figure eight

2018年数据科学家报告

天善智能

编译

呆鸟 译

天善智能

简介

近年来，Figure Eight一直在追踪数据科学的发展，自2015年发布上一版数据科学报告以来（那时我们还叫CrowdFlower），数据科学社区里发生了很多变化。机器学习技术蓬勃发展，需要越来越多的数据支持。

如今，互联网每天会产出100万亿字节以上的数据供数据科学与机器学习分析。因此，数据科学和机器学习也顺势成为领英上增长最快的工作岗位。

2015年以来出现的另一大趋势是数据科学社区比以往更加注重伦理问题，数据隐私问题越来越引人注目。随着人工智用于医学诊断、法律量刑等领域的决策，需要更加谨慎地论证这些伦理问题。

了解各领域从业者对前沿技术的想法十分重要。为此，我们调研了医护人员、神职人员及执法人员等500多位伦理专家。

本报告后面的内容，还将专门对比伦理专家与数据科学家的观点。

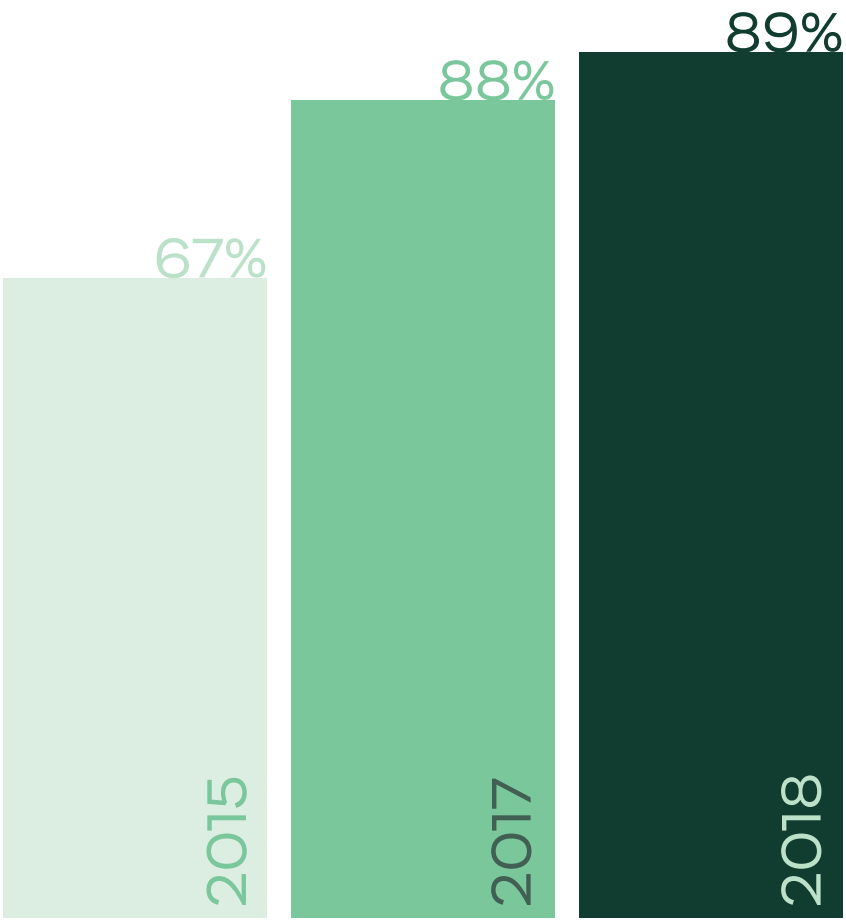
毋庸赘言，开始阅读本报告的调研结果吧。

数据科学家

喜欢

并热爱这份工作

认为当数据科学家幸福或非常幸福的占比



相信很多人都听过一句话，“干自己喜欢的事，还能挣到钱，就算成功”。假设这话说的没错，还真的很难找出比数据科学家更成功的职业。

几年来，我们一直在跟踪这个问题，并发现数据科学家非常热爱这一行，即便真正的数据科学家可能会质疑1%的增长不具备统计显著性。

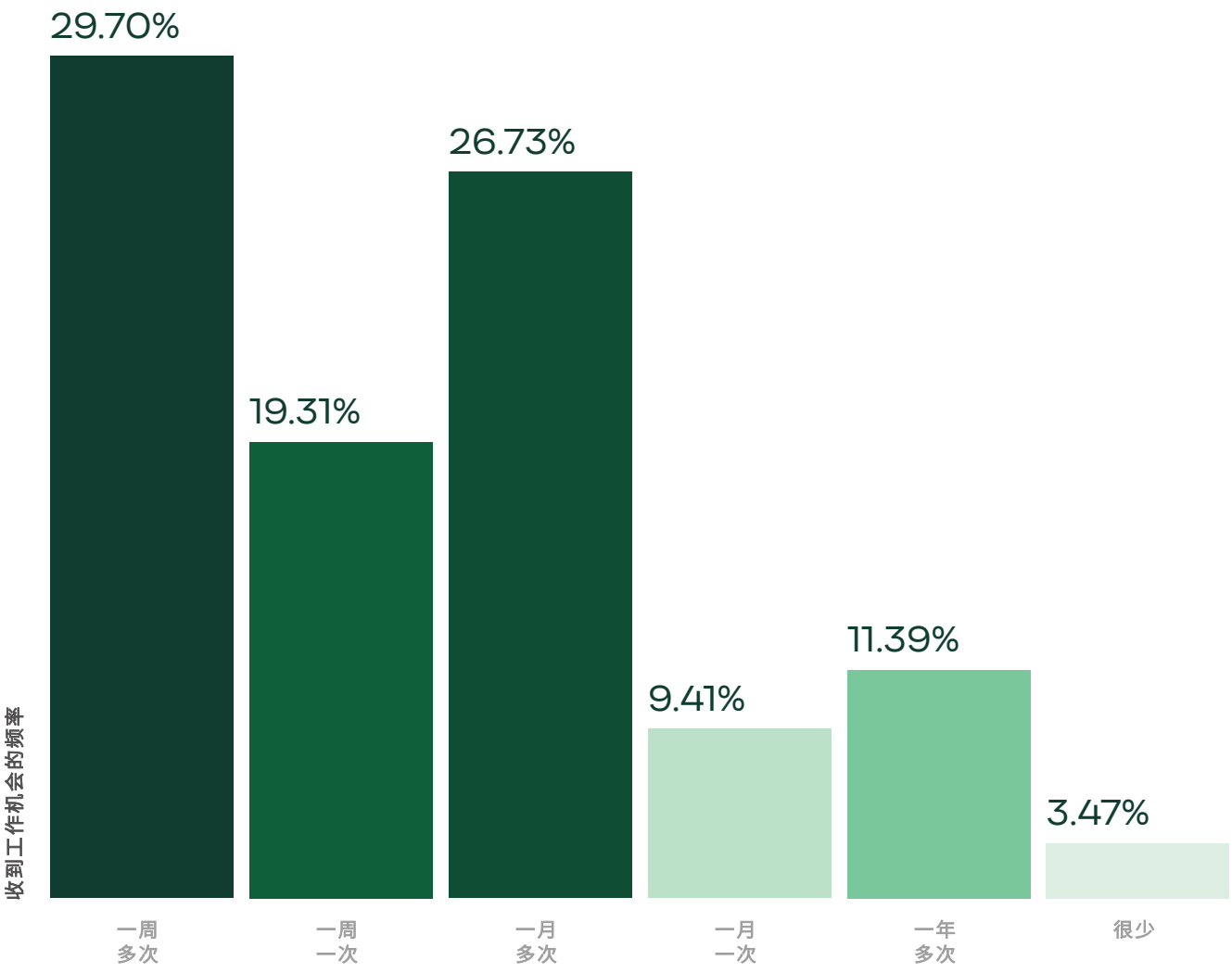
热爱数据科学？就别错过机会

这几年，数据与数据科学带来了很多热门话题，谷歌人工智能专家Peter Norvig曾提出著名的“数据非理性效果”理论，哈佛商业评论将数据科学称为“21世纪最性感的工作”，经济学家杂志甚至跳出来说“数据是新的石油”。

相信大多数人还记得大数据一夜之间就红遍全球了。

数据科学家的市场需求

收到工作机会的频率?



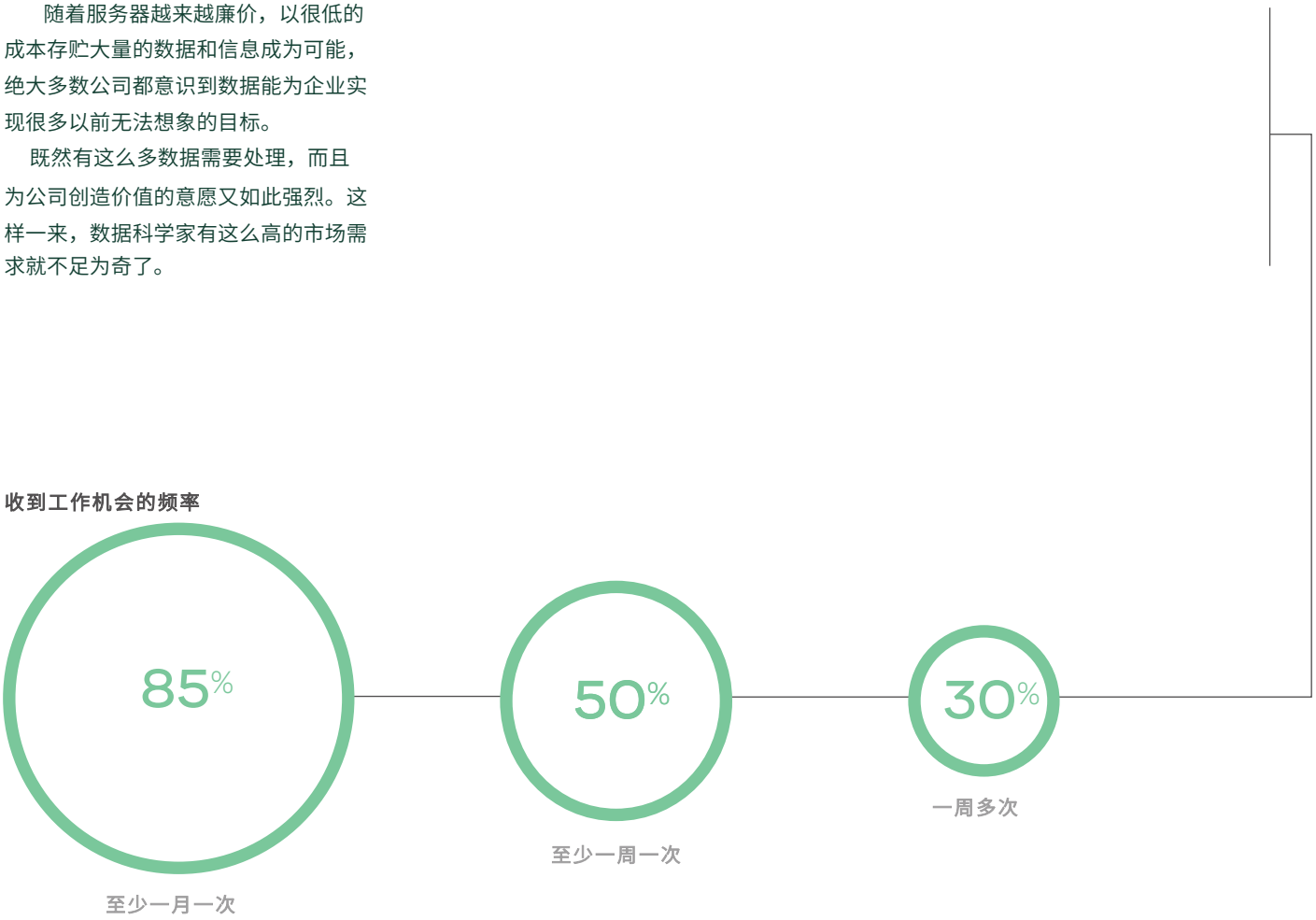
虽然，数据科学如今炙手可热，但要记住以前可不是这样。毕竟，仅仅在10多年前，大部分公司根本就不会跟踪并保存用户交互数据，但是如今，还是这些公司，他们会把认真采集这些数据，并将之作为企业的核心财富小心翼翼的看护起来。

随着服务器越来越廉价，以很低的成本存贮大量的数据和信息成为可能，绝大多数公司都意识到数据能为企业实现很多以前无法想象的目标。

既然有这么多数据需要处理，而且为公司创造价值的意愿又如此强烈。这样一来，数据科学家有这么高的市场需求就不足为奇了。

我们曾咨询数据科学家一般多久能收到一次新工作推荐，下图所示的数据说明了不少问题。大约50%的数据科学家每周都会收到一次工作机会，30%的数据科学家每周至少会收到多次工作机会，85%的数据科学家至少每个月会收到一次工作机会。

换句话说，精英数据科学家的市场需求很高。所以，如果你们公司有一名水平很高的数据科学家，一定要把他哄好，因为他还有很多选择。



什么拖了数据科学家的后腿，是数据，不是科学

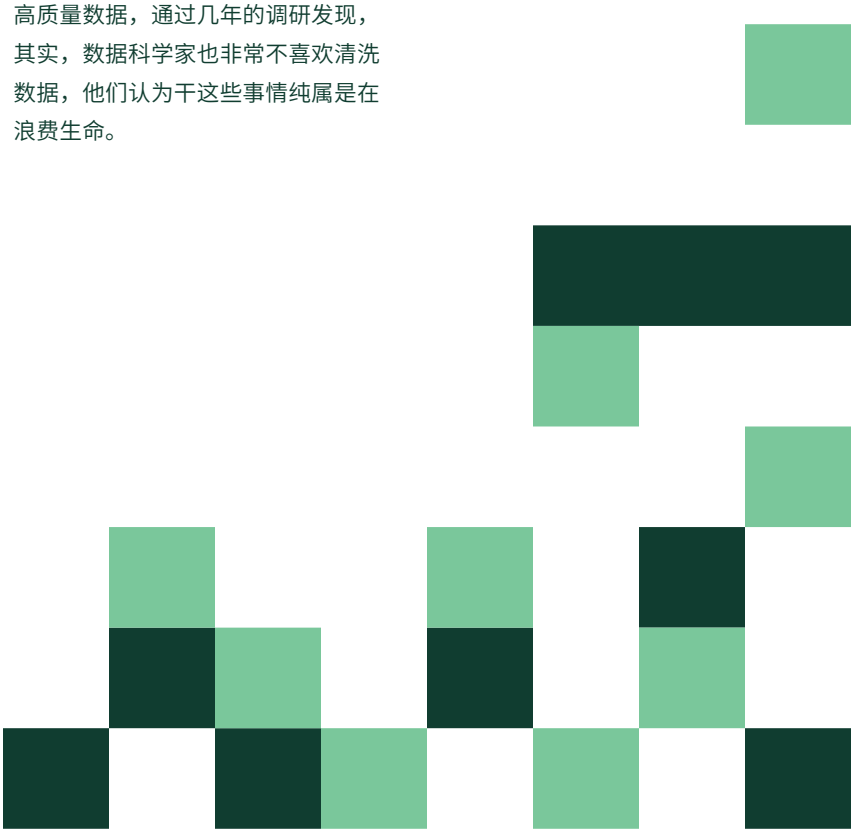
偷偷告诉大家一个关于数据科学家的小秘密，他们都非常贪得无厌。这不是说他们的坏话，实际上，很多数据科学家逢年过节都会寄给我们非常不错的礼物。但是，一旦涉及到数据，不管他们已经掌握了多少数据，还永远都觉得不够。

我们已经在数据科学社区里做了几年调研，这个问题依然是当前社区里最大的挑战。去年大约有50%的数据科学家会说，这是他们日常工作中最头疼的三件事之一，而到了今年这个数字已经增长到了55%，并被列为最头疼的事情。

数据专家非常清楚只有拥有大规模的高质量数据，才能构建精准的模型，并作出精明的决策。高质量数据越多，他们对所做的模型就越有信心。

公司能为数据科学家做的事就是提供数据，而机器学习团队拥有数据的质量会为机器学习的结果带来极大的区别，这一点是重中之重。

但是请记住，数据科学家需要的是高质量数据，通过几年的调研发现，其实，数据科学家也非常不喜欢清洗数据，他们认为干这些事情纯属是在浪费生命。

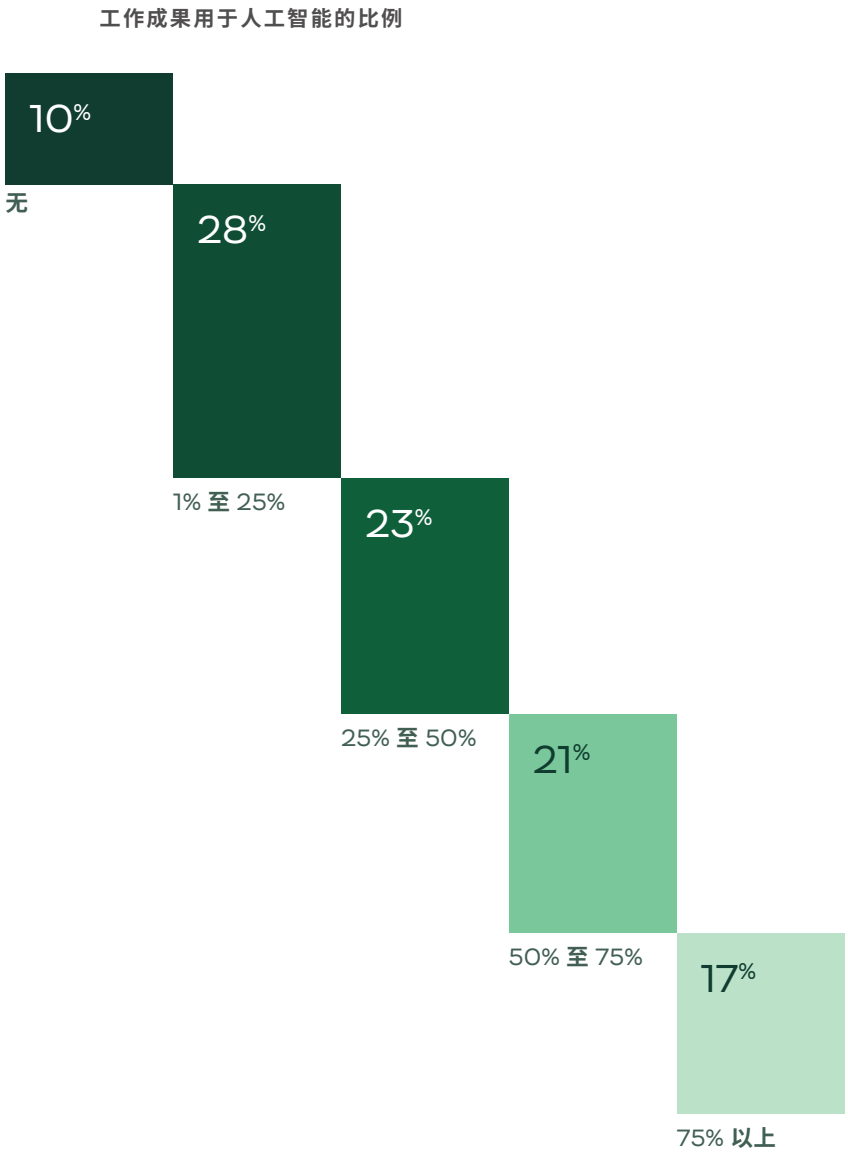


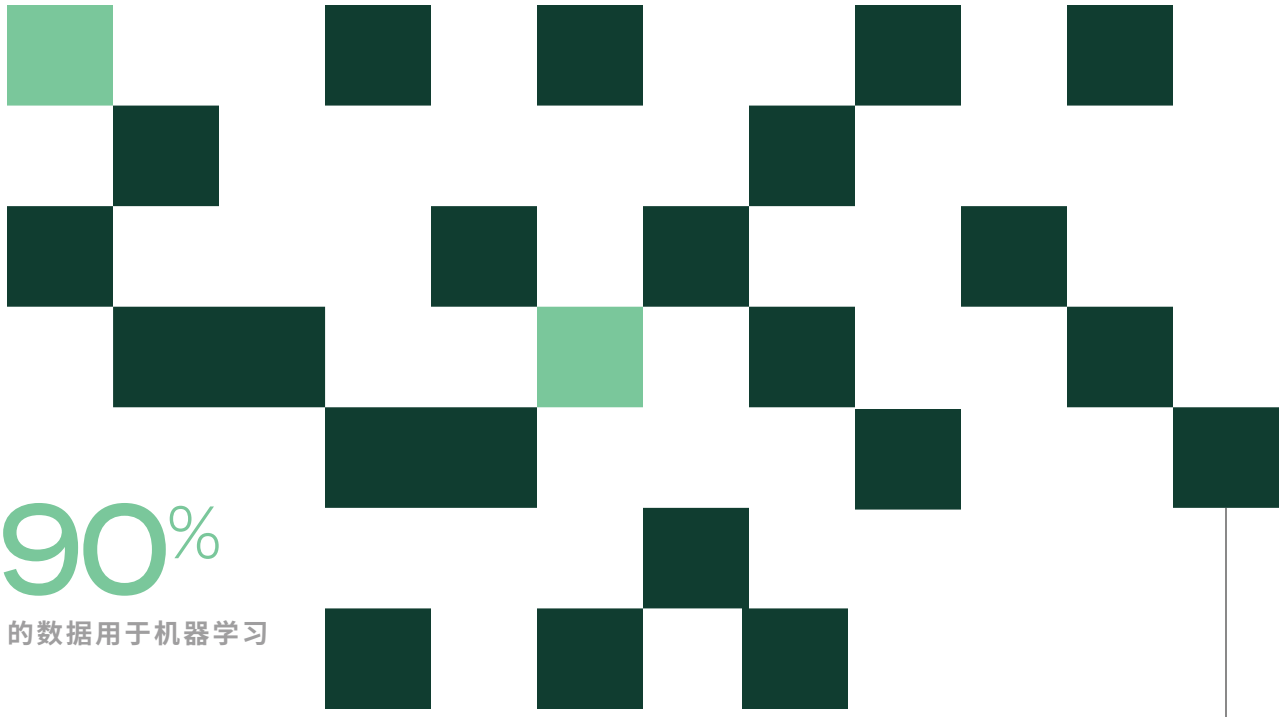
55%

的数据科学家说训练数据集的质量是他们最头疼的事情。

机器学习使用的数据

以前，我们从未问过数据科学家到底拿数据来干什么？但是，随着公司平台不断壮大，我们已经能够解开一些机器学习的神秘面纱，越来越多的数据直接从我们公司的平台传递给各种人工智能和机器学习的项目。然后我们就想，是不是应该问一下这些数据科学家，他们所做的工作到底有多少比例用于人工智能？





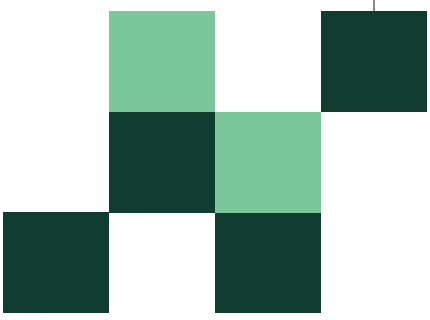
90%

的数据用于机器学习

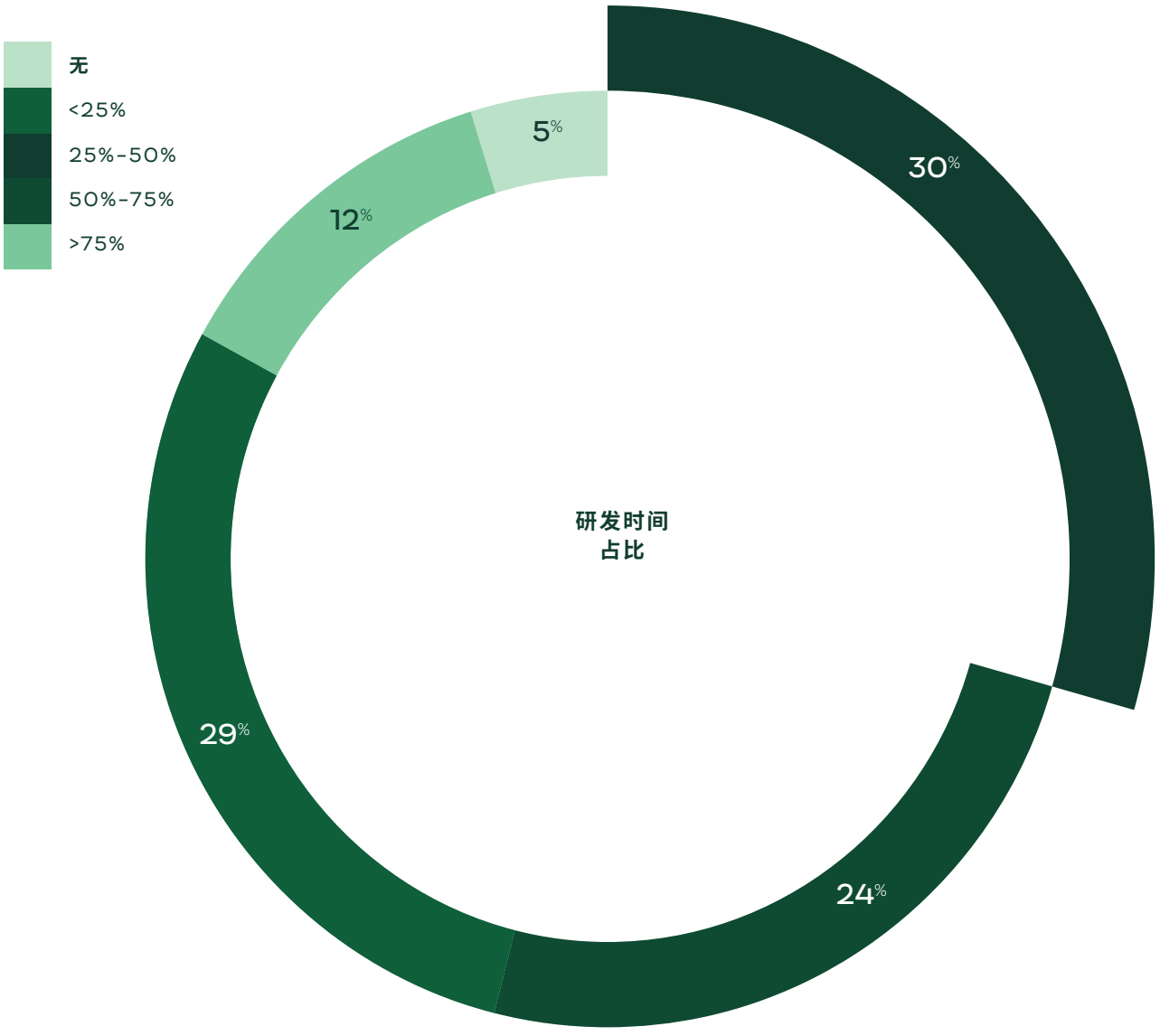
约有10%的数据科学家说他们的工作跟人工智能无关。不过，还有差不多40%的人说他们的工作和人工智能相关。

考虑到当前投资界对人工智能的投入非常之大，我们特别期待看到明年这个数字会变成什么样。不过，我们相信一定会变得越来越高。

数据科学家一般不需要干清洗日志这样的低级工作，基本上都是处理公司里最尖端的技术解决方案，难怪他们会觉得幸福。



多少时间研发？
多少时间开发？

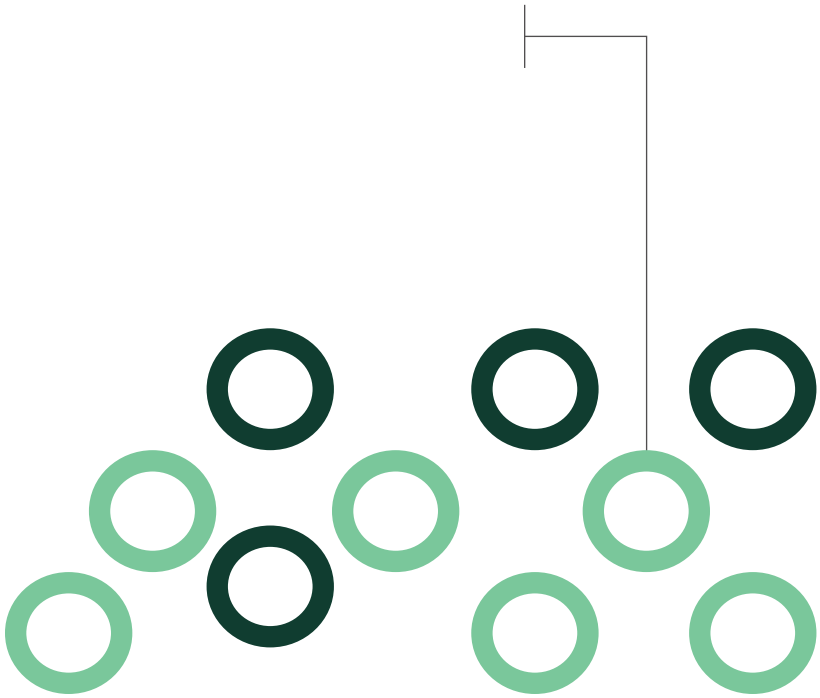


数据科学家使用哪些工具?

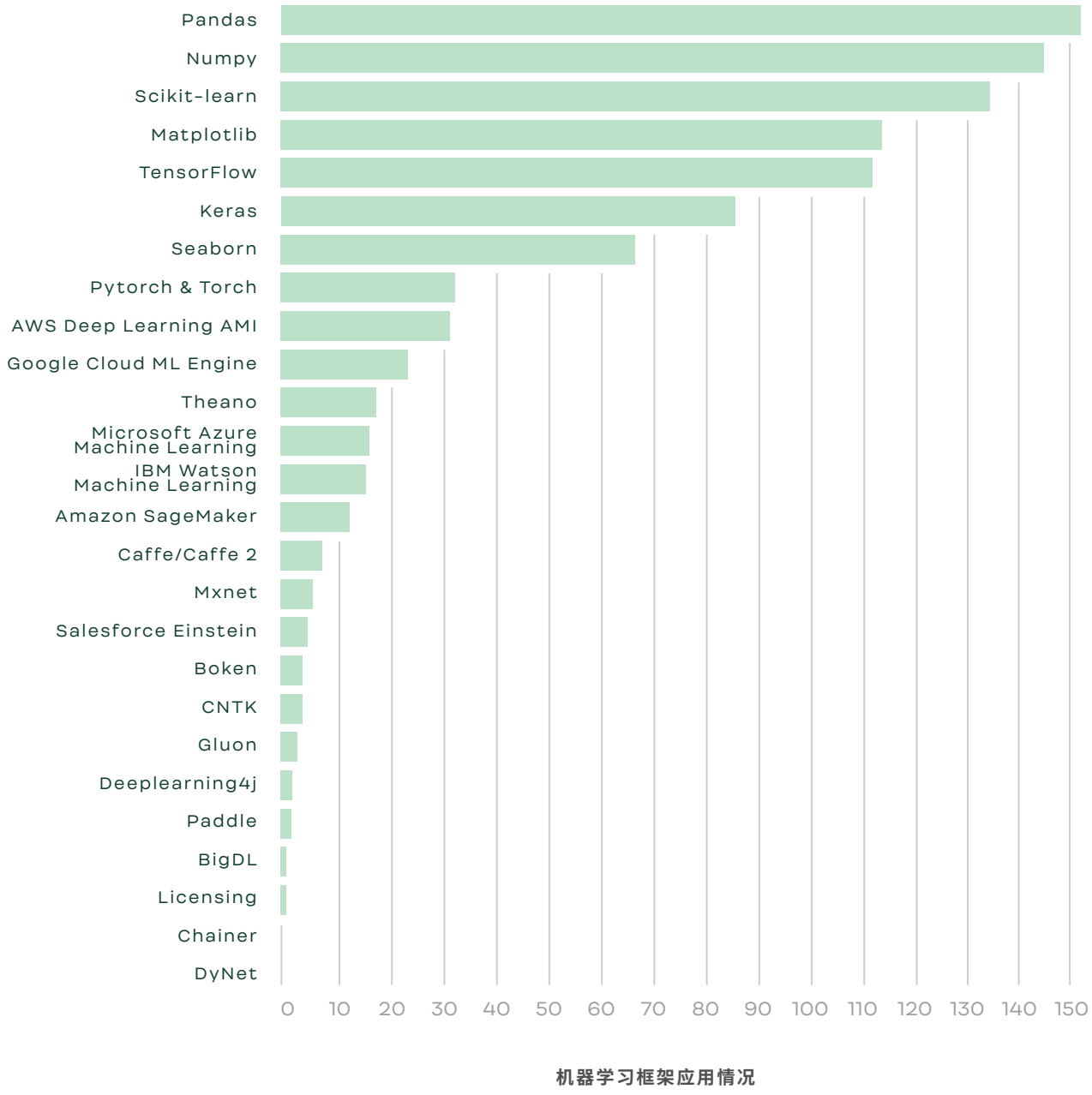
2015年，我们重点关注数据科学家使用什么工具。虽然，当时Excel还是处理数据的主流工具，但那时已经出现了很多数据工具和处理办法供数据科学家选择。实际上，Partially Derivative公司在一集叫“怪怪的数据科学”播客节目里就提到过这个问题。

他们的观点是数据科学是崭新的领域，没有哪种语言、工具或框架可以成为主流，即便现在也很难说哪种工具是最好的，数据科学家必须具备非凡的创造力，找出适于处理手头上数据科学项目的最佳工具和策略。

现在机器学习与数据科学当时的情况差不多，也没有大家公认可行的策略，但是有很多方法供人选择，用于处理以前难以解决的问题。不过，现在数据科学社区里大部分人（约61%）都选择了Python。但是，下面列出的常用Python库大多数并不是机器学习框架。



流行的机器学习框架

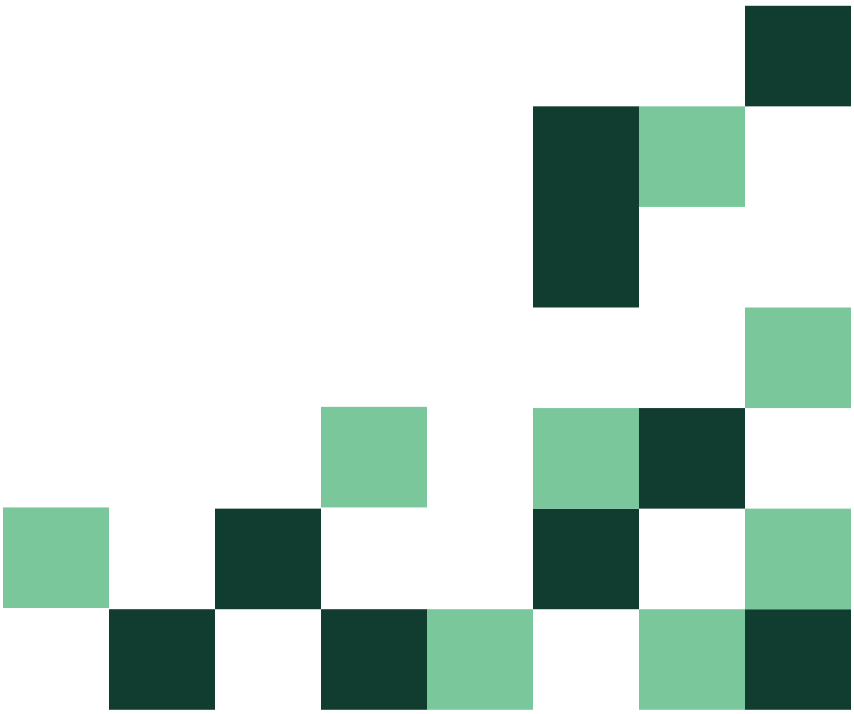


机器学习框架应用情况

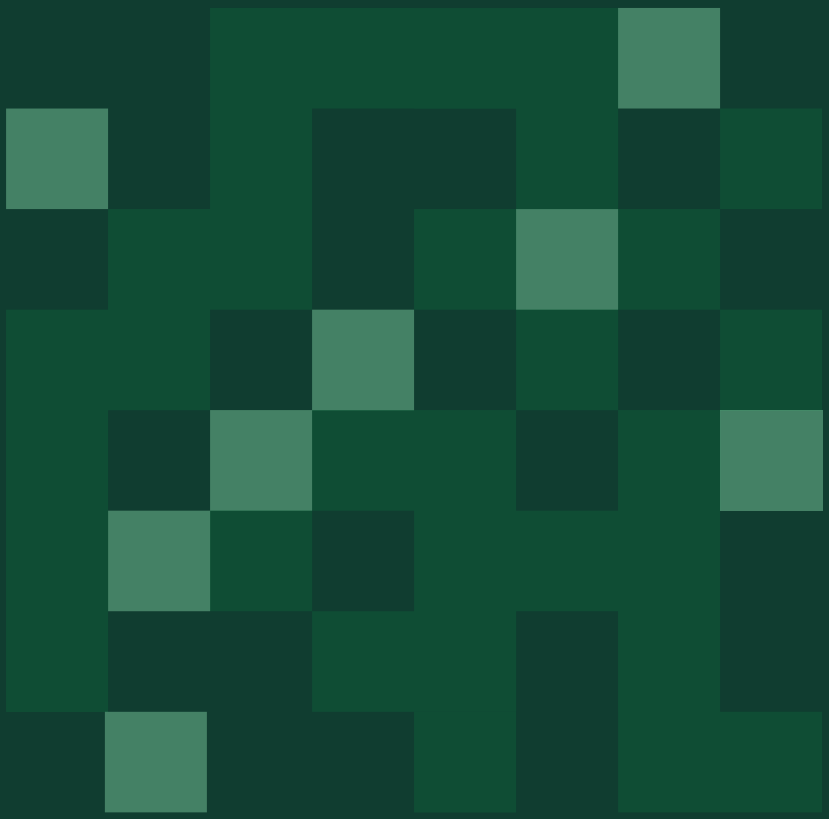
开源软件占这些工具和框架的主流。Pandas和NumPy已经推出了很长时间，此外，与之类似的还有Scikit-Learn和Matplotlib，也是老牌的Python库。

TensorFlow虽然是谷歌开发的，不过它也是开源软件。这里需要提醒的是，不能只根据数量进行判断，但另一方面，这些工具的用户确实很多，也说明了现在数据科学社区热捧开源和社区驱动的软件。

由于这些框架已经存在了很长时间，早期使用者已经对它们非常熟悉，如果新产品想取代这些老牌开源软件，恐怕还需要投入更多的时间、努力，并大力开展市场推广，比如增加更多的营销费用。



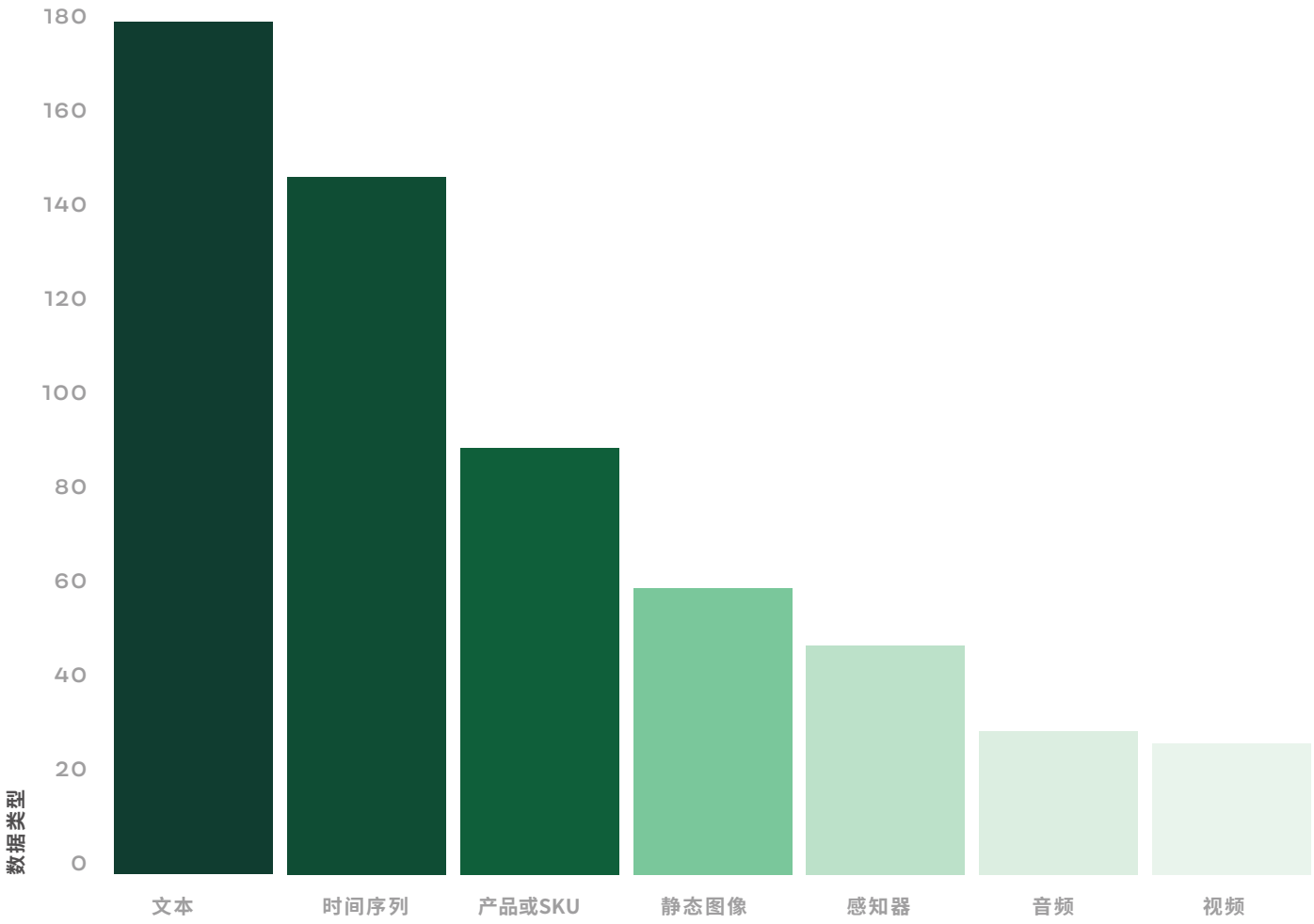
2018年
数据科学家处理哪些数据？



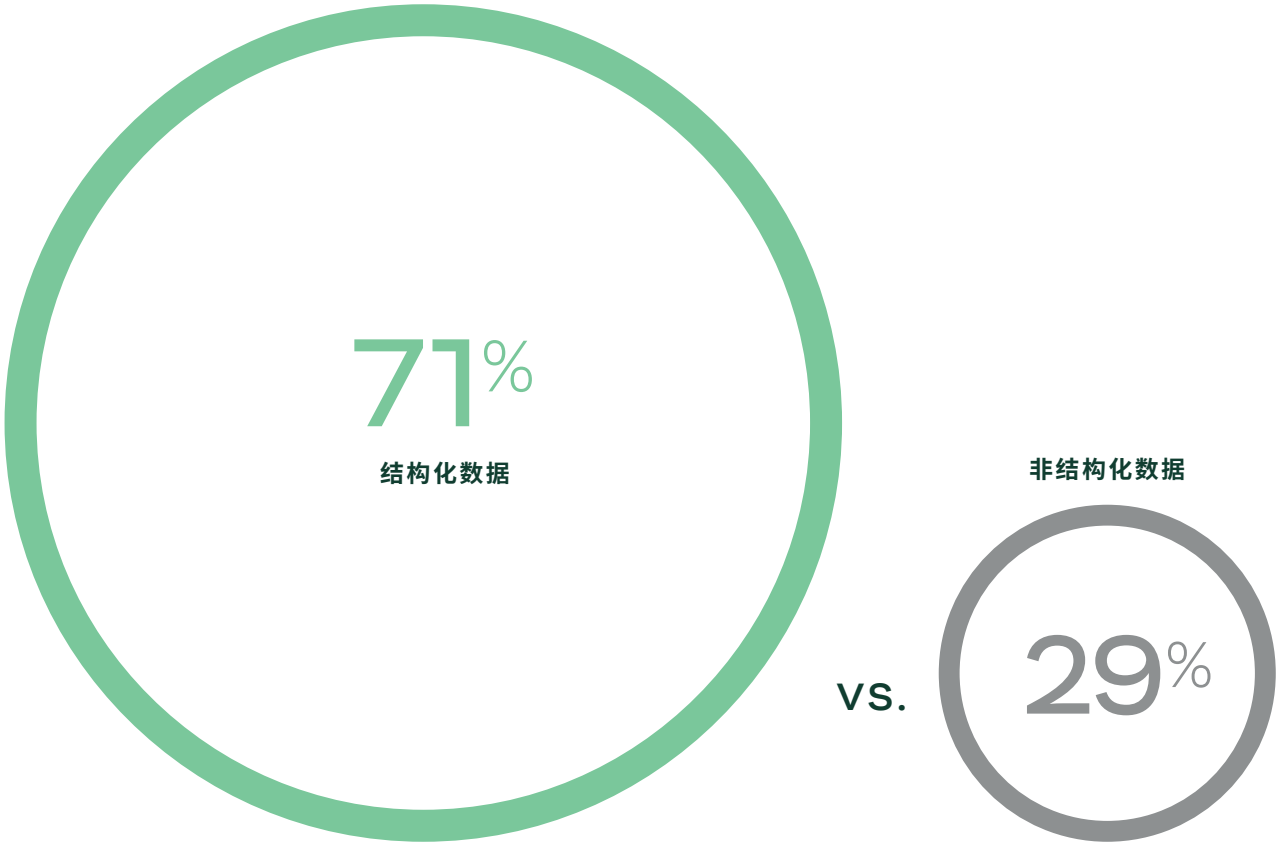
今年，媒体关注的重点是自动驾驶汽车或家庭助理等机器学习项目，但必须意识到绝大多数数据科学家处理的数据并非激光雷达和音频话语数据。

我们采访了不少数据科学家，发现日常工作中他们还是以处理文本和时间序列数据为主。很少涉及感知器、音频和视频数据，相对而言，排名第四位的是静态图片。

数据类型


180
160
140
120
100
80
60
40
20
0
数据类型
文本
时间序列
产品或SKU
静态图像
感知器
音频
视频

处理结构化数据与非结构化数据的比例？



数据科学伦理问题

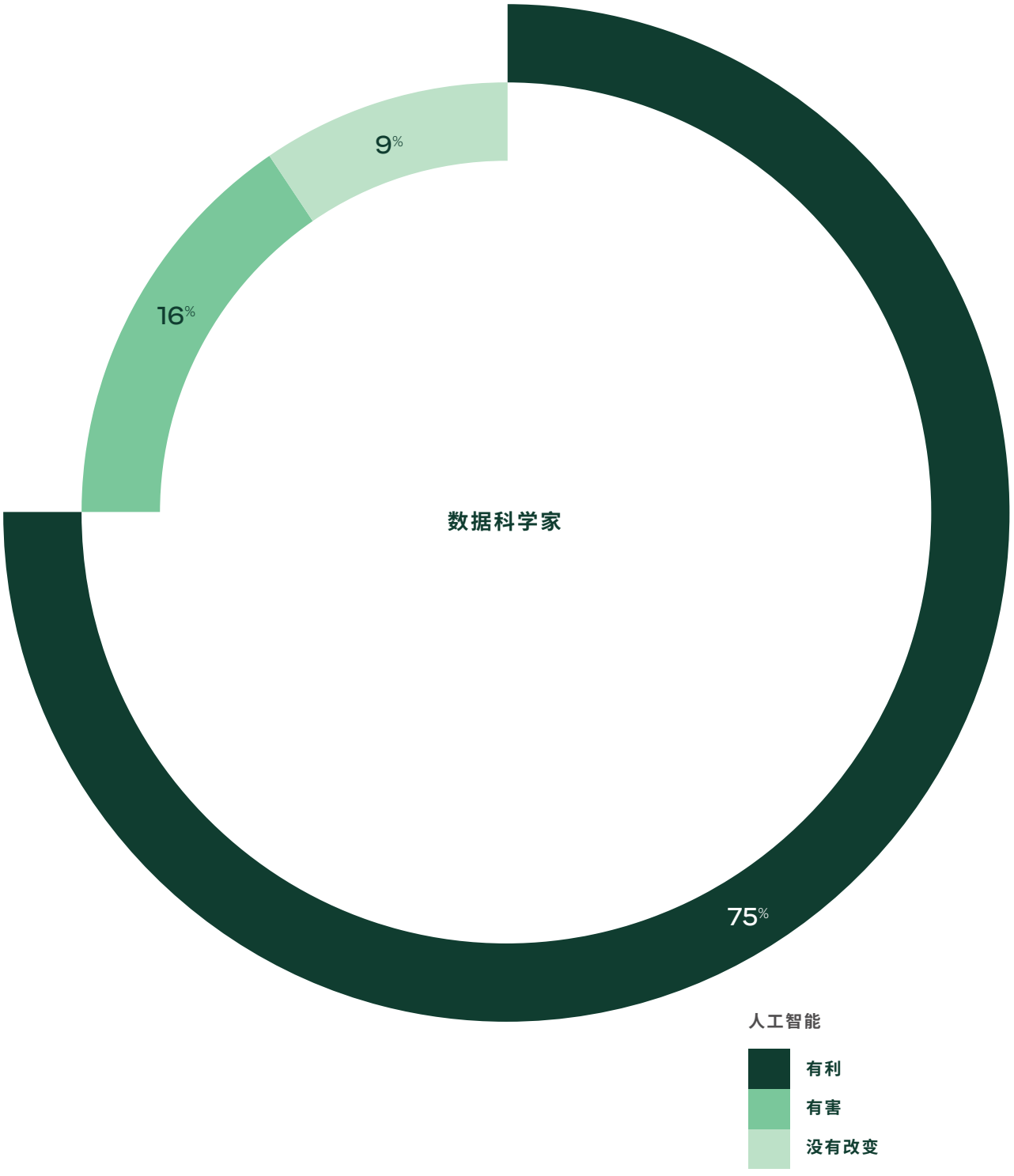
近年来，人工智能应用的伦理问题被炒得热火朝天，仅我们就了解到大量关于人脸识别、招聘审核和声音助理等子领域的算法歧视案例。去年，最高法院曾有机会处理一桩关于算法量刑的案件（详见卢米斯诉威斯康星州一案），但最高法院没有受理此案，虽然如此，也可以推断10年内很有可能出现关于机器学习的判例。

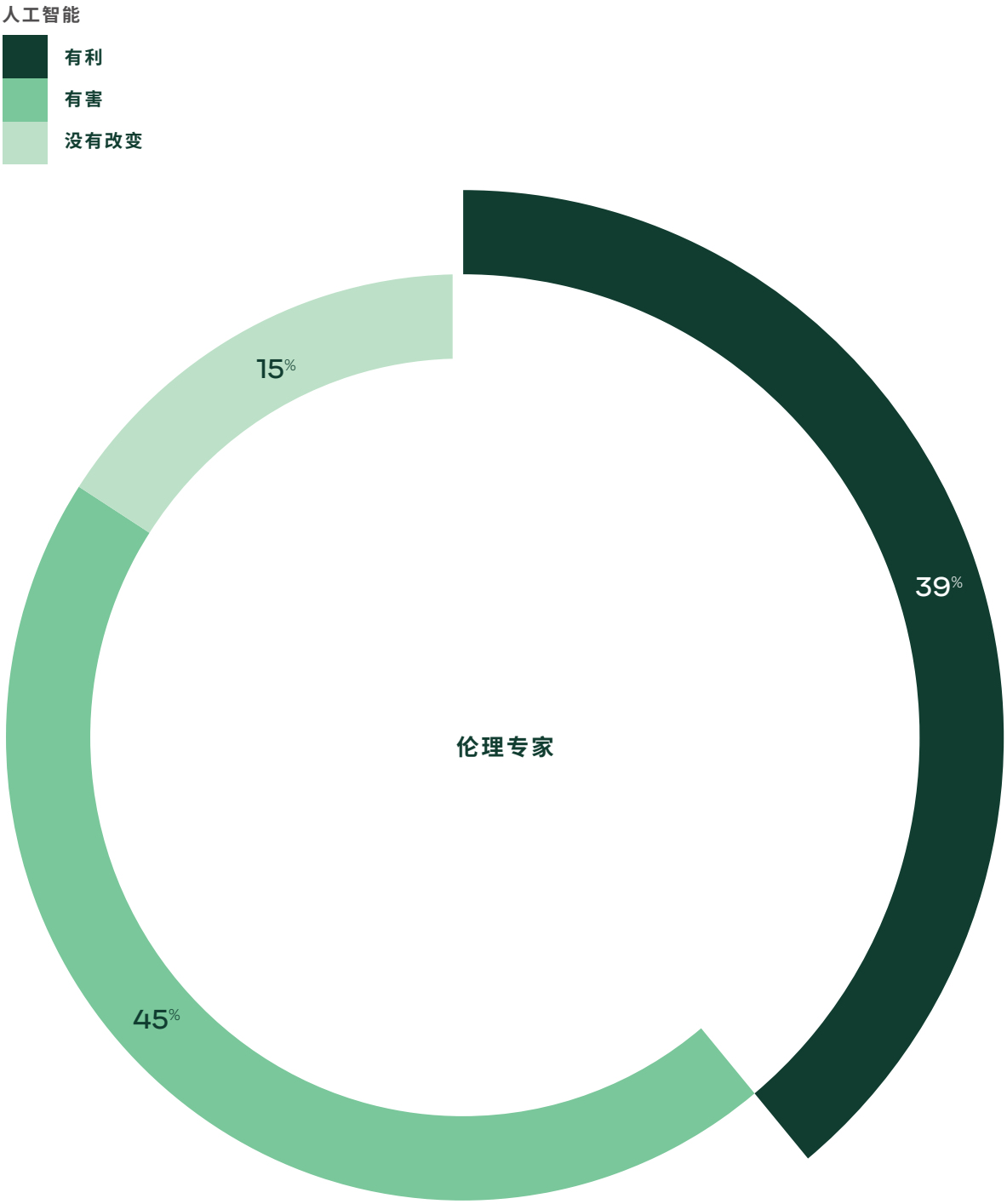
本文不关心远期的，诸如未来特工或普世智能等带有科幻色彩的，甚至有关意识边界的伦理问题，现实问题涉及的领域才是当今大众真正关心的内容，本文关注的是这类伦理问题。

之前曾说过，本次调研采访了医护人员、神职人员及执法人员等各行业的伦理专家。在这一节里，我们会把他们的观点与数据科学家的观点进行对比。

一般来说，数据科学家都看好人工智能的发展。两组专家都认为人工智能利大于弊，他们之间最大的差异在于伦理专家对人工智能可能会给社会带来的潜在挑战漠不关心。这一点倒也说的通，毕竟，大家都知道数据科学家肯定比法官对人工智能了解得更深刻。

数据科学家就身处这个领域，为人工智能的发展投入了颇多精力，因此，要说数据科学家认为人工智能不会给社会带来翻天覆的变化，那是不可能的。


9%
16%
数据科学家
75%
人工智能
有利
有害
没有改变

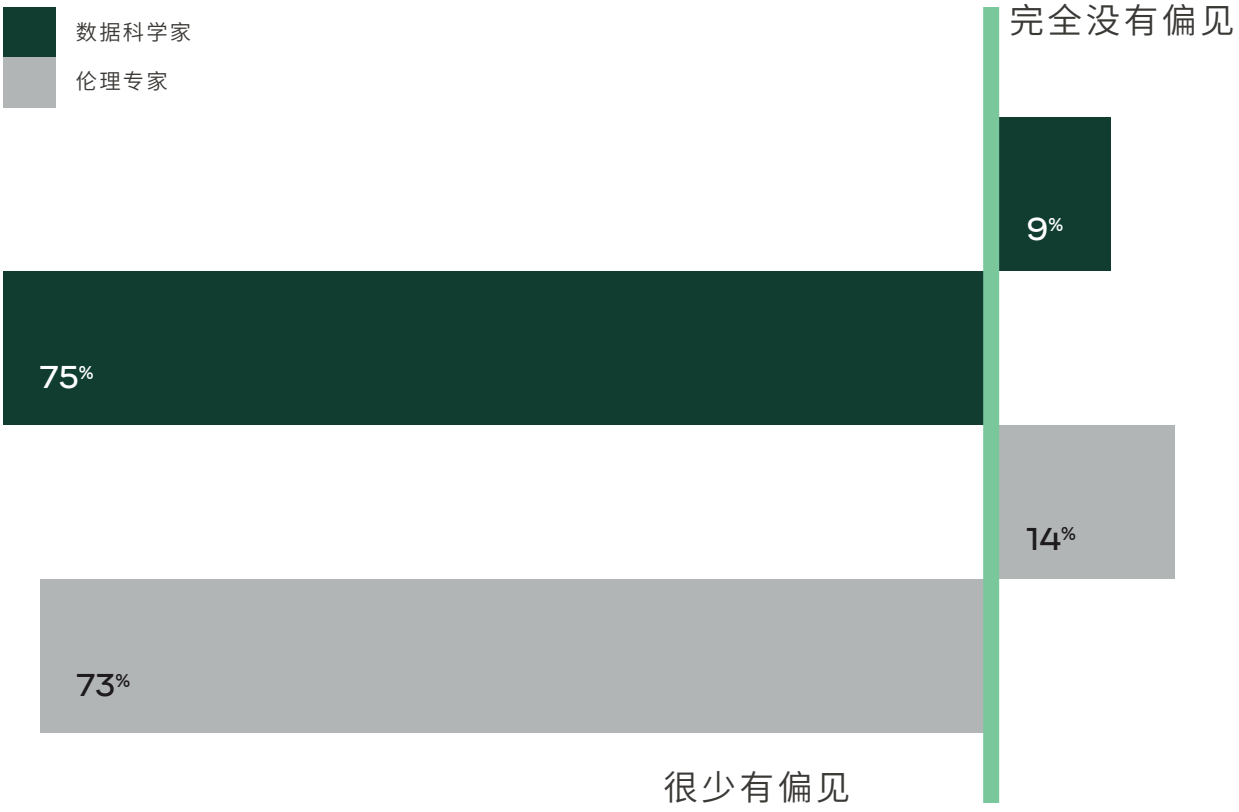

人工智能
有利
有害
没有改变
15%
39%
45%
伦理专家

还不承认算法歧视？

上一节里，我们提到了一些非常知名的算法歧视案例。实际上，麻省科技评论最近就提出了“算法歧视已经遍地都是，但是大家对此都漠不关心”的观点。

但是，当我们问及数据科学家与伦理专家是否认为人工智能比人类更容易产生歧视时，得到的答复是这样的：

人工智能比人类更容易产生歧视？还是更少产生歧视？



其实，大家都知道对比技术是否比人类更容易产生歧视这个问题本身就非常滑稽，这基于你对人类本性的认识。

归根结底，算法歧视源于人类程序员、数据及一些不可言的原因。但有趣的是，很多反馈都说算法没有那么多歧视，甚至根本就不存在歧视，然而不管怎么说，我们手里确实有大量现实中已经发生的算法歧视案例。

我们真正要解决的问题是到底为什么会出现这样的结果？要知道大部分情况下，不是算法模型本身的问题，而是模型使用的数据有问题。

算法模型的歧视是潜在、无意识的，但又是真实存在的，要解决这个问题需要花费大量的精力，还要对症下药，首先，标注数据时要认真负责，不偏不倚；然后，还要通过不断更新数据对模型进行迭代；并且还要站在最终用户的角度来思考问题。

现实世界中
人工智能到底能干什么

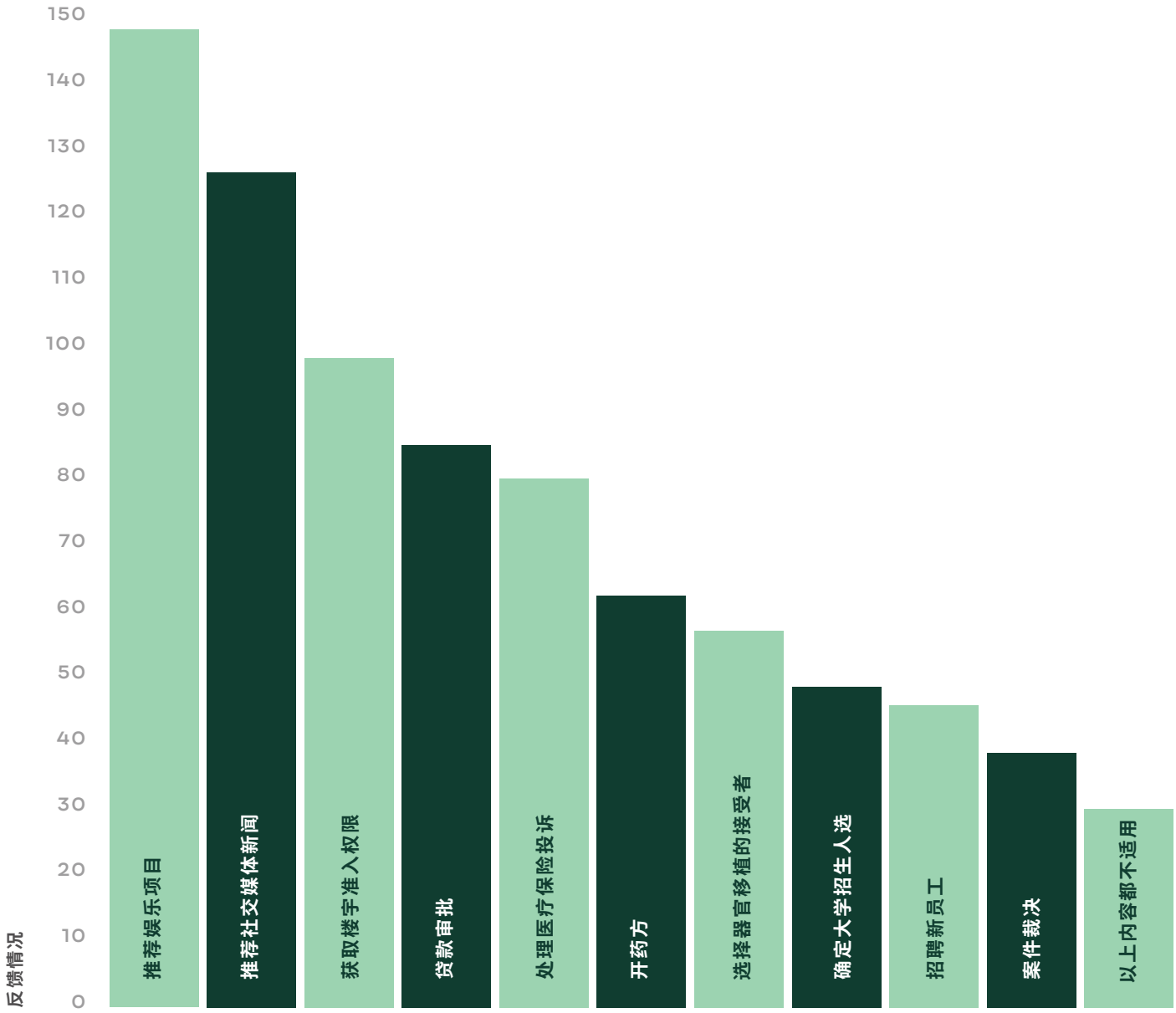
现在，绝大多数的互联网用户每天都会用到人工智能。产品和娱乐内容推荐、搜索引擎、新闻推荐，你能想到的基本上都有：机器学习的应用已经扩展到越来越多的领域。

怎么说呢？实际上，大部分数据科学家觉得人工智能参与决策这件事很正常。事情越复杂，数据科学家就会觉得越不舒服。

虽然，在一些无关紧要的场景下，人工智能的应用已经取得成功。但是，在涉及重大的关键性问题时，目前人工智能所取得的成果还不足以让人给出肯定的答案。现在只能说，数据科学家还没有那么大的胃口，将人工智能应用于社会的每个角落。如果人工智能专家要推行更稳健或更理智的解决方案，大家最好静下心来听听他们说的到底是什么。

伦理：人工智能决策

下面哪些场合可以让人工智能自行决策，无需人类干预。



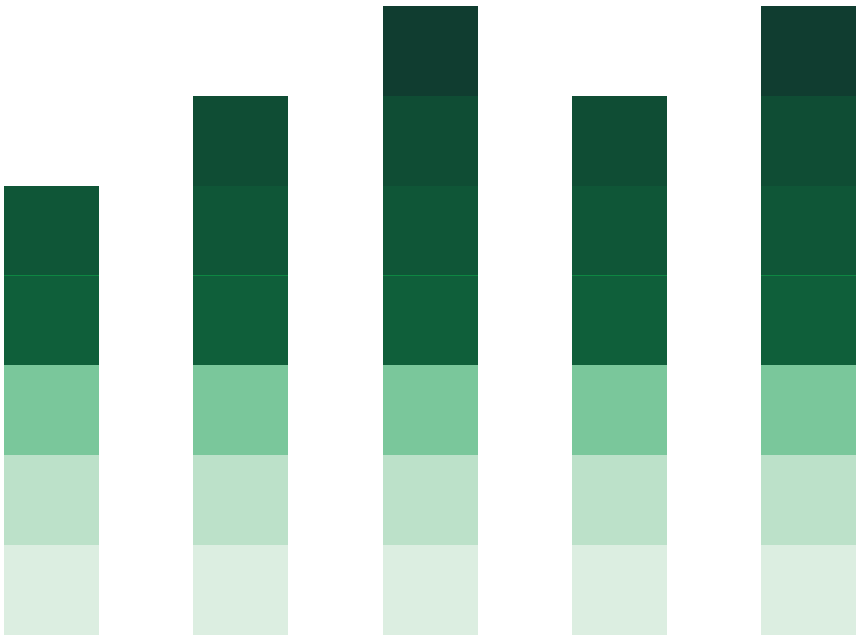
天善智能

用不用人工智能

这是个问题

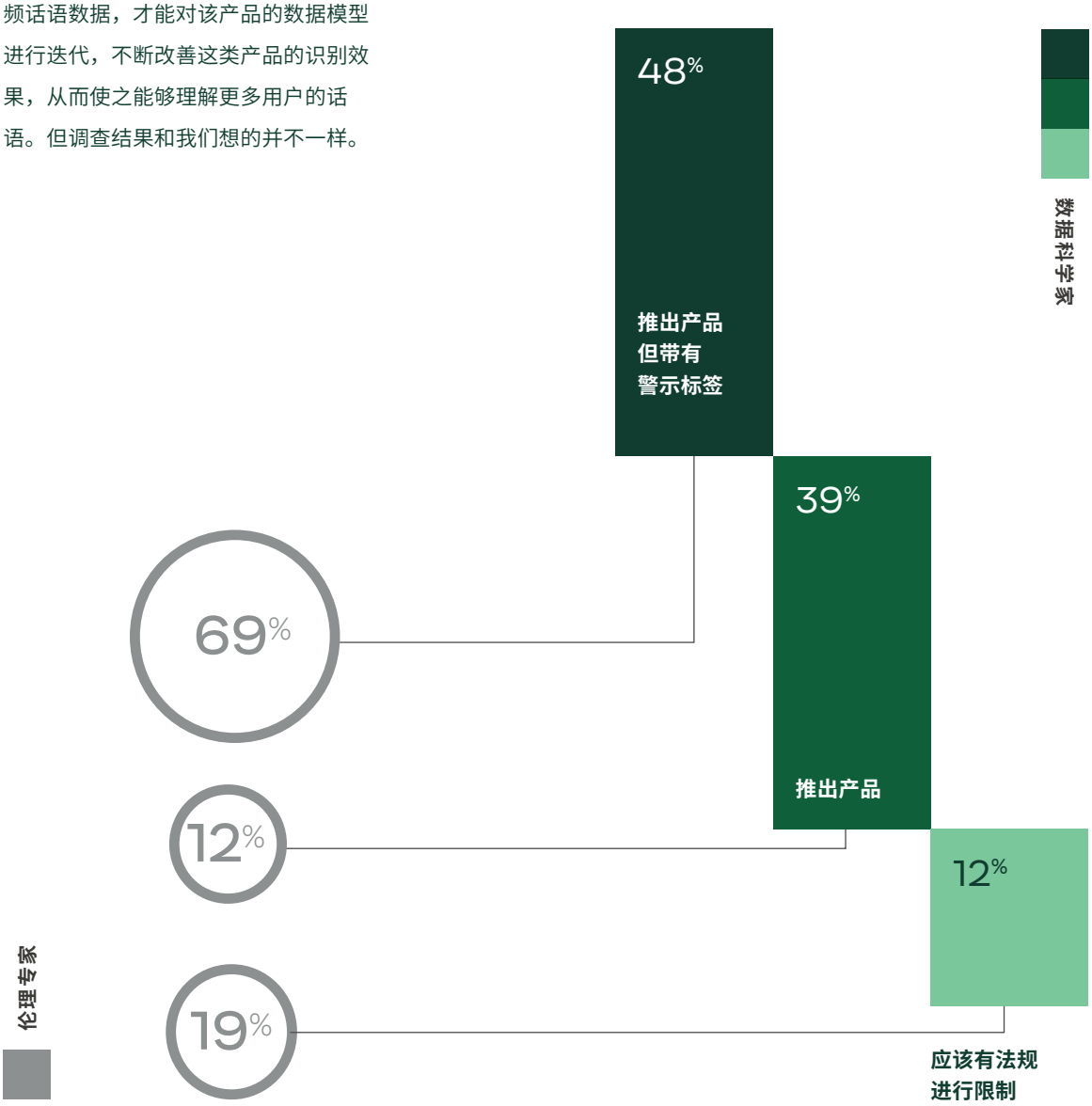
从现在开始每过去一天，音频交互界面都在变得越来越流行。Comscore公司预测2020年50%的搜索都将是语音搜索。其实即便现在，每个月都已经差不多有10亿条语音搜索了。但是，就算是最先进的语音助手仍在与每天遇到的语音作斗争。尤其是遇到说话的人讲的不是母语，或有口音、说方言的时候，这个问题就会愈发严重。

就此问题，我们特意咨询了相关数据科学家，希望了解如果推出家庭语音助理类产品，但该类产品又不能很好地理解口音和方言时，是不是仍要坚持推出该类产品，还是说要在该类产品上标明警示，提醒哪些人不适用，或者是否有相关法规会限制该类产品在某些区域销售。



坦白的说，我们希望数据科学社区能够推出这些产品。因为不管怎么说，只有把这些产品销售出去才能采集更多音频话语数据，才能对该产品的数据模型进行迭代，不断改善这类产品的识别效果，从而使之能够理解更多用户的话语。但调查结果和我们想的并不一样。

用不用人工智能
这是个问题



虽然，我们对这样的结果感到惊讶，但这和之前的调研结果也非常契合，数据科学社区对人工智能的应用非常谨慎。他们喜欢的事情搞得清清楚楚，然后再实施。

回想数据科学社区对开源平台和开源数据的热爱，就会理解为什么他们会做出这样的选择。

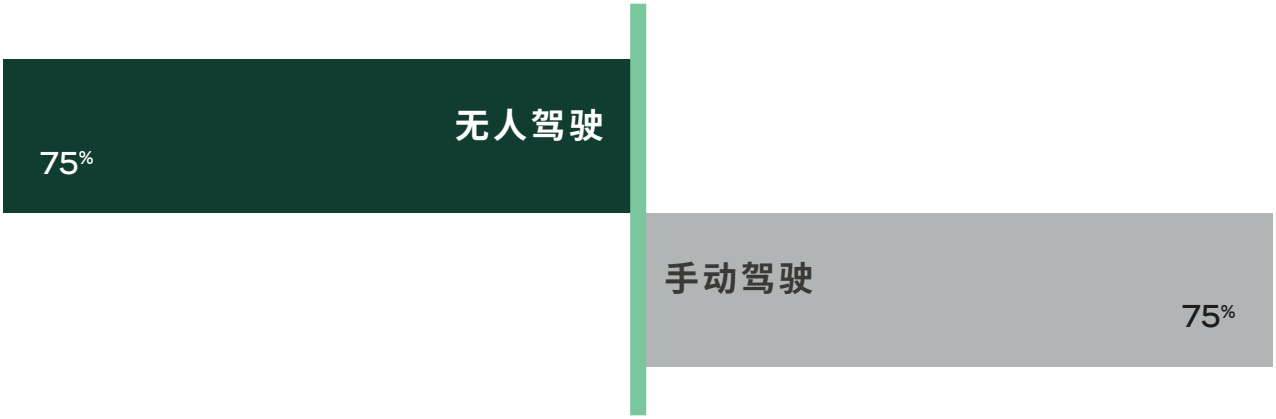
对于自动驾驶双方差异极大

我们问了伦理专家和数据科学家一个非常简单的问题。如果统计数据表明，最新的人工智能比人类驾驶汽车的平均安全系数更高，你是愿意自己驾车呢？还是愿意开自动驾驶汽车呢？

对于调研报告里面的其他内容，两组调研对象的反馈基本上都非常相似，总的来说，他们都认为人工智能利大于弊。即便是对于某些比较敏感的人工智能产品，也只需标清哪些人适用，哪些人不适用就可以了。比如，大家普遍都能接受人工智能驱动的产品推荐功能，对人工智能驱动的贷款审批或案件裁决持保留态度。

但是对于自动驾驶，两组调研对象存在严重的两极分化，这只能说明数据科学家对无人驾驶技术的运行机制比神职人员了解的更多。不过，我们确实没有预料到两极分化的情况会这么严重。我们现在还很难解释清楚为什么两组调研对象会有如此不同的反应，但如果你从事于自动驾驶汽车行业，现在就应该清楚你的营销对象是谁了吧。

自动驾驶，还是手动驾驶？



数据科学家

伦理专家

报告背景

今年，我们通过邮件和现场访谈等形式采访了240位数据科学家。

如需获取2015年版数据科学报告，请到我司官网的资源中心下载。

Figure Eight是为数据科学团队提供人际回圈型人工智能平台的公司。我们为客户的机器学习模型提供高质量的自定义训练数据，还为客户提供易于部署、便于使用的人工智能模型及整合人机回圈的工作流。

我司的软件平台支持包括自动驾驶汽车、个人智能助理、医疗图像分类、内容分类、客户支持票证分类、社交数据分析、CRM数据补值、产品分类及搜索相关性分析等众多业务类型。

我司总部位于旧金山，投资者为Canvas创投、Trinity创投、微软创投。Figure Eight是一家涉足多个行业，快速增长的数据驱动型公司，我们的客户主要是财富500强公司的数据科学团队。

figure-eight.com