Large-scale Structure: the numerical version

Lecture 3:
Gravitational lensing.
Statistical Methods in Cosmology.

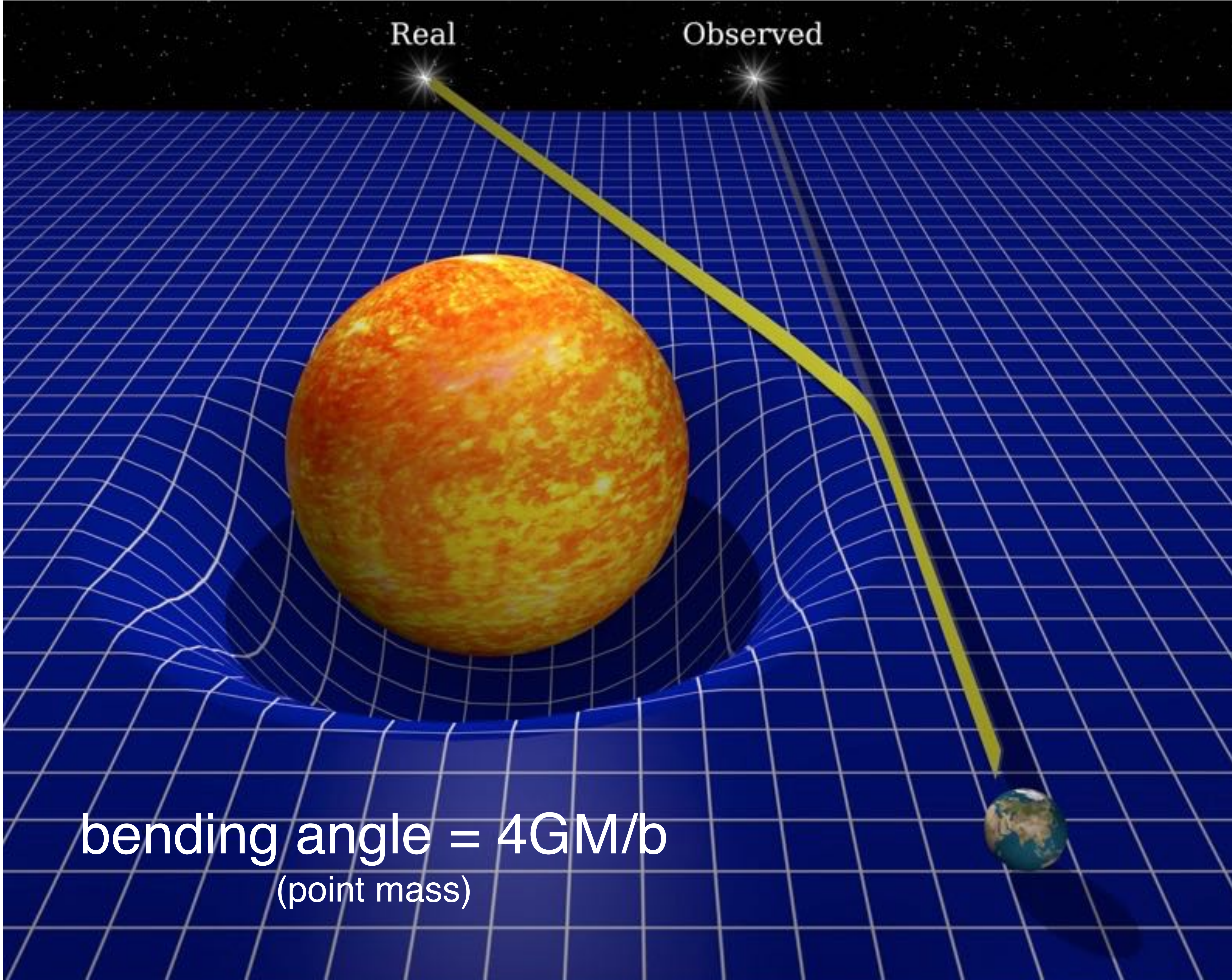
Dragan Huterer

ICTP Trieste/SAIFR Cosmology School

January 18-29, 2021

Max-Planck-Institut für Astrophysik

Alexander von Humboldt Stiftung / Foundation

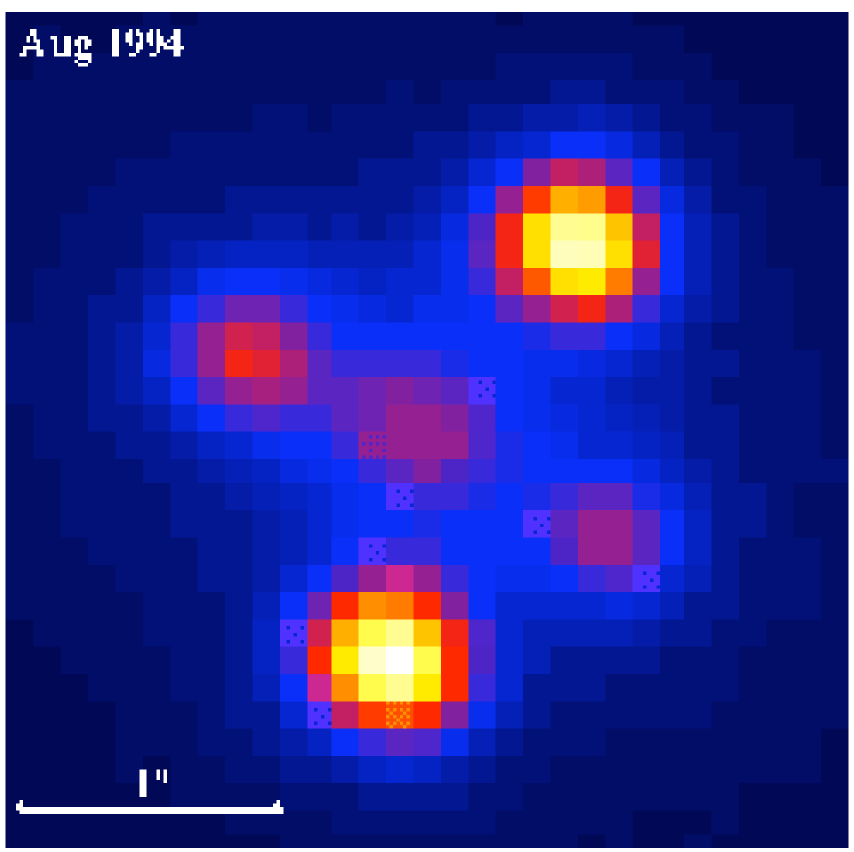


bending angle = $4GM/b$

(point mass)

Strong gravitational lensing

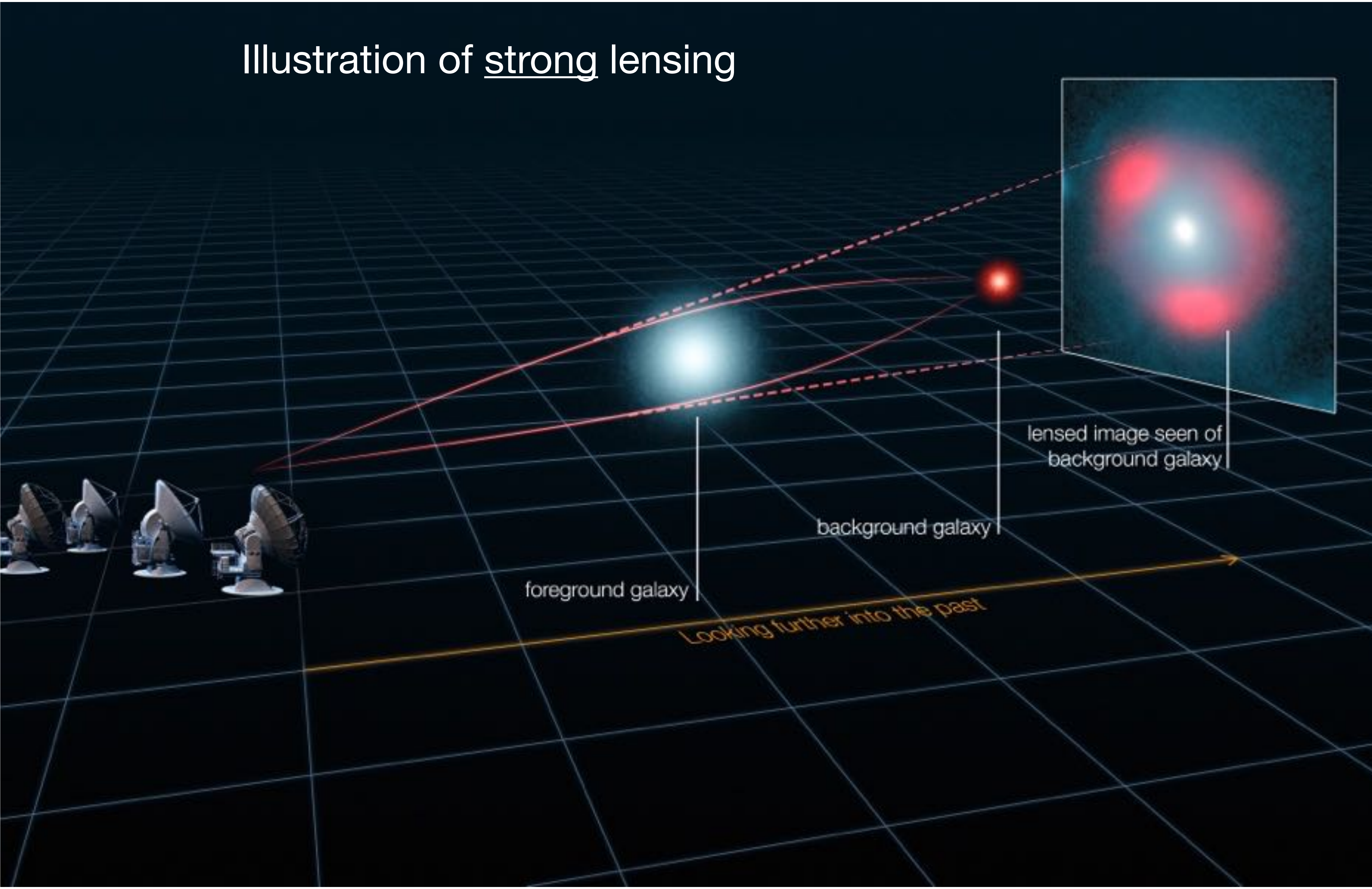
- Multiple images of a single background object (e.g. galaxy) seen
- Lens can be another galaxy or cluster of galaxies - must be massive!



One observed "Einstein cross" lens

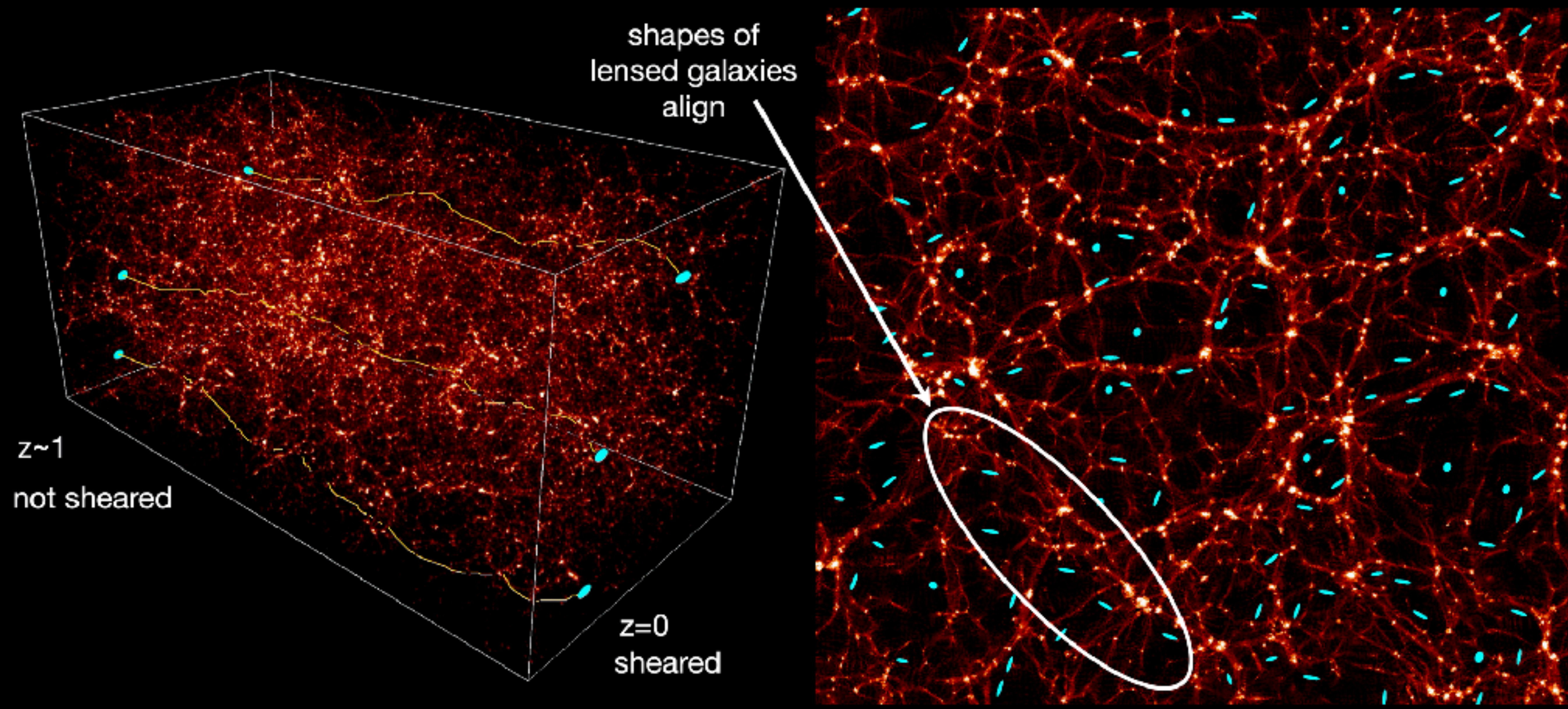
- Typical image separation: 1 arcsec (1")
- Typical distance to lens: cosmological (z~1)
- Typical lens is halfway between observer (us) and the source (images)
- Typical probability of distant galaxy being multiply-imaged: 1/1000

Illustration of strong lensing



ALMA (ESO/NRAO/NAOJ), L. Calçada (ESO), Y. Hezaveh et al.

Weak Gravitational Lensing
by large-scale structure



based off of image by Colombi & Mellier

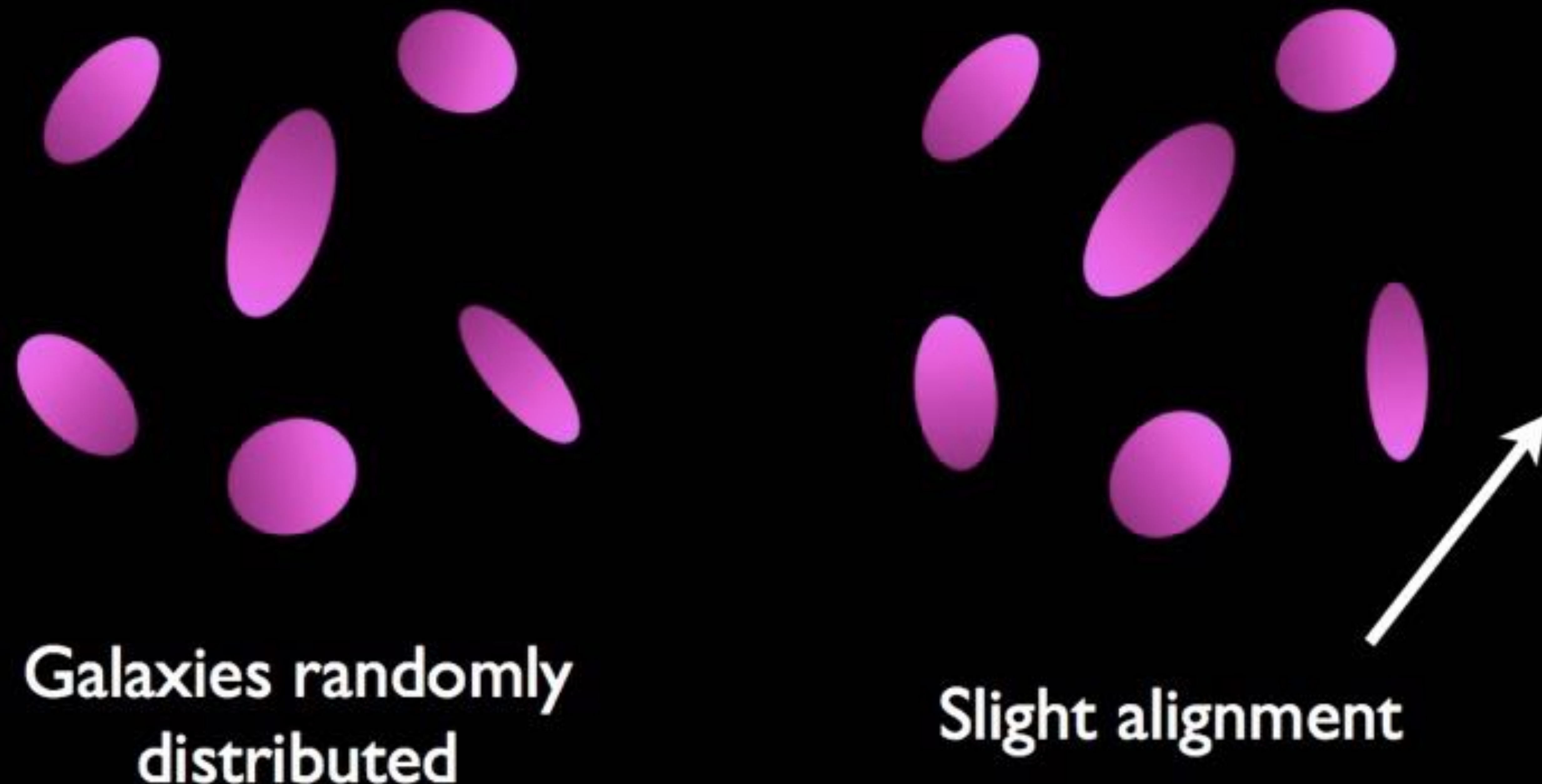


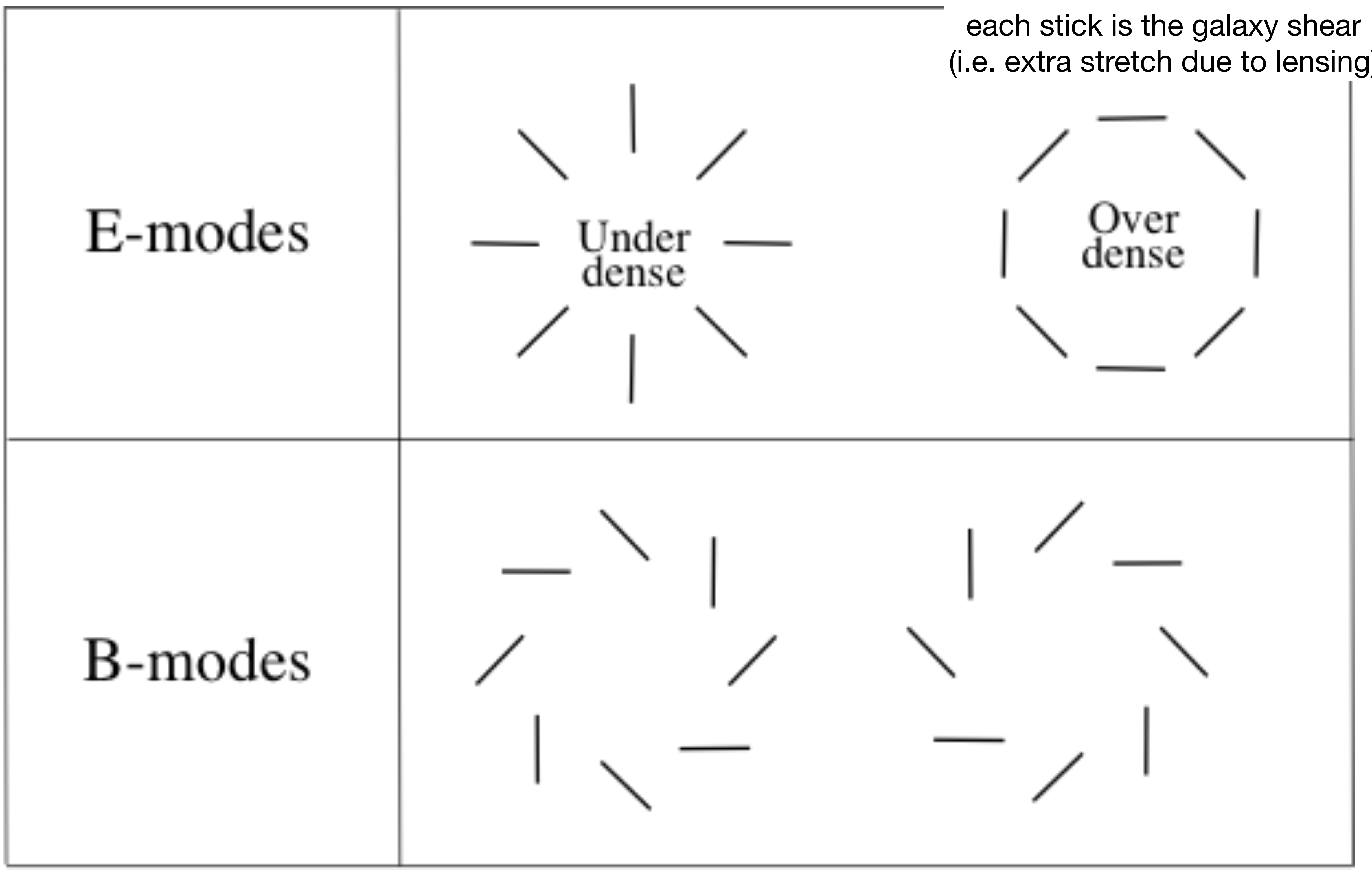
Image: E Grocutt, IfA, Edinburgh

Weak Gravitational Lensing

Credit: NASA, ESA and R. Massey (Caltech)

Key advantage: measures distribution of matter, not light

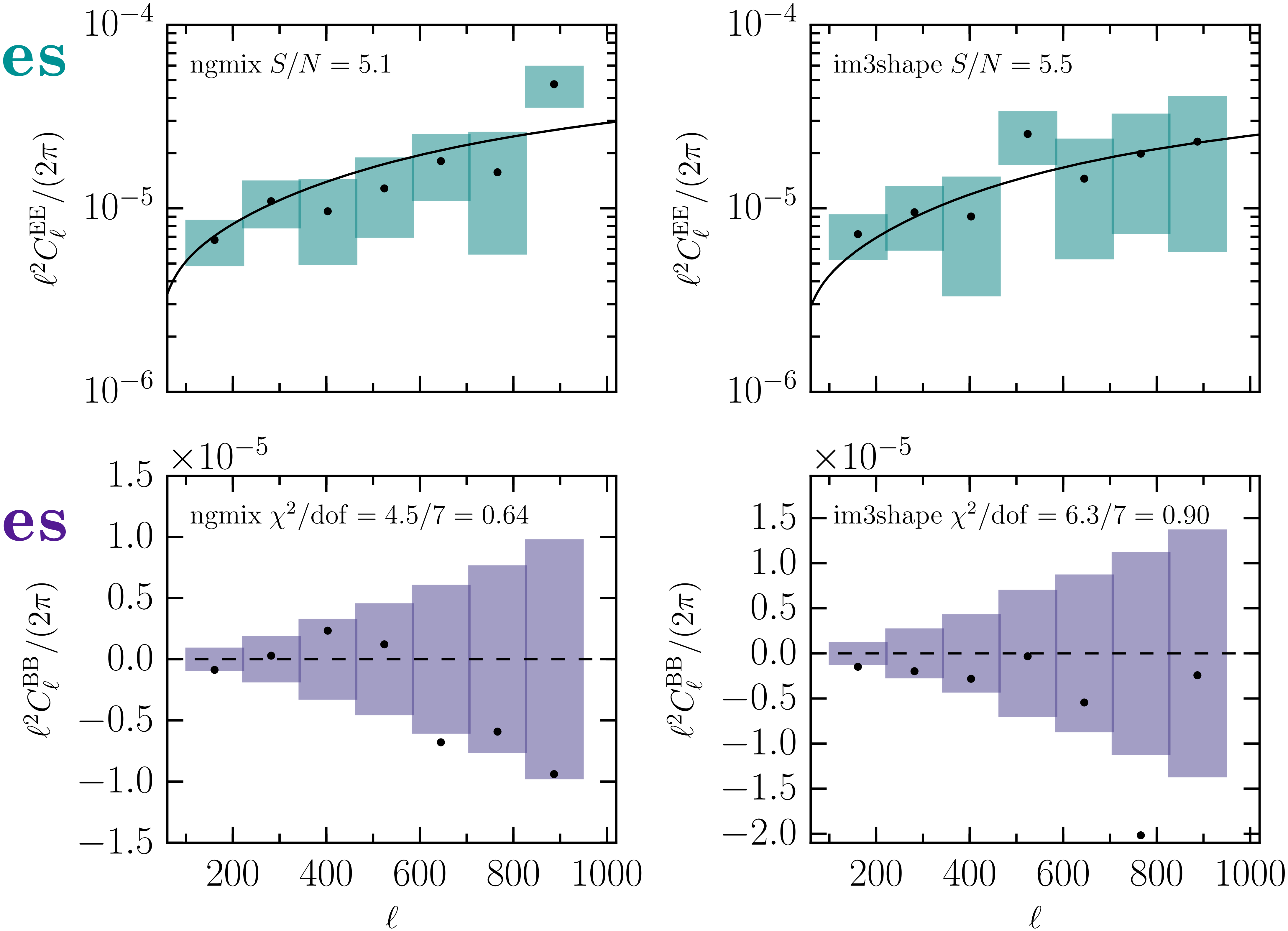
E and B modes



Gravity produces only the E-modes!!

E-modes

B-modes



Becker et al (DES collab), based on DES SV data; arXiv:1507.05598

(obsolete data, but just to illustrate the concept)

Weak Lensing and Dark Matter/Energy

WL measures integral over the line of sight:

$$P^\kappa(\ell) = \int_0^\infty dz\, \frac{W^2(z)}{r(z)^2\, H(z)}\, P\left(\frac{\ell}{r(z)}, z\right)$$

galaxy shear clustering (measure)

distance, volume factors (theory→DM,DE)

(dark) matter clustering (theory→DM,DE)

where

$$W(\chi) \to \frac{3}{2} H_0^2 \Omega_M (1+z) r(\chi) \int d\chi_s n(\chi_s) \frac{r(\chi_s - \chi)}{r(\chi_s)}$$

Weak lensing summary:

- Pros:
  - Sensitive to ALL matter!
  - No bias! (recall $P_g = b^2 P_m$)
  - Sensitive to both geometry (distances) and growth of structure

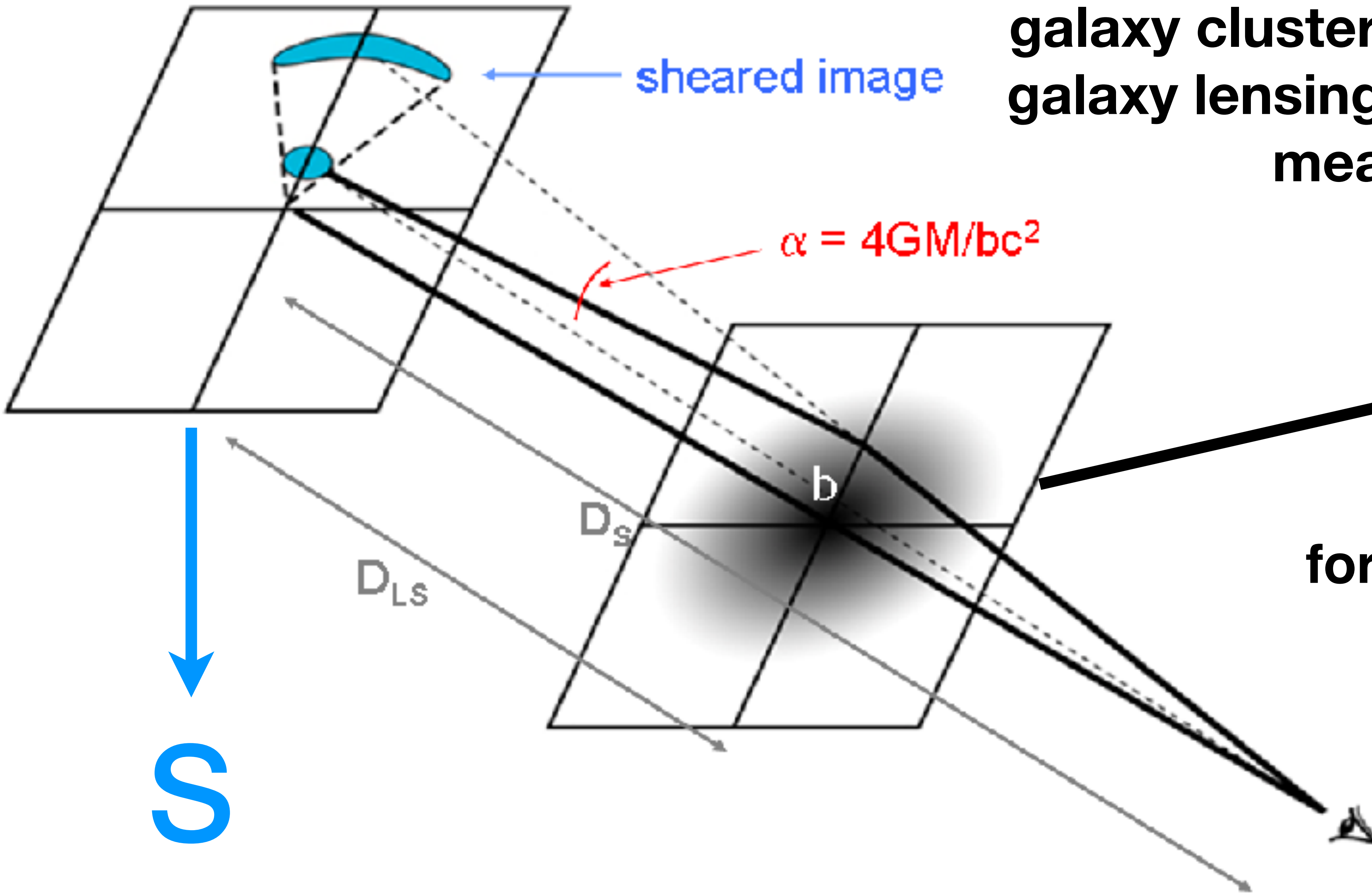
- Cons:
  - Lots of systematics! atmospheric distortions and "rounding" of shapes; intrinsic alignments, etc etc.

Galaxy-galaxy lensing

- Around each (<u>foreground</u>) galaxy, add up tangential shear of <u>background</u> galaxies seen around it
- Should really be called galaxy-galax*ies* lensing
- Then stack signal of many such foreground galaxies
- Probes relatively small scales (~0.1 to ~10 Mpc)
- Much easier to do than shear-shear weak lensing: higher signal-to-noise, fewer systematics
- Challenge: modeling theory (clustering - recall, this includes bias) at small scales

Combining cosmic shear (ss), galaxy clustering(gg), and galaxy-galaxy lensing (gs) into “3x2” type measurement

image: LSST science book



g

(**positions** of **foreground** galaxies; 4 redshift bins)

s

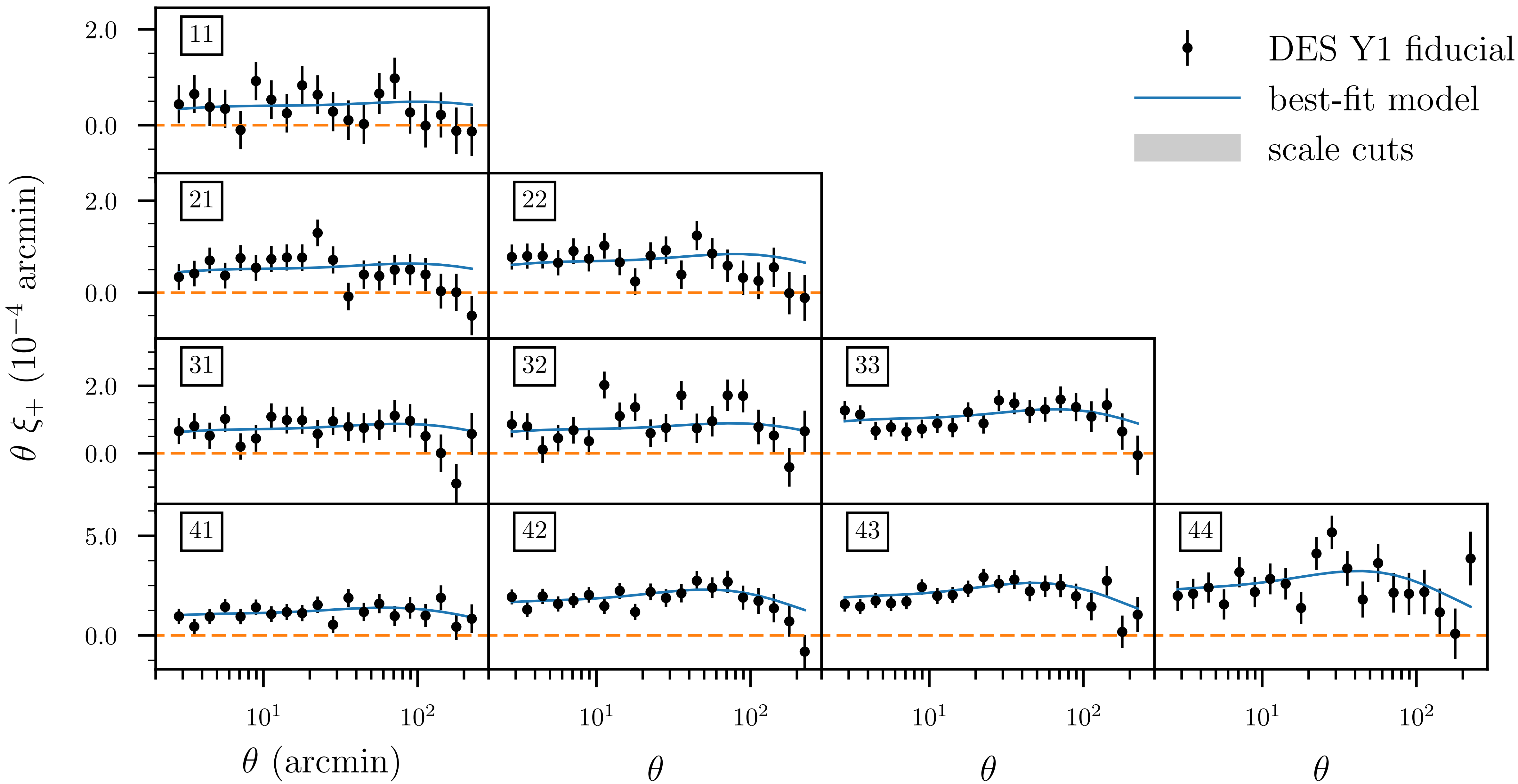
(**shear** of **background** galaxies; 5 redshift bins)

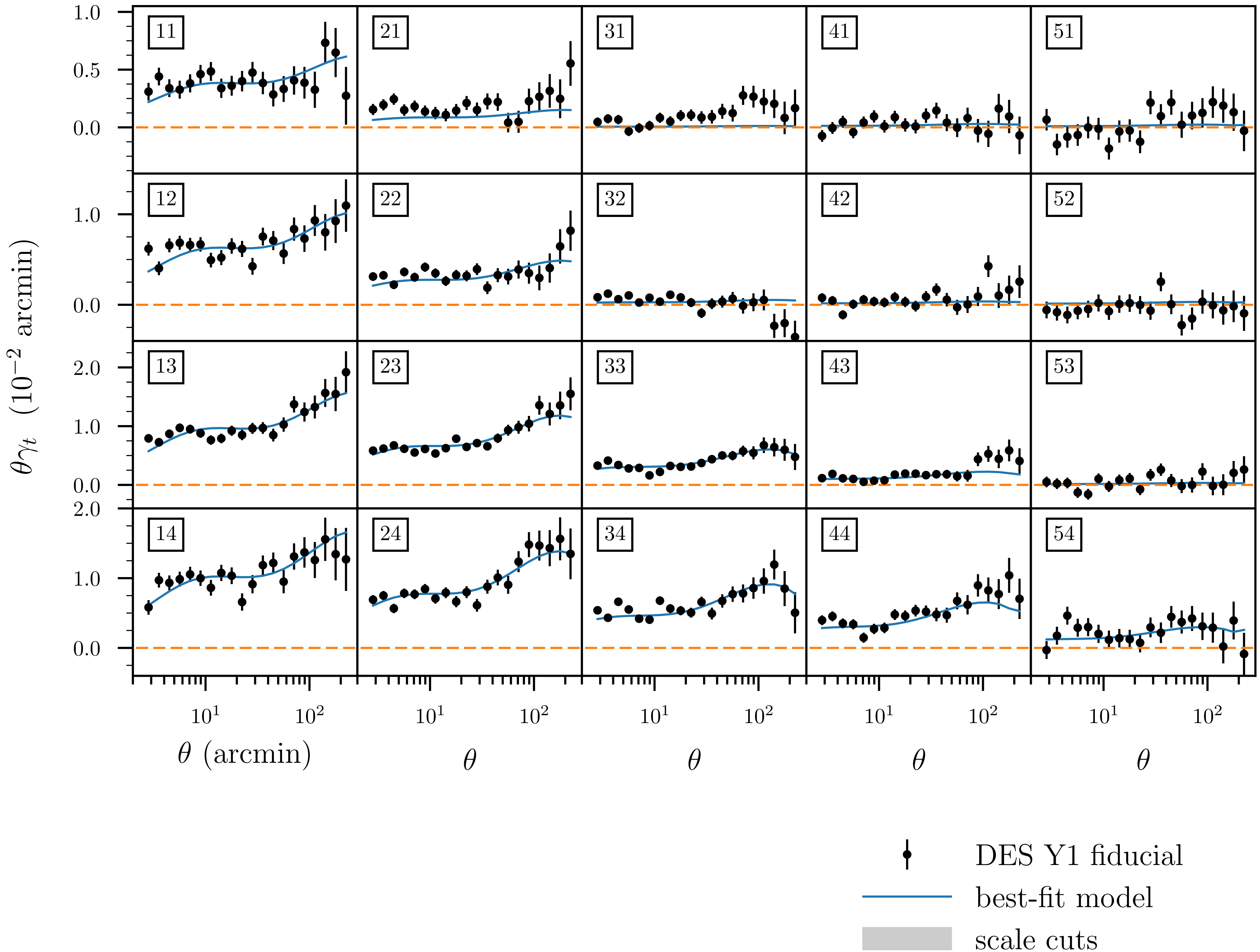
“3x2 (point-function)” clustering measurements:

$$\begin{bmatrix} gg & gs \\ gs & ss \end{bmatrix}$$

DES Y1 Measurements:
shear clustering, galaxy-galaxy lensing, gal clustering

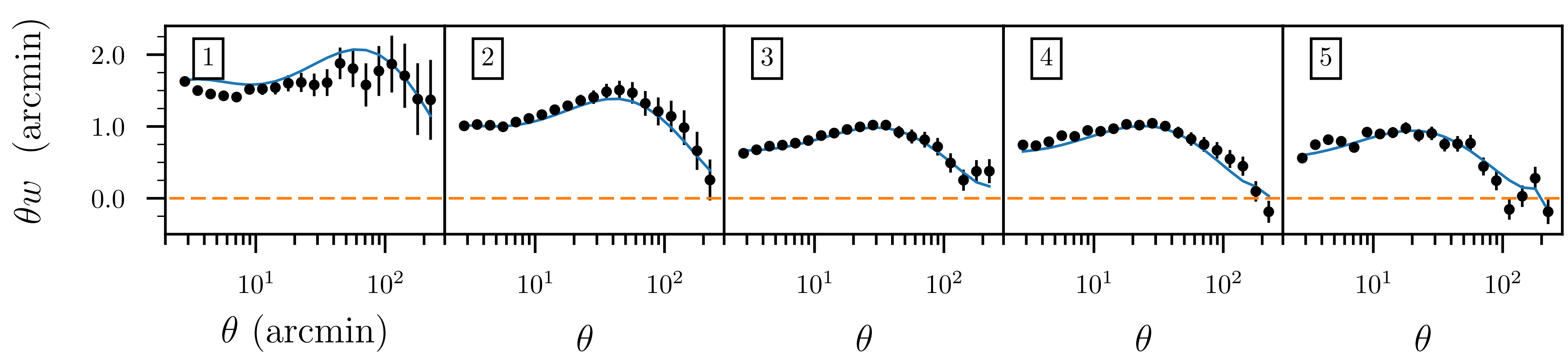
Shear clustering:





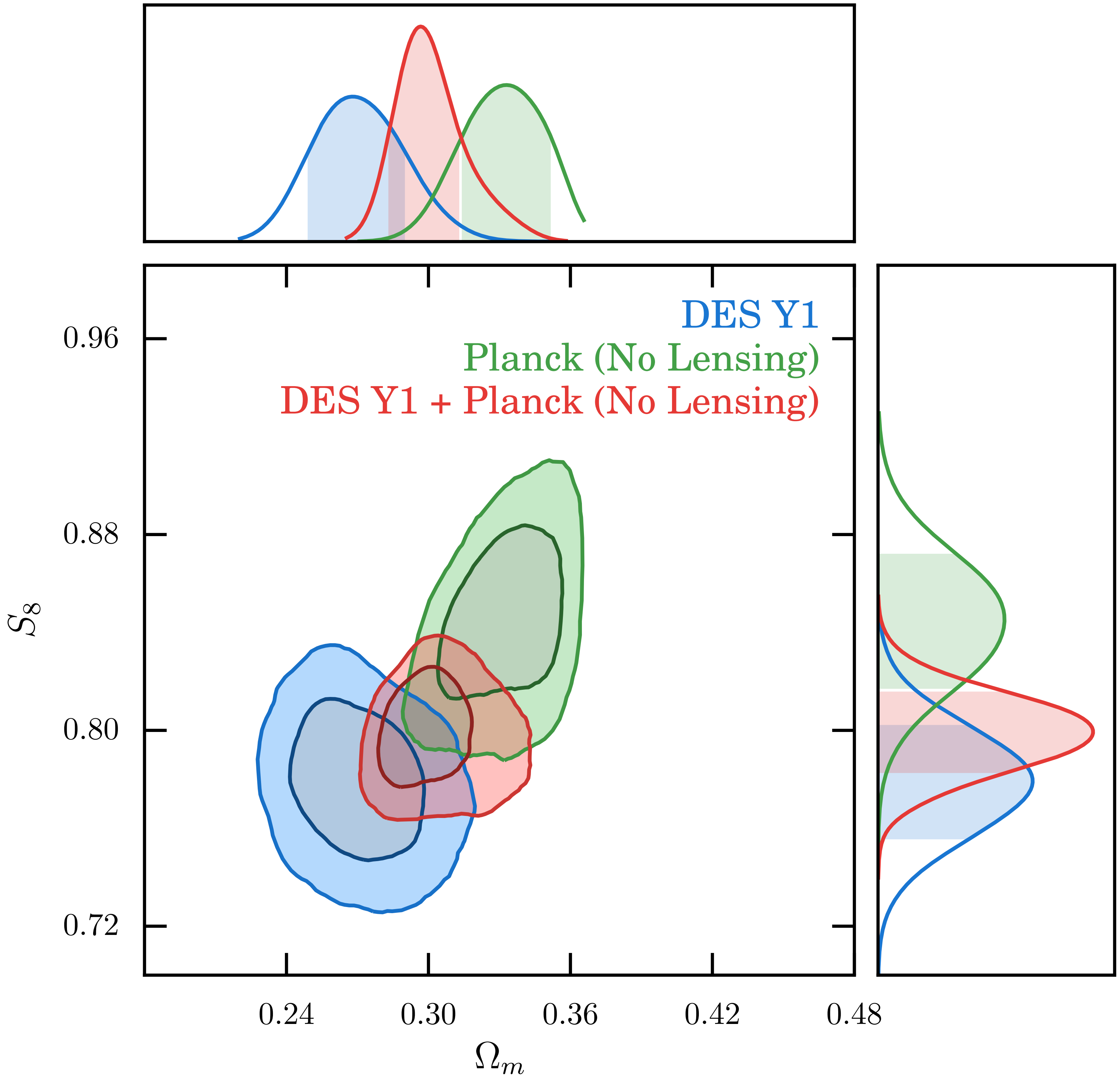
Shear-galaxy correlations

(“galaxy-galaxy lensing”)



Galaxy clustering

DES 3x2 results: Ω_m-S_8 plane



$$S_8 \equiv \sigma_8 \left(\frac{\Omega_M}{0.3}\right)^{0.5}$$

$$\Omega_m = 0.267^{+0.030}_{-0.017}$$

$$S_8 = 0.773^{+0.026}_{-0.020}$$

Abbott et al, arXiv:1708.01530

Introduction to Statistics in Cosmology

Dragan Huterer
ICTP Trieste/SAIFR Cosmology School
Jan 18-29, 2021

Basic statistics

Let P(X) be probability (likelihood) of some random (stochastic) variable X. Then:

$$P(X) \geq 0$$ non-negativity

$$\int_{-\infty}^{\infty} P(X)dX = 1$$ normalization

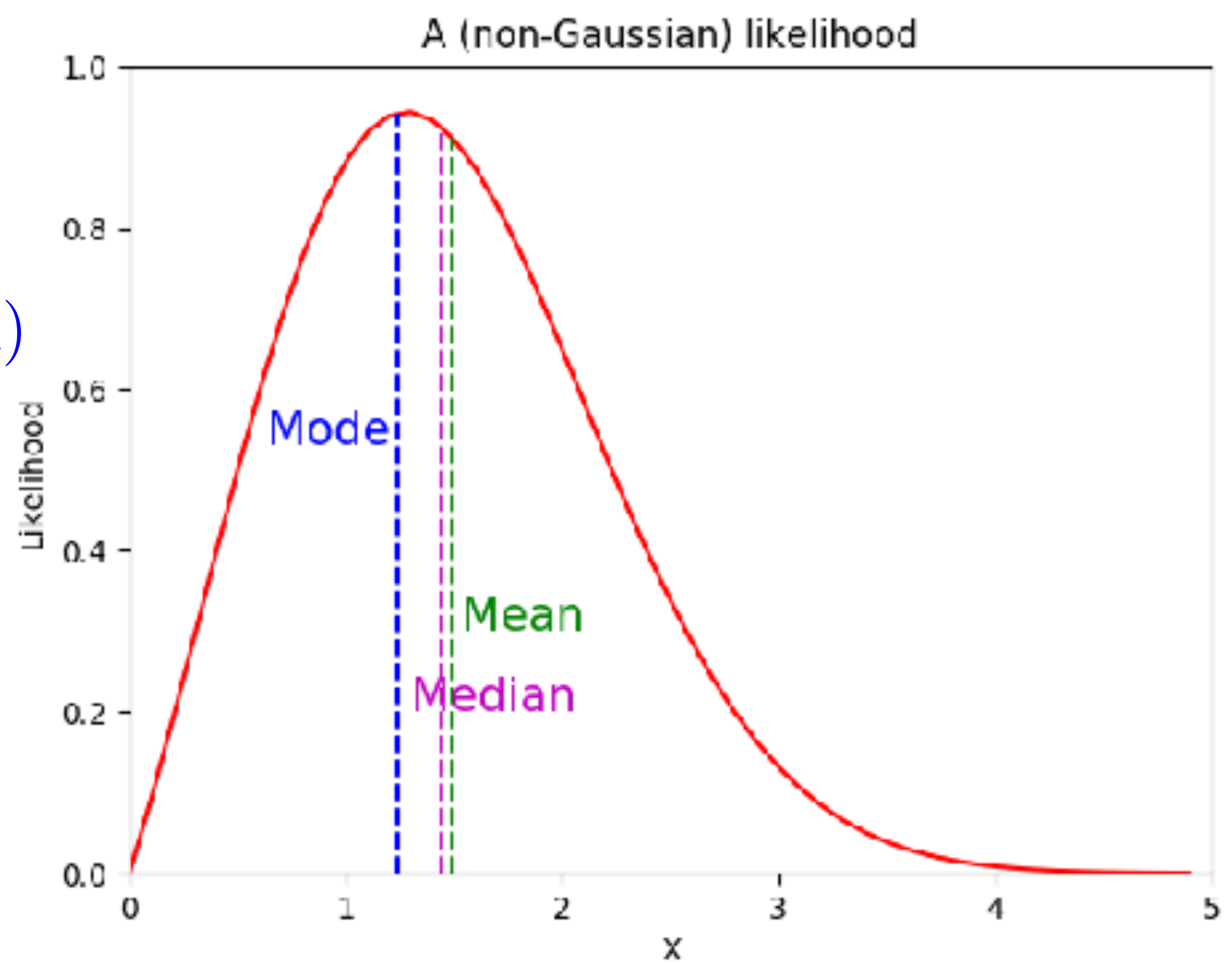
$$P(X_2) = \int P(X_1, X_2)P(X_1)dX_1$$ marginalization (over X_1)

Lowest moment is the mean:

$$\mu \equiv \bar{X} \equiv \langle X \rangle = \int_{-\infty}^{\infty} X P(X)\, dX \quad \text{(mean)}$$

$$\frac{1}{2} = \int_{-\infty}^{X_{\text{median}}} P(X)\, dX \quad \text{(median)}$$

$$\left.\frac{dP}{dX}\right|_{\mathrm{x}_{\text{mode}}} = 0 \quad \text{(mode)}$$



Basic statistics

Variance (2nd moment):

$$\mathrm{Var}(X) \equiv \sigma^2 \equiv \langle (X-\mu)^2 \rangle = \int_{-\infty}^{\infty} (X-\mu)^2 P(X)\, dX \quad \text{(variance)}$$

measures the width (squared) of the distribution

Higher moments:

$$S \equiv \left\langle \left(\frac{X-\mu}{\sigma} \right)^3 \right\rangle \quad \text{(skewness)}$$

measures the asymmetry (the "skew")

$$K \equiv \left\langle \left(\frac{X-\mu}{\sigma} \right)^4 \right\rangle \quad \text{(kurtosis)}$$

measures the peakedness (the "heavy tails")

Estimators

Given N realizations (draws) of some random variable X, can you estimate the properties of the distribution of X?

Example, to find the mean, a good estimator is: $$\hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Example, to find the variance, use: $$\widehat{\mathrm{Var}}(X) = \frac{\sum_{i=1}^{N}(x_i - \hat{\mu})^2}{N-1}.$$

Good properties of estimators:
1. unbiased
2. as minimal variance as possible

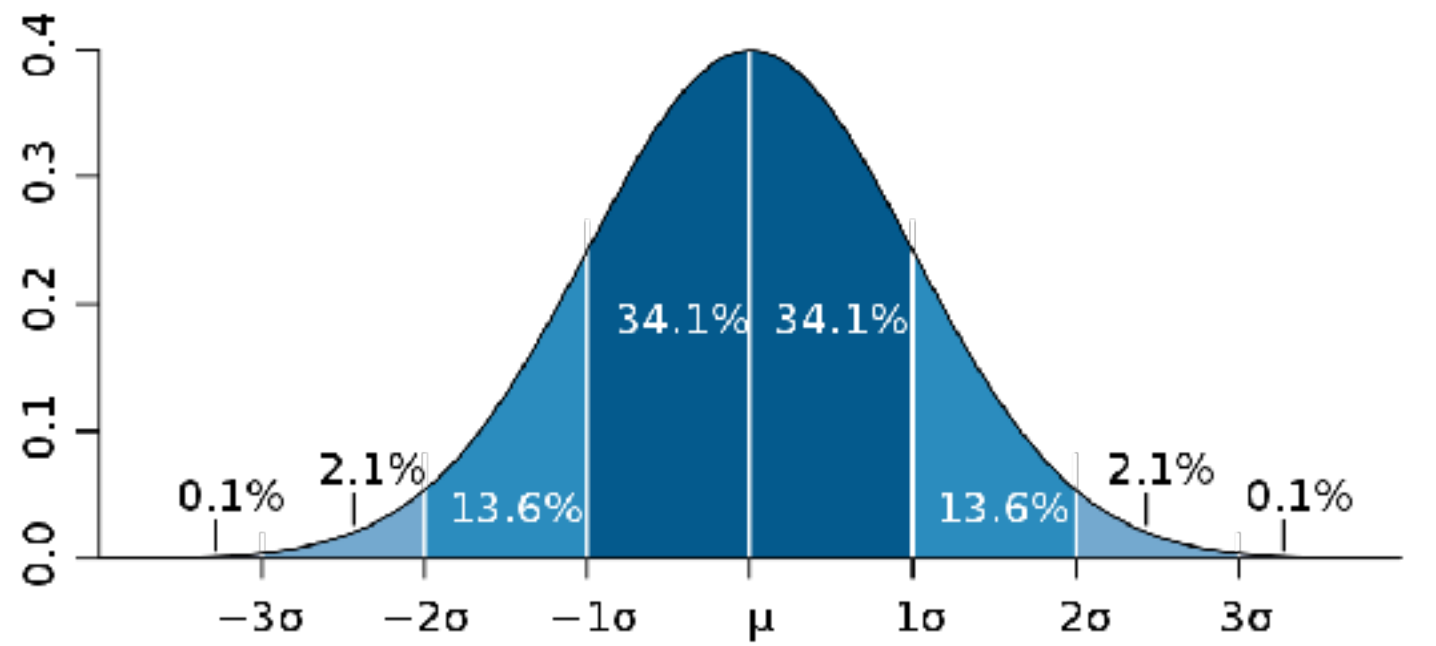
$\Longrightarrow$ hence one really wants to find+use a BUE (Best Unbiased Estimator)

Most often we want a full Bayesian posterior distribution on cosmological parameters, but

- sometimes we prefer to produce an estimator of a parameter, esp if it's super time-consuming to calculate it from a map (e.g. f_{NL} which is calculated from the 3-pt function in CMB)

Gaussian distribution

characterized by:
- mean μ (or vector of μs)
- variance σ^2 (or cov. C)



1Dim:

$$P(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right]$$

nDim:

$$P(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\det C|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T C^{-1}(\mathbf{x}-\mu)\right]$$

By FAR the most useful, simple, convenient distribution in cosmology. Notably:

Simplest* inflation predicts - and measurements so far indicate that at large scales, L ≫ 10 h^{-1}Mpc (or, equivalently, early times like the CMB) **the universe is Gaussian to 1 part in 10,000 (!!!); $f_{NL} \lesssim 5$**

Holy Grail for DESI, Euclid, WFIRST etc: is it Gaussian to 1 part in 100,000??

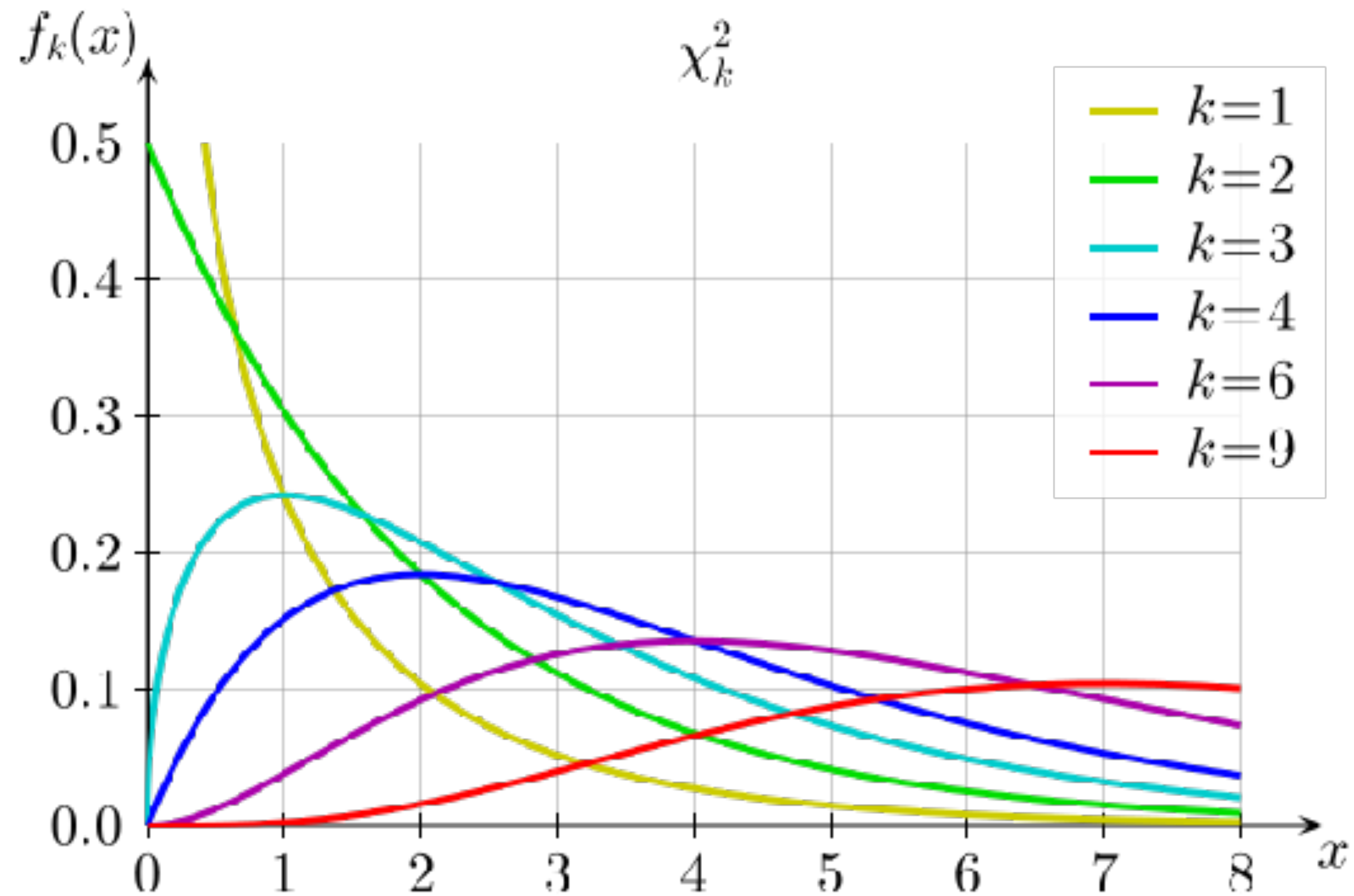
*Single scalar field, always slow-rolls, in Einstein gravity….

Chi squared distribution

If you add squares of k Gaussian variables, you get a chi-squared distribution with k d.o.f.

$$Y = X_1^2 + X_2^2 + \ldots + X_k^2$$

$$P(Y) = \frac{1}{2^{k/2}\Gamma(k/2)} Y^{k/2-1} e^{-Y/2}$$



Important properties:

- mean is k
- variance is 2k
- for k≫1, looks like Gaussian!

1. When X_i are Gaussian-distributed, then

$$\mathcal{L} \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{k}\left(\frac{X_i - \mu_i}{\sigma_i}\right)^2\right] \equiv \exp\left[-\chi^2/2\right]$$

Simplest goodness-of-fit metric: χ^2/dof ≈1 where dof = k −$N_{\text{params-fit}}$

2. If the density field δ is Gaussian-distributed, then the power spectrum is chisq-distributed;
in the CMB: $a_{\ell m}$ are Gaussian, then each C_ℓ is chisq-distributed with d.o.f. = $2\ell+1$

Likelihoods

The one we love the most is: $$\mathcal{L}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\det C|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T C^{-1}(\mathbf{x}-\mu)\right]$$

$$\equiv \frac{1}{(2\pi)^{n/2}|\det C|^{1/2}} \exp\left[-\chi^2/2\right]$$

"It's Gaussian…" but whose likelihood is Gaussian (or non-G or whatever)?
What is **x**?

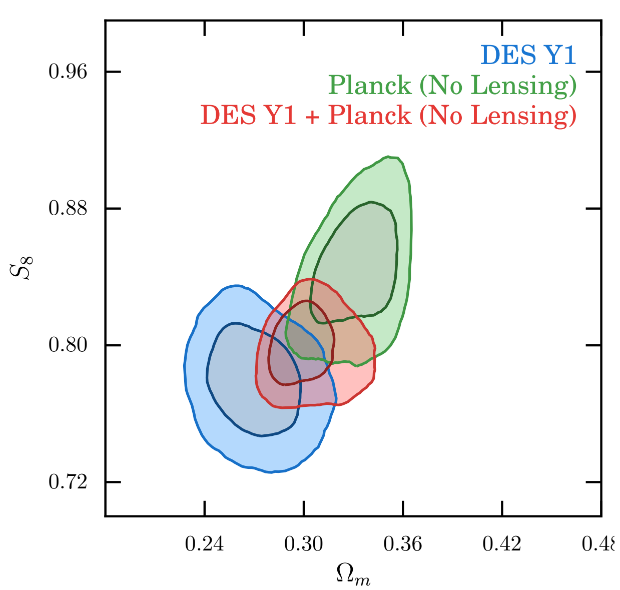
*to a theorist

"Theory"
x = Cosmological parameters **p**
(e.g. Ω_Λ, m_ν, σ_8, etc)

OR

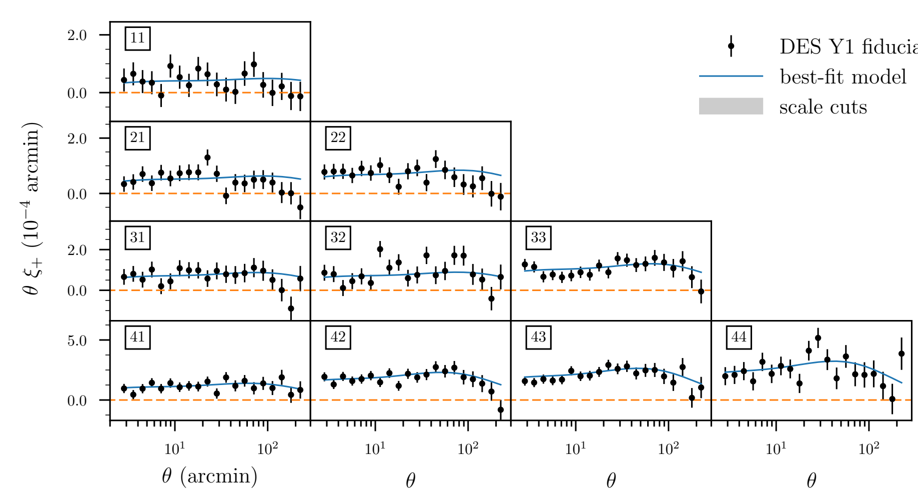
"Data*"
x = observable* quantities **d**
(e.g. P(k), ξ(r), dn/dlnM(z), etc)

$\mathcal{L}$ is simply not Gaussian in most cases



DES Y1 3x2 paper, Abbott et al.

Usually ok to assume $\mathcal{L}$ is Gaussian by the Central Limit Theorem:
as $n\to\infty$, $\mathcal{L} \to$ Gaussian with $\sigma\to\sigma_i/\sqrt{n}$



DES Y1 shear paper, Troxel et al.

Bayesian and Frequentist statistics

Lakers or Celtics? Real Madrid or Barcelona? Michigan or Ohio State?
Montagues or Capulets? Rock or Classical? Brazil or Argentina??

.....

Bayesian or Frequentist???

- Frequentist: model is fixed, data are repeatable
- Bayesian: data are fixed, model is repeatable

likelihood

prior

D = data
M = model

$$\boxed{P(M|D) = \frac{P(D|M)\,P(M)}{P(D)}} \qquad (\text{Bayes}' \text{ theorem})$$

posterior

evidence

Bayesian and Frequentist statistics

- Bayesian: data are fixed, model is repeatable
- Frequentist: model is fixed, data are repeatable

Say H_0 = (72 ± 2) km/s/Mpc. Then:

Bayesian: the posterior distribution for H_0 has 68% if its integral between 70 and 74 km/s/Mpc. The posterior can be used as a prior on a new application of Bayes' theorem.

Frequentist: Performing the same procedure will cover the real value of H_0 within the limits 68% of the time. But how do I repeat the same procedure (generate a new H_0) if I only have one Universe?

Good references:

Bayesian: R. Trotta, “Bayes in the Sky”, https://arxiv.org/abs/0803.4089

Frequentist: Feldman & Cousins, “A Unified Approach to the Classical Statistical Analysis of Small Signals”, https://arxiv.org/abs/physics/9711021

Example of one cosmology inference done both Bayesian and frequentist way: G. Efstathiou, “The Statistical Significance of the Low CMB Multipoles”, https://arxiv.org/abs/astro-ph/0306431

Bayesian and Frequentist statistics

- Bayesian:
  - can given probabilities for models
  - depends on both prior and likelihood (of data)
  - currently the dominant method in cosmology
- Frequentist:
  - doesn't give probabilities of models, only of hypotheses
  - doesn't depend on prior, just likelihood
  - currently the dominant method in particle physics

likelihood

prior

D = data
M = model

$$\boxed{\underbrace{P(M|D)}_{\text{posterior}} = \frac{\overbrace{P(D|M)}^{\text{likelihood}}\,\overbrace{P(M)}^{\text{prior}}}{\underbrace{P(D)}_{\text{evidence}}}} \quad \text{(Bayes' theorem)}$$

posterior

evidence

Which credible intervals do you report?

The overwhelming convention in cosmology is to

- Report the **peak (mode) value as the best fit**. This is peak of the posterior marginalized over all other parameters
- Report the (asymmetric) ± error bars that encompass 68.3% (and 95.4% and 99.7% of posterior volume around the peak.

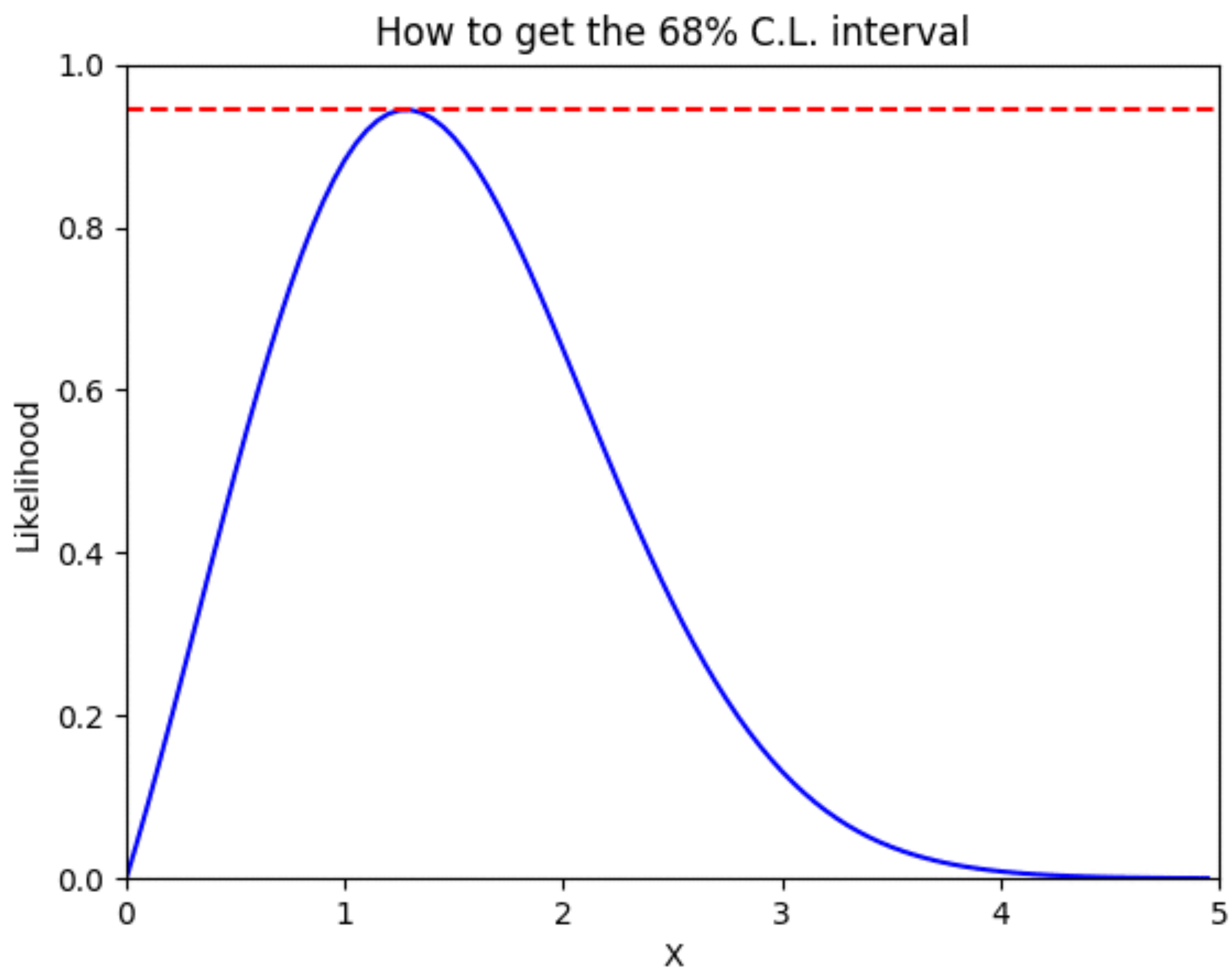
For a Gaussian distribution, $\mu\pm1\sigma$ region encompasses 68% of likelihood volume.
For a non-Gaussian it doesn't,
but we are supposed to always calculate the latter (68%) even
when we lazily speak about the former
that is, in general you <u>don't</u> quote "sigma" (error) by calculating sqrt(variance).

So… how exactly DO you get the 68.3% (and 95.4% and 99.7%) region?
Let's be super explicit!

How to calculate confidence level

Start from the peak of the posterior

"lower the water level" until you encompass 68% of the likelihood volume



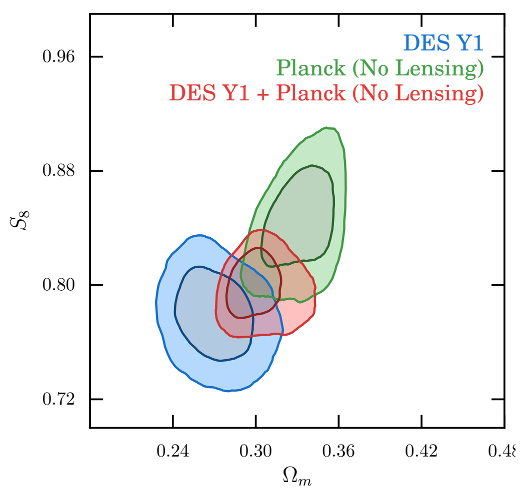
Fisher information matrix

*for a theorist

“Theory”

x = Cosmological parameters **p**

(e.g. Ω_Λ, m_ν, σ_8, etc)



DES Y1 3x2 paper,
Abbott et al.

…what are errors in theory parameters?

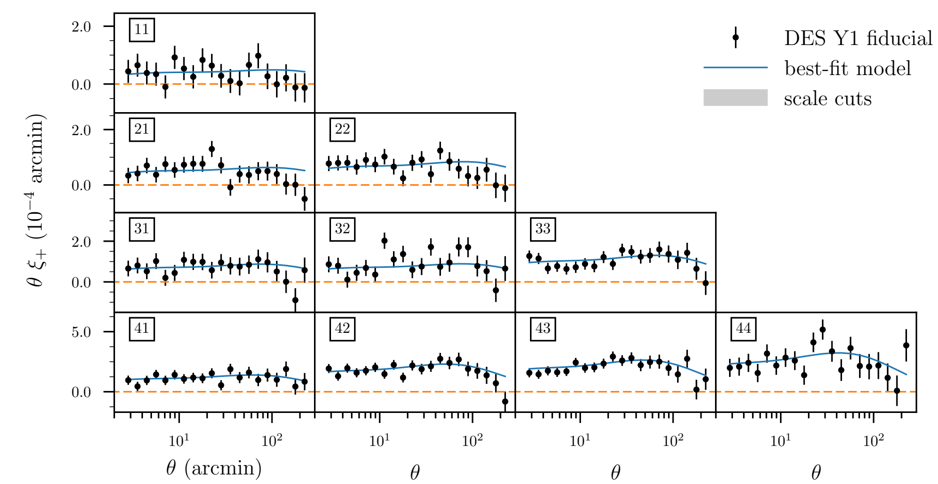
OR

⇔

“Data*”

x = observable* quantities **d**

(e.g. P(k), ξ(r), dn/dlnM(z), etc)



DES Y1 shear paper,
Troxel et al.

Given data error bars…

Normally this requires simulations (Monte-Carlos) to evaluate;
very time-consuming <u>and noisy</u>

Fisher matrix is a semi-analytical tool that gives answers instantaneously, and without stochastic noise.

Fisher information matrix

$$\boxed{F_{ij} = \left\langle -\frac{\partial^2 \ln \mathcal{L}}{\partial p_i \partial p_j} \right\rangle}$$

It's the curvature matrix (negative Hessian) of negative log likelihood around its peak

Step 1:

For a Gaussian likelihood (in parameters p_i) this evaluates to (Tegmark, Taylor, Heavens 1997)

$$\boxed{F_{ij} = \mu_{,i}^T C^{-1} \mu_{,j} + \frac{1}{2}\mathrm{Tr}[C^{-1} C_{,i} C^{-1} C_{,j}]}$$

If the *mean* of the data depends on p_i, then first term is nonzero
If the covariance of the data depends on p_i, then second term is nonzero

⇒ For clustering statistics in LSS, typically the latter, as $\mu=\langle\delta\rangle=0$, while C = P(k) or ξ(r)

Step 2: Then, the Cramer-Rao inequality says:

$$\boxed{\sigma(p_i) \geq \begin{cases} \sqrt{(F^{-1})_{ii}} & \text{(marginalized)} \\ 1/\sqrt{F_{ii}} & \text{(unmarginalized)} \end{cases}}$$

Fisher information matrix

$$F_{ij} = \mu^T_{,i} C^{-1} \mu_{,j} + \frac{1}{2} \mathrm{Tr}[C^{-1} C_{,i} C^{-1} C_{,j}]$$

$$\sigma(p_i) \geq \begin{cases} \sqrt{(F^{-1})_{ii}} & \text{(marginalized)} \\ 1/\sqrt{F_{ii}} & \text{(unmarginalized)} \end{cases}$$

Couple of examples:

1. Type Ia supernovae, where μ = m(z, Ω_M, Ω_Λ…) = m(z_n, $\{p_i\}$), then

$$F^{\mathrm{SNe}}_{ij} = \sum_{n=1}^{N_{\mathrm{SNe}}} \frac{1}{\sigma_m^2} \frac{\partial m(z_n)}{\partial p_i} \frac{\partial m(z_n)}{\partial p_j}$$

(assuming uncorrelated errors that don't depend on p_i)

2. Weak gravitational lensing, where μ = 0 but C = shear power spectrum, then

$$F^{\mathrm{WL}}_{ij} = \sum_{\ell} \frac{\partial C^{\kappa}(\ell)}{\partial p_i} \mathbf{Cov}^{-1} \frac{\partial C^{\kappa}(\ell)}{\partial p_j}$$

where

$$\mathrm{Cov}\left[C^{\kappa}_{ab}(\ell), C^{\kappa}_{cd}(\ell)\right] = \frac{\delta_{\ell\ell'}}{(2\ell+1)\, f_{\mathrm{sky}}\, \Delta\ell} \left[C^{\kappa}_{ac}(\ell) C^{\kappa}_{bd}(\ell) + C^{\kappa}_{ad}(\ell) C^{\kappa}_{bc}(\ell)\right].$$

Fisher information matrix

How do you marginalize over some parameters? Easy!

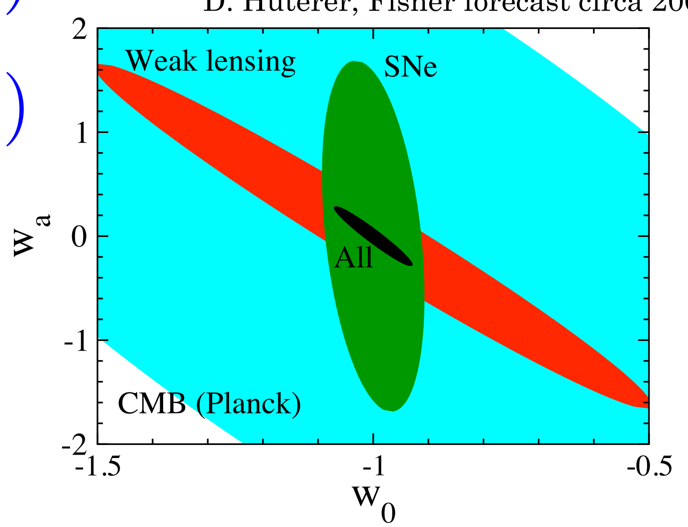
1. Calculate the original NxN Fisher matrix F
2. Take its inverse, get F^{-1}
3. Pick a subset (e.g. 2x2 submatrix), call it G^{-1}
4. Invert G^{-1} to get G
5. G is your new Fisher matrix, marginalized over other params

How do you plot a contour in 2D parameter plane? Easy!

$$G_{11}p_1^2 + 2G_{12}p_1p_2 + G_{22}p_2^2 = \Delta\chi^2_{2\,\mathrm{dof}} = \begin{cases} 2.3 & (68.3\%) \\ 6.2 & (95.4\%) \\ 11.8 & (99.7\%) \end{cases}$$

It's an ellipse (always in the Fisher approximation)

D. Huterer, Fisher forecast circa 2000



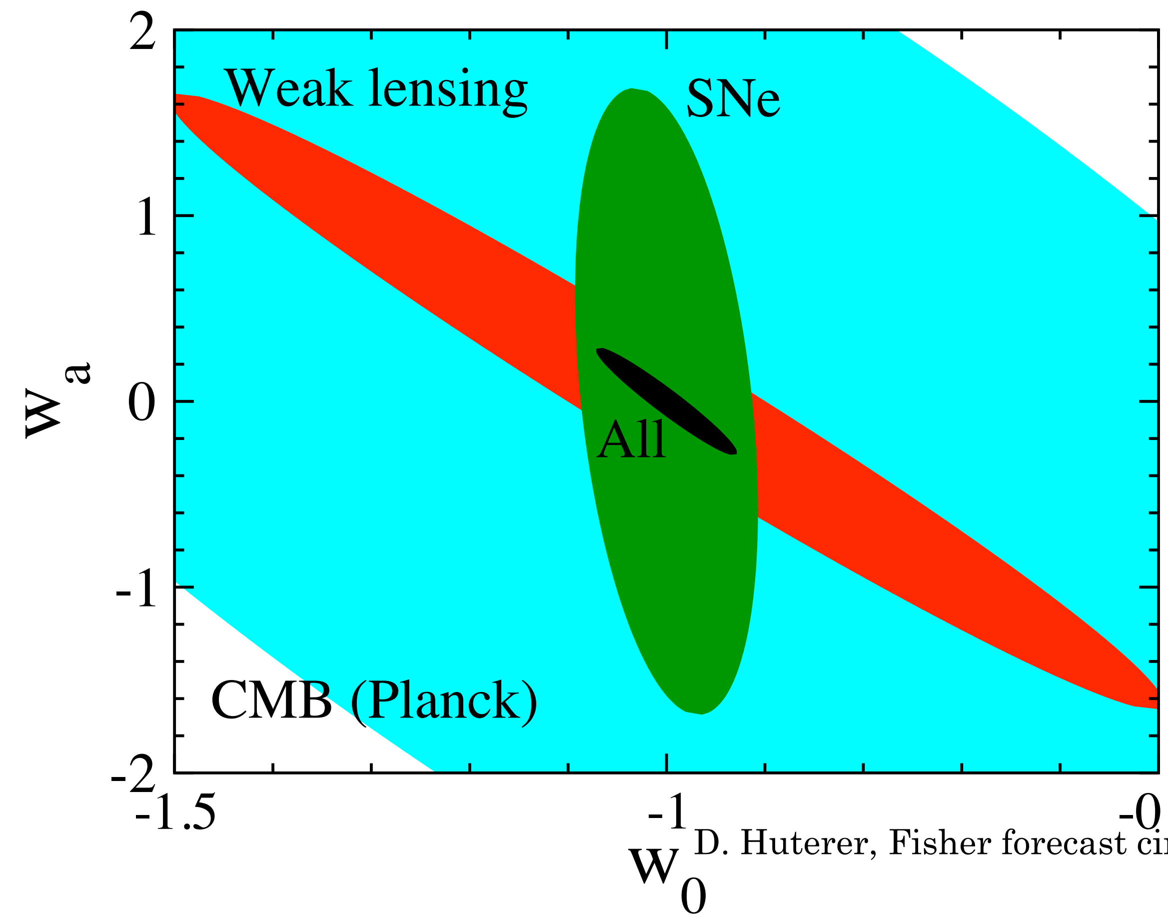
Fisher information matrix

Want area of ellipse?
$\Rightarrow A \propto (\det F)^{1/2}$

Want best-constrained directions?
$\Rightarrow$ diagonalize F

Want to add independent constraints?
$\Rightarrow$ add their Fs ($\Leftrightarrow$multiply likelihoods)

Want to add a prior on i-th parameter p_i?
$\Rightarrow$ add $1/(\sigma_{\text{prior}})^2$ to its F_{ii} (diag) element



D. Huterer, Fisher forecast circa 2000

Want to see how much the parameters shift due to a (small) shift in the data??
$\Rightarrow$ use the **Fisher bias formula**

$$\delta p_i \approx F^{-1}_{ij} \sum_{\ell,\alpha,\beta} \left[C^{\kappa}_{\alpha}(\ell) - \bar{C}^{\kappa}_{\alpha}(\ell)\right] \mathrm{Cov}^{-1}\left[\bar{C}^{\kappa}_{\alpha}(\ell), \bar{C}^{\kappa}_{\beta}(\ell)\right] \frac{\partial \bar{C}^{\kappa}_{\beta}(\ell)}{\partial p_j}$$

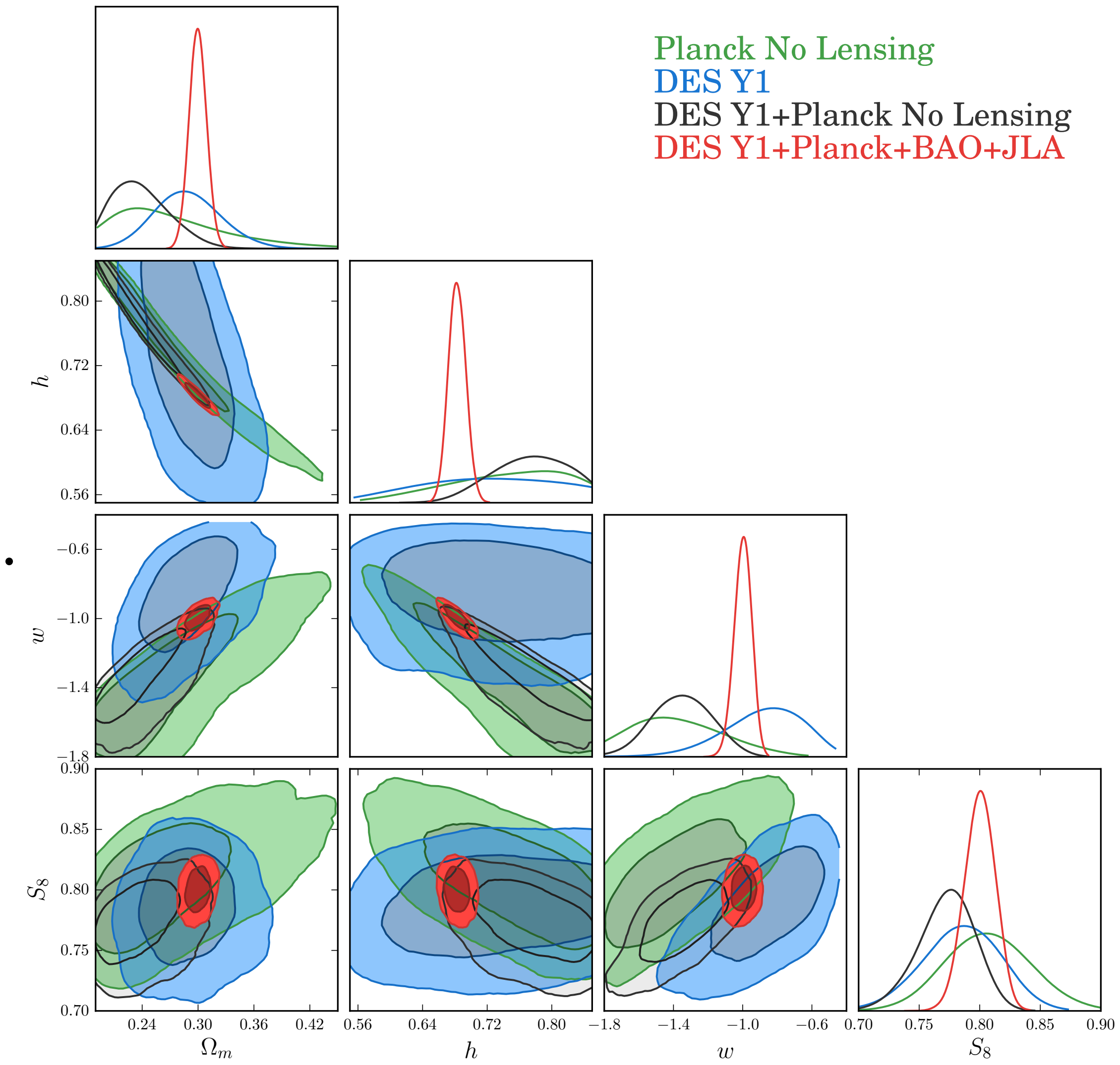
parameter bias   biased value   true (fiducial) value

Extremely useful: super fast and not subject to stochastic noise

Markov Chain Monte Carlo (MCMC)

The challenge: map out a posterior in multi-dimensional parameter space.

Example: say there are just 10 parameters.
Lets say calculation takes just 1 second/model.
Say you want a grid with 20 values in each par.
Then
$N = 20^{10} \simeq 10^{13}$
$\Rightarrow$ it would take 300,000 years to do it!
$\Rightarrow$ Totally impossible, ever!!



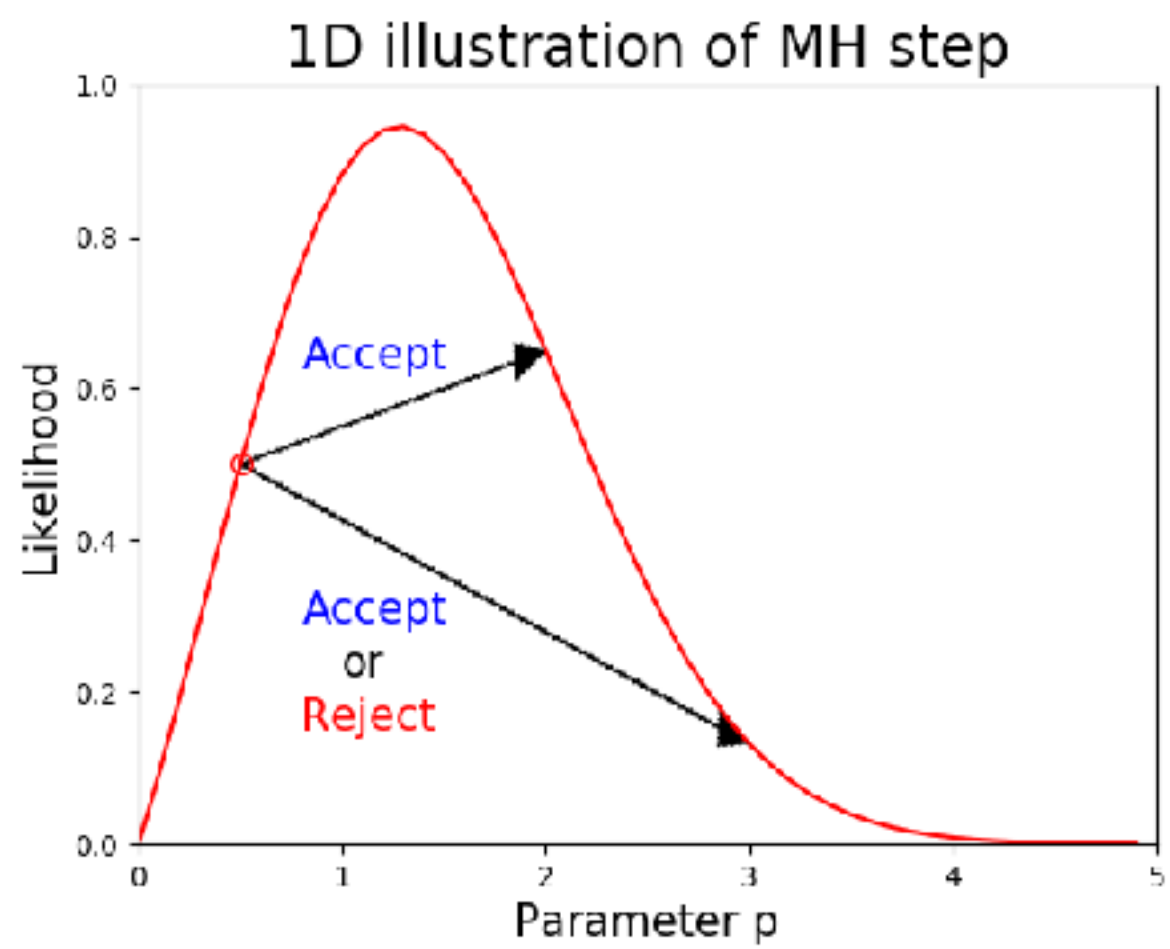
DES Y1 extensions paper (Abbot et al 2019);
the full param-space is 25-dimensional!

Amazingly clever, efficient solution to the problem:
Instead of gridding, <u>sample</u>!
"Walk" through the parameter space in a clever way in order to map out the likelihood "banana" just enough.

$\Rightarrow$ MCMC, invented at Los Alamos National Lab in 1950s.

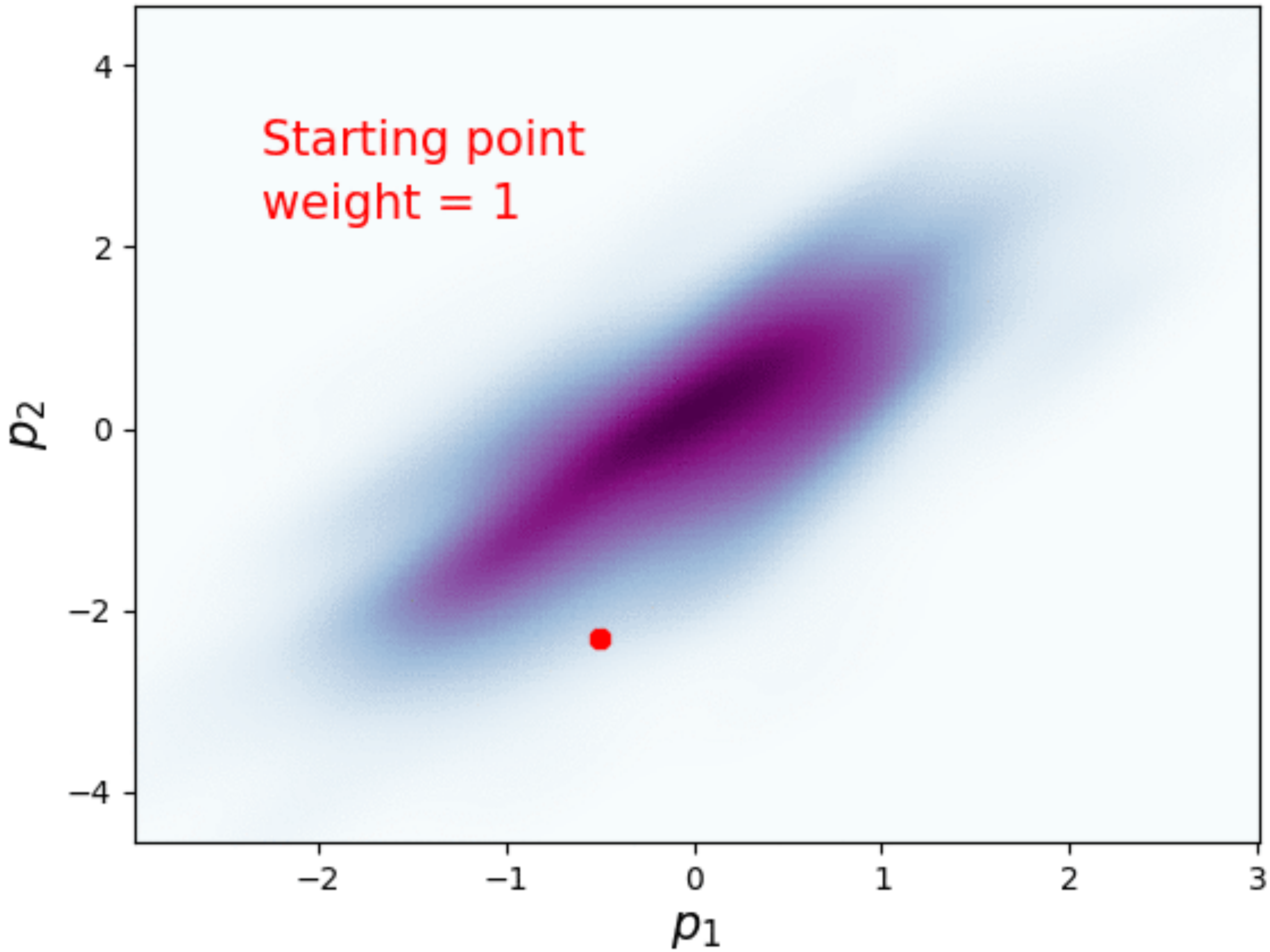
MCMC:
the Metropolis-Hastings algorithm

- at step t, at some parameters p_t
- propose move to $p_t' = p_t + \Delta p_t$ (randomly draw Δp_t)
- evaluate $r = L(p_t')/L(p_t)$
- MH step:
  - if r > 1 accept move
  - if r < 1 generate a random number $\alpha \in [0, 1]$
    - if $\alpha < r$, accept move
    - if $\alpha > r$, reject move
- t=t+1



One can prove that,
with this rule,
one asymptotically recovers the
true posterior

Illustration of the Metropolis-Hastings algorithm



MCMC: interpreting the output

WEIGHT	P_1	P_2	P_3		P_N
5	0.2	-0.3	0.15	...	2.8
1	-0.7	0.4	0.12	...	3.5
12	0.7	0.1	0.19	...	1.7
...	...	...	...	...	...
...	...	...	...	...	...

(~ MILLION ROWS)

To get the posterior probability,
simply histogram the parameter values vs weights - this is your posterior!

Want to look at posterior in p_3 marginalized over all other parameters?
Simply plot histogram of p_3 values vs weight (eaaasy!)

MCMC is an incredibly clever, powerful set of algorithms without which data-driven cosmology wouldn't have gotten far.

Suggested further reading

“Statistics in theory and practice", book by Robert Lupton

“Numerical Recipes - the Art of Scientific Computing", Press, Teukolsky, Vetterling & Flannery

“A practical guide to Basic Statistical Techniques for Data Analysis in Cosmology", L. Verde, arXiv:0712.3028, and ``Statistical methods in Cosmology", arXiv:0911:3105

"Unified approach to the classical statistical analysis of small signals", G.J. Feldman and R.D. Cousins, PRD, 57, 3873 (1998)

“Bayes in the sky: Bayesian inference and model selection in cosmology", R. Trotta, arXiv:0803.4089

Wikipedia - really good for looking up properties of functions, distributions, and other “math”.