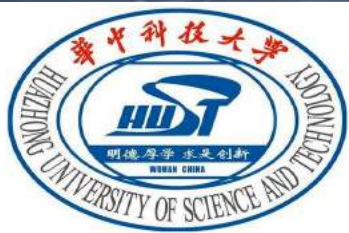


第六章 大数据治理



肖江

Mail : jiangxiao@hust.edu.cn

Office: 东五楼 222 室



目录

6.1 引言

6.2 大数据治理基本概念

6.3 数据架构管理

6.4 元数据管理

6.5 主数据管理

6.6 数据质量管理


6.7 数据标准化

6.8 数据资产化

6.9 本章小结

6.1 引言

在实际应用中常见的状况是大数据分析模型建立的非常完美，算法设计的也很漂亮，但在应用到实际的数据上并没有得到预想的效果。



**如何对大数据
进行有效治理？**

现实中的数据常来自于独立自治的数据源，缺少面向具体应用的顶层设计，缺少质量保障等机制问题。

6.1 引言

组织、行业、国家三个层面

在这三个层面定义构建一个完整的曲线

仍需要完善的法律法规，全面的标准体系支撑

治理是基础，技术是承载，分析是手段，应用是目的

目录

6.1 引言

6.2 大数据治理基本概念

6.3 数据架构管理

6.4 元数据管理

6.5 主数据管理

6.6 数据质量管理

6.7 数据标准化

6.8 数据资产化

6.9 本章小结

6.2 大数据治理基本概念

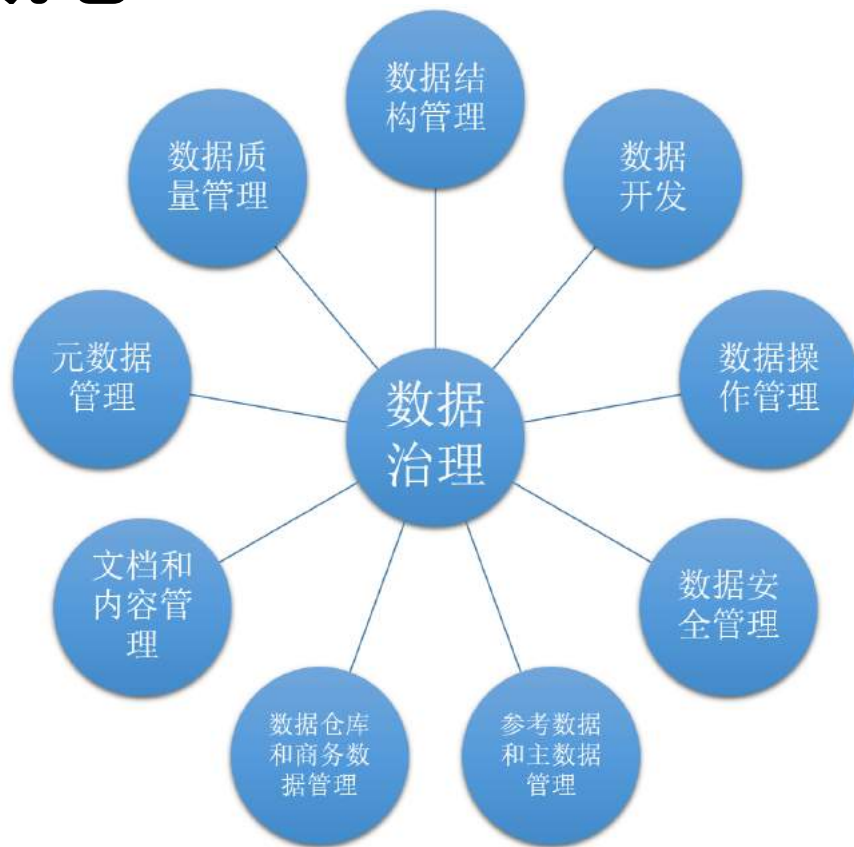
■ “治理” 源于拉丁语 “掌舵” 一词

- **宏观层面（体系框架角度）**：大数据治理是对组织的大数据管理和利用进行评估、指导和监督的体系框架，通过制定战略方针、建立组织架构、明确职责分工等，实现大数据的风险可控、安全合规、绩效提升和价值创造，并提供不断创新的大数据服务。
- **中观层面（信息治理计划和策略角度）**：大数据治理是广义信息治理计划的一部分，即制定与大数据有关的数据优化、隐私保护与数据变现政策。
- **从部署与管理角度定义**：大数据治理是企业数据可获性、可用性、完整性和安全性的部署及全面管理；在微观层从策略或程序角度定义，大数据治理是描述数据该如何在其全生命周期内有用和经济管理的组织策略或程序。



6.2 大数据治理基本概念

- **数据治理是数据管理框架的核心职能，指导其他数据管理职能如何执行。**
- **数据治理通过制定正确的政策、操作规程，确保以正确的方式对数据和信息进行管理。**



6.2 大数据治理基本概念

- 大数据治理的定义：

大数据治理是广义信息治理计划的一部分，即制定与大数据有关的数据优化、隐私保护与数据变现的政策。

6.2 大数据治理基本概念

• 大数据治理定义的内涵

大数据治理是**广义信息治理计划**的一部分。信息治理机构必须将大数据整合到既有的信息治理框架中。

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。

6.2 大数据治理基本概念

• 大数据治理定义的内涵

大数据隐私至关重要。

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时，组织必须制定适当政策，以防止大数据误用带来各种风险，包括声誉法律等。

大数据必须变现。

变现的方式可以是将数据卖给第三方，也可利用数据开发新的服务。

大数据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等。

案例1 通过大数据治理，提高运营实时性和旅客安全度

• 某高铁公司先进的状态监测实践

某高铁公司连接台湾北部的台北和台湾南部的高雄，全程214英里，用时仅90分钟。列车运行时速186英里，穿越50个隧道和大段高架桥，部分线路穿过地震多发地带。

这家公司部署传感器，采集超过32万个数据元素，包括车轮的转速和温度，以及架空线的密度。以火车轮轨为例，状态数据可以通过无线方式实时传送到中央数据库，与正常技术参数进行比较。在发生刹车磨损过皮等偏离特定参数的事件时，车载监测系统会自动将警报发送到维修系统，维修系统自动发出合适的**定期检修**、**校正维修**或**应急修复**等请求。

案例1 通过大数据治理，提高运营实时性和旅客安全度

• 某高铁公司先进的状态监测实践

- 这家公司从该计划中受益匪浅：
 - A. 运营实时性。高铁系统日均旅客吞吐量超过9万，预定计划6秒之内的到站和出站时间准确率，控制在99.15%。
 - B. 旅客安全度。通过监测各种各样的变量，高铁公司可以将旅客安全度提升到一个全新水平。相比之下，到目前为止，德国艾雪德列车出轨事件，仍是全球死亡率最高的铁路事故。该事故发生于1998年6月3日的德国艾雪镇，造成101人丧生，88人受伤。事故的起因是，一个轮子出现疲劳裂纹，导致列车在岔道口出轨。



案例2 评估数据质量和主数据对大数据计划的影响

• 某品牌零售商的大数据分析计划

受此起彼伏的促销活动影响，某大名鼎鼎的全球零售商正面临产品利润下滑的困境。为迎接这一商业挑战，该公司决定采集并分析Twitter和其他网站中客户对产品的反馈信息，以确定新产品的定价战略。如果在产品发布期间，大数据情感分析为非正面，那该公司就决定在主产品目录中及时更新定价，并提供30%的折扣，从而取代在季末以70%的折扣出售这些产品的通常做法。借此，该零售商提高了产品利润率。值得注意的是，该零售商使用了具有标准产品定义和层次的高质量主数据，以支撑利润提升计划。

6.2.2 大数据治理的挑战

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、安全等环节
数据管理专员制度	大数据成为企业的重要战略资源，需要在企业中为大数据设置数据管理专员
数据集成	由于大数据多源异构的特点，需要进行有效集成才能够得以协同工作，大数据集成，需要统一元数据标准，对大数据做统一定义
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
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。
元数据和数据定义	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。
隐私	由于大数据多源异构的特点，应为隐私保护需求，制定相应政策。
风险	大数据治理与内外部风险管控需求建立联系。

目录

6.1 引言

6.2 大数据治理基本概念

6.3 数据架构管理

6.4 元数据管理

6.5 主数据管理

6.6 数据质量管理

6.7 数据标准化

6.8 数据资产化

6.9 本章小结

6.3 数据架构管理

- 数据质量驱动的数据市场
 - 数据架构 (Data Architecture)是一套整体构件规范，用于**定义数据需求、指导对数据资产的整合和控制，使数据投资和业务战略相匹配。**
 - 数据架构包括正式的数据命名、全面的数据定义、有效的数据结构、精确的数据完整性规则和健全的数据文档等。



6.3 数据架构管理

- **数据管理架构是定义和维护如下规范的过程：**

提供标准的、通用的业务
术语/辞典

表达战略性的数据需求

为满足上述需求，概述
高层次的整合设计

使企业战略和相关业务
架构相一致

6.3 数据架构管理

• 数据管理架构具体内涵：

流程架构	业务架构	应用架构	技术架构	价值链分析
职能分解 流程工作流 信息产品 事件和业务周期 程序规则	目标和战略 组织架构 角色和职责 地点位置	应用系统组合 实施项目组合 软件组件架构 SOA	平台 网络拓扑 标准和协议 软件工具组合	数据 业务流程 组织角色 应用、地点 目标、项目 技术平台之间的关系

目录

6.1 引言

6.2 大数据治理基本概念

6.3 数据架构管理

6.4 元数据管理

6.5 主数据管理

6.6 数据质量管理

6.7 数据标准化

6.8 数据资产化

6.9 本章小结

6.4 元数据管理

技术元数据

存储关于数据仓库系统**技术细节**的数据，主要用来定义信息供应链中各类组成部分元数据结构，具体包括各个系统表和字段结构、属性、出处、依赖性等，以及存储过程、函数、序列等各种对象。

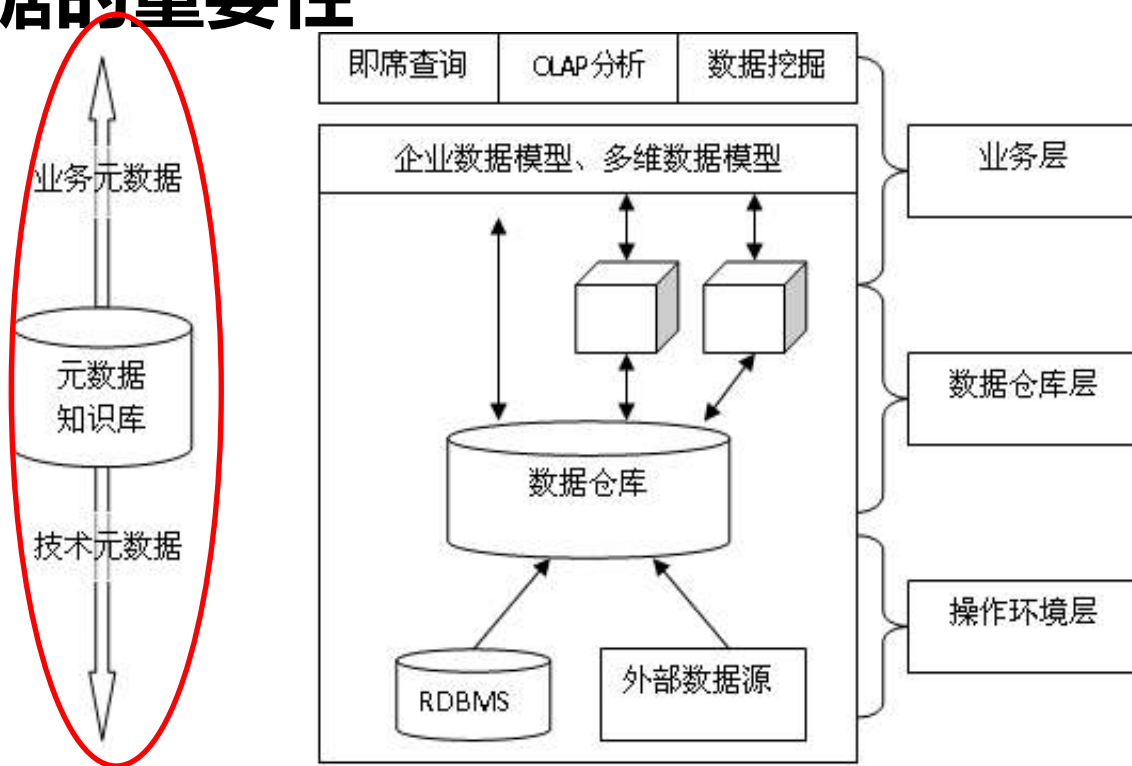
业务元数据

从业务角度描述了数据仓库中的数据，它提供介于使用者和实际系统之间的**语义层**，使得不懂计算机技术的业务人员也能够“读懂”数据仓库中的数据。它主要包括业务规则、定义、术语、术语表、运算法则和系统使用业务语言等。

6.4 元数据管理

6.4.1 元数据的重要性

元数据管理



6.4 元数据管理

6.4.1 元数据的重要性

元数据是进行数据集成所必须的

元数据定义的语义层可以帮助最终用户理解数据仓库中的数据

元数据是保证数据质量的关键

元数据可以支持需求变化

目录

6.1 引言

6.2 大数据治理基本概念

6.3 数据架构管理

6.4 元数据管理

6.5 主数据管理

6.6 数据质量管理

6.7 数据标准化

6.8 数据资产化

6.9 本章小结

6.5 主数据管理

- 主数据概念

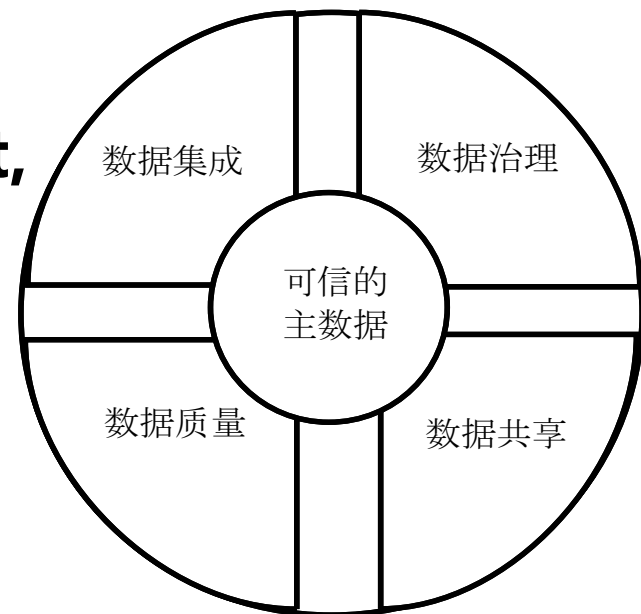
- 主数据(Master Data)是关于业务实体的数据，这些实体为业务交易提供关联环境，如当事人、产品、财务结构、位置等。主数据是关于**关键业务实体**的权威的、最准确的数据，可用于建立交易数据的关联环境。

- 在商业环境中，包括客户、员工、厂商、合作伙伴和竞争对手的数据；
 - 在公共部门，重点是关于公民的数据；
 - 在执法机构，重点是犯罪嫌疑人、证人和受害人；
 - 在非营利组织中，重点是成员和捐助者；
 - 在医疗机构，重点是病人和提供者；在教育系统中，重点是学生和教职员工。

6.5 主数据管理

6.5.1 主数据管理概述

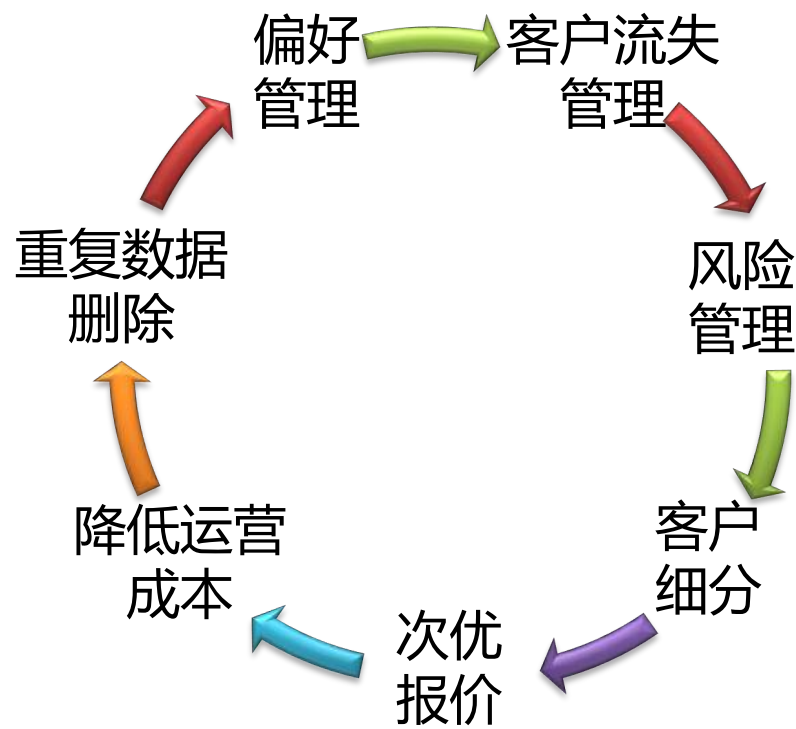
- 主数据管理(Master Data Management, MDM)是指一整套用于生成和维护企业主数据的规范、技术和方案，以保证主数据的完整性、一致性和准确性。
- 主数据管理的目的是保证系统的协调性、通用性以及主数据的正确性。



有效的主数据管理将带来四大好处：
降低运营成本、提高灵活性、提高合规性并降低风险、增加销量

6.5 主数据管理

- 大数据时代的主数据管理



6.5 主数据管理

6.5.2 主数据管理架构

- 目前业界较为常见的主数据管理解决方案主要可以分为：

依托套装软件来
实现主数据管理

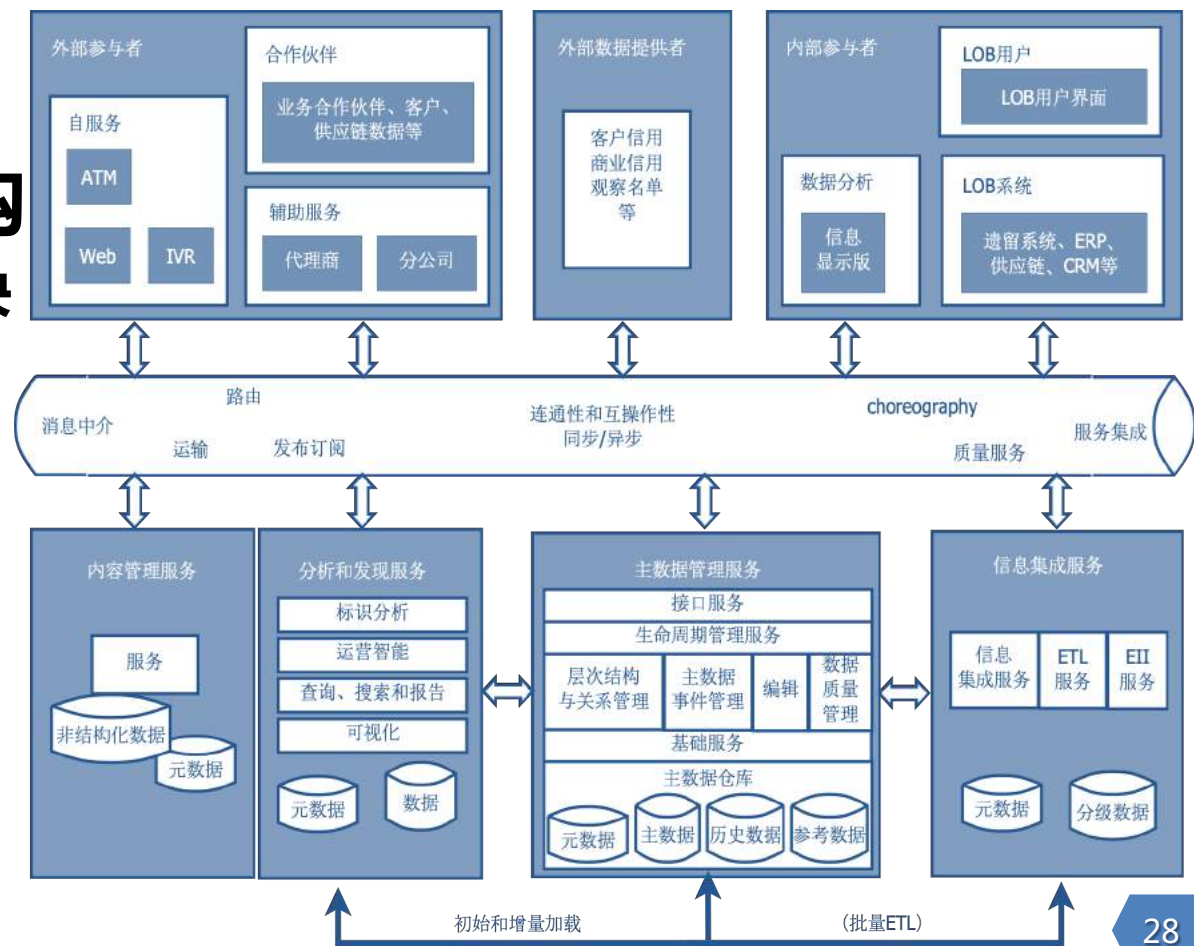
侧重于分析型应用
主数据管理

专注于主数据管理
中立的、完整的
解决方案

6.5 主数据管理

6.5.2 主数据管理架构

- 一个完整的主数据管理解决方案的逻辑架构



6.5 主数据管理

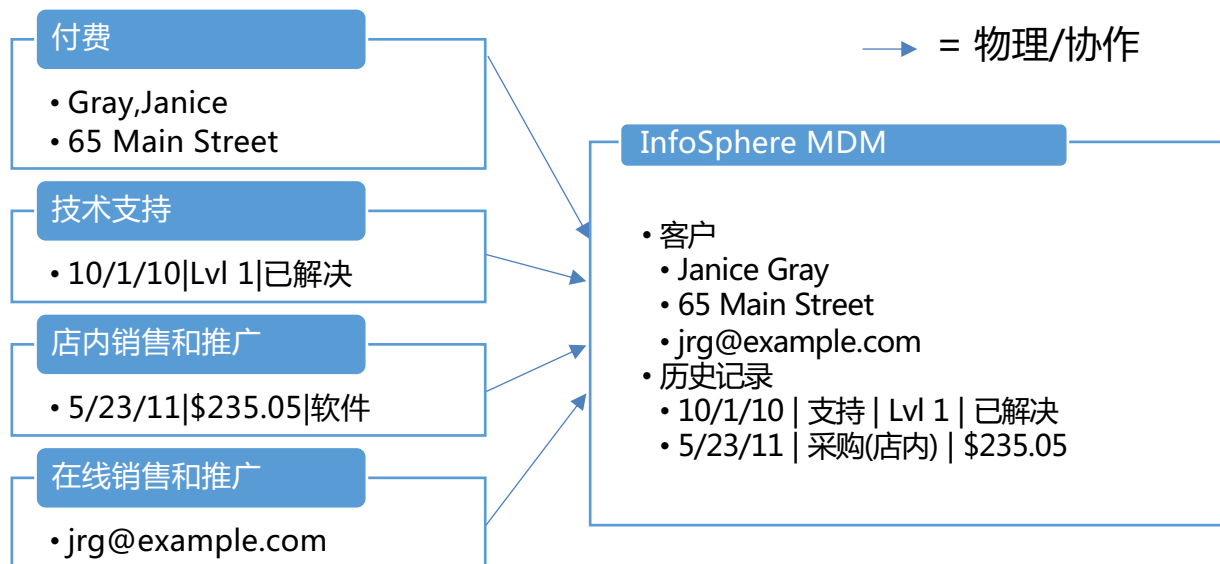
6.5.2 主数据管理架构



6.5 主数据管理

6.5.3 主数据的应用

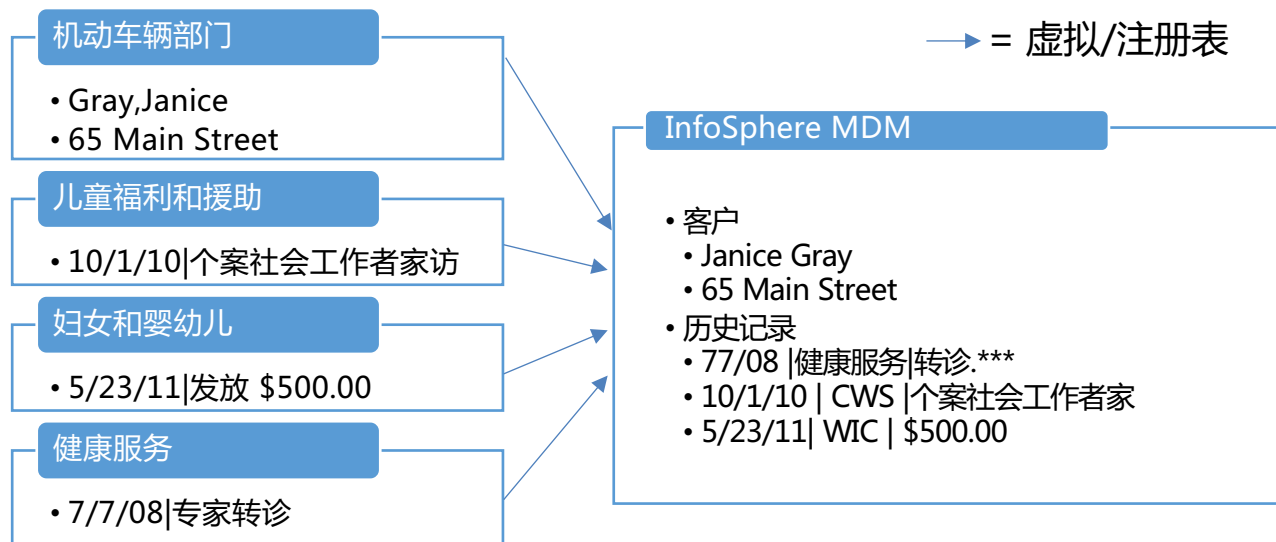
场景1：信息整合 银行、保险和电信业企业往往通过兼并和收购实现业务增长



6.5 主数据管理

6.5.3 主数据的应用

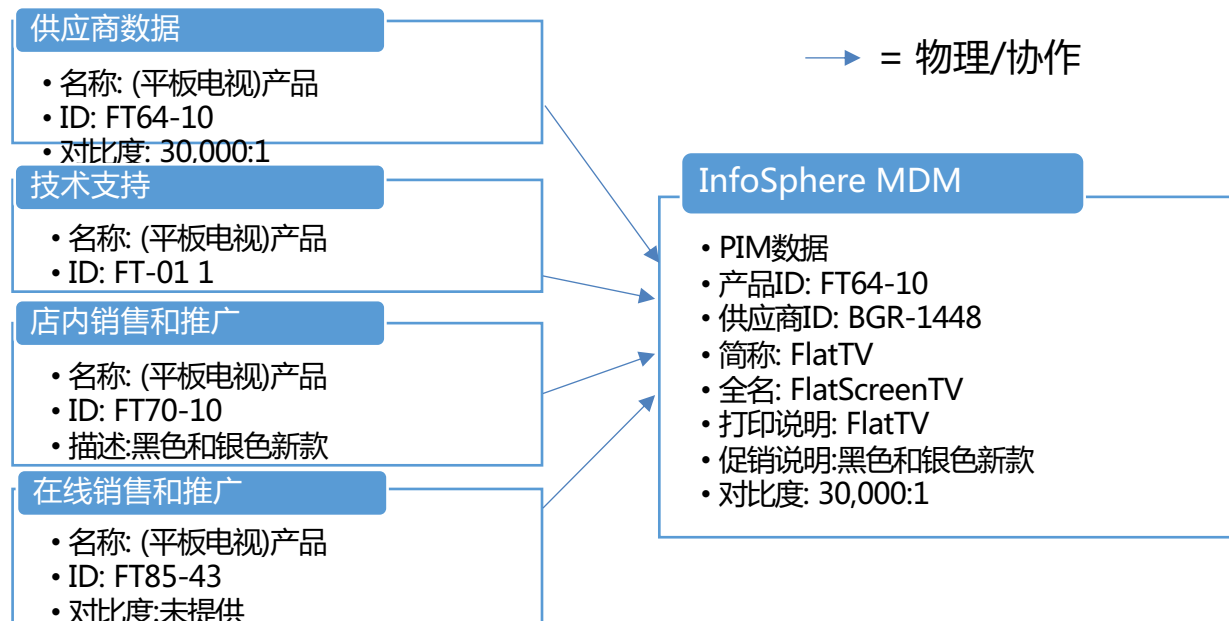
场景2：保护信息共享安全



6.5 主数据管理

6.5.3 主数据的应用

场景3：协同创作



目录

6.1 引言

6.2 大数据治理基本概念

6.3 数据架构管理

6.4 元数据管理

6.5 主数据管理

6.6 数据质量管理

6.7 数据标准化

6.8 数据资产化

6.9 本章小结

6.6 数据质量管理

6.6.1 数据质量概述

- 数据质量管理是指对数据从计划、获取、存储、共享、维护、应用、消亡生命周期的各阶段可能引发的各类数据质量问题，进行**识别、度量、监控、预警等一系列管理活动**，并通过改善和提高**管理水平**，使得数据质量获得进一步提高。

6.6 数据质量管理

6.6.1 数据质量概述

- **数据质量问题导致的恶劣后果**
 - 数据质量问题及其所导致的知识和决策错误已经在全球范围内造成了恶劣的后果。
 - 例如，在医疗方面，美国由于数据错误引发的医疗事故每年导致患者死亡人数高达98000名以上。

6.6 数据质量管理

6.6.1 数据质量概述

- 数据质量的5个维度

- 1. 数据一致性

- 数据集中，每个信息都不包含语义错误或相互矛盾的数据。
 - 例如，数据（公司="先导"，国码="86"，区号="10"，城市="上海"）含有一致性错误，因为10是北京区号而非上海区号。

公司= "先导"
国码= "86"
区号= "10"
城市= "上海"

但是上海的区号是21，而北京的区号是10。

6.6 数据质量管理

6.6.1 数据质量概述

- 数据质量的5个维度

- 2. 数据精确性

- 数据集中，每个数据都能准确表述现实世界中的实体。
 - 例如，某城市人口数量为4 130 465人，而数据库中记载为400万。宏观来看，该信息是合理的，但不精确。

城市人口 4130465

城市人口 400万

6.6 数据质量管理

6.6.1 数据质量概述

- 数据质量的5个维度

- 3. 数据完整性

- 数据集中包含足够的数据来回答各种查询, 并支持各种计算。
 - 例如, 某医疗数据库中的数据一致且精确, 但遗失某些患者的既往病史, 从而存在不完整性, 可能导致不正确的诊断甚至严重医疗事故。



6.6 数据质量管理

6.6.1 数据质量概述

- 数据质量的5个维度

- 4. 数据时效性

- 信息集合中, 每个信息都与时俱进, 保证不过时。
 - 例如, 某数据库中的用户地址在2010年正确, 但在2011年未必正确, 即这个数据已经过时。

明	相
日	逢
黄	不
花	用
蝶	忙
也	归
愁	去

6.6 数据质量管理

6.6.1 数据质量概述

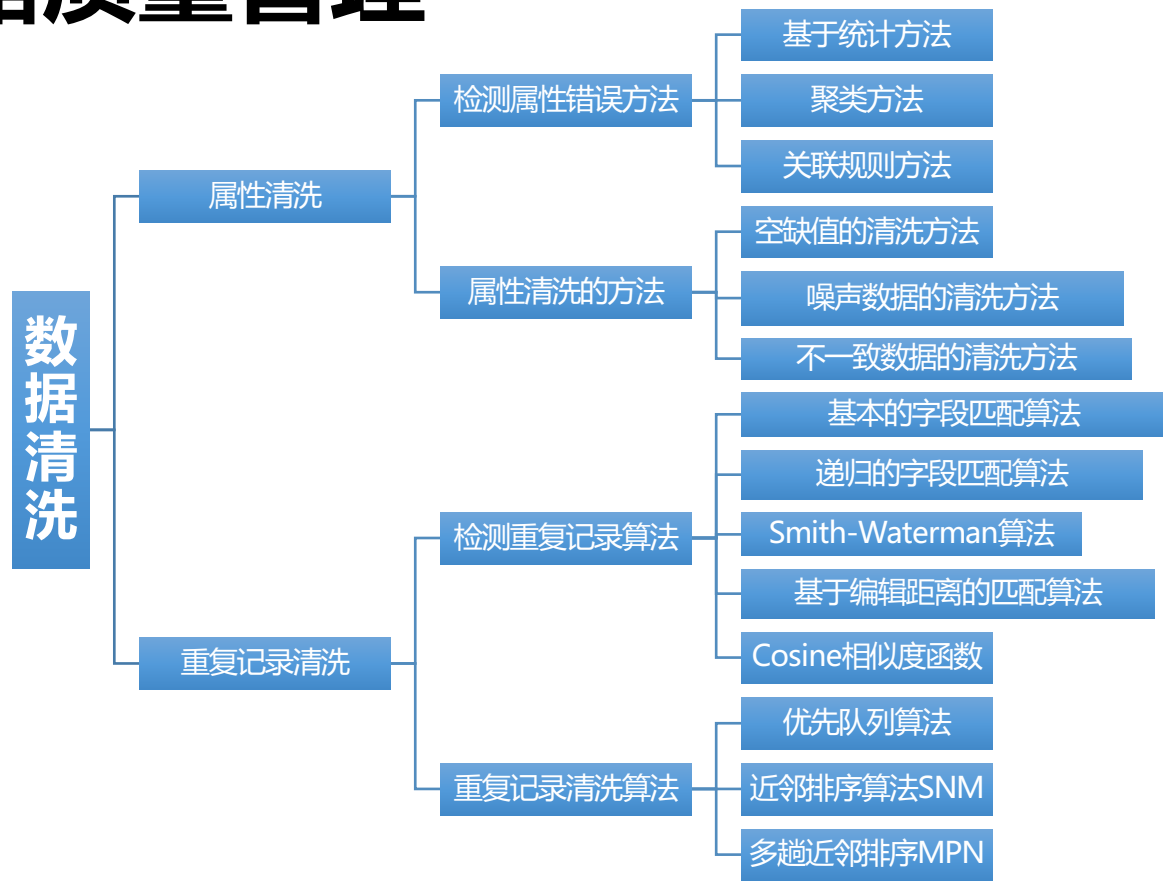
- 数据质量的5个维度

- 5. 实体同一性

- 同一实体的标识在所有数据集合中必须相同而且数据必须一致。
 - 例如, 企业的市场、销售和服务部门可能维护各自的数据库, 如果这些数据库中的同一个实体没有相同标识或数据不一致, 将存在大量具有差异的重复数据, 导致实体表达混乱。

姓名	性别	电话	出生年月
王小明	男	18277777777	1997.01
王晓明	男	18277777777	1997.01

6.6 数据质量管理



6.6 数据质量管理

6.6.2 缺失值填充

缺失值填充的方法

删除

直接删除相应的属性或样本。

统计填充

使用所有样本关于这一维的统计值对其进行填充，如平均数、中位数、众数、最大值、最小值等。

统一填充

将所有缺失值统一填充为自定义值，如“空”、“0”、“正无穷”、“负无穷”等。

平均数、中位数、众数、
最大值、最小值

空、0、正无穷、负无穷

6.6 数据质量管理

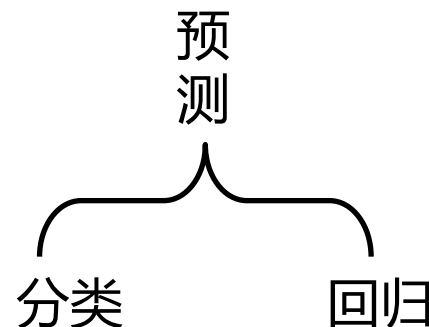
6.6.2 缺失值填充

缺失值填充的方法

预测填充

通过预测模型利用不存在缺失值的属性来预测缺失值。虽然这种方法比较复杂，但是最后得到的结果比较好。

对于类别属性，可以使用分类方法进行填充，如朴素贝叶斯方法。对于数值属性，可以采用回归的方法进行填充。



6.6 数据质量管理

6.6.2 缺失值填充

缺失值填充举例

年收入

在商品推荐场景下填充平均值，借贷额度下填充最小值。

驾龄

没有填写这一项的用户可能是没有驾照，为它填充0较为合理。

是否结婚

将是否结婚作为预测属性，构建一棵决策树，对是否结婚属性上缺失的属性值进行预测填充。

姓名	性别	电话	驾龄/年
王小明	男	18277777777	10
李刚	男	18266666666	

6.6 数据质量管理

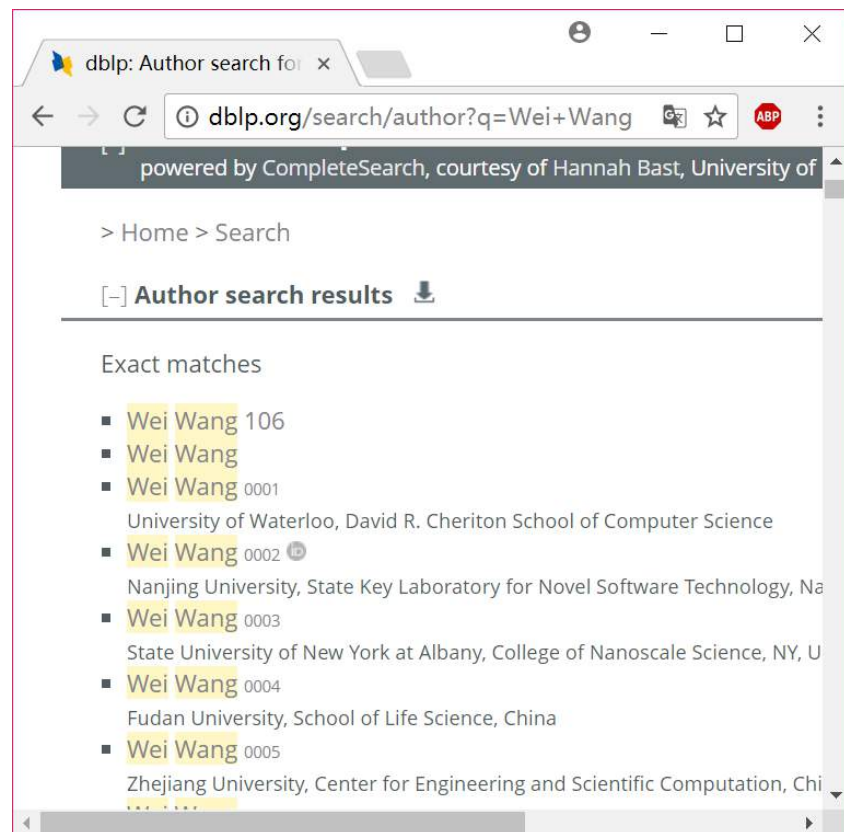
实体识别

问题

当DBLP中检索“Wei Wang”的文章时，会检索到14个“Wei Wang”的197篇文章。

什么是实体识别

在给定的对象集合中，正确发现不同的实体对象，并将其聚类，使得每个经过实体识别后得到的对象簇在将现实世界中指代的是同一实体。



6.6 数据质量管理

实体识别

实体识别解决的问题

- 1. 冗余问题: 同一类实体可能由不同的名字指代。如名字叫王伟，用英文表示可能是“Wang Wei”，也可能是“ Wei Wang”。
- 2. 重名问题: 不同类的实体可能由相同的名字指代。例如在DBLP中检索“Wei Wang”，会得到14个不同的作者。

Name	affiliation
Wei Wang	National University of Singapore
Wang Wei	National University of Singapore

Name	affiliation
Wei Wang	National University of Singapore
Wei Wang	Fudan University, School of Life Science, China

6.6 数据质量管理

实体识别

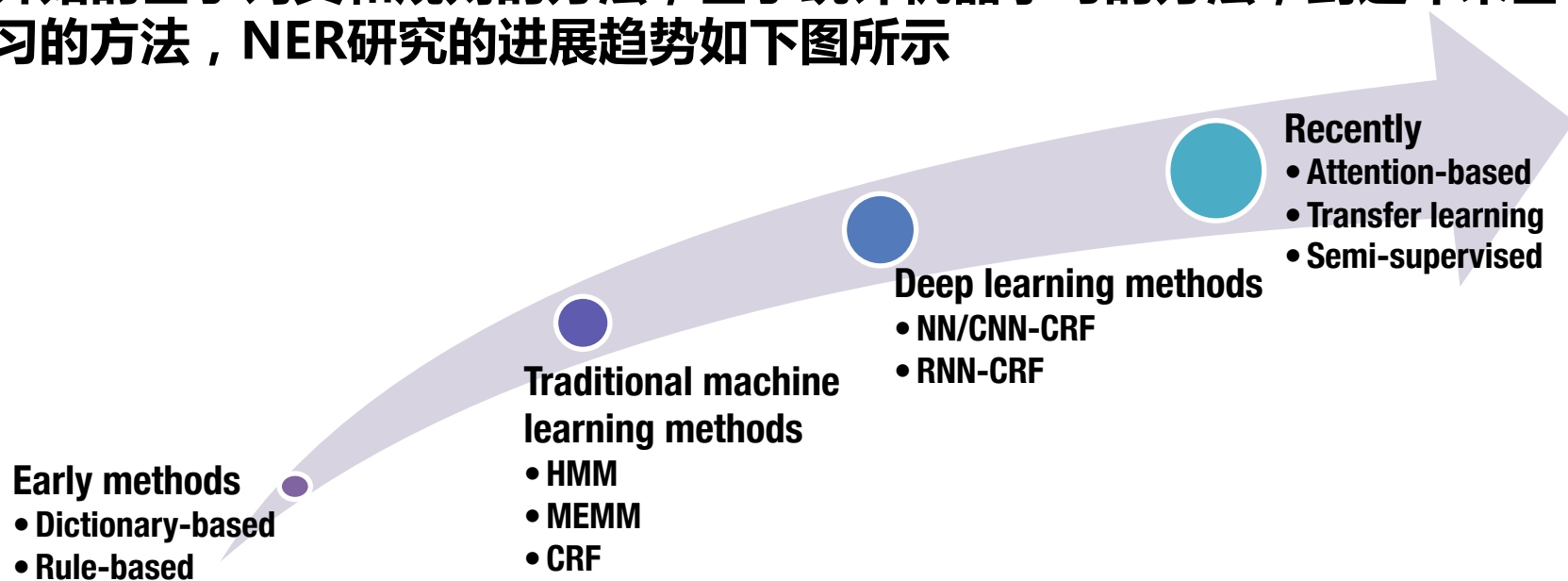
实体识别的两类技术

- 1. 冗余发现· 计算对象之间的相似性，并与阈值比较，从而判定对象是否属于同一实体类。
- 2. 重名检测· 利用聚类技术，通过考察实体属性间的关联程度判定相同名称的对象是否属于同一实体类。

Name	affiliation
Wei Wang	National University of Singapore
Wang Wei	National University of Singapore
Name	affiliation
Wei Wang	National University of Singapore
Wei Wang	Fudan University, School of Life Science, China

6.6 数据质量管理

命名实体识别 (Named Entity Recognition, NER) 一直是NLP领域中的研究热点，现在越来越多的被应用于专业的领域，如医疗、生物等。NER的研究从一开始的基于词典和规则的方法，基于统计机器学习的方法，到近年来基于深度学习的方法，NER研究的进展趋势如下图所示

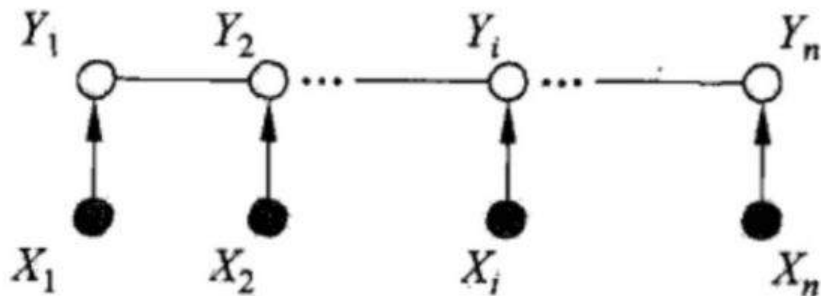


6.6 数据质量管理

- 基于统计机器学习的实体识别方法主要包括:隐马尔可夫模型(Hidden Markov Model, HMM)、最大熵 (Maximum Entropy, ME)、支持向量机 (Support Vector Machine, SVM)、条件随机场 (Conditional Random Fields, CRF) 等。
- HMM主要利用Viterbi算法求解命名实体类别序列, 在训练和识别时的效率较高且速度较快。隐马尔可夫模型适用于一些对实时性有要求以及像信息检索需要处理大量文本的应用, 如短文本命名实体识别。
- ME结构紧凑、具有较好的通用性, 缺点是训练时间复杂性高, 有时甚至训练代价难以承受, 由于需要明确的归一化计算, 导致计算开销比较大。

6.6 数据质量管理

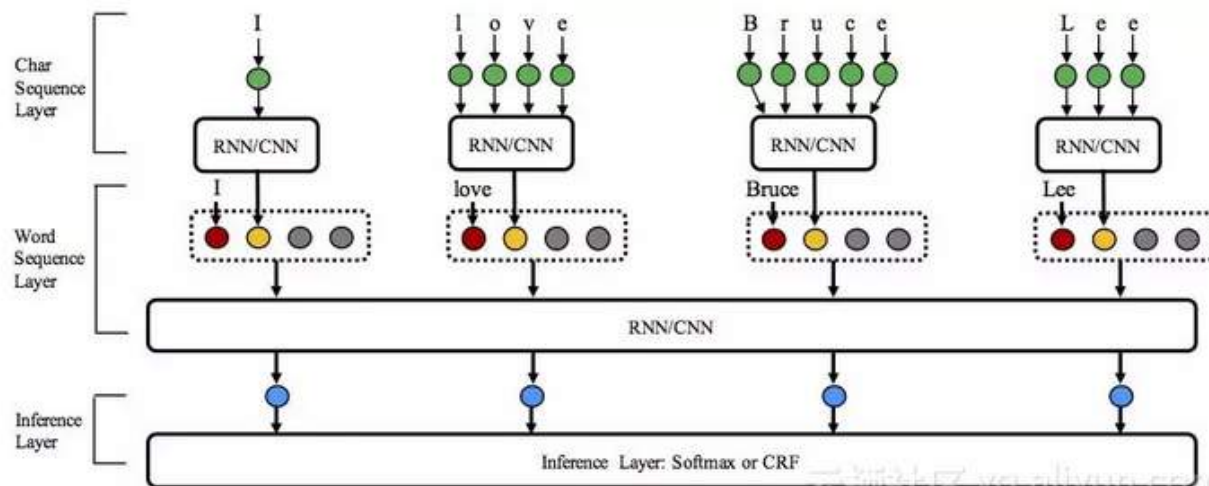
- 传统的公认较好的处理算法是CRF，它给定一组输入随机变量条件下另一组输出随机变量的条件概率分布模型，其特点是假设输出随机变量构成马尔可夫随机场，它是一种判别式概率模型，是随机场的一种。CRF常用于标注或分析序列资料，如自然语言文字或是生物序列，在NER中的基本应用是给定一系列的特征去预测每个词的标签。



- 上图中， X 可看作一句话的每个单词对应的特征， Y 可看作单词对应的标签。这里的标签就是对应场景下的人名、地名等等。

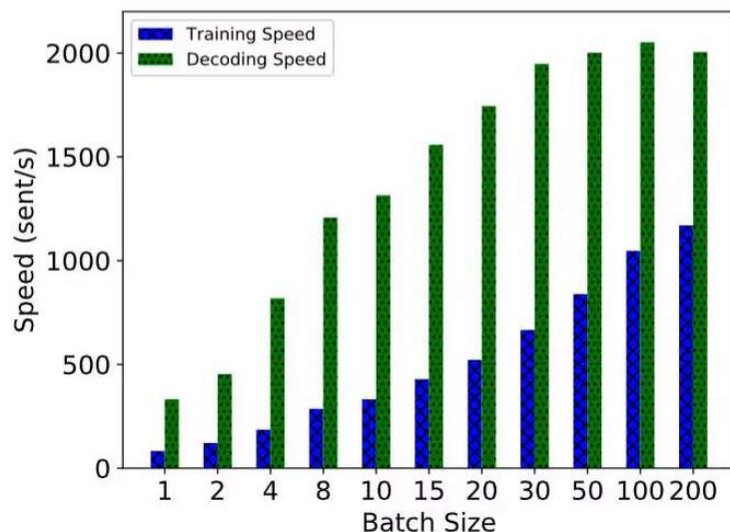
6.6 数据质量管理

- NCRF++ 算法, 是目前最好的 NER 算法, 发表在 COLING 2018上
- NCRF++ 的整体框架如下:



6.6 数据质量管理

NCRF++ 的速度表现也非常优异，在使用全批处理的情况下，在单个1080 显卡上，训练速度能到到1000句话每秒，解码速度能达到2000句话每秒。



6.6 数据质量管理

异构冲突

- 不同数据源采用不同的数据模型,树型,网型,关系表型等,可以采用分布式数据库解决

结构冲突

- 如在 Source 中 JK. Rowling 指代一个实体,而在另一个数据库中,指代的是一个属性值。

语义冲突

- 对同一个实体,不同数据源的描述在语义上有差别,传统的数据集成在解决语义冲突时采用本体论的方法,但是数据本身可能是劣质的,在劣质数据上采用本体论的方法效果并不理想。

描述冲突

- 语义上是一致的,但是两个源给的描述仍不能百分百相同,在描述冲突中,可以用自然语言处理相关技术,如短语分离,数据融合技术 Data Fusion1。

6.6 数据质量管理

- 针对异构冲突和结构冲突已有很多解决策略：分布式数据库是一种很好的方法,但能消解的冲突有限。采用数据融合技术 data fusion将不同提取模型中的数据进行融合，针对数据源提取模型,对模型进行融合。当然也可以采用模式匹配的相关技术，如模式集成 Schema。
- 真值发现是进行冲突消解的重要技术，可以作为消解语义冲突的一种。

6.6 数据质量管理

真值发现

实体识别之后

经过实体识别之后，描述同一现实世界实体的不同元祖被聚到了一起，然而这些对象的相同属性值可能包含冲突值。

真值发现

在这些冲突值中，发现真实的值。

Name	affiliation	Age
Wei Wang	National University of Singapore	41
Wang Wei	National University of Singapore	47

6.6 数据质量管理

真值发现的两个思路

投票方法

往往真值是由大多数的源提供。

Name	affiliation	Age
Wei Wang	National University of Singapore	41
Wang Wei	National University of Singapore	47
Wang Wei	National University of Singapore	47

6.6 数据质量管理

真值发现的两个思路

投票方法

往往真值是由大多数的源提供。

考虑数据源精度的迭代方法

迭代地计算数据源的可信度，进而计算事实的置信度。

Name	affiliation	Age
Wei Wang	Naonal Unrsity of Singpe	47
Wang Wei	Natiol Uniety of Sinaore	47
Wei Wang	National University of Singapore	41
Wang Wei	National University of Singapore	41

目录

6.1 引言

6.2 大数据治理基本概念

6.3 数据架构管理

6.4 元数据管理

6.5 主数据管理

6.6 数据质量管理

6.7 数据标准化

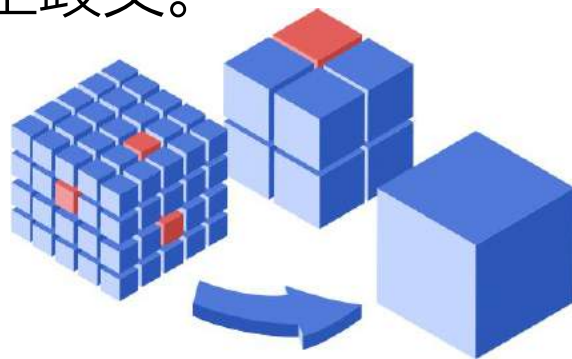
6.8 数据资产化

6.9 本章小结

6.7 数据标准化

• 概述

- 数据需求缺乏规范，造成数据对象多份存储，存储结构各异，严重影响数据共享；
- 数据标准依据各异，造成统计口径无法匹配；
- 业务口径不统一，造成沟通困难，发生歧义。



6.7 数据标准化

- **数据标准**

- 若干描述规范 and 要求的文档
- 为了使组织内部和外部使用交换的数据保持准确并且一致，经协商一致制定的并由相关主管机构批准，共同使用和重复使用的一种规范性文件
- 由管理规范、管控流程、技术工具共同组成的体系，使通过这套体系逐步实现信息标准化的过程

6.7 数据标准化

- **数据标准依据不同的实施领域可分为三类：基础类、分析类和专有类**

基础类数据标准是企业日常业务开展过程中所产生的具有共同业务特征的基础性数据

分析类数据标准指为满足公司内部管理需要及外部监管要求在基础性数据基础上按一定统计、分析规则加工后的数据

专有类数据标准是公司架构下子公司在业务经营及管理分析中所涉及的特有数据

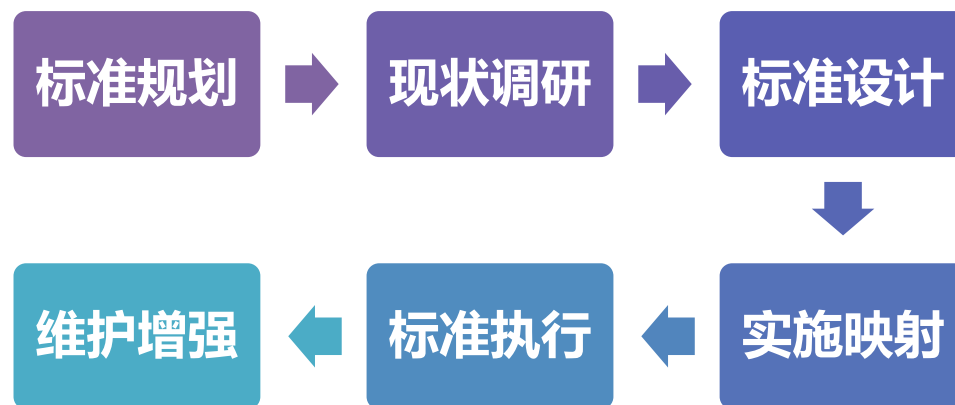
6.7 数据标准化

- 数据标准

- 数据标准包括三个文档：主题定义、信息项、标准代码
- 信息项文档是数据标准的核心，一般由信息大类、信息小类、信息项、信息项描述、信息类别、长度等6项组成。

6.7 数据标准化

数据标准化 (Data Standardization)是指研究、制定和推广应用统一的数据分类分级、记录格式及转换、编码等技术标准的过程。

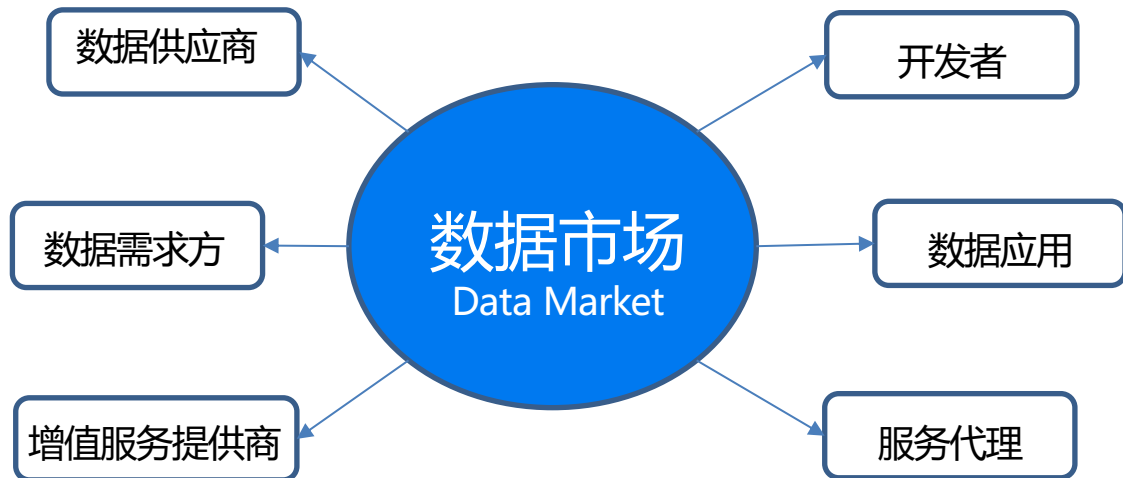


6.7 数据标准化

• 数据标准化应用案例

序号	应用领域	案例名称
1	社会管理与公共服务	案例一：地市级城市政务信息资源标准化归集、交换和应用——徐州智慧城市信息资源枢纽
2	城市运营	案例二：昆明国家经济技术开发区城市智能运营中心（IOC）
3	金融业	案例三：“数控金融”互联网金融大数据监管平台应用案例
4	农业	案例四：农业部农产品批发市场价格挖掘及可视化平台
5	制造业	案例五：海尔 COSMO Plat 空调噪音大数据智能分析
6	制造业	案例六：长安汽车智能制造技术研究所冲压质量大数据项目
7	工业	案例七：江苏省重点领域共性技术攻关项目——工业大数据元数据规范与验证技术攻关专题——工业大数据标准体系研究及重点标准编制
8	电力	案例八：国网电力大数据应用案例
9	通信	案例九：浙江移动“天盾”反欺诈系统
10	邮政	案例十：国家邮政局数据管控系统
11	科研	案例十一：重庆两江大数据科技服务双创孵化基地案例
12	卫生	案例十二：本溪大健康服务平台
13	文化	案例十三：CCDI 版权监测案例

6.7 数据标准化



- 大数据时代，数据可以作为一种资产或商品进行交易。
- 数据市场能够在不通过的行业客户之间搭建起一个数据交易平台，使数据可以作为商品进行市场流通。
- 数据市场可对这些原始数据进行初步加工处理，提供高价值、合规数据的交易支持。
- 数据需求方可通过数据市场搜索到所需数据，通过交易采购，以帮助企业补充、丰富其数据总量和范围、支持企业自身业务发展。

目录

6.1 引言

6.2 大数据治理基本概念

6.3 数据架构管理

6.4 元数据管理

6.5 主数据管理

6.6 数据质量管理

6.7 数据标准化

6.8 数据资产化

6.9 本章小结

6.8 数据资产化

- **数据资产是企业及组织拥有或者控制，能给企业及组织带来未来经济利益的数据资源。**

拥有和控制

- 表明数据资产不一定是企业在内部信息系统中拥有的数据资源，也可以是通过合作、租赁等手段，从企业外部获取使用权的各种数据形式。

带来未来经济利益

- 表明数据资产拥有或间接导致资产或现金等流入企业的潜力。这种潜力是将数据作为一种经济资源纳入企业经济活动的能力。通过为企业管理控制和科学决策提供合理依据，减少和消除企业经济活动中风险，进而为企业简介带来预期的经济利益，也可以通过交易或事项，直接为企业带来经营收入。

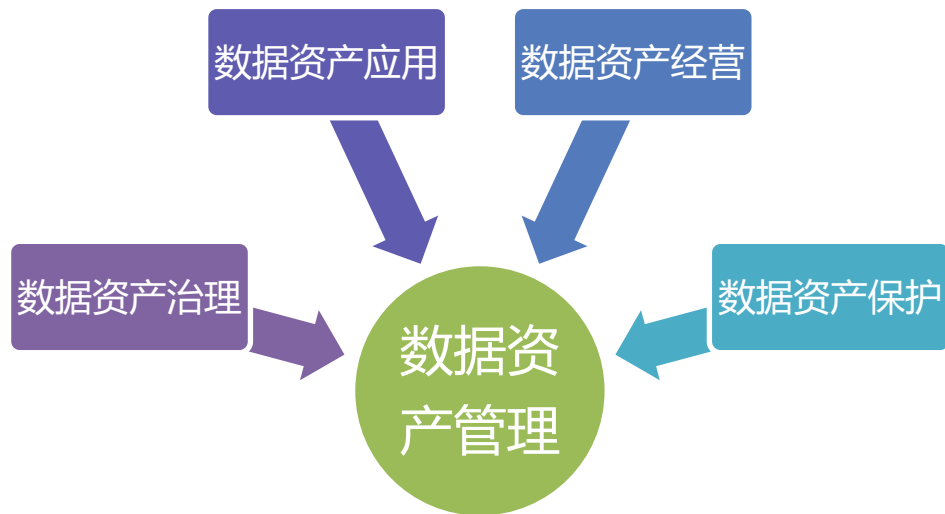
数据资源

- 指出了数据资产的具体形态，表现为以物理或电子方式记录的数据，如工作记录、表单、配置文件、拓扑图、系统信息表、数据库数据、操作和统计数据、开发过程中的源代码。

6.8 数据资产化

• 数据资产管理

- 数据资产管理是企业及组织采取的各种管理活动，用以保证数据资产的安全完整，合理配置和有效利用，从而提高数据资产带来的经济效益，保障和促进各项事业发展。一般而言，包含三个领域：数据资产治理、数据资产应用和数据资产经营。



6.8 数据资产化

参考开放组体系框架(The open group architecture framework, TOGAF)

输入	步骤	输出
<div><div>·外部参考资料</div><div>·架构参考资料</div><div>·非架构的输入</div><div>·架构工作需求</div><div>·能力评估</div><div>·沟通计划</div><div>·架构的输入</div><div>·企业架构组织模型</div><div>·定制的架构框架</div><div>·数据原则</div><div>·架构工作说明书</div><div>·架构愿景</div><div>·架构储藏库（可重用构建块、组织特定参考模型、组织标准）</div><div>·架构定义文档的草稿</div><div>·架构需求规格书草稿</div><div>·架构路线图的业务架构构件</div></div>	<div><div>·选择参考模型、视点和工具</div><div>·开发基线数据架构描述</div><div>·开发目标数据架构描述</div><div>·进行差距分析，安全和私有性影响分析</div><div>·定义路线图构件</div><div>·解决跨架构图景的影响</div><div>·进行利益相关者的正式审查</div><div>·最终确定数据架构</div><div>·建立架构定义文档</div></div>	<div><div>·优化和更新架构愿景阶段的交付物</div><div>·更新架构工作说明书</div><div>·架构定义文档草稿</div><div>·基线/目标数据架构 1.0 版（业务数据模型、逻辑数据模型、数据管理流程模型、数据实体/业务功能矩阵）</div><div>·数据架构相关视图</div><div>·架构需求规格书草稿</div><div>·差距分析结果</div><div>·技术需求</div><div>·对将要设计的技术架构的约束</div><div>·更新的业务需求</div><div>·架构路线图的数据架构构件</div></div>

6.8 数据资产化

数据资产发现与评估

Zachman框架

- 由约翰·扎科曼在1987年创立的全球第一个企业架构理论,是其他企业框架的源泉。 Zachman框架实际上是组织架构工具(用来设计文档、需求说明和模型的工具)的一种分类学,架构材料往往通过使用二维表格进行表示。

EAP框架:

- Steven H. Spewak在1998年定义了企业架构计划(Enterprise Architect Planning, EAP),它用于制定信息架构以支持业务这一过程和实现该过程的计划,它更偏重于企业架构的动态部分,包括过程、计划、阶段的划分等。EAP将企业架构的过程分为四个步骤即开始启动、现状分析、目标分析、实现和整合计划。

FEAF框架:

- 美国联邦政府CEO委员会于1999年提出了联邦政府组织架构框架(federal enterprise architecture framework, FEAF)并用于指导美国政府部门的信息化建设。FEAF首先是一种组织机制,被用来管理企业架构描述的开发和维护,而在将企业架构付诸实施方面FEAF还提供了一种结构,用于组织联邦政府资源以及描述和管理联邦企业架构的相关行为。

6.8 数据资产化

数据交易

(1) 直接销售模式

数据直接销售模式是最基础、最直接、最简单的数据交易模式，数据所有者明码标价，将自身拥有的数据交易给买方。交易过后卖方获得数据的货款，买方获得数据本身。就像去菜市场买东西，我给你钱，你给我蔬菜，通过简单的交易得到对方想要的东西。



主体

大数据
经营商

发展源
动力

数据重
复利用

业务
核心

大数据
交易

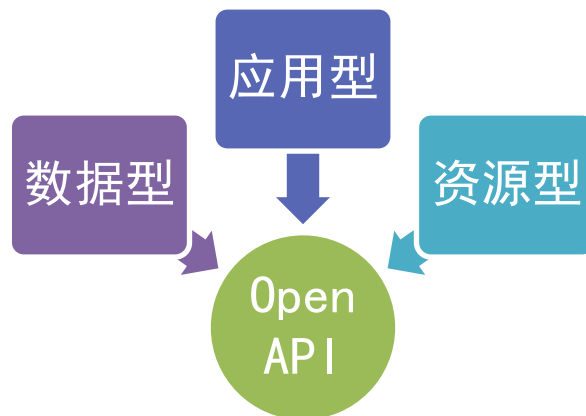
6.8 数据资产化

数据交易

(2) Open API模式

Open API 并不是新概念，在计算机操作系统出现早期就已存在。

在互联网时代，把网站的服务封装成一系列计算机易识别的数据接口进行开放，以供第三方开发者使用，称为开放网站的API，与之对应的所开放的API称为Open API。



6.8 数据资产化

数据交易

(3) 多方合作模式

多方合作模式是数据交易的一种特别方法，立足于数据加工的特性，在多方之间建立合作关系，共同提供数据、加工数据、分享信息成果，并在此基础上形成利益划分关系。

在很多企业间，建立数据合作实验室、数据沙箱、数据研究分享计划等，即为具体多方合作的体现。这种合作方式基于企业间的合作关系或框架协议，能够更好地形成信任，并且推动深度的数据共享，可视作一种更为彻底的数据交易模式。



6.8 数据资产化

大数据在价值上具有不确定性、稀缺性、多样性等特征

效用价格——估计大数据的使用价值 $P_{\max} = \sum_{i=1}^n T_i J_i - \sum_{j=1}^n Q_j H_j.$

成本价格——估计大数据的建立和维护成本 $P_{\min} = C_0 + C_0 r$

参数符号	说明
Q_j	使用大数据之前事件发生的概率
H_j	使用大数据之前事件的预期收益
C_0	生产成本
r	利润率
P	成本价格

6.8 数据资产化

数据定价

大数据的定价策略主要有静态定价策略和动态定价策略两种形式。其中，静态定价策略包括多重定价、歧视定价、捆绑定价和拉姆齐价格。动态定价策略包括协商定价、拍卖式定价、反向拍卖式定价。

- 预处理策略
- 拍卖定价策略
- 协商定价策略
- 反馈性定价策略

6.8 数据资产化

数据定价

案例1:

2007年,英国另类摇滚乐队“电台司令”(Radiohead)决定:他们最新推出的包含十首歌曲的专辑《彩虹里》将不再走传统的定价和分销模式,而**大胆地将其放在乐队主页上,由粉丝们以任意价格下载**。结果在竞争激烈且受到盗版和互联网严重冲击的音乐行业,这项“随您打赏”的实验大获成功。从10月9日实验开始到10月29日项目结束,超过180万人下载了该专辑,虽然60%下载者没有付钱,但另外40%都自愿支付了一定金额。

乐队表示:“就收入数字而言,**这张专辑赚的钱比我们以往所有专辑加起来的还要多**。”

6.8 数据资产化

数据定价

案例2:

在时尚的服装行业，衣服的零售价总要比成本高出许多，通常商场都会在一些时间里做打折促销，以吸引更多的顾客在自家店里驻足，而且也期望通过这种让利促销行为给消费者留下美好善良的印象。

不过，顾客并不能感受到充分的善意，因为她们几乎不能及时获知何时进行促销活动。客观上，这就使得商家不能将相对价格敏感的潜在购买者的价值进行变现。但是总部设在纽约的服装零售商Syms创新地引入了自动降价机制，破解了这种困局。在Syms商场，女性服装的标签上都标着三个价格：全国统一售价、Syms售价，以及日后的折扣售价，这二个价格一个比一个便宜，而且每个价格只保持10天。通过这种相对透明明的自动定价模式，Syms扩大了客户群，保持了又高又稳定的销售额。

6.8 数据资产化

数据定价

案例3:

在医药行业，日益高涨的药价已经将药品供应商推到了和消费者严重对立的位置。**消费者想要享受低价，而供应商却要高价创收**，这是一个死结。

有没有可能开创一个新的定价方式，将药品的供应端和消费端在价值上统一起来，化冲突为合作？强生公司的**“好用再付钱”模式**就做到了。

强生公司为过去不可治疗的多发性骨殖瘤研发了一种新的药物万珂 (Velcade)，并准备在英国推广。但英国国家卫生医疗质量标准署(NICE)却认为这种药纯粹是浪费政府的钱。这种药每个疗程要3000英镑，疗效却不确定。

遭到严厉拒绝后，强生公司采取了种截然相反的办法：**经过四个疗程，任何此种癌症产生的病变蛋白没有降低25%的患者，可以得到全额退款。**

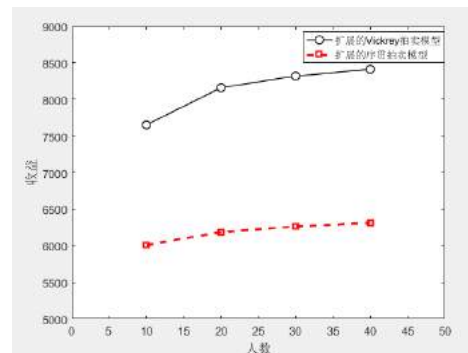
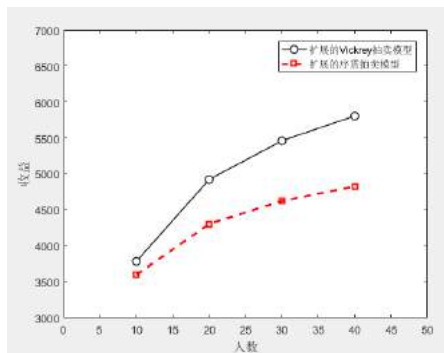
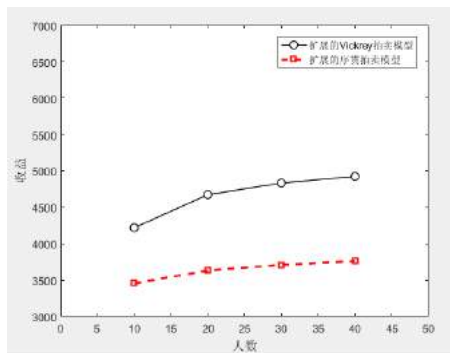
6.8 数据资产化

大数据拍卖定价策略

问题背景：大数据交易规模小、定价方法简单、价格设置较低导致卖方失去持续供给大数据商品的动力

模型1：改进Vickrey模型使得可以在拍品数量不确定情况下确定拍品数量

模型2：拍品数量不确定时的序贯拍卖，使得卖家收益最大化



结论：在拍品数量可以选择的情况下，两种模型都可以有效帮助卖家进行拍卖决策，实现自身期望收益的最大化。

6.9 小结

- **数据治理是对数据资产管理形式权力和控制的活动集合（规划、监控和执行），是大数据产生价值的关键步骤之一。**
 - ❖ **大数据治理的基本概念**
 - ❖ **数据架构管理的定义和参考模型**
 - ❖ **元数据管理的概念和作用**
 - ❖ **主数据管理的概念、主数据的架构和应用**
 - ❖ **数据质量管理的基本技术，缺失值填充、实体识别和真值发现等关键技术**
 - ❖ **数据标准化相关的概念和应用实例**
 - ❖ **数据资产化方面的内容，包括数据资产发现、评估、交易、定价等方面的概念与技术。**