

# David DeWitt 访谈录

本专访主要介绍了 David DeWitt 关于计算机学科课程的反思，以及为什么数据库研究团队应该自豪，为什么查询优化不起作用，超级计算基金有时被拙劣的花掉，以及他是一个多么不够好的编码人员，他没有那么聪明来做数据库理论，等等

玛丽安 温丝莱

这期我们采访的是 David DeWitt，时间是 2001 年的 11 月，在爾巴那平原的伊利诺斯州大学的罗伊·坎贝尔的高清晰电视直播间。再一次感谢那些帮助我们为本次专访以及其他专访提供问题的人。通常，所有手抄本的错误都是我自己一个人的原因，使用磁带作为最终的存储介质。这次专访被摄录下来，并且期望这个视频最后放到 SIGMOD 网站上，并且提供下载。

.....

**玛丽安:** 我是玛丽安·温丝莱，很高兴欢迎你们看 ACM SIGMOD Record 设置的数据库访谈专栏。今天，我们的客人是 David DeWitt，他是约翰 P. 摩格里奇教授，还是麦迪逊的威斯康星州大学的计算机科学专业的主任。他是一个国家工程院院士；他还是一个 ACM 会士；【他是今年的 SIGMOD/PODS 会议的分委会主席】并且他因为性能评价和并行数据库的工作被人所知。所以欢迎 David DeWitt。

我喜欢从学术生涯开始问一些问题。你已经和工业界走的很近，但是没有数据库企业。你觉得无经验的数据库研究者是否应该去企业，还是没有关系？

**David DeWitt:** 我不认为他们有太多的关系。我认为最重要的事情是挑选一个能够支撑起建立一个研究组的一个系。我认为可以在任何具有支撑力的学校建立一个强大的研究组。我认为我们【在威斯康星】已经展示了在我国中部可以公平有效地做数据库研究。我认为有比【东西美国】海岸还要多的强大的组。

**玛丽安:** 讲到建立一个数据库组：大多数学术系都有一到两个从事数据库研究的人，但是在威斯康星，许多年里，已经有 5 个或者更多的做同类研究的人。随着人数的增加，是否会有一个质的改变，还是会和以前一样？

**David DeWitt:** 我认为一个组有 4 或者 5 个成员会有一定的优势。如果你看威斯康星的系的设置，我们已经试着组织 4、5 个人一组。我认为两个人一起会出现一个周期性问题的：时而他们不想一起工作，时而他们想一起工作。但是当这 5 个人是 Mike、Jeff、我、Raghu 和 Yannis，如果再加上 Miron，会有很多种排列组合方式进行工作。所以 5 是一个有趣的数字，并且比 2 要好得多。总之，我认为除了量的不同之外还有质的不同。

**玛丽安:** 你已经领导完成了学术界的很多较大的软件项目，这不是经常发生的：很多钱，很多人。你怎么样把精力分散到做文章和生产软件产品两类工作中？

**David DeWitt:** 我认为做论文和做软件本来就是一个整体事件。我认为我们先看一下我最近的两个项目，Paradise 项目和 Niagara 项目。Paradise 项目——如果你看到每美元所发论文数，

就会觉得糟糕透了。我们生产了一个伟大的软件产品，很高兴做这件事情；实际上有 25 个人为了这个做出了努力，其中包括学生，也包括全职工作人员。所以这个项目对于学术圈来说太大。但是，发表的论文很少，每篇论文的花销确实很高。但是在 Niagara 项目中，生产一个可靠的软件产品很困难，但是我们已经发表了很多论文。我认为你不能计划它；你只能顺其自然，并且有时你会获得好的软件产品，并且有时你会获得很多论文，并且有可能同时获得两个，就像在 Gamma 项目发生的。但是 Gamma 项目是个例外。

**玛丽安：** 在那种项目中，论文和软件产品，哪一个更有影响力？

**David DeWitt：** 主要是依赖于你试着出售什么。

**玛丽安：** 好吧，在这三个项目中，你试着出售什么？

**David DeWitt：** 在 Gamma 项目中，我们试着指出概念的证明。最终 Naughton 和我成为了开发这个软件产品的最终的编程者。Jeff 设计了实验，并且和我写了代码。我不是一个很好的编代码者，并且直到我们写完代码，软件产品还不可用

**玛丽安：** 但是他证明概念了吗？

**David DeWitt：** 它证明了概念。项目的论文和软件产品哪一个更有影响力，主要依赖参与该项目的学生以及这些学生的能力。有时，你会有一些好的想法转化为好的软件产品，但是有的时候这些想法是坏的想法。我认为你需要利用现在所拥有的，并且开启学生的潜在能力——他们的软件开发能力和他们的做研究能力。

**玛丽安：** 稍微的偏一下，对于专业，什么应该进入引导性计算机科学课程？

**David DeWitt：** 孩子，这是一个很好的问题，很重要的问题。实际上我们正在查看这个问题，因为如果你看到在引导性计算机科学课程上的女生数量（至少在威斯康星），是很低的。并且问题是，为什么它是如此低？

**玛丽安：** 它有多低？

**David DeWitt：** 我认为在引导性课程上有 22% 的女生，但是主修的却只有 10-15%。

**玛丽安：** 我所在的系比这还低。

**David DeWitt：** 并且这个问题为什么会发生？我不知道为什么。我有两个女儿，一个主修化学，另一个主修数学，没有一个上计算机科学课程，虽然数学专业将来会用到计算机科学课程的内容。所以一定有某些事情使得高校女孩子不去上计算机科学课程，尽管这会使他们具有做计算机科学的完美能力。我不知道为什么会这样；我不知道是不是因为计算机科学被认为是男人主导的，或者是被认为是枯燥的。我认为问题的部分原因是我们先教编程。并且我认为对于大多数人来说，编程是枯燥的，并且编程不能完全体现计算机科学的发展。如果你思考化学，化学先从无机化学开始，定量分析只是引导课程中很小的一部分；引导课程的大部分还是无机化学。我们的引导课程可以先教他们一些体系结构，一些理论和一些数据库系

统。但是不用一步到位直接到数据结构和编程。

**玛丽安：**那么它们是否是实用课程？

**David DeWitt：**或许是，或许不是。我不再认为编程是计算机科学的一个巨大部分。我认为一定存在一些你可以做的计算机科学相关的事情，并且不需要你有太多的编程技能。我认为我们应该从不同角度去尝试，看看是否它们是否影响了这个领域的女人数量。当然，这样做可能没有效果。

**玛丽安：**那么你在威斯康星已经开始尝试这种新的引导性课程方法了吗？

**David DeWitt：**没有，但是我们有一个课程委员会，主要由低年级教员组成，他们刚刚获得他们的博士学位，没有更高级的人了。在最近的 25 年内我们没有改变我们的课程。引导顺序看起来还是和以前一样，什么课教什么内容。所以我们已经试着让这些低年级教员思考一些不同的想法，并且尝试这些新想法。我们将会尝试这些不同的想法在引导课程中，使其带有少量的编程，而把编程编入正式课程中。

**玛丽安：**有趣。我期待听到好消息。

**David DeWitt：**实际上，我希望其他人也思考尝试一下，如果可以的话，我们会复制他们的课程。

**玛丽安：**你可以作为我们其他人的先驱者。

很多年以前，你是最先流行的数据库测试标准之一的作者中的一个，测试标准是 Wisconsin 测试标准。关于这个测试标准，你是否有故事要告诉我们呢？

**David DeWitt：**不在录像带上。

实际上，那是一个很有趣的经历。很多人都关注它。同时它也使得很多人，包括一些朋友，都很生气。我记得 Mike Stonebraker 就因为我很抓狂，因为我们指出了 Ingres 不能很好的处理一些特殊的查询。我认为很多人开始追求性能结果，而不是拿到结果并且用结果分析他们的初衷，也就是说，我们的系统能做什么不能做什么

这有一个例子，Larry Ellison 非常非常生气——我猜这是最好的故事——试图让我被解雇。他不明白任期的概念，也不明白系主任不会开除我是因为我不会说 Oracle 正面的事情。但是我想了又想，测试标准在圈内已经发挥了很好的作用。我认为这可以帮助开发者聚焦他们的关注点。综上，我认为整个测试标准的努力对于圈内的发展起到了积极的作用。

**玛丽安：**你在暗示教授不应该去做测试标准，除非他们有固定的任期，是吗？

**David DeWitt：**（笑）是的，我没有明确暗示这点！悲伤的事情是每个数据库产品【我相信除了 DB2 以外】都有一个条款在它里面，基本上就是 Wisconsin 测试标准的结果，可以说除了运营商之外不可能发布数据。我认为这很糟糕。我认为这对于工业团队来说是一个很愚蠢的态度。如果你卖一个产品，人们应该能够评估这个产品。用户评估运营商产品使得数据库运营商有了各种各样的恐惧。

**玛丽安：**但是运营商发布的测试标准结果经常是自立地被审计。

**David DeWitt：**不是的，他们从来不被审计。有许多规则，运营商在报告他们的测试标准数字的时候必须遵守的，但是我认为大家普遍同意没有客户像运营商那样去做测试。

**玛丽安：**好吧！那个一定是真实的：一个测试标准（运营商发布的结果）保证了你的性能不会超过发布的数字。

**David DeWitt：**那是当然，那一定是上线。

我认为这个约束允许运营商关注一个特殊的数字，是否它是 TPC-A 或者是-B 或者是-C 或者是-D 或者是-H，并且它从整体上影响了这个团队或者用户，因为用户不能进行他们自己的评估并且发布他们自己的评估结果。那就允许运营商只需要关注一个单一数字并付出努力就可以，而我认为这是错误的事情。

**玛丽安：**好吧！你可以发布，你给你的数据库系统叫做 A、B.....

**David DeWitt：**.....C 或者 D。是的，这是标准托词，但是仍然奏效。

**玛丽安：**对于学术界，启动发热是一件好事还是一件坏事？

**David DeWitt：**我认为他仍然是一件坏事情，很多很棒的学生没有好好的研究下去获得博士学位。同时它也是一件好事情，很多学者已经做得很好。我认为总体上来说它是中性的。我不得不说它已经伤害了博士的质量。

**玛丽安：**如果发热持续下去会怎么样？

**David DeWitt：**我认为很好。现在每个人都想持续下去并获得一个博士学位。我认为会有一个摇摆并且学生获得硕士学位离开学术圈之后会变得更加保守。我认为近十年来这对学术圈有好处。

**玛丽安：**在美国，最近的经济衰退怎么样？你觉得这对学术圈会有什么样的影响？

**David DeWitt：**我认为同类事情会发生。我认为将会有越来越多的应用面向毕业高校。我认为吸收毕业生会更好。我认为他们会待很久。我认为这就意味着我们将生产出更多高质量的博士生，并且有望使得更多的学生对学术感兴趣并且继续从事学术研究。

**玛丽安：**可是学术资金怎么办？付钱给这些学生。

**David DeWitt：**我认为现实问题是政府经过9.11时间之后是否还有能力资助那些需要资助的事情？并且是否会对更基础的研究产生长远的影响？我认为如果你在安全中，那么你就应该高兴你在安全中。如果你在数据库系统中，它可能是件好事，因为他们将要不得不管理很多信息。问题是政府是否能承担得起，这点我不知道。数据库系统，和信息管理，在政府试图收集更多的信息过程中变得越来越重要。同时出现了隐私问题，这是我们不得不担心的。我认为这对于数据库研究团队是一件好事情。

**玛丽安:** 继续前面的数据库资金问题。我知道你是一名美国国家科学基金会（NSF）的CISE咨询顾问委员，并且CISE是数据库研究方面的NSF资金的主要来源，就像其他研究领域一样。你认为NSF应不应该资助人民，或者资助一些特殊的研发项目？

**David DeWitt:** 我认为他们应该资助尽可能多的项目。有时资助人也是很好的。有时提议比较窄。但是我认为你需要能够资助新的教员，所以有时你需要资助一些建议。资助人的确是完美可行的。

我不认为CISE顾问团对CISE所做的有过多的影响，所以人们不应该凭空认为我对谁获得资金说了很多话。

**玛丽安:** 那么你给了他们什么建议？

**David DeWitt:** 不管我们对他们说了什么，他们从来不听取我们的建议，所以没关系。我不确定为什么CISE有一个顾问团，因为我认为我们的建议被一次又一次的忽略。

**玛丽安:** 你说你认为NSF CISE应该资助更多的项目，但是你也说你认为一些项目计划太窄。

**David DeWitt:** 如果你说你想做X方面的工作，并且X实际上很宽泛，我认为很难得到该种项目的资金支持。一个典型的策略是做调研，然后写一个计划书——我认为这是不成功的。我认为人们应该能够说，我想在更宽泛的领域做工作；并且那可能就是我所说的资助人。

我认为这整个资助状态，即使是NSF的ITR，都很令人失望。在过去的一段时间里，有一个项目叫做系统实验研究（CER）；政府补助从上个世纪70年代晚期开始，每年花掉1百万美元左右，并且你能真正用一部分钱去做一些重要软件开发。现在，你可以得到的最大的ITR资助是每年1百万美元的范围内——并且20年已经过去了！你现在获得的每年一百万的投入产出要比过去的还要少。我认为这真的很不幸。从我个人角度来看，我认为CISE把太多的资金投入到了超级计算机和T数量级和网格计算和其它所有UIUC获资助的研究。

**玛丽安:** 确定我们是一个那种基金和研究的温床。

**David DeWitt:** 我认为那种基金没有资助计算机科学；他资助的是物理学家，并不是资助计算机科学家。

**玛丽安:** 很好！它在资助我。那个和我的安全方面的工作，这些你提出的课题。。。

**David DeWitt:** 好好。我认为付出了太多的钱。我认为构建2000个结点的集群并且声称它是计算机科学，这些都是废话。

**玛丽安:** 你不想它们能够模拟原子能工厂吗？

**David DeWitt:** 我认为那是在资助物理学家，而不是资助计算机科学研究。

**玛丽安:** 奥。好，我认为为了能够模拟原子能工厂他们需要很多帮助，因为写这种模拟器很困难。

**David DeWitt:** 我认为浪费了很多钱在超级计算机这个名字上。我认为PACI项目就是一个很好的没有很好的充分利用资金的例子。

**玛丽安:** 你是被采访者，你问这个问题，并且那是我所感受到的。

**David DeWitt:** 不，好好。是否有一些你所希望的特殊的東西是由PACI产出的，或者那些你所考虑的全局方向.....

我认为资助大块设备（比如说超级计算机）很好，因为你需要具有国家中心，比如伊利诺斯州【NCSA】，以及匹兹堡【超级计算机中心】和圣地亚哥【超级计算机中心】；我认为你需要有超级计算机中心，在那里人们可以脱离政府实验室来完成他们的计算。但是我不认为你应该绑定设备资助和研究资助。那是我对PACI的一点自己的看法。我认为它就是在试图绑定设备资助和研究资助及应用资助，并且我认为他们应该是三个独立的部分。实际上我比较喜欢匹兹堡的模式，把设备资助和研究资助独立开来，而不是作为一整块进行总体资助——因为我认为这样更容易说明。

**玛丽安:** 我本打算问你为什么，可是你已经告诉我：更好的可说明性。

**David DeWitt:** 对于资助代理的更好的可说明性

**玛丽安:** 所以你的意思是说，举个例子，如果他们在构建大设备时成功了，他们可能称整个项目取得了一个成功，即使.....

**David DeWitt:** 我不认为在构建大块设备的时候有什么研究工作。你可能只需要买一大堆机器，堆到计算机房，然后把它们聚在一起，把它们连接成网络。我只认为购买硬件属于资金使用范畴。明显地，如果你买硬件，那么你就应该支持维护硬件；但是你应该不必需要获得授权的人来决定哪个研究项目应该被资助。我只是不喜欢那种模式，这就是为什么我退出PACI项目。

**玛丽安:** 我知道，很有趣。

**David DeWitt:** 我不认为SIGMOD社团会对那个感兴趣。

**玛丽安:** 好吧！只是我对它比较感兴趣。那是我生活中的全部。【对于SIGMOG社团，它似乎显得很无聊，】我们可以立即从访谈录的打印版里删除这部分。

数据库研究的传统核心领域不再像以前一样被资助。是否这就代表我们这个领域比较成熟了，或者说，是否我们已经错过了一些需要更多研究的核心领域？

**David DeWitt:** 我认为我们已经错过了一些需要更多研究的核心领域。

【首先，让我们说一下】我认为这个领域已经很成熟了。我们现在已经有很能干的系统，并且这个领域也应为完成这样一个系统而自豪。我认为学术研究者和工业人士都做出了突出贡献。这些系统即可靠又可扩展，同时提供很高的性能。我认为我们就这个领域已经做了一个相对较完美的的工作，并且每个人都应该因为这个而自豪。

【但是，】我认为有很多核心领域【需要更多的关注】。我认为查询优化就是一个很大的漏洞；同时我认为IO也是一个大漏洞。我认为很多人已经进入这些热门领域。While是一个递归查

询处理，无论是面向对象数据库，无论是数据立方；因为Jim【Gray】写了一篇很好的关于数据立方的文章，那之后，我们发现300多人写了关于数据立方的文章。现在我们还有数据挖掘，KDD会议就有700多人参加。我认为人们着迷于那些热门领域——这很好，因为我认为只有一小部分人对核心问题感兴趣。

对于核心数据库研究，只有很少的资金资助。美国防御高级研究项目代理（DARPA）已经对这个不感兴趣很多年了；DARPA现在也对数据库没有投入，虽然这可能改变，并且NSF也不感兴趣，所以几乎不大可能获得资助做这些核心研究。

**玛丽安：**你说查询优化需要很多研究，那么查询优化的哪个部分需要更多的工作？

**David DeWitt：**整个查询优化！查询优化已经有22年的历史了。每个人都在做同样的事情，所有的工作都是基于Pat Selinger和系统R团队所做的工作，但是不能好好工作。数据库系统已经变得很强大。现在我们的数据库系统用户可能要去做10路连接，我们可能运行TPC-H查询（具有难以置信的复杂度的查询）于可扩展机器上的大数据集之上。如果没有手动调优，想为这些查询产生可靠的较好的计划，查询优化器实现起来就会表现的很糟糕。我认为我们需要重新考虑我们怎么进行查询优化，因为数据库其它技术已经得到很大提高，而查询优化却没有得到提高。

**玛丽安：**对于我们应该怎么样做查询优化，你是否有什么特殊的建议？

**David DeWitt：**我知道关系排序来自于Ingres怎么做查询优化，它基本上采用在优化和执行阶段进行枚举的方式。现在则是先优化数据库操作，然后执行。我们完全是基于数据统计的荒谬假设来优化九路和十路连接查询计划。现实是经过数个查询之后，我们就没办法预测有多少个元组会被查出来。你不知道连接列的属性值是否有关系；你不知道你的直方图是否还准确——或许你根本就没有直方图。所以查询优化器在处理查询树中有5、6层的连接时会做一个理想化的假设。

我个人观点是我们需要重新审视一下我们该怎么做优化和执行。现在，我们先优化然后执行。取而代之，我认为我们需要观察一些事情，比如说，优化一点，执行一点，优化再多一点，执行再多一点。我们应该从不同角度去尝试，因为这是一个没有技术提高的领域。

但是，这不代表说Pat Selinger在他的工作里没有做出巨大贡献。当你写一篇论文并且那个可能结束这个领域，明显的它就会是一片超级论文，Pat就是一个超级巨星！但是现在我们能在执行方面做点什么，我们需要回去重做查询优化。添加直方图不能够解决问题。我不知道怎么做，但是这是一个我认为很重要的方向。

**玛丽安：**当你指出查询优化和IO时，你的意思是通过IO解决吗？

**David DeWitt：**我的意思是磁盘变得越来越慢。如果你实际地看过传输率，磁盘是变快了；但是如果你让容量除以传输率，你会发现磁盘实际上是变慢了。

一些人会提倡你所应该做的是把SQL处理器放在磁盘控制器中，创建一个智能磁盘。我认为那不会帮助我们解决问题；我认为智能磁盘看起来只是像一个旧的数据库机——处理器和磁盘绑在一起。

在威斯康星，我们正在试图寻找一种解决问题的方法：我们试图看一下是否我们能够做一个虚拟分片工作。这是一个很旧的想法；MCC的Bubba项目做过这个想法，并且把它叫做分解存储模型。这个想法是，如果你只需要一个表的一个、或者两个、或者三个、或者多个列，为什么你要读整个表？垂直分片能够使得硬件缓存得到充分利用；它使得压缩更容易实现；他可能极大的增加你所用的IO设备的效率。

明显地，作为数据库人，我们不能去改变磁盘的制造过程。我们不得不与商业磁盘共存。并且最近几年他们已经达到了半T，两年之后会出现1T；到2010年会出现几个T的磁盘。数据库不会像磁盘一样增长那么快，除非你处理图像或者视频。

总之，我只认为IO是一个大问题，并且现在运营商只生产不管问题，因为磁盘变得越来越便宜。或许有一些IO方面的问题会让我们感兴趣去做一做。

**玛丽安：** 还有其他什么领域，你想提醒大家需要额外关注的吗？

**David DeWitt：** 我肯定有其他领域，但是有两个是我现在正在考虑的。

**玛丽安：** 你有没有喜欢的热门领域？你是否愿意看到人们追求流行？

**David DeWitt：** 明显XML是一个热门领域。我认为对XML感兴趣的原因是数据库研究者在研究分布式关系数据库时遇到了失败，并且我认为XML很灵巧的，因为如果这个发生了，并且人们提供XML，并且他们的网站运行XQuery，那么你就可以考虑构建一个巨大的分布式系统。我认为那是令人兴奋的研究领域，数据库团队正在研究它。我认为在大规模上解决分布式数据库系统问题，对我们来说，在接下来的几年间会成为一个有趣的挑战。但是已经有很多人在研究XML和XML数据库。

XML数据库不是一个核心领域，但是我猜测他是一个热门领域。并且那会引出新问题：我们能否考虑语义，和人工智能团队一起处理这些内容？只有XML是什么也做不了的，为了能够智能的处理大量数据，所以需要集成其他技术。

**玛丽安：** 在数据库界，很多人都有一种感受，学生发表了比以前多的增量文章，因为如果它是一个增量文章就很容易发表到顶级会议上去，因为很容易处理在增量文章中审稿人提出的所有漏洞，并且学生不得不比过去发表更多的文章以找到一份较好的工作。真的是这样吗？增量文章真的越来越多吗？如果是这样，是否这也意味会有一个问题？如果有一个问题，我们该怎么解决？

**David DeWitt：** 我不确定有越来越多的增量文章。

我认为有一个基本问题，SIGMOD和VLDB论文评阅方式。我最近写了一篇文章，被SIGMOD拒绝但是被VLDB接收并被评委最佳论文。论文在两次投稿过程中基本上未做修改。现在，就有了这种错误，如果一篇文章被一个会议拒绝，但是同一篇文章被另一个会议认为是很好的工作。

我不知道调解处理怎么了。我认为论文接收或者不接收变成了一个随机事件。我认为我们需要引进一个反馈调解处理的循环，你首先应该投稿，编委会会审阅你的稿件并且给你一些反馈意见，并且给你一次机会去反驳它，直到编委会满意；或者我们需要多轮的处理。

我认为现在处理接收一篇文章是一个走运过程。我认为这对低年级教员来说很难。作为一个高年级教员，当我论文被接受时我也会觉得有点失意，并且和我的研究前景没有关系！特别是由于我是我们系的主任，所以院长设定我的薪水并且我不由同事来审阅，并且院长也不会看我是否有两篇VLDB拒接的文章。但是对于一个不是终身教职的年轻人来说，必须小心审视你所想的是否是好文章——理由不是很清楚。

**玛丽安：** 你提到的这个方式听起来很像期刊的审阅方式。你是说让SIGMOD转化成TODS一样的？

**David DeWitt：** 当然不是，因为TODS除了理论文章没有别的。

**玛丽安:** 我希望不是这样。

**David DeWitt:** 期刊处理是开放的，但是会议的编委会不是开放的。现在这个时间线完全是荒谬的。我们第一次在11月提交论文并且在六月发表它。粗略计算一下之间有8个月时间。我们都知道论文已经录入计算机了。从快照准备（复制）到生产的整个处理过程不是一个事件。有一个很长的窗口，从11月1号到3、4月份，在这段时间我们可以执行调解处理。他不像一个期刊的，因为它只有一轮审阅和讨论。你提交你的论文；你从审阅者那里获得意见；你有一个机会写反驳意见给调解员；并且你不需要修改你的论文。然后由委员会处理。

我建议选择这个，是因为我认为，委员会成员有时因为不能很好的知道这个领域而评阅这篇文章，或者他们会误解作者意图。我认为我们应该试图去尝试改变，因为在处理论文是否应该被接受的过程中存在着太多的不确定性因素。

我还认为我们应该接收比预期多的文章。一些人给了很好的讨论；一些人给了坏的讨论。我认为这不会伤害到我们，举例来说，SIGMOD有250篇文章投稿，可能收录75到100篇成为一个论文集，然后只挑出来25到30篇论文做会议报告。我认为没有必要逐一介绍每一篇被录取的文章。一些文章会做到比其它更好的展示。

**玛丽安:** 当你挑出25到30篇文章，你怎么知道你挑出来的都有最好的展示者？

**David DeWitt:** 我还真不知道。我刚才在想，让我们做一些改动！就像引导性计算机科学课程那样：我们已经做了很久类似的事情，让我们回到SIGMOD，从1979年到现在它都没有变化——让我们做点什么让它改变。

**玛丽安:** 如果SIGMOD有现在的两倍大，并且接收的文章也是两倍，这会有帮助吗？这是否会令录取过程变得随机性越来越小？

**David DeWitt:** 如果SIGMOD一年增长两倍，我认为会有帮助，或者如果VLDB在一些比香港更公道合理的地方，或者无论下一年有什么变化，还是有很长的路要走。

**玛丽安:** 好吧！当然对于每个生活在香港的人来说地域的选择是很公道合理的。

**David DeWitt:** 是的，对于生活在香港的人来说是公道合理的，但是对于生活在美洲和欧洲的人来说是不公道的。基于现在的系统来说，一年组织两次SIGMOD会议是很困难的，因为当前我们已经安排SIGMOD会议在不同地方。巨大组织机构已经具有行业展示，并且他们还雇用别人帮助他们运作行业展示。做SIGMOD和VLDB的编委不是那么困难。困难的是处理所有地方的组织安排。我认为我们有足够的这个领域的人使得我们可以在美洲每年召开一次额外的会议。

**玛丽安:** 听起来很有趣。

你是否有什么建议的话要送给没有经验的或者是刚从业的数据库研究者和从业者？

**David DeWitt:** 我认为我的建议不同于我给那些初级教员的建议。（作为系主任，不得不担心这些事情。）我认为重要的是挑选一两个方向，并且做出一个实际上很好的工作。我认为一个初级教员可以做得最坏的事情就是泛而不精。如果你想做数据挖掘，好，那就努力成为一个数据挖掘方面最厉害的人之一。不要试图做数据挖掘、数据立方、XML和内存数据库。挑选一两个方

向，把所有注意力都集中在挑选的方向上。

我的其他建议是不要太早的带太多的学生。我认为一个初级教员的学生数量最多的时候不应该超过3到4个，因为学生是很好的资源并且如果你有太多的学生，你不能以很简单有效的方式和他們进行工作。

**玛丽安：**实际上你有多少学生？

**David DeWitt：**太多了！我现在有7、8个，并且我在试图回到3、4个的水平。

**玛丽安：**7、8个博士研究生吗？

**David DeWitt：**大部分是博士研究生和少数的几个大学生。我开始越来越多的雇佣大学生。

**玛丽安：**他们有时可能有用。

**David DeWitt：**他们很有用

**玛丽安：**如果过去工作时可以做的一件事而现在还没有做的，那是什么事情？

**David DeWitt：**我没办法回答你这个问题。。。去游泳池好好地游泳？

**玛丽安：**如果作为一个计算机科学研究者，你过去可以改变自己的一件事情，那会是什么？

**David DeWitt：**我希望我有强有力的数学背景知识。我认为有很多东西我不明白，但是又希望弄明白。我本科学的是化学，所以没有上那么多的数学课程。我认为有一个整数集的研究我不能参加。这是我猜我希望我可以改变的一件事情。

**玛丽安：**如果你有这个背景知识，你是否会做更多数据库理论知识？

**David DeWitt：**可能吧！我不可能做这种工作。我不够聪明去做数据库理论工作。我有一个PODS论文，有时人们会拿那个取笑我。但是那个是我学生的论文，不是我的。

**玛丽安：**非常感谢参加我们的节目！

**David DeWitt：**谢谢你们邀请我！

（范玉雷译，富丽贞校）