影响精度。
4. 龙格现象，维度越多，可供单一维度的数据量就也少，反而会影响分析结果。目前“小”数据是 ZW 数据分析的一个重点。老子《道德经·第六十三章》有云：天下大事，必做于细。
5. 策略方面，有分析和统计两种模式，各有优劣，建议采用统计作为匹配模型。这个也是目前大数据分析的一个趋势，人工智能领域的外语翻译项目，六十年代开始，一直采用分析模型，始终无法商业化。2000 年后，互联网的兴起，派生海量语义库，短短几年时间，人机外语翻译已经初步实用化。
6. 传统技术平台，首制于 PC 运算速度，偏重与分析，近年，伴随 CUDA 并行运算的崛起，PC 也可达到以往巨型机 10G 以上的运算速度，分析建模，逐渐被统计建库（数据库）取代。统计模型的建立、选择，实际上也融合了不同团队的策略。
7. 模型建立后，导入历史数据，进行归一化处理、统计分析、聚类分析，可生成 2-3 个维度的数据库，便有了 2D、3D 的数据节点。运行时，获取实际交易数据，按数据节点进行匹配，就可以获得实时的：盈利概率（参数 v）。参数 v，根据预设的交易阀值 K，便可进行买、卖、忽略等预设操作。注意盈利参数 V，其他都是技术细节。实际操盘，采用群组交易，测试表明，针对单一对象的分析预测，远低于多个对象的群组分析。经验表明，对整个数据级，5-8%左右的筛选结果，盈利概率（参数 v）相对较高。

目前 ZW 采用的数据源是足彩数据，原因如下：

2012 年，初期采用国内股票交易数据，自己编程并下载了国内开盘以来历年的日数据，五分钟交易数据，量太大，而且不完整。股票数据，作为数据源，有个先天缺陷，股票交易，只有时间一个维度，无法进行交叉分析，同一只股票，同一个时间节点，没有横向对比参数。

2013 年开始，采用足球博彩数据作为分析数据源，因为同一场比赛，全球有数百家公司同时提供横向的对比数据，同时，同一个公司，同样的赔率，可以提供纵向的对比数据。

当然，还有同一只球队、不同联赛等数据，并未采用。未采用，一方面是限于数据规模，运算速度，另外一方面，是实战测试，数据越多，反而会影响精度。

通过一年的盘前数据分析，相关模型不断优化，目前，盈利概率（参数 v）已经超过 95%.

近期，对比检索了国内数十家相关网站，包括百度、谷歌、微软的世界杯足彩、人工智能项目、大数据项目，以及相关的博彩分析平台。

这个指标，应该是目前行业最高的

以上是个人的一家之言，仅供参考。

技术博客：http://blog.sina.com.cn/zbrow

【补充】

基于大数据的量化投资、股市系统，验收标准，

（基于大数据的量化投资、股市系统，验收标准，摘自 QQ 对话）

注意下盈利参数 V，其他都是技术细节

目前大盘整体波动大，要和大盘平均指数比，不然没有意义

另外，注意稳定性，取 2-3 个月的周平均指数，看看系统模型有没有 bug

字王看：大数据

黑天鹅才是新常态

"啤酒和尿布有什么关系",这个十年前经典案例，目前我是作为反面课件来说的。

这个是冰岛的一个数据分析结果，至少在中国不存在。

金融市场，大家都是大数据，会反向干扰态势的现在（2015） 黑天鹅才是新常态。

看看：石油价格，瑞士法郎，日元升值，光大砸盘，黄金狂跌。全部没节操，没下限！

上海证券交易所周一称因软件设置原因，上交所市场成交金额超过 1 万亿元人民币后无法及时更新，此非技术故障。

黑天鹅算法模型（zPSO，z 粒子算法的升级版）。传统的大数据分析，像 R 语言，置信空间是 95%，也就是说，5%的小概率事件是不考虑的，属于黑天鹅事件。而实盘中，恰恰是这 5%的黑天鹅，才是真正的盈利点所在！

大数据与足彩

关于大数据、高频交易和人工智能，个人的基本观点：凡是无法通过“足彩数据”实盘测试的方案、算法，都是在耍流氓。

足彩数据是最透明的数据源，如果足彩不是，就没有更加公平的了博弈模型。如果这个都通不过，其他都是扯蛋。所以说：足彩是最合适的数据源，有历史数据，还有横向对比。其他任何数据源都没有这种实时的“矩阵”数据源。

2014 年世界杯对于大数据，人工智能是个分水岭，是元年。微软、谷歌、百度都有相关的项目，胜率<50%。根据百度世界杯 18 连胜，可以肯定的说，百度的模型，绝对有人工干预、修正。不然，百度其他业务可以全停了，只作足彩博弈这块,，就可以收割全球的资本市场了。

因为我们的用采集的数据，建模，再回溯测试，准确率也不过 92%！百度 100%的准确率，是无法长期保持的！无法第三方复制、验证！而第三方自由复制、验证，才是一个成功的算法、模型。

大数据与小数据

原本 macd,是股市一个不错的指标，大家都 macd,互相干扰, macd 就成为垃圾数据源，完全失真。信息，知识，不等于智慧。没有合适的模型 ,算法,数据越多，干扰越多。一个模型，容纳所有数据是不可能的 。

你的一个个人微信，QQ 发言，蝴蝶效应，就有可能影响大盘。所以，只能是切片分析。根据实数的概念，实数在数学上是可以无限分割。

实战测试，数据越多，反而会影响精度。目前个人数据分析的一个重点，就是“小”数据。老子《道德经·第六十三章》有云：天下大事，必做于细。我在一个 blog 上面也找到了数学支持,龙格现象，http://zh.wikipedia.org/wiki/龙格现象。维度越多，可供单一维度的数据量就也少，反而会影响分析结果。

我们做的一个模型,原来是 700 多个数据源, 精度反而没有现在 34 个数据源的高。所以，目前我们对于大数据研究的一个重要课题，就是：小数据,如何过滤掉无效的干扰数据源。干扰越多，垃圾数据越多，要求分析的节点越多，真实数据反而被掩盖。做网络危机公关就是这样操盘的，用大量的，无关的关键词，淹没事实真相。

托马斯·弗里德曼，《世界是平的》作者，美国知名时政评论家，在《弗里德曼：你们感受到颠簸，我们看到的是上升 》叙述了他眼中的大数据：

有时候，你的确在嘈杂声中发现新闻，但有时候，你要在无声处发现新闻。学会去聆听寂然无声中的声音，非常重要。人们喜欢谈论各种大数据，各种调查数据。但不要忘了，不只有数字信息才是数据。不要忘了，一句引语也是数据，一个人讲给你的故事也是数据，某个人回答问题是低头看鞋，那也是数据，他抬了一下眉毛，那也是数据。你要收集所有这些数据。

大数据与实盘

多动手，实盘验证、优化、调整如此反复几轮后，自己的观点，就不会光是书本上的了。

看书和自己做是不同的，企业很少直接用大学实验室的东西。为什么？因为，实验室的环境太干净，是没有干扰源的，实战不同的。任何有市场价值的模型, 特别是高频交易，资本市场,都不会直接出现书本上。反过来说，任何纸面上的模型，都是理论型，任何不是基于一线、实盘的、大数据分析,，都没有实际意义。

我们去年和一家高频交易的 boss 沟通，他们对标的企业，直接就是高盛。你想想， 高盛可能把自己的模型，算法，卖给第二家企业？多少钱合适？ 10 亿、100 亿？如果这个算法、软件真心能够在市场上赚 100 亿，客户才会买。不过，如果能够，在市场上赚 100 亿，高盛为什么不自己赚？ 要培养一个竞争对手？！！

大数据=哲学+数学

大数据的本质是：哲学+数学。

而易经，有可能，是唯一融合了哲学+数学的模型。只用 64 个维度 0、1（阳阳）两种状态， 就描述天下各种事态。如果能够数字化，也许是一条途径。

大数据模型的核心是：聚类分析，个人认为， 武汉 xx 大学邓聚龙教授的《灰色数学》在数据归一化， 聚类分析，方面都有独到之处

大数据与郑国渠

这两年，国内大数据貌似太阳能、风电样被炒的很火。贵州还开办了大数据交易中心，也许是全球第一个。国内政府在经济乏力，科技相对落后的情况下，强行推进大数据 ，甚至提升到国家战略层面，却有可能陷入欧美国家的战略陷阱当中。

春秋战国的郑国渠，美帝的星球大战，都是成功的经典战略欺骗案例，还有所谓的千年虫、.com 科技泡沫经济，都历历在目。一个国家的资源是有限的，战略重点也是有限的，不可能到处重金投入

国内前几年火爆的太阳能、风电新能源，目前都处于行业性崩溃，而且，在短期内，也许 20-30 年内，甚至 50 年，无法恢复元气。由此，耗费的资金、人才、资源，只能是全民买单。日本九十年代强推第五代电脑：人工智能电脑，方向错误，越努力越失败，今天的结果是，整个日本国家的 IT 产业崩溃。

国内政府强推大数据，提升为国家战略的另外两个“潜在“考虑，可能是：

- 基于大数据、信息科技的新型“计划经济”，个人对经济不熟悉，但直觉上觉得不靠谱，至少目前没看到有这方面的理论体系，而成熟的理论体系，是项目成功的基本要素。有了成熟的理论体系，未必一定成功，没有，绝对是失败
- 建立类似 1984 的社会管理体系，这个更加不靠谱，网络危机公关的经典手法就是，采用大量的关联信息，淹没负面新闻。一组（10 台）电脑，每天可以发布上亿条信息（包括填写验证码），可以模拟千万级的用户数据。
- 政府决策部门，跑步进入数据共产主义，多半是被神奇的“人脸识别”算法和淘宝、支付宝后台数据唬住了，就像古代方士们神奇的魔术表演。“人脸识别”其实是个很简单的 opencv 通用算法，普通的手机、平板都可以实现，不需要大数据、也不需要云计算，我们发布的开源项目：zwPython，就内置了相关模块和算法、以及源码。
- 淘宝、支付宝的海量数据，也没有多么神奇，余额宝的利息，目前也和普通基金、定息差不多。

大数据、云计算，看起来的确很高大上，比玩地产的土鳖“逼格”高太多了，比玩实业的工商企业轻松多了。可是，大数据的的核心硬件服务器、软件、数据库，都要进口，而硬件服务器的折旧比汽车还快，最前沿的硬件，基本 3 年就基本价值归零，就是一堆废铁。因此，目前各地政府的批量上马数据中心、计算中心，投资回报更加令人担心，一个 3-5 年，回报率无限归零的项目，而且投资总额分分钟过万亿。

顺便说一句，个人是国内首家 4A 级网络公关公司的联合创始人之一，服务过 150+国际 500 强，包括微软、奔驰、西门子。 淘宝、微信、app 市场的好评刷单，目前高达 50-80%以上，这么多的垃圾信息，将真实数据完全淹没. 政府其实也知道这点，所以提出了网络、手机实名制，以及目前的一卡通，希望能够强行绑定信息发布主体。

可是，即使 20-30 年后，一卡通完全推行，还是无法解决这些问题，至少，已经运行了十年的支付宝，目前的假号，才几元一个，最严密的银行卡也不过 200-300 元一张。更何况，数据并非越多越好，有时候数据越多，精度更低，这个数学上称为：龙格现象。实战测试，数据越多，反而会影响精度。

大数据其实并非新科技和高技术，其核心与本质，不过是数据分析，尤其是聚类分析

这点，国内武汉华中科大邓聚龙教授，1982 年提出灰色系统理论、灰色数学当中灰色聚类、数据归一化算法，目前依然是最好的分析模型之一。

大数据分析的核心，是统计分析、聚类分析，以及各种各样、五花八门的分析模型。这些分析模型与算法，大多基于传统的人工智能研究，什么啄木鸟算法、萤火虫算法、蚁群算法，大部分都是经验性、实验模型，缺乏系统的理论支持。这些模型，全部都是高次多元的，而三元以上的 n 次(n>3)模型，除了特殊的经验公式，在数学上是无解的，至少目前没有一个通用的求解算法。这些算法，看名字就知道，玄而又玄，不知所云。关键的是，这些算法都是受限模型，是基于某些特定条件下的模型，无法通用。就像冰岛的"啤酒和尿布"模型，到了中国，完全没戏，至少在沃尔玛、家乐福、华润等超市，没有看到这种模式。

对于大数据这种新产业而言，全世界都在摸索，政府做决策，必须进行调研和试点，而不是听过几个专家，尤其是某些协会的人员胡说几句，就作为国家战略操作。大数据产业，从概念到目前，不超过五年，因此试点是不存在的，以大数据作为核心战略，不要说国家，就是大企业，在全世界至今都没有一个成功的案例。

至于中国协会专家的意见，大家完全可以忽视，我的首部书籍，第二作者，现在就是中大的副院长，博导，可水平，也就哈哈而已。

关于大数据、高频交易和人工智能，个人的基本观点：凡是无法通过“足彩数据”进行实盘测试的方案、算法，都是在耍流氓。

大数据并非无用，可最多不过成立 3-5 家类似联想级别的公司即可，完全不是国家级项目，更别说国家战略级项目。作为国家战略，不管成功失败，我更担心是郑国渠效果。

郑国渠并非没用，时至今日，依然在造福国民。郑国渠,从战术讲是个成功的项目,耗费了秦国大量战略资源后，从战略讲，属于基础建设，反而增强了秦国的国力。相比郑国渠，大数据的的核心硬件服务器、软件、数据库，都要进口，而硬件服务器的折旧比汽车还快，最前沿的硬件，基本 3 年就基本价值归零，就是一堆废铁

因此，目前各地政府批量上马数据中心、计算中心，投资回报更加令人担心，一个 3-5 年，回报率无限归零的项目，而且投资总额分分钟过万亿。也许，大家会认为，这么多资金，上万亿砸下去，至少在人才方面会有收获，会培养一支自己的团队。这个，也许，不过意思不大。日本全民动员的第五代电脑，目前也有些国际上知名的 IT 项目：比特币、ruby 语言，可是对日本 IT 产业的整体盘，没有多少帮助。

希望，太阳能、风电等新能源方面的失败，能够让政府国家谨慎

因为在几个大数据群里，发现政府居然成为大数据的主力，有感而做，初稿未对郑国渠细细考究。理科生的坏习惯，不过不影响大局，谢谢几位指出的网友，不过这个是细节。希望大家多从主题方面展开讨论。

将大数据比做郑国渠，的确有些不恰当，至少郑国渠现在依然在造福国民，而大数据的投资，数年后，只是一堆废铁。于其中的团队，政府公务员，能够有什么人才，最好也不过是一群技术官僚，可能连技术两个字都称不上。

大数据项目，其实更接近日本九十年代的第五代电脑：人工智能计划至少，当年、和现在的富士，是极少数能够制造商业级小型机的企业，包括 CPU 这点，国内目前尚未这个级别的企业。天河系列，的确取得了不少成果，特别在巨型级的架构方面，不过，这个是不计成本的国家投入，商业化没有多少竞争力。 mit 的学生，当年用 ps 游戏机 cell 芯片，现在用 gpu 显卡，攒的计算集群，配合 linux，对于企业而言，性价比可能更高。把大数据和日本的第五代电脑对比下，大家会感觉更加贴切,不过，现在，谁知道小日本的这个东东？

大数据的通道是互联网，数据、信息是一次性消费产品，可以零成本传播、复制，互联网的核心只有两个字：free（免费）+open（开放)。积累的数据，一个连 pm2.5、耕地面积，都是国家机密的政府，再多的数据，缺乏流动与共享，也是死数据，有意义吗？

大数据与死数据

为什么，在得知贵州还开办了大数据交易中心，也许是全球第一个。会觉得无比别扭？？

4 月 30 日，一周后，黑天鹅又一次出现：《中国科学家难以获取高质量的国内数据科学》。

上海海事大学的 Zheng Wan 在《自然》上发表文章称，中国科学家越来越难以获得高质量的国内数据，认为这一情况可能阻碍科研和创新。他说，大部分公共数据被政府部门控制，其中一些加强了对数据的垄断，使得中国研究人员难以获取这些数据。人文科学的研究人员受影响最大，但数据访问的限制正扩大到环境科学和公共健康等领域，原因是数据具有政治敏感性。即使数据公开了，其质量也令人担忧，最明显的一个例子是全国的 GDP 数据和各省公布的 GDP 数据之间存在显著差距,国家统计局称数据差异是数据收集方法的不同导致的。在文章最后,Zheng Wan 谈论了互联网审查，称 Google 学术搜索被屏蔽对他的工作影响非常大。

大数据是互联网、后资讯时代的产物。而互联网的核心只有两个字：free（免费）+open（开放）。纽约的大学生，利用市政府的开放数据库，可以轻松制作出全市的犯罪热点分布图。而我们，就连专业科学家，都无法获得一手的数据，更何况商业应用了、BI 开发。也许，贵州的大数据交易中心，改为免费的、开源的数据共享中心，能够有一个华丽转身。

字王：大数据与黑天鹅算法 2.0

wiki 百科：“黑天鹅”隐喻那些意外事件：它们极为罕见，在通常的预期之外。如果一种理论、模型和算法，能够在一年内，捕获一只黑天鹅，无疑是成功的、科学的和实用的。如果在一个月内，那无疑是奇迹。如果一个月，甚至半个月，捕获不仅一只，而是两只黑天鹅。那只能是“神迹”了。

目前，zw 黑天鹅算法 2.0 正式 ok，首个基于 zw 黑天鹅算法 2.0 的足彩分析系统(C8)，Beta 版本也已完成，进入实盘测试阶段，速度比传统大数据方案快 50-100 倍。C8 系统，采用单台 i7 笔记本电脑，盈利概率（参数 v）高达 95%，居于业内领先水平，高于百度预测集群（集群规模>一百台），以及其他专业足彩分析平台。

人工智能永远差 500 年

1984 还早得很，人工智能连 0.1 版都没有，还属于黑暗期，前几天看到有个 MIT 的 AI 权威，也是怎么认为的，（sorry，链接没找到），反而是行外的人士信心无比，这个 AI 高潮，五六十年代，IBM-pc 刚出来时，闹过一段，现在，不过是历史再次重复吧，潮水，不过是一阵阵的，退潮只是时间早晚。

九十年代，日本最 High 的时候，也闹过一阵：第五代电脑、智能电脑时代。结果，是日本整个 IT 产业彻底崩盘，现在日本企业，自己都不好意思提。大数据的基础之一，数据分析，基本都是人工智能分析数据，人工智能连 0.1 版都没有，还属于黑暗期，大数据也还是黑暗期，将大数据产业化，纯粹是扯蛋。比较而言，目前的核能研究，反而靠谱的多，差距永远是 25 年，AI 人工智能，应该是永远差 500 年

为什么是永远差 500 年？个人认为，有两个基本因素：（一家之言，仅供参考）

- 人的智慧和进化，属于生物学上的突变现象，而电脑是严格程序化、机械化的，这个不是随机变量，能够简单解决的。
- AI 的母模板，采用谁，是电脑设计师本人，还是历史名人，希特勒和爱因斯坦的模型，肯定是不同的，发展的方向也是不同的。

当然，也可以采用混合型，大数据吗，找一百个、一千个、甚至十万个、一亿个人的性格，取中位数等等。不过到底采用多少个参数才行，这个是不靠谱的，100 个参数，还是一千个，甚至十万个、百万个。即使超算能够算的过来，这些参数的标准、选择也是问题。不要说“善”、“恶”，这种“蛋蛋鸡鸡”无法量化，涉及基本哲学的问题，就连最简单的胖瘦标准，都没法确定，多少为胖，多少为瘦，150 斤在中国是胖子，在欧美说不定可以当超模。而且，环肥燕瘦，到底是胖子取正数，还是瘦子取正数呢？

这个，貌似比五仁月饼更加难以判断？人脑都解决不了的问题，电脑如何程序化？不要说胖瘦问题是扯蛋？

大数据最成熟的领域，量化交易、高频交易，基础哲学就是混沌理论。而蝴蝶效应，是混沌理论的标志之一。真心不看好大数据产业化，吃伟哥都没用。大数据，从零开始，一下子成为国家级产业项目。大数据的硬件基础是 GPU、多核 cpu，国内非常不靠谱，目前连 486、z80 都做不出，前几天天河二号不是爆出 40%的负载事件.。软件，并行程序设计，全世界都刚起步，完全黑暗期。

大数据的核心是数据分析，基于人工智能、机器数据的智能化数据分析。说白了，就是 AI 人工智能技术，AI 这块，美国从六七十年代起步，日本九十年代纳入国家战略。至今，美国在人工智能方面的持续研究，超过了六十年，日本也超过了三十年，可是还是处于 v0.1 版，黑暗期。以国内目前的 IT 技术水准，零基础，想一下子超过别人几十年的研究，就是全体人员集体吃伟哥，也没有用，赶英超美，理想是好的，现实真心太残酷。不要说大数据、AI 人工智能，这些高大上的东西，就连最基本的 linux，100%开源的，有全部源程序，国内真心能够看懂核心模块的专家，不会超过一百人。就连带军方背景、实力最强的麒麟系统，不过也就是 ubuntu 的汉化版而已。新核能源，中国至少是全球五强，国内才 16 个（？？）试点。

大数据，要做，政府搞几家、甚至几十家企业就可以了，作为国家级项目，真心“亚历山大”

国内玩大数据，最好的是阿里淘宝，因为他们有源数据。余额宝的收益，早期，不过是因为阿里数据源不开放、其他基金嫌麻烦，再加上互联网企业贴本吆喝、花钱买客户的传统，表面收益貌似蛮高。等行业稳定下来，目前余额宝的收益，也和行业其他基金差不多，

至于所谓提前半年，一年，根据阿里大数据，布局股市，获得 70-80%的高额收益，这种案例纯是扯淡。从职业操守而言，不过是内幕交易，完全不需要大数据，哪些三线城市、乡政府的官员，根据规划局的预案，强行拆迁买卖房产，收益比这个高 N 倍，百度一下案例大把。

天河二号事件，虽然与技术无关，但反映出的问题是，相关产业链非常不靠谱。如果当年，还是邓小平时代，讲银河作为国家核心产业扶持，也许，天河这块可以与核弹并驾，其他产业可能还是一清二白。天河从银河开始，做了几十年，产业化还是这个水准，

大数据，从零开始，什么都没有，就作为国家级产业，凭什么？

关于大数据、人工智能，普通人可以看看这几部电视剧：

《机器之心》:高度人工智能的人形机器人警察；

《疑犯追踪》：超级电脑、网络监控系统和人工智能；

今年的电影《超能查派》，是南非《第九区》新锐导演的新作。

从另外一个角度解析了人工智能模板的选择问题：死狗和活狗。

在生存目前，无所谓善恶，你能说别人这种观点是错误的吗？凭什么？你难道就是是上帝？即使你真的是上帝？难道上帝比安拉更正确？更伟大？其实，政府部门，大家最熟悉的应该是《24 小时》。里面的 pda、智能手机与管理中心、卫星数据的实时对接、无缝集成、人物追踪，特别是实时性方面，可能很多目前都未完全实现。看影视的最大好处就是，相关术语影像化，而且结合了大量社会化应用场景，非专业用户，易于理解，专业人员，也可借助发散思维，举一反三。

大数据、趋势与黑天鹅

大数据的核心是关联算法，抓主流，分析趋势，一般取 95%的置信度。问题是，真正有价值的恰恰是哪些 5%，我们在实际分析时发现：黑天鹅才是新常态。

金融市场 大家都是大数据会反向干扰态势的。现在（2015） 黑天鹅才是新常态。

看看石油价格、瑞士法郎、日元升值、光大砸盘、黄金狂跌全部没节操！没下限！

07 年我就开始做舆情，而且采用的是智能语义分析模式，应该是国内最早的。基本是原创代码，后来检索资料，发现政府招标，并且有总参参与，就主动放弃了。也接触一些机构，包括广东省宣传部相关人员和深圳专业的舆情分析公司（类似香港的第三方民间评估机构）。因为这块太敏感，而且个人不喜欢与政府机构合作，政府部门往往多破坏，少建设后信息时代，创意经济，个人的主动性非常重要，甚至是第一位的，这个才是欧美目前真正的核心竞争力，国内政府必须认真解决这块，才能提升全体国民、企业的竞争力。

黑天鹅才是新常态

真正做大数据分析，和看报告是不同的，做研究，尽量使用第一手的资料和数据，转手越多，数据污染越严重。目前大数据用的比较成熟的有三块：互联网广告分析、机器翻译、量化投资。因为项目需要，早期我做过原创的 ocr 代码，这个图像分析、模式匹配、人工智能是基本功。早在 99 年，就开始用语句库、统计模式做英语翻译软件，比谷歌还早几年，素材是电影的双语字幕，当时就有百万级的语料库，国内同期的北师大等项目，不过几十万，

后来因为资源和课题发现，没做这块，附带出版了一套《魔鬼英语》教材，对于普通人而言，想把握目前大数据、人工智能的发展程度，看看百度、谷歌的中英翻译网页就可以，随便找段英文，机器翻译下，这个翻译水平，降低一个数量级，差不多就是当前大数据、人工智能的实际水平。

这几年，做量化投资方面的数据分析，越做越发现：黑天鹅才是新常态。为什么混沌理论，是量化投资的基础理论？因为市场是双向的，任何机构、个人，通过数据分析，进行决策，参与市场。对市场是会有干扰的，人少还好办，人一多，整个市场就乱套了。在所有的股票数据中，早期，macd 是比较科学的，也是非常有效的。当大家都用 macd 指标，作为投资参考，完蛋了，不是一只蝴蝶，而是所有人都成为了蝴蝶，整个市场数据，完全被污染，macd 也成为无效指标

所以说：人人都大数据，就人人都没数据。

现在的投行标配，全部是交易员自己写代码，将策略直接程序化，尽管如此，即使 100%保密，因为每家头行都以亿美元为起点，对市场影响也是超级“蝴蝶效应”，造成很多策略都是一次性的。

大数据与死数据

据说，汶川地震，药物管理问题，刺激了政府大数据战略，这个实际上，是有很大问题的

首先，这个模式类似 macd 指标，有效性，是建立在数据库封闭基础上的，只有政府和少数关联企业可以使用，普通企业、个人，没有权限使用这个数据库的。如果大家都能使用这个数据库，百度一下，分析汶川缺少板蓝根，大小老板、甚至个人投资者，全体板蓝根，几天后，汶川会成为全国、甚至地球上板蓝根密度最高的地区，这个“姜你军”要涨价,“蒜你狠”不折腾，已经有过案例。资本的力量是无法阻挡的，即使政府限制，关系企业，有关人士，也会拿到相关权限，这个毕竟只是商业数据，保密权限不可能很高，“SSS”级，和二炮一个级别。淘宝余额宝，也是一个类似的案例，早期，阿里数据源不开放、其他基金嫌麻烦，再加上互联网企业贴本吆喝、花钱买客户的传统，表面收益貌似蛮高。等行业稳定下来，目前余额宝的收益，也和行业其他基金差不多。

这种趋势，不过是价格二元化，在大数据行业的复制，与政府改革开放的出发点是相悖的

互联网的基础是：open（开放）+free（免费），基于互联网的大数据产业，如果违背这个基础，只能是空中楼阁。

这种管制模式的大数据产业，越发展，对整体经济损伤越大。首先，少数权贵部门和企业，从资本、原料等方面的垄断，会延伸到数据方面的垄断，获得不当利益，而广大普通企业、个人，却因为受限于数据，无法进行正确的商业决策、个人投资，社会的二元化分割更加严重。这个，看看现在的房屋数据库，始终无法进行全民查询。

这里多说一句，政府与其，梦想通过大数据，建立 2.0 版本的 1984 社会，不如管好全国四百个城市的局级以上官员，毕竟这个才几十万数量级。如果连几十万数量级的中高官员，而且绝大部分是党员，都无法有效管理，希望利用大数据，来管理十亿级的民众，只能是。。。。。。

其次，数据与资本、原料、设备不同，一个邮件，一张 U 盘，就可以将涉及全体国民的数据暴露给国外敌对机构。发达国家的模式是，除极少数敏感数据库外，普通数据基本免费开放，全民共享，这样才能全体国民受益，减少数据事故，减少数据意外事故，对普通企业、个人的冲击。

200 万亿数据只是小 case

政府主打的阿里健康，起点是汶川药品管理，数据库据说有 200 万亿条纪录。这个数据规模大吗？

实际上很少，药品数据库，不过是名称、价格、厂家等几十个字段，而且基本是结构化数据。1G 大约 10 亿直接，结构化数据，200 万亿，每条 50 字节，不过是 1000G（1T），1T 的硬盘，才 2-300 元。这个规模，比我们做 2000 年，做字模时少多了，国标 2 级是每套字库 6700 多个汉字，按 256x256 像素采样，每个汉字 128k（64k x 2）字节数据,一套字模差不多 700M（兆）字模的筛选率是百分之一，每套合格字模，需要处理 70G 的数据。

可能，黑天鹅算法最早的灵感和萌芽，就是不经意间源自这里。

2000 年,我们做“千禧版”版权登记，共一千套字体，数据总量超过 1000x70G=70T，是阿里健康的七十倍。当时没有超算，没有 GPU，我们是几台电脑，每天 24 小时运算，差不多半年才做完。其实，早在 92 年，我们 180 款的字模，数据量就差不多 20T，是阿里健康的二十倍。那时候 dvd 刚问世，刚开始只有视频 dvd，没有电脑的，我还特意去广州海印 xx 公司看过了 dvd 演示效果。

在大数据领域，200 万亿数据，只是小 case。吓唬外行有用，一线的，再多数据，不过是多几个索引表而已，而且现代 k-v 表，全部采用 hash 算法，与数据规模关系不大。比数据规模更重要的是，数据的实时性。可惜，这些因为公司利益，政策等原因，在国内目前基本无法操作，而国外，基于社会化数据的投资策略，已经出现 N 多模式。

大数据与数据干扰

政府大数据项目的一个“G 点”，是舆情监控。通过 QQ 纪录、微信关键词的确，可以进行一定程度的舆情监控，公共事件管理。但是，真正的破坏者，例如敌方，恐怖分子，异议者，可以用很低的成本干扰数据源。

《机器之心》里面，杀手采用口红大小喷剂，就可以屏蔽摄像头的人脸拍摄。其中提到，如何通过软件，提升关键词比重 10%-50%，这个很多 seo 教材都有。关键是，这个成本很低，稍微在网络下载一些 hack 教材，只需要一台笔记本，就可以控制成千上万台肉鸡，进行干扰信息发布。hack 违法，没关系，买套群发软件，买几台二手电脑（五百元的主机级 ok），一根网线，几千元，就可以搭建全部硬件，验证码，没关系，云打码，完全人工识别，准确率 99%。IP 限制，没关系，vpn 每个月十块钱，上千个 IP 地址，全世界都有。

文科生、易经与大数据

文科生与大数据

在网上和大家讨论大数据，有的人急眼了，说：理科生不懂趋势，看不到大数据是未来，开始玩概念、掉名词。凭什么说理科生看不懂趋势？

玩概念、掉名词，难道不知道互联网时代早 20 年，硅谷就有：magic word 之称吗？

零零散散，我也出版了二十多本书，比大部分文科生更文科生吧。我在网络上喜欢以理科生自居，是因为觉得文科生大多喜欢玩虚的，空谈无物，其实，我更喜欢做个读书人。

古代书生讲究六艺，格物骑射，都是基本功，相当于今天的理工类。网络时代，电脑是基础，文科生至少 office 应该过关。大数据，提 pandas、r 语言，gpu 超算、hadoop，有些欺负文科生，office 里面的 access 玩不转，不过至少要熟悉 Excel 数据分析模吧，知道画画图表、统计汇总这些吧。如果连这些都不懂，谈大数据、谈趋势，不纯纯是扯淡

易经与大数据

其实，殊途同归，什么东西到了极致，根源都是相通的。易经是纯文科的了，zw 小数据理论、“黑天鹅算法”，灵感就是来自：易经、阴阳、八卦。。。。。。

在大数据分析时，我们发现，所有的分析，抛开表象，以量化投资为例，到了最后，无非是两种选择：亏、赢

延伸一下，其他项目，也无非也是：

输、赢；正、负；胜、负；

涨、跌；加、减

男、女；老、少；黑、白；取、舍

这些，正好对应易经的：阴、阳。

计算机的核心是 cpu，cpu 的本质，不过是一个“加法器”，在 cpu 里面，加减其实都是加法，减不过是加上一个负数符号。这个加法器，正好对应了易经里面的：一生二：一（cpu 加法器）生二（加、减）。

易经里面的：二生三可以理解为，AB 两个对象，有三种状态，懂足球的都知道，球赛无外乎：胜、负、和；足彩就是：胜 3、和 1、负 0 三种模式。从投资较度而言，就是：盈利、保本、亏损，三种模式

八卦六爻：三生万物。A、B 两队、每队的三种状态，用数学表示，有八种变化（2^3，二的三次方），正好对应八卦。A、B 两队、每队的三种状态，对应上、下六爻，组成易经八卦卦象，64 种变化（2^6，二的六次方），正好对应易经六十四卦。易经八卦，八八重叠生六十四卦，包含宇宙万物。

小数据理论与六十四卦

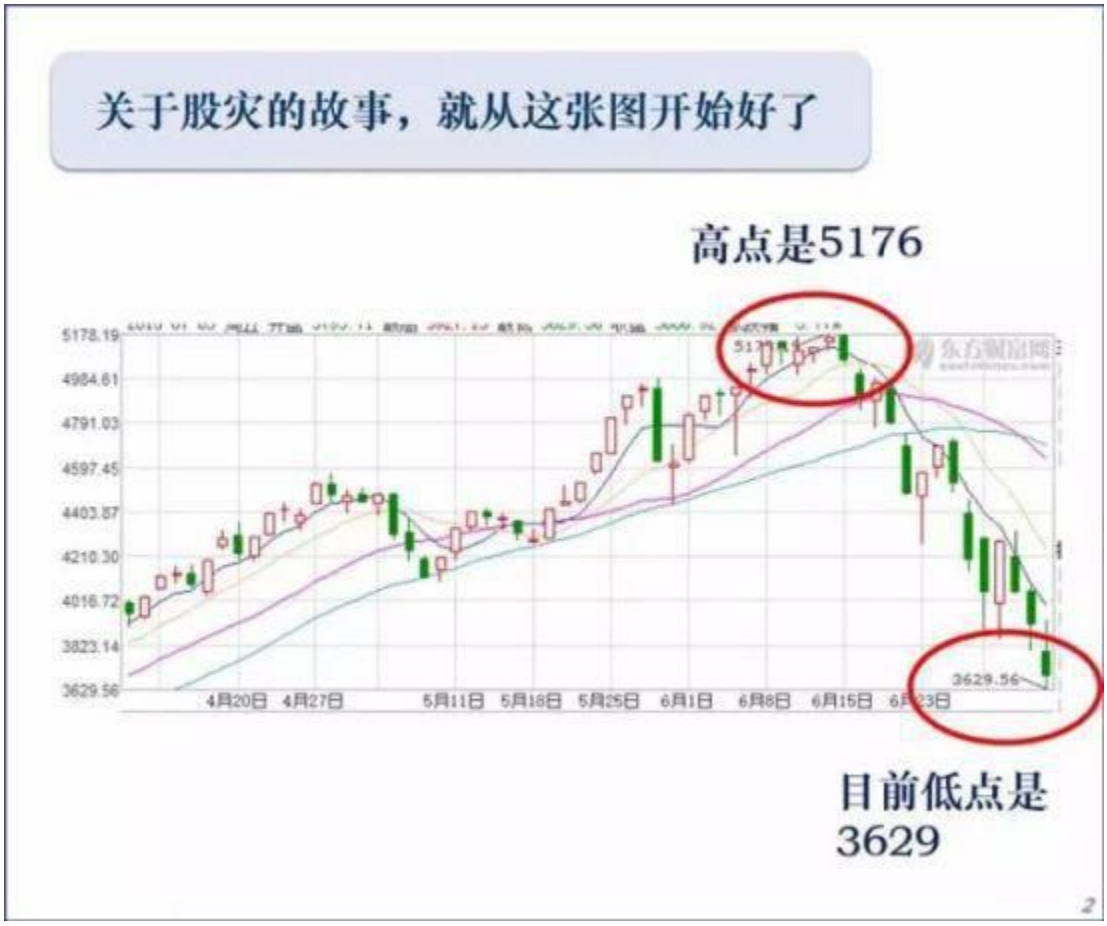
易经八卦，八八重叠生六十四卦，包含宇宙万物，why？

从数学、大数据分析的角度而言，是不是一个项目，有 64 个关键参数，就可以导出结论？这个也是目前需要大家论证的问题？

虽然暂时没有找到合适、严谨的数学理论支撑，不过我们由此受到启发，原创的：小数据理论，应用到“黑天鹅算法”实例分析当中，收到了非常好的效果。足彩，被误解的投资模式：年收益 269%的神奇投资术。269%的年收益，相信比目前 99%的大数据理论、投资公司更加靠谱。相对与其他大数据分析理论，zw 小数据理论、“黑天鹅算法”目前的一个重点就是，做减法，不断减少数据采样参数的数量。

股灾、马云、大数据

股市关头“七·七”之日，就差不多构思，因为事件敏感，一直压住没有发布。“七·七”股灾，是国家大数据战略发布后、也是本届政府最重大的事件，没有之一。涉及的领域，不仅仅是股市，以及经济、金融领域，而且将国家政治战略、社会战略甚至军事应对措施，暴露在敌对国家面前。



自六月起，得知国家大数据战略后，连续发布了二十余篇大数据的 blog，其中大多为负面。作为专业的一线 IT 业者，在大数据方面，还算 ok，不能像政府官员一样，只听忽悠。这里的 ok，说的是国内 TOP10，应该没有问题。

在大数据领域，zw 团队目前已有的：

- 自己原生的大数据理论体系：小数据理论，黑天鹅算法。
- 配套的大数据 SDK 平台：zwPython（大数据 dat 专版，正在进行 Beta 测试）
- 业内首部面向金融业者、交易员和白版用户的大数据+金融量化分析教材：《零起点，python 大数据与量化交易》，

虽然“七·七”股灾，相隔贵州全球首个大数据交易会（数博会，5 月 29 日），才一个月左右。但作为政府项目，有理由相信，国家的大数据战略，至少经过了 6-12 个月的压力测试。这种国之重器，如果不经严格压力

测试，没有 N 套灾变应对策略，就匆忙上线，所有全体相关政府官员，无论级别，都应该下台，自裁，移交司法机关。

对于大数据这种新产业而言，全世界都在摸索，政府做决策，必须进行调研和试点，而不是听过几个专家，尤其是某些协会的人员胡说几句，就作为国家战略操作。大数据产业，从概念到目前，不超过五年，因此试点是不存在的，以大数据作为核心战略，不要说国家，就是大企业，在全世界至今都没有一个成功的案例。

将大数据比做郑国渠，的确有些不恰当，至少郑国渠现在依然在造福国民，而大数据的投资，数年后，只是一堆废铁。至于其中的团队，政府公务员，能够有什么人才，最好也不过是一群技术官僚，可能连技术两个字都称不上。

“七·七”股灾前后，整个社会好像“三战”核弹爆发，经济、金融的“灭国”之战，国家经济有倒退十年的危险。连外访总理都匆忙回国，好像美国华尔街、联合欧洲、日本等全球资本，恶意做空中国，更加关键的是，这种错误的推断，引发了政府部门实质性的的救市行为，上万亿的资金被导入股市，好似当年三个代表，有关政府部门，近期言必“大数据”，这次股灾应对，必然会极大的参考大数据方面的资料。

可惜，政府主导的大数据，和其他政府项目一样，往往换来的是是十倍、百倍的失败。面对“七·七”股灾，政府种种应对措施，全面失败，而且，闹出了个世界金融史上的超级“大乌龙”事件，居然找错了“靶标”。“七·七”股灾的最终调查结果，目前虽然没有发布，但有消息称，虽然不一定正确：

前期，不过是江浙的一些土老板，为防止风险，做的空头对冲保险，因为配资杠杆，引发的技术性股市大幅度下调。

后期，则是因为程序化交易系统，对阀值 K 值的设定，引发的一系列自动抛盘，就像前几年光大乌龙事件，专业人员一听，就知道是因为交易系统，相关参数，未进行初始化设置，直接上线引发的自动抛盘也许，“七·七”股灾，根本就不关华尔街、美帝什么事？完全是躺枪。

事件后，损失惨重的投资人，有人质疑某些机构、个人，利用关系，可能可以获得恒生系统交易后台的所谓“大数据”，获得不当暴利。这个是必须的，某些政府官员，为了拆迁，就敢于不顾人命，透过交易系统，看看底牌，赚的钱毕竟要干净些，而且是千亿、万亿级的“大钱”。

资本的力量是无法阻挡的，即使政府限制，关系企业，有关人士，也会拿到相关权限，这个毕竟只是商业数据，保密权限不可能很高，“SSS”级，和二炮一个级别。少数权贵部门和企业，从资本、原料等方面的垄断，会延伸到数据方面的垄断，获得不当利益，而广大普通企业、个人，却因为受限于数据，无法进行正确的商业决策、个人投资，社会的二元化分割更加严重，这个，看看现在的房屋数据库，始终无法进行全民查询。

这里多说一句，政府与其，梦想通过大数据，建立 2.0 版本的 1984 社会，不如管好全国四百个城市的局级以上官员，毕竟这个才几十万数量级。如果连几十万数量级的中高官员，而且绝大部分是党员，都无法有效管理，希望利用大数据，来管理十亿级的民众，只能是。。。。。。

其次，数据与资本、原料、设备不同，一个邮件，一张 U 盘，就可以将涉及全体国民的数据暴露给国外敌对机构。发达国家的模式是，除极少数敏感数据库外，普通数据基本免费开放，全民共享，这样才能全体国民受益，减少数据事故，减少数据意外事故，对普通企业、个人的冲击。

至于所谓提前半年，一年，根据阿里大数据，布局股市，获得 70-80%的高额收益，这种案例纯是扯淡。从职业操守而言，不过是内幕交易，完全不需要大数据，哪些三线城市、乡政府的官员，根据规划局的预案，强行拆迁买卖房产，收益比这个高 N 倍，百度一下案例大把

这次股灾，如果当事人是一家企业，即使是“五百强”、高盛，十万亿的盘子，数千亿的亏损，分分钟倒闭。这也说明，政府的大数据战略，存在重大 bug，负责的话，政府大数据项目，应该在近期理性化，转交给几家专业机构操作，而不是全民大数据、大忽悠。

国家强力部门，公安部直接介入金融机构，这个可能比上万亿的救市资金，更加恶劣。大家不妨好好看看伦敦“金融城”，这一块被称为“一平方英里”(Square Mile)的地方，为什么要采用“国中之国”的运营模式。

虽然大伦敦统一的行政管理机构——大伦敦市政府，对包括伦敦城在内的每个郡都有约束力，但是伦敦城有自己的一套市政、警察和司法机构。重大庆典时，英国女王还要等候伦敦市长将一柄“市民宝剑”献给她以后，才能进城。

资本最重要的属性，就是安全。商人，特别是金融业，对枪杆子是最敏感的。不要说非洲、南美，这些动荡之地。希腊危机，就连欧盟的马甲都不好使。17%+利率的希腊债券，为什么不能无限印刷、发行。余额宝啦，据说就是阿里和半官方机构中信合作的产物。

马云旗下企业，特别是支付宝，是中国互联网事实上的隐形央行，作为政府电商、金融、大数据领域，最核心的技术企业。在这次股灾事件当中，也许涉及的环节、深度，比大家想象的要“深的多”。马云的恒生，作为事件操盘核心 IT 企业，涉及事件，这次，也许不是主观恶意。

孔老夫子，算是圣人了吧，中国五千年，也只出了一位，还留下了偷会“南子”小姐的野史。马云，毕竟只是商人，不是道德模范。

商人，对于利润的敏感，想必会刺激其他的“牛”云、“羊”云，组织团队，研究此次股灾。中国这么大一个盘子，居然被江浙的一小搓土豪，就差点引发经济、金融的“灭国”之战。有理由相信，美国华尔街的猎手、日本的经济学者、甚至五角大楼的专家。就在此时，就在此刻，会有不下十个，国际顶级的专业团队。从各个角度，犹如庖丁解牛，如外科手术般，在分析这次股灾的每一个细节，每一篇大 V 以上的 blog、新闻、甚至微信、帖子。从而制定更加专业的，可以操作的，股市、金融、经济、军事，“商业计划书”。也许，下一次股灾，才是华尔街专业猎手真正登场，正式引爆 xx 经济、金融的“灭国”之战。

大数据·实战个例“宏”分析

MBA 教育体系最成功之处，就在于导入了科学的个案分析。

二战最伟大的技术成功，不是原子弹、导弹、喷气机，而是流水线。流水线提供的生产力，比二战所有科技提高的总和还要高。至于“宏”，学过 c 语言的都知道宏定义、宏替换。本文不玩文字游戏，也不玩数字游戏，只是简简单单，对几个大数据实战个案，进行宏观的定性分析。

目前，大数据，和大数据分析的核心，人工智能，都处于 v0.1 的黑暗期，这个阶段，“宏”分析，可能比大量的数字堆砌更加重要。还记得量子物理学爆发前的原子轨道模型吗？当时，有几个人能够想象、理解原子轨道的跳变模型？还记得天圆地方、地球中心学吗？要不是哥伦布，“宏”分析一把，认为地球是个“球”，敢去环球探险？废话少说，言归正传。

本文“宏”分析，包括以下几个大数据案例：

1. 经典“啤酒+尿布”案例
2. 2015 中国股市“七·七”股灾
3. 国内首个大数据网络推广个案

1. 经典“啤酒+尿布”案例

“啤酒+尿布”案例，是最经典、最古老的大数据个案，其历史甚至比大数据这个名词更悠久。早在上个世纪，dbase 时代，数据仓库，数据分析，都用其做过案例。久而久之，“啤酒+尿布”案例，似乎成为了“神”一样的存在。好像三大几何原理，成为大数据的基本“公理”。不过，这个“神”，是“伪神”。"啤酒和尿布有什么关系", 这个十年前经典案例, 目前我是作为反面课件来说的这个是冰岛的一个数据分析结果，至少在中国不存在。

大数据，再多的专家，再 nb 的模型，再炫的 demo，也不如自己亲自去沃尔玛、家乐福、华润等超市亲眼看看，再回头问问这些大师们，"啤酒和尿布"模型，怎么玩砸了？

2. 2015 中国股市“七·七”股灾

对“七·七”股灾，进行“宏”分析，断定事件：“七·七”股灾，根本就不关华尔街、美帝什么事？幸运的是，对于“七·七”股灾的“宏”分析，及其推断，目前，已经证明是科学的、正确的。

“七·七”股灾前后，整个社会好像“三战”核弹爆发，经济、金融的“灭国”之战，国家经济有倒退十年的危险。连外访总理都匆忙回国，好像美国华尔街、联合欧洲、日本等全球资本，恶意做空中国。更加关键的是，这种错误的推断，引发了政府部门实质性的的救市行为，上万亿的资金被导入股市，好似当年三个代表，有关政府部门，近期言必“大数据”，这次股灾应对，必然会极大的参考大数据方面的资料。可惜，政府主导的大数据，和其

他政府项目一样，往往换来的是是十倍、百倍的失败。面对“七·七”股灾，政府种种应对措施，全面失败，而且，闹出了个世界金融史上的超级“大乌龙”事件，居然找错了“靶标”。

“七·七”股灾的最终调查结果，目前虽然没有发布，但有消息称，虽然不一定正确。前期，不过是江浙的一些土老板，为防止风险，做的空头对冲保险，因为配资杠杆，引发的技术性股市大幅度下调。后期，则是因为程序化交易系统，对阀值 K 值的设定，引发的一系列自动抛盘，就像前几年光大乌龙事件，专业人员一听，就知道是因为交易系统，相关参数，未进行初始化设置，直接上线引发的自动抛盘。也许，“七·七”股灾，根本就不关华尔街、美帝什么事？完全是躺枪。

3. 国内首个大数据网络推广个案

2004 年，我们利用 AI 人工智能和大数据分析技术，研发成功国内首个海量级社区营销软件：百万社区营销系统（软件著作权登记号：2005sr5133）。社区数据库超过一百万个，比同期类似产品，高两个数量级。

2008 年，依托百万社区营销系统，在北京联合创办国内首家 4A 级的专业网络公关公司：wowa 传媒，首年业绩突破一千万。

同年，“特仑苏” 危机公关案爆发，wowa 受中国国际公关协会委托，在北京，首度对国内大型公关公司，统一进行专业的网络公关培训，被协会誉为：中国网络公关事业的开拓者和启蒙者。

Wowa 服务过 150+国际 500 强客户；是微软公司首家官方认证的网络公关服务商；新华美通首选网络传媒合作伙伴；国内 TOP10 网络公关公司，50%采购过 wowa 的服务。

2007 年，操盘惠普笔记本“数码混搭”推广个案，成为年度十大公关行业经典案例（注意，非仅指网络公关）；被业界誉为：史上最强之网络推广案例，没有之一；不可逾越的概念营销“标杆之作”。

我们在 hp 笔记本“数码混搭”推广个案当中，首度提出的百度、谷歌搜索引擎：覆盖率指标，NLP 反向链接数，等参数，目前已经成为网络公关行业的基础指数。这个也是大数据技术，首度在网络推广方面的应用个案，我们当时的经验参数：1:1000。（百度 NLP 反向链接数-抽样采集率）。

按照客户要求，利用自行开发的 AI 智能语义分析系统，针对“笔记本电脑”，“数码混搭”两个主关键词，结合百度、搜狗的行业分类关键词，将发布主题帖，细分为数十组不同风格的软文，并在每篇软文前后，插入系统细分的关键词组。

硬件方面，我们采用了近百台 PC，组成了一个简单的发布集群系统，通过 1-2 周时间，围绕关键词：hp 笔记本、数码混搭，发布了过百万条网络推广软文。

最终，hp 笔记本“数码混搭”的网络推广方案，获得了“空前绝后满天飞”的成功，软文的存活率非常高。

“数码混搭”个案当中高至 70-80%的覆盖率，至今，国内外尚没有一家团队、公司能够超越，包括百度、谷歌自身。

史上最强网推个案，没有之一【ZW团队实战个案回顾】

ZW团队认为，互联网本质只有两个字

颠覆

- 2007年，ZW团队，操盘惠普笔记本“数码混搭”推广个案，成为年度十大公关行业经典案例（注意，非仅指网络公关）；
- 个案当中的百度、谷歌搜索引擎覆盖率指标，高到60-80%；
- 这个覆盖率，至今国内外，没有一家团队、公司能够超越，包括百度、谷歌本身。

当时，这个案例，震撼了整个公关行业，彻底颠覆了传统的营销、公关概念。

ZW团队也一举斩获150+国际500强用户，创业当年，业绩突破千万，打破了互联网企业，创业初期的烧钱模式。

2008年，受中国国际公关协会委托， ZW团队在北京对国内大型公关公司统一进行专业的网络公关培训，被协会誉为：**中国网络公关事业的开拓者和启蒙者。**

z世代，z传奇，zBrow

搜索引擎结果—Baidu(搜索”数码混搭”,每页85%的结果都是HP dv3000论坛主题贴)

沃华传媒

北京迁都,十万亿美元的大订单

1.北京的过去

明朝，最大的问题是什么？蒙古外患。所以有了：天子戍边。所以有了：北京。不管后人如何美化，没有明朝的宫斗，以及天子戍边的大义。自古以来，北京就是个鸟不拉屎,鸡不生蛋的地方。除了风沙，还是风沙。农业时代，完全没有任何自给能力。到了如今，工业化时代，更加痛苦地是，缺水，极度缺水。

2. It's the economy, stupid! （傻瓜，问题是经济!）

这是 1991 年,克林顿竞选时的名言。那么，今天，国内，甚至全球，最大的问题是什么？

“It’s the economy, stupid! ”傻瓜，问题是经济!.

今日的中国，已经是世界工厂，不缺制造能力。今日的中国，外汇傲视全球，不缺资金。今日的中国，虽然比国外，也许好一点，但经济乏力，是不争的事实。

既然，经济是今日中国最大的问题。那么，我们就来谈谈经济。2008 年，4 万亿人民币，才 5000 亿美金，就把中国的经济火车，高速推到 2015 年。

经济，就业，说来说去，根子，都是订单的问题。不管经济如何乏力，如果，有一笔：十万亿美元的超级大订单。 想必，中国经济的高速火车，再持续 10-20 年，完全没有问题。

3. 订单在哪里

那么，现在问题来了？到哪里去找一笔：十万亿美元的超级大订单？

高铁，是不错？俄罗斯莫喀高铁，泰国高铁，以及美国加州高铁，加起来的订单不到一千亿美元。一带一路，远景非常美好，可是，现在叙利亚打的一塌糊涂，投资，安全是第一位，亚投行也不知中国人一位大 boss。非洲，习老大，600 亿刚宣布，欧美就中国殖民论，就已经卷土重来了。

.....

想来想去，在中国军队的大炮，能够真正覆盖到天涯海角以前。投资，还是在国内好，对于资本，安全永远是第一位的。要烂，肉也要烂在自己的锅里面，毕竟是中国几代人的血汗钱。那么，现在的问题，还是老问题来了？到哪里去找一笔：十万亿美元的超级大订单？

迁都。迁都北京。

4. 北京的现在

首都经济圈，是一种经济聚集模式。巴黎、东京的首都经济圈，巴黎，东京，几乎占据了全国经济、入口的 30-50%。

可是，在中国，北京首都经济圈，京津冀一体化，完全是扯淡。 看看夜晚北京的卫星图，环绕北京的一个黑圈，只有环北京贫困带。周围城市的经济，入口、水电，甚至空气，全部被北京这个无底的黑洞给吸光了。

雾霭时节，珍惜生命，远离北京。在中国，要想启动首都经济圈，只有迁都。如今，中国经济，全球经济，低迷到近乎窒息的时刻。迁都，作为一个超级大项目：一笔十万亿美元的超级大订单？无疑，更是刻不容缓。

5. 未来的首都

至于迁都哪里，西安、武汉、重庆、喀什、长沙或者长株潭合并的毛泽东城，这些并不重要可以组织专门的团队进行调研。

关键要离开北京，找一个适合现代化的中国，适合工业化时代新中国的首都。至少，有足够的：水、电、空气和配套资源。适合未来、真正有活力的：中国首都经济圈

6. 十万亿的超级大订单

作为工业化国家的首都经济圈，例如巴黎、东京，一般占据国家入口的 30-50%，中国十三亿入口，实在太多了点，10%吧，1 亿人口的首都经济圈，还是蛮有弹性的。

一亿人口，按人均 10 平方米的建筑面积（包括办公、居住、商业），就是十亿平方米的超级大工程。首都的房价，按下限万元每平方，光房地产项目，这就是十万亿的超级大订单。按照房地产与 GDP 之间 1:10-20 的杠杆作用，最少也是百万亿人民币的 GDP，相当于 10-20 万亿美元。这还是最保守的估计，万元房价，不过现在北京的 20%。

如果思想开放一点，按如今北京房价的 50%估计，这个超级大订单的总额，分分钟突破一百万亿美元。这个帐很简单，也很好算。连小学生，都可以算的清清楚楚。有了这笔十万亿、一百万亿美元的超级大订单？中国的经济火车头，8%，10%，再飚上 10-20 年，绝对没有问题。

7. 北京人与上位者

至于有人说，北京人民不愿意？可能吗？连明朝，封建王朝的帝王，都知道，天子戍边的大义。北京人民的觉悟，是全中国最高的。难道真的为了一己之利，和全国人民做对。

资本的力量是无敌的。北京人，并非都是上位者，大部分不过是上位者的包衣。更何况，真正的上位者，离开北京在新的首都经济圈，依然是上位者。

ZW 点评：一个牛人的技术分析历程

原文地址：一个牛人的技术分析历程 作者：QuantWay

http://blog.sina.com.cn/s/blog_57a1cae80101aww9.html

第 1 阶段,是学习一些传统的技术分析指标，如 MACD，KDJ，RSI 等等。发现不确定性很大。

zw：其实 MACD，KDJ，RSI 这些，都是基本指标，也是真心有用的。但是，凡事都怕“但是”；但是，人人都大数据，就人人都没数据；市场是互动的，双向干扰，蝴蝶效应，混沌理论。

第 2 阶段，学习用飞狐编程序，下载个许多人编制的指标，还学习了 Vb，发现不确定性很大，虽然花样无穷，但本质上与传统的技术分析指标没有差别，都是建立在对历史数据简单的各种均线基础上而已。

zw：量化，不过是传统的技术分析，换了个新马甲，就像农民，现在都可以称为“生物工程师”。

第 3 阶段，追求更厉害的统计分析，学习了 SPASS，玩熟了时间序列分析 ARIMA。发现不确定性很大。原来 ARIMA 对白噪音的残差没有估计。

zw：靠，连德国人的 spass 都敢碰，真心佩服，不过，真正有用的算法、模型、架构是不会出现在货架上的。

第 4 阶段，学习 GARCH，该死的 SPASS 居然没有这个工具，只好学习 MATLAB7。GARCH 玩熟后，发现不确定性很大。原来，GARCH 本质上依然是线性估计，不过是将 ARIMA 的残差继续 ARIMA 了一次。晕倒。

第 5 阶段，被一些网络 N 人忽悠人工神经网络，开始玩 BP，RBF，发现不确定性很大。BP，RBF 对历史数据的拟合简直是完美，但对未来的泛化，简直是狗屁。仍然不死心，又捣鼓用遗传算法改进，用混沌理论的相空间改进，依然是狗屎。

zw：看来不只 zw 一人觉得神经网络之类不靠谱，一线实盘，大把吐槽；关键是不靠谱的神经网络，再用更加玄妙的：混沌理论来修正 真心胆大看来不知‘死’字是如何写得。

第 6 阶段，听南大的一个人工智能专家说，SVM 是目前最 NB 的，继续学习，这玩意很难，终于还是给搞定了，结果，发现不确定性很大。正确率让人失望。

zw：作者继续吐槽，学院派都是温室长大的，股市金融行业这么肮脏，进来后，100%水土不服

第 7 阶段，茫然之际，又有 N 人说，据说小波可能有用，找来书翻翻，感觉无比艰深。而此时对技术分析已经信心动摇。某日遇一朋友，实战高手，一席交谈演示，发现，靠，实战中还是传统的那几个老掉牙的指标最好，关键是是运用之妙了。

第 8 阶段，目前阶段，重新玩那传统的那几个老掉牙的指标。

zw：看来，最后，还是靠加减乘除，最简单的方式，往往是最好的，至少数据衰减少 n 个级别，这里面的为什么 why？可能价值几个博士学位

雅虎方面发表声明说，这整个科学计算是基于很基础的功能，只要你能忍受结果有一点点偏差，那么完全可以大幅度提升计算的速度。

《Yahoo 开源 Java 超快速计算算法 Data Sketches》

http://top.jobbole.com/31760/

大数据，why python

在《zwPython 3.0 初步规划》中，我们极大地强化了大数据功能，并作为首个 All-in-one 大数据分析平台。参见：http://blog.sina.com.cn/s/blog_7100d4220102vlpa.html

zwPython 3.0 目标：目前最强的集成式 Python 开发平台，大数据分析平台，没有之一：比 pythonXY 更加强大，内置全中文用户手册； 苹果“开箱即用”模式，绿色软件，解压即可，零配置。

首个 All-in-one 大数据分析平台：内置 pandas、Scala、R 语言、Q 语言、Quant、matlab、hadoop、spark 模块库和 API 接口支持（仅限 V3.0 版本）。超强功能:图像处理、AI 人工智能、机器学习、openCV 人像识别、gpu、openCL 并行超算开发、pygame 游戏设计。

为什么是 python，而不是 r 语言、Julia、matlab、Scala、Hadoop、Spark，等目前热门的解决方案。这个，主要是因为 python 发展太快，太猛，尤其是在 AI 人工智能、机器学习领域，已经超越 lisp，成为行业标准。而国内，因为中英语言、区域分隔等种种原因，通常要落后欧美 2-3 年。像大数据架构，目前欧美 IT 行业：“强烈推崇 Spark 技术，宣称 Spark 是大数据的未来，同时宣布了 Hadoop 的死刑”无他，因为 Spark 比 Hadoop 快一百倍。而国内，今天百度了一下大数据人才需求，90%还是：Hadoop

大数据的核心是数据分析，数据分析的核心是模式匹配、机器学习方面的算法模型。
简单但说，就是一个类似字符串匹配的算法，不过这个字符串是一个超长的字符串，可能超过 1000T 字节。

算法、模型，向来是 AI 人工智能理论方面的范畴，这个类似于量子理论物理学，和理论天文学。目前人工智能尚处于 0.1 版本阶段，大体上相当于哥白尼以前的天文学“地球中心说”、和量子物理以前的经典物理学阶段。因此，大数据、人工智能，基本上，就和理论物理学差不多，AI 的算法模型，99%都是靠理论推测，说白点，基本上靠“蒙”。

我说“蒙”大家可能不服气，这个却是老老实实，来自一线的实战经验，庆幸的是，国外的顶级 AI 学者的观点也差不多：

“对于（大数据、人工智能）这个词，我觉得最近社交网络上比较流行的那个笑话非常贴切，把大数据比作青少年性行为：每个人都在谈论它，没人知道到底怎么做，每个人都以为其他人知道怎么做，所以每个人都声称自己也在做。”

当然，这个 0.1 阶段，也已经能够解决 N 多实际问题了，例如目前的人脸识别、车牌识别、客户行为模式分析、网络广告点击分析、关联商品推荐等算法都比较成熟。

我谈大数据，特别是黑天鹅算法，更加强调我们提出的“小数据理论”,原因有以下两个：简单来说，国内除了进入“国际 500 强”的企业，例如阿里、百度、四大银行、移动等巨无霸企业；99%的企业，基本上并不需要大数据，这些企业所谓的大数据，其实只是最简单的数据库、数据分析。

简单做个大数据的量化门槛标准，可以分为以下两条：

1、企业的活跃用户规模超过一个亿。

2、每天的活跃用户更新数据的数据量，超过数据总量的 1%，换句话说，每天有过百万的活跃用户数据更新。

如果符合以上两条，可以导入真正的大数据平台:hadoop、spark,其他的，用 pandas、R 语言、matlab，或者其他传统数据库，可能效果更佳好。

例如淘宝、阿里的用户就完全符合以上两条，是典型的大数据企业。而中国民政部门，负责人口统计管理的信息中心，每个人的记录就那么几十条记录，例如：出生日期、籍贯、性别等，基本都是关系数据库表格可以高效处理的，即使数据库规模，超过十亿人，也无需采用什么大数据系统，一台 i7 的的笔记本，基本上就可以搞掂。

当然，这里的一亿用户、1%，日过百万活跃用户，也都是笔者根据一线实战，总结的经验参数。这些参数，不一定完全正确，但还是有过十年的专业经验做背书的。例如，笔者得知百度世界杯足球预测十八连胜，就断言，百度的大数据、人工智能算法不靠谱，里面绝对有大量的人工干扰。果然，不断两个月，百度的大数据图像识别，被 K 了，（百度在 ImageNet 图像识别测试中有违规行为）

我们强调“小数据理论”的第二个原因，以目前大数据应用最广泛的、最深入的量化交易为例。无论什么模型、算法，无论是 pc 集群、云计算、天河巨型机系统，归根到底，就是一个“涨”和“跌”的问题。简而言之，就是一个 1 与 0 的问题。这个，说来说去，又绕到了最基本的哲学问题，非三言两语能够说清。

事实上，对于大数据而言，比金融股票更好的数据算法对象，是足彩数据，因为足彩的结果是 3、1、0，胜、平、负，三种状态结果这三种状态结果，可以适用于所有的模型框架，暗合易经之道：一生二，二生三，三生万物。至于为什么，我也在研究。

关于大数据、高频交易和人工智能，个人的基本观点：凡是无法通过“足彩数据”实盘测试的方案、算法，都是在耍流氓。足彩数据是最透明的数据源，如果足彩不是 就没有更加公平的了博弈模型，如果这个都通不过其他都是扯蛋。所以说：足彩是最合适的数据源有历史数据 还有横向对比。其他任何数据源都没有这种实时的“矩阵”数据源 2014 年世界杯对于大数据人工智能，是个分水岭，是元年。微软、谷歌百度都有相关的项目胜率<50%。

AI 人工智能理论专家，和理论物理学家、理论天文学，大部分不是程序员（实验员），
因此，要求他们学习 c，可能需要等上一百年，还不一定靠谱。不过，“生命总是会找到自己的进化之路。”（侏罗纪公园）。

转来转去，AI 人工智能、大数据方面的理论专家,不约而同地找到了 python。同样的，目前量化投资领域，一线交易员必须自己 code，他们的选择也是 python。

“目前，量化投资、高频交易领域，一线操盘手自己编程，将投资策略直接程序化，已经成为国际大投行的标配。”

“在数据处理领域，特别在量化交易方面，python 已成为“统治级”编程语言。”

事实上，目前，python 已经是天文学、化学行业的标准编程语言。既然，这些地球上最聪明的家伙，都不约而同选择了 python，我们为什么不“跟进”呢？目前，python 在人工智能、机器学习方面积累的资源，可能比 c、r 语言、matlab 等加起来都要多，而且全部是“TOP one”级别的：scikit-learn、Orange、NLTK、MDP、PyBrain、BigML、PyML、Pattern 、Theano、Pylearn 等

r 语言虽然凭借统计背景，在早期的大数据、人工智能方面有些热。不过，到 2012、2013 年，涉及到深层的 AI 理论、算法模型时，r 语言就力不从心了。而此时，无论是厚积薄发的 scikit-learn、NLTK,还是 pandas、Theano、Pylearn 的异军突起，一下子,就把 r 语言上升的势头给掐死了，顺便把，matlab 给伤大了，就像加多宝 PK 王老吉，把和其正给灭了。

Hadoop、Spark 虽然都内置编程语言，特别是 spark，内置的 scala，完全 lisp 风格。lisp 近年因函数编程大热，事实上，lisp 和 prolog，也一直是 AI 人工智能的行业标准语言。lisp 的逆波兰语法虽然小众，不过作为 AI 行业的笔者，还是比较熟悉的，而且比起曾经用过的，100%纯逆波兰风格的 forth，语法要简单、传统 N 倍。尽管如此，笔者还是义无反顾，选择了 python。

因为，目前大数据、人工智能、机器学习都尚处于 0.1 版本阶段。这个阶段，需要的是，大量的建模、分析、测试。而 python，可能是地球上建模最快的编程语言，再加上，python 有这么多的数据分析、机器学习模块库，而且大部分是开源的 AI 行业，国际顶级的专家学者也如是说。

“基本上（机器学习）工具有两个推荐：Torch7（lua）、Theano + Pylearn2（python）”

python 最大的缺点是速度，一般比 c 慢十倍左右，不过大数据分析的瓶颈在 IO。目前，全内存计算是趋势，而且 intel 前几天发布的 xpoint，号称能够提高内存速度 1000 倍，基本上是 cpu 内部 cache 级别。（事实上，目前最前沿的高频、黑池交易软件，已经开始基于 cpu 的 cache 进行加速编程。）另外一个大杀器是，gpu 并行运算，无论是 cuda、opencl，2014 年，千元左右的 GPU，已经能够提速 3-500 倍，未来几年，2020 前，提速 3000-5000 倍，甚至上万倍，应该没有问题。具体到 python，虽然有衰减，不过目前，非官方的 GPU 模块库，提速 100-200 倍，已经完全 ok。

至于 gpu 并行运算的门槛，目前已经很低了，最简单的，只要在相关函数前，加一个 python 的修饰符"@jit"，就全自动加速。无需修改任何其他代码，至于超级复杂的 cpu、gpu 内存拷贝、交换，cl 异构运算语法、矢量编程，完全可以无视，比 matlab 还方便 pandas、scikit-learn 的 GPU 加速模块，也已经发布了多种版本。目前，python 与 c、fortran，已经是 cuda 官方认可的三大 gpu 并行编程语言。工业级的大数据分析，离开 gpu，即使是计算机集群，无论在投资产出、还是实时运算方面，完全就是扯淡。相比 c、fortran,无疑，python 要可爱的多，特别是“小白”般的理论学者。

python 号称：胶水语言，是目前唯一能够打通:pandas、Scala、R 语言、Q 语言、Quant、matlab、hadoop、spark。等目前、以及未来，各种大数据架构的平台。统一的开发环境、统一的数据分析平台，无论在前期的建模、测试，还是后期的数据分析、系统维护，在管理维护、培训研发成本方面，至少可以降低一个数量级。想象一下，

同时维护 windows、linux，甚至还有 ios、bsd，以及手机安卓 app，更何况，大数据往往还需要提供集群，gpu 异构运算支持。这些，仅仅是维护的硬件、软件名录清单，就可以把一个企业的 IT 部门主管，以及所有的工程师逼疯。

既然，python 如此美好？大数据， why python？应该说得通吧。

Python 可以称为大数据全栈式开发语言。因为 Python 在云基础设施，DevOps，大数据处理等领域都是炙手可热的语言。像只要会 JavaScript 就可以写出完整的 Web 应用，只要会 Python，就可以实现一个完整的大数据处理平台。

python、量化与“雅典娜”项目

大数据，量化，本质上是想相同的，都是数据分析。

Q 群里面有朋友问道：

老大，我们老师说量化投资用 python 最好，但是现在互联网上关于 python 搞量化并愿意分享经验的就您一家，而且还没正式开始，也找不到一本关于量化投资的书籍，根本无从下手。

但是 R 语言搞量化已经很成熟了，学习起来也有很多途径。关于 R 和 Python 搞量化的区别，我想听听您的观点。我的打算是先学习 R 语言搞量化投资，用 R 语言打好量化投资的基础。之后在转到 Python 来，这样不会毫无目标的学习，这个也想听听您的看法。谢谢老大。

zw 因为以 IT 和大数据分析为主，对于国内金融行业一线，的确有些疏远。国内 IT 行业，python 的趋势已经起来，金融行业，相对保守，百度一下，的确资源很少。所以，zw 在 Q 群里面，这样答复到：python 量化，也许，暂时在网上，只能找到 zw 一家或者极少的量化资料，但绝对不只这么写，大家可以搜索关键词：pandas 量化 pandas（潘达思），熊猫数据分析软件，作者本身就金融公司出来的。另外，不要太纠结于量化，量化的本质是数据分析。zw 讲课的重点也是数据分析。商场如战场，特别是如何透过（战争）迷雾,看清真正的数据，排除干扰数据。

这两年，大数据，人工智能、量化高频领域，python 已经火烧连营：matlab，R 完败。matlab，R 辛苦多年，对用户、市场进行的开拓、启蒙工作，都被 python 一家独吞。特别是 matlab，起了个大早，一觉回到解放前。

没办法 数据分析是个综合工程:人工智能 语意分析，环节众多。数据分析，其实只是：其中最简单的一个环节。做综合项目，py 真心没对手强大的没有朋友。

国内资料虽然少，还是有的，英文好的话建议直接看国外资料。国内一般滞后 2-3 年，这块应该 1 年吧。其实把 pandas 官网资料 http://pandas.pydata.org/仔细看看，还有官网的衍生项目，友情链接好好看看。特别是网站的论坛社区，追新，就只能这样了。

zwPython 升级时，就是这样，每周关注 opencv3 winpython 的最新动态。折腾了半年，总算都升级了。关于，zw 对 python 的选择，曾经做过非常多的准备工作，起步于 2012 年。

2013 年后，python、pandas 在 IT 领域的异军突起，特别是：大数据、人工智能方面，原因很多，这里就不细谈了。而金融领域，则是因为：“雅典娜”项目：

不久以前，在金融行业，Python 作为一种编程语言和平台技术还被视为异端。相比之下，2014 年有许多大型金融机构—如美国银行、美林证券的“石英”项目或者摩根大通的“雅典娜”项目—战略性地使用了 Python 和其他既定的技术，构建、改进和维护其核心 IT 系统。

众多大大小小的对冲基金也大量使用 Python 的功能，进行高效的金融应用程序开发和金融分析工作。同样，当今许多金融工程硕士课程（或者授予类似学位的课程）也使用 Python 作为核心语言之一，教授计量金融理论与可执行计算机代码之间的转换方法。针对金融专业人士的教育项目和培训也越来越多地在课程中加入 Python。有些课程将它作为主要实现语言。摘自《Python 金融大数据分析》，总计 528 页，2015 年 12 月，人民邮电出版社

以下：摘自字王前几年的 blog，关于 matlab、python、R 的抉择。

字王新一代智能字模 sdk，准备采用 1-2k 像素点阵的字模，数据处理量是 256 点阵的 10-100 倍，目前正在做前期规划，大体是以下三大框架：

- matlab，无疑是首选，支持 cuda 加速，处理速度可以提升 10-200 倍，可以达到超算中心小型机的速度，可惜软件太大，2013 的新版本，2 张 DVD，安装耗时，而且正版价格太贵。
- r 语言,是如今大数据的热选，是统计行业、精算师的 photoshop，今年最小了的 V3 版，才 50M，而且是免费开源软件，图像支持也不错，正在考察中
- python 无所不在，也是开源软件，而且有 pythonXY，专业的 python 图像、数据处理计算整合包，开源的字库设计软件 fontforge，主体模块，也是采用 python 编写的。问题是。python 目前正处于 2.0 到 3.0 版本的大变迁时代，V3 版 python 的语法虽然变动不多，但部分语句不兼容 v2，是个大问题，pythonXY 目前貌似采用的还是 2.x 版本，因为 pythonXY 集成的第三方数学模块太多了，全部升级到 3.0 版，估计至少是要 3-5 年。

目前正在恶补统计学基础，推荐几本漫画图解版本的，对于非统计专业的技术人员，设计师，无疑是大补。字库设计与统计学，表面看毫无关系。

其实不然，数字化后的字模，是 100%的纯数据处理，而数据处理，正是统计学的领域，也是 r 语言的拿手好戏。以前字王的字模 sdk，关于字型轮廓的间距模块，只涉及到：最短路径，绝对距离两种类型算法。这次看过 r 语言，才发现，统计行业的间距模块，种类甚多，而且都有现成的模块：基本距离、绝对距离、曼哈顿距离、欧氏距离、明氏距离、马氏距离、二值定性距离。

zw 足彩·大事件

大家知道，主流学者，到现在都看不起足彩，认为是赌博，这个我们不讨论。

虽然，最近国家级别的彩票大数据研究中心，也已经正式成立了，资本的力量是无穷的。下面，我跟大家介绍下自己的足彩研究历程，因为太长了，只能以大事件的方式，节选下：

- 大家都知道，zw 人虽然马马虎虎，却做过不少很 NB 的项目，象现在的：zw 黑天鹅足彩算法，就非常 NB，可以说行业内算得上 TOP10 了。
- 这个黑天鹅项目，最早源于 zw 的验证码项目，当时叫：z-SPO：z 粒子算法，2012 年的博客有提过。
- 在做验证码识别项目时，我想，如果事先，新浪、sohu、网易、天涯这些网站分下类，识别角度可能更高。因为网站、论坛很多，几十万个，不可能人工分类，就写了个自动分类算法。算法其实很简单，就是“二选一”模式，首先：是新浪的验证码，类型标记为新浪，不是的标记为：其他。然后，再对其他，进行二次分类，是：sohu 的，标记为搜狐，不是的标记为：其他。这样，几个循环下来，自动分类就完成了。令我吃惊的是，这个程序很简单吗，准确度却相当高，差不多 90%，可以完成几百种不同风格验证码自动分类。
- 因为，是“二选一”模式，很容易就联想到股市的：涨与跌。这方面，大家可以我的博客《文科生、易经与大数据》，这里就不展开说了。于是，就下载了历年的股票数据，当时是 2010，从 2004-到 2010 都有，这些数据现在还在网盘里面，不过格式忘记了，要看看源程序。
- 股市整了一年，各种公式、指标、算法，差不多都测试过，没一个靠谱的。
- 因为我做项目，喜欢快些相关的外围书，特别是相关的历史书，在整股市数据的时候，发现，股市和足彩

的发明人，都是同一个英国贵族。而且，通过维基百科，知道了必发交易所获过商业创新大奖，知道了高盛足彩套利的故事。

- 到 2012 年，朋友叫我去深圳做项目，在高丽，深圳大学城附近，到了新地方，我喜欢一个人在四周溜达溜达。结果，发现，高丽地铁站十分钟路程内，至少有三十家彩票投注站，比药店、银行还要多，仅次于手机店，要知道的，深圳的门面租金，差不多每个月一万。
- 于是，再上网溜达，发现国内正规的彩票网站，差不多上百家，而且大部分都有上市公司背景。在一看现金流，立马坐不住了。基本上都是十亿、百亿级别，很少有千万级别的。（2015 年，深圳查封一家足彩网站，一年的流水就是上千亿）。
- 我做过很多互联网公司的方案，知道，很多互联网公司，看起来风光，可实际上就几千万的盘子，几百万的现金在周转。
- 再回想到高盛、必发的故事，于是，开始研究足彩。
- 起步，很痛苦，因为我从不看足球，现在也是，更不用说 3、1、0，胜、平，负了，还有什么大球、小球，上盘、下盘。
- 刚开始，只是看，可是不掏钱，很多细节老是搞混，好在足彩盘子小，才 2 元。刚开始，自己乱投，和琪琪小姐差不多，他是看名称，我是看赔率，看数字，看那个数字顺眼，就投哪个。
- 交了不少学费后，开始知道有合买了，而且运气不错，跟单开始，连续几个 4-5 倍的胜盘，不过一个月没有，就全亏了，于是换人合买，换了 n 个，还是不靠谱。
- 这期间，交不少学费，基本上，每个环节都交过，因为是真金白银，掏的学费，虽然不多，但印象都在。所以，我跟学员说，一定要做实盘练习，一定要注意细节。没有实盘，量化训练没有任何意义，有了实盘，细节上吃几次亏，就会注意了，这个总比正式投资出问题好。
- 当然，这时候，对于足彩的数据分析也没停。足彩也有好多所谓的必胜公式，套了 n 个，不靠谱，再用经典的教科书算法，机器学习，数据挖掘，还是不靠谱。
- 最终，转来转去，还是 zw 自己的算法最靠谱。个人认为，不是 zw 的算法靠谱，可能是，一方面，靠谱的算法，一般机构不会公开；第二，公开的算法，大部分是限制模型，是有 n 多前提的；部分没有前提的，准确度又不行。
- 2014,2015，大数据风起云涌，zw 跟风，一激动，就把 zwPython 给开源了
- 于是，就有了出版社约稿，《零起点，python 大数据与量化交易》目录提纲、zw 魔鬼训练营，等一系列后面的事件。

熔断、公开课，黑天鹅，缠中说禅

熔断，缠中说禅、黑天鹅。没想到，转眼之间，一只黑天鹅落到了公开课上。随便说一句，凤凰卫视，2015年财经总结，标题里面也有黑天鹅。

元旦后这几天，zw忙于编辑公开课的同时，中国首度股市推出了“熔断”机制，差点又弄成了、“七七股灾”的翻版。完成公开课后，相对松懈点，浏览网站，发现，面对这次股市“熔断”事件，损失惨重的股民想起了中国第一操盘手，李彪先生的《缠中说禅》：

炒股票，大家一定要看缠中说禅的理论。虽然李彪先生已经不在了，但是作为中国第一操盘，至今无人可以比肩。亏钱的时候，能静下心看看书么？网上有，免费的。这是国内最好的教程，没有之一。

既然中国第一操盘手，李彪先生的心态修炼，也是主打：禅学，主打，平常心。

zw公开课里面，强调心态训炼，强调平常心的养成，看来也是符合股市的“大道”。无处不在的微操作。年轻人，喜欢效率，喜欢微操作，这个是正常的，也是合理的。不过，量化毕竟是量化，再怎么“宏”分析，其中的量化分析、微操作，也是不会少的。

在zw公开课里面，《足彩vs股票·数据源》，就是经典的量化分析手段，通过三个阶段，相关的数字，进行对比分析。

并非，只有财务数据，才是量化分析。在 zw 公开课《八里桥之战》，zw 也是通过具体数字，表明技术代差的后果：交战双方的部队数量、死亡人数、受伤人数等等。同样，在公开课里面，“如何量化操盘手”、“1w 小时黄金定律”等多个环节，zw 都是通过具体的数字，来表明相关的观点。

跳出量化做量化。zw 曾经说过：要跳出 IT 做 IT。同样，做量化，也需要跳出量化做量化。

在 zw 博客《字王看：大数据观点补充》里面，曾经提过：

人们喜欢谈论各种大数据，各种调查数据。但不要忘了，不只有数字信息才是数据。

不要忘了，一句引语也是数据，一个人讲给你的故事也是数据，某个人回答问题是低头看鞋，那也是数据，他抬了一下眉毛，那也是数据。你要收集所有这些数据。

-----《弗里德曼：你们感受到颠簸，我们看到的是上升》

托马斯·弗里德曼，《世界是平的》作者，美国知名时政评论家

并非只有具体量化的数字，才是数据，才是大数据。同样，并非，只有电脑，股市价格，才是量化分析。年轻人，喜欢效率，喜欢微操作。其实，许多真正的微操，并非只是分析数据。在 zw 公开课，“zw 足彩·大事件”环节，是这样说的：

- 2012 年，朋友叫我去深圳做项目，在高丽，深圳大学城附近。到了新地方，我喜欢一个人四周溜达。结果发现高丽地铁站十分钟路程内，至少有三十家彩票投注站，比药店，银行还要多，仅次于手机店，要知道的深圳的门面租金差不多每个月一万。
- 上网一查，大吃一惊。发现国内正规的彩票网站差不多上百家，而且大部分都是上市公司背景。

- 再一看现金流，立马坐不住了，基本上都是十亿、百亿级别，很少有千万级别的。（2015 年，深圳查封一家彩票网站，一年的流水就是上千亿）。
- 我做过很多互联网公司的方案，知道很多互联网公司，看起来风光。可实际上，就几千万的盘子，几百万的现金在周转。
- 再回想到高盛、必发的故事，于是开始研究足彩。

公开课视频里面，更是特意增加了相关的彩票店门面的画面。其实，“zw 足彩·大事件”环节，前面的验证码部分、后面的 zw 足彩历程，本身也是个量化分析的过程。

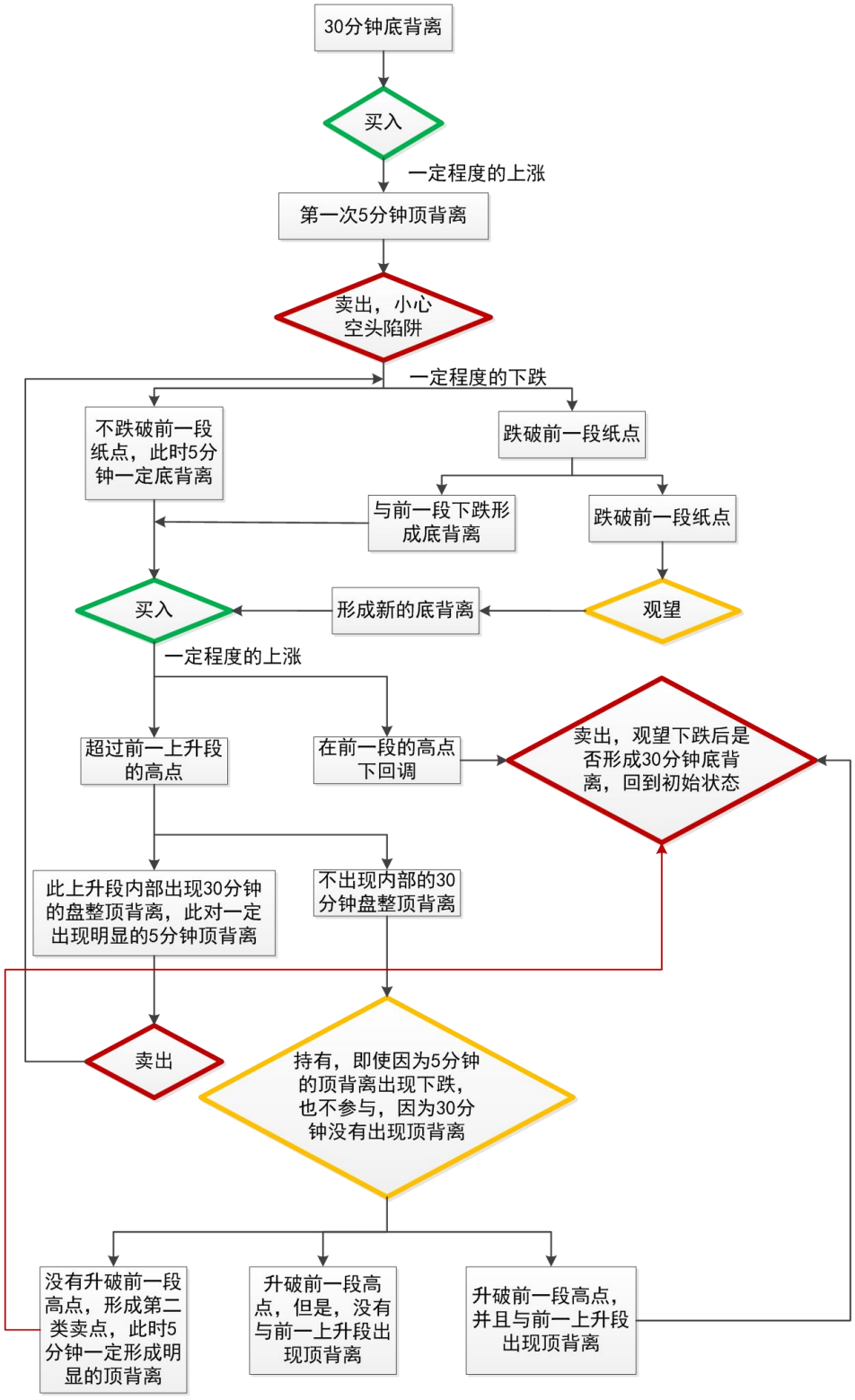
zw 说这个，大家可能认为不够权威。下面，我们再说一个，相似度高达 90%的、华尔街的案例，ps，又是一只黑天鹅。电影《大空头》，里面的主角，好像就是，“末日博士”麦嘉华(Marc Faber)。这个案例，虽然源自电影《大空头》，不过应该是源自事实。

电影《华尔街》里面，也有类似的环节，操盘手晚上，去投资对象公司：偷垃圾、偷报表。

电影是 2008 年金融危机爆发的前夕，几位华尔街大投行的投资经理、操盘手，得知有人做空的消息后，并没有只看数据。而是，亲自，去了几个最底层的放贷机构，画面上，正式这个镜头。同时，还亲眼看到，许多因为无法还贷，空无一人的豪宅，有个镜头中，游泳池里面还有宠物鳄鱼。当然，也有普通中产、蓝领的家庭。这几位华尔街经理的现场微操作，是不是和 zw 在公开课里面的环节，过于相似了

（zw 发现）高丽地铁站十分钟路程内，至少有三十家彩票投注站，比药店、银行还要多，仅次于手机店，要知道的，深圳的门面租金，差不多每个月一万。信息时代，尽管资讯无比发达。Face To Face（面对面）。依然是华尔街，最喜欢的商业模式。也许，这些年轻的学员，什么时候，知道：做量化，也需要跳出量化做量化。可能，就真正成熟了。

缠中说禅&量化交易



《缠中说禅》和李彪先生，都是公开课后才第一次接触。

不过，看百度百科，李彪先生的裁判流程图，好像是做当日交易超级短线的。里面不少是 5 分钟数据。整个流程，与量化分析的流程非常类似。年后有时间，看看能不能用 python，做个类似的交易策略模型

附录

没的选择时，存在就是合理的

- - 与李旭科书法字 QQ 聊天记录

2015-8-11，晚上与李旭科书法字作者，在 Q 上聊了下，有些资料涉及到字库设计、字库产业，对大家也有益处按惯例没细整理，直接发 blog 了。

9.11 靠，今天是 911，早上查资料，在 fonts.com 发现了这款英文，字王模板库里面，也有 n 多（n>100）这类风格的中文字库，可是嫌太丑，一直没敢发布，看来心态还是不够 open，再一次证明，存在就是合理的。

http://www.fonts.com/search/all-fonts?SearchType=WebFonts&page=1&itemsPerPage=10#languages=W44&page=1

原文:《FontForge 常见问题 FAQ》字王翻译版，http://fontforge.github.io/en-US/faq/。

有趣的是，这个 FAQ 的第一个问题就是，为什么需要新字体，经常有人问我这个错误的问题，你不会去问一位画家，他是不是已经有足够的水彩。英文毕竟十万种了，设计简单，面临的压力，是创新困难，与字王 blog《没的选择时，存在就是合理的》遥想呼应。

http://blog.sina.com.cn/s/blog_7100d4220102vqqx.html

SL** ： 请问中华大字库是你的作品吧？
字王 ： 是啊 。
SL** ： 现在网上到处都能下载到，应该是有版权的吧。
字王 ： 《字王 20 年》，http://blog.sina.com.cn/s/blog_7100d4220102vqcf.html。
字王 ： 做过版权登记，不过国内字体版权保护，目前争议很大，国外也是，我是按美术作品保护的。
SL** ： 我看过你的博客，我也认识梁老师。
字王 ： 没指望用这个赚钱，是个学术项目。
SL** ： 但你的这个应该是软件吧。
SL** ： 恩，这个特别实用对制作字体。
字王 ： 不是软件，是字模。软件是字王智能字模开发平台 zw-sdk，现在升级版是 zwPython。
字王 ： x2ttf。你说的是这个，涂鸦造字是个公益软件，全功能，全免费的，2012 的作品。
SL** ： 恩。
字王 ： 你的字是自己写的？ 有没有个人网站 blog。
SL** ： 恩。 没有，之前有一个是请人做的， 但人找不到了所以网站关闭了。
SL** ： 请教一下字模主要是干什么的。
字王 ： 开个 blog，字王网站在升级，暂时也关闭了，目前主要是更新 blog。

字王 ： 字模就是字型。黑体，宋体等都是。字王主要做个性化字体 应该是鼻祖了。

SL** ： 但的确字体盗版侵权太厉害，我的字曾经用在舌尖上的中国第二季，但并没有经过我的同意。

SL** ： 那做这个字模有收入吗？现在这个情况。

字王 ： 所有字库这块我是慢慢推，有时间就整点。不过字王云字库，会是全免费的，互联网的核心就是 free+open。

字王 ： 我做字库这块 95 年做中华大字库，当时就投入三十多万 。光碟开模费就几万，字库项目一直没赚钱。不过，这块对于我的技术、课题提升很快。目前我做互联网，网络营销 大数据，核心技术都是字王的智能字模，底层模块都是 AI 人工智能。

SL** ： 哦。具体我还是有点不太懂，对我有点太专业。

字王 ： 简单说，在中文字模这块，字王是全世界做的最好的。有了这样一个高度，看问题很发现其他行业也不过如此。08 年我做网络公关，以前从来没有做过公关。一年，成为行业老大，又是一个第一。（高度决定一起，在学术界，字库行业，至少目前没有人，敢在字王目前，称王称霸。）

SL** ： 那么现在你还在坚持做这个吗。

字王 ： 在做。github 中文云字库，字王又是第一。目前 zwPython 在升级 3.0 的。

SL** ： 云字库以前听过，但这方面市场感觉还是不够理想吧？

字王 ： 这个不需要市场，中文字体是种稀缺资源，目前全世界的市场都处于饱和超饱和状态。手机 app，免费下载，推广费都要 5 元一个。云字库，至少字王的云字库不担心市场，用户至少一个亿的用户没问题。（目前排名前三的手机输入法，市场估值都超过一亿，字库的技术含量，特别是云字库，比输入法高至少 2-3 个数量级，而且字库版权，天然具有排他性、垄断性，按目前移动互联网的 IP 模式，市场估值，至少是数百亿。不过云字库种类的起点，至少是 1K 中文字库，所以。。。。）不过启动费用也不低。所以，现在字王一方面在找钱，一方面在攒钱。

SL** ： 字王的云字库包括很多种字体吧。

字王 ： 字王独家拥有超过一千款中文个性化字库毛胚和版权，比全球其他所有字库公司加起来都多。

SL** ： 这么多。全是你和你的团队开发的，还是有合作的？

字王 ： 我们自己的。字王是通过 AI 智能技术，电脑合成，人工筛选。

SL** ： 不过即使这样也应该特别费时费工。

字王 ： 是啊，字王 20 年了。

SL** ： 对了，我请教你一个具体问题。

SL** ： x2ttf，自动生成的字体是开源免费的吗？

字王 ： 准开源，源程序。因为涉及字王的底层代码没开放。其实看源码不如看 fontforge、fonttool 的，很多都是 python。看起来更加轻松。

SL** ： 哦。 x2ttf 可以制作繁体吗？

字王 ： 没问题。不过内码必须是 GB。

SL** ： 现在你有和这些字体公司合作吗？

SL** ： 电视台，影视公司，还有现在的移动设备中大量用到字体，但平时也没有看到过字王的字体。

字王 ： 没有，谈不拢，字库公司都是传统的软件公司，很古板。我是做互联网的，操作方式不同。上海 xx 网的，老总算是年轻的，聊过 n 次 是谈不拢，思维方式不同。

SL** ： 恩。就是，我准备开发的一套字体，他们嫌笔画不匀称。本来是毛笔书法， 能看出来还是太过保守。

字王 ： 字王的最终字体产品很少，我是做企业的 。国内字体不保护，没办法，一套字体毛胚，人工修正的费用至少是五千行业平均是 5 万现在应该不低于一万。不过很多字体都是在字王的毛胚上衍生的。

SL** ： 哦。现在情况有所好转，汉仪听新闻上说最近几年开始盈利了。

字王 ： 挺好看的，我对这块不讲究，以前大家都说字王的字不好看。现在字王揭开一个日本的盗版者：日本三次元刻绘字，盗用字王 95 年的拙体。大家应该会闭嘴 好好反思一下了。

SL** ： 恩，懂字，会挖掘价值的人还是不多。

字王 ： 英文十年前就超过五万，这个只是当年 coreldraw 的配套资源光盘上面的字体，现在至少十万英文字库。目前全球的中文字库去掉简繁粗细，总和不超过一百款。所以目前是（中文字库）走量的时期，在 1 万款以前所有中文字库都是：存在就是合理的。大家没的选。（目前，大家对于中文字库，没得选择，属于短缺经济时代）

SL** ： 不过好的中文字体开发的难度就是大点。

SL** ： 字数太多了。

字王 ： 所以说字王才伟大。 ：）

字王 ： 而字王云字库，基于人工智能，提出的“智能字模”技术，是目前唯一可以突破千款瓶颈的中文字库解决方案。

SL** ： 呵呵　能不能理解成吧汉字按照规律拆分组合快速生产的一项技术？

字王 ： 有这个设想，希望能够在 100-300 个汉字内，完成所有字库部件。这个我 2012 年的 blog 就说过。目前 zwPython 里面的 opencv 就有图像识别模块，不过，黑体和草体的模块库肯定不同，可能需要开发数十甚至几百种模板，目前没时间，等机会吧。

SL** ： 美术字的可以，但书法字的这样造出来的字感觉缺点东西 。

字王 ： 300 字应该可以，覆盖一种字体了，书法字也差不多。

SL** ： 那么现在 zwPython 主要有哪些功能？

字王 ： 你自己下载吧，有中文手册。

字王 20 年

95 年，字王发布了《中华大字库》，不经意间，二十年过去了。

20 年来，一直有人问我，有上市公司的主席，也有获过国际大奖的专家学者：

“你做这个干什么？”“盗版这么多。。。”“字库又不赚钱？”。。。。。

Why?

最早，是因为在书上看到，”汉字是中国五千年唯一保留的传统文化。”网络时代，更加神奇：“中国远远早于圣经写作年份而发明的汉字里隐藏着与圣经记载相同的信息，这超出了人类的智慧。”

“中国方块字隐含着与圣经记载相同的上帝的信息、上帝的思想、上帝的教诲，是上帝的密码。”

“汉字，是目前世界上唯一保留着上帝真道教诲的文字。”

字王、字库与大数据、量化交易的关系，大家可以浏览《大数据、趋势与黑天鹅》

http://blog.sina.com.cn/s/blog_7100d4220102vn8s.html

200 万亿数据只是小 case

国标 2 级是每套字库 6700 多个汉字，按 256x256 像素采样，每个汉字 128k（64k x 2）字节数据，一套字模差不多 700M（兆）字模的筛选率是百分之一，每套合格字模，需要处理 70G 的数据。可能，黑天鹅算法最早的灵感和萌芽，就是不经意间源自这里。

2000 年,我们做“千禧版”版权登记，共一千套字体，数据总量超过 1000 x 70G=70T，是阿里健康的七十倍。

当时没有超算，没有 GPU，我们是几台电脑，每天 24 小时运算，差不多半年才做完。

其实，早在 92 年，我们 180 款的字模，数据量就差不多 20T，是阿里健康的二十倍。那时候 dvd 刚问世，刚开始只有视频 dvd，没有电脑的，我还特意去广州海印 xx 公司看过了 dvd 演示效果

在大数据领域，200 万亿数据，只是小 case 吓唬外行有用，一线的，再多数据，不过是多几个索引表而已，而且现代 k-v 表，全部采用 hash 算法，与数据规模关系不大。

这块一直都是个人资金在操作，难免有很多足，不过收获，却也是实实在在的，说起来，不比几大字库企业差。

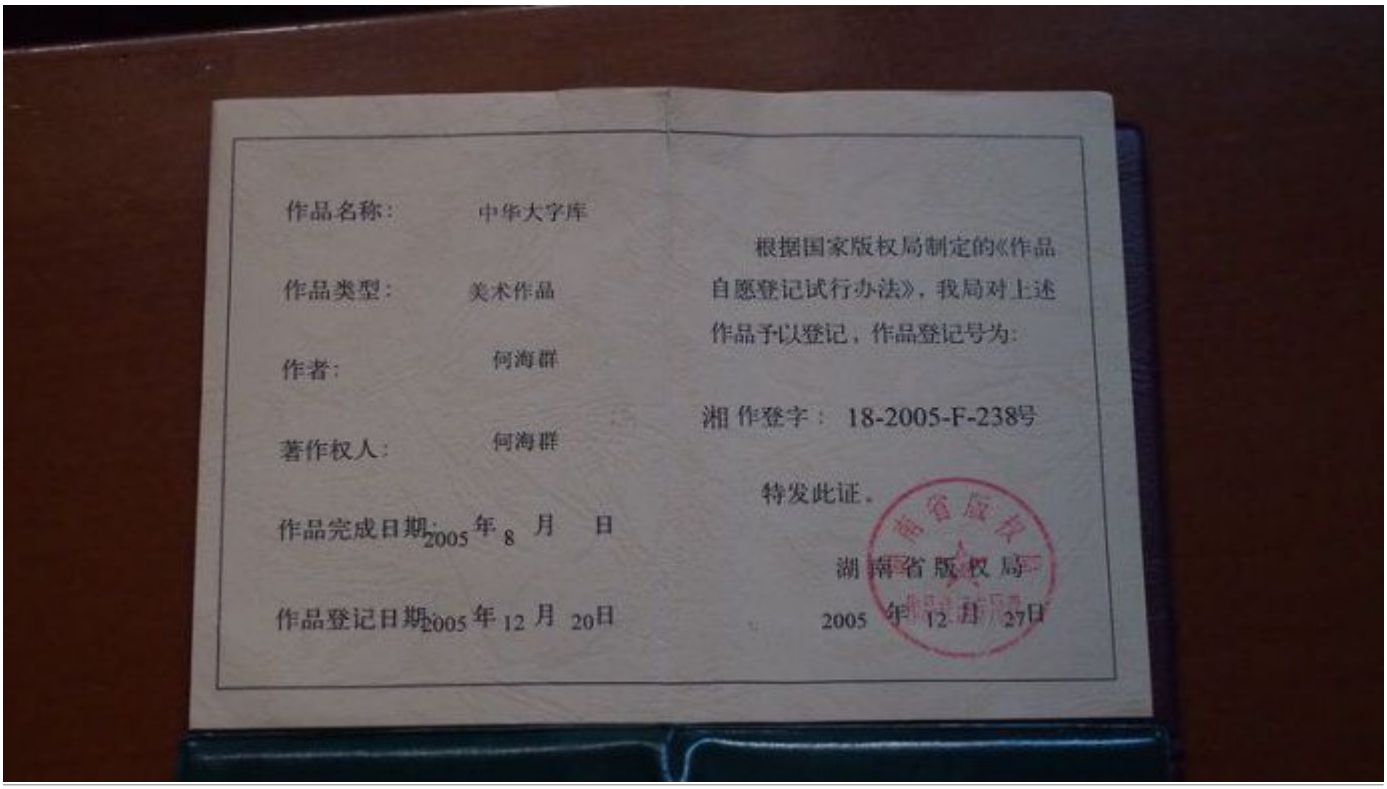
如今，回头再看，95 版的《中华大字库》，是全球首款个性化中文字库，开创了中文字体个性化时代，比文鼎类似字库要早近十年。

98 年，字王《人工智能与中文字型设计》入选《广东青年科学家论文集》。（中国科技出版社出版）

如今，字王的中文字体“智能建模”理论，已经成为字库行业三大字库建模理论之一，也是唯一具备工业化、产业化的中文字库建模理论。百度百科、互动百科的“字体”词条，均原文大段引用字王论文。以及相关文档。

2000，发布《中华大字库》千禧版。

2005，并首度对作品进行了版权登记。



2012，字王发布 x2ttf，涂鸦造字公益软件，全功能免费下载，准开源模式。

2015，字王 4K 云字库，成为首个进驻 github 的中文云字库项目。

2015，zwPython，字王 SDK 升级版，除少数字模核心模块外，全部采用开源模式，免费下载。

zwPython 3.0 目标：目前最强的集成式 Python 开发平台，大数据分析平台，没有之一。

- 比 pythonXY 更加强大，内置全中文用户手册。
- 苹果“开箱即用”模式，绿色软件，解压即可，零配置。
- 首个 All-in-one 大数据分析平台：内置 pandas、Scala、R 语言、Q 语言、Quant、matlab、hadoop、spark 模块库和 API 接口支持。（仅限 V3.0 版本）
- 超强功能：图像处理、AI 人工智能、机器学习、openCV 人像识别、gpu、openCL 并行超算开发、pygame 游戏设计等。

虽然字王网站，目前处于封闭升级、战略调整当中，但字王的口号：

我们只谈原创！我们始终第一！

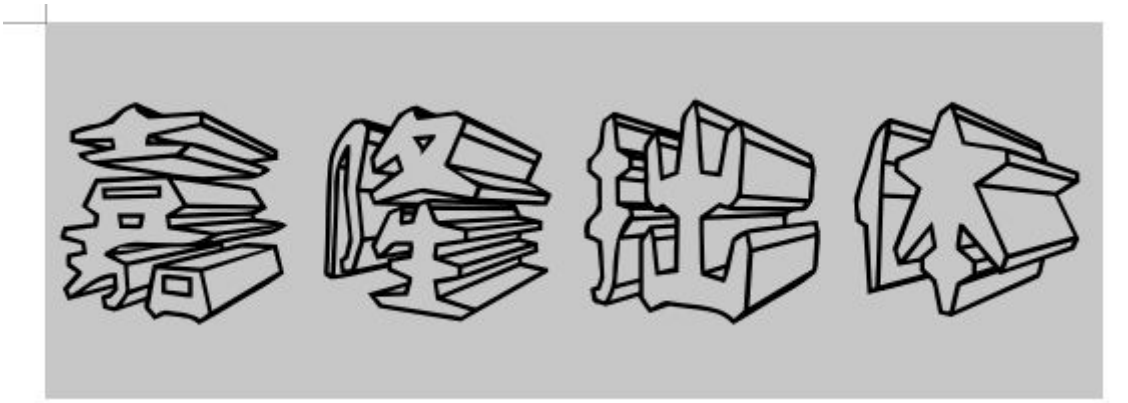
二十年，在中文字库行业，至今无人能够超越。

一直以来，总有些所谓的设计师，这些设计师，也许熟悉 ps，不过鲜见作品发布，尤其是需要扎扎实实，设计六千多字形的中文字库。这些 ps 设计师，多年以来，一直认为字王的字体缺乏美学、没有设计，作为理工背景的

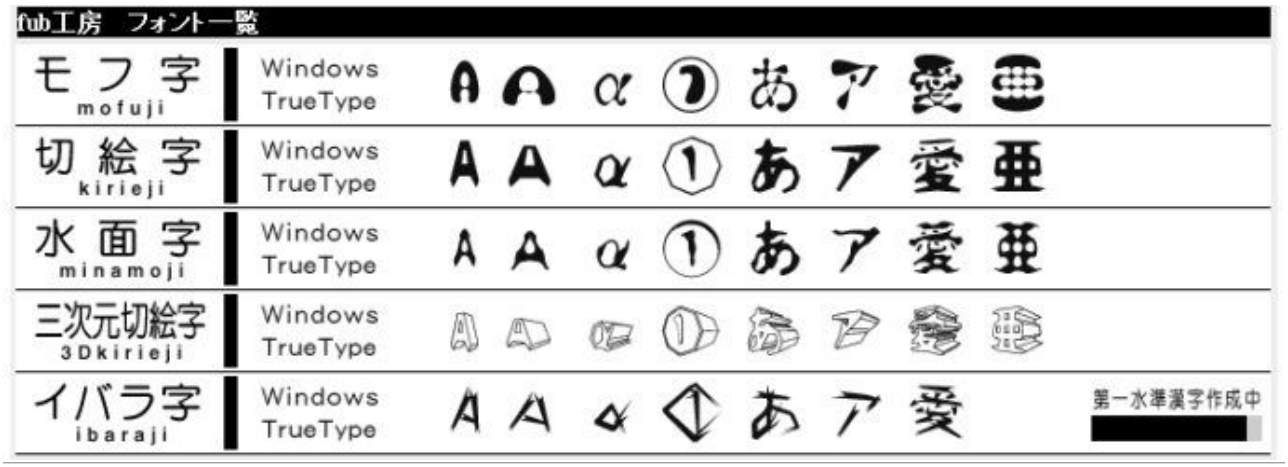
我们，一直容忍，少做回击。不过，以后再有人，特别是自己没有作品的 ps 设计师，喷字王的字体难看。请阁下先 show 下自己好看的字库，少于一百套免谈。

日本的设计，向来以注重细节，创作严谨著称。如今网络盛传的“日本三次元字体”，被业内称为“迫力满格”。据说是日本 fub 工房 2007 年的作品。http://www2s.biglobe.ne.jp/~fub/font/3Dkirieji.html

可事实上，这个所谓的“日本三次元字体”，是抄袭字王 95 版《中华大字库》“拙体字”，只是对字王“拙体字”进行了一个简单的 3D 拉边处理。虽然二十年前的资料不好查询，幸运的是，“拙体字”是笔者最喜欢的字体之一，并且是 95 版《中华大字库》的封面字体。大家不妨下载“日本三次元字体”,与 95 版嘉隆《中华大字库》的 demo 图片对比下，相似度 99%。



虽然创作也偶有“撞衫”，但无论从知识产权，还是创作角度而言，“撞衫”也是侵权。奇怪的是，不仅这款“日本三次元字体”,fub 工房的其他几款字体：mofuji，切汇字、水面字、甚至正在制作中的 ibaraji 字体，都与《中华大字库》千禧版的字形高度相似。http://www2s.biglobe.ne.jp/~fub/font/font.html



日本的工业，包括字体产业，一直强于国内，这个，至今都是如此。不过，强大的日本字库产业，居然出现了一款抄袭字王二十年前的作品，这个也是不争的事实。用这个“日本三次元字体”打脸，对于这些 ps 设计师，也算是有个交代了。为了正本清源，笔者已经给日本大使馆递交了 mail，希望能够认真处理。日本毕竟是发达国家，在知识产权方面，也应该更加规范些。如今流传甚广的剪纸体，以及 n 多类似的个性化字体，从创作角度而言，都是源自字王的拙体。

《中华大字库》，无论是 95 年 180 款的版本，2000 年千禧版的 1000 款字体，还是字王目前正在规划的 4K 云字库。字王“智能建模”,通过 AI 人工智能技术，以及目前时髦的大数据，基本上涵盖了个性化中文字体的方方面面。无他，因为字王是中文个性化字库的鼻祖，做的早，总是有些先发优势。这些，已经是历史，无需争辩。欢迎熟悉日文、知识产权的网友提供协助。

ps:给日本大使馆的公开信：

标题：关于“日本三次元字体”抄袭字王作品事件

正文：

先生，你好：

日本的字体工业，一直强于中国，这个，至今都是如此。不过，强大的日本字库产业，居然出现了一款抄袭字王二十年前的作品，这个也是不争的事实。日本毕竟是发达国家，在知识产权方面，也应该更加规范些。日本的设计，向来以注重细节，创作严谨著称。如今网络盛传的“日本三次元字体”，被业内称为“迫力满格”，是日本 fub 工房 2007 年的作品，http://www2s.biglobe.ne.jp/~fub/font/3Dkirieji.html

事实上，这个所谓的“日本三次元字体”，是抄袭字王 95 版《中华大字库》“拙体字”，只是对字王“拙体字”进行了一个简单的 3D 拉边处理。虽然二十年前的资料不好查询，幸运的是，“拙体字”是笔者最喜欢的字体之一，并且是 95 版《中华大字库》的封面字体。大家不妨下载“日本三次元字体”,与 95 版《中华大字库》的 demo 图片对比下，相似度 99%。虽然创作也偶有“撞衫”，但无论从知识产权，还是创作角度而言，“撞衫”也是侵权。

中日虽是邻国，但毕竟是千里之外，笔者又不通日文，缺乏相关资源。恳请贵机构能够妥善查处此事，正本清源。

汉字结构·熵·理论

字王·汉字结构·熵·理论--汉字结构的数字化大门，量化分析之路，终于开启。

前几天，blog《字王·百字工程》发布，百字工程正式启动。虽然，还只是最初期的数据整理阶段，也涌现了大量的创意 idea、理论突破。看来，任何事情，只要动起来，都会有收获。这几天，最大的收获，就是《汉字结构·熵·理论》。

如果说，20 年前的字王字库智能字模，只是一个概念，那么，《汉字结构·熵·理论》，可以说是套完整的“汉字结构”量化分析理论体系。也许，在五千年的历史当中，我们第一次能够，对汉字结构，进行数字化的量化分析。在信息论中，熵被用来衡量一个随机变量出现的期望值。很自然的，熵这个物理学、热学的概念，被导入了字王的汉字结构体系。

1z，一个标准单位的汉字结构的熵，是 1K（1000x1000 像素）字库的 1/n，（这个目前只是概念参数，日后会细化）汉字结构的熵，单位是英文字母：z，发音“泽”。字母：z，是 zw 字王拼音的缩写，也可以纪念太祖。泽东先生，汉字始祖“仓颉”；“颉”字，在南方某些地区，发音也有些类似。

根据这几天的数据整理，套用萌芽阶段的《汉字结构·熵·理论》，发现了几个有趣的现象：

- 500 像素版本，仿宋体最高。690z，黑体：500z，楷体：560z，新宋体/宋体：530z，雅黑：480z。
- 取样像素的大小，对字体"熵"值影响不大。雅黑 400 像素：500z，雅黑 500 像素：480z，雅黑 1K 像素：460z。
- 奇怪的是，取样精度越高，"熵"值越小，理论上，所需要的“核心字”数目也越少 。

未来数月，随着《字王·百字工程》的逐步推进，将会发布更加完整的相关文章和报告。目前，规划的方向有：

- "熵"值、"熵"密度单位的定义与延伸。
- 汉字笔画、单字、字库的"熵"值、"熵"密度 。
- 各种字形的"熵"值表。（类似早期的对数表）
- 字王·工具箱字体公益软件，也会收入"熵"值计算模块 。
- “永”字八法、“国”字结构的数学基础。

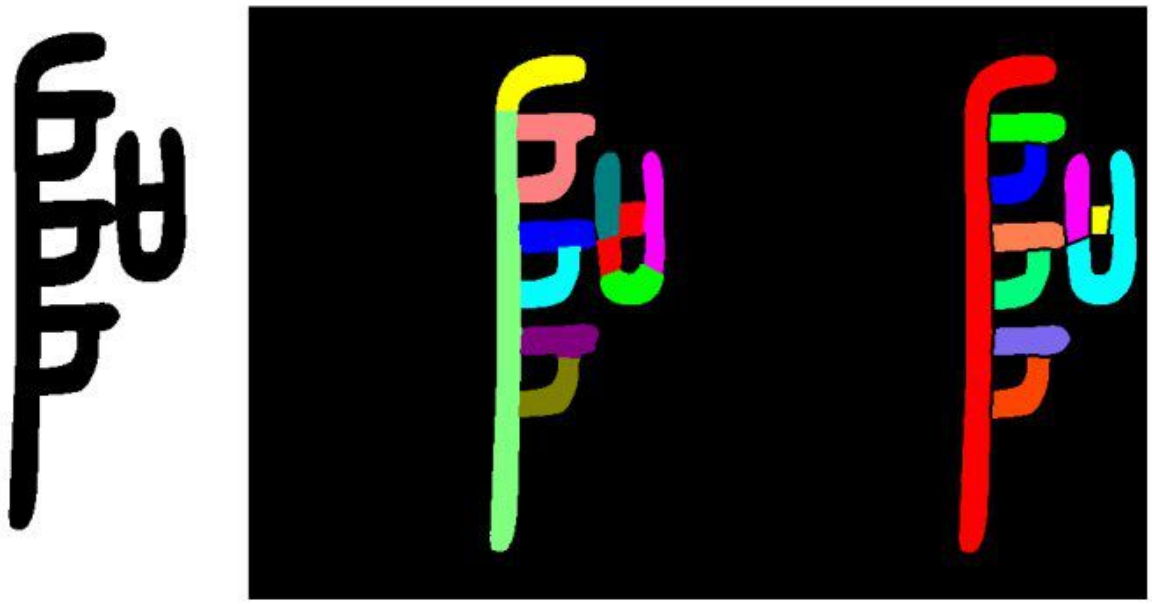
ps，补充，昨晚，导入"熵"值，连夜发布了 blog：《汉字结构·熵·理论》

睡了一觉，思路打开，继续喷。"熵"值的导入，类似现代物理学原子、电子的发现，从底层、根本上解决了汉字结构的最基本的量化单位问题。也许，目前版本的"熵"值：定义，模型、参数等，与未来，最终的数字化、汉字结构理论，天差地别。无所谓，毕竟，汉字结构的数字化大门，量化分析之路，终于开启了补充几个新的思路：

- "熵"密度，"熵"值与汉字“黑”像素的比率 。
- 统一的"熵"值，为不同汉字字体的结构分析，奠定了基础 。
- 基于"熵"值的汉字笔画排序 。
- 从"熵"值角度，分析，古代、现代诗歌、散文的形体美，包括国外作品
- 延伸下，从"熵"值角度，对圣经、古兰经、四书五经等经典作品，宗教作品进行数字化的量化分析，这个可以参考近年的定量历史学，例如从盐、布匹消耗量，分析历史人文环境 。
- "熵"值角度，对艺术作品，不同风格、流派的量化分析，尤其是黑白色调的中国山水画、书法，可以将每幅作品。（视为一种字库）
- 不同艺术家，不同阶段，作品的"熵"值演化历史，以及与其他人物作品的"熵"值对比分析 。
- 目前机器学习的热点：人脸识别、图像识别，也可以进行"熵"值处理，延伸下，这个与大数据、AI 人工智能，完全是 100%无缝对接。

（殊途同归，立地成佛，万流归宗。。。，古人诚不欺我也。这句话中的：文艺气、禅味，懂的，可。。。）

百字工程·第一阶段纪念



字王·百字工程分为三个阶段：拆字、建模、组字。三个阶段虽然都有难度，不过“拆字”阶段，相对而言，是最难的。一方面，万事开头难，想到，就等于成功了一半。

字王·百字工程，不管进度、未来如何，字王·百字工程，项目的提出，就代表了中文智能字模技术的又一个”里程碑“。从工程角度而言，“拆字”阶段的难点在于，”尽量“按汉字笔画结构，拆分汉字。古人云，画鬼易，画人难。

现代 CG，AI 智能，最大的难点就在于：拟人化。道法自然，逆向工程版本的：道法自然，更是难中之难。

为此，字王在行业内，第一次导入了机器视觉技术。利用 AI 人工智能技术，对相近的笔画、角度、矢量，进行了智能合并。目前，虽然不能 100%，还原汉字笔画顺序，也算不错了，至少，已经能够代表，目前的行业最高水平。