# 谷歌流感的寓言：大数据中的陷阱分析

# 谷歌流感的寓言：大数据分析中的陷阱

**大卫·拉泽，[1,2]*瑞安·肯尼迪，[1,3,4]加里金，[3]亚历山德罗维斯皮纳尼[5]**

[1]雷泽实验室，东北大学，波士顿，马萨诸塞州02115，美国。[2]哈佛大学肯尼迪学院，哈佛大学，剑桥，马萨诸塞州02138，美国。[3]哈佛大学定量社会科学研究所，剑桥，马萨诸塞州02138，美国。[4]休斯顿大学，休斯顿，德克萨斯州77204。[5]生物和社会技术系统建模实验室，东北大学，波士顿，MA 02115，美国。
*通讯作者。电子邮件： d.lazer@neu.edu。

2013年2月，谷歌流感趋势（GFT）成为了头条新闻，但这并不是因为谷歌的高管或流感追踪系统的创造者所希望的原因。据《自然》杂志报道，GFT预测的流感样疾病（ILI）的就诊比例是疾病控制和预防中心（CDC）的两倍多，CDC是根据美国各地实验室的监测报告估算的（1,2）。尽管GFT是用来预测CDC的报告的，但这种情况还是发生了。鉴于GFT经常被认为是大数据的典型使用（3,4），我们可以从这个错误中吸取什么教训呢？

我们所发现的问题并不仅限于GFT。关于搜索或社交媒体是否能够预测x的研究已经变得很普遍（5-7），并且经常与传统的方法和假设形成鲜明的对比。虽然这些研究显示了这些数据的价值，但我们还远没有看到它们可以取代更传统的方法或理论(8)。我们探讨了导致GFT错误的两个问题，即大数据的傲慢和算法动态，并为大数据时代的向前发展提供了经验教训。

**大数据傲慢**

“大数据的傲慢”通常是一种隐含的假设，即大数据是传统数据收集和分析的替代品，而不是补充。我们已经断言，在大数据中存在着巨大的科学可能性（9-11）。然而，数据的数量并不意味着人们可以忽略测量、构造效度和信度以及数据之间的依赖性等基本问题（12）。其核心挑战是，大多数受到广泛关注的大数据并不是旨在产生适合科学分析的有效和可靠数据的仪器的输出。

GFT的最初版本是一个特别困难的大小数据的结合。本质上，该方法是在5000万个搜索词中找到最佳匹配项，以匹配1152个数据点（13）。找到与流感倾向相匹配但结构无关且无法预测未来的搜索词的几率相当高。事实上，GFT的开发者报告说，剔除了与流感无关但与疾控中心数据密切相关的季节性搜索词，比如关于高中篮球的数据（13）。这应该是一个警告，即大数据过度拟合了少数案例，这是数据分析中的一个标准问题。

当GFT完全错过了2009年非季节性A-H1N1流感大流行时，这种抛出特殊搜索词的特殊方法失败了（2,14）。简而言之，GFT的最初版本部分是流感探测器，部分是冬季探测器。GFT的工程师在2009年更新了算法，这个模型一直在运行，并在2013年10月宣布了一些变化（11,15）。

尽管直到2013年才得到广泛报道，但新的GFT在很长一段时间内一直高估流感流行率。GFT在2011-2012流感季节也有很大的差距，从2011年8月开始的108周中有100周缺席（见图表）。1).这些误差不是随机分布的。例如，上周的误差预测了本周的误差（时间自相关），误差的方向和幅度随一年中的时间（季节性）而变化。这些模式意味着GFT忽略了可以通过传统统计方法提取的大量信息。

即使在2009年GFT更新后，该算法作为一个独立的流感监测器的相对价值仍然值得怀疑。2010年的一项研究表明，GFT的准确性并不比使用已经可用的（通常滞后2周）CDC数据的相当简单的预测好多少(4)。从那时起，比较变得更糟了，滞后模型的表现明显优于GFT（见图表）。即使是3周大的CDC数据在预测当前的流感流行率方面也比GFT做得更好[见补充材料（SM）]。

考虑到大量的方法提供了对流感活动的推断（16-19），这是否意味着当前版本的GFT没有用处？不，将GFT与其他近实时健康数据相结合可以获得更大的值（2,20）。例如，通过结合GFT和滞后的CDC数据，以及动态地重新校准GFT，我们可以显著提高GFT或单独使用CDC的性能（见图表）。这并不能替代正在进行的评估和改进，但是，通过合并这些信息，GFT本可以在很大程度上治愈自己，并很可能一直留在头条新闻之外。

## 算法动力学

所有的实证研究都建立在测量的基础上。仪器实际上捕获了感兴趣的理论结构吗？不同病例和不同时间的测量是否稳定和可比性？测量误差是系统性的吗？至少，由于算法的动态影响了谷歌的搜索算法，GFT很可能是流感流行率的不稳定反映。算法动态是指工程师为改善商业服务和消费者在使用该服务时所做的改变。谷歌的搜索算法和用户行为上的一些变化可能会影响到GFT的跟踪功能。对GFT的错误最常见的解释是上个流感季节由媒体引发的恐慌（1,15）。虽然这可能是一个因素，但它不能解释为什么GFT两年多来一直在高幅度缺失。2009年版的GFT经受住了其他与流感有关的媒体恐慌，包括2005-2006年甲型H5N1流感（“禽流感”）爆发和2009年A/H1N1流感（“猪流感”）大流行。一个更有可能的罪魁祸首是谷歌的搜索算法本身所做的改变。

谷歌搜索算法不是一个静态的实体——该公司在不断测试和改进搜索。例如，谷歌官方搜索博客仅在2012年6月和7月就（SM）报告了86个变化。搜索模式是该公司的程序员在不同的子单位和全球数百万消费者中做出的数千个决策的结果。

复制GFT的原始算法存在多重挑战。GFT从未记录过所使用的45个搜索词，而且已经发布的例子似乎具有误导性（14）（SM）。谷歌确实提供了一项称为谷歌相关的服务，允许用户识别与给定时间序列相关的搜索数据；然而，它仅限于国家层面的数据，而GFT是使用区域层面的相关性开发的（13）。该服务也没有返回在与GFT相关的出版物中报告的任何示例搜索词（13,14）。

尽管如此，使用谷歌相关性来比较GFT时间序列的相关搜索词与CDC的数据返回的相关搜索词，显示了一些有趣的差异。特别是，搜索流感的治疗方法和搜索区分感冒与流感的信息密切跟踪GFT的错误（SM）。这就表明了一种解释的可能性

相对搜索行为的变化是“蓝色团队”动态的——其中产生数据（以及用户利用率）的算法已经由服务提供商根据他们的业务模型进行了修改。谷歌在2011年6月报道说，它已经修改了搜索结果，提供建议的额外搜索词，并在2012年2月再次报告说，它现在正在返回潜在的搜索诊断，包括“发烧”和“咳嗽”等身体症状（21,22）。前者建议寻找流感的治疗方法，以应对一般流感的询问，而后者可能可以解释一些区分流感和普通感冒的搜索量的增加。我们还记录了其他几个可能影响了GFT（SM）的变化。

在改善对客户的服务方面，谷歌也在改变数据生成过程。对搜索算法的修改可能是为了支持谷歌的商业模式——例如，通过快速向用户提供有用的信息，部分是为了促进更多的广告收入。推荐的搜索，通常是基于其他人搜索的内容，会增加某些搜索的相对大小。由于GFT在其模型中使用了搜索词的相对流行率，因此搜索算法的改进可能会对GFT的估计产生不利影响。奇怪的是，GFT假设某些术语的相对搜索量与外部事件静态相关，但搜索行为不仅是外源性决定的，它也是由服务提供商内源性培养的。

蓝色团队的问题并不仅限于谷歌。推特和Facebook等平台一直在重新设计，一年前对这些平台收集的数据进行的研究是否能在后期或更早的时期被复制是一个悬而未决的问题。

虽然这在GFT中似乎不是一个问题，但学者们也应该意识到我们所监控的系统可能遭到“红队”攻击的可能性。当研究对象（在本例中是网络搜索者）试图操纵数据生成过程以实现他们自己的目标，如经济或政治利益时，红色团队的动态就会发生了。推特上的民意调查就是这些策略的一个明显例子。竞选团队和公司，意识到新闻媒体正在监视推特，已经使用了许多策略来确保他们的候选人或产品的趋势（23,24）。

推特和脸书也有类似的用途来传播有关股价和市场的谣言。具有讽刺意味的是，我们在监控使用这些开放信息源的人的行为方面越成功，操纵这些信号就越诱人。

**透明度、粒度和全数据**

GFT寓言作为一个案例研究很重要，在我们推进大数据分析时代的过程中，我们可以学习关键的教训。

*透明度和可复制性*。复制是整个学院越来越受到关注的问题。与GFT相关的论文的支持材料不符合新兴的社区标准。既没有识别出核心搜索词，也没有提供更大的搜索语料库。谷歌不可能向外界提供其完整的数据武库，考虑到隐私问题，这在伦理上也不会被接受。然而，对于导数的、聚合的数据，并没有这样的约束。即使人们可以访问谷歌的所有数据，也不可能从提供的分析信息中复制原始论文的分析。

虽然谷歌开发的谷歌表面上来自于GFT的概念是值得称赞的，但公共技术不能被用来复制他们的发现。点击标题为“匹配实际流感活动模式(这是我们建立谷歌流感趋势的方式”的链接具有讽刺意味的是，不会产生GFT的复制搜索词（14）。奇怪的是，论文（14）中提供的少数搜索词似乎与GFT或CDC数据（SM）没有强烈的关系——我们推测作者觉得没有明确的需要掩盖实际识别的搜索词。

风险是双重的。首先，科学是一项累积的努力，而要想站在巨人的肩膀上，则要求科学家能够不断地评估他们正在建立的工作（25）。第二，知识的积累需要以数据的形式存在的燃料。有一个研究人员网络在等待着提高大数据项目的价值，并从这些类型的数据中挤出更多可操作的信息。关于GFT的最初设想——对目前传染病的流行情况产生更准确的了解，可能允许采取挽救生命的干预措施——从根本上是正确的，所有的分析都表明，确实有有价值的信号需要提取。

谷歌是一项业务，但它也拥有关于人类的欲望、思想和联系的信任数据。“不做坏事”就能赚钱（意思是谷歌的座右铭）是不够的。学术界也有责任建立机构模式，以促进与这些大数据项目的合作——这是现在在大学中经常缺失的东西（26）。

*利用大数据来理解未知的数据。*因为一个简单的流感流行率滞后模型会表现得如此好，CDC数据的模型预测几乎没有改进的空间[这不适用于其他直接测量流感流行率的方法，例如。(20, 27, 28)].如果你是90%的方法，你最多可以获得最后的10%。更有价值的是了解在非常局部的水平上的流感流行，这对CDC来说广泛生产是不现实的，但原则上，可以提供更细的GFT测量。反过来，这样一个精细的颗粒状的观点将为流感传播的生成模型提供强大的输入，并提前几个月更准确地预测流感（29-33）。

*研究算法。*由于数百万工程师和消费者的行动，推特、脸书、谷歌和互联网都在不断变化。研究人员需要更好地了解这些变化是如何随着时间的推移而发生的。科学家们需要跨时间使用这些数据源和使用其他数据源来复制这些发现，以确保他们观察到的是稳健的模式，而不是消失的趋势。例如，用谷歌进行控制实验是非常可行的，例如，看看谷歌搜索结果如何根据位置和过去的搜索而不同（34）。更普遍地说，研究嵌入我们社会的社会技术系统的演变在本质上是重要的，值得研究的。谷歌、推特和脸书背后的算法帮助确定我们对健康、政治和朋友的了解。

*这不仅仅是关于数据的大小。*大数据研究和更传统的应用统计数据倾向于生活在两个不同的领域——意识到彼此的存在，但通常不太信任彼此。大数据为新的见解提供了巨大的可能性

（特别是围绕网络，空间和时间动力学），为了在系统层面上理解人类系统，并为了检测变量之间的相互作用和非线性关系。我们认为，这些都是研究人类行为的最令人兴奋的前沿领域。

然而，传统的“小数据”通常提供大数据中不包含（或可包含）的信息，而使大数据成为可能的因素正是使更传统的数据收集成为可能。互联网已经为改善标准的调查、实验和健康报告开辟了道路（35）。而不是专注于“大数据革命”，也许是时候我们专注于“所有数据革命”，在我们认识到世界的关键变化创新分析，使用所有传统和新的数据来源，并提供更深层次、更清晰的了解我们的世界。

**参考资料和注释**


1.D.巴特勒，《自然》杂志，494,155（2013）。
2.D.R.*Olson等人，PLOS Comput。比奥尔。*9, e1003256 (2013).
3.A.McAfee。Brynjolfsson哈夫。*公共汽车。发动机的旋转*90, 60, 68, 128 (2012).
4.S.Goel等，Proc。*Nat1。阿卡德。科学。U.S.A.*107, 17486 (2010).
5.A.*Tumasjan等人，在第四届国际AAAI会议上，亚治亚州，2010年7月11日至15日（人工智能促进会，2010年），p。*178 - 185.
6.J.博伦等人，J。*压缩。科学。*2, 1 (2011).
7.F.Ciulla等。，*EPJ数据科学。*1, 8 (2012).
8.P.T.*Metaxas等人，《PASSAT-IEEE第三届社会计算国际会议论文集》，波士顿，2011年10月9日至11日，（IEEE，2011），页。*165 - 1171.doi:10.1109/PASSAT/SocialCom.2011.98
9.D.*Lazer等人，《科学》323,721（2009）。*
10.A.维斯皮格纳尼，《科学》325,425（2009）。
11.G.*金，科学331,719（2011）。*
12.D.博伊德和K。克劳福德，Inf.，通讯。*& Soc.*15, 662 (2012).
13.J.金斯伯格等人，《自然》457,1012（2009）。
14.S.Cook等人，PLoS ONE 6，e23610（2011）。
15.P.科普兰等人，Int。*社会忽略。太分离*2013, 3 (2013).
16.C.Viboud等人，我。*J.流行病学。*158, 996 (2003).
17.W.W.汤普森等人，J。*感染。分离*194(增刊。2), S82 (2006).
18.I.M.*霍尔等人，流行病学家。感染。*135, 372 (2007).
19.J.B.S.Ong等人，PLoS ONE 5，e10036（2010）。
20.J.R.Ortiz等人，美国公共科学图书馆ONE 6，e18687（2011）。
21.组织相关搜索列表，谷歌；http://insidesearch.blogspot.com/2011/06/organizing-lists-of-related-searches_16.html
22.改善健康搜索，因为你的健康很重要，谷歌；http://insidesearch.blogspot.com/2012/02/improving-health-searches-because-your.html
23.E.Mustafaraj, P.Metaxas，《WebSci10论文集》，罗利，2010年4月26日和27日（网络科学信托基金，2010年）；http://journal.webscience.org/317/。
24.J.*Ratkiewicz等人，发表在第五届国际AAAI网络博客和社交媒体会议论文集，旧金山，加州，2011年8月7日至11日（AAAI，2011），p。*297 - 304.
25.G.*金，PS Polit。科学。波利特。*28, 443 (1995).

26. P. *维森，《高等教育纪事报》，2013年9月13日；* http://chronicle.com/article/Researchers-Struggle-to-Secure/141591/.
27. R. Lazarus等人，BMC公共卫生出版社，1,9（2001年）。
28. R. Chunara等人，在线J。*公共卫生通知。*5, e133 (2013).
29. D. 巴尔坎等人，Proc。*Natl。阿卡德。科学。U.S.A.*106, 21484 (2009).
30. D. L. *Chao等人，PLOS Comput。比奥尔。*6, e1000656 (2010).
31. J. 萨满。卡斯皮克，Proc。*Natl。阿卡德。科学。U.S.A.*109, 20425 (2012).
32. J. 沙曼等人。，*Nat。通勤。*4, 2837 (2013).
33. E. O. Nsoesie等人，美国公共科学图书馆ONE 8，e67164（2013）。
34. A. 汉纳克等人。，*第22届国际万维网会议会议记录，里约热内卢，2013年5月13日至17日（2013年），第3页。*527 – 538.
35. A. J. 伯林斯基等人，波利特。*肛门。*20, 351 (2012).

**致谢：这项研究部分由美国国家科学基金会资助的。**1125095，部分由情报高级研究项目活动（IARPA）通过内政部国家商业中心（DoI/NBC）合同D12PC00285。我们也感谢HRL实验室有限责任公司提供的帮助和支持。本文所包含的观点和结论是作者的观点和结论，不应被解释为必然代表NSF、IARPA、DoI/NBE、HRL或U的、所表达或暗示的官方政策或背书。S. 政府

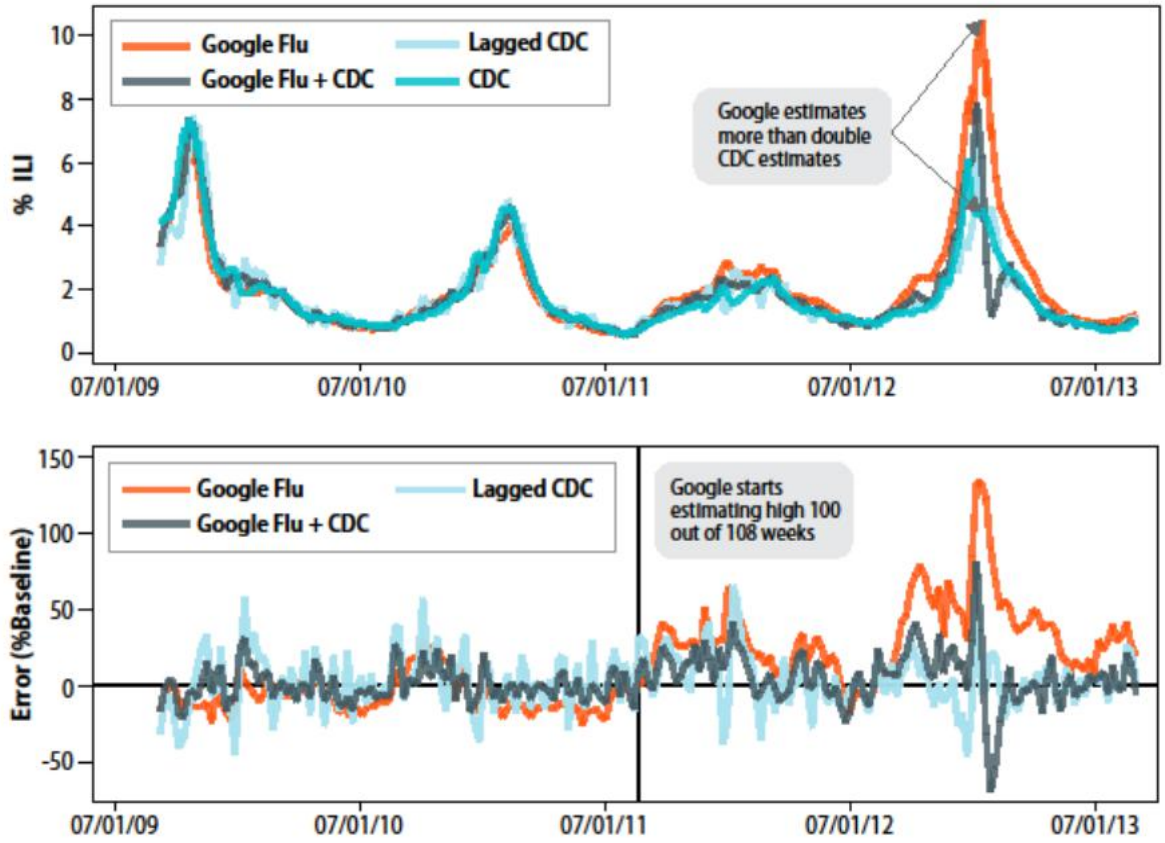


**图1.GFT高估-在上**

比2011-2012年的实际水平高出50%以上。从2011年8月21日至1

2013年9月，GFT报告了108周中有100周的流感流行率过高。（上）ILI的医生就诊估计。“滞后的CDC”包含了52周的季节性变量和滞后的CDC数据。“谷歌Flu+CDC”结合了GFT、滞后的CDC估计、GFT估计的滞后误差和52周的季节性变量。（下）错误[作为CDC基线的百分比：（由CDC估计）/CDC数据]。这两种替代模型的误差都比单独的GFT要小得多。在样本外期间，GFT的平均绝对误差（MAE）为0.486，滞后CDC为0.311，GFT和CDC组合为0.232。所有这些差异在P<为0.05时均有统计学意义。看到SM。