

eScience and Big (Data) Science

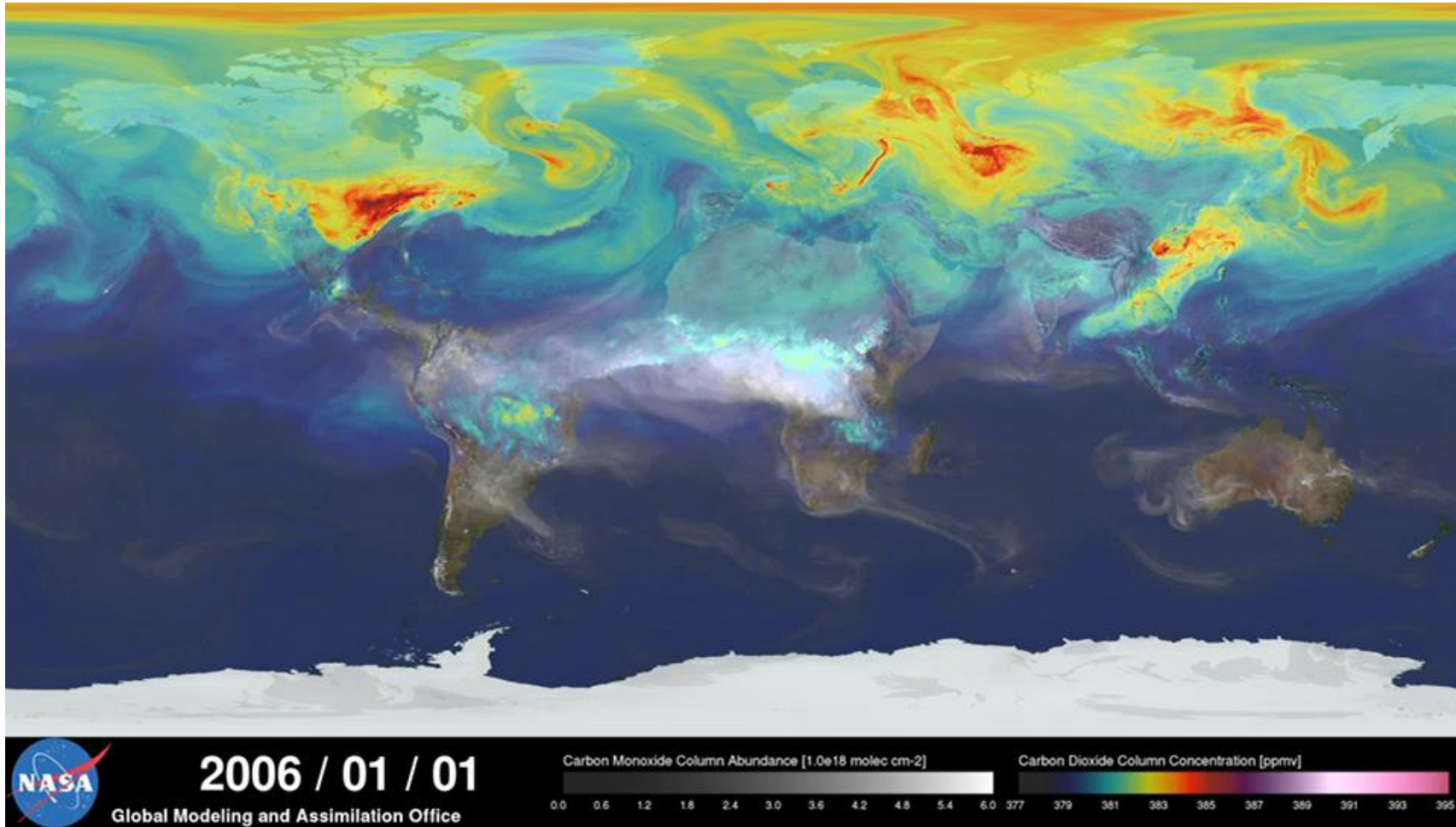
Wilco Hazeleger

www.esciencecenter.nl

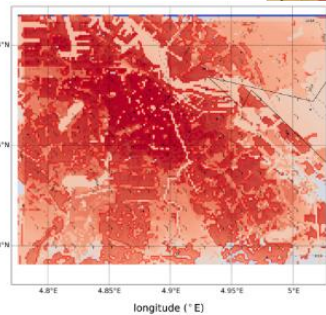
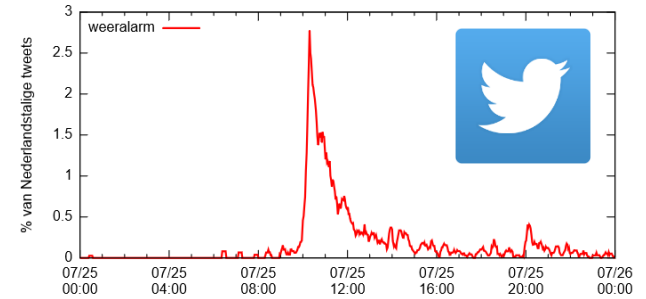
netherlands

eScience center

by SURF & NWO

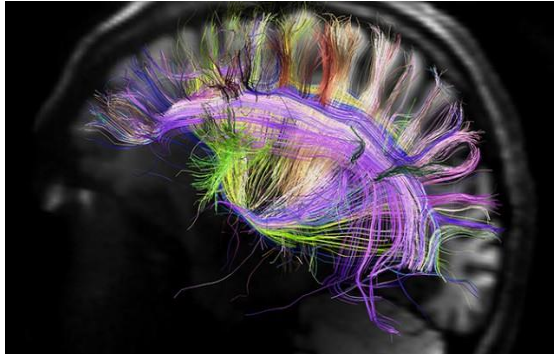
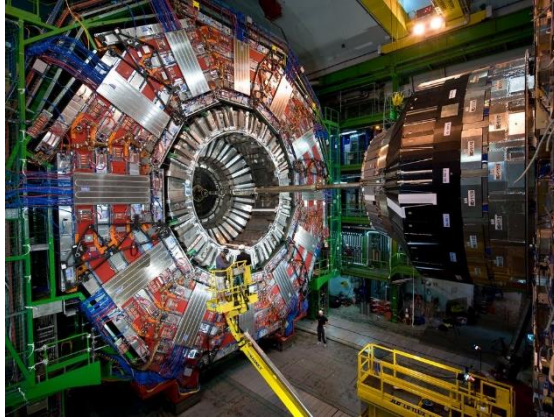


New data sources

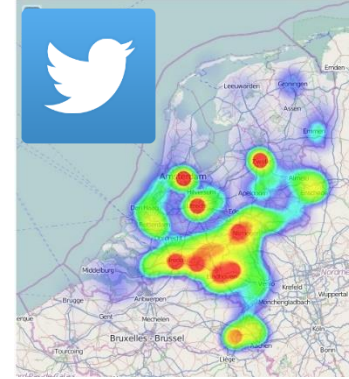
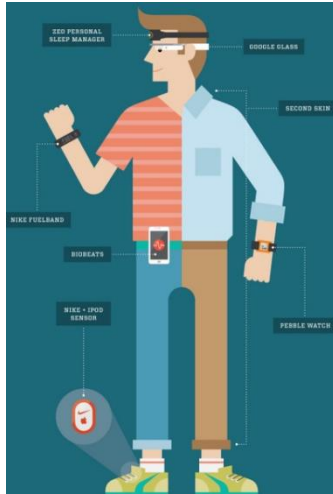


<https://www.esciencecenter.nl/project/summer-in-the-city>

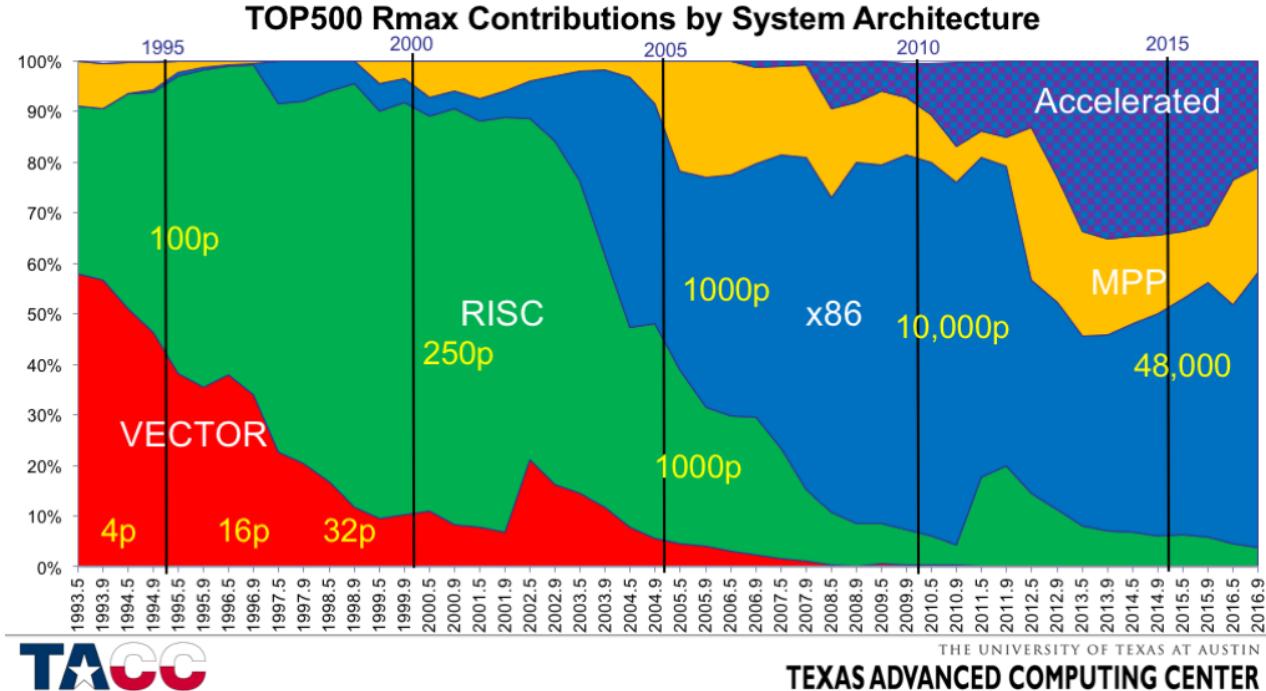
The 'Big Science' era



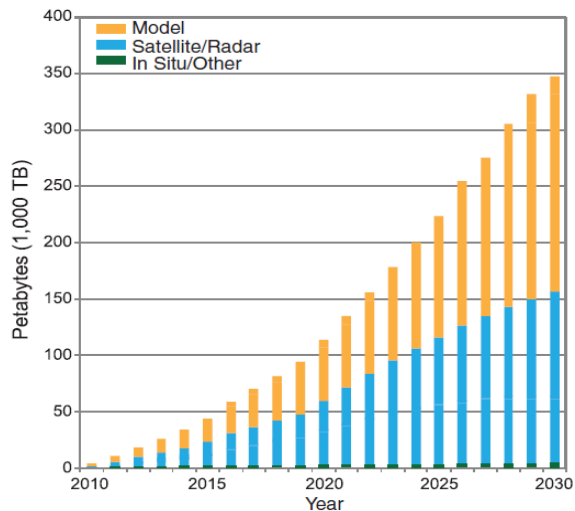
The other 'Big Data'



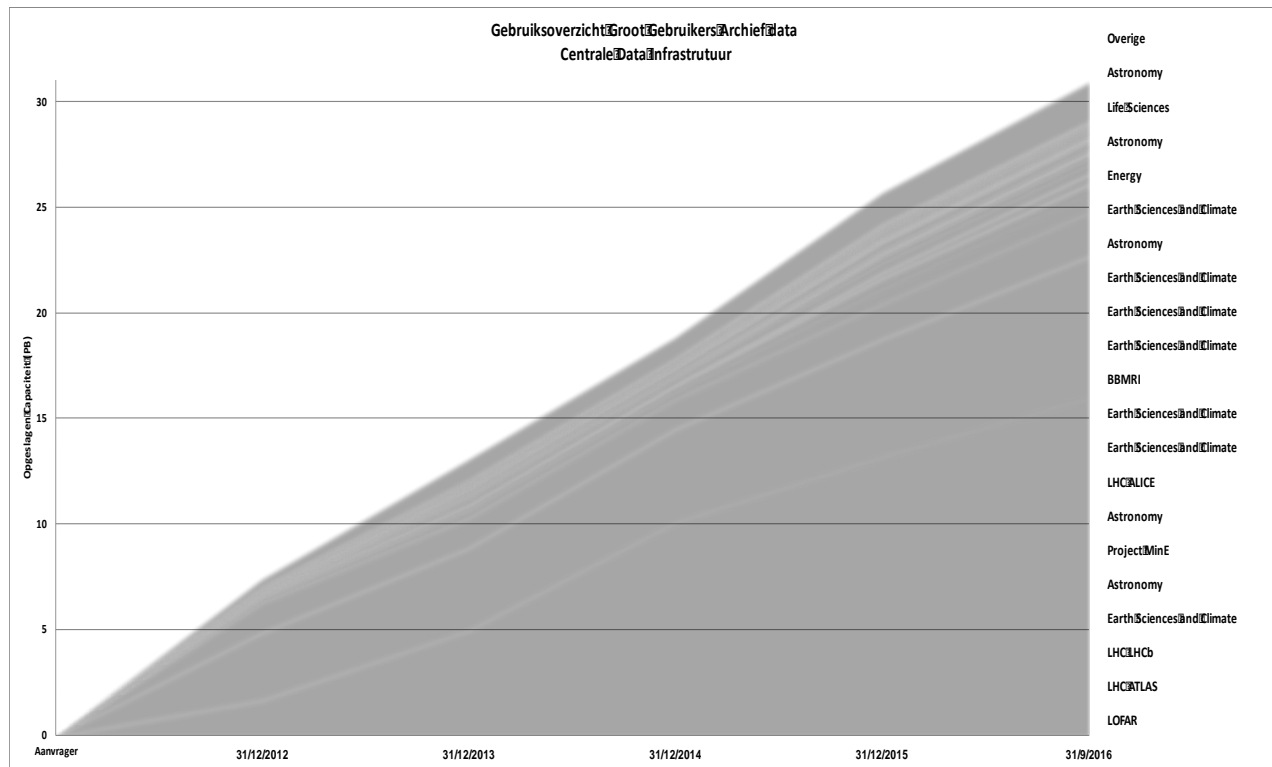
Computing challenge: towards exascale



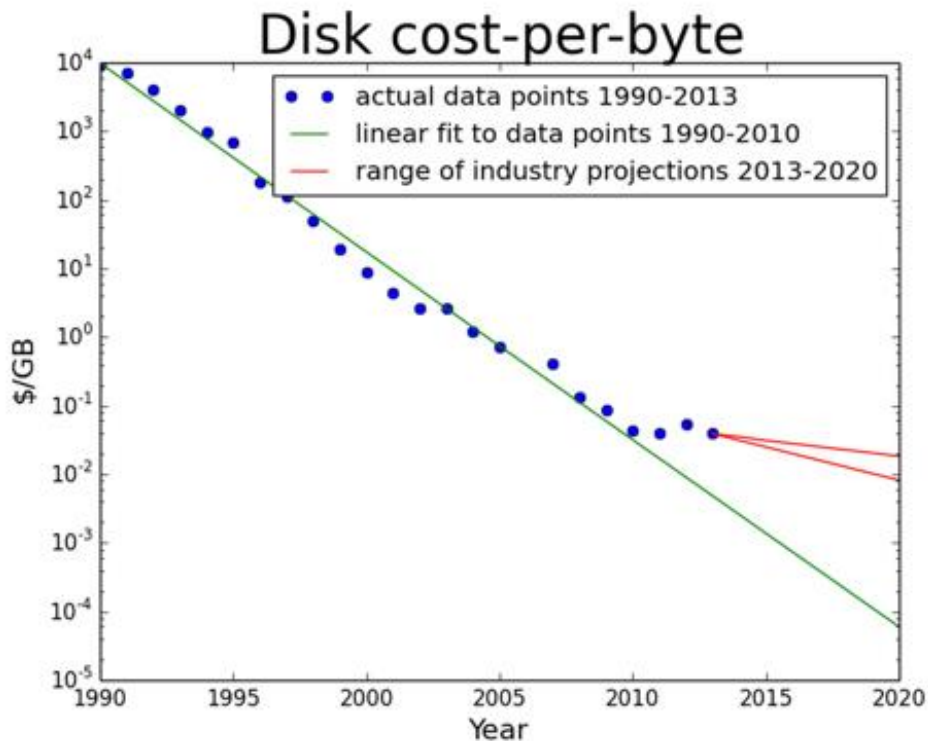
The data challenge



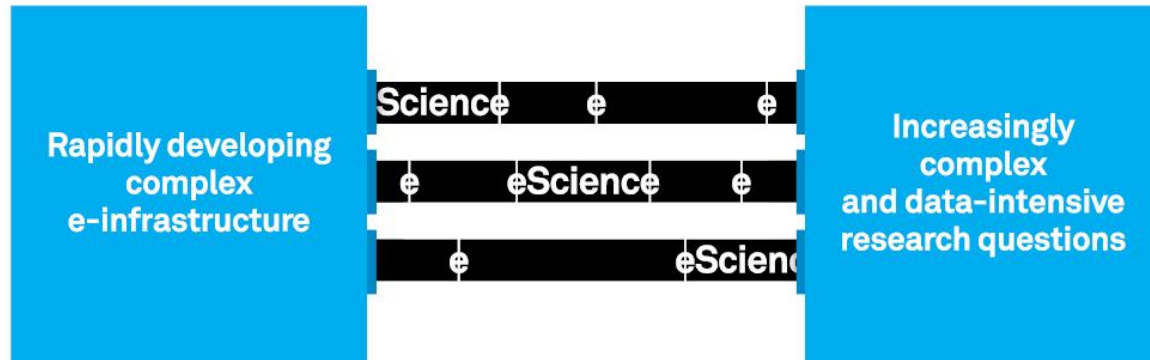
Overpeck et al, Science 2011



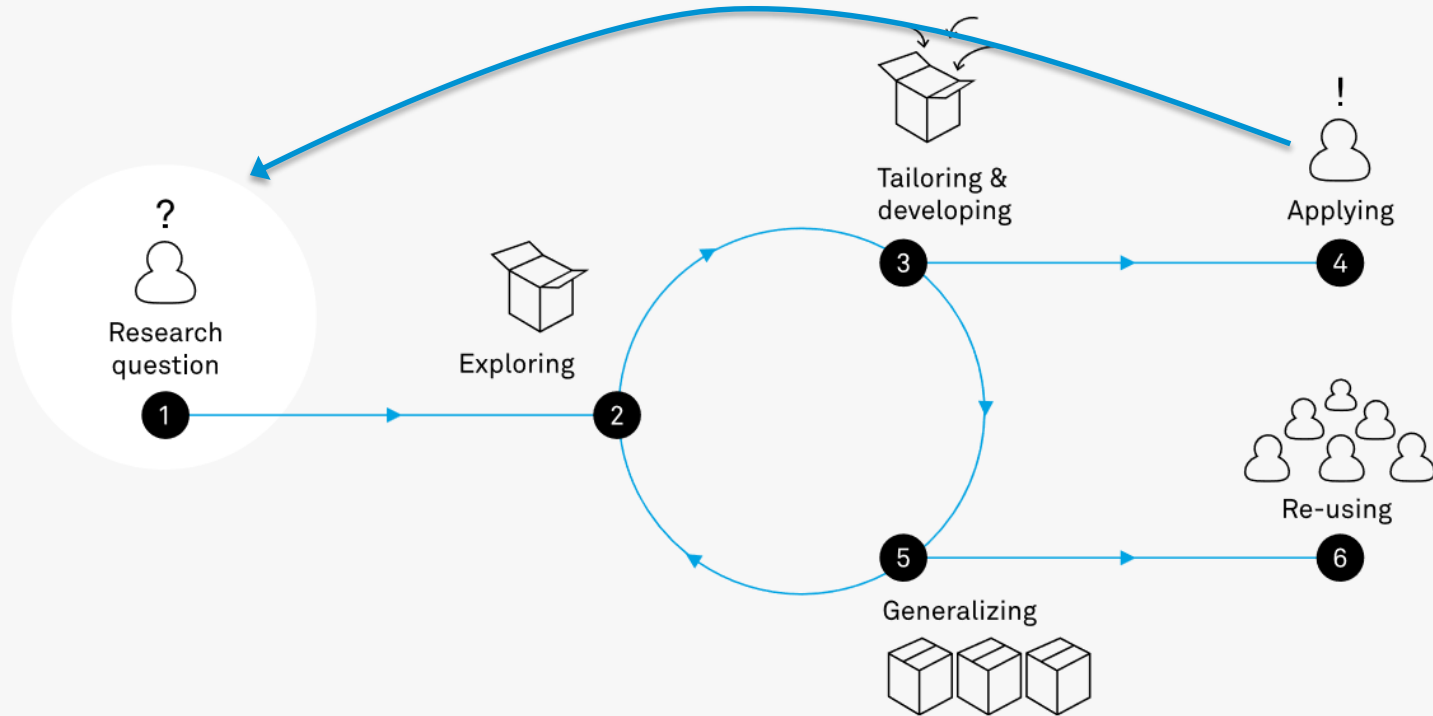
Data increasingly incomputable?



Why eScience



Developing research software: it starts with a question



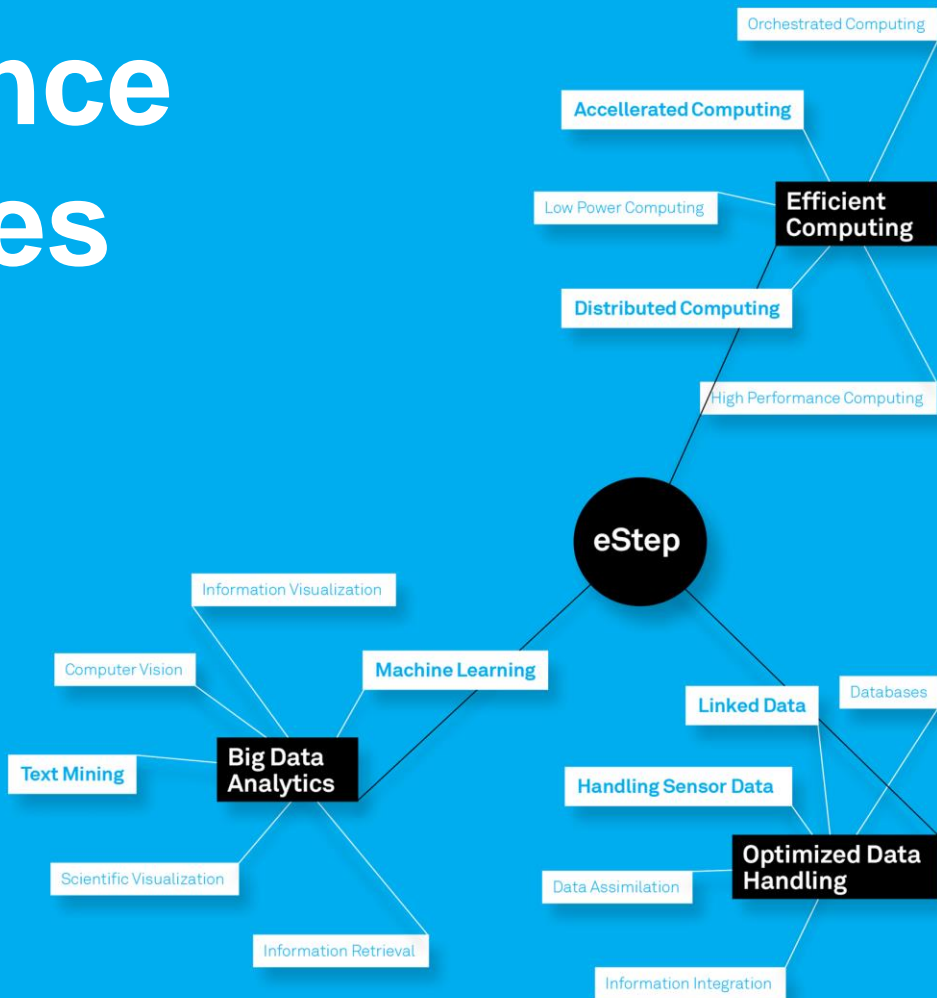
How we work

- **eScience Research Engineers**
- **Partnership with domain scientist**
 - Open calls for proposals
 - All of science
- **Public private partnerships**
- **(inter)National coordination and advocacy (PLAN-E for Europe)**
- **Generic *eScience Technology Platform*, eStep (software, open access, knowledge basis)**

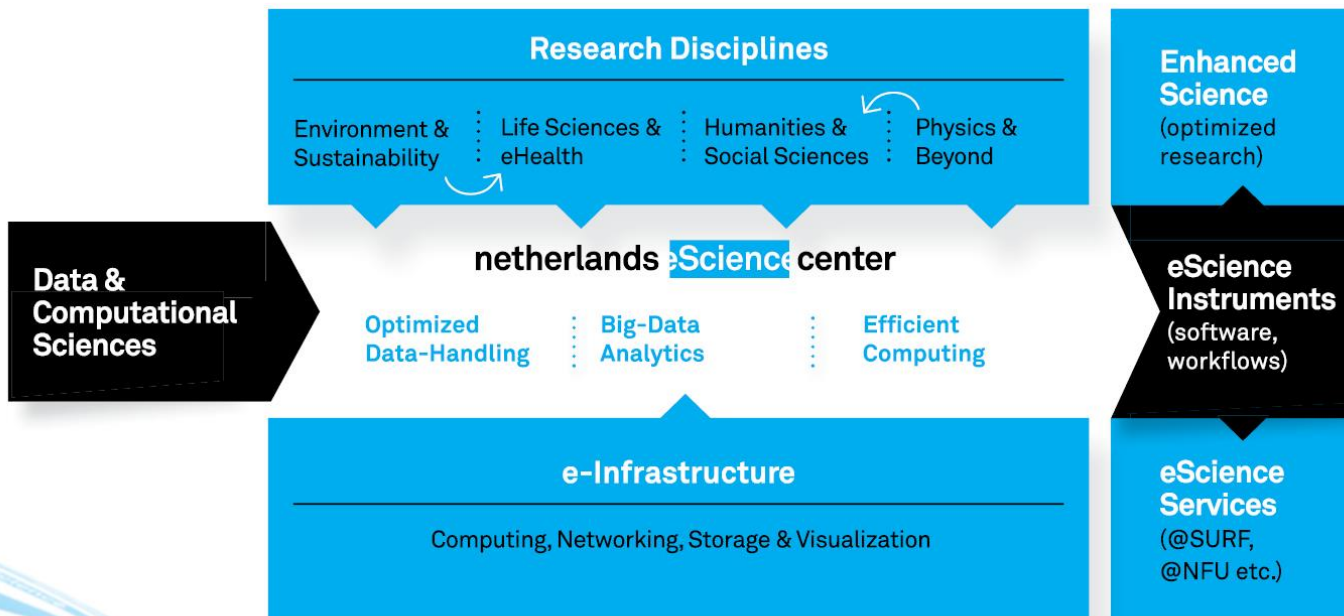


Core eScience Technologies

eStep.eScienceCenter.nl

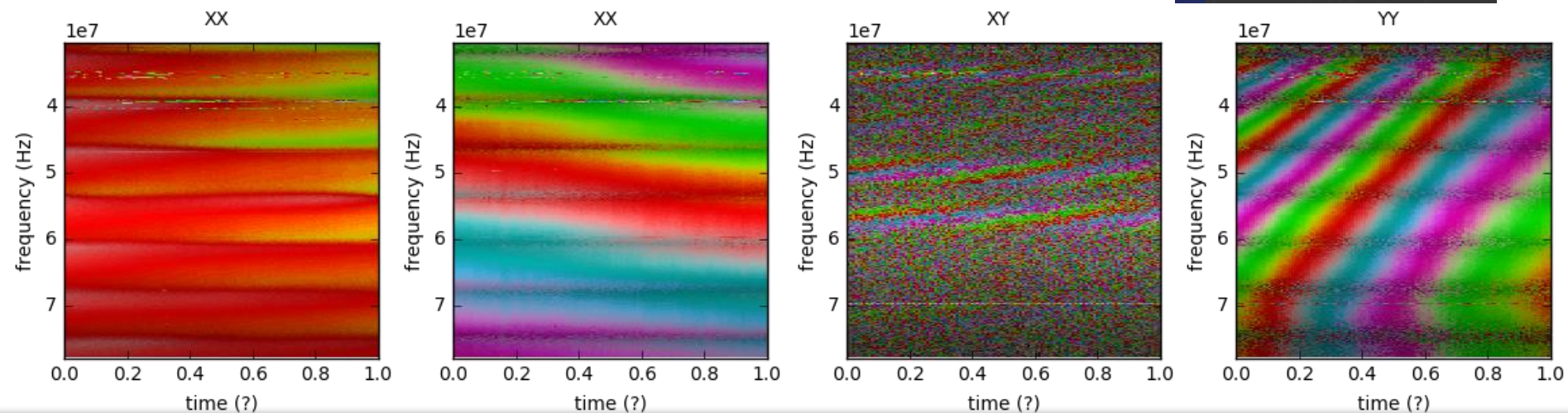
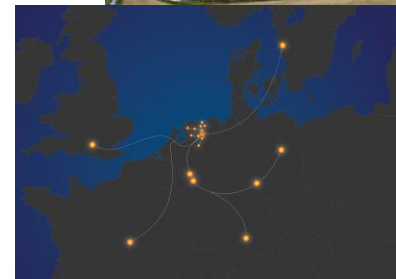


How we work



Radio telescope system health management (error detection)

Boonstra (Astron), Meijer (NLeSC) et al

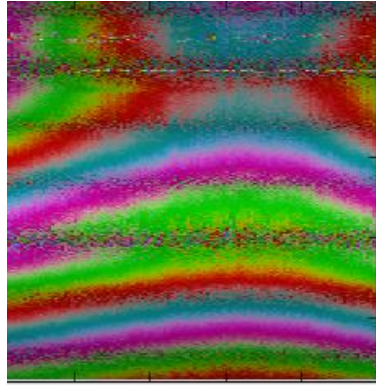


Automatic classification

- Stations not sending data
- Stations with internal errors (RFI?)
- Ionospheric scintillation
- Stations having low gain
- Non-linearities in receivers/ADC's
- Dead cobalt node
- Extreme external RFI
- Solar bursts
-

Simple distance metric

$$d = \sum |$$



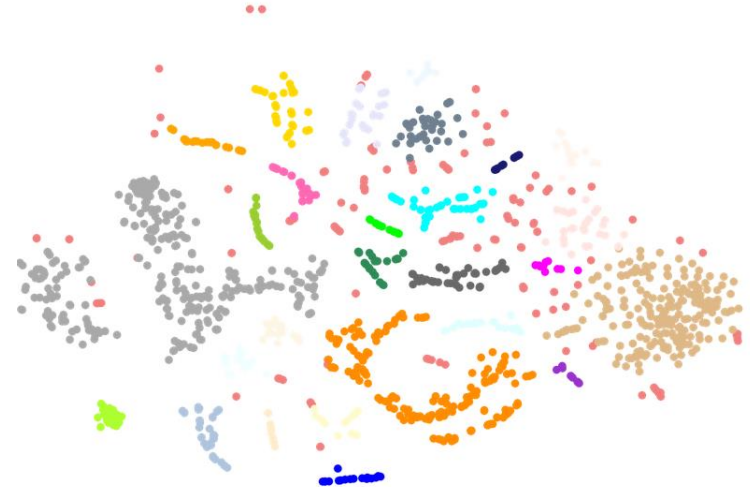
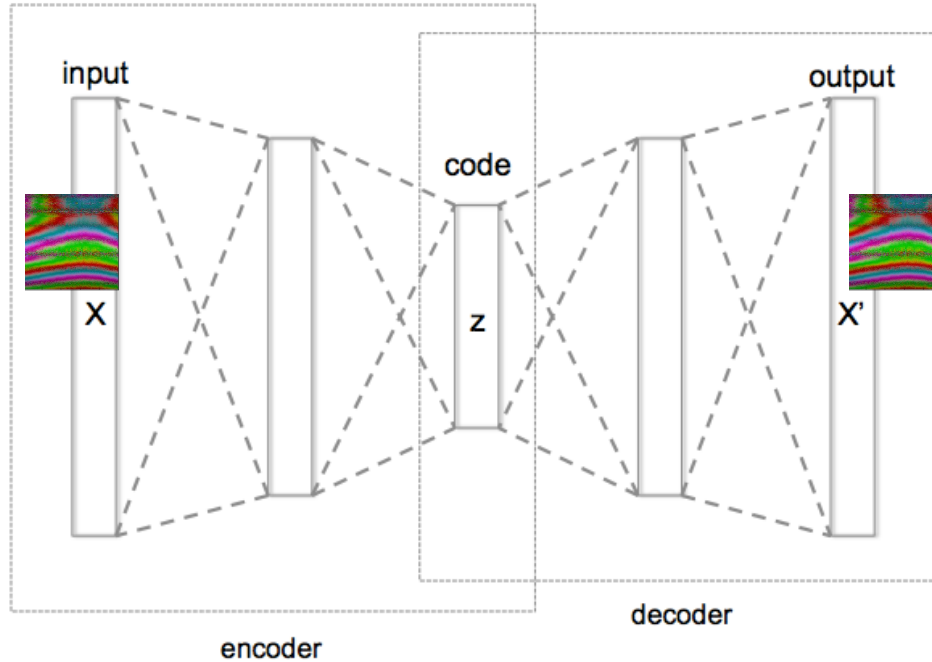
-



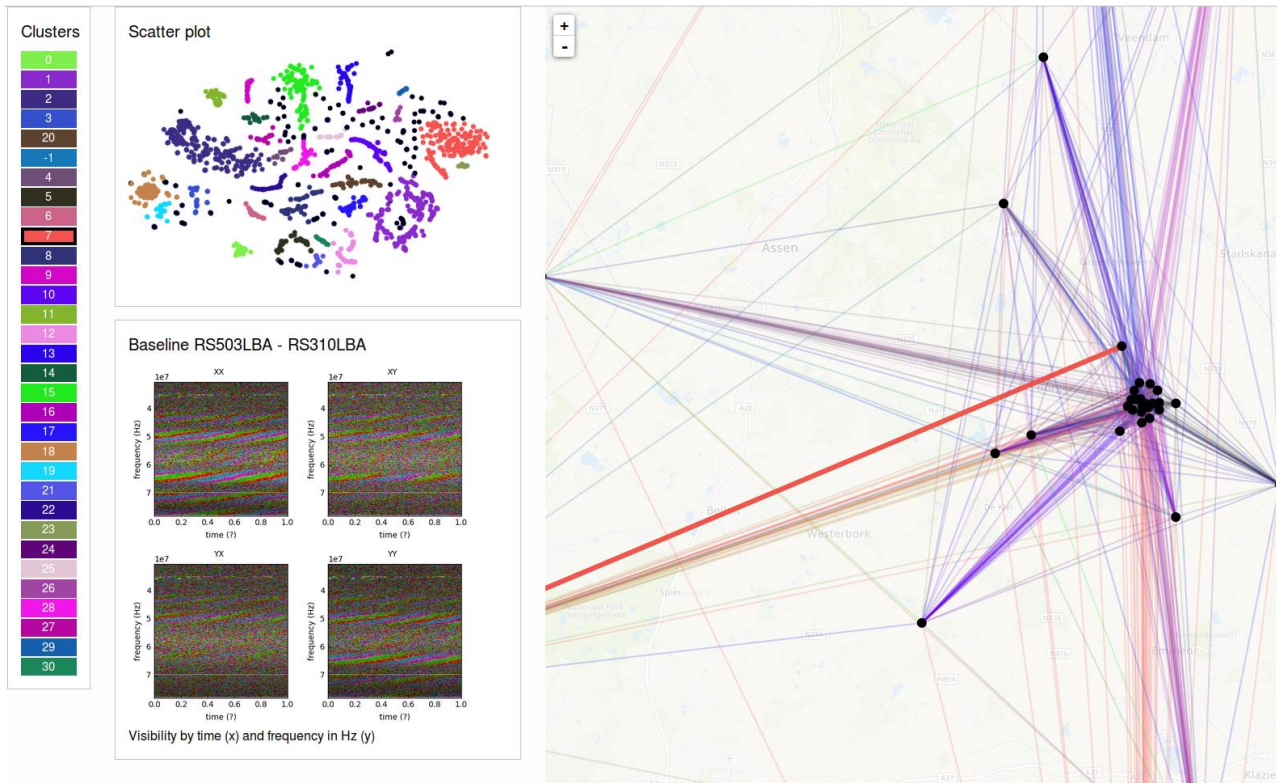
|



CNN distance metric



Results viewer



DIRAC

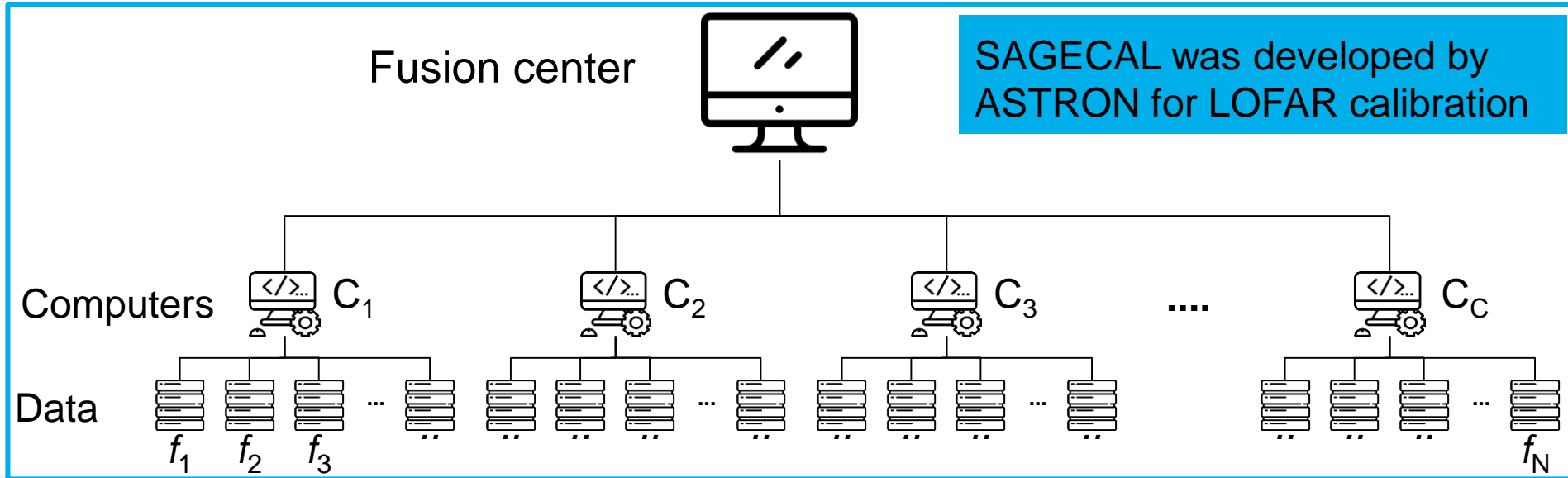
Yatawatta (Astron), Diblen, Spreeuw (NLeSC) et al.

- Software radio telescope searching for faint signals from the early universe
- The signal is **order of magnitude fainter** than the most contaminating signals
- Need to eliminate all systematic (instrumental, ionospheric etc.) errors (i.e. “**calibrated**”)
- The calibration in **parallel** on different data frequencies which requires processing of **many terabytes of data**



DIRAC - Calibration

- Complex non-linear optimization problem with **millions of unknown parameters**
- **global calibration scheme** is needed
- Solutions should be **continuous** over frequency



DIRAC - eScience + Big data

The existing code

SAGECAL

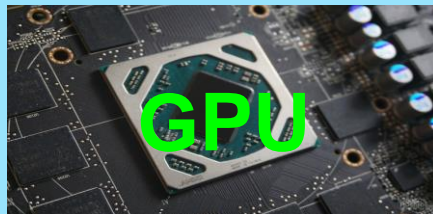
Used for calibration of
LOFAR

Candidate for **SKA**

- written in C/C++
- using MPI
- has GPU support



eScience Center contribution



- Optimization and generalization of the **existing code** for the state-of-the-art GPU architectures

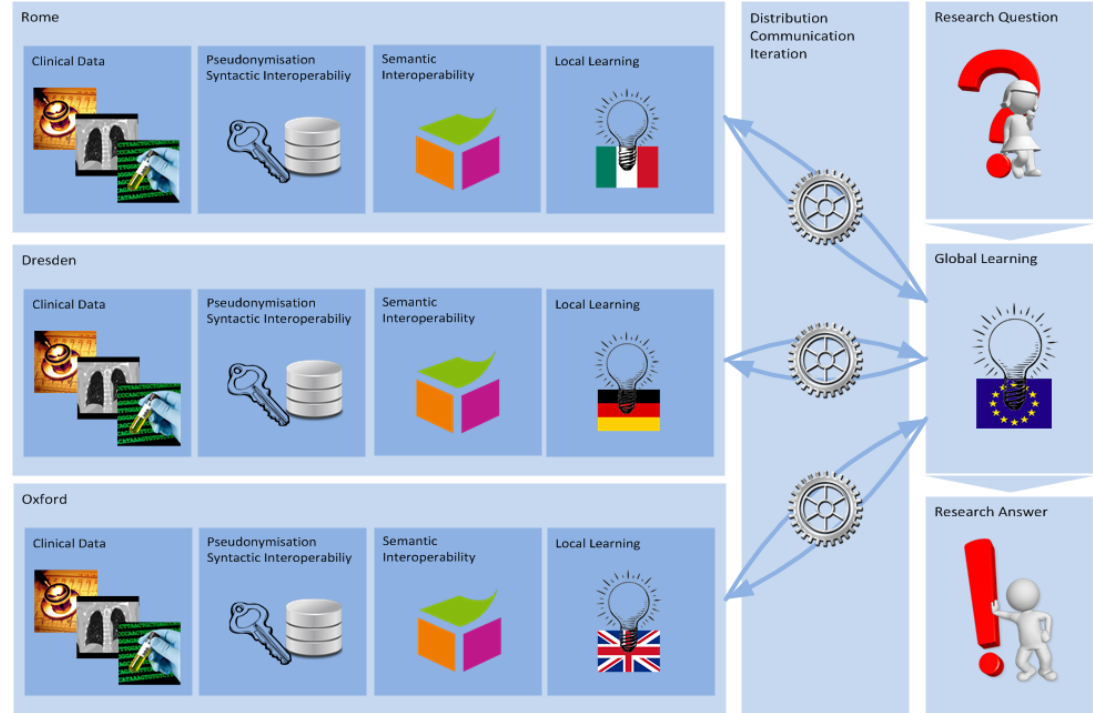


- Migrating the optimized code to **big data (Apache Spark)** platform
- Optimizing the workflow for Big data platform



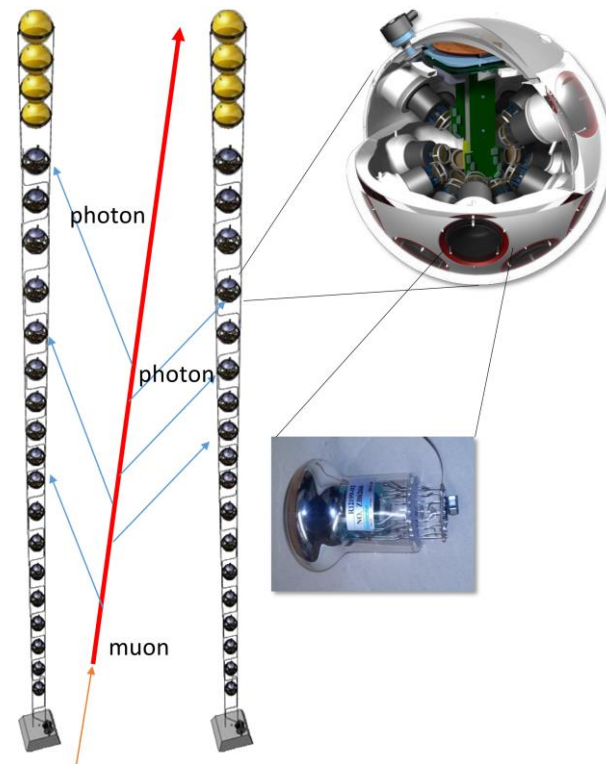
A Health research problem: Global Distributed Routine Data Registry

- **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!!**



KM3NeT – Neutrino Telescope

- Huge instrument at the bottom of the Mediterranean Sea
- Pretty high data rate due to background noise from bioluminescence and Potassium-40 decay
- Current event detection / reconstruction happens on pre-filtered data (so called L1 hits)
- Our goal: Work towards event detection based on unfiltered data (so called L0 hits)

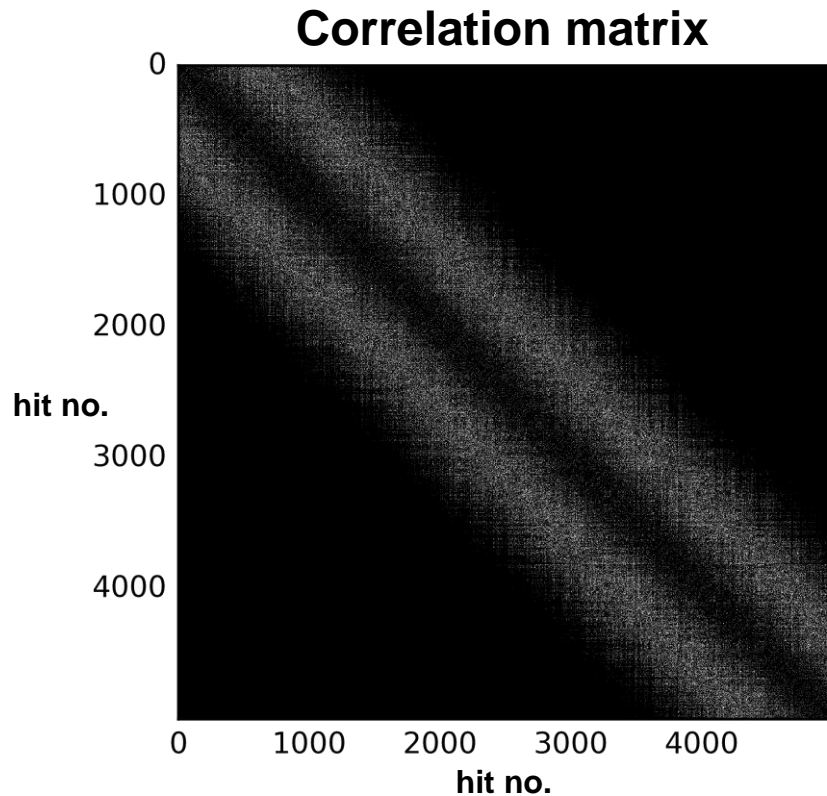


Correlating hits

- Hits are correlated based on their time and location
- Correlations can only occur in a small window of time
- Density of the narrow band depends on correlation criterion in use

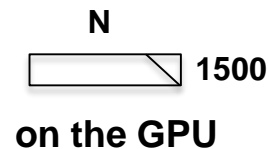
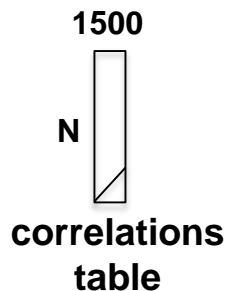
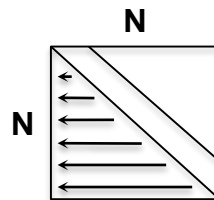
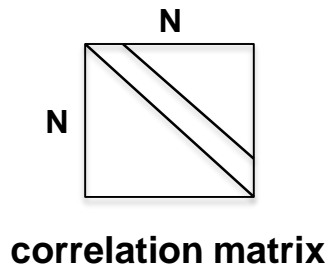
Try-out two designs:

- Dense pipeline that stores the narrow band as a table
- Sparse pipeline that stores the matrix in compressed sparse row (CSR) form

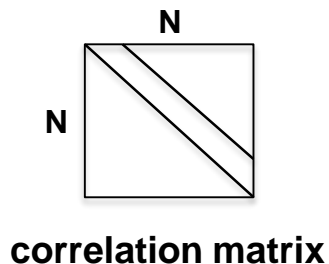


Data representation

— Dense



— Sparse



CSR format

column indices

start of row

correlations

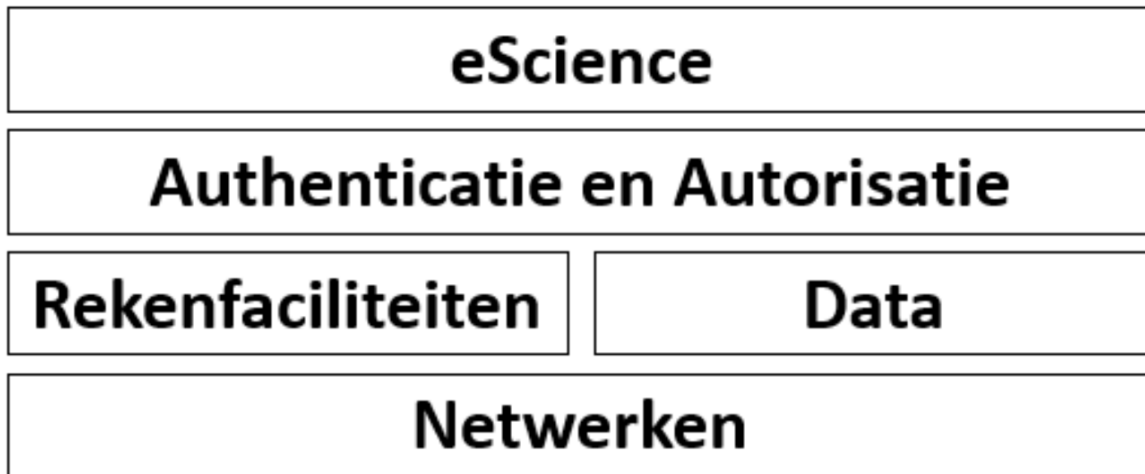


Some conclusions

- **Data and compute intensive research questions**
 - **Typically: optimization problems (parameter estimation), very large data sets, distributed data sets,....**
 - **Similar problems in all of research**
 - **Generic eScience methods: efficient computing (distributed, accelareted, orchestrated), data management (distributed databases etc.), data analytics (machine learning, deep learning, distributed learning, visual analytics)**
- eScience expertise and support & Research Software Directory (eStep)**



National e-Infrastructure



12 October 2017 / Amsterdam ArenA
National eScience Symposium



Science in a Digital World

Keynotes

Cecilia Aragon (University of Washington)
Diederik Jekel (Science Journalist)

Themes / Partners

Future of Machine Learning / [Commit2Data](#)
Internet of Things / [SURFnet](#)
Brain, Cognition & Behavior / [NeuroLabNL](#)
Energy Science / [NWO-Shell programme](#)
Natural Language / [CLARIAH](#)

Young eScientist Award 2017

Win € 50.000 worth of expertise for
a novel eScience idea!

More information & registration

www.eScienceCenter.nl/event/nlesc17

netherlands **eScience** center

