Research Data Infrastructure of CAS :Practices and Challenges

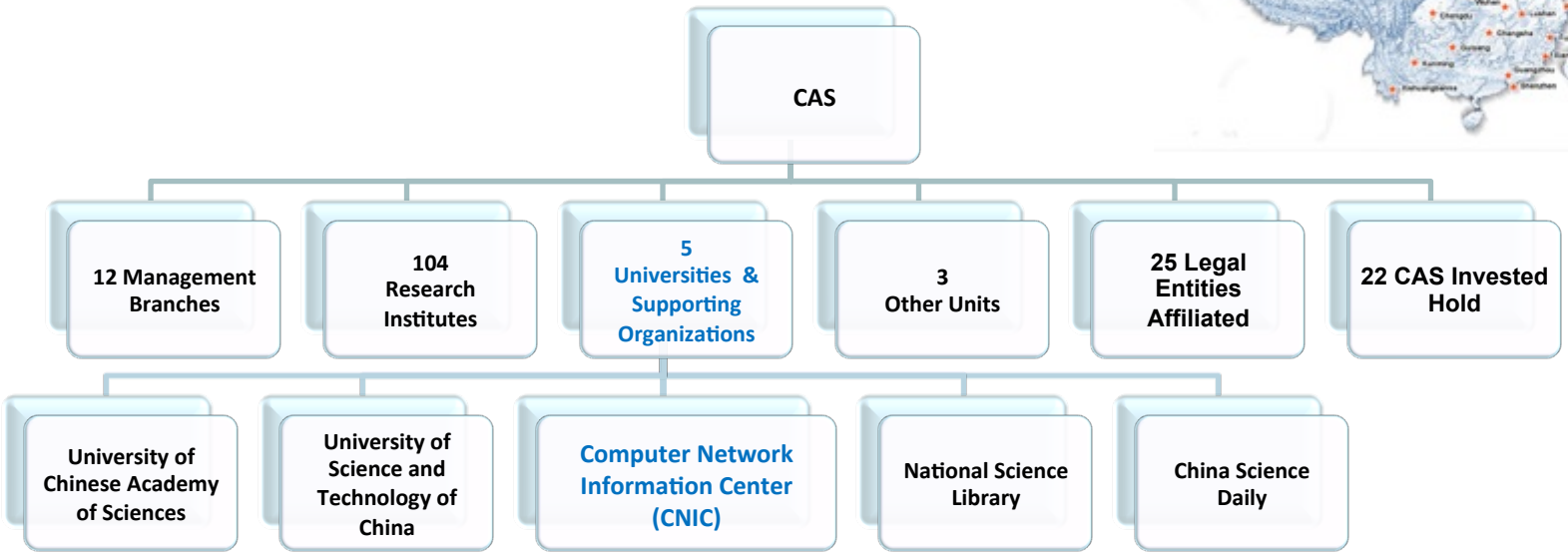
Dr. Jianhui LI(lijh@cnic.cn)
Computer Network Information Center
Chinese Academy of Sciences

中国科学院
CHINESE ACADEMY OF SCIENCES

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

CNIC in CAS

Founded in November 1949, **Chinese Academy of Sciences (CAS)** is the primary academic institution in China in the natural sciences. It is also China's largest comprehensive R&D organization in the natural sciences and high technology as well as the country's foremost science and technology advisory body.



Founded in April 1995, **Computer Network Information Center (CNIC)** is supporting institute involved in constructing and operating IT infrastructure and providing IT-related services. In addition, it serves as an R&D and demonstration base for IT technology applications.

中国科学院
CHINESE ACADEMY OF SCIENCES

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

CNIC Located in zhongguancun, Beijing

Zhongguancun, Haidian District， Beijing

Data Center located in
HuaiRou District, Beijing

中国科学院
CHINESE ACADEMY OF SCIENCES

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

CNIC's History and Mission

First router developed in China

The State Planning Commission approved the initiation of Scientific Database Project.

China's first .CN Domain Name Server

NCFC achieved full-functional connection to the Internet.

.CN Domain Name Server was established and began to provide registration service.

China's earliest HPC service

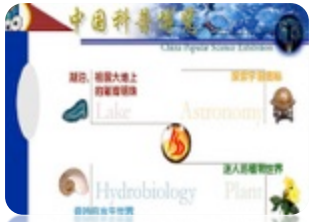
The State Science and Technology Commission approved the founding of CNIC.

CNIC started the earliest scientific computing application and service in China.

Earliest CAS Scientific Database

CNIC initiated Virtual Science Museums of China.

CNIC constructed the early version of the CAS website system.



Earliest CAS Virtual Science Museums

CNIC provided operation support to Academia Resource Planning system of the CAS.

The State Development and Reform Commission approved the initiation of China Internet of Things Name Service Platform.

1986 1994.4 1994.5 1995.3 1996 1999 2001 2002 2008 2013 2015

Internet

地理空间数据云
Geospatial Data Cloud

CHINA SCIENTIFIC DATA

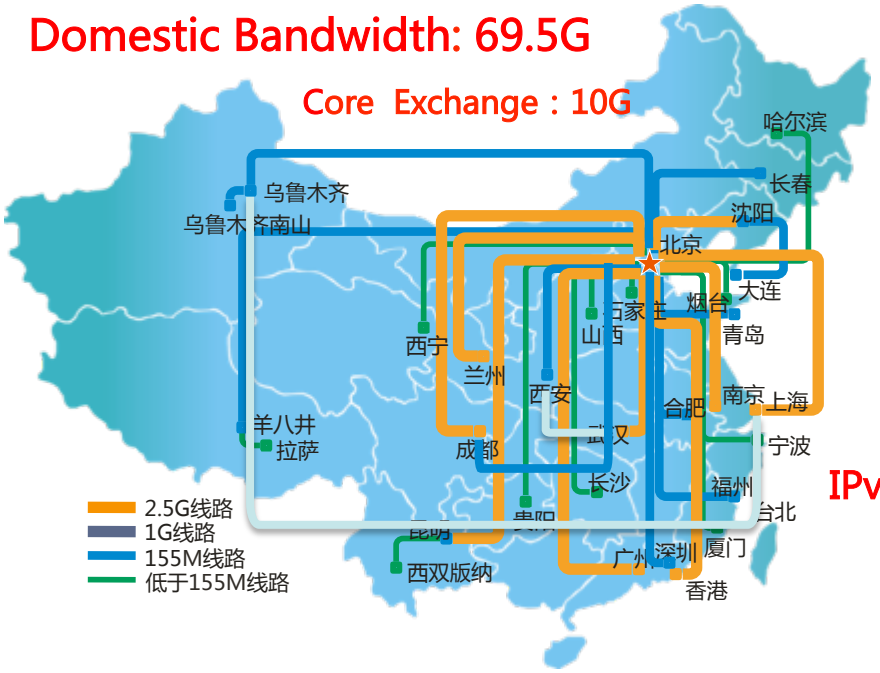
中国科学院
CHINESE ACADEMY OF SCIENCES

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

CSTNet

Domestic Bandwidth: 69.5G

Core Exchange : 10G



IPv4/IPv6

International Bandwidth : 27.5G

USA : 10G
EU : 10G

OrientPlus: 中国与欧洲科研网络连接的桥梁

（GLORIAD）

中国电信 CHINA TELECOM	中国联通 China Unicom	中国移动通信 CHINA MOBILE	中国教育网 www.edu.cn	BJNAP		Deutsche Telekom
4G	5G	2.5G	32G	10G	10G	10G

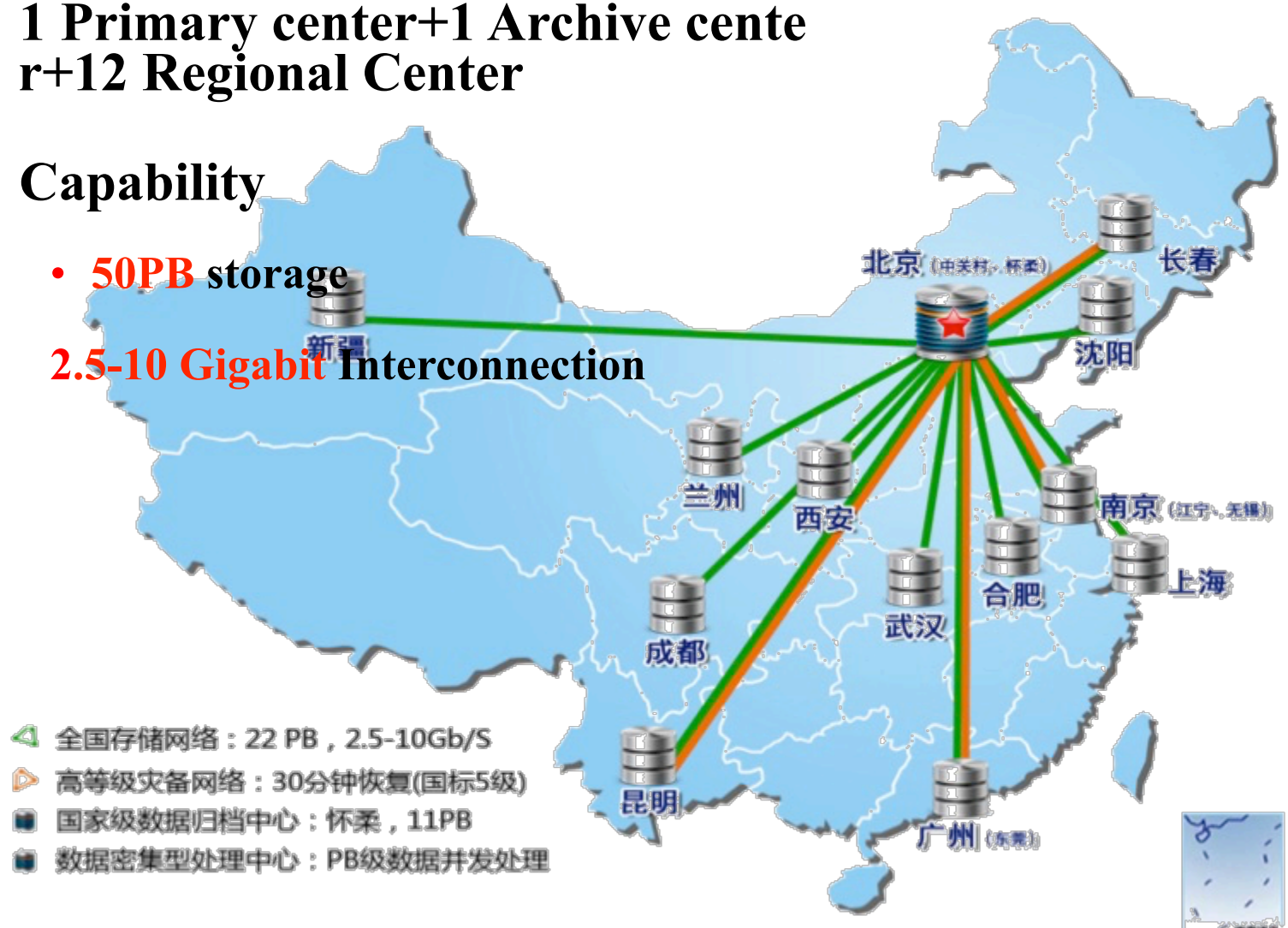
- 30 provinces
- 370 + Organizations
- Above 1M Users

Massive Storage System

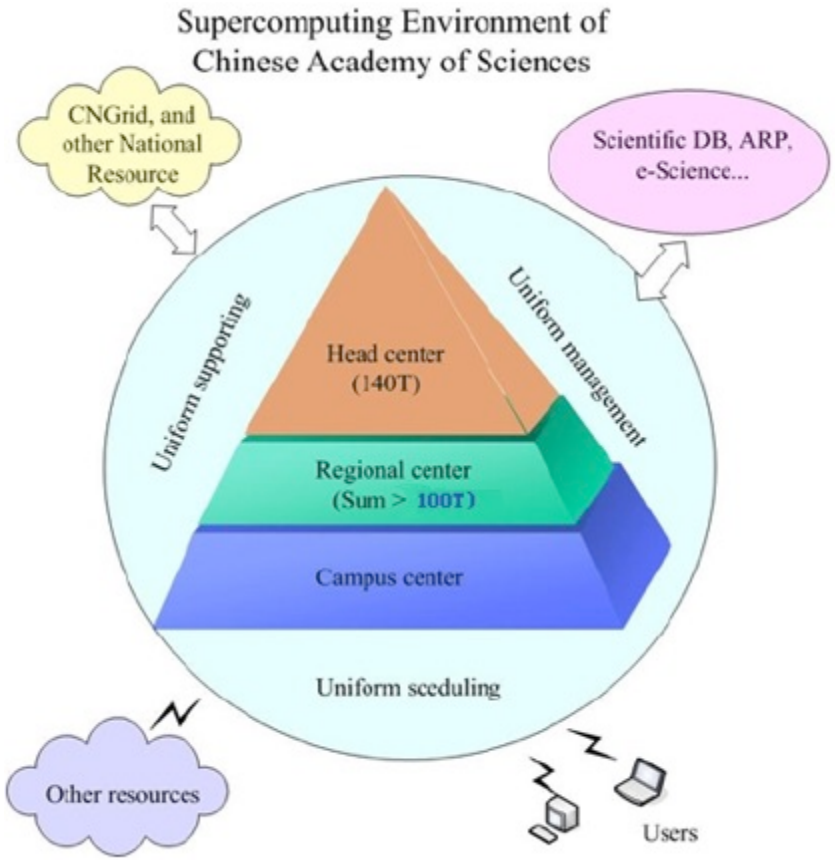
– 1 Primary center+1 Archive center+12 Regional Center

– Capability

- 50PB storage
- 2.5-10 Gigabit Interconnection



Computing Grid in CAS



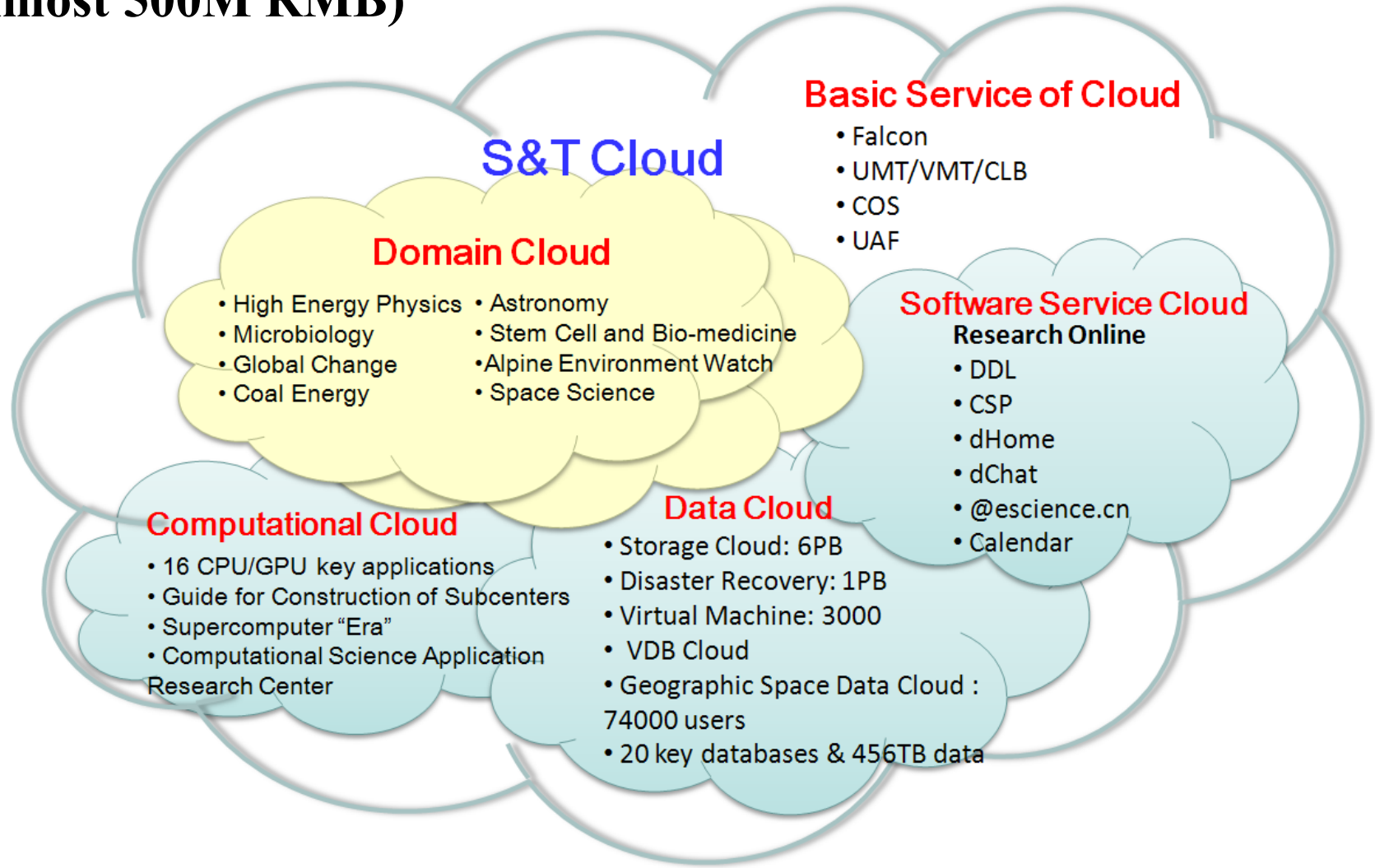
- Aggregated Computing capacity: 1.3Pflops(CPU)+3Pfops (GPU)
- Tier0: 1. 2Pflops
- Tier1: >100Tflops
  - 10 nodes
  - general or specific Purpose
- Tier2: 50Tflops
  - ~20 institutes of CAS

中国科学院 CHINESE ACADEMY OF SCIENCES

S&T Cloud of CAS

中国科学院 计算机网络信息中心 Computer Network Information Center, Chinese Academy of Sciences

Informatization Program of CAS from 2010-2015 (budget is almost 500M RMB)



Why We Need Research data Infrastructure

- Large scientific facilities produce huge data
  - +20 being operation
  - +20 under construction
- Long-Term field observation stations
  - +100 stations including Ecology, Environment, Space, etc.
- Long-Term Research data need to be archived and curation and sharing
  - 100+ institutes

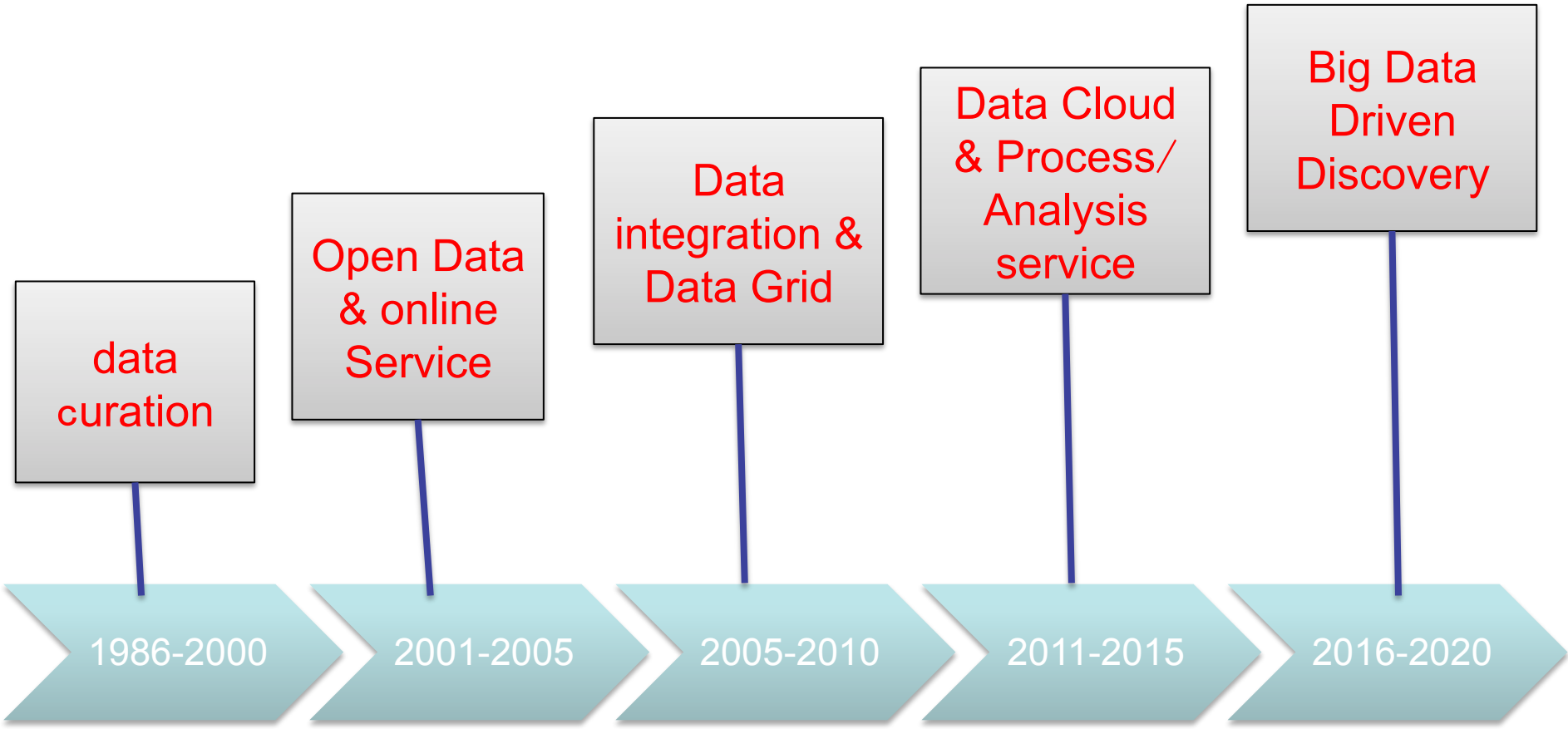
Field observation stations

Large Scientific facilities

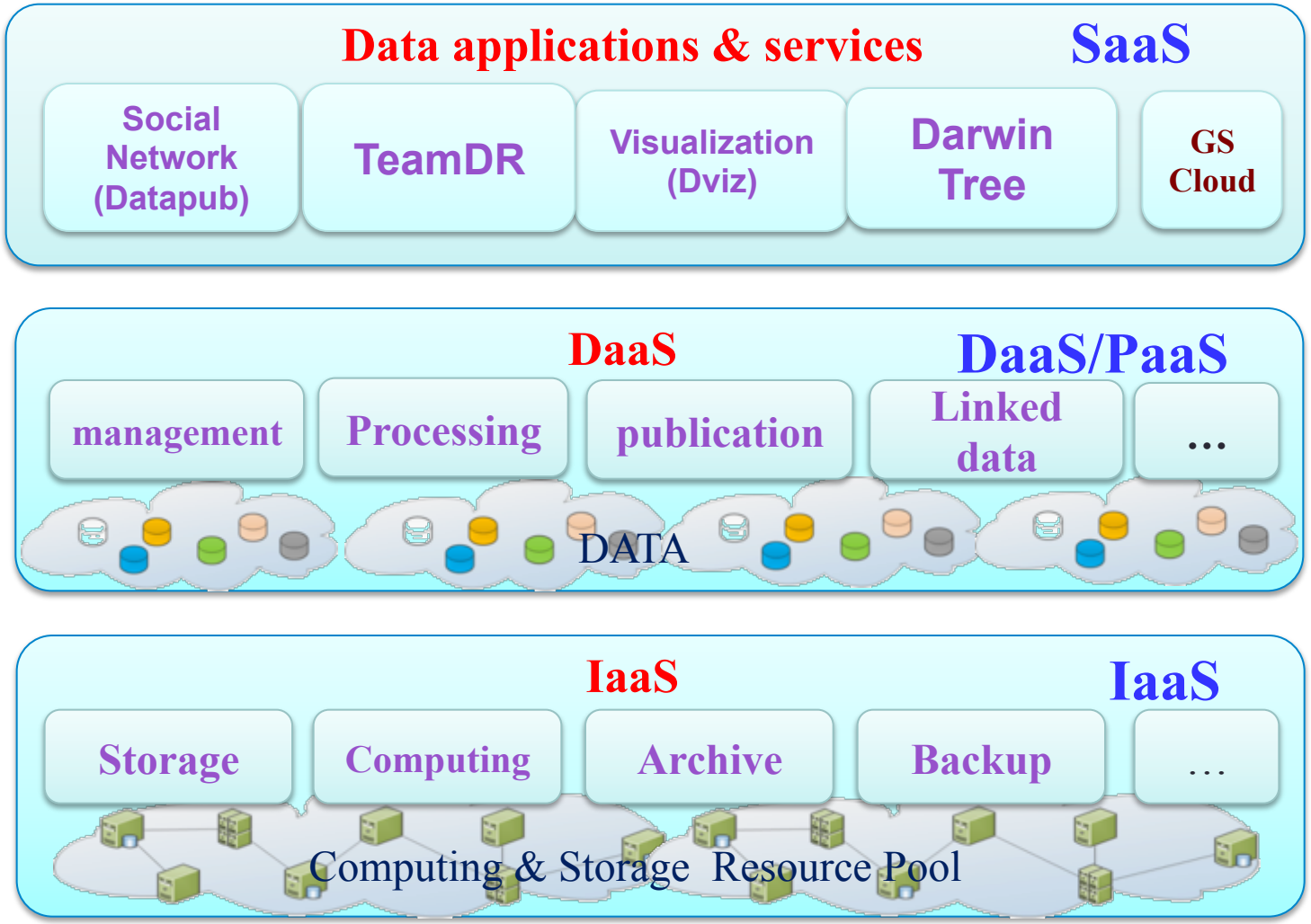
中国科学院
CHINESE ACADEMY OF SCIENCES

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

Evolution of Research Data Infrastructure in CAS



Data Cloud of CAS



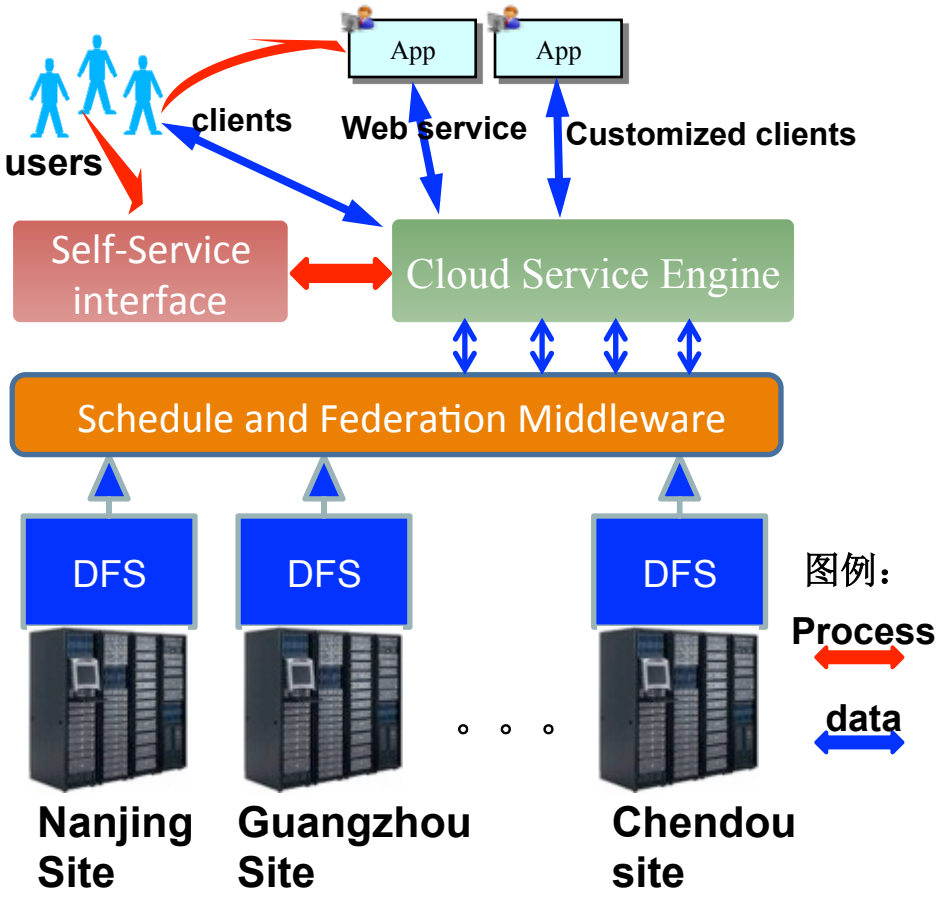
中国科学院
CHINESE ACADEMY OF SCIENCES

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

Data Cloud Portal

IaaS

- Computing Cloud
  - ZeStack based on OpenStack
  - Multi-domain resource management and Federation service
- Storage Cloud
  - Distributed File System
  - Providing one-surface federation service

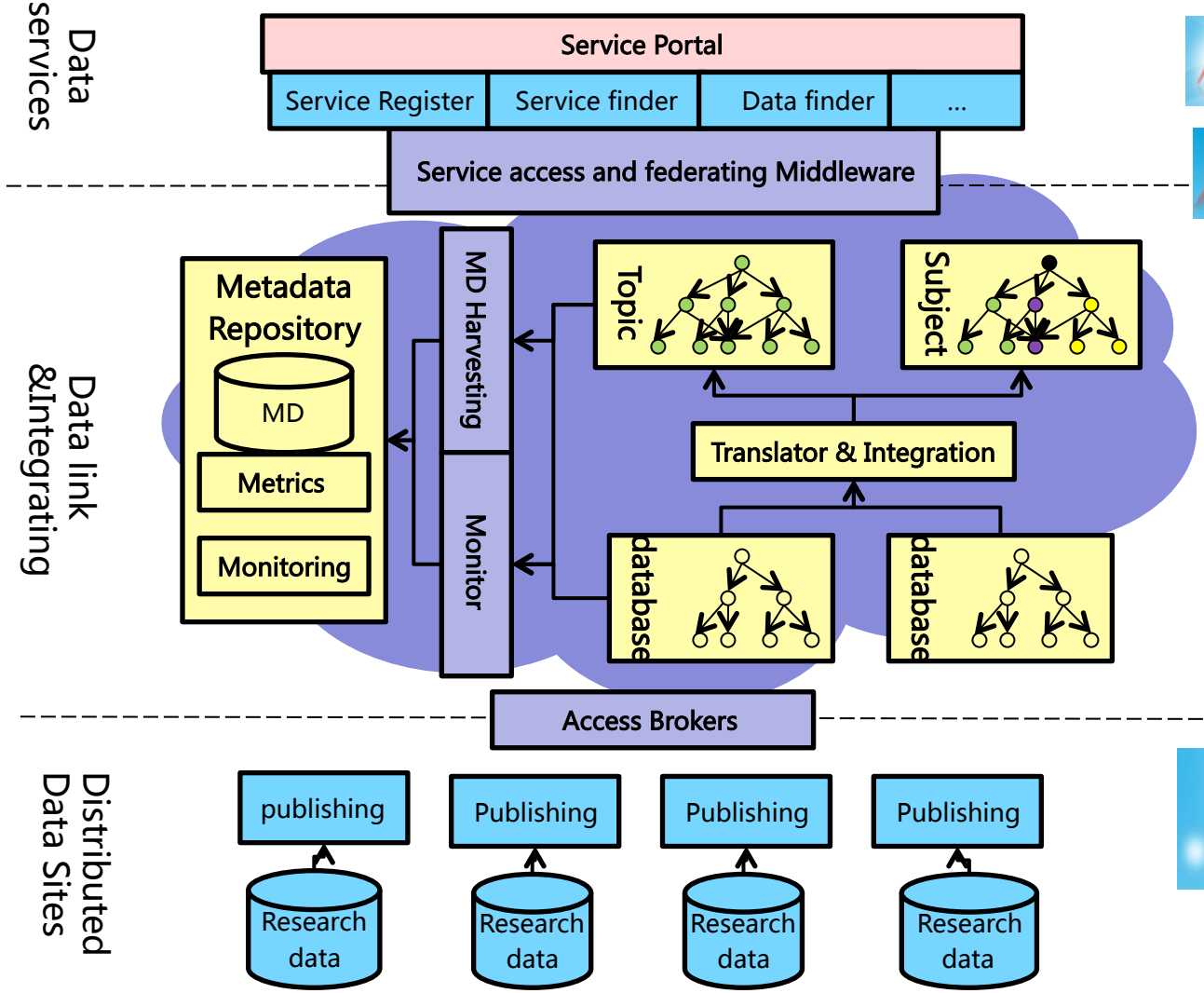


中国科学院
CHINESE ACADEMY OF SCIENCES

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

DaaS

Voovle



中国科学院数据应用环境用户统一认证系统

中国科学院数据应用环境咨询服务系统
Digital Reference Service,DAE,CAS

数据应用环境资源与服务注册系统
Resources & Services Registry, DAE

数据服务监控与统计系统
Scientific database Service monitoring & Statistics system

数据库资源量在线统计系统
Data Collection Statistic System, DataStat

VDBCloud

VisualDB
可视化关系数据库管理发布系统 2.0

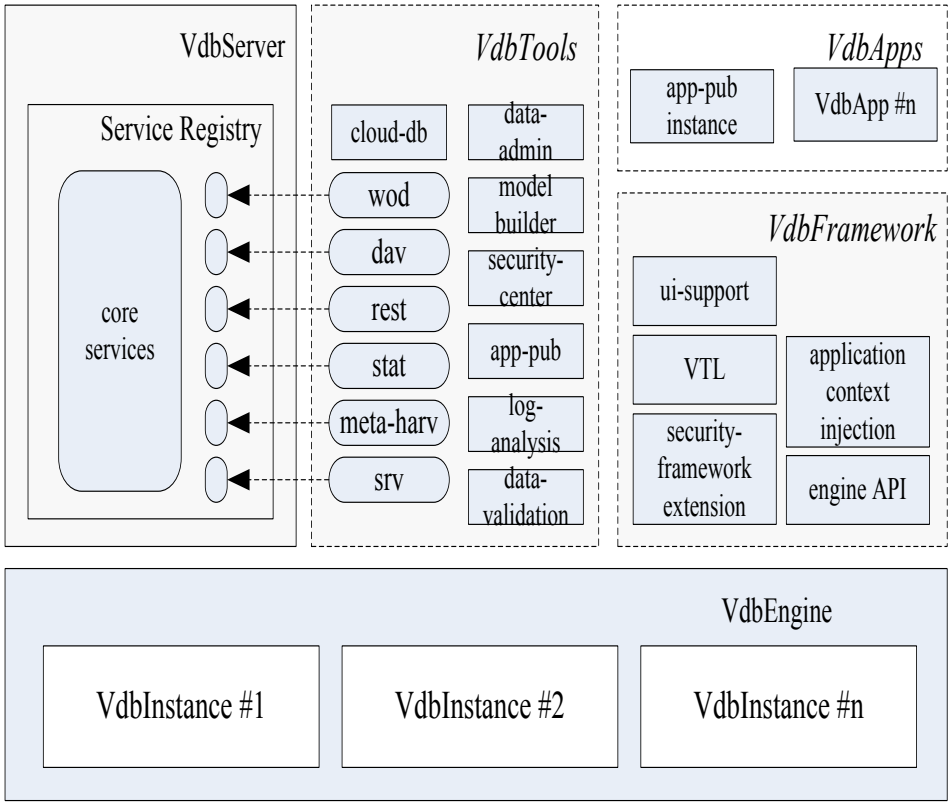
VDB/VDBCloud

① • An invisible **VdbEngine** which organizes all data into a whole database

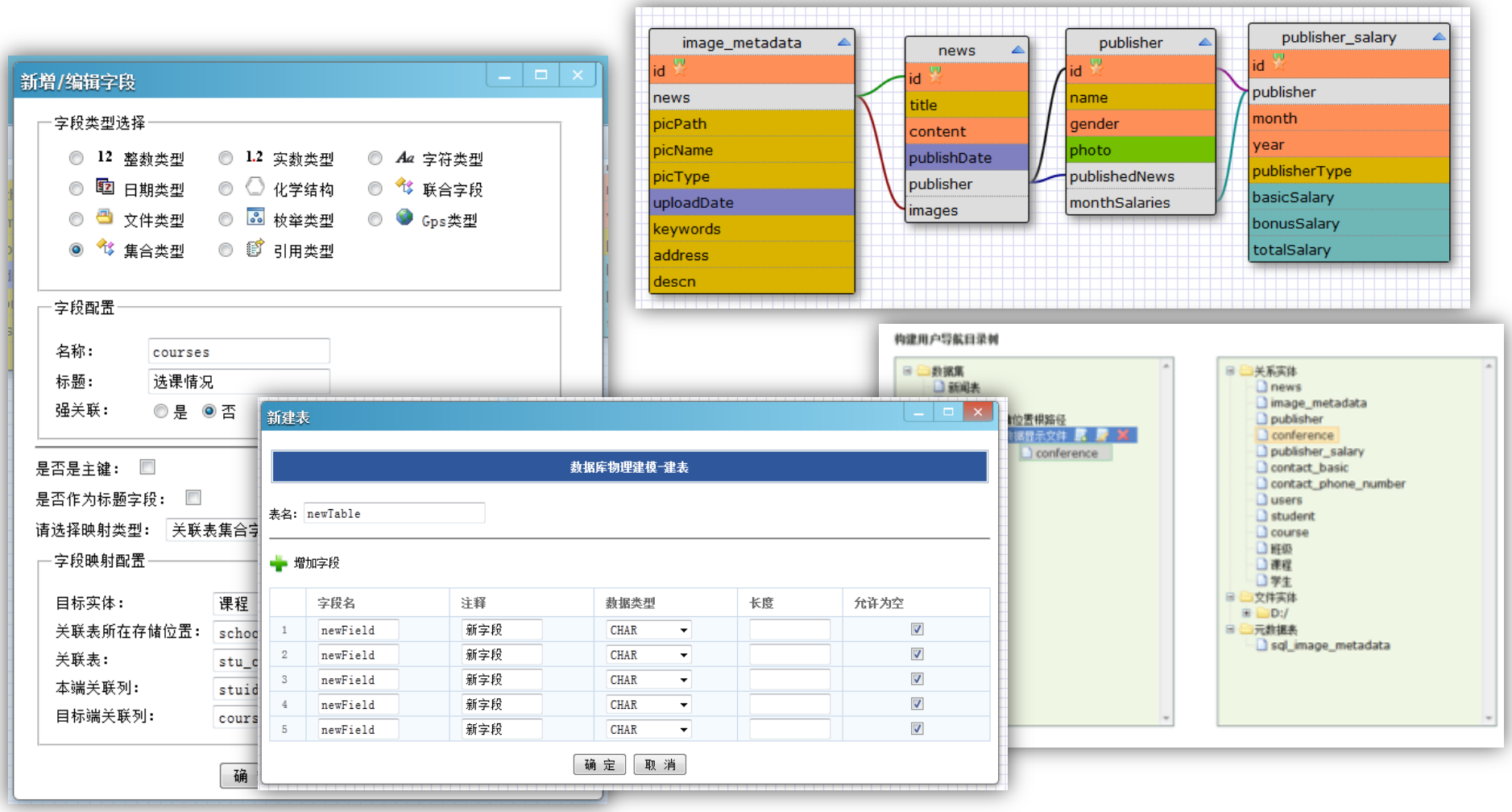
② • A set of **VdbTools** which help data managers manage and publish their data

③ • A **VdbServer** which serves local database as an accessible data source on the Web

④ • A **VdbFramework** which help developers build applications (VdbApps) based on VisualDB

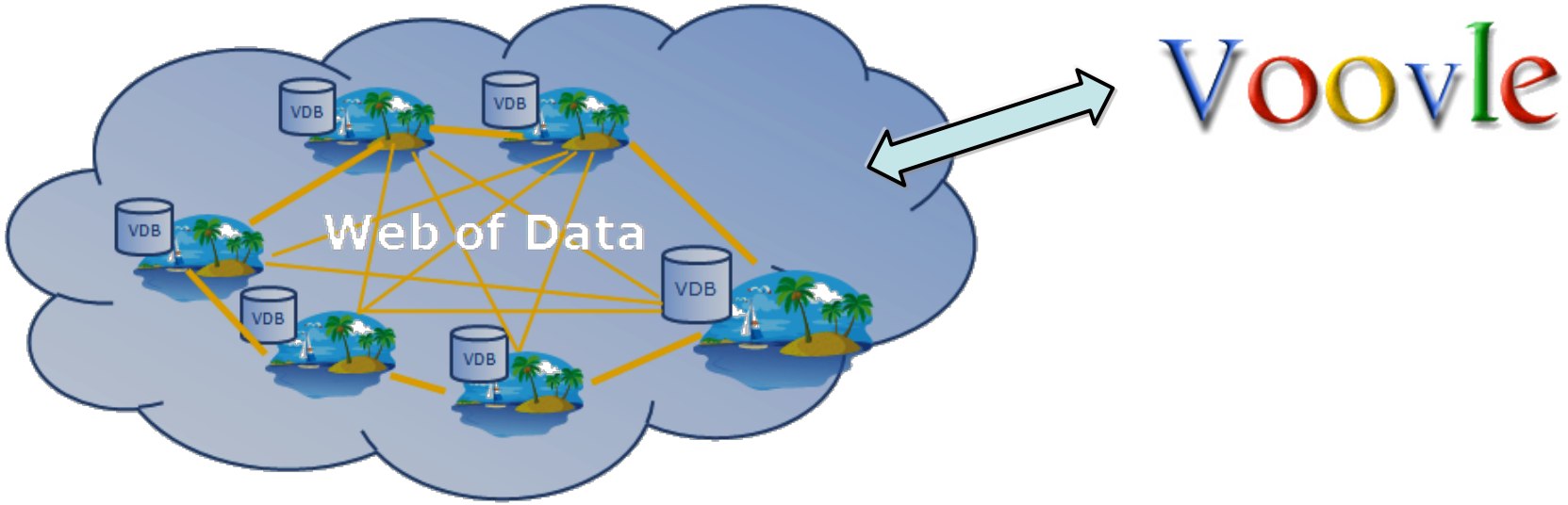


- VdbTools - **ModelBuilder**

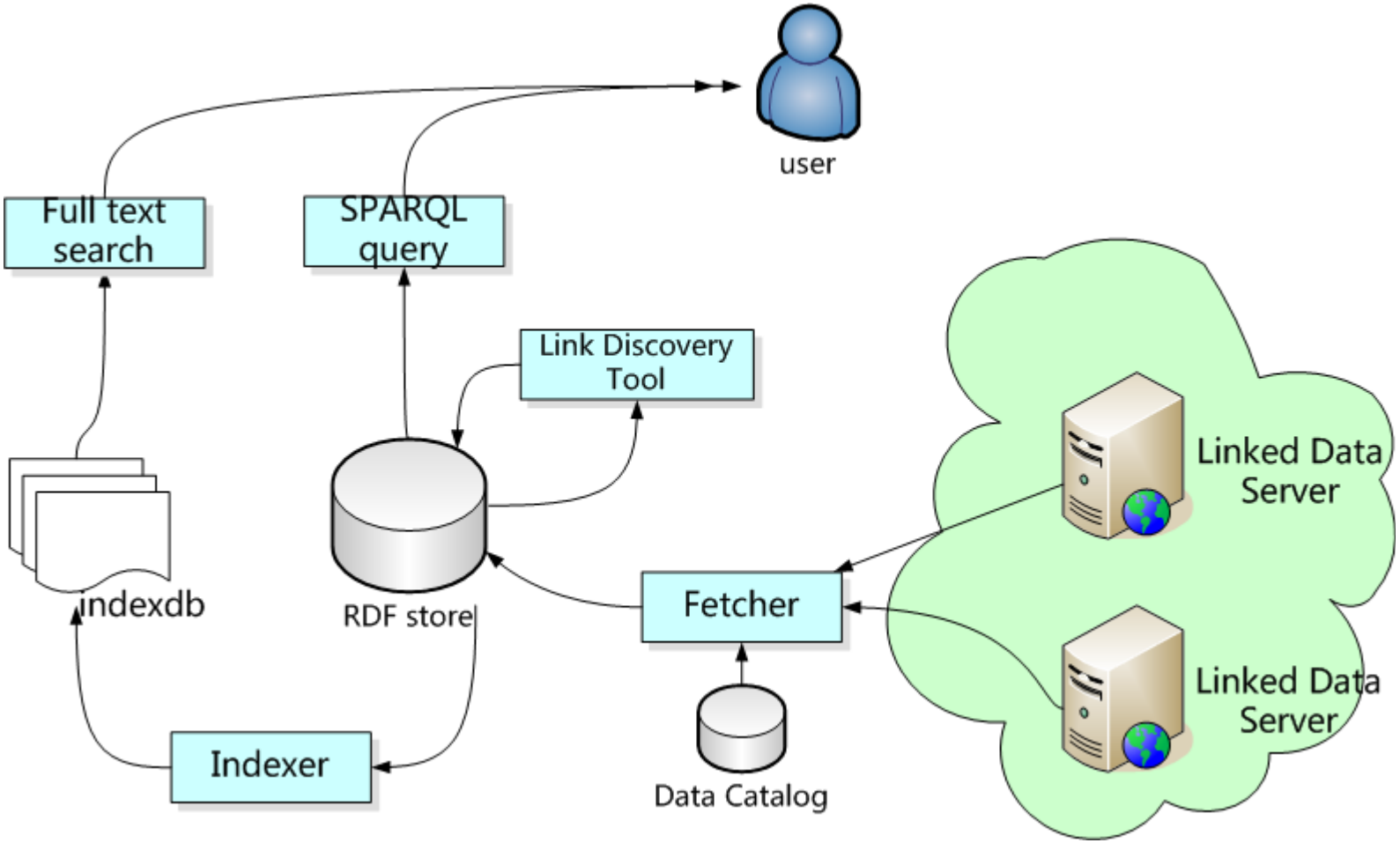


Finding Data Services

- Voovle fetches all scientific data from distributed databases via interfaces exposed by middleware, builds a large data store, creates index, and provides search interfaces for end users.



Architecture of Voovle



Finding Data

- SPARQL query interface
- free text search
  - Voovle use Jena + Lucene to index and search data descriptions

- linking data

标题	红土
form	chemphy2_2
class_id	8
valume_id	valum2_2
土种名称	红土
soil_id	20422
一般性描述	1. 归属与分布 红土，属红粘土亚类红土土属。零星分布于辽宁省丘陵地区，以大连、丹东、营口、朝阳部。现 全部开垦为耕地，面积66.7万亩。 2. 主要性状 该土种母质为红色粘土堆积物，土体深厚，层次分异不棕，红色率由表层向下逐渐增加，平均红色率3.94； A11层质地砂质粘壤土，C层壤质粘土，粘粒含量A11率2.68，硅铝率3.35，铁的游离度36.8%。土壤多呈中性反应，pH6 .5—7.8，阳离子交换量通体较高，在17个农化样分析：有机质含量1.12%，全氮0.067%，速效磷6ppm，速效钾123ppm。有效微量元素硼0.64ppm铁4ppm。 3. 典型剖面 采自大连市甘井子区苏家农校前。低丘缓坡中部，坡度10°，海拔56m。中度片蚀，水量640mm，≥10℃积温 3620℃，无霜期201天。种植玉米。 A11层：0—20cm，浊棕色(湿，7.5YR5/3)，多，pH7.5。 C1层：20—50cm，浊红棕色(湿，5YR5/3)，砂质粘壤土，块状结构，较紧，有植物根系，pH 5YR4/3)，砂质粘壤土，块状结构，紧实，植物根系极少，有少量铁锰胶膜，pH7.8。 C3层：85—130cm，构，极紧，无植物根系，中量铁锰胶膜，pH7.8。 4. 生产性能综述 该土种土体深厚，质地上壤下粘，结构出苗。另外耕作层薄，养分含量缺乏，尤其磷素极缺。 适种作物窄，产量不高。目前种植高粱、谷子、措施：应加强深翻、深松、耙地、压地、铲 等管理措施，增厚 耕作层，改良土壤物理性状；合理施用机质含量；因地制宜，发展果树生产，一般以苹果和山楂为宜。

关联数据

[1] 亚类: 红粘土. [EB/OL] http://www.soil.csdb.cn/wod/resource/cn.csdb.soil.soiltype/soilSubclass/70. [2011-01-23 17:14:33]

[2] 县市名: 大连. [EB/OL] http://www.soil.csdb.cn/wod/resource/cn.csdb.soil.soiltype/sublocation/587. [2011-01-24 05:02:33]

[3] 县市名: 丹东. [EB/OL] http://www.soil.csdb.cn/wod/resource/cn.csdb.soil.soiltype/sublocation/589. [2011-01-24 05:02:37]

[4] 县市名: 营口. [EB/OL] http://www.soil.csdb.cn/wod/resource/cn.csdb.soil.soiltype/sublocation/590. [2011-01-24 05:02:37]

[5] 县市名: 朝阳. [EB/OL] http://www.soil.csdb.cn/wod/resource/cn.csdb.soil.soiltype/sublocation/1675. [2011-01-24 05:24:44]

[6] 理化性质表: 2435. [EB/OL] http://www.soil.csdb.cn/wod/resource/cn.csdb.soil.soiltype/valum1/2435. [2011-01-24 06:13:48]

[7] 理化性质表: 2433. [EB/OL] http://www.soil.csdb.cn/wod/resource/cn.csdb.soil.soiltype/valum1/2433. [2011-01-24 06:13:44]

[8] 理化性质表: 2436. [EB/OL] http://www.soil.csdb.cn/wod/resource/cn.csdb.soil.soiltype/valum1/2436. [2011-01-24 06:13:53]

[9] 理化性质表: 2434. [EB/OL] http://www.soil.csdb.cn/wod/resource/cn.csdb.soil.soiltype/valum1/2434. [2011-01-24 06:13:47]

Descriptive information of red soil, including soil class no., soil class name, general description, etc.

subclass

geographical distribution

physical and chemical properties

中国科学院
CHINESE ACADEMY OF SCIENCES

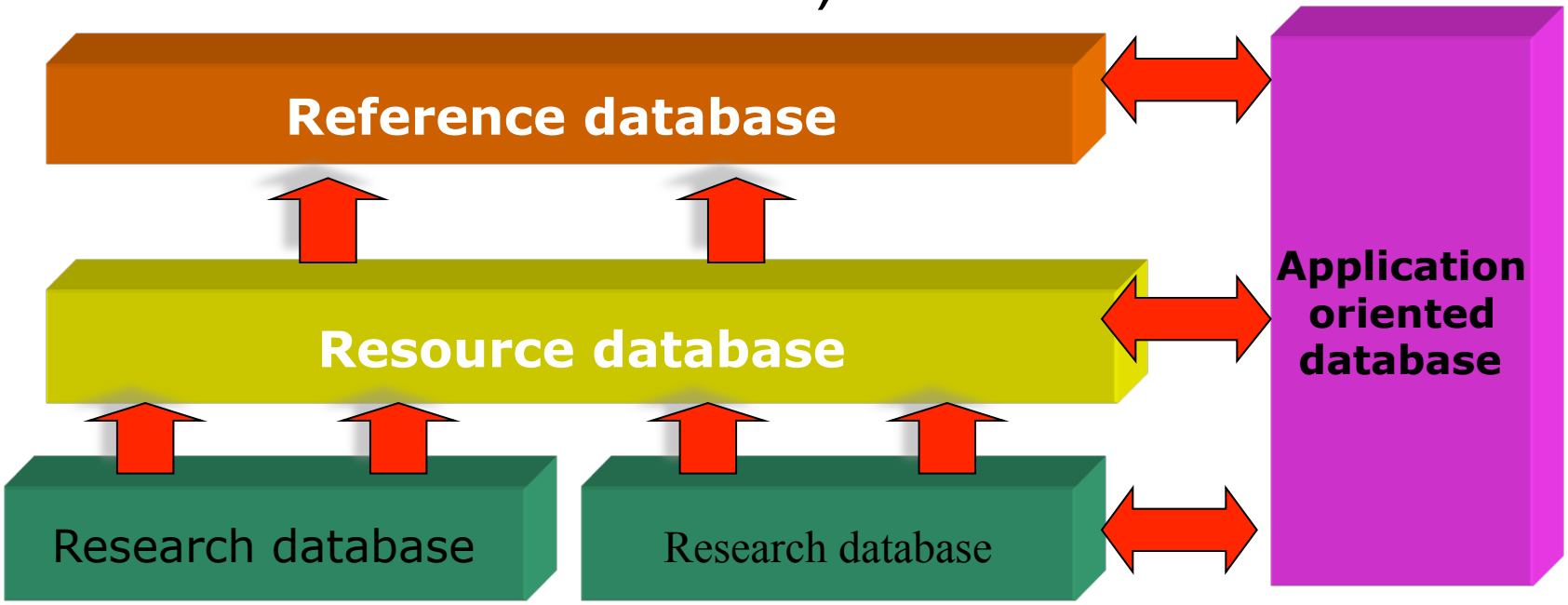
中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

Scientific Databases (SDB)

- A Long-term mission started in 1986 which funded by CAS
  - many institutes involved
  - long-term, large-scale collaboration
  - data from research, for research
- Collecting multi-discipline research data and promoting data sharing
  - More than **350** research databases and **1350** datasets by 61 institutes
  - Over 60**0TB** data available to open access and download

Scientific Databases (cont.)

- focusing on data integration and improving research database to be resource database and even reference database)



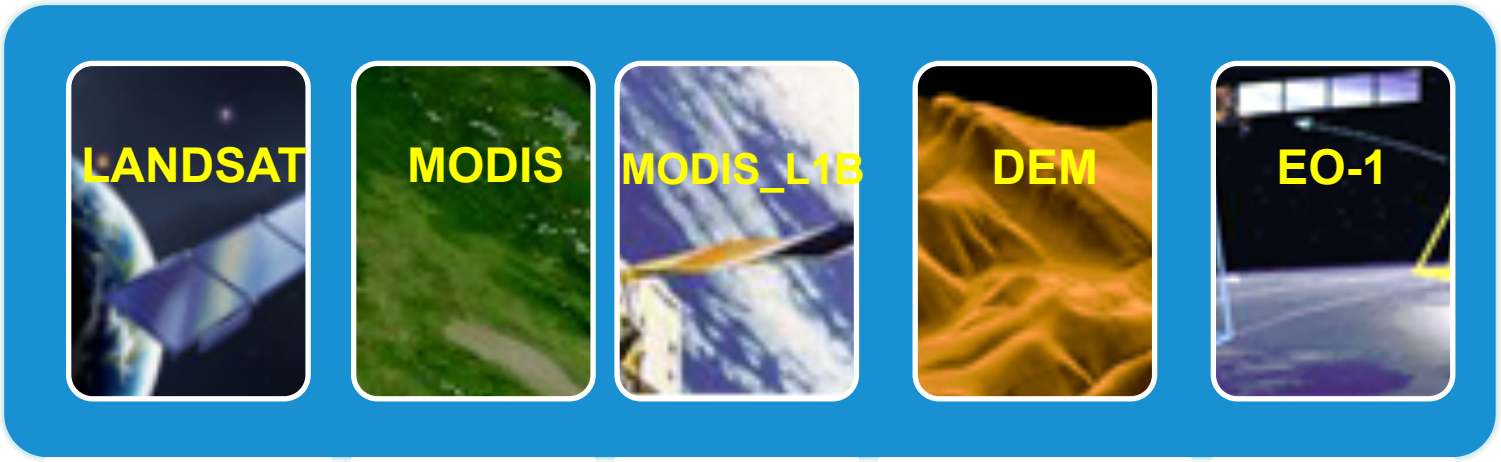
Scientific Databases (cont.)

- **Resource databases**
  - Geo-Science
  - Biodiversity
  - Chemistry
  - Astronomy
  - Space Science
  - Micro biology and virus
  - Material science
  - Environment
- **Reference databases**
  - China Species
  - Compound
- **Application-Oriented databases**
  - High Energy (ITER)
  - Western Environment Research
  - **Ecology research**
  - Qinghai Lake Resource databases
- **Research databases**
  - Physics & Chemistry, Geosciences, Biosciences, Atmospheric & Ocean Science, Energy Science, Material Science, Astronomy &Space Science

Typical Case1: Geospatial Data

- A open data cloud for geospatial data
  - Collecting Global Data sources dynamically
  - Coverage main focus on China and surrounding area
  - Make data Findable, Accessible , Interoperable Reusable（FAIR) and Citable

Data Sources



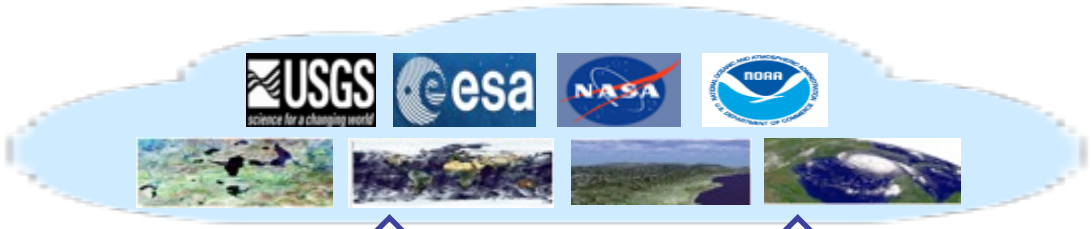
LANDSAT	MODIS	MODIS_L1B	DEM	EO-1
Landsat1-3 MSS Landsat 4-5 MSS Landsat 4-5 TM Landsat 7 SLC-on Landsat7 SLC-off Landsat 8 OLI	TERRA AQUA	TERRA AQUA	SRTM 30米 SRTM 90米 海洋高程	HYP_L1R HYP_L1G AL1

open Data which can be harvested online

中国科学院
CHINESE ACADEMY OF SCIENCES

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

GSCloud Architecture



Service Portal

Harvest Engine

Service finder

Access Data

Metadata Harvesting

Queue scheduling

Translate & filter

Task Monitor

Service Engine

Product process based on crowd sourcing

On-demand Computing Engine

temporal-spatial retrieval Engine

Metadata Repository

Data Pool

Cloud Resource Pool

中国科学院
CHINESE ACADEMY OF SCIENCES

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

Http://www.gscloud.cn

Data resources -- 438.7TB

总数据量：**438.7 TB**　　昨日下载：**475.6 GB**

#	产品名称	数据量	昨日下载
1	LANDSAT	75.3 TB	433.8 GB
2	SENTINELS	15.6 TB	18.2 GB
3	Mosaic_Landsat	305.4 GB	6.6 GB
4	MODIS_PRODUCT_CN	12.4 TB	6.3 GB
5	DEM	1.6 TB	5.3 GB
6	Mosaic_Landsat1984-1997	511.1 GB	2.8 GB
7	modis_land	96.9 TB	2.4 GB
8	MODIS_L1B	200.5 TB	235.2 MB
9	LANDSAT_WATER	2.5 GB	7.7 MB
10	EO_1	3.8 TB	0 bytes
11	ENVIRON_DATA	23.2 TB	0 bytes
12	GLS	7.1 TB	0 bytes
13	global_change	1.5 TB	0 bytes

中国科学院
CHINESE ACADEMY OF SCIENCES

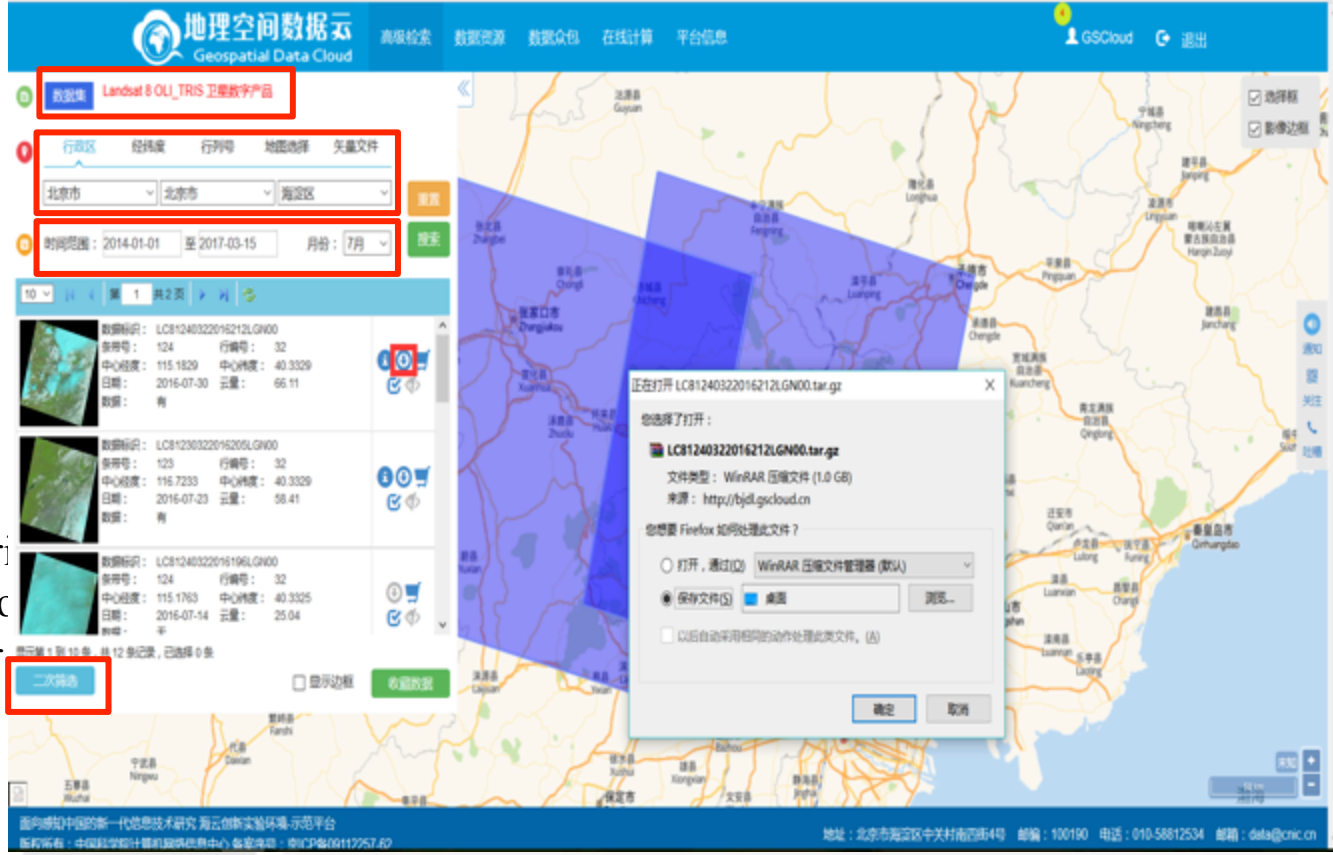
Data retrieve

Query options:

- Dataset Filter
- Geographic Filter
  - Administration boundari
  - latitude and longitude cc
  - Defined regions by user
  - Path and row
  - Shapefile
- temporal range Filter
- Metadata Filter

Download options:

- Download: HTTP
- Bulk Download: FTP



Online data processing

中国科学院 计算机网络信息中心 Computer Network Information Center

地理空间数据云 Geospatial Data Cloud

高级检索 数据资源 数据众包 在线计算 平台信息

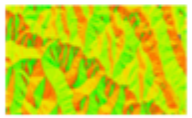
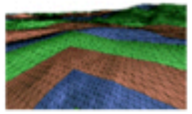
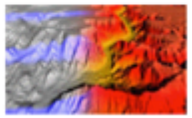
GSCloud 退出

模型分类

- 遥感数据处理
- 地形信息提取
- 遥感数据应用

平台首页 / 在线计算

10 第 1 共1页 显示第1到7条，共7条记录

编号	模型信息	说明	操作	缩略图
1	地形坡向指数 针对DEM数据的使用，基于数字高程数据，计算坡向（Aspect）。 评分：★★★☆☆ 调用次数：211次 创建时间：2016-03-16 13:42	详细	运行	
2	地形阴影指数 针对DEM数据的使用，基于数字高程数据，计算山体阴影（Shaded relief）。 评分：★★★★★ 调用次数：280次 创建时间：2016-03-16 13:42	详细	运行	
3	地形坡度指数 针对DEM数据的使用，基于数字高程数据，计算坡度（Slope）。坡度计算需要在投影坐标系统（PCS）下进行，地理坐标系统(GCS)不支持坡度计算。 评分：☆☆☆☆☆ 调用次数：706次 创建时间：2016-03-16 13:43	详细	运行	
4	地形坡位指数 针对DEM 数据的使用，基于数字高程数据，计算坡位（Topographic Position Index）。坡位系数用来描述地形部位的一个地形参数，在地貌分类中具有十分重要的意义，用于确定研究目标点与其周围地形的位置关系。 评分：★★★★★ 调用次数：546次 创建时间：2016-03-16 13:44	详细	运行	
5	地形粗糙指数 针对DEM数据的使用，基于数字高程数据，计算地形粗糙指数（Terrain Ruggedness Index），地表粗糙度是反映地表起伏变化与侵蚀程度的指标，地面越粗糙，地形粗糙指数越大。 评分：★★★★☆ 调用次数：737次 创建时间：2016-03-16 13:46	详细	运行	
	DEM高程数据切割			

On-demand data processing

1、Set parameters of the model online:
For the DEM Clip Model, parameters include task name, data source (e.g. DEM of 30m and 90m resolution), spatial region, output file format, save path

2、After executing the task, users can check the progress of the task in page “My tasks”.

3、Results can be downloaded at last.

Data processing By crowdsourcing

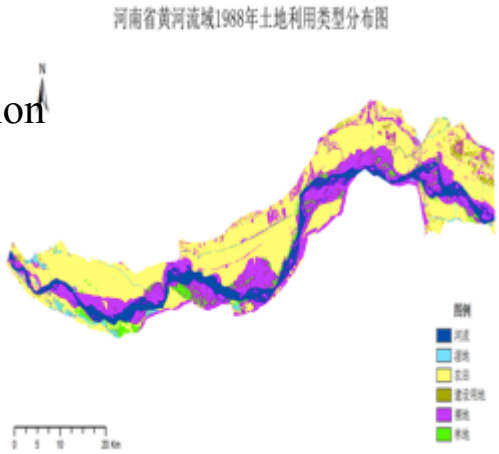
Goals：

Extracting land use information from Remote Sensing images

Problem:

Difficult to do it automatically.

It is human computation task.



Steps:

1.Task Definition and Division

2.Recruitment and Talents Selection

3.Task Execution and Time Control

4. Quality Control and Result Aggregation

中国科学院 CHINESE ACADEMY OF SCIENCES

45 tasks have been published

中国科学院 计算机网络信息中心 Computer Network Information Center, Chinese Academy of Sciences

平台首页 / 任务汇

任务类型: 不限 数据预处理 遥感影像解译 专题图制作 大规模数据获取 地图矢量化 大数据计算 商业数据采购 其他

任务状态: 不限 已发布 报名截止 任务分配 数据处理 任务完成

任务报酬: 不限 100元以下 100-500元 500-1000元 1000-5000元 5000-1万元 1万元以上

发布时间: 不限 最近一周 最近一月 最近半年 最近一年 至 确定

10 第 1 共5页 显示第 1 到 10 条，共 45 条记录，已选择 0 条

编号	名称	任务类型	时间信息	报酬（元）	操作
T2017030001 正在报名	基于Lansat影像的平谷县植被覆盖度指数专题图制作	专题图制作	报名开始:2017-03-10 报名结束:2017-03-15	1000.0	正在报名,已报14人 我要报名
T2017020001 认领结束	基于遥感影像的杭州市萧山区部分区域土地利用类型解译	遥感影像解译	报名开始:2017-02-21 报名结束:2017-03-01	4000.0	报名结束,已报19人
T2017010001 认领结束	基于Landsat影像的嘉礼县土地利用类型解译	遥感影像解译	报名开始:2017-01-04 报名结束:2017-01-09	4000.0	报名结束,已报11人
T2016120006 认领结束	基于Landsat影像的准格尔旗等区域土地利用类型解译	遥感影像解译	报名开始:2016-12-20 报名结束:2016-12-27	6500.0	报名结束,已报10人
T2016120005 认领结束	基于Landsat影像的乌审旗等区域土地利用类型解译	遥感影像解译	报名开始:2016-12-20 报名结束:2016-12-27	6500.0	报名结束,已报12人
T2016120004 认领结束	基于Landsat影像的定边县等区域土地利用类型解译	遥感影像解译	报名开始:2016-12-20 报名结束:2016-12-27	4000.0	报名结束,已报10人
T2016120003 认领结束	基于Landsat影像的平罗县等区域土地利用类型解译 基于Landsat影像解译平罗县、兴庆区、灵武市、盐池县、同心县、鄂托克前旗区域1978年、2000年、2015年三个时期的土地利用类型。	遥感影像解译	报名开始:2016-12-08 报名结束:2016-12-18	7500.0	报名结束,已报9人
T2016120002 认领结束	基于Landsat影像的鄂托克旗土地利用类型解译 基于Landsat影像解译内蒙古鄂托克旗区域1978年、2000年、2015年三个时期的土地利用类型。	遥感影像解译	报名开始:2016-12-08 报名结束:2016-12-18	5000.0	报名结束,已报12人
T2016120001 认领结束	基于Landsat影像的杭锦旗土地利用类型解译 基于Landsat影像解译内蒙古杭锦旗区域1978年、2000年、2015年三个时期的土地利用类型。	遥感影像解译	报名开始:2016-12-08 报名结束:2016-12-18	5000.0	报名结束,已报12人
T2016080002 任务结束	基于Landsat影像的山东省平度市土地利用类型解译 基于Landsat影像，对1995年，2005年，和2015年的山东省平度市进行土地利用类型解译，解译结果分为6大类，耕地（01）、林地(02)、草地(03)、居民点及工矿用地(04)、水域(05)、和未利用地及其它（06）。同时目视解译道路信息，保存为线数据。	遥感影像解译	报名开始:2016-08-19 报名结束:2016-08-28	6000.0	报名结束,已报22人

10 第 1 共5页 显示第 1 到 10 条，共 45 条记录，已选择 0 条

编号	基本信息	毕业学校/工作单位	数据处理经验	解决方案摘要	状态(得分)	备注	操作
1	潘[redacted]平(女) 硕士 157[redacted]3651		处理过landsat数据进行植被覆盖度以及相关数据的计算；进行过野外数据采集及处理	利用arcgis软件，对相关数据进行计算，得出最终需求(详细方案)	未处理(0分)		报名表 删除报名 审核报名
2	马[redacted](男) 硕士 182[redacted]8530			基于NDVI利用像元二分模型反演植被覆盖度(详细方案)	未处理(0分)		报名表 成果示例 删除报名 审核报名
3	徐[redacted](男) 本科 135[redacted]178	鲁东大学(研究生) 河北工业大学城市学院(本科)		收集2007、2012和2016年三期Landsat影像，进行大气校正和几何校正，消除辐射误差和几何畸变；计算NDVI值，根据NDVI计算植被覆盖度指数，利用ArcGIS出图(详细方案)	未处理(0分)		报名表 成果示例 删除报名 审核报名
4	屈[redacted](男) 硕士 152[redacted]911	兰州大学(研究生) 西北师范大学(本科)	参与并负责《全国水土流失动态监测与公告项目》中多个区域的土地利用解译、植被覆盖度计算、土壤侵蚀计算项目，有良好的数据处理能力和项目组织协调能力。如，黄河流域内孤山川流域、皇甫川流域、秃尾河流域、石马川流域等六条流域内基于航空遥感影像的土地利用类型解译、植被覆盖度计算、土壤侵蚀强度计算；内蒙古准格尔旗、新疆尉犁县基于资源三号、高分一号遥感影像的地利用类型解译、植被覆盖度计算、土壤侵蚀强度计算等。影像预处理包括几何校正、大气校正、影像拼接、裁剪、地理配准等。	1、影像准备：下载2007年，2012年和2016年3个时相北京市平谷区的植被覆盖度；2、数据预处理：根据下载影像，做必要的数据预处理，如几何校正、大气校正、缓冲区分析、裁剪等；3、植被覆盖度计算：确定植被覆盖度遥感估算模型，分别计算植被指数和植被覆盖度；4、制作植被覆盖度专题图；5、撰写报告：针对数植被覆盖度计算过程和结果，撰写技术报告。(详细方案)	未处理(0分)		报名表 成果示例 删除报名 审核报名

Task progress and results uploaded by task executors

中国科学院
CHINESE ACADEMY OF SCIENCES

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

registered users： 150,000+

Supporting published papers: 1783

terrain-corrected (L1T) products. They have undergone systematic radiometric and geometric corrections and overall geometric fidelity with ground control points and a digital elevation model ensured (NASA Goddard Space Flight Center, 2011). In 2003, the Scan Line Corrector (SLC) mechanism on-board Landsat 7 permanently failed. Therefore, for the ETM+ SLC-off images used for rice cropping systems classification, multi-image adaptive local regression proposed by the International Scientific Data Service Platform, Computer Network Information Center, CAS (http://datamirror.csdb.cn) was applied to fill in gaps to improve data usability. This method generally used 1−2

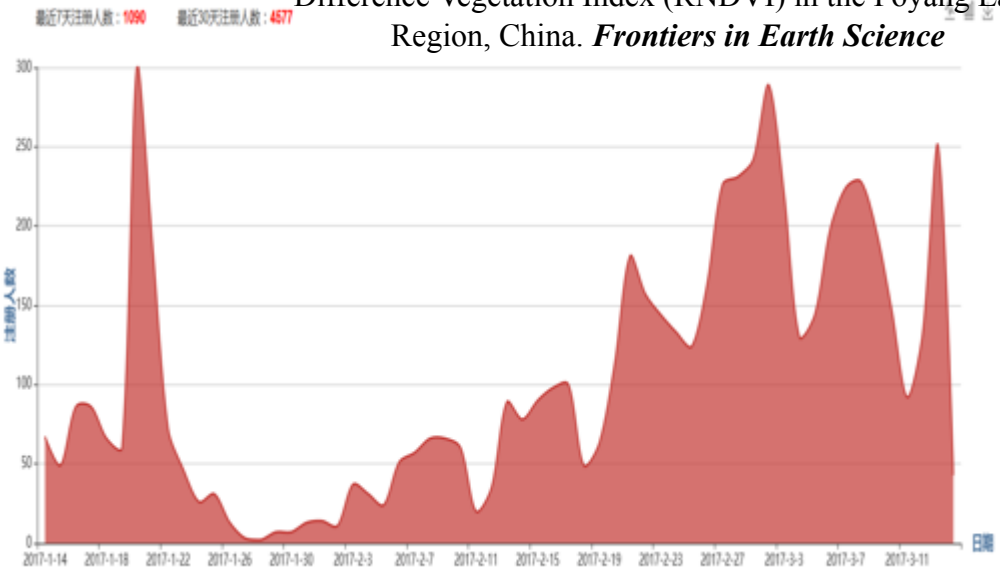
Peng Li et. al., 2015, Mapping rice cropping systems using Landsat-derived Renormalized Index of Normalized Difference Vegetation Index (RNDVI) in the Poyang Lake Region, China. ***Frontiers in Earth Science***

注册用户管理 / 用户列表

20 第 1 共 7656 页 显示第 1 到 20 条，共 153118 条记录 已选择 0 条

二次筛选

编号	邮箱	用户名	单位	学历	注册时间	上次登录
1	971612396@qq.com	t86412068		本科	2017-03-14 15:20	2017-03-14 15:22
2	zhaibing2007@163.com	uiui117		硕士	2016-06-15 09:37	2017-03-14 15:21
3	398875249@qq.com	康晓娟	上海同济城市规划设计研究院	硕士	2015-07-23 12:00	2017-03-14 15:20
4	123829335@qq.com	半久		博士及以上	2017-03-14 14:42	2017-03-14 15:20
5	1129971910@qq.com	张志赟	福州大学	硕士	2015-04-17 17:12	2017-03-14 15:19
6	4F972B5A31D6B7521514E9651C168BB4	娟子柚子			2017-03-14 15:19	2017-03-14 15:19
7	387715609@qq.com	龙泽锟	哈尔滨师范大学	本科	2016-02-04 13:39	2017-03-14 15:17
8	910950311@qq.com	星汉灿烂		本科	2016-03-31 09:44	2017-03-14 15:17
9	3149903302@qq.com	惜	池州学院	本科	2017-03-12 22:16	2017-03-14 15:17
10	790906399@qq.com	颗粒状		本科	2017-02-21 15:02	2017-03-14 15:16
11	755492936@qq.com	expo	辽宁工程技术大学	硕士	2016-09-12 20:45	2017-03-14 15:16
12	1525188076@qq.com	鱼儿		本科	2017-03-14 15:12	2017-03-14 15:14
13	793747426@QQ.COM	①帽丶个逆锛		本科	2016-09-05 21:48	2017-03-14 15:13
14	253852778@qq.com	禹言	郑州大学	硕士	2015-11-11 19:20	2017-03-14 15:12
15	jliang08@163.com	梁军	天津华北地质勘查总院	硕士	2015-03-02 15:59	2017-03-14 15:11
16	41359925@qq.com	休伯利安	生态所	硕士	2011-07-31 14:06	2017-03-14 15:11



Typical CASE: Research Data publication

- Academic journal publishing **multidisciplinary** ***data papers***
- **Quarterly** with **online** bilingual versions
- Open access,efficient processing,high exposure and rapid dissemination
- CN11-6035/N，ISSN 2096-2223

- SCOPE:Data papers describing (but not limited to) :
  - Datasets or data products generated from major scientific activities
  - Derived datasets or data products refined from raw data
  - Datasets linked to existing publications
- China Scientific Data does NOT publish new research findings, or techniques, methods and cases concerning data quality researches and data applications.

- Enhance data Quality
- Credit for data producers
- Incentive for open data

Publishing Process

Author

Editor&Data curator

Experts

Editorial board

SUBMISSION

- data paper
- datasets

REVIEW

- paper review
- data curation

REPEER REVIEW

VOTING

OTHER PROCESS BEFORE PUBLISHING

DATA PAPERS

DOI/Metrics

DATASETS

DOI/Metadata

Editorial office

Data paper

- TITLE
- ABSTRACT
- KEYWORDS
- **Database/Dataset Profile**
- **Introduction/Overview**
- **Data collection and processing**
- **Sample description**
- **Quality control and assessment**
- Values and significance (optional)
- Usage notes (optional)
- Acknowledgments (optional)
- Authors and contributions
- References
- Data citation

Dataset Profile

Chinese title	青藏高原 MODIS 逐日无云积雪面积数据集（2002－2015 年）		
English title	MODIS daily cloud-free snow cover products over Tibetan Plateau (2002 – 2015)		
Corresponding author	Qiu Yubao (qiuyb@radi.ac.cn)		
Data author(s)	Qiu Yubao, Guo Huadong, Chu Duo, Zhang Huan, Shi Jiancheng, Shi Lijuan, Zheng Zhaojun, Laba Zhuoma		
Time range	July 2002 – April 2015		
Geographical scope	The study spans an area of 25° – 45°N and 67° – 107°E, including the entire Tibet Autonomous Region and Qinghai Province, parts of Sichuan, Yunnan, Xinjiang and Gansu provinces, as well as parts of foreign territories in southern and western Tibetan Plateau		
Spatial resolution	500-meter	**Data volume**	6.9 GB
Data format	Geotiff		
Data service system	<http://www.sciencedb.cn/dataSet/handle/55>		
Source(s) of funding	Special Fund for Meteorological Scientific Research in the Public Interest "Constructing a Remote Sensing Product Dataset for Snow Pack over Tibetan Plateau" (No. GYHY201206040); State Key Program of National Natural Science Foundation of China (ABCC Grant No. 41120114001); the National Natural Foundation of China (No. 41371351); and "One-Three-Five" Planning Projects of the Chinese Academy of Sciences.		
Dataset composition	The dataset consists of two parts: the "MODIS daily cloud-free snow cover products over Tibetan Plateau from 2002 to 2015", and the vector data of the research area. They are, respectively: (1) MODIS_Dysno_Cloud-free_2002-2015.zip (6.9 GB), i.e., the daily cloud-free snow data; (2) Tibet_Range.zip (24 KB), i.e., the auxiliary vector data storing the boundary of the study area over Tibetan Plateau.		

Published Data papers over 70

中国科学院
CHINESE ACADEMY OF SCIENCES

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

Big Data for Better Science

+ BIG DATA =

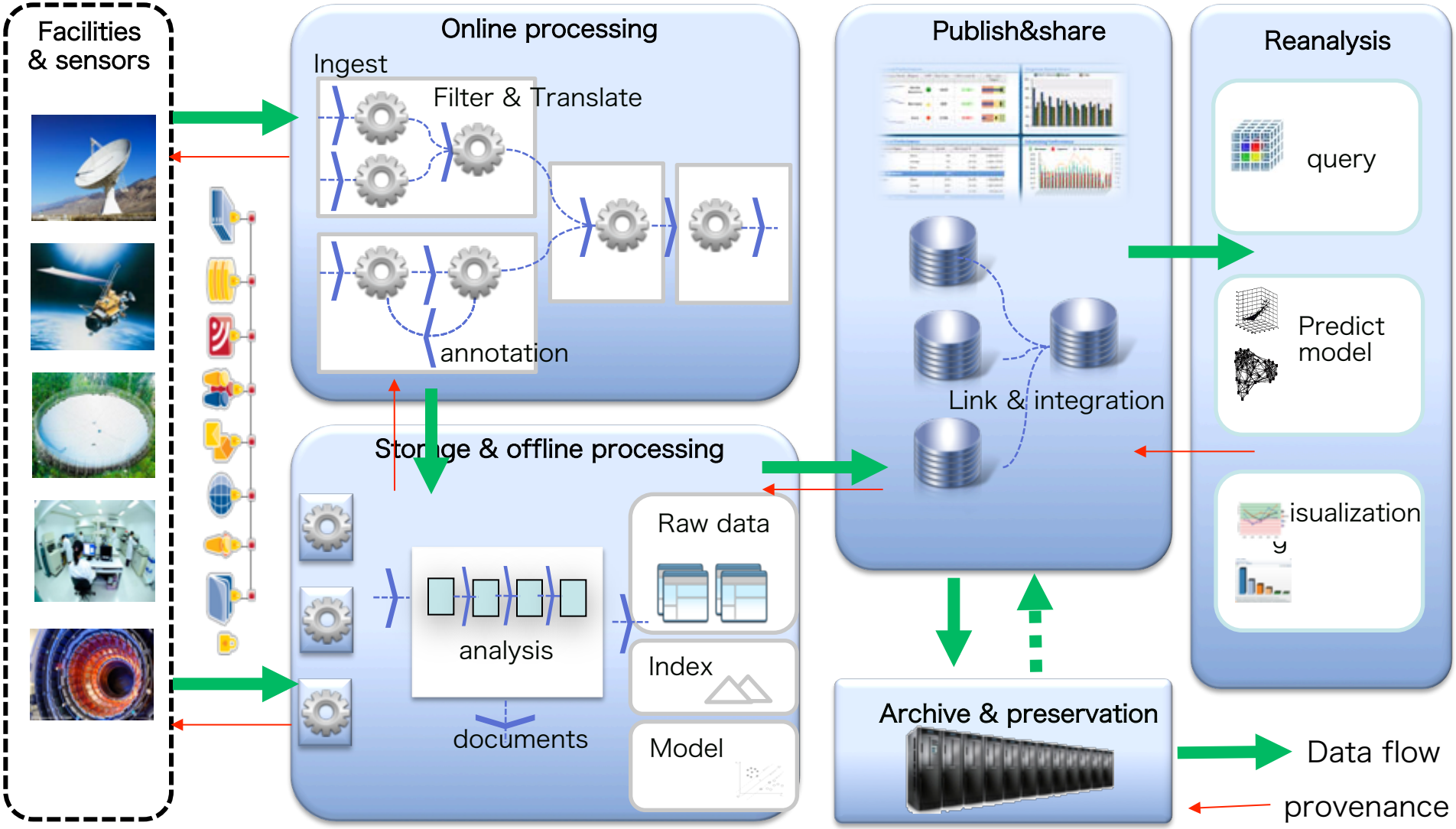
Challenges

- Make data flowing from generator to applications (Data Flow)
  - without barriers, automatically or semi-automatically, on-demand
- Make huge multi-source data linked together as a data network（knowledge graph）
  - Based on semantics, automatically or semi-automatically find invisible relationship
  - Huge graph data management
- Semi-automatically find correlation from multi-source, high dimension data based on feature (Data explore and data Intelligent)
  - Feature selection, machine learning model selection ， training and optimization
- **Make research data infrastructure scalable , evolution and invisible**

中国科学院
CHINESE ACADEMY OF SCIENCES

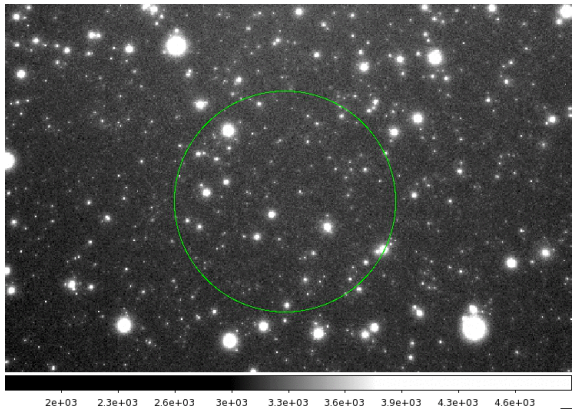
中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

Data Flow Life Cycle

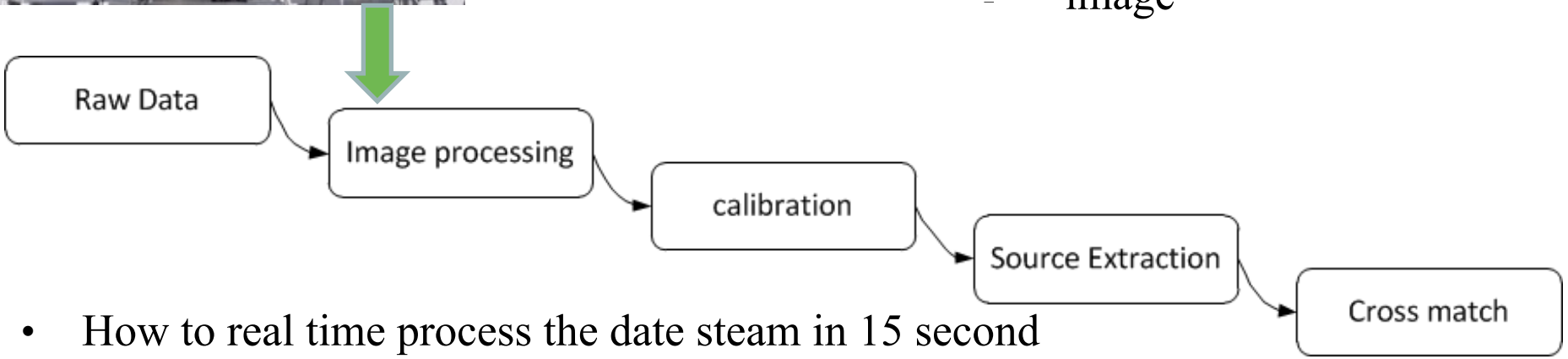


A typical use Case：GWAC

The Ground-based Wide-Angle Camera array, part of SVOM



- 40 telescopes
- 15 second： 40＊32MB Image
- 1. 7 Million object per image



- How to real time process the date steam in 15 second
- How to storage and query 100-1000 billion objects

One Size Does Not Fit All

Take away
One Size Does Not Fit All

- Column store (stupid analytics)
- Array store (smart analytics)
- Streaming (velocity solution)
- New SQL (other velocity solution)
- No SQL (low end; semi-structured data)
- Legacy stuff (in place now - but obsolete)
- One or more curation systems (800 pound gorilla)

- Use the right tool for the job!!!!!

Stonebraker,2015年

one surface fits all？ ！

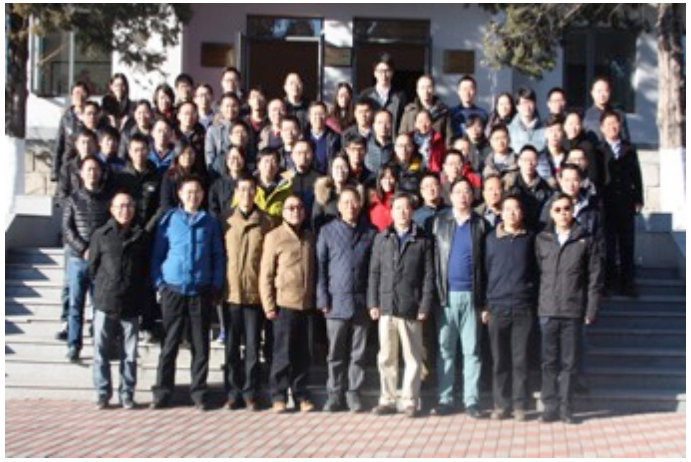
中国科学院
CHINESE ACADEMY OF SCIENCES

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

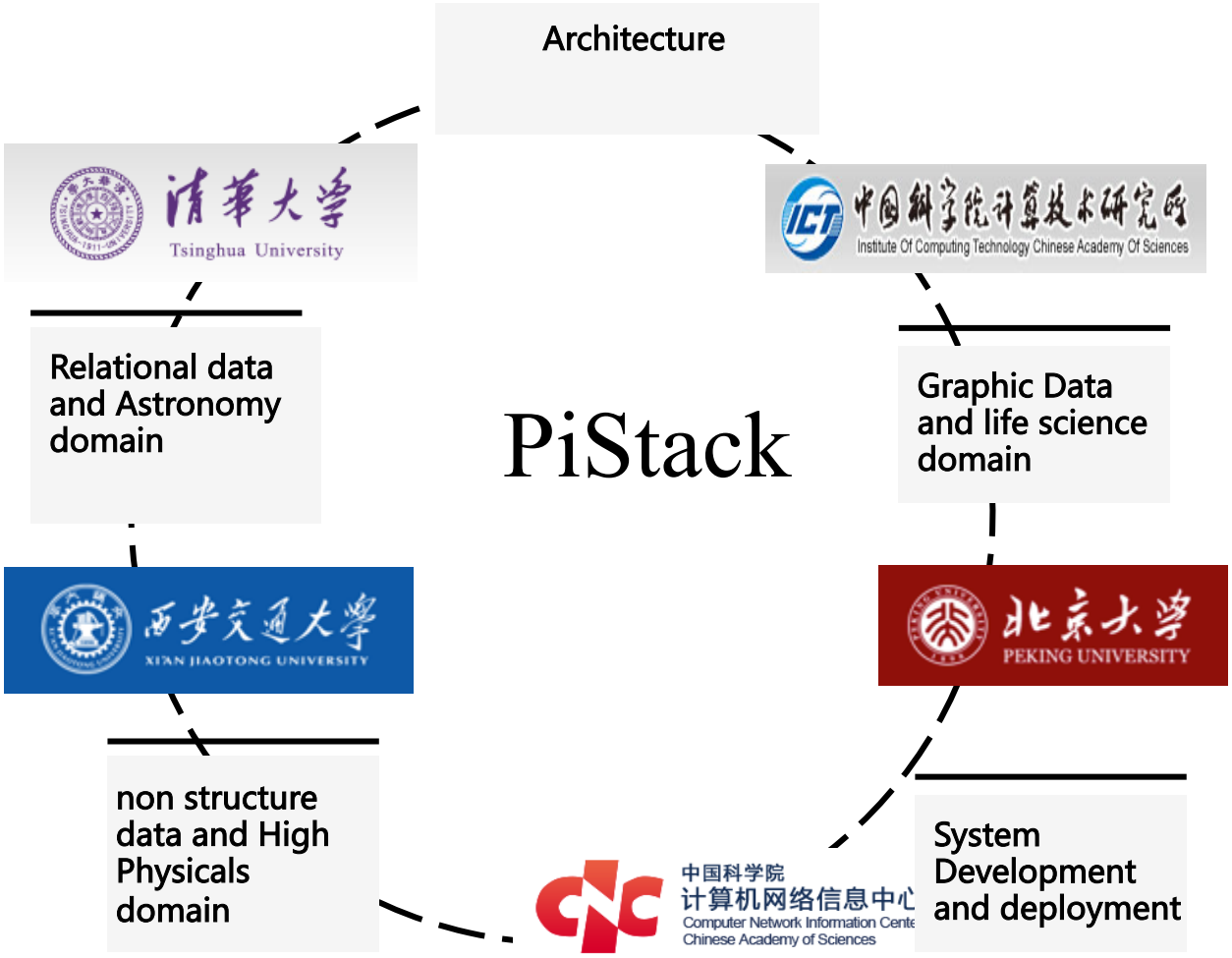
Scientific Big Data Management system(PiStack)

one surface fits all

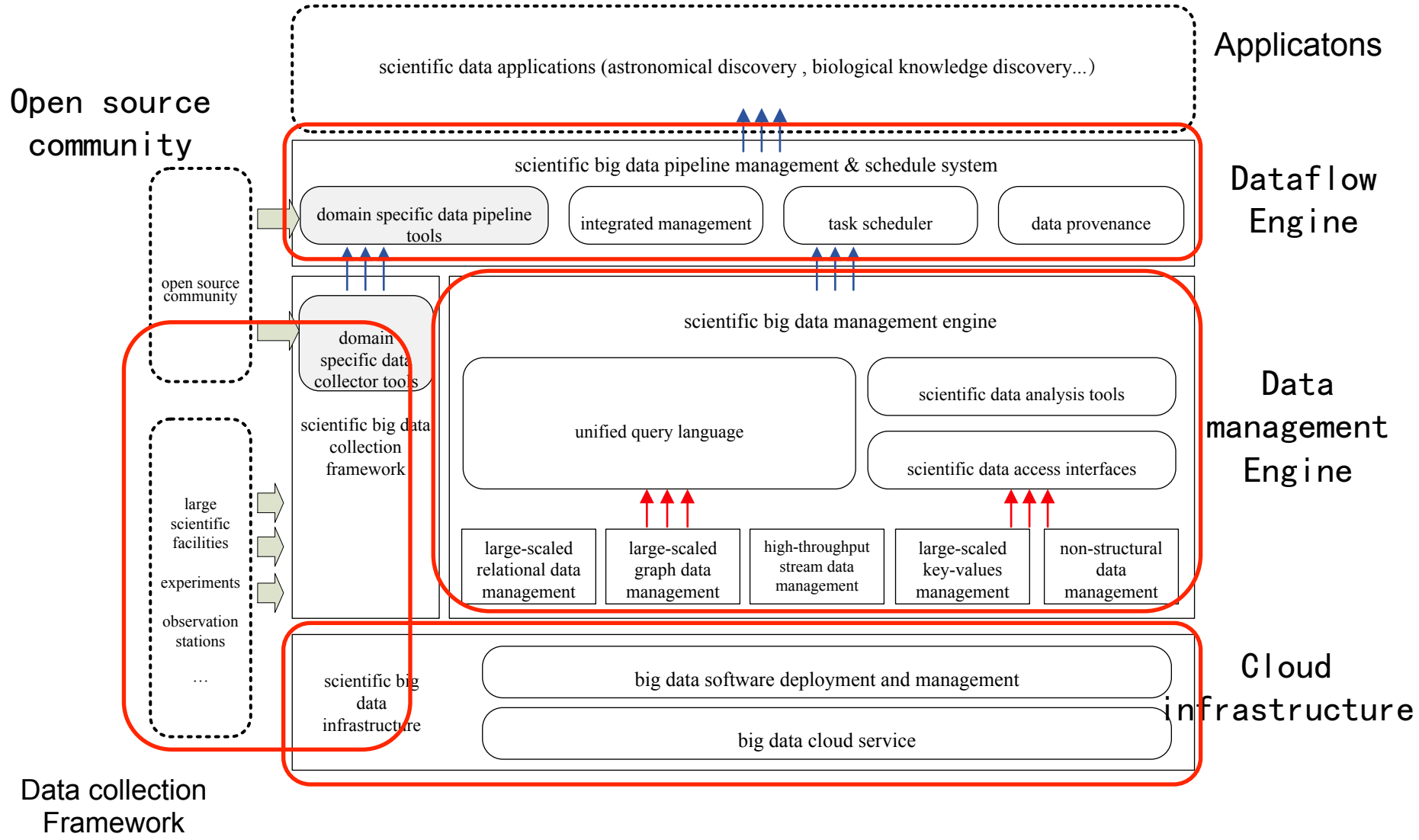
- Accelerate Data Flowing and data link
- A whole life cycle management
- 0pen architecture, Open source, and dynamic evolution
- Funded by National Key Research and Development Program(2017.7-2019.8)



Collaborators



Architecture of PiStack



Progress and outlook

- 3 specific data management model
  - AstroDB(base on relational model)
  - GraphDB(based on graph model)
  - EventDB(based on no-SQL model)
- Simba: A Unified data query and access model
- PackOne: A Cloud resource scheduling and deployment toolkits
- A big data analysis framework for discovery
  - Methodology + specification + implementation + domain specific application
- CAS's Scientific big data program start, it will push PiStack development and deployment.

Conclusion

- Science discovery has increasingly become data intensive, and it calls for reliable and easily accessible scientific data infrastructure
- CAS is always promoting to build scientific data infrastructure to driven better science, especially in big data era
- Seeking potential cooperation with all international Colleagues in big data for science

Thank you!