Yale

Methods of data manipulation in galaxy surveys

Talk at NYU, Langone Medical centre

Angela Burden, Yale

January 31, 2017

Table of contents

- Introduction , who I am and why I am here (5 mins)
- Methods- 7 subsections (5 mins each)
  1. 2D to 3D, where do we point our telescope?
  2. Highlighting the signal.
  3. Removing noise and unwanted artefacts.
  4. Filling in data/ creating simulations.
  5. Modelling linear flow.
  6. Data analysis techniques.
  7. Categorising data w/ machine learning
- Summary

Outline

- Introduction , who I am and why I am here (5 mins)
- Methods- 7 subsections (5 mins each)
  1. 2D to 3D, where do we point our telescope?
  2. Highlighting the signal.
  3. Removing noise and unwanted artefacts.
  4. Filling in data/ creating simulations.
  5. Modelling linear flow.
  6. Data analysis techniques.
  7. Categorising data w/ machine learning
- Summary

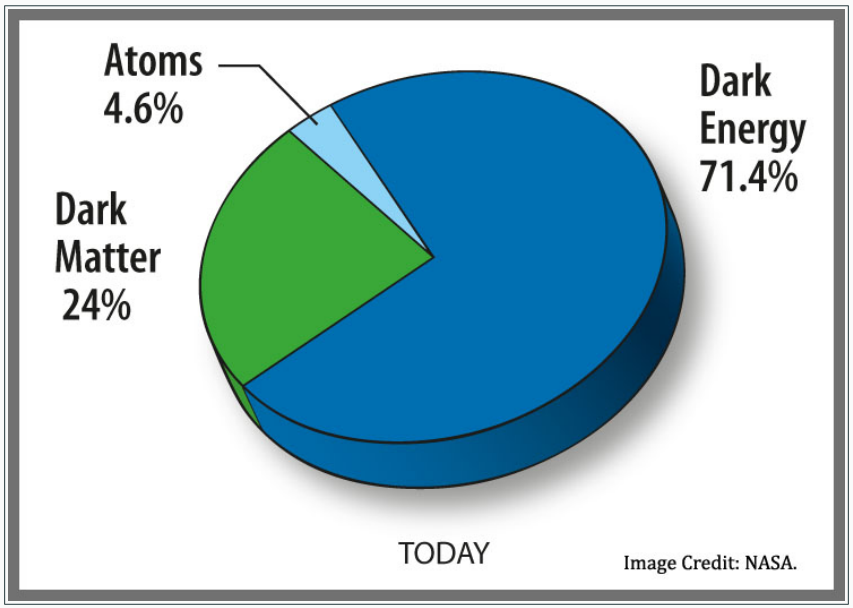
Who am I?

- Postdoc at Yale (1.5 years)
- Cosmology (PhD) at Institute of Cosmology and Gravitation, UK
- Physics Masters and BA at Sussex University, UK

-Work on galaxy surveys

-Use patterns in the galaxy distribution $\rightarrow$ expansion rate of the Universe

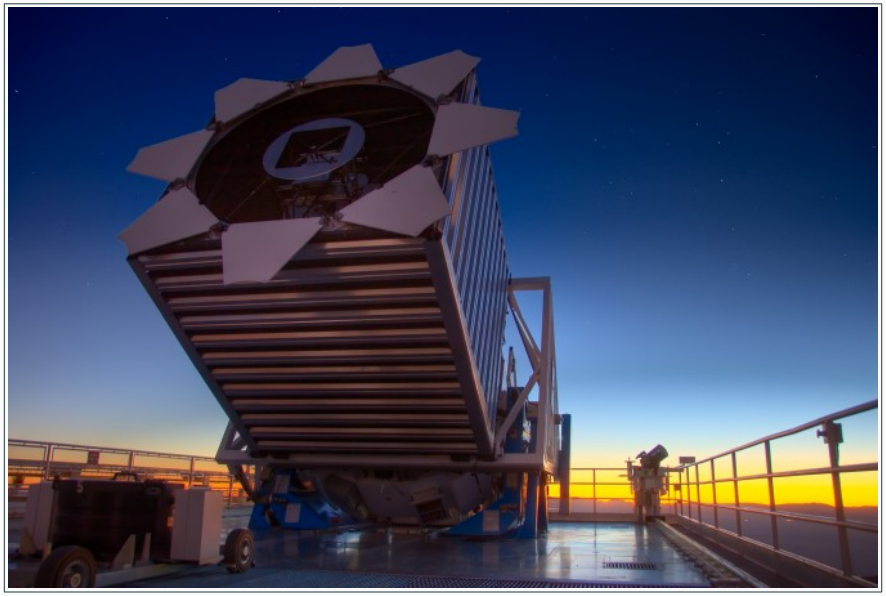
-Figure out what the stuff in the pie chart is....



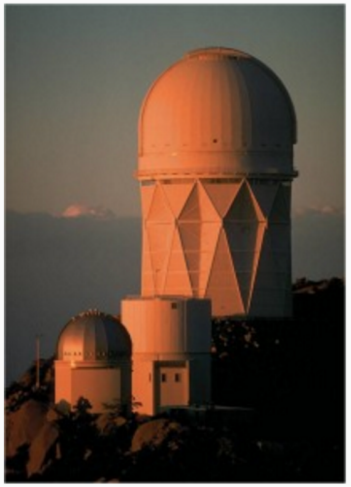
NB "Dark" = we don't know!

Data collection tools

The Sloan Digital Sky Survey (SDSS)



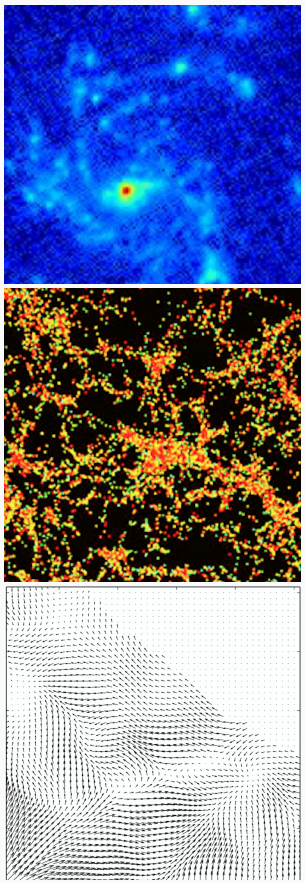
The Dark Energy Spectro. Instrument (DESI)



Exterior of Kitt Peak Mayall 4-meter telescope (Image: NOAO/AURA/NSF)

Data types

-Introducing data types.



- galaxy image
- galaxies as point sources
- flow lines (may call displacement fields)

Outline

- Introduction , who I am and why I am here (5 mins)
- Methods- 7 subsections (5 mins each)
  1. 2D to 3D, where do we point our telescope?
  2. Highlighting the signal.
  3. Removing noise and unwanted artefacts.
  4. Filling in data/ creating simulations.
  5. Modelling linear flow.
  6. Data analysis techniques.
  7. Categorising data w/ machine learning
- Summary

Section 1

2D to 3D data

1. 2D $\rightarrow$ 3D

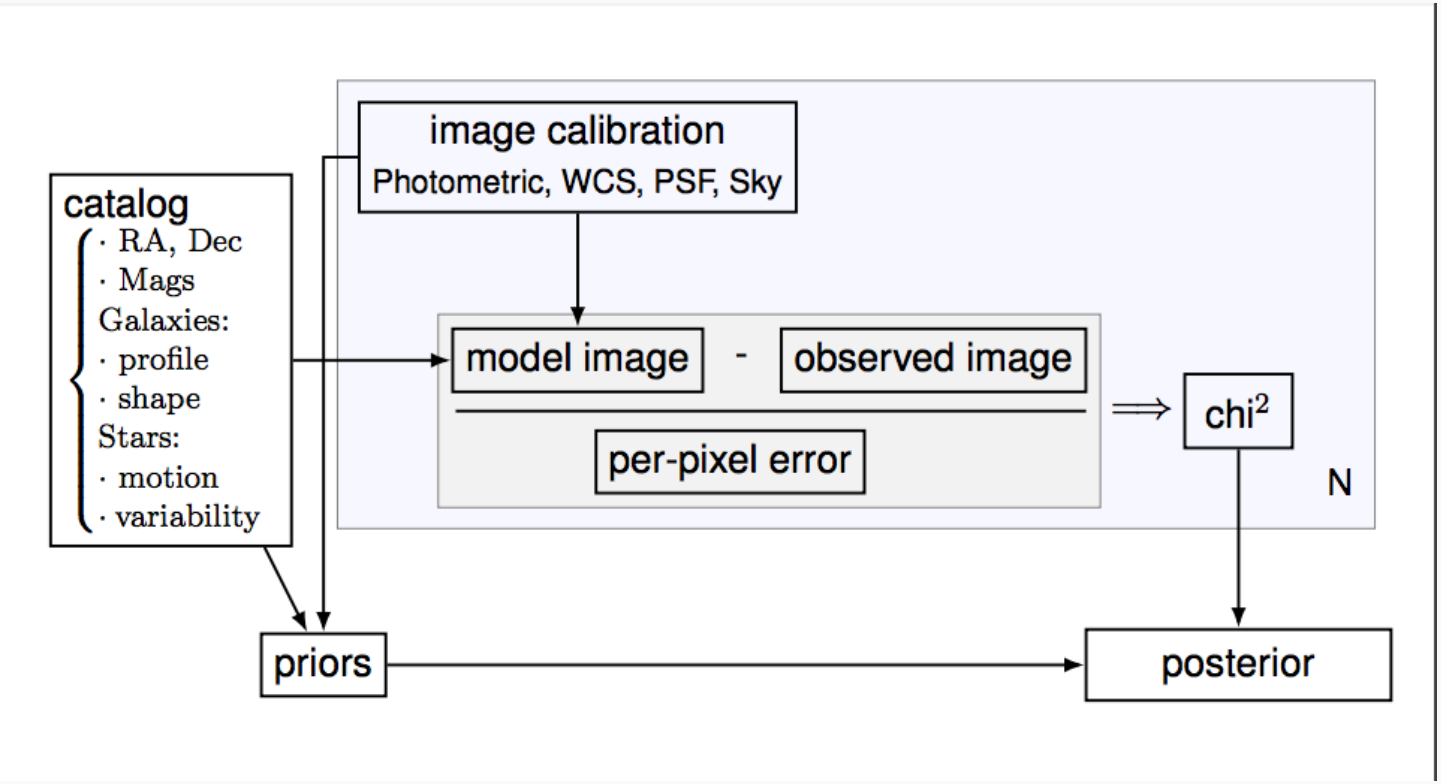
Two types of galaxy observation **photometric** (2D) and **spectroscopic** (3D).

- **photometric** - 2D pictures, angular positions on the sky- no distance information.

We use the 2D map as a target locator to indicate where we have to point our optical fibres but how do we identify targets?

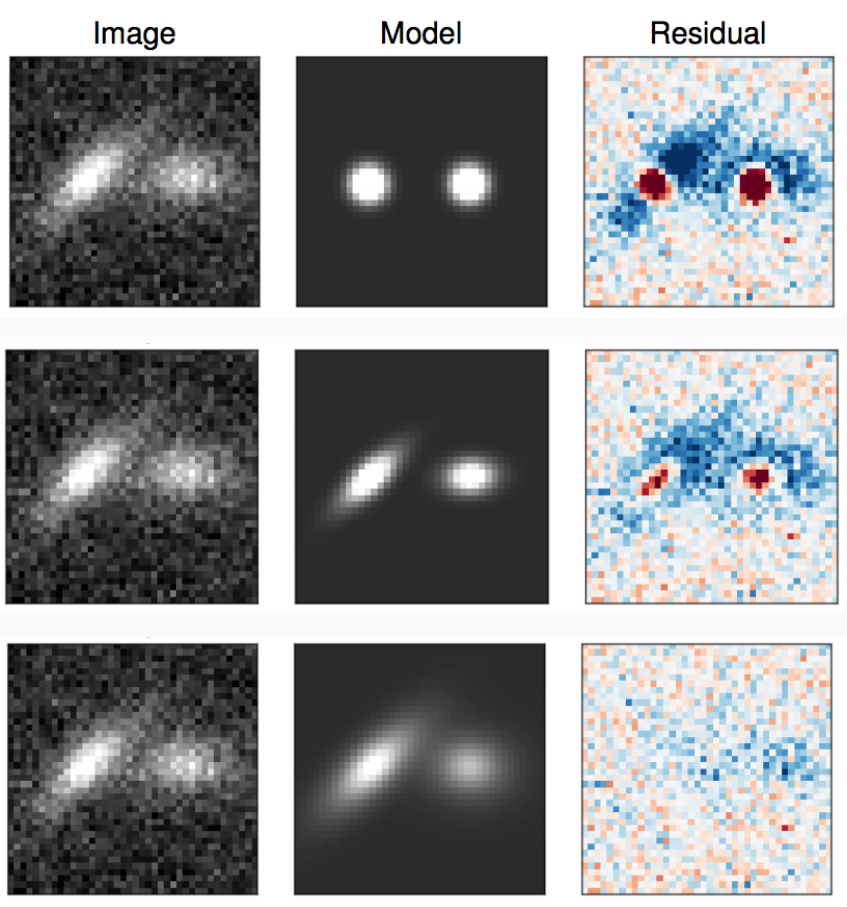
1. The tractor, machine learning code

One method is to use machine learning techniques, eg "The Tractor code" written by Dustin Lang at Berkeley.



What is the probability of matching this image given the set of parameters and model?

1. The tractor, machine learning code

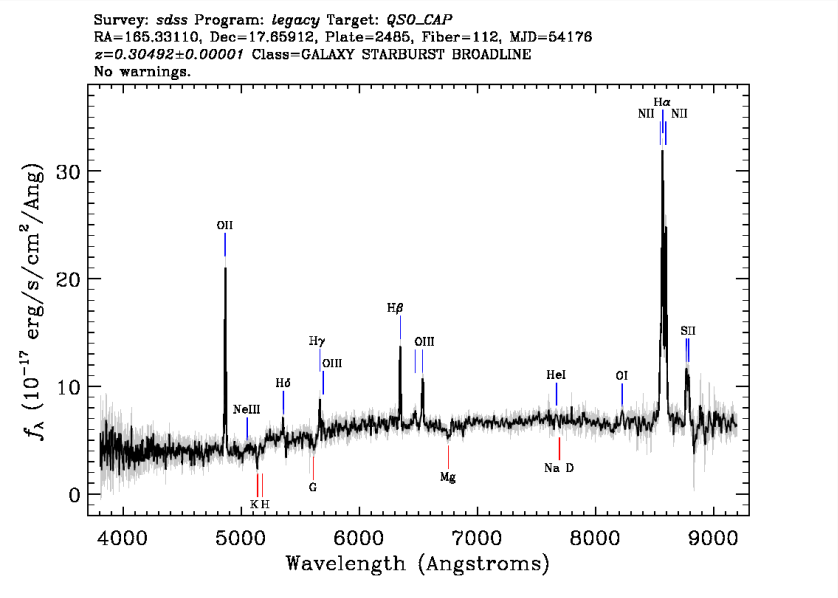


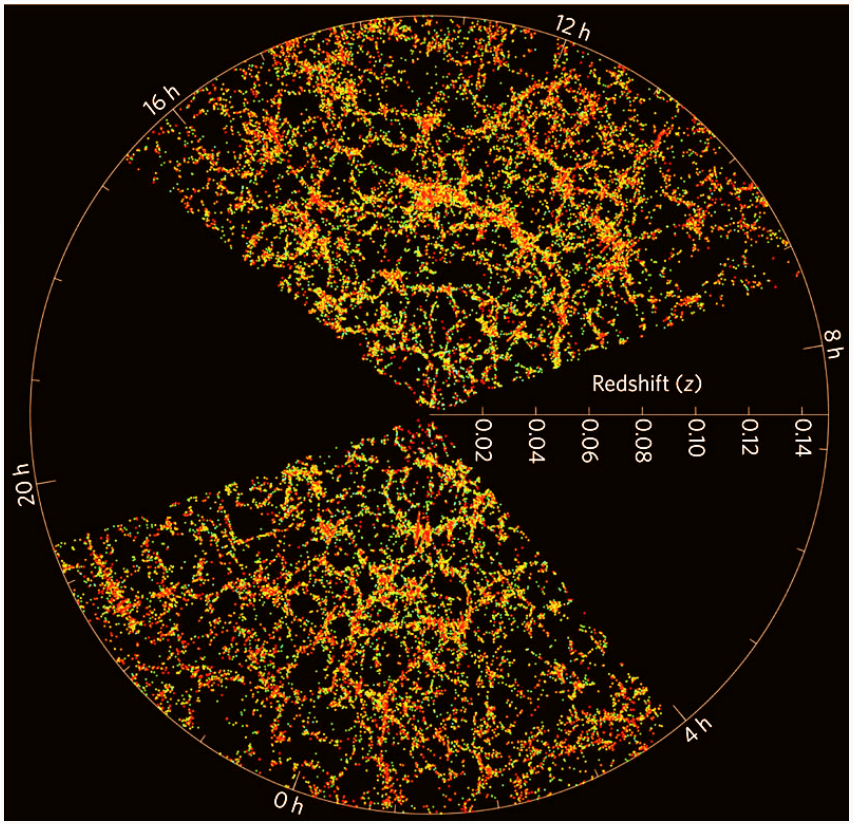
Ref :The DESI Experiment Part I: Science,Targeting, and Survey Design,
Authors: DESI collab . inc **AB**
arXiv:1611.00036

1. spectroscopic

Once we know where to point our telescope and where to position the optical fibers

- **spectroscopic** -light $\rightarrow$ redshift $\rightarrow$ distance to get the 3rd dimension.





Section 2

Highlighting the signal

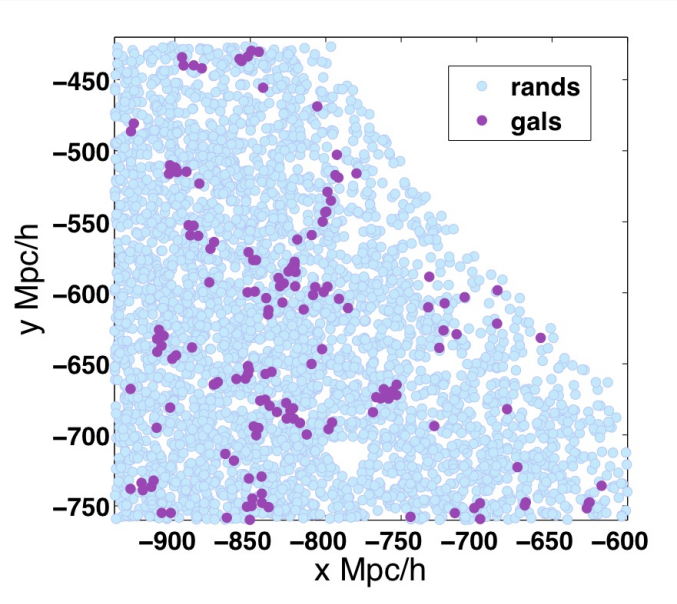
2. enhancing the signal

Galaxies $\rightarrow$ points in a cartesian grid {x,y,z}

$\rightarrow$ binned into voxels {Vx, Vy, Vz}

Previous image; how do we know if regions are empty OR we didn't collect data there

- foreground objects
- atmospheric conditions
- instrument defects



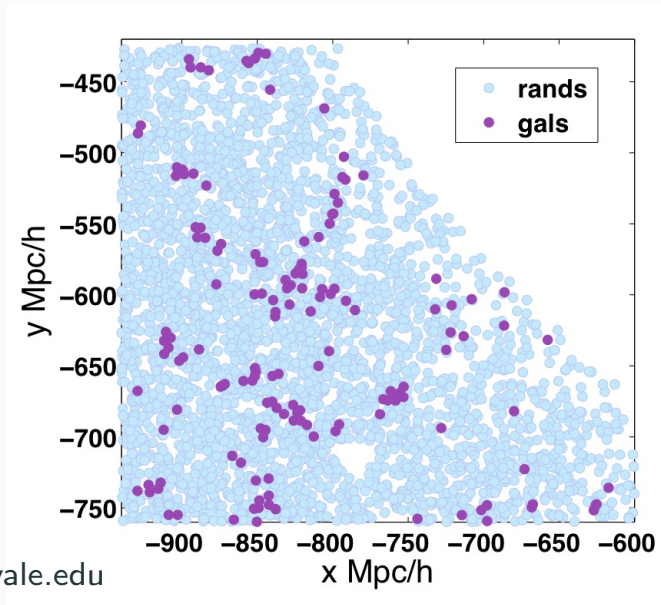
2. enhancing the signal

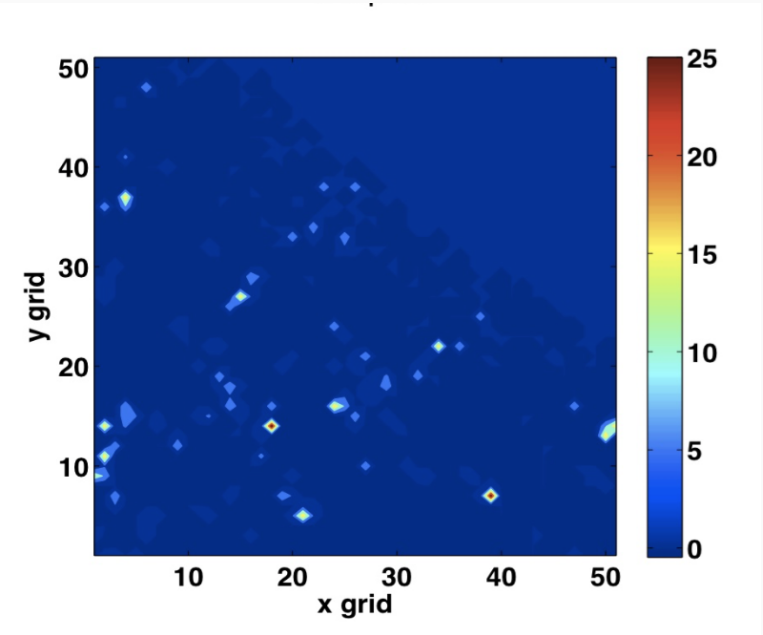
- Highlight the signal in the data.
- Use a *random* catalogue.
- IF there was no signal in the data, this is what it would look like.
- Monte Carlo random data points into the same volume as the data.
- Compute the overdensity from the density where:

$$\delta = \frac{\rho_g - \rho_r}{\rho_r}. \tag{1}$$

If no data AND no random points $\rightarrow$ no signal.

if no data BUT random points $\rightarrow$ actual under-dense part of the universe.



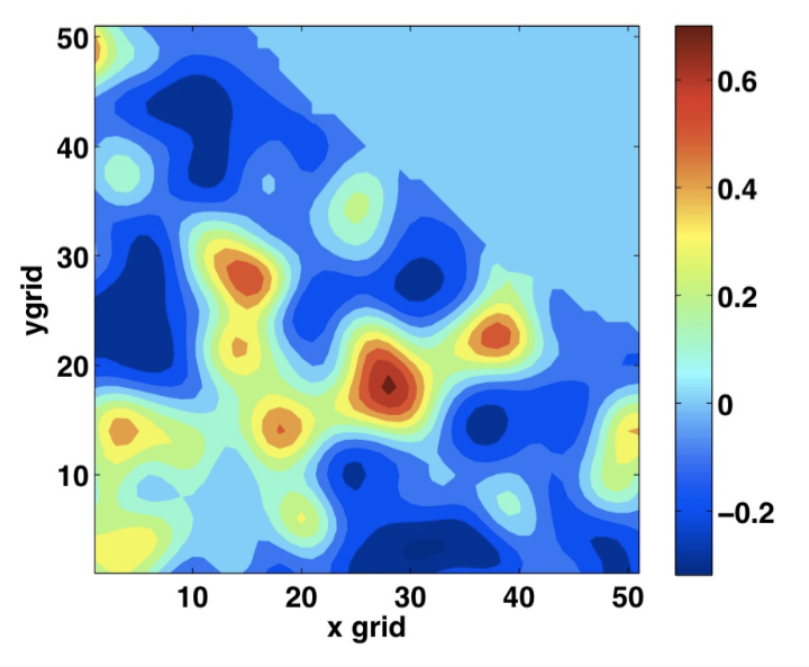


Section 3

Removing noise and artefacts

3. smoothing the data

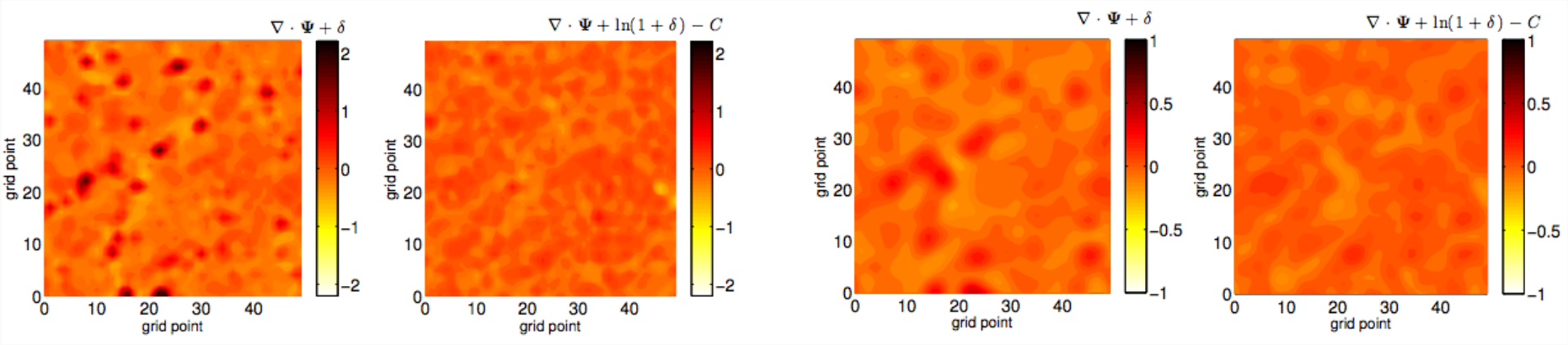
- Many ways of removing noise or sharp non-linear spikes in the signal.
- In this example, smoothing by convolving the signal with a Gaussian function gives best result.
- Make models of the expected true signal, add noise, include artefacts, move forward in time,
- Smooth the data.
- Can we retrieve the original true signal?



Smoothed with Gaussian

$$S(k) = \exp -k^2 R^2/4$$

3. smoothing the data



Optimise smoothing technique depending on

- sparsity of data
- smoothing scale
- edge to volume/area ratio.
- volume of data
- binning algorithm and grid size.

Burden et al. Mon. Not. R. Astron. Soc. 453 (2015)
Burden et al. Mon. Not. R. Astron. Soc. 445 (2014)

Section 4

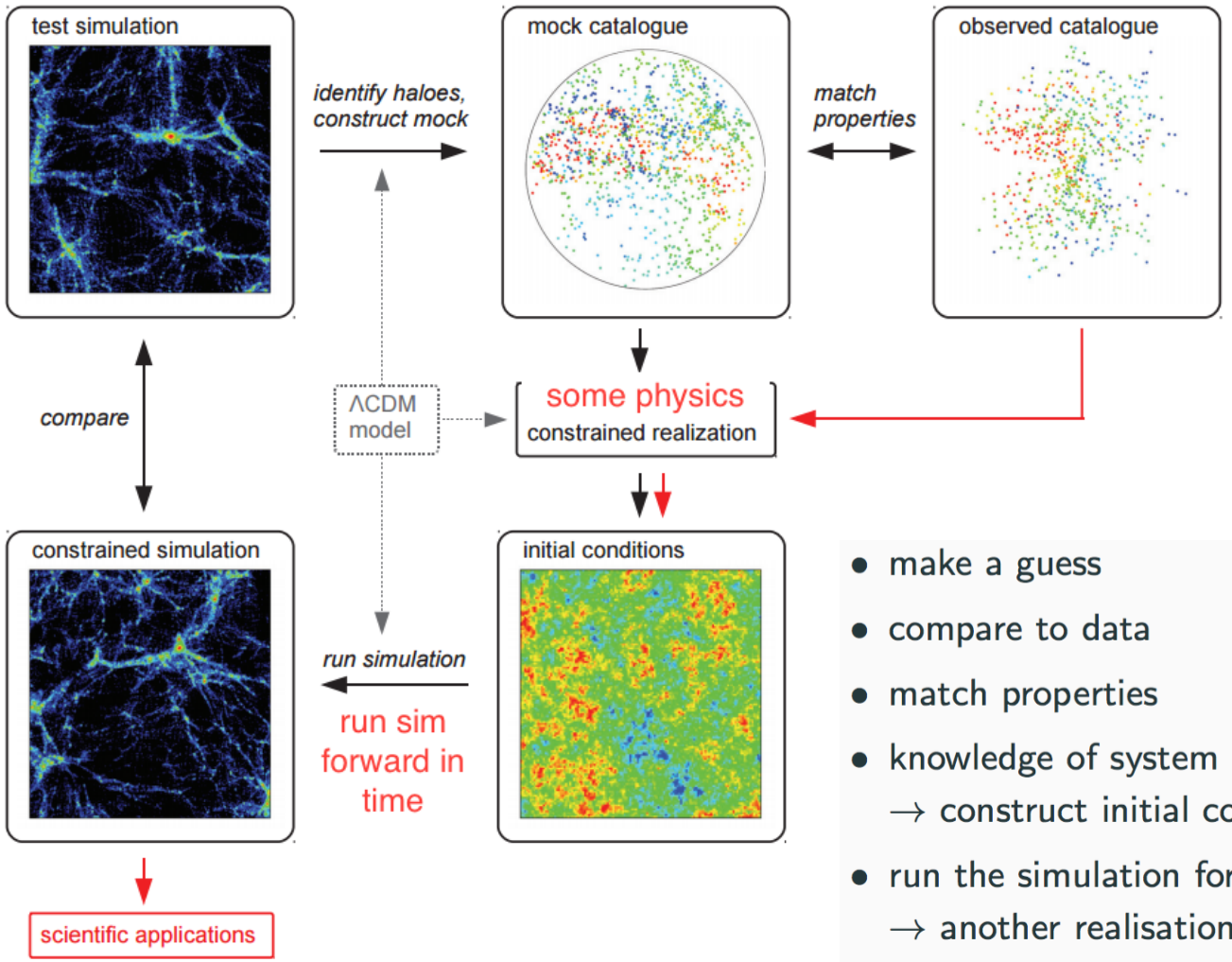
Filling in data/ creating simulations

4. Constrained realisations

Constrained realisations

- Used to **fill** sparse data, **pad** data, or **create** data simulations.
- In the example -create simulations of the data.
- Ask, if another universe/data set had the same initial conditions, what is the probability that we would recover the signal we see now?
- Is our signal real or just coincidence (a bit like p-value to reject NULL hypothesis).

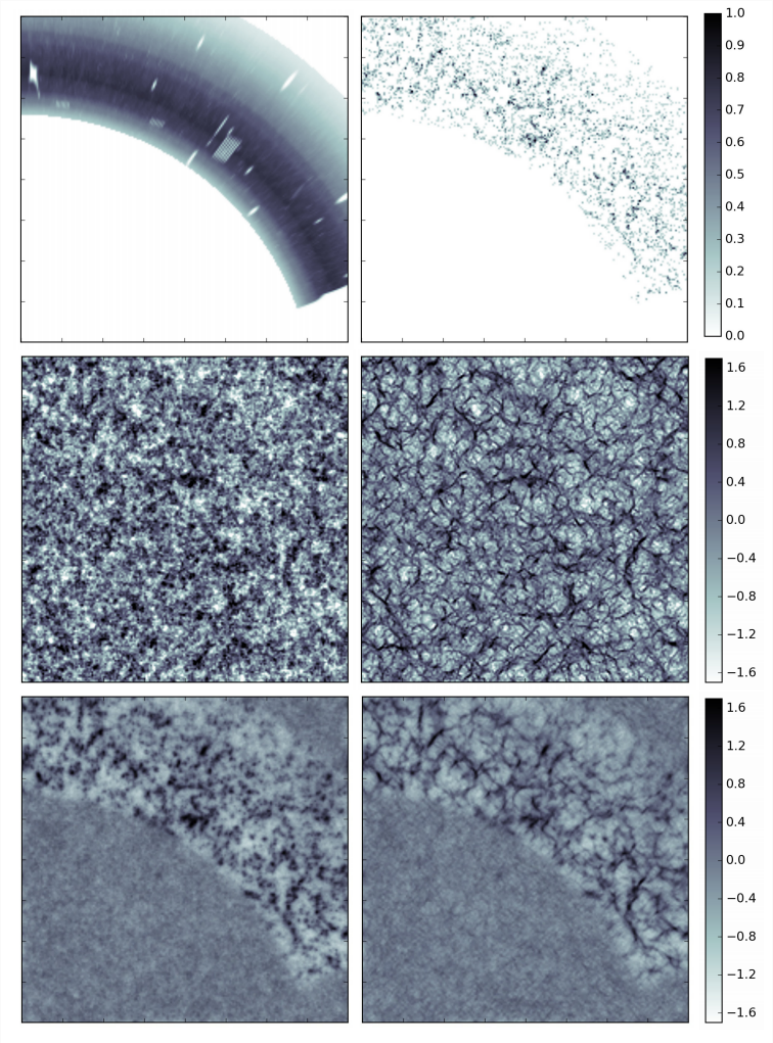
Doumler 2012



- make a guess
- compare to data
- match properties
- knowledge of system + CR
  → construct initial conditions
- run the simulation forward in time
  → another realisation.

Baysian phase space data reconstruction

- What is the probability of recovering this data given a set of parameters, i.e. $P_\delta(\delta_L|\theta)$.
- Run until convergence.
- Repeat to get many samples.



4. other methods to optimize signal

Other methods

- Sparse representations
- Component separation
- Inverse problems and sparse solutions
- Missing data interpolation
- Object detection and Poisson noise
- Compressed Sensing $\longrightarrow$

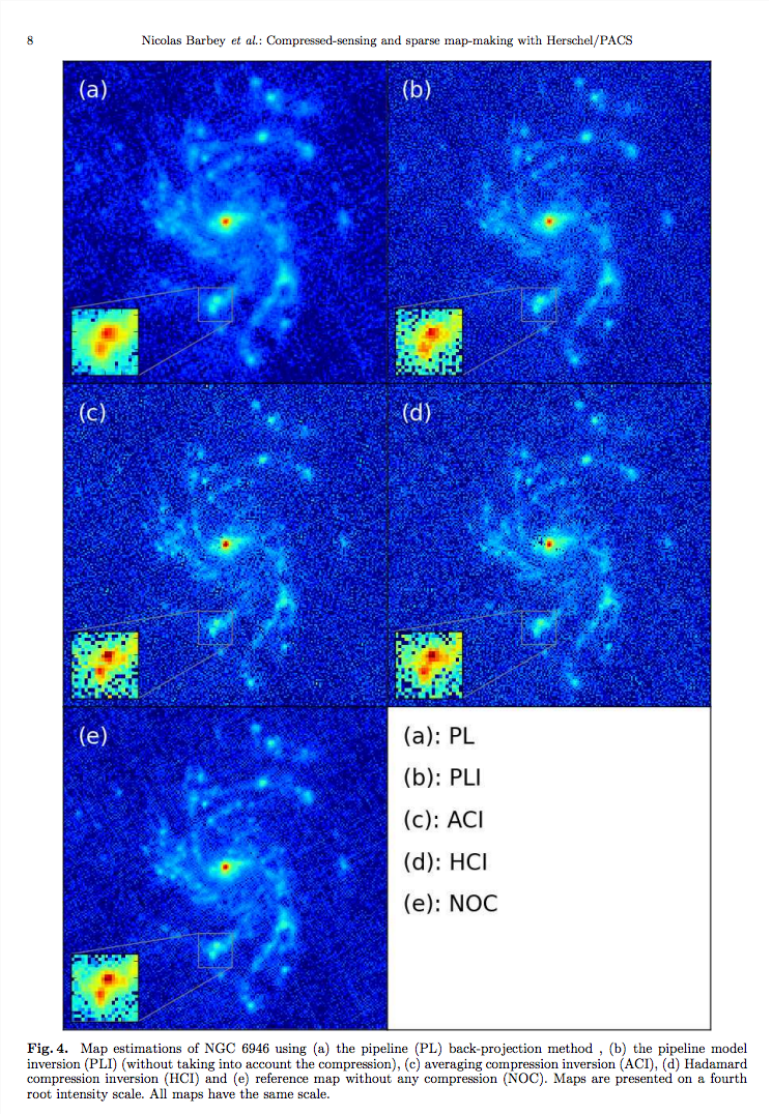


Fig. 4. Map estimations of NGC 6946 using (a) the pipeline (PL) back-projection method , (b) the pipeline model inversion (PLI) (without taking into account the compression), (c) averaging compression inversion (ACI), (d) Hadamard compression inversion (HCI) and (e) reference map without any compression (NOC). Maps are presented on a fourth root intensity scale. All maps have the same scale.

Section 5

Modelling linear flow

5. Linear flow estimate

-Back to the smoothed density slice.

-Assuming the data has changed over time

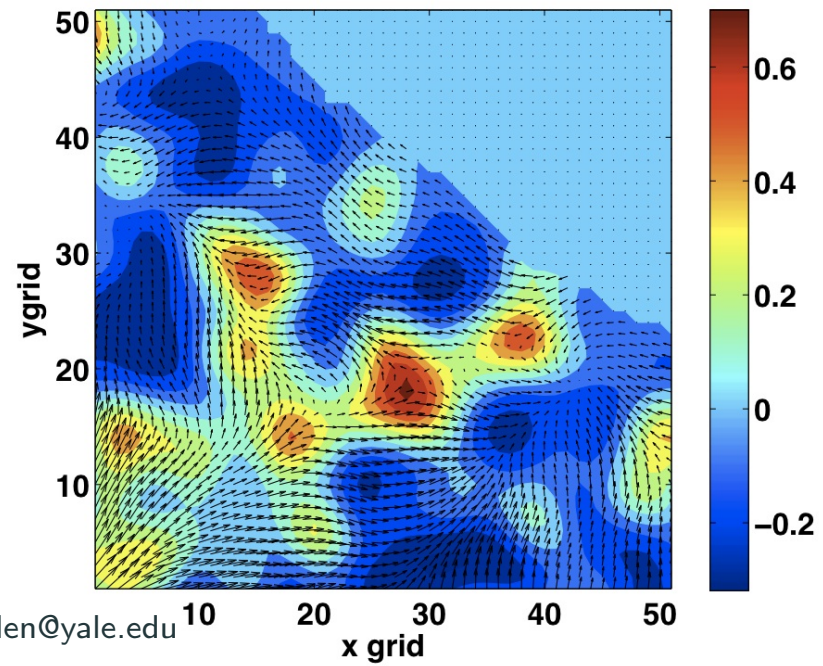
-Estimate the flow path of the data between some time and now

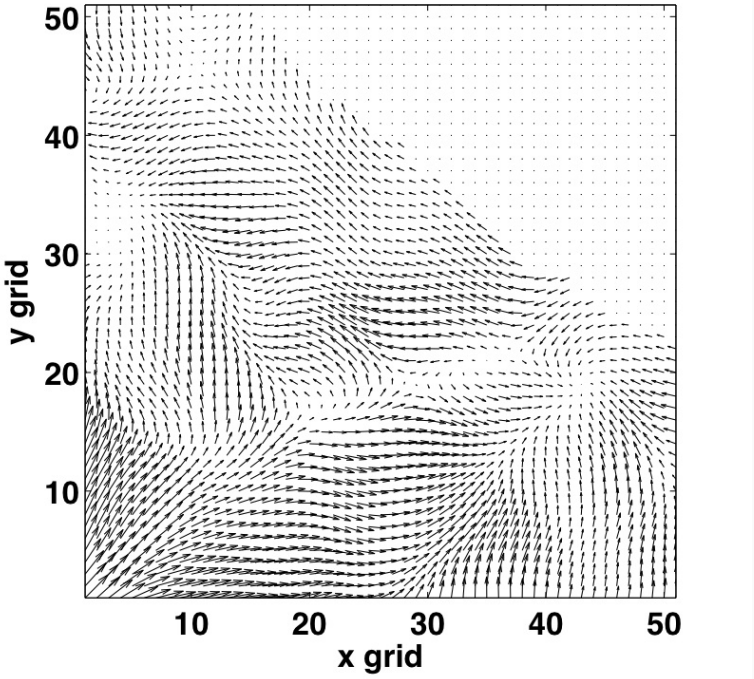
-This example, use the Poisson equation

$$\nabla^2 \phi = 4\pi G \rho$$

-combine with perturbation theory, estimate of the displacement

$$\nabla \cdot \Psi = -\delta$$





METHOD

Two standard methods used to calculate the displacement Ψ

1. Configuration space method

- Finite differences to build up a matrix A relating ϕ to $\delta \rightarrow$ $A\phi = \delta$
- Linear equation solver to get numerical values of ϕ
- Finite difference to get Ψ

(Padmanabhan et al. 06, Anderson et al.12,14, Burden et al 14)

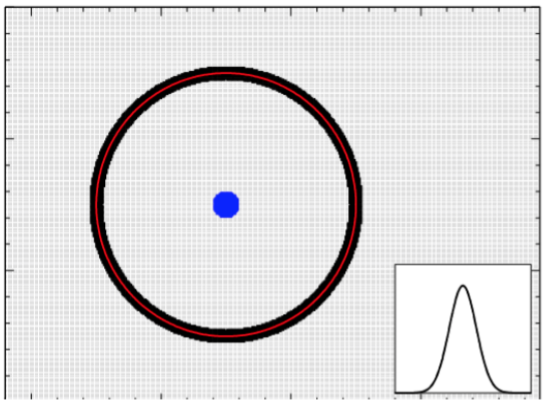
2. Fourier based method

- Use ……some physics……and Fast Fourier transforms to get Ψ from δ

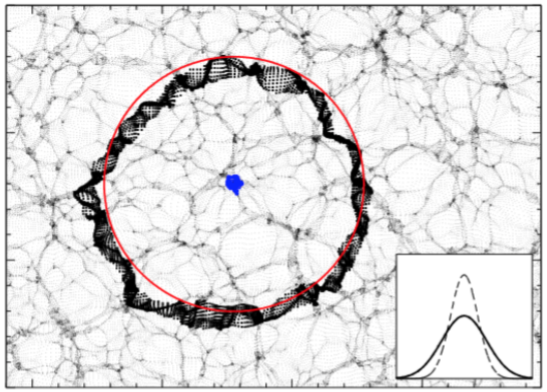
(Burden et al. 14,Tojeiro et.al 14, Ross et al. 14, Ross et al. 13)

(Burden et al. 15)

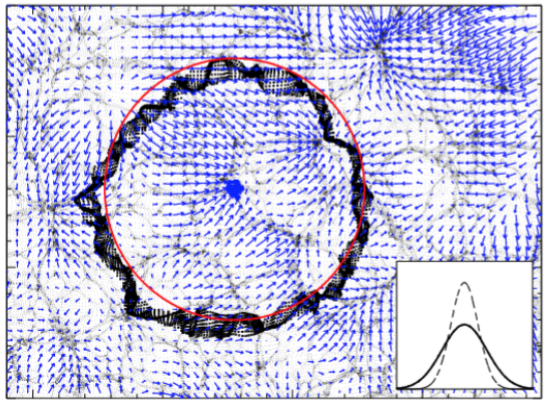
A simplified picture of how this is used to restore the signal



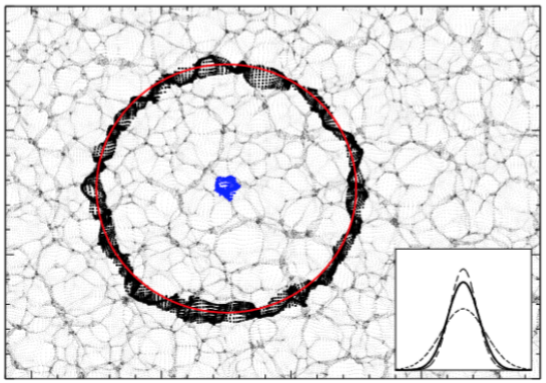
1.



2.



3.



4.

Figures from Padmanabhan et al. 2012

Padmanabhan et al. 2012

Section 6

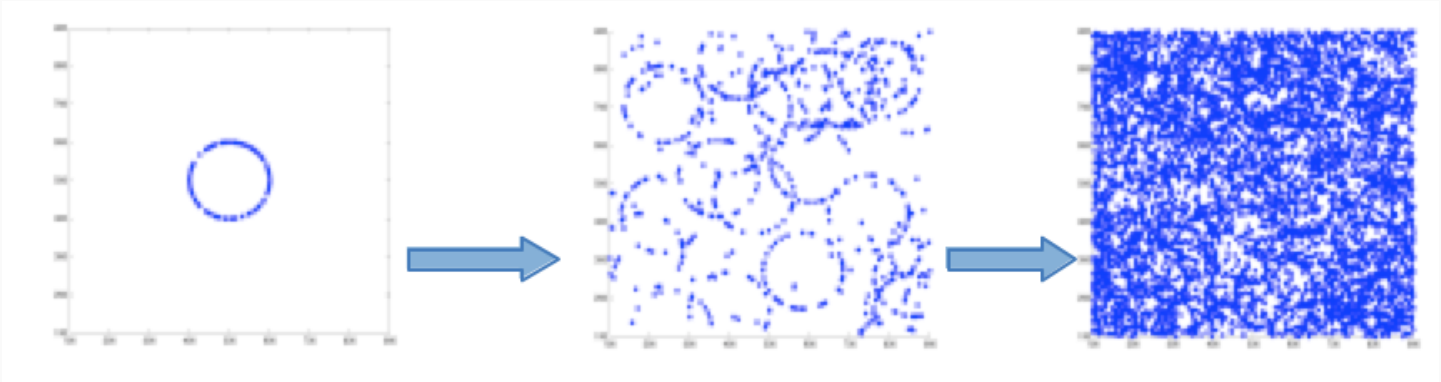
Data analysis techniques

searching for patterns

-Signal is weak but repetitive

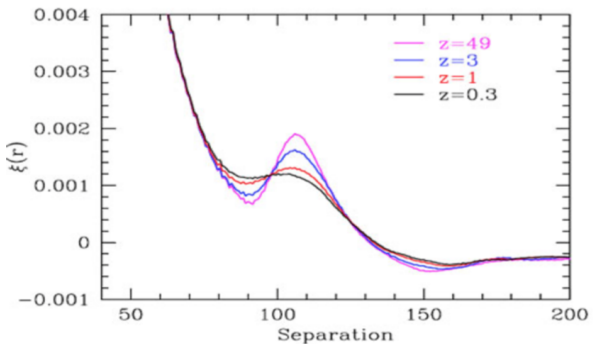
-Very large sets of data points (galaxies) covering large volumes -statistics.

-2D example of simple but effective method.



- count PAIRS of data points separated by x1, x2, x3 ....
- COMPARE to pairs of random data points at x1, x2...
- fit simple model to data, recover peak center and spread -radius of the circle.

- Extract signal statistically using



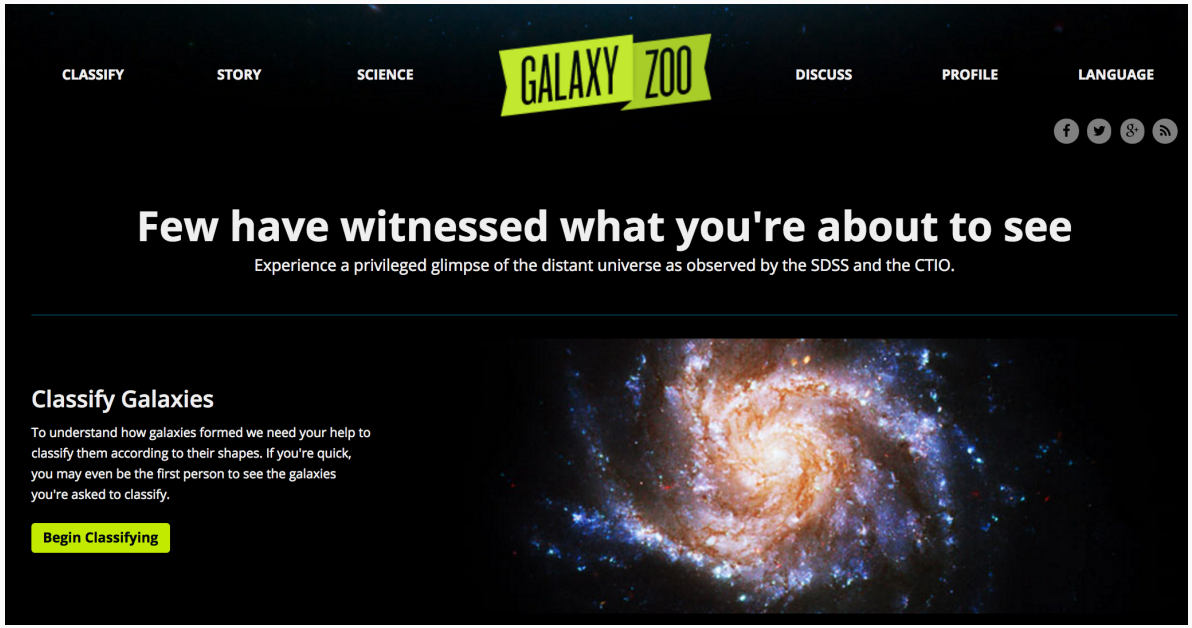
- 1. The correlation function $\xi(r)$

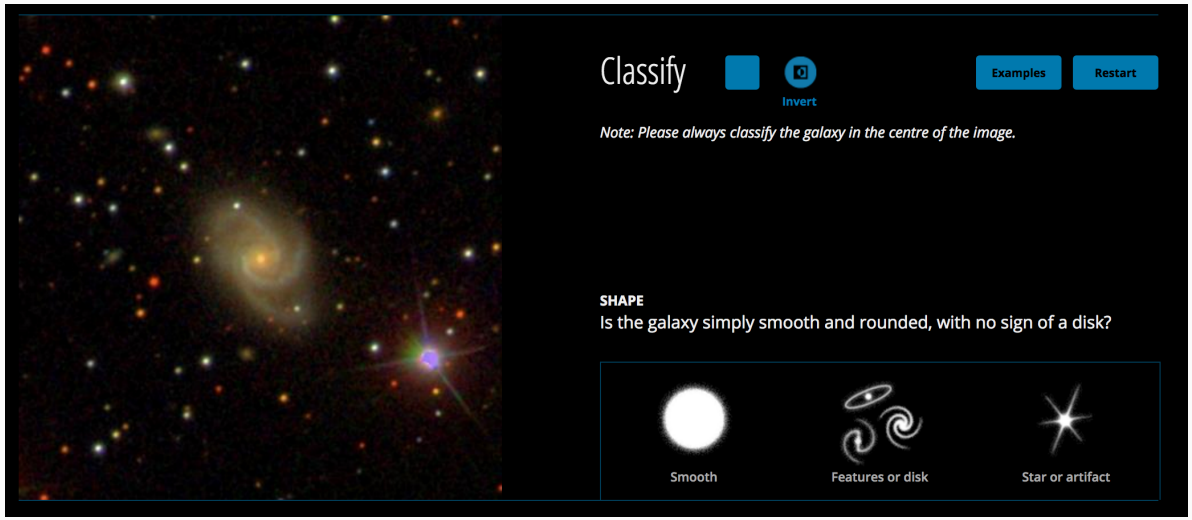
Section 7

Categorising data w/ machine learning

Data categorisation with decision tree

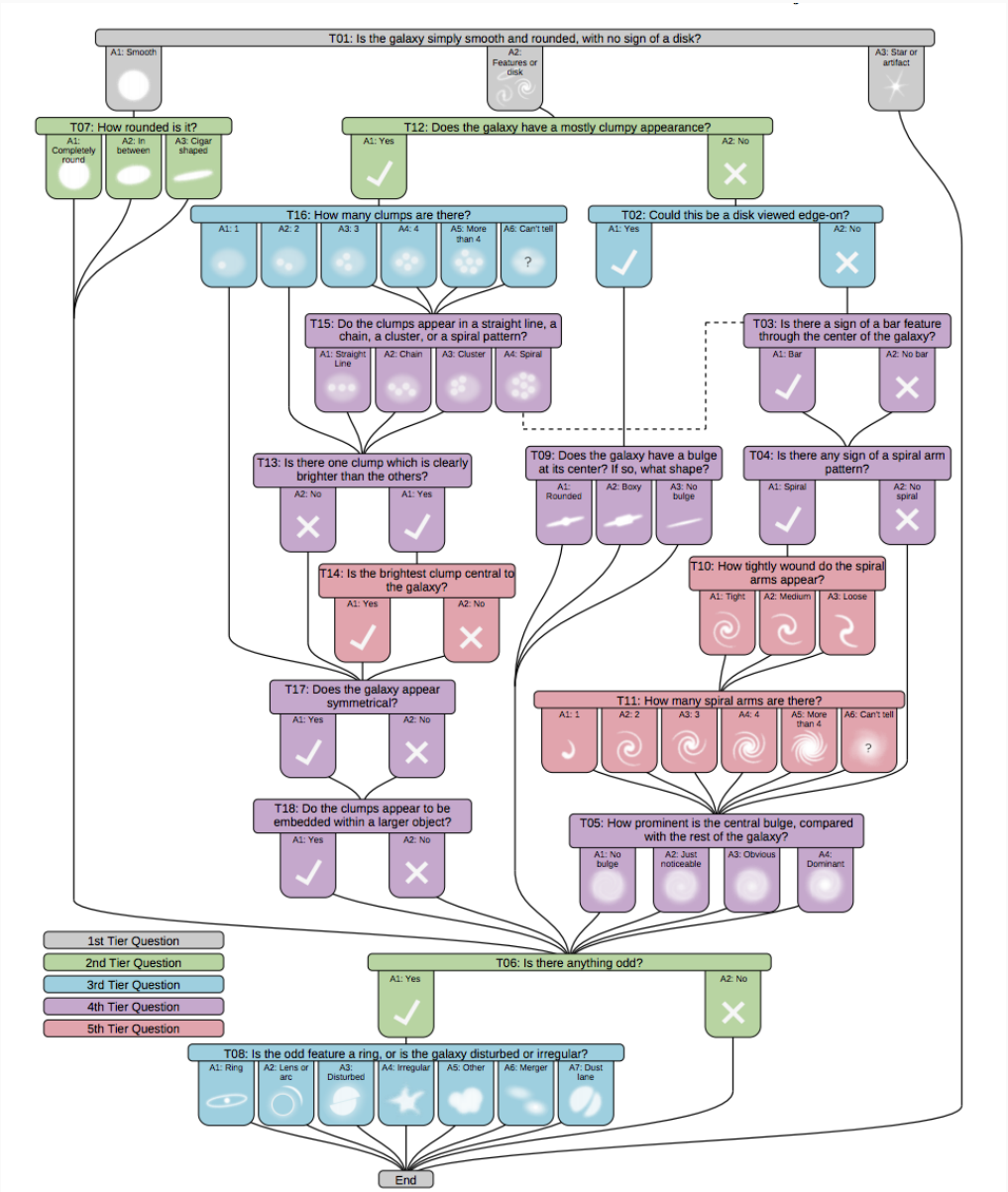
Galaxy zoo - citizen science, https://www.galaxyzoo.org/





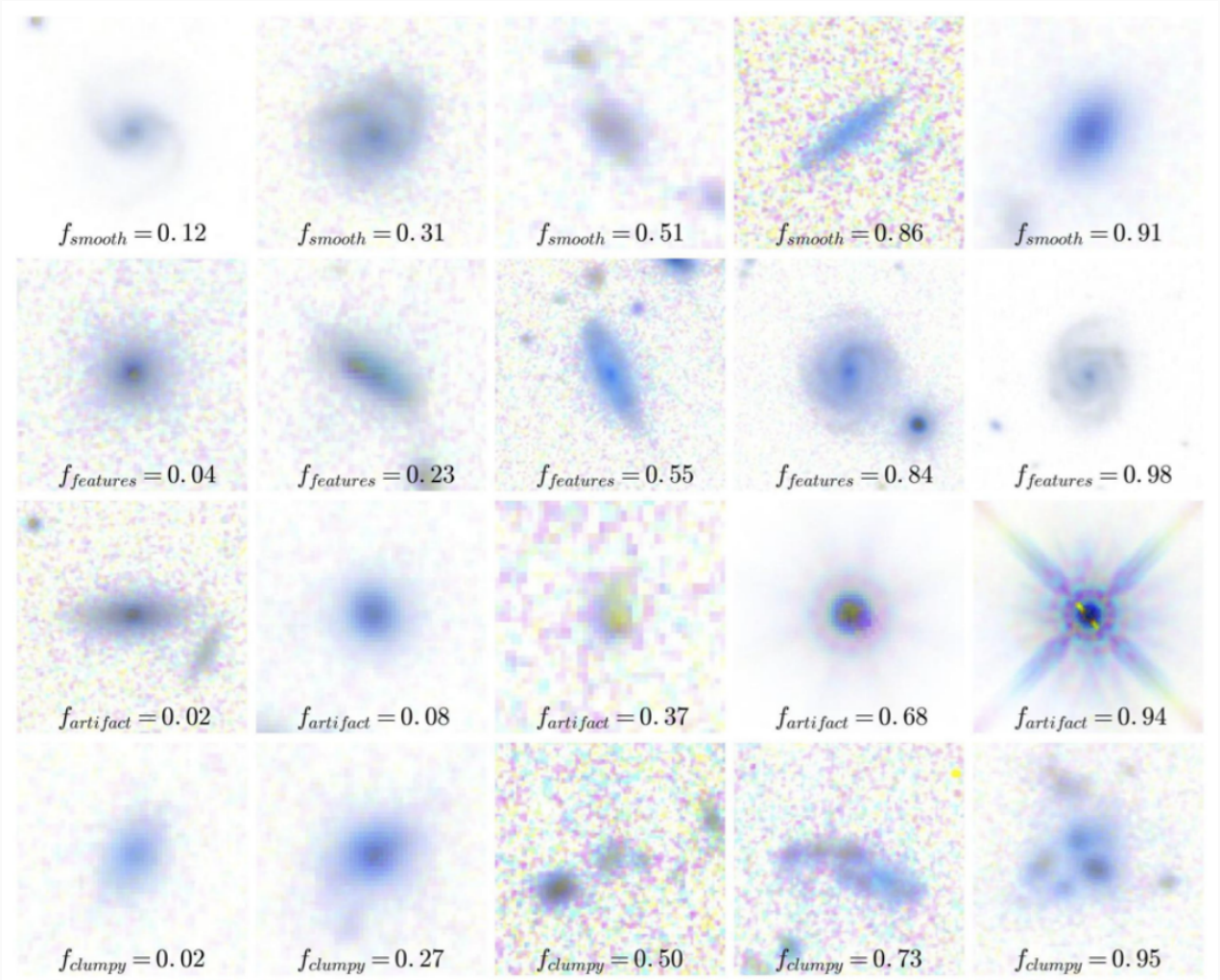
7. Data categorisation with decision tree

Decision tree - Willett et al. 2016 arXiv:1610.03068



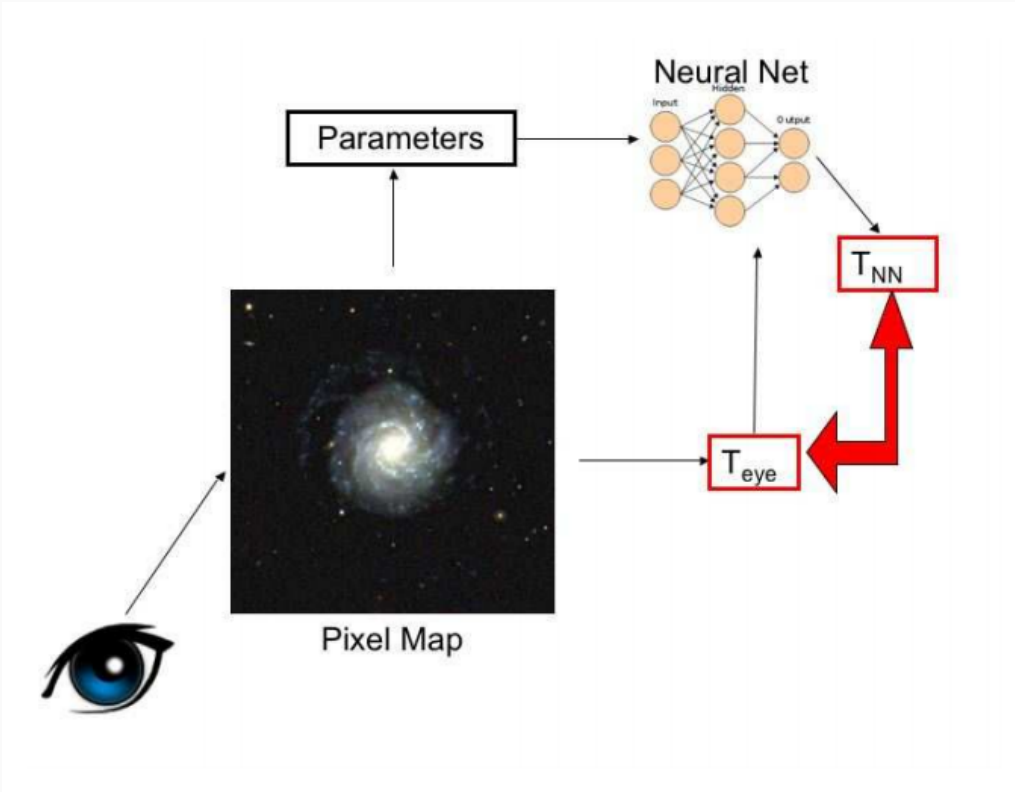
7. Data categorisation with decision tree

Classification results - Simmons et al. 2016 arXiv:1610.03070



- smoothness
- features
- artefacts
- clumpiness

7. Data categorisation with neural network



- set up network, **input, hidden layers, output**
- train the network to recognise things with known classification, i.e. data from above.
- feed it a new data set, it uses previous knowledge to predict the probability of the data belonging to a given class.

Banerji et al. 2009, arXiv:0908.2033 [astro-ph.CO]

Outline

- Introduction , who I am and why I am here (5 mins)
- Methods- 7 subsections (5 mins each)
  1. 2D to 3D, where do we point our telescope?
  2. Highlighting the signal.
  3. Removing noise and unwanted artefacts.
  4. Filling in data/ creating simulations.
  5. Modelling linear flow.
  6. Data analysis techniques.
  7. Categorising data w/ machine learning
- Summary

Summary

- Many techniques to optimise the information we gain from the data.
- Just a very simplified overview of a few.
- Share techniques and get inspiration from each others fields.
- If you see anything that might be relevant to your work let me know.
- Many publicly available codes!

Thank you

Questions?