# 预测的局限性
## *Limits of predictions*

**周涛**(1982-),男,电子科技大学教授、博士生导师,大数据研究中心主任,主要从事统计物理与复杂性、数据挖掘与数据分析方面的研究。在Physics Reports、PNAS、Nature Communications等国际SCI期刊发表200余篇学术论文,引用16 000余次,H指数为63。2009年获教育部自然科学一等奖,2011年获第十二届中国青年科技奖(系我国最年轻的获奖者),2013年获四川省科技进步一等奖,2014年获中国计算机学会自然科学二等奖(一等奖空缺),2014年起历年入选Elesvier最具国际影响力中国科学家名单(物理天文类)。2015年当选第十二届中华全国青联常务委员,并担任科学技术界别工作委员会副主任。2015年当选全国十大科技创新人物,2016年当选四川省首批杰出人才(3年评选一次)。2017年获全国创新争先奖。

中图分类号:TP399 文献标识码:A doi: 10.11959/j.issn.2096-0271.20170046

## 1 引言

从几千年前的巫祝开始，人类从来没有放弃过预测未来的追求。只不过那个时候的预测，往往是基于极其朴素的世界运行的原理，例如以水为万物生长和运动之源的泰利斯假说和以对立、协同、转换为核心的阴阳学说；又或者一些重大事件之间的偶发联系，例如某一次日食后出现了持续数月的大旱。从现代的观点来看，那不过是人类卑微理性在无常自然面前无畏的挣扎罢了。

从预测使用的理论和方法的坚实程度以及相应的对于预测结果正确性的信念，可以大致把预测分为3类。一个极端是基于严格理论的预测结果，例如基于量子力学理论预测某A粒子在一定时间内衰变成某B粒子的概率。这种预测的正确与否，可以用来检验相应物理理论是否正确。另一个极端是对未来的预言，这往往是来源于未来学家和科幻作家，例如凡尔纳在《从地球到月球》中对于人类登月和在《海底两万里》中对海底潜艇的预言，托夫勒在《第三次浪潮》中对互联网科技时代的预言以及诺查丹玛斯在《诸世纪》中对于世界末日的预言等。这类预言更像是猜测而非预测，往往只能给出定性上正确与否的判断，而不能在数字上给出精确的程度，而有些时候，因为预言可以有不同的解释方法，甚至连定性上判断正确错误都做不到，例如对李淳风和袁天罡所著《推背图》的解读。

本文所讨论的预测是介于两者之间的：既没有一套坚实的理论，也不是漫无边际的未来学说，而是基于手头已经掌握的一些数据，利用数据挖掘或者机器学习的方法，对未知的数据或者未来的发展进行的可量化的预测。得益于大数据和人工智能的发展，特别是可获取数据量的暴涨和机器学习能力的飞升，预测在现代的社会学、经济学、管理学乃至政治学中似乎已经开始并即将扮演统治性的重要地位。这些预测涉及人们可以想到和难以想到的方方面面，从一个人的民族、信仰、政治态度、是否同性恋[1]，到社交网络未来的演化[2]；从恐怖袭击最可能发生的时间和空间[3]，到各个国家和地区政治大选的最终结果[4]等。然而，从特朗普奇迹般地逆转就可以看出，这个难以捉摸的世界和尚显幼稚的技术手段，还远远不足以让预测“君临天下”。2017年2月，美国的《科学》杂志推出了一期讨论预测问题的专栏[5]，本文结合专栏的内容和笔者个人的经验，从4个方面谈一下预测的局限性。

## 2 预测中的自证陷阱和自否偏差

在很多情况下，判断一个预测是否正确不是一件容易的事情，因为预测都是某种前提或者环境下做出来的，而预测本身有可能改变这种环境。所以，有可能人们看到的正确或者错误，都是预测本身带来的，和预测结果反而没有关系。

首先，很多预测具有自证的特性[6]。举个例子，如果国家自然科学基金委员会（以下简称基金委）召集资深专家或者利用引进机器学习的方法，试图预测未来哪些研究方向中国学者能够做出颠覆性或者系统性的贡献，这就是一个典型的自证场景。因为不管是专家意见、机器学习还是随机乱说，如果基金委处在正常的逻辑下，它必然会大幅提高对这些“重要方向”的支持力度，那么我国学者更可能在这些方向做出更多贡献，写到这里有种悲伤的感觉，但

事实就是如此，而这又反过来证明了预测的正确性。

其次，很多预测具有自否的特性。比如通过对治安事件时空模式的分析，预测出了接下来一段时间最容易出现街头犯罪的场所，于是公安部门在相应的时间和地点增加了巡逻的警力，结果街头犯罪量大幅度降低。那么，这个预测本身是准确的还是错误的呢？又比如自动监测慢性肾病患者的用药和饮食习惯，发现有一个病人经常不按时按量服药并且不忌嘴，根据大量病历样本，预测他2年之后会转为尿毒症。患者震惊于这个警告，然后改变了自己的习惯，尽最大可能配合医嘱，结果5年过去了，病情也没有进一步发展。那么，这个预测本身是准确的还是错误的呢？

简而言之，在真实的场景中判断预测是否准确不是一件容易的事情，尤其当人们会基于预测结果有所行动的情况。

## 3 预测可能扼杀多样性和创新性

在绝大部分预测场景中，精确性都是评价预测效果的唯一指标。以精准广告或者个性化推荐为例，系统的核心指标是用户对广告的点击概率和推荐商品的购买概率。虽然每个人的兴趣都有不同，但显而易见的是，推荐那些原来就表现出众的产品，有更大的可能性获得用户的青睐，从而使得曾经好卖的产品越来越好卖，而初期滞销的产品找不到出路[7]。在淘宝、京东和大众点评上这样的例子比比皆是，所谓的“小而美”的推荐算法，事实上并没有效果。这就带来了两个问题，首先是初期一些随机的波动可能会让一些因为“运气”而占据上风的产品优势得以快速积累，而高质量的产品如果短期内运气不好，则可能永远没有翻身的可能性[8]。第二个更严重的问题是多样性会被打压，一些冷门小众的需求得不到响应[9]。如果人们老老实实阅读“根据预测出来的感兴趣程度”推送的新闻，购买“根据预测出来的喜爱程度”推送的商品，那么大家的兴趣、爱好和视野会逐渐趋于一致。

遗憾的是，对于机器或者对于生产商而言最优的世界，可能并不是人们最期望生活于其中的世界，因为人们的世界需要被惊喜和邂逅点亮。

再回到第2节所举的基金委预测未来重点研究方向的问题。如果要基于数据做严肃的预测，那么曾经成果斐然的研究方向肯定更有可能产生进一步的产出，而那些冷门的研究方向，甚至可能因为某种颠覆性发现而涌现的新方向，自然是不会被聪明的机器学习相中的。一些学者已经在呼吁，如果科研管理机构利用某种预测结果分配学术资源，那么这些聪明的预测很可能会扼杀创新[10]，而且如第2节所言，它们还会被证明是正确的。

## 4 预测可能加剧歧视

歧视可能来源于方方面面，例如性别、种族、年龄、身高、身体残疾等，歧视可能发生在各种地方，包括职场招聘、公共生活、社会交际等。事实上，学术界也充满了歧视，比如女性在同等水平下获得的评价和被录用的可能性都低于男性[11]，女性和有色人种在美国的基金申请中更容易被拒绝[12]。

有些歧视是可以量化的，例如求职被录用的概率、工作的年薪、申请被接收的比例等，另外一些是难以量化的，比如学小儿麻痹症患者走路的样子或学习口吃者说话

的样子。笔者主要讨论可以量化的歧视。如果用理性的眼光审视可以量化的歧视，这些歧视大体上还可以再分为两类。

第一类是歧视本身就是非理性的。譬如在Moss-Racusin[11]等人的实验中，简历对象被随机分配了一个性别和相应的名字，而实验结果显示如果你不幸被分配了一个女性的姓名，那么你被录取的意向将大大下降。这个时候，消除歧视的努力就是回归理性的判断。

第二类是歧视本身从数字上来看是理性的。举个例子，现在中国很多企业不愿意聘用年轻女性，是因为她们的感情、婚姻以及两次育儿经历都会影响她们在企业的工作情况，所以聘用同等水平甚至水平更差一些的男性，统计而言，可能给企业带来的利益更大。在这种情况下，消除歧视反而是不理性的。那为什么还要努力消除歧视呢？因为利益的最大化并不是人类全部的追求。

在上面这种情况下，消除歧视不是一种理性，而是一种温情。如果这时候祭出基于人力资源大数据的员工绩效预测算法，那么冷酷的预测结果只会进一步加剧歧视的程度。

## 5 暴力预测缺乏实际价值

在机器学习领域，有一种方法被形象地称为“off-the-shelf prediction”，中文直译为“现成的预测”，笔者为了加深大家的印象，姑且称其为“暴力预测”。顾名思义，应用暴力预测的时候，不需要对问题的背景和专业知识有深刻的理解，只需要将它转化成一个预测问题，然后从机器学习的算法货架中把各种回归分析、各种神经网络、各种支持向量机、各种决策树等取下来，依次测试预测的效果。通常，再利用集成学习的办法，融合各个预测器的结果，最终得到精度更高的预测结果。事实上，就算没有专家的知识，集成学习往往也能够给出比专家都精确的预测结果。著名的Netflix大奖赛的获胜者就是使用了集成学习的方法，而最近10年有影响的数据挖掘比赛中保守估计有不低于一半的冠军队伍都采用了这种方法。

然而，这种方法的缺点也很突出，就是没有办法给出对预测结果有效的解释。对于在计算机领域的会议和期刊上发表文章而言，解释似乎是不重要的，但是如果你希望自己的预测结果能够实实在在地影响决策，那么解释就太关键了。举个例子，在解决街头犯罪和恐怖袭击问题上，政府管理人员希望知道到底是什么因素导致了这些潜在的犯罪发生，要通过什么办法才能够消除危险。如果一个算法只能预测出袭击容易发生的地点A，那么在A处增加警力导致的结果往往是袭击最终发生在地点B。从这个意义上讲，如果要说服关键决策者，仅仅有预测结果和一些基于以往样本的测试成绩是不够的，还需要说明预测背后的主要机制、预测之所以正确或者错误的原因，甚至给出一些稳定的因果关系[13]。

尽管很多政府管理人员和社会经济学家开始认识到预测的重要性和潜在价值，但是一个完全缺乏有力解释的暴力预测而又在政府治理领域取得了重大成效的例子，笔者迄今都没有找到。

## 6 结束语

预测不代表理解，预测也不讲感情，预测甚至无法避免自己影响自己。所以，人们可以把预测当作劈开问题的一把利剑，但不能把问题的解决全盘押宝其上。

## 参考文献:


[1] KOSINSKI M, STILLWELL D, GRAEPEL T. Private traits and attributes are predictable from digital records of human behavior[J]. PNAS, 2013, 110(15): 5802-5805.

[2] LÜ L, ZHOU T. Link prediction in complex networks: a survey[J]. Physica A Statistical Mechanics & Its Applications, 2011, 390(6): 1150-1170.

[3] HELBING D, BROCKMANN D, CHADEFAUX T, et al. Saving human lives: what complexity science and information systems can contribute[J]. Journal of Statistical Physics, 2015, 158(3): 735-781.

[4] KENNEDY R, WOJCIK S, LAZER D. Improving election prediction internationally[J]. Science, 2017, 355(6324): 515-520.

[5] JASNY B R, STONE R. Prediction and its limits[J]. Science, 2017, 355(6324): 469.

[6] MERTON R K. The self-fulfilling prophecy[J]. The Antioch Review, 1948, 8(2): 193-210.

[7] LÜ L, MEDO M, YEUNG C H, et al. Recommender systems[J]. Physics Reports, 2012, 519(1): 1-49.

[8] SALGANIK M J, DODDS P S, WATTS D J. Experimental study of inequality and unpredictability in an artificial cultural market[J]. Science, 2006, 311(5762): 854-856.

[9] ZHOU T, KUSCSIK Z, LIU J G, et al. Solving the apparent diversity-accuracy dilemma of recommender systems[J]. PNAS, 2010, 107(10): 4511-4515.

[10] CLAUSET A, LARREMORE D B, SINATRA R. Data-driven predictions in the science of science[J]. Science, 2017, 355(6324): 477-480.

[11] MOSS-RACUSIN C A, DOVIDIO J F, BRESCOLL V L, et al. Science faculty's subtle gender biases favor male students[J]. PNAS, 2012, 109(41): 16474-16479.

[12] GINTHER D K, SCHAFFER W T, SCHNELL J, et al. Race, ethnicity, and NIH research awards[J]. Science, 2011, 333(6045): 1015-1019.

[13] ATHEY S. Beyond prediction: using big data for policy problems[J]. Science, 2017, 355(6324): 483-485. □