

前言：学术江湖任我行—评价学诞生记

詹剑锋博士撰写

谨以此文献给无数为梦想挣扎的年轻人和曾经的年轻人，大家终将得解放。

早上我从梦中醒来，非常懊悔。我已经梦里写了两个小时前言。要是早起床两个小时，已经完成了这个艰巨而幸福的任务。

2009 年，我接到一个艰巨的任务，来自当时的领导孙凝晖研究员。孙凝晖研究员 2019 年当选中国工程院院士。他让我负责撰写新型计算基础设施的报告。开始在内部把这种新型计算叫高吞吐计算，后来改名为高通量计算。当时我们所在的高性能计算机中心跃跃欲试，准备向新型数据中心转型。孙凝晖研究员把我介绍给了著名的普林斯顿大学教授李凯。请他指导我撰写这份报告。同时期，加入讨论和撰写报告的有中国科学院计算技术研究所（中科院计算所）和中国科学院研究生院（现在的中国科学院大学）近二十位不同方向的研究人员，详细名单可以从图 1 的致谢里找到。当中不少人已经成了知名专家和教授。

李凯是个传奇人物。他 1980 年代毕业于中科院计算所，不仅后来成为美国工程院院士，而且是成功的创业家。他创办的 DataDomain 公司被存储巨头 EMC 几十亿美金收购。

李凯教授给我的两点印象非常深刻。一是谈到创新，他喜欢算一算最新技术的成本。最新的内存多少钱？处理器多少钱？他了如指掌。这让我很惊讶。在国内，算成本似乎显示不了水平，离科学家很远。李凯教授在做研究之前，还喜欢先用最先进的工业界技术搭建一个实验系统，看看“油水”有多大。油水是他的原话。

还有一点，李凯教授至少做成功了两个有影响力的评价基准 (benchmark)。一个是评价处理器的 PARSEC，芯片巨头英特尔资助。还有一个是和传奇华裔女性教授李飞飞合作的 ImageNet。学术界和工业界普遍认为后者是促进人工智能繁荣的力量之一。

我钦佩李凯的影响力。回过头看。和李凯的交流恐怕是我创建评价学理论的开端。我将评价学定义为揭示万物影响的科学。在这里，我也用同样的方法，回顾影响我和这本书的人和事。

在这十几年岁月里，有两件事值得回顾，一件充满着幸运，另外一件则没有那么幸运。放在一起，就像评价学里面提到的对照组。

第一件是我的第一个有影响力的评价基准工作：大数据系统基准 BigDataBench。2013 年，我当时的博士生王磊和我以及别的合作者一起工作。这篇论文的撰写我们只花了两个星期时间，不过应该熬了两个夜。最终这篇文章被计算机体系结构顶级会议 HPCA 2014 工业版录用。我当时在美国旧金山机场的一架即将起飞的飞机上，颤悠悠地提交论文。飞机上网络信号很差，点完提交按钮，五分钟后，飞机起飞，回北京。

致谢

本技术白皮书是在孙凝晖研究员的具体指导下,由高性能计算中心詹剑锋负责执笔完成。普林斯顿大学的李凯教授在全文结构、内容以及 Benchmark 的重要性方面给予了指导。在完善本技术报告方面,张立新研究员提出了很多重要建议,如完善 HTC Benchmark 和增加 IBM SMT 技术内容。

2.1 节和 4.4.1 节摘自 Google 工程师的著作[LUIZ2009],由王磊助理研究员、袁琳硕士生、刘雪姬硕士生、席华锋硕士生、李勇博士生、黄雅斌硕士生翻译。3.1.1 节, 4.5.4 节和 6.3 节基于熊劲副研究员的调研和贡献。4 节部分材料来自 Wiki 等网络上的匿名作者。4.5.1 节内存系统部分和 6.1 节内存系统对体系结构的影响基于包云岗博士的调研和贡献; 4.5.2 节基于曹政博士和王凯博士的调研。4.5.5 节基于王鹏博士的调研。前瞻中心朱登明副研究员和中科院研究生院徐俊刚副教授对 HTC 的部分应用有贡献。部分有关 HTC 应用的材料来自网络中心的申请材料。有关 IBM 智能地球的材料来自孙毓忠研究员的调研。6.1 节部分材料基于张立新研究员的贡献。6.2 节的部分材料取自文献[ANDR2009]。6.5 节部分材料基于王磊助理研究员的贡献。6.6 节基于人民大学单智勇博士的贡献。

高性能计算中心、网络中心、体系结构实验室、前瞻实验室、应用集成中心和中科院研究生院的 20 个人左右的志愿者团队不同程度地参与了讨论,包括张立新研究员、陈明宇研究员、熊劲副研究员、曹政博士、王大伟博士、王磊助理研究员、朱登明副研究员、徐俊刚副教授、张军超博士、王伟平博士、张迎平博士、刘振军博士、郭明阳博士生、熊刚助理研究员、袁清波博士生、张文力助理研究员、包云岗博士、吕慧伟博士生和毛天露博士。

Figure 1: 2009 年我牵头撰写的技术报告。致谢部分提到的不少人已成为著名专家和教授。

我们非常幸运。文章得到高分。工业版主席来自芯片巨头英特尔,给了我们很大的鼓励和支持。一开始,我们就觉得这篇文章会有影响力。后来,我们团队的贾国博士去李凯教授那做博士后。他告诉我,普林斯顿的一个著名计算机教授让李凯推荐大数据基准。李凯教授毫不犹豫地推荐了 BigDataBench。

差不多相同的时间,我另外一名博士生陆钢向 ASPLOS、SOSP 和 OSDI 等顶级会议提交了一篇操作系统论文。我们一共花了六年时间,投了六次稿。前几次,第一轮分数很高,第二轮就被人干掉了。我仍记得有一年投 ASPLOS,一位审稿人甚至在第一轮给出了“强烈接收”的评价——这在系统会议中相当罕见。最终,我决定不发表这篇文章。

第二个工作是 AIBench,我们有关人工智能系统的评价基准。当时的主要贡献者是高婉铃博士生。她非常聪明和勤奋。论文提交给 HPCA 2020 的工业版。婉铃在提交论文之前,三天三夜几乎没有睡觉。

我对文章录用充满着信心,因为我们更有经验了,英文写作水平也提高很多。期待着论文有更大的影响力。评审结果也如我所料,收获了审稿人非常高的评价。五个审稿人:四个接收和一个弱拒绝。

可是我注意到一些异常。给弱拒绝的评审人提到,“人工智能系统评价基准有一个



Figure 2: Jack Clark 有关 AIBench 和 MLPerf 竞争的评论。2025 年 12 月 1 日截图。截图的目的是提供证据，以防访问链接消失。

MLPerf 就够了，没有必要有第二个工作”。MLPerf 聚集了几乎所有的美国巨头公司和顶尖大学，例如斯坦福、谷歌等，是我们的竞争者。MLPerf 当时还没有发表论文，而且也举步维艰，牵头的科学家我也认识，他还跟我抱怨得到的支持很少。

我决定给工业版主席写信，他同样来自英特尔。我感谢他付出的努力，并且提醒他：AIBench 和 MLPerf 是竞争者。我也强调同一个领域有两个独立的竞争工作，对社区有利。这不是我的一厢情愿。当时美国著名的信息技术媒体记者 Jack Clark 在评论 [1] 中将我们的 AIBench 和谷歌主导的 MLPerf 标准之争上升到中美科技竞争的高度。Jack Clark 后来成为 OpenAI Strategy and Communications Director 和斯坦福 AI Index 共同发起人。Jack Clark 的评论见图 2。

这次，主席没有给我回信。我感觉不妙，一定有事情在悄悄发生。因为工业版的评审委员会里有 MLPerf 的主要贡献者。果然不出我所料，我们的文章最后以非技术的原因被拒，有人在后面悄悄运作。

我很愤怒，给大会组织者以及更高层委员会写信。高层委员会的两个美国教授我都认

识，有一个还可以说是熟悉。有一个女教授还专门给我写信，了解详细的细节。主席给我写信，表示歉意，说让我感到不公平。但最终委员会的决定还是倾向主席。我毫不奇怪。差不多同样时间范围内，计算机体系结构圈内发生了更糟糕的事。因为评审人串通、运作，一名高大的华裔博士生的论文在顶级会议 ISCA 上录用。内容有明显的错误，他想撤稿，导师不同意，最后他选择在导师办公室自杀，震惊全球计算机体系结构圈子，而且影响在外溢。即使这样震惊的事件，博士生自杀了才换了姗姗来迟的调查和处罚。

MLPerf 的论文半年后被 ISCA 录用，成了一颗超级明星。我们的论文后来又投给了 Micro 2020，同样是高分。据程序委员会主席给我的回信，在最后讨论时，被全体评委投票给拒掉了。后来又投给 PACT 2021(都快两年过去了)，又注定被拒时，我愤怒了，给主席写信。主席是个法国科学家，为我们不平。他再找了些评委，把论文录用了。但文章的影响已消耗殆尽，我们的论文发现刚开始让评阅人很吃惊，后来慢慢成为常识。

这是一个典型的案例，证明了我们科学技术社区有多残酷；每一个从事科学技术研究的人，他或她的遭遇可能会有多么不公平。因为这件事，高婉铃有几年都感觉到深深地愤怒。有一次聊天，我提到我们有些地方做得不对。高博士跟我说，即使做对了，他们也不会录我们的论文。如果比较两篇文章或两个作者的命运，公开条件几乎一样，最终决定结果的是那些未知的，或已知却在背后运作的因素。这是我们每个人不得不面对的残酷现实，这也是评价学这本书的主题之一：准实验。

我已经不再愤怒了。我超脱愤怒靠关注更有意义的事。我觉得不受人制约的最佳方法就是做一个更深刻的工作，不随大流。这也是我开始评价学研究的动力之一。

2021 年，我意识到评价基准尽管在计算机、人工智能、金融领域广泛使用，却没有严谨的科学方法。即使是最著名的人工智能评价基准 ImageNet 也同样没有严谨的方法。我听过李凯教授的公开报告。他说，当年和李飞飞一起申请项目的时候，被人嘲笑过。李凯当时已经是非常资深的著名教授。我觉得应该发展严谨的评价方法。21 年，我在自己创刊的期刊 TBench 上写了一篇社论，呼吁建立 Benchmark 科学与工程 [223]。

但什么是 Benchmark 呢？从来没有人严谨地定义过，也不太被人重视。我曾经的博士生汤飞曾经跟我说，别人问他的研究方向，他都不好意思说自己是做 Benchmark 的。他觉得看起来好像很低级，至少一点都不高级。汤飞是一个非常勤奋、有天赋和诚恳的年轻人。他现在在浪潮工作。

22 年，我意识到了 Benchmark 和评价之间的关系。我问了自己一个非常的关键问题，“为什么药物评价用随机控制实验 (RCT)，但计算机领域仍在用经验方法，例如评价芯片的 SPEC CPU？”每个 CPU 产商都喜欢用 SPEC CPU 报告性能数据。我的博士生王晨曦早就证明：同样一个处理器，完全遵循 SPEC CPU 的方法来评价，性能数据之间可以差异几十倍，甚至百倍。也就是说，两个公司生产同样性能的 CPU，报告数据时，一个公司可以说自己的性能比另外一个公司好几十倍，而遵循的评价方法是大家认可的。这样的数据有意义吗？

我开玩笑说，报告一个处理器的性能数据，反正不会死人。评价药物则不同，它真的会杀死人。在其他领域，情况更糟糕。比如说大学排名，它甚至会让一些青春期的男孩、女孩甚至他的父母苦恼或者抑郁。多么讽刺！

2023 年，我正式创建了一个英文单词 Evaluatology 来命名评价学，并且写了一篇文章，评价学：评价科学与工程 [226]。我将这篇文章发给了一些学者。美国一个著名的教授 David Lilja 给我回了信，鼓励了我。我也收到国内不同领域的同行的反馈。比如中国民航飞行学院的首席教授李维萍教授，华东师范大学周傲英教授、钱卫宁教授和王伟教授等。2024 年，我们还在广州组织了一次评价学研讨会，很多学者热情参会，参与讨

论和交流。钱卫宁教授还跟我开玩笑说，这可能是评价学元年。

24 年至 25 年，我几乎无休息地投入了本书的写作。最初计划一个人独立完成，但很快意识到单打独斗的困难，也意识到带同事与学生一起工作更有价值。我第一批邀请了王磊博士、高婉铃博士和博士生李鸿霄加入。一个月后，进度还是滞后，我又邀请博士生王晨曦，特别研究助理范帆达博士加入。上个月，特别研究助理康国新博士对基于评价学的人工智能产生了浓厚兴趣，并付出大量努力，我认为邀请她加入才公平。

在两年研究和写作过程中，有几件事值得提及：

一位熟人受我的评价学公开演讲的启发，快速撰写了一篇文章，通过特殊渠道很快发表在一篇知名杂志上，其中有些观点明显受我的工作的影响。在我公开报告之前，我的论文早发表一个月。他是主动跟我交流的，说他的工作很早就录用了，希望我能参考他的工作。我知道这些杂志的运作方式，还认识他们的责任主编。看到我发了一些关键证据后。他又强调曾在另一篇文章中匿名提及过我的工作；另外想邀请我在他负责的一个杂志专栏写篇文章。我拒绝了他的提议，但也没有向委员会投诉，以免影响他的职业发展。

科技生涯中常遭遇这些令人失望的情况。例如我曾经写过一份技术报告，有人问我有没有发表，如果没有，他想在我的报告的基础上出本书。我对此非常反感，因此在这本专著里，添加了不少脚注，明确标注灵感的来源或者是谁的贡献，以示对他人贡献的尊重。

许多人轻视评价学，认为它是“软领域”。我完全反对此观点，坚信评价与设计同样重要，甚至可能催生基于评价学的新的人工智能范式。在一次交流中，罗纯杰博士提出评价与设计是对偶问题，通俗地讲就是同一硬币的两面。他的论证很有说服力。我在脚注里清楚地提到这个想法受他的影响。

另外闫和东曾经和我联系，他想跟我读博士，研究评价学。我让他先做员工。我们相处不太好，一年后，他走了。我之所以要提及他，是因为他带来了不少有关本书第四部分资料，尽管他对这部分内容没有具体贡献。

25 年某一天，我突然想到：评价学可被定义为“揭示事物影响的科学”。这一灵感源于朱迪亚·珀尔博士和达娜·麦肯齐的著作《为什么之书》。

在评价学发展过程中，有两位中外科学家应该提及。一位是美国的迈克尔·斯克里文 (Michael Scriven) 教授，他在 1966 年发表的文章里，就认为对不同事物的评价应该是相似的 [173]。之后他一直致力于建立评价学科，发表了大量有影响力的文章。他是最早提出要建立评价学科的人之一，我追溯到的最早文献是 1966 年。中国学者邱均平教授和他的学生们在 2000 年出版了一本评价学的书 [57]，他可能是最早在中文使用评价学这一词的人。但他们完全忽略了迈克尔·斯克里文的大量著作。

我本来一开始不想使用评价学这个词，用的是评价科学和工程 (Evaluatology)。但邱教授的书更多的是各个学科领域方法的汇总，没有统一的方法。另外，他更多的是从社会学、文献计量学的角度出发，完全忽略了实验设计 (DoE)，随机控制实验 (RCT) 等基础性的科学方法。我将评价学定义为解释万物影响的科学，内涵完全不同。所以仍然选择使用评价学 (Evaluatology)。我在书中也阐述了过去建立评价学科失败的原因；将评价定义为确认一个事物的优点和价值，容易掉入主观的陷阱；没有提出的统一的概念、公理、核心问题和方法。当然我的尝试可能也会失败。

去年，一位知名教授半开玩笑地问我：“中科院计算所为什么还没开除你？”出于尊重，我不便透露其姓名。他解释道：许多科学家忙于申请项目、争取帽子，而我似乎对此毫无兴趣。言下之意，我早该被解雇。这也正是我特别感谢计算所的原因之一。当我

从先进计算机系统研究中心调任分布式系统研究中心时，计算所主动提供了 100 万元研究经费——甚至无需我申请。对陈熙霖所长和孙凝晖院士的鼎力支持，我深表感激。

最后，我要真诚地感谢家人。用评价学的视角看，我的妻子显然是过去二十年来对我影响最大的人。我们 2001 年 10 月在香山徒步时相识，次年便在没有举办婚礼的情况下旅行结婚，直接前往九寨沟度蜜月。此后一直过着简单而幸福的生活。

女儿是我最珍爱的宝贝。非常想念你，愿你在波士顿的生活充满快乐。每个生命都是独特的，拥有与生俱来的尊严——这种尊严不在外表，而存于内心。

还要感谢过去一年里我照料的小动物、树木和花草。每当精疲力竭时，静静观察你们的成长总能带给我深沉的平静与力量。