
AI
人工
智能
指数
2017年度报告

指导委员会

Yoav Shoham（主席）
斯坦福大学(Stanford University)

Raymond Perrault
SRI International

Erik Brynjolfsson
麻省理工学院(MIT)

Jack Clark
Open AI

项目主管

Calvin LeGassick

「人工智能指数 2017 年度报告」

Yoav Shoham, Raymond Perrault, Erik Brynjolfsson, Jack Clark, Calvin LeGassick「人工智能指数 2017 年度报告」, 中文翻译今日头条、机器之心

由斯坦福大学人工智能百年研究(AI100)推出的「人工智能指数」(AI Index)是一个追踪人工智能行业动态与发展的非营利性项目，其研究覆盖了百年以来人工智能的总体情况，目标是基于数据来推动人工智能的广泛交流和有效对话。近日刚刚推出的 2017 版报告是人工智能指数的首届年度报告，它从多个角度观察和解读了人工智能领域的动态和进展。经「人工智能指数」项目委员会授权，今日头条联合机器之心对此报告做了中文翻译，译文错误由翻译方负责。

报告原文链接：https://aiindex.org/2017-report.pdf

报告引用格式：

Yoav Shoham, Raymond Perrault, Erik Brynjolfsson, Jack Clark, and Calvin LeGassick, "The AI Index 2017 Annual Report", AI Index Committee of the One Hundred Year Study on Artificial Intelligence (AI100), Stanford University, Stanford, CA, November 2017. Chinese translation by Bytedance & Synced.

目录

人工智能指数 2017 年度报告简介

当下，人工智能已然跃居为全球话题的焦点，来自开发者、业界领袖、政策制定者乃至大众的关注正与日俱增。一年以来，我们从新闻中看到了各种有关人工智能的言论与争辩，可以发现，这个领域正在被广泛地考察、研究和应用。然而，由于目前人工智能技术的发展异常迅速，即便是该领域的专家也很难理解与把握整个行业的面貌。

如果没有研究人工智能技术现状的相关数据，那我们只能在有关人工智能的讨论和决策中做出盲目的推论。

我们只能在有关人工智能的讨论和决策中做出盲目的推论。

由斯坦福大学发起并实施的「人工智能百年研究计划」(AI100)，是一个旨在追踪人工智能发展的公开非盈利项目。该项目力图让大家通过数据来了解人工智能，推动有效对话。这是人工智能指数的第一份年度报告，其从多个角度对人工智能的现状与进展展开了研究。我们首先整合了散落在网络上的各种原始数据，然后从中提取出了新的人工智能评价标准。

所有用于生成该报告的数据都将会在 人工智能指数网站上公开（地址：aiindex.org）。然而收集数据仅仅是第一步，人工智能指数需要来自更大范围的社区的帮助。从根本上来说，这篇报告号召大家参与进来。你们可以提供数据、分析已有数据，也可以提出意愿，表明自己希望追踪哪些数据。在这里，无论你是否可以提供任何答案或问题，我们都希望这份报告能让你有兴趣来了解人工智能指数，并且成为其中的一员，为奠定一个扎实的人工智能发展脉络共同努力。

报告概览

本报告前半部分展示了人工智能指数团队收集的系列数据，后半部分包含了对重要领域的讨论、各类专家的评论，以及我们的一项倡议——希望大家支持我们的数据收集工作、并加入到评估与交流人工智能技术进展的对话中来。

数据部分

本报告中的数据主要分为以下四个部分：

领域活力
技术表现
流行趋势关系研究
达到人类水平的性能？

「领域活力」这个指标被用来描绘人工智能领域中「量」的一面，如人工智能大会的参会情况、风险投资资本对人工智能创业公司的投资情况。「技术表现」这个指标力图抓住领域中「质」的一面，如计算机在理解图像和证明数学定理方面的能力。附录中提供了收集各个数据库的方法论。

前两部分数据向我们展示了领域现状：所有折线图中的线条都在朝右上方升高，这反映出人们在人工智能领域内从事的活动和相应技术的发展正在不断提升。在「流行趋势关系研究」这个部分，我们会考察当前流行趋势之间的关系。此外，我们还会引入一个探究测量方法，即「人工智能活力指数」(AI Vibrancy Index)，该指数能够结合学术界与产业界的流行趋势，对人工智能领域的活力进行量化。

在测试人工智能系统的性能时，我们会很自然地将其和人类在相应领域的表现进行对比。在第四部分「达到人类水平的性能」中，我们会概括那些和人类表现相比有显著进步或直接超越了人类表现的人工智能系统。我们也会讨论这种对比面临的困难，并适当进行说明。

讨论部分

在展示完所收集的数据以后，我们会对本报告强调的流行趋势以及报告未提及的重要领域进行讨论。

部分讨论关注了本篇报告的局限性。由于数据来源以美国为中心，使得报告并非完全客观，且因只追踪监测了定义清晰的标准，报告也可能高估了技术领域的发展。报告缺乏数据的人口统计学分类，没有提及由政府和公司联合投资的人工智能研发情况。这些方面都十分重要并将在以后的报告中有所涉及。我们在本报告的「查缺补漏」章节中会进一步讨论这些局限性以及其它的问题。

正如报告的局限性所示，人工智能指数并不能十分全面地描绘整个人工智能领域。因此，我们收集了来自交叉领域的人工智能专家的意见。专家们讨论并完善丰富了数据背后的故事，同时也补充解释了报告中缺漏的部分。

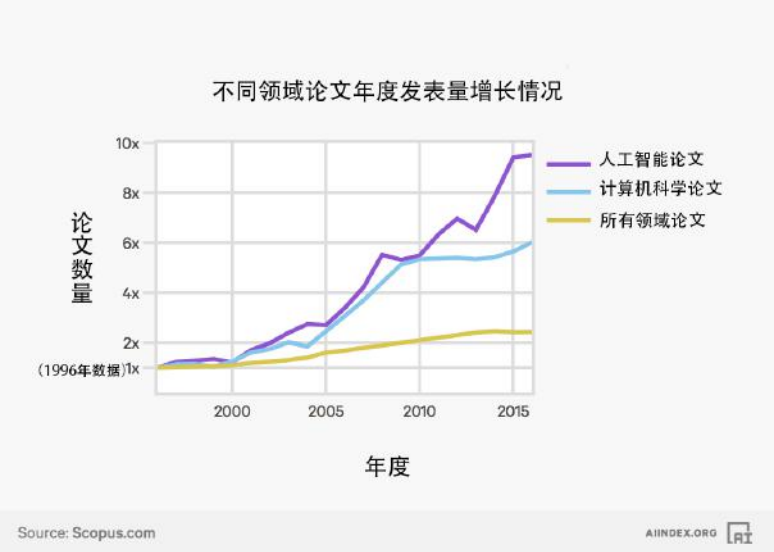
最后，在专家对话板块结束后，人工智能指数诚挚邀请各位读者参与本报告的更新与修正。我们需要来自更大范围社区的反馈和参与，来设法解决本报告中提出的问题及补充遗漏问题，从而在追踪人工智能活力与发展的道路上结出丰硕的果实。

领域活力
学术领域
论文发表数量

下图统计了 Scopus 学术论文库中标注关键词「人工智能」的计算科学论文数量。

自 1996 年至今，每年发布的 人工智能论文数量增加了 9 倍多。

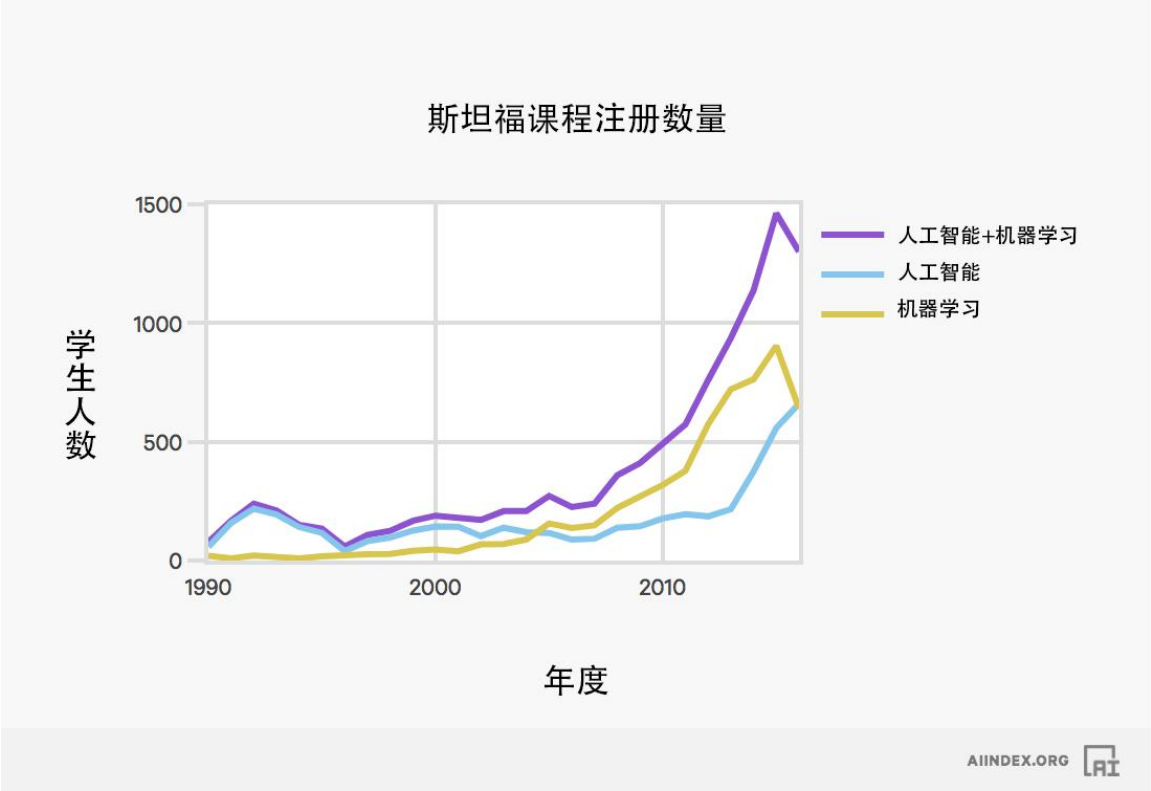
这里是各类学术论文年发表率与其 1996 年发表率的比较。图表显示了各领域论文、计算机科学领域论文以及计算机领域内人工智能论文年发表率的增速。



数据揭示了人工智能论文发表率的增长不仅仅是出于对更广泛计算机科学领域兴趣的增长。具体来说，尽管自 1996 年以来整体计算机科学领域内的论文数量已经增长了 6 倍，同时期人工智能领域每年发表的论文数量已经增长了 9 倍多。

课程选修人数

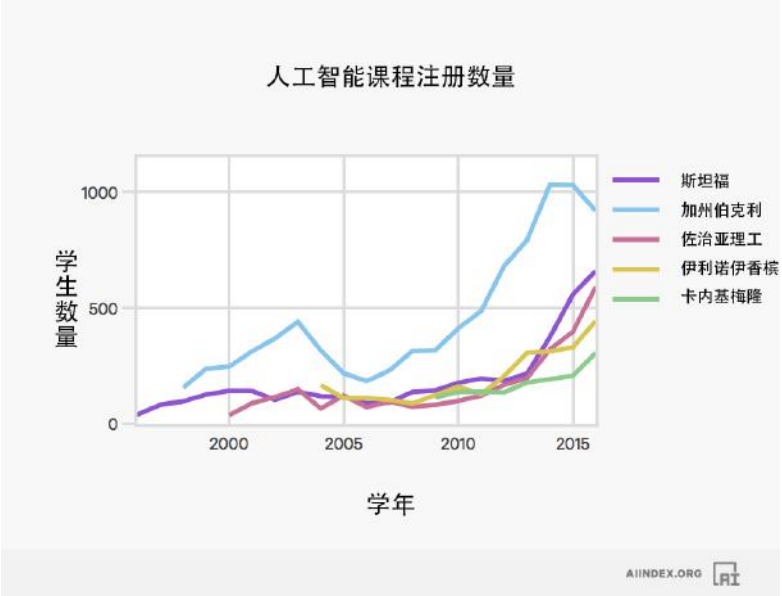
除了论文发表数以外，课程的参与人数也能体现这个领域的活力。以下展示的是斯坦福大学每年选修人工智能与机器学习导论课程的学生数量。
机器学习是人工智能的子领域。我们着重关注机器学习导论课程的参与度是因为目前人工智能领域很多成果都基于机器学习的算法与理论。



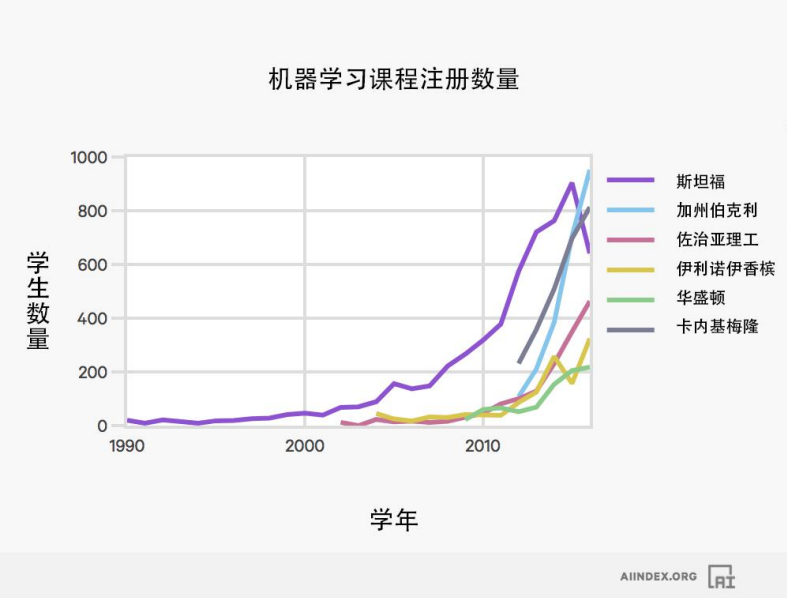
自 1996 年以来，选修斯坦福大学人工智能导论课程的人数已经增长了 11 倍。

注：斯坦福大学 2016 学年机器学习入学人数的下降是基于当年的行政问题而非学生兴趣。详情请见附录。

本报告之所以着重突出斯坦福大学导论课程的选修人数是因为其数据最全面。不过如下所示，其它高校导论课程的选修趋势也与斯坦福相似。



注：许多大学从上世纪 90 年代起开设人工智能课程。上图展示的是可获取数据的年份的情况。

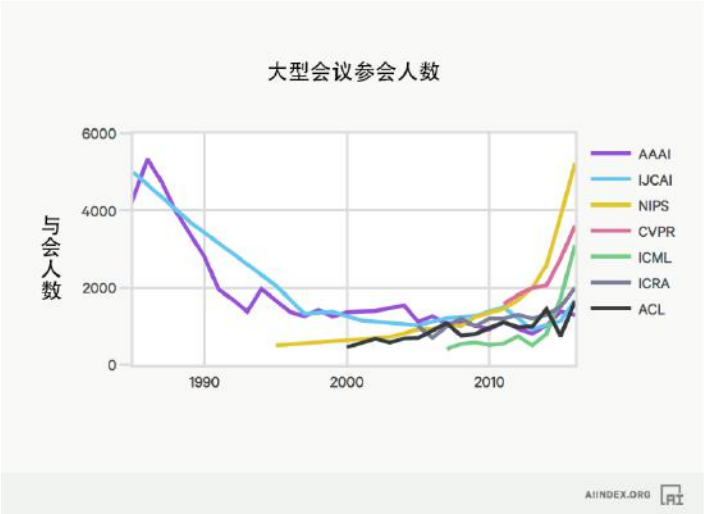


注：许多大学从上世纪 90 年代起开设机器学习课程。上图展示的是可获取数据的年份的情况。

需要注意的是，这些图表展示了高等教育领域中的一个侧面，这些数据并不一定代表学术机构总体的发展趋势。

学术会议出席情况

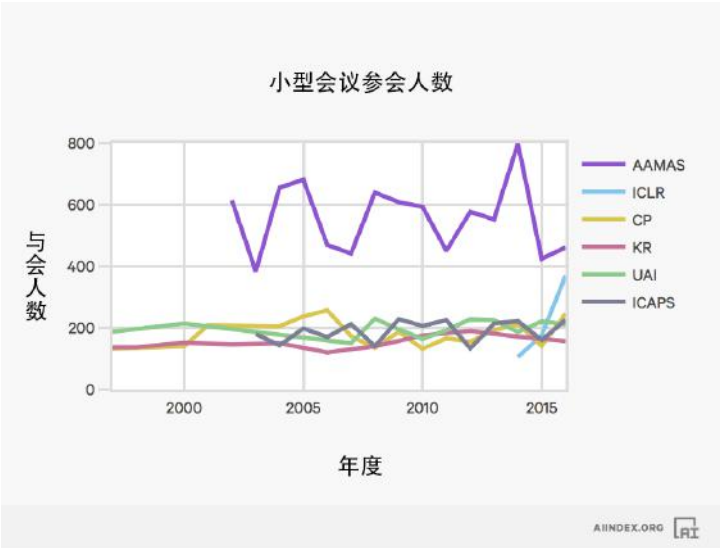
以下展示了人工智能领域有代表性的学术会议的参会情况，其中既有如 AAAI、IJCAI 和 ICML 这样的大型综合性会议（按 2016 年参会人数超过 1000 人为标准），也有像 CVPR、ACL、ICRA 那样专注于计算机视觉、自然语言处理和机器人的小型会议（2016 年参会人数不足 1000 人）。



注：大多数学术会议自 1980 年代起即开始举办，上图展示的是参会人数有记录的年份的情况。

研究重心转移：上图的参会人数同样表明了研究重点已经从符号推理转向了机器学习与深度学习。

下图展示了参会人数少于 1000 人的小型学术会议的参会情况，其中需要注意的是 ICLR，该会议专注于深度学习领域，第一次会议于 2013 年由深度学习先驱 Yann LeCun 及 Yoshua Bengio 主办。

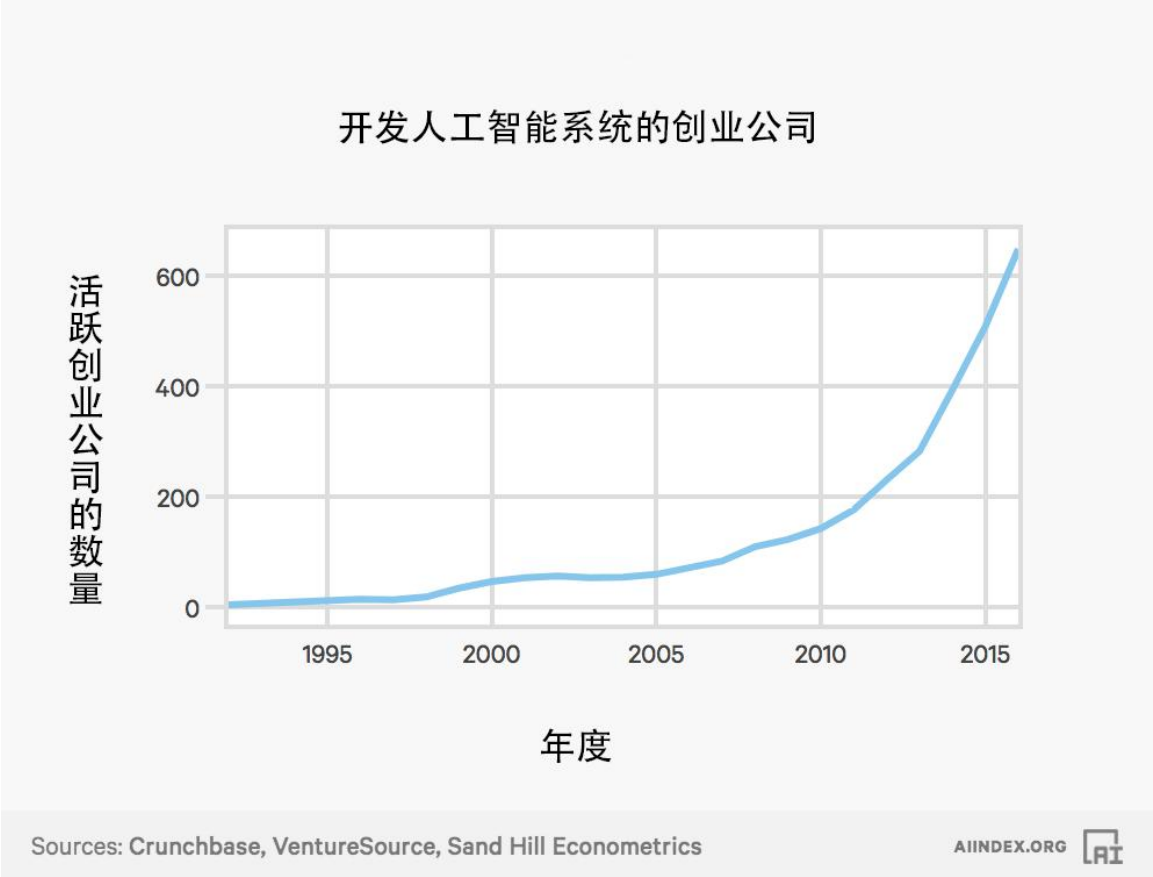


稳步前进：尽管学术界研究重点近年来已转移至机器学习及深度学习，仍有一小部分研究者继续在符号推理方法上进行探索并取得进展。

产业领域

AI 领域创业公司

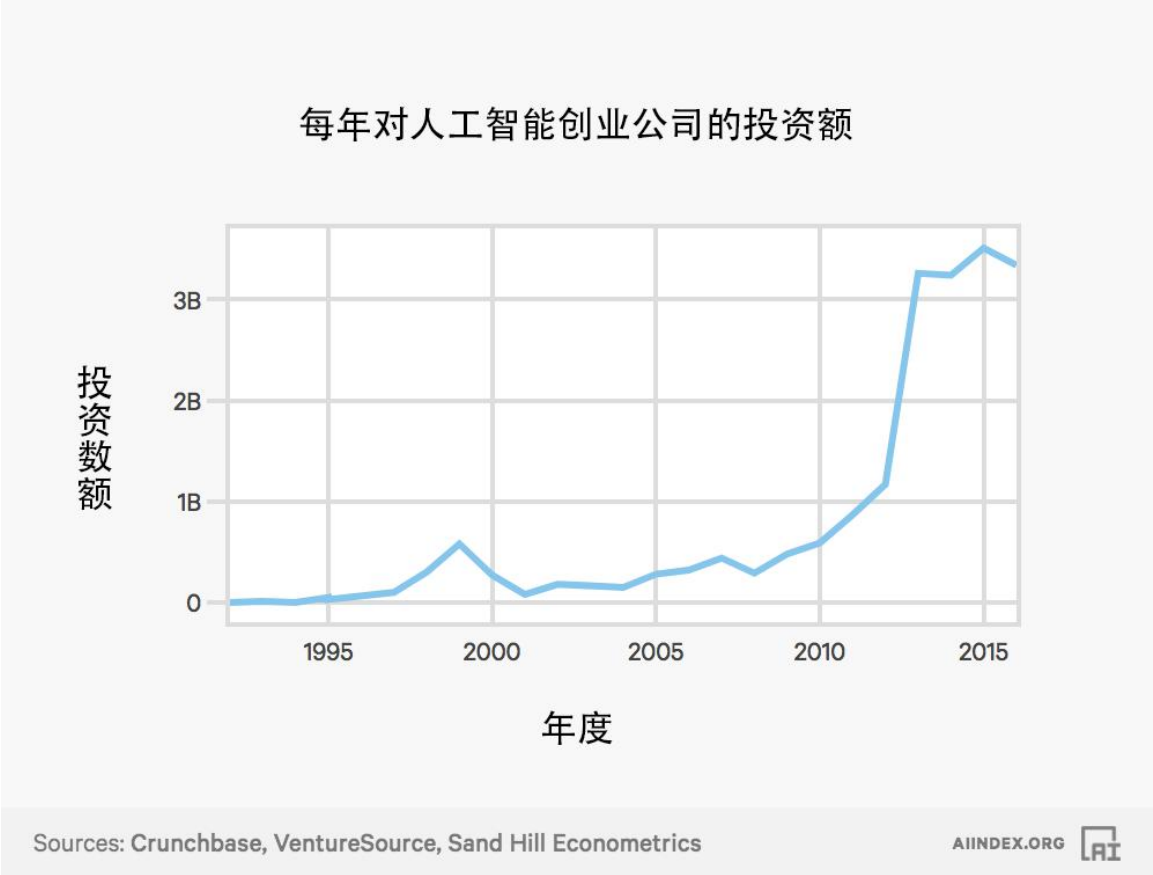
下图展示了得到风投资本支持并开发了人工智能系统的美国活跃创业公司的数量。



这一数量自 2000 年以来已增加了 14 倍。

AI 领域风险投资

下图为风投资本对美国人工智能创业公司所有融资阶段的年投资总额。



这一金额自 2000 年以来增加了 6 倍。

工作机会

下图分别展示了两个在线招聘网站 Indeed 和 Monster 上需要人工智能技能的工作数量的增长。我们通过标题和工作描述的关键词区分出需要人工智能技能的工作。

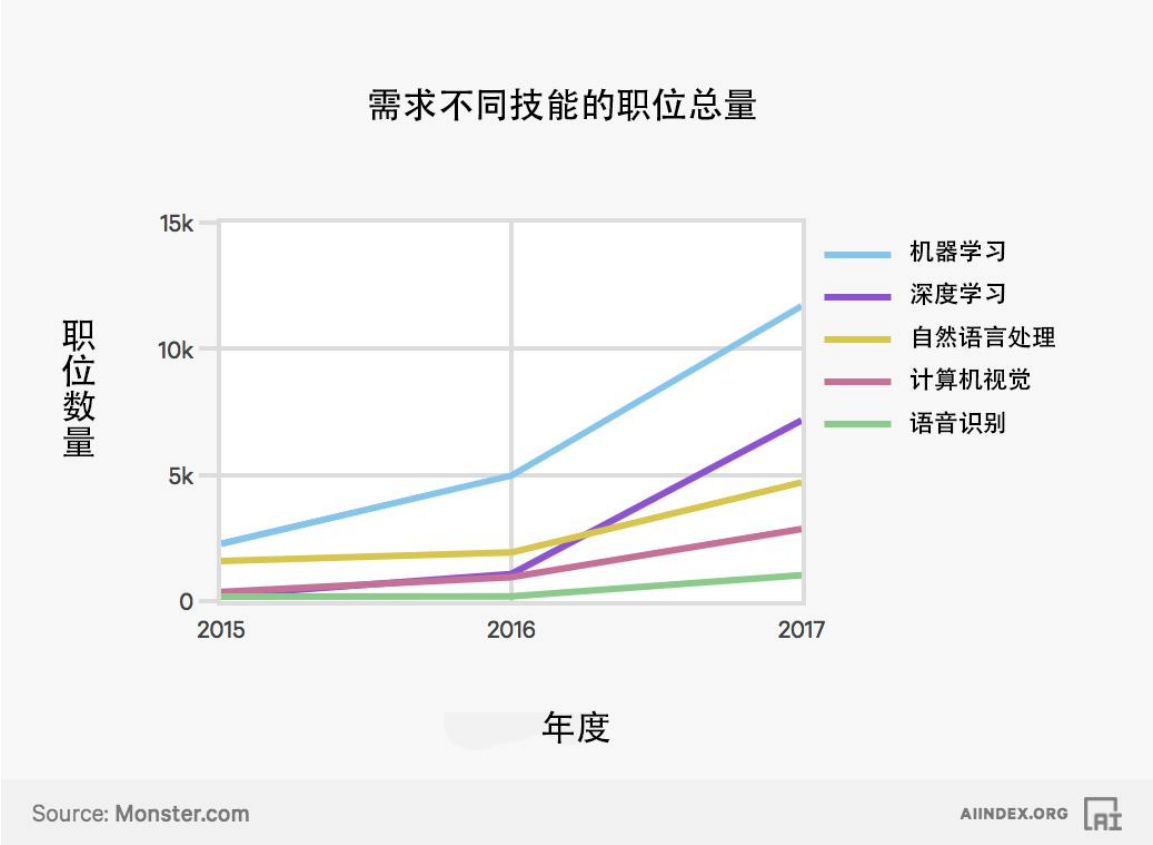
下图是 Indeed 网站上美国需要人工智能技能的工作数量的增长数据。涨幅是基于 2013 年 1 月 Indeed 网站上美国要求人工智能技能的就业岗位所占份额的增长倍数。

自 2013 年以来，在美国需要人工智能技能的工作比重增长了 4.5 倍。

下图为 Indeed.com 平台报告的多个国家需要人工智能技能的工作比重的增长趋势。

注：虽然在加拿大和英国 人工智能就业市场增长很快，但 Indeed.com (http://indeed.com/) 称相对来说它们在绝对规模上仍然只有美国AI 就业市场的 5% 和 27%。

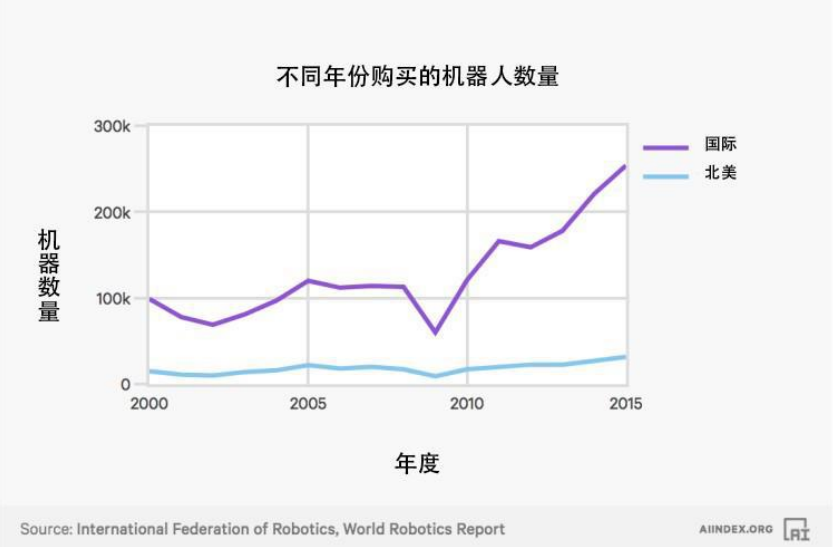
下图为 Monster 平台发布的按照所需的特定技能划分的一年内人工智能工作机会总量。



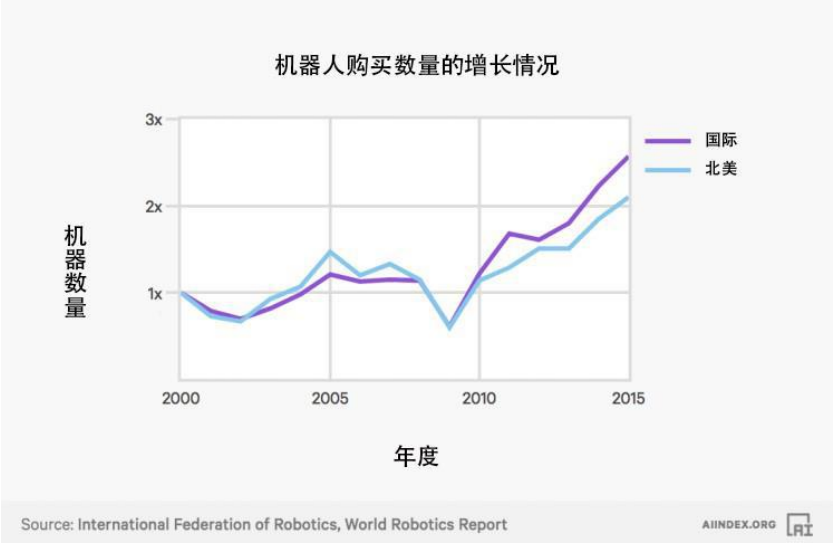
注：一份与人工智能相关的工作可能出现被计算两次的情况（属于不同的类别）。比如，一份工作可能尤其需要自然语言处理和计算机视觉两种技能。

自动化及机器人应用

工业机器人进口到北美和全球的数量。



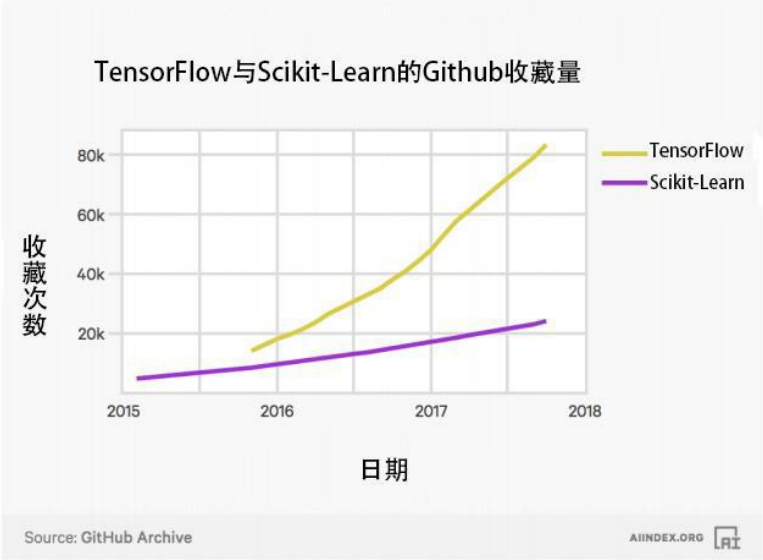
工业机器人进口到北美和全球的数量增长趋势。



开源生态

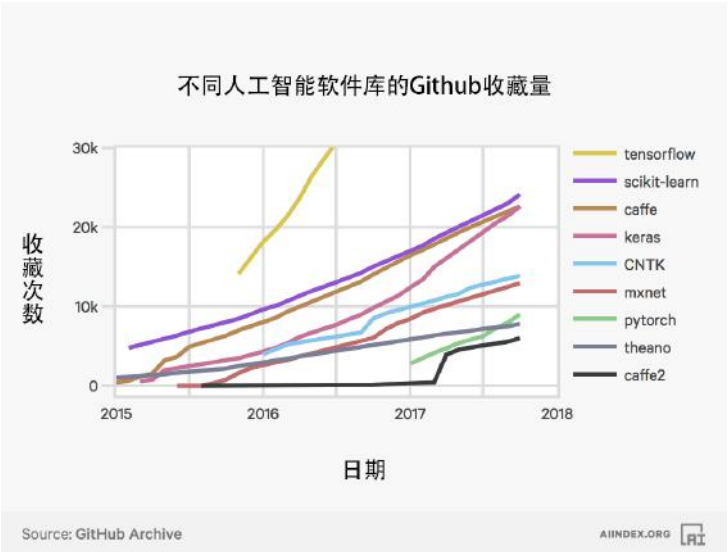
GitHub 项目统计

下图展示了 GitHub 上 TensorFlow 和 Scikit-Learn 软件包被收藏(star)的次数。二者都是深度学习和机器学习的常用软件包。



软件开发者在 GitHub 上收藏(Star)软件项目以表示感兴趣并希望快速导航至该项目。收藏可以代表开发者对软件和软件使用的兴趣。

下图展示了 GitHub 上不同人工智能和机器学习软件包被收藏的次数。

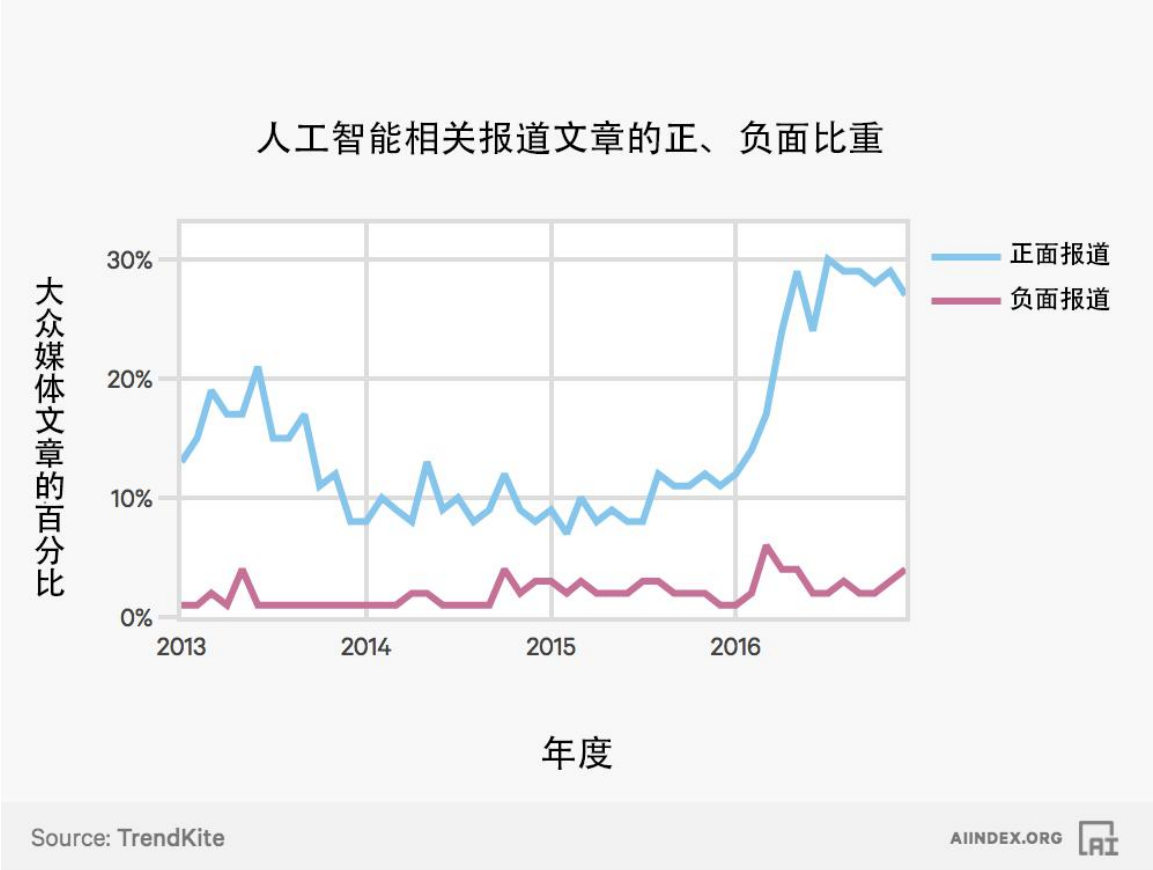


注：GitHub 库的 fork 数量遵循几乎同样的趋势（尽管每个库的 fork 量和 star 量不同）。

公众认知及媒体报道

舆论倾向

下图展示了包含关键词 「人工智能」的大众媒体文章的百分比，文章根据其意见倾向性被分为正面报道或负面报道。

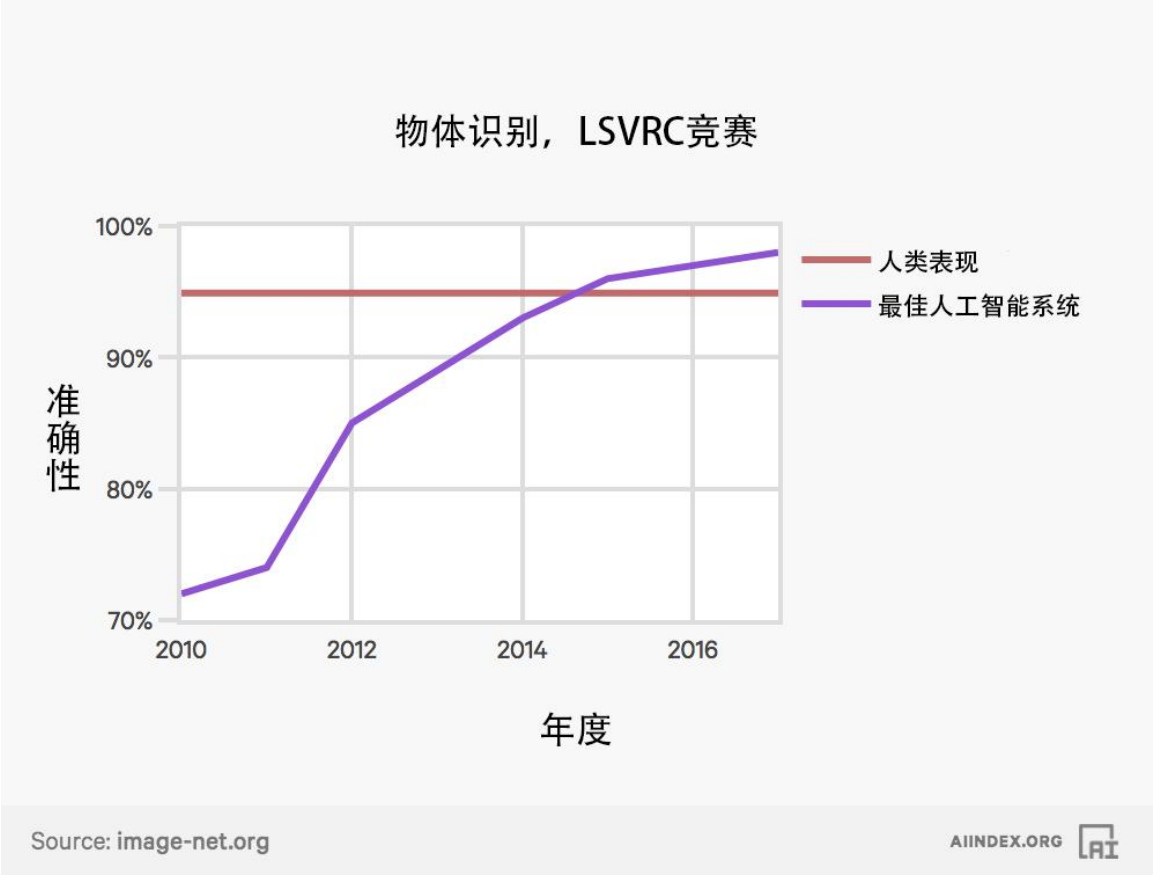


技术性能

计算机视觉

物体检测

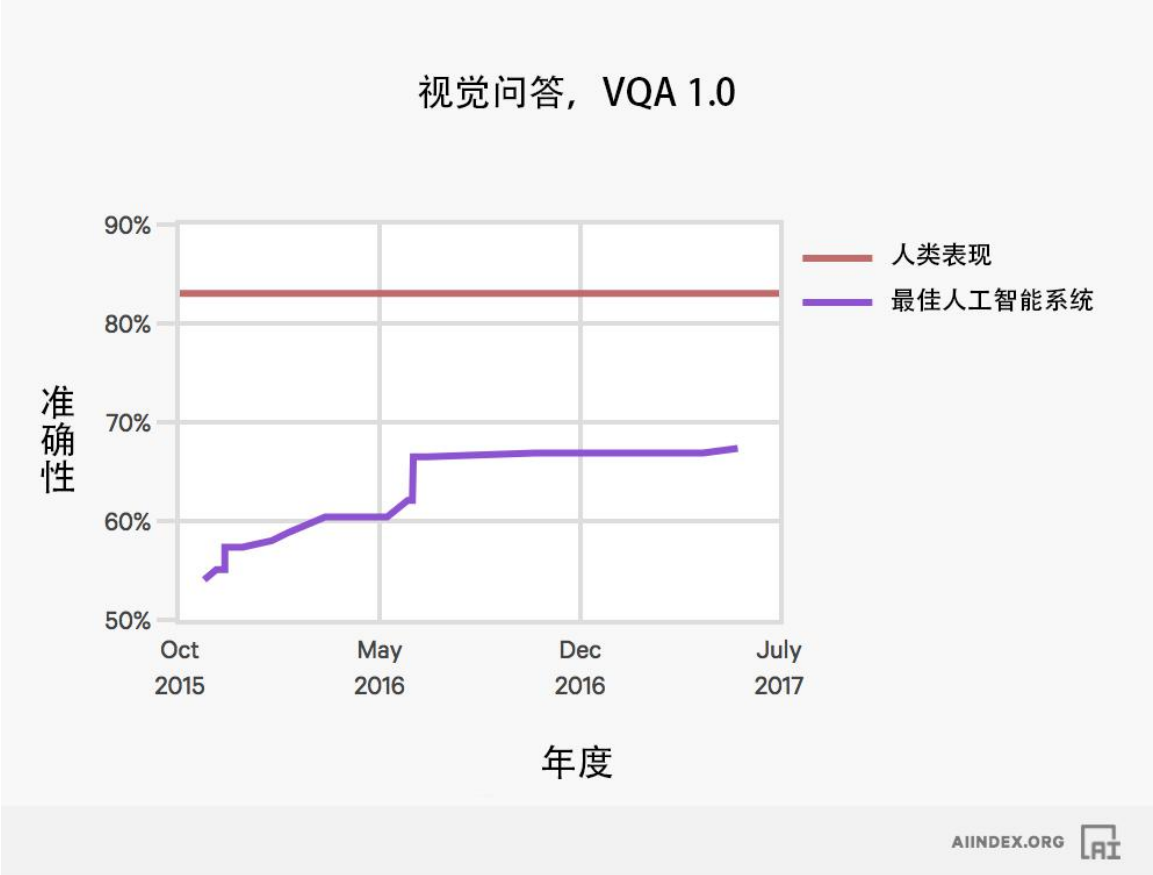
下图展示了 LSVRC 竞赛 (Large Scale Visual Recognition Challenge)中人工智能系统在物体检测任务上的性能表现。



图像标注的误差率从 2010 年的 28.5% 降至低于 2.5%。

视觉问答

下图展示了人工智能系统在针对图像问题提供开放式回答任务上的表现。

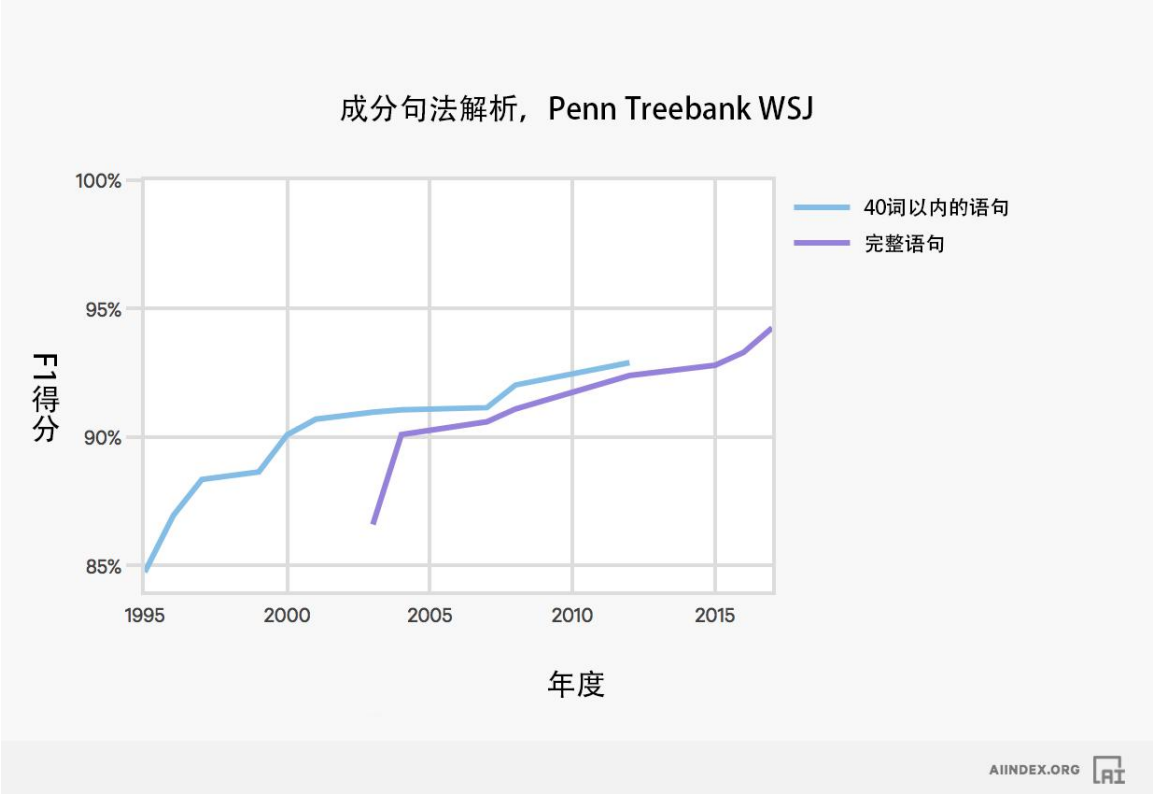


注：VQA 1.0 数据集已经被 VQA 2.0 数据集超越，目前尚不明确 VQA 1.0 数据集在未来会获得多少关注。

自然语言处理

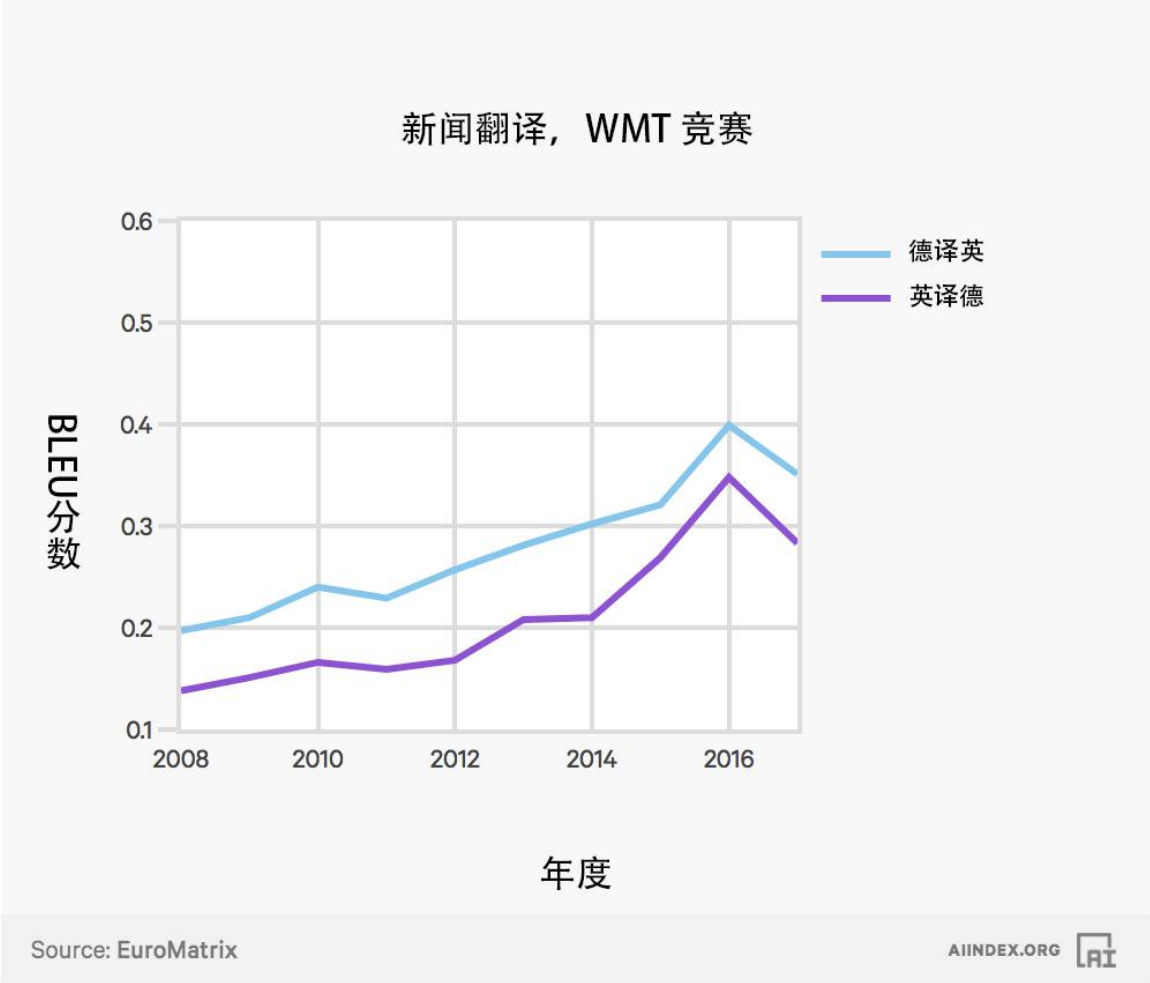
解析

下图展示了人工智能系统在确定句子句法结构任务上的表现。



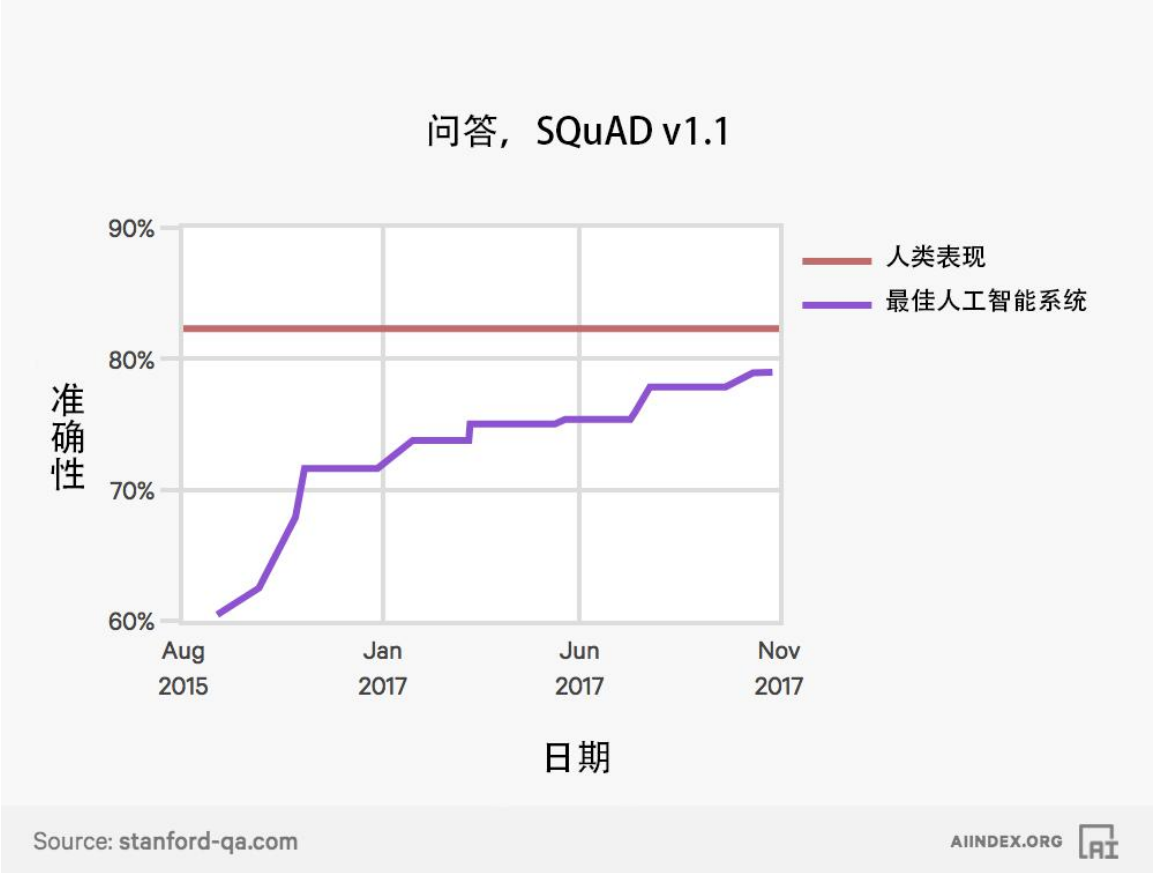
机器翻译

下图展示了人工智能系统在英德新闻互译任务中的表现。



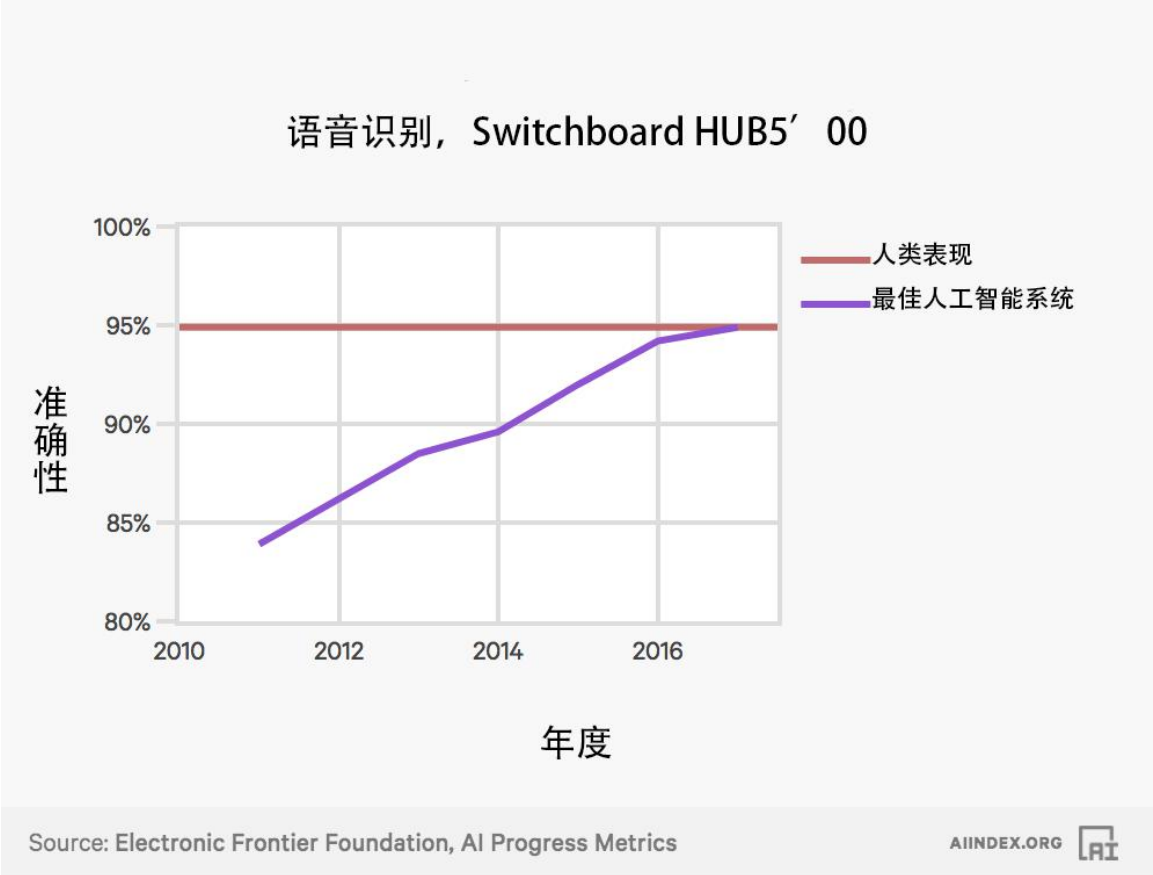
问答

下图展示了人工智能系统在从文档中找到问题答案任务上的表现。



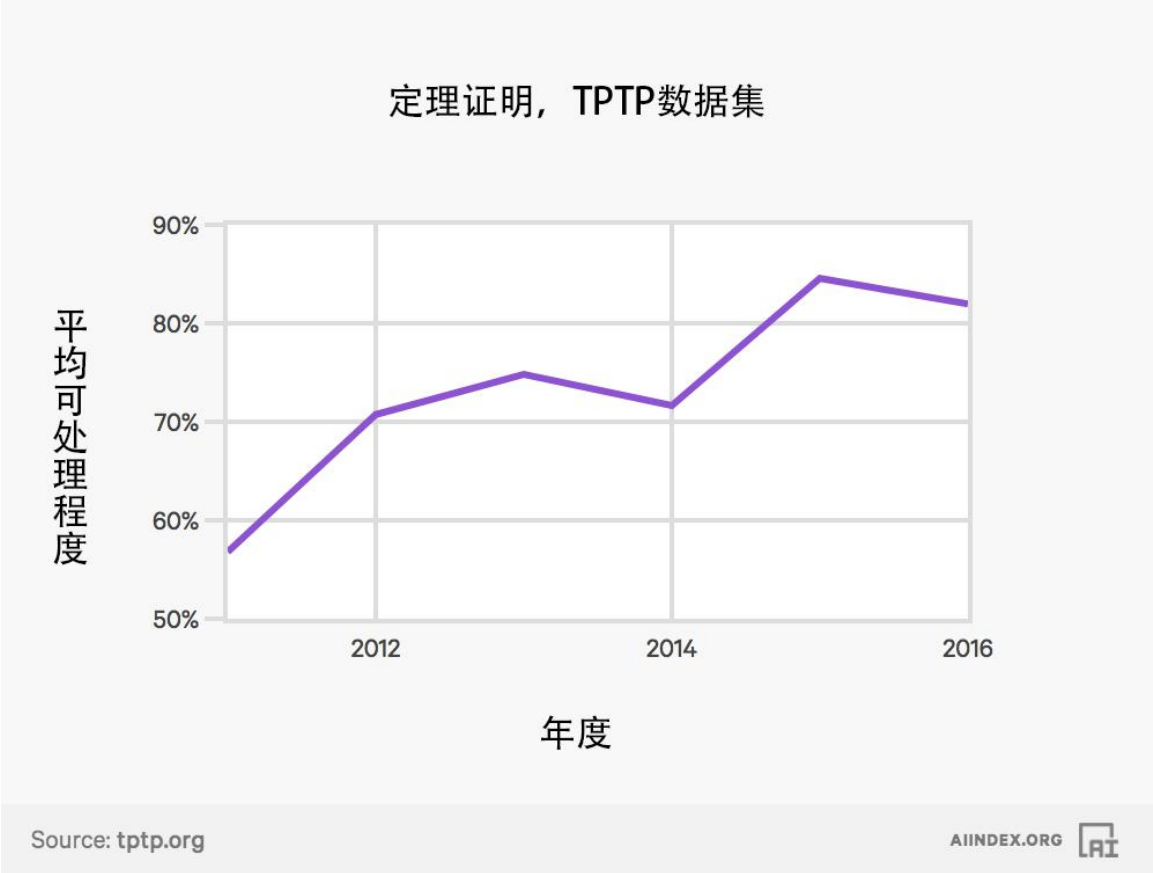
语音识别

下图展示了人工智能系统在语音识别上的表现。



定理证明

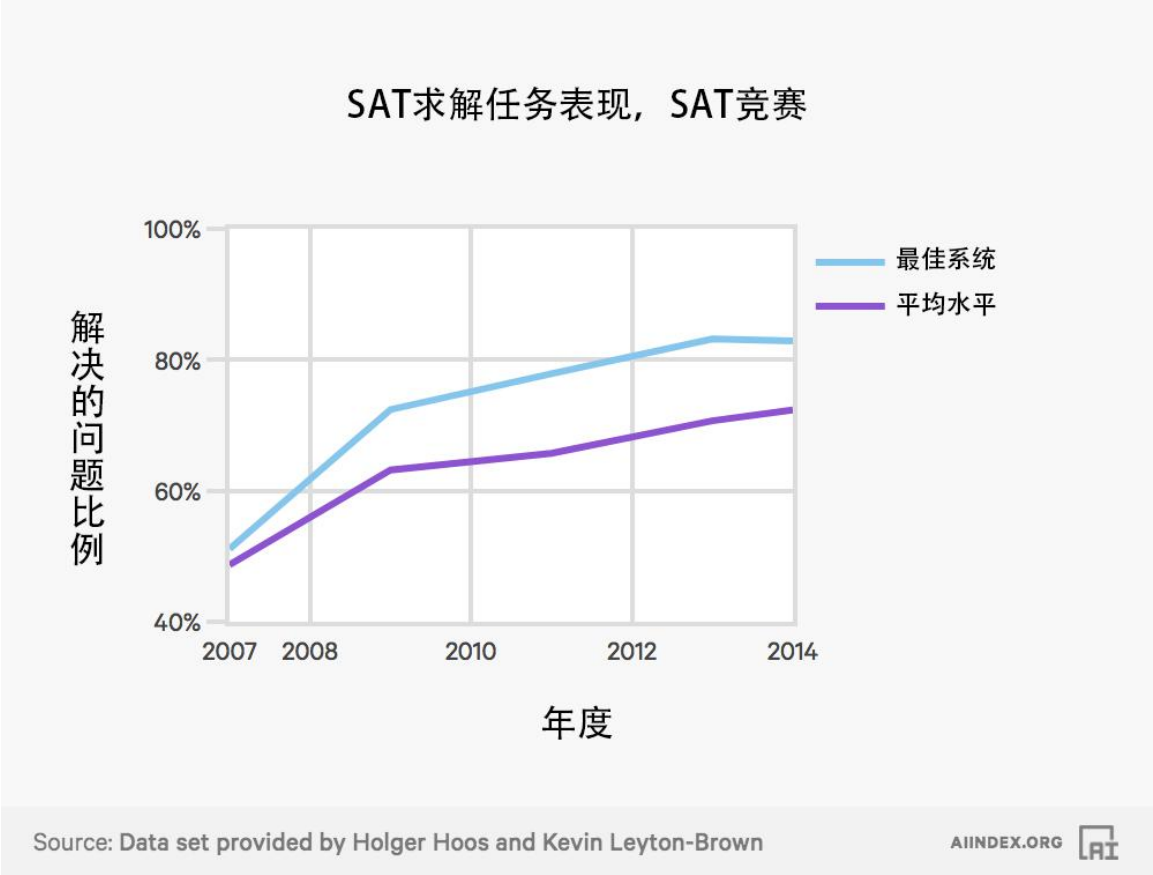
可处理度(tractability)是指自动定理证明器在大量定理的数据集上的平均可处理程度。它可以被用来衡量部分最先进的自动定理证明器。参见附录以获取与「可处理度」有关的更多信息。



注：引进最先进的证明器虽然可以解决新问题，但由于其在处理其他证明器擅长解决的问题上表现糟糕，平均可处理度可能会下降。

SAT 求解

这里指的是 SAT 求解系统解决问题（那些可应用到产业实践中的问题）的百分比。



流行趋势关系研究

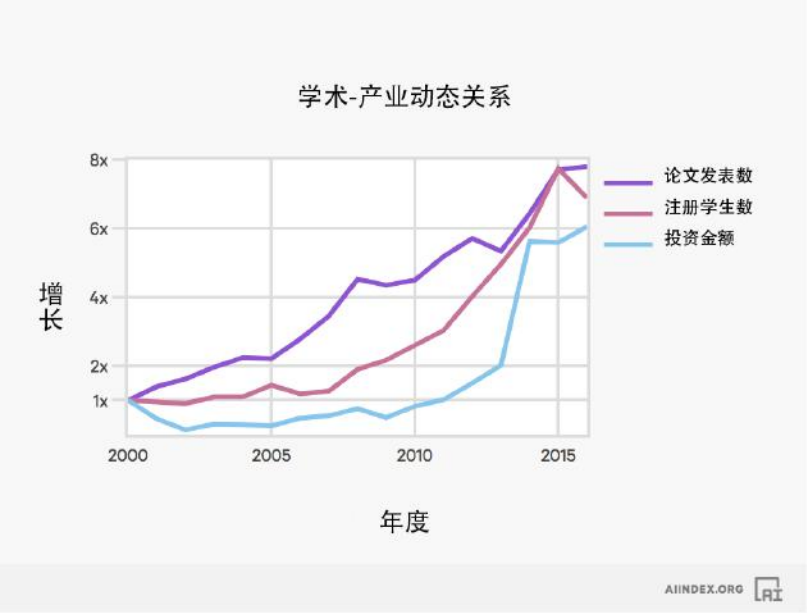
通过研究不同流行趋势之间的关系，我们可以从前述章节中的评估中获得进一步的领悟。本章展示了人工智能指数收集的数据可以如何被应用到进一步的分析中，以及这些数据如何推动了一个全新、精确的衡量指标的发展。

由于这是一个案例研究板块，我们会着眼于横跨学术圈与产业界的流行趋势去探究其之间的动态关系。进一步，我们会将这些标准整合成一个联合的人工智能活力指数。

学术界–产业界的动态关系

为了研究学术界与产业界人工智能相关活动的关系，我们首先从之前章节中选择了部分具有代表性的评估结果。特别地，我们考察了人工智能论文的发布情况与斯坦福大学人工智能与机器学习导论课程的修读情况，此外还考察了风投资本对人工智能创业公司的投资情况。

论文发表数、注册学生数和投资金额这些数量指标并不能直接比较。为了分析这些趋势之间的关系，我们首先以 2000 年为起始为每个测量指标设定了时间标准。这使得我们可以来比较这些指标随时间的增长情况变化，而不是仅仅从最后的绝对值入手分析。

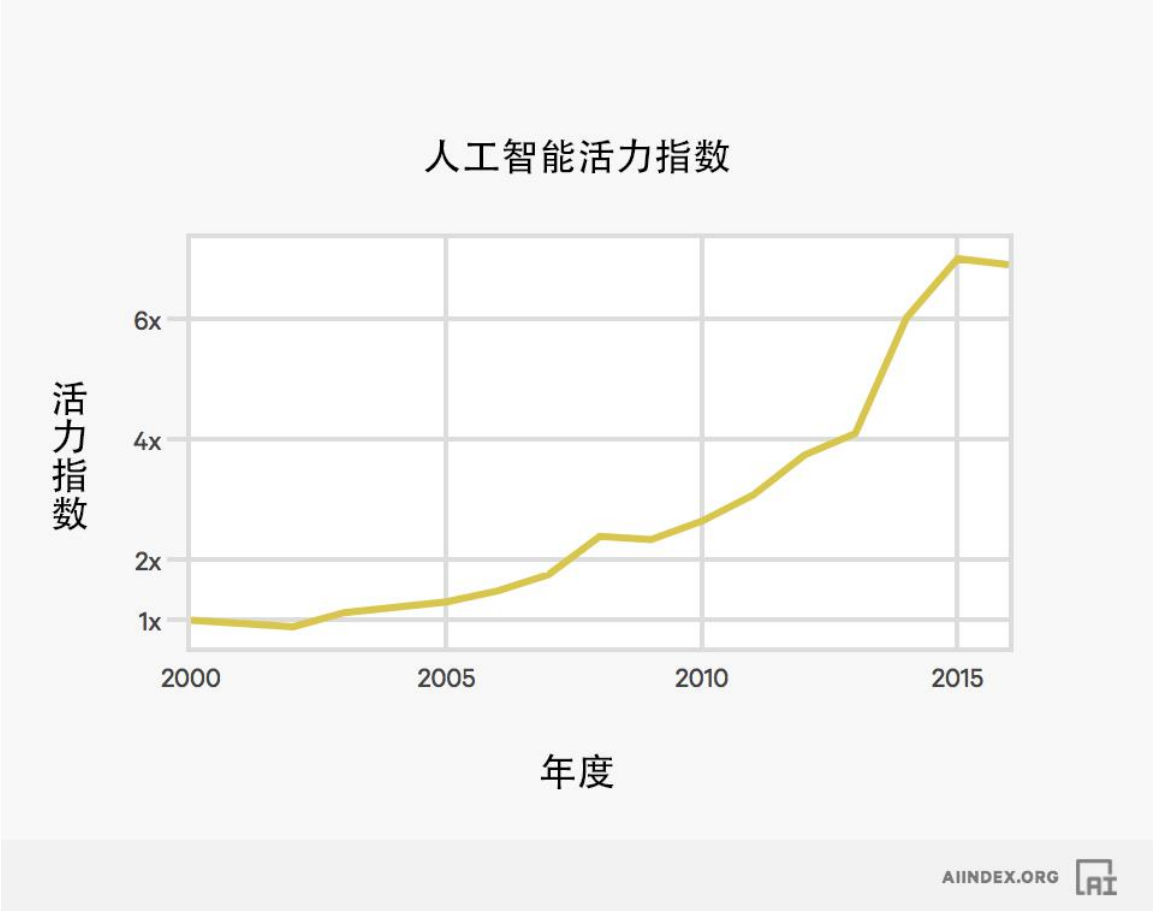


注：注册学生数在 2016 年有所下降，这反映了学校行政上的某些问题，并非没有足够的学生对课程感兴趣。具体细节可参考附录 A2。

数据显示，首先，学术活动数量（论文发表与注册学生数）在稳步上升。在 2010 年左右，投资者便开始注意到了这个领域，到 2013 年，投资者已经成为了推进该领域发展的核心驱动力。此后，学术界逐渐赶上了产业界的步伐。

人工智能活力指数

人工智能活力指数整合了来自学术界和产业界的各类数据（论文发表量、课程注册学生数、风险资本投资）来量化整个人工智能领域的活力。为了计算人工智能活力指数，我们按照时间对来自论文发表、学生课程注册和投资领域的数据进行了归一化平均处理。



我们希望这份简要调查可以激发大家在研究如何进一步分析人工智能指数中数据类别方面的兴趣，也希望可以引起讨论来研究出一个可以长期追踪的有价值的测量方法。

达到人类水平的性能？

我们很自然地会将人工智能系统和人类在同样任务上的表现进行比较。显然，在某些特定任务上计算机的确展现出了远超人类的能力；在 1970 年代，手摇式计算器便可以在算数运算上击败人类。然而，在执行通用性(general)逐渐增强的任务时，如回答问题、玩游戏以及医疗诊断，人工智能系统的能力变得越来越难以评估。

一般来说，在建构人工智能系统时，其可执行的任务都被设计得非常有限，因为这样有助于系统在特定问题或应用层面上取得突破。虽然机器可以在某一特定任务上展现出其卓越的性能，但只要任务稍加改变，其性能就会严重下降。比如，一个能够阅读汉字的人多半能理解中文发音，了解一些中国文化，甚至可以在中国餐厅里向别人推荐好吃的菜肴。相反，机器则需要完全不同的人工智能系统来进行处理这些不同的问题。

即使任务只发生了轻微变化，机器的性能也会大打折扣。

尽管在人类和人工智能系统之间进行比较存在困难，但让机器的性能赶上或超过人类仍然是件有趣的事。需要记住的重要一点是，已经取得成就并不意味着可以推广这些系统的能力。我们下面也列举了很多机器在游戏领域内的成就。这些游戏大多相对简单可控，可以很方地便应用到实验中，因此它们经常被应用于人工智能研究。

里程碑

下方简要列举了一些重要的成就及其产生的背景。部分里程碑事件意味着人工智能领域在系统性能逼近甚至超越人类水平方面取得了重大进展。

1980 – 奥塞罗

在 1980 年代，李开复和 Sanjoy Mahajan 设计了一款基于贝叶斯学习的系统 BILL，它可以玩一个叫奥塞罗的棋牌游戏。1989 年，该系统在所有计算机玩家中获得了美国全国锦标赛冠军，并以 56-8 击败了当时位列得分榜首的美国玩家 Brain Rose。1997 年，一个名为 Logistello 的程序在与奥赛罗世界冠军保持者的对战中以 6 场全胜的战绩完胜对手。

1995 – 跳棋

1952 年，Arthur Samuels 设计了一系列玩跳棋游戏的程序，并通过自我博弈来改进程序本身。但直到 1995 年，一个名为 Chinook 的跳棋程序才击败了当时的世界冠军。

1997 – 国际象棋

一些计算机科学家在 1950 年代预测计算机在 1967 年之前便可以击败人类国际象棋冠军，然而直到 1997 年，IBM 公司的深蓝系统才击败了世界冠军卡斯帕罗夫。现在，即使是在手机上运行的国际象棋程序都可以达到大师水平。

2011 – Jepoardy！

2011 年，IBM 公司研发的 Watson 计算机系统在著名电视智力游戏 Jeopardy! 上击败了卫冕冠军 Brad Rutter 和 Ken Jenning 并获得了 一百万美元的冠军奖金。

2015 – Atari 游戏

2015 年，谷歌 DeepMind 的一个团队利用增强学习系统来学习玩 49 款 Atari 游戏。系统在大部分游戏中达到了人类水平（比如 Breakout），然而在如 Montezuma’s Revenge 游戏中，系统的水平还远不及人类。

2016 – 基于 ImageNet 数据集的目标识别

从 2010 年到 2016 年，人工智能系统对 ImageNet 数据集图像的自动标注错误率从 28% 降到了 3%。相比之下，人类的错误率为 5%。

2016 – 围棋

2016 年 3 月，由谷歌 DeepMind 团队研发的 AlphaGo 系统 4-1 击败了世界围棋冠军、韩国棋手李世石。随后在 2017 年 3 月，DeepMind 发布了 AlphaGo Master，它战胜了排名第一的棋手柯洁。2017 年 10 月，《自然》杂志的一篇论文详细描述了迄今为止的另一款新一代围棋程序 AlphaGo Zero，它以 100-0 完胜先前的 AlphaGo 系统。

2017 – 皮肤癌分类

Esteva 等人在《自然》杂志 2017 年 发表的一篇论文中描述了一款人工智能系统，研究者训练该系统学习识别一个包含 2032 种不同疾病的 129450 张临床图像的数据集，并比较该系统和 21 位具备医师资格的皮肤科医生在诊断能力上的区别。结果显示，人工智能系统在皮肤癌诊断上的能力与专业医生不相上下。

2017 – 语音识别

2017 年，微软与 IBM 两家公司均在语音识别开发上取得进展，其语音识别系统性能非常接近于受限电话交换机领域内的「human-parity」语音识别能力。

2017 – 扑克

2017 年 1 月，来自卡耐基梅隆大学的程序 Libratus，在 12 万手单挑无限注德州扑克比赛中击败了 4 位顶尖的人类选手，最终赢得冠军。2017 年 2 月，来自加拿大 Alberta 大学的程序 DeepStack 分别与 11 位专业玩家进行了超过 3000 局的比赛。结果证明，算法根据统计学作出的判断超越了专业选手。

2017 - Ms. Pac-Man

Maluuba 团队设计了一款人工智能系统，它通过学习可以在 Atari 2600 游戏机上达到 Pac-Man 游戏的最高分 999900。

查缺补漏

这份初始年报覆盖了很多内容，但它并不全面。由于缺少数据或时间有限又或二者兼有，很多重要领域尚未涉及。我们希望在接下来的报告中提及这些遗漏内容。

同时我们也相信这份报告会得到源于广大社区的支持来共同参与面对这些挑战，如果你对于应对这些挑战有任何想法或是有相关数据，我们都邀请你加入人工智能指数。

技术性能

我们尚未涉及很多重要技术领域的进展，有些领域还没有明确的标准化基准 （比如对话系统、规划、连续性机器人控制）。而在如常识推理的其它一些领域，我们很难在其尚未取得显著进展的情况下去评估系统的性能。此外，还有一些领域因未收集数据而在本报告中没有被涉及（比如推荐系统，标准化考试等）。

追踪那些一直以来就缺少有效测量手段的领域也可以推动人工智能评估的严格化。人们在已取得不错进展的情况下会持续追踪发展进程。因此，本篇报告可能勾勒出了一个过于乐观的图景。

实际上，聊天机器人远远达不到人类对话水平，同时我们在这方面也缺乏一个公认的评价标准。类似地，尽管现在的人工智能系统在常识推理上的表现还远不及一个五岁孩子，但具体量化这一差距的技术标准还不明确。扩展报告未涉及的内容将可以帮助我们纠正这些过于积极的评价。此外，在这些难度更大的领域中为制定有效报告指标所做的任何努力，都会为推动该领域进一步发展做出贡献。

报告的全球性

虽然人工智能在全球范围内蓬勃发展，这份报告的研究对象却仅局限在美国。单举一个例子，现在中国的人工智能投资与活跃程度已经非常惊人，然而这部分信息却并没有写入报告。尽管在这份初始年报中我们没有时间或能力去收集相关数据，但相信今后的报告将会更加具有全球视野。

多样性与内在联系

那些研究与开发部署人工智能系统的人在塑造人工智能对社会影响方面将会扮演重要角色。如果我们打算通过数据来讨论这种影响，那么必须要量化谁能参与到人工智能的交流中来，以及衡量谁拥有影响人工智能未来研究与发展的力量。

此外，这份初始报告并没有依照性向、人种、性别、民族等其它特征对数据进行划分。有多少女性在产业界获得了人工智能研发的机会？美国的人工智能创业公司有多少是由黑人创立的？这些问题与技术产业、风投产业以及更广层面上的歧视问题都息息相关。尽管没有一项研究能够将这些复杂的动态关系纳入到讨论中来，我们依然相信在人工智能指数的努力下一定可以对这些问题展开探索，针对人工智能的社会影响提供权威分析。

政府与企业投资

此次报告涉及的风投数据仅限于美国地区，然而它们仅反映了整个人工智能研发投资的冰山一角。

政府与企业在人工智能研发过程中扮演了重要的投资角色。虽然这部分数据很难收集，但是我们可以人为通过追踪在美国和全球范围内政府与企业的投资数据来实现我们的目标。如果要想获得显著成果，这部分数据的收集则需要高度的协同与合作。

特殊领域内的影响

我们希望为医疗、自动化、金融、教育等行业提供与人工智能相关的影响力评价标准。这些领域可以说是最难以处理的。因需要对这些截然不同的领域有额外了解，相关的评价标准也难以识别与汇总。为此，我们计划与这些等等领域的专家合作来共同开展研究。

降低社会风险

此次报告并未涉及人工智能产生的潜在社会危害。我们希望在将来能为此提供评价标准，为人工智能的安全性、可预测性、算法公平性以及人工智能时代的隐私担忧、自动化技术扩张带来的伦理疑虑和其他问题提供扎实的讨论。

专家讨论

根据定义，每个数据集或是指数都会遗漏部分信息，引入一些无意识的偏见。我们的报告只能勾勒人工智能过去、现在与将来的部分图景。为扩宽视野，我们收集了来自学术界、工业界、政府与媒体相关专家的观点。

Barbara Grosz（哈佛大学）

Mind the Gap

这次发布的第一版人工智能指数报告着实让人称赞，不仅立足于坚实数据来分析人工智能的现状以及报告发布的时间，其可圈可点之处还在于作者同时注意到报告的遗漏之处。不管怎样，报告确实忽略了一些内容。有些测量标准在当下需要特别注意，因为它们不仅涉及到人工智能方法性能的测量问题，同时还与人工智能技术如何与人们发生互动、人工智能技术驱动下的整个系统如何影响人们息息相关。这在最近尤其需要重视，因为越来越多的人开始关注通过发展人工智能来补充或增强人类能力，而不是复制人类智能。IJCAI 2016 将「人类要意识到人工智能的影响」作为特别讨论主题。AAAI 2018 将会把「人与人工智能协同合作」作为新主题来讨论。此外，近年已出现了很多关注这些问题的研讨会或专题会，旨在讨论人工智能及其影响下的社会大循环中的人类境况。

报告遗漏的内容反映在自然语言处理的部分，报告提及了语法分析、机器翻译以及根据问题在指定文档中找到答案的能力，然而（正如报告自己也承认），有关对话系统或聊天机器人的部分报告却没有涉及。语法分析不需要考虑表达者的心理状态，并且机器翻译和问答的实验显示其有可能忽略话语表达者的精神状态，特别是忽略一个极其重要的目标，即理解表达者的话语含义。然而这种方法却不能使用在对话问题当中。

如这份报告强调的，只有当我们找到了一个有效的办法去测量某指标时，该指标才会被写入我们的指数报告。我们的主要挑战是测量与人和人工智能技术相关的因素，对此任何从事过机构审查研究的人都可以证明。想找到成功测量人工智能算法与系统的方法，不仅要考虑到它们的效率和计算能力，还要考虑到它们如何影响人类的生活。如果人工智能指数报告能够促进这部分测量标准的发展，那么它将为人工智能、计算机科学、乃至整个社会做出重大贡献。

找到一个测量人工智能系统影响人类生活的有效方法，是我们目前面对的重要挑战。

我希望在未来的 人工智能指数报告中，不仅能看见对人工智能课程学生注册情况的研究，同时也能看到对人工智能伦理问题课程开设数量的研究。（注明：在过去三年中，我一直在上一门名为「智能系统的设计与伦理挑战」的课程）随着人工智能驱动的系统逐渐渗透进日常生活，人工智能课程也需要告诉学生在智能系统设计之初就将伦理问题考虑进去的重要性。此外，人工智能指数接下来面对的另一个挑战是随着设计人工智能系统的公司越来越多，更多的人工智能开发者开始考虑其设计可能造成的潜在影响（这种影响与系统的设计有关），并开始思考最佳改善方法。

Eric Horvitz（微软）

我很激动看到首次年度 人工智能指数报告的发布。这个项目由斯坦福大学创立，是受斯坦福大学 AI100（研究近一百年来的人工智能）影响而产生的，并且与 AI100 的目标深度一致，通过组织定期的研究来评估和应对近一个世纪里人工智能的进展对人类生活的影响。AI100 的目的是建立一个长期存在的「连接展现」，它能够扩展人类对人工智能的理解以及人工智能对人类生活的影响。人工智能指数的倡议早在 2015 年 AI100 的常务委员会讨论中就已被提出。

人工智能指数定义了一系列关于 人工智能随时间发展的指标。首次报告提供了众多衡量人工智能发展趋势的关键指标和数据，如人工智能能力、相关活动以及探索性的「扩展」指标。这个指标展示了很多包括基于机器学习方面的最新进展，尤其是在算法进步和大规模数据资源和强大计算资源可用性加持下产生的进展。

尽管这些派生指标仅提供了大致的信号且还留有较大解释空间，我认为它们还是很有创新性和实用性的。人工智能活力指数(AI Vibrancy Index)尝试着去捕捉全局时间范围内的人工智能「活力」(Vibrancy)，并通过工业界与学术界对人工智能发展 影响力的综合测量指标反映出来。将来，这些派生指标可能会在更接近真实目标的现象中进行验证和微调，例如人工智能人才在公司中的雇佣、组成以及薪酬情况。

这份报告专门用一部分内容讨论了「人类级别」的性能，并且引出了几个比较著名且易于定义和跟踪的结果。这包括具有人类级别能力的医疗诊断（例如，通过对组织切片的视觉分析来诊断病症），在游戏中取胜（例如奥赛罗棋、西洋跳棋、国际象棋、围棋以及扑克）。该报告还提及了难以定义的人类级别能力，例如利用常识推理来理解并追踪进展的能力，这包括一个幼儿表现出来的尝试理解能力（这部分内容暂时超出目前的人工智能技术）。

我还发现了报告中重新整理的「遗漏内容」的部分。除了表达的差距之外，我希望更多的人能够在首份报告中找到更多的差距、盲点、以及定义和设计选择方面的缺陷。然而，实现从初步讨论到公开发表具体指数的跨越并不是一件小事。除了提供一系列有趣发现之外，通过发布一些指数让这些数据在社区内以合理的方式占得一席之地是完善和扩展这些指标的关键步骤。

关键的一步是通过发布这些指数使更广泛的社区参与到对话中来。

随着时间的推移，我们期望能够看到大量关于人工智能进展指数的研究。这包括朝着一个或多个方面的涉及人工智能能力、活动和影响力的深入挖掘（例如 AAAI 2017）。我认为我们应该赞扬有关人工智能的各项分析和指数，基于人工智能研究以及在多社区持续增长的影响力，我们可以期待多方面观点的百家争鸣。尽管如此，我认为一个有价值的事情是将日益全面的指标汇总起来，这可以被视作推动人工智能进步的主要贡献点，也是追踪和理解人工智能发展的「共同视角」。

人工智能指数报告的首次发布提供了关于人工智能发展最近趋势的一系列有趣见解。持续投入人工智能的分析工作令人兴奋，激发着我们对未来图表中可能出现的数据点与趋向曲线的想象。

李开复（创新工场）

当今中国的人工智能

人工智能指数是讨论人工智能的一个重要进展。年度报告中有很多关于美国市场中的重要统计数据。现在让我来补充一下当今中国人工智能方面的数据。

「数据从来不嫌多」(There's no data than more data.)。数据越多，人工智能的智能程度就越高。那么在中国到底产生了多少数据呢？

中国拥有世界上最多的移动手机用户和互联网用户，用户数量大约是美国或者印度的三倍。可能很多人认为中美之间的差距也是这么大，而事实上远比三倍大得多。在中国，使用手机付款的人数是美国的 50 倍。中国的食品运输的美国的 10 倍有余。中国的共享单车公司摩拜只花了十个月时间就从一无所有到实现每天两千万次订单（骑行次数）。每天有两千万次骑行将 GPS 和其他传感数据传送到服务器，创造了 20TB 的数据。类似地，据报道中国的出行服务公司滴滴将其数据与一些试点城市的交通管制数据联系起来。所有的网络互联都会产生有助于使现存产品和应用更加高效的数据，也会催生我们从未想到的新应用程序。

那么中国市场中的人工智能产品的质量如何呢？很多人依然记得，大约在十五年之前，中国只是山寨大厂，除此之外一无所有。然而现在聪明又热切的中国科技巨头和企业家早已被西方的创新推动甚至超越了海外同行。举一个例子，中国的人脸识别创业公司旷视(face++)最近在 3 项计算机视觉挑战赛中拿到了冠军，将 google、Microsoft 、Facebook 以及 CMU 的团队甩在了身后。

中国国务院还宣布了一项规划：在 2030 年成为全球人工智能创新中心。

中国政府对科技发展持开放态度，中国的大环境也更有利于快速启动和迭代。2017 年 7 月，中国国务院公布了「下一代人工智能发展规划」，其目标非常明确，就是要在 2030 年前成为全球人工智能创新中心。该计划有望推动人工智能成为主要行业和省部级政府工作的重中之重。如果你认为这只是空话，可以看下中国过去在高铁等项目计划与大规模创业创新运动上的政策都得到了很好的实施。我们可以期待人工智能方面的政策也会遵循相似的发展轨迹。

中国在人工智能领域所具有的前沿科技、专业实验与高速发展将为其成为强大的人工智能大国提供助力。在这个人工智能时代，我预测中美之间的两强联合是不可避免的，而事实上这种联合已经出现了。

Alan Mackworth（加拿大不列颠哥伦比亚大学）

人工智能指数的 α 版本是一个伟大的开端，其早已成为一个用来衡量人工智能进展的工具。我给出的大部分评论都是以愿望清单形式出现，而我希望涉及的范围还在扩大与重组。（这是因为）对于我希望加入的数据，很难获取数据来源。不要掉入「灯光谬论」是很重要的：不要仅仅在在「灯光」下面寻找钥匙，那里并不是钥匙最可能存在的地方。最容易到手的数据也许并不能提供最大的信息量。

最容易到手的数据也许并不能提供最大的信息量。

一个明显的缺陷是目前大多数数据以美国为中心，但美国的数据都是比较成熟的，我们希望国际人工智能社区能够通过众包来帮助填补这个漏洞。欧盟和加拿大的统计数据可能是接下来最容易获得的，比如人工智能/机器学习课程的注册人数。我们可以追踪欧盟对人工智能研究和创业公司的资助。亚洲特别是中国的数据将会非常重要，部分数据已经可以使用。

对缺乏数据源多样性的关注引起了我们对人工智能从业者和研究者性别及地域差异的重视。

在「领域活力」方面，除了学术界和工业界，政府应成为一个主要相关者，比如关于人工智能研究的经费数据。有没有关于衡量监管活动和政府治理指标方面的研究？这些活动肯定在增加，但是它们的衡量方式是否有意义？

考虑为 AI2、OpenAI、WEF 和图灵研究所等新兴组织增加一个非政府组织类别。

在学术界方面我们应该可以得到以下数据：

a) 来自学术界的人才供应，例如每年从人工智能/机器学习专业毕业的硕博士学生数量，附以人工智能和机器学习方面每年创作的论文数量和比例。

b) 学术界的人才需求，例如在人工智能/机器学习方面的特定的学术职位数量（也可以是 计算机研究协会 CRA 招聘博士后和教职人员的广告数量），以及人工智能和机器方面职位的比例。

两个非常受欢迎的指数将会是学术界和工业界的人工智能工资水平。猎头公司、咨询顾问和 CRA 可以为获取这一指数提供一定的帮助。虽然可以获得可靠数据，但棘手的是目前我们确实只有来自于纽约时报的八卦和轶事。

将发表论文数量和学术会议的出席情况放在学术界下面其实是一种错误的分类，因为工业界的研究力量也很强大，在论文发表者和参会单位中不乏科技公司的身影。大量人工智能领域的相关活动独立于学术界、工业界、政府和非政府组织，因而应解除这种分割。其他有用的测量指标包括会议论文的年提交量和人工智能书籍的年出版量，比如可以通过亚马逊的书籍分类测得。

在技术性能评估方面，人工智能对验证码识别问题的解决程度可能是一个不错的指标。先进技术列表中有一个被遗忘的重要领域就是多智能体系统(MAS，multi-agent-system)。MAS 领域两个理想的候选指标是已经设立了 2050 目标的 RoboCup 以及交易人代理竞赛(TACs)。有人认为曾获得 Loebner 奖的图灵测试会是一个不错的候选指标，这是否是个好主意将是一个值得争论的问题。事实上，更多的元评论对于确定包含指数在内的评价标准是很有用的。对于那些目前还很难量化的重要活动而言，目前是否也应该被提及以免在以后被忽略掉？

最后，元-人工智能活跃度(meta-AI activity)的指标如何呢？虽然目前还不明确具体的量化方法，但元人工智能研究所、组织、合作伙伴、智库和包括元人工智能衡量指标在内的指数都在明显地呈指数级增长，它们关心人工智能研究本身，测量人工智能，预测人工智能对社会、就业、经济、法律、政府和城市的影响。我们能够量化并预测元人工智能的增长和结果吗？或许某时就会有关于元人工智能奇点的笑话呢？

吴恩达（Coursera，斯坦福大学）

人工智能是一种新电能

人工智能是一种在改变多个产业的新电能。人工智能指数将会帮助当代人追踪并使用这种电能，它也能帮助未来的人们去追溯并理解人工智能的发展。

人工智能是一种全球现象

此外，现在人工智能是一种全球现象，人工智能指数提醒我们每一个人必须突破自己的国界去理解这种全球性进展。美国和中国拥有最大规模的的投资和急剧增长的应用程度，加拿大和英国也做出了突破性的研究贡献。从网络搜索到自动驾驶再到客服机器人，人工智能改变了很多技术系统的基础，这给了一些国家机会在某些应用领域中「挑选」现有人才。出台合理的人工智能政策的国家将会取得更加迅猛的进步，但如果政策思想不佳，则可能要面临被甩在后面的风险。

出台合理的人工智能政策的国家将会取得更加迅猛的进步，
但如果政策思想不佳，则可能要冒着被甩在后面的风险。

人工智能子领域中深度学习的转换

深度学习首先改变了语音识别，然后是计算机视觉。现在自然语言处理和机器人学也在发生相似的变化。最近语音和视觉精度的提升带来了使用语音技术和计算机视觉技术相关应用的繁荣，例如语音控制扬声器和无人车。现在，自然语言处理方面的深度学习转型也在顺利推进，这将会引领一波新的应用繁荣，例如聊天机器人。深度学习在机器人学中也提供了很显著的推动力，这也将会带来很多新的应用（例如新的制造能力）。

Daniela Rus（麻省理工学院）

人工智能：一个正向变化的向量

我们的世界已经在飞速改变了。今天，远程呈现视频技术(telepresence)能够让学生和老师在网上见面，能够让医生远程治疗他们的病人。机器人帮助工厂实现新的制造力，网络传感器使得工厂可以对设施进行监控，3D 打印可以创造定制化的商品。我们生活在一个一切皆有可能的世界中。伴随着我们展开对人工智能可应用领域的想象，这种可能性只会变得更大。

在全球规模下，人工智能会帮助我们在解决某些重大挑战问题方面提出更好的想法：如从海洋、温室气候和植被状况的监控传感器中收集大量数据，并分析与理解气候变化；通过数据驱动的决策制定来帮助政府治理；通过检测、匹配和重新规划需求关系来消除饥饿；使用信息物理传感器来预测和应对自然灾害。这将有助于我们通过配合学生进步的 MOOC 教育来实现教育民主化，并确保每个孩子都能习得找到好工作所需的技能，创造美好生活。钢铁侠不再是一个漫画人物，而是代表了甚至可以帮助孩子把童年梦想变成现实的技术可能性。

人工智能会帮助我们在解决某些重大挑战问题方面提出更好的想法。

在局部范围内，人工智能将有可能让我们的生活变得更加安全、便捷、具有满足感。这意味着自动驾驶汽车可以载着我们上下班、可以在青少年驾车时避免威胁生命安全的意外事故，意味着我们可以使用大数据学习到的知识建立起定制的医疗保健。与常识相反，我们对工作的满意度会提升而非降低，因为人工智能和机器人带来的生产力的提高使我们摆脱单调的任务，让我们能够专注于计算机无法胜任的创造性任务、社交以及其他更高端的任务。

当我们将计算能力指向人类在没有机器支持的情况下无法解决的问题时，所有的事情就变得可能了。机器人、机器学习和人工智能三个不同领域正在取得进展。机器人将计算转化为运动并赋予机器自主权。人工智能的智能水平提升，使机器能够完成推理。机器学习跳过机器人和人工智能，让机器来学习、改进和预测，各个领域正在迅速取得进展。人工智能指数引入了几个指标来跟踪这一进展，这些指标为该领域的教育、研究和创新状况提供了一个重要的量化观点，并提供了对一般趋势的洞见。

虽然人工智能有潜力成为能带来巨大积极变化的载体，了解当今的技术水平还是很重要的——今天的方法有所为、有所不为。人工智能指数定义了智能任务，并为这些任务衡量了最先进的人工智能系统的性能。它还为人工智能教育和人工智能方面的重大挑战提供了一个框架。

人工智能可以成为带来巨大积极变化的载体，了解当今的技术水平很重要——今天的方法有所为、有所不为。

智力问题、大脑如何产生智能行为、机器如何复制它，仍然是科学和工程领域的深度挑战。这需要训练有素的研究人员和持续的长期研究与创新来解决，而人工智能指数正在追踪这一进展。

Megan Smith（美国政府第三任 CTO，Shift 7） 和 Susan Alzner （联合国非政府组织联络服务）

人工智能指数：促使我们改进不足之处

本年报提出的目标很重要，特别是对于人工智能发展指数而言，「……旨在促进基于数据的人工智能的知情对话」——分享和支持跨越全球所有社区为和在其内部所需的多方会谈的新兴趋势。这个团队正勇敢地担负起责任，帮助支持合作工作，填补空白，共享世界上大多数人正在经历的可见性。它指出「……若没有关于人工智能技术状态的推理的相关数据，在与人工智能的对话和决策中，我们本质上是在盲目行动。」对于那些不直接在本领域工作，特别是那些没有计算机科学或其他背景的人即绝大多数人而言，「盲目行动」一词显得格外正确。

多样性 (Diversity)和开放性 (Inclusion)极其关键。
由于偏见、歧视性的文化模式以及系统排他性的习得行为，
我们身上的人性关怀正在消失。

该报告包括一个名为「遗漏内容」的重要开放性部分，承认仍然存在许多问题需要解决（甚至是开始考虑或优先考虑）。其中，多样性和开放性极其关键。由于大量有意识或无意识的偏见、重大的歧视性文化模式、以及存在于几乎全部社区和媒体形式当中的系统排他性的习得行为，我们正从对话和设计队伍当中不断丢失大部分的人性关怀。这些基于数据的新兴人工智能/机器学习技术对全球人类及其生活的影响已经极其显著，我们也将在未来几十年里看到巨大的改变——即使是在未来的短短几年内也将发生深刻的变化。我们迫切需要在技术部门各个层面和维度上、在对话中、在技术决策制定者之间，以及在各部门对科技的应用中，迅速和根本地提升多样性与开放性。

报告认为「人工智能领域仍然在迅速发展，甚至专家们都很难理解和追踪整个领域的发展进程。」我们感谢世界各地的专家和团队正在加紧简历有用的开放性论坛和类似我们描述的初期指标工具，从而欢迎和吸引更多人进行卓有成效的对话。以下是一些初始的言论选，以供考虑：

2016 人工智能的未来——白宫科技政策办公室 OSTP

作为开始广泛参与人工智能/机器学习议题的一部分，奥巴马前总统要求基于科技政策办公室(OSTP)的美国首席技术官团队和包括国家科学技术委员会(NSTC)在内的其他领导人合作举办一系列政府对话会议；这些活动发生在 2016 年夏季。 在 2016 年 10 月份举办的白宫前沿会议期间启动了联合举办的公开集会及后续报道，其中包括人工智能的国家发展轨道。为人工智能的未来做好准备(Preparing for the Future of Artificial Intelligence，May 3 2016 BY ED FELTEN)——研习会和跨部门工作组，旨在了解更多人工智能的利益和风险。

关于人工智能未来的行政报告(The Administration’s Report on the Future of Artificial Intelligence，October 12, 2016 BY ED FELTEN AND TERAH LYONS)——重点关注人工智能的机遇、考虑和挑战。

势在必行：拓宽参与和应用，培养伦理和价值整合

我们正处于一个深刻的变革时代，互联网就像人类一样互相连通，且数据变得比以往任何时候都更加整合。人工智能/机器学习扮演着越来越重要的角色——数据科学、大数据、人工智能/机器学习和社区连接的新兴智能是如此重要。我们应该如何避免糟糕的结果？无论是「机器人启示录」还是史蒂芬·霍金、伊隆·马斯克等人迫切谈论的大规模破坏性情景，亦或只是滥用人工智能/机器学习为改善人类生活所带来的危险可能性（如贫困、平等、饥饿、正义、抵制偏见等）。

如果能迅速扩大人工智能应用的热点话题范围，并让这些人数众多的部门更多参与进来，就可以帮助解决这些挑战。无论是何种方式，我们都需要将共同的价值观整合到这些体系当中，并尽可能扩大创造力的覆盖范围。我们应考虑一下是否真的想要通过共享数据库来训练所有这些技术？我们的确做出了贡献，但同时也产生了一些恶果。人工智能在得到良好使用的同时，也将自己武装了起来。现在，技术既不好也不坏，它只是取决于于设计和处理的方式，包括恶意应用、偏见和恶意迭代。

所以，对于这个正在发生转变的事实，我们当中更多的人必须更加「警醒」，调整自己、进行关于伦理道德的艰难对话，并且采取行动。我们需要讨论将自己武装起来的的人工智能，探索潜在的控制方法和其他选择。让我们参与进来，试着将想法变成现实。即便如此，黛安·冯·芙丝汀宝和埃隆·马斯克都提到，我们人类都会变成人工智能的「宠物」。

关于这个议题，有一些重要的文件——特别是联合国和其他具有全球议定价值的文件，包括：

- 2030 年可持续发展议程和目标
- 世界人权宣言
- 原住民权利宣言
- 北京宣言和妇女人权行动平台
- 残疾人人权公约
- 经济、社会和文化权利公约
- IEEE 伦理一致性设计通用准则——尤其是关于人类利益清单的原则 1（第 16 页）

2017 年 6 月份，在麻省理工学院的毕业典礼上，苹果公司首席执行官蒂姆·库克表示，

「我并不担心人工智能能够给予电脑如同人类一般的思考能力。我更关心的是人们像电脑一样，毫无价值观和同情心且不计后果地思考。这就是需要大家帮助我们防范的地方。这是因为，如果自然科学是在黑暗中的搜寻，那么人文科学就是一根蜡烛，向我们展示所到之处及前方面临的危险。正如乔布斯所说，单靠技术是完全不够的。技术与人文、文学的联姻，才能让我们的心灵歌唱。以人类作为工作的核心，方能产生巨大的影响。」

我们需要广泛留意到组织、协会和个人的领导工作，从而协作采取行动。例如：

- 算法正义联盟(The Algorithmic Justice League)强调算法偏见，为人们表达对代码偏见的关注和经验提供空间，并开发问责式实践。
- 向联合国请愿(The petition to the UN)要求对武装化的人工智能迅速采取行动，敦促人类在这个话题上开展全球性参与。
- 计算机科学全民运动(The Computer Science for All Movement)旨在进行美国和其他国家的科技融合。
- AI For ALL 计划(The AI4All initiative)培养未来的人工智能技术专家、思想家和领导者。
- 致力于能力建设领域，即尚未使用过多人工智能和机器学习的领域，推进解决方案——借此，任何需要的议题都可能利用这一技术产生积极影响。
- 将有主题专业知识的专家和传统上不具有此类能力的人员相结合（包括「TQ」，即各领域的技术商，比如政府中的「公共政策 TQ」）

现在发布的人工智能指数是新生的、不完善的，甚至很可能在某些我们不了解的方面是幼稚的。但是，它的伟大之处在于这是一个采用了开放性合作方式的开始，而在未来漫长的旅途中，我们很难观察到未来不同的到达终点。最近，《Weapons of Math Destruction》一书的作者 Cathy O'Neil 写了一篇专栏文章《The Ivory Tower Can't Keep Ignoring Tech》(New York Times Nov 14, 2017)，敦促各界学者参与进来。我们十分赞同，也邀请、希望每个人特别是青年人选择参与本次对话——因为最重要的是，人工智能指数的邀请是面向所有人的。

第一版 人工智能指数是新生的、不完善的，甚至很可能在某些我们不了解的方面是幼稚的。但是，它的伟大在于它是一个开始。

Sebastian Thrun （斯坦福大学，Udacity）

人工智能近期进展的重要性再强调也不为过。人工智能领域已经有超过 60 年的历史，并且已经产生了极其重大的影响。人工智能，是 Google 搜索算法，Amazon 网站设计和 Netflix 电影推荐的核心组成。但是，强大的计算机配以规模空前的数据集使得它变成了社会游戏规则的改变者。在过去的短短几年中，已经发展出了与高技能人类相当、甚至能击败人类的系统。DeepMind 的 AlphaGo 击败了世界上的顶尖围棋高手。在我们的实验室中，我们发现人工智能系统能比某些通过职业认证的皮肤科医生更准确地诊断皮肤癌的图像。我也一直认为，Google 的自动驾驶汽车能比像我这样的普通驾驶人员驾驶得更好。现在，Google 已经在公路上部署这些车辆，并不需要驾驶安全员。而初创公司 Cresta 已经证明，与人类专家合作的人工智能系统能使在线销售团队效率翻番。

我相信，在不远的将来，人工智能将会把我们从重复性工作中解放出来。人工智能系统能够通过观察专家工作，逐步掌握我们在日常工作中应该具备的技能。在这种情况下，机器将会承担更多的重复性任务，让我们能自由地进行创造性工作。这场革命在历史上也有类似的对比。在蒸汽机发明之前，大多数人都是农民。大部分人是通过自己的体力和敏捷程度（而不是他们的头脑有多聪明）在社会中占有一席之地的，他们在这些领域进行高度重复性的工作。但是，机器将曾经的农民变成了「超人」。根据 FarmersFeedUS.org (http://farmersfeedus.org/)提供的数据，一个美国农民能够供给 155 人的食物。因此。低于 2% 的美国人从事农业，使得 98% 的人能自如地找到不同工作。美国 75% 的劳动力都在办公室中工作。我们的工作包括律师、会计师、医生和软件工程师。大部分的工作都是高度重复性的。可以想象，人工智能技术能够习得重复性工作的模式，并且帮助我们更快地完成工作。最终，我们都将成为「超人」，由人工智能帮助组织建立起我们的生活和对世界的理解。

这是好事还是坏事呢？我想，如果在未来回溯历史，将会见证人类的飞速发展。当我们停止进行重复的体力劳动时，我们会接受更高程度的教育，变得更具创造力。伴随着这场变革，我相信我们将进入一个人类有着前所未有的创造力的时代。

随着这场变革，我相信我们将进入一个人类有着前所未有的创造力的时代。

但这也给人类带来了负担。据估计，如果自动驾驶出租车成为日常交通的主要方式，九分之一的就业机会将会受到威胁。为了走在这些变化的前面，我们必须终身学习。我们要习得新技能，学习掌握新技术。作为社会的一份子，我们需要找到新的方式，以帮助适应这些变化。

这份报告非常重要。它认真研究了人工智能的最新进展，并记录了其对社会的影响。我赞成作者将这样细心的调查报告整合在一起。我希望，这份报告能够为建设性地为大量的人工智能公开讨论做出贡献。如果我们能征服这场挑战，如果我们做好了准备，而且如果我们是引领者——对于我们所有人而言，未来将十分精彩。

Michael Wooldridge（牛津大学）

从我的角度出发，无论是作为牛津大学计算机科学系主任，还是人工智能国际联合会议主席，或是欧洲人工智能协会主席，这份工作报告都十分吸引人。该报告提供了令人信服的全面证据，从很多方面解释了人工智能技术在其早期产生的核心问题（游戏玩法、机器翻译、定理证明、问题解答等）上正取得稳步进展。在这些部分，人工智能已经处于或者超过了人类专业水平。报告还提供了很明确的、看似真正需要的证据——人工智能吸引了学生和业界的注意力，人工智能课程的招生人数激增，人工智能初创公司数量大幅上涨。

人工智能泡沫目前显然存在。从这份报告中我想到的问题是该泡沫是会爆发（参见 1996-2001 年互联网泡沫的繁荣时期），还是会慢慢缩减；而当这种情况发生时，又会留下什么呢？我非常担心的是在大规模投机性投资之后出现幻灭，由此引发另一个人工智能的严冬。很多江湖骗子都很高兴利用人工智能来招摇撞骗，而这也导致一些媒体开始将篇幅用于报道错误的人工智能信息，在我看来这已经接近疯狂的边缘。（最近的一个例子，见 http://tinyurl.com/y9g74kkr）

不过，虽然我认为未来数年内的泡沫是不可避免的，但有理由相信这将是一个郑重而缓慢的收缩，而非一次规模巨大的萧条。这其中的主要原因是，人工智能指数清楚地表明人工智能正在将自身能力付诸实践。在各类任务中，人工智能系统正在稳步（有时是快速）提高性能，而这些功能在很多不同的应用领域都取得了成功。换句话说，我认为目前的人工智能泡沫中其实蕴含着大量实质性的内容，而且各大公司现在都明白如何高效使用人工智能技术。正是因为有着明显的实质，且在科学的角度上有着显而易见的实际进展，所以我相信我们不会看到人工智能的严冬以及专家系统繁荣后的终结。（我期待阅读 2027 年的人工智能指数报告，去看看我的论断是如何奏效的。）

我认为目前的人工智能泡沫中其实蕴含着实质性的内容。

强人工智能（通用人工智能）的发展没有出现在报告的「技术性能」部分，这也是出于完全可以理解的原因。包括我在内没有人知道如何描述这种进展，这正是报告中没有提到强人工智能的主要原因。目前报告中所描述的与强人工智能最接近的就是问题回答，它可以被认为是表明理解力中的一种，但这还不是强人工智能。我不认为图灵测试是合适的测量强人工智能的方法，无论它有多么大的影响力和独创性，而作为人工智能的受众，其原因众所周知。那么，我们如何测试强人工智能的进展？这对于消解公众对于强人工智能实力的担忧非常重要——很多媒体仍在宣扬这种恐惧，而这种看法也影响了很多人。

加入行动

我们相信，如果没有多元化社区的参与，任何旨在理解人工智能技术发展进程和影响的倡议都不会取得成功。
如果没有多元化社区参与，任何旨在理解人工智能技术的影响的倡议都不能取得成功。

您有很多可以支持人工智能指数的方式，不管支持力度是大是小，我们都希望您能参与进来。

分享关于人工智能指数 2017 报告的反馈

我们想听到您对本报告中的数据的看法——您认为我们遗漏了什么、在收集有关沟通人工智能相关数据的信息时我们应该利用什么机会，您可以随时写一份透彻的审阅意见并通过电子邮件发送给我们，您也可以在 Twitter 上向 @indexingai 分享您的简短见解。

开放您的数据

如果您或您的组织机构有能力分享相关的数据，请与我们联系。我们与很多组织机构合作完成了这份报告，而且强有力的合作关系仍将继续是人工智能指数 的运营模式继续发展的基石。

提供领域知识

人工智能指数报告将在未来迭代版本中量化人工智能对特定垂直领域的影响方式，比如医疗健康、交通运输、农业等。我们必须与各行各业的专家合作才能完成这一目标。如果您或您所知的某个组织机构有可能成为追踪人工智能对特定领域的影响的信息源，请与我们联系。

纠正我们的错误

我们希望尽可能保证我们所提供的信息的准确性。即便如此，我们的数据有着非常广泛的来源，我们也可能会在数据聚合过程中出错。如果您看到了任何错误，请告知我们；然后我们会更新这个 PDF 的版本以及我们网站上的信息。

支持人工智能指数的数据收集

关于人工智能的数据总是多于我们可以收集和组织的量。我们希望能与您合作收集最重要的信息。

如果您知道某个我们遗漏了的有用的数据源或某个应该追踪的指标，请给我们发送信息，帮助我们改进。

帮助我们国际化

我们已经开始与国际合作伙伴合作收集美国之外的数据了。如果您有相关的国际数据，我们希望收到您的消息。

保持联系！

本报告中有什么发现令您感到惊讶？或者您惊讶于我们遗漏了哪些方面？请在 Twitter 通过 @indexingai 与我们联系或向 feedback@aiindex.org 发送邮件来联系我们。最后，如果您想获得人工智能指数的定期更新以及了解人工智能的现状，请在 aiindex.org 订阅邮件更新。

感谢

人工智能指数得到了斯坦福大学「人工智能百年研究」(AI100/One Hundred Year Study on AI)的帮助，该项目催生了人工智能指数并为其发布提供了种子基金。我们也得到了斯坦福大学下列人士的多种方式的慷慨支持：

Tom Abate、Amy Adams、Russ Altman、Tiffany Murray、Andrew Myers

我们也非常感谢谷歌、微软和字节跳动（今日头条）提供的额外启动资金。但是，人工智能指数是一项独立研究，并非一定反映了这些组织机构的观点。

在人工智能指数的启动阶段，其咨询委员会提供了非常有益的智慧和建议，该委员会的成员有：

Michael Bowling、Ernie Davis、Julia Hirschberg、Eric Horvitz、Karen Levy、Alan Mackworth、Tom Mitchell、Sandy Pentland、Chris Ré、Daniela Rus、Sebastian Thrun、Hal Varian、Toby Walsh

我们也很感谢专家论坛(Expert Forum)的贡献者，他们中有一些已经被列在了上面的咨询委员会成员列表中，他们对本报告以及对人工智能在社会中位置的持续讨论的贡献具有无法衡量的价值：

Susan Alzner、Barbara Grosz、Eric Horvitz、李开复、Alan Mackworth、吴恩达、Daniela Rus、Megan Smith、Sebastian Thrun、Michael Wooldridge

我们也很感谢下列人士提供的建议和支持：

Toby Boyd、Kevin Leyton-Brown、Miles Brundage、AJ Bruno、Jeff Dean、Catherine Dong、Peter Eckersley、Stefano Ermon、Oren Etzioni、Carl Germann、Marie Hagman、Laura Hegarty、Holger Hoose、Anita Huang、Dan Jurafsky、Kevin Knight、Jure Leskovec、Tim Li、Terah Lyons、Mariano Mamertino、Christopher Manning、Gary Marcus、Dewey Murdick、Lynne Parker、Daniel Rock、Amy Sandjideh、Skyler Schain、Geof Sutclife、Fabian Westerheide、Susan Woodward

这些人士所代表的以下组织为初始报告提供了数据：

艾伦人工智能研究所、Crunchbase、电子前线基金会、Elsevier、EuroMatrix、Google Brain、Indeed.com、Monster.com、Sand Hill Econometrics、创新工场、TrendKite、VentureSource

我们感谢下列人士帮助我们取得了会议参会数据：

Chitta Baral、Maria Gini、Carol Hamilton、Kathryn B. Laskey、George Lee、
Andrew McCallum、Laurent Michel、Mary Ellen Perry、Claude-Guy Quimper、
Priscilla Rasmussen、Vesna Sabljakovic-Fritz、Terry Sejnowski、Brian Williams、Ramin Zabih

我们也感谢下列人士帮助我们取得了课程选修数据：

Lance Fortnow、Charles Isbell、Leslie Kaelbling、Steven Lavalle、Dan Klein、Lenny Pitt、
Mehran Saham、Tuomas Sandholm、Michael-David Sasson、Manuela Veloso、Dan Weld

不管名单中各位的贡献是大是小，均对本报告的成文提供了助力。我们对他们表示感谢，并且希望能有更大范围的社区参与到人工智能相关的对话中来。

附录 A：数据描述与收集方法

A1. 论文发表数量

主要数据源和数据集

Elsevier 的学术发表数据库 Scopus，其中索引了近 7000 万份文档（69,794,685）。

参见有关 Scopus 的更多信息：https://www.elsevier.com/solutions/scopus

收集的数据的定义

每年被 Scopus 目录索引在“计算机科学”学科领域中并且还索引了关键词「人工智能」的论文的数量。这里给出一些参考信息：

整个 Scopus 数据库中计算机科学领域的论文的数量超过 200,000（200,237）篇都索引了关键词「人工智能」。
Scopus 数据库中「计算机科学」学科领域下包含了近 500 万（4,868,421）篇论文。

上面引用的这两个数字以及 Scopus 数据库中出版物的总数都是 2017 年 11 月的记录。

数据收集过程

我们向 Scopus 已发表学术论文数据库进行了查询，请求计数与人工智能相关的论文的数量、计算机科学学科领域中的论文的数量以及数据库中论文的总数。比如说，用于获取 2000 年相关论文的数量的查询是：

查询人工智能：

```
title-abs-key(artificial intelligence)
AND SUBJAREA(COMP)
AND PUBYEAR AFT 1999
AND PUBYEAR BEF 2001
```

查询计算机科学：

```
SUBJAREA(COMP)
AND PUBYEAR AFT 1999
AND PUBYEAR BEF 2001
```

查询整个 Scopus：

PUBYEAR AFT 1999 AND PUBYEAR BEF 2001

我们查询了从 1996 年到 2016 年中每一年的数据。

Elsevier 也提供了 Scopus API 的访问权限，让从 Scopus 提取数据的过程实现了自动化。

要了解更多有关 Scopus 查询语言的信息，参阅 Scopus Field Specification：
https://api.elsevier.com/content/search/fields/scopus
要了解更多有关 Elsevier 的 API 的信息，参阅 Elsevier API 文档：
https://dev.elsevier.com/api_docs.html
要了解更多有关 Scopus 搜索 API 的信息，参阅搜索 API 文档：
https://api.elsevier.com/documentation/SCOPUSSearchAPI.wadl

细节

Scopus 系统会追溯式地更新。因此，在给定一个查询时，Scopus 系统返回的论文的数量可能会随时间增大。比如说，查询“SUBJAREA (COMP) AND PUBYEAR BEF 2000”返回的论文数量结果可能会随 Scopus 覆盖得越来越全面而越来越大。

Elsevier 团队的成员评论说 1995 年之后的论文发表数据是最可靠的，而且那之后他们的系统的数据处理也更为标准化。因此，我们仅从 Scopus 源收集了 1996 年及之后的论文发表数据。Scopus 有很广泛的数据源。他们的索引技术和查询语言也让确认与特定主题相关的论文变得轻松简单。有关论文发表的其它可用数据源还包括 Web of Knowledge、微软学术、DBLP、CiteSeerX、谷歌学术。尽管每个系统中收录的论文的总数和具体来源各有不同，但我们预计论文发表的增长趋势在各个数据库中应该会基本保持一致。

A2. 课程选修

主要数据源和数据集

大学的课程选修记录。选修数据是从以下大学收集的：

加州大学伯克利分校、卡耐基梅隆大学、乔治亚理工学院、伊利诺伊大学香槟分校、麻省理工学院、斯坦福大学和华盛顿大学。

收集的数据的定义

在每个学年中所选大学的代表性本科人工智能和机器学习课程的选修学生人数。「学年」始于每年的秋季。

收集过程

我们联系了每所大学的代表，他们帮助确定了人工智能和机器学习课程并从学校记录中收集了选修数据。

细节

本报告选择的是入门级人工智能和机器学习课程，尽管很多大学都提供了入门课程以上的额外课程，但入门课程在各个大学中更为一致并且也容易区分。

很多大学的学生参加入门级人工智能和机器学习课程的需求超过了它们的支持能力。我们的数据仅代表这些大学有能力提供的课程。

特定年份之间部分尤其剧烈的上升下降是行政问题导致的结果，而不是因为学生。比如，我们的斯坦福大学联系人这样解释了 2015 年和 2016 年之间机器学习课程选修人数的下降：

「机器学习班通常每年教授一次。但在 2015-16 学年则教授了两次（秋季一次，春季又一次）。在秋季课程选修已经完成之后，春季课程才被列出来。所以我可以想见（根据直觉，而非实际数据）2015-16 学年的春季课程会吸引一些原本可能会参加 2016-17 学年秋季课程的学生，进而导致 2016-17 学年秋季课程选修人数下降。所以实际上 2015-16 和 2016-17 学年之间的选修人数应该是平滑变化的。我不相信学生对机器学习课程的兴趣真会有下降。」

A3. 学术会议出席情况

主要数据源和数据集

主办人工智能相关会议的组织结构的记录。数据收集自以下会议：

AAAI、AAMAS、ACL、CP、CVPR、ECAI、ICAPS、ICRA、ICLR、ICML、IJCAI、IROS、KR、NIPS、UAI

收集的数据的定义

与人工智能及其子领域相关的所选学术会议的参会人数。我们将 2016 年参会人数超过 1000 人的会议定义为「大型会议」(large conferences)，将 2016 年参会人数少于 1000 人的会议定义为「小型会议」(small conferences)。

数据收集过程

人工智能指数团队与会议组织方以及赞助机构的领导者合作收集了每个会议的参会数据。

数据的细节

并非所有会议组织团队都有全部参会数据。很多领导团队都指出某些年的参会数据丢失了，有的也仅能给出一个大概数据。根据我们的评估，似乎可以接受由多个领导者提供的估计数据并认为其是准确的。

并非所有会议都是年度举办的，有些会议跳过了一些年份。

A4. 人工智能领域创业公司

主要数据源和数据集

Crunchbase：https://about.crunchbase.com/about-us/
风险投资公司综合数据库 VentureSource：
https://www.dowjones.com/products/venturesource-2/
风险投资支持的私营公司的指数提供商 Sand Hill Econometrics：
http://www.sandhillecon.com/

收集的数据的定义

每年被认定为正在开发或部署人工智能系统的活跃创业公司的数量。

数据收集过程

我们首先收集了 Crunchbase 中具有与人工智能相关的类别标签的所有组织机构的列表。为了得到类别标签的集合，我们审阅了 Crunchbase 中所有类别的集合并从中选择了我们认为处于人工智能技术领域内的集合，如下列出。我们通过 Crunchbase 向我们提供的 Crunchbase API 获取了这些类别标签和组织机构列表。

然后来自 Crunchbase 的组织结构列表再与 VentureSource 数据库中的所有风险投资支持的公司的列表进行了交叉参照。只要在 Crunchbase 列表中的风险投资支持的公司得到了 VentureSource 数据库的确认，我们就将其包含了进来。VentureSource 也为每家公司关联了关键词。在 VentureSource 中所有关联了「人工智能」或「机器学习」关键词的公司也被包含进了相关创业公司列表中。

更多有关 Crunchbase API 的信息：https://data.crunchbase.com/docs
参看 Crunchbase Categories 列表：https://www.crunchbase.com/search/categories

与 VentureSource 产品的所有交互都是由 Sand Hill Econometrics 完成的。

细节

用于识别人工智能公司的 Crunchbase 「类别」标签列表：
人工智能、机器学习、自然语言处理、计算机视觉、面部识别、图像识别、语音识别、语义搜索、语义网、文本分析、虚拟助手、视觉搜索、预测分析、智能系统。

确定某家公司是否与人工智能相关并没有什么简单直接的方法。我们的探索目前以机器学习技术为重点。

A5. 人工智能领域风险投资

主要数据源和数据集

Crunchbase
风险投资公司综合数据库 VentureSource
风险投资支持的私营公司的指数提供商 Sand Hill Econometrics

收集的数据的定义

本报告给出的数据是风险投资者每年投资给创业公司的资金的数量，并且在这些创业公司的业务中，人工智能应该在一些关键功能上发挥了重要作用。

数据收集过程

这一部分使用了「人工智能领域创业公司」一节中的公司集合。然后我们从 VentureSource 中检索了这个公司集合的相关投资数据并将其聚合成了年度的投资数据。

与 VentureSource 产品的所有交互都是由 Sand Hill Econometrics 完成的。

A6. 工作机会

主要数据源和数据集

Indeed.com
Monster.com

收集的数据的定义

Indeed.com 数据代表了每个国家对人工智能技能需求的岗位的份额比重，数据归一化到了 2013 年 1 月的份额百分比。

Monster.com 数据代表了随时间变化的人工智能相关工作岗位需求的绝对数量，并且还根据工作岗位所需的具体技能将其分成了人工智能子领域。注意分解后的工作岗位可能有所重叠。比如，需要机器学习技能的工作也可能还需要自然语言处理技能。这样的工作岗位会在分解图中计算两次。

数据收集过程

我们与 Indeed 和 Monster 团队直接合作得到了这些数据。Indeed.com 和 Monster.com 使用了不同的处理方式来识别人工智能相关工作并且提供了不同类型的有关人工智能工作岗位增长的数据。

Indeed.com 首先确定了在相关工作中至少 50% 都列出了人工智能相关关键词的一系列工作。其所用的关键词为：
人工智能、机器学习、自然语言处理。

根据数据，在描述中具有其它人工智能关键词的工作岗位中，自然语言处理与其中超过 90% 有关。获得了这些工作岗位之后，Indeed 检查了每个国家中人工智能相关岗位数占总岗位数的百分比。他们追踪了这个百分比随时间的变化并将数据返回给了我们，然后我们根据 2013 年的值对数据进行了归一化。

Monster 使用 CEB 的 TalentNeuron 工具提供的数据确定了美国在 2015 年、2016 年和 2017 年（截至 11 月 10 日）的需要人工智能技能的岗位招聘数量。为了进行分解，他们还使用该工具确定了需要「人工智能」加上「计算机视觉」等其它技能关键词的岗位招聘数量。

A7. 自动化及机器人应用

主要数据源和数据集

由国际机器人联合会发布的年度《世界机器人报告(World Robotics Report)》。

收集的数据的定义

这部分给出的数据是每年北美进口的以及国际进口的工业机器人的数量。工业机器人按照 ISO 8373:2012 标准定义。

数据收集过程

国际机器人联合会的年度《世界机器人报告》包含北美和国际的机器人进口数量。我们从这些报告中提取了自 2000 年以来的出货数据。

细节

目前还不清楚这些机器人中有多大比例运行了可以被归类为「人工智能」的软件，而且也不清楚人工智能的发展对工业机器人的应用有多大贡献。

A8. GitHub 项目统计

主要数据源和数据集

GitHub Archive：https://www.githubarchive.org

BigQuery 上的 GitHub Archive：
https://bigquery.cloud.google.com/table/githubarchive:day.20150101

收集的数据的定义

各种 GitHub 库随时间被加星(Star)的数量。这些库包括：

apache/incubator-mxnet、BVLC/cafe、cafe2/cafe2、dmlc/mxnet、fchollet/keras、Microsoft/CNTK、pytorch/pytorch、scikit-learn/scikit-learn、tensorflow/tensorflow、Theano/Theano

收集过程

GitHub 归档数据存储在 Google BigQuery 上。我们通过与 Google BigQuery 交互而计数了每个相关库的「WatchEvents」数量。收集 2016 年的数据的代码样本如下：

```
SELECT
project,
YEAR(star_date) as yearly,
MONTH(star_date) as monthly,
SUM(daily_stars) as monthly_stars

FROM (
SELECT
repo.name as project,
DATE(created_at) as star_date,
COUNT(*) as daily_stars
FROM
TABLE_DATE_RANGE(
[githubarchive:day.],
TIMESTAMP("20160101"),
TIMESTAMP("20161231"))
WHERE
repo.name IN (
"tensorflow/tensorflow",
"fchollet/keras",
"apache/incubator-mxnet",
"scikit-learn/scikit-learn",
"cafe2/cafe2",
"pytorch/pytorch",
"Microsoft/CNTK",
"Theano/Theano",
"dmlc/mxnet",
"BVLC/cafe")
AND type = 'WatchEvent'

GROUP BY project, star_date
)
GROUP BY project, yearly, monthly
ORDER BY project, yearly, monthly
```

细节

GitHub Archive 目前还没提供措施来计数用户移除库的 Star 。因此，本报告给出的 Star 数量稍微有所高估。与 GitHub 上的库的实际 Star 比较表明这个数量是相当接近的，而且趋势仍然是一样的。还有一些检索 GitHub Star 数的其它方法。我们使用 star-history 工具(https://github.com/timqian/star-history)抽样检查了我们的结果。

GitHub 项目的 Fork 也值得调研。我们发现库的 Star 和 Fork 趋势是基本一致的。但是，如果你有兴趣了解具体的 Fork 数据，你可以在我们的网站 aiindex.org 上找到数据，也可自己使用 BigQuery 代码查询（将上面的 type = 'WatchEvent' 改成 type = 'ForkEvent'）。

A9. 舆论倾向

主要数据源和数据集

TrendKite：https://www.trendkite.com

收集的数据的定义

TrendKite 服务索引了一般的媒体文章并且使用了一种情绪分析分类器将这些文章分成了「正面」、「负面」和「中性」三类。我们给出了被分类为「正面」和「负面」的文章的比例（剩下的都归类为「中性」）。

收集过程

我们使用了下面的查询来确定人工智能有关文章。我们对其进行了调整，移除了一个引入了数量不成比例的具有负面情绪的不相关文章的来源。

查询
"Artificial Intelligence"
AND NOT "MarketIntelligenceCenter.com's"

NOT site_urls_ll:(
"individual.com"
OR "MarketIntelligenceCenter.com")

TrendKite 的文章来源有很多，但也提供了可使搜索更相关的过滤器。我们使用了这些功能的过滤器：

仅包含英语文章
移除新闻稿
移除财经新闻
移除讣告

我们希望通过这些数据分享公众对人工智能的兴趣程度以及公众对人工智能的了解程度。这个过滤器帮助我们简化了我们的信号的来源。

A10. 物体检测

主要数据源和数据集

2010 - 2017 年的 LSVRC ImageNet 竞赛：http://image-net.org/challenges/LSVRC
ImageNet 数据集：http://image-net.org

收集的数据的定义

自 2010 年以来在 LSVRC ImageNet 竞赛中的物体检测挑战赛上获胜团队的准确度结果。可在 LSVRC 网站上查看这些指标的定义：

数据收集过程

我们从托管在 ImageNet 网站上的每个 LSVRC 竞赛中收集了排行榜中的比赛数据。

细节

ImageNet 竞赛在 2017 年已经终止。继续调查在 ILSVRC 测试集上得到新最佳结果的文献也许是可能的，但我们很可能需要确定和跟踪新的基准。

人类水平的表现是由 Russakovsky 等人在他们 2015 年的论文中估计的：https://arxiv.org/pdf/1409.0575

A11. 视觉问答

主要数据源和数据集

Arxiv（用于文献审查）：http://arxiv.org

VQA 数据集：http://visualqa.org

VQA 数据集中包含图像、关于这些图像的内容的问题以及 10 个人类给出的这些问题的答案。

收集的数据的定义

收集到的数据表示了每个人工智能系统得出这些关于图像的问题的开放式答案的准确度（而不是得出有关图像的多个选择题的答案）。

准确度标准是根据原始的 VQA 论文定义的：https://arxiv.org/pdf/1505.00468。我们收集了当学术论文报告实现了当时的最佳结果时的准确度。

数据收集过程

我们执行了一次文献审查来确定在 2016 年到 2017 年间在 VQA 1.0 数据集上得到了新的当前最佳结果的论文。

数据的细节

在执行文献审查过程中，我们很有可能错过了一些会稍微改变新的当前最佳成果时间线的结果。我们也考虑了组合式方法，而不只是单个模型。

随着 ImageNet 不再继续充当推动视觉任务的竞赛，我们决定调查视觉问答领域的进展全景。但是，不久的将来可能会有替代 ImageNet 的基准出现；在一个更具主导性的基准出现之前，我们可能还需要继续测量中断的进展，就像我们使用 VQA 做的那样。

VQA 1.0 在其发布之后不久就退休了，以支持 VQA 2.0。VQA 2.0 有很多改进，包括增加了更多数据以去除数据集各方面的偏差。

A12. 解析

主要数据源和数据集

Penn Treebank：https://catalog.ldc.upenn.edu/ldc99t42

Penn Treebank 中《华尔街日报》这部分数据集的每个句子都标注了一个基于成分(constituency)的解析树。这个数据集的第 23 节已经变成了自动解析器研究的主要测试集。

收集的数据的定义

自动解析器的评估方式是比较自动生成的解析的成分与来自测试集的黄金解析的成分。在本报告中，生成的成分的精度和回调结合成了 F1 分数。我们报告的是解析器在 Penn Treebank 的 WSJ 部分的第 23 节的句子上得到的 F1 分数。我们报告的这些分数针对的是长度小于 40 词的句子，并且是在其中每个句子都可用的整个数据集上。参阅解析树的维基百科页面可了解更多有关基于成分的解析树的信息：https://en.wikipedia.org/wiki/Parse_tree#Constituency-based_parse_trees

数据收集过程

我们进行了一次文献审查来确定什么时候出现了提升了自动解析的当前最佳的解析器。我们收集了 1995 年以来我们确认的解析器的 F1 分数。我们也考虑了组合式方法，而不只是单个模型。

数据的细节

在自动解析研究的早些时候，由于计算和方法上的原因，对解析器的评估通常是在少于 40 词和少于 100 词的句子上评估的。我们记录了在长度少于 40 词的句子以及语料库中所有句子（当可用时）上评估的系统的 F1 分数。

A13. 机器翻译

主要数据源和数据集

机器翻译会议/研讨会(WMT)新闻翻译任务 EuroMatrix：http://matrix.statmt.org/matrix

机器翻译会议(Conference on Machine Translation)是年度举办会议，其又分离出了一个年度机器翻译研讨会(Workshop on Machine Translation)。WMT 每年都会举办一个新闻翻译任务比赛并且会提供新的训练和测试数据集。参与团队可以提交他们开发的翻译系统来参加这个翻译任务。

收集的数据的定义

WMT 所用的主要指标的目的是评比互相竞争的项目而并不允许不同年份的比较。这也是非常耗费心力的工作。我们最终选择了 BLEU，这是一种将系统翻译与许多人类生成的翻译进行粗略比较的自动方法，这是一种修改的精度版本，结果在 0 到 1 之间，越高越好。也可以计算机器翻译系统在翻译对语料库上的平均 BLEU 分数。我们记录了每一年提交给当年的英语翻德语新闻翻译任务的系统所实现的最高平均 BLEU 分数。请在下面查看 WMT 翻译任务和 BLEU 指标的细节。

数据收集过程

EuroMatrix 记录了自 2006 年以来新闻翻译任务中英语翻德语和德语翻英语语言对上的 BLEU 分数。我们选择了每年系统得到的最高 BLEU 分数，具体来说使用了 BLEU (11b)，其定义了一种用于实现句子 token 化的协议。如有可能，我们选择了在两个语言对上都具有高排名的 BLEU 分数的系统的分数作为当年的代表。

11b token 化的实现请参阅：ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13.pl

数据的细节

BLEU 可以自动计算，而且事实也已经证明这个分数与人类对翻译质量的判断有关。但是，这个指标不能跨语料库使用，因为这可能导致不同系统之间 BLEU 分数比较错误。尽管从图表上看结果呈上升趋势，但我们看到这一指标在 2017 年存在缺陷，这一年的 BLEU 分数比 2016 年显著更低（尽管 2017 年的分数仍然高于 2015 年的分数）。机器翻译系统 2017 年的表现不太可能比 2016 年还差，但这里给出的评估方案并不是完美的。

但是在更大的时间周期上看看这个趋势，BLEU 分数仍然还是能说明机器学习领域的进展。

A14. 问答

主要数据源和数据集

斯坦福问答数据集： http://stanford-qa.com

斯坦福问答数据集(SQuAD)中包含了超过 500 篇文章以及与这些文章相关的 100,000 个问答对。给定关于一篇文章的内容的一个问题，任务目标是在这篇文章中找到答案。

收集的数据的定义

在这个数据集上所选择的评估指标是 Exact Match(EM)，即由系统生成的答案和测试集中答案确切匹配的百分比。报告中给出的数据是问答系统在 SQuAD 数据集上随时间变化的当前最佳 EM 分数。

数据收集过程

我们从托管在 SQuAD 网站上的排行榜中收集了结果。

数据的细节

SQuAD 数据集中的所有答案都是从相关文章中直接引用的。因此，系统的工作实际上是确认文章中的哪一部分包含了所给问题的答案。

尽管 SQuAD 数据集易于跟踪，但目前仍不清楚该数据集还将被使用多长时间。自 2016 年 6 月 SQuAD 发布以来，其上的分数一直在迅速增长。

人类在该数据集上的 Exact Match 分数据称是 82.304。

A15. 语音识别

主要数据源和数据集

Switchboard Hub5'00 数据集（语音：https://catalog.ldc.upenn.edu/LDC2002S09；转录：https://catalog.ldc.upenn.edu/LDC2002T43）
EFF AI Progress Metrics：https://www.eff.org/files/AI-progress-metrics.html

收集的数据的定义

训练后的语音识别系统在标准的 Switchboard Hub5'00 数据集上随时间变化的当前最佳的词错率(WER)。WER 是将转录文本映射到黄金标准上所得到的错误量（包括错词、少词和多词）。我们绘制了词准确度（即 1-WER）图来表示相关的进展。

数据收集过程

电子前线基金会之前执行了一次文献审查，提取了语音识别系统在 Hub5'00 数据集上的表现。我们在本报告中直接给出了这些结果。

数据的细节

Switchboard 数据集已经得到了很长时间的应用。有人担心我们的人工智能系统可能会严重过拟合这个特定的数据集，在这一数据集上的进一步进展可能无法代表这一领域的总体进展。

人类在 Switchboard Hub5'00 数据集上的 WER 表现究竟如何最近还有一些不一致的意见。5.1% 和 5.9% 的结果都有，甚至还有低于 5% 的报告。在这份报告中，我们选择使用 5.1% 作为人类水平的标准。

A16. 定理证明

主要数据源和数据集

用于定理证明器的数千个问题(TPTP/Thousands of Problems for Theorem Provers)：
http://tptp.org/

TPTP 是一个定理证明问题实例的大型集合。

收集的数据的定义

自动定理证明(ATP)社区已经开发出了一种确定使用当前 ATP 技术解决给定问题实例的难度的方法。我们记录了 TPTP 问题的一个子集随时间的平均难度变化。这里所选择的 TPTP 问题是那些自 TPTP v5.0.0（2010 年）以来就再没更新过的问题。我们绘制了「1−难度」的图表，并将其命名为了「可解决度」(Tractability)，以保持与整个图表的持续增长方向一致。

参阅 TPTP 维护者的论文了解 TPTP 问题难度的定义：
http://www.sciencedirect.com/science/article/pii/S0004370201001138
查看如何计算给定 TPTP 问题实例的难度的可视化示例：
http://www.cs.miami.edu/~tptp/Seminars/Evaluate/Summary.html

数据收集过程

TPTP 数据集包含了每个问题每个版本的难度。我们写了一个脚本来从 TPTP 问题实例的所需子集中提取这些难度以及计算相关问题的平均难度随时间的变化。该脚本将会在 AI Index 网站 aiindex.org 上公开。

数据的细节

自动定理证明社区所用的 TPTP 问题难度的定义有些古怪。它取决于可用的 ATP 系统集合。如果有段时间创造出了很多有用的 ATP 系统但却不能解决这个问题，那该问题的难度也可能会随时间变得更大。

参阅 TPTP v6.4.0 数据集的维护者对该数据集的概述：
https://miami.pure.elsevier.com/en/publications/the-tptp-problem-library-and-associated-infrastructure-from-cnf-t

A17. SAT 求解

主要数据源和数据集

SAT 竞赛： https://baldur.iti.kit.edu/sat-competition-2017/
SAT 求解器表现数据

SAT 竞赛对有实际问题形式的问题实例有“行业”跟踪。Holger Hoos 和 Kevin Leyton-Brown 选取了 69 个求解器和 1076 个问题实例，这些是自 2007 年以来的竞赛的一部分；然后他们在同样的硬件上为每个问题都运行了所有的求解器。

收集的数据的定义

对于每一年的数据，我们取了那一年提交的求解器完成的问题的平均百分比（自 2007 年以来竞赛中的所有问题）以及由最佳求解器解决的问题所占的百分比。

收集过程

Hoos 和 Leyton-Brown 收集了每个求解器在每个问题上的表现。我们直接将数据进行了聚合，然后得出了前面描述的分数。

细节

这个指标会随处理器速度而提升，尽管 Hoos 和 Leyton-Brown 已经通过在同样的硬件运行而纠正了这一问题。

尽管这个指标本质上是追踪 SAT 求解器随时间的效率变化，但这个指标并没有量化随时间变化的新 SAT 求解器的贡献的新颖性。换句话说，这个指标也可能仅仅表示了工程上的进展（这仍然很重要），而不是算法上的突破。我们正在评估能更好地量化新创造的 SAT 求解器的基础贡献的方法。


aiindex.org