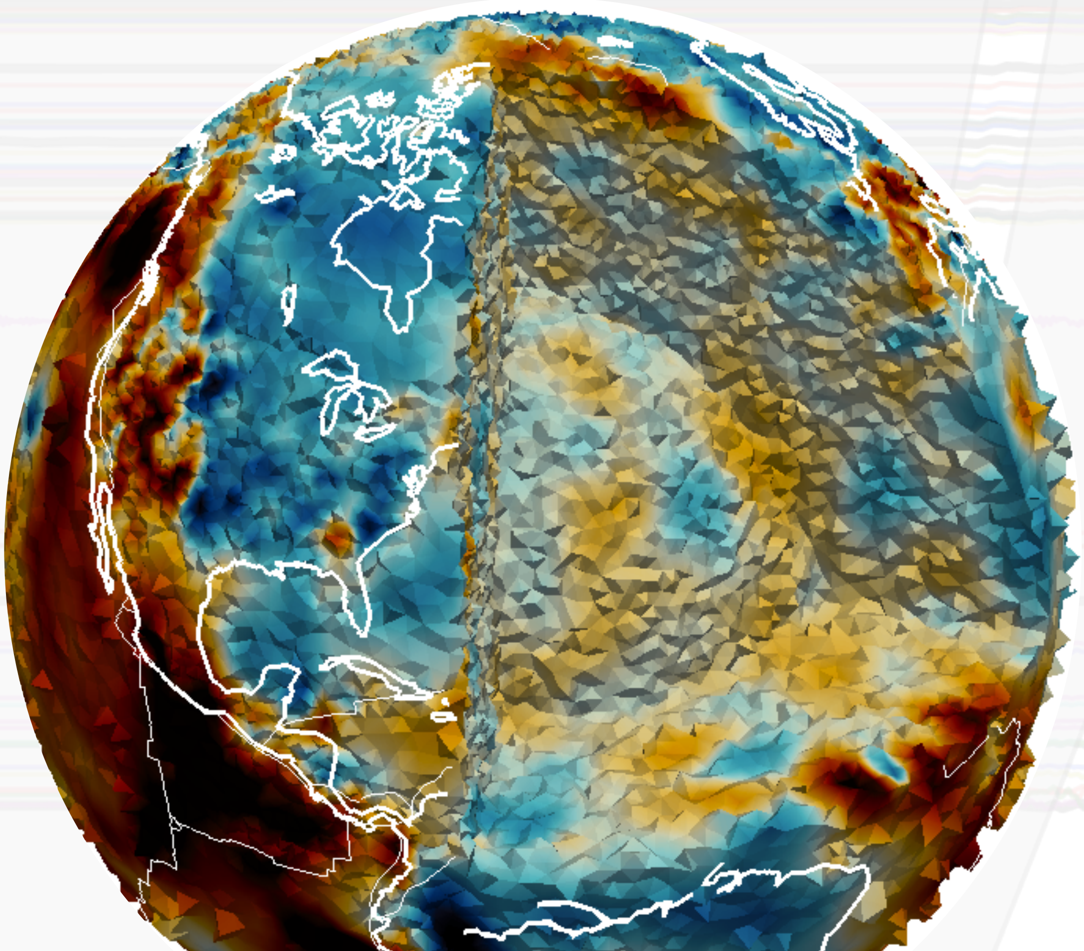
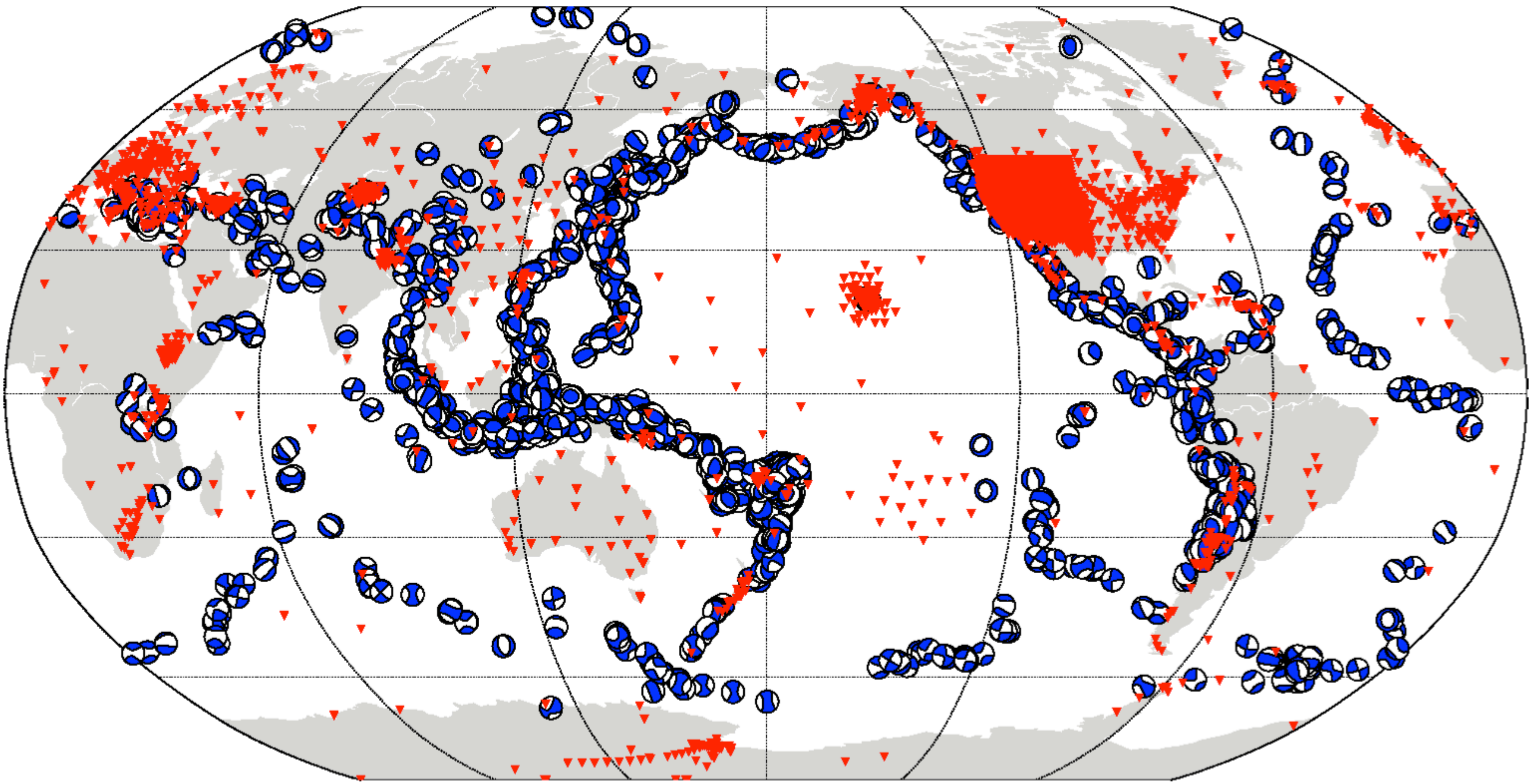
Seismic tomography: a geophysical inverse problem featuring $O(10^6)$ observations and unknowns, plus massive volumes of modelled data.



Karin Sigloch

University of Oxford

On their way from earthquakes to seismological stations, seismic waves sample the earth's interior. The 3-D structure of the interior can be inferred if enough wave paths cross at depth.

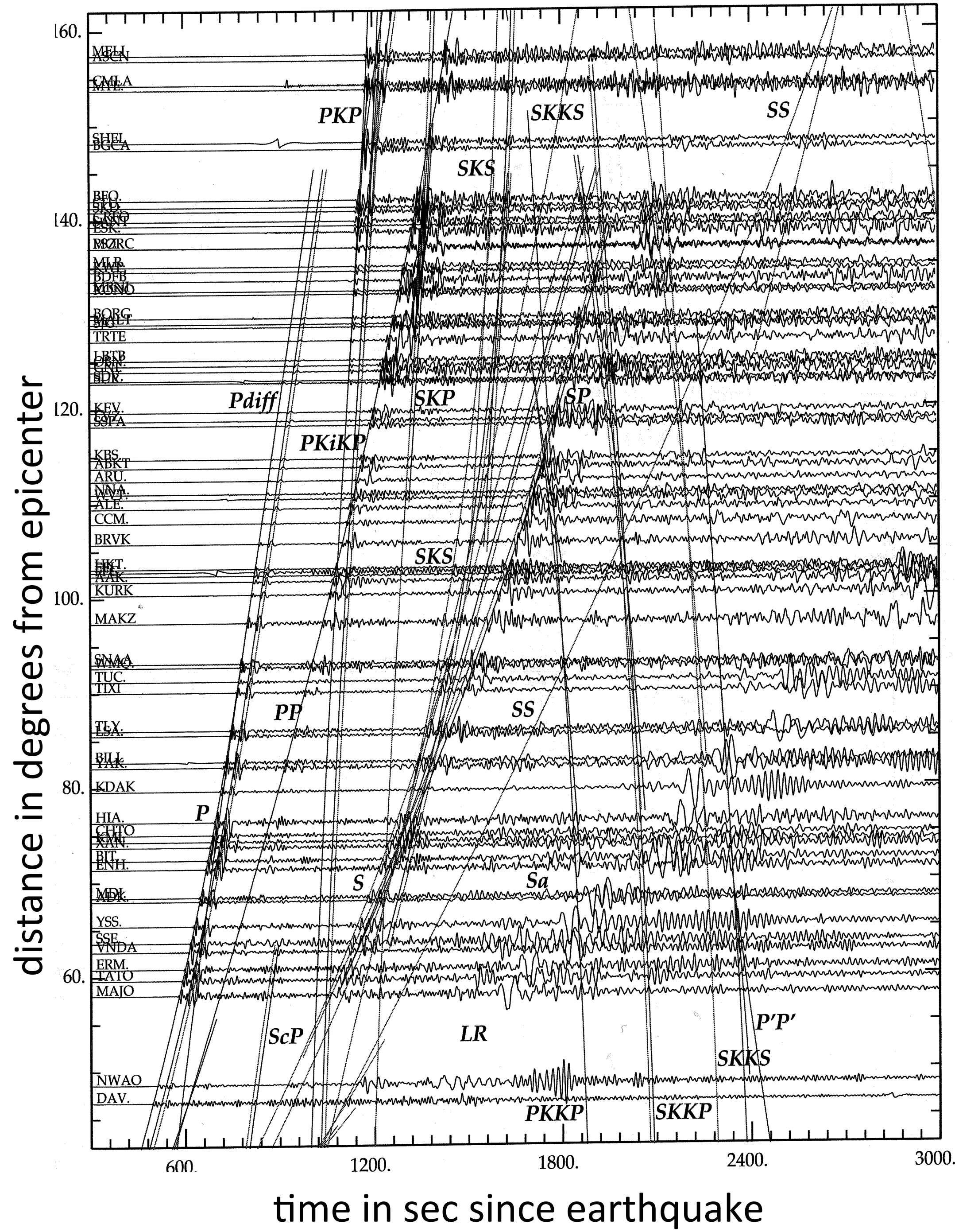


Blue: moderate to large earthquakes from 1999-2010.
Red: seismic stations that recorded them.

Large earthquakes can be measured anywhere on earth.

Data example: earthquake of magnitude 6.8 in Vanuatu, recorded by seismological stations around the world.

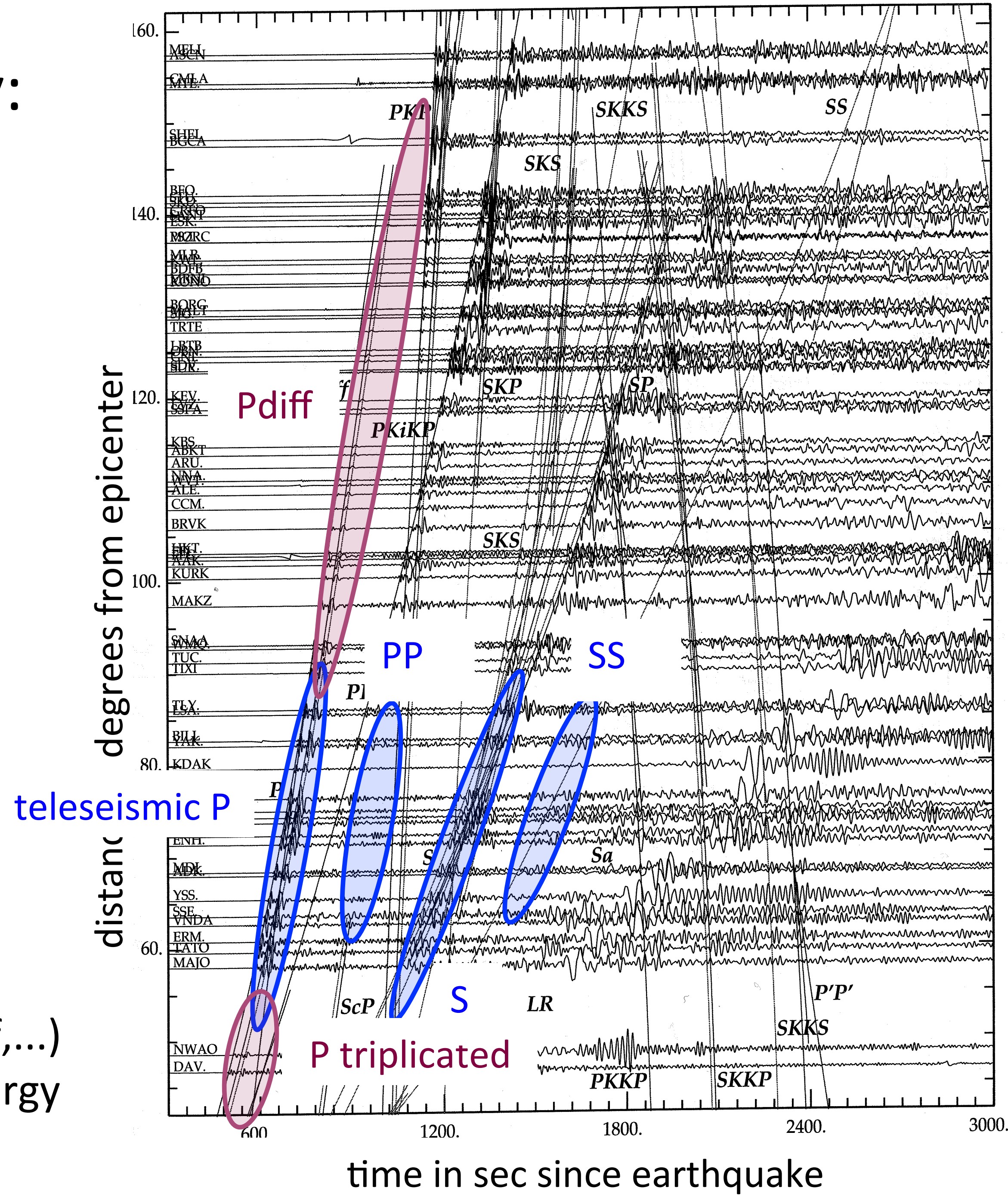
Modern broadband seismometer



Seismic tomography:

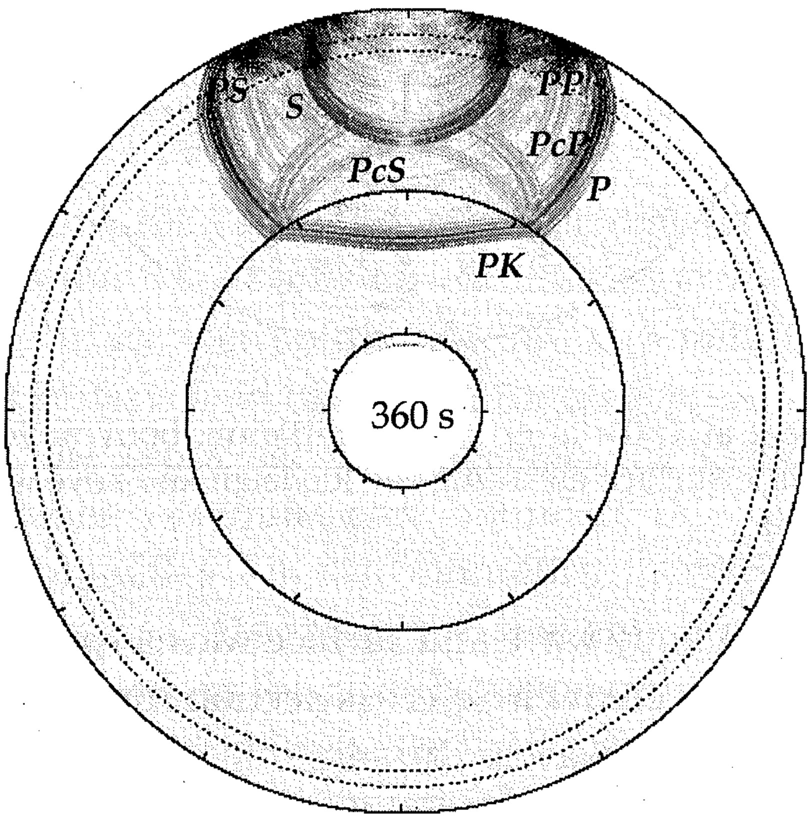
Invert for the subsurface structure that produces this observable surface wave field. Two parts:

1) Radially symmetric structure.
2) 3D deviations from radially symmetric.

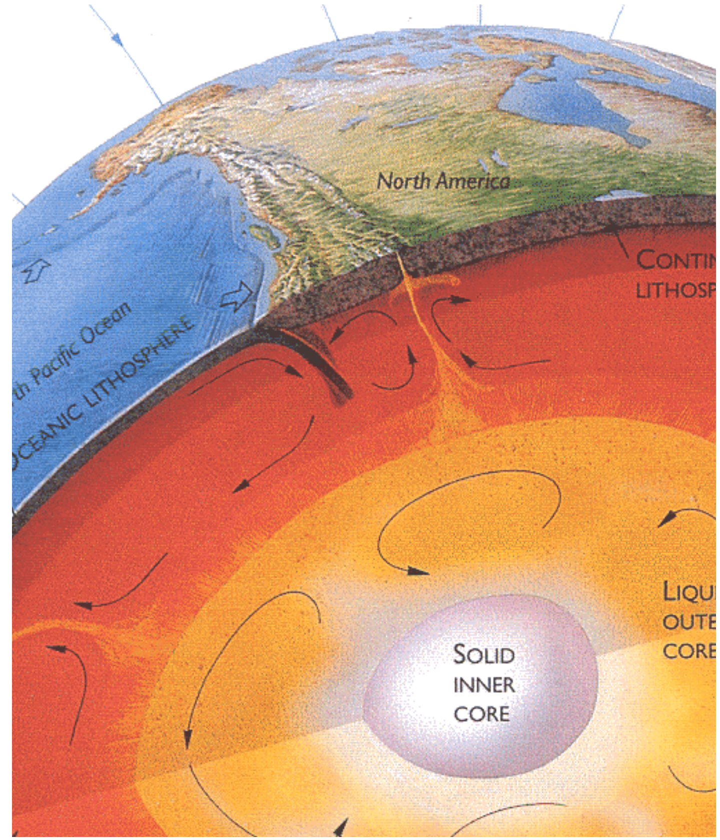


Arrivals in **phases** (P, PP, Pdiff,...) = episodic pulses of wave energy

Seismic waves sample an almost spherically symmetric planet



Simulated wave propagation, 6 minutes after an earthquake at the North Pole. Spherically symmetric models are very decent approximations…



…but we are interested in the weak lateral deviations that occur in reality. Elastic moduli and wave velocities vary by a few percent.

Inversion for weak 3D heterogeneities → a linearizable inverse problem

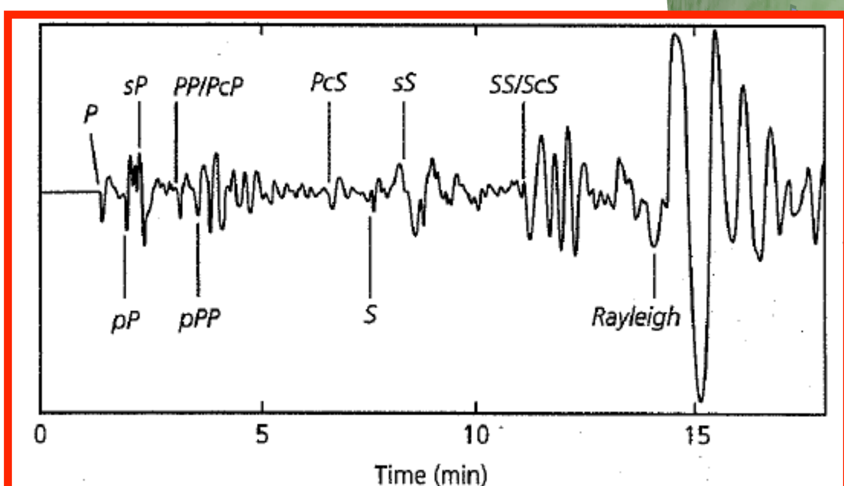
Wavefield ~15 min after earthquake at North Pole

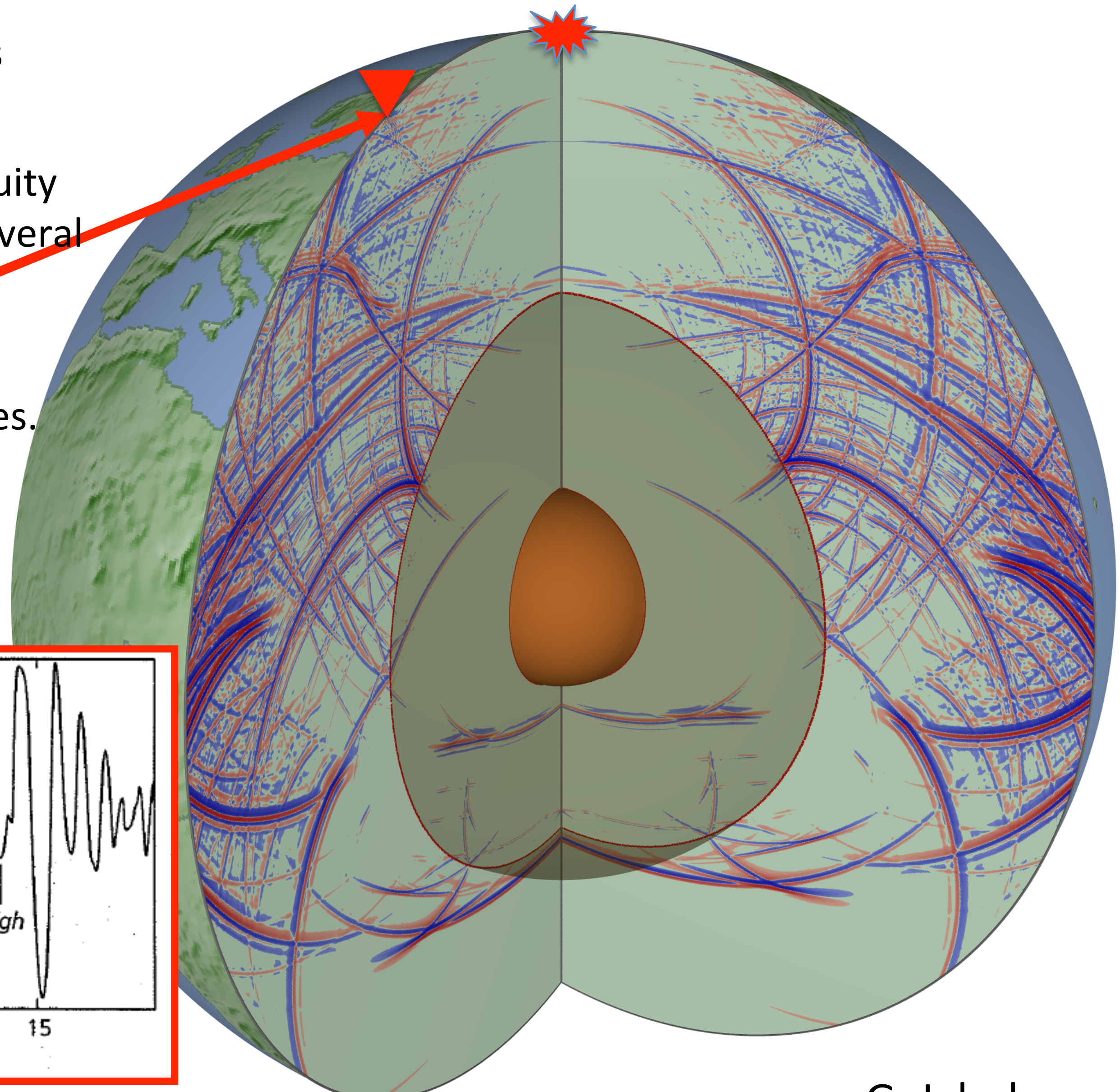
Energy travels in wave packets ("seismic phases").

When a phase hits a discontinuity (e.g. the surface), it spawns several more phases.

By now, a station here has recorded many different phases.

What can be inferred about earth structure?





G. Jahnke

The inverse problem: How do traveltime measurements (for phases P, PP,…) sense the structure of the mantle?

$$d_i = \int_V K_i(r) m(r) d^3 r$$

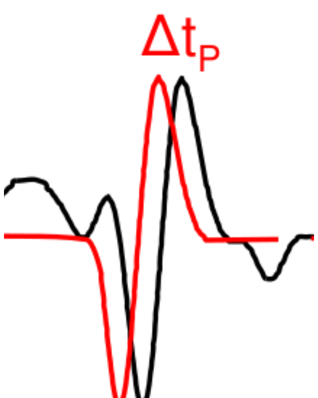
traveltime measurement

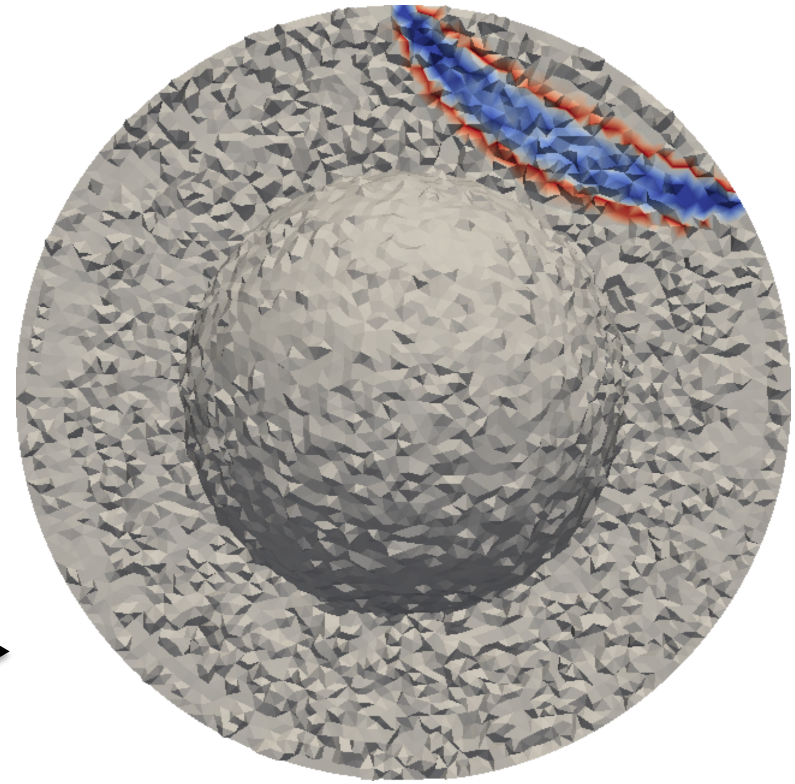
$$\frac{\Delta t_i}{\sigma_{ti}}$$

velocity structure

$$\frac{\Delta v / v}{\sigma_v}$$

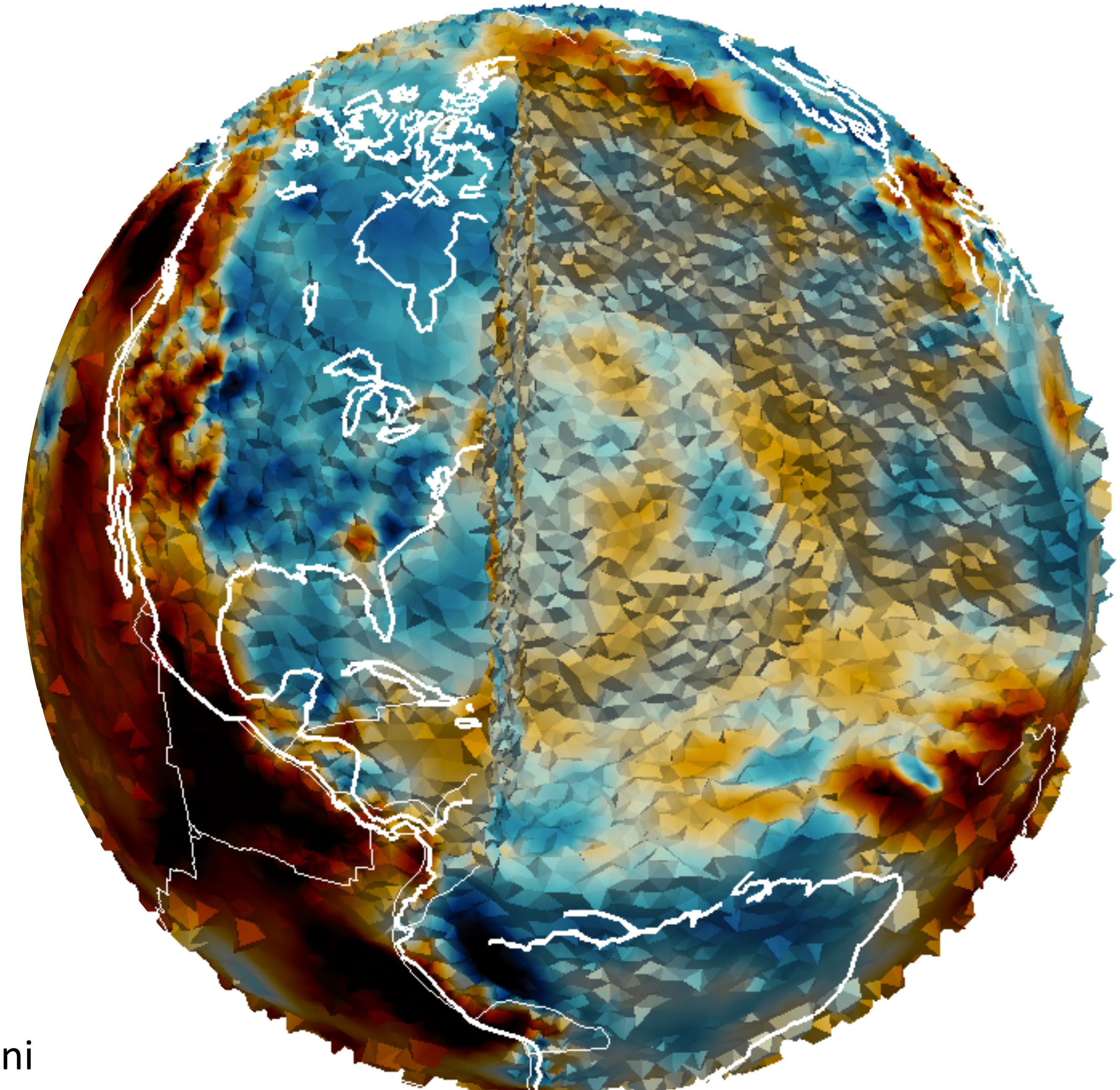
sensitivity kernel





Sensitivity K_i of a P-wave traveltime measurement d_i to velocity variations dv/v in the earth's mantle.

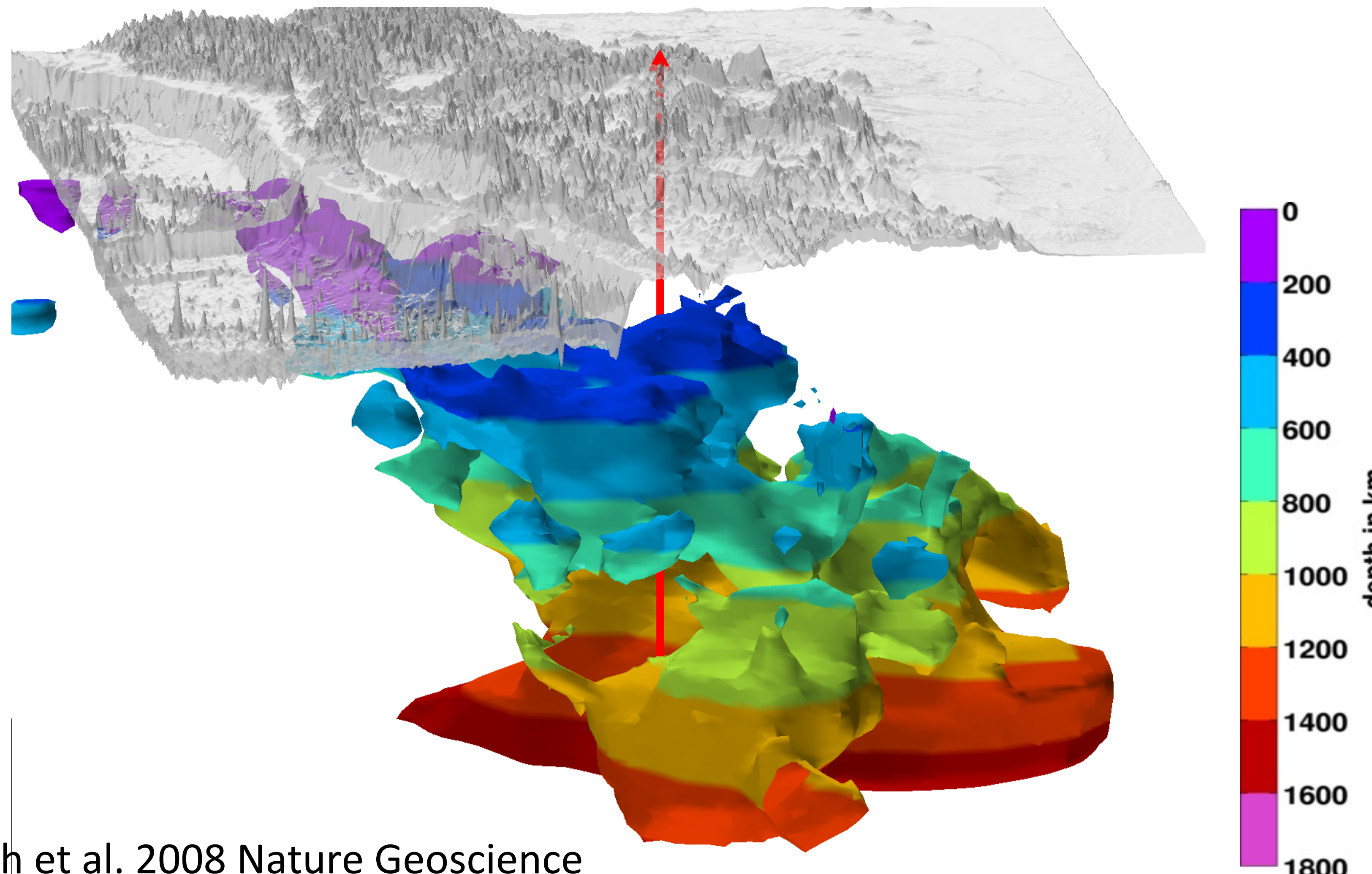
Inversion result: 3D variations in seismic wave velocity (a proxy for mantle temperature)



dv/v is on the order of a few percent

PhD work
Kasra Hosseini

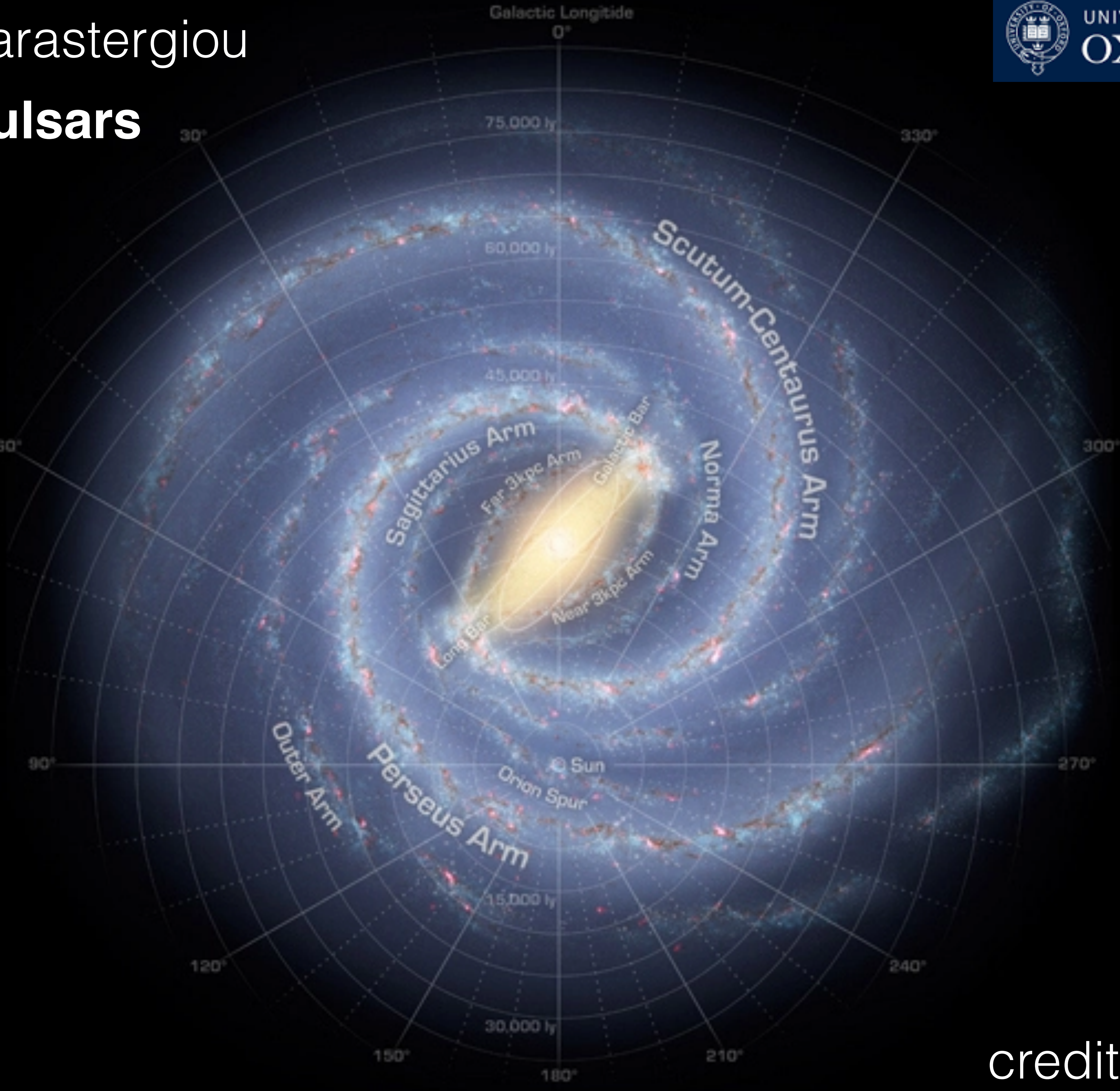
Example: 3-D mapping of seismically fast domains in the mantle. Piles of ancient seafloor (lithosphere) that have been sinking back into the mantle over the past 100-200 million years.



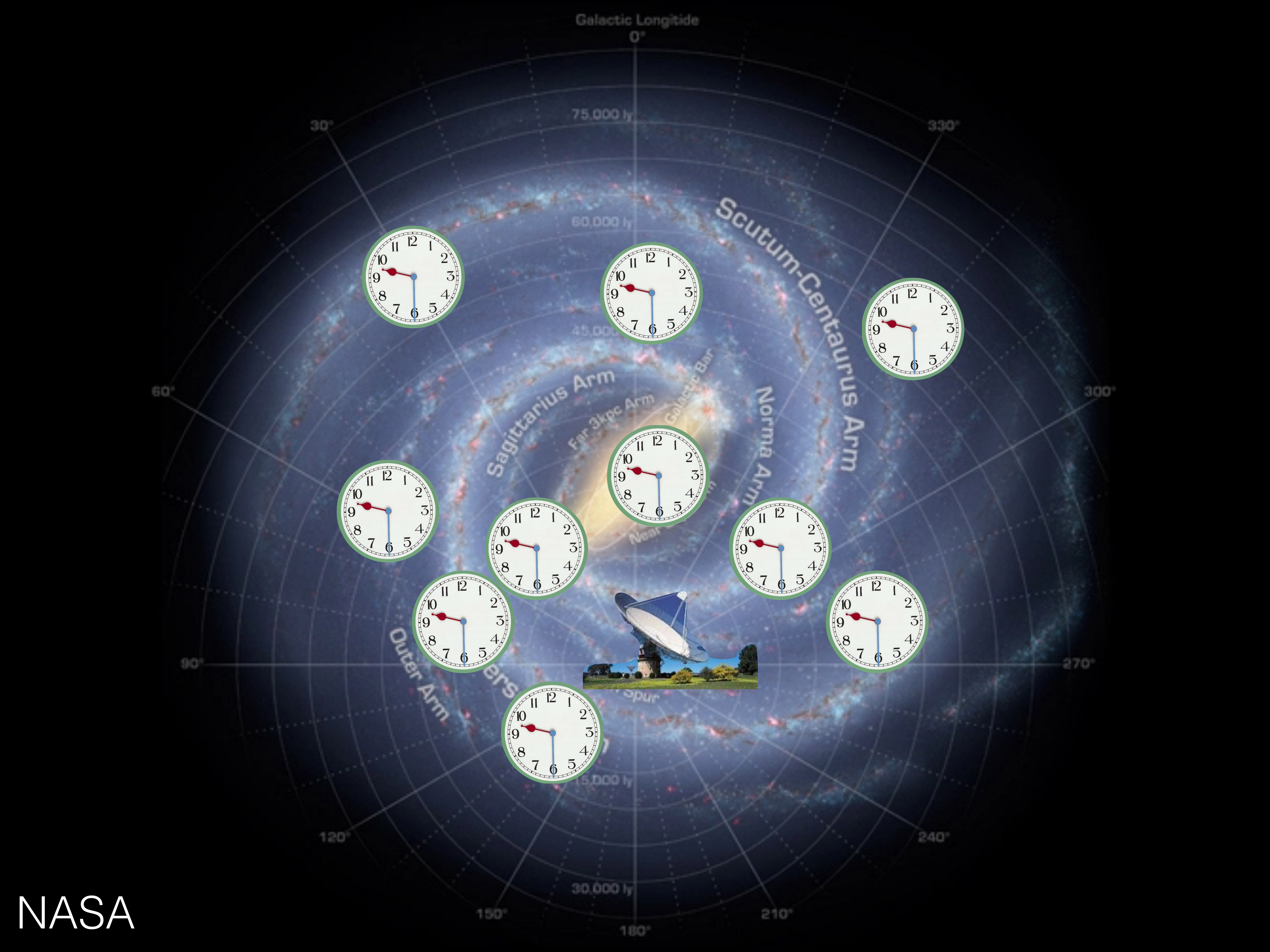
Sigloch et al. 2008 Nature Geoscience

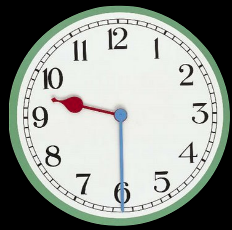
Aris Karastergiou

Pulsars

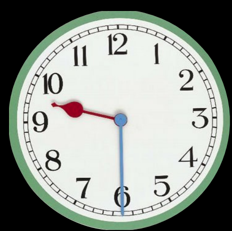


credit: NASA

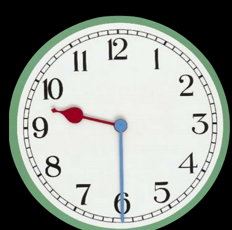
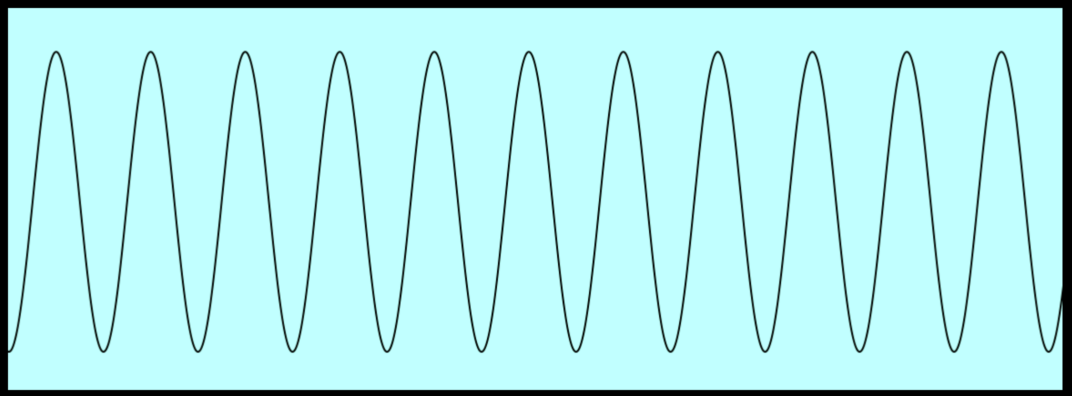

Galactic Longitide
0°
30°
330°
60°
300°
90°
270°
120°
240°
150°
210°
180°
75,000 ly
60,000 ly
30,000 ly
Scutum-Centaurus Arm
Sagittarius Arm
Far 3kpc Arm
Galactic Bar
Norma Arm
Outer Arm
NASA



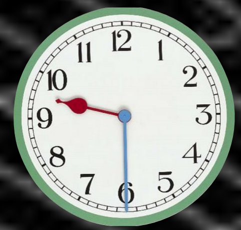

Rate of ticking tells us something about the clock




Ticks from a moving and ticking clock will be Doppler shifted

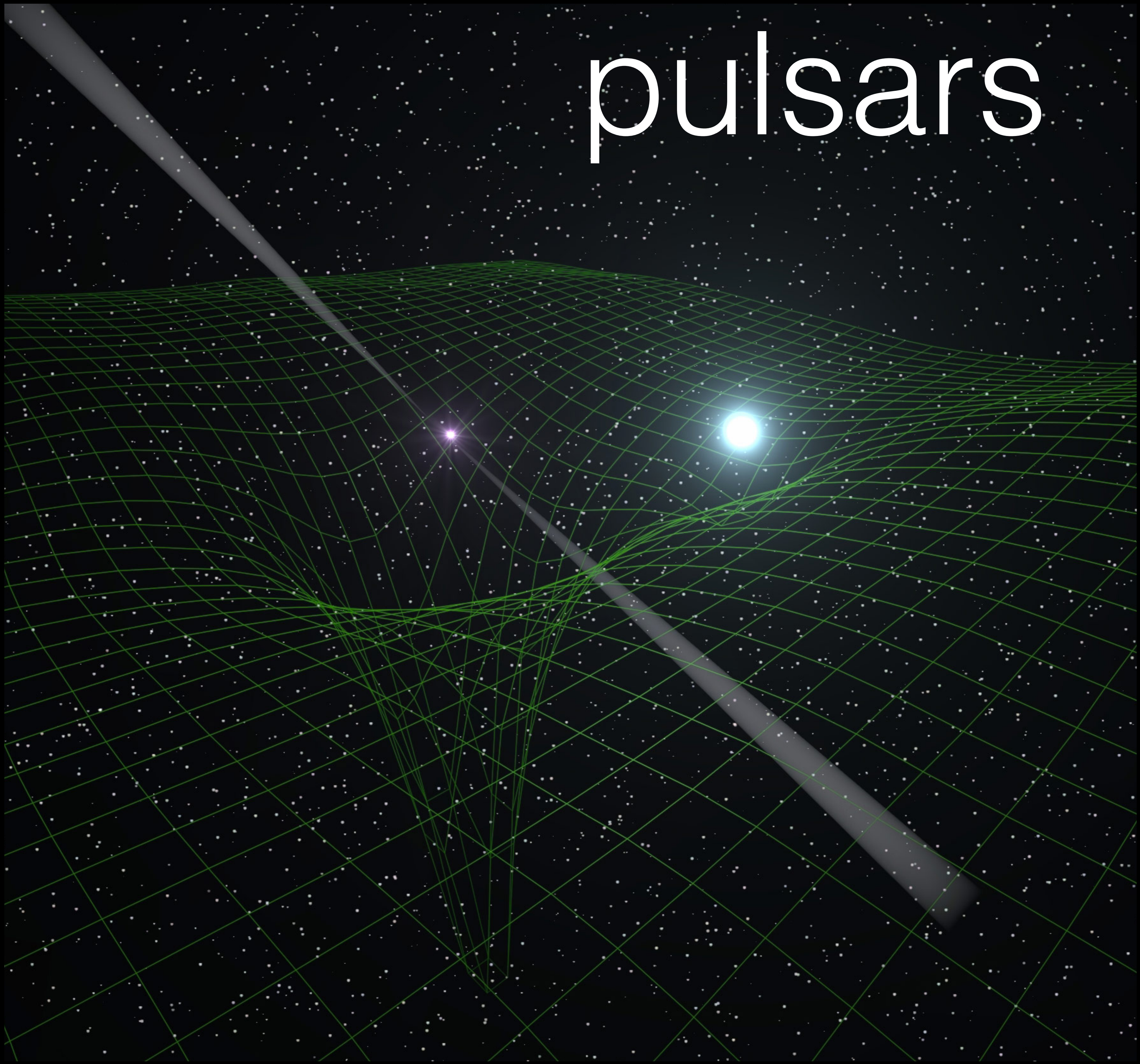




Radio waves will be delayed by dispersion effects in the interstellar medium; larger distance = larger delays



Clocks moving in a strong field of gravity will send out ticks delayed by the field

Clocks moving in a strong field of gravity will have strange orbits



Antoniadis et al.

main properties of pulsars

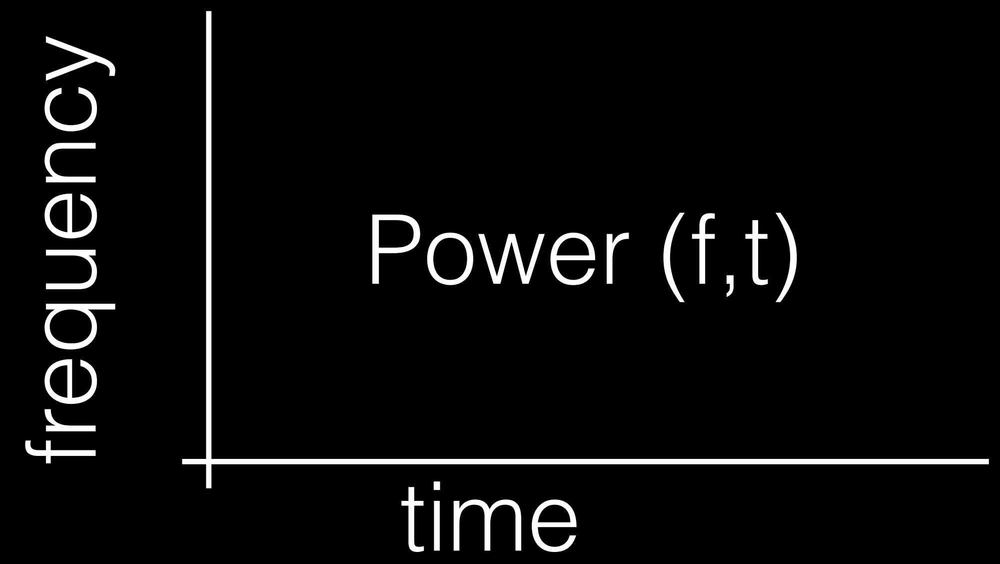
- 1.2 to 2 Solar masses of dense nuclear matter, spinning with periods between milliseconds and several seconds
- diameter of a medium sized town
- superfluid interior surrounded by crust
- super-strong surface magnetic field 10^{12} Gauss
- co-rotating charged magnetosphere & light cylinder
- high energy streaming plasma
- radio, optical, Gamma-ray, X-ray emission
- Gravitational wave radiation

Observing pulsars

Form beams from an array
or use single dish

point your telescope

Channelize your signal for
dedispersion and RFI excision



dedisperse

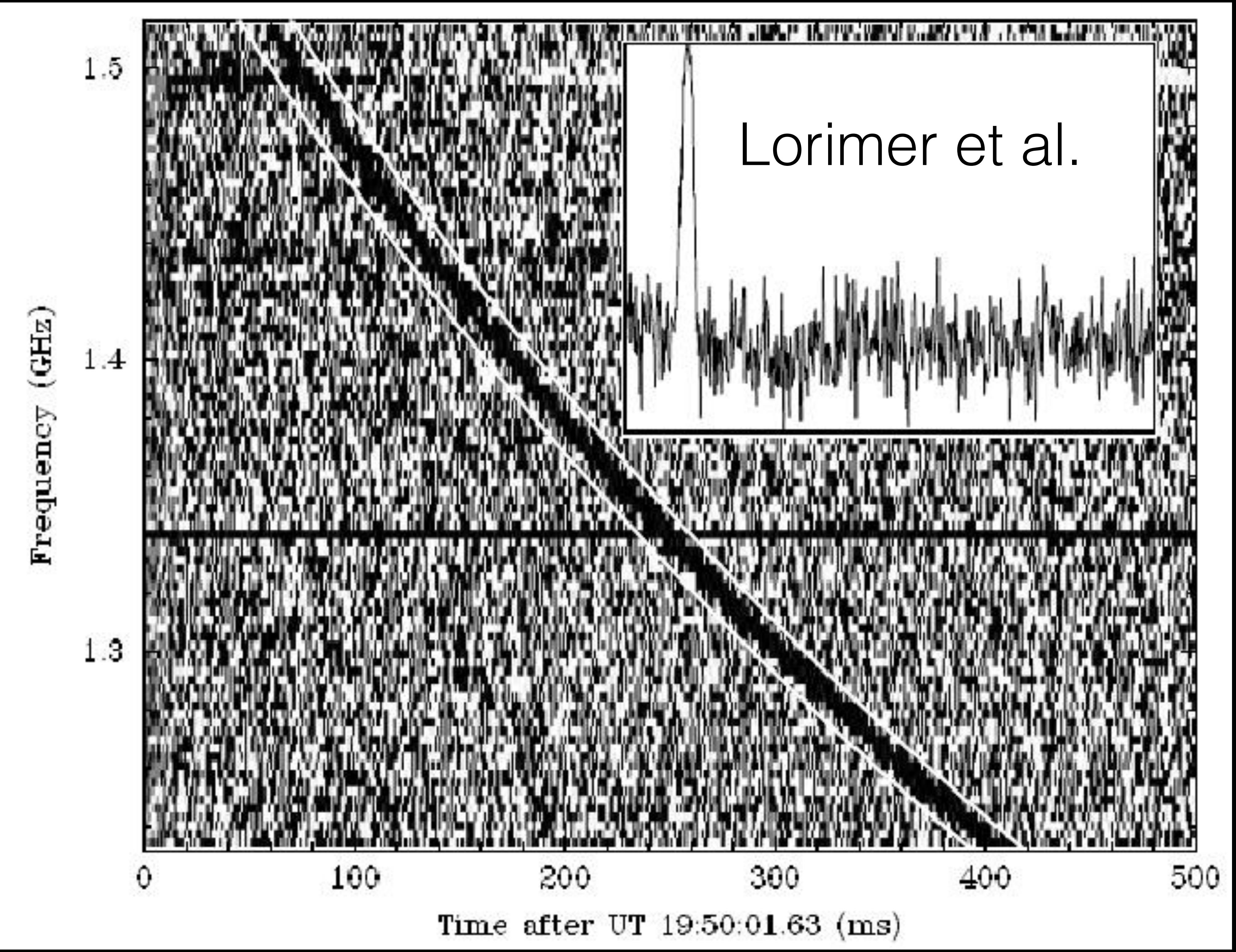
apply differential frequency delay

fold

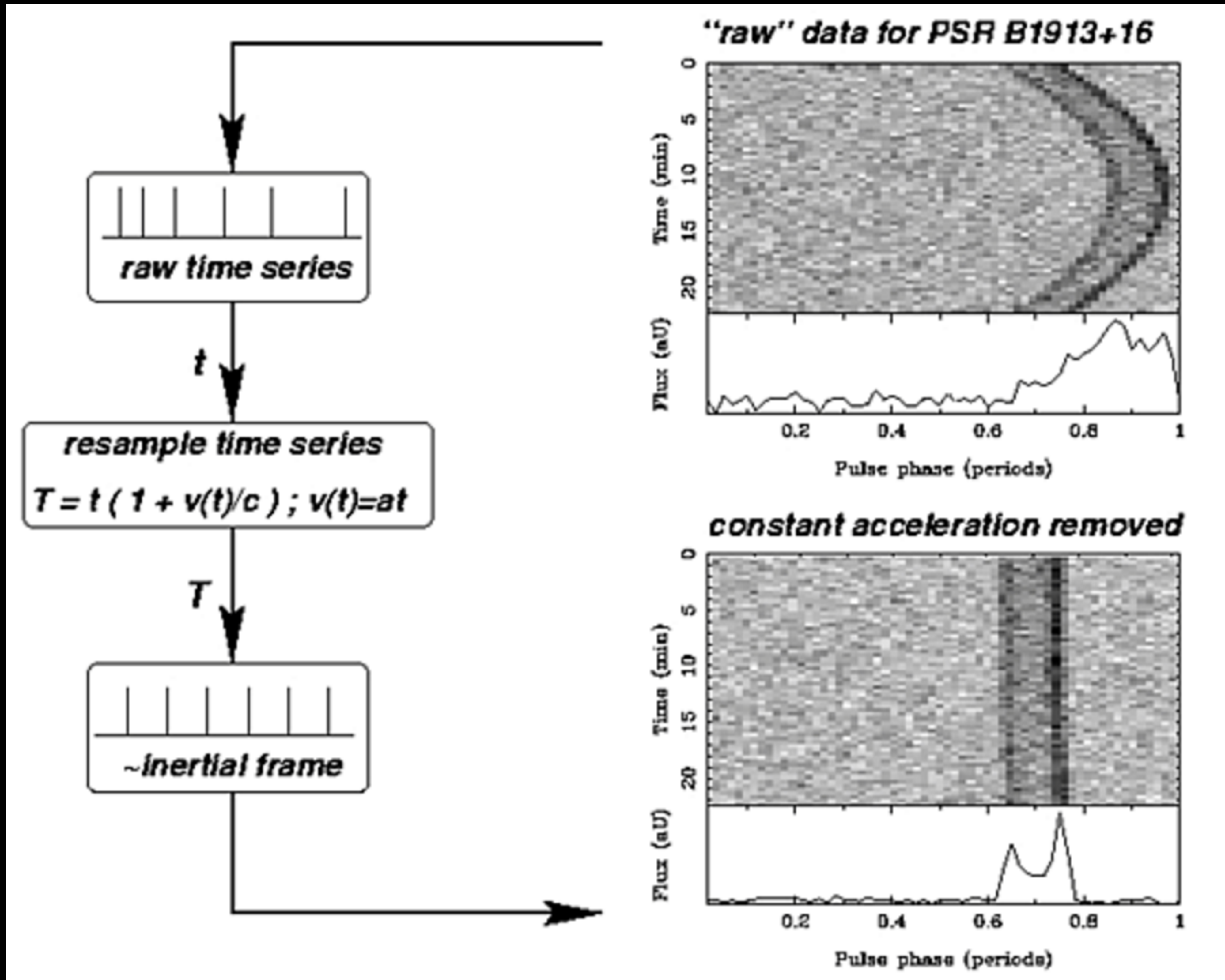
Fold at known pulsar period
or search for period

data rates

- Typical observations take place around 1 GHz or ~30 cm
- Pulsar signals are bright across the 30 MHz to 5 GHz range
- Radio telescopes in this range typically sample a bandwidth of 0.5 the observing frequency
- 500 MHz of bandwidth require a sampling rate of 1 Gsample/s
- For interferometry, multiply that rate by the number of tied-array beams
- Example: 1000 tied array beams produce 1 Tsample/s or 1-4 TB/s for 8 to 32-bit sampled data
- Data rate reduction by choosing appropriate time-frequency resolution
- Data rate reduction by integration, after considering the physical properties of the observed signal, depending on objectives



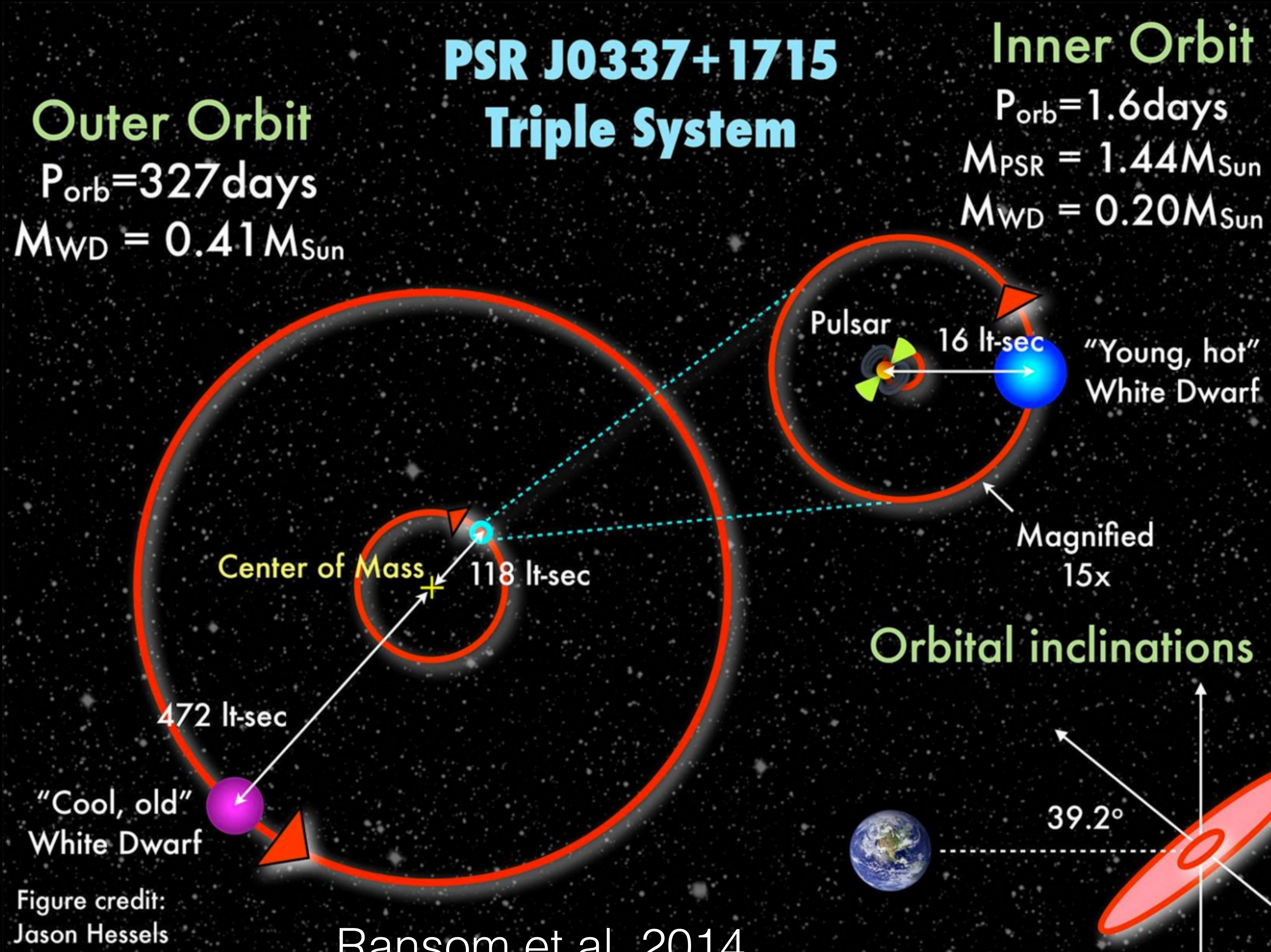
dedispersion



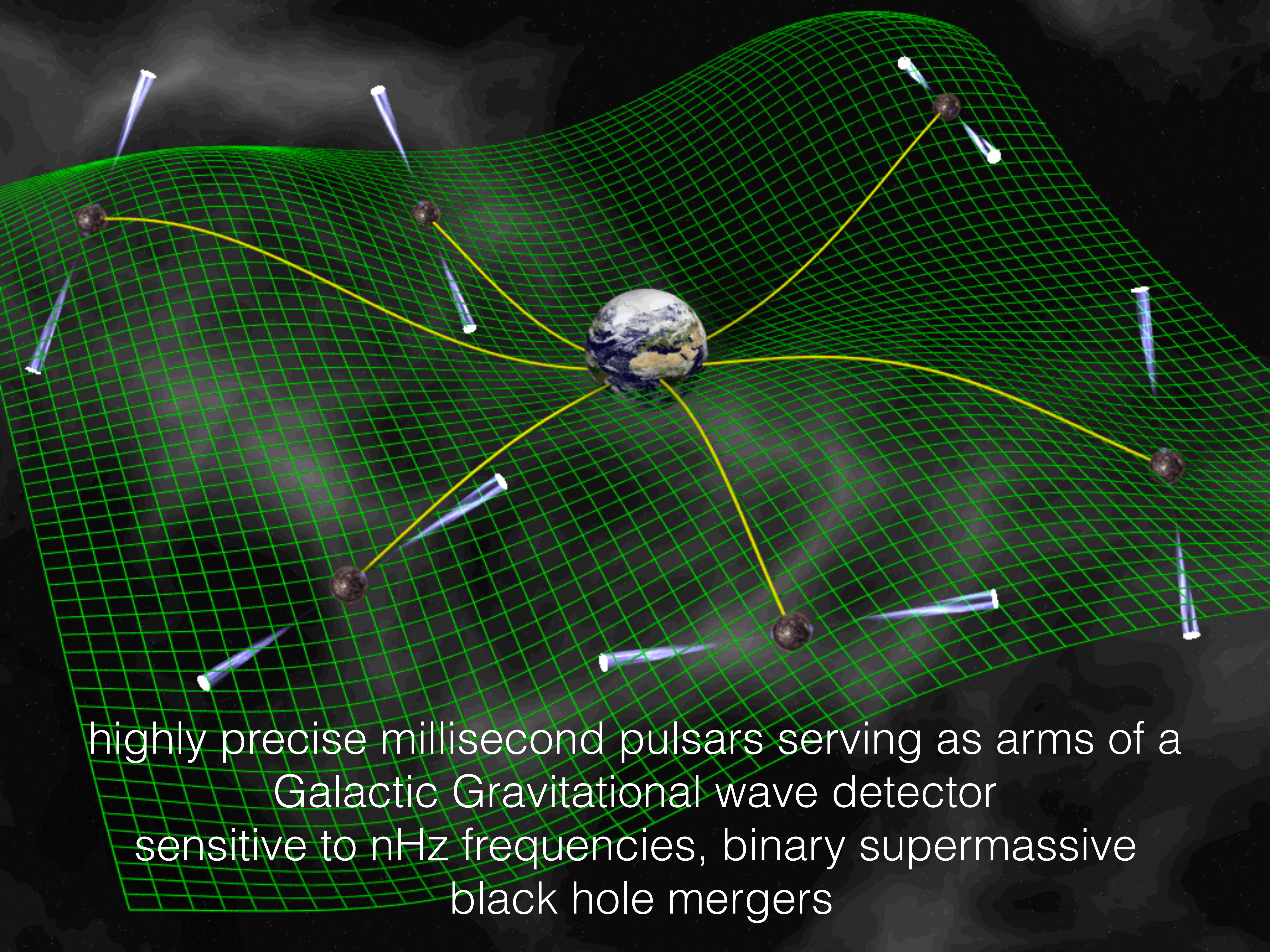
folding

Jodrell Bank Centre for Astrophysics

Visit https://www.skatelescope.org/ for details



Ransom et al. 2014



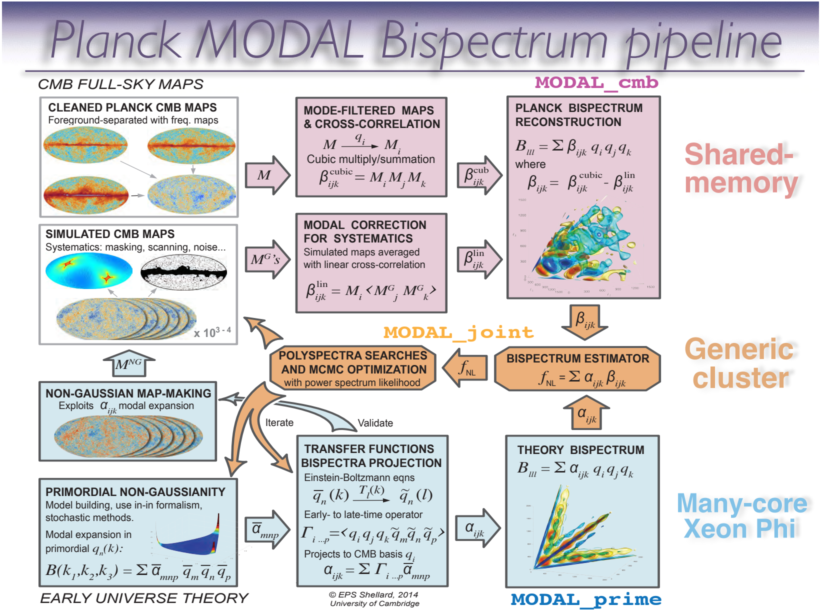
highly precise millisecond pulsars serving as arms of a
Galactic Gravitational wave detector
sensitive to nHz frequencies, binary supermassive
black hole mergers

Algorithms to Architectures

Juha Jäykkä (jj411@cam.ac.uk)
University of Cambridge, UK
London, 2016-01-13

COSMOS Intel Parallel Computing Centre

- optimise, modernise, design, make sustainable research software
- intimate link between hardware, algorithms, software
- excellent link with Intel, collaborating in all aspects of software development and two-way exchange of ideas and knowledge
- actively engaged with vendors (mostly SGI, Intel) designing new systems (e.g. the MG Blade)
- 3.5 (research) software engineers working as part of research groups
- awarded the HPCwire Readers' Choice Award (Best High Performance Data Analytic) 2015
  [J.P.Briggs et al in "High Performance Parallelism Pearls", vol 2, 171-190, Morgan Kaufman, Boston, 2015; arXiv:1503.08809]


Planck MODAL Bispectrum pipeline
CMB FULL-SKY MAPS
MODAL_cmb
CLEANED PLANCK CMB MAPS
Foreground-separated with freq. maps
MODE-FILTERED MAPS & CROSS-CORRELATION
$M \xrightarrow{q_i} M_i$
Cubic multiply/summation
$\beta^{\text{cubic}}_{ijk} = M_i M_j M_k$
PLANCK BISPECTRUM RECONSTRUCTION
$B_{lll} = \Sigma\, \beta_{ijk}\, q_i q_j q_k$
where
$\beta_{ijk} = \beta^{\text{cubic}}_{ijk} - \beta^{\text{lin}}_{ijk}$
Shared-memory
M
β^{cub}_{ijk}
SIMULATED CMB MAPS
Systematics: masking, scanning, noise...
x 10^{3-4}
M^{G}'s
MODAL CORRECTION FOR SYSTEMATICS
Simulated maps averaged with linear cross-correlation
$\beta^{\text{lin}}_{ijk} = M_i \langle M^G_j M^G_k \rangle$
β^{lin}_{ijk}
β_{ijk}
MODAL_joint
M^{NG}
POLYSPECTRA SEARCHES AND MCMC OPTIMIZATION
with power spectrum likelihood
f_{NL}
BISPECTRUM ESTIMATOR
$f_{NL} = \Sigma\, \alpha_{ijk}\, \beta_{ijk}$
Generic cluster
NON-GAUSSIAN MAP-MAKING
Exploits α_{ijk} modal expansion
Iterate
Validate
α_{ijk}
PRIMORDIAL NON-GAUSSIANITY
Model building, use in-in formalism, stochastic methods.
Modal expansion in primordial $q_n(k)$:
$B(k_1,k_2,k_3) = \Sigma\, \bar{\alpha}_{mnp}\, \bar{q}_m \bar{q}_n \bar{q}_p$
$\bar{\alpha}_{mnp}$
TRANSFER FUNCTIONS BISPECTRA PROJECTION
Einstein-Boltzmann eqns
$\bar{q}_n(k) \xrightarrow{T_l(k)} \tilde{q}_n(l)$
Early- to late-time operator
$\Gamma_{i\,..p} = \langle q_i q_j q_k \tilde{q}_m \tilde{q}_n \tilde{q}_p \rangle$
Projects to CMB basis q_i
$\alpha_{ijk} = \Sigma\, \Gamma_{i\,..p}\, \bar{\alpha}_{mnp}$
α_{ijk}
THEORY BISPECTRUM
$B_{lll} = \Sigma\, \alpha_{ijk}\, q_i q_j q_k$
Many-core Xeon Phi
EARLY UNIVERSE THEORY
© EPS Shellard, 2014
University of Cambridge
MODAL_prime

Confrontation of Observation and Theory

Cosmology

- cosmic microwave background (Planck), large scale structure (DES)
- gravitational waves (aLIGO), cosmic defects
- also needs some traditional HPC for solving PDEs numerically
- all either create and process or just process very big datasets

Astrophysics

- solar atmosphere and interior simulations
- investigations of evolution of protoplanetary disks

In all this data sizes are big enough that

- movement of data must be either avoided altogether or carefully orchestrated ($\Rightarrow$ tightly-coupled heterogeneous systems)
- processing efficiency is paramount: the work that led to the HPCwire award, provided an increase of two orders of magnitude
  - importantly, one order of mag from choosing the appropriate Algorithm to Architecture

Future Challenges

Computing Hardware Becoming Harder to Use Efficiently

- challenging nested parallelism (vector, threads, processes)
- multi-level memory hierarchy: registers, caches, fast RAM (on-chip HBM), slow RAM (classic DRAM), distributed off-node RAM, etc
- widening vector units and many-core architectures
- non-SIMD CPU performance hasn't really increased since 2008 or earlier; GPU even more disruptive to codes than SIMD
- many codes have reached their scaling limits: cannot simply add nodes to increase performance either
- increasing imbalance between non-volatile IO, memory IO and GF/s

Rapidly Increasing Size of Data

- cannot really move around any more
- often needs to be post-processed (visualised) remotely (OSPRay!)
- or even on-the-fly/in-situ: throw away uninteresting data like LHC

Research Software in the Future

Co-design Hardware and Software

- Alan Turing was a co-designer (Bombe to break Enigma)
- co-design useful for both CPU, many-core, and GPU codes
- need to engage more specialised (research) software developers
- COSMOS was involved in co-designing with SGI the "MG-blade" for Intel Xeon Phi co-processors and GPUs in SGI UV2000 systems
- currently involved in the design of an advanced hybrid "co-cluster"
- helps early adoption of new hardware

Productivity of Big Data Analysis and Computing

- typical HPC software isn't the easiest to use or maintain
- easy to use tends to be inefficient (1st vs 100th solution)
- involve software engineers to combine ease of use and efficiency
- workflow management tools can address workflow inefficiencies

Algorithms to Architectures

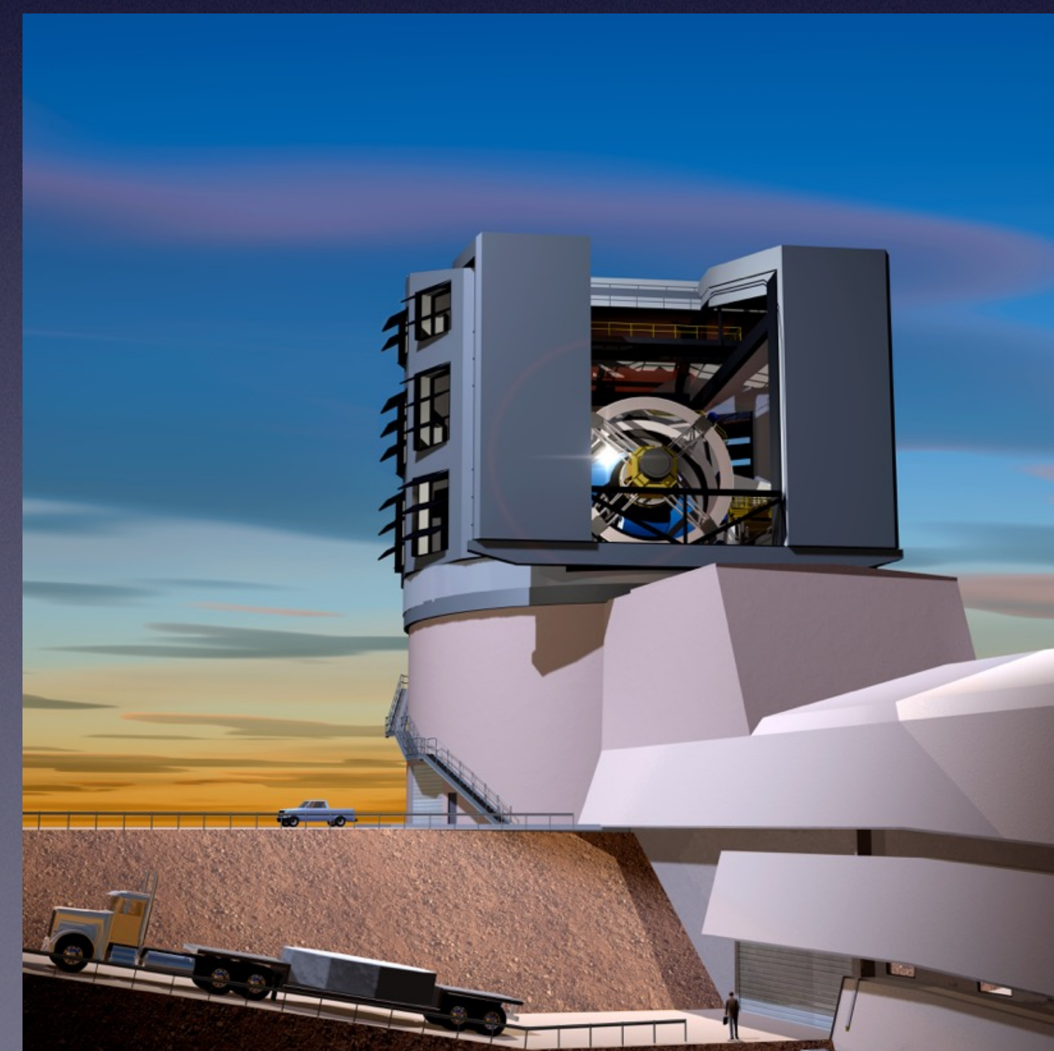
- Develop architecture-aware and architecture-specific algorithms to process Big Data and simulate faster
  - not just bigger data and bigger simulations but also present size more energy efficiently and faster (not necessarily the same thing)
- Design to preserve data locality through in-situ and on-the-fly post-processing
  - COSMOS IPCC participates in the development of Intel's OSPRay in-situ visualiser, HAM offload library, etc
- Engage with hardware vendors and co-design heterogeneous systems to ensure early adoption of next generation hardware
- Broader impact through public release of world-leading data analytic parallel software packages
  - use standards (Fortran, C, C++, OpenMP, OpenCL etc) to ensure portability
- Other external impacts
  - ISC2015 OSPRay visualisation demo with Planck data (first public demo of KNL)
  - Big Data real-time visualisation demo at SC2015 with 10TB Walls data
- Multi-disciplinary interactions essential to reach full potential

Big data problems for transient sky surveys in astronomy

S.J. Smartt, Ken Smith, Darryl Wright, D. Young (Queen's University Belfast), K. Chambers, M. Huber, E. Magnier, J. Tonry, L. Denneau, B. Stalder, A. Heinze   ++ (IfA, Hawaii)

Pan-STARRS + ATLAS (now - 2020+)

LSST : Large Synoptic Survey Telescope (2020-2030)



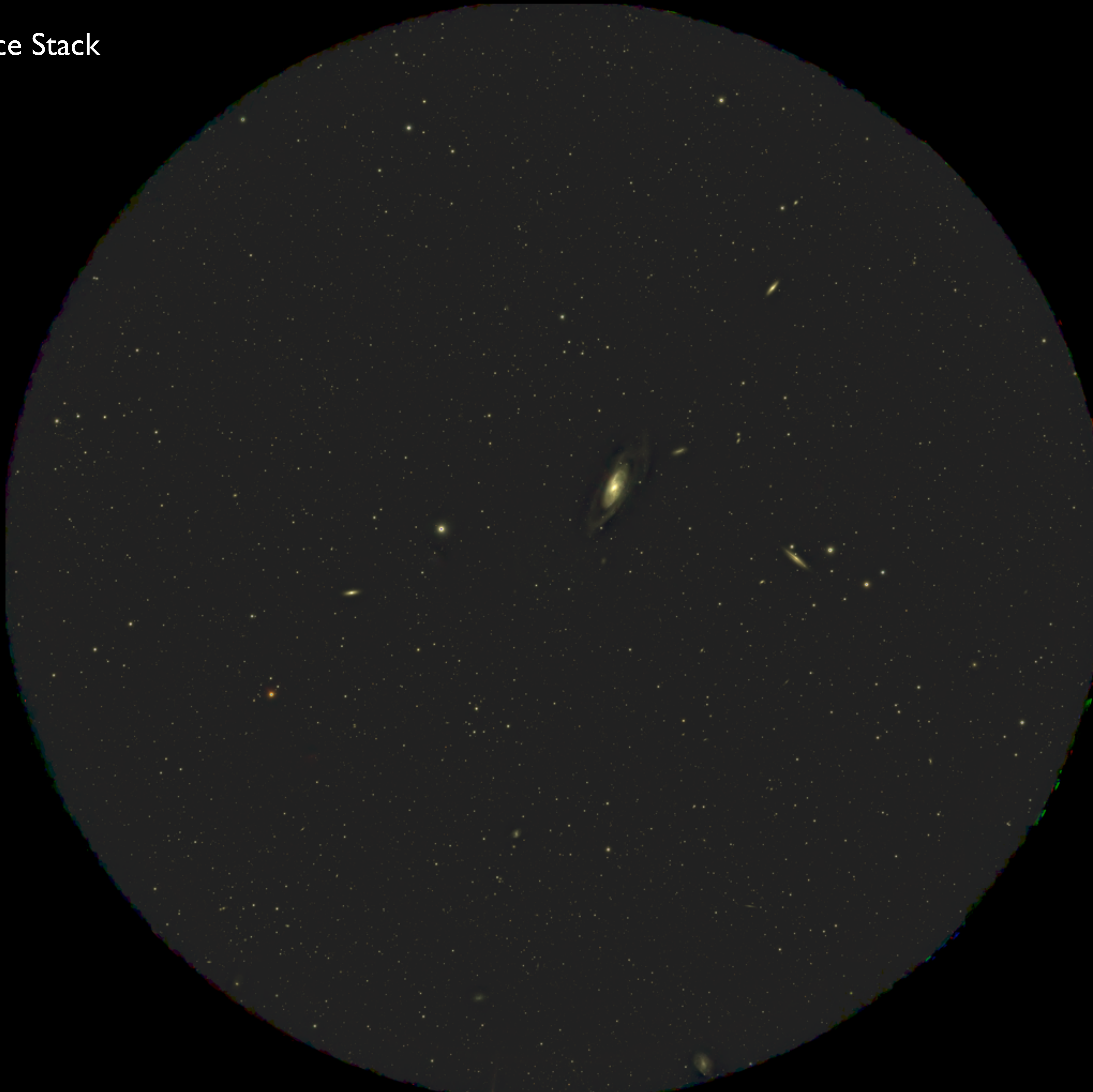
Big data problems for transient sky surveys in astronomy

1. Image recognition : real/bogus and rapid astrophysical classification (1-10TB image data per day)

2. Massive database : 1 billion objects, 10000 measurements over 5 yrs (indexing, database partitioning, database architecture)

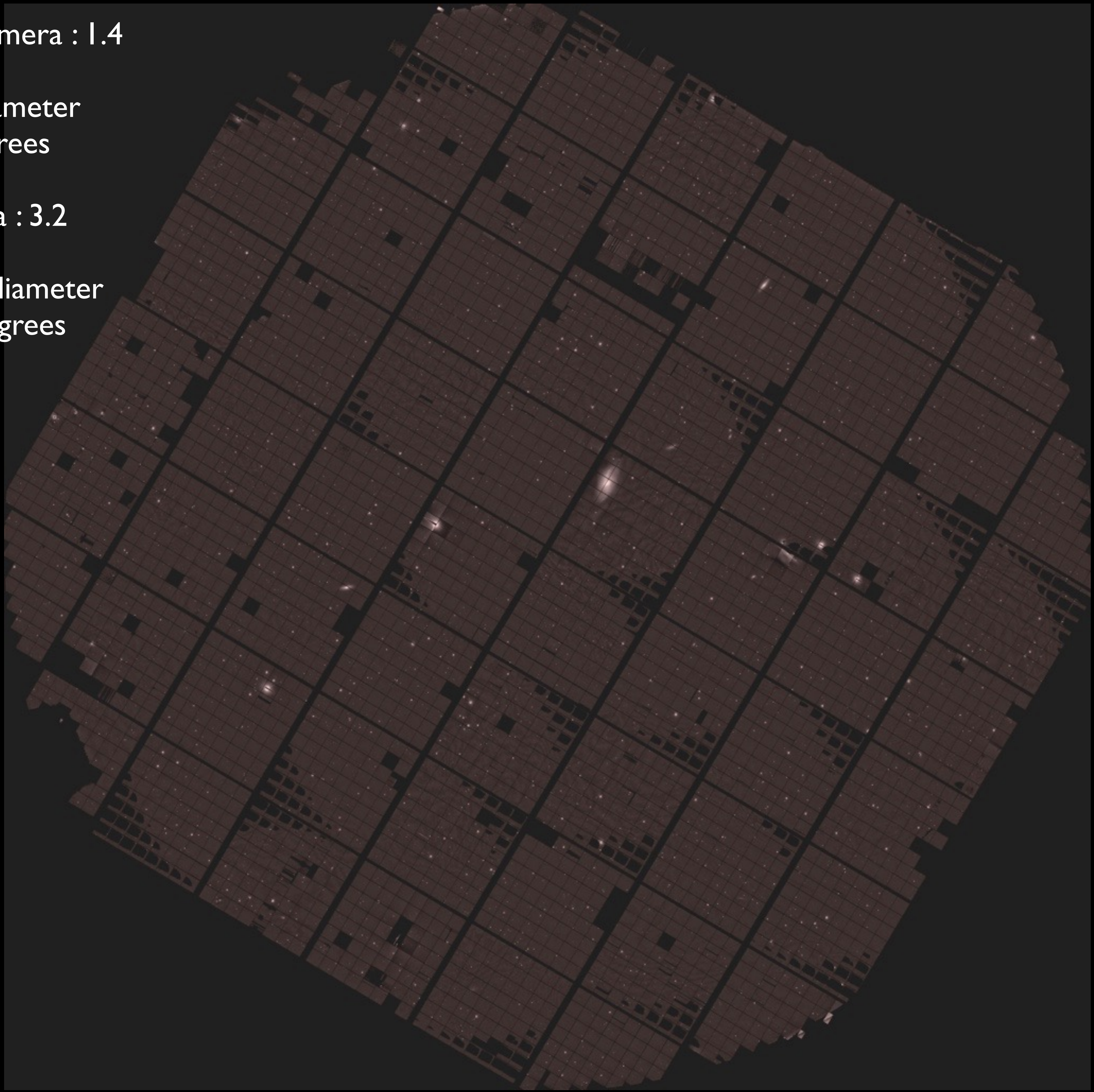
3. Turn around speed : insert 64000 per sec into database (24hr spread). Index and association

MD Reference Stack
(MD06)

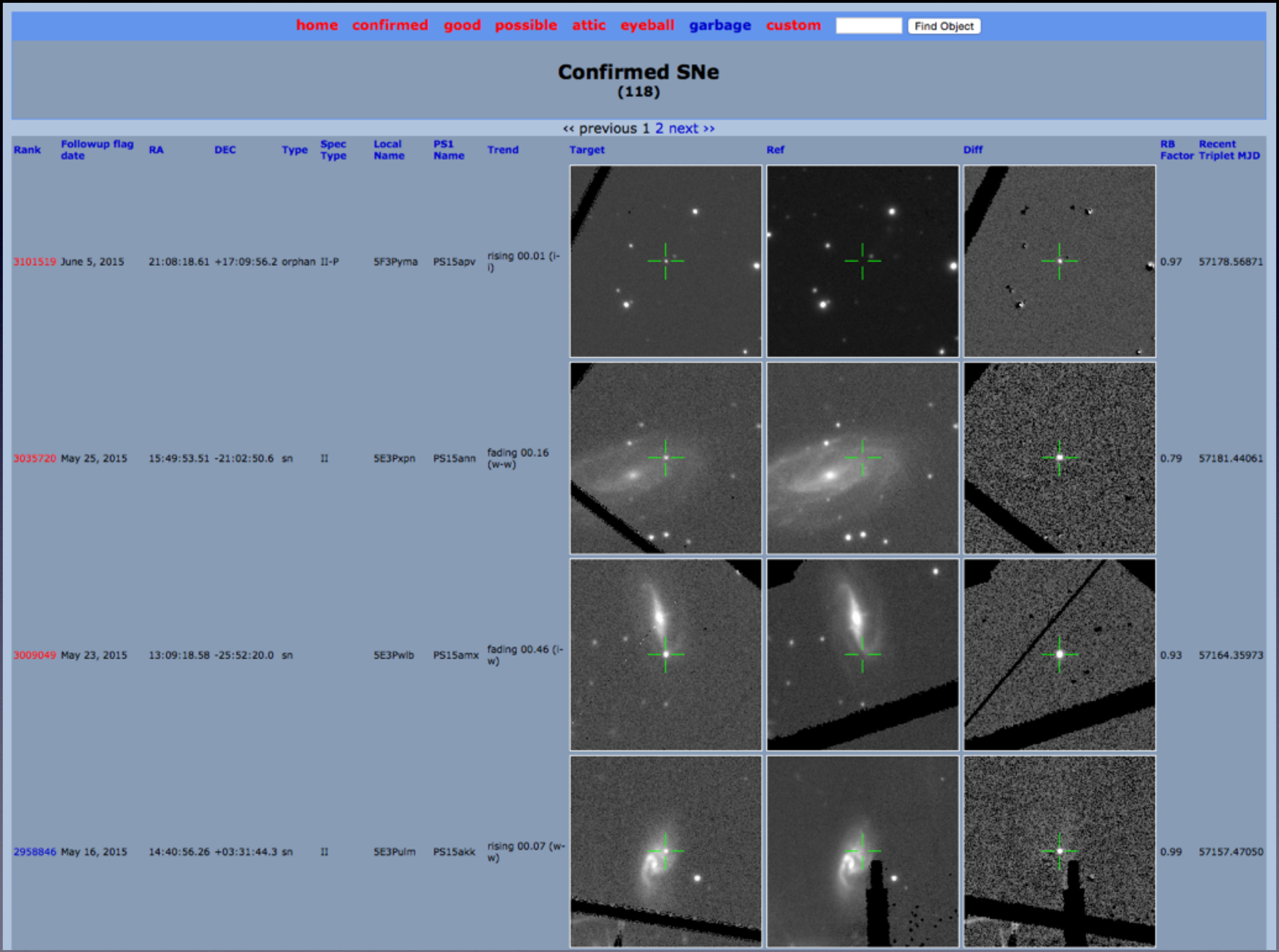


Giga-pixel camera : 1.4
gigapixels
3 degrees diameter
7 square degrees

LSST Camera : 3.2
gigapixels
3.5 degrees diameter
10 square degrees



Difference images - to find transient and variable sources



home confirmed good possible attic eyeball garbage custom Find Object

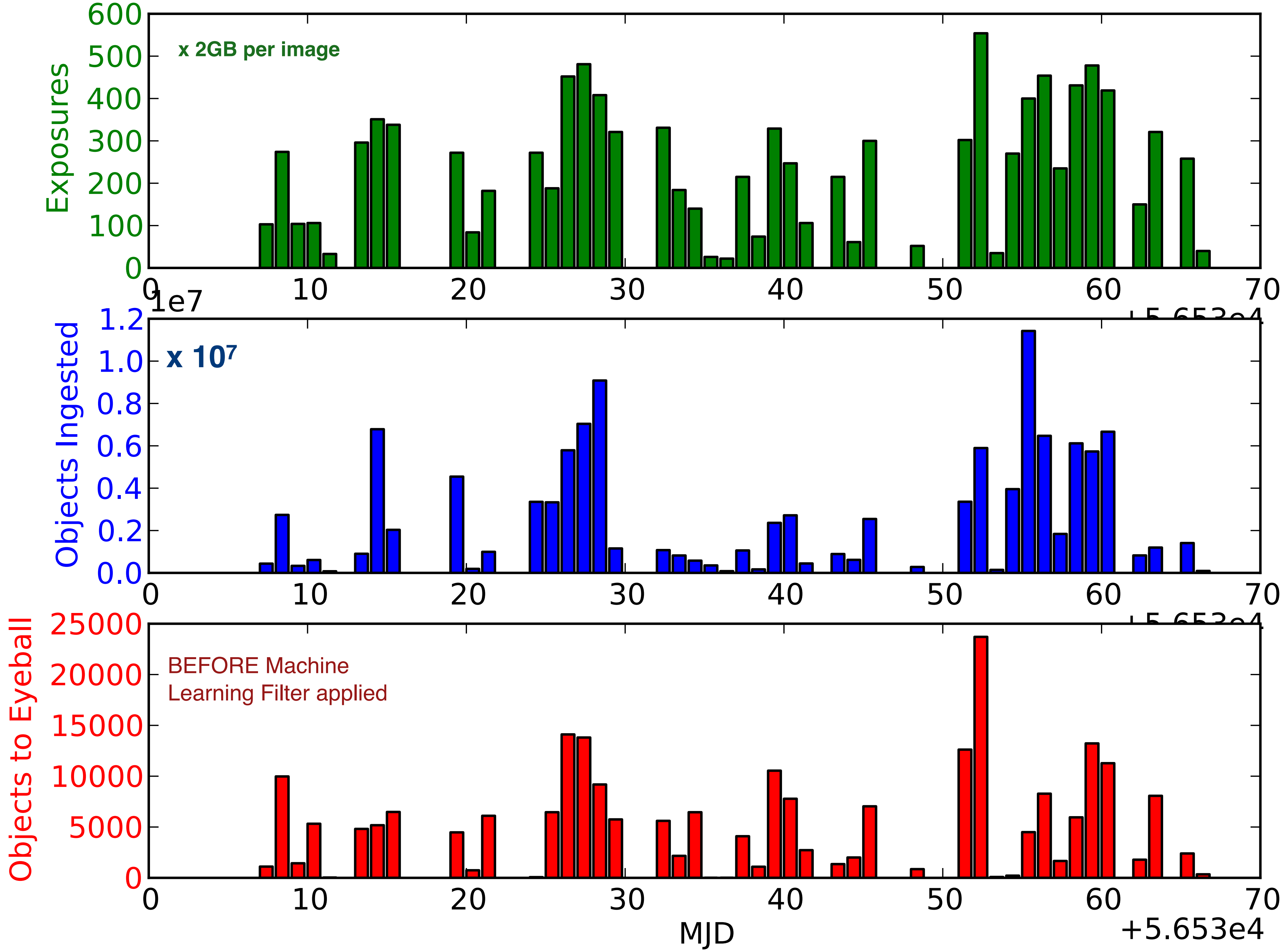
Confirmed SNe
(118)

<< previous 1 2 next >>

Rank	Followup flag date	RA	DEC	Type	Spec Type	Local Name	PS1 Name	Trend	Target	Ref	Diff	RB Factor	Recent Triplet MJD
3101519	June 5, 2015	21:08:18.61	+17:09:56.2	orphan	II-P	5F3Pyma	PS15apv	rising 00.01 (i-i)				0.97	57178.56871
3035720	May 25, 2015	15:49:53.51	-21:02:50.6	sn	II	5E3Pxpn	PS15ann	fading 00.16 (w-w)				0.79	57181.44061
3009049	May 23, 2015	13:09:18.58	-25:52:20.0	sn		5E3Pwlb	PS15amx	fading 00.46 (i-w)				0.93	57164.35973
2958846	May 16, 2015	14:40:56.26	+03:31:44.3	sn	II	5E3Pulm	PS15akk	rising 00.07 (w-w)				0.99	57157.47050

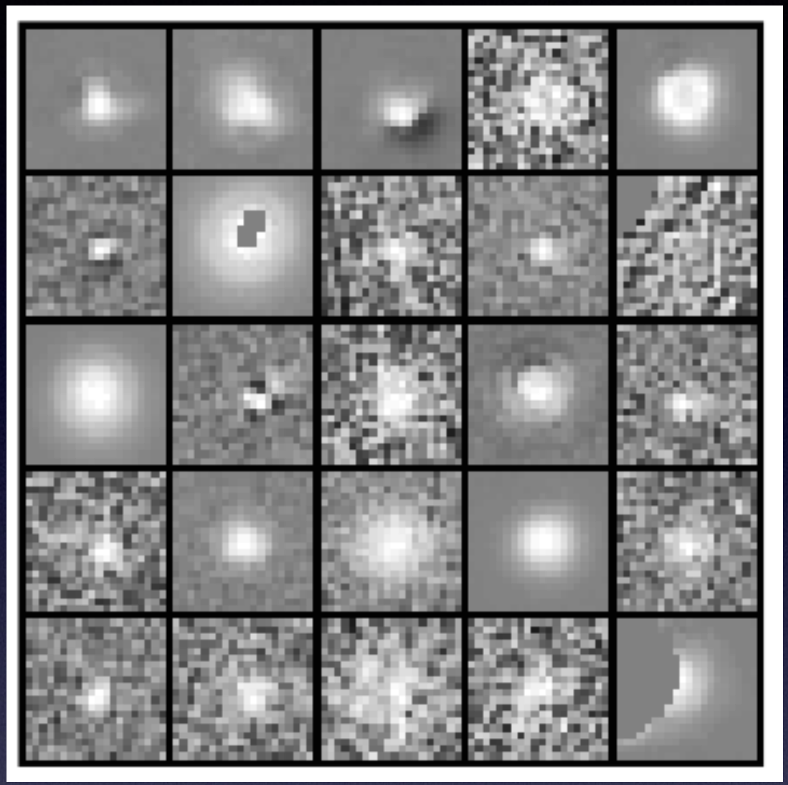
Example of current data rate processing

Average number of detections per day in PanSTARRS ~ 10 million

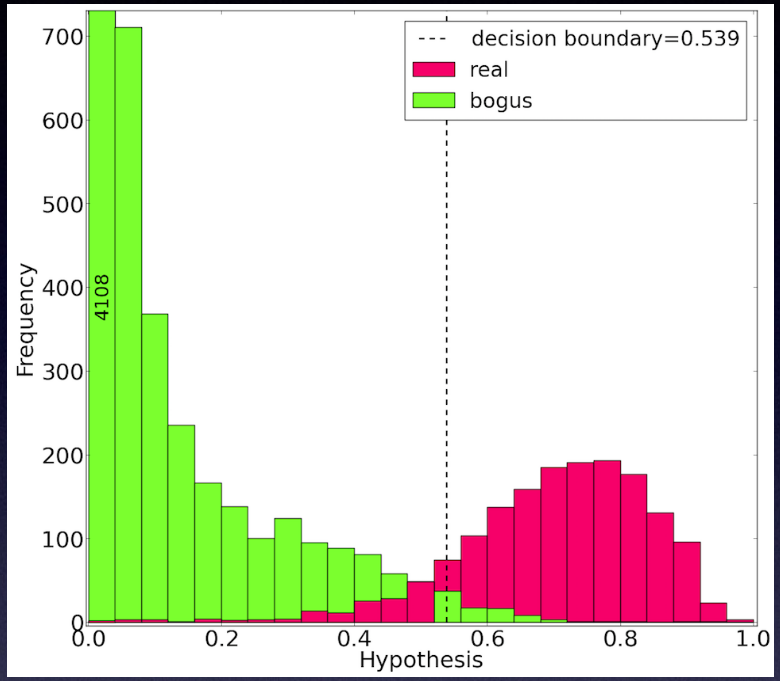


1. Image recognition : machine learning

Image input - which are real, which are not ?

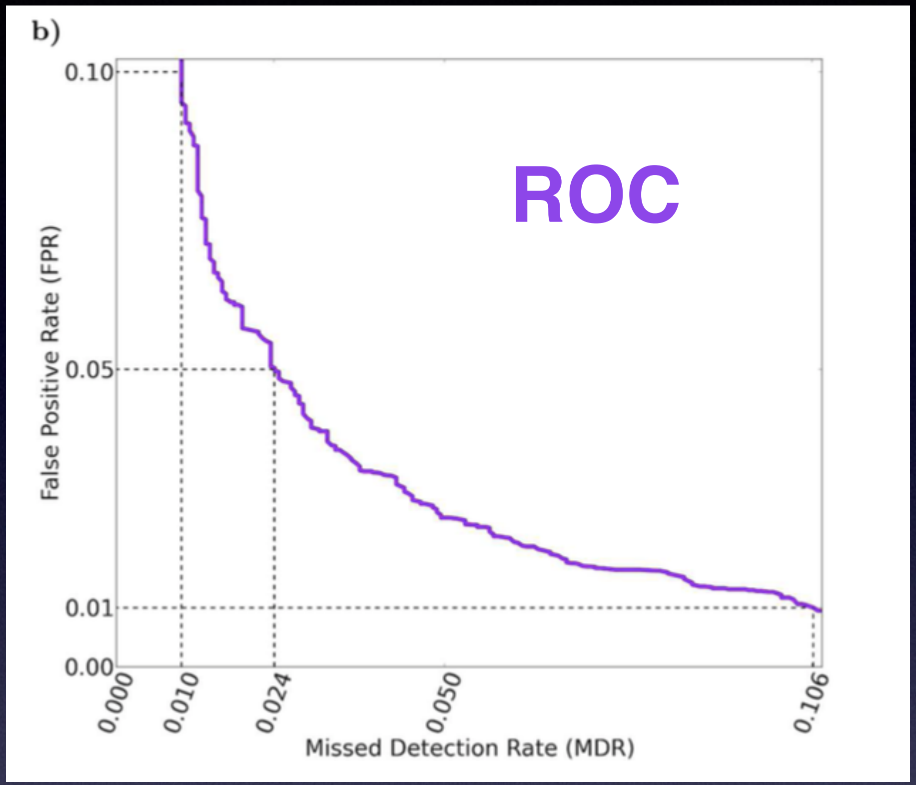


Random Forest Classifier, neural networks, support vector machines



Floating point number between 0 and 1
Bogus objects have H < 0.5
Real objects have H > 0.5

Receiver Operating Characteristic curve



Working in real time now, but two problems

- Have hit floor in performance for 1% FPR : can't do better than 5-10% MDR
- Astrophysical classification, once we decide REAL/BOGUS

Machine learning for transient discovery in Pan-STARRS1 difference imaging
D. Wright et al. MNRAS, 2015, 449, 451

http://star.pst.qub.ac.uk/ps1threepi/psdb/

2. Massive databases

ATLAS (2 x 0.5m telescopes, 20 mag, all-sky 2-4 times per night)

- Object : 100 bytes (conservative! FP number 4 bytes, double=8 bytes. Excluding indexes, overlapping partitions.)
- 1×10^9 sources
- 1×10^{12} detections per yr
- 100 TB database per yr (x 2-3 for backup)

Large Synoptic Survey Telescope

- 40×10^9 sources (after Year 1)
- 1×10^{12} detections per yr
- 100TB database per yr (but 10 yr rolling project, and "forced" measurements. Final = 15PB)

Big data problems for transient sky surveys in astronomy

1. "Small projects" now producing big data and associated problems

2. UK will play major role in LSST : both image analysis, classification and database architecture unsolved (LSST developing qserv)

3. Speed : insert 64000 per sec into database (24hr spread, so probably worse). Need to rapidly index and associate, and be querying at same time (support multiple users)

Compute and Data-Intensive Simulations, Error Analysis & Control in the Chemical Sciences

Peter Coveney

Centre for Computational Science,

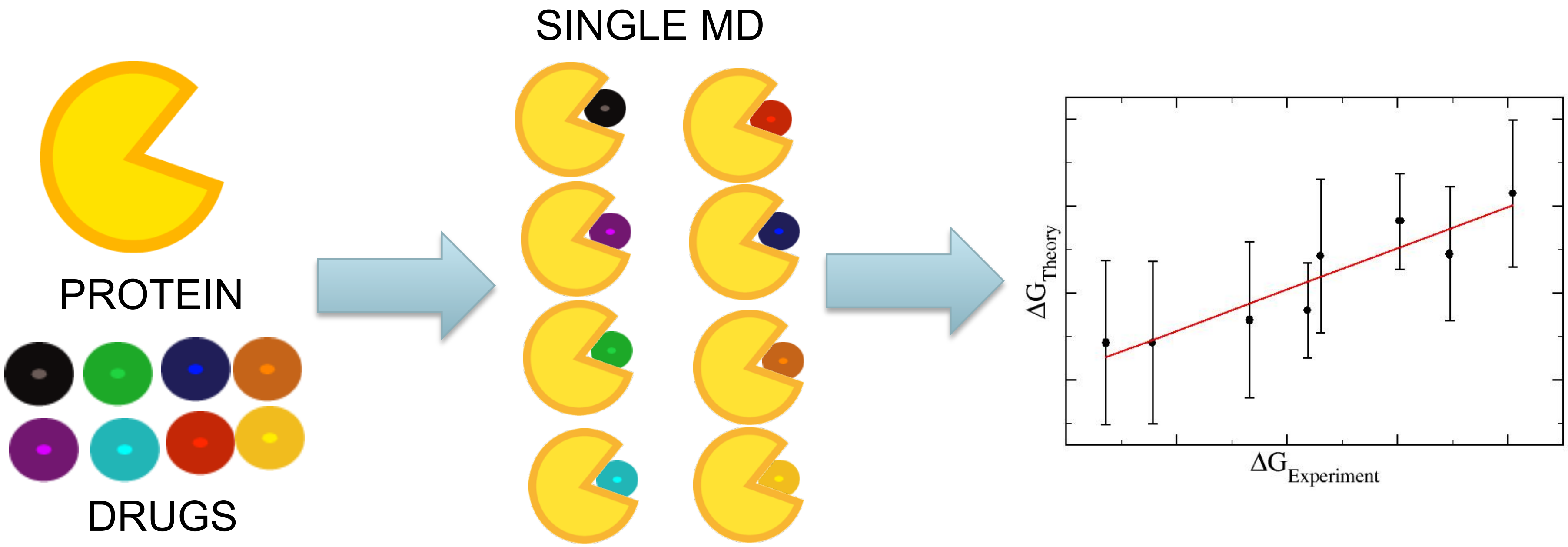
Department of Chemistry,

University College London

Alan Turing Institute Summit

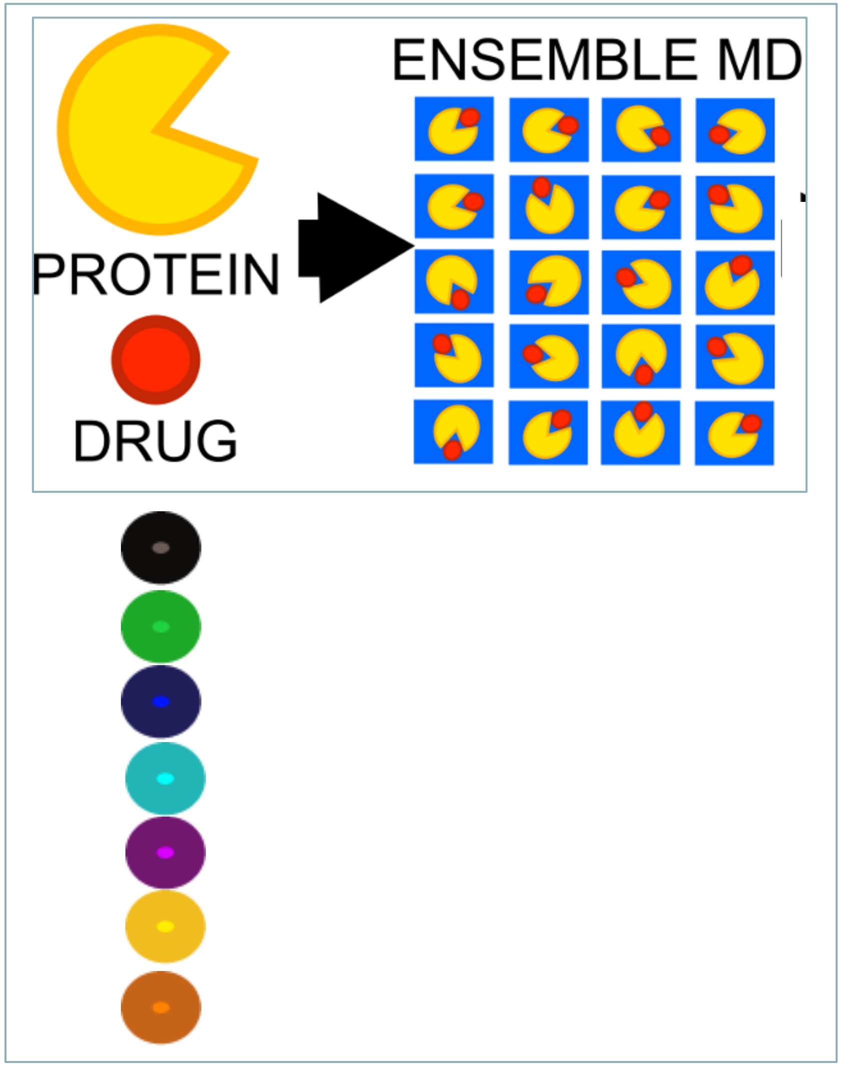
13 January 2016

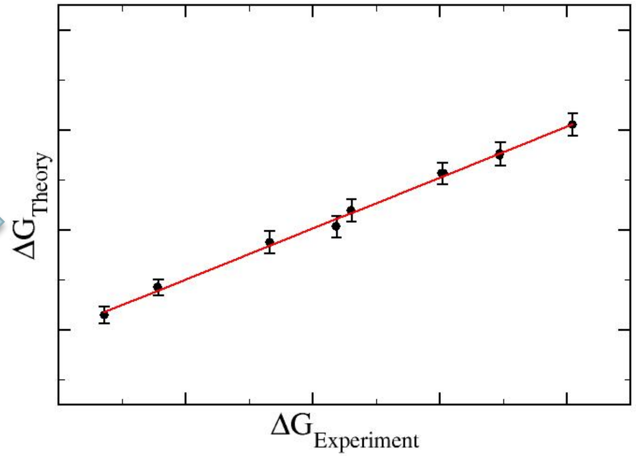
Computational Application to Drug Affinity Ranking – Single MD simulation



Errors uncontrolled
Results unreproducible

Computational Application to Drug Affinity Ranking – Ensemble Simulations





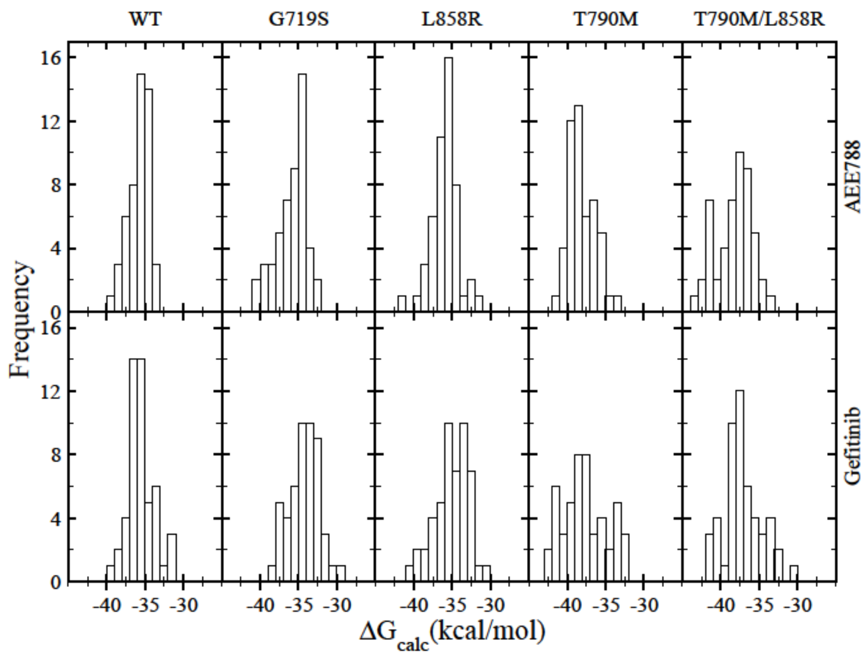
Errors fully under control;
Results reproducible.

(Data from Bcr-Abl kinase ligand binding.)

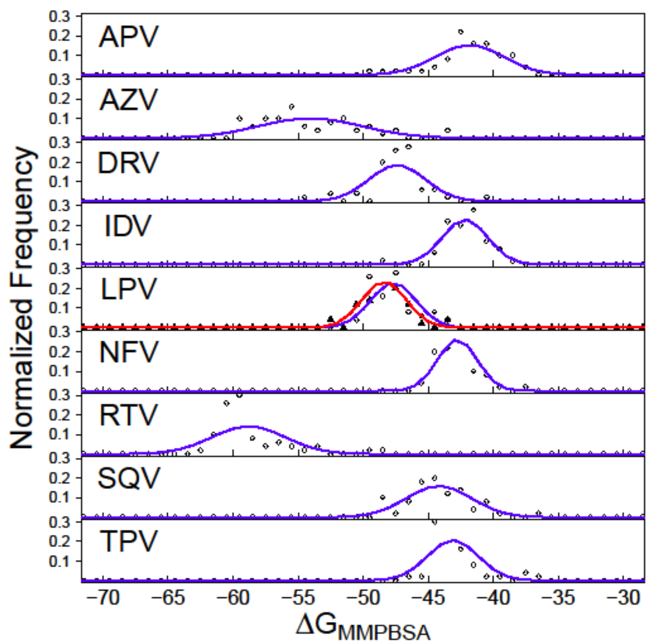
Single *vs* Ensemble MD Simulations

The binding free energy can vary widely (up to 12 kcal/mol) between two single simulations.

Single simulation: not reproducible, unscientific!



Drug – EGFR

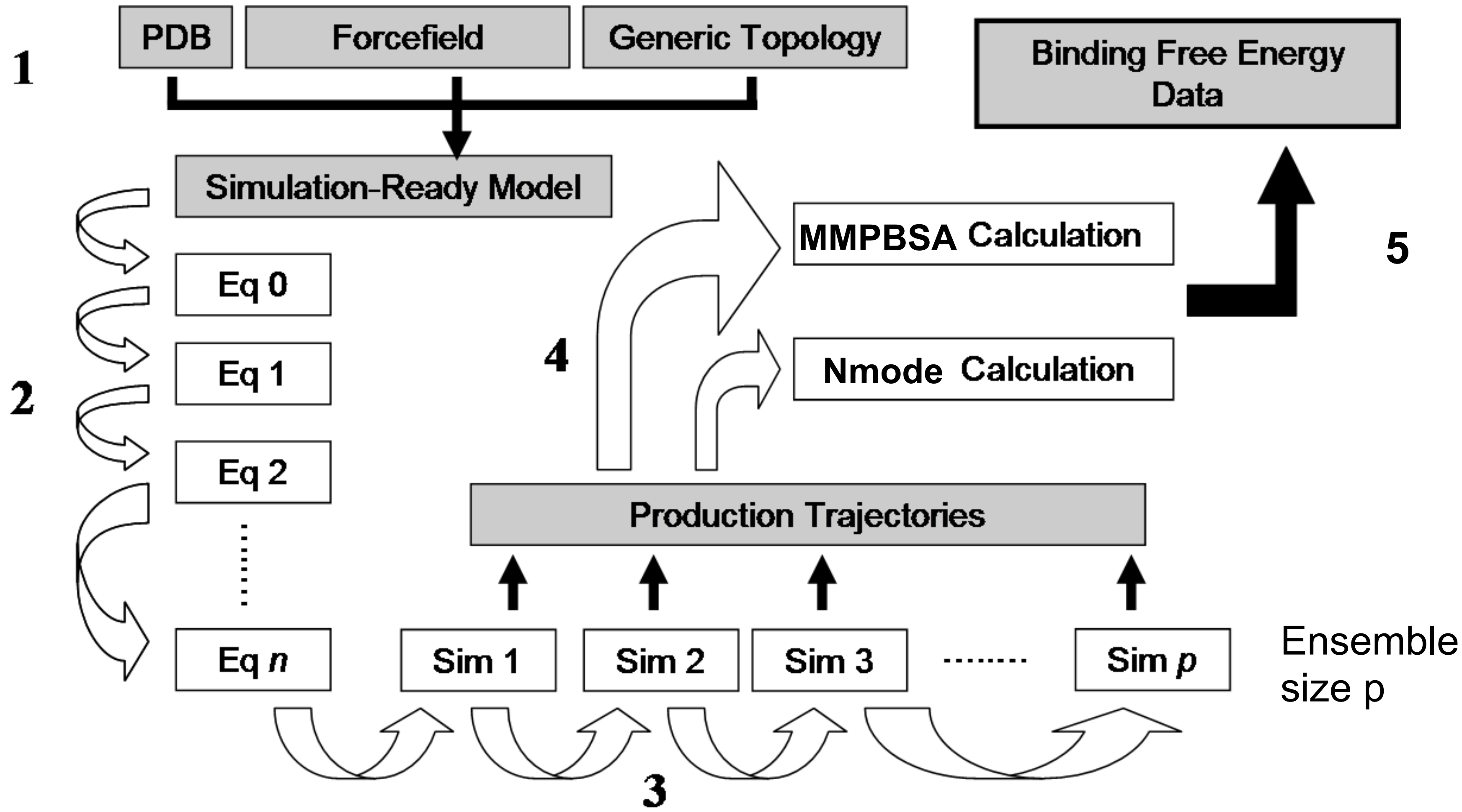


Drug – HIV-1 protease

Wan & Coveney, J. R. Soc. Interface, 8, 1114-1127, (2011).
Wright, Hall, Kenway, Jha & Coveney, J. Chem. Theory Comp. (2014), DOI: 10.1021/ct4007037.

Binding Affinity – Data Intensive Workflow

Molecular Mechanics Poisson-Boltzmann Surface Area (MMPBSA) & Entropy Calculation



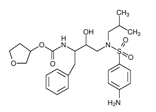
1. Model preparation; 2. Equilibration; 3. Production; 4. Free energy calculation; 5. Analyses and results

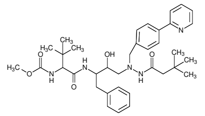
Applications used include: NAMD, CHARMM, AMBER, VMD…

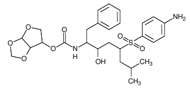
S. K. Sadiq, D. Wright, S. J. Watson, S. J. Zasada, I. Stoica, Ileana, and P. V. Coveney, *Journal of Chemical Information and Modeling*, **48**, (9), 1909-1919 (2008)

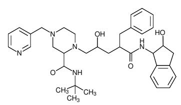
Calculating Clinically Relevant Binding Affinities

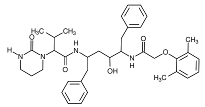
FDA-approved drugs to wild-type HIV-1 protease

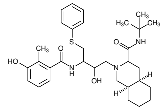

Amprenavir (APV)

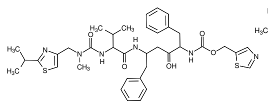

Atazanavir (AZV)

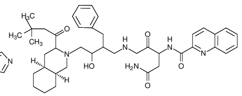

Darunavir (DRV)

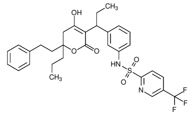

Indinavir (IDV)

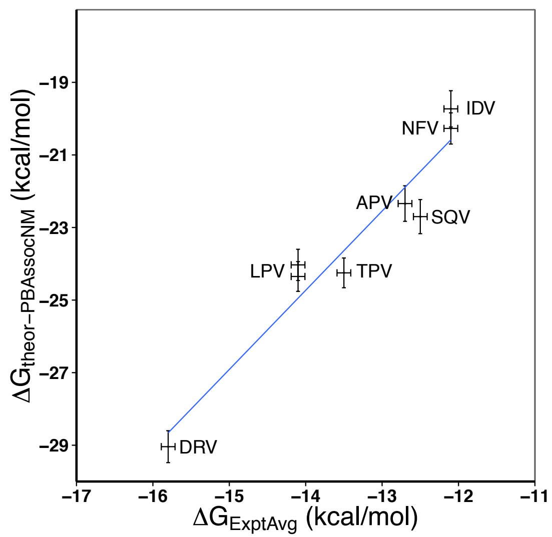

Lopinavir (LPV)


Nelfinavir (NFV)

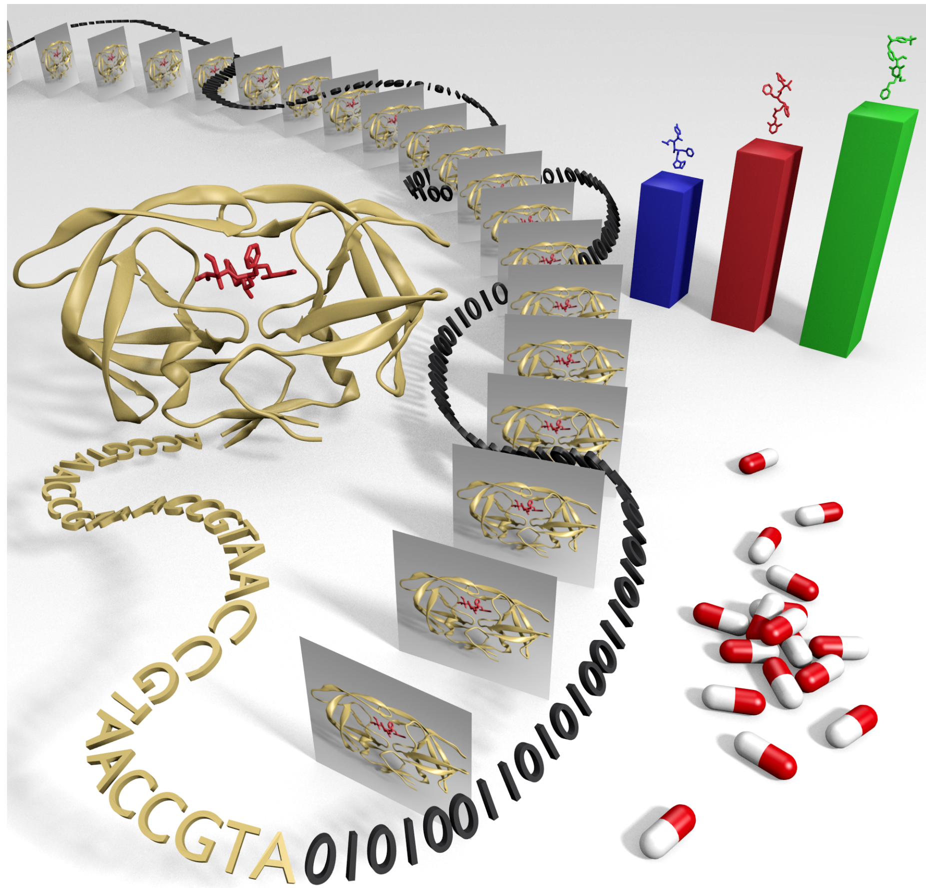

Ritonavir (RTV)


Saquinavir (SQV)


Tipranavir (TPV)



This work used several of the most powerful supercomputers in the USA, UK, and EU.

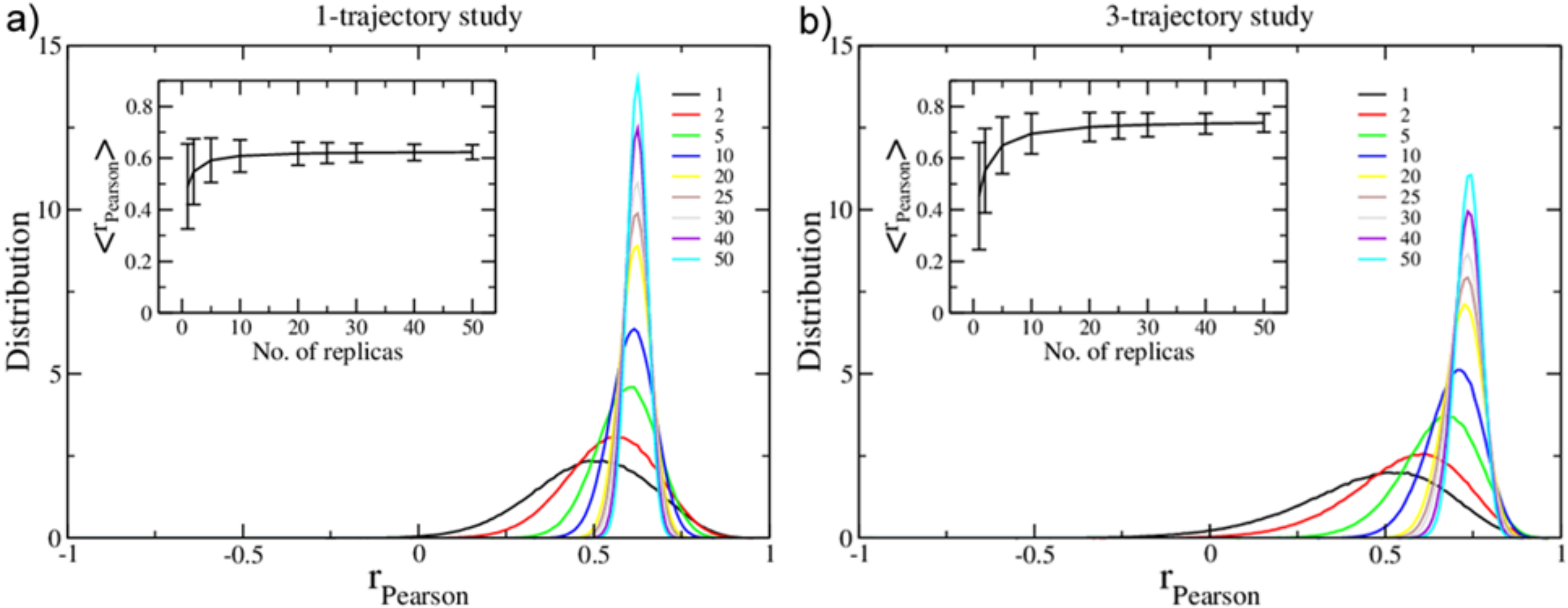


Wright, DW, Hall, BA,Kenway, OA, Jha, S and Coveney, PV, "Computing Clinically Relevant Binding Free Energies of HIV-1 Protease Inhibitors." **J. Chem. Theory Comput.,** 2014, DOI: 10.1021/ct4007037

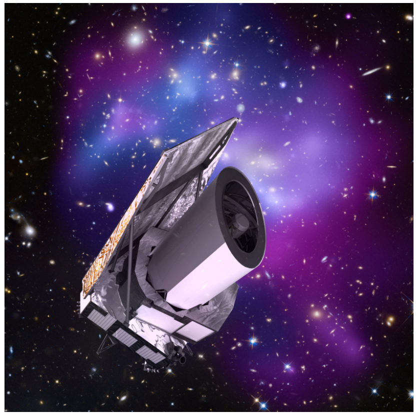
Ranking of p-MHC Binding Free Energies

The influence of ensemble size on the reproducibility

- Larger sizes of ensemble make rankings more reproducible and with lower standard deviations.
- One should use ensembles containing a minimum of 25 replicas per ensemble to provide reproducible results.



S. Wan, B. Knapp, D. Wright, C. Deane, P. V. Coveney, *J. Chem. Theory Comput.* **11** (7) 3346-3356 (2015)



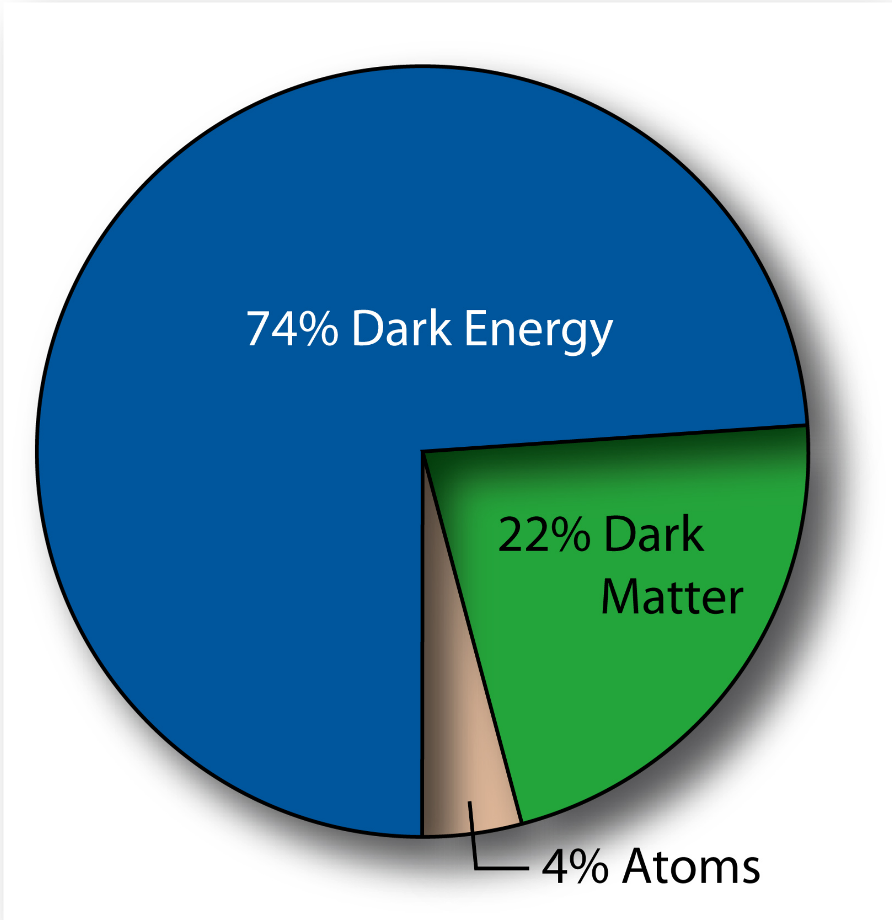
Euclid

Europe's Next Space-Based Cosmology Experiment

Tom Kitching (UCL MSSL) – Euclid Science Lead

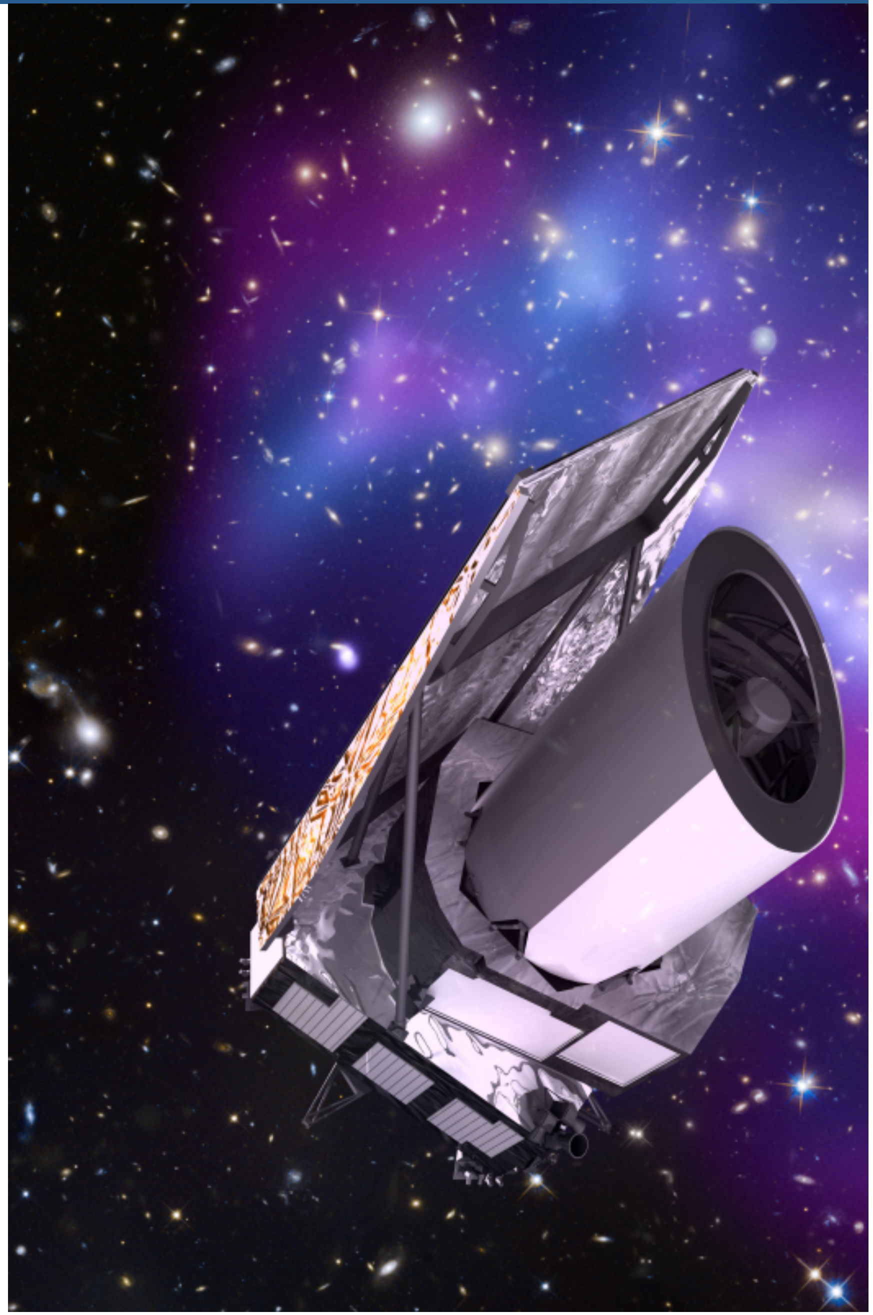
What is the Universe made of?

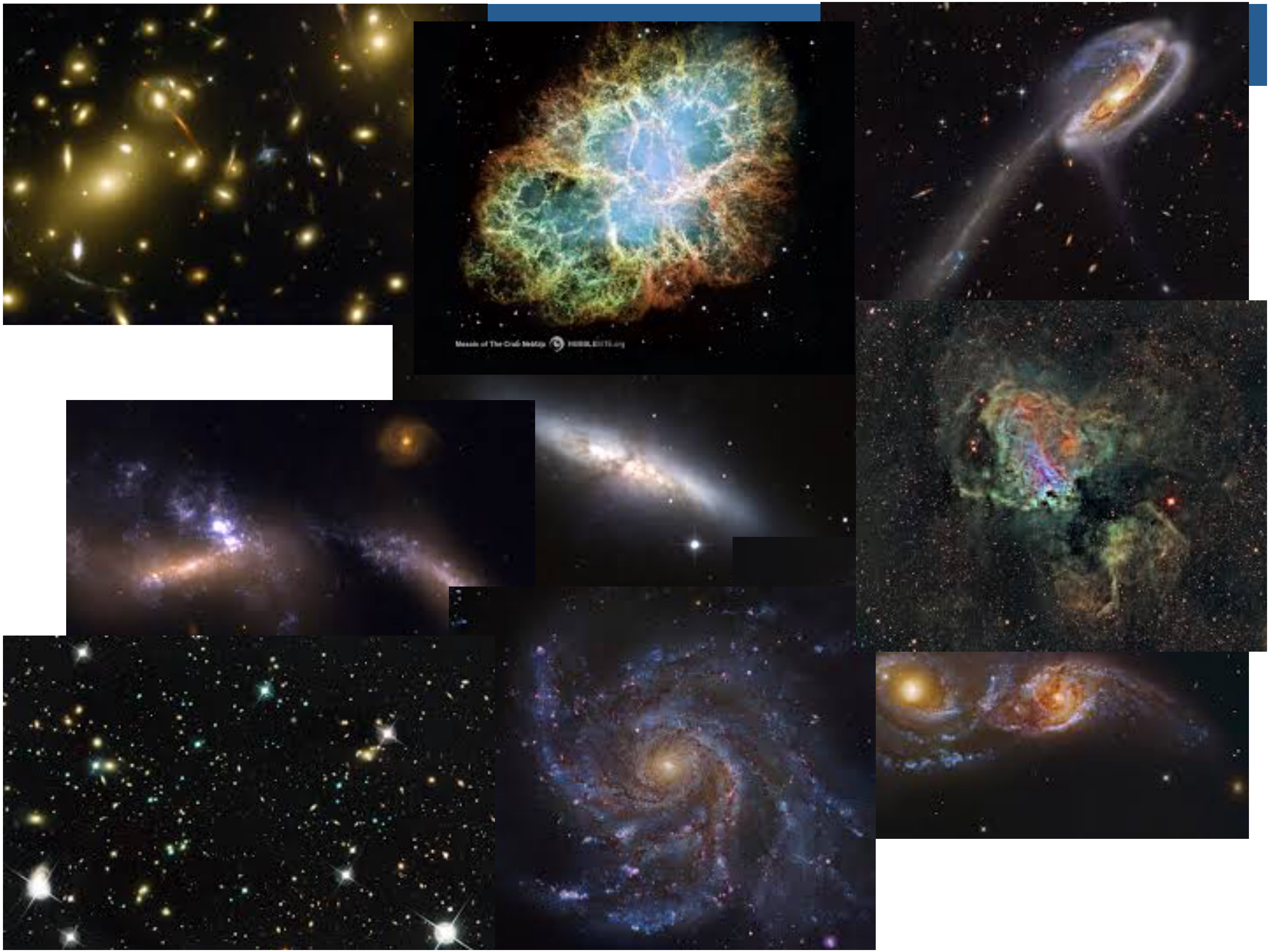
- Euclid is designed to to decisively answer this question
- Explanations require either:
  – Changing general relativity
  – A new fundamental field (like the Higgs)
  – Multiverse

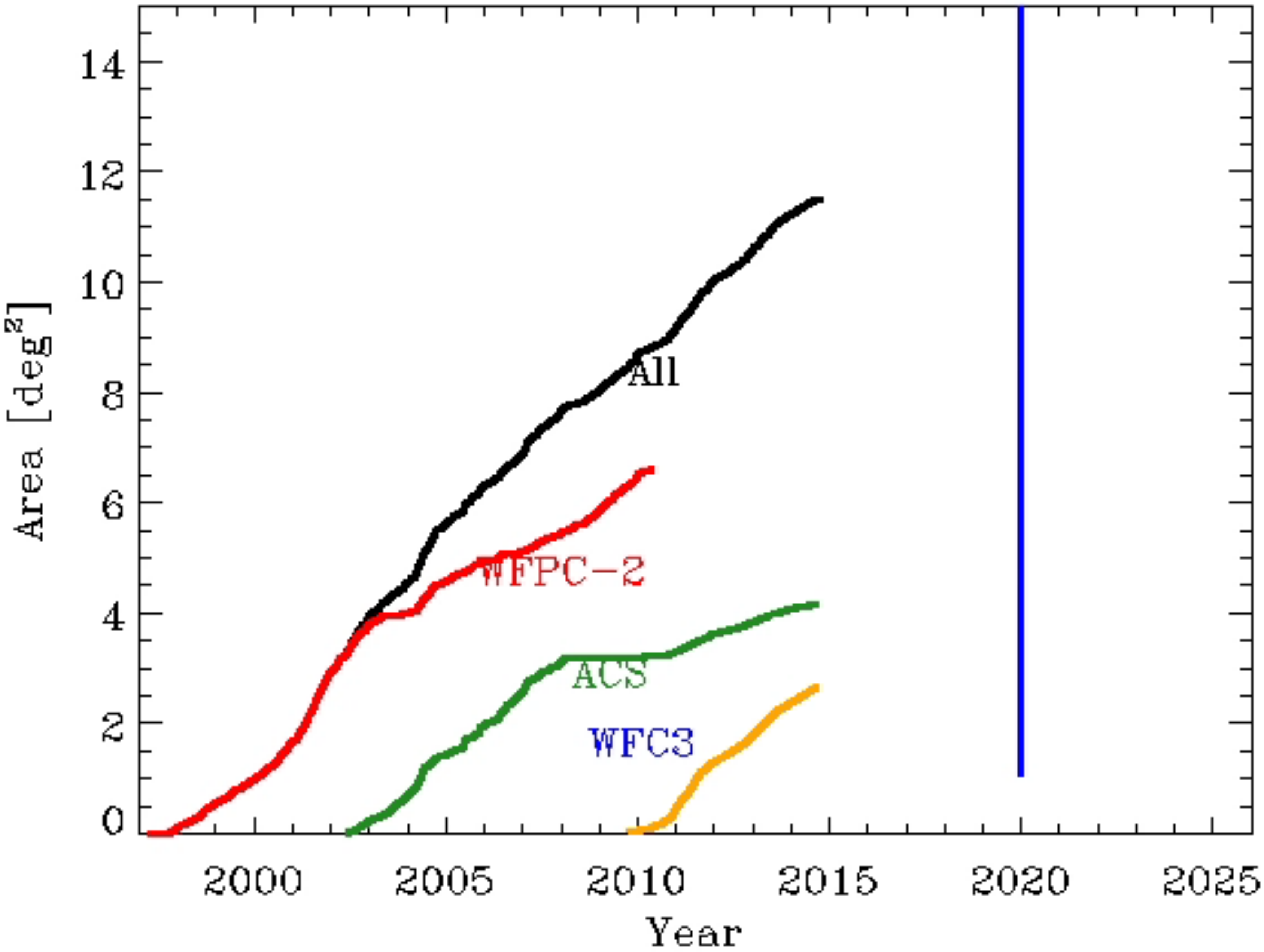


What is Euclid ?

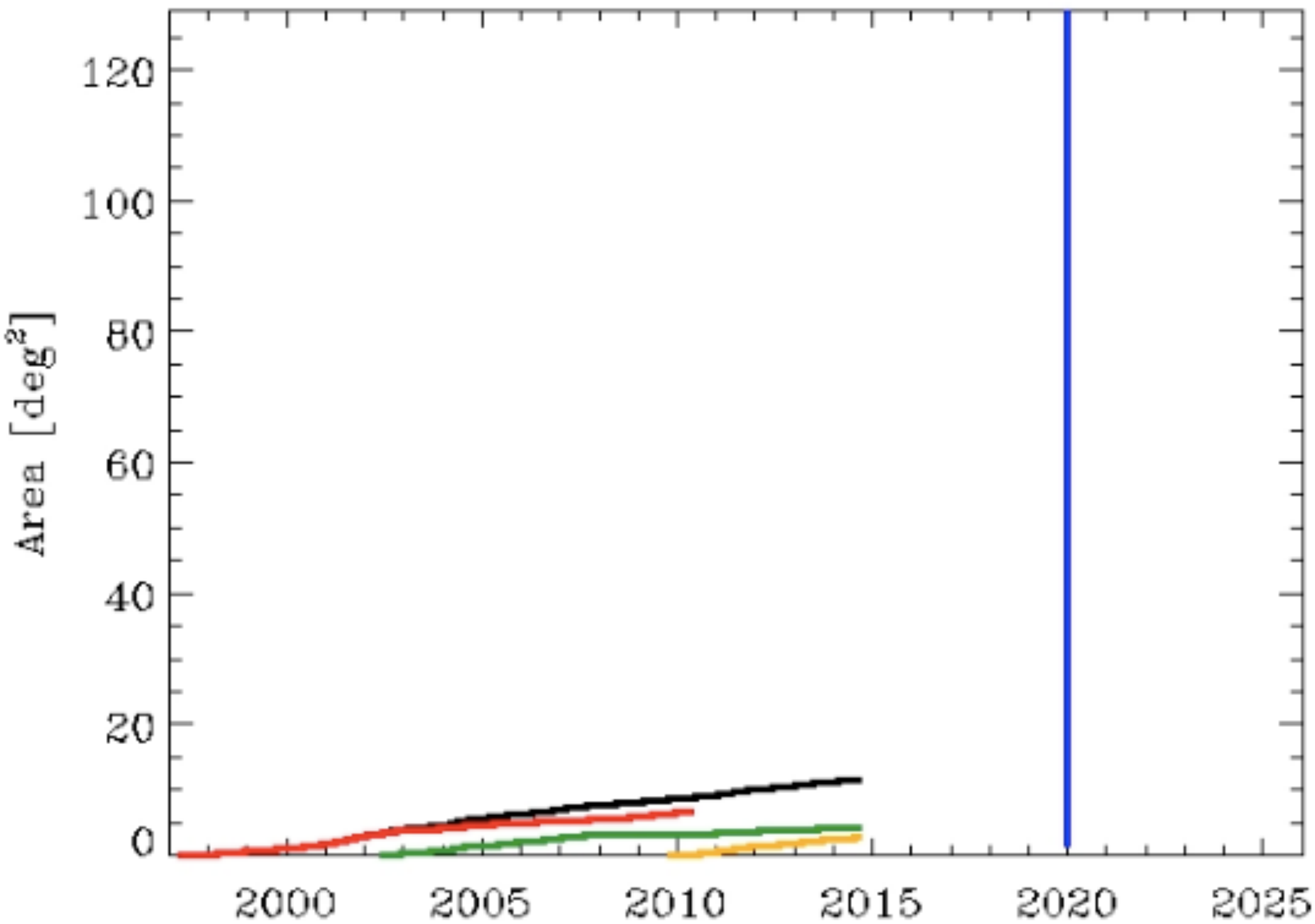
- Due to launch in 2020
- UK leads Science, Data Processing & Engineering aspects
- Product: Hubble-Space Telescope quality images over 75% the available sky over 75% the age of the Universe



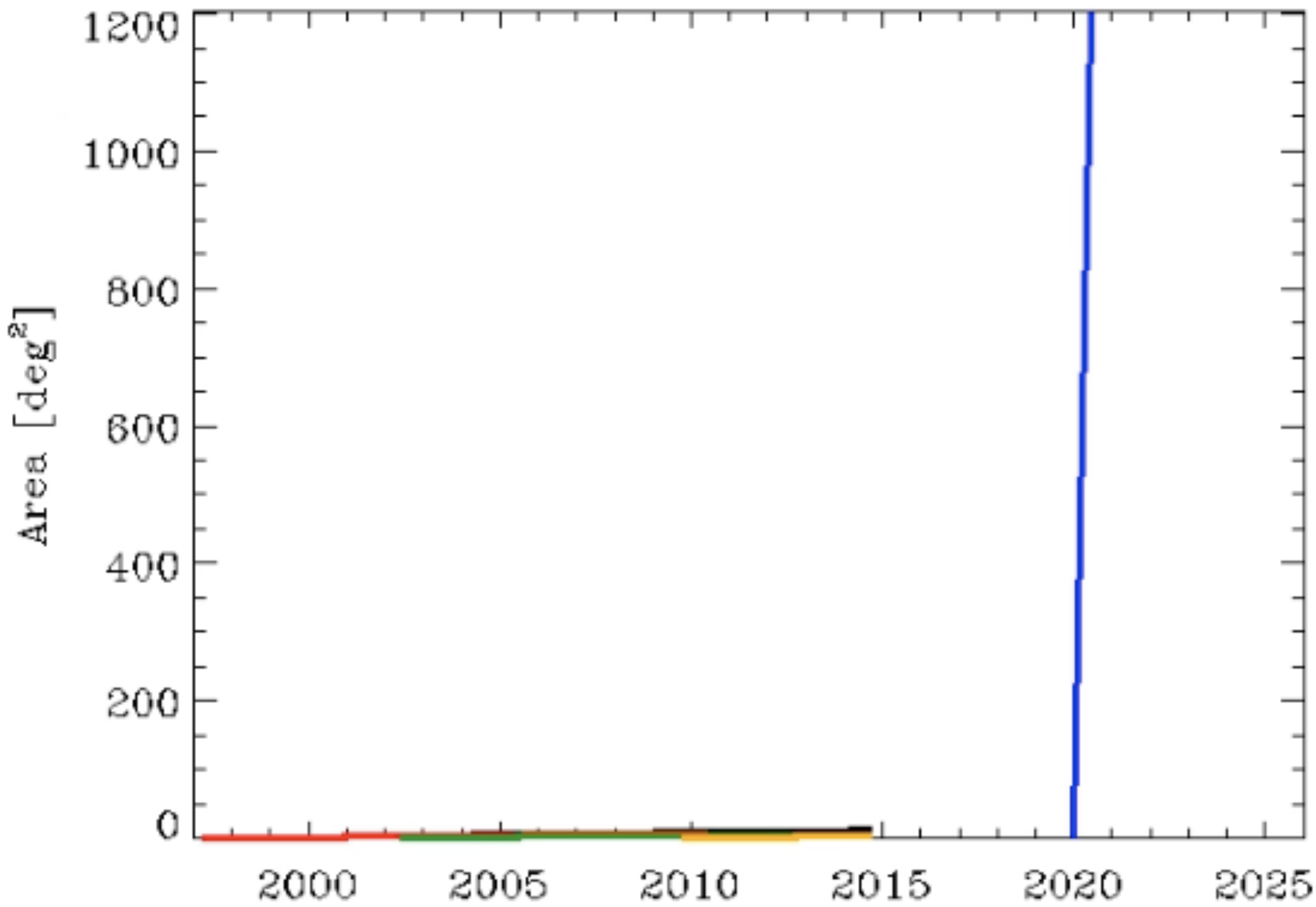



Area [deg^2]
14
12
10
8
6
4
2
0
2000
2005
2010
2015
2020
2025
Year
All
WFPC-2
ACS
WFC3

2 weeks

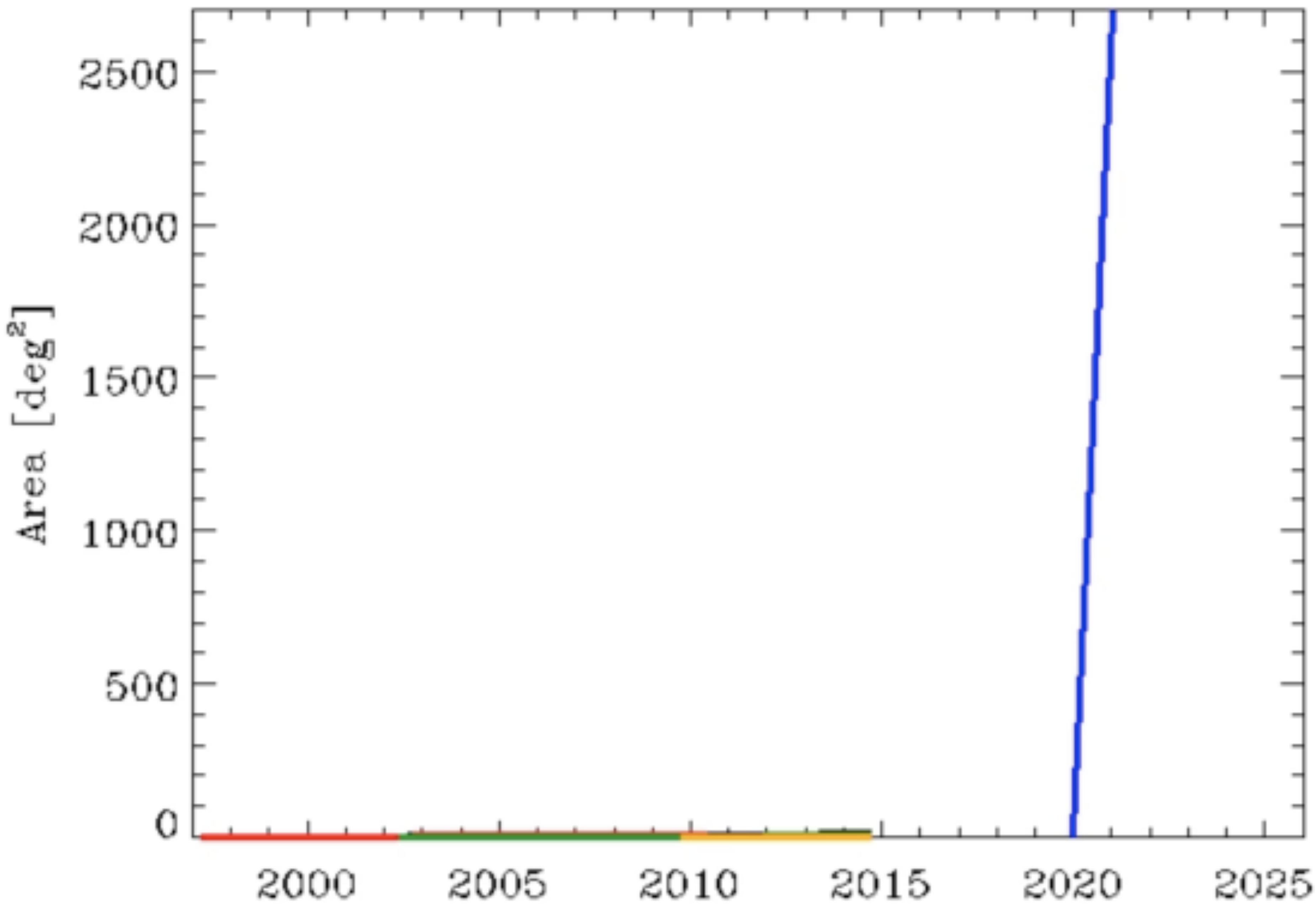


6 months

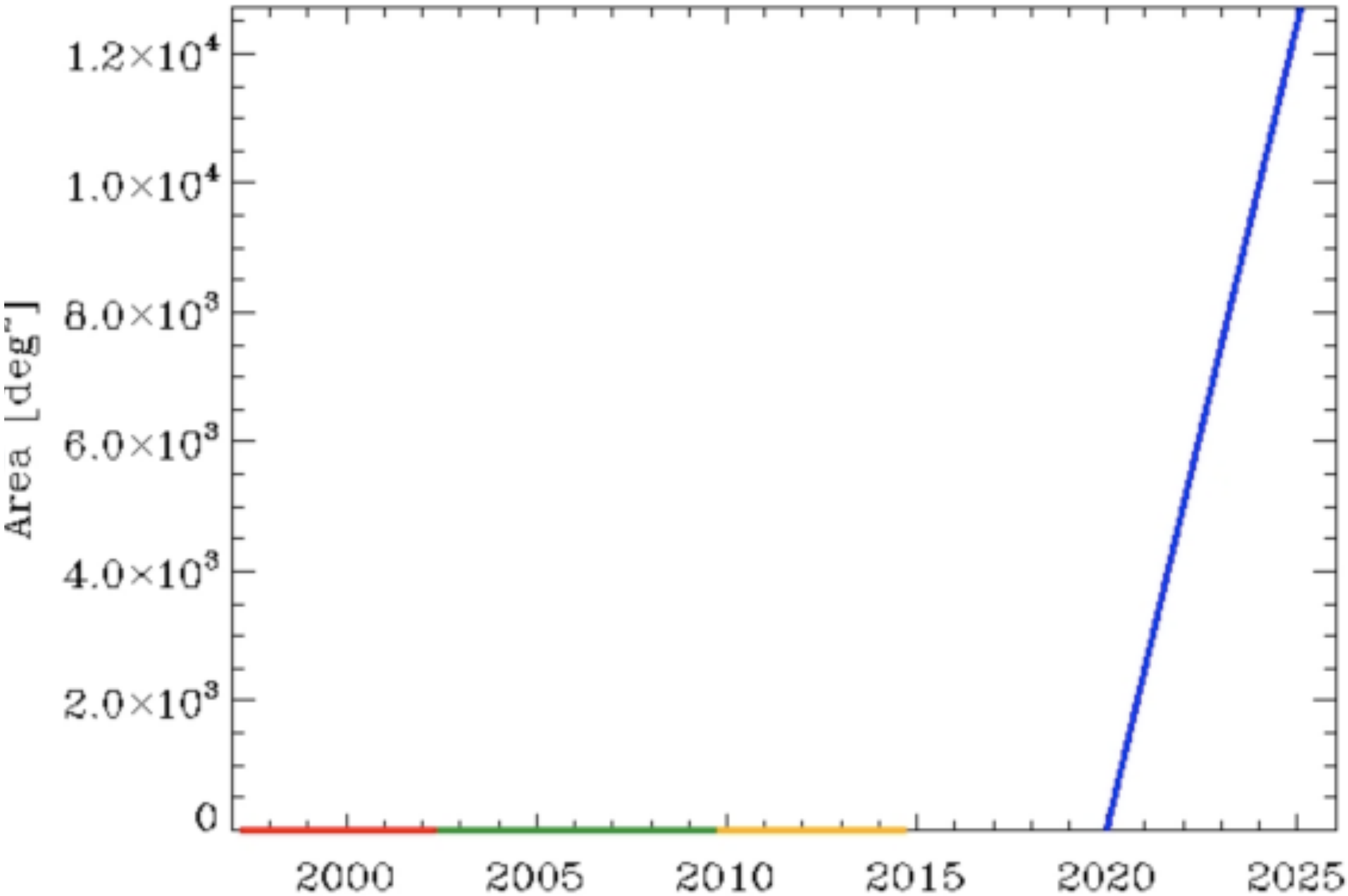


1200
1000
800
600
400
200
0
Area [deg^2]
2000
2005
2010
2015
2020
2025

1 year


2500
2000
1500
1000
500
0
Area [deg^2]
2000
2005
2010
2015
2020
2025

5 years



Area [deg²]
1.2×10⁴
1.0×10⁴
8.0×10³
6.0×10³
4.0×10³
2.0×10³
0
2000
2005
2010
2015
2020
2025

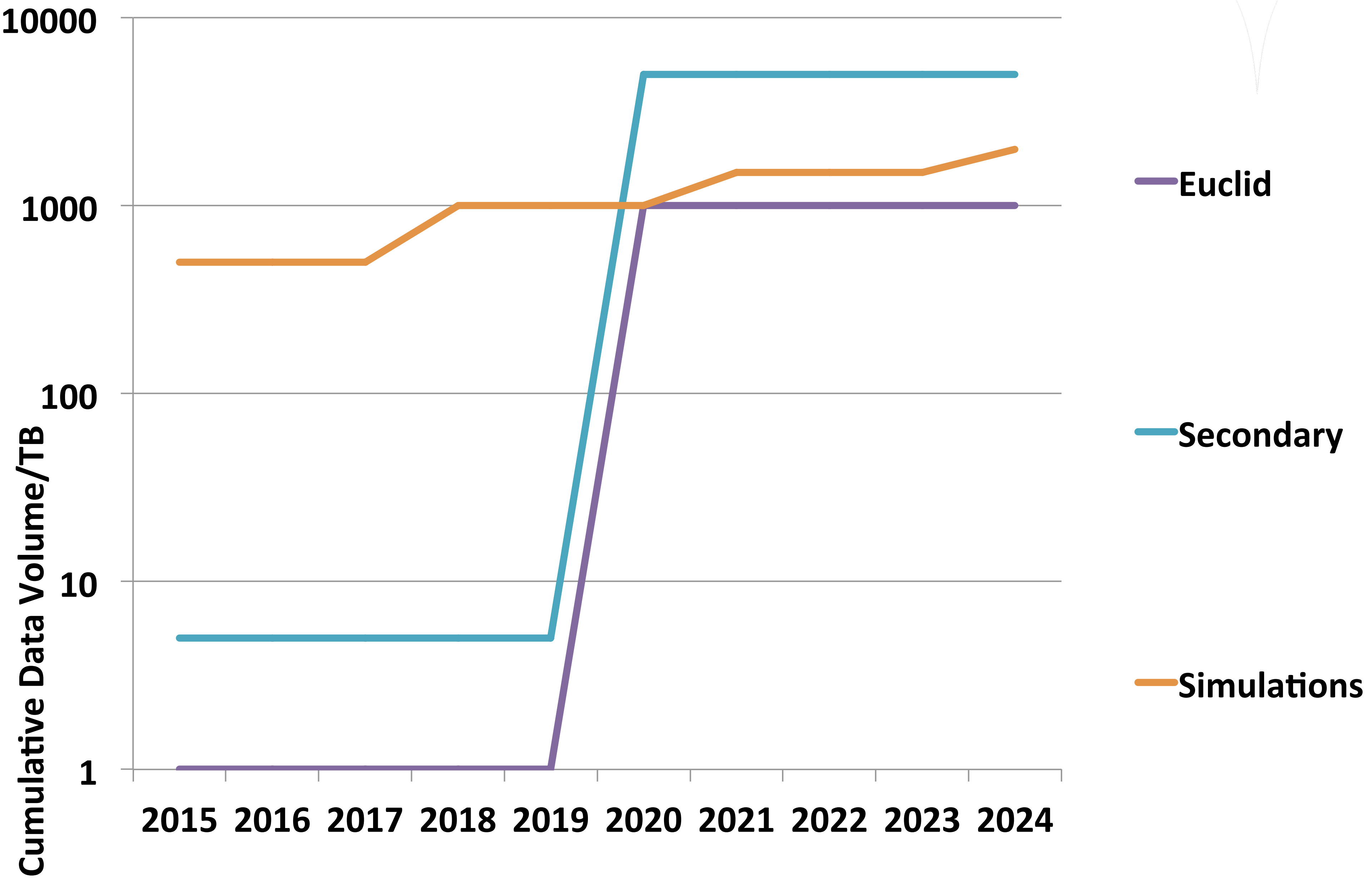
Big Data

- The largest CCD array ever flown in space
- 36 4k x 4k chips
- Need 36 stacked HD TVs to display *one* image
- Euclid will make one image every 5 minutes continuously for 6 years

Big Simulations

- Only have one Universe
- So need to re-run the experiment in simulations
- Require > 10^6 Universe simulations
- Hundreds-thousands of PB required

Total Science Storage Requirements



- Euclid will observe:

75% of extragalactic sky
over 75% the age of the Universe

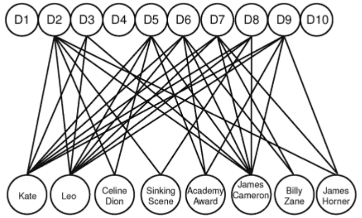
- Designed to determine nature of dark energy
- Big Data, Big Simulations
- Big Opportunity for UK & the ATI

Efficient Massive-Scale Graph Processing

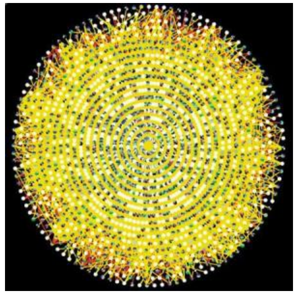
Eiko Yoneki
eiko.yoneki@cl.cam.ac.uk
http://www.cl.cam.ac.uk/~ey204

Systems Research Group
University of Cambridge Computer Laboratory

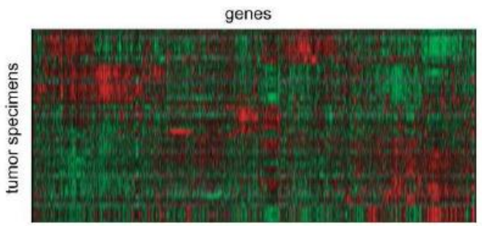
Emerging Massive-Scale Graph Data



Bipartite graph of phrases in documents



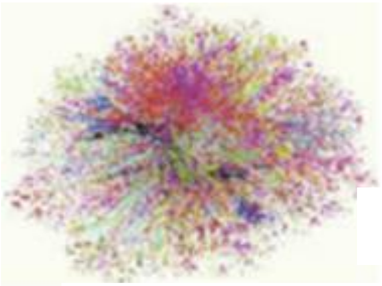
Protein Interactions [genomebiology.com]



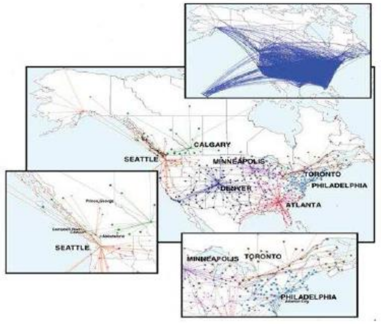
Gene expression data

Brain Networks: 100B neurons(700T links) requires 100s GB memory

Social media data



Web 1.4B pages(6.6B links)

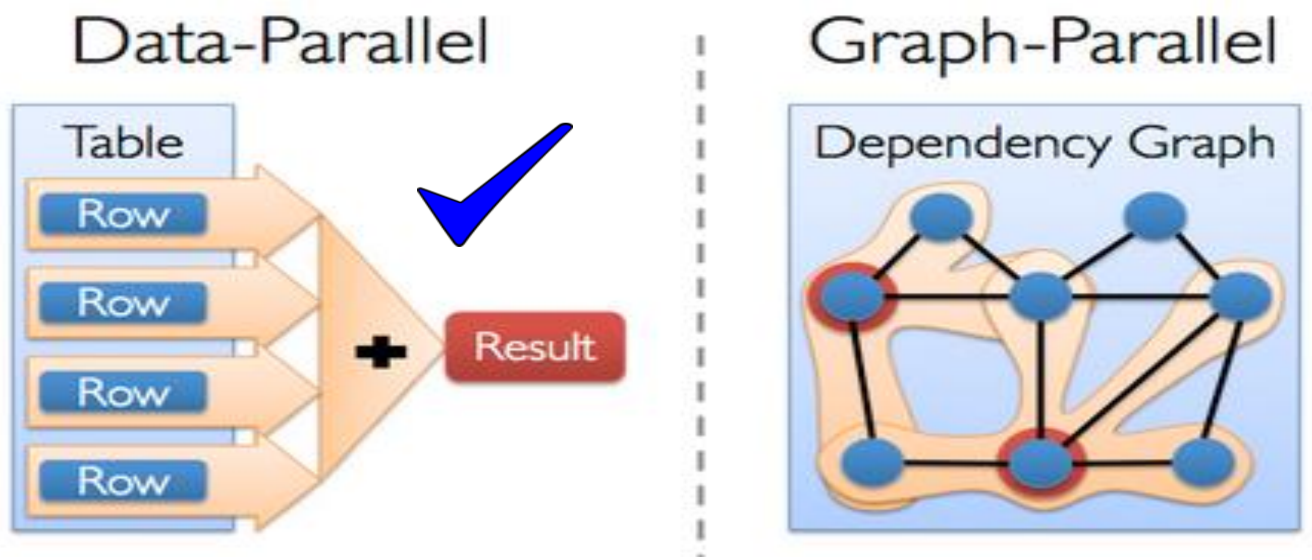


Airline Graphs

Everything will be connected in Future!

Data-Parallel vs Graph-Parallel

- Big data forms complex networks: key to solve problems in diverse fields
  - Web 1.4B pages + 6.6B links; Brains 100B neurons + 700T links → 100s GB of memory
- *Data-Parallel* for everyone? *Graph-Parallel* is hard!
  - Only for big players with HPC/Large Clusters?



- BSP: Pregel, Giraph, Graphlab
- Unifying graph- & data-parallel: GraphX/Spark
- Data-flow programming: NAIAD, DryadLINQ

Big Data: Scale-Up vs Scale-Out

- Popular solution for big data processing
  → scale and build distribution, combine theoretically unlimited number of machines in single distributed storage

- Scale-up: add resources to single node in system (e.g. HPC)
- Scale-out: add more nodes to system (e.g. Amazon EC2)

Do we really need large clusters?

- Laptops are sufficient

Twenty pagerank iterations

System	cores	twitter_rv	uk_2007_05
Spark	128	857s	1759s
Giraph	128	596s	1235s
GraphLab	128	249s	833s
GraphX	128	419s	462s
Single thread	1	300s	651s

Fixed-point iteration: All vertices active in each iteration (50% computation, 50% communication)

Label propagation to fixed-point (graph connectivity)

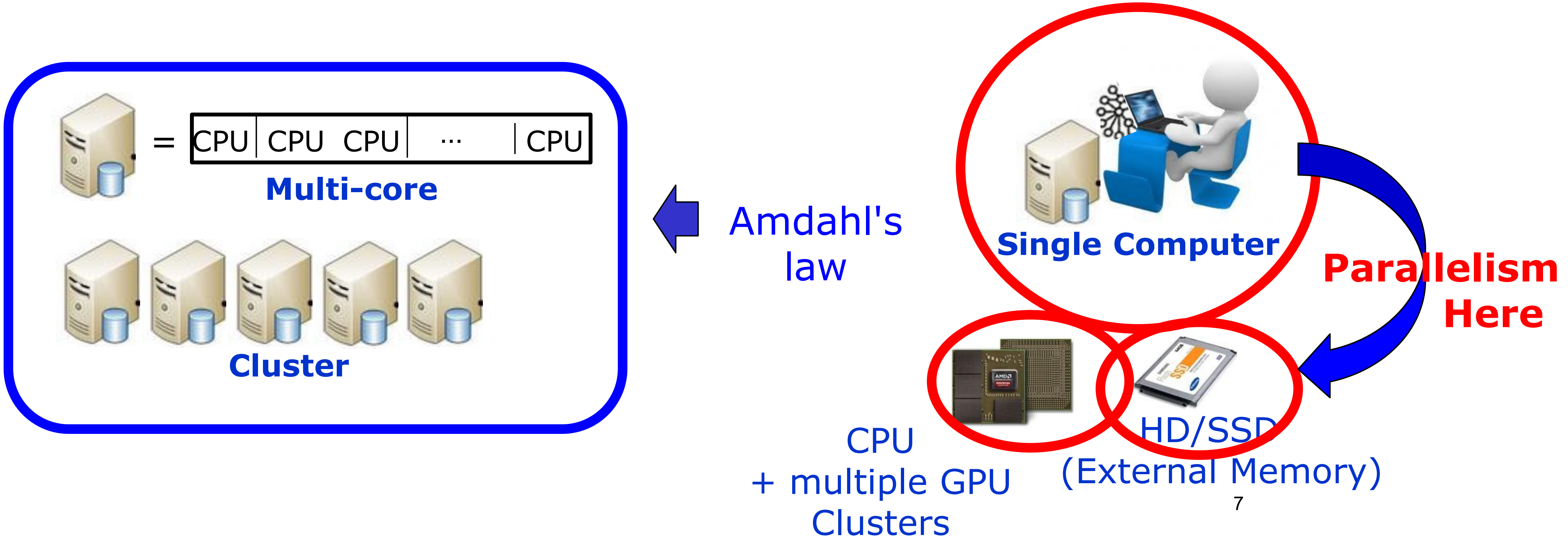
System	cores	twitter_rv	uk_2007_05
Spark	128	1784s	8000s+
Giraph	128	200s	8000s+
GraphLab	128	242s	714s
GraphX	128	251s	800s
Single thread	1	153s	417s

Traversal: Search proceeds in a frontier (90% computation, 10% communication)

from Frank McSherry HotOS 2015

Bring Big Data Processing to Single Computers

- Use of powerful HW/SW parallelism
  - SSDs as external memory
  - CPU/GPU integrated heterogeneous many core architecture
- Open up massive graph processing to everyone



Graph Computation Challenges

1. Graph algorithms (BFS, Shortest path)
2. Query on connectivity (Triangle, pattern)
3. Structure (Community, Centrality)
4. ML & Optimisation (Regression, SGD)

- Data driven computation: dictated by graph's structure and parallelism based on partitioning is difficult
- Poor locality: graph can represent relationships between irregular entries and access patterns tend to have little locality
- High data access to computation ratio: graph algorithms are often based on exploring graph structure leading to a large access rate to computation ratio

Research Vision: Synthesis of Entire Stack

- Algorithms, S/W and H/W for mainstream parallel approaches are not effective for more complex structured data from real world
- Data and algorithms dictate complex & irregular graph data processing: Utilise systems' parallelisms and resource coordination - no burden of algorithm implementation

- Close gap between domain algorithms and systems research
- Programming paradigm and model (runtime, algorithmic, query layer…)
  - Opening up fresh research areas such as algorithm independent optimisation

- Exploit different parallelism at different scales (SSD, CPU/GPU)
- Map input data structure and algorithms onto processing model
- Auto-tuning structured Bayesian optimisation for dynamic scheduling
  - Complex decision making, and resource provisioning in complex parameter space

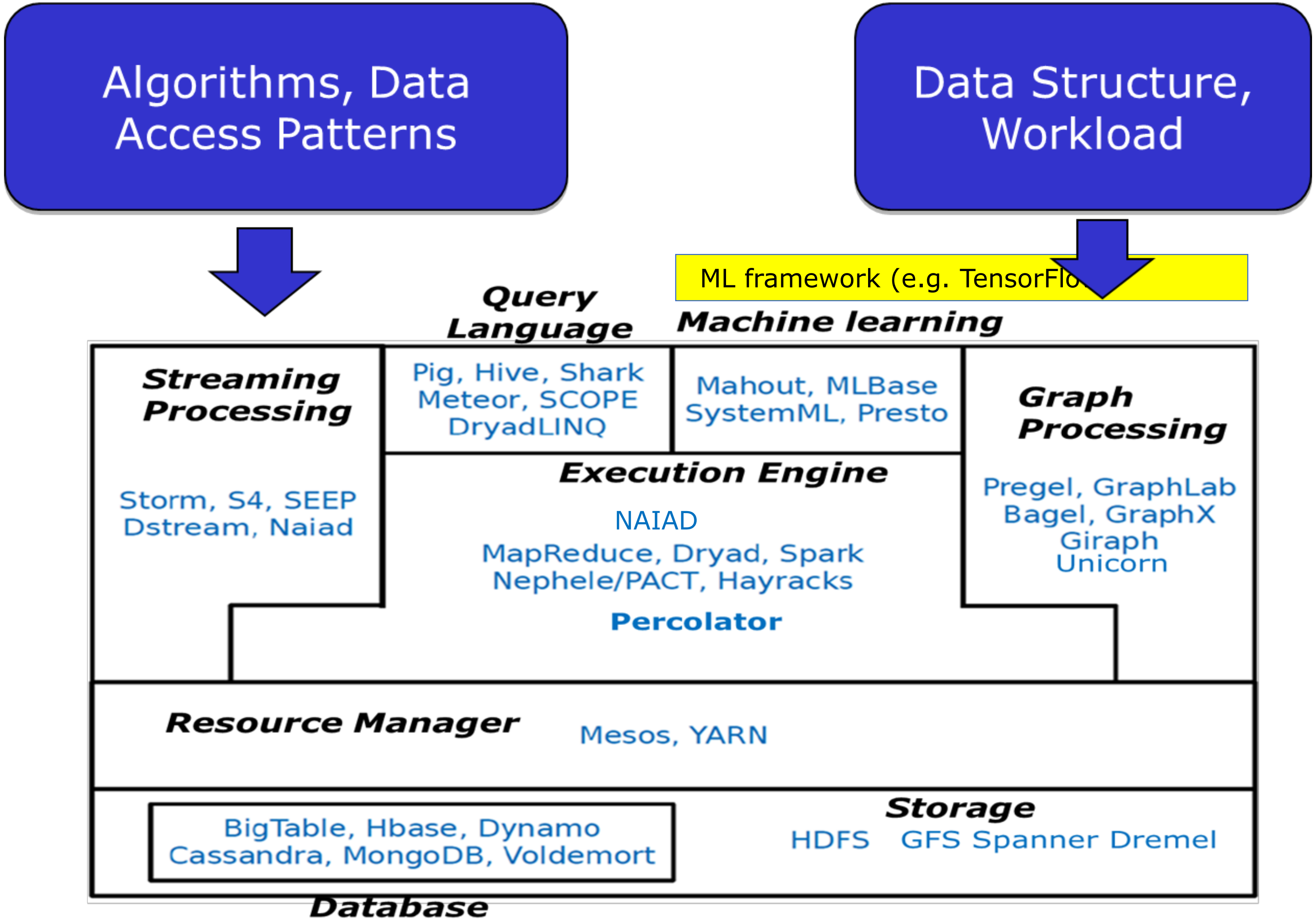
- Inter-disciplinary approach required
  (distributed systems, algorithms, statistics, computer architecture, database…)

Big Data: Technologies

- Distributed infrastructure
  - Cloud (e.g. Infrastructure as a service, Amazon EC2, Google App Engine, Elastic, Azure)

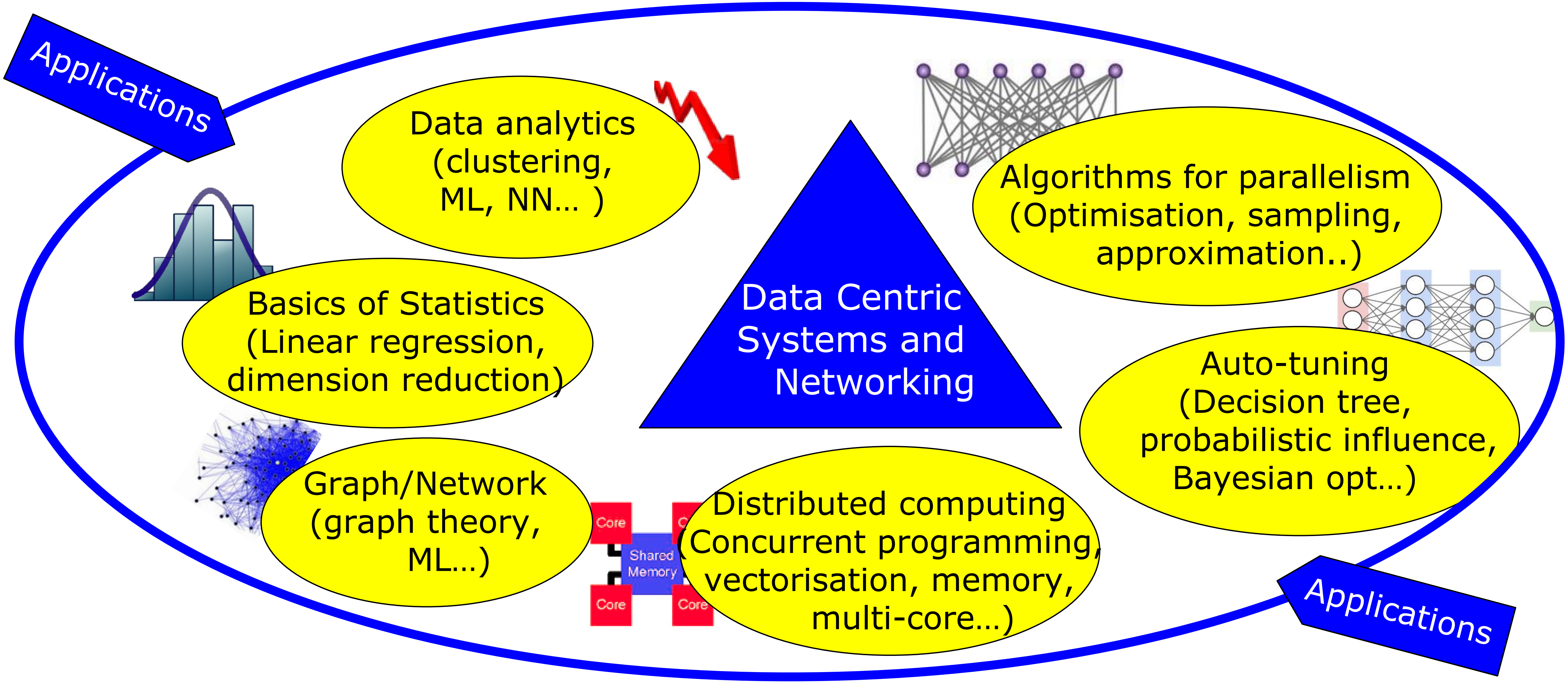
    cf. Multi-core (parallel computing)
- Storage
  - Distributed storage (e.g. Amazon S3, Hadoop Distributed File System (HDFS), Google File System (GFS))
- Data model/indexing
  - High-performance schema-free database (e.g. NoSQL DB - Redis, BigTable, Hbase, Neo4J)
- Programming model
  - Distributed processing (e.g. MapReduce)

Big Data Analytics Stack



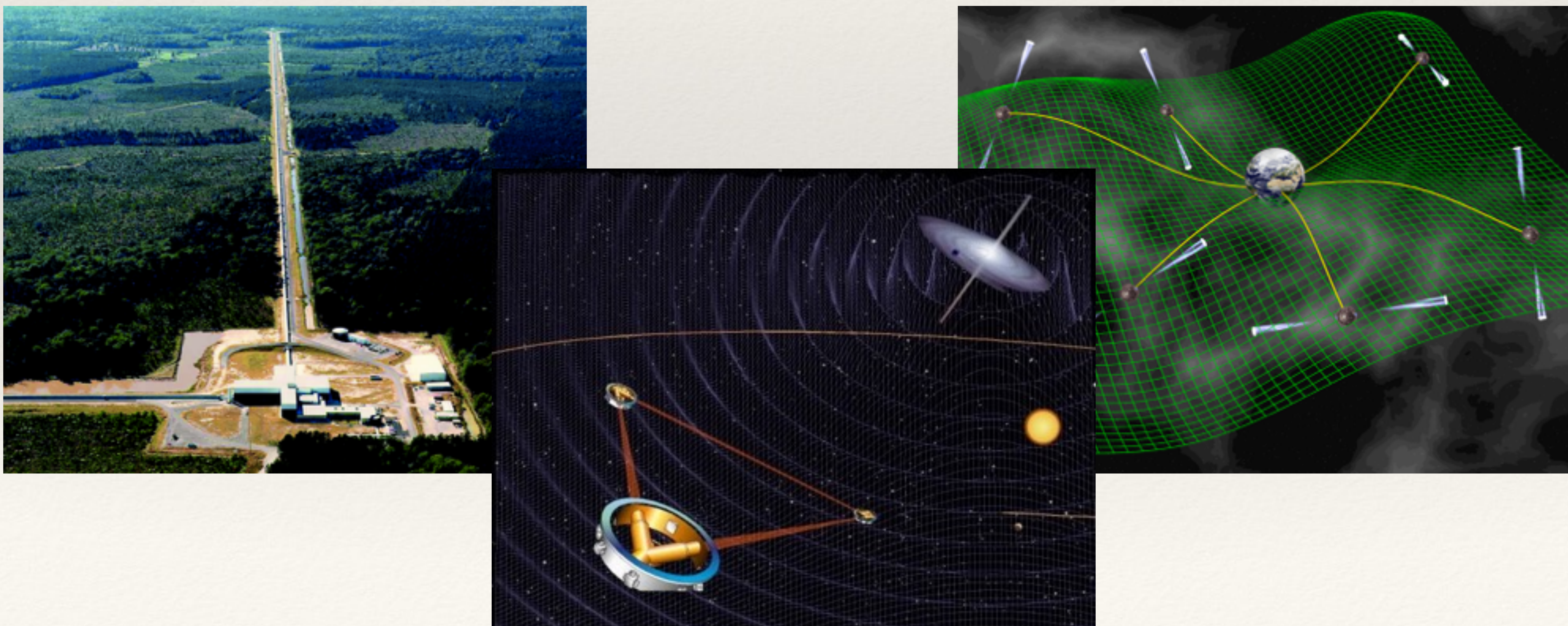
Data Centric Approach for Big Data Generation

- Data is a token in programming flow and networking, and impacts computer system's architecture



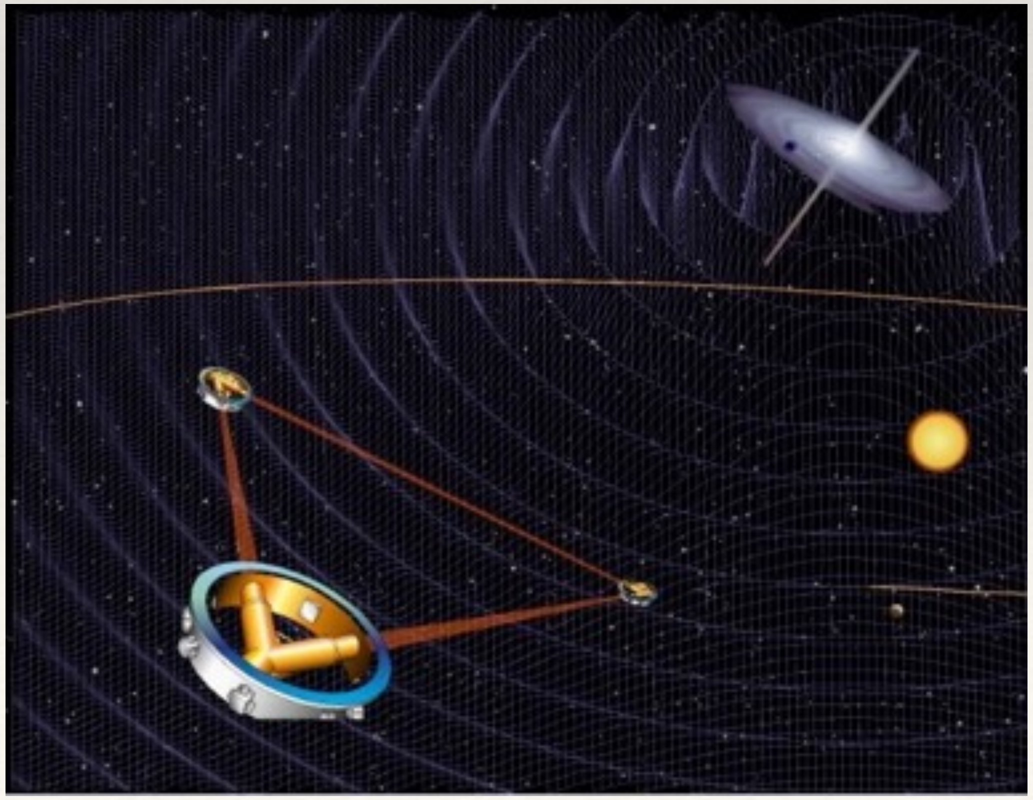
Challenges in data analysis for gravitational wave detectors

Jonathan Gair, School of Mathematics, Univ. of Edinburgh,



Gravitational wave detectors

- A major international effort is underway to detect gravitational waves (GWs) for the first time.
- A ground-based network of kilometre-scale interferometers (LIGO, Virgo etc.) is in the middle of its first observing run with “Advanced” sensitivity.
- Radio telescopes are hunting for nanohertz GWs through precise timing of arrays of millisecond pulsars (PTAs).
- A million-kilometre interferometer in space (eLISA) will be launched by ESA as the L3 mission in the Cosmic Vision programme.

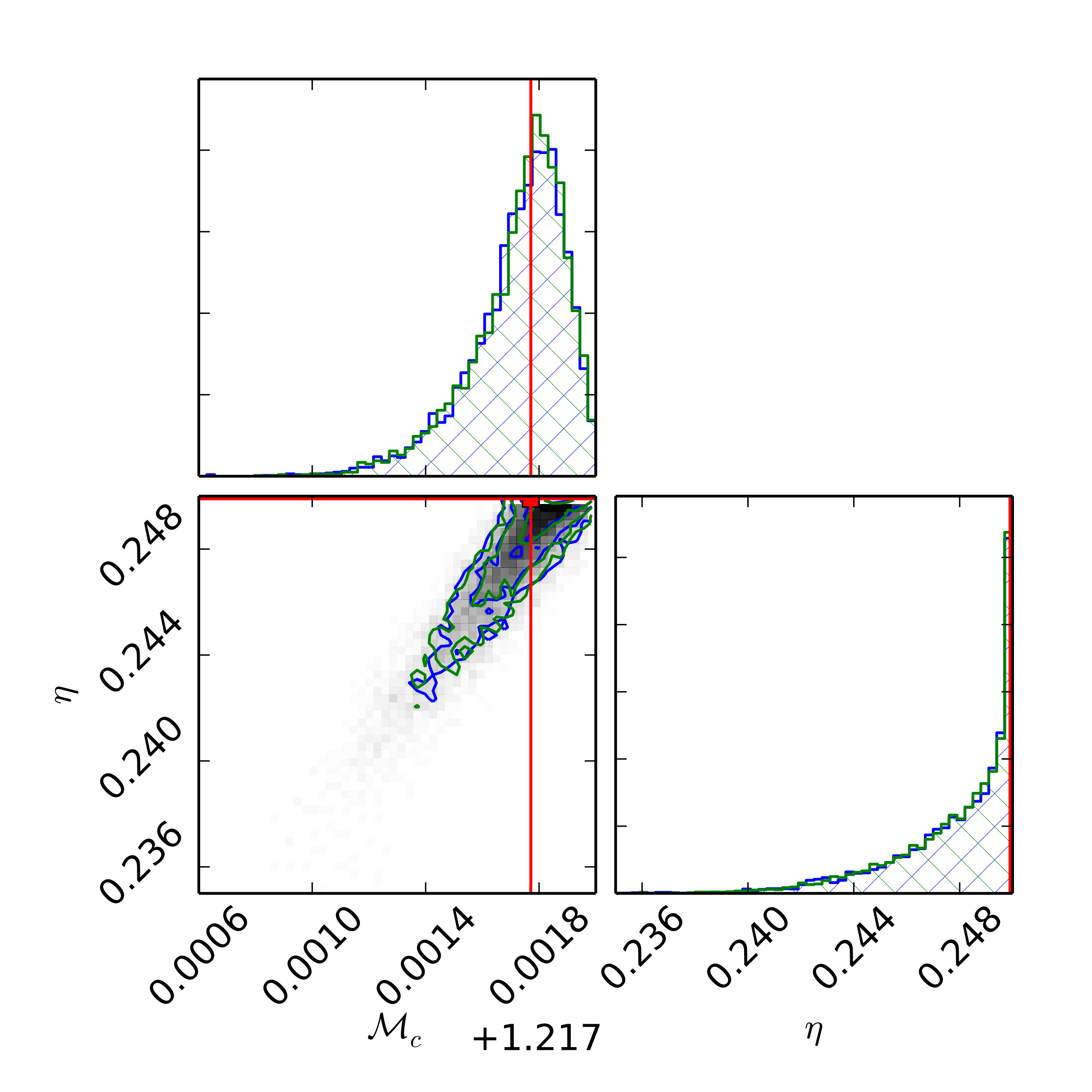


Challenges in data analysis

- These are new experiments and therefore pose new challenges.
- The raw data is not as "big" as that from some other experiments
  - LIGO/Virgo sample at ~4kHz over ~month to ~year observing runs. Terabytes of data from each observing run.
  - eLISA will have a much smaller sampling rate (~1Hz) and therefore three orders of magnitude less data.
  - The data used for GW analysis with PTAs are the residuals for each of ~50 pulsars, measured every ~2 weeks over ~10 years.
- Challenges arise from the complexity of the data and the expected signals.

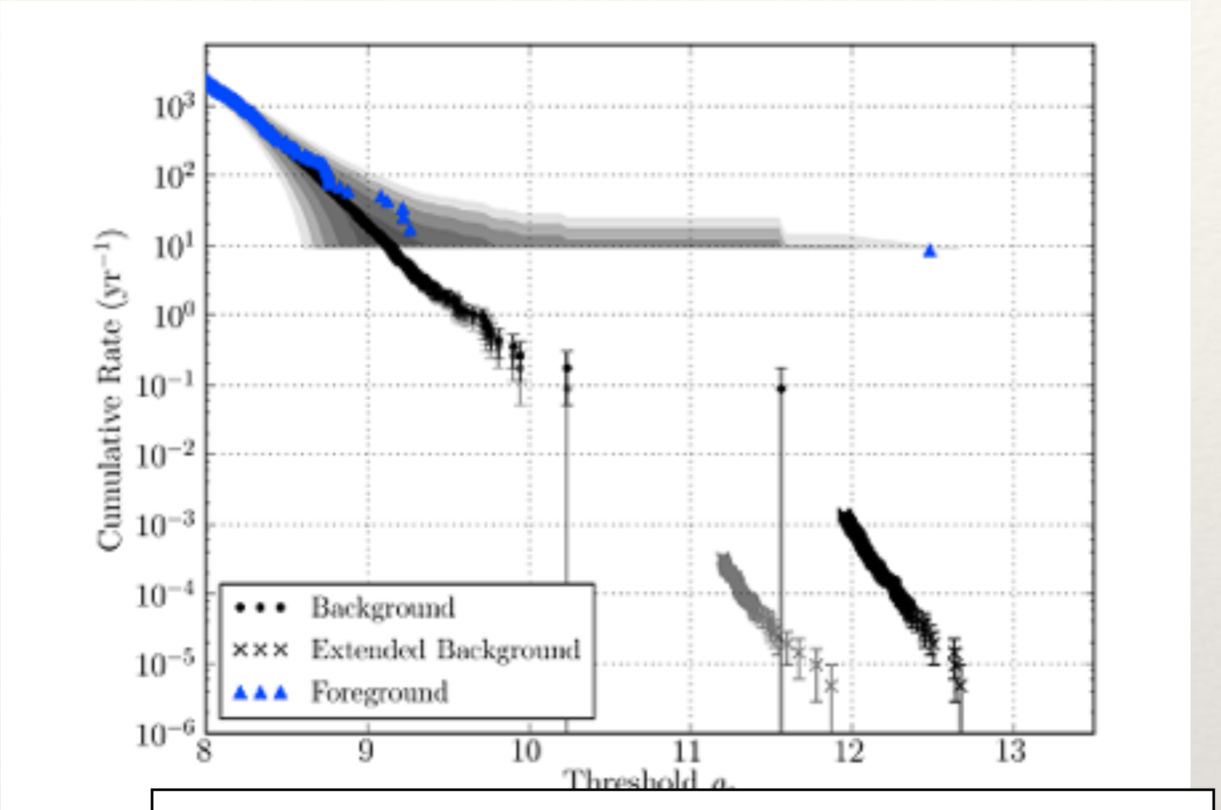
Challenges - large parameter spaces

- Many searches are performed on LIGO data, targeted at different source classes
  - Low-latency for rapid follow-up.
  - Modelled sources - binaries of different types, continuous waves etc.
  - Un-modelled sources (bursts), both targeted (GRBs or SNe) and un-targeted.
  - Stochastic background.
- The eLISA data will contain thousands of sources that overlap in time and frequency, creating a confusion problem.
- Each source is characterised by ~10 parameters that must be estimated.

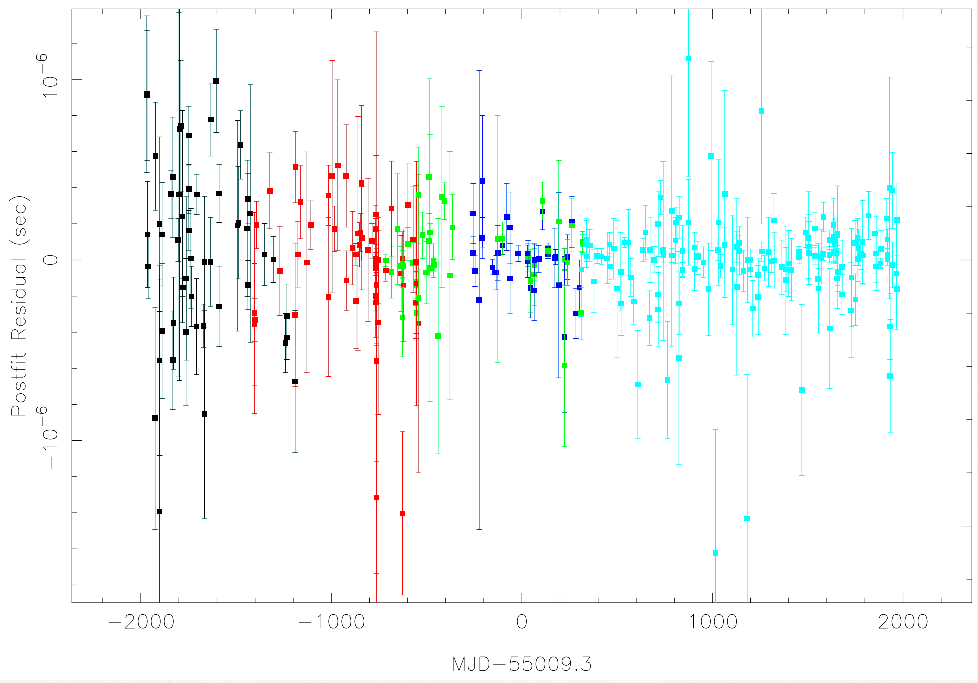


Challenges - noise characterisation

- The statistical properties of the noise in GW detectors is poorly understood.
- Background typically estimated by “timeslides”. Need to repeat analysis many times to reach desired significance level.
- For PTAs, data is collected over many years with many different instruments that have different noise properties.
- Need to fold noise measurement uncertainty into parameter estimation.

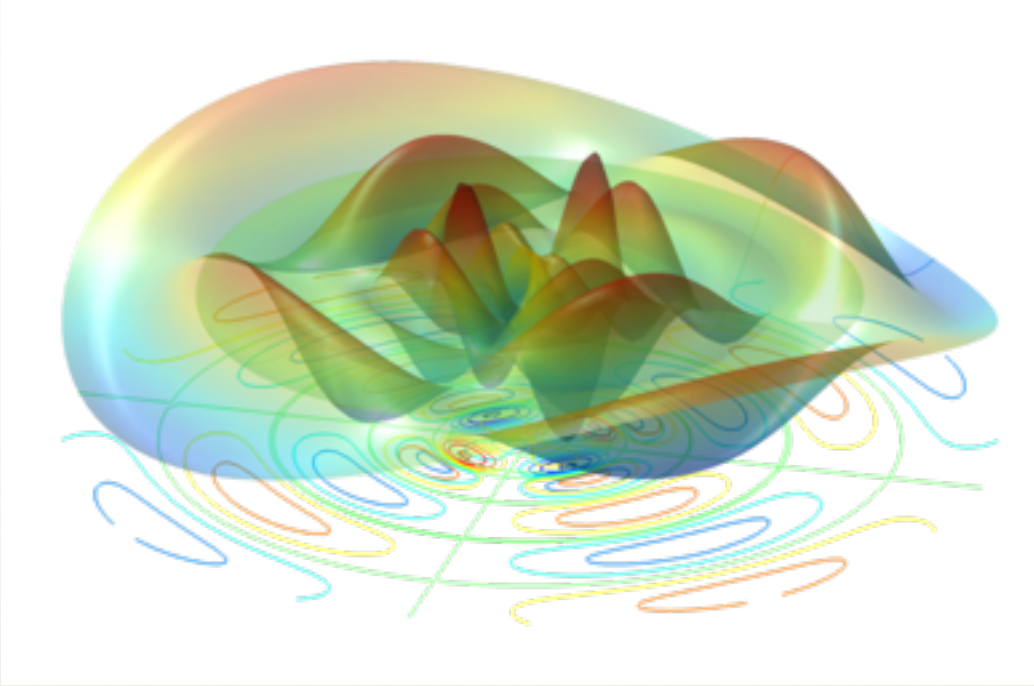


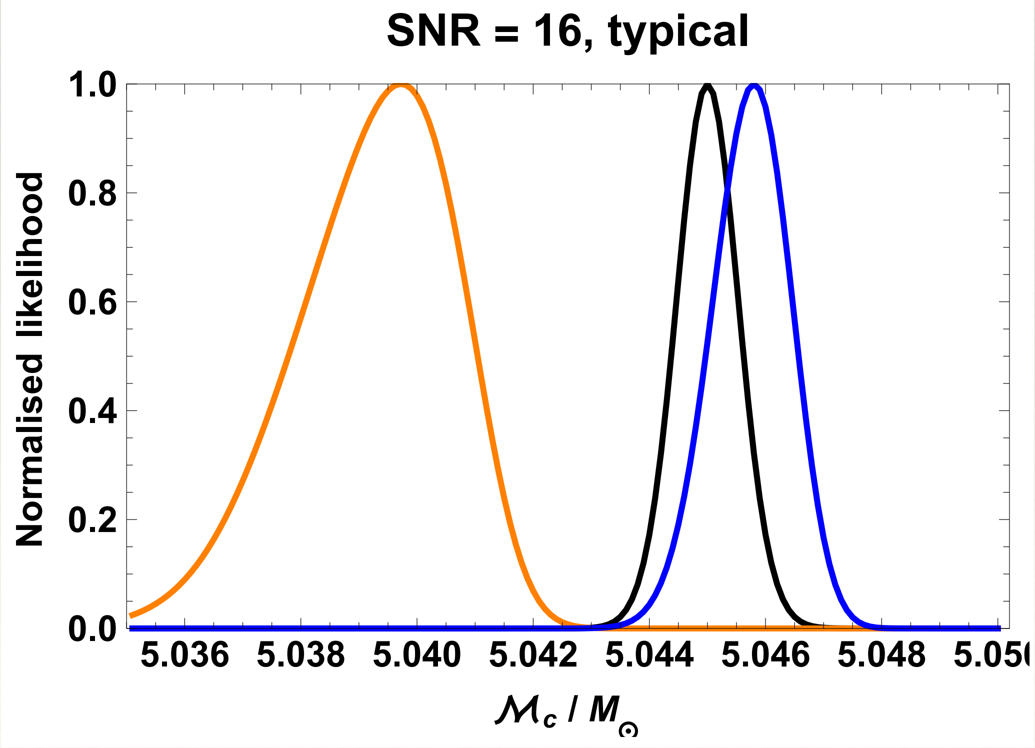
Aasi et al. (LVC), PRD **85** 082002 (2012)

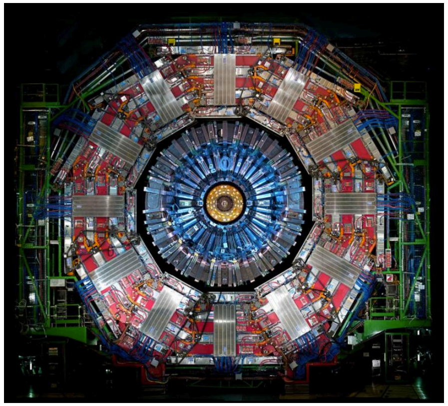


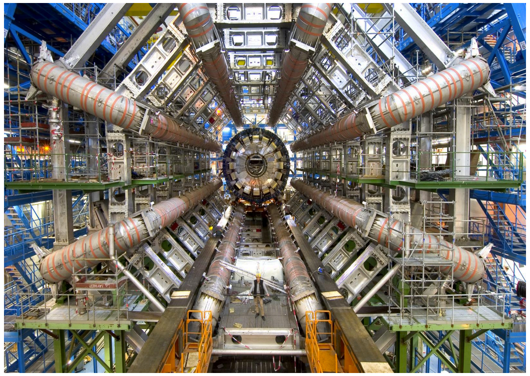
Challenges - complex signal models

- Signal models are complex and expensive to evaluate numerically.
- Inference relies on approximations.
- Bias from approximation must be folded into parameter estimation results.
- One promising approach: Gaussian process regression.
- Building the Gaussian process model is challenging, and introduces additional parameters that must be estimated or marginalised over.



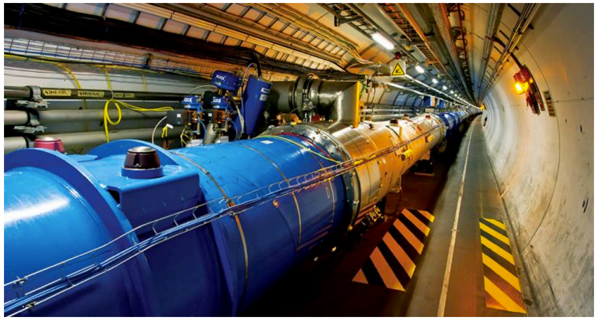






'Big Data' at the Large Hadron Collider

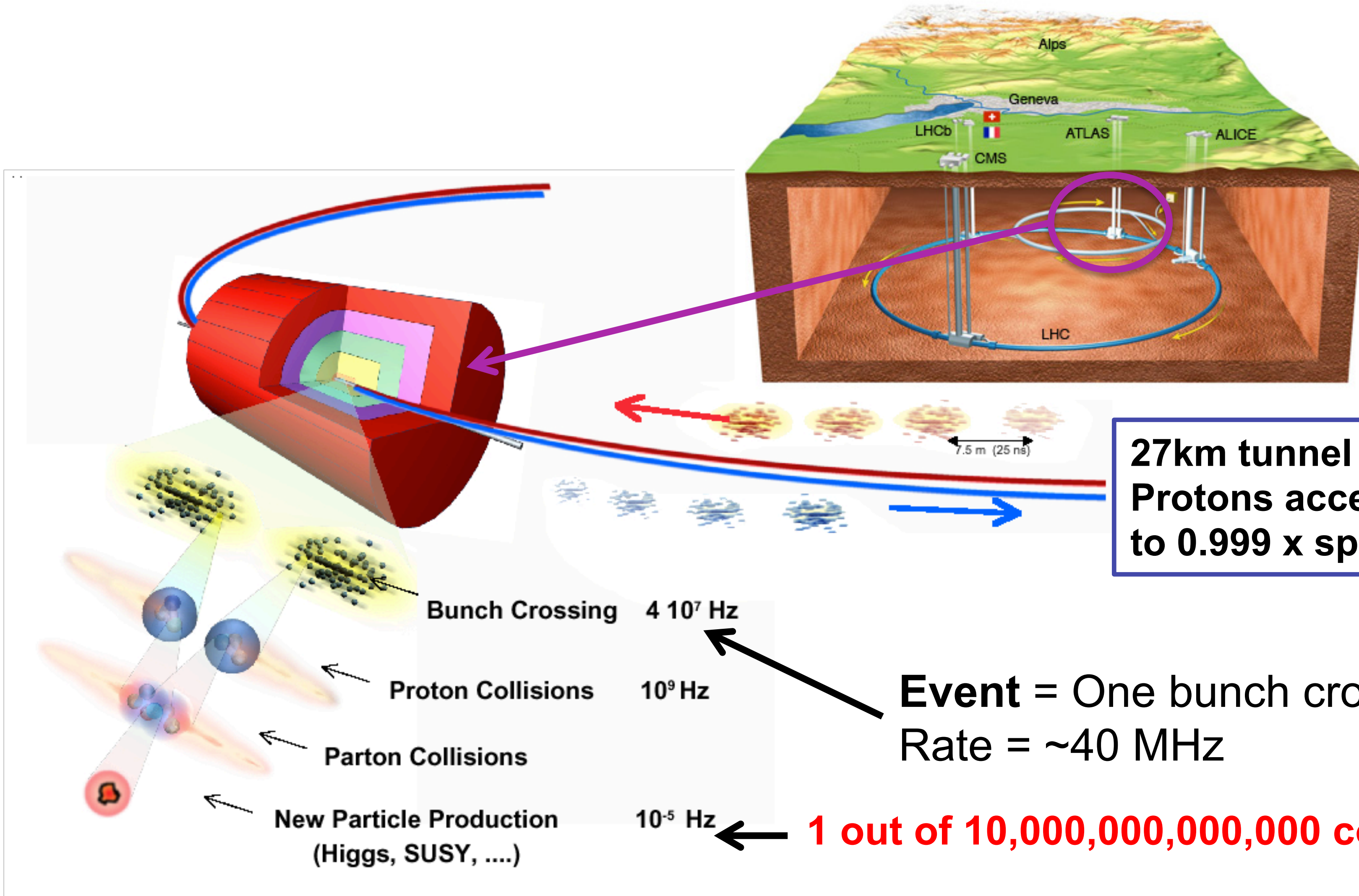
Tim Scanlon
University College London



UCL

Large Hadron Collider

Study the fundamental particles and forces of the Universe



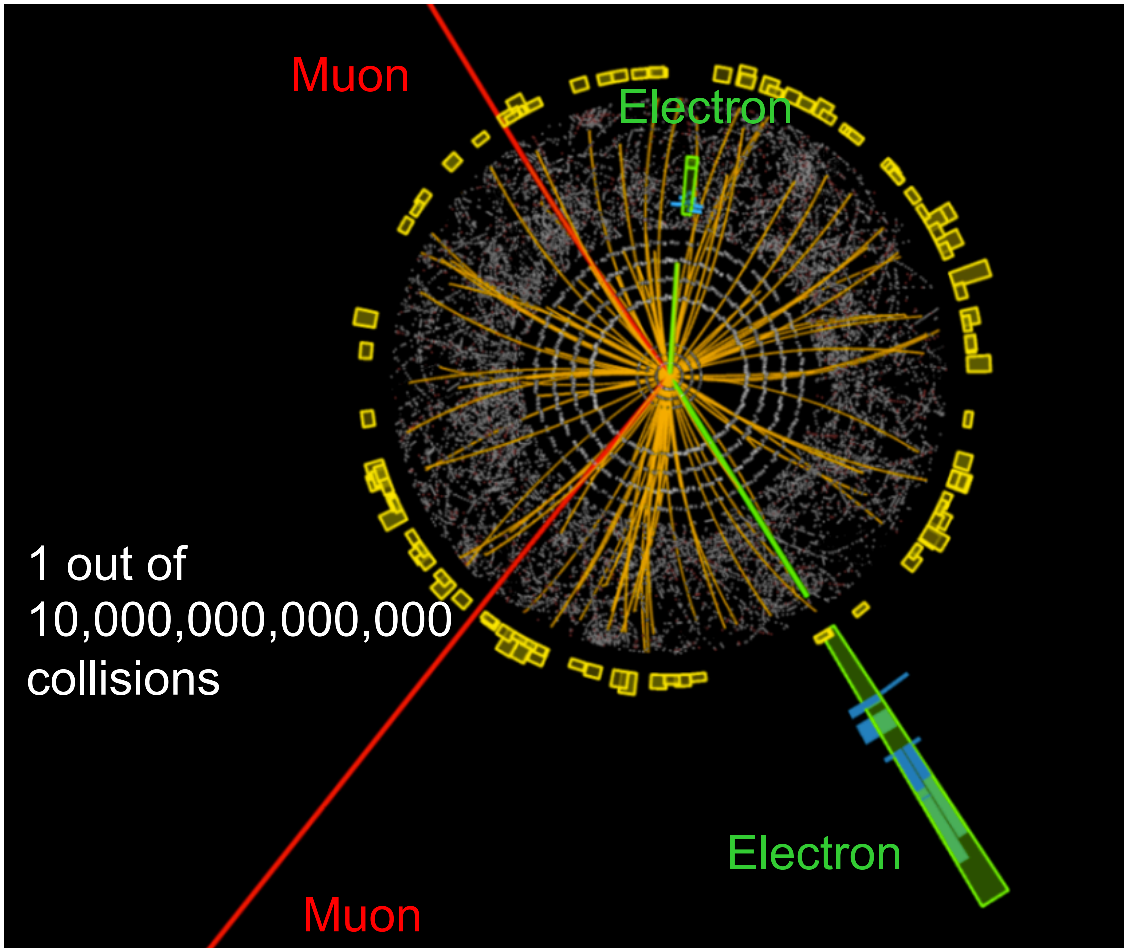
27km tunnel
Protons accelerated
to 0.999 x speed of light

Event = One bunch crossing
Rate = ~40 MHz

1 out of 10,000,000,000,000 collisions

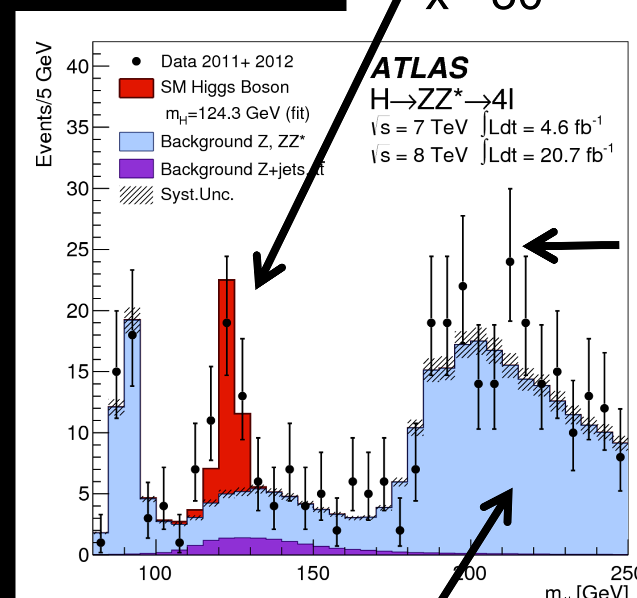
Identifying a Higgs(?) Event

Identify then combine: 2 x Muons + 2 x Electrons



1 out of
10,000,000,000,000
collisions

x ~30

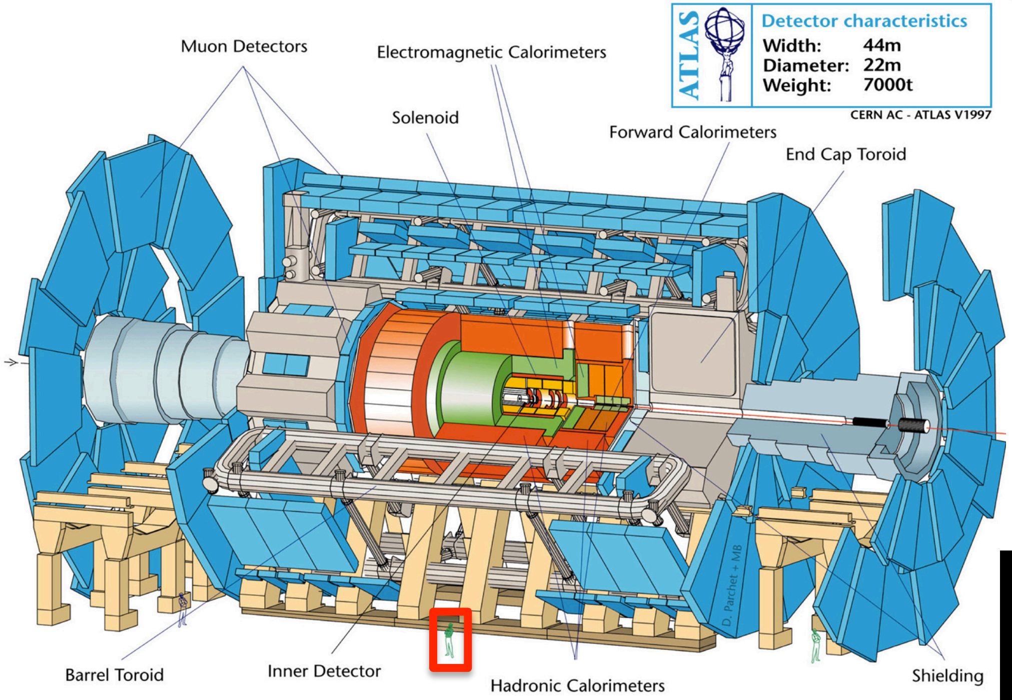


Real data

Simulated Events

Challenge to identify the **particles** and **events**

Reconstructing a Collision



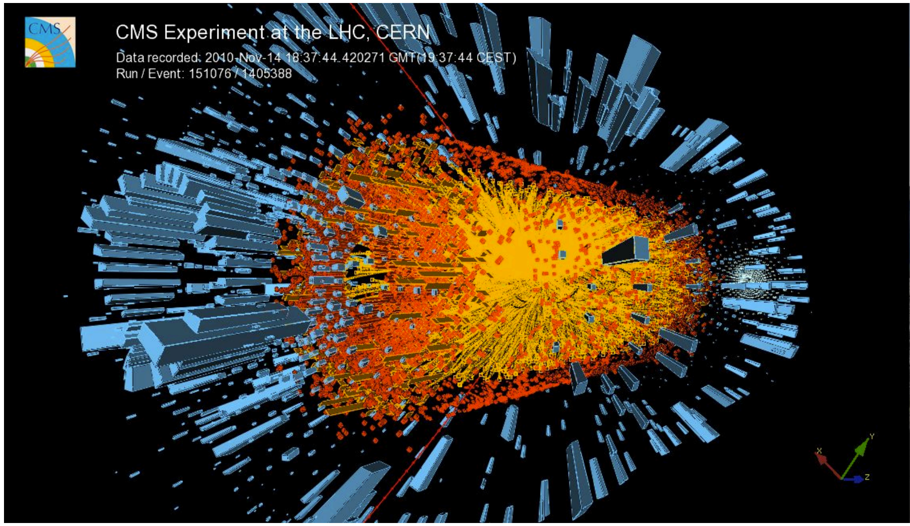
Vast 'data-creating machines'

- **Size of a six storey building**
- **160M readout channels**
- **Creates 1 PB/s data**

First level filter keeps only **~1% of events**

Complicated algorithms reconstruct collisions

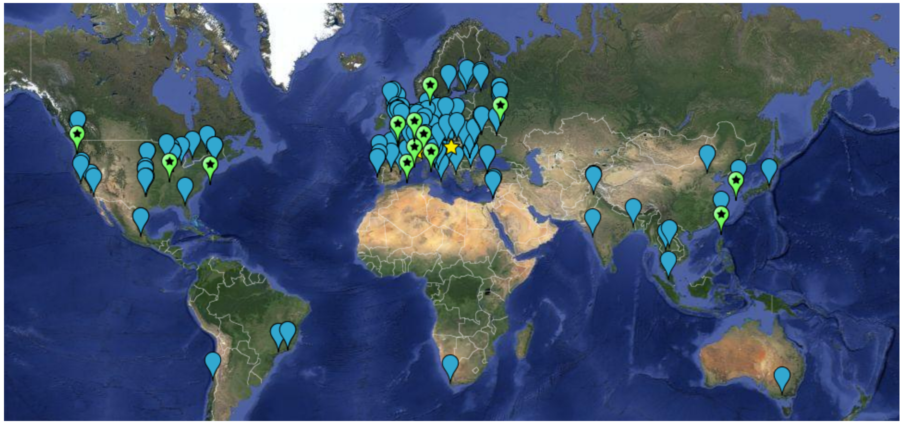
- Use **100k CPU farm** at CERN
- Can take up to **20s CPU time**

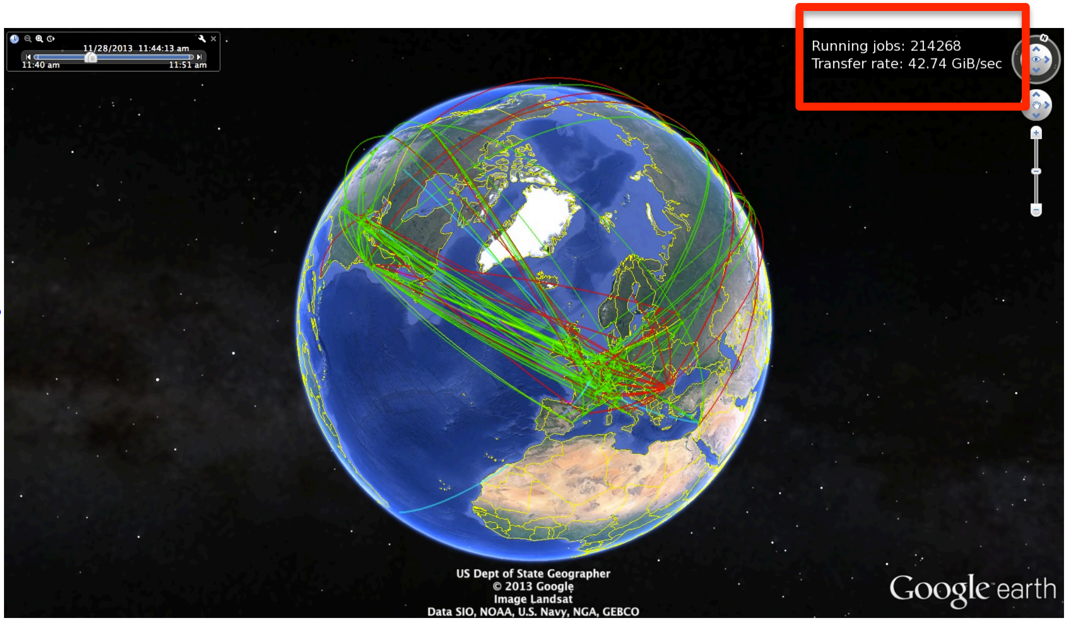


Overall reduction in data by factor of 10^6….
... still huge volumes of data (1 GB/s) and events (~billions)

Worldwide LHC Computing Grid

- The data challenge
  - **30M GB of data** per year from LHC
  - **Billions of events**
  - **10,000 physicists** worldwide
    - Need real-time access to this data
    - Shared computing resources





Worldwide LHC Computing Grid
- **42 countries**
- **170 computing centres**
- **2 million jobs run a day**

Outsourcing to home users!

LHC@home
Volunteer computing for the LHC

‘The most sophisticated data-taking and analysis system ever built for science’

Machine Learning (ML)

- Many challenges ideal for machine learning
  - ➢ Identification of **particles**
  - ➢ Selection of **signal events**

Widely used with large performance increases achieved

- ML techniques used from 90s
  - ➢ Mostly **Neural Networks (NN)** and **Boosted Decision Trees (BDT)**
  - ➢ Investigating newer techniques: Deep Learning NNs
  - ➢ Tool kit: Use TMVA/Root framework

- Outsource to ML enthusiasts
  - Discover more effective ML methods!
  - Engage people in fundamental research

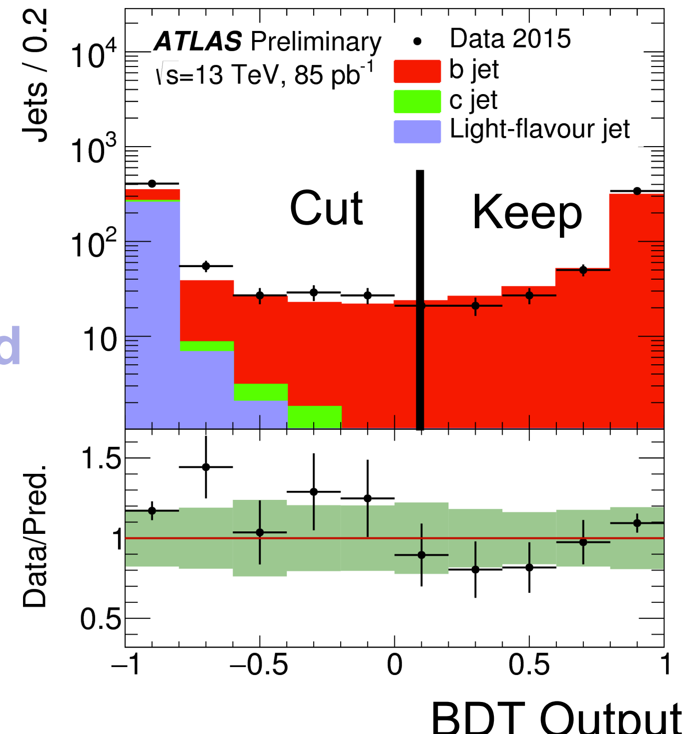
1785 Teams
1942 Players
35772 Submissions

Analysis Challenges

- Use ML to identify both **particles** and **events** of interest
  - A lot of tuning: parameters, variables, algorithms etc.
  - No data 'standard candles' for training/modelling – use simulation
    - **Need to ensure variables and correlations are well modelled**
    - **Extra uncertainties**
    - **Limited statistics**

Categorise
- **Signal**
- **Background**

Uncertainties on modelling of real data by simulation

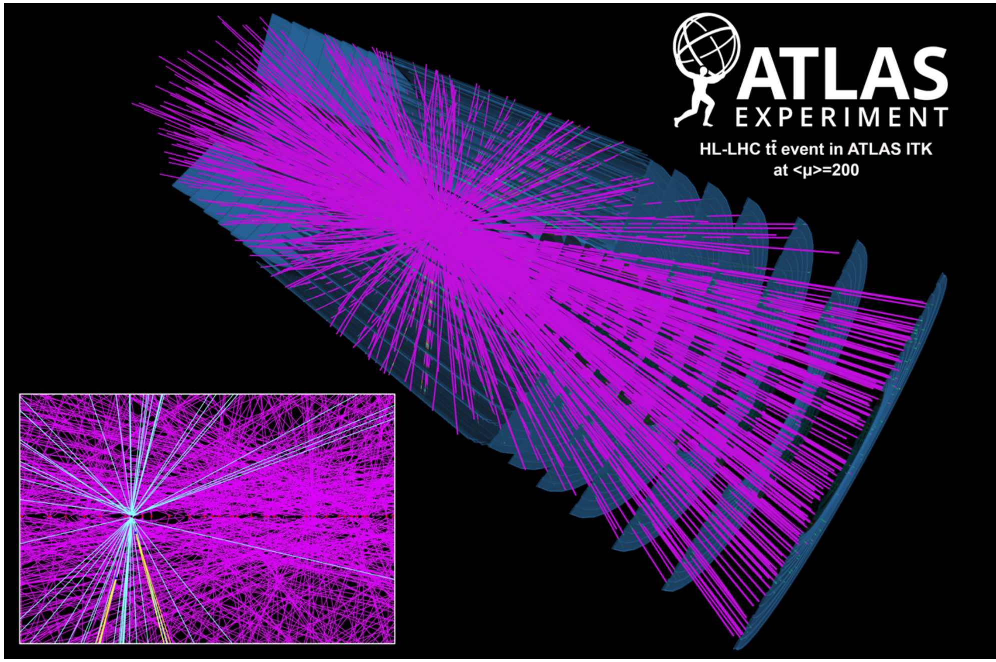


- Finally: advanced statistical techniques to quantify significances
  - **Profile likelihoods**, **Bayesian analyses**

Summary

- Many big data challenges at the LHC
  - Huge amounts of data/events, complicated algorithmic problems, difficult classification problems

~200 collisions per event



- Cutting edge tools adapted
  - Worldwide LHC Computing Grid
  - Complex reconstruction algorithms
  - ML techniques
- Greater challenges ahead
  - Data x 100
  - Event complexity x10
  - Ensure we fully exploit the data
- Collaboration between fields important to meet these challenge
  - Share experience and expertise
  - Common and improved tools
  - Fully exploit cutting-edge techniques

THE ALAN TURING INSTITUTE

Many Data: few numbers
Many Data: many numbers

Alan Heavens

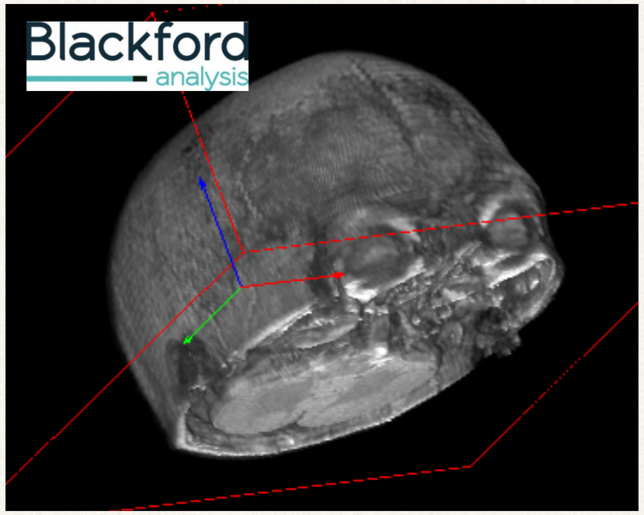
Imperial Centre for Inference and Cosmology

Imperial College London

Data? Numbers?

- **Framework:**
- **Data** interpreted in context of a **Model**
- Model has **parameters**: these are the **numbers**
- We want to know the numbers

Many Data: few numbers



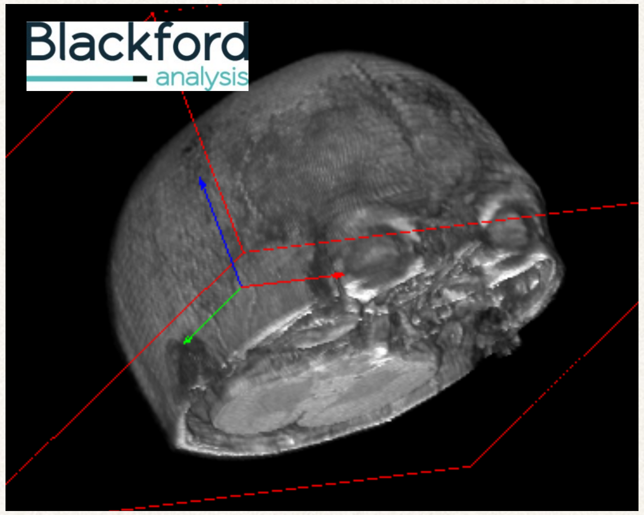
MRI scan:

512x512x100

26 million voxels

Many Data: few numbers

- **Model:** two volumetric images are (almost) the same, but rotated, shifted



MRI scan:

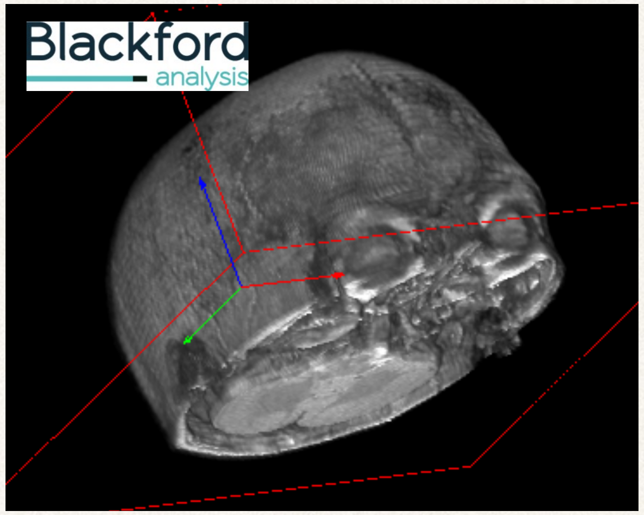
512x512x100

26 million voxels

Many Data: few numbers

- **Model:** two volumetric images are (almost) the same, but rotated, shifted

- **Data:** MRI voxel intensities



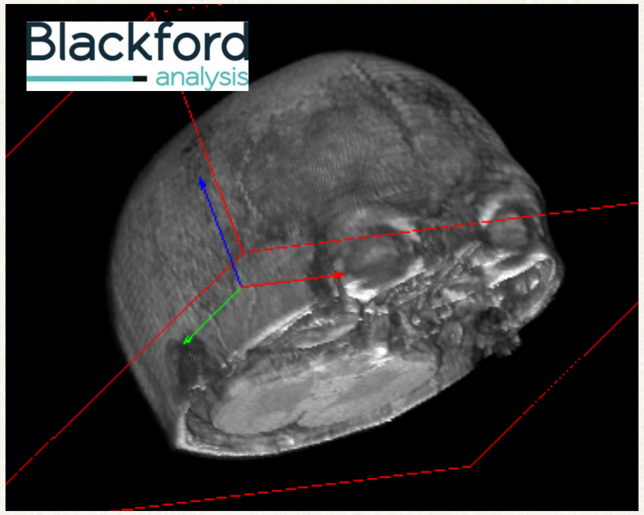
MRI scan:

512x512x100

26 million voxels

Many Data: few numbers

- **Model:** two volumetric images are (almost) the same, but rotated, shifted
- **Data:** MRI voxel intensities
- **Model parameters**: 3 rotations, 3 translations



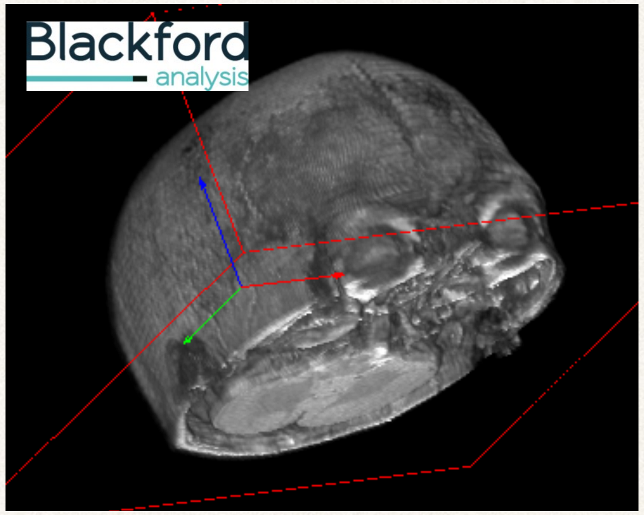
MRI scan:

512x512x100

26 million voxels

Many Data: few numbers

- **Model:** two volumetric images are (almost) the same, but rotated, shifted
- **Data:** MRI voxel intensities
- **Model parameters**: 3 rotations, 3 translations

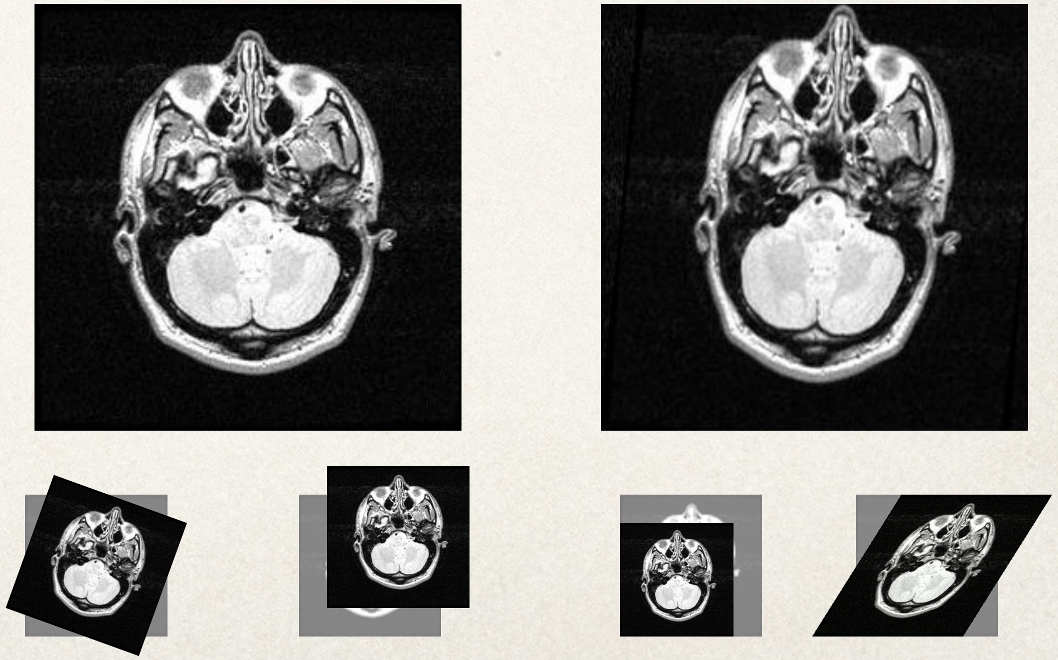


MRI scan:

512x512x100

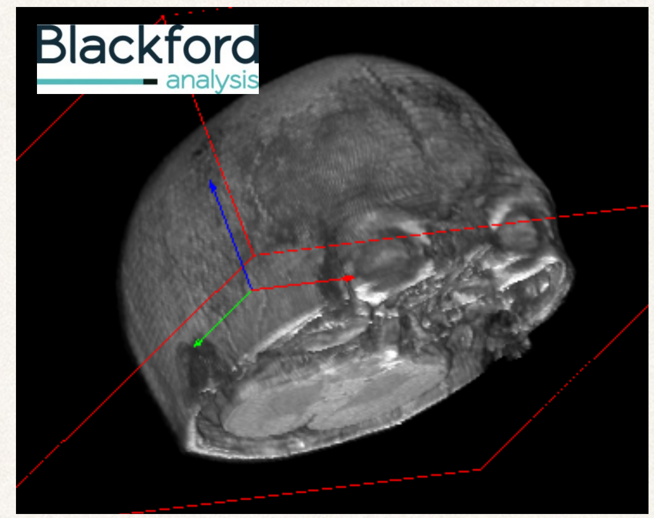
26 million voxels

Image Distortions



Many Data: few numbers

- **Model:** two volumetric images are (almost) the same, but rotated, shifted
- **Data:** MRI voxel intensities
- **Model parameters**: 3 rotations, 3 translations
- **26 Million Data: 6 numbers**

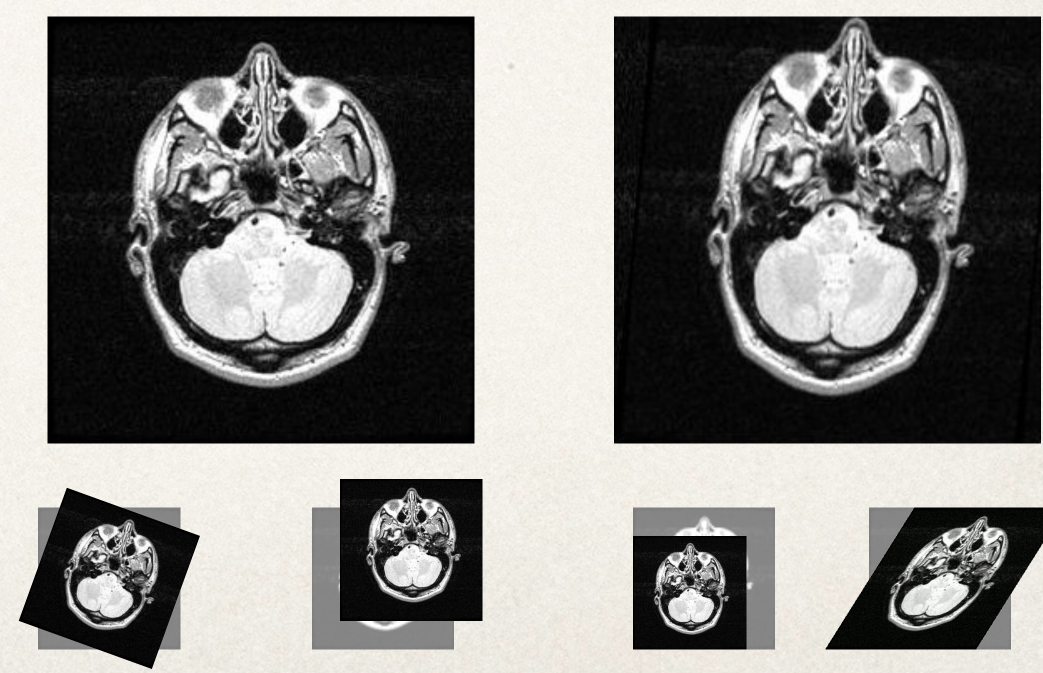


MRI scan:

512x512x100

26 million voxels

Image Distortions



ICIC

Many Data: few numbers

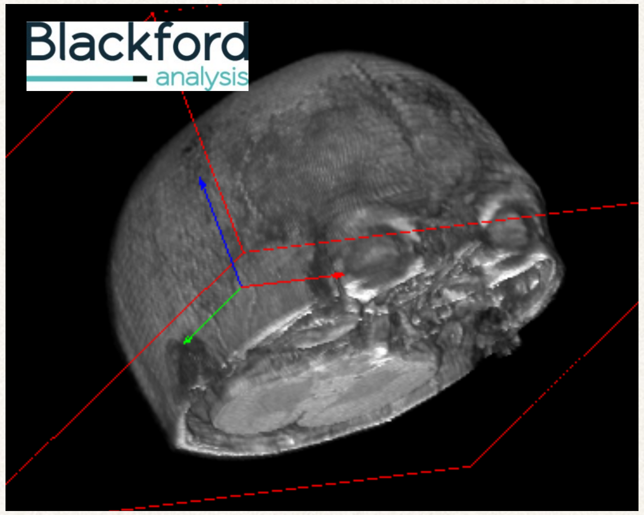
- **Model:** two volumetric images are (almost) the same, but rotated, shifted

- **Data:** MRI voxel intensities

- **Model parameters**: 3 rotations, 3 translations

- **26 Million Data: 6 numbers**

- **MOPED algorithm** **(Heavens et al 2000)**
  Compresses 26 million numbers into 6 (or 12) with no loss of precision

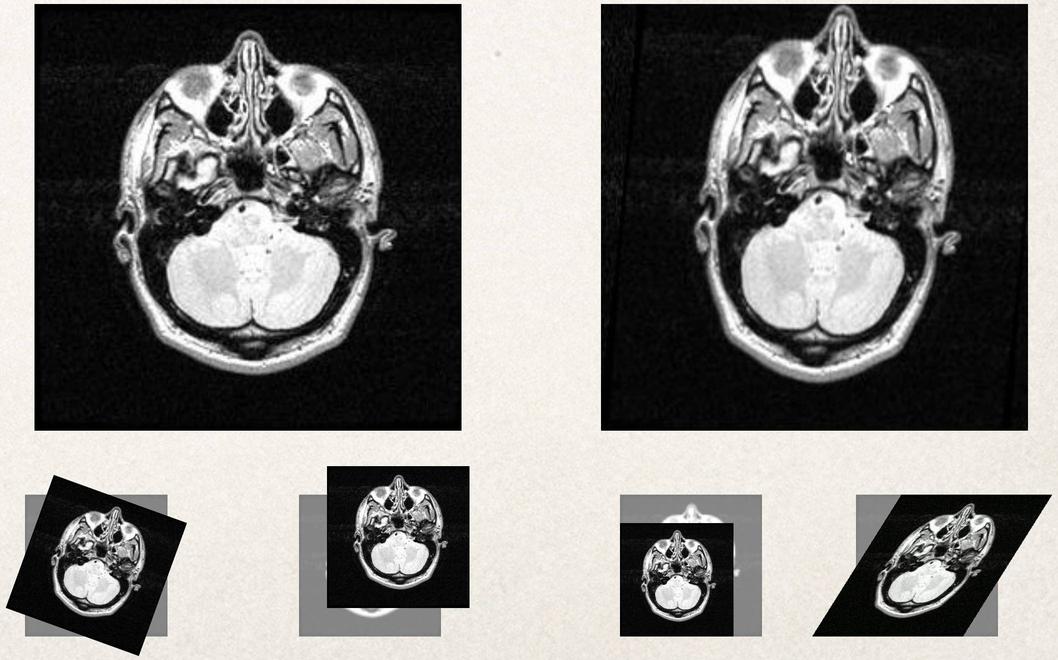


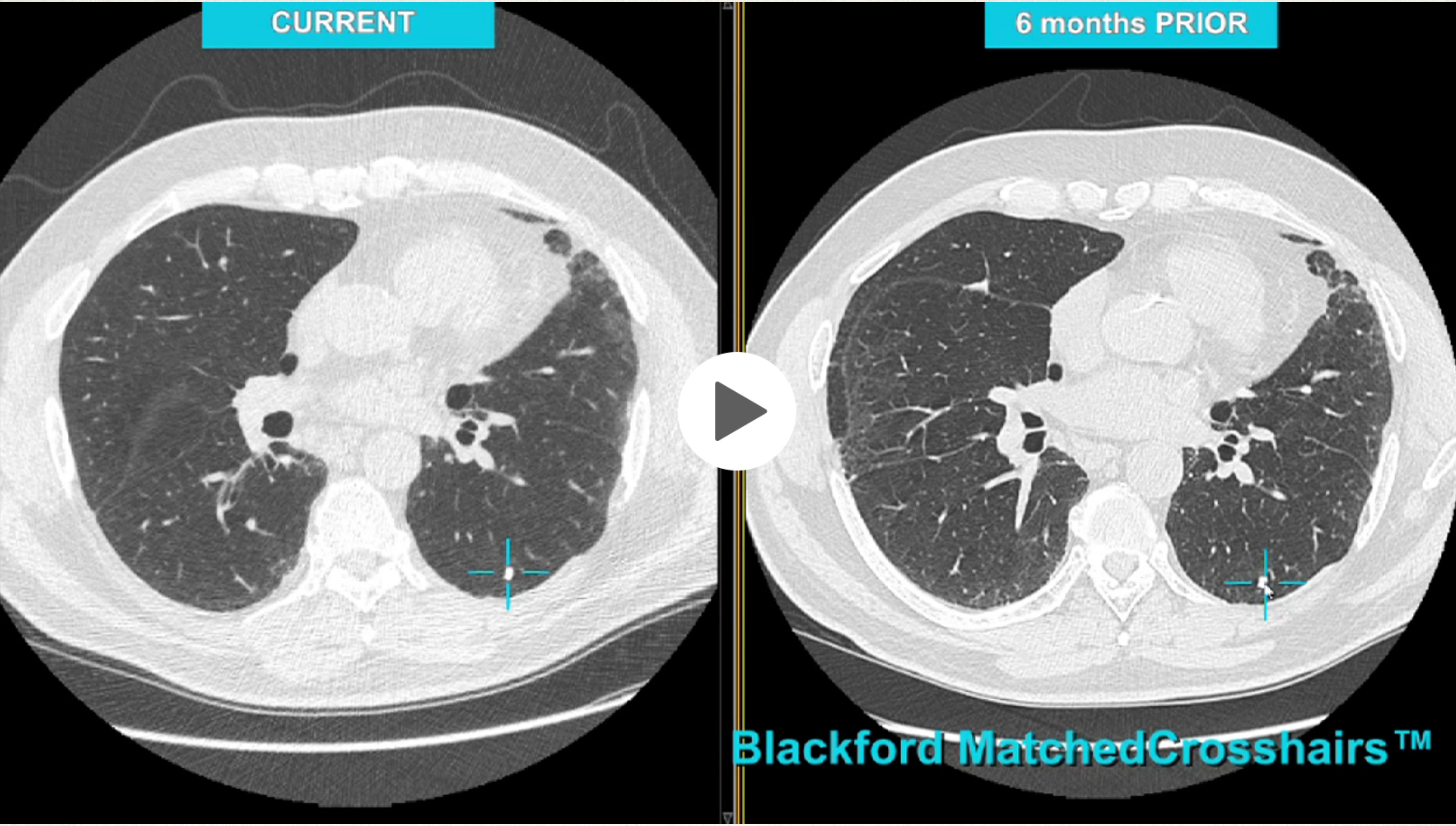
MRI scan:

512x512x100

26 million voxels

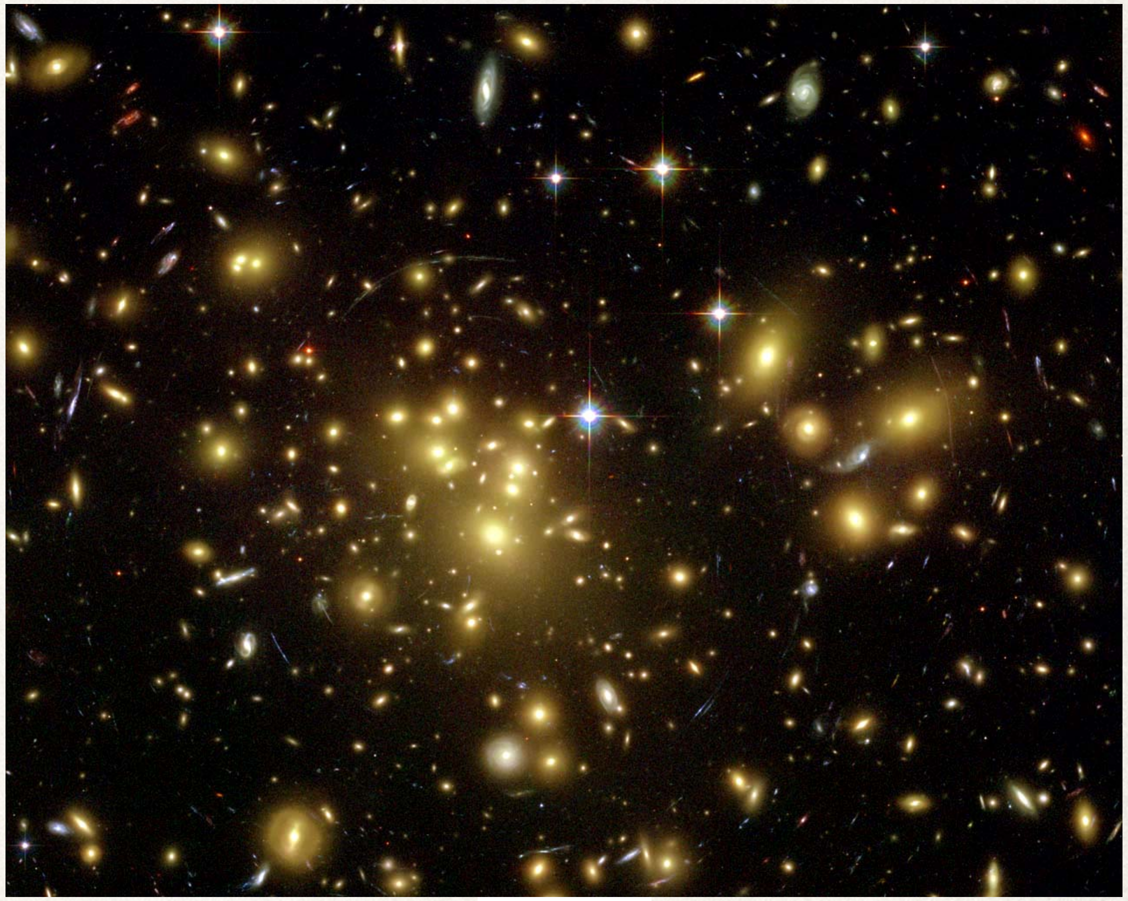
Image Distortions

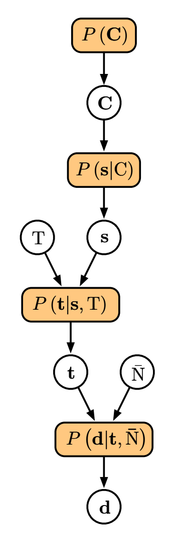




Blackford
analysis

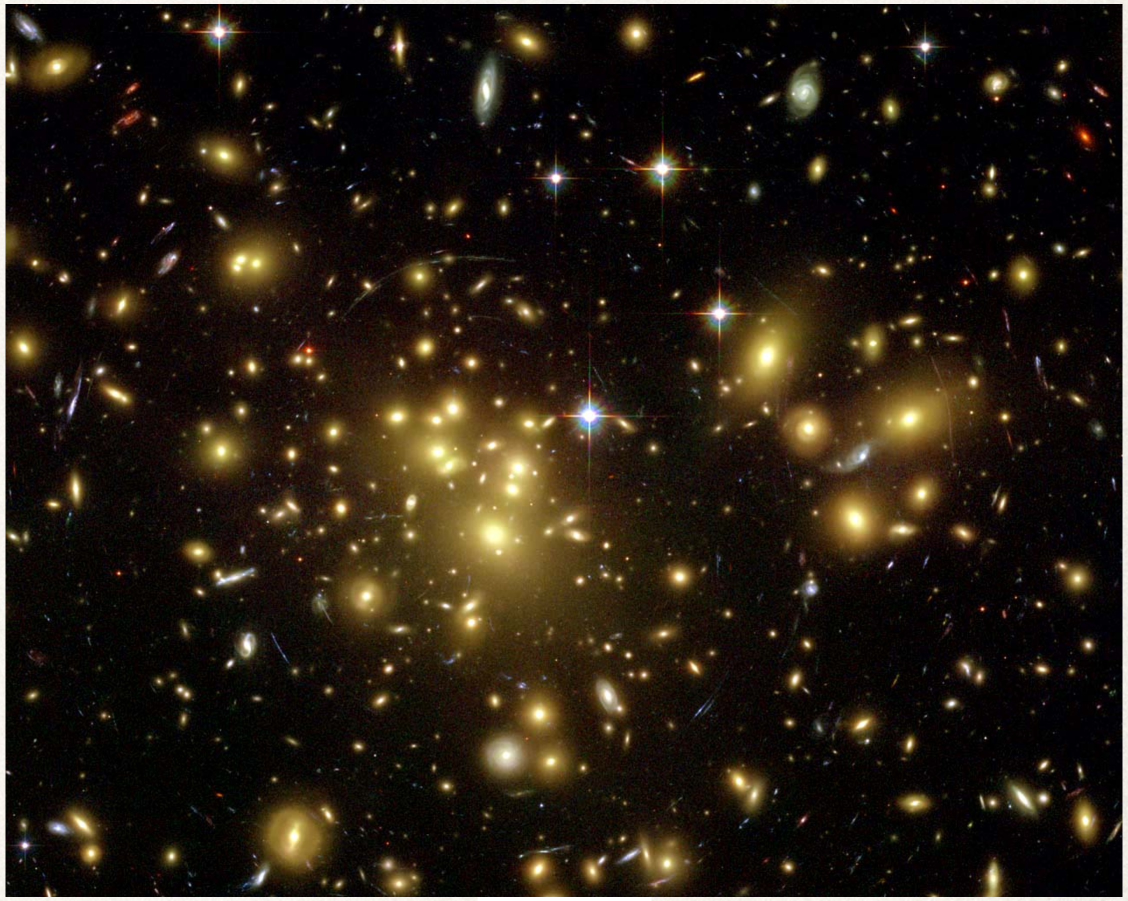
Many Data: many numbers

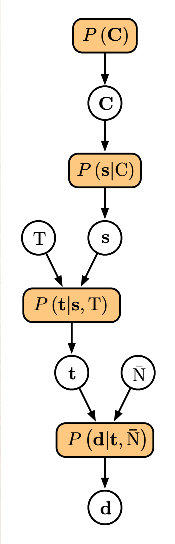




Many Data: many numbers

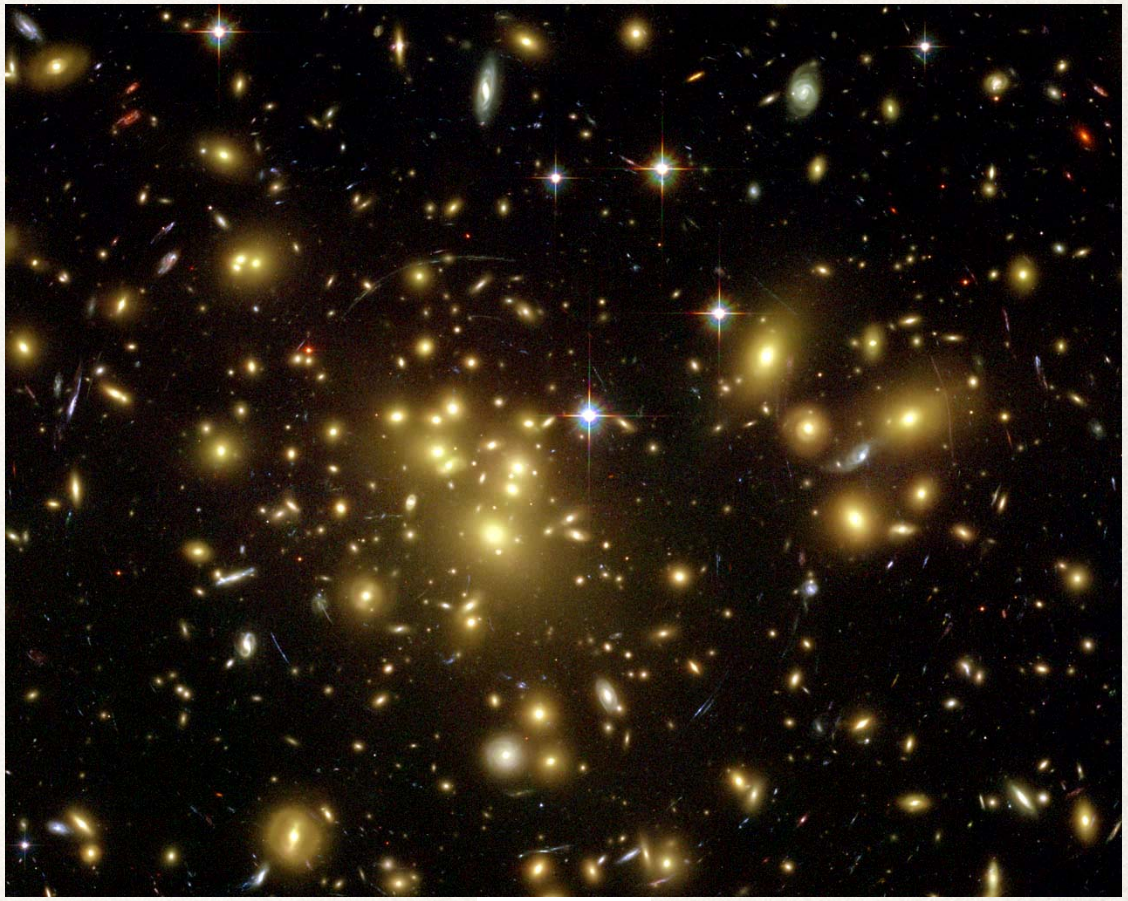
- **Model:** General Relativity ➢ Mass bends light

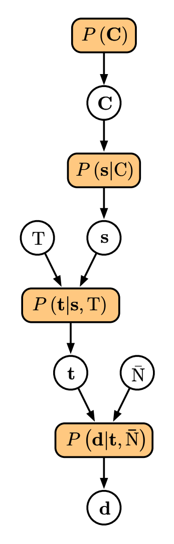




Many Data: many numbers

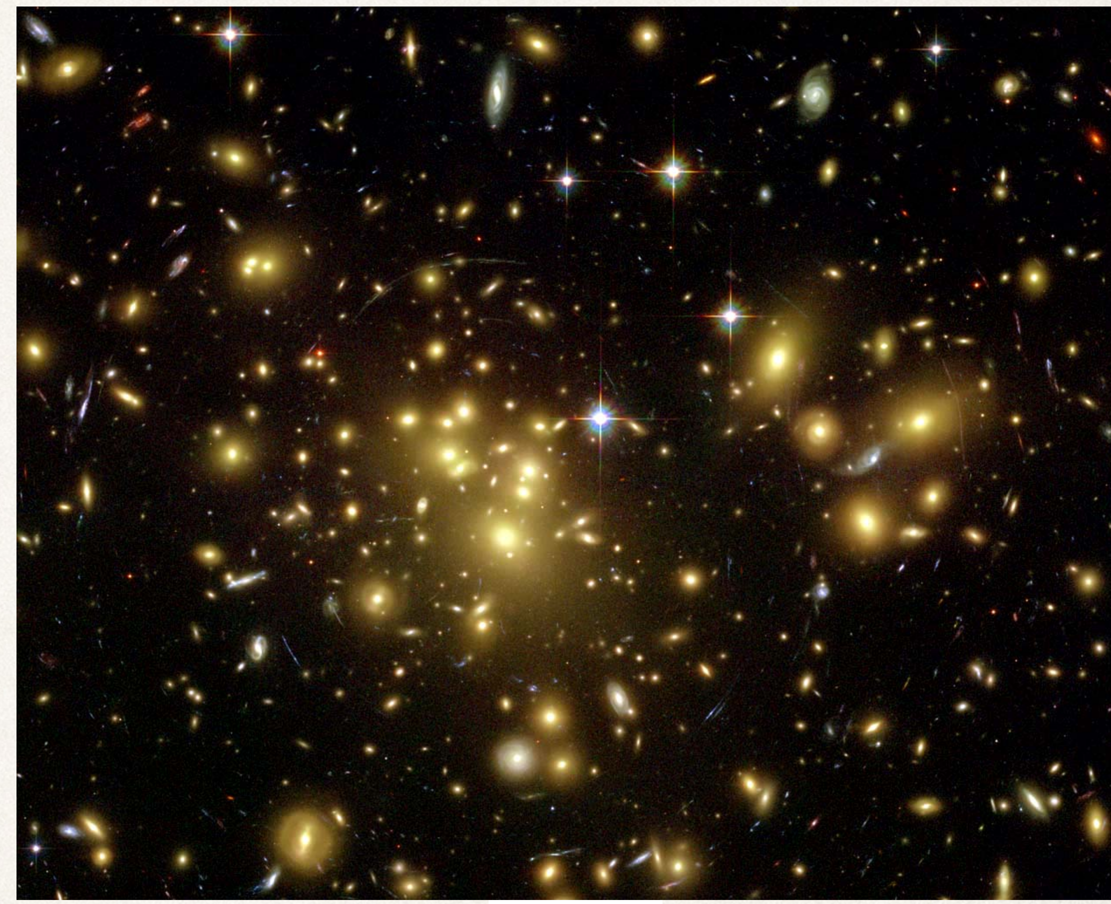
- **Model:** General Relativity ➢ Mass bends light
- **Data:** image distortions (Millions)

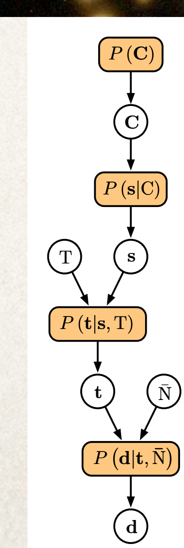




Many Data: many numbers

- **Model:** General Relativity ➢ Mass bends light
- **Data:** image distortions (Millions)
- **Model parameters:** mass distribution (>100,000 numbers)

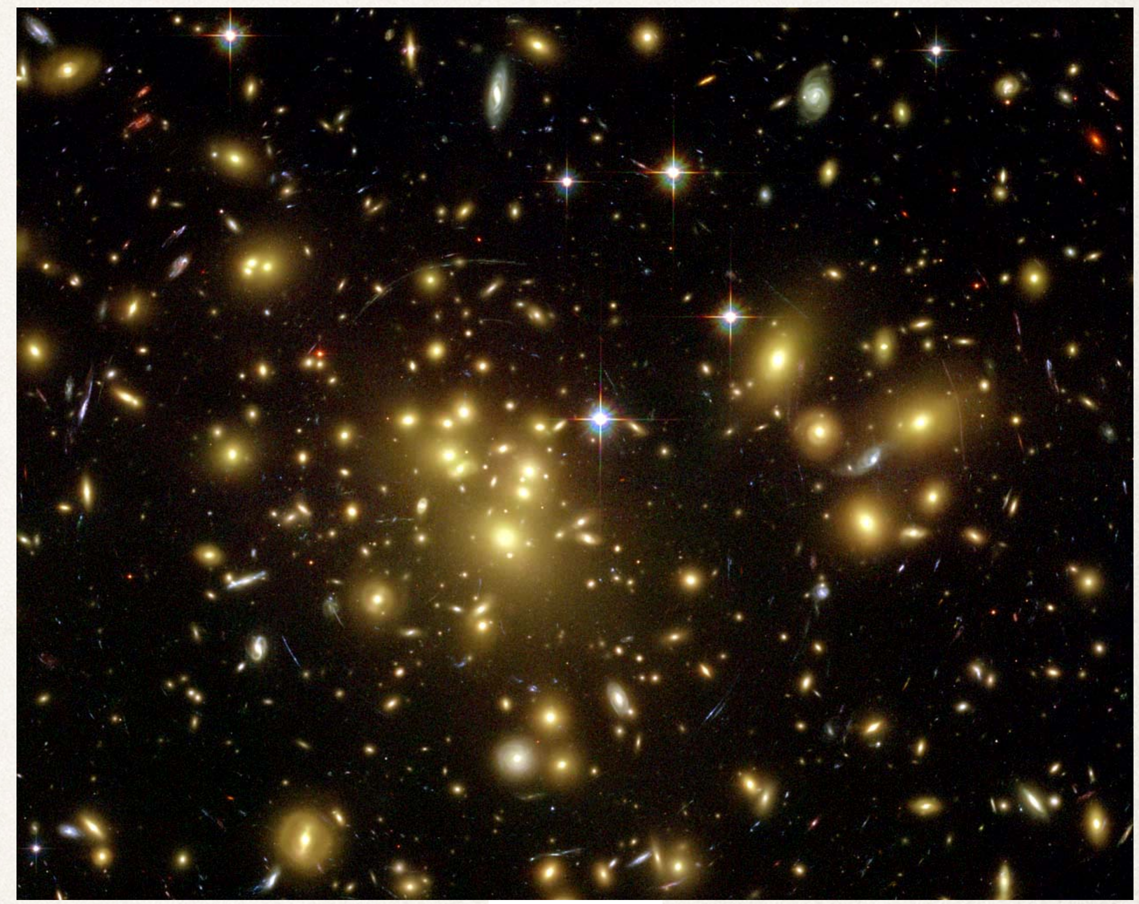


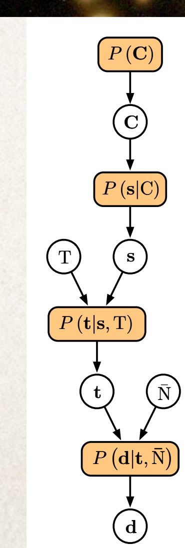


ICIC

Many Data: many numbers

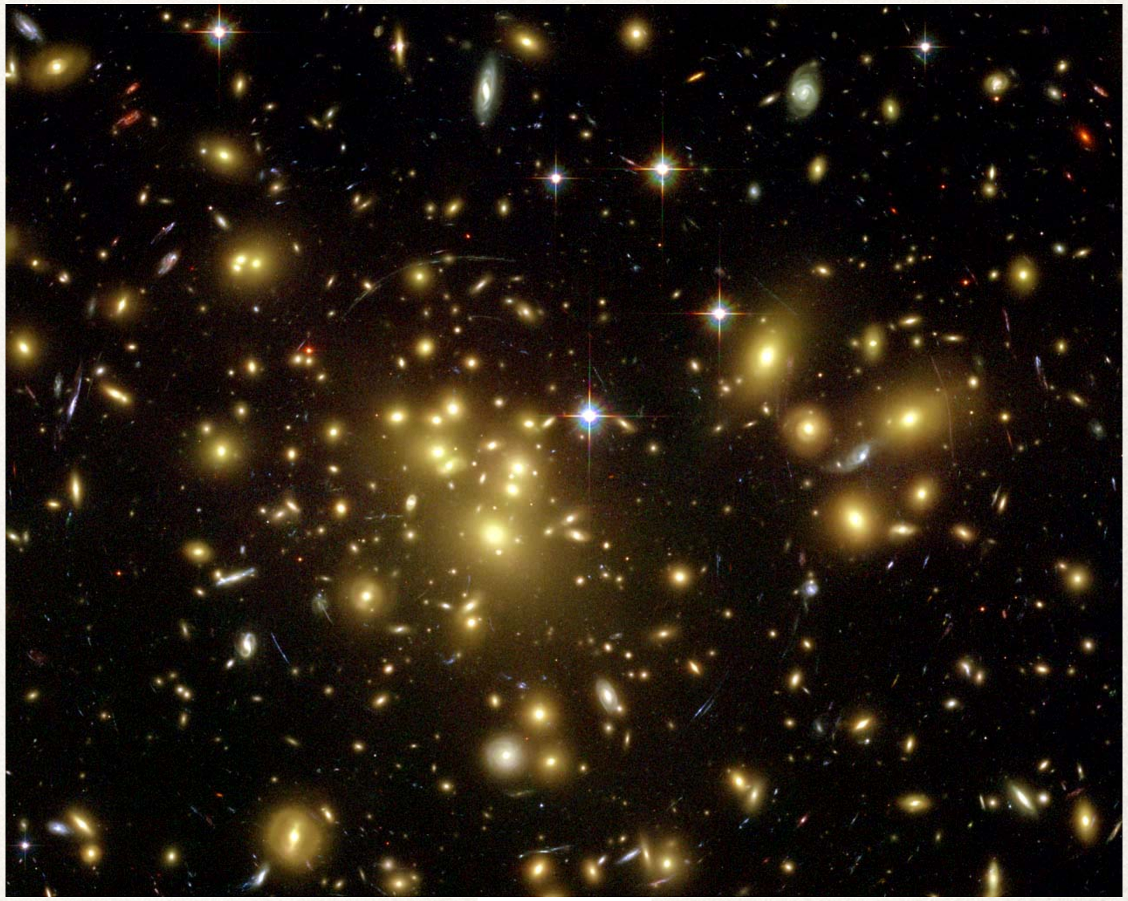
- **Model:** General Relativity → Mass bends light
- **Data:** image distortions (Millions)
- **Model parameters:** mass distribution (>100,000 numbers)
- **Bayesian Hierarchical Model**

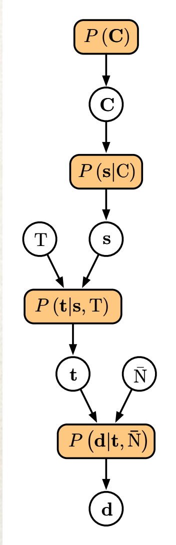




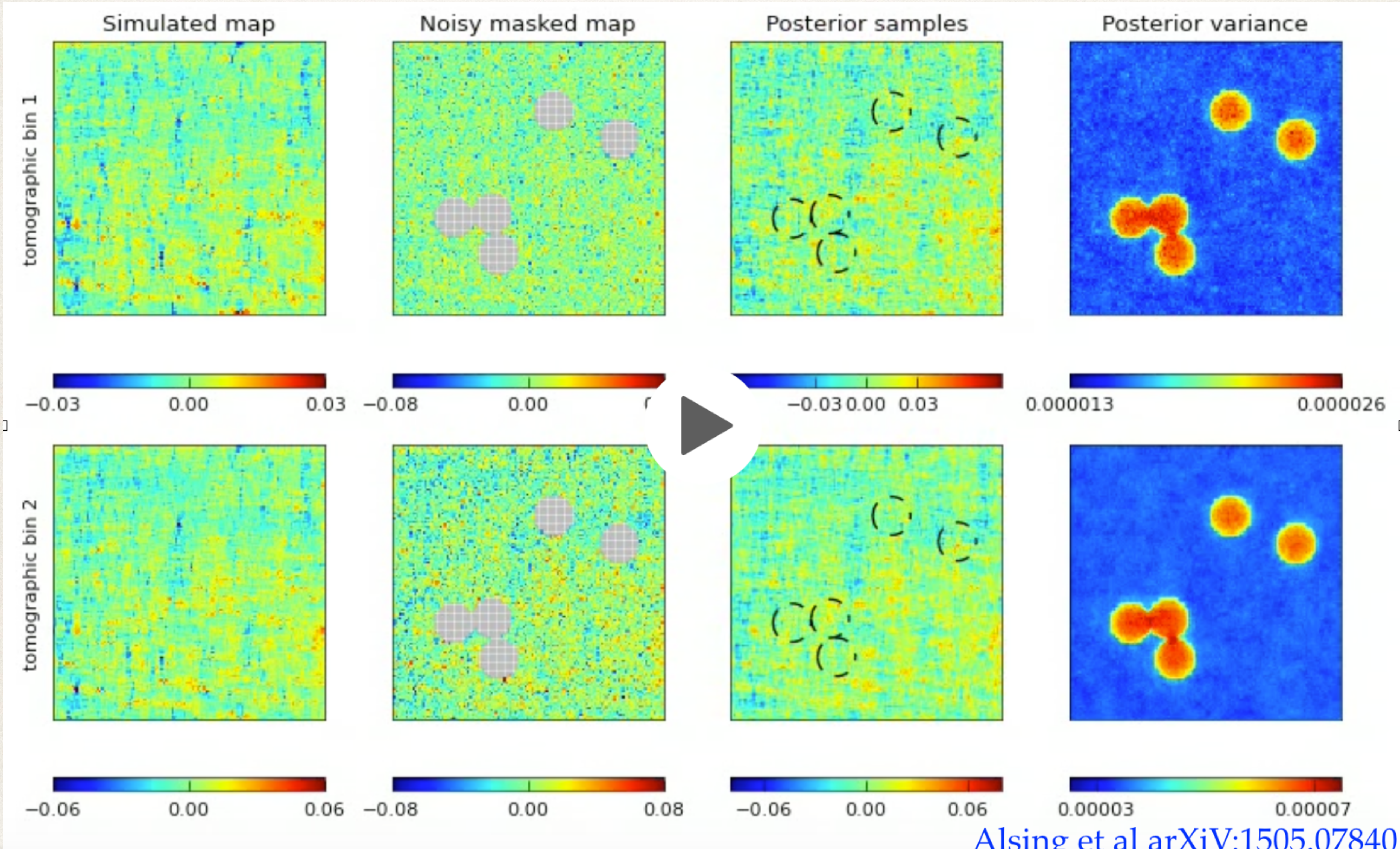
Many Data: many numbers

- **Model:** General Relativity ➢ Mass bends light
- **Data:** image distortions (Millions)
- **Model parameters:** mass distribution (>100,000 numbers)
- **Bayesian Hierarchical Model**
- 10 candidate mass maps per second on a desktop





Samples of the truth



Alsing et al arXiV:1505.07840

Conclusions

Conclusions

- **Many Data: few numbers**

Conclusions

- **<u>Many Data: few numbers</u>**
    - *May* be able to be analysed very efficiently when there is a good model for the data

Conclusions

- **Many Data: few numbers**
  - *May* be able to be analysed very efficiently when there is a good model for the data
  - MOPED

Conclusions

- **Many Data: few numbers**
    - *May* be able to be analysed very efficiently when there is a good model for the data
    - MOPED
- **Many Data: many numbers**

Conclusions

- **Many Data: few numbers**
  - *May* be able to be analysed very efficiently when there is a good model for the data
  - MOPED
- **Many Data: many numbers**
  - *May* be able to be analysed properly for the first time

Alan Heavens a.heavens@imperial.ac.uk

Imperial College London

Conclusions

- **Many Data: few numbers**
  - *May* be able to be analysed very efficiently when there is a good model for the data
  - MOPED
- **Many Data: many numbers**
  - *May* be able to be analysed properly for the first time
  - Bayesian Hierarchical Model

Analysing data from *Large N* permanent seismic stations to monitor subsurface processes

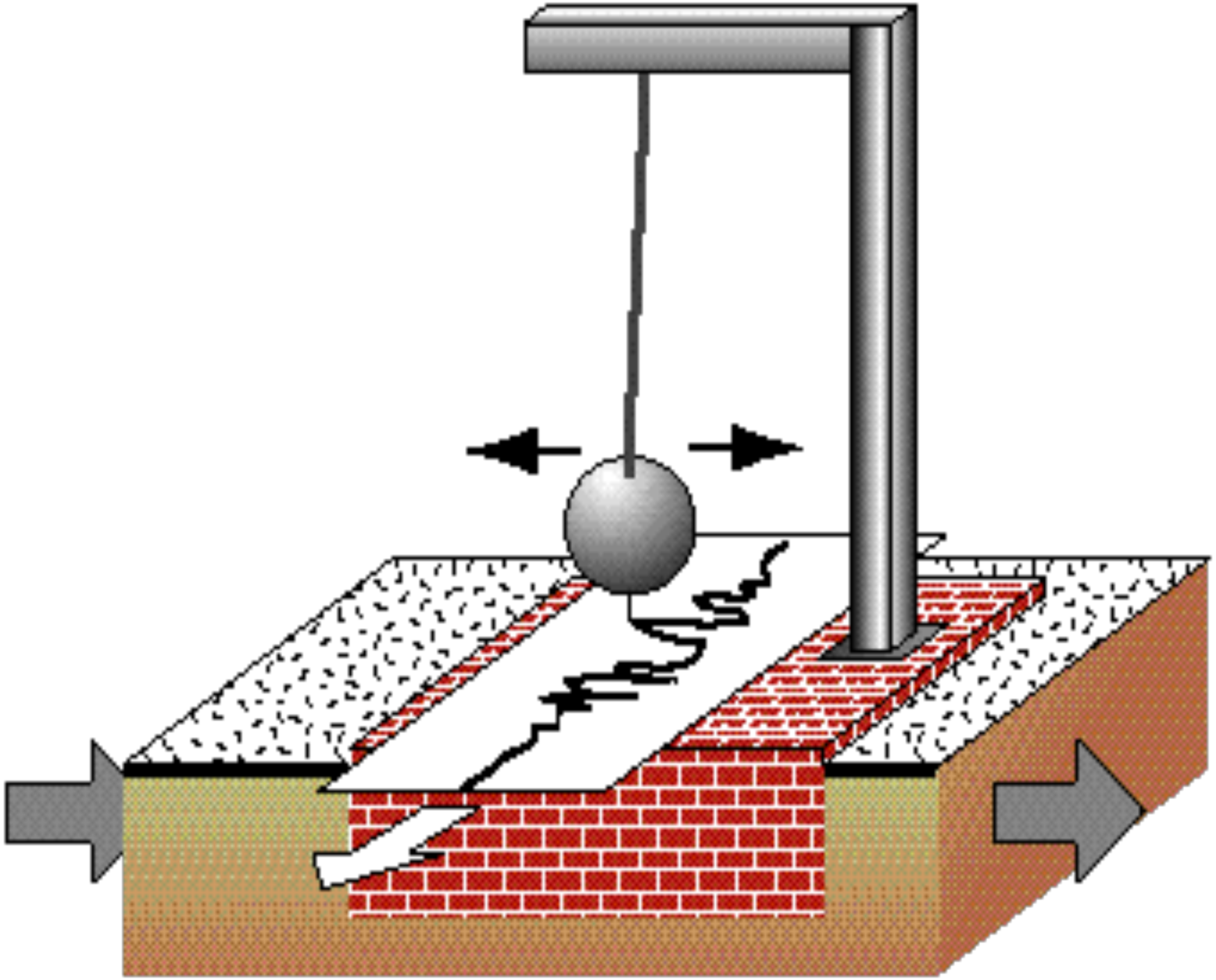
Sjoerd de Ridder and Andrew Curtis.

Take Home Message

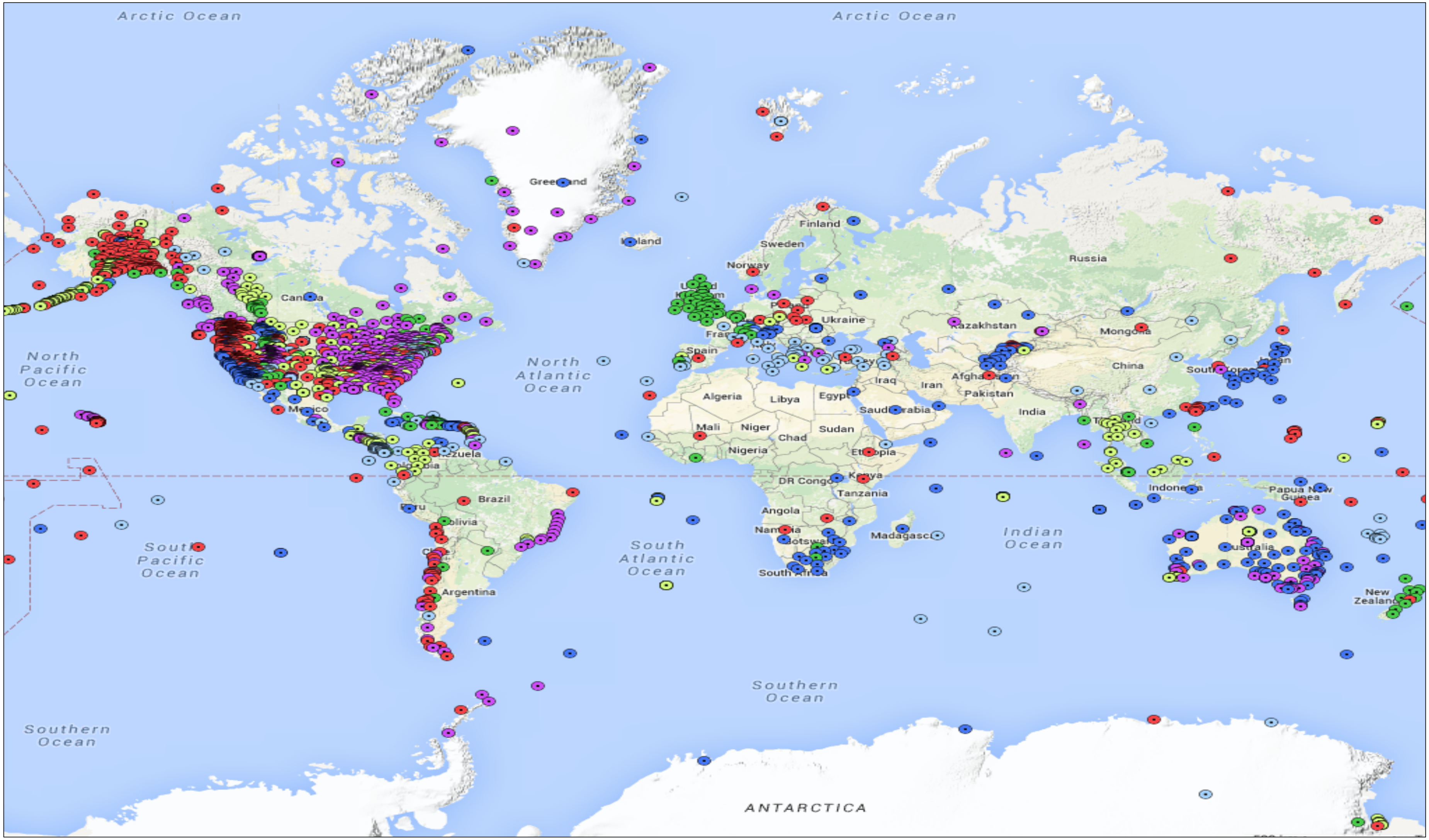
Big data and GeoSciences

Big data science key to observe and monitor the Earth in real-time

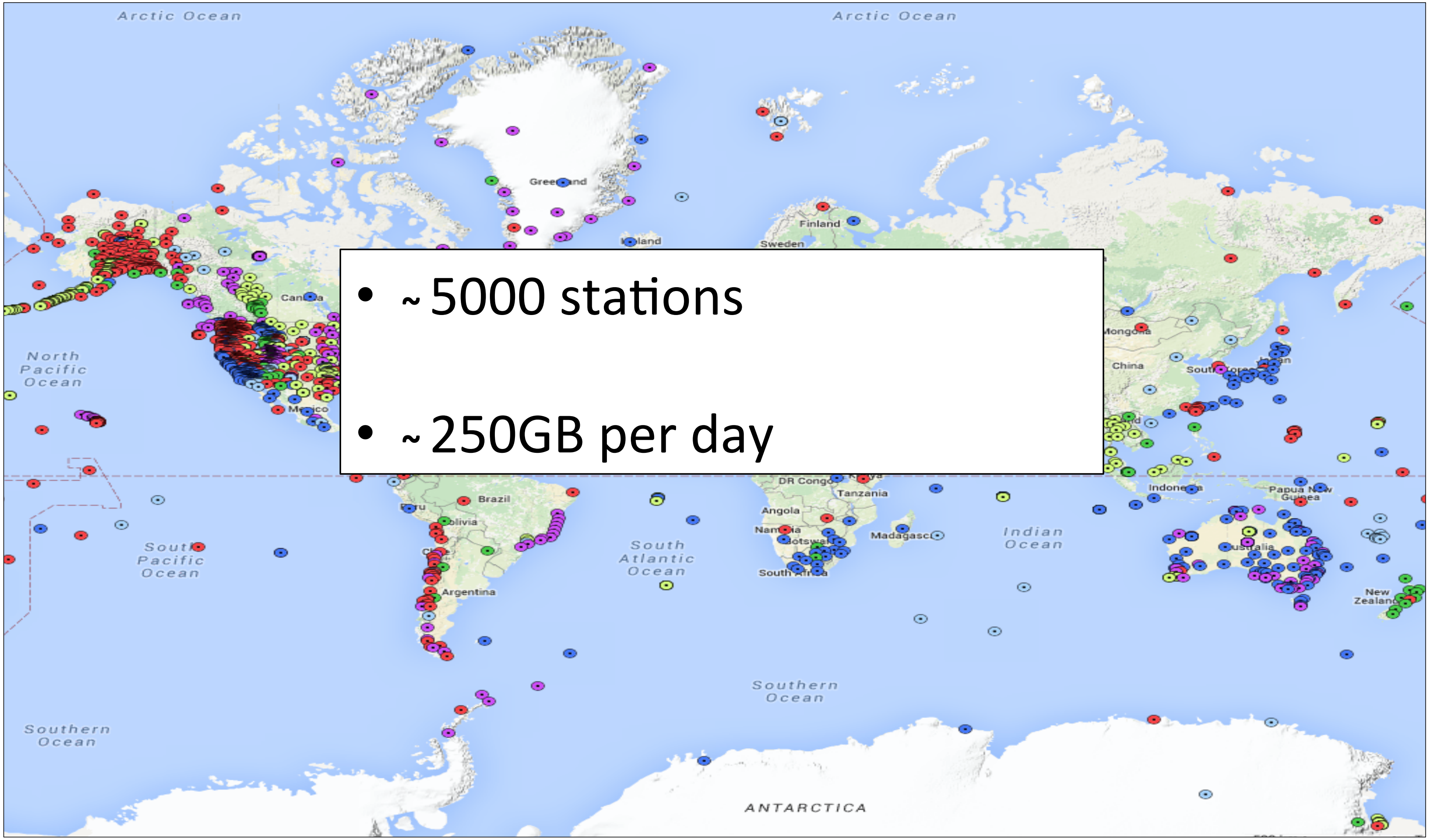
Seismograph Station



IRIS Data Center Real-Time Feeds

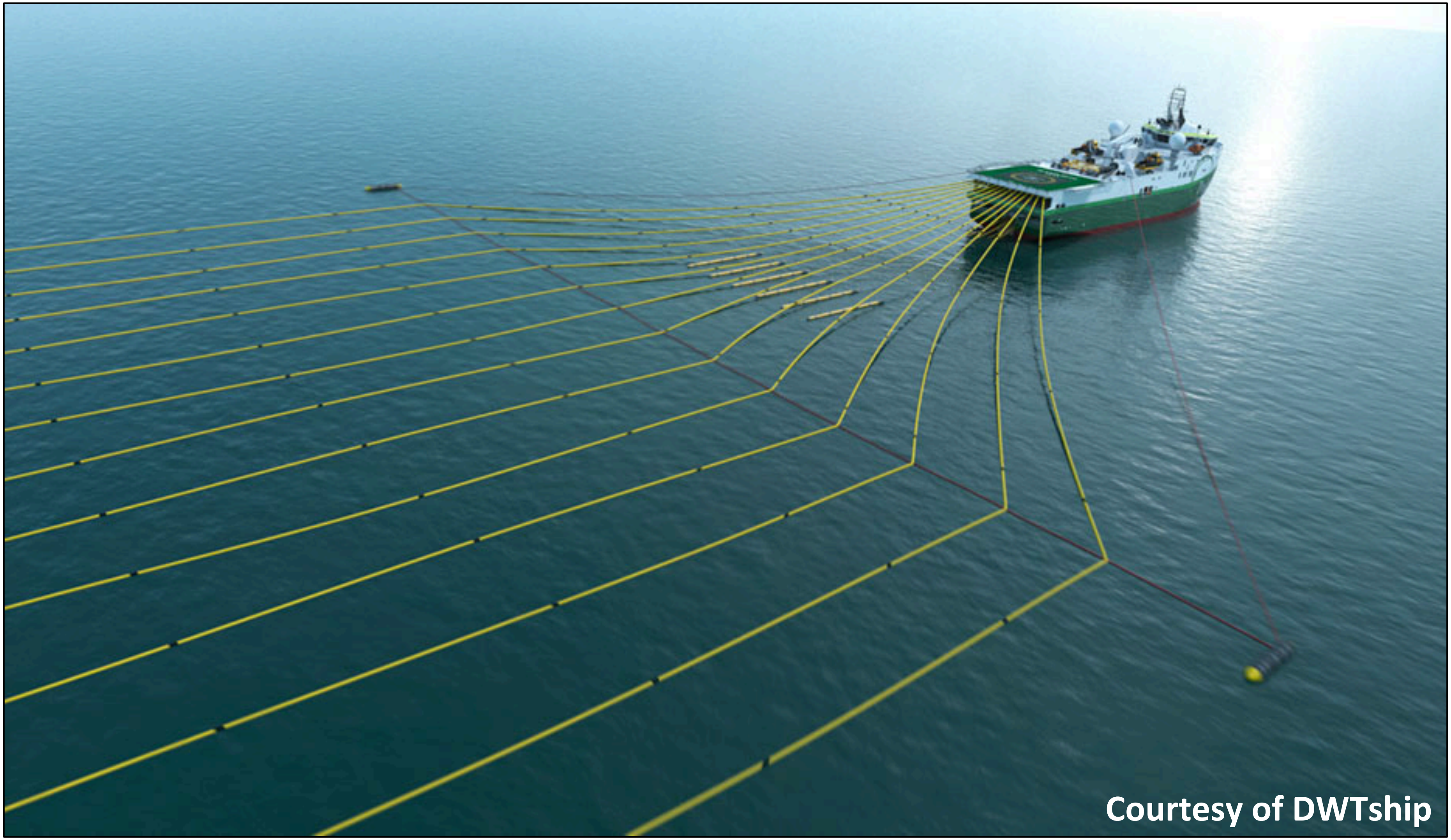


IRIS Data Center Real-Time Feeds



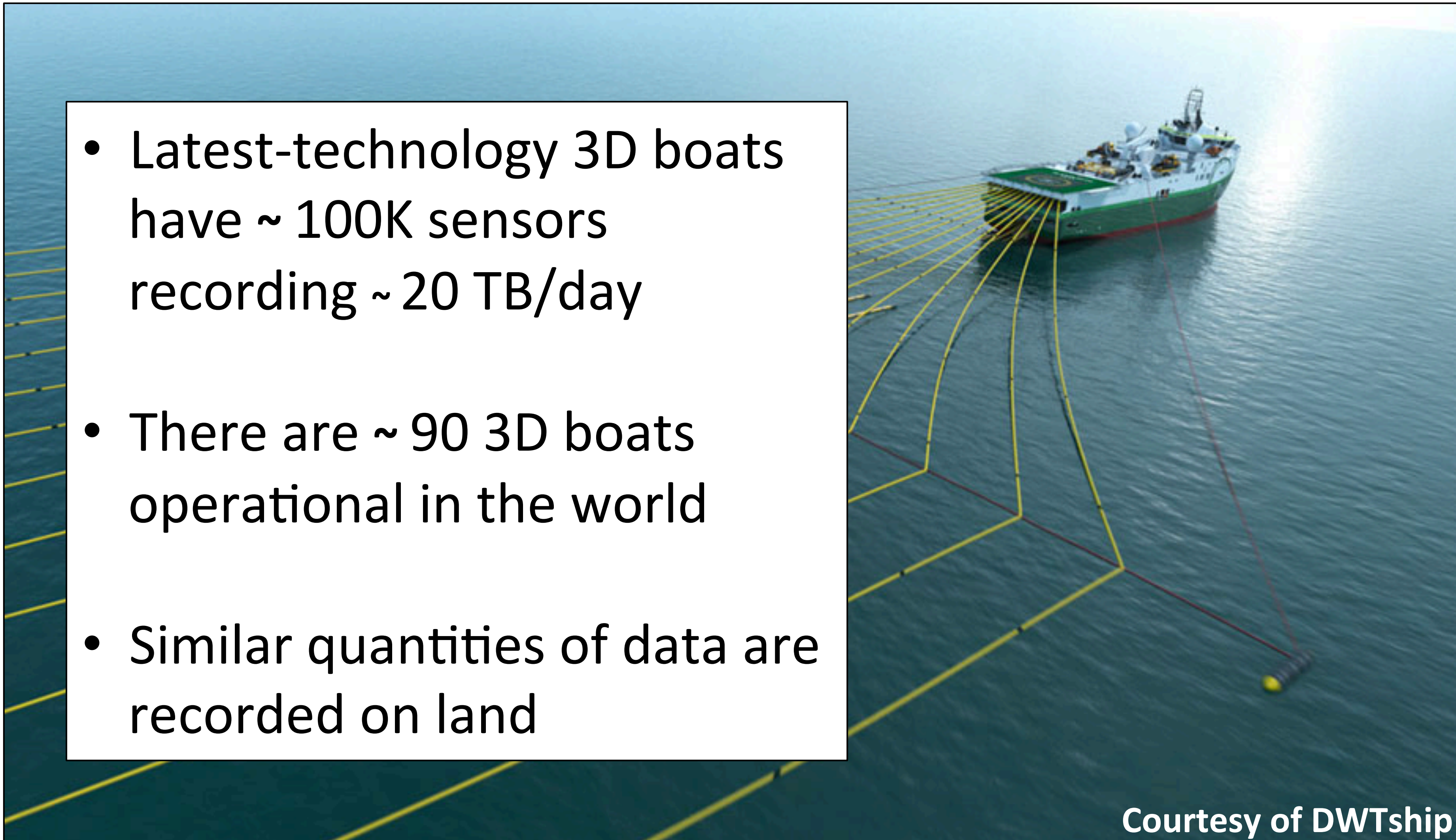
- ~5000 stations
- ~250GB per day

Big Data in Seismic Industry


Courtesy of DWTship

Big Data in Seismic Industry

- Latest-technology 3D boats have ~ 100K sensors recording ~ 20 TB/day
- There are ~ 90 3D boats operational in the world
- Similar quantities of data are recorded on land



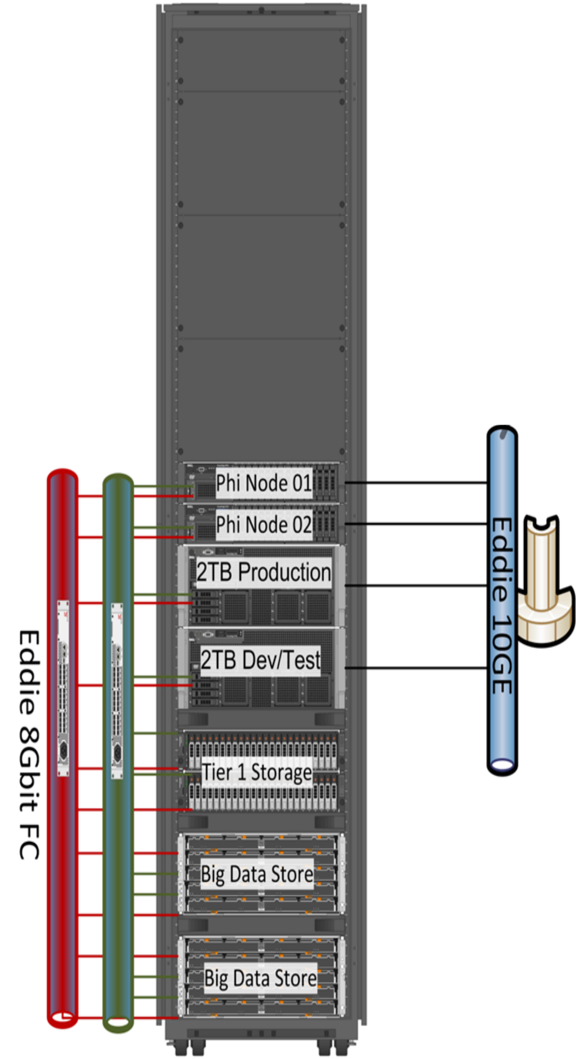
Courtesy of DWTship

TerraCorrelator Facility

- Seismic noise correlations for imaging of earth properties.

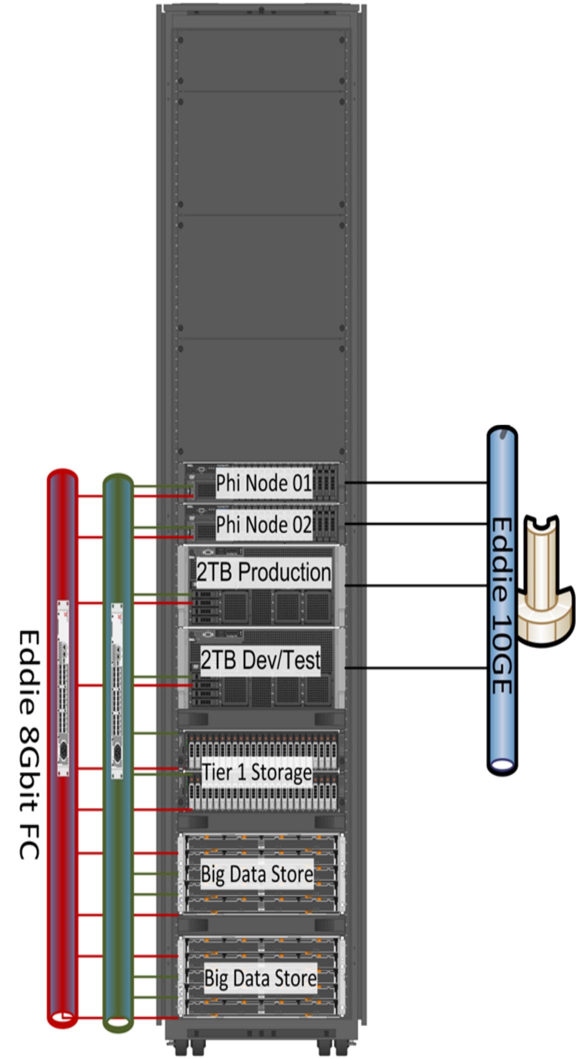
- Earthquake repeater analysis, for volcano and plate boundary study.

  ➔ Real-time risk assessment with seismic data.



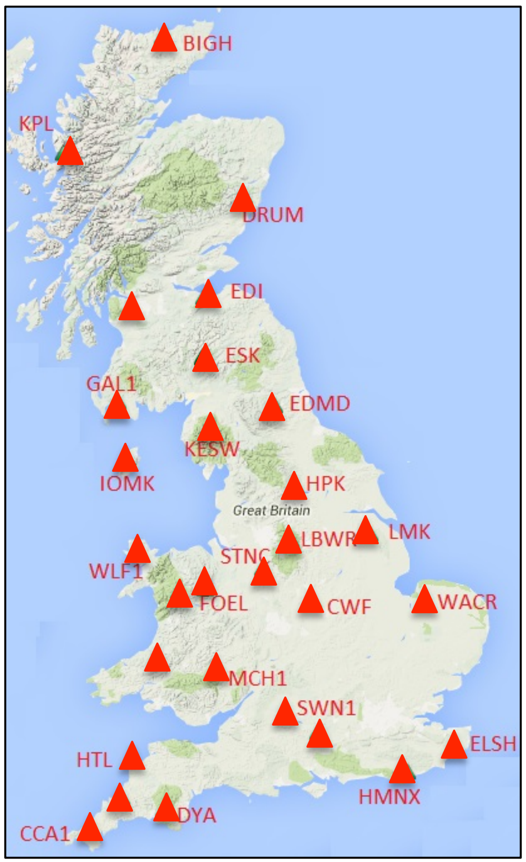
TerraCorrelator Facility

- Seismic noise correlations for imaging of earth properties.
- Earthquake repeater analysis, for volcano and plate boundary study.

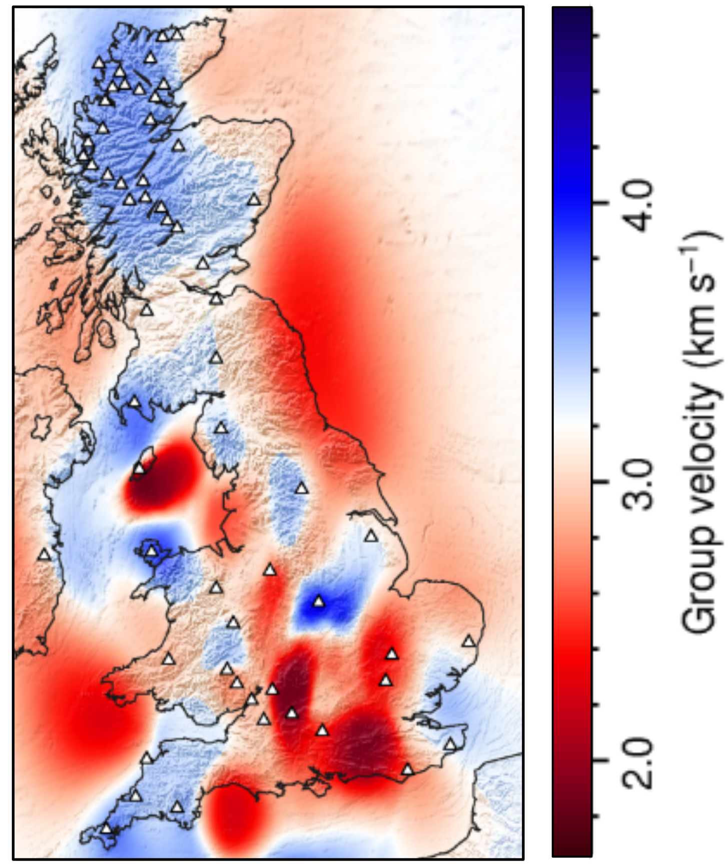


Great-Britain Seismic Network

Seismic Stations



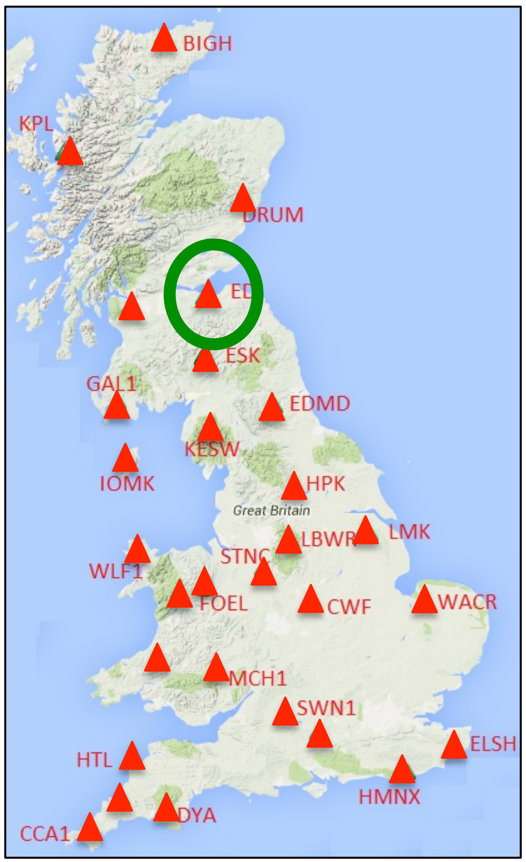
Map of Wave Velocities



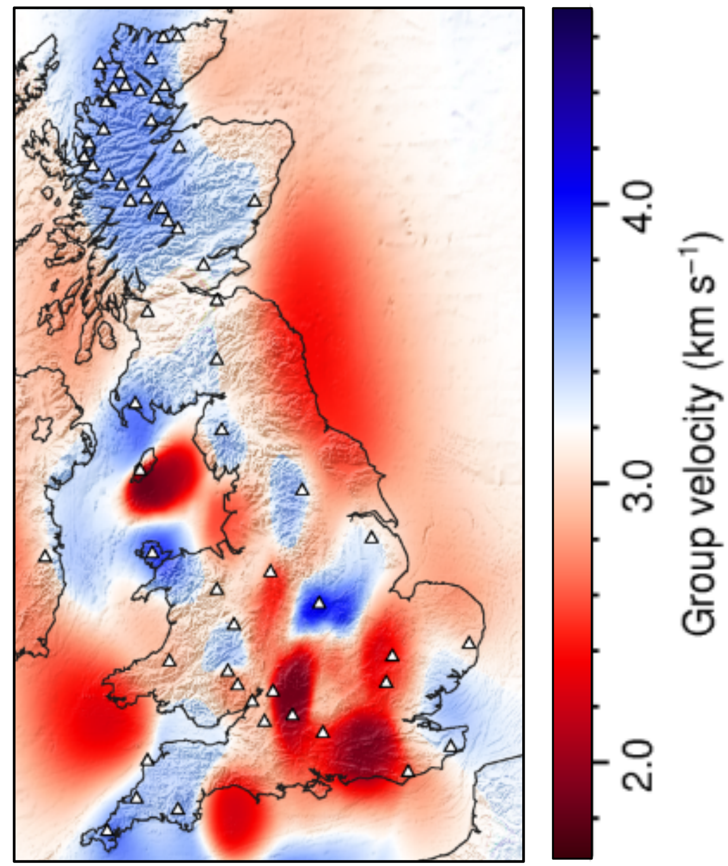
Courtesy of Erica Galetti, UoE

Great-Britain Seismic Network

Seismic Stations



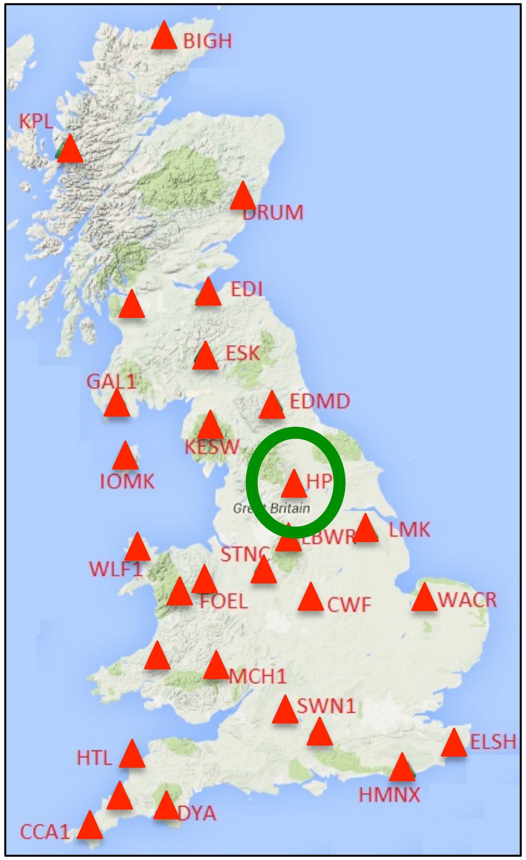
Map of Wave Velocities



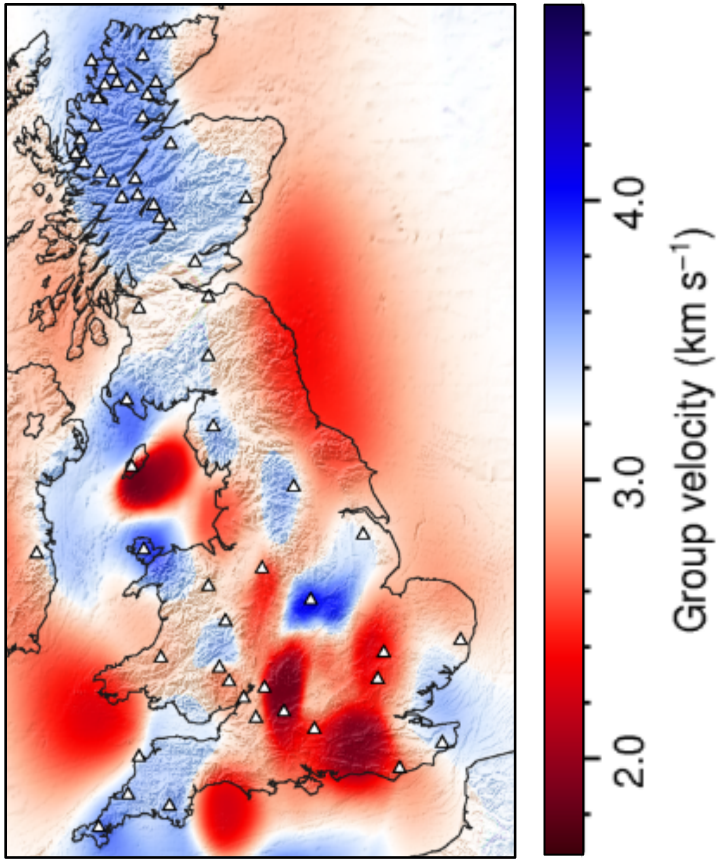
Courtesy of Erica Galetti, UoE

Great-Britain Seismic Network

Seismic Stations

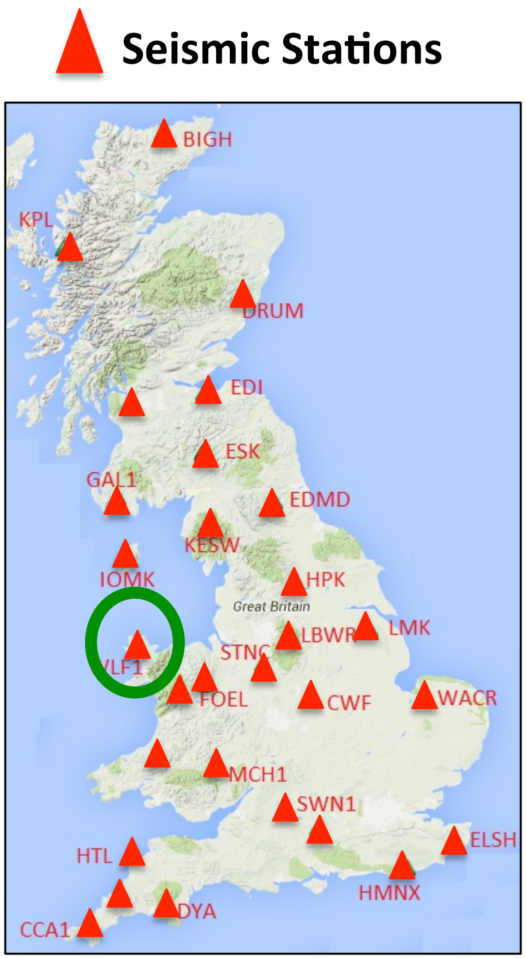


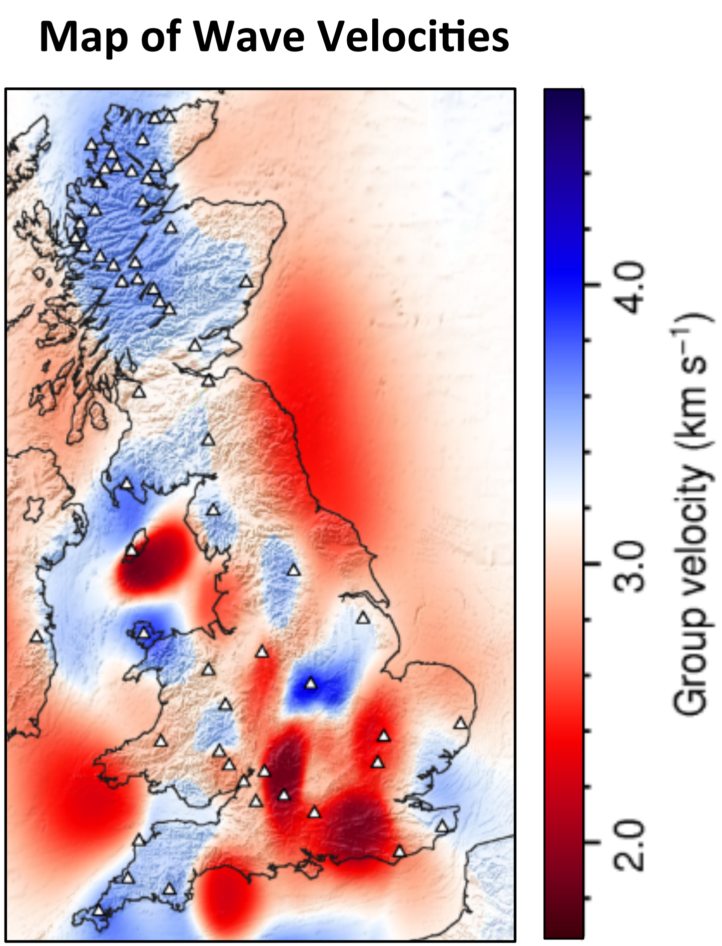
Map of Wave Velocities



Courtesy of Erica Galetti, UoE

Great-Britain Seismic Network

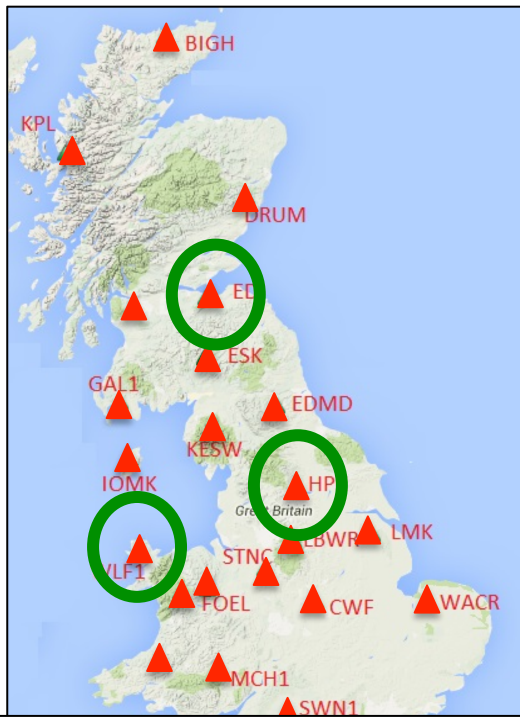




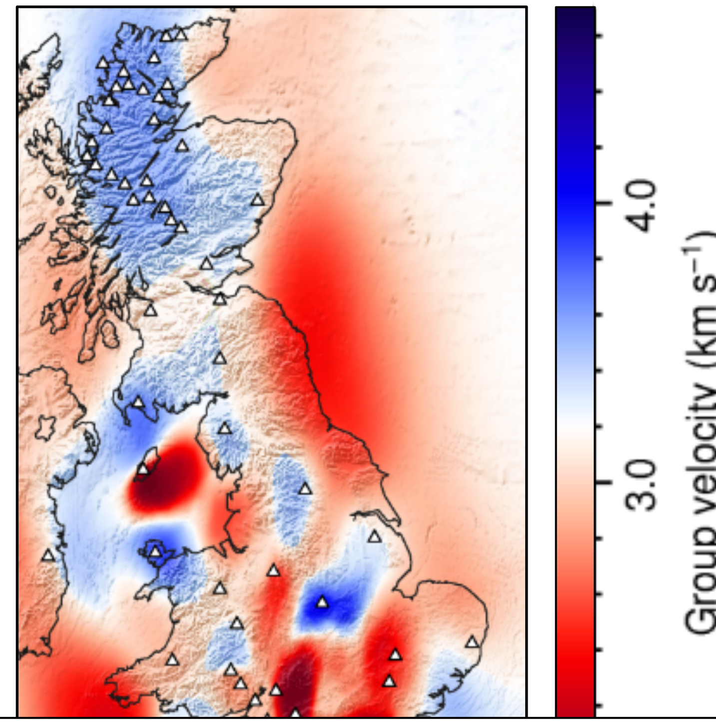
Courtesy of Erica Galetti, UoE

Great-Britain Seismic Network

Seismic Stations



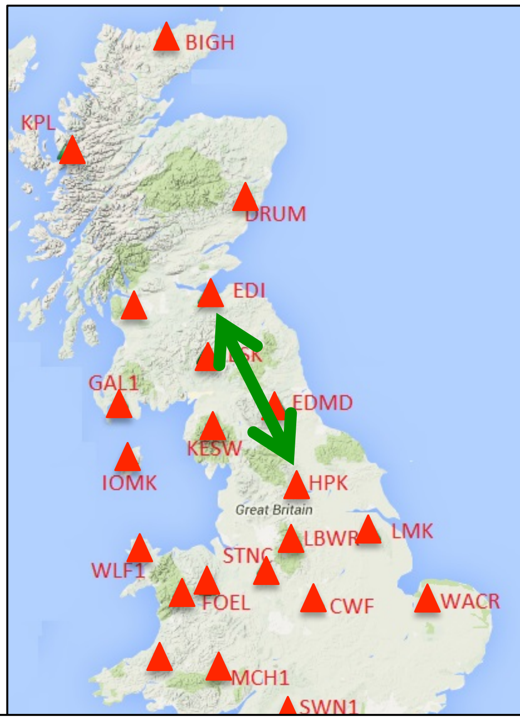
Map of Wave Velocities

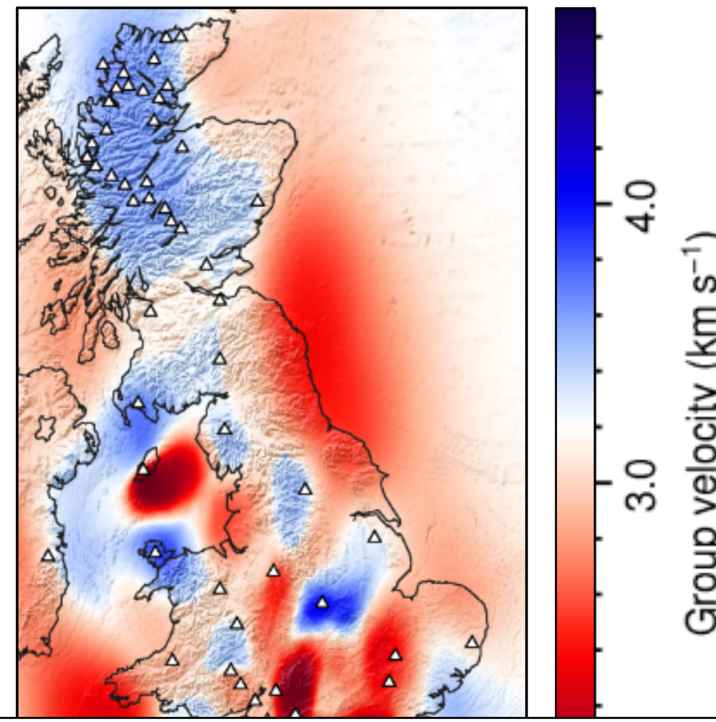


- Pre-Processing of Recordings – one station at a time

Great-Britain Seismic Network

Seismic Stations **Map of Wave Velocities**

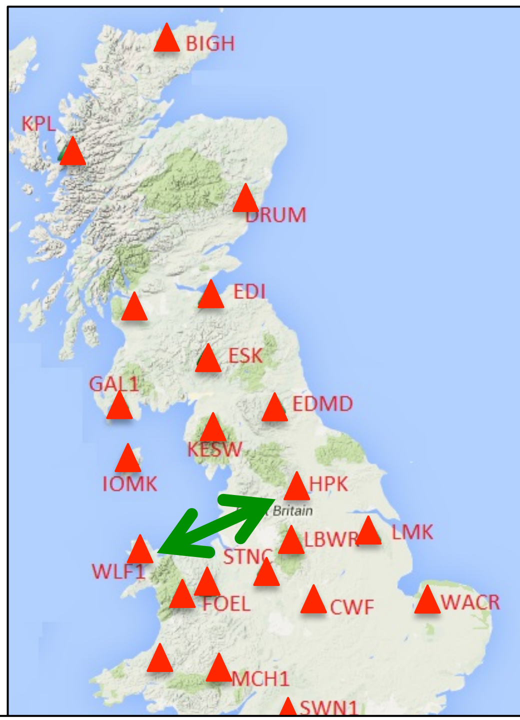




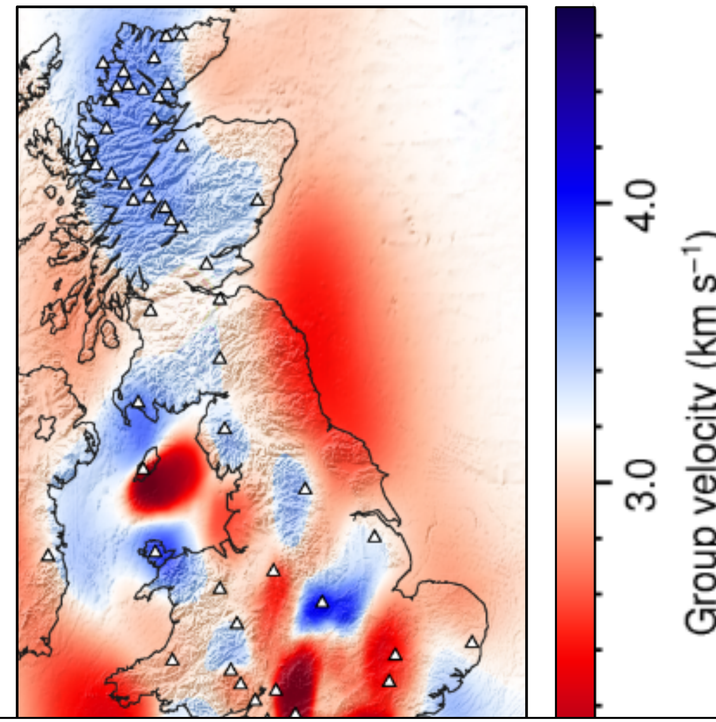
- Pre-Processing of Recordings – one station at a time

Great-Britain Seismic Network

Seismic Stations



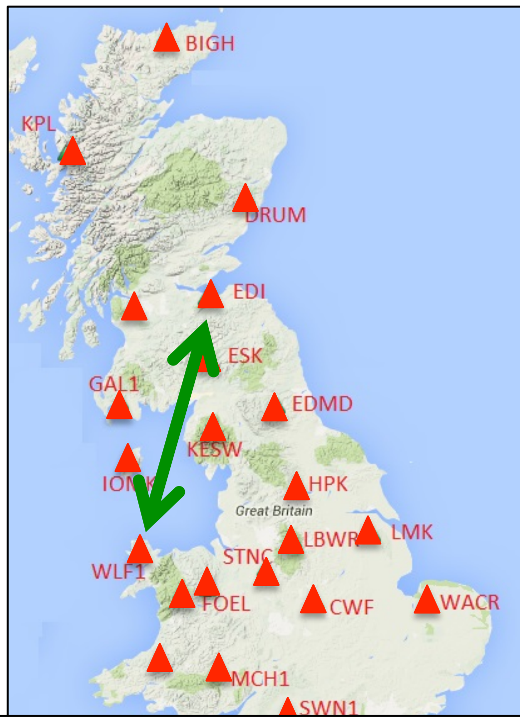
Map of Wave Velocities



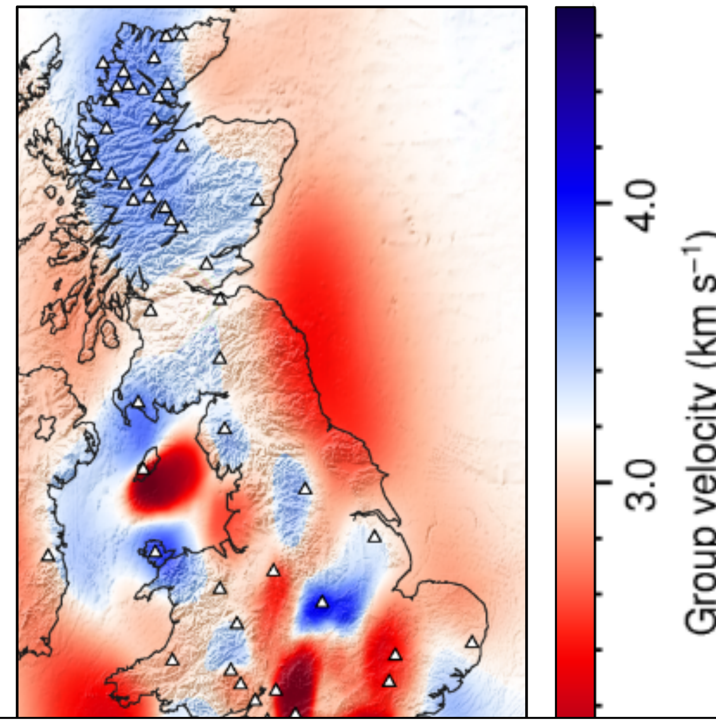
- Pre-Processing of Recordings – one station at a time

Great-Britain Seismic Network

Seismic Stations



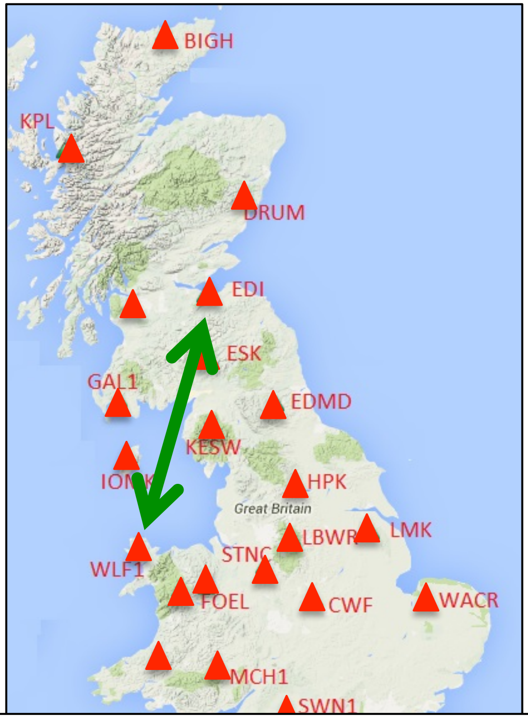
Map of Wave Velocities



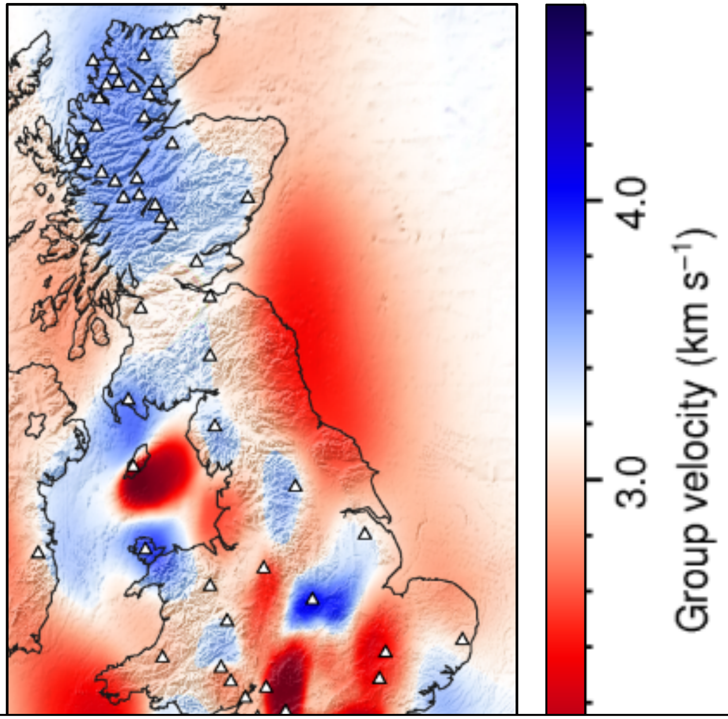
- Pre-Processing of Recordings – one station at a time

Great-Britain Seismic Network

Seismic Stations



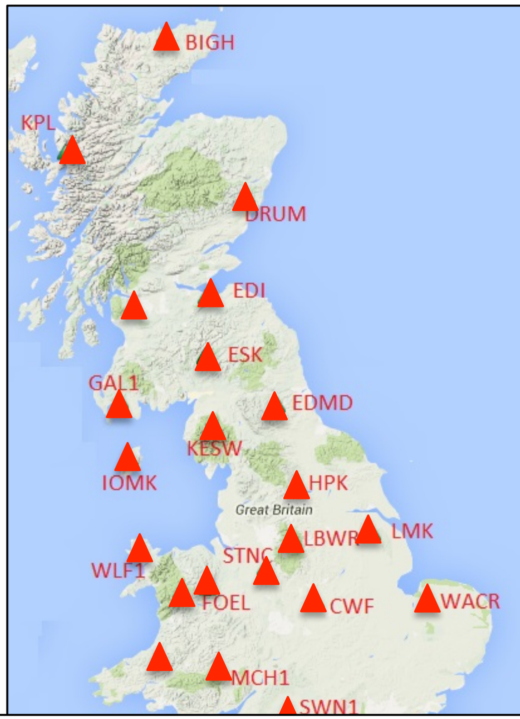
Map of Wave Velocities



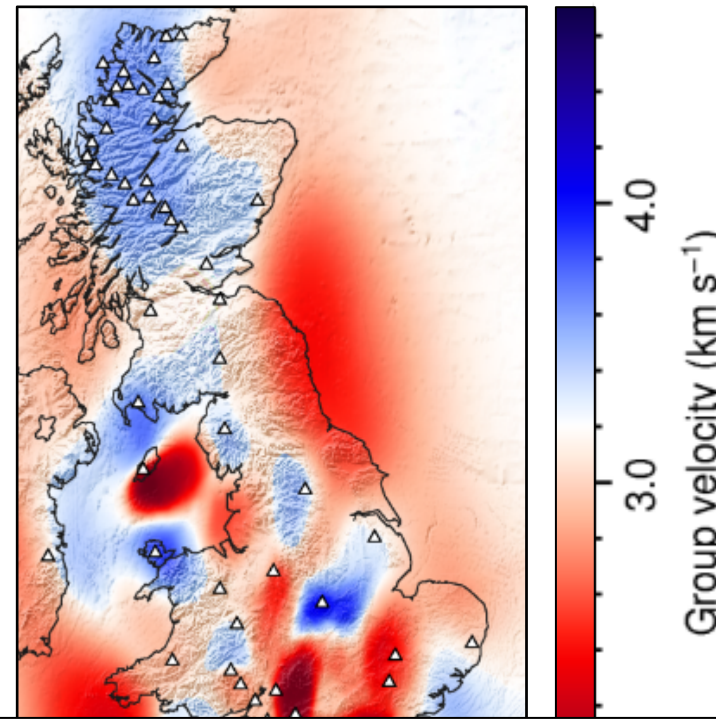
- Pre-Processing of Recordings – one station at a time
- Travel time Computations – two stations at a time

Great-Britain Seismic Network

Seismic Stations



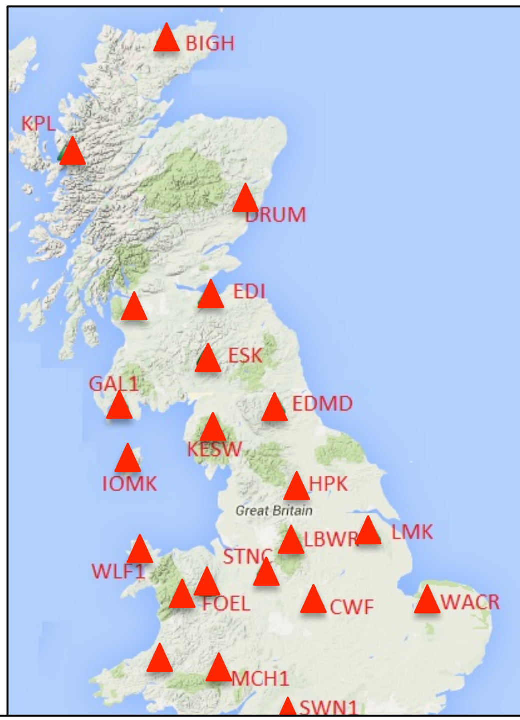
Map of Wave Velocities



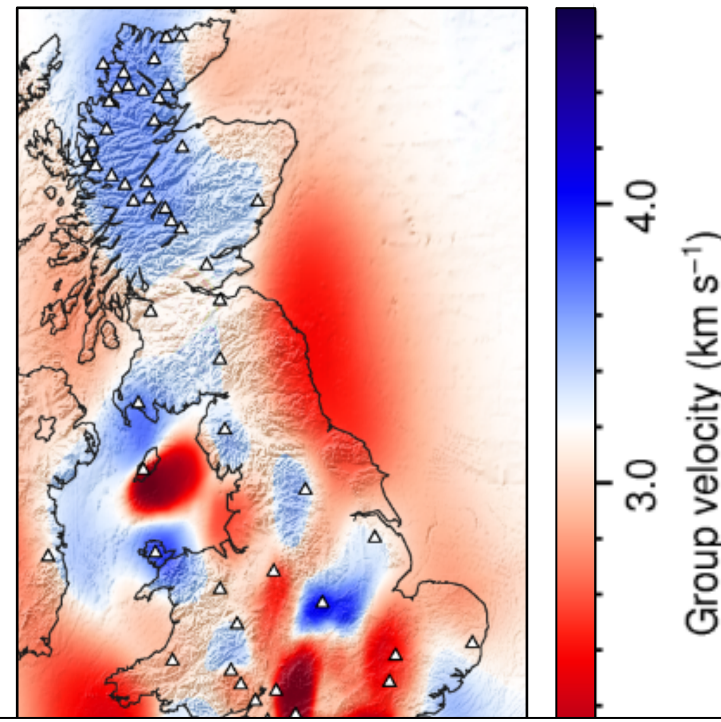
- Pre-Processing of Recordings – one station at a time
- Travel time Computations – two stations at a time

Great-Britain Seismic Network

Seismic Stations



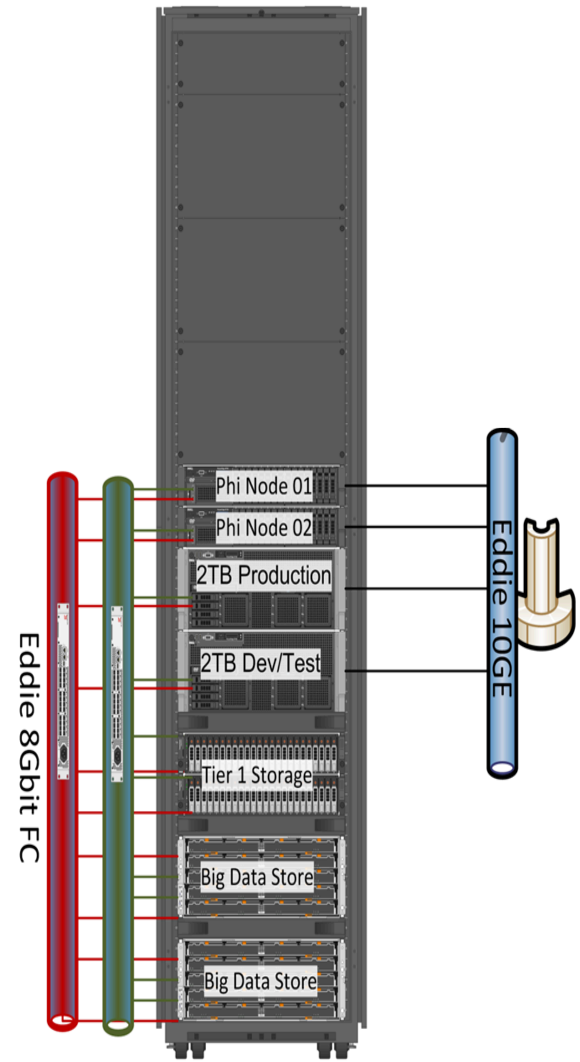
Map of Wave Velocities



- Pre-Processing of Recordings – one station at a time
- Travel time Computations – two stations at a time
- Tomographic Computation – all stations simultaneously

TerraCorrelator Facility

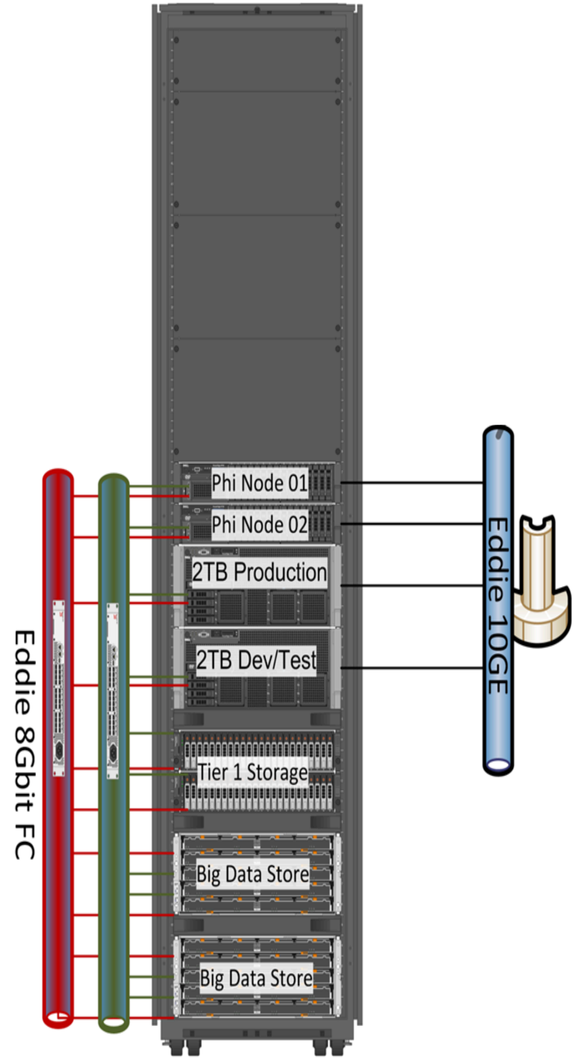
- 2 nodes with 4 Intel Xeon E7-4830 8 core processors, and **2TB RAM**.
- 2 fileservers: 208 TB.
- 1 fileservers: **28 TB high-performance SAS.**



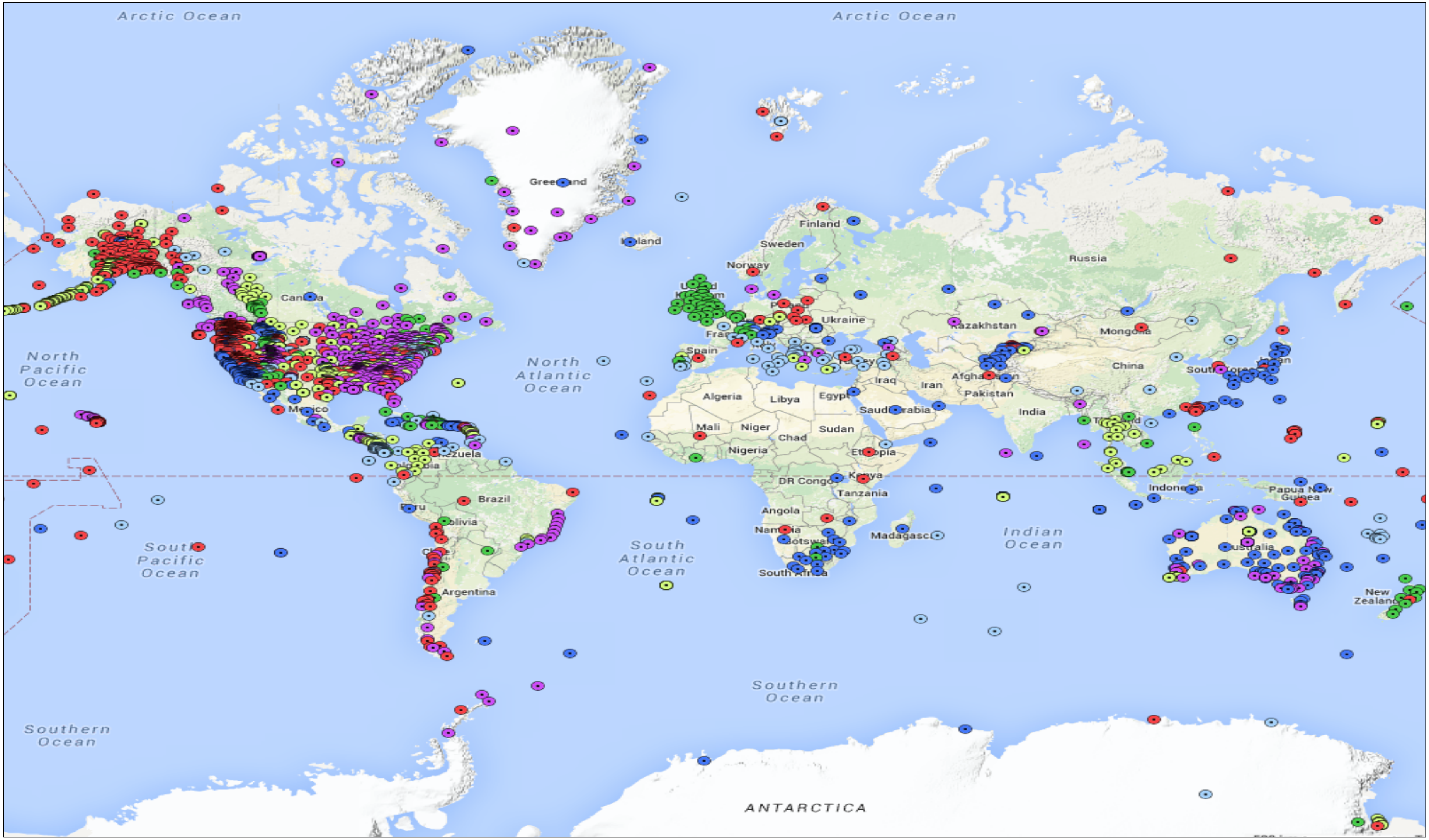
TerraCorrelator Facility

- 2 nodes with 4 Intel Xeon E7-4830 8 core processors, and **2TB RAM**.
- 2 fileservers: 208 TB.
- 1 fileservers: **28 TB high-performance SAS.**

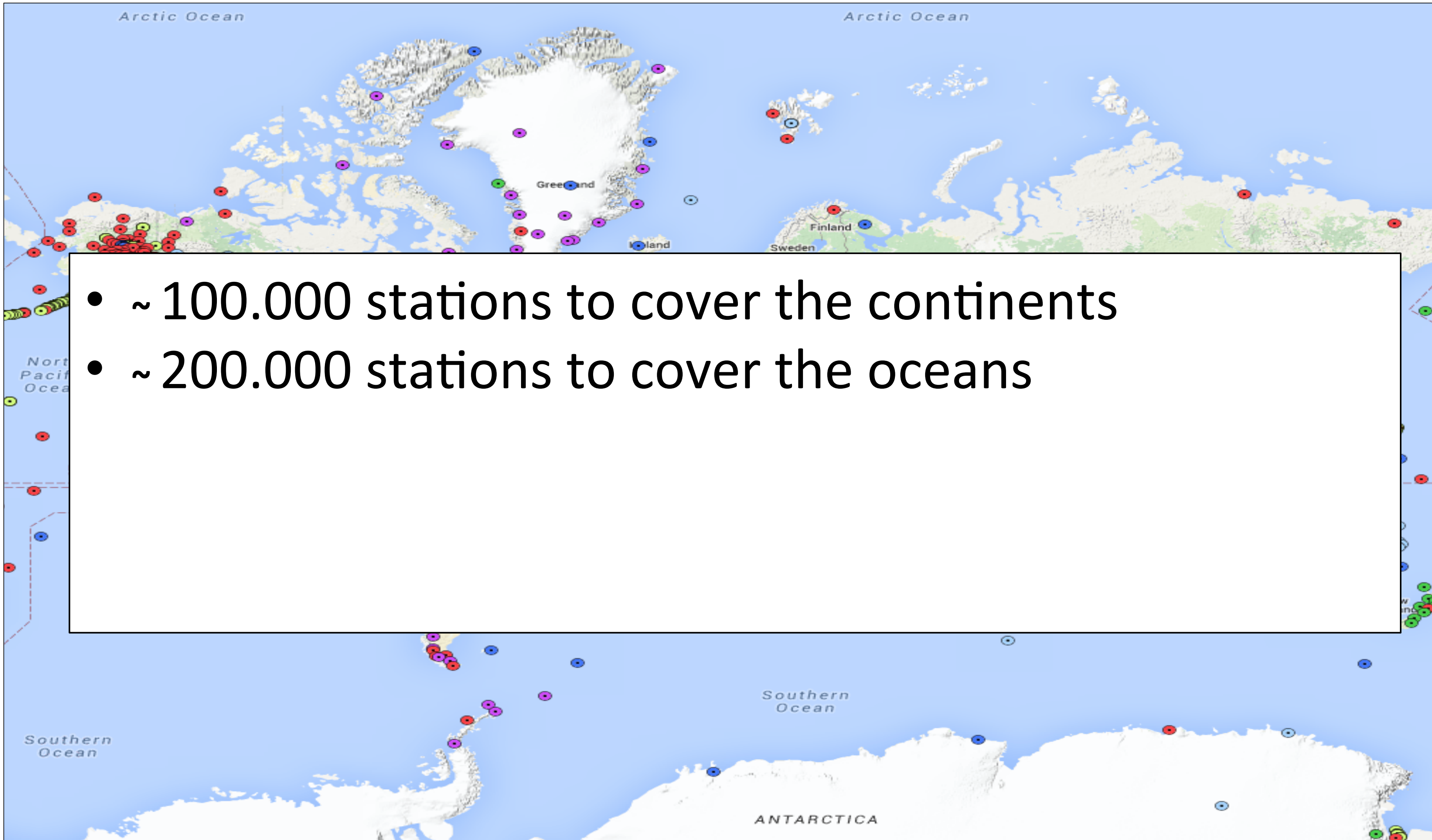
- Can handle up to 1000 stations



Vision: A Future of seismology in ATI

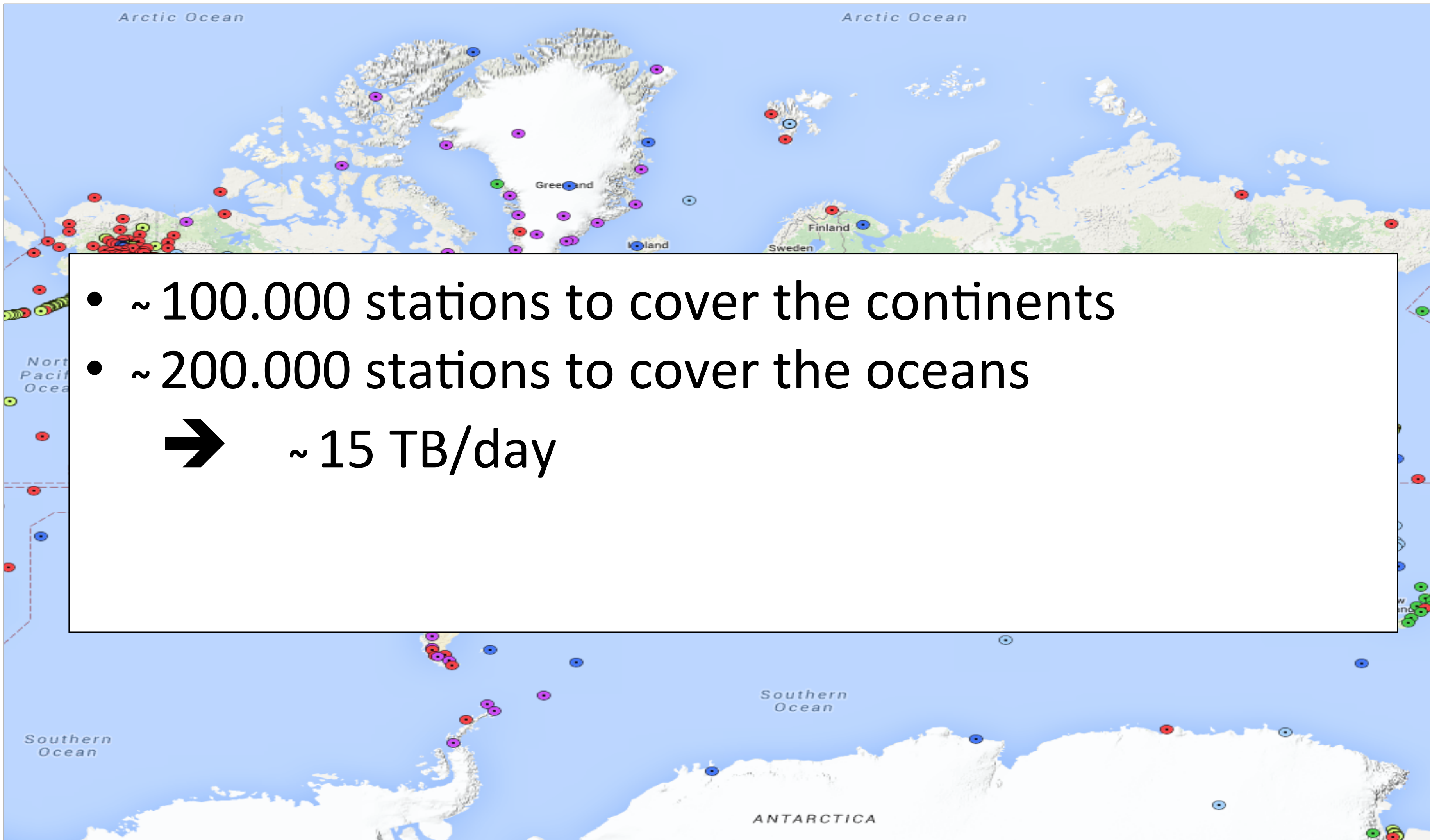


Vision: A Future of seismology in ATI



- ~ 100.000 stations to cover the continents
- ~ 200.000 stations to cover the oceans

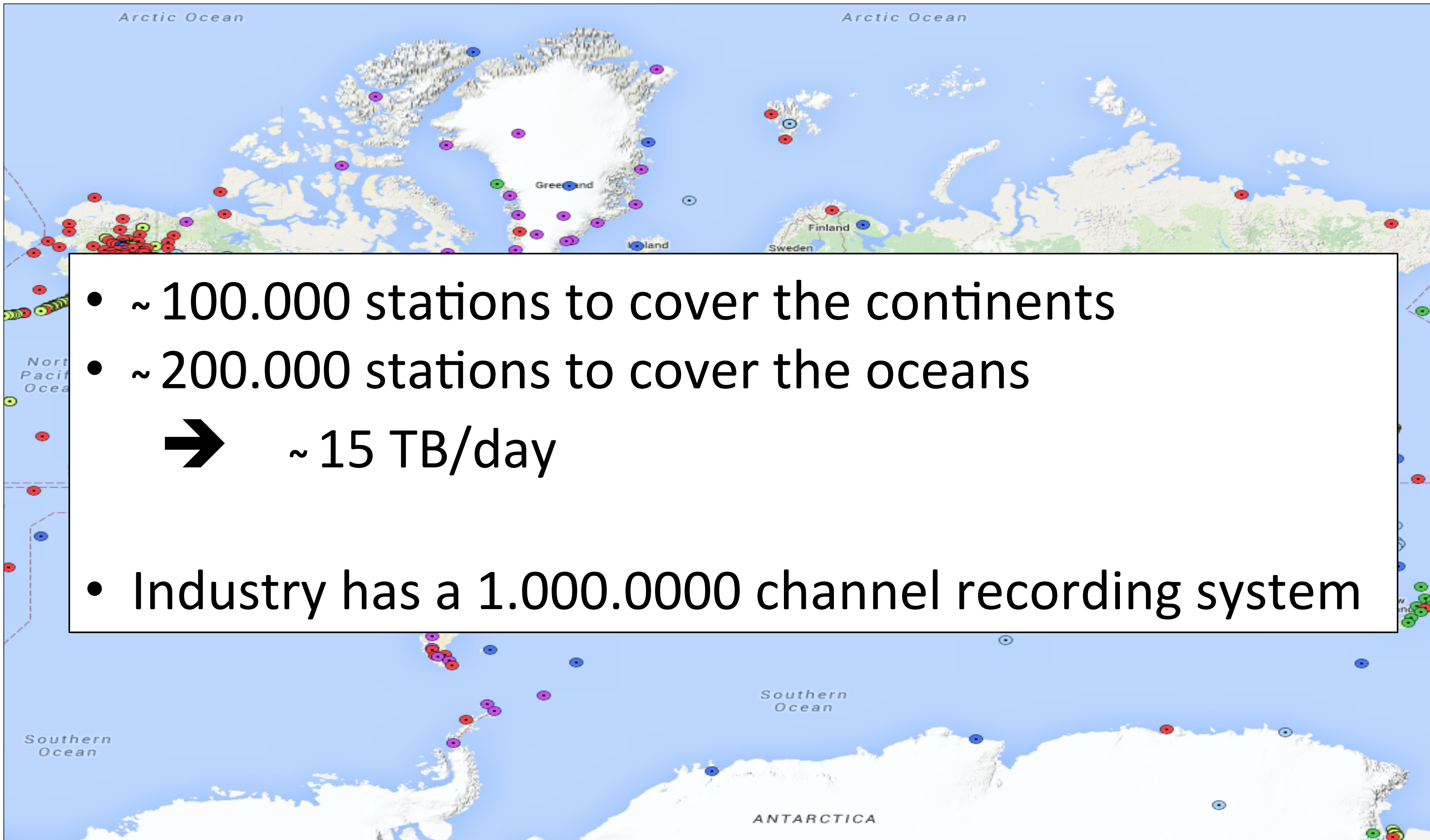
Vision: A Future of seismology in ATI



- ~ 100.000 stations to cover the continents
- ~ 200.000 stations to cover the oceans

➔ ~ 15 TB/day

Vision: A Future of seismology in ATI



- ~ 100.000 stations to cover the continents
- ~ 200.000 stations to cover the oceans

  ➔ ~ 15 TB/day

- Industry has a 1.000.0000 channel recording system

Vision: A Future of seismology in ATI

- Challenge 1: Rolling out a dense seismic network across the globe

Vision: A Future of seismology in ATI

- Challenge 1: Rolling out a dense seismic network across the globe
- Challenge 2: Obtaining the data in real-time

Vision: A Future of seismology in ATI

- Challenge 1: Rolling out a dense seismic network across the globe
- Challenge 2: Obtaining the data in real-time

➔ Relatively simple informatics problem.

Vision: A Future of seismology in ATI

- Challenge 1: Rolling out a dense seismic network across the globe
- Challenge 2: Obtaining the data in real-time

→ Relatively simple informatics problem.

→ Societal and Political science aspects to roll this out to poor and instable countries.

Vision: A Future of seismology in ATI

- Challenge 1: Rolling out a dense seismic network across the globe
- Challenge 2: Obtaining the data in real-time

Vision: A Future of seismology in ATI

- Challenge 1: Rolling out a dense seismic network across the globe
- Challenge 2: Obtaining the data in real-time
- Challenge 3: The Earth-Data-Science challenge

Vision: A Future of seismology in ATI

- Challenge 1: Rolling out a dense seismic network across the globe
- Challenge 2: Obtaining the data in real-time
- Challenge 3: The Earth-Data-Science challenge

➔ Need mathematicians, informaticians, statisticians, and physicists to join with Earth scientists.

Data science challenges and solutions in Astrochemistry

Serena Viti

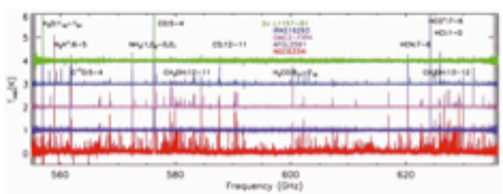
Department of Physics and Astronomy

UCL

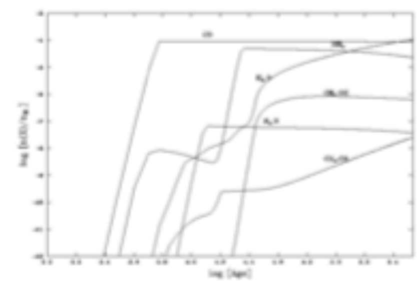
Molecular observations and interpretation:

The canonical approach and its limitations

Astrophysical object



Spectra



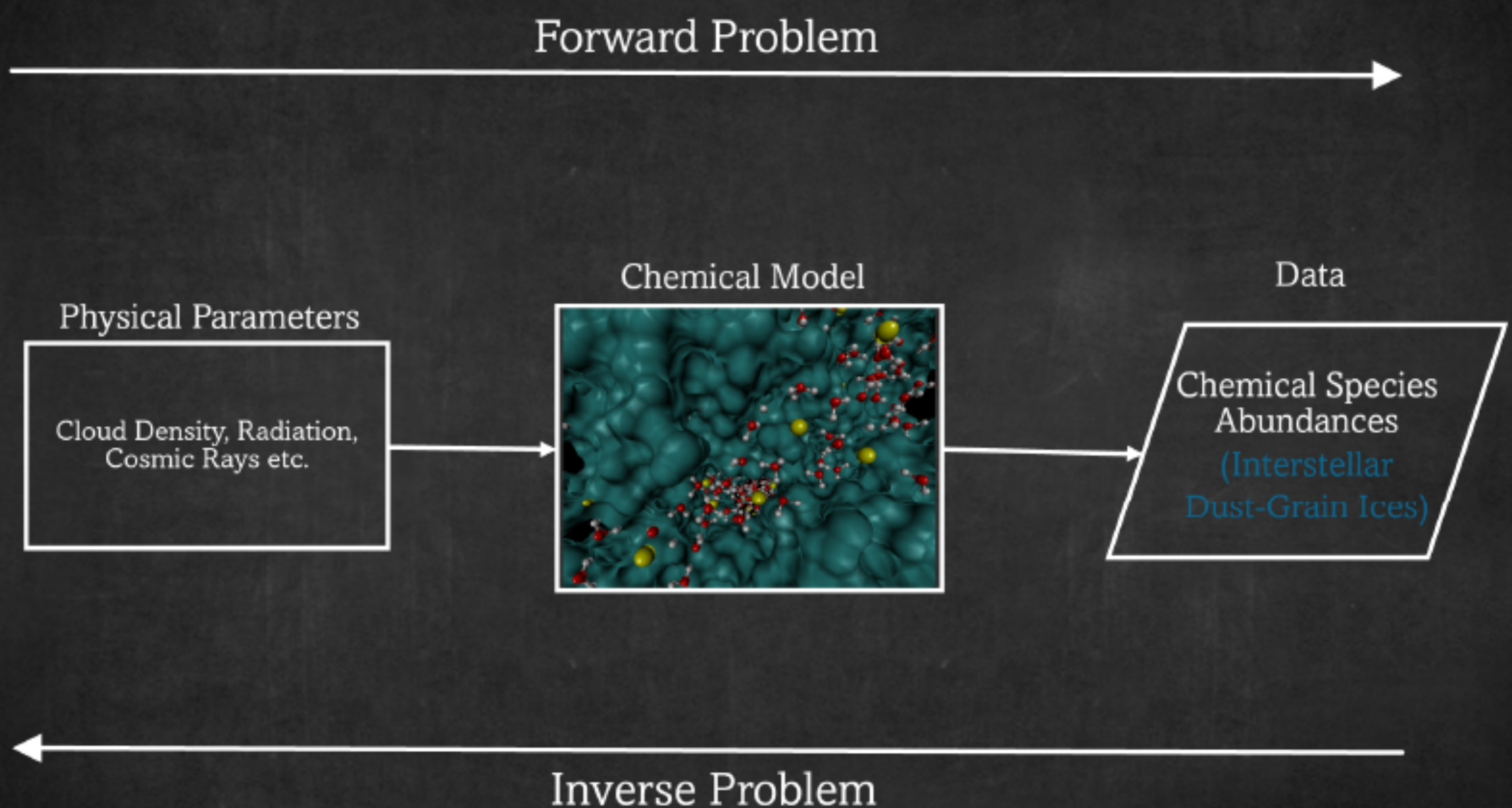
Best fit model

This 4-step procedure highlights the *inverse* nature of the problem → deriving information about molecular clouds using observational information and, even well established modelling codes, is an ***inverse*** problem that usually does not fulfil Hadamard's postulates of well posedness i.e:

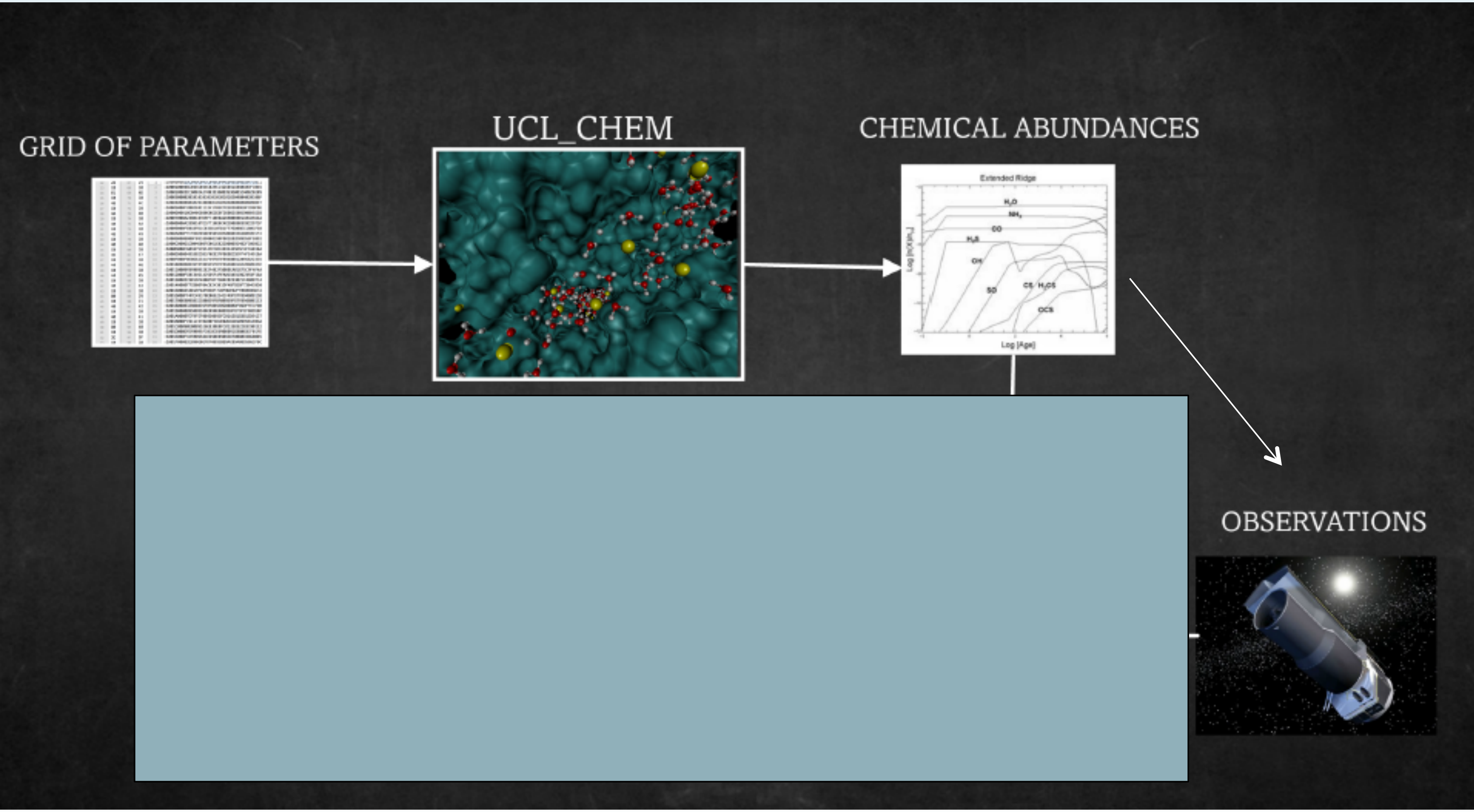
- it may not have a solution
- solutions might not be unique and/or might not depend continuously on the observational data.

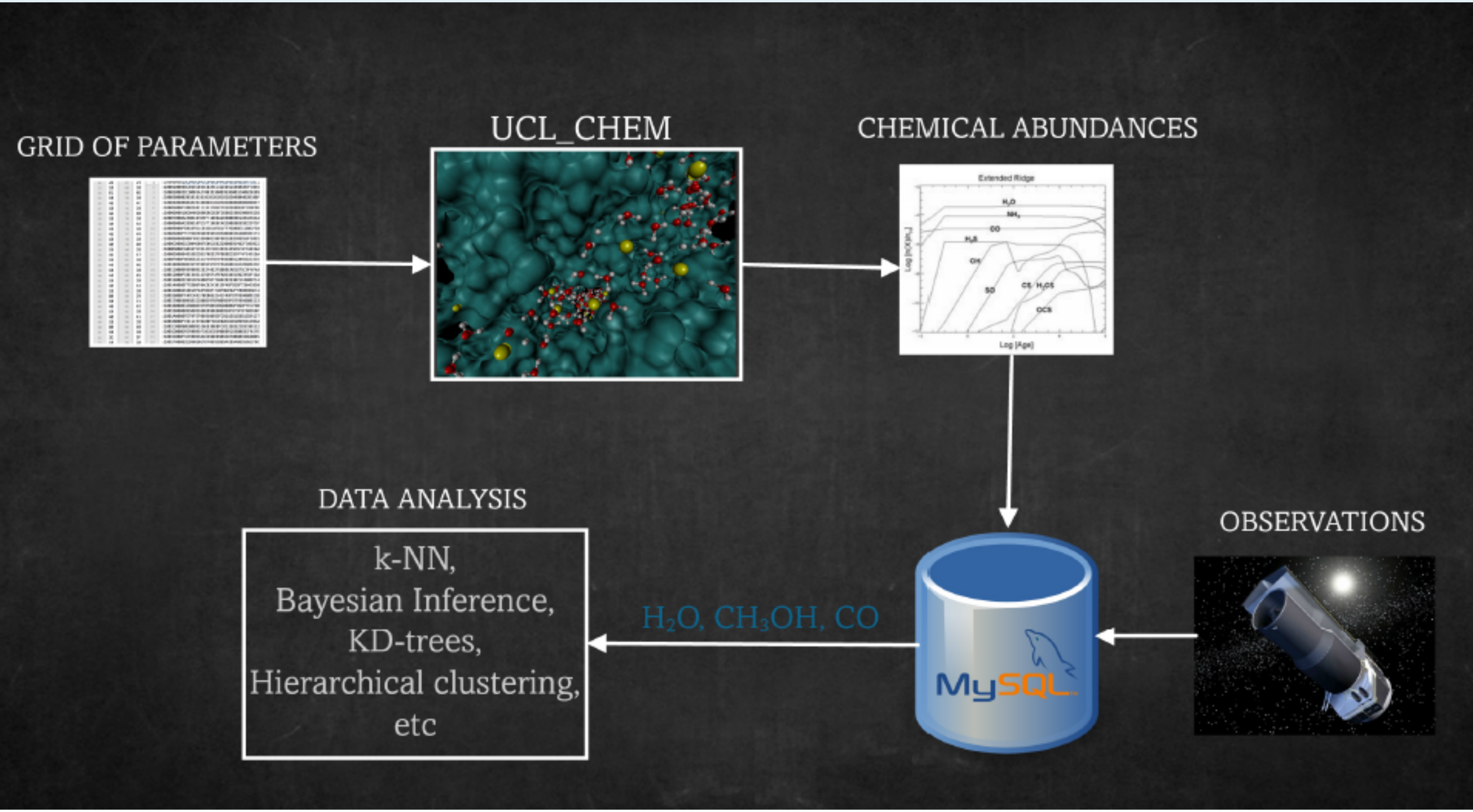
→ We have to deal with non-linear ill-posed inverse problems.

The Inverse Problem

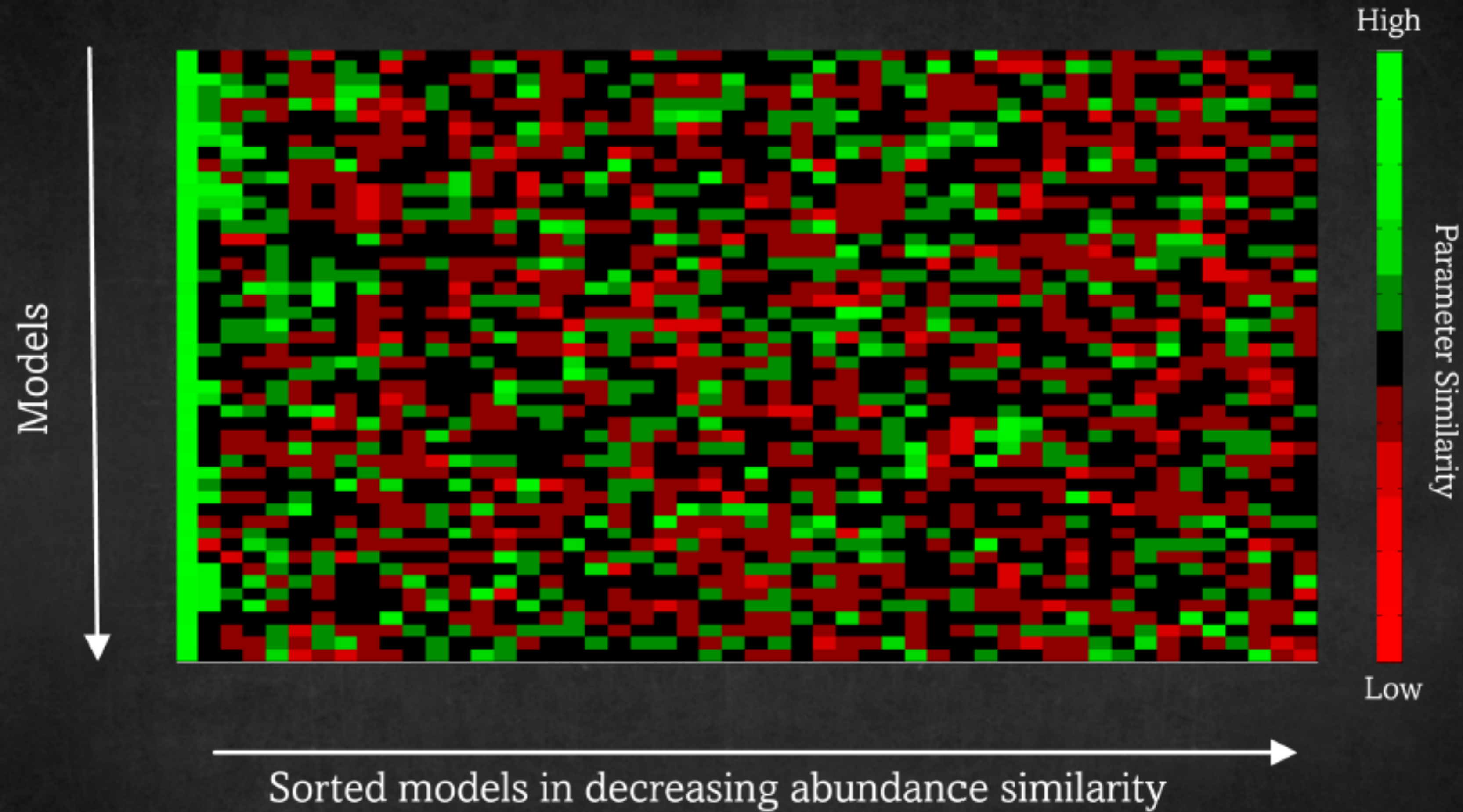


Given available observations what can we say about the physical parameters ?

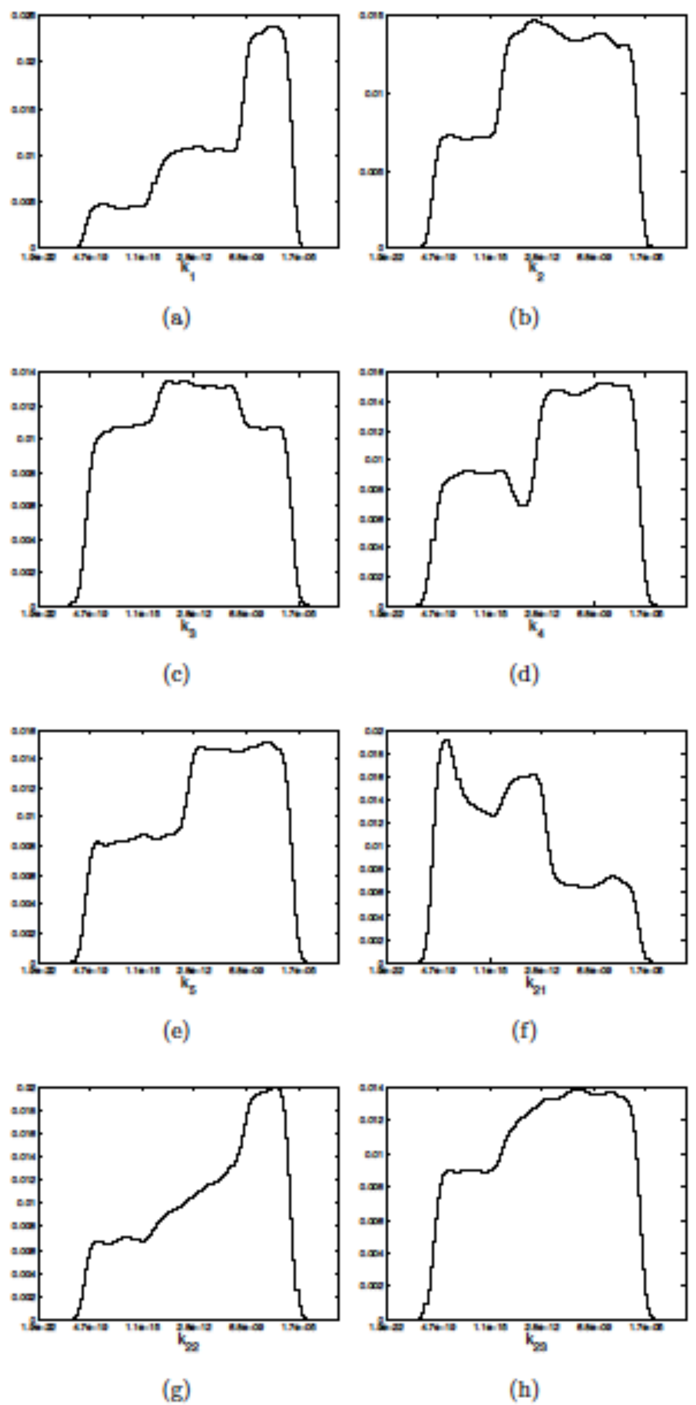

GRID OF PARAMETERS
UCL_CHEM
CHEMICAL ABUNDANCES
Extended Ridge
Log [Age]
OBSERVATIONS


GRID OF PARAMETERS
UCL_CHEM
CHEMICAL ABUNDANCES
DATA ANALYSIS
k-NN,
Bayesian Inference,
KD-trees,
Hierarchical clustering,
etc
H_2O, CH_3OH, CO
MySQL
OBSERVATIONS

The challenge of the inverse problem



Similar parameters might not give similar abundances
OR
Similar abundances might be produced by very different parameters



The first two proof of concept projects led to over a million chemical models (Makrymallis et al. 2014, 2016)

↓

This analysis led to a potential breakthrough in the way experimentalist astrochemists approach the problem of surface reactions.

e.g. 15 out of 23 reactions are not needed

No.	Reactions				
1.	O	+	H	→	OH
2.	OH	+	H	→	H_2O
3.	CO	+	OH	→	CO_2
4.	S	+	H	→	HS
5.	HS	+	H	→	H_2S
6.	H_2S	+	S	→	H_2S_2
7.	CS	+	H	→	HCS
8.	HCS	+	H	→	H_2CS
9.	CO	+	S	→	OCS
10.	OCS	+	H	→	HOCS
11.	H_2S	+	CO	→	OCS
12.	H_2S	+	H_2S	→	H_2S_2
13.	H_2S_2	+	CO	→	CS2 + O
14.	H_2S	+	O	→	SO_2
15.	CS_2	+	O	→	OCS + S
16.	CO	+	HS	→	OCS
17.	S	+	O	→	SO
18.	SO	+	O	→	SO_2
19.	SO	+	H	→	HSO
20.	HSO	+	H	→	SO
21.	CO	+	H	→	HCO
22.	HCO	+	H	→	H_2CO
23.	H_2CO	+	H	→	CH_3OH

Aims

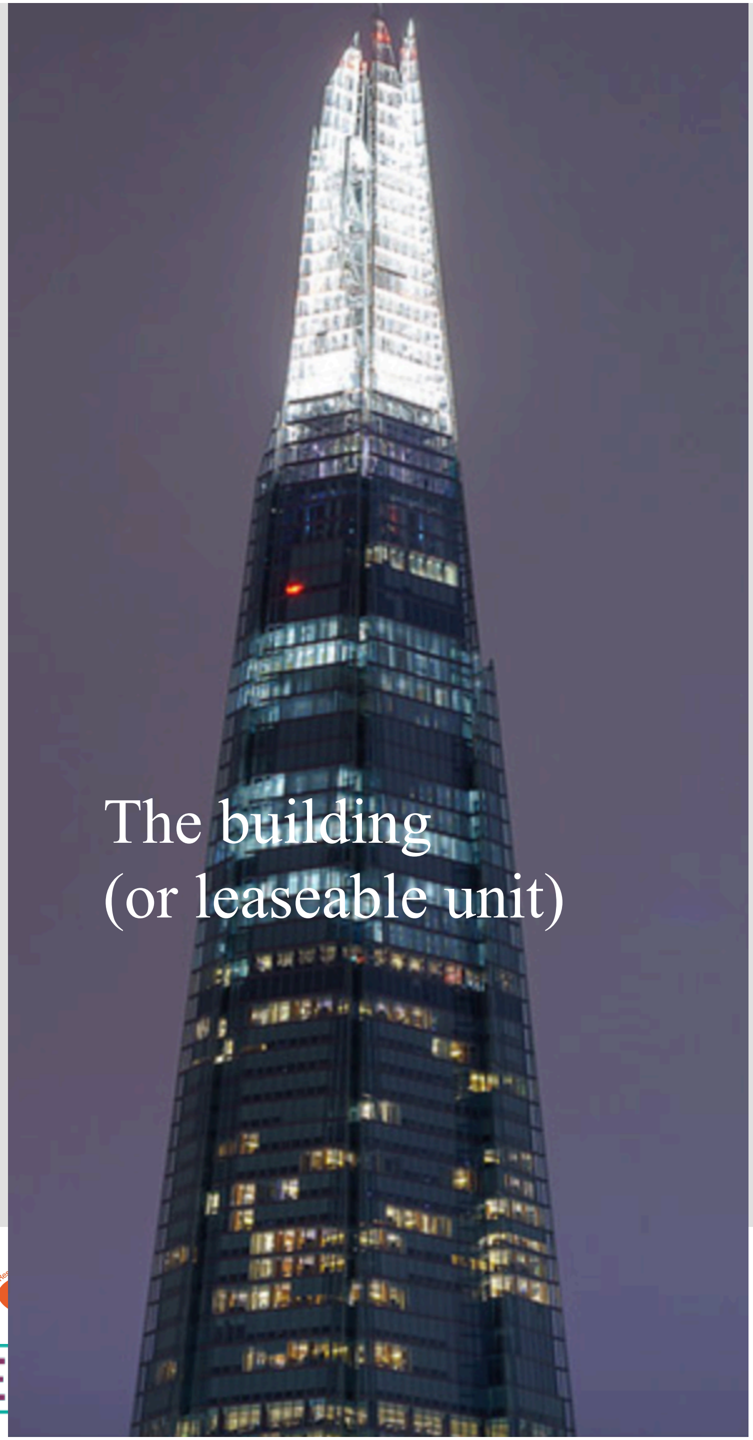
- Need to
  - Maximise the number of models we can run → essential for the accuracy and validity of statistical inferences
  - Perform rapid testing
  - Perform large scale sensitivity analyses
- In order to do that, we need to:
  - Perform innumerable simulations over a very large parameter space, generating a combinatorial explosion of model runs and large, high-dimensional data sets.

Generating Insight from Big Data in Energy and the Environment

David Wallom

Scale matters
for problems and solutions in the built environment

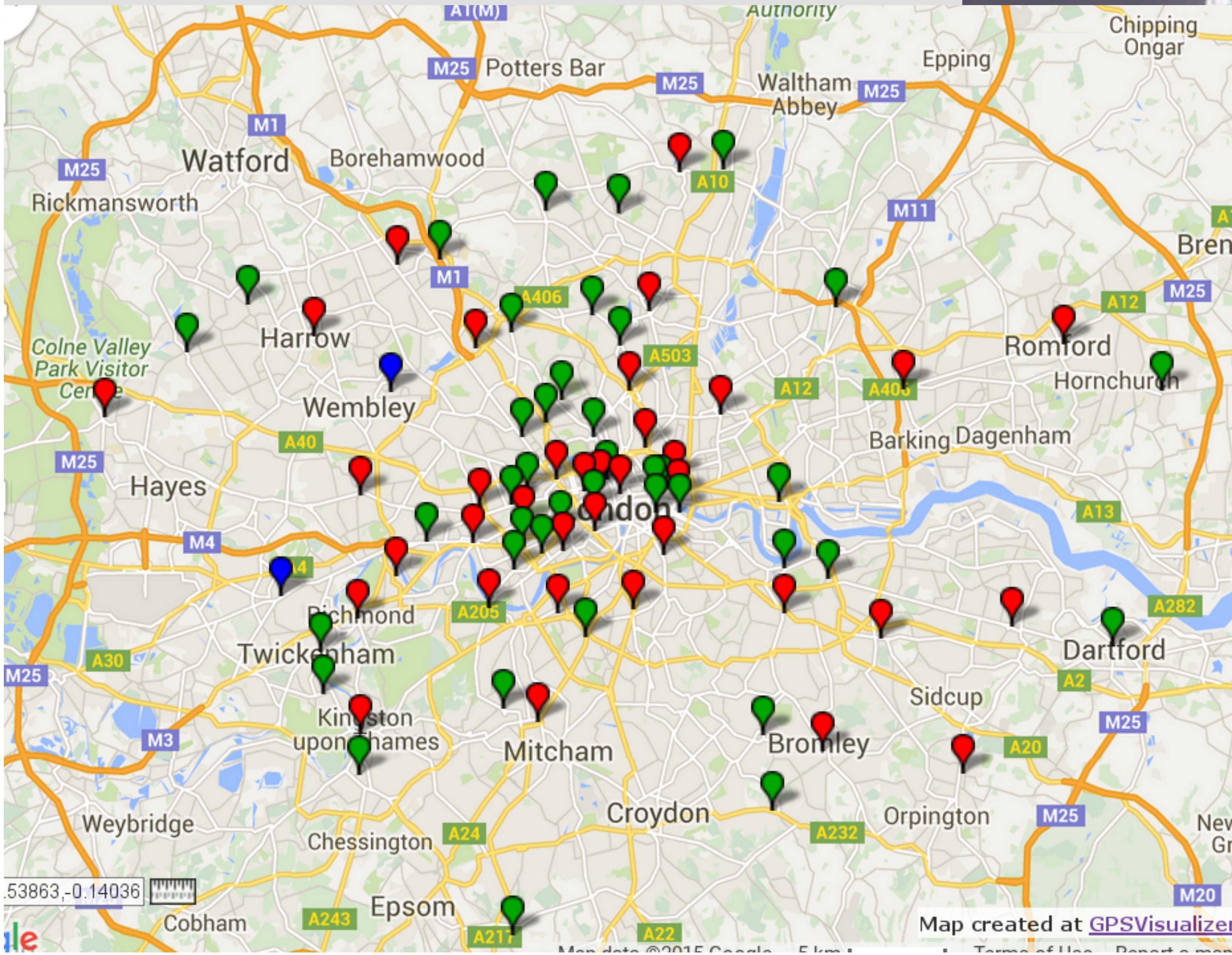
"stock" at the city, national, international scale

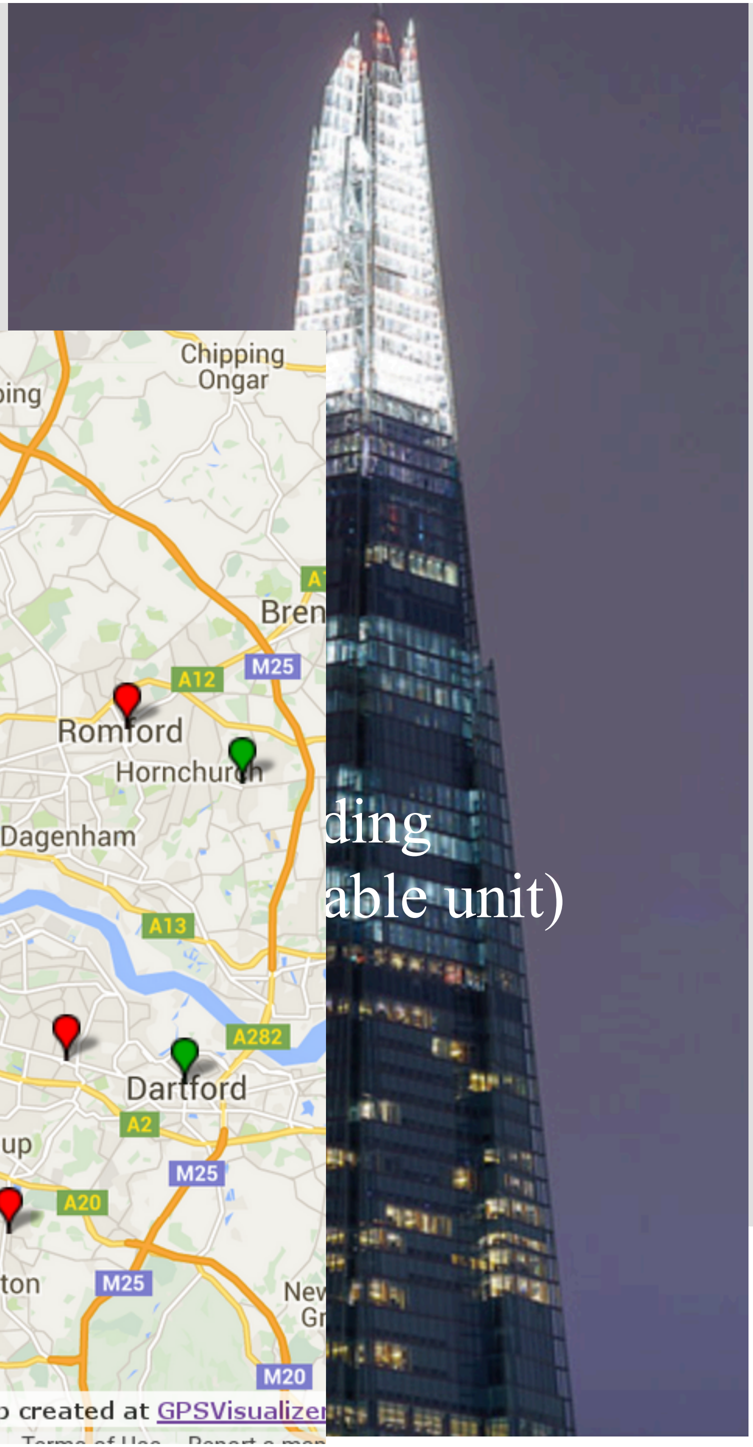


UNIVERSITY OF OXFORD

WICKED

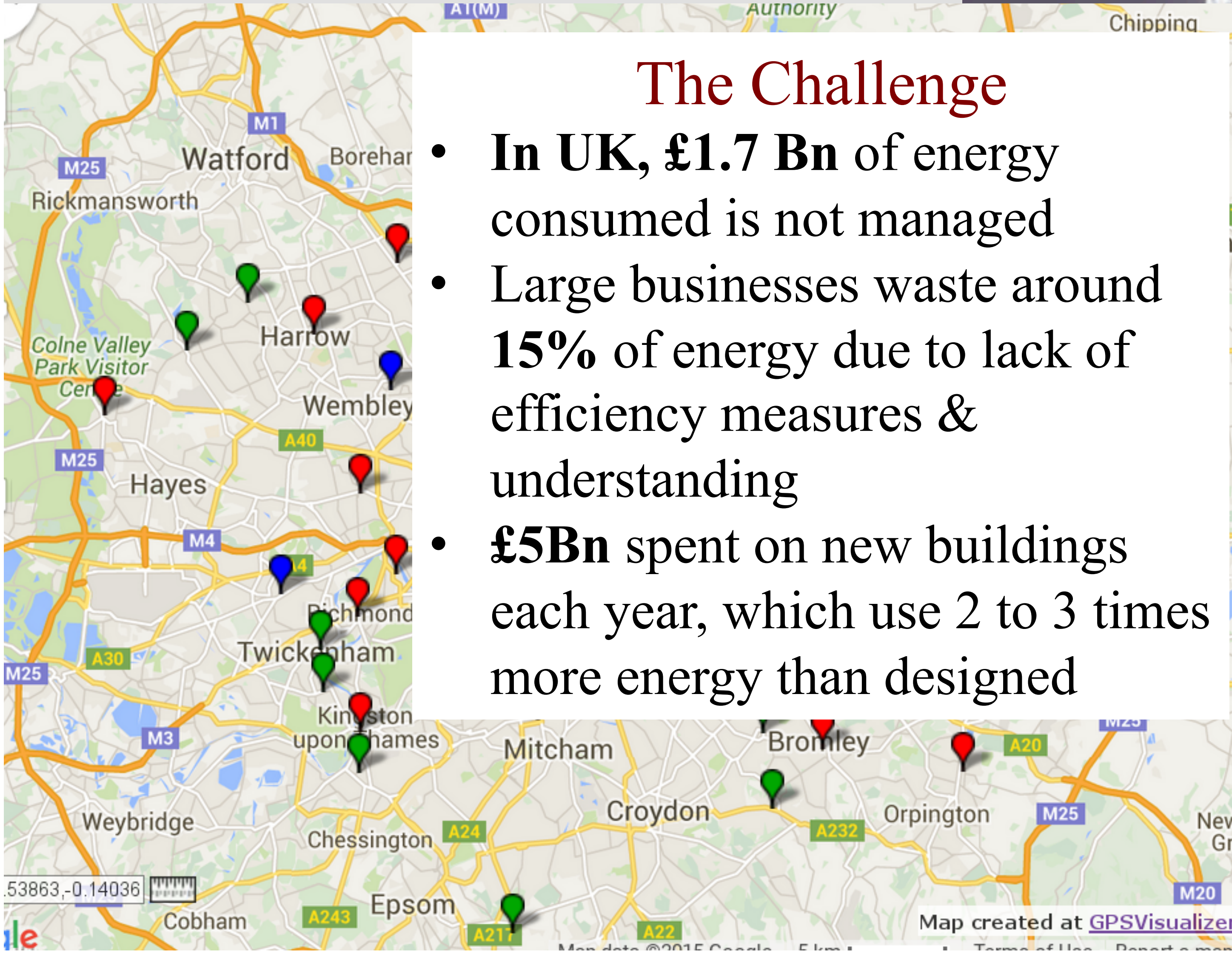
Scale matters
for problems and solutions
in the built environment

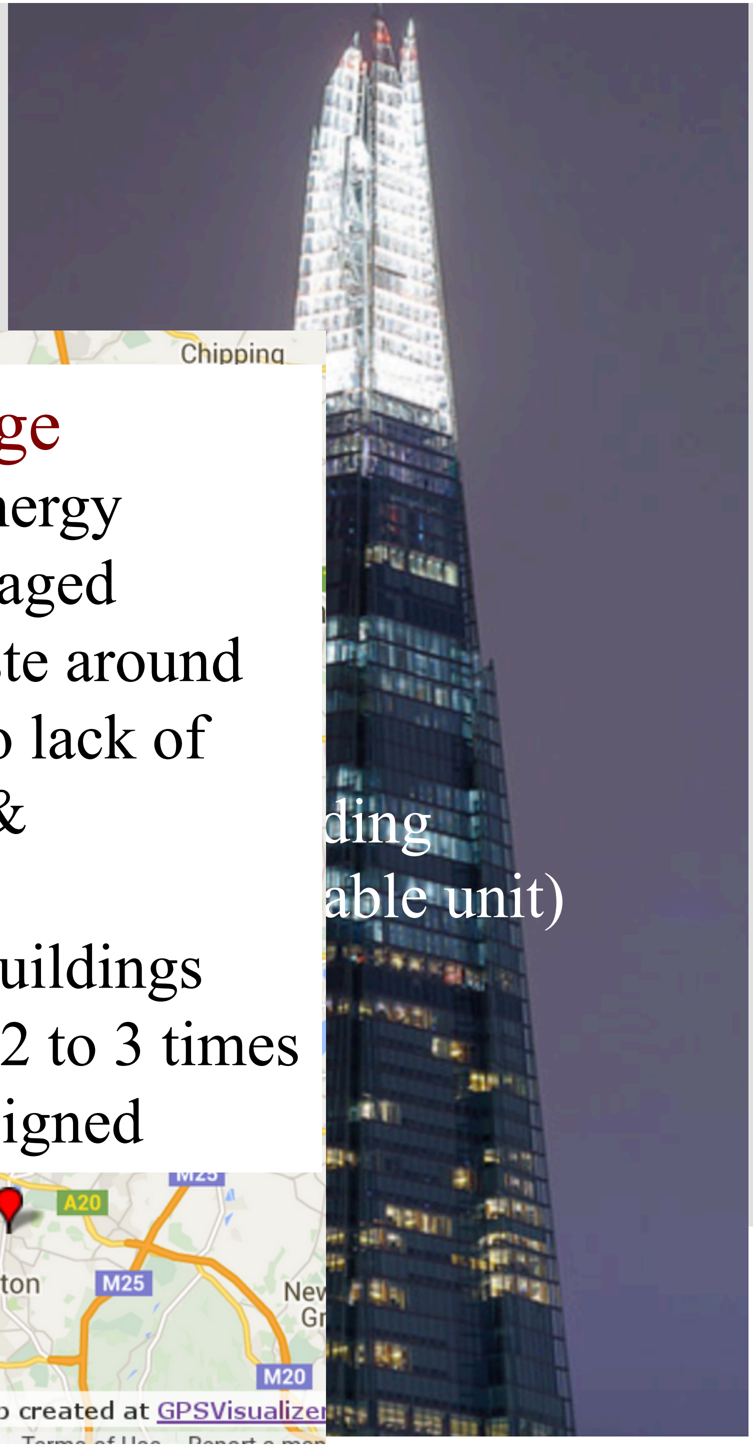




ding
able unit)

Scale matters
for problems and solutions
in the built environment

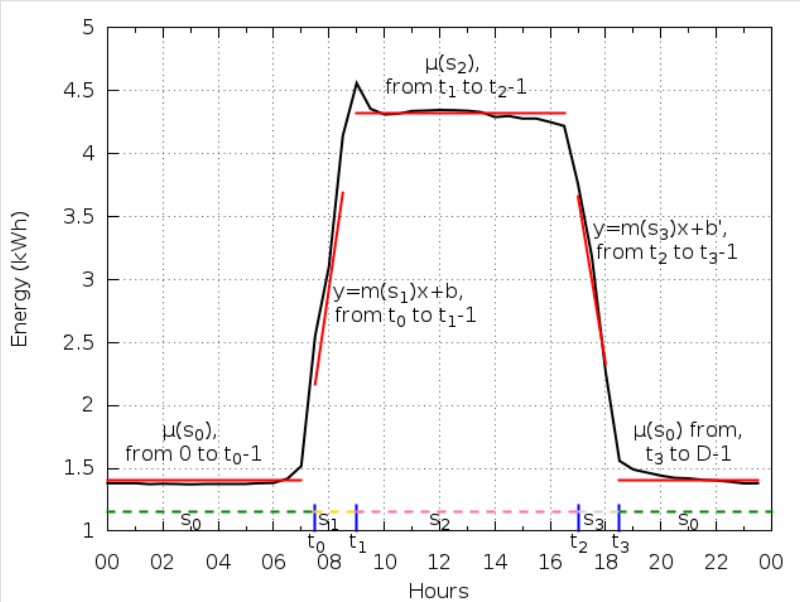


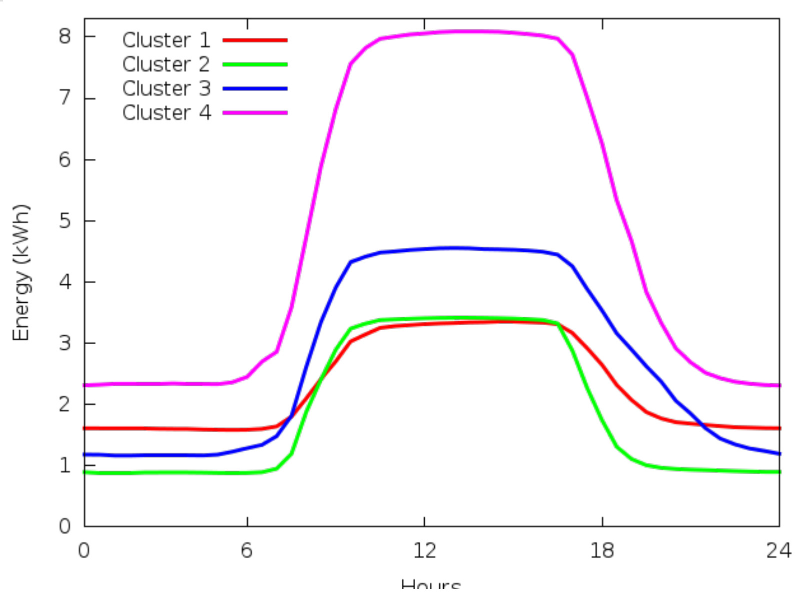


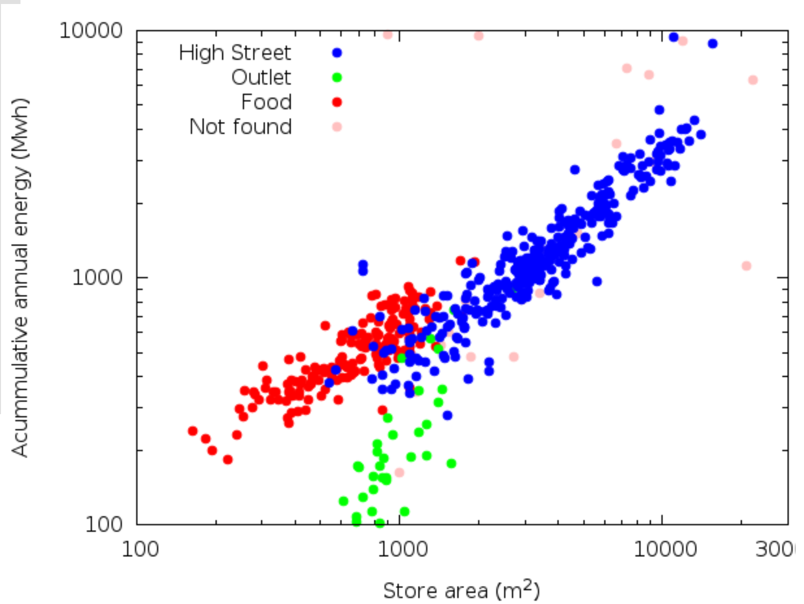
The Challenge

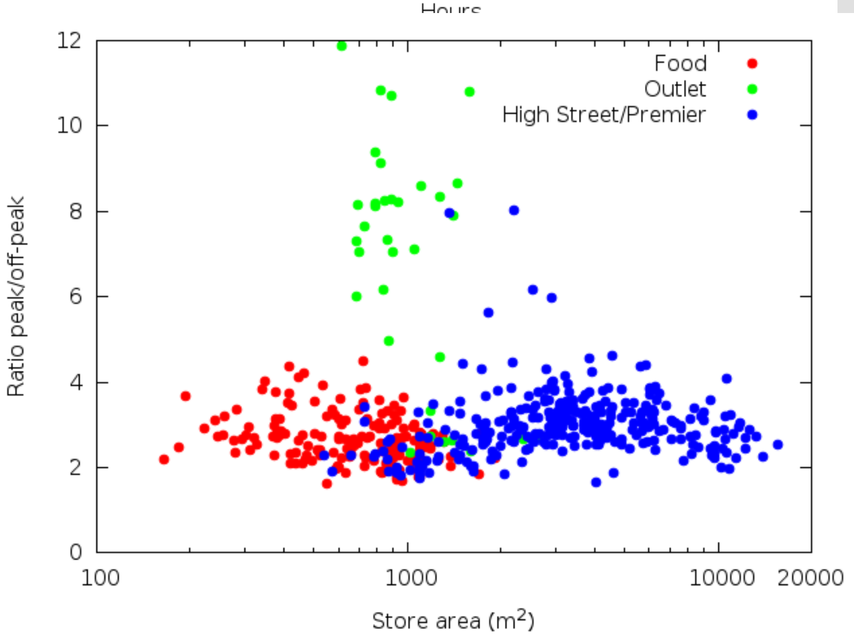
- **In UK, £1.7 Bn** of energy consumed is not managed
- Large businesses waste around **15%** of energy due to lack of efficiency measures & understanding
- **£5Bn** spent on new buildings each year, which use 2 to 3 times more energy than designed

Energy usage in retail premises

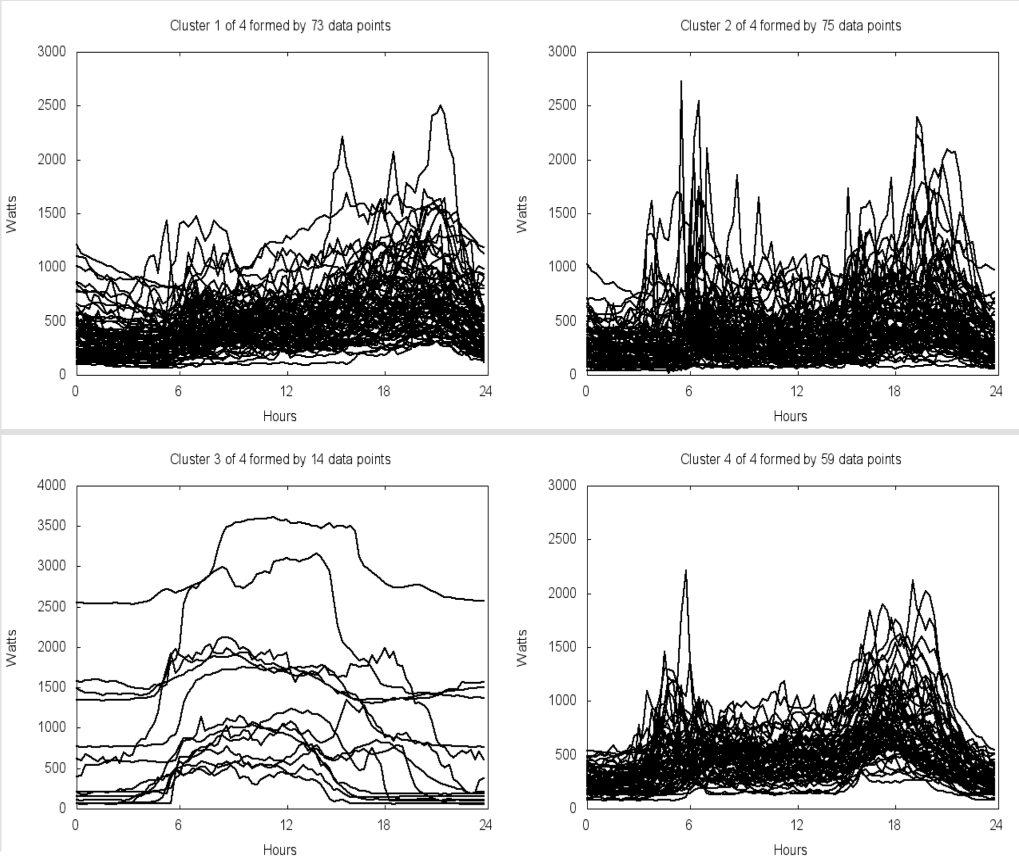


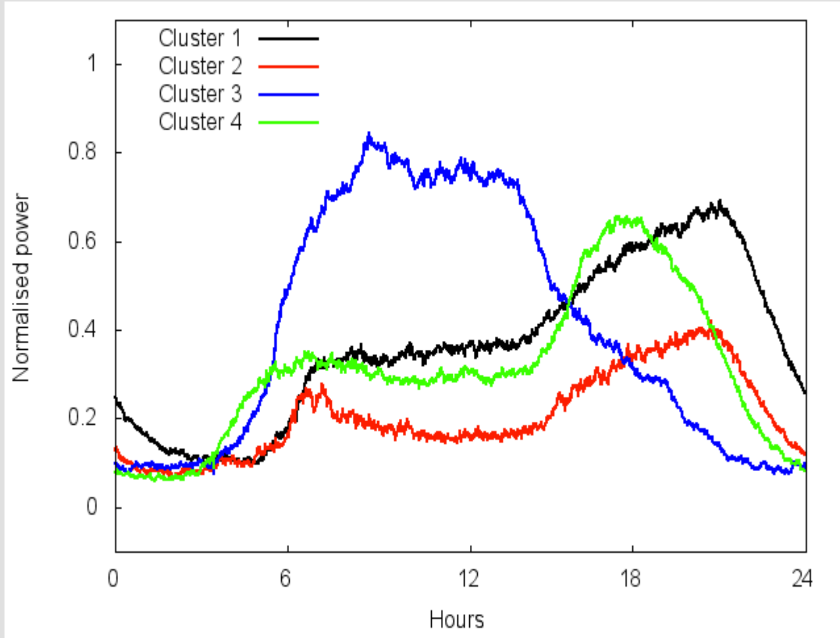






Clustering electricity load profiles using Bayesian clustering on domestic energy consumption





Data from EC FP7 DEHEMS

UNIVERSITY OF OXFORD

WICKED

Research Councils UK Energy For a Low Carbon Future

EPSRC

Clustering electricity load profiles using Bayesian clustering on domestic energy


BBC Sign in News Sport Weather iPlayer TV

NEWS ENGLAND

Home World UK England N. Ireland Scotland Wales Business Politics Health Education

20 January 2014 Last updated at 10:53

468 Share

Criminal gangs 'hotwire power supply' to help cut bills

As the row over energy prices grows ever more heated, a growing number of people are choosing to steal their gas and electricity.

Criminal gangs are helping homeowners and landlords avoid paying for power by "hotwiring" supplies for as little as £10, BBC Inside Out

Related Stories

Examples:

A black box tamper: A device, often concealed in a black box (hence the name), is fitted to an electricity meter to either stop the index, slow it down or even reverse the reading.

Index Tamper: Directly altering the recorded total consumption via meter breach

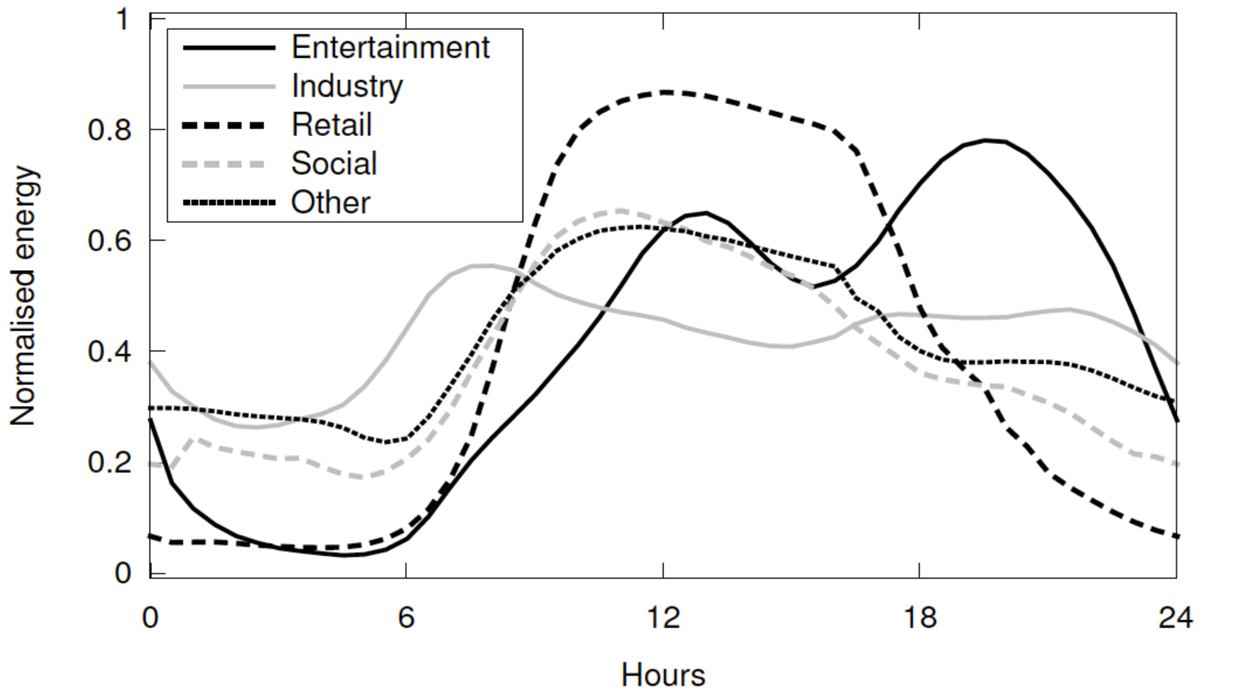
OXFORD

EPSRC

Commercial energy consumption and real time pricing

- Analyse the impact of introduction of time-of-use and real-time pricing strategies

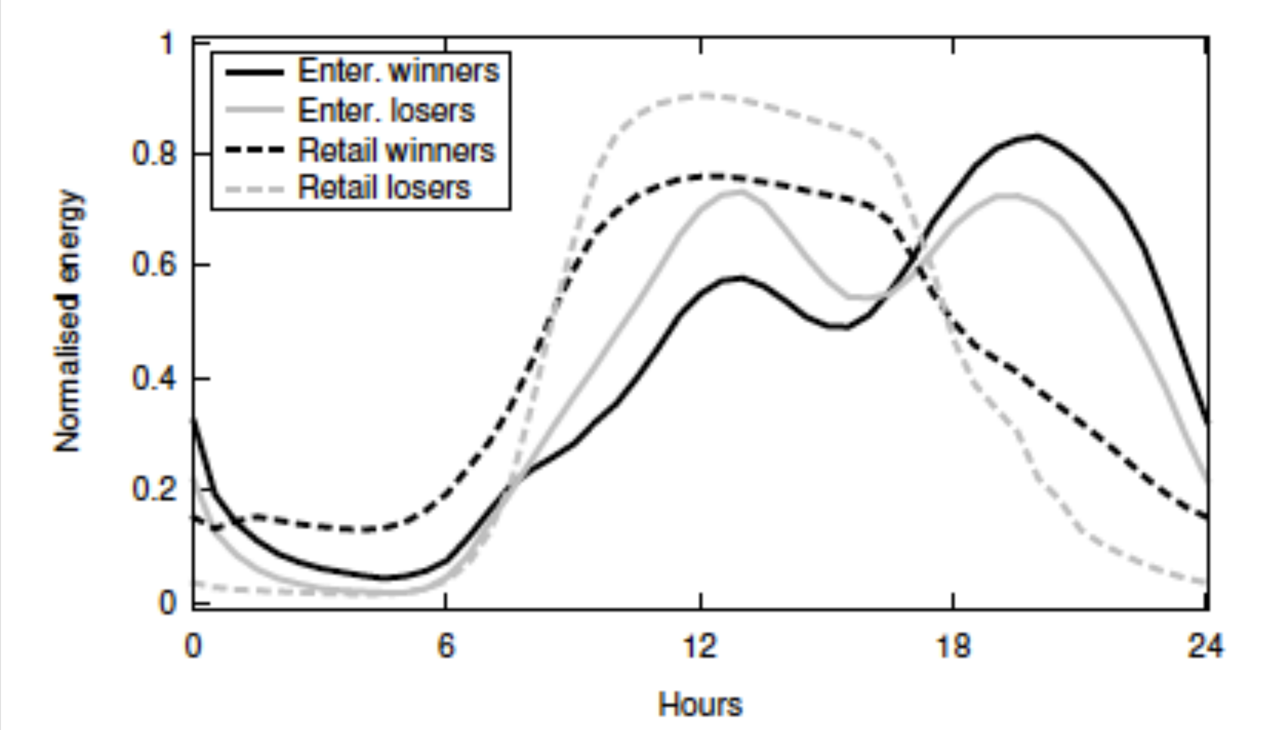
Normalised daily power demand profiles for all businesses by sector (Top Level SIC Classification)



Data from Opus Energy Ltd

Commercial energy consumption and real time pricing

- Analyse the impact of introduction of time-of-use and real-time pricing strategies

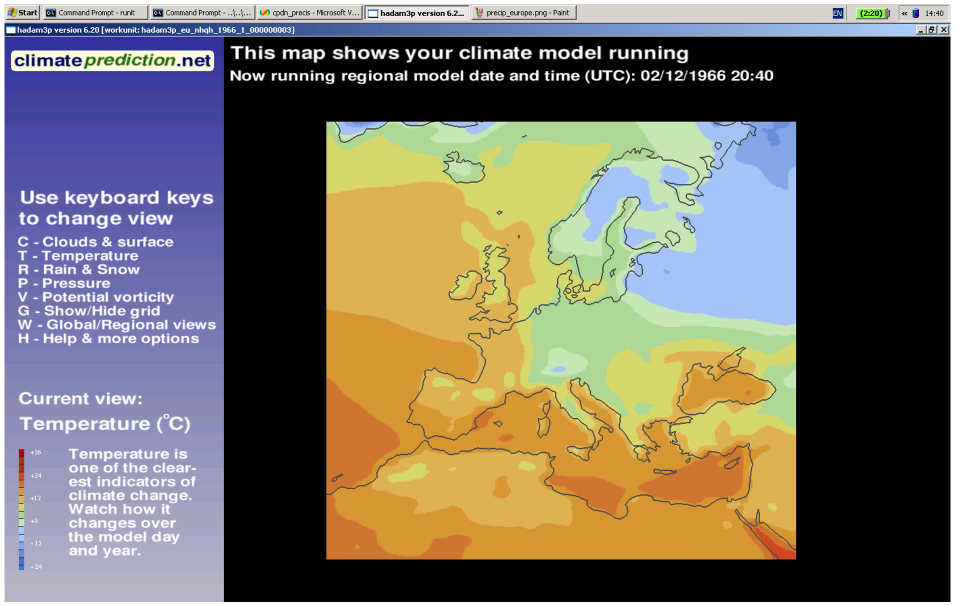


- Turning *Data into Actionable Information*;
  - Predicting and classifying costs with a shift in tariff type, e.g. shifting to a real-time tariff from a fixed price tariff,
  - Clustering of load profiles, determining behaviour type and/or consumer response, detecting energy theft
  - Determining fundamental drivers of energy consumption and improving understanding.
- Create commercial value

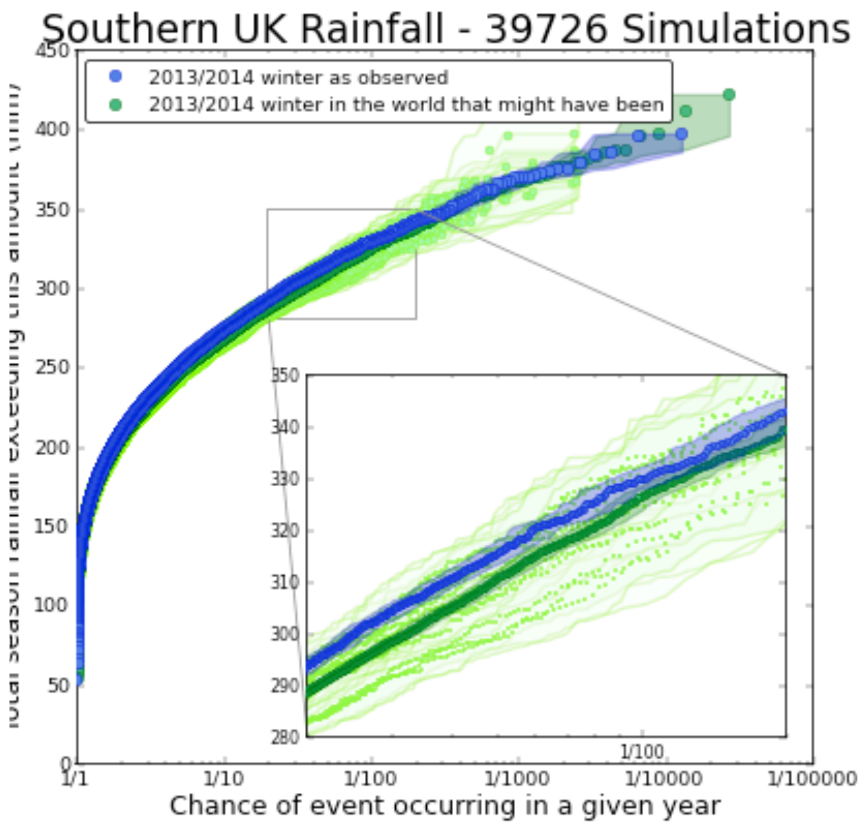
The weather@home regional modelling project

- High impact weather events are typically rare and unpredictable.
  - Flooding
  - Heatwave
  - Drought
- They also involve small scales.
- Resolution provided by nested regional model.

- Modify boundary conditions to mimic counter-factual "world that might have been".



UK Winter 2014 Floods

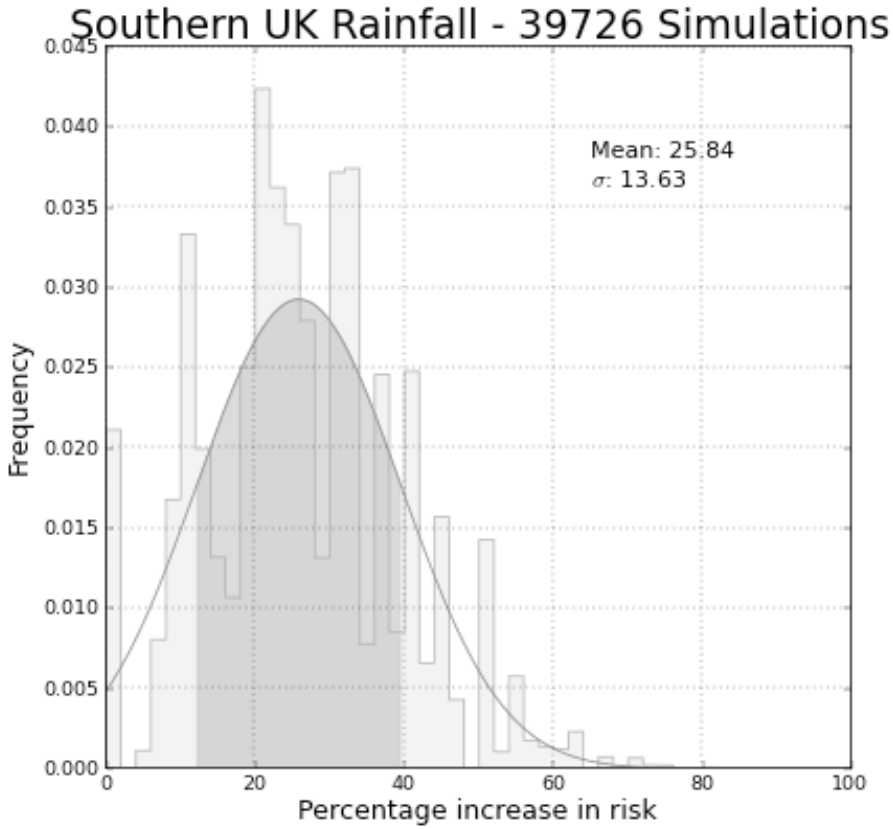


- 39726 simulations
- 2014 flooding described as a 1 in 100 year event in terms of rainfall volume
- Return time plot shows this has become a 1 in 80 year in terms of risk

climateprediction.net
the world's largest climate modelling experiment for the 21st century

UNIVERSITY OF OXFORD
eci
Oxford e-Research Centre
Department of Physics

UK Winter 2014 Floods



- 39726 simulations
- 2014 flooding described as a 1 in 100 year event in terms of rainfall volume
- Return time plot shows this has become a 1 in 80 year in terms of risk
- Risk of a very wet winter has increased by 25%

(Schaller et al, Jan 16, NCC)

climateprediction.net
the world's largest climate modelling experiment for the 21st century

University of Oxford
eci
Oxford e-Research Centre
Department of Physics

World Weather Attribution

A new international effort designed to sharpen and accelerate the scientific community's ability to analyze and communicate the possible influence of climate change on extreme-weather events such as storms, floods, heat waves and droughts.

California wildfires, 2014

A Multi-Method Approach

- Observational data, regional and global climate models.
- Provide answers about trends in risk and vulnerability, and the role of human activity in extreme weather.
- Possible outcomes of our attribution analysis of an event:
  - Global warming *increased its likelihood*.
  - Global warming *reduced its likelihood.*
  - Global warming had no detectable role.
  - Our analysis methods were unable to give information.

Malawi flood, 2015