OECD GSF
LODES Workshop

Open Data Platform in Korea

Global Science experimental
Data hub Center

September 6, 2018
Seo-Young Noh

Contents

1. Open Science: *Connected Science*
2. Data-driven R&D Era
3. Data Infrastructure in Korea
4. Linking Data Repositories: *Practical Implementation*
5. Summary: *Responses to Requests*

Open Science:

Connected Science

Open Science…

- OECD Principles and Guidelines for Access to Research Data from Public Funding (2006-07)
- Initial discussion of Open Science at CSTP in 2011
- Many Open Science related activities on-going (PSI, open gov data, open educational resources, MOOCS…)
- OECD produced the first Open Science report, mainly focusing on Open Access, Open Collaboration and Open Data (2015)
- Several expert groups in GSF have been formed to build advisory policy for Open Science: Research Infrastructure, Data Infrastructure for Open Science

OECD Principles and Guidelines for Access to Research Data from Public Funding

Acknowledgment on the importance of open

Please cite this paper as:

OECD (2015), "Making Open Science a Reality", *OECD Science, Technology and Industry Policy Papers*, No. 25, OECD Publishing, Paris.
http://dx.doi.org/10.1787/5jrs2f963zs1-en

OECD Science, Technology and Industry Policy Papers No. 25

Making Open Science a Reality

OECD

Open science is ***more than open access to publications or data***; it includes many aspects and stages of research processes. […]

Open science is ***a broader concept*** that includes

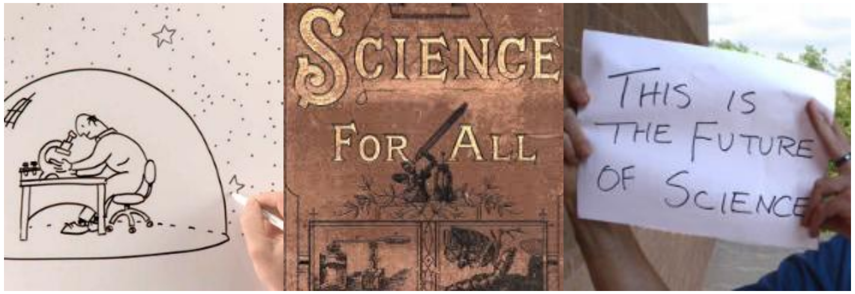
- the interoperability of scientific infrastructure
- open and shared research methodologies (such as open applications and informatics code)
- and machine-friendly tools allowing, for example, text and data mining.

Source: POLICIES TO PROMOTE OPEN SCIENCE: EVIDENCE FROM OECD COUNTRIES, Giulia Ajmone Marsan

Key features of Open Science

- **Main goal of Open Science:**
  - Provides cost-effective access to digital research data from public funding
  - Enhances utilizations of research data to scientific communities as well as societies including corporate sectors

Science cannot exist in a bubble...



should work with public communities...
The Future of Science

- **Benefits**
  - Easy Research → Efficiency, Removing Redundancy → Solving Contemporary Problems
  - Tackling Big Problems → Enabling Big Science → Solving Problems of Humankind
  - New Value Creation → Enabling Convergence → Solving Unknown Problems

- **Ways and Means**
  - Open Data, Open Access and Open Collaboration through Information and Communication Technology
  - "Open Access" and "Open Collaboration" look straightforward, but *"Open Data" is not so simple, requiring deep understanding the features of data*

THE DATA ARE MINE!
IVORY TOWER OF SCIENCE

SCENE FROM THE PAST?

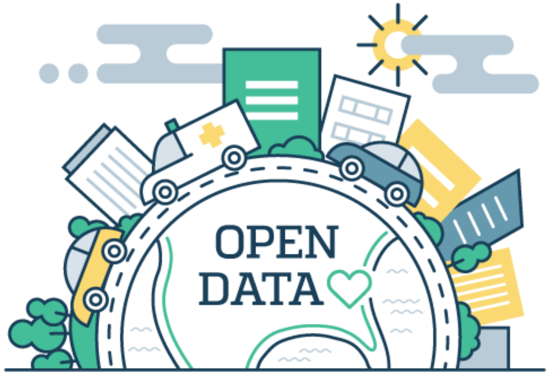
DATA TREE
FREE PICKING

THE FUTURE?

Features of data in Open Science

Open data is the key part of Open Science:

- Transparency in experimental methodology, observation, and collection of data
- Public availability and reusability of scientific data
- Public accessibility and transparency of scientific communication



Data should be valuable

- Reusable data (ex: observation data, knowledge database)
- Data requiring long time for data accumulation (ex: pathology tracking data, climate change tracking data)
- Data requiring big budget for data acquisition (ex: large equipment-based experiments)
- Data requiring huge computing power for data generation (ex: simulation data)

Interoperability – Connected Science

Why open data is matter? … many terminologies, properties

Transparency, Public Availability, Public Access, Reusable, Redistribution, Universal Participation

Interoperability

Interoperability makes it possible for diverse research groups inter-operate scientific data, intermixing different datasets, *leading to open a new way for unveiled values*

Connected Scientific Data

Connected Science

Value Creation

Social Responsibility

Connected Data

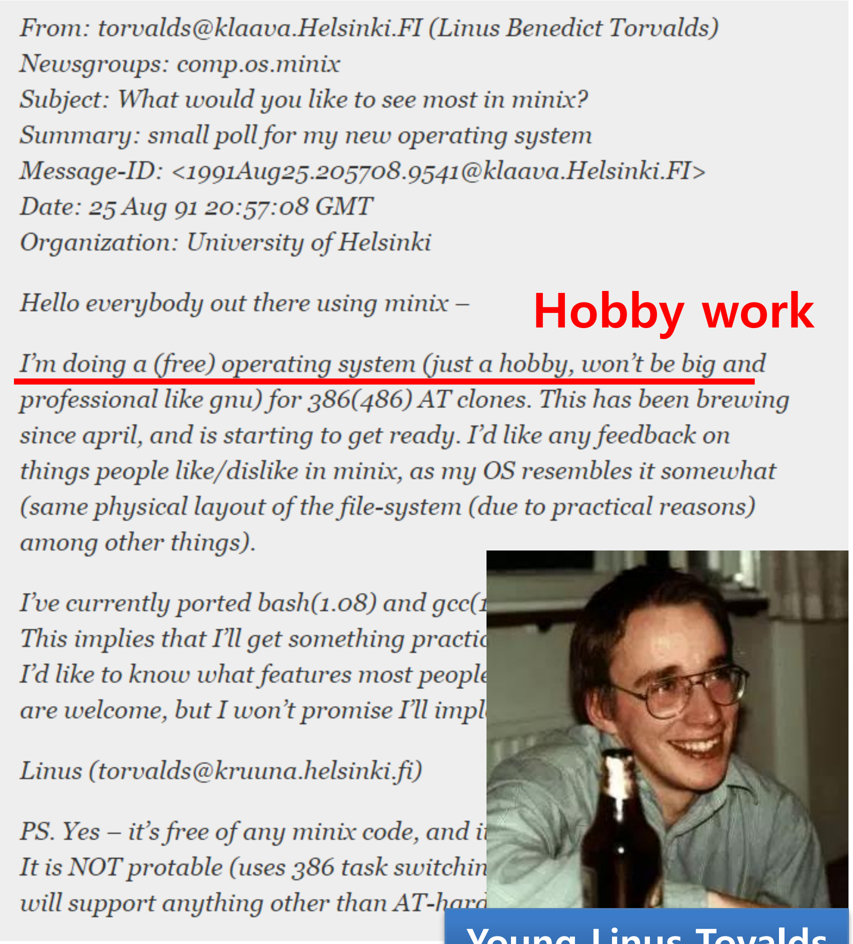
Connected Devices

Value Creation

~~Social Responsibility~~ Business

Open S/W – Good Example

Linus Tovalds opened free operating system (Open S/W), later combined with GNU project, influencing great impact



From: torvalds@klaava.Helsinki.FI (Linus Benedict Torvalds)
Newsgroups: comp.os.minix
Subject: What would you like to see most in minix?
Summary: small poll for my new operating system
Message-ID: <1991Aug25.205708.9541@klaava.Helsinki.FI>
Date: 25 Aug 91 20:57:08 GMT
Organization: University of Helsinki

Hello everybody out there using minix –

I'm doing a (free) operating system (just a hobby, won't be big and professional like gnu) for 386(486) AT clones. This has been brewing since april, and is starting to get ready. I'd like any feedback on things people like/dislike in minix, as my OS resembles it somewhat (same physical layout of the file-system (due to practical reasons) among other things).

I've currently ported bash(1.08) and gcc(1
This implies that I'll get something practic
I'd like to know what features most people
are welcome, but I won't promise I'll impl

Linus (torvalds@kruuna.helsinki.fi)

PS. Yes – it's free of any minix code, and i
It is NOT protable (uses 386 task switchin
will support anything other than AT-hard

Hobby work

Young Linus Tovalds

Sharing is good, and with digital technology, sharing is easy

– Richard Stallman

Richard Stallman

***Software** to be distributed in a manner such that its users receive the **freedoms to use, study, distribute and modify** that software*

GNU Project since 1984

GNU Public License

Data-driven R&D Era

Research Paradigm Shift

Data & Infrastructure are Key in Scientific Discovery

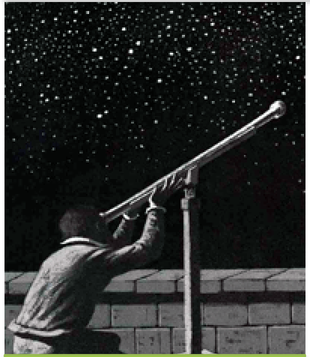
Describing natural phenomena based on **Observation**

Modeling and **Theory**

Computing Simulation

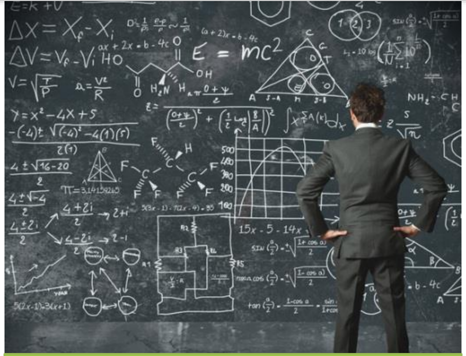
Data Analysis of tremendous **data produced from large experimental facilities**

Research Paradigm Shift to Data Intensive Scientific Discovery



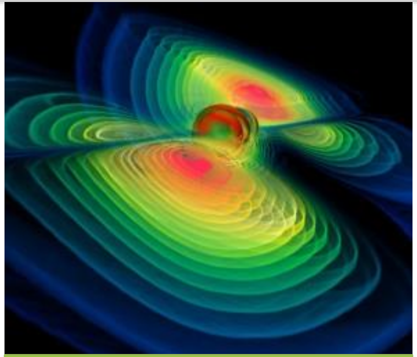
1st Generation: Observation

Galileo's telescope



2nd Generation: Theory

Higgs Theory



3rd Generation: Simulation

Black Hole Simulation

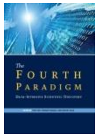
4th Generation: Data

CERN's CMS and ATLAS experiments → Higgs discovery

More chance to do research with advanced equipment, higher chance to get Nobel prize

87% of Nobel prizes have been given to researchers who produced outstanding scientific discoveries using advanced experimental equipment since 1914.

Source:
The Fourth Paradigm

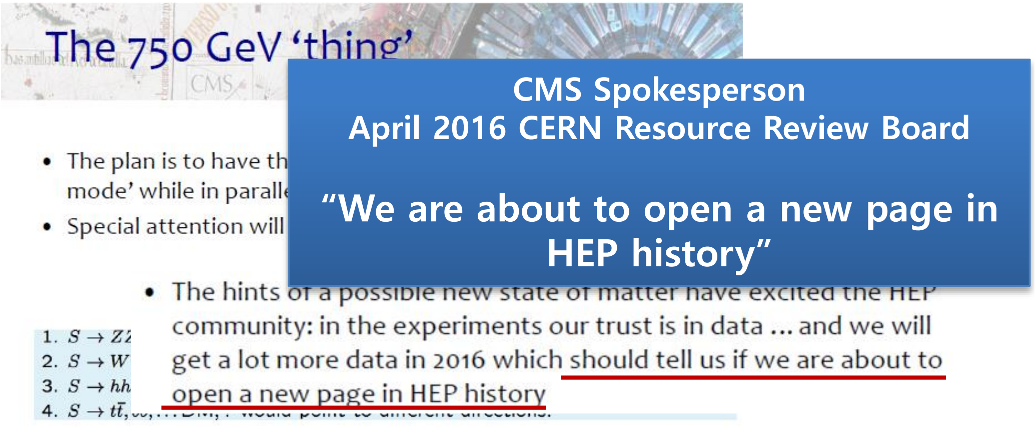


Trust in data: hidden meaning…evidence

- **CERN** noticed a signal like a new particle in CMS & ATLAS experiment in December 2015.
- The **750 GeV diphoton excess** in particle physics was an anomaly in data collected at the Large Hadron Collider(LHC) in 2015, **which could have been an indication of a new particle.**
- However, **the anomaly was absent in data collected in 2016**, suggesting that the diphoton excess was **a statistical fluctuation.**
- In the interval **between the December 2015 and August 2016 results**, the anomaly generated considerable interest in the scientific community, including about **500 theoretical studies**.

We are in data-driven science era!!!

Our trust is in data



The 750 GeV 'thing'

- The plan is to have th
mode' while in paralle
- Special attention will

1. $S \to ZZ$
2. $S \to W$
3. $S \to hh$
4. $S \to t\bar{t}$, …DM, ? would point to different directions.

- The hints of a possible new state of matter have excited the HEP community: in the experiments our trust is in data … and we will get a lot more data in 2016 which should tell us if we are about to open a new page in HEP history

CMS Spokesperson
April 2016 CERN Resource Review Board

"We are about to open a new page in HEP history"

PRL 116, 150001 (2016) PHYSICAL REVIEW LETTERS week ending 15 APRIL 2016

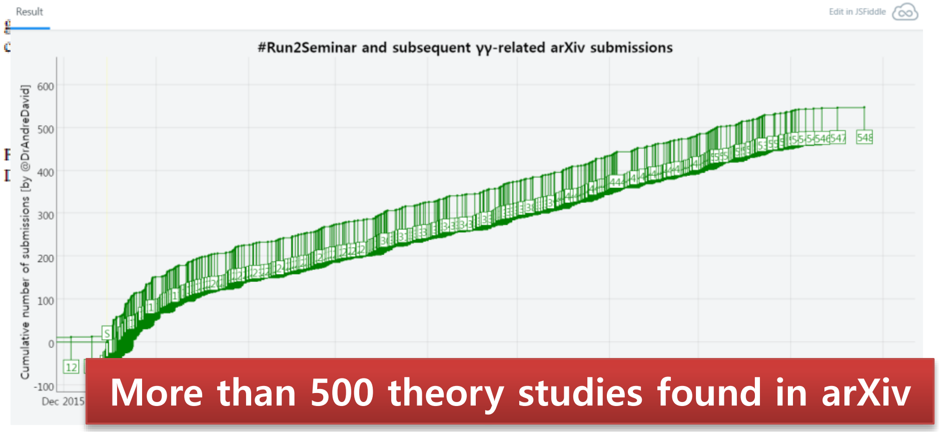
Editorial: Theorists React to the CERN 750 GeV Diphoton Data

Last December, the ATLAS and CMS Collaborations at the Large Hadron Collider reported
pr
[1
have recently reanalyzed their data [3,4], and the signal has become slightly stronger. Though the results are extremely intriguing, more data are required to establish if the excess is real, or a statistical fluctuation.

Over 250 theory papers have appeared following the December announcement, and a number of them were submitted to us. We found it appropriate to publish a small sample of them. To maximize the coherence and fairness of our choices, we obtained informal advice from several experts.

Four such Letters appear in this issue [5–8]. Others may follow, but we think that this set

More than 250 theory papers submitted to PRL

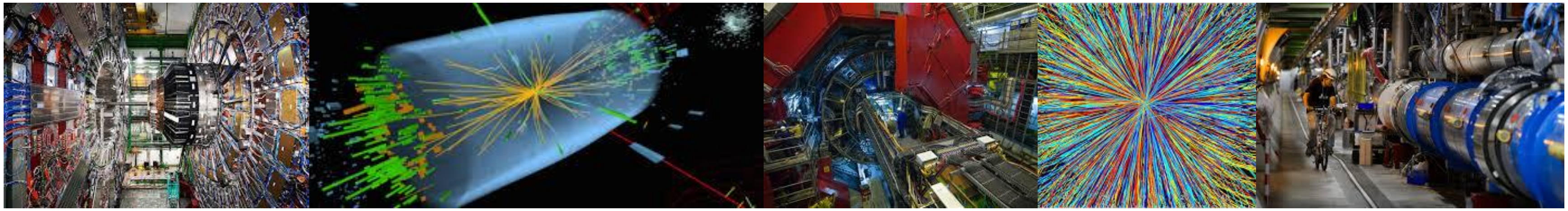


More than 500 theory studies found in arXiv

Data Infrastructure…that is what we need

Science relies on data, requiring infrastructure for data.

Data is getting more important and growing fast.



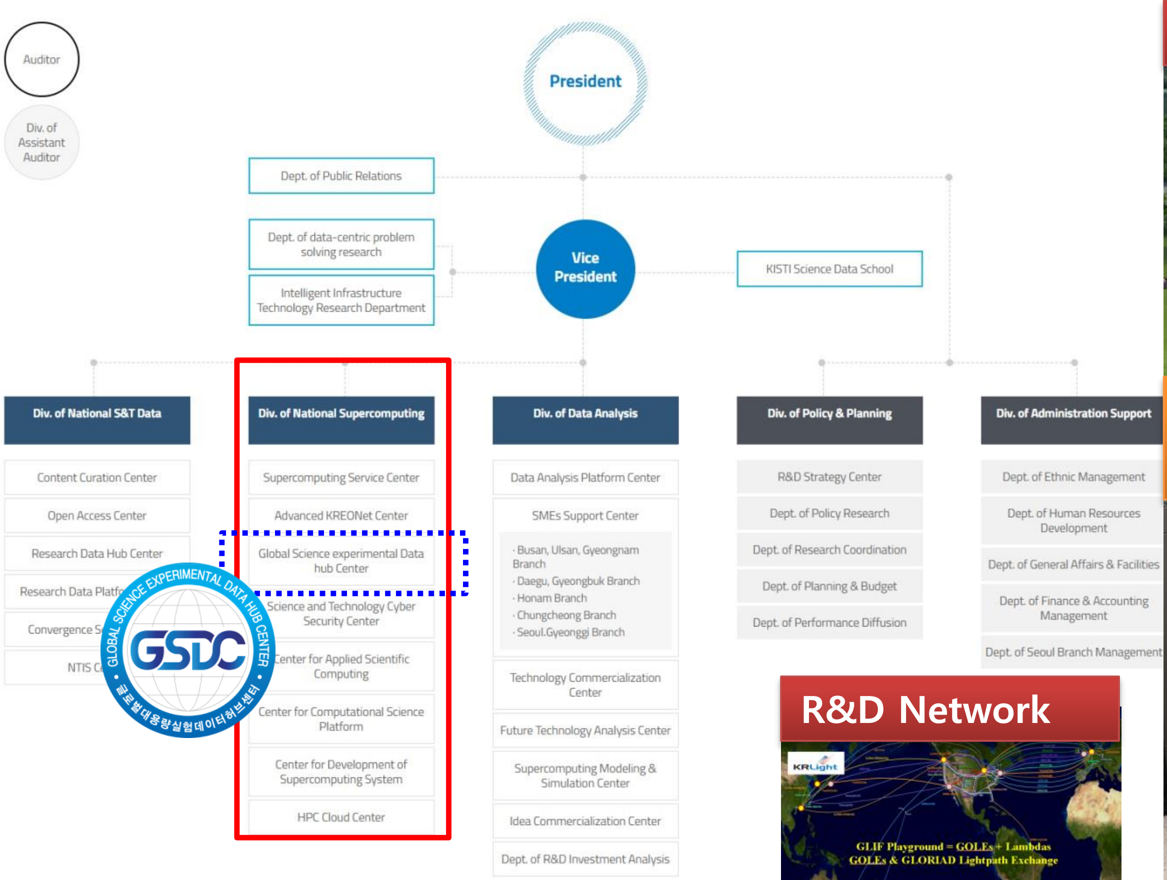
Data Infrastructure is the one of key factors for successful science and tackling big problems of humankind.

KISTI has been in preparation for big data research era. Our mission is gradually expanding to national role for data intensive research.

Data Infrastructure:

KISTI, Korea

KISTI…providing powerful ICT infra. service



Auditor

Div. of Assistant Auditor

President

Dept. of Public Relations

Dept. of data-centric problem solving research

Intelligent Infrastructure Technology Research Department

Vice President

KISTI Science Data School

Div. of National S&T Data
- Content Curation Center
- Open Access Center
- Research Data Hub Center
- Research Data Platfo
- Convergence S
- NTIS C

Div. of National Supercomputing
- Supercomputing Service Center
- Advanced KREONet Center
- Global Science experimental Data hub Center
- Science and Technology Cyber Security Center
- Center for Applied Scientific Computing
- Center for Computational Science Platform
- Center for Development of Supercomputing System
- HPC Cloud Center

Div. of Data Analysis
- Data Analysis Platform Center
- SMEs Support Center
- · Busan, Ulsan, Gyeongnam Branch
- · Daegu, Gyeongbuk Branch
- · Honam Branch
- · Chungcheong Branch
- · Seoul.Gyeonggi Branch
- Technology Commercialization Center
- Future Technology Analysis Center
- Supercomputing Modeling & Simulation Center
- Idea Commercialization Center
- Dept. of R&D Investment Analysis

Div. of Policy & Planning
- R&D Strategy Center
- Dept. of Policy Research
- Dept. of Research Coordination
- Dept. of Planning & Budget
- Dept. of Performance Diffusion

Div. of Administration Support
- Dept. of Ethnic Management
- Dept. of Human Resources Development
- Dept. of General Affairs & Facilities
- Dept. of Finance & Accounting Management
- Dept. of Seoul Branch Management

GLOBAL SCIENCE EXPERIMENTAL DATA HUB CENTER
GSDC
글로벌대용량실험데이터허브센터

Unique organization in Korea, our mission is to support various R&D activities through ICT infra. service

New Supercomputing Building

25.7PFlops and 20PB ranked 11th in Top 500

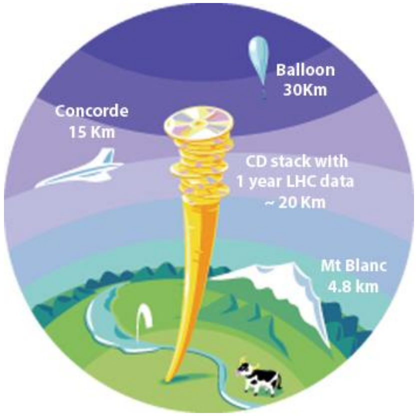
Supercomputer

R&D Network

KRLight
GLIF Playground = GOLEs + Lambdas
GOLEs & GLORIAD Lightpath Exchange
GOLE
GLORIAD
APAN
100
KREONet2 Members: ~300
Backbone: 14 GigaPOPs(20G~5G)
Over 1G access links: 50

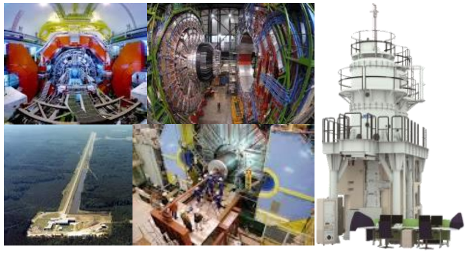
Fast and secure network, providing domestic researchers with a constraint free collaborative research environment through KREONET (locally) and GLORIAD (globally)

Global Science experimental Data hub Center



Large-scale Scientific Data:
20Km CD stack with data produced per year in CERN

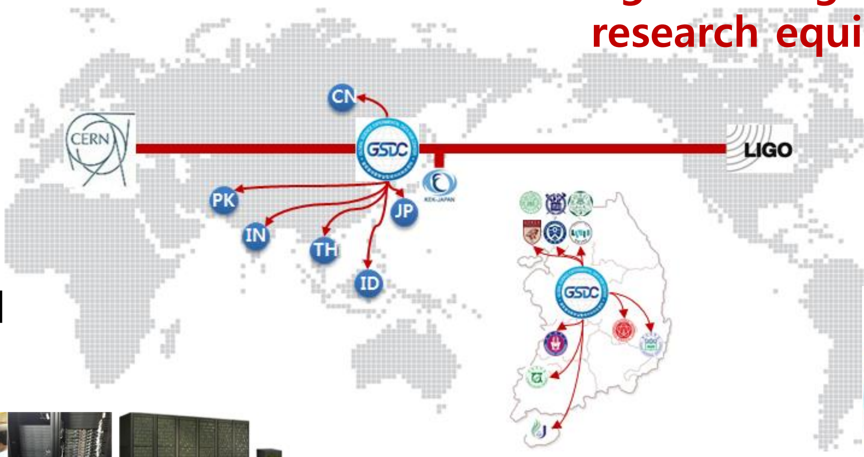
Collaboration with global laboratories



Data from **large and high-valued research equipment**

- (Global)
  Asia representative **Data Hub**

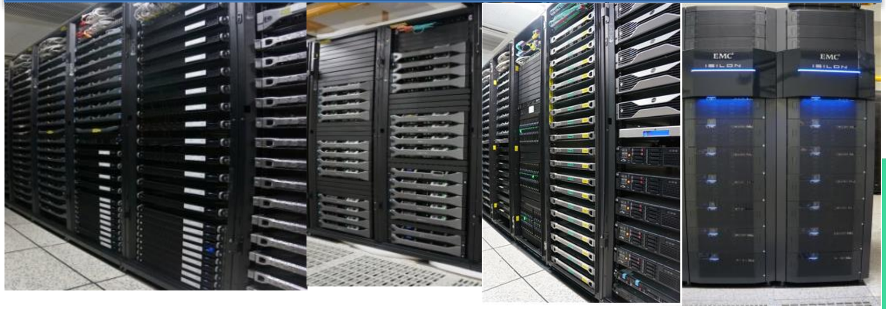
- (Domestic)
  Scientific data management and analysis platform service



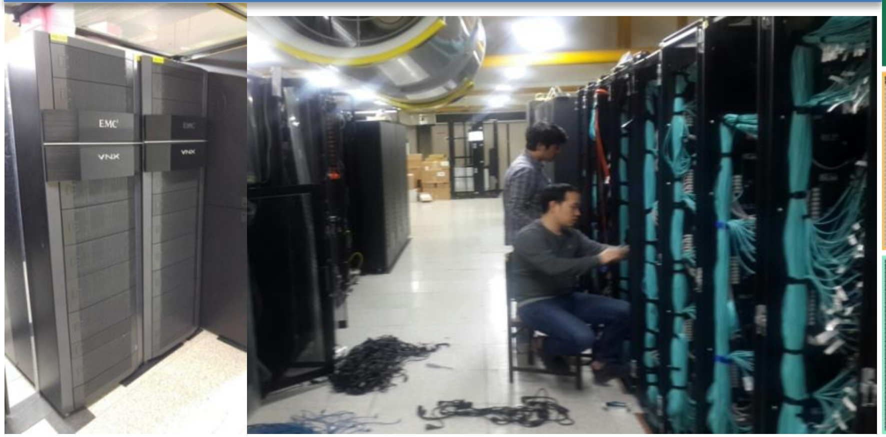
Data Repository Infrastructure

Growing every year by ~1,200 cores and ~2PB

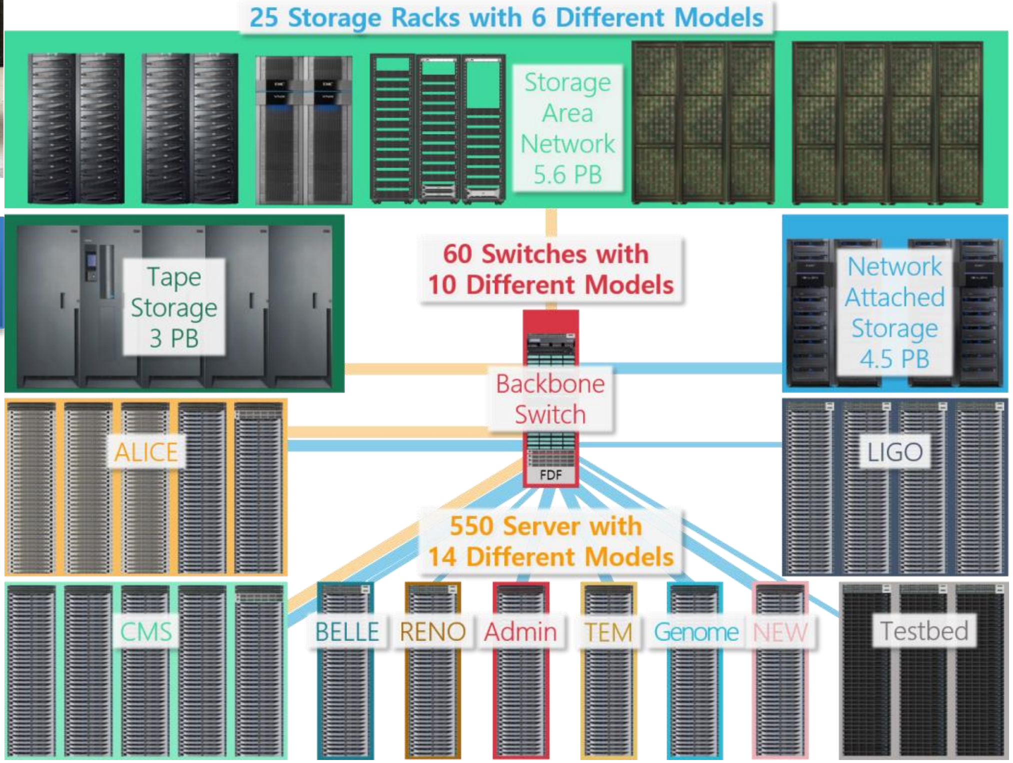
Best equipment procured every year



Interlocking with existing systems
done by 100% KISTI experts



Centralized data repository model, but **distributed model** in near future


25 Storage Racks with 6 Different Models
Storage Area Network 5.6 PB
Tape Storage 3 PB
60 Switches with 10 Different Models
Network Attached Storage 4.5 PB
Backbone Switch
FDF
ALICE
LIGO
550 Server with 14 Different Models
CMS
BELLE
RENO
Admin
TEM
Genome
NEW
Testbed

Data Repository Services (1)

G Global D Domestic

High Energy Physics

1. ALICE(A Large Ion Collider Experiment)

- ➲ To generate similar conditions that have existed a fraction of the second after the Big Bang
- ➲ 1,655 scientists from 159 institutes, 41 countries

2. CMS(Compact Muon Solenoid)

- ➲ To investigate a wide range of physics, including the search for the Higgs boson, new physics
- ➲ 3,000 scientists from 172 institutes, 40 countries

3. Belle/BelleII(Japan KEK)

- ➲ To investigate CP-violation effects and new physics
- ➲ 428 scientists from 67 institutes, 20 countries

Astro Physics

4. LIGO(Laser Interferometer Gravitational Wave Observatory)

- ➲ To detect cosmic gravitational waves and to develop gravitational-wave observations
- ➲ 1,000 scientists from 60 institutes, 16 countries

Particle Physics

5. RENO(Reactor Experiment for Neutrino Oscillation)

- ➲ To measure a limit on the neutrino mixing matrix parameter(Yeonggwang Nuclear Power Plant)

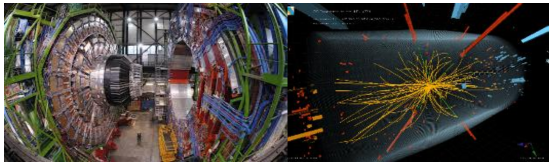
Medical Science

6. Genome Research

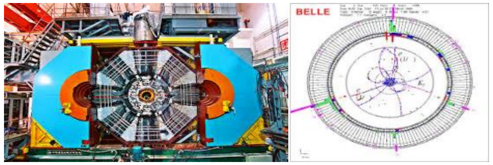
- ➲ Genome data analysis for the next generation of personalized treatment: 50 Korean researchers

1 G D

2 G D

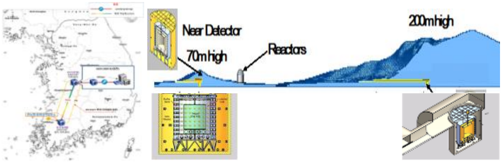


3 G D

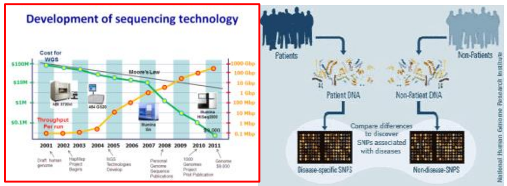


4 D

5 D



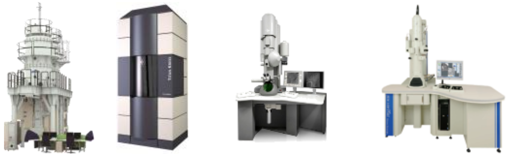
6 D



Data Repository Services (2)

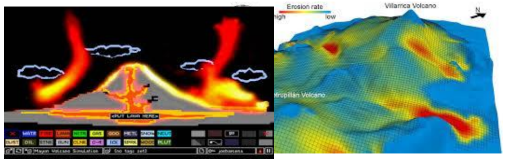
Biology

- **Structural Biology**
  - Data management and analysis service for Transmission Electron Microscope-based research



Safety & Disaster

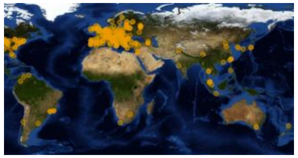
- **Volcanic Disaster Prevention**
  - Data repository service for volcano simulation
  - Helping to build a policy for volcanic disaster response scenarios



High Energy Physics

- **CMS Tier-2 Service**
  - Unified WLCG service in Korea
  - A new Tier-2 center targeting 2018 service

WLCG
Worldwide LHC Computing Grid



Astro Physics

- **KAGRA(Kamioka Gravitational Wave Detector, Japan, 2017)**
  - A project of the gravitational wave study group
  - Data transmission, repository and data analysis environment in conjunction with LIGO data analysis environment

General Purpose Accelerator

- **PAL(Pohang Accelerator Laboratory, Pilot Project, 2018)**
  - Data repository and analysis service for 4th generation accelerator
  - **Trial service for data convergence produced from heterogeneous large equipment for structural biology (Government Supported Funding)**

Medical Science

- **Brain Research (Under discussion)**
  - Platform service for nationwide management, repository, transmission, and analysis of brain data
  - Extending knowledge of HEP data analysis platform service to brain data

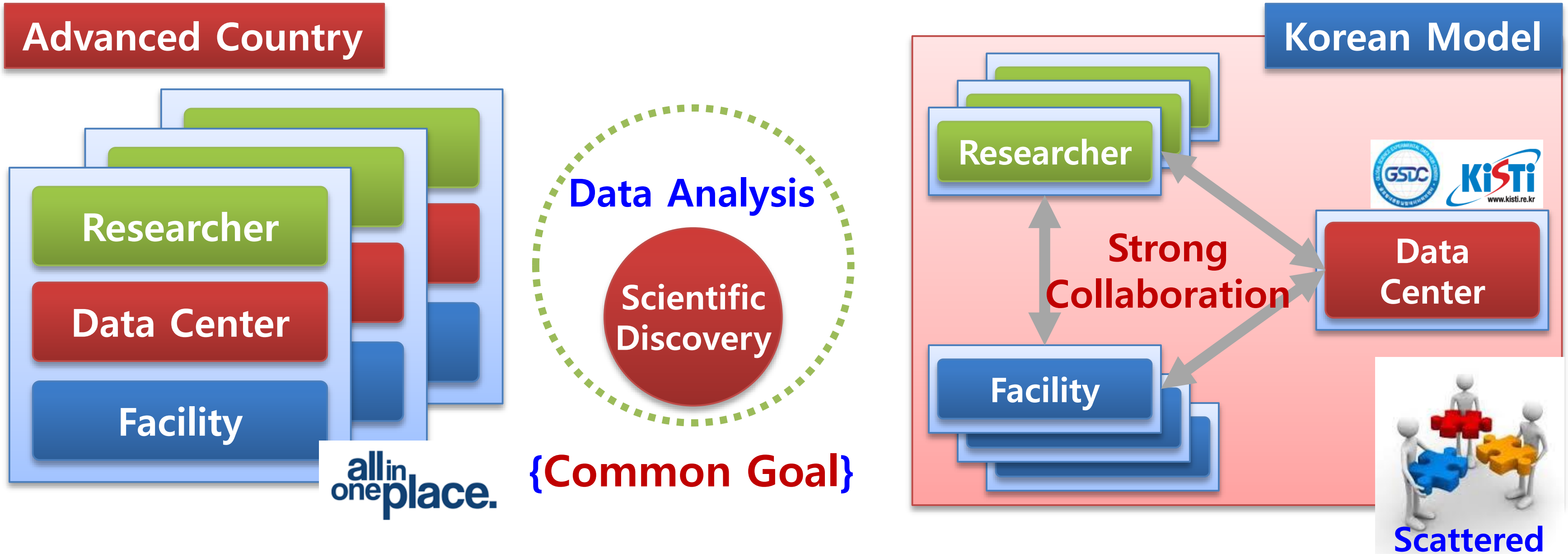
High Energy Physics

- **RISP(Rare Isotope Science Project, RAON, under discussion)**
  - Exploration of the origins of chemical elements, structural study of new isotopes and applied medical research.
  - Discussion with Government for data center role of RAON

Korean Model for Data Intensive Research

Focusing on a centralized model for data repository at the beginning

by fully utilizing ICT specialized institute like KISTI/GSDC



Advanced Country

Researcher

Data Center

Facility

allinoneplace.

Data Analysis

Scientific Discovery

{Common Goal}

Korean Model

Researcher

Strong Collaboration

Data Center

Facility

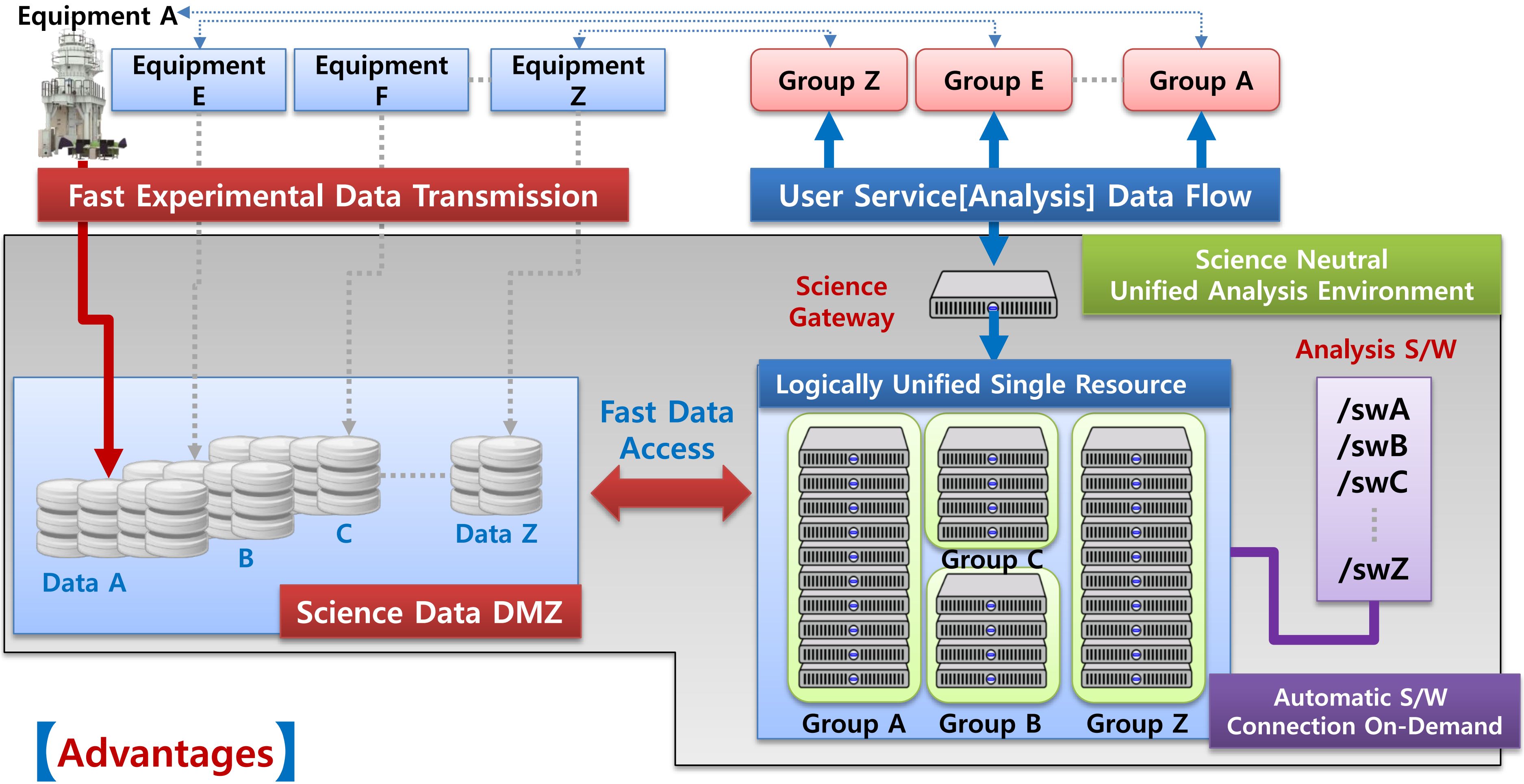
Scattered

- Large scale research group
- Large scale research facility
- Dedicated data center

VS.

- Small scale research group
- Small or medium scale facility
- Not easy to have a dedicated data center (in size and experts)

Unified Data Analysis Environment (Centralized Model)



【Advantages】

1. Pluggable Science → Supports in unified way for various groups and equipment
2. Data Infra. Sharing → Reuse and full utilization of infra. saving tax-payer's money
3. Simple R&D Process → Fast results from data acquisition to data analysis

Linking Data Repositories: *Practical Implementation*

R&D Open Data Task Force: Open Data Platform

A Task Force has been setup in Ministry of Science and ICT, focusing on …

1. Fostering data-driven R&D communities
2. National level regulation for data management
3. Scientific program development for open data

연구데이터 공유·활용체계
구축전략 공청회
2017. 12. 18(월) 14:00~16:00
대한상공회의소 국제회의장(지하 2층)
주 최 | 과학기술정보통신부
주 관 | KISTI 한국과학기술정보연구원

Ministry of Science and ICT

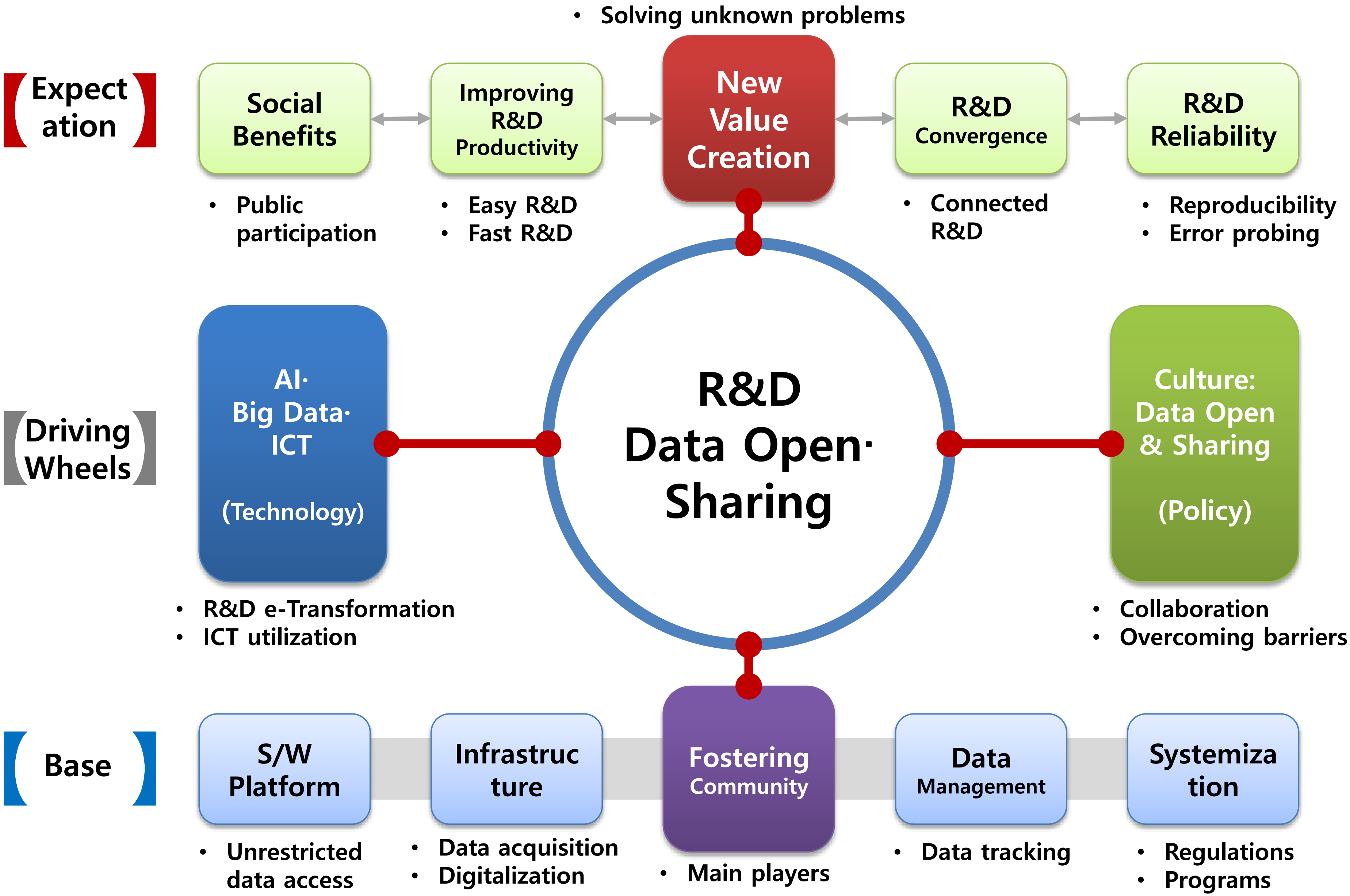
Public Hearing - Dec. 18, 2017

Main activities …

1. Analyzing best practices including European Open Science Cloud in EU, Big Data Hub Program in U.S
2. Defining priority and categorization of data intensive research fields and new program development
3. Developing R&D Open Data Platform, helping seamless data sharing, data accessing, data analysis, data linking across disciplines

Accelerating R&D productivities
through ICT-based R&D e-Transformation

Big Picture: what we expect …



【Expectation】

Social Benefits
- Public participation

Improving R&D Productivity
- Easy R&D
- Fast R&D

New Value Creation
- Solving unknown problems

R&D Convergence
- Connected R&D

R&D Reliability
- Reproducibility
- Error probing

【Driving Wheels】

AI· Big Data· ICT

(Technology)

- R&D e-Transformation
- ICT utilization

R&D Data Open· Sharing

Culture: Data Open & Sharing

(Policy)

- Collaboration
- Overcoming barriers

【Base】

S/W Platform
- Unrestricted data access

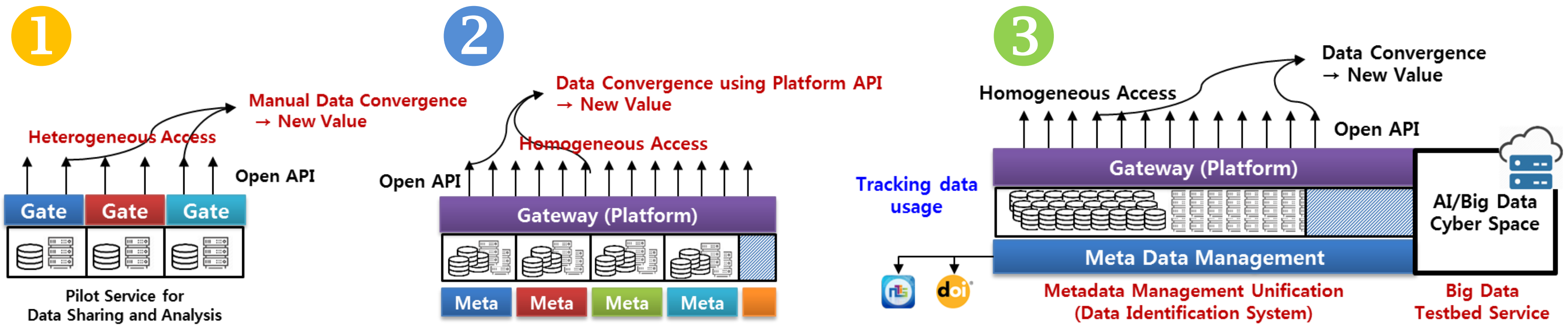
Infrastructure
- Data acquisition
- Digitalization

Fostering Community
- Main players

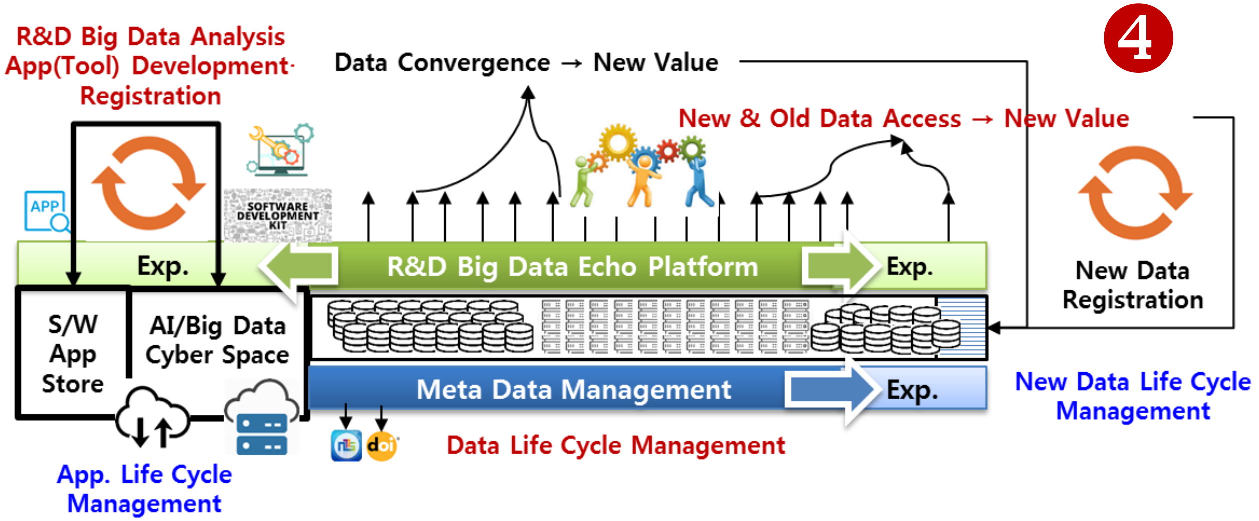
Data Management
- Data tracking

Systemization
- Regulations
- Programs

Stepwise Implementation: from part to all



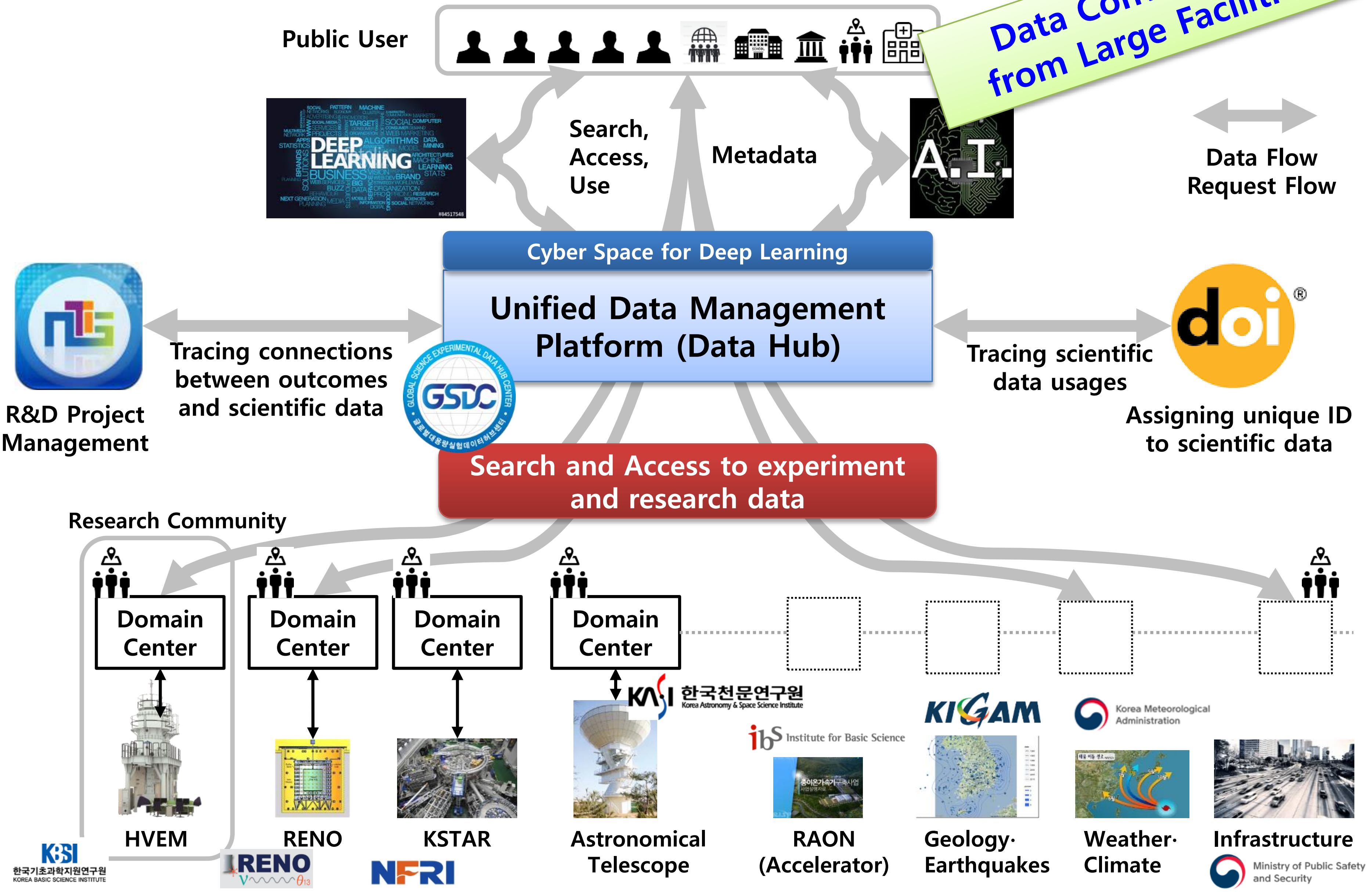
- Covering entire data management, linking data repositories, providing easy and seamless access across various R&D
- Enabling Open Science Service by stepwise implementation



4 Pilot Projects (launched in 2018)

1. Genome Data
2. Material Data
3. Data Convergence from Large Facilities (KISTI-GSDC)
4. AI/Big Data

Connected Scientific Data – External View



Data Convergence from Large Facilities

Public User

Search, Access, Use

Metadata

Data Flow
Request Flow

Cyber Space for Deep Learning

Unified Data Management Platform (Data Hub)

R&D Project Management

Tracing connections between outcomes and scientific data

Tracing scientific data usages

Assigning unique ID to scientific data

Search and Access to experiment and research data

Research Community

Domain Center

Domain Center

Domain Center

Domain Center

HVEM

RENO

KSTAR

Astronomical Telescope

RAON (Accelerator)

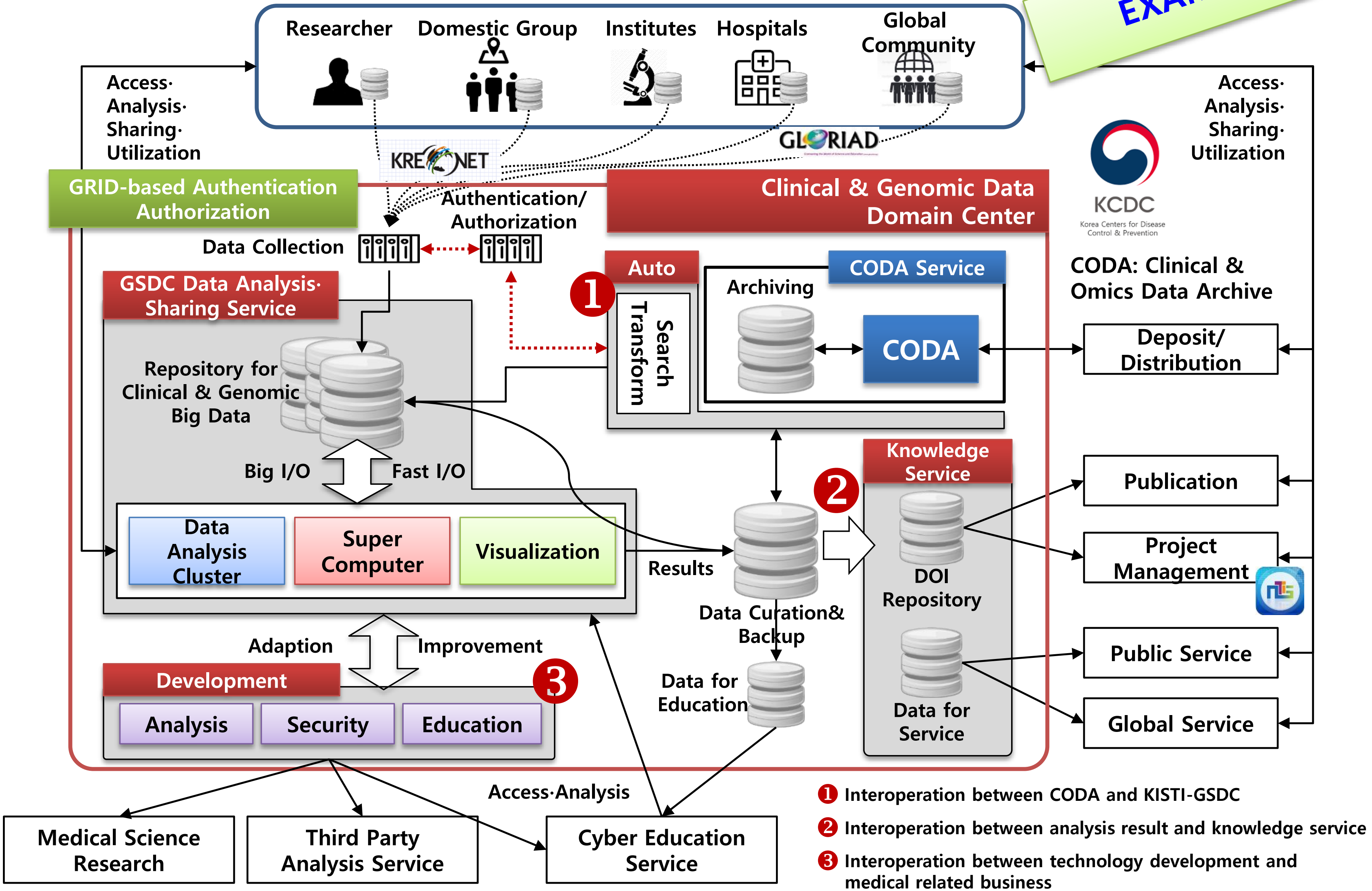
Geology· Earthquakes

Weather· Climate

Infrastructure

DEEP LEARNING
A.I.
GSDC
GLOBAL SCIENCE EXPERIMENTAL DATA HUB CENTER
doi
KBSI
한국기초과학지원연구원
KOREA BASIC SCIENCE INSTITUTE
RENO
NFRI
KASI 한국천문연구원
Korea Astronomy & Space Science Institute
ibS Institute for Basic Science
KIGAM
Korea Meteorological Administration
Ministry of Public Safety and Security

Connected Scientific Data – Internal View

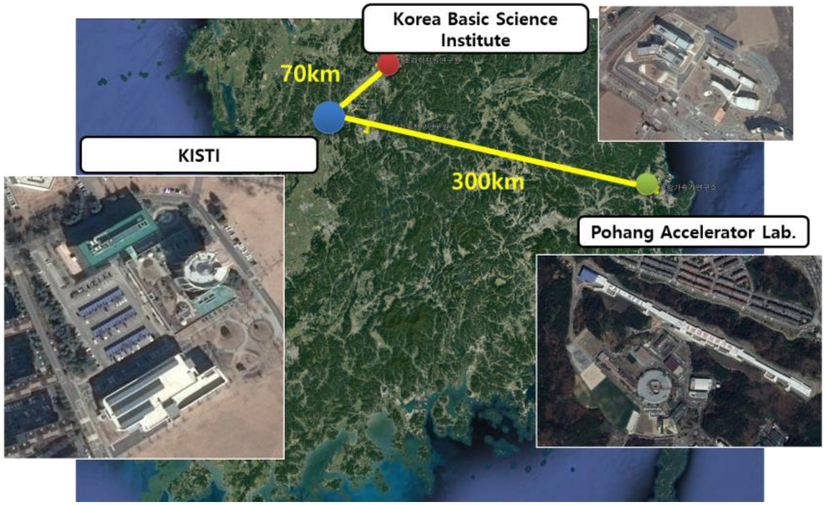


❶ Interoperation between CODA and KISTI-GSDC

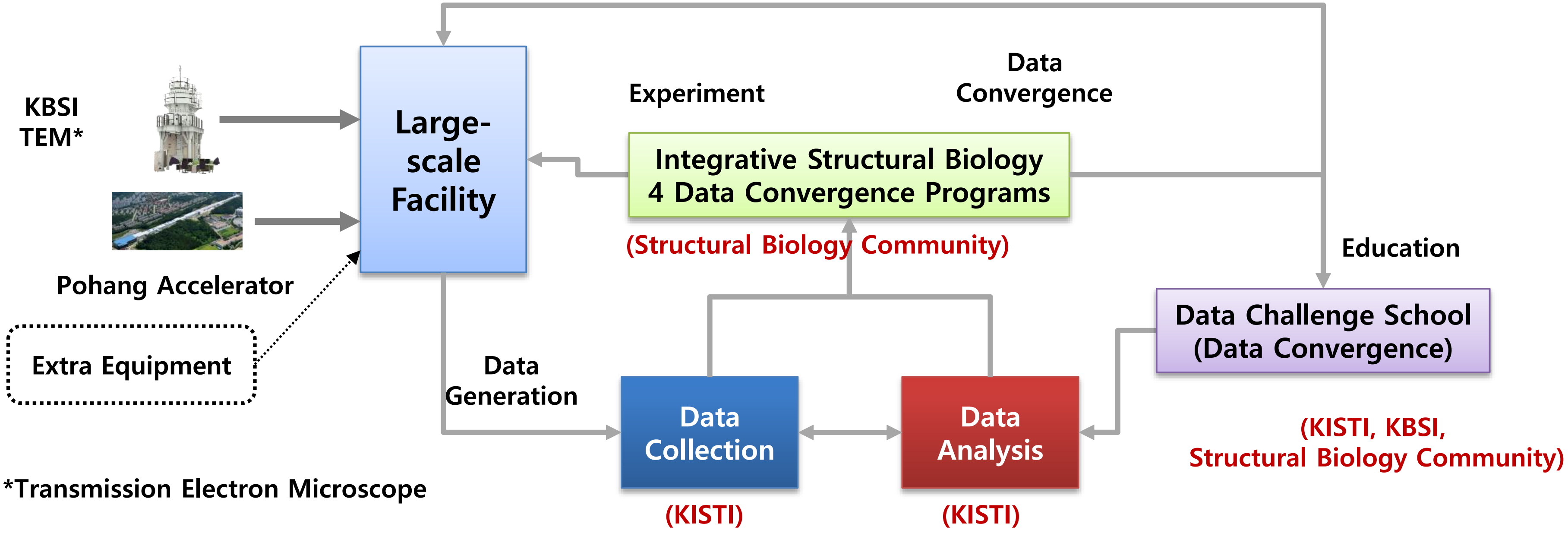
❷ Interoperation between analysis result and knowledge service

❸ Interoperation between technology development and medical related business

Pilot Project – Data Convergence (KISTI-GSDC)



- Centralized data repository is used for data collection and access
- Central data repository will be connected with Open Data Platform, which is also being developed by KISTI.
- Expect each site has own repositories interoperable with the Open Data Platform in future



KBSI TEM*

Pohang Accelerator

Extra Equipment

Large-scale Facility

Experiment

Data Convergence

Integrative Structural Biology
4 Data Convergence Programs

(Structural Biology Community)

Education

Data Challenge School
(Data Convergence)

(KISTI, KBSI,
Structural Biology Community)

Data Generation

Data Collection

(KISTI)

Data Analysis

(KISTI)

*Transmission Electron Microscope

Summary:

Response to Requests

Summary

Open Science

- Highlighted openness of access, collaborations and data
- Believed to give benefits to scientific community:
  Connected Scientific Data → Connected Science

Data-driven R&D

- Research paradigm is being shifted to data-driven scientific discovery
- Data and infrastructure are the key in scientific discovery

Open Data Platform

- Importance of Open Data is recognized in government level and the TF has launched four pilot projects this year (2018)
- Data repositories, setup by pilot projects, will be connected to the platform which makes them interoperable as a final goal.

Implementation of Open Data Platform is not an easy task…
The pilot projects are not big scale…

But, such a trial is a Big Step moving toward to making Open Science a reality in Korea

Responses to Requests

Identify your initiative as a stakeholder …

- Policymakers(Government), researchers, ICT specialists are stakeholders.
- Policymakers wants to make data, generated from public funding, traceable through Open Data Platform.
- Researchers wants to use well managed ICT infrastructure for data analysis and data sharing, in order to accelerate R&D process.
- Role of ICT specialized institute like KISTI, expected from government, is well aligned with its mission – promoting Science using ICT technologies.

Analyze how your initiative addresses a multiplicity of different cultural contexts …

- Respect differences in R&D domains and encourage scientific communities to make own standard data management plan agreed in community members.
- Structural biology community, for example, is developing a guideline including data naming convention, sharing, accessing and management policy.

Assess the proactive role of your initiative with respect to interoperability …

- Rather than making data repositories interoperable directly(1-to-1), indirect interoperability is being considered through the Open Data Platform.
- Such an approach helps to reduce the burden of standization for all R&D. Interesting data can be accessed through OpenAPI provided by the platform. Data convergence and all R&D activities with data are left to scientist.

Thank you.