Some notes about heterogeneous data sets

Yuanyuan Zhang @NOIRLab

"don't turn off your brain"

A rather biased and personal view

- Heterogeneous data sets: data sets that are different in content or in wavelength coverage. (Brunner+, Massive data sets in Astronomy)

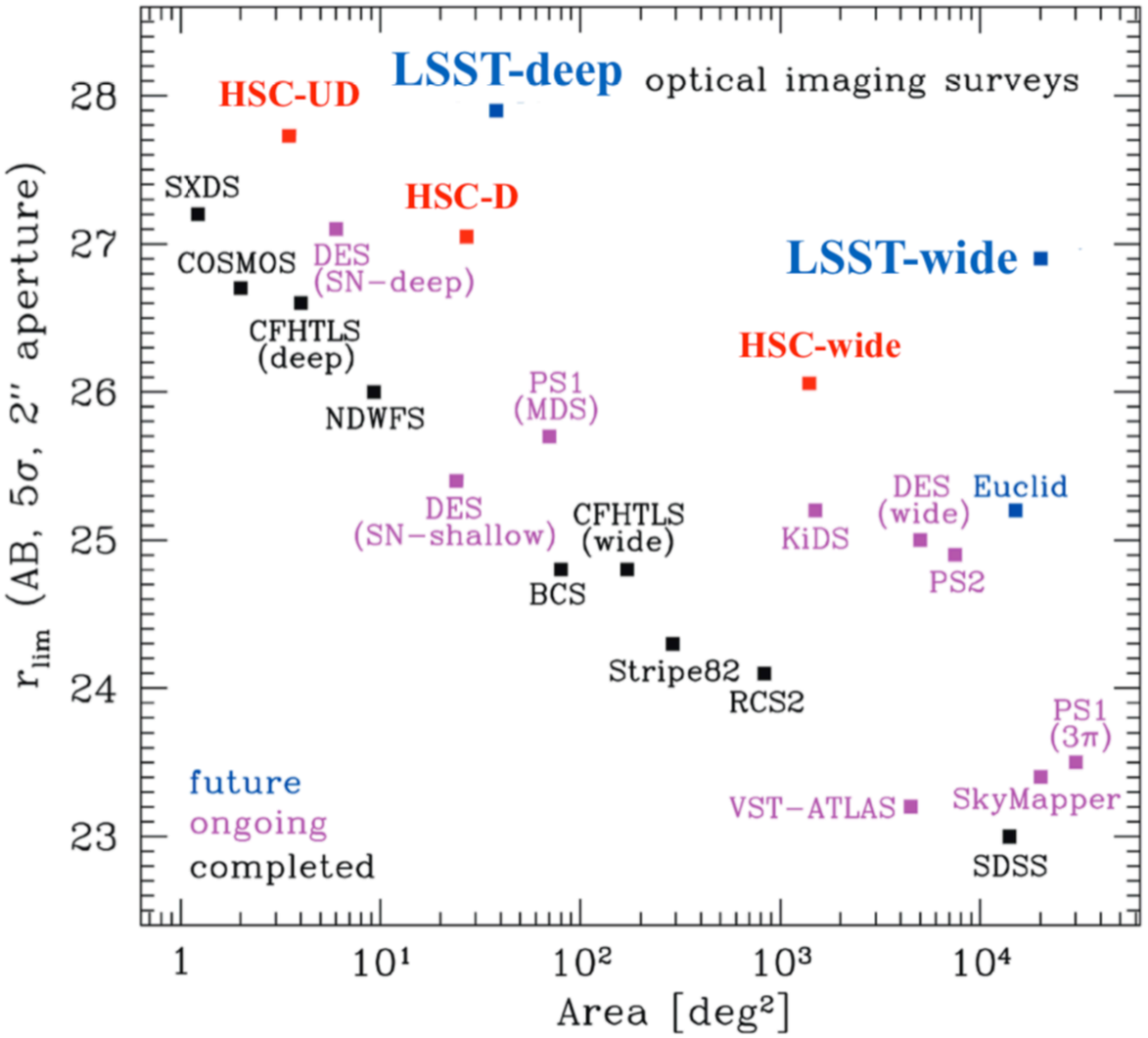
- **About me:**
  Observations, cosmology and astrophysics

    - My experience came from working with the so-called “extremely” wide-field optical surveys, and mostly working with images.

    - For example, SDSS, DES and now dabbing into LSST (sims), and multi-wavelength data sets complementary to those.

What is not a heterogenous data set?

- **Example:** “extremely” wide-field surveys
  - Optical: SDSS, HSC, DES, KiDS, DES
  - Collet a large swath of data with one telescope.
  - A pre-designed “observing strategy” that will lead to relatively uniform data.
  - Can cover thousands of degree squares of the sky.



Credit: Satoshi Miyazaki

What is not a heterogenous data set?

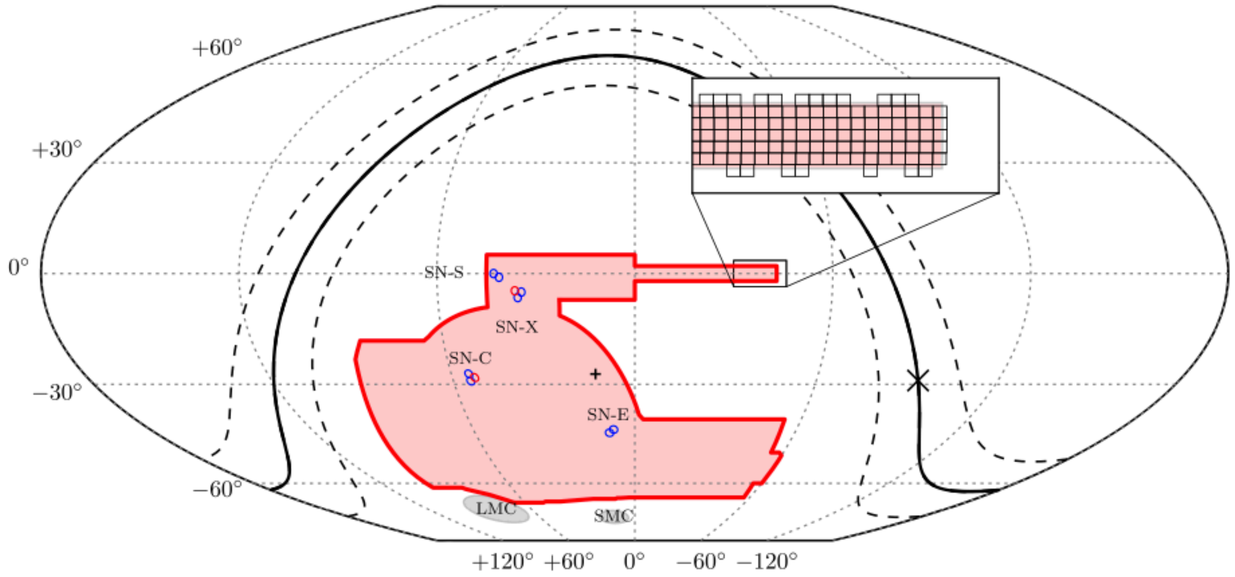
- Example: The Dark Energy Survey

- Covers a very large area 5000 deg^2, and dedicated to a variety of scientific goals:

    - Cosmology: supernovae, weak lensing, BAO, galaxy clusters

    - Galaxy physics: Milky Way, galaxy clusters, galaxy stellar mass

    - Even solar system objects: Kuiper belt objects

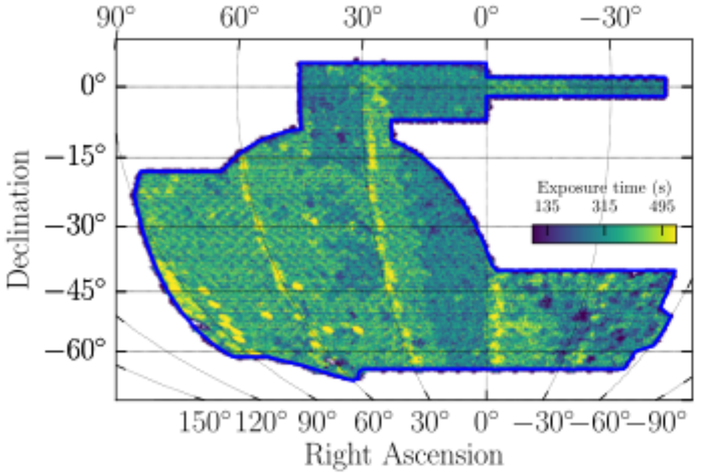


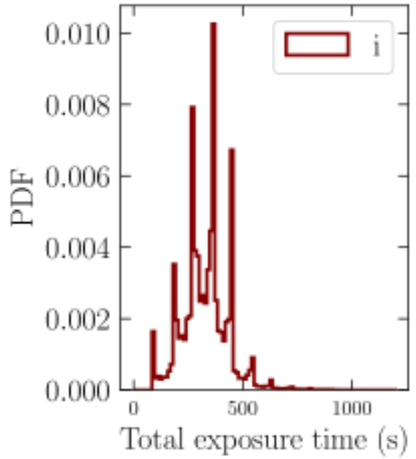
Abbott et al.
https://arxiv.org/pdf/2101.05765.pdf

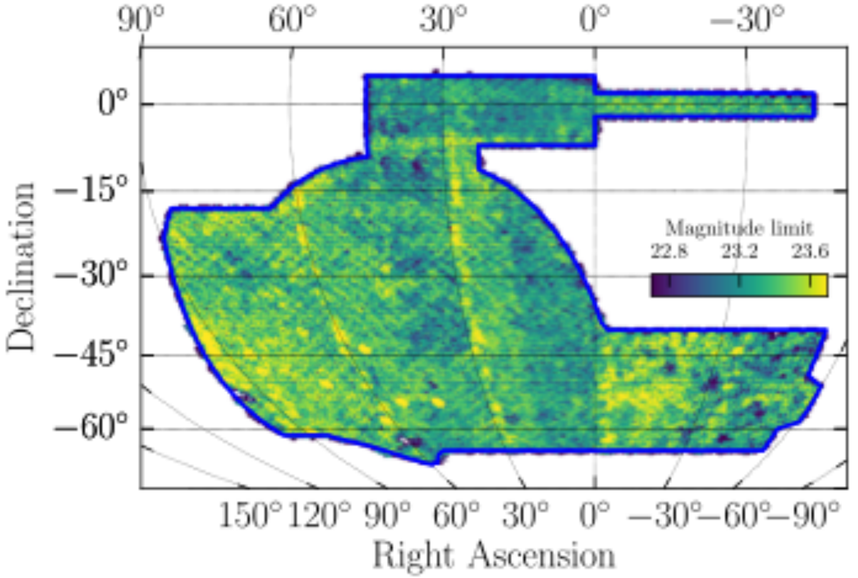
What is not a heterogenous data set?

Is DES data uniform?

- An imaging survey in g, r, i, z, Y taken over 6 years from 2013 to 2019.
- Goal was to acquire 10 exposures per band.
- Dedicated pipeline to identify galaxies, stars from the images, and measure their photometry.







Sevilla-Noarbe et al. DES Y3 gold data sets https://arxiv.org/abs/2011.03407

Data almost always need to be “cleaned”.

- Data processing can fail, yielding objects with unphysical properties.
- There are also regions in the sky that are close to bright stars, globular clusters, bright galaxies that interfere with data uniformity.

Table 3. Y3 GOLD FLAGS_GOLD bit flag variable

Flag Bit	Number of objects affected	Description
1	14185334	MOF_FLAGS != 0 or MOF_FLAGS = NULL, flag raised by MOF processing
2	6555347	SOF_FLAGS! = 0, flag raised by SOF processing
4	1532648	SOF_FLAGS == 1 or SOF_FLAGS > 2, flags for PSF fit failures
8	746568	Any SExtractor FLAGS_[GRIZ] > 3
16	3091171	Any of IMAFLAGS_ISO_[GRIZ]! = 0. †
32	152999	Bright blue artifacts in the images
64	62653	Bright objects with unphysical colors, possible transients

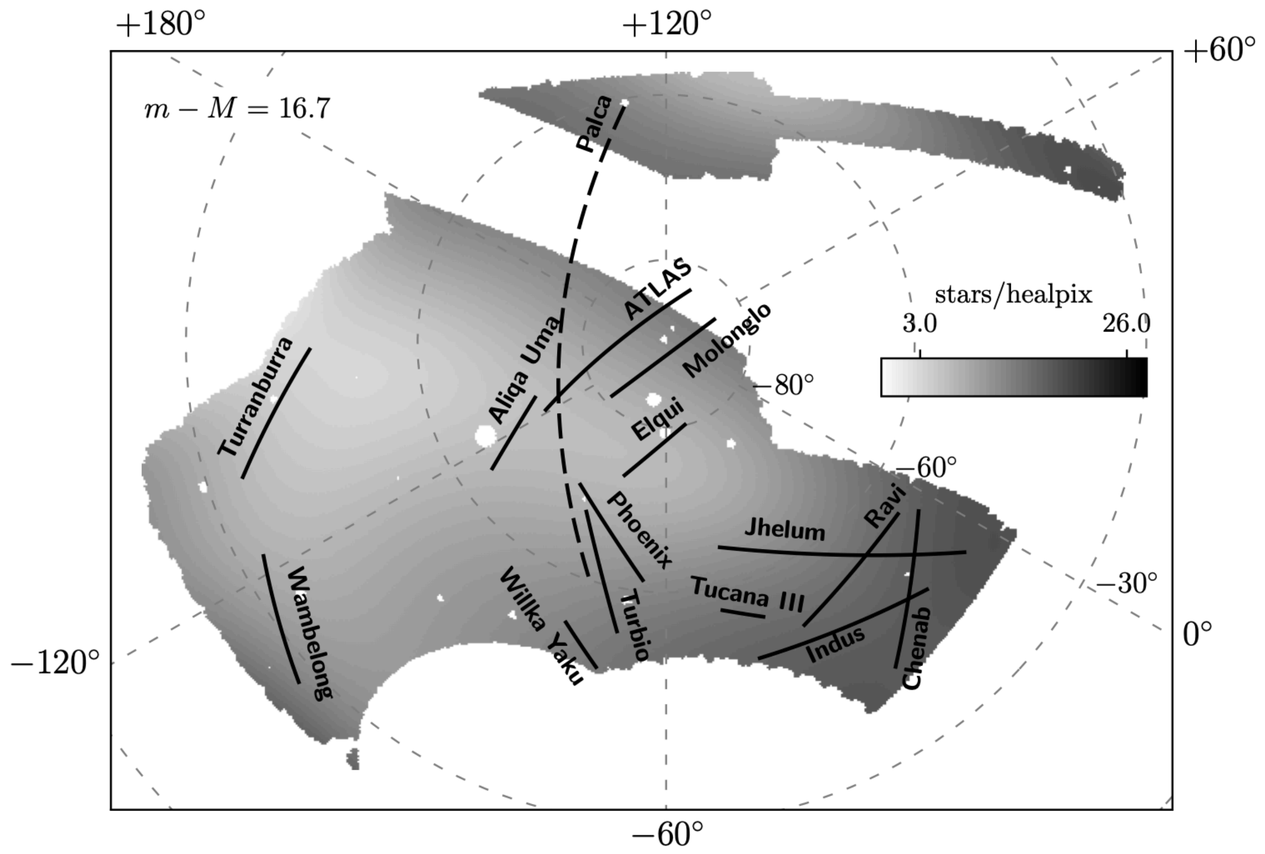
Table 4. Y3 GOLD Foreground Region Mask

Flag Bit	Area (deg^2)	Description
1	220.59	2MASS moderately bright star regions ($8 < J < 12$)
2	22.63	Large nearby galaxies (HyperLEDA catalog)
4	91.12	2MASS bright star regions ($5 < J < 8$)
8	100.61	Region near the LMC
16	86.51	Yale bright star regions
32	0.53	Globular clusters
64	61.13	Brightest stars

https://arxiv.org/abs/1801.03097

Data almost always need to be “cleaned”.

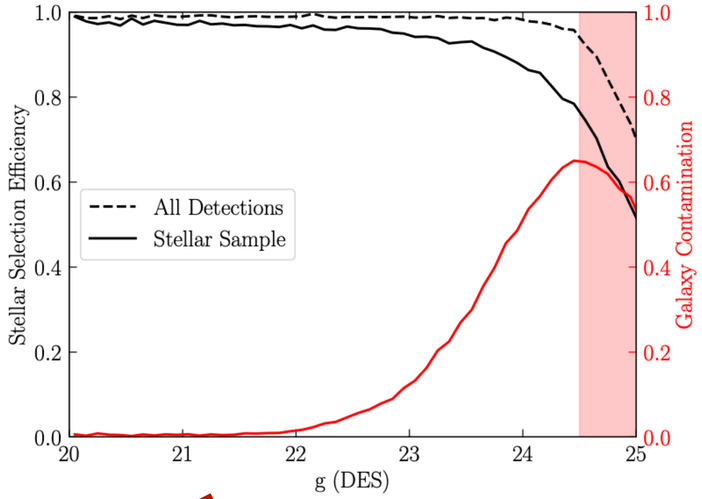
- Example: stellar stream detection in DES data. 11 new stellar streams detected. Shipp et al. 2018.

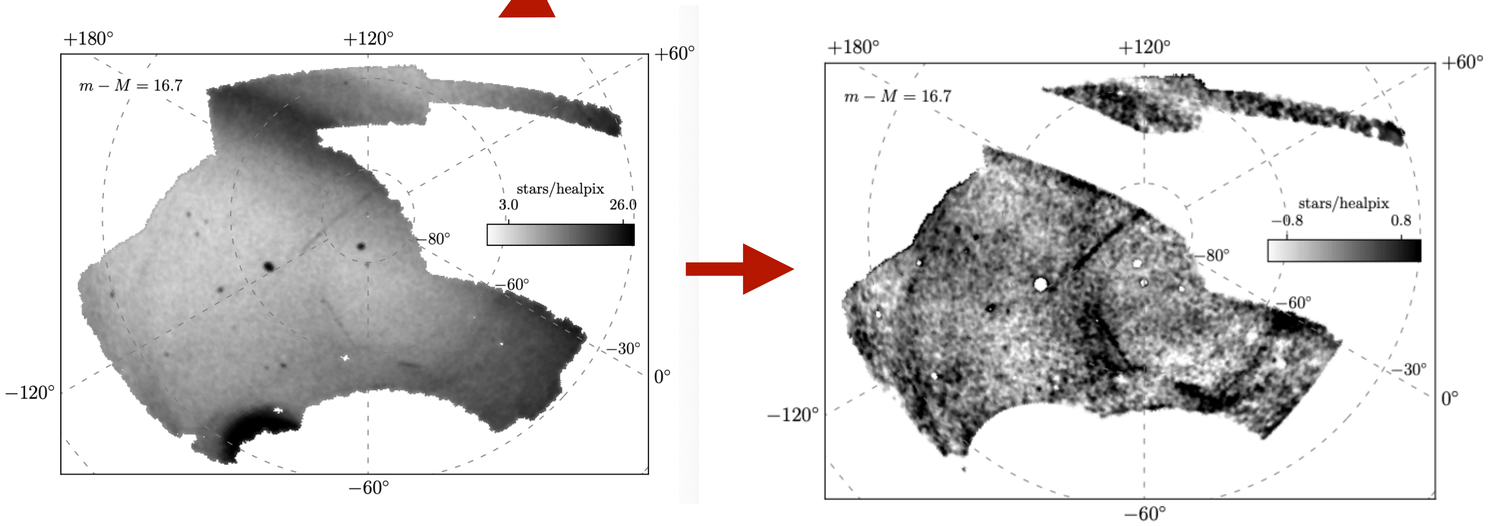


https://arxiv.org/abs/1801.03097

Data almost always need to be “cleaned”.

- Example: stellar stream detection in DES data. Shipp et al. 2018.
    - Apply stellar color cuts according to synthetic isochrone at different distance moduli.
    - Generate residual stellar density maps to detect faint stellar streams.
- Applied filtering in terms of size and magnitude to select a relatively complete and pure stellar sample.
- Certain regions of the sky are masked to improve model fits.



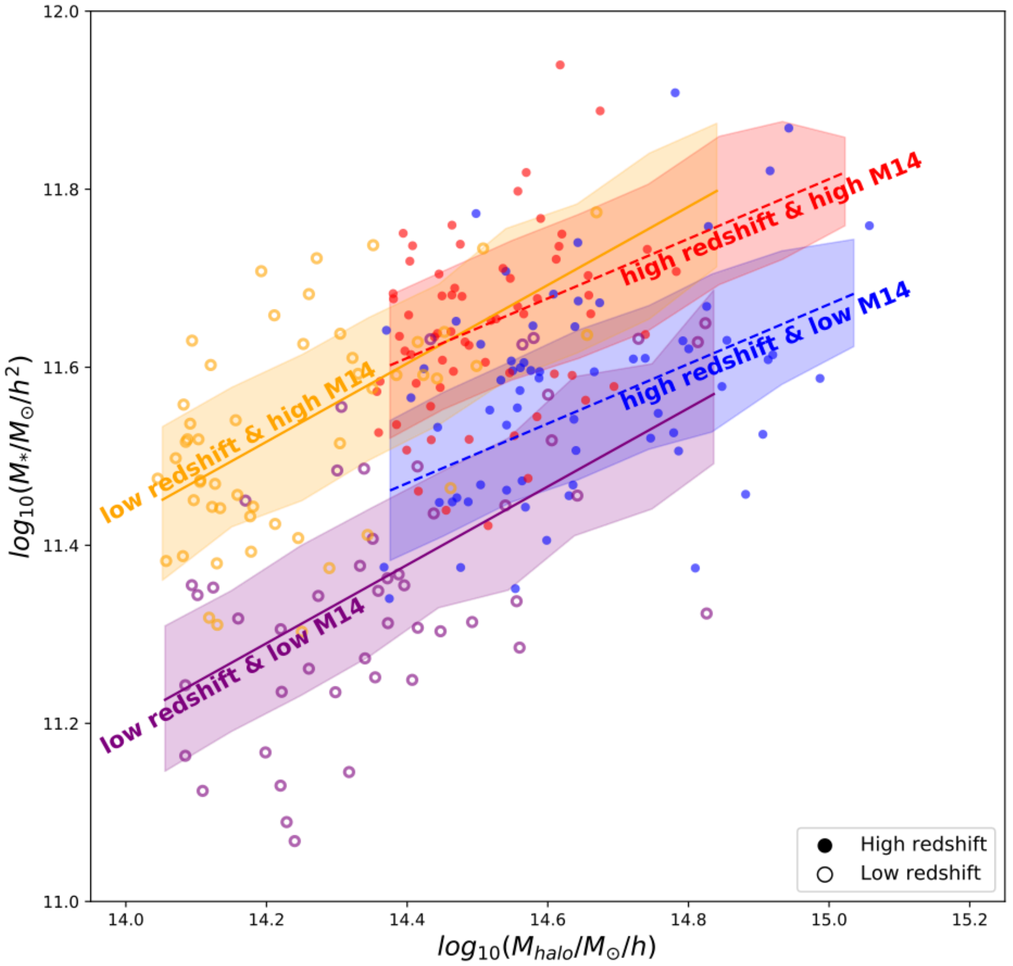


https://arxiv.org/abs/1801.03097

Combine different data sets in one analysis

- **Example:** characterizing the growth of cluster central galaxies — how their stellar mass to halo mass changes with cluster properties. Golden-Marx + 2021

- Relies on using data sets form the Sloan Digital Sky Survey (SDSS) and DES.

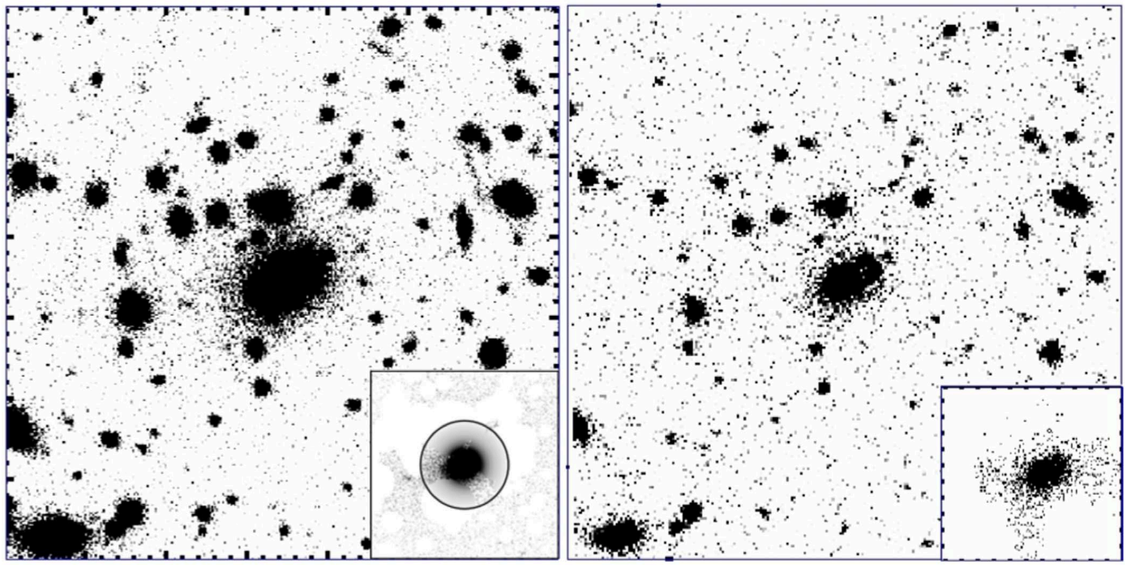


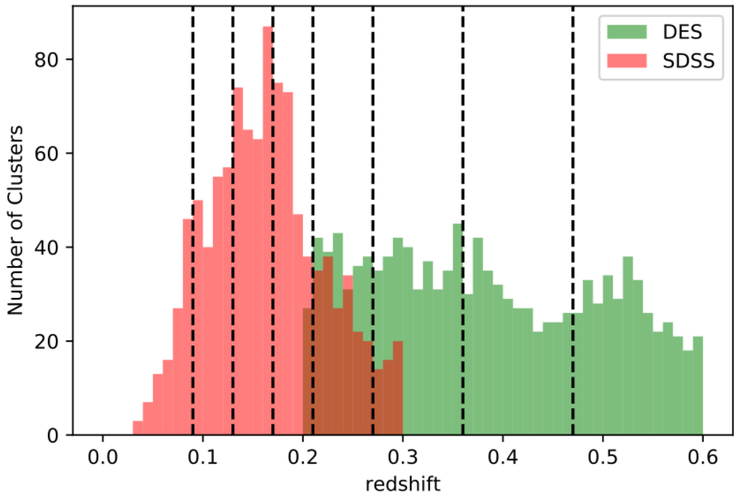
https://arxiv.org/abs/2107.02197

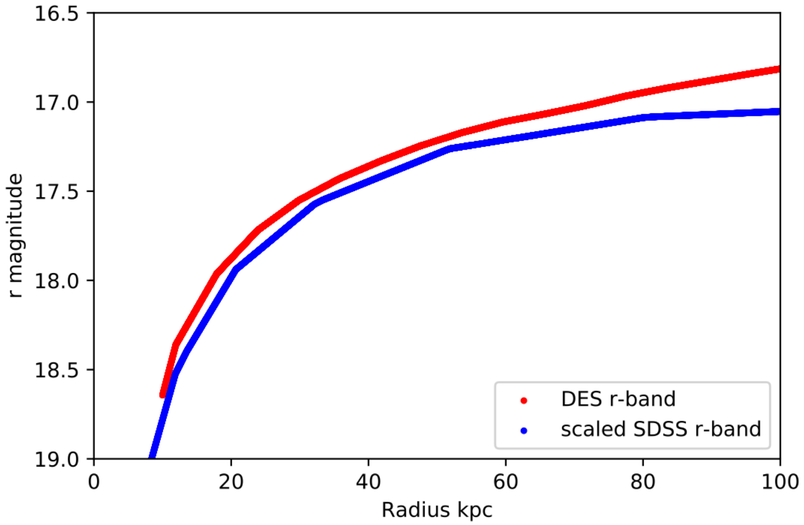
Combine different data sets in one analysis

- SDSS and DES data have very similar properties.

- An overlapping sample of central galaxies between SDSS and DES helps “calibrating” the differences between the two.



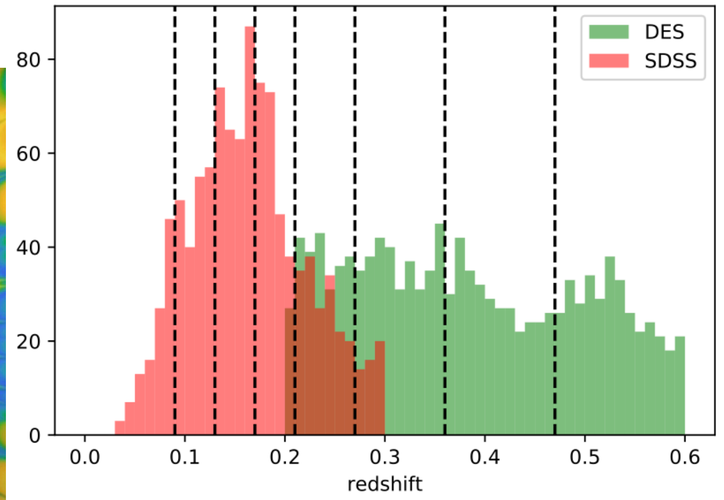


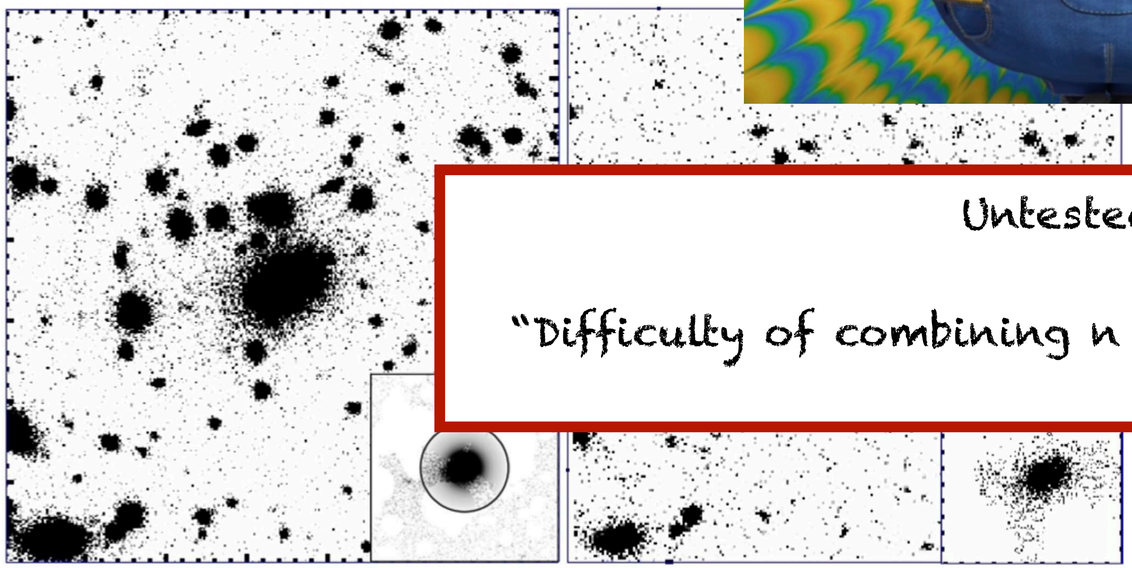


https://arxiv.org/abs/2107.02197

Combine different data sets in one analysis

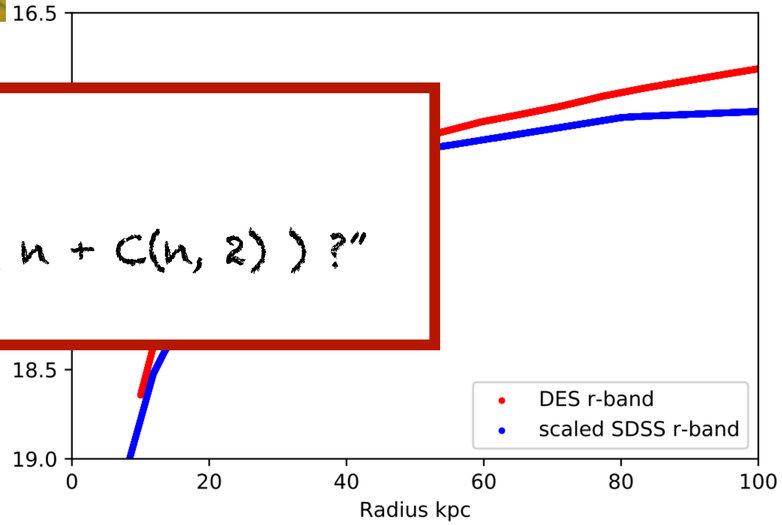
- SDSS and DES data have very similar properties.
- An overlapping sample of ce
between SDSS and DES hel
“calibrating” the differences





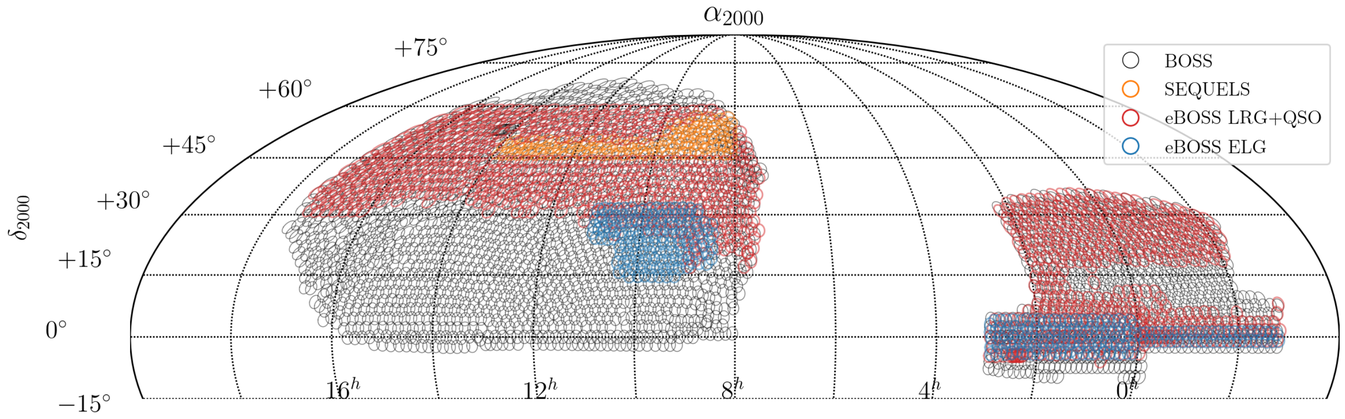
Untested theory 1:

“Difficulty of combining n data sets is O(n + C(n, 2)) ?”



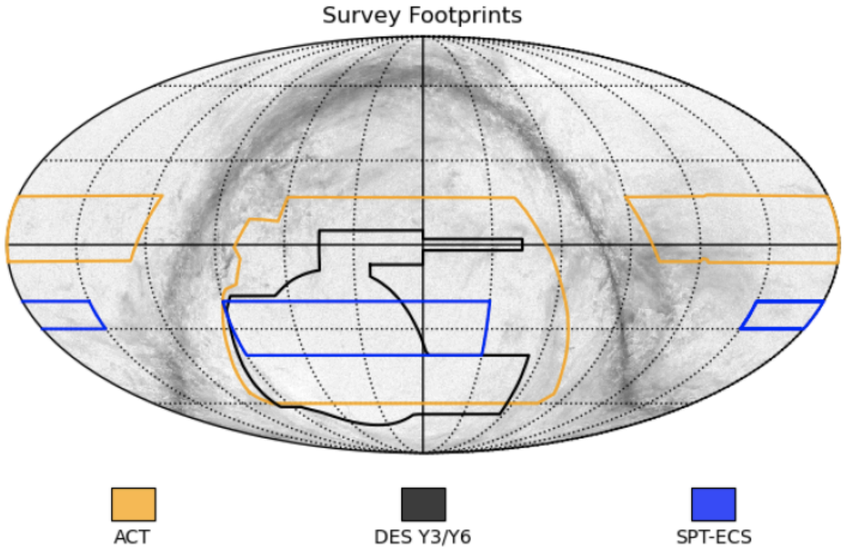
https://arxiv.org/abs/2107.02197

Multi-wavelength observations



https://www.sdss4.org/dr16/

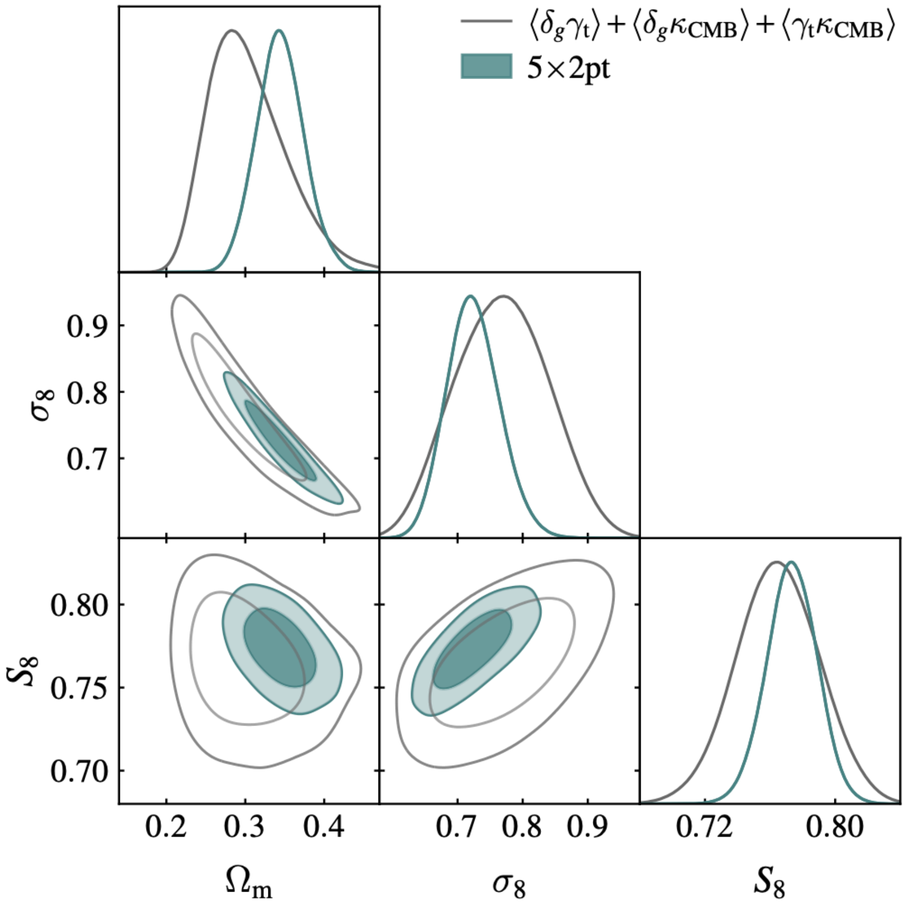
- Spectroscopic surveys: SDSS, DESI
- X-ray: Newton XMM, Chandra, eRosita …
- Cosmic Microwave Background: South Pole Telescope, Atacama Cosmology Telescope..
- H-alpha, radio observations …



https://arxiv.org/abs/2110.02418

Multi-wavelength observations

- DES+ 2022: Cross correlation between CMB lensing and galaxy positions, galaxy lensing.
- A more consistent way to combine CMB lensing and galaxy survey cosmology analysis.
- Also an important robust test.

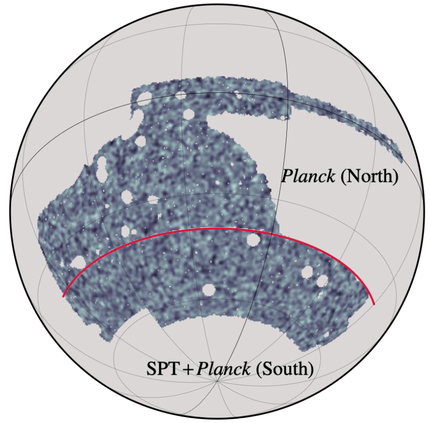


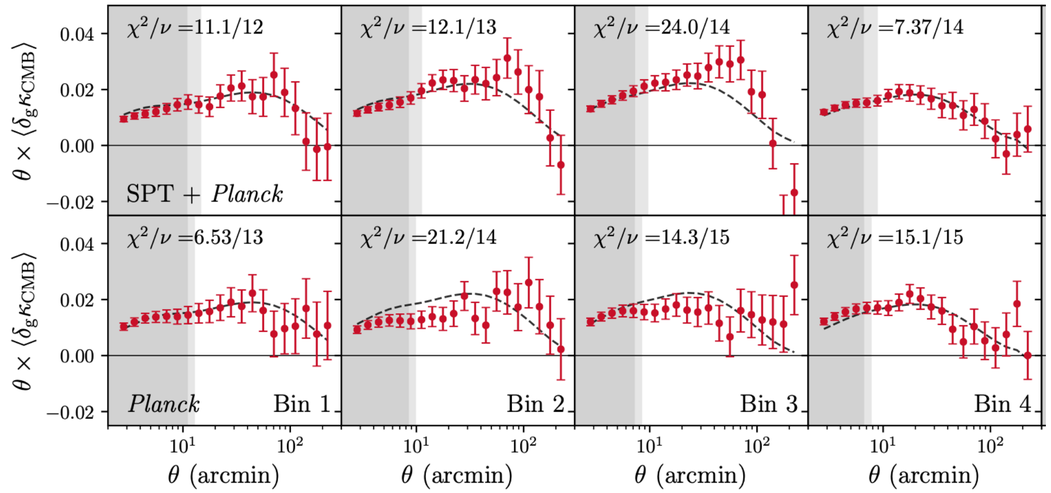
Chang+ 2022 https://arxiv.org/pdf/2203.12440.pdf
Omori+ 2022 https://arxiv.org/pdf/2203.12439.pdf
DES+ 2022 https://arxiv.org/pdf/2206.10824.pdf

Multi-wavelength observations

- The cross-correlation analysis combines DES and CMB observations.
  - Using consistent theoretical formalism toanalyze optical lensing and CMB lensing maps.
- Also needs to combine CMB lensing maps from SPT and Planck because of SPT's incomplete coverage of the DES footprint
  - SPT and Planck maps have different noise properties and filtering choices.
  - Leaving a small 0.5 deg gap between the two lensing maps to reduced the correlation between structures on the boundaries.

When in doubt, compare the measurements from different data sets.





https://arxiv.org/pdf/2203.12440.pdf
https://arxiv.org/pdf/2203.12439.pdf

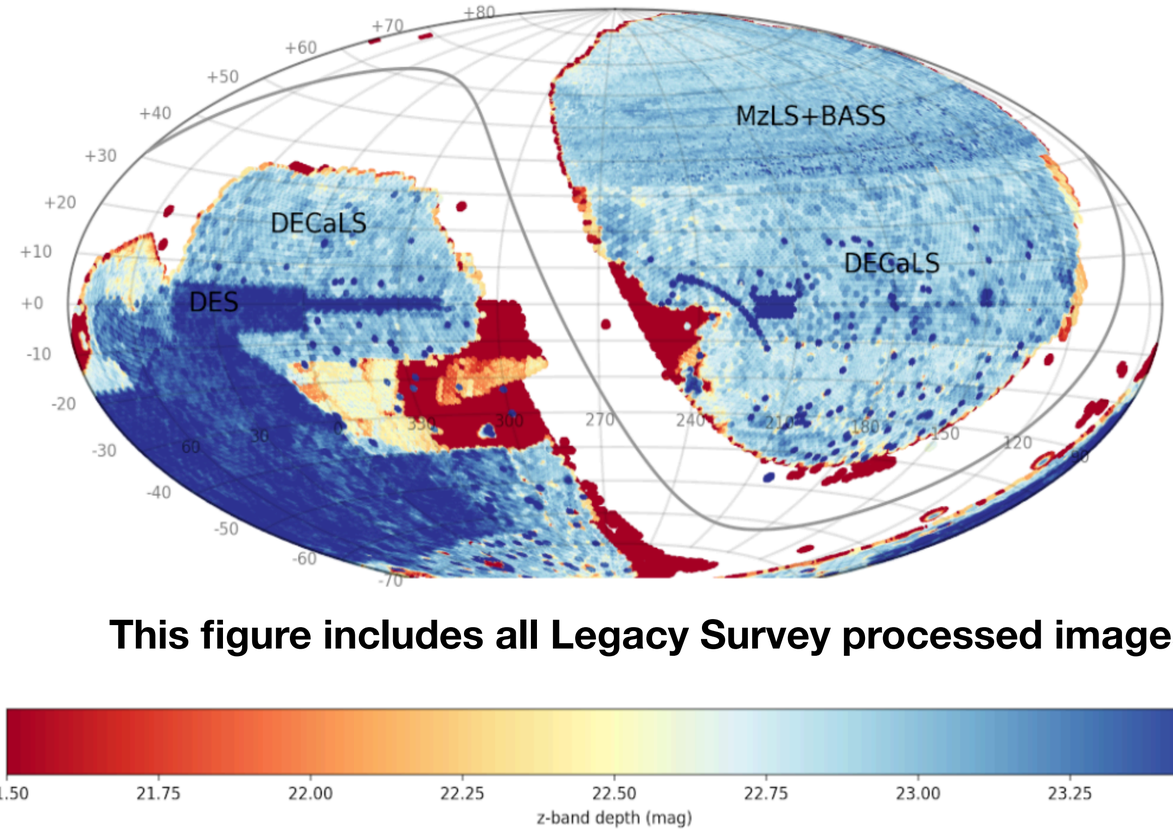
What is not a heterogenous data set?

Another example: The **Legacy Survey** and the Dark Energy Camera Legacy Survey (DECaLS)

- Imaging in g, r, and z bands.
- Main science motivation is to provide targets for the Dark Energy Spectroscopic Instrument (DESI) survey.
- When conducting the observing, the exposure times are dynamically adjusted to produce nearly uniform imaging depth.

https://www.legacysurvey.org/status/

https://arxiv.org/pdf/1804.08657.pdf

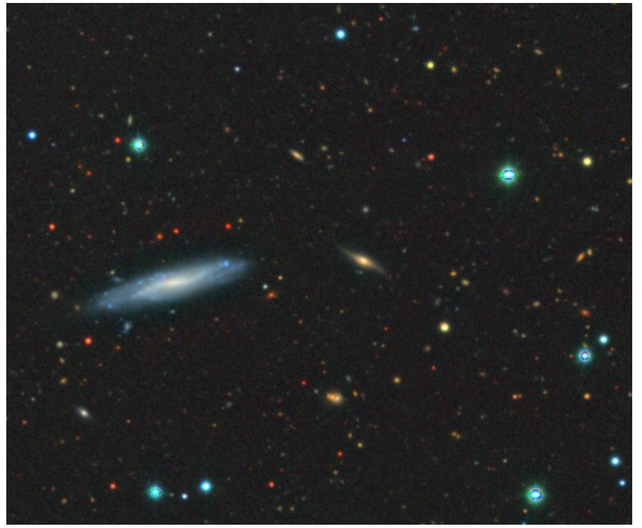


This figure includes all Legacy Survey processed images.

Legacy Survey also processes WISE data

Another example:

- The **Legacy Survey** and the Dark Energy Camera Legacy Survey (DECaLS)
- Also processing **WISE** imaging data.
- WISE: Wide-field Infrared Survey Explorer
  - almost full sky mapping at 3.4, 4.6, 12, and 22 μm, but with angular resolution of 6.1", 6.4", 6.5", & 12.0" in the four bands. (DECam: about 1" resolution)
  - sensitive to high redshift objects, dusty AGNs …



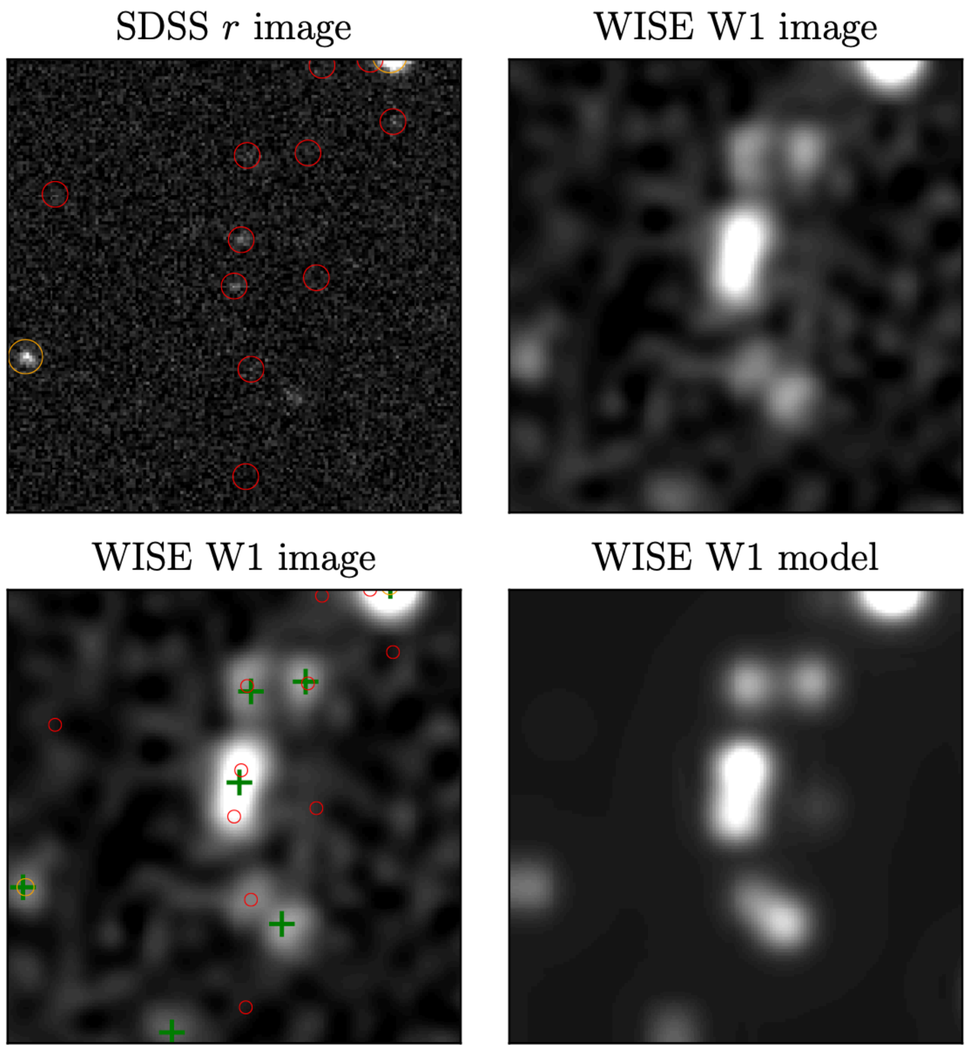
Legacy Survey image DR10

unWISE image of the same field

http://unwise.me

Legacy Survey also processes WISE data

- Legacy Survey data is processed with the tractor code (Lang + 2014).
- A multi-epoch fitting code that can do “forced photometry”.
- An image with better resolution (SDSS, Legacy Survey) can be used to help with measurements in a lower solution image (WISE data).



https://arxiv.org/pdf/1410.7397.pdf

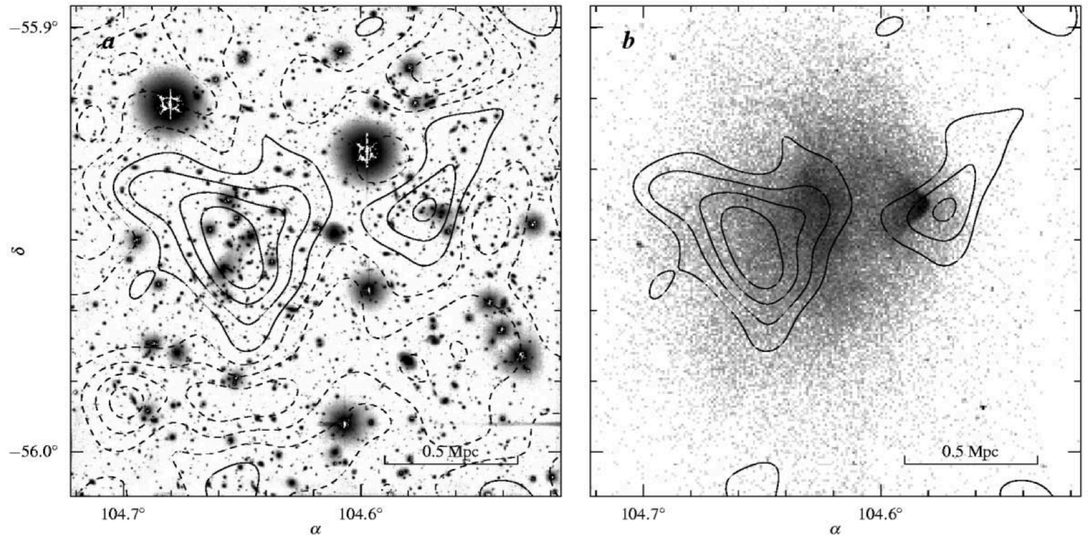
“Targeted” Multi-wavelength Observations

- Example: using Bullet cluster to constrain dark matter self-interaction cross section. Markevitch+ 2004.

  - Spectroscopic redshifts were used to derive the velocities of the two sub clusters.

  - Optical images are used to derive the dark matter mass distribution (through lensing) and galaxy density distribution.

  - X-ray observations reveal the gas morphology and positions.



https://arxiv.org/abs/astro-ph/0309303

More surveys, but not “extremely-wide” surveys

Example: GoGREEN. Gemini Observations of Galaxies in Rich Early ENvironments

- Multi-object spectroscopy of 21 galaxy clusters in the redshift range of 0.8 to 1.5
- Optical and infrared images from GMOZ z-band, Spitzer IRAC, Hubble and ground-based observatories.
- Also contains spectroscopic surveys — 2257 redshifts.

https://arxiv.org/pdf/2009.13345.pdf
https://arxiv.org/pdf/1711.05280.pdf

http://gogreensurvey.ca

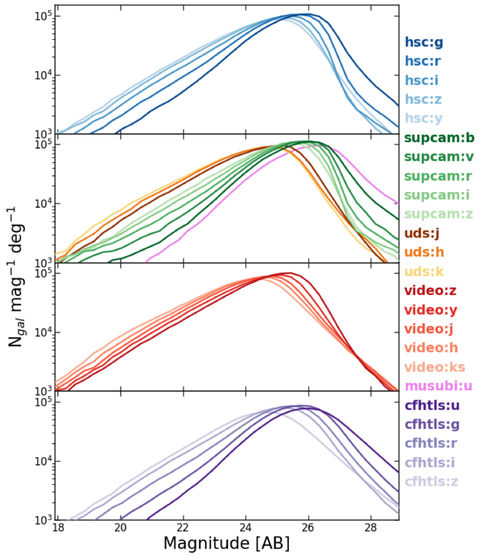
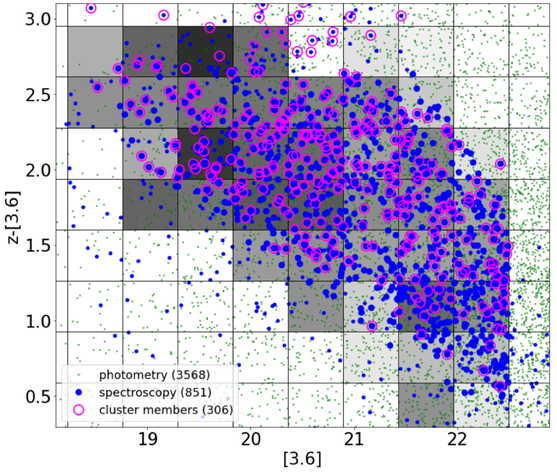
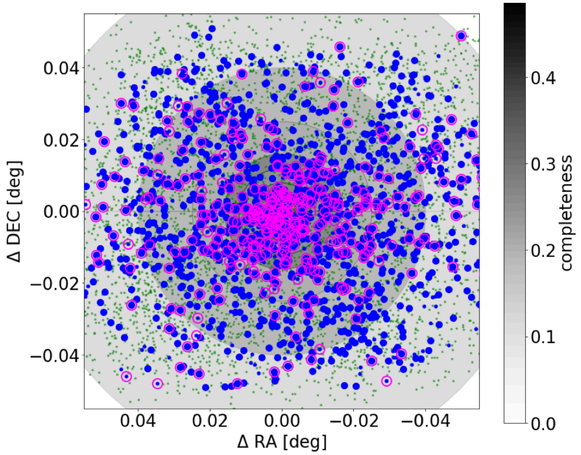


Figure 7. Number counts of sources detected using our multi-band detection image shown for each filter as a function of magnitude.





More surveys, but not “extremely-wide” surveys

Example: GoGREEN.

Gemini Observations of Galaxies in Rich Early ENvironments

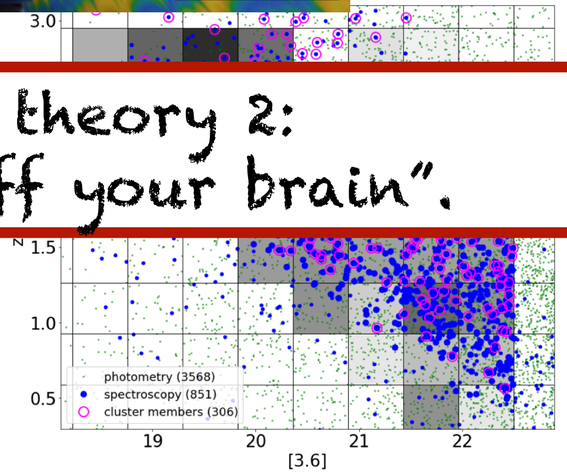
https://arxiv.org/pdf/
.pdf
rg/pdf/
.pdf

nsurvey.ca

- Multi-object spectroscopy of 21 clusters in the redshift range of 0 1.5
- Optical and infrared images from GMOZ z-band, Spitzer IRAC, Hubble and ground-based obse
- Also contains spectroscopic surveys — 2257 redshifts

awn.com

Untested theory 2:
“don’t turn off your brain”.



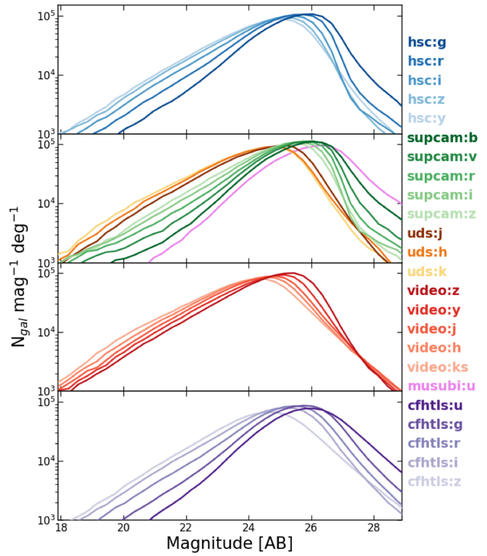
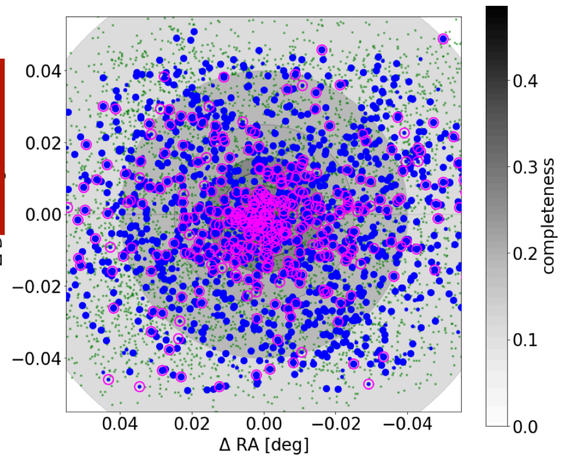


Figure 7. Number counts of sources detected using our multi-band detection image shown for each filter as a function of magnitude.



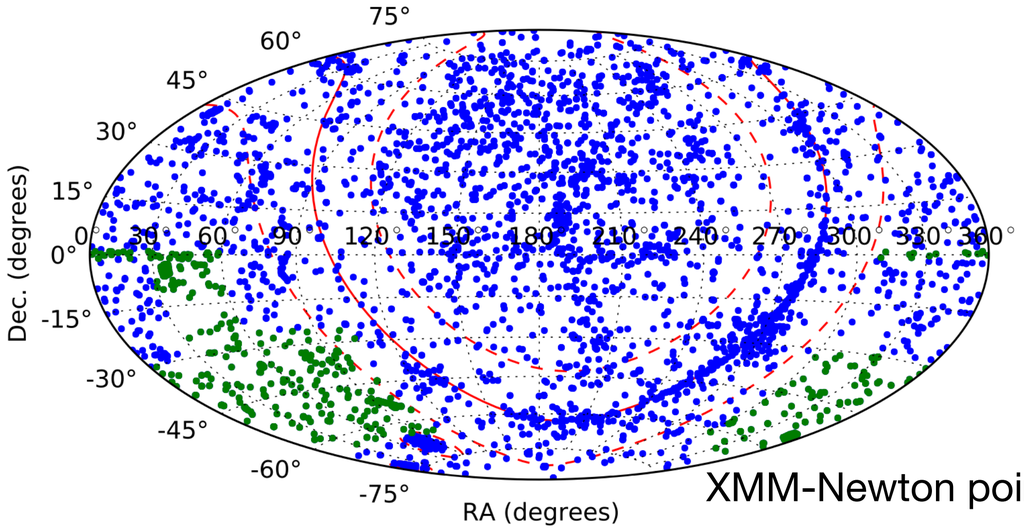
More surveys, but not “extremely-wide” surveys

Example: The XMM-Newton Cluster Survey (XCS)

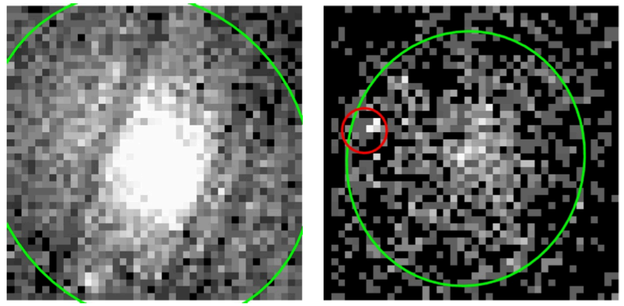
“A serendipitous search for galaxy clusters using all publicly available data in the XMM-Newton science archive. ”

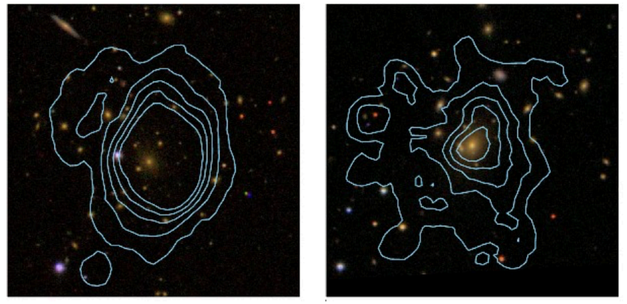
- Searching for diffuse cluster-like emissions in X-ray archival.
- Cluster candidates confirmed with optical imaging and spectroscopy observations.
- Optical images and spectroscopy came from proposed follow-up observations as well as SDSS and DES. 503 clusters identified in the first data release.

https://arxiv.org/pdf/1106.3056.pdf
https://arxiv.org/pdf/1010.0677.pdf



XMM-Newton pointings



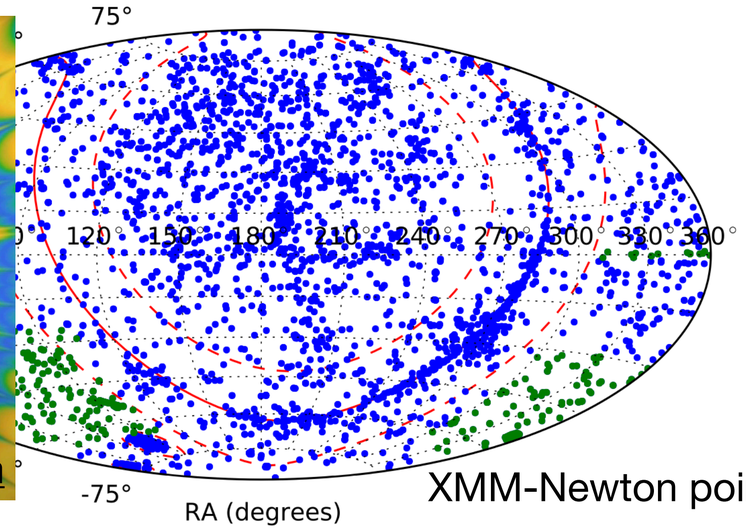


More surveys, but not “extremely-wide” surveys

Example: The XMM-Newton Cluster Survey

“A seredipitous search for galaxy clusters u
all publicly available data in the XMM-Newt
science archive. ” 503 clusters identified in
first data release.


awn.com



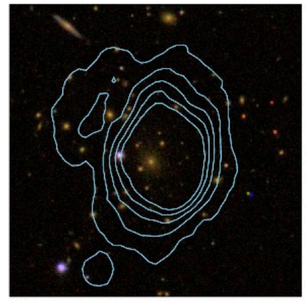
XMM-Newton pointings

- Searching for diffuse cluster-like emissio
  X-ray archival.

- Op
  obs
  car
  Op
  proposed follow-up observations as well as
  SDSS and DES.

Untested theory 3:

“it helps if the analysis is designed with a clear goal. ”



https://arxiv.org/pdf/1106.3056.pdf
https://arxiv.org/pdf/1010.0677.pdf

Questions?