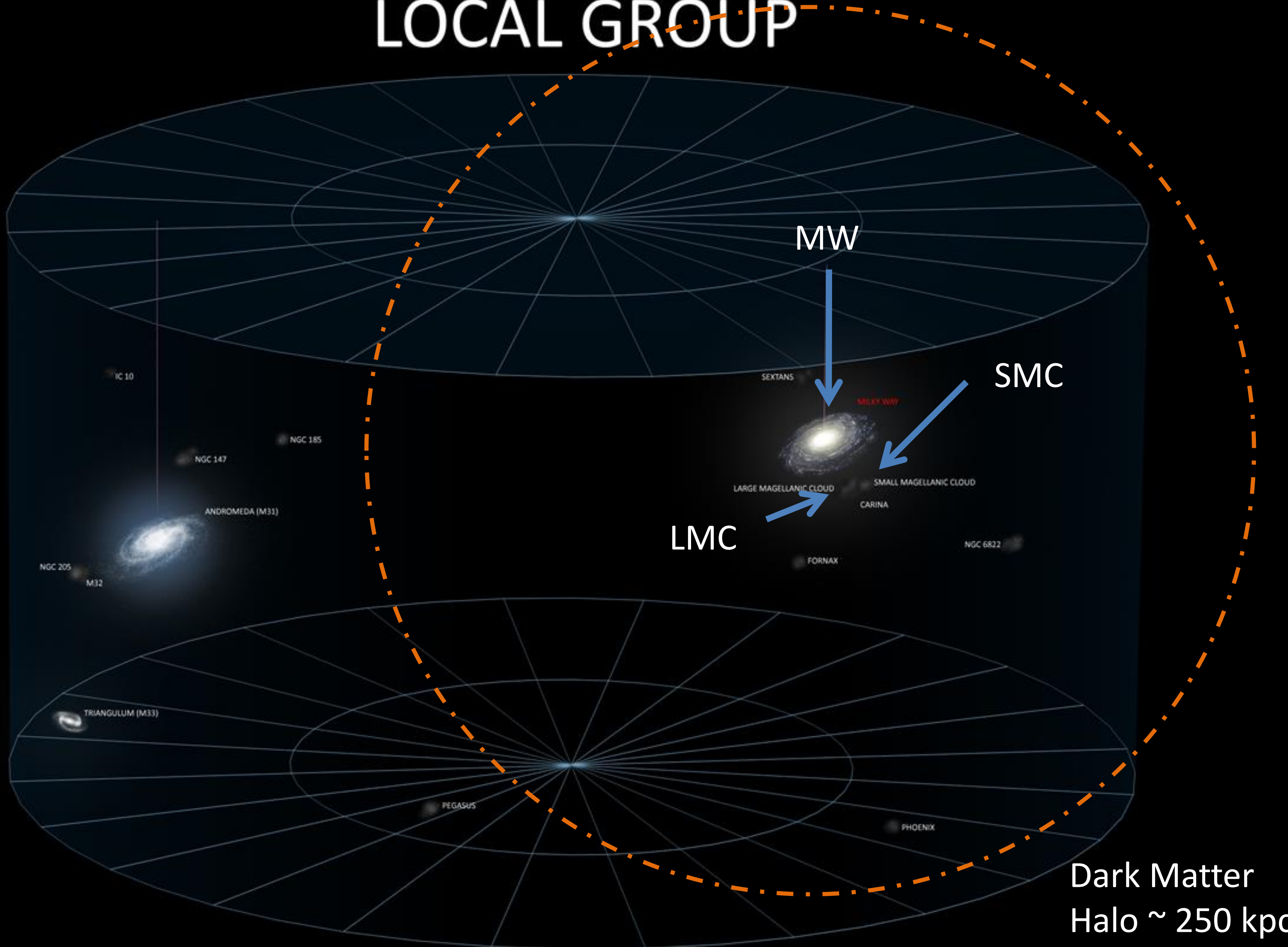
How Common are the Magellanic Clouds?

A presentation on
statistical methods in cosmology

Lulu Liu
Stanford University
Kavli Institute of Particle Astrophysics and Cosmology

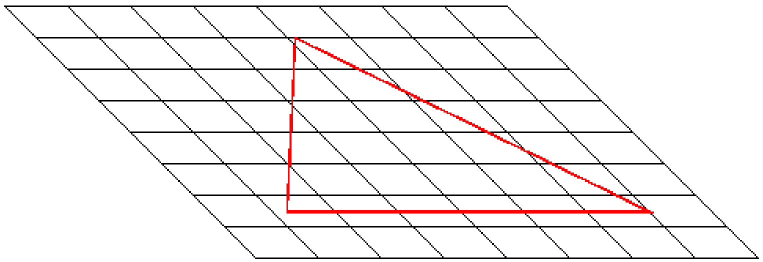
Skies over Chile– the Large and Small Magellanic Clouds

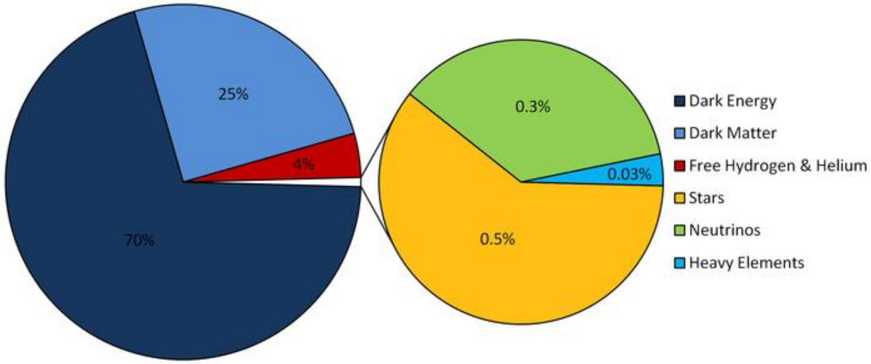
LOCAL GROUP



LMC and SMC are satellites of the Milky Way

Two too many?





- ❖LCDM: 70% DE 30% M (25% CDM) → Flat
- ❖ N-body simulations of galaxy clustering and evolution
- ❖ But simulations are not reproducing LMC/SMC, they are too bright and too close– their presence rather rare
- ❖ Is our Milky Way special or is our physics wrong?
- ❖ Look beyond MW.

Sloan Digital Sky Survey

- 100 million objects observed.
- Spectra taken for 1 million objects (redshifts known → distances known).
- Aims to map 25% of sky.
- Started in 2000 (dr7).

What’s the problem?

Traditionally:

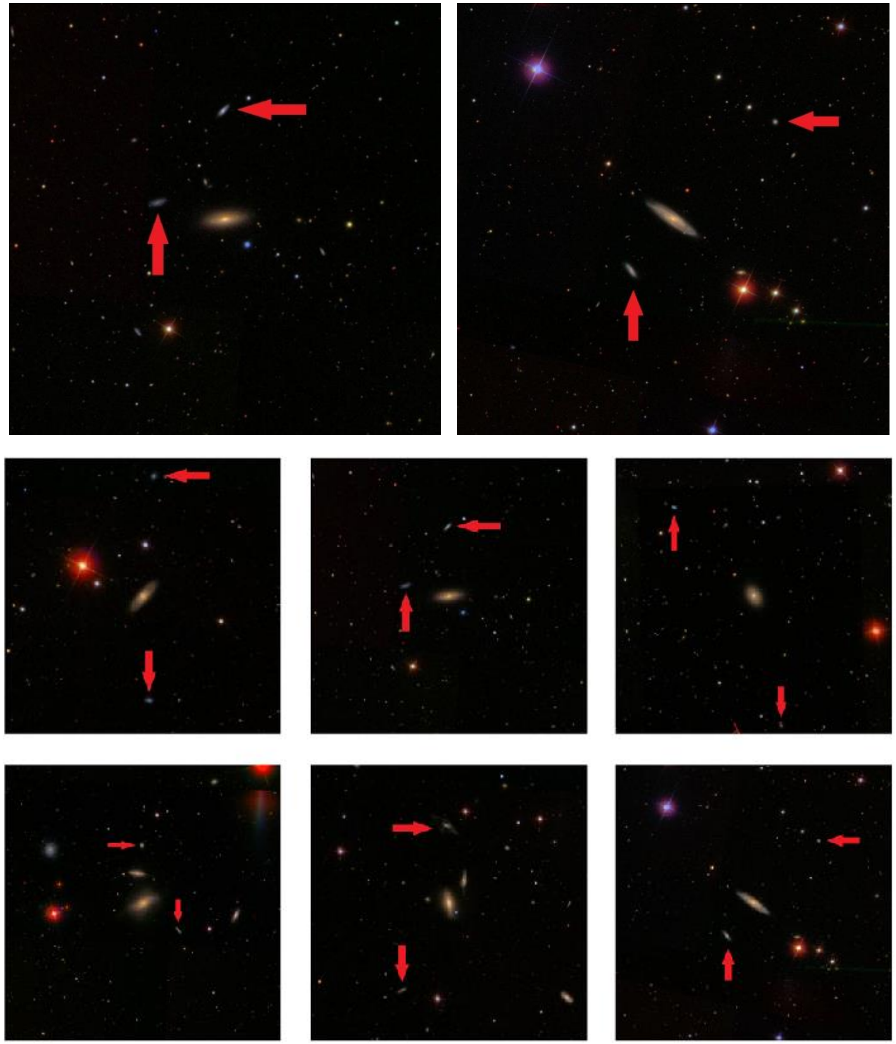
- work hard to estimate redshifts of each galaxy – best you can do is probabilities for each object, one by one analysis.
- you would constrain yourself to only objects bright enough to have complete spectra taken (sample size = abysmal)
- resign to over-counting (cylinder projection)

0.003 sq. deg.

Problem of
Signal and Noise

Why Work on Ensemble Level?

- Surrender the ability to individually identify most likely satellites →
- Can harness full statistical power of SDSS
- Precise enough measurements, well defined errors– can apply mathematical corrections
- Reproducible and widely applicable



Outline

- Question: What is the probability for a MW-sized galaxy to host S=0,1,2,… MC-like satellites?

- Methods
- Statistical Errors
- Ensemble Systematics
- Results & Simulations
- Conclusions

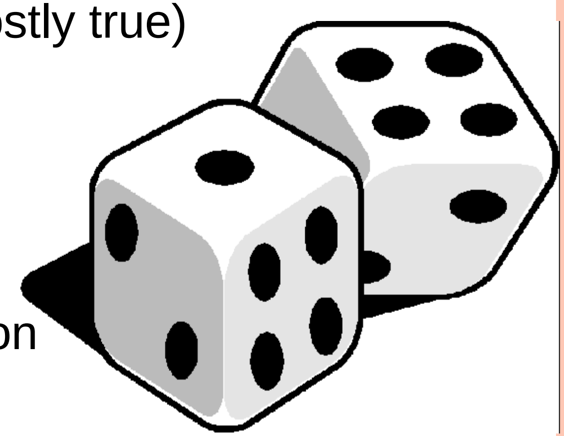
What is signal and what is noise?

- S = number of actual satellites within a predefined physical distance of host
- N = number of foreground/background galaxies projected into search aperture, assumed isotropic
- T = number of total objects which show up in a given search aperture around a MW-sized host (signal and noise)

$$T = N + S$$

LITTLE BIT OF THEORY

Granted that N and S are independent (and this is mostly true)



P(T)=P(N+S)=P(N) * P(S)

convolution

We're looking for P(S), so a deconvolution is necessary

☞ [P(T)]= ☞ [P(N)] ☞ [P(S)]

☞ [P(S)]= ☞ [P(T)] / ☞ [P(N)]

[P(S)]= ☞ $^{-1}$ [☞ [P(T)] / ☞ [P(N)]]

Finding P(N) and P(T) – count!

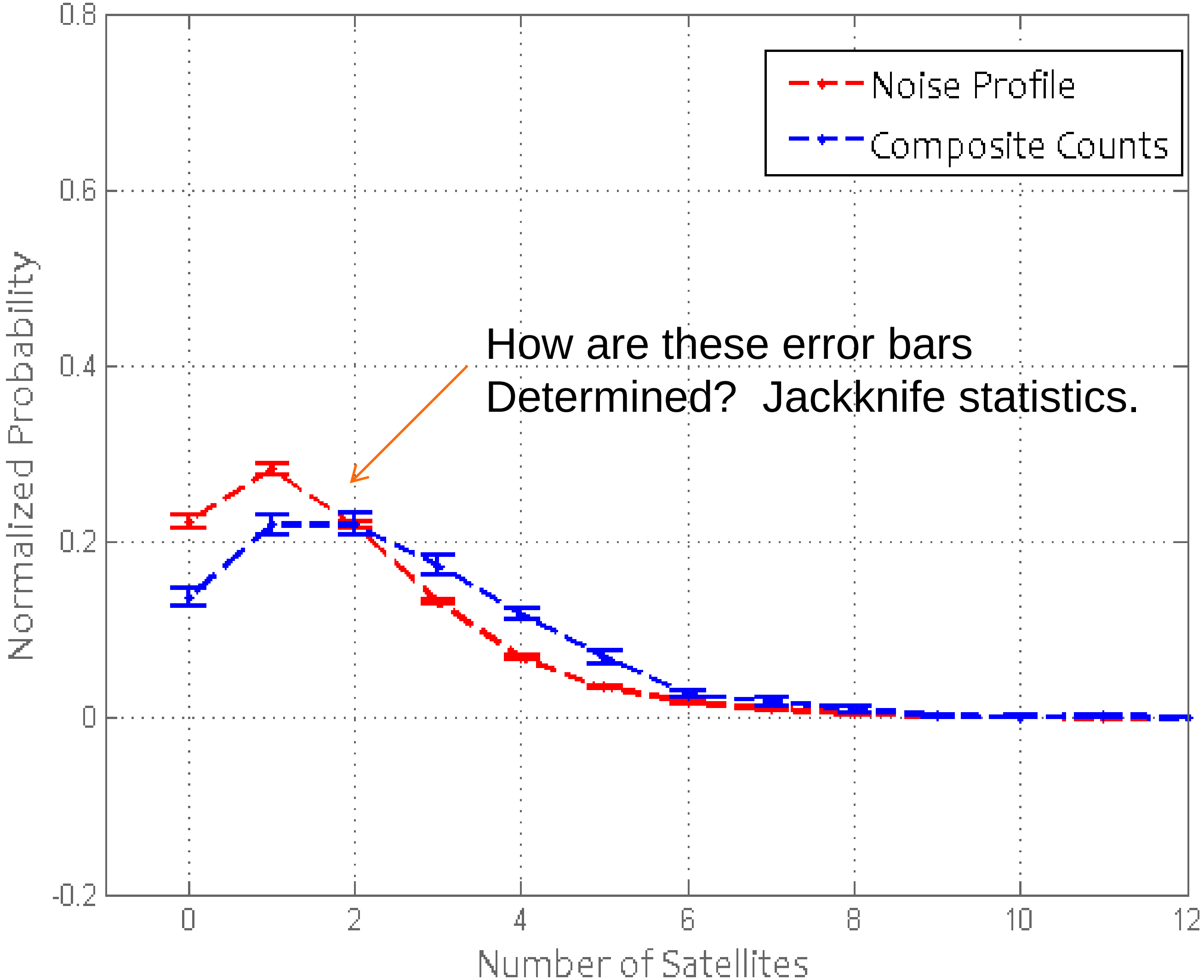
- With each host as center, draw a circle with radius corresponding to the desired physical radius – the size of this aperture (in degrees) will vary with Z

- Count every object 2-4 magnitudes dimmer than the host (this is 1/100 – 1/10,000 luminosity, range of MC's) → P(T)

- Point to random spots in the sky and do the same as if there is a host there – assume isotropic noise.

- The distribution of aperture sizes and luminosity ranges must match that of directed search → P(N)

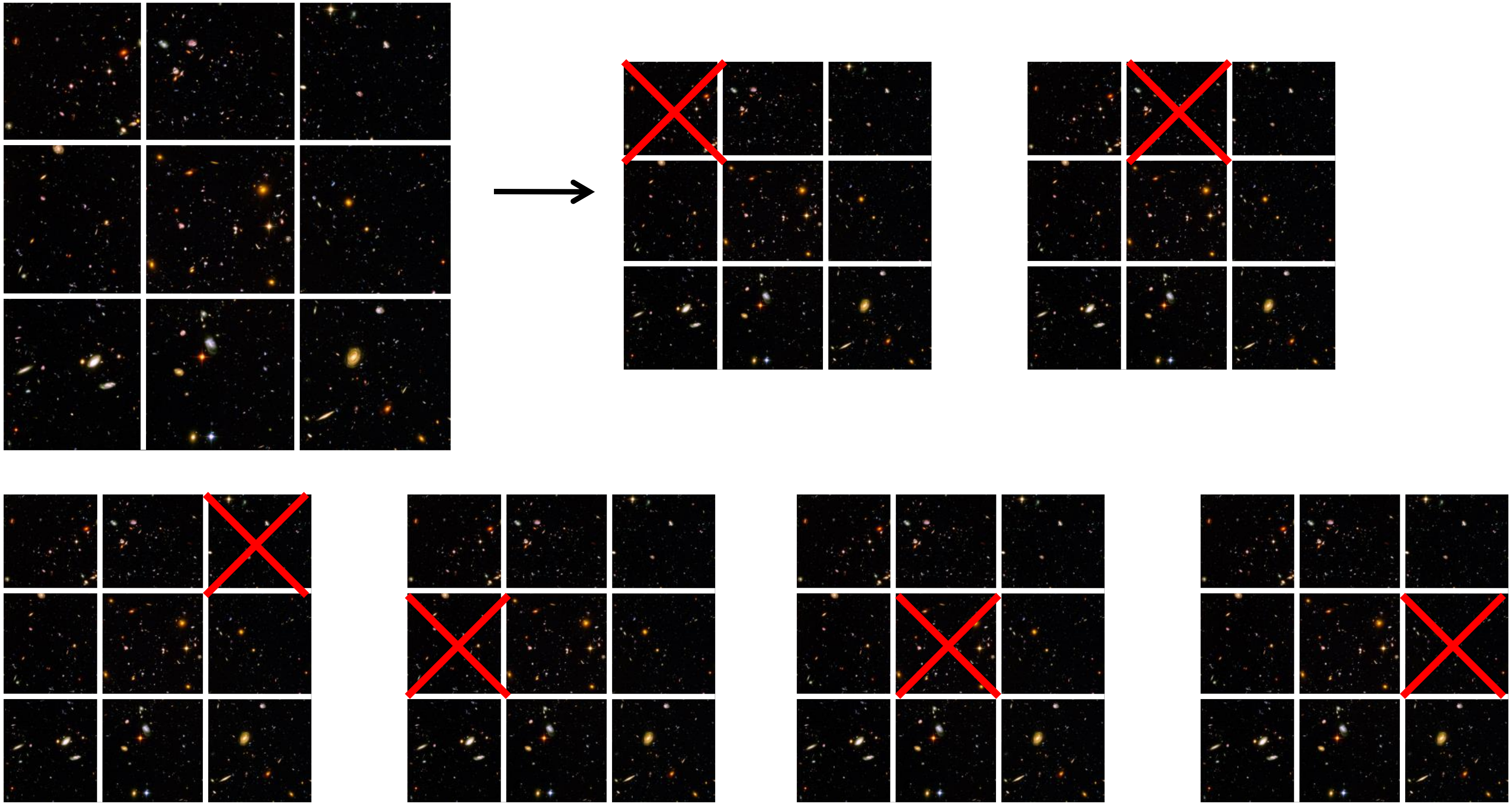
EXAMPLE



0.8
0.6
0.4
0.2
0
-0.2
Normalized Probability
0
2
4
6
8
10
12
Number of Satellites
Noise Profile
Composite Counts
How are these error bars
Determined? Jackknife statistics.

JackKnife Resampling

- Statistics on one large data sample. How do you estimate the error?

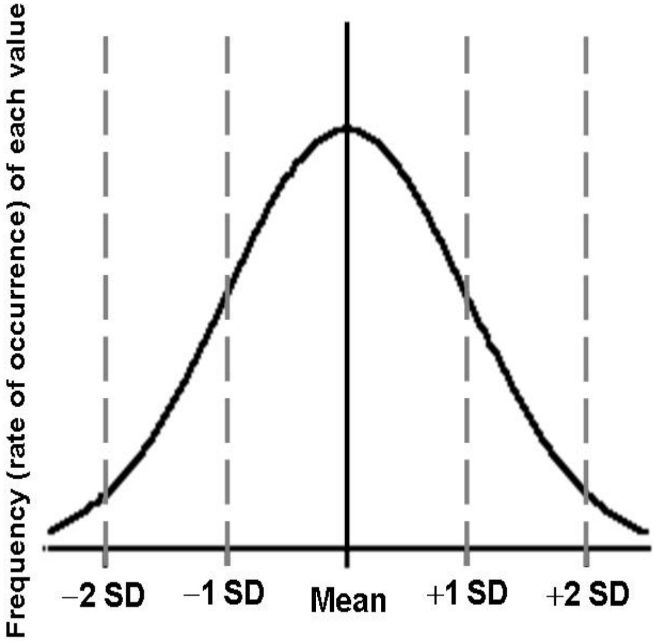


JackKnife Resampling (cont.)

- Mean PDF → Unbiased mean
- Variance * (n-1) = Variance of the large sample
- Intuitively → large sample for mean, errors scale up as sum of the error on each piece.
- Resampling techniques are crucial in cosmology.

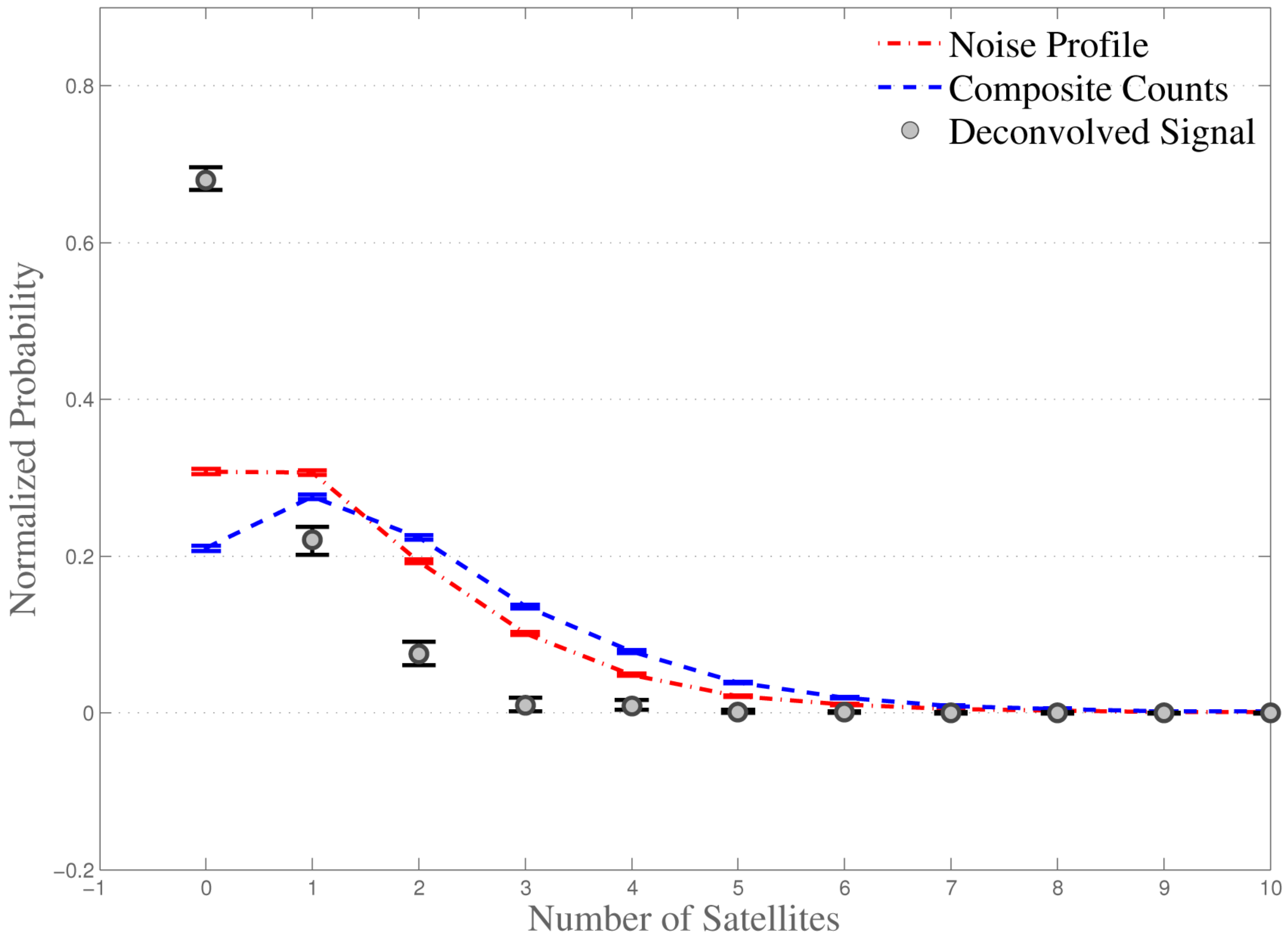
Stochastic Error Propagation

- Closed form error propagation difficult to derive, we choose a stochastic, and constrained deconvolution



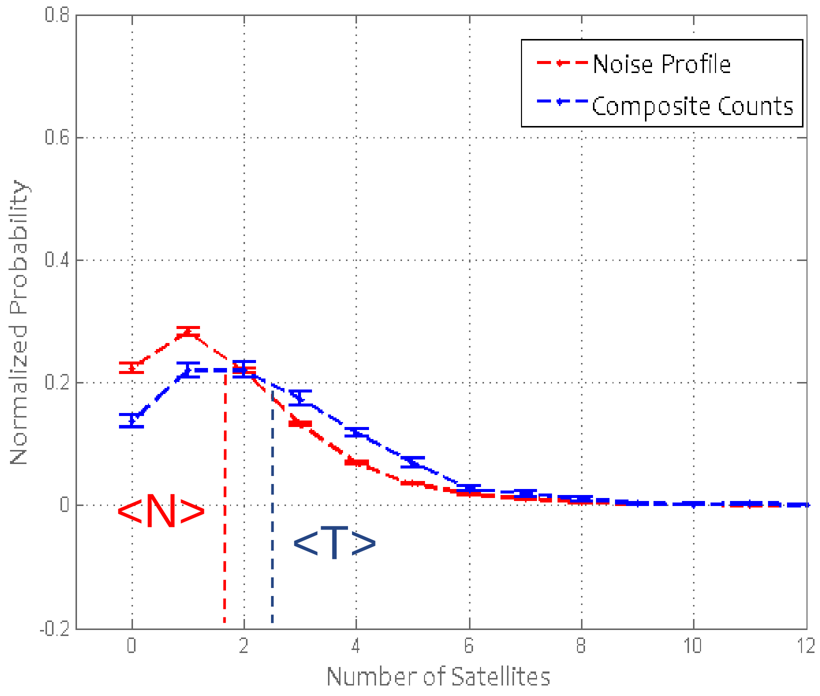
- FFT allows deconvolutions with 1 million initial P_i (N) and P_i (T) values chosen from normal distributions centered around mean value with width=error bar on each value.
- Disallow P_i (S) results with a negative value, as those are unphysical
- Mean P(S) and Variance computed

Basic Result:



Statistical Uncertainty

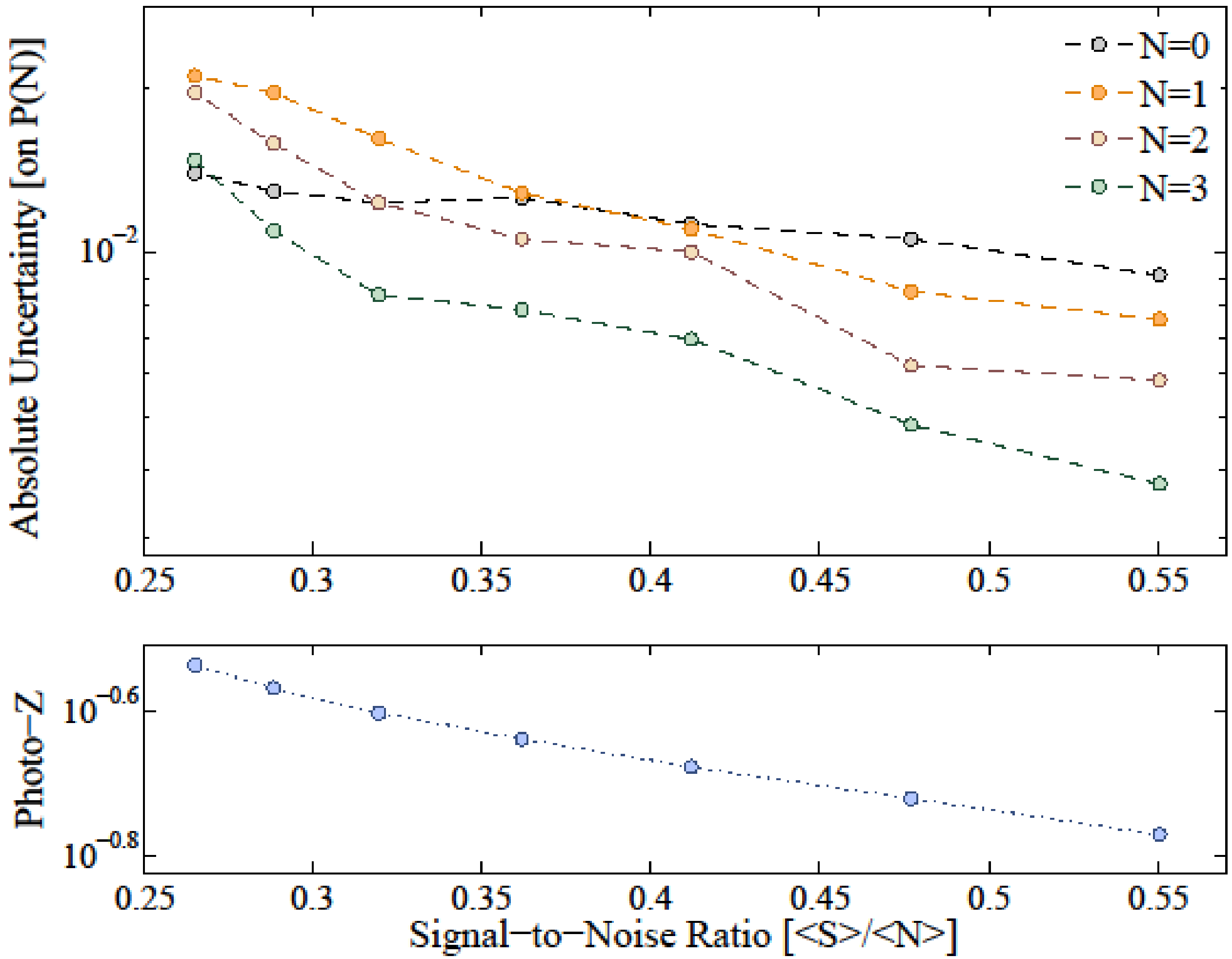
- Two contributions:
- 1) P(N) and P(T) error bars depend on sample size (# of MW-sized hosts – 25,000)
- 2) Another hidden dependence of statistical uncertainty:
  - Signal to noise ratio: <S>/<N> = (<T>-<N>)/<N>



As <S>/<N> goes down, statistical uncertainty goes up roughly exponentially.

Shows up in the stochastic error bars with fixed sample size

Statistical Uncertainty in Results As Function of Signal to Noise



Statistical Frame of Mind

- Signal to noise pictures → Cleaner way to think about errors.

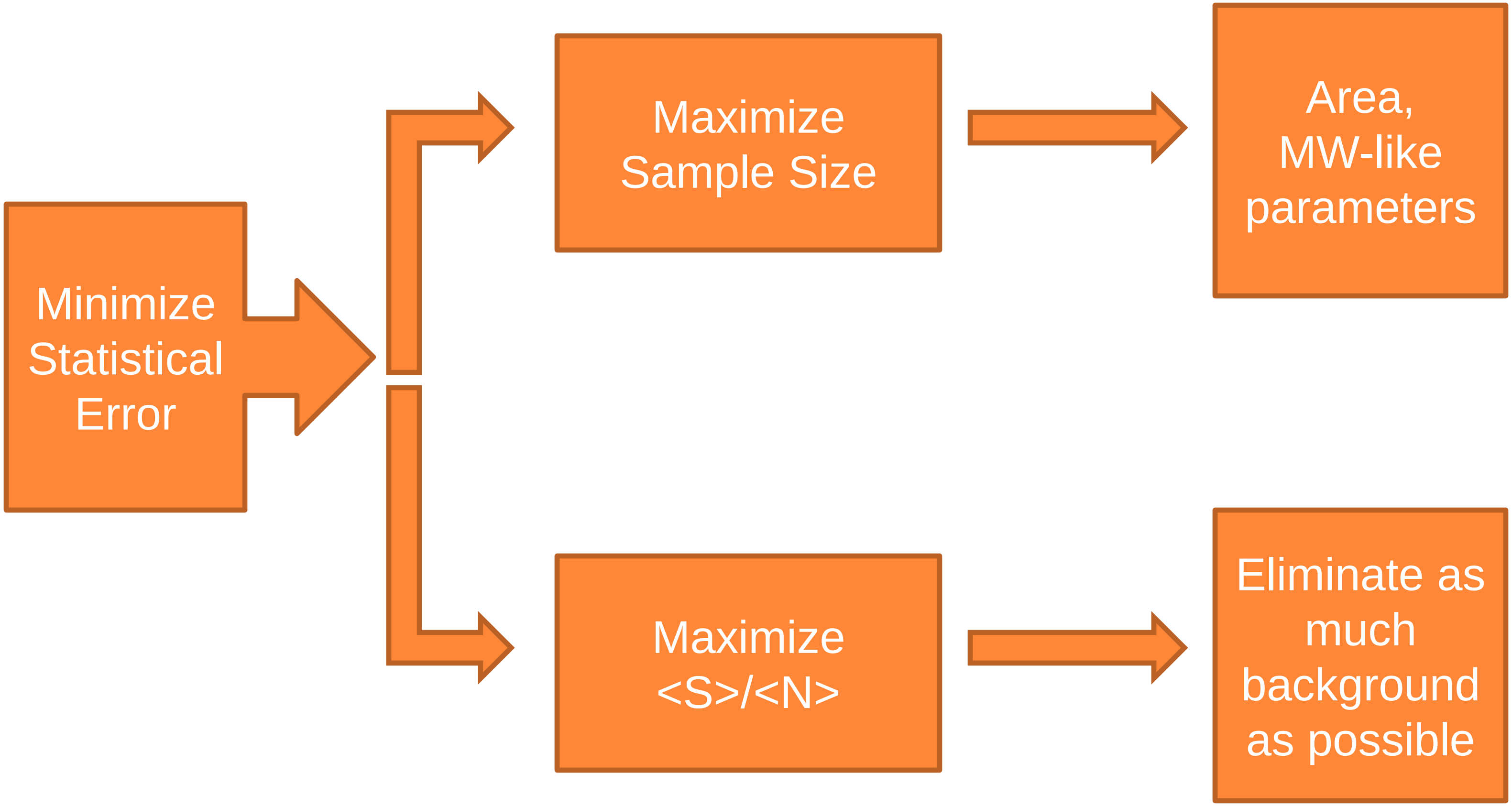


Photo-Z Trade-Off

- Details details details
- Decreases uncertainty significantly
- But introduces a systematic loss

η = probability that any single satellite is lost

Choose to keep under 15%

- Let's correct for this!
  - First of our systematic adjustments

Systematic #1: Photo-Z Loss Correction

m = number of satellites lost, T = number of true satellites

$$p_{loss}(m|T) = \eta^m (1-\eta)^{T-m} \binom{T}{m}$$

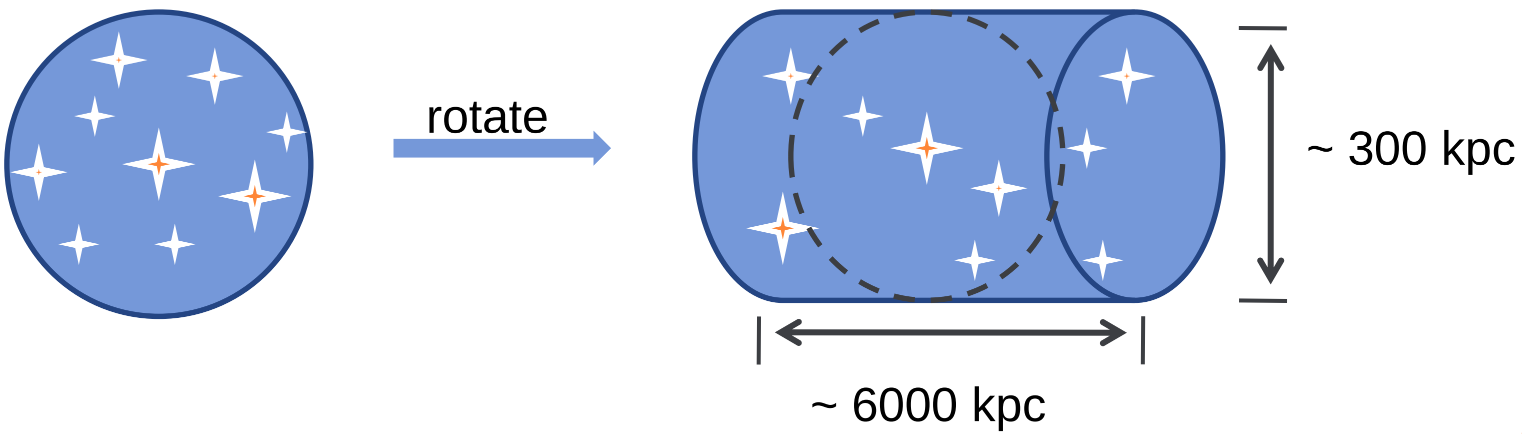
$$p_{meas}(S) = \sum_{m=0}^{\infty} p_{true}(T) p_{loss}(m|T)$$

$$p_{meas}(S) = \sum_{m=0}^{\infty} p_{true}(S+m) \eta^m (1-\eta)^S \binom{S+m}{m}$$

Where T = S+m substitution was made

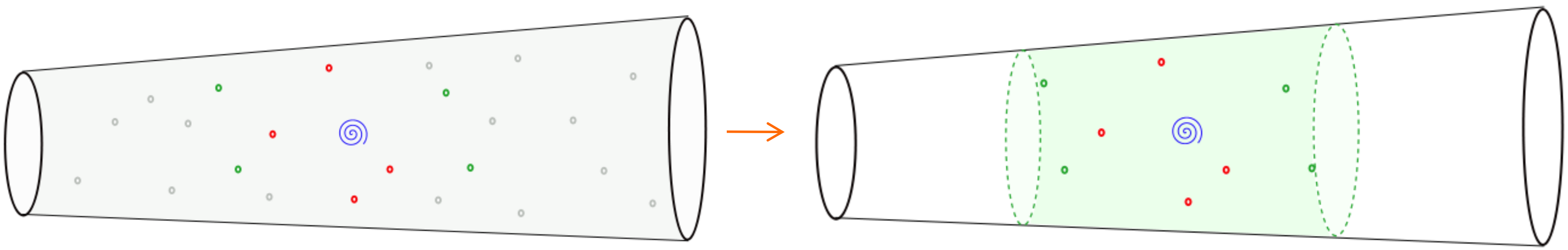
What Else Can we Do?

- Correct for cylindrical geometry.
- Normally you can't do better:
  - Line of sight motion – uncertainty in Z
  - Uncertainty in Z describes a cylinder.



Systematic #2: Correlated Galaxies Along line of Sight

- Without this correction:



Cylindrical intermediate result

- Auto-correlation function:
  - Integrate to get: given galaxy G, probability that at least n correlated galaxies will be found within a volume centered on G.

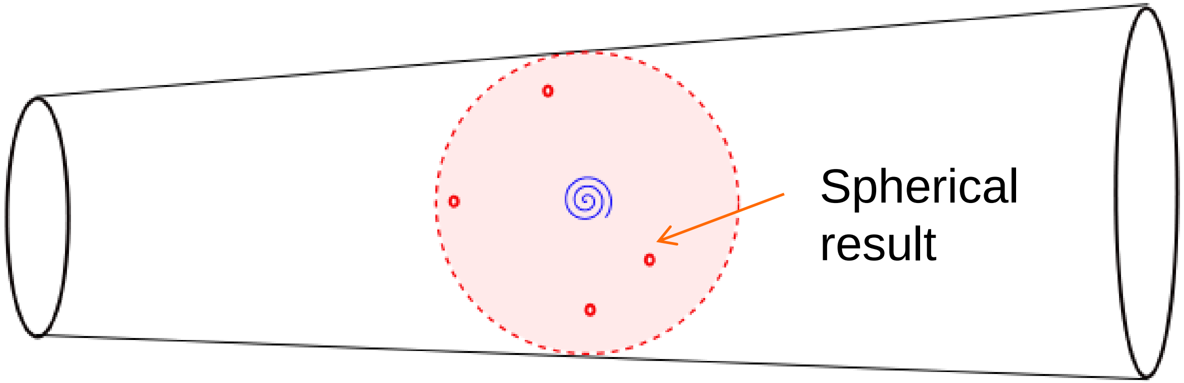
- $p_{gain}(n)$ is geometric. Integrate space inside cylinder but outside sphere.

Systematic #2: Cont.

$$p_{gain}(n) = \zeta^n(1-\zeta)$$

$$p_{meas}(S) = \sum_{m=0}^{\infty}\sum_{n=0}^{S+m} p_{true}(S+m-n)p_{loss}(m|S+m-n)p_{gain}(n)$$

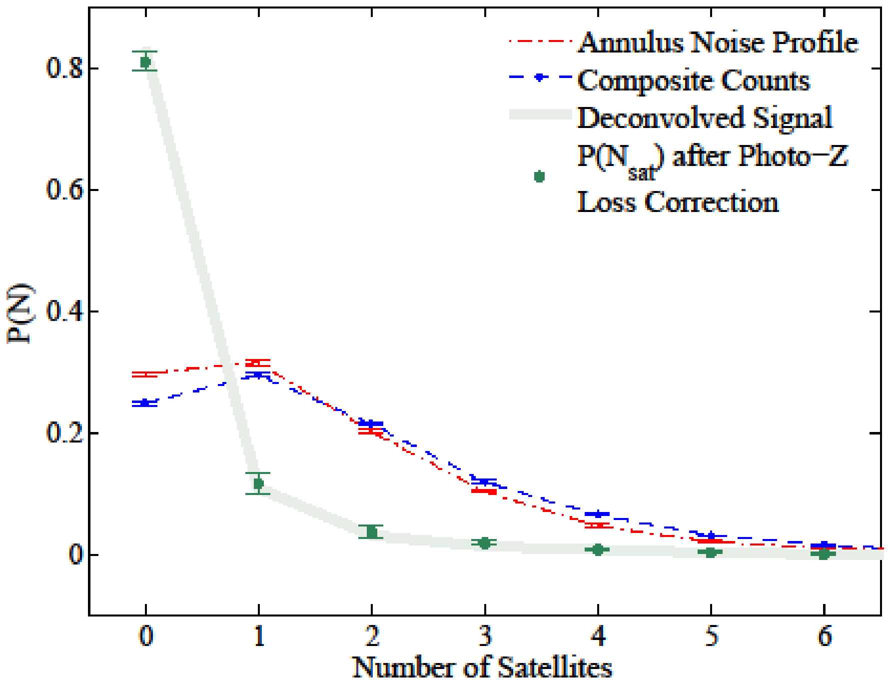
ζ is boost probability ~ 0.2

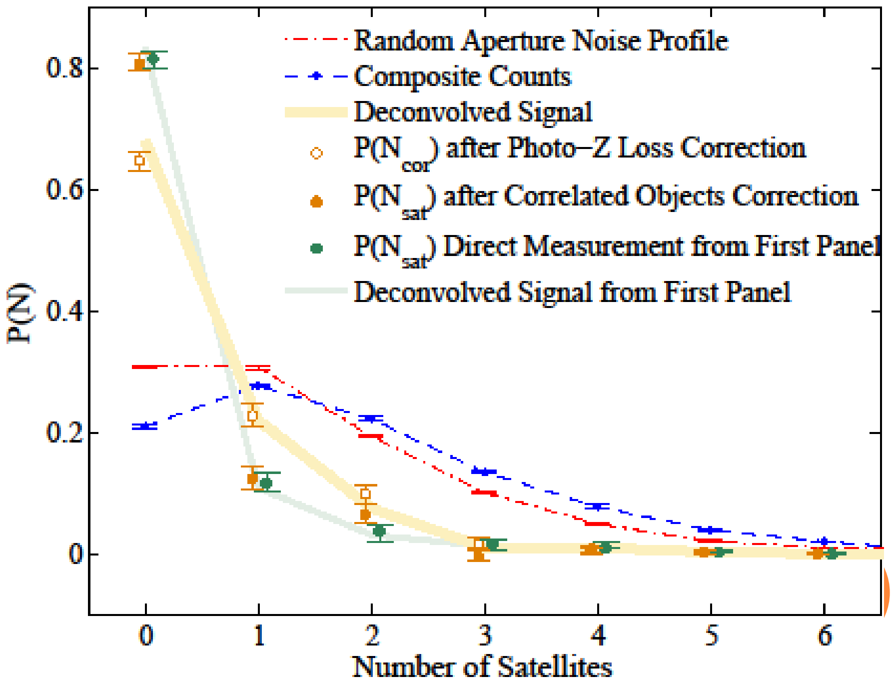


Results – Probability of Hosting N=0,1,2… MC-like satellites

Right panel shows the cylindrical intermediate result, spherical result obtained through systematic gain correction

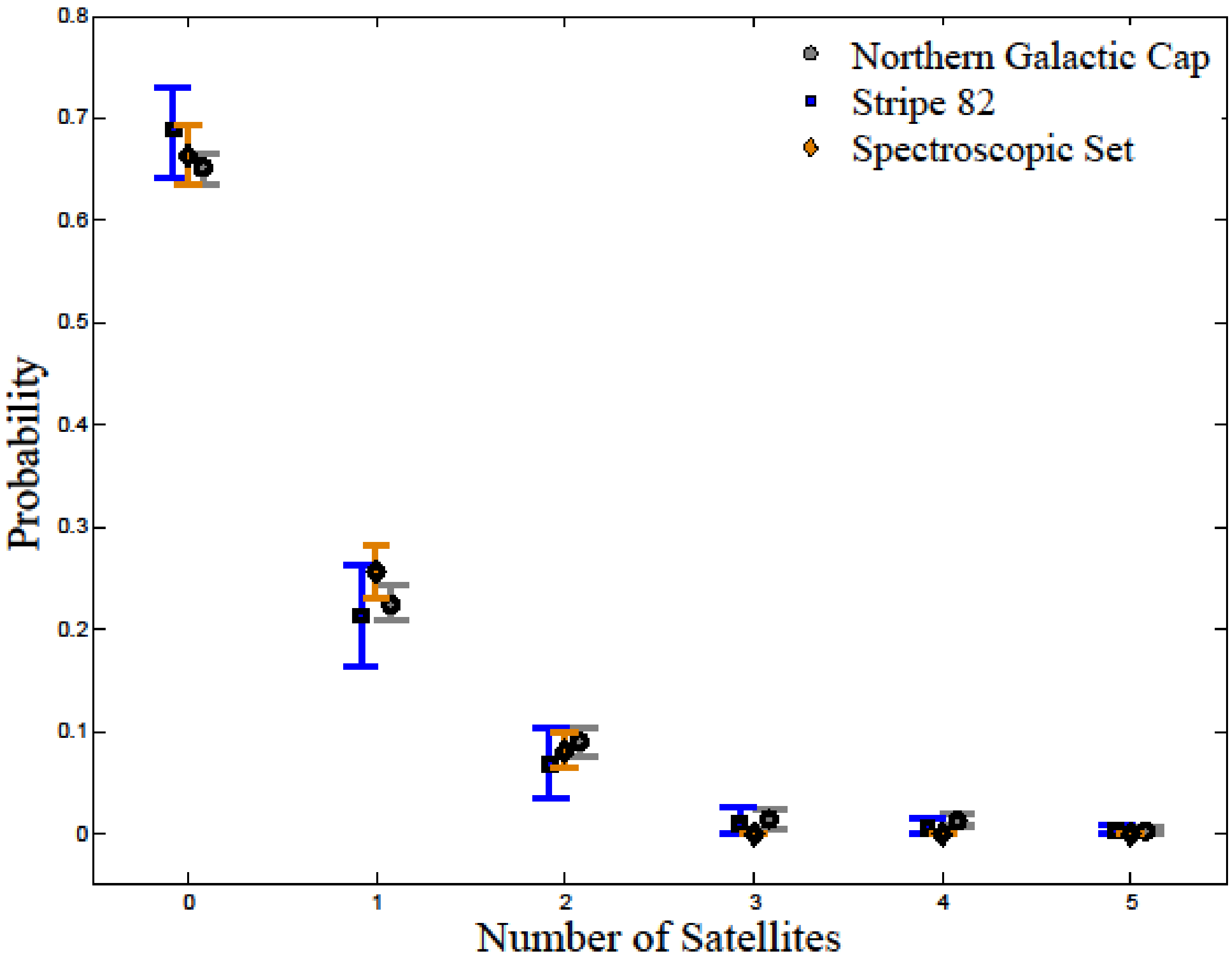
Left panel show another method of attaining spherical result directly. Compares well with the method described in this presentation





Compare with Other Methods

- Statistics Much Improved:



Cylindrical results

Stripe 82 is a trial we did with a smaller 1332 sample.

Spectroscopic set best current statistics in this field.

NGC our result. We have reduced errors by at least a factor of two.

WAIT, JUST HOW COMMON ARE THE MAGELLANIC CLOUDS??

- Cylinder

Number of Satellites	Probability %
Zero	64.6 +- 1.5
One	22.8 +- 1.8
Two	9.7 +- 1.5
Three	1.5 +- 1.2

Number of Correlated Objects	Measured Percentage of MW-sized Galaxies	Systematic Adjustment
Zero	$68.0^{+1.6}_{-1.3}$	-3.4
One	$22.0^{+1.7}_{-1.9}$	+0.8
Two	$7.5^{+1.5}_{-1.5}$	+2.2
Three	$1.0^{+1.0}_{-0.7}$	+0.5
Four	$0.9^{+0.7}_{-0.5}$	-0.3
Five	$0.1^{+0.3}_{-0.1}$	+0.1
Six	$0.1^{+0.2}_{-0.1}$	-0.1

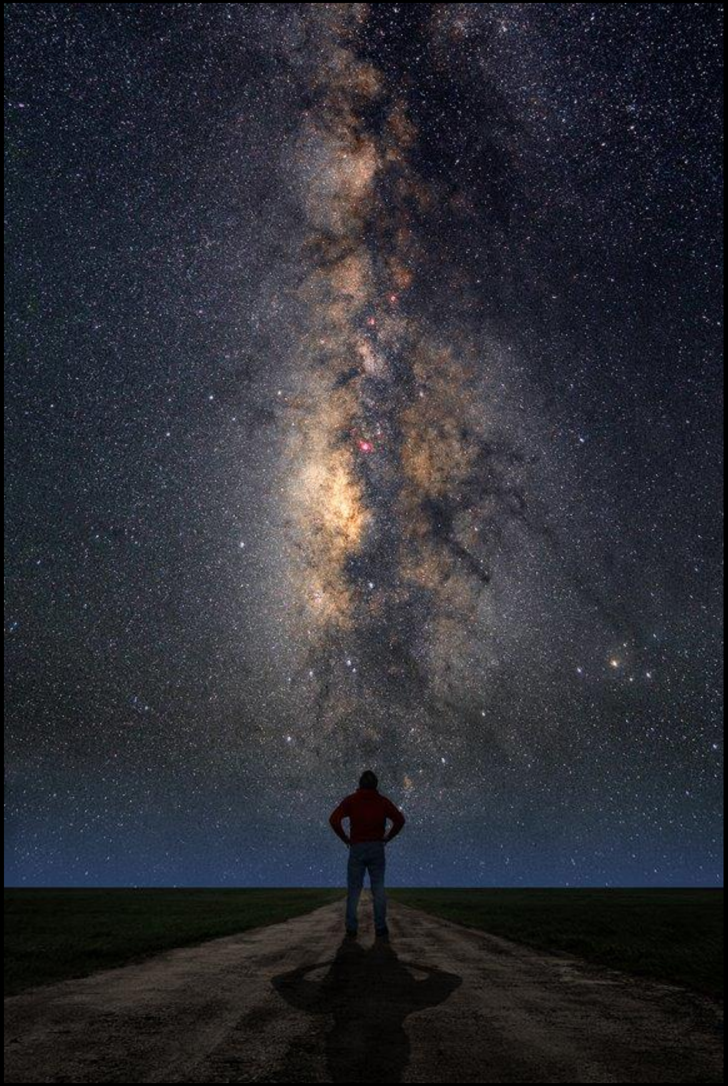
- Sphere

Number of Satellites	Probability %
Zero	81.4 +- 1.5
One	11.6 +- 1.7
Two	3.5 +- 1.4
Three	1.6 +- 1.2

Number of Satellites	Measured Percentage of MW-sized Galaxies	Systematic Loss Adjustment
Zero	$83.4^{+1.5}_{-1.4}$	-2.0
One	$10.8^{+1.8}_{-1.6}$	+0.8
Two	$3.1^{+1.3}_{-1.5}$	+0.4
Three	$1.4^{+0.9}_{-1.0}$	+0.2
Four	$0.7^{+0.6}_{-0.5}$	+0.4
Five	$0.1^{+0.2}_{-0.1}$	+0.2
Six	$0.1^{+0.2}_{-0.1}$	-0.1

Conclusions

- LCDM simulations hold up.
- Wechsler group simulations reporting ~81% for N=0, ~10% for N=1, ~5% for N=2…
- Milky Way is a special case:
  - Implies transient phase? Will be cannabalized?
  - Implies transit? Just passing through?



Power of Statistics in Cosmology

- Allows work with enormous, varied data sets.
  - Compare: 140 hosts vs. 25,000 over 3500 square degrees.
  - Time = grant $$
  - Precision, reproducibility
- Signal to Noise picture → aids in data optimization.
- Mathematical tool kit to handle systematic errors: probabilistic undercounting/overcounting.
- Flexible parameters.
- Applications: Are the defects on a wafer randomly distributed? What is the correlation between one defect and another?