2014 年《时代》杂志的年度人物称号由埃博拉患者护理人员获得，在向他们致敬的同时，让我们回顾一下去年这场备受关注并且持续到今年的全球性传染病事件。

2014 年继马航客机失联事件之后，始发于西非几内亚的埃博拉病毒疫情经由传统媒体和数字媒体走入全球公众视野。据世界卫生组织提供的资料[1]显示，埃博拉病毒最早是在 1976 年同时爆发的两起疫情中首次出现的，一起在苏丹，另一起在刚果民主共和国。后者发生在位于埃博拉河附近的一处村庄，该病由此得名。该次埃博拉疫情在刚果民主共和国爆发的是扎伊尔标准亚种，累计 318 人患病，280 人死亡，致死率 88%；在苏丹爆发的则是苏丹亚种，累计 284 人患病，151 人死亡，致死率 53%；另外还有雷斯顿、科特迪瓦、邦地布优等三个亚种，对动物和人类的危害相对温和。据悉，目前正在肆虐全球的埃博拉病毒，正是致死率最高的扎伊尔标准亚种。

埃博拉病毒疫情时隔几年便爆发一次，不过之前每次疫情规模都比较小，主要集中在一个地区爆发，并且局限在中非。特别地，刚果民众共和国史上曾多次爆发埃博拉疫情。

2014 年 3 月开始爆发的埃博拉病毒疫情的规模引起了国际社会的关注，并且被世界卫生组织列为“国际间关注的公共卫生紧急事件”（历史第三次）。首先，这次疫情涉及到了多个国家和地区。全境范围受影响的国家包括几内亚、利比里亚和塞拉利昂。部分领土受影响的地区包括马里的卡伊，西班牙的马德里，美国的达拉斯、德克萨斯州和纽约市，英国苏格兰地区的格拉斯哥市，尼日利亚的拉各斯哈科特港，以及塞内加尔的达喀尔。其次，这次埃博拉病毒疫情出现的病例和死亡数字超过了此前几次疫情的总和。截至 2014 年 12 月 31 号，累计 20206 人患病，7905 人死亡[2]。并且数字还在不断增加。而所有的埃博拉护理者，则被美国《时代》周刊选为 2014 年的年度人物。

回望过去几十年，人类无疑在信息、科技、生物、医疗等领域取得了今非昔比的成果。

1 http://www.who.int/mediacentre/factsheets/fs103/en/

2 http://apps.who.int/ebolaweb/sitreps/20141231/20141231.pdf

反观 2014 年埃博拉病毒疫情的爆发、传播、媒体报导、控制，我们不禁开始思考在这个大数据时代，数据、统计、理性思考、批判思维能为人类对疫情防控带来什么好处？这篇文章试图从三个角度去阐述大数据如何与疫情防控紧密相连。本文第一部分讨论了如何通过交通数据、移动通信数据与社交媒体数据等非传统公共卫生数据来测算乃至预测疫情风险；第二部分重点关注死亡率的不同估算方法带来的对于疫情风险的不同认知；第三部分聚焦在埃博拉病毒疫情的治疗和防控支出数据。

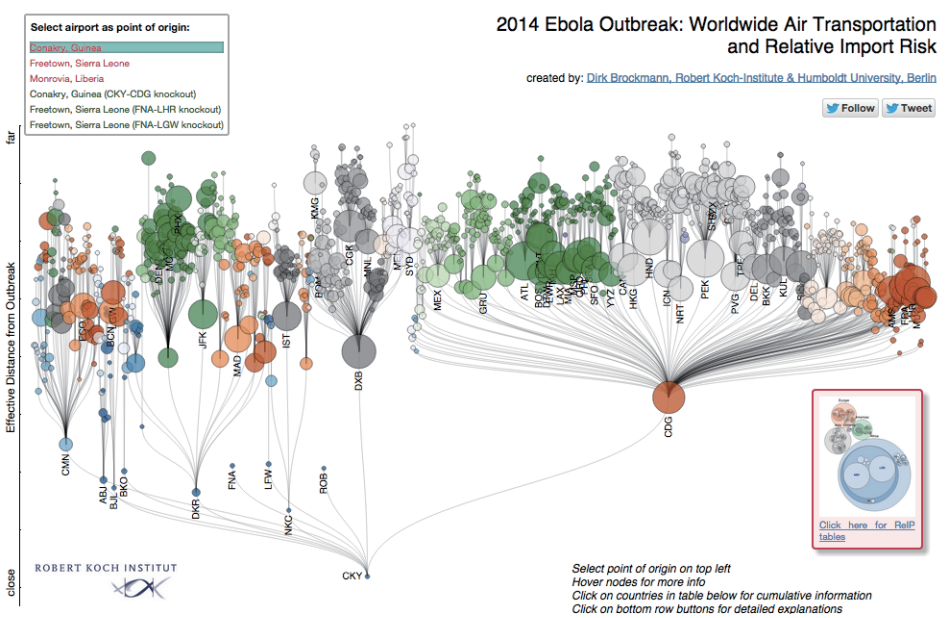
一、数据驱动的疫情预测

1. 由交通数据预测疫情 [3]

全球人口流动的日趋频繁使某区域爆发的疫情会给全球其他国家和地区带来潜在的疫情风险，于是如何对此类疫情风险进行有效地预测和评估自然就成了一个值得探究的课题。本次埃博拉爆发的一个重大特征即是，疫情通过交通运输突破国境，在源爆发地几内亚之外多国肆虐。对于国际间的人口流动，飞机显然是最主要的交通方式，因此对机场人口流动数据的统计分析也就成了重中之重。

事实上机场数据也早已被很多领域的研究者们所重视，基于此类数据分析建模的案例也已屡见不鲜。对于埃博拉疫情传播的分析，也有研究人员给出了他们的方法，其中一种就是通过估计引入风险（import risk）的方法来量化埃博拉对某一地区可能造成的影响。

对于引入风险的估计，最核心的问题便是如何通过动态模型或是统计模型将风险量化。这里介绍一种比较直观简洁的估计方式，首先把引入风险分为相对引入风险（relative import risk）和绝对引入风险（absolute import risk）。不妨假设 X 为疫情爆发区域的某个机场，而 Y 则是世界上任意一块区域，那么我们可以通过条件概率的形式来定义相对引入风险，即 $P(Y|X)$。而对于绝对引入风险我们则可以通过联合概率来定义，即 $P(X, Y)$



[3] http://rocs.hu-berlin.de/publications/ebola/index.html

$= P(Y|X)P(X)$，这里需要注意的是 $P(X)$ 往往远小于 $P(Y|X)$，因此 $P(X, Y)$ 也会远远小于 $P(Y|X)$。在实际应用中，相对风险较绝对风险而言更有价值，其主要原因在于 $P(X)$ 的估计在大部分时间都难以实现，事实上 $P(X)$ 的估计需要依赖于大量的参数来描述 X 所在地区的各类因素，而绝对风险的估计却恰恰依赖于 $P(X)$。与此相反，相对风险的计算则仅仅需要各机场人口流动的数据即可，即无需考虑地区本身的相关因素。

将全球各地的机场数据整合后，人们即可以得到各地点相对引入风险的估计值，下一步则是考虑如何将分析结果向公众展示。显然数据可视化往往是直观有效展示分析结果的最佳方式，有人提供了一种基于 D3 实现的交互式网络分析图表（上图为一个截图，有兴趣的读者可翻阅原图[4]）。

2. 由移动通信数据预测疫情

埃博拉病毒在西非的爆发引起了全世界的关注，人们纷纷把注意力投向了机场等引起人口流动的公共场所，正如上一节所说，机场数据的确对于流行病学家等研究者们而言具有极高的分析价值，但除此之外，基于手机移动端产生的数据同样极具应用潜力。

用户每次利用手机通话的过程中都会同时产生相应的通话记录数据，其中自然也包含了电话号码、通话时间以及大致的通信地点等重要信息。对于运营商而言，这些数据可以为各地基站的部署提供参考从而提升通信网络。另一方面对于城市规划者而言则可以基于该数据判断相关地点是否需要拓展相应的公共交通设施。

然而除了上述这些相对常用的应用外，在流行病学的应用却更令人期待，更激动人心。事实上到目前为止一般情况下对于疾病扩散建模的常用方法依然是基于人口普查的数据以及相关调查。然而对于通信记录数据，人们却可以得到实时更新的数据，也就是说在实际应用中无需估计某地区的人口是否会迁移。同时幸运的是，在近几年中确实并不缺乏类似的成功案例。2009 年在墨西哥爆发的猪流感，研究人员就曾利用通信数据监测公众对于政府发布的健康预警信息的反应。此后 2010 年随海地地震爆发的霍乱疫情，研究人员则同样基于手机通信数据建模并给出了最需要援助地点的最优估计。

在对于埃博拉病毒研究的实际操作中却更为复杂，一个最主要的原因在于西非大部分民众并没有手机或者其他通信设备。不过尽管如此，某种程度上它却依然优于基于陈

[4] http://rocs.hu-berlin.de/D3/ebola/

旧数据的统计分析。事实上研究人员如果可以从一个传染病爆发的地域追踪到人口的流动，那么对于下一个最有可能爆发传染病的地点就会有一个比较有效的估计和预测，从而可以提前展开合理有效的资源配置。遗憾的是，尽管很多相关的机构都做了很大程度的努力，但出于隐私等问题的考虑，电信运营商依然不允许研究人员使用这部分数据。

3. 由社交媒体数据预测疫情 [5]

网络和社交媒体数据，对埃博拉病毒的预警，发挥了重大作用。HealthMap 是一个利用大数据反应疫情的网站 / 应用，它使用一定的算法来抓取来自社交媒体网站、本地新闻和政府网站、传染病医师的社交网络和其他渠道的数据，用于探测和跟踪疾病爆发。2014 年 3 月 14 号，HealthMap 通过自己的系统，预警了几内亚境内爆发的“神秘出血热”。2014 年 3 月 19 号，HealthMap 确认其为埃博拉病毒并对世界卫生组织发出警告，还给出了其在几内亚东南部热带雨林地区传播的粗略地点和路径。2014 年 3 月 23 号，世界卫生组织正式宣布埃博拉疫情爆发并报告了第一个确诊案例。在这时，HealthMap 已经追踪了在几内亚的 29 例确诊和 29 人死亡病例——所有数据和报告都来源于社交媒体和当地政府网站等。

HealthMap 利用复杂的算法，过滤不相关的数据，结合领域内专家的帮助，再对相关的信息进行分类，确定疾病的类型并在地图上定位爆发地点。针对这次埃博拉疫情，在世界卫生组织宣布当天，HealthMap 就上线了专门的页面，其中包含一个实时可交互的地图。全球网友可以通过这个可交互地图来免费了解疫情，其中包括具体的爆发地点和跟踪新的病例和死亡人数的信息。该系统还能够记录公众的关注度。用户可以在地图上放大特定的国家和地区，上面会标记主要病例报告。用户点击标记会指向爆发的新闻报道。同时，在地图底部的滚动条可以让你通过点击关键日期，以追踪病情进展。

这不是 HealthMap 第一次立功了。这个组织成立于 2006 年，由一组研究人员、流行病学家和软件开发人员组成的团队，利用网上各式各样的数据来源，监测和预测疾病爆发，并实现对公共健康威胁的实时监控。他们汇集了各式不相干的数据源，包括网络新闻集中平台、目击者报告、专家策划讨论和官方验证的报告。除了实时和可交互的呈现数据，HealthMap 也致力于预测疾病风险。曾经有报导称，该组织成功使用 boosted regression tree 等模型成功预测了 SARS 在中国境内爆发的死亡率。

HealthMap 的官网宣称，他们主要的数据来源是 ProMED（一个国际传染病协会，成员主要为一线医生和研究人员）邮件列表、世界卫生组织官网，GeoSentinel（来自国际旅行医学协会和美国疾病预防控制中心的临床医生以个人身份的检测）、世界动物健康组织官网、联合国粮农组织、EuroSurveillance（欧洲地区以同行评审为目的的传染病监测和交流的信息平台）、Wildlife Data Integration Network（一个全球野生动物基本新闻源）、谷歌新闻搜索、百度新闻和搜搜资讯。另一份公开发表的论文显示，大部分数据来自 ProMED (61. 58%)，谷歌等其他搜索引擎新闻则贡献了 25. 24%，除此之外比较重要的来源还包括 RSS 订阅 (12. 11%)，推特等社交媒体 (8. 7%)[6]。这和新闻中极力鼓吹的完全通过社交媒体预测埃博拉疫情，似乎还是有一定的差距。这个社交媒体并不是大众所以为的普通公开社交媒体，而是全球一线医护人员建立的社交网络。事实上，谷歌以及其他一些社交媒体都试图通过抓取网络关键词来监控和预测疾病信息，但是并没有取得如此好的效果。谷歌曾经宣称自己的系统很好地预测了美国每一季的流感爆发，而实际数据显示，他们的系统常常高估了患病率。普通人对疾病的感

[5] http://www.dailymail.co.uk/sciencetech/article-2722164/Ebola-flagged-computer-software-nine-days-BEFORE-announced-HealthMap-used-social-media-spot-disease.html

[6] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4198292/