所造之境，必合乎自然

---探讨数据科学与人工智能学科的发展

大数据和人工智能是当今最为热门的科技术语。我国相关部门下发了一系列重要的指导性、纲领性文件，也启动了一大批大数据和人工智能相关的重大科技专项。同时，科教精英们积极筹备设立人工智能一级学科。迄今为止，我国已经有近700多所大专院校设立了数据科学与大数据技术、200多所高校开设人工智能或智能技术本科专业。设立学科点和启动重大科研项目通常是我国发展热门学科的两大主要法宝。

诚然，因应现代科技需要，发展数据科学(Data Science, DS)和人工智能(Artificial Intelligence, AI)是必然趋势，也是各国积极布局的重点。但是它们并非新兴学科，其发展历程也非一帆风顺。如何更好地发展它们仍然值得进一步探索。这里我们把自己多年来的一些疑惑以及思索写下来和大家探讨。

第一，设立本科专业是否必要，其条件是否成熟、可行？教育的宗旨包含启智、立德、求名、取利等几个层面。我们认为本科教育阶段应更专注于基础知识体系和综合素质培养。而高尖端的专业人才的培养则更适合于在研究生阶段进行。特别地，AI是一个高度专业化、快速迭代化、跨领域的交叉学科。因此，我们认为先开设DS或AI的硕士、博士项目，等积累一定学术资源和教学经验之后再考虑设置本科专业是更为稳妥、也更可行的方案。

无论是DS还是AI，都是数学、统计学和计算机科学的交叉学科。这些现有本科生专业能够支撑DS或AI研究生项目。反过来，DS和AI促使计算机科学和统计学也要与时俱进，拓展其领域边界。一方面，计算机科学的理论基础正在从离散、组合数学转移到以概率和统计为核心。因此，其本科专业体系需要增加概率统计的分量。另一方面，统计学本科培养应重在方法/方法论(Methodology)，且需要与现代计算思维深度融合，这包含求解连续问题的数值分析、求解离散结构的算法和大规模分布式计算等。

机器学习是数据科学和人工智能的核心，它的作用越来越重要[1]。一方面，我们可以为不同专业本科生开设机器学习导论，讲述机器学习的模型、方法和应用。代替数学模型，作为通识课。同时，可以为统计学和计算机科学高年级本科生开设机器学习基础，强调机器学习本身的理论体系和数理基础。这两门课类比于高等数学和数学分析的各自不同定位。

第二，除了启动重大专项经费支持，是否还有其它有效方式来促进数据科学和人工智能的发展？作为交叉学科，它需要不同领域背景（特别是数学、统计学

和理论计算机科学）学者深度合作，寻找新的学术方向和研究问题、开设新的课程、探索新的学术机构组织机制和运行方式。

美国自然科学基金委 NSF 设立了 TRIPODS (Transdisciplinary Research in Principles of Data Science) 项目[2] 来建立数据科学基础。自 2017 年以来该项目在大学里建立了 12 所数据科学基础机构。不同机构因其立足的学校学科优势而有不同的特色和侧重点，但都集学术研究、教育以及探索符合数据科学发展的新型研究机构的运行机制于一体。其实，这 12 个机构也聚集了各自大学的 AI 主要学者。

比如，加州大学伯克利的 TRIPODS 项目 FODA(Foundations of Data Science)研究所致力于从基础教育到交叉尖端研究等方面来深化数据科学的理论基础，并将这些基础进展转化至不同领域数据科学实践之中。该研究所 PI 包括 Michael Mahoney (主任), Bin Yu, Michael Jordan, Peter Bartlett, Fernando Perez, and Richard Karp。他们面向全校开设了《数据科学基础》课程。而 MIT 数据科学基础研究所(MIFODS)所从事的是一项跨学科的工作，旨在通过综合研究和教育活动来发展数据科学的理论基础。其目标是在 MIT 和乃至整个学术界内促进数学、统计学和理论计算机科学间的研究和教育互动。因此，MIFODS 成员都是这三个领域的一线学者。

在人工智能领域，加拿大多伦多大学和蒙特利尔大学依靠其 AI 科学家分别建立了独立运行的研究机构 Vector Institute (图灵奖 Geoffrey Hinton 领衔)和 MILA(图灵奖 Yoshua Bengio 领衔)。在英国，则以剑桥、牛津等大学相关领域教师为主体，成立了跨校“图灵研究所”。这些机构主要从事学术研究，培养博士和博士后。

第三，我们想以统计学为例，看近 20 年以来一些著名统计学家是如何引领统计学的发展和拓展经典数理统计领域边界的。

早在 2001 年，已故加州伯克利大学统计学家 Leo Breiman 在他的著名论文“Statistical Modeling: The Two Cultures”[3]中提出了统计学要吸纳计算机科学解决问题的文化，并致力于机器学习的研究和教育。CMU 统计学家 Larry Wasserman 在他的名著“All of Statistics”[4]的引言部分，反思了统计学和计算机科学界从轻视对方到相互欣赏背后的原因。正是这促使他写下了这本书。成为计算机专业的统计学教材，同时又是统计学专业的机器学习教材。

斯坦福大学统计学家 Bradley Efron 和 Trevor Hastie 为了应对大数据、数据科学和机器学习对传统统计学的挑战和机遇，撰写了“Computer Age Statistical Inference: Algorithms, Evidence, and Data Science”[5]这一现代统计推理经典教

材。斯坦福大学数学家和统计学家 David Donoho 则在 2017 年发表了文章“50 Years of Data Science”[6]，对数据科学和统计学作了全面、系统和深刻的梳理和展望。而且，他新近又在斯坦福开设了课程“Theory of Deep Learning”。

2018 年 10 月，美国 NSF 组织了题为“Statistics at A Crossroads：Challenges and Opportunities in the Data Science Era”研讨会[7]，48 位领先研究者和教育者聚集一起，讨论统计学未来 10-20 年的发展愿景。2019 年统计界一个有重要影响的事件是把统计学界最高奖 COPPS 奖颁发给了 RStudio 公司的首席科学家 Hadley Wickham，表彰其在统计应用和统计软件开发领域做出的卓越贡献。历届获奖者都是对统计理论做出杰出贡献的学者，今年首次来自业界。反映了统计学界在积极应对大数据和计算所带来的变革。

Leo Breiman、David Donoho、Bradley Efron、Trevor Hastie 和 Larry Wasserman 等是世界最著名的统计学家。他们在统计学领域不仅做出了许多奠基性的工作，且以推进统计学发展为己任，纷纷撰写教材和开设新课程来梳理学科的知识体系和开拓学科的领域边界。有境界则自成高格，自有名作，学科能常青在此。

王国维先生提出“境界”说来评价诗词，先生认为诗词的创作手法分为“造境”和“写境”两种。造境和写境分别对应于“理想”和“写实”，然而这两者常常浑然一体，颇难区分。**所造之境，必合乎自然；所写之境，亦必邻于理想**。发展学科不也应遵循此法乎！

参考文献

1. M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 2015.
2. https://nsf-tripods.org/institutes/
3. Leo Breiman. Statistical Modeling: The Two Cultures. Statistical Science, 16 (3): 199-231, 2001
4. Larry Wasserman. All of Statistics: A Concise Course in Statistical Inference. Springer, 2003.
5. Bradley Efron and Trevor Hastie. Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. Cambridge University Press, 2016.
6. David Donoho. 50 years of Data Science. Journal of Computational and Graphical Statistics, 26(4): 745-766, 2017.
7. Xuming He, David Madigan, Bin Yu, John Wellner. Statistics at A Crossroads: Who is for the Challenges. The National Science Foundation, Report 2019.

2019 年 12 月 8 日初稿 2020 年 3 月 2 日修订