GA No 777413

14th RDA Plenary co-located event

DARE Platform – Services & Tools

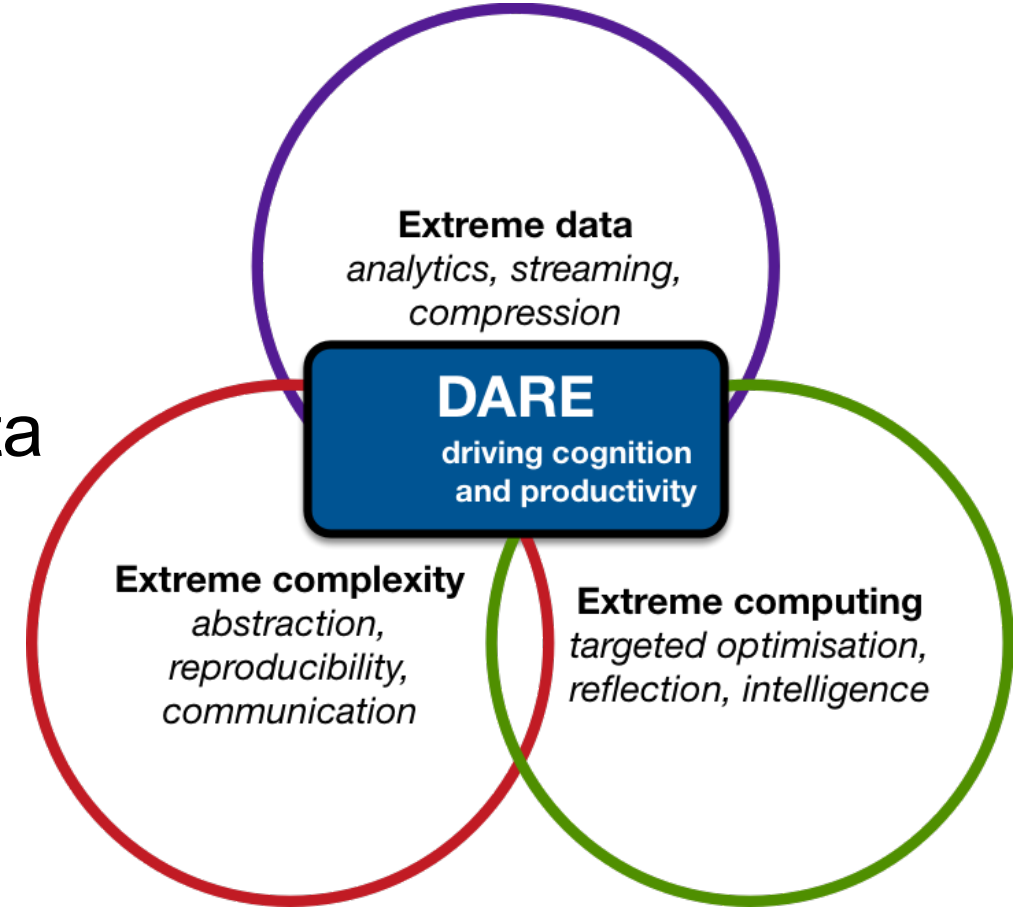
Antonis Koukourikos

NCSR “Demokritos”

DARE for data-driven science

- Increasing diversity and distribution of resources
  - Data, computation
- Increasing complexity of methods
- Increasing scale and complexity of data

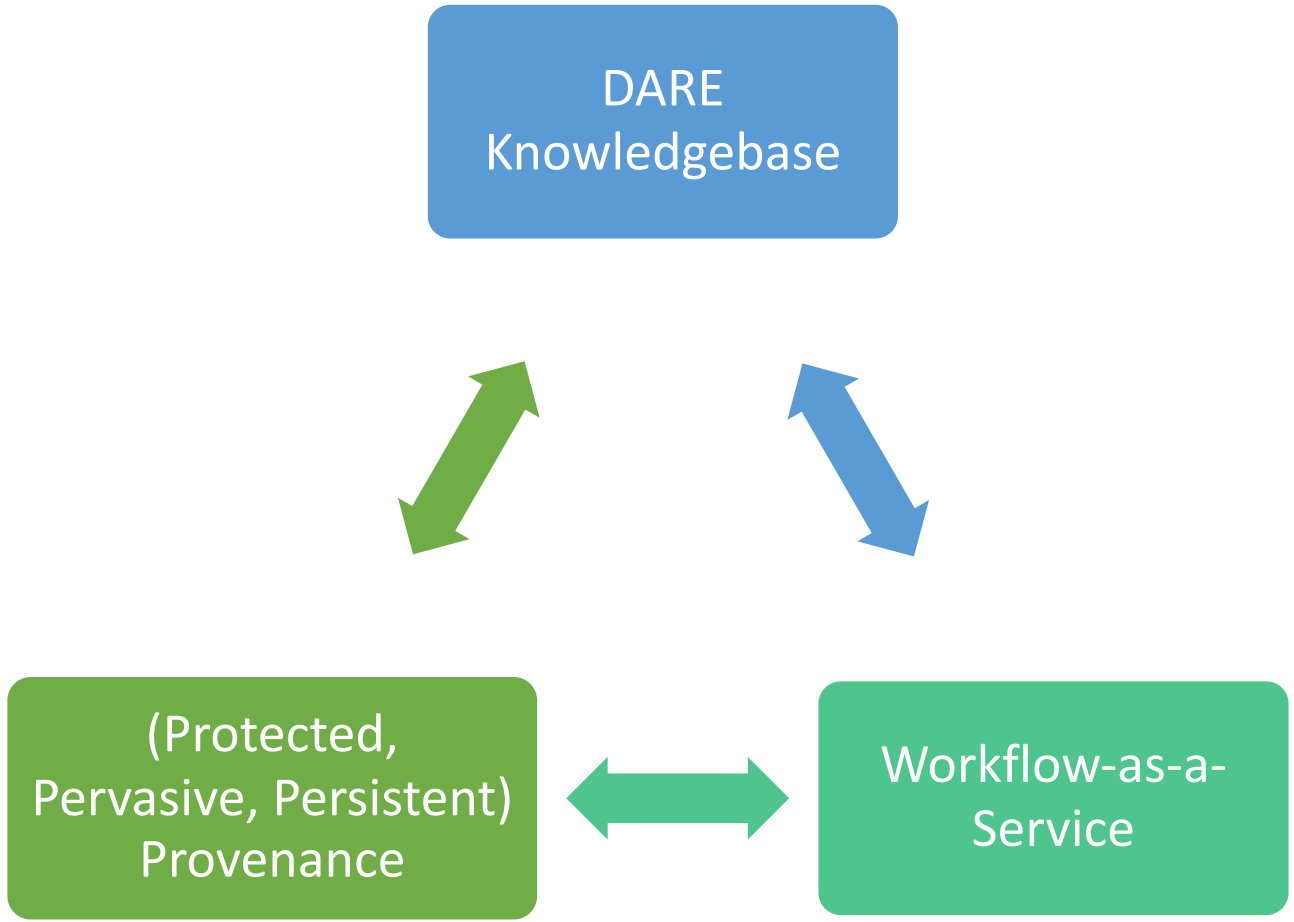
- Research developers and scientists find it difficult to cope



The DARE platform

- A cloud-ready platform for enabling data-driven science
- Provide research developers and scientists with tools to
  - Create powerful workflows in abstract terms
  - Describe and search for semantic properties of methods and data
  - Monitor the execution of parts of workflows
  - Track, query and exchange data provenance records
- Hiding technical detail
- Allow for the use of multiple underlying resources
  - Also of different types

DARE core components

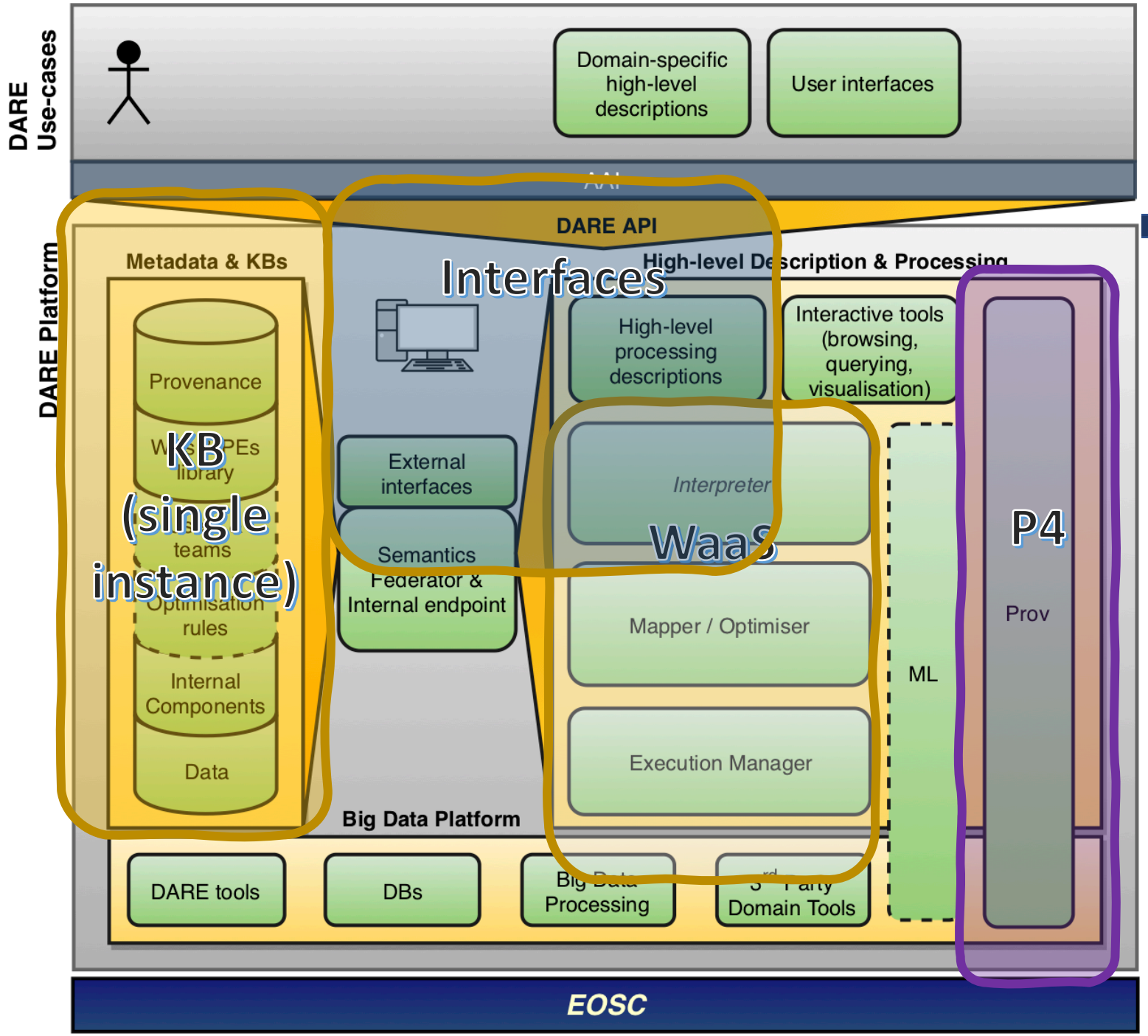


- Interfaces with the world via:
  - Tools and UIs
  - APIs
- Potentially serving
  - Humans
    - Research developers
    - Scientists
    - Data scientists
    - Policy makers
    - …
  - Systems
    - Optimisation
    - Federation
    - …

M. Atkinson et. al. "Comprehensible Control for Researchers and Developers facing Data Challenges", eScience 2019

Platform overview

- DARE KB
  - Catalogues, metadata and semantics
- Optimising WaaS
- Data provenance
- Integrator of big data tools
- RESTful API



Concepts and semantics

- Concepts re methods, data, past runs, etc. are described in the DARE KB
- The KB is the point of reference within the platform
- PEs and workflows can be identified and described
- Data provenance can be used to draw insights about the use of methods and e-infrastructures
- The data and components catalogues will allow users to abstract away from data and software details in the near future
- A graph of concepts pertaining to an experiment can be extracted and communicated

Platform components as microservices

- Each component has its own RESTful API
- Some resources are for use by other DARE components only
- Some are exposed to the outside
- The set of all resources exposed to the outside world forms the DARE API

- Decouple individual components
- Good fit for k8s deployment, scaling and management of containerized applications

The European Open Science Cloud (EOSC)

- A federation of core cloud-based services and infrastructures
- Relationships with RIs, HPC and national stakeholders

- DARE as an EOSC platform
- Exposed API to enable higher-level, user-facing application and services
- *Investigate potential collaborations with key infrastructures*

Technical priorities for the next RP

Task	Priority
AAI	Highest
Data catalogue	Highest
Components catalogue	High
Link PROV and WFs catalogues	High
WF optimisation	Highest
ML for optimising WFs	Low
Python library wrapping the DARE API	Medium
Semantification of WFs and PEs	High

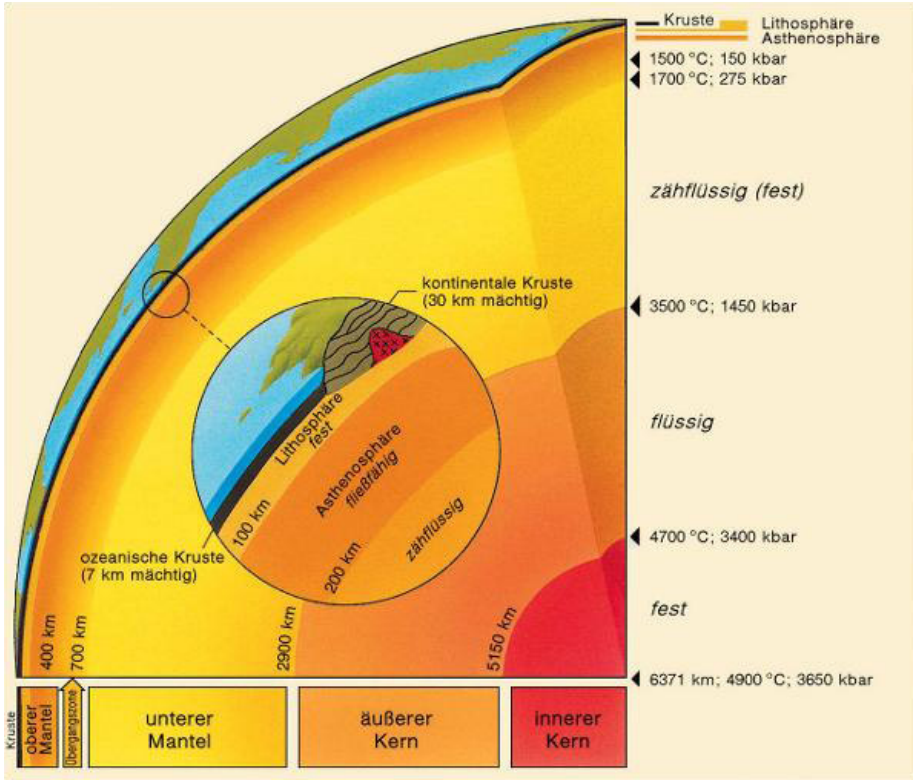
Actions to investigate with EINFRA-21 projects

Projects	Action for investigation
DARE, DEEP, XDC	INDIGO DC orchestrator
DARE, DEEP	INDIGO DC udocker
DARE, DEEP	DARE provenance as a generic service
DARE, FREYA	PIDS
XDC, DARE	Change event notification system
DEEP, DARE	DEEP-as-a-service
EUXDAT, DARE	EUXDAT data connectors

DARE Use Cases

Seismology Community

Fundamental ← **Seismology** → **Applied - Operational**



EPOS Test Cases

Rapid Ground Motion Assessment (**RA**)

Tackle a basic, urgent seismological issue

Seismic Source Analyses (**SS**)

Easily include complex seismological tools/codes from multiple research groups

Ground Motion Ensemble Simulation Analyses (**ES**)

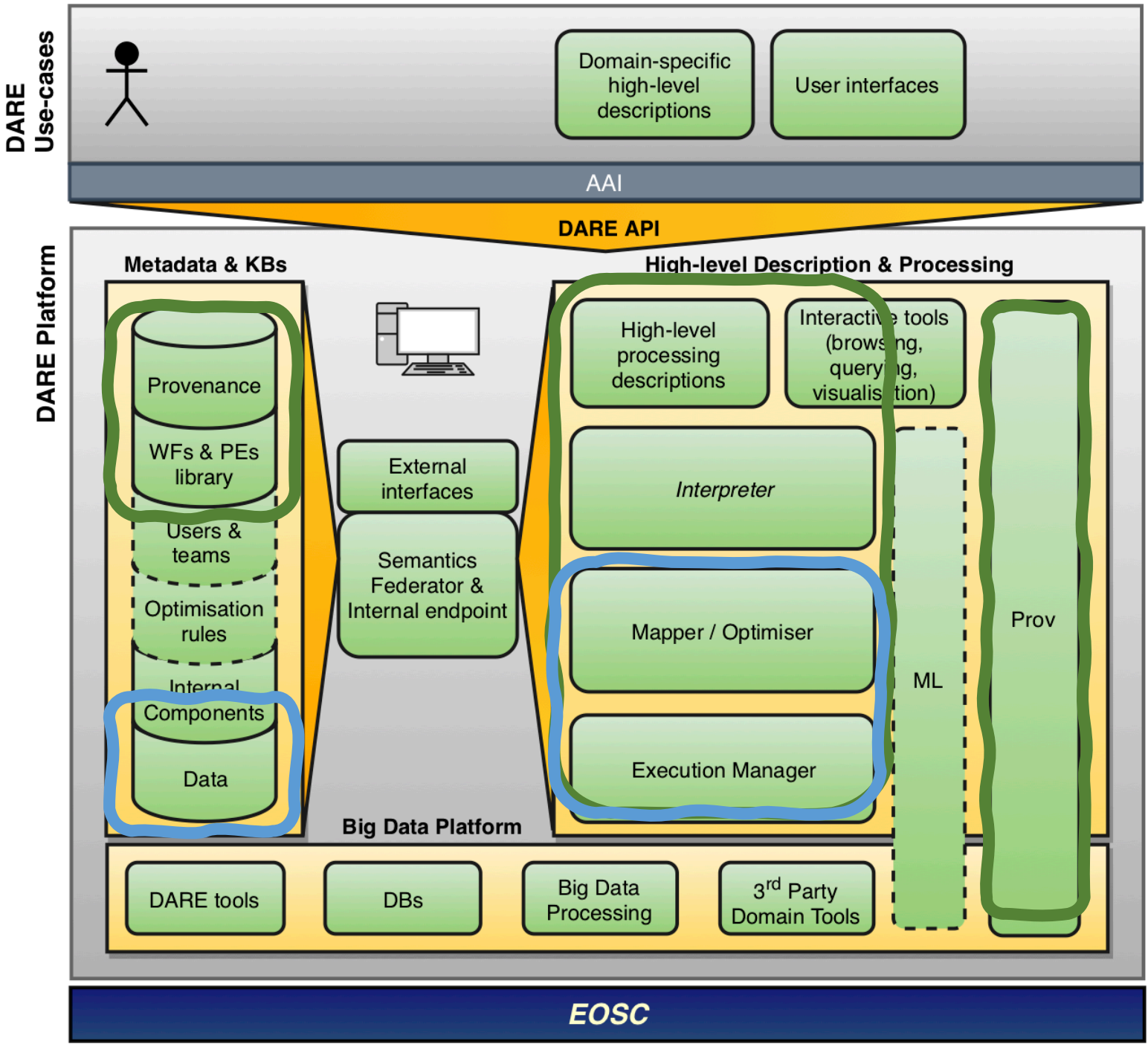
Combine actions into complex workflows that target all three extreme dimensions

EPOS

- Rapid Ground Motion Assessment (RA)

- Optimising WaaS
- Automatic spawning of MPI cluster on the cloud
- WFs & PEs Registry
- Provenance

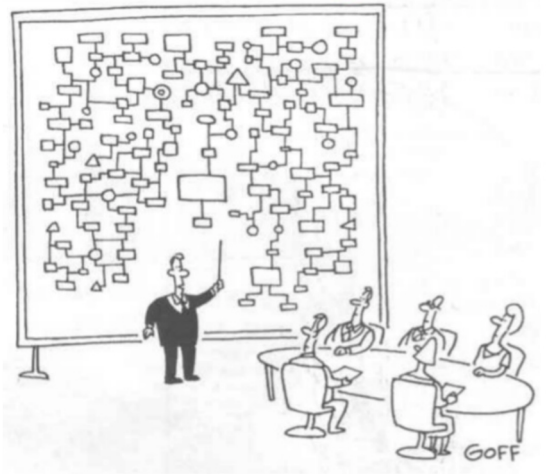
- Data catalogue
- Semantification of execution



Climate Community

- Perform efficient Data Analysis
  - Large number of realizations (ensemble of scenarios)
  - Uncertainties range estimation
  - Process Higher spatial and temporal resolution
  - Easily share intermediate results with collaborators
  - Comparisons when doing numerical model developments

- Achieve a more robust and flexible Data Life Cycle
  - More robust experiments setup
    - Explore several experiment configurations to answer scientific questions
  - Reproducible and traceable experiments
  - **Download locally then Analyze**: a workflow that cannot be sustained



IS-ENES Test Case

Objective: Generate a multi-model multi-scenario time series average of the surface temperature using CMIP5 data

Scientific Workflow

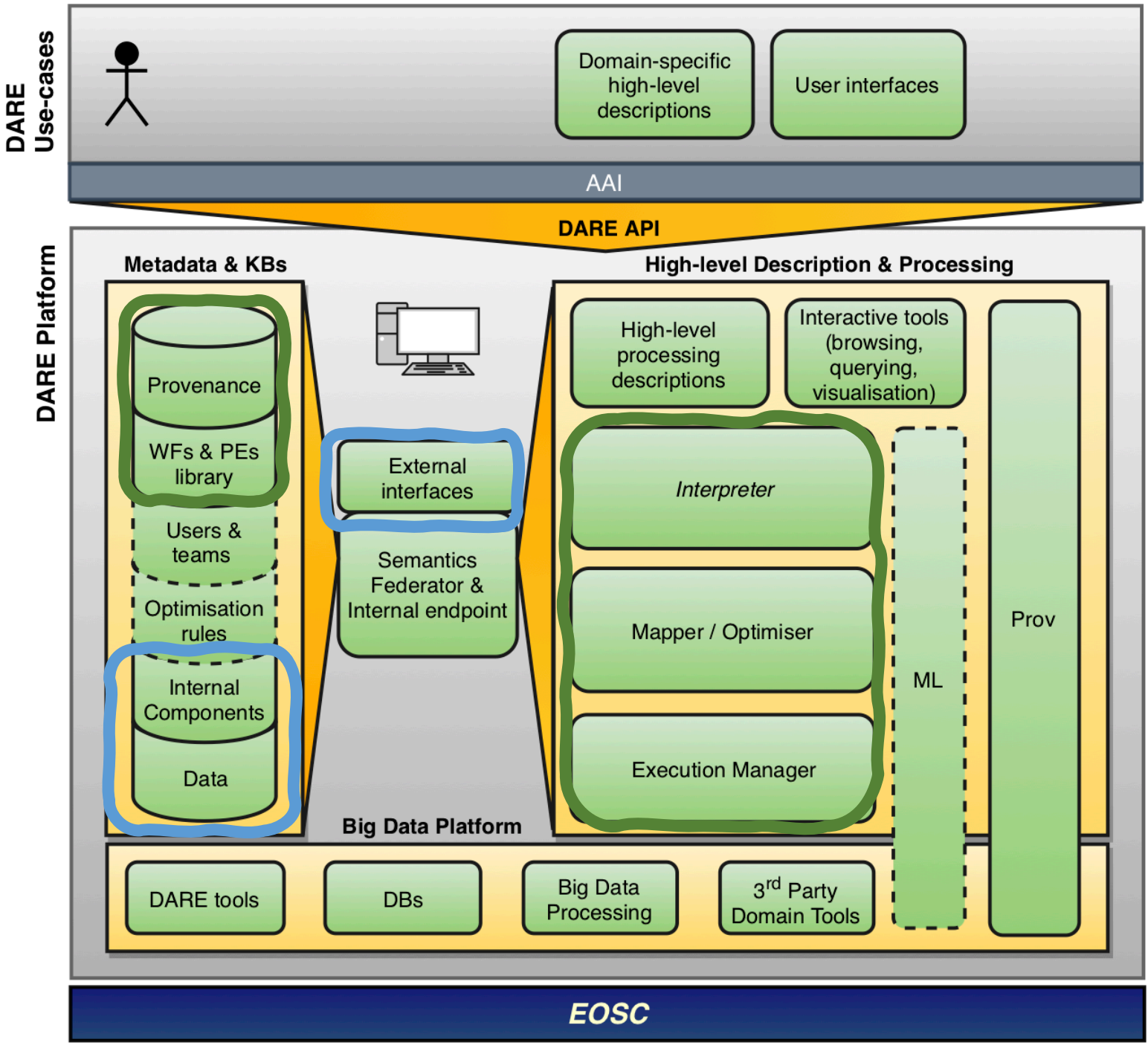
- Spatially average over Western Europe (continents only)
- Time Period 1950-2100
- RCP 8.5 GES scenario
- All Global Climate Models available
- All members available
- Calculate the average time series
- Calculate the standard deviation
- Extract separate time series of every simulation
- Plot all those time series on a single graph

IS-ENES

- C4I data preparation and basic climate modelling

- WaaS
- WFs & PEs Registry
- Provenance

- Data catalogue
- Components catalogue
- Interfaces to external RIs/ESGF



Thank You

- Project Website

http://project-dare.eu

- DARE Platform 2nd Release:

https://gitlab.com/project-dare/dare-platform/tree/v2.0