最近冠状病毒的传播成为家家户户关心的重点，同时谣言满天飞。病毒到底最后会感染多少人，我有多少风险被传染是众说纷纭。而这正涉及了一个人类历久常新的话题 - 预测未知。也是机器学习和复杂系统科学研究的重中之重。

不懂模型的朋友往往会被网上各种不同的预测搞得时而信息满满时而悲观绝望。事实上了解模型的第一步就是记住一句话：所有模型都是错的， 但是有些模型是有用的。任何模型都是一定的假设加推理的结果， 都有一定的置信区间， 不加批评的相信模型甚至还不如没有模型。了解模型的基本逻辑才是最重要的。那么我们这里了解一下基本的模型如何帮我们预测肺炎传播。

**一， 统计模型**

首先能够帮我们预测感染人数的是统计学模型，统计模型是数据驱动的预测模型，它能够帮我们解决的一个首要任务是在信息不全（小样本）的情况下， 预测样本总体的情况。比如我们知道某个飞机上武汉病人有n个人， 你可否知道整个武汉的感染人数？这看似是一个不可完成的任务，犹如通过梧桐一叶去判断天下秋至。但是从统计学角度这是可能的。

但是我们有一个数据确实已知的， 就是某个时间段内海外检测到的武汉肺病病人个数。然后我们可以从这个人数估算整个武汉肺炎的得病个数。看上去天方夜谭， 这个海外出行的人仅仅占总武汉人口的沧海一粟， 那么这样一个数据有什么特殊之处？为什么不用武汉某医院检测到的病人个数？

这里面的理由是从这个小样本到总人口存在一条数量记录清晰完整的链条， 直到得到武汉市的整个人口。为什么？首先每日出入机场的人数已知， 前往出境的人数已知， 那么计算一段时间内从机场出境的人数就是已知的。这使得计算这部分人中的染病概率变得比较简单（某段时间患病人数/总人数），一旦推测到这个概率， 那么我们就得到了这个人口总数。

这个方法取决于非常重要的假设：武汉所有地区和人群的患病概率一样。事实上我们很容易想到这个假设是错误的。能够有能力出境的人口所具备的特点显然不同于武汉的总人口。而且这个模型只能监测某个特定时间窗口的人数， 而非随时间变化的人数。

如果我们引入更复杂的统计模型， 比如贝叶斯网络，会取得比刚刚说的简单模型好的多的预测效果， 但是依然无法对感染人数进行动态预测。

最初期的对感染人数的推断来自这种由小样本反推大样本的统计模型， 用这个方法得到武汉21号左右的染病人数在4029， 但是大家千万不要对十位数认真因为这个数字的方差估计得有大几千。虽然方差巨大， 总体看在节前三天4，5000的感染人数是一个合理范围。

Estimating the potential totalnumber of novel Coronavirus (2019-nCoV) cases in Wuhan City, China

**二， 动力学模型**

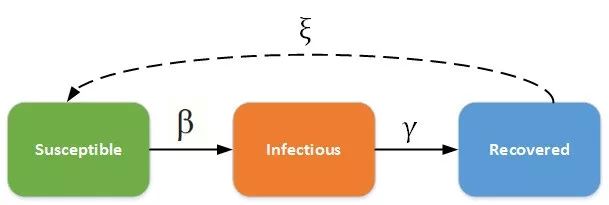
如果想预测未来的感染人数， 那么动力学模型是无法绕开的必由之路。**动力学研究的就是事物运动变化的因果关系。**在这个学科视角下，我们把世界理解成一张相互作用的大网，而事物运动变化的原因，都可以从这张大网上通过求解微分方程找出。

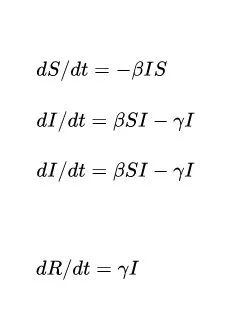
动力学最初的一鸣惊人，是空前绝后的牛顿三定律的提出，而牛顿第二定律又是三定律的核心，只要有高中数学基础的人就知道，它是力与加速度的关系，看起来不起眼，但是它包含了整个动力学的核心思维，是什么？是受力分析吗？

No。**第二定律的本质，是预测**，而预测无穷尽的未来，你不需要太多信息，**只需要知道此刻和与之最近的下一刻的关系**（微分的思维，下一刻是一个极限的概念，恰好脱离此刻的时刻），**所谓s(t+1)=f(s(t))。s代表state，即状态， f就是由此刻的状态得到下一刻的状态的迭代关系。有了f和初始时刻的s，未来即确定。**

**动力学里最重要的是了解系统的维度也就是变量的个数， 它决定了系统的复杂度。**

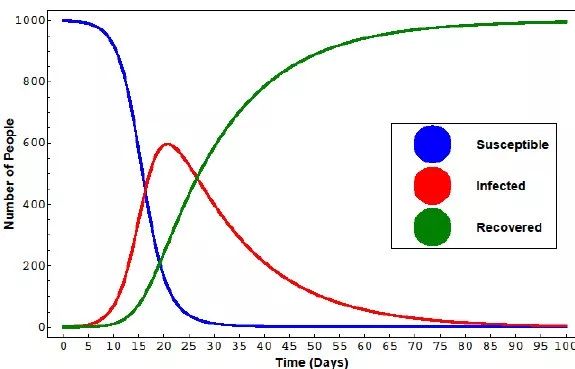
我们把类似的思维用到疾病防控的思维里来， 如果你要做一个模型预测疾病的走势， 就首先要把这个问题简化成一个维度较低的系统， 这里我们可以把人口切分为三分，易感人群， 感染人群和康复人群， 病毒传播过程里感染人群会传播易感人群会， 而某段时间又会恢复。我们把这三个变量相互影响的关系用微分方程刻画，就得到一个简单的动力系统， 这三个变量可以称为S， I 和R。I感染S， 同时I可以恢复成为R。虽然传播是复杂的， 但是抓住最核心的因素， 无非三个不同的人口群体， 易感， 感染和恢复， 好比把人群分成了三节车厢， 这三节车厢就是SIR， 通过分析人口在这三个车厢的迁徙就可以预测未来的患病人数变化（如下图）。





SIR模型示意图 和微分方程

首先我们可以看到系统的稳定点（也就是微分方程的不动点）是I=0， S=0， 也就是病毒最终会过去（好消息！）。然后呢， 我们可以预测出一个指数增长然后又降落的趋势， 传播越快， 结束越快，  整个人口被病毒过滤一边， 经历从S到I到R的流程。我们目前可能正在处于这样的指数上升期， 大家可以观测最近几天疑似的人数变化。

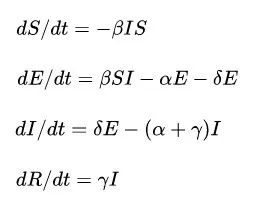
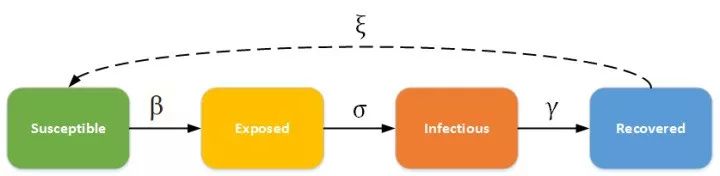


SIR模型中，处在不同状态的人数随时间变化的曲线

这样一个简单的模型告诉给我们以一种原始的信心，就是无论如何瘟疫总会过去， 而且病毒的感染率高，它起来的快过去的也快。这里的最核心参数就是传播速度beta， 它决定了多少时间整个武汉人口将被过滤一遍疫情结束。beta取决于病毒的易感性和人群的接触密度，它也是我们提出降低接触密度控制传播的最原始的基础。

当然这里没有考虑死亡率， 加上死亡率，那么最终得到一个被死亡过滤过的数量变小了的健康人口，就如同瘟疫后的幸存者组成新的聚落。

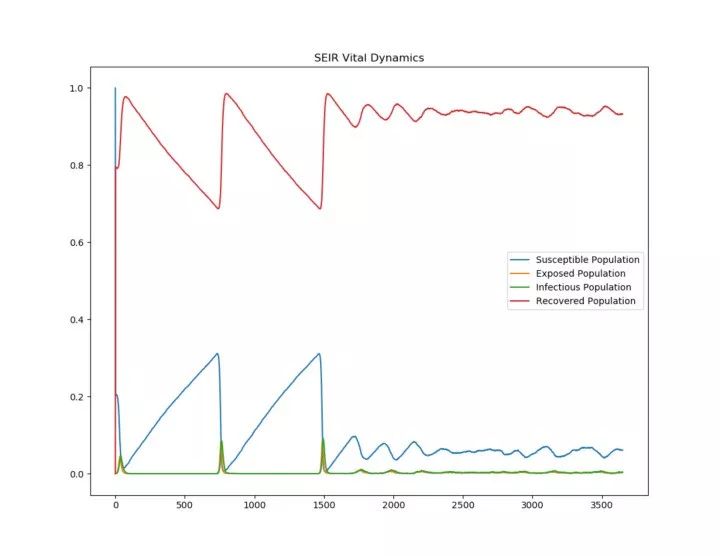
但是显然忽略了众多核心要素， 第一感染后的恢复者是否会在感染， 以及是否当接触发生，同时I就变成了S， 也就是说接触和感染发病几乎是同时的。我们知道潜伏期是存在的， 那么加入潜伏期会对整个病患产生什么样的影响？再者这里我们忽略了网络效应，假定人口是封闭的。对于病毒， 这也不是一个特别合适的买卖， 因为经过一轮传播自己也灭绝了。由此我们需要对整个疾病方程进行修正， 得到通常说的SEIR模型， 也就是在三个变量后加入一个E， exposed 也就是潜伏期的人。



SEIR模型示意图 和微分方程

新增的方程本质上加入了延迟， 它的作用使得动力学变得异常复杂， 甚至引入某种程度的波动性。它导致在疫情的某个阶段我们完全看不到人数的增加， 而后面某个时间点突然开始增加。

病毒进化出潜伏期是非常聪明的选项，这仅仅是数学的世界， 在现实世界里，这个潜伏者的网络会导致一部分人作为病毒传播的缓冲（而且通常潜伏期是一个区间而不是一个固定时间，它可以使得疾病此起彼伏的发生）， 在不被发觉的情况下移动到非常遥远的地方，从而大大加速传播的时间和空间距离。



seir模型可能的不同人口变化曲线

seir模型可能的不同人口变化曲线从病毒的角度看， 增加潜伏期也是非常明智的，因为增加的潜伏期意味着动力学可以进入一种长期震荡状态，通过潜伏可以在时间轴上传递更久的时间，有利用病毒自身的进化。利用类似SEIR的模型， 一些专家得到了武汉病毒的蔓延预测，以及平均每个人都周围人传播的R值。

Estimation of the transmission risk of 2019-nCov and its implication for public health interventions Biao Tang , Xia Wang , Qian Li , Nicola Luigi Bragazzi , Sanyi Tang , Yanni Xiao  Jianhong Wu

在这篇文章里我们加入了更精细的对易感人群和隔离人群的刻画， 从而也更贴近真实情况。最关键的利用蒙特卡洛模拟它给出了一个人平均传播病毒的个数R=6.47一个十分可怕的数字。如果相信这个数字是真实的， 那么一个人传播六个人， 不加防控的结果将十分可怕。

如果你感兴趣还可以参考一篇更为悲观的文章同样用到了SEIR， 虽然文章中对R值的估计更保守：

Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions Jonathan M. Read , Jessica R.E. Bridgen , Derek A.T. Cummings , Antonia Ho, Chris P. Jewell

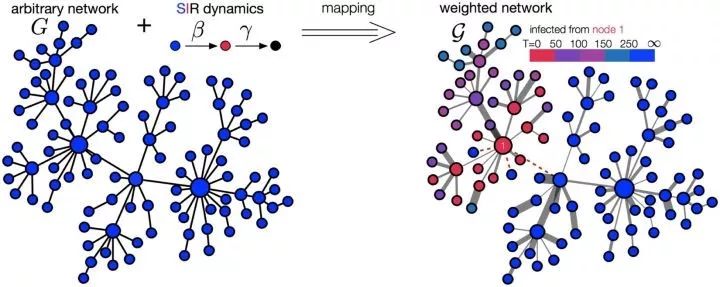
尽管如此，动力学预测对我们理解问题的意义是重大的， 我们可以通过公式了解到影响传播速度的是哪些要素， 比如疾病自身的传染性和社会网络的连接性， 这也可以指导我们做出正确的决策， 别如什么时候封城或者开城。但是， 如果要做到更精确的预测，比如疾病在某个地区的变化曲线，就不那么有效。

**三 交通网络的影响**

这时候我们必须借助网络的力量。网络即高维的动力学系统。网络有很多层的含义， 在一个小区或社区里人们相互连接形成网络，在城市和城市之间有立体的交通网络。这些网络是构成病毒传播的第一层体系。整个网络， 组成一个层级化的复杂系统。这时候整个传播过程变得异常复杂， 每个个体只能传播到相邻节点，而人员需要靠交通网络流动。这时候网络结构本身的性质对传播开始起到至关重要的作用。如果直接考虑个体， 整个模型将参数大到无法想象， 所以在考虑网络的时候，我们不妨把每个城市节点的方程看成上面的SEIR模型， 然后考虑不同的城市间的耦合， 这样我们仿佛得到了一个全部城市组成的动力学方程组。

这时候我们会发现城市交通网络在传播中的重要作用。首先我们会想到武汉是一个九省通衢的中心城市， 那么从这样的枢纽型节点爆发，就比从其它节点有深远的多的传播力。当这整个社会网络进化成无标度的小世界网络， 病毒可以无阈值传播。这突出的说明了网络结构对病毒传播的至关重要性。

Pastor-Satorras, R. and Vespignani, A. (2001) Epidemic Spreading in Scale-Free Networks. Physical Review Letters, 86, 3200.



无标度网络上的疾病传播

何为无标度网络， 在古代社会， 社会网络以局部的连接为主， 当一个社会分工变得越来越复杂， 就需要更多的枢纽承担起物流中转站的重担， 比如武汉这样的超大城市。这些枢纽会使得网络的连接不再是局部的，而是呈现出一种重要hub（枢纽）和局部连接并存的形式， 枢纽和大部分枢纽都有连接，这样的网络称为小世界网络，类似的网络结构在大脑和互联网上也是类似的。小世界网络的传播特性决定了一方面它是一种十分高效的组织形式， 一方面比起局部连接的网络它也更加脆弱，比如在病毒的传播上。我们可以想象除了病毒之外还有什么可以打击这样的网络系统，或许就可以从此次病毒防控中得到启示。

Simulating SIR processes on networks using weighted shortest paths Dijana Tolić, Kaj-Kolja Kleineberg & Nino Antulov-Fantulin

同时这个理论给截断中心城市提供了绝好的理论依据， 因为截断中心城市，就可以将网络从无标度网络退化成局部连接的网络，对病毒传播起到非常大的抑制作用。这比防控乡村或局部城市要来的高效的多。

**四 考虑信息传播的网络影响**

事实上人们在谈论病毒的过程中也改变了病毒传播的动力学本身。比如一个消息传递到一个地方， 这个地方的传播网络自然就减弱了。如果考虑到信息和病毒传播的耦合， 整个动力学系统就增添了新的性质。

在网络时代， 信息的传播是完全自下而上的自组织行为，信息流里的恐慌情绪会具有和病毒类似的传播能力， 这种网络情绪会如何和病毒传播相互作用应该是一个新的学术问题。从目前看这种非理性的情绪某种程序也起到了阻断病毒的作用， 因为过度行动肯定优于不行动， 虽然某些时候恐慌情绪引起的医院挤兑也确实加重了传播。

在很多时候， 信息传播可能会比阻断物理传播有更低的成本，也是这个时代赋予我们的超级武器。

**五， 考虑病毒自身的进化效应**

之前的模型依然考虑的是病毒自身的基因是静态的不变的， 事实上， 病毒通过每个人和整个社会网络进行传播，被杀灭的同时也迅速进化着（这点对于ncov 这样的RNA病毒体现尤为明显）， 因此这是一场人和病毒的赛跑或者博弈。

在这个层次上，如果模型里只考虑静态不变的病毒， 就没有任何说服力。事实上我们的防控手段也在影响病毒的进化。病毒的基因本身的变化也应该被看做动力学系统的一部分。

这个问题就涉及超级传播者， 超级传播者的含义是一些年老体弱的个体好比病毒的超级培养皿， 加上长时间的医学治疗， 可能会使得病毒在这些个体上传播超级多代，从而在它们身上得到更加强大的病毒。由于现代医学的发达， 在整个人和病毒斗争的治疗周期里我们也给了病毒更多的进化机会， 某种程度我们在治疗疾病的同时也在创造新的病毒。

何谓强大的病毒， 它体现在能够和人打太极而不是简单的致死率， 也就是潜伏期更长， 传播更隐蔽， 不太高的致死率， 从而取得和人类共存，继续协同进化（博弈）的机会。   在这个层级上， 我们不得不将博弈论纳入我们的模型，而使得最后的模型成为一个复杂网络演化加博弈模型。

**六， 理解传播动力学的实际意义**

如果理解动力学仅仅解释了我们已知的现象， 那么并不足以让人们觉得这种建模是有用的。事实上， 建模的意义在于更加精确的预测和防控。这个时候深度学习的武器就不能被我们忘记。   我们有了动力学模型，加上一定量的数据， 就可以建立一个可以学习的模型， 从而根据数据不停的优化预测的精度， 这个预测不只是感染的总人数，还有每个地区的人数， 以及在什么时候需要采取怎样的措施（强化学习问题）。

在病毒传播的建模过程里， 我们看到一个贯穿不同层次的核心概念-网络。病毒本身是一个RNA的分子网络， 然后病毒入侵人体网络使人得病， 而人作为社会网络里的个体又传播疾病。不同层级的网络耦合在一起，每一层的网络都是动态的，使得这个问题充满了不确定性。这也是为什么当下的所有预测数据置信区间都很大， 同时有一些问题是我们永远无法估计的， 比如病毒的变异究竟会趋于什么形态，是否某种更高致死率和传染性的变种会在一个个人体进化炉里被孵化出来。

**而最重要的一点，在于理解网络动力学对疾病控制的启发。**如果我们知道了传播的动力学， 我们就可以比较有效的控制一个复杂网络，也就是阻断疾病的传播。比如如何根据病毒的特性制定最有效的阻断策略， 同时不至于杀敌一千自损二千。**从长远看， 它要求我们把对大型疫情的控制， 看成社会网络设计的一部分**， 比如对高危传播路径（人畜）的控制， 对疫情信息的早期播放和传递， 在出现疫情后如何第一时间记录和锁定传播路径切断， 以及如何用最有效的方法构建医疗网络等等，这样的长效措施将会把未来发生比冠状病毒更可怕的传播时候的社会风险降到最低。