高校大数据教育：基础知识结构与学位设计

主讲人：艾春荣 院长
中国人民大学统计与大数据研究院

大数据战略

大数据开发与应用被各主要工业国（美国，德国，英国，欧盟、日本）视为战略资源，希望以此为支柱产业，带动经济增长和人民福利（新产品、新服务、更低成本）的提高，并将大数据的开发和应用提升至**国家意志（高于国家战略）。**

我国亦将大数据的开发与应用作为国家意志，并多次发文推进大数据及相关产业的发展。

大数据价值

这一崇高的地位决定了，大数据的开发与应用必须促进经济的持续增长。也就是说，大数据的开发与应用最终必须提高生产率和社会福利，否则，用不着以国家意志来推动大数据的发展。

那么，大数据是如何提升生产率和社会福利的呢？ 我的理解是，通过解决社会经济中的信息不对称问题。

大数据价值

例如，在生产领域，大数据中的信息可以帮助我们识别生产过程中资源错误配置和低效的地方，帮助我们做出更高效的的决策。在消费领域，大数据中的信息可以帮助我们更多的了解消费者的偏好和需求，帮助设计新服务、新产品。在流通领域，大数据能帮助我们更有效的匹配供需双方，减少流通成本。在服务领域，大数据中的信息能帮助我们更深入的了解经济个体的行为，帮助我们设计对应的服务以及风险控制。大数据中的信息还是设计人工智能产品的基础。

大数据价值

因此，大数据的价值取决于对隐含的信息的挖掘、分析与应用程度。**它不只是数据的采集、储存和简单的挖掘与关联分析，而是对隐含的行为信息做深度分析的结果。**

数据处理 ≠ 数据分析

大数据的特点是量大但所含信息相对（数据量）很小，对信息的挖掘和分析离不开应用领域的知识指导和统计学工具。

大数据 ≠ 总体

最后，分析结果的应用离不开优化与算法。

大数据知识结构

可见，大数据的开发与应用涉及到数据的采集和储存**（计算机科学）**、数据的挖掘与分析**（算法、统计学、应用科学）**、分析结果的应用（计算科学、应用科学）。表面看，这一过程是上述领域的简单组合，但简单组合无法达到整体最优。

因此，**大数据教育的知识结构**必须是，计算机科学、统计学、计算科学与应用科学的深度融合。

大数据学位设计

基于这一认识，大数据学位设计（本科、硕士、博士）必须是系统的训练，但侧重点可以不同。

工学学位可以侧重计算机科学的训练，加强算法训练，了解统计学和应用科学。

理学学位了解计算机科学，强化计算科学，侧重统计学，熟悉应用科学。

大数据学位设计

本科学位主要培养实用型人才，知识面要宽，而不是精。因此，本科学位教育要介绍上述各学科的基本知识。

- **应用学科专业**背景的学生可以通过对计算机领域、统计学领域的课程的辅修而获取大数据应用的技能。
- **计算机学科专业**背景的学生可以通过增加优化与算法、统计学和应用学科课程获取大数据技能。
- **统计学专业或其他理学背景**的学生可以通过增加计算机科学、计算科学和应用学科领域的课程获取相关技能。

大数据学位设计

总之，无论是什么背景的学生都应该修下列课程：

- 数据科学基础（Foundation of Data Science）
- 数据科学概率基础（Probability for Data Science）
- 数据科学统计基础（Statistics for Data Science）
- 数据科学计算机基础（三门课介绍数据库，C语言程序设计，JAVA, MapReduce, Hadoop, Spark, R, Python等）
- 机器学习
- 数据科学实践

大数据学位设计

更专一点的课程，如：大数据采集与处理，流数据处理，计算机网络，优化与算法，数据挖掘，文本挖掘，数据可视化原理、分布式处理与云计算，数据结构，深度学习，金融大数据分析与应用，交通大数据分析与应用，政府大数据分析与应用，法律大数据分析与应用，新闻大数据分析与应用，工业大数据分析与应用等，可根据学生的背景和兴趣，师资力量酌情开设。

大数据学位设计

硕士学位仍然培养实用型人才，但因学制短，学生背景差异，设计课程只能以特色为主，突出某一方面，其内容可在本科培养方案的基础上增加点难度。

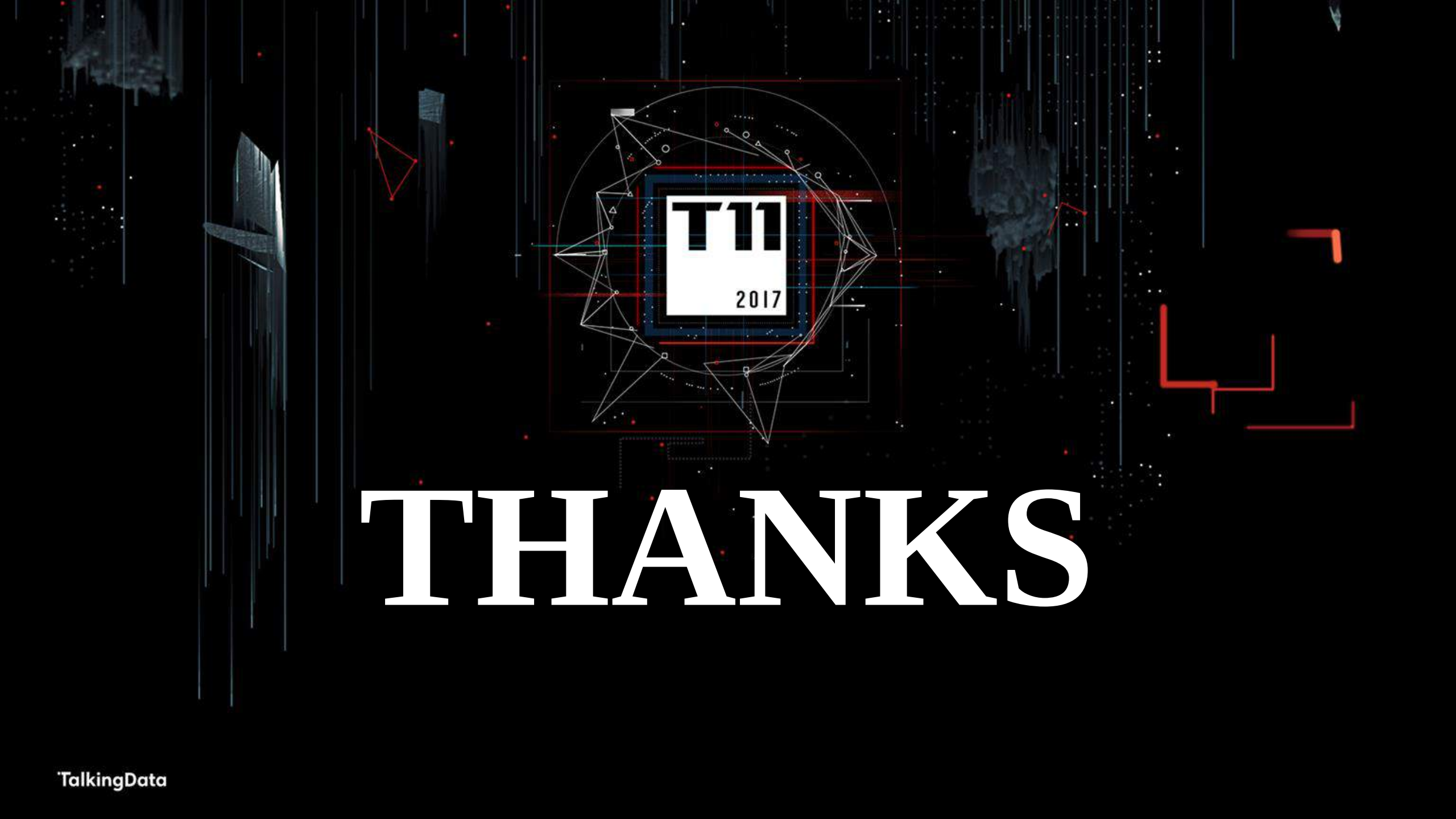
例如，如果对应用学科背景的学生，可以以统计学领域的课程为主，计算机领域课程为辅；对理科背景的学生，可以以优化与算法领域为主，强化计算机领域课程，了解应用领域；对计算机学科背景的学生，应以优化与算法为主，统计学和应用学科为辅。

大数据学位设计

统计与大数据研究院

博士学位是培养开发型高端人才，学制长，无论学生背景如何，除系统的学习外，应在更高水平差异化训练，重在开发和创新能力的培养。

例如，应用学科背景的学生，可专注大数据应用。统计学科或其他理科背景的学生科专注大数据分析。计算机背景的学生可专注计算机和算法。


T11
2017
THANKS
TalkingData