



Data Analytics and Visualization

数据分析和可视化

中国地质大学（北京）信息工程学院

叶山

yes@cugb.edu.cn

自我介绍

课号：SR043268

数据分析和可视化

叶山（信工学院计算机系）

邮箱：yes@cugb.edu.cn

办公室：教三楼221

辅助工具：学习通



讲师背景

本科：宾夕法尼亚州立大学（地理信息）& 密歇根大学（地球和环境科学）

硕士：塔尔萨大学（地质学）

博士：威斯康辛大学麦迪逊分校（地质学/地学大数据）

科学的研究的第四范式

课程概览

经验科学 (古典时期/中世纪)

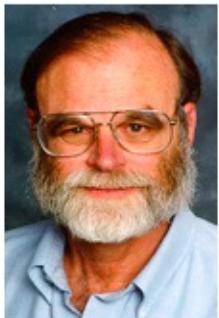
- 早期科学家通过观察和简单的实验来了解自然现象
- 阿基米德、伽利略、沈括、布冯

理论科学 (始于17世纪)

- 用定量或定性地总结归纳自然定律
- 开普勒、牛顿、开尔文、斯特诺

计算科学 (始于1950年代)

- 利用计算机的计算能力进行新的研究
- 仿生学、数值模拟、遥感图像、气候模型



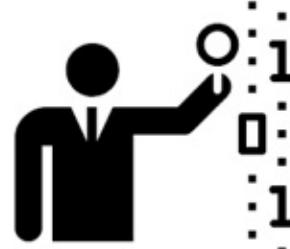
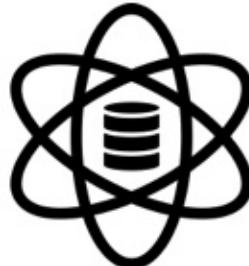
Jim Gray
图灵奖得主
微软工程师

数据科学 (始于21世纪)

- 利用人工智能、机器学习、云计算来处理大数据集
- 发现以前难以发觉的规律

课程概览

- 课程背景
 - 进入大数据时代，数据已经成为一种新的战略资产。
 - 越来越多的企业重视数据资产管理。
 - 越来越多的科研项目要依靠数据科学的技术和理论。
 - 数据的展示（可视化技术）处于重要地位。
- 课程主题
 - 数据分析的基本理论
 - 数据可视化的基本理论
 - 数据可视化的实战（R语言）
- 课程规划
 - 总学时32，其中讲课学时20，上机操作（Lab）学时12
 - 推荐教材：Winston Chang 《R 数据可视化手册》人民邮电出版社



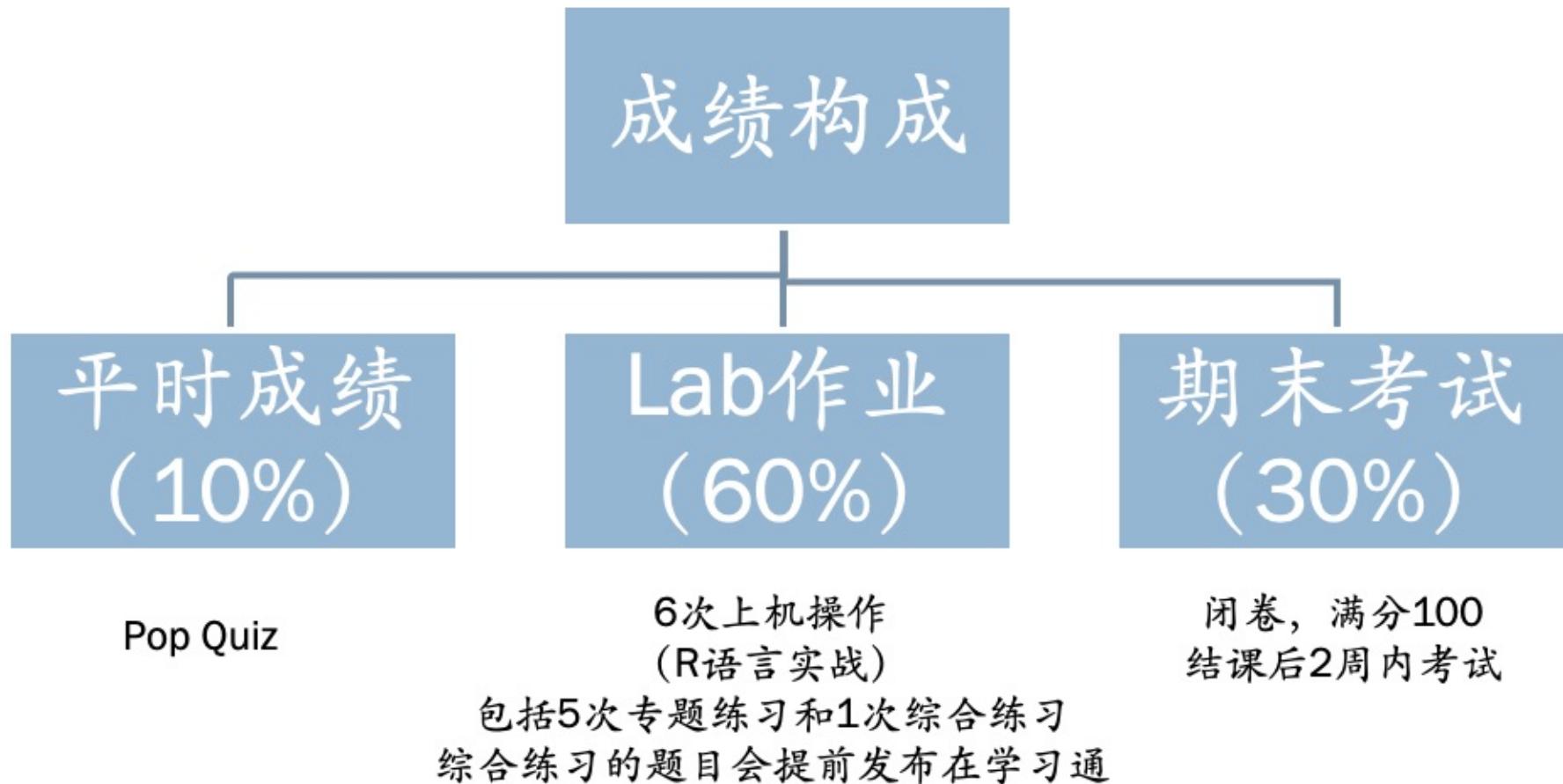
课程概览

这门课不是艺术课，而是数据科学课。
作业不以“美观”作为主要的评判标准。
在这门课上，可视化的主要目的是有效地传递数据和信息。
美观只是手段，而不是目的。

课程概览

课程目标	评价标准			
	90-100分	75-89分	60-74分	0-59分
	优	良	中/及格	不及格
1.能够综合运用所学科学原理，针对数据科学与大数据技术应用复杂工程问题，结合数据分析与可视化的理论与技术，设计合适的研究方案。	具备数据分析与可视化的理论与技术能力；能用于复杂工程问题的分析。	掌握数据分析与可视化的理论与技术能力；熟悉复杂工程问题分析的基本流程。	掌握数据分析与可视化理论与技术相关文献阅读、分析、总结能力。	不具备数据分析与可视化理论与技术能力；不能用于复杂工程问题的分析。
2. 具备自主学习和终身学习的意识。	具备自主学习和终身学习的能力。	基本养成了自主学习和终身学习的意识。	了解主动学习专业相关知识的习惯，并部分养成了学习意识。	不具备自主学习和终身学习的意识。

课程概览



Pop Quiz

课堂“突击”测验

- 每次5道选择题，每题2分，满分10分
- 内容为上一次课的知识点
- 每道题答对得2分，答错得1分，没有提交记0分
- 学期结束后，去掉一次最低分，剩下8次的平均分为平时成绩（满分80）



本节内容

数据、大数据、数据科学的概念

数据类型和数据集

数据清洗

数据预处理（数据精简）

数据 大数据 数据科学



什么是数据

- 数据是**符号的集合**，是表达客观事物的未经加工的原始素材。
- 数据的形式：图形、符号、数字、字母等。

数据模型

- 用来描述数据表达的**底层描述模型**
- 包含内容：
 - 数据的**定义和类型**
 - 不同类型数据的操作功能
 - 举例：浮点数据可以进行加减乘除等操作

概念模型

- 对目标事物的**状态和行为**进行**抽象的语义描述**
- 提供构建抽象事物、推理支持等操作
 - 举例：一维浮点数可以描述温度，三维浮点向量可以描述空间中的风向

对象和属性

数据的构成

- 数据也可看成是数据对象和其属性的集合，其中属性可被看成是变量、值域、特征或特性。
 - 例如：人类头发的颜色、人类体温等。

数据对象 Object/Datum

- 单个数据对象可以由一组属性描述
- 也被称为记录、点、实例、采样、实体等。

数据属性 Attribute

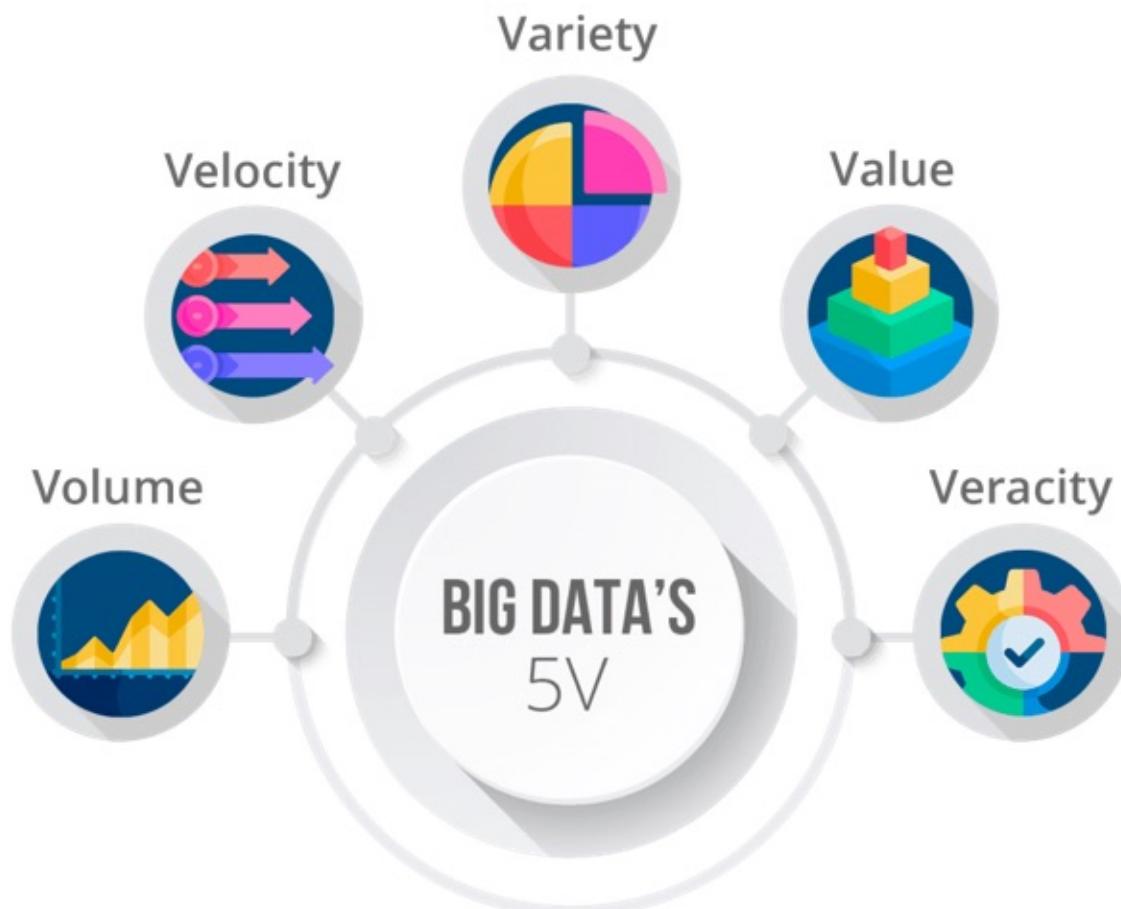
- 属性值可以是表达属性的任意数值或符号，同一类属性可以具有不同的属性值。
 - 例如：长度的度量单位可以是米或者英尺。
- 不同的属性也可能具有相同的取值、不同的含义。
 - 例如：年份和年龄都是整数型数值，而年龄通常有取值区间。

大数据的诞生

- 国际数据公司（IDC）的统计表明：
 - 大约从2006年开始，每年全世界数据总量在呈指数增长
 - 2011年全年世界创建和复制的数据总量已达到1.8ZB（1ZB=10⁹TB）
 - 预计2025年的数据量会达到80ZB。
- 更多的数据意味更多的价值，但是当前的信息处理和分析能力落后于数据生产的速度。
- 麦肯锡公司正式提出大数据概念（2011）
- 大数据的现状
 - 大数据专业在众多高校被开设
 - 大数据概念在民众间获得了极大的关注度
 - 但是，公众对于大数据的概念存误解
 - 大数据这个词有被滥用的趋势

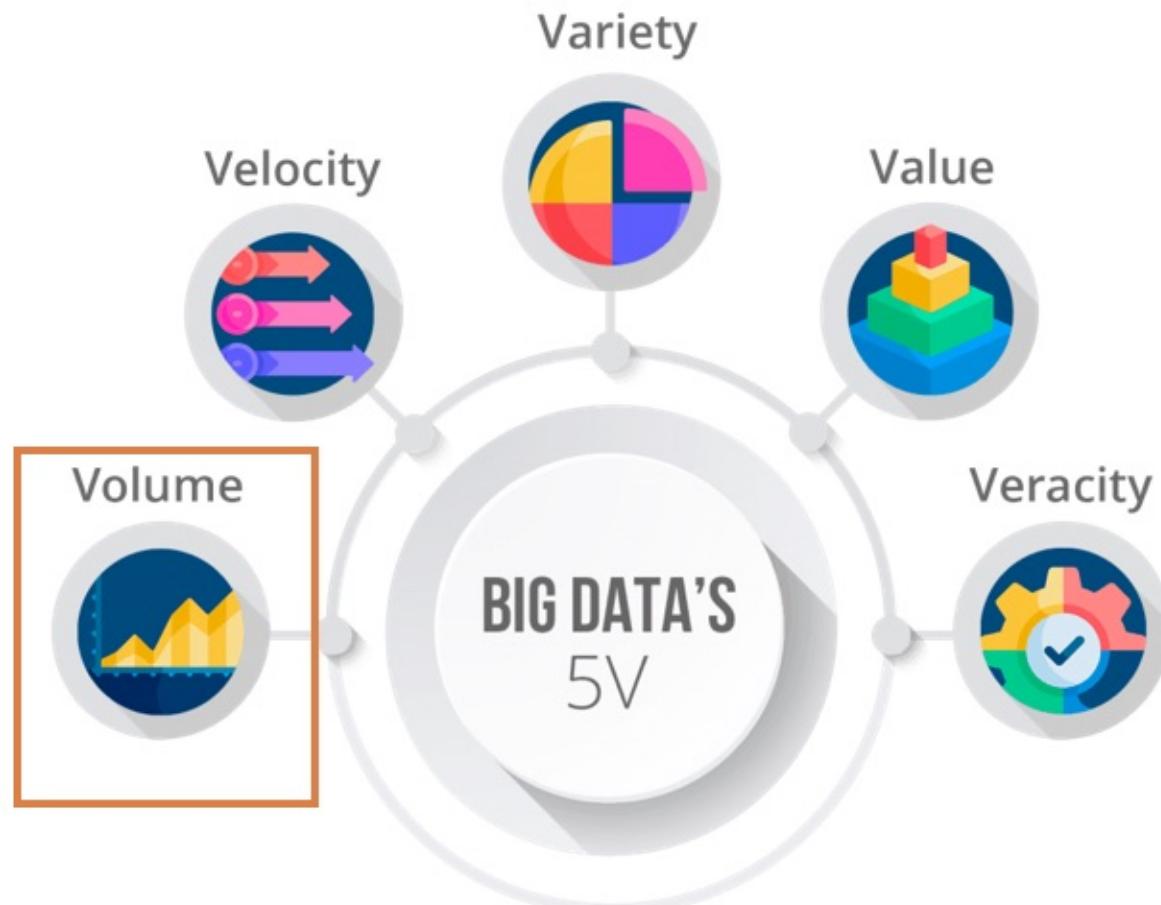
大数据的定义

- 无法利用现有的工具在可接受的时间范围内进行采集、管理和处理的数据，其特征可以用5V来概括



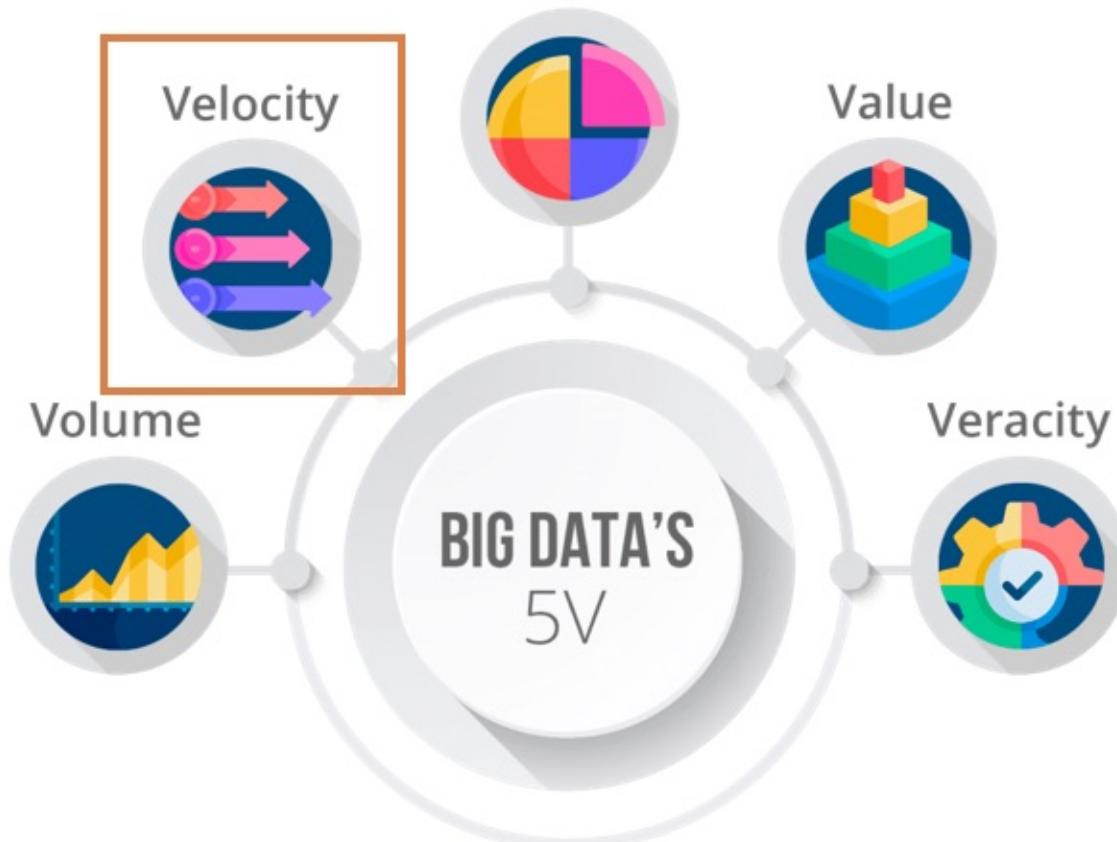
Volume

- 大量的数据规模，来源于数据采集方式和存储设备发展更新：
网页数据、电子商务数据、金融交易数据、科研数据、交通
数据、医疗卫生数据



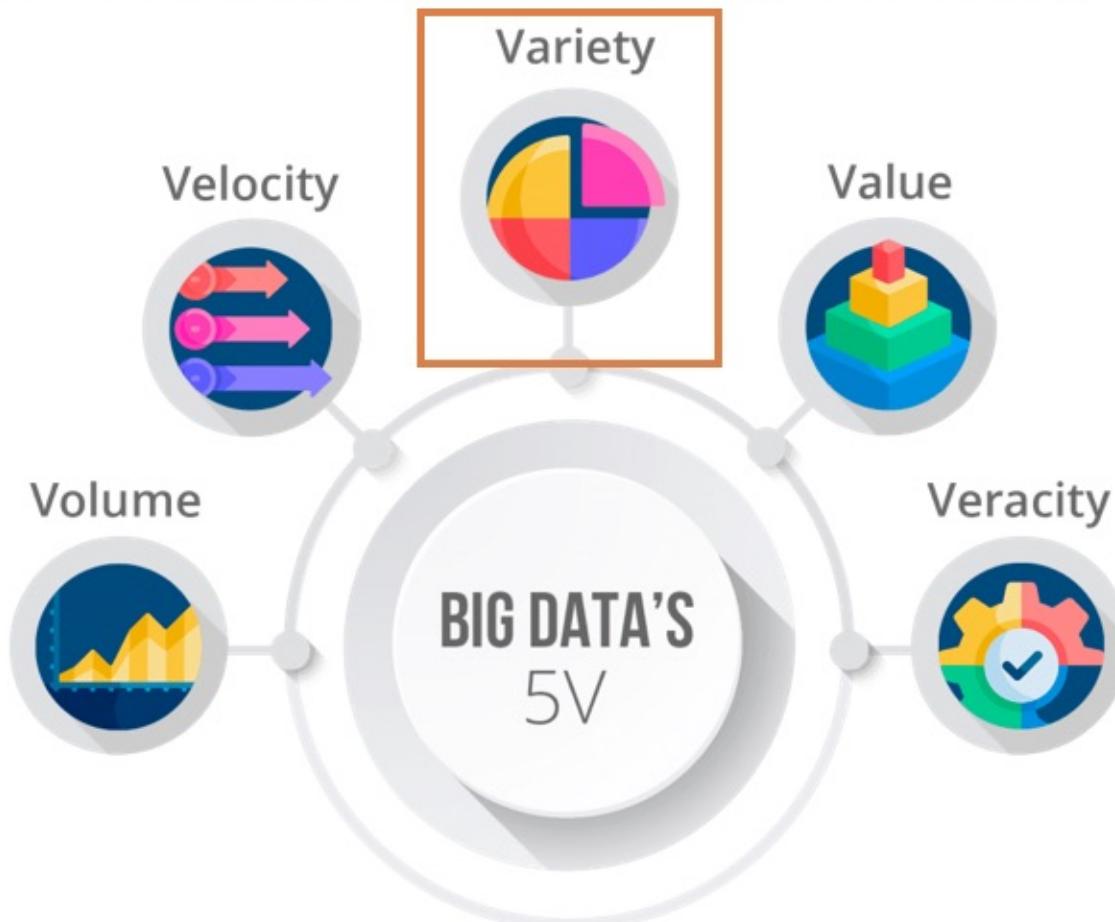
Velocity

- 数据量的增大、数据产生速度的加快，意味着在大数据时代我们需要面对快速的数据流转。社交软件、传感器、监控摄像头等数据采集设备带来巨大的数据流；互联网的内容流量构成了高速变更的数据体系。如何存储、管理、分析这些数据成了一个棘手的问题。 Variety



Variety

- 当前的大数据通常是非结构化数据，这意味着这些数据的存储格式和组织形式以及数据间的关系没有统一的数据模型来描述。如何有效地应对以结构化、非结构化数据组成的异构数据体系，是大数据时代处理复杂数据的重要议题之一。



Variety

结构化数据

- 能用二维表结构来表达的数据
- 举例： csv、 tsv等

半结构化数据

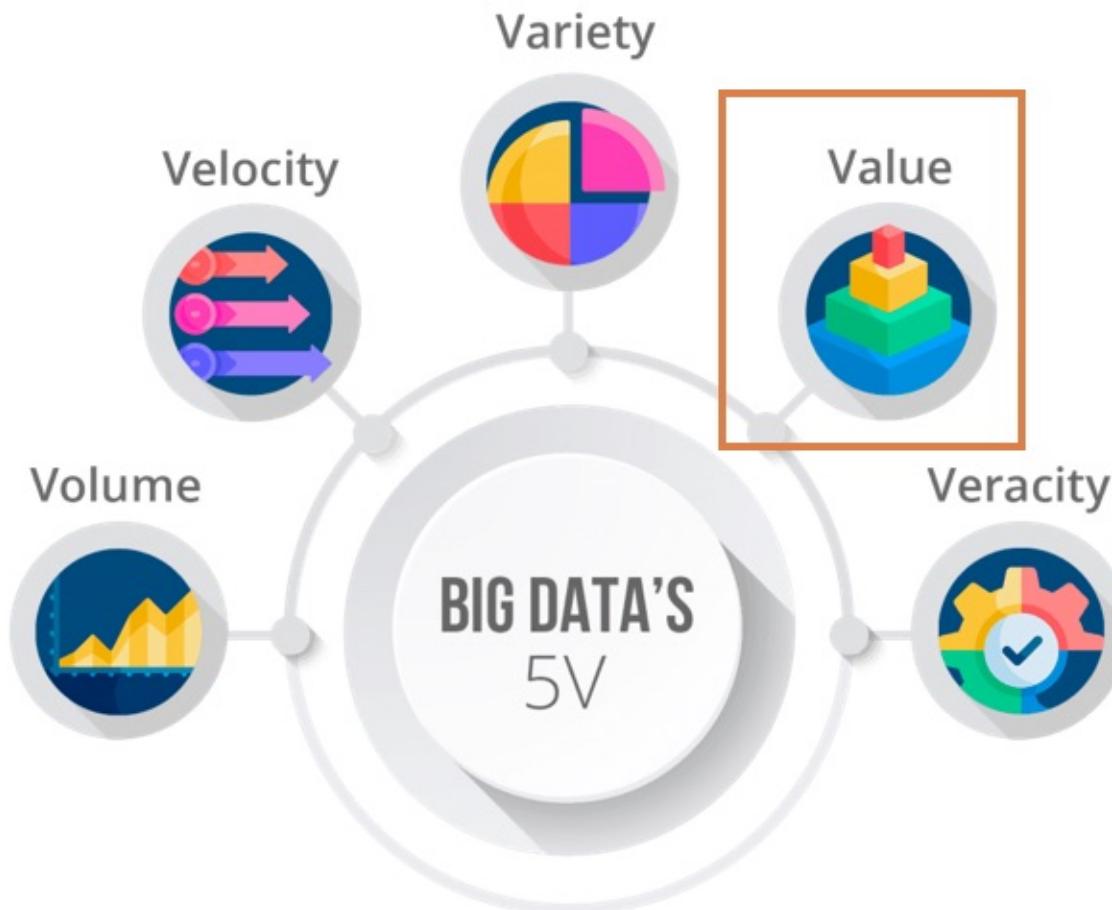
- 维度高于二维但机器可读的数据
- 举例： json、 netcdf、 html、 shapefile等

非结构化数据

- 没有固定组织原则、未经过滤的信息
- 举例： 网页日志、自然语言、地图、声音信号等

Value

- 数据获取设备和计算资源越来越廉价，使我们能以计算的手段从数据中挖掘出更多的应用价值。



- 商业价值
 - 网站可以利用用户行为数据开展个性化服务
 - 公司可以基于商业数据开发新产品和值服务等。
- 科研价值
 - 利用大数据、高性能计算、人工智能等计量化的方法，开展交叉学科的研究

Bioinformatics: 基因测序、蛋白质结构、生物大分子、生物动力学

Geoinformatics: 宏进化、气候模型、遥感、地震反演、古环境重建

Urban Informatics: 城市大数据、智慧城市、城市感知、数字孪生

Medical Informatics: 流行病、药物研发、远程ICU、免疫、肿瘤

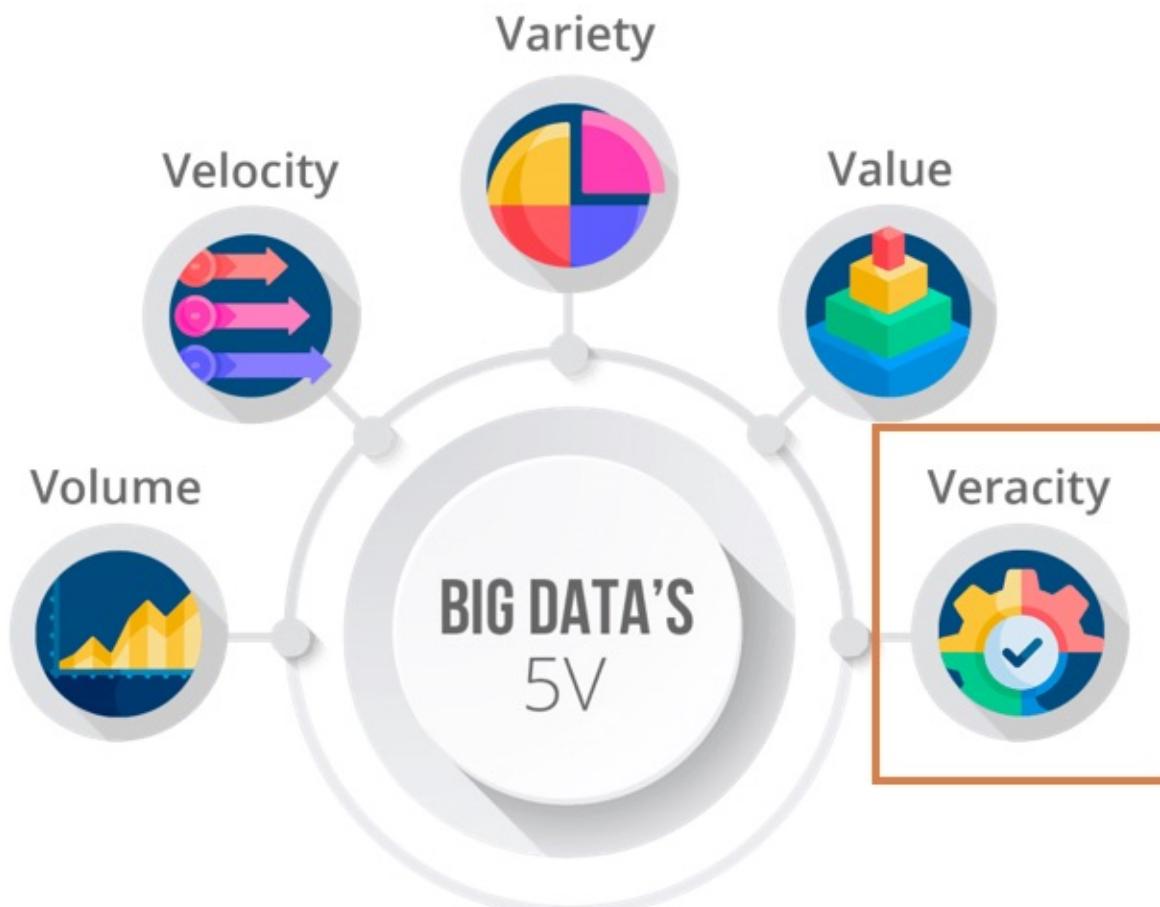
Materials Informatics: 材料识别、热力学相图识别、微观结构

Social Informatics: 人口流动、宏观经济、国际关系、大众媒体

Agricultural Informatics: 农作物监测、智慧农场、产量预测

Veracity

- 大数据的内容是与真实世界息息相关的，真实不一定代表准确，但一定不是虚假数据，这也是数据分析的基础。



Veracity

- 基于真实的科学研究、商业交易与用户行为产生的数据，才有意义。
- 为什么要反对数据造假?
 - 数据造假破坏了信息数据的真实性，导致数据指标失真。
 - 数据不可信了，大数据行业也就没了“基石”。
- 不真实数据的危害
 - 不真实的网络数据：导致价值评判标准失衡，阻碍优秀内容和产品的诞生，影响互联网行业的发展。
 - 不真实的商业数据：误导投资人和消费者，长久以往会扰乱市场秩序，对消费者乃至社会经济带来损失。
 - 不真实的科研数据：误导后续的研究，浪费科研经费、人力和时间，阻碍科学进步。

数据科学

数据科学的发展背景

- 我们来到了数据即服务（DaaS）的时代，用户可以随时随地按需求获取数据和信息。
- 传统方法难以应对海量原始数据的直接处理和分析，在很多情况下数据被没于浩瀚的“数据海洋”中，造成数据浪费。
- 急需一种能够针对大数据进行统计、分析和信息提取的方法。

数据科学的兴起

- 近年来，电子科学、信息科学、语义网络、数据组织与管理、数据分析、数据挖掘和数据可视化等领域的办法，可以有效地提取隐藏在数据中有价值的信息。
- 形成了提炼科学原理、验证科学假设、服务科学探索的新思路。
- 研究这种综合性方法的学科被称为“数据科学”，是新兴的交叉学科。

数据科学

数据科学涵盖的方面

- 数据管理、计算机科学、统计学、视觉设计、可视化、人机交互；
- 基于架构式技术和信息技术的自然科学（比如bioinformatics, geoinformatics等）。

数据科学的意义

- 改变了所有学科个人和协作工作的模式。
- 让商业和科学数据分析处理进入了“数据驱动”阶段（第四范式）。
- 能够帮助相关人员解决大尺度、高复杂度的全局问题。这些问题的尺度和复杂度超越了已有工具的承受范围。

数据科学的基础知识

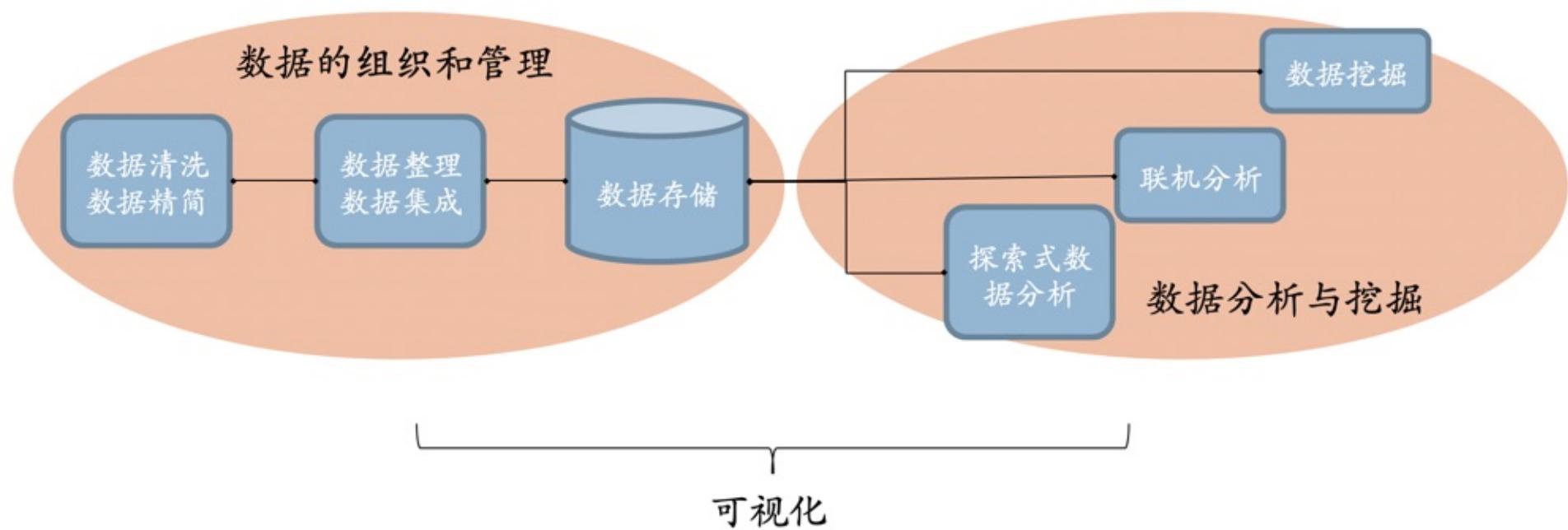
- 数据集成、数据模型和架构、元数据追踪、数据生命周期
- 统计学、人工智能的部分知识
- 数据可视化及可视化分析

数据科学与可视化

- 可视化是数据科学的核心要素之一
 - 数据信息的展示方法
 - 人机交互的接口
- 面对海量数据，直接观察或简单的统计分析难以发现数据中隐藏的信息。
 - 无法通过查看海量的服务器日志，来判断系统是否遭到攻击威，
 - 无法简单统计社交网站上的好友关系，来发掘用户的喜好。
- 可视化的作用
 - 通过可视化方法，把海量数据变成图形
 - 有助于我们对数据中的属性、关系进行深入探究
 - 直观展示数据中隐藏的规律，为科技研发、政策制定、商业决策、社会服务提供科学依据

数据的生命周期

- 可视化能在数据的生命周期中的所有阶段发挥作用
- 它作为一种人机交互手段，贯穿于整个数据处理和分析的流程里。



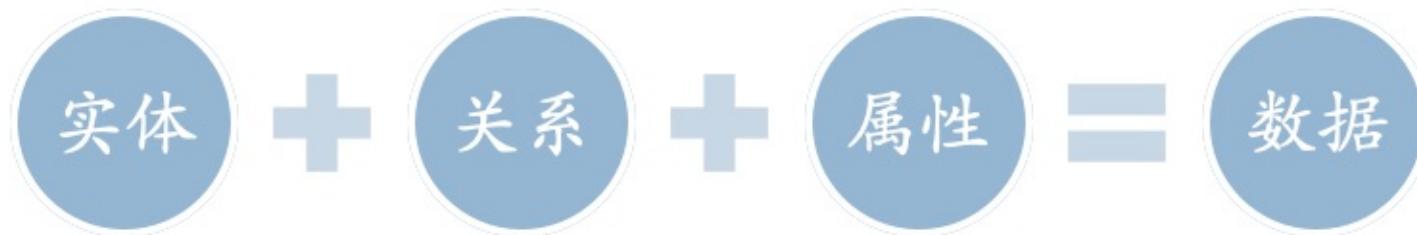
数据类型和数据集



数据分类

- 数据的分类和信息与知识的分类相关

水果包含苹果、梨子



水果、苹果、梨子

颜色、大小、包含关系



数据属性

离散属性

- 取值来自有限或可数的集合
- 例如：性别、颜色、学位、职称、籍贯

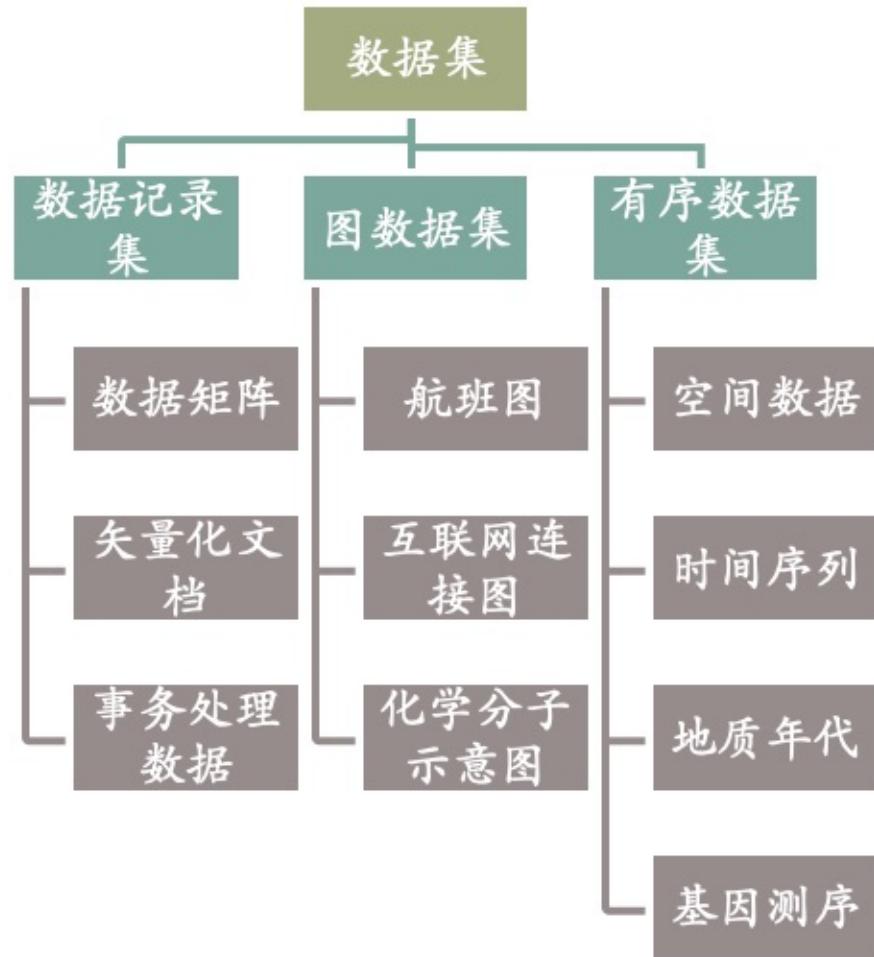
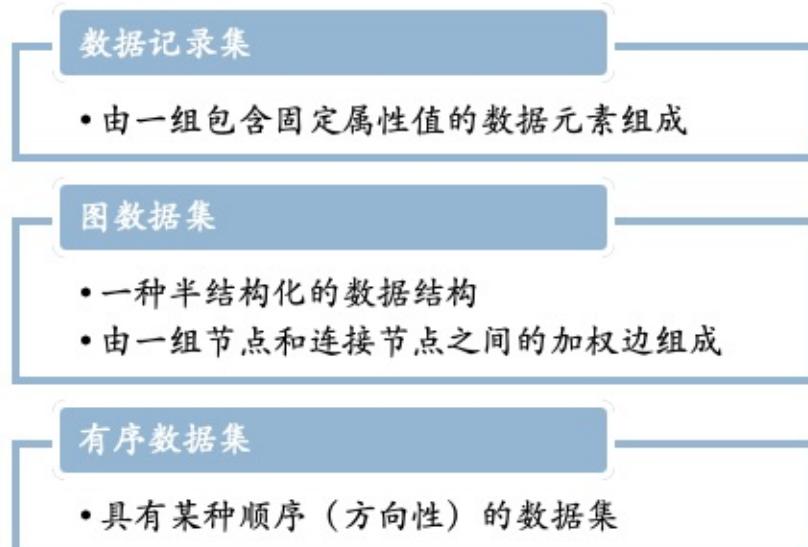
连续属性

- 取值范围对应实数域
- 例如：温度、高度、湿度

思考问题：类型为数值的属性都是连续属性吗？
错误。比如：邮政编码、年级、雅思分数

数据集

- 数据集是数据的实例，有3种常见表达形式。



数据记录集

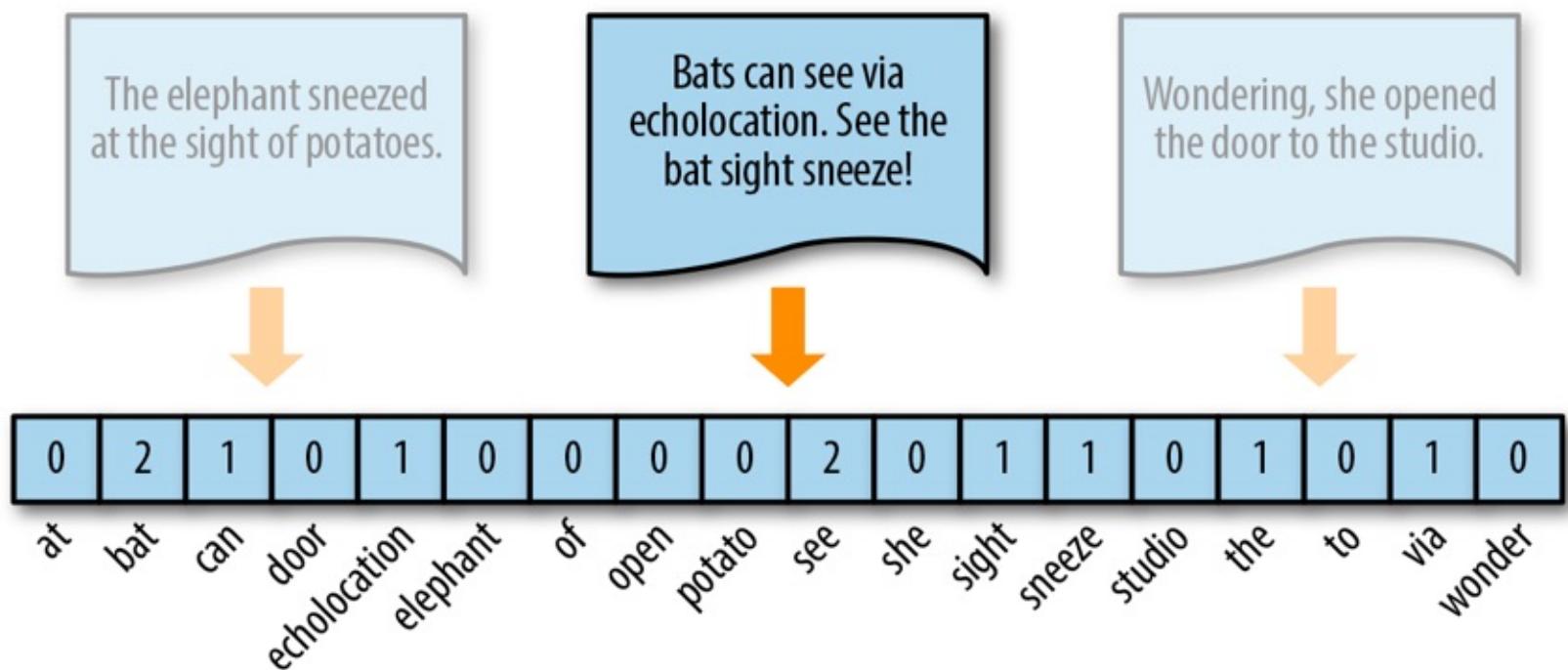
- 数据矩阵

- 如果数据对象具有一组固定的数值属性，则数据对象可视为高维空间的点集，每个维度对应单个属性。这种数据集可以直接表达为一个 $m \times n$ 的矩阵。其中，矩阵的每行代表一个对象，每列代表单个属性在数据集中的分布。这种表示方法称为数据矩阵，通常被组织为二维表格形式。

名称	人口	陆地面积	水域面积	众议院人数
亚拉巴马州	5,024,279	131,171	4,597	7
阿拉斯加州	733,391	1,477,953	245,384	1
亚利桑那州	7,151,502	294,207	1,026	9
阿肯色州	3,011,524	134,771	2,961	4
加利福尼亚州	39,538,223	403,466	20,501	52
科罗拉多州	5,773,714	268,431	1,170	8
康涅狄格州	3,605,944	12,542	1,816	5
特拉华州	989,948	5,047	1,399	1
佛罗里达州	21,538,187	138,887	31,424	28
佐治亚州	10,711,908	148,959	4,951	14
夏威夷州	1,455,271	16,635	11,678	2
爱达荷州	1,839,106	214,045	2,398	2
伊利诺伊州	12,812,508	143,793	6,202	17
印第安纳州	6,785,528	92,789	1,537	9

数据记录集

- 矢量化文档
 - 如果统计文档中所有单词出现的频率，则一个文档可以被表示为一个矢量，其长度是单词集的个数
 - 向量中的每个分量记录单词集中每个单词在该文档中的出现频率（词频统计）



数据记录集

- 事物处理数据是一类特殊的数据记录，每条记录都包含一组数据项

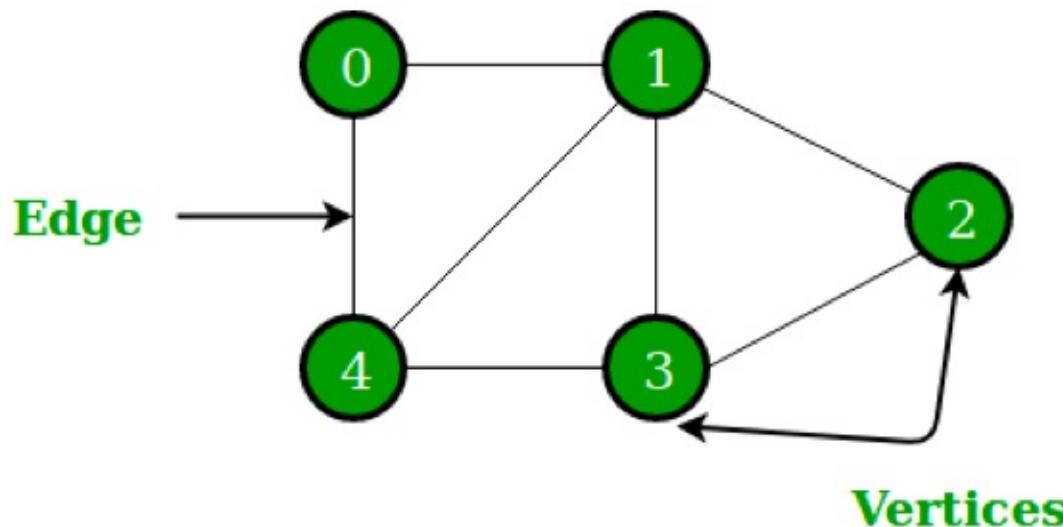


```
root@precise64:/var/log# ls
alternatives.log  boot      dist-upgrade  dmesg.1.gz  faillog   kern.log  mailing.log  syslog    ufw.log          vboxadd-install-x11.log
apt              boot.log   dmesg        dmesg.2.gz  fsck      lastlog   news       syststat  upstart         VBoxGuestAdditions.log
auth.log         btmp     dmesg.0.gz  dpkg.log   installer  mail.err  ntpstats  udev      vboxadd-install.log  vtmp

root@precise64:/var/log# grep error dmesg* | log | more
dmesg:[ 1.696121] EXT4-fs (dm-0): re-mounted. Opts: errors=remount-ro
dmesg:[ 2.243534] EXT4-fs (dm-0): re-mounted. Opts: errors=remount-ro
dmesg:[ 2.876374] EXT4-fs (dm-0): re-mounted. Opts: errors=remount-ro
dmesg:log:2012-09-14 05:46:54 install libgcc-error0 unmet: 1.10-Zubuntu1
dmesg:log:2012-09-14 05:46:54 status half-installed libgcc-error0 1.10-Zubuntu1
dmesg:log:2012-09-14 05:46:54 status unmet libgcc-error0 1.10-Zubuntu1
dmesg:log:2012-09-14 05:46:54 status unmet libgcc-error0 1.10-Zubuntu1
dmesg:log:2012-09-14 05:47:06 configure libgcc-error0 1.10-Zubuntu1 <none>
dmesg:log:2012-09-14 05:47:06 status unmet libgcc-error0 1.10-Zubuntu1
dmesg:log:2012-09-14 05:47:06 status half-configured libgcc-error0 1.10-Zubuntu1
dmesg:log:2012-09-14 05:47:06 status installed libgcc-error0 1.10-Zubuntu1
dmesg:log:2012-09-14 06:07:03 install liberror-perl unmet: 0.17-1
dmesg:log:2012-09-14 06:07:03 status half-installed liberror-perl 0.17-1
dmesg:log:2012-09-14 06:07:03 status installed liberror-perl 0.17-1
dmesg:log:2012-09-14 06:07:03 configure liberror-perl 0.17-1 <none>
dmesg:log:2012-09-14 06:07:03 status unmet liberror-perl 0.17-1
dmesg:log:2012-09-14 06:07:15 configure liberror-perl 0.17-1 <none>
dmesg:log:2012-09-14 06:07:15 status unmet liberror-perl 0.17-1
dmesg:log:2012-09-14 06:07:15 status half-configured liberror-perl 0.17-1
dmesg:log:2012-09-14 06:07:15 status installed liberror-perl 0.17-1
dmesg:log:2012-09-14 06:12:43 status installed liberror-perl 0.17-1
dmesg:log:2012-09-14 06:12:43 remove liberror-perl 0.17-1 <none>
dmesg:log:2012-09-14 06:12:43 status half-configured liberror-perl 0.17-1
dmesg:log:2012-09-14 06:12:43 status half-installed liberror-perl 0.17-1
dmesg:log:2012-09-14 06:12:43 status half-installed liberror-perl 0.17-1
```

图数据集

- 图是一种半结构化数据结构，由一组节点(vertices)和连接节点的边(edge)组成。

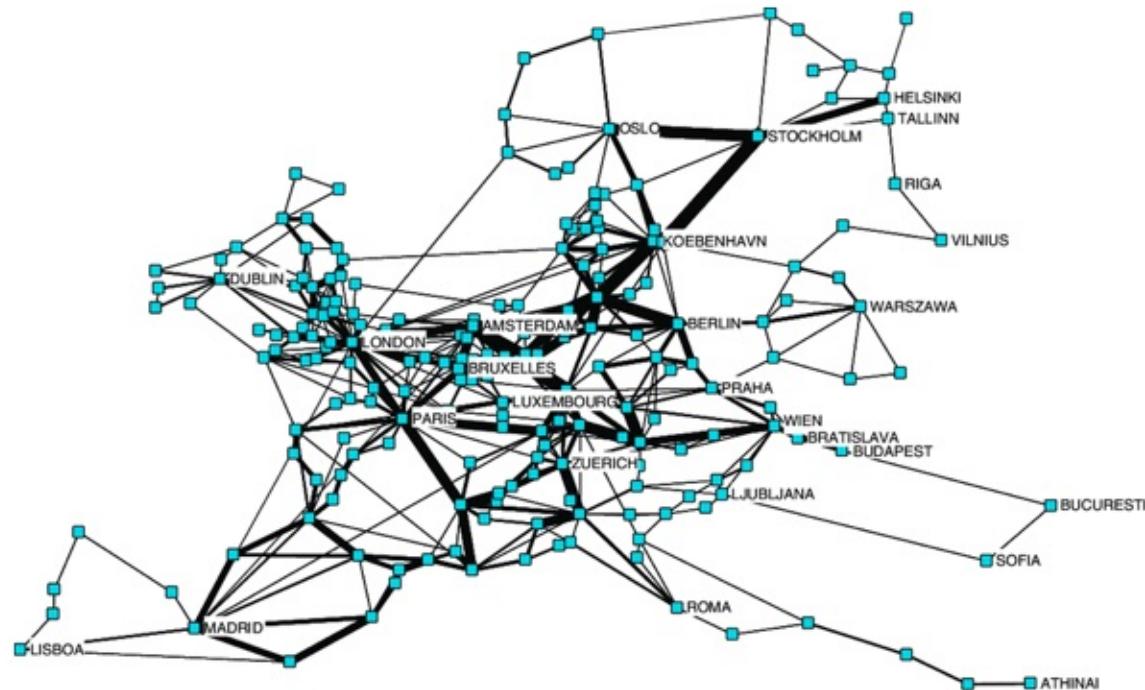


图数据集



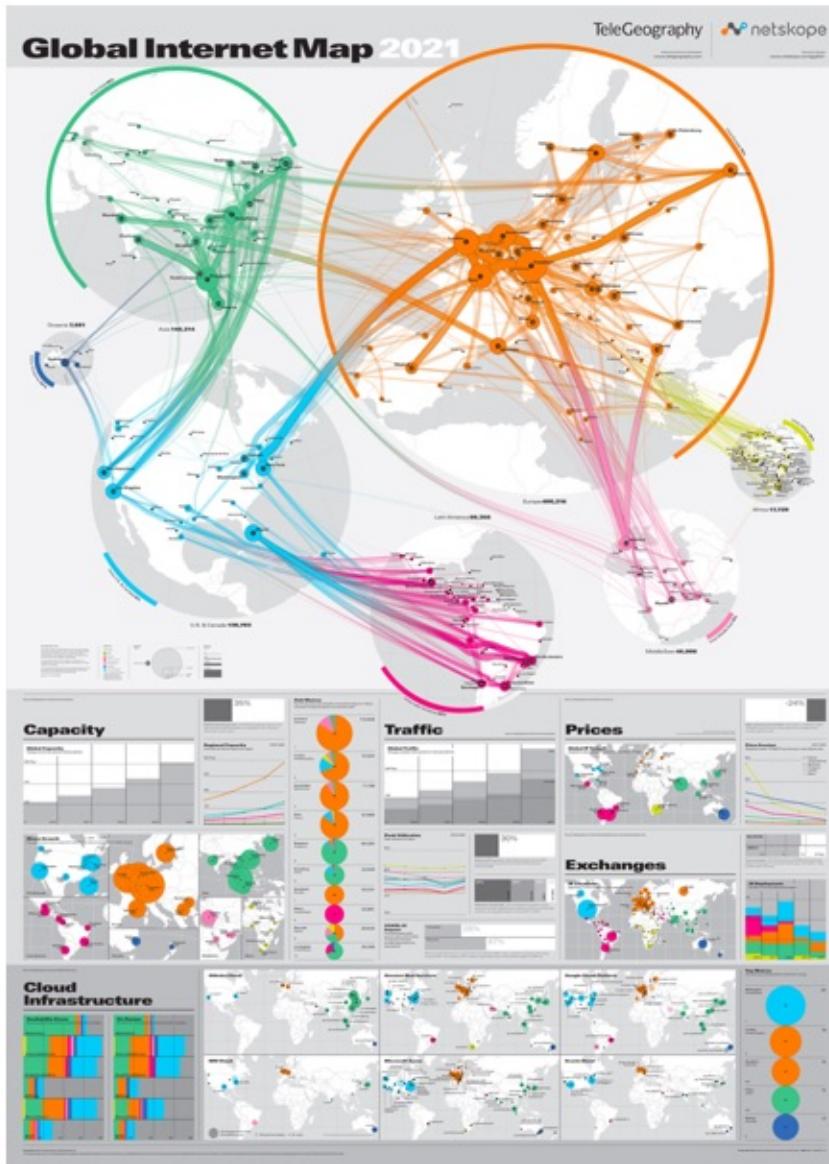
全球航班线路图

图数据集



2001 年欧洲互联网基础设施图。线条（边）的粗细代表两个节点（城市）之间的互联网服务供应商数量。

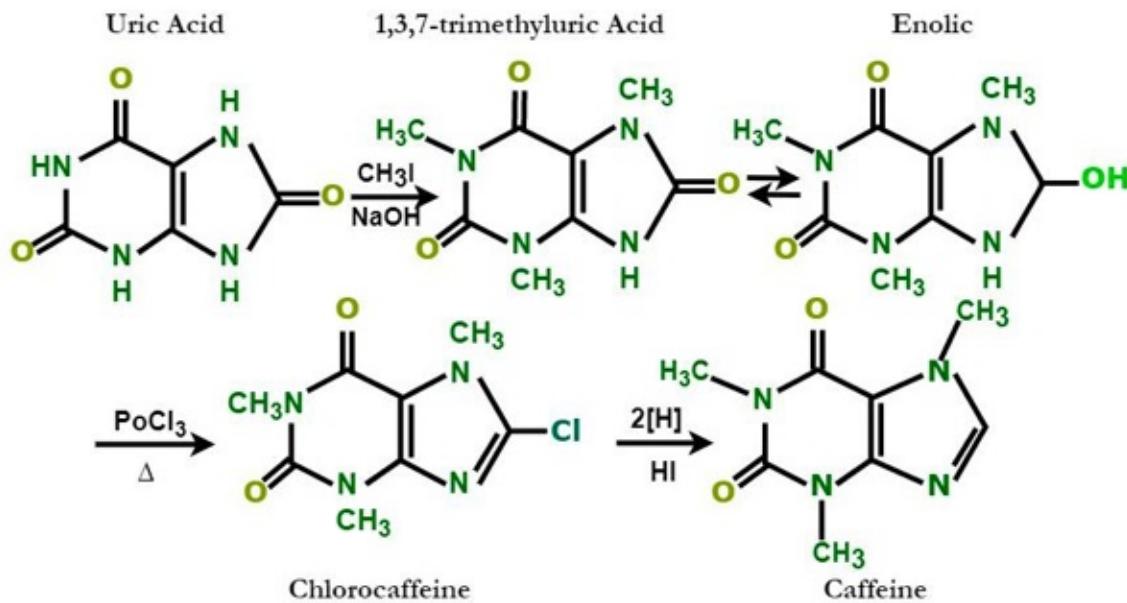
图数据集



2021年全球互联网地图（容量、流量和云基础设施）

<https://global-internet-map-2021.telegeography.com/>

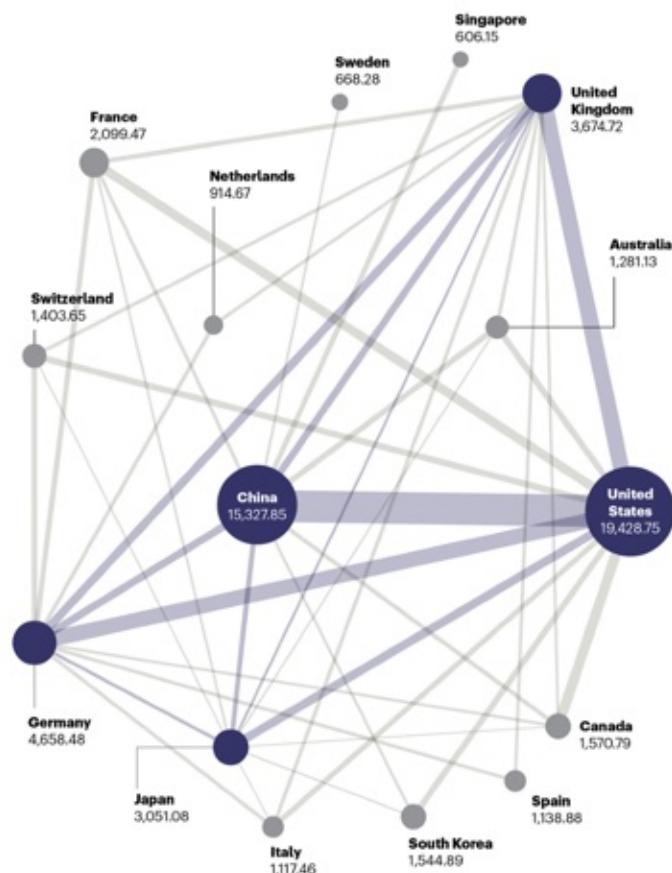
图数据集



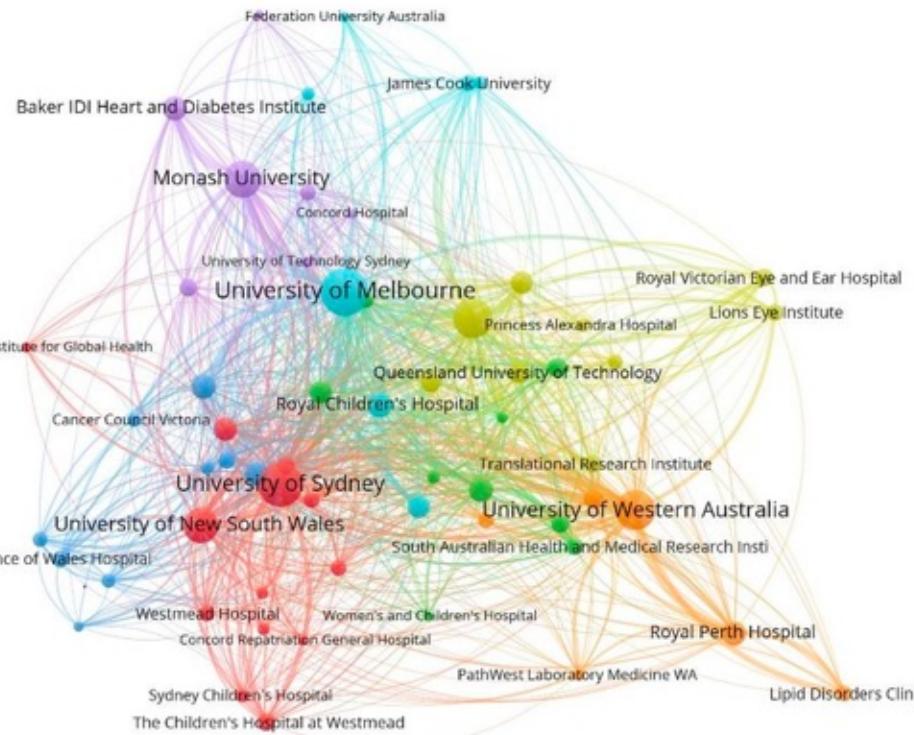
图数据集

Legend

- Big 5 country
 - Non-big 5 country
 - Big 5 collaboration
 - Big 5-non-big 5 collaboration
- 250 collaboration score, 2021*
- 500 Share, 2021*
- 1,000 Share, 2021*
- 1,000 collaboration score, 2021*
- 5,000 Share, 2021*



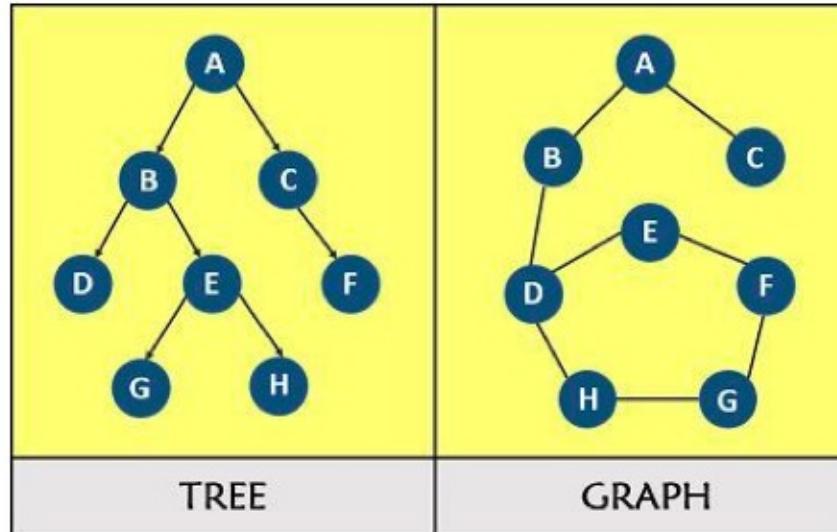
*Data for 2021 represent the period October 2020–September 2021.



学术圈的论文合作关系网

图数据集

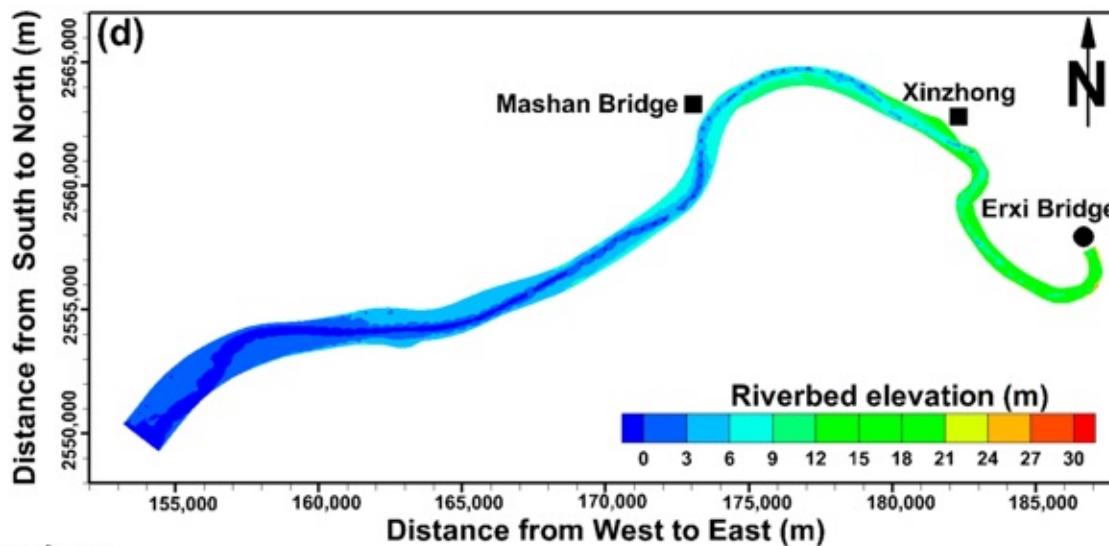
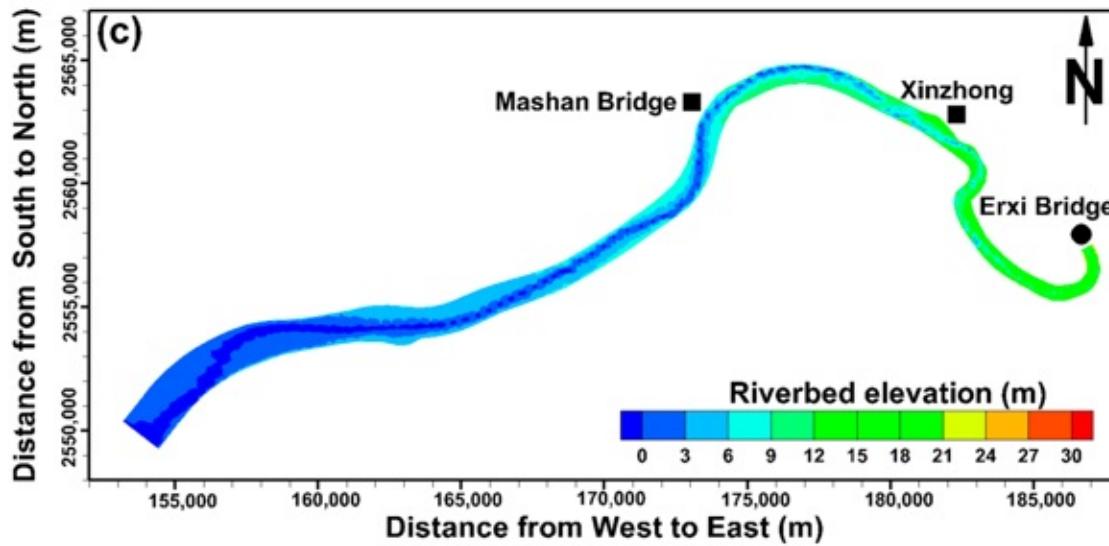
- 树 (树形结构, Tree)
 - 没有回路的图, 任意两个顶点间有且只有一条路径, 子树不相交
 - 除了根节点外, 其他节点有且只有一个父节点
- 树形结构的优点
 - 它是一种非线性数据结构, 能更方便快捷地访问数据。线性数据结构按顺序存储数据, 而树形结构允许从不同方向访问数据。



有序数据集

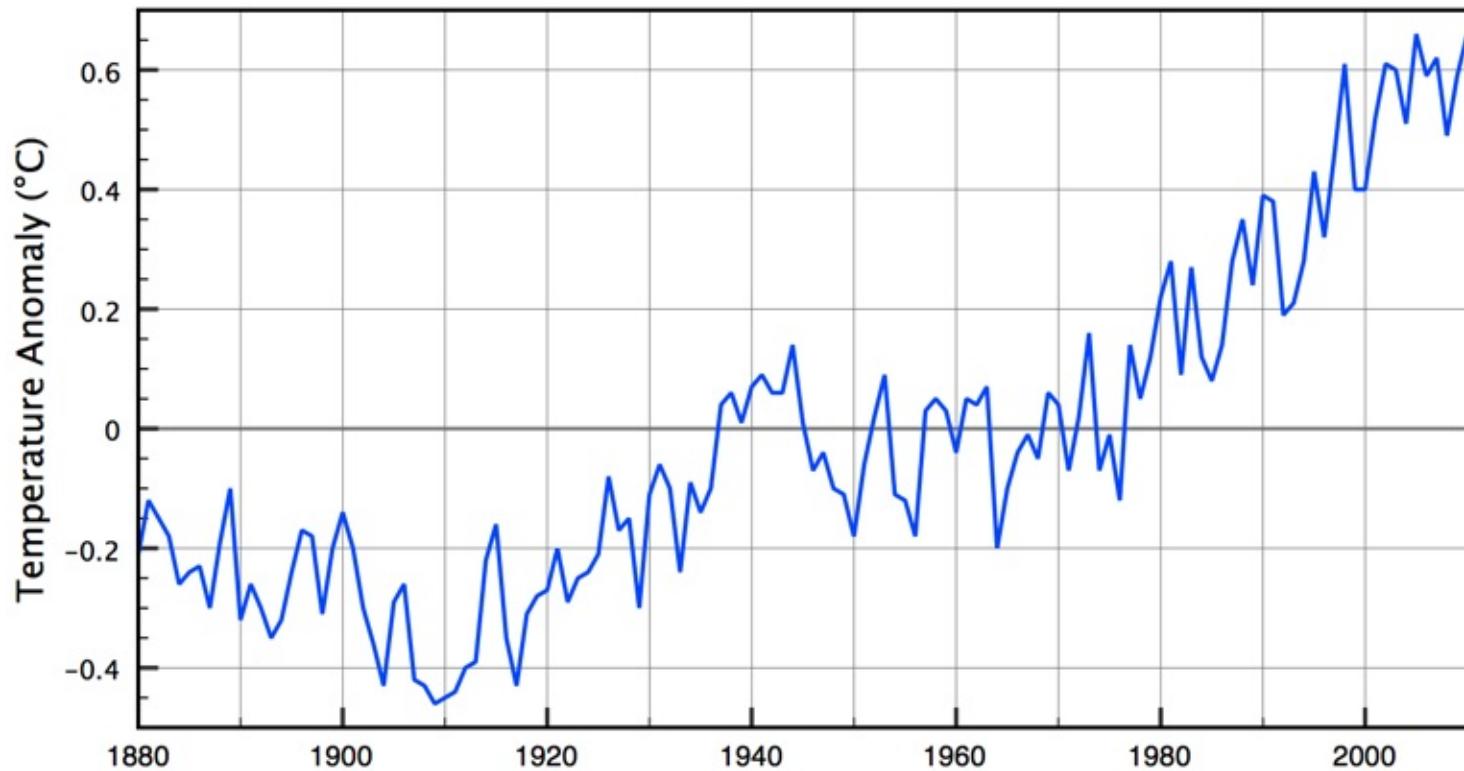
- 有序数据是具有某种顺序（方向性）的数据集。
 - 方向：时间方向、空间方向、逻辑方向、抽象方向
- 有序数据出现在许多领域，尤其是地球科学、气象学和医学。
- 有序数据是一种“区间型”数据，原点（零刻度）的位置可以任意定义

有序数据集



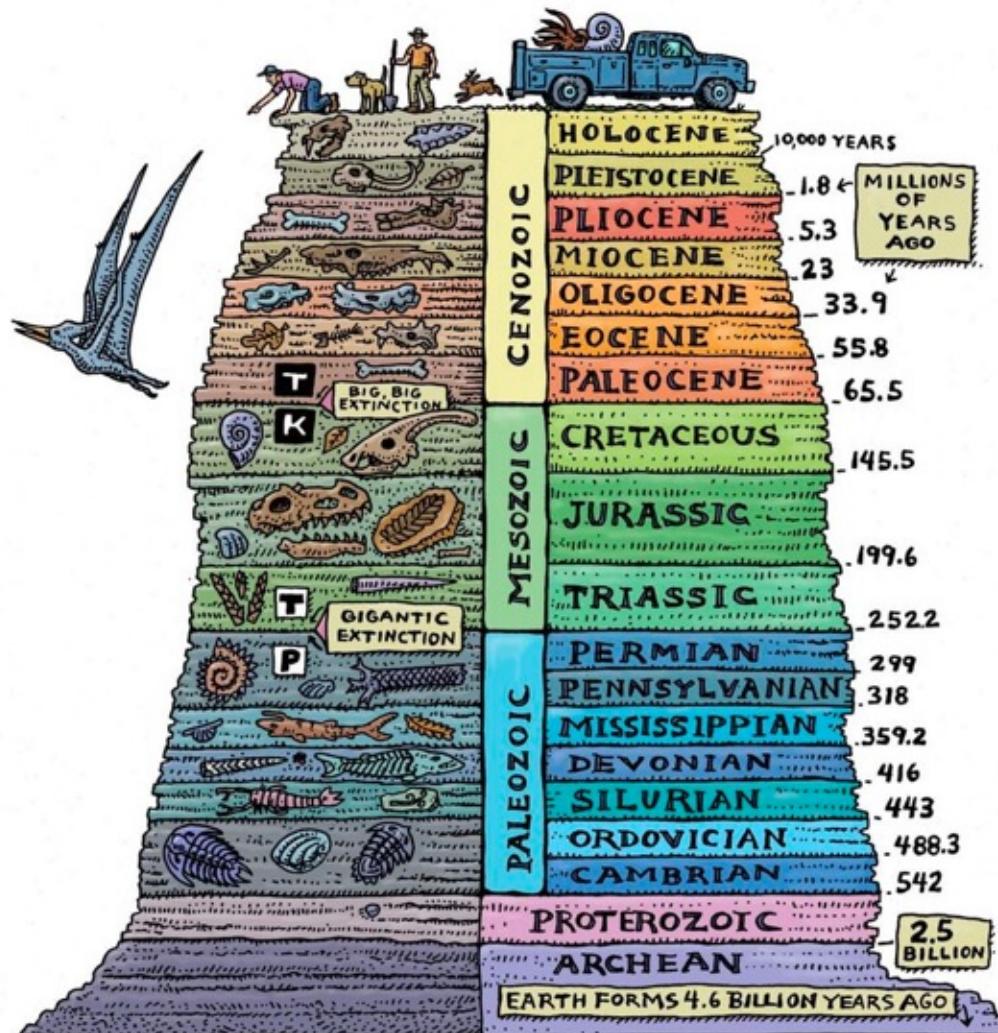
河床高度的空间变化

有序数据集



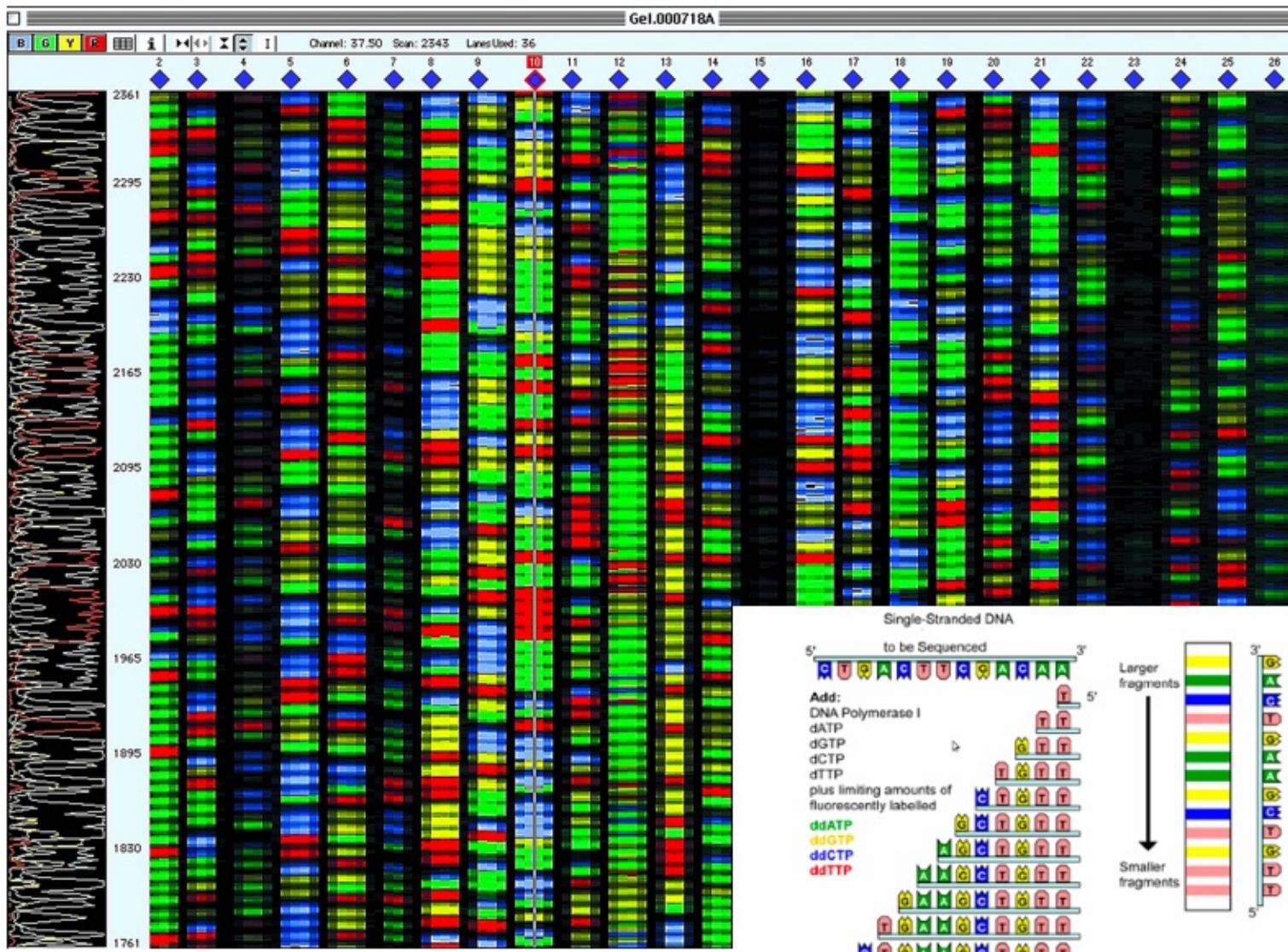
温度异常值数据（时间序列）

有序数据集



地质年代表

有序数据集



基因测序

数据清洗



数据的来源

数据的源头

实验测量或实际观测（自然科学、社会科学）

计算机仿真（气候模型、有限元模拟）

用户自发提供（社交媒体、新闻稿、电商数据传输）

数据获取方式

传统时代

- 文件输入、输出为主
- 较高的时间成本、人力成本

大数据时代

- 基于网络的多源数据交换成为了主流

数据获取的挑战

- 数据的多样性：不同的数据语义表述、储存格式
- 差异的源头包括：
 - 不同的安全要求
 - 不同的用户类型
 - 不同的数据格式
 - 不同的数据来源

数据获取协议 Data Access Protocol (DAP)



数据获取协议 DAP

- 一种通用的数据获取标准，在科研领域有广泛的应用
- 通过定义基于网络的数据获取句法来完善数据交换的机制，改善数据分享和获取的效率
- DAP是中立的协议，它在文件里提供数据格式、位置和数据提供组织的基本信息，并且以网页化的方式，和FTP、HTTP、SRB（Source Route Bridging 源路由网桥）、开放地理空间联盟等网络协议兼容
- 促成了跨领域的国际网络数据获取协议，成为NASA/ESE等热门数据源所使用的标准

数据获取协议 DAP

The screenshot shows the Earthdata website interface. At the top left is the NASA Earthdata logo with the text "EARTHDATA OPEN ACCESS FOR OPEN SCIENCE". To the right are navigation links: Data, Topics, Learn, Engage, About, and a search icon. Below the header, a breadcrumb trail reads: Earthdata / Earth Science Data and Information System (ESDIS) Project / ESDIS Standards Coordination Office (ESCO) / Standards and Practices / The Data Access Protocol - DAP 2.0. The main content area features a large image of Earth from space. On the left side of this area, the text "ESDS Program" is visible above the title "The Data Access Protocol - DAP 2.0".

Summary

This page defines the Open-source Project for a Network Data Access Protocol ([OPeNDAP[®]](#)) Data Access Protocol (DAP), a data transmission protocol designed specifically for science data. The protocol relies on the widely used and stable Hypertext Transfer Protocol (HTTP) and Multipurpose Internet Mail Extensions (MIME) standards, and provides data types to accommodate gridded data, relational data, and time series, as well as allowing users to define their own data types.

Technologies

Open-source Project
for a Network Data
Access Protocol
(OPenDAP)

Featured

NASA provides

数据获取协议

- 进一步更新后的DAP 4提供了更多的数据类型和传输功能，适用于更多样的环境。
- OPeNDAP (<http://www.opendap.org>) 开发了科学数据联网框架，以便于科学家们能快速地获取数据并进行格式转换。
 - OPeNDAP 是一个致力于简化科学数据网络的框架。
 - OPeNDAP 提供的软件使本地储存的数据（无论格式如何）可以远程访问。
 - OPeNDAP 还提供了将现有应用程序转换为 OPeNDAP 客户端的工具。
 - OPeNDAP 软件是免费的。

The screenshot shows the GitHub profile for the OPeNDAP organization. At the top, there's a logo featuring a stylized orange and yellow flame or leaf-like shape next to the text "OPeNDAP™". Below the logo, it says "OPeNDAP is a non-profit that develops Advanced Software for Remote Data Retrieval". It lists 23 followers, is located in Colorado, Montana, Oregon and Rhode Island, and provides links to their website and support email. The main navigation bar includes "Overview", "Repositories 99", "Projects", "Packages", and "People 1".

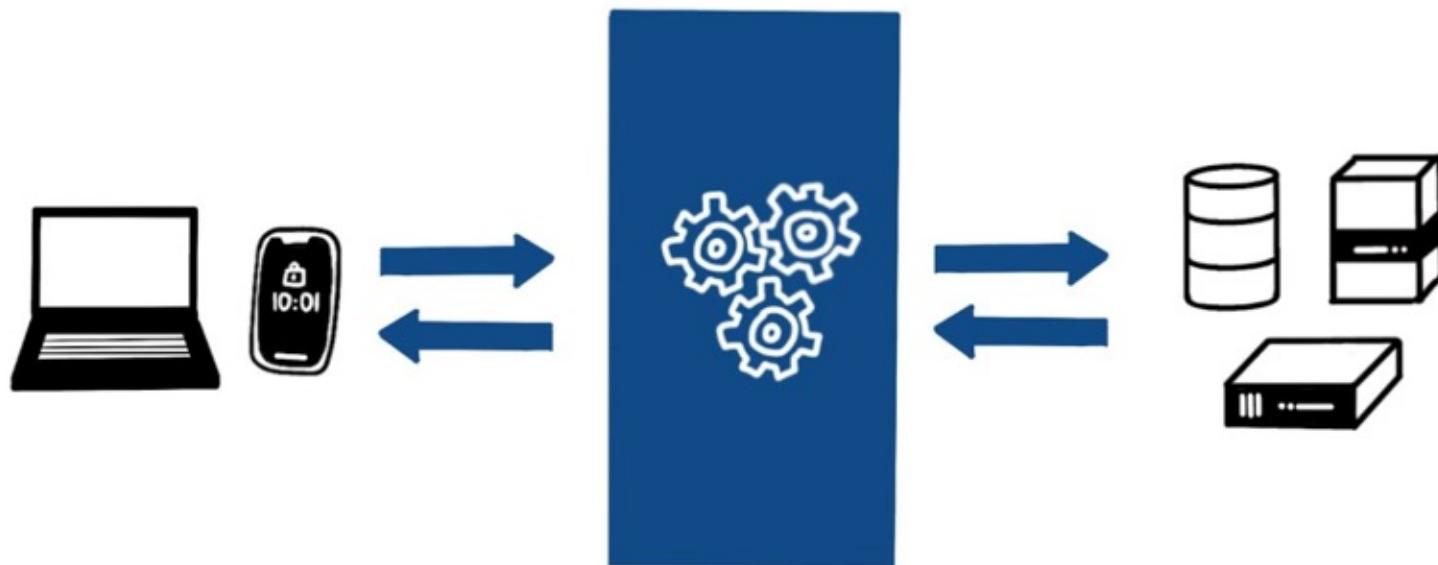
Pinned

- bes** Public
The BES framework, which forms the basis for the Hyrax server.
C++ ⭐ 12 🏷 18
- olfs** Public
The front end for the Hyrax Data Server.
Java ⭐ 5 🏷 7
- hyrax** Public
Get and build the Hyrax data server.
Python ⭐ 34 🏷 11
- libdap4** Public
A new version of libdap that contains both DAP2 and DAP4 support.
HTML ⭐ 13 🏷 12
- documentation** Public
- NASA-tutorials** Public
Tutorials for NASA's OPeNDAP services.
Jupyter Notebook ⭐ 2 🏷 3

OPeNDAP

应用程序编程接口 API

- API (Application Programming Interface)
- 定义：API 是一些预先定义的函数，目的是提供应用程序与开发人员基于某软件或硬件得以访问一组例程的能力，而又无需访问源码，或理解内部工作机制的细节。



应用程序编程接口 API

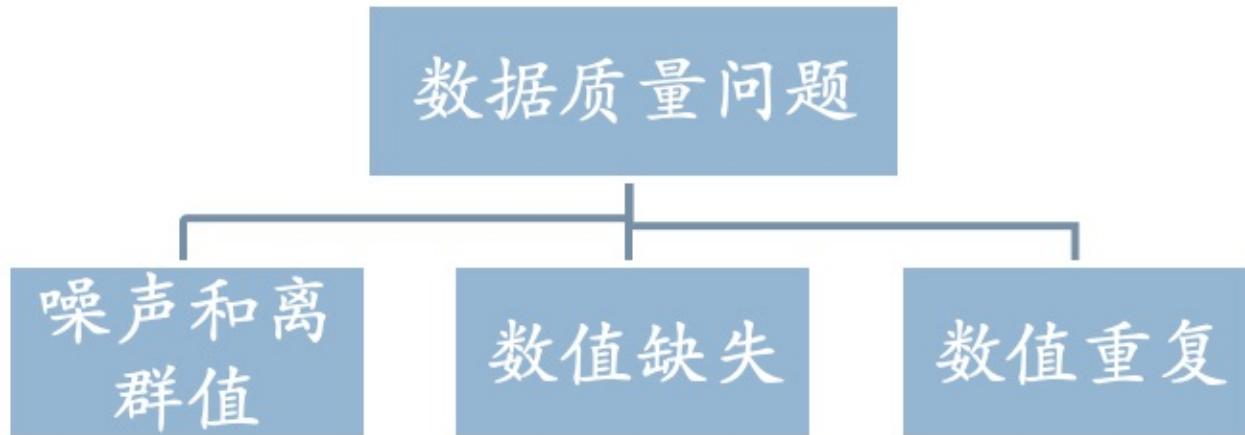
- Macrostrat (一个关于沉积地层的空间数据库) 的API
- https://macrostrat.org/api/units?lith=sandstone&interval_name=Cretaceous



```
{  
    "success": {  
        "v": 2,  
        "license": "CC-BY 4.0",  
        "data": [  
            {  
                "unit_id": 6698,  
                "section_id": 1667,  
                "col_id": 232,  
                "project_id": 1,  
                "col_area": 5425.838,  
                "unit_name": "Unnamed Rocks in Coastal Belt Franciscan Complex",  
                "strat_name_id": null,  
                "Mbr": "",  
                "Fm": "",  
                "Gp": "",  
                "SGp": "",  
                "t_age": 12.1775,  
                "b_age": 90.825,  
                "max_thick": 0,  
                "min_thick": 0,  
                "outcrop": "both",  
                "pbdb_collections": 0,  
                "pbdb_occurrences": 0  
            },  
            {  
                "unit_id": 25783,  
                "section_id": 7544,  
                "col_id": 1251,  
                "project_id": 7,  
                "col_area": 48682.523,  
                "unit_name": "Totogalpa Fm",  
                "strat_name_id": 9147,  
                "Mbr": "",  
                "Fm": "Totogalpa",  
                "Gp": "",  
                "SGp": "",  
                "t_age": 33.9,  
                "b_age": 80.4375,  
                "max_thick": 0,  
                "min_thick": 0,  
                "outcrop": "",  
                "pbdb_collections": 0,  
                "pbdb_occurrences": 0  
            }  
        ]  
    }  
}
```

数据清洗

- 未经处理的原始数据中，通常包含大量的无效数据，这些数据在存储之前，就应该被过滤掉。



数据清洗



噪声和离群值的处理

- 纠正或删除错误的数据

缺失值的处理

- 将缺失的数据补充完整

重复值的处理

- 把多余或重复的数据筛选出来并清除

修正数据格式

- 调整错误的数据格式，使其真正机器可读

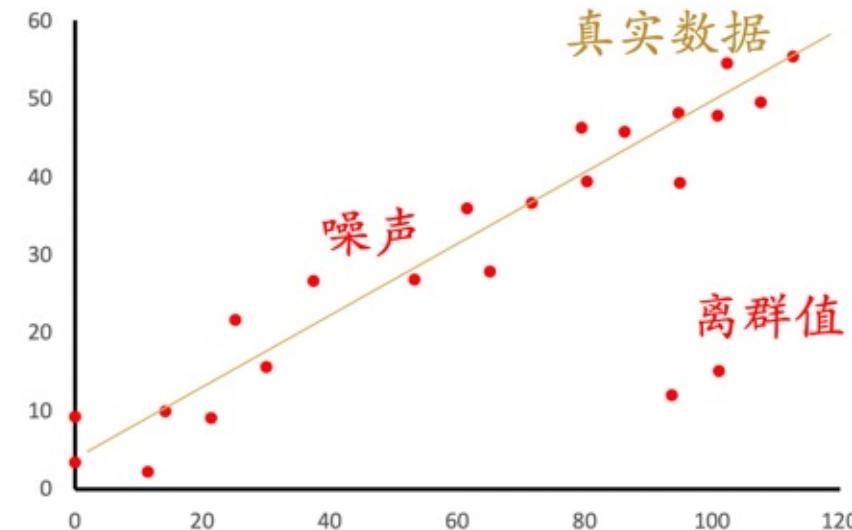
数据清洗

噪声 Noise

- 数据集中的干扰数据（对场景描述不准确的数据），它们拥有随机的误差和方差。
- 来源：实验操作的不精准、仪器精密程度有限、分辨率不足。

离群值 Outlier

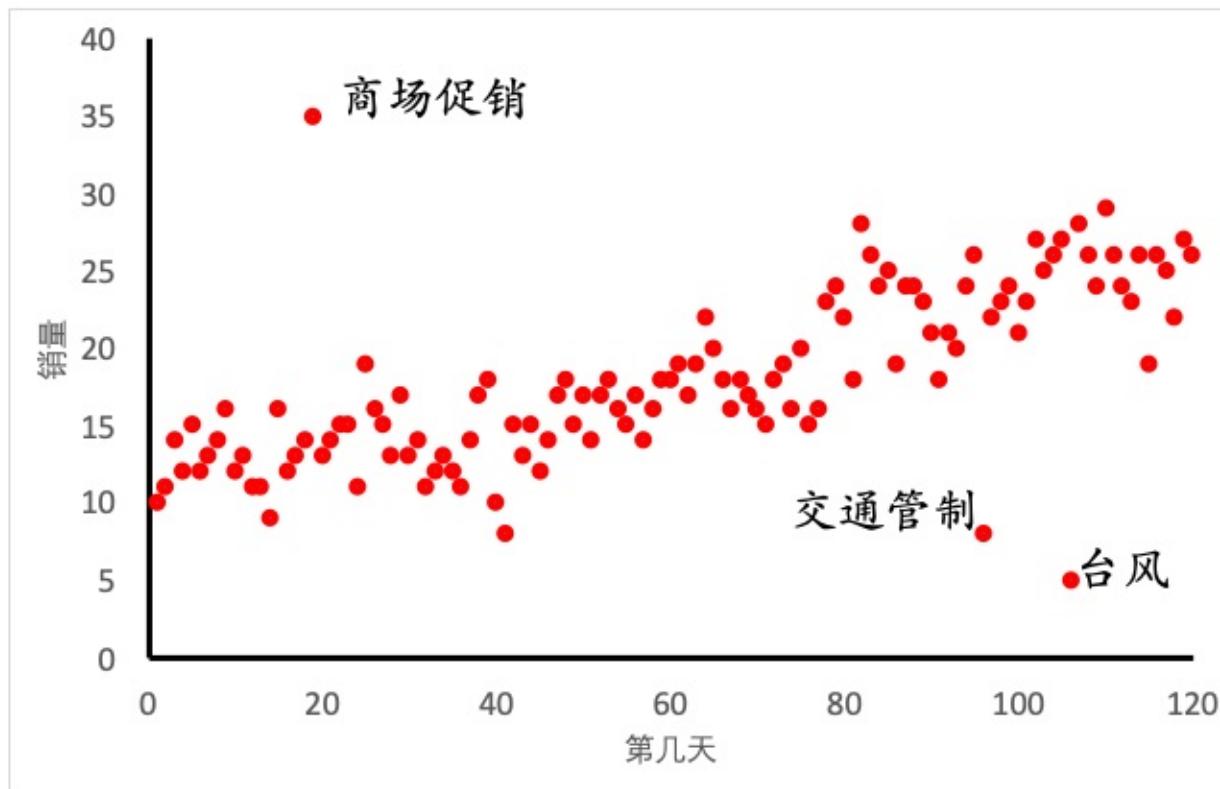
- “不合群”的数据点，与大多数观测点的数据值有较大的出入。
- 来源：可能是数据本身，也可能是噪声带来的（实验操作失误、机器故障、样本污染等）。



为什么要处理：离群值通常不具备代表性，会影响数据的收敛和模型的精确度

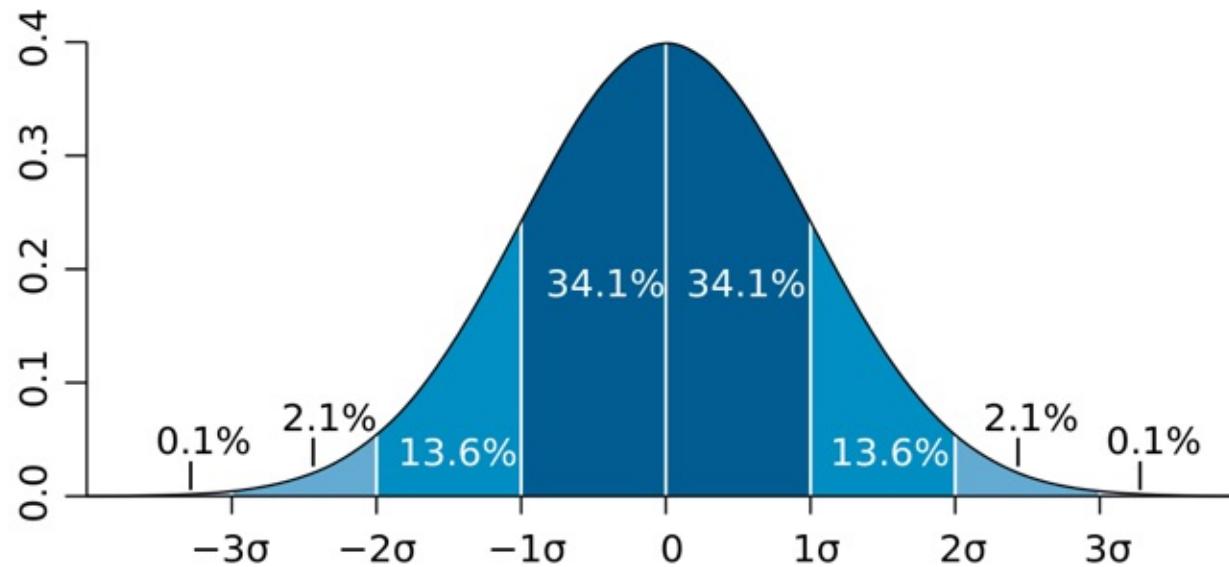
数据清洗：离群值处理

- 人工手动检查
 - 根据业务经验和对数据的理解来判断
 - 结合研究内容或商业需求，手动决定数据点的去留



数据清洗：离群值处理

- 标准差法
 - 适用于正态分布的数据
 - 在正态分布中，以平均值为准，正负3个标准差内涵盖99.7%的数据值
 - 超过三个标准差的数据可被认为是离群值



数据清洗：离群值处理

- 分箱法 (binning)
 - 适用于非正态分布的数据，通过考察相邻数据来确定最终值。
 - 按照属性的值把数据划分成不同的区间，再考察每一个区间中的数据
- 优点
 - 提高模型的稳健性和包容性
 - 还可用于平滑数据，在数据量极大的情况下有助于提高模型的运算效率

