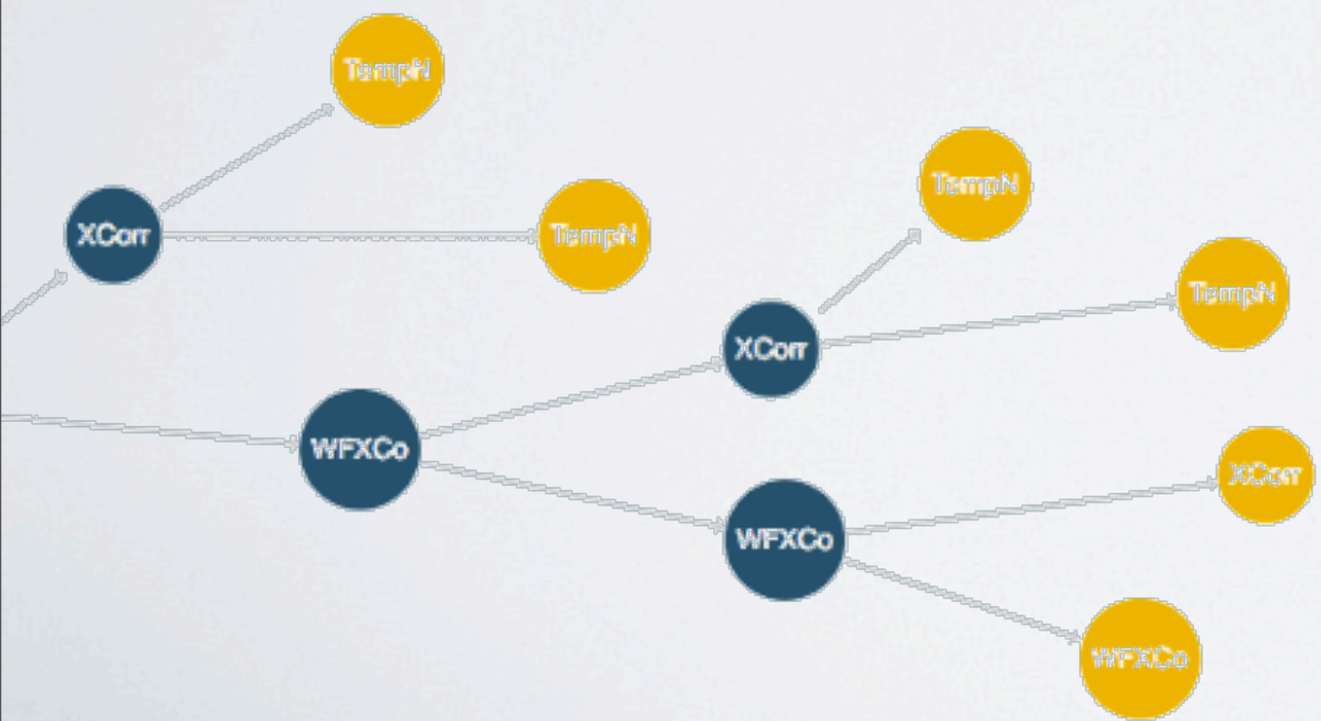
VERCE

HPC and Data Intensive Seismology

Alessandro Spinuso KNMI
and the **VERCE** team

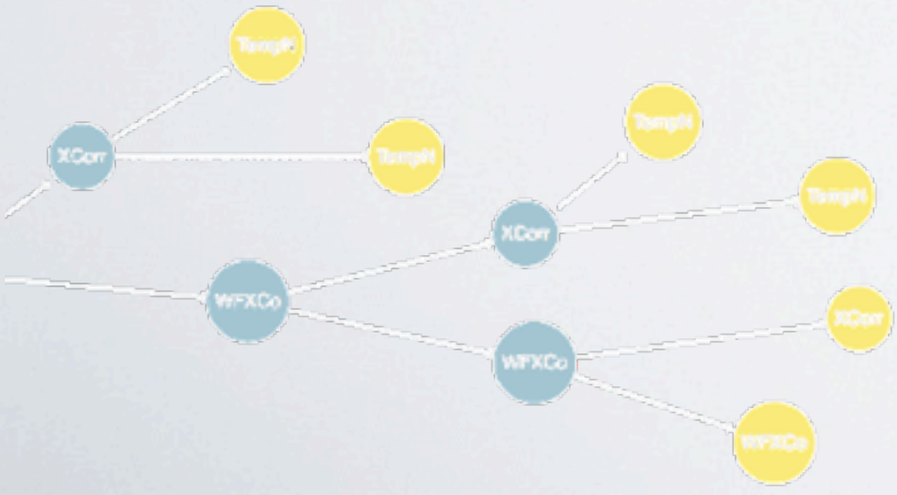


OUTLINE

Introducing VERCE

Use Cases: Data and Computational Intensive

Data Management and Provenance Challenges



Introducing VERCE

Virtual Earthquake and Seismology Research Community in Europe

International structure

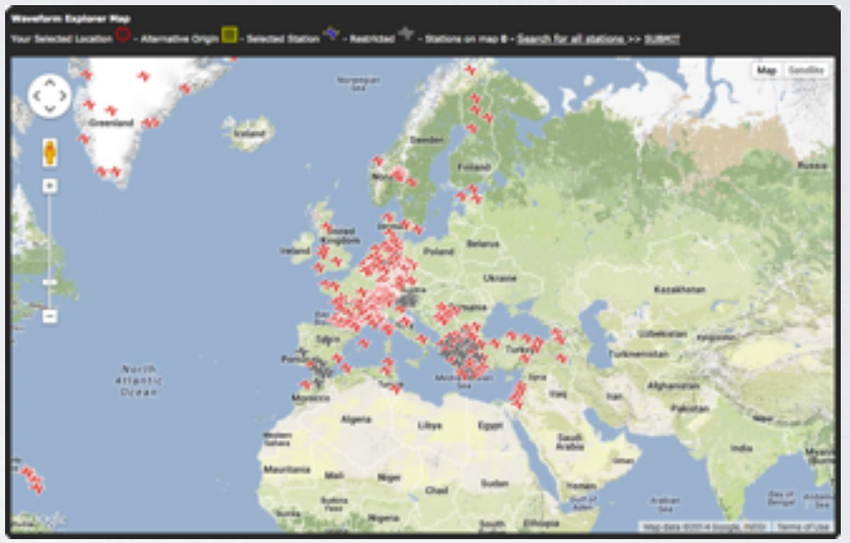
Global observations *and monitoring systems*
Integrated ***Distributed Data Archives***
Data and metadata formats

Scientific challenges

Understanding ***Earth's dynamics*** *and structures*
Imaging Earth's interior *and seismic sources*

Impact on Society

Natural hazard and risk mitigation;
Energy resources *exploration and exploitation;*
Underground wastes and ***carbon sequestration;***
Nuclear test monitoring *and treaty verifica*



Introducing VERCE

Virtual Earthquake and Seismology Research Community in Europe

International structure

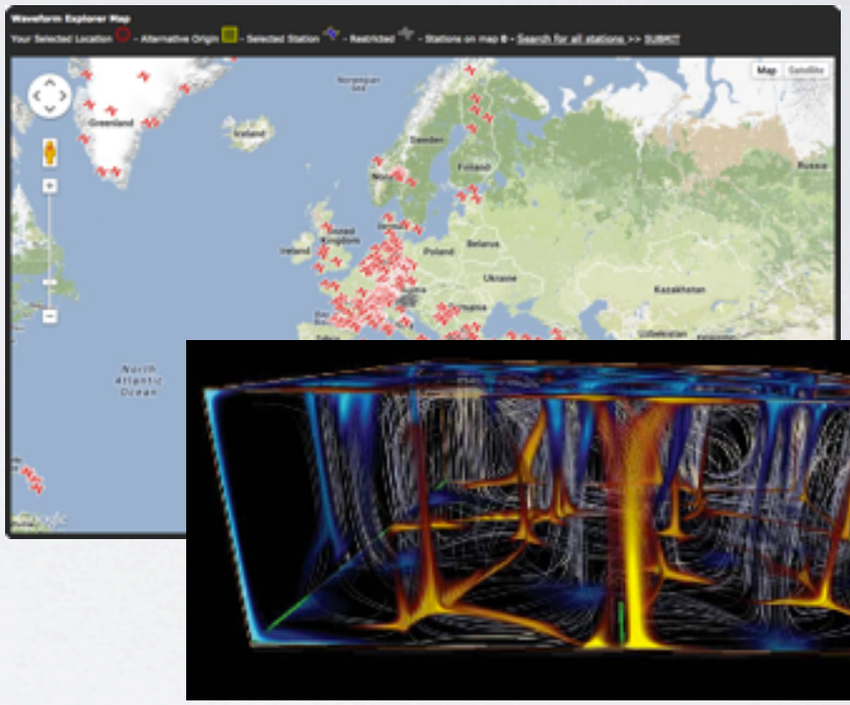
Global observations *and monitoring systems*
Integrated ***Distributed Data Archives***
Data and metadata formats

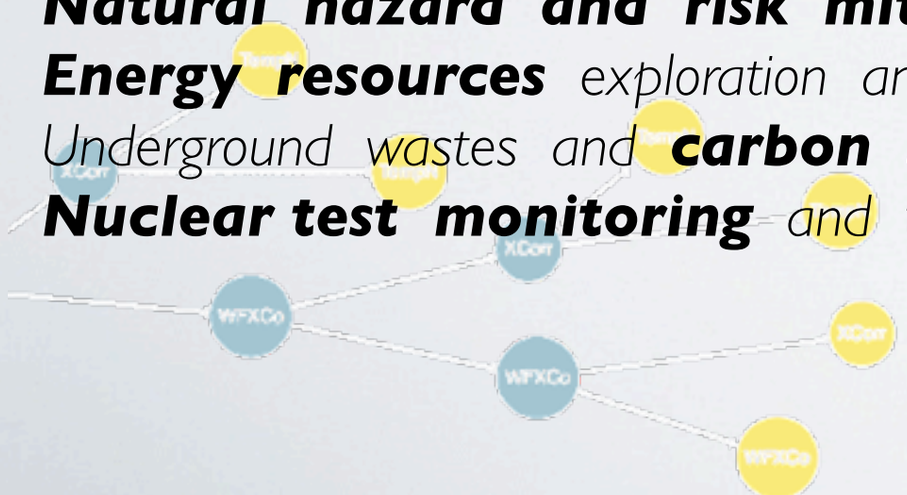
Scientific challenges

Understanding ***Earth's dynamics*** *and structures*
Imaging Earth's interior *and seismic sources*

Impact on Society

Natural hazard and risk mitigation*;*
Energy resources *exploration and exploitation;*
Underground wastes and ***carbon sequestration;***
Nuclear test monitoring *and treaty verifica*





Introducing VERCE

Virtual Earthquake and Seismology Research Community in Europe

International structure

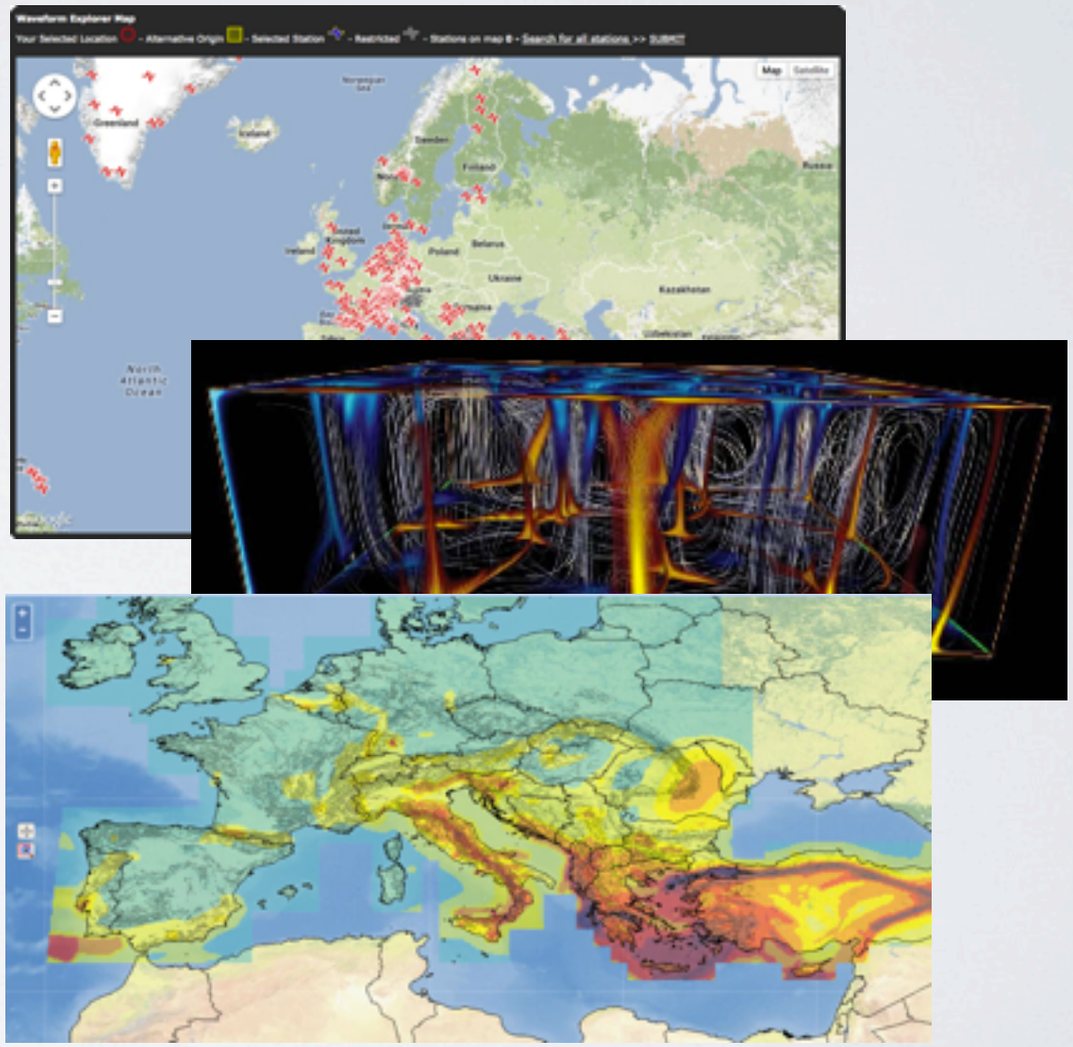
Global observations *and monitoring systems*
Integrated ***Distributed Data Archives***
Data and metadata formats

Scientific challenges

Understanding ***Earth's dynamics*** *and structures*
Imaging Earth's interior *and seismic sources*

Impact on Society

Natural hazard and risk mitigation*;*
Energy resources *exploration and exploitation;*
Underground wastes and ***carbon sequestration;***
Nuclear test monitoring *and treaty verifica*



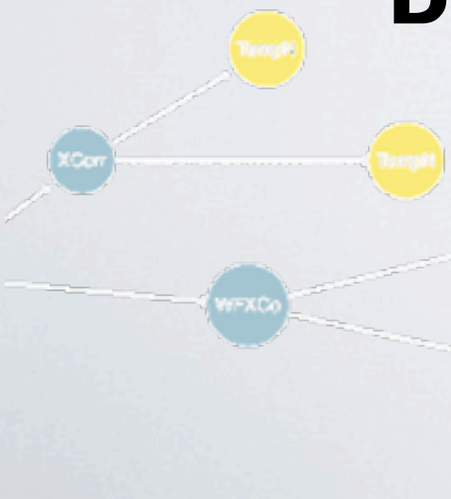
Introducing VERCE

VERCE provides to seismologists

Software as a service via the **VERCE Science Gateway**

Workflow tools and Registries

Data Management and Provenance System



Combine computing infrastructures
(EGI, PRACE, CLOUD), local resources HPC/DI

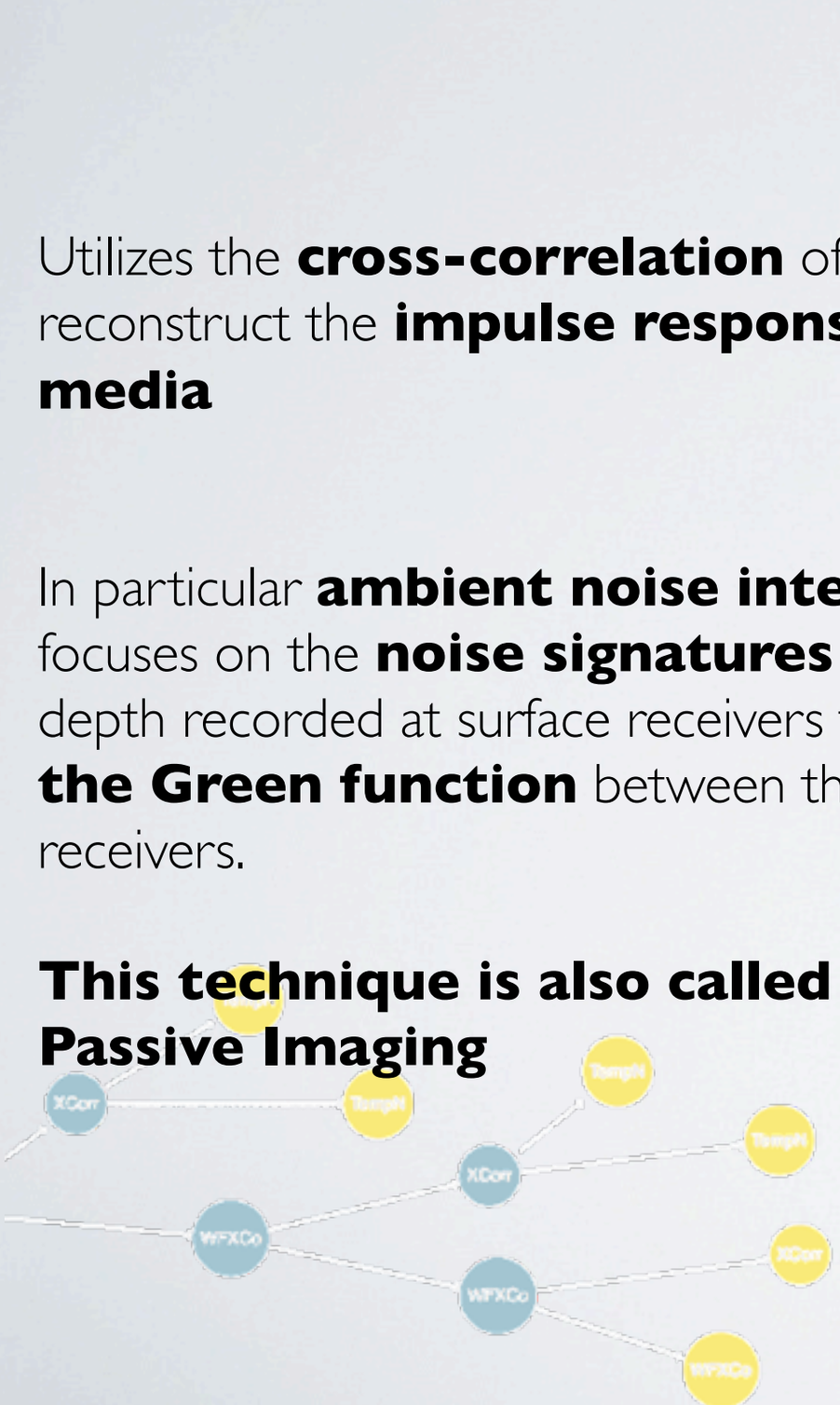
Access to **European** data archives and services

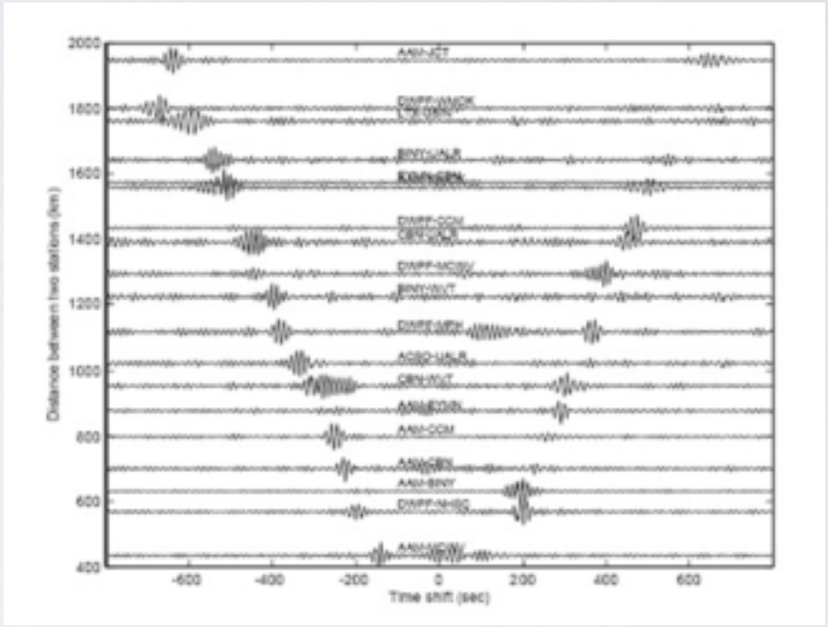
Data Intensive - Seismic Interferometry

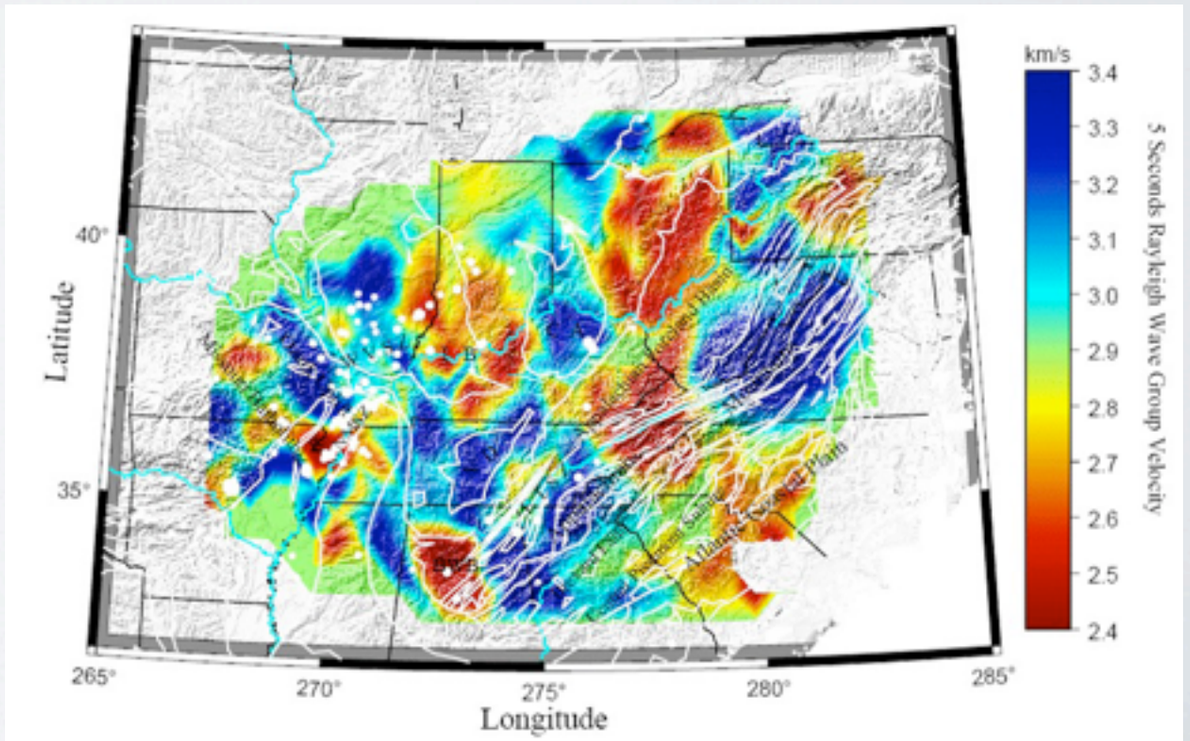
Utilizes the **cross-correlation** of signal pairs to reconstruct the **impulse response of a given media**

In particular **ambient noise interferometry** focuses on the **noise signatures** originating at depth recorded at surface receivers to **retrieve the Green function** between these receivers.

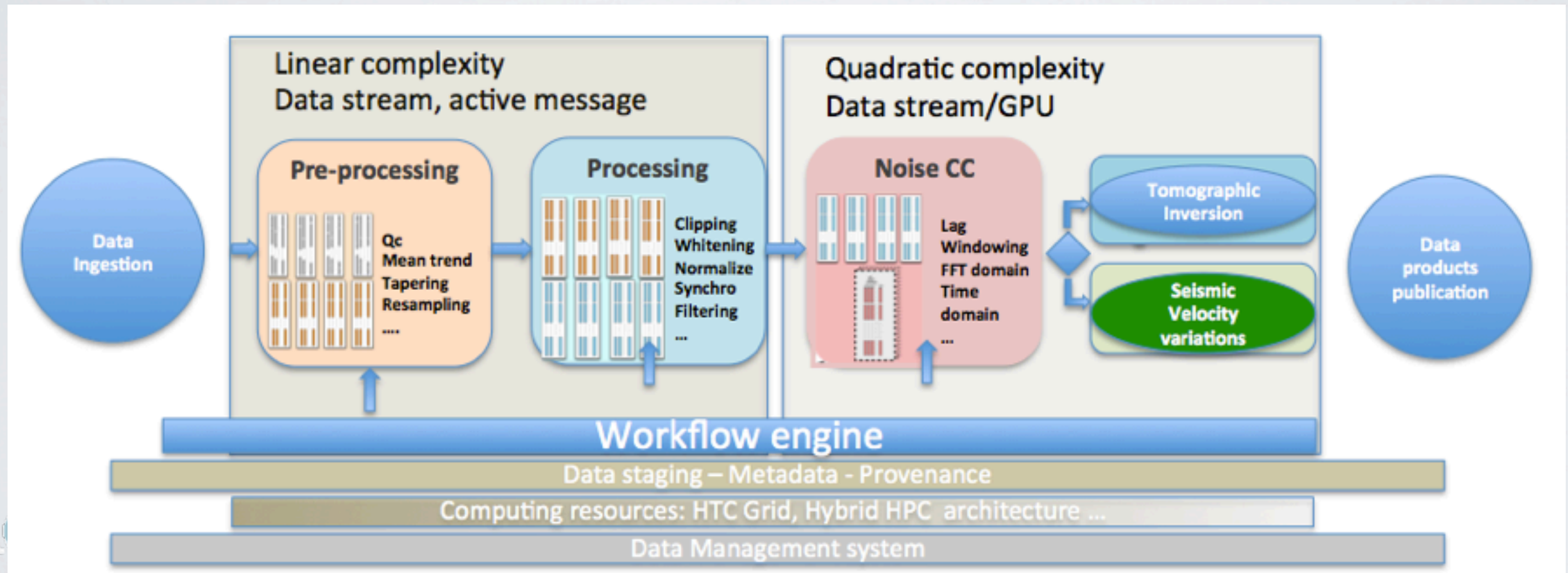
This technique is also called Passive Imaging



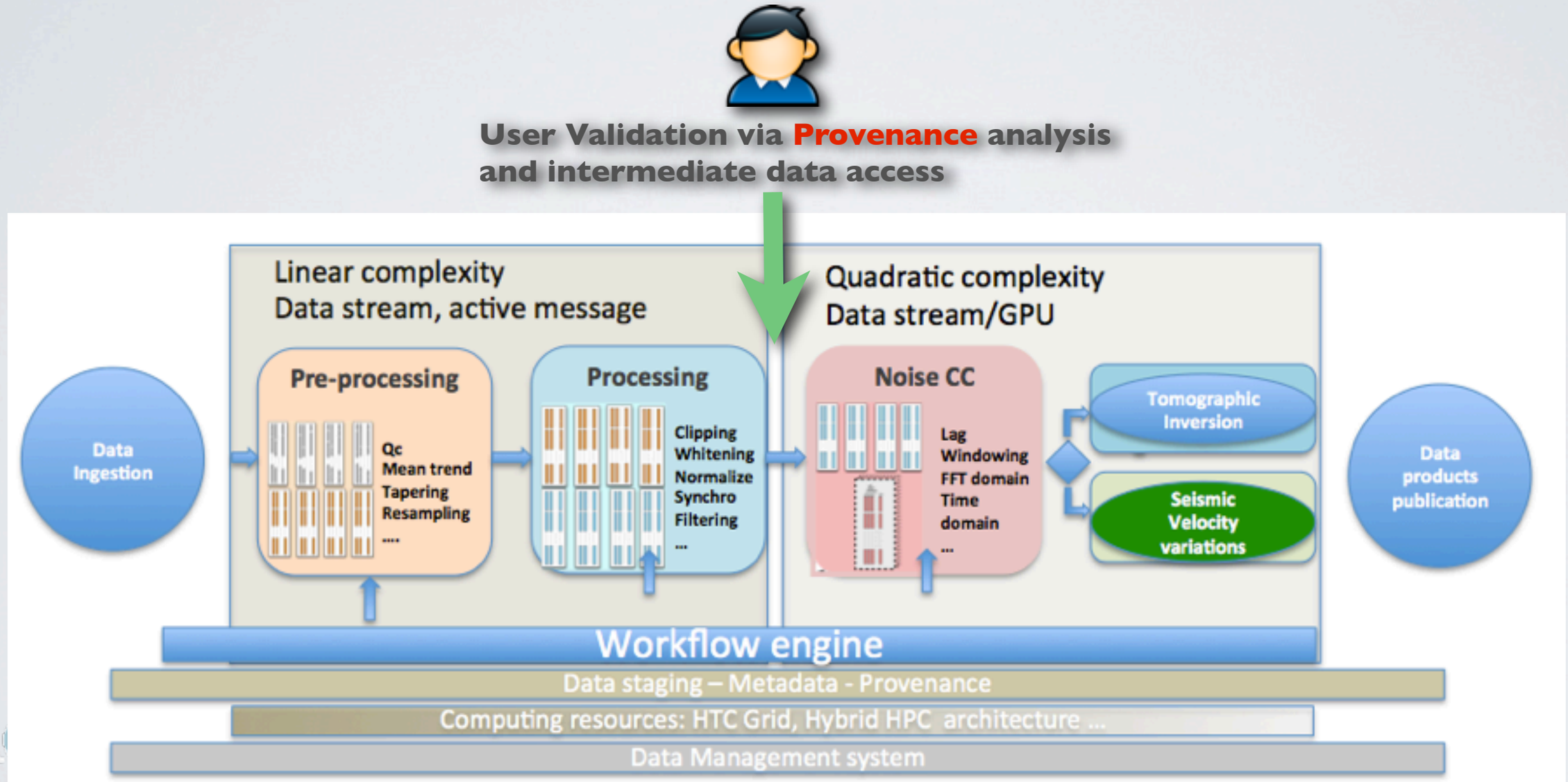




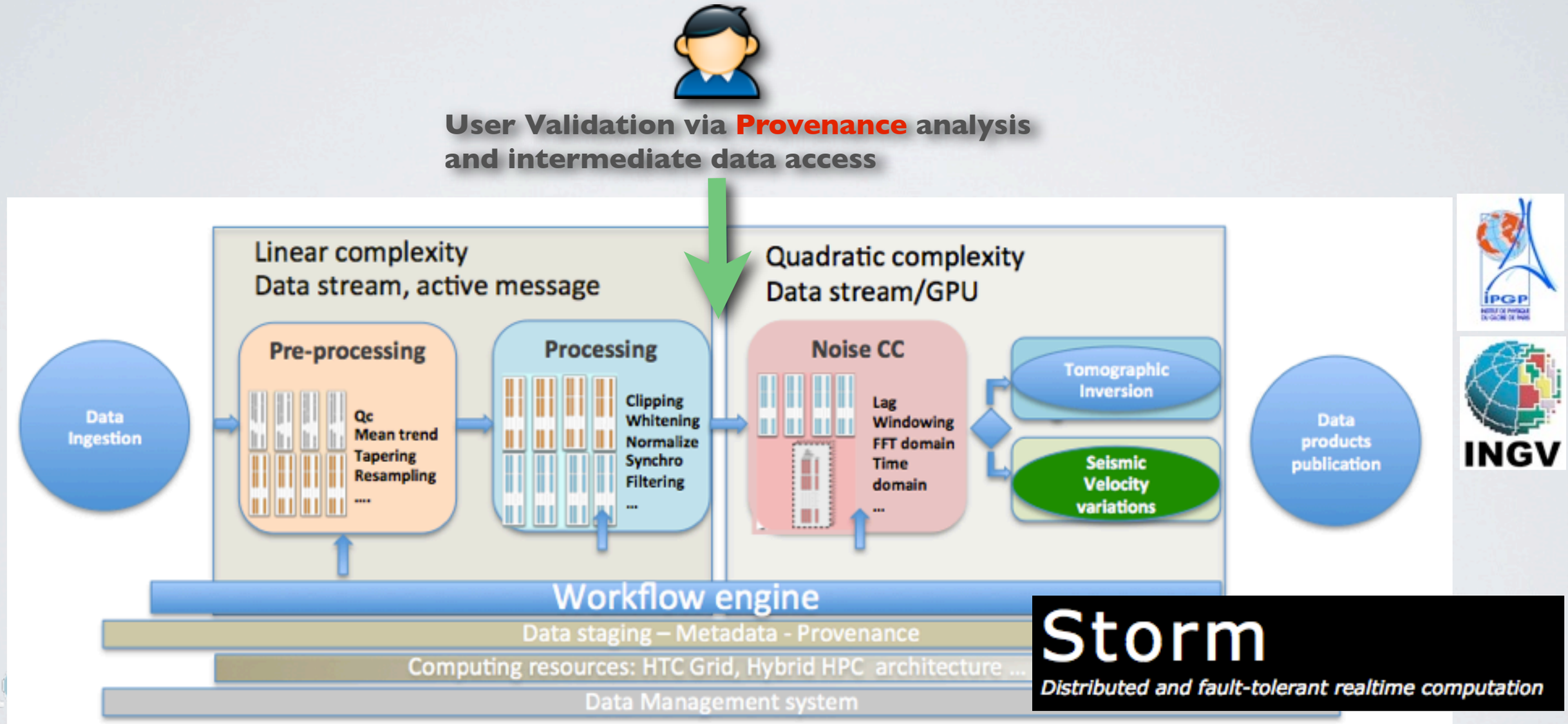
Data Intensive - Seismic Interferometry



Data Intensive - Seismic Interferometry



Data Intensive - Seismic Interferometry



Seismic Interferometry - Numbers

Data ingestion: 10 to 100 GB of input data per run.

Intermediate data products: Expected to be **as large as the input dataset**

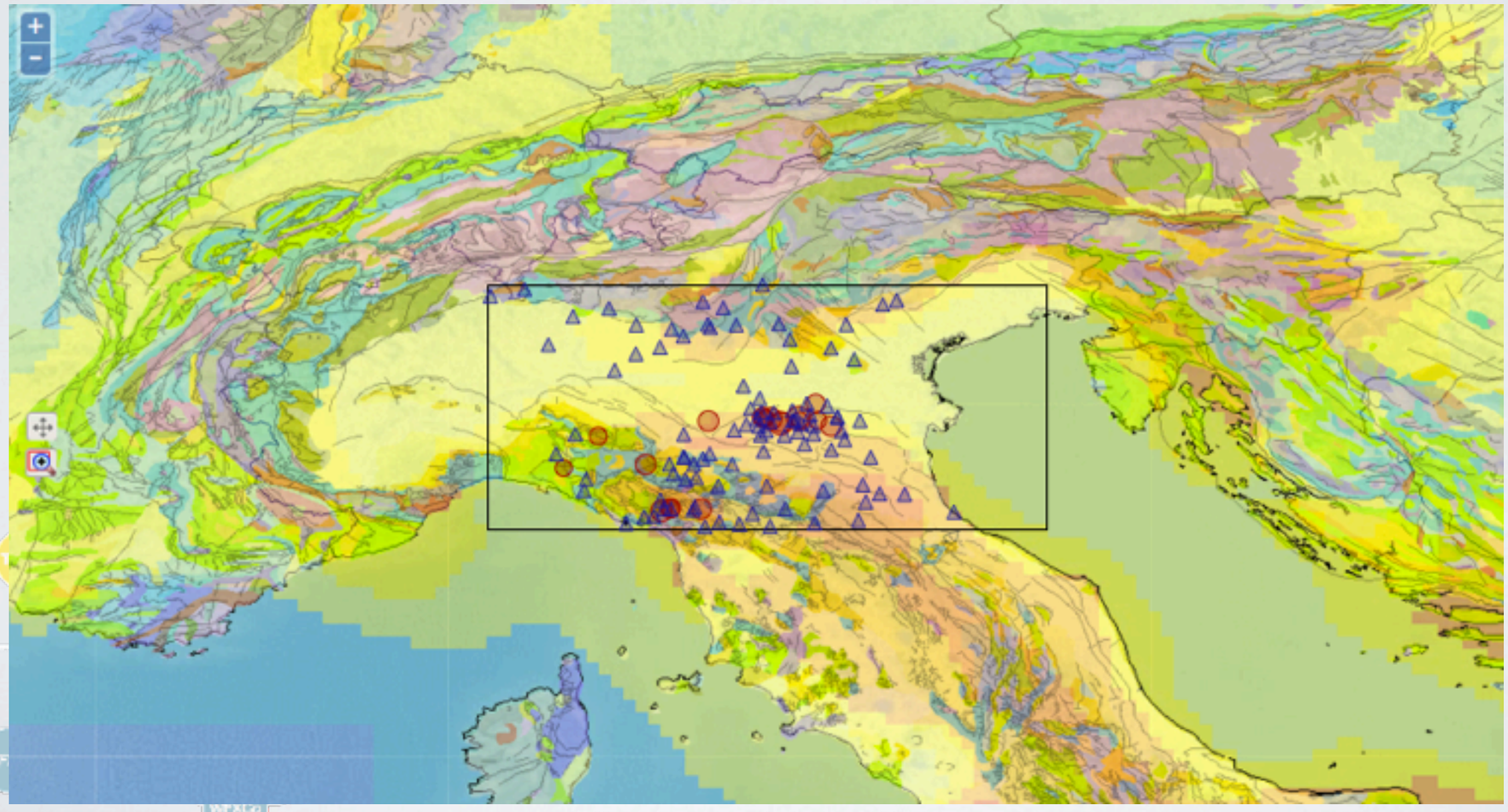
These data can be considered as part of a **preparatory phase** that will lead eventually to the execution of the complete workflow on the whole dataset.

Final results: Relatively small in size, we can consider them as **1/10 of the original Data**.

Frequency: full runs performed **hourly**, for time dependent variation analysis over **near-real time observations**

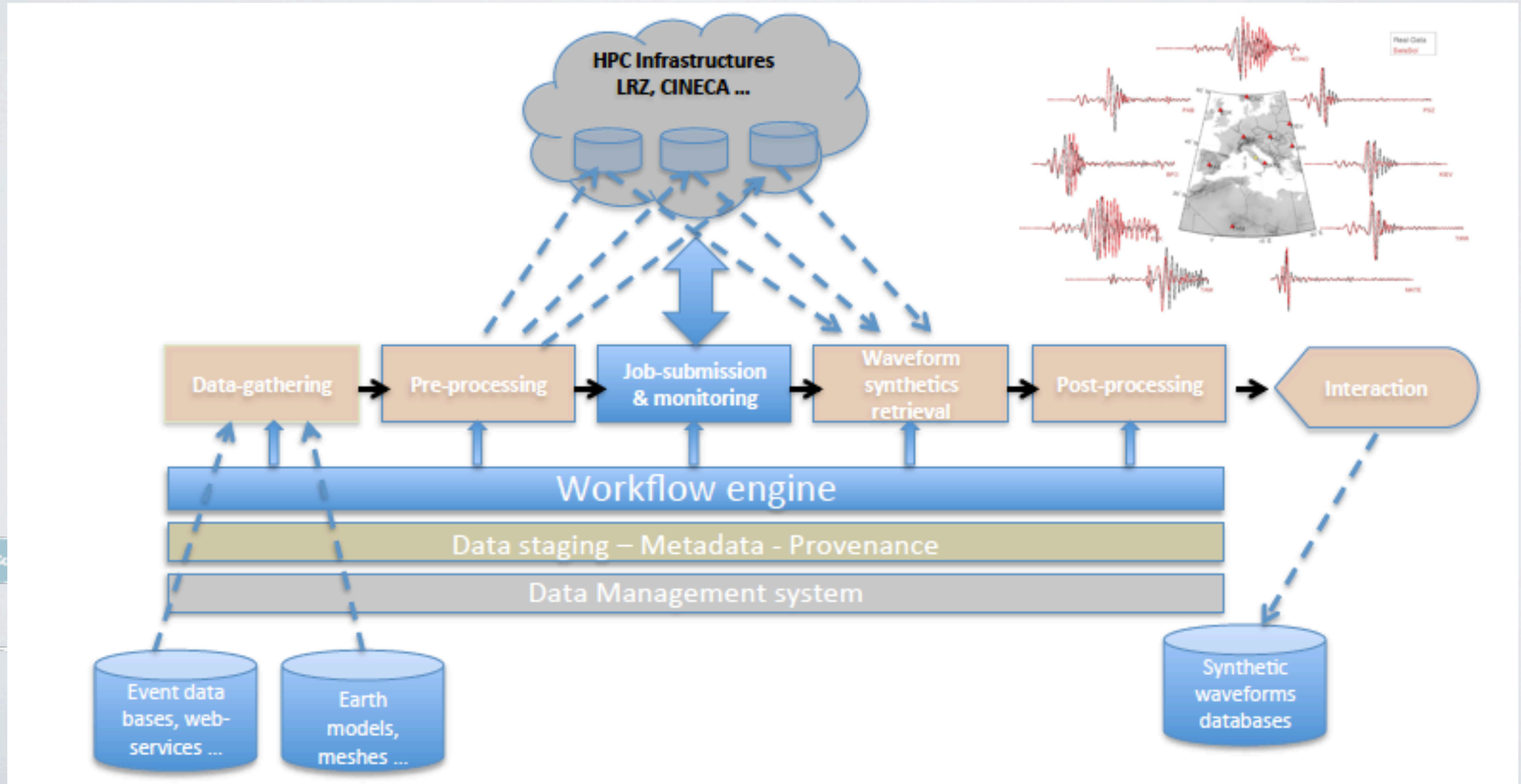
HPC - Forward Modeling

Simulation of seismic waves on HPC as a service
Develop a system which produces and **synthetic seismograms** for various **Earth models** (using forward solvers to be chosen) and eventually **compares them with observed seismograms** for **earthquakes** on a continental scale.



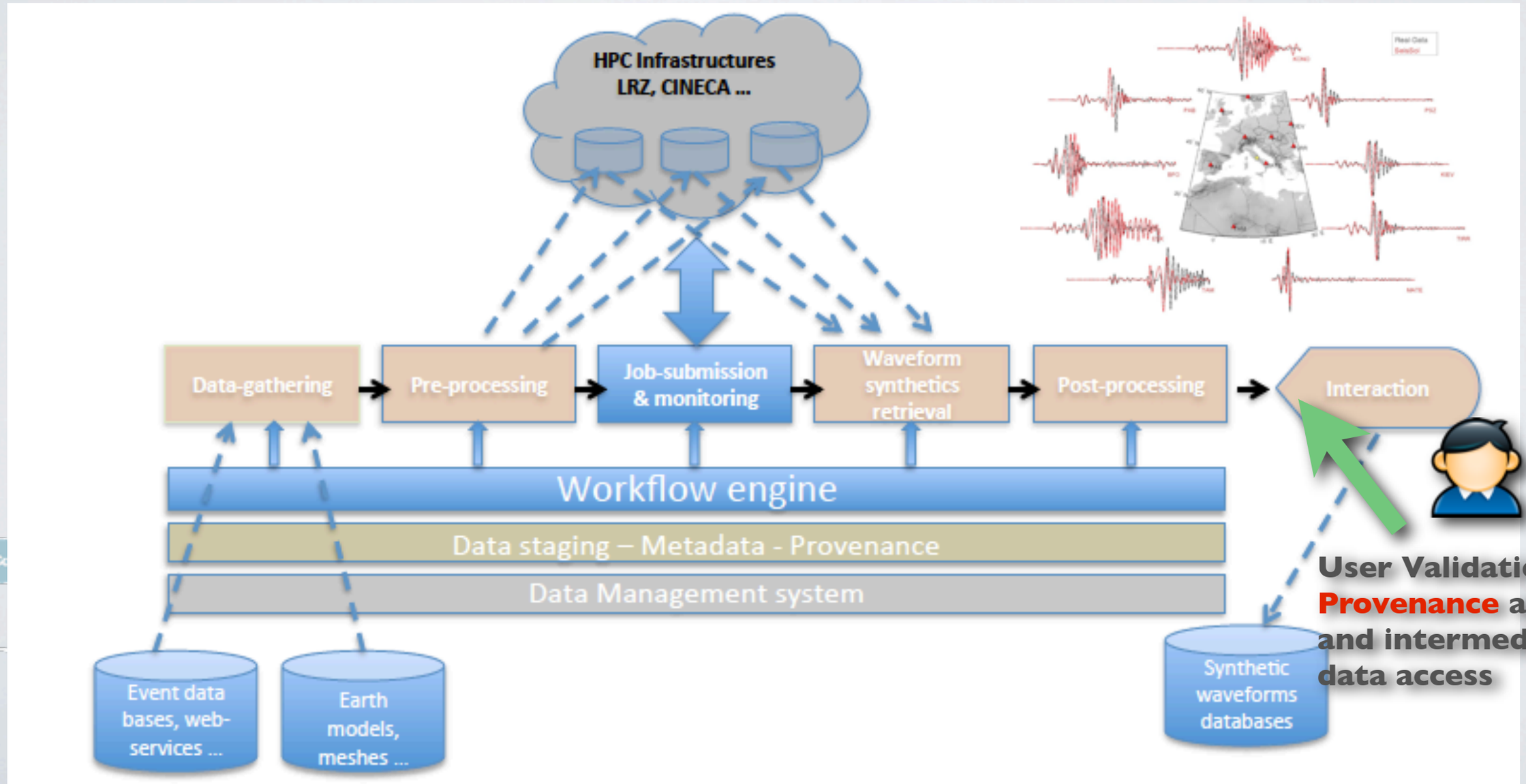
HPC - Forward Modeling

Simulation of seismic waves on HPC as a service
Develop a system which produces and **synthetic seismograms** for various **Earth models** (using forward solvers to be chosen) and eventually **compares them with observed seismograms** for **earthquakes** on a continental scale.



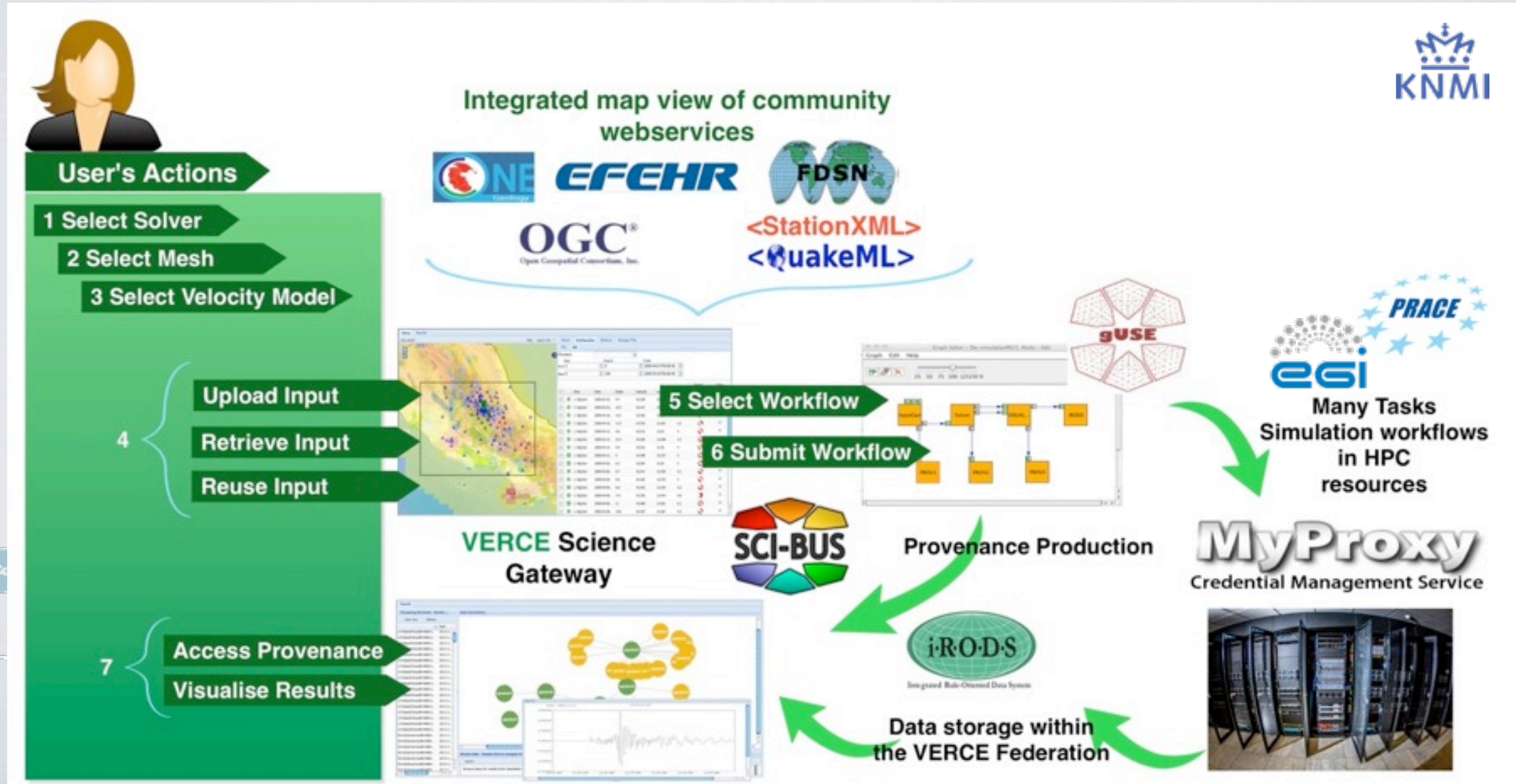
HPC - Forward Modeling

Simulation of seismic waves on HPC as a service
Develop a system which produces and **synthetic seismograms** for various **Earth models** (using forward solvers to be chosen) and eventually **compares them with observed seismograms** for **earthquakes** on a continental scale.



HPC - Forward Modeling

Simulation of seismic waves on HPC as a service
Develop a system which produces and **synthetic seismograms** for various **Earth models** (using forward solvers to be chosen) and eventually **compares them with observed seismograms** for **earthquakes** on a continental scale.



Forward Modeling Numbers

EGI - PRACE clusters (LRZ)

The data sources I: Configuration files and models which consist of roughly 300MB.

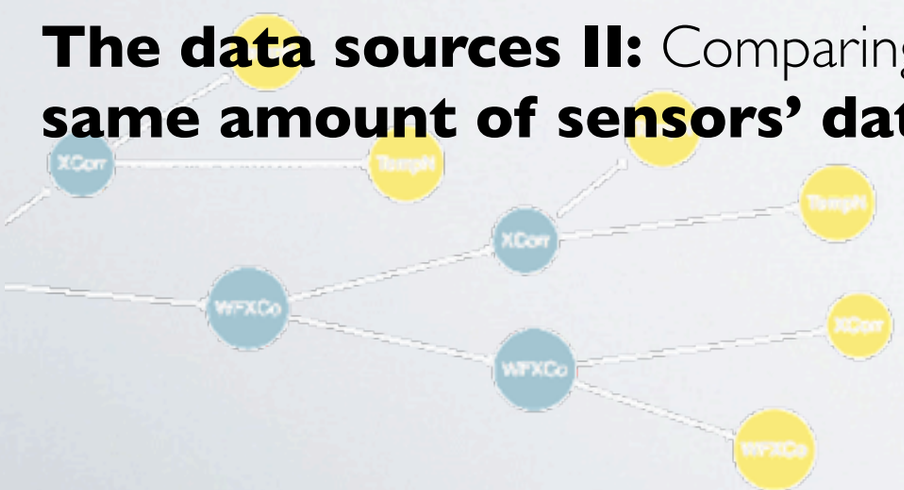
Processed data I: ~ 4GB of data to start with after **meshes processing.**

Processed data II: Thousands of **synthetic seismograms**.

Frequency: **100 stations** will produce **900 products** and **metadata** in **30 minutes**

Data types: Binary (application specific), **ascii** (application specific), **graphic visualization of output** (png/pdf)

The data sources II: Comparing **synthetic** with **real observations**. Access at least the **same amount of sensors' data.**



Challenges

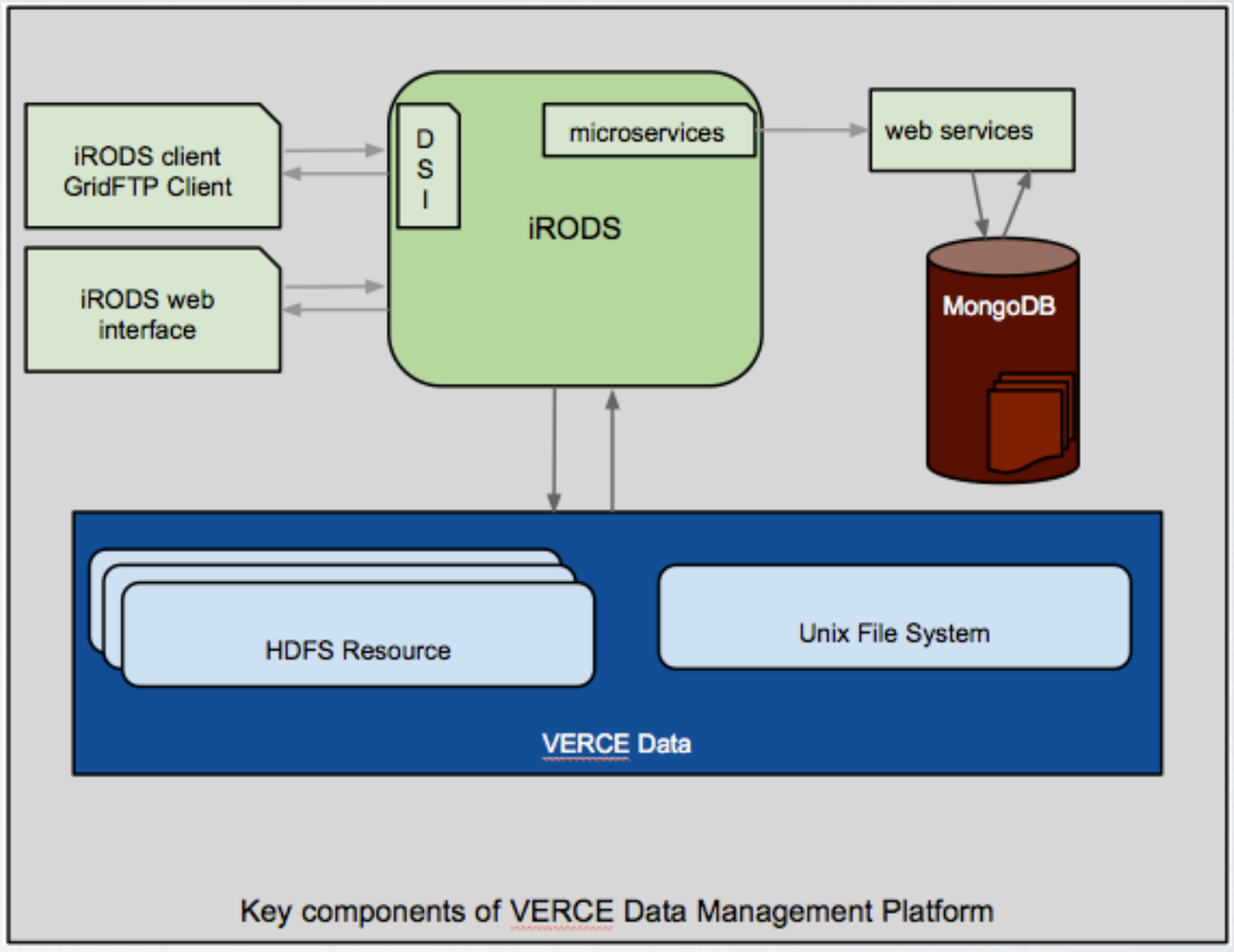
Data Management external to the computing resources

iRODS Federation allows **authenticated user** from a trusted remote site to **access shared data**

GSI Authentication **GridFTP** (DSI - CINECA)

Microservices

Exploring distributed data processing on this platform (**HDFS**)



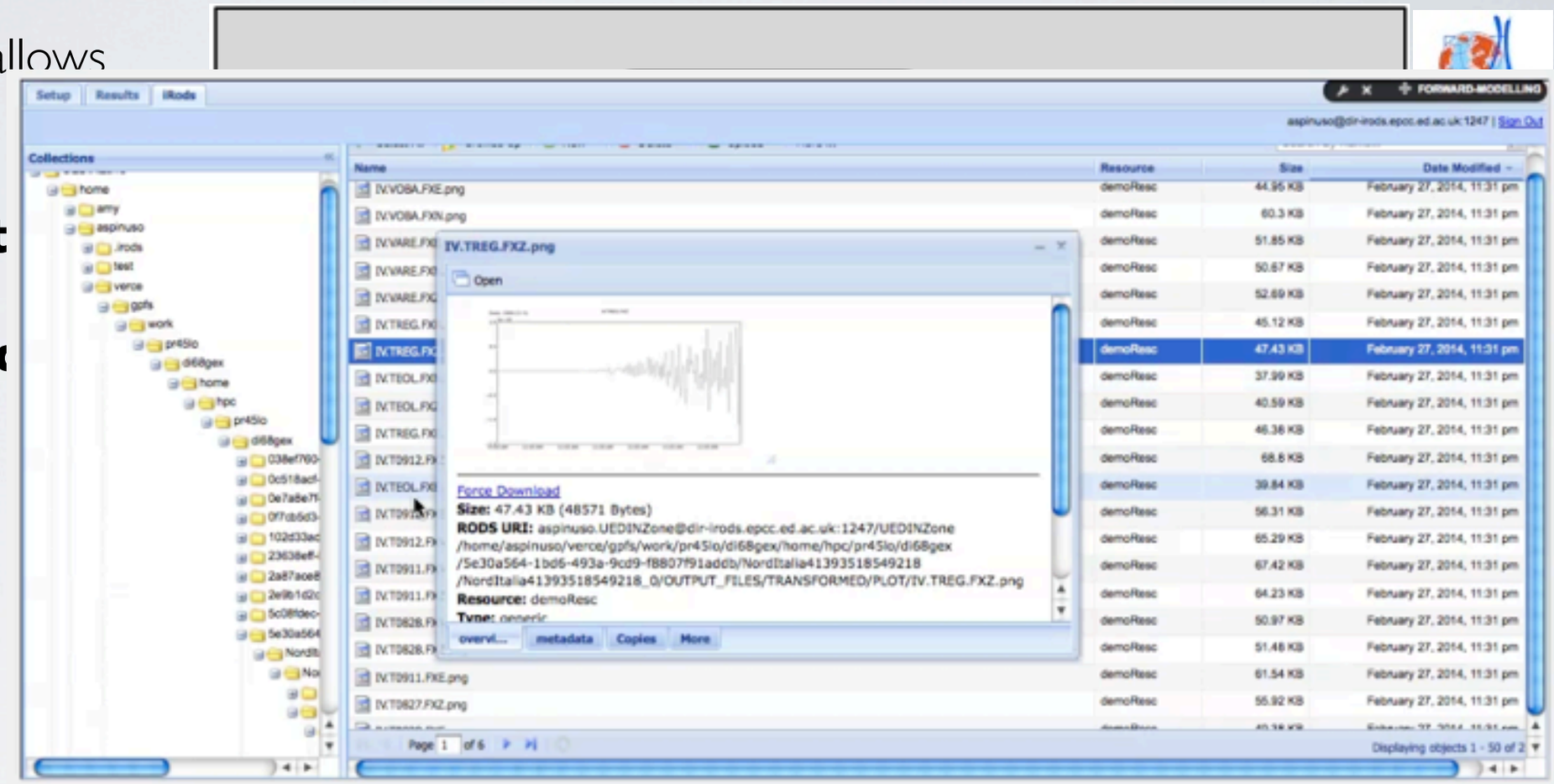
Key components of VERCE Data Management Platform

Challenges

Data Management external to the computing resources

iRODS Federation allows
authenticated user
a trusted remote site
to **access shared dat**

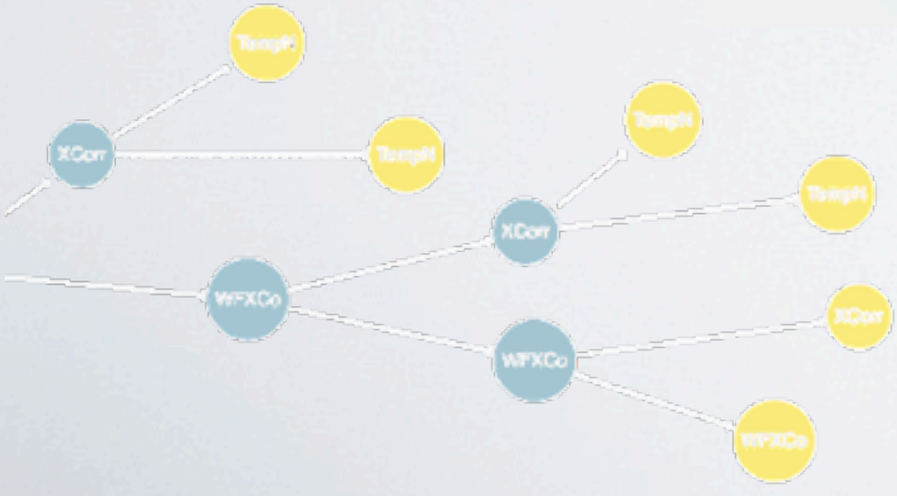
GSI Authentication **Gri**
(DSI - CINECA)



Key components of VERCE Data Management Platform

Microservices

Exploring distributed
data processing on
this platform (**HDFS**)



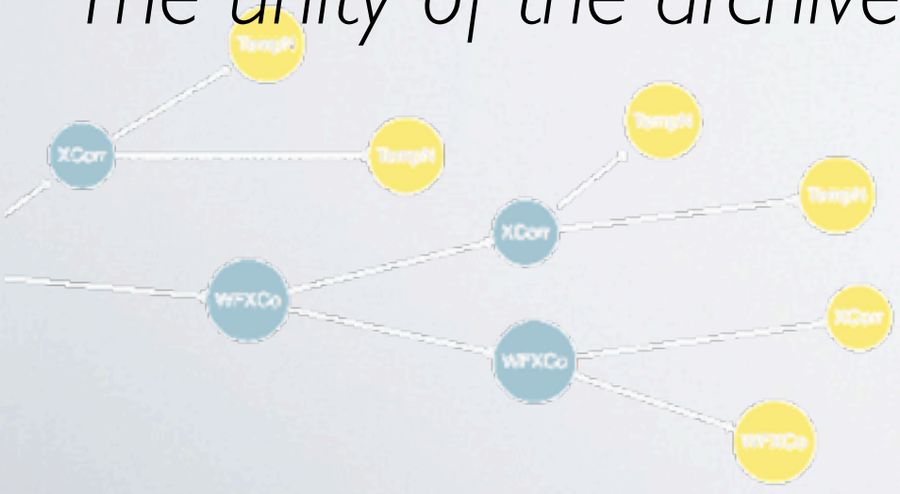
Provenance information, a point of view

France 1790, Establishment of Les Archives Nationales,
trying to merge private and public archives throughout the country

... 50 years of archivists headache trying to regroup and classify records ...

France 1841, formulation of the principle of provenance:

The unity of the archive took precedence over the material objects (respect des fonds).

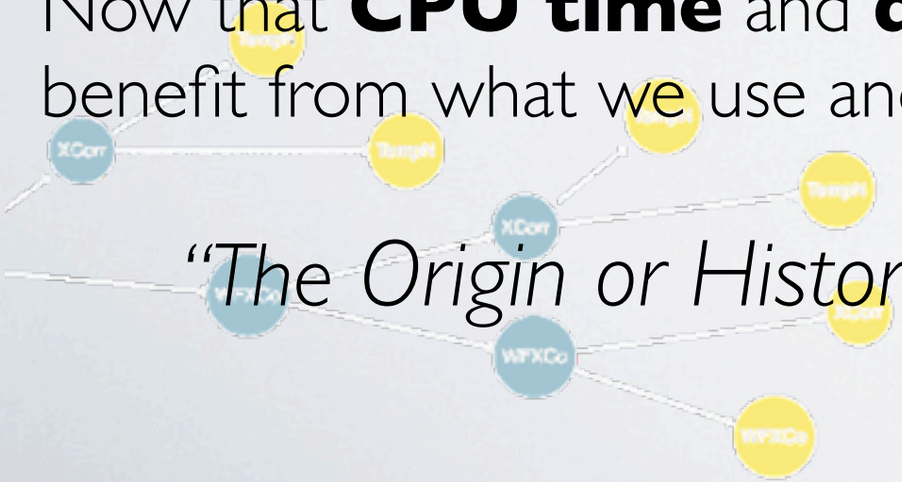


http://ingmarbergman.se/en/category/tags/principle-provenance

Provenance information, a point of view

Which is the impact on the curation of Scientific Data?

Respect des fonds: The unity of the archive takes precedence over the material objects.
Who are the archivists?
Which percentage of the actual **provenance recordings is relevant ?**
Unified models are useful, how about the **content ?**
Now that **CPU time** and **data bandwidth** is a **business model**, shall we benefit from what we use and pay ?



"The Origin or History" - That really matters - "of something"

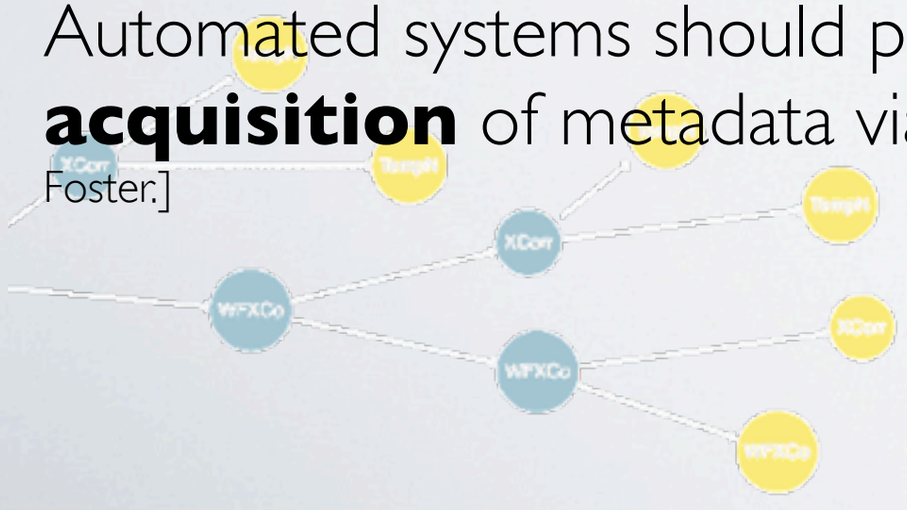
Provenance information, a point of view

What to archive and Who are the archivists?
Researchers vs Software

The Exploration of the adoption of **data stream processing engines** for **Data Intensive** computation, presents expensive requirements for provenance collection, either in terms of size or I/O [W. D. Pauw]

Researchers are part of the archiving process.

They know what is relevant to understand their results.
Automated systems should provide **support** for a consistent and effective **acquisition** of metadata via **selective** and **extensible** provenance. [A. Misra] [I. Foster.]



Challenges

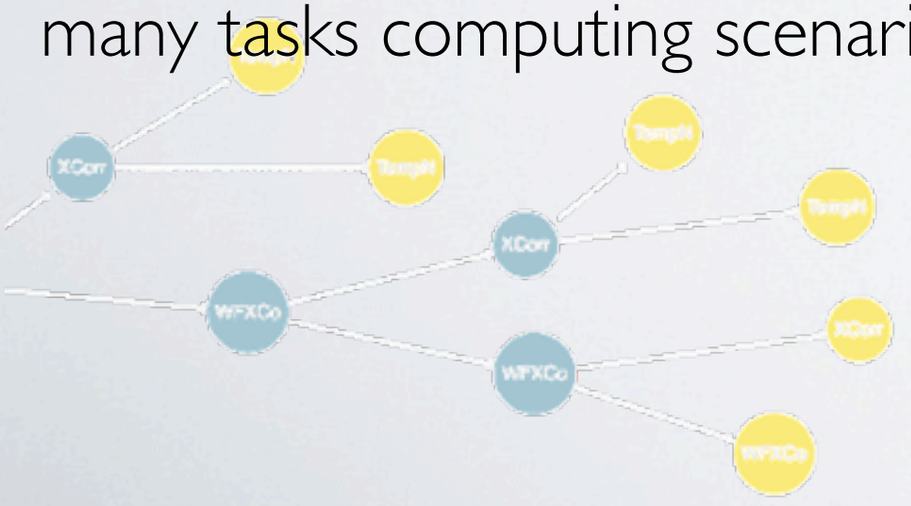
Runtime Provenance for BigData

History is now!
Provenance data should be accessible at runtime

It allows the **immediate interaction** of users who can **monitor** the production of the results, avoiding useless waits.

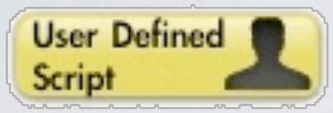
A preliminary evaluation of partial results might **suggest actions** to be taken before the termination of the computation

Runtime Provenance can be used to **recover failures**, for instance, in a many tasks computing scenarios



Challenges

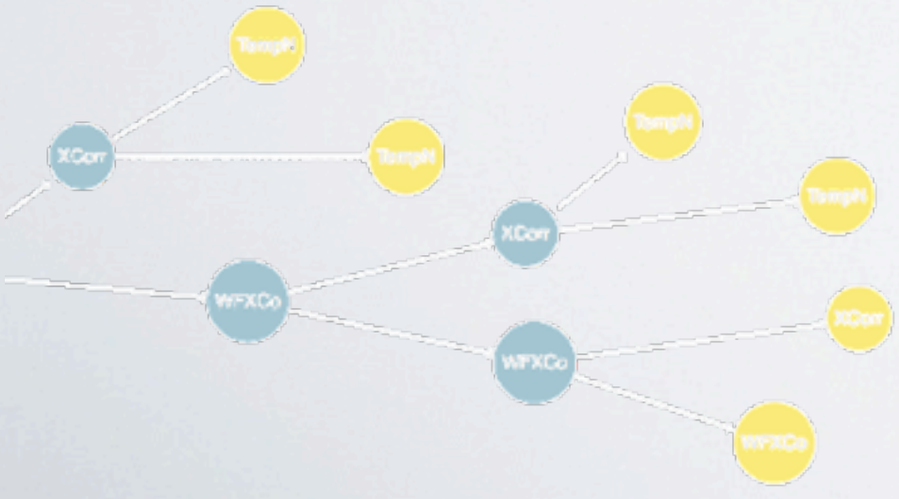
Comprehensive pre-post processing framework across DI-HPC models



User Code Integration framework:

Easy for scientists to implement, **Python Objects / Functions**
Users define which metadata has to be extracted from the data

Automated extraction of lineage rich of domain specific metadata



Challenges

Comprehensive pre-post processing framework across DI-HPC models

User D
Script

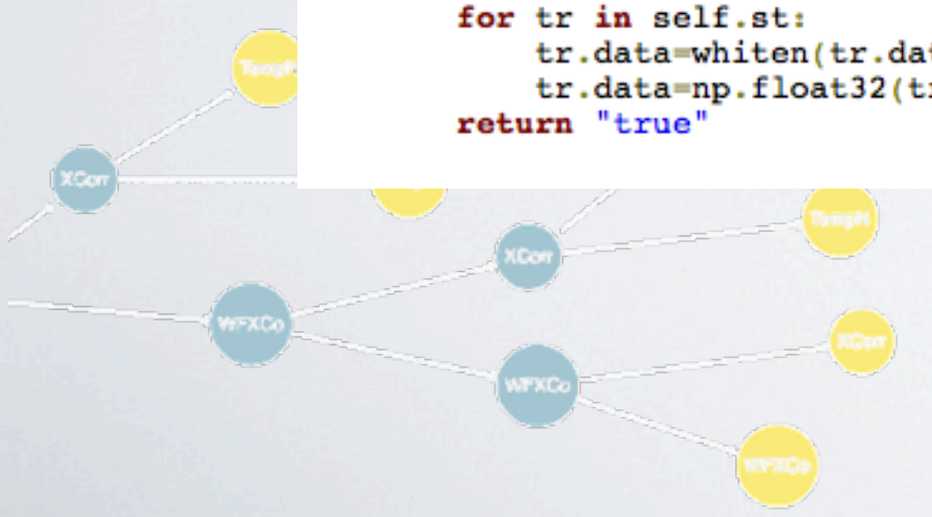
Use

Easy

Use

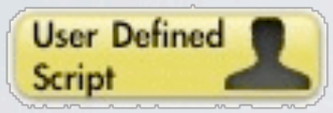
Aut
met

```
class Whiten(SeismoPreprocessingActivity):

    def compute(self):
        _ntaper=int("%s" % (self.parameters["ntaper"],));

        def whiten(data, delta, ntaper):
            """
            applies spectral whitening on signal by normalizing all freq. amplitudes
            sig: waveform array
            delta : sampling interval
            ntaper : tapering width
            """
            filtSig = np.zeros(np.shape(data))
            FFTsig = fft(data)
            if sum(abs(FFTsig)) == 0:
                filtSig = np.zeros(np.shape(filtSig))
                self.error="warning: cannot whiten, the FFT is zero !!"
            else:
                # normalize frequencies
                filtSig = FFTsig / abs(FFTsig)
                data = np.real(ifft(filtSig))

            return data

    for tr in self.st:
        tr.data=whiten(tr.data, tr.stats.delta, _ntaper)
        tr.data=np.float32(tr.data)
    return "true"
```



Challenges

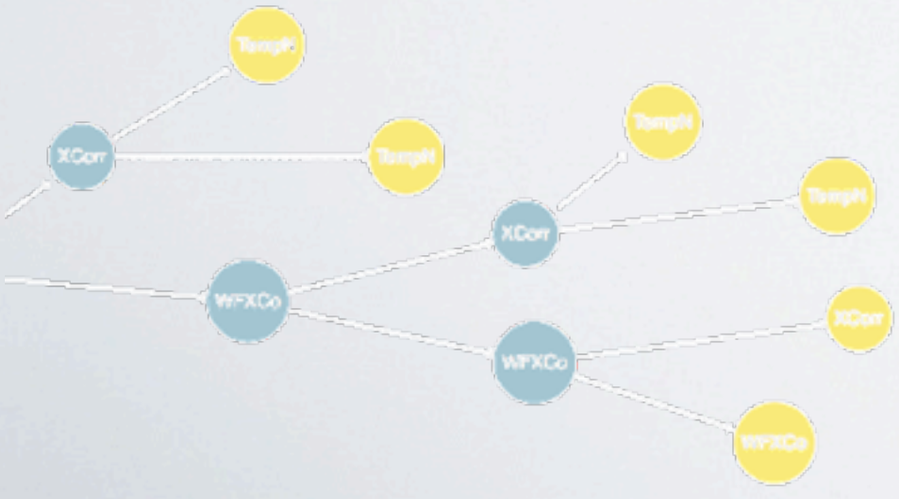
Comprehensive pre-post processing framework across DI-HPC models



User Code Integration framework:

Easy for scientists to implement, **Python Objects / Functions**
Users define which metadata has to be extracted from the data

Automated extraction of lineage rich of domain specific metadata



Challenges

Comprehensive pre-post processing framework across DI-HPC models

User D
Script

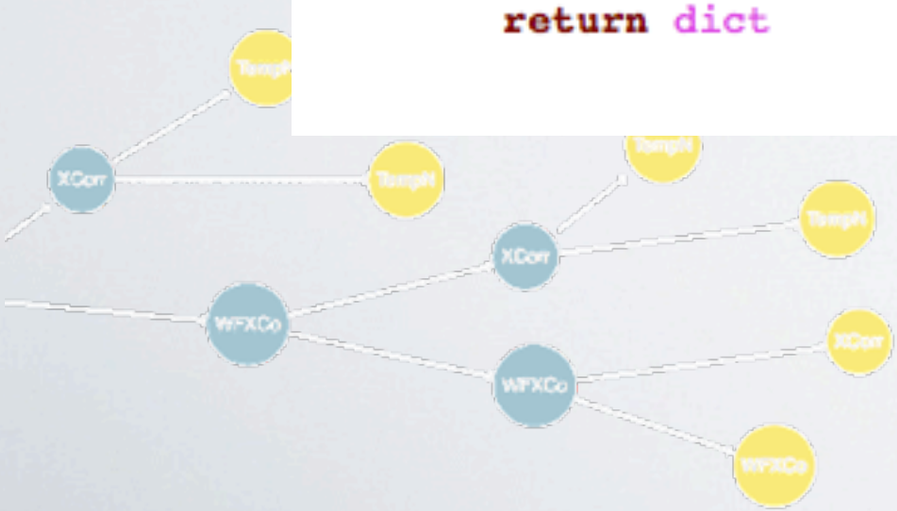
Use

Easy

Use

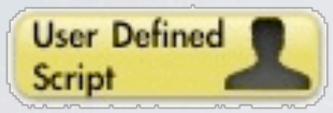
Aut
met

```
class inputGenerator(SeismoPreprocessingActivity):

    def extractItemMetadata(self,st):
        dict=""
        try:
            dict={"path":str(st[0])}

            for attr, value in st[1][0].iteritems():
                try:

                    if type(value)==obspy.core.utcdatetime.UTCDateTime:
                        dict.update({attr:str(value)});
                    else:
                        dict.update({attr:float(value)});
                except Exception,e:
                    dict.update({attr:str(value)});
        except Exception,e:
            dict={"to_xdecompose":str(st)}

        return dict
```



Challenges

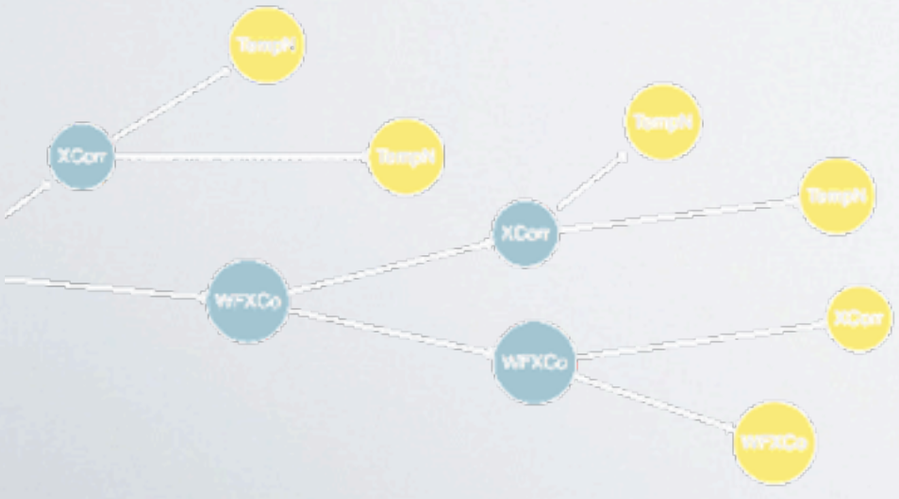
Comprehensive pre-post processing framework across DI-HPC models



User Code Integration framework:

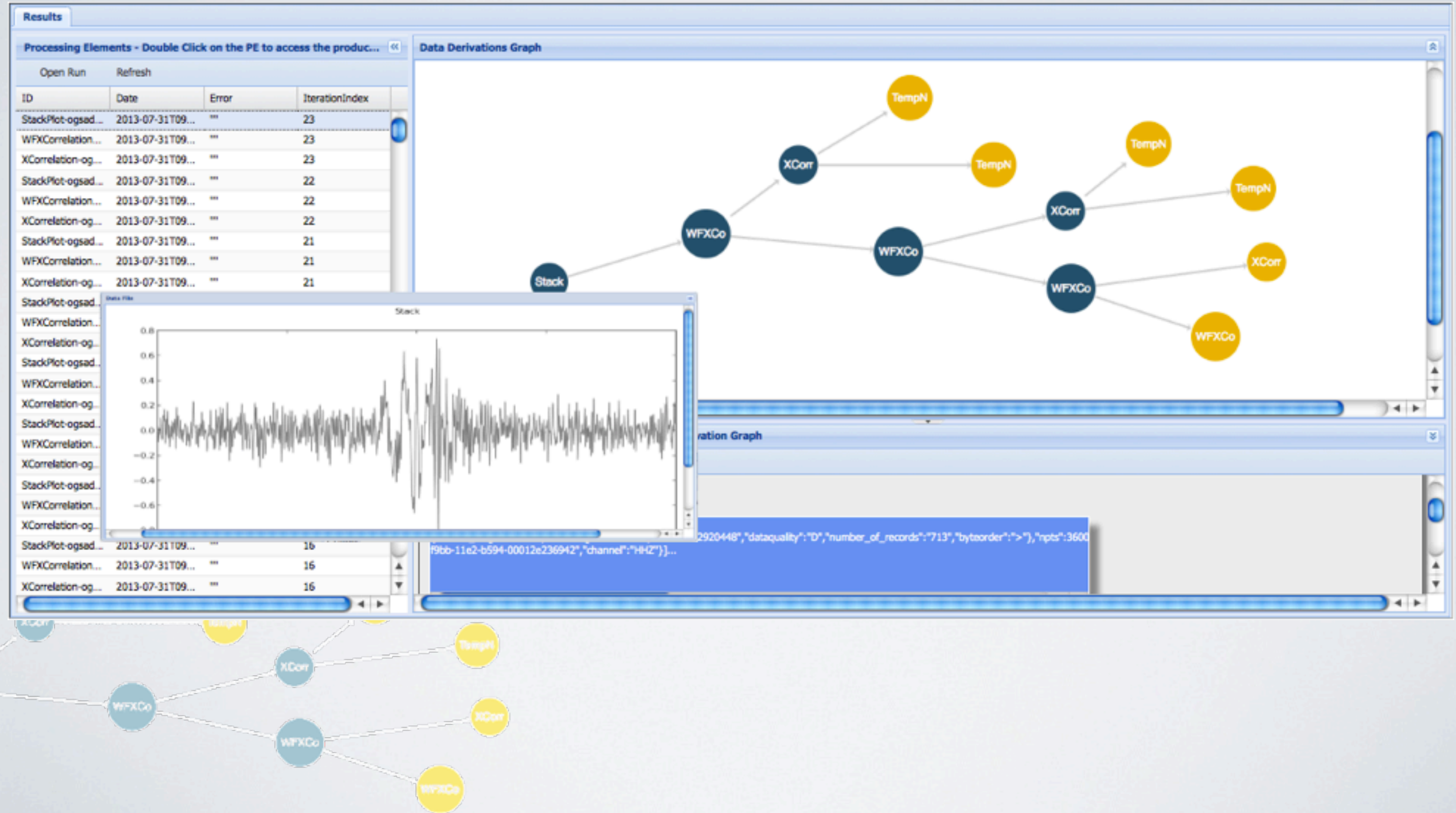
Easy for scientists to implement, **Python Objects / Functions**
Users define which metadata has to be extracted from the data

Automated extraction of lineage rich of domain specific metadata



Challenges

Comprehensive pre-post processing framework across DI-HPC models



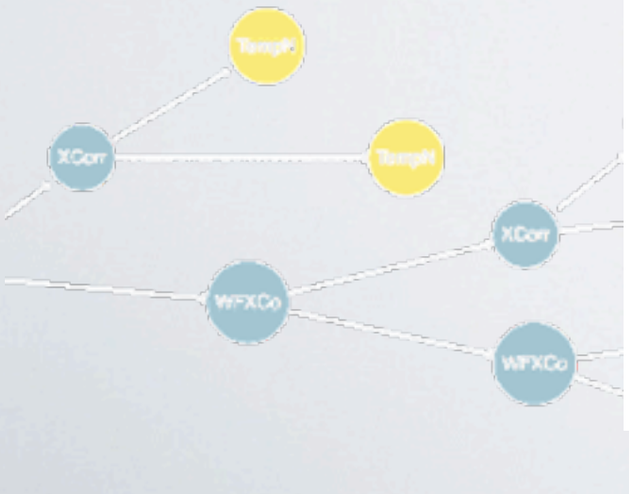
Challenges

Comprehensive pre-post processing framework across DI-HPC models

Lineage
Model Coverage
W3C-PROV

Table 1: PROV-O Coverage of the proposed model

PROV-O Terms	Covered	Desc
`prov:Activity`	Yes	A process that occurs over a period of time.
`prov:Agent`	Yes	Something that bears some form of responsibility for an activity taking place.
`prov:Entity`	Yes	A physical, digital, conceptual, or other kind of thing with some fixed aspects.
`prov:endedAtTime`	Yes	The time at which an activity ended.
`prov:startedAtTime`	Yes	The time at which an activity started.
`prov:used`	Yes	A Entity that was used by this Activity.
`prov:wasAssociatedWith`	Yes	An Agent that had some (unspecified) responsibility for the occurrence of this Activity.
`prov:wasAttributedTo`	Yes	Attribution is the ascribing of an entity to an agent.
`prov:wasDerivedFrom`	Yes	A derivation is a transformation of an entity into another.
`prov:wasGeneratedBy`	Yes	Generation is the completion of production of a new entity by an activity.
`prov:wasInformedBy`	Yes	Communication is the exchange of an entity by two activities.



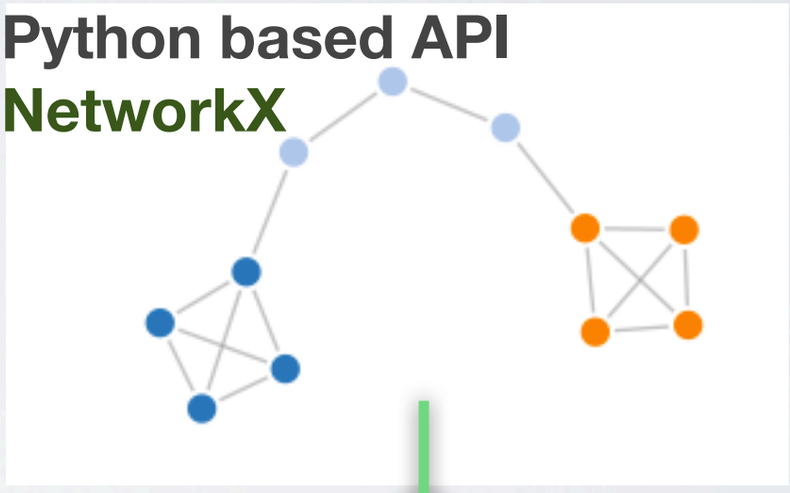
Challenges

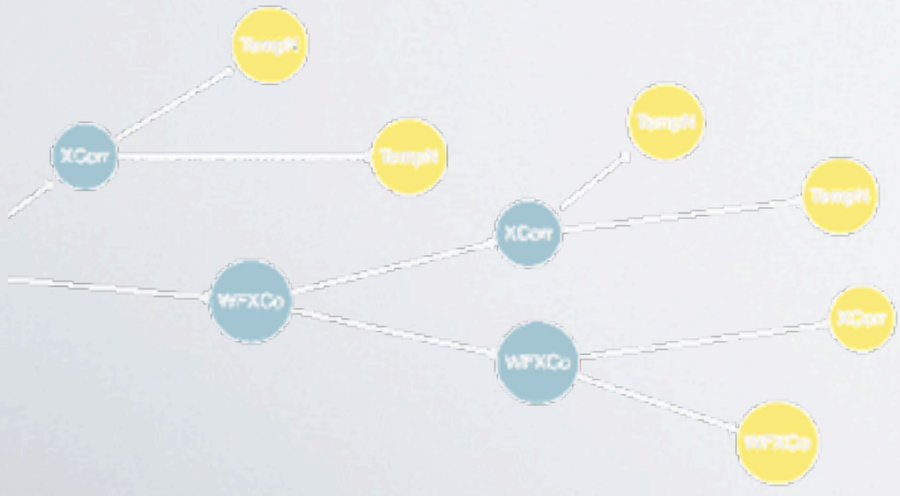
In workflows Selective Provenance

Definition of a common processing framework enabling **selective provenance tracking**, across computational models and infrastructures

Trade-offs between ease of use,
efficiency and **power of the interrogation**.

Needs the exploration of **specific use cases**
and the observation of the **measurable benefits for the users**





Storm
Distributed and fault-tolerant realtime computation

PRACE

egi

Challenges

Distributed Model, Limitations and Workflows Adaptation

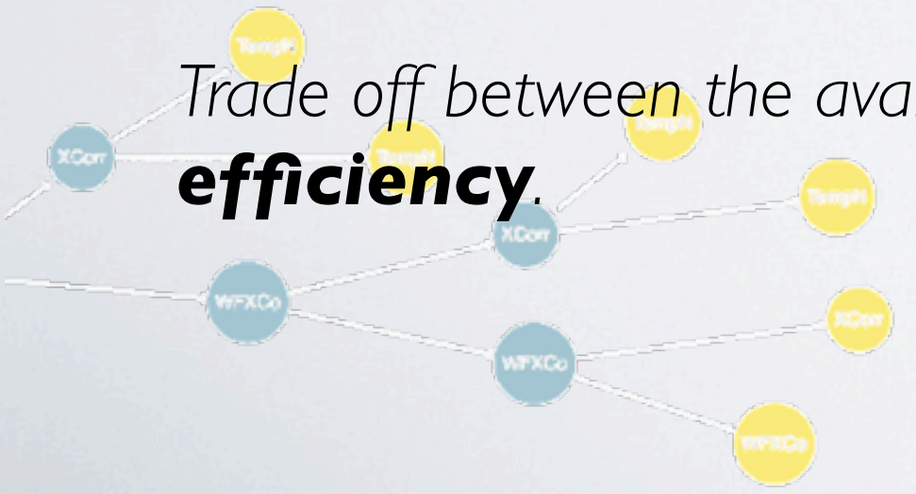
Distribution of computation across:

Computational models (e.g. HPC or Data Intensive)

Institutional infrastructures:

Security policies interfere with consistent and **rapid data transfer and provenance updates**

Towards User Driven Preservation and Cleanup strategies



Trade off between the availability of ***consistent*** *data views at* ***runtime*** *and* ***efficiency***.

Thank you!