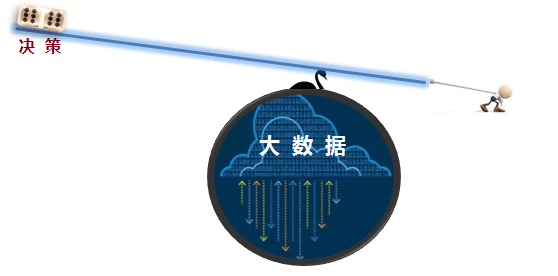
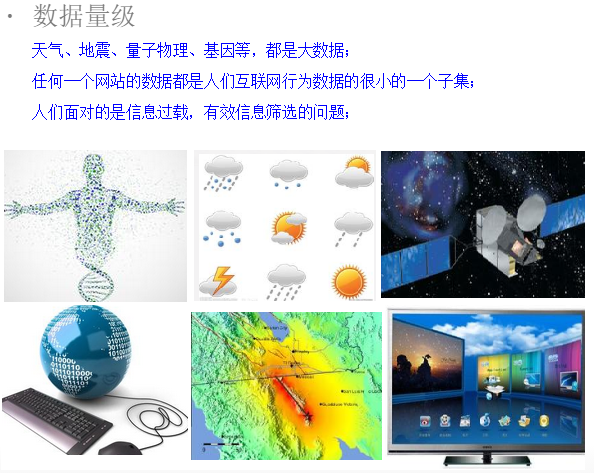
# 数据分析实践与理解-数据的大与小



数据量的增长可预见是越来越快，但我们又不能从这海量的数据直接看出端倪，要辅助决策或者智能升级，汇总、统计、分析、建模就是必经手段，数据之大摆在眼前的时候，更容易专注于解决这个大问题的手段，因为大数据大的问题是显然的，甚至大到能让我们忽视最初想要解决的问题或决策。但其实对数据大小本身的探究也是有助于解决大数据大问题的，而且在算力提升之前，我们更多的还是从这个角度来做的。所以，下面没有新瓶新酒，只是旧瓶旧酒，只想说：大并不新鲜，小亦有美点，如果大头不好解决的问题场景不妨试一下小端的魔力，也许能发现不一样的沉香。

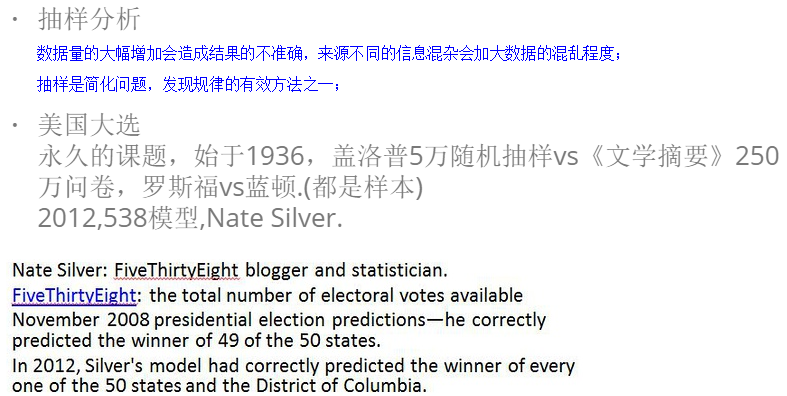
## 数据之大



在大数据之称号随着互联网之风口，占据越来越多行业制高点之前，数据之大其实是已然存在的事实。大数据的时尚，刘德寰老师早有思考（<http://weibo.com/1497587753/zxL2g0Lq2>）：关于子集，网站或公司拥有的数据都是一个子集；关于错误，混杂噪音也会随着数据增大而增大；关于抽样，也是合理有效的…

大数据并不新鲜，我们只是赶上算力爆发的临界，发现以前只能埋没角落的所谓大样本，可以用更复杂的模型来学习学习，但前辈对数据的分析之道已然是很精细的，不做借鉴甚至完全摒弃真能做得更好么。

## 数据之小



分布式的兴起，使我们得以方便高效的处理更多数据，拿更多的数据做模型。但在没有这个便利的过往，我们在小数据上也有一套行之有效的方案，那就是抽样。如果经过有效抽样设计的数据分析所做决策参考，与暴力输入所有数据全由模型诠释的决策参考，结论相同或相似或迥异，您是否有对大和小的信任偏好呢？还请注意所谓所有数据，很多情况下其实只是更大范围数据的一个小样本，比如纵向的时间选择，横向的场景限制，大小是相对的，没有小何来大，没有最大只有更大。

大小偏好的有效性，也许只能从实践中检验，不管大小，适合自己的才是最好的，只是无需一味地追求大之方案。跨越时空，美国大选不同时期的两个例子，诠释了小而美的套路，其走俏也是很有可能的。1936年盖洛普以5w样本战胜了文学摘要250w的问卷（<https://site.douban.com/146782/widget/notes/15517178/note/321469585/>），虽然它不能次次中奖；近期Nate同学基于汇总数据的统计表现，也是风头尽显，2012年50个州预测全对，并于后来公布了其背后的想法（<https://fivethirtyeight.com/features/how-the-fivethirtyeight-senate-forecast-model-works/>）。

## 数据之路

数据已日渐成为生活的必需品，不论我们是否愿意或主动参与，我们都已成为的数据的生产者和消费者。出门选择数据推荐的最优路线，饮食选择数据推荐的新开餐厅，健康状况也由随身监测的智能手环提示，在使用与反馈的循环中构建越来越智能的生活。也许不久，一个人自我评估的方式是你给这个世界留下了多少有效数据，过去有名人著书立说，以后你的历史足迹也可能被很多研究所用，基因、行为等等，用于复原疾病、社会演变的数据模型，你的数据被用的频次高了，被单拎出来复原一下你的典型人生，比著书什么的厉害多了。要理解和分析我们自己参与生产的大数据，在“大”行其道的当前，期望给大家回顾一个从“小”出发的视角，以期能更好地服务于数据生产价值。