



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

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和科技创新能力的持续提升, 海量科学数据不断产生, 逐步成为重要的科技战略资源, 推动了科学研究向数据密集型科学模式的转换^[2]. 作为科研活动的灵魂, 科技创新越来越依赖于海量数据的集成、分析和利用. 因此, 积极开展科学数据治理和开放共享, 不仅能提升科学数据的服务和开发能力,



《中国科学》杂志社
SCIENCE CHINA PRESS



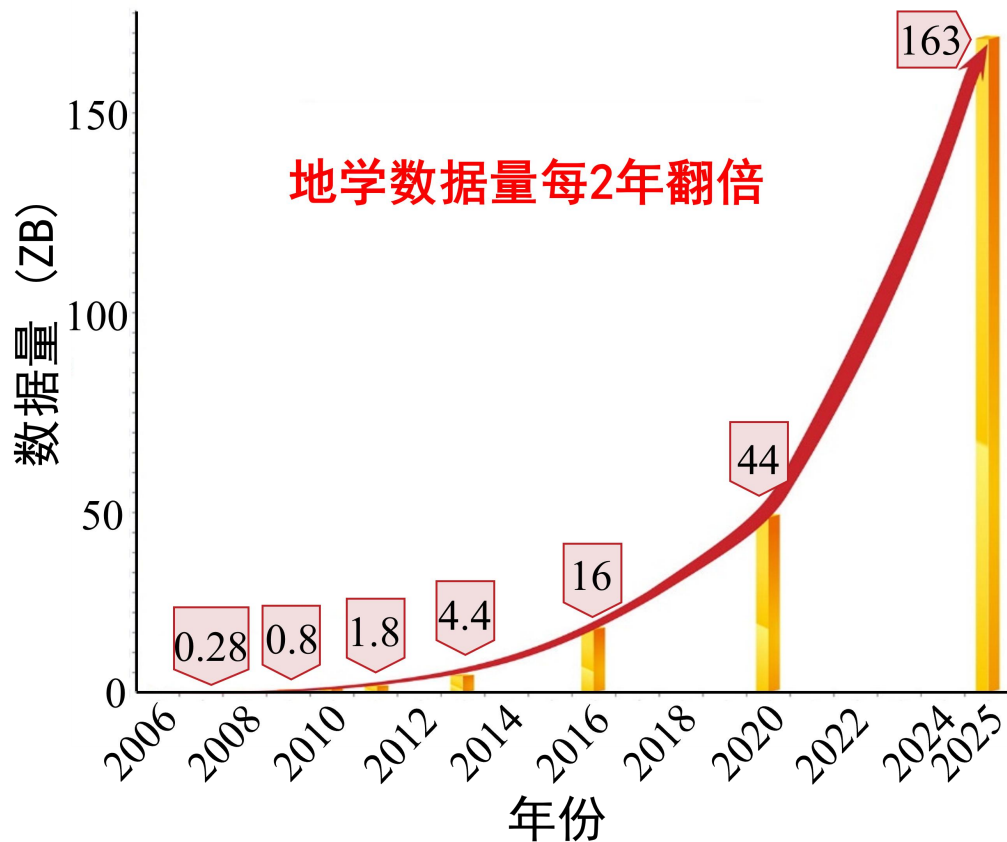
CrossMark
← click for updates



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.

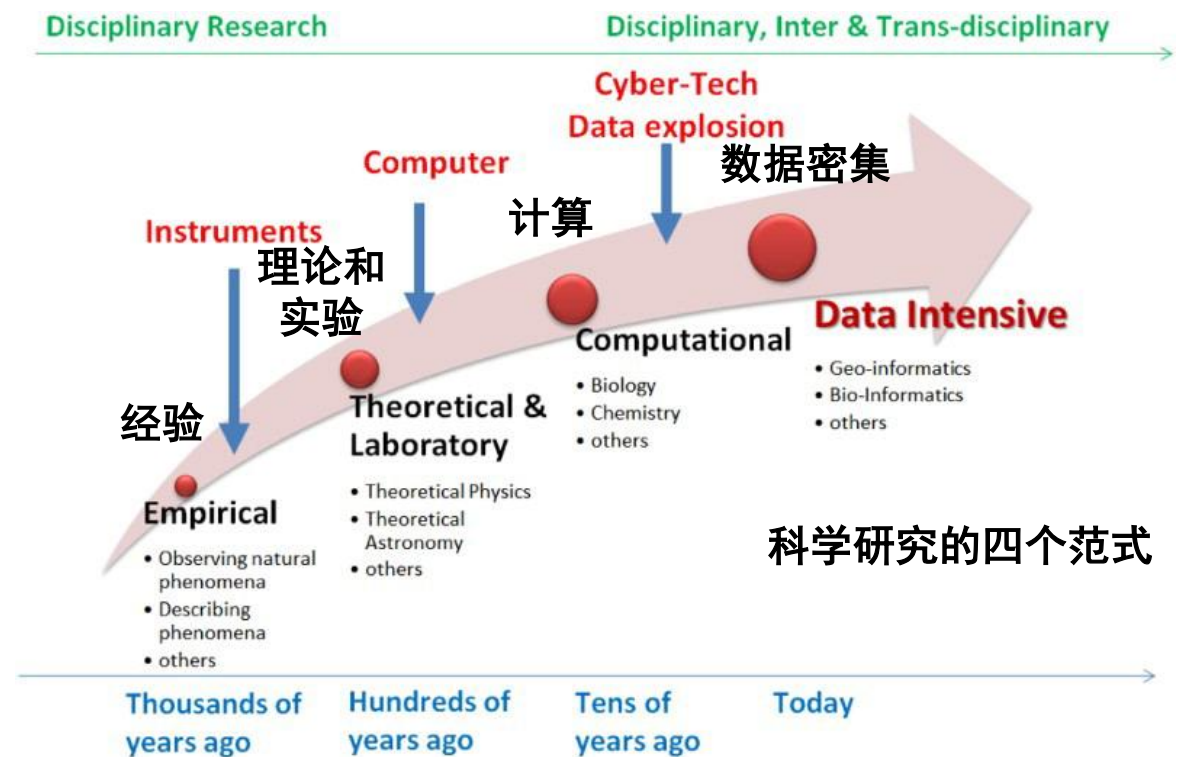
数据驱动的新地学

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

EDITORIAL
Making Data Maximally Available

SCIENCE IS DRIVEN BY DATA. NEW TECHNOLOGIES have increased the amount of data collected and consequently the amount of data independently mined and reanalyzed by others. Advances in data, for example, in responding to disease, to climate change, and in improving transportation are an essential element of scientific research. It is our basic responsibility toward transparency, as pointed out in a special section of this issue (pp. 692-729), to ensure that the huge amount, complexity, and variety of the data that are available after publication. Thus, Science and our strengthened their policies regarding data, and as per our online, added supporting online material (SOM) to ensure that data are available after publication. It is a growing challenge to ensure that the data produced during the course of reported research are deposited, standardized, archived, and available to the community. Science's policy for some time has been that “to understand, assess, and extend the conclusions, data must be available to any reader of Science” (see our article feature contribution). Besides publishing in our published papers (including those described above) we have encouraged authors to comply in one of two ways: depositing their data in public databases that are reliably maintained or, when such a database is not available, including their data in the SOM. However, only a few journals are not equipped to curate data without a plausible home, we have therefore required agreement, in which the author commits to archive a copy of the data held at Science. But such an agreement for permanent, community-maintained archiving is not sufficient to address the growing complexity of data and access requirements listed above to include computer analysis of data. To provide credit and reward, we want to produce a single list that combines references from the print list will be available in the online version of it will provide a template to curators in content to be sent to reviewers and readers. We will also ask authors to include the availability and curation of data as part of their cover letters. We consider this a responsibility of the author and not the publisher. We thus provide, for example, or in some cases when data or materials are obtainable reasons. But we expect these exceptions to be rare. As gatekeepers to publication, journals clearly have a responsibility to ensure that data are science, and thus provide for, and just improved data curation. —Brook

www.sciencemag.org SCIENCE VOL 331 111



Research cannot flourish if data are not preserved and made accessible...

Nature, 10 September 2009, special section



科学数据的开放、共享和应用，会促进科学界带来新的知识。而大数据、人工智能和大模型的融合发展，也将激活科学研究的创新力和生命力，破解更多科学密码。

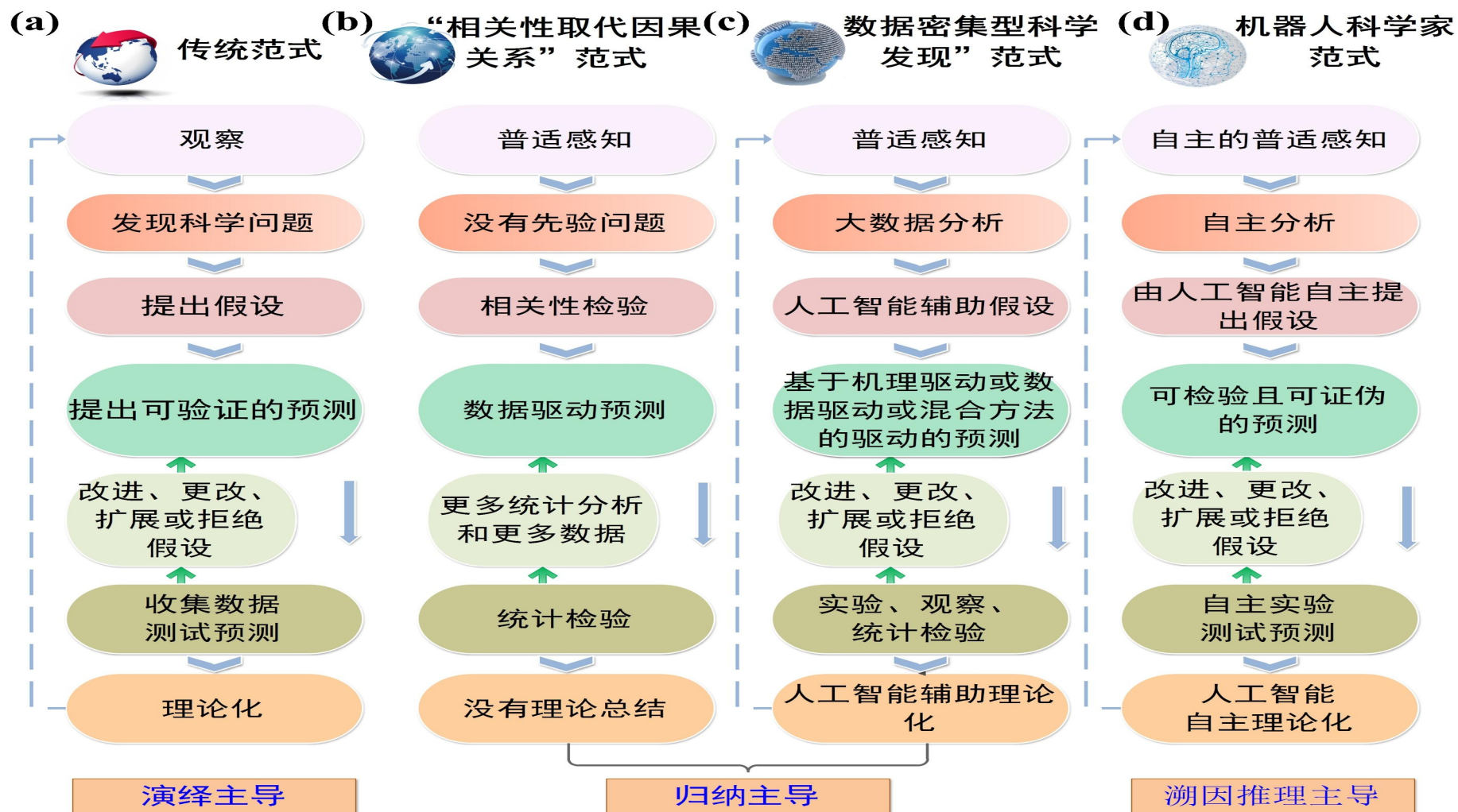
——陈润生 中国科学院院士 中国科学院生物物理研究所研究员

开放数据状况报告是一项全球调查，旨在深入了解科研人员对于开放数据的态度和体验。我们很高兴与中国科学院计算机网络信息中心携手，合作发布一份有关中国开放数据情况的报告，以便出版机构、科研资助机构和科研机构更好地了解科研人员的看法，以及需要以哪些支持来帮助他们将数据公开。作为科研界的积极合作伙伴，施普林格·自然致力于开创数据共享的新方法，并支持科研人员使数据共享成为新常态。

——STEVEN INCHCOOMBE (史蒂文·印驰库姆) 施普林格·自然 科研市场总裁

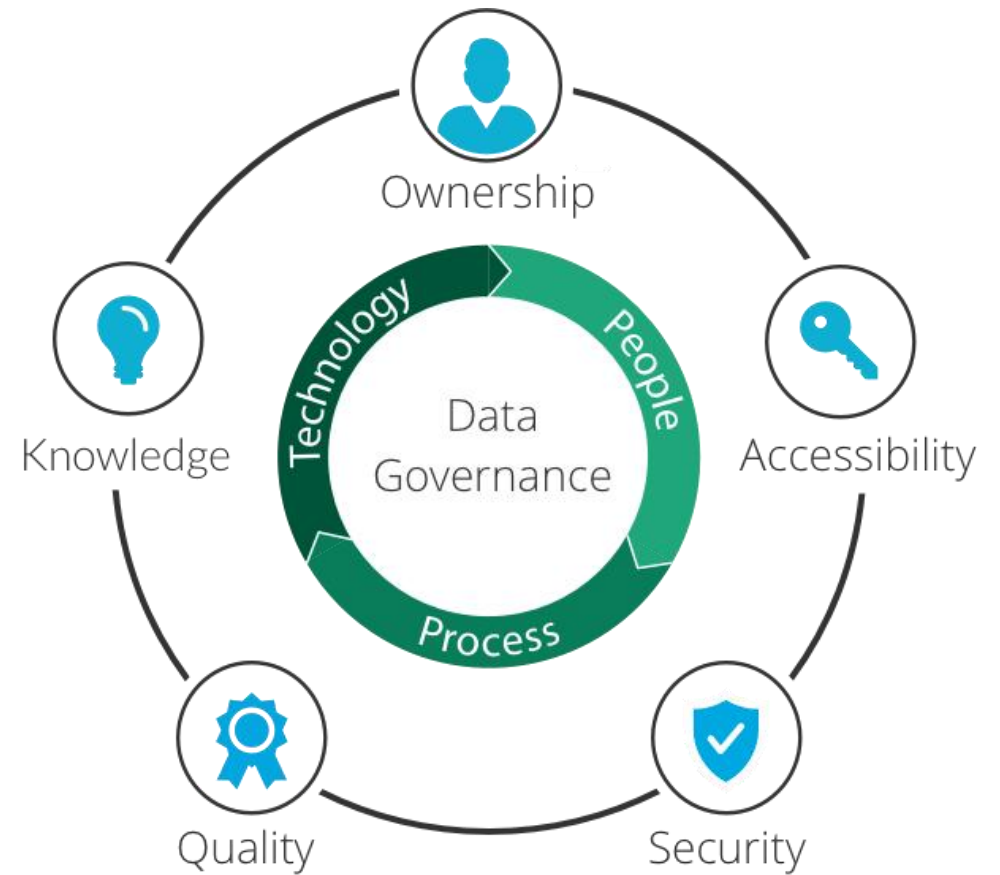
《中国开放数据白皮书2023》

地球系统科学研究中的范式转换

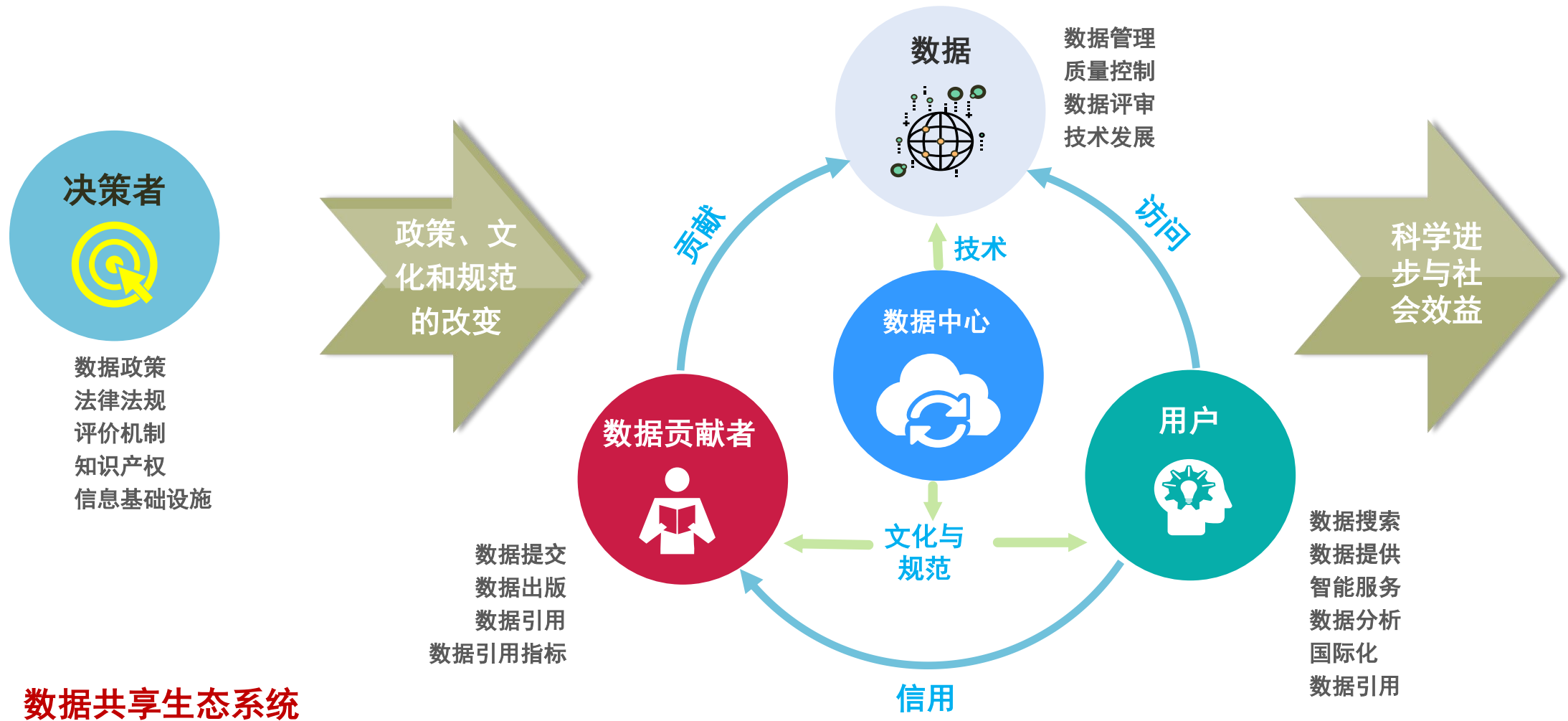


挑战1：从科学数据管理到科学数据善治

- **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](#))



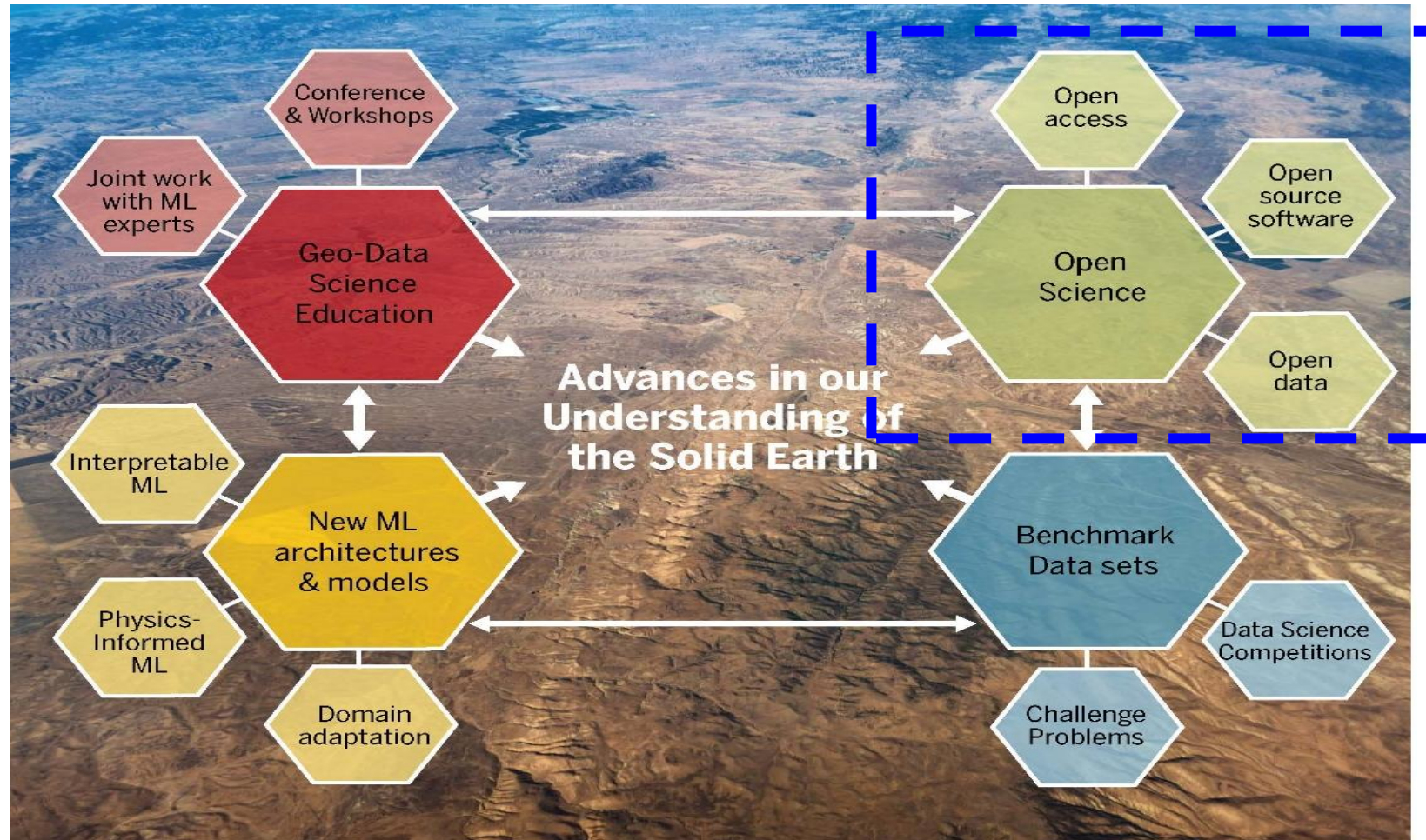
自上而下的政策约束 + 激励机制



地学数据善治：应对举措建议

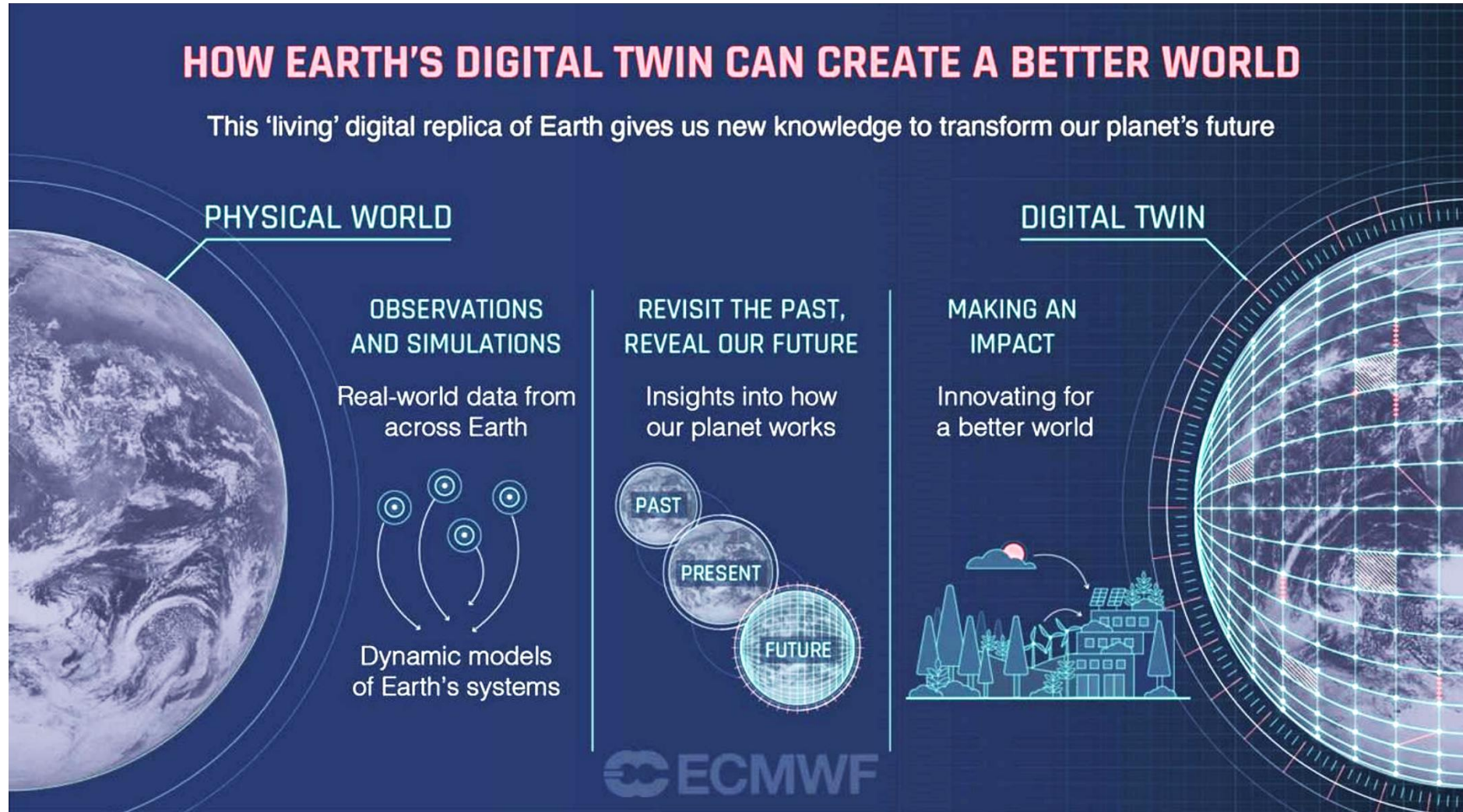
- 制定数据质量控制的度量指标和标准
- 启动数据引用，作为激励机制的杠杆
- 加强与期刊（优先国内的一流期刊，如NSR, Science Bulletin, 中国科学）之间的合作，加快论文关联数据仓储
- 深度参与国际上地学数据治理（FAIR, TRUST, CARE）
- 细粒度、可操作地贯彻**FAIR**原则
- 鼓励无差别、不设限的数据访问

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



SDG大数据平台

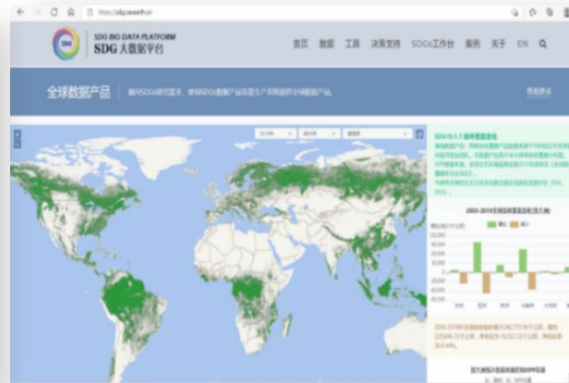
促进我国开放数据基础设施不断产出重要成果，服务SDG目标监测评估

- 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



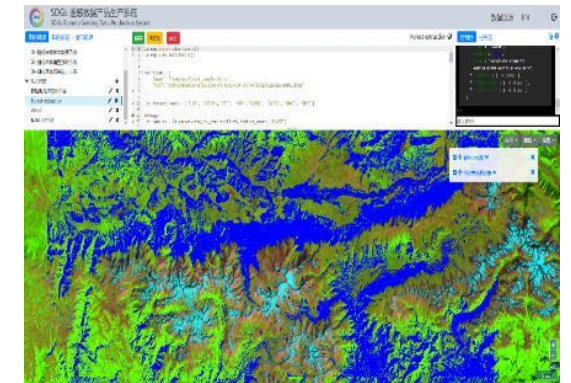
SDG大数据平台



SDG决策支持系统



SDG学科交叉协同

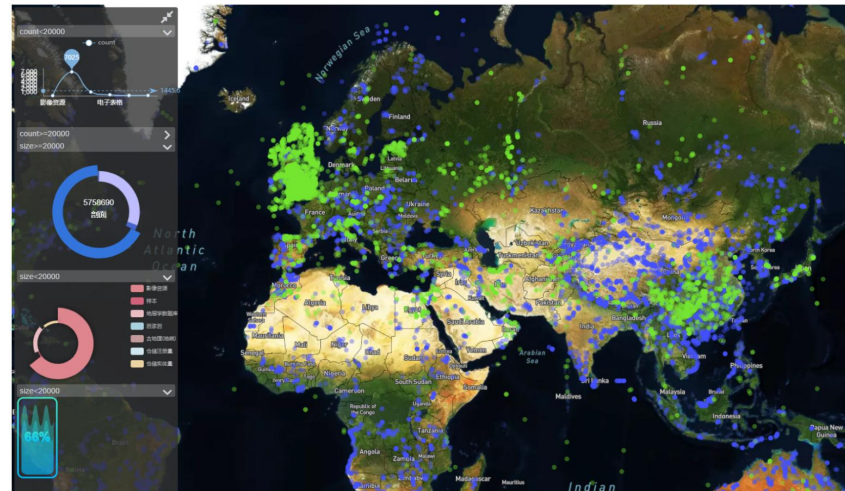


“深时数字地球” 国际大科学计划

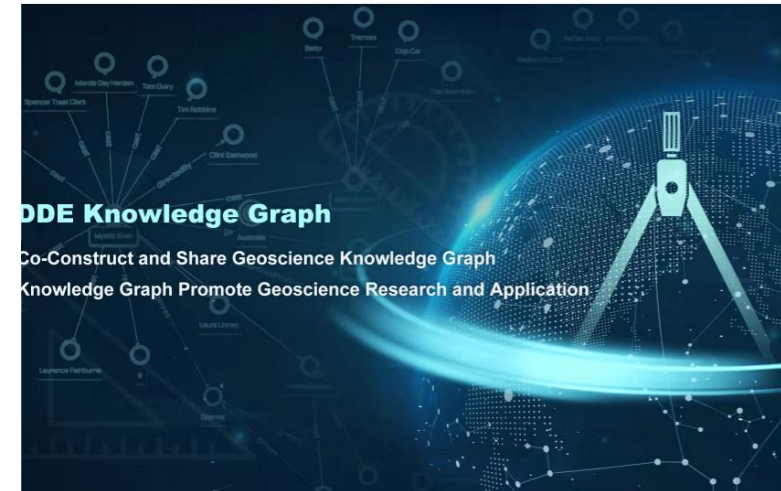
整合地球演化全球数据，打造数字时代具有竞争力的全球科技公共产品



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

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”

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

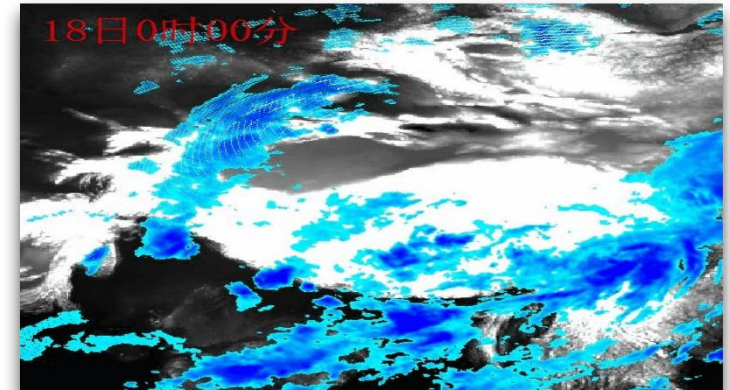
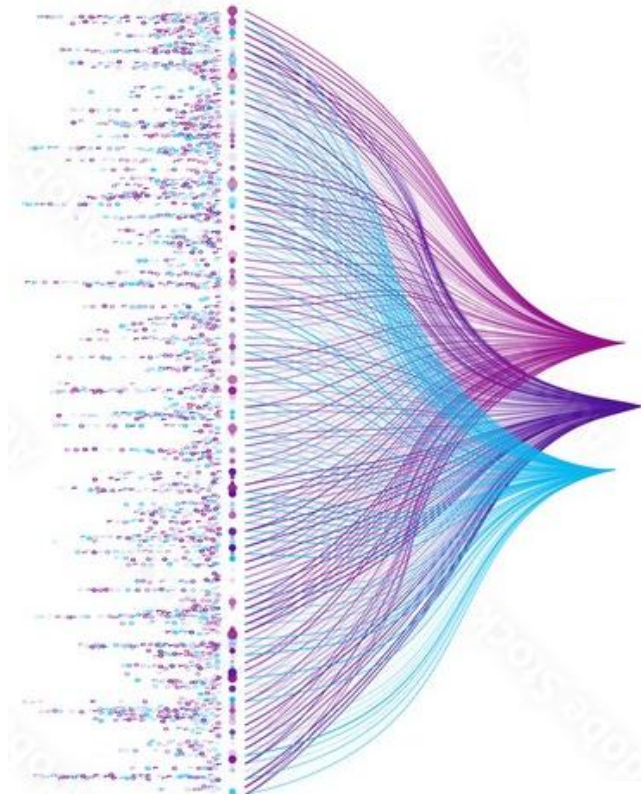


TPDC国家青藏高原科学数据中心

第二次青藏科考西藏24个湖泊水质数据集 (2019)

Water quality data sets of 24 lakes in Tibet for the second Qinghai Tibet scientific research (2019)

2019年8-9月第二次青藏科考共采集色林错、纳木错及周边共计24个湖泊的水质样品，分析了叶绿素 (CHL, 单位为微克每升)、总氮 (TN, 单位为毫克每升)、总磷 (TP, 单位为毫克每升)、溶解性总氮 (DTN, 单位为毫克每升)、溶解性总磷 (DTP, 单位为毫克每升)、硝态氮 (NO₃-N, 单位为毫克每升)、亚硝态氮 (NO₂-N, 单位为毫克每升)、铵态氮 (NH₄-N, 单位为毫克每升) 及磷酸盐 (PO₄-P, 单位为微克每升)、总悬浮颗粒物 (TSS, 单位为毫克每升)、有机悬浮颗粒物 (OSS, 单位为毫克每升)、无机悬浮颗粒物 (ISS, 单位为毫克每升)。同时提供样点所在湖泊名称、湖泊简写及点位所在经纬度数据，数据格式为xlsx。数据均为实验室手工分析，并经科研人员反复核验，真实可靠。



现代气候变化过程

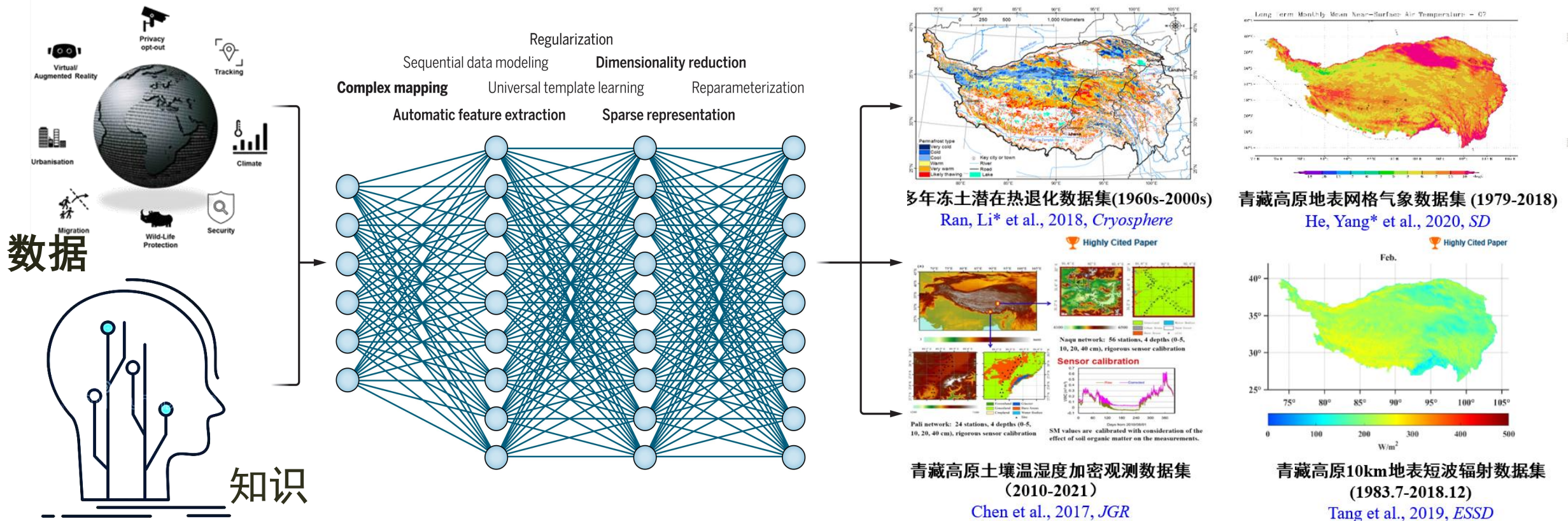


动力演化过程

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

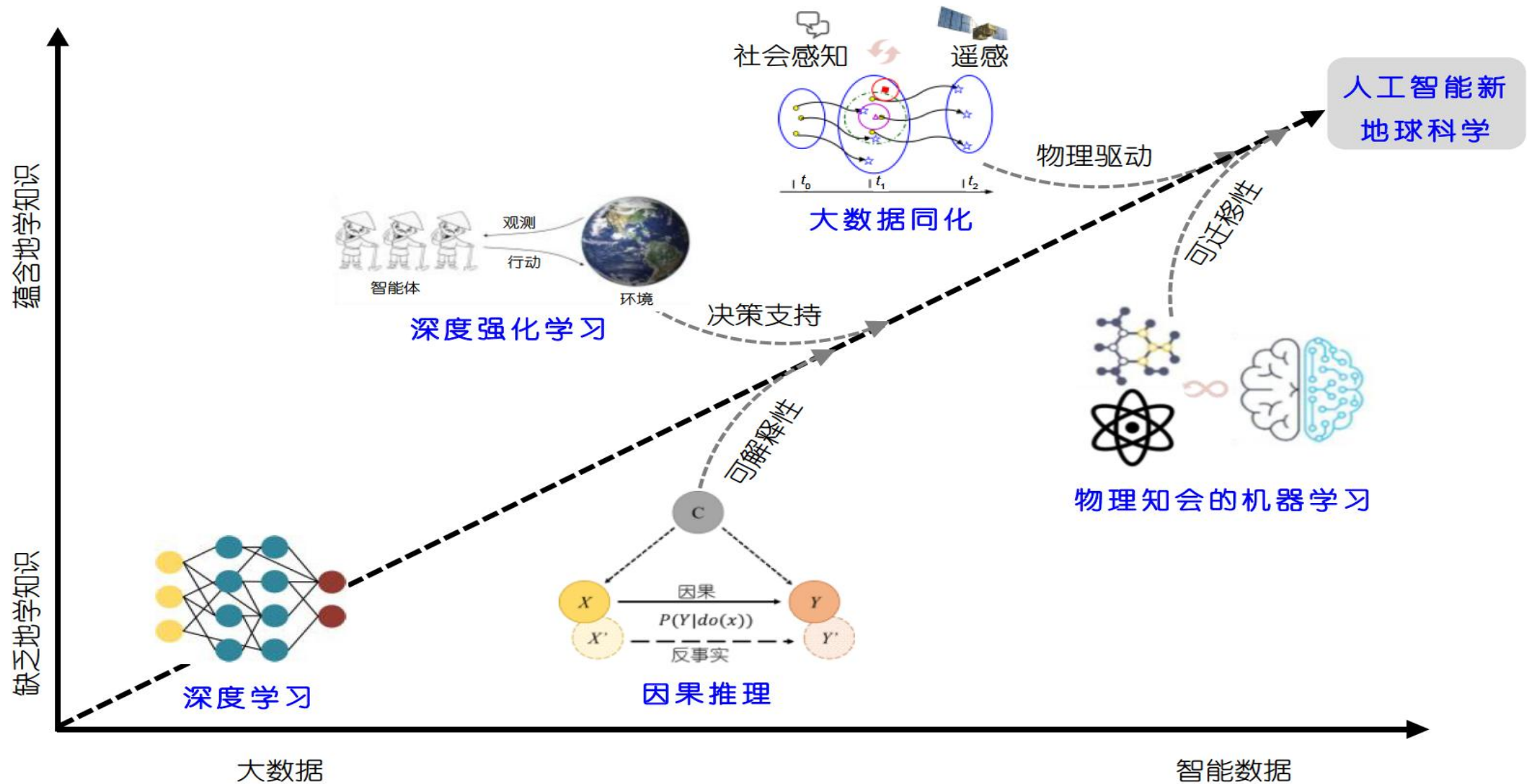
从原始数据资源向高质量数据产品的转化，逐步打造“智能生产”、“快速定制”技术，形成支持地球系统科学和可持续发展的旗舰数据产品



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

挑战4：前沿大数据分析方法，走向AI for geoscience



大数据分析的前沿

动力系统
Dynamic system

优化
Optimization

Bayes框架

$$\begin{aligned} \operatorname{argmax} \mathcal{P}(\mathbf{x}_{K:0} | \mathbf{y}_{K:1}) &\propto \prod_{k=1}^K \mathcal{P}(\mathbf{y}_k | \mathbf{x}_k) \mathcal{P}(\mathbf{x}_k | \mathbf{x}_{k-1}) \\ &= \mathcal{P}(\mathbf{x}_0) \prod_{k=1}^K \mathcal{P}_\epsilon[\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)] \mathcal{P}_\eta[\mathbf{x}_k - \mathcal{M}_{k:k-1}(\mathbf{x}_{k-1})] \end{aligned}$$

数据 $\bar{\mathbf{y}}$

模型 $\bar{\mathcal{M}}$

误差 ϵ

统计学
Statistics

反演问题
Inversion

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



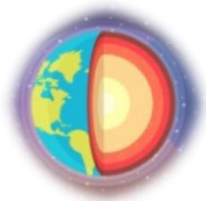
陆地

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

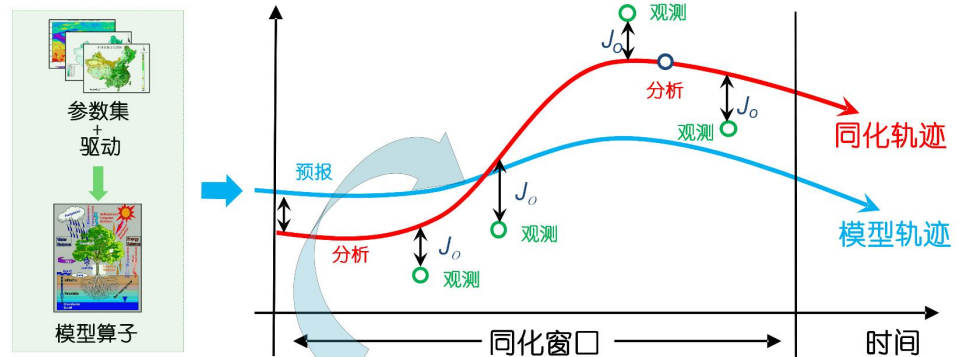


固体地球

地幔动力学
地震
地球物理
地球化学

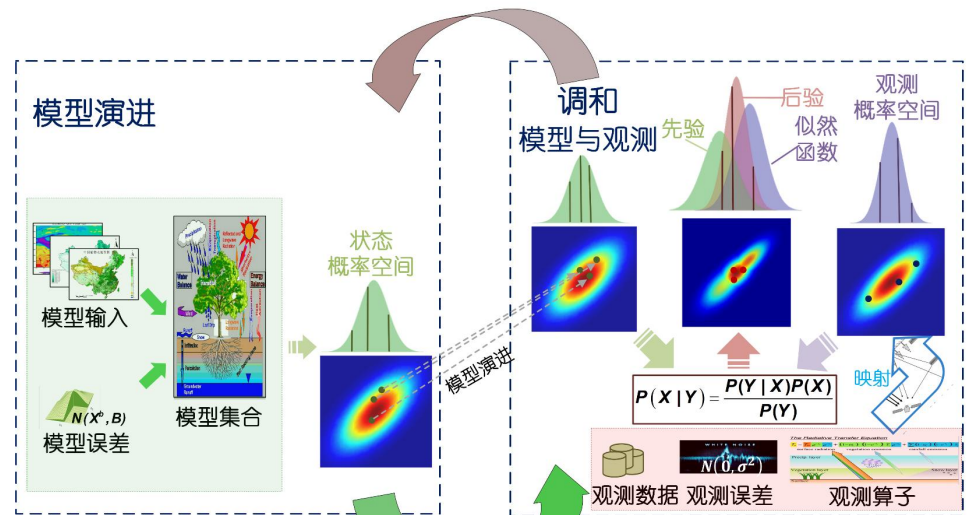
李新 等, 2020. 模型与观测的和弦: 地球系统科学中的数据同化. 中国科学: 地球科学

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



$$J_o(x_0) = \frac{1}{2}(x_0 - x_0^b)^T B^{-1}(x_0 - x_0^b) + \frac{1}{2} \sum_{i=0}^n (H_i\{[M(x_0)]_i\} - y_i^o)^T R_i^{-1}(H_i\{[M(x_0)]_i\} - y_i^o)$$

(a)



(b)

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

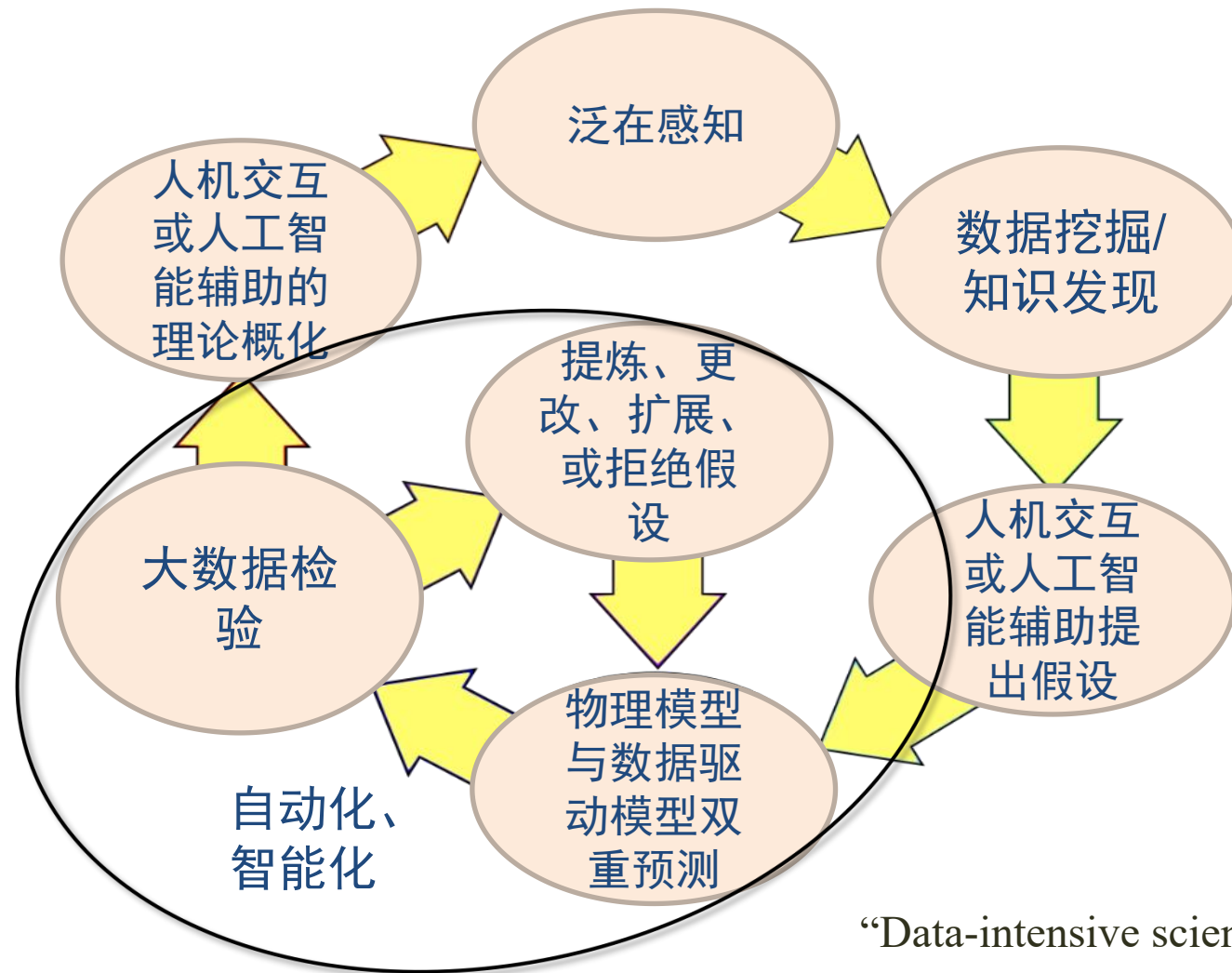
-
- ❑ Innovator and leader of data science
 - ❑ Booster of big data revolution
 - ❑ Data platform to facilitate sustainable development
 - ❑ Data banker of national R&D programs
 - ❑ Data bridge that facilitates international cooperation
 - ❑ Data hub outreaches to public
 - ❑ Data engine to promote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



李新 等, 未发表