大数据视角下的复杂数据与复杂模型

李丰

https://feng.li/

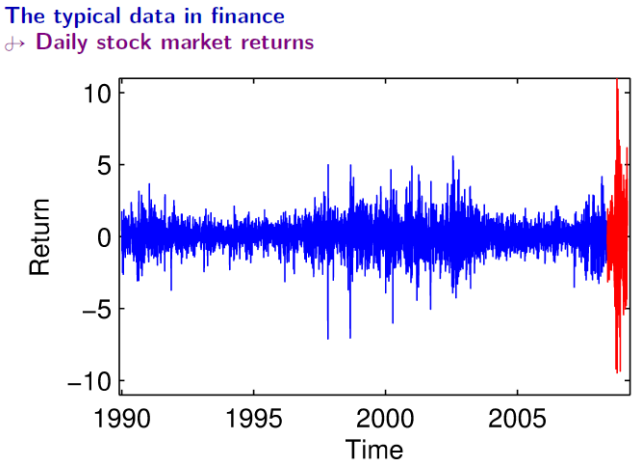
中央财经大学

今天介绍的是复杂模型与复杂数据，主要基于大数据的视角。生活中，我们分析数据时总会想到一些问题，比如说天气预报为什么准或是不准？再比如一封垃圾邮件有没有被识别。大家记得上次金融危机是什么时候，2007-2008 年间。那我们能不能预测下次金融危机？用什么数据来预测下次金融危机？之前预测金融危机时，我们总是采用金融市场数据那现在能否加入其他数据综合考虑？基于以上，我认为**模型和数据的结合是建模的基础，也就是说这一过程在寻找数据关系。**现在的数据日益复杂，数据间关系也如此，因此需要更复杂的模型。

第一部分：日益发展中的数据和模型

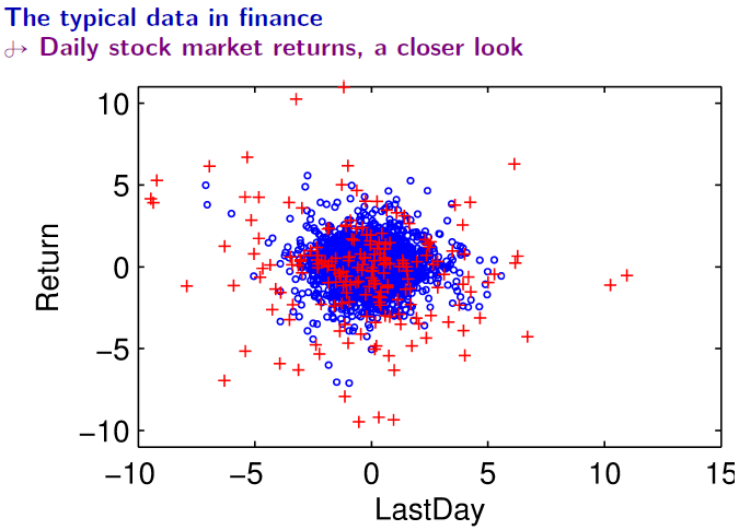
一说到模型，大家会想到曾经学的回归分析是一种简单模型，但回归分析真的如此简单吗？50 年代时，回归分析一度被认为极其高深，比如衡量父亲和孩子身高的关系，又或是衡量家庭的收入和消费，都能通过回归模型描述。基于这些，学者延伸和发展了许多重要理论，如经济理论方面，以及将此运用于时间序列。在当时它们被定义为复杂模型，经过了几十年，模型会越来越复杂，而我们曾认为的复杂模型现在变成了简单模型。

回顾以前我们所认为的复杂模型，现在看来其实往往很简单。这些模型有很简单的假设，如正态性，而且这些模型容易被计算机实现，比如用 R 软件或 Stata 都可以做回归模型，甚至短至一秒能出结果。很多理论成果都建立在这些简单回归模型的扩展上。用金融数据举例，下图是 1990-2006 年股票数据，红色线是 2006 年之后出现金融危机的状态，左边是金融危机未出现时。数据呈上下波动，但这个波动越往右越来越大。波动越来越大是什么原因造成的？当我们仍处于历史位置时，能否发现将来要出现金融危机呢？

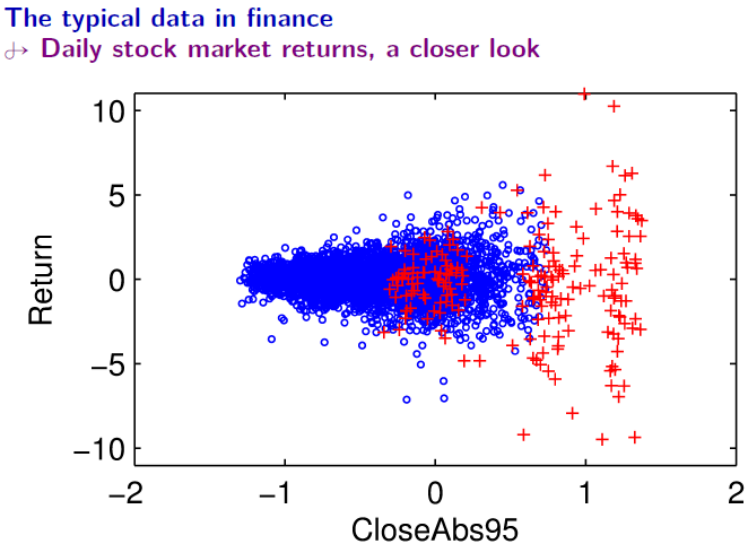


我们做经济模型或统计模型，会有各种各样的假设。经济理论中有一个著名假设——有效市场假说，指今天股票的收益仅和昨天的股票收益有关（给定所有历史的信息）。人

们它描述了很多重要的经济问题，但有时市场却不是有效的。比如下图，它描述了过去的和今天的股票收益，画出散点图，会发现没有经济危机时段内，数据分布很集中，但金融危机时，红色数据会发散，故有效市场假说在这此不再适用。之前建立在有效市场假说上的那些理论到底还能不能用，不能用的话会造成什么问题?



如果我们不仅仅研究过去的股票收益，还研究波动性，在下图，CloseAbs95 描述了过去的波动性，左边是它的收益，你会发现出现金融危机时，过去的波动性确实在不断增大但是存在较好的方式描述这些过去的波动性吗？回顾之前学过的时间序列模型，如自回归模型、移动平均模型，但是比这些更复杂的模型在经济场景中用得多吗？答案往往是否定的。



当经济活动和联系变得越来越频繁和交互，我们会发现仅通过正态或是这些最基础的假设来描述数据间关系远远不够。我们更关心反常情况，比如股票市场正常情况下大盘是

很好的，突然十年中有一年出现了大波动，那么如果哪一天股票波动很大了，我们就应该关心那是不是出现了尖峰厚尾，这些在正态假设里无法识别，所以我们需要更好的特征描述方法，也就是怎么描述它的尖峰及偏态形式。在关心数据的正态与否时，只有在金融危机时人们才意识到“正态”是不正常的，“不正态”才是正常的。事实上，我们之前认为所有数据都是正态的，这本来就有误，很多情况下数据并非我们想象的那样完美，看似一个钟形曲线有可能是左偏的，有可能是右偏，也可能是厚尾。当维度变高时越能体现出来比如金融危机的数据，当股票收益在金融危机期间时，你会发现能够刻画他的自由度，波动性、以及方差会越来越大。

所以在 2007 年金融危机前，人们描述股票市场所用的正态模型，不能够检测到这些尖峰厚尾的形式。也就是说反常情况发生时，我们仍被蒙在鼓里，用错误的模型做出了错误结论。本来已经出现了很强的自由度的变化，它的峰态变化、偏态变化时，我们没有去捕捉这些信息，因为它们的均值和方差仍然在一定的可接受范围内。

很显然如果模型有误，我们会被结论误导。但人们总是这么做，这是为什么？因为正态性的模型很简单，易操作。然而事实上并不知道数据有偏，是非正态的，但我们没有更好的工具描述这些非正态性，而且工作量比实现一个简单正态假设下的回归模型要复杂。

有时我们描述一只股票或单个变量的波动相对还算容易，但描述多只股票的联合波动就会很复杂。但随着全球化进程，我们会发现这种多边联合波动似乎变得习以为常了。比如我们经常关心的是，如果我是一个投资者，有 100 万，想买两只股票，最想看到一只股票涨停，另外一只也跟着涨停，最不想看到一只股票跌停，另外一只也跟着跌。所以我特别关心股票不要同时被跌停，或是不要同时损失很多钱，那怎样来描述？**我们需要新的工具。**

在多元数据里，我们不仅会描述每一变量的尖峰厚尾，包括原有正态性等特征，还要描述数据之间的相互关系。正常情况下有何种相关性，非正常情况下又是什么关系。这种非正常情况下，它们共同发生某种波动时的关系，我们把它叫做尾部相依。尾部相依是金融危机后很多人都在研究的一个问题。之前人们更多关心的是两只或多只股票在正常情况下的静态相关性。尾部相依的关联能够真正地刻画风险，尤其在发生危机的时候的风险。

基于以上，你会发现**建模随着数据、随着维度的考量会越来越复杂，因为考量的角度会增多，数据描述特征也会变多。而且现在数据变得越来越不一样了。**甚至在经济学里边出现了一个新分支——神经经济学（Neuroeconomics），通过研究大脑的三维彩超图及核

磁共振图来描述看到某一个特殊广告场景时，你要买或是不买所推荐的产品。有人会研究你看到两个不同的广告时，把一个核磁共振机放到你头上来检测，看大脑在什么地方会变得兴奋，看到什么样的图会更想去买这个产品。很早之前我们处理的数据也许就几个、几十个变量，或仅仅几个观测值。但现在的数据常常数以百万计的出现。我们之前研究数据时可以用 Excel，鼠标上下拖拽就能大致了解数据规模。现在很多数据却不能用肉眼直接观测。仅仅可视化数据就变成一门单独的学问，所以说**高维且复杂的数据不再罕见，而变成了新常态**（New Normal）。

如果把数据和模型比作道和魔，用魔高一尺，道高一丈来形容他们的发展关系会很贴切。数据变得越复杂，模型就会更复杂。以前用线性回归模型、交叉列链表，现在用 GARCH 模型、更多的高维模型、Copula 模型。

不管是简单模型还是复杂模型，都需要经历五个阶段：数据收集；模型估计；模型评估；模型比较；模型预测。我在读本科时，这里很多部分都没涉及过。第一部分数据收集老师给大家准备好 Excel 数据。或者干脆把数据写在课后的习题处，我们誊抄在 SPSS 软件中。这一部分通常比较简单，直接用 SPSS 把数据放进去即可。后面的部分（模型评估、模型比较）都没有，因为那时我们只有一个模型就是线性模型。预测也很少涉及。但现在完全不一样了，**数据收集变成了整个数据分析的重中之重**。没有收集到好数据，下面就难以进行。因为模型评估、模型比较已经有很多现成的方法，很多非统计、经济学人士都会用。他们只是比我们更会收集数据，比如计算机学家，比如网络工程师，那我们做经济、统计还有用武之地吗?

以前我们通常学习模型时是怎么重视做预测的，但现在大大不同了。如果你有机器学习的背景，会发现现整个机器学习理论最终是以预测为基准的。**现在数据分析过程中，我们把模型预测作为所有的模型比较、评估的主要方法。数据收集和模型预测**这两样成了我们整个统计、经济建模中的两大方面，如果这两方面做不好，对于应用而言，中间再好也无济于事。

我们都听过盲人摸象，不同的人在不同的角度摸着象的感知不一样。其实它完全可以类比现在的大量高维度高通量的数据分析的情况下，**当一个模型仅能描述数据一小部分特征时，就好比盲人摸象，你摸到均值，就好比摸到大象的鼻子，你摸到方差，好比摸到大象的尾巴，但你真正了解这个数据吗**？只有把所有特征都描述清楚，才能知道大象是什么样的。

现在问题来了，数据如此复杂，是不是可以拥有一个巨大无比的模型覆盖到所有数据特征。从机器学习的角度认为这不可行的，因为参数如果比模型所描述的数据量多很多的话，自由度就会差。通俗讲，就是指模型太复杂，模型复杂度超过你的数据数据复杂度了就描述不出什么来了。

随着我们的关注的角度不同，每一个模型对于数据的不同关注点，其特色优势也是不一样的。所以我们应该用特定模型使之能够捕捉数据的某一方面特征并利用它。关注均值往往一个回归模型即可；关注波动性，用一个 GARCH 模型就可以；如果对尾部效应感兴趣，可能需要一个 Copula 模型，但并不是说 Copula 这样的复杂模型就一定比简单线性回归更高级。总之，**我们应该用正确的模型来描述我们关注的重点**，之前更关注均值与方差，现在更关注风险，若更多关注其他维度，用复杂模型是有必要的，但我们还没跨过这个鸿沟，还不能用很简单的措施把复杂的模型应用出来。

一个阻碍就是现在的计算机还没有足够快，不能够把任何模型在很短时间内实现。其次，现在很多计算机上的工具并未广泛应用于经济学统计学模型，比如计算机里常用的并行计算、分布式计算，仍然没有被统计学家和经济学家普遍接受。再次，我们还没有真正把大数据中的“大”和统计模型里的精髓内容相结合，而更多停留在描述数据特征上。

第二部分：大数据基础

总结第一部分，经济发展、模型的变化（如数据变化、数据和模型之间的特征变得越来越复杂）告诉我们：**我们现在缺乏一个合适的工具能真正地把这两方面很好的结合起来**在过去十年内出现一个新名词——大数据，这是计算机学家提出的，他们认为大数据描述了多个不同维度，以及新时代下数据采集、应用、展示的方式。**那么大数据对我们传统的这种学科有什么影响？我们能把我们现在的模型应用其上吗？**这是我们第二部分要讲的。

在大数据时代，人们关心这样几件事，什么是大数据，它能做什么？统计学家、经济学家用大数据能做什么？大数据是不是只存在于互联网企业，我们能用吗？用 GDP、GNP 这类数据的传统学科还有用武之地吗？我们能用新的模型吗？这些模型能用在这些大数据上吗？需要些什么工具？

早前 IBM 对很多高管做过调查，大都认为大数据是一个极有用的工具（这里的“大数据”是一个名词），其中不少认为大数据将彻底改变企业管理甚至人们的生活方式。但**如果我们缺乏利用大数据的工具，那将直接阻碍大数据在企业和日常生活中的应用。也就是我们在很多领域现在还不能把“大数据”当作一个“动词”来使用。**为什么很多大数据工

具、应用出现在互联网企业里？这是由于互联网企业掌握了更多处理大数据的工具，但它们用的这些工具是不是经济、统计模型里最好的工具？那可不一定。

以前很多统计人和经济人认为大数据来了，我们都会丢掉饭碗。但现在想来，其实大可不必这么说。因为这些大数据工具终有一天所有人都会用，就像人们都会用 ffice，当这个工具被所有人掌握时，我们就有了优势，我们能不能把我们的数据在这样的工具下给他玩转了？能不能真正从中发现我们的价值？从经济学、统计学的角度来描述数据特征？虽然现在我们还没有达到那个阶段。

为什么说现在还没有呢？因为这方面教育太少，我们没受过这样的训练，甚至不编程，很多经济学家统计学家用的是 Stata/ SPSS（现在的处理大数据的方式不适用）。而企业（如互联网企业）数据都存储在分布式仓库，通过 Hive 提取，Hadoop 实现数据模型的操作，Spark 执行机器学习算法，Stata、R 无处可用。所以**要想让我们在这些数据上有用武之地，必须要做一件事：我们得和他们对话，得了解这样的数据。**

那么怎么了解这个数据？**有必要用数据的大数据特征来描述，用五个 V**（Volume、Velocity、Variety、Value、Veracity）**描述大数据的特点。**举例而言，一个 TB 的下一个数量级是一个 PB，一个 PB 是多少呢？是一千个 TB，相当于你有一千个笔记本的硬盘，那一千个笔记本硬盘能装多少？《阿凡达》里用一个 PB 数据来存储所有的 3D 文字，美国的国会图书馆用十个 PB 数据存储很多资料，目前世界印刷材料总容量也就 300 个 PB 左右。

从前有新闻播过人类基因组 1%基因的测序，1%的基因测序是多少呢？1 克 DNA 可以存储 360 个 EB 的信息，1 个 EB 是 1000 个 PB，1 个 PB 是 1000 个 TB，1 个 TB 是一个笔记本的硬盘。所以在 2000 年左右我们能完成 1%的基因测序，很了不起。现在的基因测序可以做全系列测序，如果家里有孕妇，你会发现医院经常让孕妇做婴儿或母婴的基因的测序，现在这很普遍，通过大规模信息筛查看有没有疾病表象。

世界海量大数据里，哪些数据占很大比重呢？之前有一个误导认为社交媒体上的数据占很大比重，事实上正好相反。真正的数据被存在商业数据里。有统计表明全世界 90%的数据都是在过去几年被创造的，包括人类自有文明以来所有数据，即人类文明 5000 年的数据比不上过去两年创造的数据。所以说这种**大数据变革必然带来方法大变革，必须要有新工具来处理**，很早之前这种工具只能存在于高技术公司，如谷歌、百度，后来发展到社交网络或电商平台，而现在是全民大数据，至少在中国，现在人人都在说大数据。

那么有多少人真正了解大数据呢？我认为现在提大数据，起码在中国的很多传统行业，可以把“大”字去掉。十年前，我们开始谈数据，而之前是不谈数据的，也就是说现在做任何一件事，要考虑有没有数据支撑，而之前想做就一股脑做了。所以过去人们投资时并不像现在这么理性，因为现在数据运用得很广泛，很多有趣的领域被开发了。

大数据各维度的数据可以被应用在各领域。最近听说一个有意思的数据，某个在线商城一个专门卖骑行设备的平台，无意中发现有人买骑行设备时会买面膜，当时不明白。后来发现，户外骑行皮肤会晒黑，所以人们喜欢晚上敷面膜，这就是新商业契机。之前人们摆放产品时按不同的品类，比如自行车区就只摆放自行车相关商品，但现在不一样了，会把你需要的一篮子商品重新摆放，摆放在最容易拿到的地方，所以商场里如何规划商品的摆放能直接影响商品销量。

当然**大数据也是双刃剑**，比如社交软件，以前一大创新叫“发现你的好朋友”，如果张三有个好朋友是李四，王五有个好朋友也是李四，然后就把王五推荐给张三，这能扩展社交圈。但数据是没有感情的，比如一个人刚把他的前女友删掉，而他最近认识了不久要结婚的未婚妻，那么数据分析算法就会根据他们三个有历史关系甚至有可能会把前女友推荐给他的未婚妻。这类情形是大数据所不能够解决的问题，所以**大数据对于这种数据感情的描述还不足。**

此外，很多传统方案都受到大数据的冲击，比如门店里销售的药品，我们往往不能精准控制库存。现在却能结合很多因素综合考虑，比如根据当地的天气预报预测当地感冒流行程度，据此精准进药，这样库存会变得更动态。这一系列的解决方案不仅仅依托大数据而是把其他数据和内部的数据结合，得到新数据的解决方案，极大提高了预测能力。

第三部分：如何掌握大数据分析技能

上一节讲到，大数据背景下，数据变复杂了，但是我们没有合适的工具。接下来呢我们再回到统计学家经济学家的角度，我们需要补一些什么课程？这些课程能帮助我们提高什么样的认知？即能不能学到那些获取大数据的方式、跟传统的数理经济模型结合？

我初来中央财经大学时，周围许多老师同学都会遇到各种问题。有老师关心企业财务报表，但很不幸财务报表都通过 PDF 格式发布出来，不能从中迅速找到信息，比如每一只股票里的信息通报，是被惩罚了还是有些信息披露，怎么从网站的公告里获取；有人想研究散户情绪如何影响大盘，去哪里找这些数据？一个传统的学者往往无从下手。对同学来说，想做这样的研究，能否得到相应的大数据；还有人想研究最近兴起的一个产业 P2P 平

台的借贷融资方向、形式，能否这样的数据。如果得不到所需数据，那这些新兴事物是不是就和我们无关？我认为**我们有必要懂大数据的语言，需要掌握数据科学工具，去和大数据对话，让大数据能为我们的模型、数据的理解服务。我们现在缺乏的是在大数据时代利用现代的数据科学工具独立地获取一手数据的能力。**

说到获取一手数据，哪里有一手数据？是不是统计局或各个数据库里的才是一手的呢？并不是。我们的互联网大数据，比如互联网搜索引擎的数据从何而来？其实也是一点点加工得来的，但并不是说互联网的数据都是互联网程序员去网站上复制粘贴的，他有一套现代数据科学工具，能够帮他们整理数据，把非结构化数据变成结构化数据，再把它们转化成价值。这也是我们需要训练自己的地方，**掌握这些工具链，利用数据分析优势，来发掘复杂数据中的价值，也是在锻炼我们这种数据科学的思维模式。**只有把这两点结合起来，才能把我们掌握的经济的洞察的模型、经济统计里独特的知识应用在上面，然后才有我们的用武之地。**如果没有一手数据，那我们并不比任何人有优势**。

但这并非易事，现在的数据处理模式不像过去“一个软件打天下”，而是一个软件对应一个专长，比如数据处理有专门的数据处理软件，数据的分布式存储需要分布式存储的软件，调取这些分布式存储里的数据还需额外软件，需要一个**工具链**。这给学习增加了挑战，因为我们没经历过这样的训练模式。我们之前训练模式是老师告诉大家模型在哪里，数据也已经备好，学生把数据放到模型里得到结论。但当数据还没有准备好的时候，能不能训练自己去获取？尤其对于现在的本科生，如果想学习一个新技术，一定要抓紧学习，否则以后再想学习会很困难。过去我教过研究生和本科生大数据的相关课程，大概总结了一下，我认为学生需要掌握以下八个方向。

首先我们得明白**数据核心价值到底在哪里？**不是说把所有的数据堆积拼凑在一起才能够得到它的核心价值，我们需要对数据价值有新认识。其次我们要**理解现在的数据生产环境和工具**，尤其是在数据生产环境里，数据来源五花八门，现在数据采集和传统的调查问卷和入户调查不再一样（其实这些传统工具也变得不一样），我们需要理解这些工具是如何产生和存储数据的，而数据存储后又如何存放。也就是第三点，我们应该**有能力创建自己的大数据仓库**。这里的大数据仓库不是像百度一样的大数据仓库，而是建立一个你感兴趣的数据容器。在这一点上，尤其对低年级学生而言，当你对某一个特定方向感兴趣，你应该从一开始就积累，积累越多，数据自然而然会扩充。百度的大量数据也不是一天采集而成，而是经过数十年收集和分析得到的，所以如果我们可以建一个能够代表自己研究学

习方向的有用的数据库，那将来你就是这个方向数据库的创造者和数据理解的第一个使用者。

同样的当数据变复杂后，你会发现原始 GDP、GNP 这样的数值性数据已不再是唯一的数据类型，还有图像数据、声音数据、文本数据等，**如何把非结构化数据整理成模型能够辨认的数据，即非结构化的数据处理也是我们要掌握的，所以需要新工具，**如 Python 等。

以前数据可视化并不是一个重要领域，但现在不一样了，数据源于各领域，且越来越复杂，维度越来越高，如果没有一个好的数据透视工具就无法直观理解数据模式，无异于盲人摸象；所以**要学会如何透视数据，也就是做模型前，在揣摩这个数据的时候，可视化是必经的一条路。**

我们学过很多模型，但之前的模型大多在单机上执行，如 stata、R 软件，如果现在有五个 T 的阿凡达数据，你想知道这段时间里其中某一个统计量，是无法完成的。所以**有必要学习如何把统计模型应用在这样分布式存储的数据上面，如何把一个回归分析放在一个 T 数据上运行，数据量大了又该如何做，是抽样还是把模型做分布式，解决这些问题需要掌握相应的技术**，做到这六条，基本能力也就有了，接下来就是**发挥自身优势，把经济上统计上的、我们独特的见解和和大数据工具结合，实现大数据的价值。在这**之前我认为数据科学和我们自己无关，但如果前六条被所有人都学会了，那么我们的优势则在第七条。

如何把上述七条学好，可以从一些简单的数据科学的应用案例入手。你觉得这个案例需要用到什么工具，遇到什么难点的时候去学习，缺什么补什么，这可能是最快也是最有效的，让你能够把这七点全部结合，如果单独研究每一点，那每一点都能写成博士论文了所以要想真正实现数据的价值，不是要把这七点全掌握，而是要把它们间的工具学会。

大数据背景下一种新的科学素养，或新的一体化的知识结构正在被人们逐渐认可，因为数据科学在不断改变业界学者及学生的思维模式，加强对数据科学工具的使用能极大提高我们自身的专业技能和专业价值，这既是对大数据时代的融入，也是未来发展的方向。**未来专业将不再是割裂的，必然是各个方向的大融合**，我们要把这些工具不断结合、综合起来。在现在大数据背景下，其实更多强调的是以实际应用为导向的知识思维模式和学习模式，所以作为初学者，我个人建议一定要**避免盲目，应有效平衡数据科学内容繁琐和学习精力有限的矛盾，所以刚刚提及的从数据科学的案例入手是一个有效的解决方法。**

其次我认为做数据科学或利用数据科学解决专业问题，很多时候应该将数据科学知识和专业知识结合，，同现代的数据对话进而使之更好地用于所学，这样才能培养跨界融合

的创新精神。譬如教学中我们老师试着引入现代交互式的教学模式，老师可以告诉学生学习目标是什么，需要什么工具，让学生自己寻找，能不断校正学习模式和学习方法，从而提高我们用一手数据探究问题本质的能力。

知识变得越来越多，要掌握的知识结构也趋于复杂，老师不可能教会学生所有工具，更多应该引领学生，告诉他们正确的研究方向。这样他们会发展得更快，互相促进，迈向更多新领域。

感谢各位老师和同学的聆听，今天跟大家分享的就是这些，还请大家多指正。谢谢！

整理：巩林静、张咪、余璐、楚伊晨

校对：姜嘉琪