开放科学时代的科学数据治理

李新、苏建宾

中国科学院青藏高原研究所

国家青藏高原科学数据中心

2024年9月20日，成都

走向数据善治: 以地球科学数据治理为例

科学通报 2024 年 第 69 卷 第 9 期: 1149 ~ 1155

香山科学会议专栏 观 点

第S69次学术讨论会·科学数据治理与利用的前沿和热点

《中国科学》杂志社
SCIENCE CHINA PRESS

走向数据善治: 以地球科学数据治理为例

李新*, 苏建宾

中国科学院青藏高原研究所, 青藏高原地球系统与资源环境国家重点实验室, 国家青藏高原科学数据中心, 北京 100101

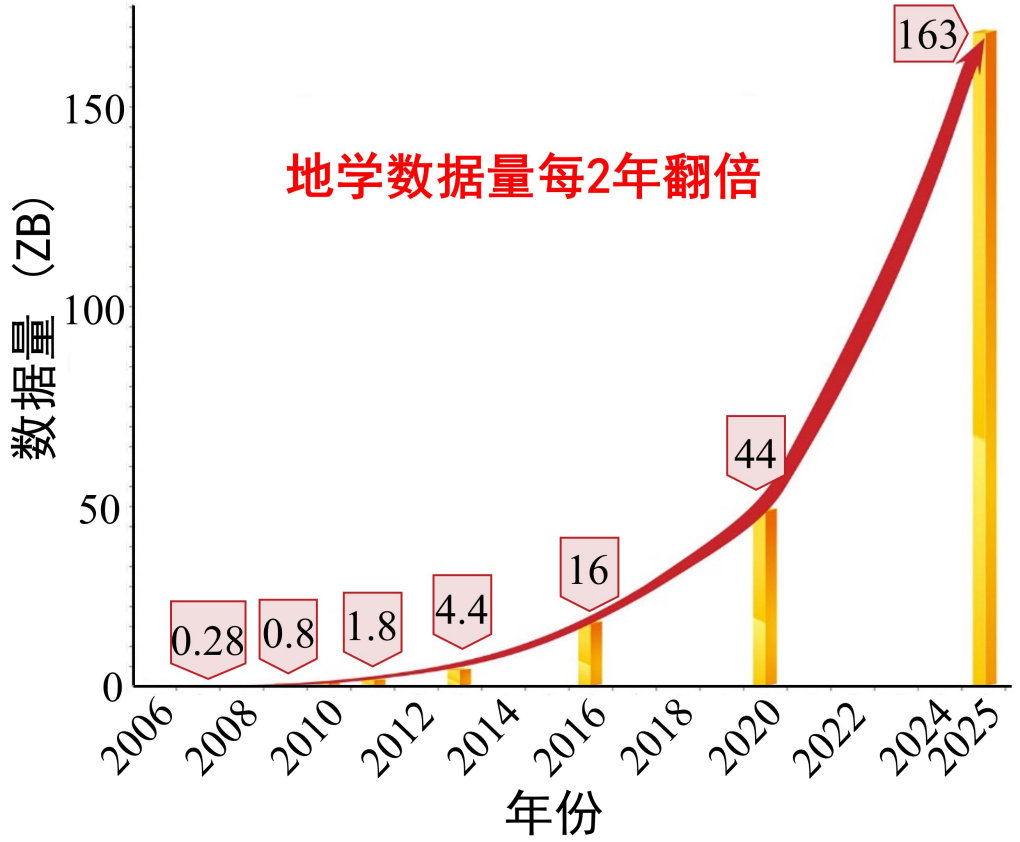
* 联系人, E-mail: xinli@itpcas.ac.cn

科学数据是科技活动中获取的客观数据, 也是科研
不可或缺的组成部分[1]. 随着我国科技投入的不断增长
技创新能力的持续提升, 海量科学数据不断产生, 逐步
重要的科技战略资源, 推动了科学研究向数据密集型科
式的转换[2]. 作为科研活动的灵魂, 科技创新越来越依赖
海量数据的集成、分析和利用. 因此, 积极开展科学数
理和开放共享, 不仅能提升科学数据的服务和开发能力,

李新, 苏建宾, 2024. 走向数据善治: 以地球科学数据治理为例. 科学通报. 69(9), 1149-1155.

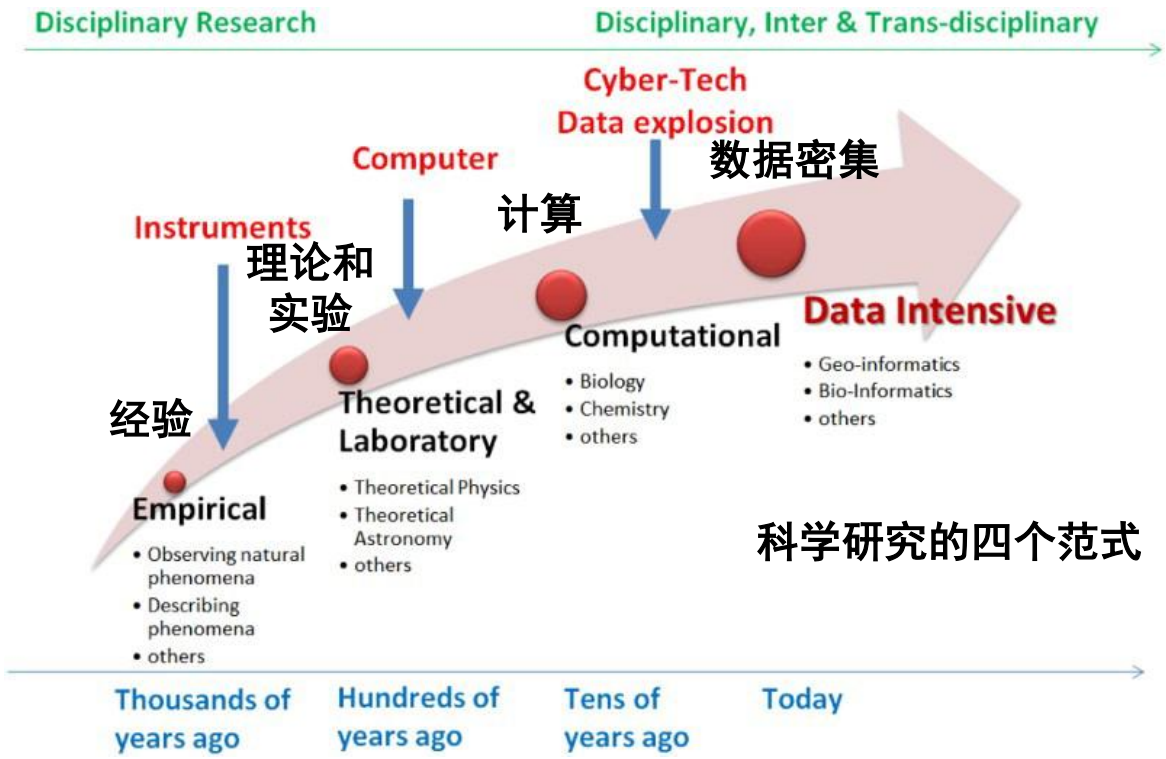
数据驱动的新地学

地学发展日益呈现密集数据驱动和学科交叉的趋势



新技术和新方法的出现和应用推动了高质量地学数据呈指数级增加

(Guo et al., 2017)

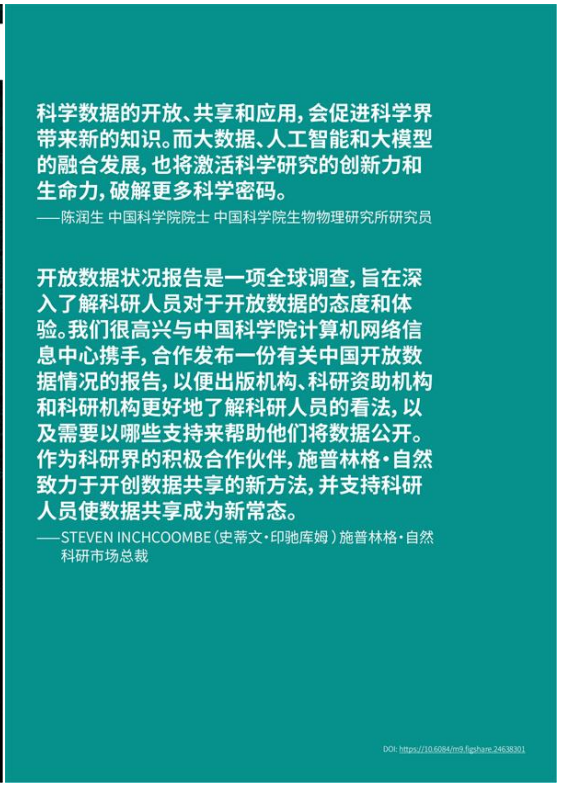


地球大数据和人工智能技术的发展推动了新研究范式的产生

(Hey et al., 2009)

开放数据共享的重要性

科学数据的开放共享和自由流通对于科学发展非常关键，而其与人工智能的融合，能够有力地激活科研创新力、生命力，破解更多科学密码

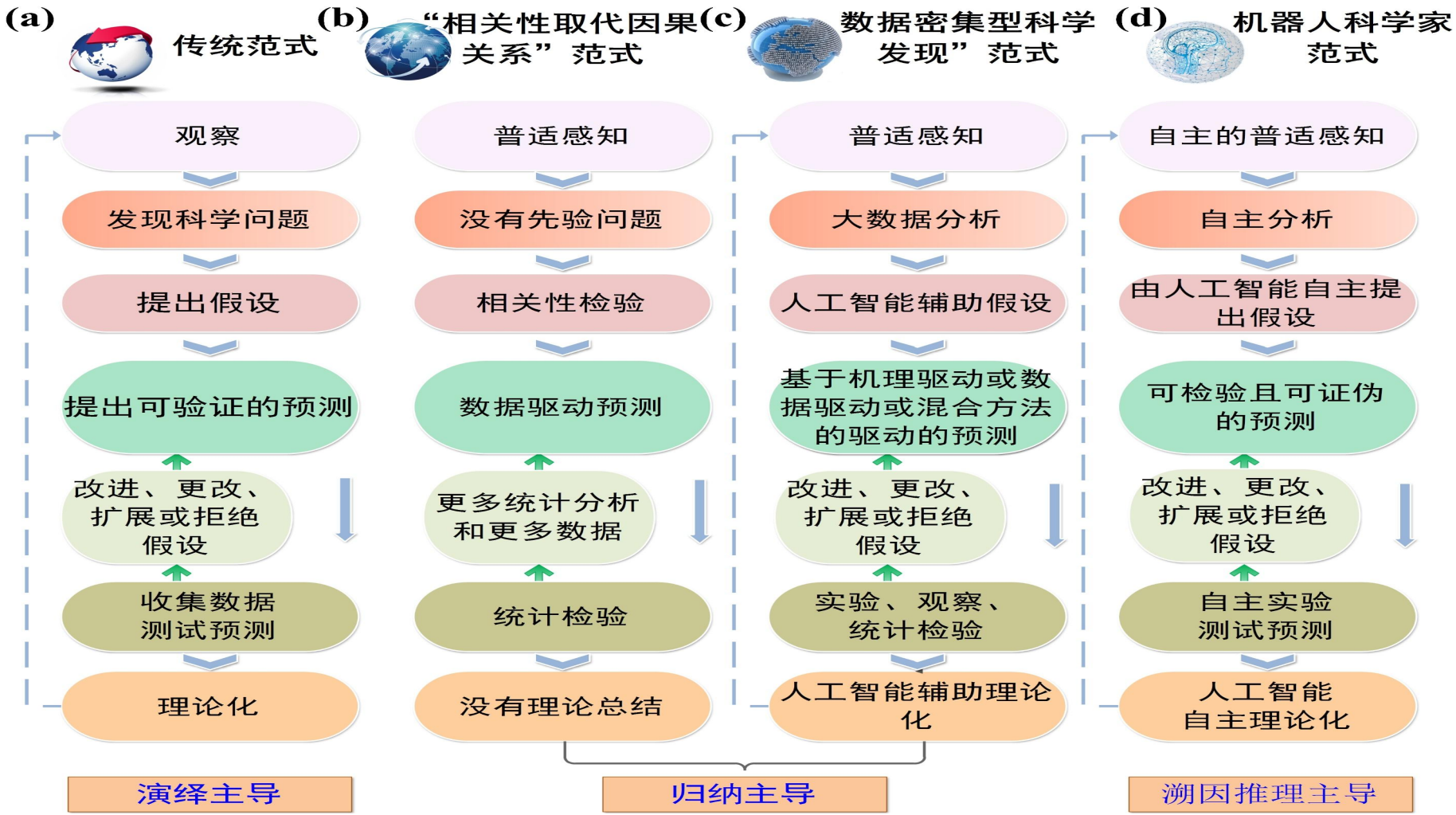


“Science is driven by data.
We must all accept that science is data and that data are science, …”
Science, 2011, special section, pp. 692-729

Research cannot flourish if data are not preserved and made accessible…
Nature, 10 September 2009, special section

《中国开放数据白皮书2023》

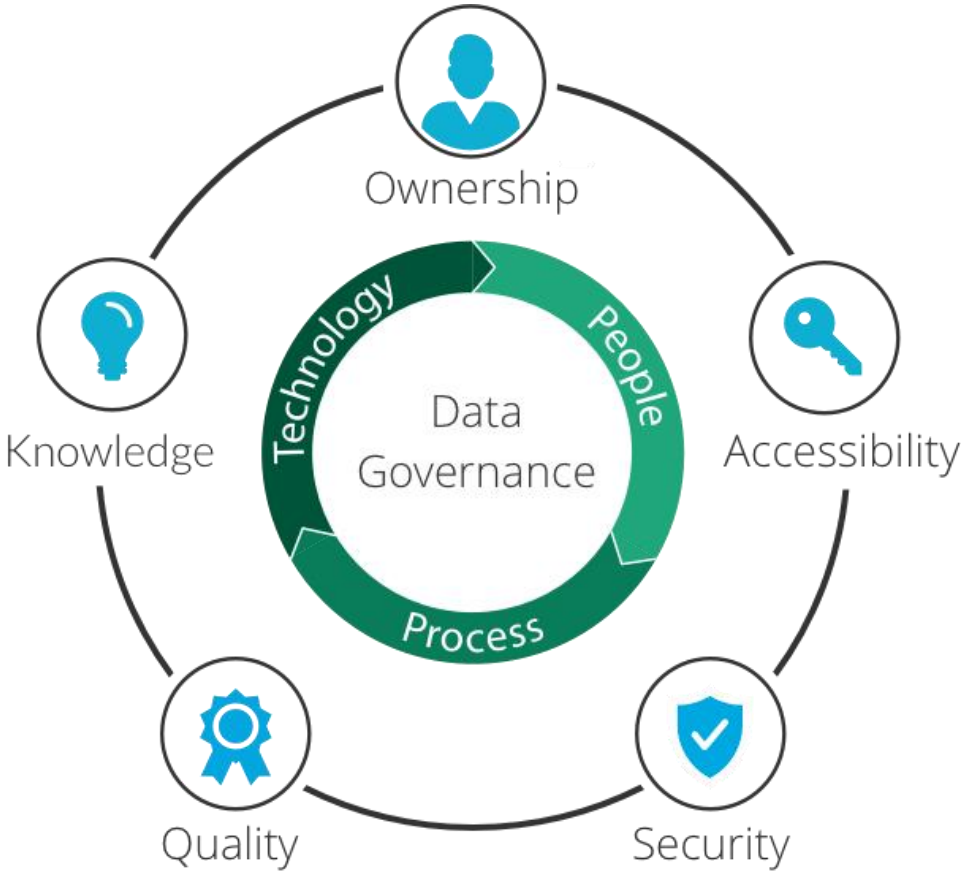
地球系统科学研究中的范式转换



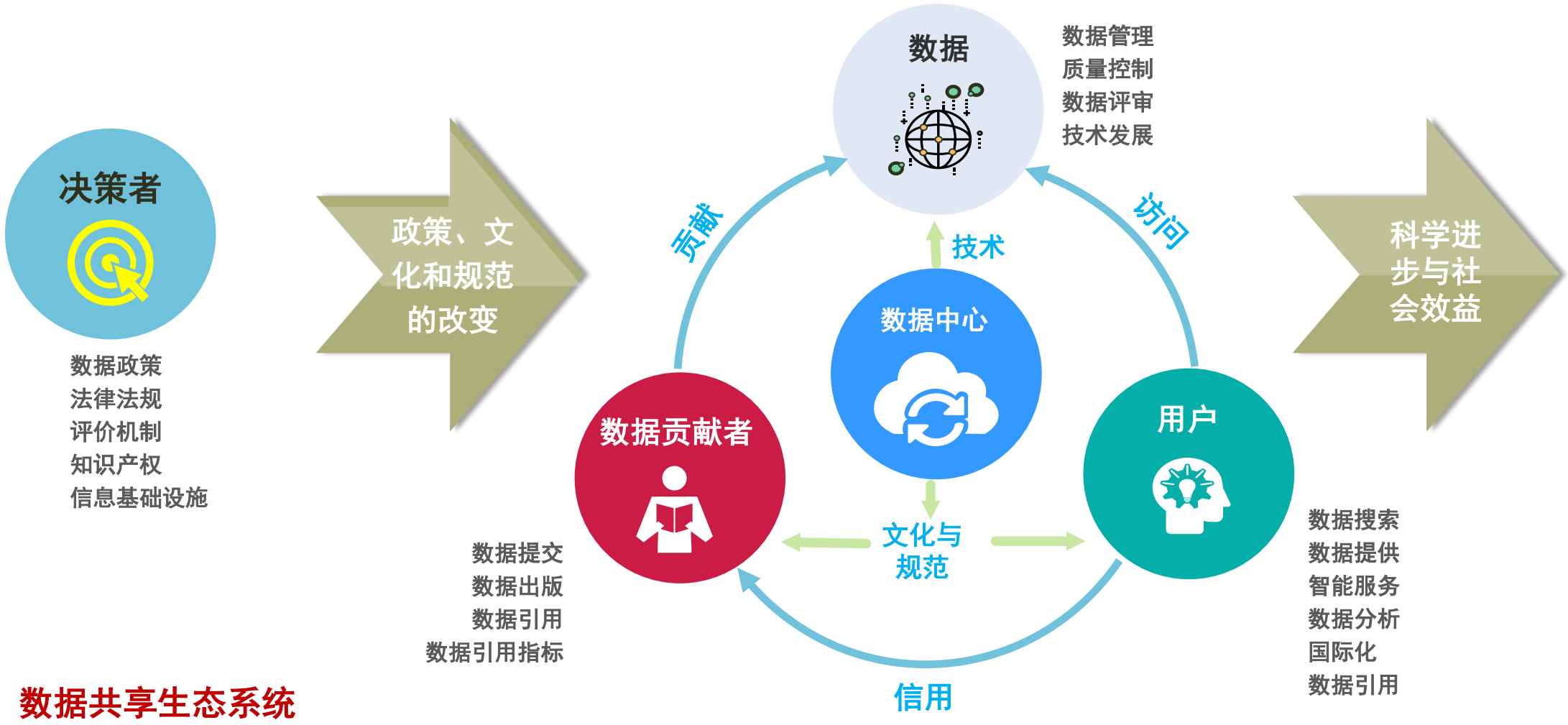
Li and Guo, 2024, Science Bulletin

挑战1：从科学数据管理到科学数据善治

- **数据治理**的目的是增加数据的价值，并将数据相关的成本和风险降至最低（Abraham et al., 2019）
- 确保高数据质量贯穿于数据的整个生命周期（wikipedia），可获取性、可用性、一致性、集成度、安全、标准
- 数据治理是一个**集体行动**问题（Benfeldt et al., 2020）
- 贯彻FAIR and CARE原则
- 加持 AI for Science, FAIR: **F**indable and **AI** **R**eady (Scheffler et al., 2022)



自上而下的政策约束 + 激励机制



数据共享生态系统

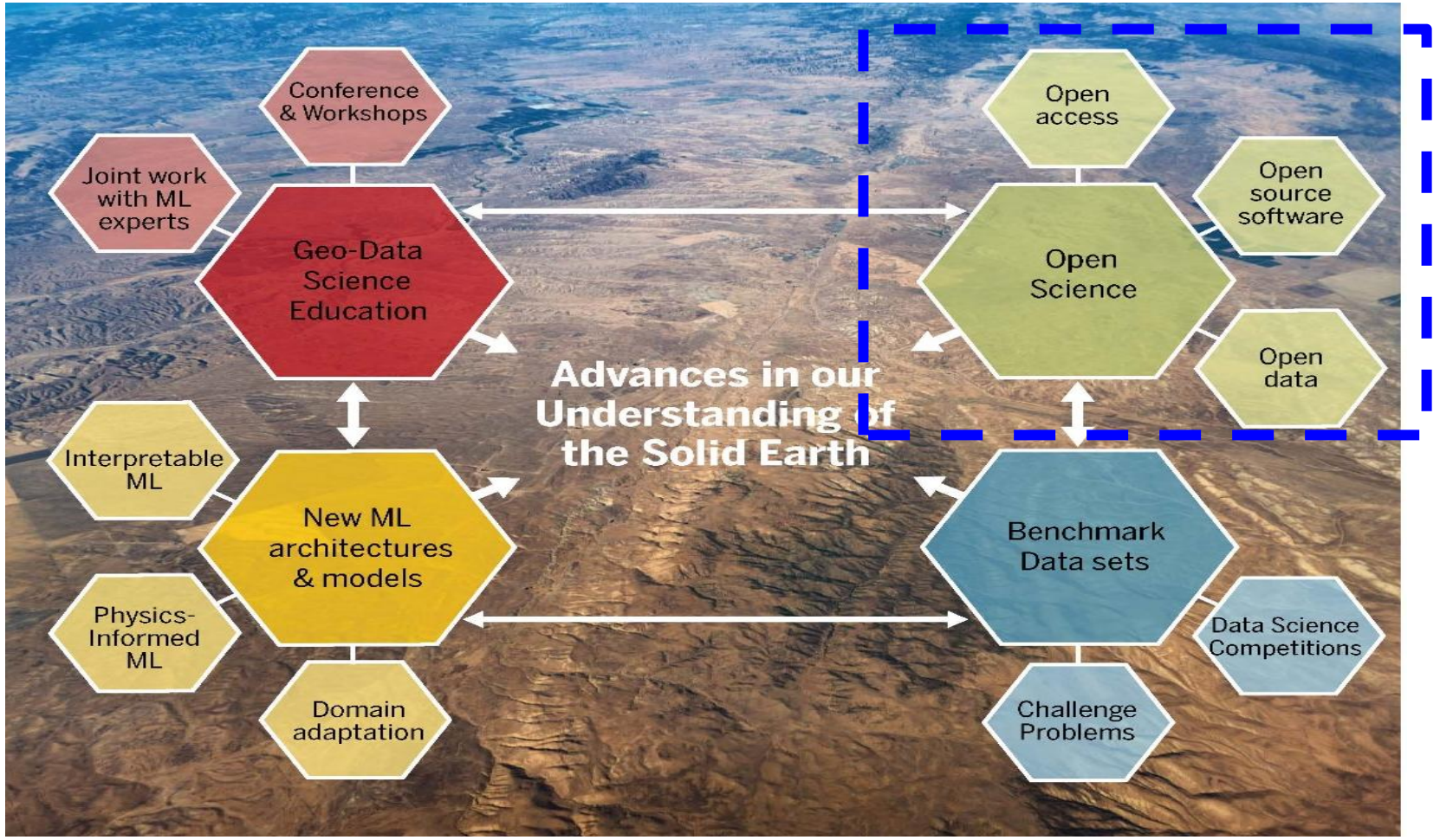
Li X, Cheng GD, Wang LX, Wang JL, Ran YH, Che T, Li GQ, He HL, Zhang Q, Jiang XY, Zou ZM, Zhao GF. Boosting geoscience data sharing in China. *Nature Geoscience*, 2021, 14(8): 541-542, http://10.1038/s41561-021-00808-y

地学数据善治：应对举措建议

- 制定数据质量控制的度量指标和标准
- 启动数据引用，作为激励机制的杠杆
- 加强与期刊（优先国内的一流期刊，如NSR, Science Bulletin, 中国科学）之间的合作，加快论文关联数据仓储
- 深度参与国际上地学数据治理（FAIR, TRUST, CARE）
- 细粒度、可操作地贯彻**FAIR**原则
- 鼓励无差别、不设限的数据访问

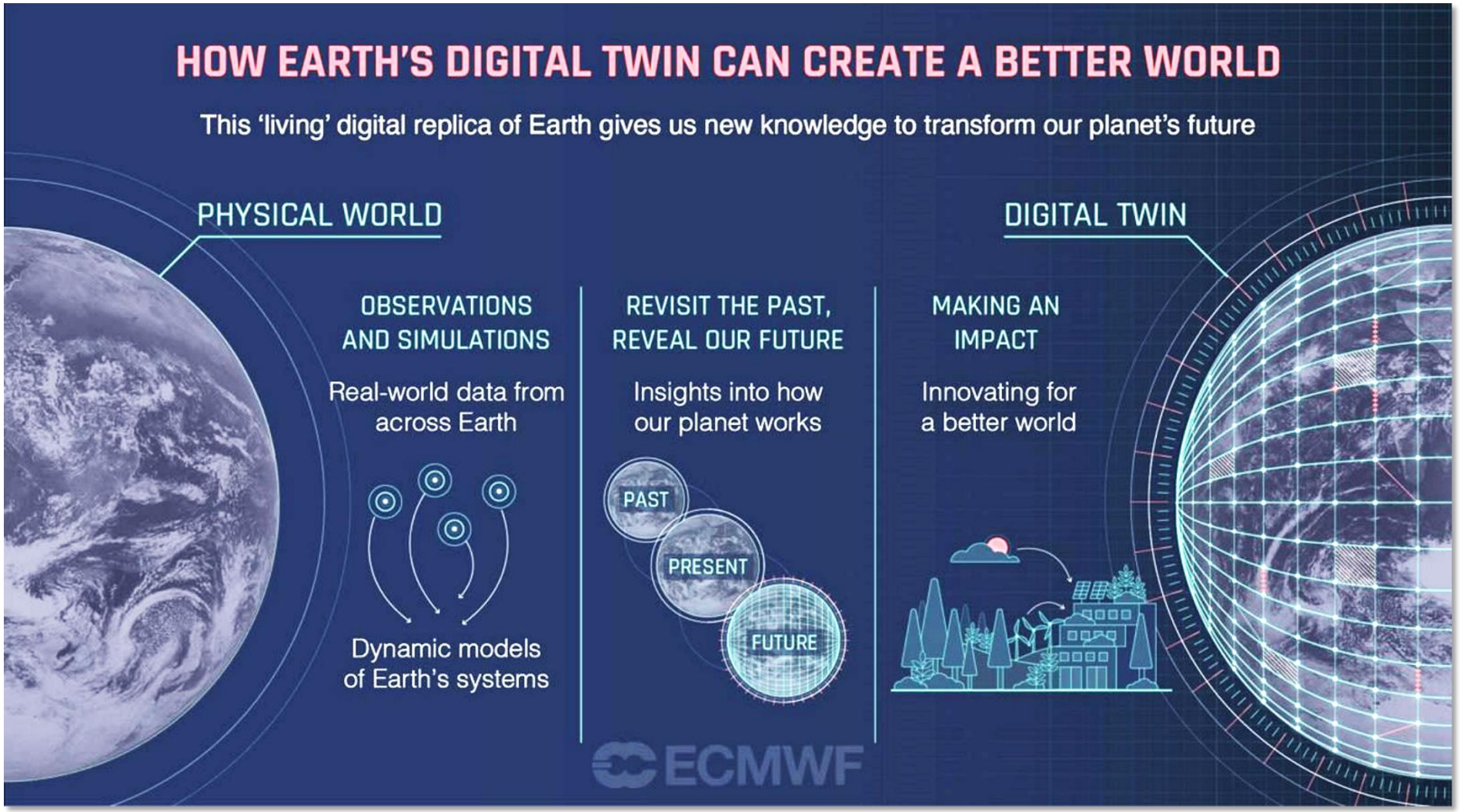
李新、苏建宾, 2024. 走向数据善治: 以地学数据治理为例. 科学通报

挑战2：打造Open science 时代的地球科学信息基础设施



Bergen KJ, Johnson PA, Maarten V, Beroza GC. Machine learning for data-driven discovery in solid Earth geoscience. *Science*, 2019, 363(6433): eaau0323

数字孪生地球：Destination Earth 项目



Bauer et al. 2021, NCC; Voosen 2021, Science

SDG大数据平台

促进我国开放数据基础设施不断产出重要成果，服务SDG目标监测评估

- 建成SDG大数据云基础设施：50PB 存储，2PF 高性能计算、10000CPU 核心云计算
- 突破SDG大数据系列关键技术，自主研发数据管理引擎Databank、智能分析引擎EarthDataMiner和可视化引擎DESP等系统软件，开发了120个栅格和矢量计算函数算子和15个机器学习模型，150个左右的算法工具，云化集成了100余款工具软件
- 提供了大规模计算模拟、人工智能模型训练、数据产品按需生产、指标计算与综合分析等服务

SDG大数据云基础设施

SDG大数据平台

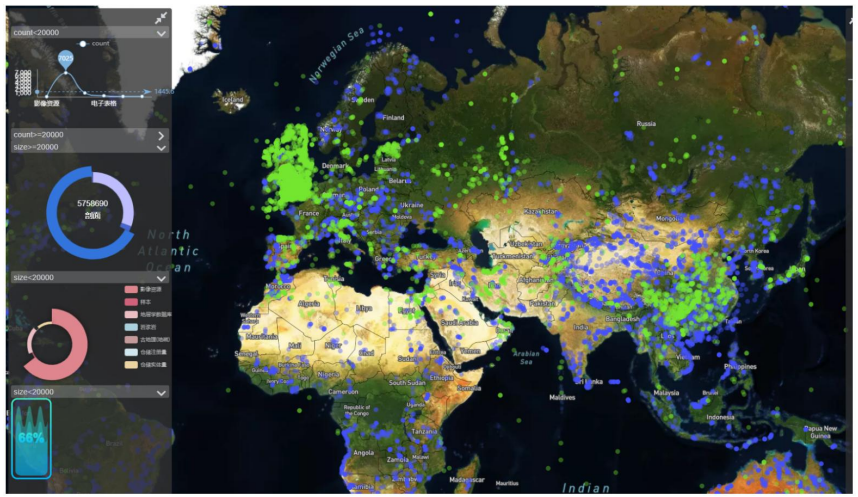
SDG决策支持系统

SDG学科交叉协同

Guo et al., 2021, 2022, Science Bulletin

“深时数字地球”国际大科学计划

整合地球演化全球数据，打造数字时代具有竞争力的全球科技公共产品



以DDE云为基础，通过整合各种技术，进行“知识-数据-算法-算力”协同优化，从而实现科技创新与数字世界的打通

以DDE数据为核心引擎，通过使用开放数据政策促进地球演化数据的聚合，提供长期数据管理，支持数据驱动的地球演化发现，从而支持DDE愿景

DDE Knowledge是创造知识的工具，通过知识图谱等表达复杂的地质实体之间的关系，致力于数据驱动的知识发现

Wang et al., 2021, NSR; 王成善 等，2022

地学信息基础设施建设：应对举措建议

数据杂货铺

数据图书馆

数据实验室

数字孪生地球

过去 现在 将来

- 数据中心应全面转型到数据实验室阶段，并向数字孪生地球进军
- 打造中国制造的“数字孪生地球”

李新、苏建宾, 2024. 走向数据善治: 以地学数据治理为例. 科学通报

挑战3：地学数据集成

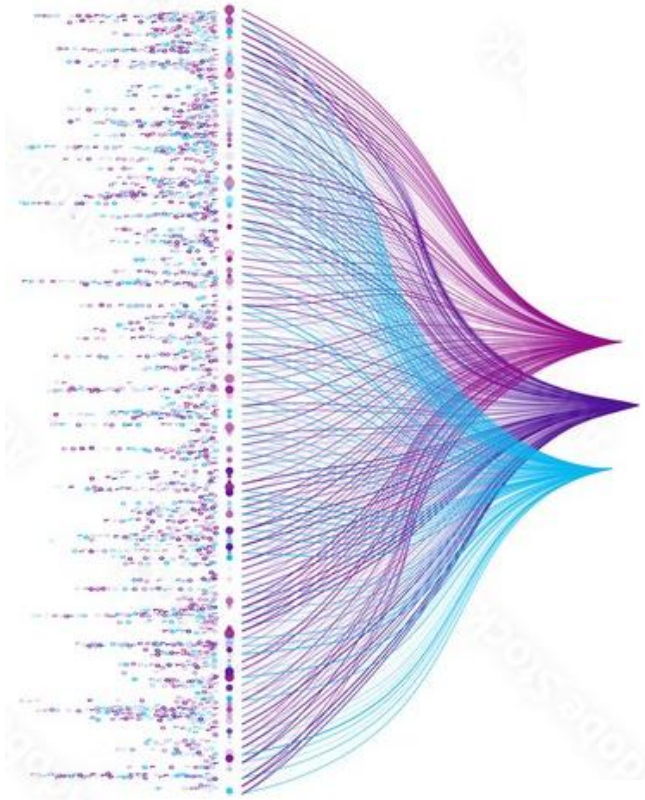
再分析数据产品	发布时间	Google scholar 引用	Web of Science 引用
ERA5	2020	16,146	12,976
MERRA-2	2017	6,389	5,125
JRA-55	2015	4,391	3,536
ERA-Interim	2011	25,842	21,040
NCEP/NCAR 40-year	1996	35,423	27,434

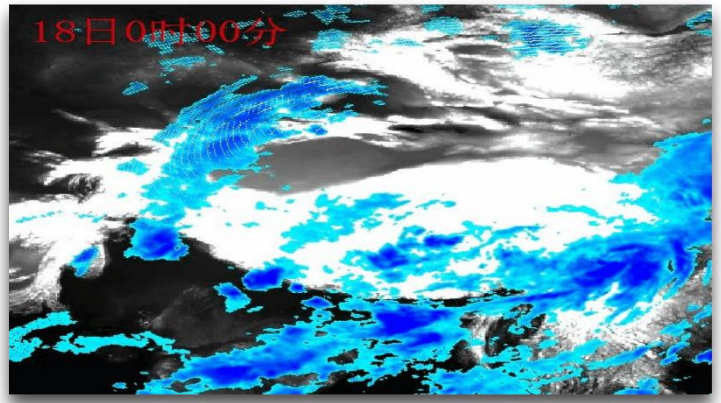
"*ERA5 thus benefits from a decade of developments in model physics, core dynamics and* ***data assimilation***" -- Hersbach et al., 2020

引用数更新于2024年9月18日

从分散到关联的深度数据集成

传统数据组织方式（文件、记录）难以适应地学发展的需要，只有新的集成方法才能更有效制备数据产品，支持地学新发现





现代气候变化过程

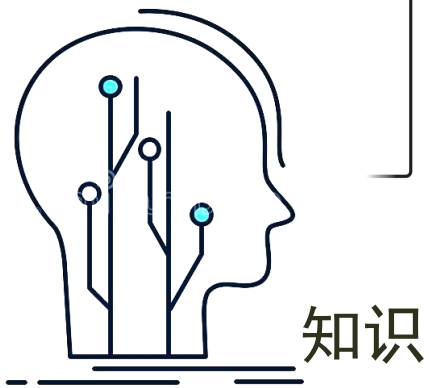
动力演化过程

第二次青藏高原科学考察项目已发布765个数据集，含101.6万个文件，总数据量87.6T；平均每个数据集包括了超过1300个文件

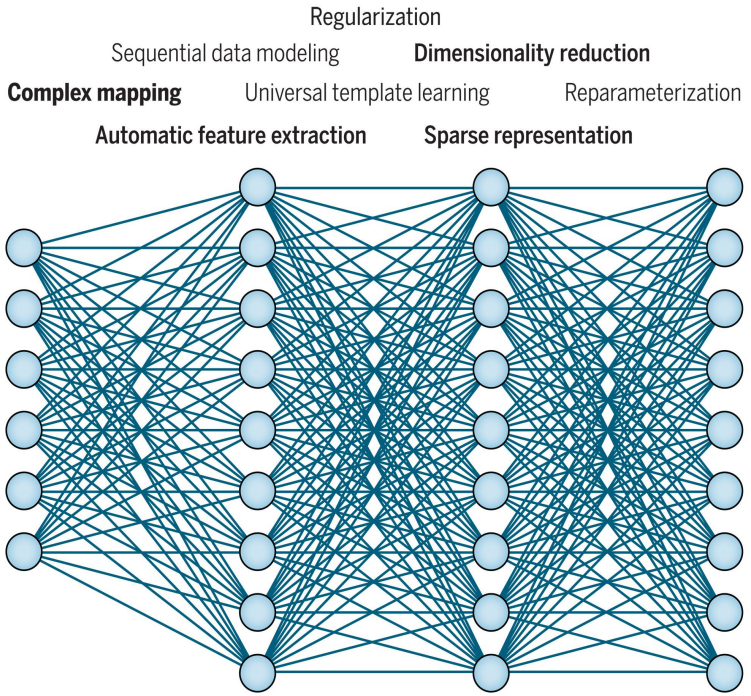
研发旗舰数据产品

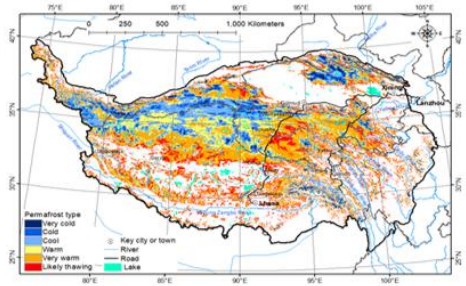
从原始数据资源向高质量数据产品的转化，逐步打造“智能生产”、“快速定制”技术，形成支持地球系统科学和可持续发展的旗舰数据产品

数据



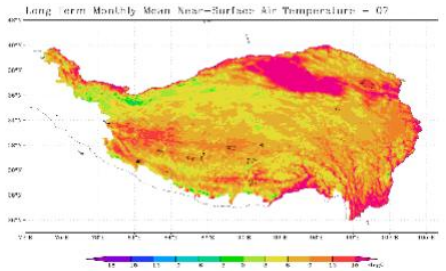
知识





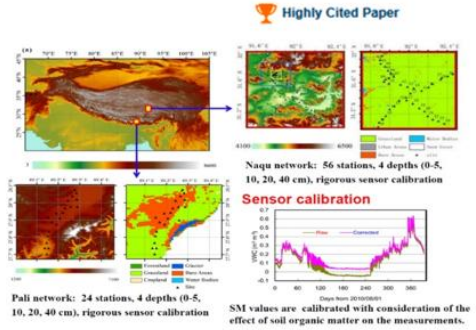
多年冻土潜在热退化数据集(1960s-2000s)

Ran, Li* et al., 2018, *Cryosphere*



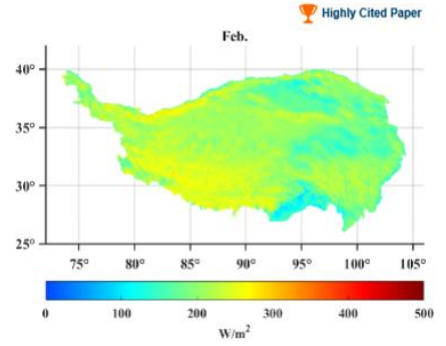
青藏高原地表网格气象数据集 (1979-2018)

He, Yang* et al., 2020, *SD*



青藏高原土壤温湿度加密观测数据集（2010-2021）

Chen et al., 2017, *JGR*



青藏高原10km地表短波辐射数据集 (1983.7-2018.12)

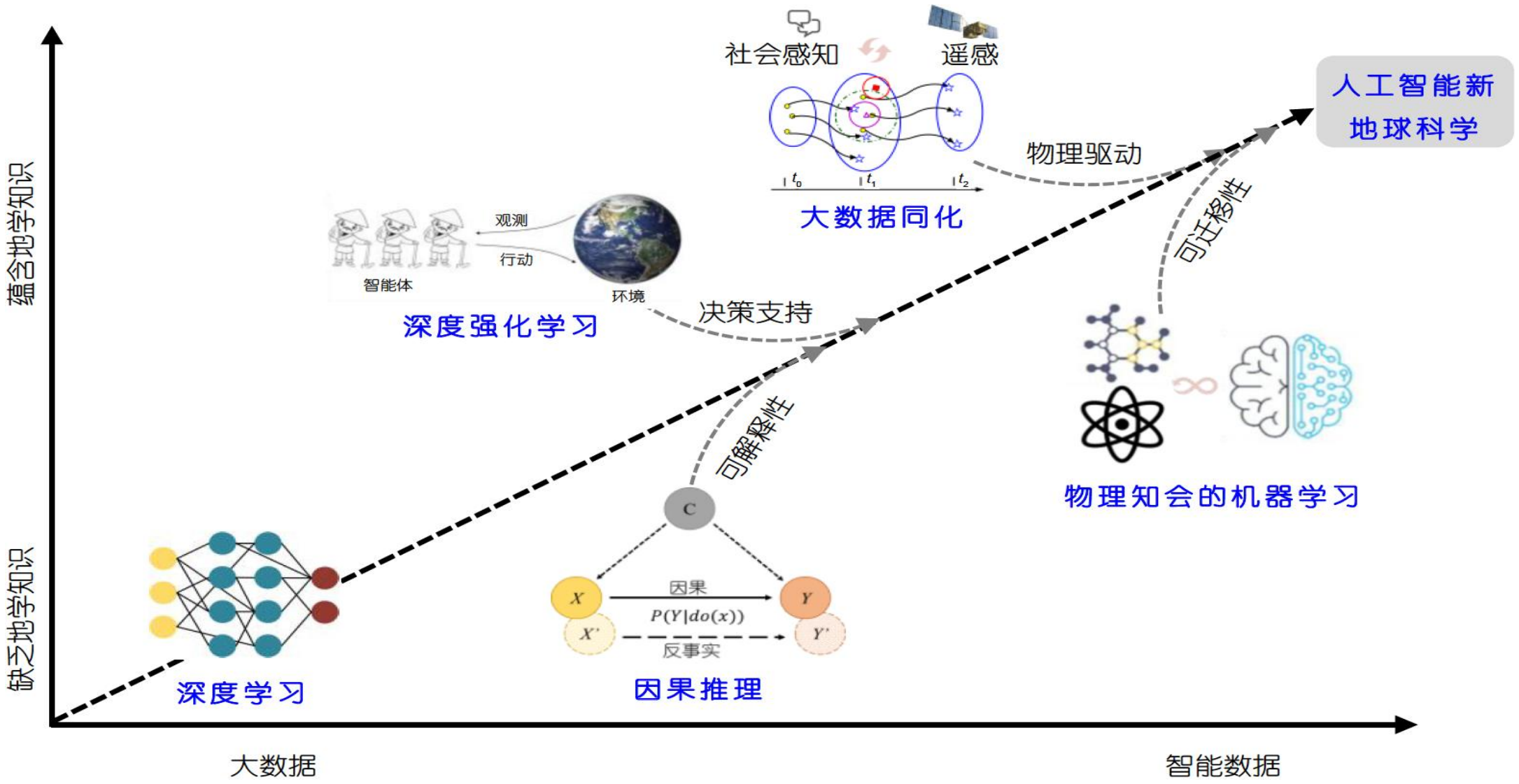
Tang et al., 2019, *ESSD*

地学数据集成：应对举措建议

- 形成从分散的数据集数据集抽取数据、质量控制、形成数据产品，并科学分析的工具
- 研发高质量基准（benchmark）数据集
- 形成地学数据产品“智能生产”技术
- 研发地球系统再分析数据产品

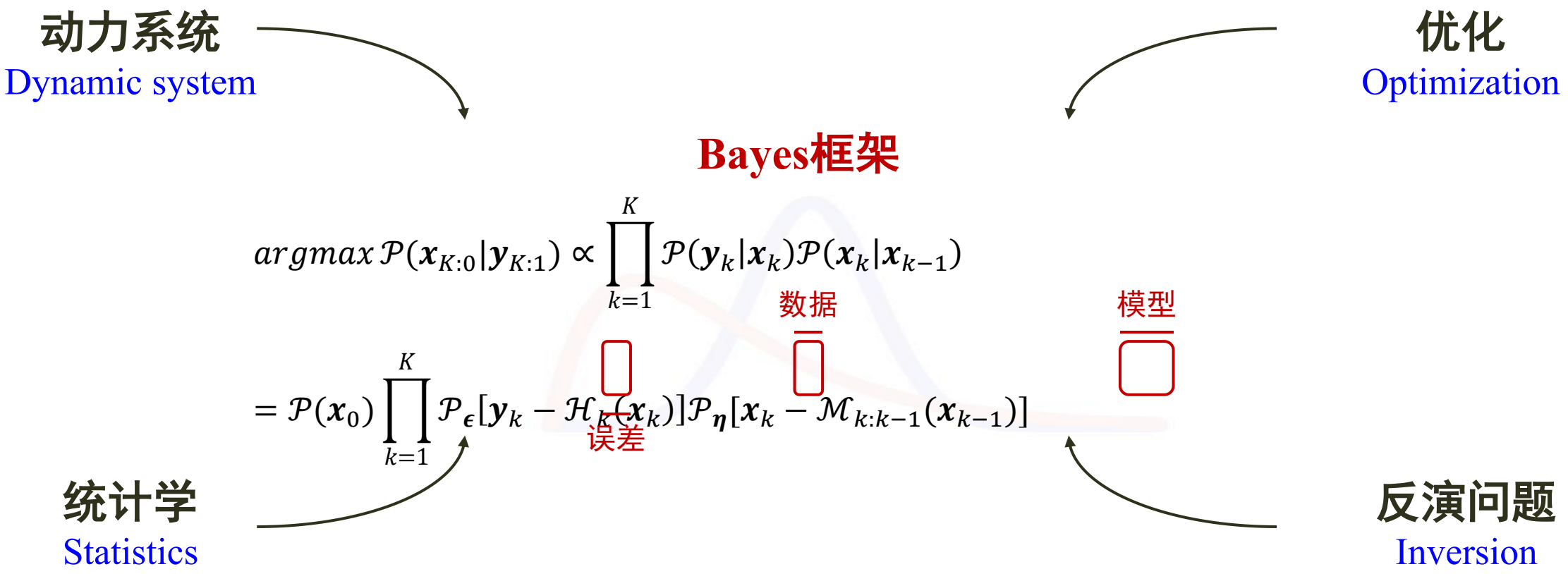
李新、苏建宾, 2024. 走向数据善治: 以地学数据治理为例. 科学通报

挑战4：前沿大数据分析方法，走向AI for geoscience



Li et al., 2023, Nature Reviews Earth & Environment

大数据分析的前沿



Li et al., 2024, Reviews of Geophysics

大数据同化

大气

天气预报
气候再分析
古气候再分析
大气化学

陆地

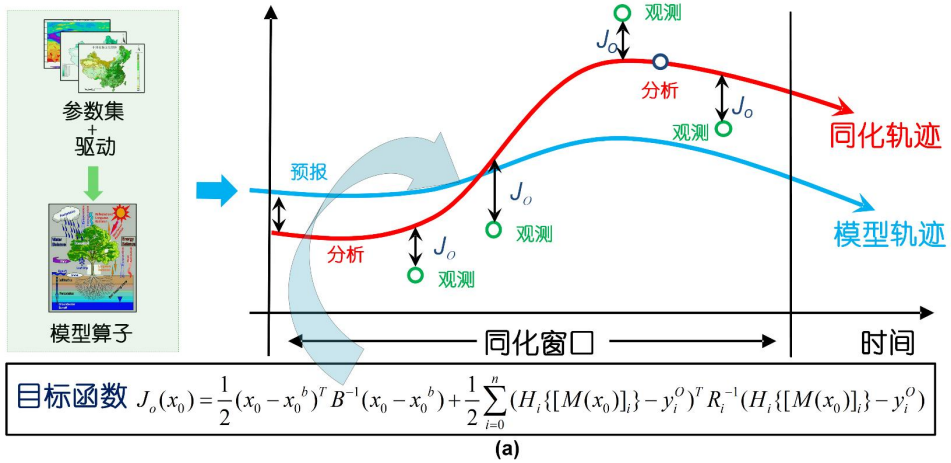
陆面过程
水文
生态
关键带

海洋

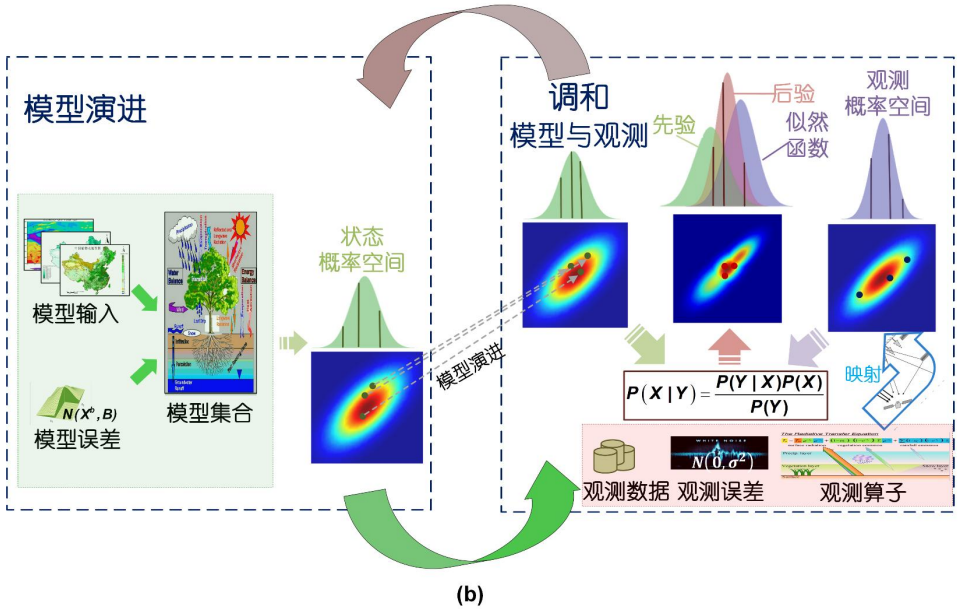
海洋动力学
海啸
海洋温度
海洋盐度

固体地球

地幔动力学
地震
地球物理
地球化学



(a)



(b)

李新 等, 2020. 模型与观测的和弦：地球系统科学中的数据同化. 中国科学: 地球科学

Li Xin et al., 2024. Land Data Assimilation: Harmonizing Theory and Data in Land Surface Process Studies. Reviews of Geophysics

AI for Geoscience：应对举措建议

- 数据中心应成为大数据革命的创新者，成为 AI for geoscience 的引擎
- 发展大数据同化方法，实现地球系统模型和地球大数据的和谐，为数字孪生地球提供动力核
- 发展内嵌地学知识、可解释的机器学习方法，在因果推理、深度强化学习上发力，走向人工智能新地学

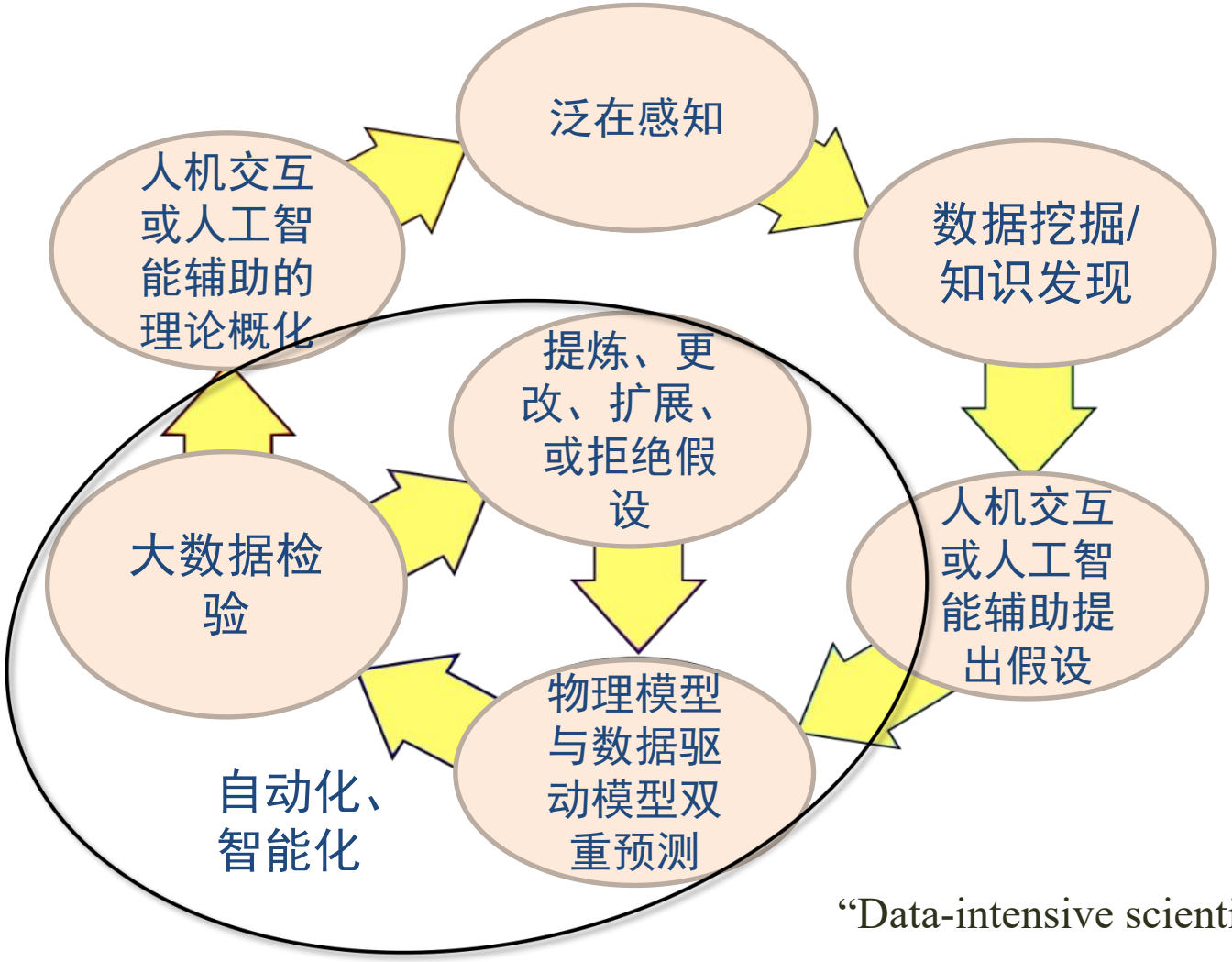
李新、苏建宾, 2024. 走向数据善治: 以地学数据治理为例. 科学通报

- Innovator and leader of data science
- Booster of big data revolution
- Data platform to facilitate sustainable development
- Data banker of national R&D programs
- Data bridge that facilitates international cooperation
- Data hub outreaches to public
- Data engine to promote entrepreneurship

谢谢！

敬请批评指正！

数据驱动地球系统科学新发现、新进展



“Data-intensive scientific discovery ” paradigm

李新 等，未发表