



数据科学导论-导读

王小宁

中国传媒大学数据科学与智能媒体学院

2025 年 02 月 24 日



目录

大纲

数据科学的历史

大模型和数据科学

数据科学的前身：统计学

数据科学研究的主要问题



大纲



主要内容

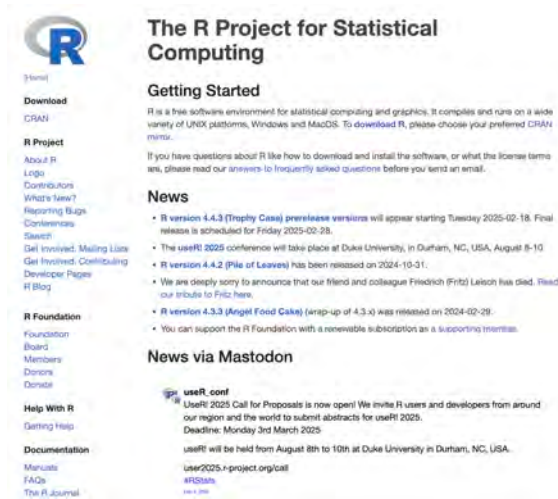
- 介绍数据科学的基本思想、发展历程。
- 介绍数据科学中常见的分析方法、建模思想、常见算法和基本工具。
- 结合案例讲解数据分析在实际生活中的应用。
- 介绍数据科学和传媒行业的结合应用。
- 大模型时代数据科学的一些思考。



考核

- 平时作业：30%，布置 4-5 次课程作业
- 期中测试：40%，线下，考试（闭卷）
- 期末测试：30%，实验报告
- 注：本课程主要参考书《R 语言实战（第 2 版）》，且有大量的阅读资料和相关资源推荐，详细信息：<https://github.com/xiaoningwang/IntroductionofDataScience>
- 智能助教平台：书卷侠 (<https://scholarhero.cn/>)

教学软件 (R VS Python)



The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

News

- R version 4.4.3 (Trophy Case) prerelease versions will appear starting Tuesday 2025-02-18. Final release is scheduled for Friday 2025-02-28.
- The [useR! 2025](#) conference will take place at Duke University, in Durham, NC, USA, August 8-10.
- R version 4.4.2 (Pile of Leaves) has been released on 2024-10-31.
- We are deeply sorry to announce that our friend and colleague Friedrich (Fritz) Leisch has died. [Read our tribute to Fritz here](#).
- R version 4.3.3 (Angel Food Cake) (wrap-up of 4.3.x) was released on 2024-02-29.
- You can support the R Foundation with a renewable subscription as a [supporting member](#).

News via Mastodon

[useR!_conf](#)
useR! 2025 Call for Proposals is now open! We invite R users and developers from around our region and the world to submit abstracts for useR! 2025.
Deadline: Monday 3rd March 2025

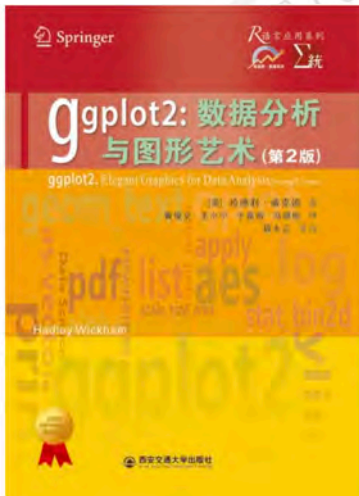
useR! will be held from August 8th to 10th at Duke University in Durham, NC, USA.

[useR2025.r-project.org/call](#)
[#RStats](#)
2025-03-03

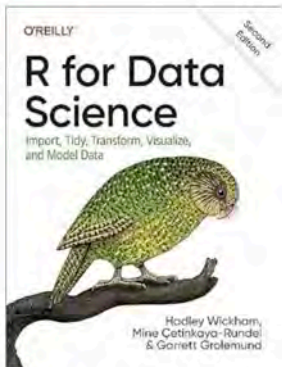
智能助教-书卷侠



主要参考资料



主要参考资料 2



《R数据科学（第2版）》
Hadley Wickham 等 | 著
张敬信, 王小宁, 黄俊文 | 译



一个小问卷





数据科学的历史

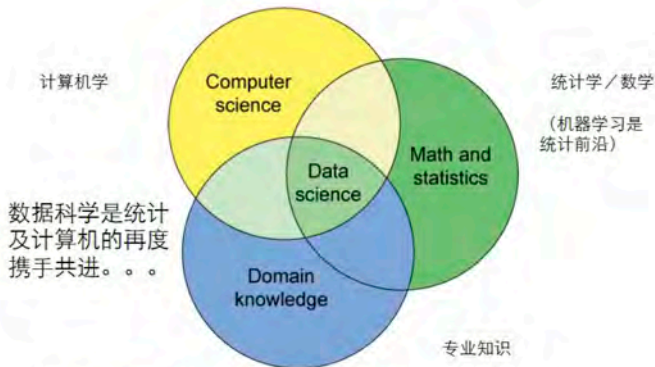


数据科学

- 数据科学是一门交叉学科，主要研究如何利用科学的方法、过程、算法或系统，从结构化的或非结构化的数据中提炼知识、洞察规律、获得见解，基本内涵：
 - ① 用数据的方法研究科学问题。在科学研究的历史长河中，经过多年的发展，形成了从实验归纳到模型推演，再到计算机仿真的三种科学研究范式。在如今这个数据爆炸的时代，数据驱动来推进相关原理和方法发现的科学研究方法被称为科学研究的第四范式，比如生物信息学、天体信息学等等。
 - ② 用科学的方法研究数据。我们对于数据的研究不是靠经验或者感觉，而是把数据的研究看作一个具有生命周期的过程，包含数据的采集、管理、分析，到可视化呈现，以及数据如何进行有效的治理，甚至数据的分析过程是不是涉及伦理问题等，都采用一种科学的方法来进行研究。这就是数据科学的另一层含义，用科学的方法研究数据。

数据科学三元素

数据科学三元素



<http://www.ibm.com/developerworks/jp/opensource/library/os-datascience/figure1.png>

6



数据的方法研究科学

- 科学研究的第四范式也称作数据密集型科学。它是将海量数据放入庞大的计算机集群中，只要数据间存在着一定的相互关系，那么就能找到相应的模型和算法，来发现传统的科学方法发现不了的新模式、新知识，甚至是新规律。
- 数据科学对科学研究产生了重要的影响，当前它已经成为科研体系的重要组成部分。随着未来的发展，它将取得与物理、化学、生命科学等自然学科同等重要的地位。数据科学也促使科学研究与市场产业、行业的联系更加密切，缩短了从基本原理的发现，到产生经济效益的产业化周期。除此之外，数据科学相关的研究和应用，与社会的发展以及人们日常生活的联系也将会越来越紧密。

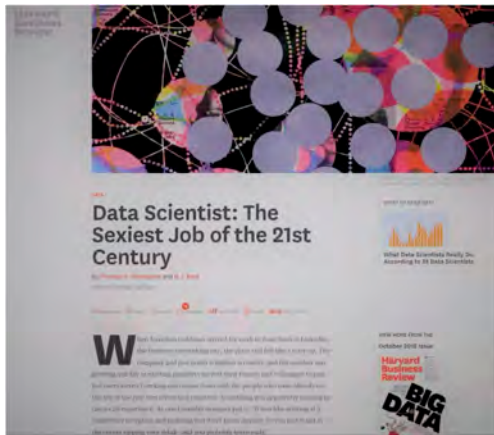


核心

一句话

- 数据科学是解决问题的科学
- 数据科学三元素 + 给力团队 + 验证 = 解决数据问题/获取知识/做决定

论证



- 图片说明：2012 年 DJ · Patil 在《哈佛商业评论》上发表文章“Data Scientist: The Sexiest Job of the 21st Century” 图片来源：Harvard Business Review)



数据产品经理的技能

数据产品经理





大模型和数据科学



什么是大模型？

- 大模型（Large Model）是指参数量庞大的深度学习模型
- 主要基于 Transformer 结构，具备强大的学习与推理能力
- 典型应用领域包括自然语言处理（NLP）、计算机视觉（CV）等



大模型的关键特性

- 海量参数：如 GPT-4、PaLM2，参数规模达千亿级
- 大规模数据训练：基于互联网数据进行预训练
- 强泛化能力：可迁移到不同任务，如文本生成、代码补全
- 推理能力强：可以进行复杂推理，如数学计算、逻辑推演



典型大模型

模型	机构	主要应用
GPT-4	OpenAI	语言生成、问答
PaLM 2	Google	多语言处理
LLaMA 2	Meta	开源语言模型
Claude	Anthropic	对话 AI
Gemini	Google DeepMind	多模态 AI
Kimi	月之暗面	联网、长上下文
通义千问	阿里	开源大模型
DeepSeek	深度求索	思考、成本低、开源



大模型的应用

- 自然语言处理：
 - 机器翻译 (DeepL、Google Translate)
 - 智能对话 (ChatGPT、Claude)
 - 文本生成 (写作辅助、代码生成)
- 计算机视觉：
 - 图像识别 (CLIP、DINO)
 - 生成式 AI (Stable Diffusion、DALL·E)
- 科学计算与医疗：
 - 药物研发 (AlphaFold)
 - 天文数据分析



大模型的挑战

- ① 计算资源消耗：训练需要大量 GPU/TPU
- ② 高能耗问题：对环境有较大影响
- ③ 数据隐私与安全：可能泄露敏感信息
- ④ 幻觉 (**Hallucination**)：生成错误或不真实信息
- ⑤ 成本高昂：训练与推理费用较高



数据科学与大模型的联系

数据是大模型的基础

- 训练数据：大模型的性能高度依赖数据质量和规模
- 数据预处理：数据清洗、标注、归一化等数据科学方法是训练大模型的重要步骤
- 特征工程：在中小型模型中仍需要数据科学家设计有效特征，大模型则自动从数据中学习特征



大模型提升数据科学效率

- 自动特征提取：如 Transformer 模型能自动学习复杂特征，无需手动特征工程
- 数据生成：大模型可以进行数据增强，合成更多训练样本
- 数据分析自动化：大模型可自动生成报告、数据洞察，提升效率



大模型推动数据科学发展

- 更强的建模能力：大模型在 NLP、CV、时间序列等任务中大幅超越传统模型
- 跨模态分析：文本、图像、音频等多模态数据的联合建模与分析
- 小样本学习：通过预训练模型实现迁移学习，在少量数据下取得优异表现



数据科学在大模型中的角色

数据科学环节	大模型中的作用
数据采集与清洗	确保训练数据质量
数据分析	评估数据分布与模型表现
特征工程	辅助模型理解复杂特征
模型评估与调优	使用数据指标指导模型优化
结果解释与可视化	辅助模型输出合理解释



挑战与思考

- 数据隐私与安全：如何在大模型训练中保护敏感数据？
- 数据质量控制：低质量数据如何影响大模型表现？
- 模型可解释性：如何让大模型决策更加透明？
- 计算资源需求：如何降低大模型训练与推理成本？

GPT-3 数据集

- GPT-3 模型由 OpenAI 于 2020 年发布。论文阐明了所用训练数据集的 token 数量，但训练数据集的内容和大小尚不清楚（Common Crawl 的数据集大小除外）

Dataset	Tokens (billion)	Assumptions	Tokens per byte (Tokens / bytes)	Ratio	Size (GB)
Common Crawl (filtered)	410B	-	0.71	1:1.9	570
WebText2	19B	25% > <i>WebText</i>	0.38	1:2.6	50
Books1	12B	<i>Gutenberg</i>	0.57	1:1.75	21
Books2	55B	<i>Bibliotik</i>	0.54	1:1.84	101
Wikipedia	3B	<i>See RoBERTa</i>	0.26	1:3.8	11.4
Total	499B				753.4GB

科学的方法研究数据

- 数据科学在数学、统计学、计算机科学的多学科的支撑下，从数据采集、数据管理、数据治理、数据分析、数据可视化、数据伦理等众多的方面来开展科学的研究，涵盖了数据全生命周期的流程和相应的处理链条。



图 1: 数据科学基本内容

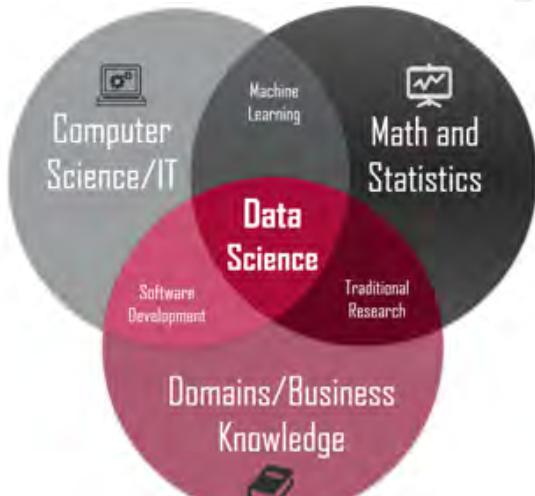


具体内容

- 数据采集：借助相关的技术和手段来进行数据的收集；通过将收集的数据存储在介质中，来对数据进行管理和维护；
- 数据治理：通过对数据进行有效的组织，可以有效提升数据的质量，以便为后面的分析过程提供更好、更可用的数据；
- 数据分析：在数据分析环节，通过对数据进行详细的研究和概括总结，提炼有价值的信息来洞察规律，是最为重要的环节；
- 数据可视化：数据可视化，就是指运用图形、图表等多种有效的可视化方法来展示数据，以便更清晰明确地传递数据中所蕴含的价值，也帮助人们更好的理解数据。
- 数据安全：在分析和运用数据的过程中，是否会产生数据安全问题？是否会侵犯用户的隐私？用算法得出的结论，是否会对特定群体产生不公平现象？是否会存在认知上的偏见？这都是数据伦理问题。
- 数据应用：通过对数据的分析，得出的知识、见解、原理，或者是相关关系，必将对相关的行业领域产生影响，也为相关的领域带来应用价值。

数据科学：交叉学科

- 数据科学是一门交叉学科，德鲁·康为（Drew Conway）的韦恩图展示了数据科学多学科交叉的特性。





大数据

- “大数据” (Big Data) 这个词近年来十分火爆。大数据是指无法在一定时间内用常规的软件工具对其内容进行获取、管理和处理的海量数据的集合。大数据具有“5V”特性：
- 规模性 (Volume)：形容数据量非常大。
- 多样性 (Variety)：指数据的类型众多，有结构化的，也有非结构化的。
- 高速性 (Velocity)：是指数据产生的速度非常快。由于在我们的日常生活中，每天都会快速产生大量的数据，所以要求我们处理数据的能力要强、处理的速度要快，这样才能快速发现数据中的价值。
- 真实性 (Veracity)：指的是从互联网或是智能传感器等数据收集工具得到的数据，是人们真实行为的一个体现，通过对这些数据的分析，可以洞察人们的行为规律。
- 价值性 (Value)：指的是大数据的价值密度低。价值隐藏在海量的数据中，我们要通过分析挖掘才能得到这样的价值。
- 这就是大数据的“5V”特性。



数据科学与大数据

- 数据科学是一门以数据，尤其是大数据作为研究对象的学科。大数据最大的特点就是数据的规模大，而数据科学本身它并不强调数据规模的大小，在大数据出现之前，数据科学也有着广泛的应用。
- 针对大数据所带来的这种挑战，数据科学更拥有了用武之地，它为在海量的数据中挖掘价值，构建相应的规律，提供了新的思维、新的思路和新的方法。



数据科学的前身：统计学



统计学

- 统计学作为一门学科已有三百多年的历史。按统计方法及历史的演变顺序，通常可以将统计学的发展史分为三个阶段，分别是古典统计学时期、近代统计学时期和现代统计学时期。
- 古典统计学的萌芽最早可以追溯到 17 世纪中叶，此时的欧洲正处于封建社会解体和资本主义兴起的阶段
- 政治改革家们急需辅助国家经营和管理的数据证据以适应经济发展需要，此时一系列统计学的奠基工作在欧洲各国相继展开。
- 在这一时期，以威廉配第和约翰格朗特为代表的政治算术学派与海尔曼康令创立的国势学派相互渗透和借鉴，服务与指导了国家管理和社会福利改善。

统计先驱

统计先驱Herman Hollerith: 发明早期计算机 Hollerith Machine



统计学家，发明家
(1860-1929)

现代机器数据处理之父
他的公司是IBM前身公司之一



他受聘与美国人口统计局. 为了解决人口统计问题，他发明了计算机...

为解决问题，超越前人的框架。
问题必须驱动统计进化，前行。



统计发展

- 18 世纪末至 19 世纪末为近代统计学发展时期。
- 这一百年间欧洲各国先后完成了工业革命，科学技术开始进入全面繁荣时期，天文、气象、社会人口等领域的数据资料达到一定规模的积累，对统计的需求已从国家层面扩展至社会科学各个领域。
- 对事物现象静态性的描述已不能满足社会需求，数理统计学派创始人凯特勒率先将概率论引进古典统计学，提出了大数定律思想，使统计学逐步成为揭示事物内在规律，可用于任何科学的一般性研究方法。
- 一些重要的统计概念也在这一时期提出，误差测定、正态分布曲线、最小二乘法、大数定律等理论方法的大量运用为社会、经济、人口、法律等领域的研究提供了大量宝贵的指导。



描述统计

- 20 世纪科学技术的发展速度远超过之前的时代，以描述性方法为核心的近代统计已无法满足需求，统计学的重心转为推断性统计，进入了现代统计学阶段。
- 随着 20 世纪初细胞学的发展，农业育种工作全面展开。1923 年，英国著名统计学家费雪（**R.A.Fisher**）为满足作物育种的研究需求，提出了基于概率论和数理统计的随机试验设计技术以及方差分析等一系列推断统计理论和方法。推断性统计方法的进步对工农业生产和科学研究起到了极大的促进作用。

统计先驱

统计先驱J.W. Tukey (1962): Future of data analysis



A mathematician (数学家)
(1915 – 2000)

U.S. Medal of Science
IEEE Medal for co-invention of FFT

It will still be true that there will be aspects of data analysis well called technology, but there will also be the hallmarks of stimulating science: **intellectual adventure**, **demanding calls upon insight**, and a need to find out "**how things really are**" by investigation and the confrontation of insights with experience.

Tukey's definition of "data science"?

数据科学既是技术，更是科学。

统计先驱

统计先驱Leo Breiman (2001): Statistical modeling: the two cultures



A probabilist, statistician, and
machine learner

(概率学家，统计学家，机器学习家)
(1928 – 2005)

CART, Bagging, Random Forests

"If our goal as field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a diverse set of tools."

统计应该不限于建模型，应该采用多样工具。



统计应用

- 自 20 世纪 30 年代，随着社会经济的发展和医学先进理念的吸收融合，人们对于医疗保险和健康管理的需求日益增长，统计思想渗透到医学领域形成了现代医学统计方法。例如在生存质量 (Quality of life) 研究领域，通过分析横向纵向资料，逐步形成重复测量资料的方差分析、质量调整生存年 (QALYs) 法等统计方法。这一阶段，统计在毒理学、分子生物学、临床试验等生物医学领域获得了大量应用，这些领域的发展又带动统计方法不断创新，主成分估计、非参数估计、MME 算法等方法应运而生。
- 20 世纪 80 年代开始，随着现代生物医学的发展，计算机技术的进步，人类对健康的管理和疾病的治疗已进入基因领域，对基因数据分析产生了大量需求。高维海量的基因数据具有全新的数据特征，变量维度远远大于样本数，传统的统计方法失效了，因此一系列面向高维数据的统计分析方法相继产生，比如著名的 Lasso 方法。



突飞猛进

- 20 世纪 90 年代以来，随着互联网的发展，数据库中积累了海量的数据，如何从海量的数据中挖掘有用的信息就变得越来越重要了，数据挖掘（Data Mining）也就应运而生了。
- 数据挖掘又称数据库中的知识发现（Knowledge Discover in Database, KDD），是目前人工智能（artificial intelligence）和数据库领域研究的热点问题，所谓数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。
- 与数据挖掘比较接近的名词是机器学习（Machine learning），机器学习被看作是人工智能的一个分支，主要研究一些让计算机可以自动“学习”的算法，是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。因为机器学习算法中涉及了很多的统计学理论，与统计学的关系密切，也被称为统计学习（Statistical learning）。



大数据

- 麦肯锡全球研究院（MGI）在 2011 年首次提出了大数据时代（age of big data）概念。依照美国咨询公司麦肯锡（McKinsey）的定义，大数据是指那些规模超出了典型的数据库软件工具的能力来进行捕获、存储、管理和分析的数据集。
- 与传统数据相比，大数据的大不仅仅是体量上的扩充，数据的结构、形式、粒度、组织等各方面都更加复杂。不过我们认为，大数据并不是从方法论角度提出的，研究大数据的方法主要还是数据挖掘和机器学习方法。



数据科学

- 近几年数据科学（Data Science）的概念被提出，这是一门分析和挖掘数据并从中提取规律和利用数据学习知识的学科，因此其概念也更广，包含了统计、机器学习、数据可视化、高性能计算等。
- 近几年，数据科学家这个词也跟着火起来，成为职场中的香饽饽。德勤（Deloitte）预测 2018 年全球企业将至少需要 100 万数据科学家，大学培养的数据科学家数量远远不能满足市场需求，按照目前数据科学家的培养数量来看，这个缺口是很大的。
- 我们真正的数据科学家人才是比较短缺的。数据科学家需要有良好的统计学、机器学习功底，能理解模型背后的原理和算法，熟练的编程能力以及熟悉业务知识。



数据科学研究的主要问题



聚焦

核心定义

- 只要和数据收集、清洗整理、分析和挖掘有关的都是数据科学要研究的问题。
- 数据科学所研究的问题，应该是从实际业务需求中提炼出来的问题。

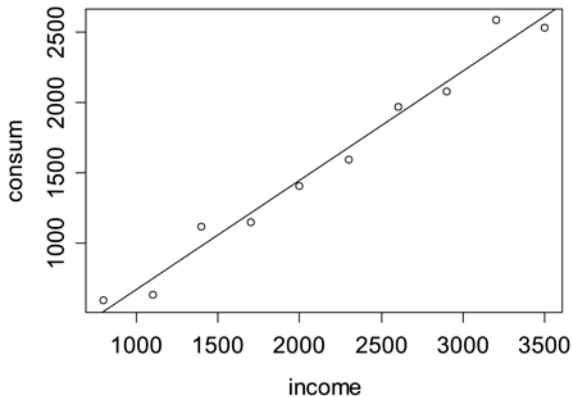


家庭收入与消费支出

- 为了研究某社区家庭月消费支出与家庭月可支配收入之间的关系，随机抽取并调查了 12 户家庭的相关数据。
- 通过调查所得的样本数据能否发现家庭消费支出与家庭可支配收入之间的数量关系，以及如果知道了家庭的月可支配收入，能否预测家庭的月消费支出水平呢？



消费 vs 收入





消费贷公司对客户的信用评分

- 客户申请消费贷的时候，公司收到客户的收入、工作年限、职业等数据以及从其他渠道获取的数据，消费贷公司需要评估客户的信用评分，以便决定是否给予核准贷款。
- 那么，该如何预测客户借钱了是否会违约？该如何给每位客户评分？

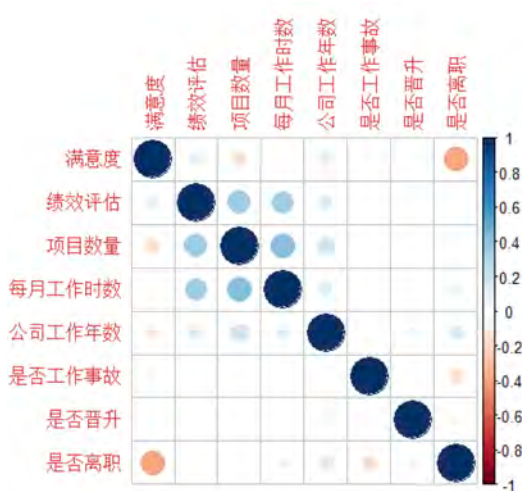


员工离职预测

- 一定的员工流动率能够为企业注入新鲜的活力，增强组织的创新能力，但过多的员工离职特别是核心员工的离职则会导致企业人力资本投资的损失、员工士气低落、破坏企业建立的竞争优势等消极影响，甚至对社会稳定也会造成一定的威胁。
- 因此，通过对离职影响因素的分析，企业管理者可以有效地对员工的离职行为进行管理。
- 比如收集了员工满意度、绩效评估、完成的项目数量、每月工作时数、工作年数等因素，如何预测员工是否离职，以便提前做好准备？



离职预测





购物篮分析

- 某超市顾客购买记录的数据库，包含 6 个事务，其中项集 = {面包，牛奶，果酱，麦片}。现在要分析已购买面包的顾客，有多大可能会买牛奶？如何根据顾客的过去购买记录，推荐其感兴趣的商品？

TID	Date	Items
T100	6/6/2024	{面包，麦片}
T200	6/8/2024	{面包，牛奶，果酱}
T300	6/10/2024	{面包，牛奶，麦片}
T400	6/13/2024	{面包，牛奶}
T500	6/14/2024	{牛奶，麦片}
T600	6/15/2024	{面包，牛奶，果酱，麦片}



花卉细分

- 测量了 18 种花卉的 8 个指标，这 8 个指标包括是否能过冬、是否生长在阴暗的地方、是否有块茎、花卉颜色、所生长泥土、花卉高度、花卉之间所需的距离间隔等。
- 如何根据这 8 个指标对 18 种花卉进行细分？该分为几类比较合适？



文本挖掘

- 从网上收集 20000 多篇关于房地产的相关新闻，如何分析这 20000 篇新闻里都在讨论哪几个主要话题？
- 如何有效地把这些新闻聚为几类？如何提取新闻的情感倾向并编制成指数？



预习

- R 语言实战（第 2 版）-第 1 章和第 2 章
- 下载 R
- 下载 Rstudio
- 《统计与真理，怎样运用偶然性》or 《女士品茶》