STATISTICS

DAVID BACON

INSTITUTE OF COSMOLOGY AND GRAVITATION, PORTSMOUTH

ACKNOWLEDGEMENTS

Thanks to

DISCnet (Data Intensive Science Centre in SEPnet)
Jon Loveday (Sussex)
Enrico Scalas (Sussex)
Rob Crittenden (Portsmouth)

INSTITUTE OF COSMOLOGY AND GRAVITATION

Well-known centre for cosmology in the UK

Leading roles in big projects: Dark Energy Survey, LSST, Euclid, SKA…

M E

YOU

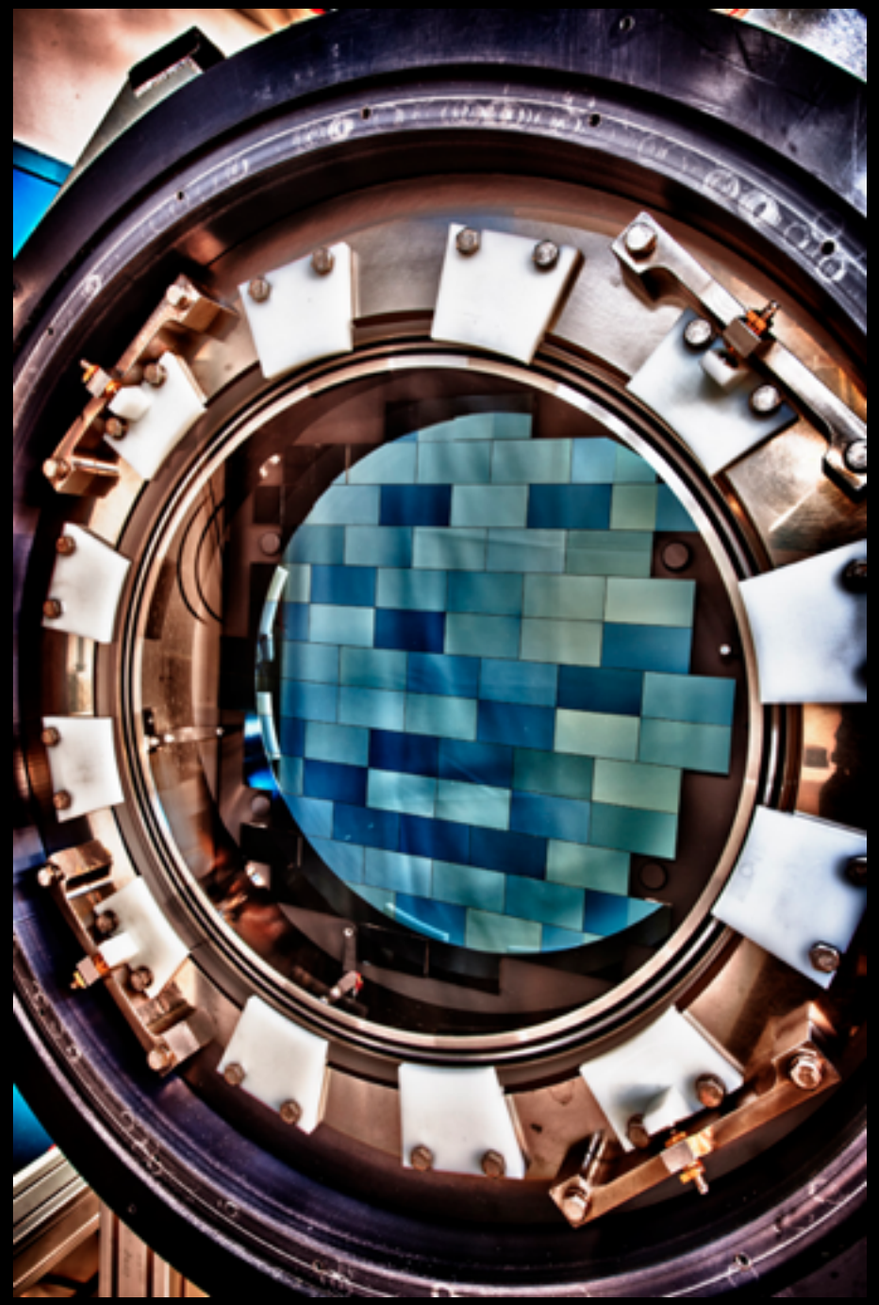
Get into groups of ~5 (preferably with people you don’t yet know well).

Introduce yourselves, and say a little about how the topic of statistics affects you.

What would your group like to get out of the day?

EXAMPLE OF USE OF STATISTICS IN MY FIELD: DARK ENERGY SURVEY

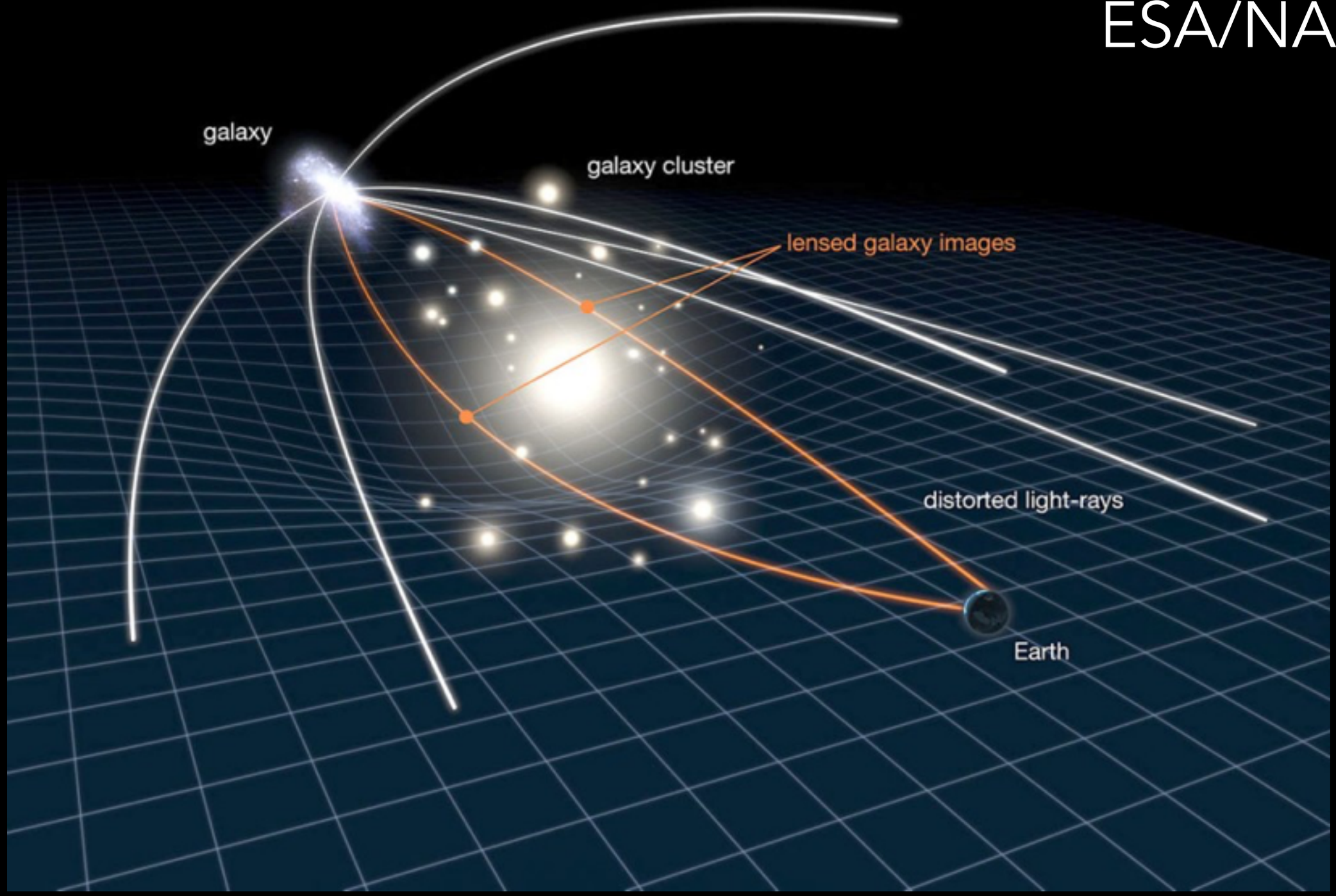
CTIO, Chile
5000 sq degree map, 8 billion ly deep



570Megapixels, 3 sq deg field
0.26” per pixel

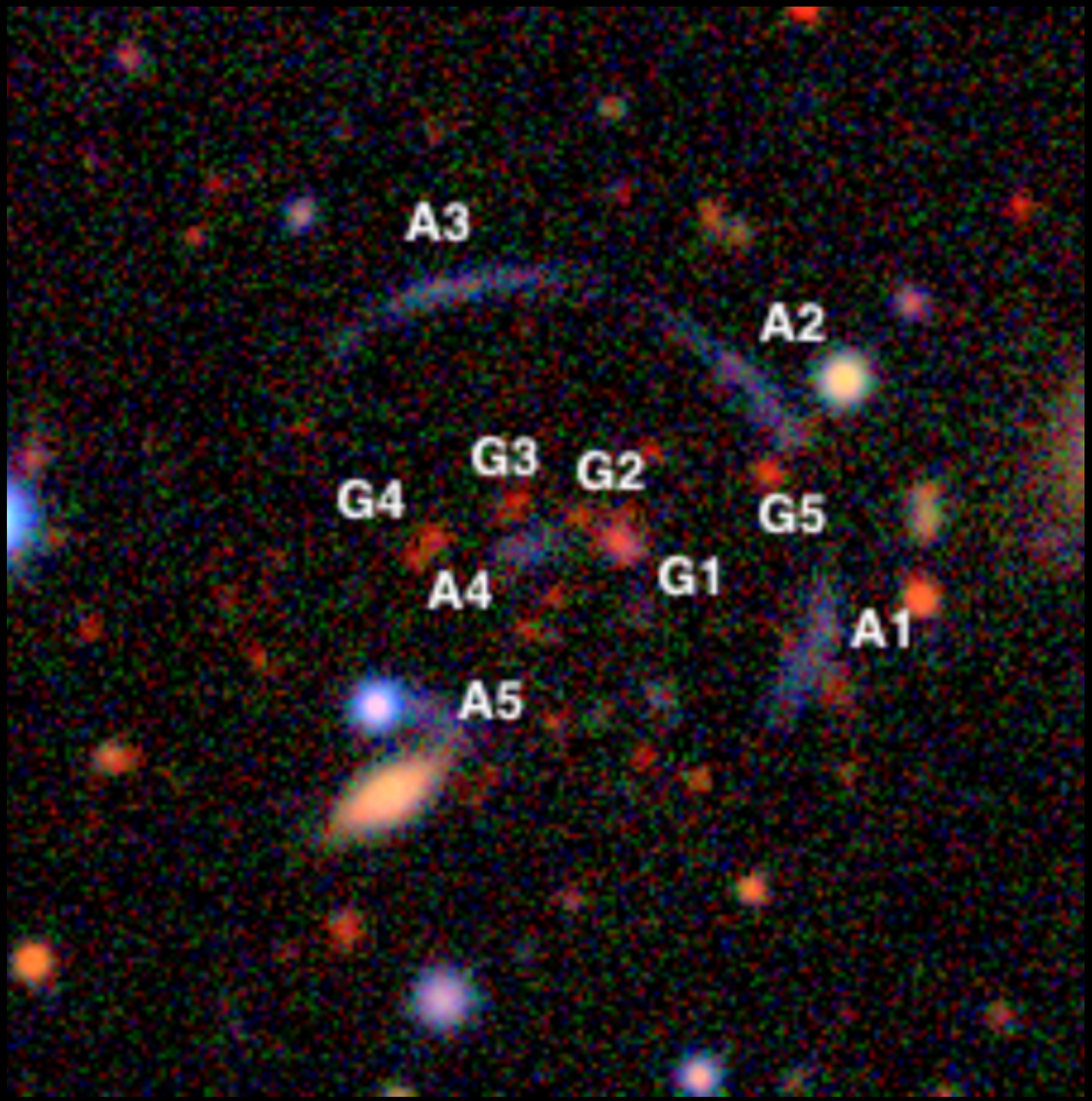
GRAVITATIONAL LENSING

ESA/NASA



Allows detailed ‘view’ of blobs of gravity

STRONG GRAVITATIONAL LENSING

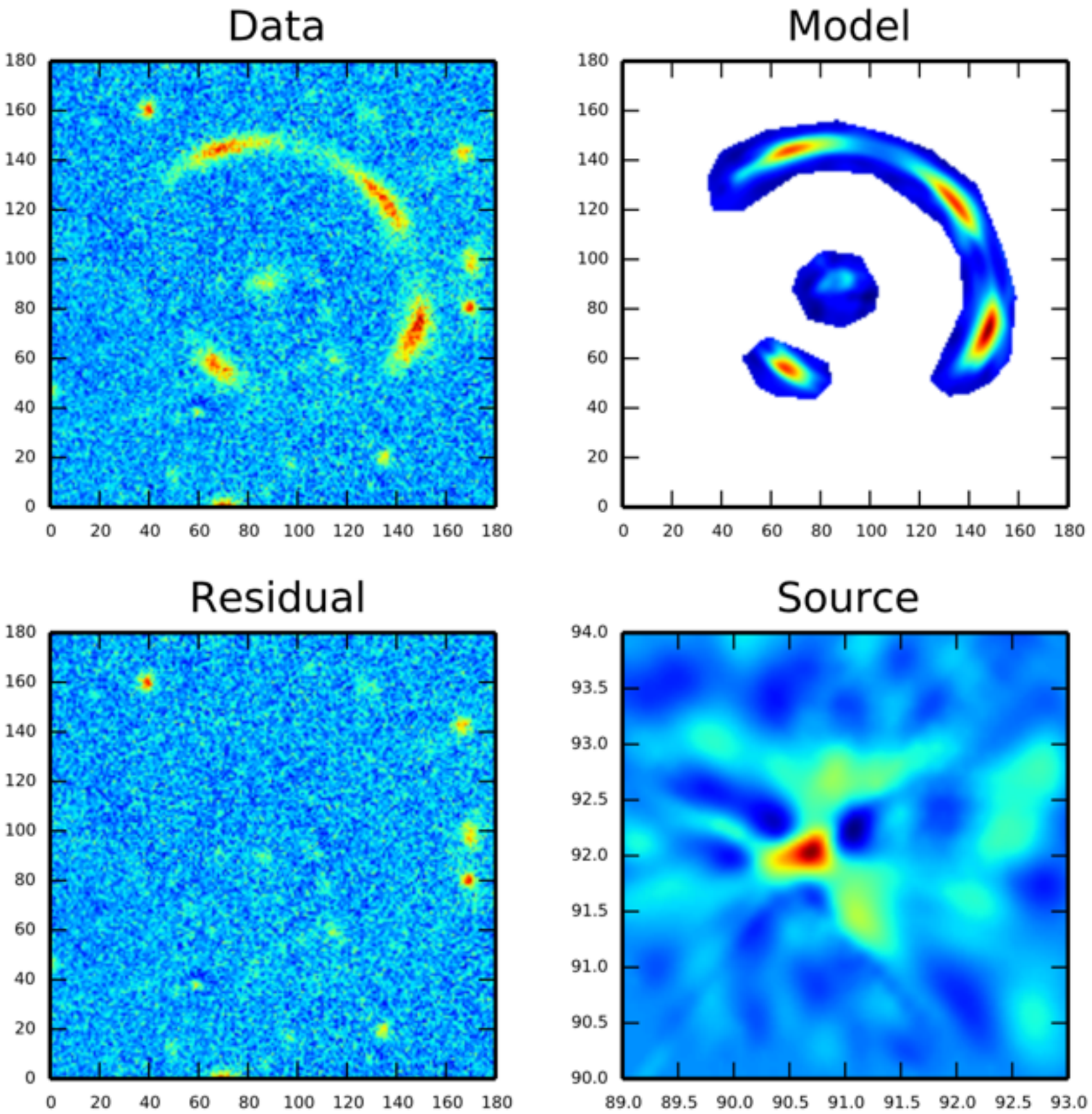


z_l=1.06

z_s=2.39

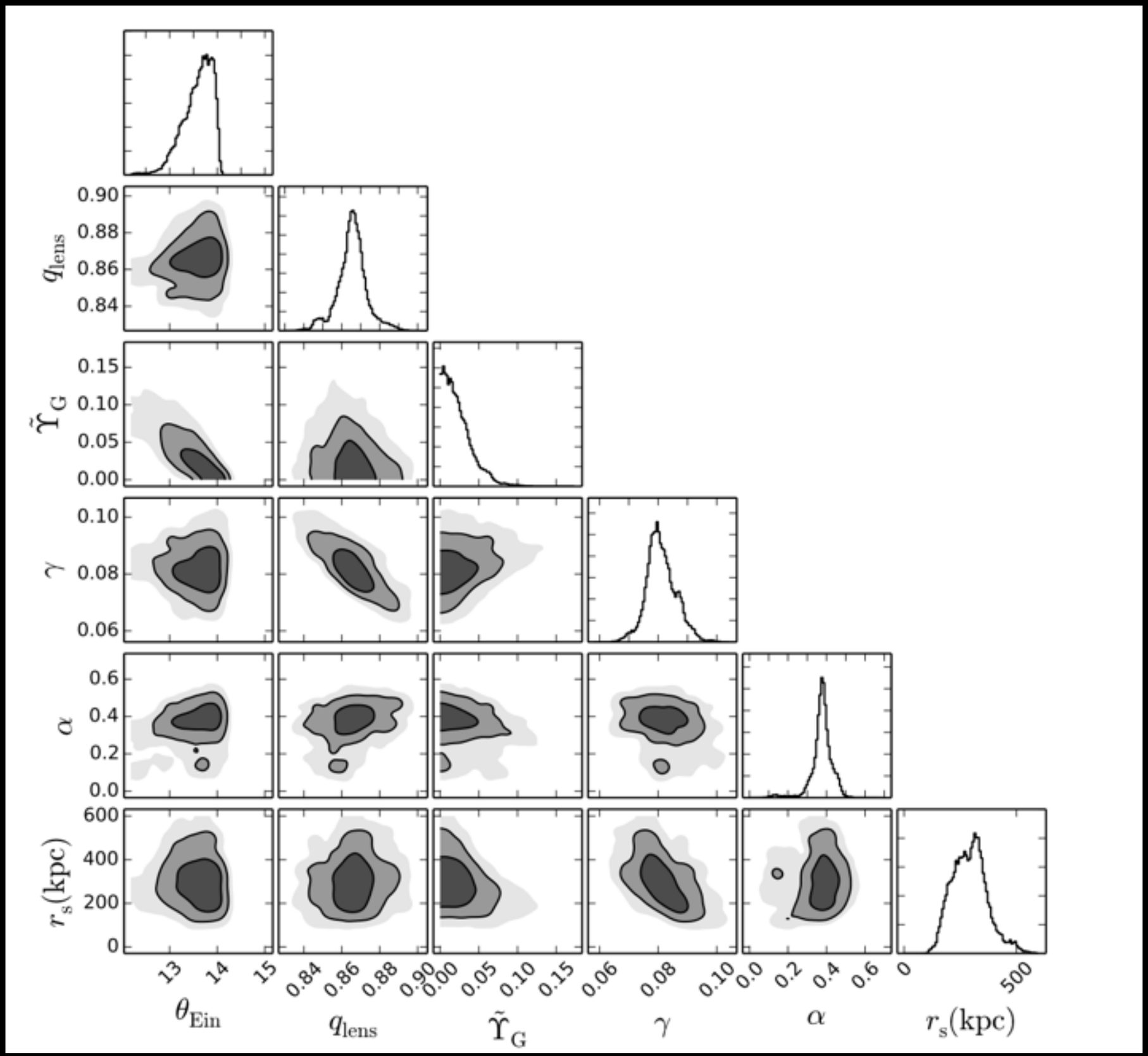
Collett et al 17

MODELLING



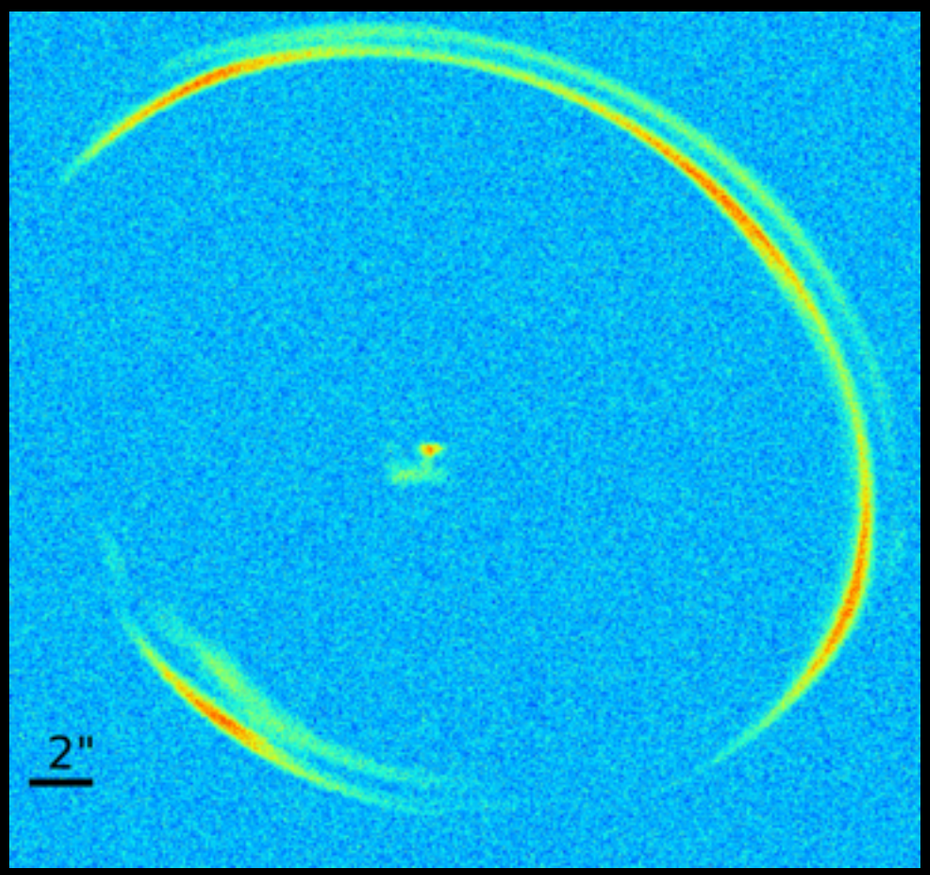
Collett et al 17

MODEL PARAMETERS

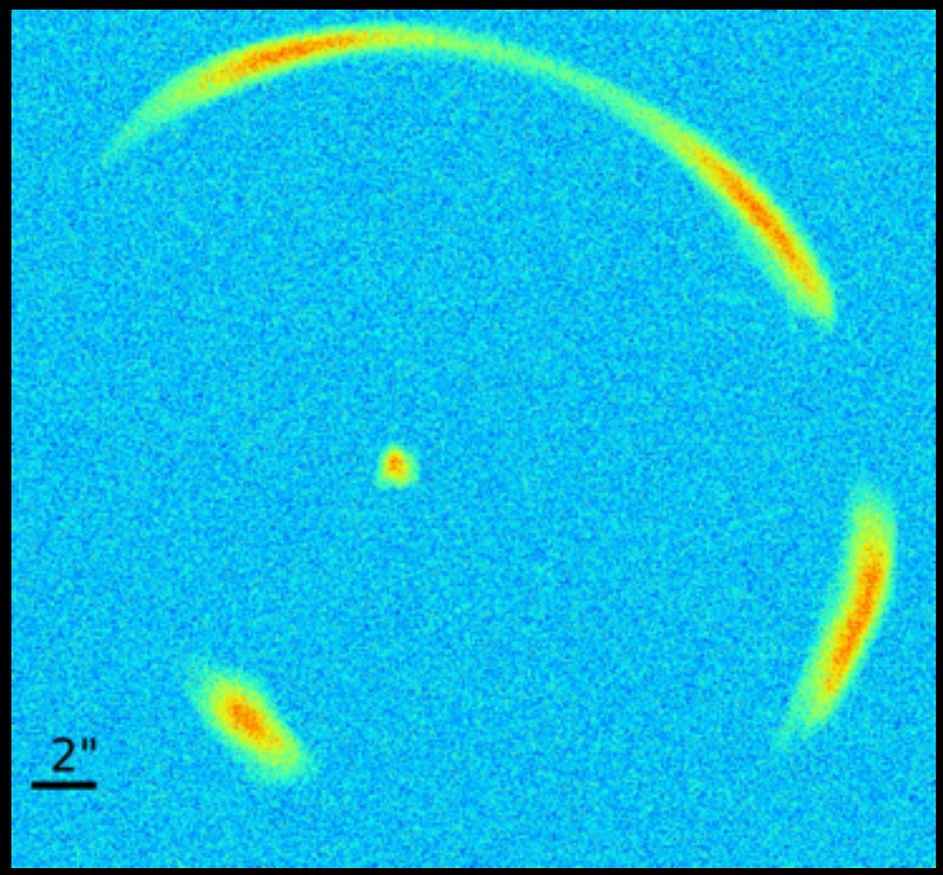


Collett et al 17

PREDICTIONS



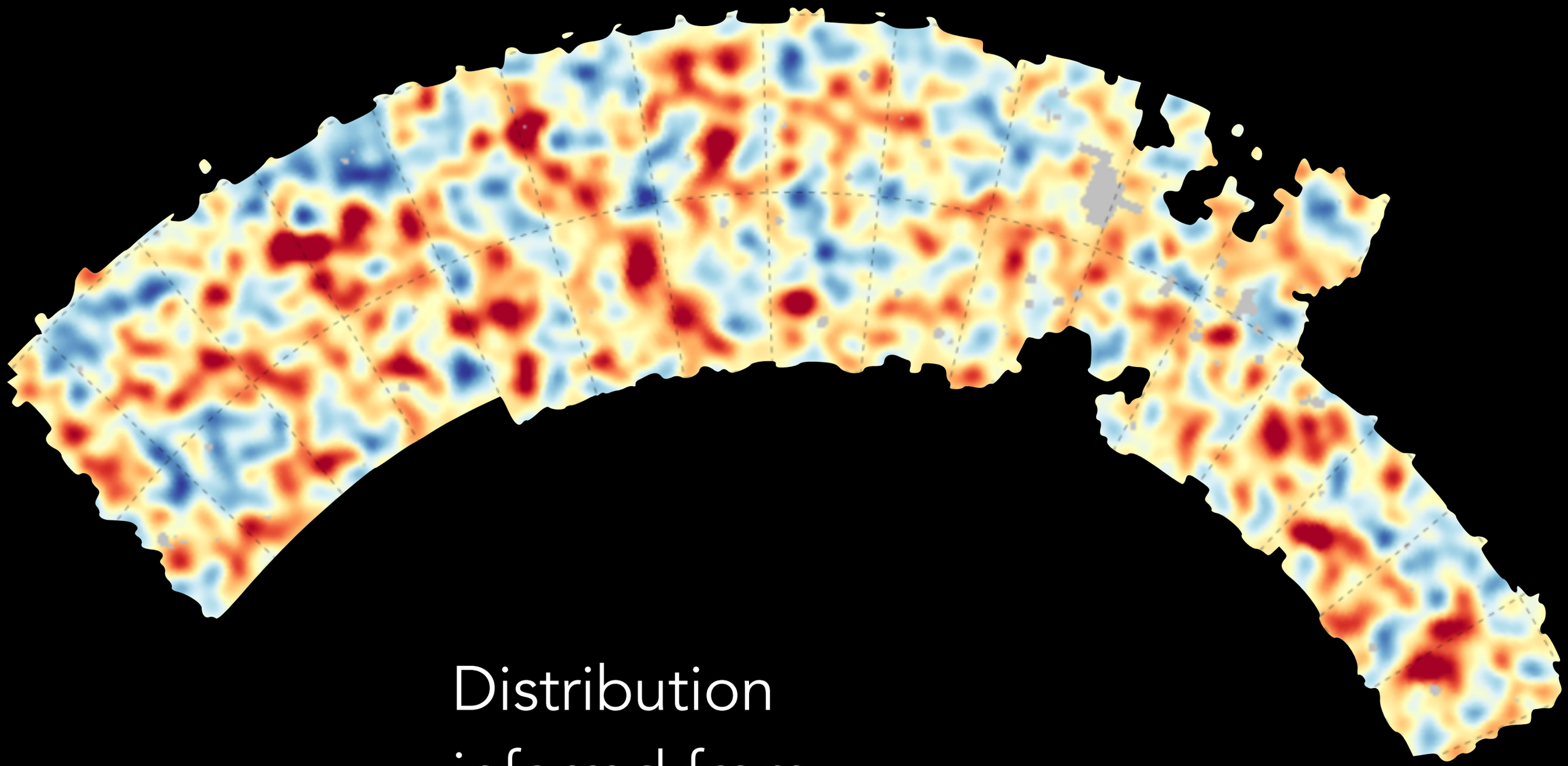
1 blob of matter



2 blobs of matter

Collett et al 17

LENSING TO MAKE A MAP

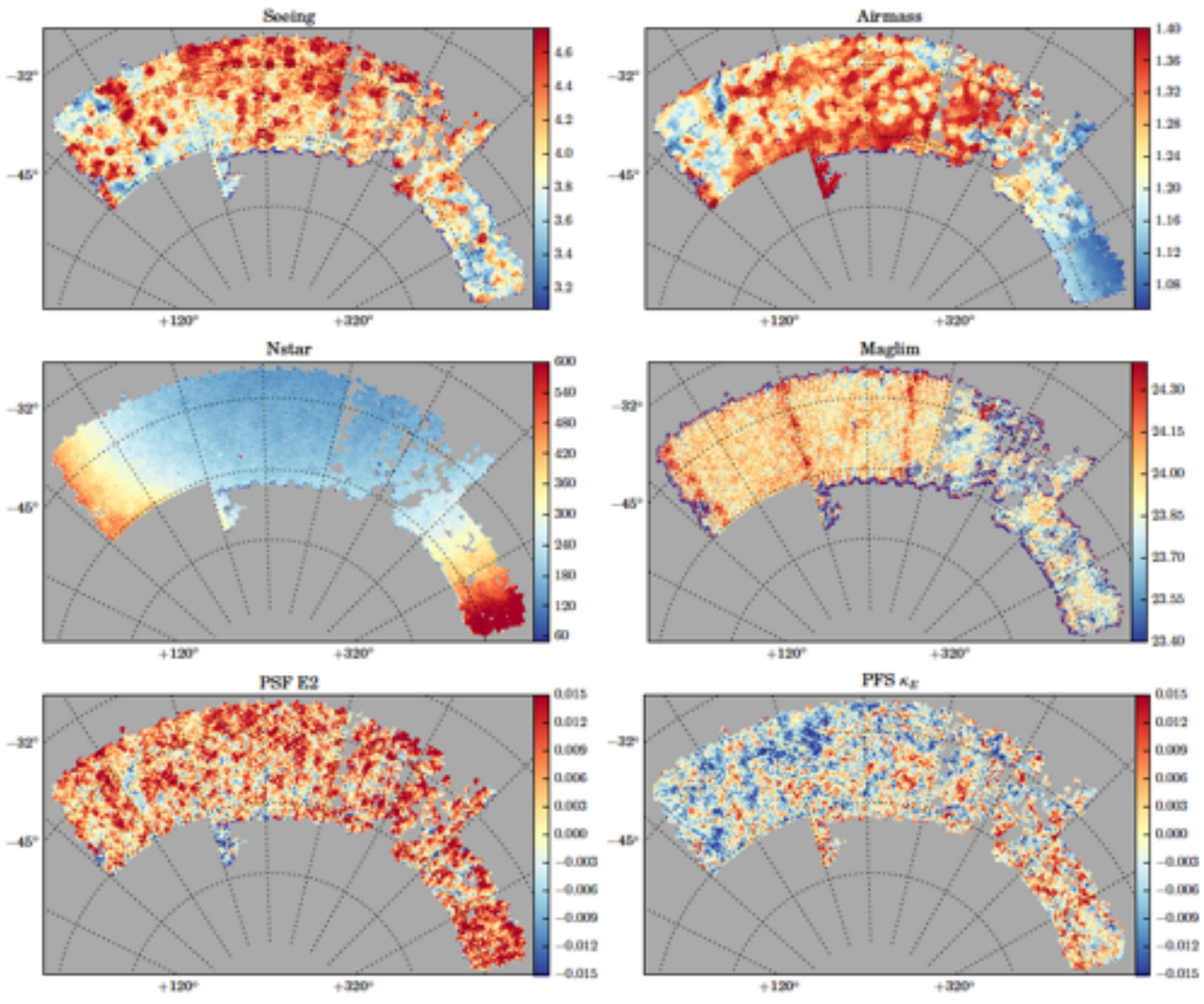


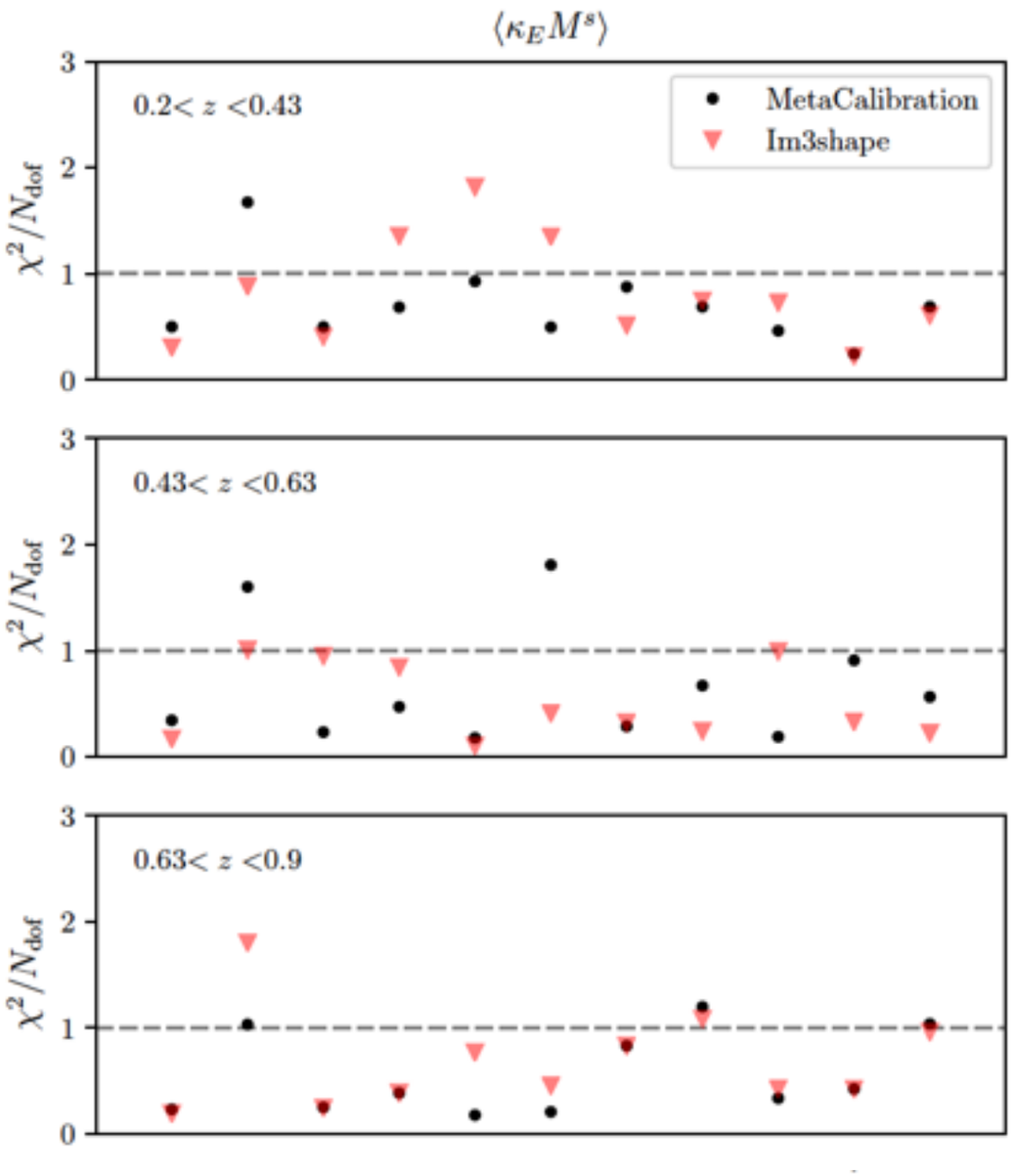
Distribution
inferred from
30M galaxies

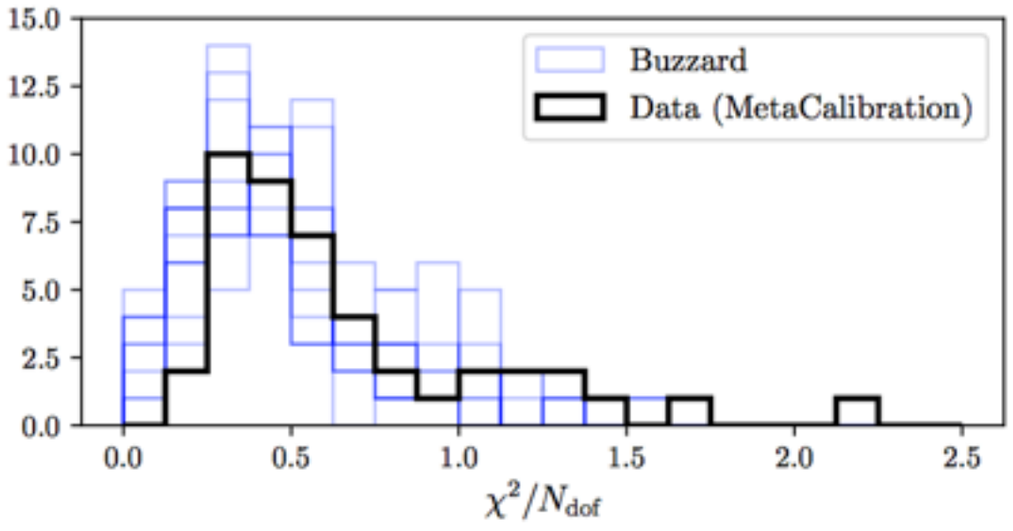
Chang et al 17

Wide-Field Mass Maps

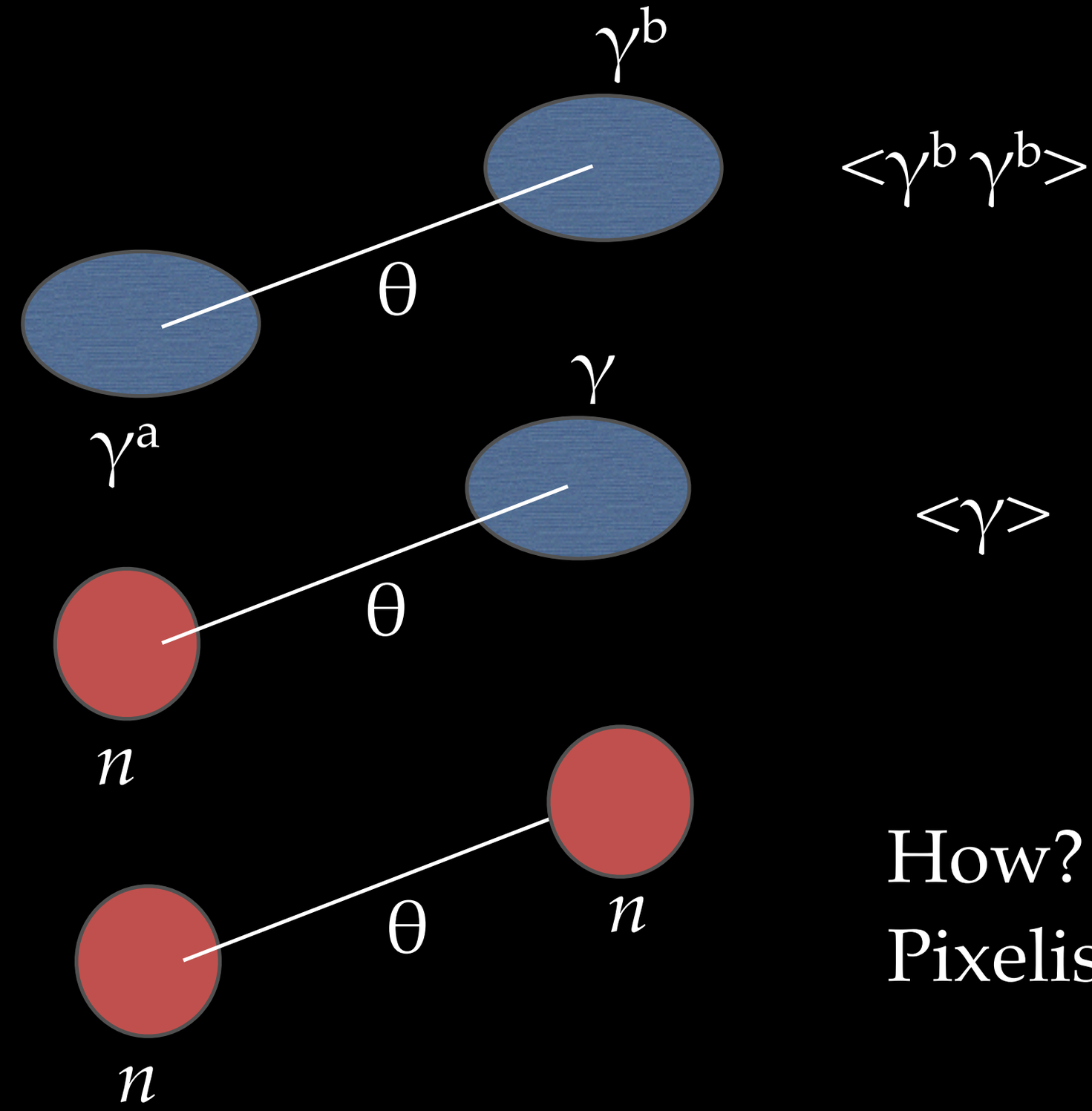
Cross-correlation with systematics maps







TWO POINT STATISTICS

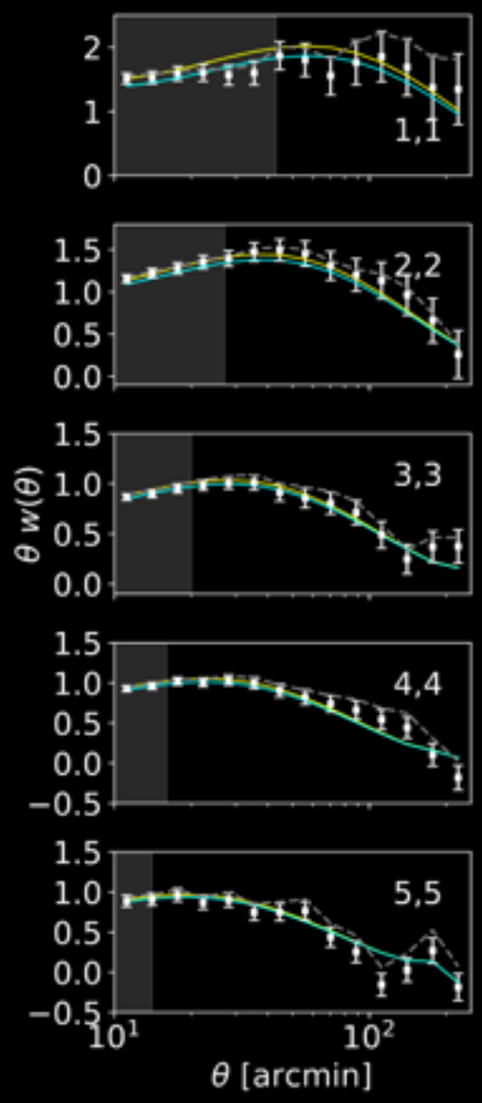


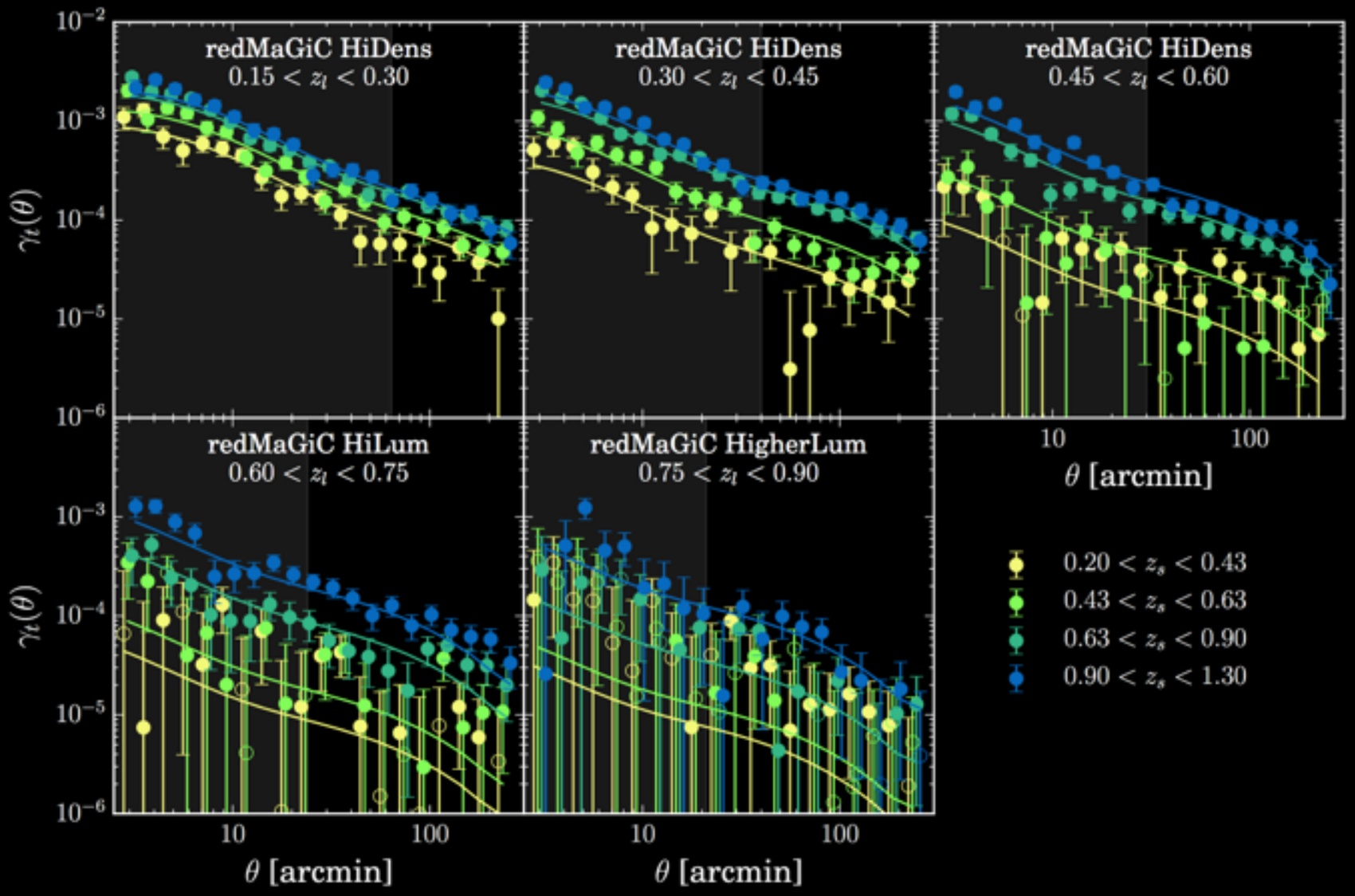
$<\gamma^b \gamma^b>$

$<\gamma>$

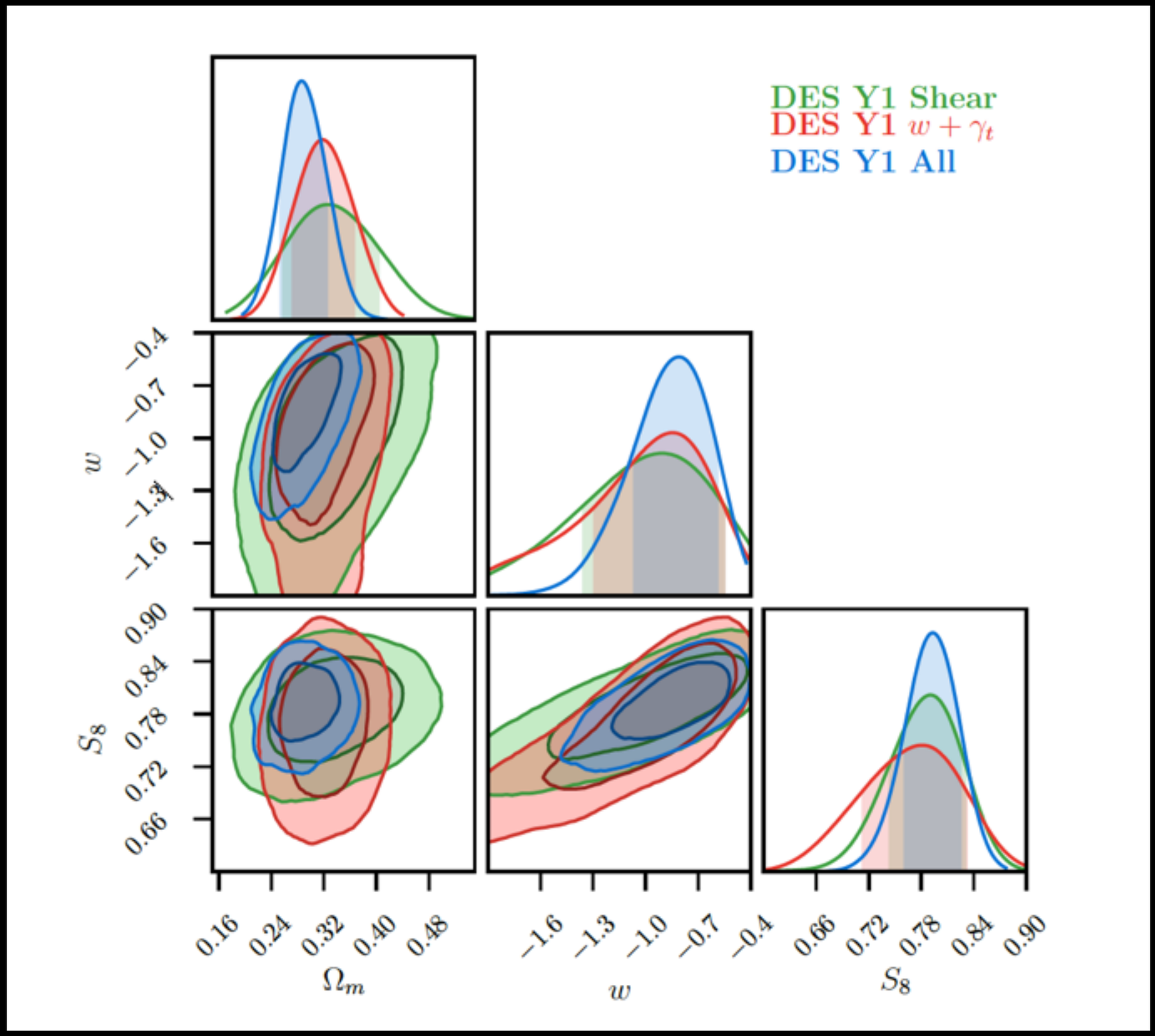
How?
Pixelise?

TWO POINT STATISTICS: RESULTS





FIT PARAMETERISED MODELS TO THESE - HOW?



DES 17

APPROACH TODAY

We’re a diverse group today - in terms of our academic disciplines and statistics background.

My aim is to be useful to all of you, whether you consider yourself an expert or less so!

So minimally: when I introduce an idea, make a note to discover more yourself at your own rate. Maximally, understand everything and have a party!

The tutorials are for you to experiment and explore.

FREQUENTIST VS BAYESIAN STATISTICS

There's an ongoing debate about how to approach statistics!

Probability as the fraction of times something will happen given repeated trials? So make estimators from samples to infer underlying properties (Frequentist)

What level of belief should we have in a particular model for this data? (Bayesian)

We won't be too strict about this, but will tend towards being Bayesian.

PROBABILITY DENSITY FUNCTION

$$\mathcal{P}(x \leq X \leq x + dx) = f_X(x)dx$$

PDF

Cumulative distribution function:

$$\mathcal{P}(X \leq x) = F_X(x) \equiv \int_{-\infty}^{x} f_X(x')dx'.$$

RULES OF PROBABILITY

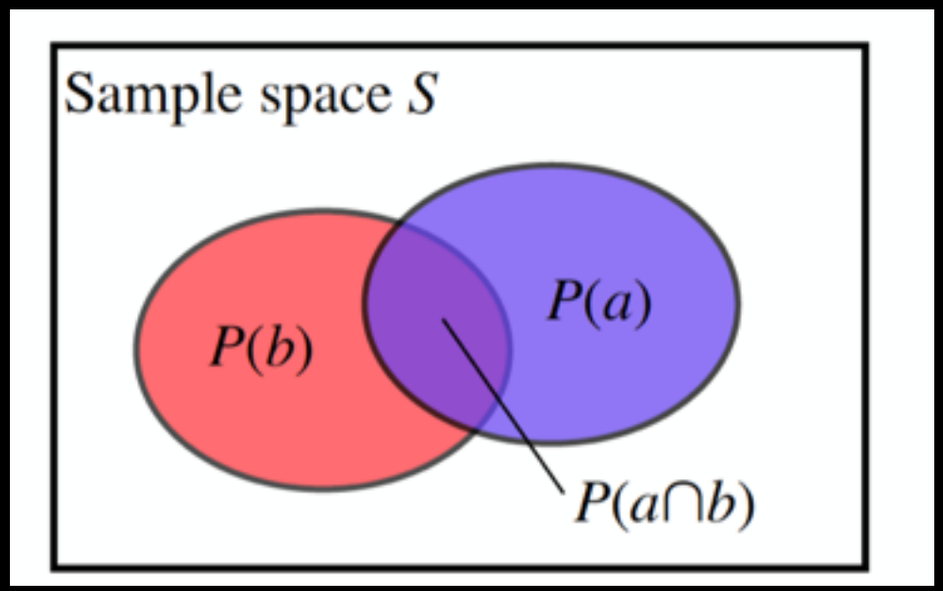
Probabilities of all possible outcomes add up to one:

$$\sum_i \mathcal{P}_X(x_i) = 1$$

$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

Conditional probability:

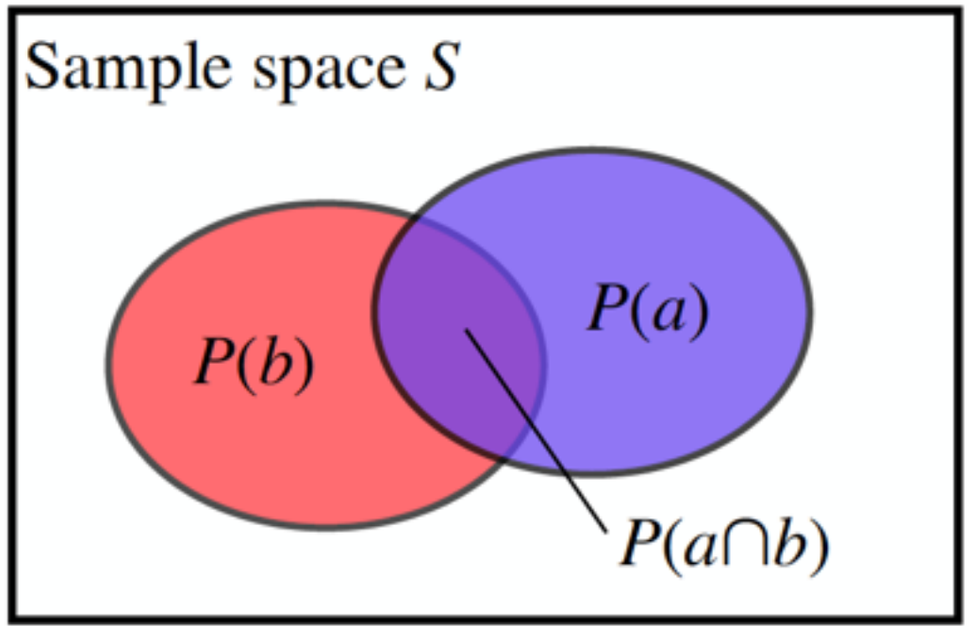
$$\mathcal{P}(X|Y) = \mathcal{P}(X,Y)/\mathcal{P}(Y).$$



Bayes' theorem (Thomas Bayes, 1763)

- Conditional probability $P(a \mid b)$ = probability of a given that b is true
- Bayes' theorem:

$$P(a \cap b) = P(a \mid b)P(b) = P(b \mid a)P(a)$$



- Note: If and only if a and b independent:

$$P(a \cap b) = P(a)P(b) \quad \text{and so} \quad P(a \mid b) = P(a)$$

Bayes' theorem

$$P(a \cap b) = P(a \mid b)P(b) = P(b \mid a)P(a)$$

- Useful to rearrange:

$$P(a|b) = \frac{P(b|a)P(a)}{P(b)} \qquad (1)$$

- Also very useful to rewrite $P(b)$ to state explicitly that it is the probability of b whether or not a is true

$$P(b) = P(b|a)P(a) + P(b|\bar{a})P(\bar{a}) \qquad (2)$$

and remember that

$$P(\bar{a}) = 1 - P(a)$$

Bayes' theorem

- Combining (1) and (2) from previous slide, Bayes' theorem is often used in the format

$$P(a|b) = \frac{P(b|a)P(a)}{P(b|a)P(a) + P(b|\bar{a})P(\bar{a})}$$

- Bayes' theorem provides a consistent way of incorporating prior knowledge of parameter values into their estimation from new data
- Subjectivity comes in deciding what prior knowledge to assume

Example: TV game show

- Three doors; £10,000 prize behind one of them
- You guess which door (but do not open it)
- Game-show host opens one of the other doors with no money behind it
- Offer: stick with your initial guess, or, for £10, choose 3rd door
- Should you?

Example: TV game show

- Initial probability that prize is behind any given door A, B or C is 1/3
- Let's say you choose A and host opens C
- Let $P(X)$ be probability that prize is behind door X

$$P(A \mid \text{host opens } C)$$

$$= \frac{p(\text{h. o. } C \mid A)P(A)}{P(\text{h. o. } C \mid A)P(A) + P(\text{h. o. } C \mid B)P(B) + P(\text{h. o. } C \mid C)P(C)}$$

$$= \frac{1/2 \times 1/3}{1/2 \times 1/3 + 1 \times 1/3 + 0 \times 1/3} = 1/3$$

- So the probability that the prize is behind door B is now 2/3 → clearly worth £10 to double your chance to win!
  - If you're not convinced, then consider the case of 10,000 doors, where after you choose a door, the host opens 9,998 doors!

Theory testing

- Subjective probability pushes Bayes' theorem further by applying it to statements that are "unscientific" in the frequentist definition

Subjective "degree of belief" or *prior*.
Can be modified by subsequent experimental evidence

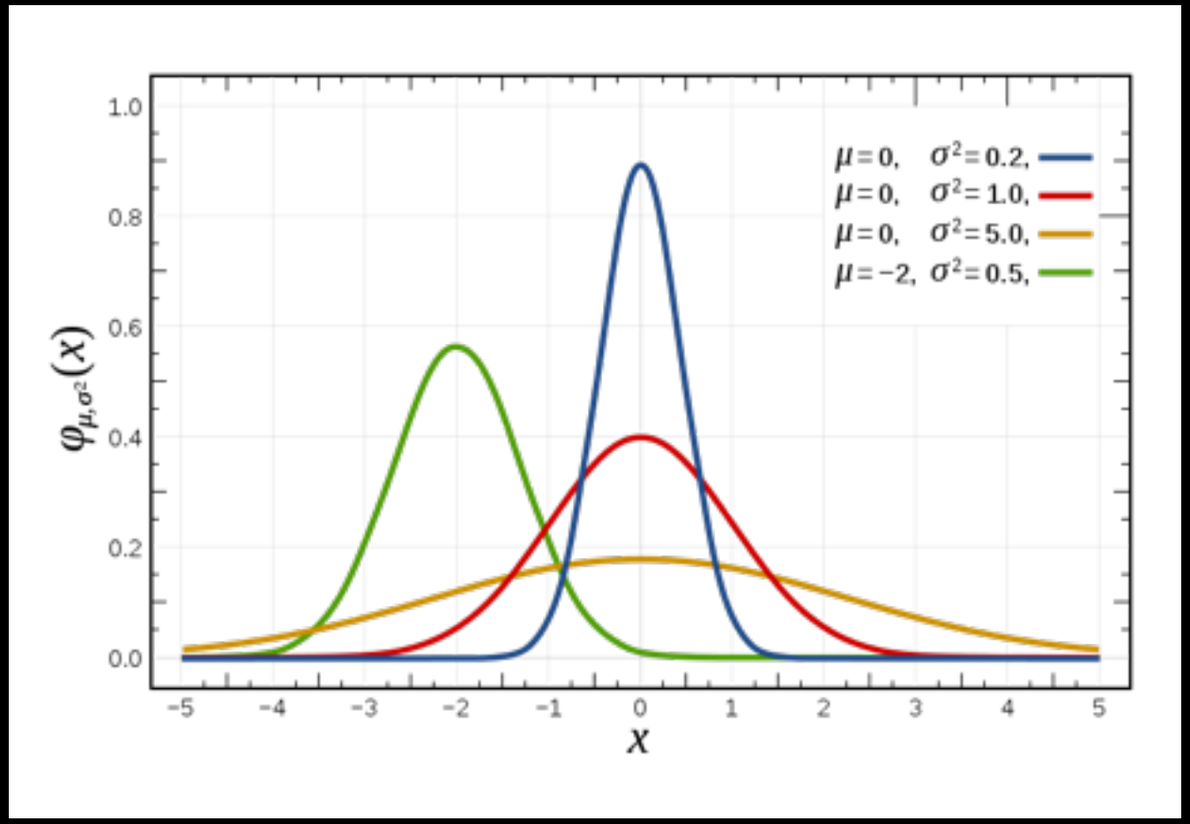
$$P(\text{theory}\,|\,\text{result}) = \frac{p(\text{result}\,|\,\text{theory})\,P(\text{theory})}{P(\text{result})}$$

A result that is likely to happen for other reasons does not provide strong support for a theory that predicts it

WE WILL NEED:
GAUSSIAN (NORMAL) DISTRIBUTION

$$f_X(x) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} e^{-\frac{1}{2}(x-\bar{X})^2/\sigma^2}$$

mean $\bar{X}$ and variance σ^2



Example: the mass of the electron

- You measure the mass of the electron as $m_e = 520 \pm 10$ keV
- You may say "there is a 68% probability that the value lies between 510 and 530 keV"
- In frequency interpretation this makes no sense: electron has just one mass — 511 keV
  - Either it is within your error bars or it isn't

- Assuming normal distribution of measurements around the true (but unknown) mass m_e, expect

$$P(\text{result}|\text{theory}) = P(m|m_e) \propto \exp[-(m - m_e)^2/2\sigma^2] \qquad (1)$$

- Bayes' theorem:

$$P(\text{theory}|\text{result}) = \frac{P(\text{result}|\text{theory})P(\text{theory})}{P(\text{result})}$$

Example: the mass of the electron

- For uniform prior, can turn (1) around to give

$$P(\text{theory}|\text{result}) = P(m_e|m) \propto \exp[-(m - m_e)^2/2\sigma^2]$$

  - NB Based on the initial distribution $P(\text{theory}) = P(m_e)$, the prior, being uniform — assume all possible values of m_e are equally likely
    - $P(\text{result})$ taken care of in overall normalisation
    - Assume instead a non-uniform prior, e.g. $P(m_e^2) = \text{const}$, then will get a different result

- Safer interpretation:
  - If the true value m_e is as high (low) as 530 (510) keV, there is a 68% probability of a measurement as low (high) as that obtained

Measurement of constrained quantities

- Weigh 0.1 g of powder on a balance with 0.2 g standard deviation
- Expect measurement ± 0.4 g of true value 95.4% of the time (2σ confidence)
- Some possible outcomes:
  - 2.3% probability that we measure a value > 0.5 g
    - e.g. (0.6 ± 0.4)g [2σ limits exclude true value]
  - Measurement in range 0.4 g to 0.5 g
    - Limits include true value
- Measurement is, say, 0.3 g
  - Our limits are –0.1 g and 0.7 g
  - Based on common sense, we modify lower limit to 0.0 g
- Measured value is –0.39 g
  - Limits are –0.79 g and 0.01 g
  - If we modify –0.79 g to 0.0 g, we'd get a range 0.0 g to 0.01 g with 95.4% confidence — clearly incorrect

Measurement of constrained quantities

- Solution: use Bayesian statistics
  - P(result | theory) = P(measurement x arising from true value X) = Gaussian of standard deviation σ
  - P(result) taken care of in overall normalisation
    - P(theory) = intrinsic probability distribution (prior) for X
- First assume uniform prior $P(X)$ = const

$$P(\text{theory}|\text{result}) \equiv P(X|x) = \frac{\exp[-(x-X)^2/2\sigma^2]}{\sigma\sqrt{2\pi}}$$

Measurement of constrained quantities

- Now incorporate additional knowledge that $X > 0$:
  $P(X)$ a step function: zero for $X < 0$ and const for $X > 0$
- Denominator in Bayes' theorem can be written

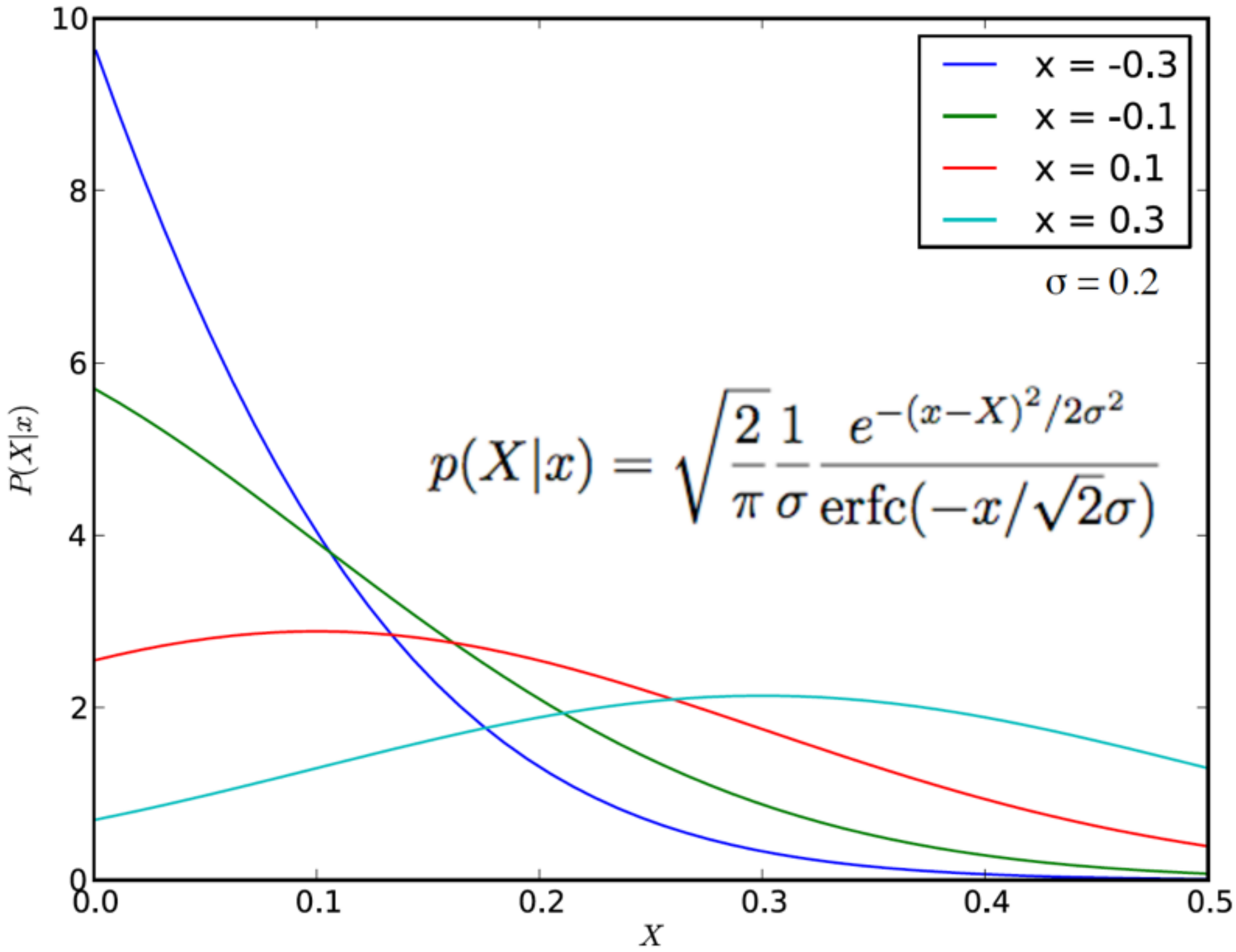
$$p(x) = \int_{-\infty}^{\infty} p(x|X)p(X)dX = \int_{0}^{\infty} p(x|X)dX$$

- Then $$p(X|x) = \frac{e^{-(x-X)^2/2\sigma^2}}{\int_0^\infty e^{-(x-X)^2/2\sigma^2} dX}$$

$$p(X|x) = \sqrt{\frac{2}{\pi}} \frac{1}{\sigma} \frac{e^{-(x-X)^2/2\sigma^2}}{\mathrm{erfc}(-x/\sqrt{2}\sigma)}$$

for $X > 0$, zero otherwise

Measurement of constrained quantities



Notes on error function

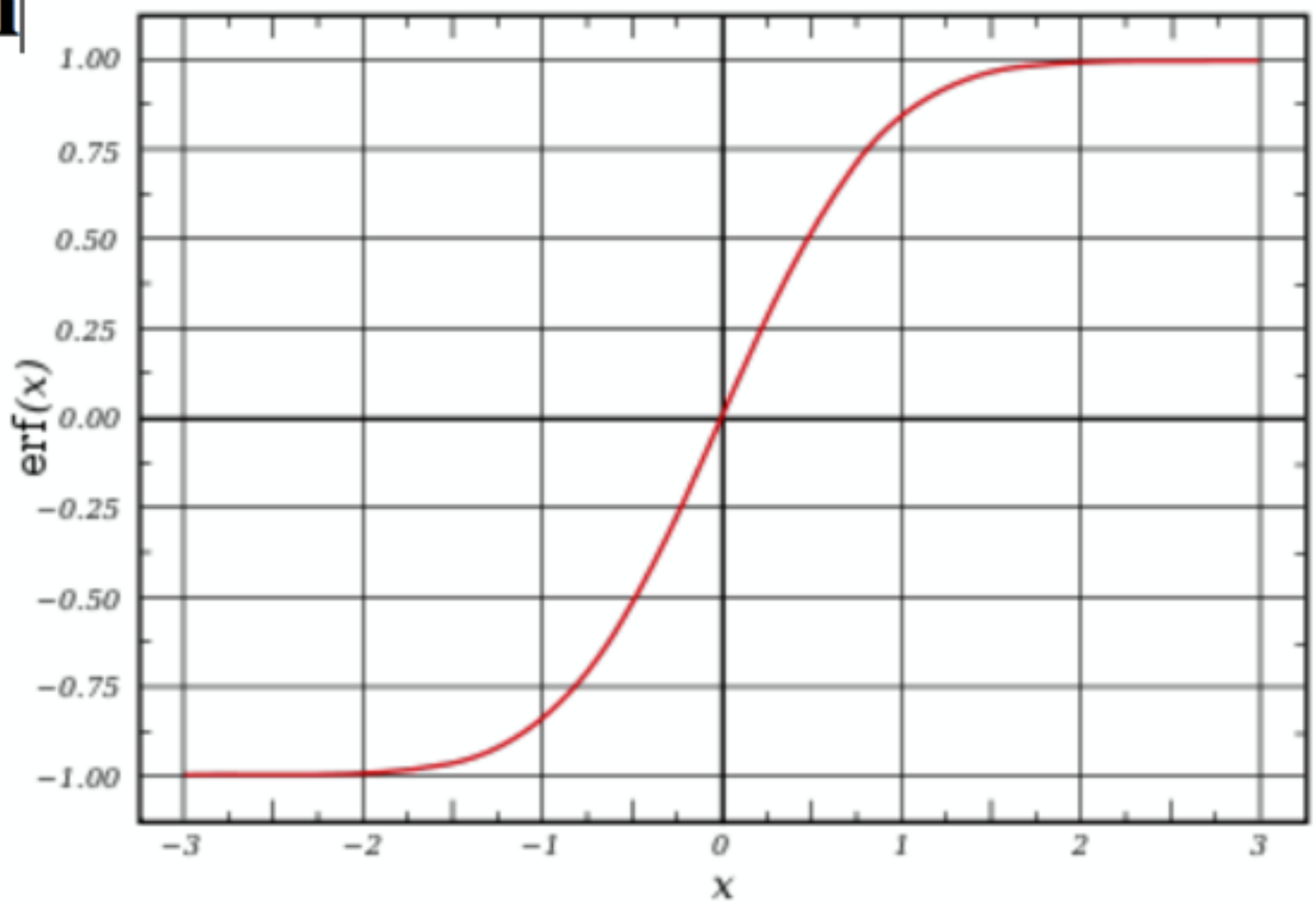
- Error function

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

  - No analytic solution

- Related to integral of standard normal distribution Φ (as given in Gaussian probability tables) by

$$\Phi(x) = \frac{1}{2} + \frac{1}{2}\operatorname{erf}(x/\sqrt{2})$$



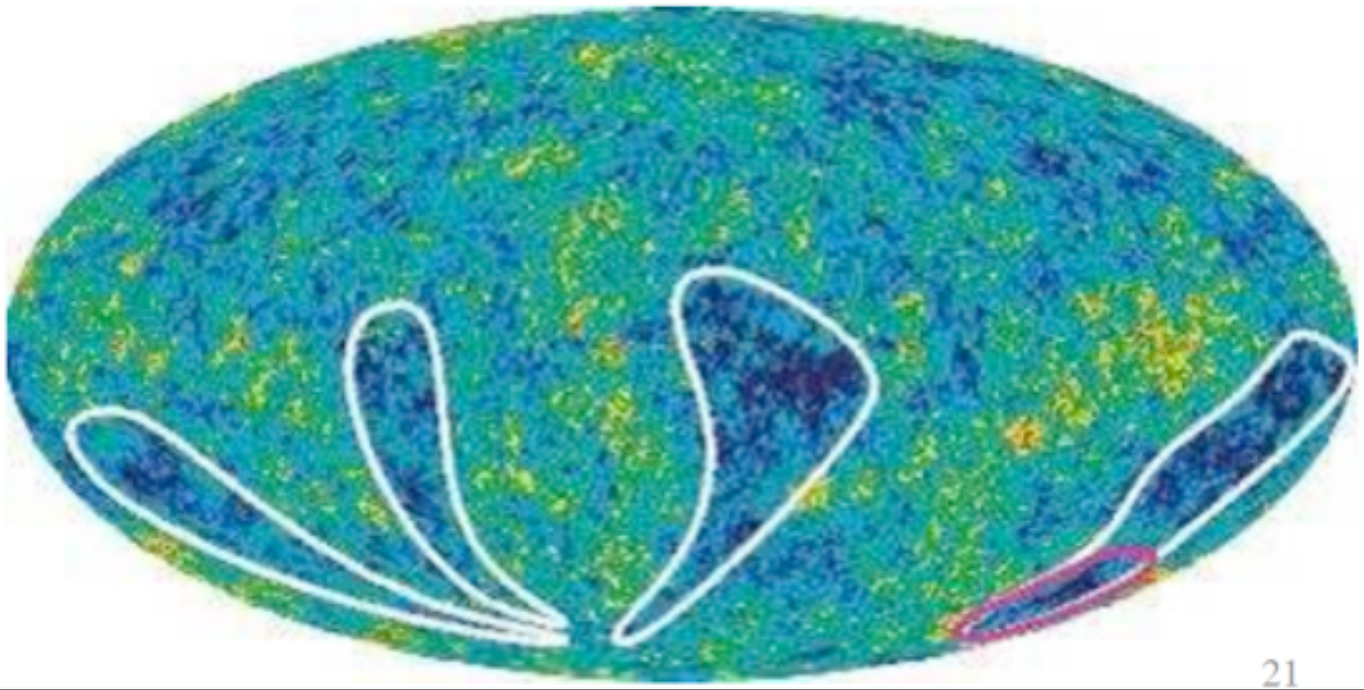
- Complementary error function

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt$$

Warning: *a posteriori* statistics

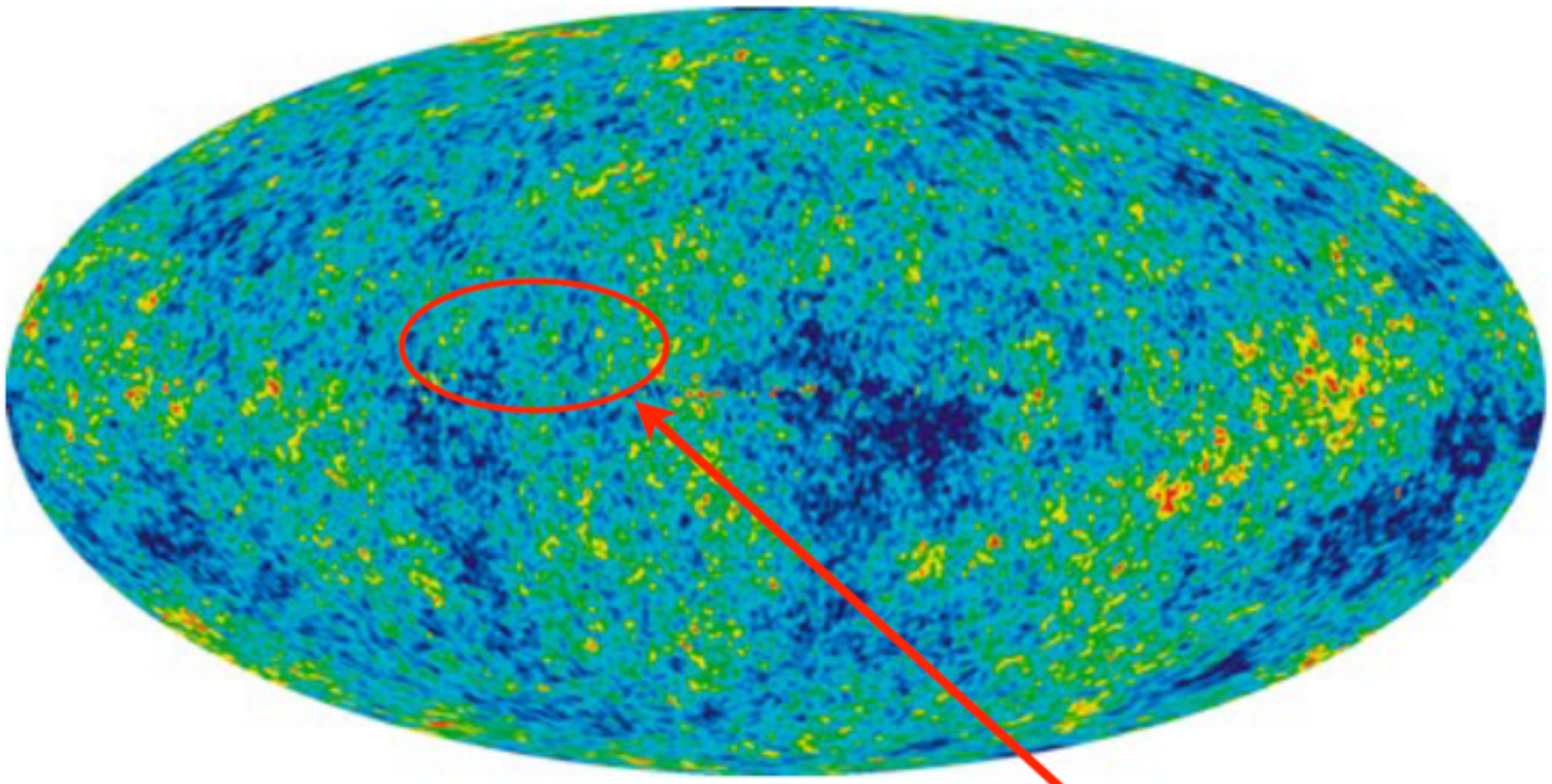
- Suppose I toss a coin 10 times and obtain the sequence HTHTHTHTHT
- Cannot claim *a posteriori* that probability of this sequence $0.5^{10} \approx 10^{-4}$ is significant − probability of *any* given sequence is the same
- Outcome needs to be predicted in advance (*a priori*) to be significant

- Example:
  - Alignment of low-order multipoles in CMB temperature map claimed to be very unlikely in standard ΛCDM model
  - However, one needs to predict these anomalies *a priori* if one is to assess their significance



Warning: *a posteriori* statistics

- One will always find a number of extremely rare events in a large enough dataset (~50,000 pixels in this case)



USING BAYES FOR PARAMETER INFERENCE

Data: d

e.g. surface temperature

Model: M

e.g. climate model

Parameters of the model: θ

e.g. warming parameter, greenhouse gas effects

The likelihood is $L(d \mid \theta) = p(d \mid \theta, M)$

The posterior is $p(\theta \mid d, M)$

The prior is $\pi(\theta)=\pi(\theta \mid M)$

BAYES' THEOREM

$$p(\theta \mid d, M) = \frac{p(d \mid \theta, M)\, p(\theta \mid M)}{p(d \mid M)}$$

p(d | M) is the EVIDENCE - important in model selection, but not here.

FINDING THE POSTERIOR

If we make M implicit,

$$p(\theta \mid d) = \frac{L(d \mid \theta)\, \pi(\theta)}{p(d)}$$

We will often want $p(\theta \mid d)$, the posterior.

So we need to consider our prior, data, and likelihood.

PRIOR

Describes our state of knowledge before the new data, or more generally our degree of belief.

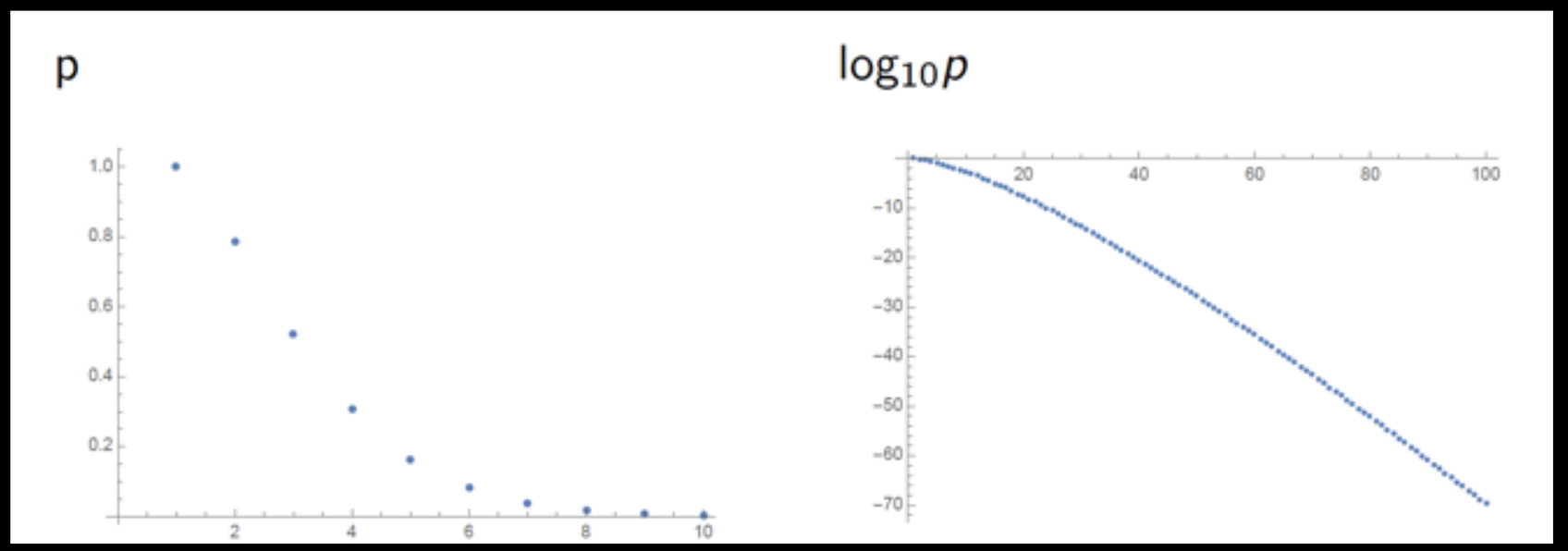
As we gather data, the likelihood will dominate over the prior, so it should cease to be important (for parameter estimation - still important for model selection).

Subjective priors are set previously and are independent of the experiment.

FLAT PRIOR

If we don’t know the value of a parameter θ, it may seem that choosing π(θ)=constant is reasonable.

But imagine a flat prior in an N-cube compared with an N-sphere of same diameter - massively different range of parameters allowed!



Heavens

CHOICE OF PRIORS

If you're looking at a location parameter, where you expect the likelihood to depend on the distance of the observation from the parameter, choose a flat prior. (e.g. mean of a Gaussian, or position of a galaxy).

If you're looking at a scaling parameter s, e.g. unknown variance of a distribution, then consider choosing a prior proportional to 1/s (i.e. uniform in the log of the parameter).

JEFFREYS PRIOR

Can we choose an objective prior - one that has the least effect on the final outcome possible?

For 1 parameter this is possible - called the Jeffreys prior.

$$\pi(\theta) \propto \sqrt{I(\theta)}$$

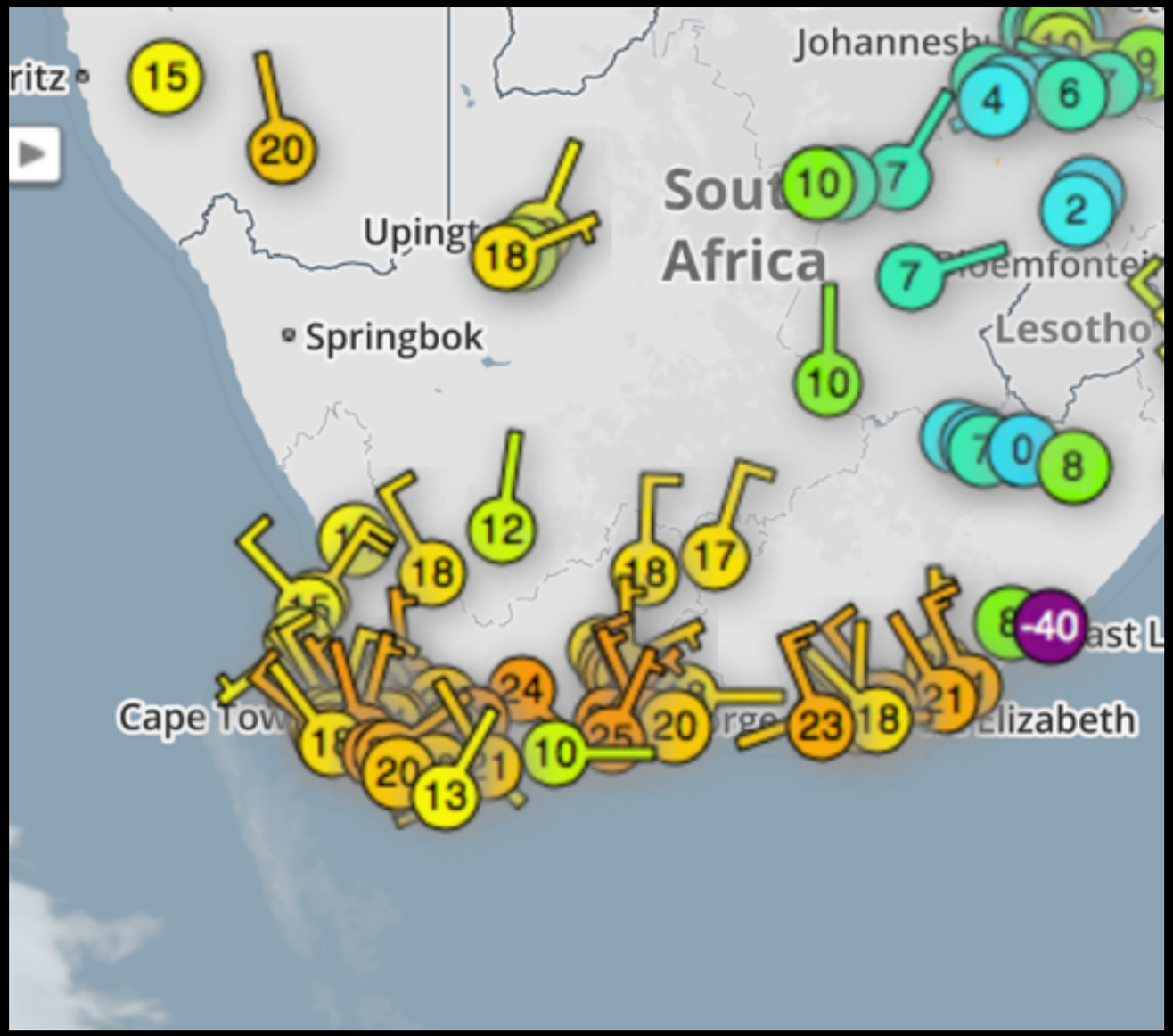
$$I(\theta) = \left\langle \frac{d^2 \ln L(d|\theta)}{d\theta^2} \right\rangle$$

Fisher information

But there are problems with this - the average is taken over likelihood involving data, so not independent of experiment.

DATA

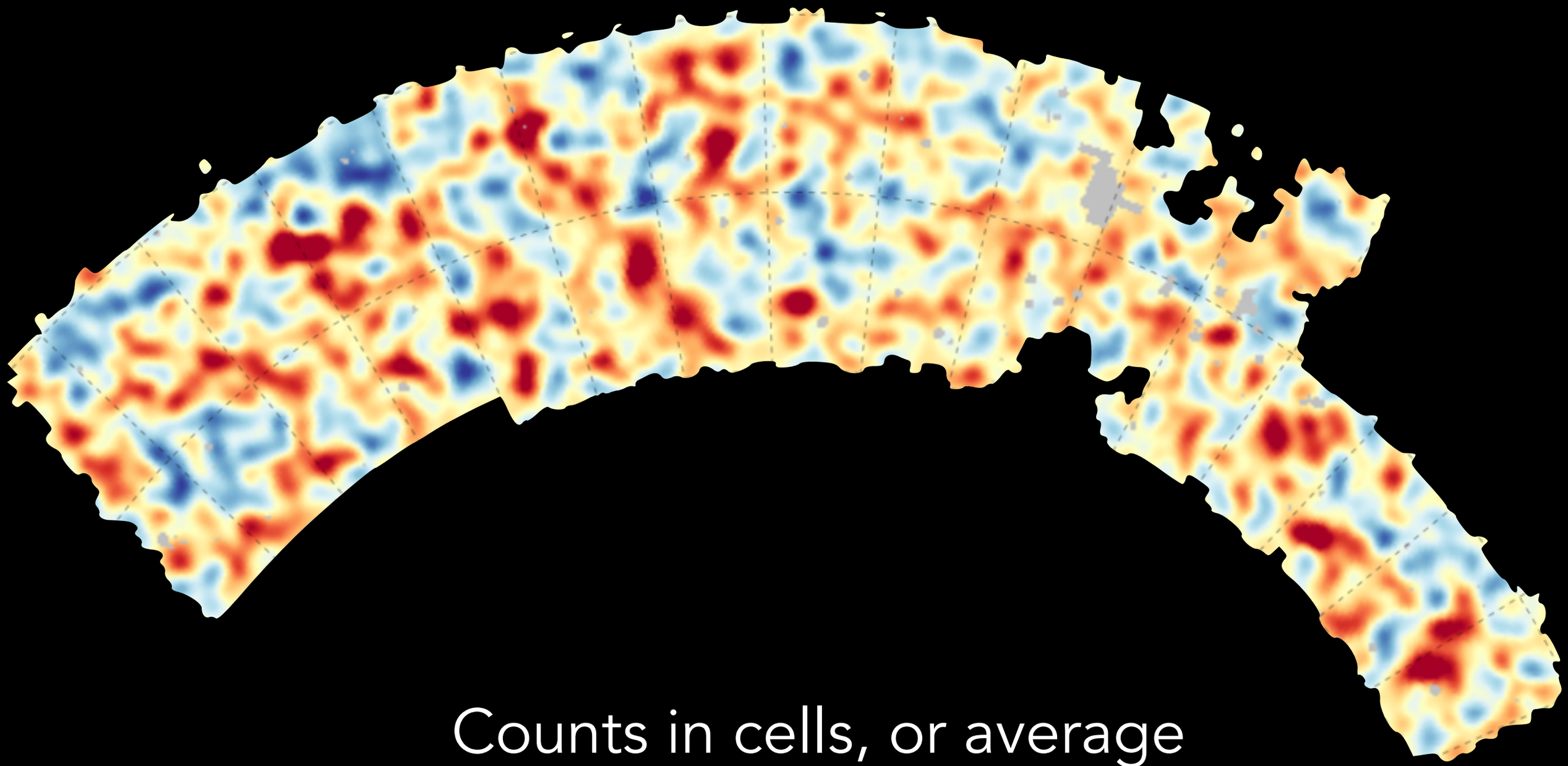
Often the dataset will be enormous! Don't want to calculate likelihood for every single datapoint.



Instead find a way of compressing the data.

Summary statistics:
e.g. binned data; maps; correlation function.

PIXELISED MAPS



Counts in cells, or average
of quantities in cells.

MOMENTS OF A DISTRIBUTION

Central moments:

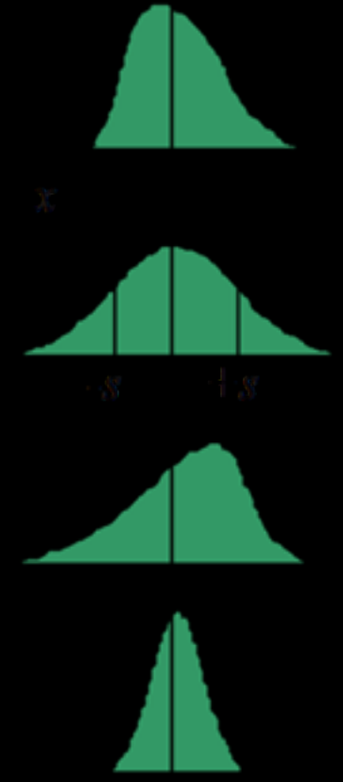
$$\mu_n = \langle (X - \bar{X})^n \rangle = \int_{-\infty}^{\infty} f_X(x)(x - \bar{X})^n dx$$

Skewness:

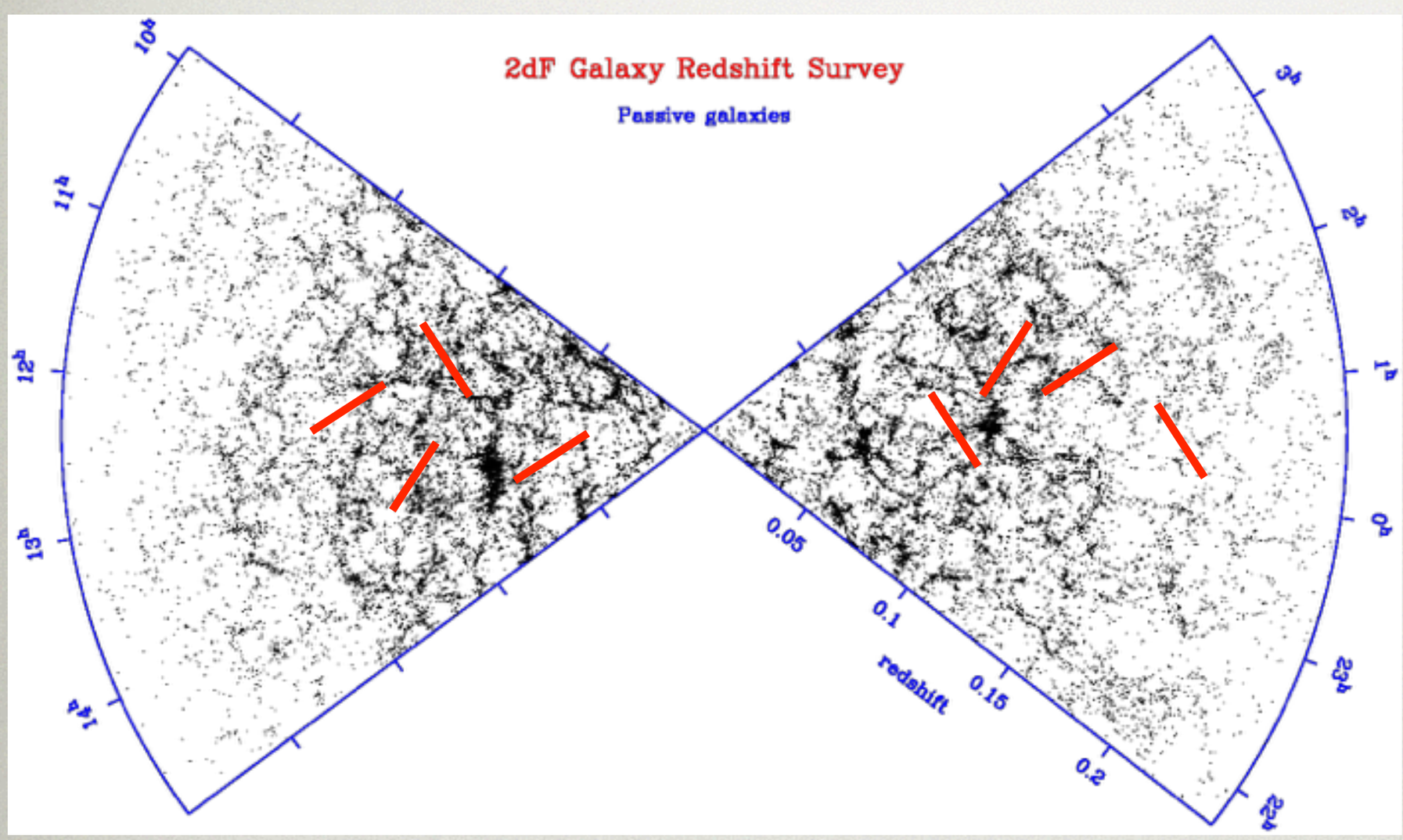
Kurtosis:

$$\gamma_1 \equiv \frac{\mu_3}{\mu_2^{3/2}}$$

$$\gamma_2 \equiv \frac{\mu_4}{\mu_2^2} - 3$$



Measuring clustering



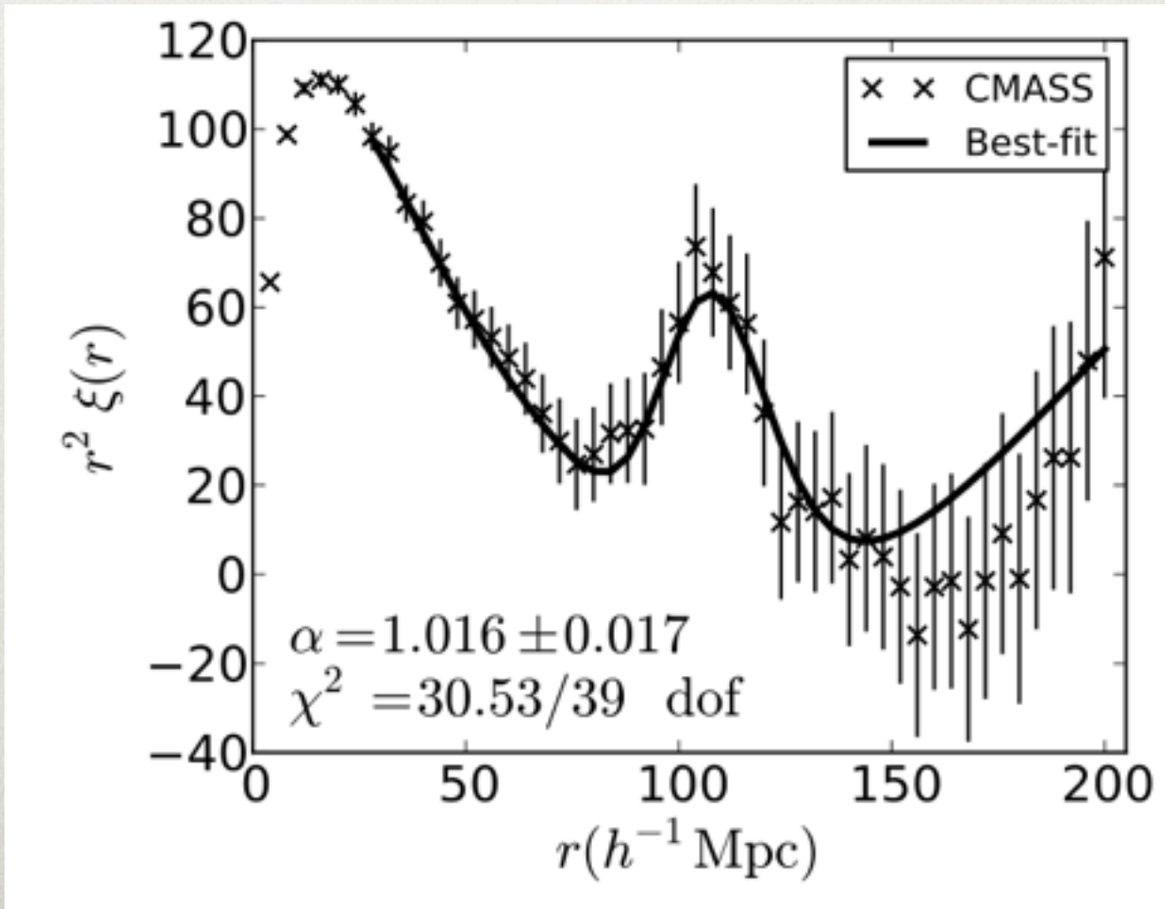
Degree to which number of pairs is in excess of that expected by laying objects down at random

$$dP = \rho_0^2 \left[1 + \xi(r)\right] dV_1 dV_2$$

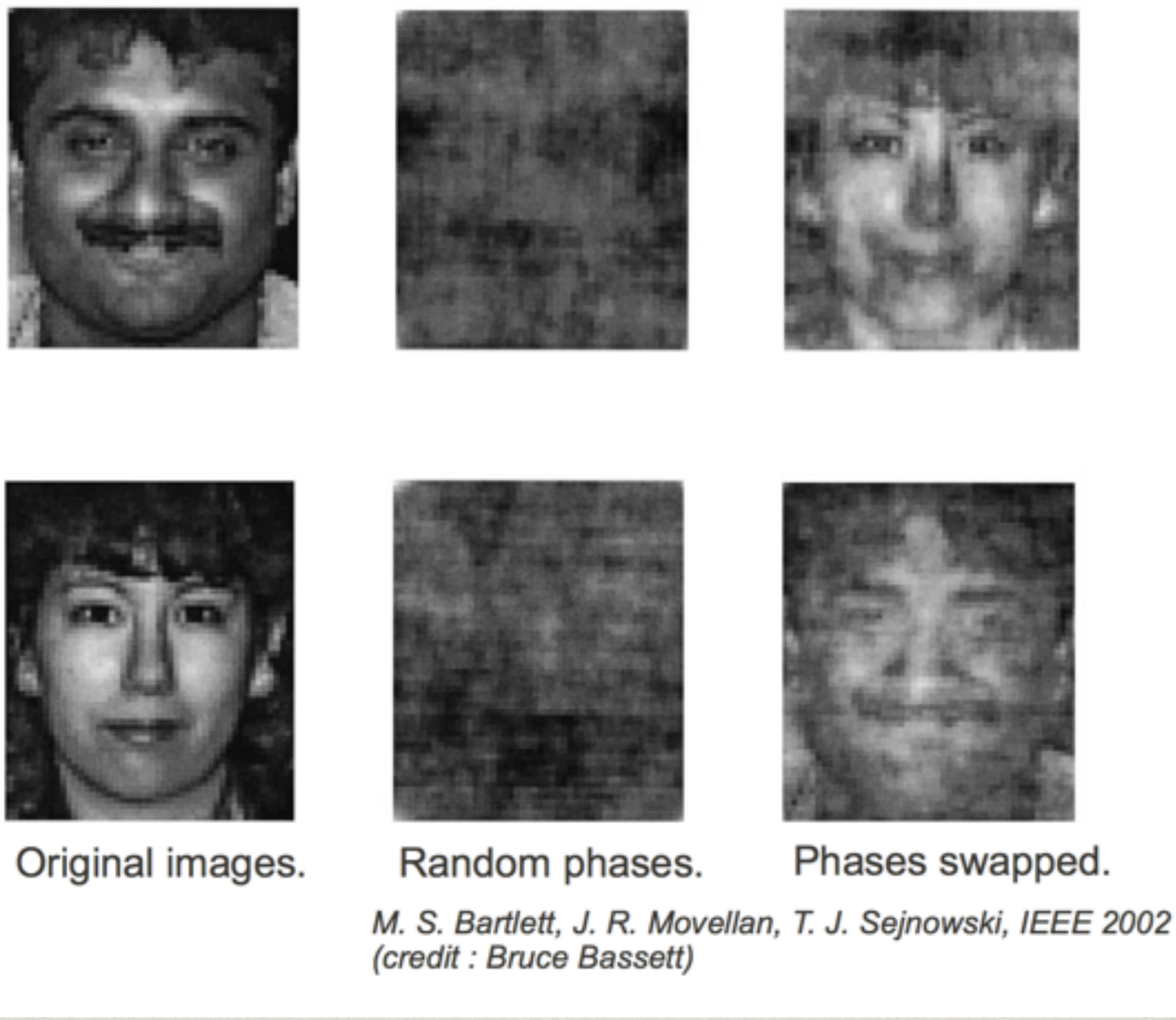
Correlation Function

We examine the overdensity, $\delta = \dfrac{\rho - \rho_0}{\rho_0}$

The correlation function is $\xi(r) = \langle \delta(\mathbf{x}) \delta(\mathbf{x} + \mathbf{r}) \rangle$



There is more to a field...



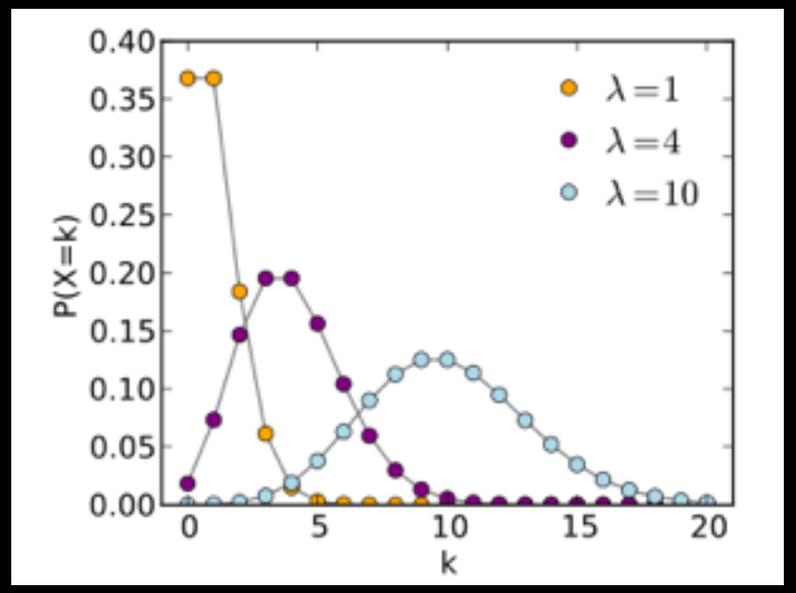
Original images. Random phases. Phases swapped.

M. S. Bartlett, J. R. Movellan, T. J. Sejnowski, IEEE 2002
(credit : Bruce Bassett)

If the field isn't Gaussian, there is more information in higher order statistics, phase correlations

ANOTHER USEFUL PROBABILITY DISTRIBUTION: POISSON DISTRIBUTION

$$\mathcal{P}_X(x,\lambda) = \frac{1}{x!}e^{-\lambda}\lambda^x,$$



x = 0, 1, 2...,

λ is expected value.

Mean and variance are λ.

Probability of x events occurring in fixed time/region, if events occur at a constant rate and independently of the time/place of the previous event.

Likelihood

- If a certain hypothesis is correct, what is the probability that our data set would result from a series of measurements?
- Given a hypothesis A:
  - pdf = $f_A(x)$
  - Measure X; obtain result x_1
  - If hypothesis A correct, then probability dP_1 of result being between x_1 and $x_1 + dx$ is given by

$$dP_1 = f_A(x_1) \cdot dx$$

- For N measurements of X, probability of our outcome is

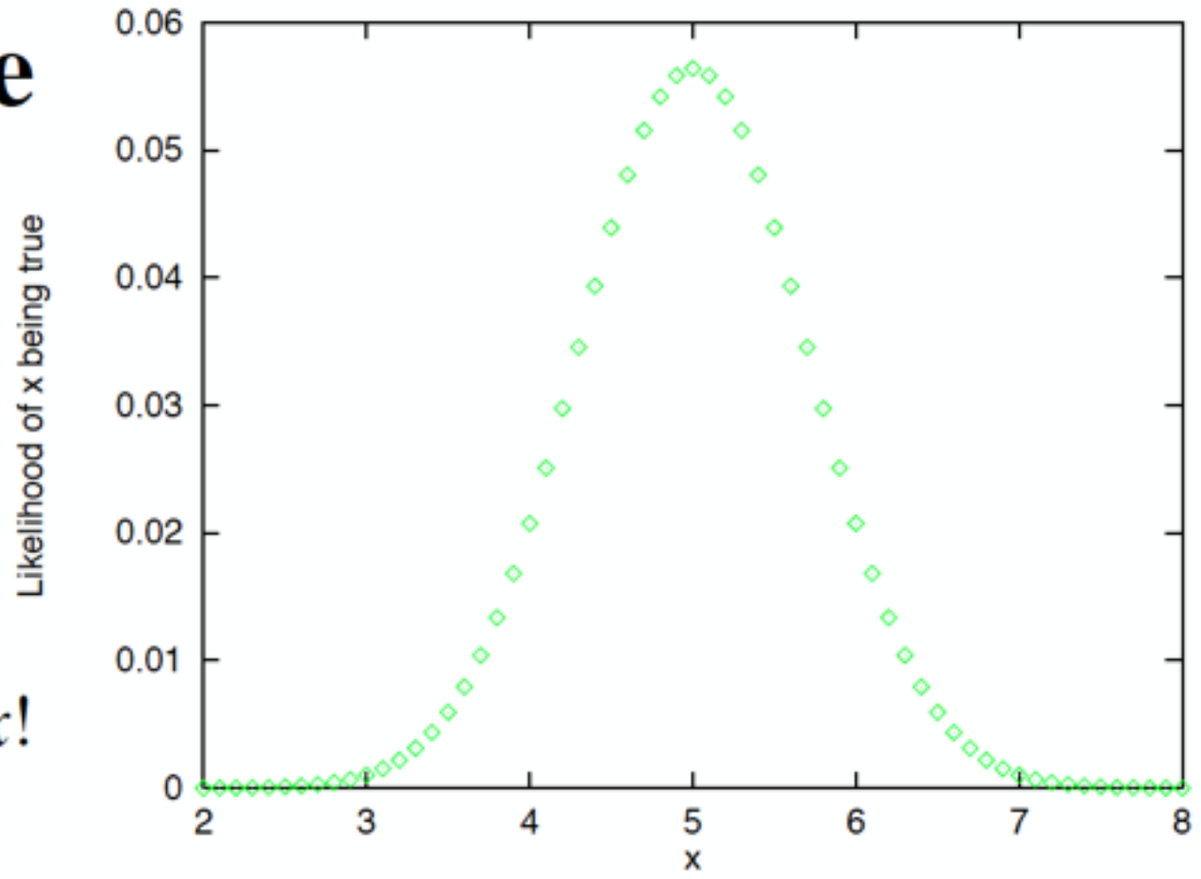
$$dP_A = f_A(x_1)dx \cdot f_A(x_2)dx \cdot \ldots f_A(x_N)dx = \prod_{i=1}^{N} f_A(x_i)dx$$

- Likelihood defined as

$$\mathcal{L} = \prod_{i=1}^{n} f_A(x_i)$$

Example: Length of a table

- Make N measurements x_i of table length
- What "true" length x_0 is the most likely to have resulted in our dataset $\{x_1, x_2, \ldots x_N\}$?
- Need to consider all possible values of x!
  - Some will be more "likely" than others

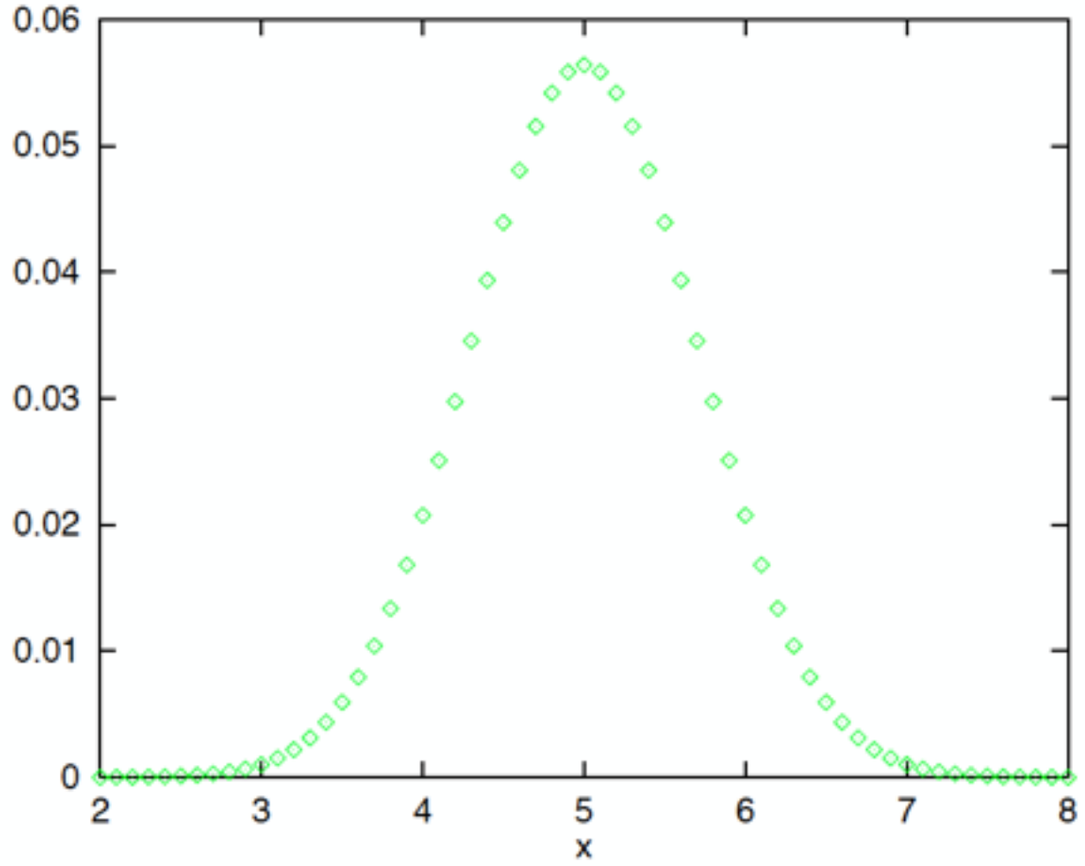


- Construct and plot the likelihood as a function of x, based on our data set
  - For each possible value of x:
    If this were the correct value, what is the probability that we would have obtained this set of results?
  - N large $\Rightarrow$ $L(x)$ normally distributed (central limit theorem)
- Most "probable" value of x, denoted x^*, is the best estimator of length x_0
  - Width Δx is a measure of accuracy with which x_0 is determined

Example: Length of a table

- Let each measurement be $x_i \pm \sigma_i$
- Assume x_i normally distributed about x with pdf:

$$f(x_i; x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i - x)^2}{2\sigma^2}\right]$$



- Likelihood of a particular outcome:

$$\mathcal{L}(x) = \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{(x_i - x)^2}{2\sigma_i^2}\right]$$

- x_0 is the value at which $\mathcal{L}(x)$ is maximised

Example: Length of a table

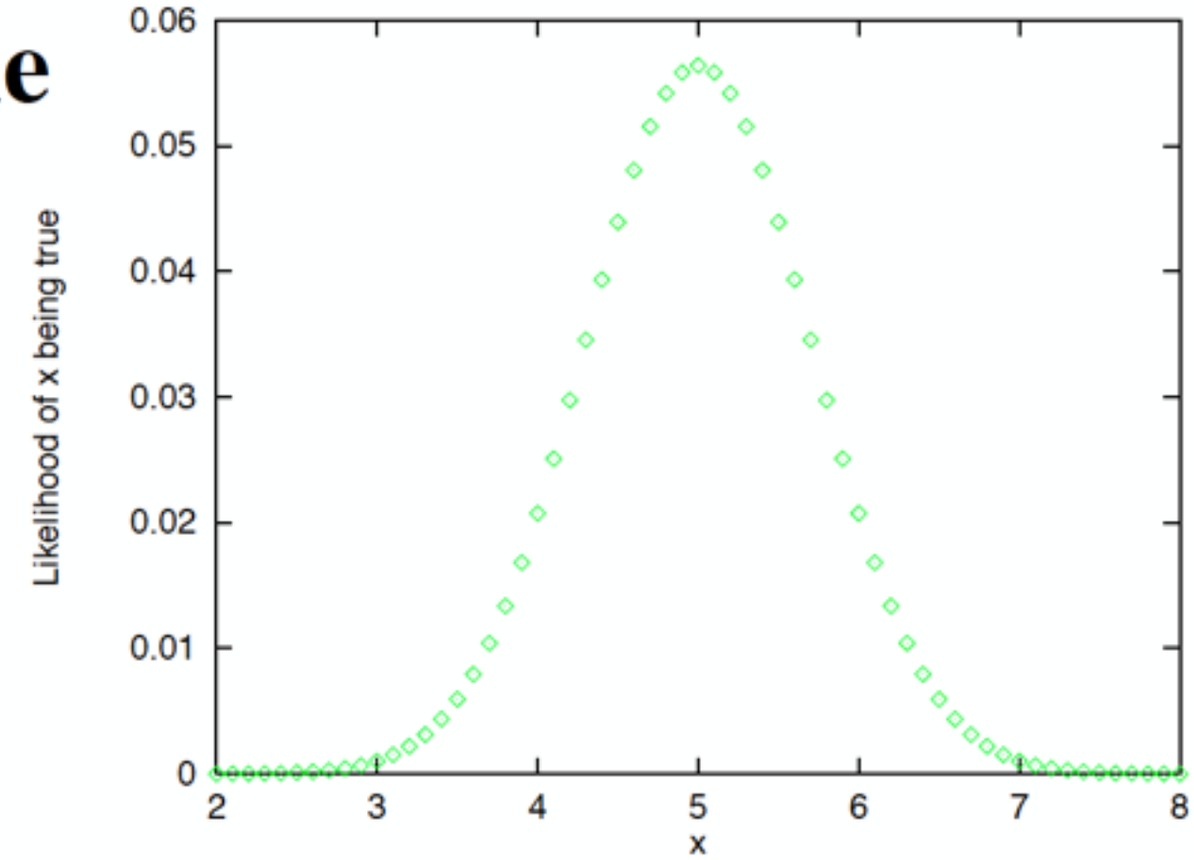
- In practice it is easier to maximise the log likelihood

$$l(x) \equiv \ln(\mathcal{L}(x))$$

Maximise by setting $dl/dx = 0$

$$l(x) = K - \frac{1}{2}\sum_i \left[\frac{(x_i - x)^2}{\sigma_i^2}\right]$$

$$\frac{dl}{dx} = \sum_i \frac{(x_i - x)}{\sigma_i^2} = 0$$



Example: Length of a table

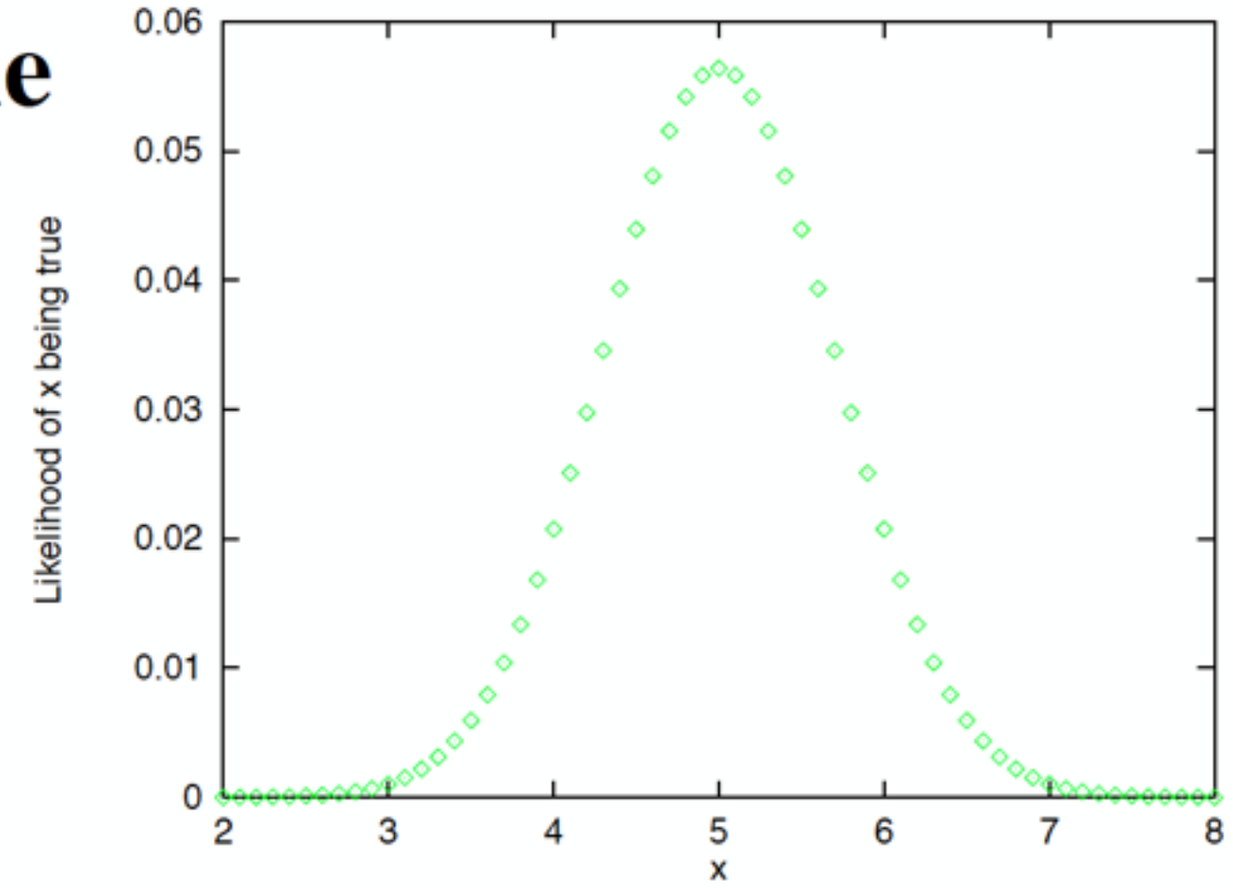
- At maximum:

$$x^* = \frac{\sum_i x_i/\sigma_i^2}{\sum_i 1/\sigma_i^2}$$

  - Weighted mean!

- Error in weighted mean is simply the standard deviation of x^*, σ_{x^*}:

$$\sigma_{x^*}^2 = \left[\sum_{i=1}^{N} \frac{1}{\sigma_i^2}\right]^{-1}$$



LIKELIHOOD

The likelihood p(d | θ, M) is calculated for d our summary statistics. This is relatively easy if d has Gaussian distributed errors.

$$L(d|\theta) = |2\pi\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(d-\mu)^T\Sigma^{-1}(d-\mu)\right]$$

μ is the mean expected for the model with parameters θ. We can get this from a particular theoretical model.

Σ is the covariance matrix, taking into account how data points do not have independent errors.

But are the errors Gaussian? Often not to very good precision.

CENTRAL LIMIT THEOREM

Consider the sum of a large number of identically distributed independent variables.

Its distribution will approach a Gaussian as the number of variables grows bigger.

This is independent of the distribution function of the variables.

Covariance and correlation

- Consider a set of pairs (x, y) of measurements, e.g.
  - (height, weight) of pupils

$$\{(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)\}$$

- Covariance defined as

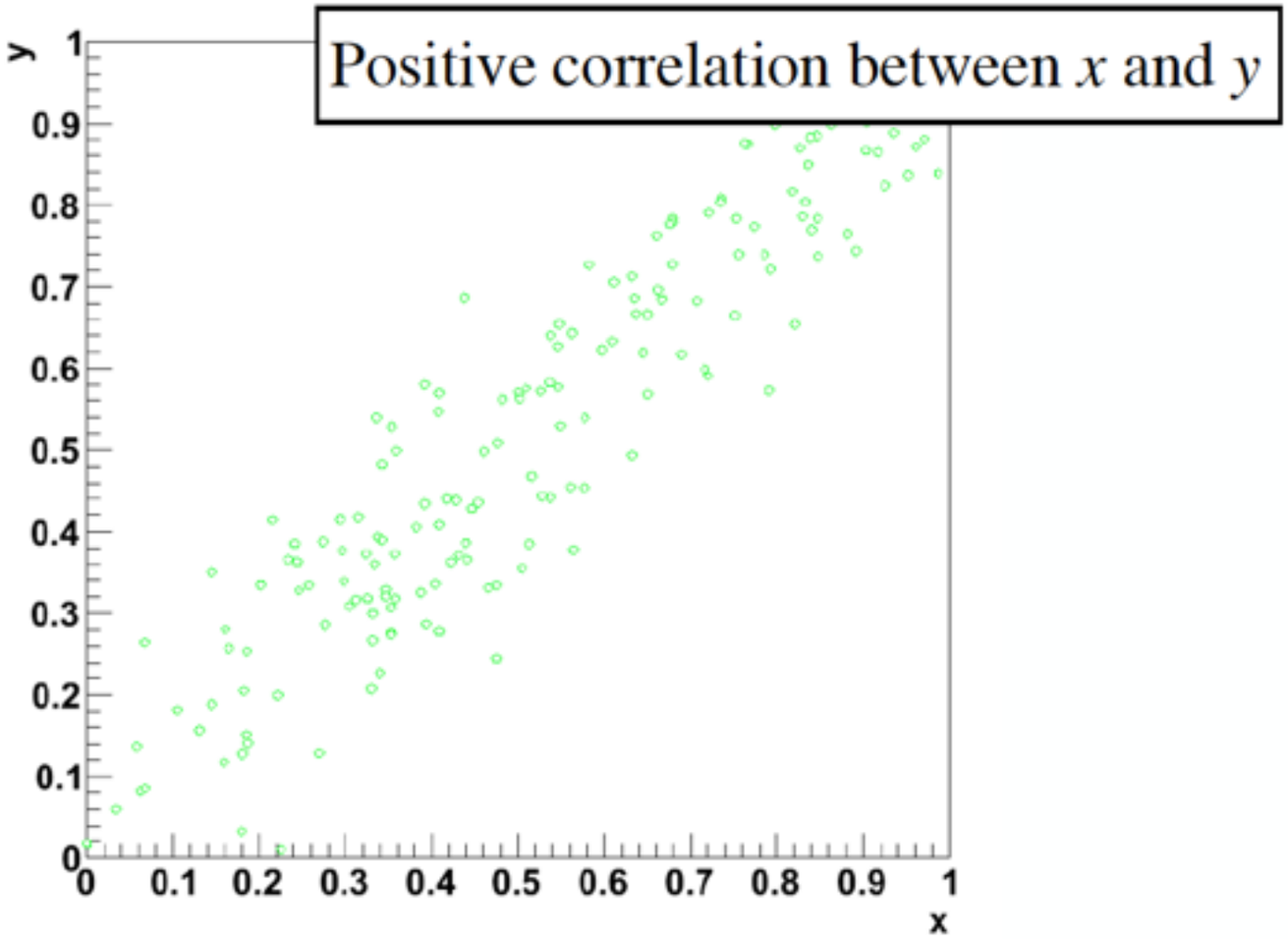
$$\begin{aligned} \mathrm{cov}(x, y) &= \frac{1}{N} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \\ &= \overline{xy} - \bar{x}\bar{y} \end{aligned}$$

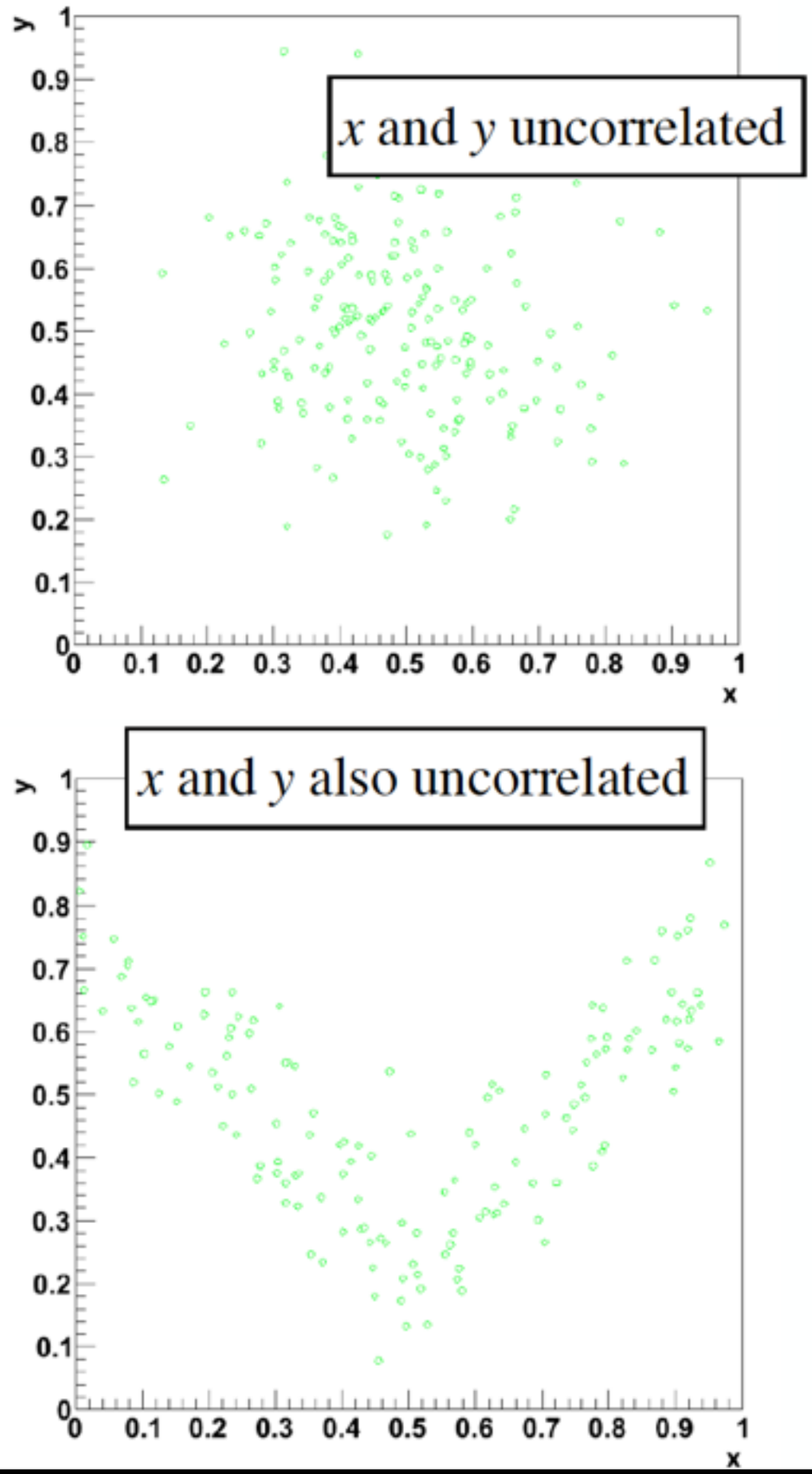
  - Compare with definition of variance

- Dimensionless coefficient of correlation:

$$\rho(x, y) = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y} \qquad -1 \leq \rho(x, y) \leq 1$$

Covariance and correlations





- NB: uncorrelated does not mean independent!

Uncorrelated vs. independent variables

- Uncorrelated variables: covariance is zero
- Independent variables if, and only if, $f(x,y) = f_x(x)f_y(y)$
- Independent $\Rightarrow$ uncorrelated
- Uncorrelated $\not\Rightarrow$ independent

- If A correlated to B, and B correlated to C $\not\Rightarrow$ A correlated to C
  - For example, $B = A + C$ where A and C are independent random variables

The covariance matrix

- Two variables:
$$V = \begin{pmatrix} \sigma_x^2 & \mathrm{cov}(x,y) \\ \mathrm{cov}(x,y) & \sigma_y^2 \end{pmatrix}$$

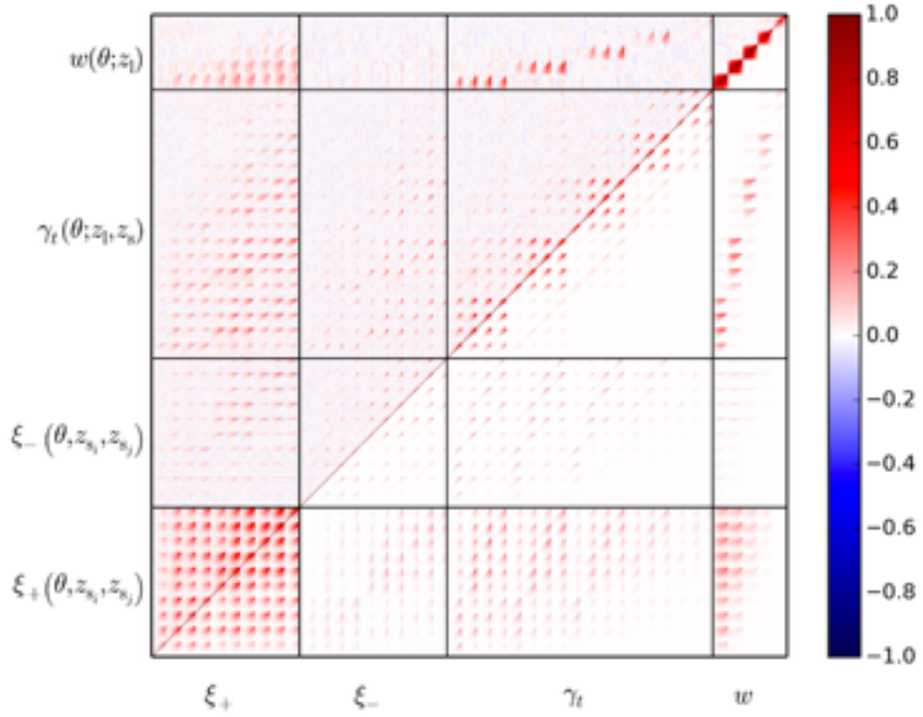
- Multiple variables:
  - Notation: $x_{(1)} = x;\ x_{(2)} = y; x_{(3)} = z; \ldots$
  - Covariance matrix:
$$V_{ij} = \mathrm{cov}(x_{(i)}, x_{(j)})$$
  - Correlation matrix:
$$\rho_{ij} = \frac{\mathrm{cov}(x_{(i)}, x_{(j)})}{\sigma_i \sigma_j}$$
  - Alt. covariance matrix:
$$V_{ij} = \rho_{ij}\sigma_i\sigma_j$$

Brackets for variable index,
No brackets = individual measurement
NB: Brackets often omitted

Matrix notation

- Use vector notation for random variables x_i - or, correctly, $x_{(i)}$:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \qquad \overline{\mathbf{x}} = \begin{bmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_N \end{bmatrix}$$



- Covariance:

$$V_{ij} = \mathrm{cov}(x_i, x_j) = \langle (x_i - \bar{x}_i)(x_j - \bar{x}_j) \rangle$$

- Matrix obtained by multiplying vector of residuals with its transpose (outer product)

$$V(\mathbf{x}) = \left\langle (\mathbf{x} - \overline{\mathbf{x}})(\mathbf{x} - \overline{\mathbf{x}})^T \right\rangle$$

- Covariance or error matrix, real and symmetric

Large errors

- error propagation formulae assume that errors are ‘small’, so that only first-order term in Taylor expansion is used
- For larger errors (e.g. a few per cent or more), error propagation is best done empirically via Monte Carlo simulation

Example: the propagation of large errors

- An experiment in 1980[1] claimed:
  - Quantity R can be measured
  - Neutrino oscillates if $R \leq 0.44$
  - Experimental results:

$$R = \frac{a}{\frac{d}{k^2 e}(b-c) - 2\left(1 - \frac{kd}{e}\right)a}$$

where

$a = 3.173 \pm 1.353$ $\quad d = 0.112 \pm 0.009$

$b = 60.77 \pm 3.57$ $\quad e = 0.32 \pm 0.02$

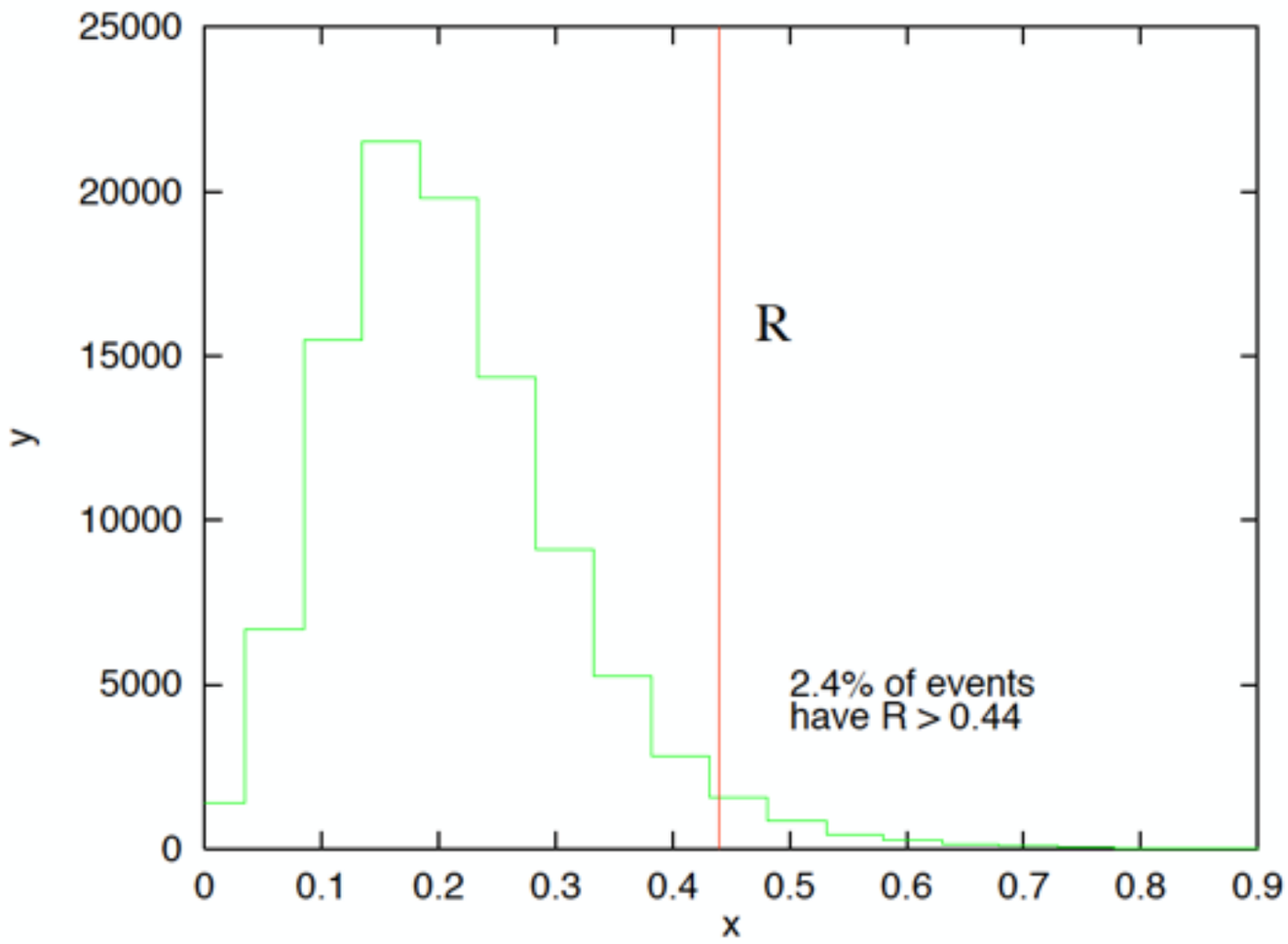
$c = 9.1 \pm 0.6$ $\quad k = 0.89$

- Linear approximation for error propagation yields
  - $R = 0.17 \pm 0.09$
  - So $(0.44 - 0.17)/0.09 = 3\sigma$ away from expected result for a stable neutrino
  - With Gaussian errors would be a 0.27% probability of being 3.1σ
  - Authors claimed this was evidence for neutrino oscillations
- However, both a and c have large errors ~ 30% and ~7% respectively...

[1]PRL 45, 1307 (1980)

Example: the propagation of large errors

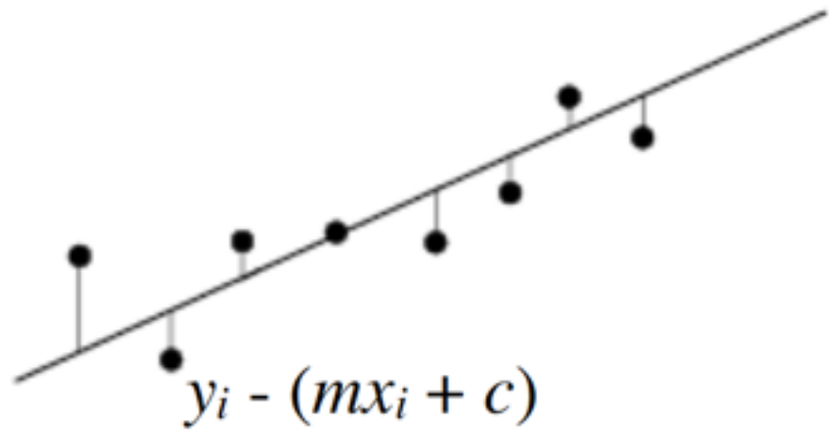
- Instead:
- Generate 10,000 values each of $a, b, \ldots k$
  - Each centred at the measured values and with Gaussian errors
- Calculate R each time
- Plot the distribution of R
- 2.4% of the measurements fall above the 0.44 threshold
  - An order of magnitude larger than the authors claimed, and statistically definitely possible!



- Subsequent experiments did not confirm this "discovery", but *have* seen neutrino oscillations, although not consistent with the claims from 1980

Least squares fitting

- Most commonly used technique for parameter estimation
- Special case of MLM
  - In general has no optimum properties (min. variance, lack of bias etc.)
  - Optimum when function is linear in the parameters



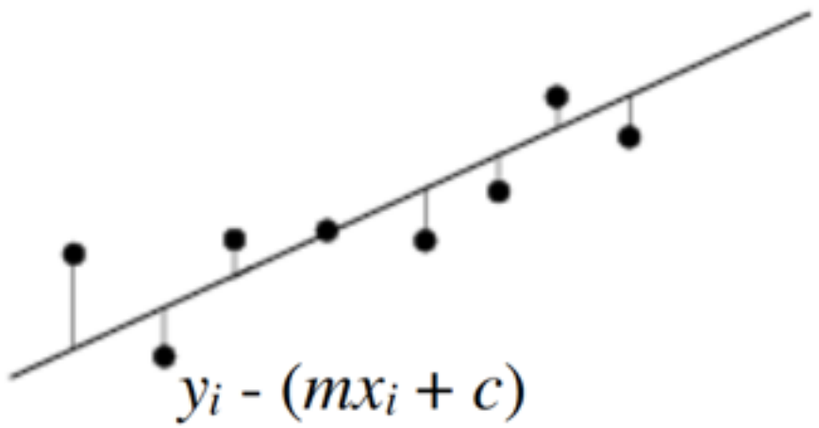
- Typically used with two variables
  - Variables x (known "precisely") and y, with precision σ

$$y = f(x; m, c)$$

- Minimise the sum of the squared deviations

Straight-line fit

- Measure pairs of values (x_i, y_i)
  - x_i assumed infinite precision
  - y_i normally distributed about $y(x_i)$ with σ_i



- Fit straight line:

$$y = mx + c$$

- Fit m and c by minimizing the squared deviations

Straight-line fit

- To within some numerical constants, construct a likelihood function (assuming gaussian errors):

$$\mathcal{L}(m,c) = \prod_i \exp\left[-\frac{(y_i-(mx_i+c))^2}{2\sigma_i^2}\right]$$

- Maximise the log likelihood by minimising $\chi^2 = \text{const} - 2\log L$

$$\chi^2 = \sum_i \frac{(y_i-(mx_i+c))^2}{\sigma_i^2}$$

- χ^2 (chi-squared) is the goodness-of-fit parameter, minimised when

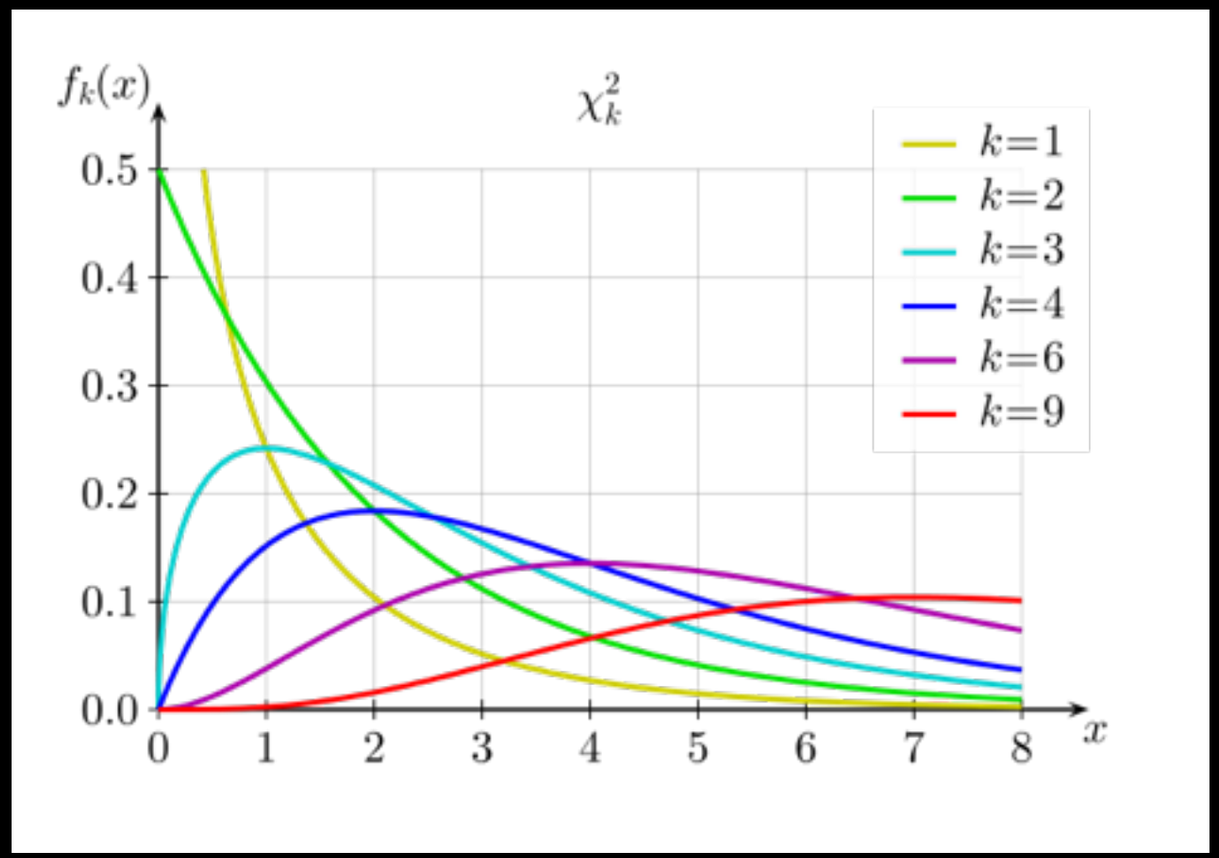
$$\frac{\partial\chi^2}{\partial m} = 0 \qquad \frac{\partial\chi^2}{\partial c} = 0$$

This has a solution!

CHI SQUARE

Distribution expected for the sum of the square of a number k of Gaussian distributed variables.

It has mean k and variance 2k.



CHI SQUARE

Affords us a goodness of fit statistic:

$$\chi^2 = \sum_{i=1}^{N} (y_i - g(x_i))^2 / \sigma_i^2,$$

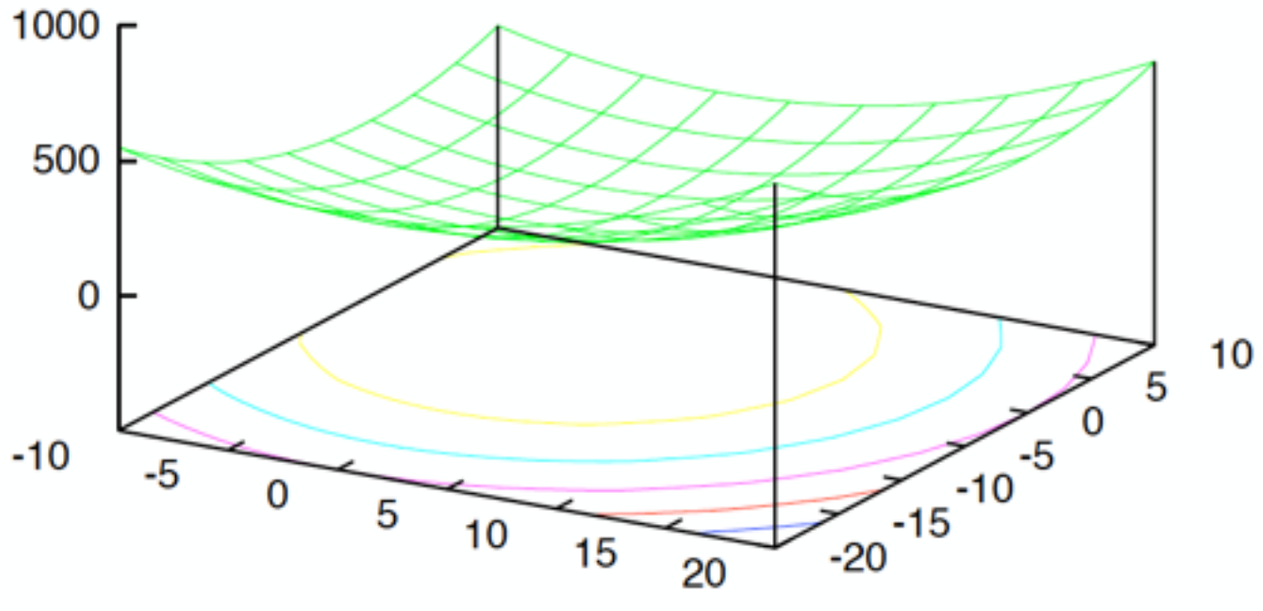
This should be distributed like χ^2 with p=no.data points - no. parameters

Can talk about “reduced χ^2” where we divide by p, then a good fit has $\chi^2_{red}=1$.

Non-linear fitting

- Fitting to non-linear functions
  - Not linear in parameters a_k,
- With normally distributed errors on y_i and χ^2 as before:

$$\begin{aligned}\chi^2 &= (\mathbf{y}-\bar{\mathbf{y}})^T V^{-1} (\mathbf{y}-\bar{\mathbf{y}}) \\ &= \sum_{ij} (y_i - y(x_i;\mathbf{a}))\, V_{ij}^{-1}\, (y_j - y(x_j;\mathbf{a}))\end{aligned}$$



- This is an m-dimensional hyper-surface in the space of the parameters **a,** conceptually the same as the the linear case
- However in this case, $\partial\chi^2/\partial a_i$ yields a set of coupled, non-linear equations
  - Solve numerically by searching the parameter space
  - Analytically — if the problem is tractable (unlikely!)

FINDING THE POSTERIOR: GRID

We could now calculate L for every set of values θ on a grid, then use Bayes to calculate p(θ | d).

$$L(d|\theta) = |2\pi\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(d-\mu)^T\Sigma^{-1}(d-\mu)\right]$$

The problem is that if you have e.g. 5 parameters in θ, each examined at 10 values, that's 10^5 grid points!

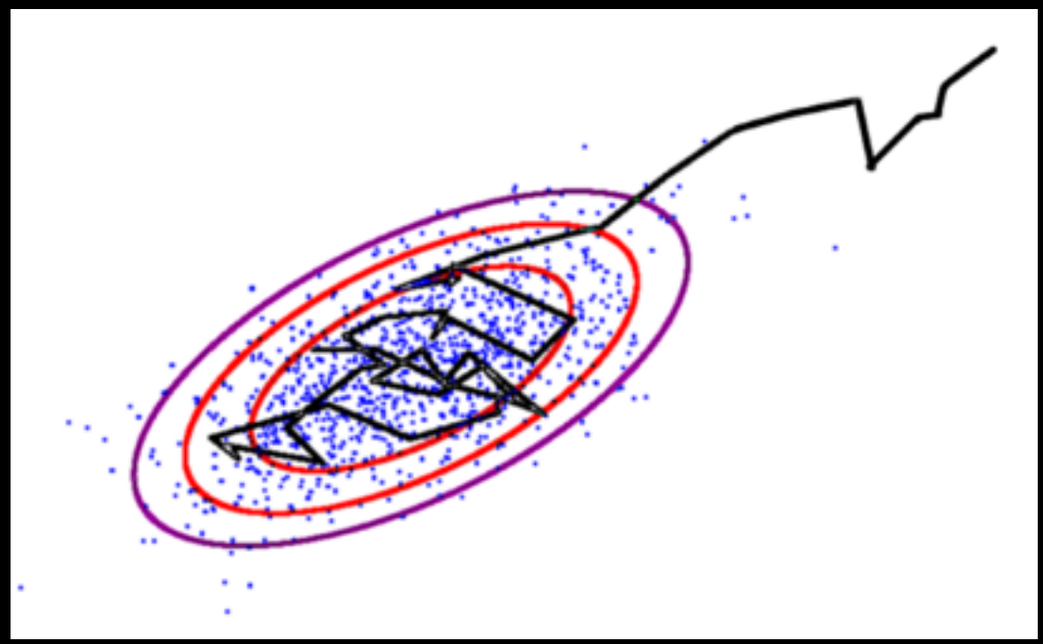
So we need something more efficient.

FINDING THE POSTERIOR: MCMC

- Monte Carlo - so there's randomness involved rather than solving exactly
- Markov Chain - nth point depends on n-1th point, but not previous points (very short memory).

FINDING THE POSTERIOR: MCMC

Evaluate likelihood, then take a step according to a proposal distribution.



This step is accepted/ rejected according to a recipe (Metropolis Hastings).

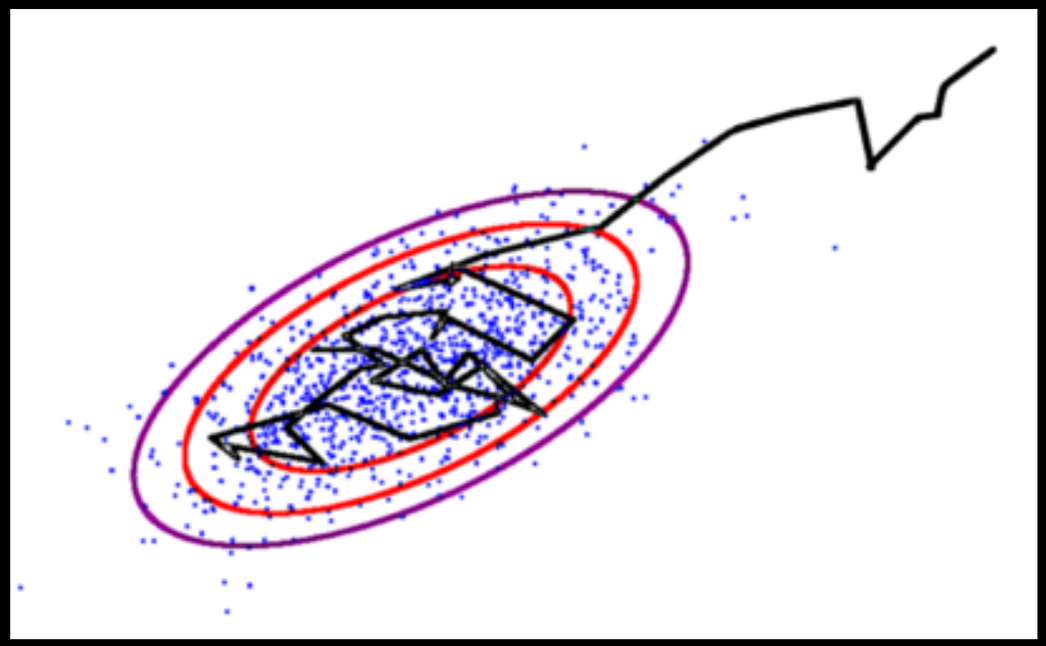
The chain wanders on, and the density of points follows the posterior.

Metropolis Hastings

- Start at a random place x_0
- Determine next x_t using proposal density $Q(x_0,x_t)$. Any fixed function, but often a Gaussian centred on current position.
- Calculate probability for new position. If it's greater than old position, accept it as next step.
- If it's lower than present likelihood, accept it sometimes, with probability $P(x_t)/P(x_0)$.

MCMC

- At first we may be well away from the peak of the distribution, so chains may be biased. Allow a burn-in phase to be thrown away.



- Stopping point needs care - you could split the chain and see how the estimate changes; or have multiple chains and examine the variance between estimators for different chains.

Proposal density

- Needs care to choose right width - too small and it'll take forever to sample the likelihood; chain could get stuck in local max. Too large and everything is rejected!

- One can examine the local curvature and use this to inform the choice of proposal density.

Finding expectation values from MCMC

- For any function of parameters, its expectation can be estimated as

$$\langle f(\mathbf{x}) \rangle = \int d^N \mathbf{x} P(\mathbf{x}) f(\mathbf{x}) \simeq \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{x_i})$$

MODEL COMPARISON

Suppose we have two different accounts (models) for the data.

Which one should we prefer?

What if one model is more complex than the other?

MODEL COMPARISON

We should look at the Bayesian evidence. Recall

$$p(\theta|d, M) = \frac{p(d|\theta, M)\pi(\theta|M)}{p(d|M)}$$

The evidence is the denominator:

$$p(d|M) = \int d\theta \; p(d|\theta, M)\pi(\theta|M)$$

MODEL COMPARISON

Suppose the models are M and M′, with parameters θ and θ′, length n and n′.

We want

$$p(M|d) = \frac{p(d|M)\pi(M)}{p(d)}$$

The probability ratio of two models is

$$\frac{p(M'|d)}{p(M|d)} = \frac{\pi(M')}{\pi(M)} \frac{\int d\theta' p(d|\theta', M')\pi(\theta'|M')}{\int d\theta\, p(d|\theta, M)\pi(\theta|M)}$$

MODEL COMPARISON

If we don't have a prior preference between the models, the ratio becomes the *Bayes factor:*

$$B = \frac{\int d\theta' p(d|\theta', M')\pi(\theta'|M')}{\int d\theta\, p(d|\theta, M)\pi(\theta|M)}$$

MODEL COMPARISON

This can be interpreted with the Kass and Raftery scale:

ln B	
<1	Not notable
1 to 3	Positive
3 to 5	Strong
>5	Very strong

NESTED MODELS

What about comparing two models, one of which is a simpler model nested in a more general one?

i.e. M’ is simpler, n’ parameters
M has these parameters, plus more, total n parameters

M will fit better, surely! But M’ might win as it is simpler - how does this appear?

NESTED MODELS

Assume uniform priors for each parameter. So

$p(\theta \mid M) = 1/(\Delta\theta_1 \Delta\theta_2 \ldots \Delta\theta_n)$

So

$$B = \frac{\int d\theta' p(d|\theta', M')}{\int d\theta\, p(d|\theta, M)} \frac{\Delta\theta_1 \ldots \Delta\theta_n}{\Delta\theta'_1 \ldots \Delta\theta'_n}$$

Suppose likelihood integrals are similar. Then we see that M’ will be favoured due to larger volume of prior for M.

SUMMARY

- Today we have examined two key areas:

- Parameter inference, using Bayes to find p(θ | d, M).
  - Often need to use summary statistics;
  - Calculate likelihood p(d|θ, M);
  - MCMC is a useful approach to this;
  - Care needs taking with choice of prior.

- Model selection, using Bayes to find p(M'|d)/p(M|d).
  - Affected by number of parameters (i.e. prior volume).