Interoperable Machine Learning for Earth Observation and Climate in Federated Cyberinfrastructures

Tom Landry, M. Sc.
Manager, geospatial platforms
Vision and Imaging Team, CRIM

Computer Research Institute of Montreal

CRIM is a not-for-profit applied research center.

- In operation for more than **30 years**
- **56 employees** (15 Ph.D., 16 Masters)
- **80 to 100 projects**, ~50 papers a year

With financial support from:

Économie et Innovation Québec

Introduction

How can AI & Machine Learning Help?

- Earth Observation data: *in situ (sensors), satellite (optical, radar)*
- Earth Systems Science: *health of Earth and its “spheres”, impact of humans*
- Climate and weather: *climate model outputs, weather forecasts, reanalysis*
- GeoInt: *what’s happening now? how to adapt?*

Big Earth Machine Learning challenges

- Training on large datasets require significant processing power (GPUs)
- Huge datasets = remote processing (Cloud)
- How general models are the models? a.k.a overfitting
- Lack of generic geospatial annotation tools, taxonomy management
- No standards for ML models, modernization of geospatial standards needed

Time to unite

Global issues often requires **federated infrastructures**. Enables interstate or international collaborations, sharing of resources, complementarity of efforts. Numerous challenges:

- Heterogeneous architectures (HPC, Cloud)
- Interoperability
- Quoting and billing
- Application and data discoverability
- Cybersecurity

Earth System Grid Federation

Develops, deploys and maintains software infrastructure for management, dissemination and analysis of climate model output and observational data. Serves model simulations, satellite observations, reanalysis products.

Primary goal of ESGF is to facilitate advancements in Earth System Science.

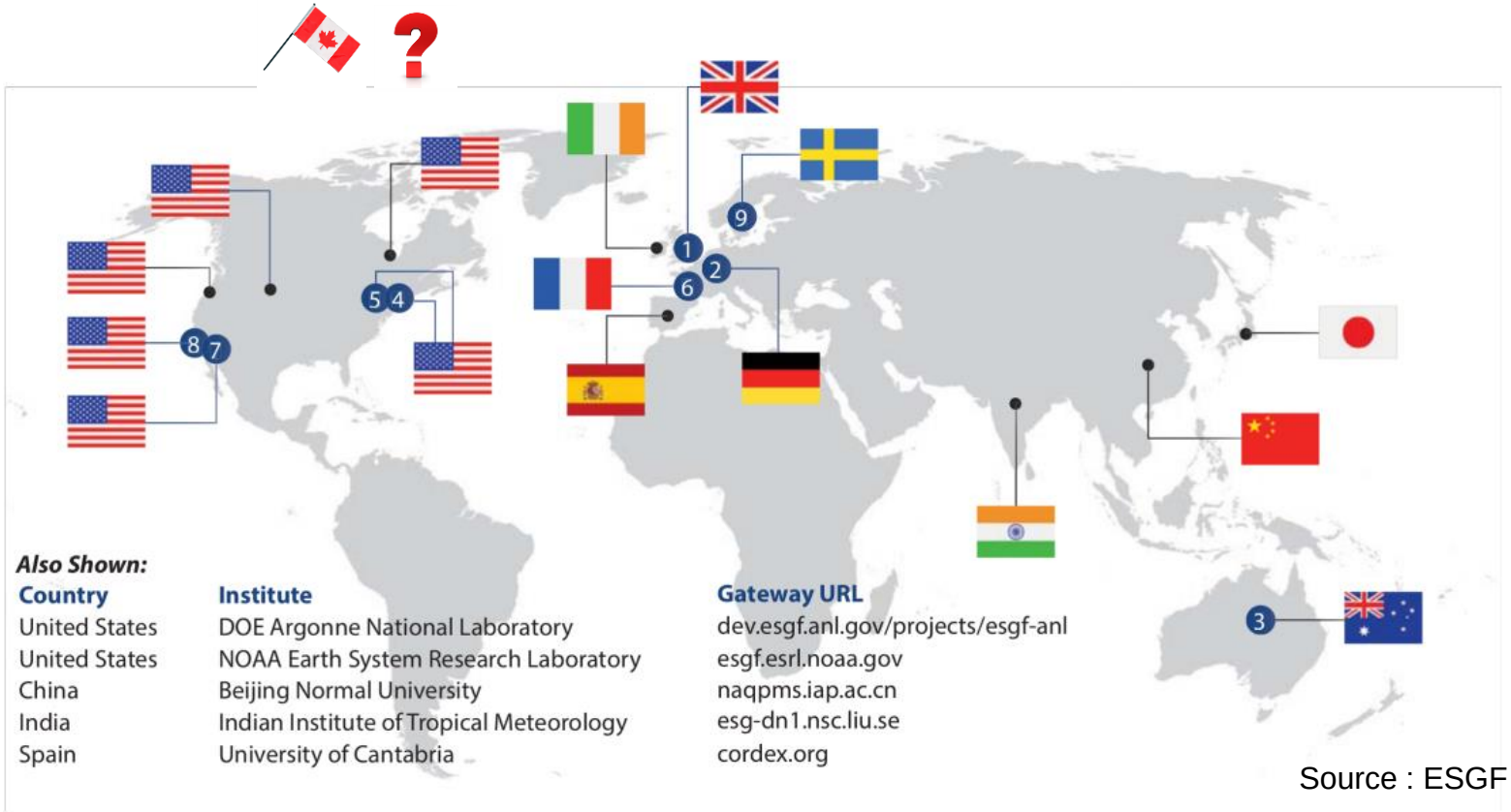
Meetings:

- 9th Face to Face meeting in March 2020, in Europe for the first time.
- Architecture workshop in planning, Nov 2019 in UK.

#C3DIS2019 #Australia #NCI #CSIRO #climate #data #CMIP6 #compute #laboratory #FAIR #Earth

- Preparation for **CMIP6**: how to deal with a multi-petabyte climate data collection
- Helping researchers who work on ‘Understanding the Earth’ to better understand the new FAIR publication requirements
- The Australian Climate Science Data-enhanced Virtual Laboratory
- Increasing scientific productivity through scalable computation and data

Major ESGF Nodes



Source : ESGF

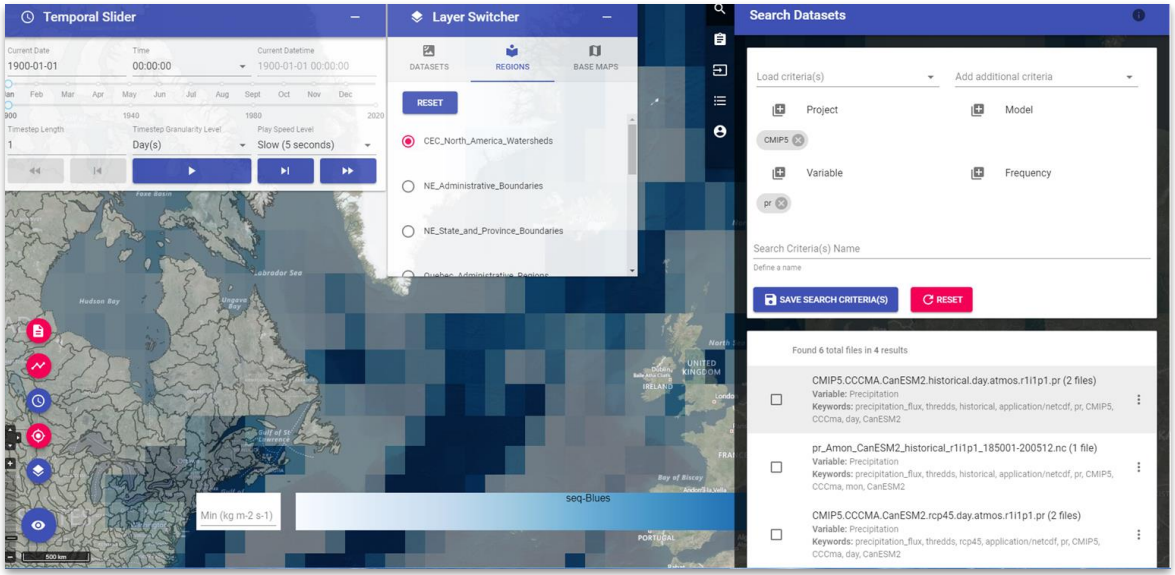
Platform for Analysis and Visualisation of Climate Science

PAVICS

PAVICS is a research platform that streamlines climate scientists *workflows* and provide tools to analyze climate data.

The platform speeds up the analysis of climate data and the creation of **climate scenarios** for impact studies.

Phase 1 completed: 2015-2018



Funded by: canarie

OURANOS

CRIM

Data Analytics for Canadian Climate Services (DACCS)

A Workflow-based Science Gateway (virtual laboratory).
Adds several additional ESGF nodes, with CMIP6 support.
Several new climate services and applications:

- Sea ice from observations and model simulations
- Ensemble diagnostics
- Climatic niches
- CO2 and methane concentrations measurement
- Climate extremes and cyclone tracking
- Coastal vulnerability analysis
- Deep Learning-based Land Cover Mapping
- EO Datacube

Development from **Sept 2019 to 2022**
Maintenance for at least 6 years afterwards

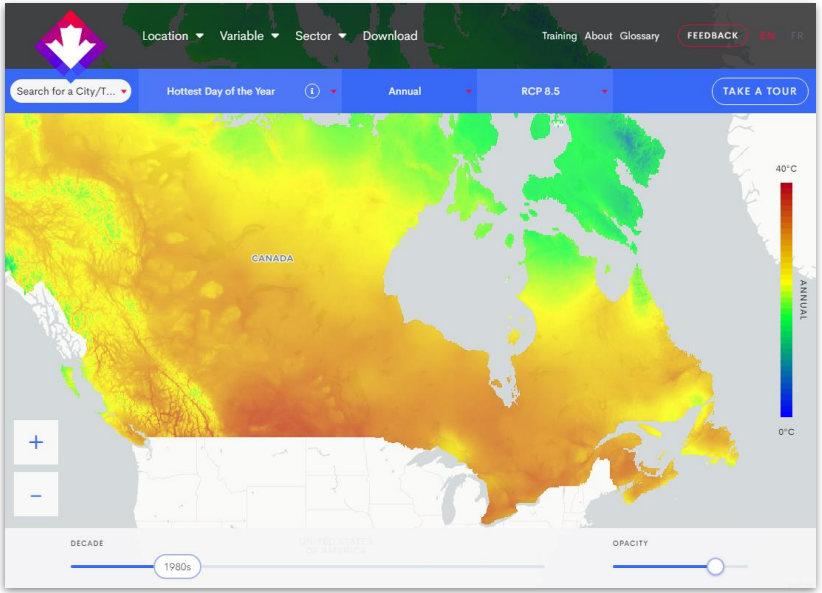
Funded by:

Climate Data Portal

Regional climate services consortia are partners of Canadian Center for Climate Services (CCCS), and new consortia are being fostered.

Collaborative development of a Canadian climate data portal led by CRIM. Launch **June 2019**.

CCCS welcomes collaboration with countries sharing needs for **user-driven** climate services for population, **indigenous** or **remote** communities.



Environment and Climate Change Canada

Environnement et Changement climatique Canada

CRIM

Prairie Climate Centre

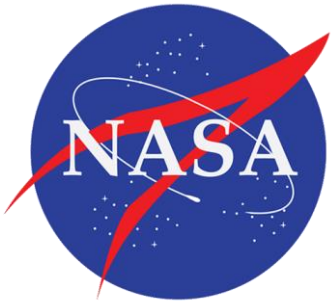
PACIFIC CLIMATE IMPACTS CONSORTIUM

OURANOS

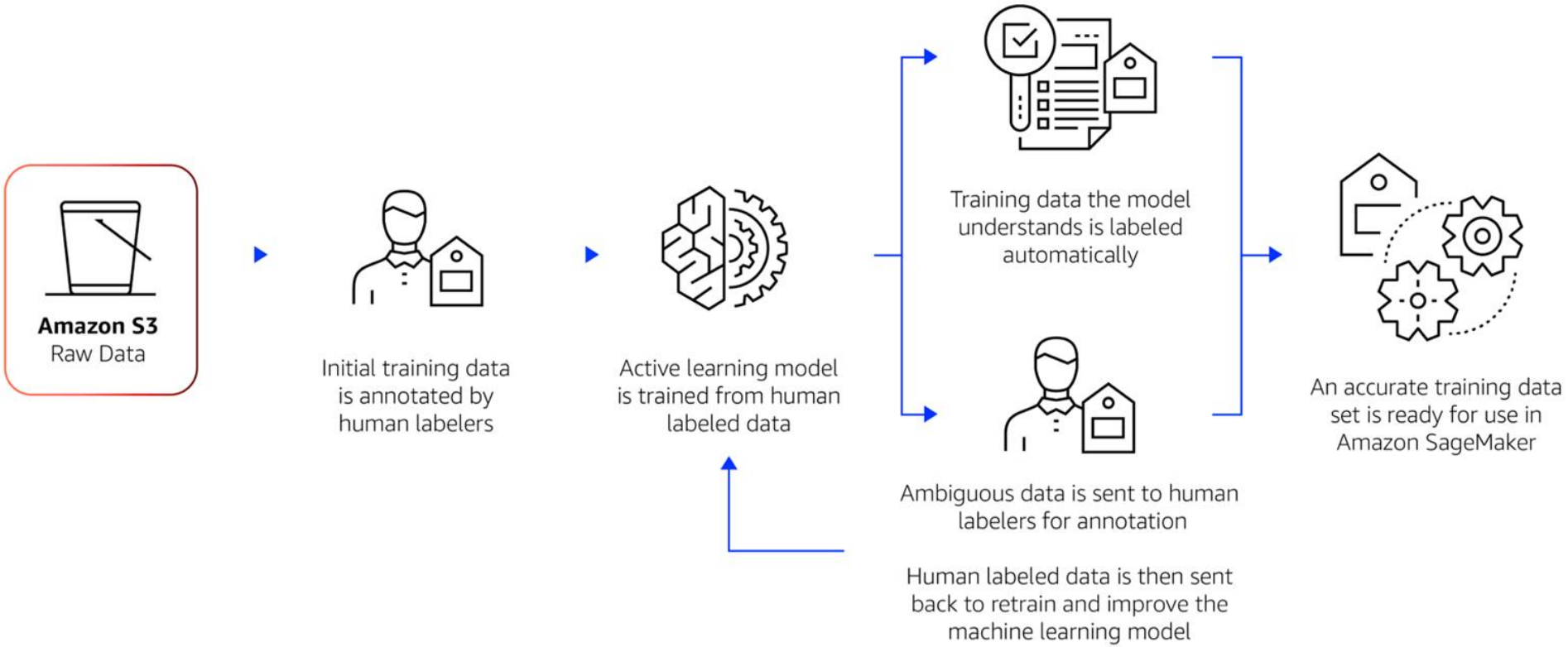
Enabling analytics in the Cloud for Earth Science Data

Deep Learning (DL) adoption for NASA Earth Science data:

- DL has tremendous potential to enable new science and apps
- Challenge is dearth of sufficient labeled training data in Earth science
- Need to systematically create, distribute, and archive training/labeled data
- Manually creating labeled training data is still a bottleneck, expensive
- New strategies to increase training size need to be explored
  - data augmentation
  - transfer learning
  - **active learning**

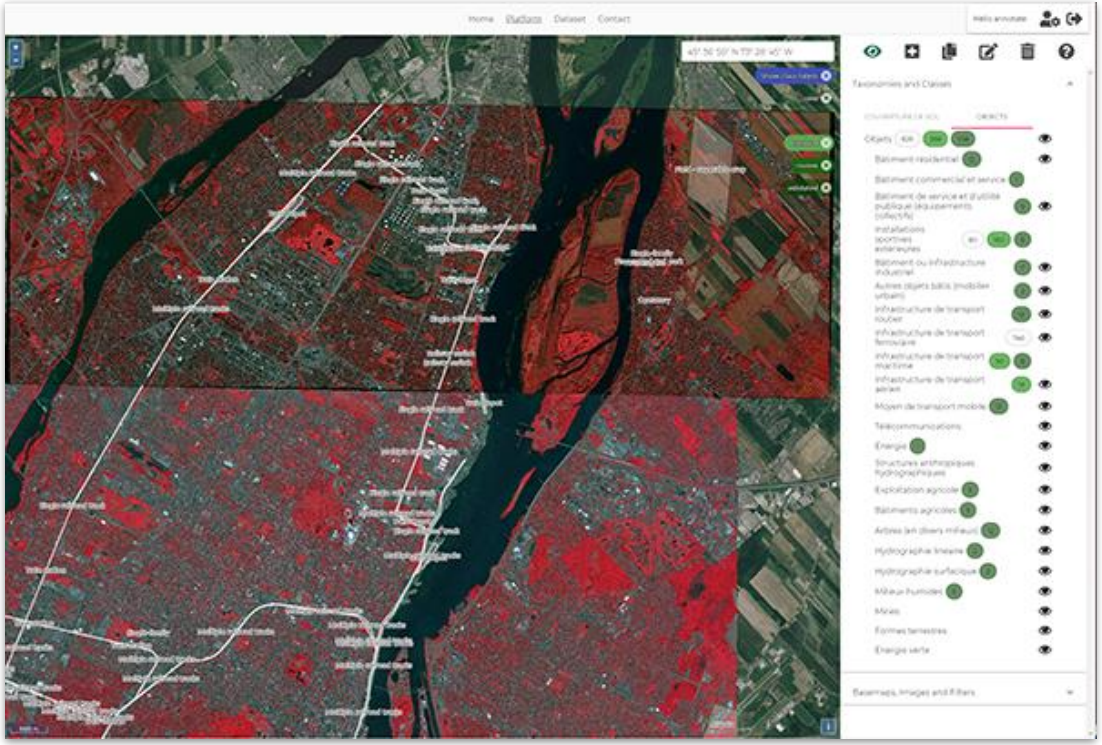


Amazon SageMaker



GeoImageNet

- Stems from PAVICS (2015-2018)
- Annotation of VHR imagery for Deep Learning applications
- Taxonomies: land cover, objects
- In operation late 2019
- Future work
  - **active learning**
  - more data sources



Funded by: canarie

UNIVERSITÉ DE SHERBROOKE

CRIM

effigis GEO SOLUTIONS

Natural Resources Canada | Ressources naturelles Canada

Canada

Deep Learning helper libraries

- Goal: package models and frameworks as **applications,** before deploying to infrastructures.
- Frameworks (PyTorch, TensorFlow, etc.) usually not sufficient. Helper libraries are required.

thelper facilitates model exploration and the development process. It provides:
1) a centralized interface for the control of experiment settings;
2) debugging tools to help understanding a model's behavior.

geo-deep-learning allows use of Convolutional Neural Networks (CNN) with georeferenced data sets. It comprises: 1) data preparation; 2) training & validation, and 3) inference.

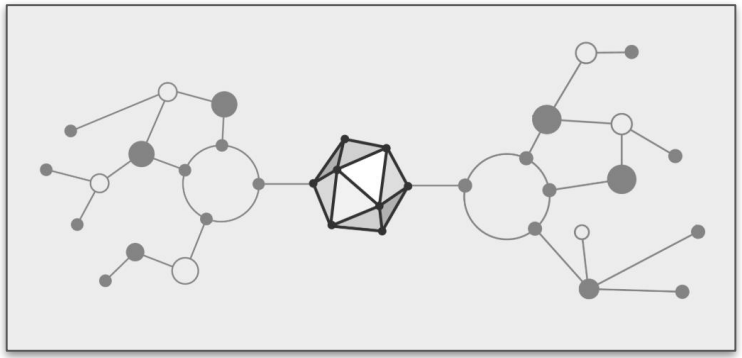
CRIM

Natural Resources Canada | Ressources naturelles Canada

Canada

Open Neural Network Exchange (ONNX)

- There’s a need for greater interoperability of AI tools.
- Support for frameworks, converters, runtimes to avoid getting locked in to one framework or ecosystem.
- ONNX defines:
  - an extensible computation graph model
  - built-in operators and standard data types
- Offers Model Zoos:
  - store, search, exchange trained models



OGC testbed - Machine Learning

Holistic approach to determine best practices of OGC web services for AI/ML.

- Testbed-14:
  - Scenario of data annotation in an EO platform, used by analyst
  - Main WPS operations: Training, retraining and execution
  - DL-based semantic segmentation, transfer learning for flood detectors
- Testbed-15:
  - Various models: arctic services, forestry outputs, land cover, lake-river classification, cloud-free mosaics, etc.

Sponsors Testbed-14

Sponsor Testbed-15

OGC testbed - EO platforms

User-facing Thematic Exploitation Platforms (TEP) relying on Mission Exploitation Platforms (MEP) for data and computing.

- Testbed-13, 14
  - Use of Common Workflow Language (CWL) for application chaining
  - Use of an Execution Management System (EMS) on TEP
  - Use of an Application Deployment and Execution System (ADES) on MEP
  - WPS 2.0 REST interfaces includes quoting, billing
- Testbed-15: Application discovery
- Pilot project “EO Big Data architecture” expected soon

Sponsor Testbed-13,14

Natural Resources Canada | Ressources naturelles Canada

Canada

Sponsor Testbed-13,14,15

esa

Sponsor Testbed-15

NASA

OGC testbed: ESGF compute challenge

Three main objectives:

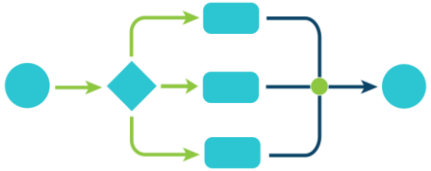
- Test common ESGF climate processes and their compute nodes
- Demonstrate TB-14 TEP architecture for climate processes, applications and workflows
- Advance application packaging, deployment and execution in clouds

Main results:

- Use of CMIP6 data from ESGF and downscaled climate scenarios from Canada portal
- Workflows conducted with **LLNL** AIMS2 , **NASA** EDAS and **CRIM** PAVICS using ESGF Compute Working Team (CWT) API
- Uses an EMS/ADES with WPS 2.0 REST, adds support **pre-deployed** WPS 1.0 and ESGF CWT API

Sponsored by:

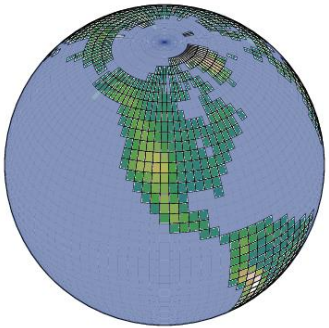
Workflows



- Scientific workflows
  - Enables provenance tracking, transparency, reproducibility
  - Enables multidisciplinary and transdisciplinary work
- Computational workflows
  - Distributes the work over regions, sites, infrastructures
- Operational workflows
  - Allows larger adoption of products, improves and maintains quality
    - mature devOps required
    - integrates with commercial clouds

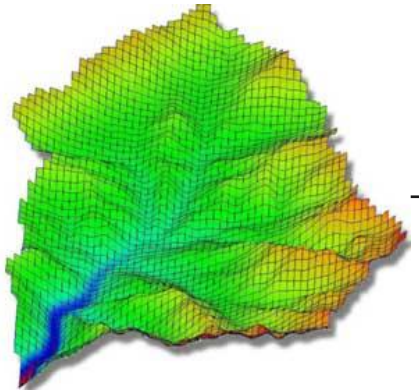
Analysis Pipeline Example

Climate model



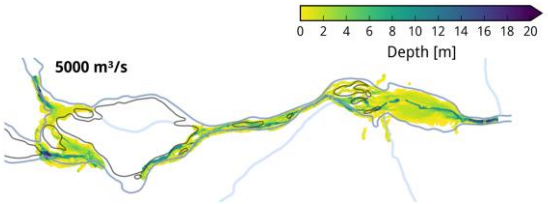
Processing

Hydrological model



Processing

Hydraulic model



Processing

Service / Product

Source: D. Huard, Ouranos, 2018

Enter the ML orchestra

AI as composer: determine what the analyst is asking using NLP, disambiguate and identify specific interests.

AI as conductor: orchestrates federated queries across domain-specific information subsystems and engines.

User queries generate intricate series of activities:

- decompose the query
- identify the applicable data sources
- perform complex analysis via disparate AI-enabled subsystems
- fuse data and query results
- generate series of highest probability responses for the analyst

Source: Sarojak M. and al., An Orchestra of Machine Intelligence, in USGIF State and Future of GeoInt Report 2018

Various federated instruments

- Copernicus ecosystem (ESA)
  - Data, Ressources, Platform and Exploitation layers: Services, TEP, MEP, Sentinels
  - Weaver: an Execution Management System (CRIM)
- Intelligent workflows (Badia R., 2017)
  - dynamically instantiated according to the needs of application objective
  - analytics-as-a-Service that can be used by the workflows
- Security blueprint architecture (AARC)
  - set of software building blocks to implement federated access management solutions
  - mix and match tried and tested components
- Smart and distributed proxying (CloudFlare)
  - content delivery networks
  - DDoS protection and other cybersecurity measures

Conclusion

- ML system, ML model and platform interoperability
  - Deployable apps, or pre-deployed services
- Make applications, ML models and applications FAIR
  - Findable, Accessible, Interoperable, Reusable
- Make it open
  - Open... source, data, standards, innovation
- Play well with commercial offerings
  - Marketplace for applications, services
  - Allow quoting, billing, technology transfer
- Seeking international and national collaborations
  - Advancement of science and research
  - Technology is an enabler for “climate diplomacy”
  - Still a need to “play it by ear”

References

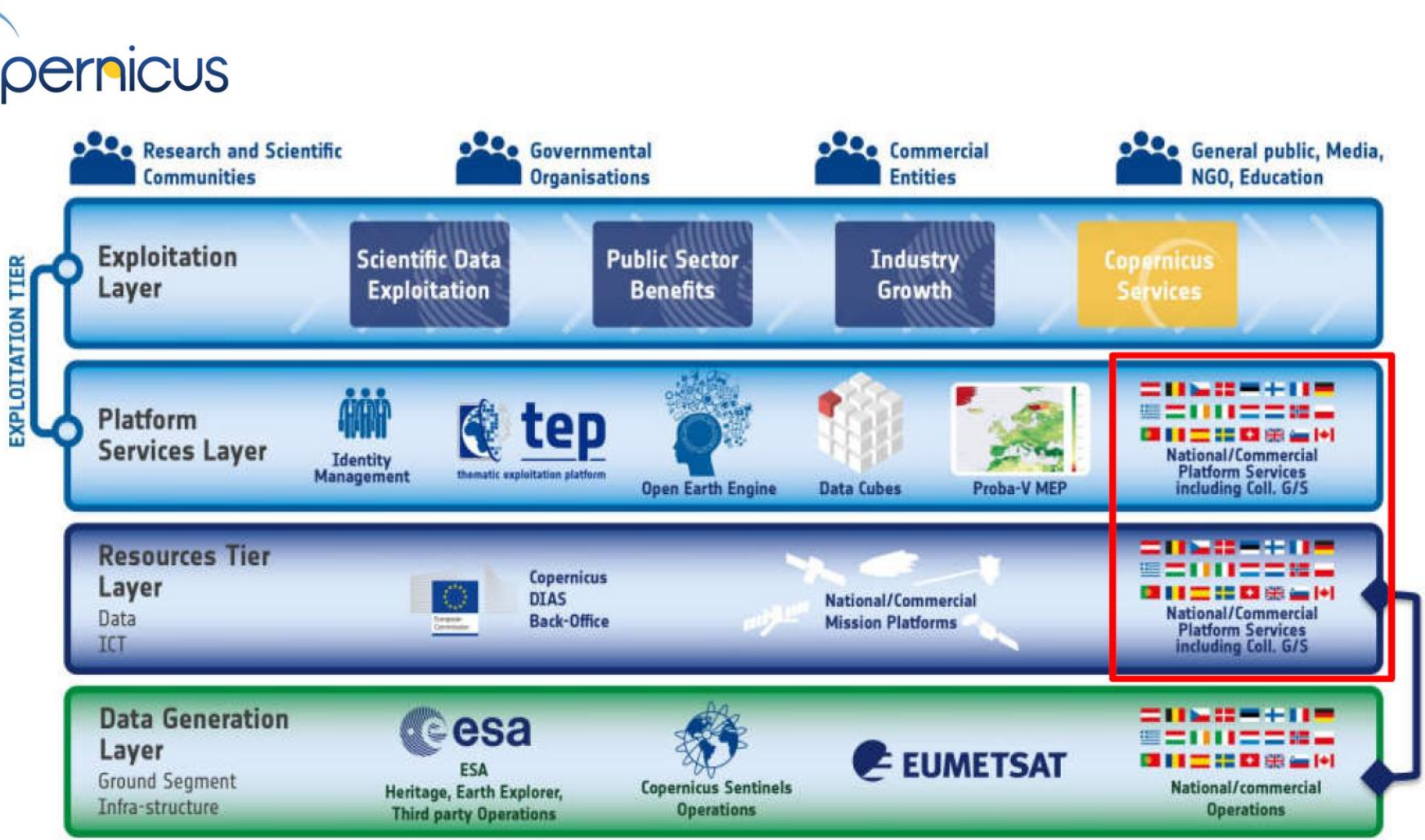
- AARC Blueprint Architecture - AARC website
- Amazon SageMaker Neo - AWS News blog
- Caffe Model Zoo - Github
- CloudFlare - Whitepaper
- Earth System Grid Federation - ESGF Brochure 2017
- GeoImageNet press release - CRIM
- Enabling Analytics in the Cloud for Earth Science Data - Workshop report
- OGC Engineering Reports - OGC Website
- OGC Call For Information - EO Big Data architecture pilot
- Open Neural Network Exchange - Github
- Weaver Execution Management System - GitHub
- 2018 State and Future of GeoInt report - USGIF 2018

Tom Landry, M.Sc.
Manager, geospatial platforms

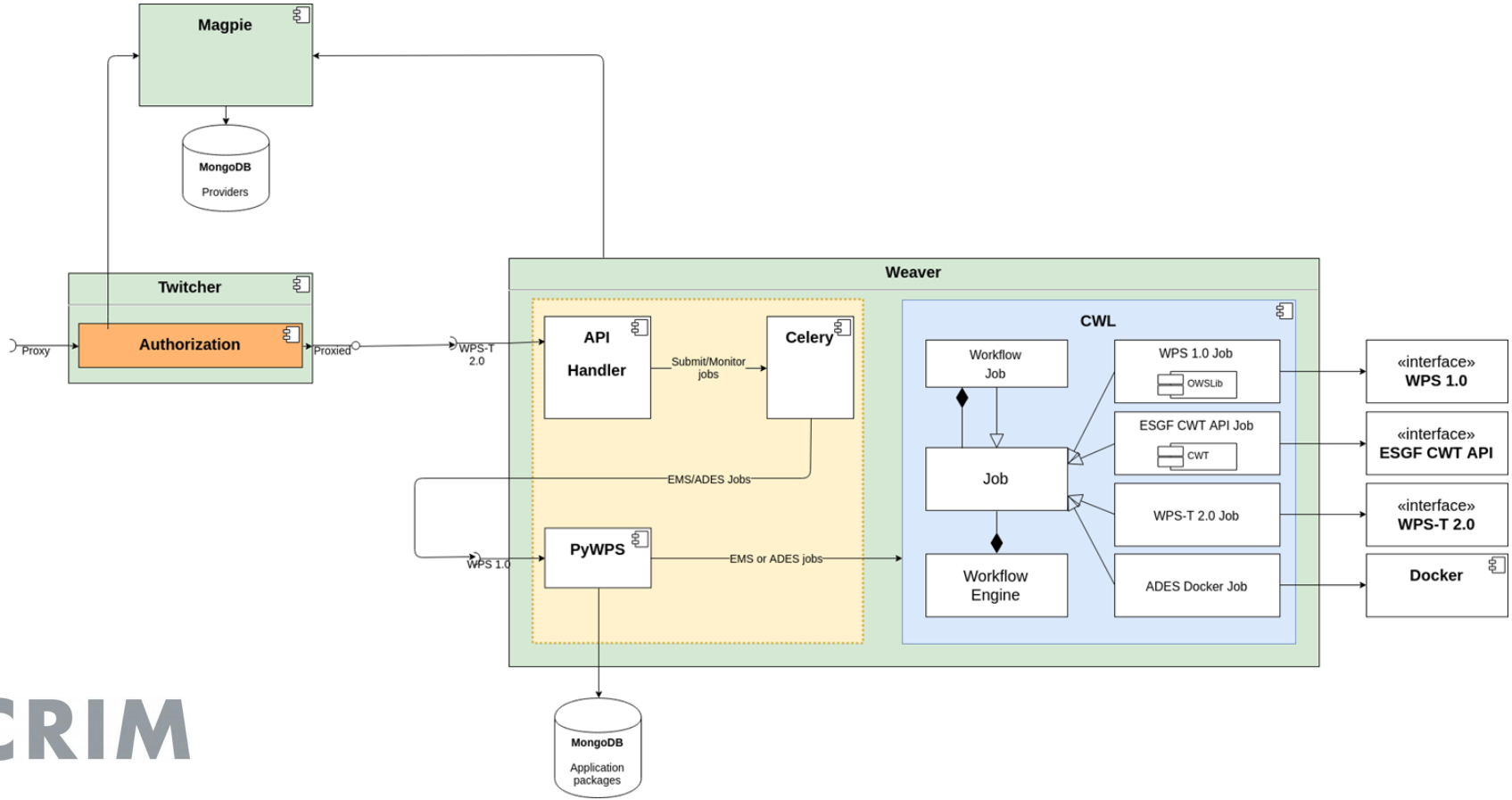
tom.landry@crim.ca
www.crim.ca

101 - 405, avenue Ogilvy, Montréal (Québec) H3N 1M3 514 840-1234 / 1 877 840-2746

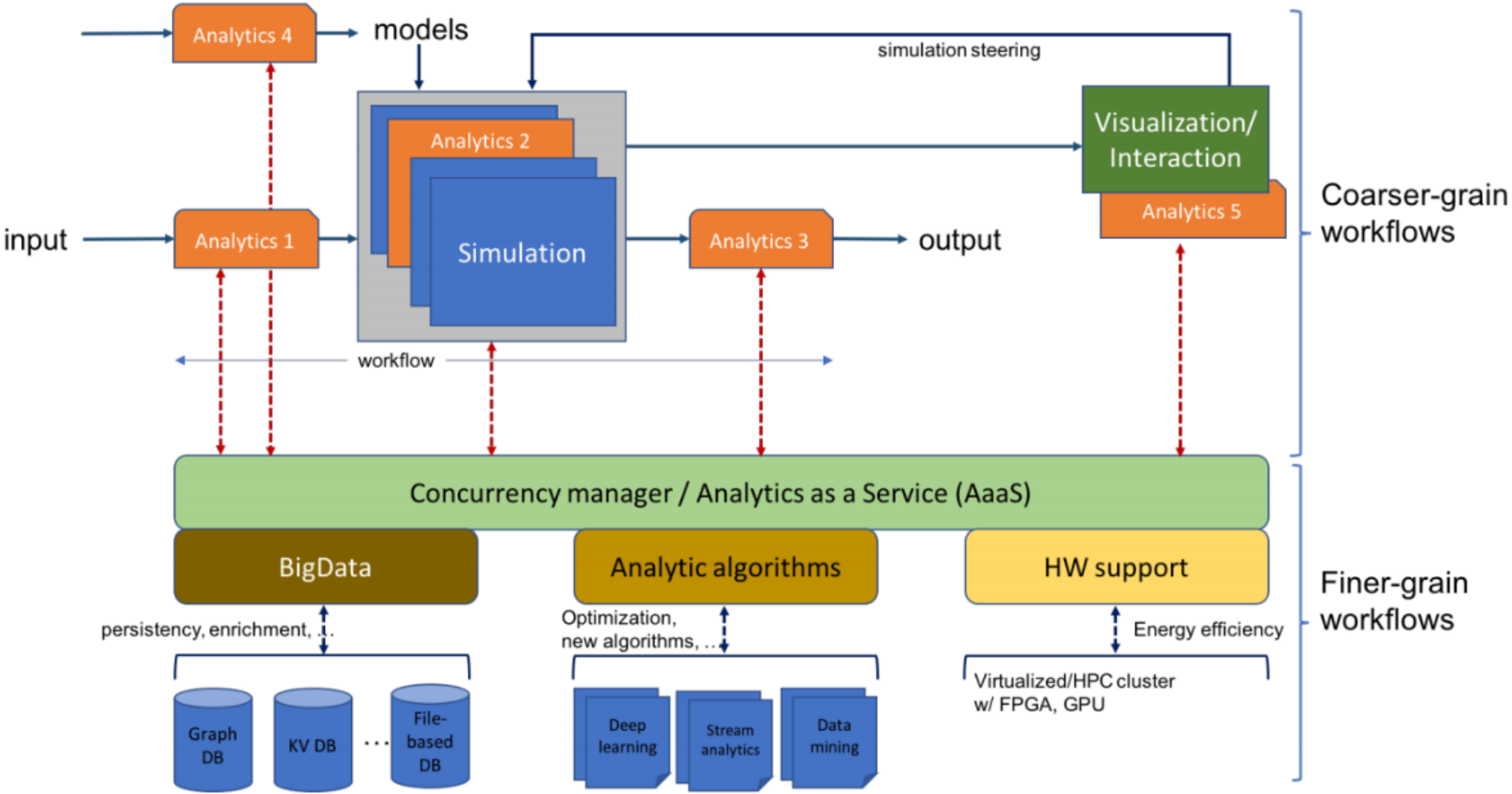
Backup slides


Research and Scientific Communities
Governmental Organisations
Commercial Entities
General public, Media, NGO, Education
EXPLOITATION TIER
Exploitation Layer
Scientific Data Exploitation
Public Sector Benefits
Industry Growth
Copernicus Services
Platform Services Layer
Identity Management
tep
thematic exploitation platform
Open Earth Engine
Data Cubes
Proba-V MEP
National/Commercial Platform Services including Coll. G/S
Resources Tier Layer
Data
ICT
Copernicus DIAS Back-Office
National/Commercial Mission Platforms
National/Commercial Platform Services including Coll. G/S
Data Generation Layer
Ground Segment
Infra-structure
esa
ESA
Heritage, Earth Explorer, Third party Operations
Copernicus Sentinels Operations
EUMETSAT
National/commercial Operations

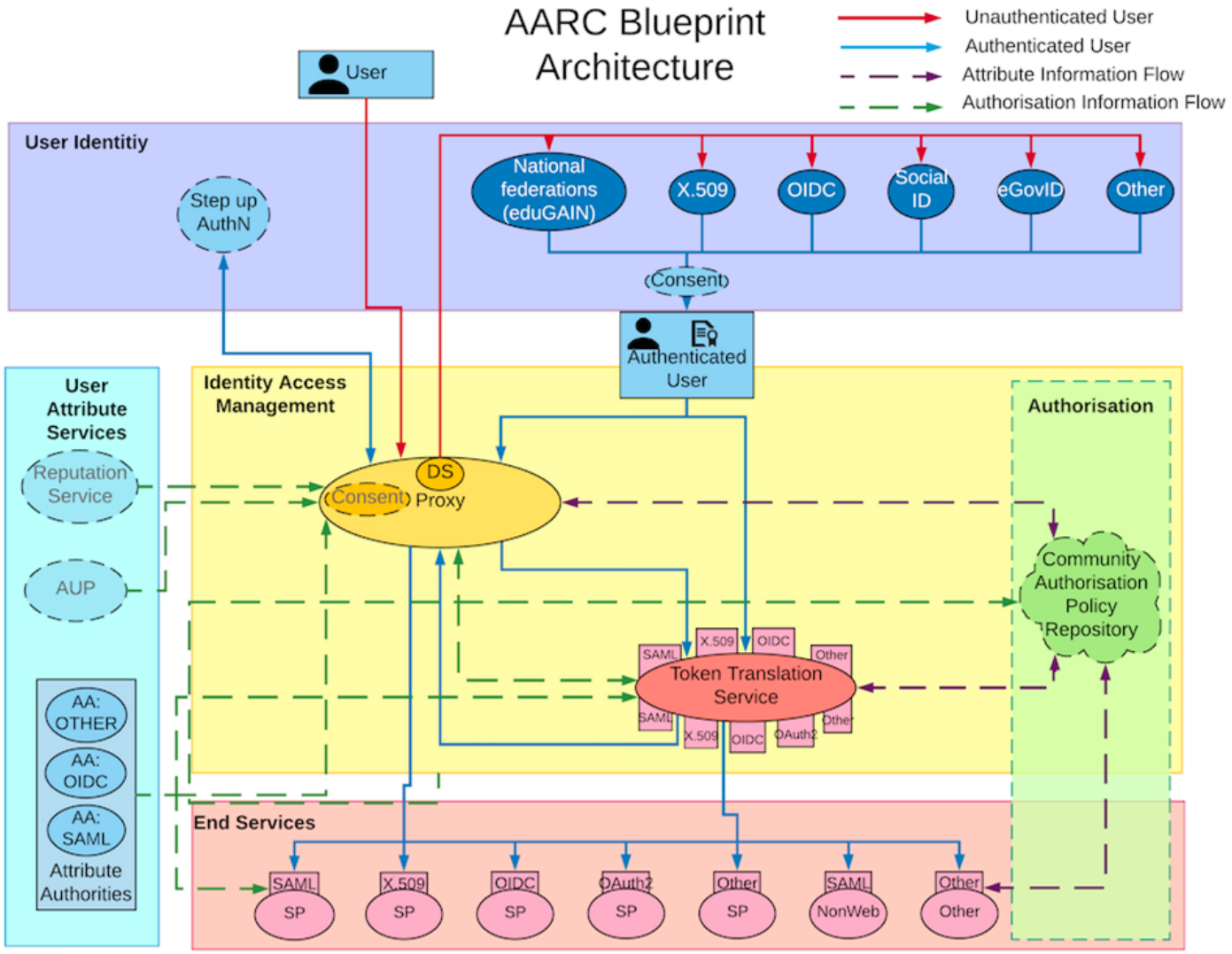
Execution Management System (EMS)



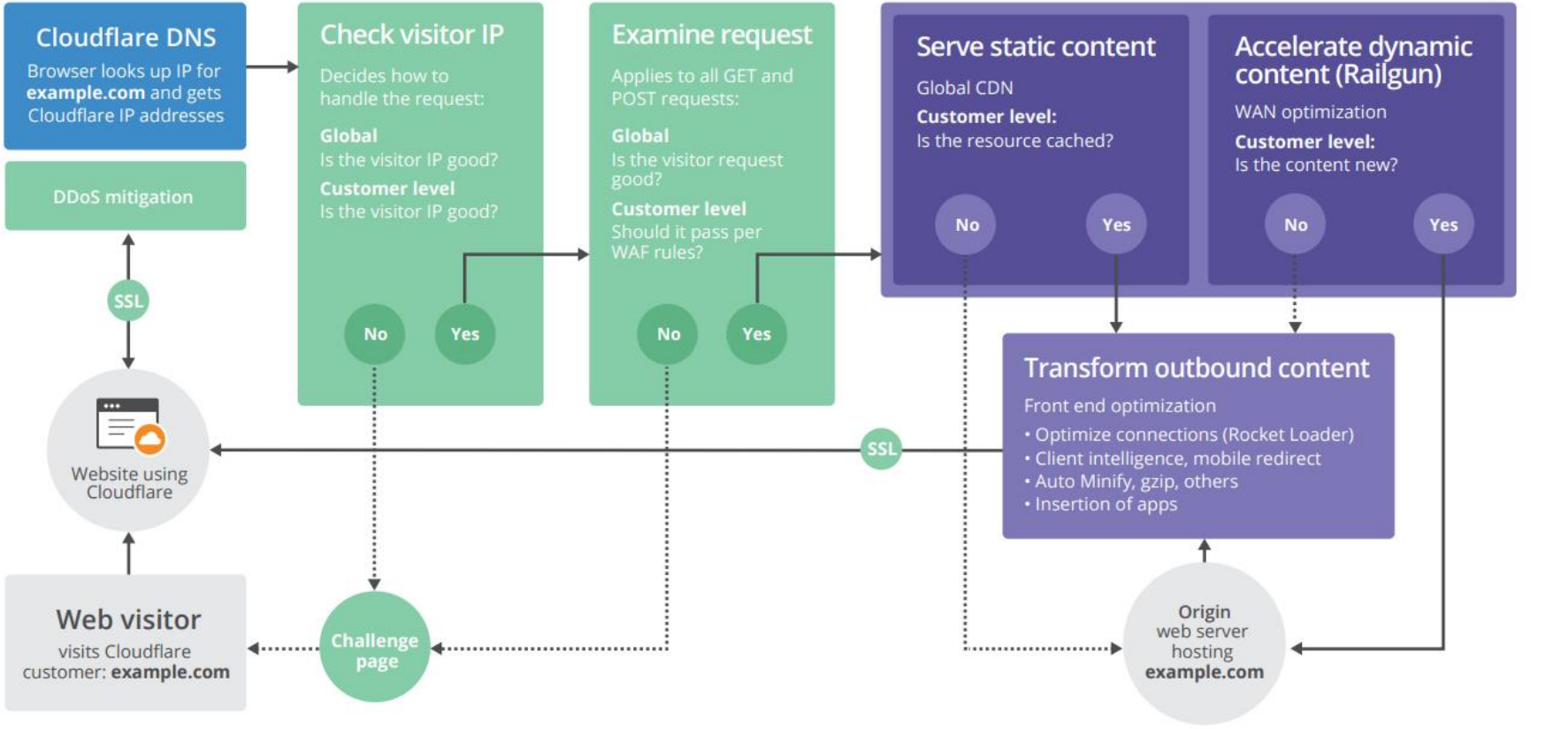
CRIM



Source: Badia, R. and al., Workflows for Science: a Challenge when Facing the Convergence of HPC and Big Data, in Supercomputing Frontiers and Innovation, 2017


AARC Blueprint Architecture
Unauthenticated User
Authenticated User
Attribute Information Flow
Authorisation Information Flow
User
User Identitiy
Step up AuthN
National federations (eduGAIN)
X.509
OIDC
Social ID
eGovID
Other
Consent
Authenticated User
User Attribute Services
Reputation Service
AUP
AA: OTHER
AA: OIDC
AA: SAML
Attribute Authorities
Identity Access Management
DS
Consent
Proxy
Token Translation Service
SAML
X.509
OIDC
Other
OAuth2
Authorisation
Community Authorisation Policy Repository
End Services
SAML SP
X.509 SP
OIDC SP
OAuth2 SP
Other SP
SAML NonWeb
Other Other

Cloud Flare


Cloudflare DNS
Browser looks up IP for example.com and gets Cloudflare IP addresses
DDoS mitigation
SSL
Website using Cloudflare
Web visitor
visits Cloudflare customer: example.com
Check visitor IP
Decides how to handle the request:
Global
Is the visitor IP good?
Customer level
Is the visitor IP good?
No
Yes
Challenge page
Examine request
Applies to all GET and POST requests:
Global
Is the visitor request good?
Customer level
Should it pass per WAF rules?
No
Yes
Serve static content
Global CDN
Customer level:
Is the resource cached?
No
Yes
Accelerate dynamic content (Railgun)
WAN optimization
Customer level:
Is the content new?
No
Yes
Transform outbound content
Front end optimization
• Optimize connections (Rocket Loader)
• Client intelligence, mobile redirect
• Auto Minify, gzip, others
• Insertion of apps
SSL
Origin web server hosting example.com
DNS
Security
Performance
Note: SSL is optional and supported for all traffic. Every node on the Cloudflare network can perform any task.

Amazon SageMaker Neo

Train Your Machine Learning Models Once, Run Them Anywhere