データ活用社会創成プラットフォームmdxにおけるマテリアルズ・インフォマティクス研究・共創に向けて

Accelerating interdisciplinary research between material science and computer science on the *mdx* data science platform

Toyotaro Suzumura and Masatoshi Hanai
Information Technology Center
The University of Tokyo
鈴村豊太郎・華井雅俊
東京大学情報基盤センター

データ活用社会創成プラットフォーム mdx の設計・実装・運用
～多様な学際領域における共創に向けて～

鈴村 豊太郎[1)]，杉木 章義[2)]，滝沢 寛之[3)]，今倉 暁[4)]，中村 宏[1)]，田浦 健次朗[1)]，
工藤 知宏[1)]，塙 敏博[1)]，関谷 勇司[1)]，小林 博樹[1)]，松島 慎[1)]，空閑 洋平[1)]，中村 遼[1)]，
姜 仁河[1)]，川瀬 純也[1)]，華井雅俊[1)]，宮嵜 洋[5)]，石﨑 勉[5)]，下徳 大祐[5)]，関本義秀[6)]，
樫山武浩[6)]，合田 憲人[7)]，竹房 あつ子[7)]，政谷 好伸[8)]，栗本 崇[9)]，笹山 浩二[9)]，
北川 直哉[9)]，藤原 一毅[10)]，朝岡 誠[10)]，中田秀基[11)]，谷村 勇輔[11)]，青木 尊之[12)]，
遠藤 敏夫[12)]，大島 聡史[13)]，深沢圭一郎[14)]，伊達 進[15)]，天野 浩文[16)]

1）東京大学 情報基盤センター
2）北海道大学 情報基盤センター
3）東北大学 サイバーサイエンスセンター
4）筑波大学システム情報系
5）東京大学 情報システム部情報基盤課
6）東京大学空間情報科学研究センター
7）国立情報学研究所 アーキテクチャ科学研究系
8）国立情報学研究所クラウド基盤研究開発センター
9）国立情報学研究所学術ネットワーク研究開発センター
10）国立情報学研究所 オープンサイエンス基盤研究センター
11）産業技術総合研究所 デジタルアーキテクチャ研究センター
12）東京工業大学 学術国際情報センター
13）名古屋大学 情報基盤センター
14）京都大学 学術情報メディアセンター
15）大阪大学 サイバーメディアセンター
16）九州大学 情報基盤研究開発センター

suzumura@ds.itc.u-tokyo.ac.jp

The *mdx* Project Report 2021
A Large-Scale Platform for Accelerating Cross-Disciplinary Research Collaborations Towards Data-Driven Societies

Toyotaro Suzumura[1)], Akiyoshi Sugiki[2)], Hiroyuki Takizawa[3)], Akira Imakura[4)], Hiroshi Nakamura[1)], Kenjiro Taura[1)], Tomohiro Kudoh[1)], Toshihiro Hanawa[1)], Yuji Sekiya[1)], Hiroki Kobayashi[1)], Shin Matsushima[1)], Yohei Kuga[1)],Ryo Nakamura[1)], Renhe Jiang[1)], Junya Kawase[1)], Masatoshi Hanai[1)], Hiroshi Miyazaki[5)],Tsutomu Ishizaki[5)], Daisuke Shimotoku[5)],Yoshihide Sekimoto[6)],Takehiro Kashiyama[6)], Kento Aida[7)], Atsuko Takefusa[7)], Yoshinobu Masatani[8)], Takashi Kurimoto[9)], Koji Sasayama[9)], Naoya Kitagawa[9)], Ikki Fujiwara[10)], Makoto Asaoka[10)], Hidemoto Nakada[11)], Yusuke Tanimura[11)],Takayuki Aoki[12)], Toshio Endo[12)], Satoshi Ohshima[13)], Keiichiro Fukazawa[14)], Susumu Date[15)], Hirofumi Amano[16)]

1) Information Technology Center, The University of Tokyo
2) Hokkaido University Information Initiative Center
3) Cyberscience Center, Tohoku University
4) Faculty of Engineering, Information and Systems, University of Tsukuba
5) Division for Information and Communication Systems, The University of Tokyo
6) Center for Spatial Information Science, The University of Tokyo
7)Information Systems Architecture Science Research Division, National Institute of Informatics
8)Center for Cloud Research and Development, National Institute of Informatics
9)Research and Development Center for Academic Networks, National Institute of Informatics
10)Research Center for Open Science and Data Platform, National Institute of Informatics

Outline

- **mdx: A platform for the data-driven future**
- **Material Science on *mdx***

Motivation

mdx

- **Accelerating inter-disciplinary research and collaborations**
  - Data science has become a key driver to advance science and technology, but more collaborations among various domain experts in both academia and private sectors should be formulated to solve complex problems in real-world society (e.g. material discovery, carbon-neutral society).
- **How can we accomplish this ?**
  - Agility : Need a national-wide cloud platform (IaaS and PaaS/SaaS) that enables researchers to promptly set up their own data analytics environments by allowing them to install/configure their required analytics softwares, and publish data repository services as needed
  - Connecting with Edge Devices (e.g. Electron Microscopes, IoT sensors) via High-Speed Network:
  - Coupling with HPC environments :
    - Some applications need big computing powers for such applications as large-scale simulations and large-scale AI/ML training thus the platform should be coupled with supercomputers when needed.
- **Why can't we rely sololy on supercompters or public clouds?**
  - These requirements can not be fulfilled by current types of supercomputers mainly targeting big science, and also public clouds from private sectors

Status in the U.S.

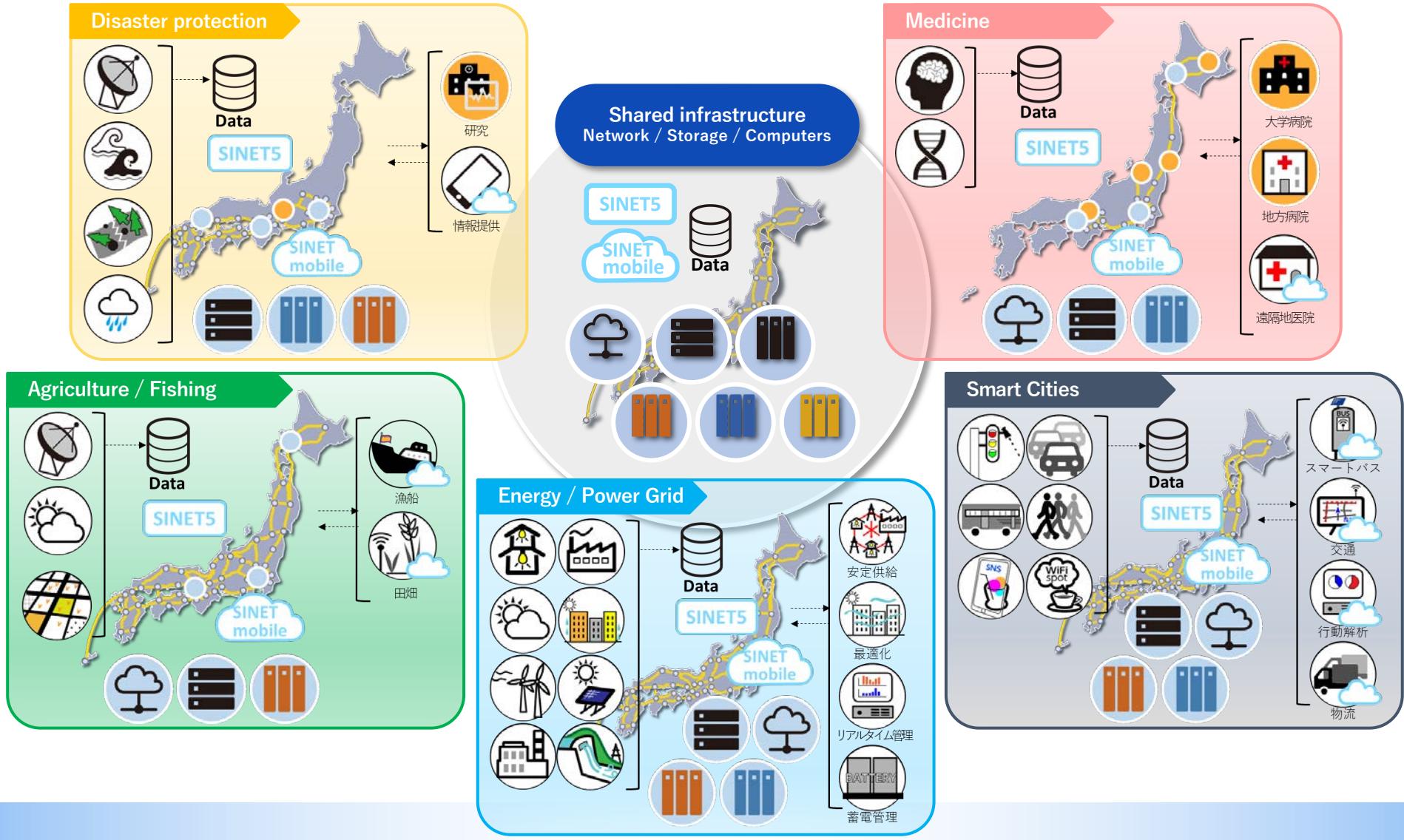
- Ian Foster, Daniel Lopresti, Bill Gropp, Mark D. Hill, Katie Schuman. "***A National Discovery Cloud: Preparing the US for Global Competitiveness in the New Era of 21st Century Digital Transformation***"
  - 2021 Apr https://arxiv.org/abs/2104.06953
  - The need for national discovery cloud:
    - *DOE and NSF supercomputers provide access to powerful simulation capabilities, but with access limited to small communities.*
    - *With a few notable exceptions, AI-ready datasets for research use are lacking. Commercial clouds are accessible to anyone with a credit card, but there is little of the coordination needed to create nationally useful discovery cloud services.*

mdx: A Platform for the Data-Driven Future

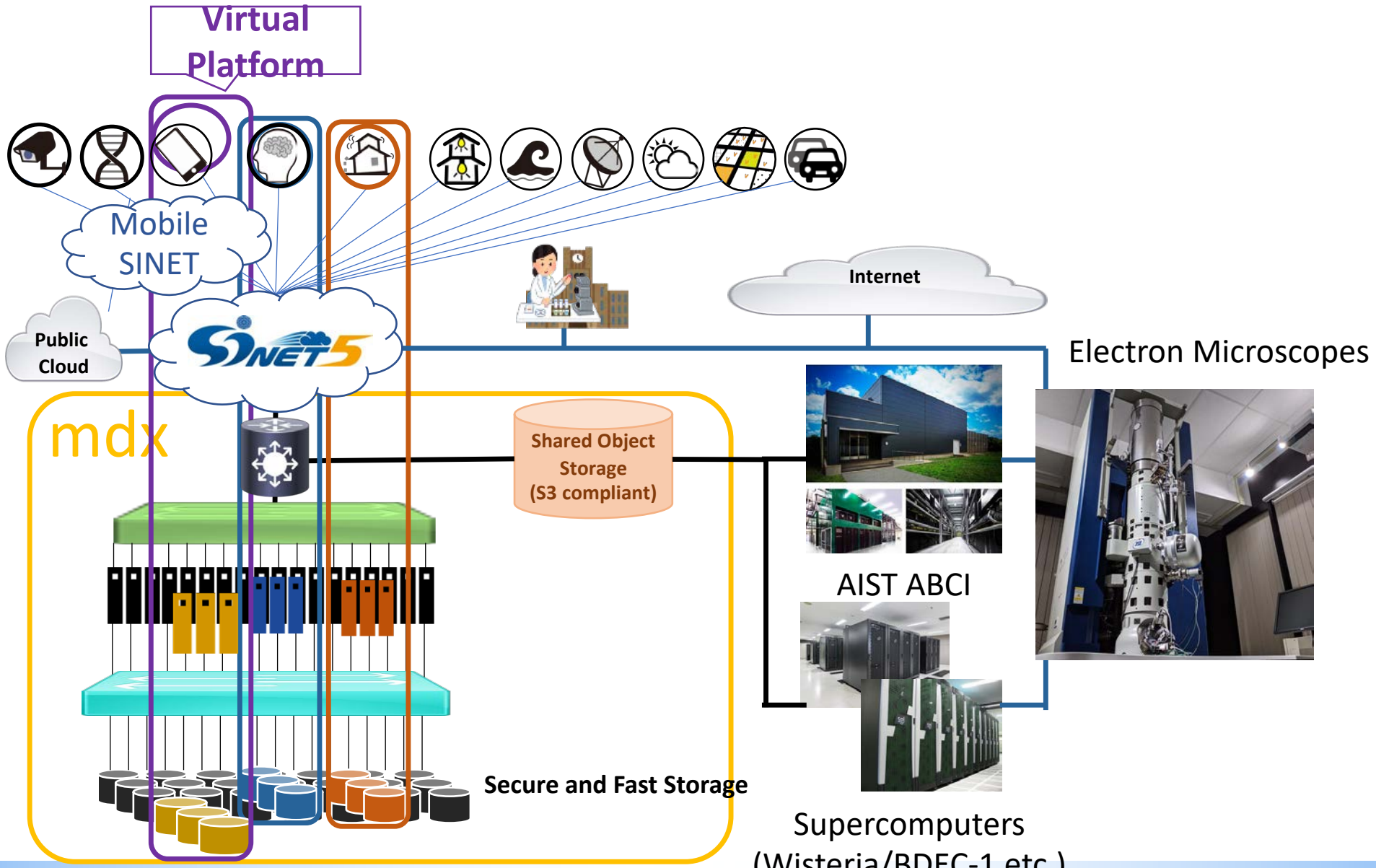
- Target is to leverage data utilization at all over Japan making full use of high performance R&E network “SINET” an R&E network of Japan operated by NII (National Institute of Informatics)
- Project supported by the Japanese government
- Currently jointly being operated by:
  - 9 National Universities (Tokyo, Hokkaido, Tohoku, Tsukuba, Tokyo Tech, Nagoya, Kyoto, Osaka, Kyushu)
  - NII (National Institute of Informatics)
  - AIST (National Institute of Advanced Industrial Science and Technology)
- Invite universities and public research institutes of all over Japan to use the platform for industry-academia and local government-academia collaboration activities.
- Production-level operation has been started since March 2021

mdx as On-demand Collaborative Platform

Conduct data analytics from various locations via SINET or Mobile SINET



mdx : Providing Secure and On-demand Virtual Platform
(Optionally) with access to Supercomputers and Edge Devices



mdx: A Platform for the Data-Driven Future

- Agility : Provide **a rapid PoC environment** to accelerate R&D data utilization as well as industry-academia collaboration projects.
  - Sharing: Shared platform for various data utilization activities
  - Network: Combine a high-performance wide-area academic network called “SINET” with high performance computing and storage infrastructure
- Seamless Integration with Edge Devices:
  - Users can use wide bandwidth low latency “slices”
  - Wide-area virtual infrastructure isolated from “the internet”
  - Connect edge devices with high performance computing and storage infrastructure and supports real-time data processing
- Matching: Will provide matching function of:
  - Data providers: various data and their owners, and
  - Data scientists and researchers who have skills/tools to analyze data

Hardware Overview

mdx

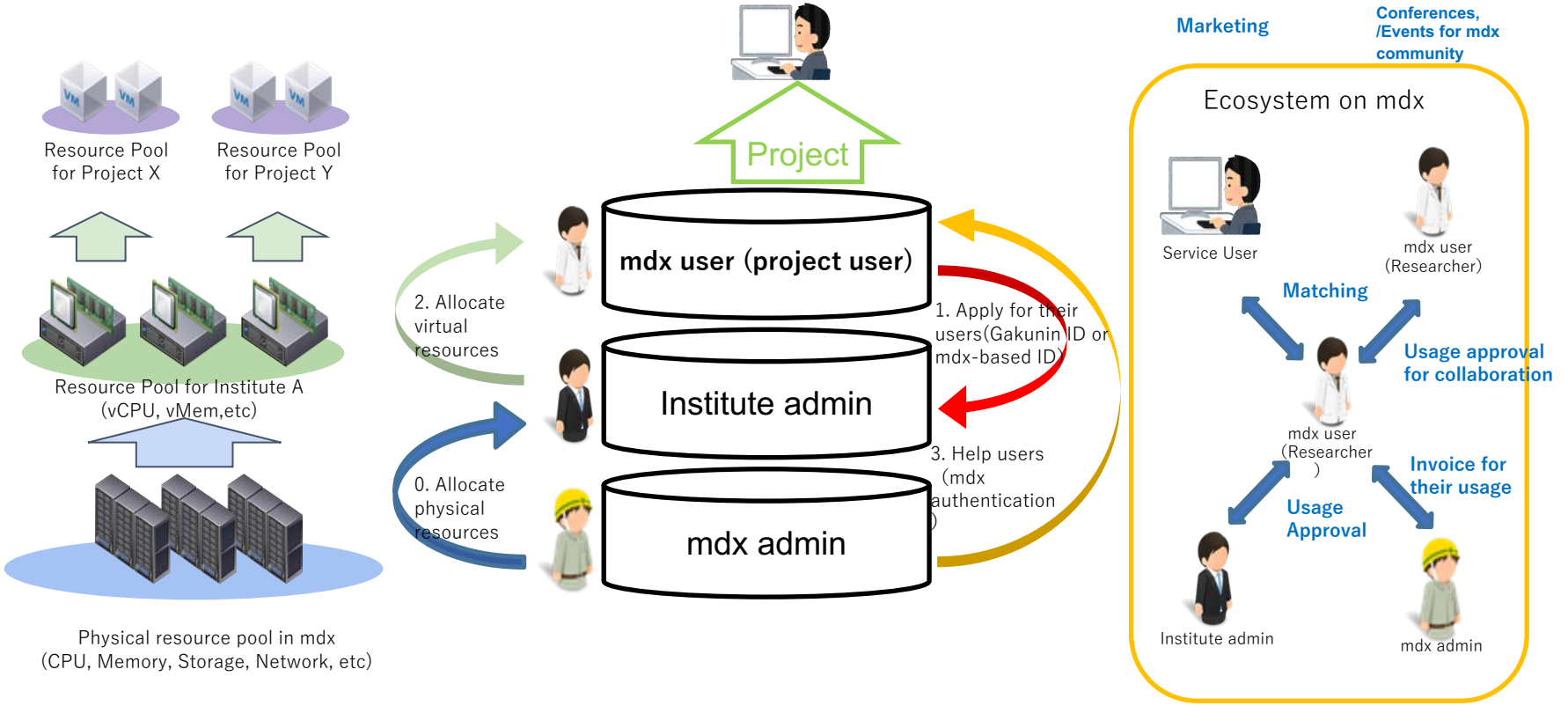
- **Facility**
  - < **2.0 MW** including Cooling, <170 m^2
  - Same location with Wisteria/BDEC-01, same campus with AIST's ABCI supercomputer
- **Compute nodes (CPU)**
  - 368 nodes : Each node : Intel Xeon (IceLake-SP, 38 cores) x 2 CPU sockets/node
  - 2.1 Peta flops (double precision)
  - Total memory bandwidth : 150 TB/sec
- **Compute nodes (GPU)**
  - 40 nodes, Intel Xeon (IceLake-SP) x2 socket
  - NVIDIA A100 x 8 GPUs/node
  - 6.4 Peta Flops (FP64), 6.7 PF (FP32), 100 PF (FP16),
  - Total memory bandwidth: 496 TB/sec
- **Storage**
  - Fast Storage with NVMe SSD: 1.0 PB, 250 GB/sec
  - Large Storage with HDD: 16.3 PB. 157 GB/sec
  - Shared Object Storage (S3): 10 PB, 63 GB/sec
- **File System** : Lustre

- **Network**
  - Frontend (Juniper) : 25 Gbps Ethernet
    - 100G to SINET
      - 400G to Wisteria/BDEC-01
  - Storage, RDMA (Mellanox/NVIDIA) : 100G Ethernet with RoCEv2
  - Overlay with EVPN-VXLAN
- **Software, etc.**
  - VM & Container (VMware vSphere)
  - IaaS like management
  - High security, high availability

New building in Kashiwa II Campus

Resource Management and Scheduling

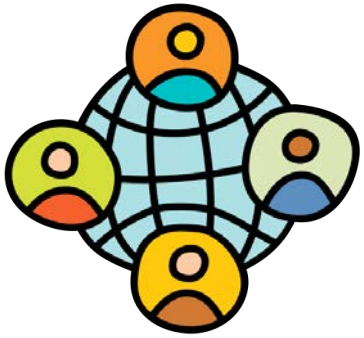
- Each institution is responsible for approving project applications and their required hardware resources



Data Community: Matching Data Providers and Data Scientists

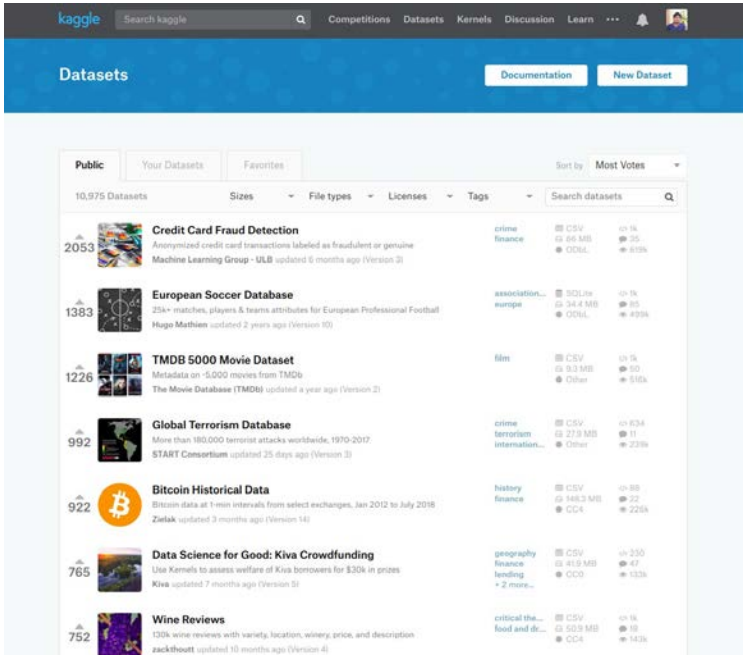
- **Data Community on mdx**
  - As of Jan 2022, we are currently in the midst of designing a PaaS level platform that establishes a community between data providers and data scientists
  - The alpha version: January 2023 (Plan)
- **Requirements**
  - **Data providers** can easily upload their data to ”mdx” by specifying the spec of data and data usage conditions (e.g. only for research purpose)
  - **Data scientists** can easily find data based on their interests, and launch Jupyterlab

Data Providers



Data Scientists

Example) Kaggle-like community

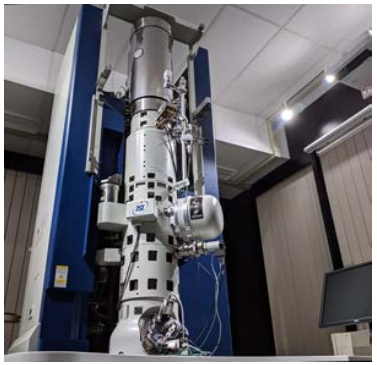


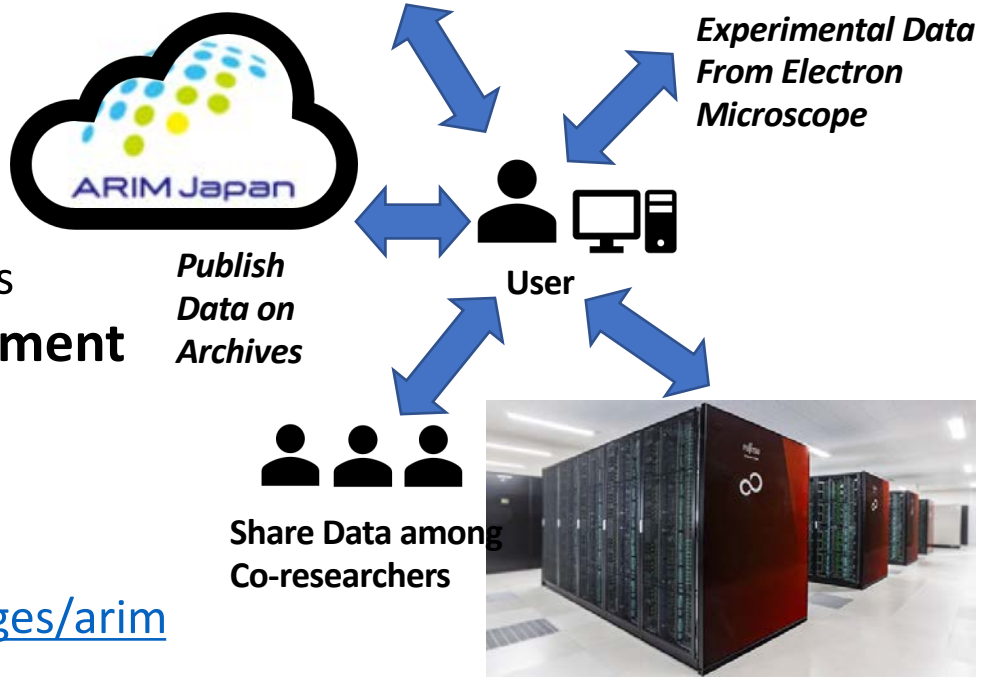
Outline

- **mdx: A platform for the data-driven future**
- **Material Science on *mdx***

Usecase: Material Science and Engineering

- Key Challenges
  1. **Big data** from laboratory instruments
     - Image data from electron microscope
     - up to 1TB / experiment
  2. **High computational-power requirement**
     - Physics simulations (e.g., First-principle calculation)
     - Machine learning / data analysis
  3. **Flexible & secure data management**
     - Share confidential data among co-researchers
     - Publish open data on archives (e.g., ARIM Japan https://www.nanonet.go.jp/pages/arim/index.html)

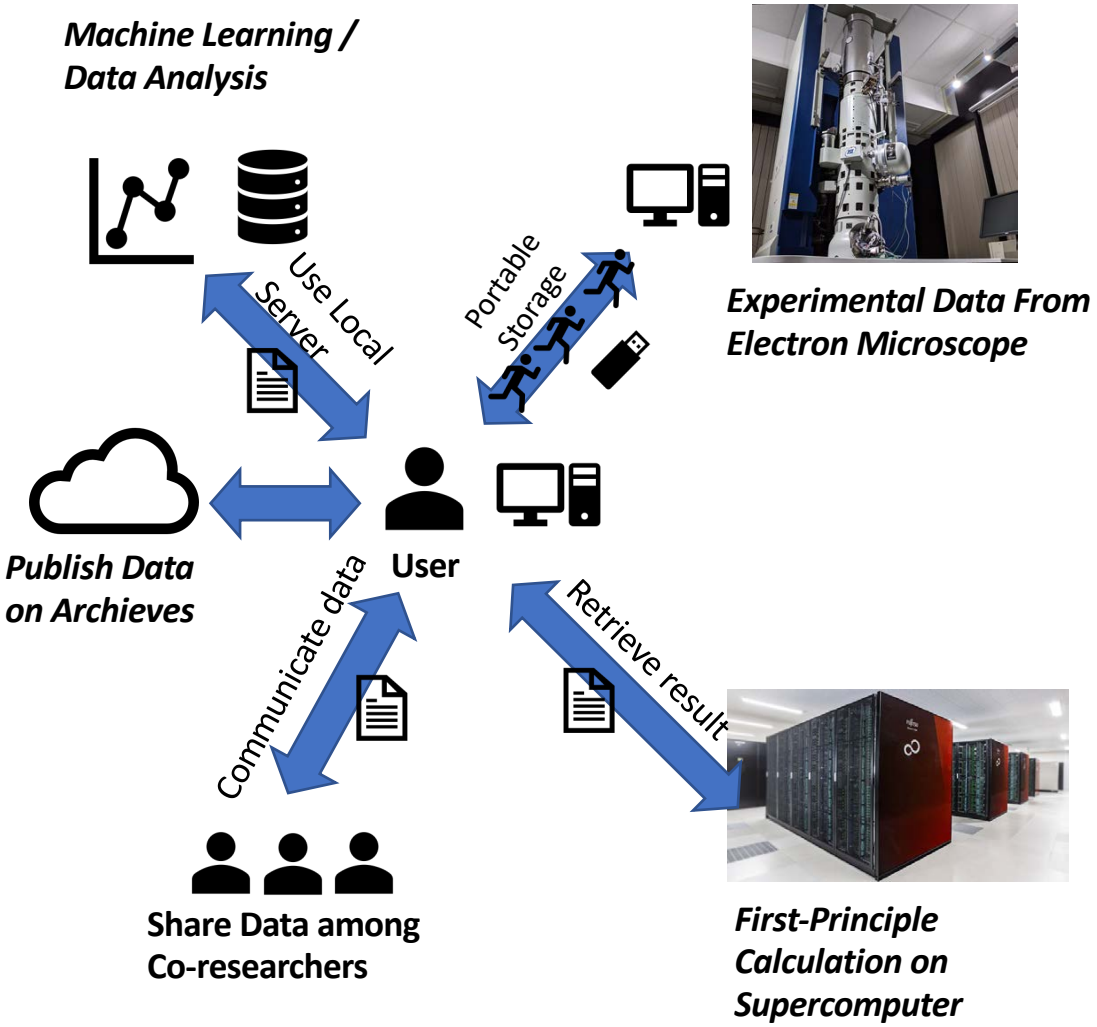




Usecase: Material Science and Engineering

Traditional ways

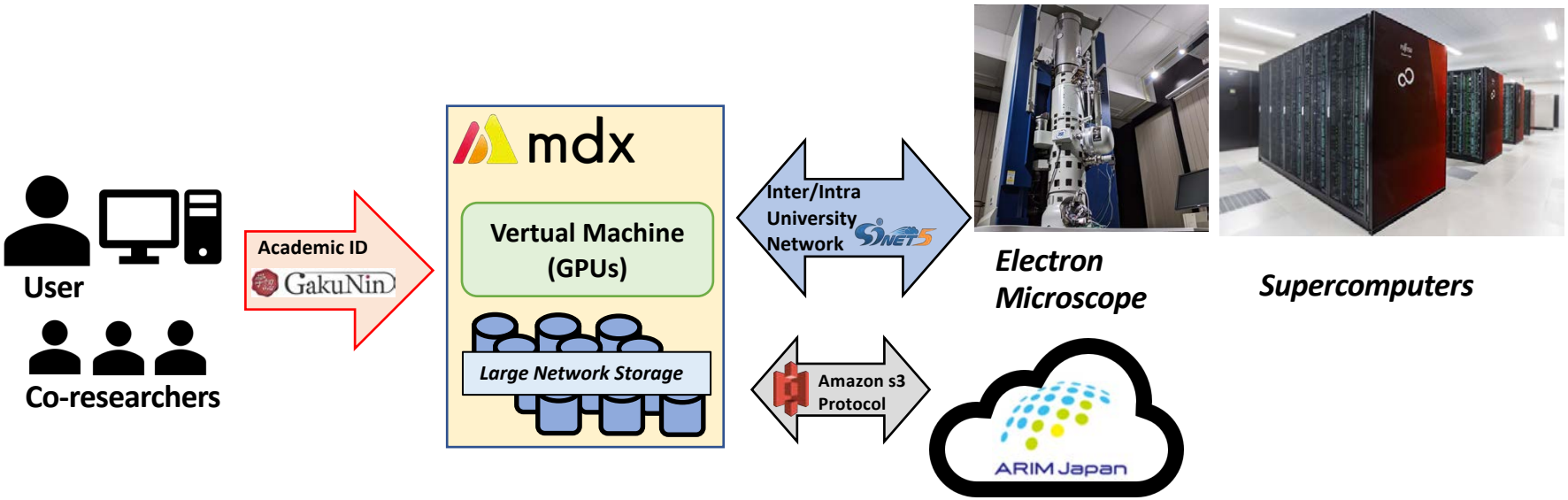
- Machine learning / data analysis on local servers
- Get data from laboratory instruments via portable storage
- Access supercomputers for First-principle calculation
- Send data to co-researchers
- etc.



Usecase: Material Science and Engineering

System Integration with mdx:

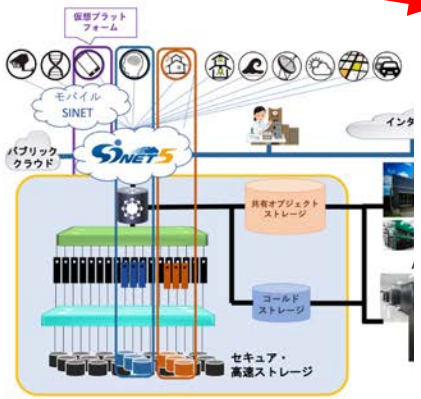
- Install **mdx** as the extension of **"local sever"** with high performance and large storage
  - Customizable VM environment (GPU-available)
  - Store all data in the large storage on mdx (Lustre)
- **Secure & High-performance Inter-University Network** (SINET)
  - Data from laboratory instrument are via SINET (or intra-university network)
  - Seamless workload extension to academic supercomputers
    - Machine learning (GPUs on mdx) ⇔ First-principle calculation (Supercomputer)
- **Publish data via Amazon S3 protocol**



Comparison to Other Services

	OS-level Management	Customizable Resources	High Computational Power	Large-Scale Accessible Storage	Secure & High-End Academic Network	Open Data Publication
mdx	✓	✓	✓	✓	✓	✓
Local Lab PC/Server	✓				✓	
Supercomputer			✓		✓	
Enterprise Cloud	✓	✓		✓		✓

mdxは、高性能な計算機と大容量のストレージを備え、国立情報学研究所が運用する学術情報ネットワークSINET5（2022年度から次期システムに更新予定）と連携することで、広域からのデータ収集機能と、データ集積・処理機能を、企業や自治体との共同研究も含めた全国の大学・公的研究機関が関与する様々なデータ活用の取組に提供し、さらにはデータ活用のコミュニティーを形成して分野・セクタを横断した連携を触媒するハブとなることを目指します。



mdxについて

2021年に稼働を予定している、データ利活用、データ科学、に重点を置いた計算基盤を紹介します。

運営組織

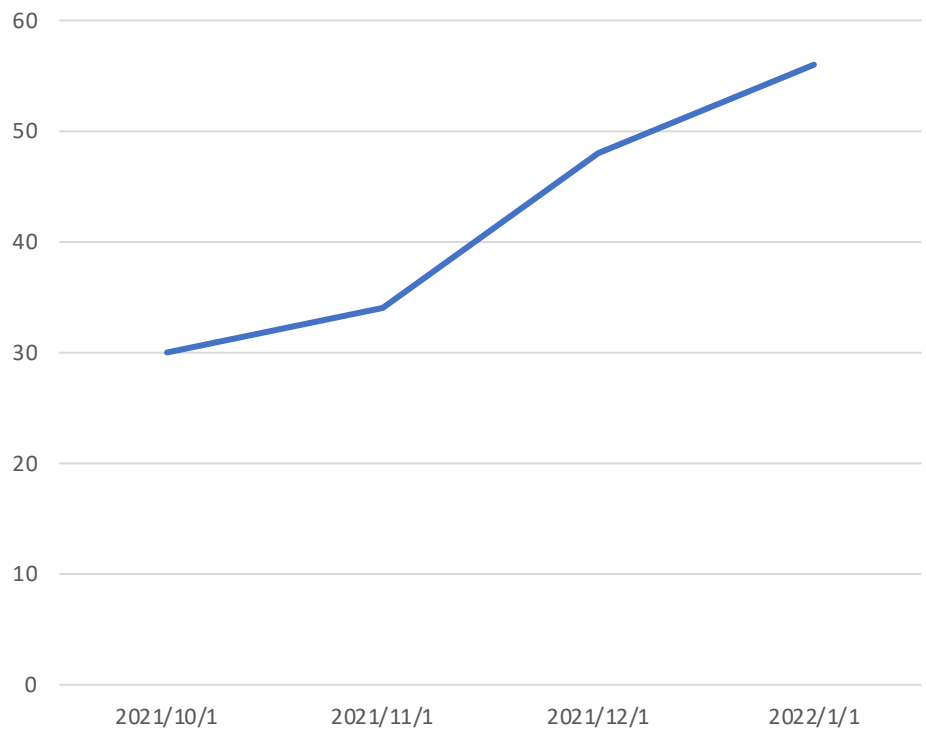
mdxは、データ活用社会創成プラットフォーム共同研究基盤（共同研究基盤）の構成機関で運用されています。

問い合わ

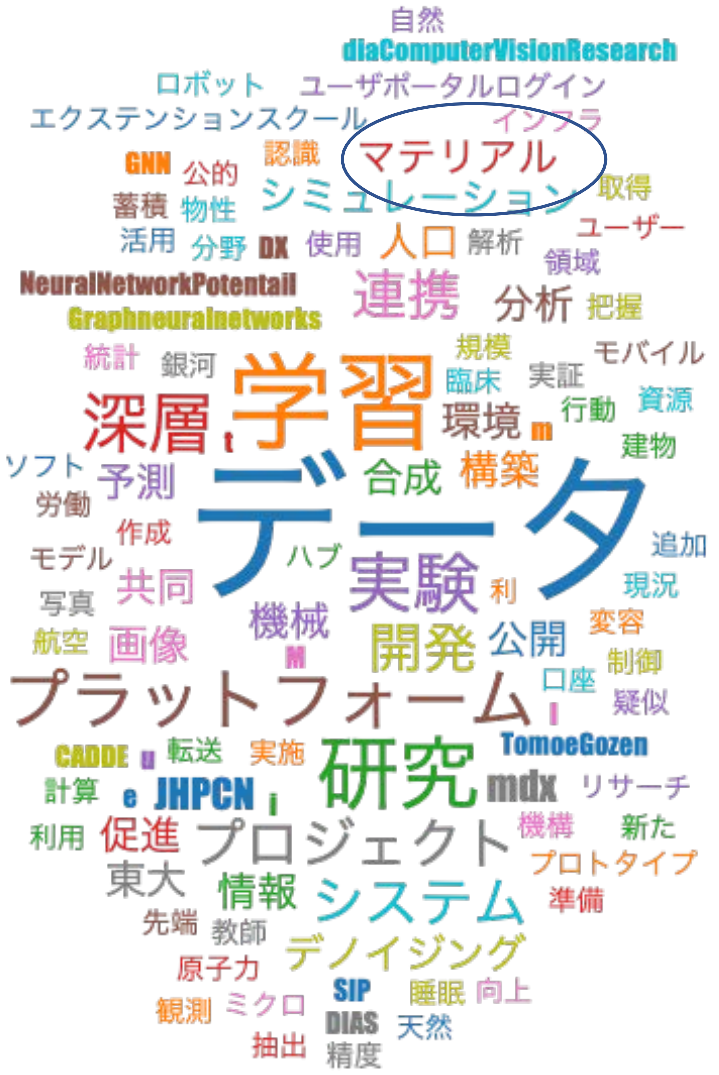
お問合せは、郵便または電子メールにてお願いいたします。

https://mdx.jp/

of Launched Projects on mdx since October 2021



Word Cloud from Project Names



https://mdx.jp/