第六届中国R语言会议(上海会场)

R与大数据对统计教育的影响

林祯舜

先锋信息科技
中山大学数学学院/兰州商学院

上海
2013-11-02

关于这个报告

我已经好几年没有和统计学界的师长及朋友们交流了，我的研究领域已经较偏向将数量方法应用在市场营销的领域。会触动我在这个会议上报告的原因，主要是从业界的角度思考统计学的发展，毕业之后都和教育界有联系，在国内各大学校讲课，尤其是在大数据的背景条件下，我觉得有必要从一个非统计教育工作者(虽然我是统计学博士)的角度，给大家一些建议，也希望这些建议对统计学科建设有一定的帮助，略尽一份百年树人之志。

这个报告的题目是我目前与袁卫教授及傅德印教授一起合作关于统计教育的文章，目前还未完成，还请各位统计教育的先进们多提建议与讨论

林祯舜

- 先锋信息科技
- 网略智慧

人民大学统计学博士

吉林大学商学院市场营销系兼任教授

中山大学数学学院统计系专业硕士导师

兰州商学院统计学院兼任教授

专长/经验

技术专业领域：

- 数据挖掘
- 机器学习
- 统计计算
- 网站效果测量
- 贝叶斯统计

营销应用领域：

- 营销模型
- 营销计算
- 市场研究方法与技术
- 固定样本研究
- 顾客关系管理

行业经验：

- 邮购/型录
- 网站/电子商务
- 汽车
- 医药
- 快速消费品

文章发表期刊：

- Journal of Advertising
- Journal of the American Society for Information Science and Technology
- Information Research
- 营销科学学报

报告题纲

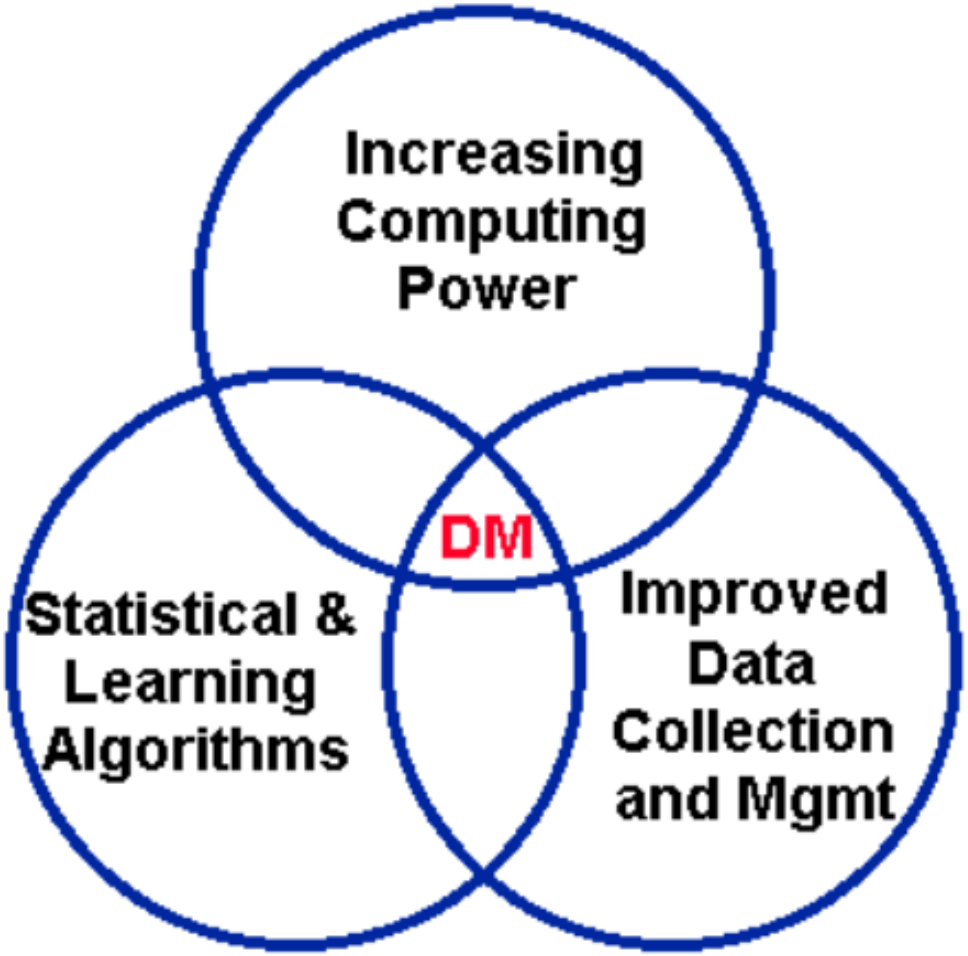
1. 从数据挖掘到大数据
2. 统计学科与计算机学科的边界
3. 大数据会让统计学科和其他学科的边界越渐模糊
4. 数据与分析话语权的转移
5. 分析技术及开发流程的碎片化，网络化与社会化编程
6. 大数据背景下，对统计人才的新要求
7. 对统计教育变革的建议-统计学科与其他学科之间的协作
8. 大数据背景下，对统计课程的的建议-培养数据科学家
9. 我的心得总结
10. 发展相关理论
11. 参考材料

从数据挖掘到大数据

从数据挖掘到大数据

Data Mining (DM)

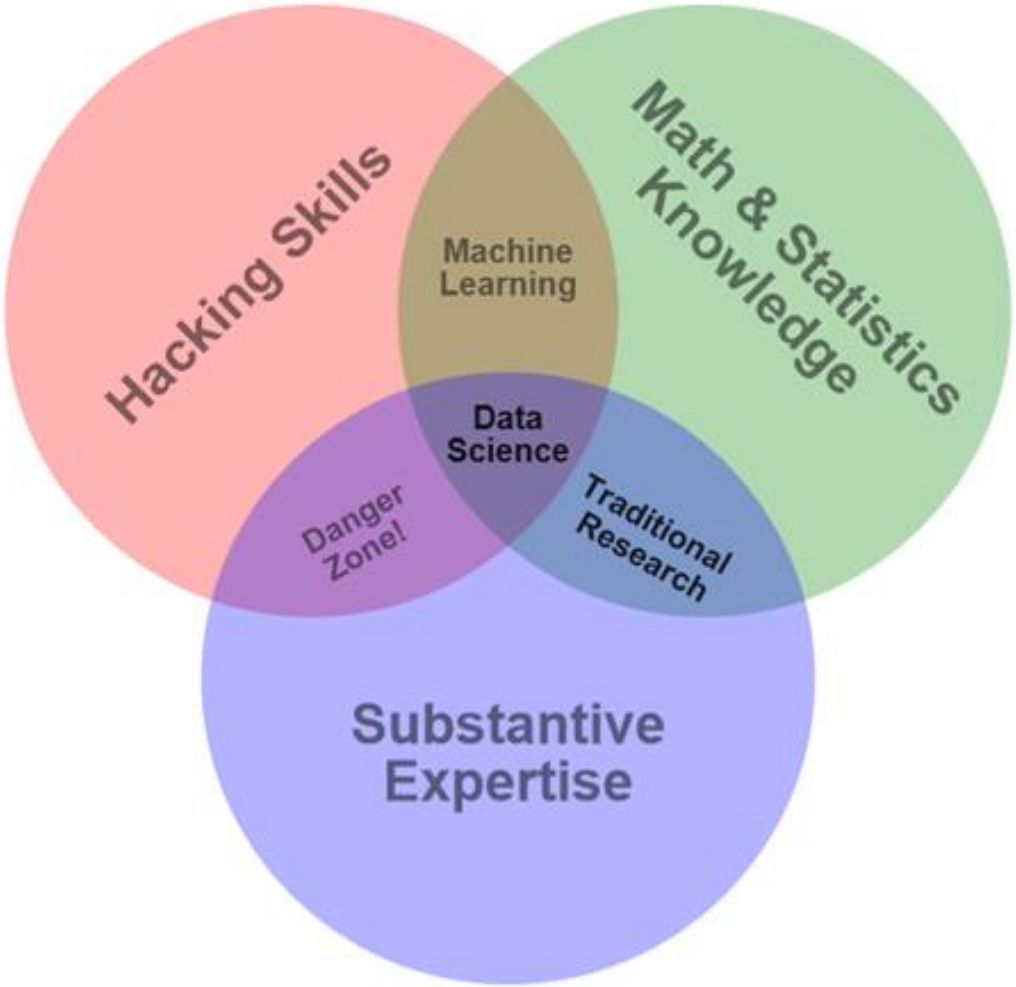
数据仓库
商业智能
OLAP



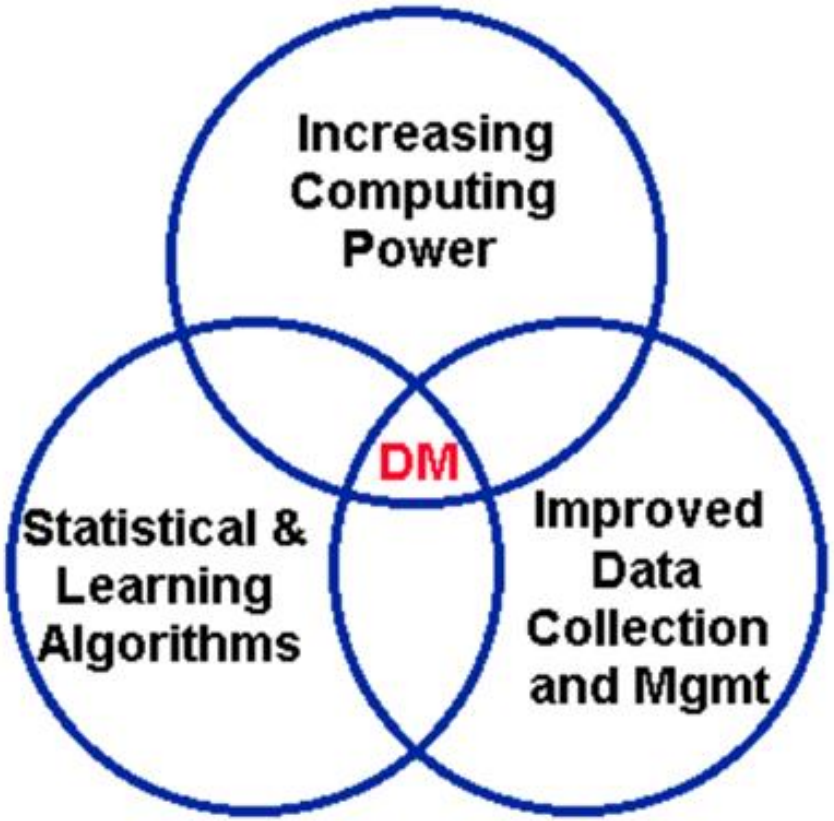
从数据挖掘到大数据

非结构化数据
平行运算
可视化

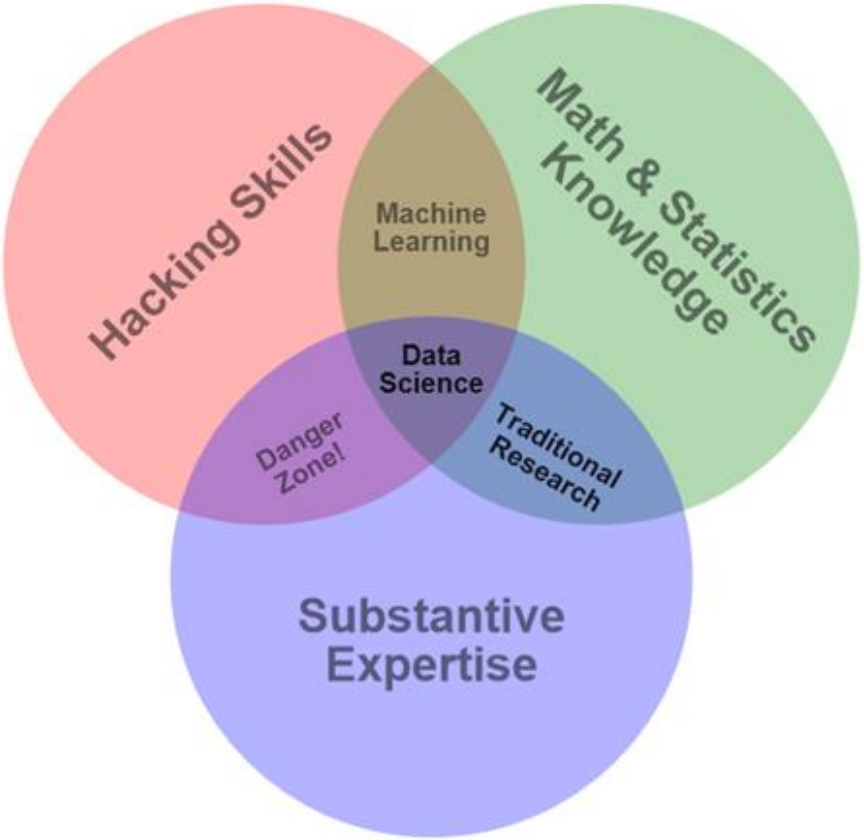
Data Science



Data Mining



Data Science



回顾过去**数据挖掘**的发展到**大数据**，整体而言，统计学科的**本质是没有变的**，**分析的核心观念**没有因为数据量的多寡而有改变，而是应用方面更强调交叉学科(学科间的协作)

从数据挖掘到大数据-影响

我依稀还记得在十多年前，”数据挖掘”这个时髦的名词出现之后，许多统计学家开始思考统计科学的课程体系(请参考2000年中华数据挖掘(资料采矿)协会成立大会的研讨会各个专家的发言及台湾辅仁大学统计系更名为统计信息系的缘由)，同时有更多的统计学家开始担忧由计算机科学家开发的数据挖掘算法及工具对传统统计数据分析方法及理论的**冲击**。

从数据挖掘到大数据-融合

然而十多年过去了，在现在回望这段发展过程及历史，我们发现当初统计学界的有识之士对于计算机学界的跨界而大声疾呼”狼来了”的忧患意识，以及担心会动摇统计根基的事并没有发生，相反的，这两个领域还互相学习、借鉴与融合。

数据挖掘的应用在计算机学家的努力开发下，将许多统计和机器学习的算法集成为方便使用的数据挖掘软件，让数据分析及数据挖掘更加的**普及**，对于海量数据在商业应用的相关观念能更快的被企业接受，而统计的观念及应用也得到**推广**。

从数据挖掘到大数据-其他学科借鉴统计学

然而，如果回望三十几年前，当时是统计学去影响并渗透进入其他学科，例如:经济学、社会学、管理学、心理学、流行病学等学科，而统计学是做为重要且基本的研究方法或工具为这些学科提供研究的分析基础。当时学科之间的边界还是非常清楚的，统计学只是做为基础的学科建立一定理论基础进而服务这些学科最后对该学科产生一定的影响。

当时统计学影响较多的学科都是容易收集到数据的学科，例如: 金融、经济、生物统计、市场调查.....等

从数据挖掘到大数据

现今，大数据这个名词又非常的火，“大数据”首先是指规模大且形式多样的数据，但又不仅如此。“大数据”还有“交叉复用”和“全息可见”两个特征。(周涛)

“交叉复用”和“全息可见”这两个特征会深深的影响统计学科未来的发展。

交叉复用强调的是数据处理及分析过程的弹性，全息可见会影响统计对于抽样、因果关系及精确性等传统统计基础观念。

从数据挖掘到大数据

因果关系和相关性

讨论因果关系或者是相关性，这只是路径选择的问题(或是典范选择的问题)，因果关系是top-down(从上而下)的方式，相关性是button-up(自下而上)的方式，以前自下而上在统计学界并不能够完全被大家所接受(因为没数据，以及学科的发展有一定的路径)。

从数据挖掘到大数据

以研究”关系”为例，计算机科学家有数据(最近数据量更大了)，所以直接用算法计算，讨论的核心重点是相关性，心理学家或社会学家没有数据，所以需要做实验(或调研)收集数据，讨论的核心重点是因果关系，强调在心理学理论上构建互相影响的因果关系，因为要讨论因果关系，所以需要有较理论的假设前提，也就造成应用方面假设太多，有许多限制。相关性是观察到的某种规律(现象)，但是这种规律能否上升到因果关系，就需要做更多的研究(归因理论)，以找到可能的原因。其实在研究发展的过程，许多领域是互相借鉴的。

从数据挖掘到大数据

我的观点是大数据也是一个媒体炒作出来的概念，对于推进算法的普及和分析型思维有很大的帮助。其实和十几年前的数据挖掘一样，这是这个概念的普及对统计学科很重要，因为公众及企业等外在环境的改变及重视，可以让我们能够思考改变。更可以让统计的核心思想得到更多的普及。

从数据挖掘到大数据

统计学家的观点

俱备定量思维的志同道合之士，总是能够解决重要的问题，这不是什么新鲜事，这一直是数据的作用，所以，这真的不是什么新玩意儿。但是从我们生活的方方面面，从宏观到微观所产生大量复杂的数据是新的玩意。我们的数据有来自政府，金融，教育，环境，社会福利，健康，娱乐以及互联网等产业，这些数据将来可以用来帮助制定政策，并发展成产品并回馈到我们的文化脉络之中。

从数据挖掘到大数据

业界的观点

数据挖掘是让用户感受及理解到工具和方法(算法)的价值，大数据是让用户感受到数据(体系)及分析思维的价值。

统计学科与计算机学科的边界

统计学科与计算机学科的边界

当初统计学家所忧虑的统计科学会被计算机科学”占领”(Dominate)的情况并没有发生，却产生计算机学科与统计学科相互融合并均衡的有趣的现象，这其中的原因，笔者认为主要是学科边界在当时还很清楚，几个原因如下:

统计学科与计算机学科的边界

- 算法被打包并集成在软件中，无法单独使用，因此让相关深层次应用的扩展受到限制，相关前沿应用普及速度缓慢。

- 数据挖掘或统计学多强调分析或挖掘”数”据，基本的数据形式都是数字为主(“数据’近似”数值”近似”信息”)且储存数据的工具为关系型的数据库，其实统计学家不关心数据库，只要数据能符合需要分析或建模的格式及要求即可。

统计学科与计算机学科的边界

- 统计学家只关心建模及相关假设，对于模型的布署的相关效率并不是太重视，以致于模型的能见度(被大家理解)及效率都不理想。

- 由于分析的过程已经被整合在数据挖掘的软件中，相关算法的灵活性及在商业中的相关应用有很大的局限性，算法及其应用还没从软件中解放出来。

统计学科与计算机学科的边界

- 当时的硬件成本还是太高，软件成本也高(虽然和二十年前相比，已经降低很多了)

- 工具的设计思路没有很大的进步，在旧的架构下，工具是封闭的，以致于统计分析并没有变成有效率的统计分析语言。而只是软件的概念。(例如: SPSS软件与语言的认知比例约为9:1，SAS软件与语言的认知比例约为7:3)

统计学科与计算机学科的边界

核心原因是数据或信息流在这两个学科的分工还很清楚。

多年以来，统计学家与产生大量数据，如:天文学，生物信息学和数据挖掘等不同领域的研究人员一起进行分析工作。以前每个领域的数据基本都是独立的，而且与该数据收集产生的学科有很强的关联，可以说每个领域的数据都是一个专业孤岛，只能在其直接关联的领域发挥自身的价值，而学科的边界也因为方法和数据的不同而界线明确。

大数据会让统计学科和其他学科的边界越渐模糊

未来大数据会让统计学科和其他学科的边界越渐模糊

大数据是不同的，因为它产生了人与人之间大规模的无法计算的在线互动(交互)关系，人与系统之间大量的交易信息，人与系统及传感设备之间的大量信息(包括人造数据、自然数据)。当数据之间的关联越来越多，数据的价值与外部性也就逐渐体现，这才是值得让人兴奋的所谓“大数据”，数据所含的信息从稀缺走向丰富，更可能泛滥。

未来大数据会让统计学科和其他学科的边界越渐模糊

我们要找到和实现数据之间一加一远大于二的价值，未来很可能经济价值、真理都在数据中，数据又带来很大的发展空间，当数据的外部性得以发挥及扩大后，数据将成为资本、人之后的第三个重要资源。

数据与分析话语权的转移

数据与分析话语权的转移

随着时间进入户联网时代，数据的存储成本更低，数据越来越容易收集及获取，数据的形式越来越多样(有数字、文字、声音、图形、视频)，处理数据的工具及技术越来越多元化，以前统计学所学的处理数据的技术已经逐渐显露出不足，而计算机领域因应不同的数据形式及需求，已经发展许多新的工具及有创意的观念。

数据与分析话语权的转移

在这段时间，统计的许多方法及核心观念已经深化入计算机科学的领域，随着需求的多样，更多开放的、有弹性又符合需求的开源工具逐渐的普及和流行，算法的模块化及方便打包让更多人可以使用，使应用及开发人员可以将关注的重点放在解决实际的应用问题上。

数据与分析话语权的转移

再加上数据量已经从海量到巨量，数量的大小呈几何级数的增长，计算机科学家在储存的方法及工具上已经发展出多种的解决方法，这意味着不管是要做数据分析、数据挖掘或是建模也好，最重要的数据来源以及数据的整理，这些所谓的数据操作能力(或权力)已经完全由计算机科学家掌控。

数据与分析话语权的转移

这意味着统计学家如果不加强对计算机的储存及计算的理解，一开始的数据开采权将由计算机科学家(大多数为研究数据库及并行计算)所掌握，其实以前也一直由他们掌握，只是以前工具较单一，计算速度不够快，统计学家介入还容易，然而随着NoSQL等数据库的增加，并行计算及云计算的普及，工具的多样性，统计学家如果不加强这方面的知识及动手能力，在未来会逐渐落后而逐步在数据话语权的博弈中屈居劣势。

分析技术及开发流程的碎片化，网络化与社会化编程

分析技术及开发流程的碎片化，网络化与社会化编程

更让人忧心的是，以前统计学家所自豪的算法及统计分析能力，在数据量增加，实时运算的应用要求越来越多的情况下，有许多算法已经不符需要，这也需要统计学家加以改正，由于分析工具及观念的落后，这个原本多属于统计学家的领域，已经逐步的被计算机科学家渗透，主要原因是统计学家在解决实际问题的过程中依赖统计软件而欠缺编程能力，如果加上对该应用领域的不熟悉，在解决问题的方面(编程、分析、解释)会趋于弱势。

分析技术及开发流程的碎片化，网络化与社会化编程

然而从数据分析的角度看，由于数据越来越大及越来越多样，处理这种类型的数据已经不是统计软件能解决的，这需要编程的语言，这种统计编程语言已经在计算机及统计等领域逐渐被大家接受并使用，随着开源统计编程语言的普及，统计分析语言已经增加编程及开源的观念，而不是封闭式的分析软件，许多的分析问题已经不再是打开软件，用鼠标点选几个分析选项就能完成的。

分析技术及开发流程的碎片化，网络化与社会化编程

然而从编程的角度看，计算机科学家在这方面更是得心应手，而传统统计教育仍在强调软件的使用，而不是编程能力及对计算机编程的语言的理解。虽然这是统计科学和计算机科学的学科训练造成的差异，如果将分析(或数据挖掘)看成是一个循环的过程(process)，这个过程每一步骤的工具及方法都在改变，统计学家也需要解放思想，与时俱进，多吸收计算机科学的长处及方法，虽然这些多是工具性的技能，毕竟这是数据分析的第一步，而模型的建立毕竟还是在数据获取及整理之后才进行的步骤。

分析技术及开发流程的碎片化，网络化与社会化编程

在软件开发速度越来越快的今天，新的开发工具不断推陈出新，数据已经逐渐的隐藏到产品背后，有许多的应用其实背后的核心是数据，但是用户看到的却是一个软件或应用，因此未来统计学的训练也要让学生具备走到前端开发数据产品的能力。

大数据背景下，对统计人才的新要求

大数据背景下，对统计人才的新要求

现今，在大数据的浪潮下，统计教育如何自处及变革，以因应未来的变化?在学科交叉的环境下，现在的外在环境已不像十年前，统计学家发展算法及理论，计算机科学家将这些算法编成软件。

大数据的核心是分析，而分析是关于人才能力的培养

大数据背景下，对统计人才的新要求

未来的情形是计算机科学家凭借着对大数据储存及处理的优势，首先接触数据，再加上开源的统计编程语言，已经让计算机科学家对大数据的整理分析从而理解数据并具备深入分析的能力，从而完成以前是由统计学家完成的事，如果要将分析的成果开发成小工具或应用并集成在系统中，这又是统计学家不擅长的事。从分析的全过程来看，统计学家在

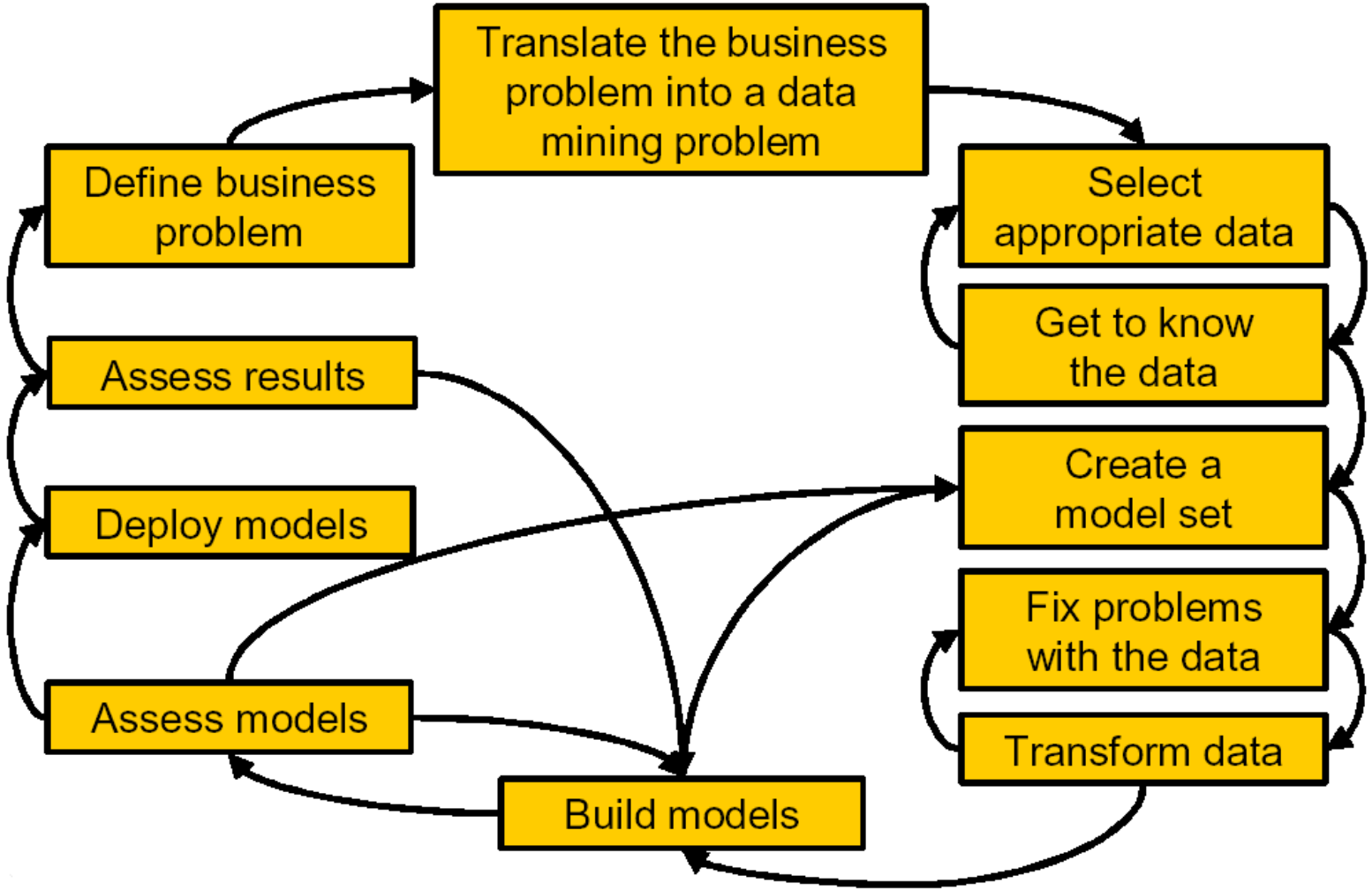
1. 数据的获取方面
2. 分析全流程各个步骤的技术方面

都明显的落后于计算机科学家

大数据背景下，对统计人才的新要求

这是归结于统计学家的传统养成过程中，并未将分析视为一个过程(虽然在数据挖掘领域一直强调数据挖掘是一个流程)，而只是将分析视为一个核心步骤，所有的统计方法大多都专注且集中(围绕)于这个核心步骤，以前数据量小的情况下，还没体现出不足，现在数据量大了，这个想法就限制了统计的发展。

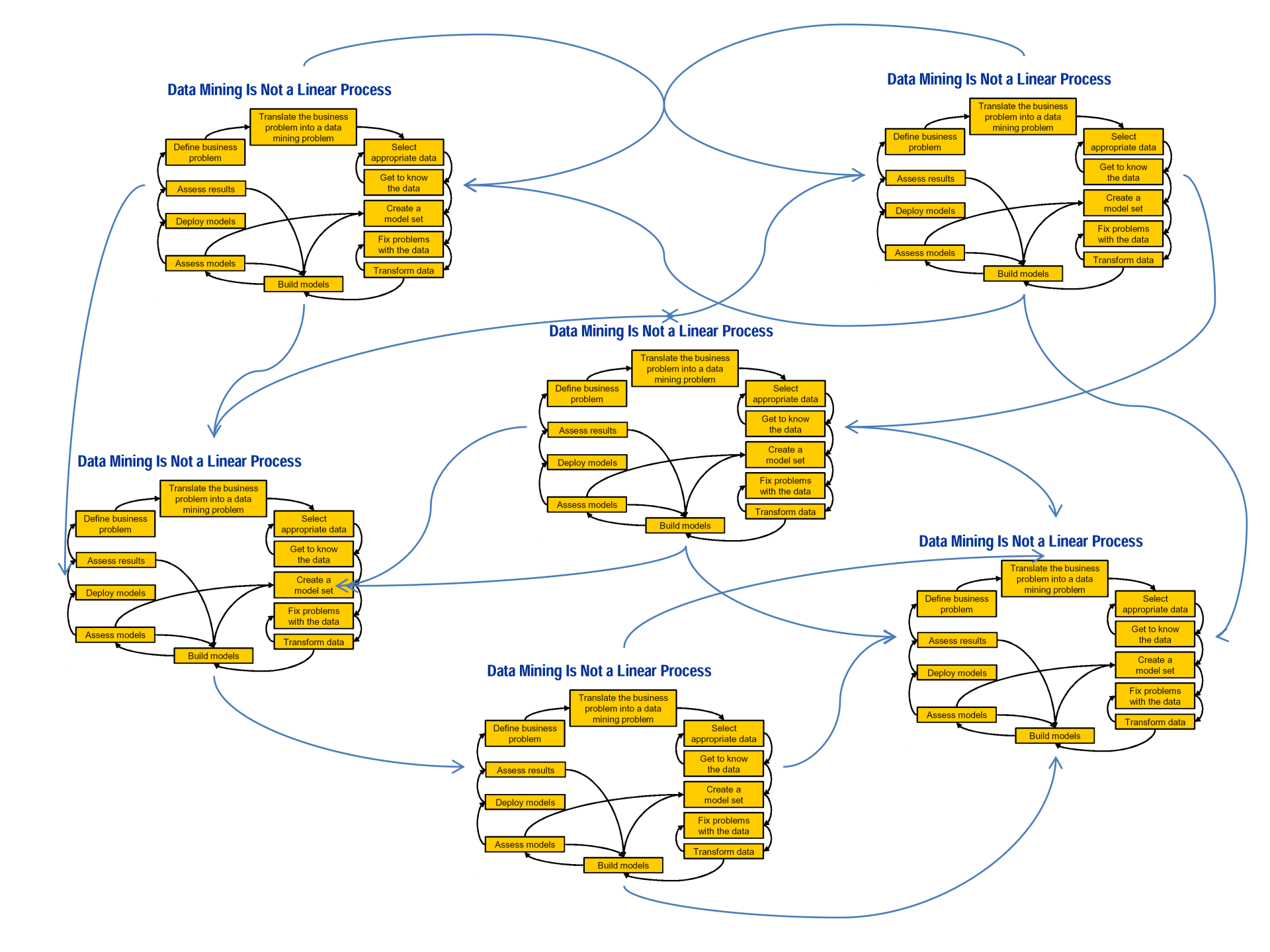
Data Mining Is Not a Linear Process



来源: SAS数据挖掘培训材料

大数据背景下，对统计人才的新要求

总的来说，其实是统计学的养成教育中，缺少了一种全局的观念(全局观)，这个观念的改正，需要引入工程学的观念，将分析的全过程视为一个工程的流程，就像建一座房子及桥梁一样，分析也是一个工程(尤其是未来处理大量数据和工具及需要与设备打交道)，将分析的过程像设计工程一样，分成许多步骤流程，再将流程按照分析的方法逐步整合完成，最后完成一个”分析工程”。


Data Mining Is Not a Linear Process
Translate the business problem into a data mining problem
Define business problem
Select appropriate data
Get to know the data
Assess results
Create a model set
Deploy models
Fix problems with the data
Assess models
Transform data
Build models
Data Mining Is Not a Linear Process
Translate the business problem into a data mining problem
Define business problem
Select appropriate data
Get to know the data
Assess results
Create a model set
Deploy models
Fix problems with the data
Assess models
Transform data
Build models
Data Mining Is Not a Linear Process
Translate the business problem into a data mining problem
Define business problem
Select appropriate data
Get to know the data
Assess results
Create a model set
Deploy models
Fix problems with the data
Assess models
Transform data
Build models
Data Mining Is Not a Linear Process
Translate the business problem into a data mining problem
Define business problem
Select appropriate data
Get to know the data
Assess results
Create a model set
Deploy models
Fix problems with the data
Assess models
Transform data
Build models
Data Mining Is Not a Linear Process
Translate the business problem into a data mining problem
Define business problem
Select appropriate data
Get to know the data
Assess results
Create a model set
Deploy models
Fix problems with the data
Assess models
Transform data
Build models
Data Mining Is Not a Linear Process
Translate the business problem into a data mining problem
Define business problem
Select appropriate data
Get to know the data
Assess results
Create a model set
Deploy models
Fix problems with the data
Assess models
Transform data
Build models

大数据背景下，对统计人才的新要求

在这种思路下，统计教育就不应只强调概率论、统计理论、多元方法、回归分析、时间序列、统计计算等传统课程，当然这些课程还是很重要，只是在新的思路下(将分析看成是一个工程)，传统的方法只是训练学生分析不同类型数据所需要的基本技巧，这也很重要。

大数据背景下，对统计人才的新要求

但是在数据多样化及海量数据的发展下，统计学科的训练也要重视分析过程中其他的能力(例如：数据的存储及整理、计算的效能、分析结果的工具化及可重复性、分析结果的展示及自动化等)，让这些能力均匀的分布在统计学科的养成过程中，进而让学生能有较好的分析全局观及具备动手解决问题的能力和技术。

大数据背景下，对统计人才的新要求

不能将软件的操作理解为学会统计技术，而是要真正的会相关算法，并深入思考相关理论应用，养成学生思考问题的习惯。

对统计教育变革的建议-统计学科与其他学科之间的协作

对统计教育变革的建议-统计学科与其他学科之间的协作

由于网格运算系统在数据中心和云计算服务中普遍存在，许多统计学家会看到伴随着数据量的不断增加进而对分析能力的期望也跟着提高。因此我们需要新的数据管理技术和新工具进行数据分析和可视化。因为有这么多的数据的来源，如手机，社交网站，和健康记录，我们还需要采集和分析非结构化的文本数据的方法。
如果要符合未来潮流的发展，统计教育需要做一个变革，而变革的方向需要在课程设计中加强学生如下的基本技术技能并与其他学科协作

对统计教育变革的建议-统计学科与其他学科之间的协作

数据处理的能力(数据库及处理海量数据的能力)

其实大数据的概念中，有一个数据多样性的概念，这个概念在以前的统计学中叫杂乱的数据(Massive Data)只是以前统计学家不重视，因此这个词还是被计算机科学家发扬光大，最后衍生出大数据的观念。另外未来处理实时的数据对计算机科学家而言也是一个很大的挑战，统计学家也需要介入并具备处理这种应用需求的能力。

对统计教育变革的建议-统计学科与其他学科之间的协作

实现分析算法及编程的能力，而不是依赖软件及拿鼠标操作软件的技巧

这要加强计算的能力，传统的统计理论的训练有强调这一环节，但是局限在小量的数据，现在要将这种计算能力拓展到能够处理大量的数据。另外编程能力的加强也可以在数据分析后软件化及自动化的实现提供一定的技术训练，这样统计学家和计算机科学家才有共同的语言，进而开启统计学家了解计算机科学的一扇窗。

对统计教育变革的建议-统计学科与其他学科之间的协作

加强数据展示的能力

数据获取容易后，大家有更多的精力能专注在数据分析后的解释，数据的展示是很重要的环节，以前在统计的训练过程中，总是觉得这很简单，但是随着数据的多样及数量的增多，如何将成果展示并说明已经不像三十年前统计学家认为的那么简单，在User experience的发展下，信息的有效展示也逐渐在信息科学的领域成为显学。

大数据背景下，对统计课程的的建议-培养数据科学家

大数据背景下，对统计课程的的建议-培养数据科学家

在大数据的环境下，统计学家需要思考未来统计课程需要培养出那些人才?这些人才是否需要有不同的类型?需要具备那些能力?这些问题，从最近的发展来看，我们可以理解为对未来数据科学家的培养。

大数据背景下，对统计课程的的建议-培养数据科学家

对于数据科学家有共识的特性是数据科学家们必须专精在统计建模和机器学习，擅长编程技能，并能扎实的掌握核心问题等专业知识的有创新精神的问题解决者。

数据科学家应该拥有哪些能力才能顺利完成任务？不妨把他们想成集”数据黑客”、”分析师”、”沟通者”、受信赖的”顾问”于一身的人。这样的组合，力量极为强大，而且难得一见。

大数据背景下，对统计课程的的建议-培养数据科学家

下一代数据科学家的各种技能，包括编程，统计，机器学习，可视化，沟通，数学。

下一代数据科学家不会迷信(执着)于工具，方法或学术理论。他们是多才多艺的和跨学科的。

心得总结

总结

从思想上和技术上都解放出来后，对于统计学课程体系的变革就有较清晰的方向及思路，这个报告先不牵涉具体的课程设计，因为每个学校可以根据自身发展的重点，在上述的原则下调整，只要大方向是对的，不管将统计科学应用在那个方向，都会培养出符合未来需要的前瞻性人才。在学科的发展过程中具备一定的战略高度。

总结

其实在大数据的浪潮下，我们相信这两个学门还是会和十几年前数据挖掘很火的时候一样，在未来大数据的应用需求下，彼此找到自己的定位，在这个过程中，两个学科互相成长融合，为科学的发展提供新的方法和应用方向。

回到本质，统计教育的变革需要依循数据的变化而调整，其实数据的变化也是社会变化和科技进步的体现。所以统计教育变革的核心战略是数据。

建议-避免分析话语权的转移

统计需要与其他领域的学者多交流，

统计需要走出去，然后引进来

从工程学的角度，是统计嵌入其他学科，这个概念就类似嵌入式系统或软件的观念

已经不像从前是将整个统计思想及方法论应用在别的领域(别的领域从统计学科借鉴统计方法)

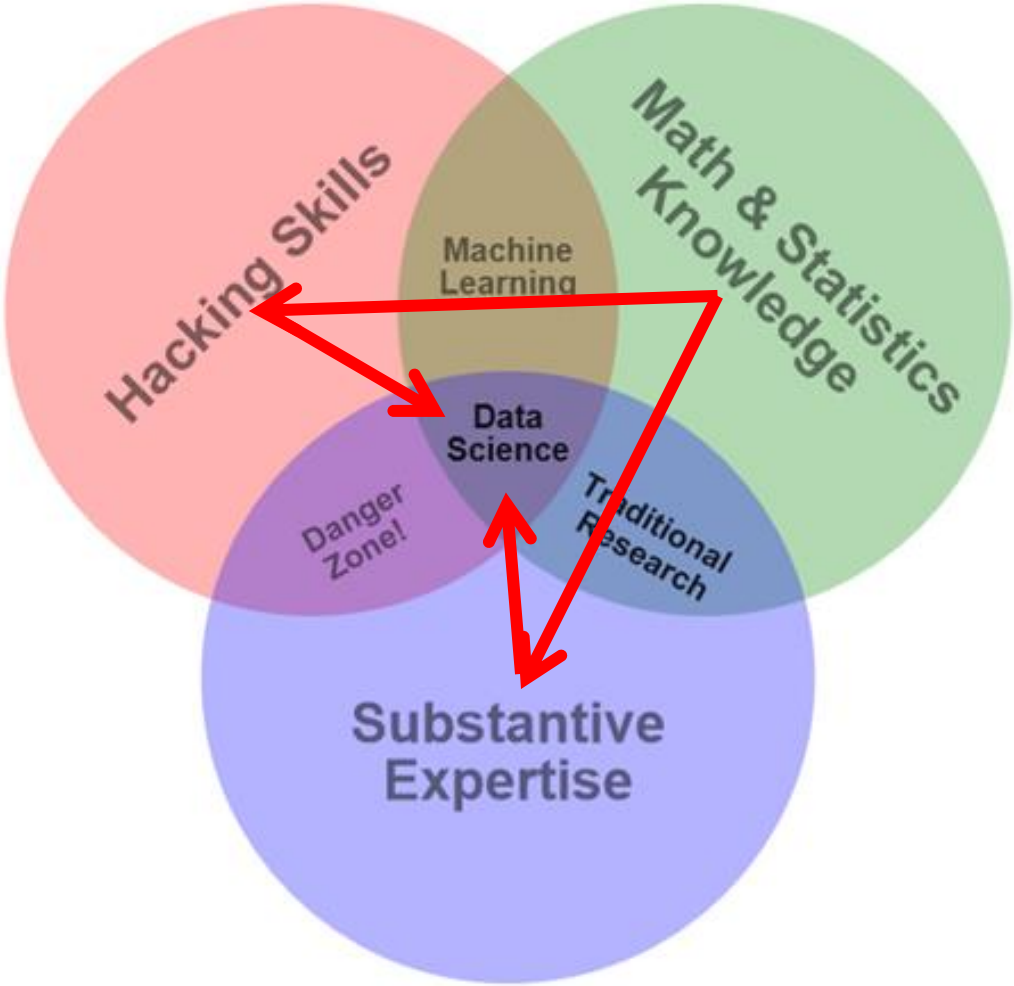
建議-避免分析話語權的轉移

统计学需要借鉴其他领域学科的方法，以前都是别的领域借鉴统计方法，未来统计学也要学会借鉴别的学科的方法。为什么?

因为大数据的特性。

计算机科学学系多是打群架(以研究群或实验室为单位)，统计学系多为单打独斗。因为统计学科发展出来的方法多是被别的学科拿去用，如果再不借鉴其他学科的思想并多和其他学科交流，以后统计学科可能会被拆分，嵌入其他学科。

Data Science



发展相关理论

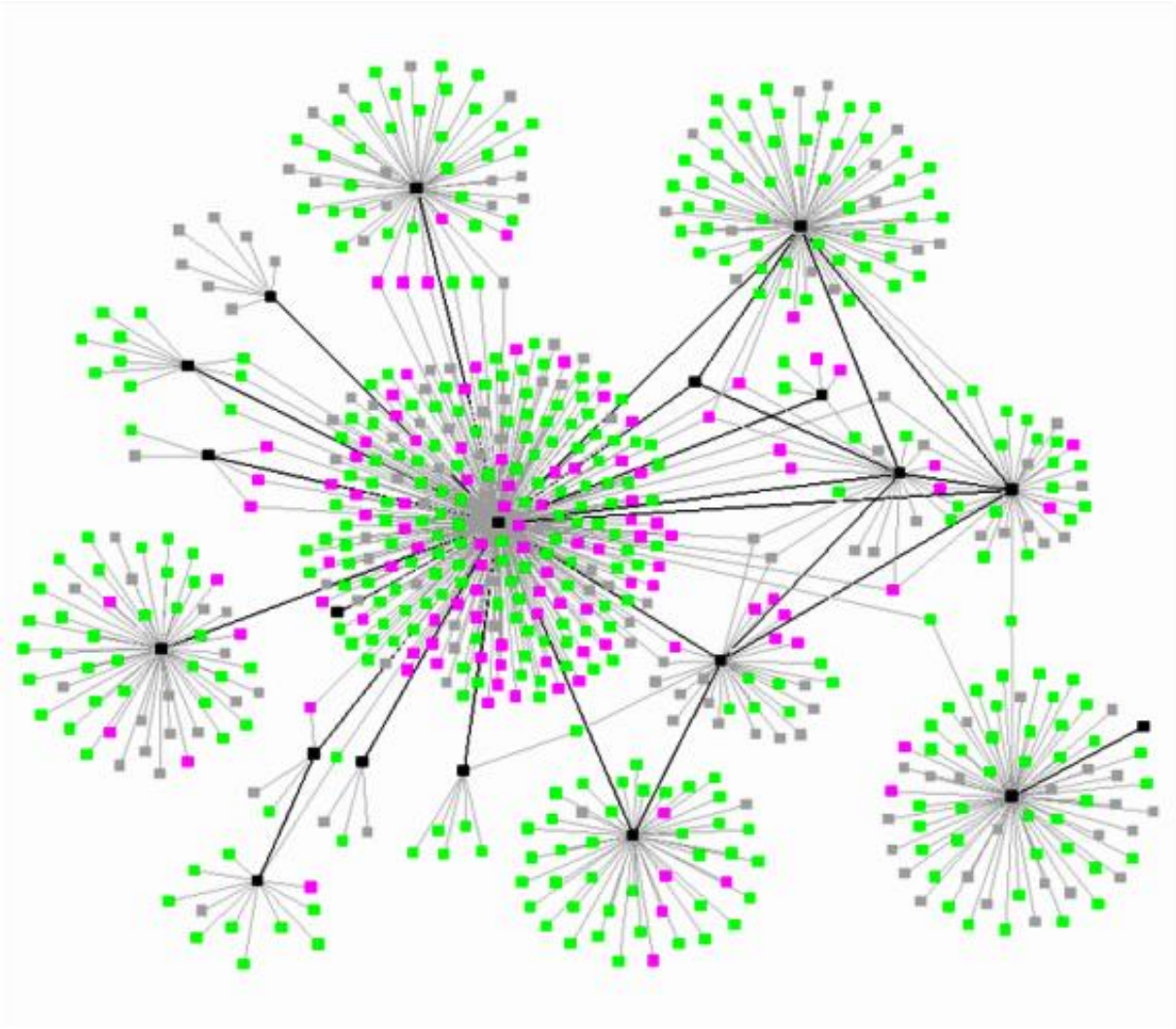
镶嵌 Embeddedness

我们需要从网络连结的角度思考统计方法的Embeddendness

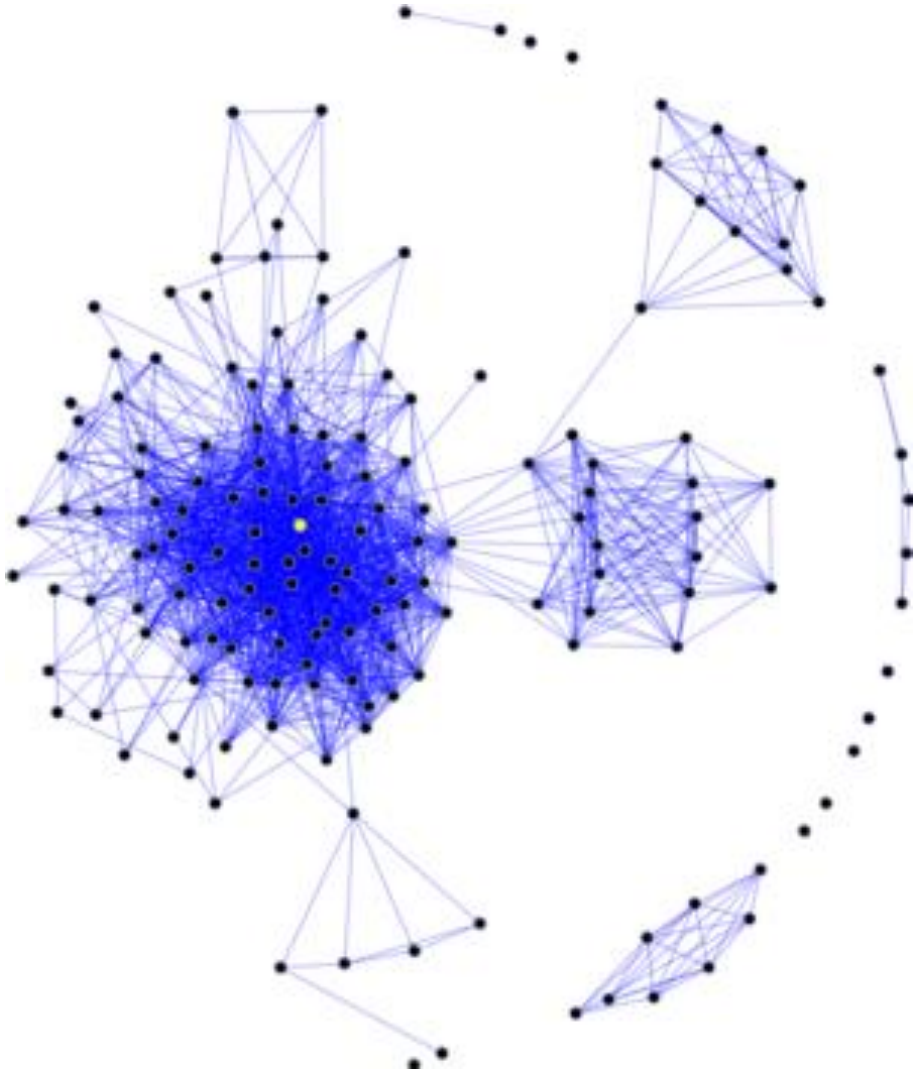
从多层（阶层）Multilevel到网络Netowrked的发展

鼓励统计系的学生多和其他领域的学生交流，学校和学院要多创造条件，特别是和计算机科学的领域

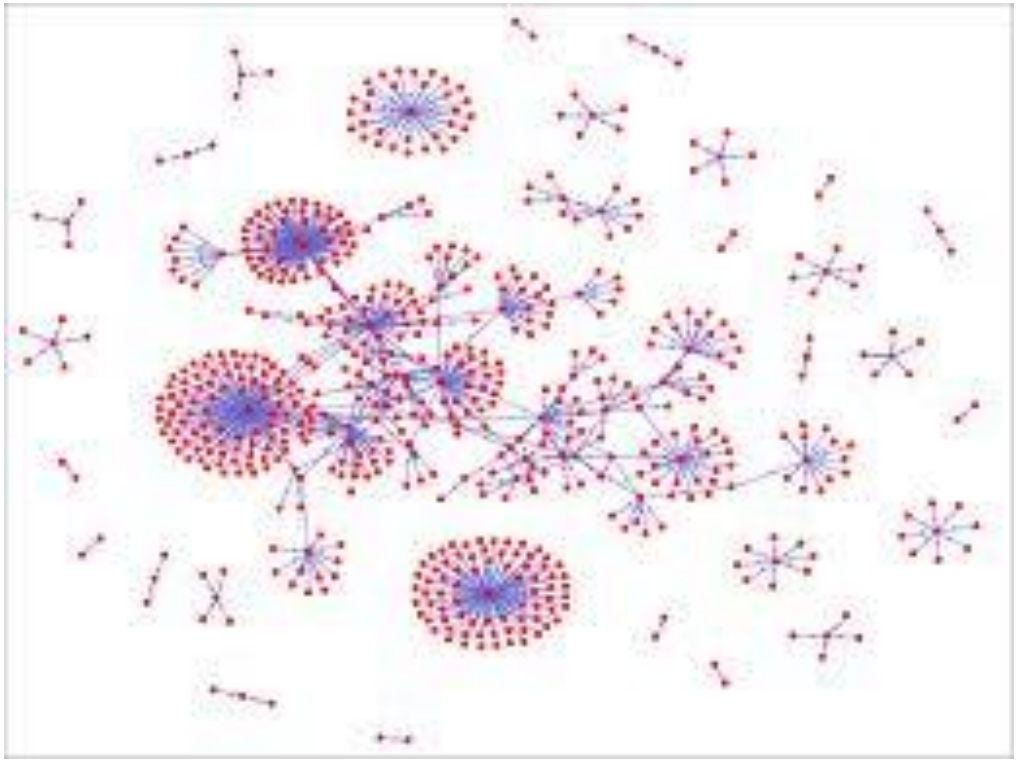
镶嵌 Embeddedness



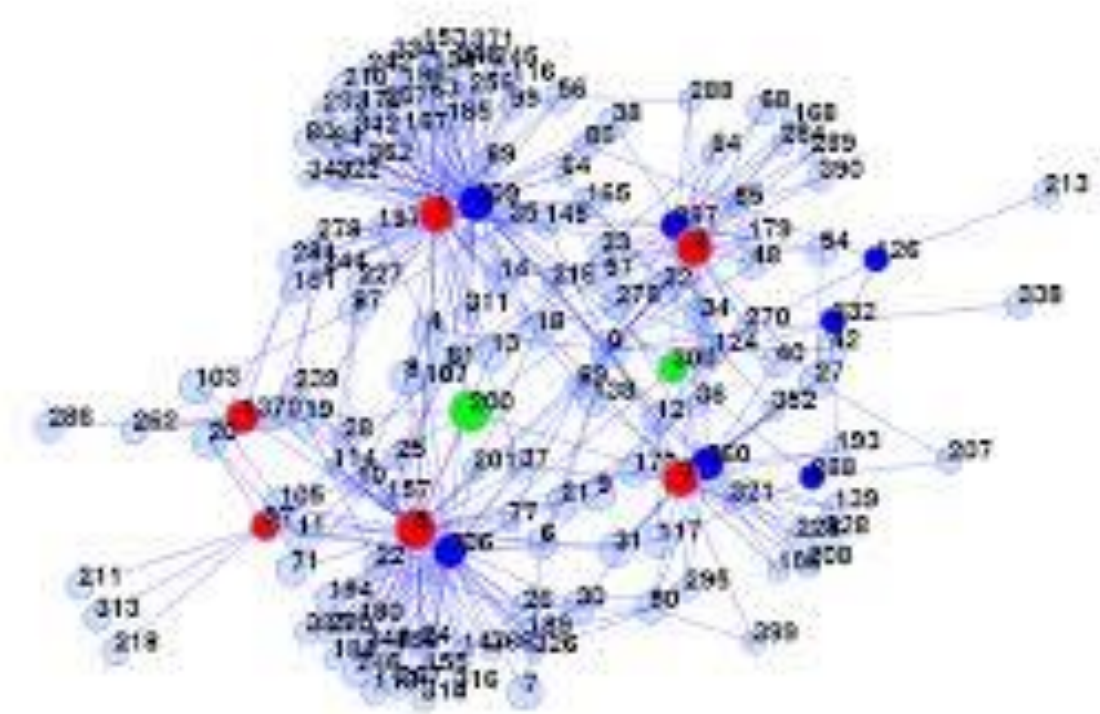
镶嵌 Embeddedness



镶嵌 Embeddedness



镶嵌 Embeddedness



最后一个问题留给学界思考

那些是因应未来发展需要的统计学科的核心理论?

Can Big Data Motivate New Theories and Methods?

大数据可以激发出新的理论与方法吗

参考材料

参考材料–参考书

1. 大数据时代
2. Strength in Numbers - The Rising of Academic Statistics Departments in the US ,编者Alan Agresti, Xiao-Li Meng
3. 驾驭大数据
4. 大数据：正在到来的数据革命

参考材料–参考文献

- ✓ Wolfgang Pietsch , Big Data – The New Science of Complexity
- ✓ Hal R. Varian , Big Data: New Tricks for Econometrics
- ✓ Michael Jordan , On statistics, computation and scalability
- ✓ Michael Jordan , Big Data: The Computation/Statistics Interface
- ✓ Alexander W. Blocker and Xiao-Li Meng , The potential and perils of preprocessing : Building new foundations
- ✓ Jianqing Fan, Fang Han, Han Liu , Challenges of Big Data Analysis