

第1章 数据可视化简介

1 可视化释义

释义

➤ 视觉是获取信息的最重要通道

- 超过50%的人脑功能用于视觉的感知，包括解码可视信息、高层次可视信息处理，和思考可视符号

➤ 人眼(视觉感觉器官)是一个高带宽的巨量视觉信号输入并行处理器

- 最高带宽为每秒100MB
- 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多个数量级
- 且大量的视觉信息的处理发生在潜意识阶段

➤ 可视化对应

➤ Visualize, 动词

- 意即“生成符合人类感知”的图像；
- 通过可视元素传递信息

➤ Visualization, 名词

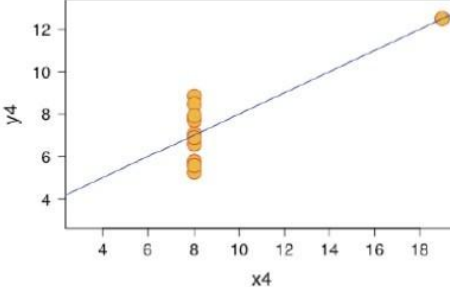
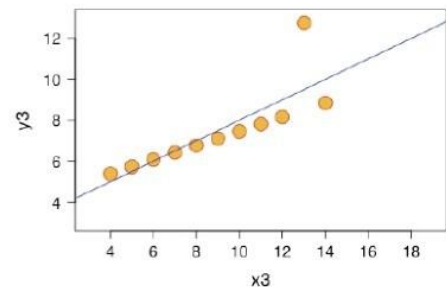
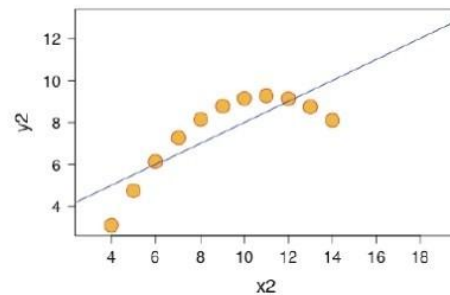
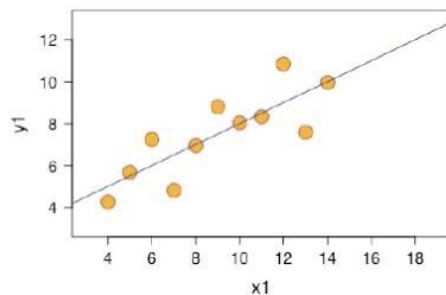
- 表达“使某物、某事可见的动作或事实”
- 对某个原本不可见的事物在人的大脑中形成一幅可感知的心理图片的过程或能力
- Visualization也可用于表达对某目标进行可视化的结果，即一帧图像或动画

可视化释义

► 可视化技术

- 利用人眼的感知能力对数据进行交互的可视表达以增强认知的技术
- 将不可见或难以直接显示的数据，转化为可感知的图形、符号、颜色、纹理等，增强数据识别效率，传递有效信息

x1		x2		x3		x4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



可视化的理解

- 可视化通常被理解为一个生成图形图像的过程。更深刻地，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
 - 可视化的终极目的，是对事物规律的洞悉，而非所绘制的可视化结果本身
 - 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“通过可视表达增强人们完成某些任务的效率”
- 信息加工角度
 - 丰富的信息消耗大量的注意力。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

可视化作用

➤可视化的作用体现在多个方面

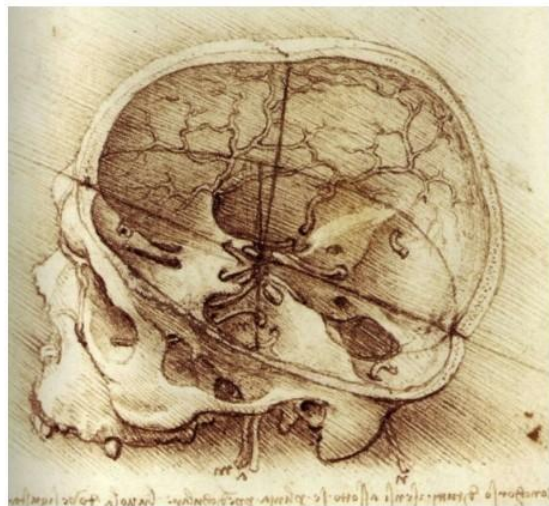
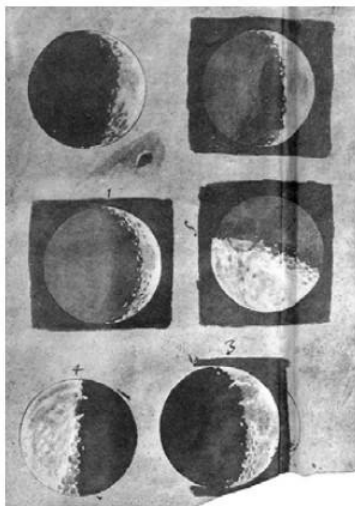
- 揭示想法和关系、形成论点或意见、观察事物演化的趋势、总结或积聚数据、存档和汇整、寻求真相和真理、传播知识和探索性数据分析等

➤从宏观的角度看，可视化包括三个功能

- 信息记录
- 支持对信息的推理和分析
- 信息传播与协同

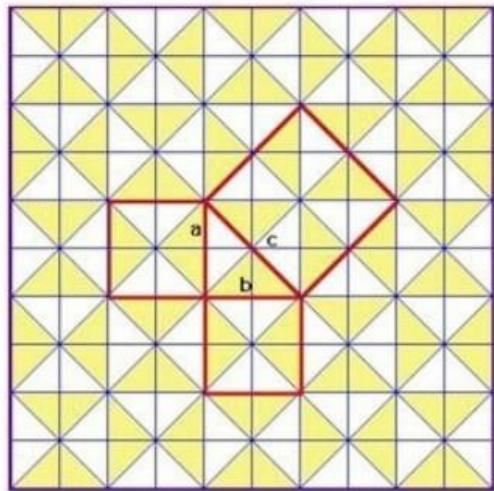
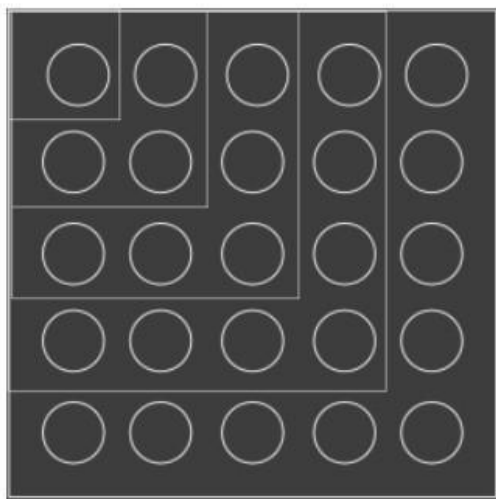
信息记录

- 将信息记录成文、进行传播的有效方式之一是将信息成像或采用草图记载
- 可视化图绘能极大地激发智力和洞察力，帮助验证科学假设

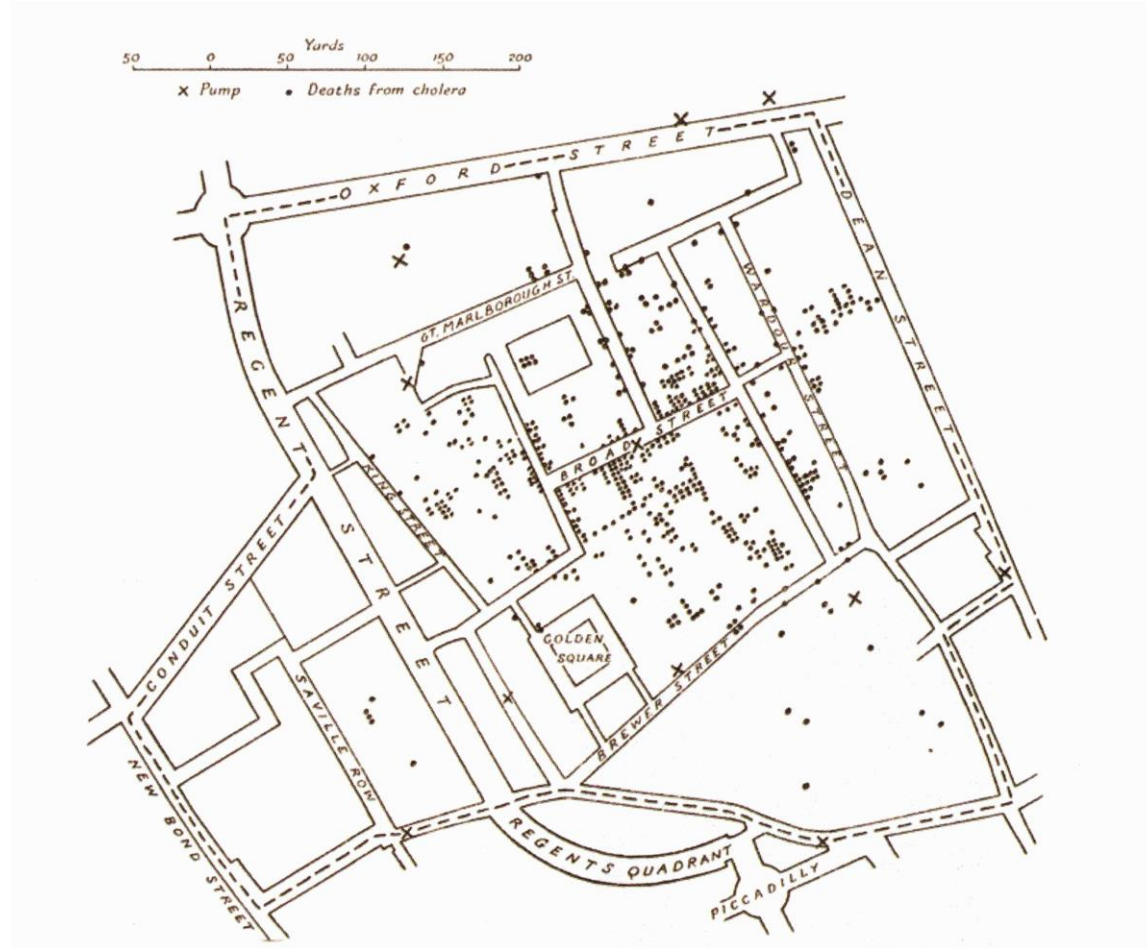


支持对信息的推理和分析

- 数据分析的任务通常包括定位、识别、区分、分类、聚类、分布、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
- 直观的信息感知机制，极大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。

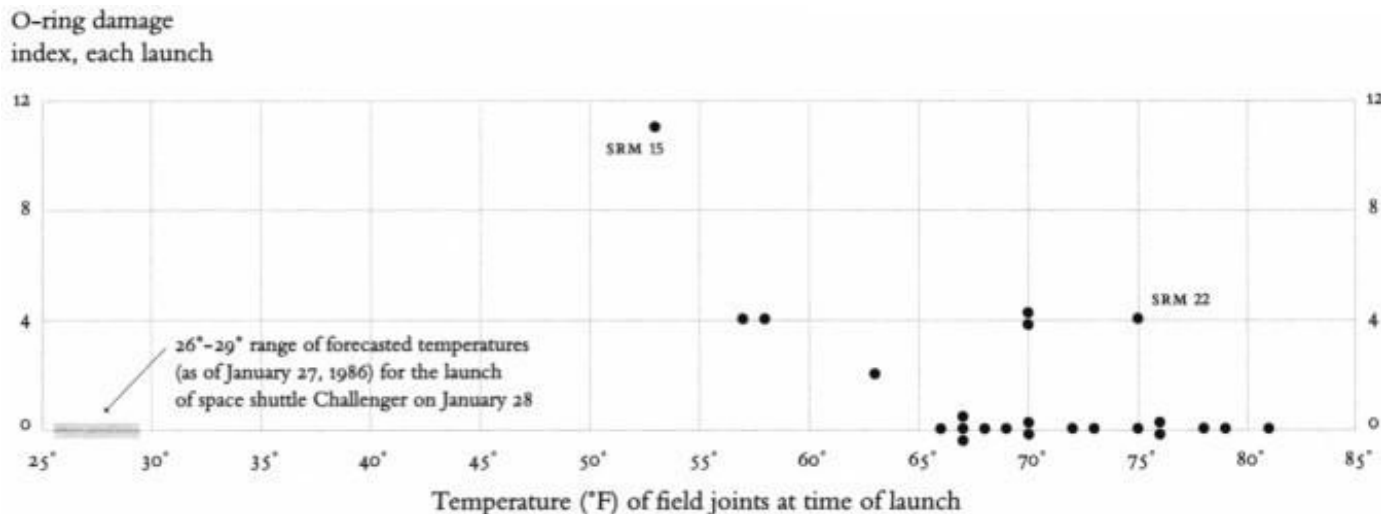


Ghost Map (1831)



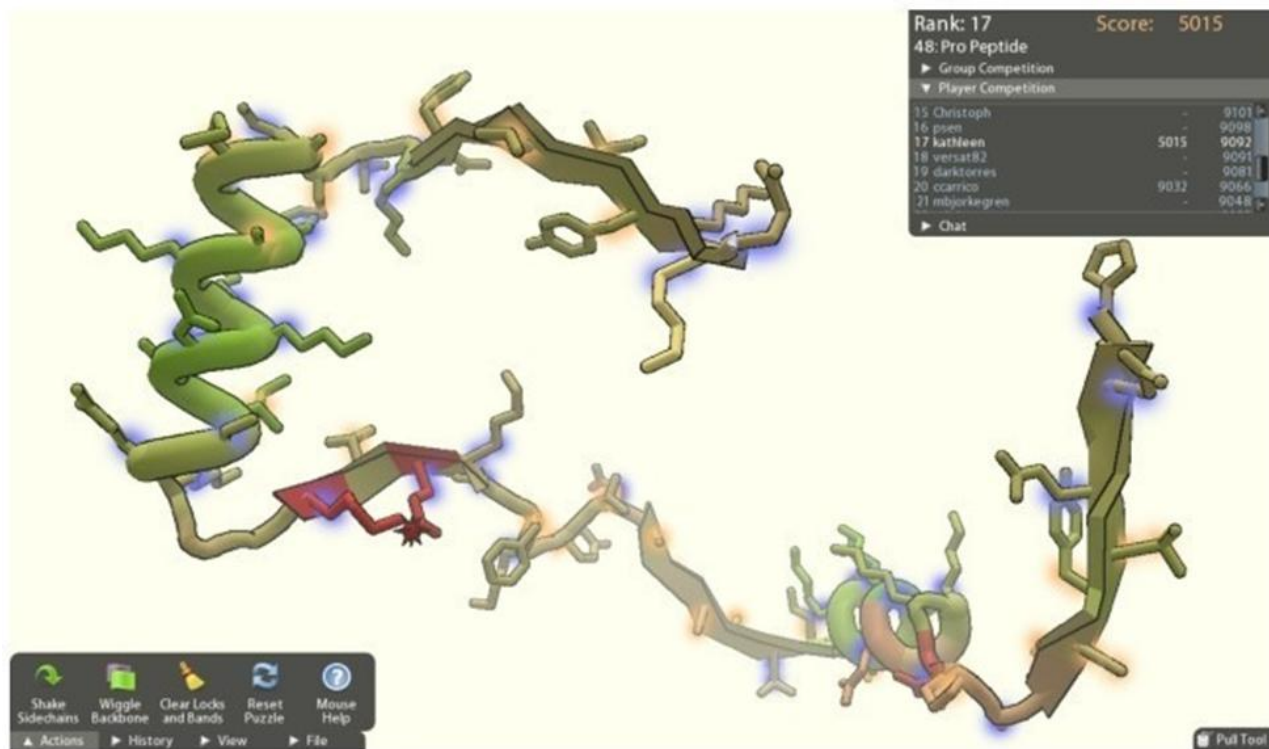
信息传播与协同

- 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%信息。
 - 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
 - 达到信息共享与论证、信息协作与修正、重要信息过滤等目的



在线游戏Fold.it

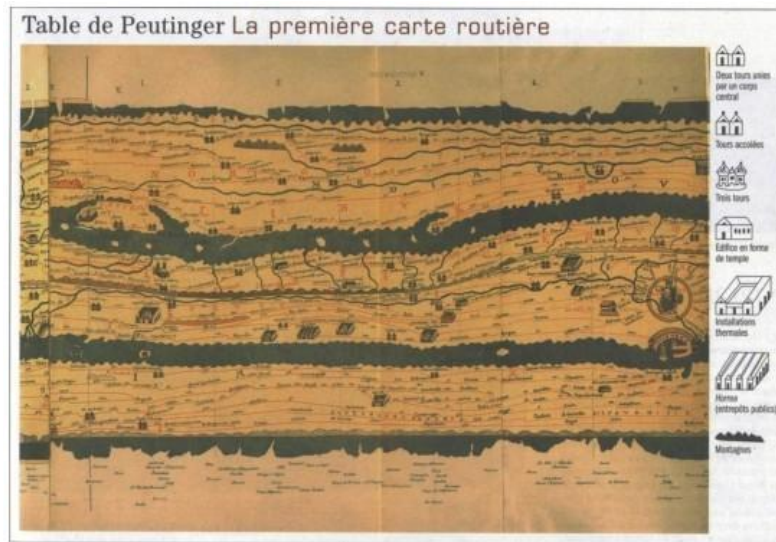
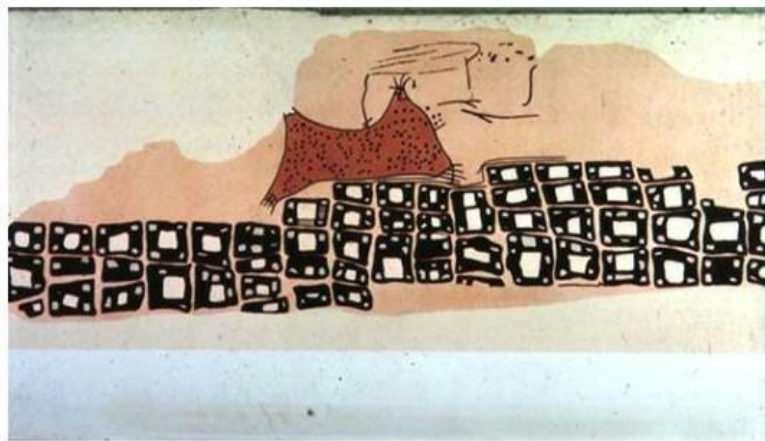
- 从半折叠的蛋白质结构起步，根据简单的规则扭曲蛋白质使之成为理想的形状



2 可视化简史

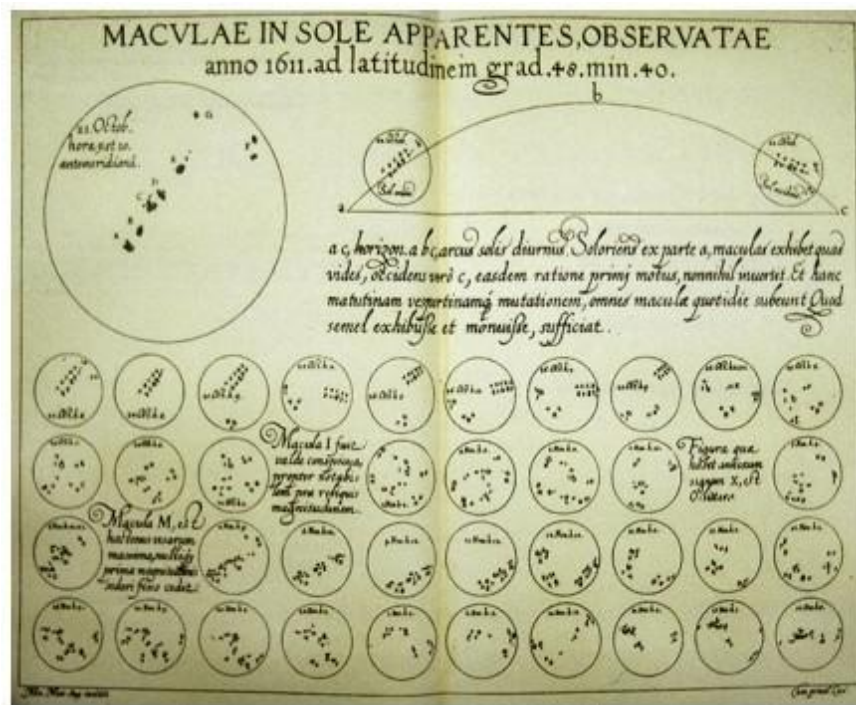
17世纪之前：图表萌芽

- 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可视化作品
- 可视化的萌芽出自几何图表和地图生成，其目的是展示一些重要的信息



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、距离和空间）的测量设备与理论的完善
- 制图学理论与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



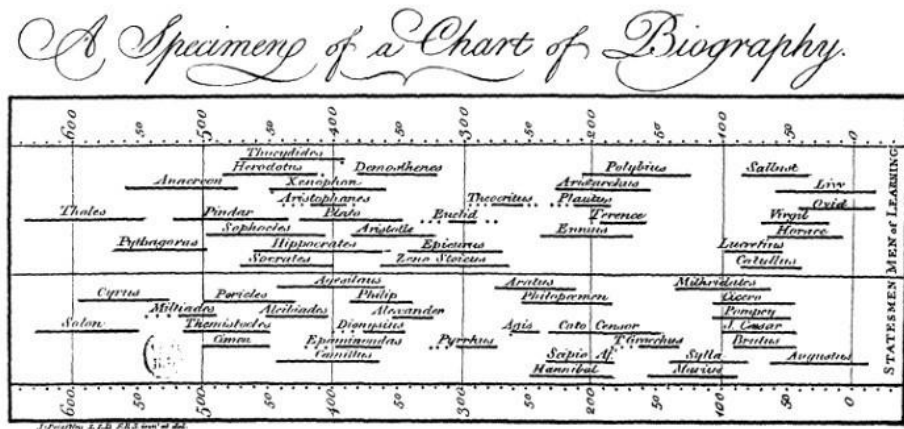
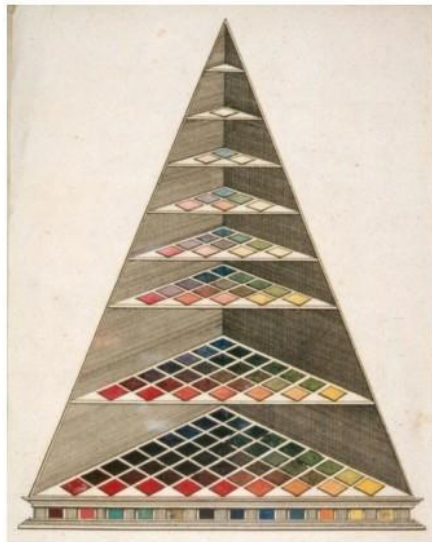
17世纪之前：图表萌芽

- 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



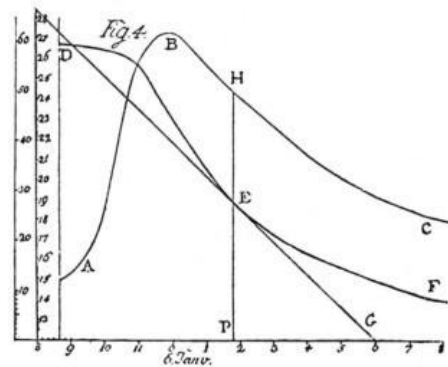
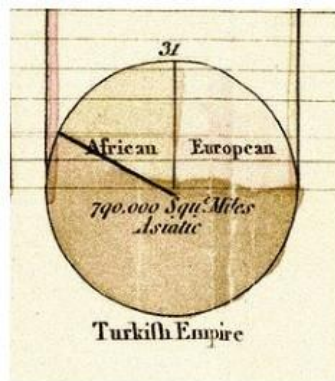
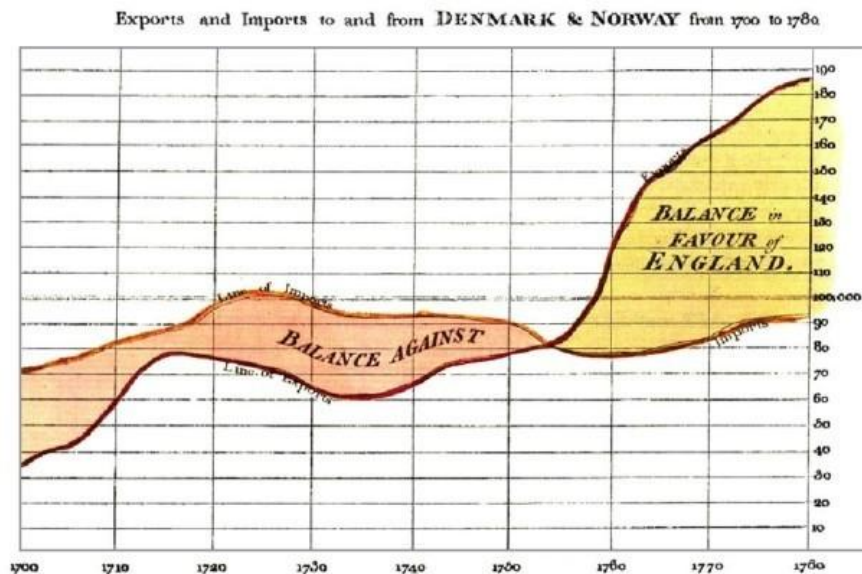
1700 - 1799年：图形符号

- 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）



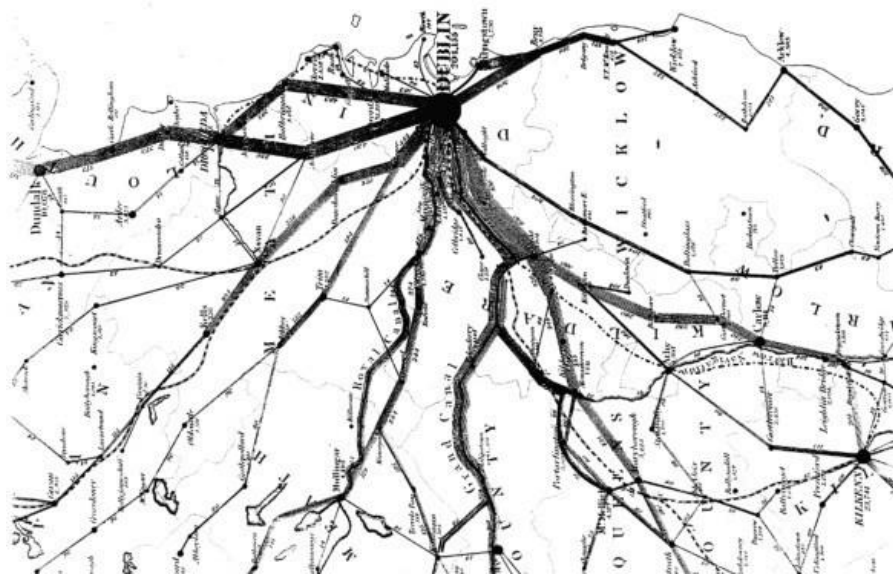
统计图形学

➤ 随着统计理论、实验数据分析的发展，抽象图和函数图被广泛发明

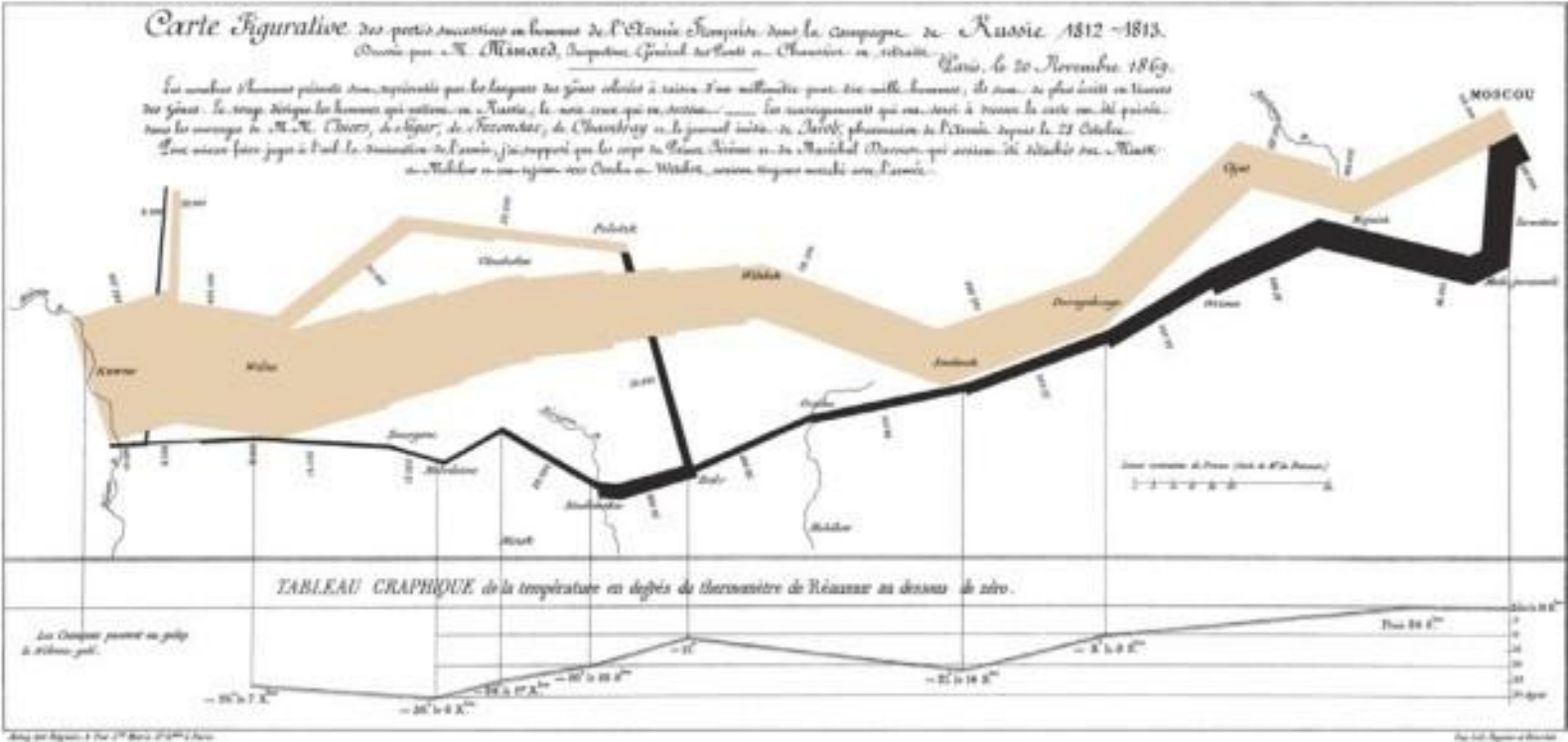


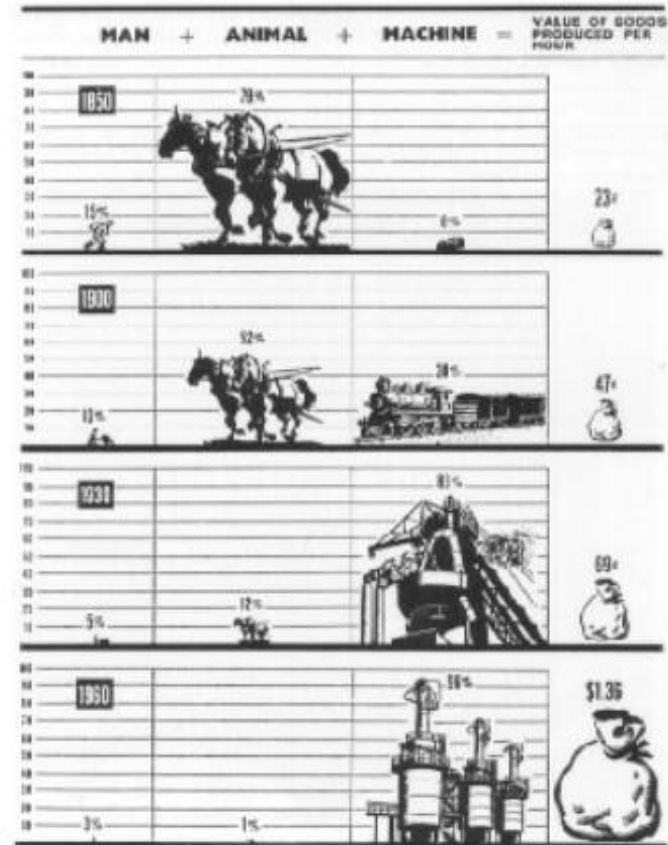
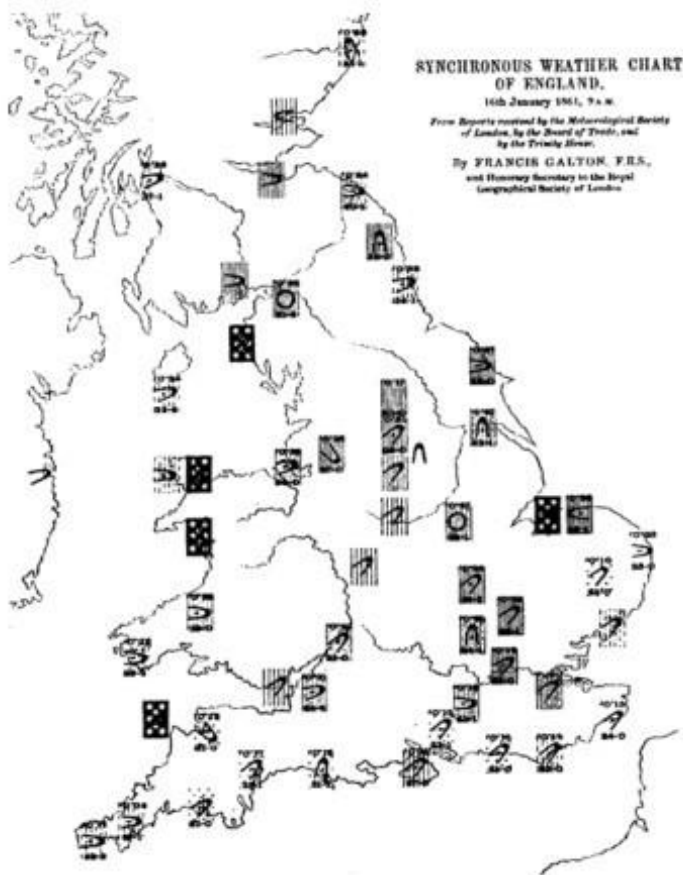
1800 - 1900年：数据图形

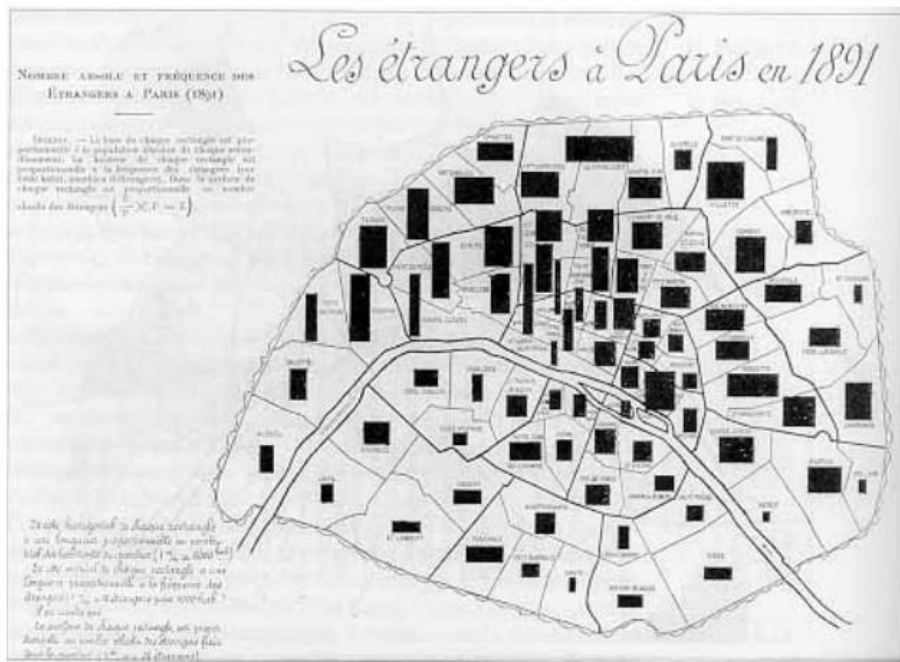
- 掌握了整套统计数据可视化工具。统计数据数据越来越多
- 将统计数据和其可视表达放在地图上，产生了概念制图的新思维，其作用开始体现在政府规划和运营中
- 采用统计图表来辅助思考的诞生同时衍生了可视化思考的新方式
 - 图表用于表达数学证明和函数
 - 列线图用于辅助计算
 - 各类可视化显示用于表达数据的趋势和分布，便于交流、获取和可视化观察



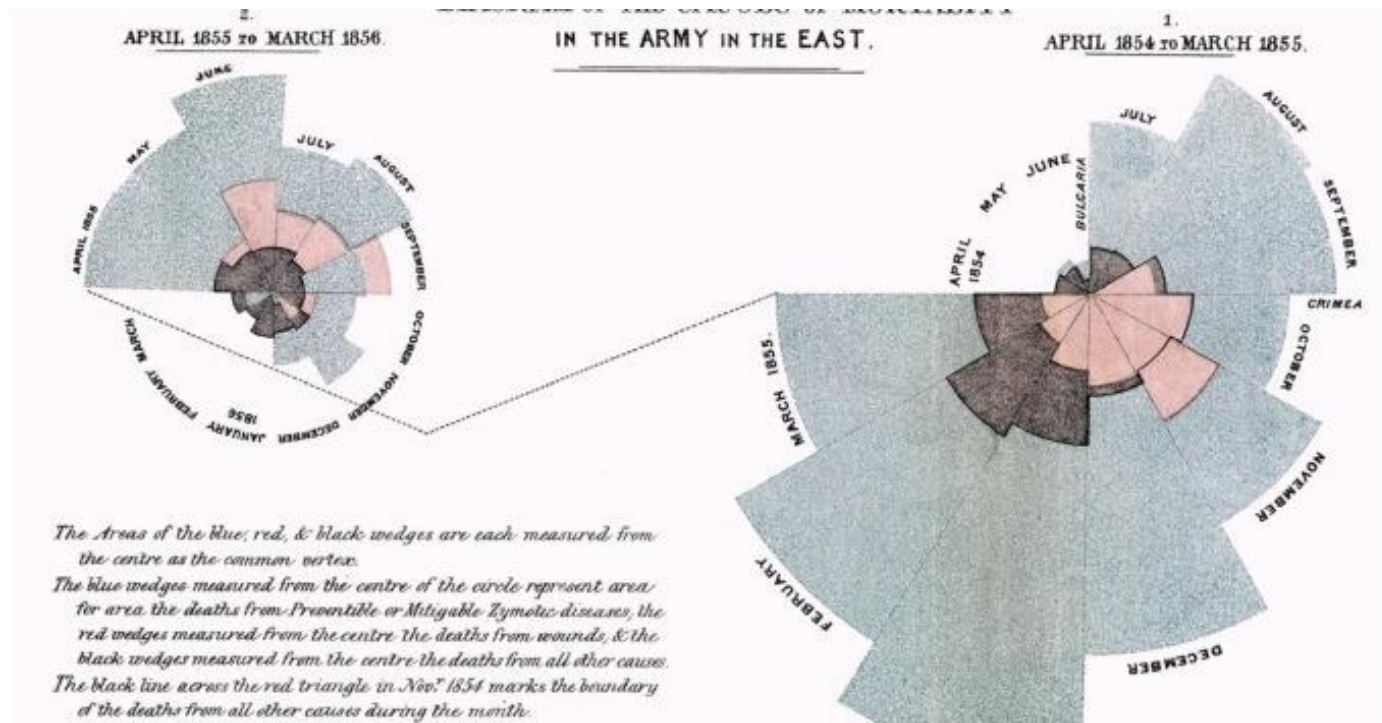
历史事件的流图可视化

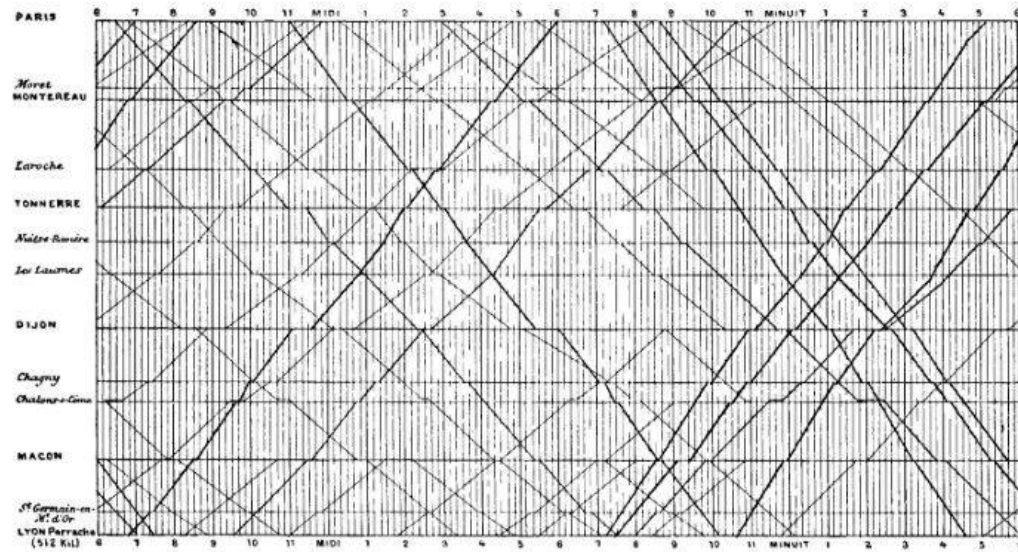






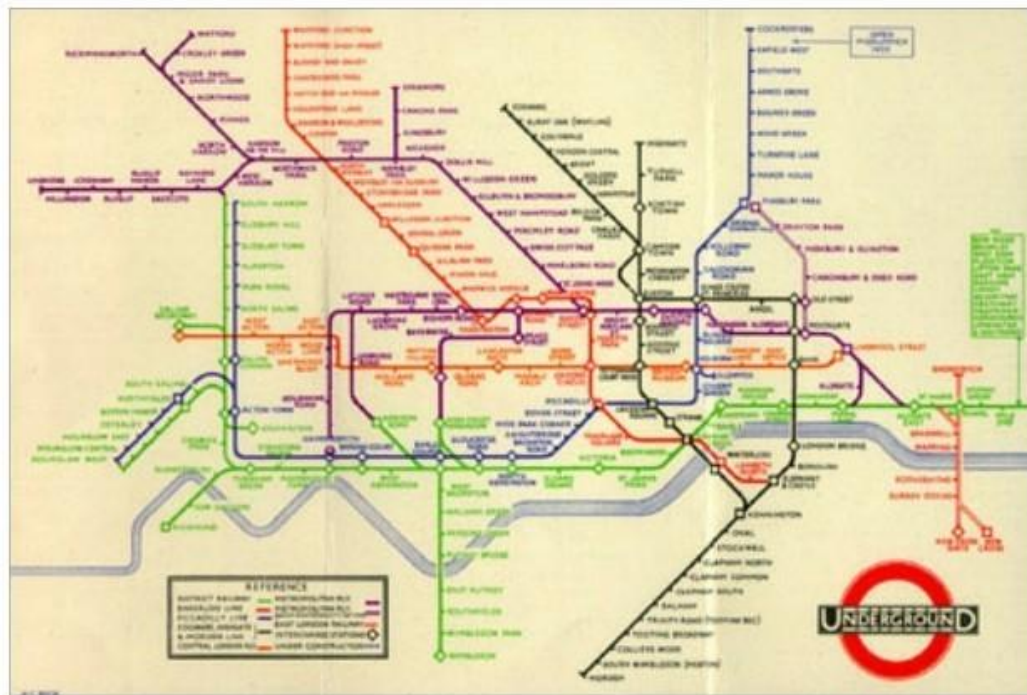
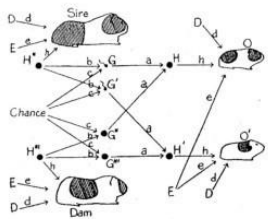
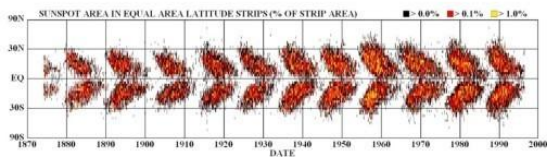
堆叠饼图（玫瑰图）





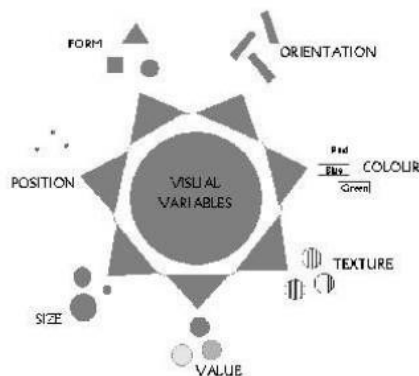
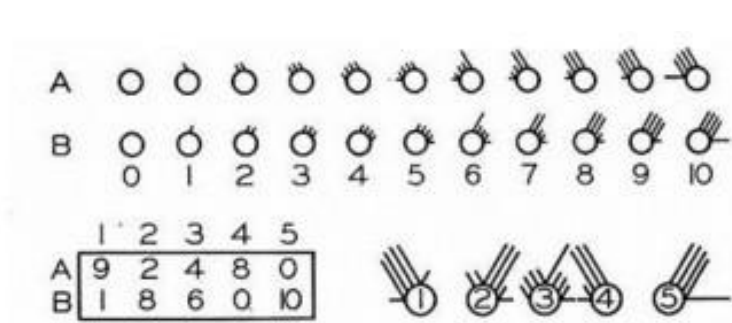
1900 -1949年：现代启蒙

- 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
- 多维数据可视化和心理学介入



1950 - 1974年：多维信息的可视编码

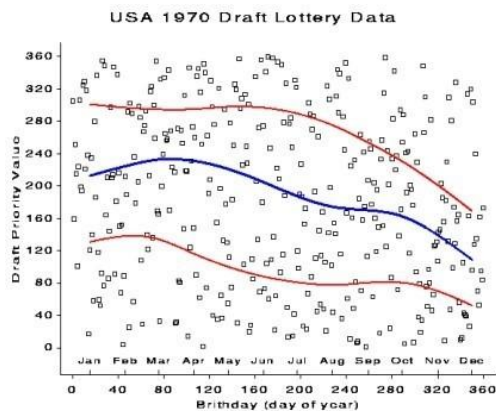
- 1967年，法国人Jacques Bertin出版了Semiology of Graphics（《图形符号学》）
- 奠定信息可视化的理论基石。
- 人们逐渐开始采用计算机编程生成可视化



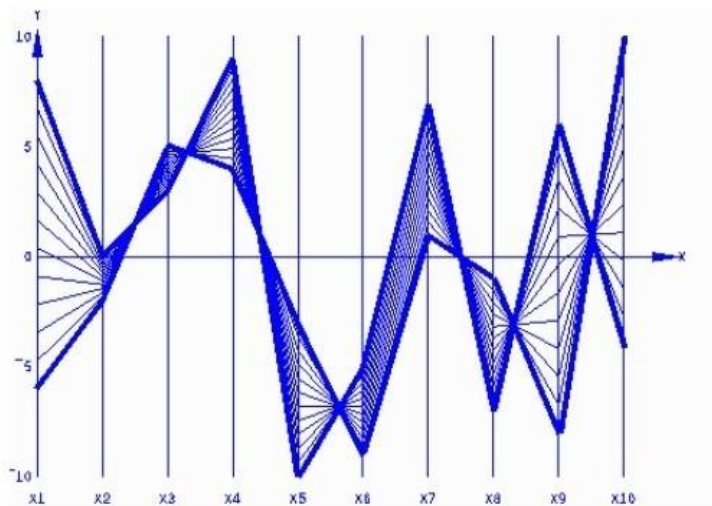
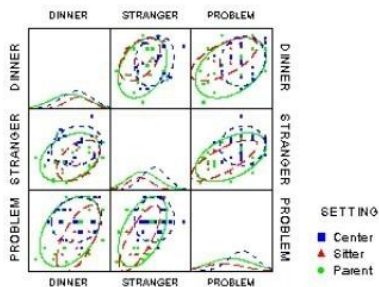
Marks	Points	Lines	Areas
LES VARIABLES DE L'IMAGE			
POINTS			
XY 2 DIMENSIONS DU PLAN	x	x	x
Z			
TAIILE			
VALEUR			
LES VARIABLES DE SÉPARATION DES IMAGES			
GRAIN			
COULEUR			
ORIENTATION			
FORME			

1975 - 1987年：多维统计图形

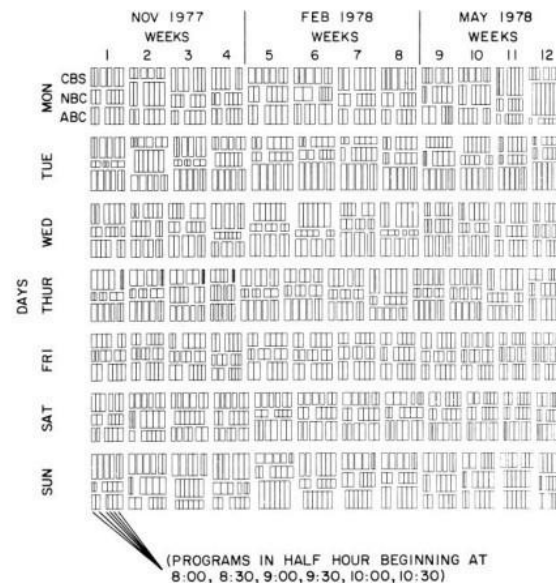
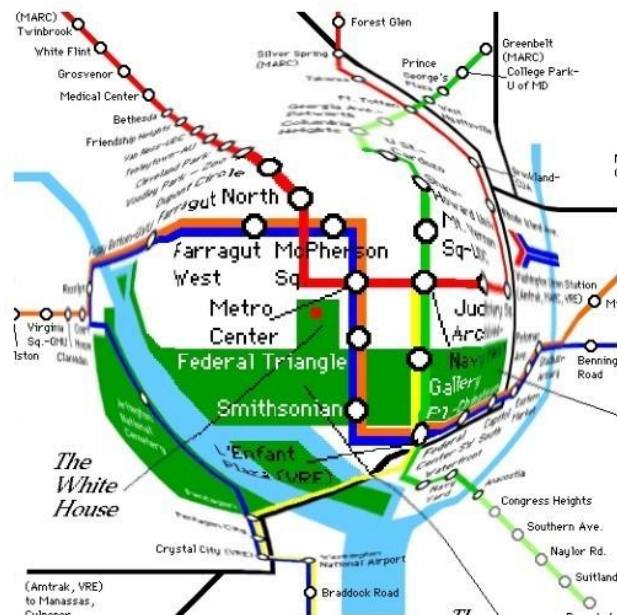
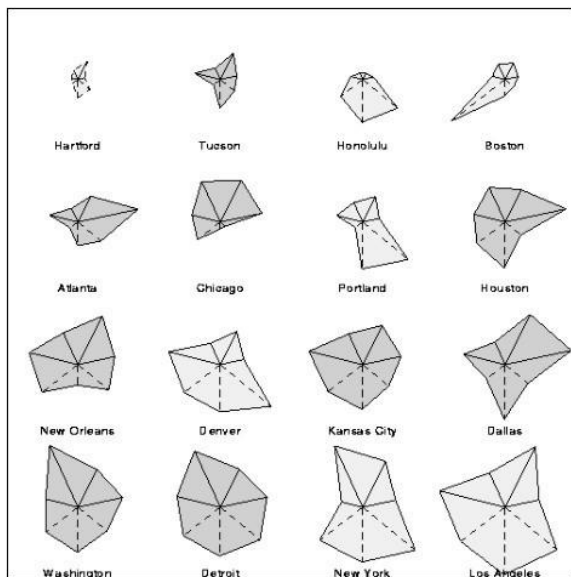
- 数据密集型计算开始走上历史舞台，也造就了对于数据分析和呈现的更高需求
- 1977年，美国著名统计学家John Tukey发表了“探索式数据分析”的基本框架。将可视化引入统计分析，促进对数据的深入理解
- 1982年，Edward Tufte出版了The Visual Display of Quantitative Information，构建了关于信息的二维图形显示的理论。Jacques Bertin的图形符号学一起，逐渐推动信息可视化发展成一门学科



Social Competence Measures Across Settings



多维统计图形

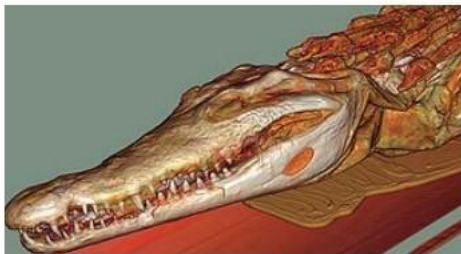


1987 - 2004年：交互可视化

- 1986年10月，美国国家科学基金会主办“图形学、图像处理及工作站专题讨论”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
 - 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，简称ViSC）
- 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
 - 会议报告正式命名并定义了科学可视化（Scientific Visualization）
- ACM SIGGRAPH 1987上，来William Lorensen和Harvey Cline发表了（Marching Cubes）一文，开创了科学可视化的热潮
 - 这篇论文是有史以来ACM SIGGRAPH会议被引用最高的论文
- 1989年，国际期刊Computer发表了一期关于科学计算中的可视化研究的专刊

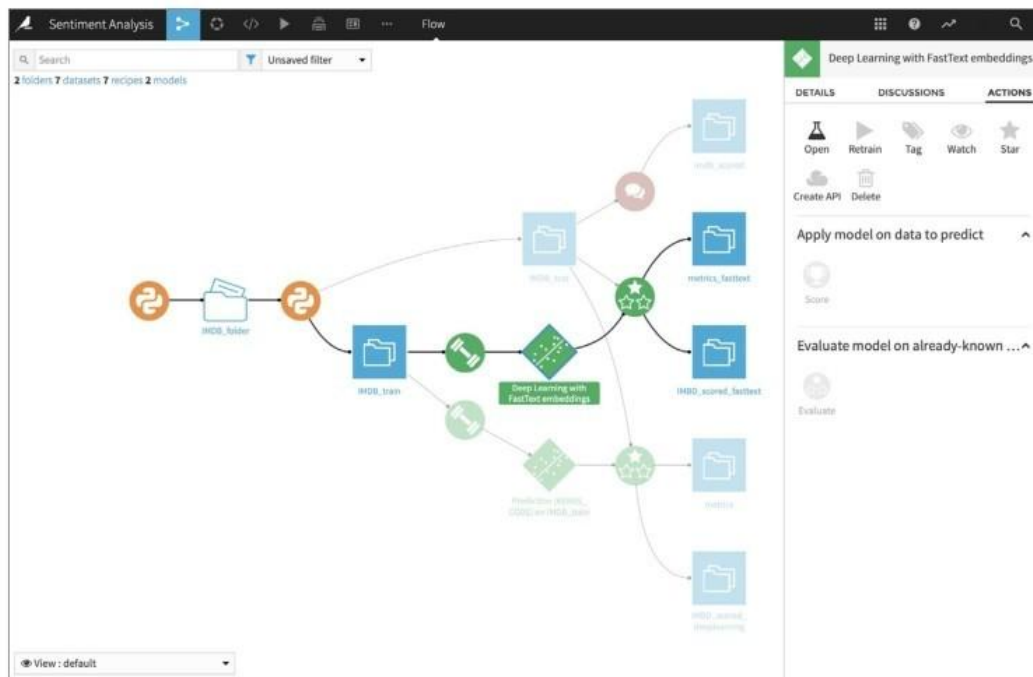
科学数据可视化

- 1989年，美国国家医学图书馆（NLM）实施可视化人体计划
- 1990年，IEEE举办了首届IEEE Visualization Conference
- 1995年始，IEEE Information Visualization会议，它以研讨会（Symposium）的形式隶属于IEEE Visualization会议，2007年，研讨会改名为IEEE Conference on Information Visualization
- 2012年，为突出科学可视化的内涵，会议更名为IEEE Conference on Scientific Visualization



2004年至今：可视分析学

- ▶ 针对海量、高维、多源和动态数据的分析挑战，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



可视分析

- 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来描述大规模数据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
- 2006年，IEEE开设了国际会议（IEEE Symposium on Visual Analytics Science and Technology）。2012年更名为IEEE Conference on Visual Analytics Science and Technology
 - 同时，欧盟欧洲可视化年会EuroVis从2010年起，专门举办可视分析研讨会EuroVAST
- 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
- 我国的数据采集、分析与应用应当通过自主研发，不能任由国外垄断公司来采集和处理，否则将危及国民生活与国防安全

3 数据可视化详解

数据科学的发展

➤ 数据爆炸

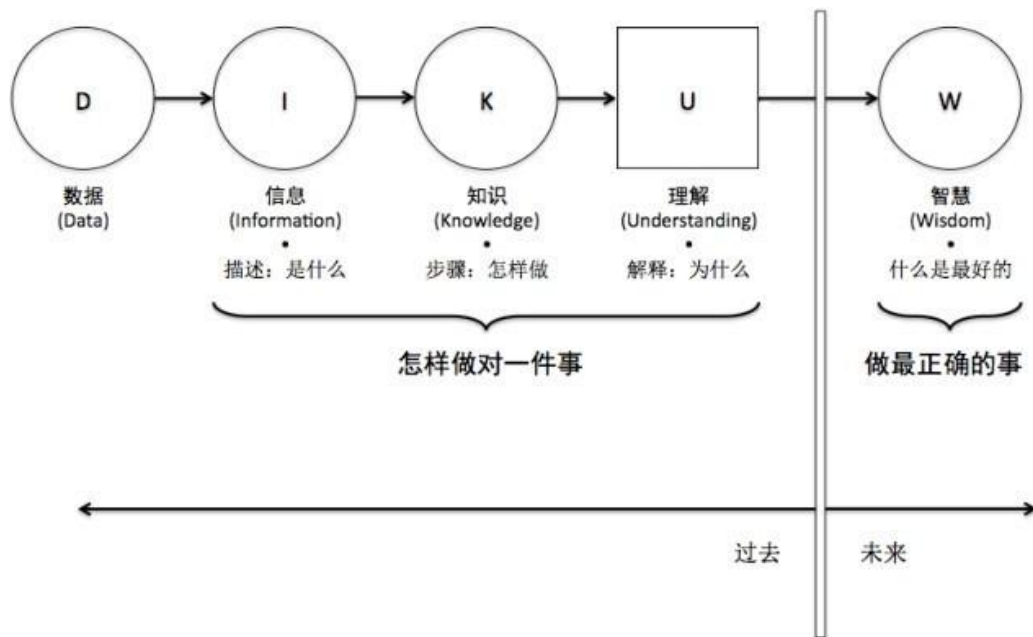
➤ 人类分析数据的能力已经远远落后于获取数据的能力。这个挑战不仅在于数据量越来越大、高维、多元源、多态，更重要的是数据获取的动态性、数据内容的噪声和互相矛盾、数据关系的异构与异质性等

➤ 2012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”

➤ 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

层次模型

- 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
- 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，该模型也作为一种数据处理流程，完成从原始数据的转化



数据

- 从信号获取的角度看，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。在表达为有用形式之前，数据本身没有用途
 - 数据即事实：数据是未经组织和处理的离散的、客观的观察
 - 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
 - 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

信息

- 信息是被赋予了意义和目标的数据

- 信息和数据的区别

- 信息是有用的、有意义的，可以回答诸如谁、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。

- 信息可以采用描述的方式定义知识

- 信息特征

- 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有关联，因此是有意义的、有价值的、有关联的。从这个意义上说，信息和数据的差别在于结构，而不是两者的功能

- 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

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
 - 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

智慧

- 智慧是启示性的，本意是知道为什么，知道如何去做
- 智慧与信息区别
 - 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，它是一种对什么的欣赏
 - 智慧是被评估过的理解
- 智慧可增加有效性和价值，它蕴含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

数据可视化意义

- 在DIKW模型所定义的数据转化为智慧的流程中，可视化借助于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
- 数据可视化的适用范围存在着不同的观点
 - 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
 - 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

数据可视化意义

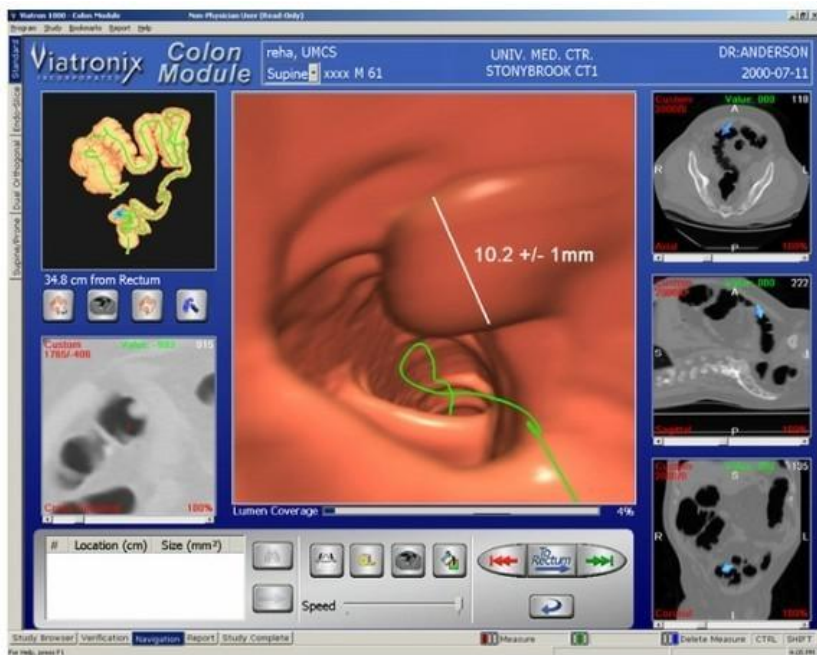
- 数据可视化的作用在于视物致知，即从看见物体到获取知识
 - 对于复杂、大尺度的数据，已有的统计分析或数据挖掘方法往往是对数据的简化和抽象，隐藏了数据集真实的结构
 - 数据可视化则可还原乃至增强数据中的全局结构和具体细节

- 数据可视化两个误区
 - 为了实现其获取知识的功能而令人感到枯燥乏味
 - 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含的信息、知识与思想，实现设计与功能之间的平衡

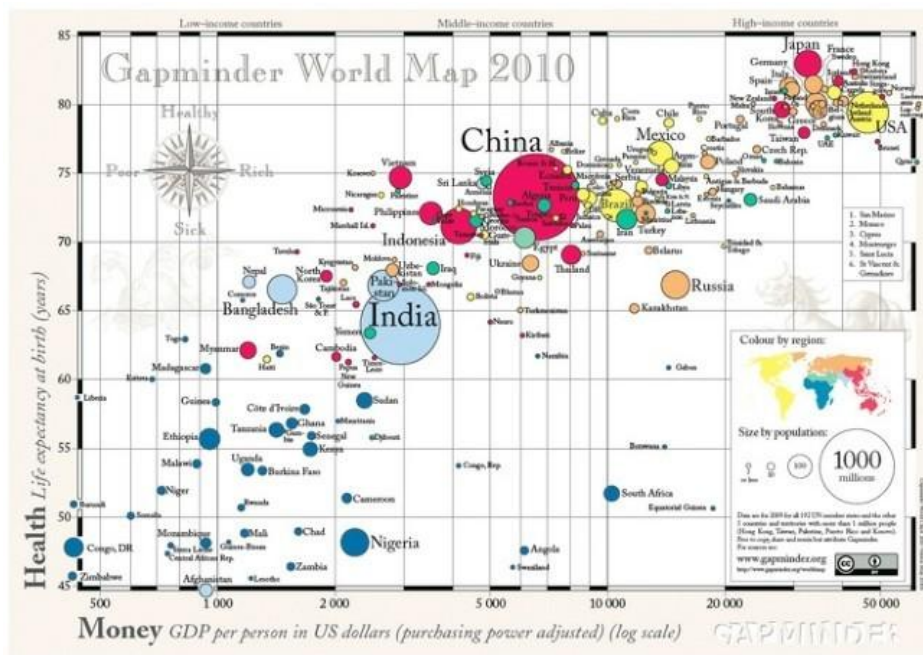
真

- 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



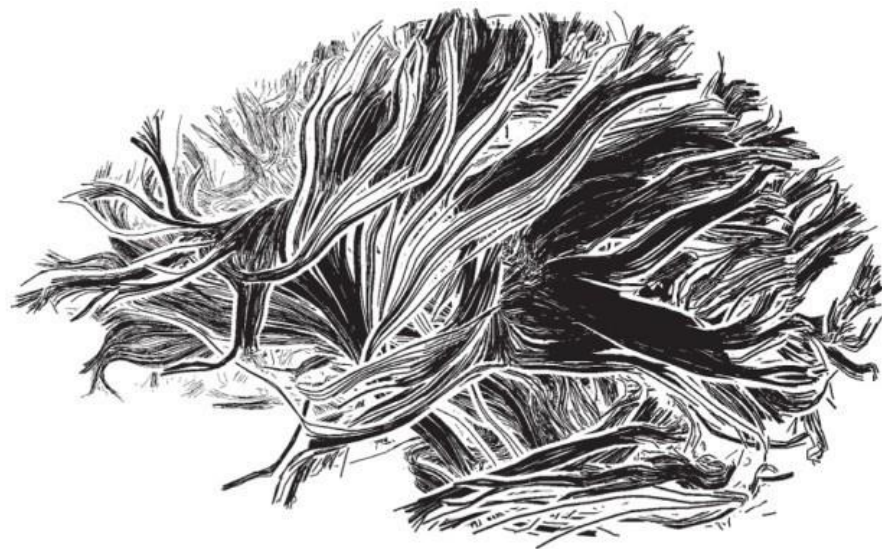
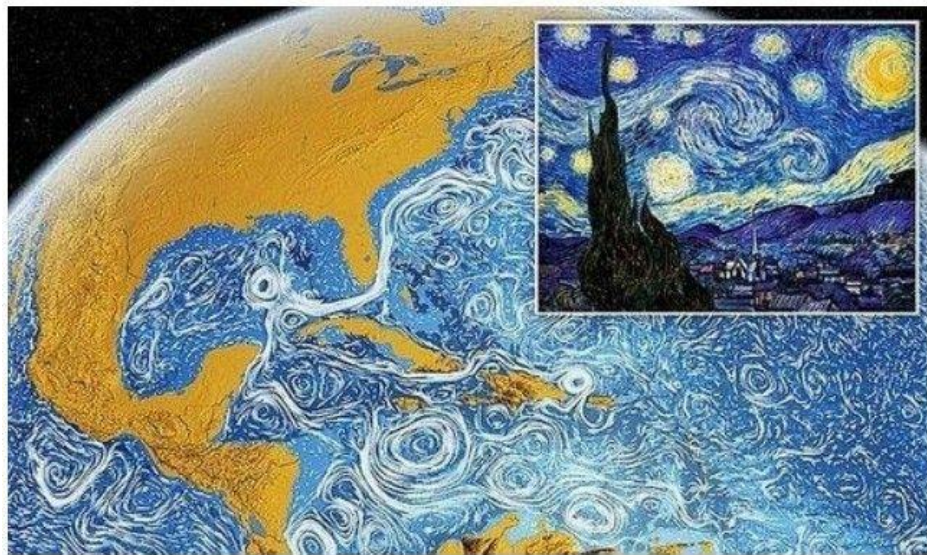
善

- 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
- 加拿大可视化专家Tamara Munzner认为，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

- 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



数据可视化分类

➤ 数据可视化

- 处理科学数据的科学可视化
- 处理抽象的、非结构化信息的信息可视化
- 将可视化与分析结合的新学科，可视分析学

数据可视化与其他学科领域的关系

➤ 图形学、人机交互

- 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，形成一门新的学科
- 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
- 数据、人、机器之间的交互是数据可视化的核心
- 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。

➤ 数据库与数据仓库

- 现代数据库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
- 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，用以支持决策制订过程

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
- 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
 - 数据可视化更擅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
 - 对数据模型没有一个预先的探索假设
 - 探寻数据中到底存在何种有意义的信息

面向领域的可视化方法与技术

- 生命科学可视化。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产生的数据的可视化方法。本质上属于科学可视化
- 表意性可视化。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
- 地理信息可视化。数据可视化与地理信息系统学科的交叉方向，它的研究主体是地理信息数据，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
- 产品可视化。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。
- 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。
- 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
- 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。
- 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

4 数据可视化研究挑战

数据可视化研究挑战

➤ 数据可视化的研究实质

- 理解可视化如何传递到观者，即人们感知和理解什么，可视化是如何对应于数据和数据模型的
- 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

- 计算能力的可扩展性
- 感知和认知能力的局限性
- 显示能力的局限性

➤ 未来的数据可视化的挑战主要在于两个方面

- 大数据可视化
- 以人为中心的探索式可视分析