netherlands eScience center

WAGENINGEN UR
For quality of life

Doing FAIR (software)

in Environmental and Life Sciences

Wageningen, 12-12-2018

Acknowledgement : team at Netherlands eScience Center

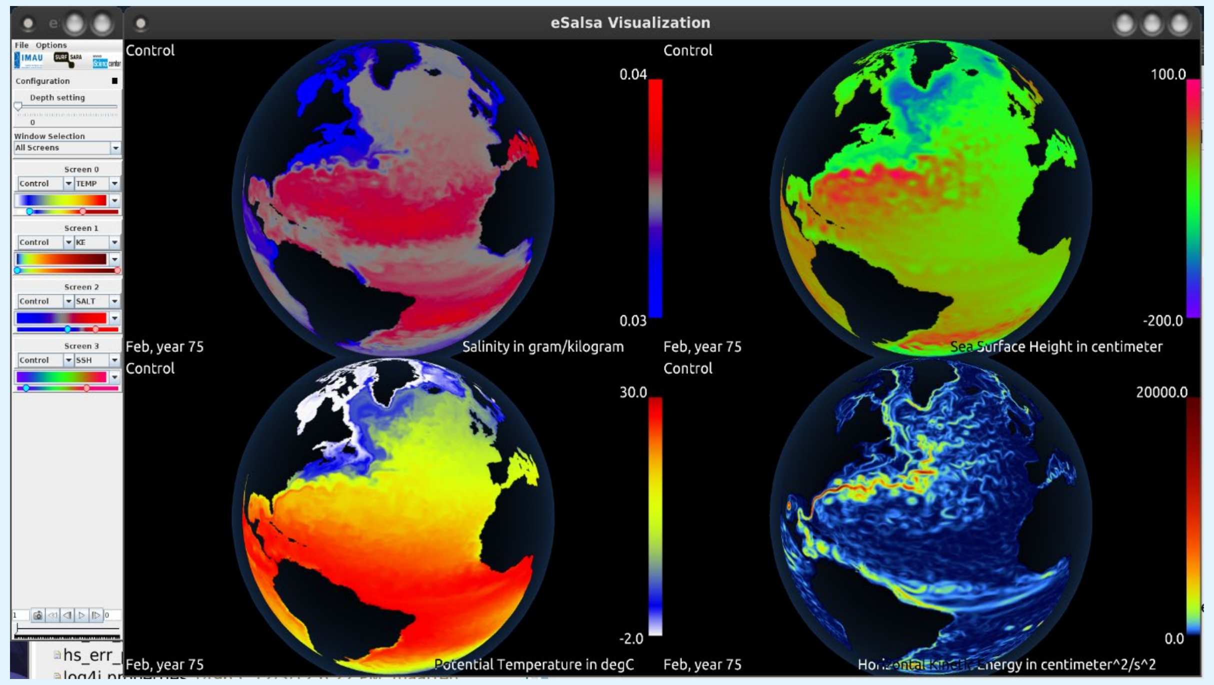
100+ projects

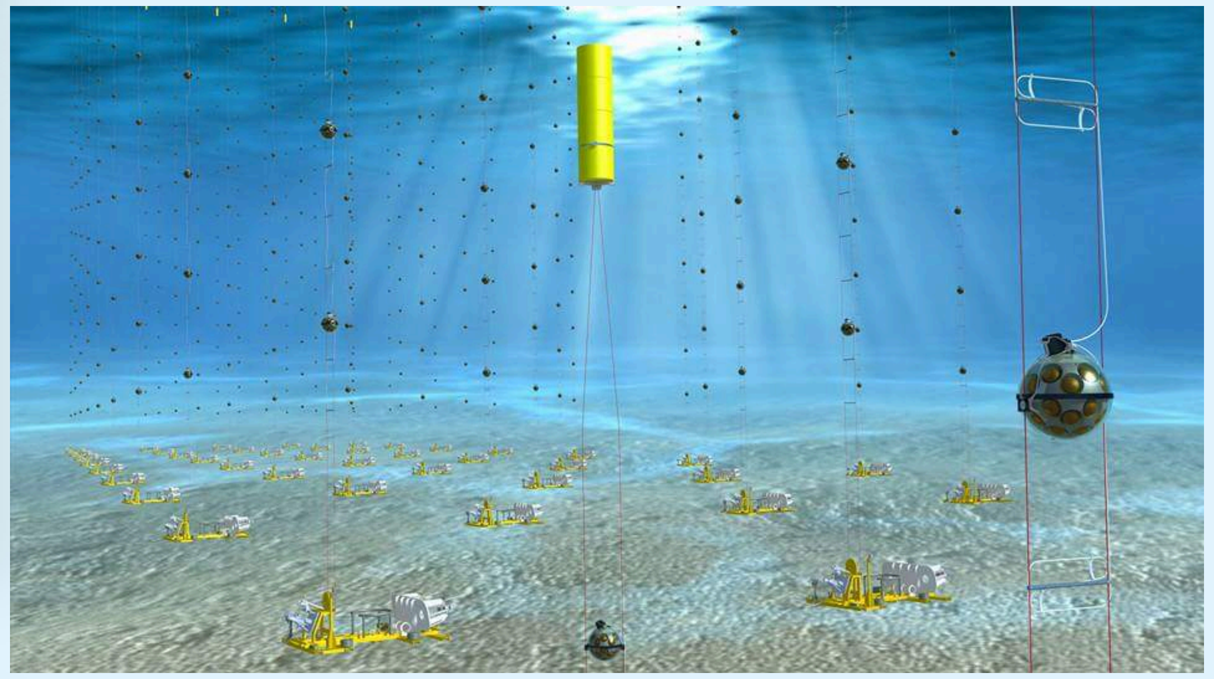
Humanities & Social Sciences

incl. SMART cities, text analysis, creative technologies

Physics & Beyond

incl. astronomy, high-energy physics, advanced materials



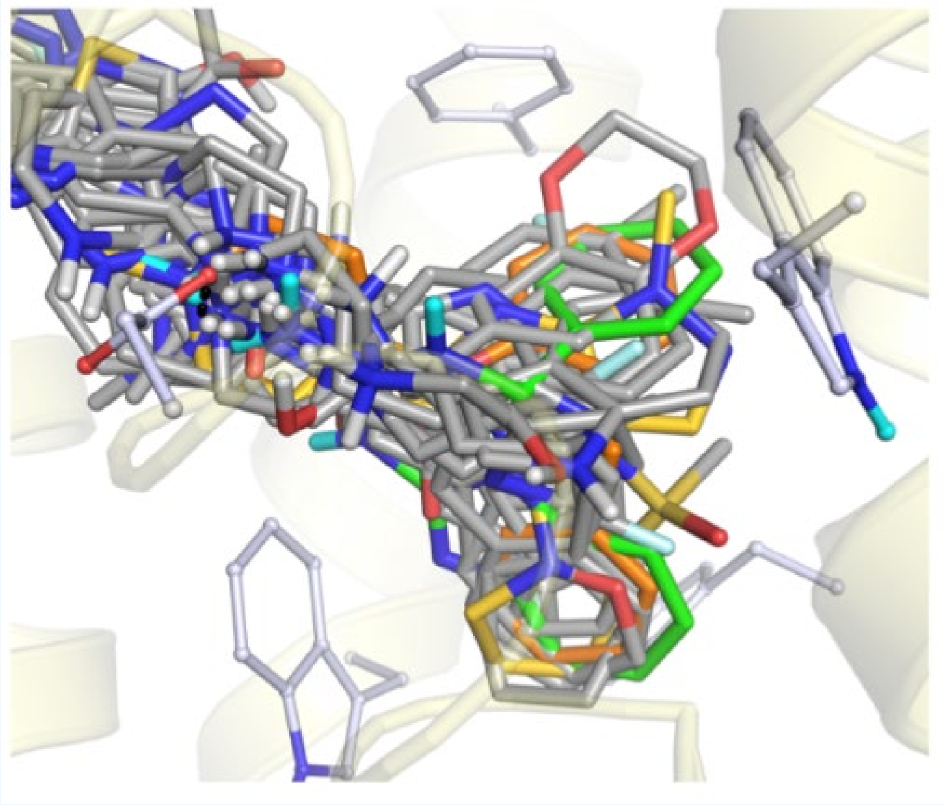


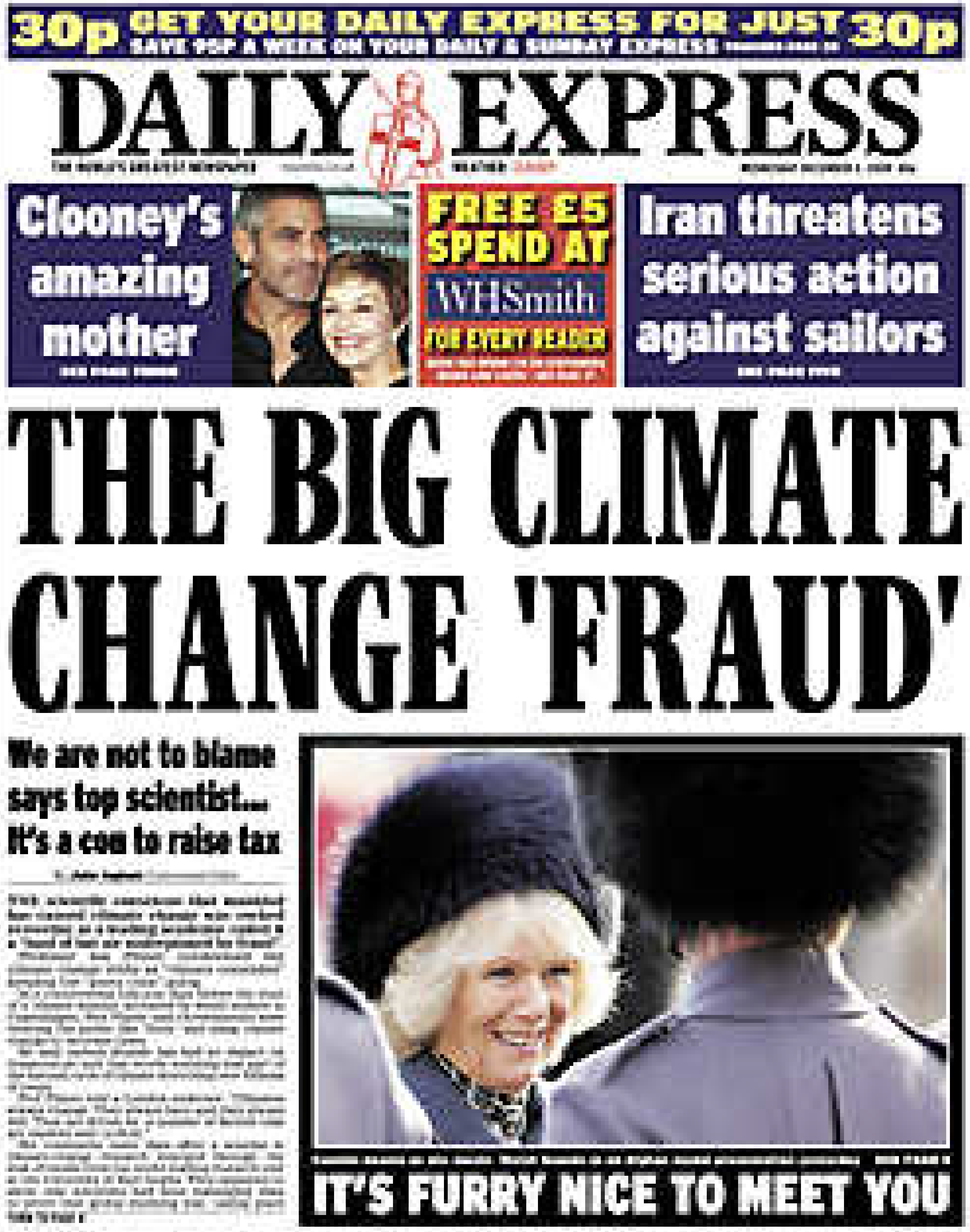
Sustainability & Environment

incl. climate, ecology, energy, logistics, water management

Life Sciences & eHealth

incl. bio-imaging, next generation sequencing, molecules





DAILY EXPRESS
THE WORLD'S GREATEST NEWSPAPER [illegible] WEATHER [illegible] [illegible]

Clooney's amazing mother
[illegible]

Iran threatens serious action against sailors
[illegible]

THE BIG CLIMATE CHANGE 'FRAUD'

We are not to blame says top scientist... It's a con to raise tax

[illegible]

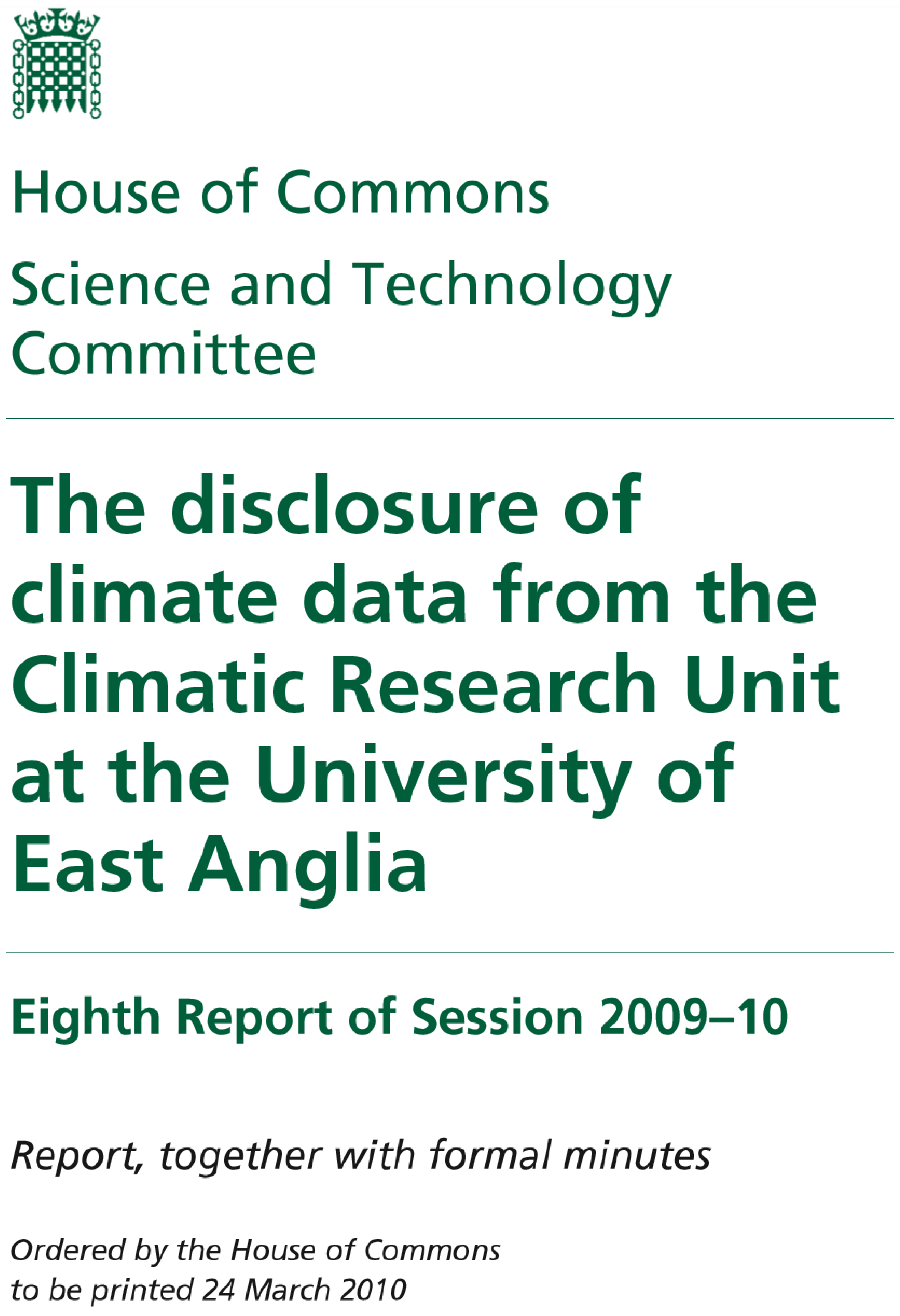
[illegible]

[illegible]

IT'S FURRY NICE TO MEET YOU

- In the context of the sharing of data and methodologies, …. Professor XX’s actions were in line with common practice in the climate science community.

- It is not standard practice in climate science to publish the raw data and the computer code in academic papers. However, climate science is a matter of great importance and the quality of the science should be irreproachable. We therefore consider that climate scientists should take steps to make available all the data that support their work (including raw data) and full methodological workings (including the computer codes).



House of Commons

Science and Technology Committee

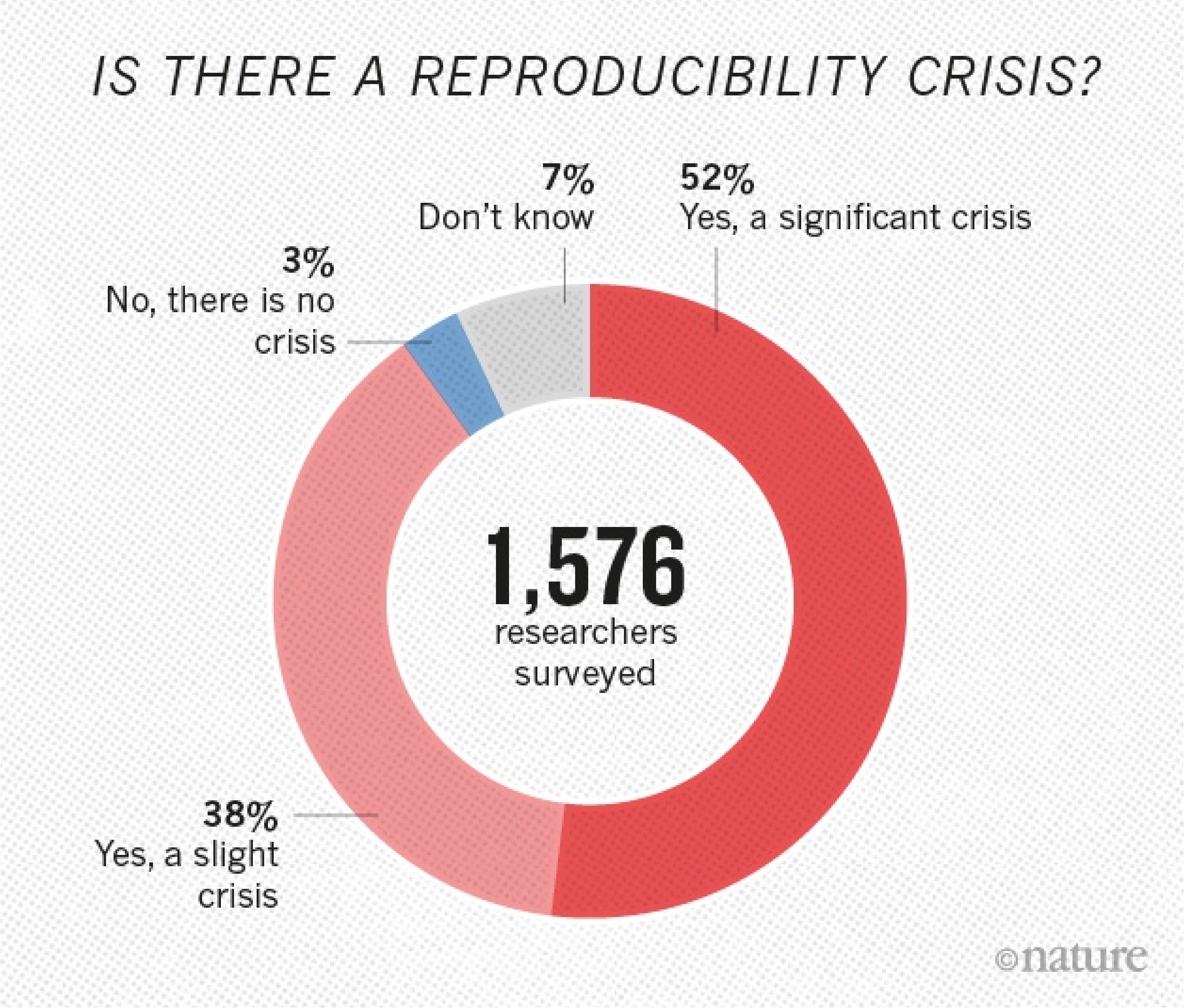
The disclosure of climate data from the Climatic Research Unit at the University of East Anglia

Eighth Report of Session 2009–10

Report, together with formal minutes

Ordered by the House of Commons
to be printed 24 March 2010

Replication and reproducibility crisis



Baker, 2016, Nature

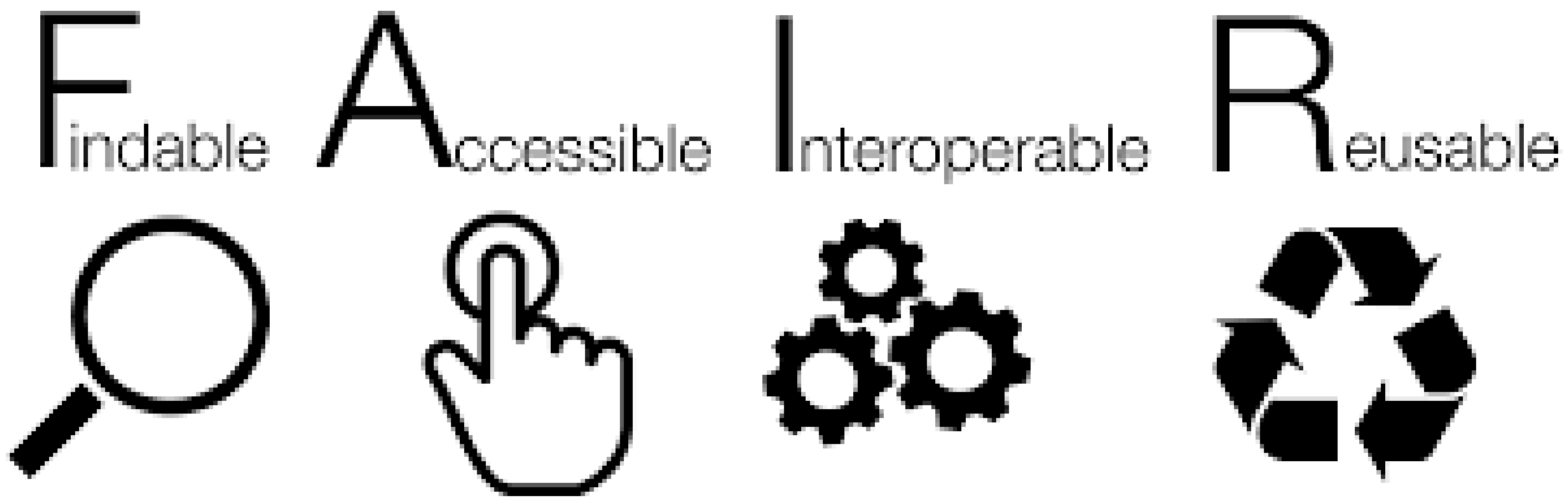
About science and scholarly research

- Increasingly problem -driven on big societally relevant themes
- Increasingly inter -, multi -, trans-disciplinary
- Grand challenges: clean energy, safe societies, healthy societies, etc.

Open Science is about **extending the principles of openness to the whole research cycle** , fostering sharing and collaboration as early as possible thus entailing a systemic change to the way science and research is done

FAIR data principles and FAIR software principles contribute to open science

FAIR software for open science (also…workflows, methodologies)



Findable : sufficiently rich metadata and unique persistent identifier

Accessible : metadata is in machine and human readable format

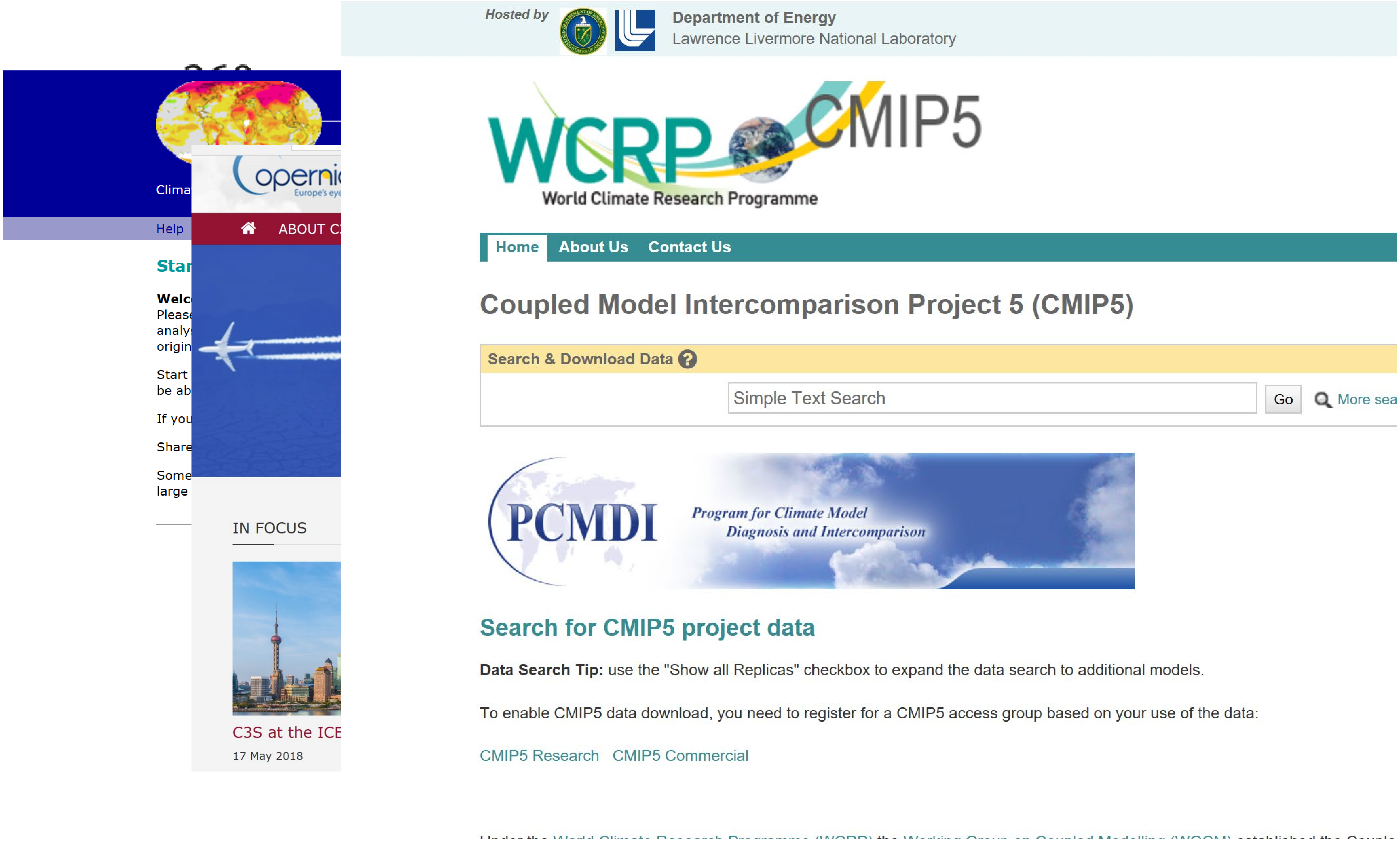
Software and metadata is deposited in trusted community approved repository

Interoperable : uses community accepted standards and platforms, making it possible for users to run the software

Reusable: has clear license and documentation

Examples of FAIR in weather & climate research

Findable open data in climate research



Hosted by Department of Energy
Lawrence Livermore National Laboratory

WCRP
CMIP5
World Climate Research Programme

Home About Us Contact Us

Coupled Model Intercomparison Project 5 (CMIP5)

Search & Download Data ?

Simple Text Search Go More sea

PCMDI
Program for Climate Model
Diagnosis and Intercomparison

Search for CMIP5 project data

Data Search Tip: use the "Show all Replicas" checkbox to expand the data search to additional models.

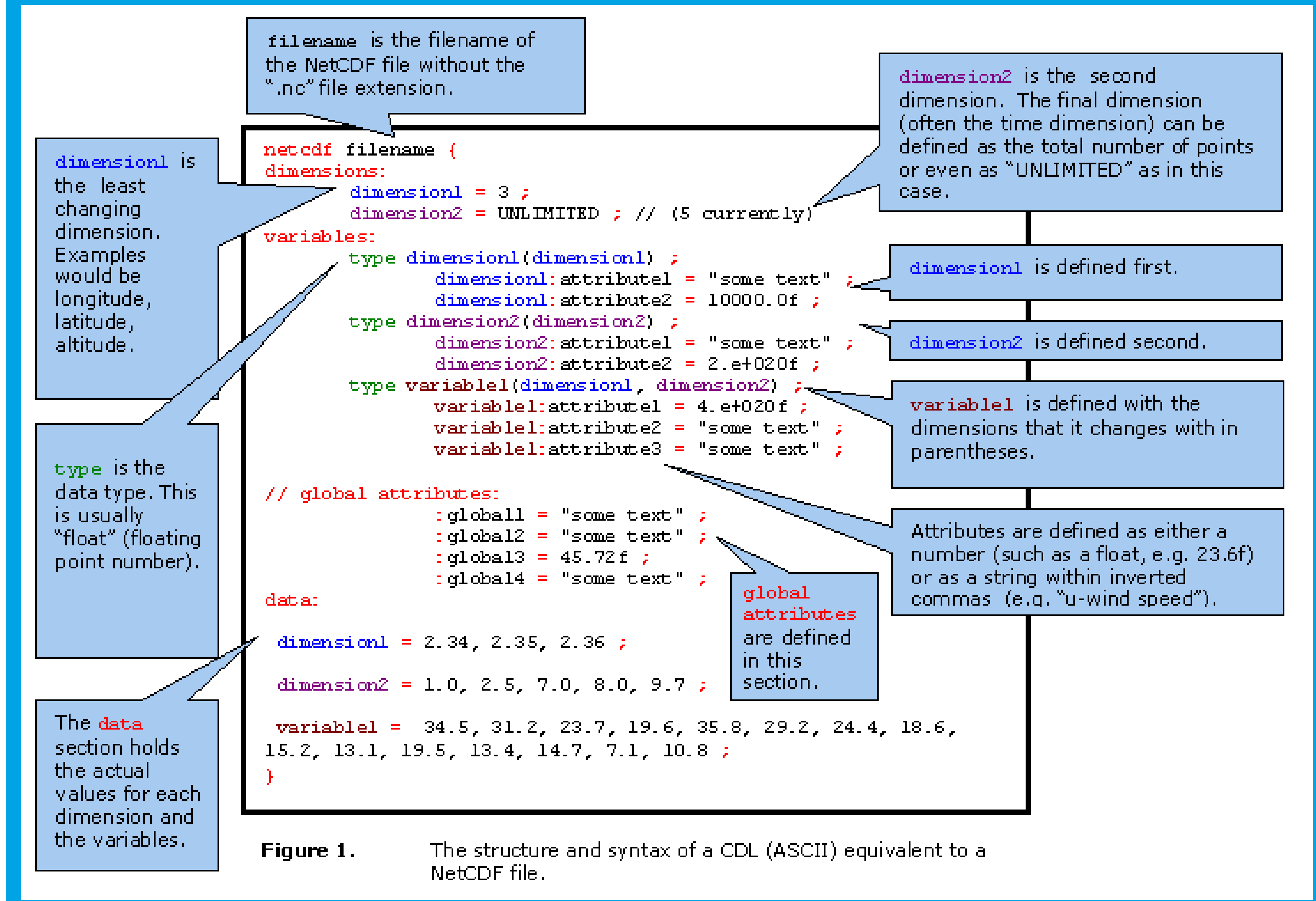
To enable CMIP5 data download, you need to register for a CMIP5 access group based on your use of the data:

CMIP5 Research CMIP5 Commercial

Under the World Climate Research Programme (WCRP) the Working Group on Coupled Modelling (WGCM) established the Couple

Interoperable in climate research

- Netcdf : is a set of software libraries and self-describing, machine -independent data formats that support the creation, access, and sharing of array -oriented scientific data.
- Climate and Forecasting conventions
- CMOR: climate model output writer



filename is the filename of the NetCDF file without the ".nc" file extension.

dimension1 is the least changing dimension. Examples would be longitude, latitude, altitude.

dimension2 is the second dimension. The final dimension (often the time dimension) can be defined as the total number of points or even as "UNLIMITED" as in this case.

```
netcdf filename {
dimensions:
        dimension1 = 3 ;
        dimension2 = UNLIMITED ; // (5 currently)
variables:
        type dimension1(dimension1) ;
                dimension1:attribute1 = "some text" ;
                dimension1:attribute2 = 10000.0f ;
        type dimension2(dimension2) ;
                dimension2:attribute1 = "some text" ;
                dimension2:attribute2 = 2.e+020f ;
        type variable1(dimension1, dimension2) ;
                variable1:attribute1 = 4.e+020f ;
                variable1:attribute2 = "some text" ;
                variable1:attribute3 = "some text" ;

// global attributes:
                :global1 = "some text" ;
                :global2 = "some text" ;
                :global3 = 45.72f ;
                :global4 = "some text" ;
data:

 dimension1 = 2.34, 2.35, 2.36 ;

 dimension2 = 1.0, 2.5, 7.0, 8.0, 9.7 ;

 variable1 =  34.5, 31.2, 23.7, 19.6, 35.8, 29.2, 24.4, 18.6,
15.2, 13.1, 19.5, 13.4, 14.7, 7.1, 10.8 ;
}
```

dimension1 is defined first.

dimension2 is defined second.

variable1 is defined with the dimensions that it changes with in parentheses.

type is the data type. This is usually "float" (floating point number).

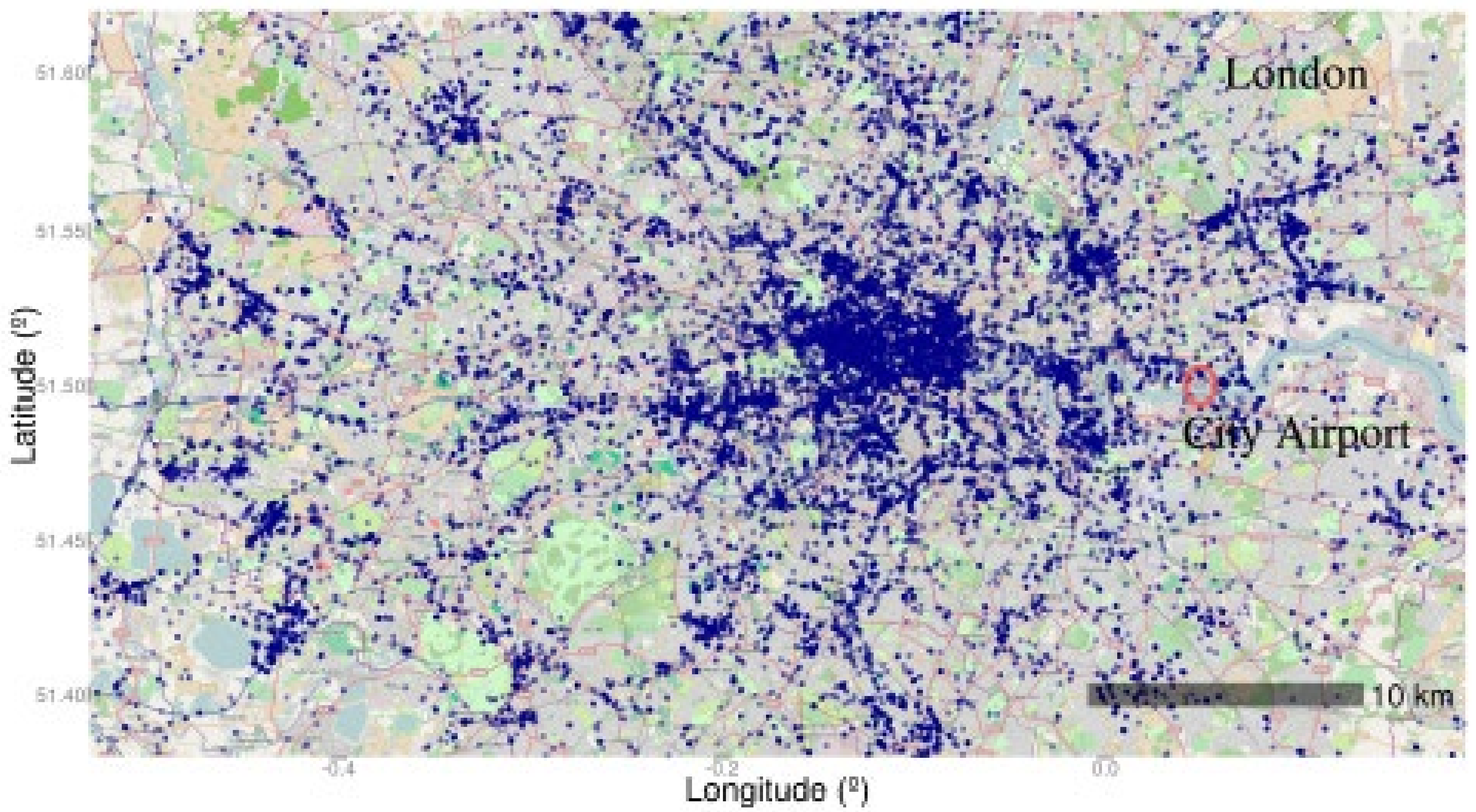
Attributes are defined as either a number (such as a float, e.g. 23.6f) or as a string within inverted commas (e.g. "u-wind speed").

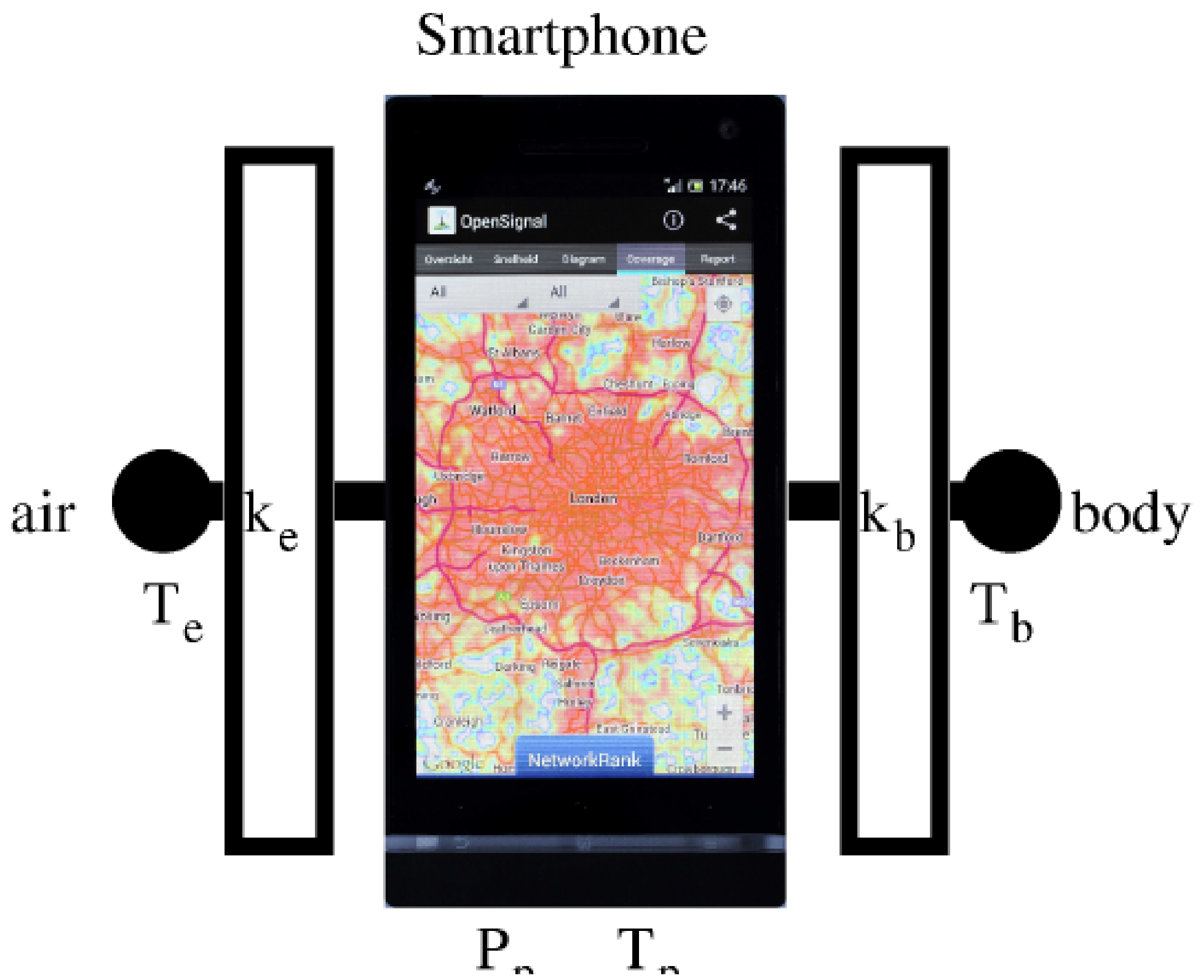
global attributes are defined in this section.

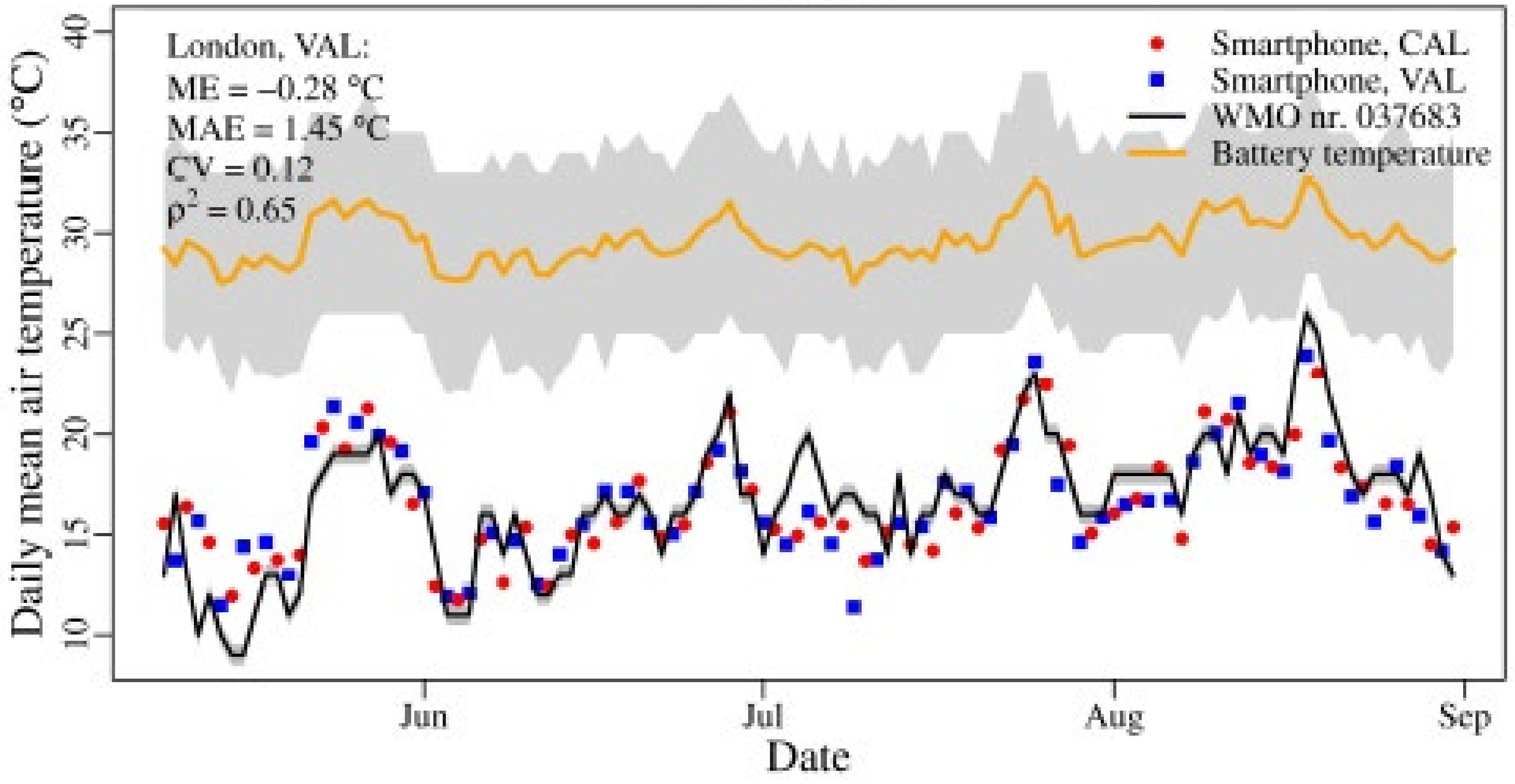
The data section holds the actual values for each dimension and the variables.

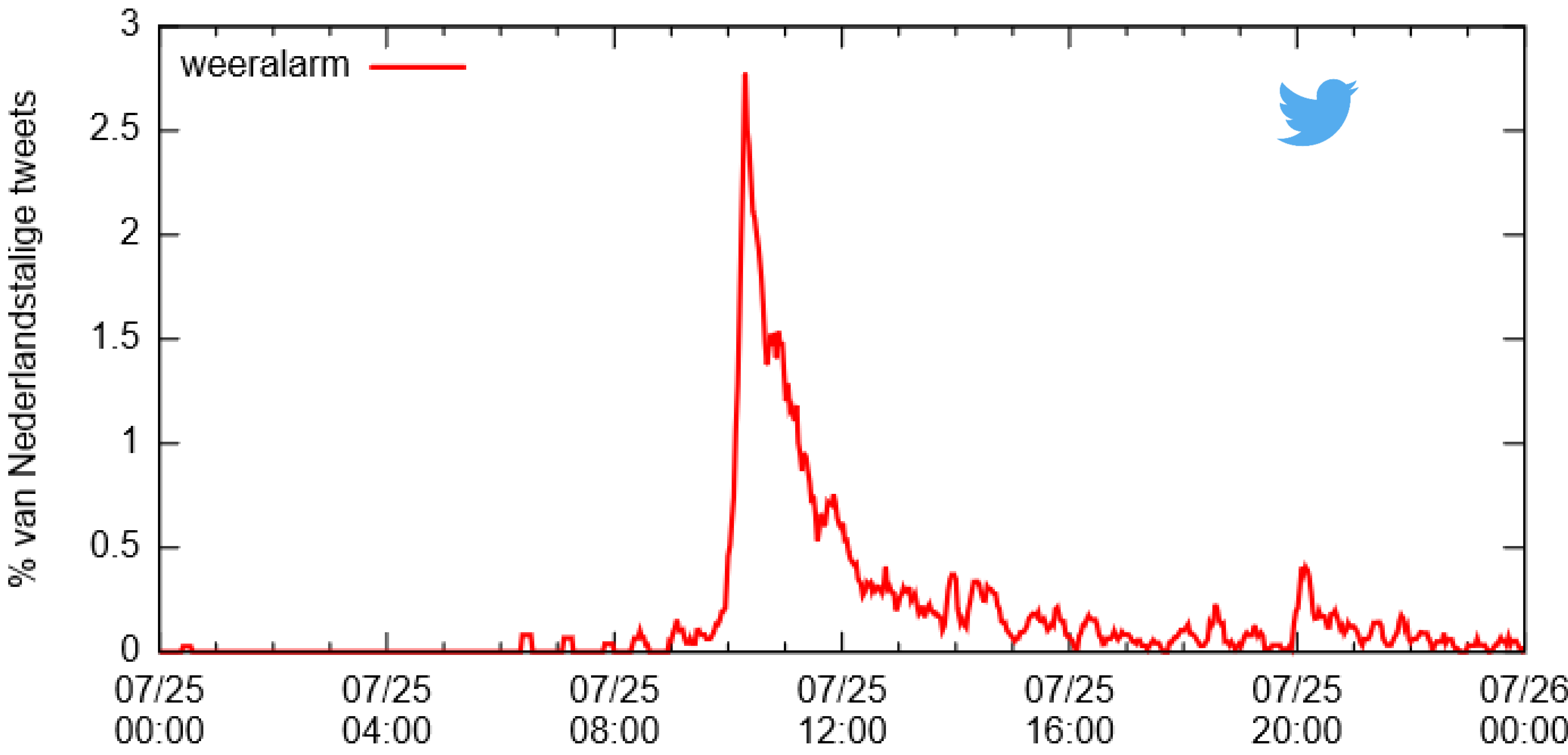
Figure 1. The structure and syntax of a CDL (ASCII) equivalent to a NetCDF file.

Other big data in meteorology









Overeem et al GRL 2013

Downscaling

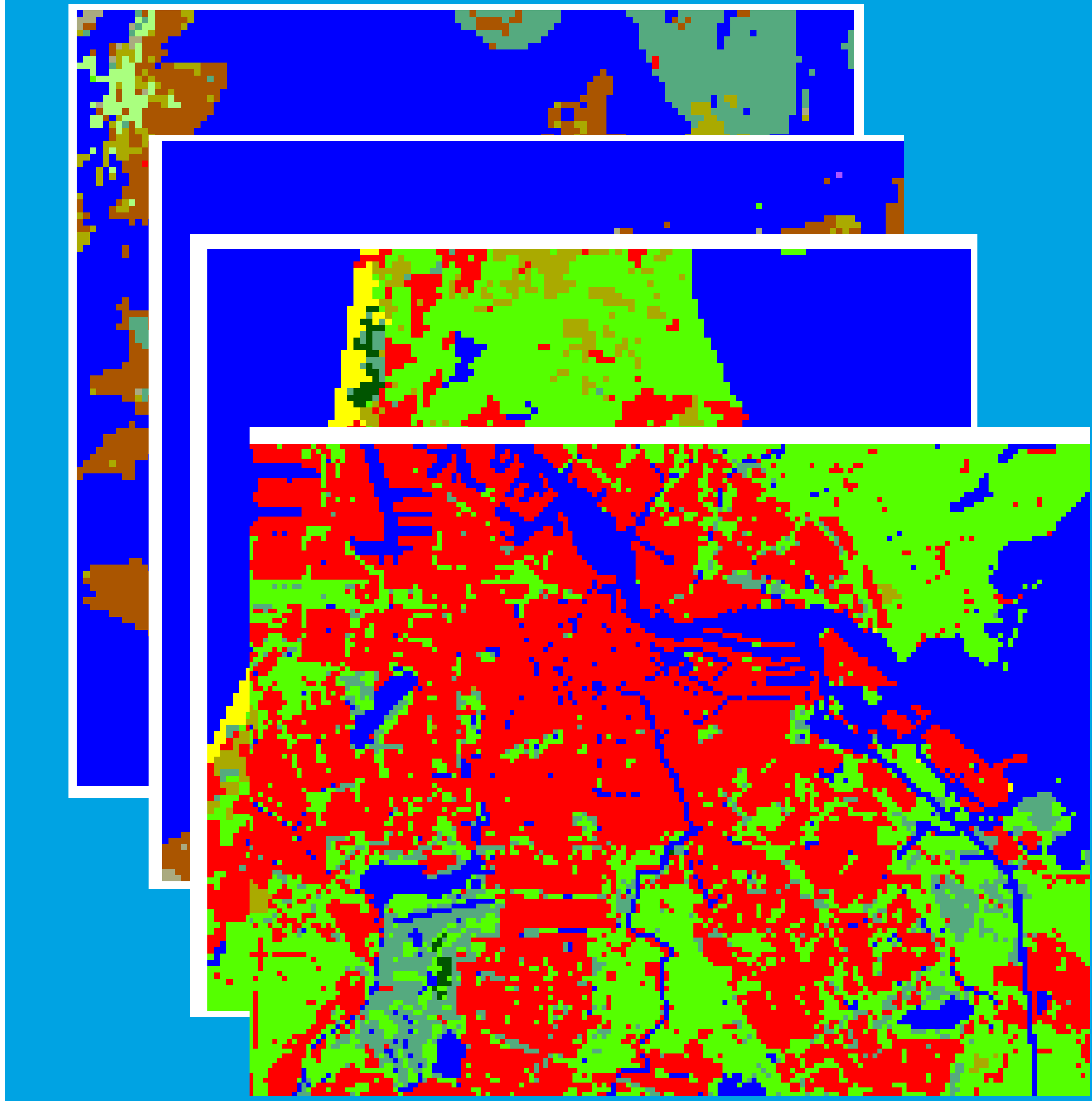
Daily forecasts
WRF3.5 + urban module (SLUCM)
48 hour runs, 24 hour spin-up

Domain 1: 12.5km
default setup

Domain 2: 2.5km
default setup

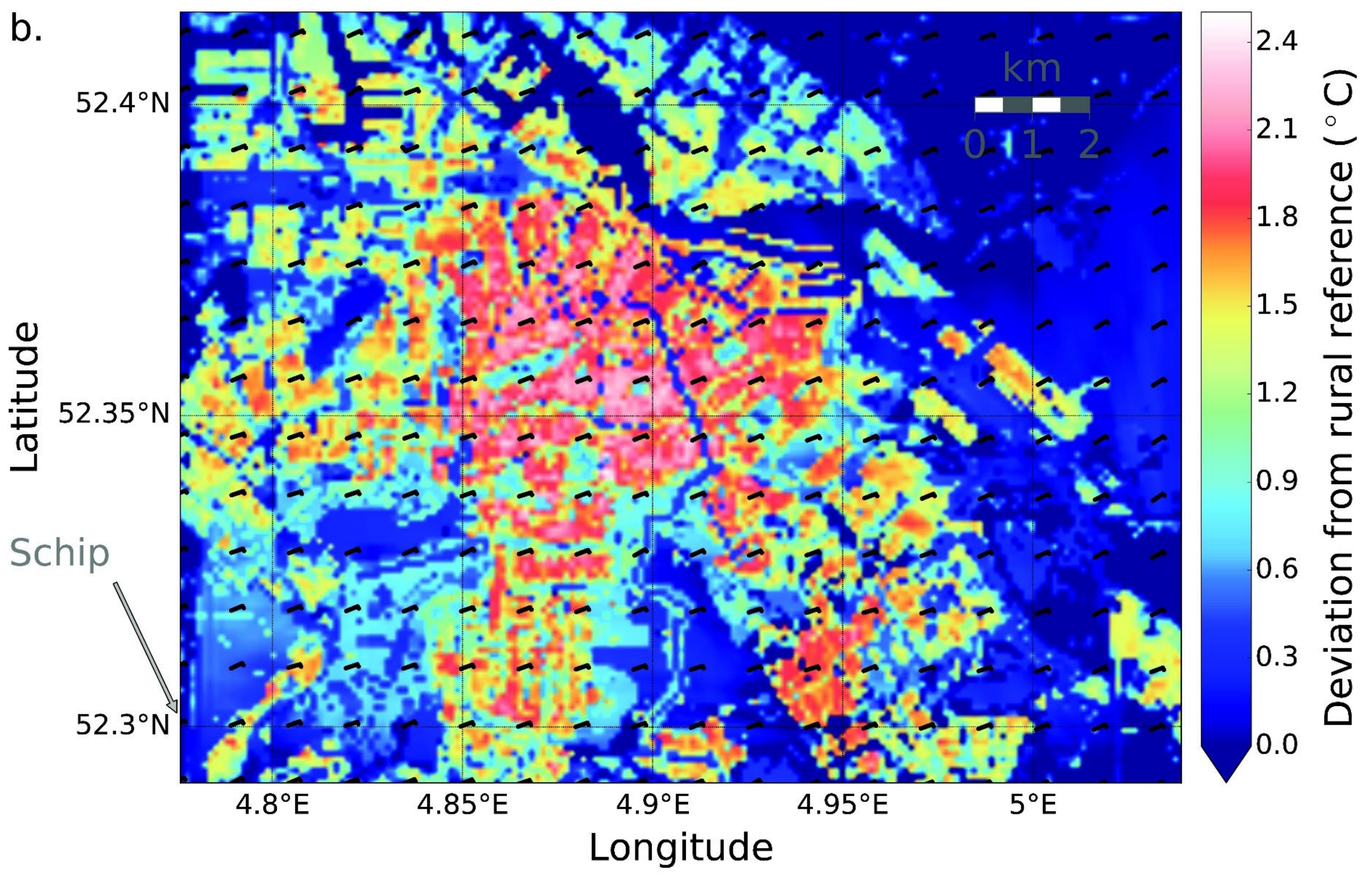
Domain 3: 500m
hi-res landuse,
Rijkswaterstaat river temperatures

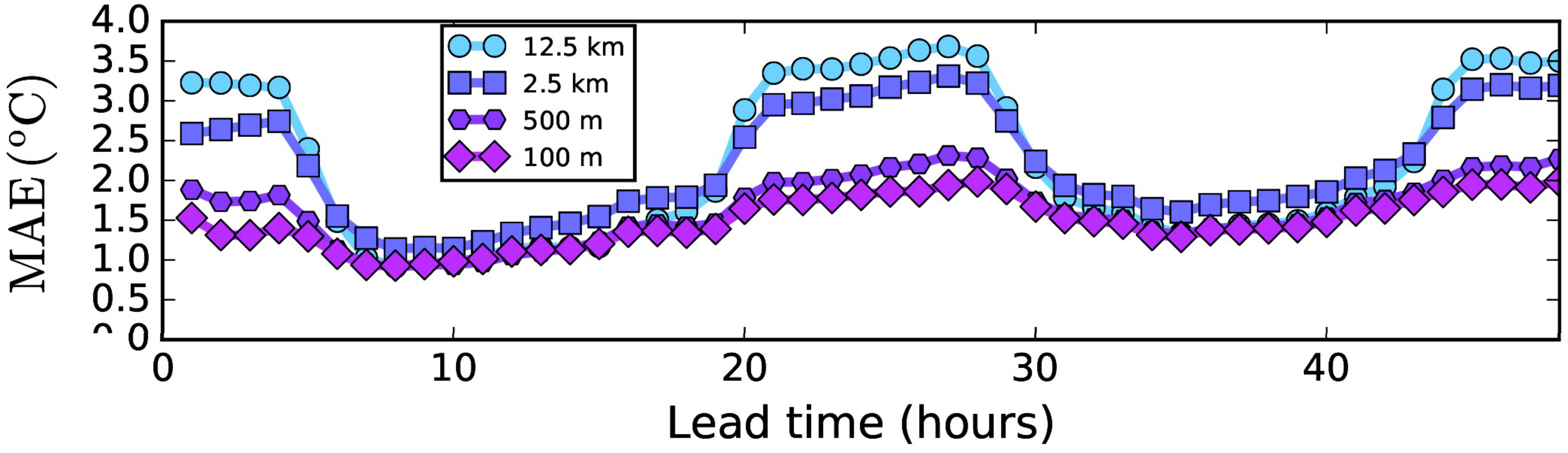
Domain 4: 100m
Rijkswaterstaat river temperatures,
TOP10NL, satellite imagery, AHN2 (height map), CBS data



Attema et al, IEEE eScience, 2015

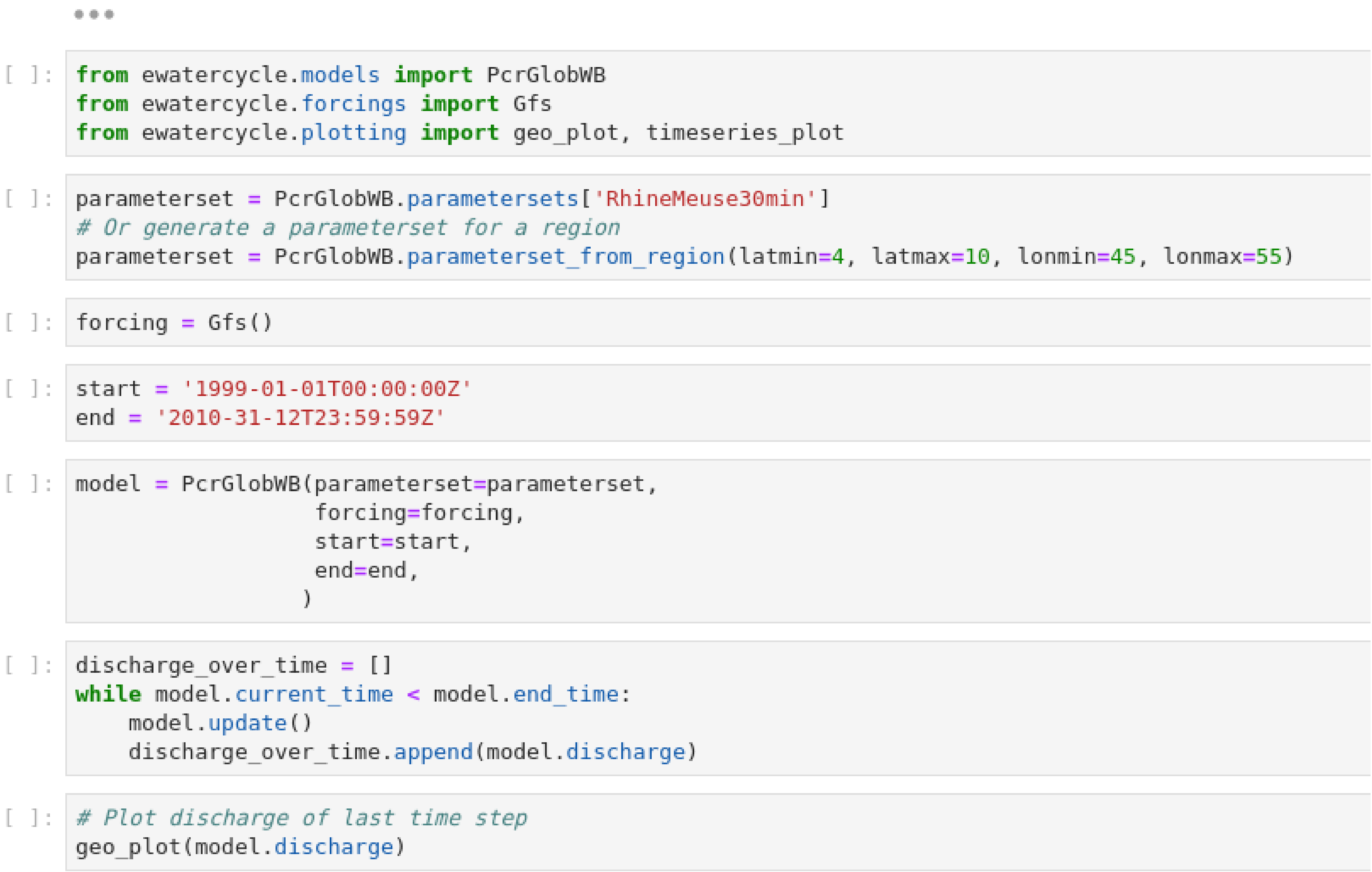
Short range weather forecasting at street level





Ronda et al BAMS 2017

Flexible steering, execution of models and data handling



•••

[ ]:

```
from ewatercycle.models import PcrGlobWB
from ewatercycle.forcings import Gfs
from ewatercycle.plotting import geo_plot, timeseries_plot
```

[ ]:

```
parameterset = PcrGlobWB.parametersets['RhineMeuse30min']
# Or generate a parameterset for a region
parameterset = PcrGlobWB.parameterset_from_region(latmin=4, latmax=10, lonmin=45, lonmax=55)
```

[ ]:

```
forcing = Gfs()
```

[ ]:

```
start = '1999-01-01T00:00:00Z'
end = '2010-31-12T23:59:59Z'
```

[ ]:

```
model = PcrGlobWB(parameterset=parameterset,
                  forcing=forcing,
                  start=start,
                  end=end,
                 )
```

[ ]:

```
discharge_over_time = []
while model.current_time < model.end_time:
    model.update()
    discharge_over_time.append(model.discharge)
```

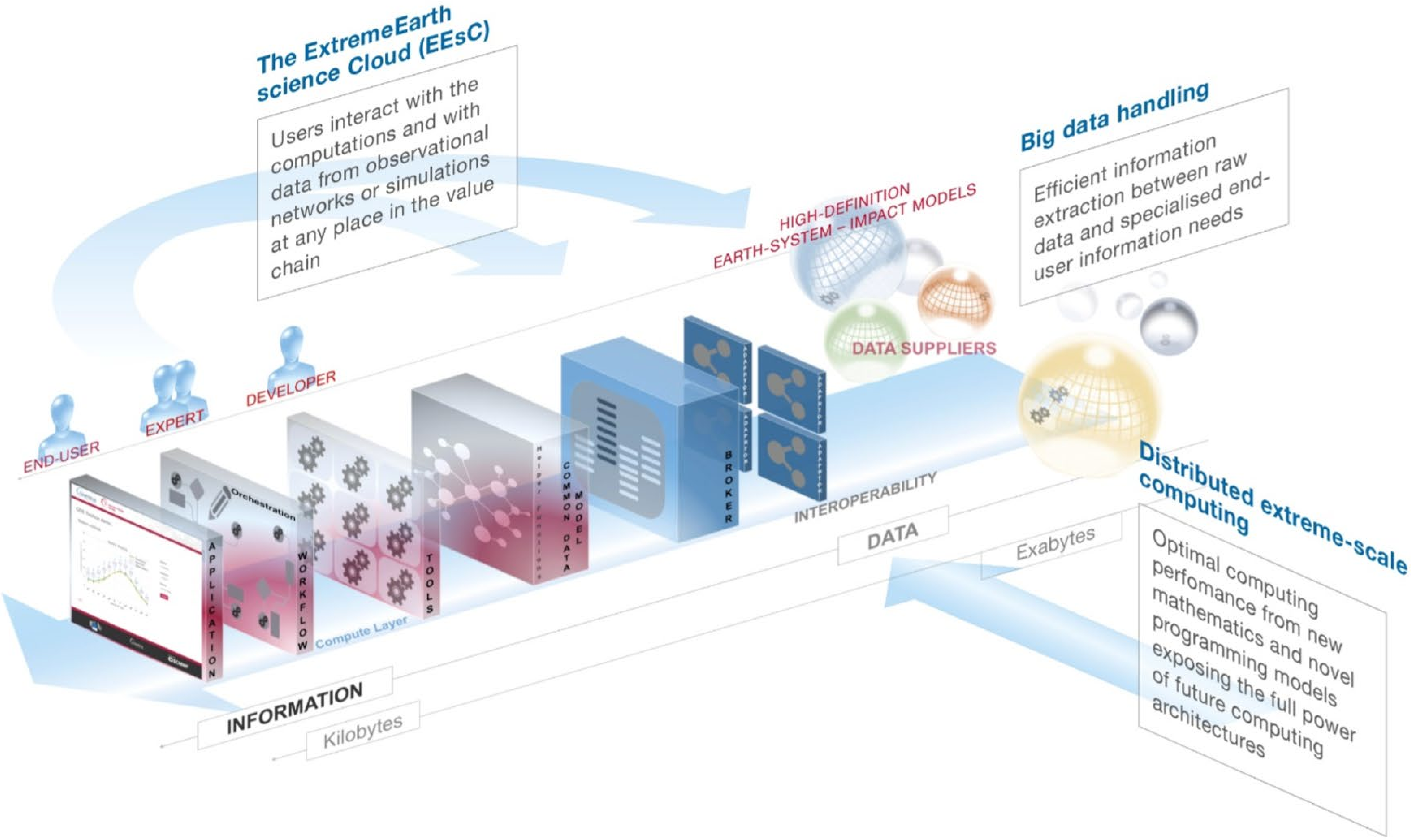
[ ]:

```
# Plot discharge of last time step
geo_plot(model.discharge)
```

Niels Drost, pers. Comm, NLeSC/TUD/UU/WUR/Deltares eWatercycle II project

Still…far from actionable information for decision making

What e-infrastructure does it take?



The ExtremeEarth science Cloud (EEsC)

Users interact with the computations and with data from observational networks or simulations at any place in the value chain

Big data handling

Efficient information extraction between raw data and specialised end-user information needs

Distributed extreme-scale computing

Optimal computing perfomance from new mathematics and novel programming models exposing the full power of future computing architectures

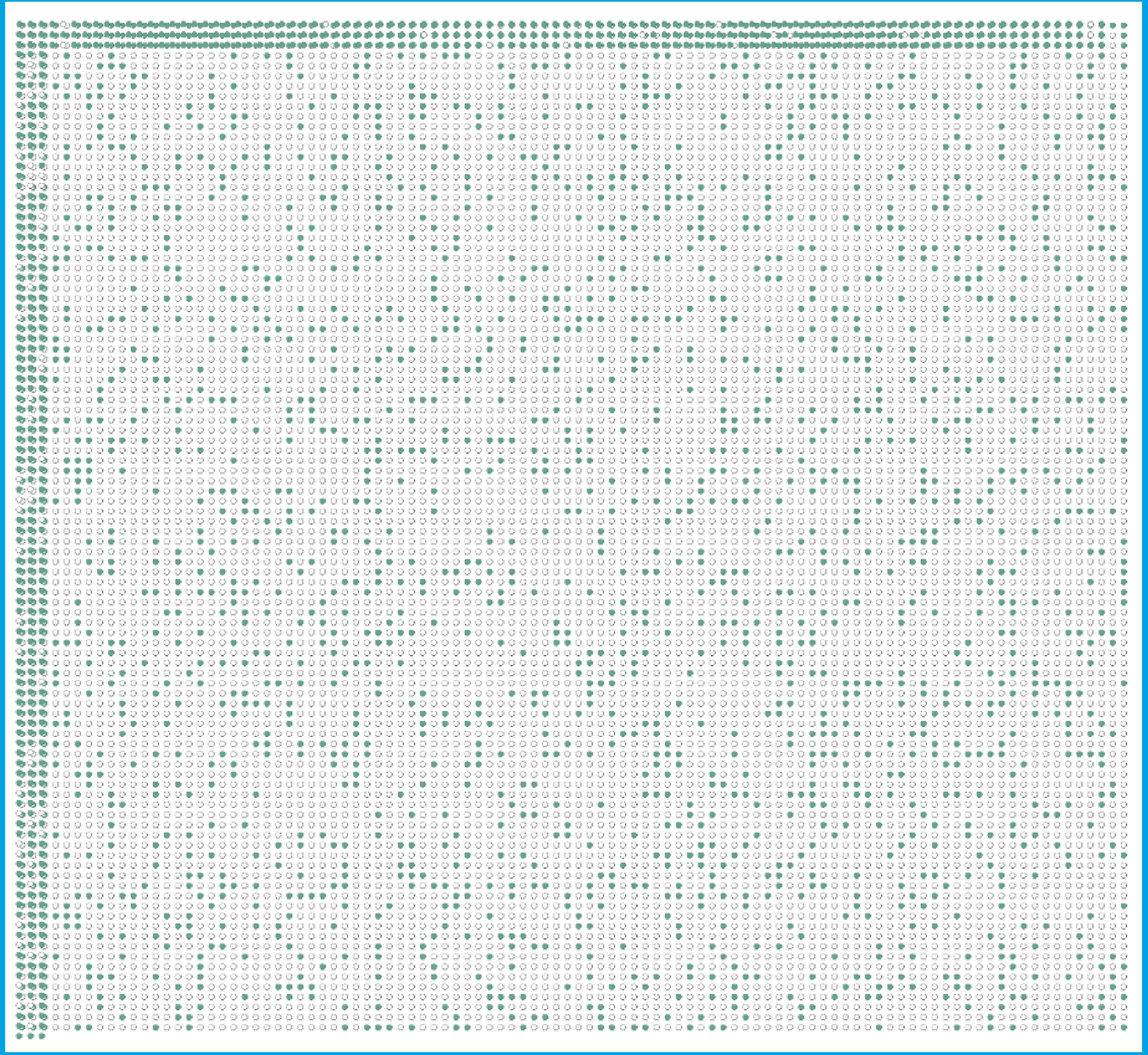
Examples of FAIR in life sciences

Big data landscape in health care

- **Clinical research**
  - **3% of patients**
  - **100% of features**
  - **5% missing**
  - **285 data points**

- **Clinical registries**
  - **100% of patients**
  - **3% of features**
  - **20% missing**
  - **240 data points**

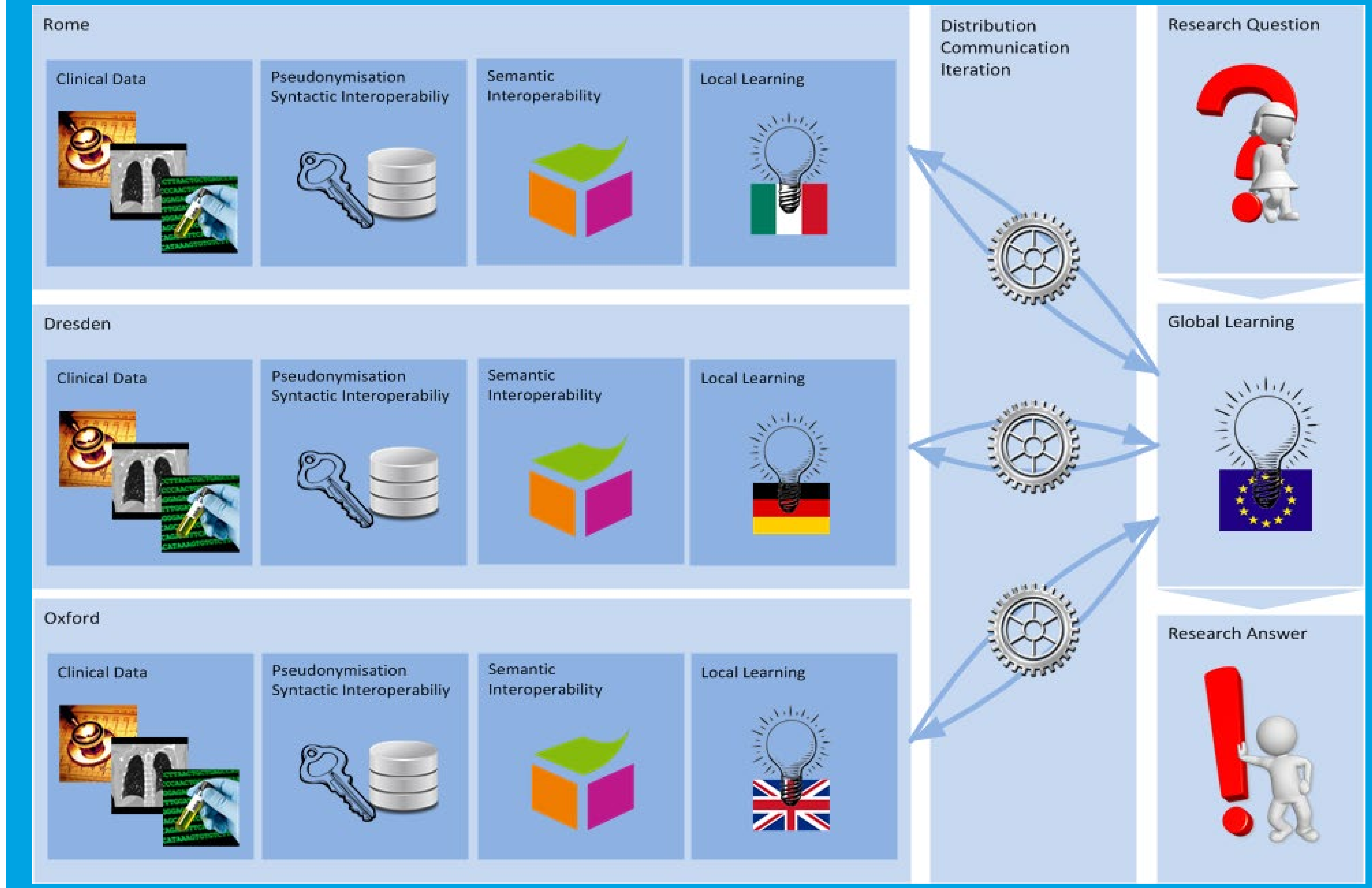
- **Clinical routine**
  - **100% of patients**
  - **100% of features**
  - **80% missing**
  - **2000 data points**



MAASTRO Maastricht UMC+ **Andre Dekker**

A Global Distributed Routine Data Registry

- Keep data locally
- Standardize it according to an ontology
- Make and send around learning and quality indicators
- Share the results & quality indicators – not the data!!


Rome
Clinical Data
Pseudonymisation Syntactic Interoperabiliy
Semantic Interoperability
Local Learning
Dresden
Clinical Data
Pseudonymisation Syntactic Interoperabiliy
Semantic Interoperability
Local Learning
Oxford
Clinical Data
Pseudonymisation Syntactic Interoperabiliy
Semantic Interoperability
Local Learning
Distribution Communication Iteration
Research Question
Global Learning
Research Answer

Reproducible science

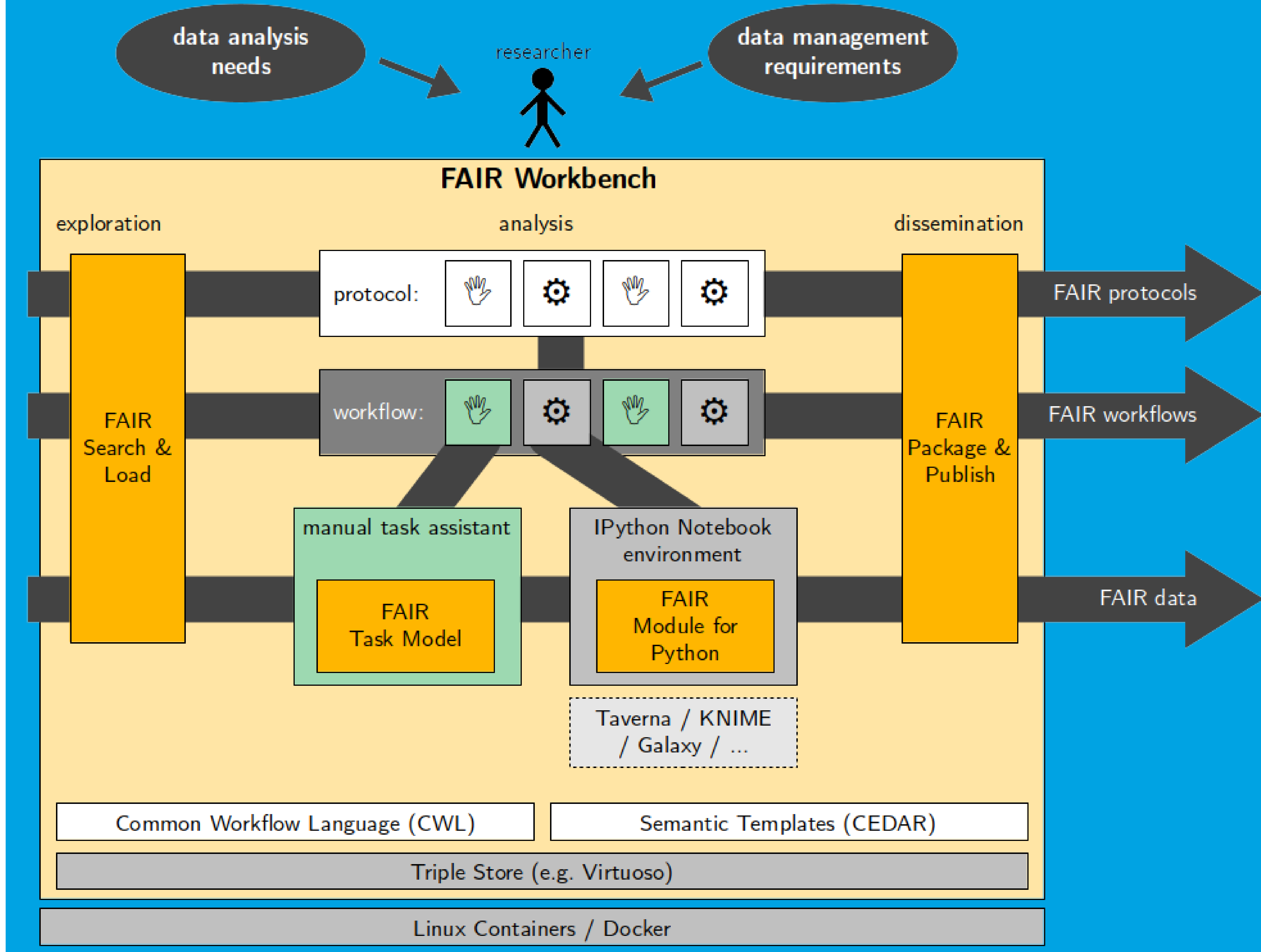
Requires not only **data** to be FAIR but also

Software:

- Research Software Directory (F + A)
- Use standard file formats, Docker, API's, etc. (I)
- NLeSC guide: https://guide.esciencecenter.nl/ (R)

Workflows:

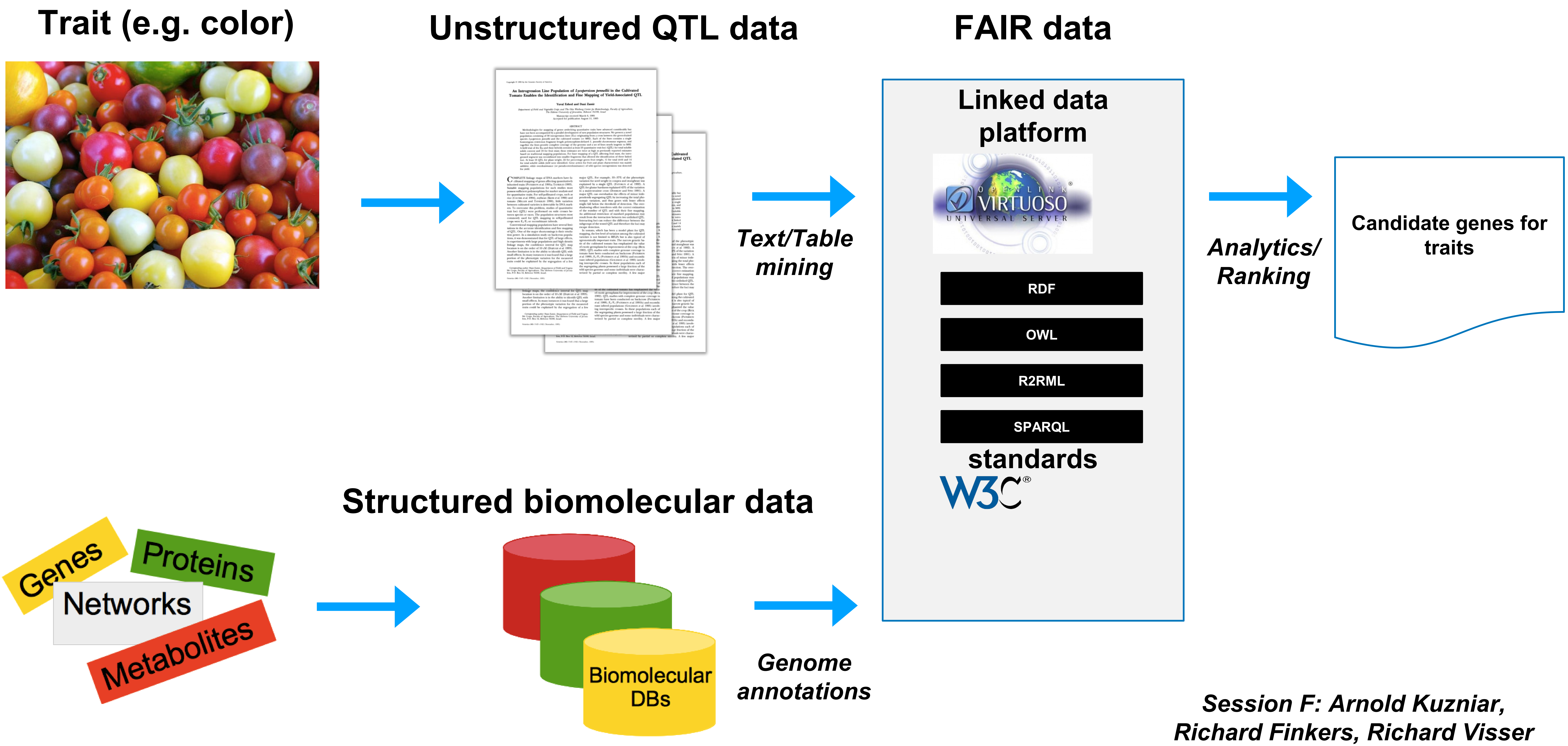
- Common Workflow Language
  - platform independent workflow definition and execution



FAIR Workflows project

Collaboration with Tobias Kuhn, Michel Dumontier

Linked data platform to relate genes with traits



Trait (e.g. color)

Unstructured QTL data

Text/Table mining

FAIR data

Linked data platform

RDF

OWL

R2RML

SPARQL

standards

Analytics/ Ranking

Candidate genes for traits

Structured biomolecular data

Genes

Proteins

Networks

Metabolites

Biomolecular DBs

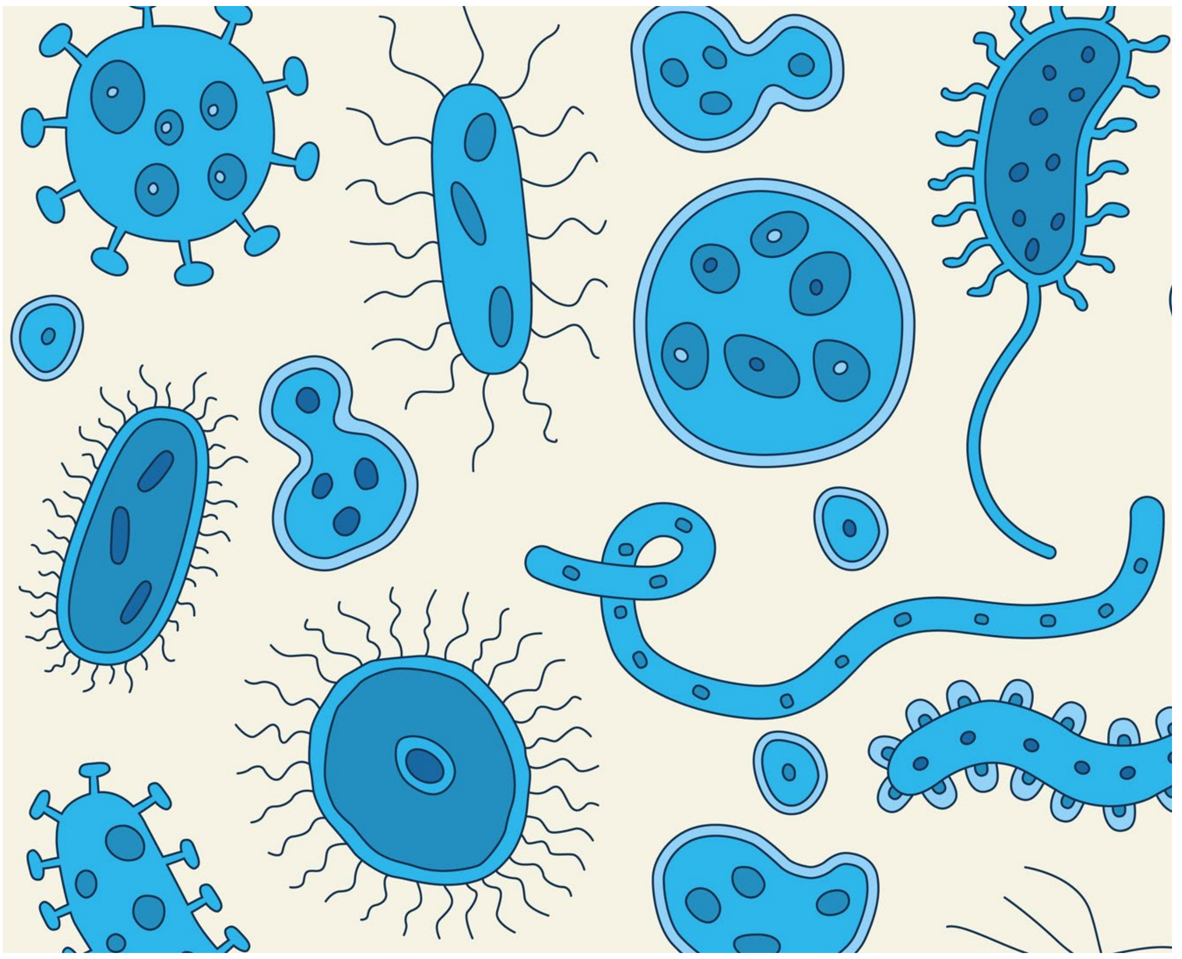
Genome annotations

Session F: Arnold Kuzniar, Richard Finkers, Richard Visser

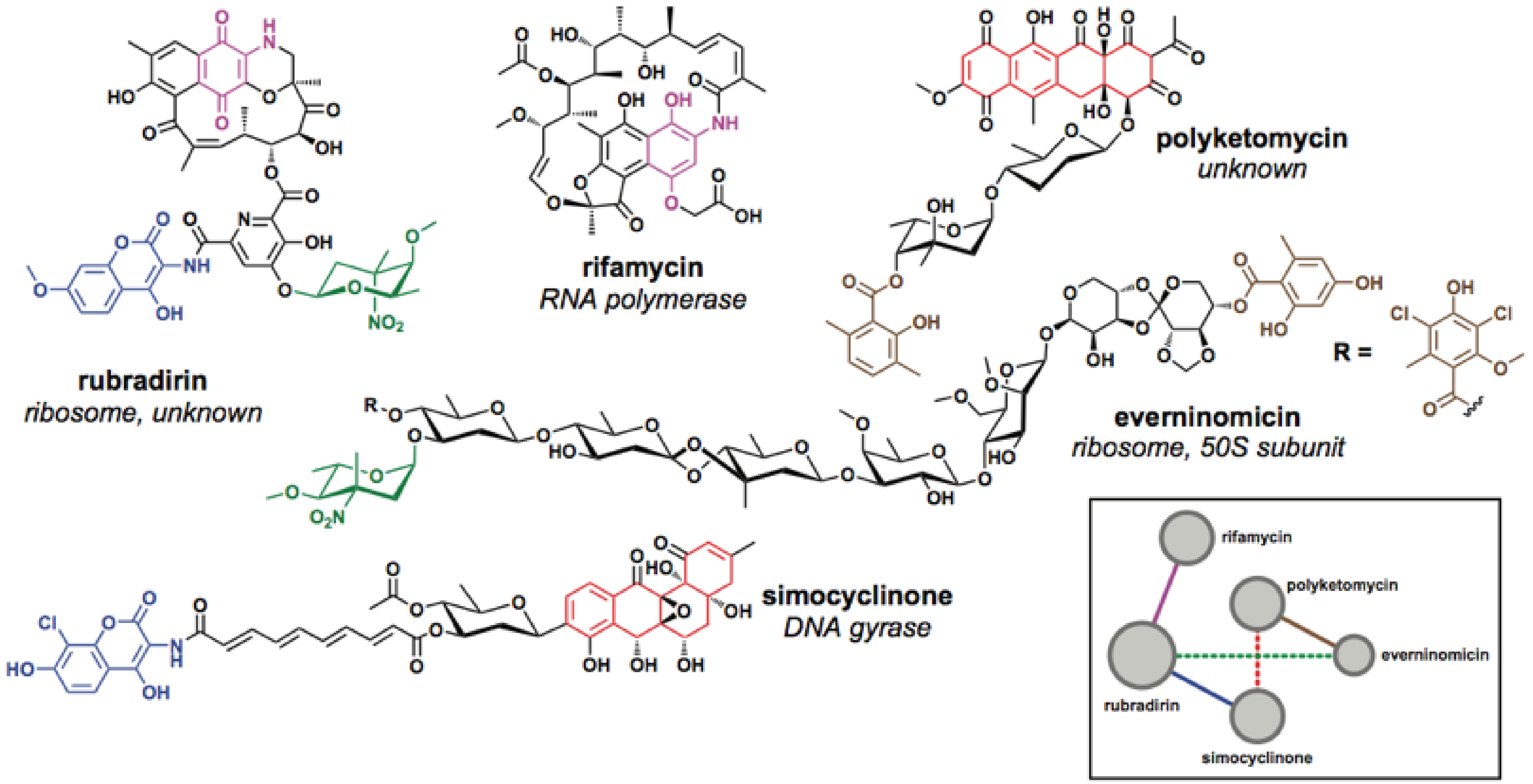
https://github.com/candYgene

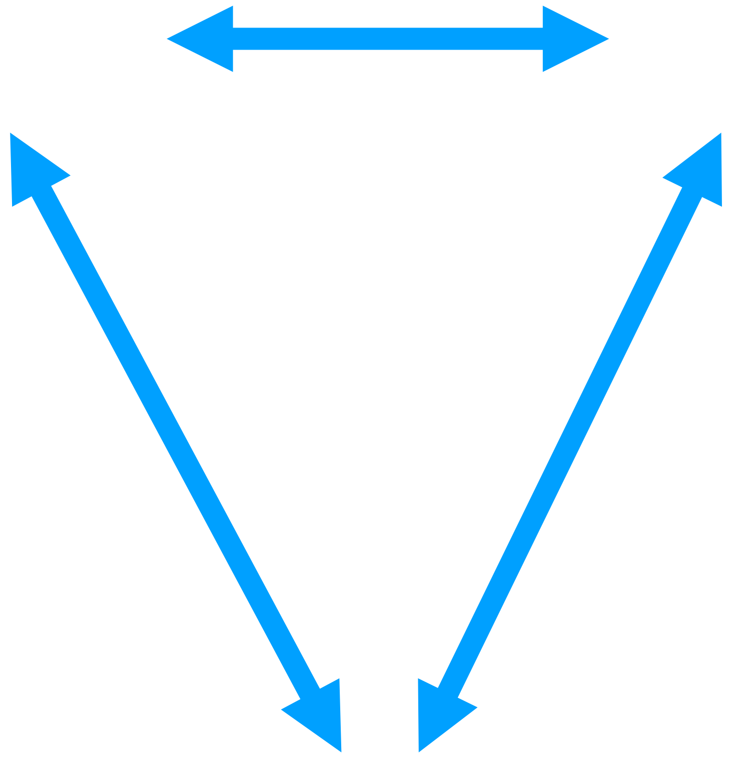
Integrated analysis and annotation of multi -omics data

Microbiome (e.g. in the human gut)

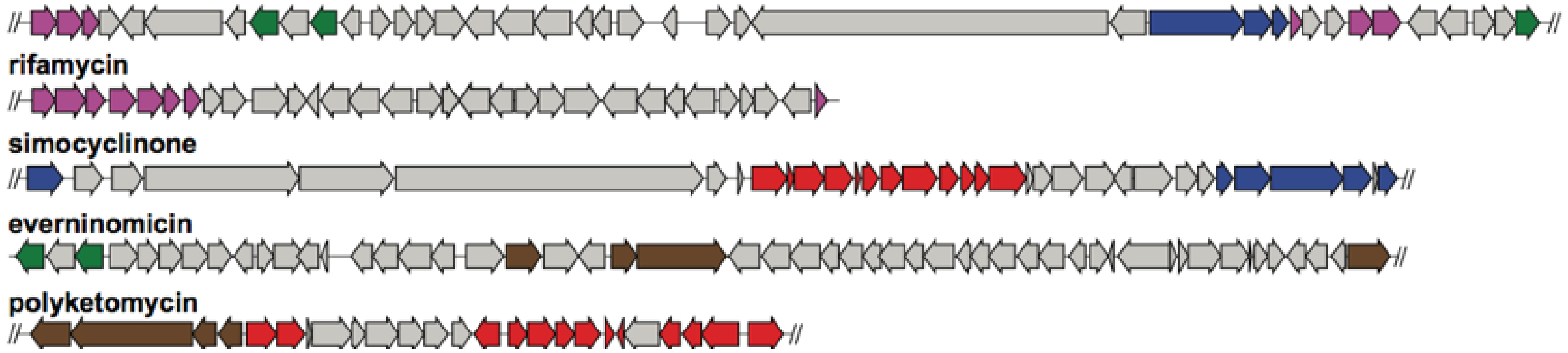


Metabolomes (Natural products)





Genomes (Biosynthetic gene clusters)



Session F: iOMEGA project
Justin van der Hooft, M. Medema
S. Verhoeven, F. Huber, L. Ridder

FAIR: Just do it!

- Supports working across domains of research
- Requires domain knowledge, digital competences & digital infrastructures, hence an collaborative work environment!
- Absolutely necessary for evidence - based (and transparent) decision making
- Not only data, but also software, worfklows , methods..

Thank you