目录

第 8 章　可视化 3D 图 / 141

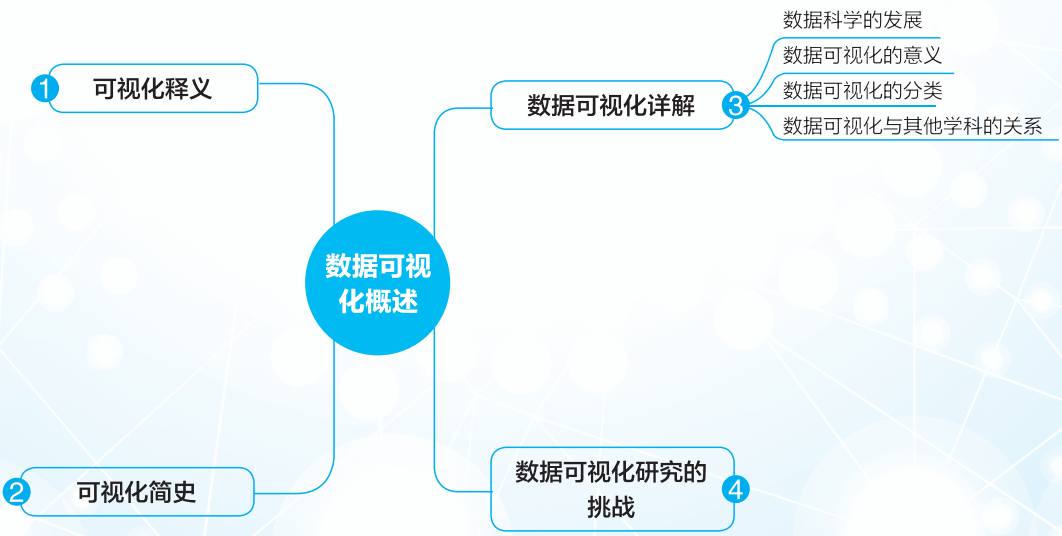
第 9 章　词云 / 165

第 1 章 数据可视化概述

学习目标

1. 了解可视化的定义。
2. 了解数据可视化的发展。
3. 了解数据可视化的分类。
4. 了解数据可视化的应用。

知识导图



笔记

本章导读

数据可视化是关于数据视觉表现形式的科学技术研究。其中，这种数据的视觉表现形式可定义为：一种以某种概要形式抽提出来的信息，包括相应信息单位的各种属性和变量。

它是一个处于不断演变之中的概念，其边界在不断扩大，主要指较为高级的技术方法，这些技术方法允许利用图形图像处理、计算机视觉以及用户界面，通过表达、建模，以及对立体、表面、属性、动画的显示，对数据加以可视化解释。与立体建模之类的特殊技术方法相比，数据可视化所涵盖的技术方法要广泛得多。

数据可视化与信息图形、信息可视化、科学可视化以及统计图形密切相关。

1.1 可视化释义

人眼是一个高带宽的巨量视觉信号输入并行处理器，最高带宽为 100 MB/s，具有很强的模式识别能力，对可视符号的感知速度比对数字或文本快几个数量级，且大量的视觉信息的处理发生在潜意识阶段。其中的一个例子是视觉突变：在一大堆灰色物体中能瞬时注意到红色的物体。由于在整个视野中的视觉处理是并行的，不论物体所占区间多大，这种突变都会发生。视觉是获取信息最重要的通道，超过 50% 的人脑功能用于视觉的感知，包括解码可视信息、处理高层次可视信息和思考可视符号。

可视化与山岳一样古老。中世纪时期，人们就开始使用包含等值线的地磁图、表示海上主要风向的箭头图和天象图。可视化通常可理解为一个生成图形图像的过程。更深刻的认识是，可视化是认知的过程，即形成某个物体的感知图像，强化认知理解。因此，可视化的终极目的是对事物规律的洞悉，而非绘制的可视化结果本身。这里包含多重含义：发现、决策、解释、分析、探索和学习。因此，可视化可简明地定义为“通过可视表达提高人们完成某些任务的效率”。

从信息加工的角度看，丰富的信息将消耗人们大量的注意力。精心设计的可视化可作为某种外部内存，辅助人们在人脑之外保存待处理信息，从而补充人脑有限的记忆内存，有助于将认知行为从感知系统中剥离，提高信息认知的效率。另一方面，视觉系统的高级处理过程中包含一个重要部分，即有意识地集中注意力。人类执行视觉搜索的效率通常只能保持几分钟，无法持久。图形化符号可高效地传递信息，将用户的注意力引导到重要的目标上。

可视化的作用体现在多个方面，如揭示想法和关系，形成论点或意见，观察事物演化的趋势，总结或积聚数据，存档和汇总，寻求真相和真理，传播知识和探索性数据分析等。从宏观的角度看，可视化包括以下三个功能。

1. 信息记录

将浩如烟海的信息记录成文、世代传播的有效方式之一是将信息成像或采用草图记载。可视化图能极大地激发人们的智力和洞察力，帮助验证科学假设。例如，20 世纪自然科学最重要的三个发现之一——DNA 分子结构的发现起源于对 DNA 结构的 X 射线图

片的分析：从图像形状确定 DNA 是双螺旋结构，且两条骨架是反平行的，骨架在螺旋的外侧等这些重要的科学事实。

笔记

2. 支持对信息的推理和分析

数据分析的任务通常包括定位、识别、区分、分类、聚类、排列、比较、内外连接比较、关联、关系等。通过将信息以可视的方式呈现给用户，将直接提升对信息认知的效率，并引导用户从可视化结果分析和推理出有效信息。这种直观的信息感知机制极大地降低了数据理解的复杂度，突破了常规统计分析方法的局限性。

如图 1-1 所示，将实际的数据分布情况用二维可视化呈现时，观察者可迅速从数据中发现它们的不同模式和规律。

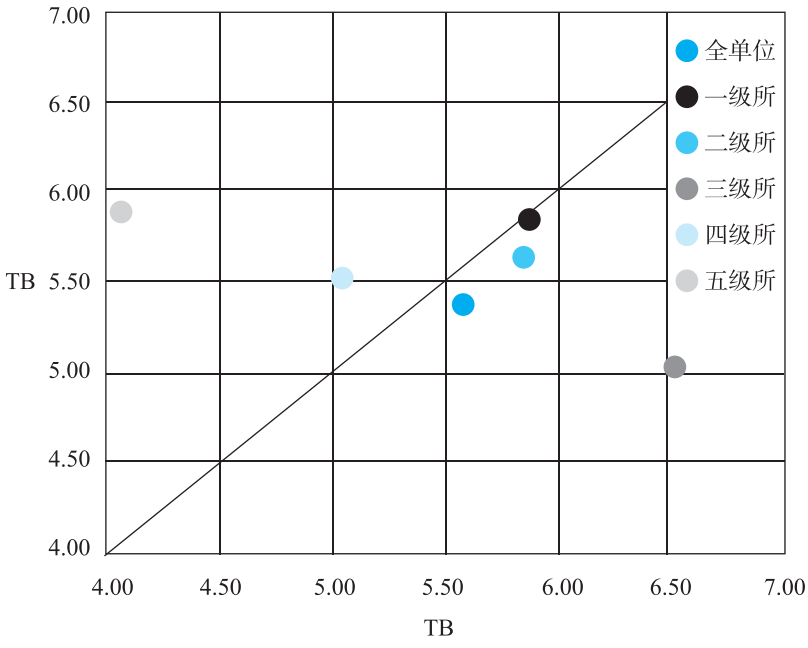


图 1-1　二维可视化呈现数据分布

可视化能显著提高分析信息的效率，其重要原因是扩充了人脑的记忆，帮助人脑形象地理解和分析所面临的任务。

3. 信息传播与协同

人的视觉感知是最主要的信息界面，它输入了人从外界获取的 70% 的信息。因此，俗语说“百闻不如一见”“一图胜千言”。面向公众用户传播与发布复杂信息的最有效途径是将数据可视化，达到信息共享与认证、信息协作与修正、重要信息过滤等目的。

在移动互联网时代，资源互联和共享、群体协同与合作成为科学和社会发展的新动力。美国华盛顿大学的可视化专家教授与蛋白质结构学家开发了一款名叫 Fold.It 的多用户在线网络游戏。该游戏让玩家从半折叠的蛋白质结构起步，根据简单的规则扭曲蛋白质使之成为理想的形状。实验结果表明，玩家预测出正确的蛋白质结构的速度比任何算法都快，而且能凭直觉解决计算机没办法解决的问题。这个实例表明，在处理某些复杂的科学问题上，人类的直觉胜于机器智能，也证明可视化、人机交互技术等协同式知识传播在科学发现中的重要作用。

笔记

1.2 可视化简史

可视化发展史（见图 1-2）与测量、绘画、人类现代文明的启蒙和科技的发展一脉相承。在地图、科学与工程制图、统计图表中，可视化理念与技术已经应用和发展了数百年。

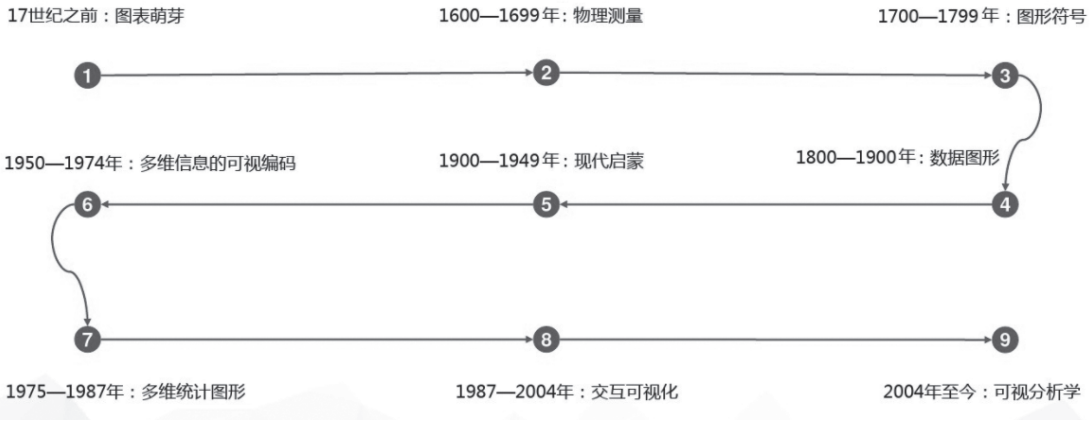


图 1-2　可视化发展史

1. 17 世纪之前：图表萌芽

16 世纪时，人类已经掌握了精确的观测技术和设备，也采用手工方式制作了可视化作品，可视化的萌芽出自几何图表的地图生成，其目的是展示一些重要的信息。图 1-3 所示为公元前 6200 年人类制作的地图。

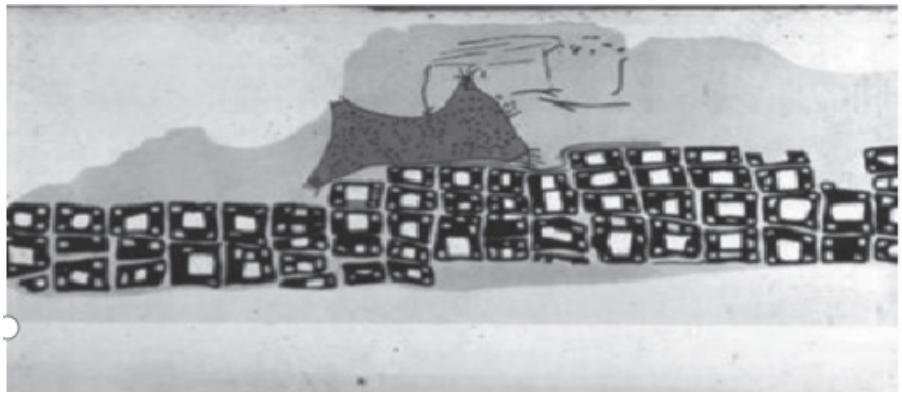


图 1-3　公元前 6200 年人类制作的地图

2. 1600—1699 年：物理测量

17 世纪最重要的科学进展是对物理基本量（时间、距离、空间）的测量设备与理论的完善，它们被广泛用于航空、测绘、制图、浏览和国土勘探等。同时，制图学理论和实践也随着分析几何、测量误差、概率论、人口统计和政治版图的发展而迅速成长。17 世纪末，甚至产生了基于真实测量数据的可视化方法。从这时起，人类开始了可视化思考的新模式。

笔记

在一个视图上同时可视化多个小图序列，是现代可视化技术中称为邮票图表法的雏形。图 1-4 为 1686 年绘制的历史上第一幅天气图，显示了地球的主流风场分布。这也是向量场可视化的鼻祖。

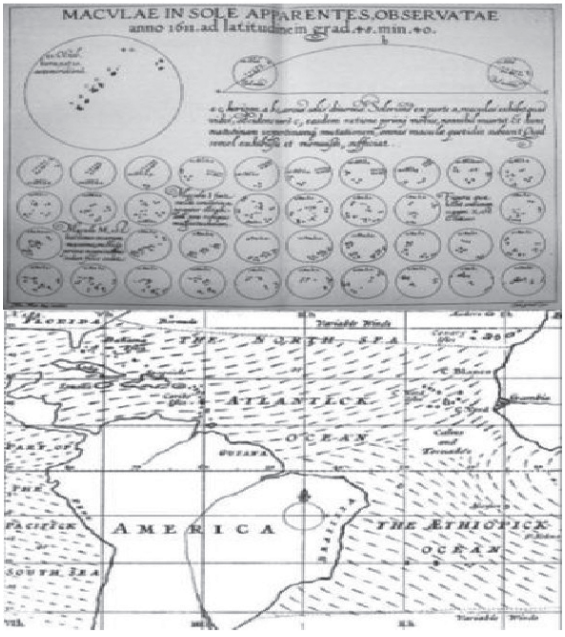


图 1-4　历史上第一幅天气图

3. 1700—1799 年：图形符号

进入 18 世纪，绘图师不再满足于在地图上展现几何信息，发明了新的图形化形式（等值线、轮廓线）和其他物理信息的概念图（地理、经济、医学）。随着统计理论、实践数据分析的发展，发明了抽象图和函数图。图 1-5 表示 1701 年地球等磁线的可视化。

图 1-5　地球等磁线的可视化

笔记

18 世纪是统计图形学的繁荣时期，其奠基人 William Playfair 发明了折线图、柱状图、显示局部与整体关系的饼状图和圆图等今天最常用的统计图表。

4. 1800—1900 年：数据图形

随着工艺设计的完善，19 世纪上半叶，统计图形、概念图等迅猛爆发，此时人们已经掌握了整套统计数据可视化工具，包括柱状图、饼图、直方图、折线图、时间线、轮廓线等。关于社会、地理、医学和经济的统计数据越来越多，将国家的统计数据和其可视表达放在地图上，产生了概念制图的新思维，其作用开始体现在政府规划和运营中。采用统计图辅助思考的诞生同时衍生了可视化思考的新方式：图表用于表达数学证明和函数；列线图用于辅助计算；各类可视化显示用于表达数据的趋势和分布，便于交流、获取和可视化观察。

19 世纪下半叶，系统地构建可视化方法的条件日渐成熟，进入了统计图形学的黄金时期。

图 1-6 为 1837 年人类历史上第一幅流图，用可变宽度的线段显示了交通运输的轨迹和乘客数量。

图 1-6　人类历史上第一幅流图

5. 1900—1949 年：现代启蒙

20 世纪上半叶对于可视化而言是一个缺乏创新的时期，但是可视化随着统计图形的主流化开始面向政府、商业和科学走向应用普及，人们第一次意识到图形显示的方式能为航空、物理、天文和生物等科学与工程领域提供新的洞察和发现机会。多维数据可视化和心理学的介入成为这个时期的重要特点。如图 1-7 所示，关于太阳黑子随时间扰动的蝴蝶图验证了太阳黑子的周期性。

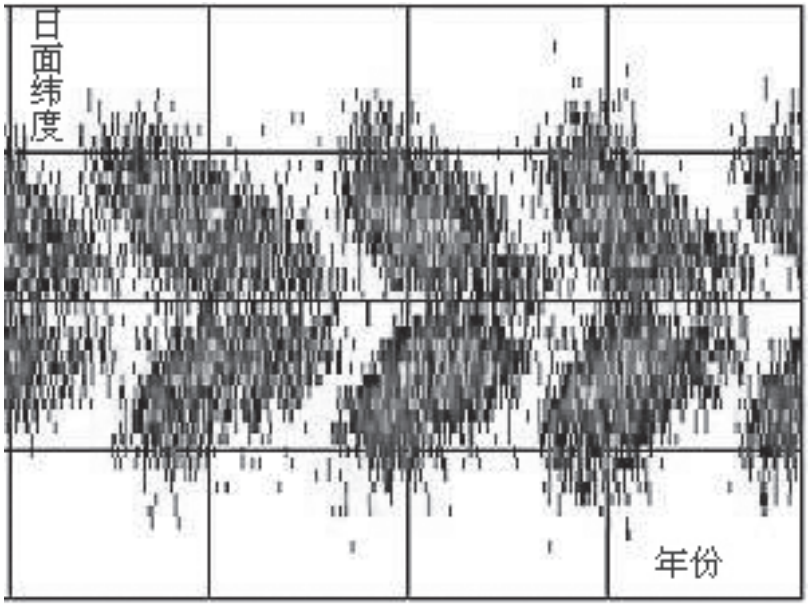


图 1-7　太阳黑子随时间扰动的蝴蝶图

6. 1950—1974 年：多维信息的可视编码

1967 年，法国人出版的《图形图像学》一书中确定了构成图形的基本要素，并且描述了一种关于图形设计的框架。这套理论奠定了信息可视化的理论基石。随着个人计算机的普及，人们逐渐开始采用计算机编程生成可视化。

图 1-8 为 1957 年发明的图形图表（采用线性及其朝向编码多维数据）和 1973 年 Herman Chernoff 发明的表达多变量数据的脸谱编码。

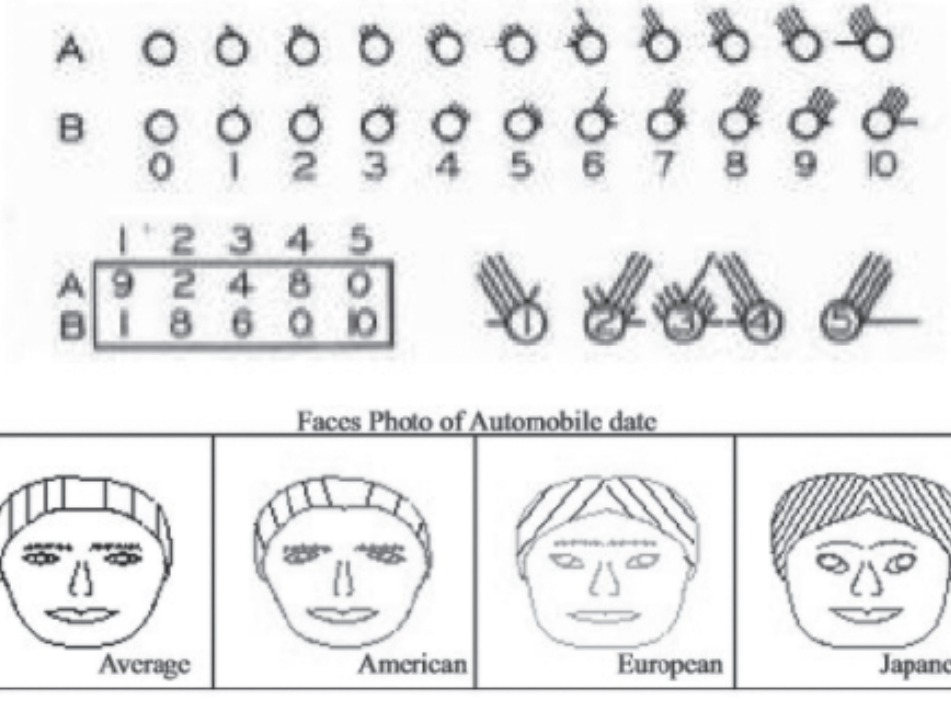


图 1-8　多维编码可视化

笔记

7. 1975—1987 年：多维统计图形

20 世纪 70 年代以后，桌面操作系统、计算机图形学、图形显示设备、人机交互等技术的发展激发了人们编程实现交互式可视化的热情。处理范围从简单的统计数据扩展为更复杂的网络、层次、数据库、文本等非结构化与高维数据。与此同时，高性能计算、并行计算的理论与产品正处于研制阶段，催生了面向科学与工程的大规模计算方法。数据密集型计算开始走上历史舞台，也造就了对于数据分析和呈现的更高需求。

1977 年，美国著名统计学家 John Tukey 发表了“探索式数据分析”的基本框架，它的重点并不是可视化的效果，而是将可视化引入统计分析，促进对数据的深入理解。1982 年，Edward Tufte 出版了 *The Visual Display of Quantitative Information* 一书，构建了关于信息的二维图形显示的理论，强调有用信息密度的最大化问题。

图 1-9 为 1981 年发明的鱼眼方法，模拟鱼眼效果对重要细节予以关注，对其他区域予以简化。

图 1-9　鱼眼视图可视化法

8. 1987-2004 年：交互可视化

1986 年 10 月，美国国家科学基金会主办了一次名为“图形学、图像处理及工作站专题讨论”的研讨会，旨在为从事科学计算工作的研究机构提出方向性建议。会议将计算机

图形学和图像方法应用于计算科学的学科称为“科学计算之中的可视化”（Visualization in Scientific Computing，VISC）。

笔记

1987 年 2 月，美国国家科学基金会召开了首次有关科学可视化的会议，召集了众多学术界、工业界及政府部门的研究人员，会议报告正式命名并定义了科学可视化，认为可视化有助于统一计算机图形学、图像处理、计算机视觉、计算机辅助设计、信号处理和人机界面中的相关问题，具有培育和促进科学突破和工程实践的潜力。

1990 年，IEEE 举办了首届 IEEE Visualization Conference，汇集了一个由物理、化学、计算、生物医学、图形学、图像处理等交叉学科领域研究人员组成的学术群体。2012 年，为突出科学可视化的内涵，会议更名为 IEEE Conference on Scientific Visualization。

自 18 世纪后期统计图形学诞生后，针对抽象信息的视觉表达手段仍然在不断发展，被用于揭示数据及其他隐匿模式奥秘。与此同时，数字化的非几何的抽象数据（如金融交易、社交网络、文本数据等）大量涌现，促生了多维、时变、非结构化信息的可视化需求。

20 世纪 80 年代末，视窗系统的问世使得人们能够直接与信息进行交互。1988 上，著名的统计图形学学者 William Cleverland 在著作 *Dynamic Graphics for Statistics* 中详细总结了面向多变量统计数据的动态可视化手段。从 1995 年开始，出现了单独面向信息可视化的会议。

图 1-10 表示采用直接体可视化技术绘制的鳄鱼木乃伊。

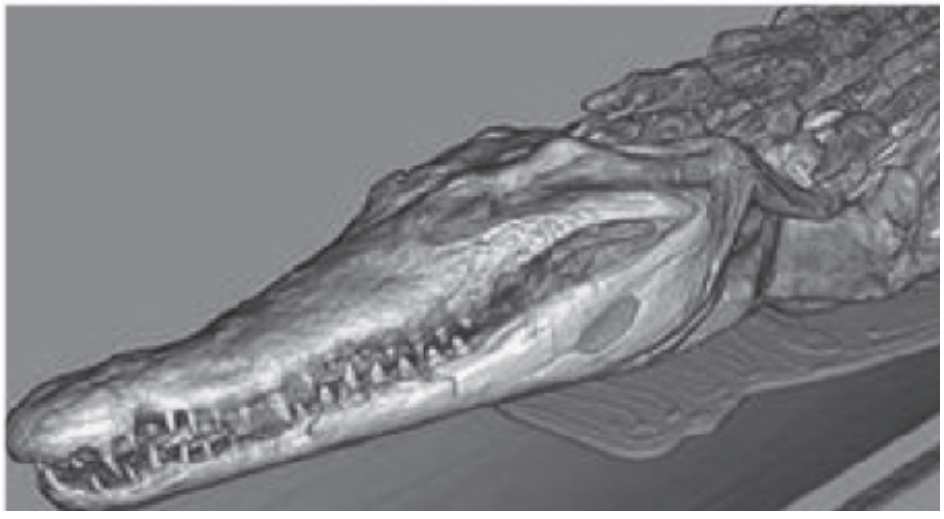


图 1-10　绘制鳄鱼木乃伊

9. 2004 年至今：可视分析学

进入 21 世纪，现有的可视化技术已难以应对少量、高维、多源和动态数据的分析挑战，需要综合可视化、图形学、数据挖掘理论与方法，研究新的理论模型、新的可视化方法和新的用户交互手段，辅助用户从大尺度、复杂、矛盾甚至不完整的数据中快速挖掘有用的信息，以便做出有效决策，这门新兴的学科称为可视分析学。

可视分析学是一门新兴的学科，其核心理论基础和研究方法尚处于探索阶段。从 2004 年起，研究界和工业界都朝着面向实际数据库、基于可视化的分析推理与决策、解

笔记

决实际问题等方向发展。

2005 年，美国国家科学基金会联合美国国家卫生研究所召集了一个新的专题小组，讨论可视化研究的现状和面临的挑战，并于 2006 年发布了一个专题报告描述大规模可视化所面临的挑战。与此同时，2004 年美国国土安全部为了应对恐怖袭击，成立了国家可视分析中心，2005 年发布的“可视分析的研究和发展规划”报告全面阐述了可视分析的挑战。

值得注意的是，可视分析的基本理论与方法仍然是正在形成且需要深入探讨的前沿科学问题。从 20 世纪 90 年代开始，我国的各大科研单位和科研人员已经在可视化领域投入了极大的精力，为应用领域认识和使用可视化奠定了坚实的基础。尽管如此，先进的可视分析软件和算法在国内仍尚未得到普遍的理解。注意，我国的数据采集、分析与应用应当自主研发，不能任由国外垄断公司采集和处理，否则将危及国民生活与国防安全。我国急需对可视分析的基础理论和方法展开研究，对涉及国家大工程、国家安全、国民经济等重要领域数据的可视分析研究应自主进行。

1.3 数据可视化详解

数据可视化旨在借助图形化手段清晰有效地传达与沟通信息。但是，这并不意味着数据可视化就一定因为要实现其功能用途而令人感到枯燥乏味，或者是为了看上去绚丽多彩而显得极端复杂。为了有效地传达思想概念，美学形式与功能需要应齐头并进，通过直观地传达关键的方面与特征，从而实现对相当稀疏而又复杂的数据集的深入洞察。设计人员往往由于把握不好设计与功能之间的平衡，从而创造出华而不实的数据可视化形式，无法达到主要目的，也就是无法传达与沟通信息。

1.3.1 数据科学的发展

信息科学领域面临的一个巨大挑战是数据爆炸。然而，人类分析数据的能力已经远远落后于获取数据的能力。这个挑战不仅在于数据量越来越大、高维、多元源、多态，更重要的是数据获取的动态性、数据内容的噪声和互相矛盾、数据关系的异构与异质性等。2013 年 3 月，美国政府发布了“大数据研究和发展倡议”，提出“通过收集、处理庞大而复杂的数据信息，从中获得知识和洞见，提升能力，加快科学、工程领域的创新步伐，强化美国国土安全，转变教育和学习模式”。至此，学术界达成共识，即关于数据的特定科学研究成为一门新的学科：数据科学。

在信息管理、信息系统和知识管理学科中，最基本的模型是“数据、信息、知识、智慧（Data, Information, Knowledge, Wisdom, DIKW）”层次模型。它以数据为基层架构，按照信息流顺序依次完成数据到智慧的转换。四者之间的结构和功能方面的关系构成了信息科学的基础理论。在数据科学中，这种模型也作为一种数据处理流程，完成从原始数据的转化。

笔记

1. 数据

从信息获取的角度看，数据是对目标观察和记录的结果，是关于现实世界中的时间、地点、事件、其他对象或概念的描述。在表达为有用的形式之前，数据本身没有用途，关于数据，不同的学者给出了不同的定义，大致分为以下 3 类。

（1）数据即事实：数据是未经组织和处理的、离散的、客观的观察。由于缺乏上下文和解释，所以数据本身没有含义和价值。如果将事实定义为真实的、正确的观察，那么并不是所有的数据都是事实，错误的、无意义的和非感知的数据不属于事实。

（2）数据即信号：从获取的角度理解，数据是基于感知的信号刺激或信号输入，包括视觉、听觉、嗅觉、味觉和触觉。由于每种感官都对应了某个信息通道，所以数据也可定义为某个器官能接收到的一种或多种能量波或能量粒子（如光、热、声、力和电磁等）。

（3）数据即符号：无论数据是否有意义，数据都可定义为表达感官刺激或感知的符号集合，即某个对象、事件或所处环境的属性。代表性的符号，如单词、数字、图表和图像视频等，都是人类社会中用于沟通的基本手段。因此，数据就是记录或保存的事件或环境的符号。

2. 信息

信息和数据的区别在于信息是有用的、有意义的，可以回答诸如认证、什么、哪里、多少、什么时候等问题，因此可以赋予数据生命力，辅助用户决策或行动。进一步讲，信息可以采用描述的方式定义知识。信息有如下两类特性。

（1）结构性与功能性：信息是组织好的结构化数据，与某个特定目标和上下文关联，因此是有意义的、有价值的、有关联的。从这个意义上说，信息和数据的判别在于结构，而不是两者的功能。

（2）象征性或主体性：信息是通用的、以符号和信号形式存在的数据。另一个观点则认为，信息具有主体性，符合所依附的对象。

3. 知识

知识是一个隐晦的、意会的、难以描述和定义的概念，是被处理、组织过、应用或付诸行动的信息。知识又是框架化的经验、价值、情境信息、专家观察和基本直觉的流动的混合，它提供了一个环境和框架，用于评估和融入新的经验和信息。知识是原语，应用于知识者的意识之中。知识通常体现于文档和资料的描述中，也流转于组织机构的流程、处理和实践中。

（1）知识即处理：与信息是组织化或结构化数据的定义相似，知识既是多个信息源在时间上的合成，也是情境信息、价值、经验和规则的混合，也可看成互联的信息。

（2）知识即过程：知识是一个通过实践经验了解如何做、是谁、什么时候等“Know-How”的过程。知识从经验背景中引申出一个连贯和自我一致的协调性行为。如果信息是描述性的，那么知识并不是对行动的描述，而是意味着行动。也有人将知识定义为数据和信息的应用。

笔记

（3）知识即命题：知识有时候被认为是信念的构建、与认知框架有关的外部化。知识的另一个定义是，主观的关于世界和所在环境的感知：关于对象（整体、联合）的独特性观察。

4. 智慧

智慧是启示性的，本意是知道为什么，知道如何去做。智慧和信息的区别等价于为什么做和为什么是。在知识和智慧之间存在一种状态：理解，它是一种对为什么的欣赏。智慧可增加有效性和价值，它蕴涵的伦理和美学的价值与主体一脉相承，并且是独特和个性化的。

1.3.2 数据可视化的意义

在 DIKW 模型所定义的数据转化为智慧的流程中，可视化借助人眼快速的视觉感知和人脑的智能认知能力可以起到清晰有效地传达、沟通并辅助数据分析的作用。现代的数据可视化技术综合运用计算机图形学、图像处理、人机交互等技术，将采集或模拟的数据变换为可识别的图形符号、图像、视频或动画，并以此呈现对用户有价值的信息。用户通过对可视化的感知，使用可视化交互工具进行数据分析，获取知识，并进一步提升为智慧。

关于数据可视化适用范围，存在不同的观点。例如，有专家认为数据可视化是可视化的一个子类目，主要处理统计图形、抽象的地理信息或概念型的空间数据。现代的主流观点将数据可视化看成传统的科学可视化和信息可视化的泛称，即处理对象可以是任意数据类型、任意数据特性，以及异构异质数据的组合。大数据时代的数据复杂性更高，如数据的流模式获取、非结构化、语义的多重性等。

数据可视化的作用在于视物致知，即从看见物体到获取知识。对于复杂、大尺度的数据，已有的统计分析或数据挖掘方法往往是对数据的简化和抽象，隐藏了数据真实的结构，而数据可视化则还原乃至增强数据中的全局结构和具体细节。当然，数据可视化经常会陷入两个误区：为了实现其获取知识的功能而令人感到枯燥乏味；或者为了画面美观而采用复杂的图形。如果将数据可视化看成艺术创作过程，则数据可视化需要达到真、善、美的均衡，达到有效地挖掘、传播与沟通数据中蕴涵的信息、知识与思想，实现设计与功能之间的平衡。

真，即真实性，指是否正确地反映了数据的本质，以及对所反映的事物和规律有无正确的感受和认识。数据可视化之真是其基石。例如，在医学研究领域，数据可视化可以通过可视化不同形态的医学影像、化学检验、电生理信息、过往病史等，帮助医生了解病情发展、病灶区域，以拟定治疗方案。

善，即倾向性，也就是可视化所表达的意象对于社会和生活具有什么意义和影响。某专家认定，可视化的终极目标在于帮助公众理解人类社会发展和自然环境的现状，实现政府与职能部门运行的透明。

美，即可视化的艺术完美性，指其形式与内容是否和谐统一，是否有艺术个性，是否有创新和发展。

笔记

1.3.3 数据可视化的分类

数据可视化的处理对象是数据。自然地，数据可视化包含处理科学数据的科学可视化与处理抽象的、非结构化信息的信息可视化两个分支。广义上，面向科学与工程领域的科学可视化研究带有空间坐标和几何信息的三维空间测量数据、计算模拟数据和医学影像数据等，重点探索如何有效地呈现数据中的几何、拓扑和形状特征。数据可视化的处理对象则是非结构化、非几何的抽象数据，如金融交易、社交网络和文本数据，其核心挑战是如何针对大尺度高维数据减少视觉混淆对有用信息的干扰。另一方面，数据分析的重要性将可视化与分析结合形成一个新的学科：可视化分析学。科学可视化、信息可视化和可视化分析学三个学科方向通常被看成可视化的三个主要分支。

1. 科学可视化

科学可视化是可视化领域最早、最成熟的一个跨学科研究与应用领域，面向的领域主要是自然科学，如物理、化学、气象气候、航空航天、医学、生物学等各个学科，这些学科通常需要对数据和模型进行解释、操作与处理，旨在寻找其中的模式、特点、关系以及异常情况。

科学可视化的基础理论与方法已经相对成形。早期的关注点主要是三维真实世界的物理化学现象，因此数据通常表达在三维或二维空间，或包含时间维度。鉴于数据的类别可分为标量（密度、温度）、风向（风向、力场）、张量（压力、弥散）三类，科学可视化也可粗略分为三类。

（1）标量场可视化。标量是指单个数值，即在每个记录的数据点上有一个单一的值。标量场指二维、三维或四维空间中每个采样处都有一个标量值的数据场。标量场的来源分为两类：第一类从扫描或测量设备获得，如从医学断层扫描设备获取的 CT、MRI 三维影像；第二类从计算机或机器仿真中获得，如从核聚变模拟中产生的壁内温度分布。

（2）向量场可视化。向量场在每一个采样点是一个向量（一维数组）。向量代表某个方向或趋势，例如，来源于测量设备的风向和旋涡等；来源于数据仿真的速度和力量等。向量场可视化的主要关注点是其中蕴涵的流体模式和关键特征区域。在实际应用中，由于二维或三维流场是最常见的向量场，所以流场可视化是向量场可视化中最重要的组成部分。

（3）张量场可视化。张量是矢量的推广：标量可看作 0 阶张量，矢量可看作 1 阶张量。张量场可视化方法分为纹理、几何和拓扑三类。基于纹理的方法将张量场转换为静态图像或动态图像序列，图释张量场的全局属性。其思路是将张量场简化为向量场，进而采用线积分法、噪声纹理法等方法显示。基于几何的方法显式地生成刻画某类张量场属性的几何表达。其中，图标法采用某种几何形式表达单个张量，如椭球和超二次曲面：

笔记

超流线法将张量转换为向量（如二阶对称张量的主特征方向），再沿主特征方向进行积分，形成流线、流面或流体。基于拓扑的方法计算张量场的拓扑特征（如关键点、奇点、灭点、分叉点和退化线等），依次将感兴趣区域剖分为具有相同属性的子区域，并建立对应的图结构，实现拓扑简化、拓扑跟踪和拓扑显示。基于拓扑的方法可有效地生成多变量场的定性结构，快速构建全局流场结构，特别适合于数值模拟或实验模拟生成的大尺度数据。

以上分类不能概括科学数据的全部内容。随着数据的复杂性提高，一些带有语义的信号、文本、影像等也是科学可视化的处理对象，且其呈现空间变化多样。

2. 信息可视化

信息可视化处理的对象是抽象的、非结构化数据集合（如文本、图表、层次结构、地图、软件、复杂系统等）、传统的信息可视化起源于统计图形学，又与信息图形、视觉设计等现代技术相关，其表现形式通常在二维空间，因此关键问题是在有限的展现空间中以直观的方式传达大量的抽象信息。与科学可视化相比，信息可视化更关注抽象、高维数据，此类数据通常不具有空间中位置的属性，要根据特定数据分析的需求决定数据元素在空间的布局，因此信息可视化的方法与所针对的数据类型紧密相关。按数据类型划分，数据可视化大致分为如下 4 类。

（1）时空数据可视化。时间与空间是描述事务的必要因素，因此，地理信息数据和时空数据的可视化也显得至关重要。对于地理信息数据可视化来说，合理地选择和布局地图上的可视化元素，从而呈现尽可能多的信息是关键。时变数据通常具有线性和周期性两种特征，需要依此选择不同的可视化方法。

（2）层次与网络结构数据可视化。网络（图）数据是现实世界中最常见的数据类型之一。人与人之间的关系、城市之间的道路连接、科研论文之间的引用组成了网络。层次结构（树）则是有一个根节点，并且不存在回路的特殊网络，例如公司的组织结构、文件系统的目录结构、家谱等。层次与网络结构数据通常都使用点线图来可视化，如何在空间中有效地布局节点和连线是可视化的关键。

（3）文本和跨媒体数据可视化。随着网络媒体，特别是社交媒体的迅速发展，每天都会产生海量的文本数据，人们对于视觉符号的感知和认知速度远远高于文本，因此，通过可视化呈现其中蕴涵的有价值的信息将大大提高人们对这些数据的利用率。我们需要从非结构化文本数据中提取结构化信息，并进行可视化。

（4）多变量数据可视化。用于描述现实世界中复杂问题和对象的数据通常是多变量的高维数据，如何将其在二维屏幕上呈现是可视化面临的挑战。多变量数据的可视化方法包括将数据降维度空间，使用相互关联的多视图同时表现不同维度，等等。

在数据爆炸时代，信息可视化面临巨大的挑战：在海量、动态变化的信息空间中辅助人类理解、挖掘信息，从中检测预期的特征，并发现未预期的知识。

3. 可视化分析学

笔记

可视化分析学可定义为一门以可视交互界面为基础的分析推理科学。它综合了图形学、数据挖掘和人机交互等技术，以可视交互界面为通道，将人的感知和认知能力以可视的方式融入数据处理过程，形成人脑智能和机器智能优势互补和相互提升，建立螺旋式信息交流与知识提炼途径，完成有效的分析推理和决策。

新时期科学发展和工程实践的历史表明，智能数据分析所产生的知识与人类掌握的知识的差异正是导致新的知识发现的根源，而表达、分析与检验这些差异必须充分利用人脑智能。另外，当前的数据分析方法大都基于先验模型，易于检测已知模式和规律，对复杂、异构、大尺度数据的自动处理经常会失效，例如，不知道数据中蕴含的模式；搜索空间过大；特征模式过于模糊；参数很难设置，等等。而人的视觉识别能力和智能恰好可以辅助解决这些问题。另外，自动数据分析的结果通常带有噪声，需要人工干预排除。为了有效结合人脑智能与机器智能，一个必经途径是以视觉感知为通道，通过可视交互界面，形成人脑和机器智能的双向转换，将人的智能，特别是“只可意会，不能言传”的人类知识和个性化经验，可视地融入整个数据分析和推理决策过程中，使得数据的复杂度逐步降低到人脑和机器智能可处理的范围。这个过程逐渐形成了可视分析这一交叉信息处理的新思路。迄今为止，可视分析的基本理论与方法仍然是一个有待解决的新课题，值得深入研究。

可视化分析学可看成集成可视化、人的因素和数据分析的一种新思路。其中，感知与认知科学研究人在可视化分析学中的重要作用：数据管理与知识表达是可视分析构建数据转换的基础理论；地理分析、信息分析、科学分析、统计分析、知识发现是可视化分析学的核心分析论方法；在整个可视分析过程中，人机交互必不可少，用于驾驭模型构建、分析推理和信息呈现等整个过程；可视分析流程中推导出的结论与知识最终需要向用户表达和传播。

可视化分析学是一门综合性学科，与多个领域相关：在可视化方面，有信息可视化、科学可视化与计算机图形学；与数据分析相关的领域包括信息获取、数据处理和数据挖掘；而在交互方面，则有人机交互、认知科学和感知等学科融合。

科学可视化、信息可视化和可视分析三者之间没有清晰的边界。科学可视化的研究重点是带有空间坐标和几何信息的医学影像数据、三维空间信息测量数据、流体计算模拟数据等。由于数据的规模通常超过图形硬件的处理能力，所以如何快速地呈现数据中包含的几何、拓扑、形状特征和演化规律是其核心问题。随着图形硬件和可视化算法的迅猛发展，单纯的数据显示已经得到了较好的解决。信息可视化的核心问题主要有高维数据的可视化、数据间各种抽象关系的可视化、用户的敏捷交互和可视化有效性的诊断等。可视分析侧重于从各类数据综合、意会和推理出知识，其实质是可视地完成机器智能和人脑智能的双向转换，整个探索过程是迭代的、螺旋式上升的过程。

笔记

1.3.4 数据可视化与其他学科的关系

数据可视化既与信息图、信息可视化、科学可视化以及统计图形密切相关，也是数据科学中必不可少的环节。数据科学在研究、教学和工业界等领域方兴未艾，数据可视化是一个活跃且关键的方面。下面简单总结数据可视化与其他学科领域的关系。

1. 图形学、人机交互

计算机图形学是一门通过软件生成二维、三维或四维动态影像的学科。起初，可视化通常被认为是计算机图形学的子学科。通俗地说，计算机图形学关注数据的空间建模、外观表达与动态呈现，它为可视化提供数据的可视编码和图形呈现的基础理论与方法。数据可视化则与具体应用和不同领域的数据密切相关。由于可视化分析学的独特属性以及数据分析之间的紧密结合，数据可视化的研究内容和方法已经逐渐独立于计算机图形学，形成一门新学科。

计算机动画是图形学的子学科，是视频游戏、动漫、电影特效中的关键技术，它以计算机图形学为基础，在图形生成的基本范畴下延伸出时间轴，通过在连贯的时间轴上呈现相关的图像表达某类动态变化。计算机动画主要包括二维动画、三维动画、非真实感动画等门类。数据可视化采用计算机动画这种表现手法展现数据的动态变化，或者发掘时空数据中的内在规律。

计算机仿真指采用计算设备模拟特定系统的模型。这些系统包括物理学、计算物理学、化学以及生物学领域的天然系统；经济学、心理学以及社会科学领域的人类系统。它是数学建模理论的计算机实践，能模拟现实世界上难以实现的科学实验、工程设计和规划、社会经济预测等运行情况或者行为表现，允许反复试错，既节约了成本，又可提高效率。随着计算硬件和算法的发展，计算机仿真能模拟的规模和复杂性已经远远超出传统数学建模所能企及的高度。因而，大规模计算仿真被认为是继科学实验与理论推导之后，科学探索和工程实践的第三推动力。计算机仿真获得的数据是数据可视化的处理对象之一，而将仿真数据以可视化形式表达是计算机仿真的核心方法。

人机交互指人与机器之间使用某种语言以一定的交互方式，为完成确定任务的信息变换过程。人机交互是信息时代数据获取与利用的必要途径，是人与机器之间的信息通道。人机交互与计算机科学、人工智能、心理学、社会学、图形、工业设计等相关。在数据可视化中，通过人机界面接口实现用户对数据的理解和操纵，数据可视化的质量和效率需要最终的用户评判。因此，数据、人、机器之间的交互是数据可视化的核心。

2. 数据库与数据仓库

数据库是按照数据结构组织、存储和管理数据的仓库，它高效地实现数据的录入、查询、统计等功能。尽管现代数据库已经从最简单的存储数据表格发展到海量、异构数据存储的大型数据库系统，但是它的基本功能中仍然不包括复杂数据的关系和规则的分析。数据可视化通过数据的有效呈现，有助于对复杂关系和规则的理解。

面对海量信息的需要，数据库的一种新的应用是数据仓库。数据仓库是面向主题的、集成的、相对稳定的、随时间不断变化的数据集合，用以支持决策制订的过程。在数据进入数据仓库之前，必须经过数据加工和集成。数据仓库的一个重要特性是稳定性，即数据仓库反映的是历史数据。

笔记

数据库和数据仓库是大数据时代数据可视化方法中必须包含的两个环节。为了满足复杂大数据的可视化需求，必须考虑新型的数据组织管理和数据仓库技术。

3. 数据分析与数据挖掘

数据分析是统计分析的扩展，指用数据统计、数值计算、信息处理等方法分析数据，采用已知的模型分析数据，计算与数据匹配的模型参数。常规的数据分析包含三步：第一步，探索性数据分析，通过数据拟合、特征计算和作图造表等手段探索规律性的可能形式，确定相适应的数据模型和数值解法；第二步，模型待定分析，在探索性分析的基础上计算若干类模型，通过进一步分析挑选模型；第三步，推断分析，使用数理统计等方法推断和评估选定模型的可靠性和精确度。

不同的数据分析任务各不相同。例如，关系图分析的 10 个任务是：值检索、过滤、衍生值计算、极值的获取、排序、范围确定、异常检测、分布描述、聚类、相关值。

数据挖掘指从数据中计算适合的数据模型，分析和挖掘大量数据背后的知识，它的目标是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、未知的、潜在有用的信息和知识。数据挖掘的方法可以是演绎的，也可以是归纳的。数据挖掘可发现很多类型的知识——反映同类事物共同性质的广义型知识；反映事物各方面特征的特征型知识；反映不同事物之间属性差别的差异型知识；反映事物和其他事物之间依赖或关联的关联型知识；根据当前历史和当前数据推测未来数据的预测型知识；揭示事物偏离常规出现异常现象的偏离型知识。

数据可视化和数据分析与数据挖掘的目标都是从数据中获取信息与知识，但手段不同。两者已成为科学探索、工程实践与社会生活中不可缺少的数据处理和发布的手段。数据可视化将数据呈现为用户易于感知的图形符号，让用户交互地理解数据背后的本质；而数据分析与数据挖掘通过计算机自动或半自动地获取数据隐藏的知识，并将获取的知识直接给予用户。

数据挖掘领域注意到了可视化的重要性，提出了可视数据挖掘的方法，其核心是将原始数据和数据挖掘的结果用可视化方法予以呈现。这种方法糅合了数据可视化的思想，但利用的仍然是机器智能挖掘数据，与数据可视化基于视觉化思考的大方针不同。

值得注意的是，数据挖掘与数据可视化是处理和分析数据的两种思路。数据可视化更善于探索性数据的分析，例如，用户不知道数据中包含什么样的信息和知识；而数据模型没有一个预先的探索假设，只是探寻数据中到底存在何种有意义的信息。

4. 面向领域的可视化方法与技术

数据可视化是对各类数据的可视化理论与方法的统称。在可视化历史上，与领域专家

笔记

的深度结合导致面向领域的可视化方法与技术。

生命科学可视化，指面向生物科学、生物信息学、基础医学、转化医学、临床医学等一系列生命科学探索与实践中产生的数据的可视化方法。它本质上属于科学可视化。由于生命科学的重要性，以及生命科学数据的复杂系统，生命科学可视化已经成为一个重要的交叉型研究方向。2011 年，IEEE VIS 举办了面向生命科学的可视化研讨会。

表意性可视化，指以抽象、艺术、示意性的手法阐明、解释科技领域的可视化方法。早期的表意性可视化以人体为描绘对象，类似于中学的生理卫生课本和大专医科院校的解剖课程上的人体器官示意图。在科学向文明转化的传导过程中迸发了大量需要表意性可视化的场合，如教育、训练、科普和学术交流等。在数据爆炸时代，表意性可视化关注的重点是从采集的数据出发，以传神、跨越语言障碍的艺术表达力展现数据的特征，从而促进科技生活的沟通交流，体现数据、科技与艺术的结合。例如，*Nature* 和 *Science* 杂志大量采用科技图展现重要的生物结构，澄清模糊概念，突出重要细节，并展现人类视角所不能及的领域。

地理信息可视化，是数据可视化与地理信息系统学科的交叉方向，它的研究主体是信息数据，包括建立于真实物理世界基础上的自然性和社会性事物及其变化规律。地理信息可视化的起源是二维地图制作。在现代，地理信息数据扩充到三维空间、动态变化，甚至包括在地理环境中采集的各种生物性、社会性感知数据（如天气、空气污染、出租车位置信息等）。

产品可视化，指面向制造和大型产品组装过程中的数据模型、技术绘图和相关信息的可视化方法。它是产品生命周期管理中的关键部分。产品可视化通常提供高度的真实感，以便对产品进行设计、评估与检验，因此支持面向销售和市场营销的产品设计或成型。产品可视化的雏形是手工生成的二维技术绘图或工程绘图。随着计算机图形学的发展，它逐步被计算机辅助设计替代。

教育可视化，指通过计算机模拟仿真生成易于理解的图像、视频或动画，用于面向公众教育和传播信息、知识与理念的方法。教育可视化在阐述难以解释或表达的事物（如原子结构、微观或宏观事物、历史事件）时非常有用。美国宇航局等机构专门成立了可视化部门，制作传播自然科学的教育可视化作品。

系统可视化，指在可视化基本算法中融合了叙事型情节、可视化组件和视觉设计等元素，用于解释和阐明复杂系统的运行机制与原理，向公众传播科学知识的方法。它综合了系统理论、控制理论和基于本体论的知识表达等，与计算机仿真和教育可视化的重合度较高。

商业智能可视化，又称为可视商业智能，指在商业智能理论与方法发展过程中与数据可视化融合的概念和方法。商业智能的目标是将商业和企业运维中收集的数据转化为知识，辅助决策者做出明智的业务经营决策。数据包括来自业务系统的订单、库存、交易账目、客户和供应商等，以及其他外部环境中的各种数据。从技术层面上看，商业智能是数据仓库、联机分析处理工具和数据挖掘等技术的综合运用，其目的是使各级决策者获得知

识或洞察力。自然地，商业智能可视化专门研究商业数据的智能可视化，以增强用户对数据的理解力。

笔记

知识可视化，采用可视表达表现与传播知识，其可视化形式包括素描、图表、图像、物件、交互式可视化、信息可视化应用以及叙事型可视化。与信息可视化相比，知识可视化侧重运用各种互为补充的可视化手段和方法，面向群体传播认识、经验、态度、价值、期望、视角、主张和预测，并激发群体协同产生新的知识。知识可视化与信息论、信息科学、机器证明、知识工程等方法各有异同，其特点是使发现知识的过程和结果易于理解，且在发现知识过程中通过人机交互界面发展发现知识的可视化方法。

5. 信息视觉设计

面向广义数据的视觉设计，是信息设计中的一个分支，可抽象为某种概念性形式，如属性、变量的某种信息。这又包含了两个主要领域：统计图形学和信息图。它们都与量化和类别数据的视觉表达有关，但被不同的表述目标驱动。统计图形学应用于任意统计数据相关的领域，它的大部分方法（如散点图、热力图等方法）已经是信息可视化最基本的方法。

信息图受限于二维空间上的视觉设计，偏重艺术的表达。信息图和可视化之间有很多相似之处，它们的共同目标是面向探索与发现的视觉表达。特别地，基于数据生成的信息图和可视化在现实应用中非常接近，且有时能互相替换，但两者的概念是不同的：可视化指用程序生成的图形图像，这个程序可以应用于不同的数据；信息图指为某一数据定制的图形图像，它是具体化的、自解释性的，而且往往是设计者手工定制的，只能应用于特定数据。由此可以看出，可视化的强大普适性能够使用户快速地将某种可视化技术应用于不同数据，但选择适合的数据可视化技术却依赖于用户的经验和运气。

与视觉设计相关的图形学是一个传统的基础性研究方向，它关注图、树等非结构化数据结构，设计表达力强的可视表达与可视编码方法。

将视觉设计、社会媒体与营销结合，则产生一个新的学科方向：视觉传播。它通过信息的可视化展现沟通与传播创意和理念，在网页设计和图形向导的可用性方面作用明显。视觉传播与艺术和设计关联度高，通常以二维图表形式存在，包括字符艺术、符号、电子资源等。

考虑到非空间的抽象数据，数据可视化的可视表达与传统的视觉设计类似，然而，数据可视化的应用对象和处理范围远远超过统计图形学、视觉艺术与信息设计等学科方向。

1.4　数据可视化研究的挑战

人类有史以来，可视化的理念就伴随着形象思维、图画、摄像等方法不断演化。现代意义上的可视化是计算机与计算机显示方法与设备发展到一定阶段后的新兴技术。尽

笔记

管显示方法和技巧各有差异，但是数据可视化的研究实质仍然是两个方面：理解可视化如何传递到观者，即人们感知和理解什么，可视化是如何对应数据和数据模型的；开发能有效地创造可视化的原理与技术，即增强认知与感知，增强可视化与数据模型之间的联系。

分析可视化系统时，设计者至少要考虑三个不同方面的约束：计算能力、感知和认知能力以及显示能力。

（1）计算能力的可扩展性。可视化系统与设计目标的应用场合有关。在大数据时代，具备处理海量的复杂数据的可扩展性始终是可视分析系统关注的中心议题。由于有限的时间和存储资源，通常可视化的效率受限于可用的时间和存储资源，面向大数据的数据清洗、转换、布局和绘制算法的计算复杂度是主要关注对象。

（2）感知和认知能力的局限性。人类的记忆容量和注意力是宝贵的、有限的资源。尽管可视化充分利用人类视觉的感知能力，但是人类大脑对事物的记忆终究是不可分的，而且记忆容量极其有限。这种有限性不分视觉和非视觉，也不分长期和短期的记忆。人类的注意力非常有限，例如，在有意识地查找某项内容时，随着检查项数量增加，任务变得非常具有挑战性。另一方面，警觉性同样是高度有限的资源，前几分钟的警觉性要远超之后时间段的警觉性，因此执行视觉搜索任务的能力只能维持数分钟。

（3）显示能力的局限性。可视化设计者往往“执行于像素之外”，屏幕的分辨率已经不能同时显示所有想要表达的信息。单个像素的信息密度表示为编码后信息的数量与未使用空间数量的比值。一次尽可能多地显示以减少导航，而一次显示太多代价又比较高，用户会产生视觉混乱，这需要权衡。

围绕这三方面的局限性，未来数据可视化的挑战主要在两个方面。

1. 大数据可视化

数据密集型科学成为继实验理论和计算仿真之后，科学研究手段的第三种范式。从海量涌现的数据中获取知识，验证科学假设，是科学前进和社会发展的驱动力。大数据的研究需要从国家战略高度认识大数据并开始行动，其着力点不仅在于进一步推进信息化建设，更在于以数据推动科研和创新。显而易见，大数据将引发新的智慧革命；从海量、复杂、实时的大数据中可以发现知识、提升智能并创造价值。面向大数据，需要发展新的计算理论、数据分析、可视分析和数据组织与管理方法，并围绕实际科学和社会问题的求解设计新的工作流程和研究范式。

2. 以人为中心的探索式可视分析

发展到 21 世纪的可视化是一个涉及数据挖掘、人机交互、计算机图形学、心理学等的交叉学科。在信息科学领域，分析可定义为一个“从数据中洞悉规律，以便更好地决策的科学过程”。如何将可视化与分析有机地结合，开发高度集成的可视分析系统是未来一个重大的研究课题。

可视化分析学的基本要素包括复杂数据的表示与变换、可扩展的数据智能可视化和支

持用户分析决策的交互方法与集成环境学等。它引导的分析推理模式是探索复杂数据中蕴含的新规律和新现象的催化剂。21 世纪以来，国际上逐步形成可视化分析学的研究热潮。可视化分析必将在国民经济、社会生活和国防安全的各个领域引申出重大应用难题，如天气预报、防灾减灾、数字城市、金融安全、社会网络等。如何结合相关学科的方法，研发面向各个应用领域的高效可视分析系统是一个持久的研究话题。

笔记