

# Euclid Big Data

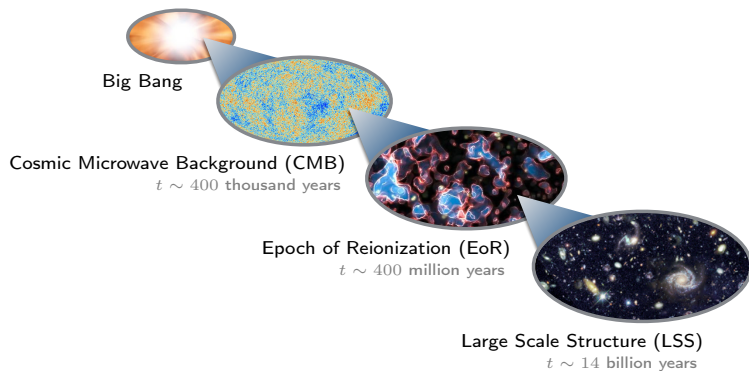
## Data Science for Science

Jason McEwen  
[www.jasonmcewen.org](http://www.jasonmcewen.org)  
[@jasonmcewen](https://twitter.com/jasonmcewen)

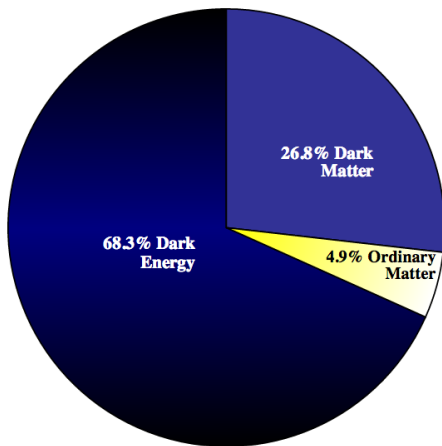
*Mullard Space Science Laboratory (MSSL)*  
*University College London (UCL)*

Big Data – A Space Perspective  
University College London (UCL), April 2018

# Cosmic evolution of our Universe

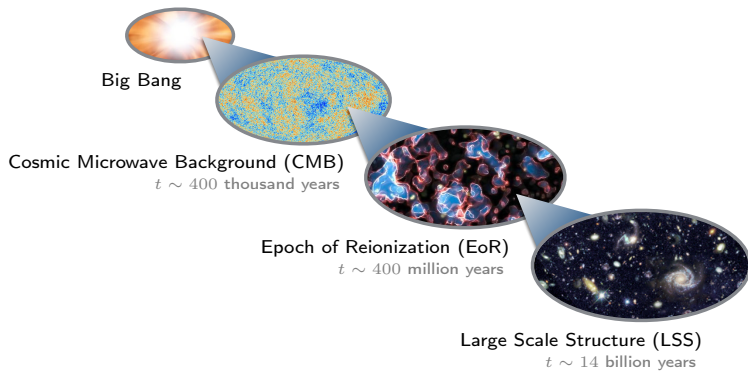


## Content of the Universe



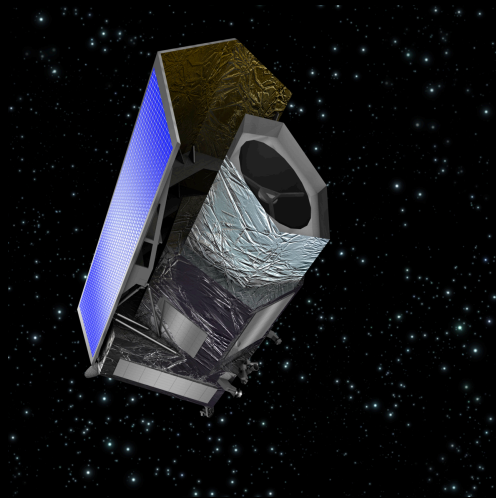
Credit: Planck

# Unanswered fundamental questions



What is the nature of **dark energy** and **dark matter**?

# Euclid satellite



Credit: Euclid



# What is Euclid?

- Euclid is the next space-based cosmology experiment.
- ESA Medium-Class Mission due for launch in 2021.
- Largest astronomical consortium: 15 countries, ~2000 scientists, ~200 institutes.
- Science objective: to understand the origins of the Universe's accelerated expansion.
- Controlling systematics to unprecedented level of accuracy (space mission critical).
- UK leads science, data processing and engineering aspects.

# What is Euclid?

- Euclid is the next space-based cosmology experiment.
- ESA Medium-Class Mission due for launch in 2021.
- Largest astronomical consortium: 15 countries, ~2000 scientists, ~200 institutes.
- Science objective: **to understand the origins of the Universe's accelerated expansion.**
- Controlling systematics to unprecedented level of accuracy (space mission critical).
- UK leads science, data processing and engineering aspects.

# What is Euclid?

- Euclid is the next space-based cosmology experiment.
- ESA Medium-Class Mission due for launch in 2021.
- Largest astronomical consortium: 15 countries, ~2000 scientists, ~200 institutes.
- Science objective: **to understand the origins of the Universe's accelerated expansion.**
- Controlling systematics to unprecedented level of accuracy (space mission critical).
- UK leads science, data processing and engineering aspects.



# What is Euclid?

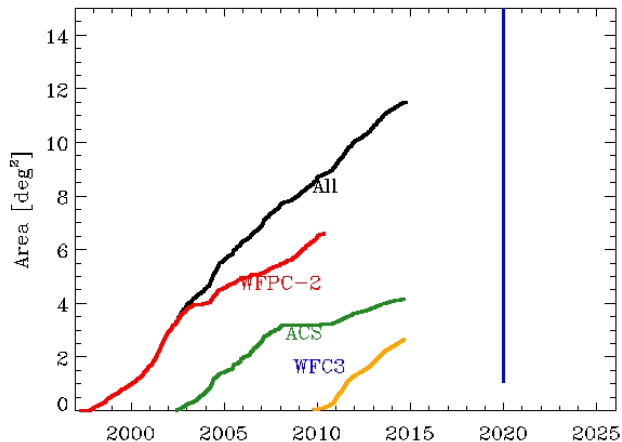
- Euclid is the next space-based cosmology experiment.
- ESA Medium-Class Mission due for launch in 2021.
- Largest astronomical consortium: 15 countries, ~2000 scientists, ~200 institutes.
- Science objective: **to understand the origins of the Universe's accelerated expansion.**
- Controlling systematics to unprecedented level of accuracy (space mission critical).
- UK leads science, data processing and engineering aspects.

# What is Euclid?

- Will image  $\sim 1$  billion galaxies.
- Observe to redshift  $z \sim 2$ , *i.e.* looking back  $\sim 10$  billion years.
- Highest ever download rate from space: 850 Gb/day.
- Observations per mission: 10 PB.
- Big Sims also required:  $10^4$ – $10^6$  N-body simulations.

# Euclid sky coverage

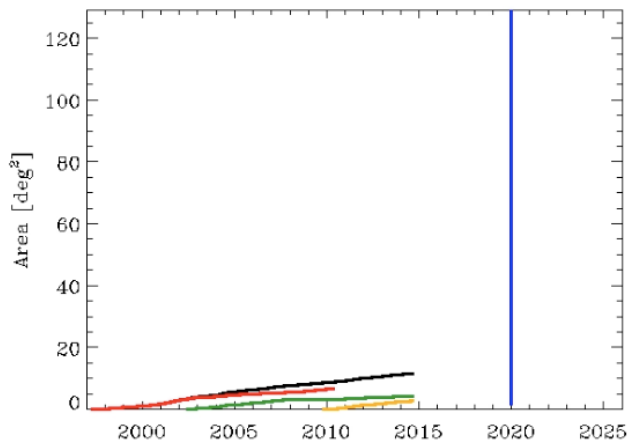
Switch on



Credit: Tom Kitching

# Euclid sky coverage

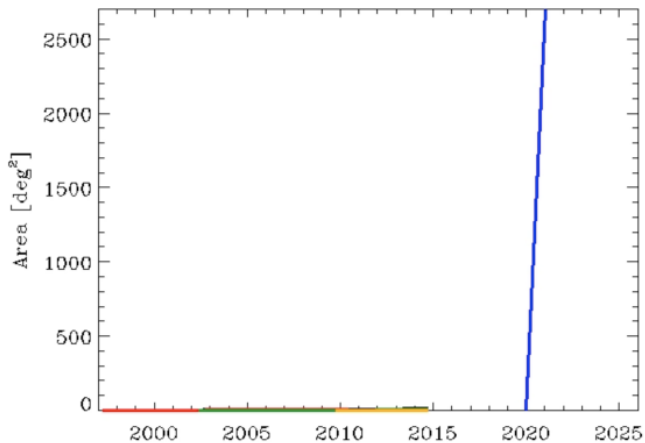
2 weeks



Credit: Tom Kitching

# Euclid sky coverage

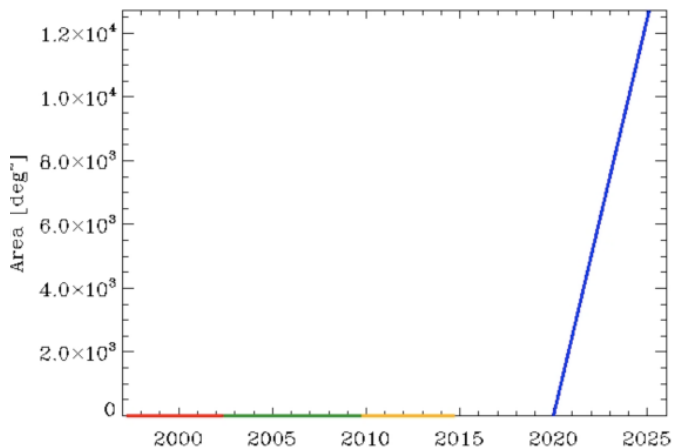
1 year



Credit: Tom Kitching

# Euclid sky coverage

5 years



Credit: Tom Kitching

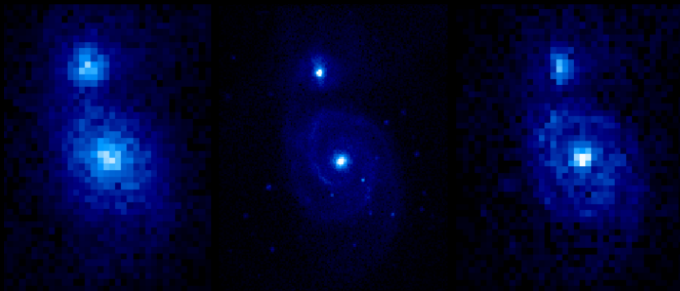
# Euclid: optimised for shape measurements

## Euclid: optimised for shape measurements

Euclid  
consortium

Courtesy J. Brinchmann,  
Steve Warren

M51



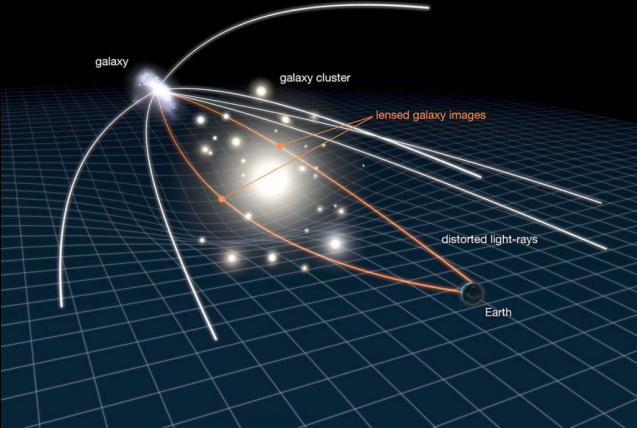
SDSS @  $z=0.1$

Euclid @  $z=0.1$

Euclid @  $z=0.7$

- Euclid images of  $z \sim 1$  galaxies: same resolution as SDSS images at  $z \sim 0.05$  and at least 3 magnitudes deeper.
- Space imaging of Euclid will outperform any other surveys of weak lensing.

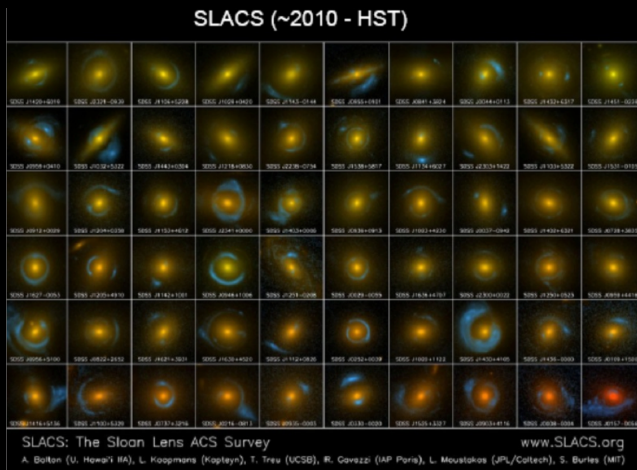
# Gravitational lensing



Credit: CFHTLenS



# Strong gravitational lenses



Credit: Koopmans

### Will become an industry

Substructure study; high-z normal galaxies...

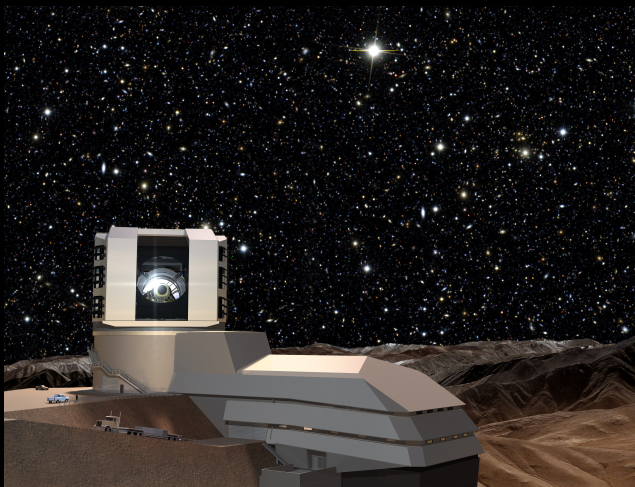
SLACS

Euclid Legacy : after 2 months  
(66 months planned)

Credit: Koopmans

# Synergies with ground-based astronomical big-data experiments

## Large Synoptic Survey Telescope (LSST)



Credit: LSST



# Synergies with ground-based astronomical big-data experiments

## Large Synoptic Survey Telescope (LSST)

### Data Releases:

Number of Data Releases = 11

Date of DR1 release = Date of Operations Start+ 12  
months

Estimated numbers for DR-1 release

Objects = 18 billion

Sources = 350 billion (single epoch)

Forced Sources = 0.75 trillion

Estimated numbers for DR-11

Objects = 37 billion

Sources = 7 trillion (single epoch)

Forced Sources = 30 trillion

Visits observed = 2.75 million

Images collected = 5.5 million

### Alert Production:

Real-time alert latency = 60 seconds

Average number of alerts per night= "about 10 million"

### Data and compute sizes:

Final image collection (DR11) = 0.5 Exabytes

Final database size (DR11) = 15 PB

Final disk storage = 0.4 Exabytes

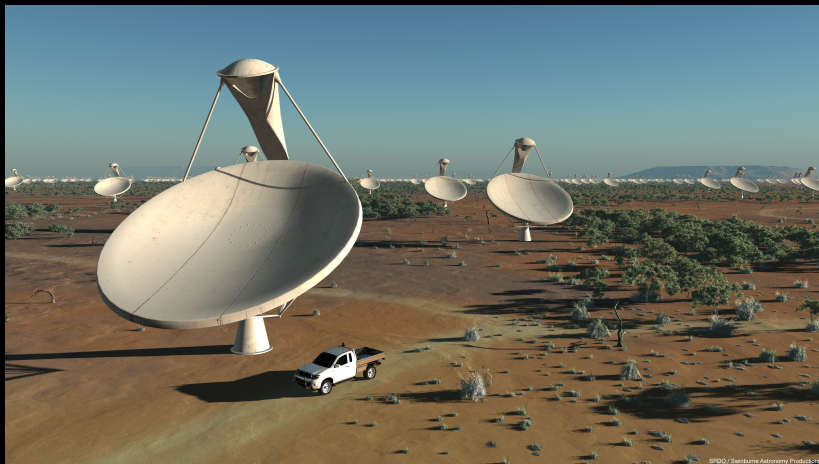
Peak number of nodes = 1750 nodes

Peak compute power in LSST data centers = 1.8 PFLOPS

Credit: LSST

# Synergies with ground-based astronomical big-data experiments

## Square Kilometre Array (SKA)




Credit: SKA



# Synergies with ground-based astronomical big-data experiments

## The SKA poses a considerable big-data challenge

The SKA will use enough optical fiber to wrap twice around the Earth!



2x

SKA

The SKA will be so sensitive that it will be able to detect an airport radar on a planet tens of light years away.



Tens of light years

SKA

The SKA will generate enough raw data to fill 15 million 64GB iPods every day!



64GB  
x 15 MILLION

DATA

SKA

The dishes of the SKA will produce 10 times the global internet traffic.



10x

SKA

The aperture arrays in the SKA could produce more than 100 times the global internet traffic.



100x

SKA

The SKA central computer will have the processing power of about one hundred million PCs.



SKA  
super computer

x 100,000,000  
Personal Computers

SKA

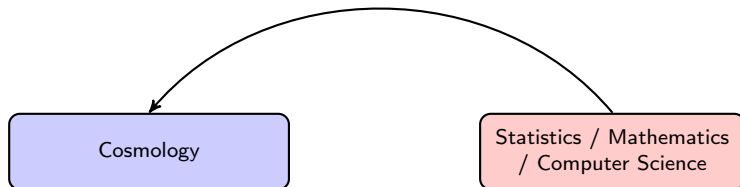
Credit: SKA



# Astrostatistics & Astroinformatics

## Closing the loop

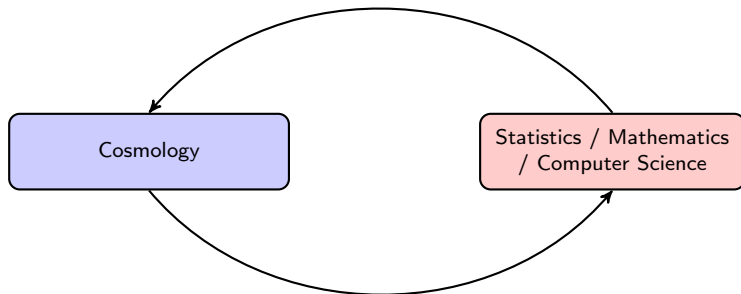
*Extracting weak observational signatures of fundamental physics from complex data-sets requires sensitive, robust and principled analysis techniques.*



# Astrostatistics & Astrominformatics

## Closing the loop

*Extracting weak observational signatures of fundamental physics from complex data-sets requires sensitive, robust and principled analysis techniques.*



*Constructing appropriate analysis techniques requires a deep understanding of cosmological problems and methodological foundations.*



# UCL Centre for Doctoral Training (CDT) in Data Intensive Science (DIS)

- UCL won bid to host **STFC's first CDT**.  
<https://www.hep.ucl.ac.uk/cdt-dis/>
- Focused on **Data Intensive Science (DIS)**.
- Aims:
  - Train next generation of leaders in the field of DIS (in both academic and industry).
  - Promote development and application of novel DIS techniques.
  - Promote **knowledge transfer**:
    - between academic fields;
    - between non-academic and academic organisations.
- Unique opportunity to **bring together DIS research** from perspective of **applications, methodologies, and theoretical foundations**.



Science & Technology  
Facilities Council

# UCL Centre for Doctoral Training (CDT) in Data Intensive Science (DIS)

- UCL won bid to host **STFC's first CDT**.  
<https://www.hep.ucl.ac.uk/cdt-dis/>
- Focused on **Data Intensive Science (DIS)**.
- Aims:
  - **Train next generation of leaders** in the field of DIS (in both academic and industry).
  - Promote development and application of **novel DIS techniques**.
  - Promote **knowledge transfer**:
    - between academic fields;
    - between non-academic and academic organisations.
- Unique opportunity to **bring together DIS research** from perspective of **applications, methodologies, and theoretical foundations**.



Science & Technology  
Facilities Council

# UCL Centre for Doctoral Training (CDT) in Data Intensive Science (DIS)

- UCL won bid to host **STFC's first CDT**.  
<https://www.hep.ucl.ac.uk/cdt-dis/>
- Focused on **Data Intensive Science (DIS)**.
- Aims:
  - **Train next generation of leaders** in the field of DIS (in both academic and industry).
  - Promote development and application of **novel DIS techniques**.
  - Promote **knowledge transfer**:
    - between academic fields;
    - between non-academic and academic organisations.
- Unique opportunity to **bring together DIS research** from perspective of **applications**, **methodologies**, and **theoretical foundations**.

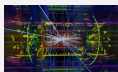


Science & Technology  
Facilities Council

# UCL Centre for Doctoral Training (CDT) in Data Intensive Science (DIS)

Who we are

**Particle Physics**  
Dpt. of Physics and  
Astronomy  
(20 CDT Staff Members)



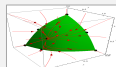
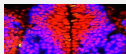
**Astrophysics**  
Dpt. of Physics and  
Astronomy  
(20 CDT Staff Members)

Department of  
**Space and Climate  
Science**  
(20 CDT Staff Members)



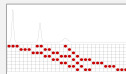
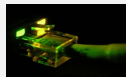
**Atomic & Molecular  
Physics**  
Dpt. of Physics and Astronomy  
(2 CDT Staff Members)

Department of  
**Computer Science**  
(8 CDT Staff Members)



Department of  
**Mathematics**  
(9 CDT Staff Members)

Department of  
**Electrical Engineering**  
(3 CDT Staff Members)



Department of  
**Statistical Science**  
(5 CDT Staff Members)

Aim to **foster closer collaboration** between these areas to aid the development of novel DIS techniques or applications to new areas.

# UCL Centre for Doctoral Training (CDT) in Data Intensive Science (DIS)

## Industrial partners



- Students will undertake **6 month internships** with partners on a DIS project
- Promote **knowledge transfer** between academic and non-academic organisations.
- More organisations joining (UKAEA, Asos, GroupM, S&P, Illuminas, ASI, ...).

## Summary

- Euclid will usher in a **new paradigm** for galactic surveys, in order to address fundamental question about the **nature of dark energy**.
- Paradigm shift in size, complexity and structure of data.
- Existing methods simply not feasible.
- Paradigm shift in analyses required.
- **Multi-disciplinary approach** will be critical, drawing on expertise from different disciplines.