

Astronomy 503

Observational Astronomy



Prof. Gautham Narayan

Lecture 05: Series and Fourier Transforms and Power Spectrums and Noise
oh my.

Announcements

- 2nd HW is on the github
- General tip: Read an assignment fully before you start to tackle it because the hints (and the traps) are frequently buried in the question

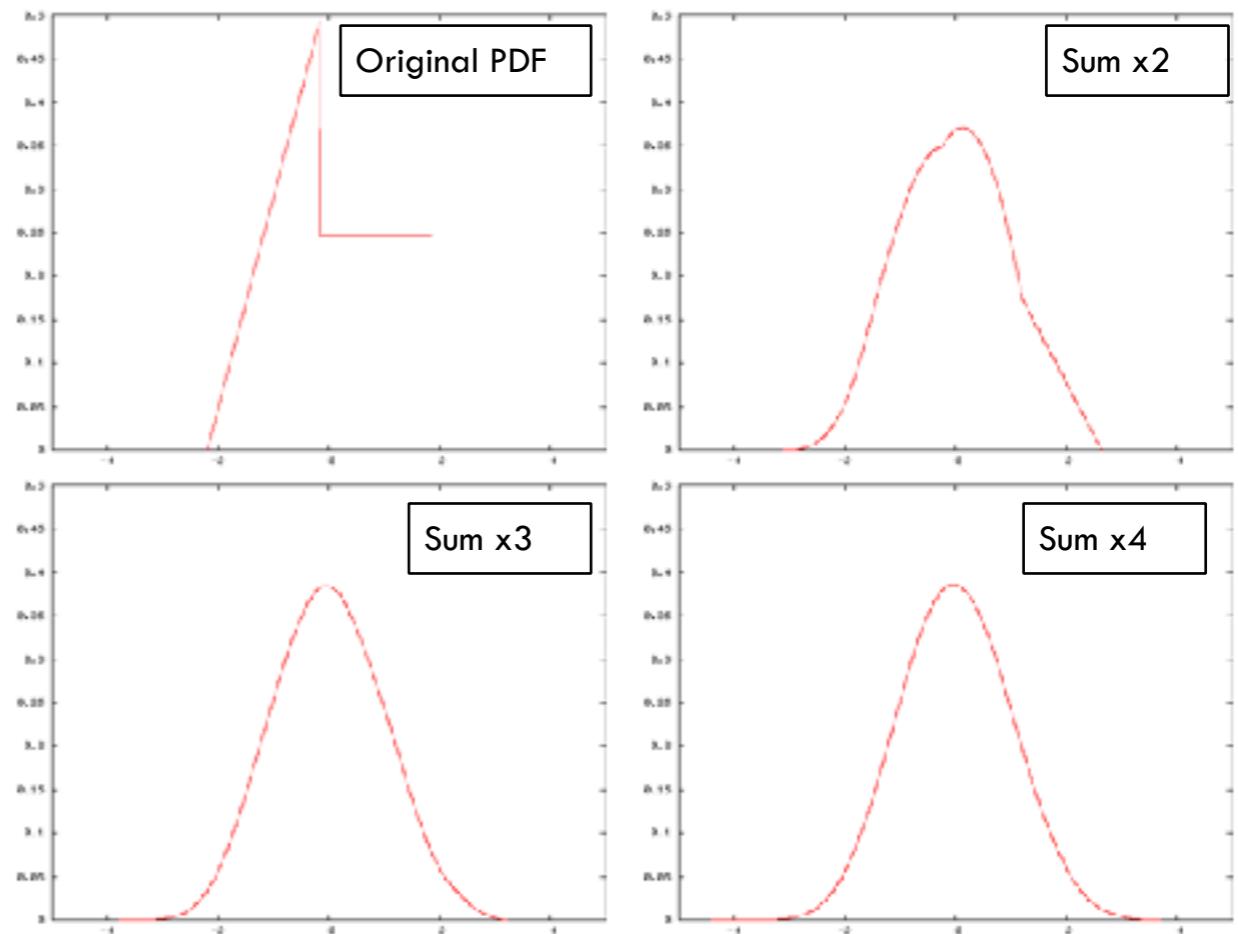
Best practice: statistical data analysis

- Understand your data model.
 - Propagation, calibration, and noise.
- Examine your implicit statistical assumptions.
 - E.g. all noise is Gaussian.
 - Errors are uncorrelated
- Always quote errors and uncertainties.
- Explain your error calculations.
 - Be conservative if errors are unknown.
- Keep science conclusions within the uncertainties.
 - Standard is observational honesty, and adherence to best community practice.
 - Absolute certainty isn't possible – we proceed incrementally from one reproducible result to another.

Why are Gaussian statistics so pervasive? —> Central Limit Theorem

4

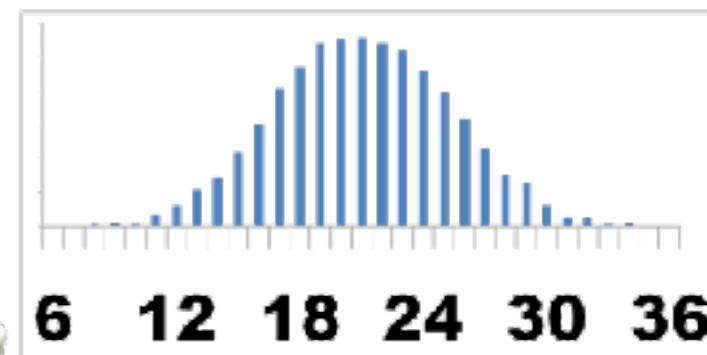
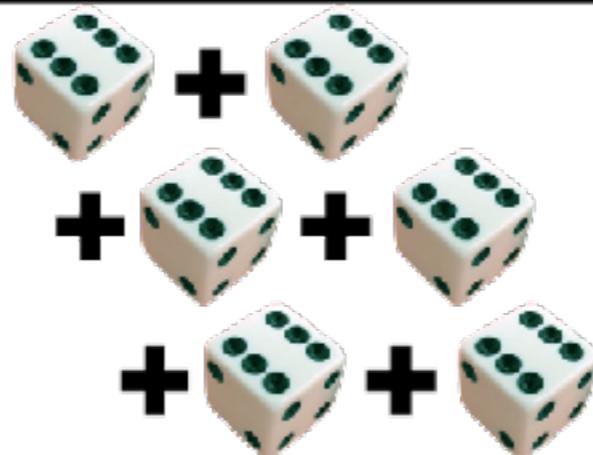
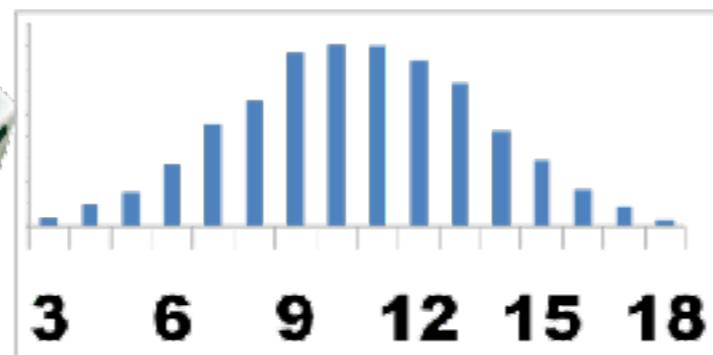
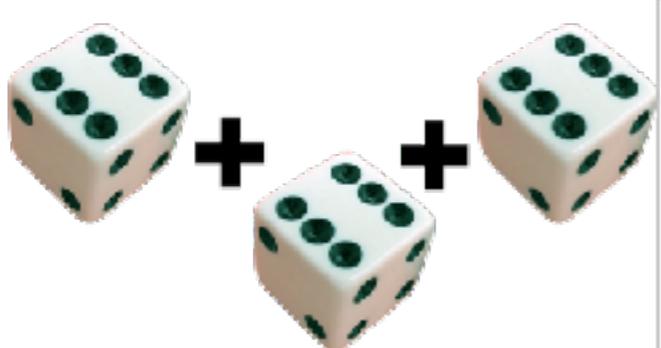
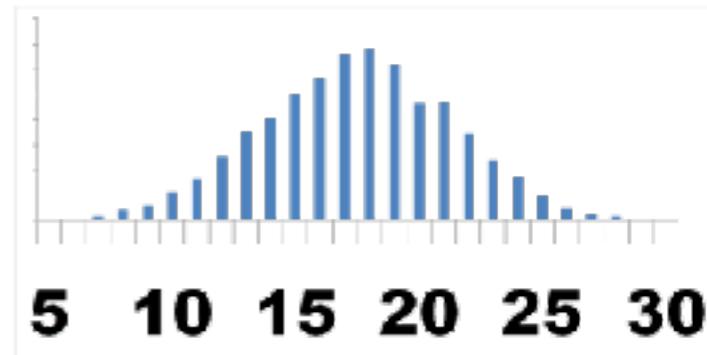
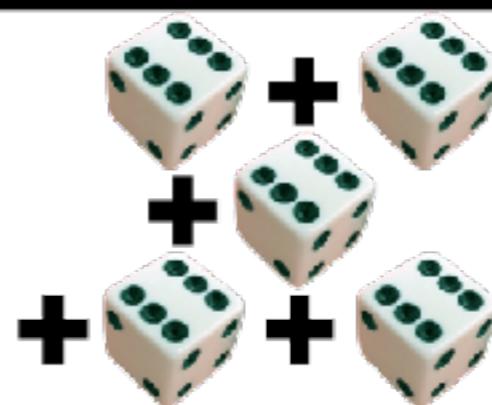
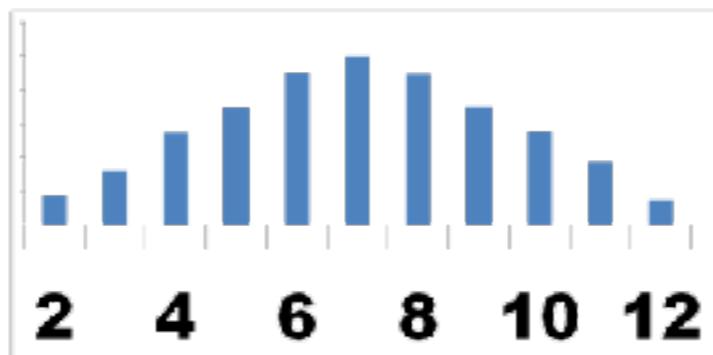
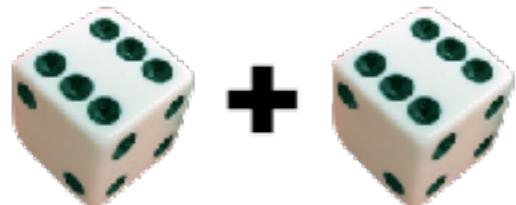
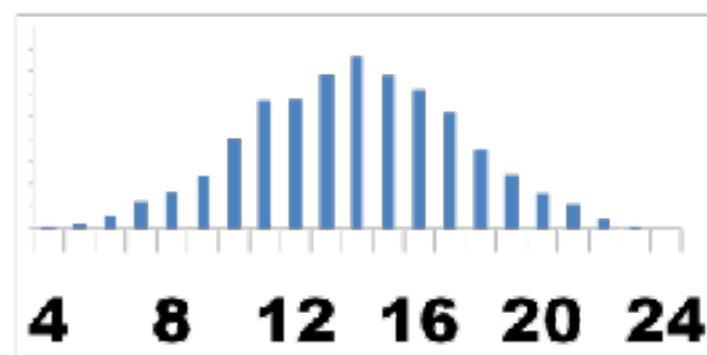
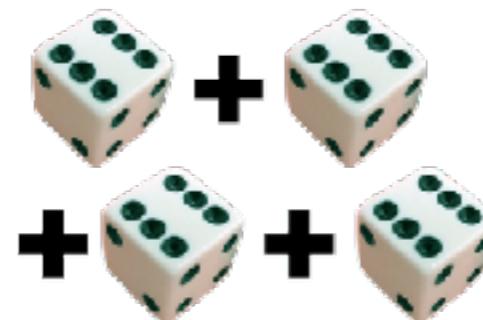
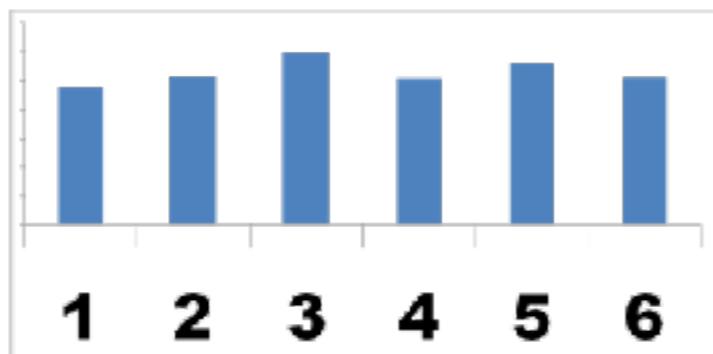
The distribution of an average tends to be Normal, even when the distribution from which the average is computed is decidedly non-Normal.



If x is the average of n independent random variables of expected value μ and variance σ^2 , then:

$$\text{As } n \rightarrow \infty \quad p(x) \rightarrow \frac{1}{\sqrt{2\pi} \frac{\sigma^2}{n}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

(All scaled to have $\mu=1$, $\sigma=1$; CC)



The error propagation equation

$$x = f(u, v)$$

Consider a Taylor-series expansion about mean values (\bar{u}, \bar{v}) :

$$(x_i - \bar{x}) \approx (u_i - \bar{u}) \left(\frac{\partial x}{\partial u} \right) + (v_i - \bar{v}) \left(\frac{\partial x}{\partial v} \right)$$

But,

$$\sigma_u^2 = \langle (u - \bar{u})^2 \rangle = \lim_{n \rightarrow \infty} \left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right]$$

Therefore:

$$\sigma_x^2 = \sigma_u^2 \left(\frac{\partial x}{\partial u} \right)^2 + \sigma_v^2 \left(\frac{\partial x}{\partial v} \right)^2 + 2\sigma_{uv} \left(\frac{\partial x}{\partial u} \right) \left(\frac{\partial x}{\partial v} \right) + \dots$$

Covariance and independence

- Covariance σ_{uv} : measure of the correlation between fluctuations in u and v :

$$\sigma_{uv} = \langle (u - \bar{u})(v - \bar{v}) \rangle$$

- Covariance is zero for **uncorrelated** statistical errors:

$$\sigma_x^2 = \sigma_u^2 \left(\frac{\partial x}{\partial u} \right)^2 + \sigma_v^2 \left(\frac{\partial x}{\partial v} \right)^2 + 2\sigma_{uv} \left(\frac{\partial x}{\partial u} \right) \left(\frac{\partial x}{\partial v} \right) + \dots$$

- We will make this assumption in what follows.

Covariance and independence

- Covariance σ_{uv} : measure of the correlation between fluctuations in u and v :

$$\sigma_{uv} = \langle (u - \bar{u})(v - \bar{v}) \rangle$$

- Covariance is zero for **uncorrelated** statistical errors:

$$\sigma_x^2 = \sigma_u^2 \left(\frac{\partial x}{\partial u} \right)^2 + \sigma_v^2 \left(\frac{\partial x}{\partial v} \right)^2 + \cancel{\sigma_{uv} \left(\frac{\partial x}{\partial u} \right) \left(\frac{\partial x}{\partial v} \right)} + \dots$$

- We will make this assumption in what follows.

$$\sigma_{uv} = 0$$

Error propagation: sums and differences

- Consider $x = u + \text{const.}$

$$x = f(u) = u + a$$

$$\sigma_x^2 = \sigma_u^2 \left(\frac{\partial x}{\partial u} \right)^2 = \sigma_u^2$$

- Consider $x = u + v$

$$x = f(u, v)$$

$$= u + v$$

$$\sigma_x^2 = \sigma_u^2 \left(\frac{\partial x}{\partial u} \right)^2 + \sigma_v^2 \left(\frac{\partial x}{\partial v} \right)^2$$

$$= \sigma_u^2 + \sigma_v^2$$

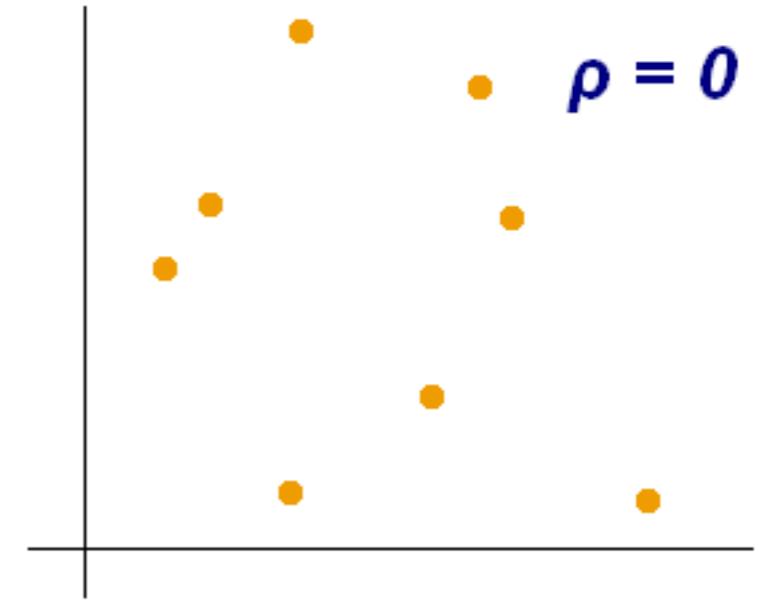
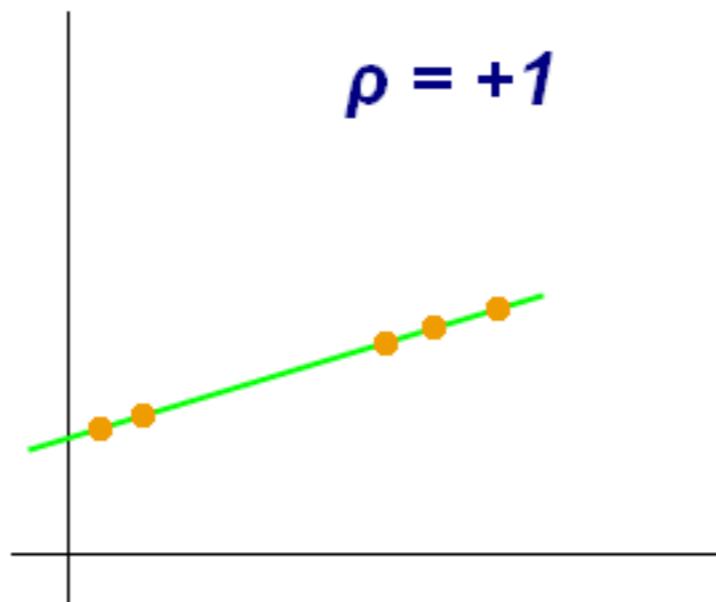
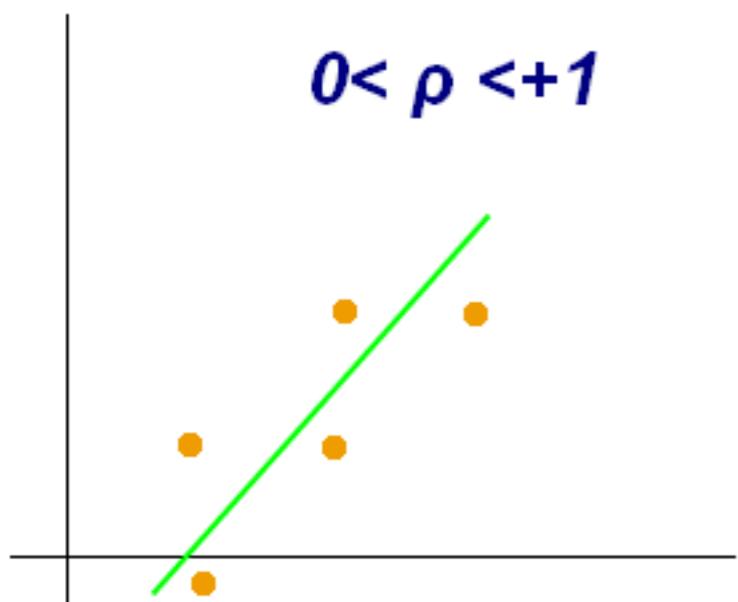
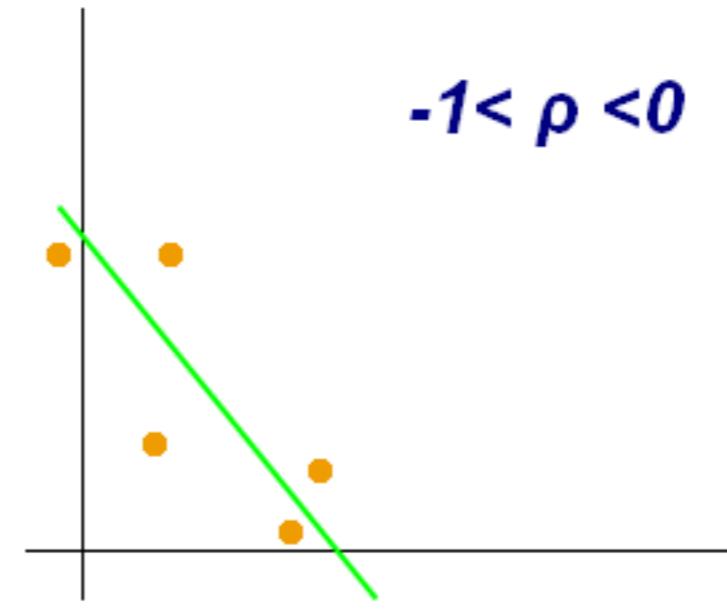
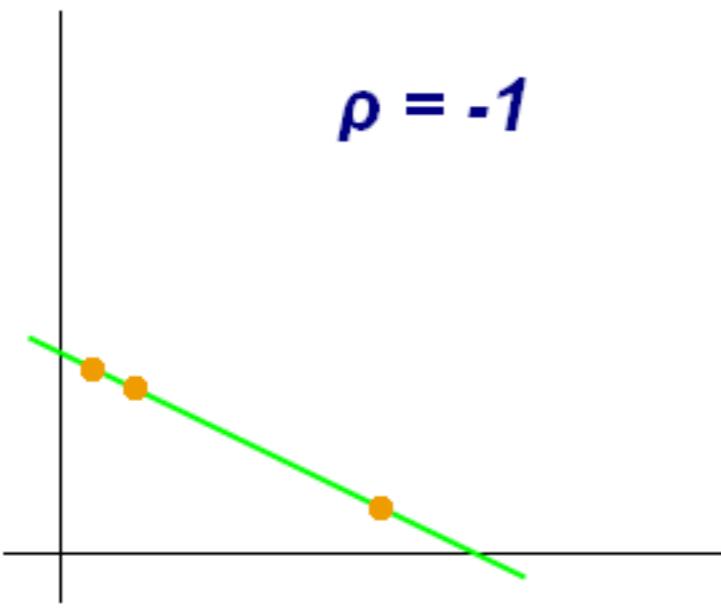
The quadrature rule

Some useful formulas

| Function | Variance | Standard Deviation |
|----------------------|---|--|
| $f = aA$ | $\sigma_f^2 = a^2 \sigma_A^2$ | $\sigma_f = a\sigma_A$ |
| $f = aA + bB$ | $\sigma_f^2 = a^2 \sigma_A^2 + b^2 \sigma_B^2 + 2ab \text{cov}_{AB}$ | $\sigma_f = \sqrt{a^2 \sigma_A^2 + b^2 \sigma_B^2 + 2ab \text{cov}_{AB}}$ |
| $f = AB$ | $\sigma_f^2 \approx f^2 \left[\left(\frac{\sigma_A}{A} \right)^2 + \left(\frac{\sigma_B}{B} \right)^2 + 2 \frac{\text{cov}_{AB}}{AB} \right]$ | $\sigma_f \approx f \sqrt{\left(\frac{\sigma_A}{A} \right)^2 + \left(\frac{\sigma_B}{B} \right)^2 + 2 \frac{\text{cov}_{AB}}{AB}}$ |
| $f = \frac{A}{B}$ | $\sigma_f^2 \approx f^2 \left[\left(\frac{\sigma_A}{A} \right)^2 + \left(\frac{\sigma_B}{B} \right)^2 - 2 \frac{\text{cov}_{AB}}{AB} \right]$ [11] | $\sigma_f \approx f \sqrt{\left(\frac{\sigma_A}{A} \right)^2 + \left(\frac{\sigma_B}{B} \right)^2 - 2 \frac{\text{cov}_{AB}}{AB}}$ |
| $f = aA^b$ | $\sigma_f^2 \approx f^2 \left(b \frac{\sigma_A}{A} \right)^2$ [12] | $\sigma_f \approx f \left(b \frac{\sigma_A}{A} \right)$ |
| $f = a \ln(bA)$ | $\sigma_f^2 \approx \left(a \frac{\sigma_A}{A} \right)^2$ [13] | $\sigma_f \approx \left a \frac{\sigma_A}{A} \right $ |
| $f = a \log_{10}(A)$ | $\sigma_f^2 \approx \left(a \frac{\sigma_A}{A \ln(10)} \right)^2$ [13] | $\sigma_f \approx \left a \frac{\sigma_A}{A \ln(10)} \right $ |
| $f = ae^{bA}$ | $\sigma_f^2 \approx f^2 (b\sigma_A)^2$ [14] | $\sigma_f \approx f (b\sigma_A) $ |
| $f = a^{bA}$ | $\sigma_f^2 \approx f^2 (b \ln(a)\sigma_A)^2$ | $\sigma_f \approx f (b \ln(a)\sigma_A) $ |
| $f = A^B$ | $\sigma_f^2 \approx f^2 \left[\left(\frac{B}{A} \sigma_A \right)^2 + (\ln(A)\sigma_B)^2 + 2 \frac{B \ln(A)}{A} \text{cov}_{AB} \right]$ | $\sigma_f \approx f \sqrt{\left(\frac{B}{A} \sigma_A \right)^2 + (\ln(A)\sigma_B)^2 + 2 \frac{B \ln(A)}{A} \text{cov}_{AB}}$ |

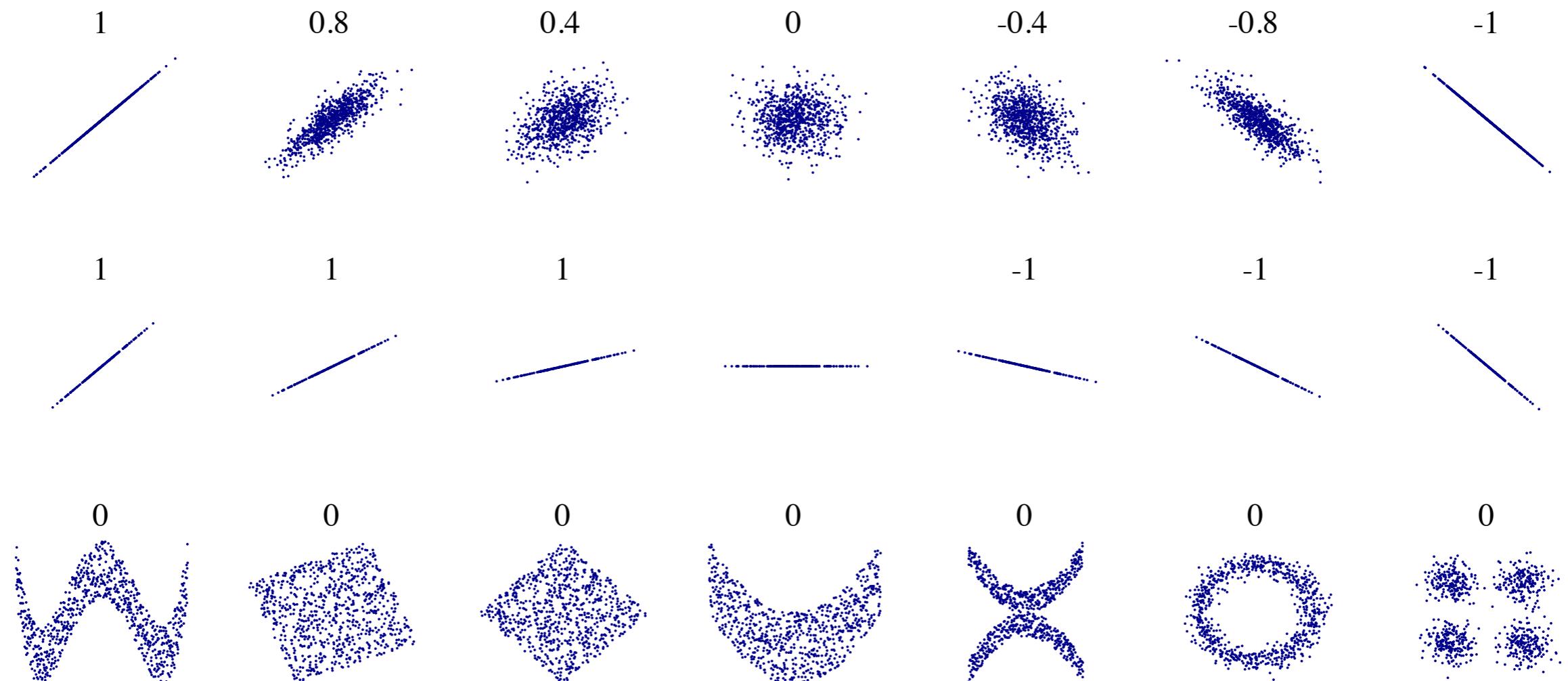
From [wikipedia.org](https://en.wikipedia.org)

Correlation



- correlation coefficient (PPMCC), is a measure of the linear dependence (correlation) between two variables X and Y. It has a value between +1 and –1 inclusive, where 1 is total positive linear correlation, 0 is no linear correlation, and –1 is total negative linear correlation.

Correlation

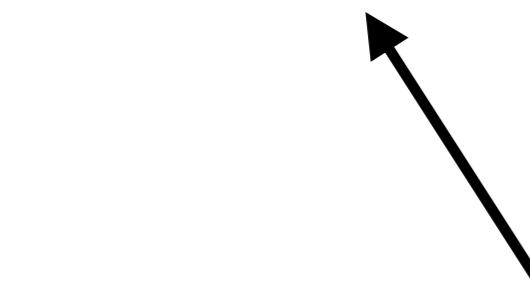


RECAP

13

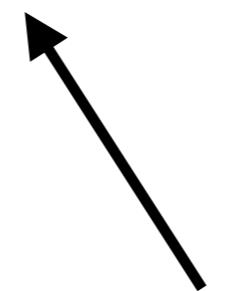
Posterior

How probable is the hypothesis given the data we observed



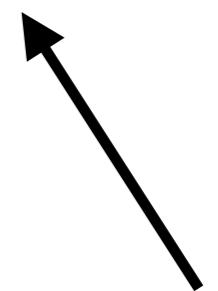
Likelihood

How probable is the data given the hypothesis is true



Prior

How probable was the hypothesis before we observed anything



$$p(\text{Hypothesis}|\text{Data}) = \frac{p(\text{Data}|\text{Hypothesis})p(\text{Hypothesis})}{p(\text{Data})}$$

Evidence

How probable is the data over all possible hypotheses



- ▶ In this view of the Universe, there is some underlying truth.
- ▶ If we measure the photon flux F from a given star, then measure it again, then again, and so on, each time we will get a slightly different answer due to the statistical error of my measuring device.
- ▶ In the limit of a large number of measurements, the frequency of any given value indicates the probability of measuring that value.
- ▶ For frequentists probabilities are fundamentally related to frequencies of events i.e. **P(D|H)**.
- ▶ This means, for example, that in a strict frequentist view, it is meaningless to talk about the probability of the true flux of the star: the true flux is (by definition) a single fixed value.
- ▶ To talk about a frequency distribution for a fixed value or model parameter is nonsense.

- ▶ In the **frequentist** school of statistics, **parameters do not have probability distributions**.
Probability can only be used to describe frequencies, not degrees of belief (or odds).
- ▶ In the **frequentist view**, it's only the data that can be modeled as having been drawn from a probability distribution, because **we can imagine doing the experiment or observation multiple times, and building up a frequency distribution of results**.
- ▶ This PDF over datasets is the sampling distribution, e.g. $P(m^{\text{obs}}|a,b,H)$, and **as long as the priors are uninformative/flat - this is the same as the Bayesian posterior, assuming the same hypothesis**.
- ▶ Given an assumed model, the frequentist view is that there is only one set of parameters, the true ones, and our job is to estimate them.
- ▶ Derivation of good estimators is a major activity in frequentist statistics, and has led to some powerful mathematical results and fast computational shortcuts - some of which are useful in Bayesian inference

- ▶ Frequentists seek to *transform* the frequency distribution of the data into a frequency distribution of their estimators, and hence *quantify their uncertainty in terms of what they expect would happen if the observation were to be repeated*
- ▶ Bayesians seek to *update their knowledge* of their model parameters, and hence quantify their uncertainty in terms of *what might have been had the observation been different, and what they knew before the data were taken*

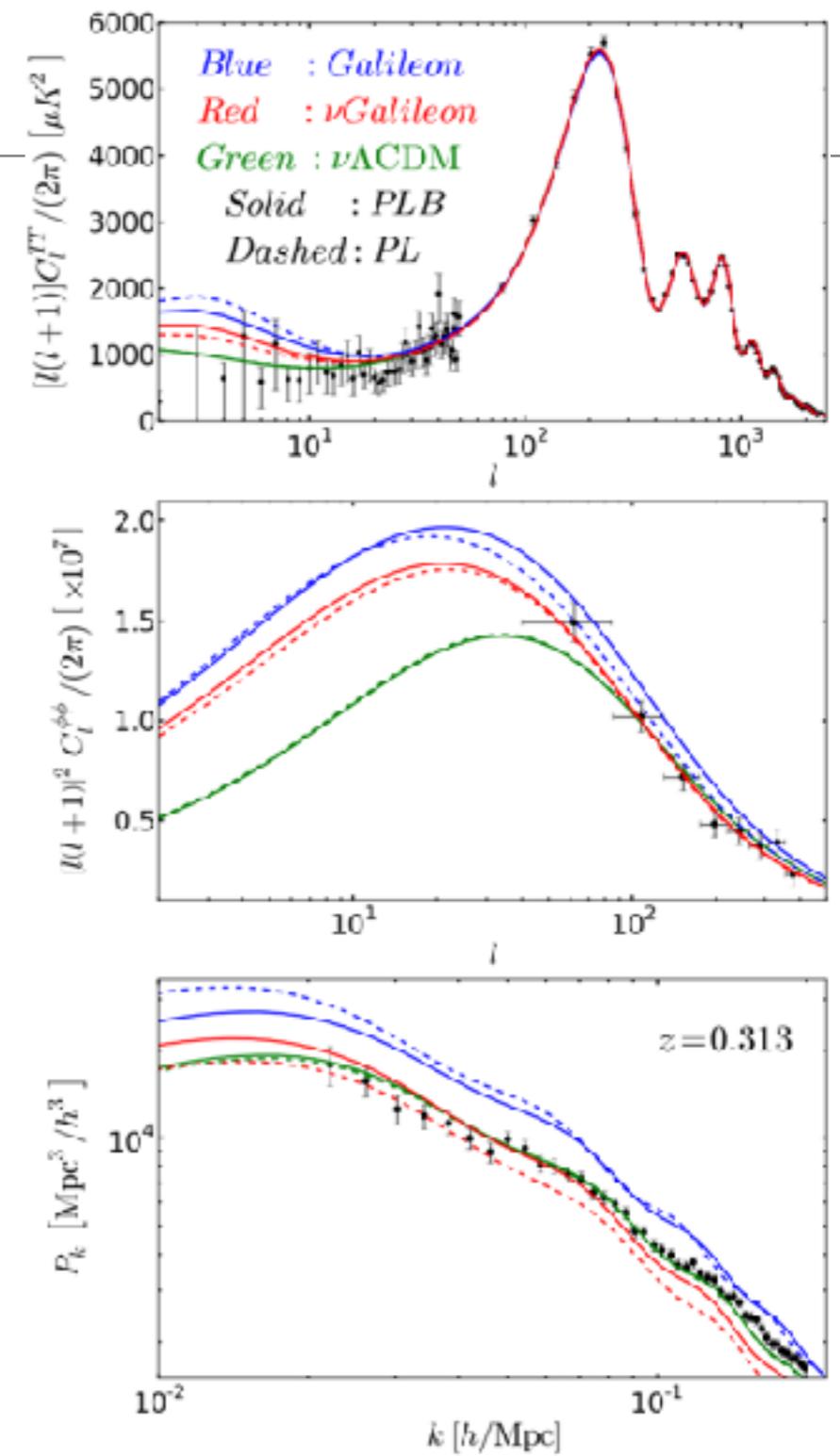
THIS IS KINDA REVOLUTIONARY

17

- ▶ To most scientists, there is Truth, and the data are the noisy realization of the truth
- ▶ You use the noisy data to say the Truth lies in some confidence interval - if that confidence interval is too large, well get better data - it should shrink as \sqrt{N})
- ▶ **Now we're treating the data as the fixed thing** - just whatever we recorded
- ▶ and instead claiming the Truth is unknowable, but rather given some model, which may not be right
- ▶ They can only be determined given the data, and the observations are themselves random variables
- ▶ **So it's reasonable to ask how confident we are in our estimate - what is the uncertainty?**

LET'S LOOK AT A CONCRETE EXAMPLE

- Fig 4: These plots illustrate the differences between Λ CDM and Galileon models (see Sect. 7.3.1), with **(GN: Solid lines)** and without massive neutrinos **(GN: Dashed lines)**. The Galileon models have background Friedmann equations that contain a scalar-field energy density contribution that generates late time cosmic acceleration and has an evolution consistent with observations and thus similar to that of a Λ CDM model.
- The Top: CMB temperature power spectra showing the ISW effect at low multipoles.
- Middle: CMB lensing potential spectra.
- Bottom: linear matter power spectra.
- The models plotted in dashed lines indicate their best fit models to Ade et al. (2014c) temperature data, WMAP9 polarization data (Hinshaw et al. 2013), and Planck-2013 CMB lensing (Ade et al. 2014d).



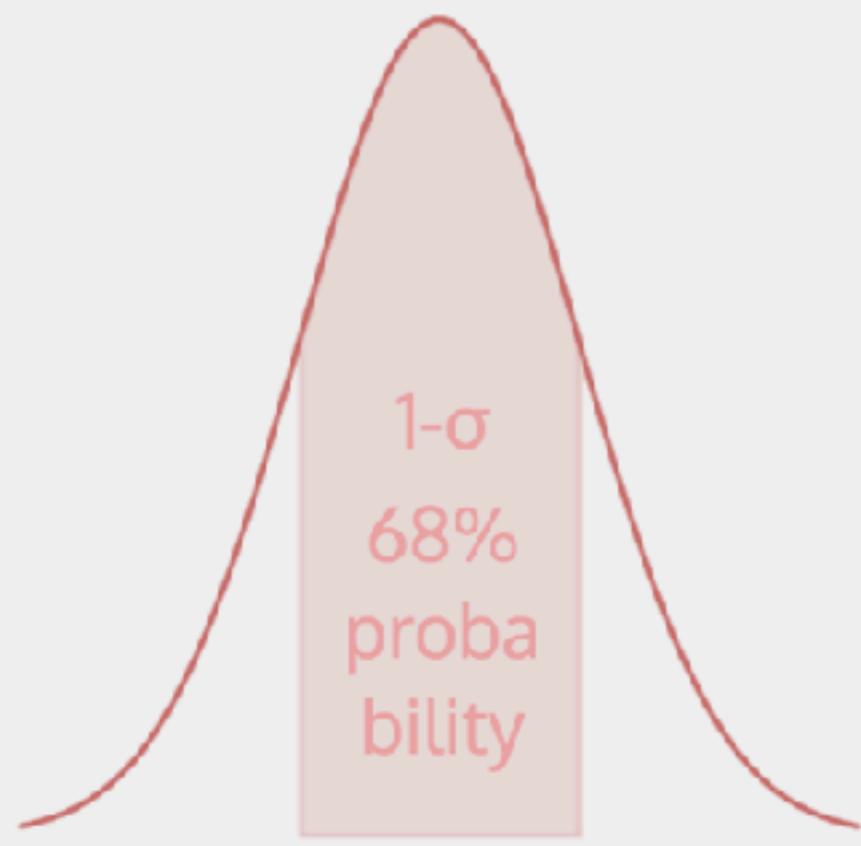
<https://link.springer.com/article/10.1007/s41114-018-0017-4>

TESTING IF A MODEL IS CONSISTENT WITH DATA IN THE FREQUENTIST WORLD

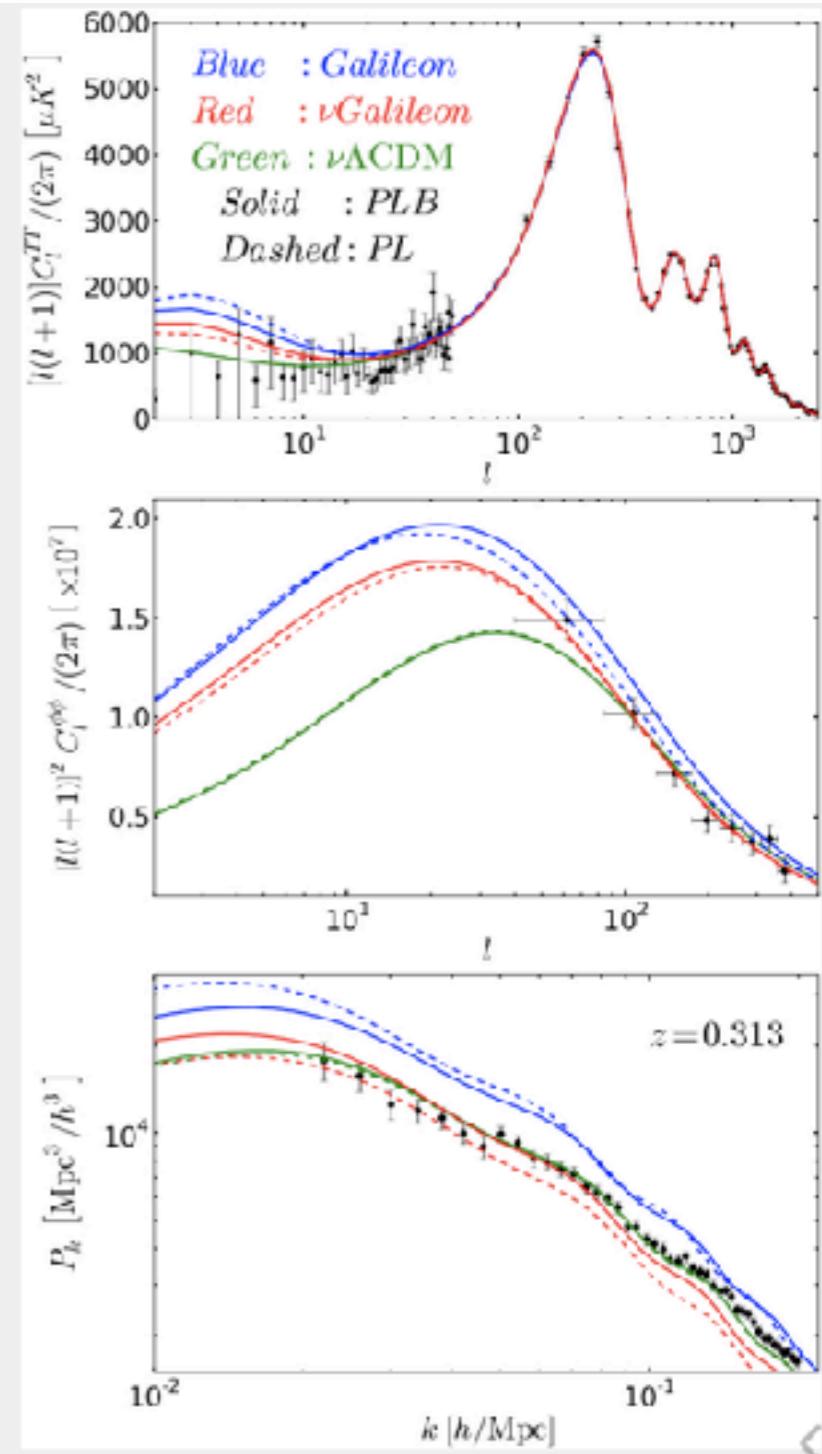
19

- ▶ The thing to note here is that there is a model that has been "fit" to some noisy data, but the model is taken as "Truth."
- ▶ There is no uncertainty reported about the model.

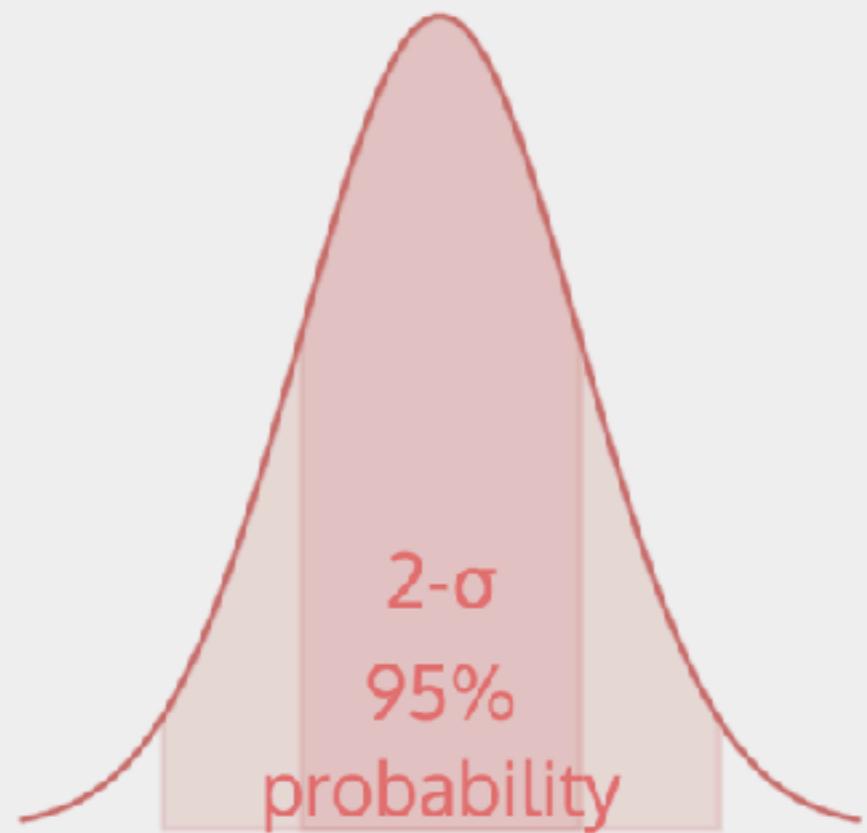
uncertainties: $1-\sigma$



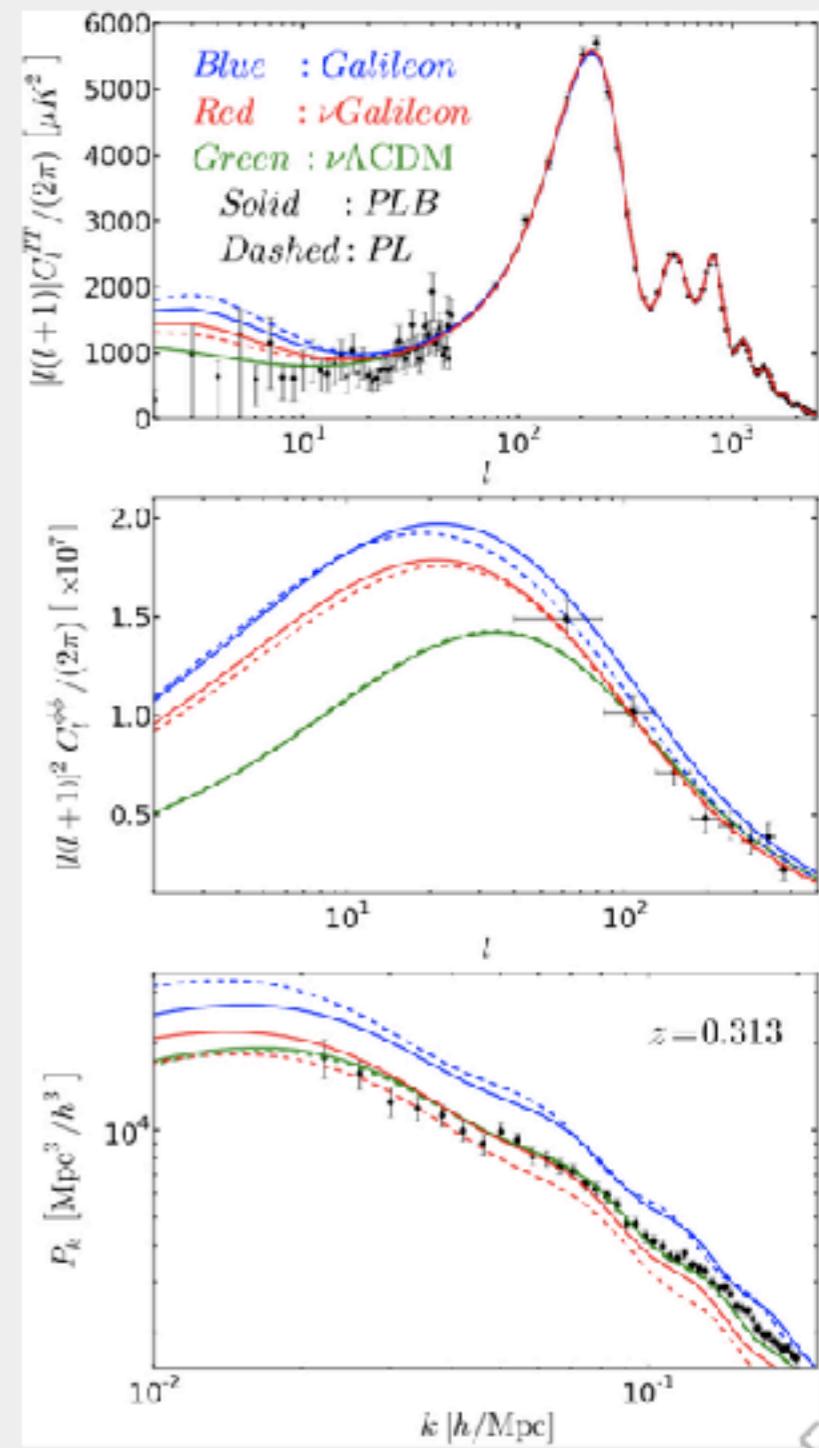
3 points out of 10 can be outside of the uncertainties



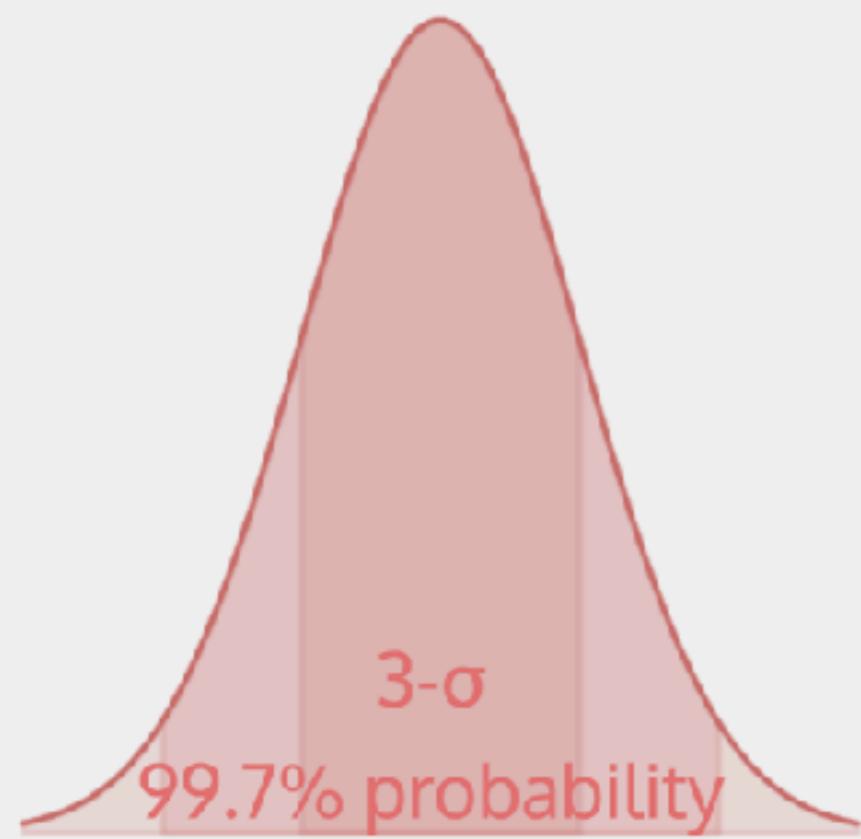
uncertainties: 1σ



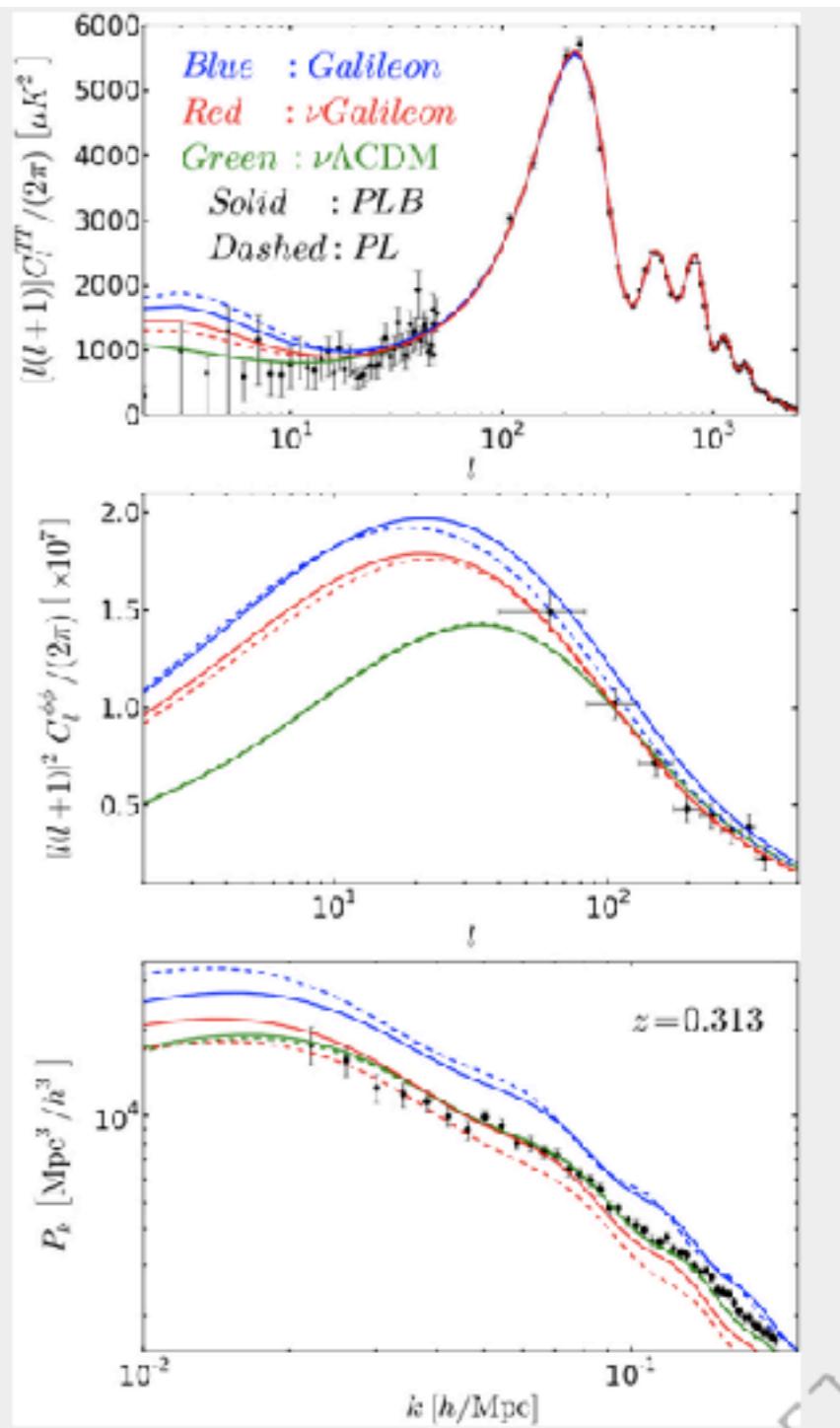
5 points out of 100 can be outside of the uncertainties



uncertainties: $1-\sigma$



3 points out of 1000 can be outside of the uncertainties

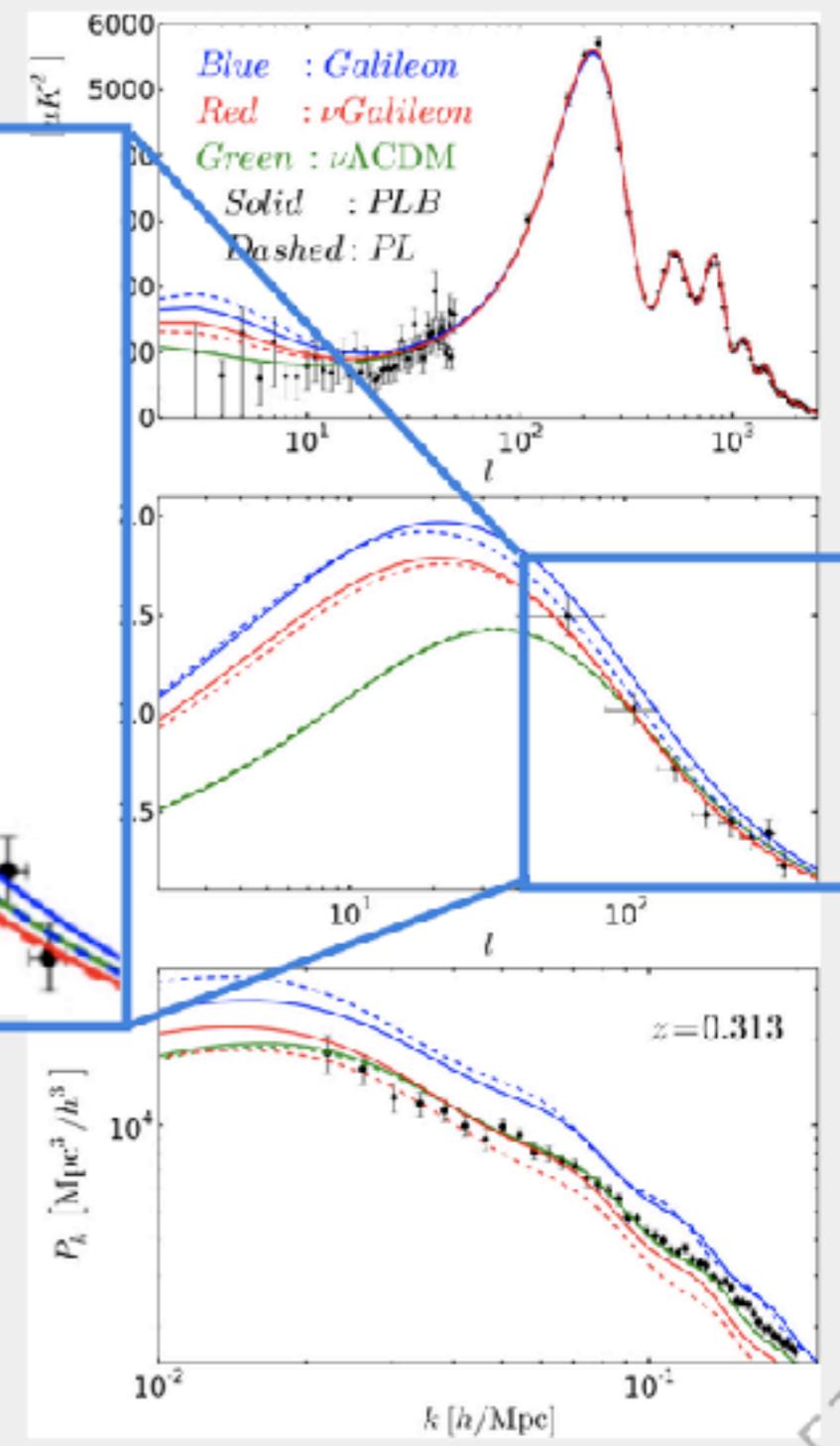
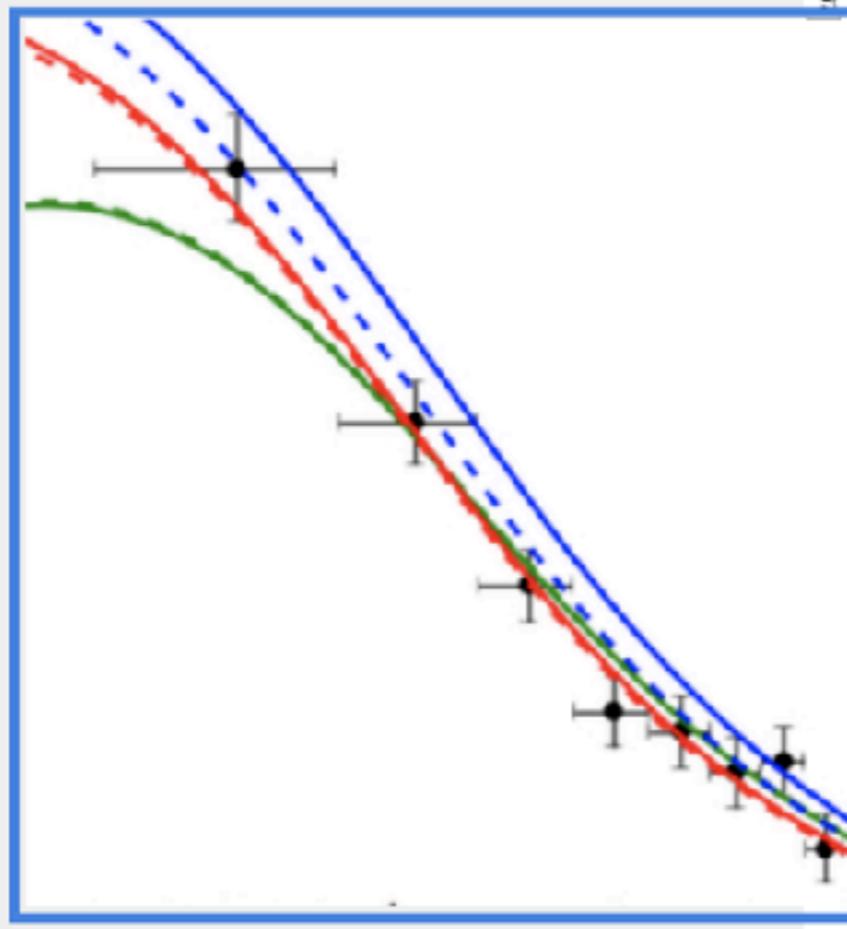


uncertainties: $1-\sigma$



1- σ
68%
probability

3 points out of 10 can be
outside of the uncertainties

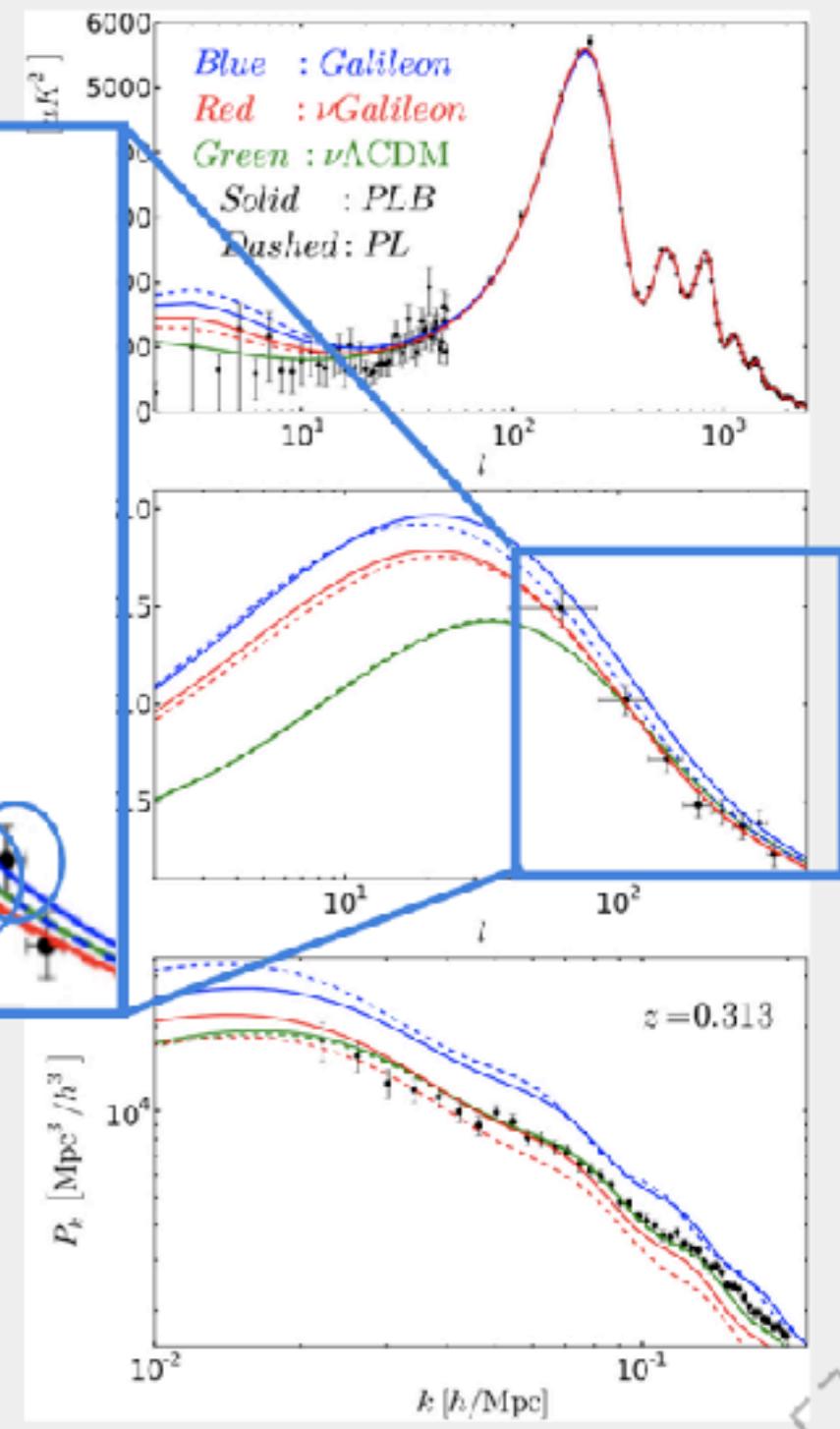
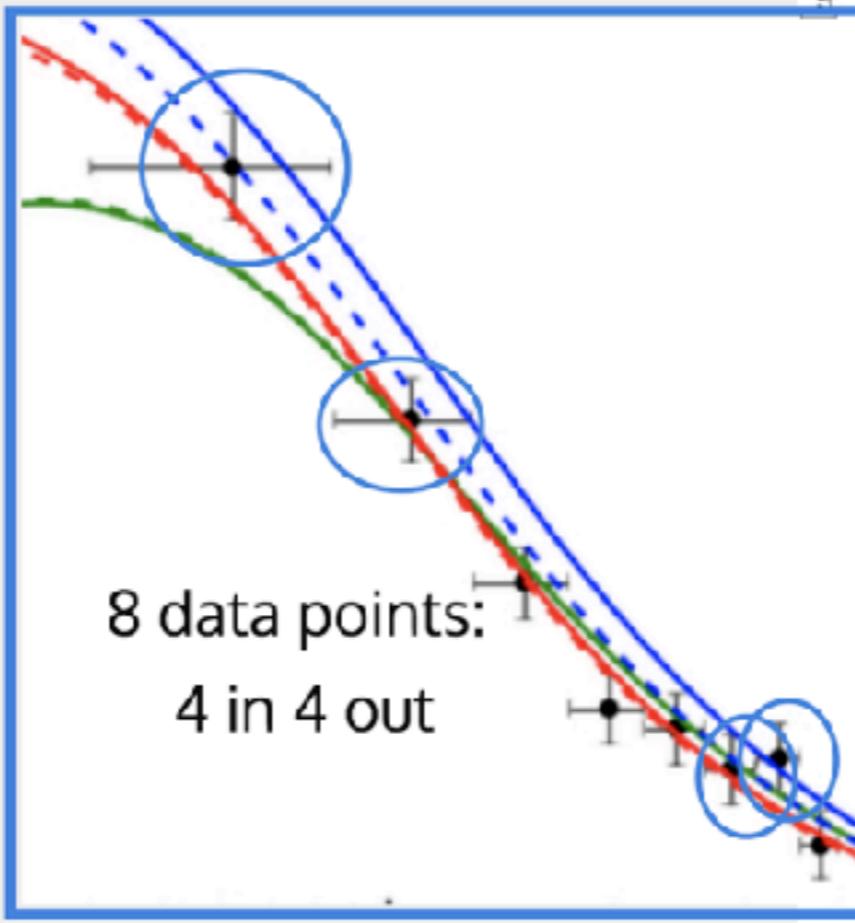


uncertainties: $1-\sigma$



1- σ
68%
probability

7 points out of 10 should be
inside the errorbar

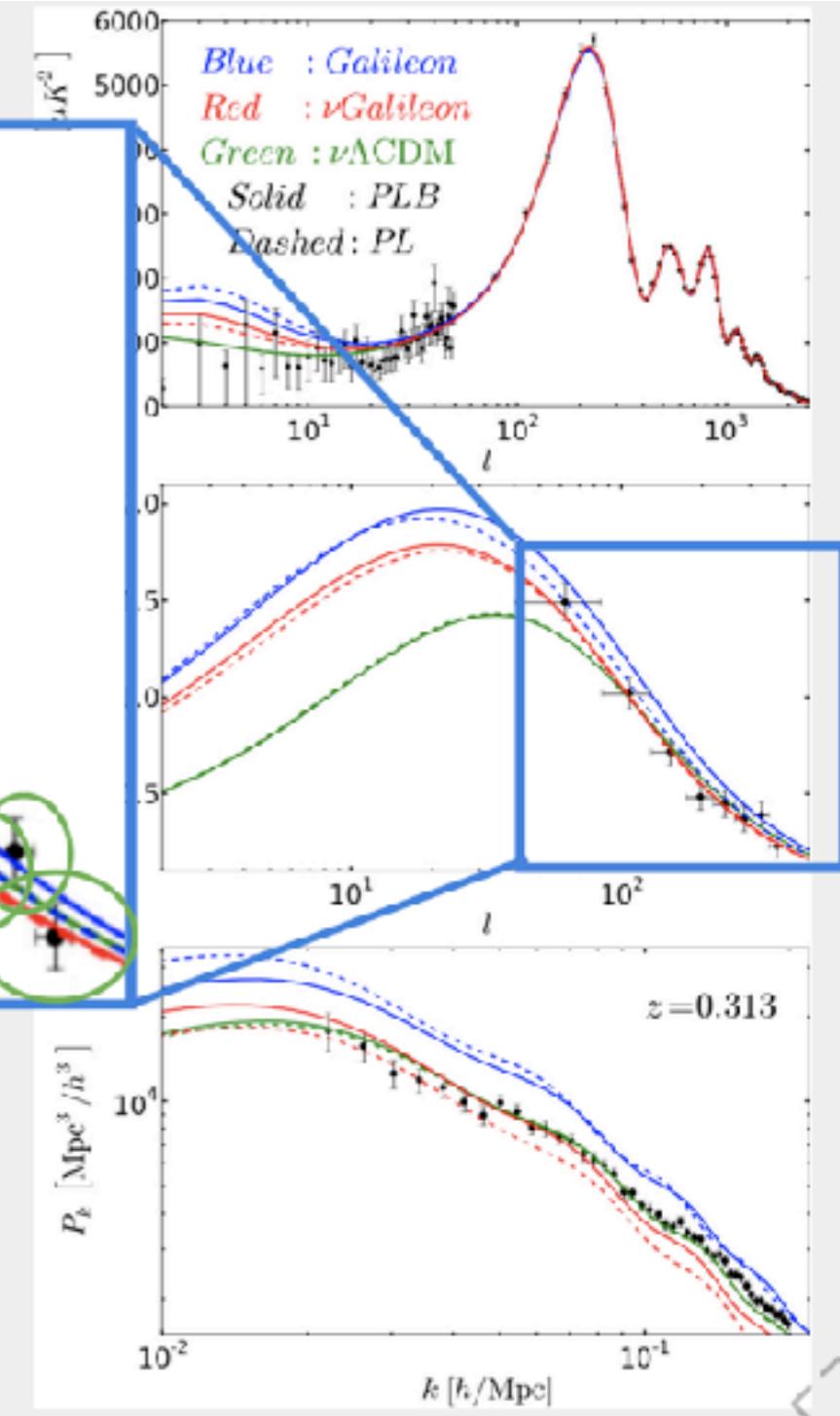
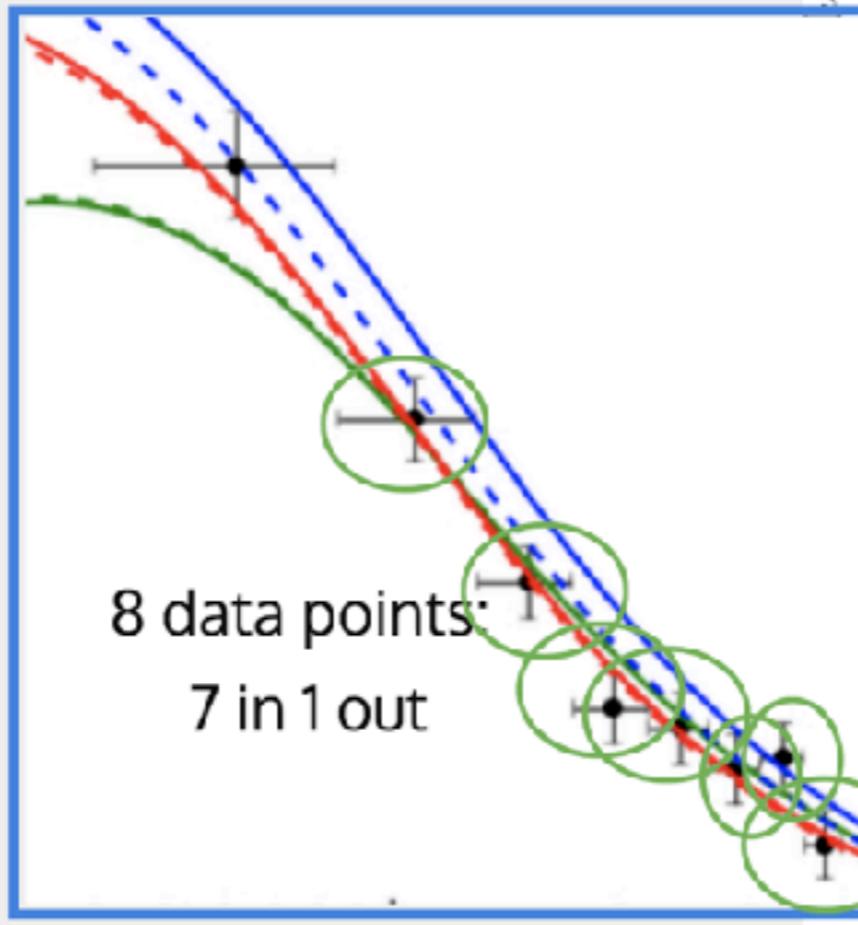


uncertainties: $1-\sigma$

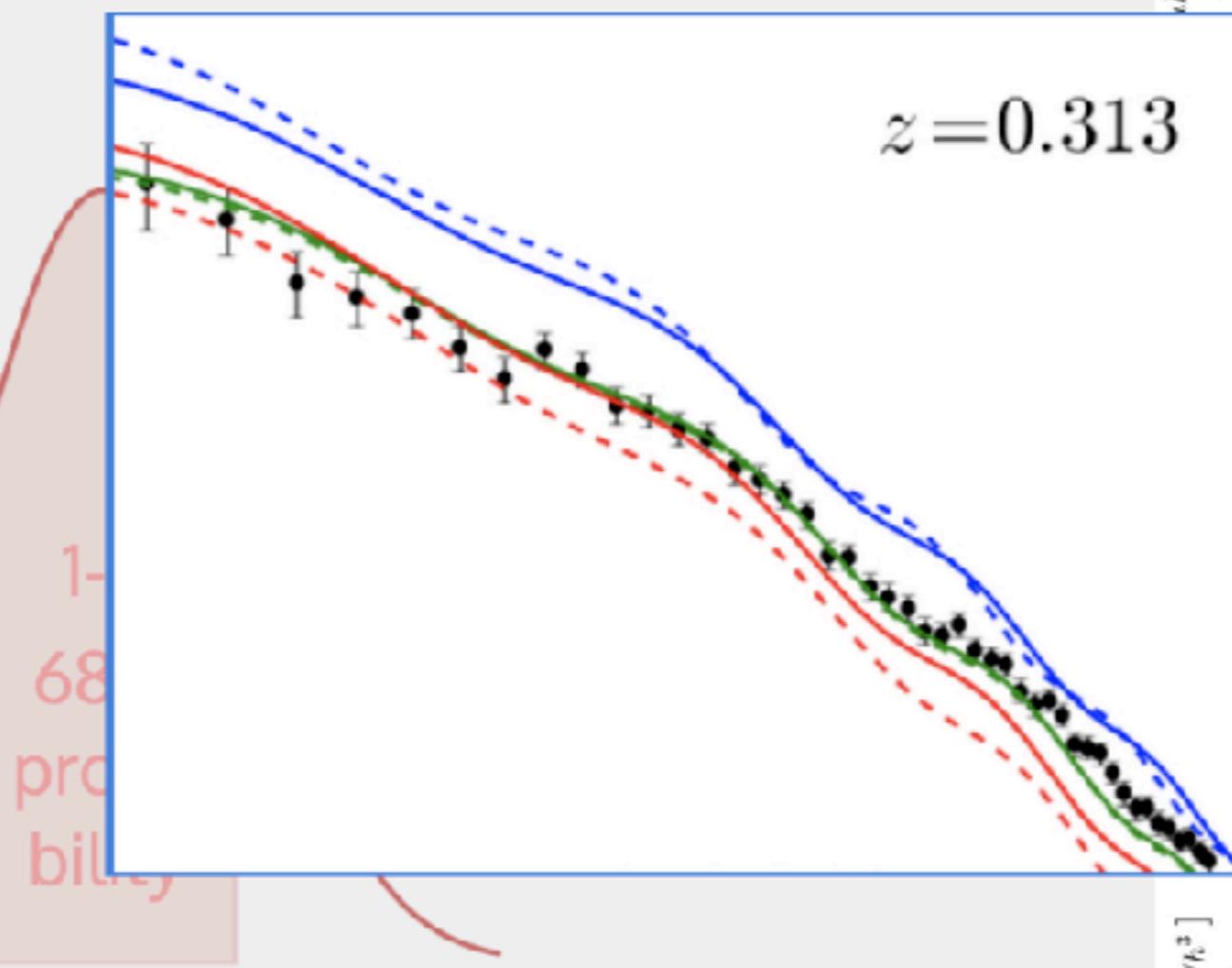


1- σ
68%
probability

7 points out of 10 should be
inside the errorbar

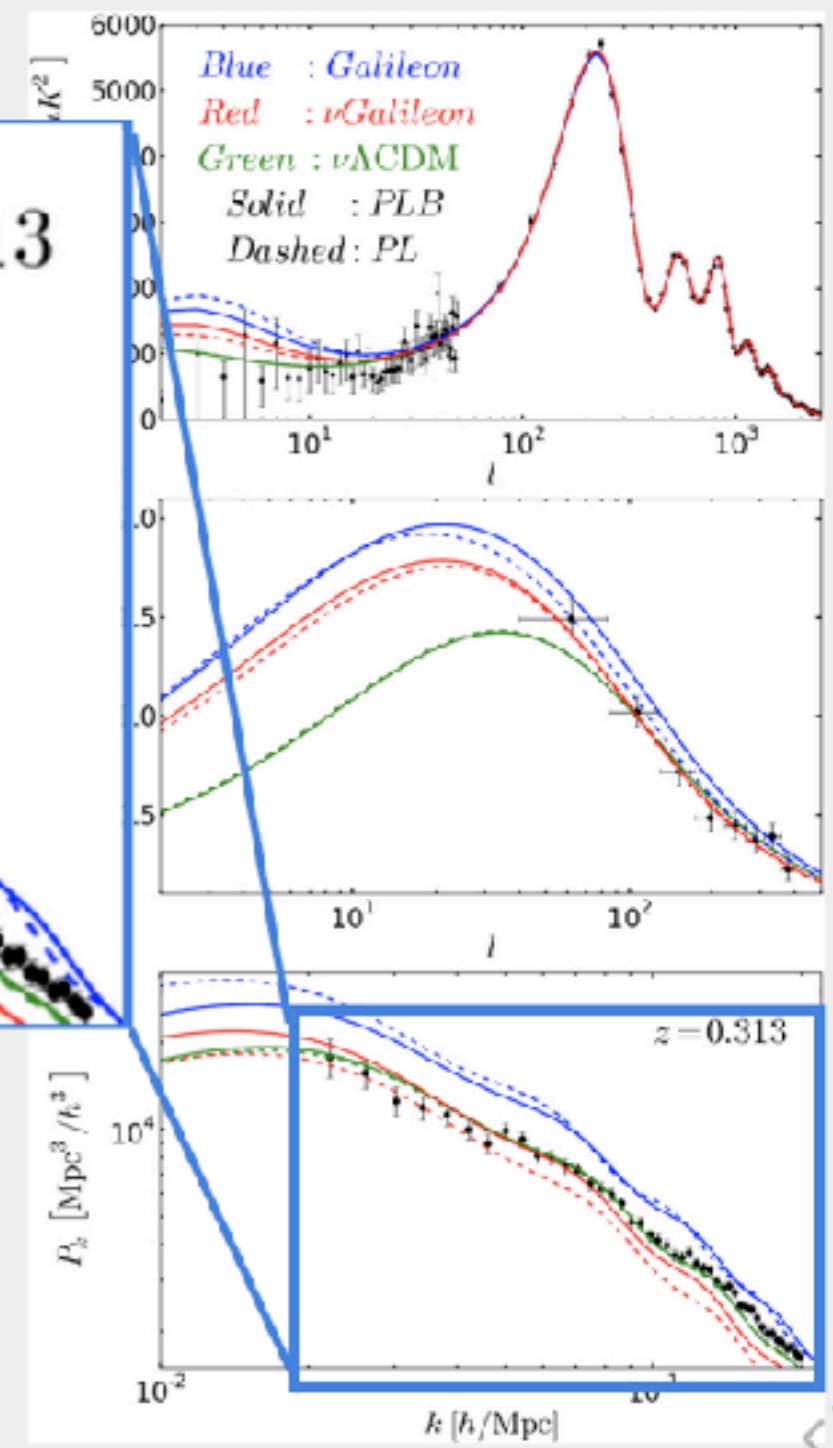


uncertainties: $1-\sigma$



7 points out of 10 should be
inside the errorbar

$z=0.313$



SO YOUR PROBLEM THEN IS HOW TO EVALUATE THE POSTERIOR

27

Posterior

How probable is the hypothesis given the data we observed

Likelihood

How probable is the data given the hypothesis is true

Prior

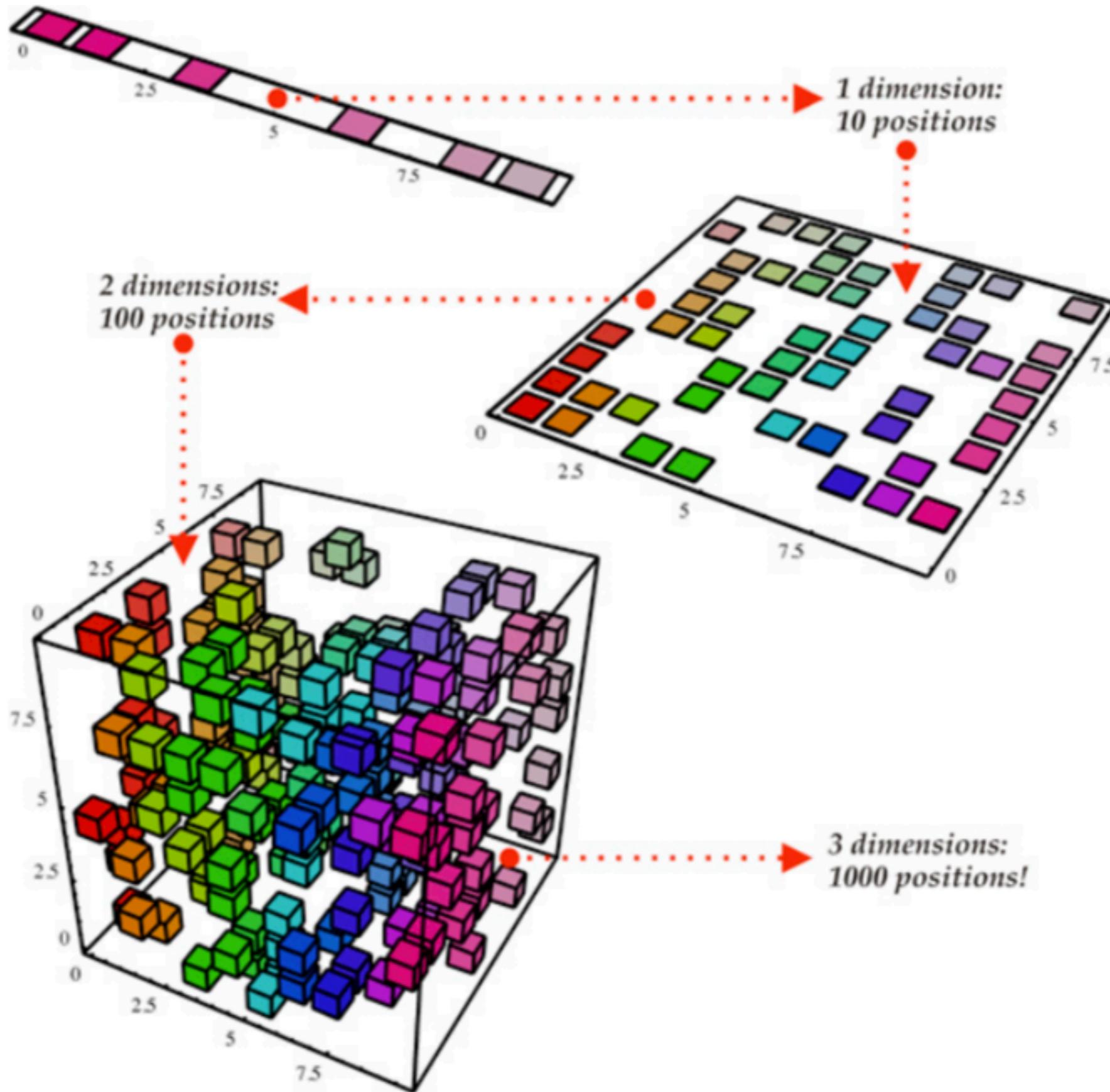
How probable was the hypothesis before we observed anything

$$p(\text{Hypothesis}|\text{Data}) = \frac{p(\text{Data}|\text{Hypothesis})p(\text{Hypothesis})}{p(\text{Data})}$$

Evidence

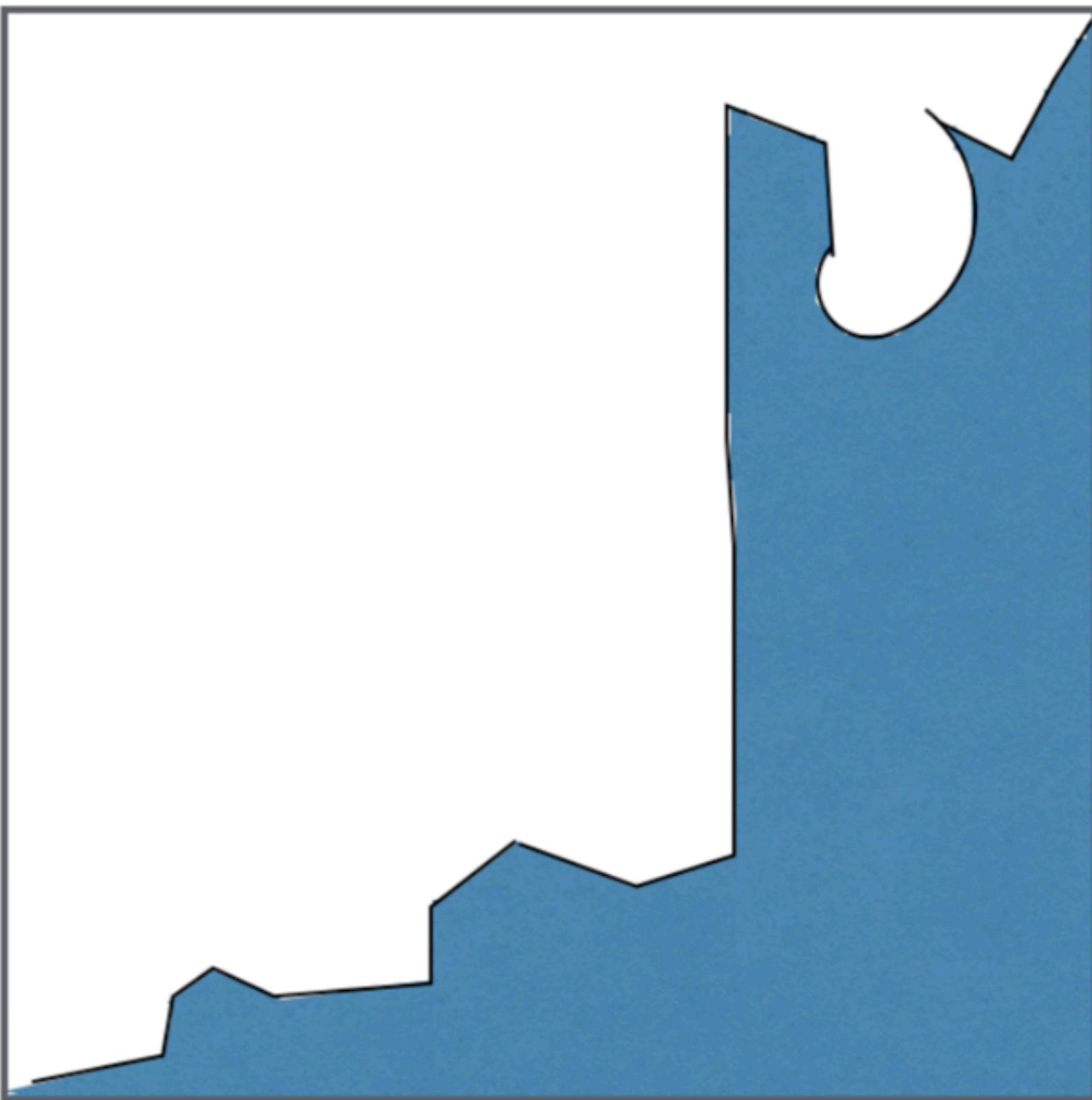
How probable is the data over all possible hypotheses

Curse of Dimensionality

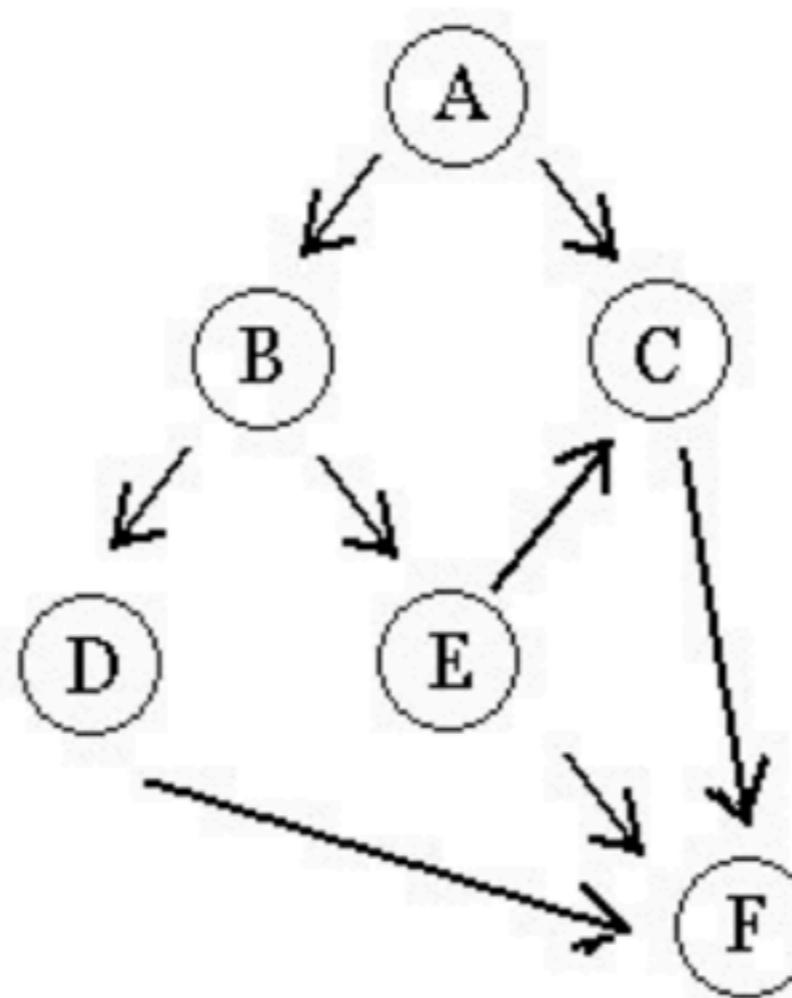
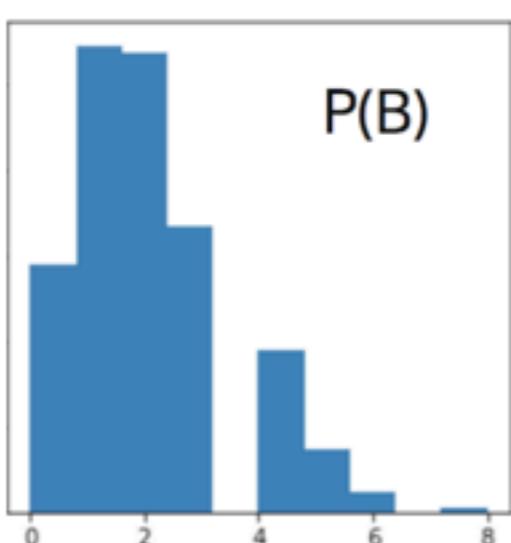
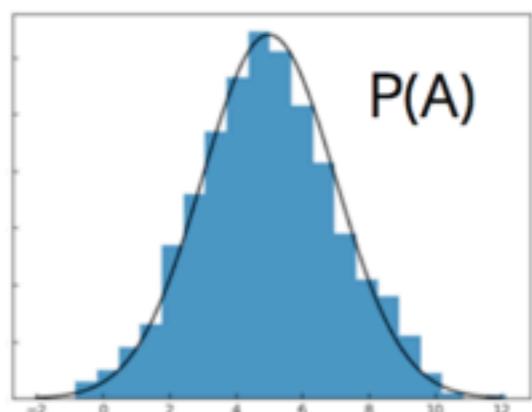


Simple Example

Area of a rectangle = Base × Height Area of a triangle = $1/2$ Base × Height What about the area of this?



Why am I bothering with areas? - Expectation values are related to areas



$$A \sim P(A)$$

$$B \sim P(B|A)$$

$$C \sim P(C|A,E)$$

$$D \sim P(D|B)$$

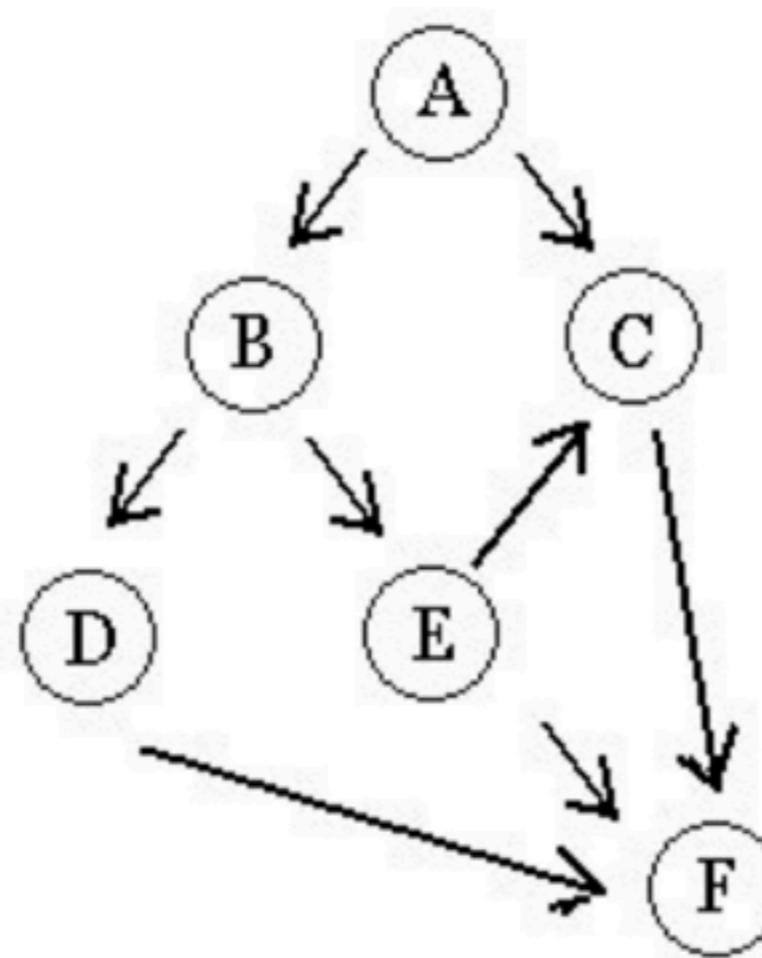
$$E \sim P(E|B)$$

$$F \sim P(F|C,D,E)$$

Why am I bothering with areas? - Expectation values are related to areas



The final probability is likely very complicated (especially if this is a complex system with feedback loops as many physics systems, e.g. radiative transfer!). It may not be tractable analytically but can be simulated



$$A \sim P(A)$$

$$B \sim P(B|A)$$

$$C \sim P(C|A,E)$$

$$D \sim P(D|B)$$

$$E \sim P(E|B)$$

$$F \sim P(F|C,D,E)$$

Why am I bothering with areas? - Expectation values are related to areas

A person's depth

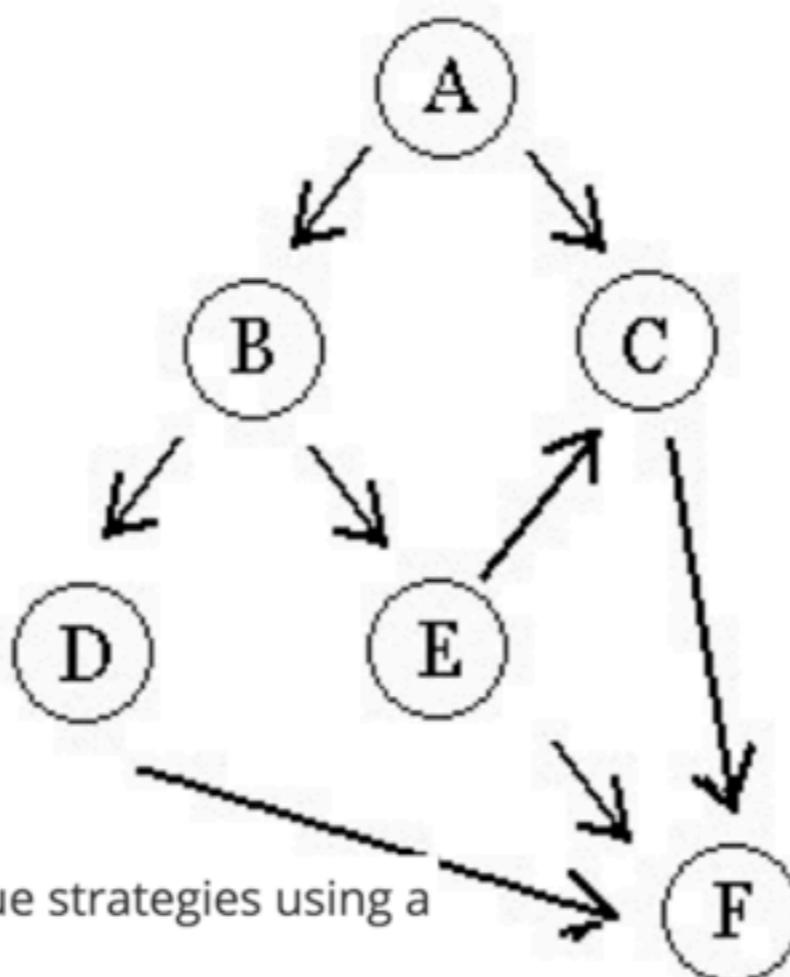
B prob to find them at time t

C they survive the avalanche

D they are still alive at t

E can be resuscitated at time t

F person survives



$$\begin{aligned} A &\sim P(A) \\ B &\sim P(B|A) \\ C &\sim P(C|A,E) \\ D &\sim P(D|B) \\ E &\sim P(E|B) \\ F &\sim P(F|C,D,E) \end{aligned}$$

A concept for optimizing avalanche rescue strategies using a Monte Carlo simulation approach

Ingrid Reiweger , Manuel Genswein , Peter Paal, Jürg Schweizer

Published: May 3, 2017 • <https://doi.org/10.1371/journal.pone.0175877>

Courtesy: Federica Bianco

A history of the Monte Carlo method and the Manhattan project:

<https://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-88-9068>

In general, evaluating the posterior throughout the entire parameter space is too costly.

We want to focus resources on mapping the posterior where it is non-tiny.

Generating samples from the posterior itself automatically accomplishes this.

Sampling and numerical integration

Almost always, we are ultimately interested in *integrals* of the posterior, i.e. marginal distributions of parameters. The tasks of Monte Carlo sampling and **Monte Carlo integration** are essentially indistinguishable. (Similar machinery is useful for difficult optimization problems.)

The essence of MC integration:

$$\int w(x) p(x) dx = \int w(x) dP(x) \approx \overline{w(x_i)}; \quad x_i \sim P$$

i.e., if we can factor the integrand into a PDF and a weight, and sample from the PDF, then our integral becomes an *average over the samples*.

In other words, given a list of samples of θ from $p(\theta)$,

- the marginalized 1D posterior for θ_0 is estimated by making a histogram of θ_0 samples
- the marginalized 2D posterior for θ_0, θ_1 is estimated from a 2D histogram of θ_0, θ_1 samples
- statistics like the mean or percentiles of the posterior are estimated directly from the samples

All of these computations would be weighted if $w(\theta) \neq 1$.

Random number generation

Useful terms to know:

- Random: predictable only in the sense of following a PDF
- Pseudorandom: not random, but "unpredictable enough" for practical purposes. Various computer algorithms produce pseudorandom sequences that approximate the uniform distribution on [0,1).
- Quasirandom: sequence that doesn't even pretend to be random, but does converge to a target PDF *more quickly* than a random or pseudorandom process would

Series

We've quietly moved from talking about single measurements to ensembles of measurements

Specifically, we introduced the concept of quantities as draws from some underlying distribution - i.e. “random variable” and multiple measurements as I.I.D random variables.

So we need a way to keep track of N measurements

The obvious solution is to introduce an index
 $x_0, x_1, x_2, x_3, x_4\dots$

- ▶ We've come up with a model, an objective/loss/likelihood function and some priors
 - ▶ Now we actually want to evaluate the posterior $P(\theta|D)$
 - ▶ analytically is too hard, so we resort to numerical techniques
 - ▶ **Option 1: Evaluate the function on some grid of parameter values - doesn't scale**
 - ▶ **Option 2: we draw samples (i.e. Monte Carlo)**
 - ▶ **convert messy integrals to sums over the samples**

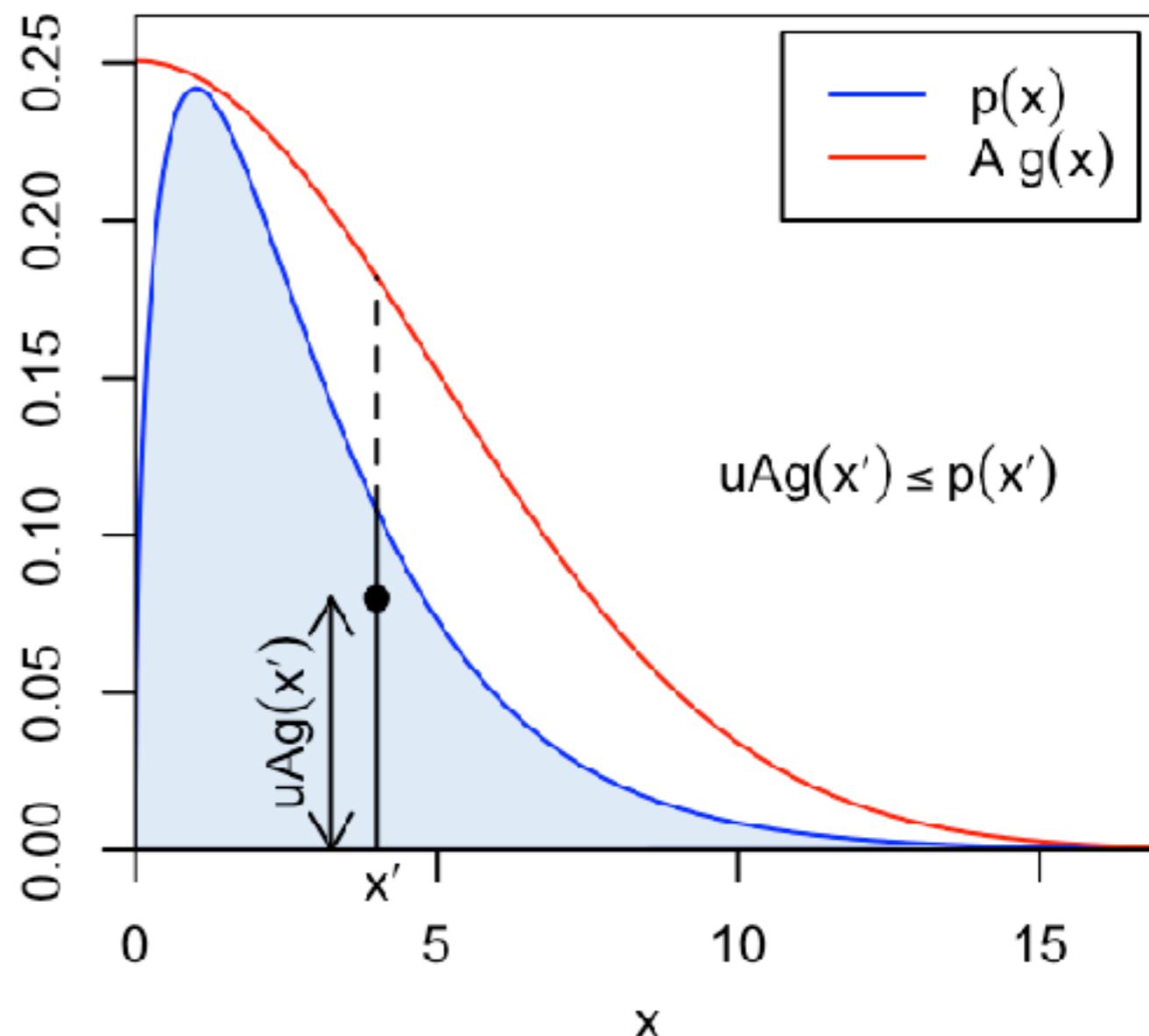
REJECTION SAMPLING

39

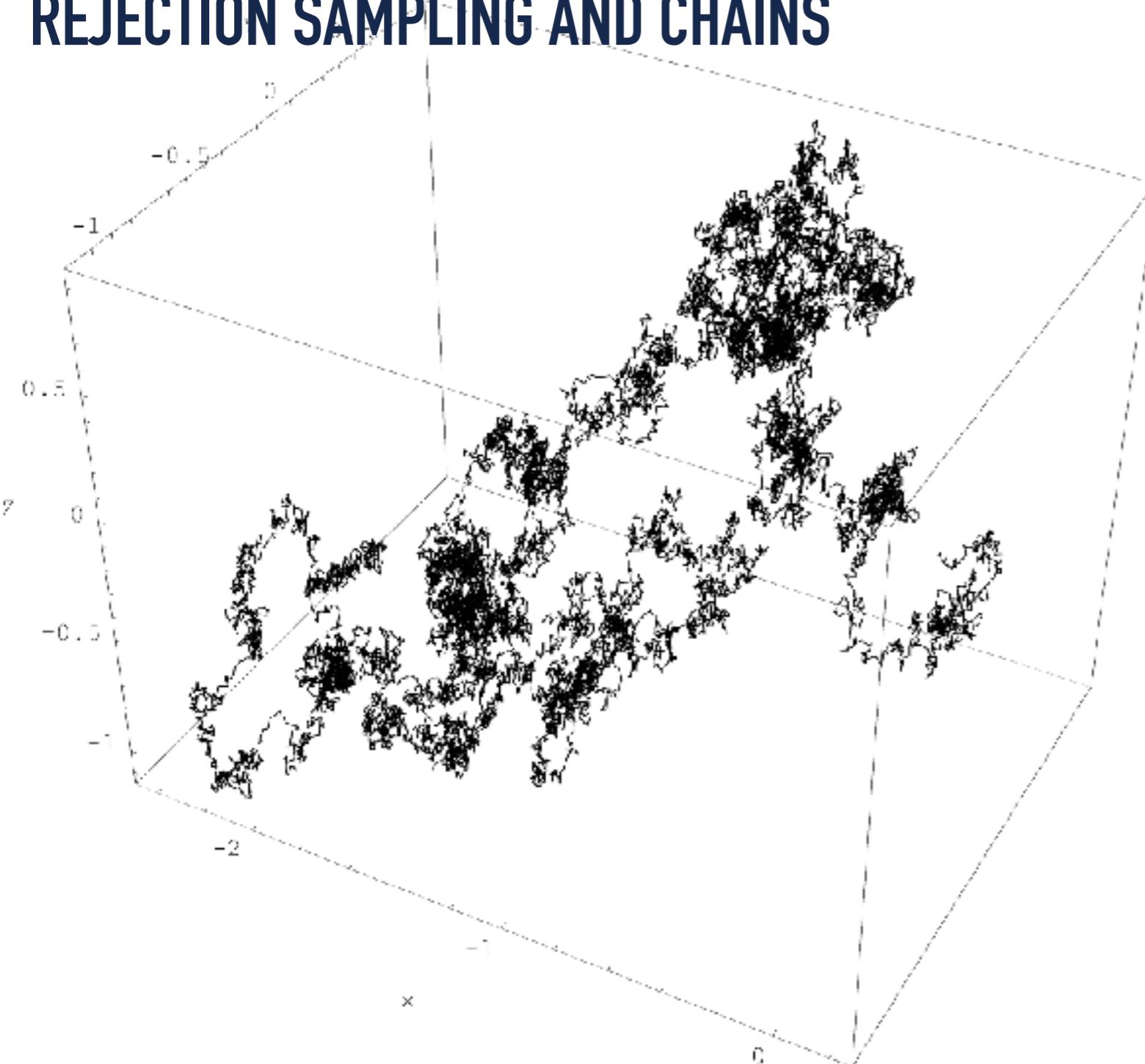
- ▶ Define an envelope function which everywhere exceeds the target PDF, $p(x)$, and can be sampled. Let this be $A.g(x)$ where A is a scaling factor and $g(x)$ is a PDF we know.

Then the algorithm is:

- ▶ while we want more samples
 - ▶ draw a random value for x from $g(x)$
 - ▶ draw u from Uniform(0,1)
 - ▶ if $u \leq p(x)/A*g(x)$, keep the sample x
 - ▶ otherwise, reject x



REJECTION SAMPLING AND CHAINS



- ▶ Start sampling in your parameter space following a prior distribution at \mathbf{x}
- ▶ Compute the likelihood at this location $p(\mathbf{x})$
- ▶ Move to a new location \mathbf{x}'
- ▶ Compute the likelihood at the new location $p(\mathbf{x}')$
- ▶ If it's higher at the new location, keep the new sample \mathbf{x}'
- ▶ If it's lower, check against a random Uniform number draw u , and maybe keep the sample
- ▶ The list of all samples is a **chain** - but this isn't enough to make a *Markov chain*

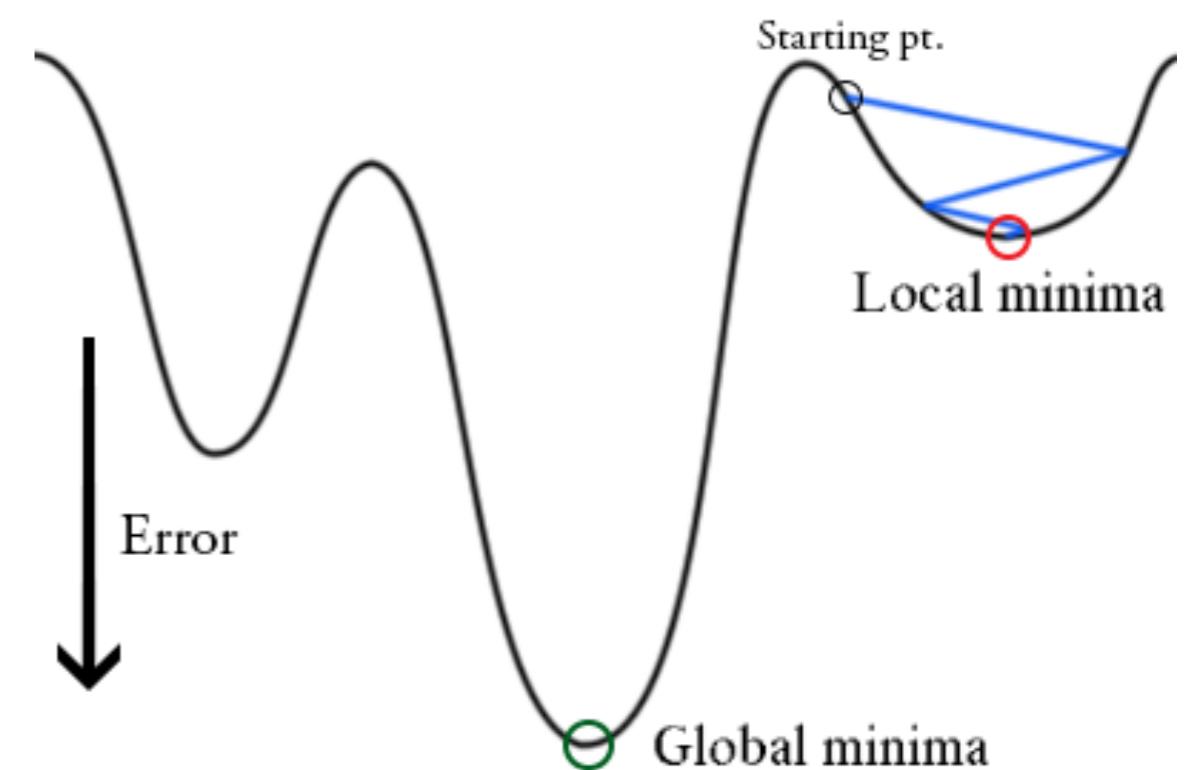
**This comparing against a random number may seem...
random**

41

But the stochasticity is the point!

**Stochasticity allows us to explore the whole surface but
spend more time in interesting spots**

**This will avoid the issues with
local minima**



-
- ▶ We've come up with a model, an objective/loss/likelihood function and some priors
 - ▶ Now we actually want to evaluate the posterior $P(\theta|D)$
 - ▶ analytically is too hard, so we resort to numerical techniques
 - ▶ **Option 1: Evaluate the function on some grid of parameter values - doesn't scale**
 - ▶ **Option 2: we draw samples (i.e. Monte Carlo)**
 - ▶ **convert messy integrals to sums over the samples**
 - ▶ Simple Monte Carlo is wasteful because we don't want to draw samples over all parameter space
 - ▶ one way to make it more efficient was make **samples that are correlated with each other**
 - ▶ keep sampling in regions of high probability, don't sample in regions with low probability
 - ▶ a) If we are lucky enough to have nice functions, we can use **inverse transform sampling**
 - ▶ b) alternately we can use **rejection sampling**
 - ▶ Markov Chains are series of samples where every sample **only** depends on the previous one
 - ▶ **If we build our Markov Chain using rejection sampling, we're doing Markov Chain Monte Carlo (the particular strategy here is called Metropolis-Hastings)**

- ▶ **Ergodic** - given enough time* the entire parameter space

$X_1, X_2, \dots, X_n, X_{n+1}, \dots$



X_{n+1} depends only on X_n
(and not on X_1, X_2, \dots, X_{n-1})

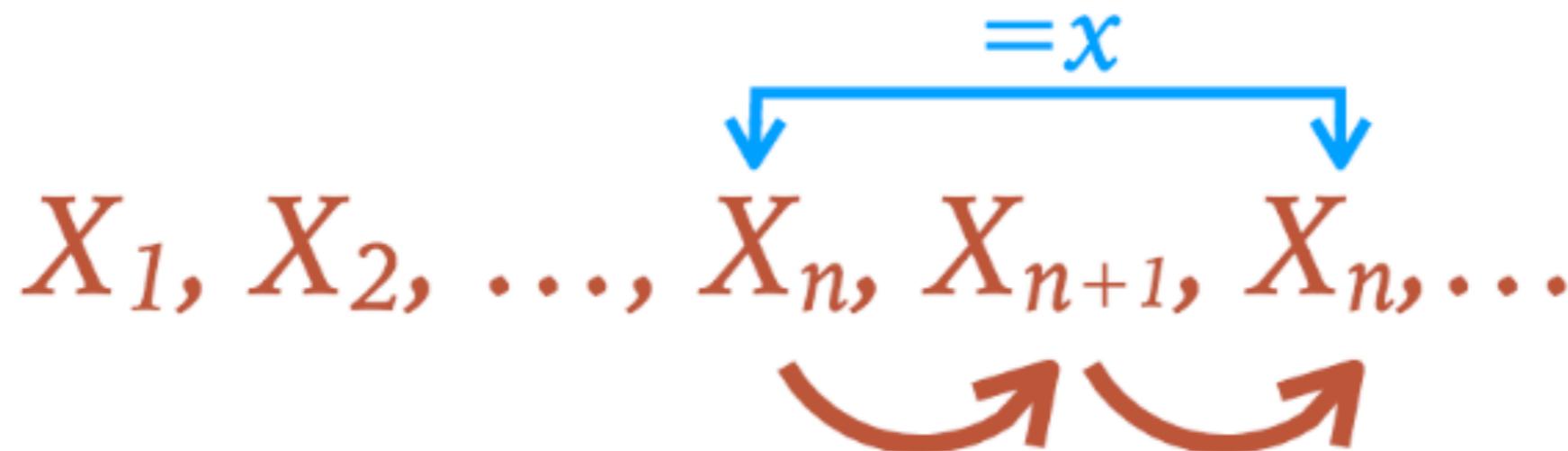
- ▶ **Stationary** - As long as the Markov chain is **positive recurrent** (i.e. you can get to any parameter in a finite number of steps) and is **irreducible** (you can get to every parameter value from every other parameter value) then it has another nice property - it is **stationary**

$X_1, X_2, \dots, X_n, X_{n+1}, \dots, X_m, X_{m+1}, \dots$

$$P(X_{n+1} | X_n) = P(X_{m+1} | X_m)$$

~time invariant

- ▶ **Reversible** - If the probability of getting from point x to x' is the same as the probability of getting from x' to x , then the chain is reversible. This happens if a condition called **detailed balance is satisfied**



$$P(X_{n+1} | X_n = x) = P(X_{n+2} = x | X_{n+1})$$

~time-reversal invariant

- ▶ Detailed balance: the probability of getting from point x to x' is the same as the probability of getting from x' to x - how do we go about getting this property?
- ▶ The Metropolis-Hastings algorithm consists of these steps:
 - ▶ given some $x, p(x)$
 - ▶ and a transition matrix probability $T(x'|x)$, draw a proposed value for x'
 - ▶ compute probability $p(x')$
 - ▶ draw a random number u between 0 and 1 from a uniform distribution; if it smaller than $p(x')$, then accept x'
- ▶ if x' is accepted added it to the chain, if not, add x' to the chain.
- ▶ ***This process is NOT stationary*** - $T(x'|x)$ and $T(x|x')$ don't have to be the same in principle

- ▶ The probability of an arbitrary point from such a chain being located at x' is (marginalizing over the possible immediately preceding points)

$$p(x') = \int dx p(x) T(x' | x)$$

- ▶ where $T(x'|x)$ is the transition probability of a step from x to x' .
- ▶ **We want to have detailed balance**

$$p(x)T(x' | x) = p(x')T(x | x')$$

- ▶ We'll break the transition $\mathbf{T}(\mathbf{x}'|\mathbf{x})$ into two steps:
- ▶ A proposal, $\mathbf{g}(\mathbf{x}'|\mathbf{x})$ and
- ▶ An acceptance ratio, $\mathbf{A}(\mathbf{x}'|\mathbf{x})$
- ▶ i.e.

$$T(x' | x) = A(x' | x)g(x' | x)$$
$$\frac{A(x' | x)}{A(x | x')} = \frac{p(x')g(x | x')}{p(x)g(x' | x)}$$

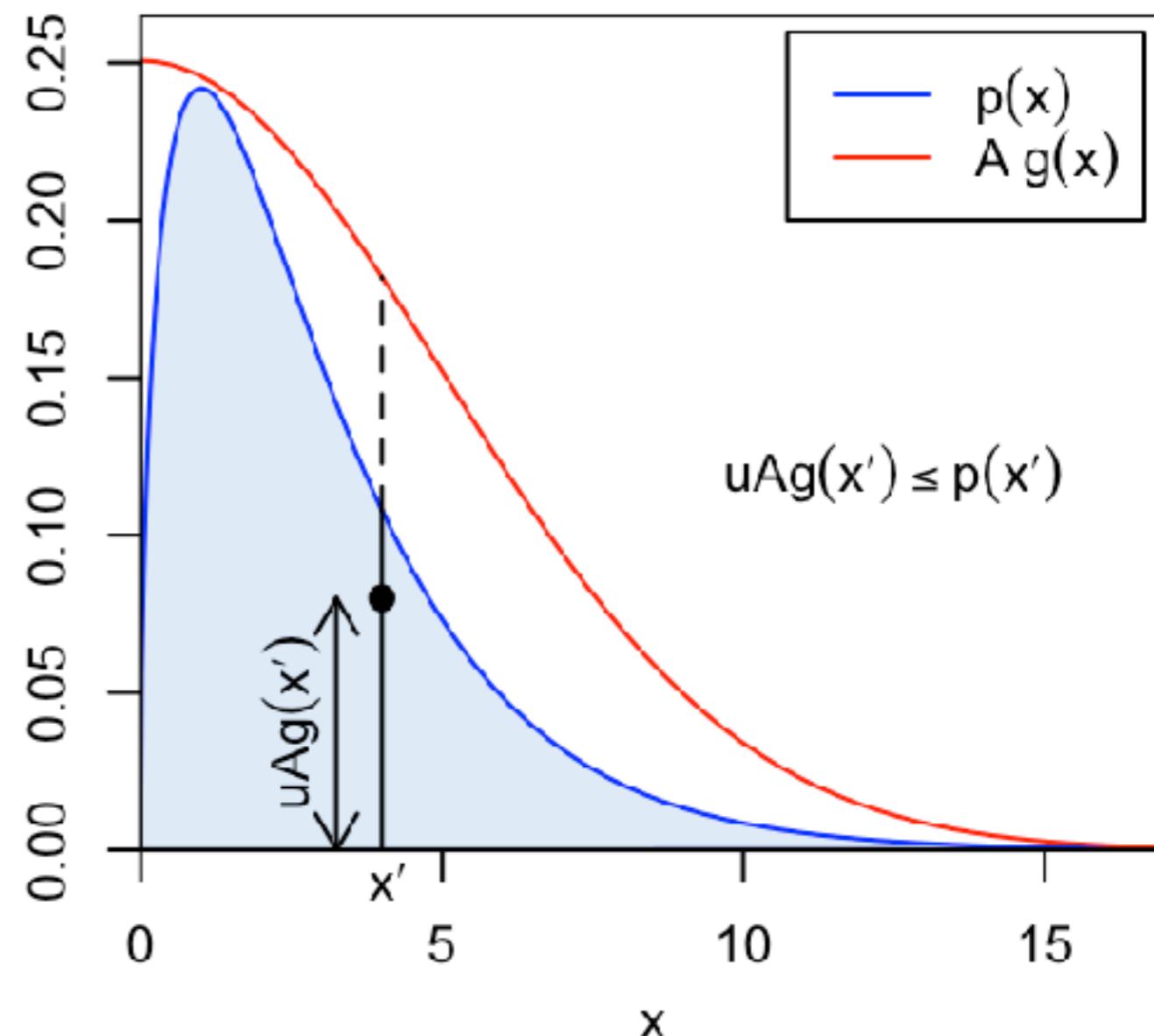
HEY THIS IS JUST REJECTION SAMPLING

- This notation of $\mathbf{p(x)}$, $\mathbf{A(x)}$, $\mathbf{g(x)}$ again is deliberate - this is just rejection sampling at each position in the chain.

$$\frac{A(x' | x)}{A(x | x')} = \frac{p(x')g(x | x')}{p(x)g(x' | x)}$$

- The probability of accepting a proposed step from x to x' then

$$A(x', x) = \min \left[1, \frac{p(x')g(x | x')}{p(x)g(x' | x)} \right]$$



HEY THIS IS JUST REJECTION SAMPLING

So if we identify:

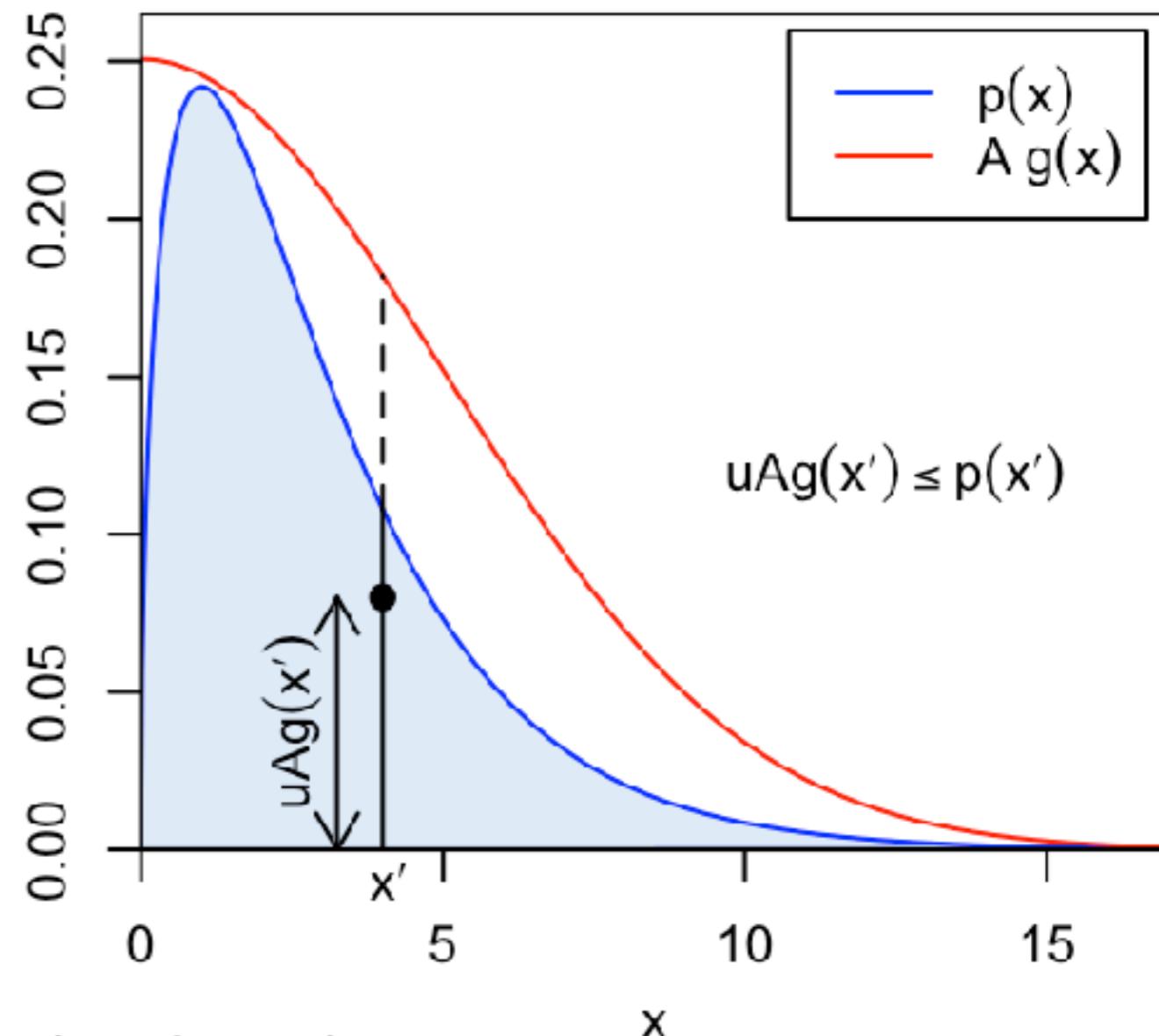
$p(x)$ is the posterior density (probability of being at x , if we're sampling P properly)

$g(x'|x)$ is the proposal distribution (probability of attempting a move to x' from x)

$A(x',x)$ is the probability of accepting the proposed move

With this definition of A , detailed balance is automatically satisfied, and we can **guarantee our chain is time-reversal invariant**

$$p(x)g(x' | x)A(x', x) \equiv p(x')g(x | x')A(x, x')$$



DETAILED BALANCE GIVES YOU REVERSIBLE MARKOV CHAINS

51

$$p(x)g(x' \mid x)A(x', x) \equiv p(x')g(x \mid x')A(x, x')$$

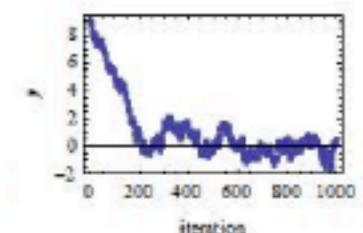
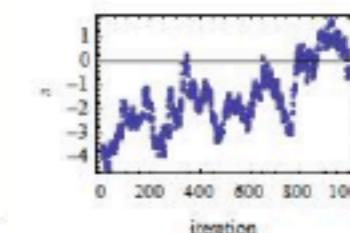
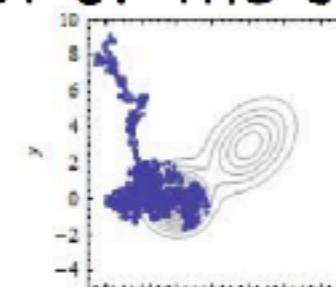


IMPORTANT:

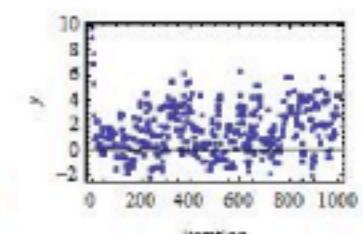
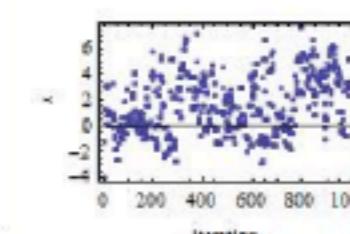
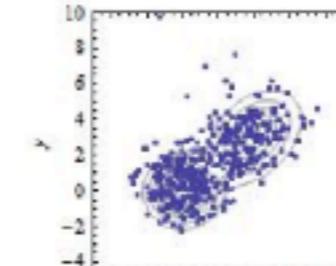
- ▶ Even if a step is rejected, we still keep a sample (i.e. the original state \mathbf{x} , without moving).
- ▶ The difficulty of finding a temptingly better point is important information!

Effect of the sampling distribution

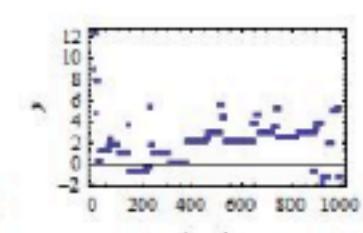
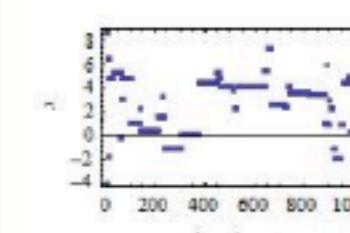
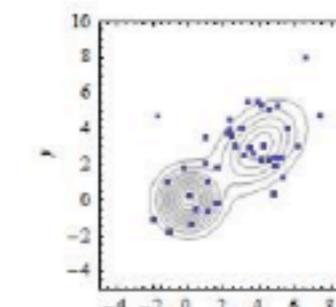
Gaussian proposal distribution with $\sigma = 0.2$, acceptance rate = 85.1%



Gaussian proposal distribution with $\sigma = 2.2$, acceptance rate = 37.9%



Gaussian proposal distribution with $\sigma = 10.2$, acceptance rate = 4.1%



- ▶ The equilibrium state of this chain is the stationary distribution of samples we desire - the posterior
- ▶ If the chain isn't in equilibrium, well tough luck. You're guaranteed it'll get there eventually... you need to tune your chain so it gets to equilibrium reasonably **efficiently**.
- ▶ So we need some metrics to diagnose if:
 - ▶ the chain has converged to the posterior distribution - **i.e. is the chain stationary?**
 - ▶ the chain provides enough effectively independent samples to characterize the posterior

THIS BOILS DOWN TO:

54

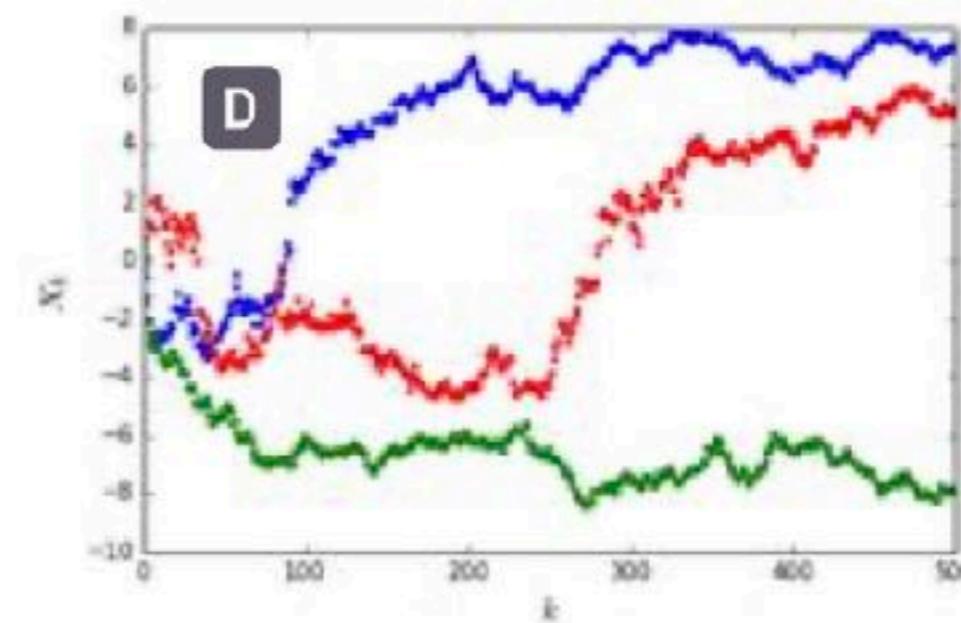
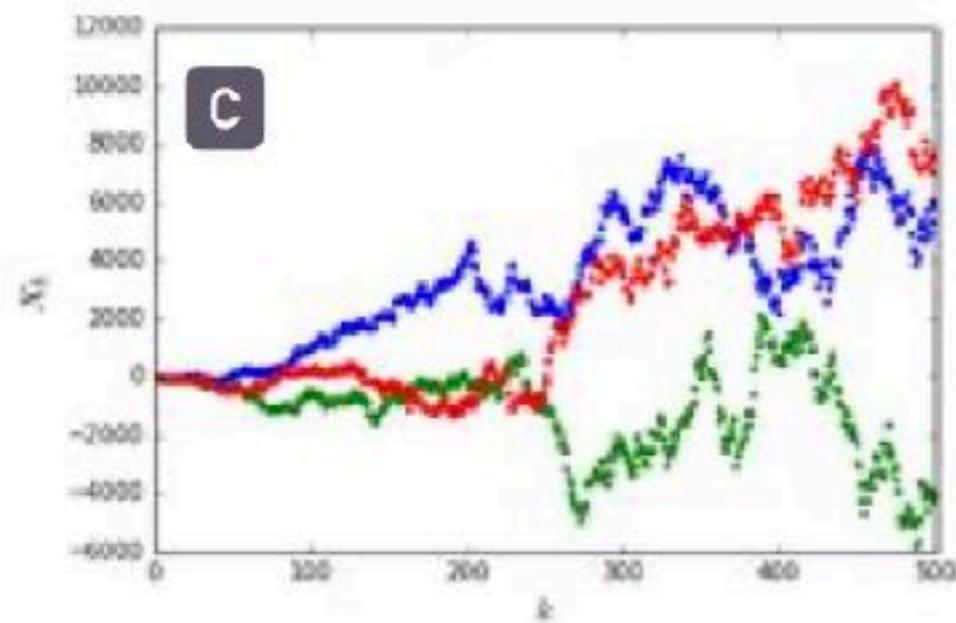
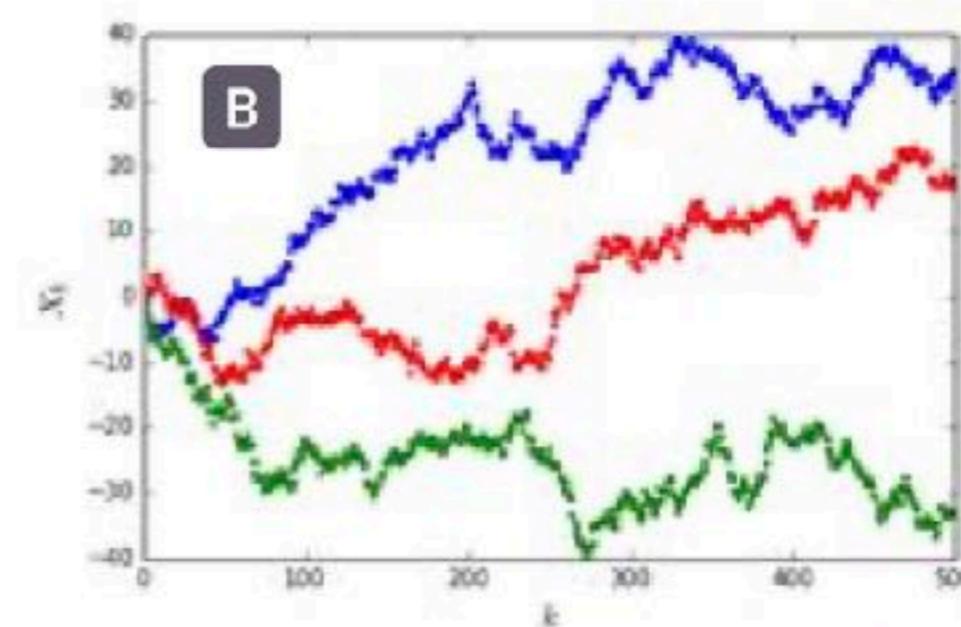
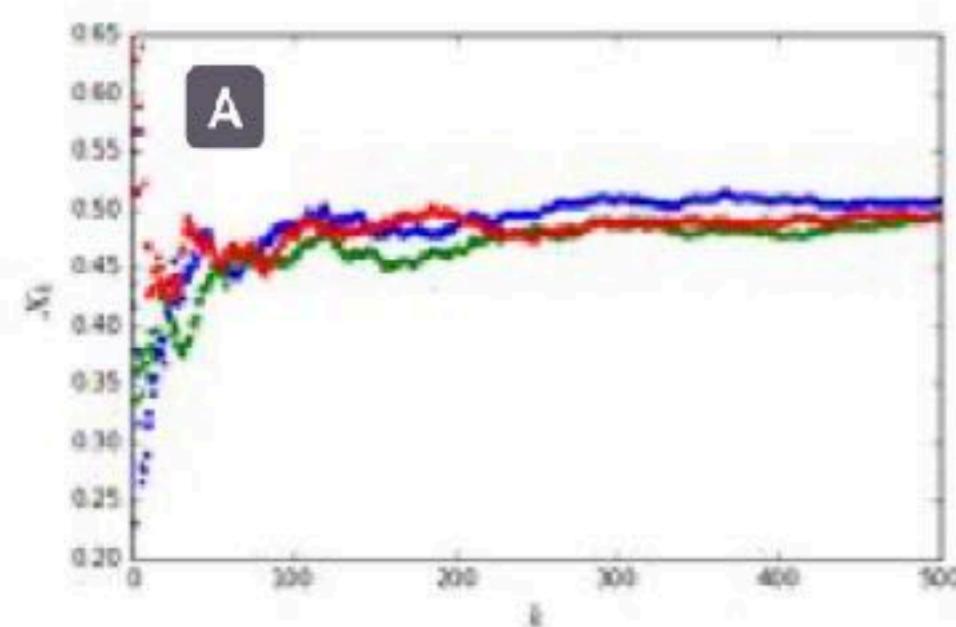
- ▶ A) How do you choose $g(\mathbf{x}, \mathbf{x}')$
 - ▶ This effectively determines what the algorithm's name is - many options...
 - ▶ Gibbs sampling - move along one parameter at a time but you know conditional distribution of each variable and always accept every sample
 - ▶ Simulated annealing - start one chain exploring, but have its step size depend on a "temperature" of the system that cools - initially starts in high temperature - bad moves accepted, but eventually cools, and moves that don't increase the posterior are less likely to be accepted
 - ▶ Parallel tempering - start many parallel chains at once, run some hot (large steps) and some cold (https://www.youtube.com/watch?v=J6FrNf5_G0&list=PLgArfv_fOU5dwjeP_57NO_jnRJ7cWCe6J - you might want to mute this)

- ▶ <http://chi-feng.github.io/mcmc-demo/>
 - ▶ Fiddle on your own time
 - ▶ Sampling efficiency often starts as the primary concern - your code needs to work in a reasonable amount of time
 - ▶ But once you've got something acceptable, the primary concern becomes **Diagnosis** - How do you know that your MCMC is providing you reasonable, independent samples drawn from the posterior

- ▶ What would make us confident of convergence?
 - ▶ Is the chain stationary?
 - ▶ Do independent chains started from overdispersed positions end up in the same stationary state?
- ▶ There's also a trick here - we wanted i.i.d samples from the posterior, but to make our Markov chain efficient, we violated the "independence" criteria a little
 - ▶ Each sample of the chain depends on the previous sample through $\mathbf{g}(\mathbf{x}, \mathbf{x}')$
- ▶ So how do we guess the number of independent samples?
 - ▶ Check how well the chain appears to exploring the distribution
 - ▶ Compare the autocorrelation length scale of the samples with the chain length

ACTIVITY: STATIONARY / NON-STATIONARY / NON-REVERSIBLE?

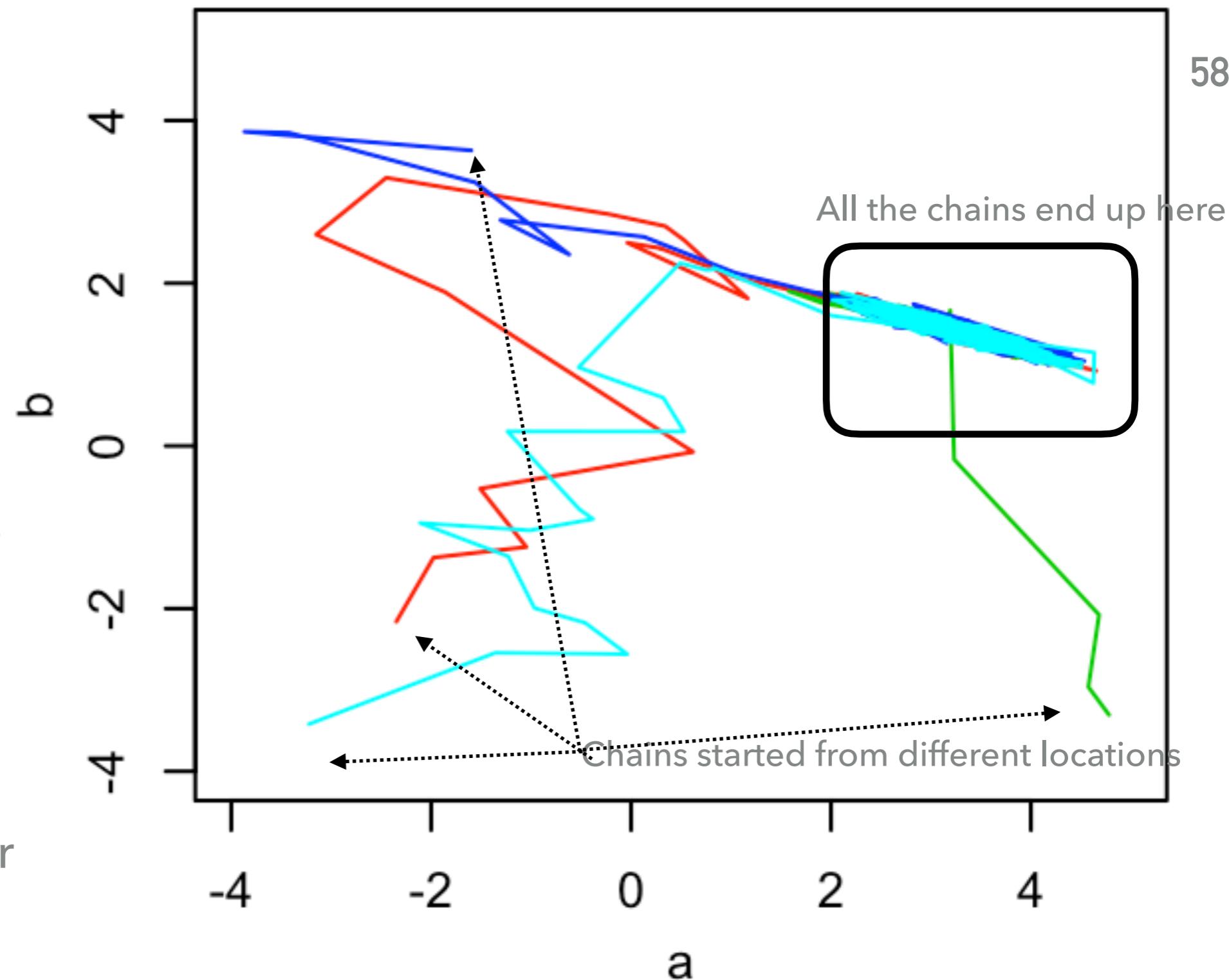
57



10

Setup: here's some chains from a simple linear problem

I've started 4 Metropolis-Hastings chains (colors) from different locations, and had them all sample the posterior distribution

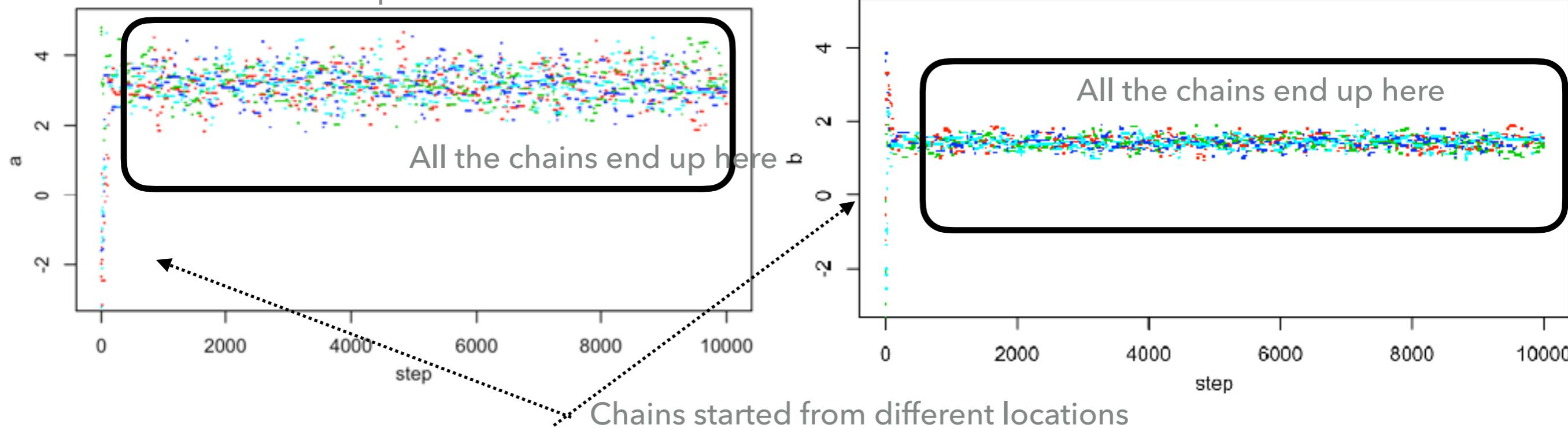


MCMC DIAGNOSTICS

59

- ▶ So you can tell by eye when things are behaving reasonably - how do we take that intuition and express it with mathematics

All the chains end up here



- ▶ **Gelman-Rubin convergence statistic:** This approach tests the similarity of independent chains intended to sample the same PDF. To be meaningful, they should start from different locations and burn-in should be removed.

- ▶ For a given parameter, θ , the R statistic compares the variance across chains with the variance within a chain.
- ▶ Intuitively, if the chains are random-walking in very different places, i.e. not sampling the same distribution, R will be large
- ▶ **Step 1:** Get the variance between each chain's estimate of the parameter and the global estimate, where $\bar{\theta}_j^-$ is the average θ for chain j and $\bar{\theta}^-$ is the global average.

$$B = \frac{n}{m-1} \sum_j (\bar{\theta}_j^- - \bar{\theta}^-)^2$$

- ▶ For a given parameter, θ , the R statistic compares the variance across chains with the variance within a chain.
 - ▶ Intuitively, if the chains are random-walking in very different places, i.e. not sampling the same distribution, R will be large
- ▶ **Step 2:** Get the average variance of the individual-chain variances for θ , where s^2_j is the estimated variance of θ within chain j .

$$W = \frac{1}{m} \sum_j s_j^2$$

- ▶ For a given parameter, θ , the R statistic compares the variance across chains with the variance within a chain.
 - ▶ Intuitively, if the chains are random-walking in very different places, i.e. not sampling the same distribution, R will be large
- ▶ **Step 3:** Get the overall estimate for the variance of θ

$$V = \frac{n-1}{n} W + \frac{1}{n} B$$

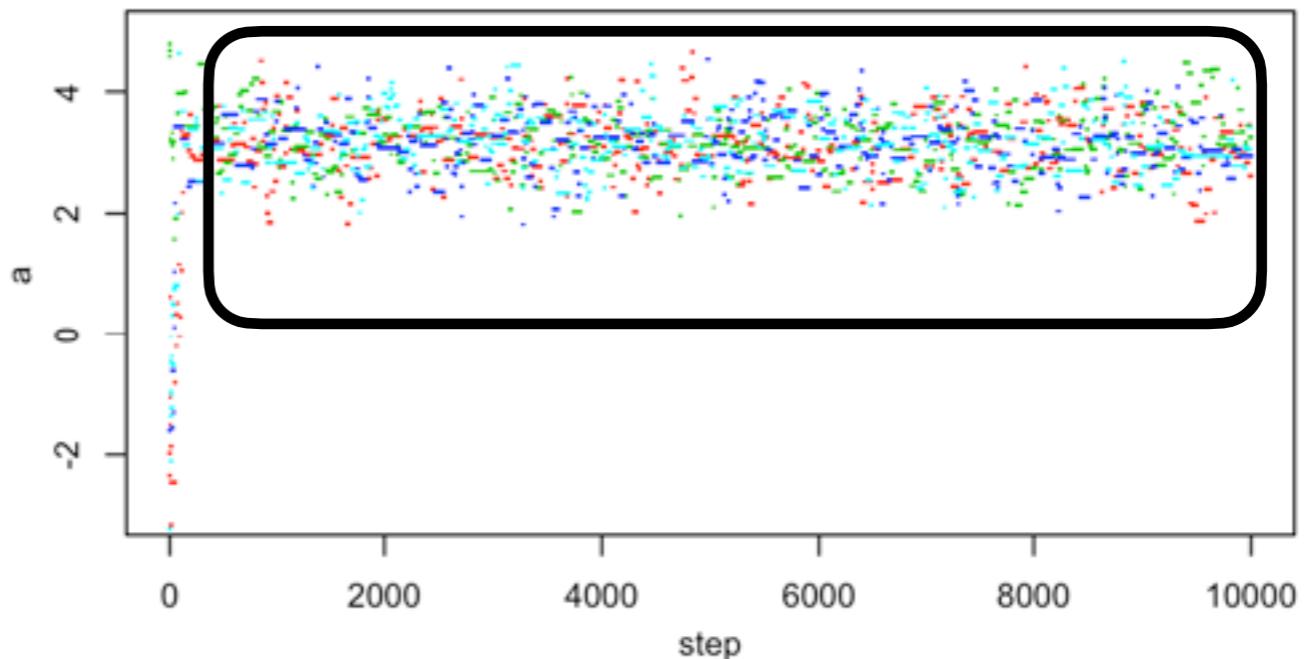
- ▶ For a given parameter, θ , the R statistic compares the variance across chains with the variance within a chain.
 - ▶ Intuitively, if the chains are random-walking in very different places, i.e. not sampling the same distribution, R will be large

- ▶ **Step 4:** Finally

$$R = \sqrt{\frac{V}{W}}$$

- ▶ We'd like to see $R \approx 1$ (e.g. $R < 1.1$ is often used). Note that this calculation can also be used to track convergence of combinations of parameters, or anything else derived from them. **Iff $R \approx 1$, then the chains are described as well mixed.**

- ▶ We also gave up a nice property of Simple Monte Carlo moving to Markov Chain Monte Carlo.
 - ▶ Our samples are now correlated - i.e. you literally took a position and adjusted it by a small amount.
 - ▶ This means that when you request 10,000 samples from Metropolis Hastings, you aren't actually getting 10,000 i.i.d samples precisely **because they aren't independent.**



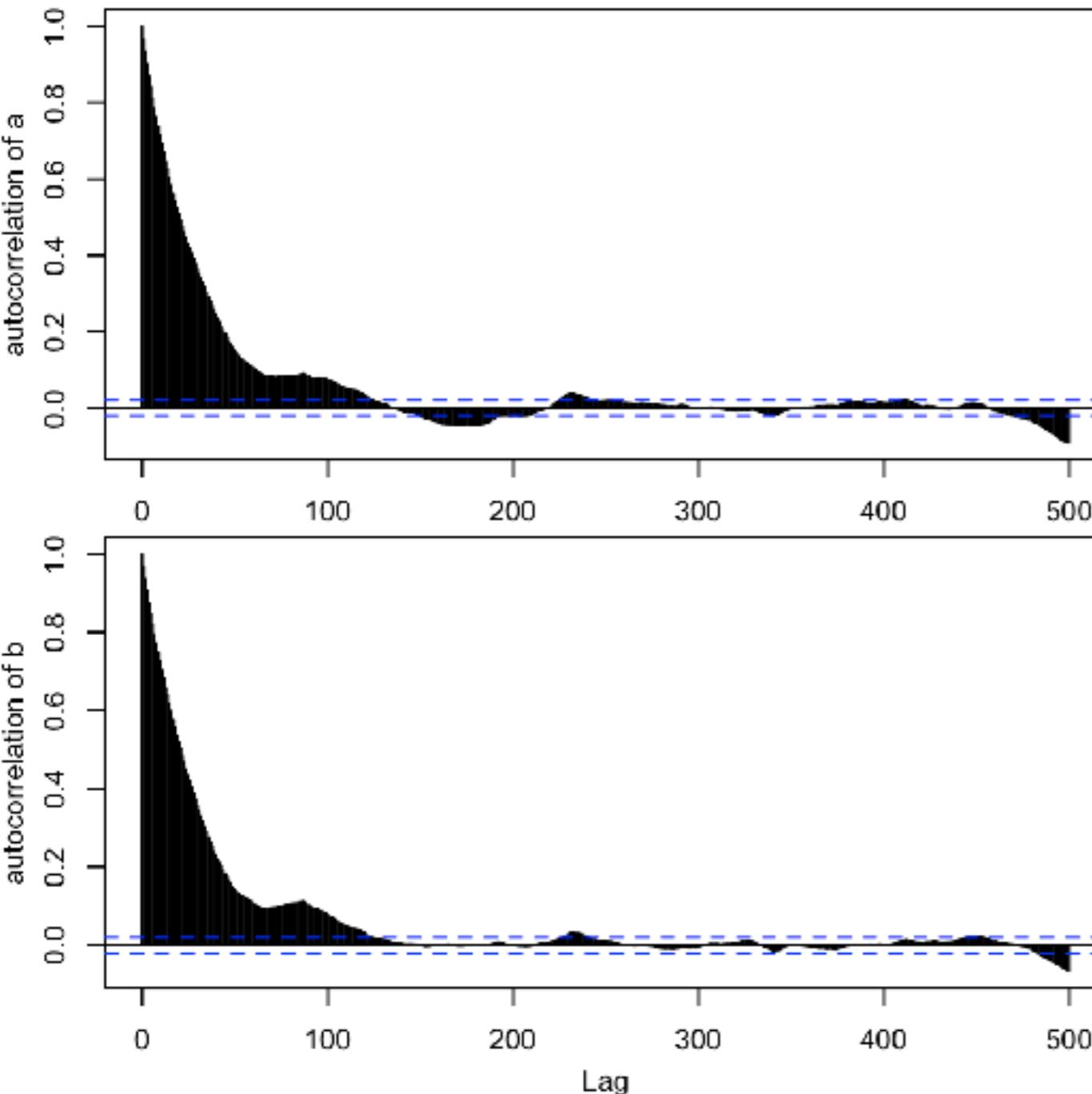
- ▶ The autocorrelation of a sequence (after removing burn-in), as a function of lag, k , is

$$\rho_k = \frac{\sum_{i=1}^{n-k} (\theta_i - \bar{\theta})(\theta_{i+k} - \bar{\theta})}{\sum_{i=1}^{n-k} (\theta_i - \bar{\theta})^2} = \frac{\text{Cov}_i(\theta_i, \theta_{i+k})}{\text{Var}(\theta)}$$

- ▶ The larger lag one needs to get a small autocorrelation, the less informative individual samples are.
- ▶ The pandas function `autocorrelation_plot()` may be useful for this.

CORRELATION TESTS

66



Note that the positive/negative oscillations basically tell us when the lag is so large compared with the chain length that the autocorrelation is too noisy to be meaningful.

We would be justified in **thinning** the chains by a factor of ~150, apparently!
(i.e. take every 150th sample)

EFFECTIVE NUMBER OF SAMPLES

67

- ▶ From m chains of length n , we can estimate the effective number of independent samples as: (though we never actually sum to infinity - it's cut off as $\hat{\rho}_t$ gets numerically unstable to calculate)

$$n_{eff} = \frac{mn}{1 + 2 \sum_0^{\infty} \hat{\rho}_t}$$

- ▶ with (V - no subscript t - is the same as the Gelman-Rubin calculation on slide 27) $\hat{\rho}_t = 1 - \frac{V_t}{2V}$

$$V_t = \frac{1}{m(n-t)} \sum_{j=0}^m \sum_{i=t+1}^n (\theta_{i,j} - \theta_{i-t,j})^2$$

- ▶ and

Noise

- **General types of noise:** white, thermal, shot, brownian, 1/f
- **General categories of noise:** Statistical v. Systematic; correlated v. uncorrelated
- **Types of noise in astrophysics:** instrumental noise, atmospheric noise, confusion noise, astrophysical noise
- **Measurements:** relative v. absolute
- **Catologs:** completeness and purity
- **Bias:** flux, Malmquist, Eddington

Photon Noise

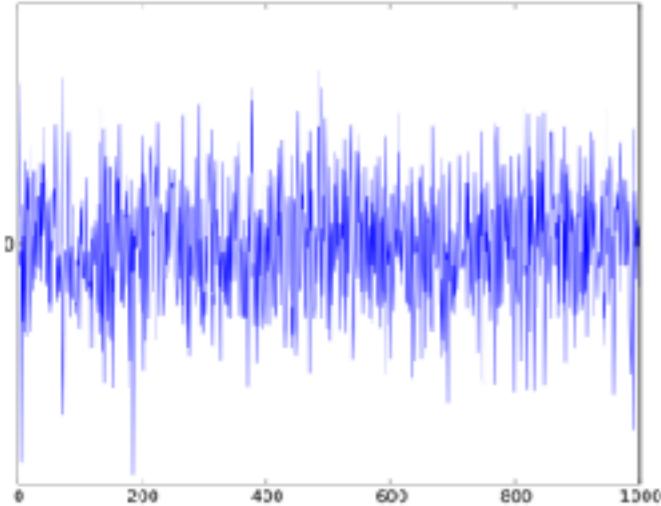


Photon Noise

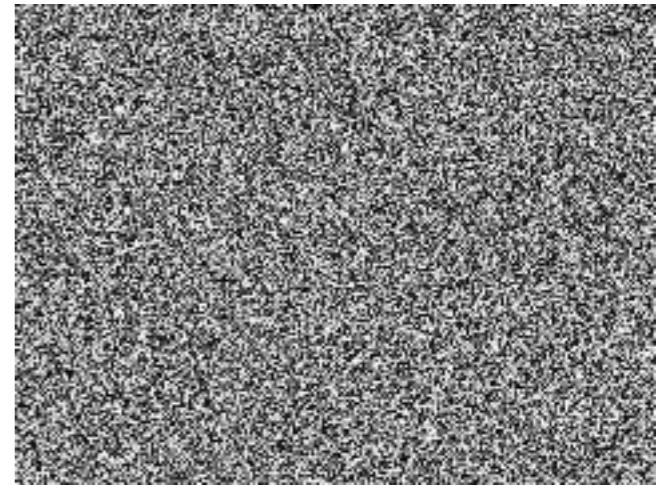


Photon Noise





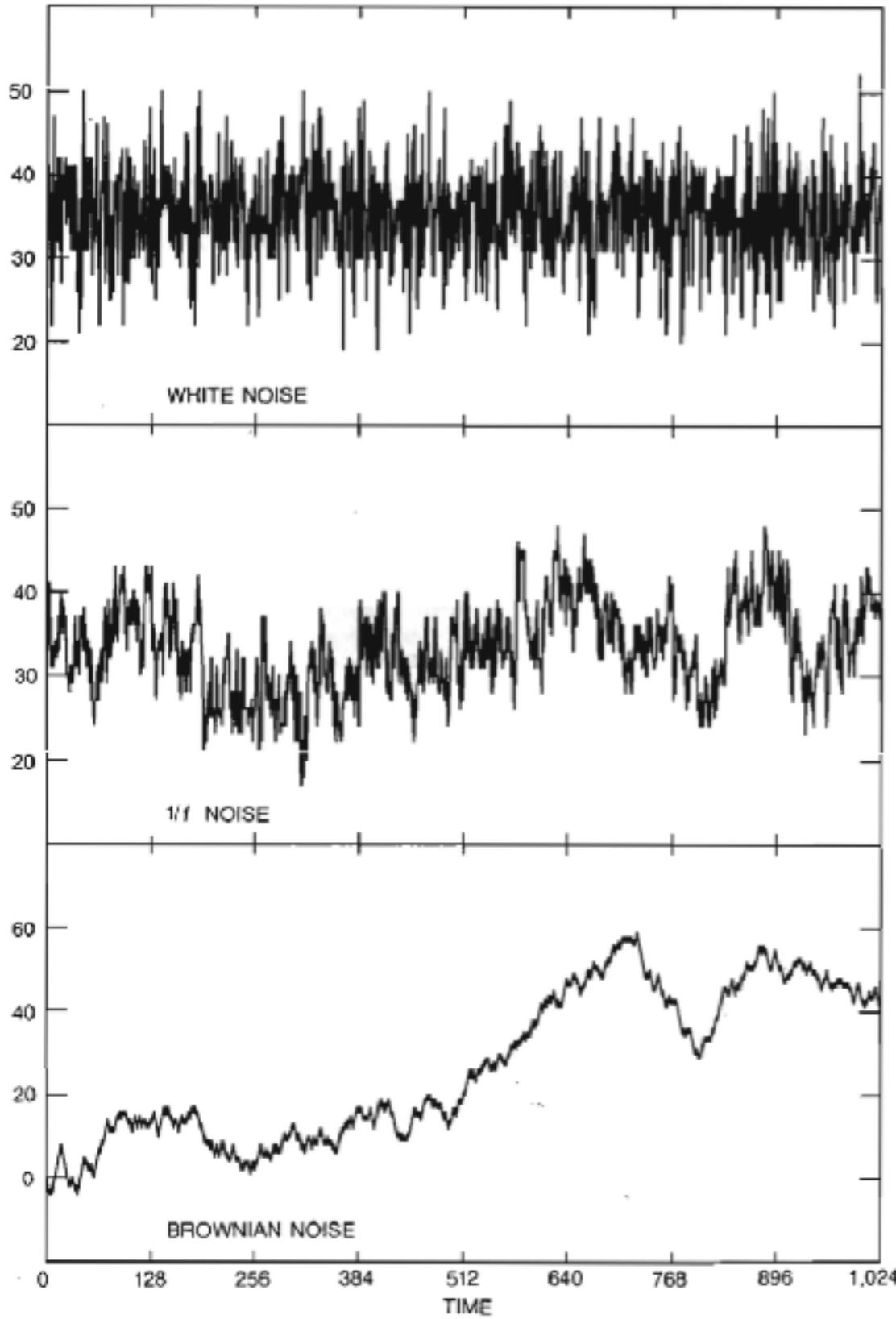
White Noise



- random signal with constant power spectral density
- each point is uncorrelated.
- Between 20 to 20,000 Hz, sounds like ssshhhhh
- mean is zero, but power is proportional to the RMS.
- Can be characterized by a constant variance or power.
- When you use a random number generator, you are simulating white noise.
- Thermal (Johnson) noise is Gaussian
 - Voltage noise $V_{\text{noise}} = \sqrt{4k_B T R}$
 - Voltage noise $V_{\text{noise}} = \sqrt{4k_B T R \Delta f}$ for RMS within a given bandwidth in nV/ $\sqrt{\text{Hz}}$
 - power noise $P = V^2/R = 4k_B T \Delta f$
- Shot noise is Poisson

Noise

$$N(f) \propto \frac{1}{f^n}$$



White or Poisson

$$N_{\text{white}}(f) \propto \frac{1}{f^0} = 1$$

stochastic

1/f or pink or flicker

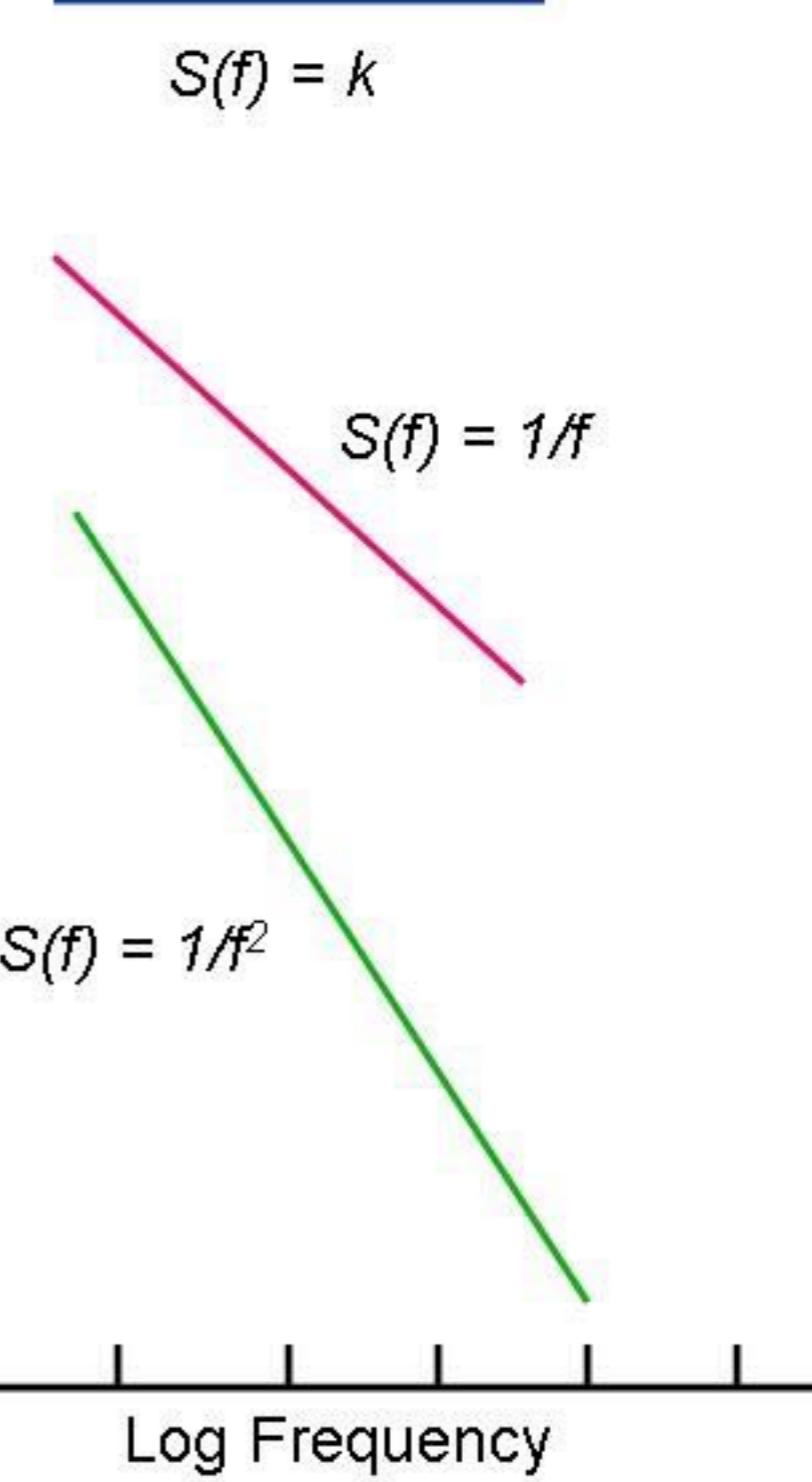
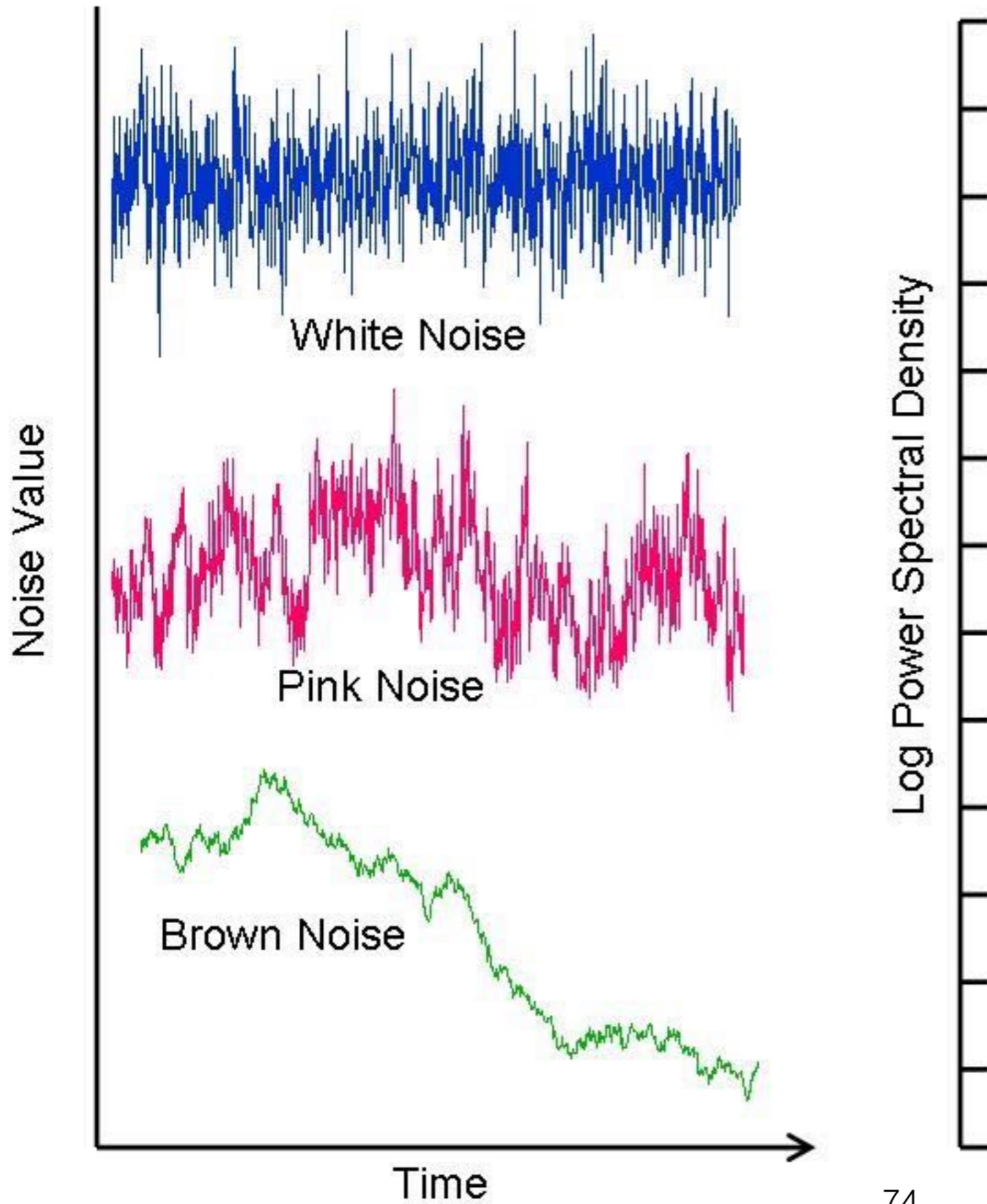
$$N_{1/f}(f) \propto \frac{1}{f^1} = \frac{1}{f}$$

occurs in nature

Brownian

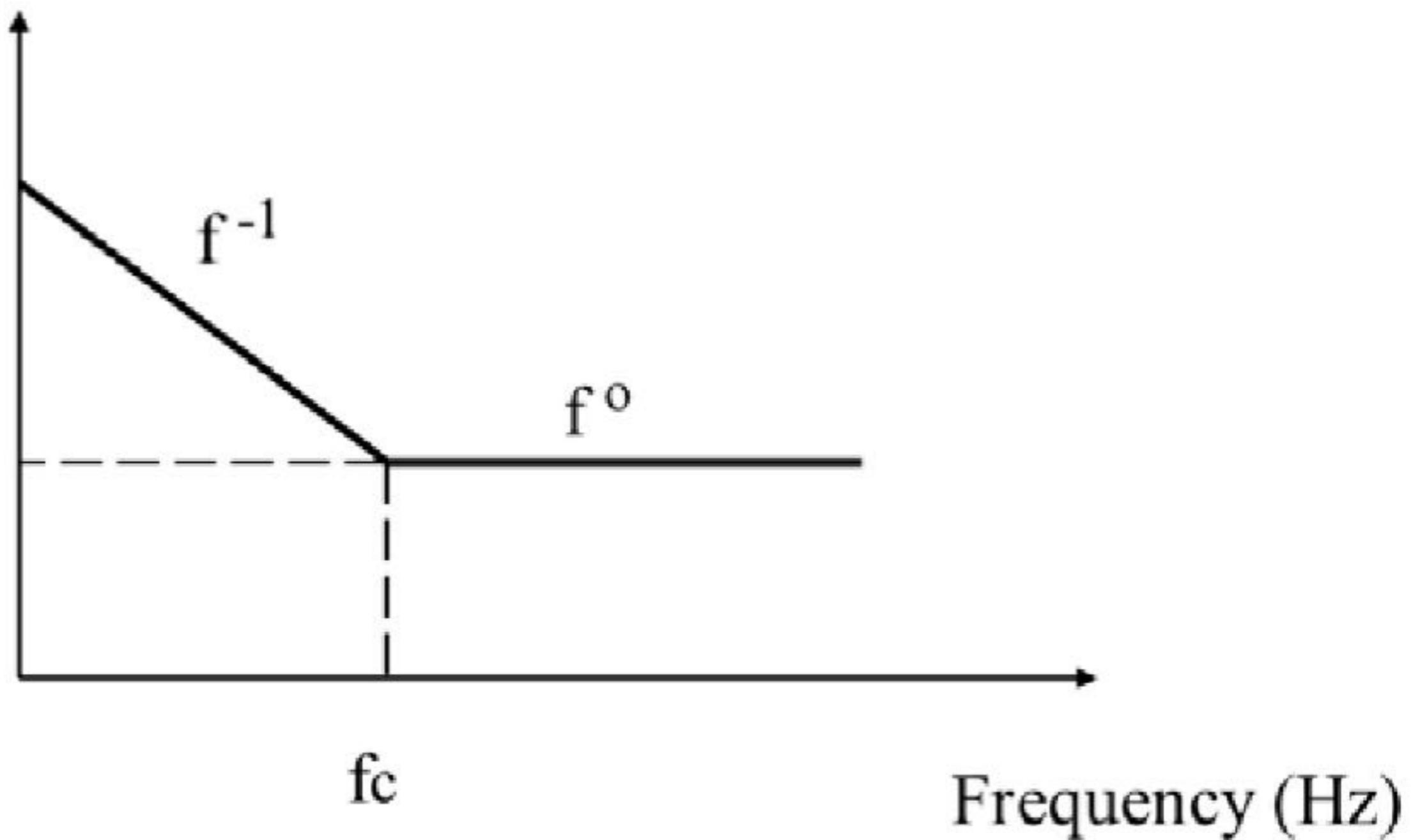
$$N_{\text{Brownian}}(f) \propto \frac{1}{f^2}$$

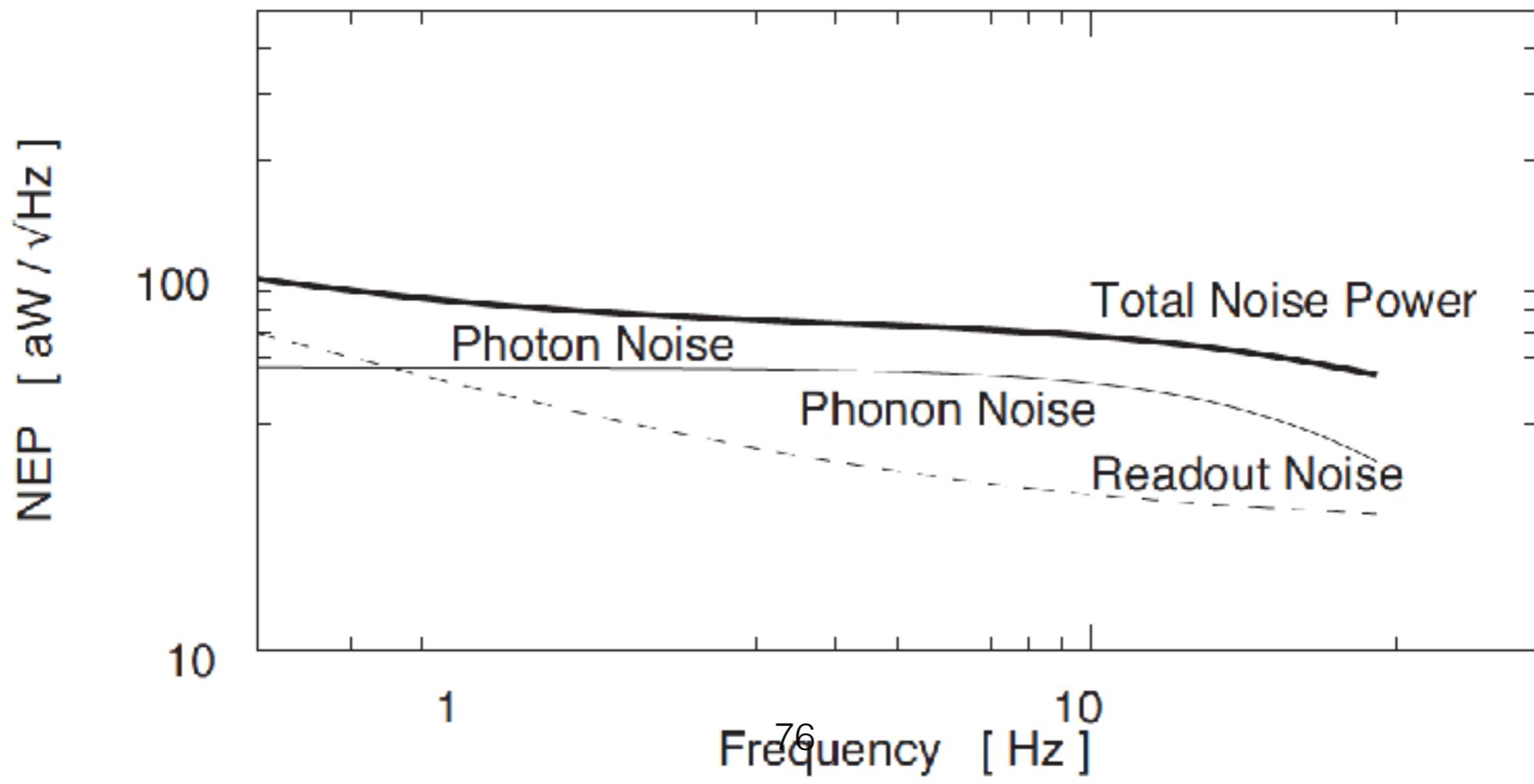
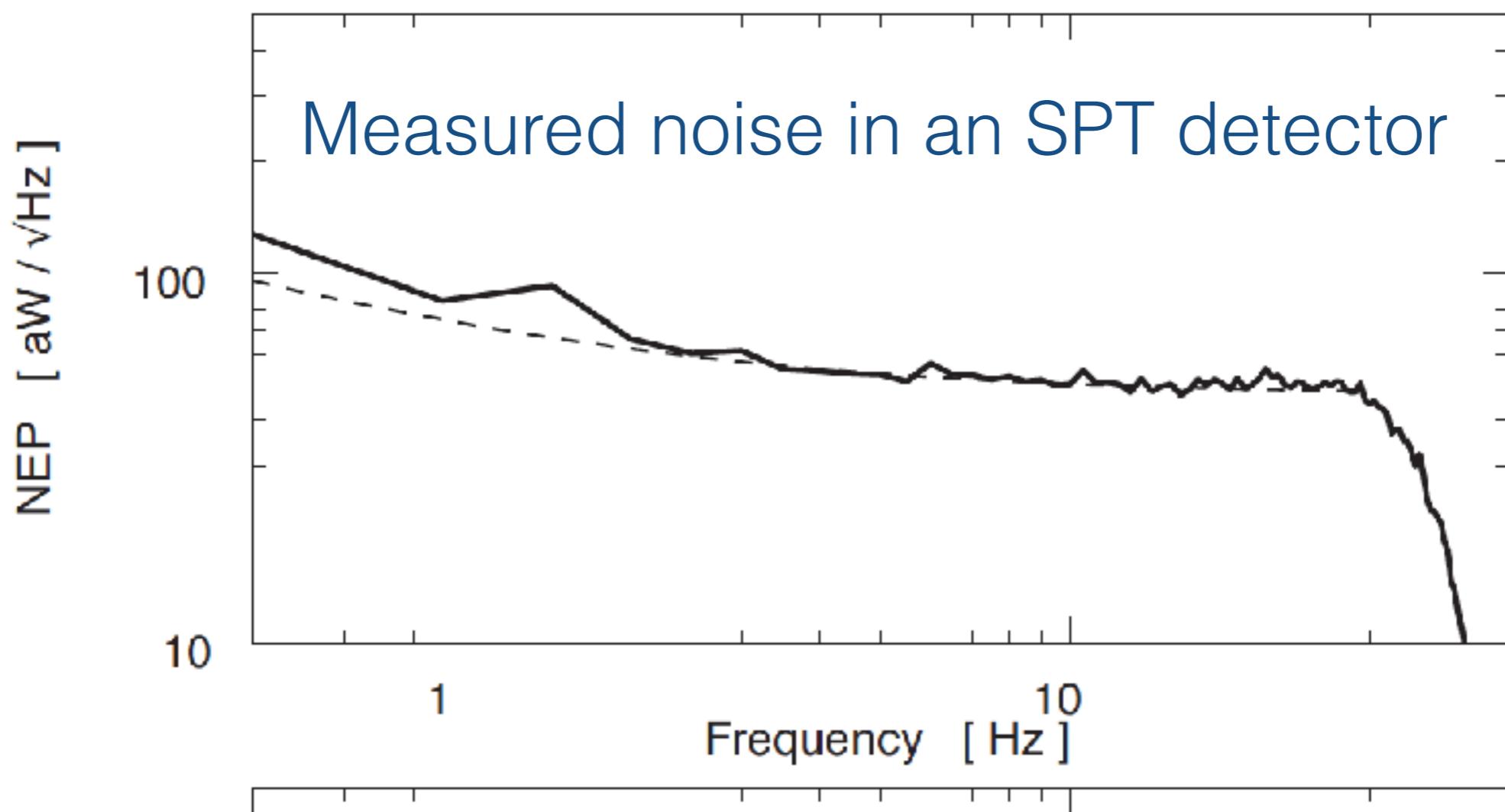
Cumulative integral of white noise



1/f noise

Noise Power





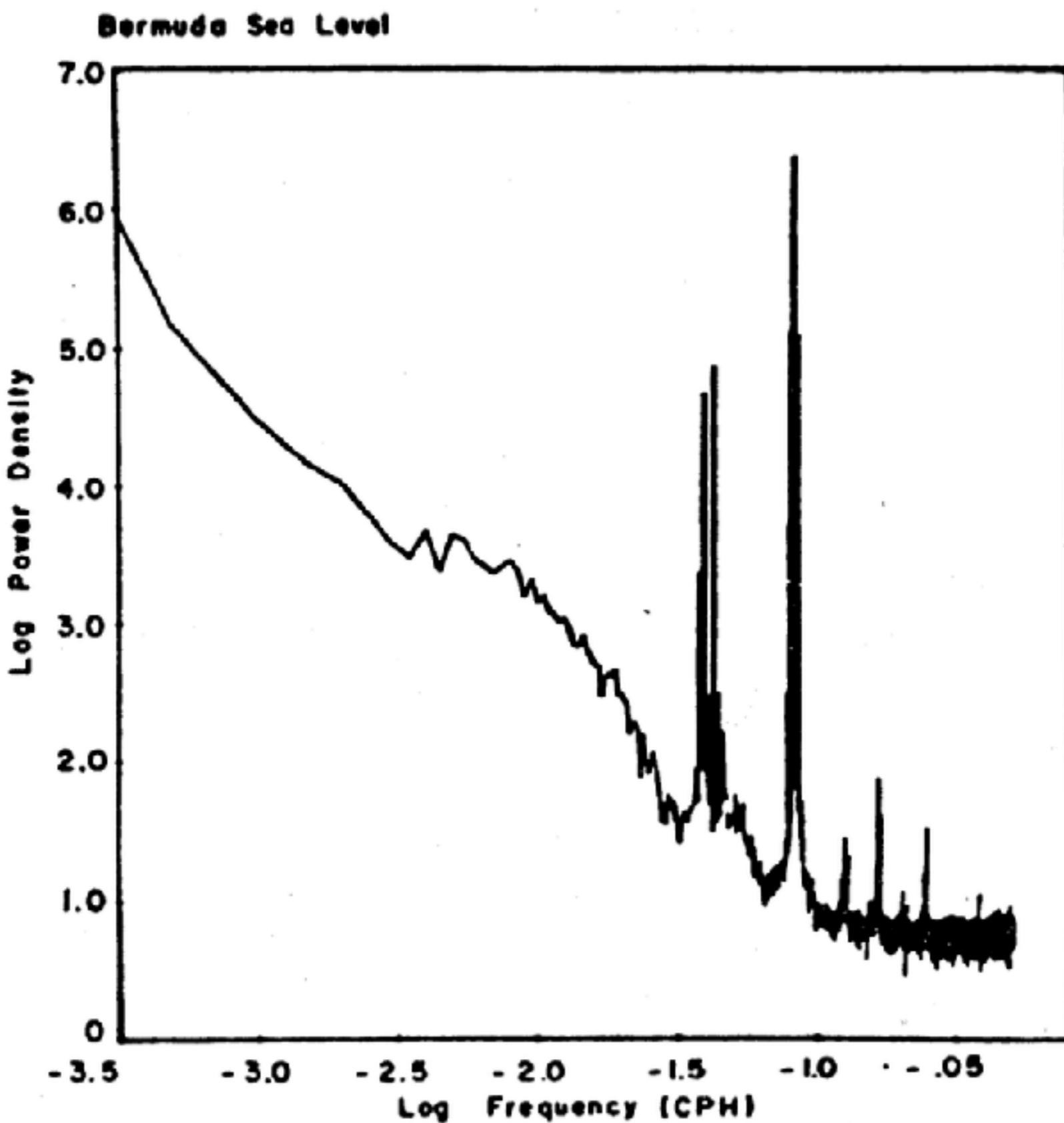


FIGURE 8 Power spectrum of sea level at Bermuda. Note the strong tidal components at 12 hours and 24 hours. The rise of power at low frequencies is here steeper than $1/f$ (and is close to $1/f^2$). (Reproduced from Wunsch.²⁵)

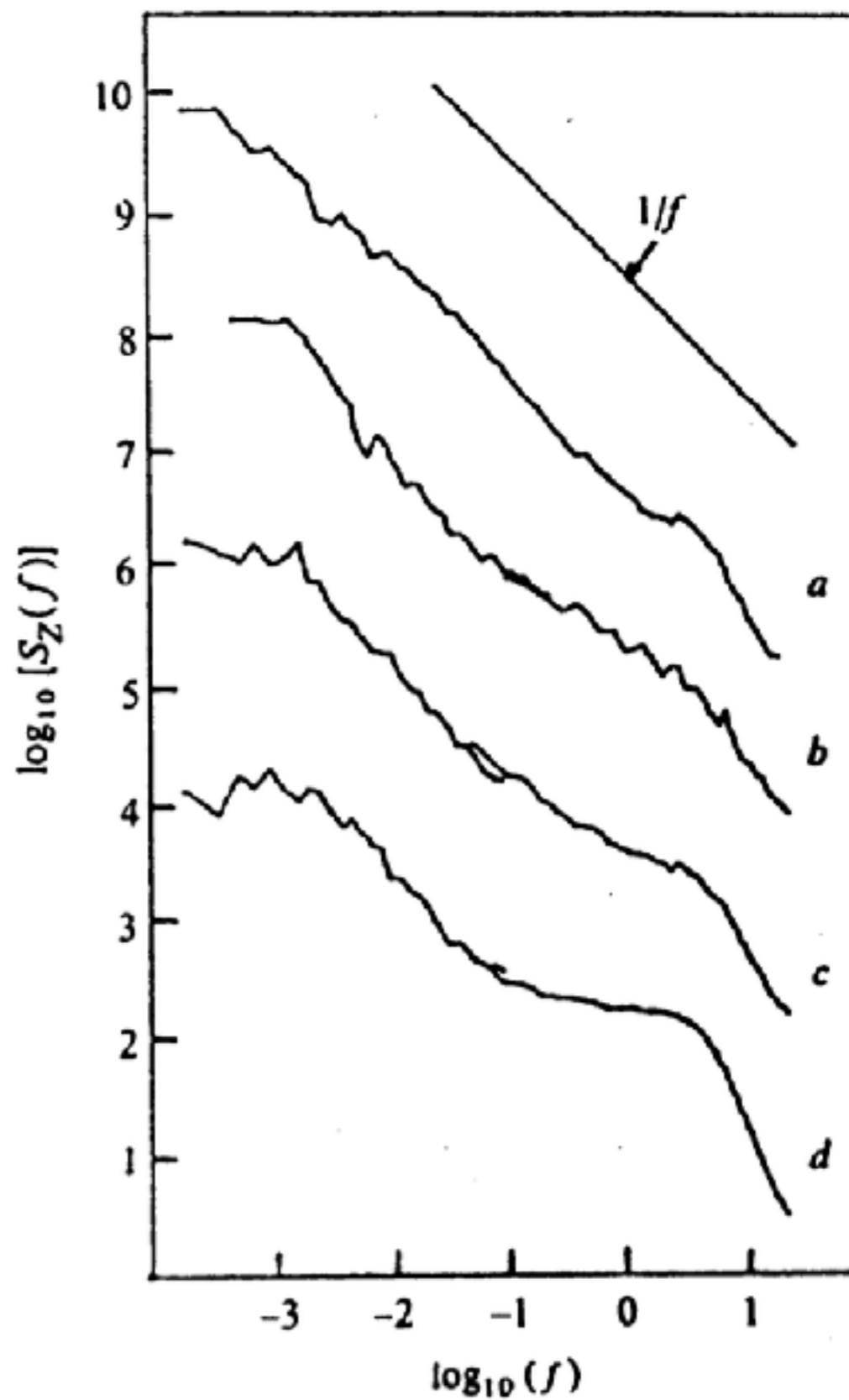
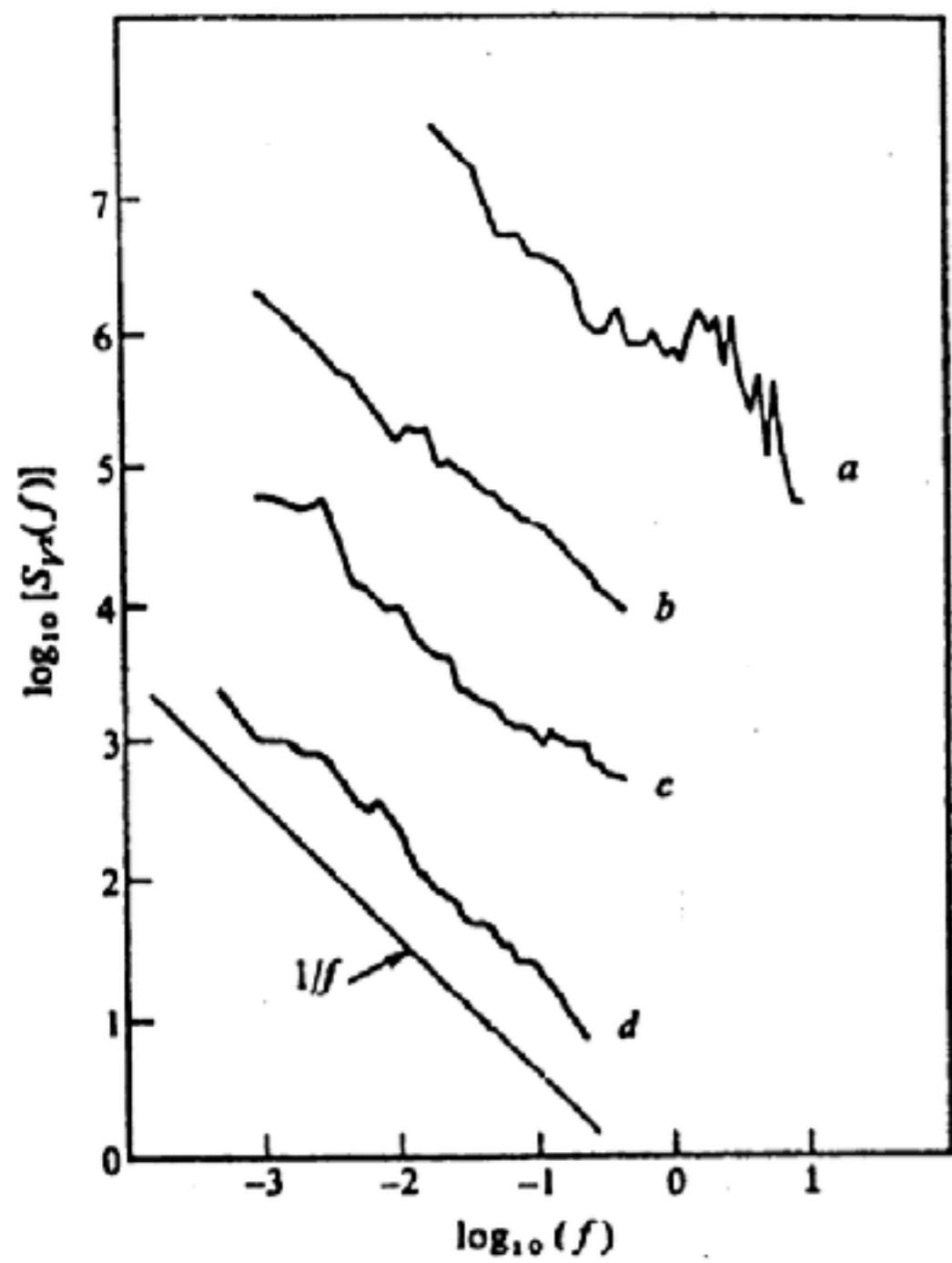
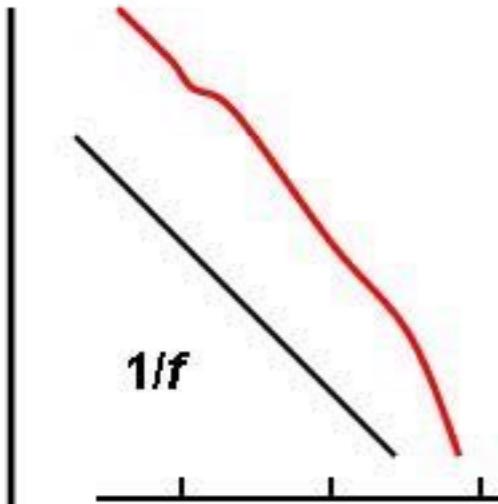
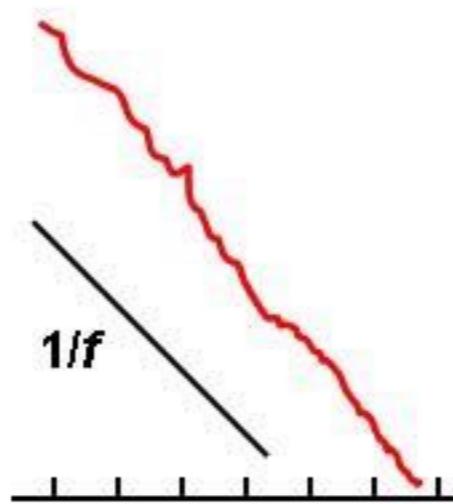


FIGURE 9 (left) Loudness fluctuation spectra as a function of frequency f (in Hz) for: *a*, Scott Joplin Piano Rags; *b*, classical radio station; *c*, rock station; *d*, news-and-talk station. (Reproduced from Voss and Clarke.²³) (right) Power spectra of pitch fluctuations for four radio stations: *a*, classical; *b*, jazz and blues; *c*, rock; *d*, news-and-talk. (Reproduced from Voss and Clarke (1975).)

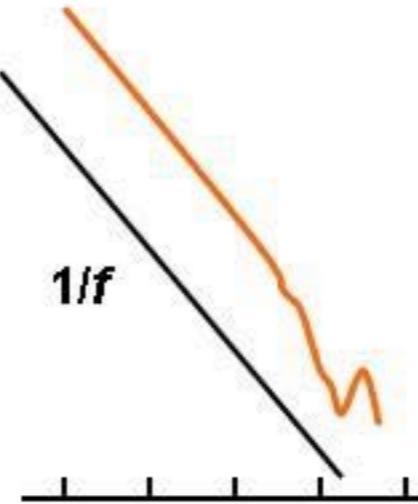
Log Spectral Power Density



A. Vacuum tube
(Johnson, 1925)



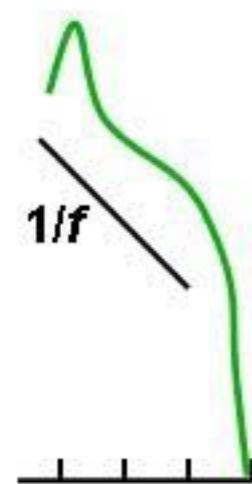
B. Semiconductor
(Caloyannides, 1974)



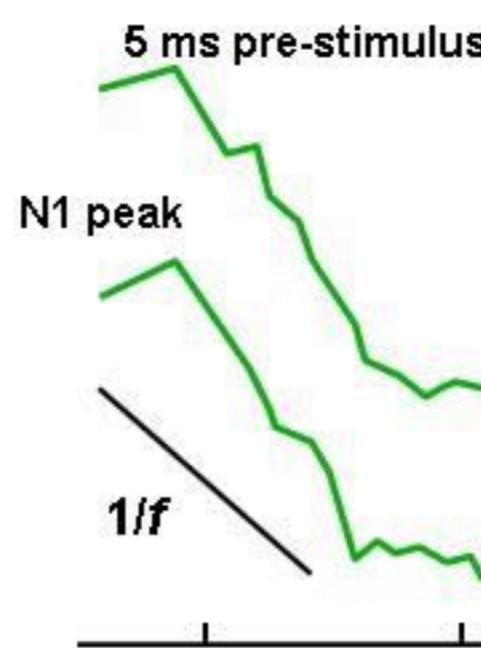
C. Human heart
(Musha, 1981)



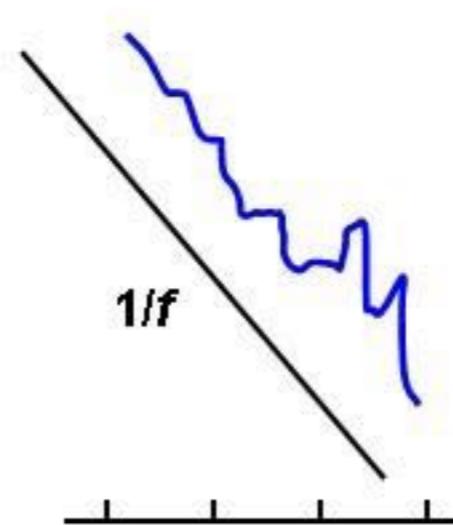
D. Squid giant axon
(Musha, 1981)



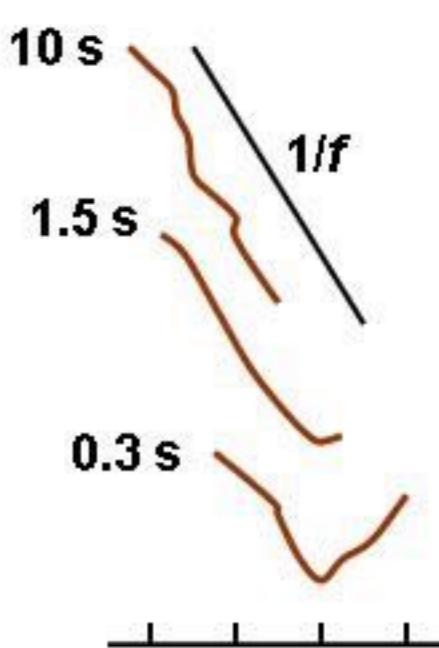
E. Brain MEG
(Novikov, et
al., 1997)



F. Brain EEG
(McDonald, Ward,
1998)

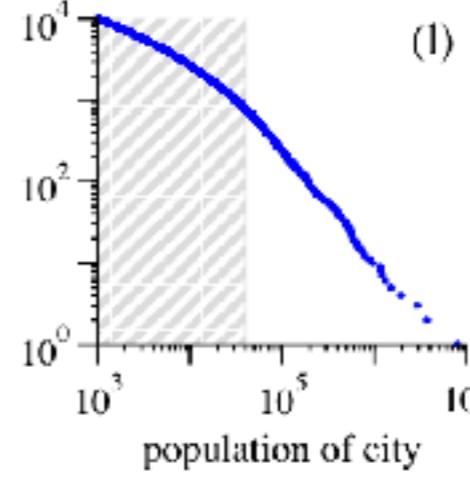
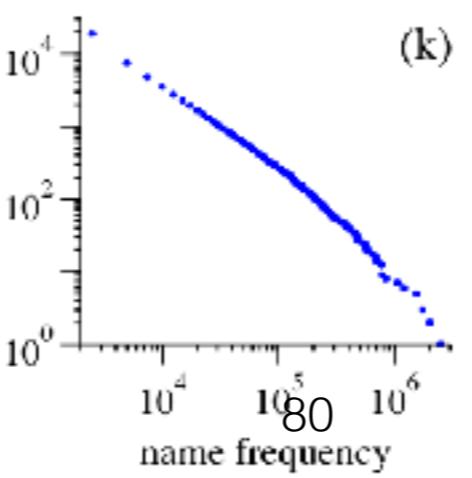
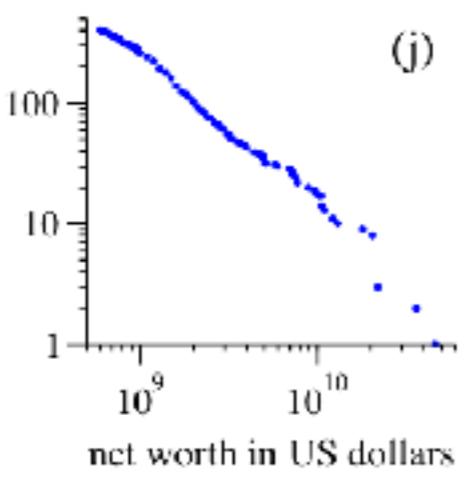
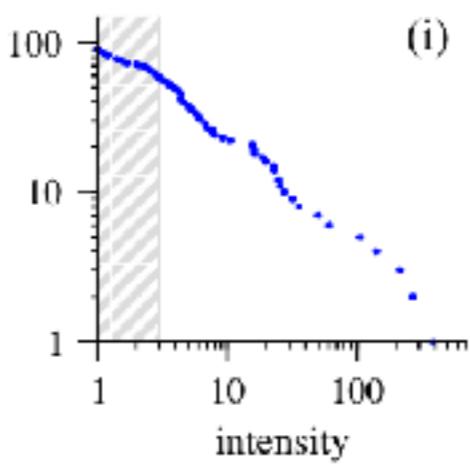
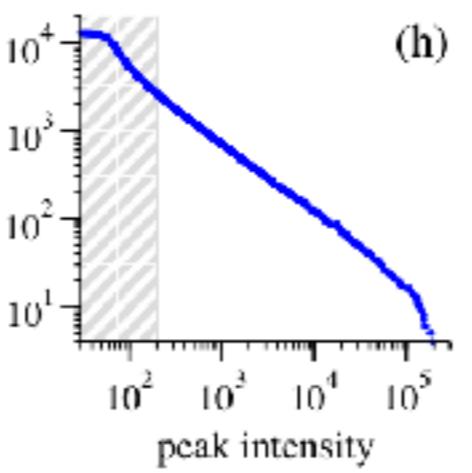
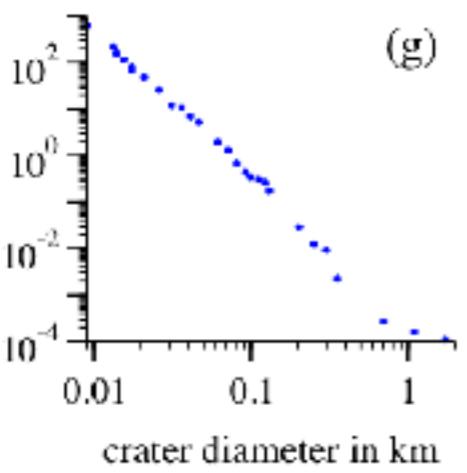
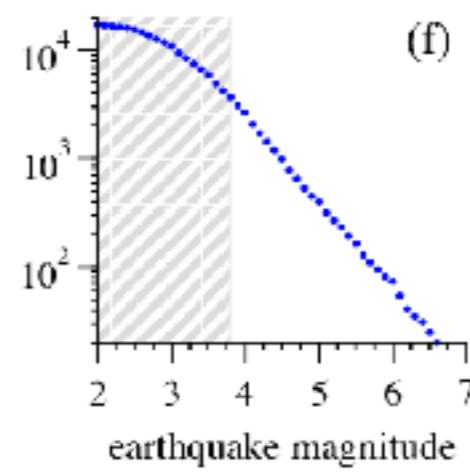
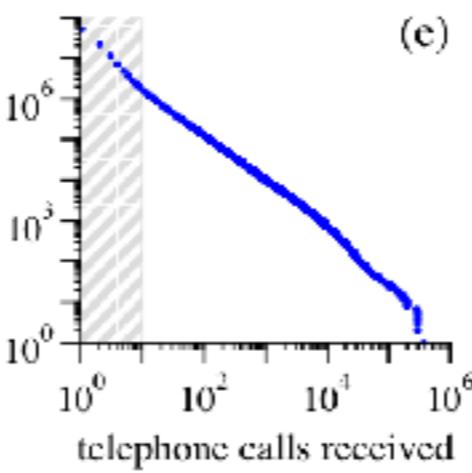
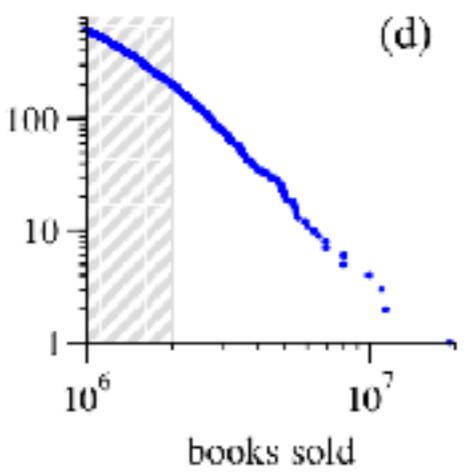
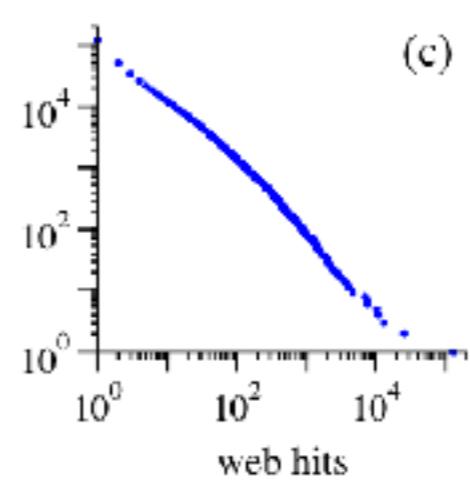
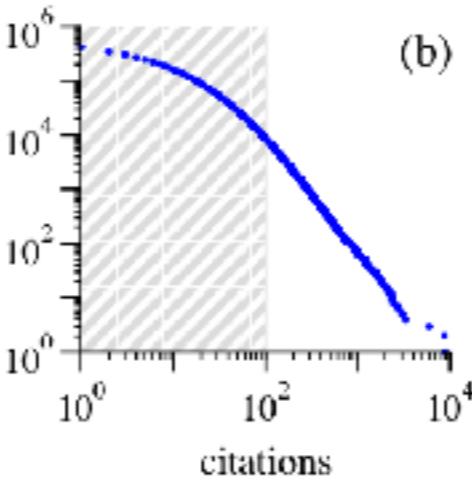
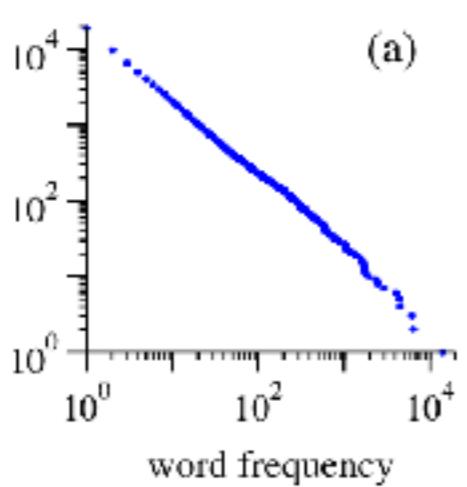


G. Brandenberg
Concerto # 1 (Voss,
Clarke, 1975)



H. Human time
estimation (Gilden
et al, 1995)

Log Frequency



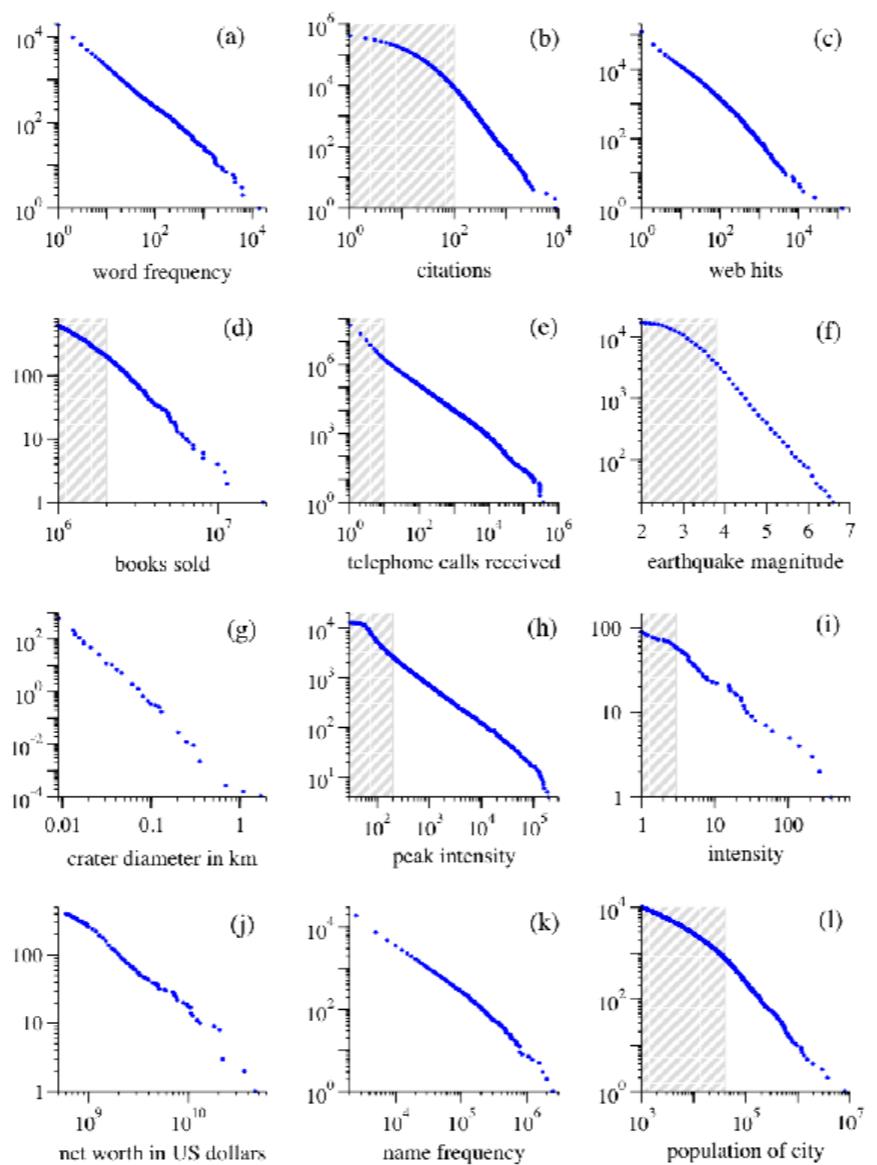
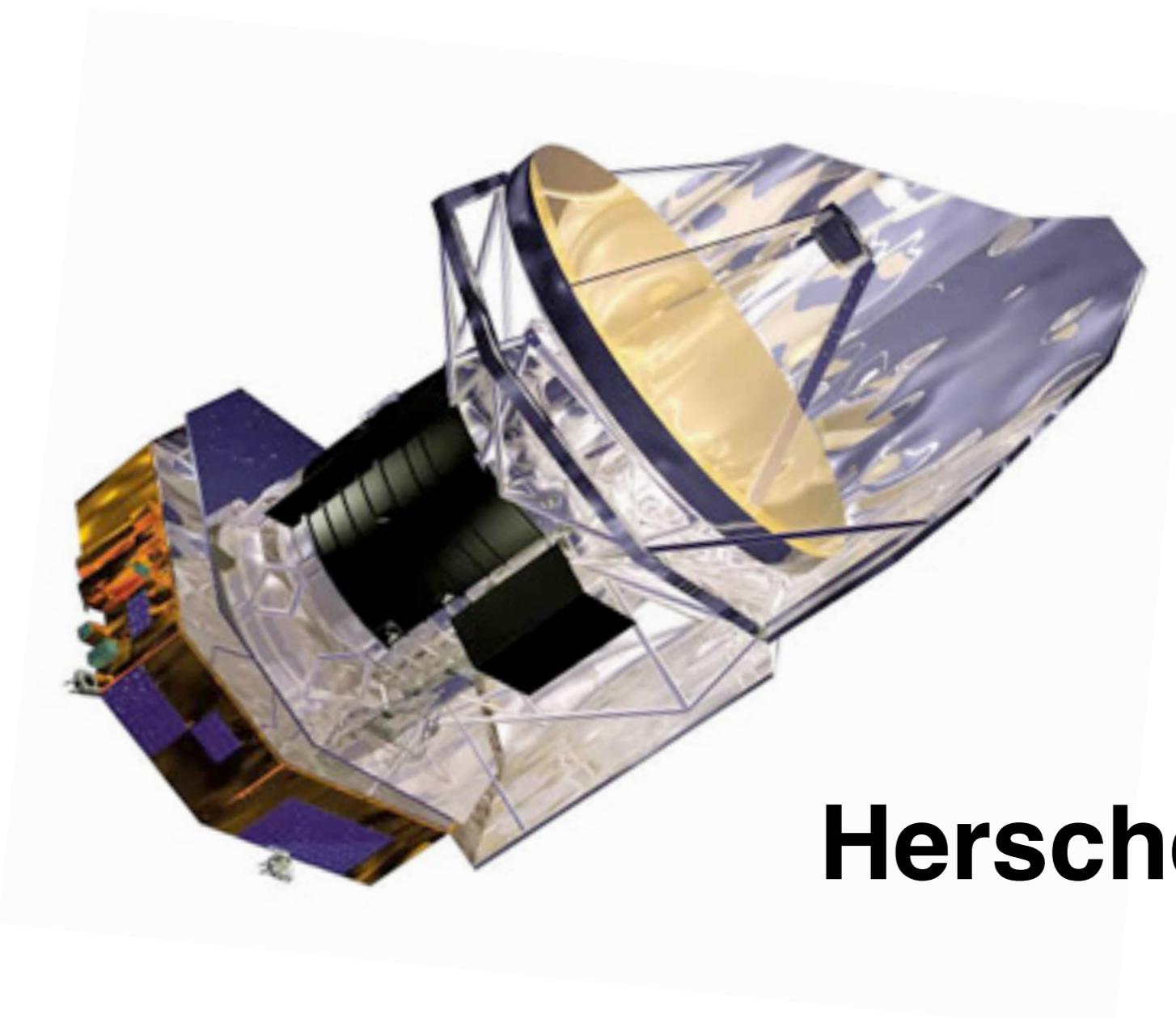
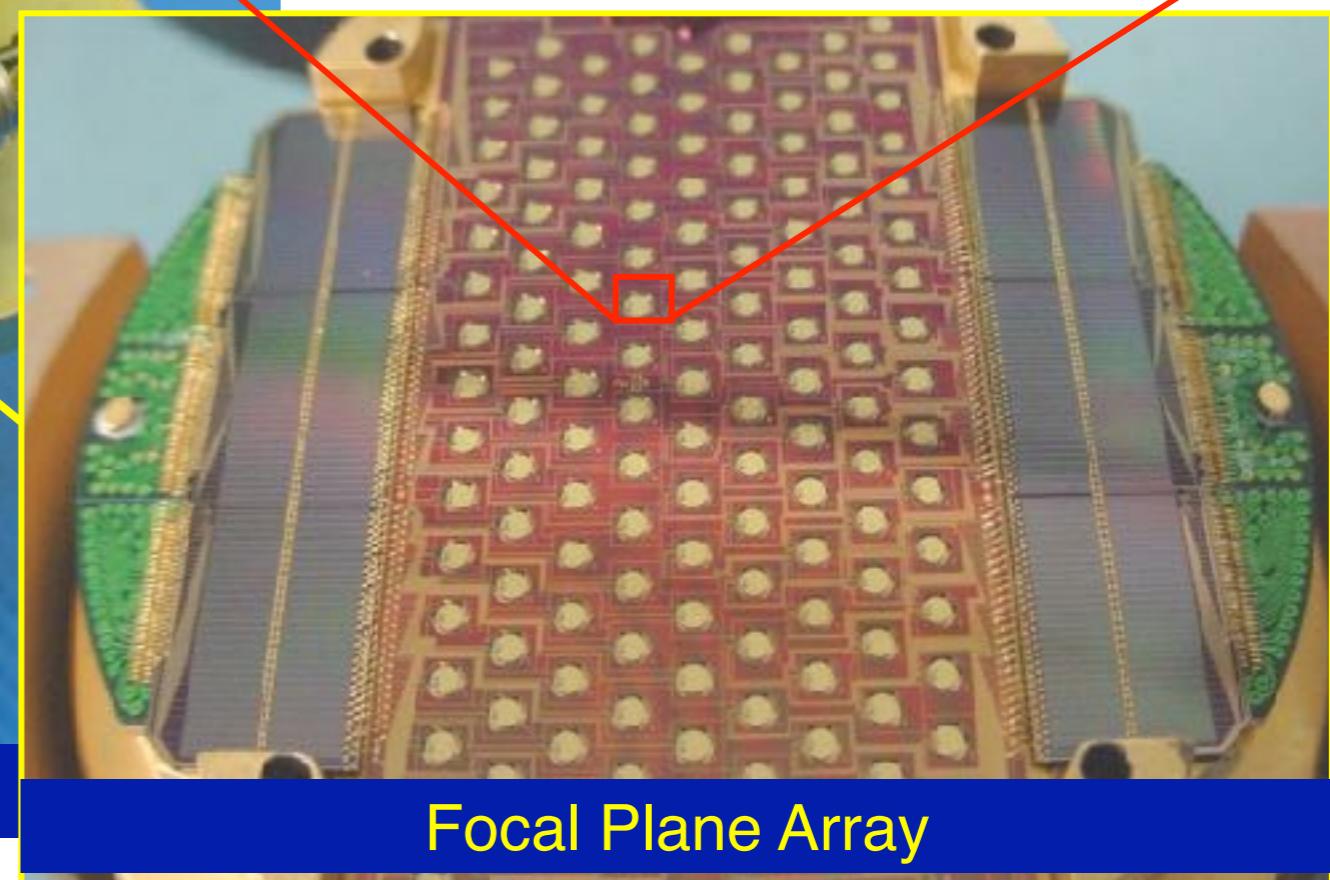
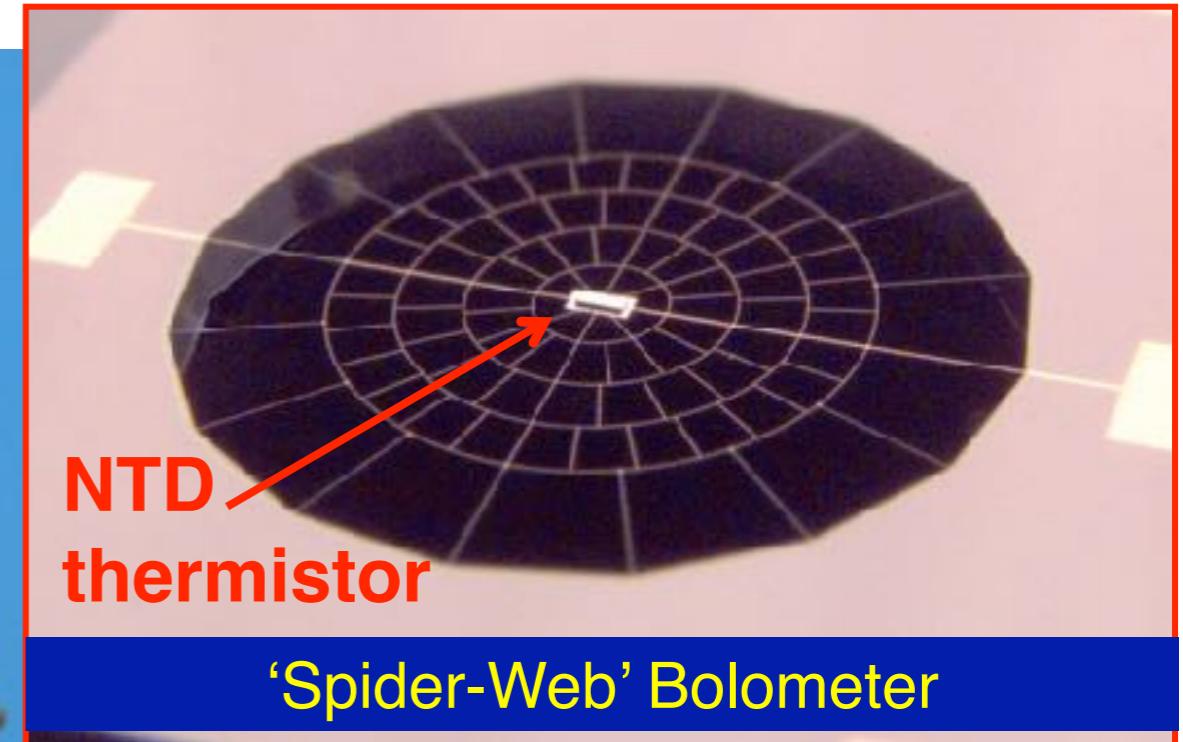
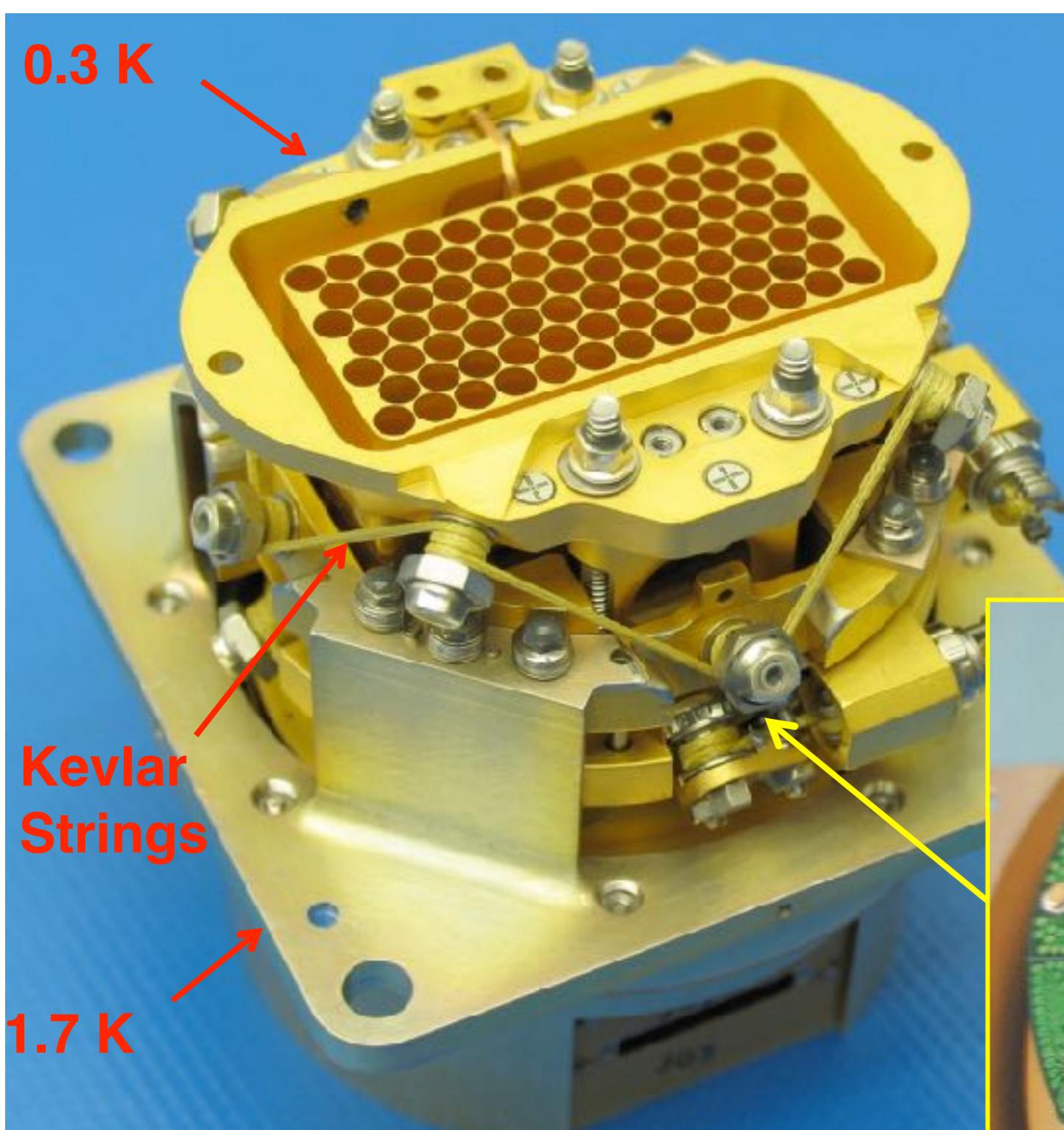


FIG. 4 Cumulative distributions or “rank/frequency plots” of twelve quantities reputed to follow power laws. The distributions were computed as described in Appendix A. Data in the shaded regions were excluded from the calculations of the exponents in Table I. Source references for the data are given in the text. (a) Numbers of occurrences of words in the novel *Moby Dick* by Hermann Melville. (b) Numbers of citations to scientific papers published in 1981, from time of publication until June 1997. (c) Numbers of hits on web sites by 60 000 users of the America Online Internet service for the day of 1 December 1997. (d) Numbers of copies of bestselling books sold in the US between 1895 and 1965. (e) Number of calls received by AT&T telephone customers in the US for a single day. (f) Magnitude of earthquakes in California between January 1910 and May 1992. Magnitude is proportional to the logarithm of the maximum amplitude of the earthquake, and hence the distribution obeys a power law even though the horizontal axis is linear. (g) Diameter of craters on the moon. Vertical axis is measured per square kilometre. (h) Peak gamma-ray intensity of solar flares in counts per second, measured from Earth orbit between February 1980 and November 1989. (i) Intensity of wars from 1816 to 1980, measured as battle deaths per 10 000 of the population of the participating countries. (j) Aggregate net worth in dollars of the richest individuals in the US in October 2003. (k) Frequency of occurrence of family names in the US in the year 1990. (l) Populations of US cities in the year 2000.

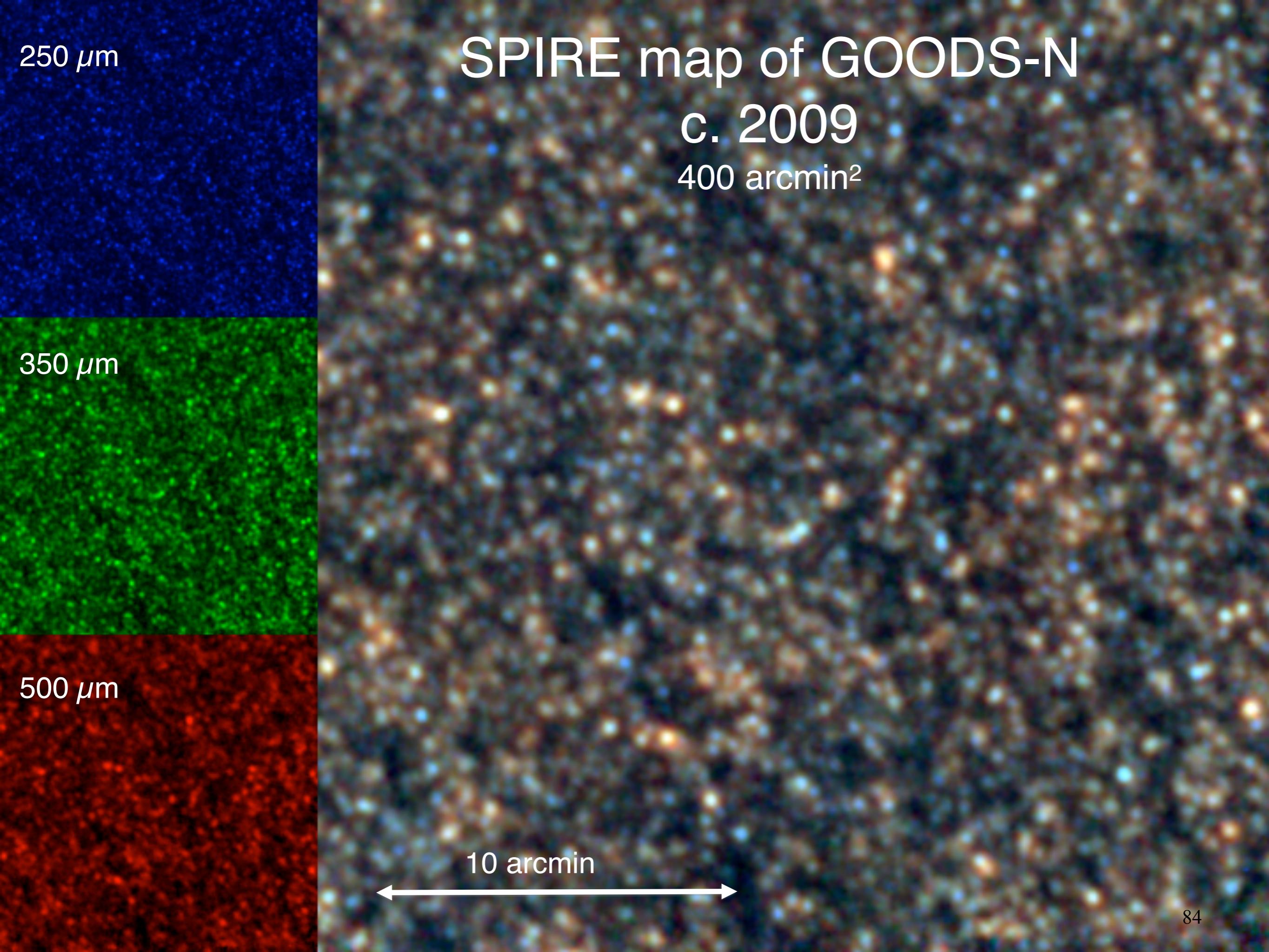
Example: Far Infrared



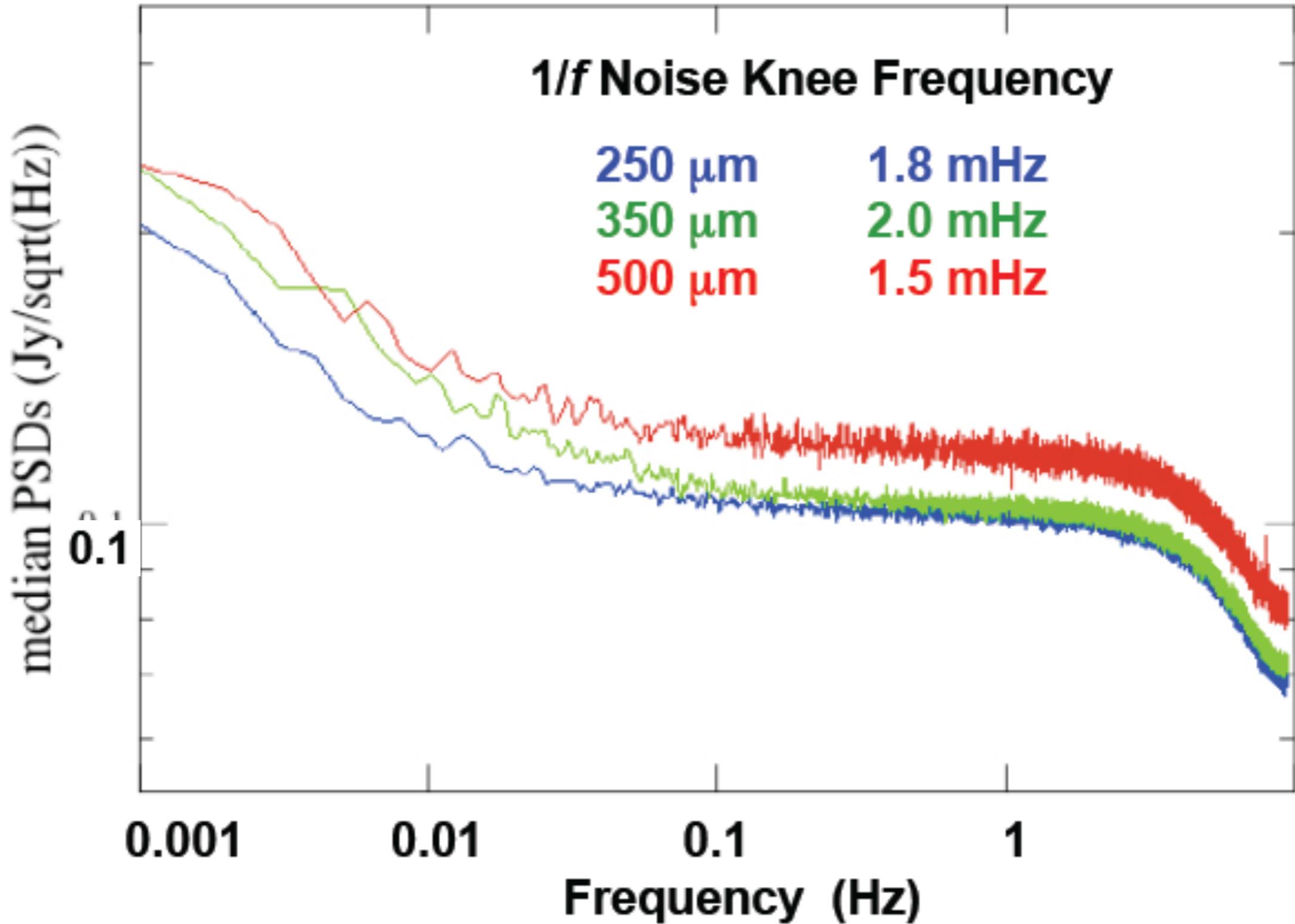
SPIRE Instrument: NASA/JPL



These detectors are background-limited !

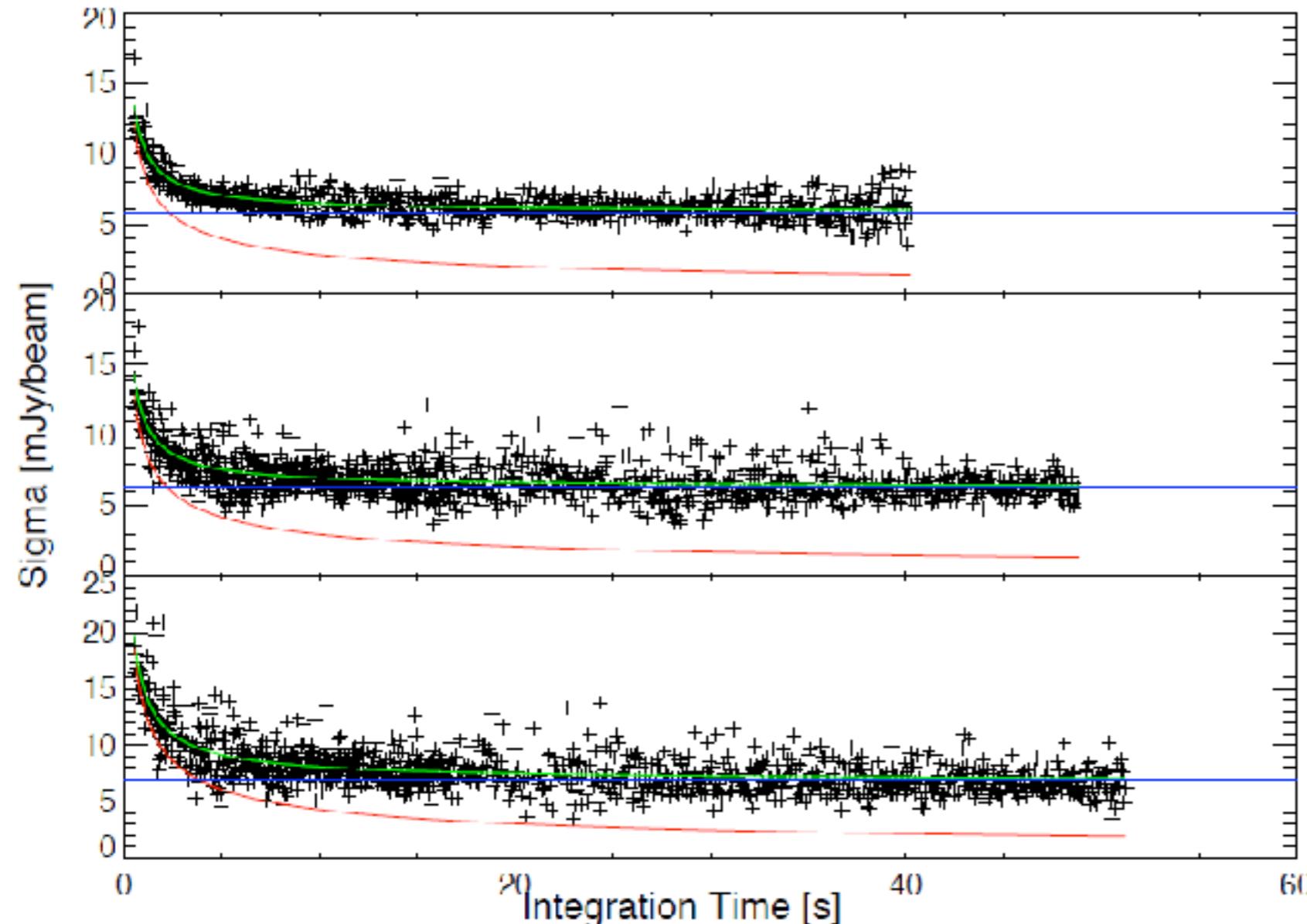


Scan Mapping with SPIRE



*SPIRE uses scan map to cover large areas
Low 1/f noise = high fidelity on all angular scales*

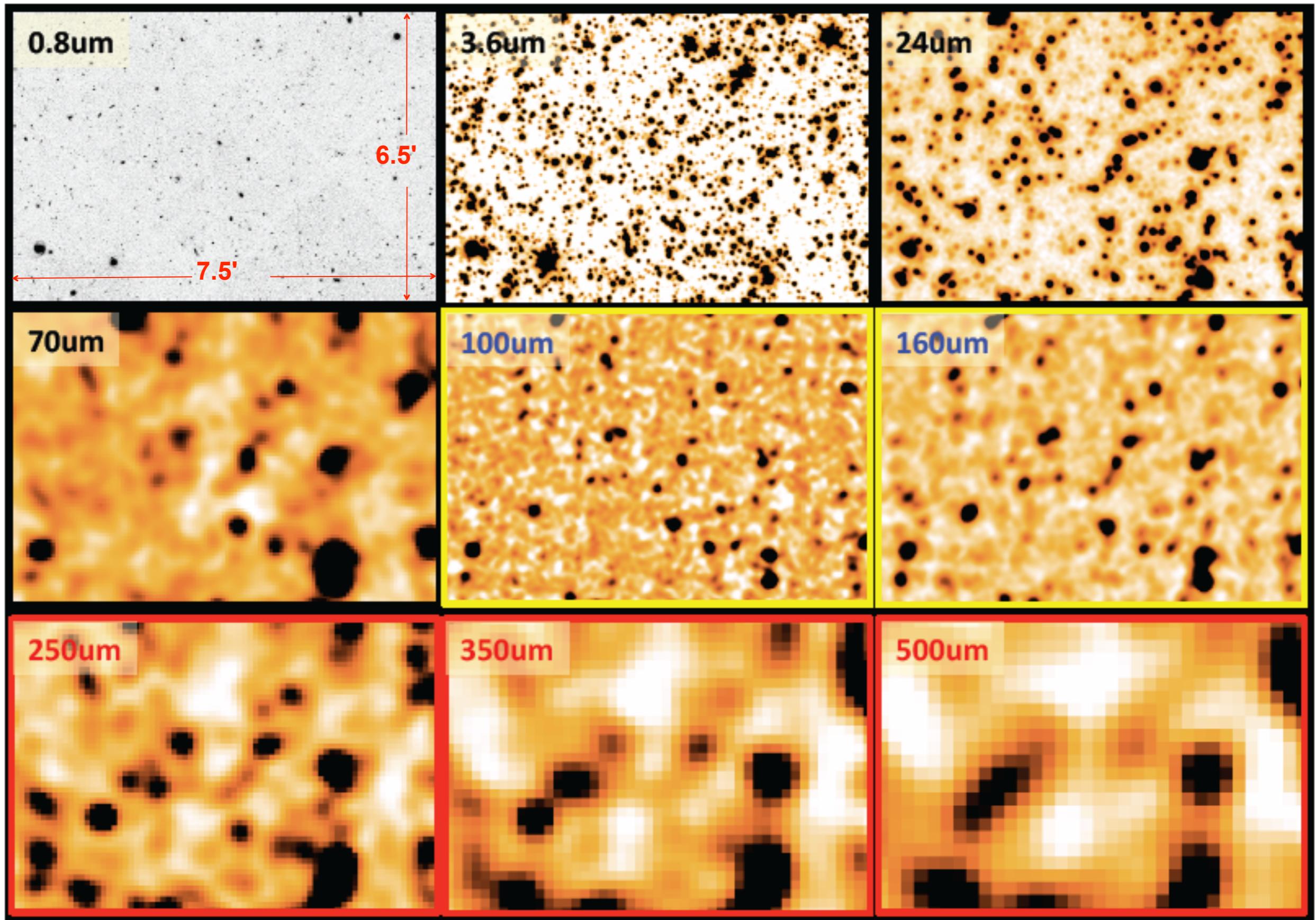
Mapping to the Confusion Limit



| Band | Conf. ($1\sigma^*$) | T/sq. deg |
|--------|-----------------------|-----------|
| 250 um | 4.8 | 2.8 h |
| 350 um | 5.5 | 1.4 h |
| 500 um | 6.1 | 2.4 h |

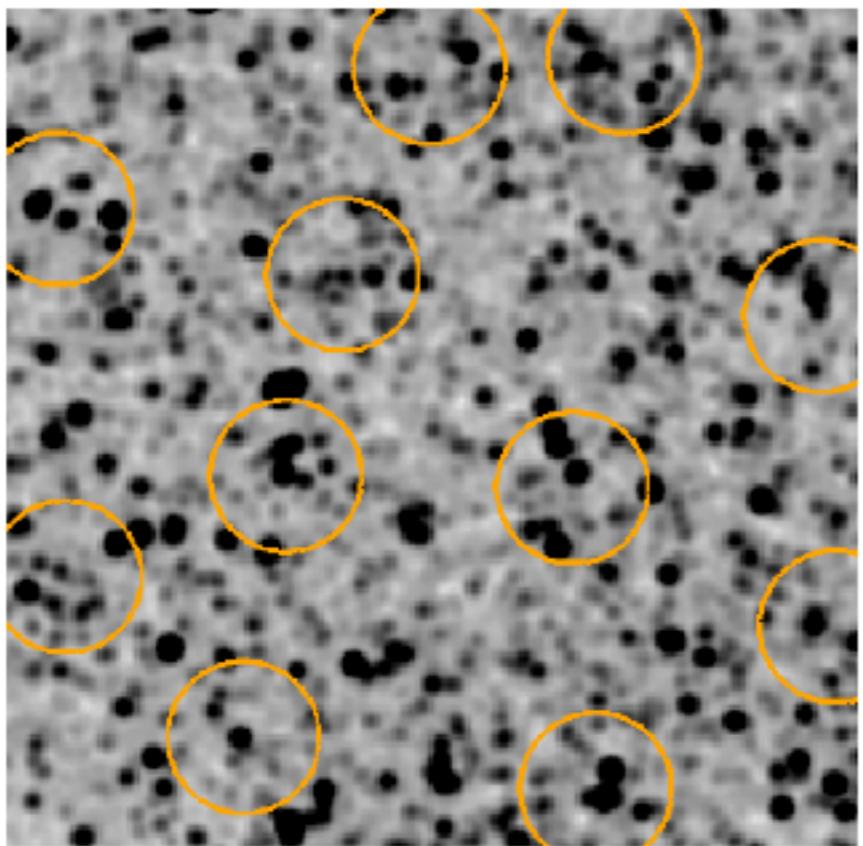
*after an iterative 5σ clip

Large Beam + High Sensitivity = Confusion

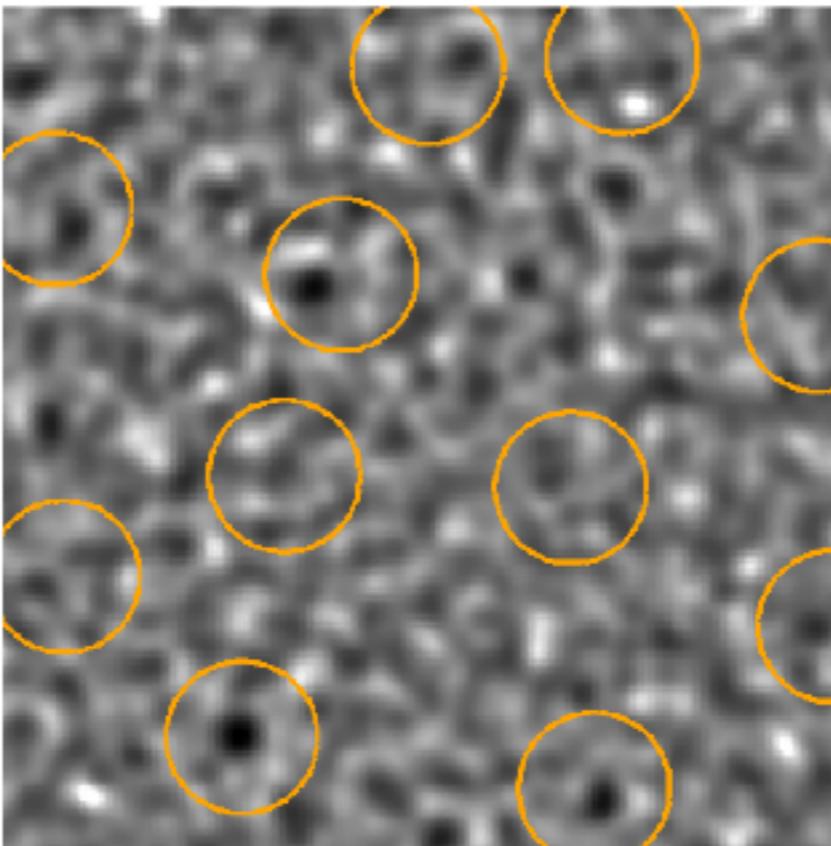


Confusion noise

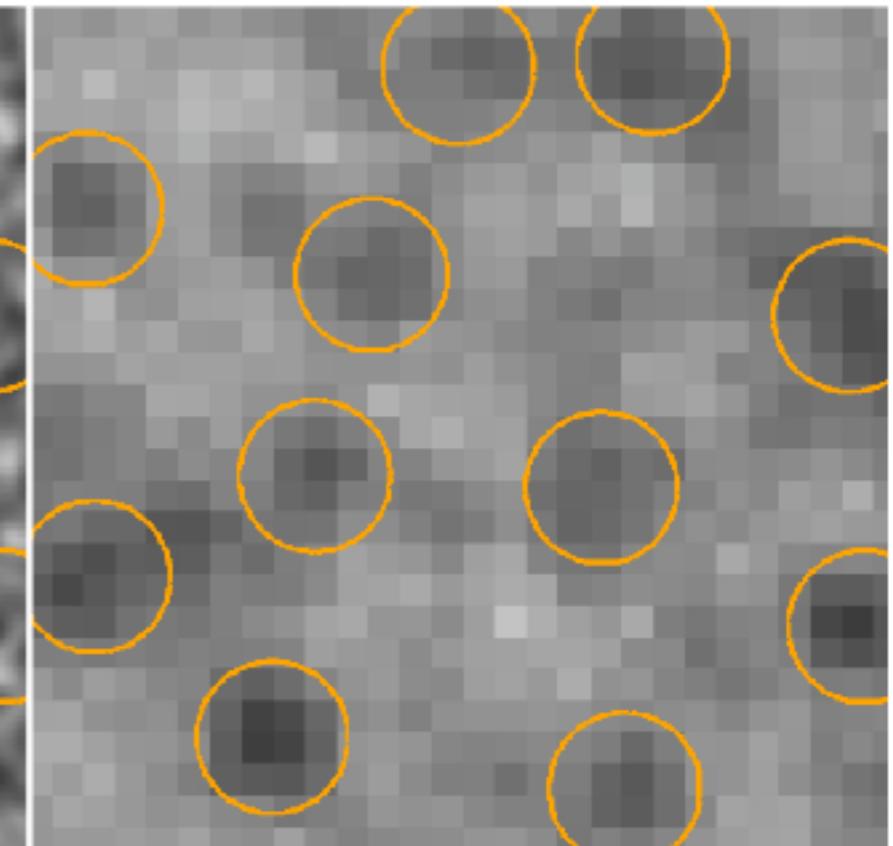
CCAT 350 μ m



JCMT 450 μ m



Herschel 350 μ m



25 m

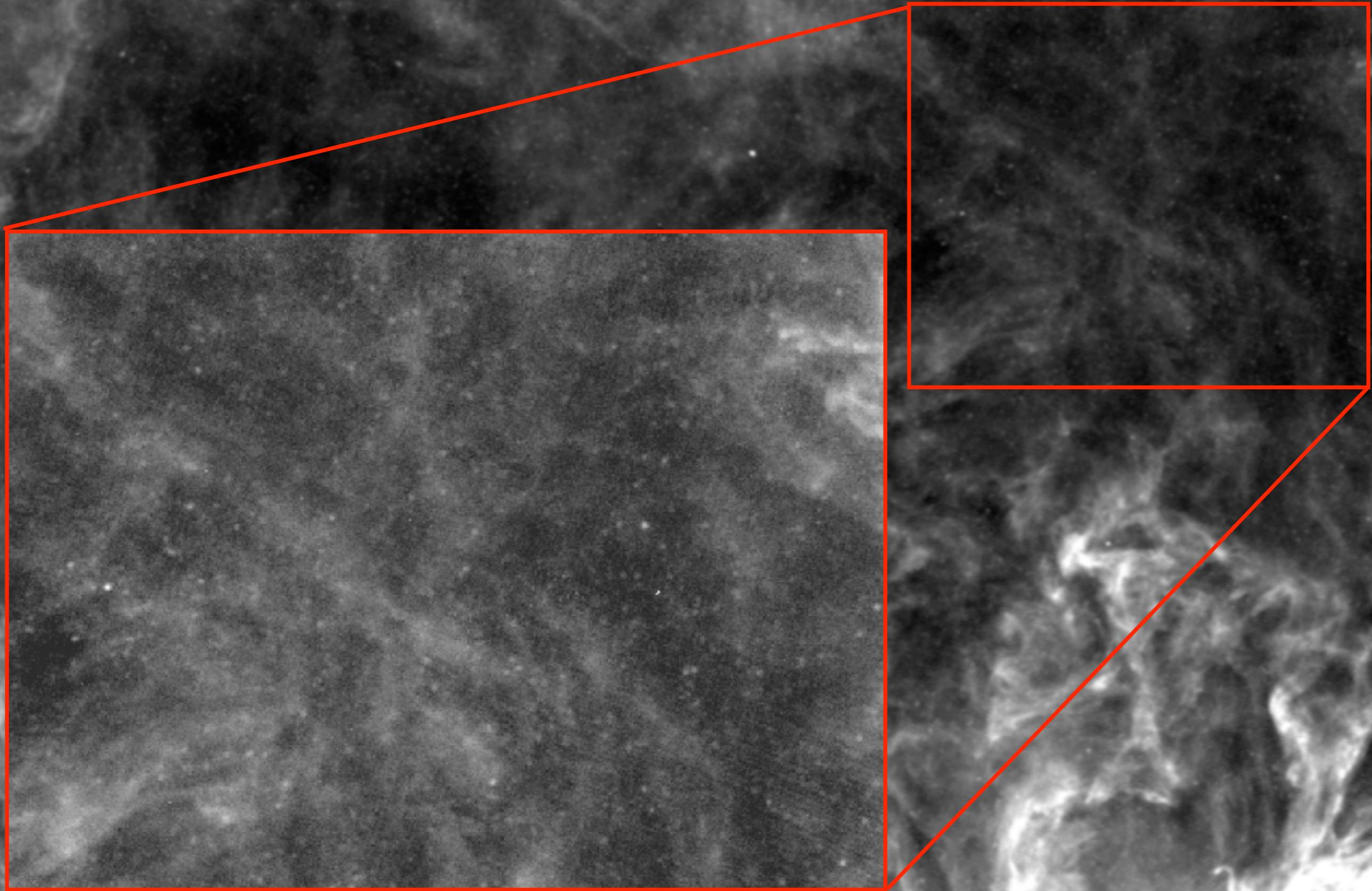
15 m

3.5 m

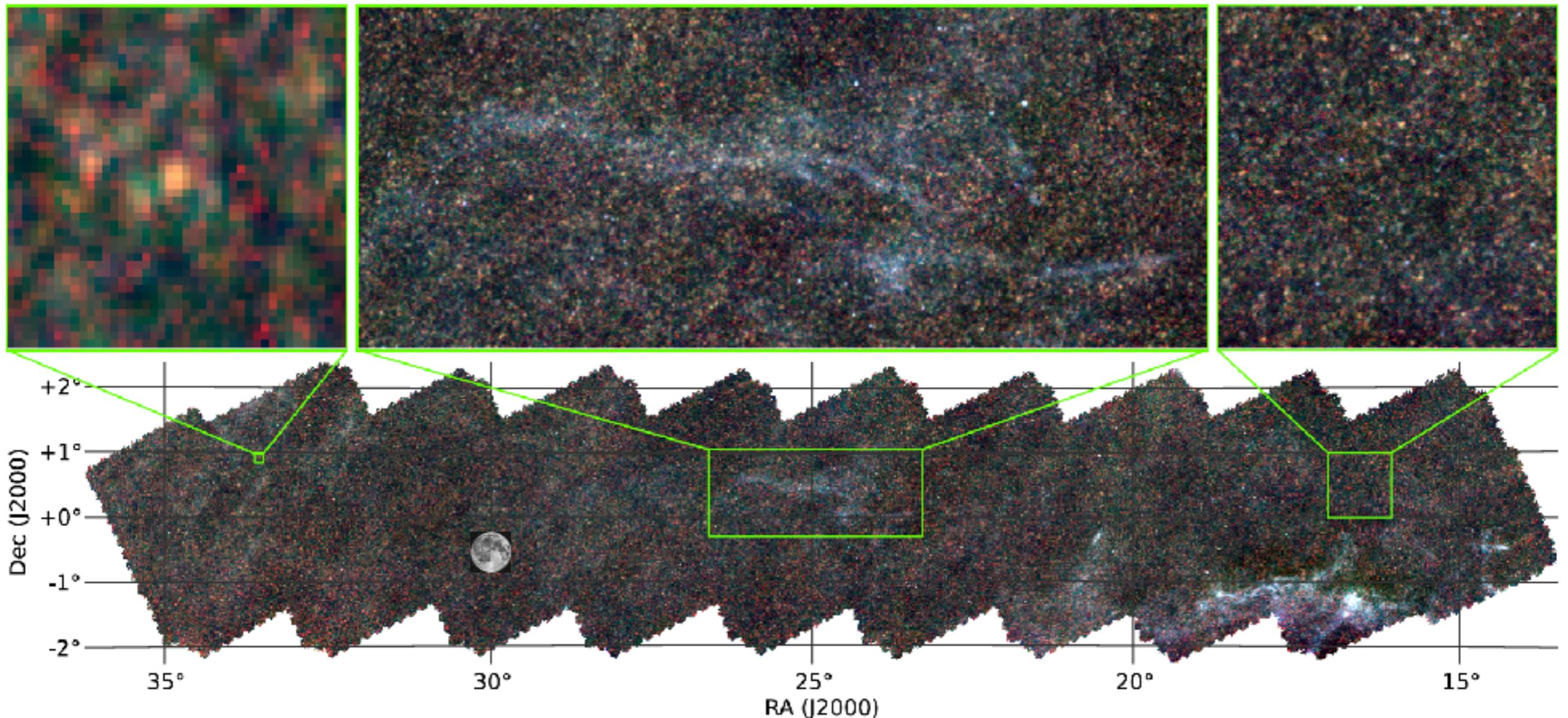
How to beat down confusion noise?

- smaller beam! (bigger telescope or shorter wavelength)
- Use positional prior and extract fluxes.
- Don't worry about it and use the Fourier Transform

SPIRE 250 μ m



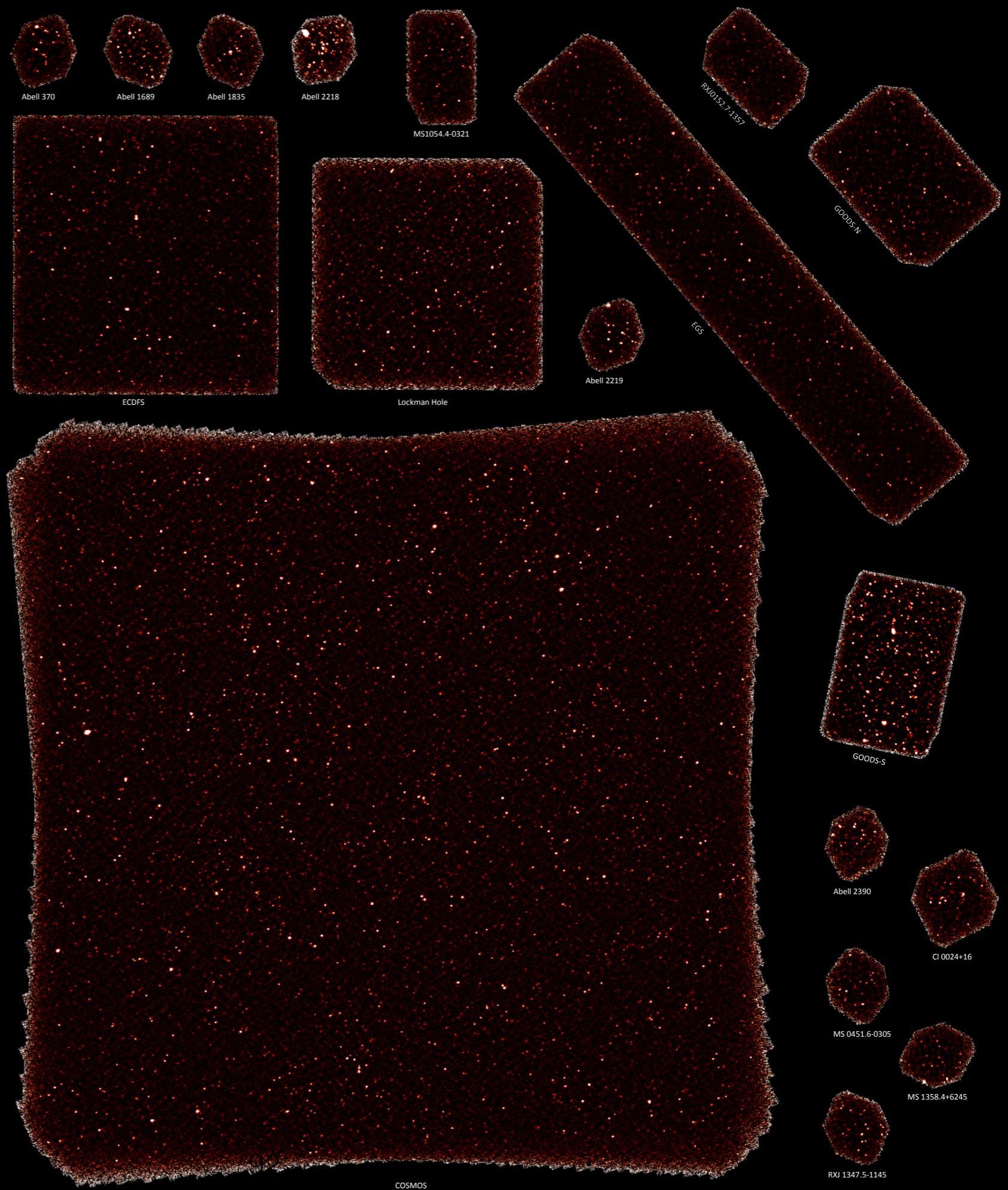
90 deg² Herschel field with cirrus



What are statistical errors? What are systematic?

Herschel/PACS

60, 100, 160 μm

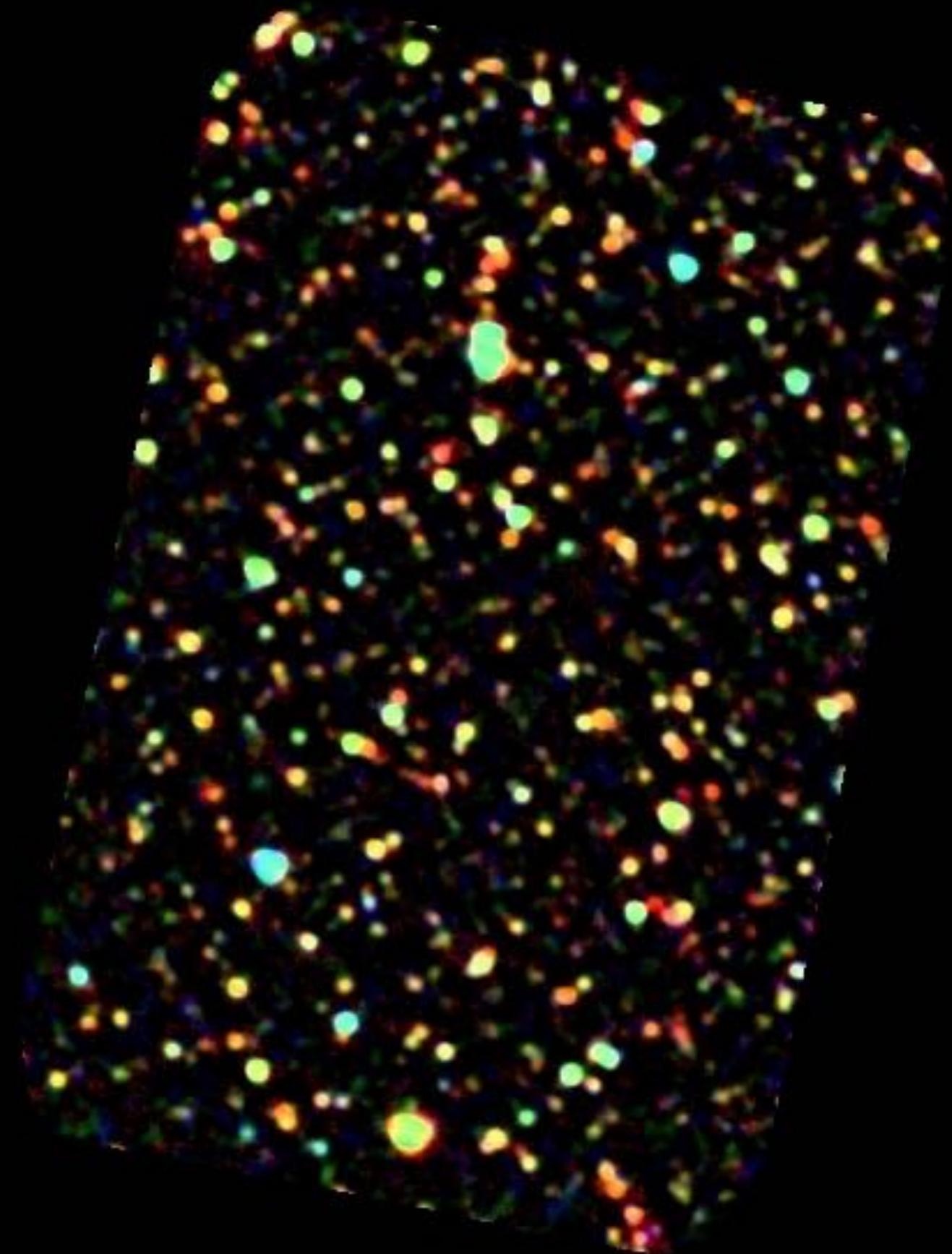


Herschel/PACS
60, 100, 160 μm

Only detector noise !

NOT background limited!

Bad !

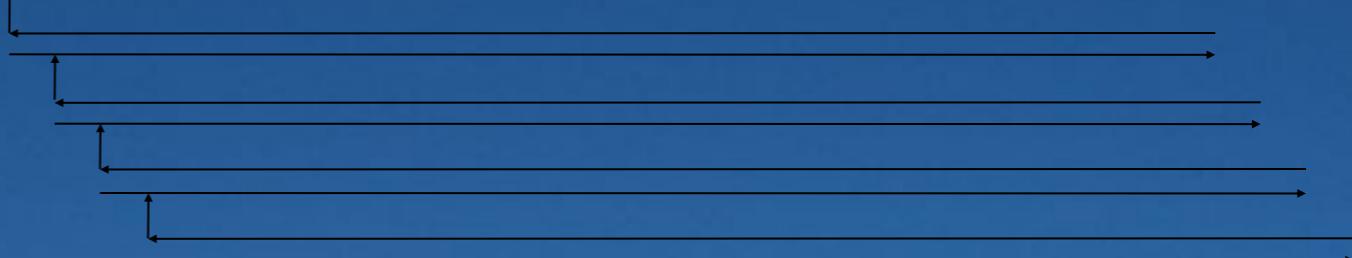


The GOODS-S field as seen by PEP

Problem: low frequency 1/f noise from the instrument and atmospheric fluctuations contaminate the astrophysical signal.

Solution: fast scanning – noise at low *temporal* frequencies contaminates only low *spatial* frequencies in the data

...continue to top of field, then repeat



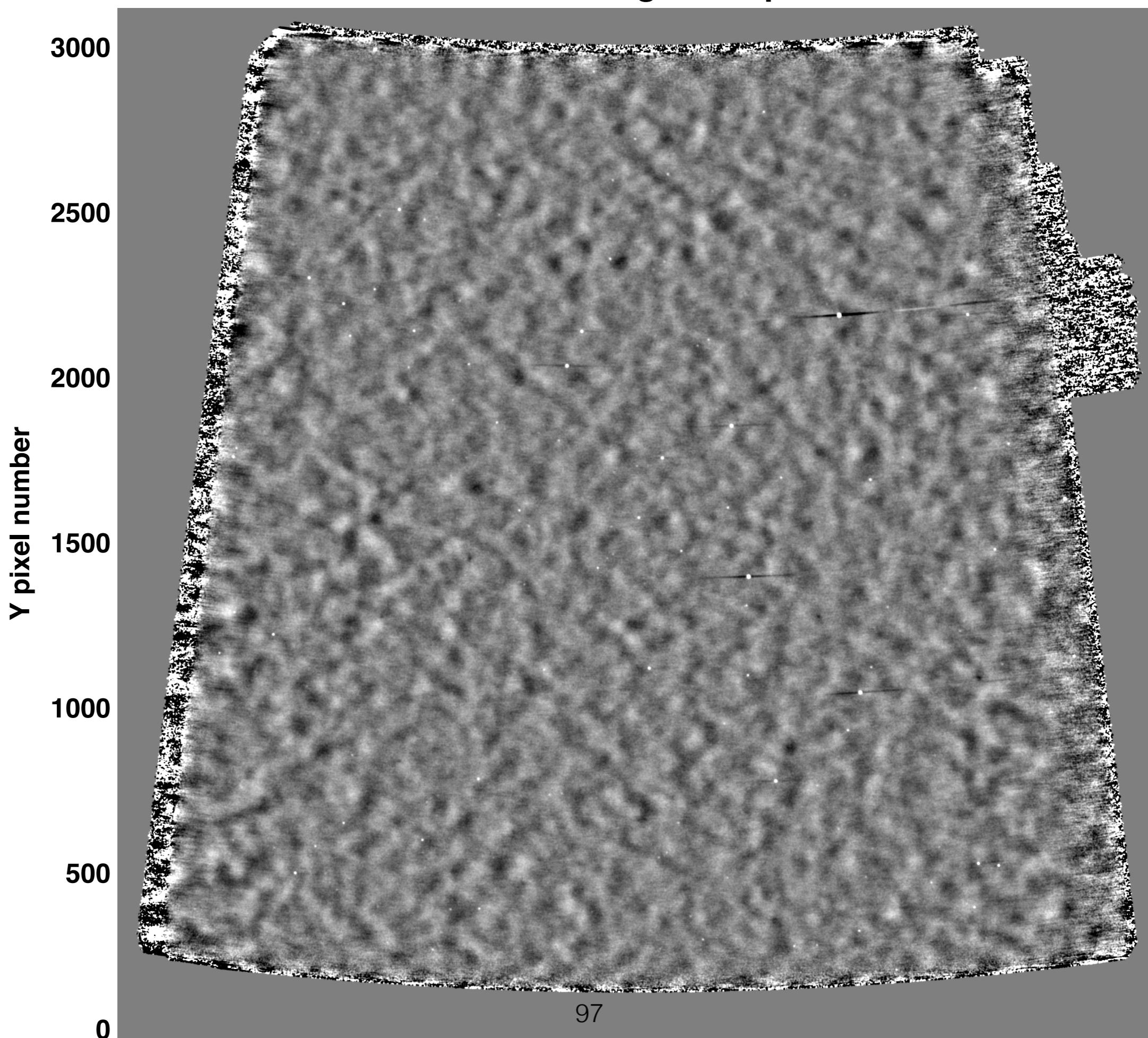
94



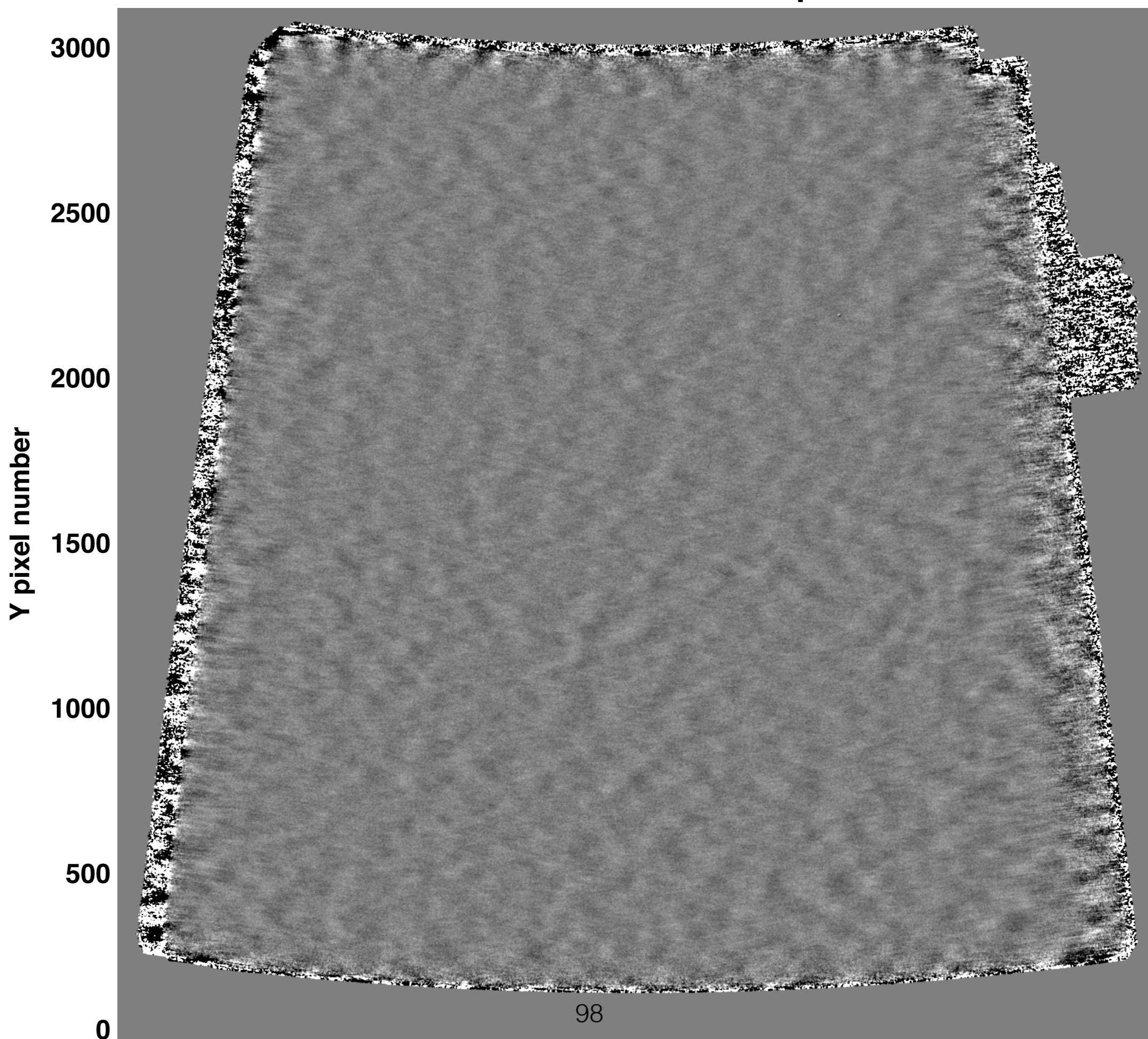


SPTpol
150 GHz.
30 deg²

2.0 mm signal map



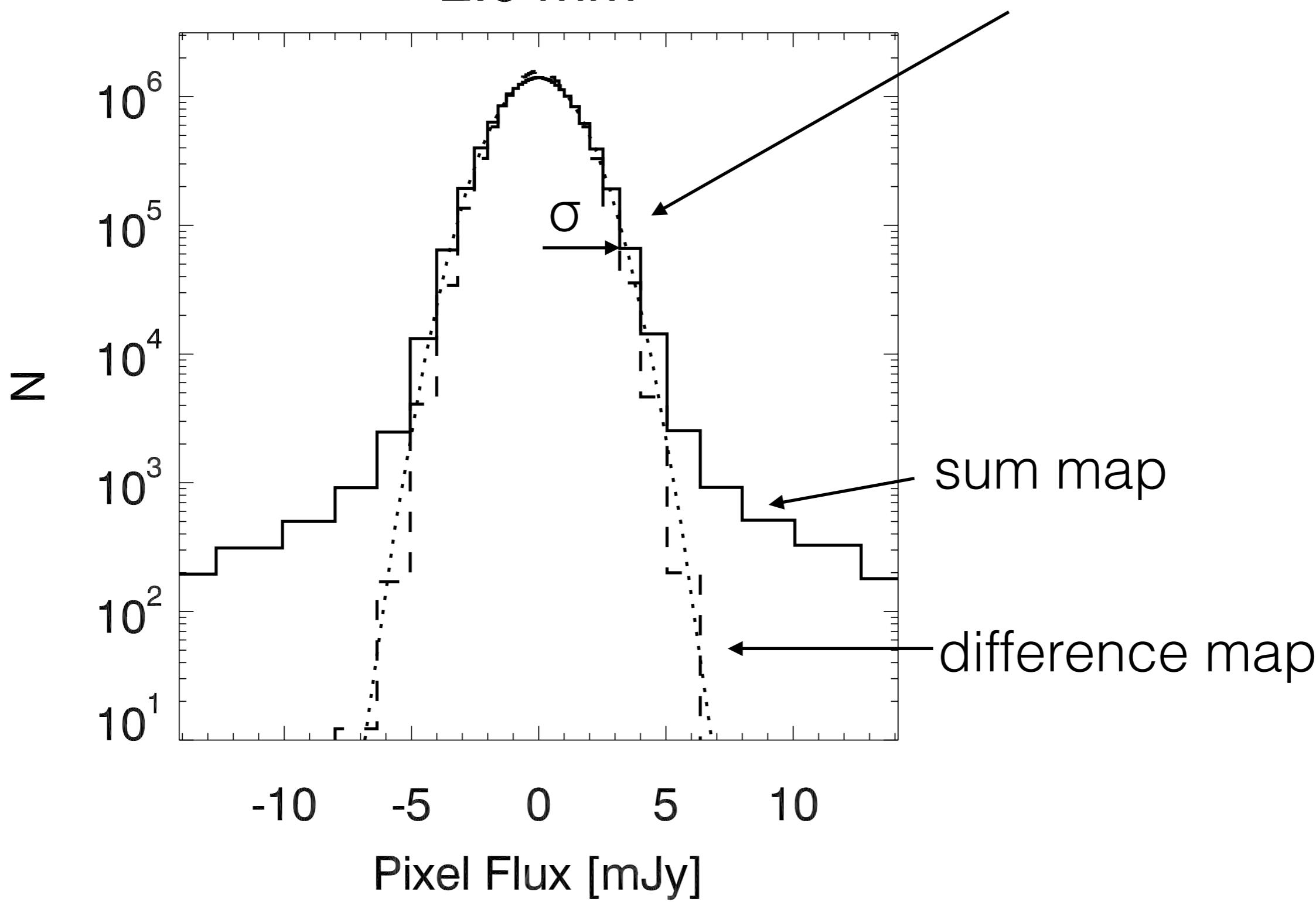
2.0 mm difference map



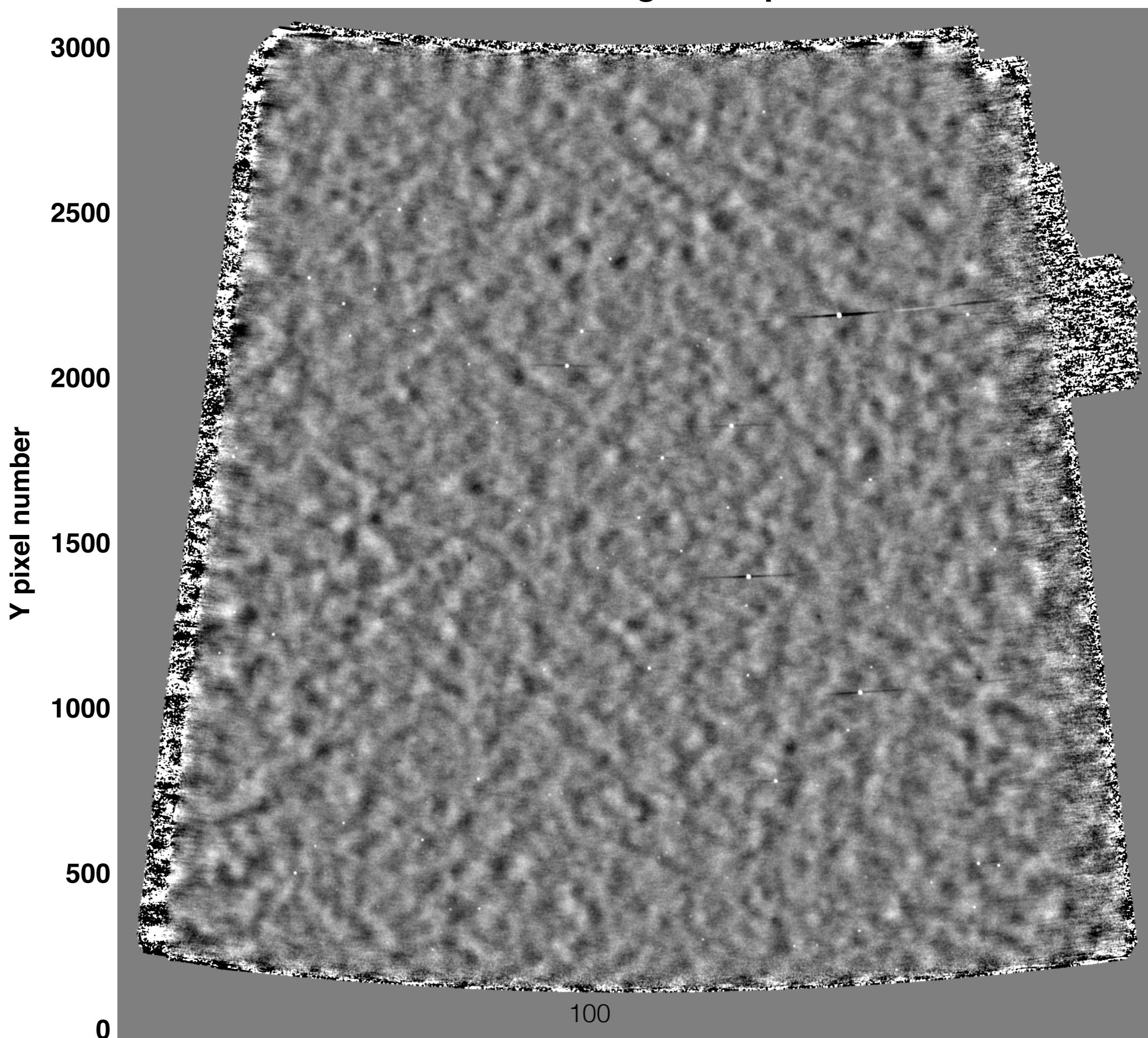
map pixel histogram

2.0 mm

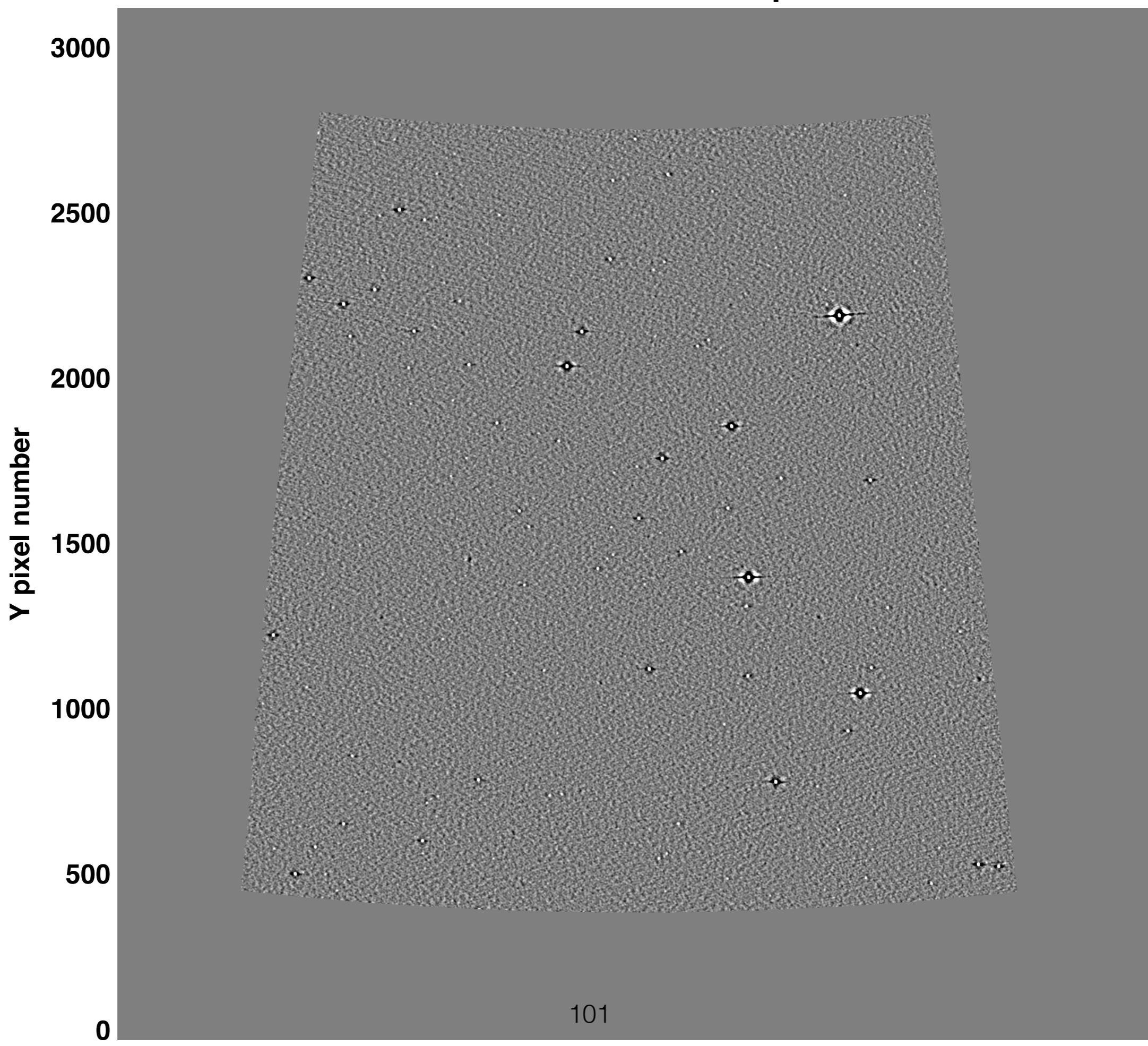
this is the noise in the map



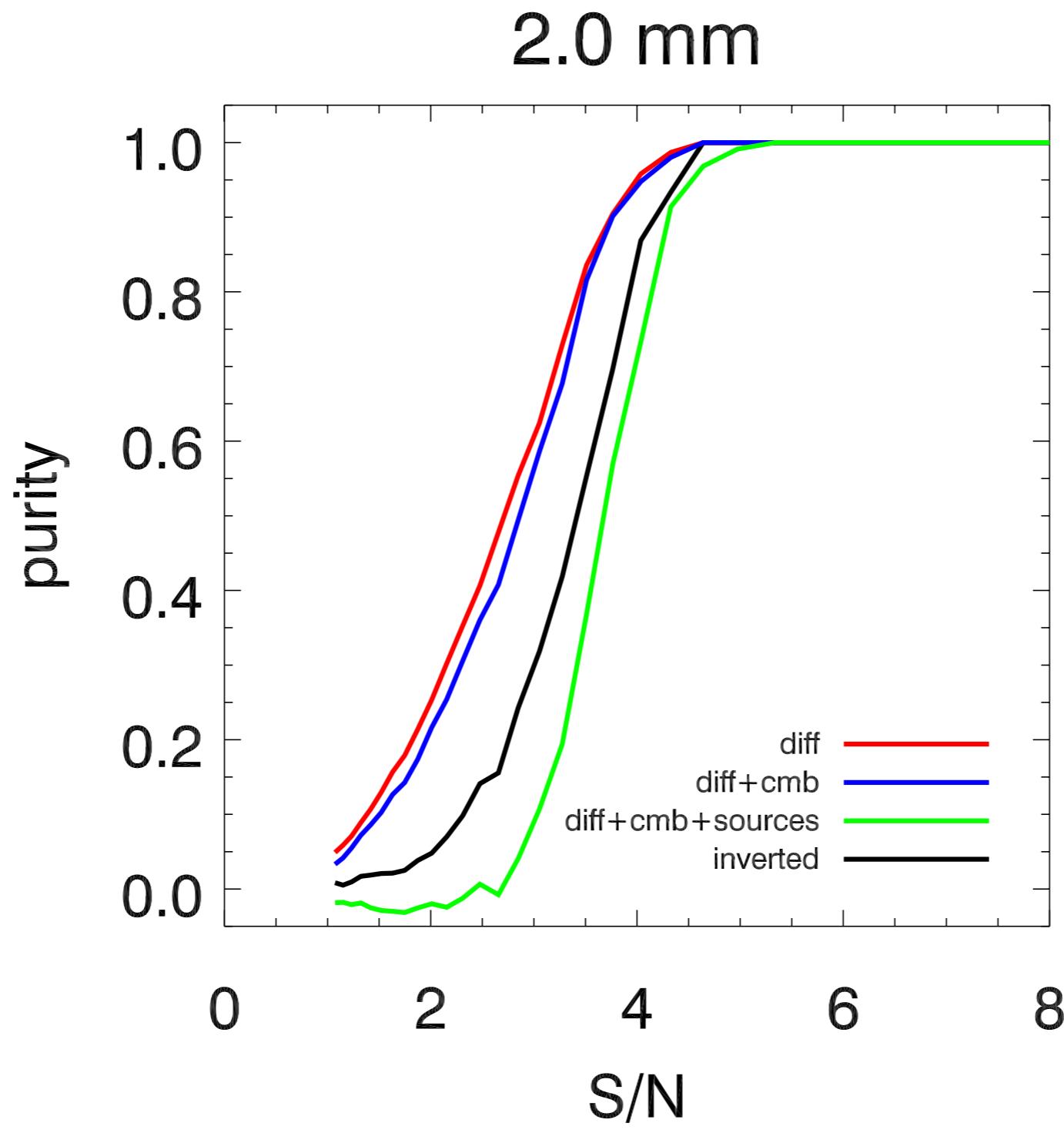
2.0 mm signal map



2.0 mm filtered map



Catalog Purity



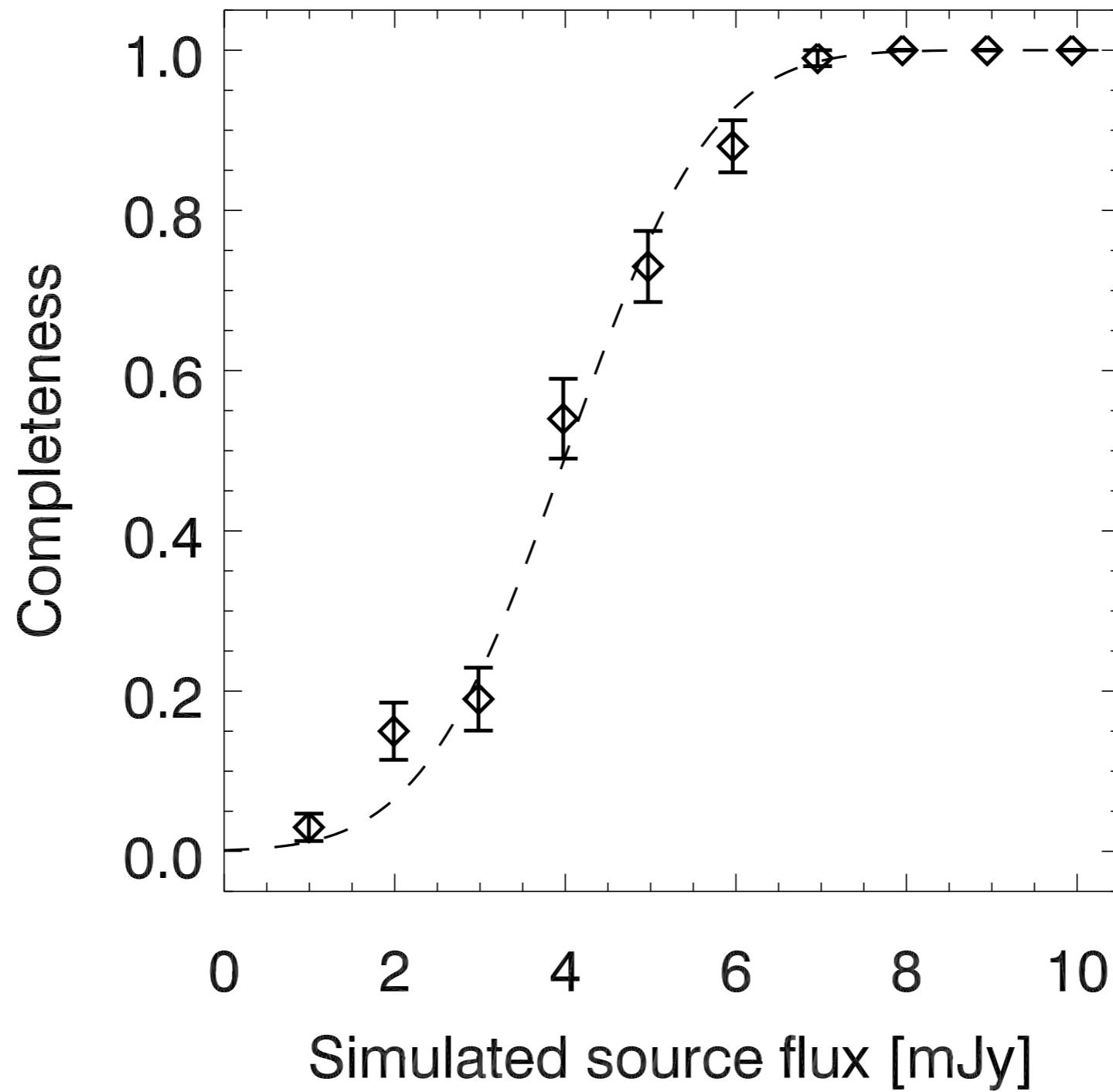
At a given S/N,
what fraction of
sources are real?

multiple ways of
testing this.

Which is best?

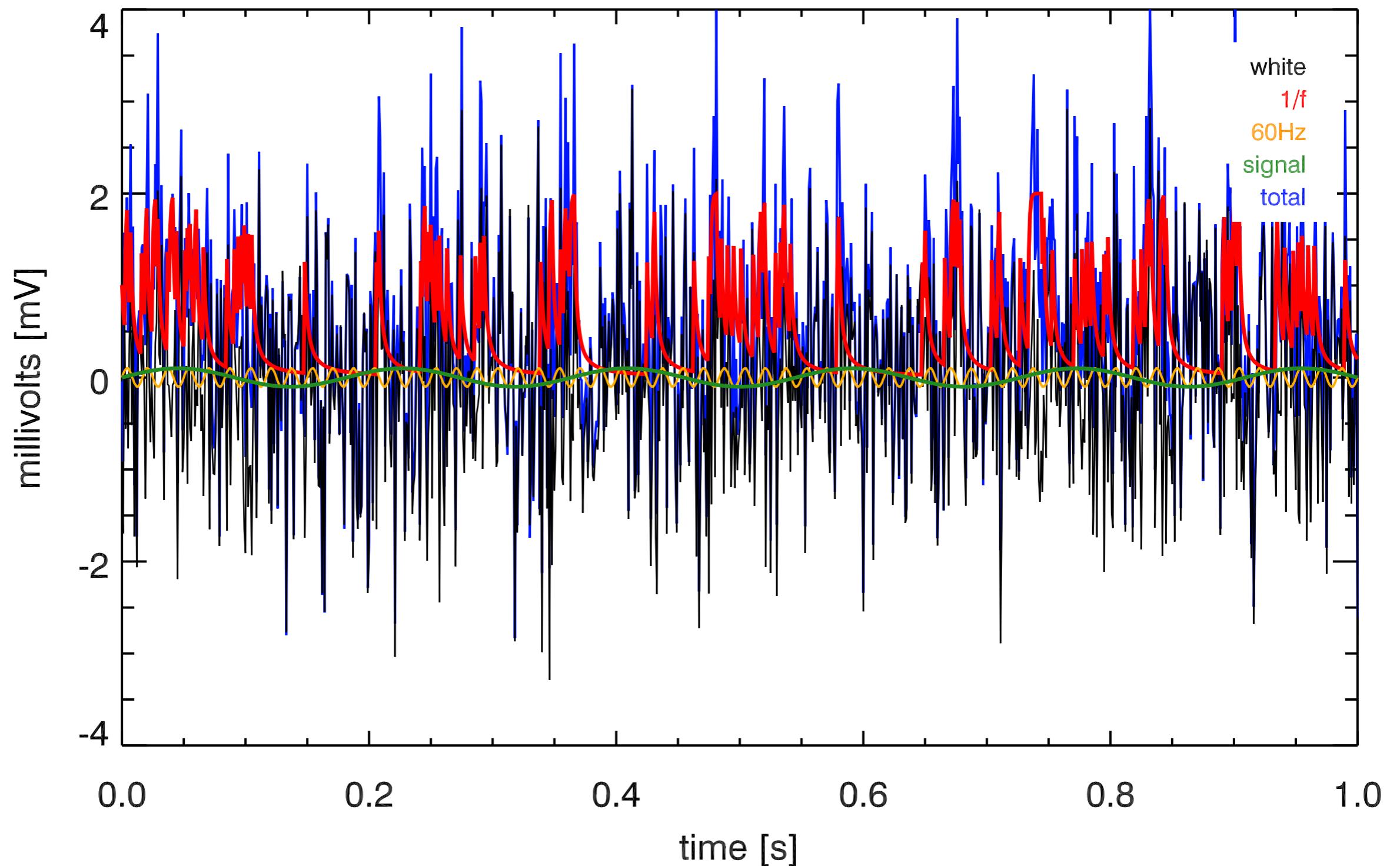
Catalog Completeness

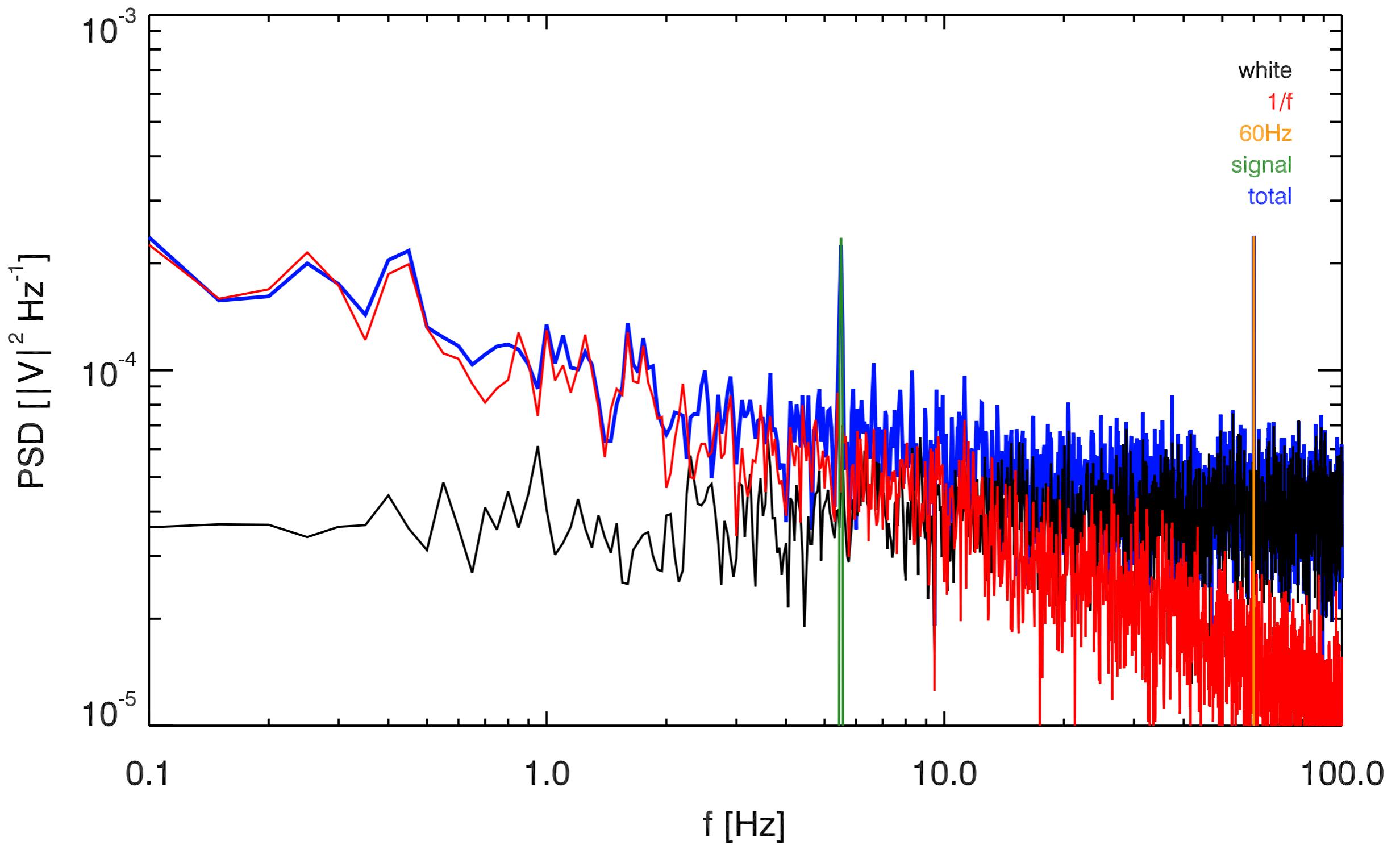
2.0 mm

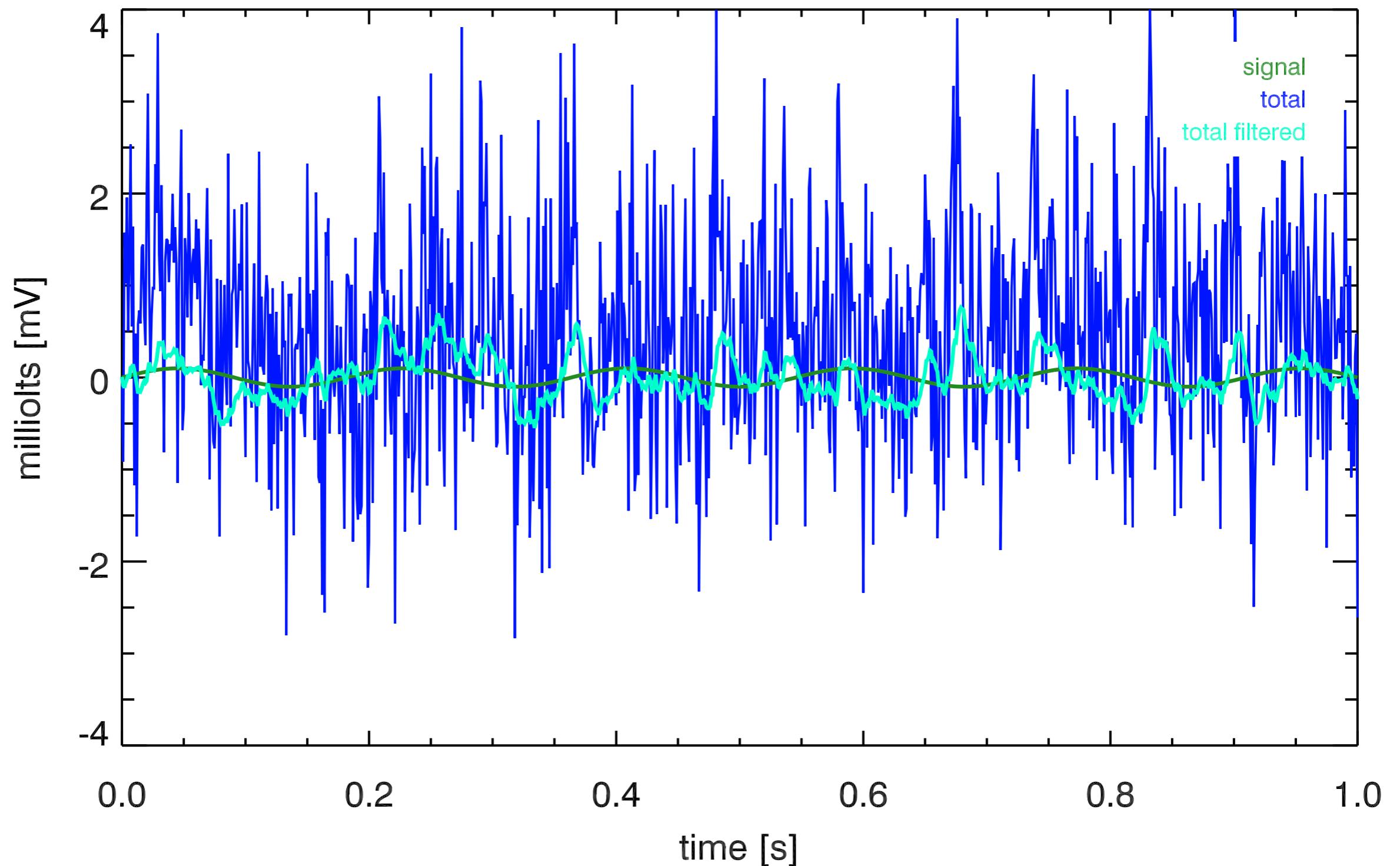


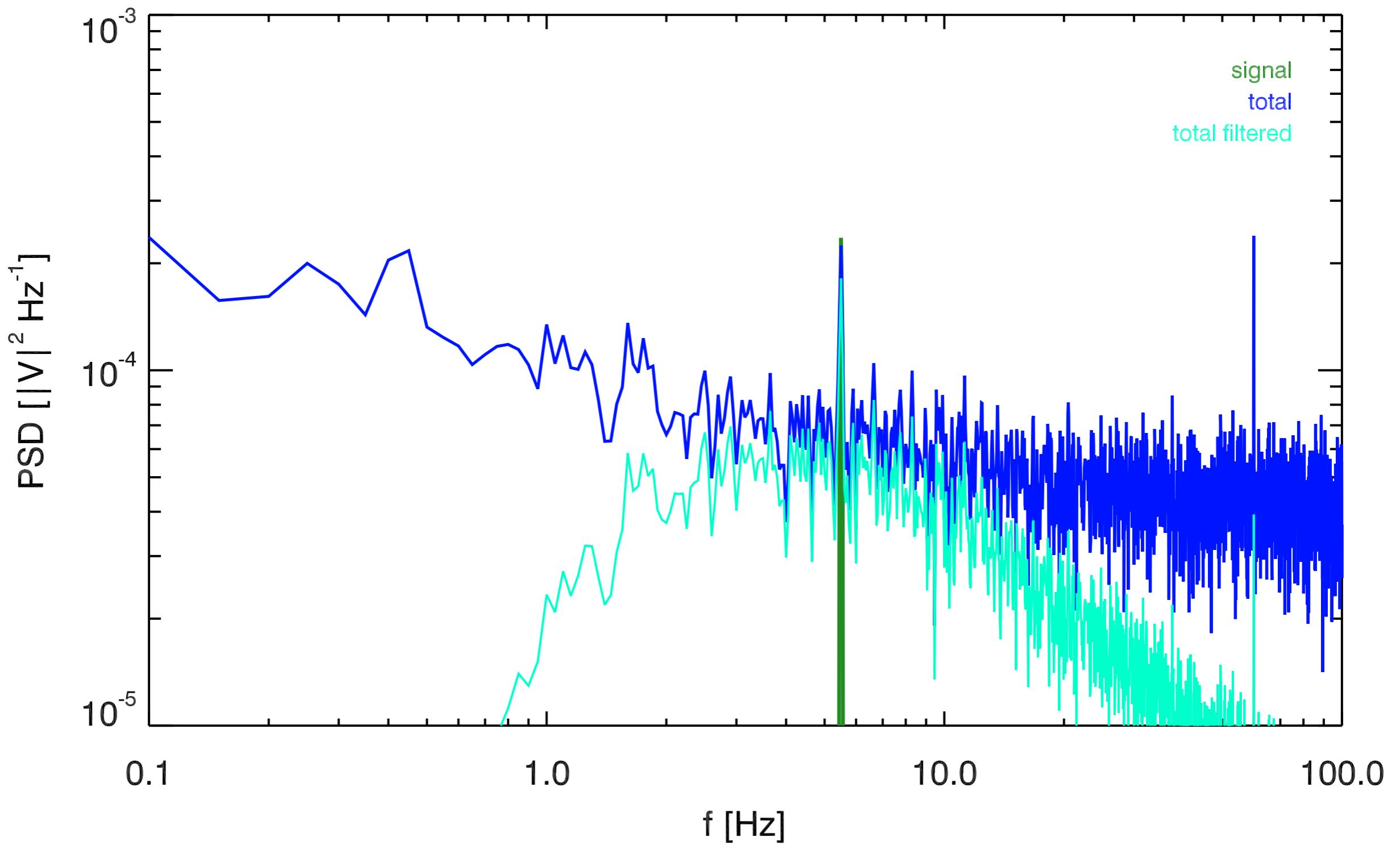
How many true sources do you expect to recover at a given flux level?

erf with 50% at the detection threshold and width the noise

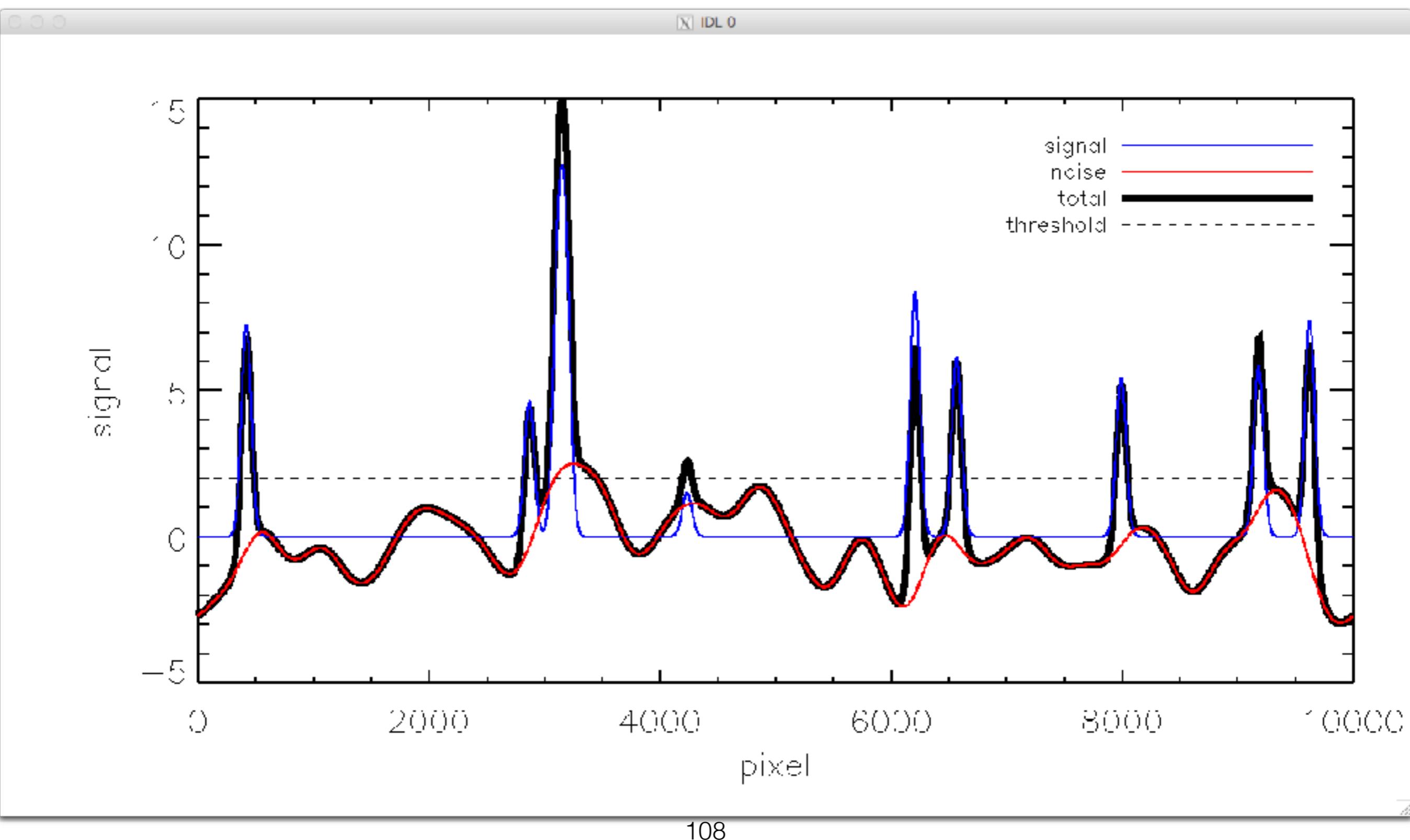








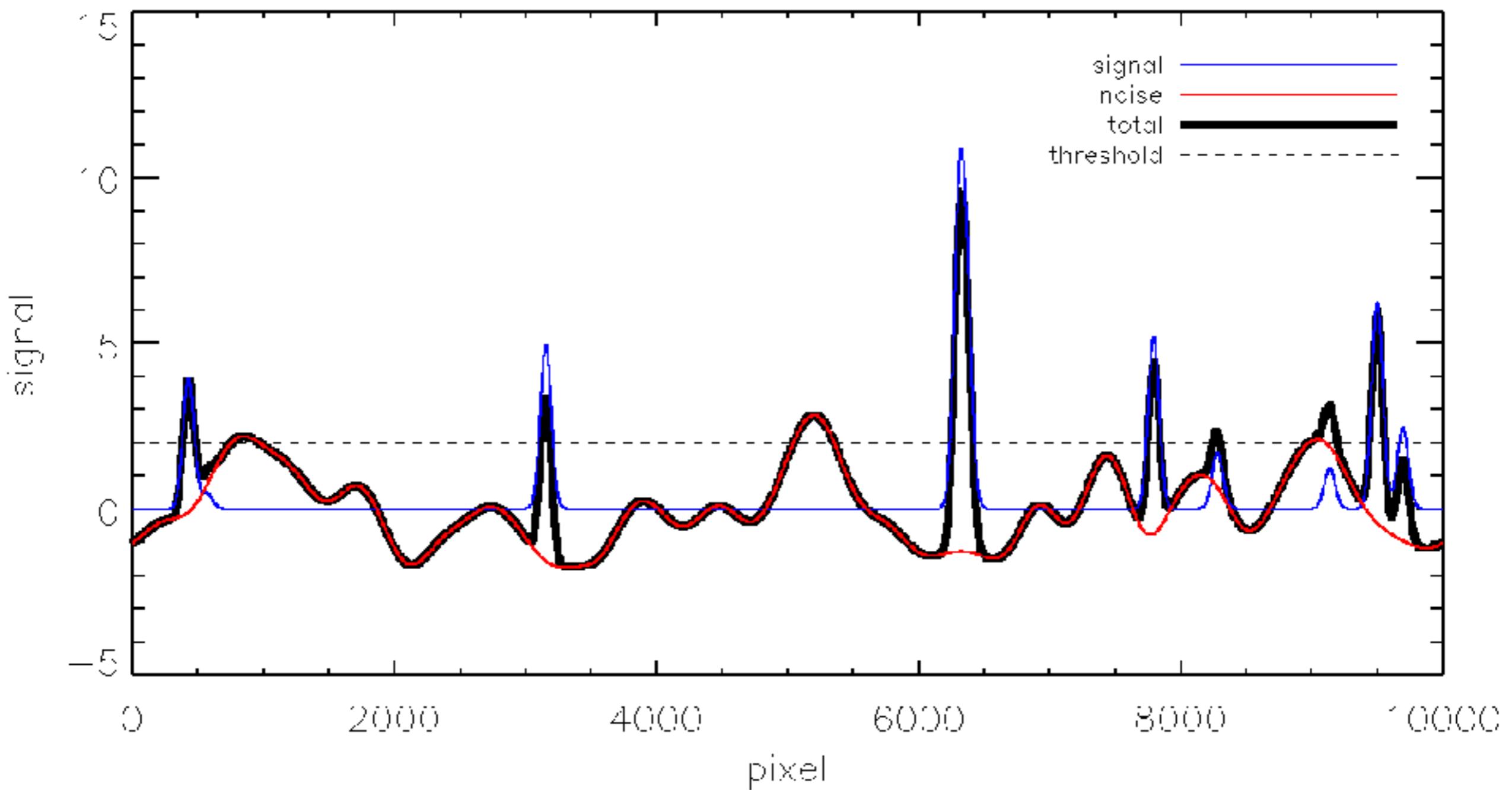
simulation exercise for Thursday



discussionquestion

- A. 6
- B. 7
- C. 9
- D. 18

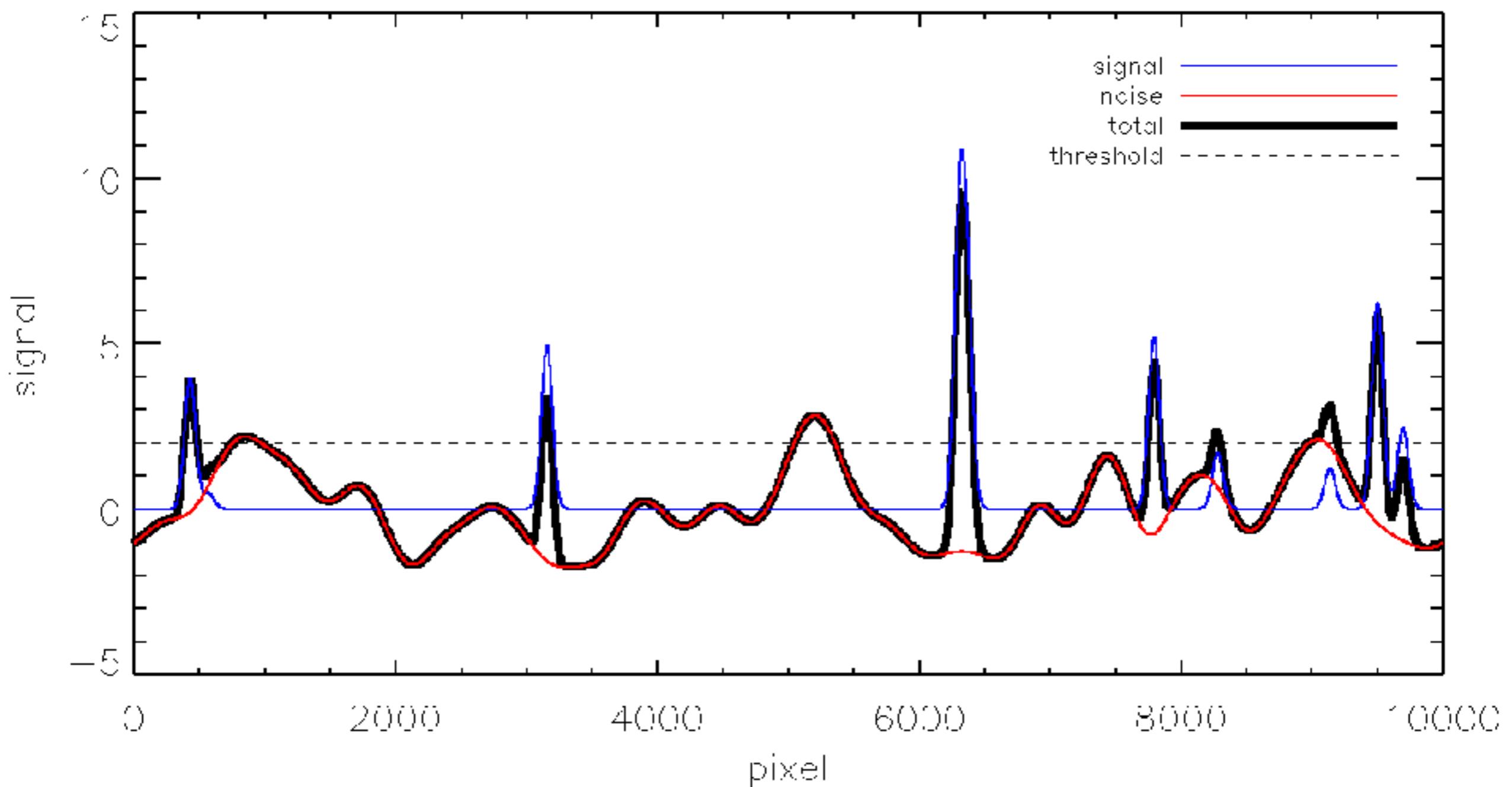
How many sources detected?



discussion question

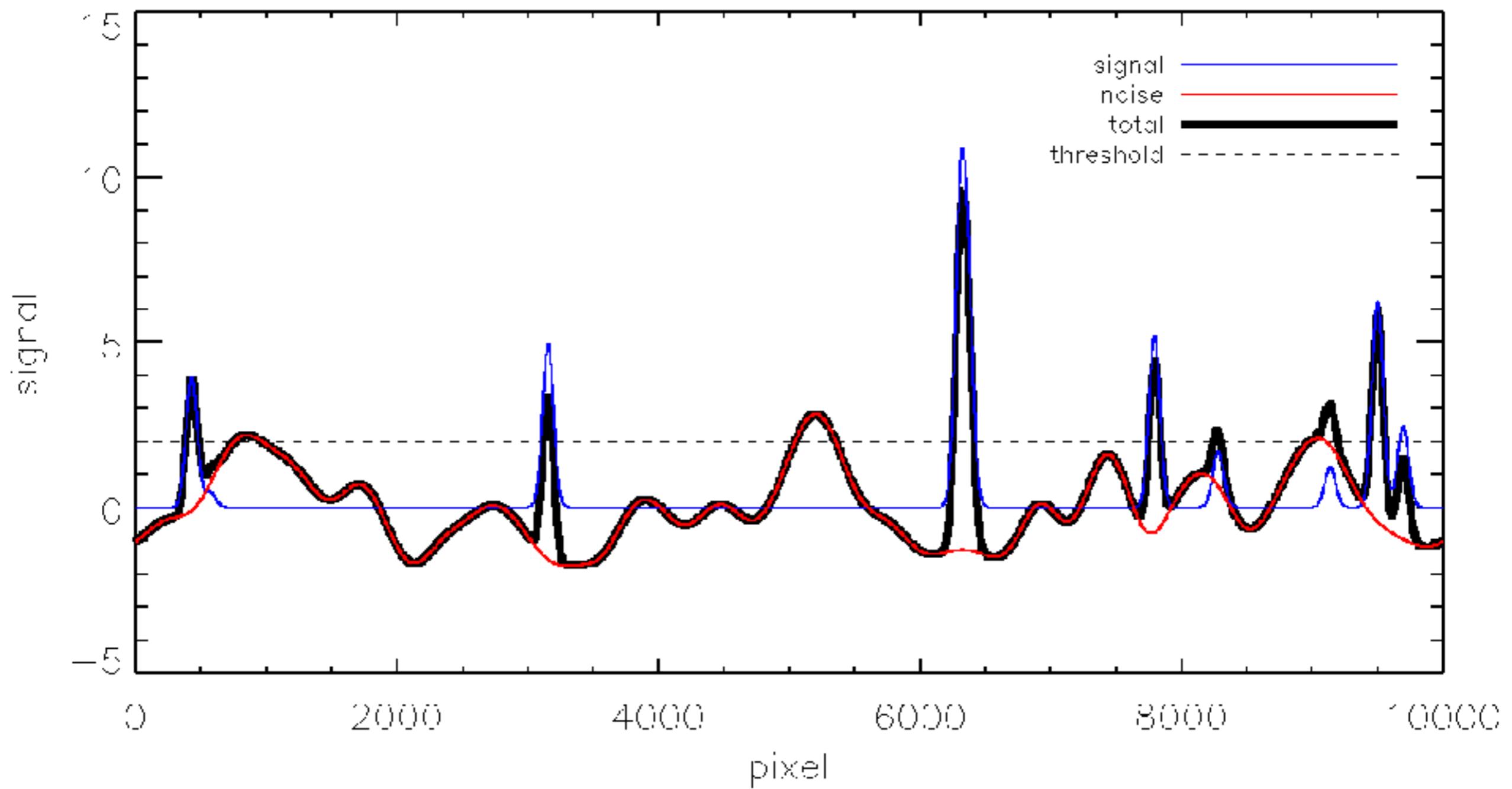
- A. 0
- B. 1
- C. 3
- D. 6

How many real sources that should have been detected are missed?



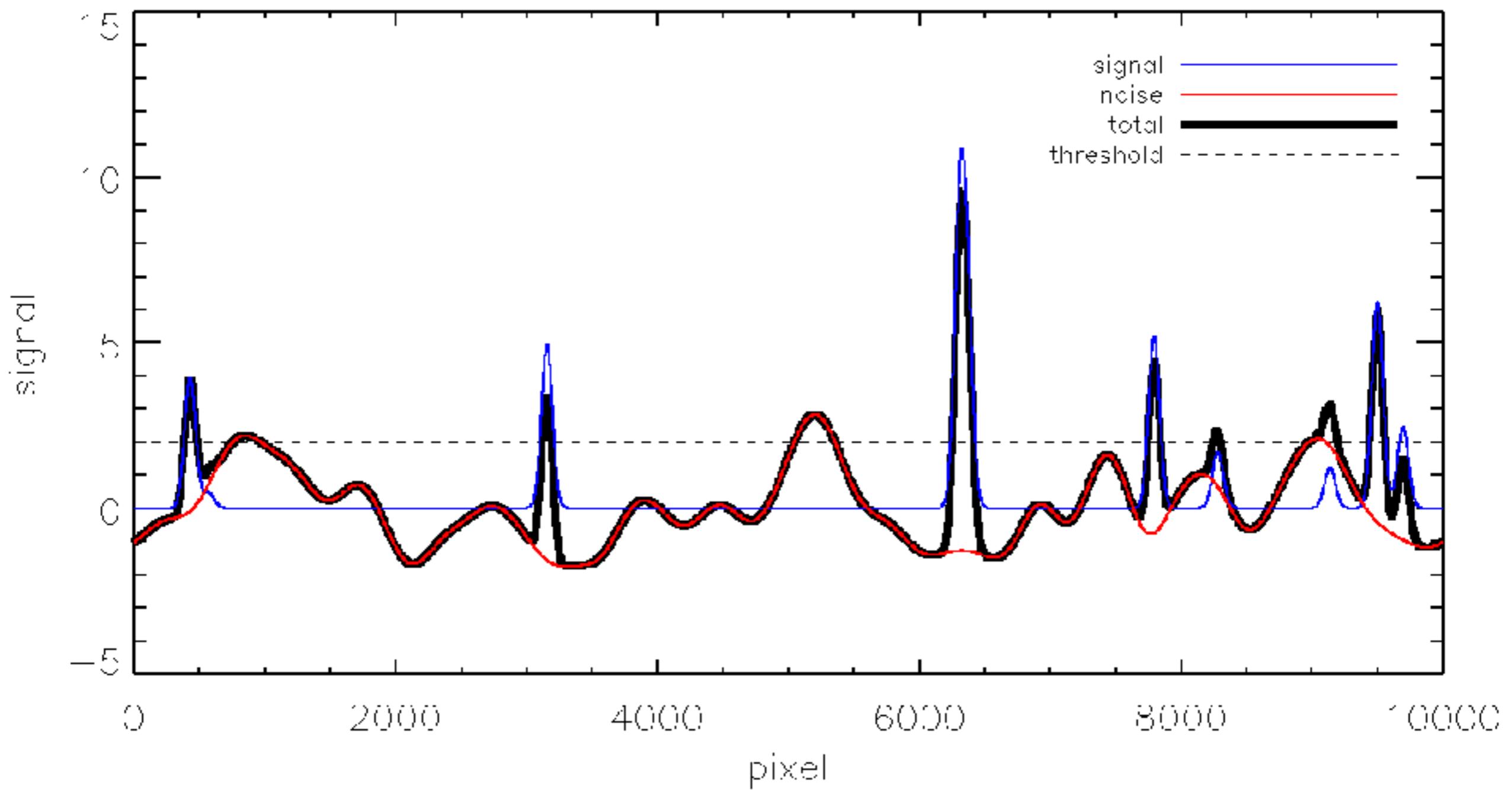
Completeness

n real sources detected / n real sources above threshold = $5/6$
= 83%

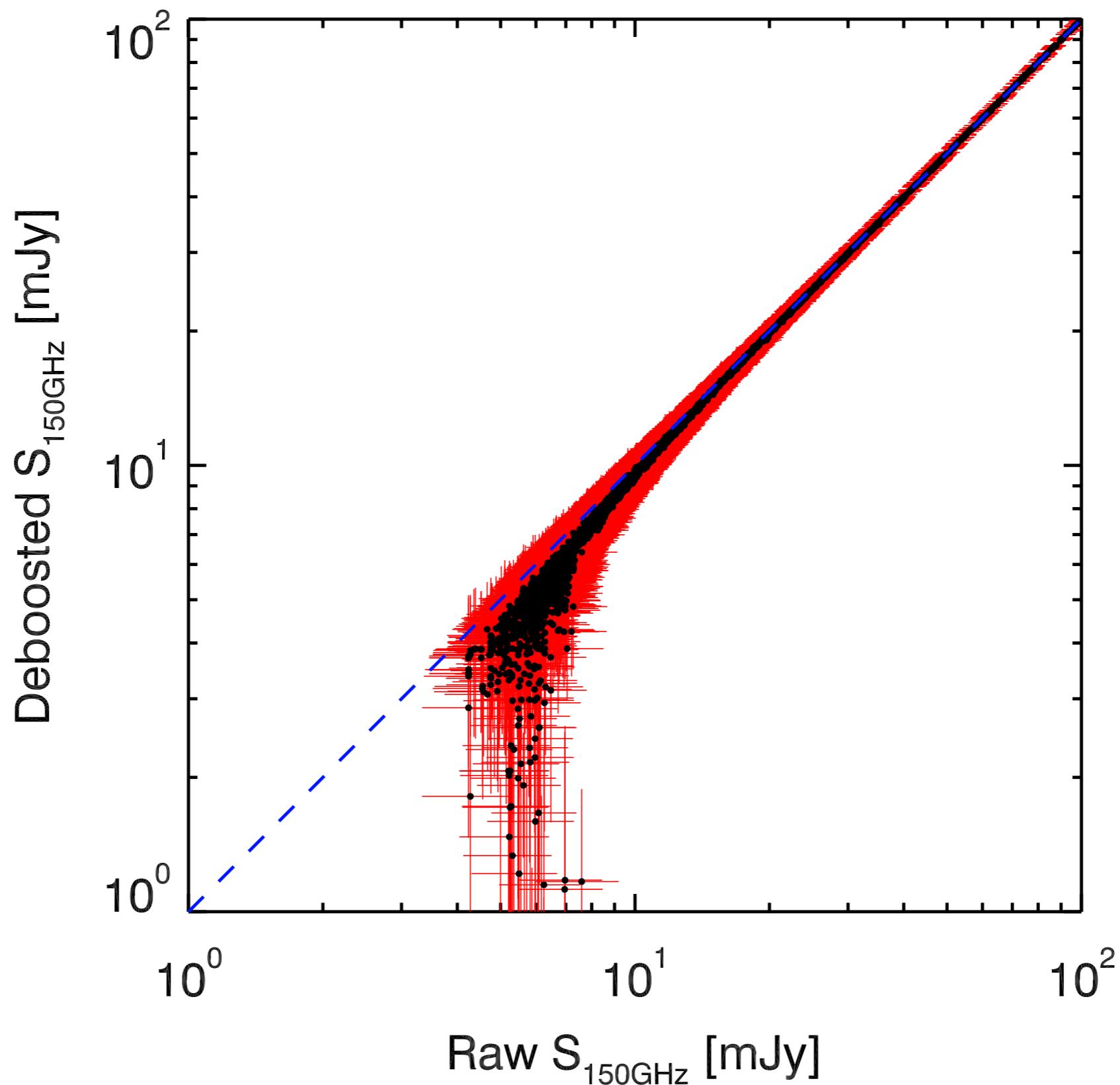


Purity

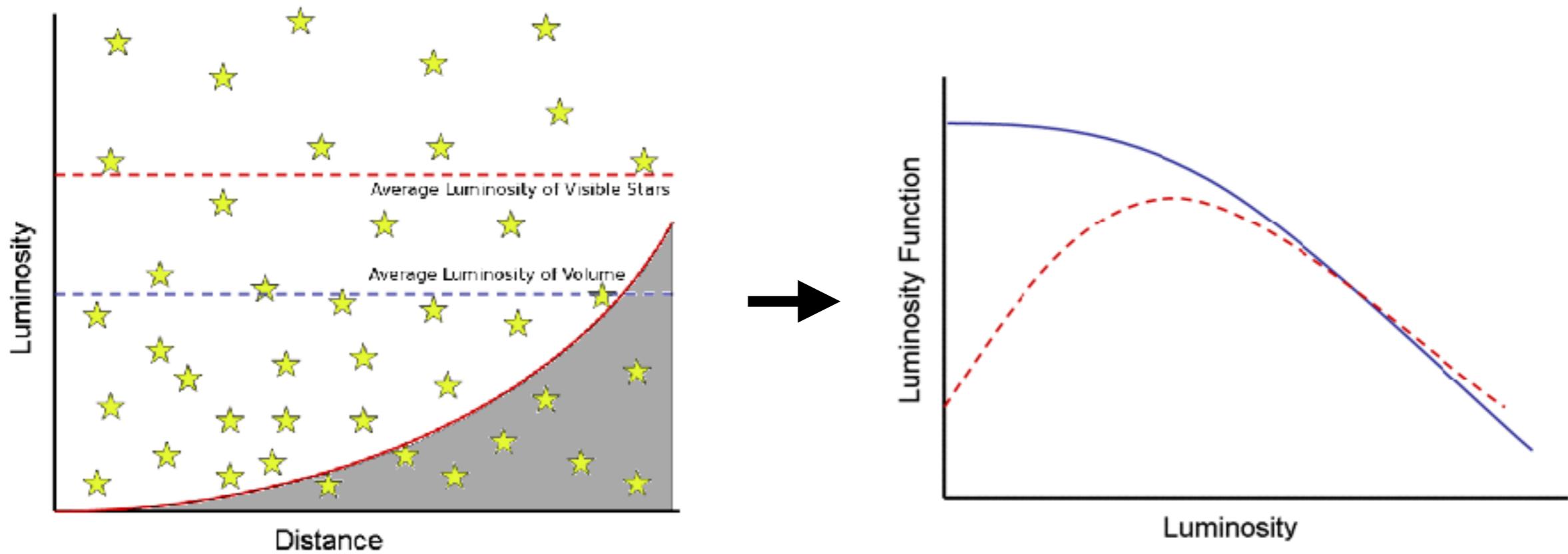
n real sources above threshold / n sources detected = $5/9 = >50\%$



Flux boosting bias

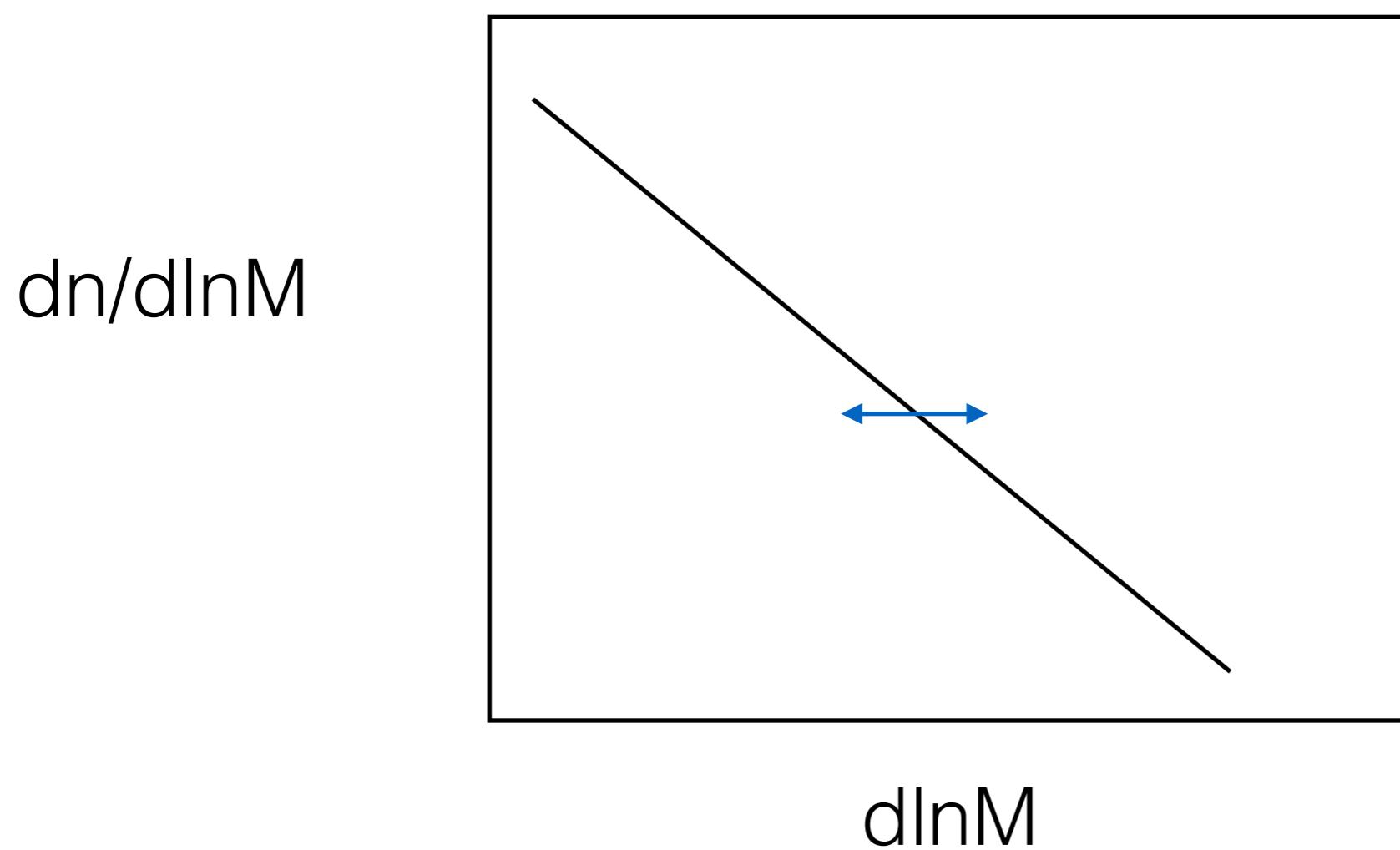


Malmquist bias



things farther away tend to be more luminous
not account for this can bias your luminosity function

Eddington Bias



There tend to be more faint things than bright things.

Flux boosting will make more faint things brighter
than bright things fainter

discussion question

Q: Which of these sources of noise is statistical and which is systematic?

Detector Noise

A: Statistical.

B: Systematic.

C: Depends on what you are trying to do.

discussion question:

Q: Which of these sources of noise is statistical and which is systematic?

1/f Noise

A: Statistical.

B: Systematic.

C: Depends on what you are trying to do.

discussion question:

Q: Which of these sources of noise is statistical and which is systematic?

Confusion Noise

A: Statistical.

B: Systematic.

C: Depends on what you are trying to do.

discussion question:

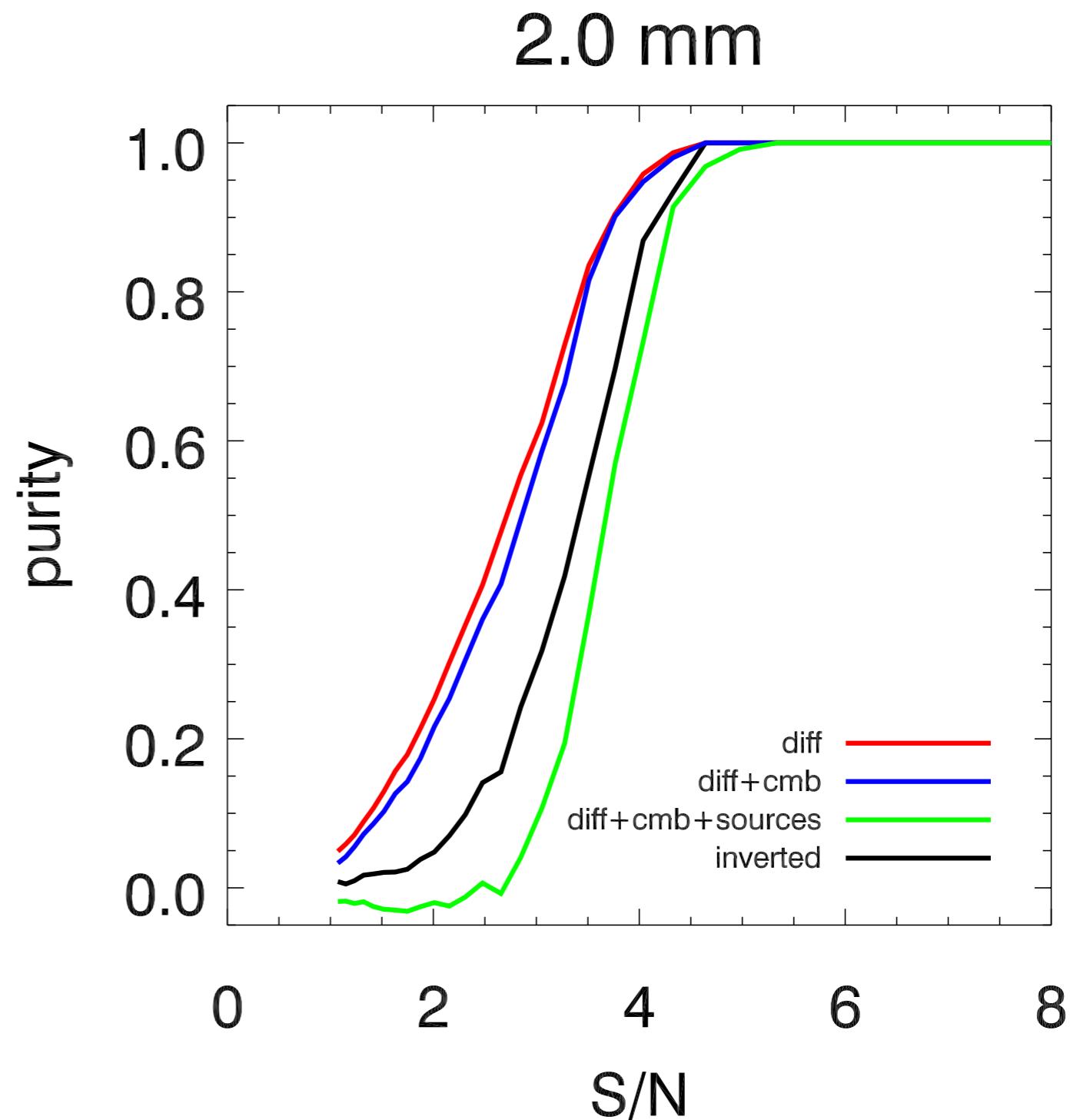
Q: What can you tell from the measurement of catalog *purity* in the plot on the right?

A: We don't know how to tell which sources are real.

B: 50% of the sources are false.

C: 100% of the sources at $S/N > 5$ are real

D: There is a 50% chance of extracting a real source.



discussion question:

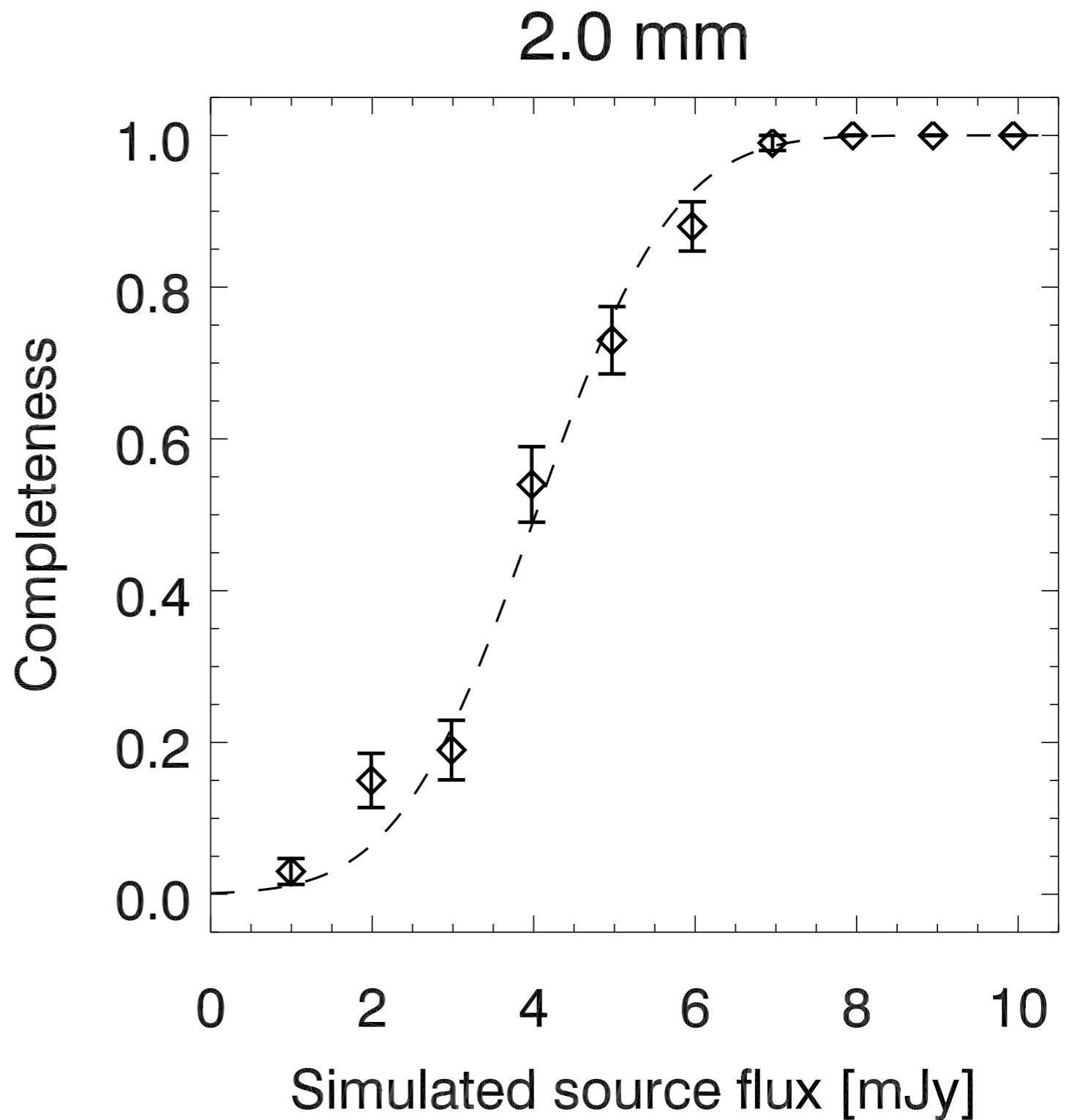
Q: What can you tell from the measurement of catalog completeness in the plot on the right?

A: We don't know how to tell which sources are real.

B: 50% of the sources are false.

C: 100% of the sources at $S > 7\text{mJy}$ are real

D: There is a 50% chance of extracting a real source above $S > 4\text{mJy}$.



Noise

- **General types of noise:** white, thermal, shot, brownian, 1/f
- **General categories of noise:** Statistical v. Systematic; correlated v. uncorrelated
- **Types of noise in astrophysics:** instrumental noise, atmospheric noise, confusion noise, astrophysical noise
- **Measurements:** relative v. absolute
- **Catologs:** completeness and purity
- **Bias:** flux, Malmquist, Eddington