06 全体大会

麻省理工学院教授 Alex Pentland：下一代通用人工智能——分布式、长尾、非静态、隐私保护和加密

整理：智源社区 高洛生

Alex Pentland本次演讲主题是《Next Generation AI: Distributed, Long-Tailed, Non-Stationary, Privacy-Preserving and Encrypted》。

Alex Pentland，美国国家科学院院士，麻省理工学院 Connection Science 实验室主任，世界上被引用最多的计算科学家之一。2012 年，与谷歌联合创始人 Larry Page 一起被《福布斯》列入“世界上最强大的七位数据科学家”。2013 年获《哈佛商业评论》颁发的麦肯锡奖。帮助麻省理工学院创建媒体实验室和位于印度的亚洲媒体实验室，受到包括法国、意大利、澳大利亚、哥伦比亚等国家和世界银行、万事达等组织和机构的资助。作为计算科学家之一，福布斯曾将他与 Google 创始人拉里佩齐等人称为世界上最具影响力的七大数据科学家。此外，Alex Pentland 教授也是计算社会科学、组织工程、可穿戴计算、图像理解和现代生物识别学的先驱，被誉为“可穿戴设备之父”。曾经共同领导达沃斯经济论坛的讨论，通过与世界各地的领导人进行对话，了解目前政府和工业界对人工智能的需求。并且担任美国律师协会、美国电话电报公司的董事会成员。

智源社区编辑根据 Alex Pentland 的现场演讲，在不改变原意的基础上整理如下。

一、安全与隐私

隐私是每个人都非常关心的，隐私不是加密的问题，而是哪些事情该做、哪些事情不该做的问题。Pentland 介绍，他已经说服许多欧盟国家和大型企业广泛地共享数据资源，但这种共享并非直接共享数据本身。GDPR (General Data Protection Regulation，通用数据保护条例) 的关键核心在于没有数据库，不将数据移动到中央存储器。因此这就是人们常说的开放算法，提出问题的人和拥有数据的人必须理解算法的作用，知道数据是否安全，知道数据是否被正确地应用，而不是把来自不同国家、不同公司的数据资源放在一起。开放算法没有真正分享数据，只是分享数据的认证证书。当然，随之而来的就是综合学习 (Comprehensive Learning)，能够把人工智能算法分解成在本地运行的小块然后进行组合，很多技术都可以实现这一点，例如安全的多方计算，使用非常安全的硬件，进行同态加密等。Pentland 认为该方面的研究尚未充足，因此其非常热衷于此。

接下来，Pentland 分享了他对 AI 未来机遇的思考，特别是在金融方面。此次新冠疫情让世界的贸易体系有了巨大的改变，人们不再频繁地旅行，导致在商业领域数字技术开始有了较大的发展。此次疫情让人们真正想要学习如何应用数字技术，改变政府、企业和个人的工作方式，由此也给 AI 带来了巨大的机遇。同时，现在网络犯罪也非常频繁，许多专家认为未来将会有更多网络犯罪，以后 AI 的很多领域都会和 5G 联系在一起，所以 5G 将会发挥很重要的作用。如今人们开始越来越多地接受新兴技术，同时需要更多的安全性，包括金融安全、隐私安全等等。此外，教育领域和商业领域也在发生着巨大的改变。每年 MIT 都会举办一场经济论坛，邀请世界各地的商业领袖谈一谈他们如何应对新兴趋势，上一场是在今年 1 月份举办（详情可在 http://imaginationinaction.xyz 中查看）的，当时疫情刚刚开始，而现在商业界（包括很多大型银行）正在制定新的数字化计划。尽管商界都有在未来十年进行数字化的计划，但 Pentland 希望可以在几个月就能完成这一转变。

最近 Facebook 推出了一款新型虚拟加密货币 Libra，它确实是一款非常棒的产品，目前已经在许多城市试点，如果未来几年想应用于更广泛的领域，重点应在于如何与比特币相结合。Libra 是基于比特币建立起来的一种虚拟加密货币，但是如何让全世界的商业巨头使用呢？这些商业巨头掌控着海量的消费者金钱，同时也掌握着海量的消费者数据，不断分发着金钱和数据，这个过程就涉及到非常重要的隐私。如今美国、中国和巴基斯坦等国家的人们也在使用类似的应用，例如 Tradecoin 作为电力交易工具已经在巴基斯坦应用，这种应用并不属于巴基斯坦政府，而是属于这些能源的使用者，所以并不一定要依赖政府资助，可以通过社区建立起这种应用。

图 1：人工智能在金融领域的应用

图 1 所示是 Pentland 参与建立的世界银行应用系统。大多数国家的政府都没有非常完善的数字化系统，Pentland 所做的就是创建开源系统，通过区块链技术让政府体系变得更加高效，也让运作成本相比之前更低。另外一个例子是 http://law.mit.edu，该网站展示了有关计算机方面的一些法律案例。很多人都会问这些数字活动是不是合法的？有没有跨越法律的边界？是否需要改变现有法律来让 AI 得到进一步的应用？这些疑问也是 Pentland 团队所聚焦的重点，如何能够抓住这些监管体系改变 AI 带来的机遇，比如长尾配送体系、保证这些金融数据的安全性等。现在越来越多的人认为当下所用的人工智能这个词是错误的，应该叫做增强智能。在过去“人工智能”这个词被发明出来主要是为了博得人们的关注，实际上现在所做的大多数事情都不能算是人工智能，而是自主智能，AI 真正需要做到的是更加智能化、更加人性化，比如用于政府体系和商业体系的应用，这也为监管体系带来了更大的挑战。不过随着现在应用场景的不断变化，更需要用人类的智能去找出人工智能在理解上的各种错误。

二、Human Network Dynamics

如今伴随着人们在世界各个城市之间不断流动，几乎所有的 Human Network 模型都是建立在人类活动不断重复的基础上，而且实际情况更加复杂。例如观察一个人的行踪，有些人群的行踪是比较相似的：有些地方他们特别喜欢，有些事情他们特别热衷，但如果这类人在同一件事情上花了同样多的时间，说明他们可能都会热衷于此事。这是一种行为的相近，包括疾病、金融亦是如此，他们的行为正在变得越来越近似。

Pentland 团队也对该问题发表了大量论文，观察世界各个主要城市人群（并不针对个人的观察，而是针对整个

街区的人群）的活动特征，希望了解人们感兴趣事情的相似性，由此建立起不同地域之间的联系。Pentland 希望通过观察了解人们进行购买或者生病的相似性，最后通过社会关联分析得出更好的结论。因为数据分析方式可以获得更多人群的行为特征，因此这种分析方法需要受到更加严格的法律监管。

Pentland 认为，可以通过观察城市当中的某个特定的地域，比如通过观察某个商场的多样性，以及来到该商场人群的多样性，来判断出哪个社区的人群增长要高于其它社区等，这种判断的准确性极高，至少可以达到 50% 以上，高于其它的所有方法。Pentland 认为这种方法可以应用于亚洲、欧洲和美国等，它们几乎每个地区都可以应用。Pentland 还将该方法用于基础设施建设的分析上，例如应该在哪里建造地铁才会最有利于城市建设，甚至可以在商店建立之前就预测出商店全天的营业额，因为通过分析可以得到某个地域人群的出行频率。

此外，Pentland 又列举了长尾配送体系的例子，认为真正的配送体系会有更多的长尾效应，单看每个样本似乎都很正常，实则会有不同的变体，这些变体也会发生更多的改变，所以很难抓住这种长尾配送体系的特征。所以如果要应用长尾配送体系，还需要进行更多的训练。通过分析配送体系的特征，实际上长尾体系已经非常普遍，在人们的日常工作当中随处可见，每时每刻都在发生变化，但人类未必能够及时做出调整。

三、增强智能与人工智能

与人工智能相比，增强智能可以得到什么？ Pentland 认为实际上目前使用的大多数人工智能算法并不尽如人意，需要人的介入和辅助，否则就不会达到预想的工作结果，所以人类的监督是至关重要的，同时又必须确保人工智能算法所做是合法的。如何才能做到这一点？必须要有一系列的方法确保 AI 的思维和人一样，这也是 Pentland 团队正在研究的方向。比如人类如何做出金融决策？人类往往会通过过去的表现对未来做出判断，AI 则是根据流行性来决定是否适用自己。人类不断采用不同新的工作方式，根据具体事物的价值进行投资。人类总是不停地在做出选择，特别是对于人类来说这让变数变得更小了，可能会得出很糟糕的结果。

接下来，Pentland 举了一个团队成长的例子。加入现在想组建一个工作团队，可以找到表现最好的员工，然后将其天赋才能传授给其他人，最后整合成为一支最棒的团队，这种方法听起来十分合理，但是通过不断评估团队每个成员的表现，然后又不断有新的成员加入团队，这些成员可能刚开始表现不是那么很好，通过对整个团队进行评估，就可以得到相对更好的结果，也能够培养出表现更好的人员。目前，Pentland 团队已经在《科学》杂志上面发表了一篇关于该方面的文章。

Pentland 通过这种方法帮助客户更好地制定策略，做出更好的金融决策并取得更好的回报，他认为尤其是在如今这样一个不断变化的市场环境当中，选择对自己的品类最有用的策略是非常有用的。

在演讲最后，Pentland 分享了加密数据模式的话题。现在用户需要时刻小心网络攻击和网络犯罪，因为大多数时候用户看不到数据本身，只能看到和数据相关的主体。当两个主体产生互动，双方可能对对方的行为均作出反应，也有可能单方面作出反应，例如老板与员工之间的互动。这种互动架构有许多因素需要进行判断，一个工作人员的行为可能影响到另一个工作人员，如果这种行为是非常具有影响力的，那么就可以影响到大多数工作人员，通过这种方法可以提高工作效率。假设需要观察 6 个工作人员的表现，用户仅知道他们之间是有联系的，但用户不知道如何把他们联系在一起，所以通过该方法可以同时观测这些工作人员的行为，然后得出他们之间的工作特征，能够把他们的每个个体从整体当中分离出来。

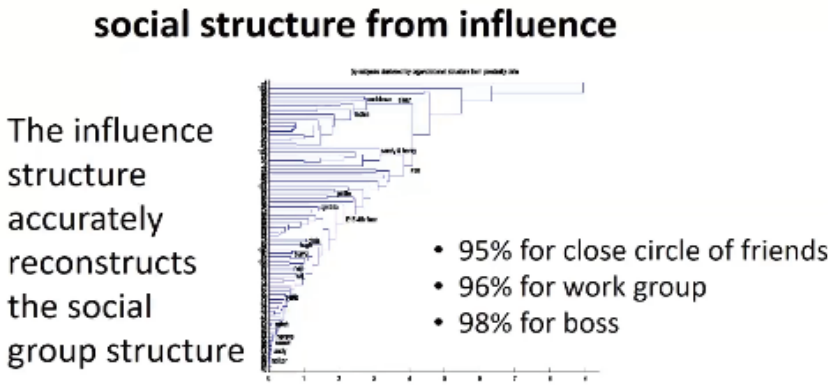

图 2：受影响的社会结构

从图 2 可以观察出一个人如何影响另一个人。如果一个人出现，另一个人是否出现？实际上在一群人当中可以判断谁是员工，谁是老板。没有做标注的数据，人们对这些数据完全不懂，但是这些数据可以反应潜在的社会关系架构。另外，还可以通过该项技术进行新冠疫情的判断，因为许多社区之间是互相联系的，可以通过观察各大药店和社区，了解这些地方是否会出现疫情的爆发。

最后，Alex Pentland 介绍了其出版的新书《Building The New Economy》，已发布了网络版，欢迎广大读者阅读、评论。

美国工程院院士 Anil K. Jain：模式识别——从统计学到深度网络

整理：智源社区 韩鹏飞

Anil K. Jain 本次的报告主题为《Pattern Recognition: Statistics to Deep Networks》。

Anil Kumar Jain 是美国密歇根州立大学 (Michigan State University) 杰出教授，美国工程院院士，印度工程院外籍院士，中国科学院外籍院士，发展中国家科学院院士。研究领域包括模式识别、计算机视觉和生物特征识别，是多个国际著名学术组织如 ACM、IEEE、AAAS、IAPR、SPIE 等的 Fellow。曾获得的荣誉包括 Guggenheim、Humboldt、Fulbright、King-Sun Fu Prize 等。曾担任模式识别领域最权威的学术期刊《IEEE Transactions on Pattern Analysis and Machine Intelligence》主编。目前已经出版了《Handbook of Face Recognition》、《Handbook of Fingerprint Recognition》和《Handbook of Multibio metrics》等多部专著，以及数百篇高水平学术论文，其中包括《Nature》论文 1 篇，IEEE Tran. PAMI 论文 95 篇。他在人脸识别、指纹识别等方面的多项研究成果被 NEC、Morpho 等国际生物特征识别公司使用，在学术界和工业界具有极高的知名度和影响力，他是全球计算机学科论文引用率最高的学者，Google h-index 为 181，Google 引用次数超 21 万次。个人主页 http://www.cse.msu.edu/~jain/

现如今，生物特征识别、机器学习、深度学习和计算机视觉等领域研究的本质都是相似的，就是要让机器可以做到一些我们觉得智能的事情。这一概念实际上是五六十年前提出的，当时主要是想创造一种智能机器人，但这种尝试并不是很成功，最后导致所谓的“人工智能的寒冬”。我们所关注的重点是人工智能到底可以解决什么样的问题？数据模型的建立是很困难的，我们需要模型告诉我们更加准确的结果，这样才能了解模型什么时候有效什么时候无效。Deep Network 是不是这个产业的终结？会不会再次迎来一场“人工智能的寒冬”？

在演讲中，Anil K. Jain 结合人工智能的发展历史，介绍了模式识别的理论和技术演进，其经历了由模型驱动的统计方法，到现在数据驱动的深度神经网络方法，但后者还面临一些挑战亟待我们去解决，比如可解释性、鲁棒性等。

一、人工智能技术的起源

Anil K. Jain 介绍，人工智能 [1] 一词是在 1956 年被 McCarthy 等人提出的，他们将能够让机器像人一样思考和行动的方式称之为智能。不过这种机器能够做所有智能的行动的这个愿望似乎已经落空了，到现在也没有实现，特别是在自然语言处理等领域做的还不是很好。

Anil K. Jain 认为重要的是需要了解人工智能的具体作用是什么，它可能是多方面的，同时想让人工智能掌握一种知识，就需要了解很多相关的领域知识 (Domain Knowledge)，比如隐私和安全等，而且需要大量的 Label Training，从而提高判断的准确度。

实际上早期的人工智能也做了一些模式识别 [2] 的工作，但人工智能更关注的是通用智能；而模式识别关注的是在一些具体的领域来实现智能。

接下来，Anil K. Jain 列举出了过去 15 年中，人工智能领域具有较大影响力的一些工作 [3]，如特斯拉的自动驾驶、Apple 公司的指纹识别与人脸识别等都是在具体应用中完成了智能任务。具体如下图 1 所示：

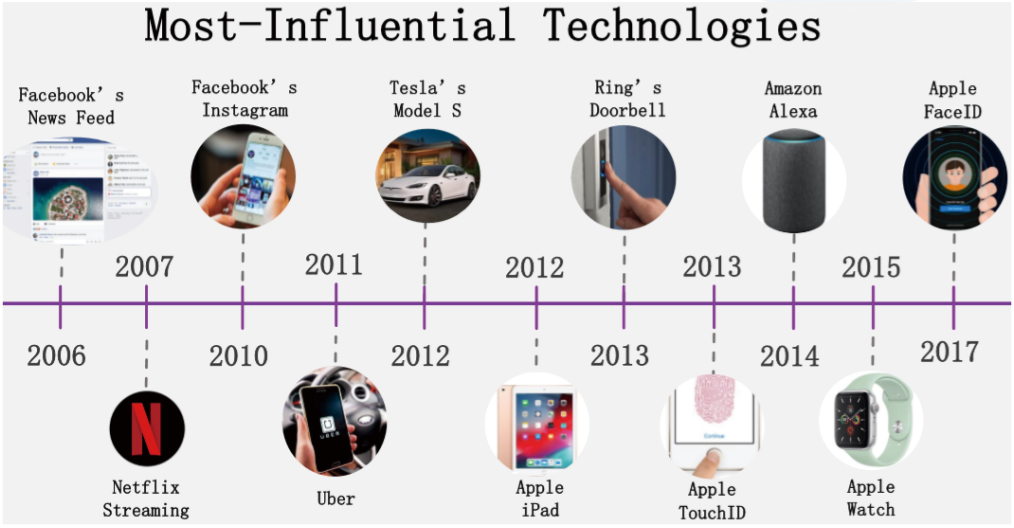

图 1：AI 代表性工作发展历程图

但是随即而来的，是人们对于人工智能这一概念的过度炒作，其现有水平还无法达到人们的预期，比如 Google 和 Uber 每年都在宣传人工智能，但实际上他们都对智能驾驶的预期过于乐观了[4,5]。

二、模型驱动到深度网络

接下来，Anil K. Jain 介绍了模式识别的几个相关概念。模式识别是指从不相关的细节背景中提取显著特征，比如脸部识别、动物识别、指纹识别，这些都是基于对某种特征的提取，包括新冠肺炎患者也有自己的特征，但在几个月之前可能识别不出来。

类是把相似的但不一定完全相同的要素放到一起的集合，其可以有不同的形式。在计算机视觉和机器学习中，模式类是由模型或示例来定义的。机器可以通过对这个类的学习，从新的样本中找出一些这样的模式类别，正如我们教育孩子们，如何正确认出狗、汽车、帽子等不同的类别。

相似度就是比较两个事物的相似性，是智能系统的一个重要基础。比如我们没有猫和狗的特征定级，但我们要让人工智能区分到底什么是猫或者狗，实际上这就是人工智能需要做的工作，而这种模型是很难建立起来的，因为我们所涉及的每个领域都有一定的相似性。比如你的朋友走在大街上，如何让人工智能区分哪个人是你的朋友或者不是你的朋友？因为你的朋友可能和很多人都很相似，这就是最具挑战性的地方。

之后，Anil K. Jain 为我们展示了如何从一个简单的人脸识别问题，上升到一个复杂的识别工作。

下图 2 所示的是印度伟人甘地的一些不同特征，大家看到后肯定会说这些照片都是甘地，但是对于人脸识别系统来说就很难准确判断了，它仍然面对着一些挑战需要解决，可能需要应用更多以数据为驱动的技术。比如我们如何确定哪个人属于哪个等级？如下图 3 所示的这些孩子的脸长得都一样，因为他们是四胞胎，但当我们把他们的头发剃光，然后把四个孩子标注一二三四，人工智能就识别不出来了，所以这是一个比较有挑战性的任务，也会带来很多问题。要么使用模型，要么使用数据，但必须确保这些数据和模型的准确性和可靠性。

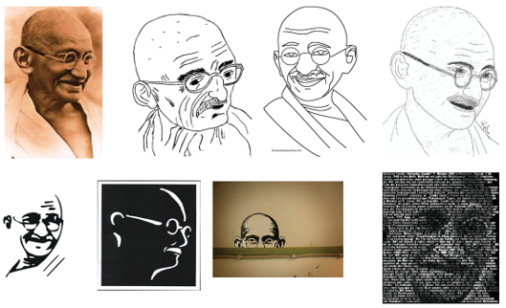

图 2：类内变异性

图 3：类间相似性

这里，我们要解决现在所面临的问题，就必须进行最为真实的表征，利用领域知识进行表征，必须明确 Domain Expertise。以指纹为例，目前世界上共有 76 亿人口，每个人的指纹都是不同的，所以要想对 76 亿人的指纹进行特征抓取也是很困难的。过去指纹的表征都是基于 Flow Pattern，全局 1 级特征（如下图 4 所示），局部 2 级特征（如下图 5 所示）。

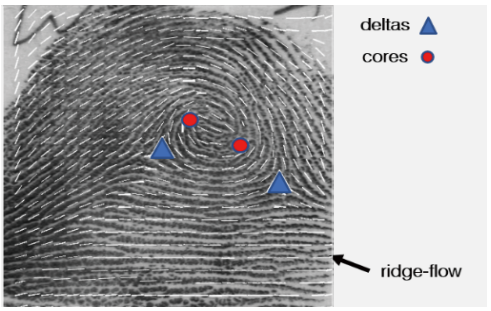

图 4：全局 Level–1 特征

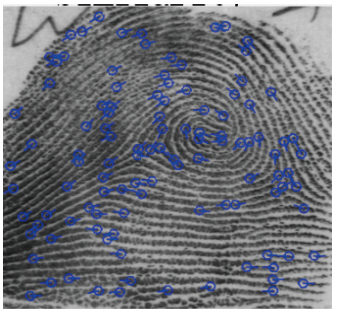

图 5：局部 Level–2 特征

现在深度学习中采用的图表征（如下图 6 所示），但它并不知道领域知识，而是根据不同的图去学习、进行表征，得到一些矢量。然后我们将多个表征进行融合（如下图 7 所示），来提高指纹识别的质量。

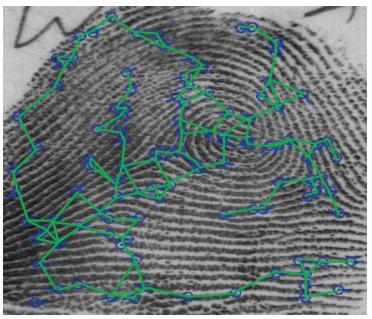

图 6：图表征

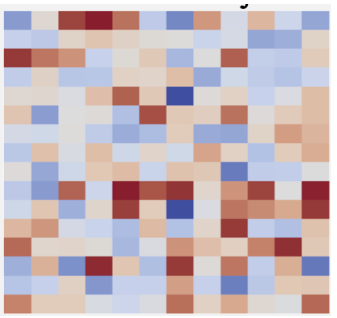

图 7：定长表征

学习能力是任何人工智能进步的基础，但如何进行学习呢？有些是在监督下的学习，有些是无监督的学习。监督学习，简单来说就是给定一定的训练样本（这里一定要注意，这个样本是既有数据，也有数据相对应的结果），

并利用这个样本进行训练得到一个模型（也可以说就是一个函数），然后利用这个模型，将所有的输入映射为相应的输出，之后对输出进行简单的判断从而达到了分类（或者说回归）的目的。如下图 8 所示。

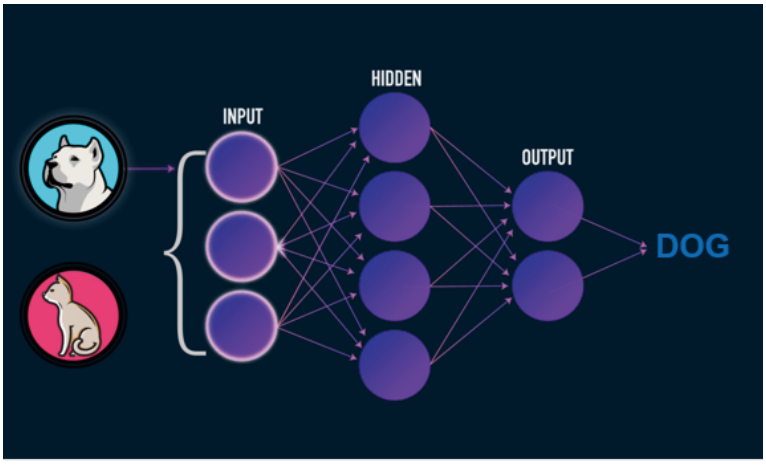

图 8：监督学习（分类）

无监督学习，则是我们提供大量数据，但是这些数据没有对应的标签，由算法来提取具体结构进行分类。聚类算法就是无监督学习的一种，如下图 9 所示，系统会自动把所有看着像猫和看着像狗的事物放在一起。

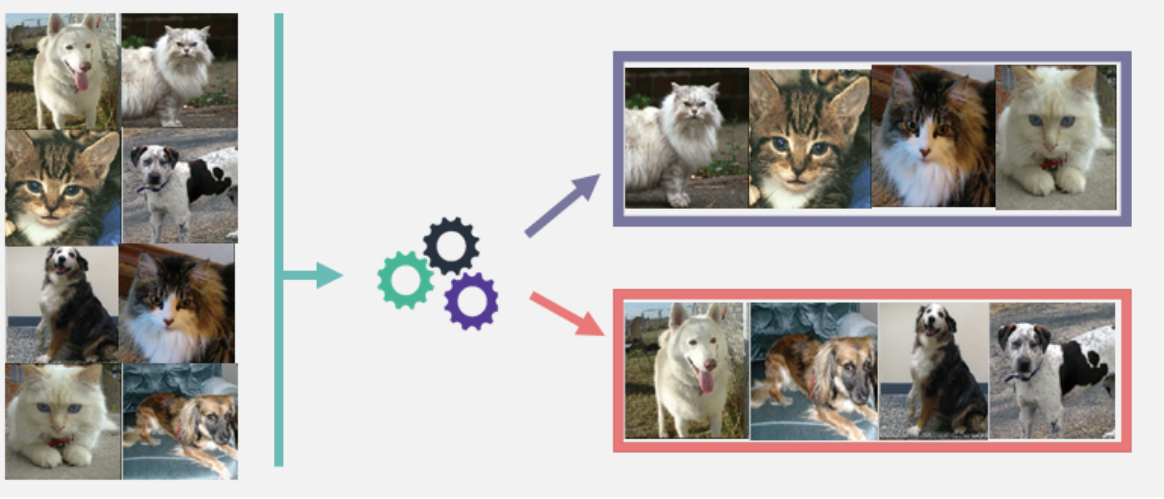

图 9：无监督学习（聚类）

下面，Anil K. Jain 开始介绍模式识别技术的发展。最初的识别技术，是一种模型驱动的方式，如 Linear Discriminant[6]。Linear Discriminant 的思想：给定训练样本例集，设法将其投影在一条直线上，使得同类例的投影点尽可能近，异类样例尽可能地远离；在对新样本进行分类时，将其投影到同样的这条直线上，再根据投影点的位置来确定新样本的类别。其具体计算方式如下图所示。

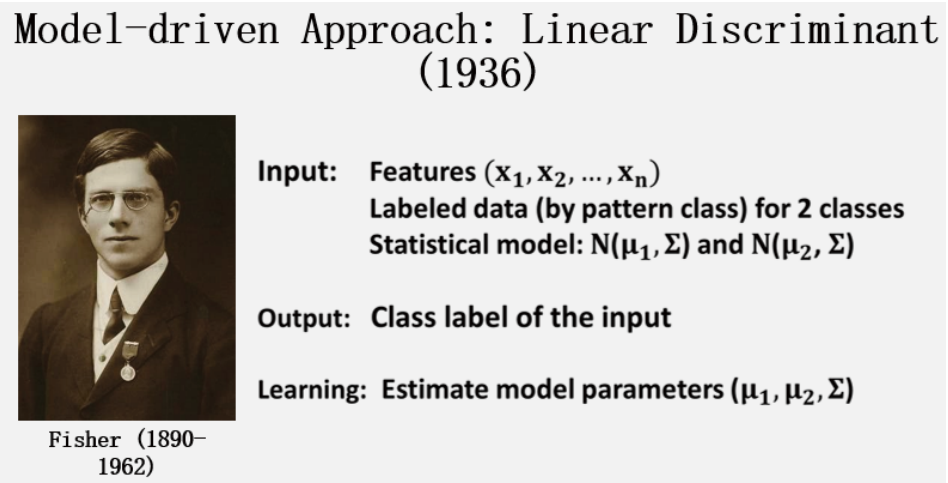

图 10：Linear Discriminant 算法流程

另一种方式是数据驱动，最开始的技术被称之为感知机 [7]。感知机是二类分类的线性分类模型，旨在求出将训练数据进行线性划分的分离超平面，从而导入基于误分类的损失函数，利用梯度下降法对损失函数进行极小化，最后求得感知机模型。其计算如图 11 所示:

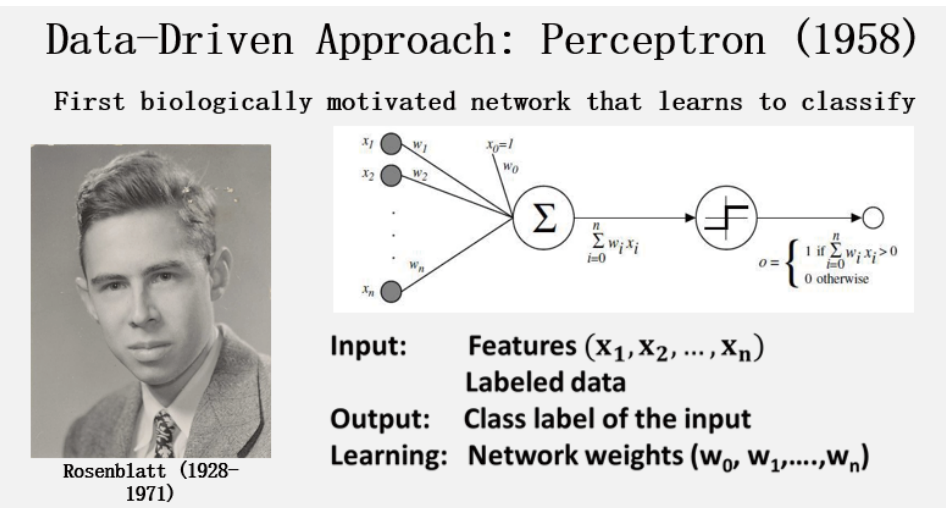

图 11：Perceptron 算法流程

Anil K. Jain 认为线性判别法和感知机方法都具有其局限性，都对非线性可分离数据不起作用，接下来他开始讨论线性到二次分类器和支持向量机 [8,9]，认为首先要找到一个机制，以便统计模型可以看到不平等的矩阵，并利用非线性的线性核来界定数据，然后可以把数据转换成线性可分离的空间。如图 12，13 所示。

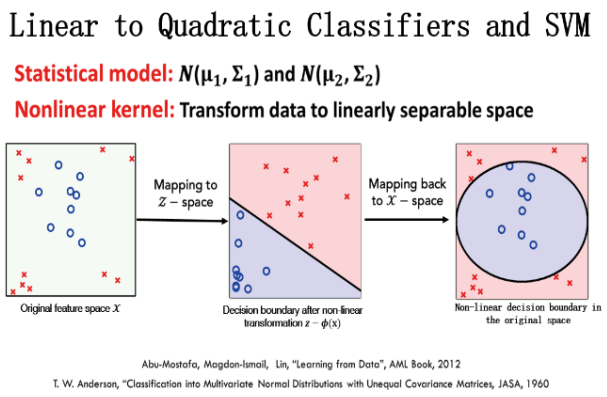

图 12：线性到二次分类器和支持向量机

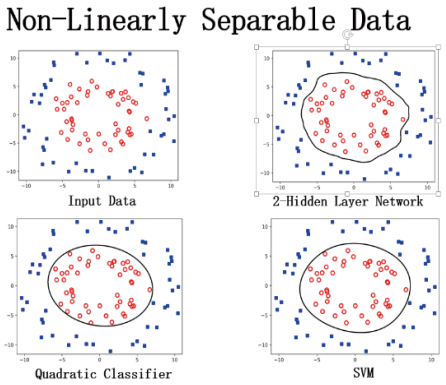

图 13：非线性可分离数据

接下来，Anil K. Jain 的视野从感知机 [10] 拓展到多层神经网络 [11]。如下图 14 所示图中的神经网络中有很多的非线性，在感知机中我们需要学习的参数只有 7 个，而在神经网络当中有 47 个参数要去学习，有很多不知道的因素需要去考虑。

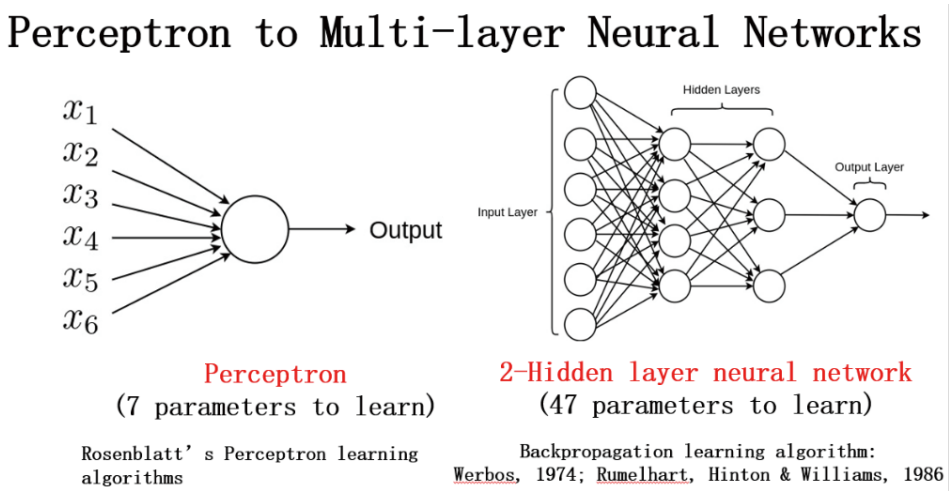

图 14：感知机与多层神经网络的区别

2014 年，Anil K. Jain 团队提出了一些新的端到端方法，基于学习的特征来进行预测和分析。下图 15 中的上半部首先使用手工提取特征，接下来他们用深度网络进行特征学习，然后进行结果的预测，效果非常受欢迎。

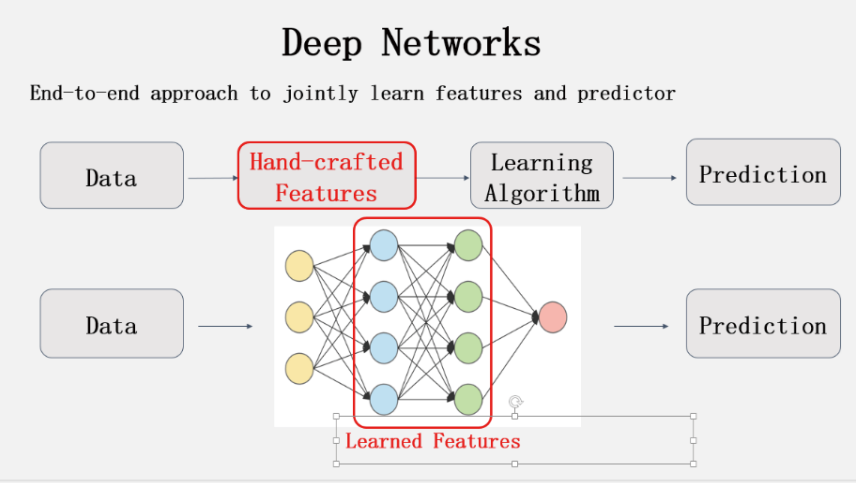

图 15：深度网络

Anil K. Jain 认为深度网络受欢迎的原因主要有以下几点：(1) 有大规模的标准数据，如 Image-Net；(2) 计算速度更快，如 GPU 大幅提高了 CPU 的运算性能；(3) 深度网络提供了更高的精确度，更有效应对各种各样的调整。

三、深度网络的主要挑战

Anil K. Jain 举了一些有关面部识别的例子，他介绍过去几年如果你想要进入美国境内或者从美国出境，比如在底特律，你需要站在摄像头前面照相，航空公司就会知道谁将会来到这里乘坐飞机，然后就会把他们自己的数据和这个面部识别进行比对，确保你是那个要坐飞机的人。还有美国的一些销售机器，也采用了面部识别的技术，大家可以通过训练集来确认你是否是被感兴趣的。如下图 16 所示。

图 16：人脸识别算法应用实例

2014 年 Facebook 提出了基于深度学习的人脸识别系统 DeepFace[12]，这一技术时深度学习人脸识别的开山之作，其精确度已达到 97.25% 的准确度。算法框架如下图 17:

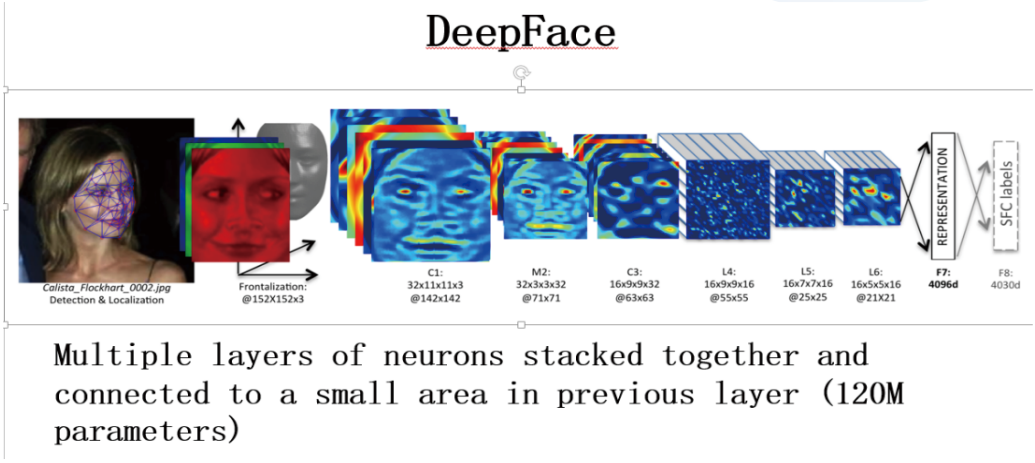

图 17： Deepface 算法框图

Anil K. Jain 认为深度网络方面目前还存在着许多挑战，主要包括如下几个方面。

3.1 类内差异性

首先是类内的差异性，目前容忍度还是比较低。如图 18 我们可以在 2009 年做的一个面部识别，正确的接受率达到了 99.2%，但 2018 年美国发布的国家技术标准，正确接收率只达到了 4.86%，所以这里还是存在一个比较大的挑战，包括我们需要建设一个更大的数据库，比如可能达到 1000 万或者是更多的图片等。

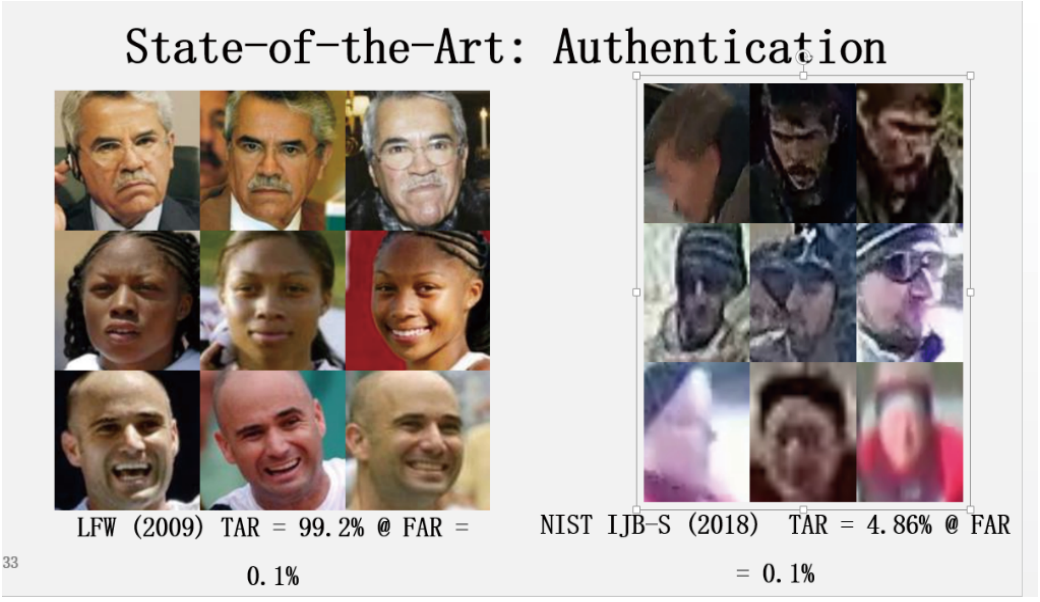

图 18：最新技术——认证

3.2 可解释性

Anil K. Jain 认为，深度网络的一个局限性就是可解释性——怎么去解释它。很多人已经在此方面做了很多工作，希望能够去说明高质量的图片。比如从面部深度特征当中重构潜在的外貌，识别出面部的一些特点，但如

果这个图片的质量很差的话，这个时候机器模型就很难去对这个面部进行很好的了解，也没有办法去很好地识别这个面部，因为我们不知道如何信任输出项。

3.3 公平性（人群偏见）

比如亚马逊等公司提出的一份报告，针对于 100 个不同肤色不同性格的人脸数据进行识别评估，发现不同种族和性别之间的精确度最多相差了 1%。如图 19 所示。

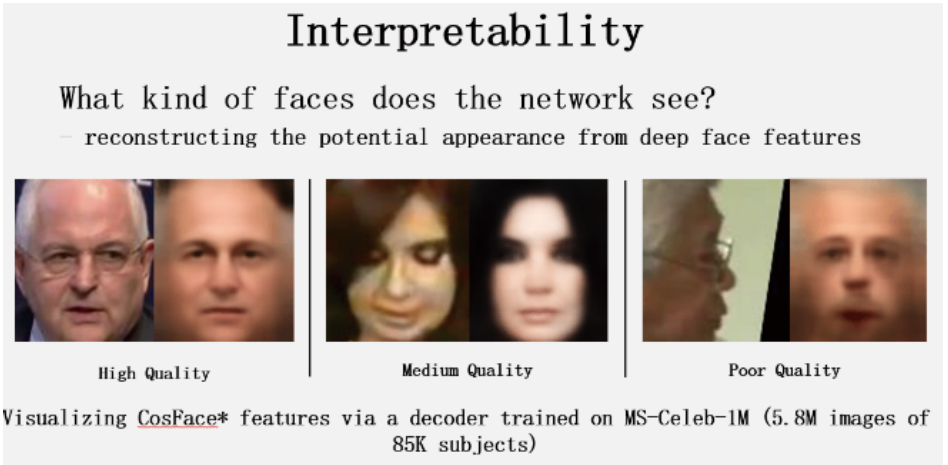

图 19：可解释性部分例子——人群偏见

3.4 鲁棒性

比如给定一张数字图片，我们有这个数字图片的一个模样，也会有一个测试集，看一看它们之间的一个匹配度，这个匹配在这个例子当中是比较高的，我们可以做一个对抗训练[13]，其中的一个照片上改变几个小像素，图片跟原照片看起来是差不多的，但是在这个数据库中图像无法进行匹配。说明此深度网络还没有足够的鲁棒性和稳健性。如下图 20 所示:

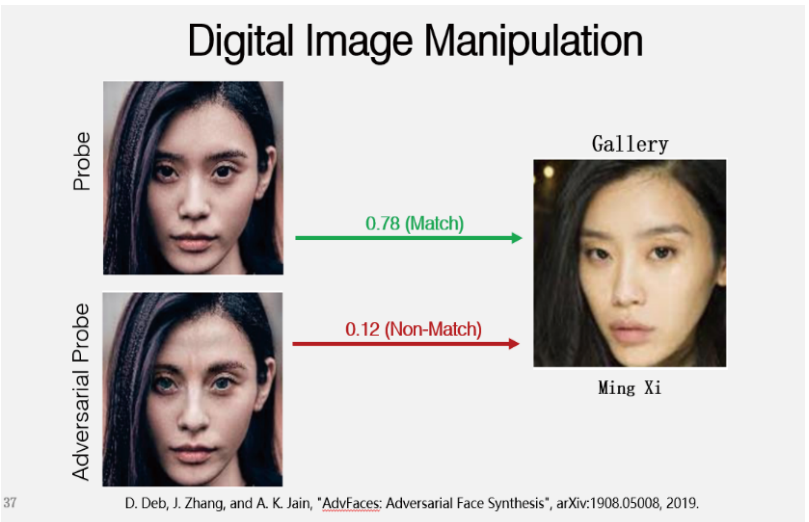

图 20：数字图像处理示例

3.5 安全与隐私

Anil K. Jain 指出，人们对隐私的定义是不一样的，安全和隐私之间有一个权衡和平衡的。人脸识别的体系当中，我们需要保证社会是安全的，我们希望能够去识别有关恐怖主义的活动或者其它的犯罪活动，比如有人抢劫商店等等，但是收集到数据库将来会做什么？永久保存还是什么？这里涉及到一个隐私问题。

四、人工智能的下一个十年

Anil K. Jain 最后展望了人工智能的下一个十年，认为我们应该关注以下几个方面：

1) 访问标记数据：利用合成和未标记数据
2) 领域知识：自上而下和自下而上相结合
3) 网络容量：它可以分离多少个模式类?
4) 对抗性攻击：脆弱到强大的网络
5) 可解释性：网络是如何做出决定的?
6) 用户隐私：保护用户隐私
7) 全球公益：设计人工智能改善极端贫困人口的生活 (约 10 亿)

参考文献：

[1] A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955, AI Magazine, Vol. 27(4), 2006

[2] Selfridge, "Pattern recognition and modern computers." In Proceedings of the Western Joint Computer Conf, pp. 91–93. March 1–3, 1955.

[3] https://www.washingtonpost.com/technology/2019/12/26/we-picked-most-influential-technologies-decade-it-isnt-all-bad/

[4] http://www.lgnewsroom.com/2019/09/lg-washing-machines-with-artificial-intelligence-and-direct-drive-motor-roll-out-region-wide/

[5] https://emerj.com/ai-adoption-timelines/self-driving-car-timeline-themselves-top-11-automakers/

[6] R.A. Fisher, The Use of Multiple Measurements in Taxonomic Problems, Annals of Eugenics, 1936

[7] F. Rosenblatt. The perceptron, a perceiving and recognizing automaton Project Para. Cornell Aeronautical Laboratory, 1957

[8] Yaser S AbuMostafa, Malik MagdonIsmail, HsuanTien Lin. Learning from Data: A Short Course[J]. AmIbook, 2012.

[9] T. W. Anderson, Classification into Multivariate Normal Distribution with Unequal Covariance Matrices. JASA, 1960

[10] Rosenblatt's Perceptron learning algorithms

[11] Backpropagation learning algorithm: Werbos, 1974; Rumelhart, Hinton & Williams, 1986

[12] Taigman, Yaniv, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. "Deepface: Closing the gap to human-level performance in face verification." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1701–1708. 2014.

[13] D. Deb, J. Zhang, and A. K. Jain, "AdvFaces: Adversarial Face Synthesis", arXiv:1908.05008, 2019.

剑桥大学教授 Zoubin Ghahramani：概率机器学习与人工智能

整理：智源社区 李维

Zoubin Ghahramani 本次的演讲主题是《Probabilistic Machine Learning and AI》。

Zoubin Ghahramani，剑桥大学信息工程系教授、剑桥大学 Alan Turing 研究所创始人之一、Uber 首席科学家、Uber 人工智能实验室联合创始人。他的研究方向包括：统计机器学习、贝叶斯非参数化、扩展推理和概率规划等，已发表相关研究论文 250 余篇。为表彰其在机器学习领域中的杰出贡献，于 2015 年被选为英国皇家学会院士。

在这场由 Zoubin Ghahramani 教授所带来的视听盛宴中，其介绍了机器学习和人工智能的基础与应用；分析了深度学习的特性、成功要素和它的局限性；强调了概率对机器学习和人工智能发展的重要性；回顾了其在概率人工智能研究中的一些前沿领域；此外，还谈到了人工智能和机器学习在 Uber 中扮演的重要角色等。

一、机器学习和人工智能的高光时刻

Zoubin Ghahramani 指出，人们常用的术语诸如人工智能 (Artificial Intelligence)、机器学习 (Machine Learning)、数据科学 (Data science)、数据分析 (Data analytics)、数据挖掘 (Data mining)、自适应控制 (Adaptive control) 等等并不是孤立存在而是彼此间互相联系的领域。正如图 1 所示，它们有着共同的理论基础，包括统计和机器学习。通常来讲，统计主要专注于拥有较少参数且有理论保障的简单模型。然而，机器学习则不然，其主要关注点则在那些有较多参数的复杂模型上。尽管如此，二者之间仍有密切联系。

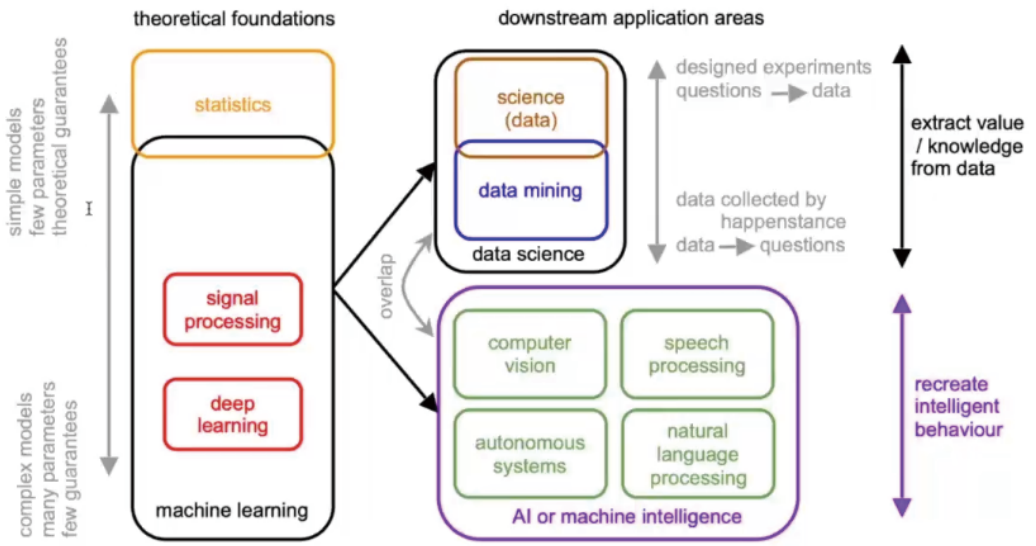

图 1：理论基础和主要应用

至于这些方法如何应用，则仁者见仁，智者见智，取决于你是想从数据中直接获取准确、有价值的信息，还是

重新创建某种智能行为。若是前者，则属于以数据挖掘为主的数据科学的范畴；若是后者，则属于以计算机视觉（Computer Vision）、自治系统（Autonomous Systems）和自然语言处理（Natural Language Processing）等为主的人工智能和机器智能的范围。不过，数据、模型、预测、决策等仍是这些方法和其所属领域共同的关键组成，是它们所绕不开的关键词。

图 2：AlphaGo 对战李世石
（https://www.alphagomovie.com/gallery）

之所以说目前是机器学习和人工智能的高光时刻，是因为它们已有很多成功的应用和突破，诸如乓（Pong）、打砖块（Breakout）、激光骑士（Beam Rider）、太空侵略者（Space Invaders）等游戏领域。我们所熟知的 AlphaGo 更是典范之一，其作为第一个击败人类职业围棋选手和战胜围棋世界冠军的人工智能机器人，是将人工智能推向普罗大众并将其置于聚光灯下的有力推手，一度成为人工智能的代名词。

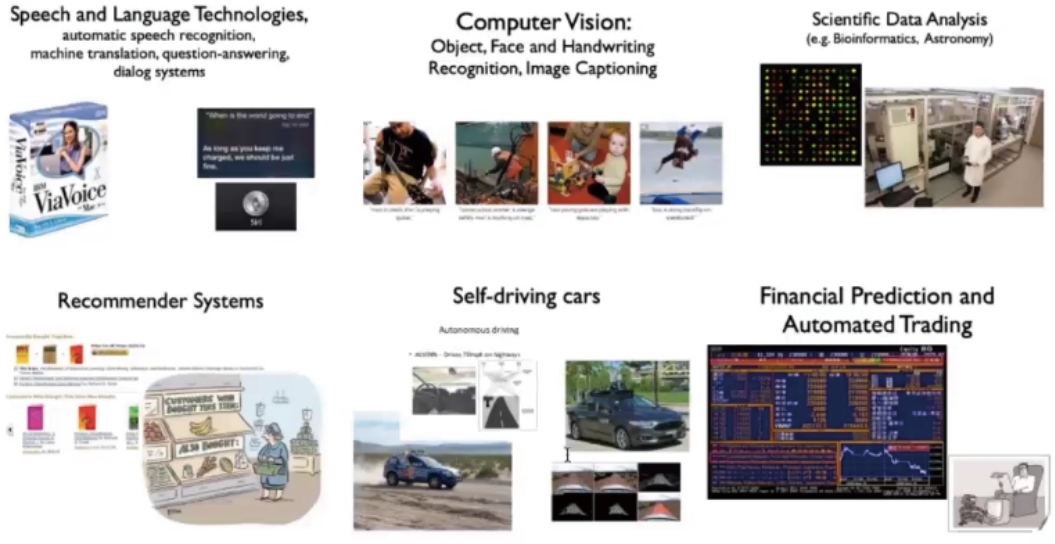

图 3：人工智能和机器学习的应用

抛开 AlphaGo 所带来的对人工智能和深度学习的思考与议论的浪潮来看，在解决现实生活中人们关心的问题方面，人工智能和机器学习也功勋卓著。比如语音识别技术、计算机视觉、科学数据分析、自动驾驶以及网购等实用性很强的关键技术中都有人工智能和机器学习的身影。最近在医学领域，应用人工智能和机器学习方法研究新冠肺炎病毒的传播甚至一度成为热门，且从结果来看也可谓卓有成效。可以说人工智能和机器学习早已“飞入寻常百姓家”，融入到人们生活的方方面面。

二、深度学习及贝叶斯规则

作为人工智能和机器学习革命性变革的幕后推手，你是否曾深入地想过究竟何为深度学习？深度学习对于我们以及同其息息相关的诸多领域来说又意味着什么呢？深度学习会成为未来十年我们思索的主题吗？

2.1 深度学习

Zoubin Ghahramani 认为深度学习是神经网络（Neural Networks）这一经典想法的重塑。至于神经网络，则主要由图 3 中所示的输入层、隐藏层和输出层三大部分构成，其中隐藏层可以有多个，而介于输入层和隐藏层、隐藏层和输出层之间的则是权值信息。从数学角度来看，神经网络的本质便是拥有多个参数的可调非线性函数，而多层神经网络在数学上则可表示为如图 4 所示的某些函数构成的多层组合函数。通常，这种多层组合函数可由多种随机梯度下降（Stochastic Gradient Descent）优化算法进行训练。

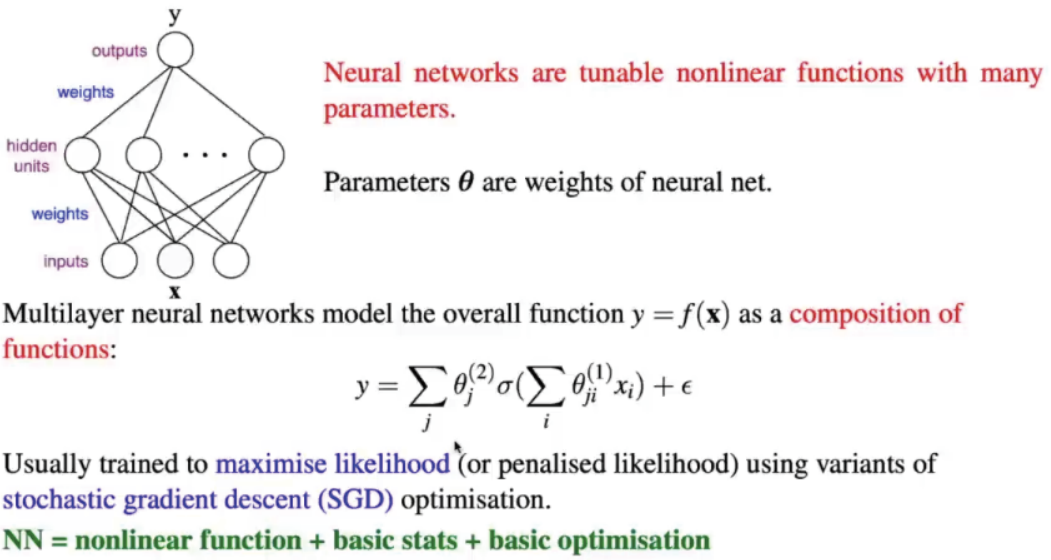

图 4：神经网络及其函数表示

Zoubin Ghahramani 指出深度学习系统事实上是一种类似于上世纪八九十年代流行的神经网络模型。不过相较后者而言，深度学习拥有以下得天独厚之处：

1）具有新的架构和创新算法；
2）庞大规模的网络数据集；
3）GPU 和云等海量计算资源；
4）PyTorch，TensorFlow 和 MxNet 等更好的软件工具；

5）高速增长的行业投资和媒体宣传。

除此之外，Zoubin Ghahramani 也认为以下技术方法的创新对促使深度学习繁荣发展功不可没，是其成功的关键因素：

1）自动微分法（Automatic Differentiation）；
2）线性整流函数（ReLU）、长短期记忆网络（LSTMs）、门控循环单元（GRUs）、残差网络（ResNets）；
3）随机优化，随机梯度下降法（SGD）；
4）更优的初始化；
5）卷积，递归网（Recursive Nets）；
6）大数据集。

尽管当前深度学习的发展也算是如日中天，但其也不是尽善尽美，Zoubin Ghahramani 认为其具有以下局限性：

1）过于数据饥渴（Data Hungry），依赖海量样本；
2）训练时所需计算量极大；
3）容易被对抗性样本（Adversarial Examples）误导；
4）对优化要求严苛，挑剔学习过程和初始化；
5）缺乏透明度，难以信赖；
6）不易将先验知识和符号表示相结合；
7）在不确定性表示方面黔驴技穷。

2.2 贝叶斯规则

作为《概率论》中的一个基本定理以及贝叶斯机器学习的基石，贝叶斯规则给出了事件 X 在事件 Y 发生条件下的概率与事件 Y 在事件 X 发生条件下的概率两者之间的确定关系。就图 5 所示公式而言，对一个概率模型进行试验，其实验结果可由一组 Data 表示，Hypothesis 则为导致实验结果的各种可能原因，P（Hypothesis）表示试验前预知的各种原因发生的可能性大小，故称为先验概率。当实验产生了结果 data 之后，结合贝叶斯公式，将获得对各种原因发生可能性的新认识 P（Hypothesis|data），称为后验概率。这一更新过程就是一种学习过程，而贝叶斯公式在其中所起作用就是指导我们如何从已知（Data）中窥探未知（hypothesis）。

$$P(hypothesis \mid data) = \frac{P(hypothesis)P(data \mid hypothesis)}{\sum_{h} P(h)P(data \mid h)}$$

当用概率来表达与我们模型相关的所有形式的不确定性和噪声时，贝叶斯规则就允许我们推断未知量，调整模型，做出预测并从数据中学习。虽然像贝叶斯机器学习这种基于概率的方法往往缺乏确定性结果，但将概率应用到人工智能领域这一思想仍具有很强的指导意义，这一点不难从概率模型已成为人工智能和机器学习领域的热门中看出来。尽管依据概率框架预测出的数据和对应做出的决策具有不确定性，但概率对人工智能发展的重要性仍能在以下几个方面得到体现：

1）校正模型与预测不确定性；
2）模型复杂度自动控制与结构学习；
3）建立做出合理决策的系统；
4）作为一种将先验知识构建到学习系统中的方法，并确保在获得更多数据时知识更新具有一致性和较强鲁棒性；
5）确保学习算法在大、小数据集上皆能奏效。

三、当前和未来的方向

在谈论现在和未来的一些研究方向的时候，Zoubin Ghahramani 围绕自动机器学习做了介绍。他首先提到了贝叶斯深度学习（Bayesian Deep Learning）以及深度和积网络（Deep Sum–Product Networks），并简要列举了贝叶斯深度学习的实现方式，如拉普拉斯近似（Laplace Approximation）、变分近似（Variational approximation）以及深层核学习（Deep Kernel Learning）等。在谈及深度和积网络时，他则条陈了它的一些关键特性，诸如：

1）可以用作生成模型或分布模型；
2）预测结果可与深度神经网络相媲美；
3）更好的校准不确定度；
4）评估似然性的能力；
5）有效的边缘化和条件反射；
6）处理缺失输入并检测异常值。

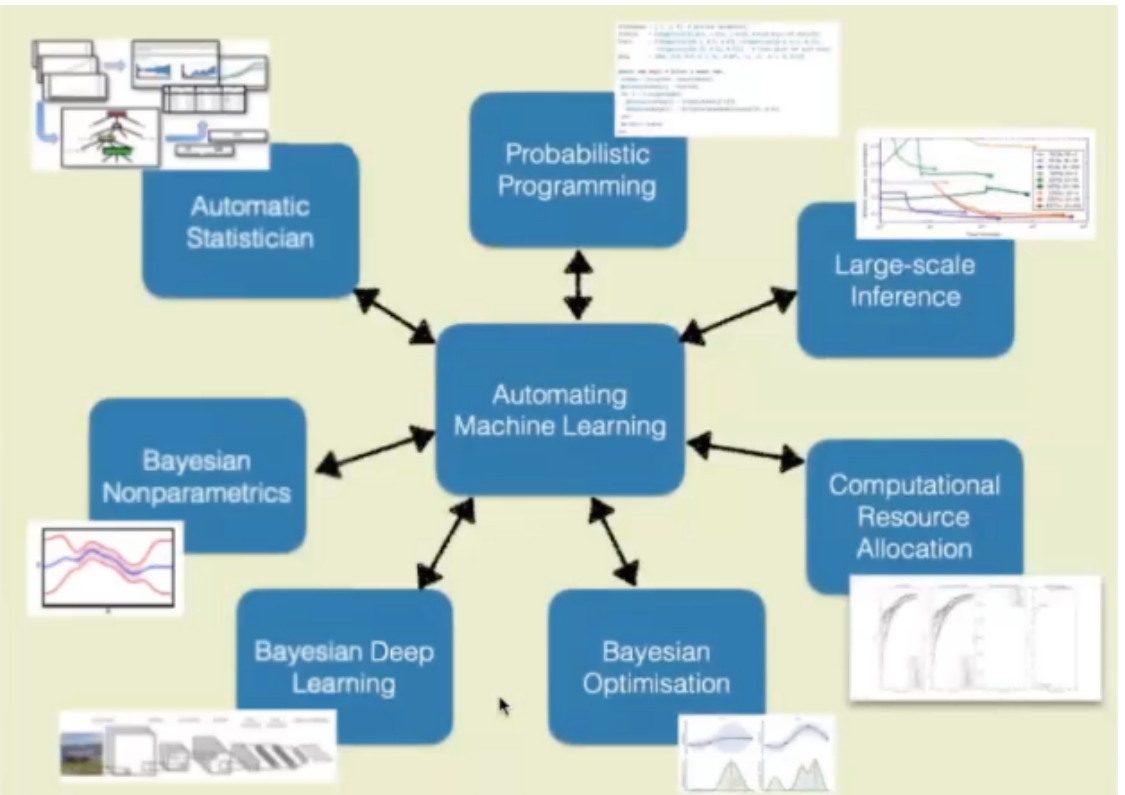

图 5：自动机器学习及其相关

除上述之外，Zoubin Ghahramani 还介绍了以下几个方面：

3.1 自动推理：概率编程

就概率编程（Probabilistic Programming）而言，Zoubin Ghahramani 指出了发展概率模型和解决推理算法推导耗时、易错问题的两种基本解决方式：其一是发展概率编程语言用以将概率模型表示为生成数据的计算机程序，例如 Edward，Pyro，STAN 以及 Turing 等语言；其二是为这些语言开发具有普遍意义的推理引擎，对给定的观测数据进行程序跟踪推断，例如 MCMC 采样（Particle MCMC）、变分推断（Variational inference）、序列蒙特卡洛（Sequential Monte Carlo）等。

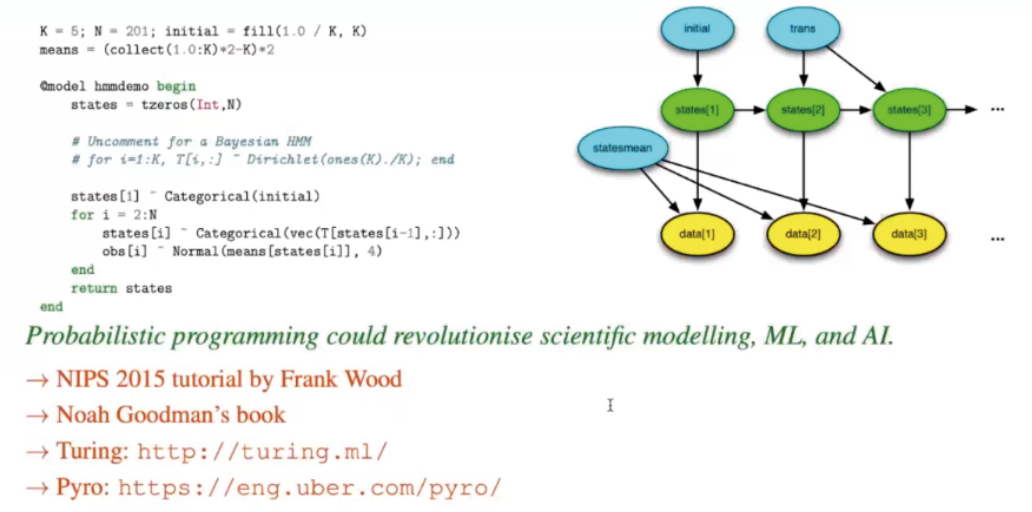

图 6：概率编程

3.2 自动优化：贝叶斯优化

贝叶斯优化（Bayesian Optimization）得名于其优化过程中使用的著名“贝叶斯定理”，这是一种十分有效的全局优化算法，具有令人瞩目的发展前景。黑箱函数（Black-Box Functions）全局最优化求解计算代价高昂，试图解决这一问题的贝叶斯优化的思想是将函数看作序列决策和模型不确定性问题，且这一思想已在机器人、药物设计以及神经网络的超参数选择等领域都有着广泛的应用。此外，贝叶斯优化方法作为一种基于模型的序贯优化，其在一次评估之后才进行下一次评估，能够在很少的评估代价下得到一个近似最优解，这使得该方法在业界赞誉颇高。

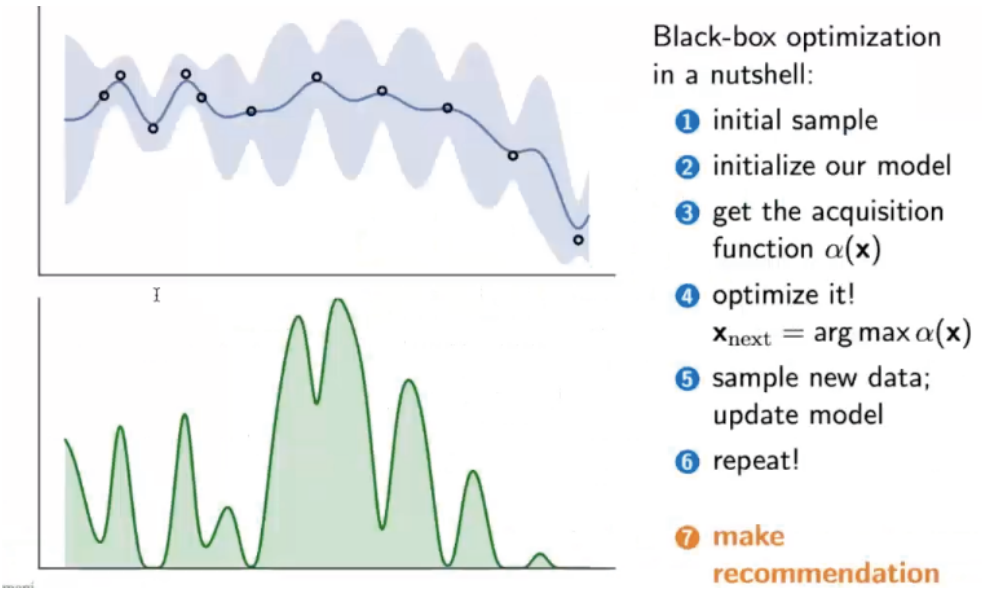

图 7：贝叶斯优化

之所以称贝叶斯优化是概率机器学习和人工智能领域中几种最先进、最有希望的技术之一，是因为它是一个思考任何优化问题的很好的框架，并且在以下情况中起着举足轻重的作用:

1）评估函数成本较高时；
2）导数难以评估甚至无法评估时；
3）函数评估中存在噪音时；
4）存在噪音约束时；
5）有关于函数的先验信息时；
6）需要优化多个相仿的函数时。

3.3 自动统计：数据科学的人工智能

数据无处不在，了解这些数据、建立模型并作出预测具有很大的价值，然而却匮乏能处理如此之多数据的数据科学家、统计学家和机器学习专家。如此以来，开发一种自动从数据中发现模型的系统或是解决之道，这种系统最好能具有处理数据、搜索模型、发现好的模型以及向用户解释发现了什么的能力，这也将大大解决人手不足问题。

Zoubin Ghahramani 认为自动统计应具备一些基本要素:

1）开放式的模型语言，其表现力足以捕捉真实世界的现象以及人类统计学家所使用的一些技术；
2）一个搜索程序，用以有效搜索语言模型；
3）一种评价模型的原理方法，用以权衡复杂性和适应数据。
4）自动解释模型的过程，使模型的假设能以非专业人士皆可理解的方式呈现出来。

四、关于 Uber

作为 Uber 的首席科学家，Zoubin Ghahramani 接下来介绍了 Uber 的相关背景情况。Uber，中文译作“优步”，是一家位于美国硅谷的科技公司，其旗下拥有风靡世界的同名打车 APP。除此之外，Uber 还有拥有 Rides，Uber Eats，Jump bikes，scooters，Uber Air 以及 Freight 等众多应用。作为打车应用的鼻祖，Uber 目前已覆盖超过 60 多个国家和 700 座城市，月活跃用户约 9300 万，每天提供近 1700 万次出行服务。

Uber 的科学家团队，除了 Zoubin Ghahramani 之外，还拥有一众出色的数据科学家、经济学家、计算机科学家以及大批的人工智能、机器学习和自动化研究员。此外，Uber 还拥有用于分布式深度学习的开源代码库 Horovod 以及用于深度概率编程的开源代码库 Pyro。后者基于 Python 与 PyTorch，专注于变分推理，也支持可组合推理算法，具有灵活、通用、可扩展的特点，能够实现灵活且富有表现力的深度概率建模，将现代深度学习和贝叶斯建模的优点相结合。

图 8：Horovod 和 Pyro 的标志

Zoubin Ghahramani 援引 Uber CEO Dara Khosrowshahi 的话说“Uber 本身就是一个巨大且颇具挑战性的机器智能问题，因其正试图优化现实世界以及它所带来的不确定性。”故其最后强调：只要我们想优化现实世界人和物的流动，解决实际存在的问题以及同这个拥有大量人群和复杂经济行为的网络互动，那么人工智能对 Uber 来说就依然具有特殊的重要性。

五、总结

在本场报告中，Zoubin Ghahramani 以对机器学习和人工智能理论基础与主要应用的介绍为开篇，详细地条陈了深度学习的特性、得以繁荣发展的要素、以及它的局限性，并且简要回顾了其在概率人工智能研究中的一些前沿领域，包括贝叶斯深度学习、概率编程和自动化等。从本场报告“概率”一词出现的频率以及报告标题《概率机器学习与人工智能》中，不难发现概率对人工智能未来发展的重要意义。正如 Zoubin Ghahramani 所言“概率建模为构建人工智能系统提供了一个框架，可以演绎推理不确定性并从数据中学习，它与决策理论相结合，形成理性决策系统的基础。”

冯诺伊曼奖得主 Jorge Nocedal：增强学习中零阶优化方法及其应用

整理：智源社区 钱小鹅

Jorge Nocedal 本次演讲的主题为《Zero-Order Optimization Methods with Applications to Reinforcement Learning》(增强学习中零阶优化方法及其应用)。

Jorge Nocedal，美国西北大学教授，曾在非线性优化、应用数学和运筹学等领域获得无数奖项。2009 年获查尔斯 – 布罗伊登奖；2010 年，他还被评为美国工业和应用数学学会院士；2012 年获乔治 –B– 丹齐格奖；2017 年，被授予冯・诺依曼理论奖。2020 年当选美国工程院院士。Nocedal 主要的研究方向为确定性和随机性设置中的非线性优化，他目前进行的算法和理论研究的动机源于图像和语音识别，推荐系统和搜索引擎中的非线性优化问题。

Nocedal 在演讲中指出，在函数优化的过程中，我们通常可以使用梯度下降的方法来获得目标函数的最值，但其实这需要依赖许多最值搜索的“运气”，其中包括：良好的初始化、步长，迭代方向计算的精度，搜索空间的结构等，但是在深度学习中，尤其是增强学习中，这些“运气”并不一定可以满足，那么我们如何通过其他的方法来做深度学习中目标函数的优化呢，本次演讲中 Nocedal 给我们分享了他的独到思路——零阶优化。所谓零阶优化算法即不利用一阶导数信息，在一定次数的抽样基础上，拟合目标函数的最值。零阶优化方法通过对目标函数逼近或对目标函数加罚函数的方法，将约束的优化问题转换为非约束的优化问题。

一、函数优化与深度学习

深度神经网络主要基于两个核心思想：其一是适合生成表示的预测函数结构，其二是在合适的空间中帮助寻找到合适的预测函数的反向传播算法。这里，反向传播算法通常意味着两件事：1）可以进行链式微分；2）可以使用梯度下降的方法进行优化。然而，优化过程中使用梯度下降法并不一定保证获得的解能够收敛到我们所期望的最小值。如下图所示，我们发现在优化过程中使用梯度下降方法，如果想要获得我们期望的结果，其依赖的条件有很多，比如：初始值、迭代步长的选择、迭代方向的计算等。

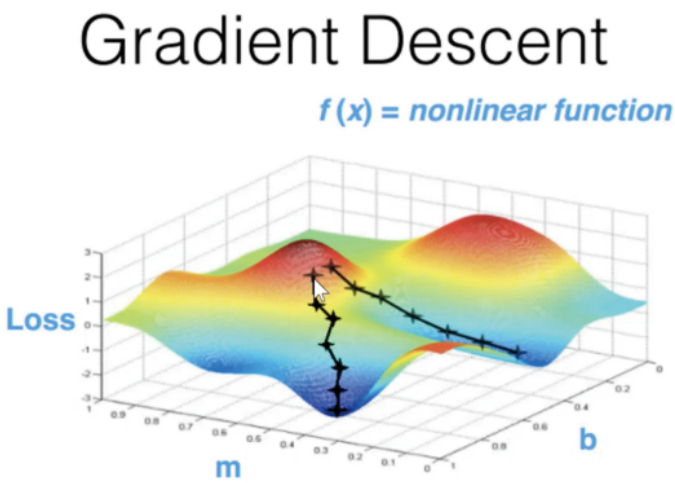

图 1：从不同初始值开始，非凸优化将得到不同的优化结果

上述我们提到的对函数优化的疑问，不少知名的数学界学者同样也对此表示怀疑，例如:

- Minsky 1961

  I doubt that in any one simple mechanism, e.g., hill–climbing, will we find the means to build an efficient and general problem–solving machine. (我怀疑，在任何一个简单的机制中，例如爬山，我们是否能找到建立一个高效和通用的问题解决机制的方法。)

- Minsky and Papert 1998

  If we can detect relative improvement, then "hill–climbing" may be feasible, but its use requires some structural knowledge of the search space. And unless this structure meets certain conditions, hill–climbing may do more harm than good. (如果我们能够检测到相对的改进，那么"爬山"可能是可行的，但是它的使用需要一些搜索空间的结构知识。除非这种结构满足某些条件，否则爬山弊大于利。)

事实上，在不同的搜索空间结构情况下，梯度下降法获得的效果不尽相同，甚至有时弊大于利。但对于"幸运"的深度学习来讲，我们经常遇到的是凸优化问题，因而梯度下降法取得了良好的结果。但对于强化学习，我们通常遇到很多非凸函数，并且由于网络很深，所以我们无法判断有多少个非凸函数。

优化问题在深度学习中的作用远不止求解最终结果这么简单，它在网络架构的设计中同样起到不容忽视的作用。例如我们熟知的残差网络，其设计的初衷即为了简化优化，换句话说即回答了为什么识别函数难以训练。Nocedal 表示探索这个问题的动机包含了如下三方面：1）计算噪声；2）深层神经网络的对抗训练；3）解决增强学习以及深度神经网络的优化问题。

那么如何解决这些情况下的优化问题呢？具体来说，假设我们希望最小化一个非线性函数，这个函数需要是光滑的（但并不需要是凸的），我们可以获得函数的估值但不知道它的梯度，同时，函数估值包含了噪声，那么，对一个有着上千个变量的这样的函数来说，是否存在一种算法能够很好的处理这类函数的优化问题?

目前来说，这仍是一个十分前沿的问题，我们还不能获得"最好算法"的确切答案。但 Jorge Nocedal 教授提出了解决这一问题的一种思路。这一思路将尝试计算梯度和噪声的近似值，并通过对噪声的拟牛顿法更新建立二次模型。Nocedal 给我们列举了一个黑箱的例子：假设有一个光滑的函数 (x)，但我们无法直接观察 (x) 而只能观察到包括了 (x) 和噪声的 f(x)，那么我们希望在只观察 f(x) 的情况下最小化 (x) 并计算出梯度近似 g。

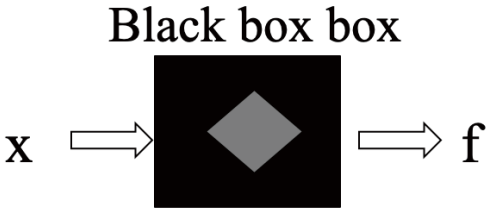

图 2：黑箱子实例

在进行噪声计算时，我们需要考虑一些不同的场景。其中一个场景是，我们需要用自适应的方法或是迭代线性求解，而另一个场景是，我们需要考虑包括舍入误差在内的随机误差。我们希望我们的方法可以适用于这些不同的场景。同时，对深度神经网络的对抗训练，我们可以观察深度神经网络的输入和输出，但不进行反向传播计算，而是对灵敏度进行分析。如下图 3 所示，这里假设是一个图像分类问题，那么，我们可以通过改变图像的一些部分获得完全不同的分类结果。我们主要对这些变化的部分进行分析，而不需要知道模型的导数，或是神经网络的各种公式。

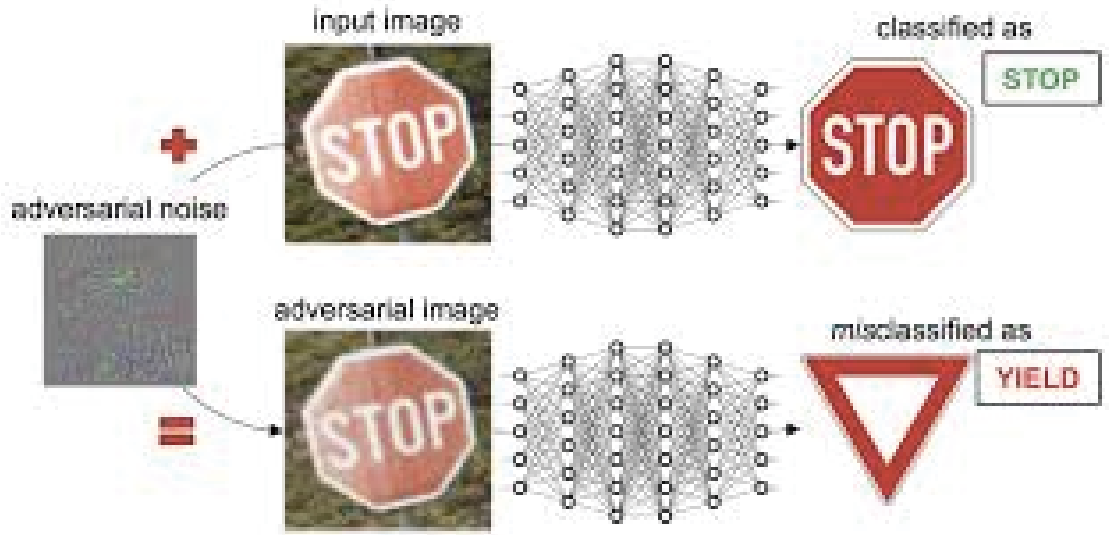

图 3: 灵敏度分析

二、零阶优化方法综述

所谓零阶优化方法，其本质为不计算目标函数的导数来计算目标函数的最值问题。在过去的二十年里，数学研究者已经设计了大量的无导数优化方法，最著名的包括直接搜索法和函数信赖域插值法。早期的方法包括：Nelder–Mead 方法、模拟退火和遗传算法。而相比于直接搜索法而言，在噪声存在的情况下，函数信赖域插值法比其他无导数优化技术鲁棒性更强。More’和 Wild 同时在其发表的文章中论述到 [2]，直接搜索方法速度慢，不能很好地适应问题的维数； 且函数插值法在最小化噪声函数方面更有效。但 Nocedal 指出，这些方法都是对离散的函数起作用，对连续的函数而言这些方法并没有很好的扩展性。2010 年，More’和 Wild 结合多年的研究经验，在文章中 [2] 大胆提到:

careful study dispels many myths about such methods. They found that the best method was one learn from the function values observed and creates a model of the objective. (仔细研究消除了许多关于这些方法的神话。他们发现最好的方法是从观察到的函数值中学习并建立目标的模型。)

如下图所示:

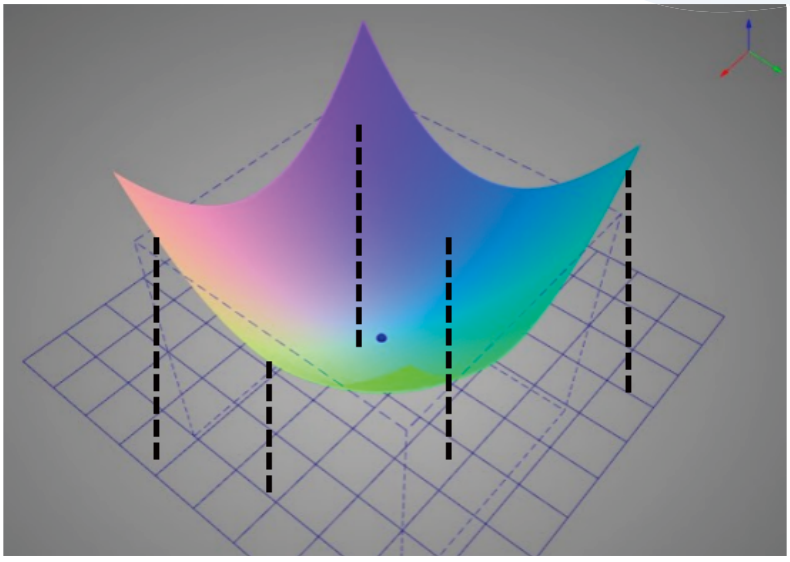

图 4：利用多点真实值构造二次插值函数

我们不妨假设我们可以获取到原始函数的五个真实值，那么我们可以根据这五个真实值来建立二次插值函数，并在信赖域中（信赖域一般不建议设置的过大）求该函数的最小值，函数形式为：

$$x_{k+1}: \quad \min \quad m(x)=x^T Bx+g^T x \quad \text{s.t.} \quad \|x\|_2 \le \Delta$$

由上式我们不难看出，如果我们想要利用纯插值方法来构建二次函数模型，并且在构建的模型中防止引入太多的噪声而导致函数非凸，那么：

- 至少需要 (d+1)(d+2)/2 个函数值，来确保我们可以获得一个完整的 Hessian 矩阵；
- 假设最小值的范数由 Hessian 矩阵更改，那么可以使用 O(d) 个点；
- 运算成本高；
- 插值点倾向于位于子空间上。

Nocedal 在报告中提到，虽然上述构建二次目标函数的方法运算成本较高，但是由于构建方法简单直观，因此他一度认为这种构建方法是正确的。但是，随着研究的深入，他发现该方法也有自己比较突出的问题：

- 不需要特别努力来计算好的梯度估计，但我们需要特别注意整个二次模型的质量；
- 信赖域估计的依赖性较强，如果信赖域较小，那么步长需要设计的较小，如果信赖域较大，那么步长需要较大，但因此在信赖域中会更容易引起震荡；
- 不可并行化。

因此 Nocedal 及其合作者 Berahas, Byrd 在 2018 年的文章“Derivative-Free Optimization of Noisy Functions via Quasi-Newton Methods”中将该方法进行了改进，改进后的方法：

- 努力逼近梯度；

- 将模型的构造委托给拟牛顿法 (BFGS);
- 恢复哈明提出的想法。

那么噪声函数的导数是什么意思呢？ Nocedal 在本次讲座中为我们分享了两种方法:

方法一：高斯平滑（Gaussian Smoothing）

假设函数中带有噪声，如下个沿高斯方向的随机小位移，那么形成的新函数和原始的函数非常近似。接着计算平滑函数的导数，

$$f^{\delta}(x)=\mathbb{E}_u[\frac{f(x+\delta\mu)}{\delta}] \qquad u\sim N(0,I_a)\ \ \delta>0 \qquad \text{Smoothed function}$$

Define

$$\nabla f^{GS}(x)=\frac{1}{m}\sum_{j=1}^{m}\frac{f(x+\delta u^j)-f(x)}{\delta}u^j \qquad u^j=e^j \text{ for finite differences}$$

计算的公式非常像有限差分，那么我们如何精确的计算梯度呢？如下图所示，我们首先给出一个随机的初始值，接着从该初始值开始，沿着高斯方向逐步移动，用有限差分的近似值乘以我们找到的方向（当然，我们的方向也可以取高斯方向的平均值），这样就可以计算出最后的导数。这样的计算方法在一些机器学习中是非常有效的，但是对科学中的其他应用并不是很友好，所以我们还可以采用第二种方法。

Algorithm Zero-order Stochastic Gradient Method

Input: $x_0, \delta>0$

for $k=1,\ldots$ do

compute $g_k \leftarrow \nabla f(x_k)^{GS}$

$x_{k+1}=x_k-\alpha_k g_k$

方法二：带有噪声估计的有限差分法（Finite Differences with Noise Estimation）

在一些科学应用领域，有限差分法是更为经典的求解方法，当然，我们考虑的并非简单呐的差分，而是更近一步，将噪声估计包含在计算中。More'–Wild 在 2012 年发表的著作中提到:

if we can estimate the noise level, we can compute a good finite–difference interval h

如果我们能估算噪声，我们将可计算出有限差分中好的步长 h.

$$\frac{\partial f(x)}{\partial x_i}\approx\frac{f(x+he_i)-f(x)}{h}$$

由于噪声水平通常是随机噪声的标准差，因此一旦我们可以估算出噪声，那么前向差分 h 的表达式如下所示:

$$h=8^{1/4}(\frac{\epsilon_f}{\mu_2})^{1/2} \qquad \mu_2=\max_{x\in I}|f''(x)|$$

例如，如下图 5 所示，我们已知噪声的标准差为 0.025，x = 0.12; 那么由上述前向差分 h 的表达式我们可以计算出 h_correct = 0.28，如果选择的 h 更接近于正确的值，那么拟合的效果会更好，如果相差较远，那么拟合效果会大打折扣。

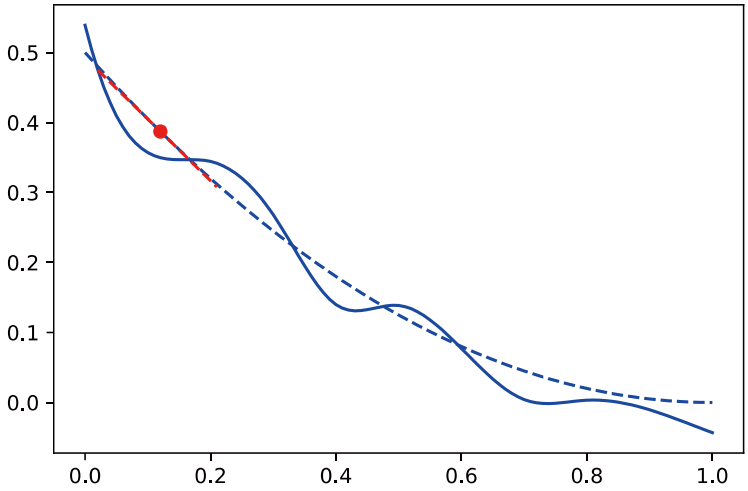

图 5: 不同的 h 对最终拟合结果的影响

那么问题来了，我们该如何估算出噪声，并且估算方法既适用于随机性，又适用于确定性呢？ Nocedal 在本次讲座中也为我们分享了他对函数噪声估算的想法。

Noise estimation (for deterministic or stochastic noise)

为了估算函数的噪声水平，也就是

$$\text{i.e.,} \qquad \sigma = [var(\epsilon(x))]^{1/2}$$

在 x 处，选择随机的方向 v, 估算 f 在同等空间的 q+1 个点的值 x+ibv:

Compute function differences:

$$\Delta^0 f(x) = f(x)$$

$$\Delta^{j+1} f(x) = \Delta^j[\Delta f(x)] = \Delta^j[f(x+\beta)] - \Delta^j[f(x)]]$$

$$\Delta^k \phi(x) = f^{(k)}(\xi)h^k$$

然后，我们根据 Hamming Difference Table，其中

$$\min f(x) = \sin(x) + \cos(x) + 10^{-3}\,\mathrm{U}(0, 2\sqrt{3}) \qquad q = 6 \qquad \beta = 10^{-2}$$

搜索在计算中需要的值，

x	f	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$	$\Delta^5 f$	$\Delta^6 f$
$-3\cdot 10^{-2}$	1.003	$7.54e-3$	$2.15e-3$	$1.87e-4$	$-5.87e-3$	$1.46e-2$	$-2.49e-2$
$-2\cdot 10^{-2}$	1.011	$9.69e-3$	$2.33e-3$	$-5.68e-3$	$8.73e-3$	$-1.03e-3$	
-10^{-2}	1.021	$1.20e-2$	$-3.35e-3$	$3.05e-3$	$-1.61e-3$		
0	1.033	$8.67e-3$	$-2.96e-3$	$1.44e-3$			
10^{-2}	1.041	$8.38e-3$	$1.14e-3$				
$2\cdot 10^{-2}$	1.050	$9.52e-3$					
$3\cdot 10^{-2}$	1.059						
σ_k		$6.65e-3$	$8.69e-4$	$7.39e-4$	$7.34e-4$	$7.97e-4$	$8.20e-4$

图 6：Hamming Difference Table

我们发现，光滑函数的高阶差分很快趋于零，而差分在噪声中是从零开始的。我们将其看作是一个可被观察的改变标志。同时，我们看到，整个过程是尺度不变的。继续看上图所示的汉明差分表，我们看到最后一行的数值是不同的。这是由于，我们的函数中带有随机的噪声（如果没有随机噪声，那么在同一点上采样的结果应该是相同的），因此我们使用这些带有噪声的值（或平均值）带入到数值算法中，这些随机的噪声将会对结果起作用。

那么有读者会产生疑问，一旦我们由上述提及的方法获取到了相对应的梯度，那么为什么我们不使用拟牛顿模型呢？如噪声有限差分 BFGS 模型？大家的想法非常正确，但实际上目前还没有人这么做。主要原因是对噪声函数的差分是十分危险的，不好的迭代可以造成灾难性后果。

在 Nocedal 的算法中，我们将对每次迭代的噪声进行估算，并根据估算结果计算有限差分的步长 h，在获取到对应的梯度后，我们就可以使用拟二阶牛顿法作为模型进行线性搜索，总结一下，Nocedal 的整体算法流程如下:

- 在每一次迭代的过程中估算噪声;
- 根据噪声估算有限差分的步长 h;
- 由有限差分公式计算梯度;
- 计算二次拟牛顿法的搜索方向;
- 进行线性搜索;
- 如果搜索值没有收敛，那么继续重复上述的步骤。

这里线搜索起到了两个作用： 其一是在有限差分区间合适的情况下帮助决定步长，其二是帮助决定是否需要重新估计噪声水平。

对于零阶优化的算法验证，Nocedal 及其合作者也非常有信心，在 Nocedal 近期发表的论文 [1] 中，我们看到如下图 7 的实验结果：在使用相同的初始值进行迭代的不同算法中，Nocedal 提出的算法在迭代 12 步之后，会更快的收敛，达到先验的噪声水平。

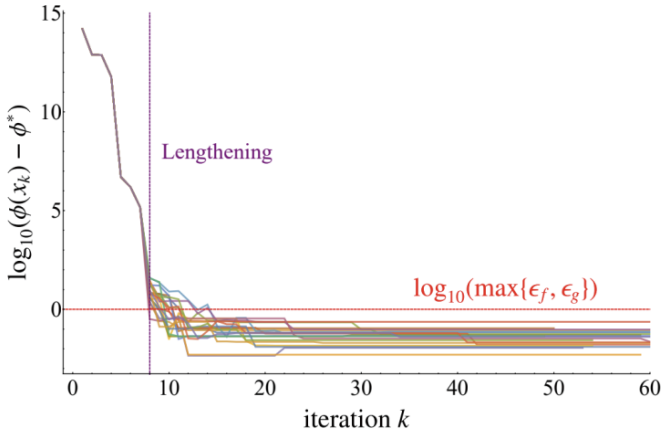

FIG. 4.1. *Results of 20 runs of Algorithm 2.1. The graph plots the log of the optimality gap for the true function,* $\log_{10}(\phi(x_k) - \phi^*)$*, against the iteration number* k*. The horizontal red dashed line corresponds to the noise level* $\log_{10} \max\{\epsilon_g, \epsilon_f\} = 0$*. The vertical purple dashed line marks the first iteration at which lengthening is performed* ($k = 8$).

图 7: 算法验证实验分析结果

三、结语

Nocedal 在本次讲座中为大家分享了深度学习中隐形的“核武器”--- 函数优化。不少做深度学习工作的读者都了解，在做许多问题中，例如： 目标检测和识别、人体跟踪、语音识别、广告推荐等，设计一个好的损失函数与设计一个好的网络几乎同样重要。损失函数，实际就是我们优化中的目标函数，而如何寻找损失函数的解，其本质就是优化问题。Nocedal 教授为我们讲解了噪声估计、步长的精确设置、梯度的计算以及搜索方向的计算，将优化问题庖丁解牛般拆分开来，可谓“功力十足”。同时他还提出了一些非常前沿的思想，例如： 优化方法也会潜在的影响网络的结构、我们是否可以将求解偏微分方程的过程与优化方法进行结合，从而设计求解偏微分方程的深度神经网络、如何将噪声的分析应用于深度学习的灵敏度分析等等，都非常值得大家深入思考。非凸优化的求解方法中仍然存在不少难点，感兴趣的读者可进一步阅读教授及其合作者的文章，进行深入研究和探索。

参考文献

[1] Yuchen Xie, Richard H. Byrd, and Jorge Nocedal. (2020) Analysis of the BFGS Method with Errors. SIAM Journal on Optimization 30:1, 182–209.

[2] Jorge J. More and Stefan M. Wild ´ , Benchmarking derivative–free optimization algorithms, SIAM Journal on Optimization, 20 (2009), pp. 172—191.

[3] Jorge J. More and Stefan M. Wild ´, Estimating Computational Noise, SIAM J. Sci. Comput., 33(3), 1292—1314. (23 pages)