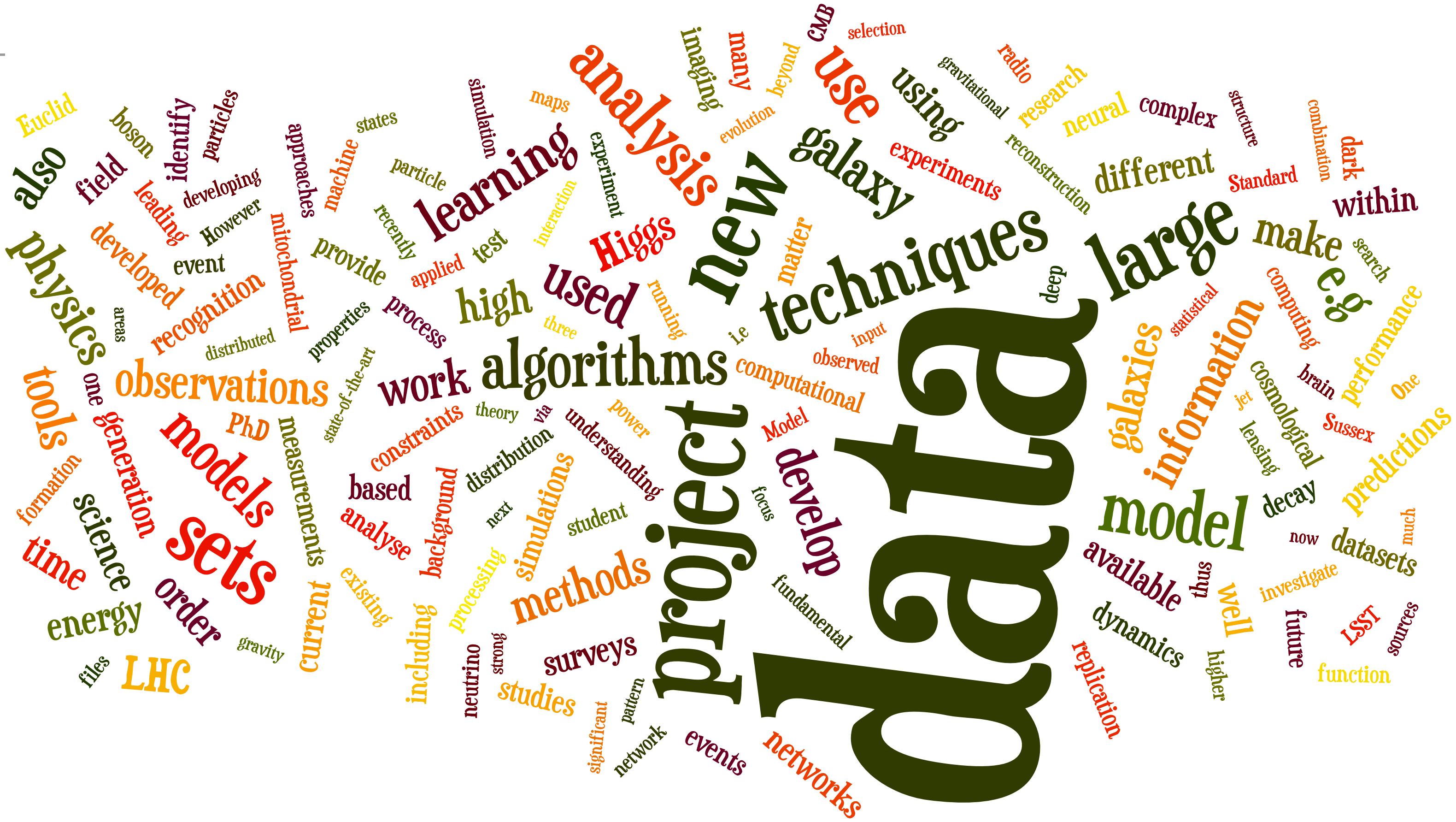
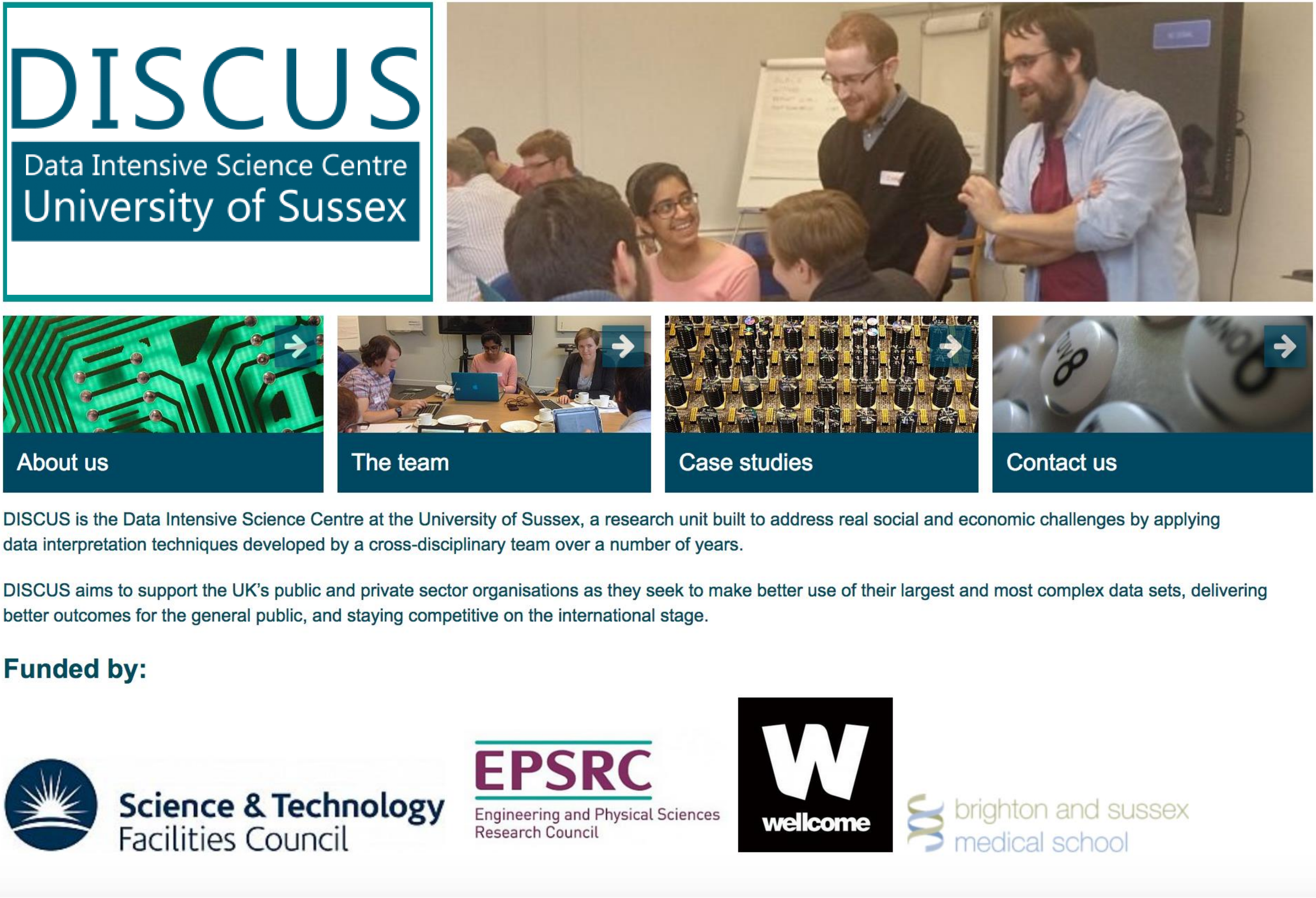
Science & Technology
Facilities Council

Data Science Skills from Astro, Particle and Nuclear Physics

Seb Oliver, University of Sussex

Science
Skills
Applications
How can we help





DISCUS

Data Intensive Science Centre
University of Sussex

DISCUS is the Data Intensive Science Centre at the University of Sussex, a research unit built to address real social and economic challenges by applying data interpretation techniques developed by a cross-disciplinary team over a number of years.

DISCUS aims to support the UK's public and private sector organisations as they seek to make better use of their largest and most complex data sets, delivering better outcomes for the general public, and staying competitive on the international stage.

Funded by:

Science & Technology Facilities Council

EPSRC Engineering and Physical Sciences Research Council

wellcome

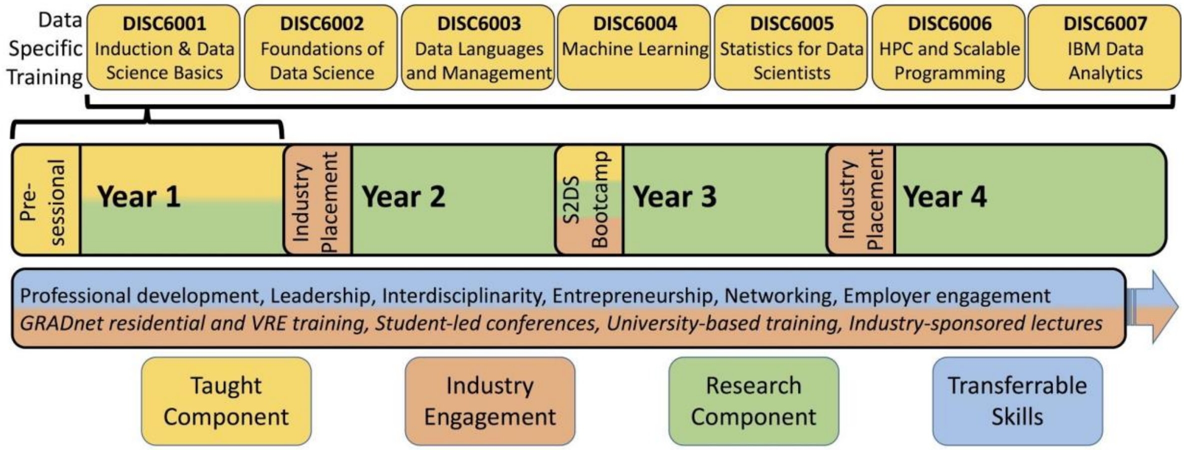
brighton and sussex medical school

DISC*net:* Data Intensive Science Centre in SEPnet

South East Physics Network

- STFC Data Intensive Science Centre call October 2016
- Sussex, Soton, QMUL, OU, Portsmouth, building on GRADnet
- 9% of all STFC activity in UK
- Leveraging 44 PhD years of STFC → 228 PhD years Training
- Data intensive training e.g. from IBM
- 132 months of commercial placements in 27 companies
- 5-week commercial transfer with pivigo
- Not just STFC. Sussex DISCUS interdisciplinary e.g. linking to GCRF
- Director Prof. Seb Oliver, University of Sussex (Also STFC ETCC Chair)
- Ranked 2nd in STFC competition
- **DISC***net* pilot being developed 56 PhD students and postdocs registered interest

UNIVERSITY OF SUSSEX



Data Specific Training

DISC6001 Induction & Data Science Basics

DISC6002 Foundations of Data Science

DISC6003 Data Languages and Management

DISC6004 Machine Learning

DISC6005 Statistics for Data Scientists

DISC6006 HPC and Scalable Programming

DISC6007 IBM Data Analytics

Pre-sessional | Year 1 | Industry Placement | Year 2 | S2DS Bootcamp | Year 3 | Industry Placement | Year 4

Professional development, Leadership, Interdisciplinarity, Entrepreneurship, Networking, Employer engagement

GRADnet residential and VRE training, Student-led conferences, University-based training, Industry-sponsored lectures

Taught Component

Industry Engagement

Research Component

Transferrable Skills

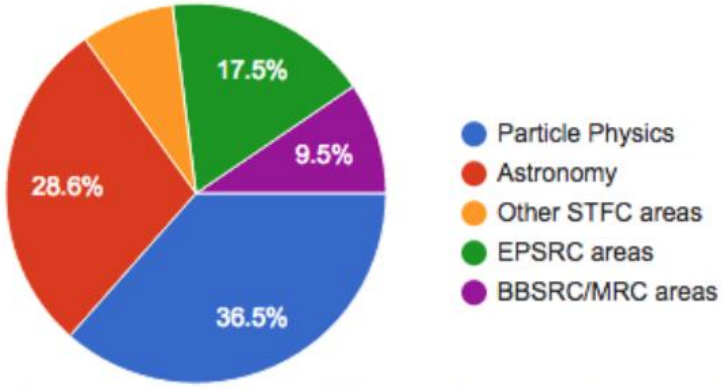
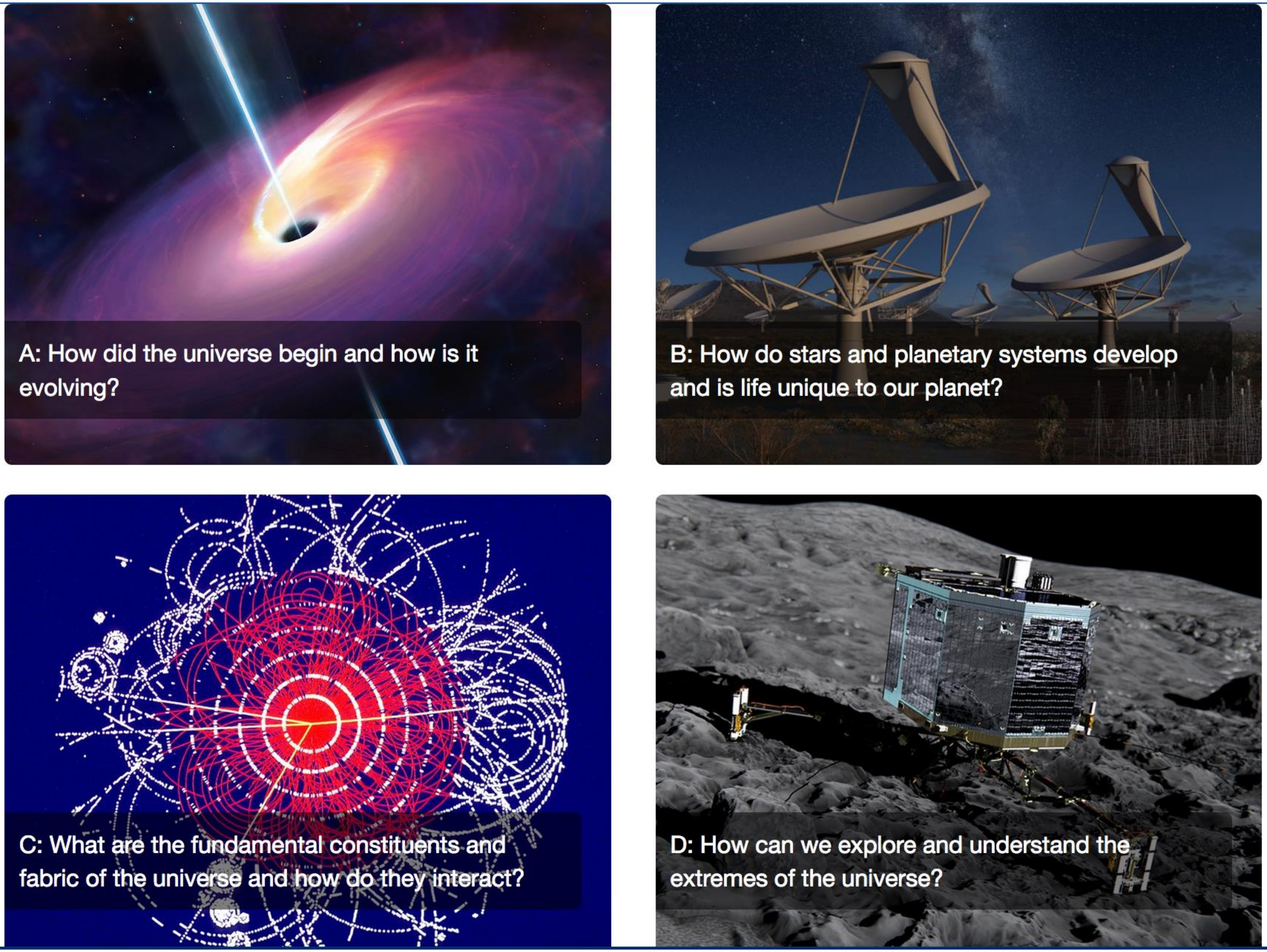


Figure 1: Breakdown of the broad research areas for 77 projects proposed for DISCnet

Partner
Ageas
Adler Planetarium
Ambiental
CCFE
Critical
DEIMOS Space
HMRC
IBM UK
IEA
Kent CC
Knowwow
Lein
MeVitae
MP Capital
NAOC
National Crime Agency
NPL
Ordnance Survey
Pivigo
PRDICT
Rank
SAC
Senseye
STFC-SCD
Thales
TUI
Viridan

STFC Science Challenges
http://www.stfc.ac.uk/research/science-challenges/

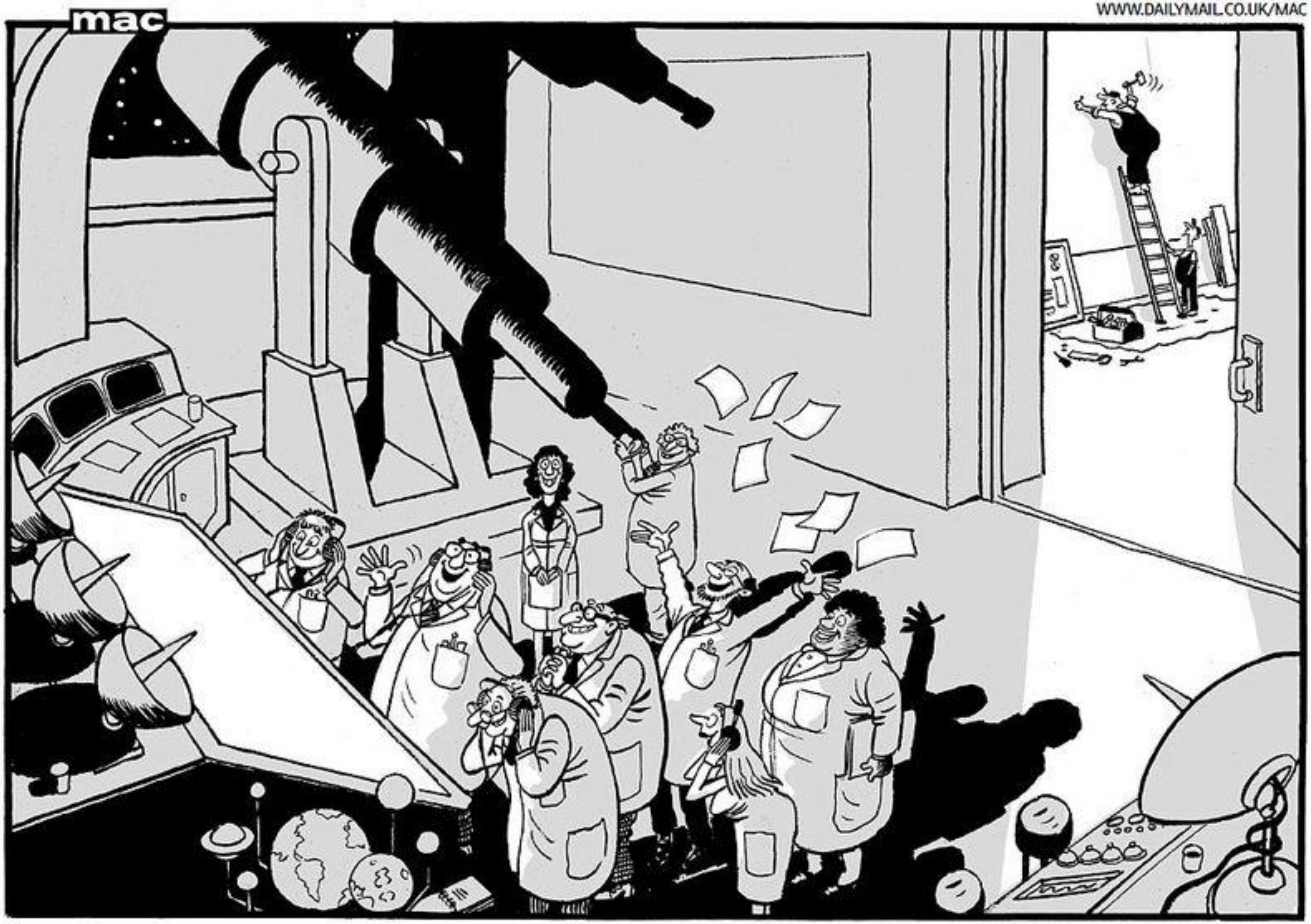


A: How did the universe begin and how is it evolving?

B: How do stars and planetary systems develop and is life unique to our planet?

C: What are the fundamental constituents and fabric of the universe and how do they interact?

D: How can we explore and understand the extremes of the universe?



'Listen! There they are again – echoes of the Big Bang. The beginning of creation!'

Astronomy & Particle Physics key Data skills

- Used to handling very big data sets
- Cradle to grave
- Raw data processing
- Image analysis
- Object detection
- Object classification
- Machine learning for classification and regression
- Application of statistics to research problems
- Bayesian methodologies for
  - Model parameter estimation
  - Hierarchical probabilistic modelling
  - Model selection
- Numerical simulation on massive scales
- In general highly numerate and computationally skilled with access to a vast “toolkit” of methods and experienced in applying these methods to real research problems

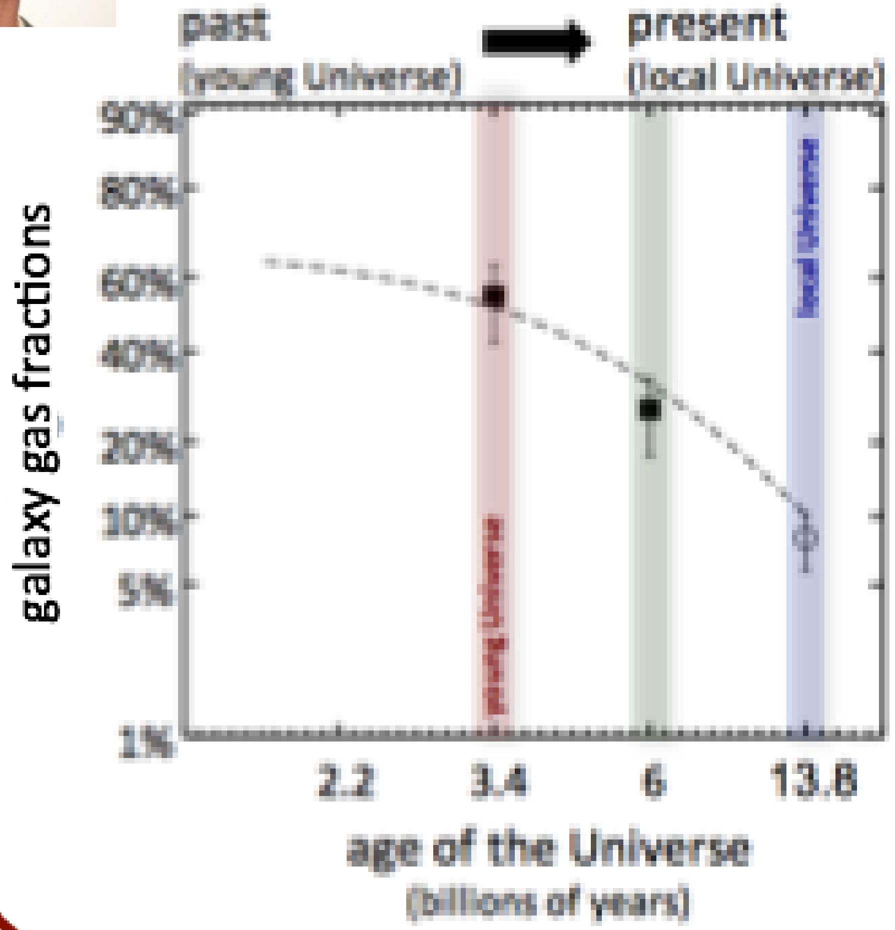
CRADLE TO THE GRAVE

Gas and star formation in galaxies through cosmic time

(supervisor: Mark Sargent)

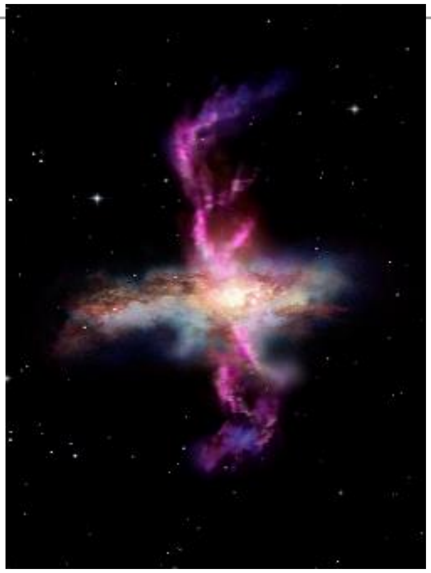
Gradual gas consumption in spiral galaxies?

Central question: *How did it come to this…?*



Rapid/efficient gas consumption in merger-induced starbursts?

Removal of gas by feedback, e.g. from AGN or star-formation?

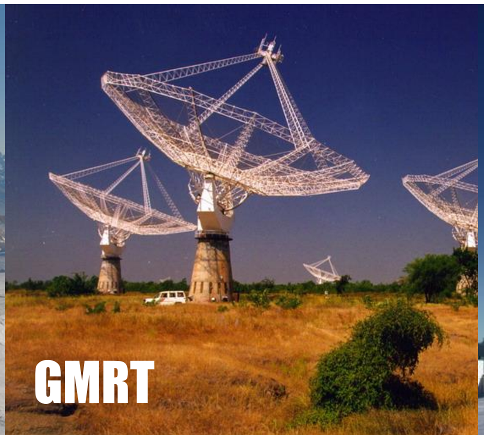


Stripping of gas reservoirs in dense environments, e.g. in galaxy groups?

JCMT

JVLA

IRAM/NOEMA


GMRT

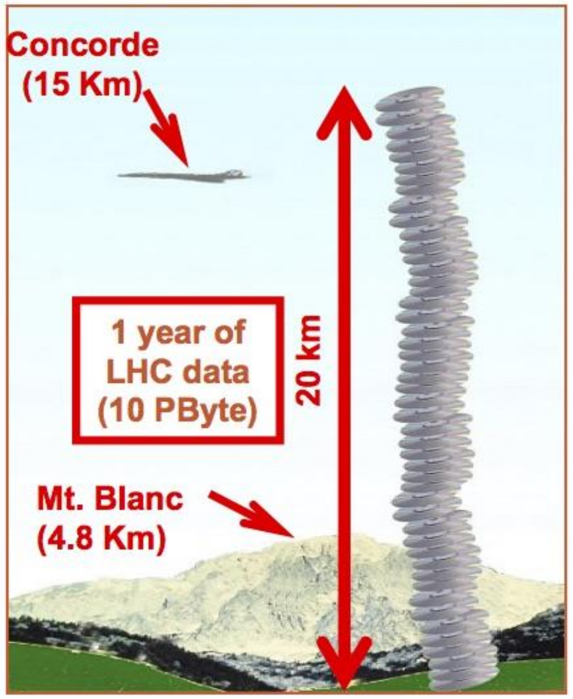
IRAM/30m

ALMA

BIG DATA SETS
RAW DATA PROCESSING &
IMAGE ANALYSIS

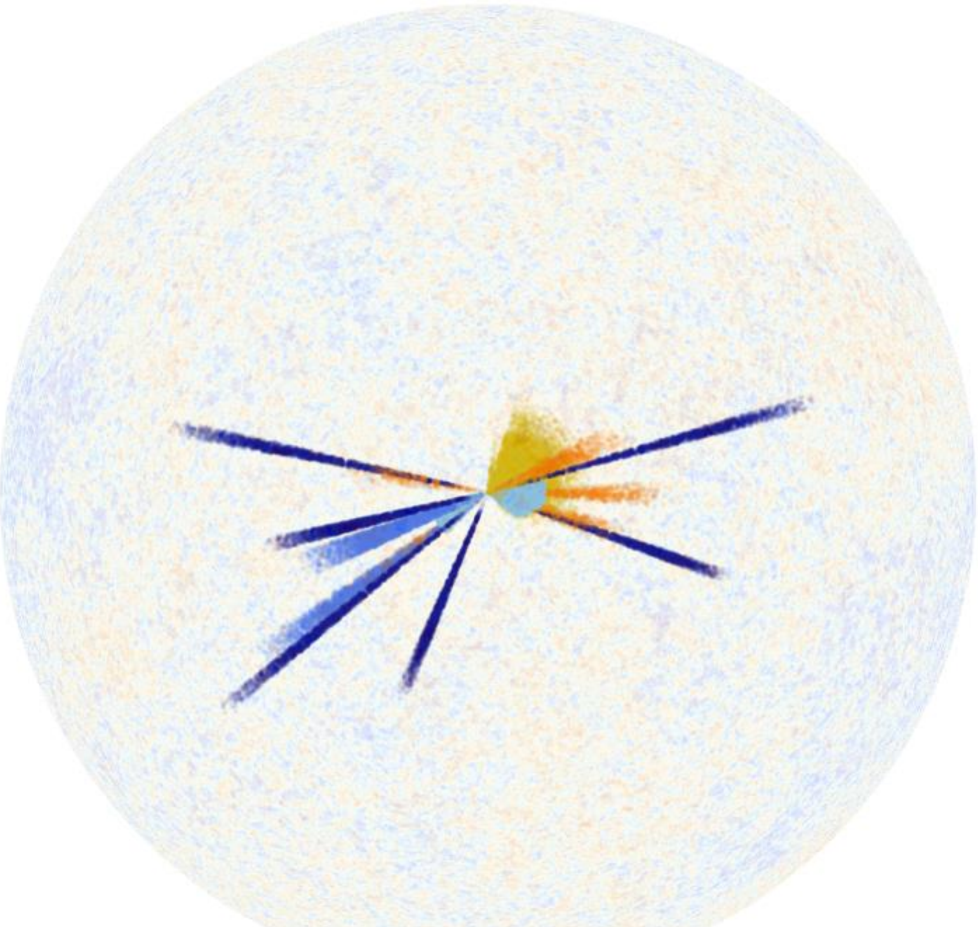
Dealing with large data sets…

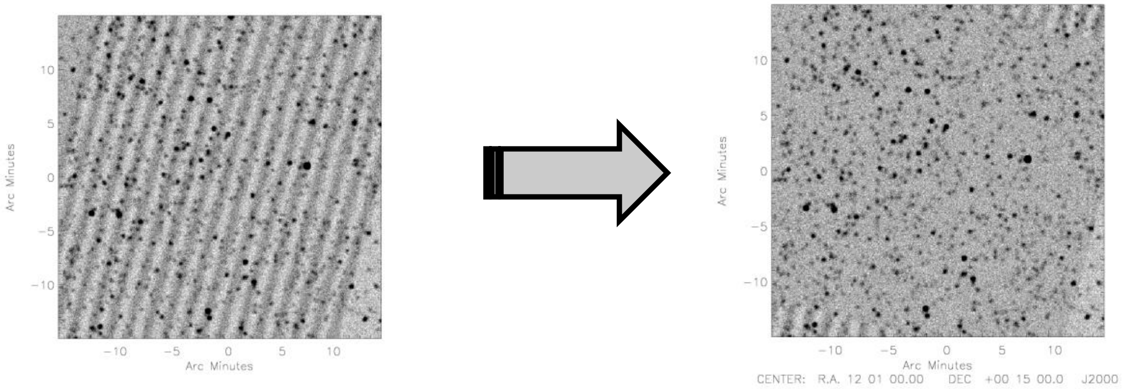
- In every single research discipline there is more and more discussion about big data
  - Try googling ‘Big Data’ !!



- In High Energy Physics there is a long tradition of dealing with large data sets
  - Big HEP Experiments – e.g. LHC, Lep (CERN), Tevatron (Fermilab), etc.
  - Simulation of Monte Carlo events for future studies – e.g. Linear Collider, future neutrino programme, etc.
  - Grid computing – interconnected computers used to analyse large data sets

- During the years there has been a lot of development of advanced statistical techniques to deal with the analysis of large data sets





OBJECT DETECTION, MEASUREMENT & CLASSIFICATON

Kathy Romer: Cosmology using observations of clusters of galaxies (Dark Energy Survey)

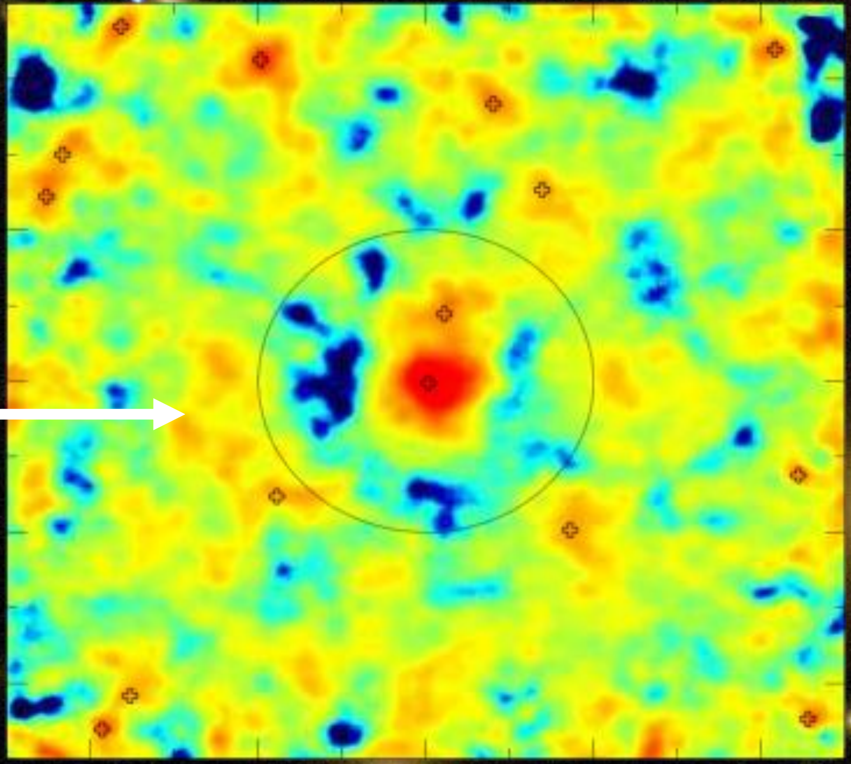
- DES is the leading Dark Energy experiment in world
- 5000 square degrees of deep imaging (optical and near IR)
- =

Combining cosmic shear and large-scale structure data to constrain the acceleration of the Universe. Supervisor: Dr Robert E. Smith

We will explore how the combination of weak lensing aperture mass statistics and redshift space distortions can help shed light on this great mystery.

Is Dark Energy quantum vacuum, scalar field or a modification to General Relativity?

Aperture mass map showing the presence of a cluster in the mock CFHTLens data

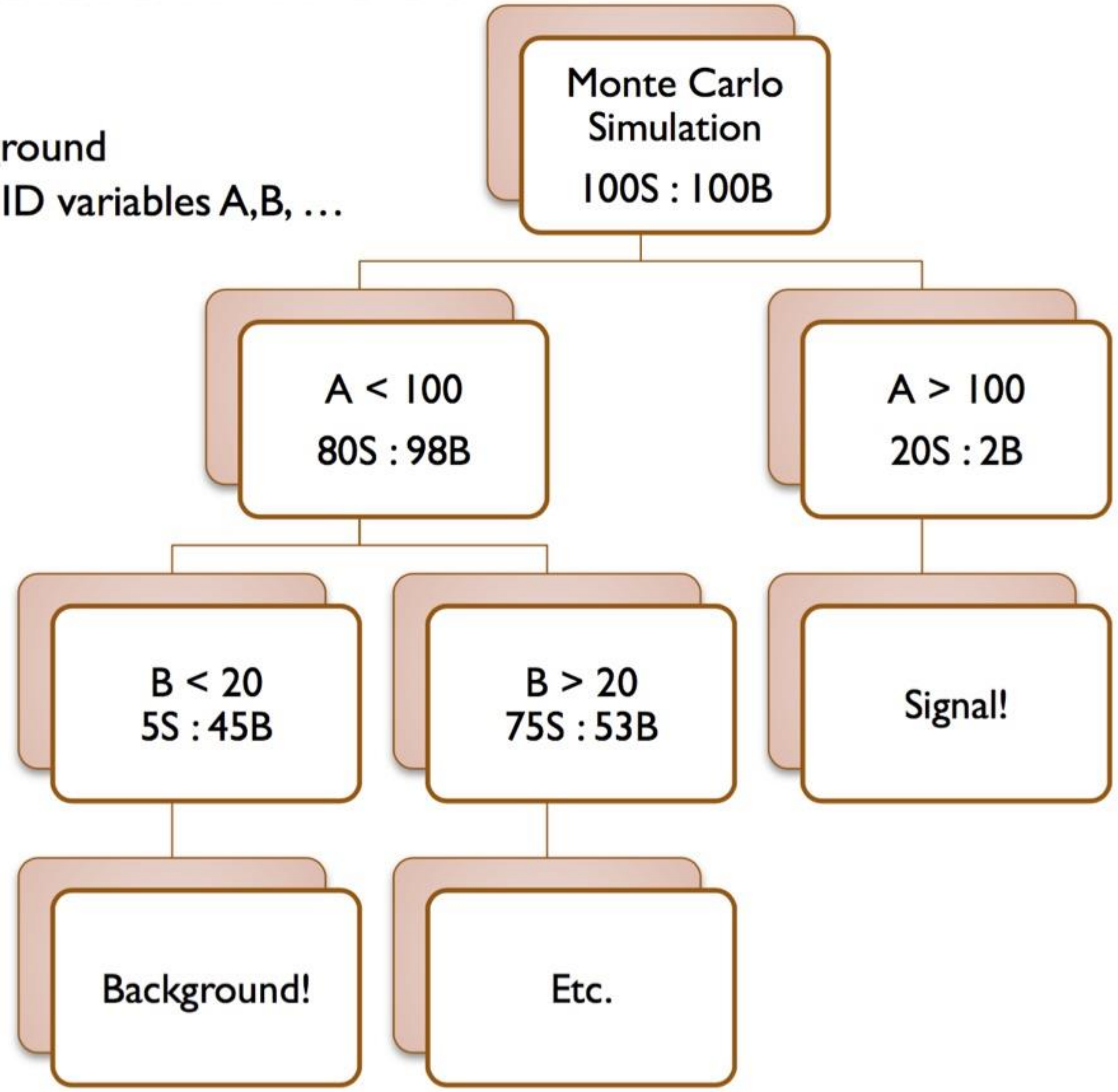


Project will involve analytic calculations, numerical simulations and data analysis.

MACHINE LEARNING

Boosted Decision Trees (BDT)

S: Signal
B: Background
Particle ID variables A,B, …



Monte Carlo Simulation
100S : 100B

A < 100
80S : 98B

A > 100
20S : 2B

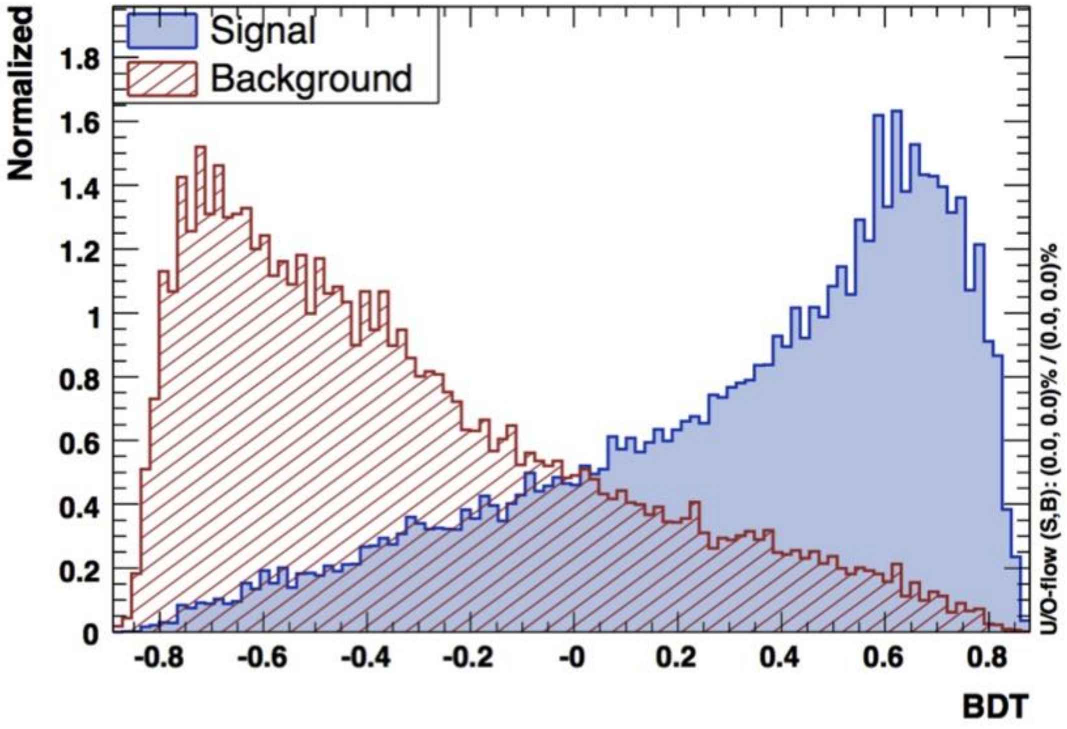
B < 20
5S : 45B

B > 20
75S : 53B

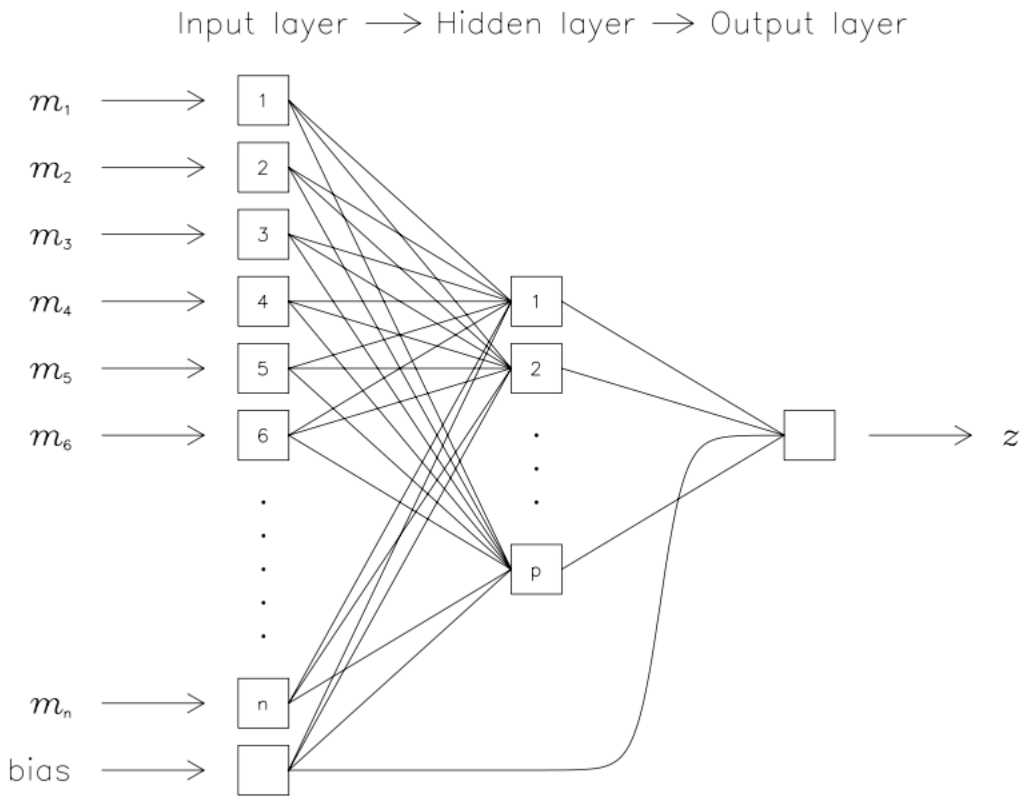
Signal!

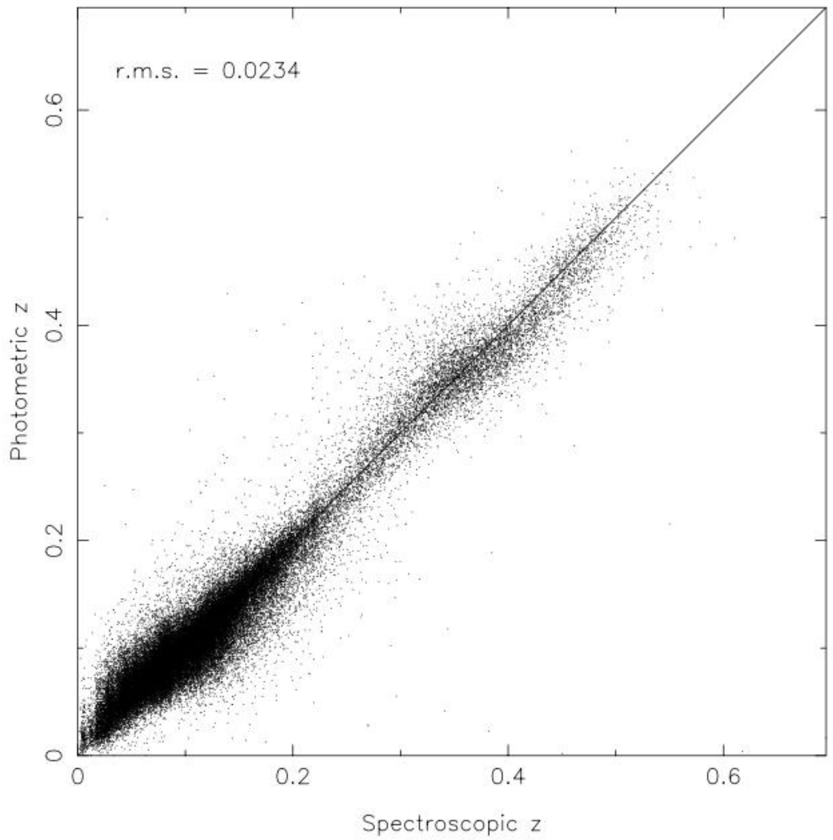
Background!

Etc.



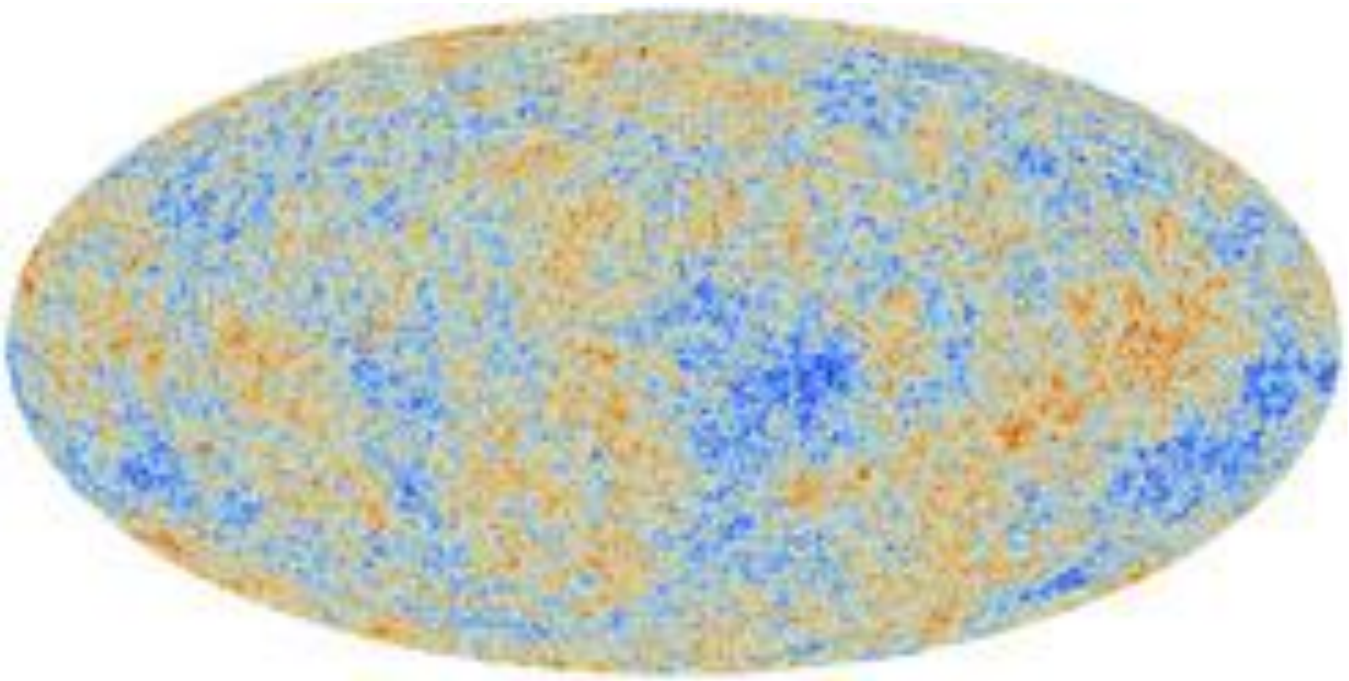
Neural Networks for galaxy Distances



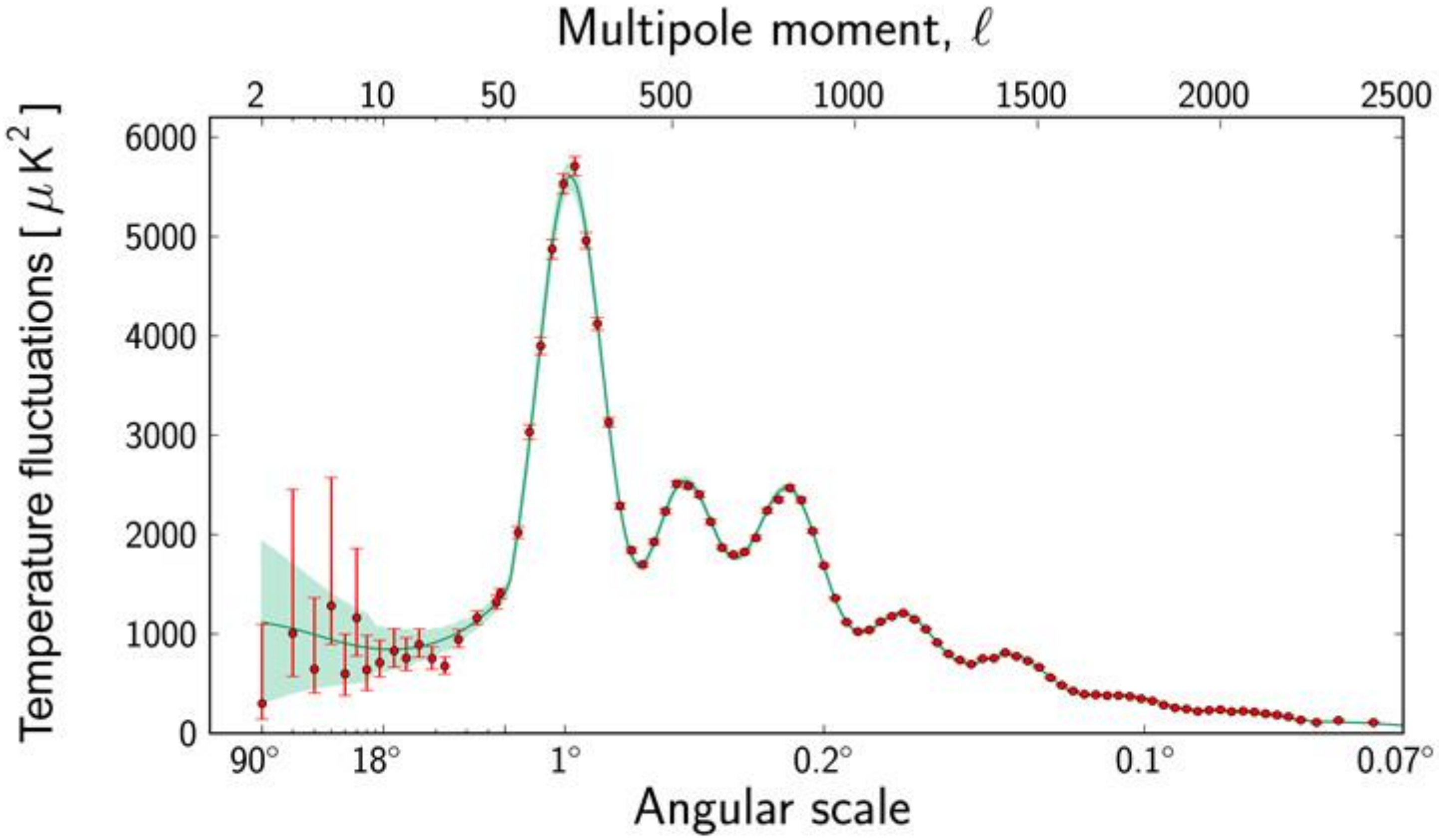


ANN-z Collister et al.

APPLICATION OF STATISTICS TO RESEARCH PROBLEMS



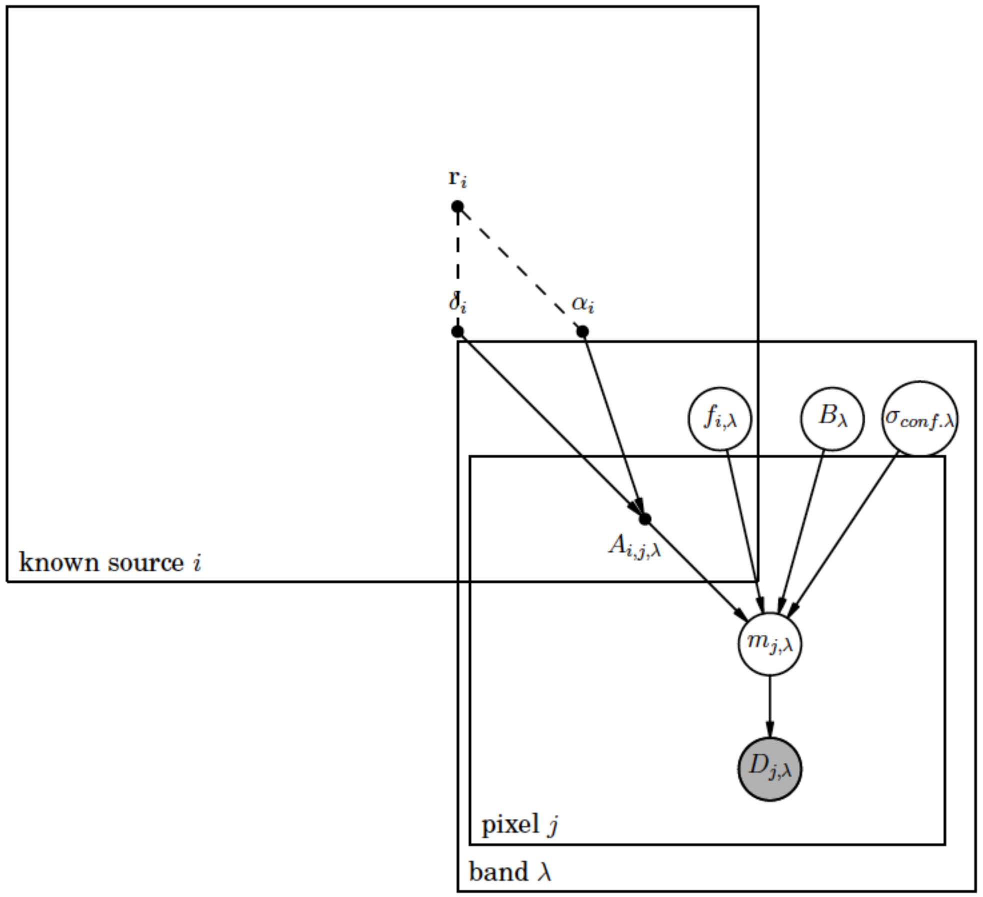
Map of the Cosmic Infrared Background

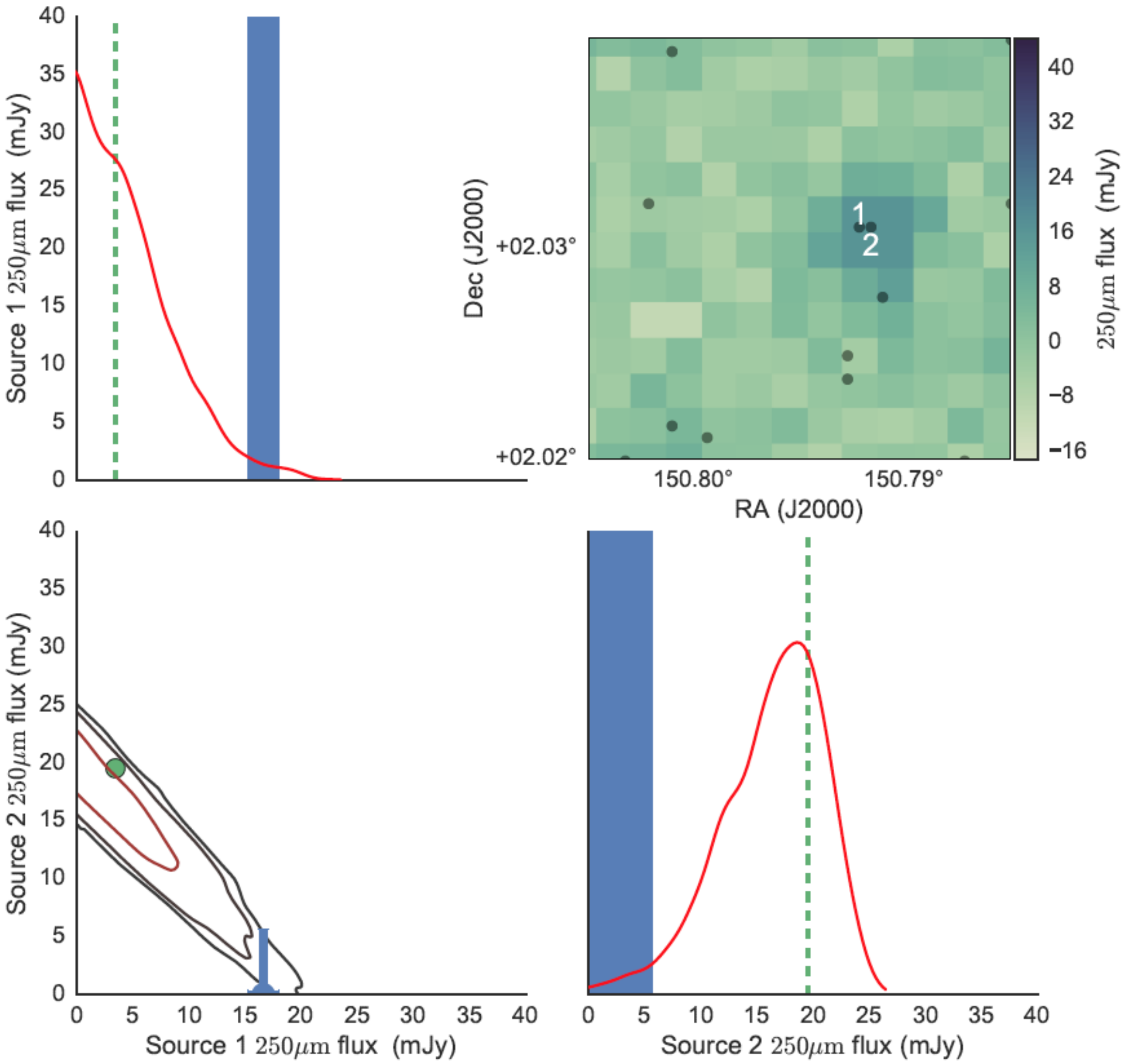


Planck team

BAYESIAN METHODOLOGIES

Bayesian Hierarchical Probabilistic Modelling



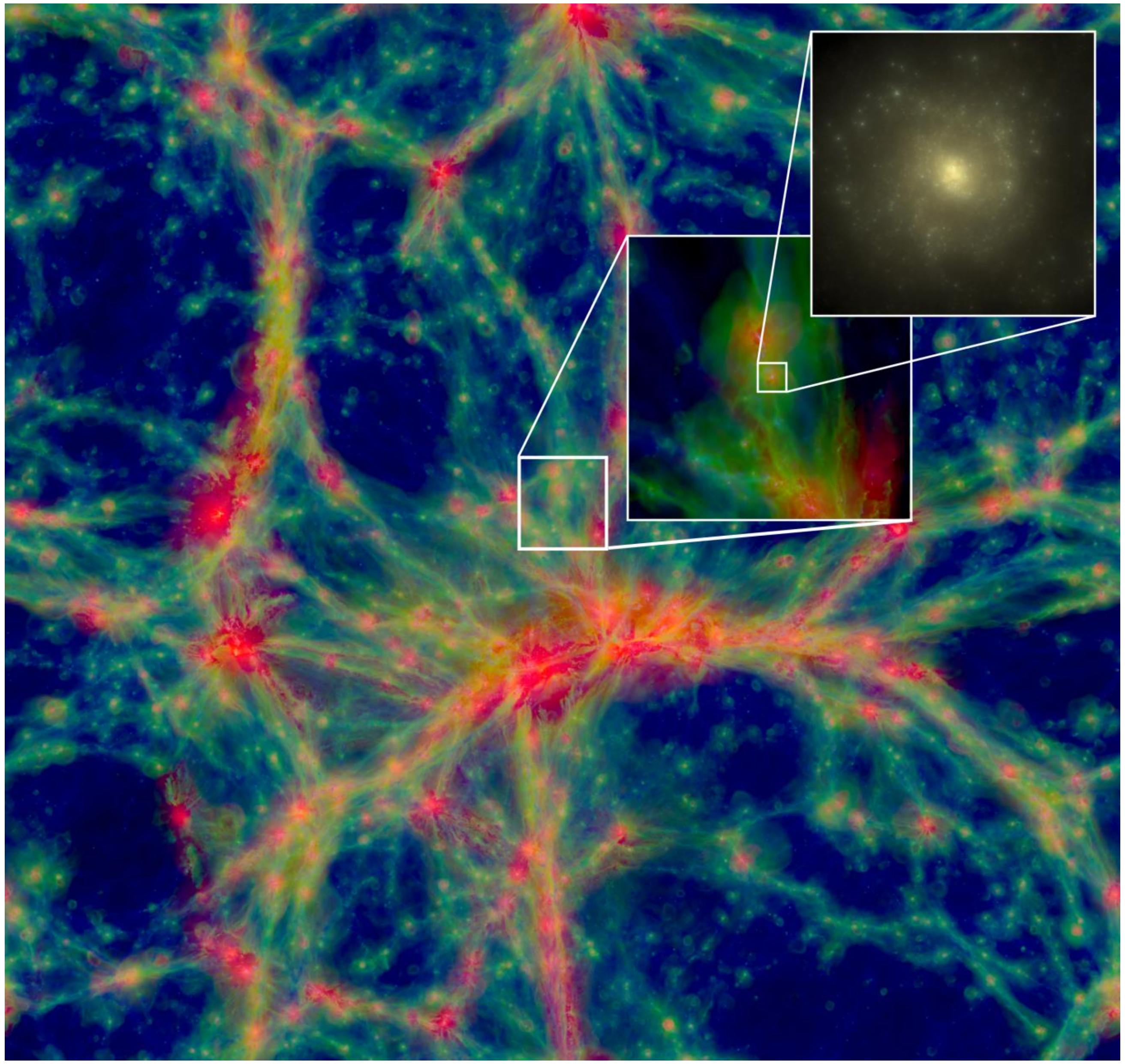


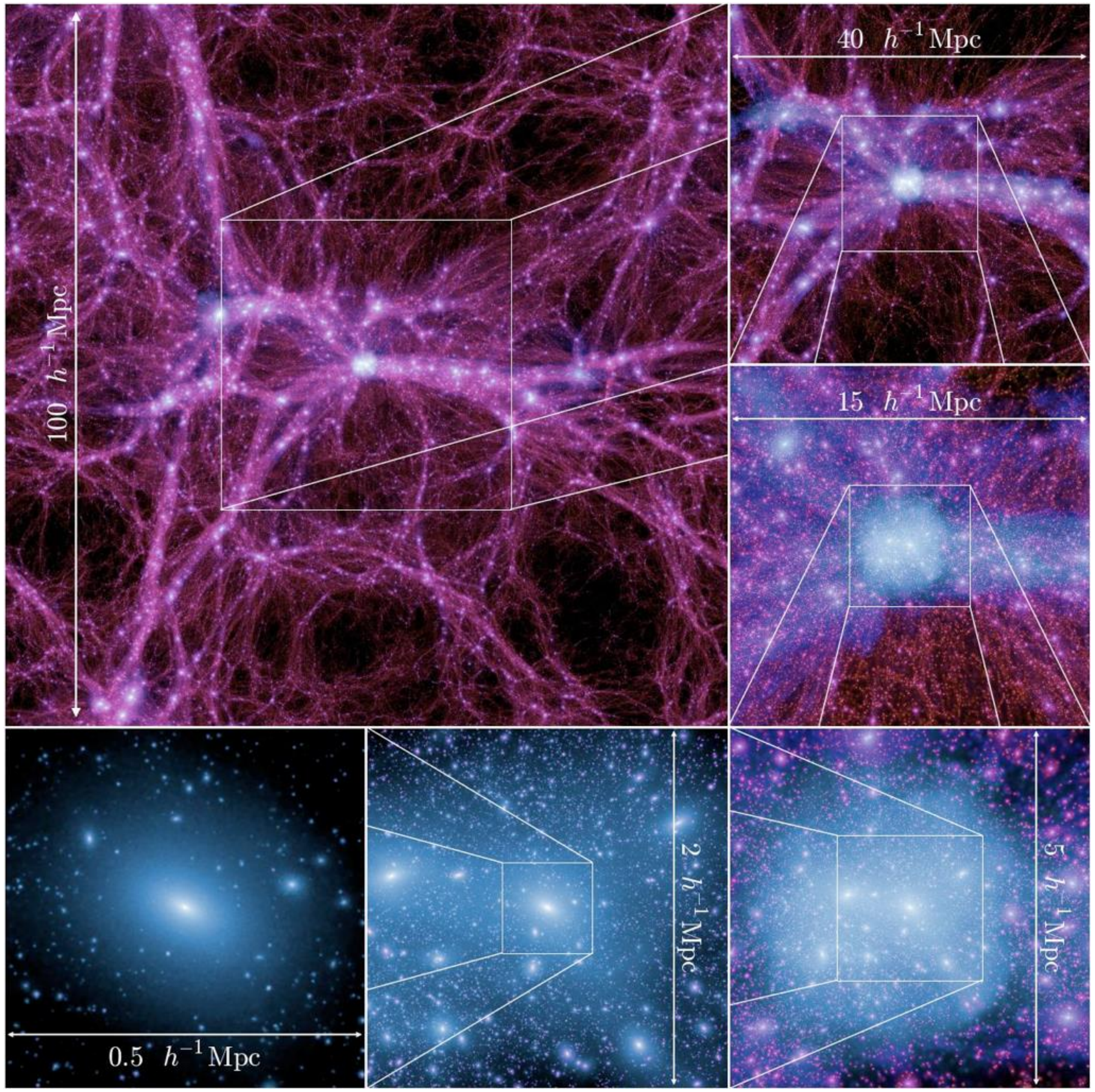
XID+, Hurley et al. 2016

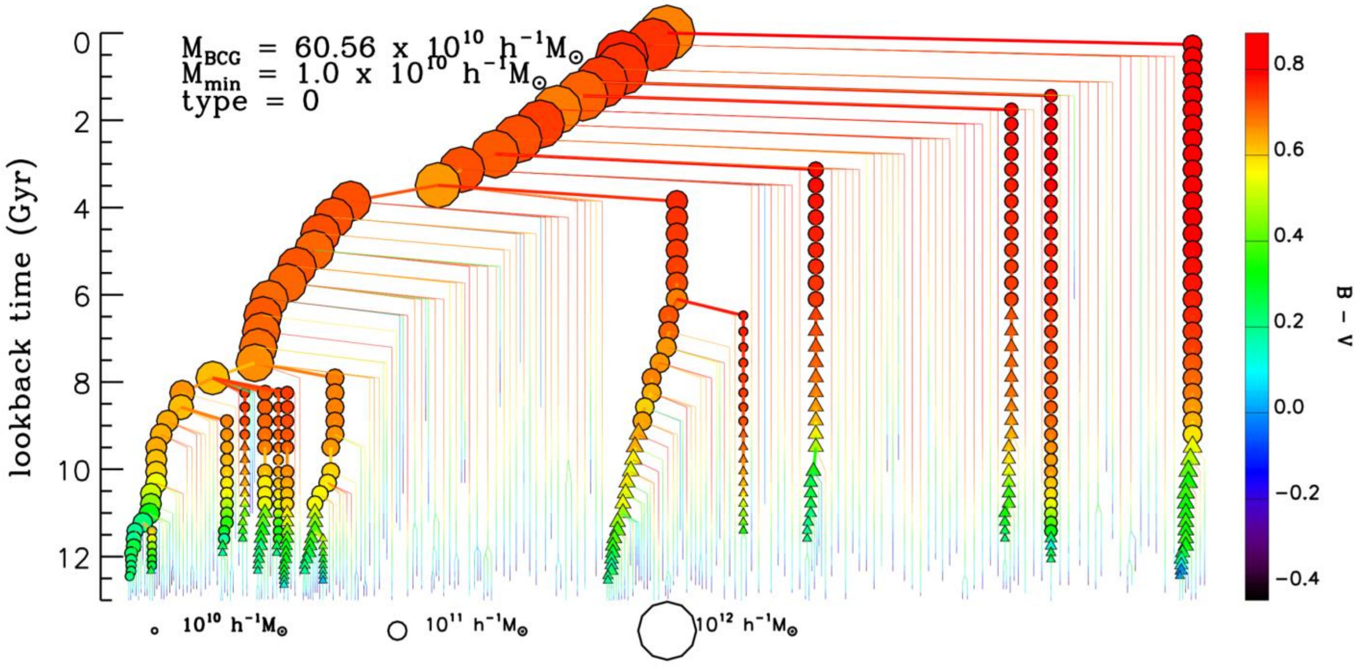
NUMERICAL SIMULATION ON MASSIVE SCALES

Making galaxies:

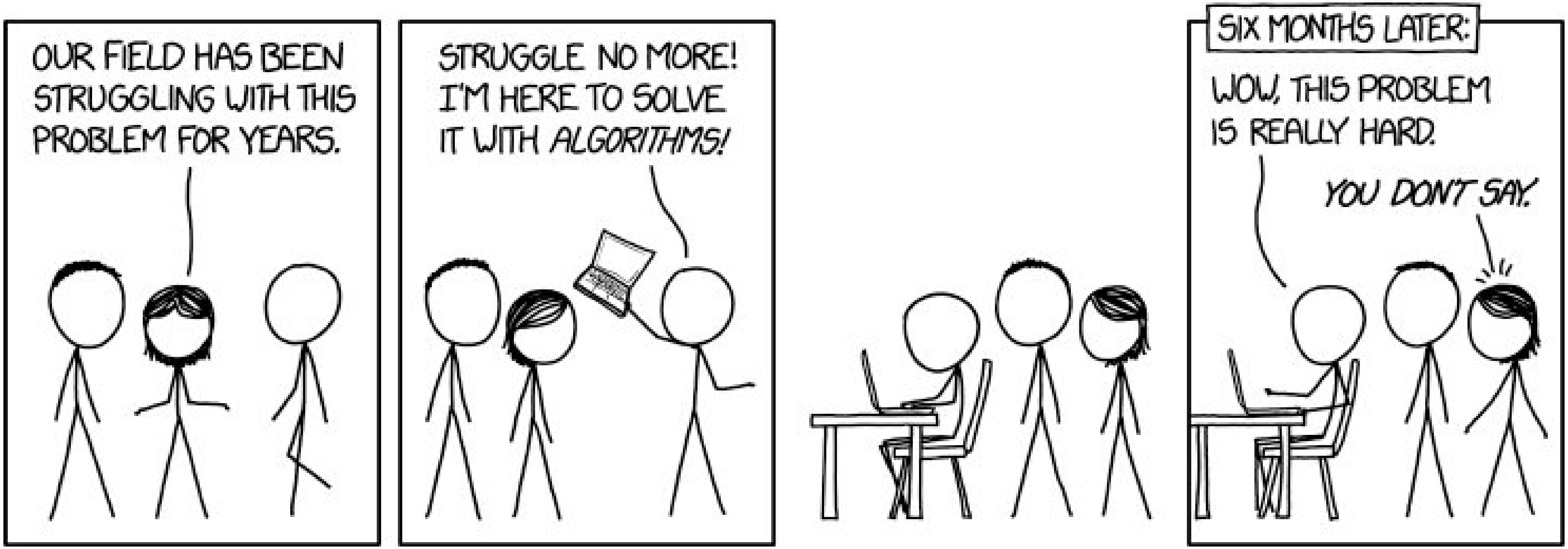
- semi-analytics
- simulation





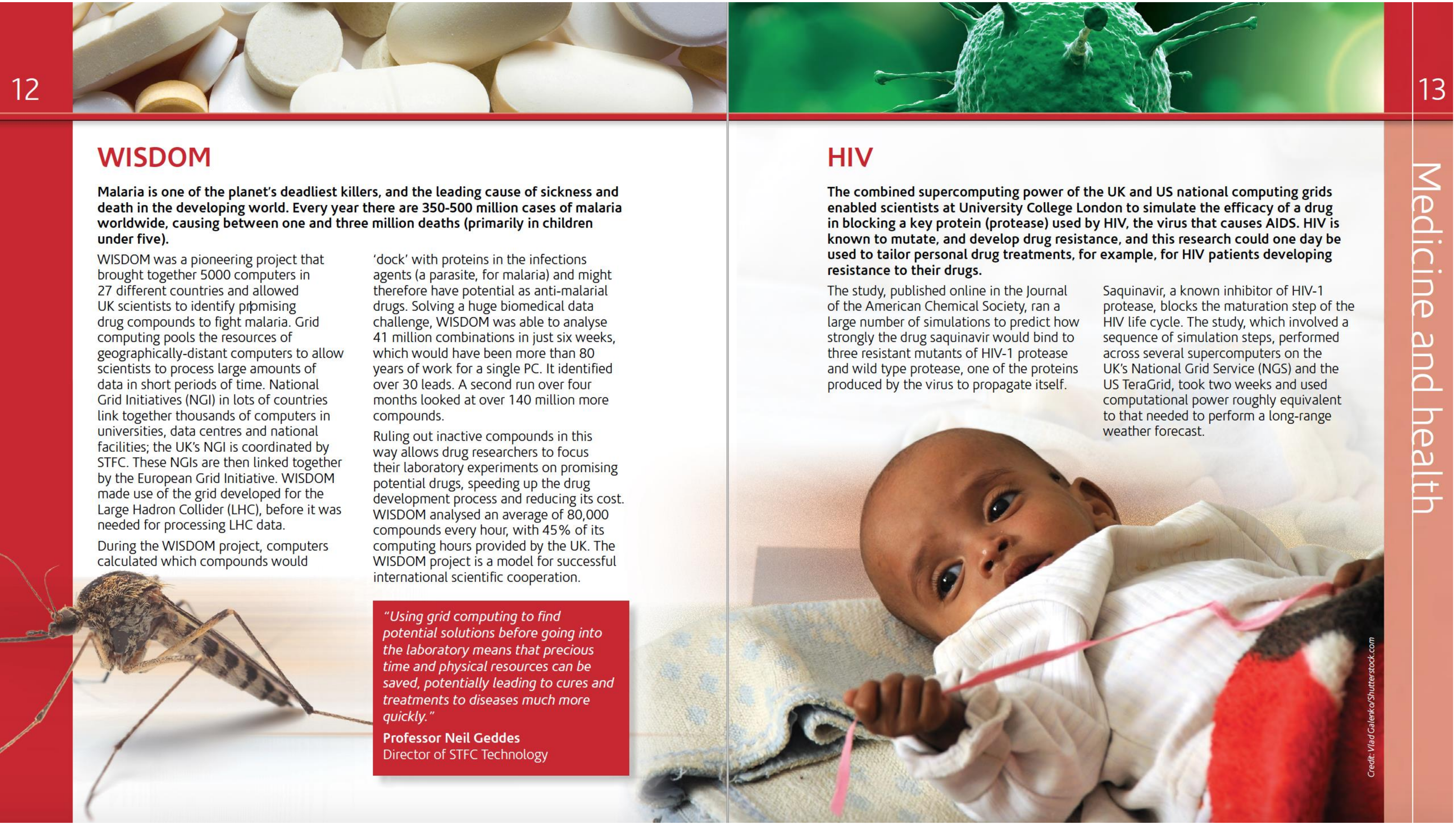


Application of data analysis


OUR FIELD HAS BEEN STRUGGLING WITH THIS PROBLEM FOR YEARS.
STRUGGLE NO MORE! I'M HERE TO SOLVE IT WITH ALGORITHMS!
SIX MONTHS LATER:
WOW, THIS PROBLEM IS REALLY HARD.
YOU DON'T SAY.

STFC Big Data Big Impact

http://www.stfc.ac.uk/files/impact-publications/big-data-big-impact/



WISDOM

Malaria is one of the planet's deadliest killers, and the leading cause of sickness and death in the developing world. Every year there are 350-500 million cases of malaria worldwide, causing between one and three million deaths (primarily in children under five).

WISDOM was a pioneering project that brought together 5000 computers in 27 different countries and allowed UK scientists to identify promising drug compounds to fight malaria. Grid computing pools the resources of geographically-distant computers to allow scientists to process large amounts of data in short periods of time. National Grid Initiatives (NGI) in lots of countries link together thousands of computers in universities, data centres and national facilities; the UK's NGI is coordinated by STFC. These NGIs are then linked together by the European Grid Initiative. WISDOM made use of the grid developed for the Large Hadron Collider (LHC), before it was needed for processing LHC data.

During the WISDOM project, computers calculated which compounds would 'dock' with proteins in the infections agents (a parasite, for malaria) and might therefore have potential as anti-malarial drugs. Solving a huge biomedical data challenge, WISDOM was able to analyse 41 million combinations in just six weeks, which would have been more than 80 years of work for a single PC. It identified over 30 leads. A second run over four months looked at over 140 million more compounds.

Ruling out inactive compounds in this way allows drug researchers to focus their laboratory experiments on promising potential drugs, speeding up the drug development process and reducing its cost. WISDOM analysed an average of 80,000 compounds every hour, with 45% of its computing hours provided by the UK. The WISDOM project is a model for successful international scientific cooperation.

> *"Using grid computing to find potential solutions before going into the laboratory means that precious time and physical resources can be saved, potentially leading to cures and treatments to diseases much more quickly."*
>
> **Professor Neil Geddes**
> Director of STFC Technology

HIV

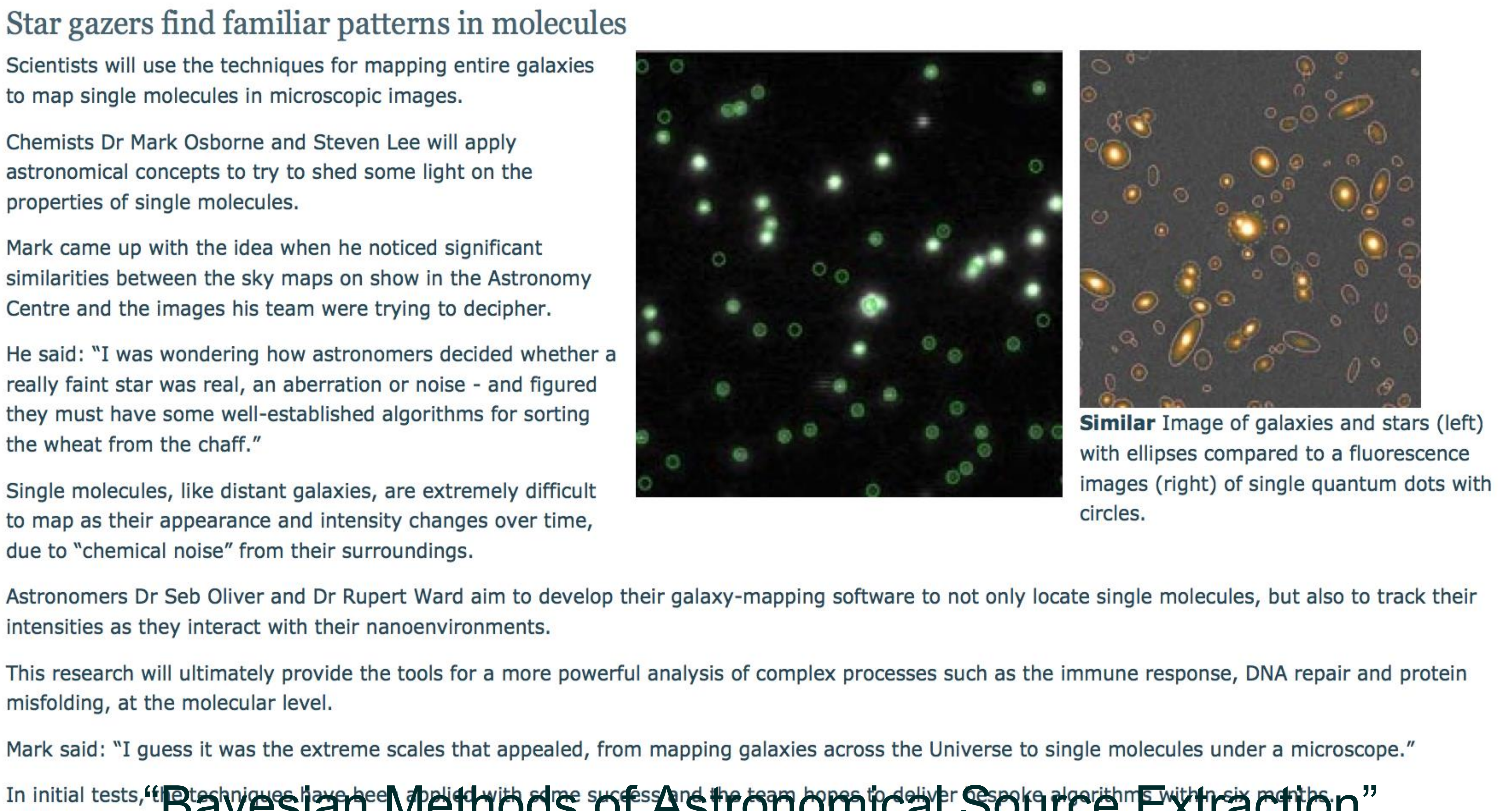
The combined supercomputing power of the UK and US national computing grids enabled scientists at University College London to simulate the efficacy of a drug in blocking a key protein (protease) used by HIV, the virus that causes AIDS. HIV is known to mutate, and develop drug resistance, and this research could one day be used to tailor personal drug treatments, for example, for HIV patients developing resistance to their drugs.

The study, published online in the Journal of the American Chemical Society, ran a large number of simulations to predict how strongly the drug saquinavir would bind to three resistant mutants of HIV-1 protease and wild type protease, one of the proteins produced by the virus to propagate itself.

Saquinavir, a known inhibitor of HIV-1 protease, blocks the maturation step of the HIV life cycle. The study, which involved a sequence of simulation steps, performed across several supercomputers on the UK's National Grid Service (NGS) and the US TeraGrid, took two weeks and used computational power roughly equivalent to that needed to perform a long-range weather forecast.

Credit: Vlad Galenko/Shutterstock.com

Quantum Dots



Star gazers find familiar patterns in molecules

Scientists will use the techniques for mapping entire galaxies to map single molecules in microscopic images.

Chemists Dr Mark Osborne and Steven Lee will apply astronomical concepts to try to shed some light on the properties of single molecules.

Mark came up with the idea when he noticed significant similarities between the sky maps on show in the Astronomy Centre and the images his team were trying to decipher.

He said: "I was wondering how astronomers decided whether a really faint star was real, an aberration or noise - and figured they must have some well-established algorithms for sorting the wheat from the chaff."

Single molecules, like distant galaxies, are extremely difficult to map as their appearance and intensity changes over time, due to "chemical noise" from their surroundings.

Similar Image of galaxies and stars (left) with ellipses compared to a fluorescence images (right) of single quantum dots with circles.

Astronomers Dr Seb Oliver and Dr Rupert Ward aim to develop their galaxy-mapping software to not only locate single molecules, but also to track their intensities as they interact with their nanoenvironments.

This research will ultimately provide the tools for a more powerful analysis of complex processes such as the immune response, DNA repair and protein misfolding, at the molecular level.

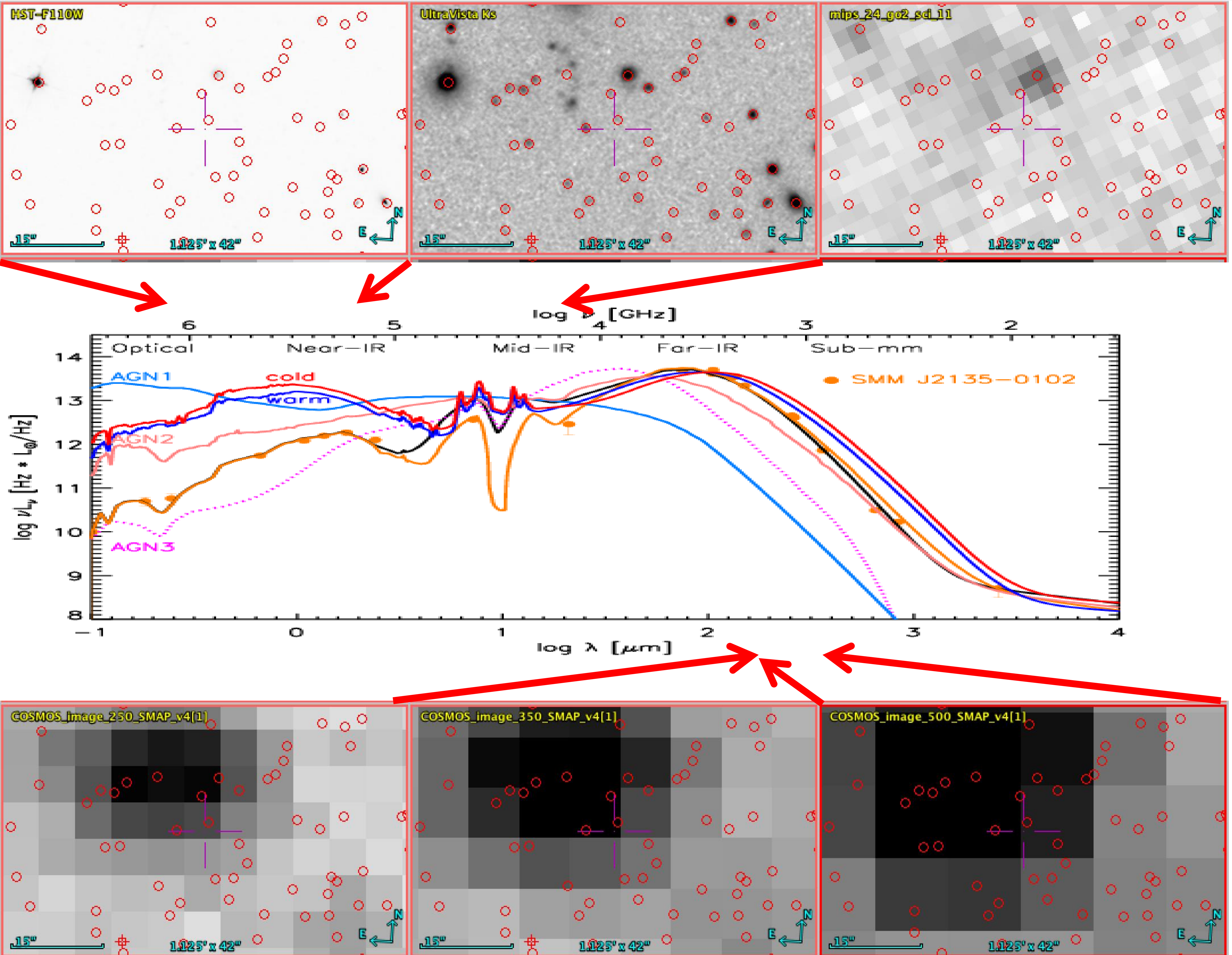
Mark said: "I guess it was the extreme scales that appealed, from mapping galaxies across the Universe to single molecules under a microscope."

In initial tests, the techniques have been applied with some success and the team hopes to deliver bespoke algorithms within six months.

"Bayesian Methods of Astronomical Source Extraction"

Savage & Oliver 2007ApJ...661.1339S

Cross identification in Extragalactic Astronomy:



Astronomy – ECFSPR
The analogy

Galaxies	→	Patient
Wavelength	→	Time
Redshift	→	Age
Position	→	e.g. Gender, Genotype, Age
Intensity	→	e.g. BMI

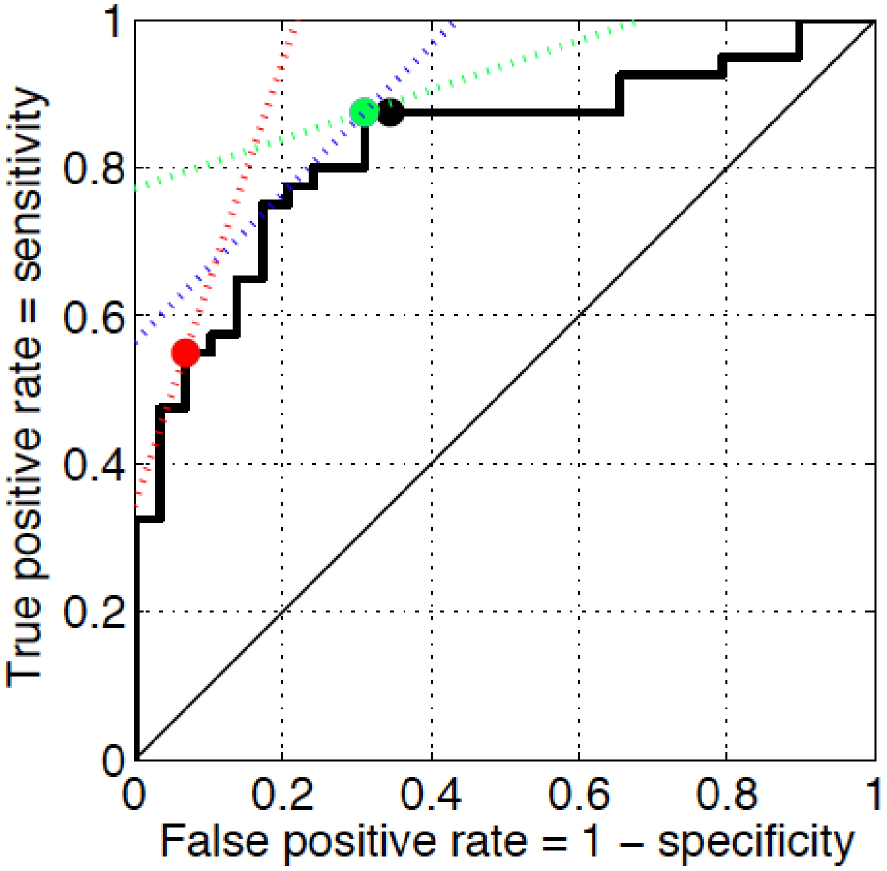
Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting state fMRI

University of Sussex

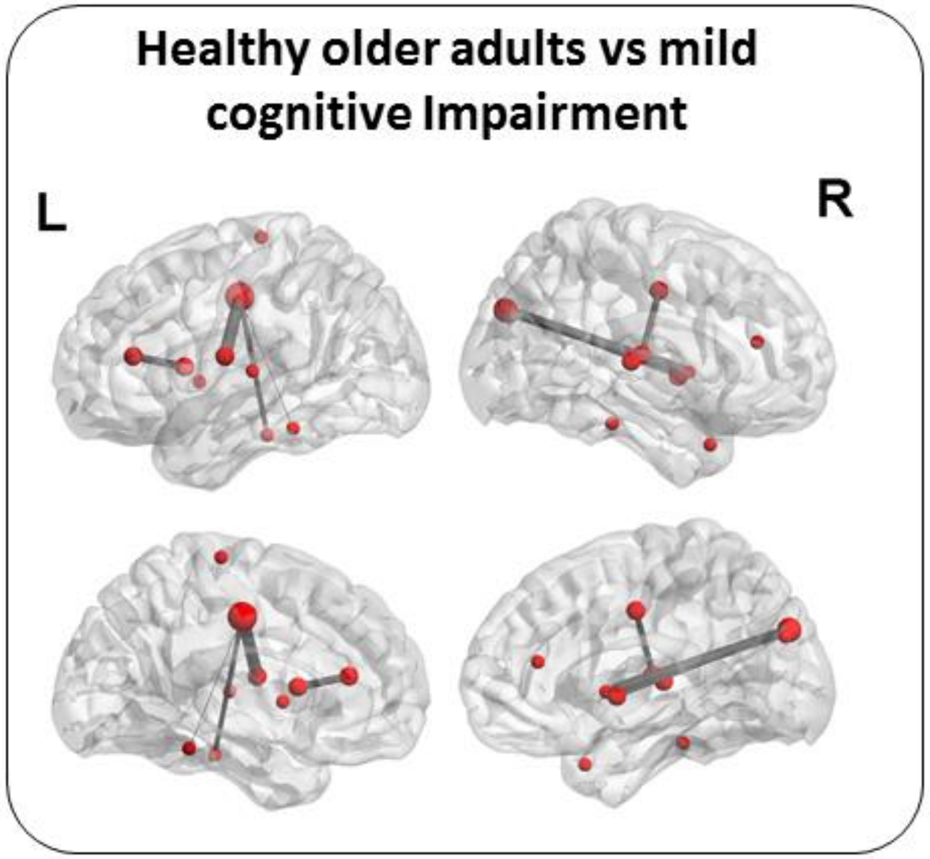
Challis E, Hurley P,
Serra L, Bozzali M,
Oliver S, Cercignani M
Neuroimaging, 2014

Tunable, e.g.

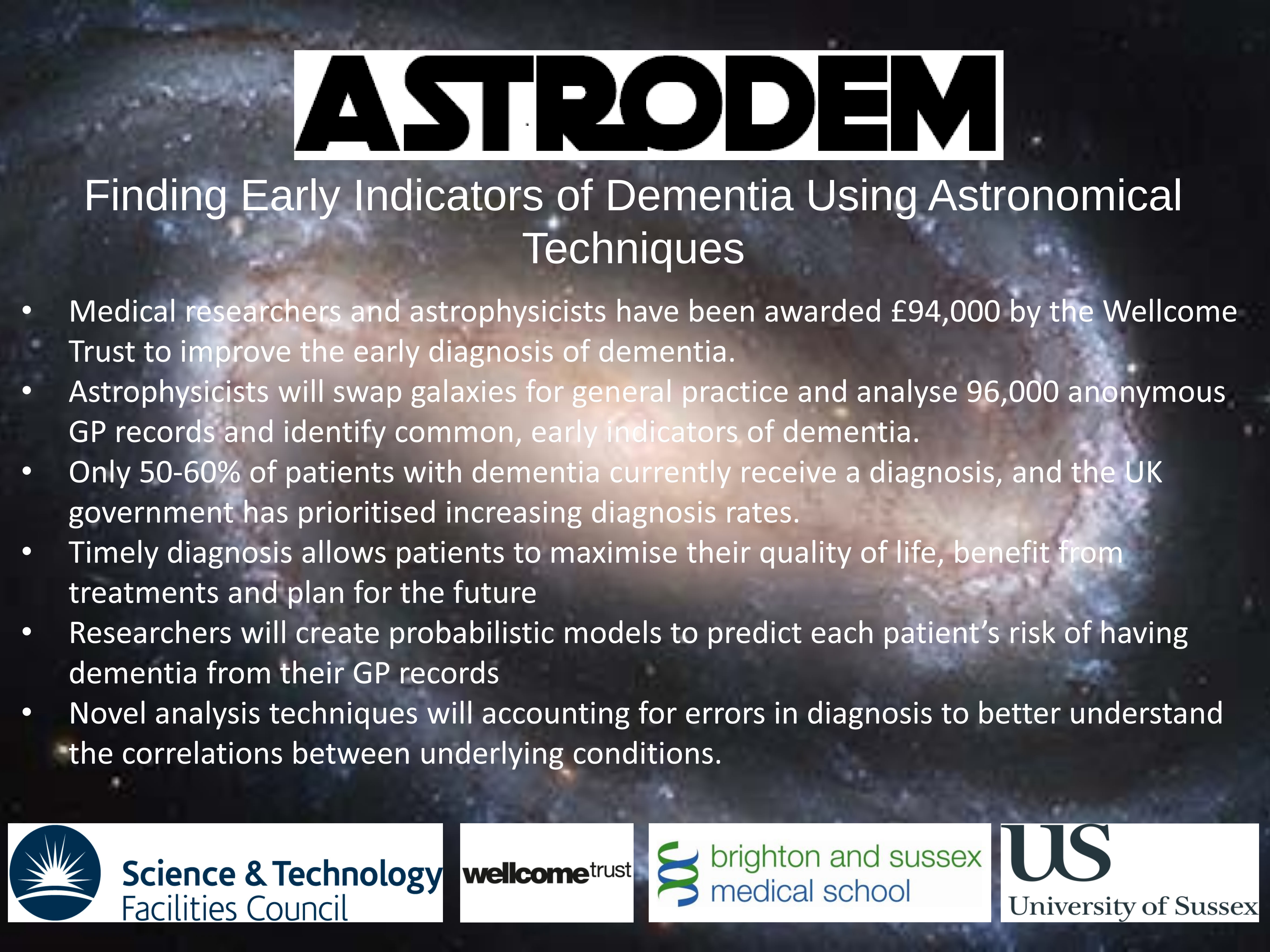
- 🟢 Sensitivity = 88%
- 🟢 Specificity = 69%
- 🔴 Sensitivity = 55%
- 🔴 Specificity = 93%



(a) NC vs a-MCI



Healthy older adults vs mild cognitive Impairment



ASTRODEM

Finding Early Indicators of Dementia Using Astronomical Techniques

- Medical researchers and astrophysicists have been awarded £94,000 by the Wellcome Trust to improve the early diagnosis of dementia.
- Astrophysicists will swap galaxies for general practice and analyse 96,000 anonymous GP records and identify common, early indicators of dementia.
- Only 50-60% of patients with dementia currently receive a diagnosis, and the UK government has prioritised increasing diagnosis rates.
- Timely diagnosis allows patients to maximise their quality of life, benefit from treatments and plan for the future
- Researchers will create probabilistic models to predict each patient’s risk of having dementia from their GP records
- Novel analysis techniques will accounting for errors in diagnosis to better understand the correlations between underlying conditions.

Science & Technology Facilities Council

wellcome trust

brighton and sussex medical school

US University of Sussex

How can we help?

- Scientific Computing Department, at STFC
- 8 new Centres for Doctoral Training involving 19 Universities

- 85 STFC funded 4 year PhD students
- Remote sensing analysis
- Infrastructure and data logistics
- Modelling and simulation of complex proceses