

第 1 章

绪 论

1.1 智能制造与工业大数据

制造业一直是国民经济的重要支柱,是工业化和现代化的主导力量。自 18 世纪中叶工业革命以来,制造业的发展成为衡量一个国家或地区综合经济实力和国际竞争力的重要标志,也是国家安全的重要保障。事实上,历史上世界强国的兴衰与中华民族的奋斗历史一再证明,没有强大的制造业,就难以实现国家和民族的强盛。因此,打造具有国际竞争力的制造业,对于提升一个国家的综合国力、保障国家安全、建设世界强国来说,是必不可少的。

当前,随着新一轮科技革命和产业变革的兴起,以信息技术与制造业加速融合为主要特征的智能制造已成为全球制造业发展的主要趋势。中国机械工程学会在其《中国机械工程技术路线图》中指出,智能制造是制造自动化、数字化、网络化发展的必然结果^[1]。智能制造的主线是智能生产,其中智能工厂和车间是智能生产的主要载体。新一代智能技术的应用及其与先进制造技术的融合,使生产线、车间、工厂发生了革命性大变革。这些智能车间具有自学习、自适应、自控制等特征,从根本上提高了制造业的质量、效率和企业竞争力。

智能制造技术的应用和普及使制造业的整个价值链和产品的全生命周期都涉及大量的数据,如虚拟化和仿真等。这也导致制造企业数据呈现爆炸式增长趋势。如今数据已然成为现代社会的新型基础设施和核心资产,是继土地、劳动力、资本、技术之后的第五大生产要素。数据之于当今社会,正如石油之于工业革命时期,成为推动经济增长和社会进步的关键驱动力。数据的价值不仅体现于其自身,还在于分析和应用所能产生的洞察力和创新能力。行业领导者希望利用数据的力量支持制造业的发展,解决以往难以解决的问题,引导制造业向高附加值的深入转型。数据在制造业中的应用不仅有助于解决传统问题,还能发掘新的业务机会,比如通过数据分析优化生产流程、降低成本、提高效率和创新产品设计。

工业大数据是指从制造车间生产现场到制造企业顶层运营所有生成、交换和

集成的数据,包括所有与制造相关的业务数据和衍生信息。这些数据在工业设备的生产、采集和处理过程中产生,具有强烈的实时性。它们是具备海量、高增长率和多样化特征的信息资产,广泛存在于企业各类应用系统中,如产品生命周期管理(product lifecycle management, PLM)、供应链管理(supply chain management, SCM)和制造执行系统(manufacturing execution system, MES),以及集成企业资源计划(enterprise resource planning, ERP)系统中。这些数据大多针对具体应用需求而采集存储,互通互用性相对较低,却是企业信息化逐步深入、数据积累到一定程度的必然产物。工业大数据需要新的处理模式,使制造企业具有更强的问题洞察能力、实时决策力和流程优化能力。

在智能制造中,大数据技术拉近了消费者与设计师之间的距离,使预测型制造成为可能,并有助于实现产品的智能化运维。它还能产品的精准营销和成本控制提供支持。目前,大数据技术在制造业中的应用包括供应链管理^[2]、产品生命周期管理^[3]等方面。然而,大数据技术在制造业的应用仍面临诸多挑战,如思维变革、技术发展、商业模式创新、基础设施建设、信息安全和人才及管理问题。这些挑战需要制造业在技术应用和组织管理上进行深入探索和不断创新。

此外,随着云计算、物联网、人工智能等技术的快速发展和广泛应用,制造业的数字化转型正在加速。数字孪生技术,即通过创建物理资产的虚拟副本优化产品设计、生产和运维流程,也成为智能制造的重要组成部分。这些技术不仅提高了制造效率和产品质量,还使制造过程更加灵活且市场变化的响应能力更强。环境可持续性也成为制造业转型的重要考量。随着全球对环境保护和可持续发展的重视,绿色制造和循环经济的理念正逐渐融入制造业发展。这不仅涉及生产过程中资源的高效利用和环境影响的最小化,也包括产品设计的生态友好性和整个产品生命周期的环境影响。

智能制造和工业大数据正深刻改变着制造业的面貌。这些技术不仅提高了制造业的效率和质量,也为制造业带来了新的商业模式和增长机会。制造业的未来将更依赖技术创新、数据驱动的决策和对环境可持续性的重视。在此背景下,数据治理、数据安全和隐私保护等问题也日益突出。如何平衡数据的利用与个人隐私保护、确保数据安全和公平竞争,成为数字经济发展中不可忽视的重要议题。

1.2 工业大数据的定义

关于工业大数据的定义,目前学术界与工业界还没有形成共识。本书采用《工业大数据白皮书》中关于工业大数据的定义:

工业大数据指在工业领域中,围绕典型智能制造模式,从客户需求到销售、订单、计划、研发、设计、工艺、制造、采购、供应、库存、发货和交付、售后服务、运维、报废或回收再制造等整个产品全生命周期各个环节所产生的各类数据及相关技术和

应用的总称。

由以上定义可以看出,工业大数据以产品数据为核心,极大延展了传统工业数据范围,同时包括工业大数据相关技术和应用。

工业大数据具备双重属性:价值属性和产权属性。一方面,通过工业大数据分析等关键技术能够实现设计、工艺、生产、管理、服务等各环节智能化水平的提升,满足用户定制化需求,提高生产效率并降低生产成本,为企业创造可量化的价值;另一方面,这些数据具有明确的权属关系和资产价值,企业可决定数据的具体使用方式和边界,数据产权属性明显。工业大数据的价值属性实质上是基于工业大数据采集、存储、分析等关键技术,对工业生产、运维、服务过程中的数据实现价值的提升或变现;工业大数据的产权属性则偏重于通过管理机制和管理方法帮助工业企业明晰数据资产目录与数据资源分布,确定所有权边界,为其价值的深入挖掘提供支撑。

1.3 工业大数据的特性分析

随着制造业信息化和自动化技术的飞速发展,特别是数控机床、传感器、数据采集装置和其他具备感知能力的智能设备在离散车间底层的大量使用,车间数字化和智能化程度越来越高,制造数据的规模和种类也逐具规模,车间运行过程中产生的制造数据呈现以下特性。

(1) 规模性。以半导体制造为例,单一晶圆检测如今每个晶圆可产生 15~16GB 数据,在某些先进情况下接近 0.25TB。一个快速自动检测设备每年可产生数百 TB 的数据。以上海某晶圆企业 2024 年数据为例,全球检查超过 10.2 亿晶圆,涉及各种参数监测,数据总量达到 Exabyte 级。据报告显示,2024 年全球工业企业平均每年产生约 2.1ZB 的数据,而这个数据量预计每两年增长约 1.13 倍。GE 报告^[1]显示,未来 10 年工业数据增速将是其他大数据领域的 2 倍。由于晶圆制造车间在制品多,工艺多,产品多,设备数量大,参数多,数据采集频繁,使得记录的数据量巨大。

(2) 多样性。主要体现为类型多样、量纲多样、尺度多样。类型上,车间生产涉及的产品物料清单(bill of materials, BOM)结构表、工艺文档、数控程序、三维模型、设备运行参数等制造数据往往来自不同的系统,具有完全不同的数据结构,可表现为数值、时间,也可表现为检测时的图像。即使都为数值类型,也分为整数类型、浮点数类型、布尔值类型等;量纲上,如参数中的温度、压力、流量,度量单位各不相同;尺度上,如刻蚀的宽度为纳米(10^{-9})级别,而晶圆搬运距离为千米(10^3)级别,尺度跨度达 10^{12} ;运行数据采集的频率不尽相同,从 3~5min 一次到每秒数十次,如晶圆刻蚀过程中反应腔温度采样频率为每秒 1~10 次,而刻蚀深度只有在—批晶圆加工完成后才检测,时间尺度为小时级别。

(3) 高速性。机台加工、测量与缺陷数据随制造过程的进行不断产生,并由相应传感器实时采集。设备数据波动、制程工艺变化、随机性缺陷等随晶圆加工批次不断变化。采集到的数据,尤其是机台数据,主要为机台运行控制提供支持,要求具有时间敏感性。因此,要求数据的采集间隔周期小,处理速度快,响应时间短。

对制造车间数据做进一步分析,发现其在 3V(volume, velocity, variety, 大量、高速、多样)特性基础上,兼具多来源(multi-source)、多维度(multi-dimension)、多噪声(multi-noise)的 3M 特性。

(4) 多来源。制造车间的产品订单信息、产品工艺信息、制造过程信息、制造设备信息分别来源于排产与派工(product planning & scheduling, PPS)系统、产品数据管理(product data management, PDM)系统、制造执行系统和制造数据采集(manufacturing data collection, MDC)系统、数据采集与监控(supervisory control and data acquisition, SCADA)系统等。近年来,随着工业大数据技术的发展,制造车间逐渐普及了主数据系统(master data system, MDS)、商务智能系统(business intelligence system)等。这些系统基于原始生产过程数据,经过一系列手段处理,产生了新的数据和信息。因此,制造车间中原始数据和中间数据并存,不同的信息来源决定了不同的数据结构和存储方式。

(5) 多维度。晶圆制造在产品、系统和设备维度体现出了多维度特点。产品维的性能参数有晶圆良率、测试寿命、电子性能与成品良率等,系统维有晶圆工期、库存水平、设备空载率、等待队列长度等,设备维有设备利用率、工序良率、设备平均失效前时间和平均故障间隔时间等。

(6) 多噪声。车间生产运行中产生的大量数据来自可编程逻辑控制器(PLC)、传感器和其他智能感知设备对制造过程的不断采样,生产中的电磁干扰和恶劣环境使感知数据具有高噪声的特点。以晶圆卡的搬运过程为例,RFID 阅读器对附于高速移动的搬运小车上的芯片进行读取时,易出现漏读和误读现象;等离子刻蚀腔中充斥高温高压等离子体,导致腔内传感器老化快,原点漂移严重。

1.4 大数据带来的新变革

1.4.1 大数据对科学研究的影响

大数据对科学研究的影响深远且广泛,其重要性也日益凸显。在科学研究领域,大数据不仅为研究者提供海量的信息,还极大地改变了研究方法和思维模式。图灵奖得主、数据库领域的权威人物吉姆·格雷(Jim Gray)博士曾对科学研究的演变历程进行了深刻剖析,并将科学研究的发展划分为 4 个关键阶段:实验科学阶段、理论科学阶段、计算科学阶段和数据密集型科学阶段^[5],为深入理解大数据在科学研究中的角色和重要性提供了宝贵的框架。

1. 实验科学阶段

科学研究的早期主要以实验方法为主导,这一阶段,科学家通过设计各种实验、收集各类数据验证或推翻假设。例如 1590 年伽利略的比萨斜塔实验,以实际观察和验证的方式推翻了亚里士多德关于物体下落速度与重量成正比的理论,纠正了长达 1900 年的错误结论^[6]。

2. 理论科学阶段

实验科学受限于实验条件、观察误差等因素,往往难以得出完全准确的结论。随着科学的进一步发展,科学家开始运用数学、几何、物理等理论工具构建模型,解释自然现象。牛顿的三大定律、爱因斯坦的相对论等理论成果都是这一阶段的杰出代表。理论科学为我们提供了更深入、系统的认识自然世界的方式,但同样存在着局限性和不确定性。

3. 计算科学阶段

1946 年,第一台计算机 ENIAC 的诞生标志着计算科学时代的开始^[7]。计算科学主要通过计算机模拟和其他计算形式解决科学问题,计算机的快运算、高精度、大存储和可重复性等特点使其成为科学研究的有力工具。从天气预报到基因测序,从物理模拟到经济预测,计算科学在各领域都发挥着越来越重要的作用。然而,计算科学仍然受到算法复杂度、计算资源等限制,难以处理海量数据和复杂系统。

4. 数据密集型科学阶段

大数据时代,数据密集型科学成为新的研究范式,数据也成为科学研究的核心驱动力。与传统的实验、理论和计算科学相比,数据密集型科学更注重对数据的深度分析和挖掘,从中发现新的规律、趋势和模式。

过去,由于数据获取和存储的限制,很多研究只能基于有限的数据样本进行分析。然而,随着大数据技术的发展,研究者可以获取海量数据,使研究的深度和广度得到极大的拓展。同时,大数据改变了科学研究的方法。传统的科学研究方法往往依赖假设驱动的实验设计,而大数据则使数据驱动的研究成为可能。通过对大数据进行挖掘和分析,研究者可以发现新的现象和规律,进而提出新的科学问题和研究假设。此外,大数据还促进了科学研究领域的交叉融合。不同领域的数据在大数据的框架下得以整合,使研究者可以从多个角度对同一个问题进行分析,进而得出更全面、更深入的结论。这种交叉融合不仅有助于推动科学研究的全球化和协同化,还有助于培养具有跨学科背景的复合型人才。然而,大数据也带来了一些挑战。例如,数据质量问题、隐私保护问题及数据分析技术问题等。因此,在利用大数据进行科学研究时,还要谨慎对待数据的选择和处理,确保研究的准确性和可靠性。

1.4.2 大数据对思维方式的影响

牛津大学教授、著名的大数据专家维克托·迈尔·舍恩伯格博士深入探讨了大数据时代思维方式的三种转变：全样而非抽样、效率而非精确、相关而非因果^[8]。

1. 全样而非抽样

在大数据时代之前,受限于数据存储与处理能力,科学分析通常依赖抽样方法。这种方法通常从大量数据中提取一部分作为样本,通过对样本进行分析推测整体数据的特性。然而,随着大数据技术的兴起,特别是分布式文件系统和分布式数据库的应用,数据存储能力理论上达到了近乎无限的程度。同时,MapReduce等分布式并行编程框架极大提升了处理海量数据的能力。因此,在大数据支持下,科学分析可以直接针对全集数据进行,实现快速而全面的分析。

2. 效率而非精确

过去由于抽样分析的局限性,精确性成为科学分析的首要追求,这是因为抽样分析中的微小误差放大到整体数据时,可能造成巨大误差。因此,传统数据分析方法着重于提高算法的精确性。然而大数据时代全样分析避免了误差被放大问题,因此高精确性不再是首要目标。反之,大数据分析强调“秒级响应”,即快速处理海量数据并及时提供结果,以保持数据价值,从而将效率置于核心位置。

3. 相关而非因果

在传统数据分析中,解释事物的发展机理(因果关系)或预测未来事件是主要目的。但大数据时代的关注点从“因果”转向了“相关”。例如,电商平台系统会根据购物行为显示相关商品推荐,而非解释为什么购买特定商品的客户还会购买其他商品。大数据分析更多地关注在海量数据中发现相关性,而非深入挖掘因果关系。

1.4.3 大数据驱动的智能制造科学范式

在数据科学的范式下,制造系统的运行分析与决策方法也将发生变革。围绕制造系统的效率、质量与稳定性等性能指标,“关联+预测+调控”的决策新模式正在形成,如图 1-1 所示^[9]。关联是指通过数据的关联分析,量化数据间的影响机理,从数据角度探索其在制造系统运行中的关系。预测是指在关联分析的基础上,进一步描述数据与系统性能指标间的内在关系,揭示系统性能指标的演化机理。调控是指在预测的基础上,针对具体的业务进行优化调控,从而使系统性能满足要求。

在数据科学范式的发展下,大数据驱动的智能制造理论方法体系进一步完善,按照数据科学的方法论体系,可分为数据融合处理、关联分析、性能预测与优化决

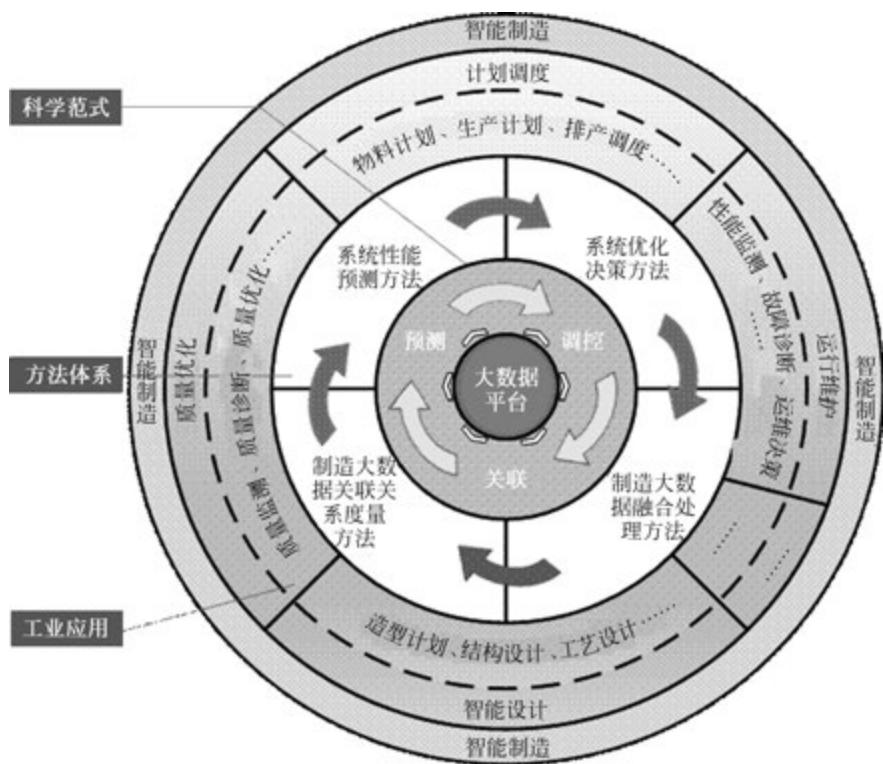


图 1-1 大数据驱动的智能制造方法与应用

策 4 部分。

(1) 数据融合处理。首先通过制造大数据融合处理方法对制造系统运行过程中产生的海量、高维、多源异构、多尺度和高噪声制造数据进行多级过滤、清洗去噪、建模集成与多尺度分类等操作,为制造系统的关联、运行分析与决策提供可靠、可复用的数据资源。

(2) 关联分析。在数据融合的基础上,针对产品、工艺、设备、系统运行等制造数据相互影响呈现的复杂耦合特性,通过制造大数据关联关系度量方法对工艺参数、装备状态参数等制造数据进行关联分析,并利用复杂网络等理论度量制造数据之间的关联程度和相关系数,挖掘影响车间性能指标的相关参数。

(3) 性能预测。在获取车间性能指标影响因素后,通过智能车间性能预测方法分析车间制造系统内部结构的动态特性与运行机制,从海量制造过程数据中学习与挖掘车间运行参数与车间性能的演化规律,实现车间性能的精确预测。

(4) 优化决策。在对车间运行情况进行分析预测后,将车间性能的预测值与目标决策值进行实时比对,运用智能车间运行决策方法对广泛存在的动态扰动条件下的关键制造数据进行定量调整,实现车间性能动态优化与决策,使制造系统始终保持最优稳定运行。

1.5 大数据与新质生产力

在信息化高速发展的今天,大数据以海量的数据资源、高效的处理能力和精准的分析能力,成为推动社会前进的关键力量,深刻改变着传统产业的运营模式,催生出新质生产力的蓬勃发展。

以知名电商平台淘宝为例,其通过大数据技术深入挖掘用户的行为数据,不仅实现了个性化商品推荐,还精准预测了市场趋势,从而快速调整商品结构和营销策略。这样的革新不仅提升了用户的购物体验,还显著提高了企业的市场竞争力。不仅如此,淘宝还借助大数据深度剖析消费者需求,推动了供应链的优化和库存的精细化管理,有效降低了资源浪费和环境污染,在追求经济效益的同时,积极履行社会责任。例如阿里巴巴的智慧城市项目,通过大数据技术整合城市中的各种资源信息,实现了交通、医疗、教育等公共服务领域的智能管理。在交通领域,该项目能实时监测和分析交通流量,预测交通拥堵,并优化交通信号控制,从而提升道路的通行效率。这不仅缓解了城市交通压力,也提升了市民的出行体验。同时,智慧城市项目还通过数据分析为城市规划提供科学依据,推动了城市的可持续发展。

大数据的应用打破了信息孤岛,促进了信息的共享与开放,实现了数据资源的合理配置,推动了社会生产力的深刻变革,使生产方式更智能化、精细化。面对技术发展带来的就业结构变化、数字鸿沟扩大等挑战,我们要不断调整思维方式和价值观念,以适应新技术带来的挑战和机遇。

总体而言,大数据作为新质生产力的代表,正以其独特的魅力和深远的影响力引领社会生产力的革新与发展。我们应充分认识大数据技术的价值和意义,积极拥抱新技术,不断提升个人素质和能力,为推动社会进步和发展贡献智慧和力量。同时,也要关注技术发展背后的社会责任问题,推动技术与社会、环境和谐发展。

本章小结

本章介绍了智能制造与工业大数据的演进历程,凸显了数据在制造业中的关键作用,以及大数据技术给制造业带来的深刻变革。首先,给出工业大数据的定义,强调其价值属性和产权属性。其次,深入分析工业大数据的六大特性,包括规模性、多样性、高速性、多来源、多维度和多噪声,并以半导体制造企业的晶圆数据为例,详细列举了不同数据类型及其特点。再次,探讨了大数据对科学研究、思维方式和智能制造科学范式的影响。最后,通过引入思政内容,展现了智能制造与工业大数据的发展不仅是技术革新的胜利,更是社会主义核心价值观在现代化进程中的具体体现。通过本章内容,希望激发读者对智能制造与工业大数据的深入思

考,并在制造业发展中充分利用大数据技术的探索,为实现中华民族的伟大复兴贡献智慧和力量。

习题

1. 智能制造和工业大数据是什么?它们之间的关系如何?
2. 工业大数据的特点有哪些?能否举例说明?
3. 大数据技术对当前社会产生了哪些重大影响?请结合实例进行回答。
4. 在当前背景下,智能制造和工业大数据面临哪些挑战?如何有效应对这些挑战?

参考文献

- [1] 周济. 智能制造——“中国制造 2025”的主攻方向[J]. 中国机械工程, 2015, 26(17): 2273.
- [2] 马士华, 林勇, 陈志祥. 供应链管理[M]. 北京: 高等教育出版社, 2003.
- [3] 黄双喜, 范玉顺. 产品生命周期管理研究综述[J]. 计算机集成制造系统, 2004, 10(1): 1-9.
- [4] Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition and productivity[M]. Washington, DC: Mckinsey Global Institute, 2011.
- [5] 王元卓, 靳小龙, 程学旗. 网络大数据: 现状与展望[J]. 计算机学报, 2013, 36(6): 1125-1138.
- [6] Buchwald J Z, Fox R. The Oxford handbook of the history of physics[M]. Oxford: OUP Oxford, 2013.
- [7] 石钟慈. 第三种科学方法: 计算机时代的科学计算[M]. 北京: 清华大学出版社, 2000.
- [8] Mayer-Schönberger V, Cukier K. Big data: A revolution that will transform how we live, work, and think[M]. Boston: Houghton Mifflin Harcourt, 2013.
- [9] 张洁, 汪俊亮, 吕佑龙, 等. 大数据驱动的智能制造[J]. 中国机械工程, 2019, 30(2): 127-133, 158.

第2章

大数据预处理技术

2.1 引言

大数据不仅改变了人们的生活方式,还为企业和组织带来了前所未有的机遇和挑战。然而要充分挖掘大数据的价值,首先要对数据进行有效的预处理。本章将介绍大数据预处理的基本概念、方法和技巧,帮助读者掌握数据预处理的核心技能。

数据预处理是大数据分析的基础,它涉及对原始数据的清洗、转换和整合,以便为后续的数据分析和挖掘提供高质量的数据输入^[1]。数据预处理的主要对象包括结构化数据和非结构化数据。在数据结构方面,预处理主要包括数据结构调整、聚合、连接、拆分和扩展等操作。在数据内容方面,预处理则包括归一化、异常值清洗、数据采样和填充等技术。这些操作可以帮助我们从原始数据中提取有用的信息,将其转换为适合分析的形式。

2.2 数据预处理的概念

2.2.1 数据预处理的定义

在实际工程应用中,获得的数据可能包括缺失值、重复值等问题,因此在进行数据分析之前,需要对数据进行预处理。这个过程旨在减少数据分析和建模过程中的错误和偏差,提高数据的质量和可靠性。值得注意的是,数据预处理并没有固定的标准流程,通常需要根据不同的任务和数据集特性制定相应的处理策略^[2]。简言之,数据预处理是对原始数据进行一系列必要的操作(清理、集成、转换、离散和规约等)的过程。