18 AI 防疫

帝国理工郭毅可：抗疫政策的评估，数据科学的应用

整理：智源社区 蒋宝尚

在第二届北京智源大会的“AI 防疫”专题论坛上，英国皇家工程学会院士、英国帝国理工计算机系教授郭毅可做了题为《抗疫政策的评估：数据科学的应用》的报告。

在报告中，郭毅可详细介绍了如何根据数据建立数学模型预测疫情。郭教授表示：传染病学、流行病学是一门标准的数据科学，同时数据科学并不一定非常复杂，换句话说，针对传染病的数据科学研究都是比较直观的。

在报告中，郭教授还具体介绍了动态和静态两种传染病模型的细节，也对之前的建模预测效果进行了展示，同时对中国和欧洲的两种抗疫政策进行了简单评价。

以下是文字整理：

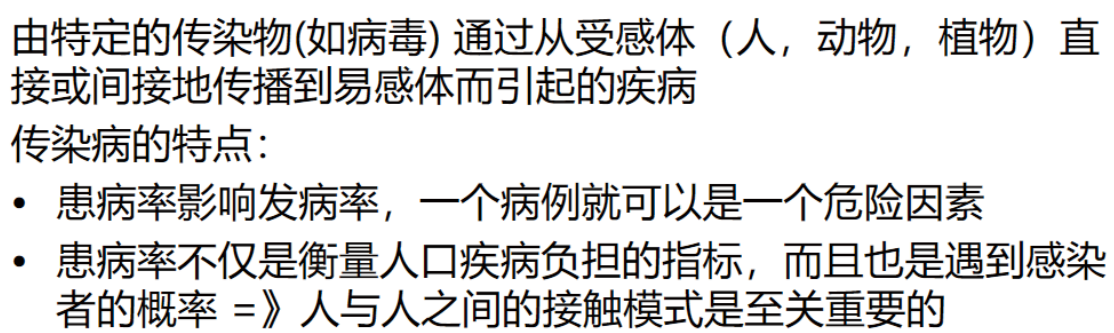


图 1：什么叫流行性传染病?

今天演讲的题目是《抗疫政策的评估：数据科学的应用》，我先从流行传染病开始谈起。传染病与普通疾病的最大区别在于：普通疾病的患者影响自己，传染病由于其传染性在，会进一步影响患病率，而患病率影响发病率，一个病例就可能是一个危险因素。另外，患病率不仅是衡量人口疾病负担的指标，而且也是遇到感染者的概率，其和人与人之间的接触模式高度相关。

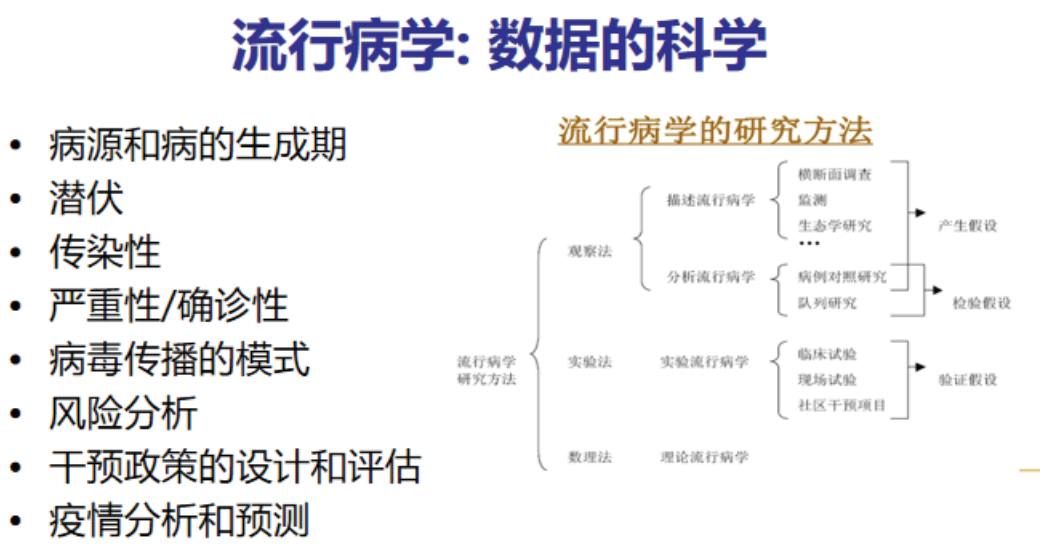


图 2：流行病学是一门关于数据的科学

因此，传染病一旦出现，没有患病的人们关心自己会不会受到传染，政府则关心如何让更少的人被传染。这时，首先需要进行的是数据科学研究，即通过观察有多少人患病（例如进行关于病毒的核酸检测），判断病毒的潜伏期、传染性和传播模式。政府也应出台相应的防疫政策，虽然政策能在一定程度上限制传染病的流行，但却给经济发展带来了风险，这时则需要对政策进行全面的分析和评估。因此，传染病学、流行病学是一个标准的数据科学。

一、静态传染病模型：求解概率方程

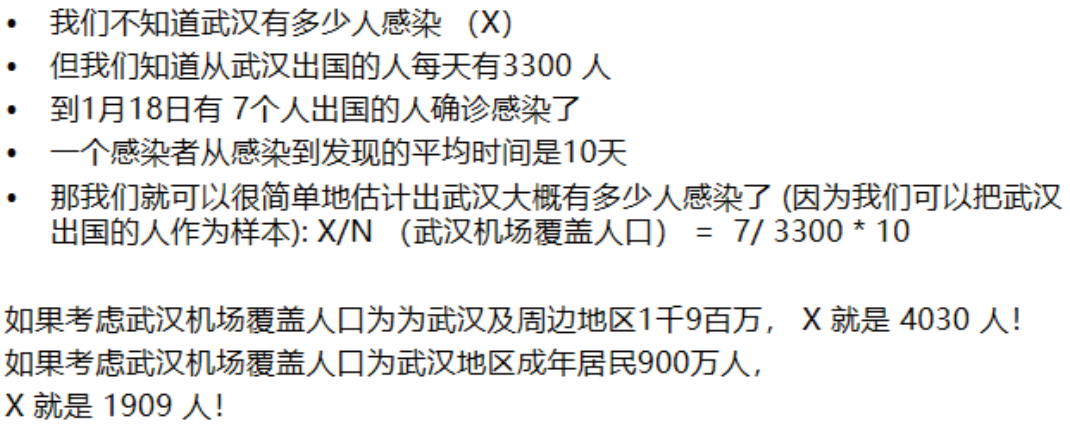


图 3：帝国理工对武汉疫情的早期预测

数据科学并不一定复杂，换句话说，针对传染病的数据科学研究都是比较直观的。例如帝国理工对武汉疫情最早的预测：我们不知道武汉有多少人感染，但我们知道从武汉出国的人每天有 3300 人；到 1 月 18 日有 7 个出国的人确诊感染了；另外，一个感染者从感染到发现的平均时间是 10 天，那么我们就可以简单地估算出武汉大

致感染的人数。因为我们可以把武汉出国的人作为样本，通过求解概率方程式就能得出感染人数。

具体而言，如果考虑武汉机场覆盖人口为武汉周边地区 1 千 9 百万，那么感染人数就是 4030 人，如果考虑武汉机场覆盖人口为武汉地区成年居民 900 万人，那么感染人数就是 1909 人。

图 4：2020 年 1 月 28 日湖北省新冠肺炎疫情情况

这个简单的预测非常有价值，因为当时抗疫政策的制定必须要知道感染的规模。该预测结果也在 1 月 29 日得到了验证，根据当时卫健委的报告，湖北省有 3500 人感染，武汉有 1900 人感染，由此可见当时的预估是较为准确的。

二、动态传播模型：用流动数据讲述生命

图 5：传播动态系统数学模型的发明者

上面是简单的数据科学，下面介绍复杂一些的动态变化。关于动态变化，传染病学最重要的一个工作，是所谓的 SIR，即传播动态系统的数学模型，由两个物理学家发明。此传染模型带有普遍性，是一个 General Model，其概括了传染过程中的三个人群：易感者、感染者，恢复者。

先来看一个简单的动态模型，即假设治愈的患者能获得“终身免疫力”，不再传染其他人； 再者，还要去除干扰因素。这时所研究的对象是两个部分，一是从易感者到感染者的传播速率，二是从感染者到免疫 / 死亡者的周期。针对这两部分，动态模型会描述三个状态： 易感状态、感染状态、治愈状态。

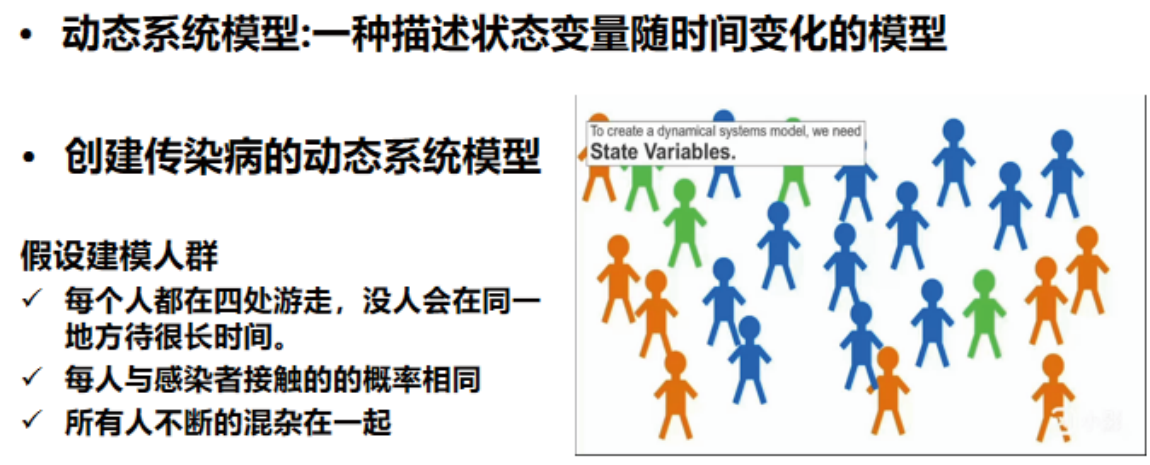


图 6： 传染病动态系统模型

另外，此动态模型对人群也有三个假设：1. 每个人都在四处游走，没人会在同一地方待很长时间；2. 每人与感染者接触的概率相同；3. 所有人不断地混杂在一起。

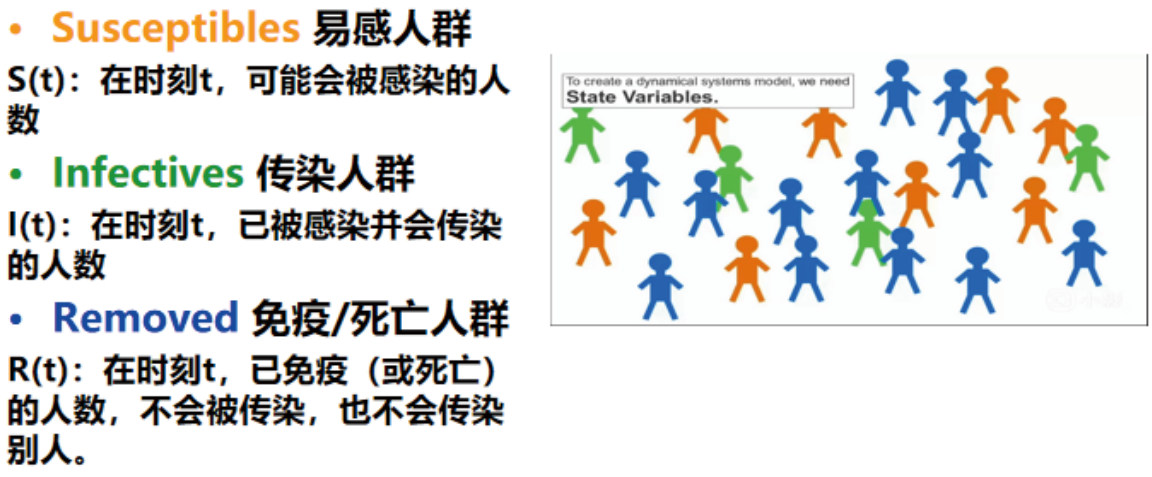


图 7： 三个动态变化的人群：三种状态变量

三个动态变化的人群对应三种状态变量，其中易感人群： 在某时刻可能会被感染的人群； 传染人群： 在某时刻已被感染并会传染的人数； 免疫、死亡人群： 在某时刻已免疫（或死亡）的人数，不会被传染，也不会传染别人。

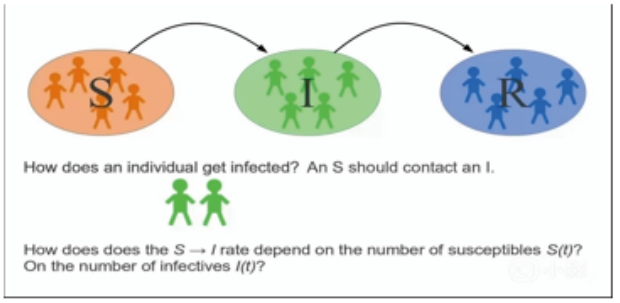
有了假设、变量之后，然后根据 SIR 模型找出动态变化规律，即需要确定个体状态如何转换、两个转化速率:

不感染到被感染、感染到治愈。

- S→I rate取决于什么?
- ✓ 单一个体被感染的可能性与I(t)感染人群数量成正比
- ✓ S→I rate也包含任意易感染者S(t)的影响
- ✓ 个体被感染的可能性与S(t)和I(t)的数量都成正比。
- 所以，在时刻t

$$S \rightarrow I \text{ rate} = b\,S(t)I(t)$$

*此处b为一个感染速率参数



- ✓ 参数b包含2方面影响：有效接触的可能性和接触导致感染的可能性
- ✓ b数值大对应人群比较活跃的社会互动，疾病高度传染的。
- ✓ b数值小对应隐居的群体，疾病难以传播。

图 8：感染速度：S → I rate

以上就是基本模型的基本思想，虽然会涉及一些数学，但非常简单。例如从易感到被感染取决于什么？当然取决于被感染人数，如果感染人数很多，或者很多人被感染，感染速度就会加快； 或者被易感暴露的人群很多，那么感染速度也会加快。

此外，还会涉及到一个参数，即针对速度的参数，其受两方面的影响，一是，有效接触的可能性； 二是，接触导致感染的可能性。也就是说，如果对应人群比较活跃的社会互动，那么参数就大； 如果对应隐居的群体，疾病难以传播，那么参数就小。

- I→R rate取决于什么?

只取决于感染的数量 I，I→R rate与I成正比

- 所以，在时刻t，

$$I \rightarrow R \text{ rate} = a*I(t)$$

*此处a为一个恢复速率参数: 如何让感染得不再感染了

图 9：恢复速率：I → R rate

上面是感染的速率，那么恢复速率由什么决定呢？其实，只取决于一个变量，就是感染的数量。

假定不考虑采取的医疗措施（因为医疗可以加快治愈率），感染的数量与恢复速率成正比。但是数学表达式中也

会有一个回复速率参数，取决于如何让之前感染的不再感染了。这里有两种方式，一种是治好，一种是死亡。

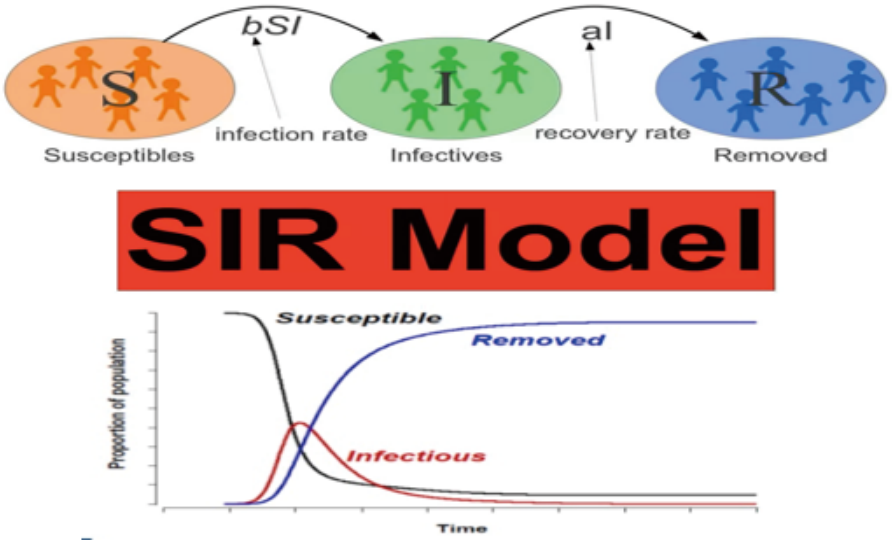


图 10：SIR 模型及转化速率

其实，这一非常简单的传染病模型预示着会有“疾病峰值”的出现： 刚开始的时候随着时间的推移，很多易感人群被感染了，感染率就上去了，到了一定规模以后所有人都生病了，没有人可以感染了，感染率就会下降，随后出现的都是治愈病例。其实，所谓的群体免疫基本上符合这种思想。

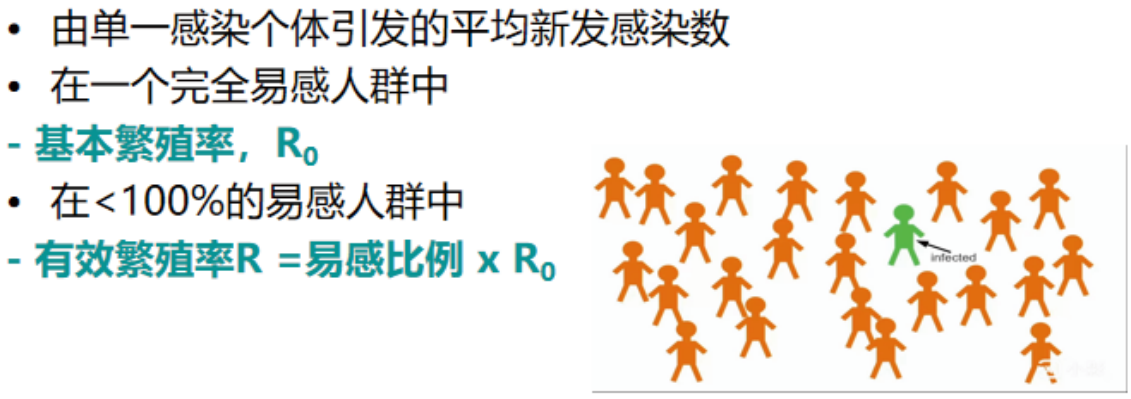


图 11：基本繁殖率

有了这个模型以后，其实可以做很多事情，其中一项非常重要的工作就是实时模拟各国现在处于各种各样干预政策中的传染率。这里会涉及一个重要的参数，叫做基本繁殖率 R0，即单一个体可以引起多少个新的感染者。

如果没有干扰政策，此 R 就是基本繁殖率，如果存在干扰政策，那么模型的易感比例就会改变，此时的 R 叫做有效繁殖率。

如果 R0 大于 1，那么传染模型就是指数型，R0 等于 1 则感染保持恒定，R0 小于 1 则感染消失。

对于一种直接人传人的病原体

$$R_0 = \beta c D$$

其中

β:每次受感染者和易感者之间接触后，病毒的传播概率
c: 接触率
D:感染持续时间

图 12：基本繁殖率的决定因素

R0 由什么决定呢？基本上是三个参数：1. 每次受感染者和易感人群接触后，病毒的传染率（因为接触不一定传染）；2. 接触率，接触的频率；3. 感染持续时间。而 R0 的数学表达式则是由这三个参数相乘。

具体而言，传染率如果是 0.15，接触的频率是 12，感染的时间是一周，那么基本繁殖率是 1.8。

三、干预政策分析：中国和欧洲对比

如果有了干预政策会改变什么呢？改变有效繁殖率，即把易感的人群控制住，把受感染的人群控制起来。

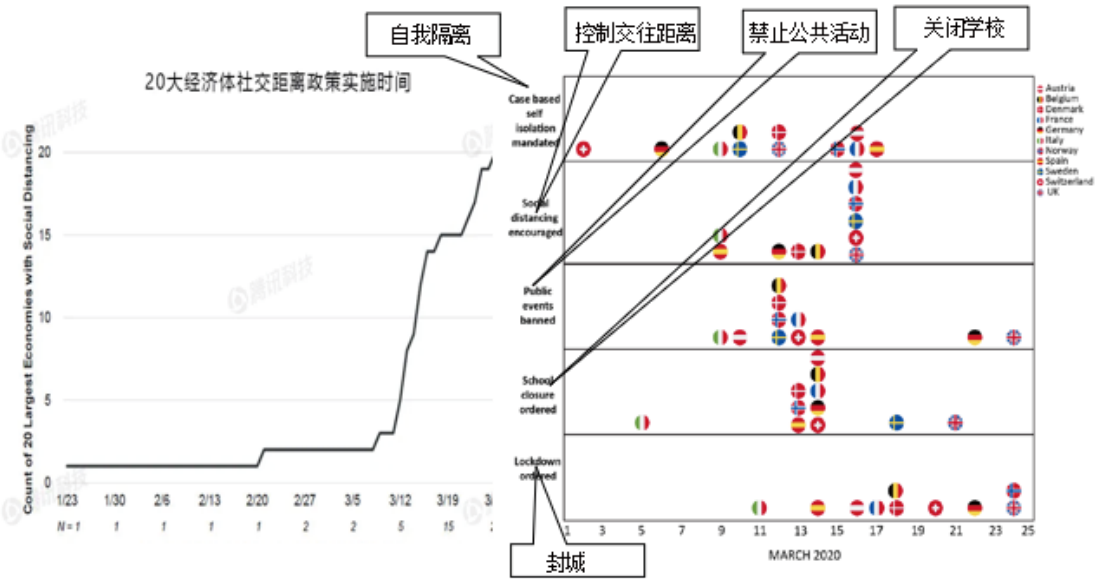


图 13：各国的干预政策

下面讲一讲对各国干预政策的研究。疫情爆发之后，二十大经济体社交政策各有不同。例如中国非常严格，瑞典、英国宽松。所有的政策都有其道理，即如何平衡生命与生活。举个例子，过马路有被车撞死的风险，但不能因为这个风险把交通封锁。因此，生命和生活平衡点的把握非常重要，如果照顾了生命死亡率和患病概率，那么人民可能在经济和人民生活上会有巨大损失，且其他原由的死亡率就会增加。

所以，这就要求我们仔细分析干预政策，然而定性分析无法得到想要的结果，最好是用定量分析，能够“互相比较”。定量之后，那么就能够将经济损失纳入模型，则会得到优化版本。当然，听起来非常科学和理性，但有的时候也很残酷，因为要考虑的因素非常多。

有多少干预政策呢？大概有自我隔离、控制交往距离、禁止公共活动、关闭学校，更严酷的就是封城。

图 14：通过人群流动的数据来决定干预政策

这二十大经济体中，干预政策的实施，在各国的基础是不太一样的。例如中国的有利条件是健康码，它可以帮助政策实施到细微处，即可以通过人与人接触的模式和个体的运动轨迹建立整个干预政策。所以这也是我们国家恢复正常秩序后没有出现太大疫情反弹的原因之一。毕竟，如果有一个人被确诊，那么其整个交往人群都能获知，则可以通过溯源对其进行控制，这样就不需要采取特别严控的政策来进行社交距离或者封城，经济可以更加放开一些。

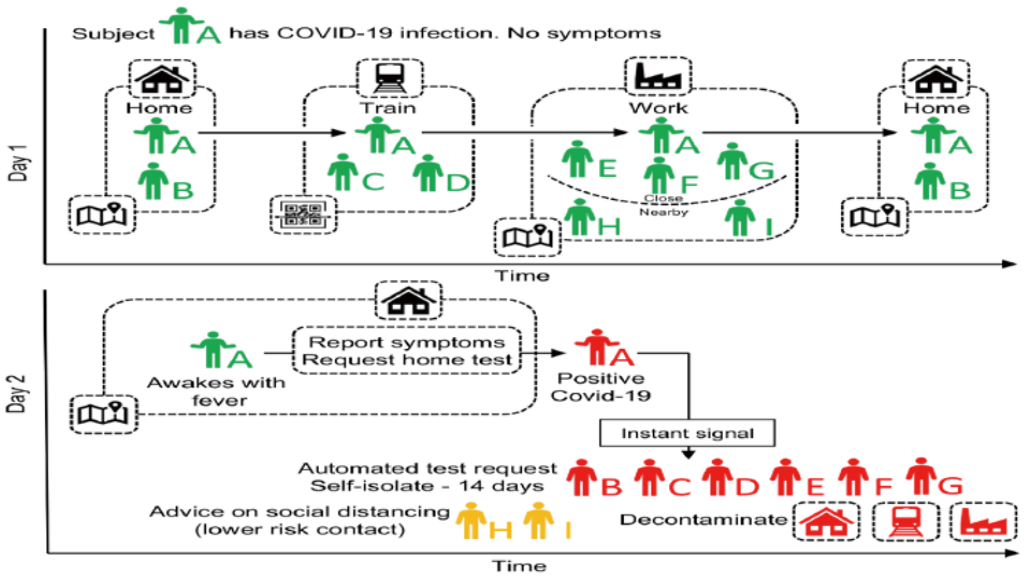


图 15：通过了解人与人接触的模式来优化干预

这种模式也存在缺点，即牺牲了个人的隐私。现在有很多研究都是希望通过在不影响隐私的情况下能够完成工作，例如通过 WiFi 可以建立联系基础（contact base）。

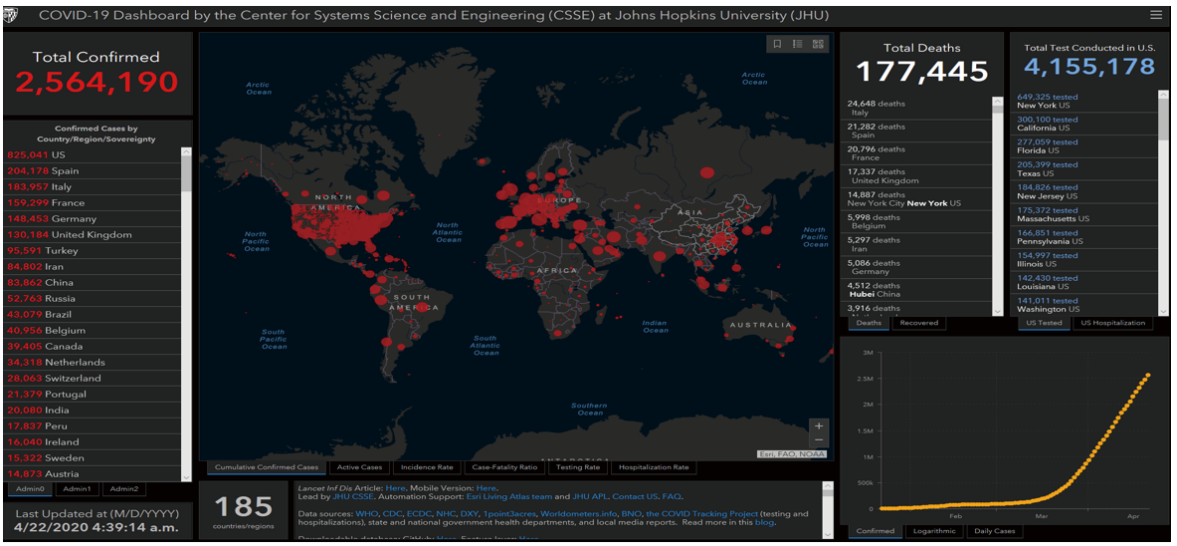


图 16：全球疫情网站

有了这些政策之后，便每天看上述网站，了解一下多少新增患者、新增死亡人数。最主要的是通过观察数据反推各国的政策有效性。

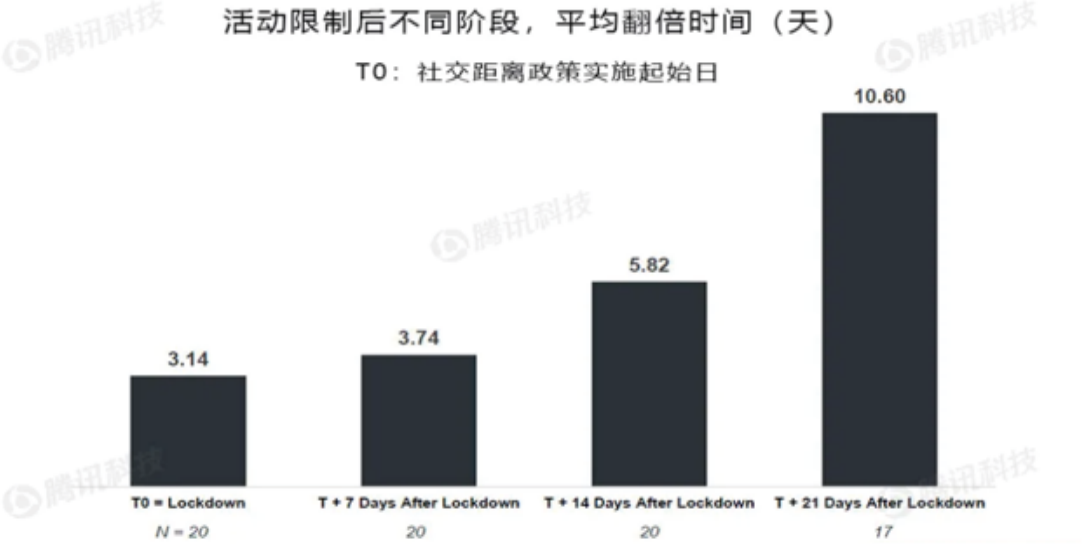


图 17：数据反映的政策干预结果：平均翻倍时间

其实，干预是有效果的。如上图所示，干预结果通过平均翻倍时间表现出来，即从 4000 人增长到 8000 人花费了多少时间。其实，封城之前基本上三天翻一倍，封城之后大概二十天才翻倍。这大大降低了病毒的传播率。

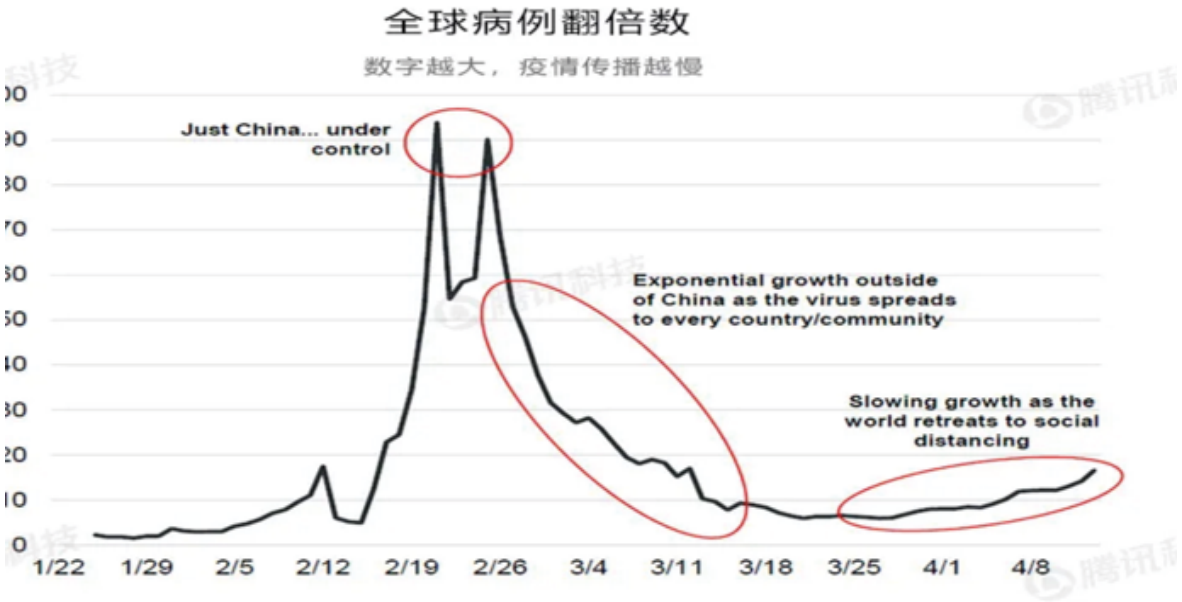


图 18：数据反映的政策干预结果：全球病例翻倍数

如上图所示，通过大规模的观察也可以了解这一现状。开始的时候“翻倍”花费的时间很短，大概为十天左右。2 月 1 日武汉封城，翻倍的时间增加到了 90 天。3 月初的时候欧洲爆发，翻倍的时间又缩短了。当然，这种衡量干预的方法都比较初级，并没有看到一个国家和一个城市的变化。

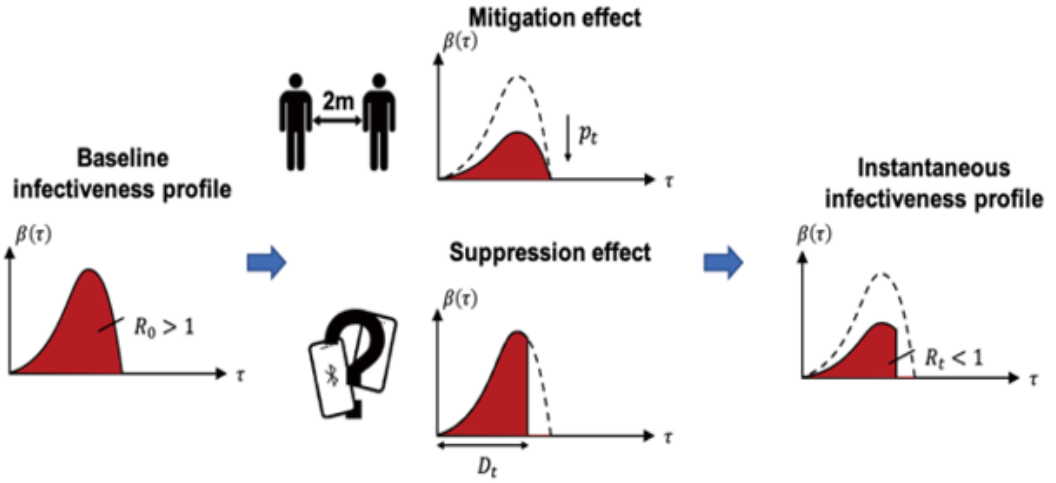


图 19：精确地量化干扰政策的作用

因此，为了精确地量化干扰政策的作用，我们用模型进行了模拟，模拟的结果就是 R 繁殖率（几何上表示的 R 就是传播曲线下的面积），即一定的时间内一个人能传播多少人。当然，传播的概率就是参数 B（Beta），下面的面积就代表传播的人数。

我们把政策的作用分成两大类：一类就是 S 到 I，只有缓和系数，就是我把 B 参数 降低下来，降低它的传染率。怎么降低？用保持社交距离、戴口罩等干扰方法。第二类是阻断，改变 I 到 R，即让 Dt 缩短（如果发现感染者，立即隔离，阻断其传染率）。

实际上任何一个政策都有两面性，既有阻断传染期的效果，也有降低传染率的效果，所以组合起来就是可以计算这个 R（繁殖率），希望能够小于 1。

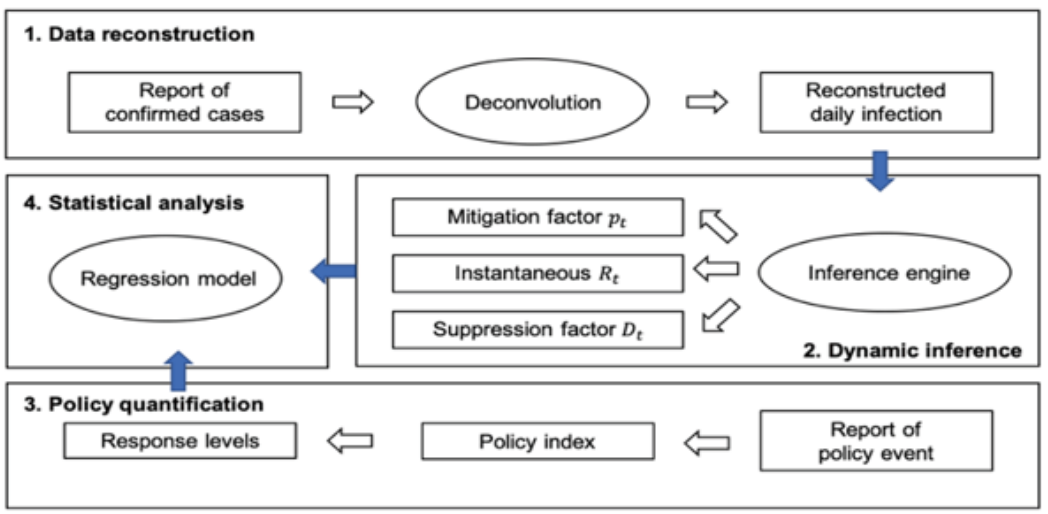


图 20：量化系统示意图

其实，整个观察以及模型结果就是报告的确诊病例。其实我们不可能知道这个国家到底有多少人感染了，除非做全民的测试，否则只能预测。那么如何从确诊病例预测全部感染者规模？我们的做法是把时间向后推，然后重建传染模型。

随后，就可以用贝叶斯的理论推断三个参数，一个参数是缓解系数（Mitigation factor），另一个是有效繁殖率（instantaneous），最后一个是阻断系数（Supperssion Factor）。

有了这些以后，就可以构建前面提到的 SIR 模型，从而把它和政策进行结合分析。即回答一个政策到底影响了哪些因素这些问题。

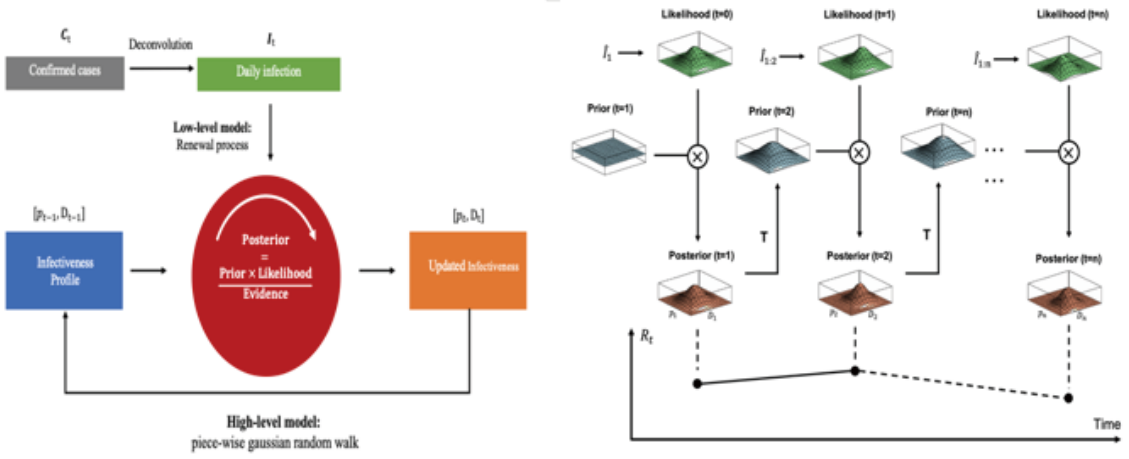


图 21：用贝叶斯估计来计算 Rt

整个模型最难的部分是用贝叶斯进行计算 Rt。毕竟，我们要求的对 RT 的估算是每天实时进行，所以由于政策具有延迟性，导致推算变得异常艰难。

这里我们采用的思想是：能够观察到的都是发生在过去的，所以还需要往后推。具体而言，我们通过贝叶斯的方法，通过顺序的学习，即根据前几天算出来的 RT 改变后验概率，也就是期望改变参数，把后验概率最大化。综上，整个过程等于是进行了一个顺序的贝叶斯推理。强调一下，我们并没有使用神经网络，而是运用贝叶斯的高斯过程。

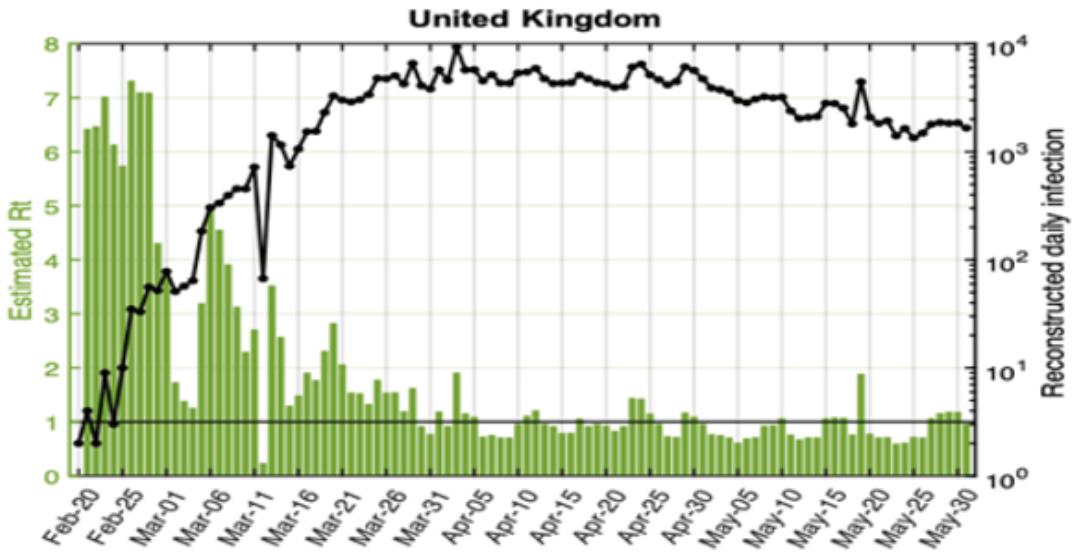


图 22：揭示伦敦的 Rt 变化

运用上述推理过程，就可以得到一个非常漂亮的关于各个城市的描述，如上图所示伦敦 RT 的变化，黑线就是根据确诊病例估算出的感染，可以看到伦敦到了 4 月 5 日以后的 RT 基本上在 1 以下，所以实际上传播已经开始控制住了。

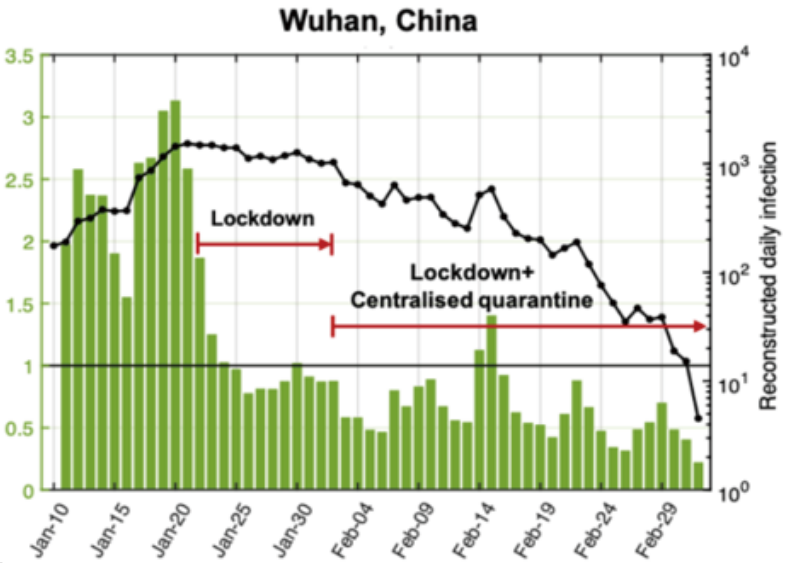


图 23：揭示武汉的 Rt 变化

武汉则更明显，封锁之前 R 值非常高，春节的时候 R 达到了 3。23 日封城以后 R 值迅速降低（因为采取的不仅是封城政策，还不让上街）。因此，1 月 23 日到 2 月 15 日中，实际上新患病的很少，几乎所有的病例都是前面感染的，只不过那个时候爆发了症状。

如上图所示武汉的 Rt 曲线有一个跳跃，因为武汉更换了市长，市长所采用的防疫政策是将过去的所有疑似都当成“确诊病例”进行控制，这作为一个人为的动作，发挥了很大作用。随后加上方舱医院的使用，可以看出武汉从 2 月 3 日疫情得到控制。这里我们也可以明确了政策对于整个抗疫的影响。

esponse	Representative Measures	Impact of measures R_t relative reduction	Suppression effect D_t relative reduction	Mitigation effect p_t relative reduction
evel 0 inimal response	No mandatory restrictions	0	0	0
evel 1 oft response	Closing schools, International travel controls.	35% CI:[25%, 45%]	22% CI: [17%, 27%]	29% CI: [18%, 38%]
evel 2 trong response	Cancel public events, Restrictions on gathering, Restrictions on internal movement.	60% CI:[54%, 65%]	26% CI:[21%, 30%]	56% CI:[50%, 61%]
evel 3	Close workplace, Close public transport, Stay-at-home requirements.	71% CI:[68%, 74%]	37% CI:[35%, 40%]	67% CI:[64%, 70%]
evel 4 mergent response		74% CI:[71%, 77%]	40% CI:[37%, 42%]	70% CI:[66%, 73%]

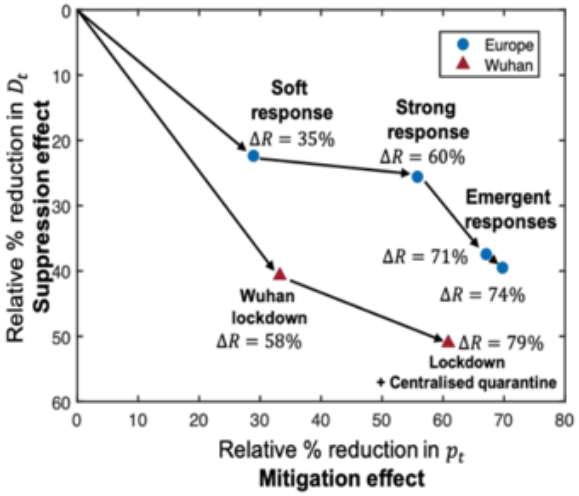


图 24：欧洲政策的评价结果

我们也对欧洲的所有政策进行了评估。其实，欧洲和中国采取两种不同的政策，欧洲的政策基本上没有强制性的封城，但也关闭了学校、工作场所，控制了交通。其实，每个政策我们都对应 R 的变化。其中，Soft Response 可以降低 30%，Strong Response 可以降低 70%，所以还是非常有效果。而中国的做法截然不同，因为中国采取的是非常严格的政策，“一下子”就把 R0 减掉 79%–80%。

因此可以看到各国政策的评价是不一样的。当然，我们没有分析具体“成本”，如果这个评价再加上“成本”，然后进行比较，可能就比较有意思了。

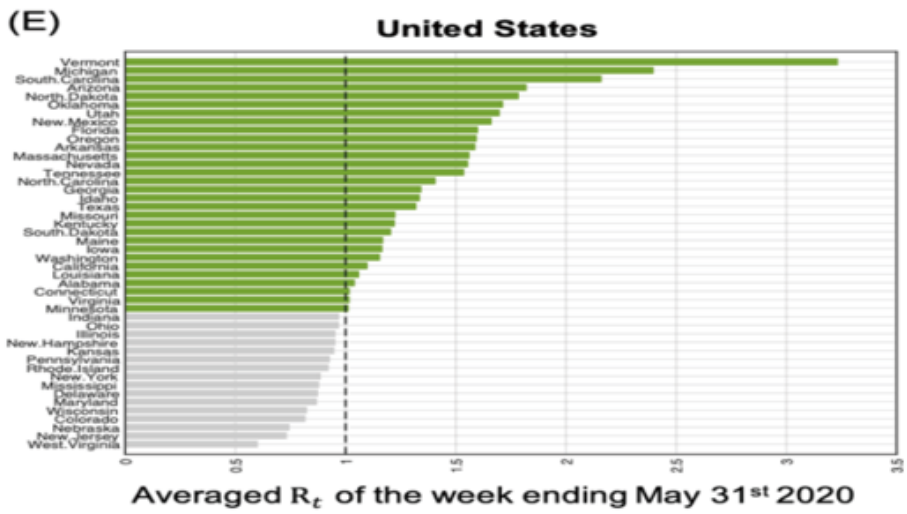


图 25：美国疫情的第二次爆发

RT 确实在一定程度上能代表预测性，例如上图我们在 5 月 31 日计算的美国 RT，那个时候就看到美国毫无疑问在面临着第二次爆发，因为其 Rt 的数值在 30 个州全部超过了 1。

由于当时 R 的效果在大概 10 天后才能观察到，所以到了 6 月 12–13 日的时候，30 个州当中的 29 个州都有第二次疫情爆发，其中 5 个州是高爆发（Record High），因此预测比较准确。

四、总结：模型参数改变，背后是人命代价

总结一下，将人工智能、数据科学应用在传染病防控当中，效果非常棒。什么叫做人工智能？就是用物理来描述一个世界，把它变成变量之间的转换，然后通过观察得到数据，通过观察推断模型，这才是真正人工智能的伟大意义。

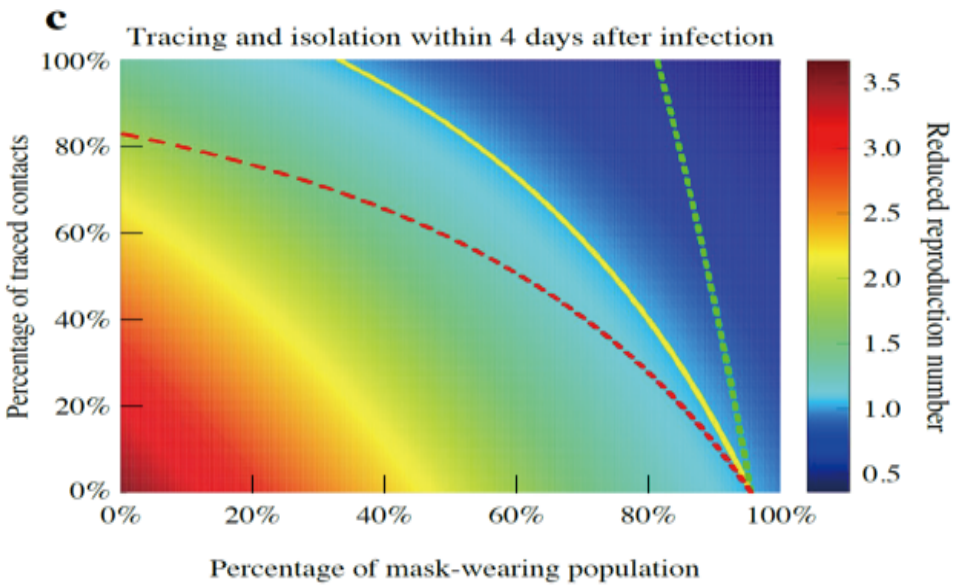


图 26：干预与否：改变的是模型参数

在建模过程中，我们需要更多的是关于模型的定义和模型参数的学习，例如上图戴口罩的模拟，通过戴口罩改变 R0、RT 的数值。

其实，干扰与否对于数据科学和人工智能来说，可能就是改变一些模拟的参数。但是对于社会，其背后的代价会是生命。

图 27：干预与否：代价是生命

如上图 4 月 20 日的数据，香港和纽约的面积差不多大，香港感染了 1000 多人，一共死了 4 个人，纽约 4 月 20 日的死亡是 1.3 万人，而整个大纽约最近的一个数字 3 万多，如果只看纽约城（New York City）这一数字也达到了 1.8 万。

那么如何评价强烈干预的政策？其实，这个问题是可以讨论的，如果是像香港这样非常强的干预，到今天为止还是封闭状态，与全世界不能正常交往。那么，这对香港的 GDP、经济、民生的副作用是非常大的。因此付出的代价换来了低死亡率，是不是完全合理？这个问题也是可以讨论的。

最好的做法是在生命和经济当中取得平衡，即在保证生命救助同时，也能保证生活。这种平衡也是我们想要的未来社会，要让城市有免疫力也要有抗灾力，如果有一套规矩，能够指导制定平衡政策，那么这座城市就是一个真正的智慧城市。

这一次我们对智慧城市的定义和理解有了很大的进步，以前我们理解智慧城市就是交通不堵，污染更少等等。实际上这些东西都是皮毛，真正一个城市的智慧在于能够抗灾，能够免疫，能够保证人民的生活，能够保护人民的健康，这才是最重要的。

人大教授文继荣：疫情突如其来，大数据人工智能如何沉着应对？

整理：人大高瓴人工智能学院

在 2020 北京智源大会“AI 防疫”专题论坛中，中国人民大学文继荣教授分享了题为《大数据人工智能技术 + 疫情防控：经验和反思》的报告。

文继荣，中国人民大学信息学院院长，中国人民大学高瓴人工智能学院执行院长，智源首席科学家。主要研究方向是互联网搜索与数据挖掘等，获得过 50 多项美国专利，同时在国际著名会议和期刊上发表了一百多篇论文，也是信息检索领域主要期刊 ACM Transactions on Information Systems(TOIS) 的副主编 (Associate editor)。

2020 年春节期间，新型冠状病毒疫情爆发，在智源人工智能研究院的部署下，文继荣教授和许多研发人员都参与到了北京市疫情防控的相关工作中，并研发出了新型冠状病毒肺炎防控智能追踪服务系统和时空足迹近距定位系统，为北京市的抗疫工作做出了一定贡献。报告中，文继荣教授分别介绍了两个系统的主要原理及功能。

一、新型冠状病毒肺炎防控智能追踪服务系统

该系统主要是通过汇聚用户散布在各个互联网公司的片段式数据，利用用户在移动互联网的“数字脚印”，从而围绕风险人员绘制轨迹信息，利用算法对高危人员进行分析、追踪与筛查，最终确定密切接触人员名单，达到快速定位到重点防范区域、防范人群的作用。自 1 月 28 日启动使用到 2 月 2 日运行生成首批数据，该系统已经持续发现疑似接触人超过 10000 名，并且能够根据人工智能算法自动生成活动报告，目前该系统仍然在持续为北京市的防控工作提供保障。

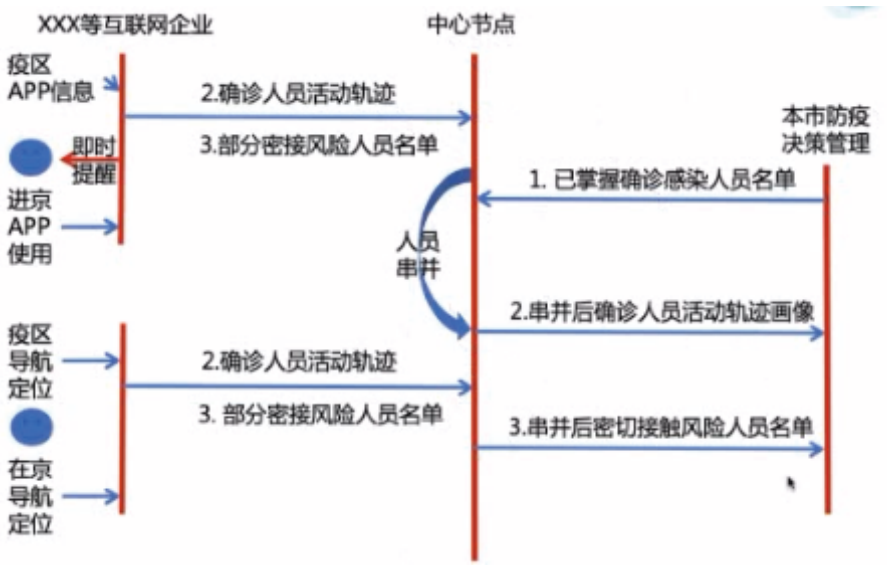


图 1：基本数据的处理流程

图 2：系统首页示意图

二、时空足迹近距定位系统

该系统是解决 GPS 在室内的定位问题，可以实现精准定位到城市中每个房间的风险防控服务，通过蓝牙、局域网信号、气压计等信号，对人员时空轨迹数据进行采集，追溯近距离接触人员，借助大数据和人工智能技术更精准的进行疫情防控，从而降低隔离成本，为企业复工复产提供支持和保障，因此，这个系统最典型的应用场景便是人员较为密集的室内办公场所。

另一方面，文继荣教授还提出了在此次疫情防控过程中所产生的三点思考:

1. 下一次公共卫生事件突发，我们能不能做的更好?

在应对此次新冠病毒疫情的工作之下，文继荣教授发现了暴露在疫情下的“数据孤岛”、“数据安全性”、“数据动态性”和“决策及时性”等问题，疫情虽然尚未结束，但我们已经要开始思考应对此类事件的解决方法和措施，保证数据信息的全面准确等，为未来的挑战做好准备。

2. 完全的隐私保护是不可能的!

文继荣教授以苹果谷歌联合推出的疫情追踪系统为例，虽然苹果谷歌在操作系统层面上进行设计该系统，并且声明不会泄露用户隐私信息，但是文继荣教授针对现实生活中人们获取隐私数据的一种经典方法的变形，提出了“完全的隐私保护是不可能的”这一见解。

3. 公共利益和个人隐私的让渡。

随着我们的世界从物理世界逐渐过渡到数字世界，在网络世界中会逐渐分化出另一个包含真实性格、记忆等的数字分身，而当用户的数据和信息在互联网中留下痕迹后，文继荣教授表示我们其实已经在不知不觉中让渡个人隐私，而对于公共利益和个人隐私的冲突问题，文继荣教授表示需要政策和法律来界定边界，并且需要公开进行讨论该问题。

北大教授宁毅：科学、精准的公共卫生 - 应对慢性病和传染病双重威胁

整理：智源社区 贾伟、王静

在智源大会“AI 防疫”专题论坛上，北京大学宁毅教授做了主题为“科学和精准的新公共卫生应对慢性病和传染病双重威胁”的报告。

宁毅教授是北京大学公共卫生学院教授、美年大健康首席科学家、公共卫生专家，并且在弗吉尼亚做过医学院的助理教授、博士生导师，也曾在跨国公司 GSK 任职。

报告中，宁毅教授提到，全球疫情不太乐观，如果能够较好地防控外源性输入的病例，中国境内疫情反复的压力就会急剧降低。这种防控范围较大，在斩断传播途径这样的过程中，每个个人、家庭以及单位发挥着重要的作用，甚至是主战场的作用，国家教育、健康促进、家庭行为改进等等，这些是保证生活和防够保持常态化的关键。相对来说，应当把防控做得更精准，可以局限到某些人群或地域进行防控。

以下演讲正文——

从去年一月份开始，一直到今天，疫情的变化每天都发生着戏剧性的变化，这吸引着我们每一个人。今天，我把近一段时间对于过去的一些回顾，以及我们对未来的一些思考与大家进行讨论。

一、疫情概况

首先看一看今天（2020 年 6 月 23 日）疫情的状况，全球如今有大约 907 万感染者，死亡 47 万人。过去的一个月，大家可能对这个数字已经麻木了，但我们还是需要去想，这样的数据究竟反映了什么？

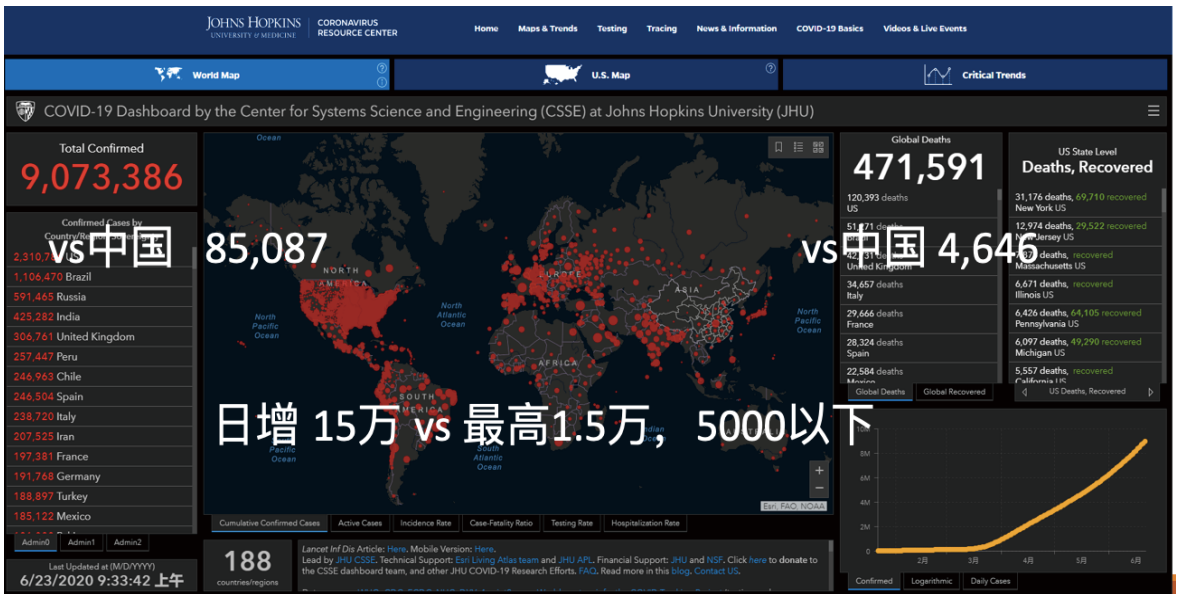


图 1：6 月 23 日，世界范围内的疫情对比（中国 vs 世界）

比较一下中国。中国在整个疫情期间一共有 8 万多人感染，但现在全球感染者已经是这个数字的 100 多倍了。按照这样的发展趋势，到本月底，感染人数将突破 1000 大关。从死亡人数上看，中国死亡人数是 4646 人，全球死亡现已超 47 万人，是中国死亡人数的 100 多倍。从新增感染人数上来看，目前整个世界还仍以每天 15 万人的速度增加； 而中国感染人数最高的一天达到 1.5 万，之前与之后大致都在 3000–5000 人之间变化，后来稳定在 1400–1200 人变化。从这个对比中，可以看出，现在全球疫情的压力非常大。之前我曾经讲过，中国能不能消灭或者消除这个疾病，不取决于中国自己，而取决于全球的疫情控制情况，所以到目前为止我们可以很肯定地说，这个疾病会长期和我们共存。

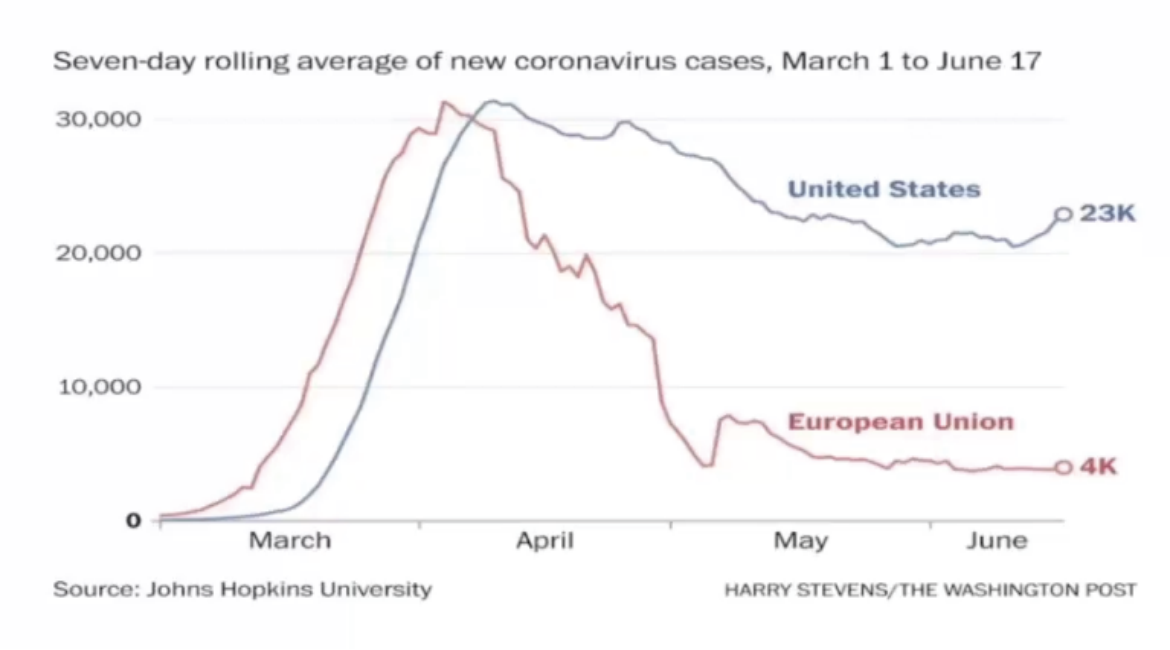


图 2：美国和欧洲的比较：相似的基础，不同的结果。

各个国家都采取了不同的措施，欧洲和美国在基本条件相似的情况下，两个国家的患病率和发病情况却表现得非常不一样。

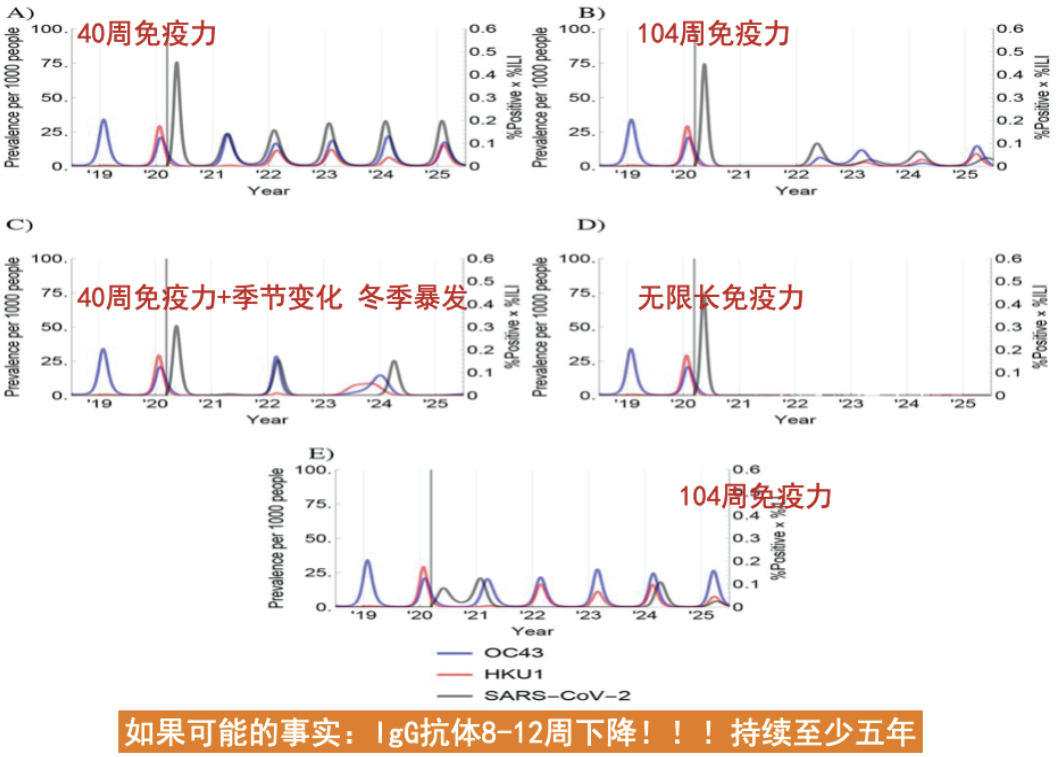


图 3：各种情况下新冠流行的预测

哈佛大学的研究小组进行了一个预测，即根据人们在产生病毒抗体之后能够维持的时间来预测疫情未来的发展。从现在的数据来看，真实情况更接近于左上角的数字变化，也即感染之后有 40 周的免疫力，预测结果是： 到 2025 年，新冠疫情仍会流行。这里提到 40 周的免疫力，也就是免疫力能够保持 10 个月； 但今天有些报道中又提到免疫力维持的时间只有 8–12 周，说明我们的抗体水平明显下降了，未来三到五年疫情很难看到明显的控制。

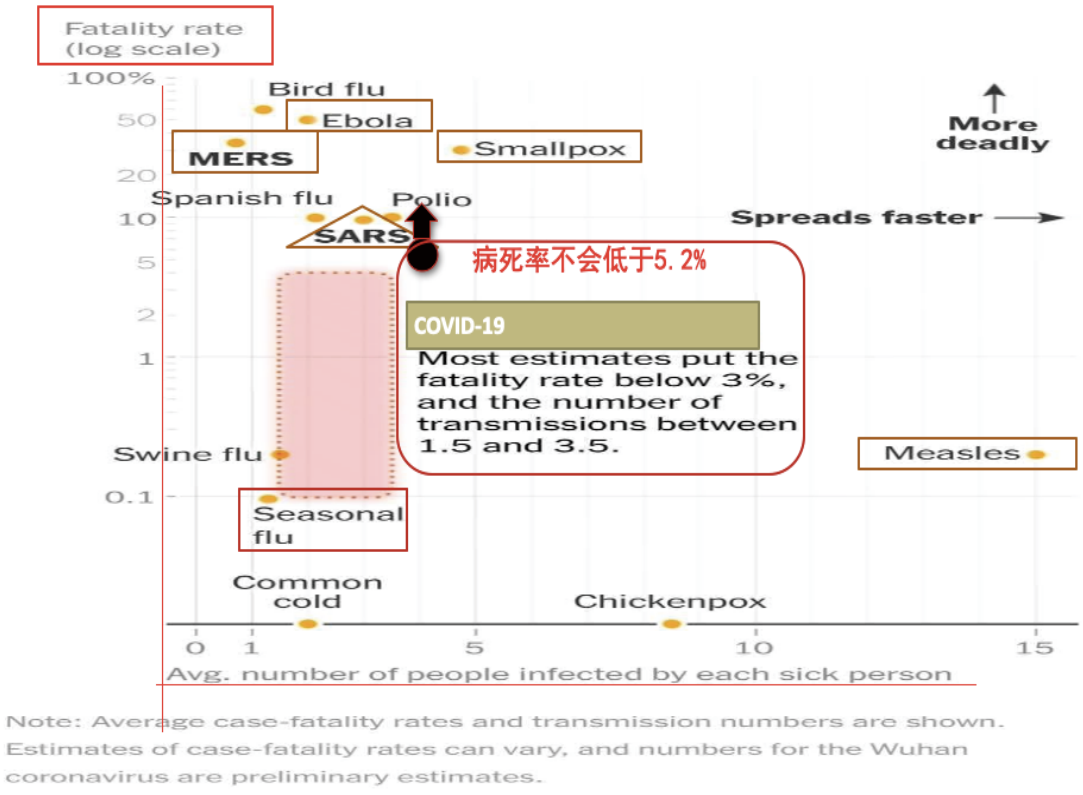


图 4：曾经和现在认知的基本参数：不支持中国群体免疫（国外不得已）

现在国内面临的最大争议是，中国要不要做群体免疫。我们必须回顾一下和其它疾病的比较。图 4 中横轴表示 R0，也就是传播系数，过去我们认为 COVID–19 的传播系数大约是 3–4，但现在来看，人口密集的情况下可能会达到 5.7。如果做出很好地控制，R0 可以接近 0。但 0 是有前提条件的，就是需要严格防控。我们曾经认为这是一个大号的流感。流感在美国的病死率是 0.1%，但目前 COVID–19 在美国的病死率是 5.2%，也就是说病死率是流感的 50 倍还要多；其他国家的情况也大体相近。针对我们国家疫情，早期我曾经判断病死率不会低于 3.5%，因为当时我们国家的数据还是低于 1%，当时“知识分子”采访我，问“为什么”，我给出的原因是“很多病重的患者还没有死亡”，随后统计上去之后应该能达到 3.5%。现在看来，当时还是比较保守，但那个时候也是国内甚至国际上，认为最高的一个估计。而现在，各个国家还都有大量的新发病人，相对第一波传播是比较发达的国家，第二波非发达的国家病死率会怎么样？我预计病死率会超过 5.2%。在这种情况下，任何一个国家思考群体免疫的时候都要想：发病阶段有没有足够的医疗资源去应对？这样的死亡率能不能接受？

回到动态的过程中来看各个国家的变化。我们国家继承原来苏联传统的疾病控制理论，提出了三点：控制传染源、切断传播途径、保护易感者。从图 4 中可以看到，这三点并不是相互独立的。传染源如果多，传播途径就要加力，也可以减少可能的易感者。

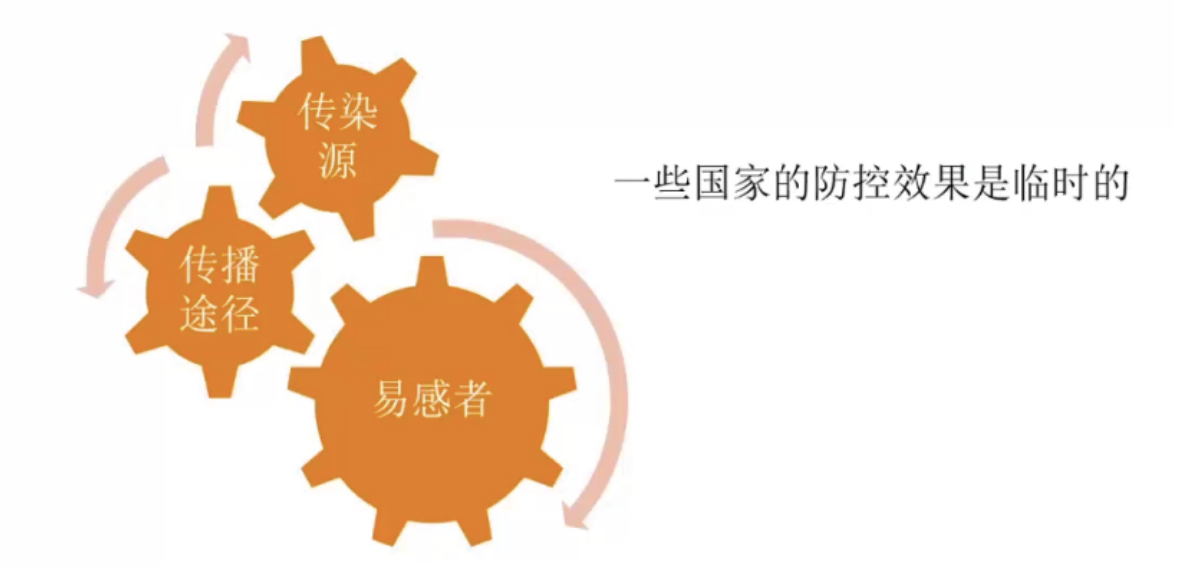


图 5：传染病的动态中，思考各个国家的疫情变化

现在来看，美国和欧洲的一些国家认为现在似乎达到了群体免疫，疾病得到了很好的控制，疾病的新增感染人数也下降了。但要注意到，这只是暂时的，因为前一段时间隔离措施比较严格，传播途径上有所着力，很多易感者并没有发病转化为传染源，从而使传染源降低了。一旦未来各个国家恢复到正常生活状况，疫情必然会再次反弹。

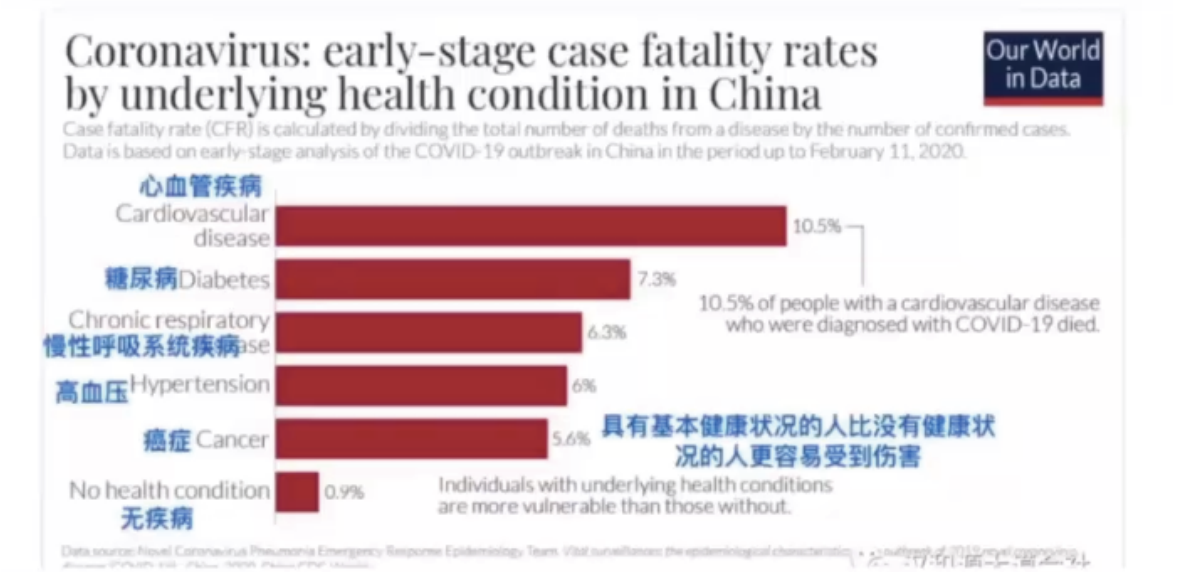


图 6：新冠中的脆弱人群

整个疫情易感人群主要是脆弱人群。心血管病的病人，世界各个国家的报道非常吻合，死亡率达到了 10%； 糖尿病、慢性病呼吸系统疾病、高血压和癌症晚期患者，也都达到了 5%–10%。相对来说，如果自身没有什么固

有疾病，病死率一般很低。

二、精准防控

我们来看，在北京这种情况下如何做到精准防控；或者在未来其他地方有疫情发生时如何做到精准防控。我们国家在防疫上的一个大的策略，基本上就是：外防输入，内防反复。我们从北京、黑龙江、天津的数据可以看出，基本上黑龙江、天津的病例都是输入的。

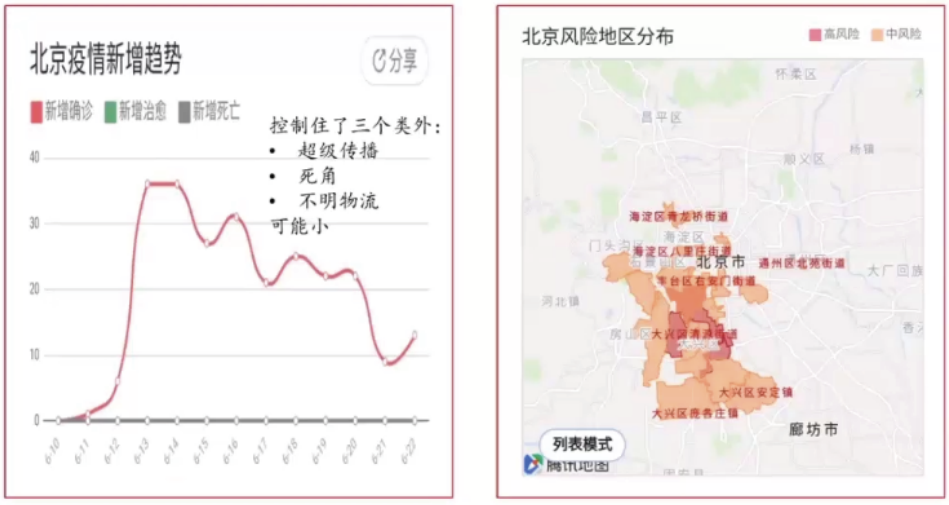


图 7：北京疫情的大致情况

我们从北京病例的基因序列发现，它和欧洲的基因序列更接近。具体是欧洲输入还是从吉林延展过来的，我们还不清楚。天津这边，已经发现了和北京的疫情有关联，因为在发病人群和北京流行的人群之间有一个无症状感染者。我们可以说，北京的疫情目前是处于积极的控制下，于 13–14 日达到高峰，之后就是稳定的下降了，18 日之后，CDC 专家谈到北京的疫情得到了很好的防控，为什么呢？因为 18 日之后，所有病人的发病和感染都确定是和新发地市场密切相关，后来发病的人群应该是间接接触，疫情的传染源是非常明确的。在这种情况下我们把新发地以及相关地区控制住，疫情防控就有保障了。从图中也可以看到，18 日以后我们看到疫情是相对稳定下降的。现在有三个意外情况可能会发生，1）出现超级传播者；2）出现监控死角；3）有不明原因的物流传播病毒。那么这会引起新的疫情，但这种疫情的可能性比较小。所以我个人支持已经看到希望的观点，让民众感觉到我们的疫情有控制的希望，让他们抱有乐观而不是恐惧的心态来面对疫情。

非常感谢文继荣教授和北京的团队，能够把不同的区域分为低风险、中风险和高风险，甚至局限在一些街道上，在不同的区域采取不同的防控措施。大家可以看到基本上在西南城区的丰台、大兴等地区的风险比较高。同时对于北京的疫情能够做到这样精准的防控，是非常不容易的，因为它的人流量非常大，而且短时间内流动快，流动半径大。这个时候追踪技术就发挥了重要、关键的作用。

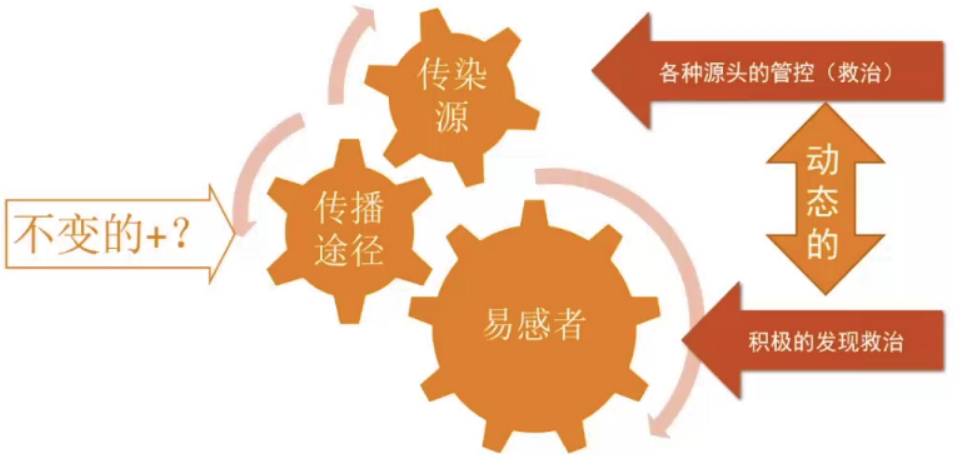


图 8：确定性中寻求突破口：传染病的动态 + 相互减压

什么叫科学？什么叫精准？怎样管控这次疫情？我们还是要回到传染源、传播途径和易感者来讲。对于传染源就是源头的管控； 易感者就是对重点人群，如年老、患病的人群，要进行重点保护。我们可以看到，基本上发病人群以及感染疾病未发病的人群是在一个动态的过程。在传染源的管理方面国家是非常发力的，每个小时都有可能有新的疫情发生，因此时间的动态是非常关键的，如何在传染源、传播途径和易感者之间减压，这是我们的一个思考。

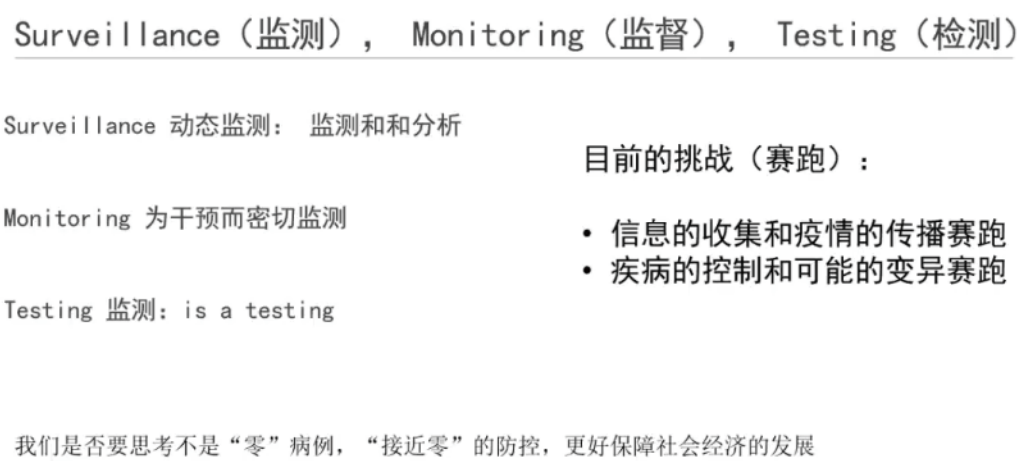


图 9：寻找精准的靶子：基础数据和信息

目前北京正在进行大面积检测，武汉也在进行全城检测，那么就要讲三个基本概念：Surveillance（检测），即要了解疾病现状； 另一个 Monitor（监督）对疾病的现状进行监督，甚至采取一些干预措施改变疫情的动态； 再就是我们常说的 Testing（检测），应该是监测和监督疫情的重要内容。但是我们从公共卫生的角度思考，如何在监测、监督这两个方面来下功夫，找到最关键的高危人群、最可能的传染源，然后有重点地实施公共卫生措施？疫情的变化基本上是每三到六天会感染一拨病人，每天的疫情分析、甚至时时刻刻的分析都非常重要。数据的收集和分析是就是在和疫情的传播赛跑，我们的控制措施也在和疫情的传播赛跑。今天有报道说我们的病毒株

在变化，对细胞的亲和力增加了 10 倍。中国这种情况下，如果能有效地控制疫情的发展，病毒变异的可能性就可以大幅度降低；如果疫情还在全球这样流行，那么变异的可能性就会增加。当然，我觉得一年以内大的变异还不会有，变异会随着传播力明显增加。其致病率是降低还是增加，现在我们还不明确，比较理想的状态就像流感这样，但现在我们还有不可确定的因素。

对于未来疾病的控制，我们一定要考虑对社会经济的影响。毫无疑问，我们控制疫情是保障社会经济的发展，所以现在的目标基本上是奔着“零病例”的方向发展，如果能够加强 Surveillance 和 Monitor，能不能考虑接近零病例的情况下让社会和生活正常运转？同时即使有小的疫情发生，我们也能够快速扑灭。当然，对于新发地这样的疫情属于难防难控，一般可防可控的情况下我们要控，同时也要防控相结合，这样能够减轻社会经济发展的压力。

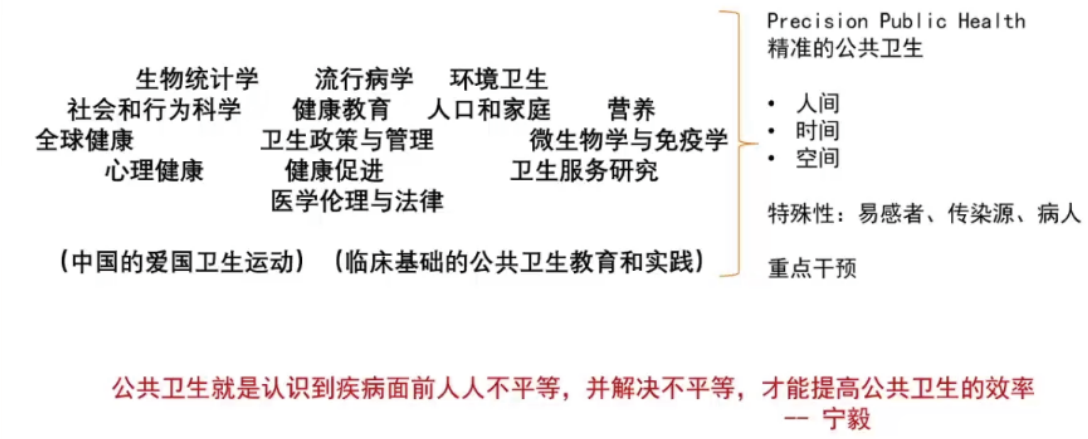


图 10：公共卫生到精准公共卫生

刚才提到我们通过监测以及监督来对高危人群、疫情可能的发展进行预判的时候，在公共卫生领域，对传染源和易感人群、传播途径的管理是不是可以做得更好？公共卫生包括生物统计流行病和环境卫生，可以看到这一次新发地疫情与环境是密切相关的，所以个人和家庭能不能考虑社会和行为学的干预建立起健康的生活习惯？如果有健康的生活习惯，我们就不用担心疫情的传播与危害。因此，如何让普通民众通过健康教育接受、强化这种健康行为？这就涉及到人口和家庭，后面再讲从营养上怎么支持这次疫情的控制，接着又涉及到全球疫情的防控、海关的管理、卫生政策的管理，还有对疾病的基本了解，包括一些社会学的因素等。这次三文鱼不能吃了，三文鱼产业在中国有多大，鱼类的产业有多大，肉食类产业有多大，这些都有很大的影响，我们应该怎么解决？同时，涉及到心理健康和健康促进等，如何防控人们的恐惧心理？每个决策者在每一次把控平衡这种精准和扩大范围防控的时候，也在把控会不会发生可能的一些风险。这就非常需要我们在专业上甚至是 Precision Public Health，涉及到什么样的人需要检测？在动态中如何把握？哪些空间要下功夫？涉及到易感者、传染源和病人，要运用怎样的综合公共卫生手段？在我看来，**公共卫生就是认识到疾病面前并非人人平等，我们解决公共卫生问题就是解决不平等的问题，提高公共卫生的效率。**对于高危人群、高危时间、高危空间、脆弱人群如何解决，要把这些问题解决得好，防控效率就会相应提高。

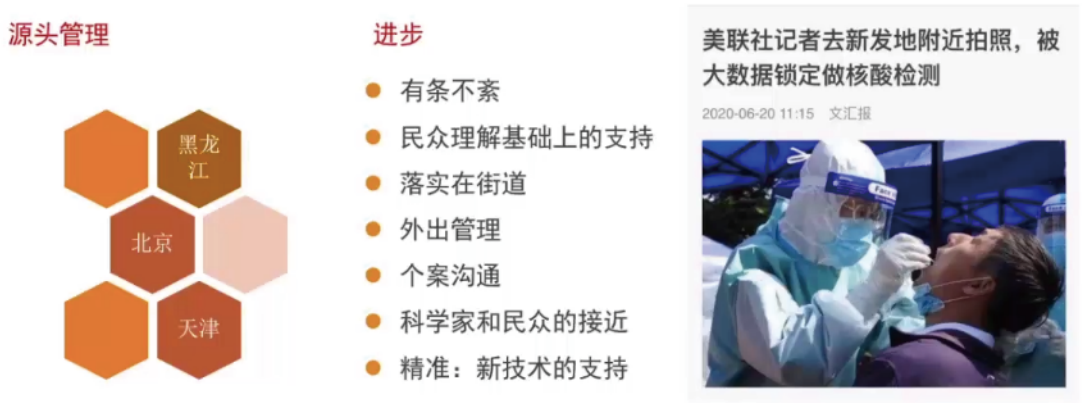


图 11：外防输入，内防反复：病毒流动和有序人口流动

从源头的管理来看，黑龙江、北京和天津都处于较难防控的状况，我们应该在控上下功夫，对于难防的部分以及重点的区域，比如农贸市场，现在对已经确定的重点区域要进行严格地管理，同时加强控制。在北京，我们感觉到有些防控还有进步的空间，表现在整个防控的过程当中太紧张，因为新发地是非常不一般的地方。不过整体来说是有条不紊的，民众能够理解和支持，并且能做到自愿去进行核酸检测，社区对于外出的管理也非常完善。另外在精准预防上，也运用了很多新技术，例如美联社的记者去新发地拍了张照片，结果被大数据锁定，要求他必须做核酸检测。

三、新冠和非传染性疾病

高发病率和高病死率是
- 老年人
- 基础疾病人群
- 社会脆弱人群

四个阶段

- 双重负担： A double burden of diseases
- 向慢性病倾斜： Shifting from infectious diseases to non-communicable diseases?
- 关键问题：卫生资源匮乏 In and Out of the basket of public health
- **殊途同归：解决慢性病舒缓传染病压力Common risk factors**

图 12：同时思考传染病和非传染病

这次疫情可以分为三个阶段： 疫情发生的阶段； 疫情发生之后现在大部分省份都处在缓冲期的黄金阶段； 接着在秋冬季来临时可能是疫情高发的阶段。

现在我们在想如何共同解决传染病和非传染病的问题。可以从这次疫情看到，高感染率人群以及高病死率人群基本上都是老年人、有接触疾病的人群和相对脆弱人群（经济收入各方面比较低）。在这种情况下，我们认为传

染病和慢性病共同构成了公共卫生的重要部分，尤其是中国这样的发展中国家，对于健康资源长期投入的不足，那么有限的资源究竟是优先覆盖慢性病还是优先覆盖传染病，这是公共卫生的焦点问题。回到今天，我们认为整个新冠疾病的高危人群就是慢性病的人群，这两个疾病都不能轻视。

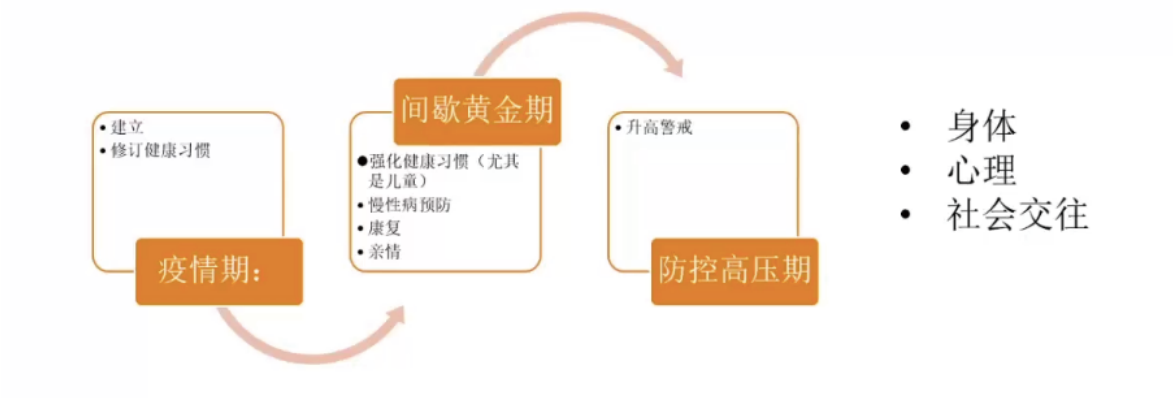


图 13：疫情期、间歇黄金期、防控高压期（秋冬季节，或者提早）

在未来的疫情期、间歇黄金期以及防控高压期阶段，我们可以做哪些事情？首先，疫情防控在秋冬季会有大的压力，也就是防控的高压期。北京目前是处在疫情期，可以看到我们要建立和修正原来的一些健康习惯，比如戴口罩、勤洗手、少去拥挤的地方，保持 1 米以上的社交距离。针对防控的高压期，需要强化健康行为习惯，包括分餐、疾病预防康复，让一些疾病康复，从而达到理想的身体状况。

在夏季，我国一些地区发生了疫情反复，预计秋冬季的疫情防控压力会更大，每一波疫情的影响都值得我们关注，我们要改变一些行为，从而降低患病风险。我们也要体会饮食、心理状况、运动和社会交往对健康的影响。因为我本人是学营养的，因此我希望能够给大家推荐一些良好的饮食习惯，即三多两少。如下图所示。

图 14：生活方式建议

四、总结

对于本次演讲，总结如下:

1. 不能过度依赖国家的组织动员，要科学精准实施防控;
2. 重在外防，同时在内防中动态地思考防控使用公共卫生技术手段;
3. 发挥个人和家庭主战场作用，生活和防疫两个常态化;
4. 预防优先，思考慢性病和传染病的防控，管控风险。

精鼎医药冯胜：真实世界数据（RWD）和 AI 如何炒好临床研究这盘“菜”？

整理：智源社区 张弛

在 2020 北京智源大会“AI 防疫”专题分论坛中，我们邀请到了冯胜教授进行题为《COVID-19: Real-World-Data (RWD) & Advanced Learning in Clinical Studies》的分享。

冯胜，精鼎医药真实世界数据亚太区总裁。中国科技大学分子生物学学士，美国北卡州立大学生物统计和生物信息学博士。

在分享中，冯胜提出了真实世界数据和临床试验数据的两条相对独立的研究处理路径，而此次 COVID-19 疫情将两条数据处理路径错综复杂地联系在了一起。针对 COVID-19 疫情背景，并结合自身多年企业工作经验，冯胜对 AI 处理 RWD、RWD 进入政府和 RWD 进入药物公司三个方向展开阐述，分析了 RWD/AI 在 COVID-19 疫情中进入临床研究所面临的普遍、特殊挑战是什么，最后为 AI engineers 和 IT 行业从业人员等将真实世界数据和 AI 技术带入临床研究提出行之有效的建议。

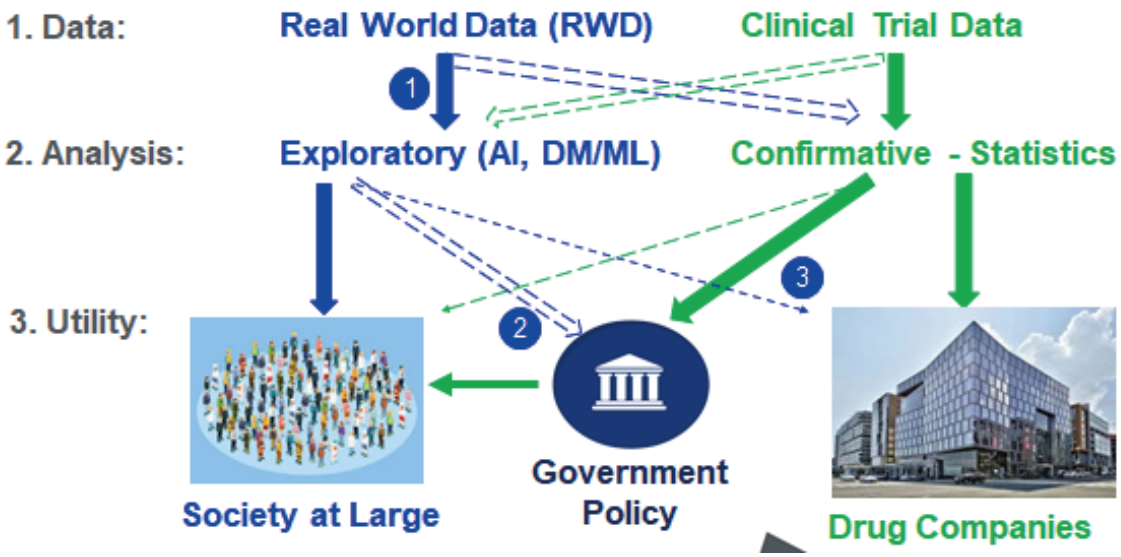


图 1：真实实验数据和临床试验数据的处理

一、RWD/AI 在 COVID-19 疫情中进入临床研究所面临的普遍、特殊挑战是什么？

值得一提的是，企业生产范围内所定义的 AI 除了包括学术范围内的神经网络、深度学习等，还包括数据挖掘、统计分析方法等，是广义的 AI，而在接下来的分享中也将在此基础上进行讨论。

1. RWD/AI 所面临的普遍性挑战——数据质量的划分

当我们从数据的应用场景横向展开，以数据质量标准（Data Quality Standard）、数据分析性质（Analytic Styles）、临床应用价值（Clinical Market Value）三个维度纵向进行分析时，我们会发现真实世界数据面向大众，主要应用于社交媒体、新闻等场景，且多为探索性数据研究，对于数据质量标准的划分并不明确，因而在临床研究中的价值甚微。然而在临床试验、公共政策制定等场景中所需要的数据多为确定性分析数据，要求对每一个数据都需要很强的质量控制，所以这些数据在临床研究中价值也是最大的。

Data Utility Scenarios	Newspapers, TV-Spotlights, Online blogs, Twitters/Facebook/WeChat	COVID-19 Scientific Papers FDA/CTAP Review, Clinical Trial Design Simulation	Public Policy Making, Clinical Trial Designs, Insurance Decision
Data Quality Standard	Mostly Unknown/untested; **no standard**	Varying: from low to high, **In general, in the mid-ground**	**Strong Quality Control** not good news for RWD/Big Data
Analytic Styles	Most Unknown. **For knowns: Exploratory 80%**: AI/DM/ML **Confirmative 20%:** Statistics	**Exploratory: 50%** **Confirmative: 50%**	**Exploratory: 5%** **Confirmative: 95%**
Clinical Market Value	$	$$$	$$$$$$$$$$$$$$$$$$$ $$$$$$$$$$$$$$$$$$$

图 2：数据在不同应用场景下的分析

2. RWD/AI 在 COVID-19 疫情背景下所面临的特殊性挑战

- COVID-19 疫情来势汹汹，在全球忙于应对疫情的情况下，收集到的数据质量参差不齐、无法控制，不同于癌症、糖尿病等数据集，疫情数据快速产生、变化迅速，这就给数据的采集和分析造成了很大的困难。
- 疫情初期，最好的数据分析师和最好的数据源之间存在鸿沟，存在于企业中的优秀的数据分析师无法获取到置于政府内的公共数据集，导致数据分析的滞后。
- 疫情突如其来，无法制定临床数据的标准，因而也就无法评判数据分析质量。
- 全世界优秀的数据分析师和强大的资源处于分散状态，无法集结起来进行数据分析。
- 疫情期间的一些虚假数据、探索性模糊数据和确定性数据交融错杂，而这些数据的不公开、不透明性导致无法辨别数据的真伪，这也就为数据处理增加了一定的困难。

分析完 RWD/AI 在 COVID-19 疫情中进入临床研究所面临的普遍和特殊挑战，冯胜介绍了一个数据集在临床研究中的应用例子，利用贝叶斯方法分析包括 local policy/law 在内的多个公共数据集，使得数据分析能够突破实时监测的限制，可以达到预测一个月后的效果。

二、政府在 RWD 进入临床研究中扮演的角色

2020 年 5 月 19——29 日，十天之内，FDA 对羟氯喹两次授权又取消，这次事件的背后，针对“羟氯喹对于疫情的治疗是否有用”这一观点出现了很多支持和反对的临床证据，而这些证据大多都表现出小样本、不严谨、发布于社交平台等特点，不符合严格的临床试验的特点，那么针对这个事件，冯胜提出问题：疫情之下，一定要大样本、随机、双盲、安慰剂对照的严格的临床试验吗？监管机构能够接受的最低力度证据（Minimal Strength Evidence）是什么？能否有指南 / 指导原则？真实世界数据到底行不行？

为了回答上面的问题，冯胜通过一个 RWD 应用于临床的例子——埃博拉和棕榈树平台试验进行说明。棕榈树实验是一个十分不严格的临床试验，它没有双盲，没有安慰剂对照，孕妇、小孩都可以入组，可以中途叫停，并将患者换入治疗效果好的组内，尽管如此，棕榈树实验还是大获成功。从以往多次惨烈的教训中，科学家、医生和世界的领导者们逐渐意识到，疫情爆发的时候，不是进行严格的药物临床试验的理想时机，如何快速的找到一种有效的药物治病救人才是首要任务。而在这次 RWD 成功应用于临床试验的例子中，人们意识到：RWD 到底可不可以应用于临床试验取决于多个因素，但是其中关键的一点是——“领袖”是否介入，通过国家或政府将有效的资源进行整合，这是 RWD 能够成功的关键。

三、RWD 应用于临床研究，AI engineers 应该怎么做

冯胜认为 AI engineers 和 IT 从业者们无法将 RWD 应用于临床的一个原因在于沟通不畅，AI 工程师在于临床医生交流时应该注重于物资分配、人员部署的讨论，而不是执着于介绍算法功能的强大，同时应该更多的需要了解适应性实验设计的要求，构筑起大数据和临床之间的桥梁，巧妙把握好实验性（Explanatory）和适应性（Pragmatic）是 RWD 助力临床试验的重要一环。

四、总结

RWD 要成功应用于临床试验，需要国家或政府的介入进行资源整合；AI engineers 要使 RWD 被临床医生所接受，便要将知识体系和思维架构逐渐向实用型临床知识过渡，掌握适应性实验设计的要求，使得科学型 AI 和实用型临床知识达到平衡，相信 RWD/AI 的助力一定会使未来临床试验的研究再创辉煌！

北大副教授边凯归：新冠疫情下接触者精准追踪技术的思考：如何最大化用户隐私的保护？

整理：智源社区 季葛鹏

在第二届北京智源大会“AI 防疫”专题论坛中，来自北京大学信息科学技术学院网络与信息系统研究所副所长、副教授边凯归，对团队近期关于“流行病接触者精准追踪技术”的研究工作做了一个详细的报告分享。整个报告分成四个主要部分，包括：流行病接触者追踪技术的基础概念和背景简要介绍，管控措施粒度与接触者追踪精度层次，产品落地过程中的三个约束条件，随后介绍了智源研究院开发的保护隐私的流行病接触者精准追踪系统——智源蓝保（Blue Bubble）。

一、流行病接触者追踪的技术背景

图 1：接触者追踪的定义

什么是**接触者追踪（Contact Tracking，CTC）**？其包含四个内容：第一个功能是追溯与感染者密切接触的人员；第二是所提供的结果是否支持对某些人员进行隔离；第三点是扩大对更多流调人员的支持力度；最后一点是需要使用更为高效的数字化的工具。**CTC 最终的目的是将密切接触者的近距离感染者找出来，保持一定的检测精度，尽可能地减少非必要隔离人员带来的损失。**

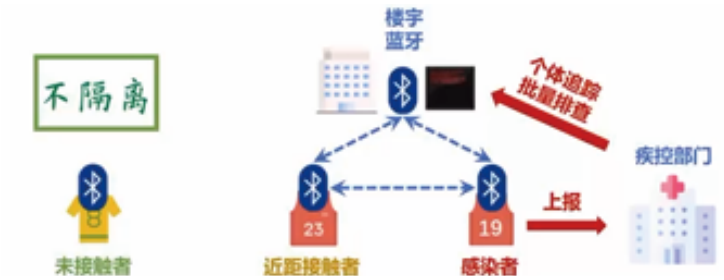


图 2：智源研究院启动流行病接触者追踪技术研发

智源研究院积极响应国家号召与社会即使需求，在二月底 3 月初便启动了接触者追踪技术的研发，用于解决复工、复产、复学过程中**精准追踪**感染人员的位置，**快速排查**与感染者有近距离接触人员，最终目的是**大幅度降低隔离人员比例。**

边凯归补充道，我们的技术方案经过几轮的迭代确定，最后决定采用蓝牙近距离感知技术方案，其有两种模式：一种是楼宇发现的模式，一种是个人发现的模式（后文会详细地介绍）。4 月 15 日这个系统已经部署在知春路智源研究院写字楼上，楼宇上面的边缘存储会有感染者以及近距离接触的蓝牙记录，通过比对数据就可以写字楼内部找出潜在的近距离接触者。一旦有人感染会把自己手机的蓝牙信息上报给疾控部门，部门会把包含曾经去过的地点的信息发回与上报者或相关部门，最终目的是降低隔离人员的比例。

图 3：新加坡 TraceTogether 系统

3 月 20 日新加坡上线了 TraceTogether 系统，每个人手机上安装了 App，两个人距离较近的时候手机会自动广播蓝牙信息，比如自动随机生成的个人 ID，这样的好处就是可以保护个人隐私，即便他人通过搜索蓝牙 ID 也无法获取个人信息。因为该系统上线较早，所以在一定程度上是有用的。边凯归接着指出其缺陷所在：“由于注册时需要个人手机号和身份信息，所以该技术在比较关注隐私的研究者的角度来看可能会有不一样的看法，后面会有专门的章节进行相关讨论。”

4 月 10 日 Apple 公司和 Google 公司宣布合作，将从底层打通个人设备和蓝牙的接口。由于两家不同的公司为保护用户隐私采用不同通讯接口的定义，终于 Google 选择迁就 Apple，计划统一接口为用于研发基于“个人设备蓝牙接口”的接触者追踪技术。但是在没有实现成功之时，就指出个人隐私信息存在被 iOS 和 Android 系统以及第三方应用截取滥用的风险，这也是当时无数学者批评该做法的原因。

4 月 19 日，来自全世界多所大学的 300 名国际学者针对基于蓝牙的近距离接触追踪技术发表联合申明，明确并提出四点要求： 第一，必须使用独立的系统开发具有独立功能的公共卫生检测手段，避免数据被第三方 App 滥用； 第二，解决方案必须公开透明； 第三，必须默认地最大化保护用户隐私； 第四，技术的使用必须是用户自愿的。这些研究者一致认为，必须使用**一个统一而又去中心化的计算框架**，这是给出的一些框架设计的原则。

图 4：Exposure Notification 框架的发布

5 月 20 日，Apple 与 Google 终于联合发布了名为 Exposure Notification 的框架，支持最新的 iOS 13.5 或 Android 系统，但不支持大陆用户。

6 月 15 日，德国也出了一套名为 Corona Warn 新冠警告系统，就是使用了 Apple 和 Google 公司在 5 月 25 日发布的框架，流程大致为：生成一个随机 ID，并且同时扫描和记录周围设备的 ID 也是把这个 Key 上传到中心服务器，服务器会将感染者的 ID 发给所有安装 App 的设备以示警告。

边凯归针对当前国内国际疫情局势发展进行总结：每一个国家都需要一套符合国家法律法规的、高度保护用于隐私的感染 AppAPP，而且由于数据安全因素，每个国家都需要使用自己的技术力量去完成这样一套独立于国外的东西。所以，基于当前市面占有率最高的两款操作系统（iOS 和 Android），我们亟待研发一套真正属于自己的通讯协议，做出属于自己的精准追踪产品。

二、管控措施粒度与接触者追踪精度

管控粒度	管控方法	借助技术手段	问题
城市级别	封城	无任何防备的情况下，难以提前进行技术部署	不得已而为之，经济损失巨大
街道 社区级别	街道、社区只进不出	大数据分析，手机连接基站数据分析，app数据	粒度较粗，各方数据难以协调融合，经济损失较大
楼宇 楼层级别	封楼 分楼层人流控制	???	需要更合理的技术，以及更精细化的管控措施
家庭 个人级别	追溯个人轨迹	新加坡TraceTogether、Apple暴露日志、德国Corona Warn	商业公司做公共卫生监测比较尴尬

图 5：管控措施粒度分级

针对疫情爆发后的管控措施粒度，边凯归将整体分成了四个级别：城市级别、街道社区级别、楼宇级别、家庭个人级别。前两个级别所带来的经济损失都非常大，但由于在疫情爆发的初期，这种方式是最为简单有效的。**楼宇级别的管控，则需要更合理的技术、更精细化的管控措施，边凯归和北京智源联合团队便是基于这个层级来完成追踪的功能（详见后续介绍）。**最后就是家庭和个人级别的追溯个人的轨迹，新加坡、Apple 与 Google 都完成了个人 APP 产品的开发，但是其必须由政府发布。新加坡可以说做的十分成功，因为新加坡是由政府发布，但如果是由商业公司完成必然会引发一些社会问题，因为以商业公司的角色来完成公共卫生监测显得比较尴尬。那么有些同学会问支付宝和微信是否能够完成类似的任务？答案也是可以的。但是这类软件属于社交软件，不能直接抓取数据拿来防疫使用，用户条款中仅仅表述了是一个社交用途，并没有说明能拿来用作防疫使用，所以这类软件在接触者追踪的过程中显得同样尴尬。

如上所述，接触者追踪系统需要满足街道社区级别、楼宇级别、个人级别，不管哪个级别的管控粒度，需要解决三个问题：第一是追溯，第二是发现，第三是决策。最难做的也就是“追溯”，这一步的核心就是采集位置信息（绝对位置与相对位置），通过位置信息即可判断感染者和接触者之间的距离，从而得出最后的决策是否需要为密切接触者采取相应的措施。

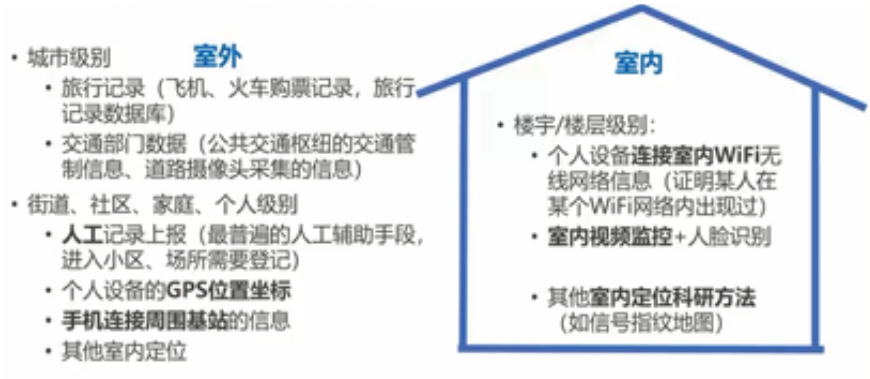


图 6：追踪绝对位置信息的常用手段（室内与室外）

边凯归列举了追踪绝对位置信息的常用手段，如图 6 所示。对于室外级别的，在之前的论述中已经略有谈及，包括：旅行记录、交通数据、个人设备的 GPS 坐标等。而对于室内级别的定位，现用最多的是基于科研的方法，大规模的工业界的室内定位产品暂时还没有。以往会做一些基于 WIFI、监控视频的绝对位置的采集，这些有利于疫情爆发初期的信息采集，快速锁定感染者和近距离接触者。**而对于常态化防控方式，通过汲取国内外的方法，边凯归团队建议采集相对位置信息，可以理解为一种间接证明感染者和接触者距离的信息。**最终，团队决定选择蓝牙 4.0 技术来实现，其优点就是可近距离判断与功耗低。**值得注意的是，由于所采集到的是相对位置信息，避免了绝对位置信息暴露用户隐私的痛处，而相对位置只是判断你是否近距离接触，可以最大化的保护用户的隐私。**

图 7：集合所有技术的接触者追踪系统

那么假设我们可以设计一个“All in One”的系统，即将所有的国内外的方法都集成为一体，我们就可以实现出一个非常完美的接触者追踪系统。这种系统的优点肯定是大而全且实用，但是里面有很多的技术因为各种担心却不太推荐去使用，例如：中国还有很大一部分的人没有智能手机，即根本没有从传感器获取来的数据是无法完成上述大系统中的某一些分支功能的。

三、三个约束：隐私、计算、成本

超越之前讨论的研究层面，在实际产品的落地应用中，还会面临着三个约束条件：

- 隐私：保护责任不够清晰、数据使用和存放不安全、绝对位置信息的滥用。
- 计算：多源数据的融合困难程度高、人工智能背景下视频监控跟踪计算成本巨大且延迟问题严重。所以推荐主要使用独立的系统采集的无线信号作为防疫数据，使用视频监控作为辅助数据加以支撑。
- 成本：普及安装 App 成本高、基于 WIFI 的室内定位算法开发成本高、视频监控系统造价高。

四、保护隐私的流行病接触者精准追踪系统

图 8：智源蓝保（Blue Bubble）

随着社会的需求和技术的发展，Blue Bubble 这一款最大化保护隐私的流行病接触者追踪系统应运而生，这是一款独立的、非商业的系统。这种基于蓝牙开发的系统支持人与人、楼宇之间互相发现，另外包含视频测温系统。该系统有如下特点： 用户数据分布式去中心化存储，从而可以有效地保护隐私； 独立采集相对位置信息，不牵涉多源数据的融合； 双工作模式，可不安装 App； 低功耗、低成本、快速部署； 可恢复待追踪者的行为轨迹。边凯归对两种工作模式进行了展开说明:

图 9：楼宇蓝牙发现模式

- 楼宇蓝牙发现模式： 楼宇蓝牙发现个人蓝牙，并存储于楼宇的边缘设备中，数据不上传中心，可基于蓝牙发现数据进行感染者轨迹生成。

图 10：个人蓝牙发现模式

- 个人蓝牙发现模式： 独立设计的一套不依赖于底层操作系统 API 的协议，包括 iOS 和 Android 系统的互相发现。可通过蓝牙互传、数据存储于个人设备。

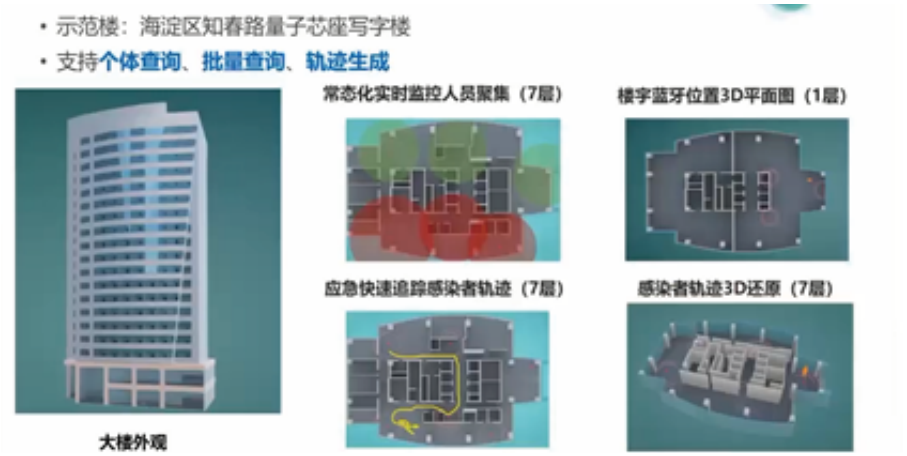


图 11：3D 可视化监控

另外，该系统提供的可视化监控可以整合到已有的视频监控系统中，支持个体查询、批量查询和特殊感染者的轨迹生成，常态化防控的时候生成一张热力图，可以看到室内的聚集度。

管控粒度	管控方法	借助技术手段	问题
城市级别	封城	无任何防备的情况下，难以提前进行技术部署	不得已而为之，经济损失巨大
街道 社区级别	街道、社区只进不出	大数据分析，手机连接基站数据分析，app数据	粒度较粗，各方数据难以协调融合，经济损失较大
楼宇 楼层级别	封楼 分楼层人流控制	**智源蓝保楼宇模式**	需要更合理的技术，以及更精细化的管控措施
家庭 个人级别	追溯个人轨迹	新加坡TraceTogether、Apple暴露日志、德国Corona Warn **智源蓝保个人app模式**	商业公司做公共卫生监测比较尴尬

图 12：智源蓝保楼宇模式

该系统还可以用于补充楼宇级别的管控力度，最大化地保护数据不离开本人的手机，因为蓝牙的近距离关系判断可以达到米级。

五、总结

习近平说过，要把人民的生命健康放在第一位，其次还要同步加快经济恢复的速度，最终观察的结论就是各国都推出了具有本国特色的防疫追踪系统。**正因为这是涉及国家的数据安全，不能用别的国家的，并且还要独立于其他国家的技术框架，拥有独立采集技术系统，这样可以避免很多问题。**最后我们需要做分布式的数据存储，最大限度保护隐私，隐私数据不能离开楼宇或者用户手机，甚至在不安装 App 的情况下可以跨设备、跨操作系统地追踪楼层级别的接触者。

最后，边凯归还留下一个思考题： 如果没有智能手机怎么办？欢迎大家进入我们的讨论区展开讨论：hub.baai.ac.cn。