# 大数据与地质学的未来发展

吴冲龙[1,2],刘 刚[1,2]
WU Chonglong[1,2], LIU Gang[1,2]

1. 中国地质大学(武汉)计算机学院,湖北 武汉 430074;
2. 智能地学信息处理湖北省重点实验室,湖北 武汉 430074
1. *School of Computer Science, China University of Geosciences, Wuhan* 430074, *Hubei, China;*
2. *Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan* 430074, *Hubei, China*

**摘要**:地质学定量化是地质学自身发展臻于成熟的重要标志。地质学家们经过长期艰难的探索和尝试,扫清了许多障碍并取得了令人瞩目的进展,但并未越过定性描述和不确定性门槛。在人类进入信息化和大数据时代的今天,地质学家们发现并找到了越过定量化之门的捷径。在以大数据和数据密集型计算为基础的第四范式支配下,地质学家有可能突破各种主客观因素的限制,使地质学进入更全面的定量化发展阶段,并取得地质科学原理和规律方面的新发现。换言之,在地质信息学的引领和支撑下,地质学将在新世纪得到快速发展。地质学家需要逐步建立与第四范式相适应的新地质科学观,即以查找和揭示隐藏于大数据中的多种地质要素关联关系为主要目标,然后在此基础上追究成因关系。

**关键词**:定量化;信息化;大数据;地质科学;地质信息科学;地质信息学
**中图分类号**:P628 **文献标志码**:A **文章编号**:1671-2552(2019)07-1081-08

**Wu C L, Liu G. Big data and future development of geological science. *Geological Bulletin of China*, 2019, 38(7):1081-1088**

**Abstract**: Quantification of geology is an important indicator of the development of geology itself. After long and hard explorations and attempts, geologists have overcome many obstacles and made remarkable progress, but they have not crossed the threshold of qualitative description and uncertainty. Today, when human beings enter into the era of informatization and big data, geologists have discovered the shortcut to quantification. Under the domination of the Fourth Paradigm which is based on big data and data-intensive computing, all kinds of limitations from both subjective and objective factors may be overcome by geologists. Geological science will be brought into a more comprehensive stage of quantitative development, and new discoveries in principles and laws will be acquired. In other words, guided and supported by geoinformatics, geology will develop rapidly in the 21st Century. Geologists need to gradually establish a new concept of geological science that is compatible with the Fourth Paradigm, that is, to find and reveal the relationship between various geological factors concealed in big data as the primary goal, and then to identify the genetic relationships on such a basis.

**Key words**: quantification; informatization; big data; geological science; geological information science; geoinformatics

## 1 从地质学的定量化说起

地质学定量化的涵义十分丰富,并且随着科学研究范式的发展而不断拓展和充实。它既非数学方法的简单应用,也非数学方法的组合应用,而是采用数学方法、数学模型和计算工具,对地质现象、

**收稿日期**:2019-03-20;**修订日期**:2019-06-18
**资助项目**:国家自然科学基金项目《城市地质环境时空透视与大数据融合关键技术》(批准号:U1711267)、《基于RASC优化模型和空间序列分析的多参数定量地层学综合方法研究》(批准号:41172300)、地学长江计划核心项目《多尺度三维地质体模型库关键技术与长江流域应用示范》(编号:CUGCJ1810)
**作者简介**:吴冲龙(1945- ),教授,博士生导师,从事矿产资源勘查和地质信息科学与技术领域的研究与教学工作。
E-mail:804077427@qq.com

地质过程和地质规律进行挖掘、演绎、推理和表达。

地质学定量化是地质学家长期追求的目标，也代表了地质科学自身的发展方向[1]。换言之，地质学定量化是地质学理论研究和技术应用的发展趋势。在地质学由经验上升到理论的飞跃中，不但需要更多、更好的探测技术与分析方法，也需要数学的介入和定量化手段的支持。从地质体和地质现象的化学、物理学和几何学测量、分析，到各种地质变量的时空变化规律研究，再到矿产资源和地质环境的质量和数量评估，地质学不断向前发展，还催生了地球化学、地球物理学和数学地质学。这些新学科的出现和发展，是地质学定量化的重要引擎，使地质学从早期单纯的现象描述进入了“物质科学”范畴，随后又使之拓展到了“能量科学”范畴。目前，随着地矿工作信息化的推进和地质信息科技的兴起，地质学已经涉足“信息科学”范畴。这一切，虽然使地质学产生了质的变化，其定量化成分也迅速增加，但总体上仍呈现出定性的描述性科学形态。

主要受制于三大因素：①地质作用和地质过程的初始条件不同、影响因素众多，导致其发展演化轨迹有确定性(机械性)、不确定性(随机性)和确定随机性(混沌)的，但以确定随机性为主[2]；②地质体的主要部分深埋于地下，而地质作用所涉及的时空域巨大，其现象极复杂，存在结构信息不完全、关系信息不完全、演化信息不完全和参数信息不完全的特征[3]；③地质科学研究和资源勘查工作的数据采集和处理方式，有定性、定量和半定量的，但以定性为主，其数据类型有结构化、非结构化和半结构化的，但以非结构化(文字、图件等)和半结构化的为主。

在阻碍地质学定量化的上述三大因素中，①是地质作用和地质过程内在本质的反映，②是地质作用和地质过程外在形态的表现，而③是地质作用和地质过程科学研究的方式。显然，①和②为客观因素，③为受客观因素制约的主观因素。这些因素导致对地质作用和地质过程的认知存在极大的不确定性。目前，多数地质作用和地质过程，例如构造作用、岩浆作用、沉积作用、变质作用、成矿作用、成煤作用、油气成藏作用等，都无法用单纯的确定性数学模型来描述和演绎，也无法用单纯的随机性数学模型描述。这也是迄今为止，地质学仍属于描述性科学范畴的原因。显然，要真正地实现地质学定量化，首先必须正视并排解这3大因素。

要排解第一因素，需采用复杂性科学的理论、方法及其数学模型。然而复杂性科学刚刚兴起，许多基本理论和方法问题还没有答案，通过复杂性科学来实现地质学定量化，其道路可能是十分漫长的。要排解第二因素，则需使地质体及其各种现象裸露于地表，或者得到物探、化探和遥感手段全面、清晰的揭示。使地质体及其各种现象完全裸露于地表是不可能的，要用物探、化探和遥感技术全面、清晰地揭示它们，其困难巨大，道路也将是十分漫长的。不能坐等复杂性科学和物探、化探、遥感科技的发展，而应当借助地质信息科技，在地质数据获取方式方法和研究方式方法上下功夫，即充分发挥第三因素的作用，促进地质学的定量化发展。具体而言，就是借助基于大数据和密集型计算的“第四范式”[4]，调整对地质学定量化的认识，即改变单纯追究因果关系的定量演绎推理科学观，建立以揭示关联关系为主要目标的新科学观[5]，在此基础上再追究因果关系。

## 2 地质信息学的发生发展

信息技术的广泛应用和信息化时代的到来，给予了地质数据获取方式和地质科学研究方式改进的条件和可能。实现地质数据采集、管理、处理、应用的数字化和信息化，可以使地质体和地质现象观察的定性描述，转化为可以用计算机采集、存储、管理、处理和应用的符号运算，从而在很大程度上改进地质数据的获取方式和地质学的研究方式。就此意义而言，地质工作的数字化和信息化本身就是地质学定量化的基础。为了加速地质学定量化进程，应当重视并加强地质矿产调查与勘查工作的数字化和信息化进程。

地质环境调查与矿产资源勘查工作信息化，简称地质工作信息化，是指在基层勘查单位采用信息系统对传统的地质工作主流程进行了充分改造，实现全程计算机辅助化，数据在各道工序间流转顺畅、充分共享[6]。这里面包含3个相互密切关联的内容：①建立以主题式地矿点源数据库(包括空间数据库和属性数据库)为基础的共用数据平台，有效地避免了系统内出现大量的数据冗余；②进行基于“多S”(DBS、RS、GPS、GIS、CADS、MIS、DDS、OAS)的技术集成、网络集成、数据集成和应用集成，使各部

分有机结合、相互衔接,数据在其中流转顺畅、充分共享;③利用信息系统技术对地质工作主流程进行充分改造,实现从野外数据采集到室内整理、处理、编图和三维可视化建模,再从地质分析、矿产预测、资源评价到成果保存、管理和使用的全程计算机辅助化。这3项内容相辅相成,既是推进地质工作信息化、建立和完善地质信息技术体系所必须的工作内容,也是衡量勘查单位信息化程度的基本标志。

为了满足地质工作信息化的需求,并解决面临的各项问题,一门崭新的边缘学科——地质信息科学正在逐步形成[7]。其核心部分是地球信息学或称地质信息学(GeoInformatics)。赵鹏大院士称其为“信息地质”,与“数学地质”一起并称为“数字地质”[8]。这不是计算机和信息技术在地质调查和矿产勘查领域的简单应用,而是一个研究地质信息本质特征及其运动规律和应用方法的综合性学科领域。地质信息科学既是地球信息科学的重要组成部分,也是地球信息科学与地质科学交叉的边缘学科(图1)。

地质信息科学的发生和发展,是地球信息科学与地质科学结合的产物。其内部条件是地质学定量化和地矿勘查信息化的自身需要,而外部条件是计算机科学和地球空间信息科学的兴起和发展。其研究对象是岩石圈的地质信息,理论框架的核心是地质信息学,包括地质信息的本质、运动规律、传输机制、信息流的形成机理等。其方法论体系包含主题信息管理法、信息分析综合法、行为功能模拟法和系统整体优化法[7]。其技术体系由地质数据采集、地质数据管理、地质数据处理、地质图件编绘、地质数据挖掘、地质过程模拟、地质资源预测、地质资源评价、地质信息传播及其多S集成化技术组成。已有的实践表明,地质信息科学的形成和发展,将使地质科学研究和勘查工作的数据采集和处理方式发生根本改变,为结构化、非结构化和半结构化的多源多类异质异构数据的有效融合、处理提供方便条件。不仅如此,由于信息本身就是用来消除随机不确定性的东西,地质信息科学与信息技术的发展,还能为排解阻碍地质学定量化的2大客观因素提供方便而有效的途径。

迄今为止,地质信息科学的理论、方法和技术体系框架已初步形成[9]。地质信息科学理论体系框架的结构和组成,主要体现在研究对象、任务、内容等方面。其研究对象是岩石圈(含地壳)的地质信息,主要任务是通过研究地质信息的本质,探索人类获取、分类、变换、传播、存贮、管理、处理、解译、表达和利用各种地质信息的一般规律,而主要研究内容包括:①地质信息的本质、特征(结构、性质)和度量(基准与标准);②地质信息的运动规律,即研究地质信息在壳幔之间、水岩之间和地质体之间的传输机制、过程、增益与衰减及信息流的形成机理;③地质信息的产生、表现和认知的一般规律,以及不确定性与可预见性;④地质信息的采集、分类、变换、加工、整理、存储、管理、统计、分析、处理、解译、反演、正演、建模、模拟、表达和可视化的理论、方法和技术;⑤地质信息传播、交流与社会化服务的途径、方法和技术;⑥利用地质信息进行地质资源和地质环境管理、预测、评价、决策、开发、保护,并实现最优化的原理和方法;⑦地质信息市场、信息商品、信息产业的特征、结构、功能及其发展机制等。

在地质信息科学方法论体系中,主题信息管理

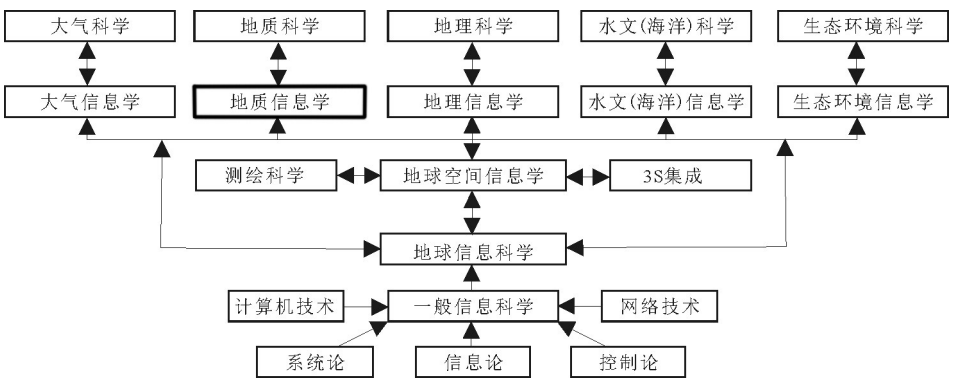


图1　地质信息科学(地质信息学)的学科地位图解[7]

Fig. 1　Illustration of the academic status of geological information science (geoinformatics)

法指建立以主题式点源地矿数据库为核心的信息系统，开展数据集成、技术集成、应用集成和网络集成，实现地质数据的充分共享；信息分析综合法指通过地质数据的分析、挖掘、反演、正演、综合，提取健全而有用的信息，并建立相应的地质模型，描述地质体、地质现象、地质过程、地质资源和地质环境，乃至整个地质系统；行为功能模拟法指利用各种地质过程模拟和预测评价系统，对各种地质模型进行解算、模拟、评价、预测和决策分析，揭示地质系统演化的影响因素及其相互间的控制与反馈控制关系，进而总结地质系统演化的内在规律；系统整体优化法指从部分与整体之间、整体与外部环境之间、资源开发利用与生态环境保护之间、人与自然的协调发展之间等相互联系中，系统、全面地考察研究对象，进行多目标的最优化决策，将知识转化为可操作的开发、保护和治理能力和行为，达到可持续发展的目的。

地质信息科学的技术体系发端于20世纪60年代初。早期靠引进统计学、地质统计学、遥感(RS)图像处理技术、图件辅助设计(CAD)技术、数据库系统(DBS)技术、地理信息系统(GIS)技术、空间定位(GPS)技术，以及GPS、RS、GIS集成技术等。这些技术基本来自于测绘和地理信息技术领域，由于地质对象的特殊性和复杂性，所引进的各种信息技术成果都经过了改造和再开发。20世纪90年代以来，地质信息科技领域的广大研发人员，经过了长期的艰苦努力，建立了以三维可视化平台为基础、以点源地矿数据库为核心，可实现地质资源勘查-评价-预测全流程计算机辅助化，并能支持科学研究的地质信息系统，形成了较完善的地质信息技术体系。

## 3 科学研究新范式的转换

随着大数据时代的到来，出现了新的科学研究思路和方法论，使得科学研究有可能跨越实验、理论、模拟3个范式，进入以大数据为基础、以数据密集型计算为手段的第四范式[4]。与第一、二、三范式相比，第四范式有许多特点和优势(表1)，能有效规避前述3项不利因素，推进地质学定量化。科学方法论的中心问题是：①如何发现和证明？②发现和证明定律和定理要依据什么推理规则？对于不同的科学研究范式，做法和规则不同。在19世纪中叶前，第一范式的简单实验和第二范式的固定推理规则，能容易地发现和证明科学真理。随着科学和技术的发展，科学发现和证明越来越困难，科学方法论的价值也因此显得越来越重要。怎样获得和接受定律或定理，成为科学方法论的中心问题，同时也催生了科学研究的第三范式——通过计算机进行动力学数值分析和过程模拟仿真，从而揭示、认识并证明自然和社会规律及其发展过程的计算范式。

这3个科学研究范式，在地质学研究和矿产资源勘查领域都有广泛的应用，但在各个分支学科的应用情况不均衡。目前，在地球物理学、地球化学、数学地质学、矿物学、结晶学、探矿工程学等领域，以第二范式为主导；在构造地质学领域，第一范式和第二范式参半；在岩石学、沉积学、地层学、古生物学、矿床学、油气地质学、煤地质学、工程地质学、水文地质学、区域地质学、矿产资源勘查学等领域，则以第一范式为主导。至于第三范式，目前较大规模的应用，只在于个别学科领域的某些研究方向上，例如石油地质学的盆地模拟、油气成藏过程动力学模拟和水文地质学的地下水动力学模拟。因此，就地质学研究总体而言，所采用的研究范式仍以第一范式为主，即以定性现象观察分析为主兼有定量测算。需要说明的是，传统概念中的所谓定量化科学研究范式，实际上是指第二范式。

随着地质探测技术的不断进步，地质科学研究从纯现象描述进入“物质科学”和“能量科学”范畴，所积累的数据越来越多，近几十年来的地质工作信

**表1 科学研究的4个范式及其特征比较**

**Table 1 Four paradigms of scientific research and the comparison of their characteristics**

| 序号 | 范式 | 属性 | 方式 | 方法 | 参照 | 目标 | 认知 | 结果 |
|---|---|---|---|---|---|---|---|---|
| 第一 | 经验 | 描述 | 定性+定量 | 综合 | 概念模型 | 协调 | 模糊 | 功效 |
| 第二 | 理论 | 演绎 | 定量 | 分析 | 理论模型 | 因果 | 清晰 | 规律 |
| 第三 | 计算 | 仿真 | 定量 | 模拟 | 数值模型 | 过程 | 系统 | 模式 |
| 第四 | 数据 | 本体 | 定量+定性 | 挖掘 | 无模型 | 关联 | 透明 | 智能 |

息化发展,更使多源多类异质异构地质数据呈爆炸性的增长。在地质科学大数据面前,数据分析和处理能力不足的特点凸显出来,多数地质数据没有得到及时和妥善处理,许多隐含信息未被识别、提取和利用。地质科学研究面对的挑战不仅来自数据量方面,而且来自数据类型方面。地质学向"信息科学"范畴拓展,正是为了应对挑战并从地质大数据中高效感知并提取信息。大数据的特点在于所容许的数据集合"不是随机样本,而是全体数据",所注重的数据品质"不是精确性,而是混杂性",所揭示的数据内涵"不是因果关系,而是关联关系"[10]。显然,这 3 个特点,也正是大数据的三大优势。以数据密集型的第四范式为主导、多种范式结合的科学研究,能够把这 3 个优势充分发挥出来,突破前述主、客观因素的限制,促进地质学定量化并取得地质科学原理和规律的新发现。

以矿产资源定量预测评价为例,阻碍定量化的主、客观因素可归结为:①可靠和普适性的地质作用机理、因果关系、动力学等理论模型缺乏;②受已有知识模型束缚较多、采样点随机、数据内容主观限定、样品空间狭小、精确数据少;③海量的多源多类异质异构数据难以一体化存储、管理和利用;④在多数情况下,只能凭借少量观测数据和定性模式进行分析、类比和预测。这些不足,正好可以用大数据的 3 个优势弥补,因此,以第四范式为主导的多范式结合,可以有效地突破主、客观因素的限制。

科学研究范式从经验归纳到理论演绎,再到计算模拟,现在又发展到基于大数据和数据密集型的第四范式,每一个发展阶段都有各自的特征和范例,但相互之间是承接和补充的关系,是不可偏废的[11]。在地质科学领域中实践并发展第四范式,关键在于开发多源多类多维异质异构大数据处理和挖掘的软件工具。

## 4　基于大数据的新地学观

科学研究范式是科学研究赖以运作的理论基础和实践规范,是从事某一学科研究的科学家群体共同遵从的世界观和行为方式[12]。其基本原则可以在本体论、认识论和方法论 3 个层面表现出来,分别回答事物存在的真实性问题、知者与被知者之间的关系问题,以及研究方法的理论体系问题。这些理论和原则,对特定的科学家共同体起规范的作用,协调他们对世界的看法及其行为方式。在第四范式支配下,地质科学家或地球科学家需要逐步建立与该科学范式相适应的新地学观。这种新地学观以查找和揭示多种地质要素之间的关联关系为主要目标,然后在关联关系中探寻因果关系,其研究方法以"全体"数据为依据,用数据驱动,通过无模型的大数据挖掘发现新知识。这种新地学观与以追求并阐述因果关系为目标、以假说或模型为依据的传统地学观有显著的差别。

如何理解这个问题? Anderson[13]认为,"数据洪流使(传统)科学方法变得过时","获得海量数据和处理这些数据的可能性,提供了理解世界的一条完整的新途径。PB 级数据让我们能够说:关联关系已经足够,……关联关系取代了因果关系,没有一致性的模型、统一的理论和任何机械式的说明,科学也可以进步"。Norvig[14]甚至认为,"所有的模型都是错误的,没有模型你也可以成功,……PB 级数据能够做到没有模型和假设也可分析数据。只要把相关的数据丢进计算机群,就可采用数据挖掘算法,发现过去发现不了的新模式、新知识甚至新规律"。二者的观点在当前大数据研究领域中具有代表性,虽因过于偏激而遭到科学界的质疑,却可作为上述大数据和第四范式 3 个特点和优势的说明。

这里需要解决一个问题,即关联关系是否是科学理论的重要组成? 以往的物理学、化学、天文学及其他科学的基本定理、定律,都是基于小数据的因果分析而得到的,人们担心大数据关联分析会陷入"知其然而不知其所以然"的境地。然而,有证据表明,对于自然界的复杂巨系统,因果关系是隐藏在系统中的。各个组成部分之间相互影响,互为因果关系,传统的因果分析实际上也难以奏效。况且,近期的一些理论物理学的重要成果,例如欧几里德量子引力学,也并不包括因果规律[15]。

各种地质作用系统,包括构造作用、岩浆作用、沉积作用、变质作用、古生物演替、成矿作用、成煤作用、成藏作用、成灾作用等系统,都是复杂的巨系统。显然,查明其中的关联关系,并且基于这种关联关系开展地质演化规律研究,以及金属资源、油气资源、煤炭资源和地质灾害预测,比单纯追求其因果关系更重要。许多实际研究结果表明,关联关系正是寻觅因果关系的基础。这里有一个基于大数据和第四范式查明关联关系,进而揭示了因果关

系的实例，即国际地球化学学会《Geochemical News》发布的2017年度十项最有影响力的地球化学研究进展的第九项："计算机分析地质学基础数据可揭示古海洋的地球化学性质"。这项研究采用文本大数据挖掘和机器学习方法，读取并分析了数十年来300多万份地学论文和专著，发现叠层石发育与海水的碳酸盐浓度有高度相关性，从而揭示，常见的叠层石在晚古生代末走向衰亡的真正原因是海水中碳酸盐浓度降低。由此，解决了长期困扰古生物学界的重大生物演化疑难问题。大数据和第四范式在地质科学研究领域的应用刚刚开始，随着时间的推移，重要成果将不断涌现出来。

大数据的基本功能是预测，因此地质科学大数据的主要用途，应当包括矿产资源、地质环境和地质灾害及其时空分布的预测。在大数据和第四范式支配下，进行地质科学研究的主要方法，就是基于无模型的数据挖掘发现地质资源与地质环境发展变化的新知识。数据挖掘的概念和方法，起源于从关系数据库和数据仓库中发现知识，随后拓展到文本和空间数据库中[16]。由于地质领域具有海量文本数据和空间数据，因此得到了广泛的应用。在未来的研究中，需要妥善解决全体数据与抽样数据、有模型与无模型、关联关系与因果关系的一体化存储、管理、处理、应用和可视化问题，特别是其深度挖掘和广度聚联方法。

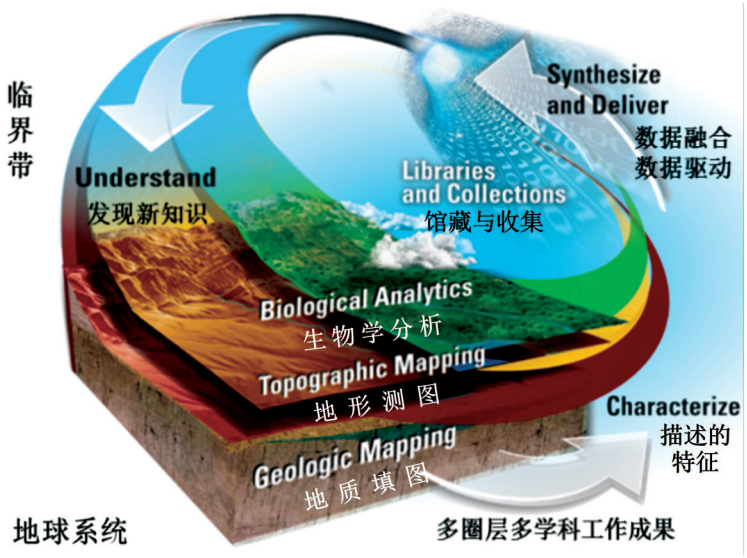


图2 美国地质调查局基于大数据的核心科学体系[19]

Fig. 2 Big data-based core scientific system of the US Geological Survey

目前，科学研究第四范式的概念已经得到广泛的认同，学术界多将其理解为"大数据范式"。这个崭新的科学研究范式，也已经影响到大数据充斥的地球科学研究领域[17]。信息科学技术和互联网、物联网等的快速发展，不但提升了数据处理的速度，将大数据的处理进程推向实时，而且拓展了数据处理的类型，增强了对快速积累的高维复杂非结构化数据、半结构化数据和结构化数据的一体化处理能力。地质信息科学和技术的进步，正是对地质科学乃至整个地球科学领域迫切需求的响应。地球科学的分析、计算和可视化方法远远落后于创造数据的能力的状况，亟待改变。

实际上，科学研究第四范式的提出，也标志着数据密集型的科学革命启动和新科学共同体形成。美国地质调查局总结了建局近150年的经验及当前的新形势、新挑战，在连续2次制定的科学战略及10年(2010—2020和2013—2023)规划中[18-19]明确提出了基于大数据的核心科学体系(图2)。该科学体系以岩石圈、水圈、生物圈和大气圈构成的地球系统为对象，以传统地质学、地理学和生物学为基础，采用数据密集型工作方法，即基于大数据的科学研究第四范式，在数据高效管理和调度的条件下，实现多源多类异质异构大数据融合，进行多样化的地球科学数据挖掘和相关的知识发现，促进交叉和综合学科的发展，以解决复杂的地球科学问题和社会问题。该核心科学体系的战略愿景，是建立基于大数据和第四范式的模块式"地球系统科学"框架，把气候变化、土地利用、生态系统、能源和矿产、环境、自然灾害和水资源7个科学战略领域无缝整合起来，更有效地解决各种问题。其任务是将数据、科学、技术、方法及模型，在恰当的时空尺度中组织起来，驱动科学的合成，促进对整个地球系统运转的理解，增强对临界带的科学认识和决策支持，包括：①填制四维地质图并构建"玻璃地球"；②实现地质信息采集管理和发布现代化；③发展数据挖掘、可视化和信息处理技术；④以数据驱动科学，推进对合成数据产品的应用。其数据管理、处理和应用，由下而上分为3个层次：第一层是数据管理

层,包括数据采集、入库、管理;第二层是数据操作层,包括数据挖掘、数据分析等;第三层是知识发现层,包括对地球系统的认知、理解、解释等。其战略目标是:①为临界带描述、认识及制图提供调查研究的途径;②通过科学服务扩展美国地质调查局调查成果的应用;③进行科学分析与合成,提高信息覆盖面、科学质量、实用性和及时性。

这种基于大数据和第四范式的核心科学体系、战略愿景和战略目标,在一定程度上体现了大数据时代的新地学观和地球科学的发展方向,可以借鉴。

## 5　地质信息学的引领作用

地质学在由经验上升到理论的过程中,得益于数学、化学和物理学的支撑。从地质体和地质现象的几何尺寸、化学成分和物理特征的测量、换算、分析,到各种地质变量的时空变化规律统计和矿产储量计算,再到各种高精尖化学分析手段和物理探测手段的相继引入,促使地球化学和地球物理学的形成和发展。而地球化学和地球物理学的形成和发展,反过来又促进了地质学的发展。20 世纪以来,地质科学的若干重大进步,包括对岩石矿物成分和来源的认识、对地球深部结构和成分的了解、对地壳中化学元素迁移聚集规律的认识、对地球动力学的认识、对岩石圈和地壳构造运动机制的判断,以及板块学说的兴起和由此引起的“地学革命”,都缘自于地球物理学和地球化学的贡献。

正如前面已经谈论过的,随着地球物理和地球化学理论、方法和技术的不断发展,以及各种微观超微观高新测试技术手段的相继出现,获取数据的手段越来越多,数据的类型越来越复杂、数据的维度越来越高、数据的数量也越来越庞大,以至于有了多源、多类、多量、多维、多尺度、多时态和多主题特征,地质数据呈现出一种爆炸的态势,成为名副其实的地质大数据[20]。为了从这些数据中获得更多的有用信息,以便深刻地认识地质体、地质现象、地质过程和地质规律,更好地利用和保护地质资源和地质环境,人们越来越多地求助于地质信息技术,从而催生了地质信息科学,推动了地质矿产勘查信息化的进程。地质工作信息化是地质学定量化的基础。地质信息科学与技术的形成和发展,突破了多源多类异质异构地质时空数据统合应用的障碍,给地质学的定量化发展提供了机遇和途径。特别是大数据理论、方法和技术的引进,对于突破采样随机性和样品空间狭小、大量非结构化和半结构化数据无法利用,以及可靠的作用机理、因果关系和动力学模型缺乏[21],仅凭少量观测数据和旧模式进行判断、预测等限制,无疑有极大的好处。

需要着重指出,地质科学问题具有显著的高维度、高复杂性和高不确定性特征,目前的一些研究仅仅开展单一的数据挖掘,或者单一的数据深度挖掘,而未考虑对整个研究对象时空域内的全体多源多类异质异构数据进行深度融合与广度聚联,更未对相关的跨界数据进行全面的采集、处理、融合和广度聚敛。这样做在特定领域的特定问题的研究和解决方面是可行的,但可能不适合于整个地质资源与地质环境的预测、评价,否则与原有的数学地质工作没有区别。显然,大数据工作方式或基于大数据的第四范式,应当具有更广泛和深刻的内涵和多样化的工作方式。

在大数据和数据密集型的第四范式支配下,有可能突破三大主客观因素的阻碍,加快地质学定量化的进程。在今天,谁能够获取、掌握、处理和利用地质大数据,并综合应用多种科学研究范式,谁就能取得地质科学的重大新发现。需要应对的挑战是实现结构化-半结构化-非结构化数据、大数据与小数据、混杂性数据与精确性数据、模型与数据、静态勘查模型与动态监测模型等的一体化存储、管理、处理和应用,实现相关关系与因果关系的统一。进行地质时空大数据统合利用,涉及一系列理论、方法和技术问题,其中包括:建立地质时空大数据的一体化空间参考体系,开展数据空间基准、时态、尺度和语义的一致性处理和数据融合;探索对各类静态地质勘查数据进行集成化、结构化、可视化转换,并与各类分布式动态地质观测数据进行一体化存储、管理的“玻璃地球”方式;研究地质时空大数据存储、智能处理、数据挖掘和云服务技术。

Watts[22]认为,借助于社交网络和计算机分析技术,21 世纪的社会科学有可能实现定量化的研究,从而成为一门真正的自然科学。我们也有理由推测,随着大数据时代的到来和地质工作信息化的发展,21 世纪的地质学有可能在地质信息学的引领下快速发展,在介入“信息科学”范畴的同时,跨入全面定量化阶段。

## 6 若干基本认识的归纳

(1)地质学定量化的内涵十分丰富,并且随着科学研究范式的发展而不断充实。它既非数学方法的简单应用,也非数学方法的组合应用,而是采用数学方法、模型和计算工具,对地质现象、地质过程和地质规律进行挖掘、演绎、模拟和表达。

(2)地质学定量化是一个漫长的过程,在各分支学科中的表现很不平衡,其发展阶段的划分不应仅以某些数学方法在个别分支学科的应用情况为依据,而应以在地质现象、地质过程和地质规律的挖掘、演绎、模拟和表达的总体状况为依据。

(3)为了满足地质工作信息化的迫切需求,一门崭新的边缘学科——地质信息科学(核心是地质信息学)正在形成。这是一个研究地质信息本质特征及其运动规律和应用方法的综合性学科领域,是地球信息科学与地质科学相结合的产物。

(4)大数据时代的到来,使地质学研究跨越了实验、理论和模拟3个范式,进入了基于大数据的数据密集型计算第四范式,综合采用4个范式可以突破各种主、客观因素的限制,促进地质学的定量化发展,并取得科学原理和规律的新发现。

(5)在大数据和科学研究第四范式的支配下,地质科学家需要逐步建立与该科学范式相适应的新地学观,即以查找和揭示多种地质要素之间关联关系为目标,以"全体"数据为依据,在数据驱动下通过深度挖掘和广度聚联来发现新知识。

(6)20世纪引领并支撑地质学发展的学科主要是地球物理学和地球化学;新世纪,引领并支撑地质学发展的学科,将是地质信息科学(地质信息学)。在基于大数据的数据密集型计算第四范式支配下,地质学有可能全面地向定量化方向迈进。

## 参考文献


[1]Merriam D F. Roots of quantitative geology[C]//Merriam D F. Down-to-earth statistics: solutions looking for geological problems New York: Syracuse Univ.. Geology Contribution, 1981,(8):1-15.

[2]赵鹏大, 孟宪国. 地质学的定量化问题[J]. 地球科学——中国地质大学学报,1992, 17(增刊): 51-56.

[3]吴冲龙, 张洪年, 周江羽. 盆地模拟的系统观与方法论[J]. 地球科学——中国地质大学学报, 1993, 18(6): 741-747.

[4]Gray J, Szalay A. eScience—A Transformed Scientific Method[C]// Presentation to the Computer Science and Technology Board of the National Research Council, Mountain View, CA, 2007.

[5]Hey T, Tansley S, Tolle K. The fourth paradigm: Data-Intensive Scientific Discovery[M]. Redmond: Microsoft Research, 2009.

[6]吴冲龙, 刘刚, 田宜平. 地矿勘查工作信息化的理论与方法问题[J]. 地球科学—中国地质大学学报, 2005, 30(3): 359-365.

[7]吴冲龙, 刘刚, 田宜平, 等. 论地质信息科学[J]. 地质科技情报, 2005, 24(3): 1-8.

[8]赵鹏大. 数字地质与矿产资源评价[J]. 地质学刊, 2012, 3: 225-228.

[9]吴冲龙, 刘刚, 田宜平, 等. 地质信息科学与技术概论[M]. 北京: 科学出版社, 2014.

[10]Mayer-Schonberger V, CuKier K. Big Data: A Revolution That Will Transform How We Live, Work and Think[M]. New York: Houghton Mifflin Harcourt Publishing Company, 2013.

[11]Hey T, Tansley S, Tolle K. Jim Grey on eScience:A transformed scientific method[C]//Hey T, Tansley S, Tolle K. The Fourth Paradigm: Data- Intensive Scientific Discovery. Redmond: Microsoft Research, 2009: xvii-xxxi.

[12]邓仲华 李志芳. 科学研究范式的演化——大数据时代的科学研究第四范式[J]. 情报资料工作, 2013, (4): 19-23.

[13]Anderson C. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete[EB/OL] (2019- 03- 20)http://www.wired.com/2008/06/pb-theory/. 2016.

[14]Norvig P. All we want are the facts, ma'am[EB/OL](2019-03-20) http://norvig.com/fact-check.html). 2009.

[15]李国杰,程学旗. 大数据研究: 未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J]. 中国科学院院刊, 2012,27(6): 647-657.

[16]李德仁, 王树良, 李德毅. 空间数据挖掘理论与应用(第二版)[M]. 北京:科学出版社, 2013.

[17]Goodchild M, Guo H D, Annoni A, et al. Next-generation Digital Earth[M]. Proceedings of the National Academy of Sciences, 2012: 1-7.

[18]Linda G, Belnap J, Goldhaber M, et al. Geology for a changing world 2010- 2020 Implementing the U.S Geological Survey science strategy: U.S. Geological Survey Circular 1369[M]. U.S. Geological Survey, 2011.

[19]Bristol R S, Euliss N H, Booth N L, et al. Science strategy for core science systems in the U.S. Geological Survey, 2013- 2023[M]. U.S. Geological Survey, 2012.

[20]吴冲龙, 刘刚, 张夏林,等. 地质科学大数据及其利用的若干问题探讨[J]. 科学通报, 2016, 61(16): 1797- 1807.

[21]Fabbri A G. Quantification and geology: methods of pattern defection and of integrating multi- disciplinary knowledge[J]. Netherlands: Enschede, l990:1.

[22]Watts D J. A twenty-first century science[J]. Nature, 2007, 445 (7127): 489.