西安电子科技大学
XIDIAN UNIVERSITY

大数据治理

马 晶

经济与管理学院 信息管理系

Email: majing@xidian.edu.cn

西安电子科技大学
XIDIAN UNIVERSITY

前言 课程简介

马 晶

经济与管理学院 信息管理系

Email: majing@xidian.edu.cn

课程性质

大数据治理（Big Data Governance ）课程是大数据管理与应用专业的专业核心课（必修），2学分，32学时。

课程的教学目标与任务

大数据治理是连接大数据科学和应用的桥梁，要实现大数据的变现，就离不开科学的大数据治理，离不开与时俱进的管理变革。

本课程是大数据管理与应用专业的核心理论课程，从数据架构管理、元数据管理、主数据管理、数据集成、数据质量管理、数据标准化、数据安全与伦理、法律保障体系8个方面对大数据治理进行系统、全面的介绍。课程采用丰富的案例，让学生直观感受相应理论的具体内涵。

通过本课程的学习旨在让学生了解大数据治理的必要性与意义，让学生对大数据治理的场景与任务有一个较为全面的认知，帮助学生理解大数据治理的框架、流程、技术和典型应用等相关基本知识，为进一步深入探索大数据领域奠定基础。

课程主要内容和教学安排

序号	课程内容	教学方式
1	大数据治理的背景和基本概念	讲授+讨论
2	大数据架构管理	讲授+讨论
3	元数据管理	讲授+讨论
4	主数据管理	讲授+讨论
5	数据集成	讲授+讨论
6	数据质量管理	讲授+讨论
7	数据标准化	讲授+讨论
8	数据资产化	讲授+讨论
9	数据安全与隐私保护	讲授+讨论
10	法律保障体系	讲授+讨论

教材及参考书目

教材：

《大数据治理》，顾东晓、刘鲁宁 主编，清华大学出版社，2023.9

参考书目：

1. 《大数据治理理论与方法》，王宏志、李默涵 编著，电子工业出版社，2021.10
2. 《一本书讲透数据治理》用友平台与数据智能团队，机械工业出版社
3. 《大数据治理》,桑尼尔 索雷斯 著，匡斌 译, 清华大学出版社
4. 《数据治理之论》,梅宏 主编 ,中国人民大学出版社

考核及成绩评定方式

1.平时作业成绩：20%，线上作业成绩。

线上作业：章节结束后布置线上作业，学生通过学在西电平台提供的教学资源对课程内容进行复习，并完成相应作业。

2.项目学习成绩：20%，组建5人项目小组，根据老师提供的学习内容要求，通过查阅文献、案例资料搜集，形成项目学习报告，并在课堂进行汇报。

3.期末考试成绩：60%，对学习情况进行全面检查，包括基本概念及知识点理解，方法应用和案例分析等。

西安电子科技大学
XIDIAN UNIVERSITY

第一章 大数据治理的背景和基本概念

马 晶

经济与管理学院 信息管理系

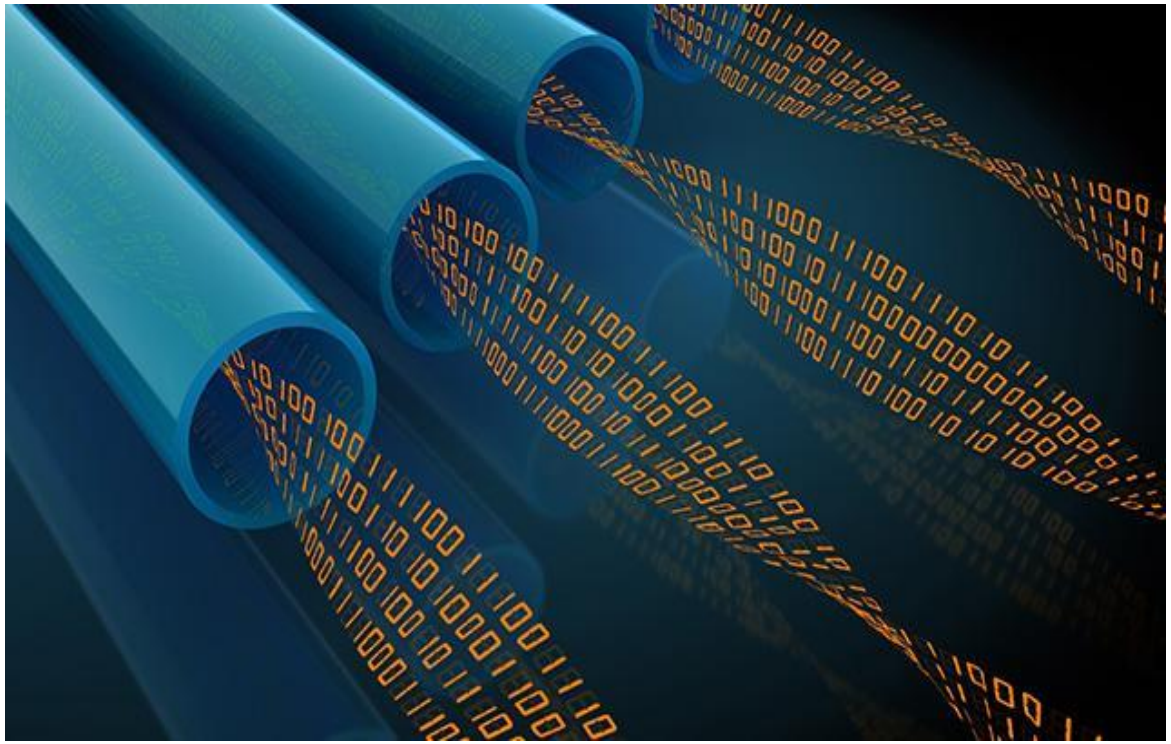
Email: majing@xidian.edu.cn

课前讨论：数据的价值

人类进入信息社会以后，数据以自然方式增长，其产生不以人的意志为转移。

从1986年开始到2010年的20余年时间里，全球数据的数量增长了100倍，今后的数据量增长速度将更快，我们正生活在一个**“数据爆炸”**的时代。

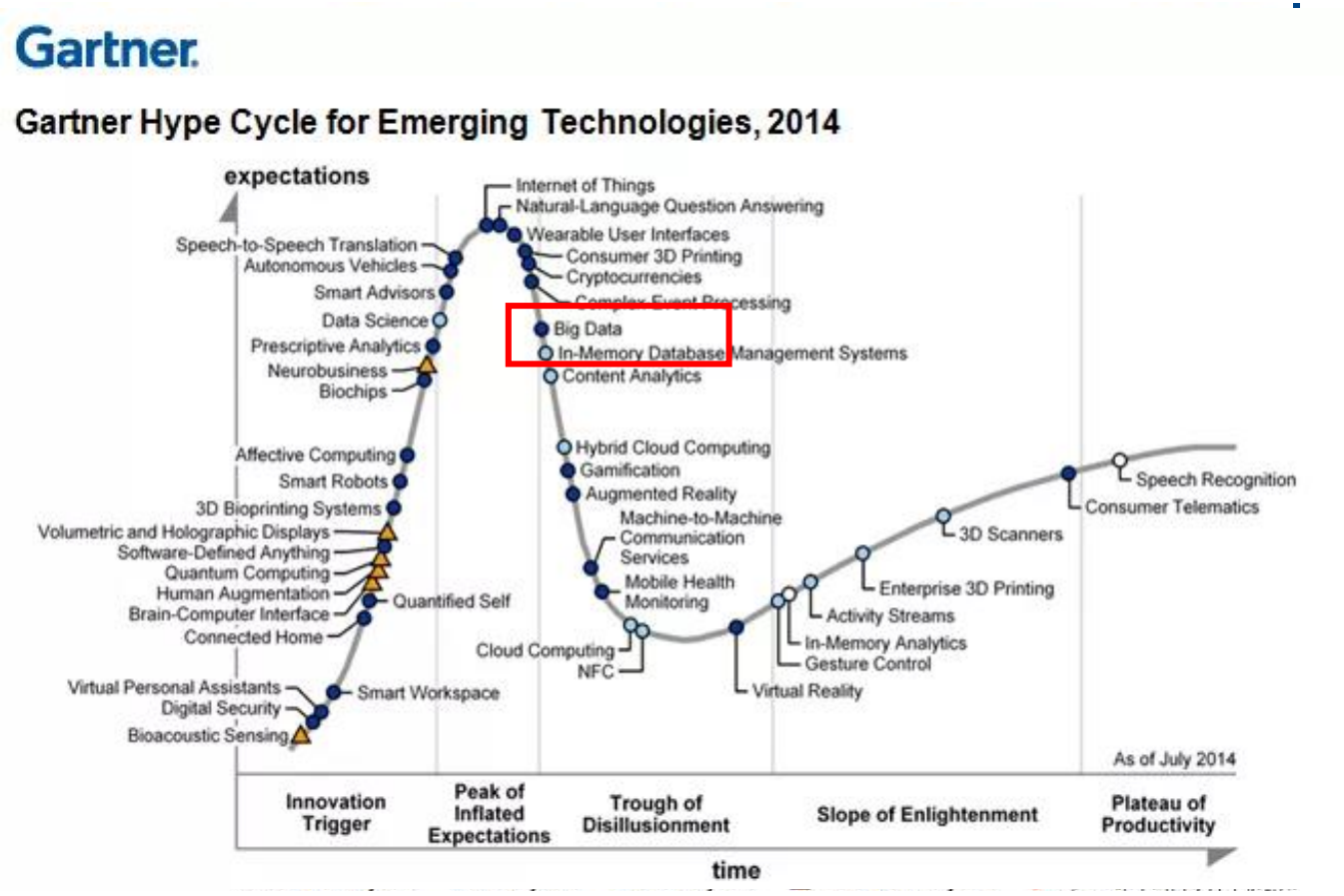
课前讨论：数据的价值



在大数据时代以前，最有价值的商品是石油，而今天和未来则是数据。目前占有大量数据的谷歌、亚马逊等全球前五大公司，每个季度的利润总和高达数十亿美元，并在继续快速增加，这都是数据价值的最好佐证。因此，要实现大数据时代思维方式的转变，就必须要正确认识数据的价值，数据已经具备了资本的属性，可以用来创造经济价值。

从Gartner技术成熟度曲线看大数据发展

Gartner Hype Cycle, 又被称为技术成熟度曲线，是 Gartner 对各种新技术或其他创新的典型发展过程的图形化的描述。Gartner 自 1995 年起，每年都针对各种技术和应用领域创建 90 多张技术成熟度曲线，用来帮助客户跟踪技术的成熟度和未来潜力。Gartner 将每项技术的发展过程分为五个阶段：创新萌发期/技术萌发期；顶峰期/过热期/期望峰值期；低谷期/幻灭期；爬升期/复苏期；稳定期/成熟期



Gartner 自 2011 年起开始将大数据“big data” 纳入 Emerging Technologies Hype Cycle 并进入到上升期/创新萌发期 (当时还被称为 “big data and extreme information processing and management”)

在 2012 年 “big data” 开始进入顶峰期
在 2013 年达到顶峰期的顶峰
在 2012 年和 2013 年，Gartner 还专门发布了针对大数据领域 hype cycle

并在 2014 年最后一次被收录 Emerging Technologies Hype Cycle 的顶峰期;
在此之后历年的 Emerging Technologies Hype Cycle 中就再也没有大数据 “big data”的身影了。

https://www.gartner.com/en/research/methodologies/gartner-hype-cycle

从Gartner技术成熟度曲线看大数据发展

Gartner 在 2014 年后不再将大数据收录到 Emerging Technologies Hype Cycle 中，也不再有专门针对大数据行业的 hype cycle, 并不能说明大数据的没落，恰恰相反，这正说明大数据已经快速步入到了稳定的成熟期，也已经渗透到了各行各业的方方面面，藏身于很多技术领域的 hype cycle 身影之后。

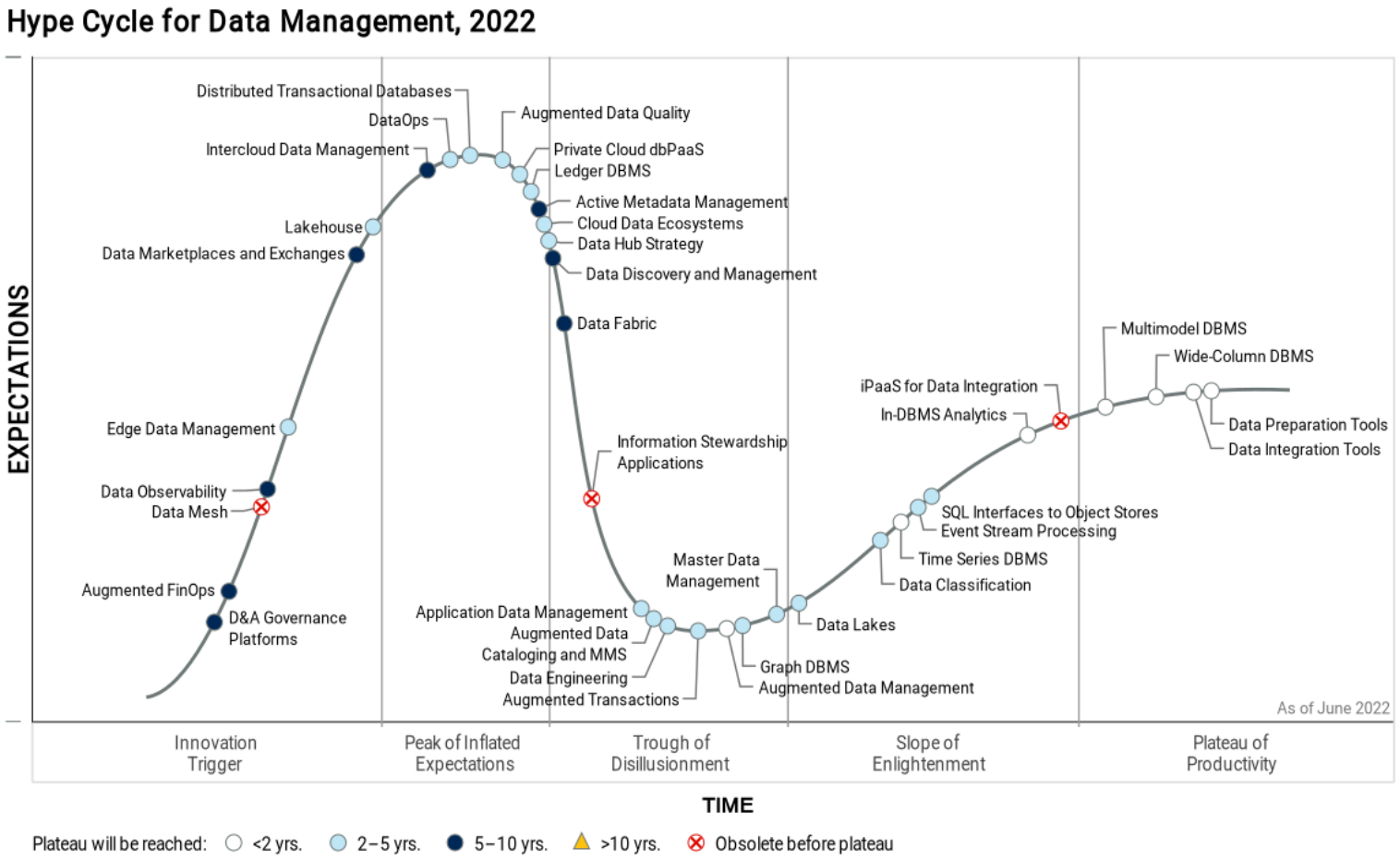
在 2020 年，Gartner 针对数据和分析领域，发布了以下多个 hype cycle，这背后其实都是大数据的身影:

- Hype Cycle for Enterprise Information Management
- Hype Cycle for Analytics and Business Intelligence
- Hype Cycle for Data Management
- Hype Cycle for Data Science and Machine Learning
- Hype Cycle for Artificial Intelligence
- Hype Cycle for the Internet of Things
- Hype Cycle for Data and Analytics Governance and Master Data Management
- Hype Cycle for Natural Language Technologies
- Hype Cycle for Data Security
- Hype Cycle for Customer Experience Analytics

“There’s a couple of really important changes, We’ve retired the big data hype cycle. I know some clients may be really surprised by that because the big data hype cycle was a really important one for many years. But what’s happening is that big data has quickly moved over the Peak of Inflated Expectations, and **has become prevalent in our lives across many hype cycles.** So big data has become a part of many hype cycles.”, "I would not consider big data to be an emerging technology, this hype cycle is very focused. I look at emerging trends.”

--Gartner分析师Betsy Burton

从Gartner技术成熟度曲线看大数据发展



- Hype Cycle for Data & Analytics Programs and Practices, 2022
- Hype Cycle for Data and Analytics Governance, 2022
- Hype Cycle for Analytics and Business Intelligence, 2022
- Hype Cycle for Data Science and Machine Learning, 2022
- Hype Cycle for Artificial Intelligence, 2022
- Hype Cycle for Data Security, 2022
- Hype Cycle for Privacy, 2022

数据资源与管理

数据：

数据是事实或观察的结果，是对客观事物的逻辑归纳，是用于表示客观事物的未经加工的原始素材。

数据资源：

数据资源是可供人类利用并**产生效益**的一切记录信息的总称，并属于一种社会资源。

数据资源管理：

数据资源管理是应用信息技术和软件工具完成组织数据资源管理采用文件处理方法，在这种方法中，数据根据特定的组织应用程序的处理要求被组织成特定的数据记录文件，只能以特定的方式进行访问。

数据资源与传统资源区别：

1. **无形性：**即非物质性和无形性使得数据资源被传统物权所排斥，因而无法成为传统物权的客体。基于此，数据资源可以被他人近乎零成本、快速地、无次数限制地复制，可以跨越时空限制而为社会公众所共享共用，且不会发生有形的损耗。

2. **可变性：**即数据资源形成和流通的过程意味着数据资源总是处于变化之中。数据流通过程中的每一个事物特征和活动状态也都可能形成新的数据资源。

3. **社会性：**即传统意义下的自然资源具有社会性，自然资源的开发利用及消耗最终追求的都是社会福利的增加。数据资源尽管也参与到了整个社会关系中，但数据资源的获取、处理及利用总是与对数据资源有需求的社会主体密切相关。当讨论数据资源归属时更多的是需要考虑数据资源的持有、使用和经营，而非所有。

4. **共享性：**数据资源在使用上具有非竞争性和非排他性，即额外用户使用它们的边际成本可以为零。从一定意义上，数据资源可以说是一种公共物品，其本质就是分享与流通。

案例：关于数据治理的讨论

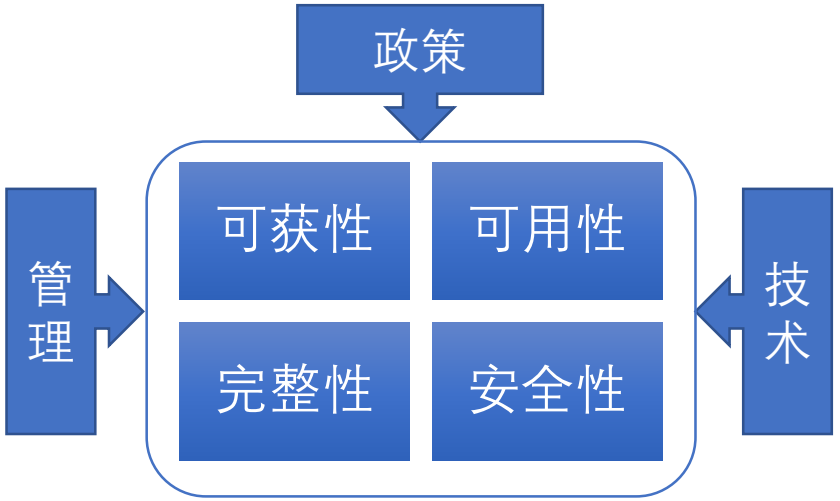
某汽车零配件制造公司正在召开有关公司数字化转型的动员会，CEO、信息技术部门员工A和B，财务以及生产部门员工代表参加了此次会议。会上他们就公司的数据治理问题阐述了自己的看法。。。

- 请抽签决定参会者身份。
- 会议有CEO主持，各代表依次发言，阐述自己的观点。
- 会议围绕与会代表的观点展开讨论，你需要尽可能争取其他代表的支持和理解。
- 会议结束后，围绕以下内容总结会议要点。

- 数据治理的任务可能包括哪些？
- 数据治理过程中涉及哪些角色和参与者？
- 过程中可能存在哪些难点？

第一章 大数据治理的背景和基本概念

- **狭义的数据治理**指对数据质量的管理、专注于对数据本身的分析，包括数据资源及其应用过程中相关管控活动、绩效和风险管理的集合，以保证数据资产的高质量、安全及持续改进。
- **广义的数据治理**是指对数据资产管理行使权力和控制的活动集合（规划、监控和执行），指导其他数据管理职能如何执行，在高层次上执行数据管理制度。
- 在大数据战略从顶层设计到底层实现的过程中，治理是基础，技术是承载，分析是手段，应用是目的。
- 大数据治理涉及组织、行业、国家三个层面，在这三个层面定义、构建一套完整的体系，不仅需要成熟的模型和算法，还需要完善的法律法规、全面的标准体系等。
- 所需的技术支撑涵盖大数据管理、存储、质量、共享与开放、安全与隐私保护等多个方面。



上述定义包括以下内涵：

- **大数据治理是广义信息治理计划的一部分**。
- **大数据治理关乎政策制定**。这里的政策是指人们在特定情形下采取的措施。如大数据治理政策可能申明：“未经顾客知情并同意，组织不得将顾客的Facebook资料整合到其主数据记录中”。
- **大数据必须优化**。与企业对实物资产的优化管理类似，组织必须对大数据进行优化，包括元数据管理、数据质量管理、信息生命周期管理等。
- **大数据必须变现**。变现的方式既可以是直接将数据卖给第三方，也可以是利用数据开发新的服务。
- **大数据的安全隐私至关重要**。在处理社交媒体、地理定位、生物计量学和其他形式的个人可识别信息（Personally Identifiable Information，PII）时，组织必须制定适当的政策，以防止大数据误用带来的声誉、法律等方面的各种风险。
- **大数据治理必须对各种冲突进行协调**。基于不同目标，大数据往往会带来多种冲突，如客户隐私与企业利益之间的冲突、计算代价和服务质量之间的冲突等。

案例：公路、铁路和航线上的交通堵塞，浪费时间、增加污染，造成极高的社会成本

- **通过大数据治理，如何提高运营效率，降低运营成本？**

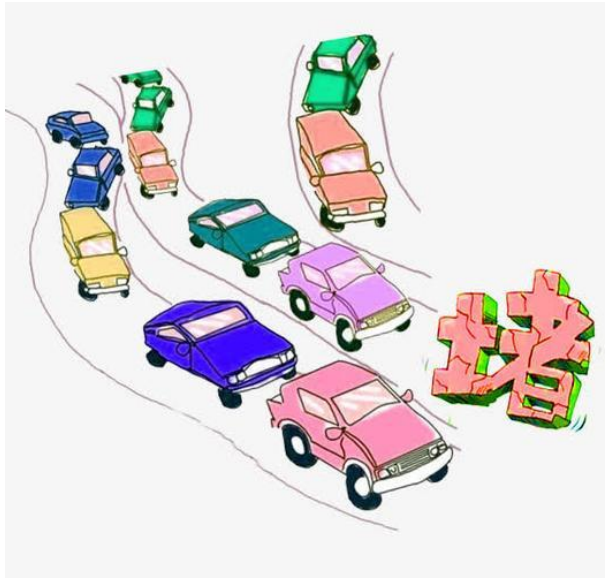
从整个欧洲来看，基础设施的交通堵塞给国内生产总值带来近1%的耗损。在美国，单单航班延误一项，每年就会造成大概约60亿美金的损失。2013年，麦肯锡全球研究所公布的一项研究报告表明，通过改良需求管理和维护以增强对现有基础设施的利用，全球每年可节省4000亿美元。

- **巴西利用GPS数据优化可用空域使用率系统**

在巴西，航空流量在过去的10年中增长迅速，预计到2030年每年乘客数量将翻番，达到3.1亿多人次。航空拥堵问题日益受到关注也就是顺理成章的事情了。为了应对该问题，**巴西开始引进一套利用GPS数据优化可用空域使用率的系统，减少飞机之间的距离以及缩短航线。**

常规的方法是准备降落的飞机在空中排成一队。使用新系统后，每架飞机将拥有自己的航道。这听起来可能简单，但是却需要大量的数据作为支撑，以及对这些数据快速而复杂的测评。每架飞机的距离、速度和载荷能力数据被加以处理使得航线最短。飞机可以在距离机场近得多的地方「曲线」着陆，而不是在即将着陆的时候在空中排队。

巴西利亚国际机场首次采用了这套系统，每次着陆节省了7分半钟和约300升燃油，同时每架飞机平均少飞了约40千米。巴西计划在全国的10家最繁忙的机场部署这套系统。据初步估计在北美机场部署这套系统可以使机场的运力提升16%到59%，具体视机场实际条件将有所不同。



案例：尽管有这样那样振奋人心的将信息和基础设施加以整合的范例，但总体进展却往往不尽如人意

信息透明度

- 交通基础设施涵盖了复杂的网络和众多的参与者。拿机场举例，不同的航空公司、地勤公司和零售商，再加上航空管制海关和机场自身。每一方都在收集各自的数据，且未必愿意拿出来共享。但是提升追踪乘客的能力可以让每个人都受益。举例来说，如果知道了客流量的分布情况和移动情况，就可以优化登机口排布和资产配置。这不止可以提高机场的运力，还可以提高零售收入。而实现的前提条件，就是所有的数据要整合到一起。

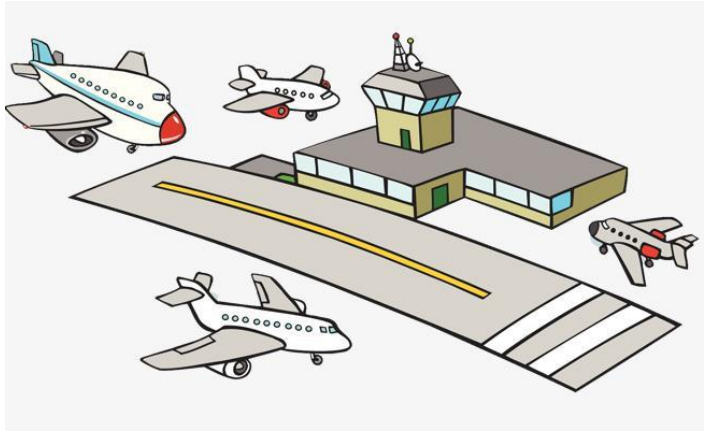
成本和收益分摊

- 航空公司需要的是更短的中转时间以最小化乘客的在途时间，但是零售商则希望旅客多逗留一会来提高店铺销售。机场可能比较倾向于资产的高利用率，但是他们也可能愿意牺牲一些利用率以换取灵活性，以便有什么突发事件后能够快速恢复原状。综合这些想找到一个解决办法，使得所有利益相关者都成为赢家，并不是一个简单的任务，这需要一定程度的相互信任。

监管

- 很多时候基础设施是天然的垄断行业。因此政府承担着重要的职责，确保运作的公平与有效，同时创造出一种监管环境，允许数据收集和使用的同时保护机密和隐私。但是在此之前，必须让竞争监管机构和数据保护机构确信数字化的益处。一个巨大挑战是，通过清楚地说明什么样的数据会被采集，数据将如何使用，从数据洞察中得出的解决方案将会最终给消费者带来什么好处，来打消对用户隐私的担忧。

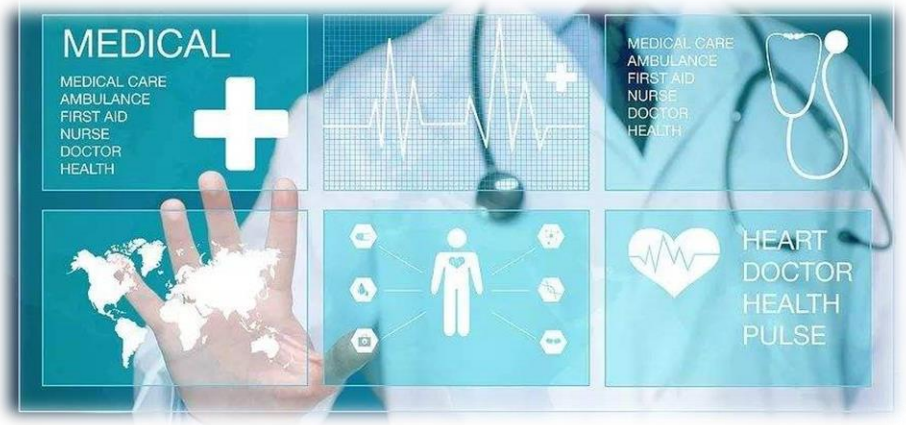
在机场行业会议上，人们对利用追踪乘客的移动设备获取的大数据来提供定制信息和管理的热情不减。创意包括，**综合考虑乘客步行速度，以短信通知乘客何时登机；基于更好的短期需求预测以及定制化购物建议缩短安检队伍。**



案例：基于开放架构与国际标准的大型智慧医院建设

传统医院“蜘蛛网”集成模式存在弊端

- 烟囱工程众多
- 临时建立的系统集成就如同“蜘蛛网”，接口错综复杂，信息交换规范不统一，数据交换不畅通，系统间存在信息孤岛、版本更新影响范围广、成本大、危险性高。
- 从长远发展来讲，系统间接口管理混乱及难以复用，医院被“蜘蛛网”所绑架，会带来高昂的维护成本问题。基于这样的情况，医院需要引入统一的信息平台技术来取代传统的集成模式，实现有效的信息交换和共享，让数据“活”起来，达到通过数据利用提高运营管理水平，提高诊疗效率和患者的就医体验。



北大国际医院系统在开业前整体平移完成，5大领域，214个系统一次点火上线成功，保证了开院时所有系统运行正常、业务操作顺利完成。目前业务运行稳定，信息化部门实现可视化管理，在整个运营的过程中大幅度降低了成本。截止目前为止，北大国际医院信息平台的每日交互数据量达到12万条/日。

在建设信息集成平台项目时，通过投入70%的精力在业务的梳理、业务流程的梳理形成了二十几个标准化的业务流程，形成了标准的数据字典。对业务模型高度的抽象化、概念化，形成了数据模型并建设了CDR临床数据中心。

采用了国际HL7 V3信息交换标准，构建以临床数据中心（CDR）为核心的大型医院信息集成平台，实现来自不同厂商的HIS、LIS、RIS、CIS、ERP等的数据整合、信息共享、流程协同，并同步推进临床信息化建设，通过电子病历浏览器和医生门户实现全流程患者信息实时同步共享。北大国际医院通过集成平台整合各个业务系统，实现了医疗行为的闭环管理。

大数据治理	数据架构管理	架构设计、实现、修正
	元数据管理	业务元数据、技术元数据
	主数据管理	架构设计、管理模式、使用模式、实现风格等
	数据集成	传统数据集成、跨界数据集成等
	数据质量管理	缺失值填充、实体识别、错误检测与修复等
	数据标准化	国际标准、国家标准、行业标准等
	数据资产化	发现与评估、交易与定价
	数据安全与隐私保护	安全存储、传输、访问、检索、处理、隐私保护

大数据治理框架

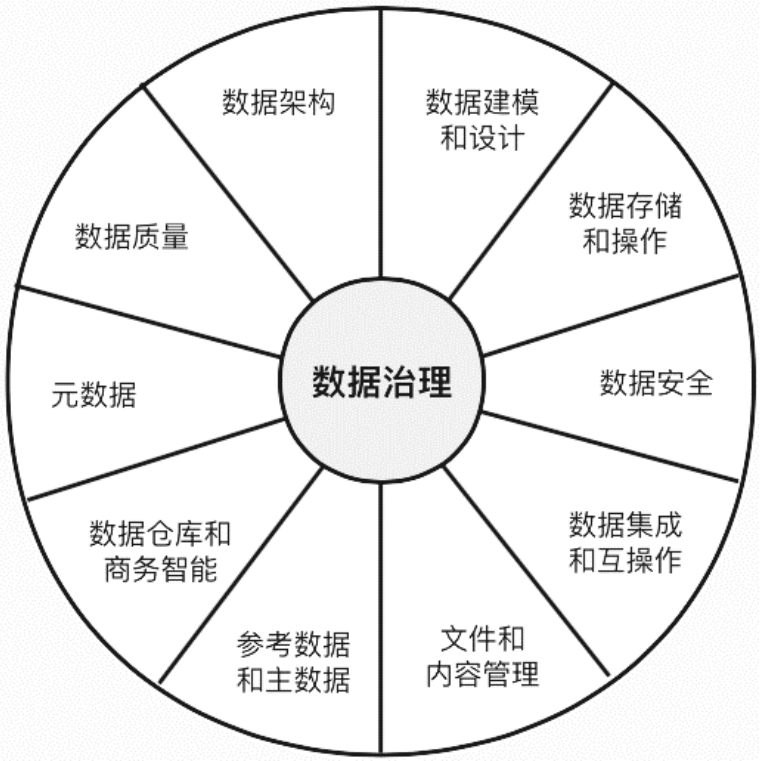
随着人类社会数字化水平的提升，数据规模不断跃迁，使得数据治理的难度不断加大。为有效的进行数据治理，国内众多组织基于数据治理理论和实践，制定了多种数据治理框架和标准，对于组织数据治理体系的建设和数据治理实践有着重要的参考意义，包括：

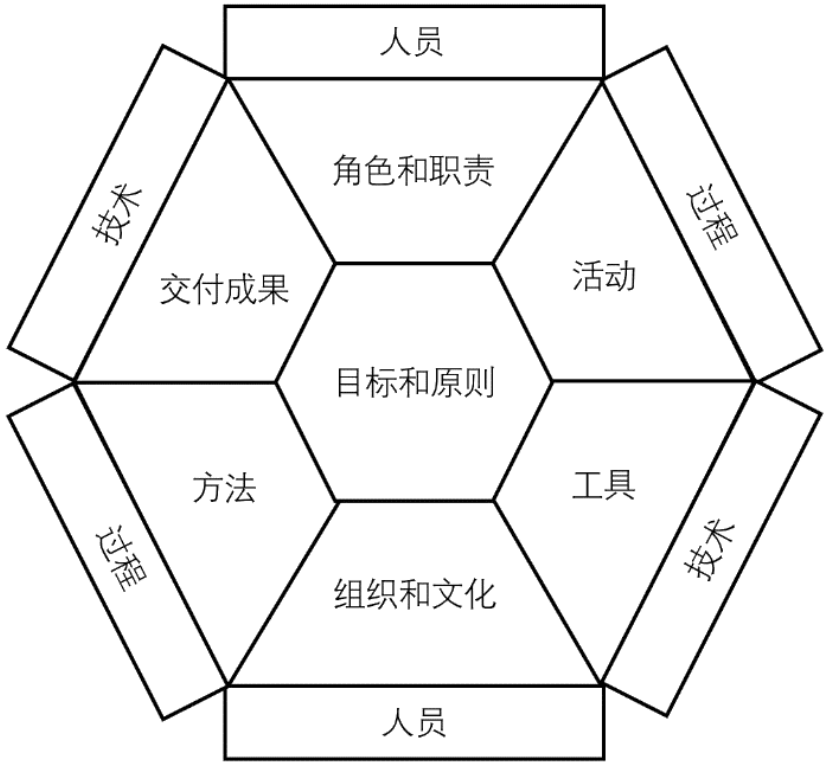
国际：ISO数据治理标准；2. DGI数据治理框架；3. DAMA数据管理框架

国内： GB/T34960规定的数据治理规范；数据管理能力成熟度评估模型

DAMA（国际数据管理协会）数据管理框架

DAMA-DMBOK2理论框架由11个数据管理职能领域和环境因素共同构成“DAMA数据管理知识体系”。每项数据职能领域都在7个基本环境要素约束下开展工作，按照一定的逻辑结构进行分析，保证数据治理的目标和实际商业过程的贡献。

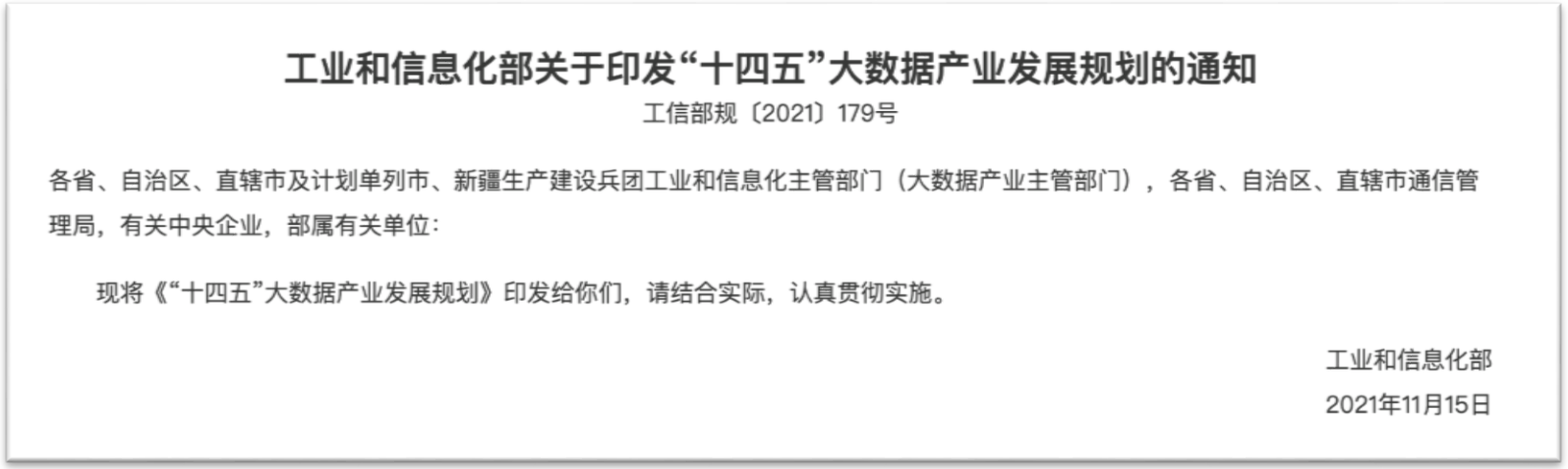




第一章 大数据治理的背景和基本概念

大数据治理体系是涉及国家实施大数据战略的重要基础和保障，也是发挥大数据作用、做大做强大数据产业的重要因素。

- 数据是新时代重要的生产要素，是国家基础性战略资源。
- “十三五”时期，我国大数据产业快速起步。据测算，产业规模年均复合增长率超过 30%，2020 年超过1 万亿元，发展取得显著成效，逐渐成为支撑我国经济社会发展的优势产业。
- 到 2025 年，大数据产业测算规模突破 3 万亿元，年均复合增长率保持在25%左右，创新力强、附加值高、自主可控的现代化大数据产业体系基本形成。



工业和信息化部关于印发“十四五”大数据产业发展规划的通知

工信部规〔2021〕179号

各省、自治区、直辖市及计划单列市、新疆生产建设兵团工业和信息化主管部门（大数据产业主管部门），各省、自治区、直辖市通信管理局，有关中央企业，部属有关单位：

现将《“十四五”大数据产业发展规划》印发给你们，请结合实际，认真贯彻实施。

工业和信息化部

2021年11月15日

主要任务：

1. 加快培育数据要素市场
2. 发挥大数据特性优势
3. 夯实产业发展基础
4. 构建稳定高效产业链
5. 打造繁荣有序产业生态
6. 筑牢数据安全保障防线

当下面临的挑战

1. **政策/流程** 大数据处理流程复杂，每个过程的问题都有可能影响大数据的应用，因而大数据治理应覆盖大数据的获取、处理、存储、安全等各个环节；
2. **数据管理专员制度** 大数据成为企业的重要战略资源，因而，需要在企业中为大数据设置数据管理专员；
3. **数据生命周期管理** 大数据的有效使用需要对数据的全生命周期进行管理，包括存储、保留、归档、处置等步骤，在数据生命周期管理的过程中需要有效平衡时间与存储空间；
4. **数据架构设计** 数据作为重要的资源和服务，必须要配以精心设计的数据架构，数据架构作为整个大数据治理的骨架，在保证治理任务顺利实施中扮演着重要的角色；
5. **元数据管理** 大数据需要与内容相关的元数据，需与传统数据定义标准保持一致；术语字典应包含大数据的术语；需要为非结构化数据提供分类、语义支持，Hadoop、NoSQL数据库等面向大数据技术的元数据需要纳入元数据存储库管理；
6. **主数据管理** 主数据是所有数据中最具价值的且被多个部门反复使用的数据，它们是公司的基本业务数据，良好的主数据管理可以为企业或组织节省大量数据整理的时间，并且能够提高数据质量；

当下面临的挑战

7. **数据集成** 大数据时代，很难保证企业或组织只需要处理单一来源的数据，大数据多源异构的特点使得其需要进行有效集成才能够得以协同工作，而大数据集成需要统一元数据标准，对大数据做统一定义；
8. **数据质量** 大数据规模大、变化快、多源异构等特点导致其有更大可能存在数据质量问题，因此应识别对业务有关键影响的数据元素，检查和保证数据质量；
9. **数据标准化** 在大数据时代，数据的交换、传递、共享非常重要，数据标准化能够使各个应用系统对客观实体的分类和描述手段一致，或者提供相应的转换接口。为了能够更好地建立良好的大数据共享与开放环境，数据标准化势在必行。
10. **数据资产化** 数据作为一种新型资产，如果不能被良好管理，也可能变成一种“负债”。如何管理数据资产，“盘活”数据以充分释放其附加价值非常重要。
11. **安全和隐私** 作为以互联网为依托的大数据，它将面临着网络带来的各种安全风险，这些风险威胁到大数据的安全，并可能给用户造成利益损失。在科学研究、产品开发、数据公开的过程中，算法需要收集、使用用户数据，因此，一些个人或敏感数据数据就不可避免的面临安全风险。如何保护数据安全和用户隐私是顺利实施大数据治理的最基础的问题之一。

讨论：身边的大数据治理问题

自“通信行程卡”上线应用以来，中国信息通信研究院（以下简称“中国信通院”）按照《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等法律法规有关规定，采取严格的技术和管理措施，依法依规保障个人信息安全。在提供行程卡服务过程中，中国信通院不存储用户个人行程数据，在此期间产生的运维数据滚动删除、销毁。

根据国务院联防联控机制综合组有关要求，2022年12月13日0时起，“通信行程卡”服务正式下线，中国信通院已按照有关法律法规规定，同步删除了行程卡相关所有数据，切实保障个人信息安全。

中国信通院

2022年12月13日

本章要点

- 数据资源与传统资源的区别
- 大数据治理的概念内涵
- 大数据治理的任务框架