一、1.

数据分析过程

1 收集5000万个最常见的搜索查询按每周计数的时间序列，一周内某一特定的搜索查询与搜索查询的总数之比作为一个查询分数

2 收集特定地区ILI相关的医生就诊占总就诊的百分比，该百分比作为医生就诊与ILI相关的概率

3 建立上述查询分数的对数概率与ILI相关的医生就诊对数概率之间的线性关系，系数待定

4 将某一个特定的搜索查询作为解释变量Q，以相关程度测量准确度

5 在对5000万个查询候选进行测试后，选择相关性最高的45个用于模型当中

6 使用2003-2007年间的数据进行拟合，得到适用于各个区域的系数

7 运用模型生成2007-2008年ILI估计数

所犯的错误

在上述第5步中直接使用了相关性最高的45个查询，但高相关性不代表这些查询一定与ILI的发生有因果关系。文中给出的后面55个查询可以说明这个问题。

尽管美国疾病控制与预防中心报告在年度流感季节以外的几周内没有提供任何数据，模型拟合中也排除了这些日期，但模型被用于生成这几周的未经验证的ILI估计数。

一、2.

不严谨之处

使用搜索数据进行分析时，未考虑到一些其他情况，例如，出现与流感相近的症状时，人们可能会进行搜索，但此时他们很可能并未患上流感。或是新闻报道中的相关内容引发了搜索量激增。

一、3.

评论类：

有学者认为谷歌的预测是一种好的尝试。Google Uses Searches to Track Flu's Spread一文作者认为，谷歌的流感趋势预测是一种值得尝试的数据利用方法，可以将用户在日常搜索中无意产生的数据充分利用起来，并且能较为准确地预测一些我们不曾设想能预测的事件，这是一种大数据的智慧思维。

但是更多学者基于谷歌预测的不尽人意的结果持批判态度，The Parable of Google Flu: Traps in Big Data Analysis一文中指出，谷歌流感趋势的预测有时非常不准确，特别是在2011-2013年期间，谷歌一直高估相对流感发病率，在2012-2013年流感季节的一个时间间隔内，预测的医生就诊次数是CDC记录的两倍。基于此类问题，该文章认为谷歌追对其踪的搜索词（如“发烧”和“咳嗽”）的分析，以及搜索算法随时间变化的影响，仍然不够到位，以致于不能精准的判断用户行为和患上流感的关系。

在Google Flu Trends: A case of Big Data gone bad?一文中，也给出了产生上述问题的部分原因的解释：在谷歌搜索与流感相关的人可能对如何诊断流感知之甚少，搜索流感或流感症状很可能是在研究与流感相似但实际上不是流感的疾病症状。但是谷歌的大数据算法没有考虑这些因素，从而根据通常的隐含的假设，形成大数据傲慢，给出错误判断。

实验类：

不少研究者基于谷歌的已有系统作出了改进方案。The Parable of Google Flu: Traps in Big Data Analysis一文中提到，通过将GFT和滞后的CDC数据相结合，以及动态重新校准GFT，可以显著提高GFT或CDC单独的性能。

Adaptive nowcasting of influenza outbreaks using Google searches一文中指出，考虑到样本内预测具有局限性，研究者使用滑动窗口在样本外构建了一个提前一步的基线模型，并采取和谷歌流感趋势相同的时间序列进行分析，结果证明谷歌搜索数据确实可以用于改进估计，将仅使用CDC数据的模型中看到的误差降低了52.7%。

Reappraising the utility of Google Flu Trends". PLOS Computational Biology的研究者则称，他们通过使用基于流感样疾病发病率和原始GFT模型输出的随机forrest回归模型，随访工作能够显著提高GFT的准确性。

二、

辛普森悖论：

举世瞩目的2016美国总统大选。特朗普和希拉里分别代表共和党和民主党角逐总统大位。

选票最终开票结果显示，希拉里在全美普选选票上大胜特朗普200多万张，但是最终问鼎白宫的却是特朗普。为什么低票数的人反而获得了选举的成功呢？

因为美国的选举制度并没有我们想象那么简单。

总统由各州选出的选举人团投票产生，选举人代表选民意愿投票。总统候选人获得超过半数选举人票（至少270张）即可获胜。美国大选一州为一个选举人团单位，根据“胜者全得”原则，美国全体选民在总统候选人之间投票，随后各州统计每名候选人在本州所获实际选票数，超过半数的候选人即获得本州全部选举人票（注：除缅因和内布拉斯加两个州是按普选票得票比例分配选举人票外，其余48个州和华盛顿特区均实行“胜者全得”制度）。在选举人票方面，希拉里仅仅拿下232张，远不及特朗普的306张。74张票的落差，让希拉里无力可回天。希拉里赢得比特朗普更多的民众投票。她获得了超过6500万张选票，是此前美国历史上总统大选第二多的选票，仅次于奥巴马。但因为特朗普赢得了30个州的胜利，在选举人团票上，特朗普却占据了优势，最终逆袭赢得大选。

我们看到，辛普森悖论现象在2016年美国大选上出现了。

伯克森悖论：

伯克森悖论产生的原因是我们忽略了一部分统计数据，没有考虑全部的样本，导致出现以偏概全的现象。

比如，出现“寒门出孝子，豪门出逆子”的原因，大概是因为寒门受到的关注度比较小，很多寒门中的逆子都被大家忽略了，大家只关注寒门中的孝子，而豪门的一举一动都是大家关注的，所以看上去豪门的逆子就比较多了。

伯克森悖论里的错误，还出现在很多地方。比如，很多家长和学生有一个错误印象，学霸似乎学习很轻松，自己的孩子或者自己很努力却学不好，于是就得出结论，自己不是读书的料，努力也没有用。

真实情况是什么样的呢？我们假设一个年级500人，一半人用功，一半人不用功；再假设全年级有100名学霸，其中80名用功，20名不用功。在这里，用功和成为学霸是正相关关系，但在旁人的眼中是怎么样呢？

大家只盯着学霸看，就会发现有20%的人明明不用功也能学好，就觉得学好和用不用功之间没有必然联系；然后大家又看那些用功学习的人，就会发现就算用功了，也有170/250=68%的人学不好。于是有些人就觉得，用功也没有用啊，学不好就是学不好，也不是我努不努力能改变的。

但认真算一下就会发现，用功的学生成为学霸的概率是80/250=32%，而不用功的人成为学霸的概率只有20/250=8%。不论资质怎么样，比起不用功的人，用功学习的人成为学霸的概率足足高出了4倍。之所以会产生前面的错觉，是因为我们忽略了那些既不用功、学得又差的人。

柏克森悖论时刻提醒我们，我们的直观感觉总是建立在我们能够注意到的现象之上的，但真实的世界里包含了很多被我们忽视的现象，要了解真实的情况，就需要把那些忽视掉的情况都考虑进来。

强加因果：

某大学的选修课实行挂牌授课制，即学生通过试听两周课来决定是否选修某位教师主讲的某一门课，结果有位教师主讲的某门课的选修人数达到了本年级在校生的85％。假设导致这一结果的真实原因是：这位教师对学生是否来听课和考试要求不严，学生易于蒙混过关。但在这种情况下，这位教师或教务部门对这一结果会解释说：学生选修的人数较多反映了这位教师的教学质量较高