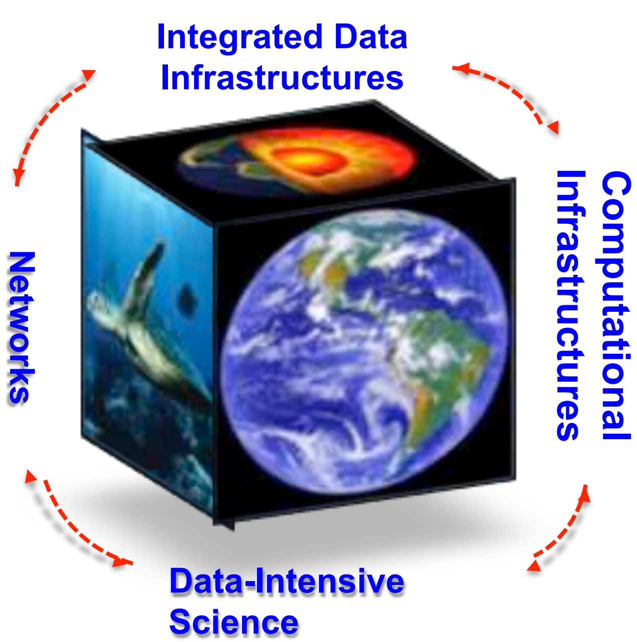
HPC/HDA: Multi-source Data Analysis and Data assimilationChallenges in Earth Systems and Universe Sciences

Jean-Pierre Vilotte

Scientific Deputy (DS) Intensive Computing and Data, Institut des Sciences de l'Univers (INSU), CNRS (France)
Institut de Physique du Globe de Paris

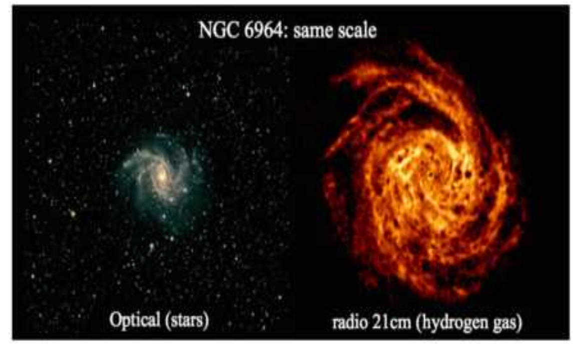


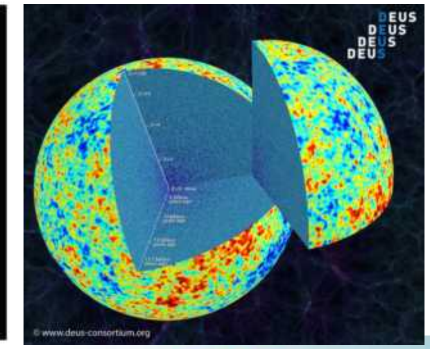
French-German-Japanese workshop
HPC/HDA convergence
Tokyo, 6-8 November 2019

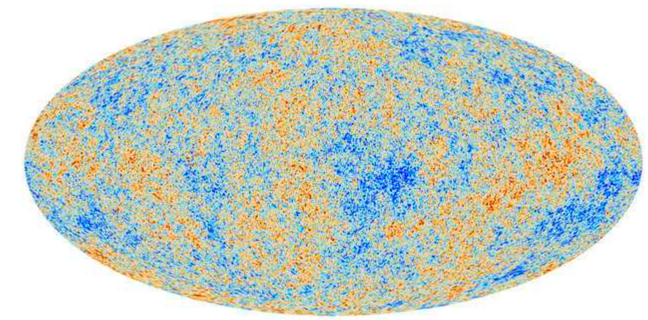
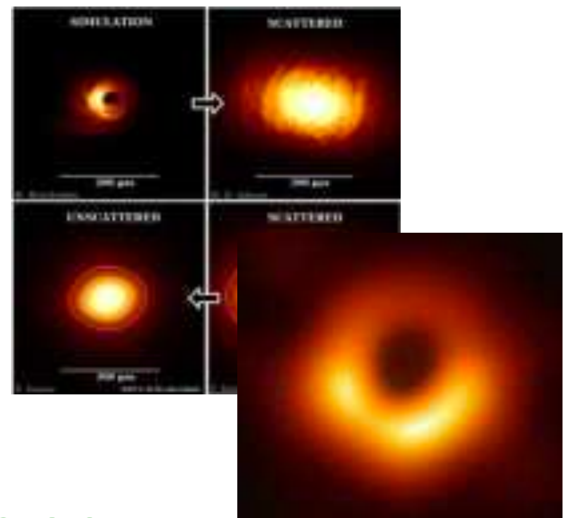
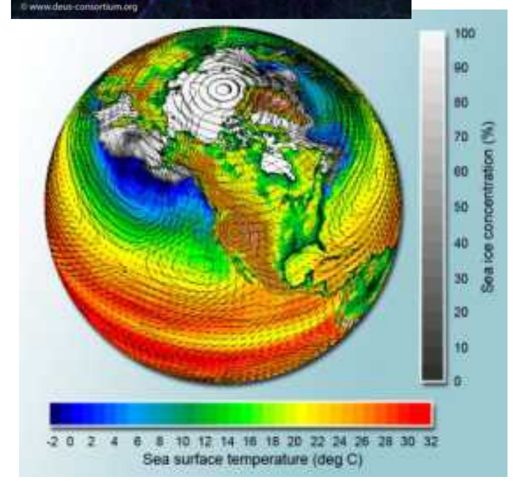
CNRS-INSU: Fundamental knowledge to sustainability

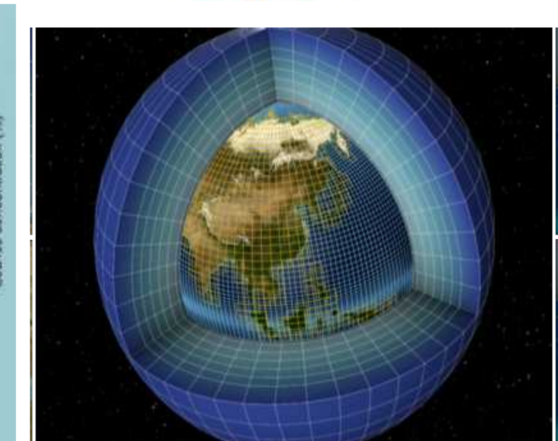
Scientific Discoveries

- Observation and monitoring
- HPC Simulations
- Multi-source data analysis
- Inversion/assimilation
- Machine learning & AI
- Uncertainties & extreme events



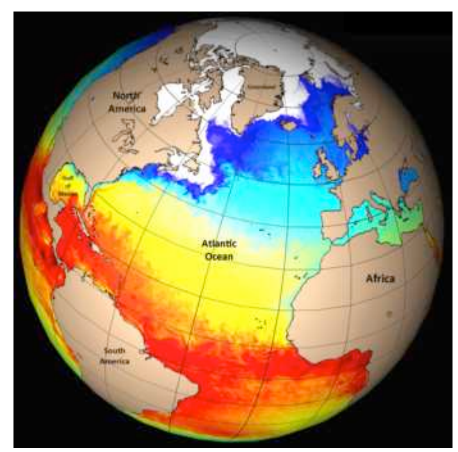


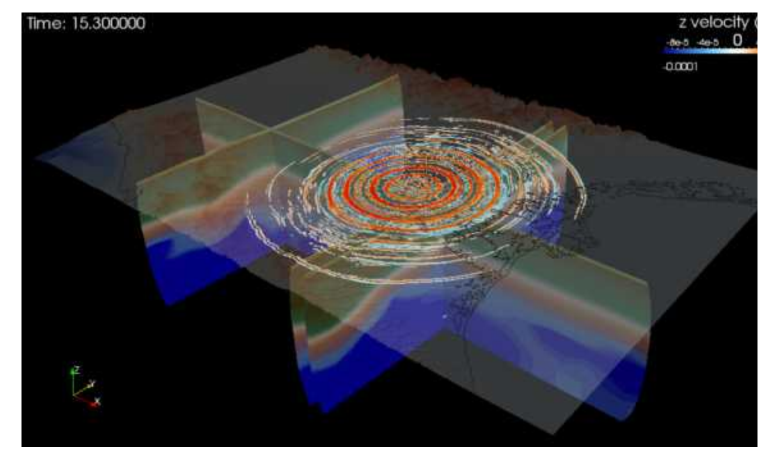






Transversal HPC/HDA challenges

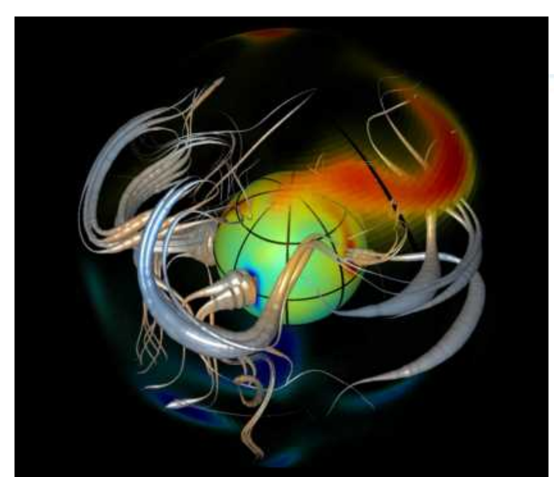
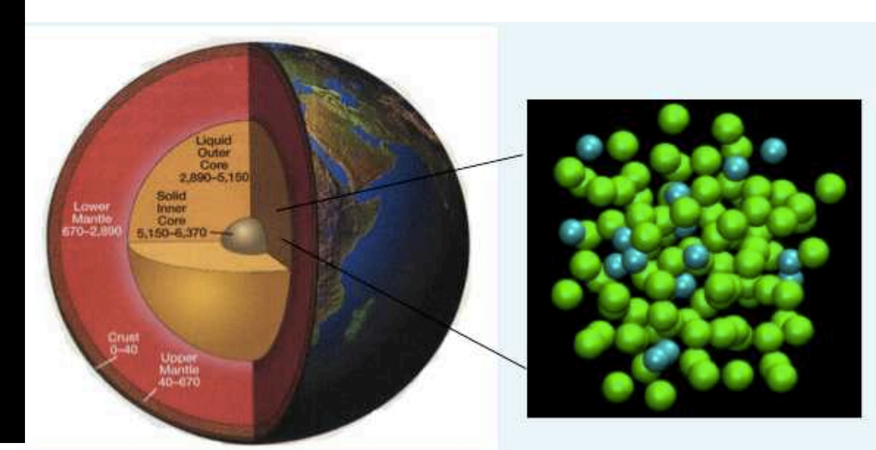
- Astronomy & Astrophysics
- Climat, Atmosphere, Ocean
- Solid Earth Sciences
- Continental surfaces and interfaces





Socio-economical applications

- Climate evolution
- Natural Hazards & Environmental changes
- Energetic resources
- Sustainable environmental goals

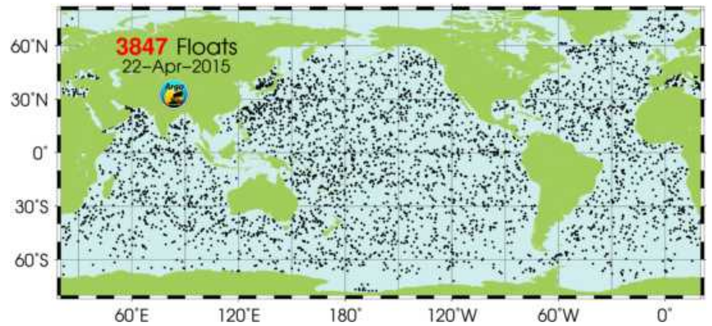



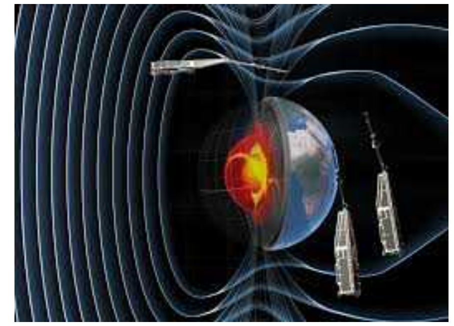
Data flux explosion and diversity

Ubiquity and explosion of data

NenuFAR/SKA

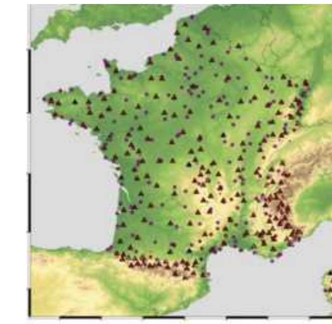
ARGO





Swarm mission

SVOM

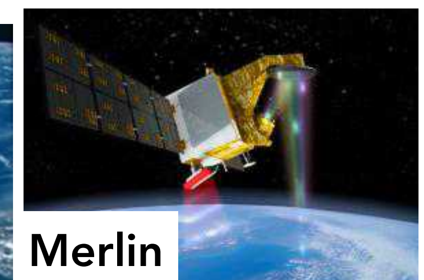


Seismic/geodesy

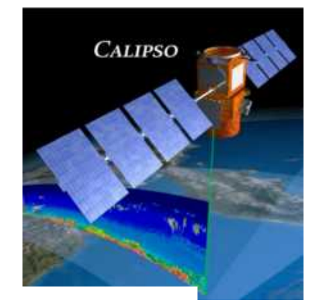
CFHT

Hayabusa2-Mascot module

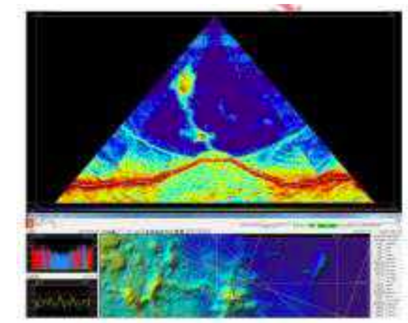
Copernicus



Merlin

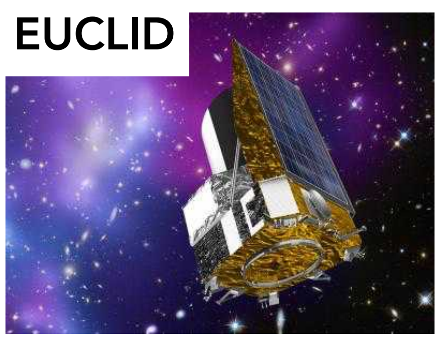


Calipso



Active volcanos

EUCLID



InSight mission

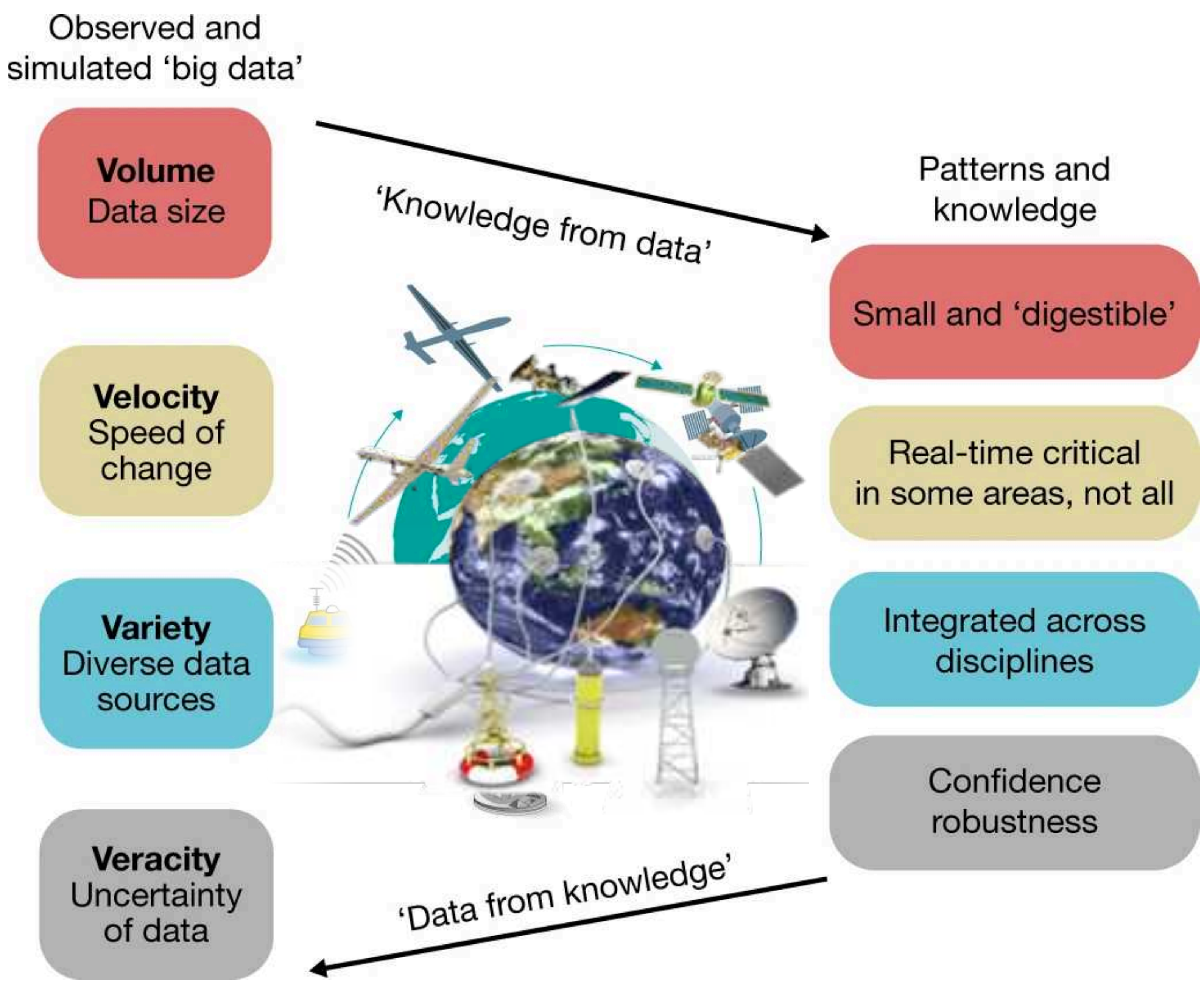
Data explosion (rate, volume, diversité):

- **Edge environments**: observation acquisition systems
- **Centralised environments** (Cloud and HPC): large ensemble simulations, HDA, data assimilation

New challenges:

- **Acquisition**: streaming data processing/reduction/compression -> primary data delivery
- **Data Management**: long-term archiving & curation (metadata, provenance, distribution)
- **HDA**: multi-source distributed data statistical analysis, ML
- **HPC**: ensemble of multi-physics and multi-scale simulations, data assimilation, ML
- **Data Distribution:** multi-source FAIR data services, virtual observatories

Big Data Challenges



BigData Challenges

- Flux rate, volume, diversity
- Multi-source, multi wavelength
- Reprocessing and versioning
- Large ensemble simulations
- Interdisciplinary and transdisciplinary

Data Policy and management

- Open Data by default, FAIR data services
- Long-term archiving and curation
- Data veracity, certified repositories
- Software management and certification

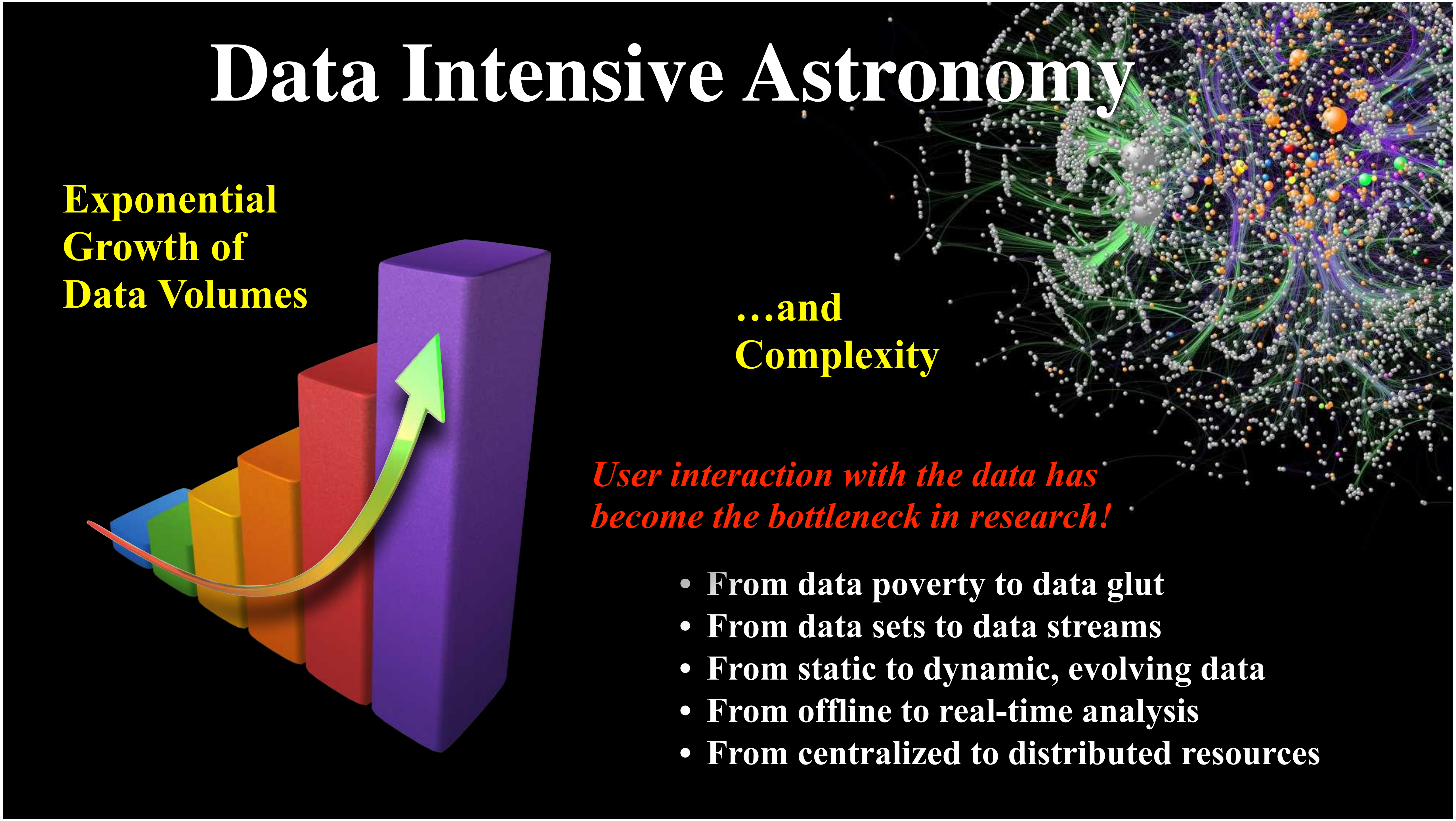
Statistical challenges

- Multi-temporal, multi-angular, multi-source
- Non-linear and non-Gaussian
- Data and systemic uncertainties,
- Extreme events

Machine learning challenges

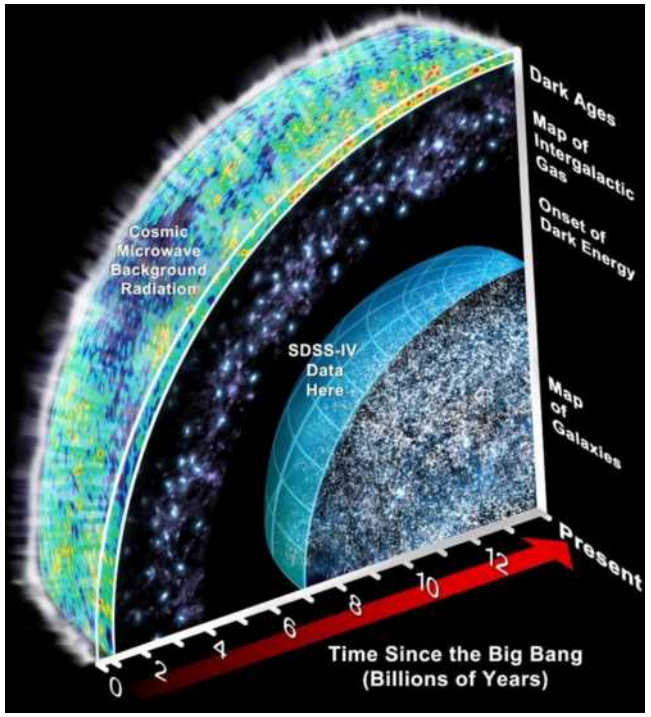
- Few supervised information available
- Computationally intensive and timeliness
- Consistency, learning and interpretability
- Multi source uncertainty propagation

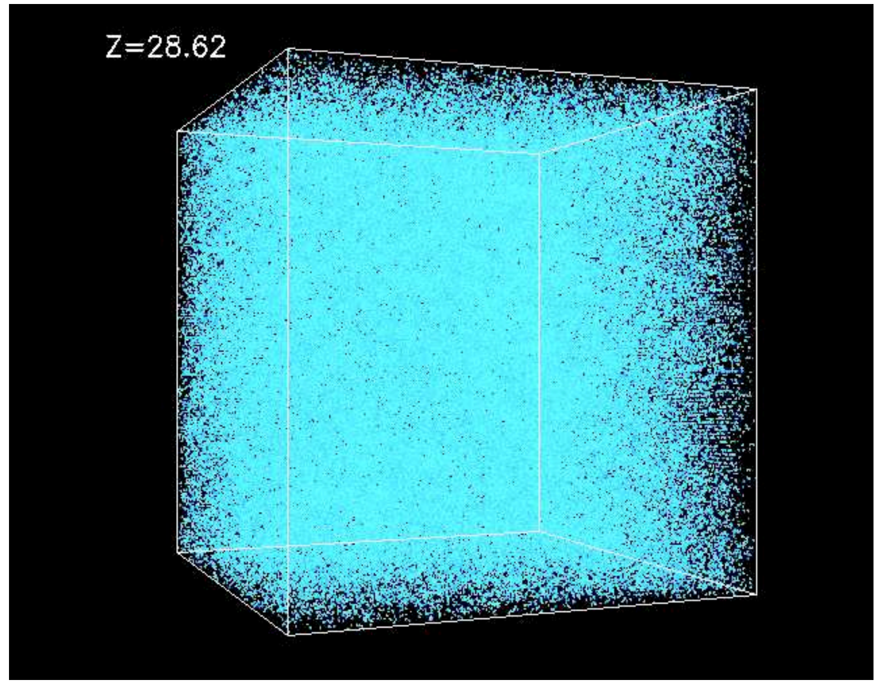
Data-intensive astronomy



- Science increasingly driven by large data sets; massive multi-source, multi-wavelength data
- Large interdisciplinary scientific collaboration
- Science extraction: distributed FAIR data services across instruments (multi-messager)
- Increasing use of ML/DL: data analysis and HPC simulations

Astronomy and SKA





Cosmic dawn
(First stars & Galaxies)

Cosmology
(Dark matter, Large-scale structures)

Galaxy evolution
(gas content & new stars)

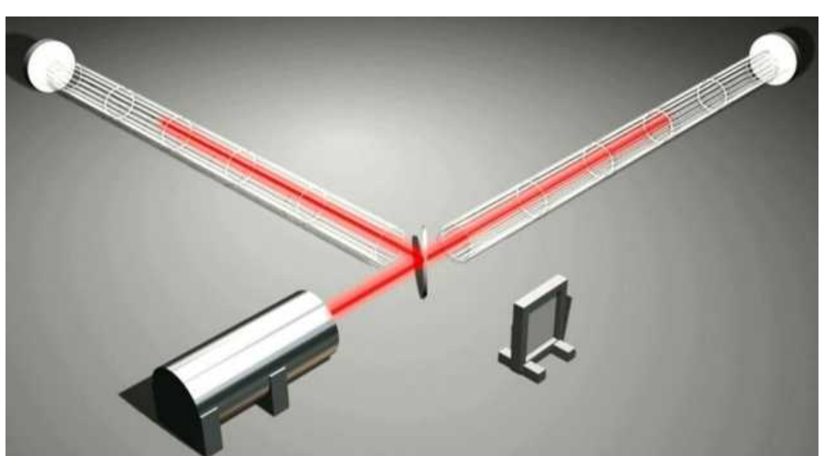
Era of big surveys

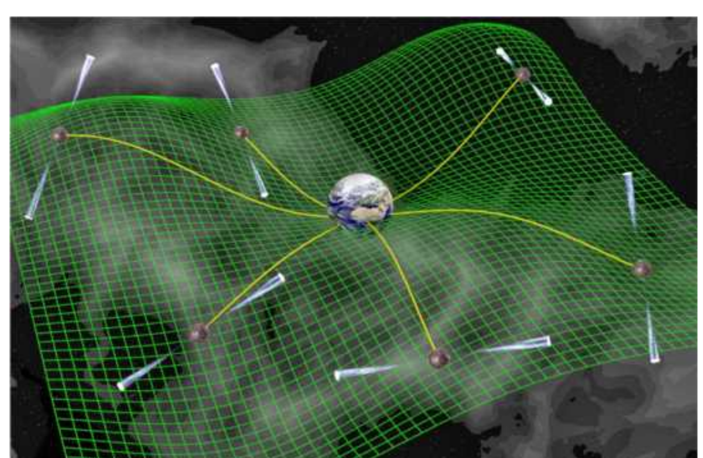
LSST: 160 MB/s, ~1.3 TB/night ,~ 30 PB over 5 yrs archived data
LOFAR: ~100 TB/night, ~6-10 PB/yr archived data
CTA: 3-10 PB/yr archived data
SKA: 0.1-3 EB/yr archived science-ready data

Cosmic magnetism
(origin & evolution)

Fundamental physics
(gravitational waves & compact objects)

Cradle of life
(Planets, Molecules, SETI)

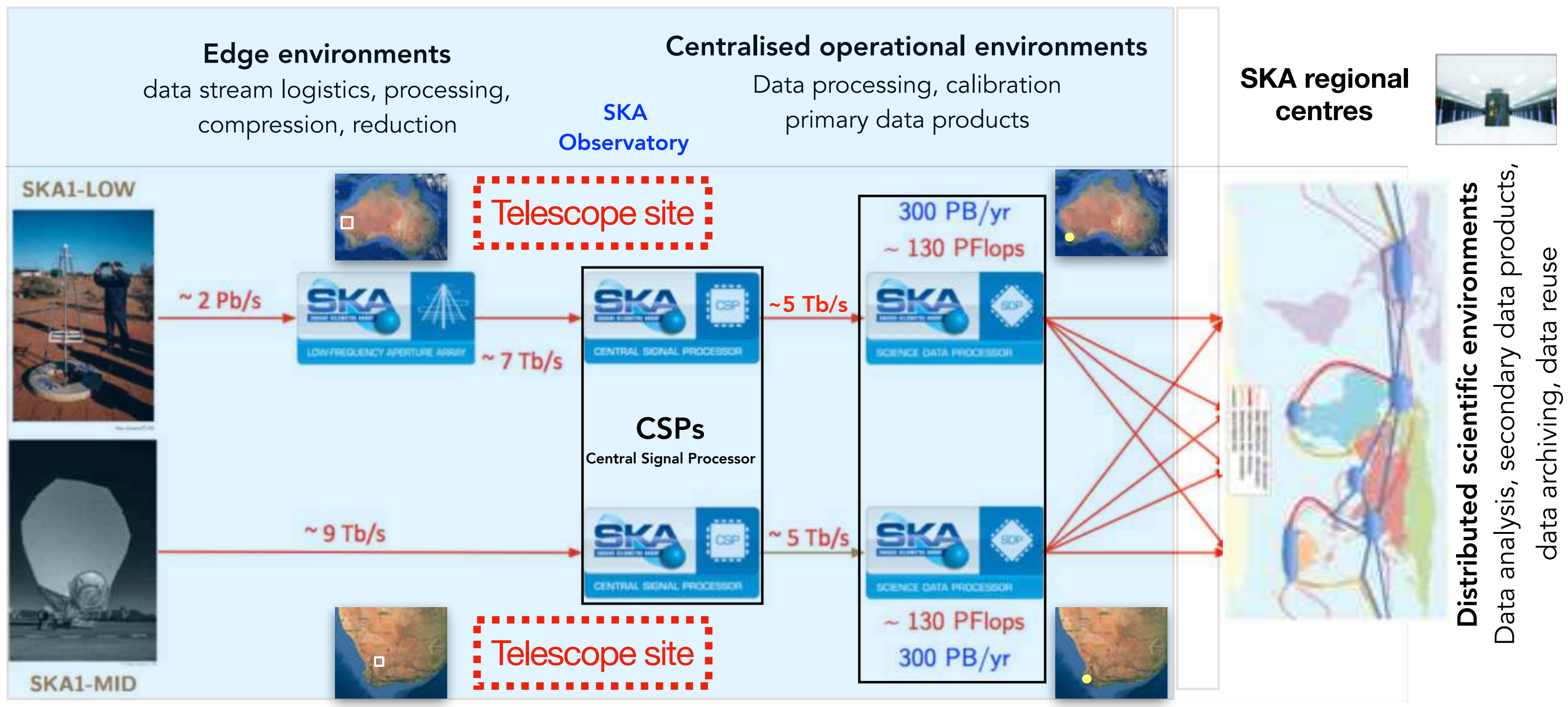




Modern sky surveys: ~10^{12} - 10^{18} bytes images; Catalog: ~10^{8}-10^{9} objects (stars, galaxies, etc.)

SKA: community driven BigData pathfinder

On Line processing: high-rate data streams



SKA observatory

From edge -> centralised infrastructures

- High-rate data stream logistics
- Stateful network services : caching/buffering
- Edge computing: numerical beam forming of signals ; removal of radio-frequency interference
- Data loss-compression and reduction
- Dynamic stream structures: observation dependent

Centralised HPC/HDA operational infrastructures

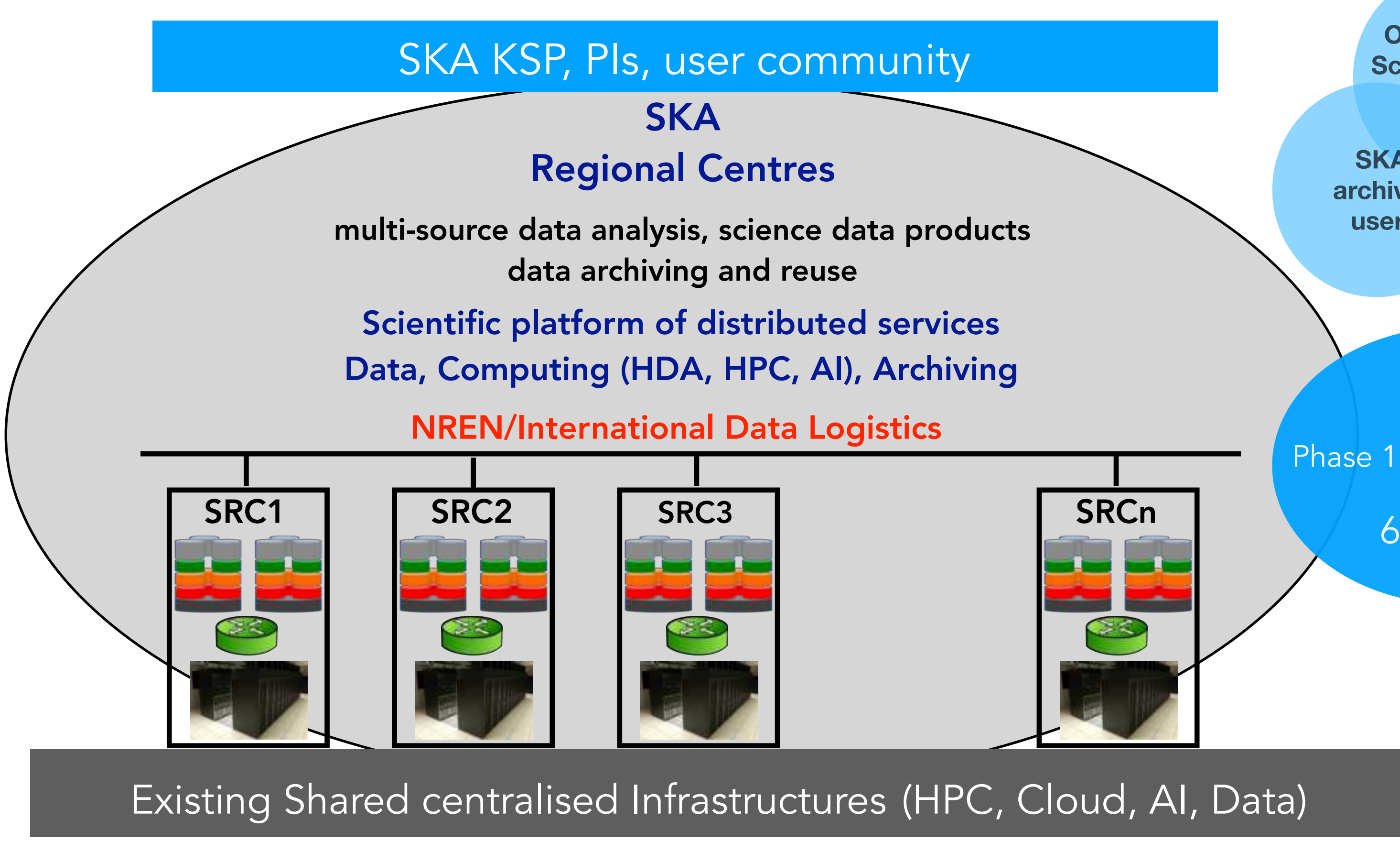
- Storage and computing capabilities/capacities
- High-rate data processing
- Complex HDA workflows (processing & calibration)

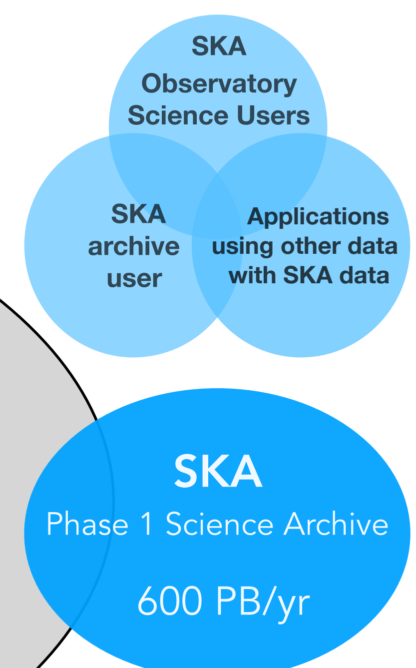
Primary data productS (events, images, cubes)

- Data models (standards, metadata, provenance)
- Archiving and dynamic distribution (data placement)

- > Machine Learning moving to the edge

Off Line SRC processing: multi-providers context





SKA Regional Centres (SRCs)

New organisational, operational, business model

- Community-driven shaping strategy
- Co-designed (SKAO, providers, scientific users)

Scientific software platform

- Distributed services across shared infrastructures
- Multi providers (Cloud, HPC, Data), Federated AAI
- Application-dependent global resource optimisation

Application workflows

- Diversity of complex workflows (HDA, HPC, AI)
- Data logistics all along in multi-provider context
- Workflow management and provenance system

Data archiving, curation and reuse

- Primary and secondary scientific data products
- FAIR multi-source data services (federated)

Scientific Users

- Key SKA Projects and PI granted observation projects
- Reuse of SKA data products: multi-messenger and multi-wavelength approaches

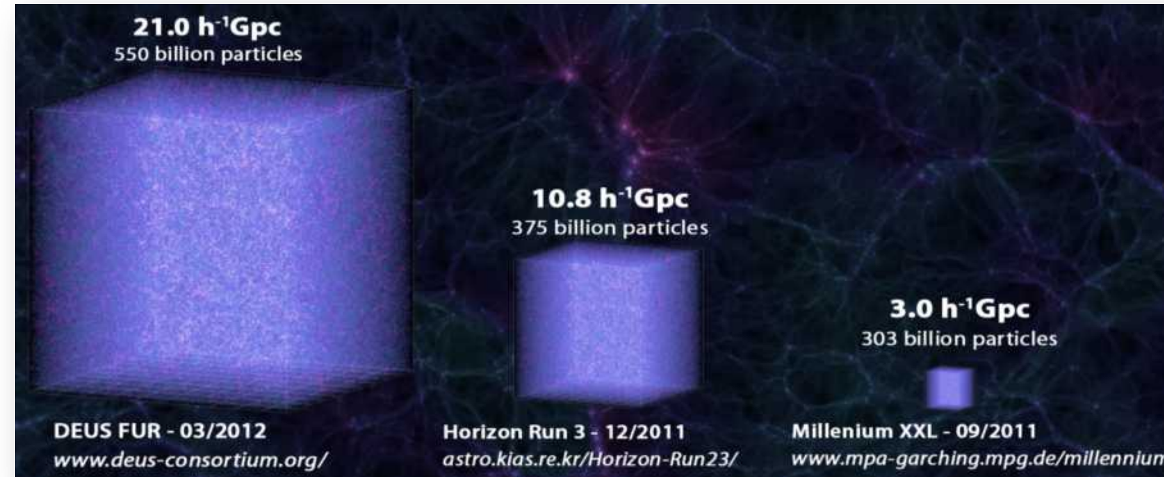
-> HPC/HDA in centralised infrastructures

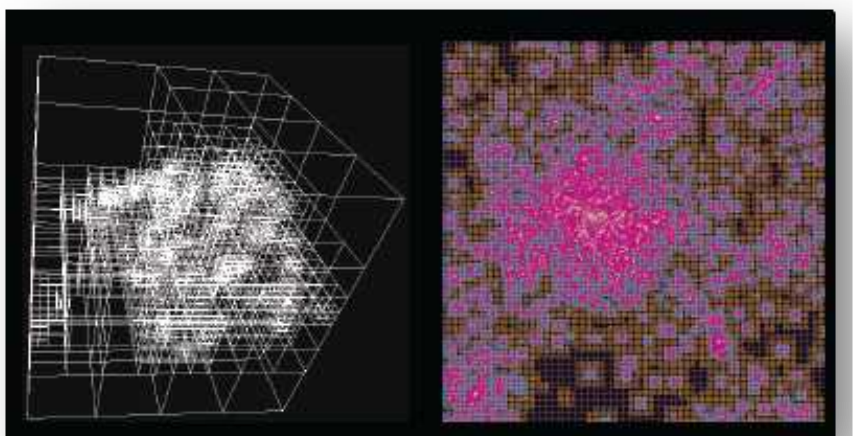
Shared with other communities: Space Observation, Earth Systems Observation, HEP

AMA-DEUS: N-Body simulation

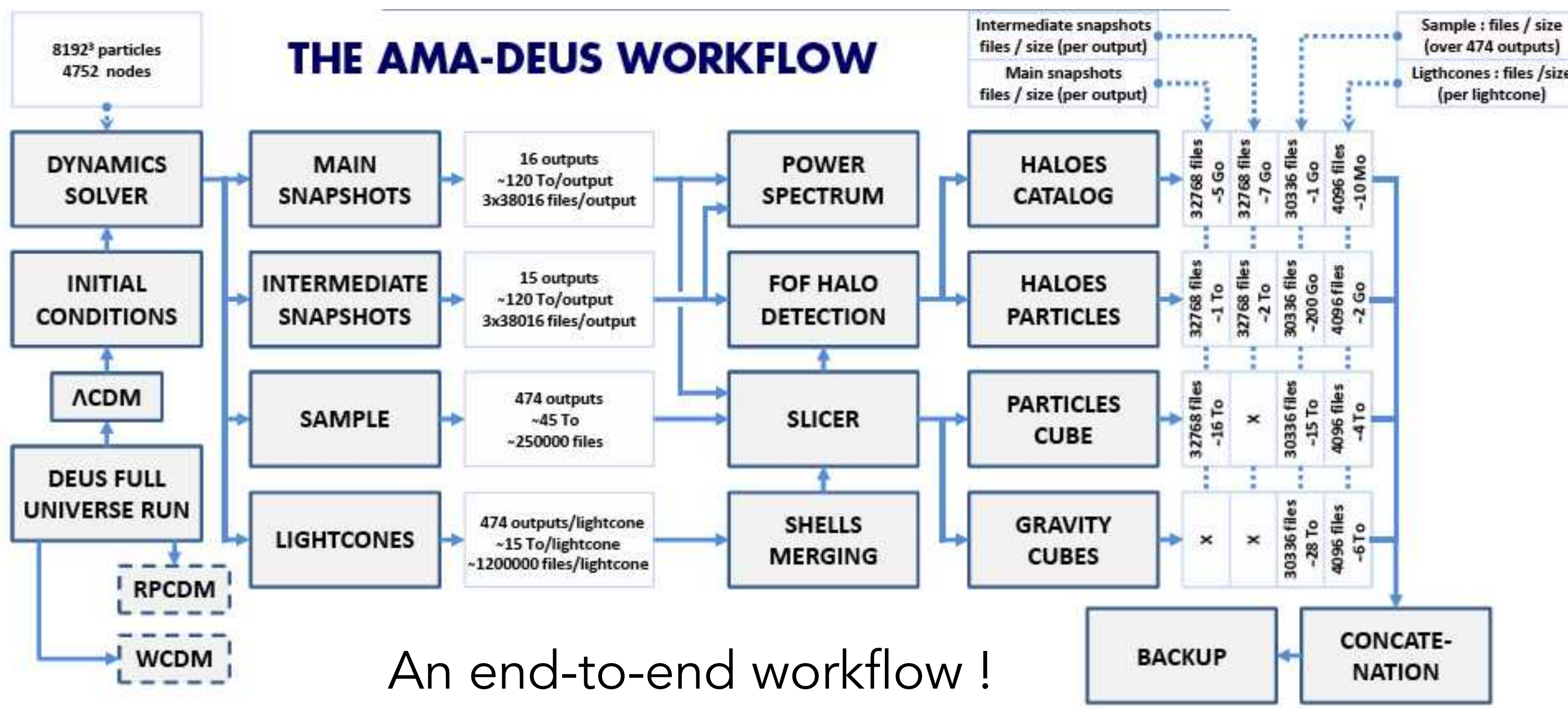
HPC grand challenge

- 550 billion particles
- 2.5 trillion computing points
- 50 million CPU hours (> 5700 years)
- 76 032 cores & 300 Tb memory
- > 50 Gb/s data throughput (PFS)
- 1 500 Pbs reduced on fly to 1 500 Tbs





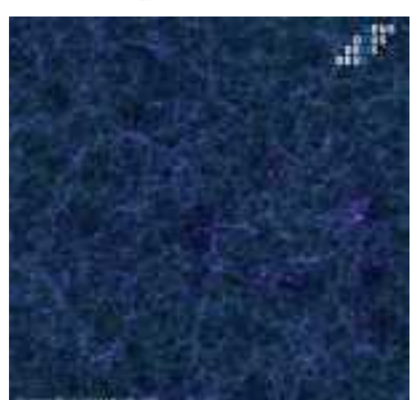
Alimi et al



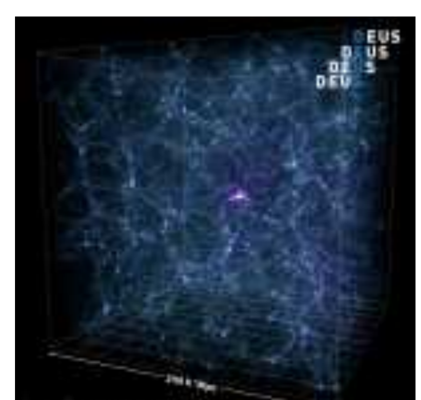
An end-to-end workflow !

Challenges

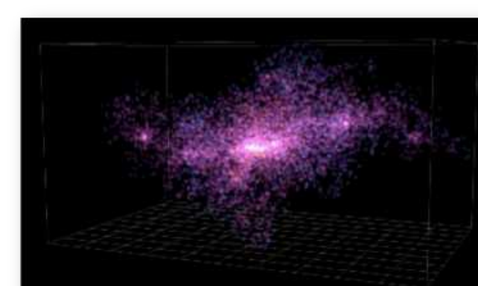
- dynamic load balancing
- smart parallel I/O optimisation
- reduction of raw data (time) -> in-situ & post processing
- physical objects -> on-the-fly processing workflow



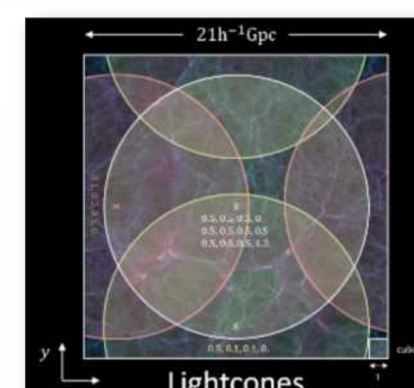
Snapshots ~16 x 16 TB



Samples ~40 TB

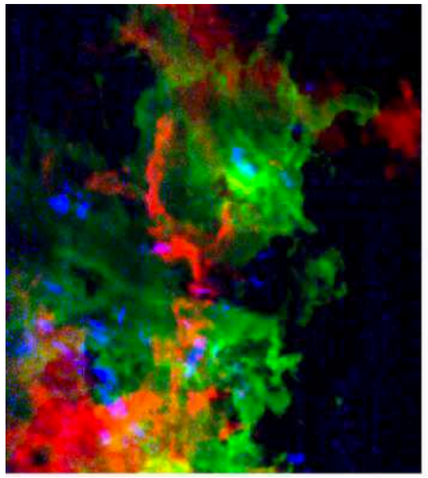


Halos/catalogs ~50 TB



Lightcones ~ 5x10 TB

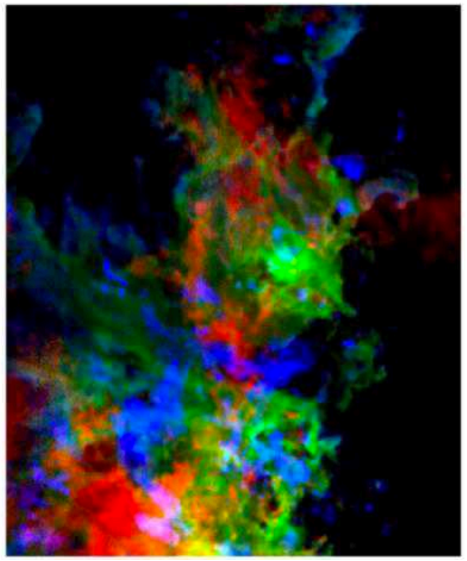
DL: de-noising & analysis hyper-spectral imaging radio astro



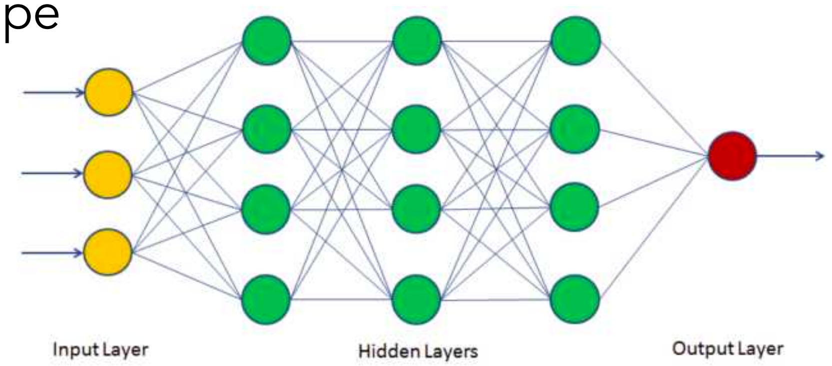
Observed image
spectral band 30/40/50
IRAM 30 m Telescope

Orion-B
Pety et al, 2017

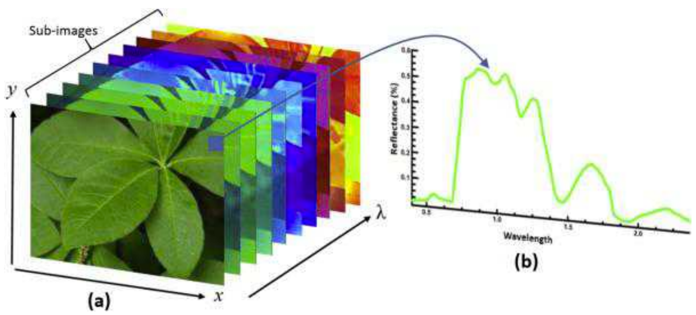
Reconstructed image
spectral band 30/40/50

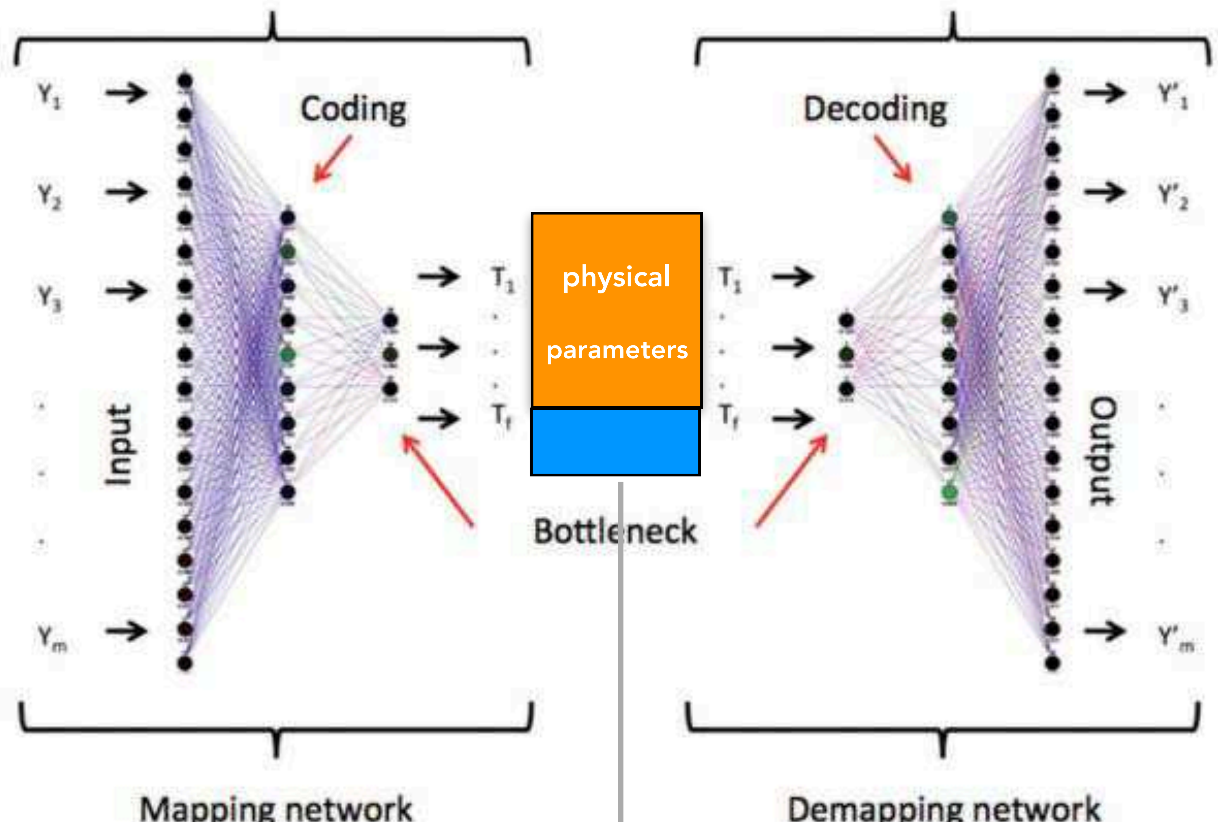


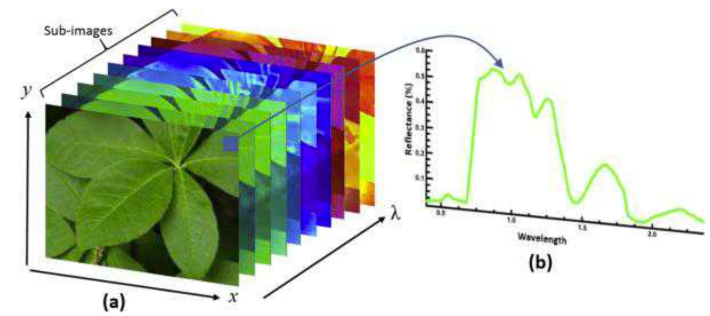
$$\overrightarrow{\text{sample}\, n} = \begin{pmatrix} \text{Band}_1 \\ \text{Band}_2 \\ \dots \\ \text{Band}_{79} \\ \text{Band}_{80} \end{pmatrix}$$



$$\overrightarrow{\text{sample}\, n} = \begin{pmatrix} \text{Band}_1 \\ \text{Band}_2 \\ \dots \\ \text{Band}_{79} \\ \text{Band}_{80} \end{pmatrix}$$



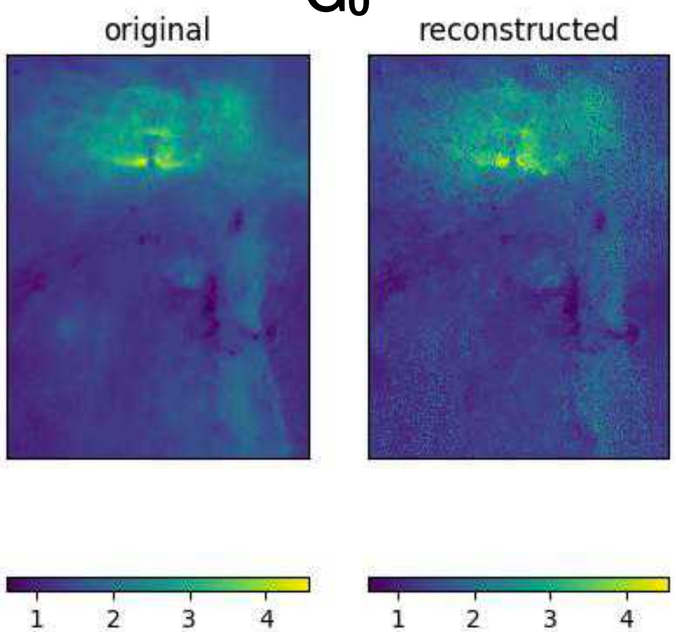




80 spectral bands
332 * 551 points
>160 000 images

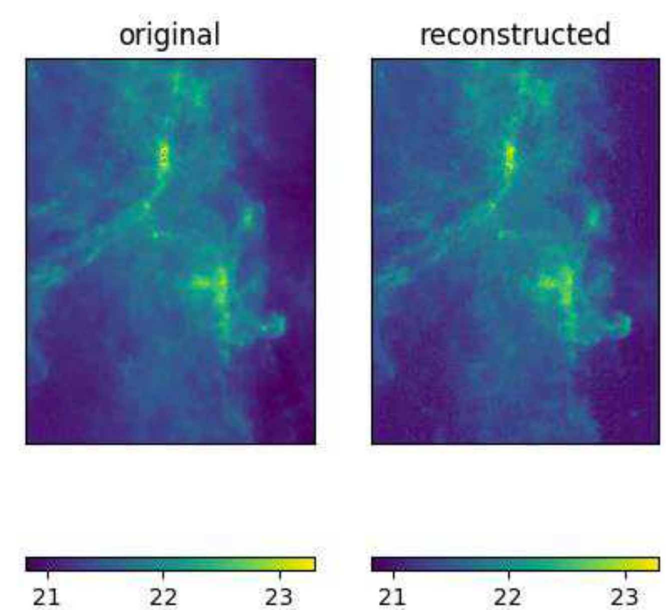
Vandame, Chanussot, Pety (2019)

G_0



$$\overrightarrow{\text{sample}\, n} = \begin{pmatrix} \text{column}-\text{density} \\ G_0 \\ \text{volume}-\text{density} \end{pmatrix}$$

column density



40 000 parameters >> 10 015 200 information

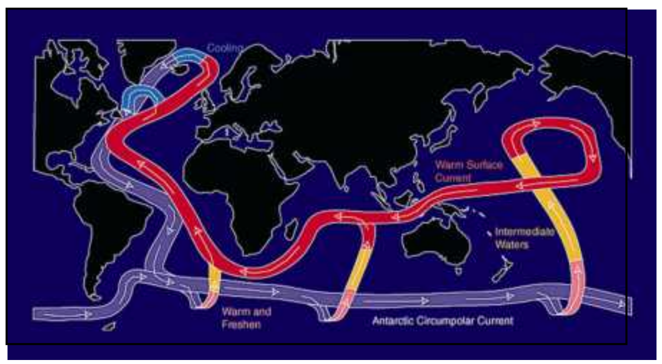
NLCA (noise) - Hybrid NLCA (Physical data)

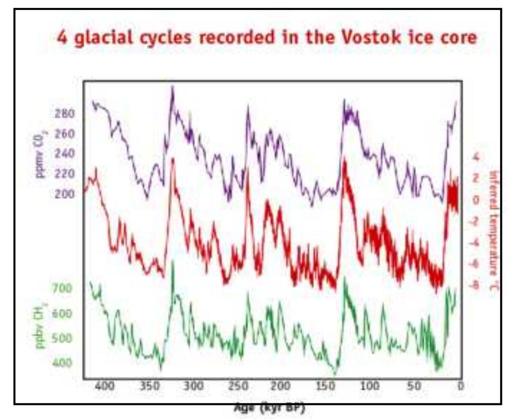
Climate system: a scientific and societal chalenge

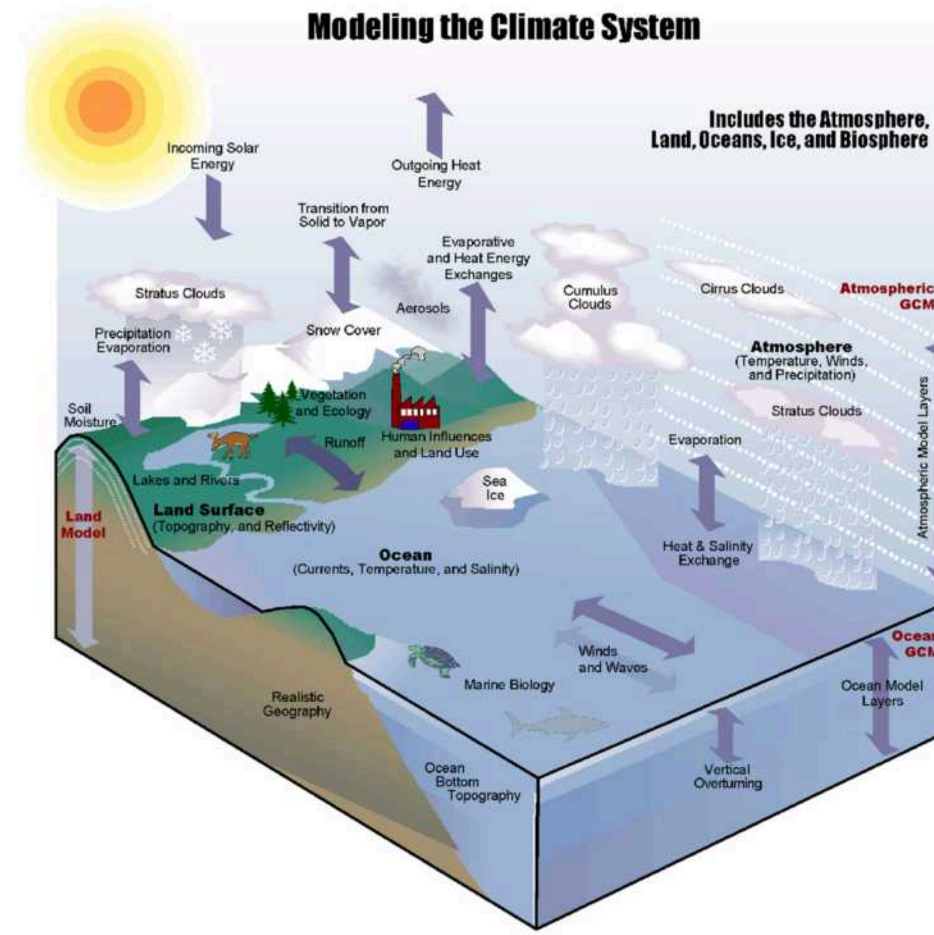
Several complex and multi-physics processes to be simulated

Several interacting processes

Large range of time scales: from days to months, years, decades and millennia

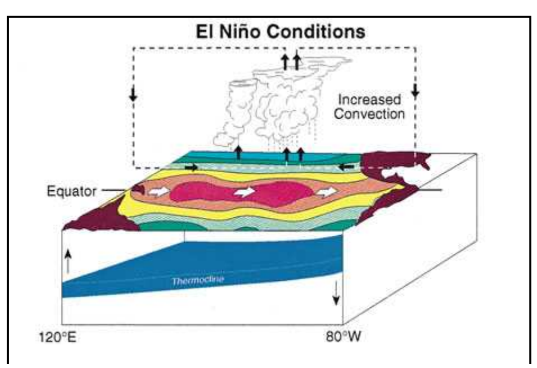


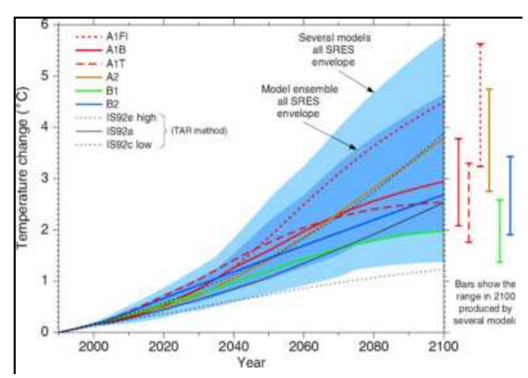




Comprehensive modelling of climate systems and variability

Large range of space scales: from local to regional, continental and global





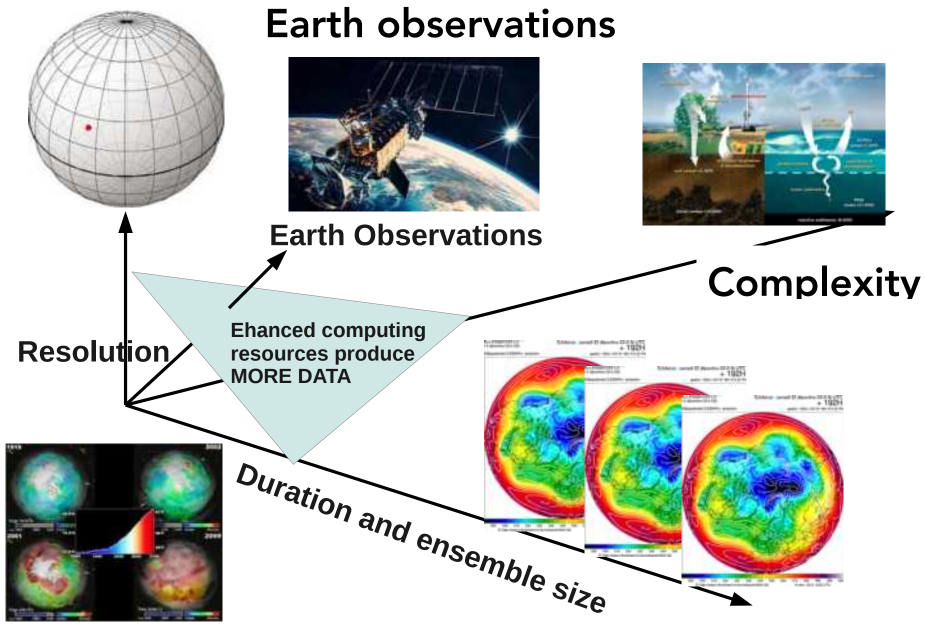
Understanding detection, attribution and prediction of extreme events and modes of climate variability

Inherently non-linear dynamical Earth systems

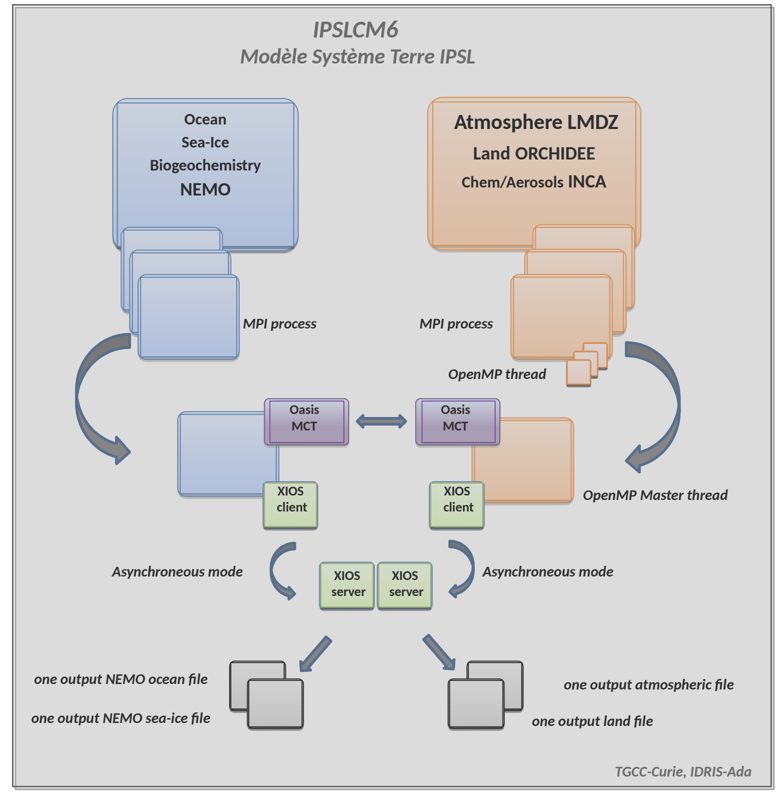
CPU demanding <-> large volume of data

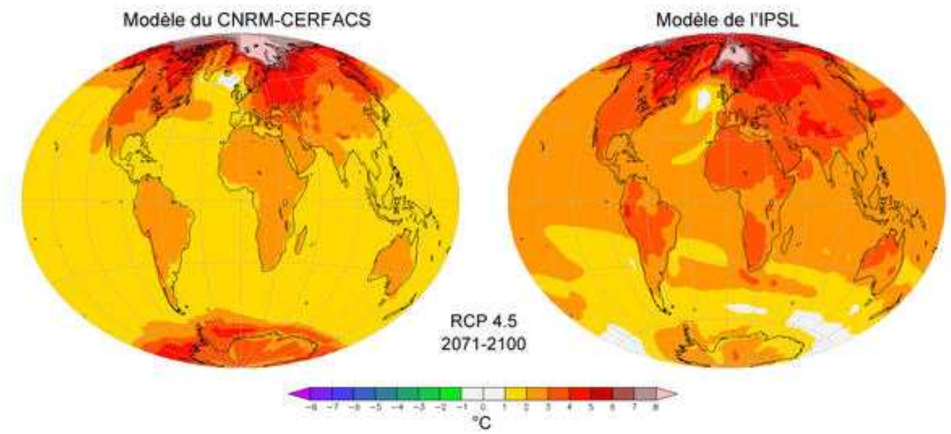
Climate science, impacts and societal services

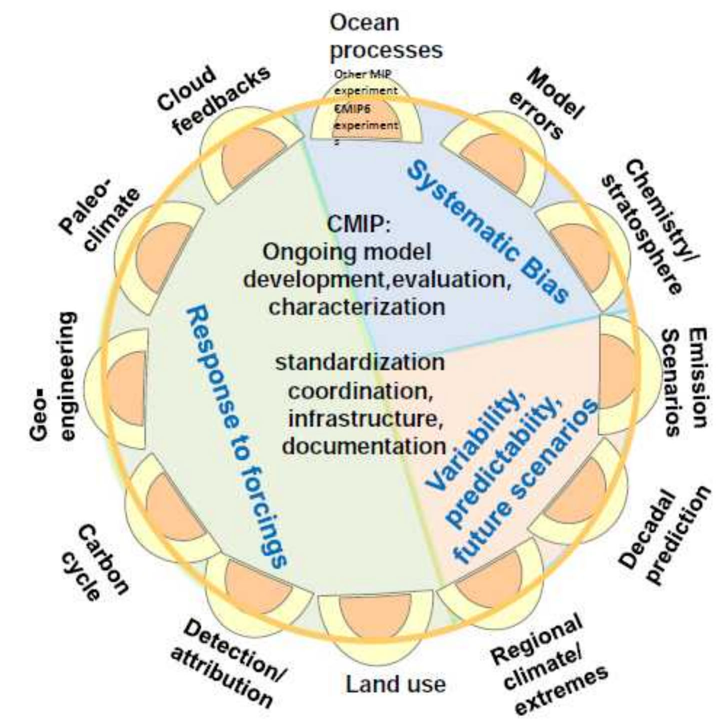
Climate simulations and observations



~ 10 PBs scale

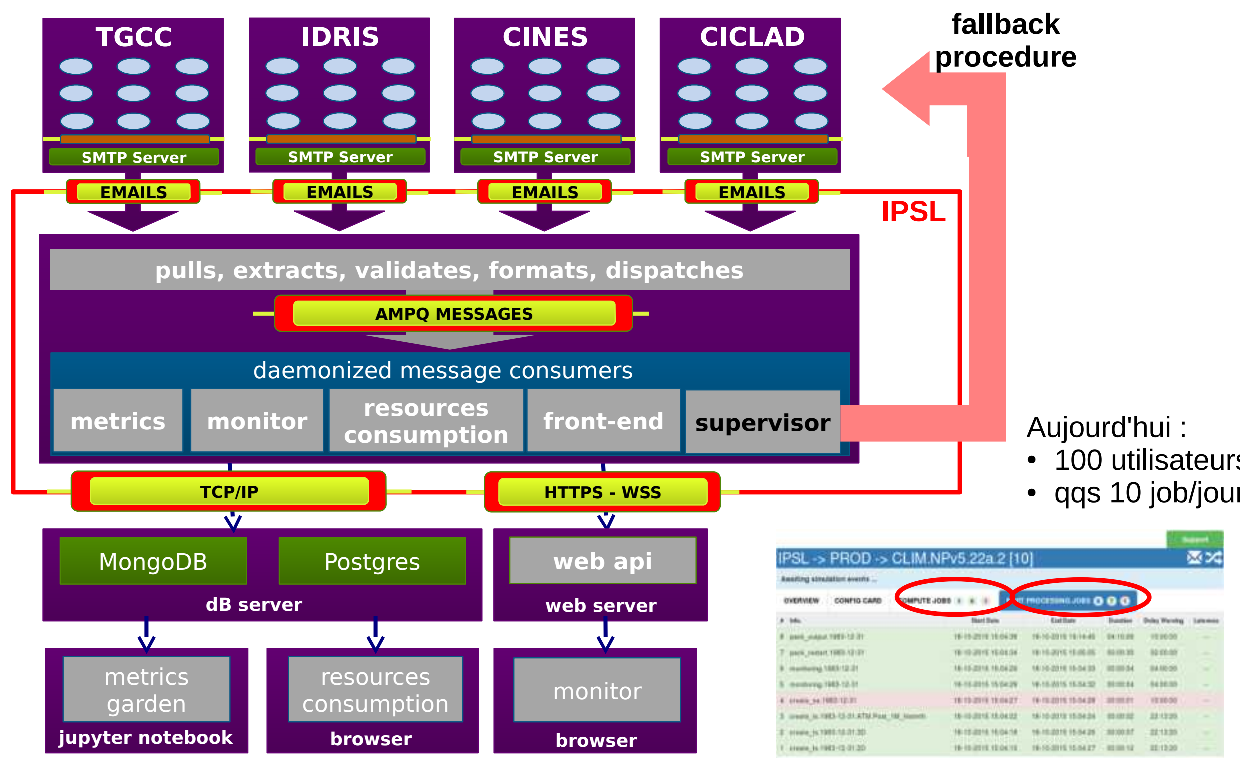






A **number of** **models:** *configurations* (parameterisation), *experiences* (scenarios), *ensemble of realisations* (uncertainty)

Large number of **variables:** large volumes of **data** and number **of files**

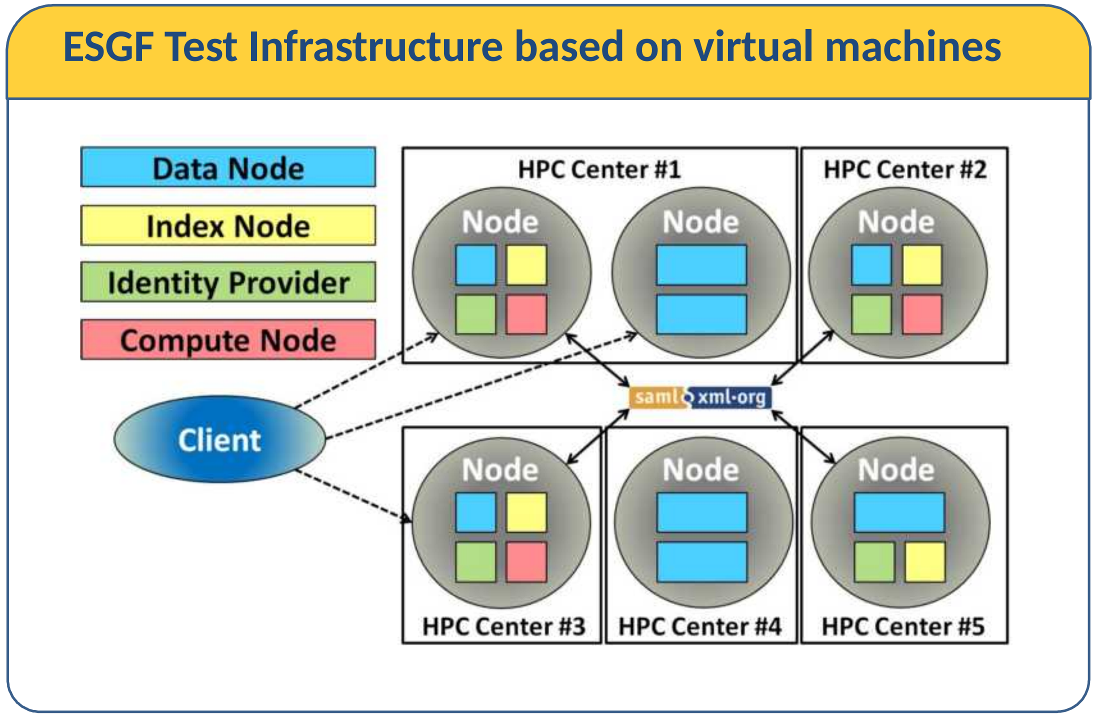


Aujourd'hui :
- 100 utilisateurs
- qqs 10 job/jour

Flexible **provenance-driven** system

- Provides ***run-time feedback*** with ***tuneable metadata*** and ***provenance-driven controlled data movement***
  - ✳ Avoids useless waits for long and unfruitful runs
  - ✳ Fosters dynamic steering, diagnostics (saving computing cycles, storage and energy!)

Numerical laboratory: Earth System Grid Federation

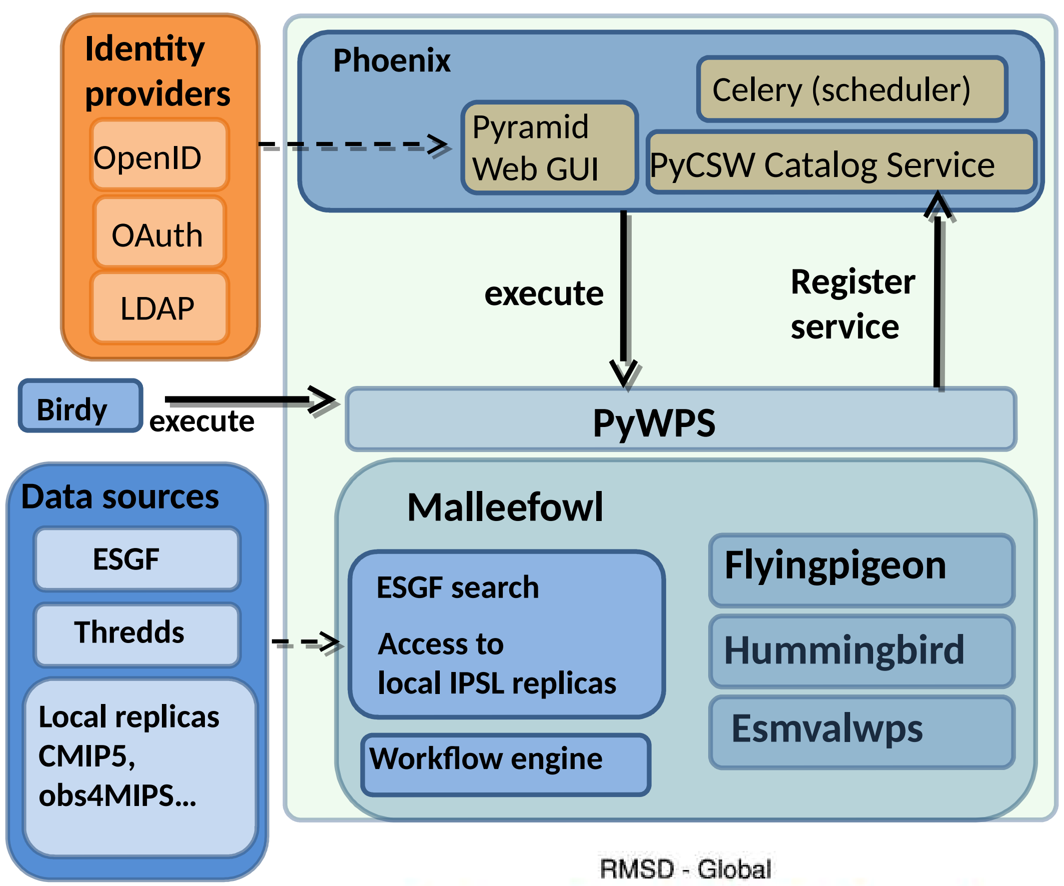


~ PBs scale

International Climate Networking Group

Climate Model Assessment Framework (CLiMAF)

Web processing services (WPS)

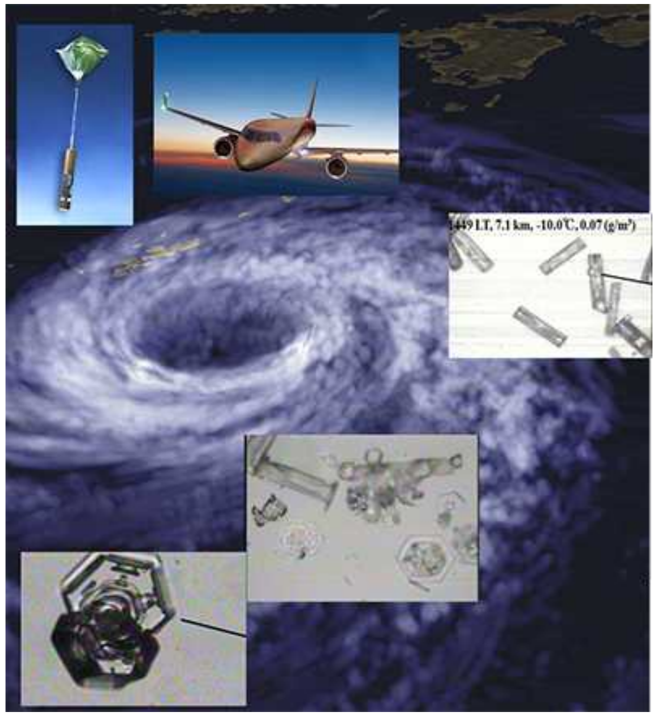


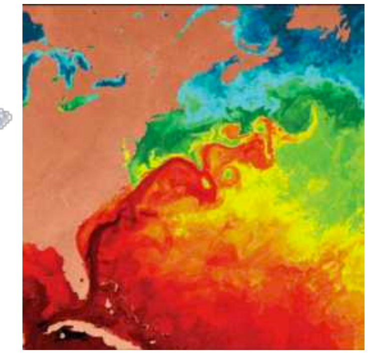
from S. Denvil

- Access to models, simulations and observations
- Share data analytic methods and tools
- Advanced management and documentation of models, simulations (indexation, metadata, provenance)
- Induction of a broad research and user community
- Data analysis platforms and web services
- Pervasive provenance system

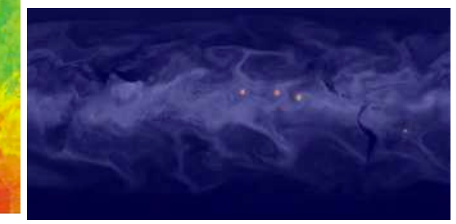
Paleoclimate Modelling Intercomparison Project PMIP3

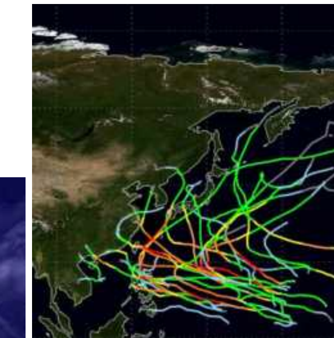
Observation: in situ (land/sea), air and space



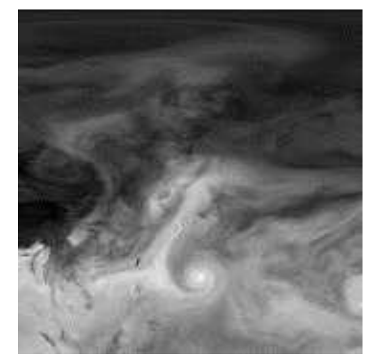


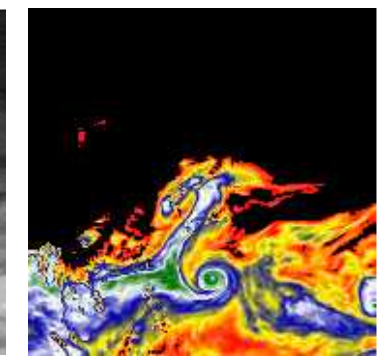
Cyclones

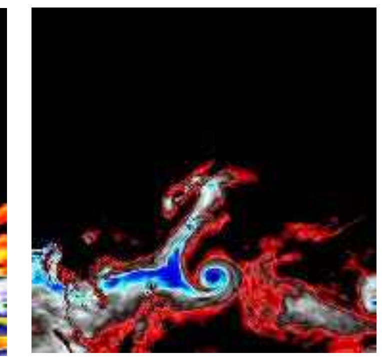




Storms







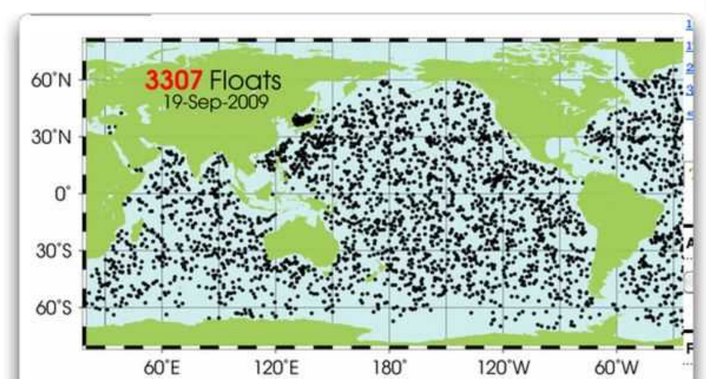
Cloud, aerosol, extreme weather radar

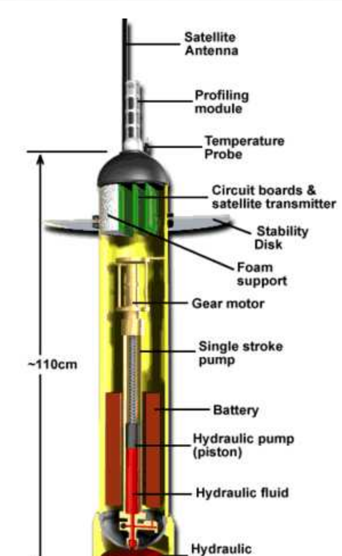
- Ground and satellite cloud observations
- Identify atmospheric instability (convective, baroclinic)
- Monitor various data (e.g. temperature, pressure, humidity)
- Track precipitable water (weather radar) and extreme phenomena (e.g. storm, cyclones)
- Machine learning, Deep learning

Ocean satellite and in situ Argo observation analysis

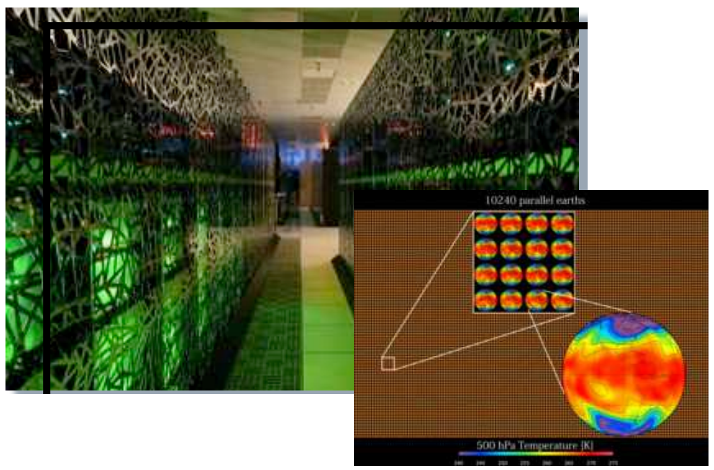
- Large amount of 4D in-situ data (3D space + time)
- Non stationary (mean and covariance) and non Gaussian
- Combined with satellite SSH and mooring data
- Spatial-temporal modelling (adding vertical dimension)
- Machine learning, Deep learning

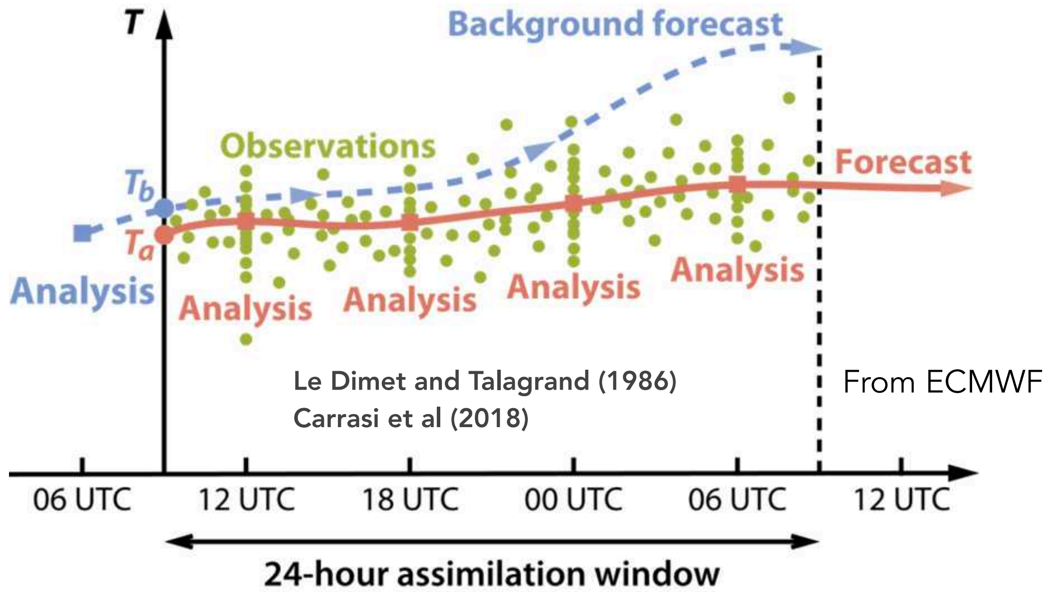
Ocean/Argo

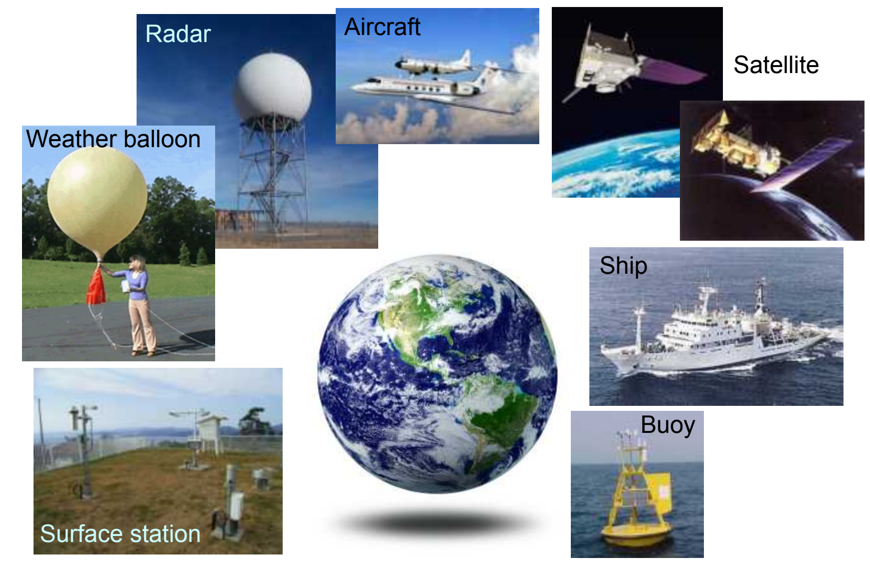




Data assimilation: numerical weather prediction





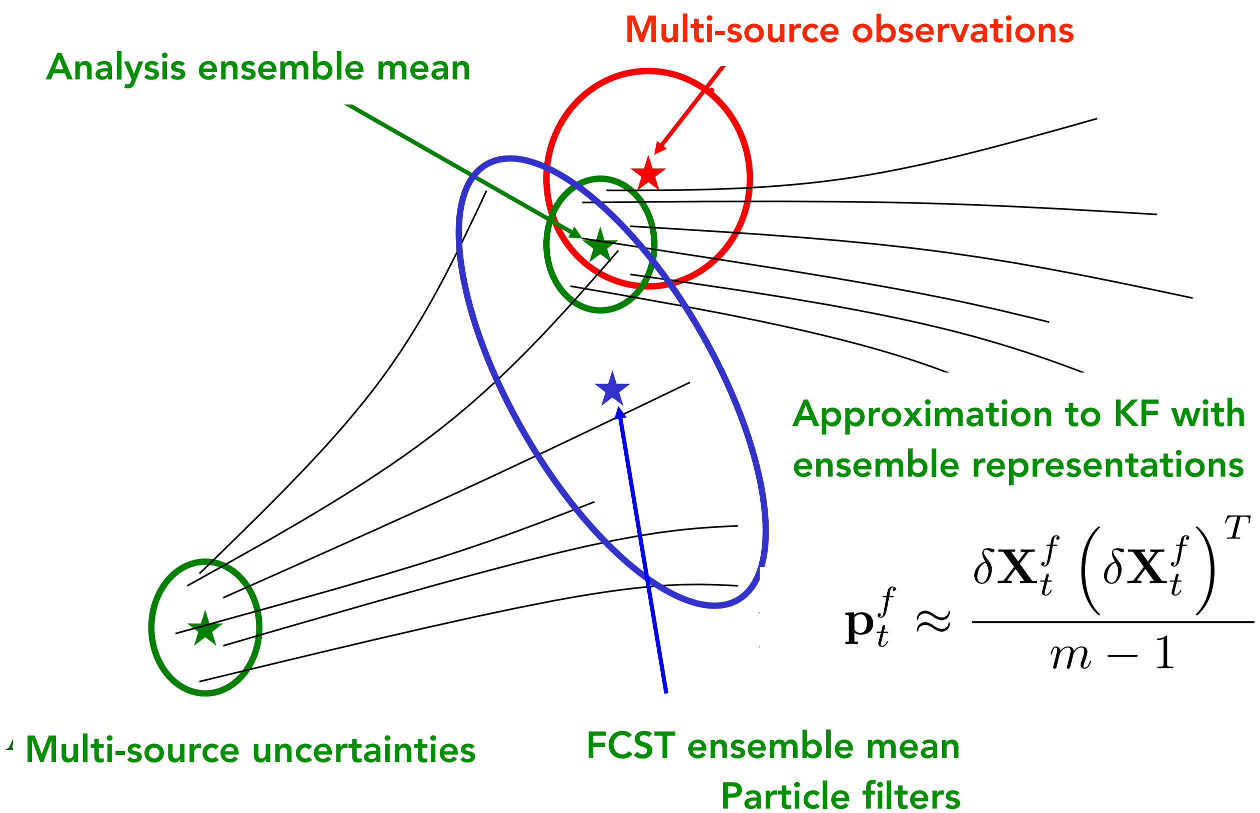


From edge: streaming data processing and reduction

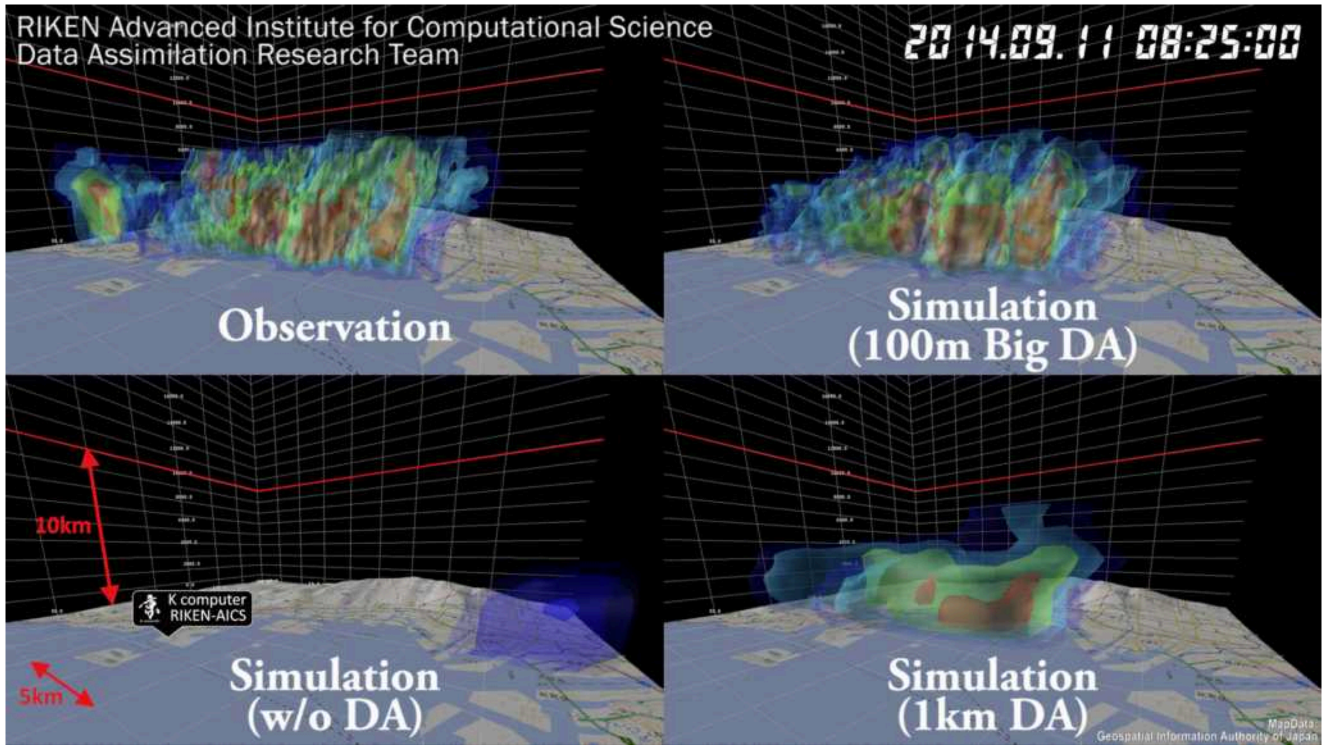
to centralised infrastructures (HPC, Cloud):

large ensemble simulations & Bayesian inference

Multi-source and multi-scale data



$$\mathbf{p}_t^f \approx \frac{\delta \mathbf{X}_t^f \left(\delta \mathbf{X}_t^f\right)^T}{m-1}$$

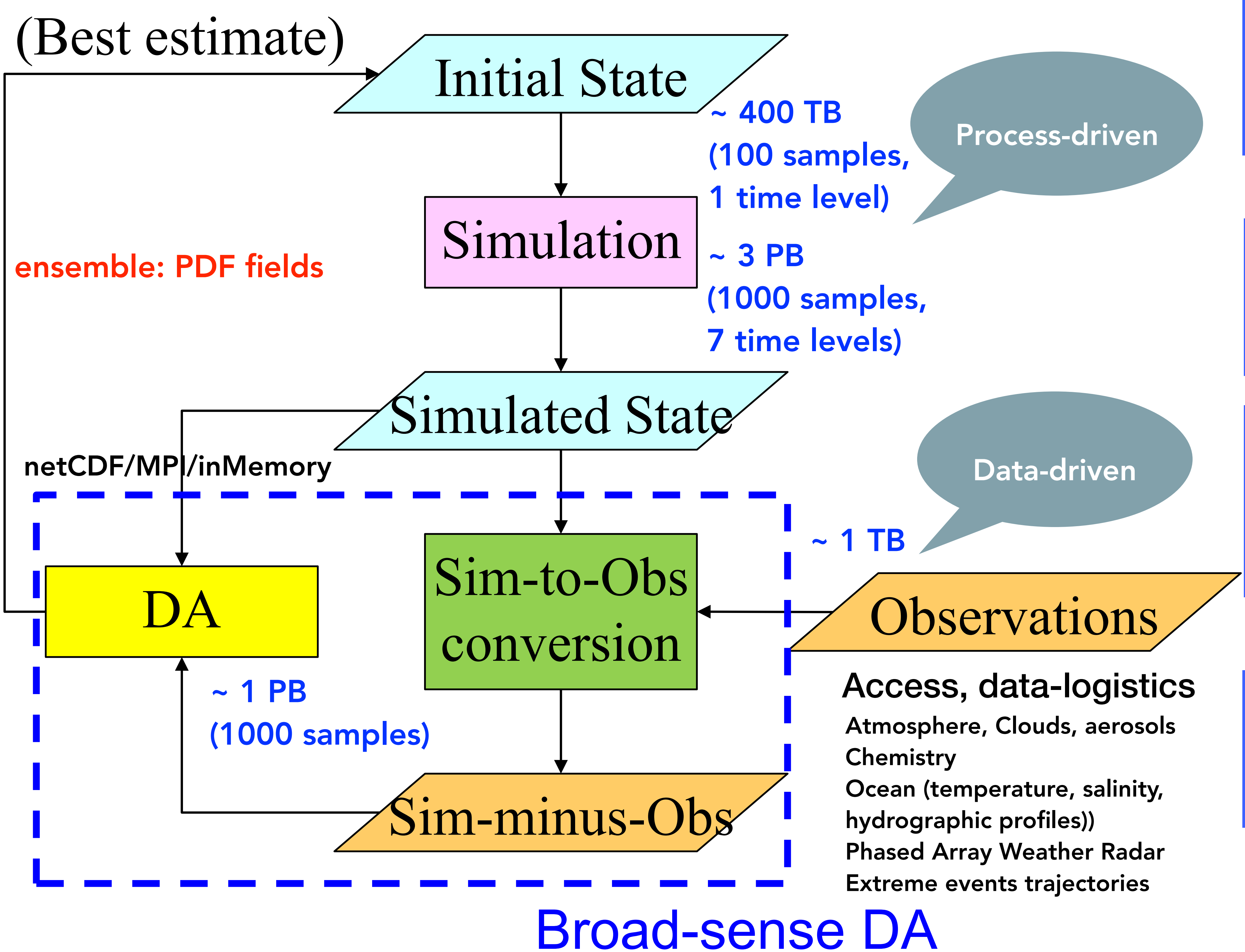


30 minutes forecasting

T. Miyoshi, Riken aics

- **Data assimilation is equivalent to a machine learning problem (Abarbanel et al (2018), Bocquet et al (2018)**
- **Artificial Intelligence: a natural framework to take up challenges of Earth Observation and Modelling**

ML accelerated workflow, data logistics



Accelerating through machine learning

Model statistics (in-situ processing)

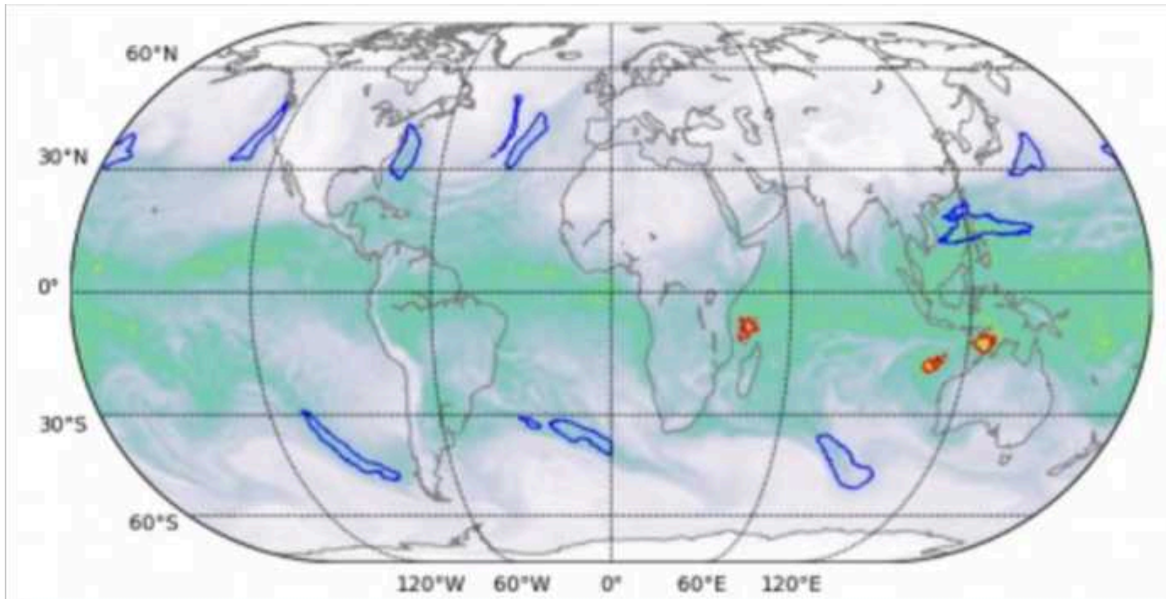
Machine learning pattern recognition smoothing (Kernel-based)

Machine learning optimisation

adapted from Miyoshi et al

Machine learning - Data driven Earth Science

Analytical task	Scientific task	Conventional approaches	Limitations of conventional approaches	Emergent or potential approaches
Classification and anomaly detection				
	Finding extreme weather patterns	Multivariate, threshold-based detection	Heuristic approach, ad hoc criteria used	Supervised and semi-supervised convolutional neural networks[41,42]
	Land-use and change detection	Pixel-by-pixel spectral classification	Shallow spatial context used, or none	Convolutional neural networks[43]
Regression				
	Predict fluxes from atmospheric conditions	Random forests, kernel methods, feedforward neural networks	Memory and lag effects not considered	Recurrent neural networks, long-short-term-memories (LSTMs)[89,99,100]
	Predict vegetation properties from atmospheric conditions	Semi-empirical algorithms (temperature sums, water deficits)	Prescriptive in terms of functional forms and dynamic assumptions	Recurrent neural networks[90], possibly with spatial context
	Predict river runoff in ungauged catchments	Process models or statistical models with hand-designed topographic features[91]	Consideration of spatial context limited to hand-designed features	Combination of convolutional neural network with recurrent networks
State prediction				
	Precipitation nowcasting	Physical modelling with data assimilation	Computational limits due to resolution, data used only to update states	Convolutional–LSTM nets short-range spatial context[92]
	Downscaling and bias-correcting forecasts	Dynamic modelling and statistical approaches	Computational limits, subjective feature selection	Convolutional nets[72], conditional generative adversarial networks (cGANs)[53,93,101]
	Seasonal forecasts	Physical modelling with initial conditions from data	Fully dependent on physical model, current skill relatively weak	Convolutional–LSTM nets with long-range spatial context
	Transport modelling	Physical modelling of transport	Fully dependent on physical model, computational limits	Hybrid physical–convolutional network models[68,94]

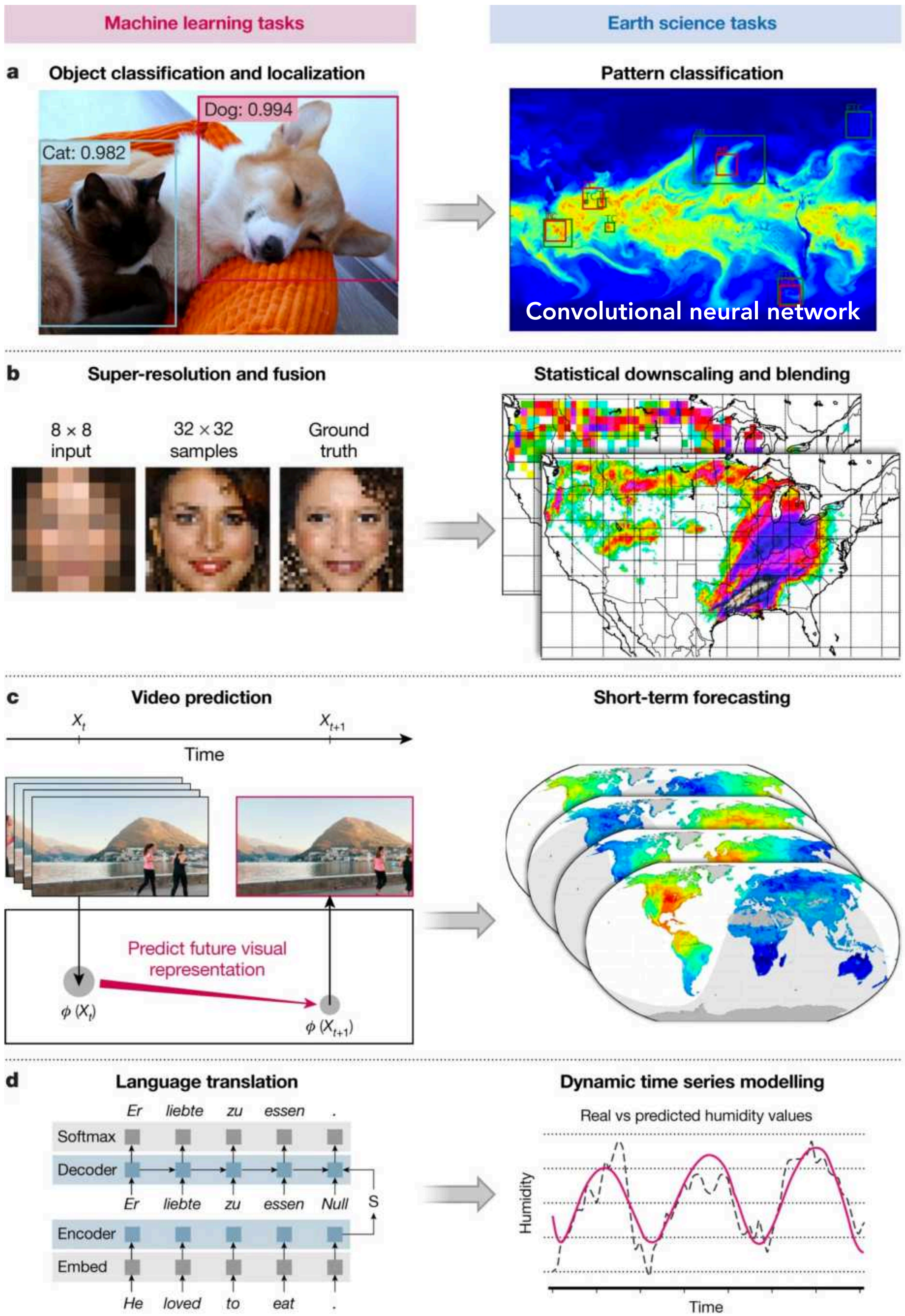


High-quality segmentation results produced by deep learning on climate datasets.

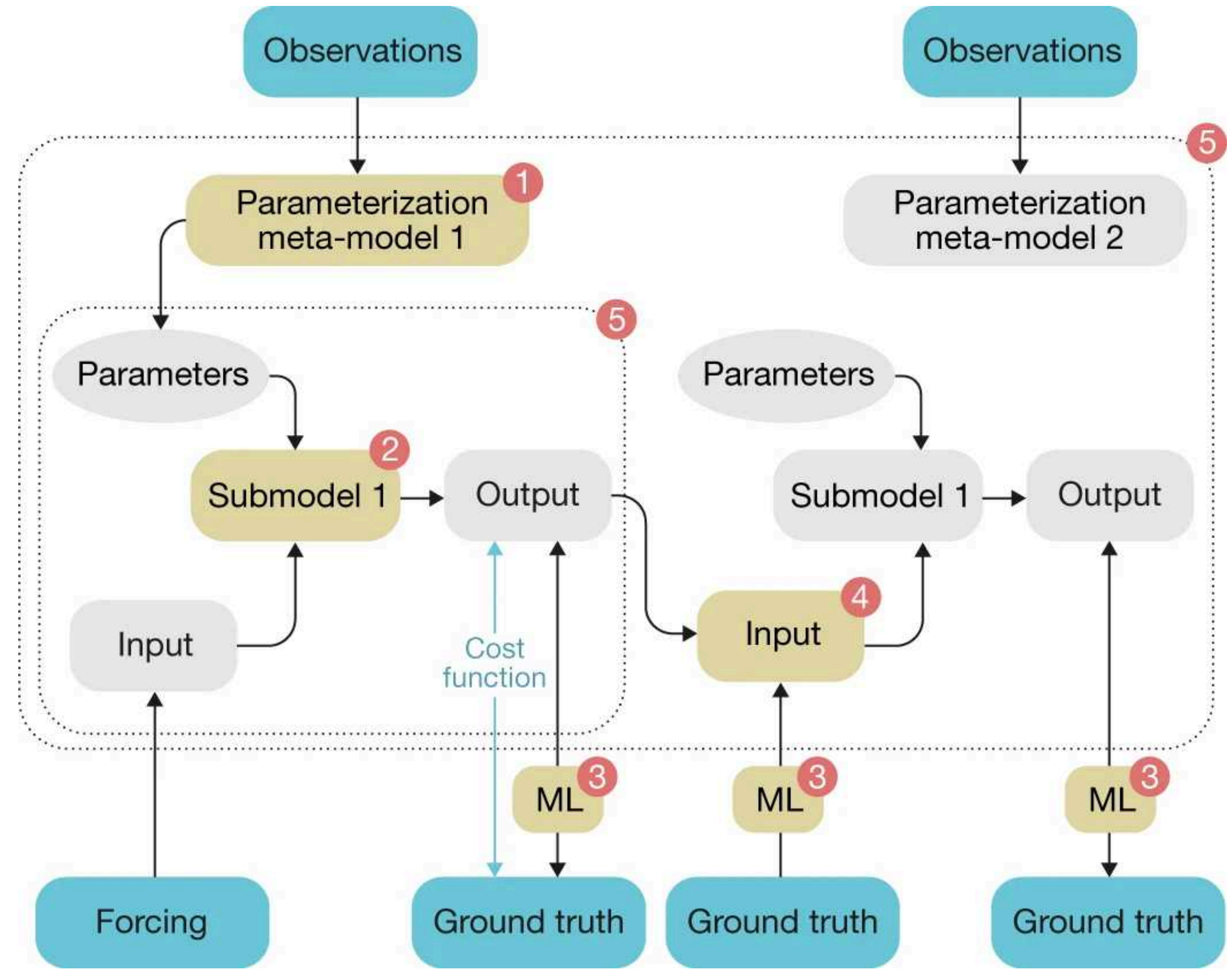
Deep-Learning Methods to Understand Weather Patterns (LBL), 2018 Gordon Bell Prize (https://bit.ly/2X42Vur)

Reichstein et al, 2019

ML & physical modelling



Reichstein et al, 2019



1. Improving parameterisations (global atmospheric modelling)
2. Physical sub-models -> ML models
3. Analysis Model-Observation mismatch
4. Constraining sub-models (from ML)
5. Surrogate modelling or emulations (ML emulators)

- **Interpretability, Physical consistency**
- **Data complexity, uncertainty and noise**
- **Limited available labelled data sets**
- **Extrapolation versus prediction**
- **Computational cost & time: transfer learning**

ML classification of volcanic deformation: InSAR data

- **Earth Observation (routinely)**
  - Volcanoes in remote regions
- **InSAR satellite remote sensing**
  - High-resolution deformation signal
  - Large geographic areas, large coverage
  - Strong statistical link to eruption
- **Increasingly large data sets**
  - Sentinel-1 A and B with 6-day repeat cycle
  - More than 10-TB/day, 2 PB (2014-2017)
  - Challenge manual inspection
  - Timely dissemination of information
- **ML & satellite-based volcano geodesy**

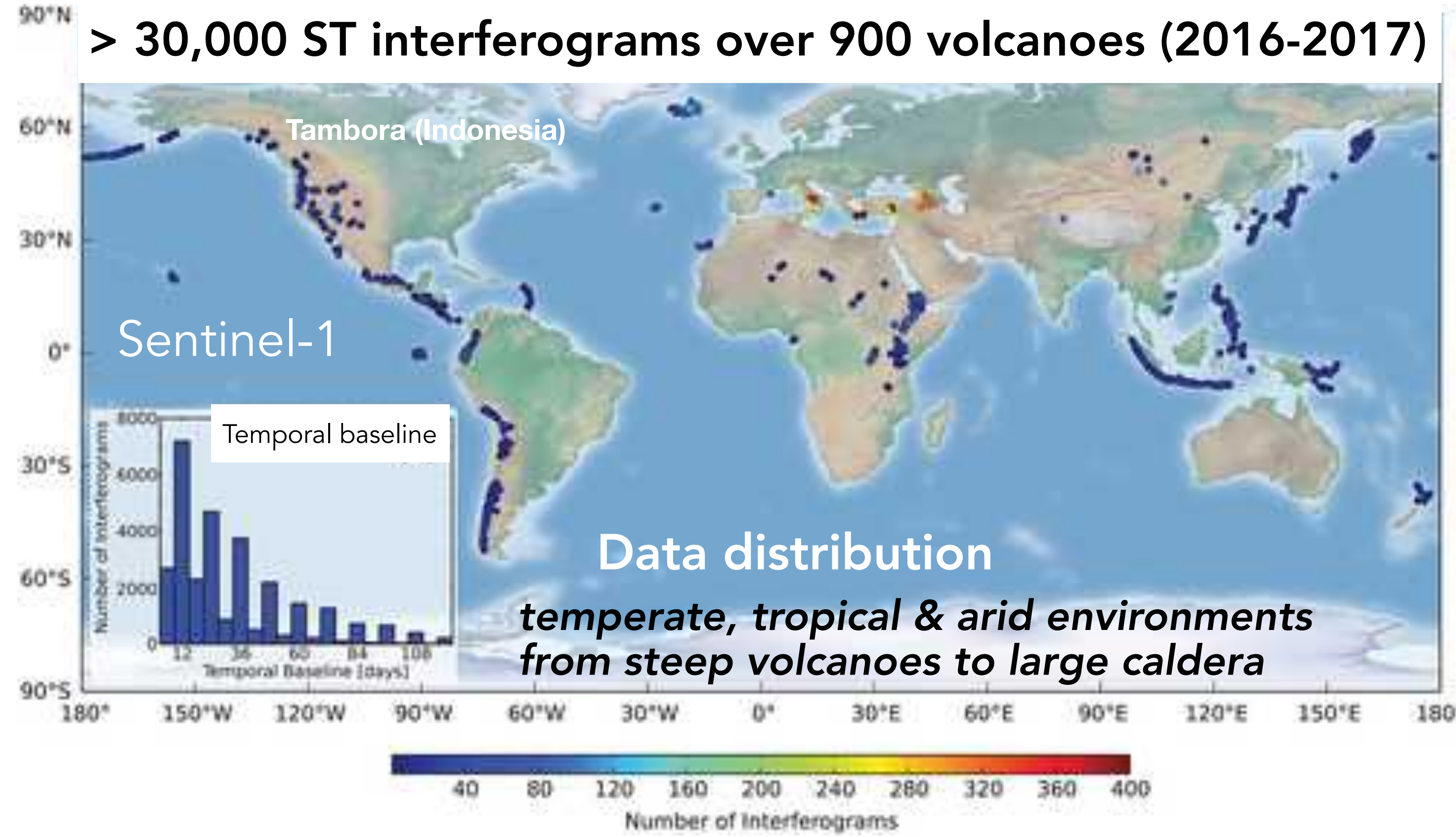
  Automatic detection of deformation patterns associated to volcanic activity

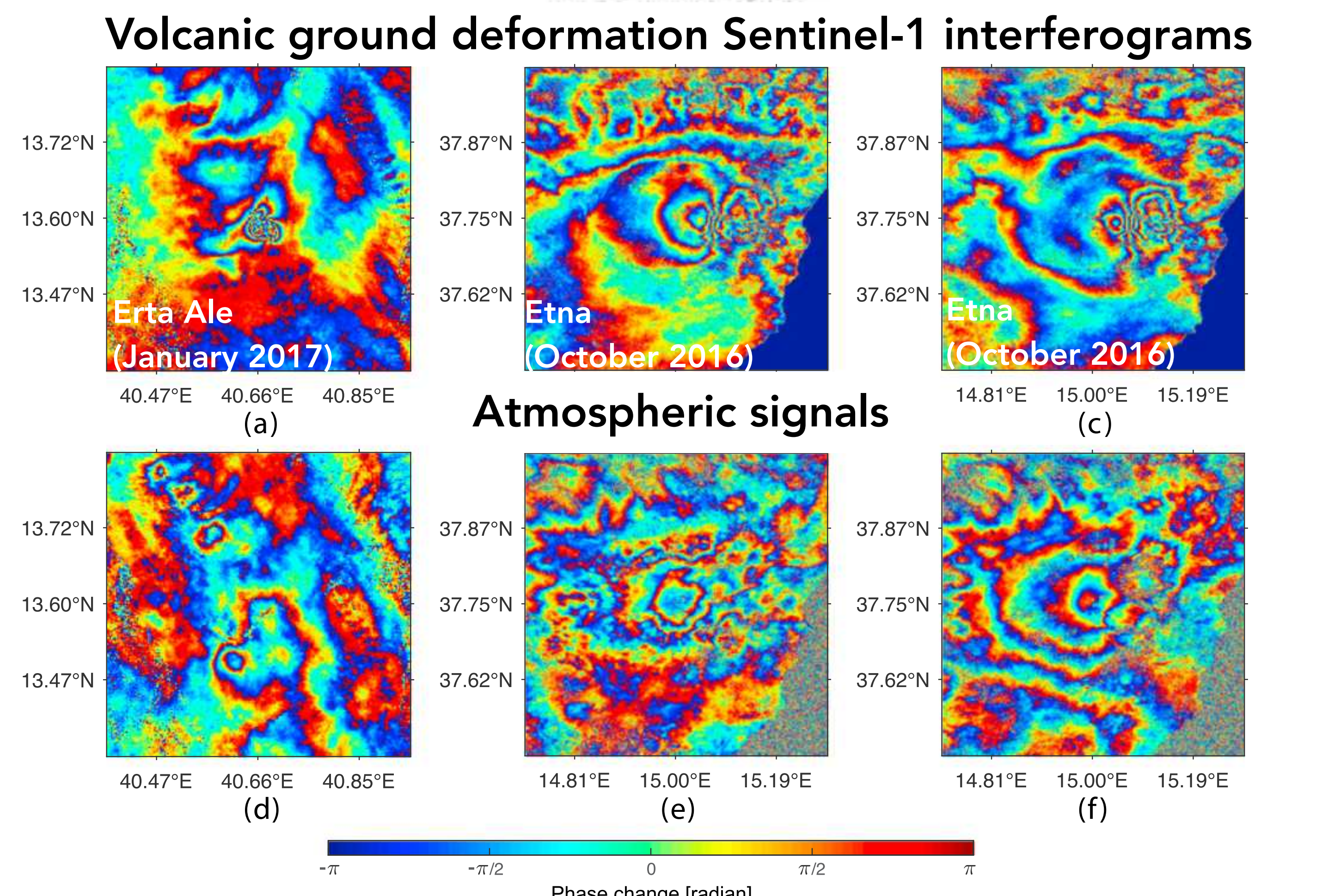
  Classify interferometric fringes in wrapped interferograms (no atmospheric corrections)

  Transfer learning strategy with pre-trained networks (AlexNet)

  Detection of large, rapid deformation signals in wrapped interferograms

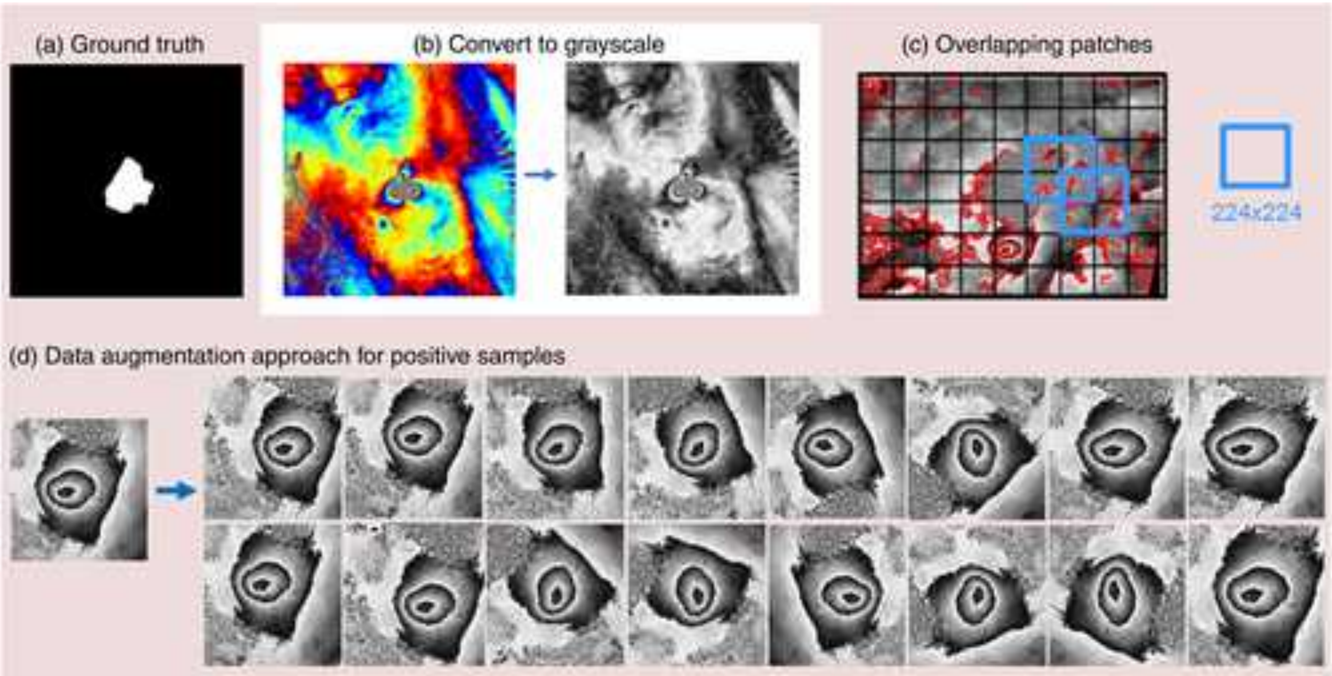
  Further developments
  - ***slow- or small-deformation*** patterns (no multiple fringes in ST interferograms),
  - ***uncertainties quantification***

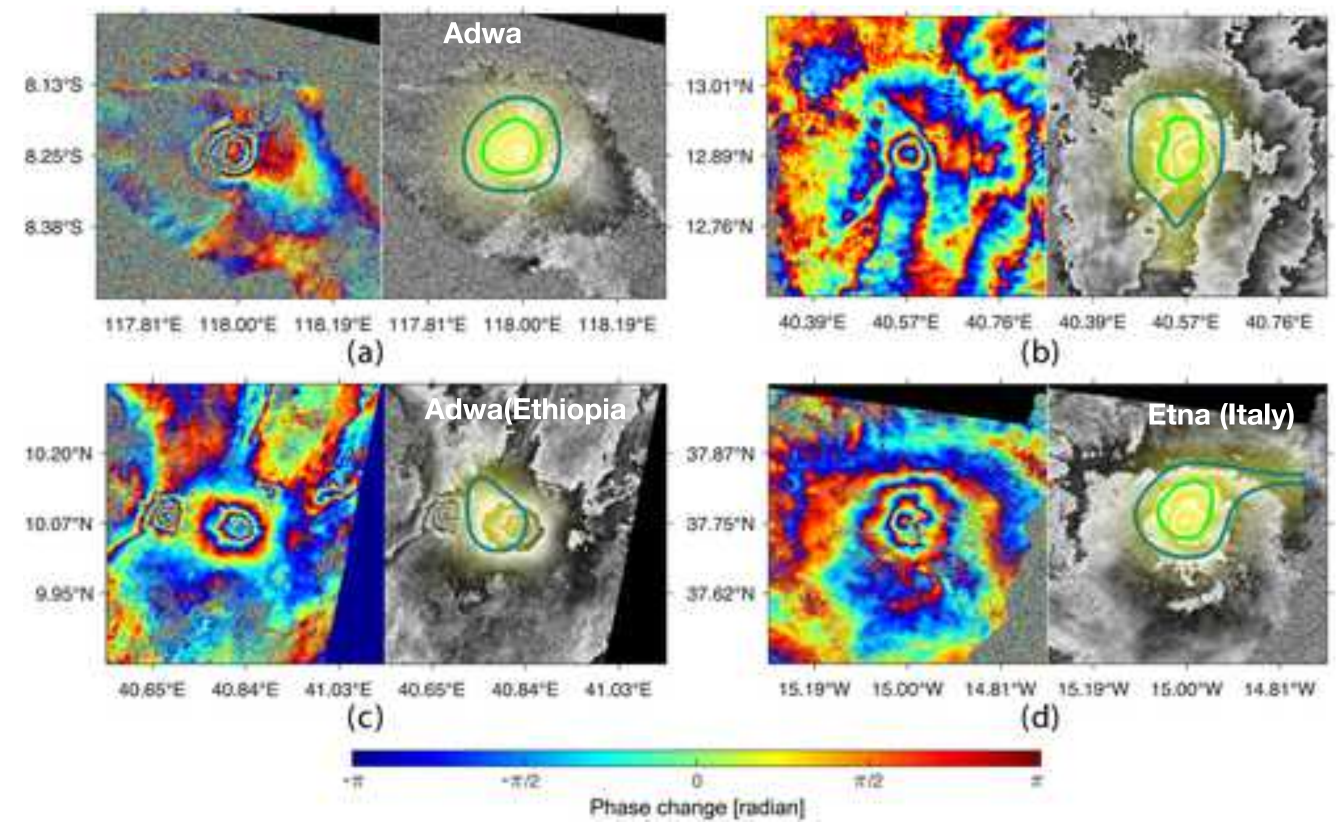




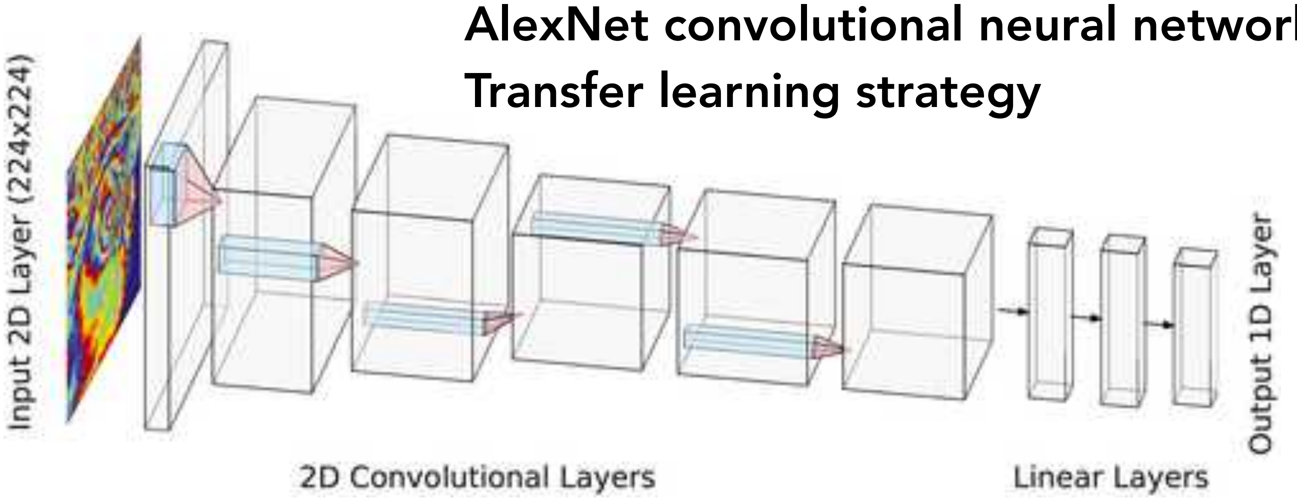
Anantrasirichai et al. (JGR,2018)

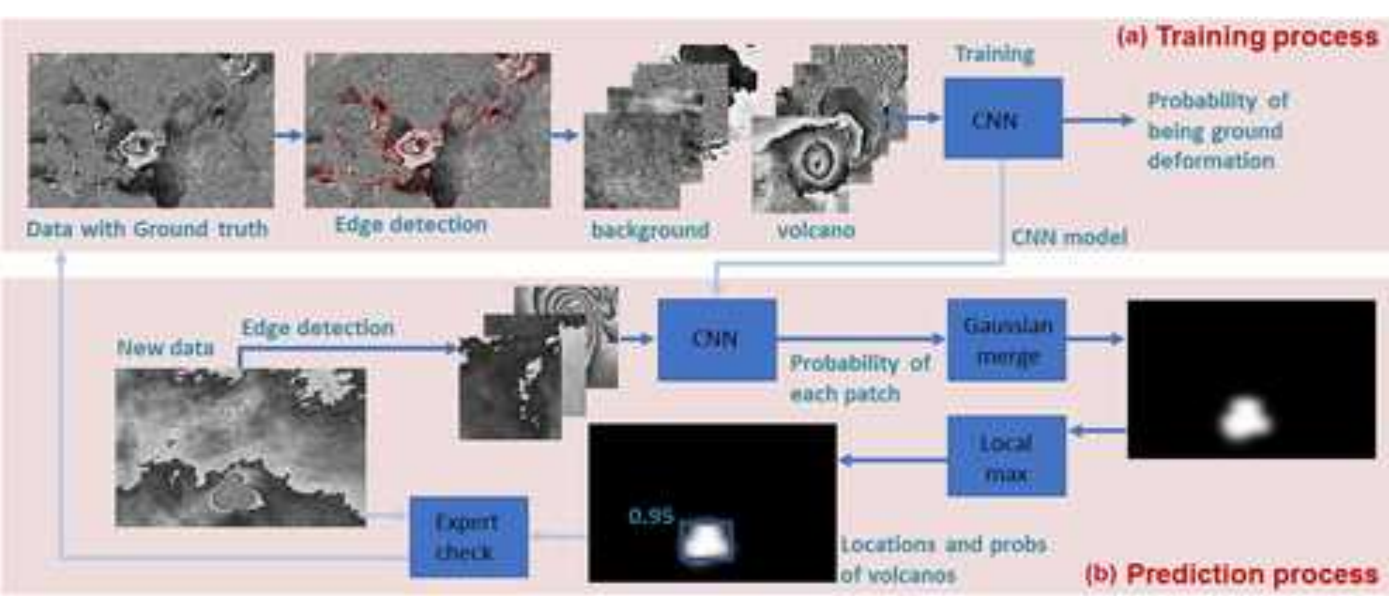
ML classification and detection: SAR data





AlexNet convolutional neural network
Transfer learning strategy





- **Training:** small archives (Envisat) and data sets (Sentinel-1)
  - Wrapped interferogram -> grayscale
  - Training images divided in overlapping patches
  - Edge detection (canny operator: Gaussian filter + double thresholding)
- **Data augmentation** (rotations, flips, distortion, pixels shift): increase number of positive patches
- **Transfer learning strategy**: fine-tuning of pre-trained CNN networks (AlexNet)
- **Retraining strategy**: unconfirmed positive and negative results

Anantrasirichai et al. (2018)

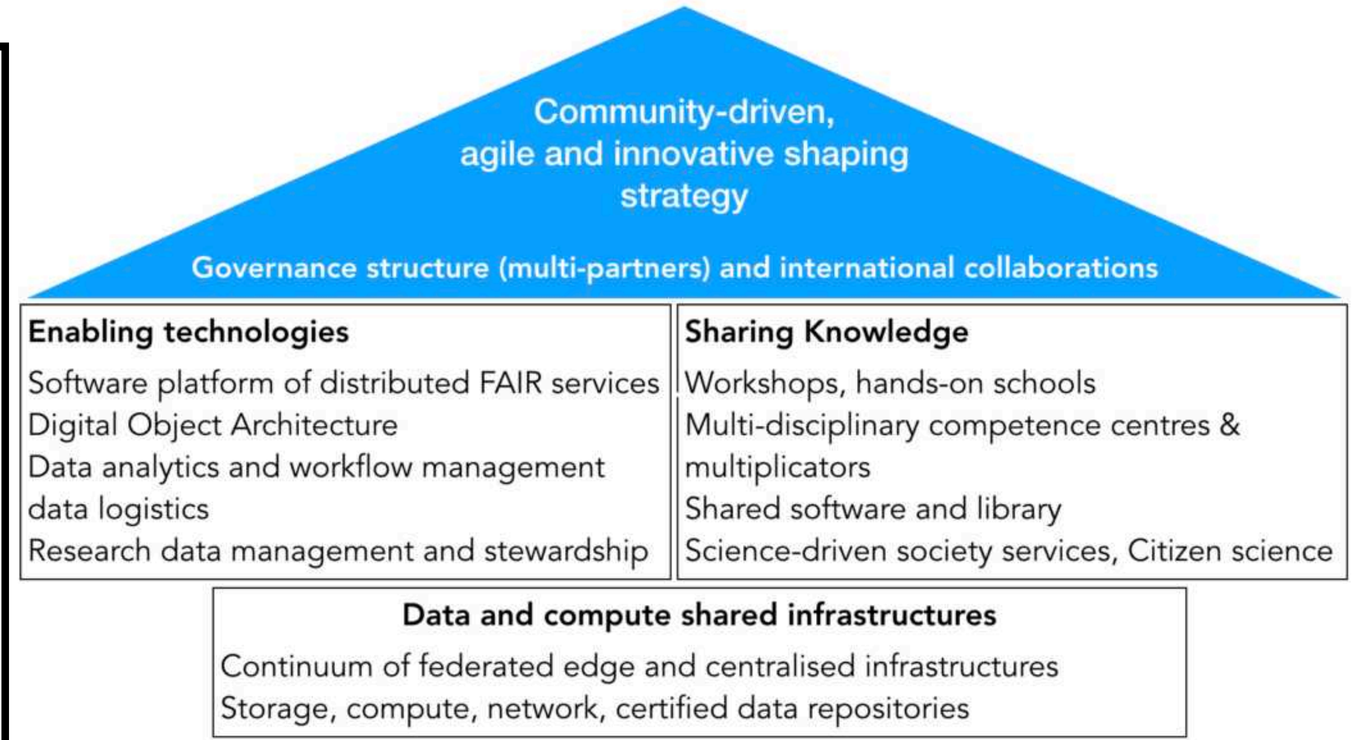
A Digital Object Architecture with a spanning Layer

Software Platform of services

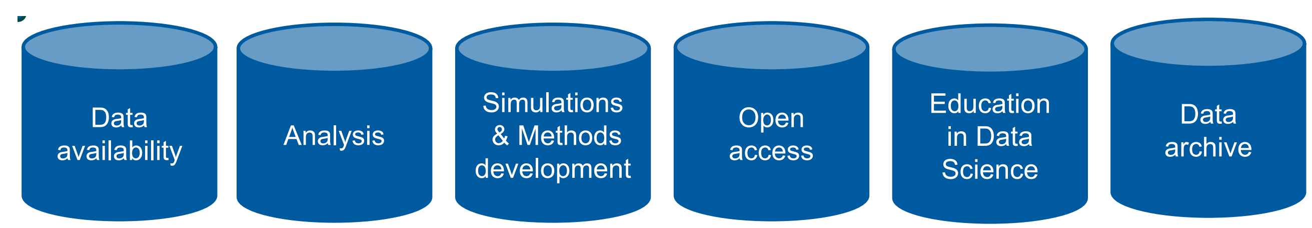
- across edge and centralised computing environments (HPC, Cloud), and Science Data Centres
  - ✳ Persistent/transient storage (variable data life cycles)
  - ✳ Batch and streaming execution models
  - ✳ Containers technology (Kubernetes, Singularity)
  - ✳ Big Data environment
  - ✳ Data logistics across streaming workflow stages
- Data logistics and data reduction across these infrastructures
- Flexible services (storage, compute, communications)
- Rendering services (visualise, analyse)

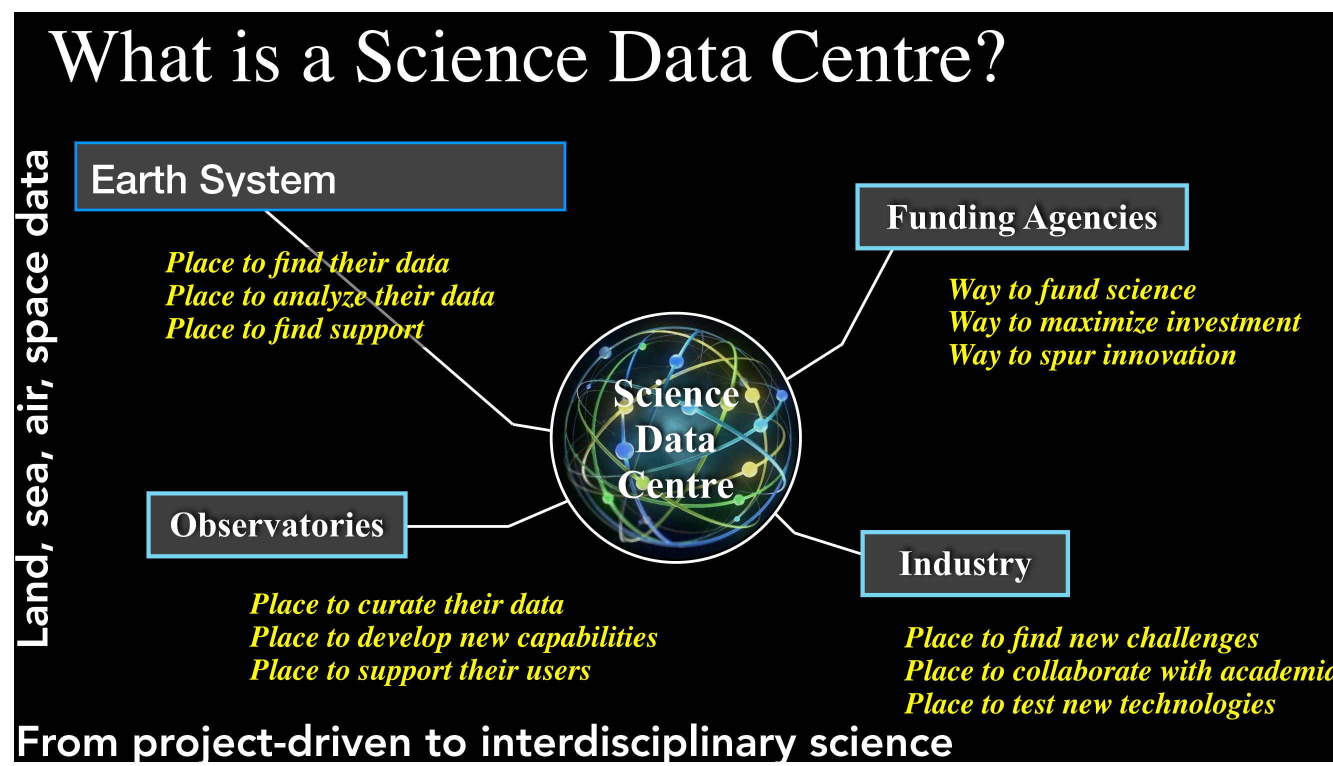
Centralised Environments (HPC, Cloud)

- Concentrate high-performance and resource capabilities (storage, compute, communication)
- Multiple research communities
- Convergence between HPC and HDA
  - ✳Data reduction (in-situ) a fundamental pattern
  - ✳Interoperable execution models (batch, streaming)
  - ✳Integrate different programming models and data formats
  - ✳Federated software stack with provenance systems
  - ✳HPC/HDA workflows including machine learning
  - ✳Leveraged HPC libraries for HDA
- Collaborative, flexible and resilient environments

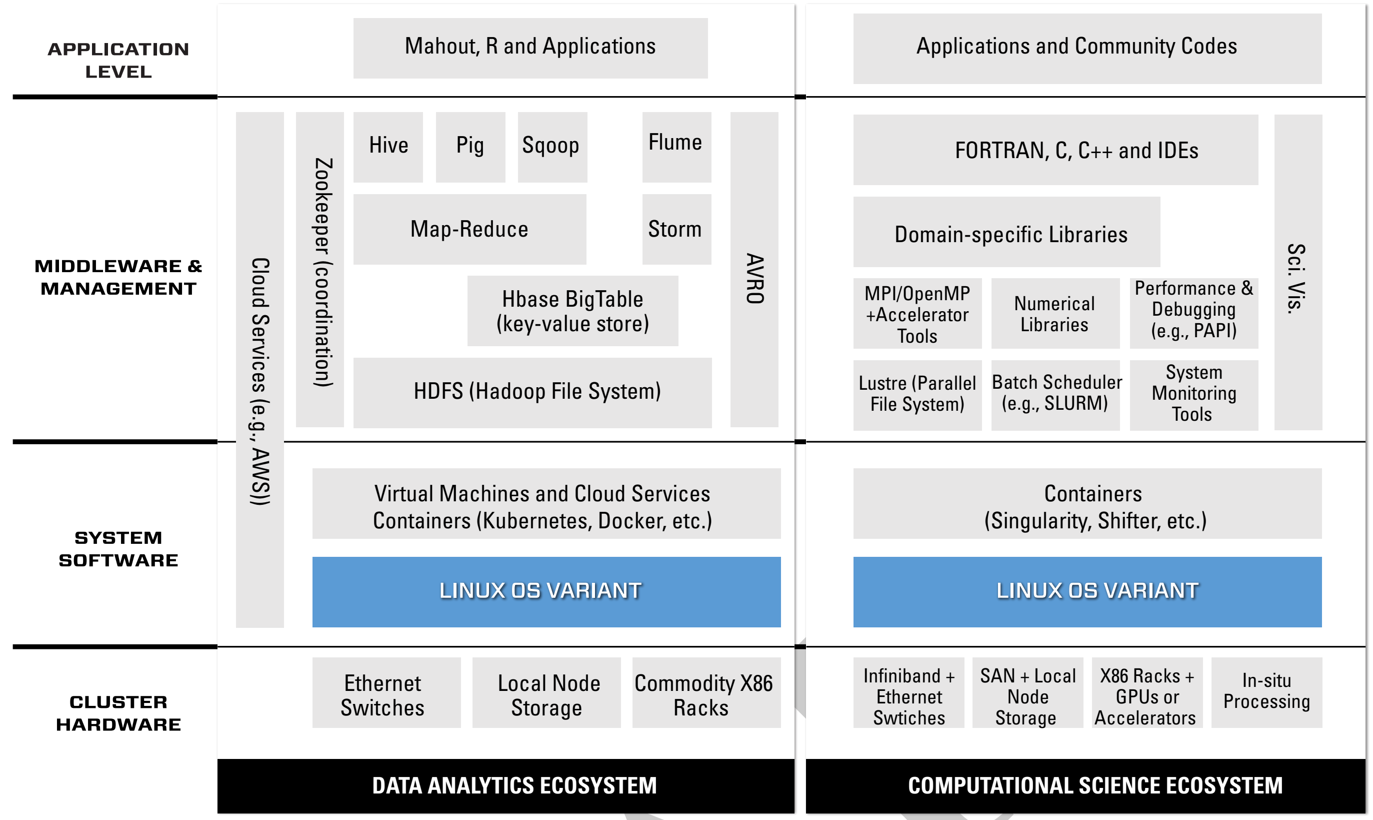


Digital Object Architecture and software services



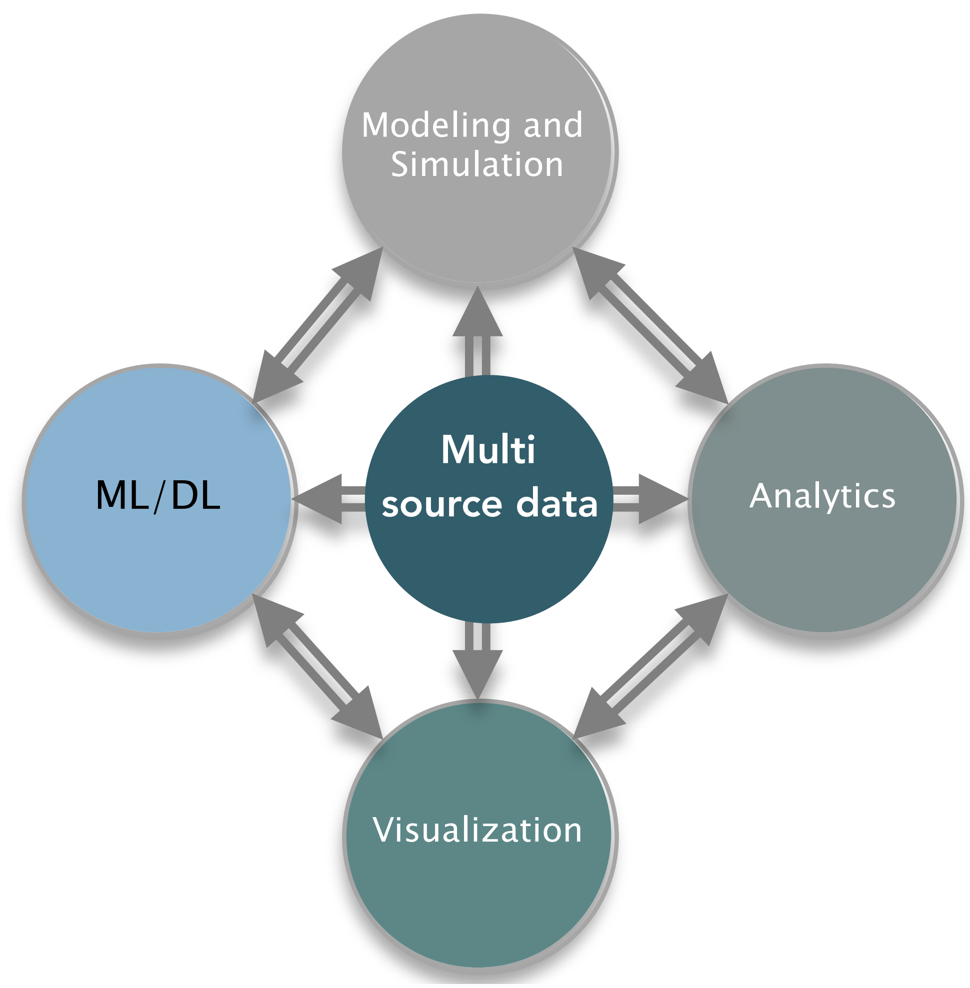


Some Challenges ahead



AI/ML in Earth and Universe Sciences

- Interpretability, adaptability, physical consistency
- Multi-source uncertainty: complex noisy data
- **AI for HPC**: multi scale & multi-physics ensemble simulations, probabilistic inference
- **HPC for HPDA/AI**: multi-wavelength, multi-source data, transfer learning limitations
- **Increasing ML/DL use**: interdisciplinary collaboration & mutualised expertise
- **FAIR software services** and support



HPC and HDA convergence

- Access policy (FAAI) & security
- Data logistics (in-coming, in, out-coming)
- Resources management and execution environments
- Persistent/temporary data storage over data lifecycle
- **Digital Object Architectures** (PiDs, meta data, registries, resolution system)
- **Software and library heritage**: evolution and new architecture adaptation