论传统研究与大数据研究的关系(代序)

张　旗[1,2],朱月琴[3],焦守涛[3]
ZHANG Qi[1,2], ZHU Yueqin[3], JIAO Shoutao[3]

1. 中国科学院地质与地球物理研究所,北京 100029;
2. 中国科学院地质与地球物理研究所岩石圈演化国家重点实验室,北京 100029;
3. 中国地质调查局发展研究中心,北京 100037

1. *Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing* 100029, *China;*
2. *State Key Laboratory of Lithospheric Evolution, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing* 100029, *China;*
3. *Development Research Center, China Geological Survey, Beijing* 100037, *China*

中图分类号:P628　　**文献标志码**:A　　**文章编号**:1671-2552(2019)12-1939-04

Zhang Q, Zhu Y Q, Jiao S T. Discussion on traditional research and big data research. *Geological Bulletin of China*, 2019, 38 (12): 1939-1942

地质学的研究现状如何？每个人有不同的标准。不可否认,地学研究已经取得了很多的成绩,但是不能津津乐道于目前的成绩,应当从更深层次去思考,从更高的境界去发现存在的问题和可能的危机。地质学天天在进步,这是毫无疑问的,但是,地质学的进步是非常缓慢的,相比大多数其他学科,地质学缺少创新的理论发现。

横向比较,中国地质界在国际上,在某些方面已经很辉煌了,例如在文献数量上遥遥领先,有不少高被引的论文。但是,大量文献中真正的创新内容并不多,全球地质学新理论、新术语、新方法中,由中国人提出来的有多少?

纵向比较,现在与民国时期、解放初期相比,地质技术、分析测试手段、物化探技术的进步可说是如日中天。但是,在学术思想、学术氛围、基础理论上,有多少进步？现在的学术界项目多,经费多,但自由创造、自由竞争的氛围如何?

一部地质发展史揭示,地质理论的进步依靠2个方面:一是地质学科自身发展的动力;二是外来科技进步的动力。最早的地质依靠野外观察和探险,显微镜的发现,使地质首次进入微观领域,发展出岩相学、晶体光学;化学分析测试技术的进步,推动了地球化学的飞速发展,开发出地球化学、同位素年代学;地球物理探测技术的进步则推动了矿床勘察技术的进步,发展出地球物理学,创造了板块构造理论。上述现象说明,地质的发展离不开科学技术的进步,依靠地质自身,基本上是渐进式的进步;而外来科技的引入带给地质的则是跨越式的进步。

目前,全球有两大技术:一个是大数据,另一个是量子力学。大数据正在被引入地质学,虽然目前效果并不显著,但是,其意义将使地质学从观察学科转变为数据科学。量子力学争论很多,但是,量子力学理论是科学界最完美的理论。虽然量子力学理论和量子纠缠技术目前还没有引入地质界,如

收稿日期:2019-04-23;**修订日期**:2019-07-28

资助项目:中国科学院地质与地球物理研究所岩石圈演化国家重点实验室项目《镁铁-超镁铁岩大数据研究》(编号:81300001)、国家重点研发计划项目《基于"地质云"平台的深部找矿知识挖掘》(编号:2016YFC0600510)和《基于地质云的地质灾害基础信息提取与大数据分析挖掘》(编号:2018YFC1505501)

作者简介:张旗(1937-),男,研究员,岩石学和地球化学专业。E-mail: zq1937@126.com

何引入也不清楚。但是,按此规律,总有一天地质学会借助量子纠缠技术迈上一个新台阶[1]。

目前正逢多事之秋,特朗普对中国的打压惊醒了中国人。更严峻的是,这种打压不是短期行为,而可能是美国持久的国策。遗憾的是,中国早先没有注意抓理论上和源头上的创新,仅满足于跟踪式的研究。今天中国学者如果还不在理论上、在核心技术上有所建树,是要受制于人的。科学无国界这句话要重新认识。抓创新,抓从0到1的创新,是刻不容缓的任务。

1 大数据与传统研究的不同

大数据研究是当今研究的热点,大数据与传统研究是什么关系?存在哪些问题?二者如何融合?是目前备受关注的问题。大数据与传统研究的不同之处表现在下面几个方面。

(1)研究的出发点不同。传统研究是理论驱动模式,大数据是数据驱动模式。理论驱动模式的研究是先有思路,后开始研究验证,通俗说就是先立项后研究。立项没有雄厚的理论基础,没有明确的目标、方法和思路,没有对研究结果的预见性,是不能开展研究的。按此设计,传统研究原则上是不允许失败的,虽然说科学是探索,允许失败,但如果研究不成熟,可行性有问题,一般是不会批准立项的。而大数据不同,大数据可以有思路,也可以没有思路,大数据允许采用"试试看"的方法,而且可以不必介意结果如何。因为某项研究从来没有用大数据方法研究过,或没有研究的先例,所以,对研究结果成功还是失败说不出个所以然来。没有思路就可能具有盲目性,大数据允许这种盲目性,允许失败。因此,大数据可以有相当大的自由度,且大大降低了研究的门槛,使初学者也能研究大问题,能够去探索专家学者都解决不了的问题。这是大数据的魅力及研究者的贡献。

(2)研究的思路不同。传统研究追求因果关系,而大数据专注相关关系。在地质研究中,人们最关注因果关系,一个问题来了,首先想到的是:它是哪儿来的,源区是什么?例如文献中大家感兴趣的岩石成因、变质成因、沉积成因、矿床成因等,人们的思想模式几乎已经固定,似乎舍此(查明任何事物的因果关系)而外就无所适从了。舍恩伯格和库克耶指出[2],一直以来,人类是通过因果关系来理解和认识世界的,主要采用2种基本方法:一种是建立在假想之上的因果关系分析,另一种是逻辑思考的因果关系分析。上述作者还进一步指出,因果关系一般来源于人类经验中的信念及直觉,有些经不起实证的检验[2]。大数据研究相关关系,不考虑因果关系,因为,因果关系需要论证,有些以为是因果关系,实际上可能并不是,而相关关系是无需证明的,大数据就是从对相关关系的筛选中获取价值的。许多研究已经表明,研究相关关系是非常有价值的,有时甚至可以沙里淘金[3-4]。

(3)数据处理的方法不同。传统研究只能处理有限数量的数据,即通常所谓的抽样数据。抽样的数据具有局限性,有时并不具备代表性,会导致结论的失真,或引向错误的结论。大数据则可处理多维、多元、混杂的数据,它摒弃抽样的数据,大多着眼于全数据模式。全数据模式的代表性是独一无二的,它的一个特点是具有包容性,能够容忍垃圾数据,能够从垃圾数据中提炼出有价值的数据。例如许多人相信TTG=adakite,张昌振等[5]的研究否定了上述认识,指出TTG=adakite是一个伪命题;刘欣雨等[6]汇总了全球岩浆岩数据发现,新生代的岩浆活动主要集中在中新世;张明明等[7]通过对全球碱性岩的数据分析发现,太古宙、古生代、中生代碱性岩分布很少,元古宙有一些,新生代最多。上述结果如果采用抽样的方法是很难发现的,这就是全数据模式的优势。

(4)人为因素的影响不同。传统研究很难摆脱人为因素的影响,因为传统研究的主体是人,演绎和推理的主角也是人。而由于不同人的视角不同、水平不同、习惯不同、知识面不同,造成研究的结果往往具有强烈的个人色彩。研究应当是客观、真实的,传统研究很难做到这一点,而大数据则可以,因为大数据是以计算机作为研究的主体,让计算机代替人去思考。因此,这里鲜有人为因素的影响,许多需要人去干预的地方,也应尽量减少人为干预的程度。故大数据的结果是真实的,也是科学、可信的。

2 大数据与传统研究的关系

大数据是最近才出现的术语,经由《大数据时代》这本书的介绍[2],才逐渐为中国人所知。但是,数学应用于地质并不是现在才有的,很早就有数学地质这门学科(Mathematical Geology)。地质大数据与数学地质有相同也有不相同的地方,相同之处是都利用数据研究地质问题,不同主要表现在处理

复杂多源混杂数据的能力和运算速度上;二者最大的区别在于科学范式上的不同:大数据属于数据密集型科学,而数学地质一直跳不出传统研究的思路,因此,多年来进展并不理想。大数据不同,大数据出来以后及大数据应用于地质学,改变了地质学研究的思路和方向。大数据不是附属于地质,大数据能改变地质。大数据对地质学的改变之大,许多人目前可能还认识不到。大数据对地质学最大的冲击是将地质学从科学降格为学科。在大数据眼里,地质学属于观察学科。只有将大数据应用于地质学,才有可能将地质学从学科转变为科学[4]。

关于传统研究与大数据研究的关系,目前的现状并不乐观,可以用三句话概括:①少数人认为大数据很重要,值得重视并予以推动和支持;②多数人认为地质没有海量数据很难研究;③个别人对大数据心存疑虑,不认为仅靠一堆数据就能够得出什么结论。上述情况表明,学术界对大数据还存在许多误解。他们不明白,大数据的"大"不是指数据绝对量的大,而是指大数据能够处理关于某个现象的"所有数据"。

大数据开辟了地质学研究的新篇章。那么,在大数据时代,传统研究是不是就不重要了呢?恰恰相反,大数据的到来不是跟传统研究对着干的,而是为了解决地质研究的困惑,给传统研究带来福音的。大数据不排斥传统研究,与传统研究不是对立而是互补的。

因为,首先,大数据研究要以传统研究为基础,没有了传统研究,也就无所谓大数据研究。大数据技术恰恰是因为传统研究的不足而诞生的,是科学发展进步的产物。

其次,大数据的结果需要传统研究来解释,如果没有传统研究,大数据的价值就很难知道。前面提到地质学进步迟缓的问题,其实地质学已经面临危机了[4]。自从板块构造理论问世以后,地质学就没有新的理论诞生。而大数据由于处理数据方面的优势,是传统研究所不及的。大数据的结果大多数属于相关关系的范畴,这种相关关系许多可能是以前不知道的,没有研究过的。这时就需要求助于传统研究,需要利用传统研究积累的知识来解释。

第三,有时大数据得出的结果是传统研究从来没有接触过的、也无法解释的[3-4],这就给传统研究提出了问题,推动传统研究去探讨和解决这些问题,从而促进传统研究的进步。

地球是一个复杂系统,由于地质三维空间巨大,影响因素众多,过程曲折反复,许多信息深埋于地下,使地质信息具有高维度、高复杂性和高不确定性,导致地质大数据具有多类、多维、多量、多尺度、多时态和多主题特征[8-15]。

大数据有广义与狭义之分:狭义的大数据以5V特点为标志,适合处理海量复杂多元的数据,需要云计算和专业人才,门槛很高。而符合大数据3个技术取向的(重全体不重抽样,重效率不重精确,重关联不重因果)、从数据出发的研究是广义的大数据研究[4]。

地质问题非常广泛,有些并不复杂,许多可以加以分解和简化,不需要复杂的数学方法即可解决。其次,在某些情况下,简单的方法也能解决大问题,问题是你把问题抓住了没有。大数据方法本身是需要研究和发展的,但是,当务之急是要面对地质、矿床研究和实践中所提出来的问题,利用已知的或现成的或成熟的方法开展研究。按照这个思路,地质问题解决了,方法也提高了,而且出成果也快了[16]。大数据出的成果越多,发挥的效益越好,自然会受到关注和认可。大数据也会因此而越来越普及,越来越受到重视,其发展道路也会越来越顺畅。

地质大数据是跨界的研究,大数据与地质是一个问题的2个方面,实际情况是谁也离不开谁。但是,二者的结合不容易,结合得好更不容易。如果大数据与地质研究沟通得好,可以得到1+1>2的结果;如果沟通不好,则可能是1+1<2[17]。

学术界与大数据存在隔阂,传统研究与大数据研究存在矛盾。笔者认为,这一对矛盾的主体不是在地质界,而是在大数据研究群体自身。那么,如何解决这个矛盾,如何拉近学术界与大数据之间的距离呢?笔者认为,一个重要的方面是大数据要尽快出成果。不必纠缠于大数据的优点如何如何,而是要埋头苦干,尽快拿出扎扎实实的成果来,让事实说话,这才是最有力的回答。

传统研究与大数据研究目前之所以沟通不畅甚至很难沟通,在于大数据目前没有多少拿得出的、实实在在的、让地质学家服气的结果。这需要一个过程,这里也存在2个问题:一是地质大数据研究的难度大。为什么难度大,主要是因为地质的数

据化程度低,且不是一般的低,而不能数据化的资料是无法进行大数据研究的。要把地质观察、报告、图件全部转化为数据,这是一个浩瀚的工程,不可能一蹴而就。因此,地质云是极其有意义的,尽可能把地质云建好并实际加以运行,是地质大数据研究的前提。另一个问题是对于目前能够研究的地质问题,要加大研究力度,力争尽快取得成果。

出成果不容易,首先需要数据。想做大数据,没有数据也无奈。有了数据,数据能否共享,也是一个问题。因此,在中国大数据研究不是一般的难。多年来,学术界积累了不少数据,但是,对数据的利用并不充分,许多数据处于长期未被开发的状态。例如一件样品的地球化学和同位素分析可以得出几十项结果,而真正用到的仅是几或十几项,其余的就被忽略了。又如矿区尤其老矿区积累了大量的地质物化探采选冶数据,真正开发利用的有多少?有多少数据被扫进垃圾行列?因此,如果能够很好地利用上述数据尤其是垃圾数据,或许能够挖掘出数据中隐含的某些价值。

在地质大数据研究中,中国学者已经开始起步并取得某些明显的进步。如自上而下逐级建立了若干数据库,制定了各门类数据的国家标准,出现了各种尚处于雏型的大数据平台,各种数学方法跃跃欲试(与路来君教授的交流)。从学科门类看,古生物学科的大数据研究取得了不菲的成绩,而地质学的其他学科则步履蹒跚。由此看来,加快推进地质大数据的研究是一项刻不容缓的任务。幸运的是,深时数字地球(DDE)大科学计划恰逢其时来到中国。全球 DDE(Deep- time Digital Earth (DDE) Big Science Program)香山科学会议于2019年2月25—28日在北京召开,会议由王成善院士主持,邀请了国际地层委员会(ICS)、国际大地测量学与地球物理学联合会(IUGG)、世界地质图委员会(CGMW)、国际地质形态学家协会(IAG)、美国石油地质学家联合会(AAPG)、国际矿床成因学会(IAGOD)、国际沉积学家协会(IAS)、国际数学地质协会(IAMG)、地球科学信息管理和应用委员会(CGI),以及两大联合会——国际地质科学联合会(IUGS)、国际科学理事会(ISC)的负责人和全球知名学者代表参会。DDE计划是由国际地科联推动的首个大科学计划,旨在推动地球大数据研究,鼓励地球科学和跨学科领域的广泛合作,推动地球科学取得重大突破。DDE计划是及时雨,将引领中国地质大数据进入快车道。传统研究不可妄自菲薄,大数据研究需要加快步伐,二者紧密融合,才能推动地质学跨越式的进步。

致谢:感谢吉林大学路来君教授对本文提出的建议。

参考文献

[1]张旗, 焦守涛, 李明超, 等. 量子纠缠技术在地质上应用的可能性[J]. 地学前缘, 2019, 26(4): 159−169.

[2]维克托−迈尔−舍恩伯格, 肯尼思−库克耶. 大数据时代: 生活、工作与思维的大革命[M]. 盛杨燕, 周涛, 译. 杭州: 浙江人民出版社, 2013.

[3]刘欣雨, 张旗, 张成立. 大数据方法建立大洋安山岩构造环境判别图[J]. 地质通报, 2019, 38(12): 1963−1970.

[4]张旗, 周永章. 大数据助地质腾飞: 岩石学报2018第11期大数据专题"序"[J]. 岩石学报, 2018, 34(11): 3167−3172.

[5]张昌振, 张旗, 金维浚, 等. 太古宙 TTG 能否与埃达克岩对比? ——全球数据给出的结果[J]. 地质科学, 2018, 53(4): 1254−1266.

[6]刘欣雨, 张旗, 张成立, 等. 中新世全球重要事件及其意义:数据挖掘的启示[J]. 科学通报, 2017, (15): 1645−1654.

[7]张明明等. 全球新生代碱性岩大爆发及其意义(审稿).

[8]吴冲龙, 何珍文, 翁正平, 等. 地质数据三维可视化的属性、分类和关键技术[J]. 地质通报, 2011, 30(5): 642−649.

[9]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.

[10]Lu L J, Liu W B. Digital spectral analysis method of geological space[J]. ICIC Express Letters, 2015, 9: 1699−1706.

[11]赵鹏大. 大数据时代数字找矿与定量评价[J]. 地质通报, 2015, 34(7): 1255−1259.

[12]赵鹏大. 地质大数据特点及其合理开发利用[J]. 地学前缘, 2019, 26(4): 1−5.

[13]朱月琴, 谭永杰, 张建通, 等. 基于Hadoop的地质大数据融合与挖掘技术框架[J]. 测绘学报, 2015, 44(S0): 152−159.

[14]黄少芳, 刘晓鸿. 地质大数据应用与地质信息化发展的思考[J]. 中国矿业, 2016, 25(8): 166−170.

[15]陈建平, 李靖, 谢帅, 等. 中国地质大数据研究现状[J]. 地质学刊, 2017, 41(3): 353−366.

[16]葛粲, 张旗, 李修钰, 等. 低维到高维密度分布函数及其可视化在大数据分析中的应用——以苦橄质玄武岩等为例[J]. 地质通报, 2019, 38(12):2043−2052.

[17]张旗, 焦守涛, 李承东, 等. 花岗岩与大陆构造、岩浆热场与成矿[J]. 岩石学报, 2017, 33(5): 1524−1540.