



大数据可视化

挑战与趋势

陈为

chenwei@cad.zju.edu.cn

浙江大学CAD&CG国家重点实验室



数据可视化

五维统计数据的生动可视化



数据可视化

- 创建并研究数据的**视觉表达** (Visual Representation)
 - 输入：数据 (data)
 - 输出：视觉形式 (visual form)
 - 目标：深入理解 (insight)



数据



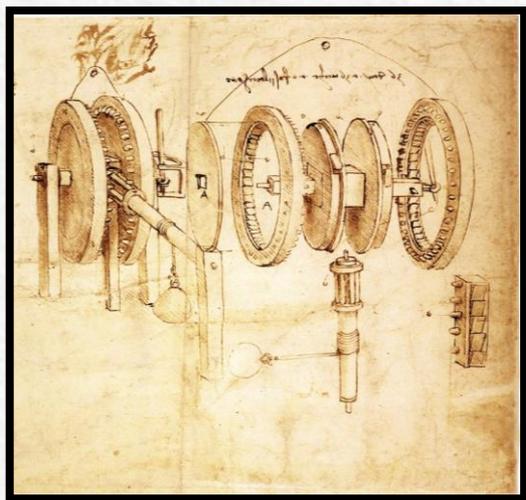
视觉形式



深入理解

数据可视化的主要任务

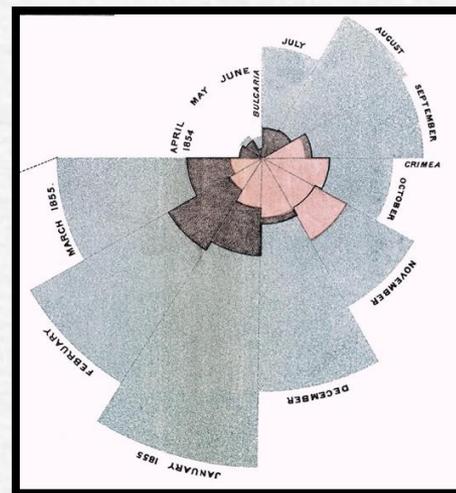
- 表示数据 - Represent
- 分析数据 - Analyze
- 交流数据 - Communicate



三维素描图



霍乱病例的分布



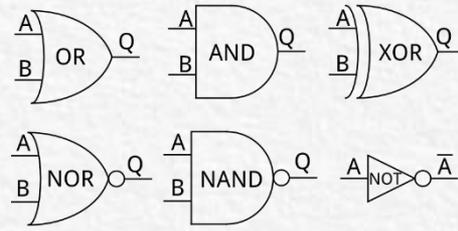
英国东征士兵死亡原因

思维系统

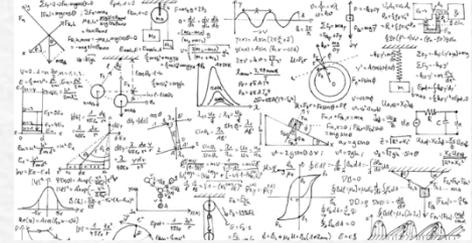
表达 (符号) + 操作规则



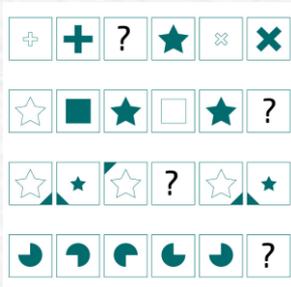
语言



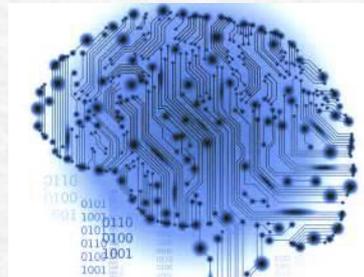
逻辑



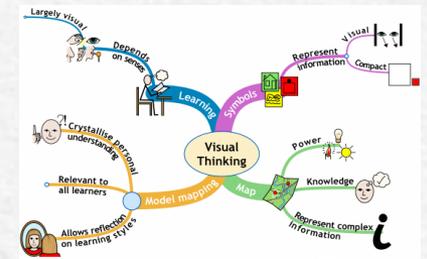
数学



推理与统计



计算



可视化与形象思维 (认知)

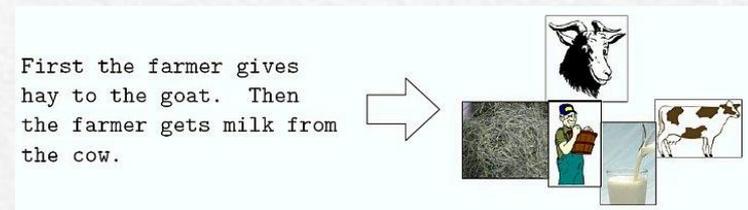
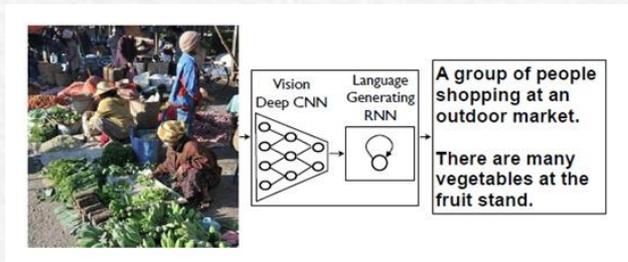
可视化与形象思维 (认知)

Why is a Diagram (Sometimes
Worth 10,000 Words

Larkin and Simon, Cognitive Science, 1987

一图胜千言

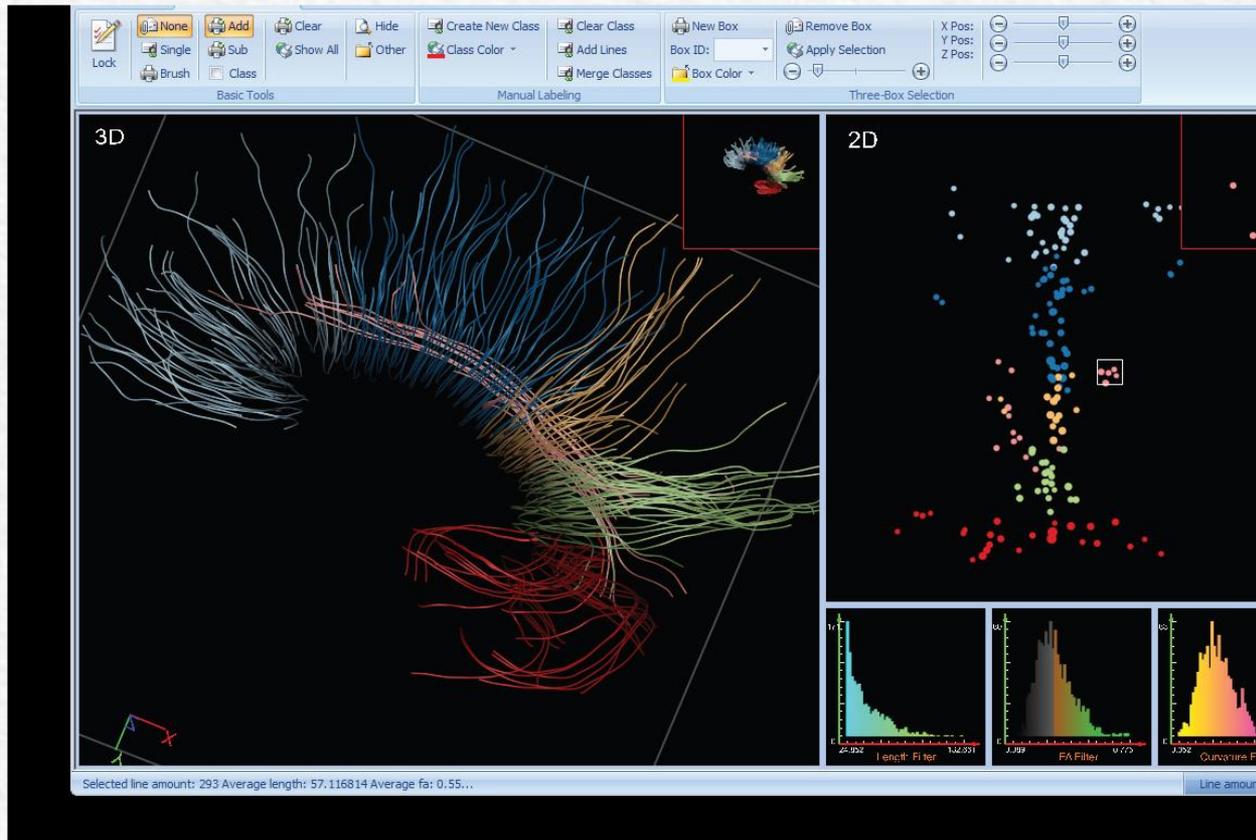
图表表达与句型表达具有信息和计算上的等价性
[Simon 1978]



图像到文字的互相转换

一图胜千言

图表具有拓扑和几何的关联，将信息基于位置进行索引，所见处即所得；



一图胜千言

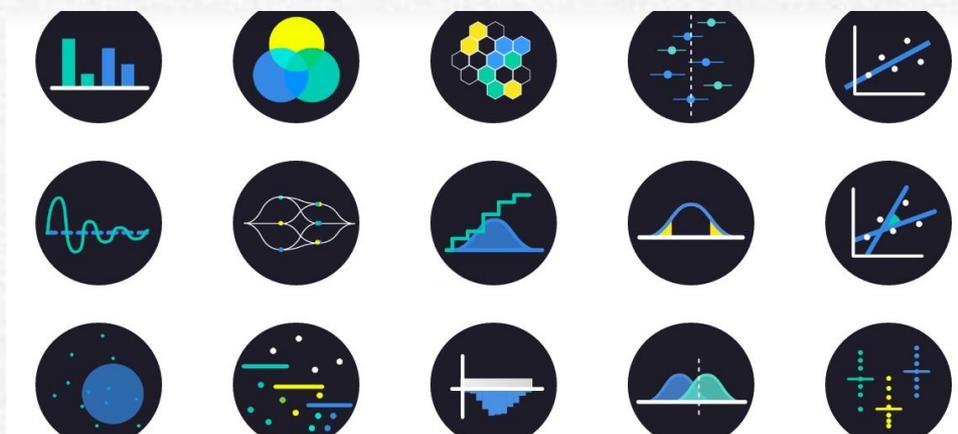
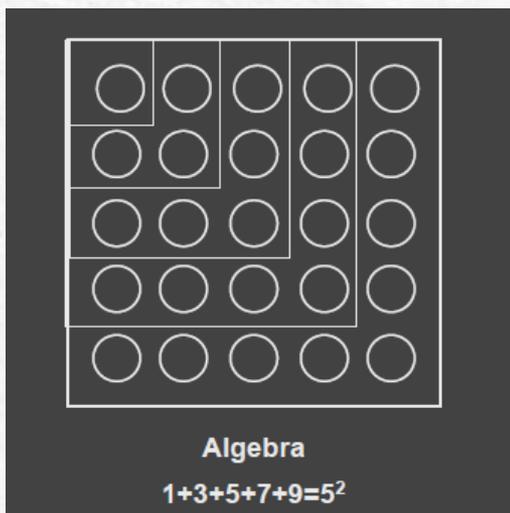
- 句型表达具有时间或逻辑方面的序列，显式地表达了单个元素。
- 句型表达假设每句话是串行阵列；而图表表达有一个简洁的语义网络，认知时只需要在不同的节点间定位。



重要人物的关系分析系统

一图胜千言

在求解问题时，图表表达可以提供搜索与认知的便利；句型表达在搜索时需要记住更多的信息。



<https://seeing-theory.brown.edu/bayesian-inference/index.html>

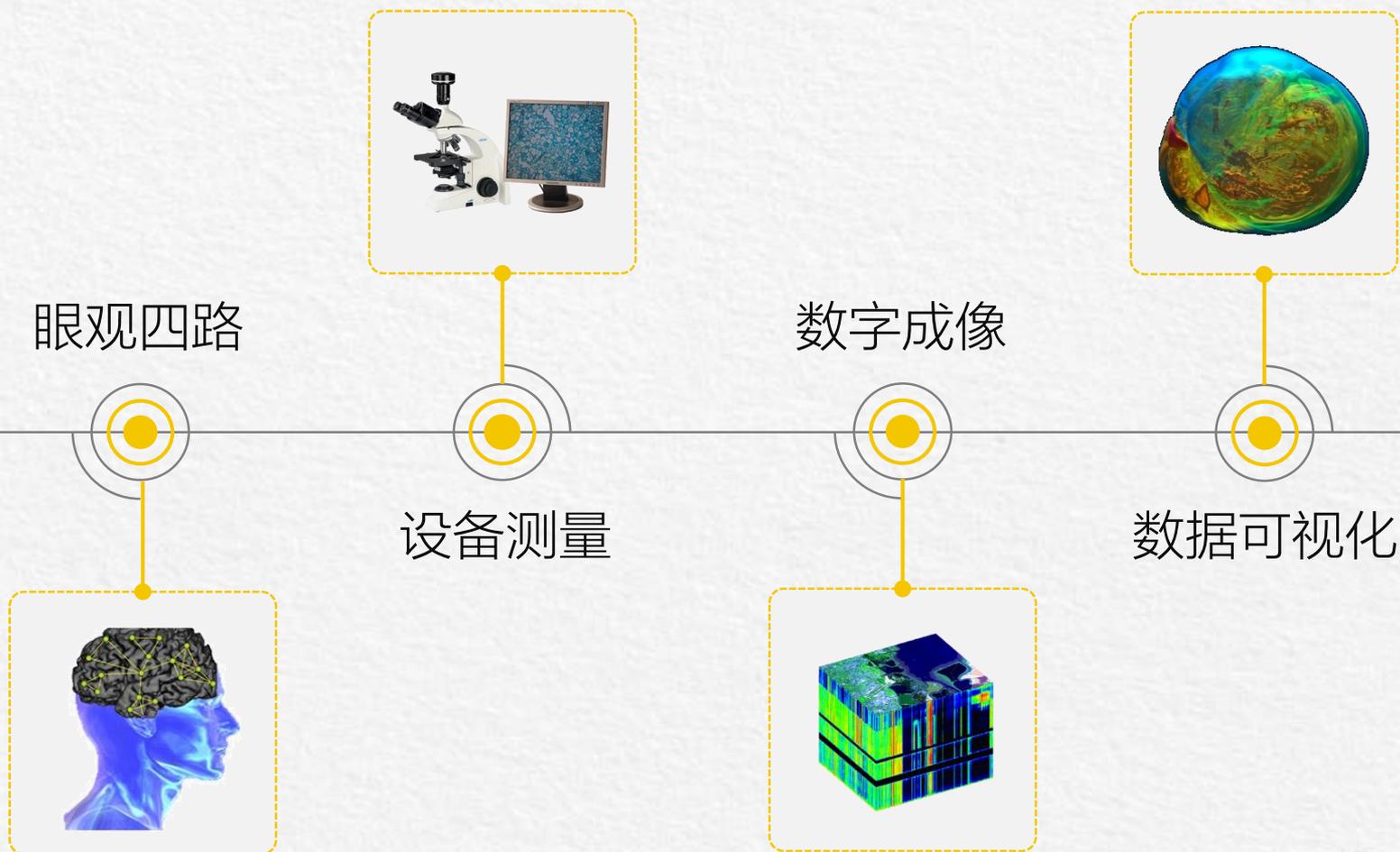
可视化是一种外部认知的方法，即：
如何利用人眼的感知能力和人脑之外的
资源，提升人脑的认知能力。



Stuart Card, 1946-

美国工程院院士
1970年代就职于美国施乐公司，
鼠标和GUI的主要推动者

可视化：大数据时代的成像利器



交互智能可视分析



不可信

不存在可信的自动分析方法



未定义

分析任务没有良好的定义,甚至不知道任务是什么



人更有效

应急、复杂环境和对抗事件等条件下,人的智能更有效和可靠



大数据可视化 现状

新一代人工智能规划中大数据智能

可视化与可视分析是人类理解数据的导航仪：运用与人类视认知相一致的图形展示数据内在结构与规律，增强理解和分析效率。



科学性： Science连续发文指出：借助可视化手段将人机智能有机结合，形成可视分析环境，可有效提升数据关联分析的效率

2014

1. 大数据从“概念”走向“价值”
2. 大数据应用广泛
3. 大数据成为国家战略
- 4. 大数据分析可视化**
5. 大数据产业成为战略性产业
6. 数据商品化与数据共享联盟化
7. 基于大数据的推荐与预测流行
8. 深度学习与大数据智能成为主流
9. 数据科学成为重要学科
10. 大数据生态

4. 大数据分析可视化

2015

1. 大数据分析成为数据价值化的热点
2. 数据科学带动学科融合，但自身尚未成体系
3. 与各行业结合，跨领域应用
4. “物云移社”融合，产生综合价值
5. 大数据成为国家战略
6. 大数据成为国家战略
7. 计算模式：深度学习、众包计算
- 8. 可视化分析与可视化呈现**
9. 大数据人才与教育
10. 开源系统将成为主流选择

8. 可视化分析与可视化呈现

2016

- 1. 可视化推动大数据平民化**
2. 多学科融合与数据科学的兴起
3. 大数据成为国家战略
4. 大数据成为国家战略
5. 大数据成为国家战略
6. 《促进大数据发展行动纲要》驱动产业生态
7. 深度分析推动大数据智能应用
8. 数据权属与数据主权备受关注
9. 互联网、金融、健康、城市、企业数据化增长点
10. 开源、测评、大赛技术生态

1. 可视化推动大数据平民化

2017

1. 机器学习继续成智能分析核心技术
2. 人工智能和脑科学相结合，成大数据分析领域的热点
3. 大数据的安全和隐私持续令人担忧
4. 多学科融合与数据科学兴起
5. 大数据处理多样化模式并存融合，成大数据分析主流
6. 大数据成为国家战略
7. 大数据成为国家战略
8. 大数据成为国家战略
9. 推动数据立法，重视个人数据隐私
10. 可视化技术和工具提升大数据分析工具的易用性

10. 可视化技术和工具提升大数据分析工具的易用性

重要性： 中国计算机学会每年发布十大大数据发展趋势报告

新一代人工智能规划中大数据智能

中国科技创新2030 “新一代人工智能” 和 “大数据” 专项均将**可视化**和**可视分析**列为**大数据智能急需突破的关键共性技术**。

2017-07/20/content_5211996.htm

2. 建立新一代人工智能关键共性技术体系。
 围绕提升我国人工智能国际竞争力的迫切需求，新一代人工智能关键共性技术的研发部署要以算法为核心，以数据和硬件为基础，以提升感知识别、知识计算、认知推理、运动执行、人机交互能力为重点，形成开放兼容、稳定成熟的技术体系。
 知识计算引擎与知识服务技术。重点突破知识加工、深度搜索和可交互核心技术，实现对知识持续增量的自动获取，具备概念识别、实体发现、属性预测、知识演化建模和关系挖掘能力，形成涵盖数十亿实体规模的多源、多学科和多数数据类型的跨媒体知识图谱。

2008年后，美、欧盟、日均成立国家可视分析研究中心。国内外著名企业成立独立部门，研发新兴可视化与可视分析技术。



周鸿祎：“看见”才是首要职责，都看不见还谈什么防火墙
 2015-09-29 18:01 稿源：钛媒体 3条评论 撤稿纠错



阿里全力投入网络可视化技术，行业发展潜力巨大
 2017-12-01 09:20 概念股 区块链



急迫性：2018年11月，美国拟提议的最新14类技术出口管制之第六条：数据分析（可视化、自动分析算法、上下文感知计算）

DEPARTMENT OF COMMERCE
 Bureau of Industry and Security

FOR FURTHER INFORMATION CONTACT:
 Kirsten MacLean, Office of National Security and Technology Transfer Controls, Bureau of Industry and Security, Department of Commerce, Phone: (202) 482-0092; Fax: (202) 482-2933; Email: Kirsten.MacLean@is.dhs.gov.

REGULATORY INFORMATION:
 Background
 As part of the National Defense Authorization Act (NDAA) for Fiscal Year 2018, Public Law No. 115-232, Congress enacted the Export Control Reform Act of 2018 (the Act or ECRAR). Section 1708 of the Act authorizes Commerce to establish appropriate controls, including license controls, on the export, reexport, or transfer (in-country) of emerging and foundational technologies that are essential to the national security of the United States and are not described in Section 710(a)(4)(C)(i) of the Defense Production Act of 1950, as amended. Emerging and foundational technologies, as defined by ECRAR, will be determined by the interagency process that will consider both public and classified information as well as information from the Emerging Technology Technical Advisory Committee of the technology and countries to which exports from the United States are controlled for export controls, at a minimum must require identification of the emerging and foundational technologies, including those subject to a U.S. embargo, including Responses to the ANPTM will help Commerce and other agencies identify and assess emerging technologies for the purpose of updating the export control lists without impacting national security or hampering the export of U.S. commercial sectors to keep pace with international advances in emerging fields.

Emerging Technologies
 To assist BIS in identifying emerging technologies that are essential to the national security of the United States, this ANPTM seeks public comment on criteria for defining and identifying emerging technologies. This ANPTM describes certain categories of technologies that are currently subject to the EAR list controlled only to multilateral countries, countries designated as supporters of international terrorism, and restricted site users or end users. These categories are a representative list of the

新一代人工智能规划中大数据智能

可视化与可视分析的核心挑战是：如何利用人的**感知能力**，增强**有限的认知能力**，以应对理解和分析**复杂数据的迫切需求**。

新挑战

数据复杂

数据尺度**海量**
时域演化**多变**
内蕴动态**关联**

认知低效

显示空间**固定**
视觉感知**有限**
任务交互**无限**

新理论

高效**可视表达与呈现方法**

探索式**可视分析理论与方法**

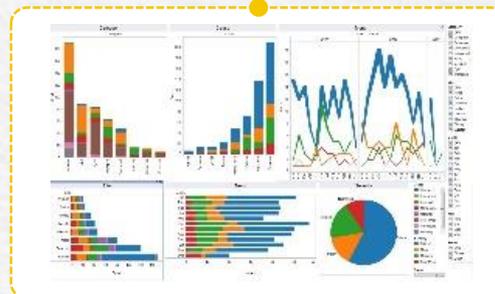
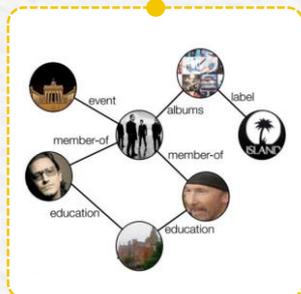
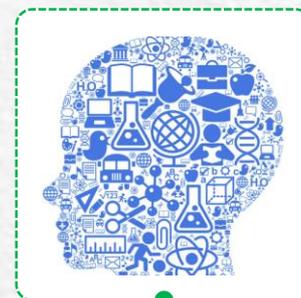
新技术

通用信息**可视化框架与软件**

沉浸式**可视分析软件与装备**

可视化与人工智能 1.0

可视化与从数据到知识的链条中的每个环节互助互利



可视化与人工智能 1.0：可视数据清洗



TRIFACTA

www.trifacta.com

探索

结构化

清洗

丰富

验证

发布

The screenshot displays the Trifacta data cleaning workflow. It features several panels: 'Job Results' showing a 94% valid rate, 'Data Exploration' with a bar chart for 'created_date', and a detailed 'Export Results' dialog. The dialog shows the export of 8.61MB of JSON data to a Hive table named 'Tweets_trifacta' in 'avro' format. The background interface also shows a 'View Script' panel with a table of data including columns for '#', 'friends_count', 'location', and 'zip'.

Export Results Dialog:

- Format: avro
- Publish to Hive: Database: default, Table: Tweets_trifacta
- Format: ppt

View Script Panel:

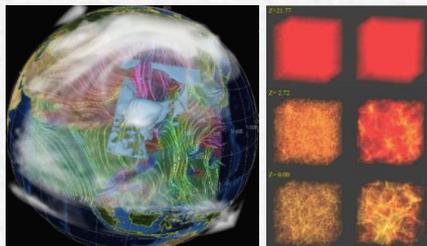
#	friends_count	location	zip
Valid	108,545	Valid	14,475
Mismatched	0	Mismatched	509
Empty	0	Empty	147

Top 10 values (location):

- Stockholm (240)
- Sweden (240)
- USA (760)
- USA - New York (240)
- Stockholm, Sweden (240)
- New York (240)
- Göteborg (443)
- Sverige (381)
- Istanbul (410)
- Uppsala (687)

可视化重要应用

大科学



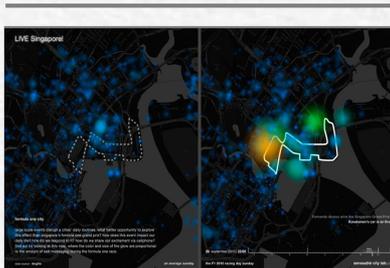
大安全



大工程



物联网与智慧城市

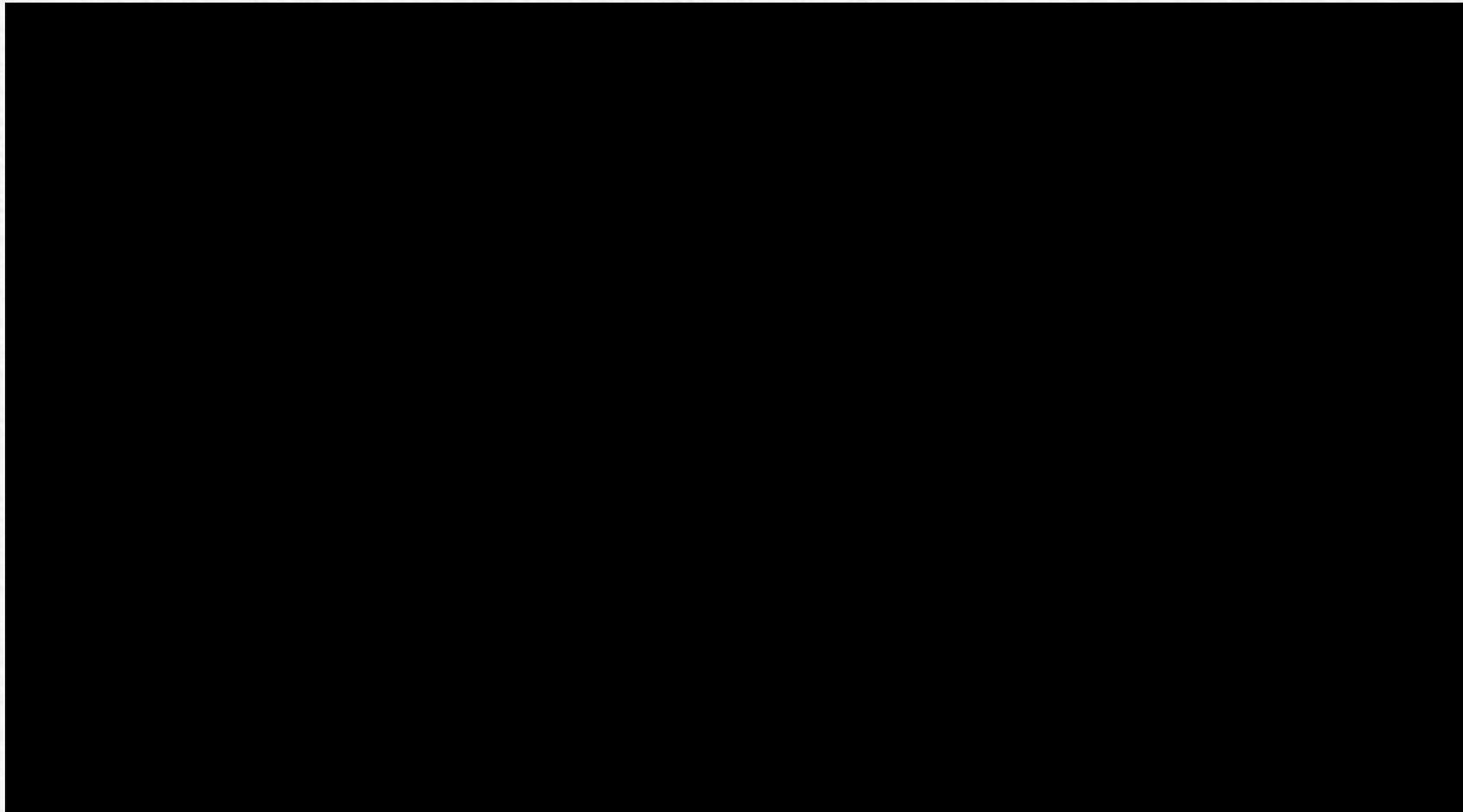


互联网与社交媒体



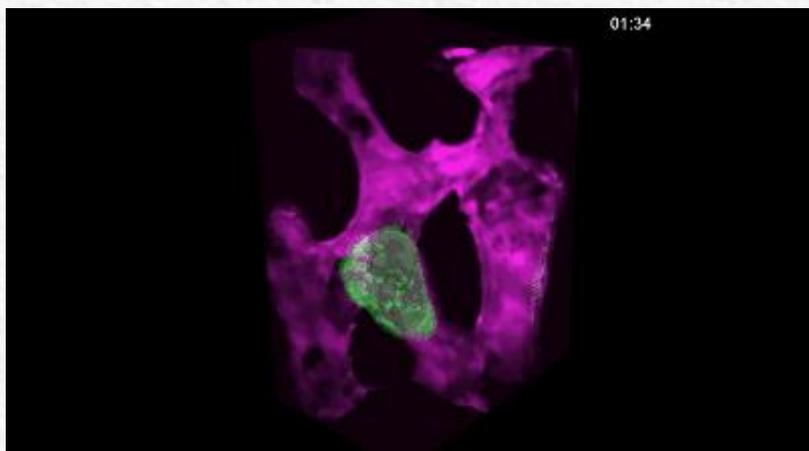
可视化重要应用：大科学

可视化成为基础自然科学研究的**必要手段**，是科学大数据发展的**必需**

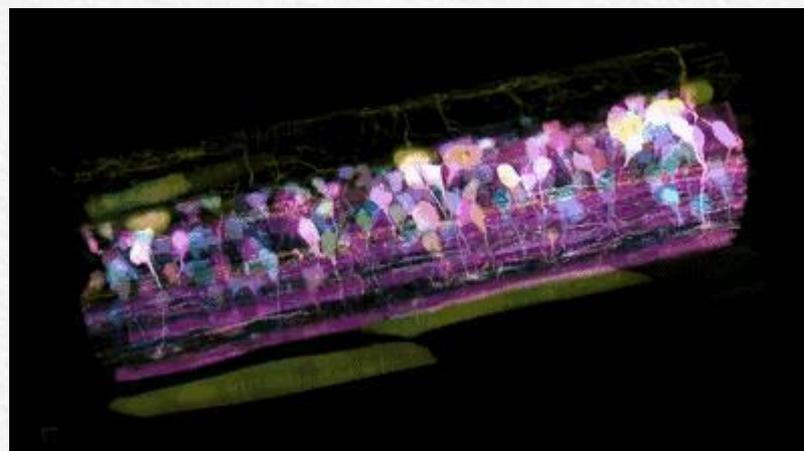


海洋

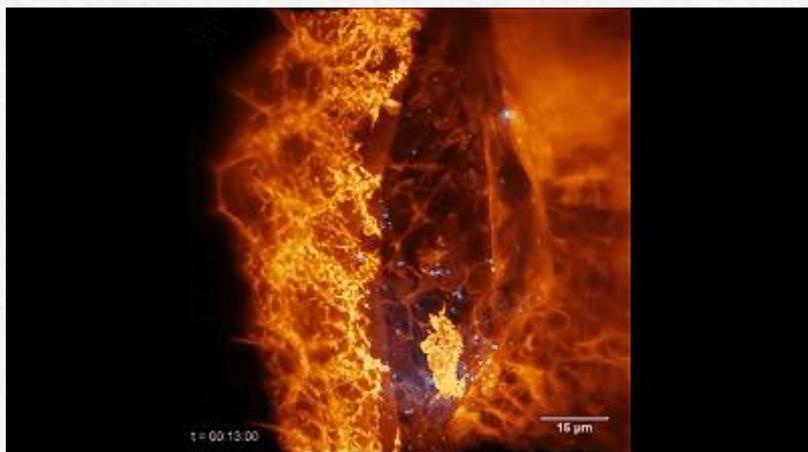
活组织细胞运动三维影像可视化



乳腺癌细胞转移



脊髓神经回路



免疫组织穿梭



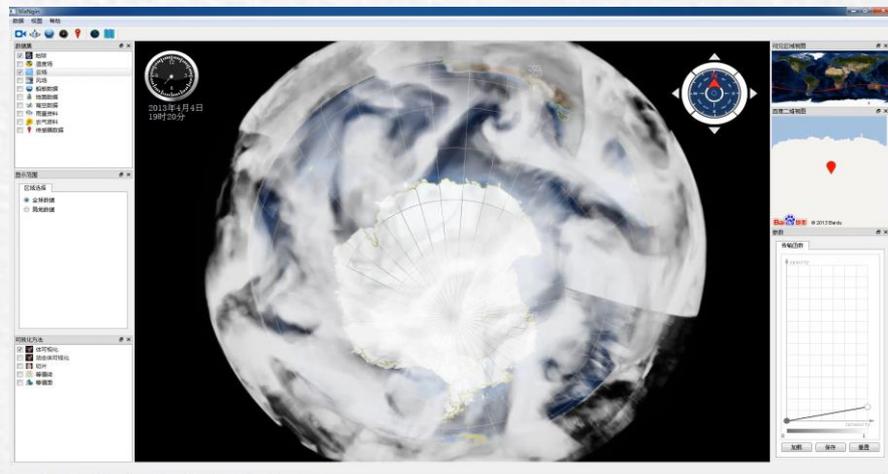
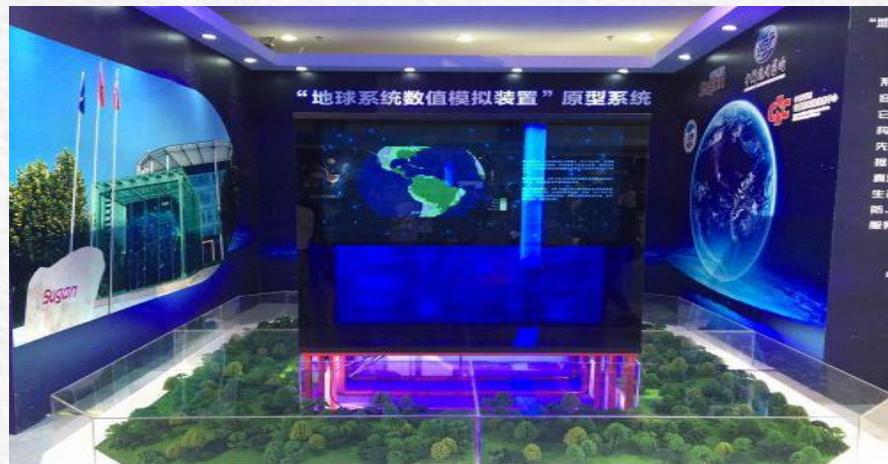
免疫组织穿梭（换个角度）

Liu T, Upadhyayula S, Milkie D E, et al. Observing the Cell in Its Native State: Imaging Subcellular Dynamics in Multicellular Organisms. Science, 2018.

可视化重要应用：大科学

“地球系统数值模拟装置”

国家十三五重大科学装置项目，
拟投资13亿，其中可视化部分
7000万。



国家卫星气象中心全球大气数据可视化平台
(浙江大学)

可视化重要应用：大工程

可视化是对大工程仿真、实测、融合、预测、测试等不同环节产生的信息进行综合理解与分析的**必要手段**



智能交通



智能电网



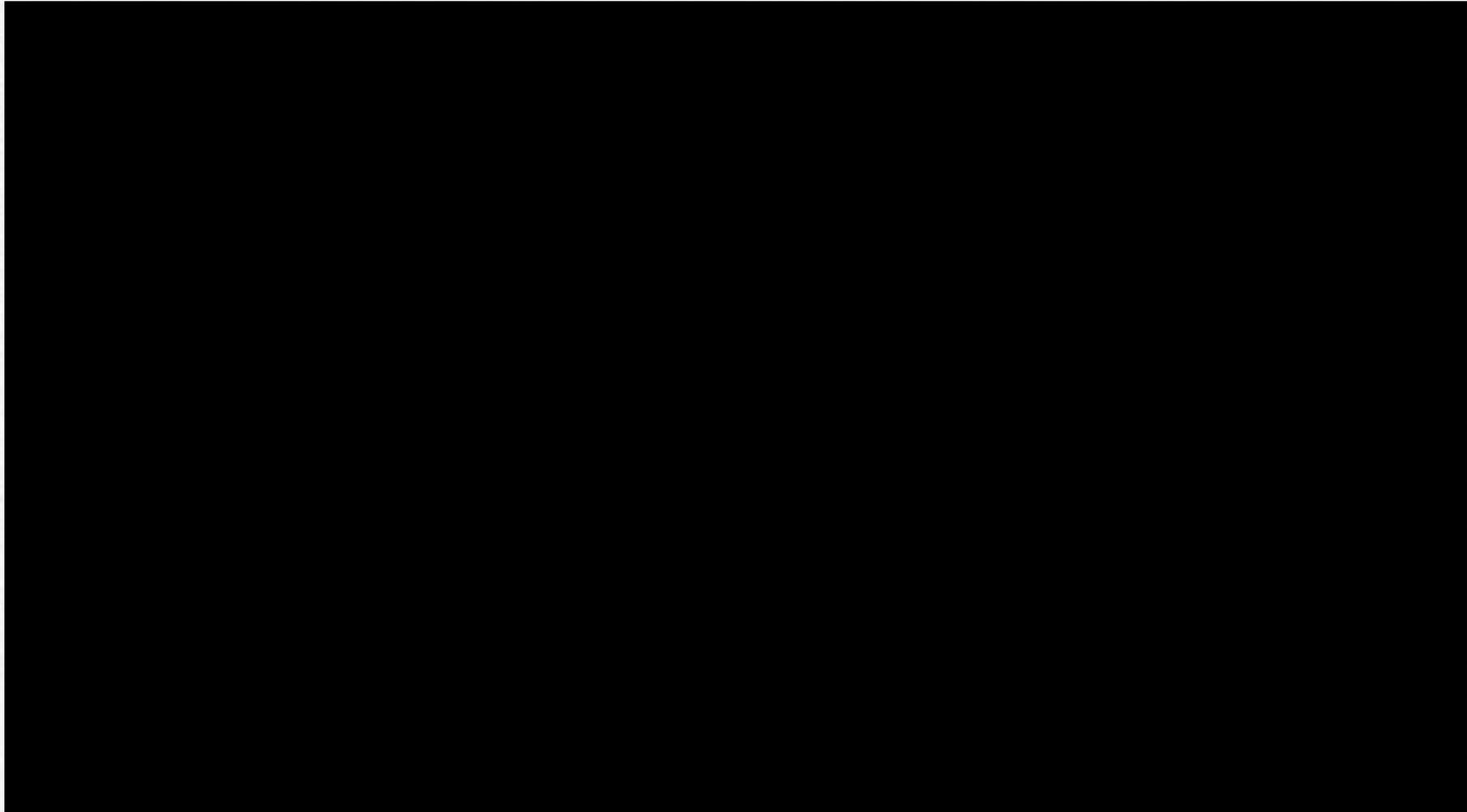
智能物流



智能制造



可视化重要应用：大工程



博世智能工厂数据可视化分析
(美国博世研究院, 浙江大学)

可视化重要应用：大安全

可视化是面向与人博弈任务的智能分析的最主要的交互界面



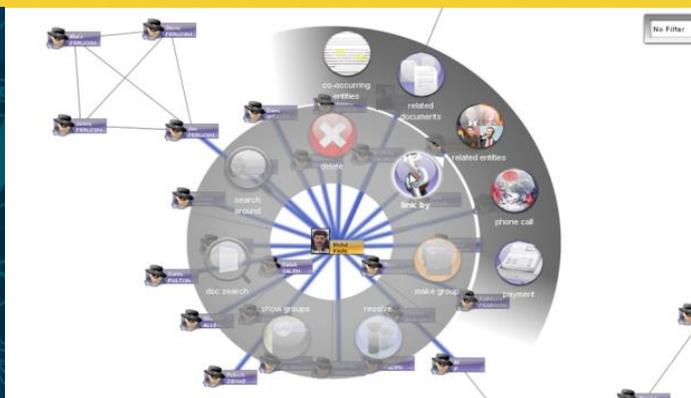
国土安全



网络安全



公共安全



金融安全



Condition "2"

Condition list

Source: [Icons]

ID = 杏花路百花苑

which when where what +

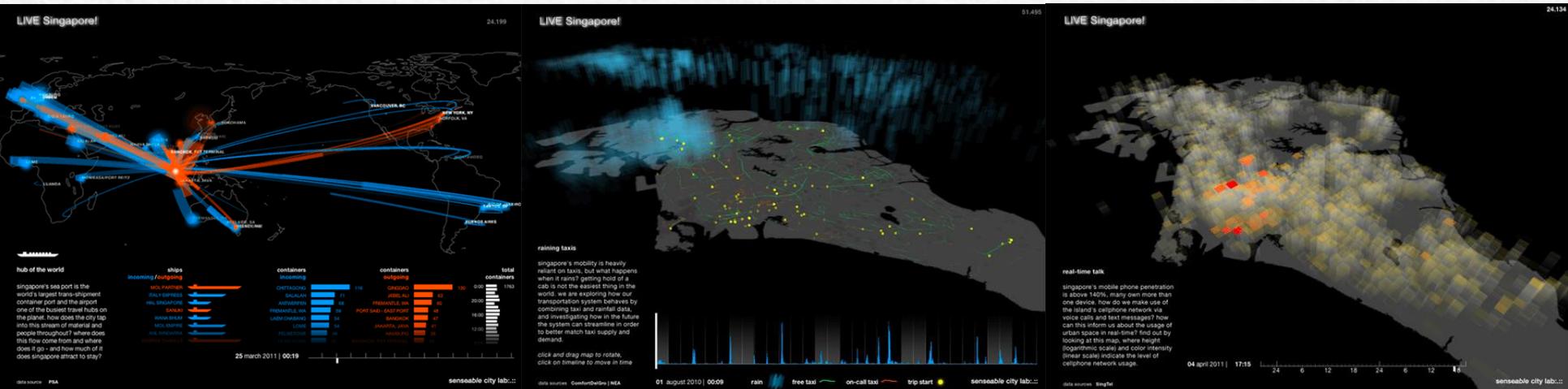
ID: 杏花路百花苑 OK

SEARCH

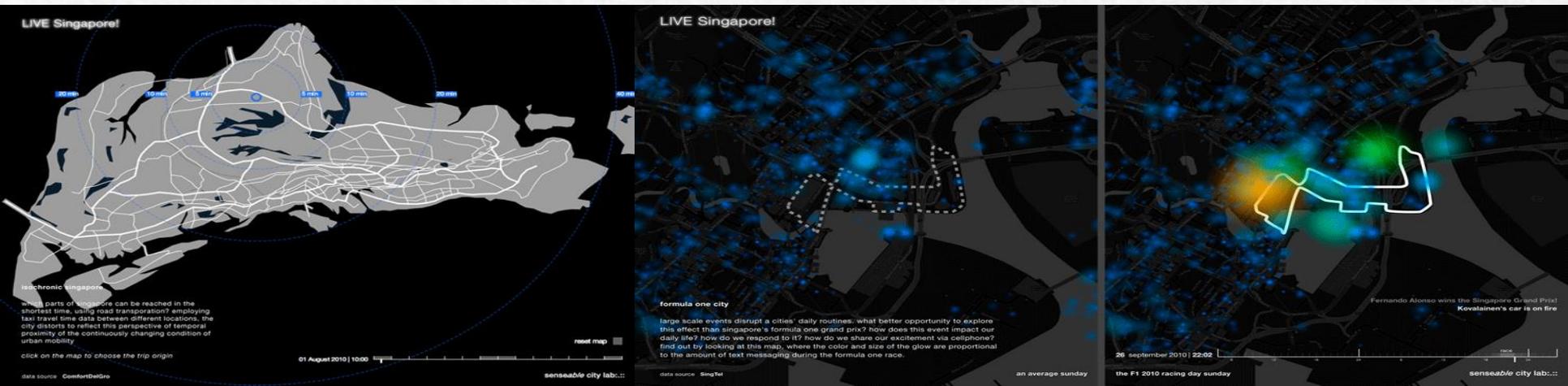
Manually extracting the names of two POIs from the post

基于定位、网络和社交数据的时空城市数据可视查询 (浙江大学)

可视化重要应用：物联网与智慧城市



可视化是基于CPS数据进行规划、理解、决策的敏捷分析途径



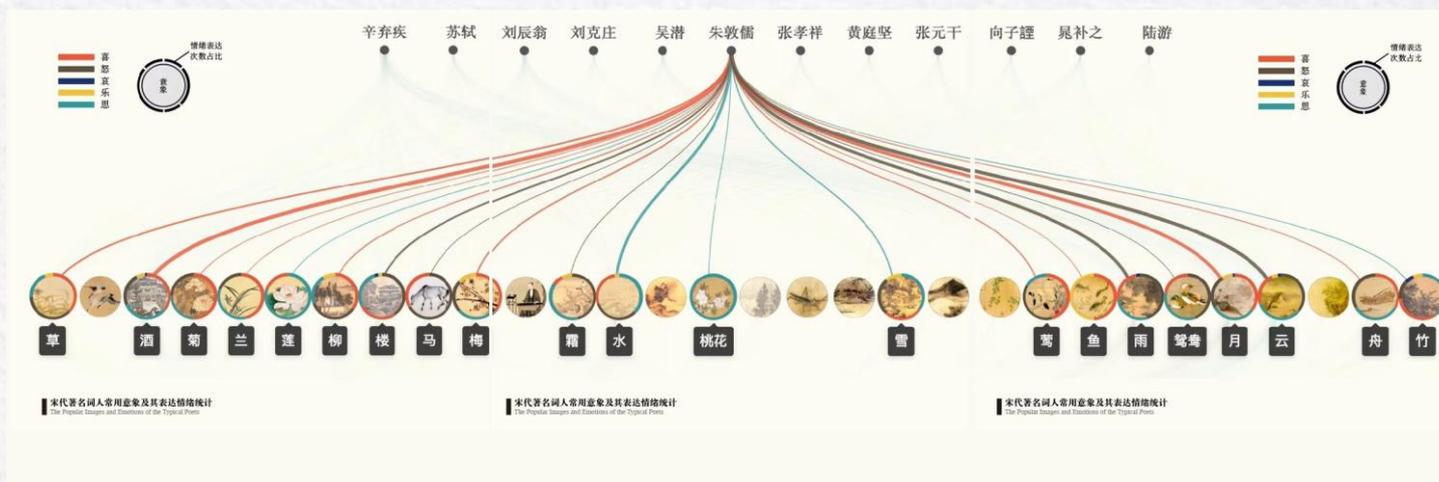
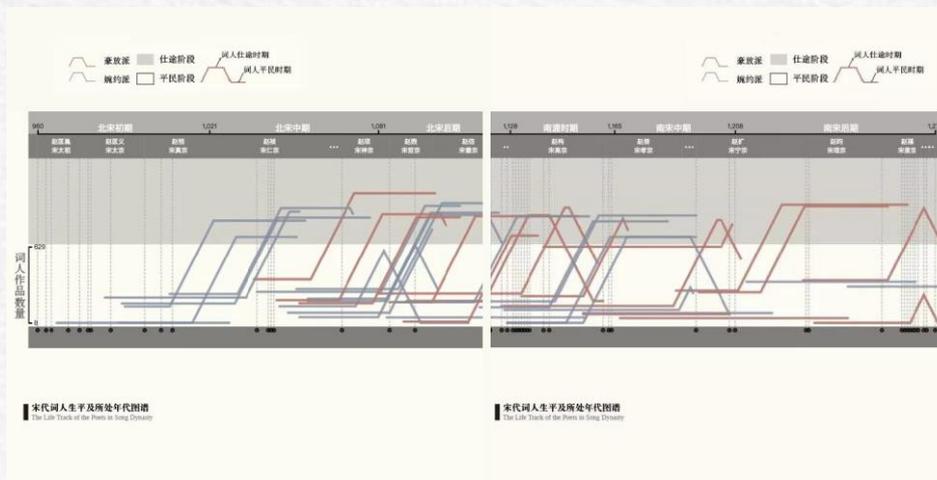
智能新闻与短视频（新华智云）



The screenshot shows an Excel spreadsheet with a table of data. The table has two columns: 'x1' and 'y1'. The data points are as follows:

x1	y1
1978	355
1979	454.6
1980	570
1981	735.3
1982	971.3
1983	1265.1
1984	1703
1985	2066.71
1986	2580.4
1987	3084.2
1988	3821.8
1989	4155.9
1990	5560.12
1991	7225.75
1992	9119.62
1993	11271.02
1994	20381.9
1995	23499.94
1996	24133.86
1997	26967.24
1998	26849.68
1999	29896.23
2000	39273.25
2001	42183.62
2002	53328.13

浙大-新华网制作的数据新闻：宋词文化可视化



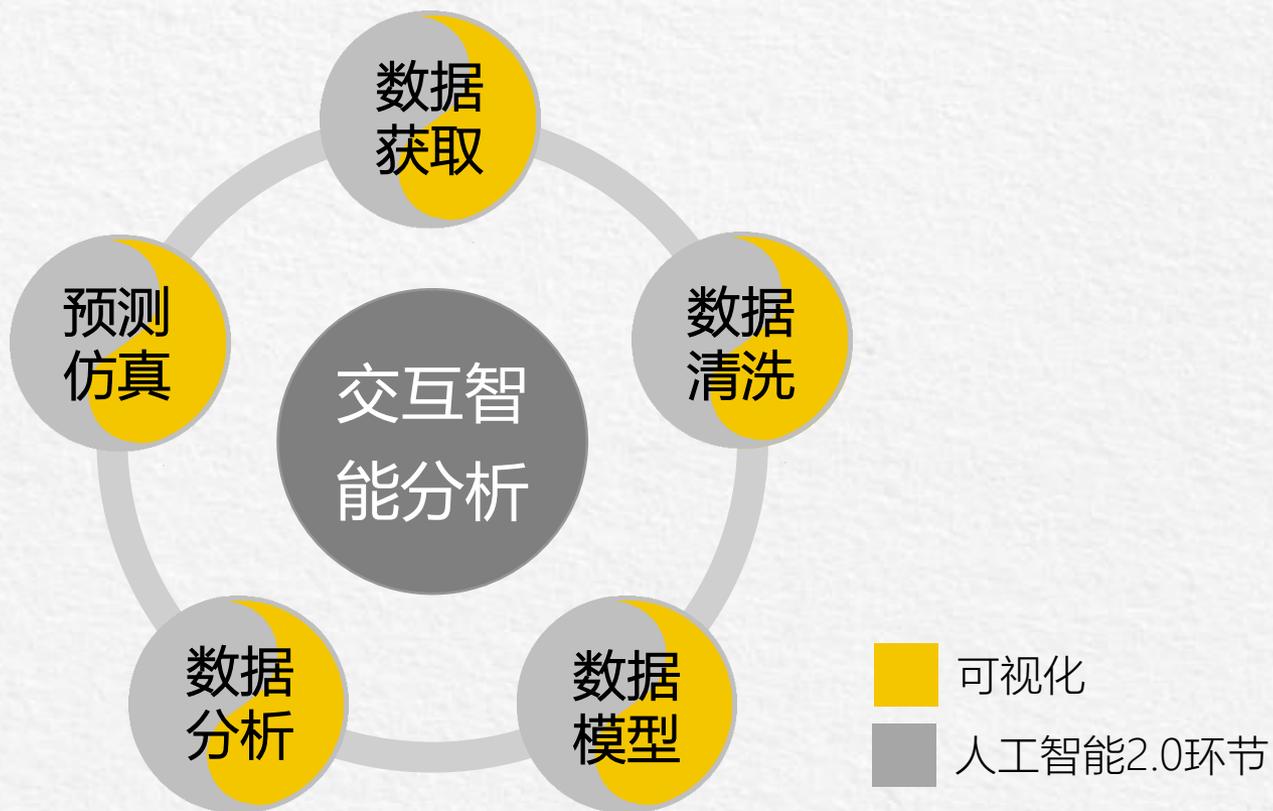
互动版 (pc端支持更多交互) : http://fms.news.cn/swf/2018_sjxw/quansongci/index.html#/
图文版链接 http://www.xinhuanet.com/video/sjxw/2018-09/05/c_129947285.htm
微博链接 <http://t.cn/Rsfnsnp>



大数据可视化 展望

智能交互 2.0

人工智能2.0的每个环节与**可视化**融合





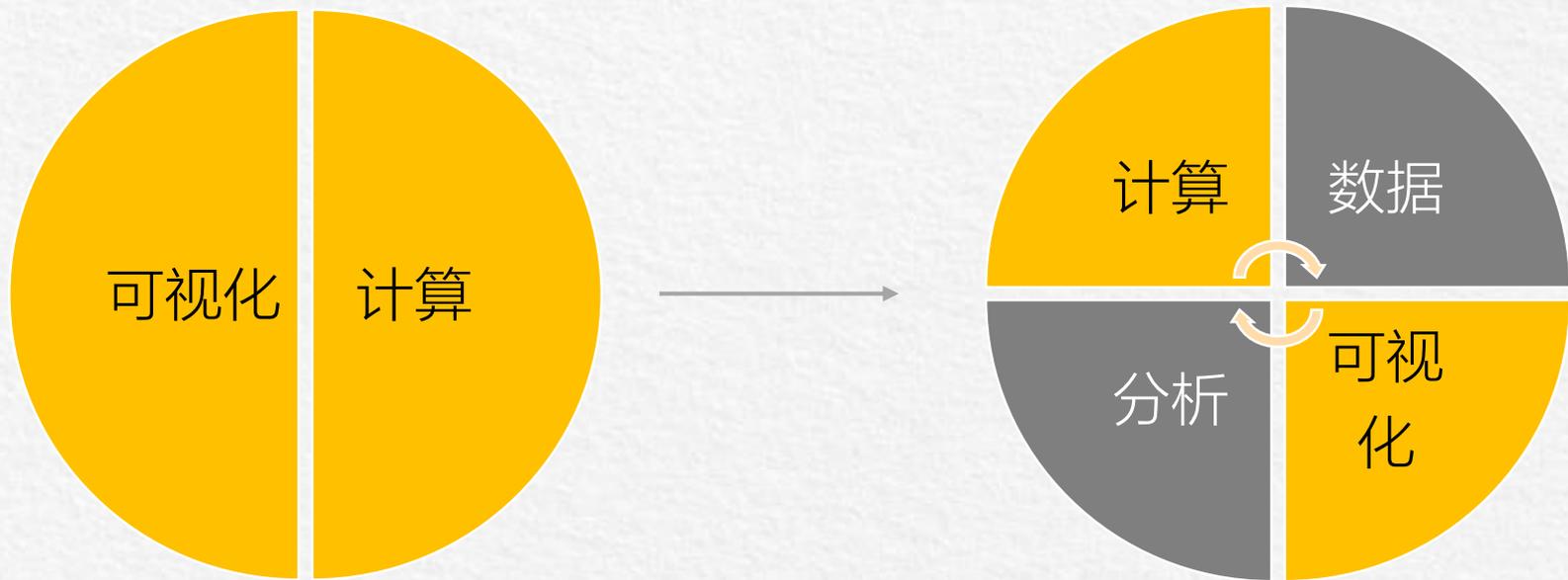
基于预计算的表格型数据快速查询

Scenario 3
Taxi trajectory dataset



面向科学研究的智能交互计算

科学研究中的**计算、数据、分析和人智**的交互融合



Usage Scenario

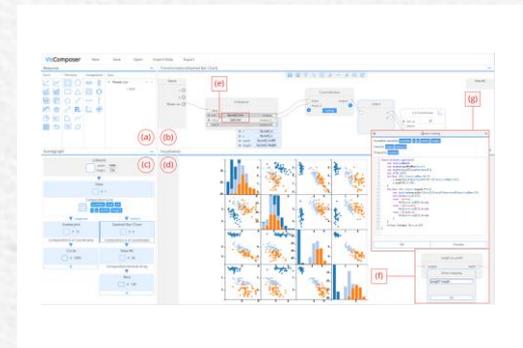
Dataset: Bitcoin trading networks(207689 nodes, 547500 links)

面向行业的通用可视化软件与组件库研发

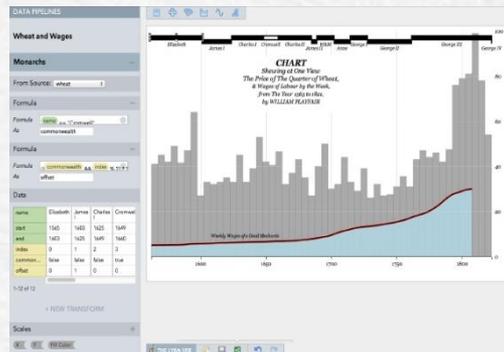
以数据模态或应用场景区分的可视化**标准件**与**通用系统软件**



软件



编程



交互开发环境



NBA比赛进程可视化

NBA Games Viewer

真实应用场景下的态势感知与临场决策（沉浸式）环境



大屏拼接沉浸式环境（浙江大学）



谢谢

陈为

chenwei@cad.zju.edu.cn

浙江大学CAD&CG国家重点实验室