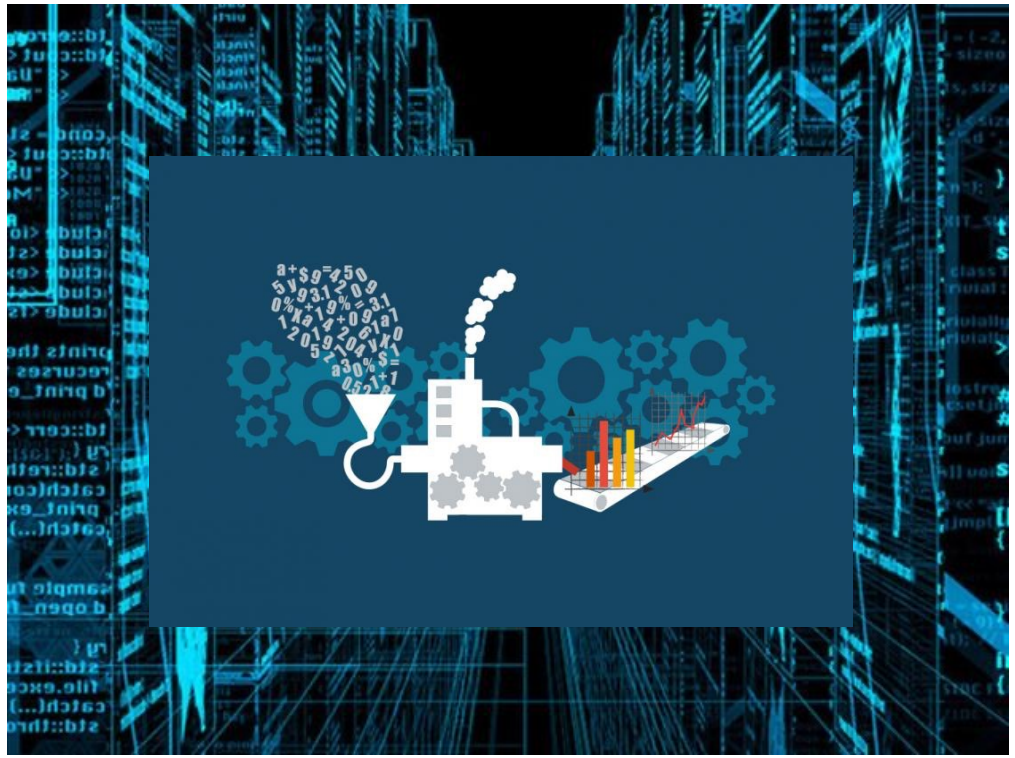
Traitement massif de données à l'époque de l’Entrepôt-cène :

Comment mouliner des données sans réinventer la roue ?



“BIG DATA” ?

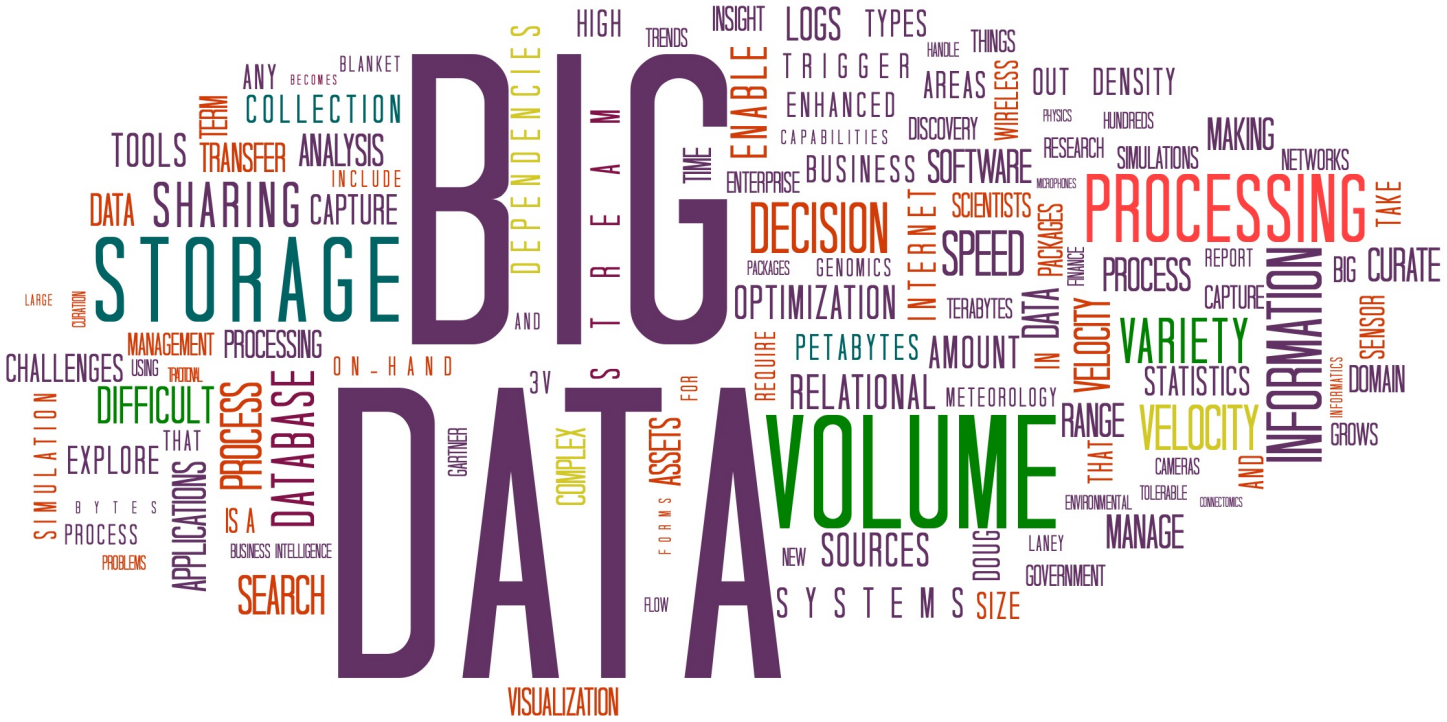
De gros VOLUMES c’est évident

Mais aussi :

- **Grande VITESSE**
- **Grande VARIÉTÉ**

Ainsi que:

- **VÉRACITÉ (qualité, incertitudes)**
- **VALEUR**
- **…**



“BIG” c’est combien ?

Qu’est-ce qu’un gigabyte ?
1 gigabyte équivaut à 1000 megabytes.
Qu’est-ce qu’un terabyte?
1 terabyte équivaut à 1000 gigabytes.
Qu’est-ce qu’un petabyte?
1 petabyte équivaut à 1000 terabytes.
Qu’est-ce qu’un exabyte?
1 exabyte équivaut à environ 1000 petabytes.
Qu’est-ce qu’un zettabyte?
1 zettabyte équivaut à environ à un trillion de gigabytes (1 milliard de milliards de GB).

44 zettabytes : volume total de l’univers digital en 2020 - **En moyenne, chaque humain génére 1.7 MB de données par seconde in 2020**

463 exabytes : volume de données générées chaque jour en 2025 avec plus de 200 zettabytes de données stockées dans le cloud au niveau Mondial

Ordres de grandeurs pour notre domaine

ICARE : 6 Pbytes cumulés online début 2022

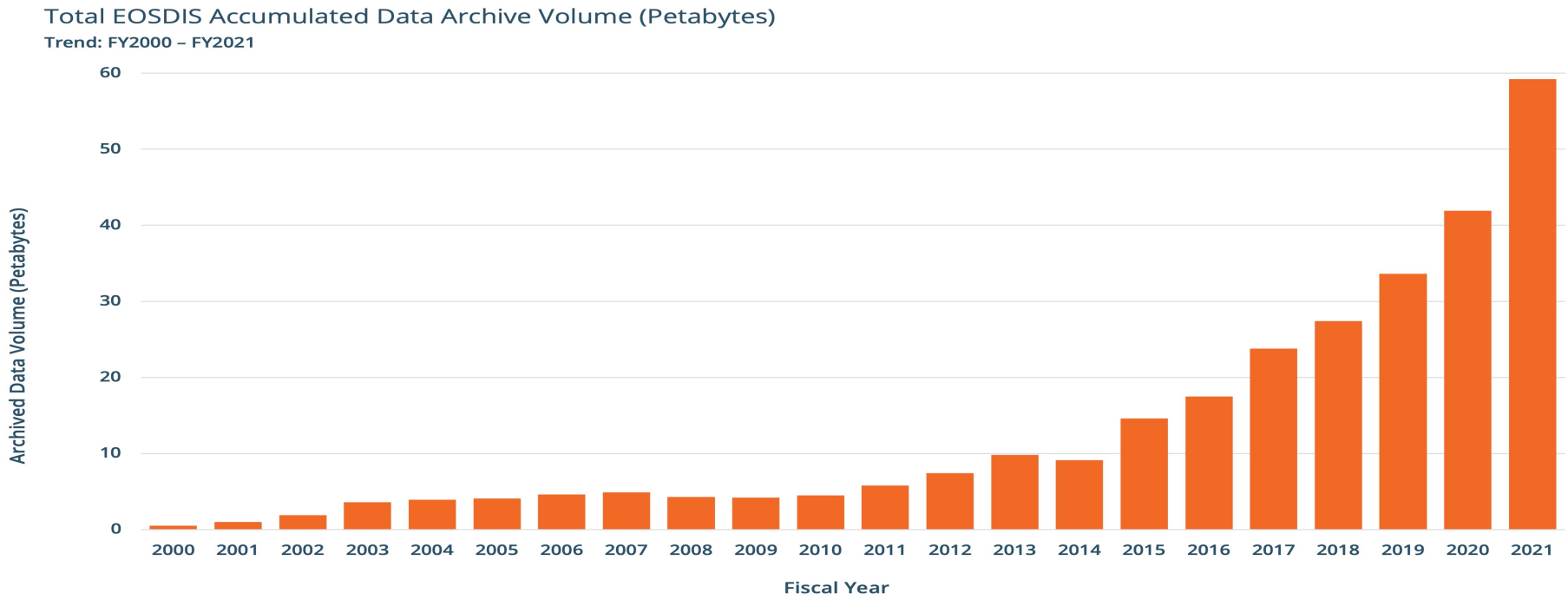
COPERNICUS : 8 TB / jour = 3Pbytes/an (S1, S2 et S3 uniquement)

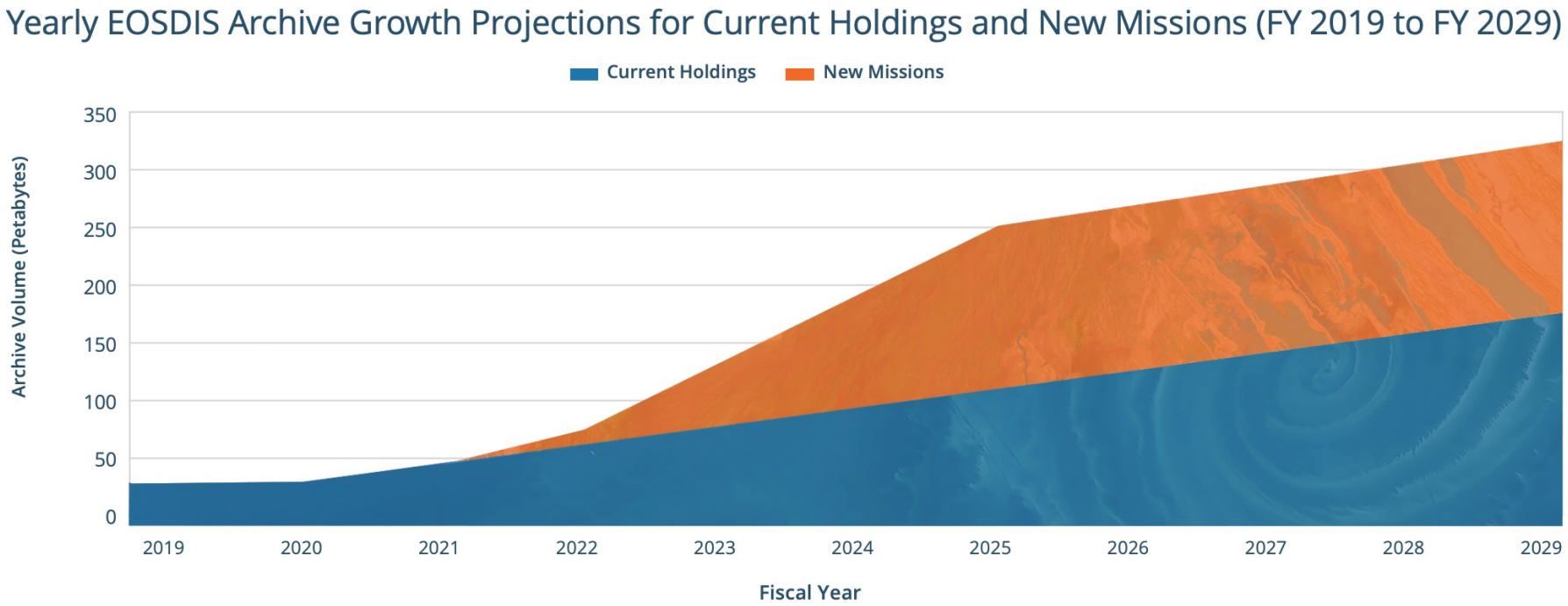
MTG Imageurs et Sondeurs: ~400 Tb / an
EPS-SG: IASI-NG, 3MI, S5, VII : >300 Tb/an

Earth Science Data Systems / NASA :
34 Petabytes en 2020 -> 250 Petabytes en 2025 !

Anthropocène, Entrepôt-cène ou “En trop”- Po -Cène

Exemple :
Croissance volumétrie
EOSDIS / NASA

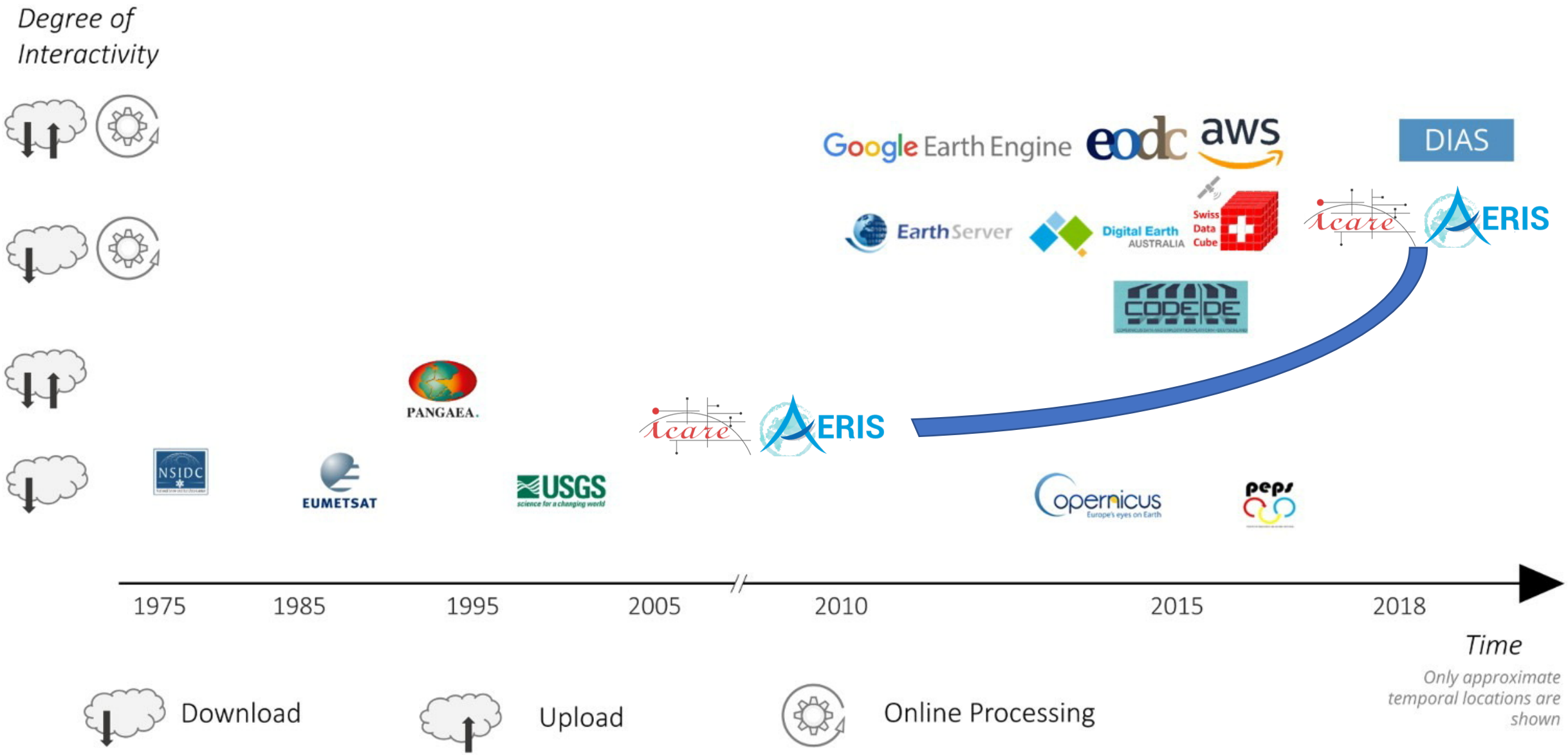




Le paradigme du Big Earth Data

- Exploiter, analyser et valoriser des volumes croissants de données et nécessitant des entrées de plus en plus variées.
- Adapter les “workflows” traditionnels et les méthodes d’analyses pour permettre des traitements dans des temps “raisonnables” (depend de l’application)
- Gérer la répartition des ressources entre les traitements massifs “off-line”, le processing en Near-Real Time, et l’utlisation croissante des interfaces web pour les utilisateurs finaux et les data analysts.

Evolution des services de traitement de données



Les DIAS : Data Integration & Analysis System

Data Integration & Analysis System (DIAS) visent à collecter, archiver, intégrer et analyser les grandes masses de données issues de l'observation de la Terre et à convertir ces données en information utiles à la société pour faire face aux problèmes environnementaux à l'echelle globale, à la gestion des risques associés aux événements extrèmes, etc ..

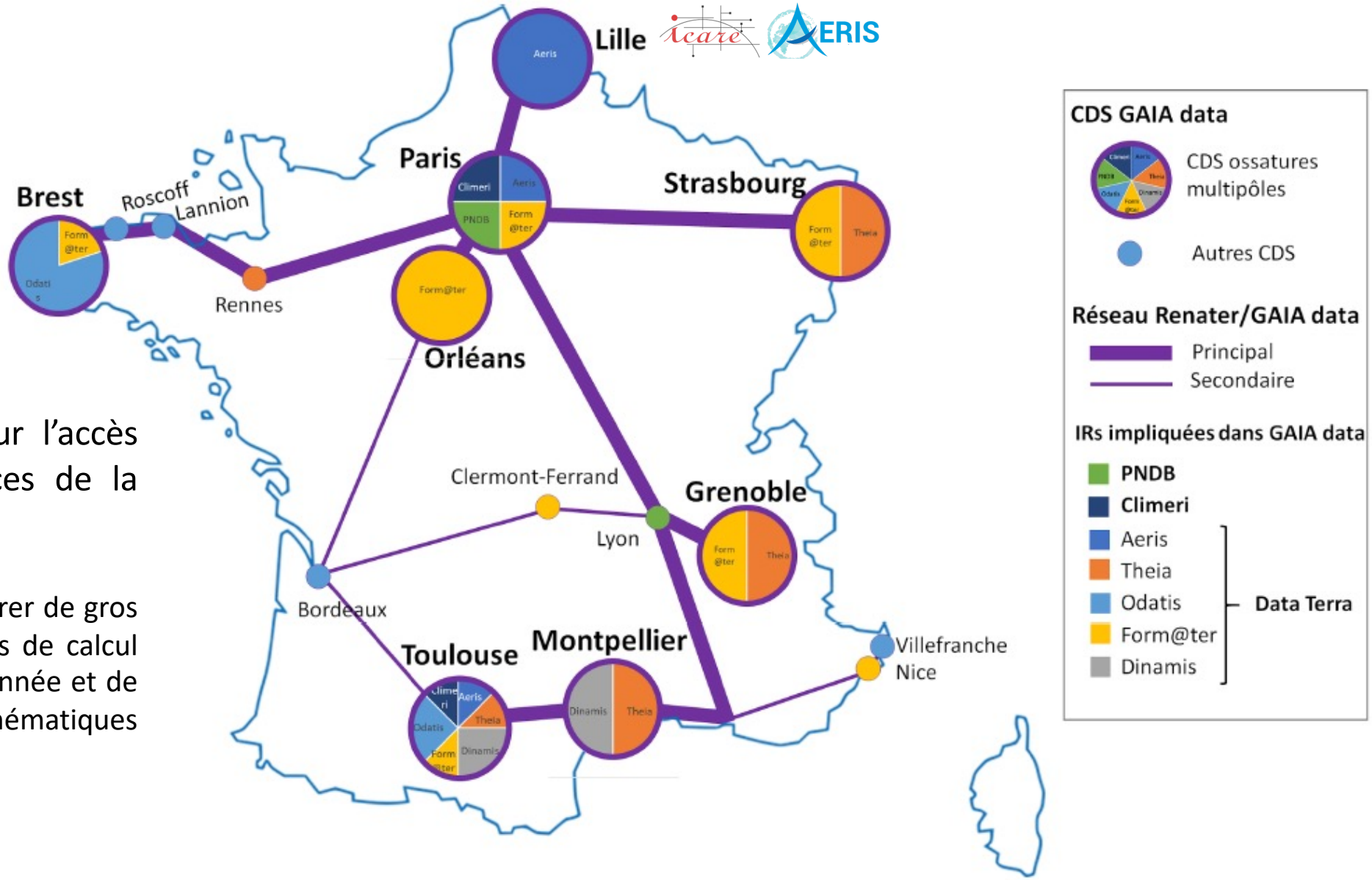
Pour faciliter et normaliser l'accès aux données, la Commission européenne a financé le déploiement de cinq plateformes basées sur le cloud offrant un accès centralisé aux données et informations Copernicus, ainsi qu'aux outils de traitement. Ces plates-formes sont connues sous le nom de DIAS, ou services d'accès aux données et à l'information.

Les cinq plateformes en ligne DIAS permettent aux utilisateurs de découvrir, manipuler, traiter et télécharger des données et des informations Copernicus. Toutes les plateformes DIAS fournissent un accès aux données de Copernicus Sentinel, ainsi qu'aux produits d'information issus des six services opérationnels de Copernicus, ainsi qu'aux outils basés sur le cloud (open source et/ou basés sur un paiement à l'utilisation).

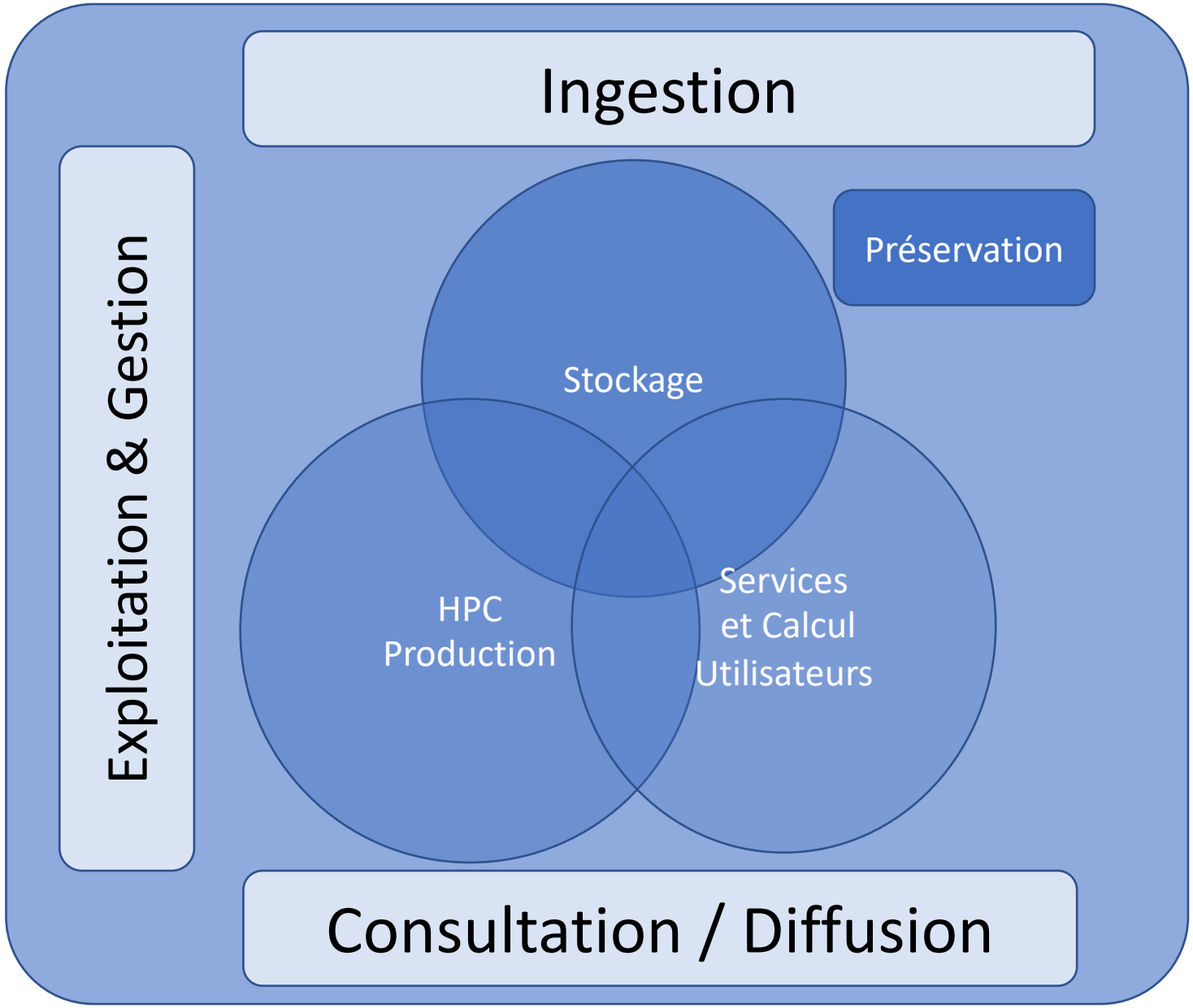
IR DATA TERRA - PROJET GAIA-DATA

8 CENTRES OSSATURES pour l'accès aux données et aux services de la plateforme.

Ils disposent de la capacité de gérer de gros volumes de données, de moyens de calcul significatifs, et d'experts de la donnée et de l'informatique des différentes thématiques de GAIA data.



Architecture AERIS/ICARE pour le Big Data



Les principes de base ICARE :

- Toutes les données sont accessibles en direct et en ligne
- Tous les moyens de calcul voient toutes les données
- Tous les workflows sont gérés par un système centralisé de déclenchement des ordres et d'allocation des ressources

Infrastructure du CDS ICARE : les chiffres

Moyens de Calcul

- 126 serveurs : 43 serveurs racks + 86 lames blade
- 252 CPU pour 2571 cœurs
- théoriquement 74.3 Tflops DP

Stockage

- 9 baies reliées aux serveurs GPFS par FiberChannel
- 1358 disques pour un volume brute de 12 Po soit 10 Pio utiles

Réseau

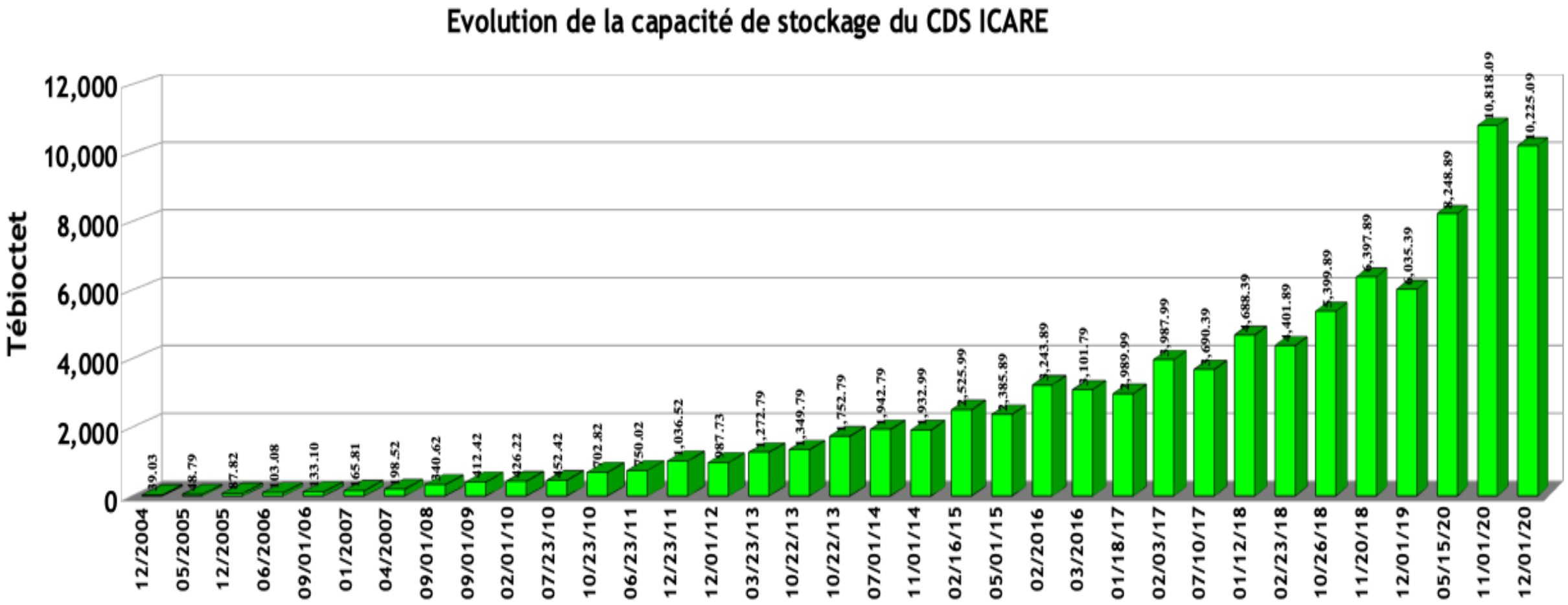
- Connexion 10Gb bidirectionelle directe à RENATER (composante française de GEANT)
- Réseau interne 10GbE/40GbE/100GbE
- 13 switchs FC 8/16Gb et 19 swtichs pour la partie cuivre en 10GbE/40GbE/100GbE

Autres

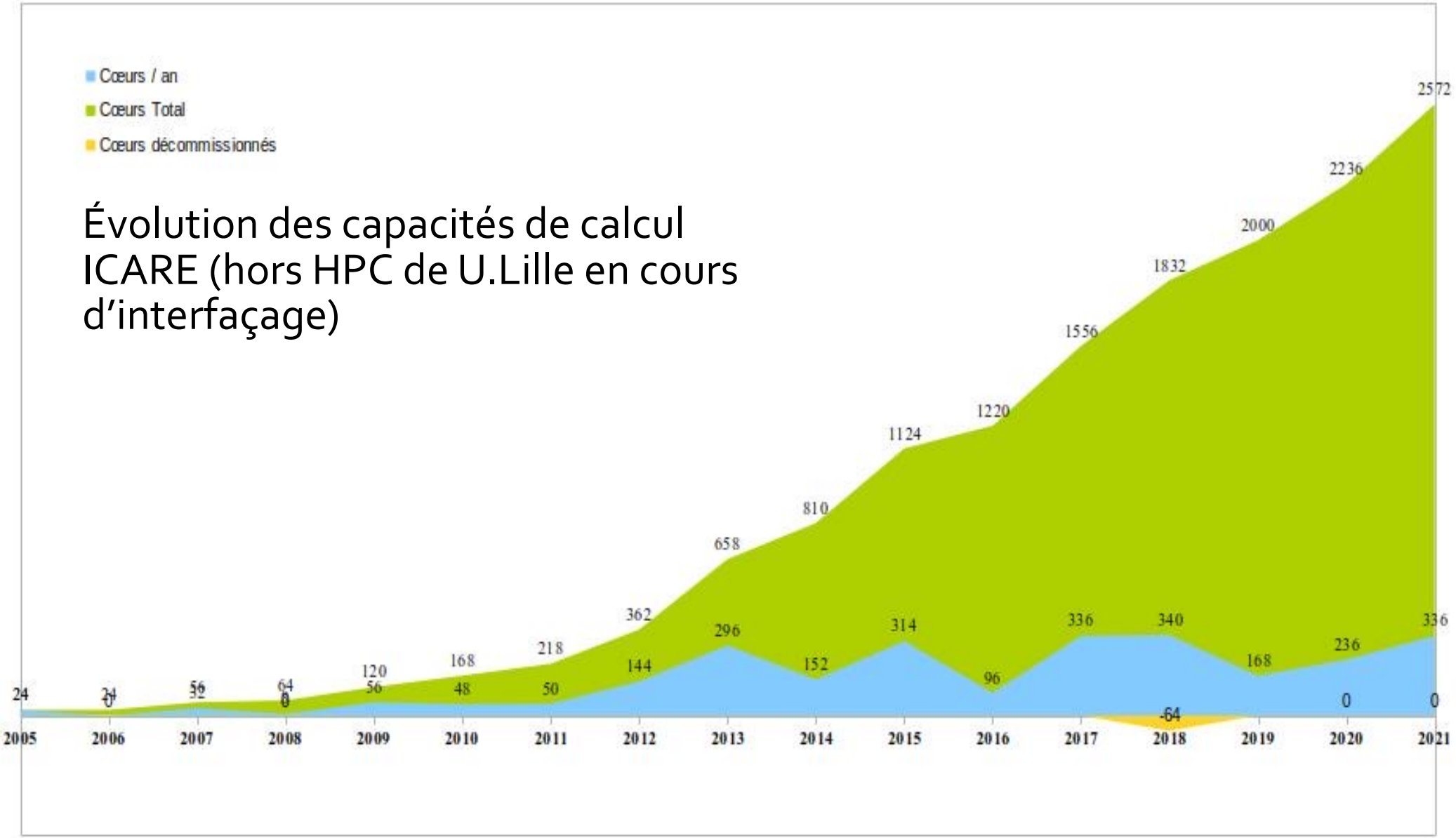
- 4 km de câbles cuivre et 4 km de fibre
- consommation moyenne de 28 000 kWh / mois

Couloir froid 18 racks de 47 U hébergé au sein du Data Center Régional ESR - U.Lille – 47,5 m²

BIG DATA : Volume

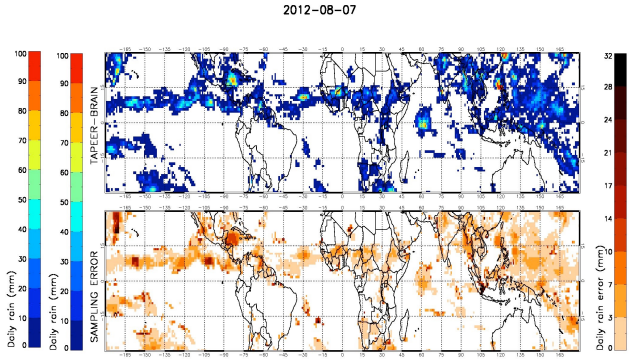


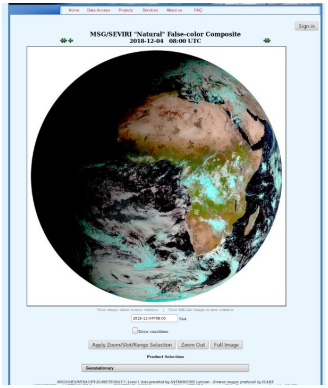
BIG DATA : Vitesse

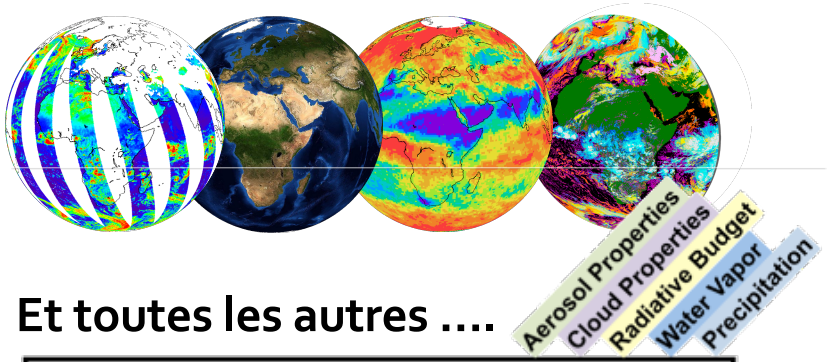


BIG DATA : Variété

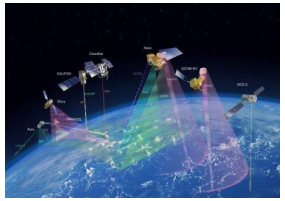
- **Megha-Tropiques**

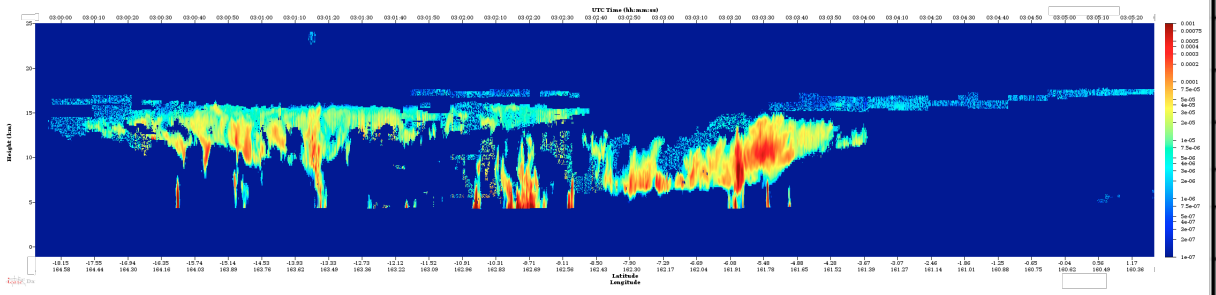




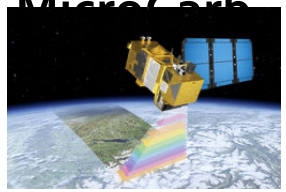


- **CALIPSO / IIR**

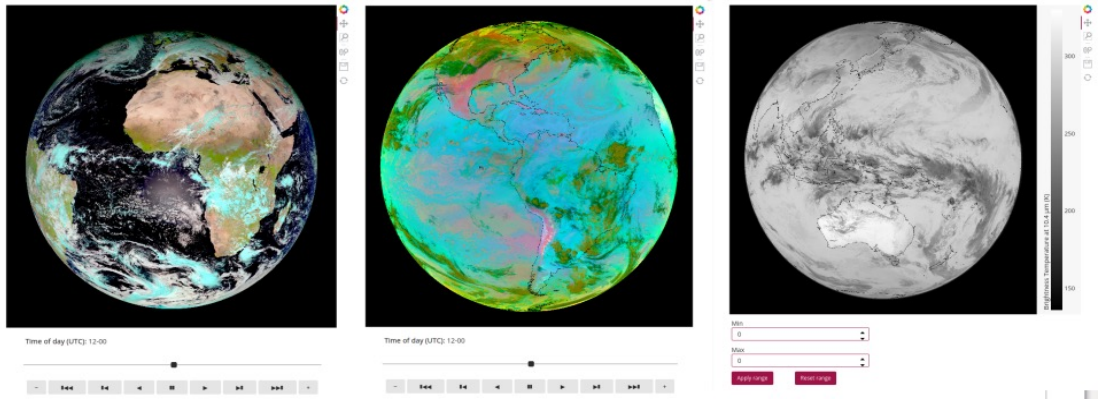




- **MicroCarb**



- **Données GEO**



Et toutes les autres ....

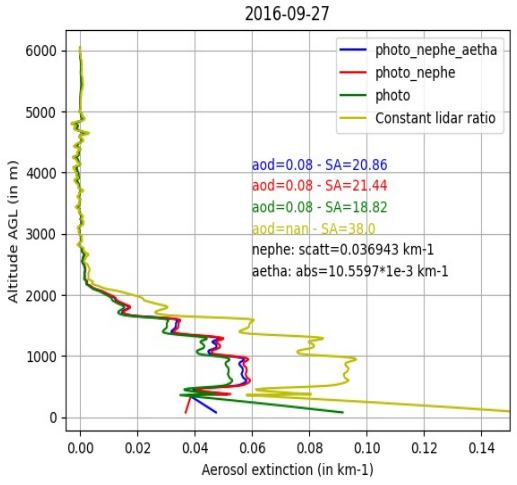
	Aerosol Properties	Cloud Properties	Radiative Budget	Water Vapor	Precipitation
Satellite Observations					
PARASOL/POLDER-3, POLDER-1, POLDER-2	X	X	X	X	
Terra/Aqua/MODIS, Aqua/AIRS, MSG/SEVIRI, Suomi-NPP/VIIRS	X	X		X	
CALIPSO/CALIOP/IIR/WFC, Terra/MISR, Aura/OMI, Envisat/ATSR/AATSR	X	X			
Envisat/MERIS	X				
CloudSat/CPR		X			
METEOSAT-7, GOES-E, GOES-W, GOES-SA, MTSAT, HIMAWARI		X		X	
Megha-Tropiques/SAPHIR, NOAA/AMSU-A , METOP/AMSU-A, NOAA/AMSU-B, METOP/AMSU-B/MHS, Aqua/HSB, NOAA/HIRS, METOP/HIRS, Suomi-NPP/ATMS				X	
Megha-Tropiques/ScaRaB, CERES			X		
Megha-Tropiques/MADRAS, TRMM/TMI, DMSP/SSMI, DMSP/SSMI/S, Aqua/AMSR-E					X
Ground-Based Observations					
AERONET, IAOOS	X				
ACTRIS, ORAURE	X	X			
ROSEA				X	X
Model Analyses and Forecasts					
MACC, CAMS	X		X		
ECMWF, GMAO, MERRA		X		X	X

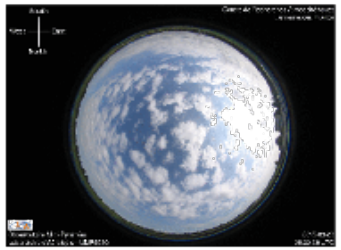
BIG DATA : Variété

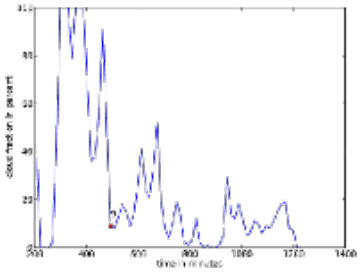
Flux et base de données sols

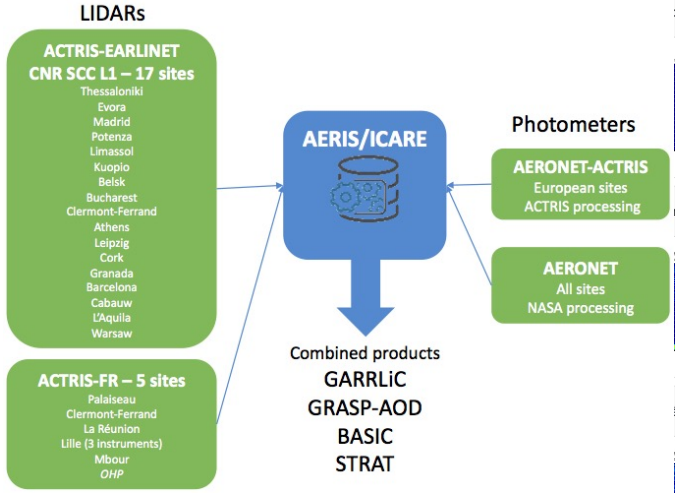
Statut : Chiffres 2021 (2020)

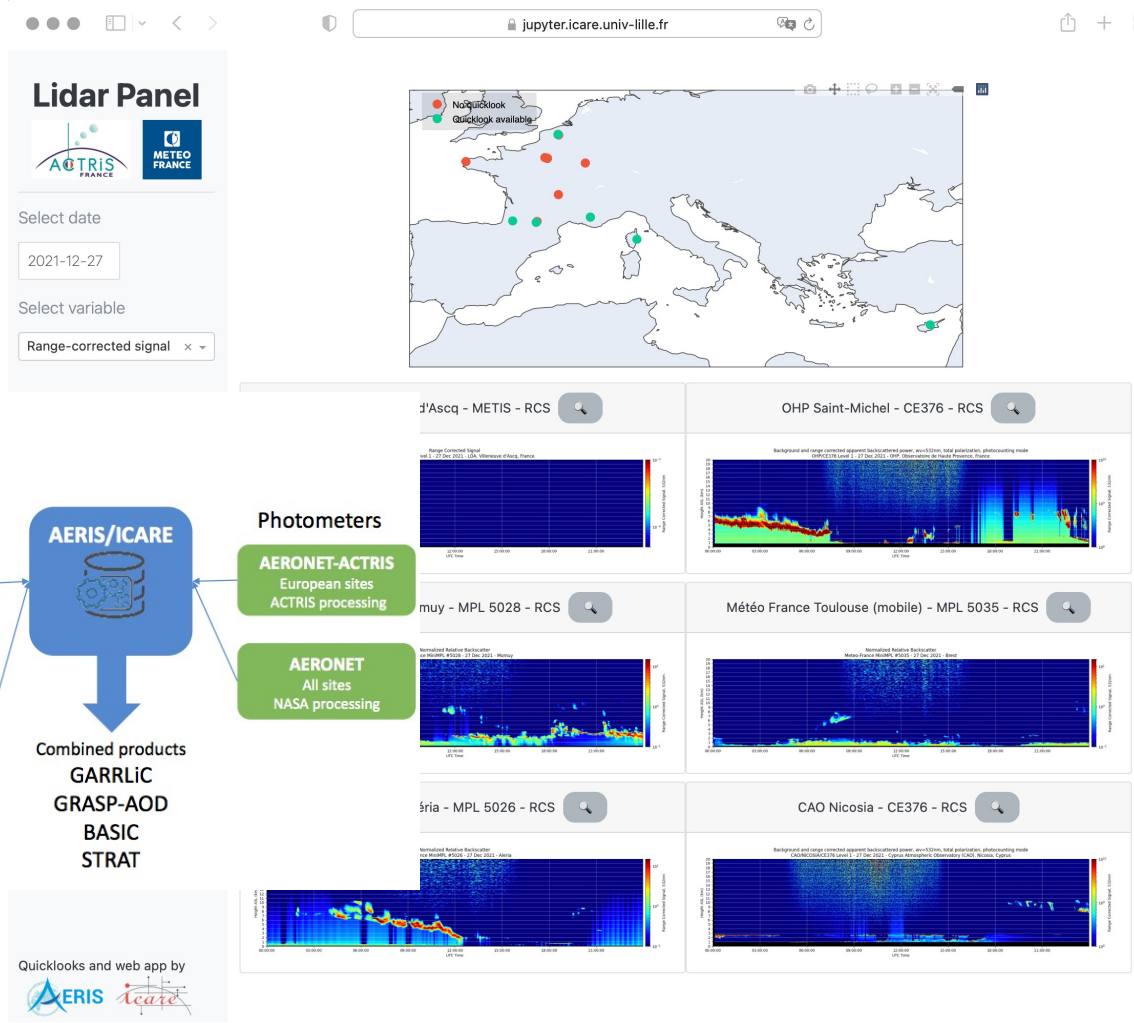
- Sources : 24 (+2 : GEODAIR et Lidar Chypre)
- Sites : 39 (31)
- Instruments : 174 (158)
- Produits : 480 (448)
- Fichiers : ~46 M ~(+7 M / 2020)
- Volume : 9,8 To (+1,4 To / 2020)





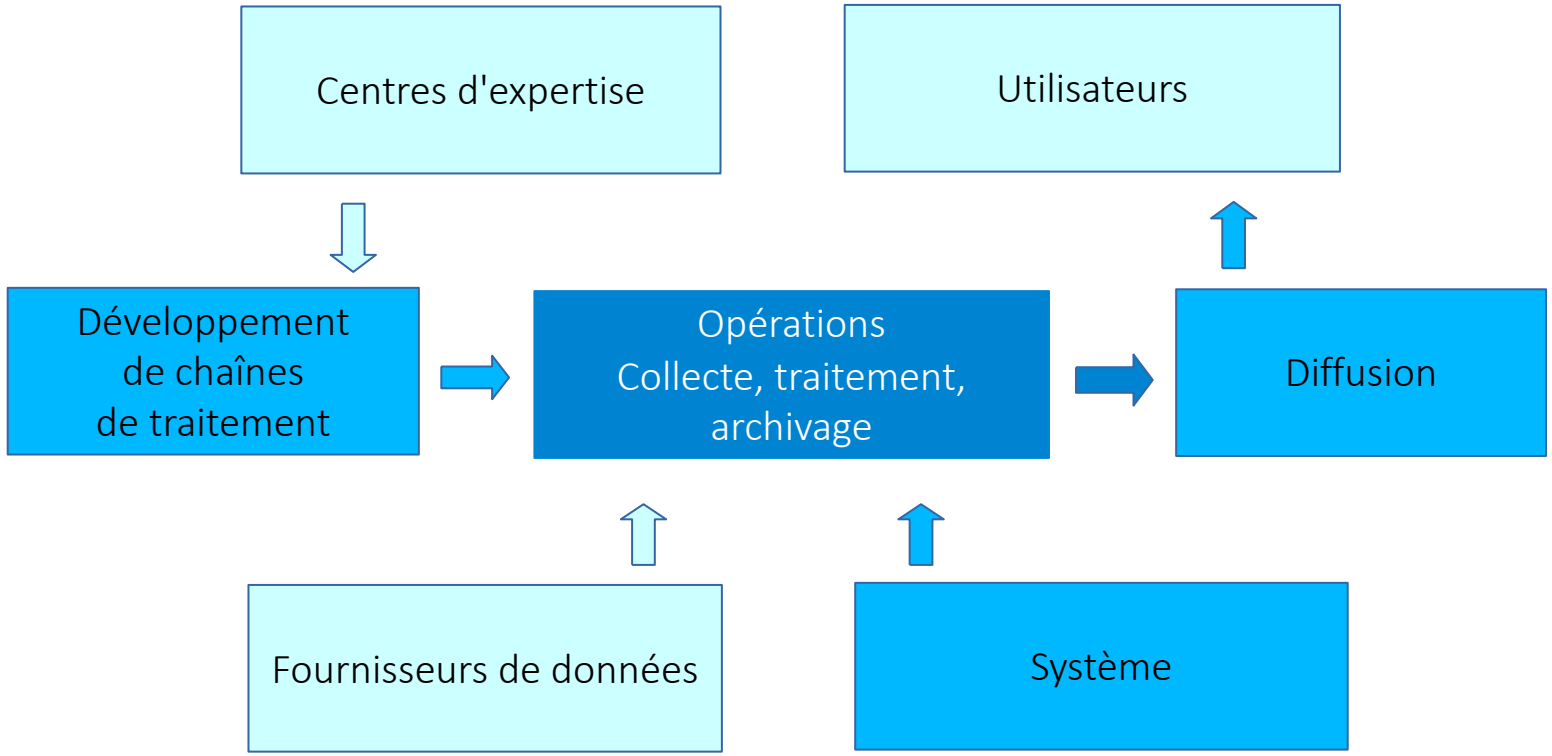






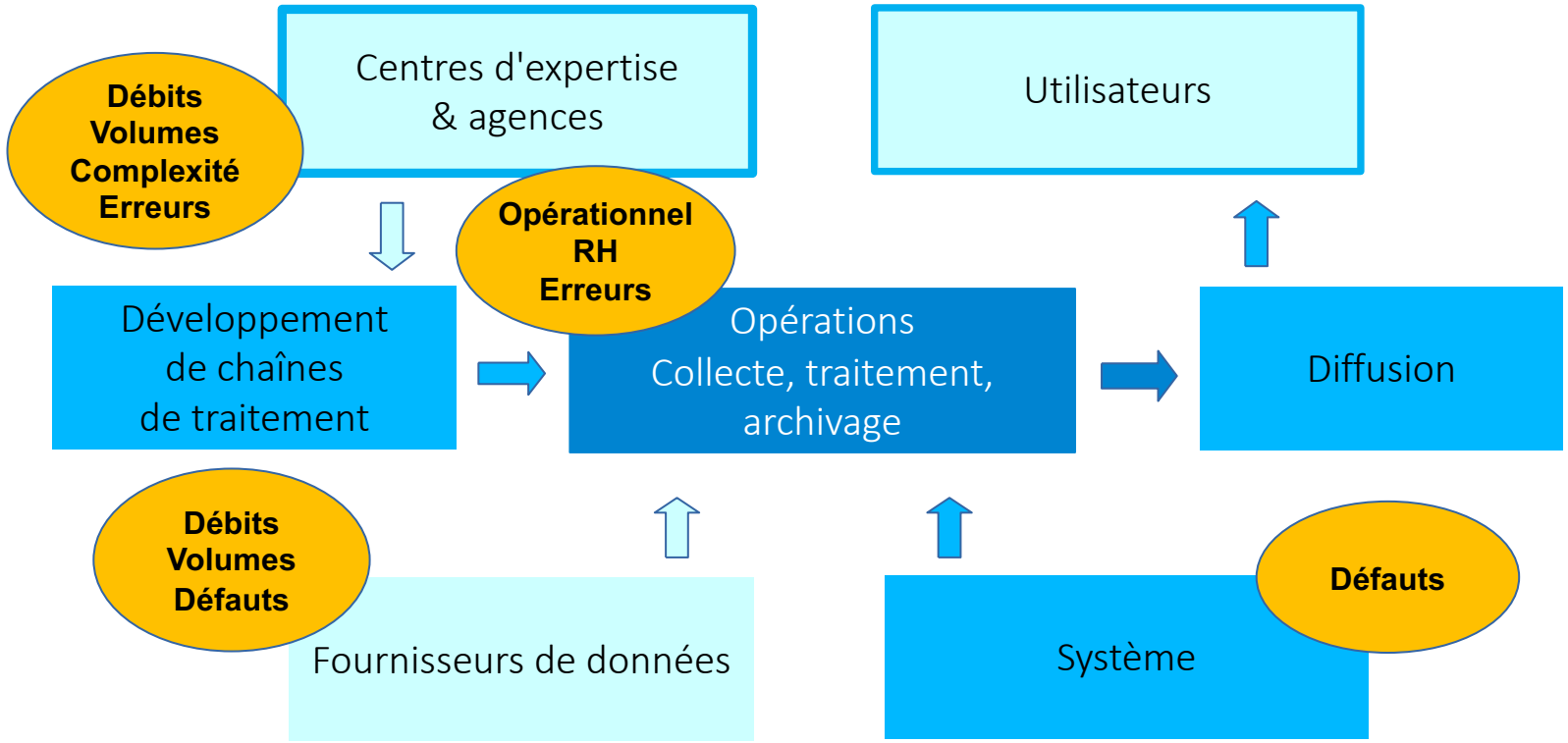
OPERATIONS: définition

Ensemble des actions de développement et d'exploitation de la couche logicielle qui permet la **collecte**, le **traitement** et l'**archivage massifs** des données ICARE.



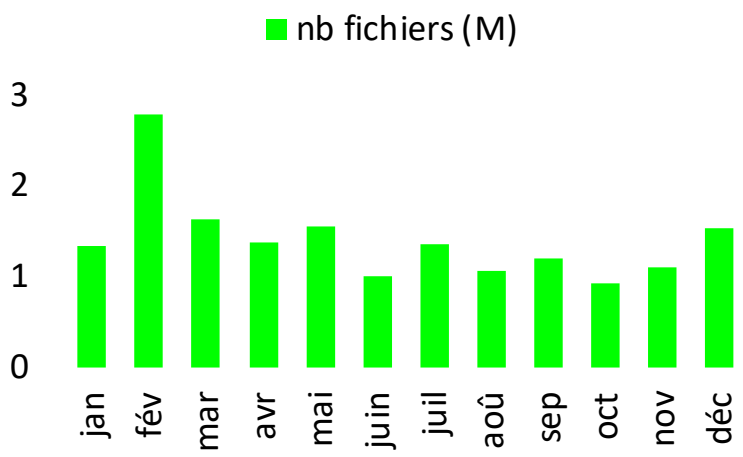
OPERATIONS: les enjeux

Volumes + débits + complexité + RH limitées → outils de gestion → développements

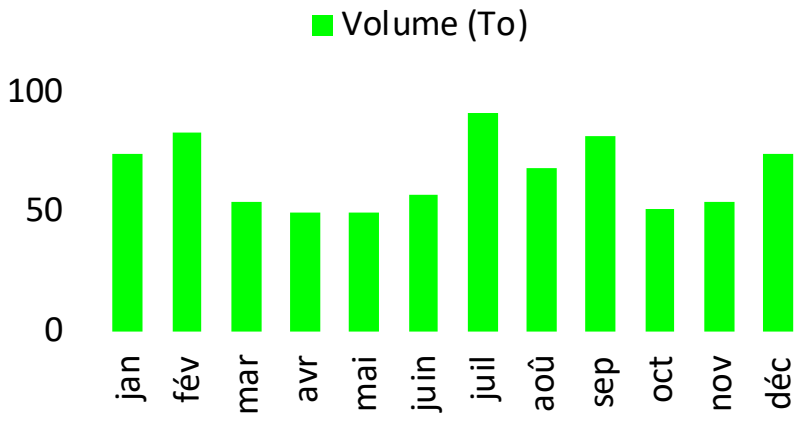


Opérations: Archivage

Fichiers archivés:
Total : 160 M
2021 : 17 M

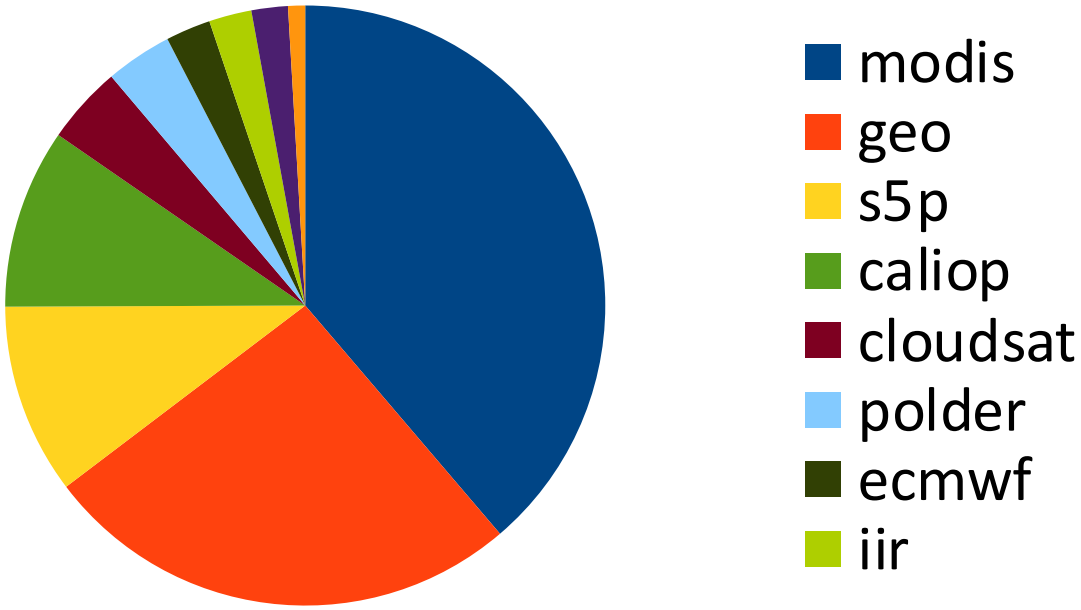


Volume archivé:
Total : 4,4 Po
2021: 0,8 Po



Significatif en 2021 :

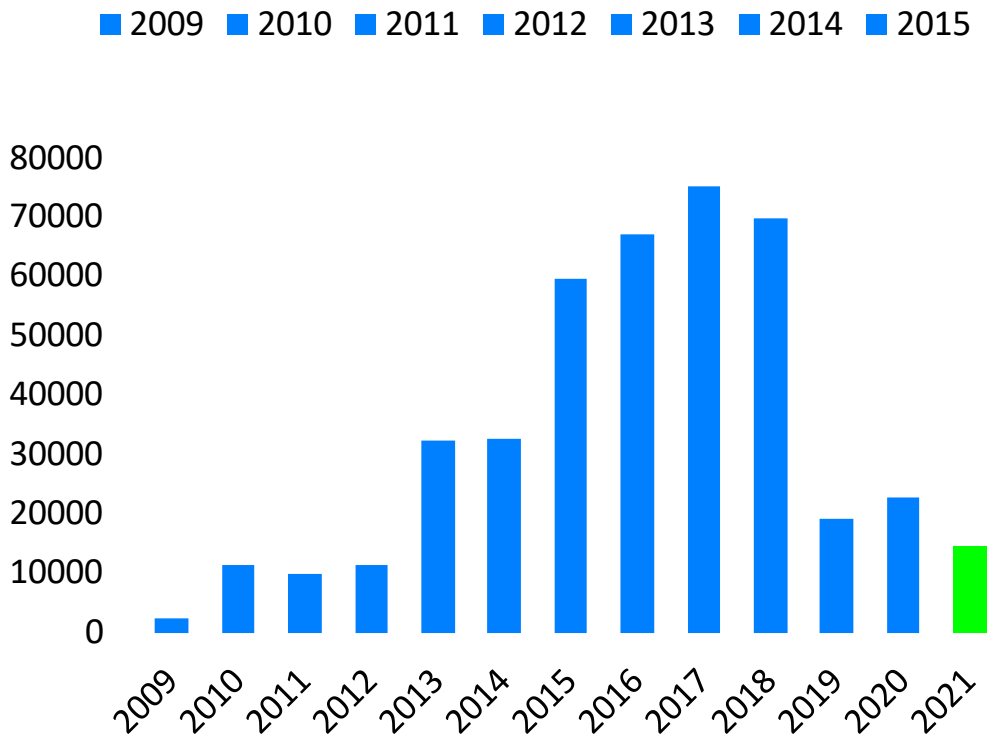
- MODIS C6: complétude de l'archive Aqua/Terra
- Données S5P TROPOMI (PHIDIAS / VOLCPLUME)
- Évolutions du GEO-ring



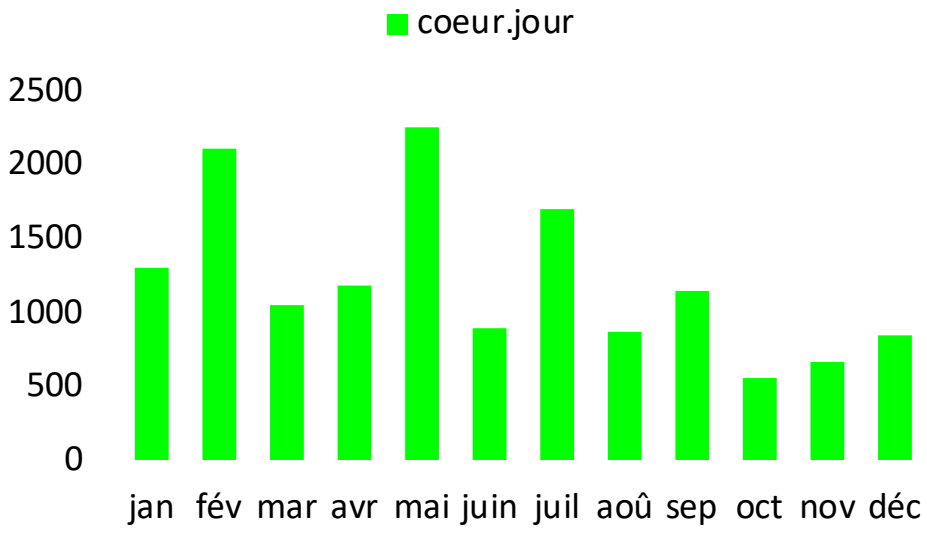
Composition archive

Opérations: Production

CPU depuis 2009 : 429 K coeur.jour



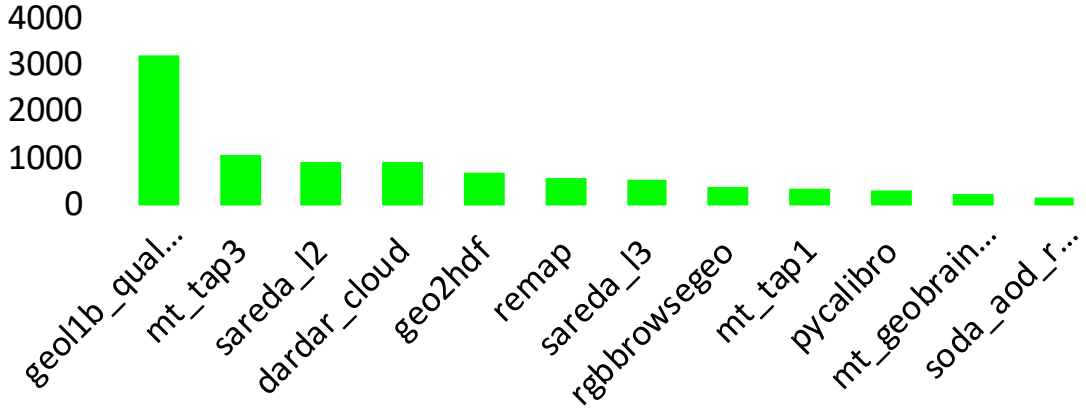
CPU 2021 : 14,5 K coeur.jour



38 chaînes actives (total 90)

amsu	1
calipso	4
geo	14
iir	2
modis	2
mt	15

coeur.jour par chaîne



Exemple de GRASP : Bilan volume / conso CPU

year	core days	Process time on 200 cores
2014	10586.8	53
2015	9958.0	50
2016	8629.6	43
2017	52900.4	264
	82074.8	**410**

product ID	Volume (TB)	number of files
grasp_l2_data	14.5	8312834
parasol_grasp_d3_cyl_proj_6480x3240_data	2.5	6122
parasol_grasp_d3_sin_proj_6480x3240_data	2.5	6122
grasp_loa_l2_data	1.4	17458
grasp_l2_dev_data	0.3	588115
grasp_l2_aux_data	0.0	12278
	21.3	**8942929**

Vitesse de traitement archive POLDER3/PARASOL : 1 mois de calcul / jour / core

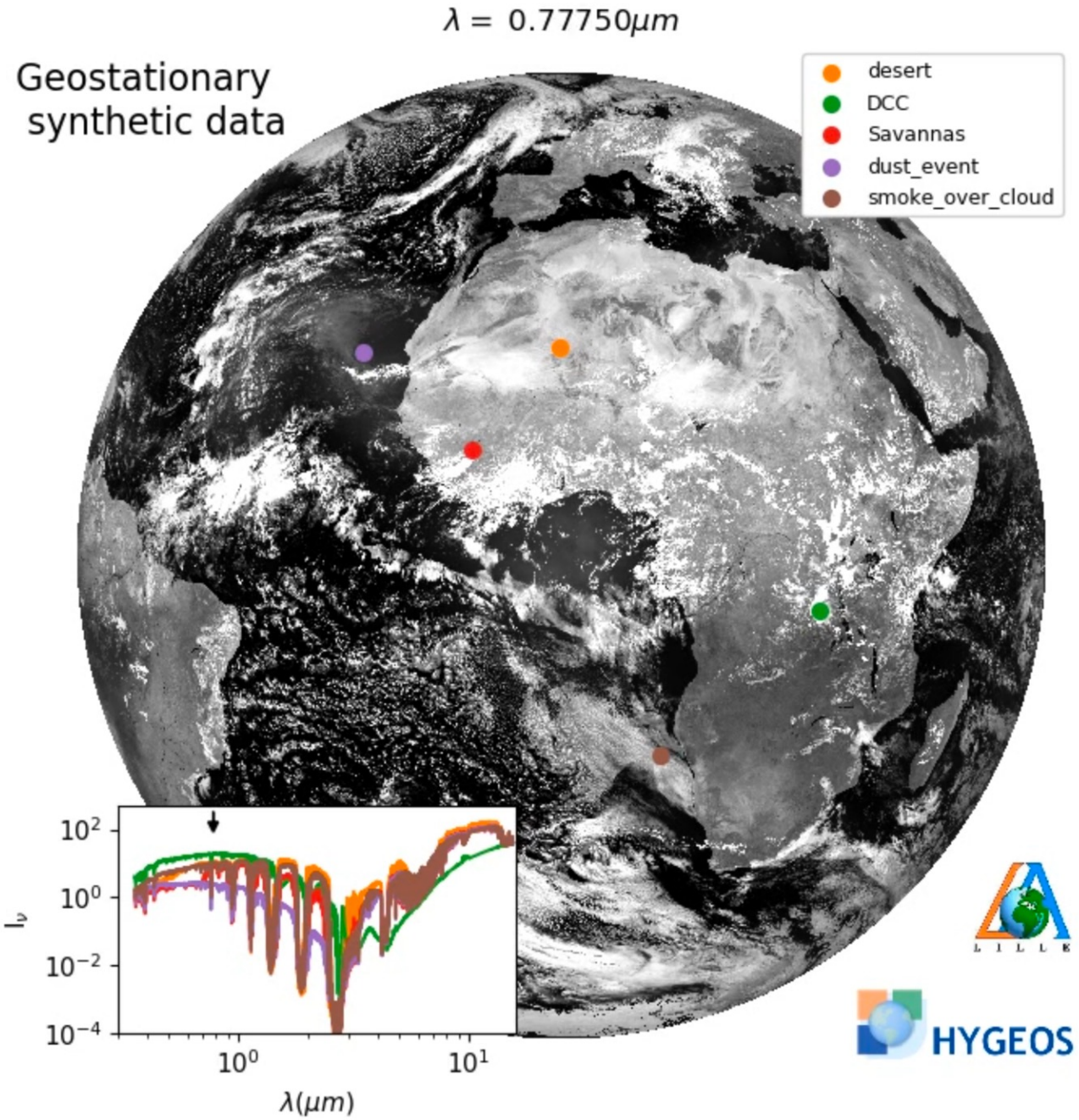
Archive de 8 ans : 365 x 8 x 30 = 87 600 jours de calcul sur 1 core (438 jours sur 200 core)

Exemple de simulation numérique : GEO-SDS

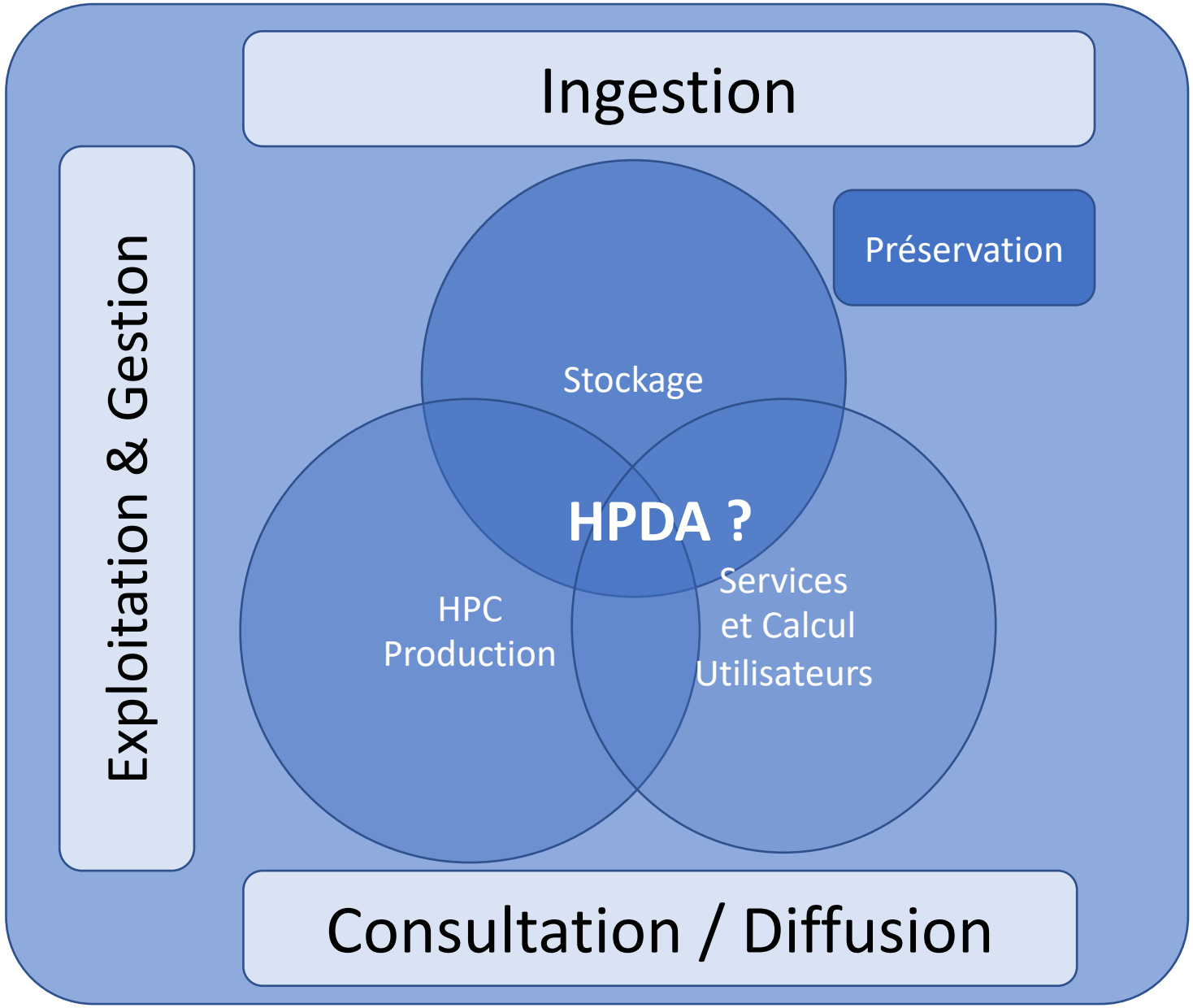
Simulation par HYGEOS/LOA pour EUMETSAT de 2 disques d'observation MSG/SEVIRI (3km au nadir) à haute resolution spectrale (entre 0.5nm et 0.25 cm^{-1}) pour la gamme spectrale de 0.354 to 15.5 µm.

Coût de calcul : 900 000 heures de CPU (HPC U.Lille)

Projet MAP/CLIM pour la mission CO2M -> 8 millions d'heures de calcul



Architecture AERIS/ICARE pour le Big Data



AERIS/ICARE prêt pour le HPDA ?

- Ingestion automatique des données ? ✓
- Logicels de traitement interopérables ? ✓
- Outils “Data Science” (AI, Machine learning, Analytics, …) ✓
- Applications, visualization, valorisation ✓

Développement intense du service JupyterLab et applications Dashboard

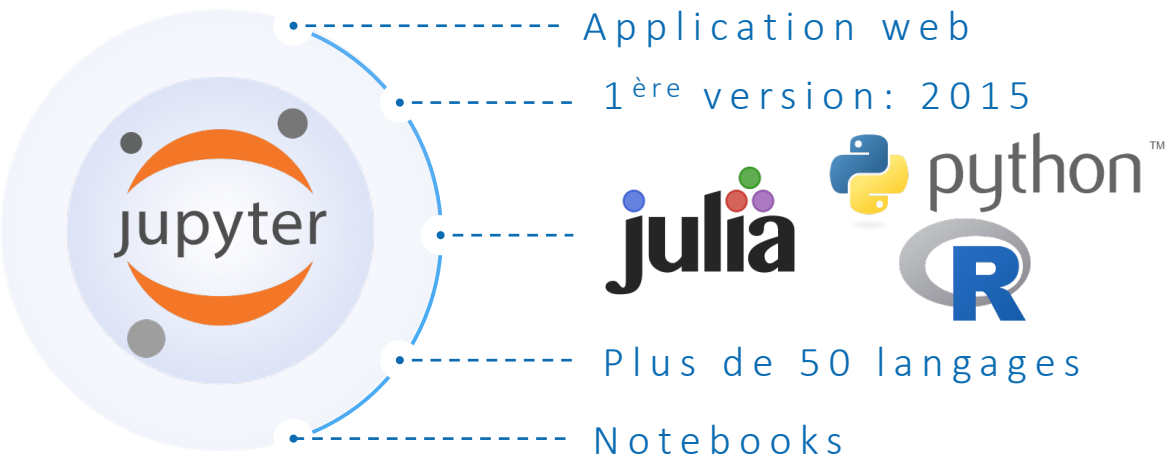
Service Jupyter et mise en place du serveur d'applications

- *Service disponible vers l'extérieur depuis le mois d'octobre 2020 :* **https://jupyter.icare.univ-lille.fr**

Architecture Hardware:

R640, 2x Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz 16 Core
384 Go RAM
4,4 To utiles

- Couplage à prévoir avec d'autres services (cluster de calcul slurm, services web etc)

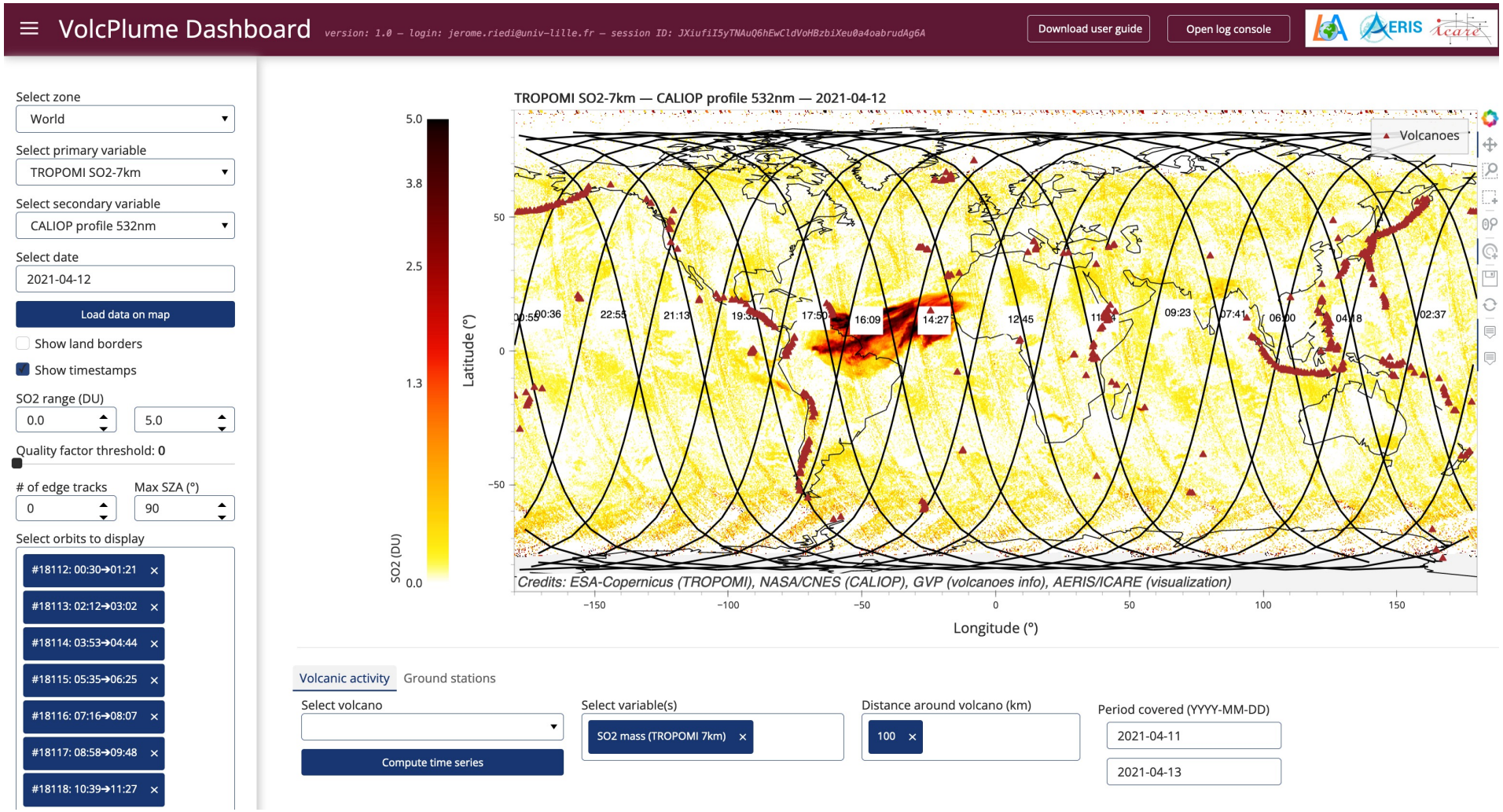


Objectif : Mise en place d'une infrastructure virtuelle de services pour l'analyse et l'exploration de données atmosphériques à l'aide de JupyterLab

En 2022 : Déploiement du serveur dédié aux applicatifs

Plateforme VolcPlume

Portail web pour le suivi multi-échelle des émissions volcaniques



Plateforme VolcPlume

Portail web pour le suivi multi-échelle des émissions volcaniques

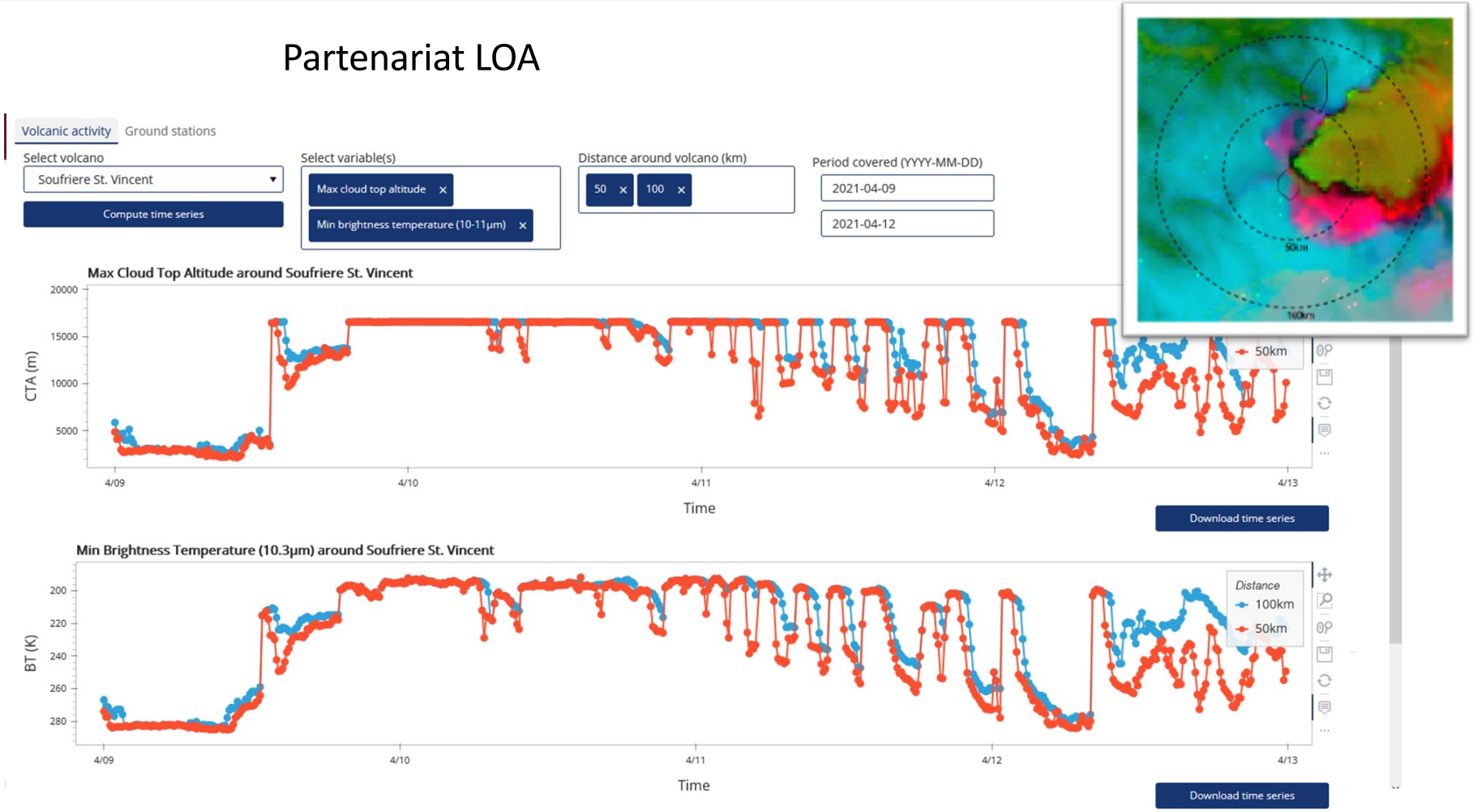
Visualisation pour analyse croisée de données

- Satellites : TROPOMI, IASI, MODIS, POLDER, CALIOP, géostationnaires
- Sol : AERONET, GEODAIR, LIDARs
- Modèle : ERA5

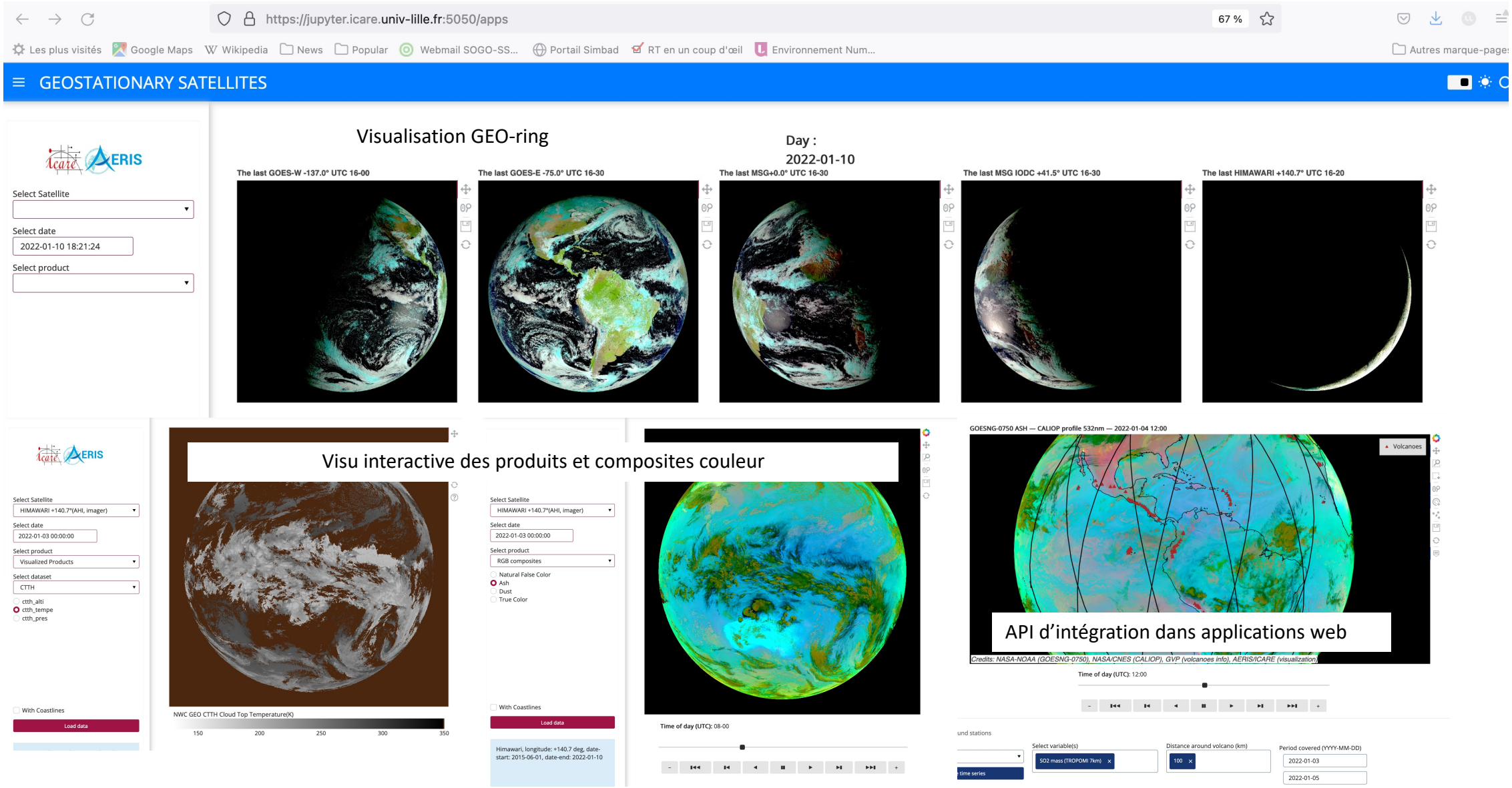
Génération à la demande de figures à valeur ajoutée

Mise en production prévue au 1er trimestre 2022

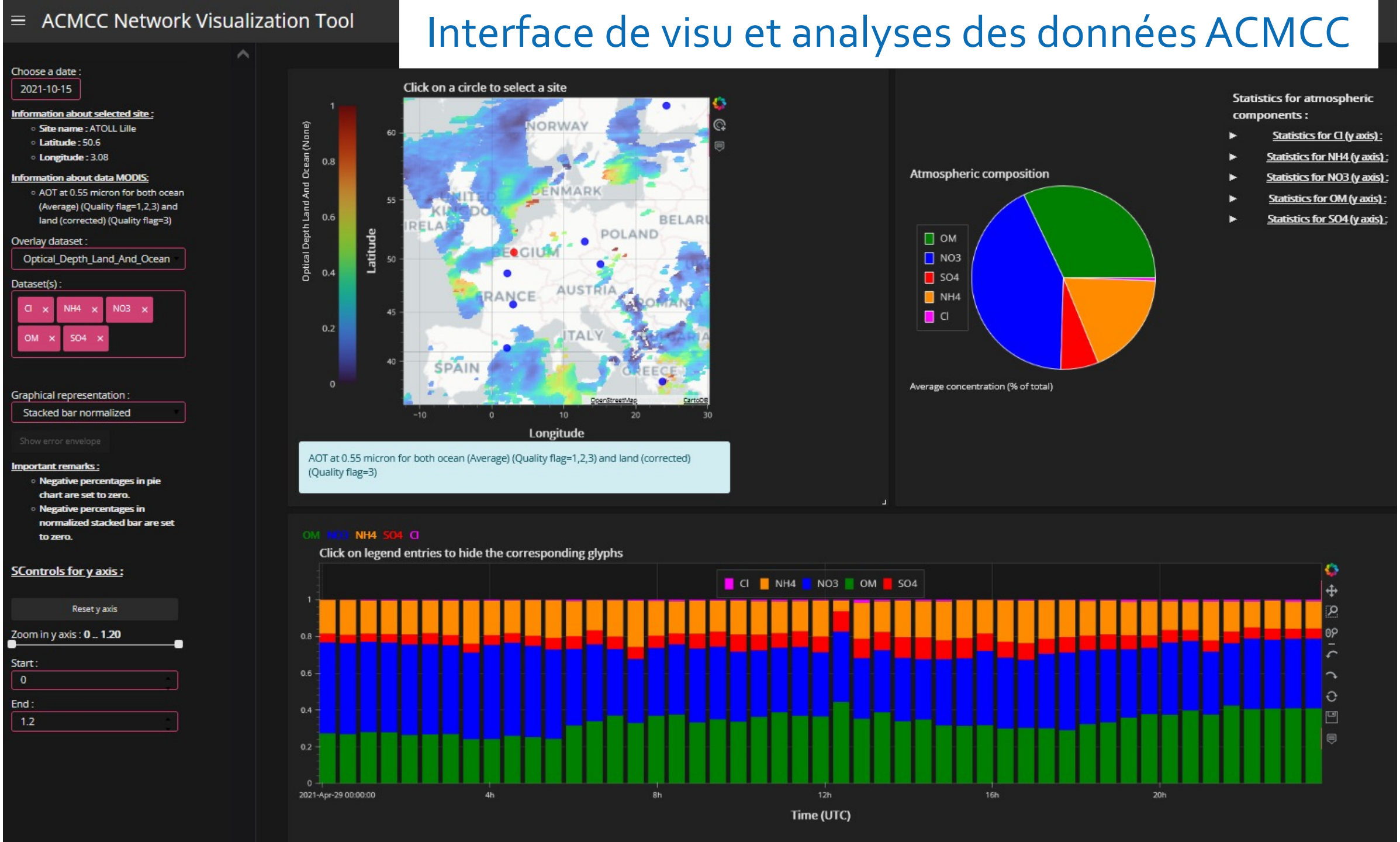
Partenariat LOA



Nouvelle interface de visu et API pour les données et produits GEO



Interface de visu et analyses des données ACMCC



Le challenge du traitement interactif : Quelle est la contribution du Machine learning et de l'IA ?

Big Data

High dimensions of Data (hyperspectral)

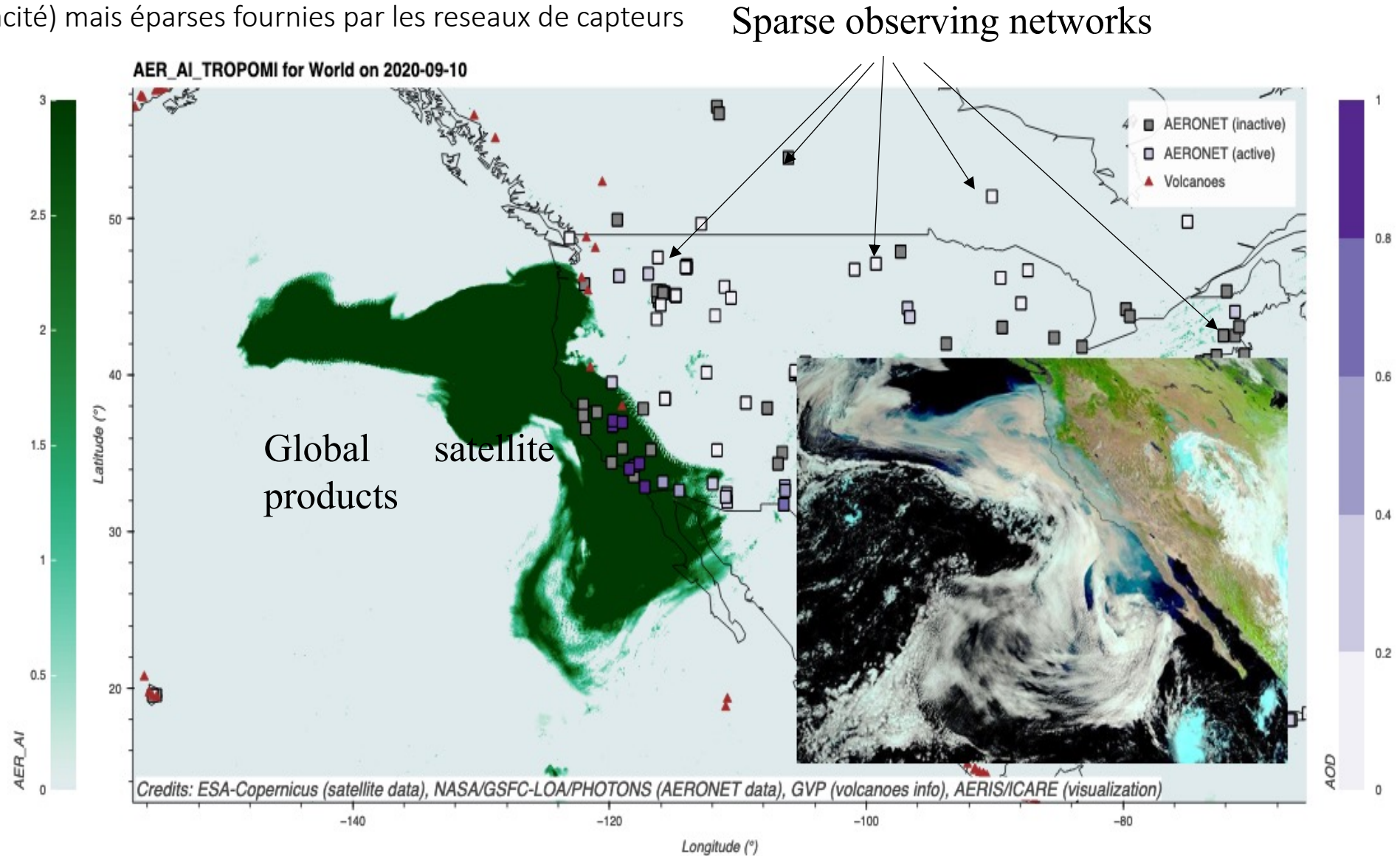
Need of new public datasets and benchmarks

Data without ground truth or with very limited ground truth

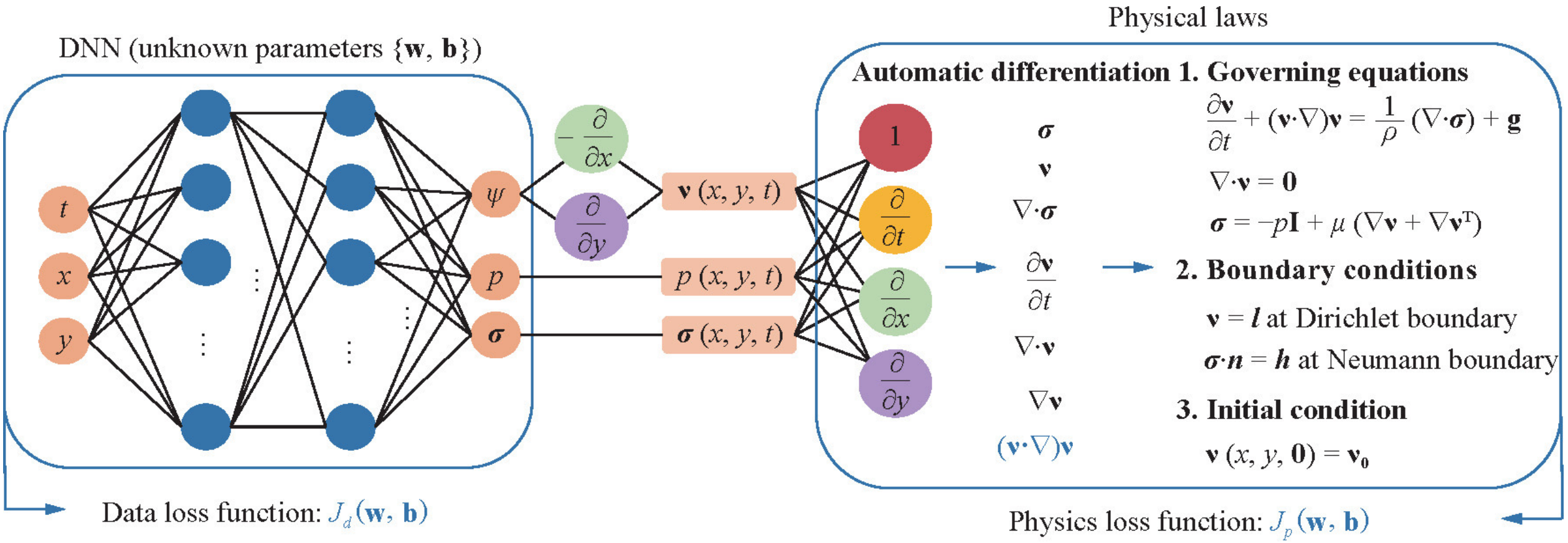
Physics informed machine learning

Apprentissage automatique sur données faiblement indexées

Challenge : relier les données globales issues du spatial (grand volume / validation limitée) aux données localisées à haute valeur (Véracité) mais éparses fournies par les reseaux de capteurs



Apprentissage automatique informé par les modèles physiques



$$\{\mathbf{w}^*, \mathbf{b}^*\} = \underset{\{\mathbf{w}, \mathbf{b}\}}{\operatorname{argmin}} \{J_d(\mathbf{w}, \mathbf{b}) + \alpha \cdot J_p(\mathbf{w}, \mathbf{b})\}$$

From: Chengping Rao, and al., *Physics-informed deep learning for incompressible laminar flows,* Theoretical and Applied Mechanics Letters, Volume 10, Issue 3, Elsevier, 2020

Conclusions

- Pas besoin de réinventer la roue pour les traitements massifs bien définis : Les données proches du calcul restent la solution la plus efficace
- Gros challenge sur le data analytics interactif : besoin d'optimiser les outils d'analyses et développer les moyens de traitement à la demande
- Rôle de l'IA : quels usages ? À quel coût ? Intérêt des travaux actuels sur les architectures bio-inspirées (ex: spikes neural networks)
- Question du coût énergétique
  - Doit guider les développements d'architecture logicielles et matérielles à faible consommation (GPU à développer en fonctions des applis)
  - Besoin d'information (a minima) des usagers "consommateurs" de kWh