数据人生

台大资讯(信息)管理系教授 陈文贤

本文以 隐喻（Metaphor）的方式说 数据科学，写给 台大资管系 三十年系刊。

1. **数据科学**
2. **数据江湖**
3. **斜杠老人**
4. **数据模型**
5. **数据料理**
6. **元宇宙**

1. 数据科学

信息(information)的原料(原始材料/料理食材)是资料(data)或称数据。数据科学(data science)需要有三个领域的知识：统计演算知识、计算机科学知识、产业专业知识。产业专业知识是领域知识(domain knowledge)。机器学习是要有统计演算知识和计算机科学知识。危险区域是包括黑客、电商等变动很快的知识，会有错误判断的危险。数据科学虽然是这三个领域的交集，实际上是要包括这三个领域的知识。

2. 数据江湖

金庸《笑傲江湖》说：「只要有人的地方就有恩怨，有恩怨就会有江湖，人就是江湖。」

恩怨改为 数据：「只要有人的地方就有数据，有数据就会有江湖，人在江湖。」

武侠小说是在写江湖传奇，通常的故事是：主角经过奇遇如灵丹怪兽，遇到师父传授功力招式，得到武功秘籍，学成武功，然后快意恩仇，行侠仗义，消灭恶徒，称霸江湖。

大数据(big data)的江湖故事是：企业得到珍贵数据，机器学习数据挖掘方法，获得信息、知识、智能，创造市场份额和优势，打败竞争对手。

《笑傲江湖》将华山派武功分为剑宗和气宗，剑宗是注重剑法招式，气宗是注重气功内功。大数据分析、数据挖掘、机器学习就是大数据的剑宗。

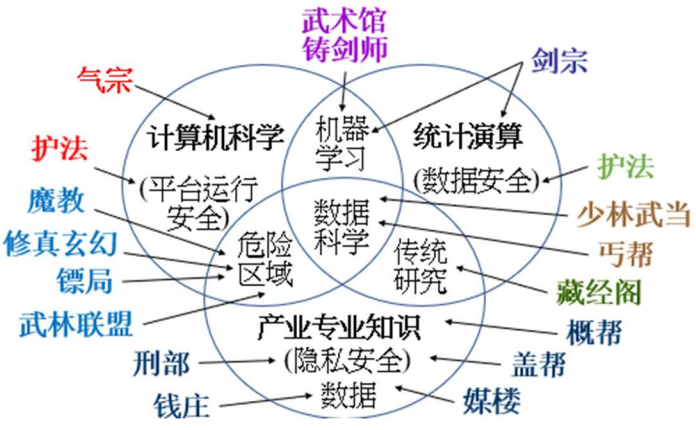


图 1 数据江湖（《大话数据科学》 图 1.23)

以下是大数据的江湖门派：

1. 华山派剑宗 --- 大数据分析的招式，数据挖掘十大算法是独孤九式。大数据分析的分类、回归、聚类、关联规则等是数据挖掘机器学习。

2. 华山派气宗 --- 大数据技术，处理程序，计算框架，存储文件系统，分布式并行处理，Hadoop, MapReduce，Spark 等。大数据的计算能力，就是内功。

3. 铸剑师 --- 大数据分析的函数与程序包(package)，以 R 语言和 Python 语言为工具。R 语言的程序套件的开发者，Python 应用在神经网络、深度学习等平台的开发商。

4. 武术馆 --- 大数据分析平台。Google 的 Tensorflow，FB 的 PyTorch，Alibaba PAI 机器学习平台。铸剑师和武术馆有要付费的和免费的分享平台开发者。

现代大数据平台不只是武术馆，而好像是武器馆，只要会：选择武器如手枪(模型)，会装子弹(数据)，会瞄准(调参)，会扣板机(指令)，检查命中率(验证)，就可以杀敌(应用)。于是：手枪原理(模型理论)，弹道理论(算法过程)，装拆手枪(程序设计处理)，是黑箱可交给专家或学术机构(铸剑师/武术馆)处理。

5. 少林武当派 --- 中国 BAT：百度、阿里巴巴、腾讯；美国 FAANG：Facebook、Amazon、Apple、Netflix、Google，这些可以说是大数据的少林武当派。

6. 丐帮 --- 数据和程序共享平台。R 语言是开源免费共享平台，R 提供 2 万个以上的套件，这些套件就像是丐帮的大小分舵，有数据有算法函数。而 Python 语言的框架，一样是免费共享平台，例如 Tensorflow 有谷歌的支持，就像是少林武当的大寺庙。

7. 概帮 --- 大数据概念帮，介绍大数据应用在医学、保险、零售、会计、工业、制造、农业、金融、电商、地理、运动等各行业。多数是概念，纸上谈兵。

对于概帮，我们要问：大数据的问题种类，数据来源，数据类型，分析方法，模型和算法，信息结果，验证评价，应用价值，这是 CRISP-DM 跨行业数据挖掘标准过程。如果无法回答上述问题，就是概帮。当然，有些概帮是因为商业机密，无法提供这些说明。

8. 盖帮 --- 在台湾“盖”是骗人、唬弄人的意思。盖帮的分析结果是常识，不用计算就已经知道的结果，或者是编造诈骗的结果。数据科学的计算结果应该是未知的、潜在的、可理解的、有价值的、和有用的信息。

9. 藏经阁 --- 大数据相关书籍和出版社，大数据案例探讨。

10. 媒楼 --- 大数据的宣传机构，帮助 盖帮宣传、炒作(如虚拟货币或 NFT)的媒体。

11. 魔教(邪派) --- 制造假数据，窃盗数据。

12. 修真玄幻(穿越) --- 虚拟货币，区块链。

13. 镖局 --- 大数据保护，数据安全。

14. 武林联盟 --- 电子商务，共享平台。

15. 钱庄 --- 大数据存储，云计算。

16. 刑部神捕司 --- 大数据执法的政府机构，维护国家和个人隐私安全，个资保护。

17. 护法 --- 门派内大数据的安全保护，大数据平台运行安全，企业的法务部门。

大数据目前没有倚天剑、屠龙刀 (武林至尊，宝刀屠龙，号令天下，莫敢不从！倚天不出，谁与争锋？)。没有一个天下无敌的招式，没有一个招数可以打败所有的武功。天下没有一个药方疫苗可以治百病防千毒。大数据没有一个模型(或算法)可以解决所有的数据分析。所以，应用数据挖掘，每个方法都有优点缺点，有适用环境和范围，实战需要经验和商业知识。

大数据和武侠世界有一点不同的是，武侠的内功(气宗)是基本功比较不会变，剑招(剑宗)是会改变的，要讲“无招胜有招”是有些过分。相对来说，大数据的气宗(计算机技术)比剑宗(数据挖掘技术)容易创新改变的，因为计算机科学的技术(量子计算机计算，不只是武侠而是仙侠)，可以说是日新月异。数据挖掘已经有二三十年的历史，是因为网络和计算机技术才有大数据。

三四十年前的算法求解，因为计算机的速度和储存能力，所以斤斤计较于计算的复杂性(Computational complexity)。现在用分布式并行处理，就可以解决很多计算的问题。所以，因为计算机的快速能力，使得以前 统计学、数据挖掘、人工智能(记得有 AI 之冬)，无法处理的模型，现在可以用训练和验证数据解决。这就说明了武侠小说的一句话：

天下武功，无坚不摧(数据)，唯快不败(气宗)，唯准能胜(剑宗)，唯狠无敌(无友)，
唯义称王(共享)。

3. 斜杠老人

我求学从数学系到工业工程所，教职到退休经过 法商学院、国防决策所、企业管理系、工业管理系、资讯(信息)管理系等，经历下列学院：

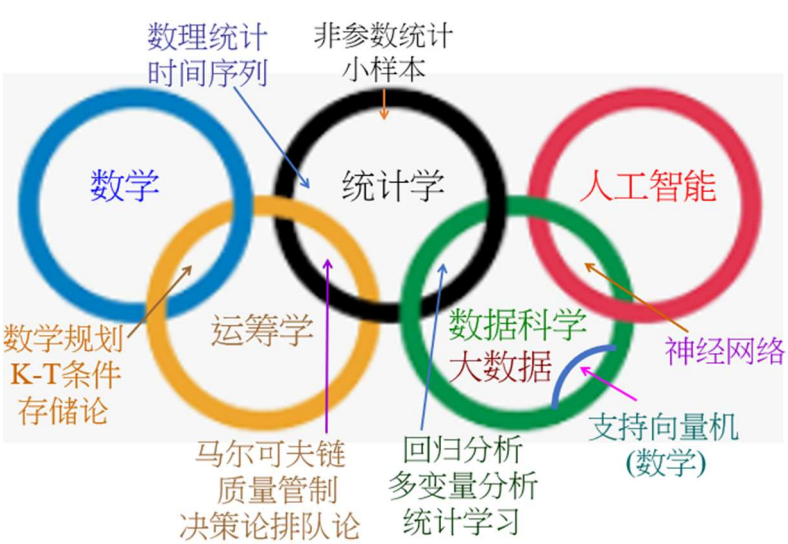
理/工/法/商/军/管理/信息/医(未来大体老师)。

教过下列课程：

管理数学/生产管理/统计学/作业研究(运筹学)/信息管理/电子商务/网络营销/电子化企业等。

出版下列书本：

《资讯管理》 (2002 年) / 《管理科学》 (2010 年) / 《统计学》 (2012 年) /
《大话数据科学 ：R 语言》 (清华大学出版社 2020 年) /
《大话统计学 ：R 语言 + 中文统计》溢彩实训版 (清华大学出版社 2022 年 4 月) /
《运筹学 ：R + Python + 运筹学 2.0》 (预计 2023 年出版) /
《人工智能 ：Python》 (预计 2024 年出版)。



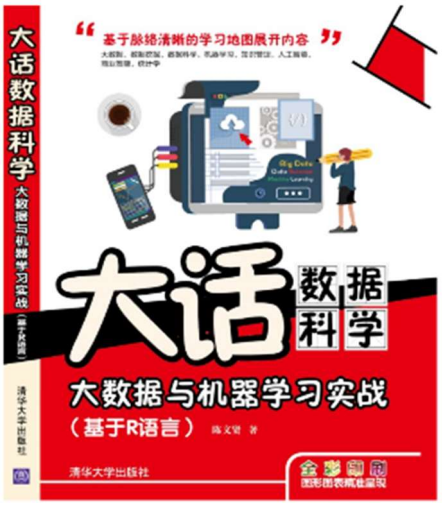


图 2　数据学科的关联图

4. 数据模型

数据科学除了数据的取得很重要，求解的方法也很重要，求解方法就是模型(model)，模型不是唯一的。利用数据科学模型，要注意是否符合假定条件(assumption)，不要削足适履非穿不可，不要因为「这个」方法比较熟悉、比较容易用，就要用它来找答案，结果找到的答案根本不对。统计学通常是抽样数据的模型选择，数据科学有训练和验证数据的模型评价。

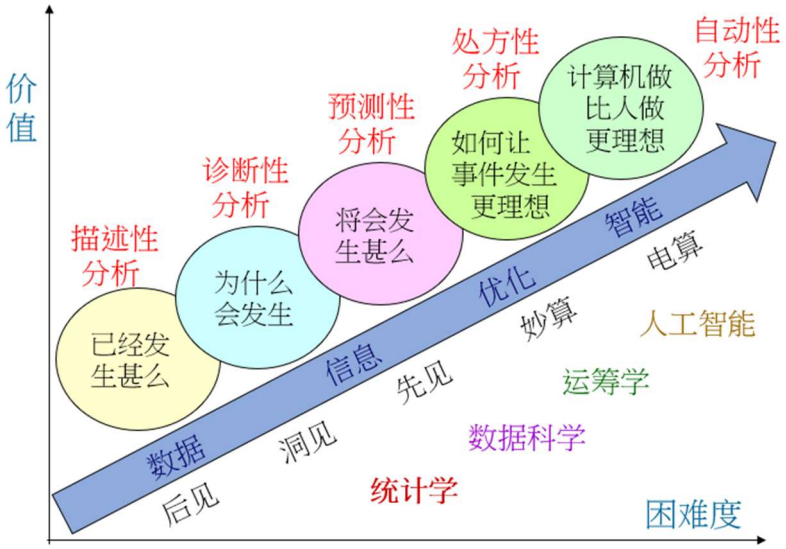


图 3　数据分析的类型（《大话统计学》 图 1.8)

Wonnacott 说：“He uses statistics as a drunken man uses lampposts --- for support rather than for illumination.” (人们利用统计，就好像醉汉利用路灯，是为了支撑，而不是照明。)

一个醉汉在夜晚的路灯下找钱包。有路人帮他找，找了很久，路人问：「你确定是掉在『这里』吗？」醉汉说：「我不知道掉在『哪里』。」路人问：「为什么要在『这里』找？」醉汉说：「因为『这里』有路灯比较亮。」

George Box 说：“Statisticians, like artists, have the bad habit of falling in love with their models.” (统计学者像艺术家，有坏习惯：会爱上他们的模型(模特儿))。

Box 又说：“All models are wrong, but some are useful. ” (所有的模型都是错误的，但是有些是有用的。)

斜杠的苏轼是：（苏东坡 才是真正的斜杠，元 Meta 的境界，我只是在数据打转）
文学家 / 诗人 / 词人 / 画家 / 书法家 / 哲学家 / 政治家 / 犯官 / 农夫 / 建筑师 / 工程师 / 美食家 / 厨师。

苏东坡说：“横看成岭侧成峰，远近高低各不同。不识庐山真面目，只缘身在此山中。”

上述是模型的隐喻。

商业模式或商业模型(business model)分三大部分：价值主张(顾客价值与获利公式)、关键资源(设备技术伙伴顾客关系)、关键流程(因果与活动)。人的三观：价值观、人生观、世界观。商业模式的 三观 ：价值主张(价值观、政绩观)、关键资源(人生观、事业观)、关键流程(世界观、工作观)。所谓，**羊毛**(价值主张) **出在狗身上**(关键资源)，**猪来买单**(关键流程)。

图 4 是基于“关系和因果”的统计学 元模型 (提升高度的模型)，请参考《大话统计学》。

图 4 基于“关系和因果”的 统计学 元模型 (《大话统计学》 图 1.6)

5. 数据料理

民以食为天，以食物来比喻，数据是食材，数据模型就是食谱，不同的食材(例如数据尺度：比率、区间、顺序、分类，或正态分配)，有不同的调味(参数选择)和烹煮(算法步骤)。

《中文统计》《运筹学2.0》(基于 Excel 2019, 2021的加载项)是 有菜单料理，有菜单料理是亲切友善的选择画面输入数据。R和Python是 无菜单料理，以函数/指令操作，无菜单料理有丰富多样的自助加料(程序)，但是要熟悉 厨师 (铸剑师、R或Python的包或库)。

《中文统计》、《运筹学2.0》、R 语言和Python都是免费的软件，天下有白吃的午餐。R好像是丐帮有许多分舵(package)，Python 像是免费的少林/武当/大饭店师傅。那么，高价的统计数据分析 商业软件怎么经营？大概要走向摆饰漂亮的米其林餐厅。你可以做气宗、武馆或钱庄，如：厨师技巧烹饪教室、厨房的设备或食材的供货商；也可以做武术分享平台，如大众点评、外送平台；还有自动化烹饪机器人。这就是 商业模式。

图 5　中文统计菜单

图 6　运筹学 2.0 菜单

6. 元宇宙

2021 年 10 月底 Facebook 脸书集团名字要改名为 Meta，造成“元宇宙” (Metaverse)的热火朝天。

以下引用：陈文贤《大话数据科学》，清华大学出版社 2020 年，第 30 页。

希腊语：μετά（metá），意思是“之后”、“之上”、“超越”、“关于”、“整合”、“变化”、“再转换”、“再诠释”，翻译为「元」或「后设」。meta 是 关于什么的什么 。

元模型/后设模型(meta model)是模型之上，超越模型的模型、关于模型的模型。

元分析/后设分析(meta-analysis)是指将多个研究结果加以整合、再诠释的分析方法。

形而上学(Metaphysics)：超越自然之上，易经：「形而上者谓之道，形而下者谓之器」。
元数据(Meta data)：关于数据的数据、超越数据的数据。
元知识(Meta knowledge)：关于知识的知识、超越知识的知识。
元语言(Meta language)：描述语言的语言。
元文法(Meta grammar)：描述文法的文法。
后设理论(Meta theory)：解释理论的理论。
后设认知(Meta cognitive)：认知自己的认知。
后设学习(Meta learning)：整合学习的学习，数据科学的 集成学习(ensemble learning)。
生物学的世代交替(Metagenesis)，蜕变(Metamorphosis)。
后设大数据 Meta Big data？后设人工智能 Meta AI？后设元宇宙 ？

本文开头说的 隐喻 Metaphor 的 meta 有「再转换」的意思。

通常学者做学术研究如 钻地机（挖井机、钻油机），往下挖细部的问题，例如机器学习的"深"度学习。我写书是期许为 学习地图和空拍机（航拍机），站在 提升高度(meta)，去看数据科学和统计学等学科的模型关系(元模型)。诚如苏东坡说：只缘身在此山中。但是，高度提升太高，可能遇到雾茫茫，会成为 "概帮"(概念帮)，例如元宇宙的概念股。所以，提升高度也要落地实战，同时要有称手好用的兵器 -- 软件工具和程序语言，如《中文统计》、R 和 Python。

百度或谷歌地图右下角的 "+" 号，是使地图更 "加" 详细，但是高度降低，范围更小，像钻地机，这就是通常的学术研究，把简单的变复杂，在数据科学/机器学习，这是 过拟合(overfitting)。百度地图右下角的 "−" 号，是 "减" 去无关因素，提升高度，范围变大，视野更广，像空拍机，把复杂的变简单。减法的人生，会提升人的高度。因为 无欲则刚。

因为 verse 是诗，universe 是宇宙。Meta verse 应该是 元诗：超越诗的诗，圣经的诗篇。翻译为「元宇宙」，有点抬举：整合宇宙的宇宙？脸书新名 Meta Platforms(整合平台的平台)。

元宇宙希望要整合 区块链 Block chain，虚拟现实 VR，增强现实 AR，混合现实 MR，人工智能 AI，5G 与人的互动等平台。请见 **表 1** 元宇宙价值链(Metaverse Value Chain)。

Metaverse 源自科幻小说《Snow Crash》(1992 年)，讲的是虚拟网络和现实世界的互动。现实与虚拟的关键技术、交易机制、规范准则、经济社交、互通互补、商业模式、价值伦理等，需要有新的定义、普世公认、不断修改。艺术品的 非同质化代币(Non-fungible token, NFT) 也许是元宇宙的一个 规范准则。但是在股票市场，小型股(小盘股)比较容易炒作(控盘容易)，NFT 是独一无二的不可互换，又没有公开的市场如股市，不是更容易炒作吗？

宋真宗赵恒：「富家不用买良田，书中自有千钟粟。安居不用架高堂，书中自有黄金屋。出门莫恨无人随，书中车马多如簇。娶妻莫恨无良媒，书中自有颜如玉。」

从前，虚拟世界（**书**）中的 千钟粟、黄金屋、车如簇、颜如玉，转换成为现实世界的良田(食)、高堂(住)、出门(行)、娶妻(育乐/成家)，其 规范准则 是 科举考试。透过 科举制度，可以当官取得 **功名**和**俸禄**。

现在或未来，虚实之间的 规范准则与商业模式，在 食衣住行育乐名利 哪部分的 元宇宙 是可行？

在元宇宙的爆火中，现实的 食衣住行育乐 都往元宇宙里装，问题是其 商业模式 是什么？元宇宙的 虚拟食品和饮料有何 价值？画饼充饥、望梅止渴、想象美味、气氛环境、欢乐共享、品牌虚荣？可以吃的 NFT 食物？吃饭要戴虚拟现实 VR 眼镜 ；虚拟衣服设计可能有 NFT 的

价值；虚拟住房有 地点(Location) 物以稀为贵的价值；旅行、教育、会议和娱乐有 虚拟现实或增强现实 VR/AR 的效果价值。

元宇宙在虚拟生活的 食衣住行育乐 之上，还有 **安全、名、利、情感和成就** (人生追求的需求层次)，后两者是 网络游戏 和 虚拟小说(fiction)的世界。因为，在网络游戏里可以满足刺激性和成就感，在虚拟小说中可以找到 爱恨情仇。

元宇宙是否会成为 暴发户的丐帮？中实户的概帮？还是 诈骗户的盖帮？

四十多年来，信息管理在产业界、学术界、顾问界的推波助澜，不断的创造新名词，有的名词可以风行很久，有的名词只是昙花一现，有的名词是流行、负面、沉寂、再爆发。

元宇宙会如何？我不知道，让子弹飞一会儿吧。

2022 年 3 月 于 澳大利亚 悉尼

表 1 元宇宙价值链 (Metaverse Value Chain)：元宇宙的元模型(概念模型)

武林江湖	元宇宙 价值链	内容	重点	实例
江湖历练 恩怨情仇	体验应用	食衣住行育乐、 社群、名、利	商业模式、 数据收集	游戏、沉浸体验、 会议、购物、教育
寻师门派 武林秘籍	探索搜寻	搜寻体验应用的 平台	整合平台、 流量、飞轮效应	腾讯、App store、 Meta、Google Play
剑宗招式	创造开发	设计工具、动画系统、 图形工具、软件供应	软件技术	Shopify、Roblox、 Autodesk、Epic
气宗内功	空间计算	3D 引擎、手势识别、 空间映射	人工智能 AI、 视频会议	微软 Teams、百度、 Google AI、Unity
炼丹画符 灵药毒蛊	去中心化	区块链、NFT、 数字资产、以太坊	规范、真实价值、 炒作、交易安全	艺术品、球员卡、 图片、音乐、宝物
倚天屠龙 刀剑兵器	人机交互	VR/AR/XR、 智能眼镜、穿戴设备	标准的竞争、 网络效应	Oculus、Vive、 Apple
神鵰奇兽	基础设施	5G/6G、IC 芯片、 云计算、边缘计算	网络通信基础	华为、腾讯、阿里、 AWS、Nvidia