Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu

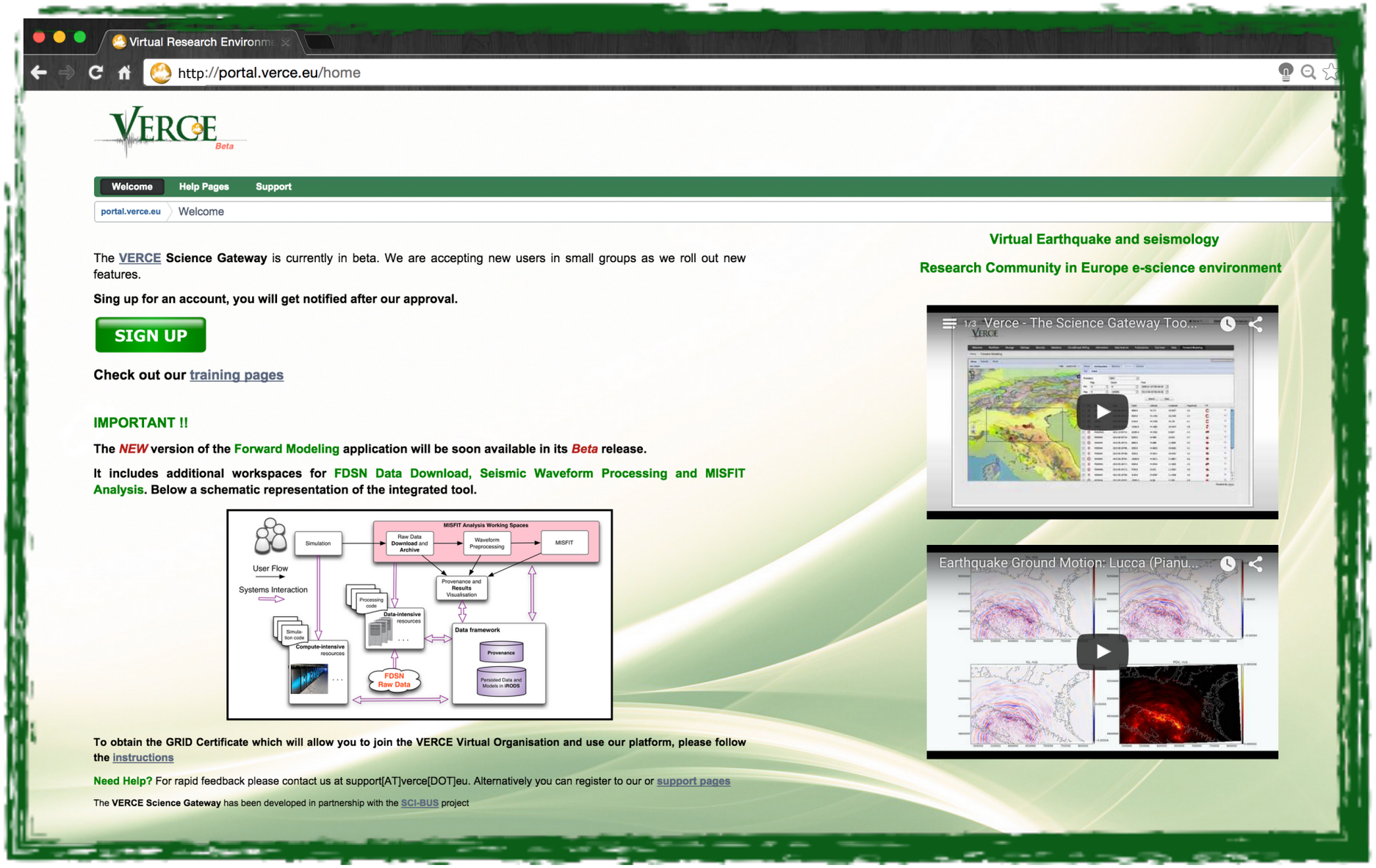
The VERCE Science Gateway

(..and data-intensive provenance)

Alessandro Spinuso
spinuso@knmi.nl

"Democratising Computational Seismology Research in Europe"

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu



INGV

INSTITUT DE PHYSIQUE DU GLOBE DE PARIS

Fraunhofer SCAI

UNIVERSITY OF LIVERPOOL

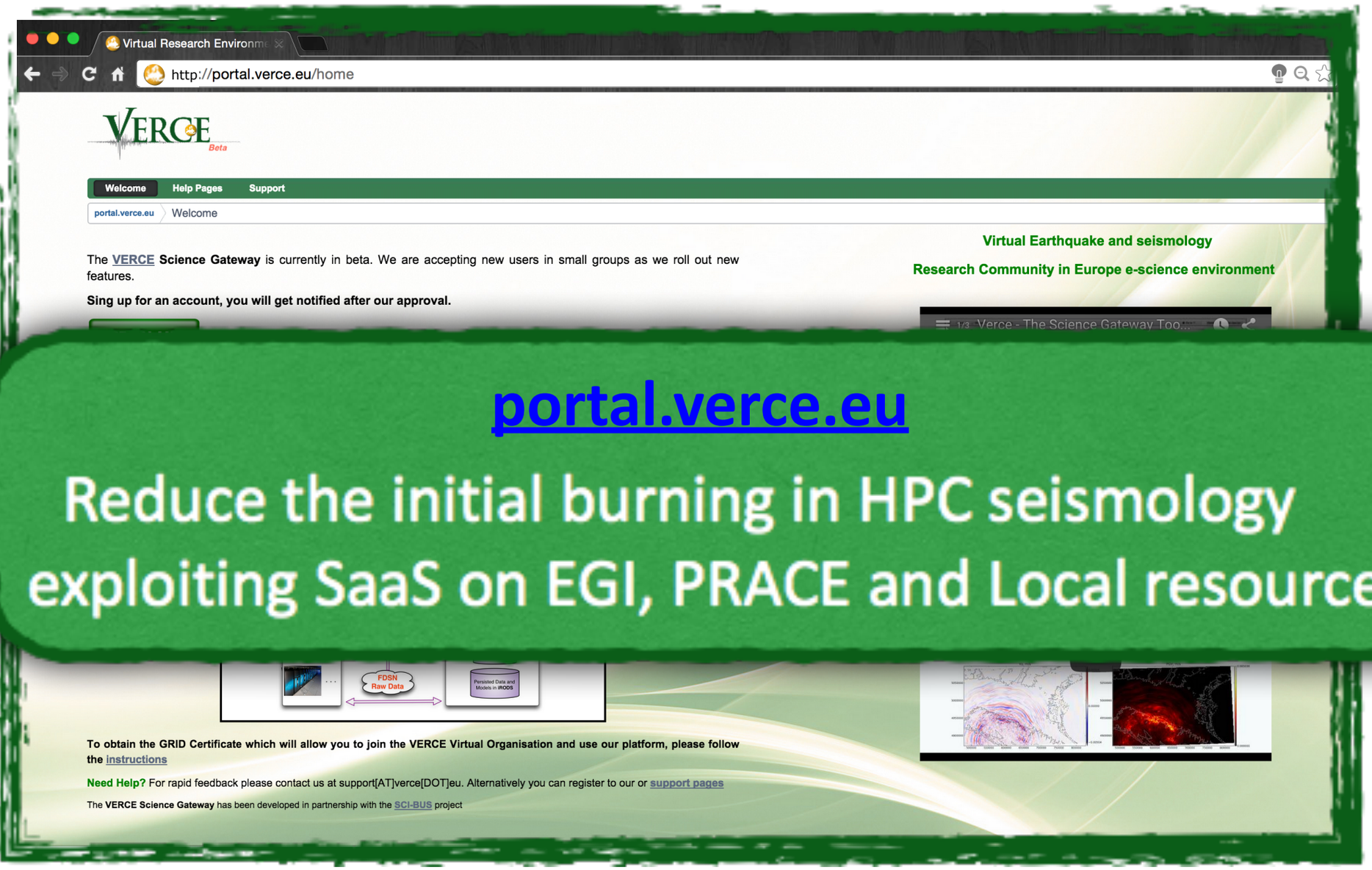
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

CINECA Consorzio Interuniversitario

KNMI

SCI-BUS

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu



portal.verce.eu

Reduce the initial burning in HPC seismology exploiting SaaS on EGI, PRACE and Local resources

- **Earthquake Simulation:** **Synthetic Seismograms** for **public** and **custom Earth models and Earthquakes** via the execution of **HPC simulation codes** called *solvers (SPECFEM3D - SPECFEMGLOBE)*

- **Raw data acquisition & Misfit:** The synthetic data may be **compared with real observations stored in distributed archives (FDSN), adopting Data Intensive methods**

VERCE Recipe

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu

SEISMOLOGICAL INGREDIENTS

MESH

MATERIAL PROPERTIES (velocity model)

NUMERICAL CODE (SPECFEM3D)

EVENTS

STATIONS

HPC INGREDIENTS

HPC CONNECTION
HPC CENTER POLICIES
HPC CENTER

SEISMOLOGICAL OUTPUT

SYNTHETIC SEISMOGRAMS

VOLUMETRIC DATA

Shakemap

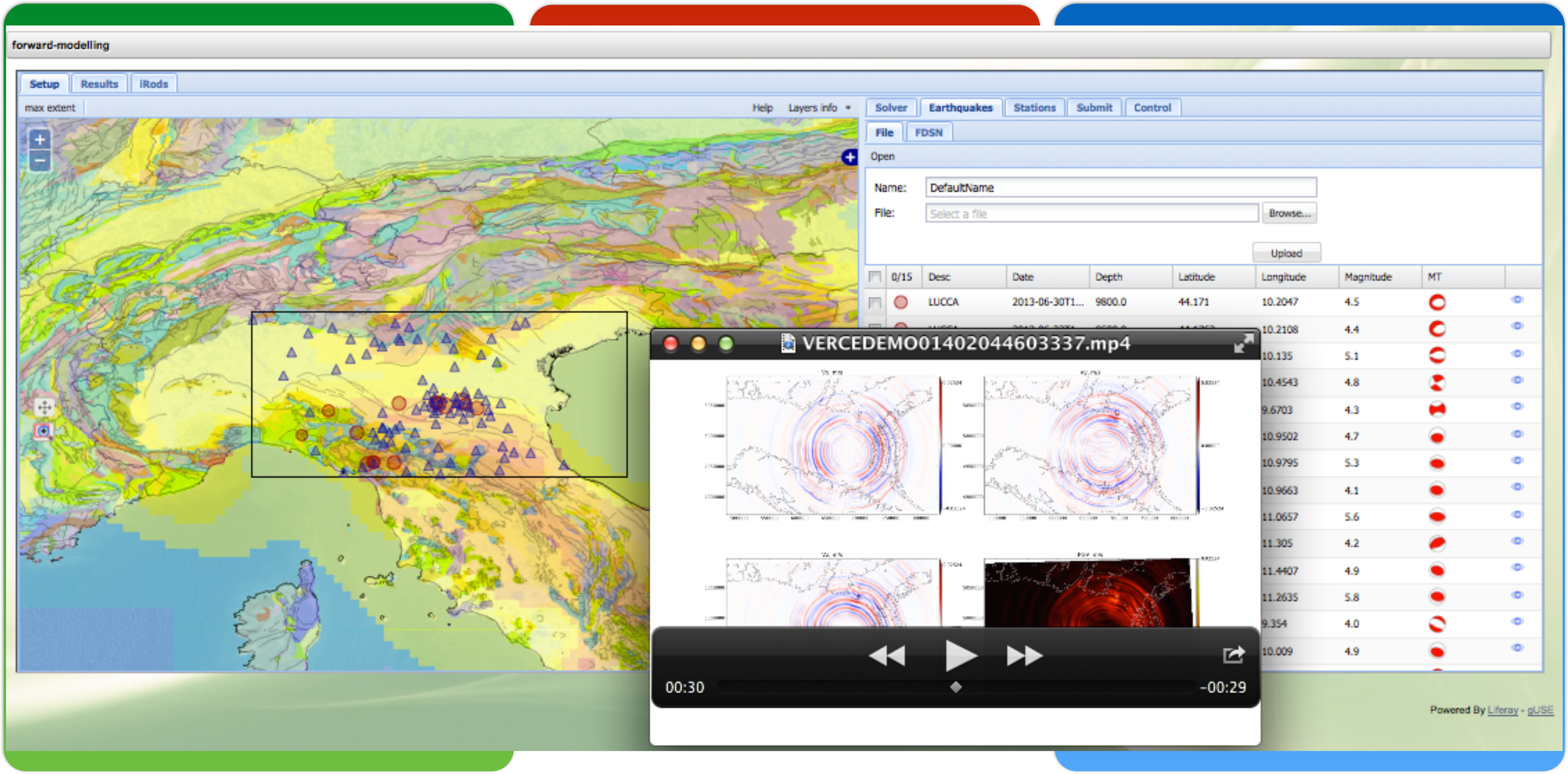
Movie

SYNT - OBS MISFIT

VERCE

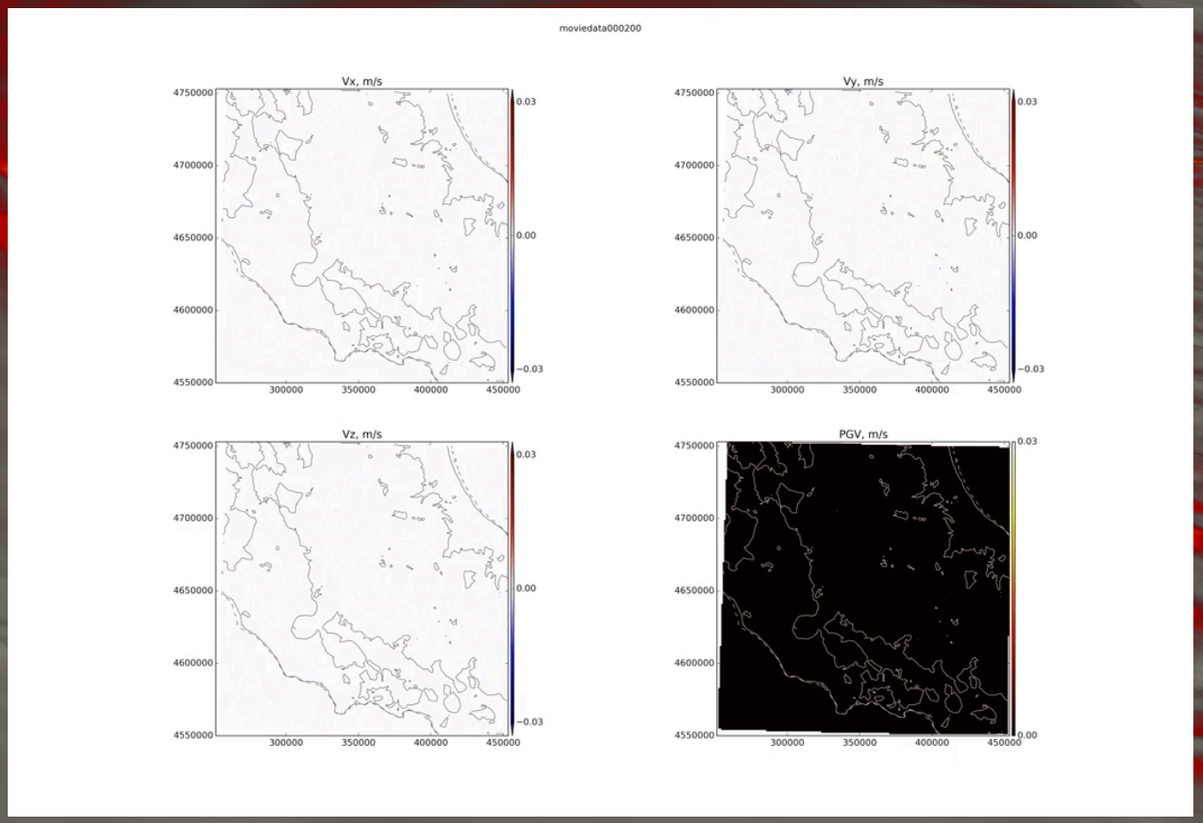
VERCE Recipe

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu



VERCE

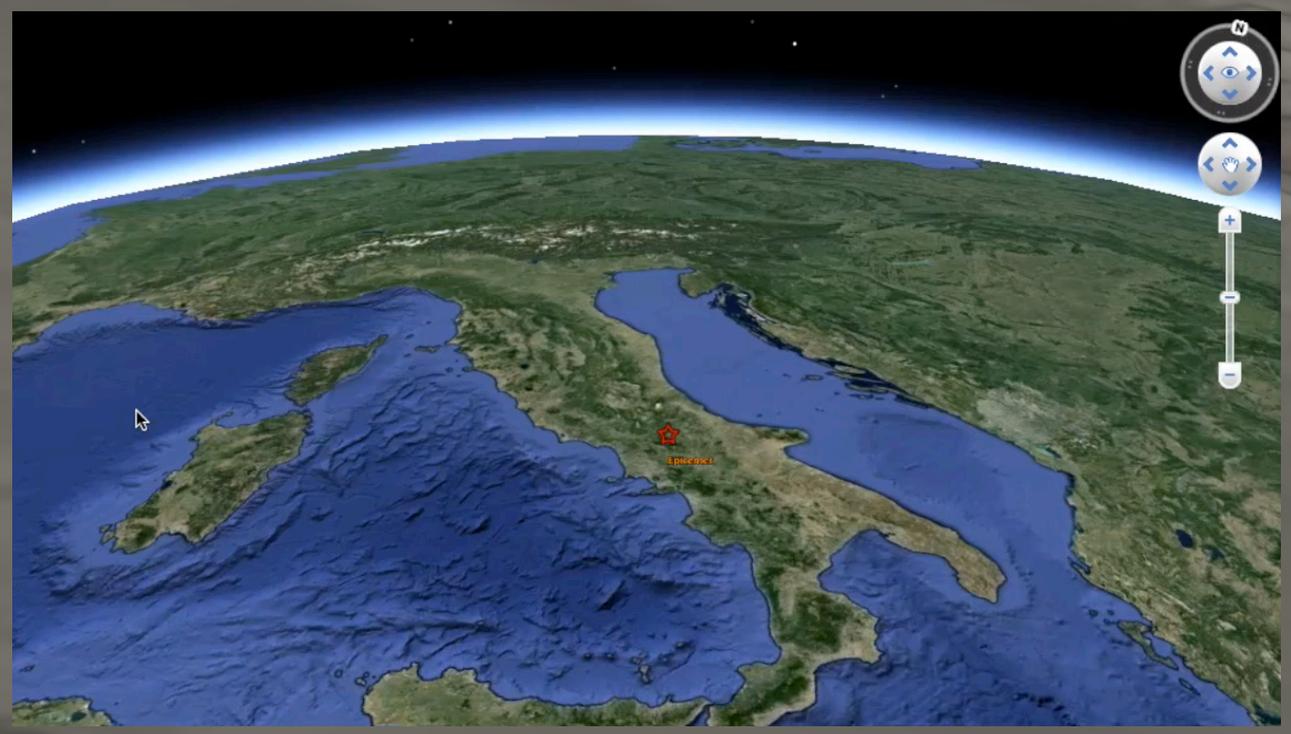
AMATRICE Mw 6.0 24/08/2016



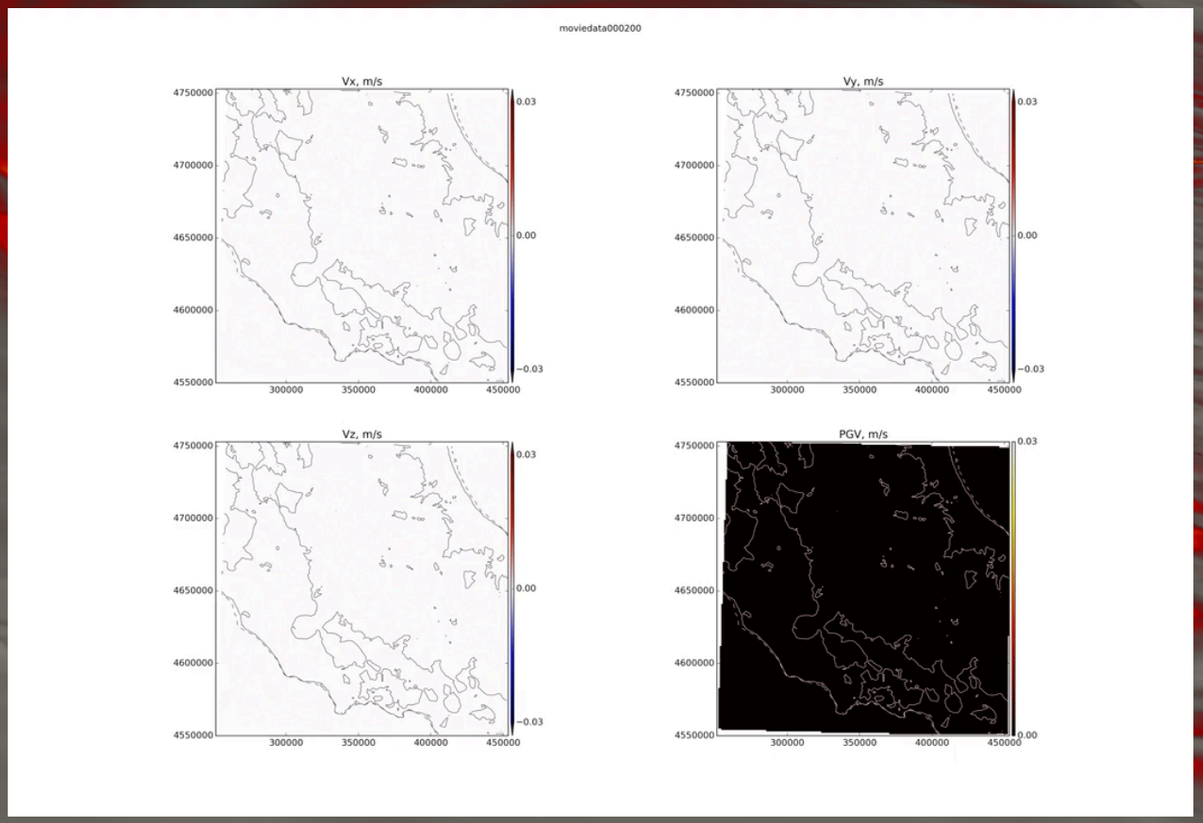
OUTPUT
seismograms, plots, 3D Geometry, Videos, KMZ packages, meshes and model**s > 10 GB**

HPC Resources @ Fraunhofer SCAI

VR Products
3D Geometry for CAVE Visualisation experimented at LRZ



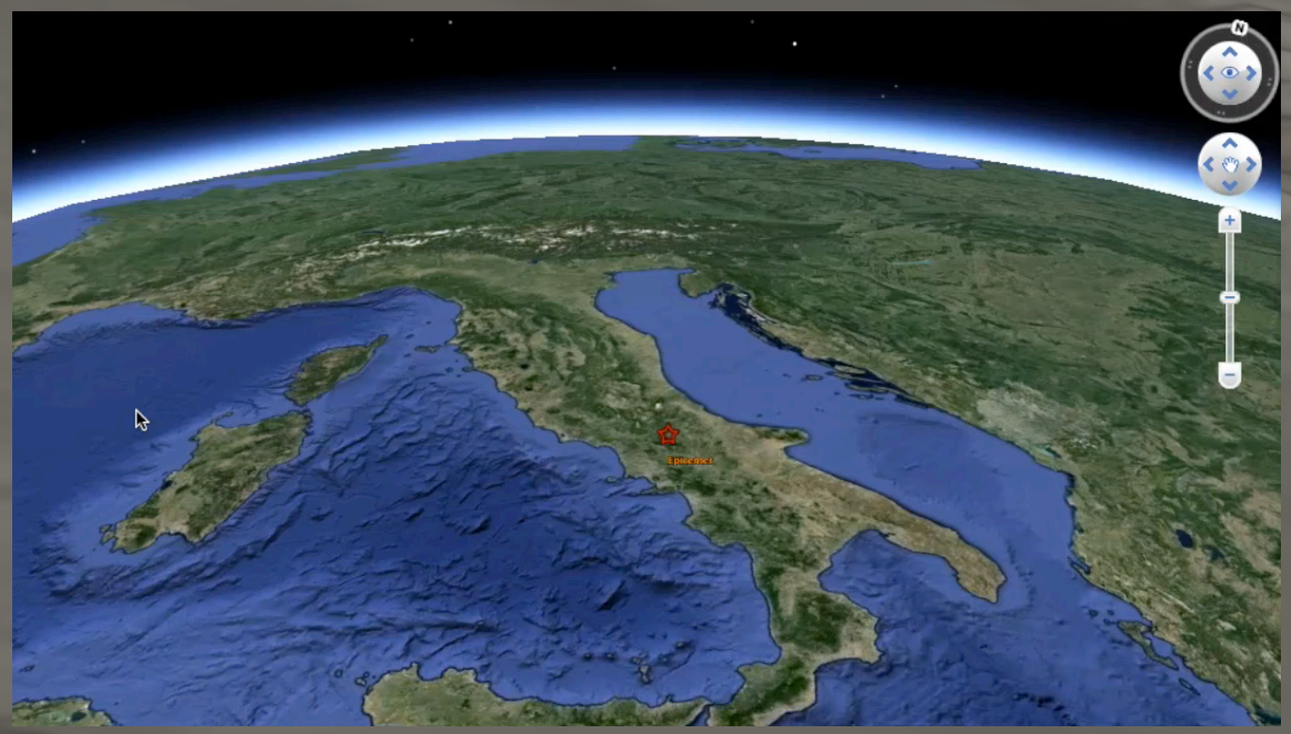
AMATRICE Mw 6.0 24/08/2016



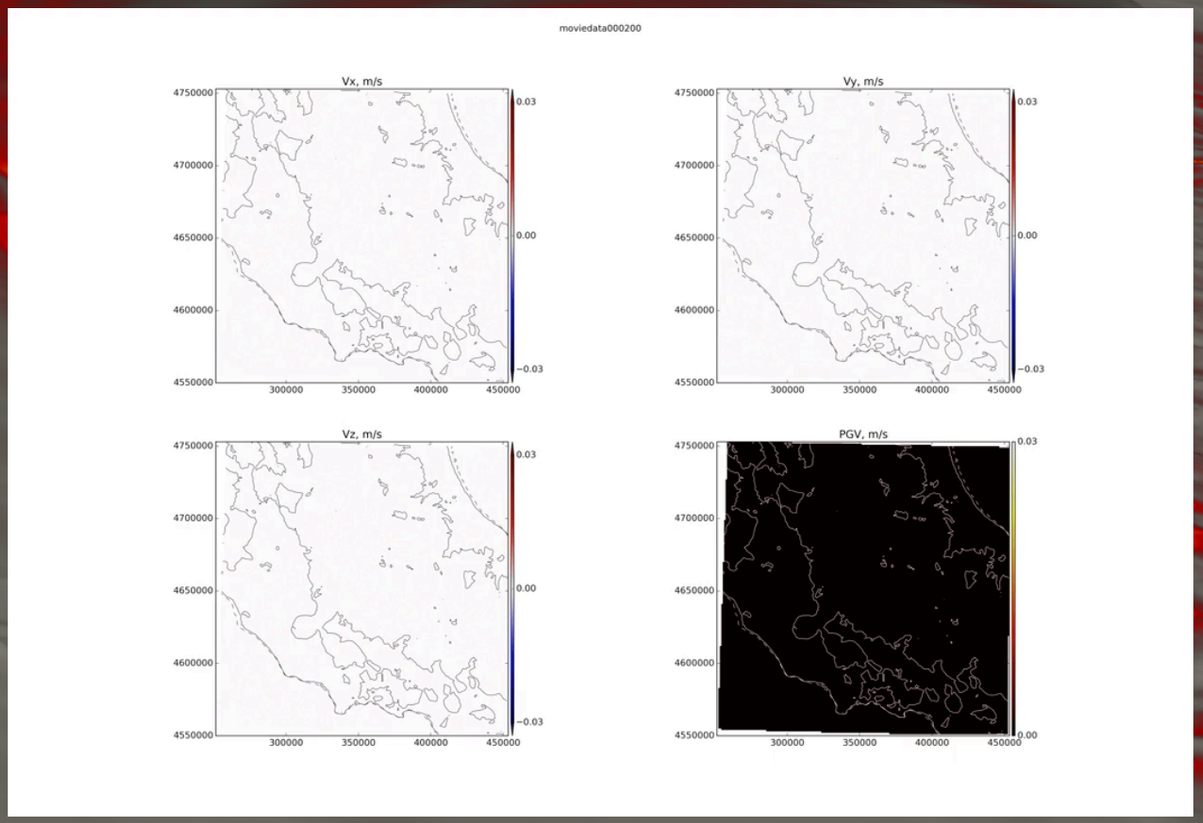
OUTPUT
seismograms, plots, 3D Geometry, Videos, KMZ packages, meshes and model**s > 10 GB**

HPC Resources @ Fraunhofer SCAI

VR Products
3D Geometry for CAVE Visualisation experimented at LRZ



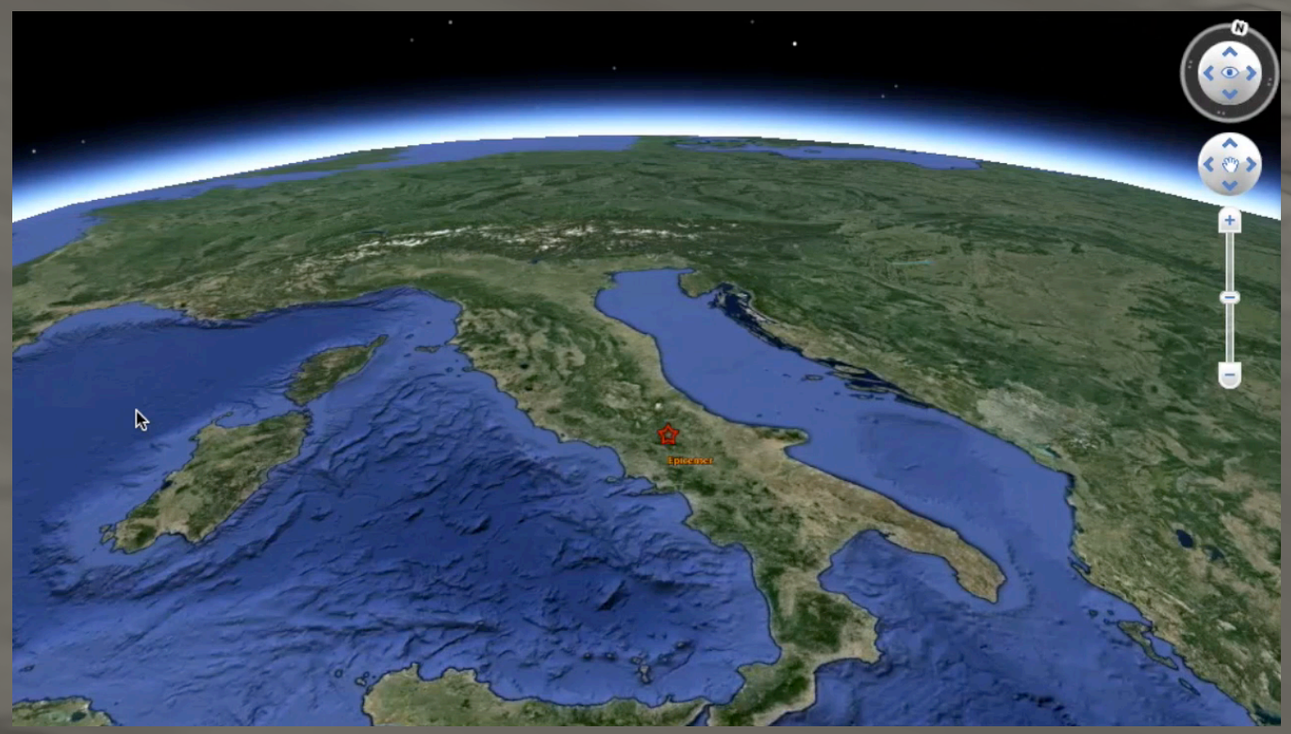
AMATRICE Mw 6.0 24/08/2016



OUTPUT
seismograms, plots, 3D Geometry, Videos, KMZ packages, meshes and model**s > 10 GB**

HPC Resources @ Fraunhofer SCAI

VR Products
3D Geometry for CAVE Visualisation experimented at LRZ



User Perspective

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu

Earthquake Simulation

Integrated view of community webservices via the Forward Simulation GUI

NERIES Geology · EFEHR · FDSN · OGC® Open Geospatial Consortium, Inc. · <StationXML> · <QuakeML>

User's Actions

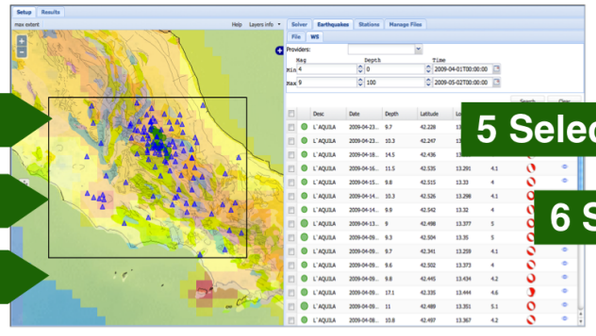
1 Select Solver

2 Select Mesh

3 Select Velocity Model

4

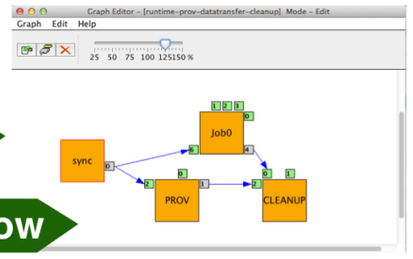
- Upload Input
- Retrieve Input
- Reuse Input



VERCE Science Gateway

5 Select Workflow

6 Submit Workflow



SCI-BUS

dispel4py

Provenance Production

Many Tasks Simulation workflows in HPC resources

MyProxy
Credential Management Service

7

- Access Provenance
- Visualise Results

i·R·O·D·S
Integrated Rule-Oriented Data System

Data storage within the VERCE Federation

VERCE

User Perspective

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu

Misfit between SYNTHETICS and DATA

Simulation | Download | Processing | Misfit | Results

Setup | Submit | Control

Name	Desc	Workflow	Date
simulation_abruzzo13430000000_1485181669330	15 stazioni	SCAI_mpi_SPECFEM_PRODUCTION__2015-03-03-120220_2015-11-19-151723	23 - 01 - 2017
simulation_abruzzo1343000000_1485180676096	15 stazioni	SCAI_mpi_SPECFEM_PRODUCTION__2015-03-03-120220_2015-11-19-151723	23 - 01 - 2017
simulation_abruzzo134300000_1485175329249	15 stazioni	SCAI_mpi_SPECFEM_PRODUCTION__2015-03-03-120220_2015-11-19-151723	23 - 01 - 2017
simulation_abruzzo13430000_1485173261691	15 stazioni	SCAI_mpi_SPECFEM_PRODUCTION__2015-03-03-120220_2015-11-19-151723	23 - 01 - 2017
simulation_abruzzo1343000_1485172391804	15 stazioni	SCAI_mpi_SPECFEM_PRODUCTION__2015-03-03-120220_2015-11-19-151723	23 - 01 - 2017
simulation_abruzzo134300_1484940067541	15 stazioni	SCAI_mpi_SPECFEM_PRODUCTION__2015-03-03-120220_2015-11-19-151723	20 - 01 - 2017
simulation_abruzzo134300_1484938577498		2015-03-03-120220_2015-11-19-151723	20 - 01 - 2017
simulation_abruzzo134300_1484938138425		2015-03-03-120220_2015-11-19-151723	20 - 01 - 2017
simulation_abruzzo13430_1484937623558		2015-03-03-120220_2015-11-19-151723	20 - 01 - 2017
simulation_abruzzo13430_1484936993904		2015-03-03-120220_2015-11-19-151723	20 - 01 - 2017
simulation_abruzzo13430_1484936730412		2015-03-03-120220_2015-11-19-151723	20 - 01 - 2017
simulation_abruzzo13333000000000_14849358		2015-03-03-120220_2015-11-19-151723	20 - 01 - 2017
simulation_abruzzo1333300000000_148493546		2015-03-03-120220_2015-11-19-151723	20 - 01 - 2017
simulation_abruzzo133330000000_1484750697		2015-03-03-120220_2015-11-19-151723	18 - 01 - 2017
simulation_abruzzo13333000000_14829339375		2015-03-03-120220_2015-11-19-151723	28 - 12 - 2016
simulation_abruzzo13333000_1481628715387		2015-03-03-120220_2015-11-19-151723	13 - 12 - 2016
simulation_macerata00_1481563368758		2015-03-03-120220_2015-11-19-151723	12 - 12 - 2016
simulation_abruzzo1333300_1481559849742		2015-03-03-120220_2015-11-19-151723	12 - 12 - 2016
simulation_abruzzo133330_1481300770050	15 stazioni	SCAI_mpi_SPECFEM_PRODUCTION__2015-03-03-120220_2015-11-19-151723	09 - 12 - 2016
simulation_abruzzo133330_1481300521985	15 stazioni	SCAI_mpi_SPECFEM_PRODUCTION__2015-03-03-120220_2015-11-19-151723	09 - 12 - 2016
simulation_macerata00_1481300109953	italia	SCAI_mpi_SPECFEM_PRODUCTION__2015-03-03-120220_2015-11-19-151723	09 - 12 - 2016
simulation_macerata00_1481299879297	italia	SCAI_mpi_SPECFEM_PRODUCTION__2015-03-03-120220_2015-11-19-151723	09 - 12 - 2016
simulation_abruzzo16000000000_1481299507926	15 stazioni	SCAI_mpi_SPECFEM_PRODUCTION__2015-03-03-120220_2015-11-19-151723	09 - 12 - 2016
simulation_abruzzo1600000000_1481299057509	15 stazioni	SCAI_mpi_SPECFEM_PRODUCTION__2015-03-03-120220_2015-11-19-151723	09 - 12 - 2016
simulation_abruzzo160000000_1481298000065	15 stazioni	SCAI_mpi_SPECFEM_PRODUCTION__2015-03-03-120220_2015-11-19-151723	09 - 12 - 2016
simulation_abruzzo16000000000000_1481119958747	15 stazioni	SCAI_mpi_SPECFEM_PRODUCTION__2015-03-03-120220_2015-11-19-151723	07 - 12 - 2016
simulation_macerata0_1477641667271	italia	SCAI_mpi_SPECFEM_PRODUCTION__2015-03-03-120220_2015-11-19-151723	28 - 10 - 2016

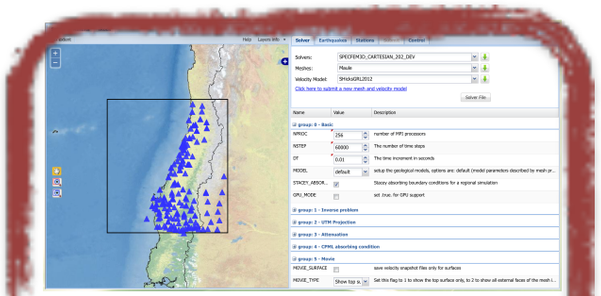
Select Previous Simulation Run for the parametrisation of RAW Data Caching Workflow

VERCE

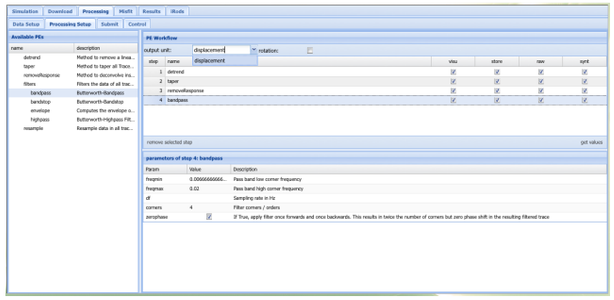
User Perspective

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu

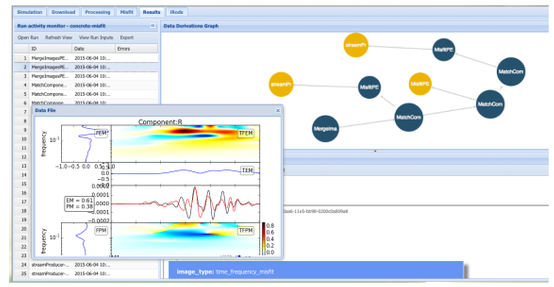
Misfit between SYNTHETICS and DATA



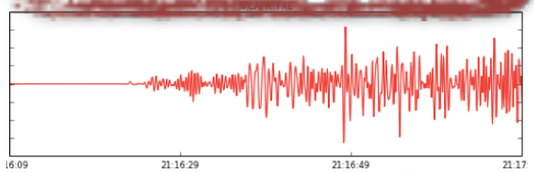
Simulated Synthetics

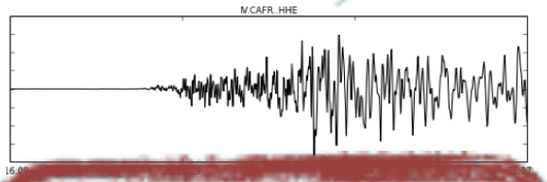


Data/Synt Processing

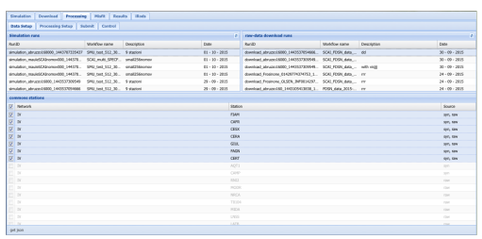


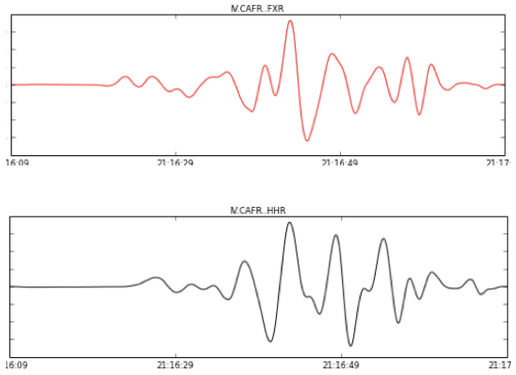
Misfit Analysis

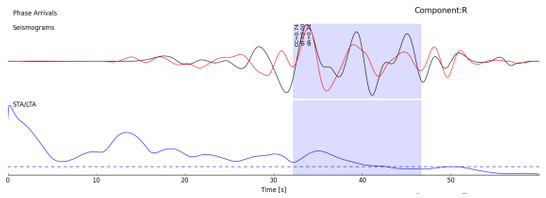


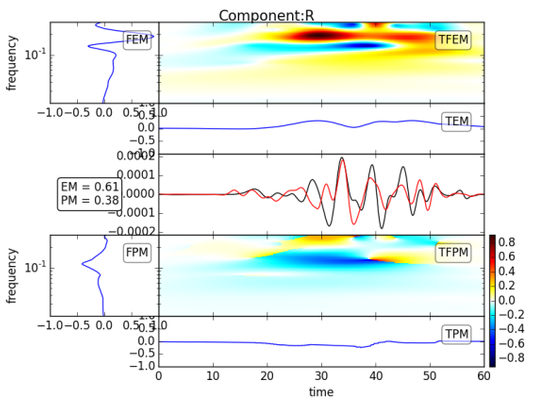


Data Download (FDSN)









dispel4py

ObsPy
A Python Framework for Seismology

VERCE

User Perspective

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu

Misfit between SYNTHETICS and DATA

Simulation Runs

FDSN Data Download Runs

Stored in VERCE and Reusable!!

			Date
simulation_abruzzo160000_1443707335437	SMU_test_512_30...	9 stazioni	01 - 10 - 2015
simulation_mauleSCAInomov000_144370...	SCAI_multi_SPECF...	small256nomov	01 - 10 - 2015
simulation_mauleSCAInomov000_144370...	SMU_test_512_30...	small256nomov	01 - 10 - 2015
simulation_mauleSCAInomov000_144370...	SMU_test_512_30...	small256nomov	01 - 10 - 2015
simulation_abruzzo16000_1443537309549	SMU_test_512_30...	9 stazioni	29 - 09 - 2015
simulation_abruzzo16000_1443537054666	SMU_test_512_30...	9 stazioni	29 - 09 - 2015

raw-data download ru

RunID			
download_abruzzo16000_1443537054666...	SCAI_FDSN_data_...	dd	30 - 09 - 2015
download_abruzzo16000_1443537309549...	SCAI_FDSN_data_...		30 - 09 - 2015
download_abruzzo16000_1443537309549...	SCAI_FDSN_data_...	with visjjj	30 - 09 - 2015
download_Frosinone_01429774374753_1...	SCAI_FDSN_data_...	rrr	24 - 09 - 2015
download_Frosinone_OLSEN_INF0014297...	SCAI_FDSN_data_...	rrr	24 - 09 - 2015
download_abruzzo160_1443105413838_1...	FDSN_data_2015-...	rrr	24 - 09 - 2015

commons stations

Network	Station	Source
IV	FIAM	syn, raw
IV	CAFR	syn, raw
IV	CESX	syn, raw
IV	CERA	syn, raw
IV	GIUL	syn, raw
IV	FAGN	syn, raw
IV	CERT	syn, raw
IV	AQT1	syn
IV	CAMP	syn
IV	RNI2	raw

get json

Component:R

TFEM
TEM
TFPM
TPM

time

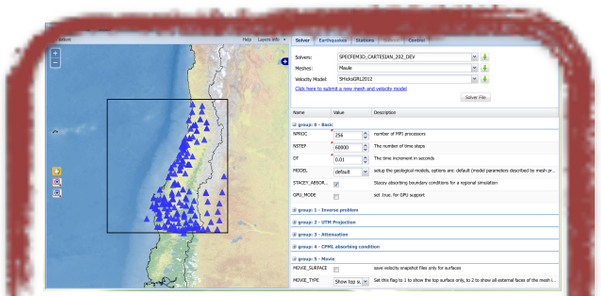
Sensors' data with common properties for Misfit analysis (time-range, event)

VERCE

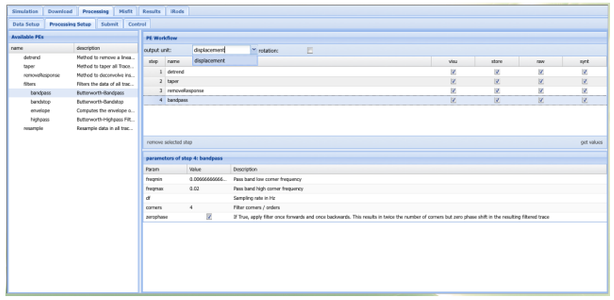
User Perspective

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu

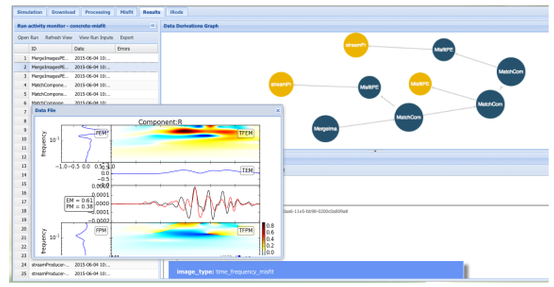
Misfit between SYNTHETICS and DATA



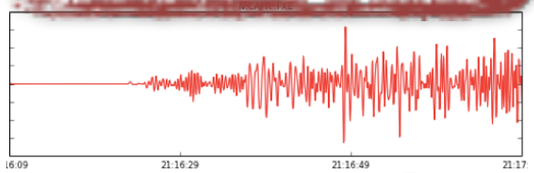
Simulated Synthetics

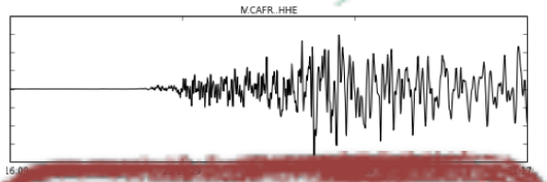


Data/Synt Processing

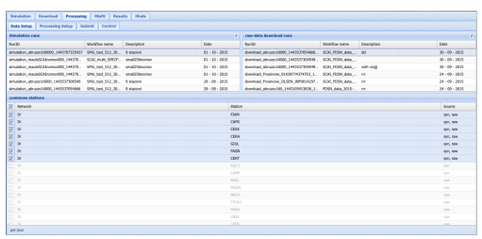


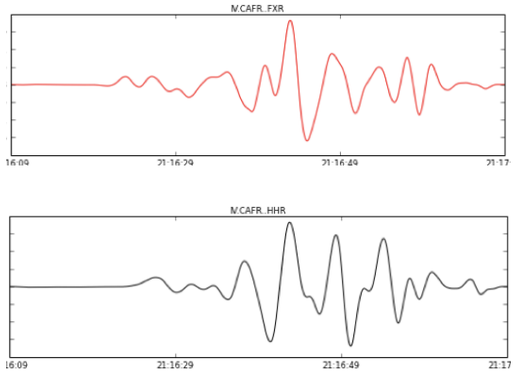
Misfit Analysis

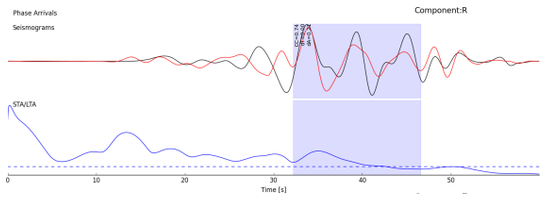


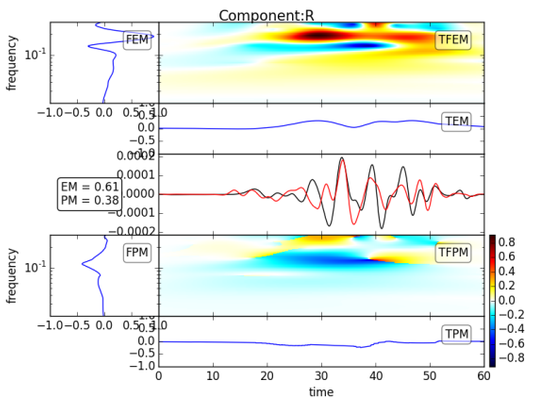


Data Download (FDSN)









dispel4py

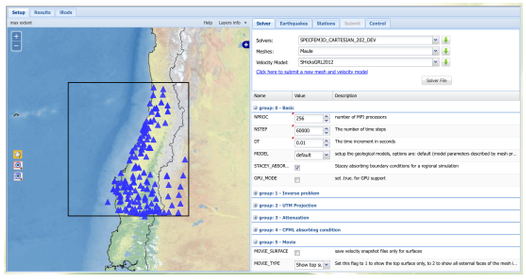
ObsPy
A Python Framework for Seismology

VERCE

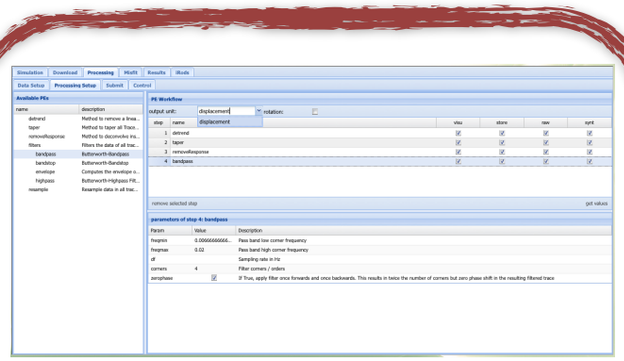
User Perspective

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu

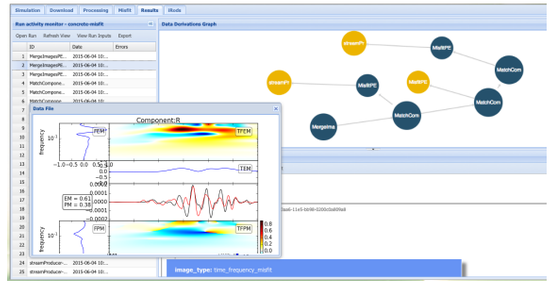
Misfit between SYNTHETICS and DATA



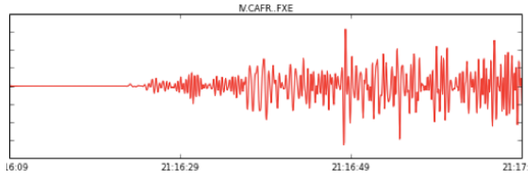
Simulated Synthetics

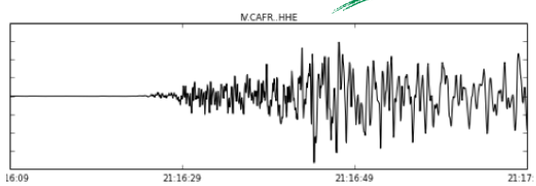


Data/Synt Processing

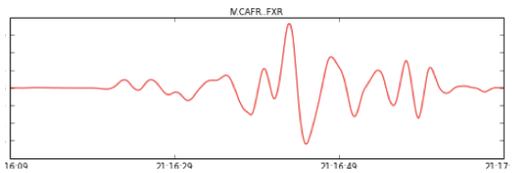


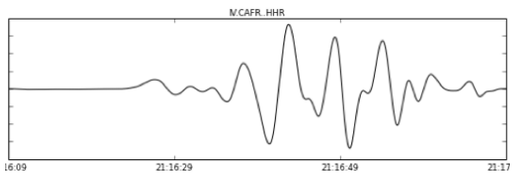
Misfit Analysis

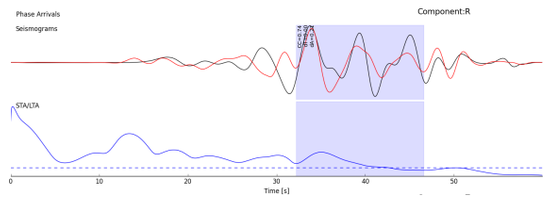


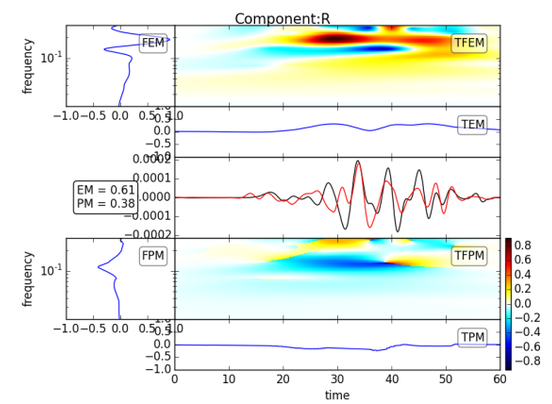


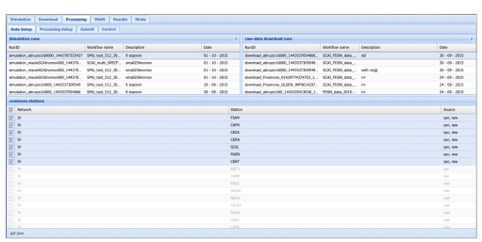
Data Download (FDSN)











dispel4py

ObsPy
A Python Framework for Seismology

VERCE

User Perspective

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu

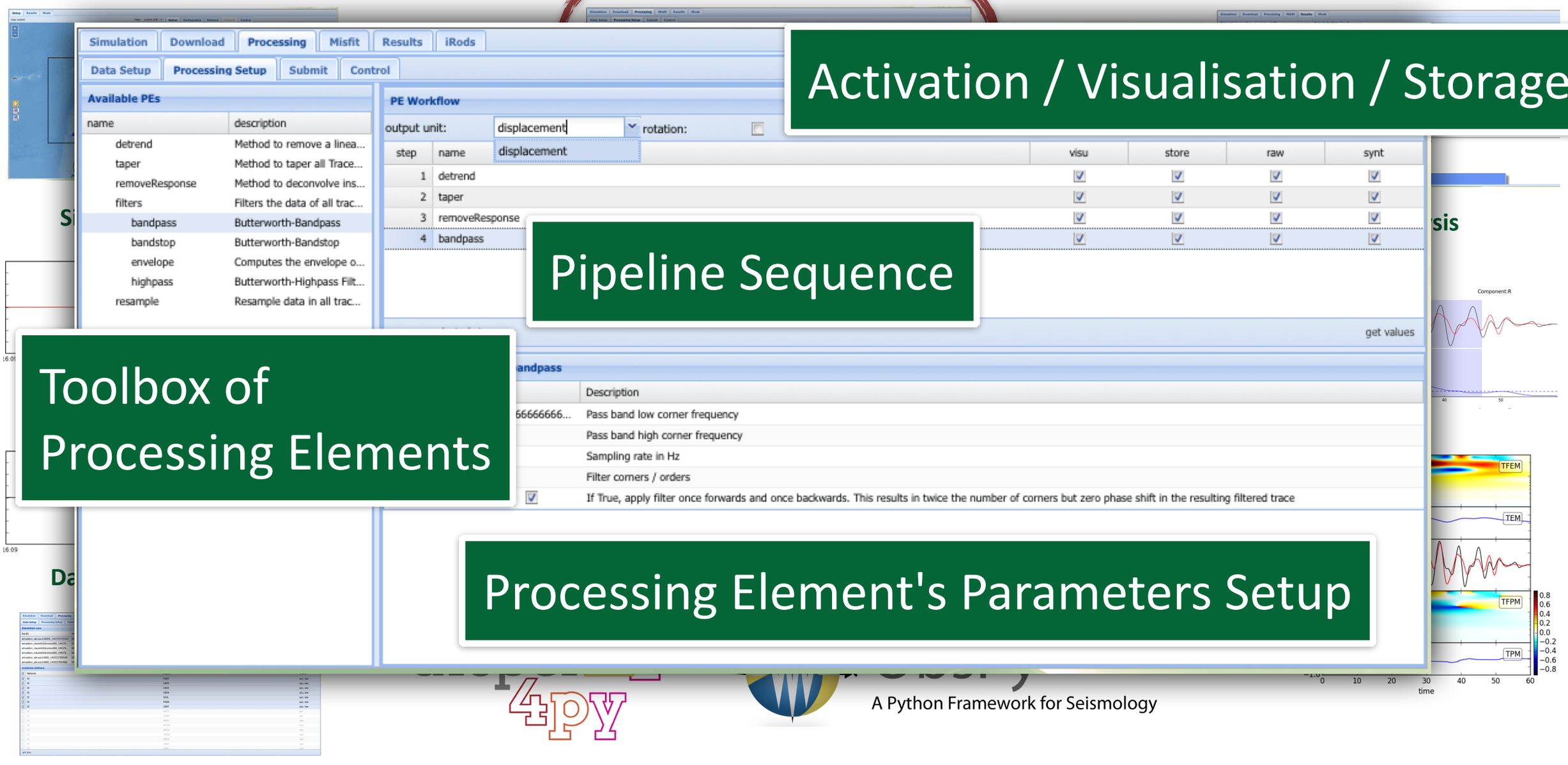
Misfit between SYNTHETICS and DATA

Activation / Visualisation / Storage

Pipeline Sequence

Toolbox of
Processing Elements

Processing Element's Parameters Setup



4py

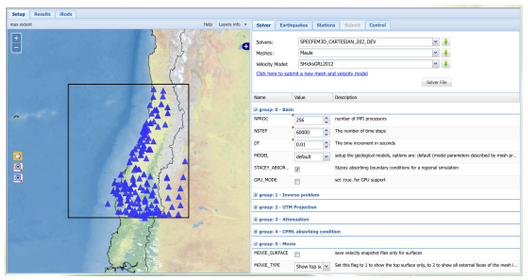
A Python Framework for Seismology

VERCE

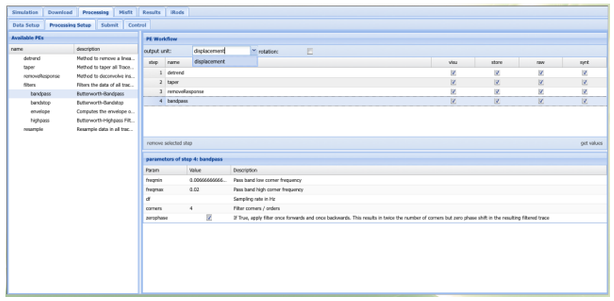
User Perspective

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu

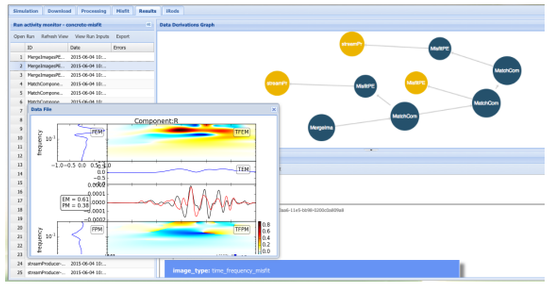
Misfit between SYNTHETICS and DATA



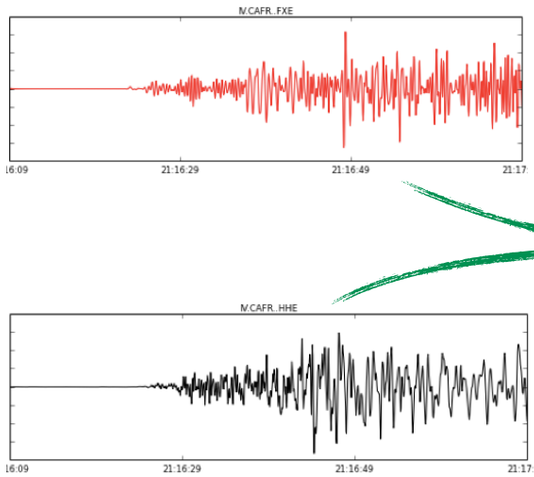
Simulated Synthetics



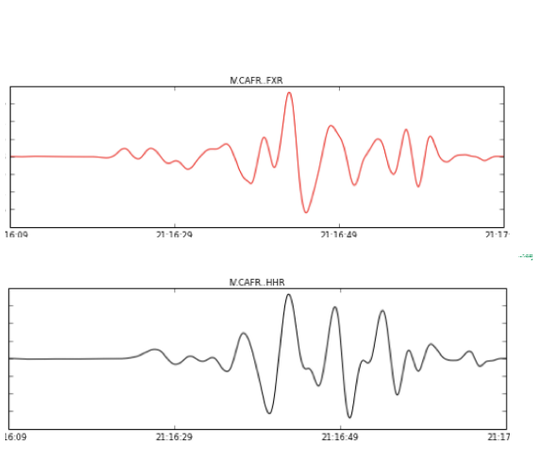
Data/Synt Processing

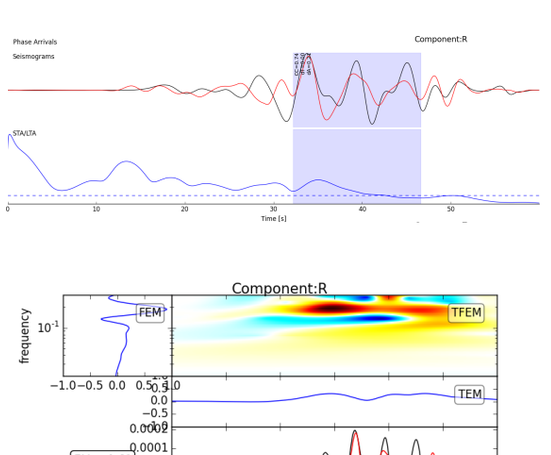


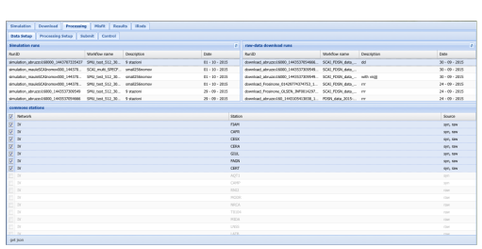
Misfit Analysis



Data Download (FDSN)







dispel4py

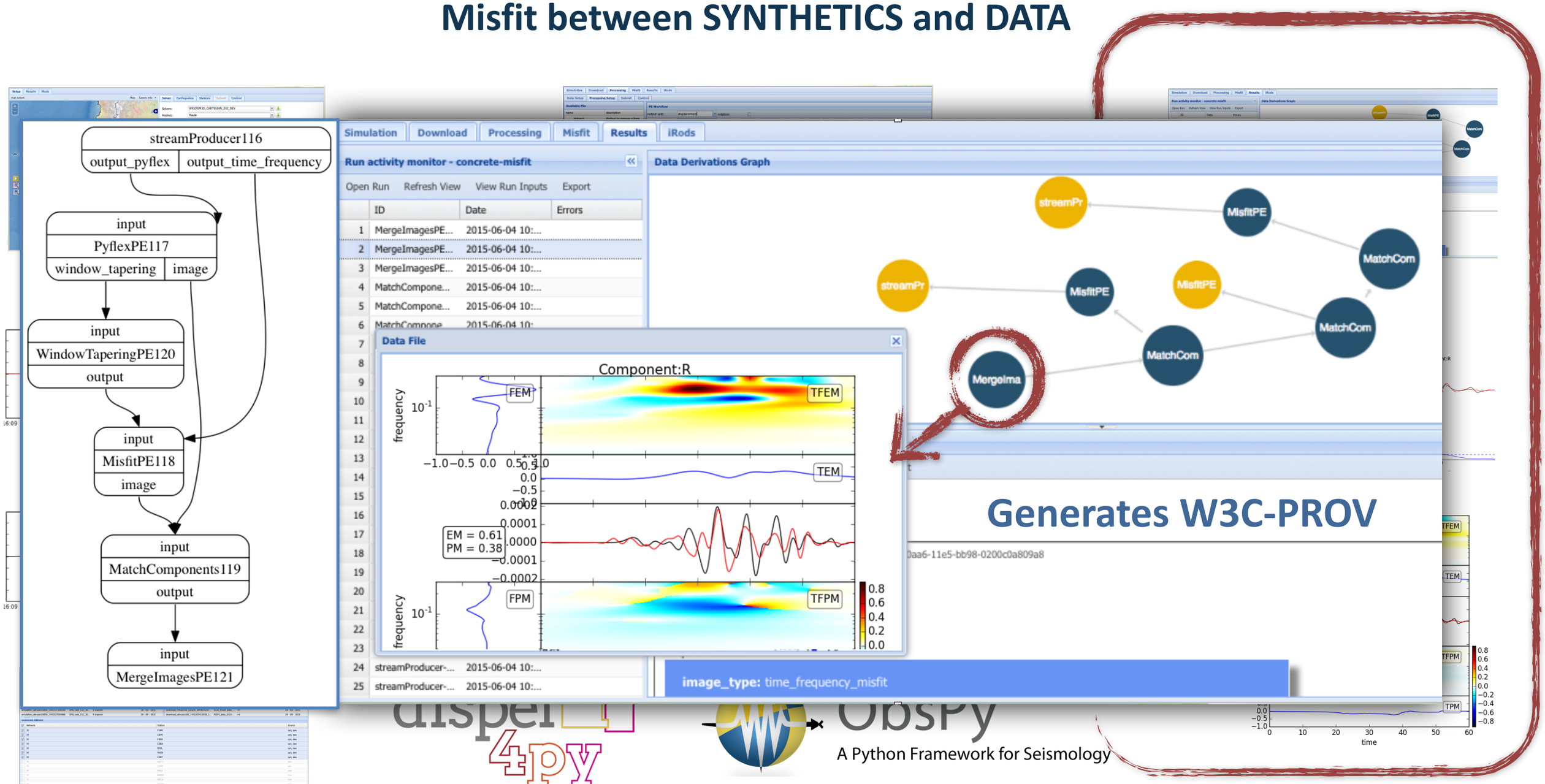
ObsPy
A Python Framework for Seismology

VERCE

User Perspective

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu

Misfit between SYNTHETICS and DATA



dispel4py

ObsPy
A Python Framework for Seismology

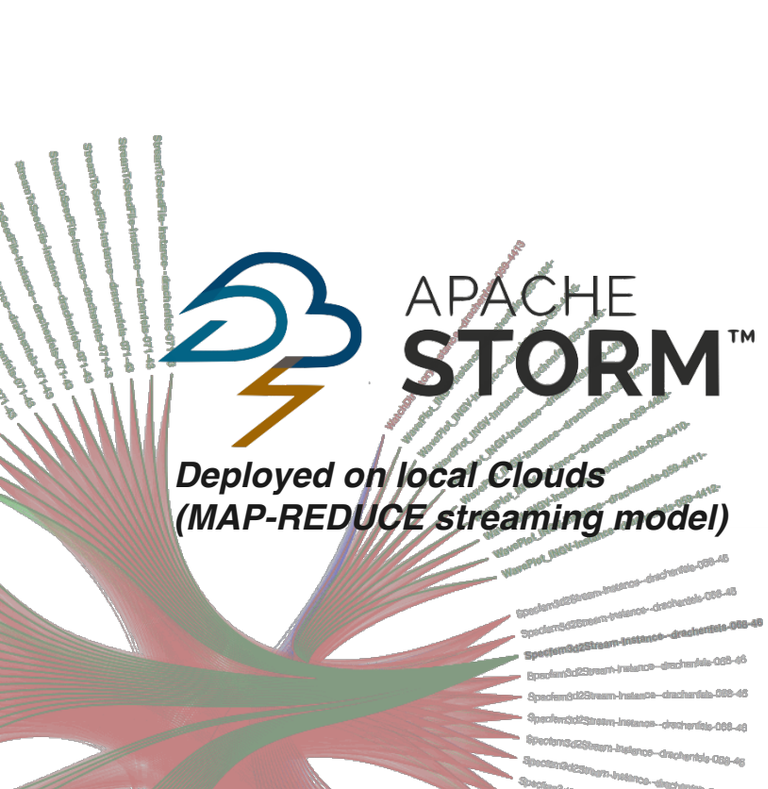
VERCE

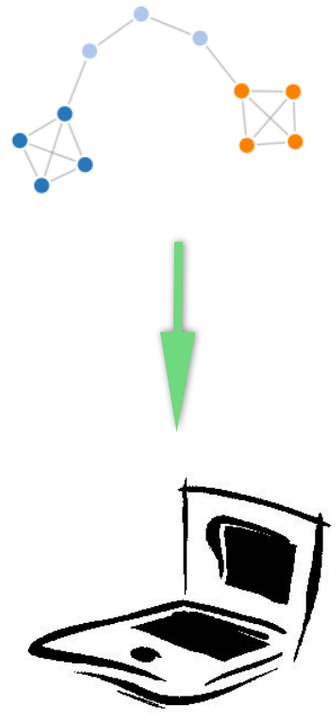
Python library used to describe **abstract workflows** for distributed data-intensive applications.

Support for composition: Single components may be defined by having their own internal workflows.

Workflows described in **dispel4Py** can be automatically executed in numerous parallel environment.

Docker containers available supporting multiple execution environments (MPI, SharedMemory) and the integration with other workflow systems (eg. Pegasus)





MPI

EGI

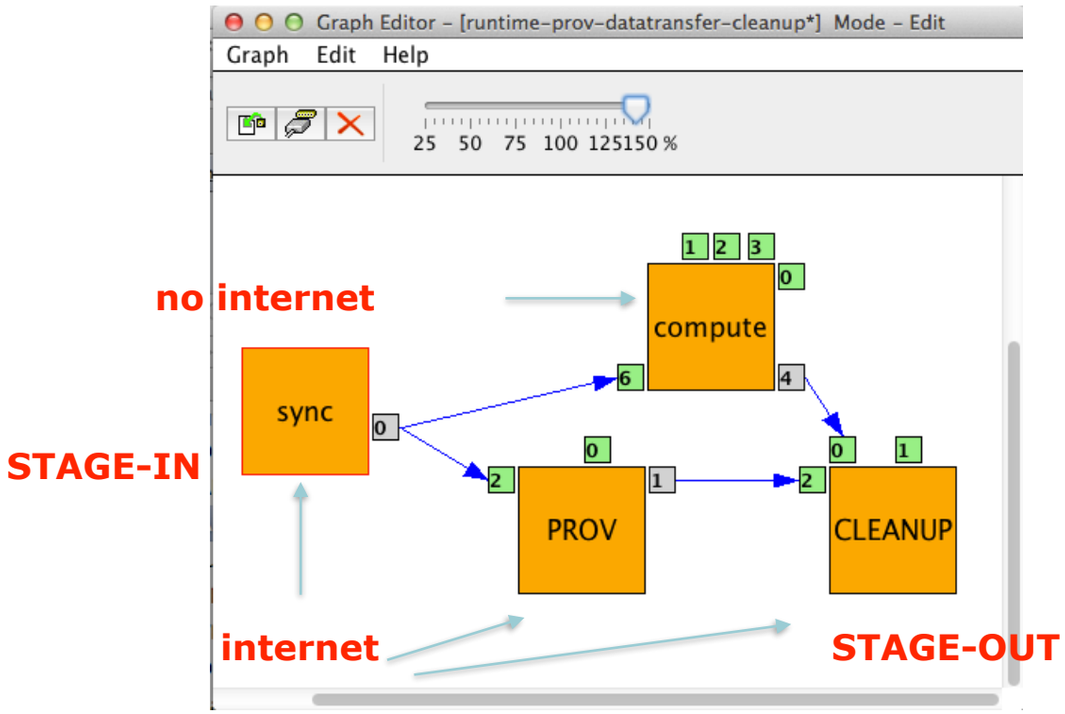
PRACE

Platform's Control Workflows (WS-PGRADE)

HPC Control Workflow

GLOBUS & UNICORE jobs to Grid and HPC Clusters.

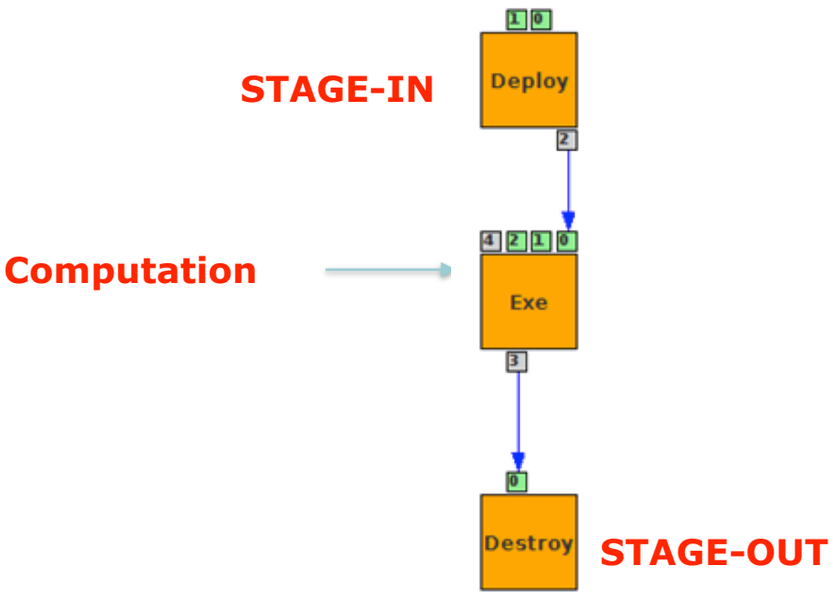
Frequent limitations in connectivity require special serial jobs for runtime data and metadata management (Sync, PROV, Cleanup)



Cloud Control Workflow

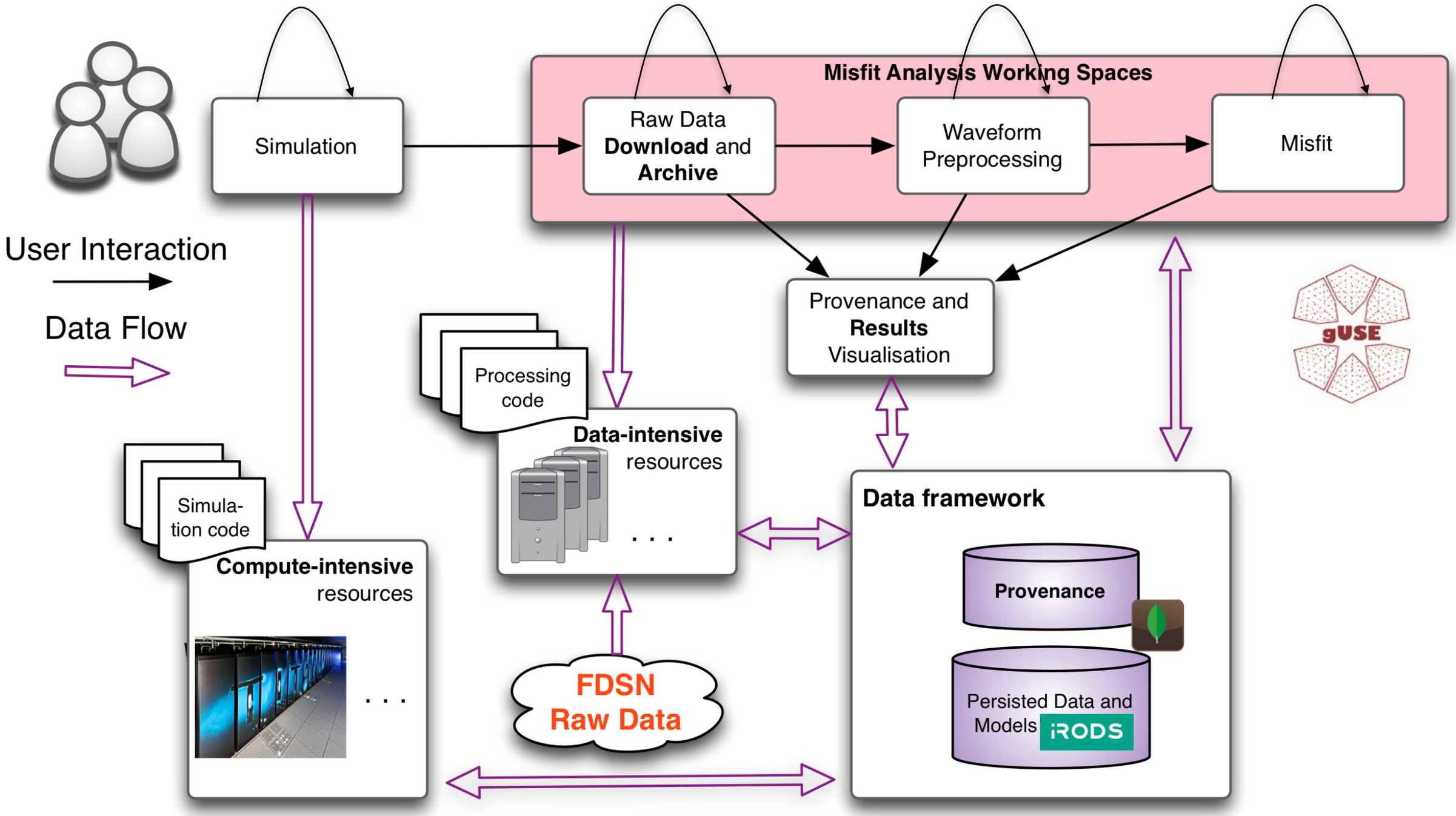
(Deploy - Compute - Destroy)

Less limitations in outbound connectivity, easier control workflow.



System & E-Infra Perspective

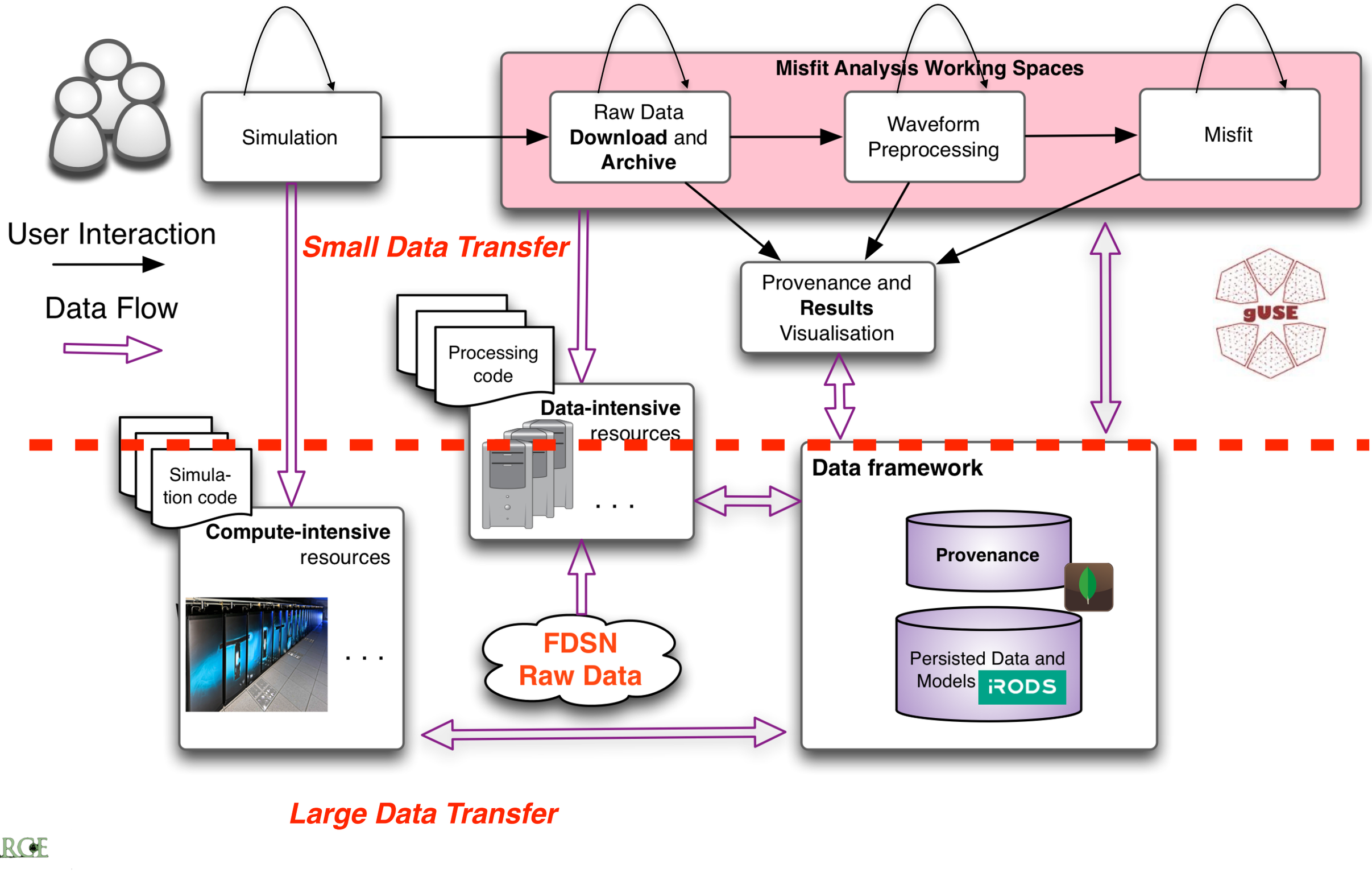
Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu



VERCE

System & E-Infra Perspective

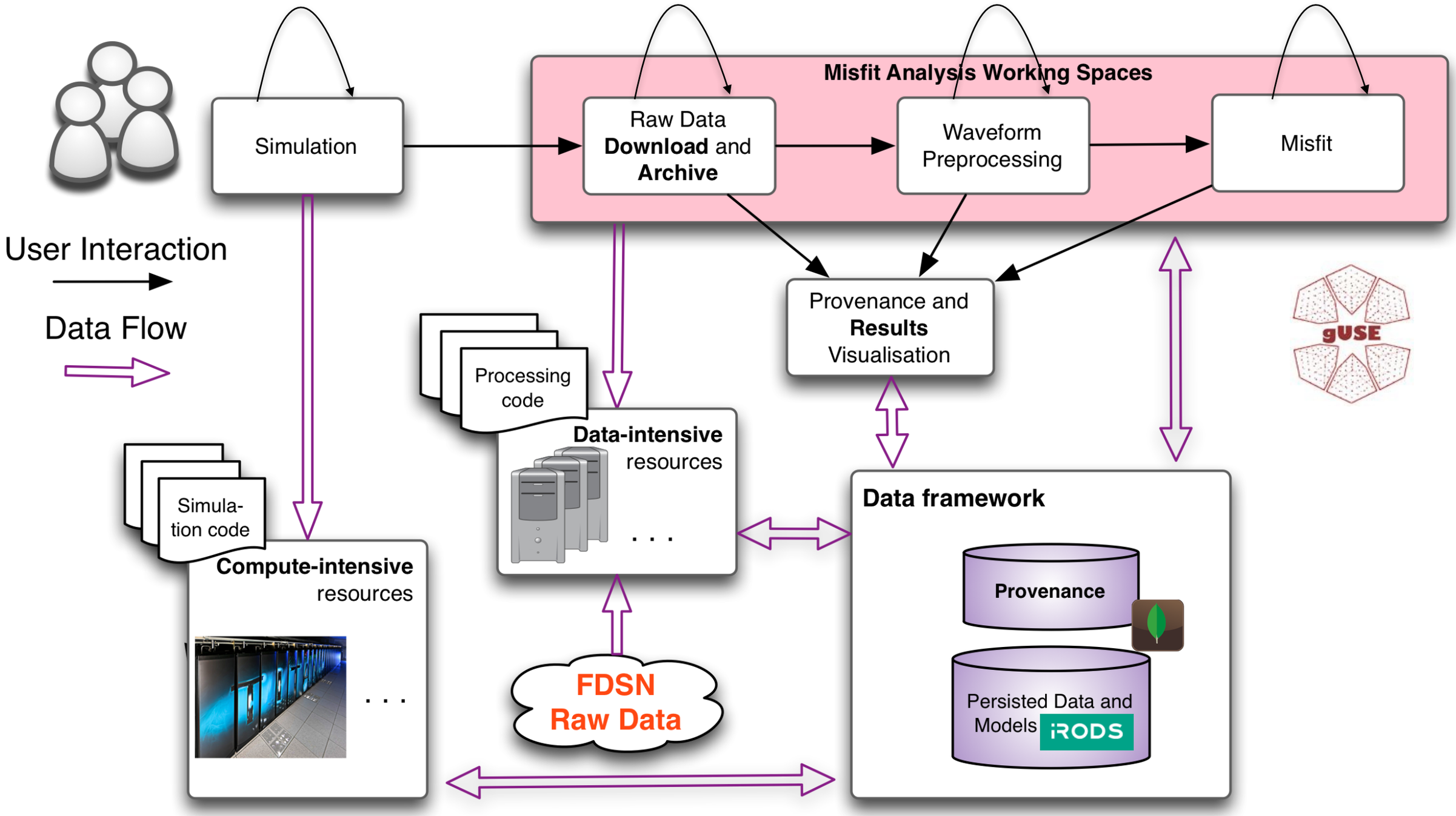
Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu



VERCE

System & E-Infra Perspective

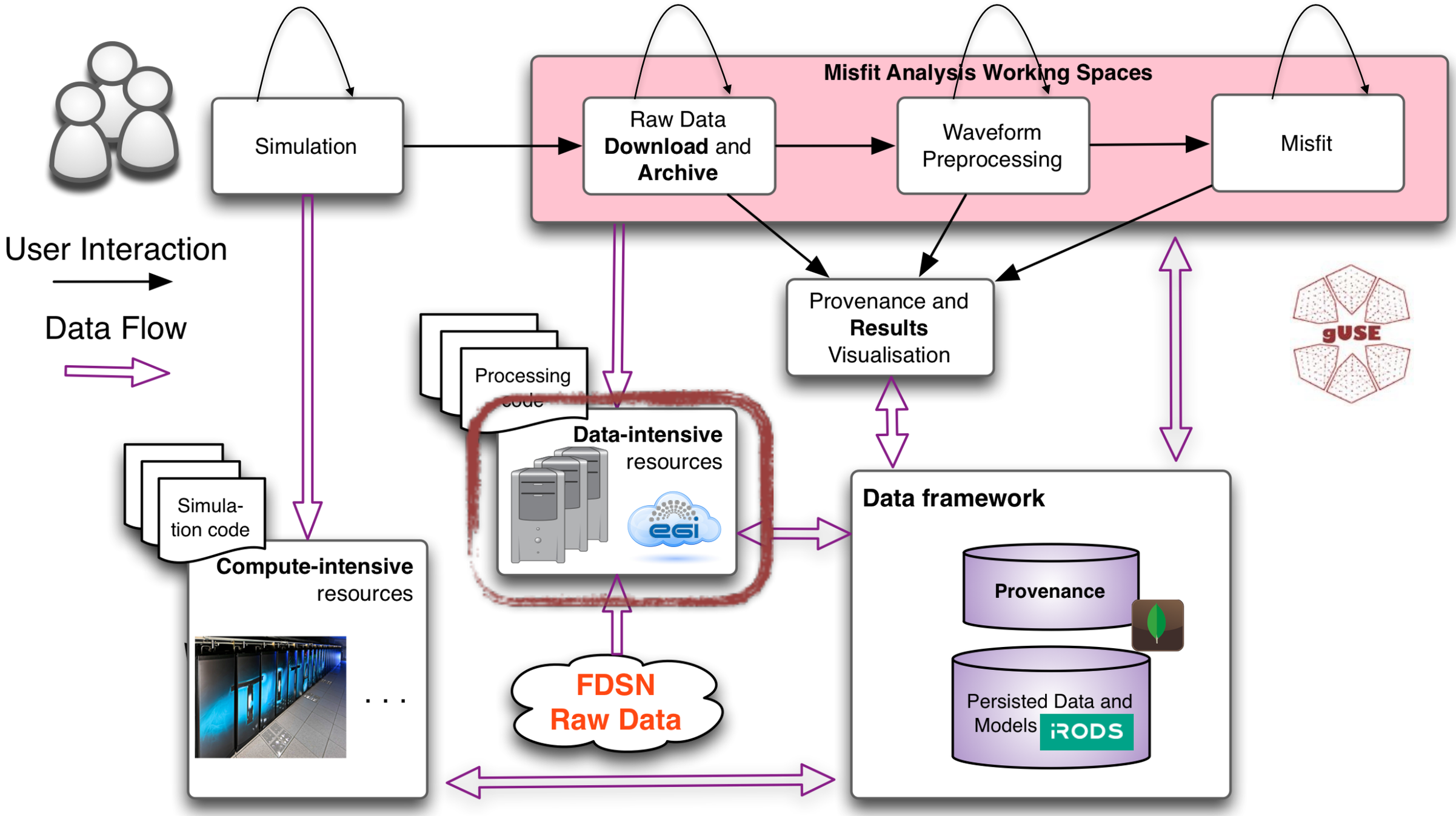
Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu



VERCE

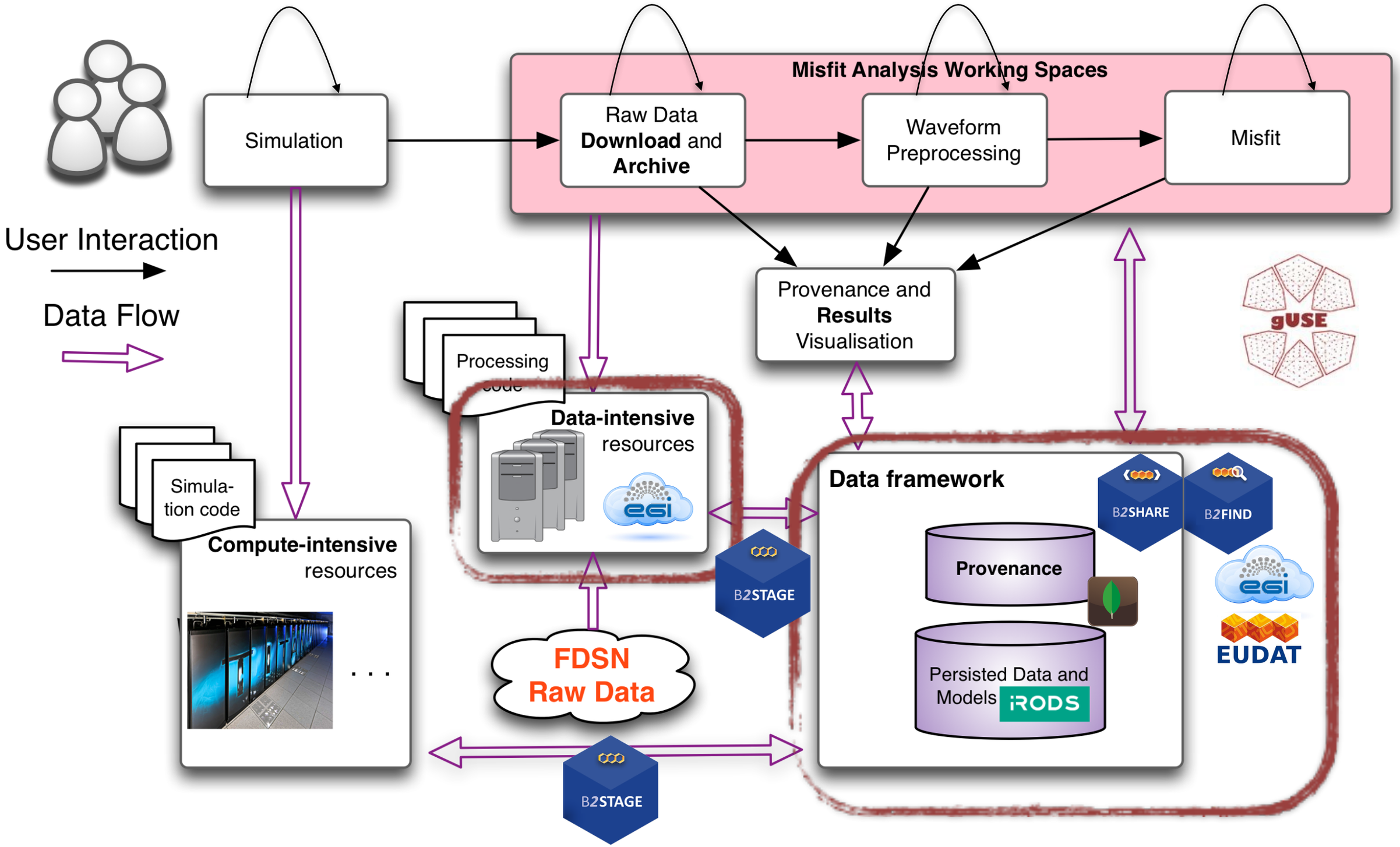
System & E-Infra Perspective

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu



System & E-Infra Perspective

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu

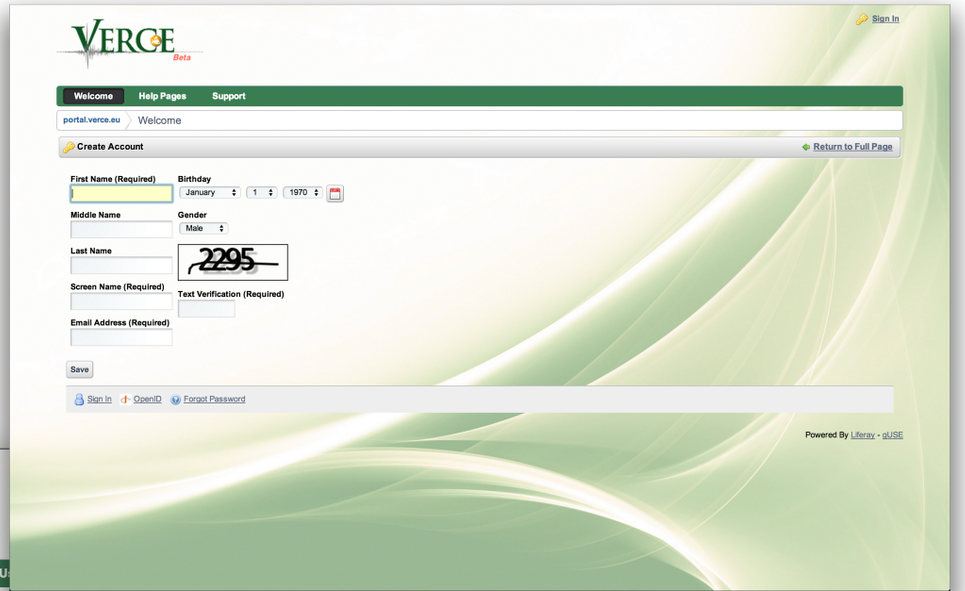


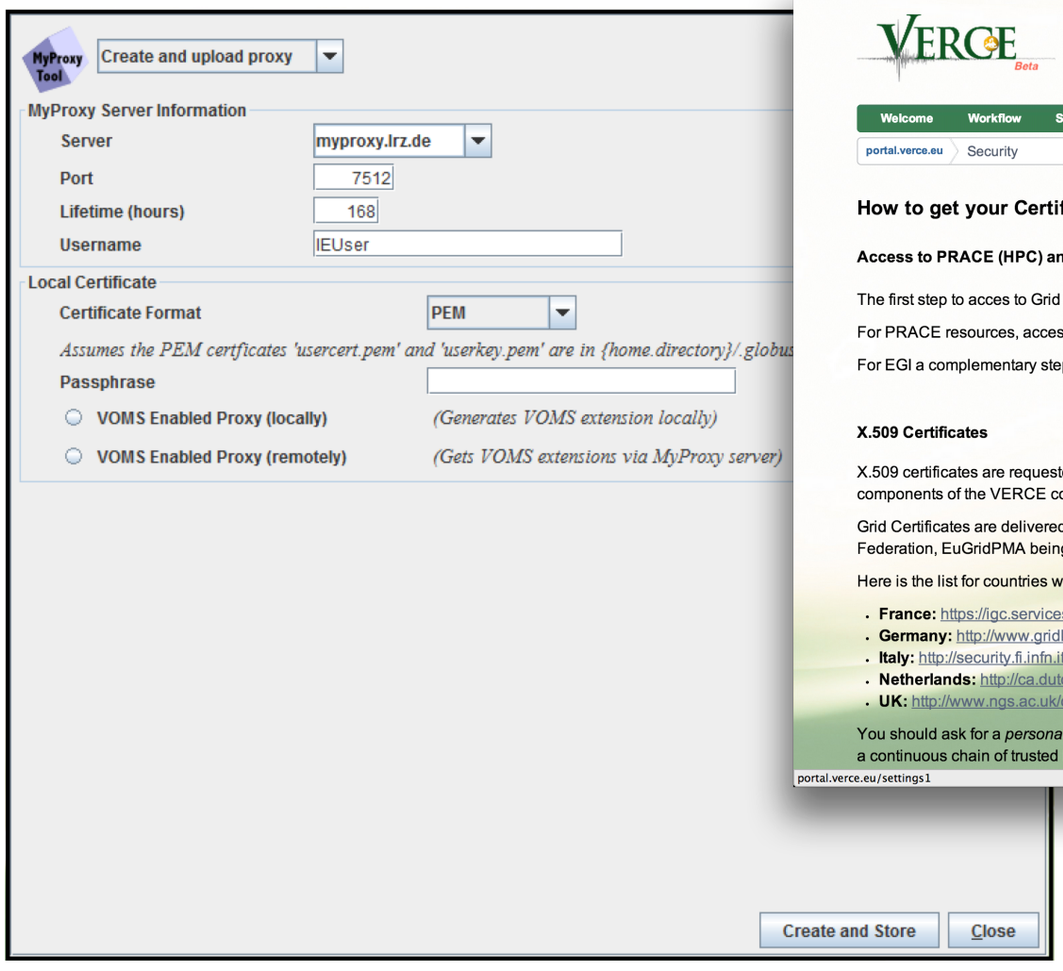
VERCE

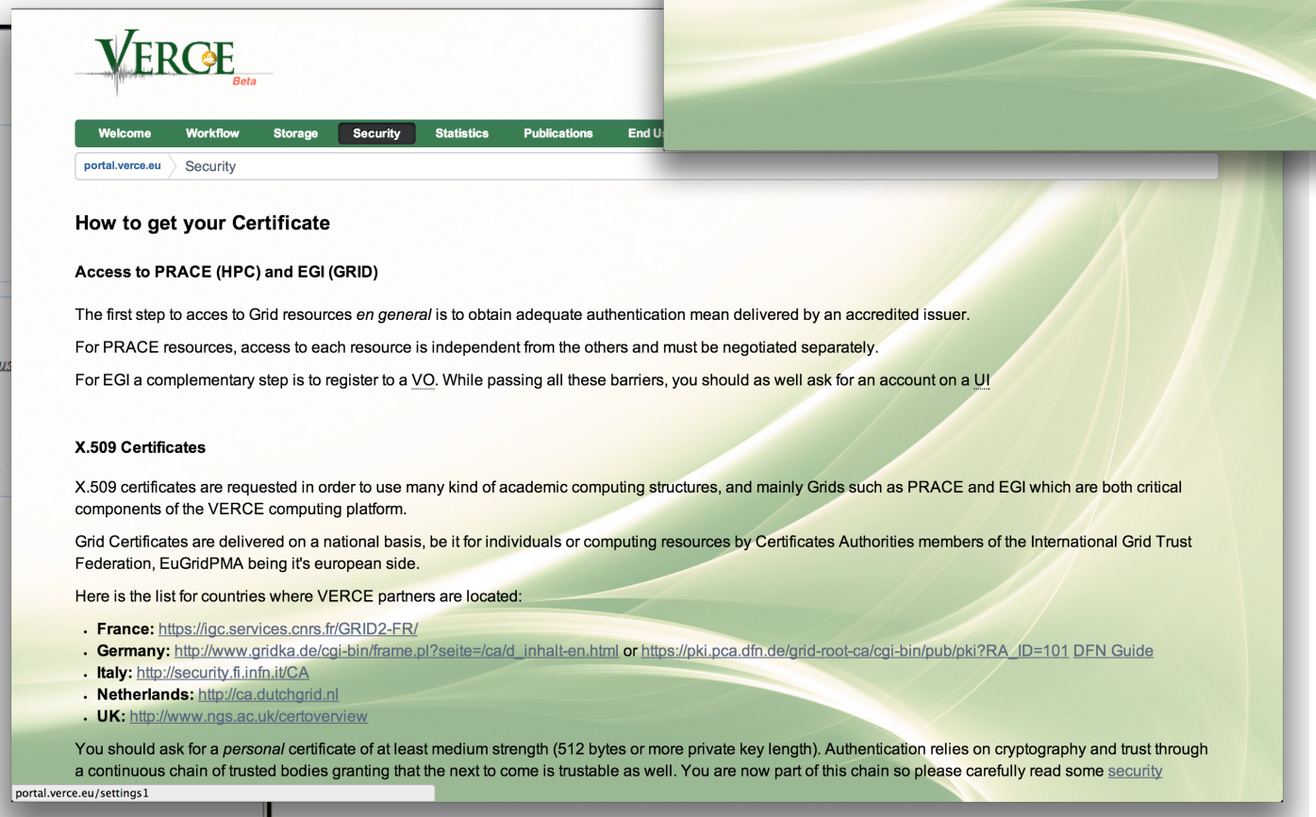
ACCOUNT creation and GRID Authentication

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu

- Anyone can sign up and browse data, but real use of the scientific tools will require a GRID Certificate
- Extensive documentation for acquiring certificate
- Clear procedures for infrastructure access
- Built in tool for proxy management







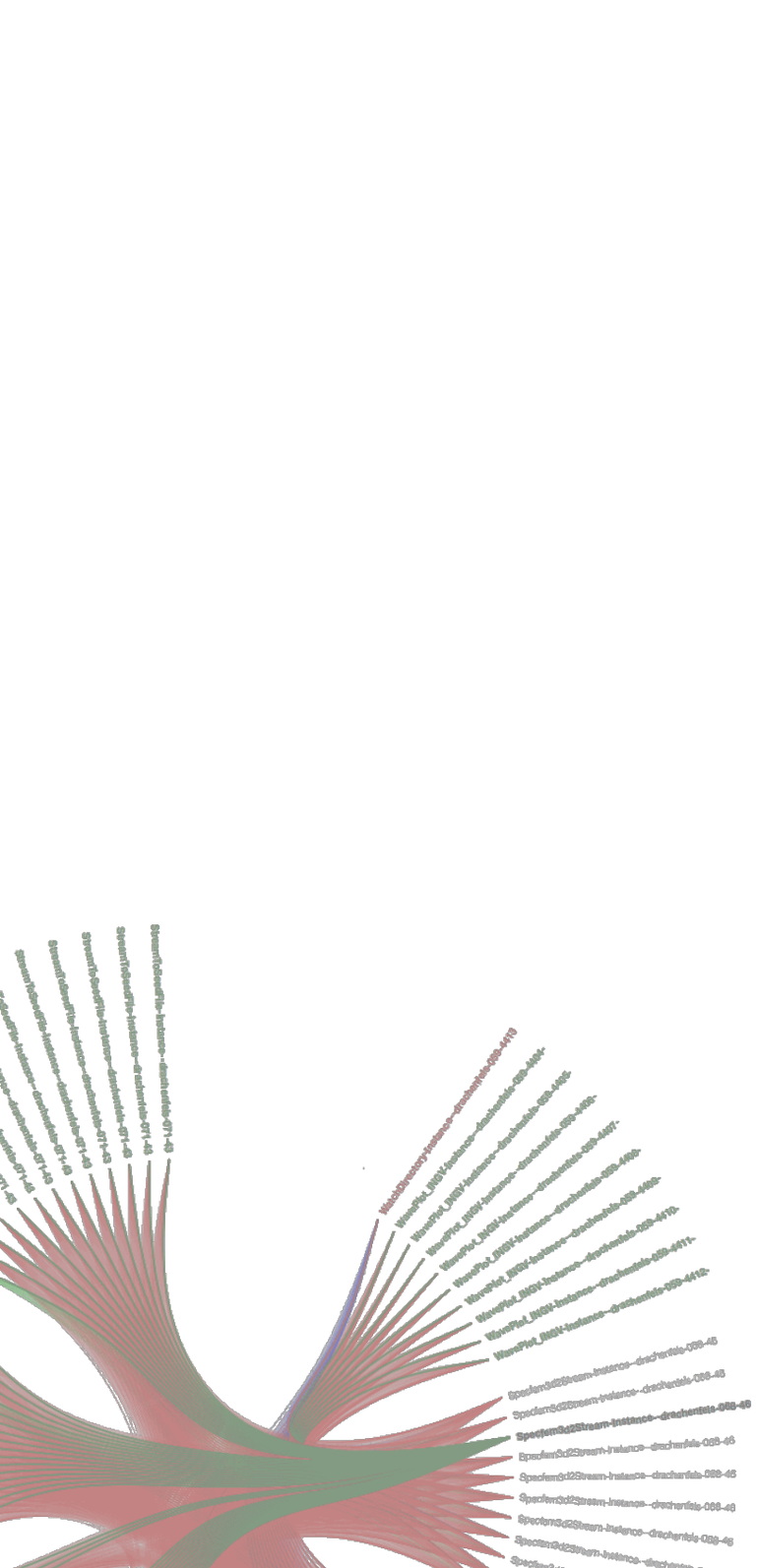
VERCE

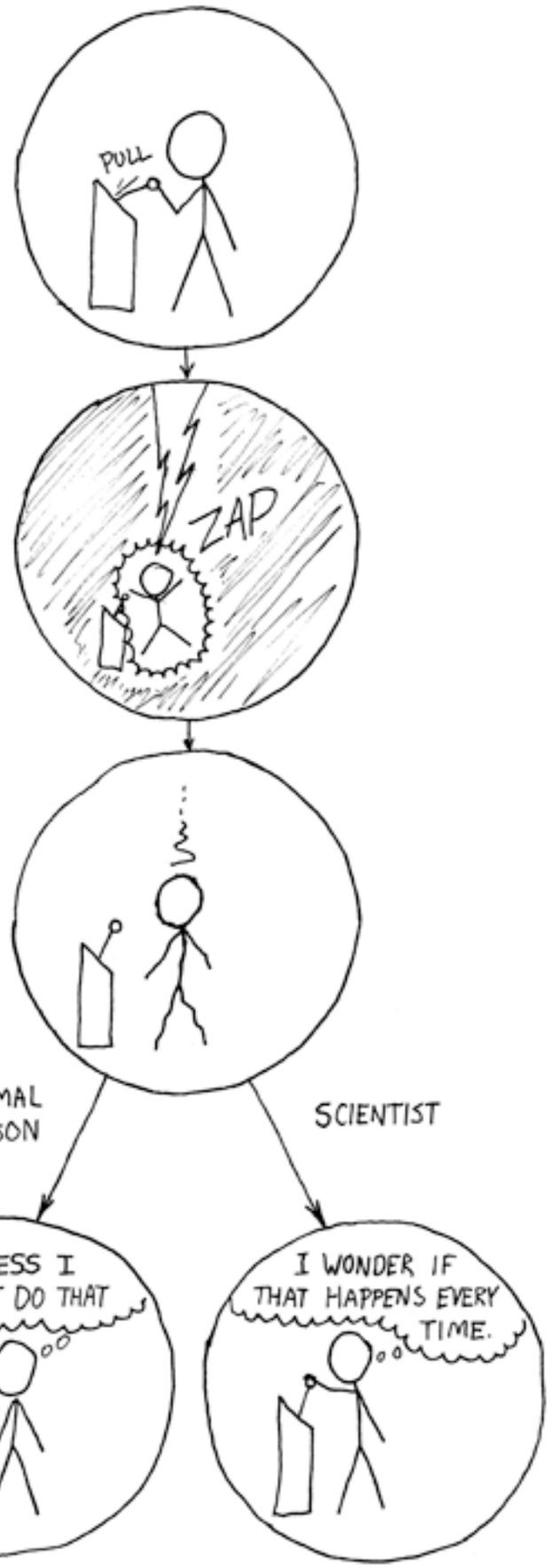
ACCOUNT creation and GRID Authentication

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu

- Anyc
  tools
- Exter
- Clea
- Built

Reproducible Science



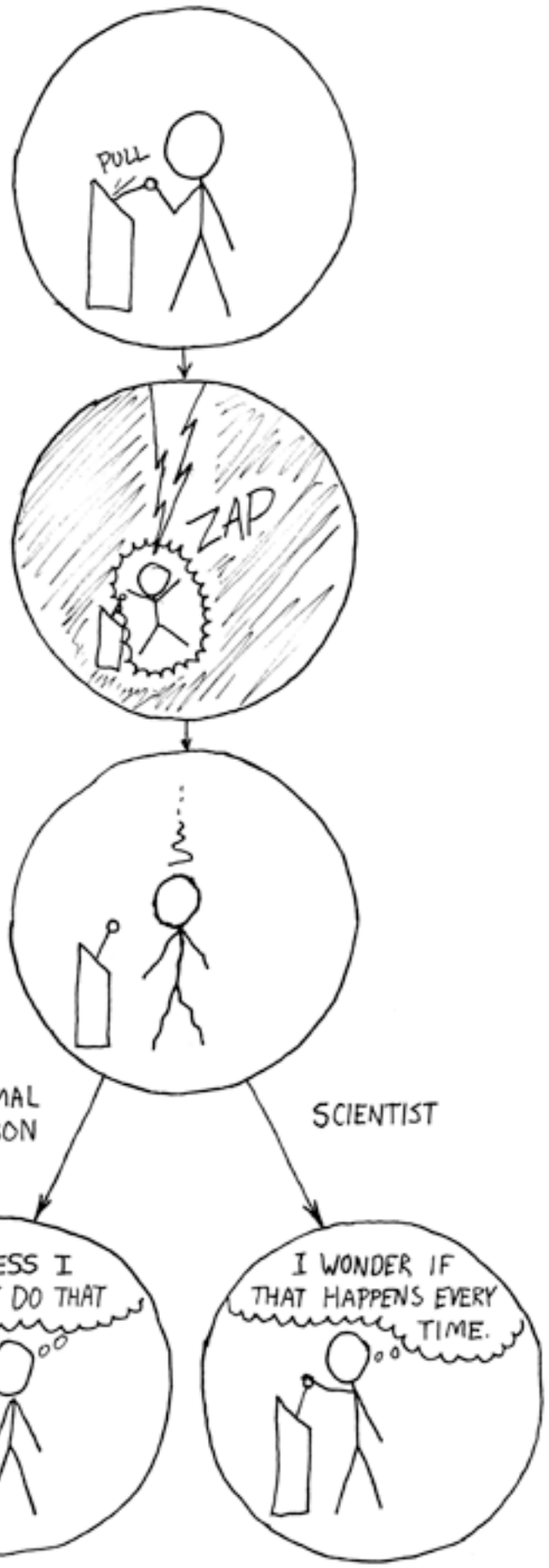


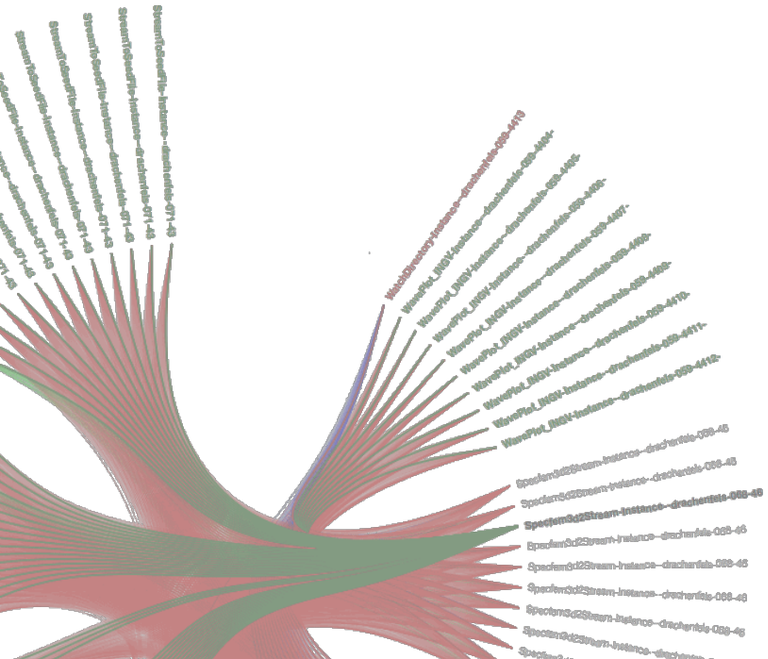
Reproducible Science

What does this suggest?

Scientists aren't normal..

Reproducibility is masochism..





Reproducible Science

What does this suggest?

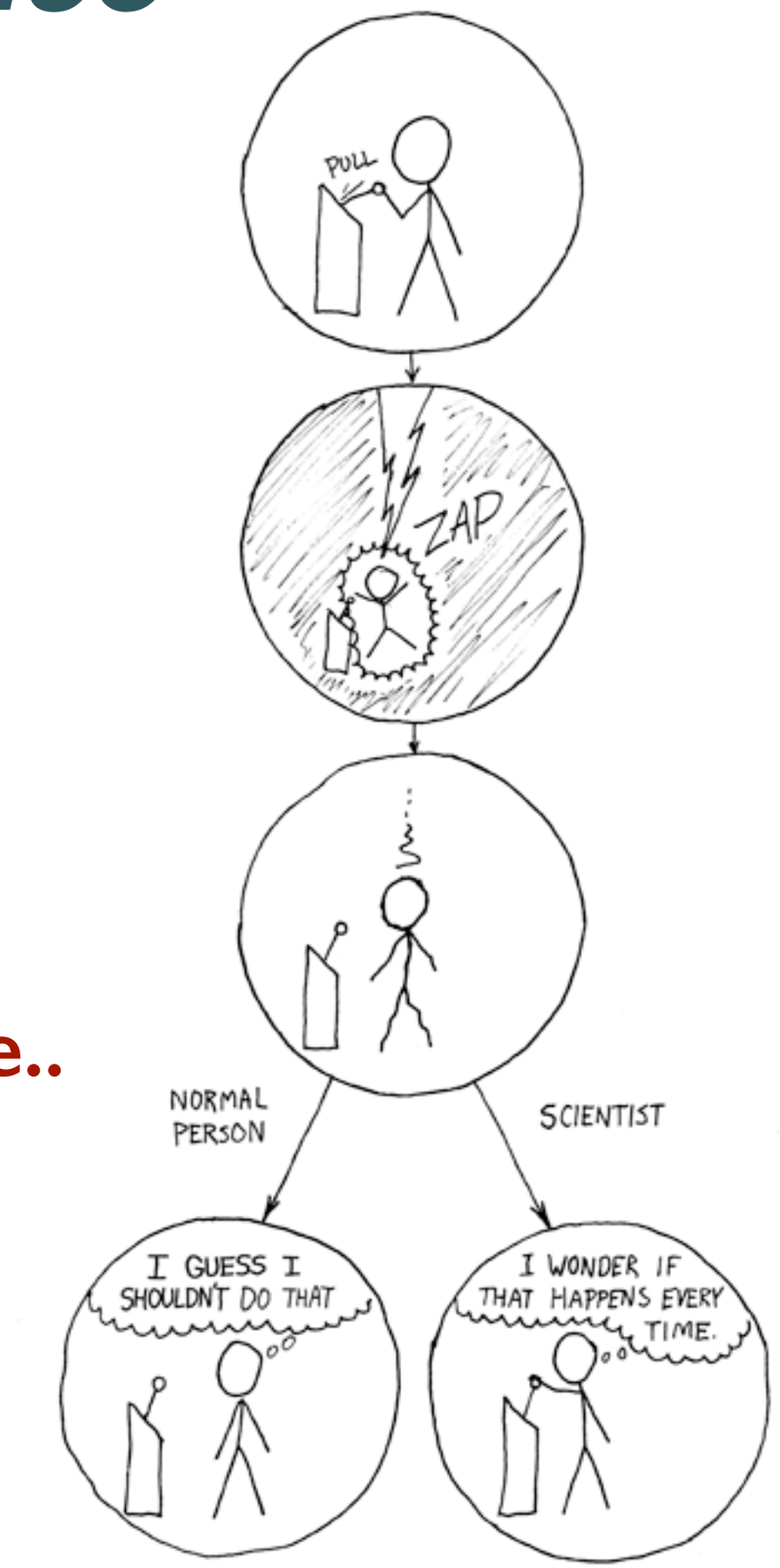
Scientists aren't normal..

Reproducibility is masochism..

Reproducibility is fundamental but difficult to achieve.

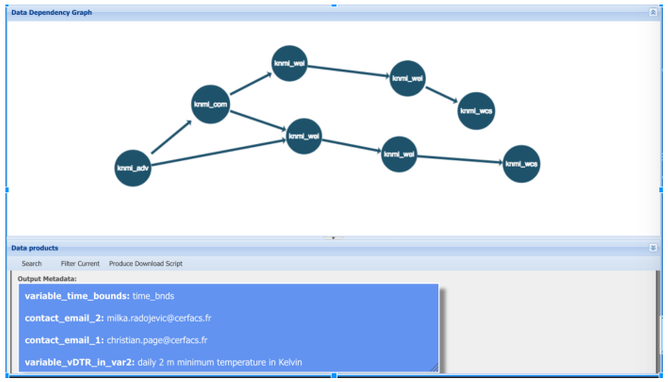
Not always convenient/possible..

We need more than just rerun

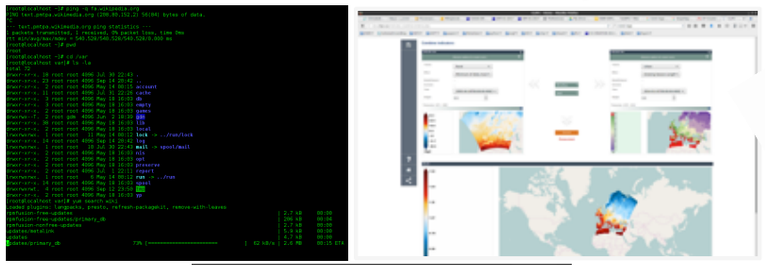


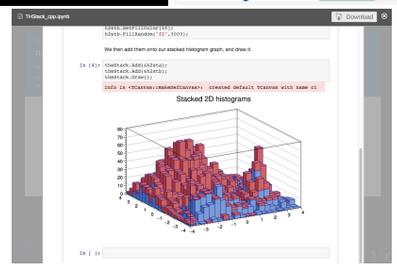
Reproducibility Cycle(s)

Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu



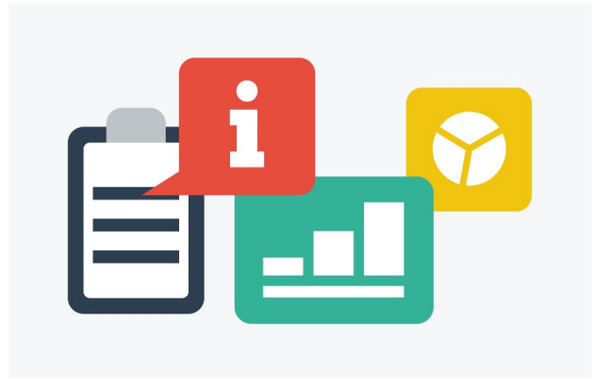
Validate





Procedures / Tools / Development

Analyse / Discover



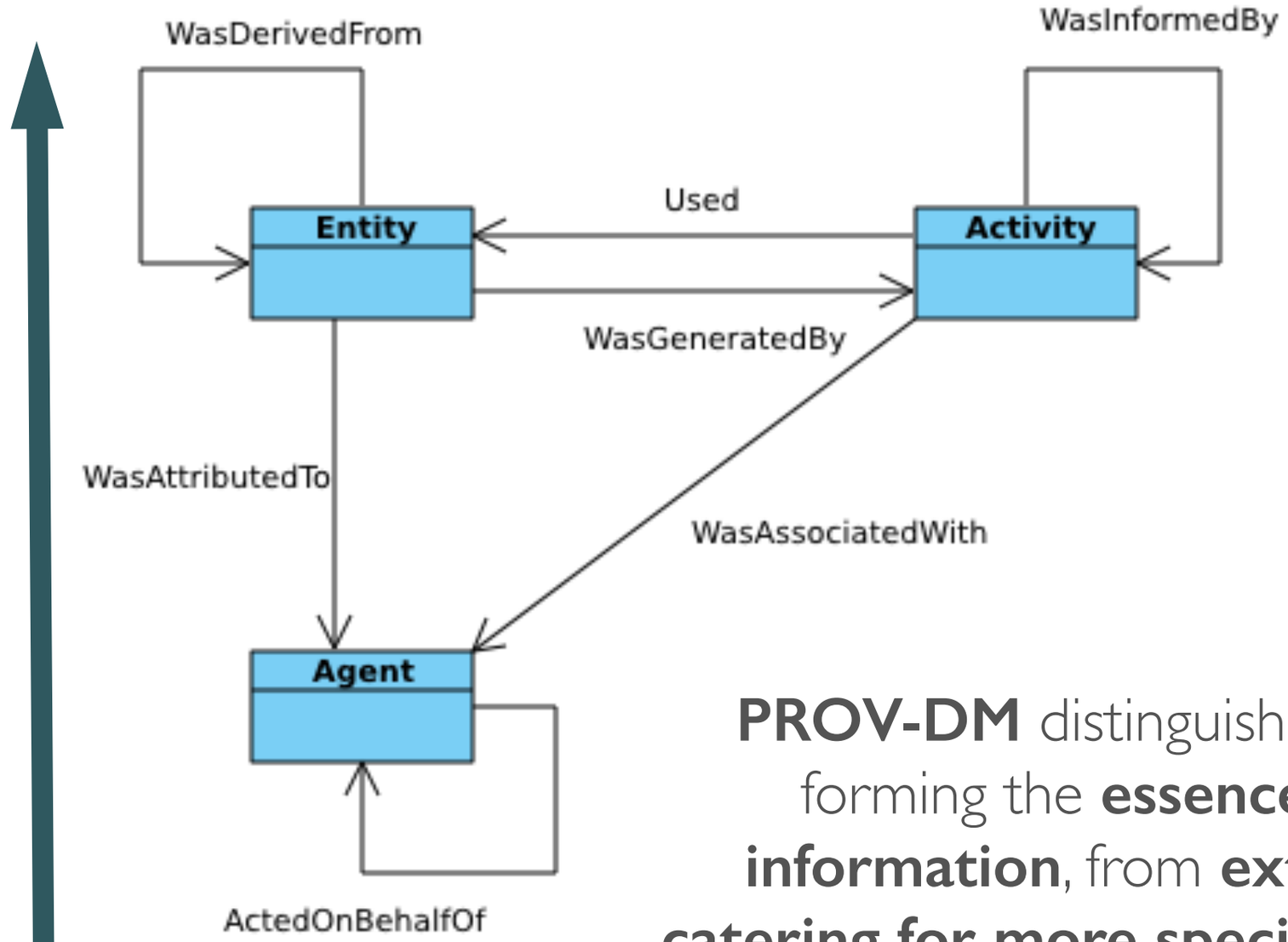
Report / Outreach / Preserve

Repeat / Verify

VERCE

Provenance Model

Data-Lineage → ***Knowledge***



W3C PROV

PROV-DM distinguishes **core structures**, forming the **essence of provenance information**, from **extended structures catering for more specific** uses of provenance.

Enactment ← ***Machinery & Data***

Challenges

*"**Automated** system should provide **support** for a **consistent and effective acquisition** of provenance metadata."* *[A. Misra] [I. Foster.]*

Expert users are part of the process - configuration and contextualisation:

"How much domain metadata should be contained?"

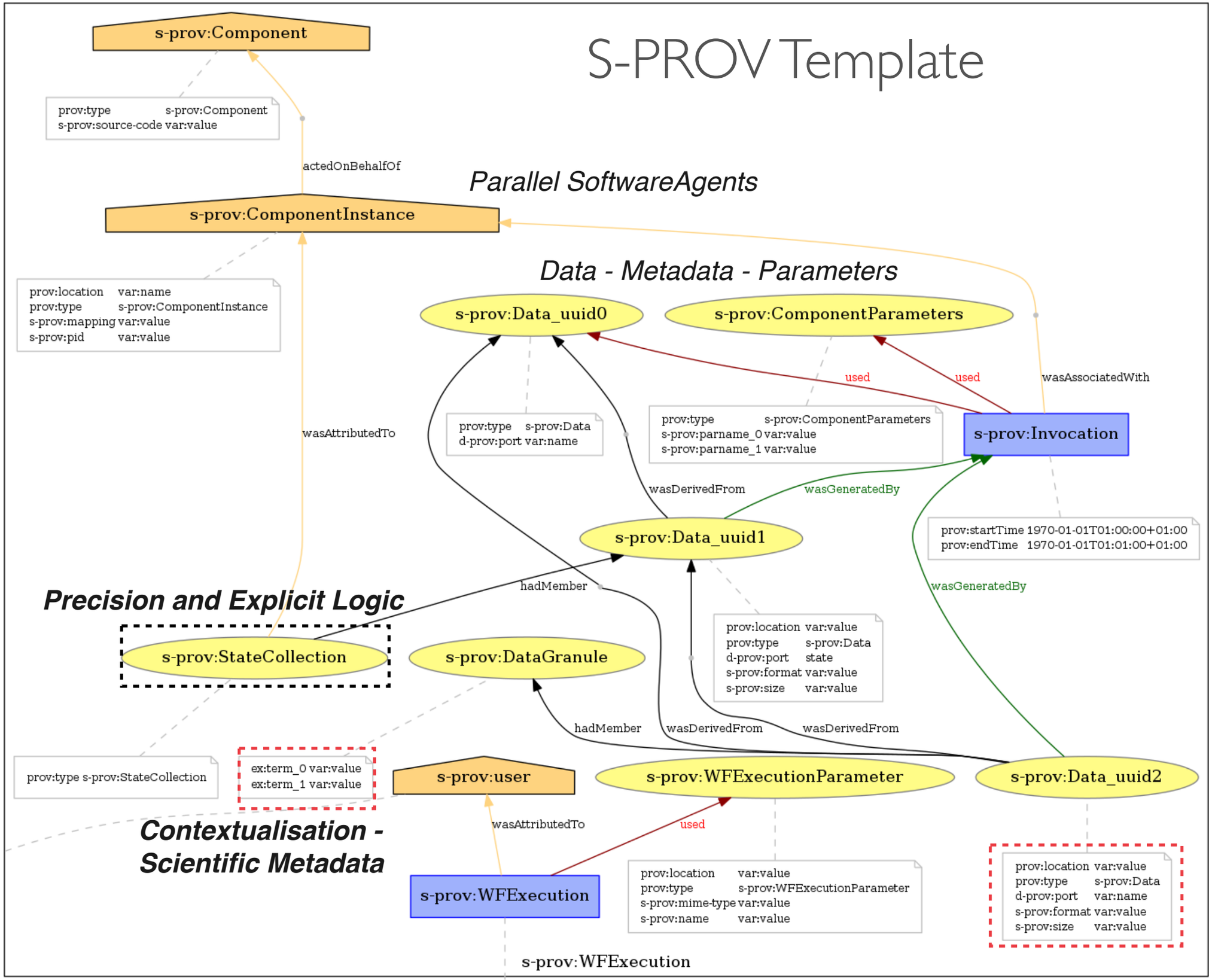
Scale of the provenance records:

"What level of granularity is needed to describe provenance of complex objects?
"Manage the scale of the provenance records to be recorded and processed"

Multiple levels of understanding:

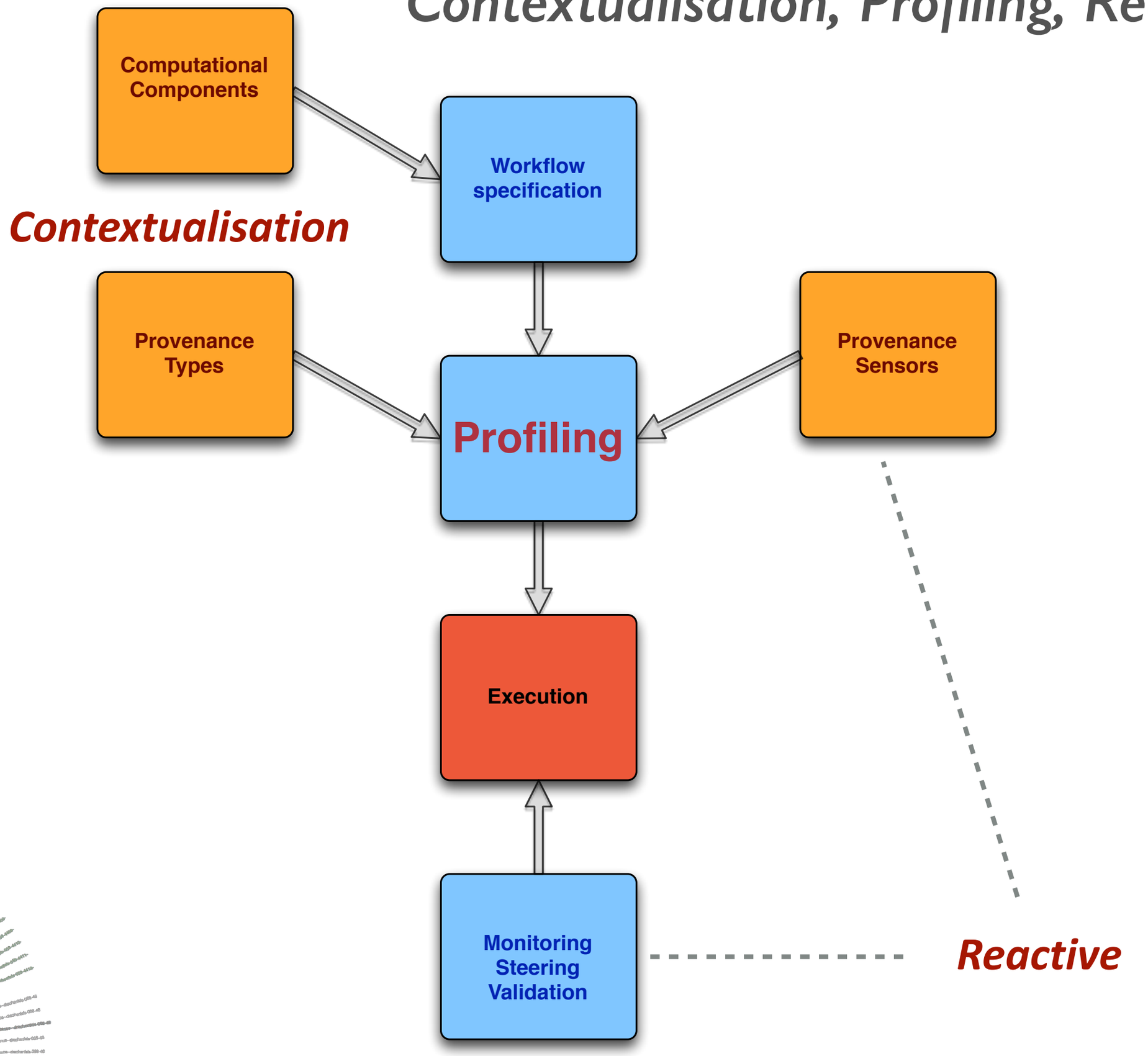
"Provenance at different levels of abstraction, extract high-level summaries of provenance from detailed records."

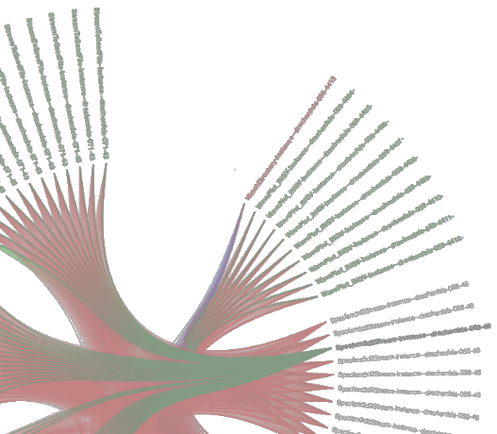
S-PROV Template



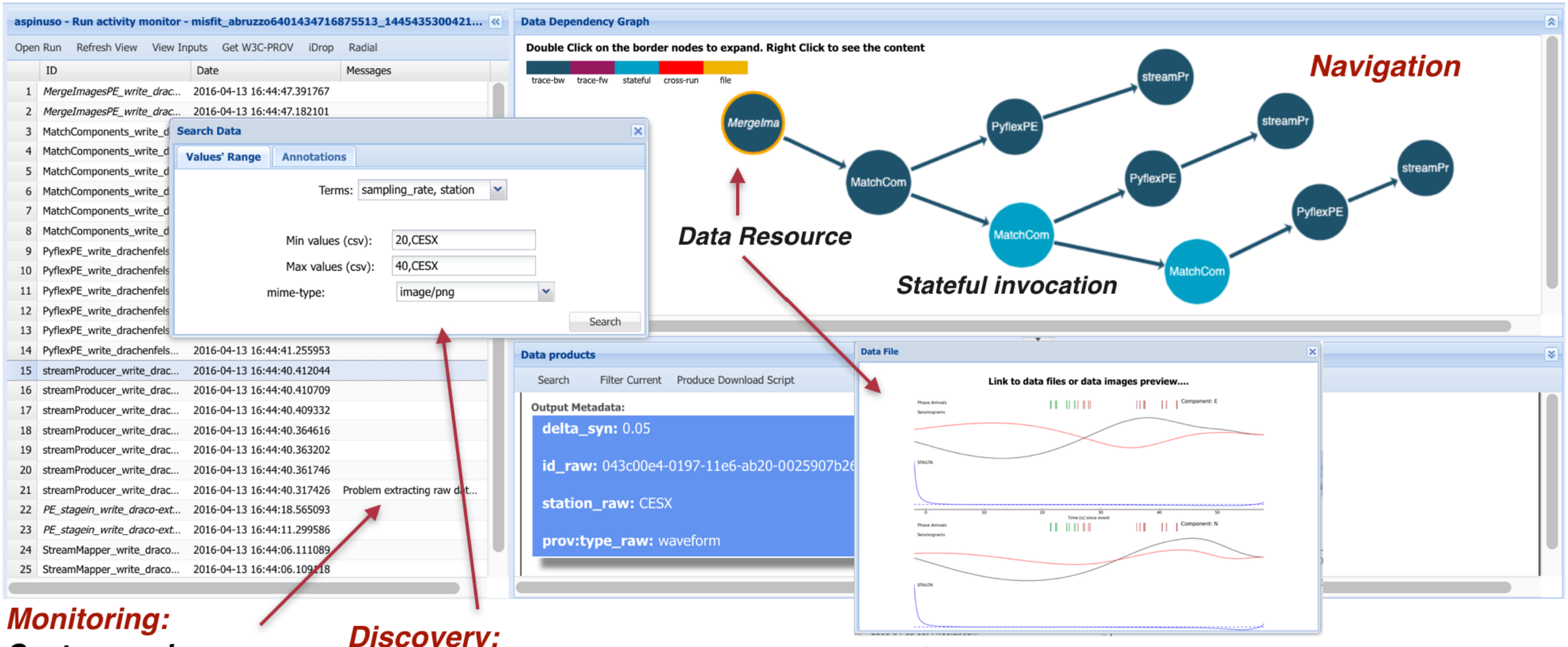
Active Provenance framework

Contextualisation, Profiling, Reactive





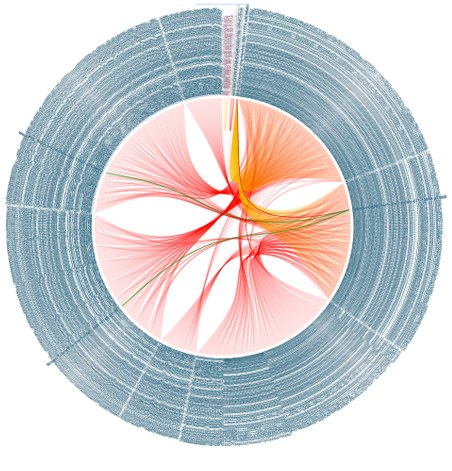
Tools: S-ProvFlow - Reproducibility as a Service

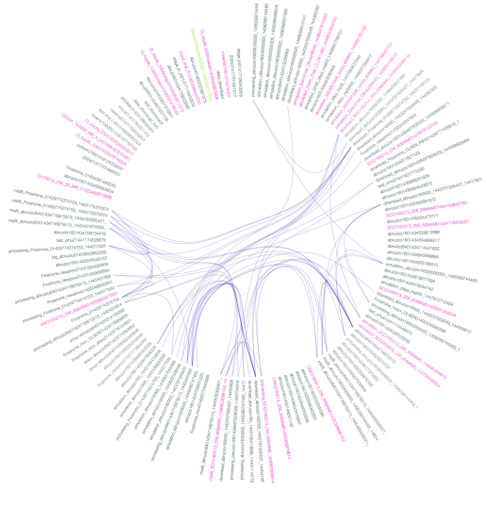


Monitoring:
System and user Messages

Discovery:
Search on contextual metadata (within and across runs)

Preview and Download





Visual analytics techniques on provenance

Perspectives on:
- data-intensive processes
- users and applications interactions
- data-reuse
- exploitation of resources..

Visual Analytics: Computation Distribution and execution modes

Edges: Data Transfer (bytes)

100 500 1000 5000 10000 >10000

Nodes: Mappings

simple multi mpi

Level: **iterations**

Select Level

Iteration Range: minidx 0 maxidx 10

Grouping: **worker**

Select Grouping

drachenfels-login

drachenfels-061

drachenfels-060

drachenfels-070

drachenfels-071

Visual Analytics: Collaborative Perspectives

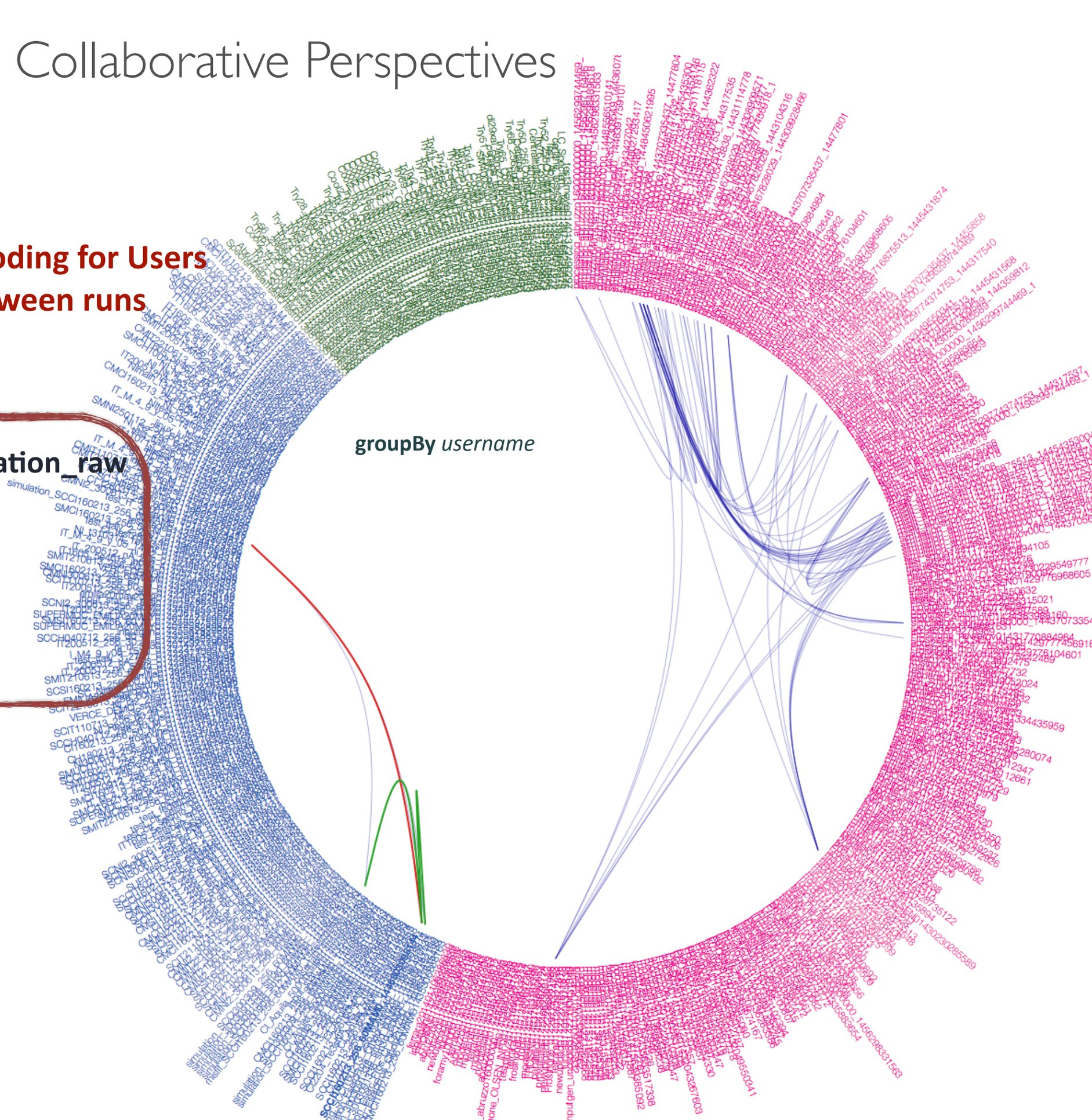
Data-reuse across Users

Vertex: Runs with color-coding for Users
Edges: Data exchange between runs

keys:**magnitude**,**station**,**station_raw**
minvalues: **4**,**CERA**,**CERA**
maxvalues: **6**,**CERA**,**CERA**

users: **aspinuso**,**fmagnoni**
cerlane

groupBy *username*



Visual Analytics: Infrastructure Perspectives

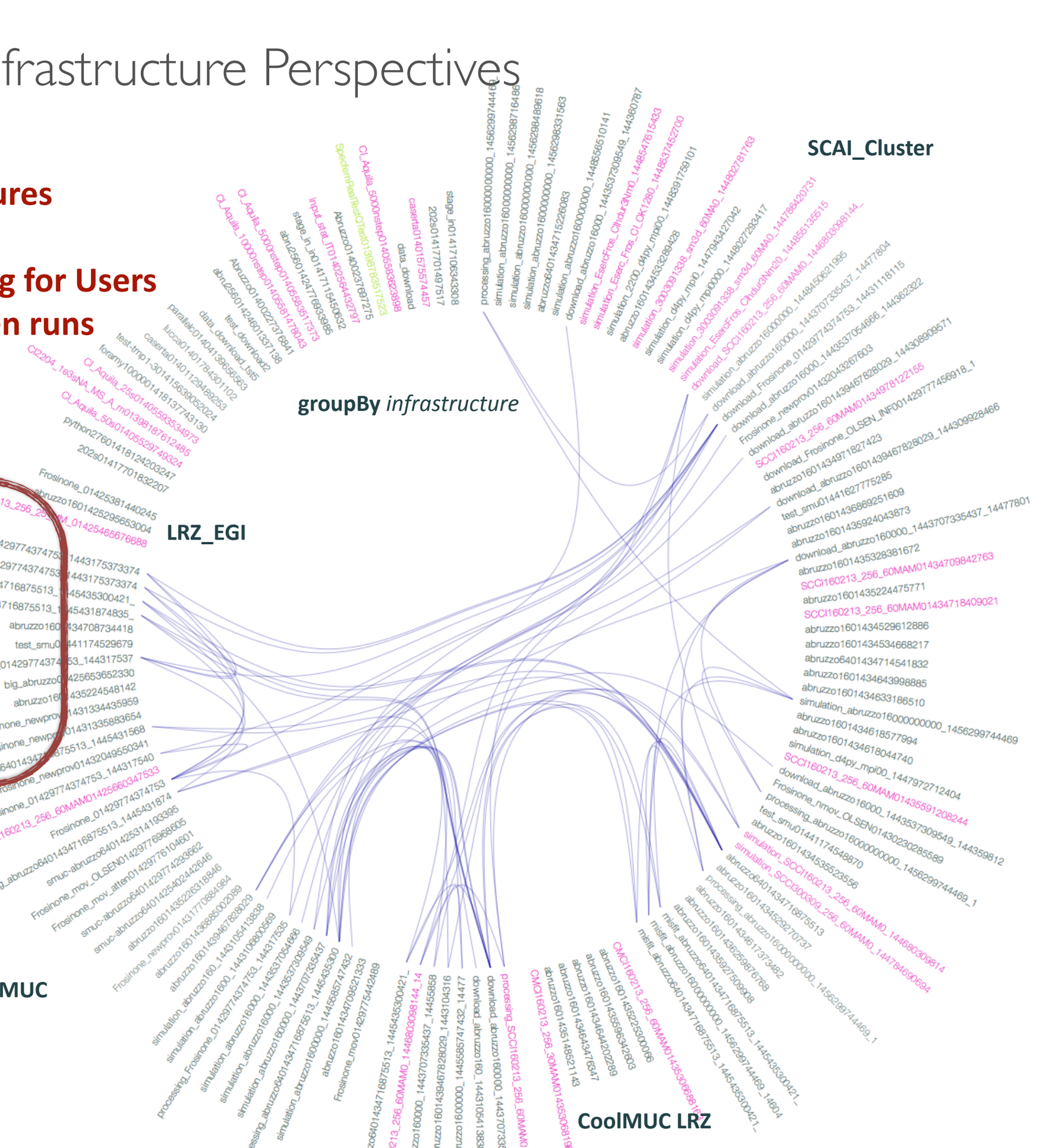
Data-reuse across Infrastructures

Vertex: Runs with color-coding for Users
Edges: Data exchange between runs

keys:**magnitude,station_raw**
minvalues: **4,CERA,CERA**
maxvalues: **6,CERA,CERA**

users: **aspinuso,fmagnoni cerlane**

groupBy *infrastructure*



jupyter ProvenanceDemo Last Checkpoint: 02/28/2017 (autosaved)

File Edit View Insert Cell Kernel Help Python 2

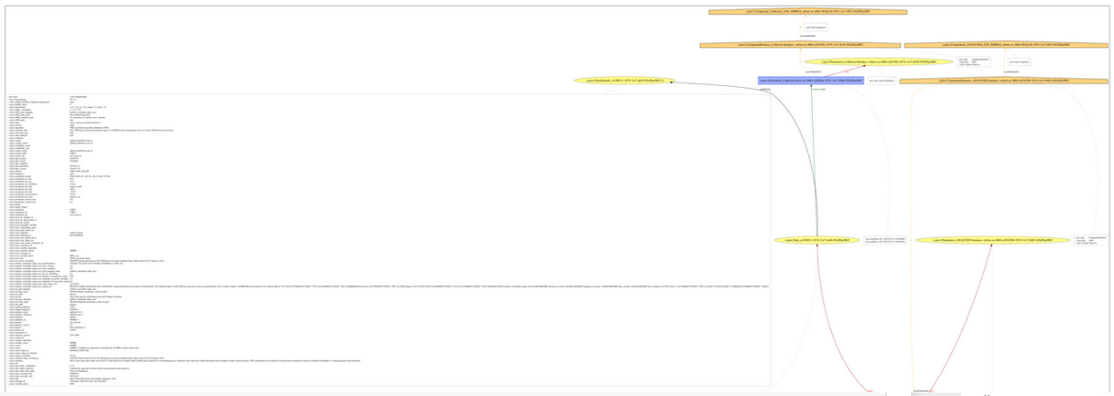
Code Cell Toolbar: None

2.5 - Explore the resulting provenance with static and interactive visualsations

The W3C-PROV provenance trace for a target data element is visualised.

In [14]:
```
import prov
import io
import StringIO
from prov.model import ProvDocument, ProvBundle, ProvException, first, Literal
from prov.dot import prov_to_dot

def provToSvg(xml,output_f):

    xml_doc = StringIO.StringIO()
    xml_doc.write(str(xml))
    xml_doc.seek(0, 0)
    #print xml_doc
    doc=ProvDocument.deserialize(xml_doc,format="xml")
    dot = prov_to_dot(doc)
    return dot.create(format=output_f)


#prov_doc=open(prov).read()

#print prov1

svg_content=provToSvg(prov1,"png")

with open("PROV.png","w+") as text_file:
    text_file.write(str(svg_content))

from IPython.display import Image
Image("PROV.png")


# visualse NetCDF provenance in PNG
```

Out[14]:



jupyter Prove

File Edit View Inse

1.2 Const

Instantiates t

In [3]:
```
#Initialis
def create
    readX.

    analys
    analys
    analys

    write
    write


    graph
    graph.
    graph.

    return
```

```
graph = createWorkflowGraph()


from dispel4py.visualisation import display
display(graph)
```

```
SETTING NAME: Read
SETTING NAME: Analysis
SETTING NAME: Write
```


dispel4py

jupyter

https://github.com/KNMI/wps_workflow/blob/master/ProvenanceDemo.ipynb

```
/" xmlns:prov="http://www.w3.org/ns/prov#" xmlns:s-prov="http://s-prov"

i.nl-10439-0758e376-d367-11e6-9b5f-901b0e877023">
e376-d367-11e6-9b5f-901b0e877023</s-prov:_id>
58e376-d367-11e6-9b5f-901b0e877023</s-prov:runId>
ov:startTime>


owId>
ne>

396.knmi.nl-10439-0758e376-d367-11e6-9b5f-901b0e877023">
knmi.nl-10439-083c05d4-d367-11e6-9b5f-901b0e877023">
.nl-10439-083c05d4-d367-11e6-9b5f-901b0e877023"/>
tor-pc150396.knmi.nl-10439-083b344c-d367-11e6-9b5f-901b0e877023"/>

knmi.nl-10439-083c1c04-d367-11e6-9b5f-901b0e877023">
.nl-10439-083c1c04-d367-11e6-9b5f-901b0e877023"/>
low-pc150396.knmi.nl-10439-083c1326-d367-11e6-9b5f-901b0e877023"/>

d="s-prov:wgb_pc150396.knmi.nl-10439-0840b278-d367-11e6-9b5f-901b0e877023">
prov:Data_pc150396.knmi.nl-10439-0840b278-d367-11e6-9b5f-901b0e877023"/>
s-prov:Invocation_Write-pc150396.knmi.nl-10439-083c294c-d367-11e6-9b5f-901b0e877023"/>

rov:Invocation_collector-pc150396.knmi.nl-10439-083b344c-d367-11e6-9b5f-901b0e877023">
5T16:50:15.456858</prov:startTime>
16:50:15.462283</prov:endTime>

rov:Invocation_Write-pc150396.knmi.nl-10439-083c294c-d367-11e6-9b5f-901b0e877023">
5T16:50:15.463131</prov:startTime>
16:50:15.493009</prov:endTime>

rov:Invocation_Workflow-pc150396.knmi.nl-10439-083c1326-d367-11e6-9b5f-901b0e877023">
5T16:50:15.462562</prov:startTime>
16:50:15.462899</prov:endTime>

v:Parameters_Write-Instance--pc150396.knmi.nl-10439-083b30e6-d367-11e6-9b5f-901b0e877023">
QName">s-prov:Parameters</prov:type>

v:Data_pc150396.knmi.nl-10439-0840b278-d367-11e6-9b5f-901b0e877023">
ksd:string"></prov:location>
QName">s-prov:Data</prov:type>
-prov:annotations>
ormat>
nl-10439-0840b278-d367-11e6-9b5f-901b0e877023</s-prov:id>
port>
ze>
```