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

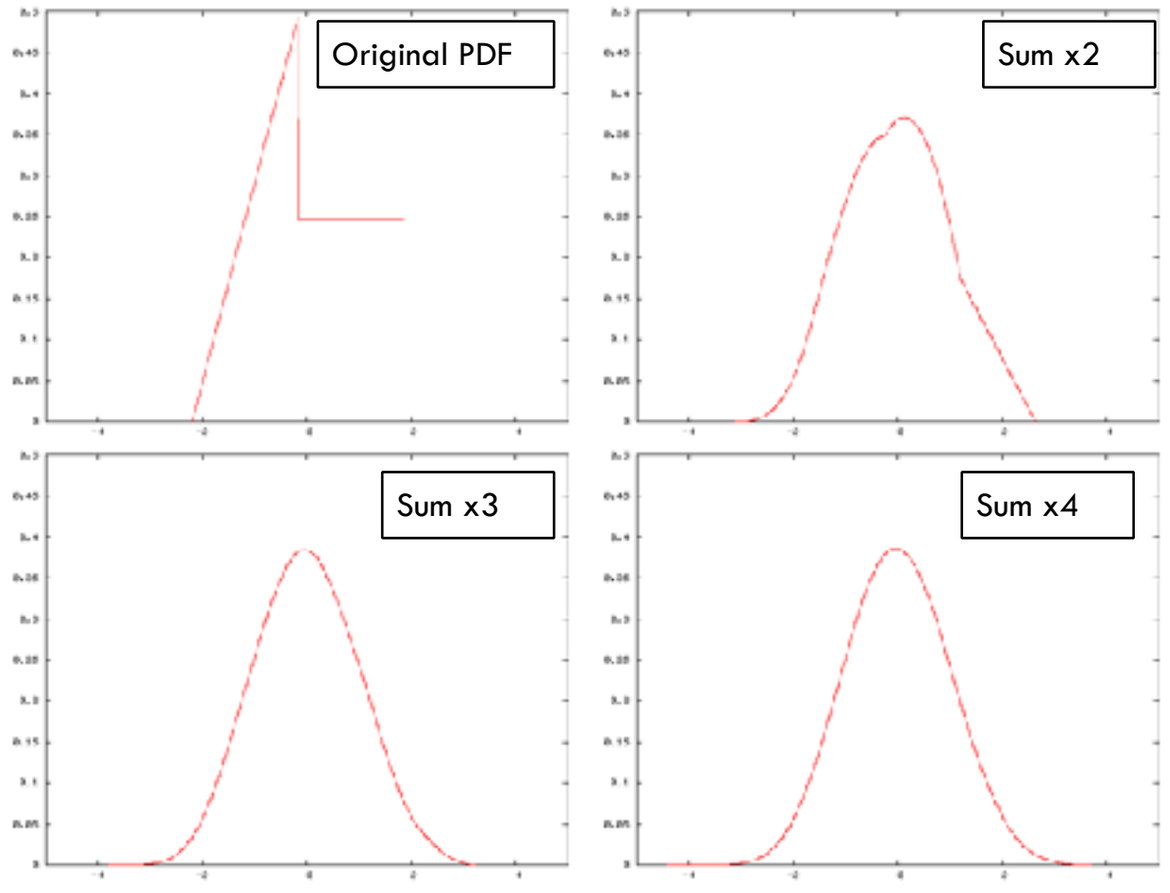
- General tip: Read an assignment fully before you start to tackle it because the hints (and the traps) are frequently buried in the question

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

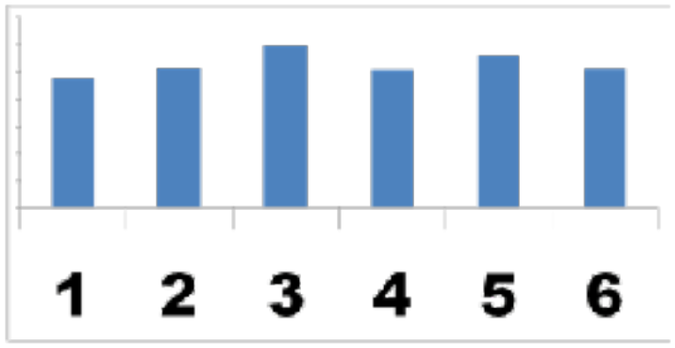
The distribution of an average tends to be Normal, even when the distribution from which the average is computed is decidedly non-Normal.

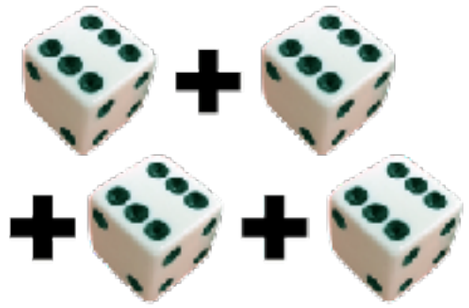


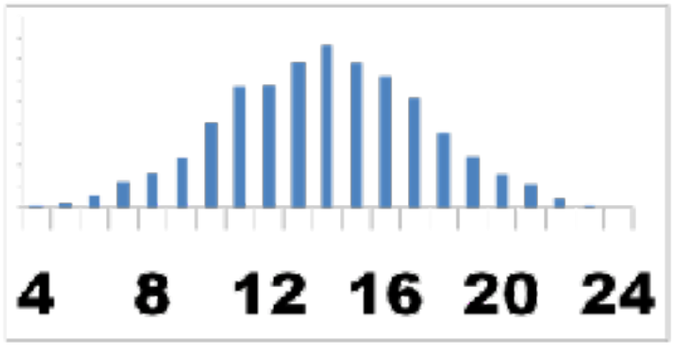
(All scaled to have μ=1, σ=1; CC)

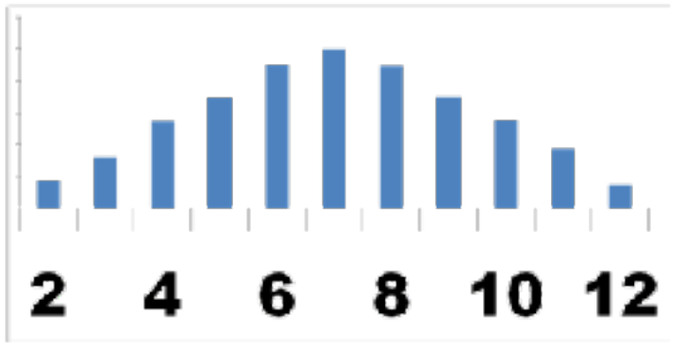
If x is the average of n independent random variables of expected value μ and variance σ^2, then:

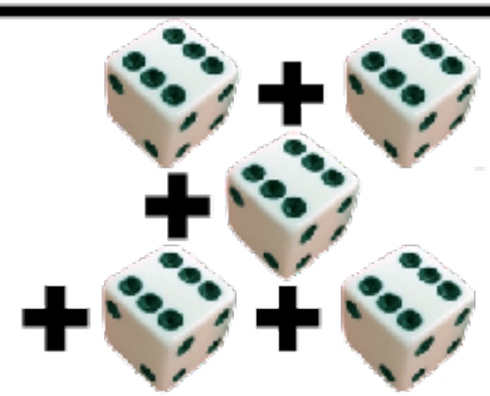
$$\text{As } n \to \infty \quad p(x) \to \frac{1}{\sqrt{2\pi \frac{\sigma^2}{n}}} e^{-\frac{(x-\mu)^2}{2\frac{\sigma^2}{n}}}$$

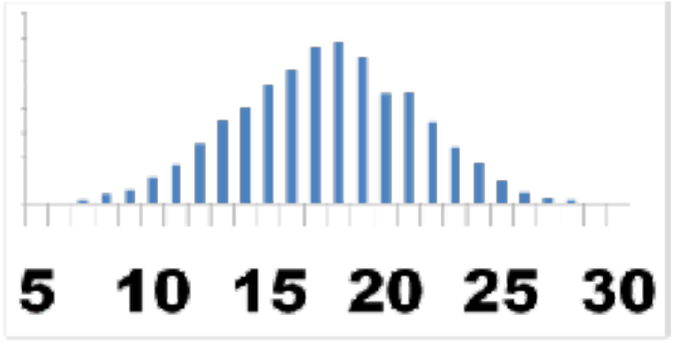

1 2 3 4 5 6

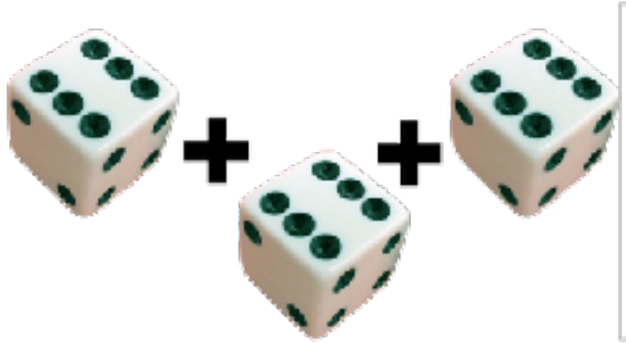


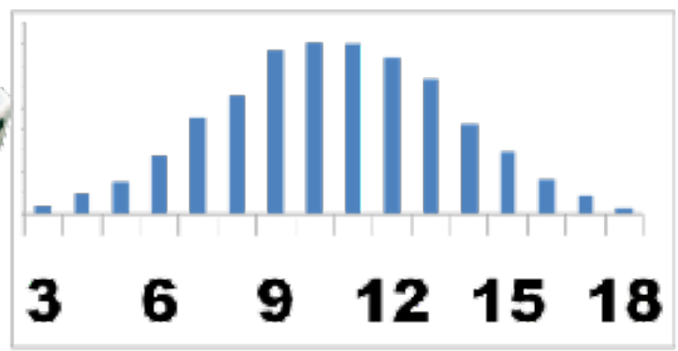

4 8 12 16 20 24

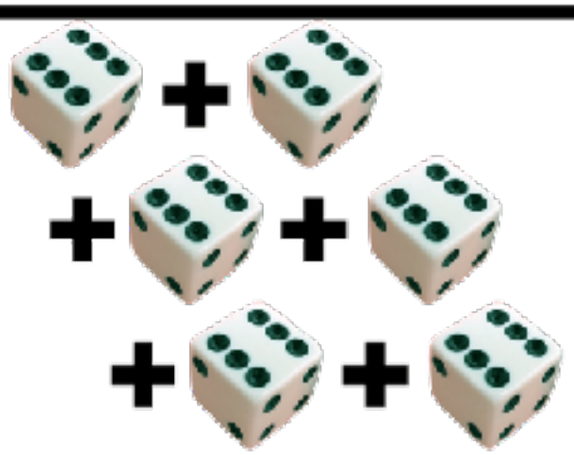

2 4 6 8 10 12

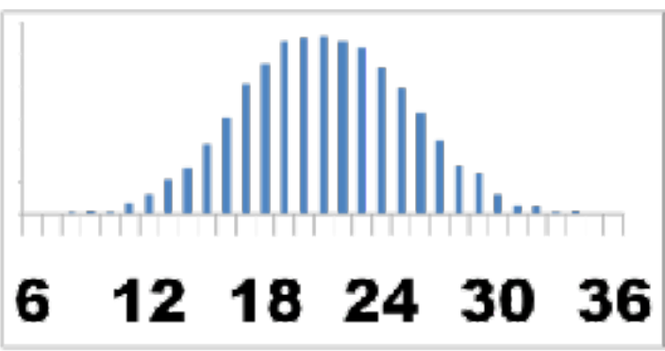



5 10 15 20 25 30




3 6 9 12 15 18




6 12 18 24 30 36

The error propagation equation

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

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

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

But,

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

Therefore:

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

Covariance and independence

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

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

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

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

- We will make this assumption in what follows.

Covariance and independence

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

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

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

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

- We will make this assumption in what follows. $\sigma_{uv} = 0$

Error propagation: sums and differences

- Consider $x = u + const.$

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

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

- Consider $x = u + v$

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

$$= u + v$$

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

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

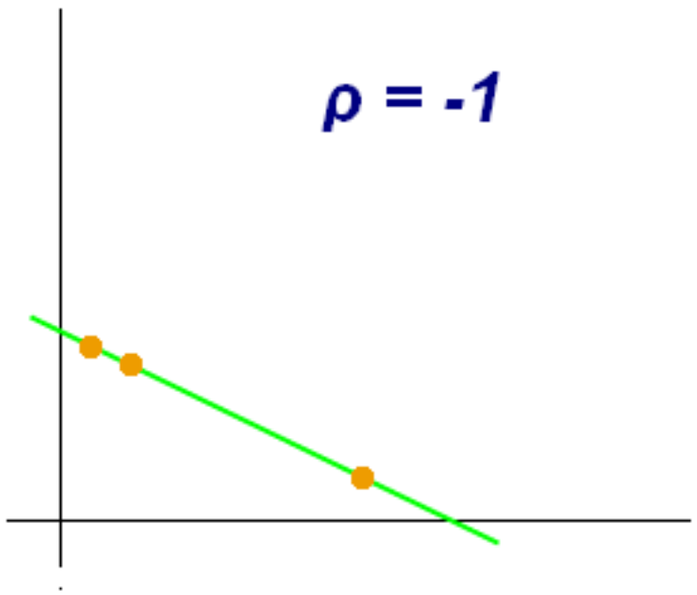
The *quadrature* rule

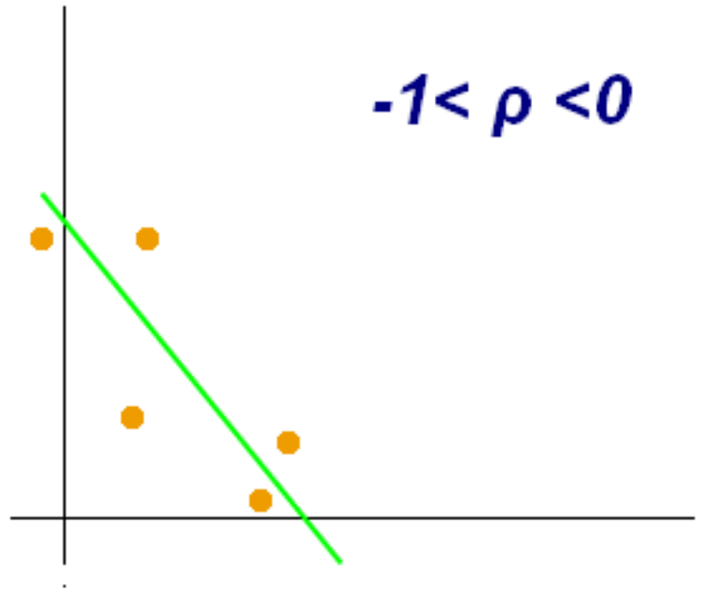
Some useful formulas

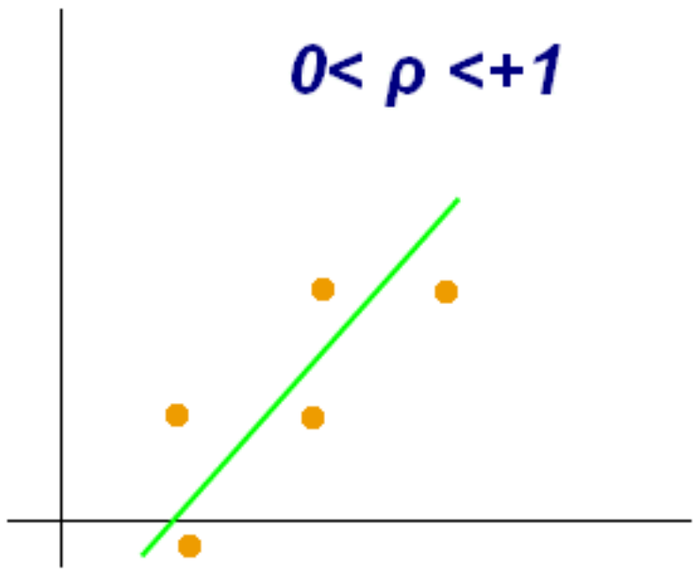
Function	Variance	Standard Deviation
$f = aA$	$\sigma_f^2 = a^2\sigma_A^2$	$\sigma_f = a\sigma_A$
$f = aA + bB$	$\sigma_f^2 = a^2\sigma_A^2 + b^2\sigma_B^2 + 2ab\,\mathrm{cov}_{AB}$	$\sigma_f = \sqrt{a^2\sigma_A^2 + b^2\sigma_B^2 + 2ab\,\mathrm{cov}_{AB}}$
$f = AB$	$\sigma_f^2 \approx f^2\left[\left(\frac{\sigma_A}{A}\right)^2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

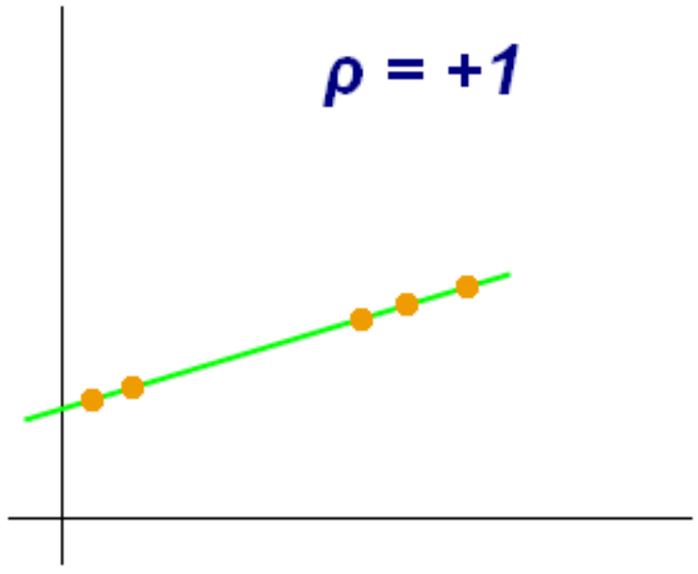
From wikipedia.org

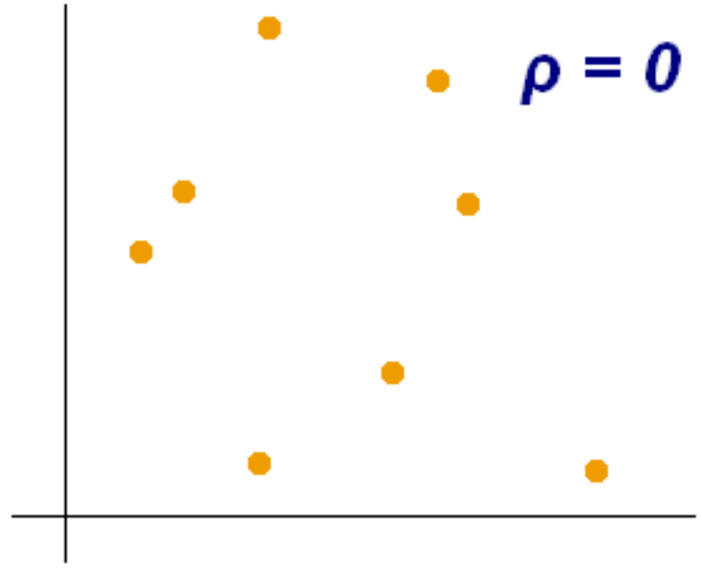
Correlation





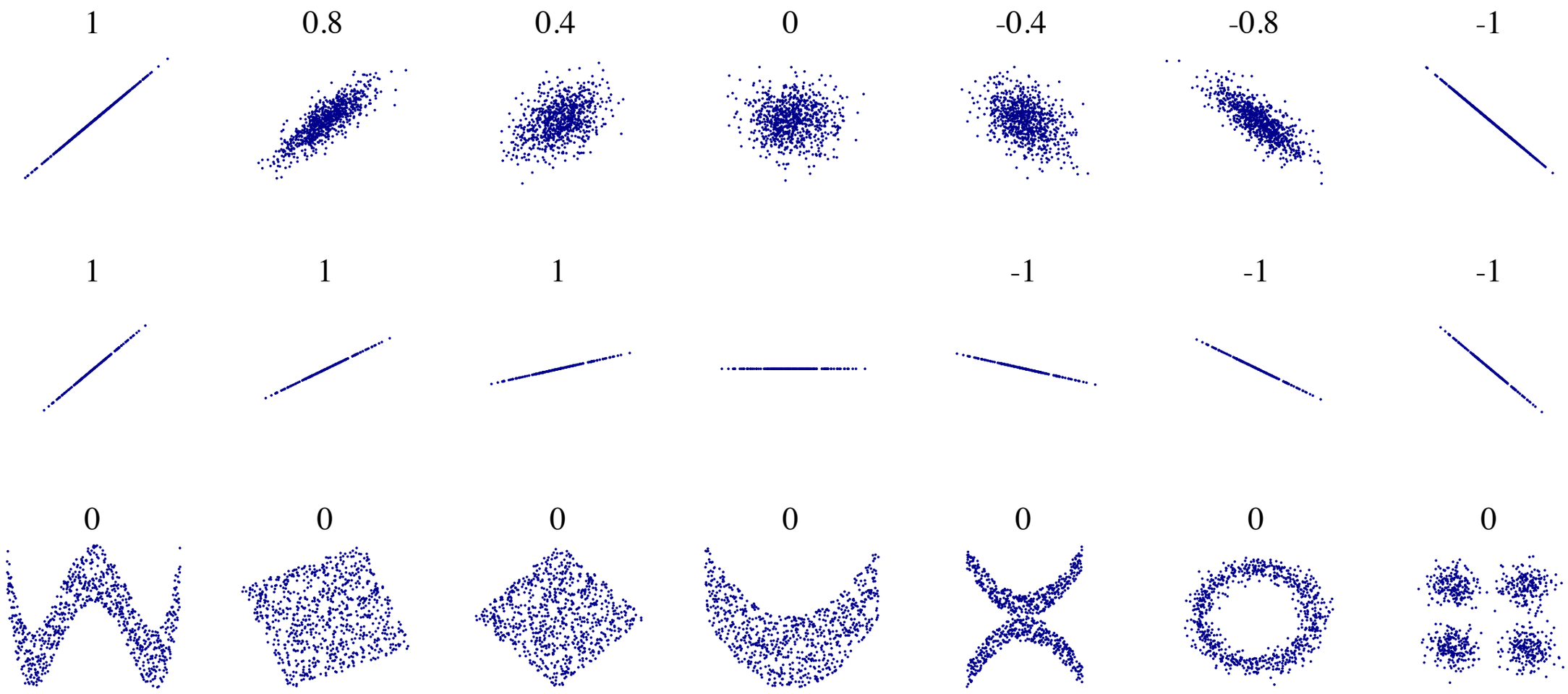






- correlation coefficient (PPMCC), is a measure of the linear dependence (correlation) between two variables X and Y. It has a value between +1 and −1 inclusive, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

Correlation



RECAP

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 hypothe
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

FREQUENTISM

- In this view of the Universe, there is some underlying truth.
- I we the photon flux F from a given star, then measure it again, then again, and so on, each time I 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.

FREQUENTISM

- 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. $P(\boldsymbol{m}^{\text{obs}}|a,b,H)$, 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

THE BAYESIAN VIEW

- 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*

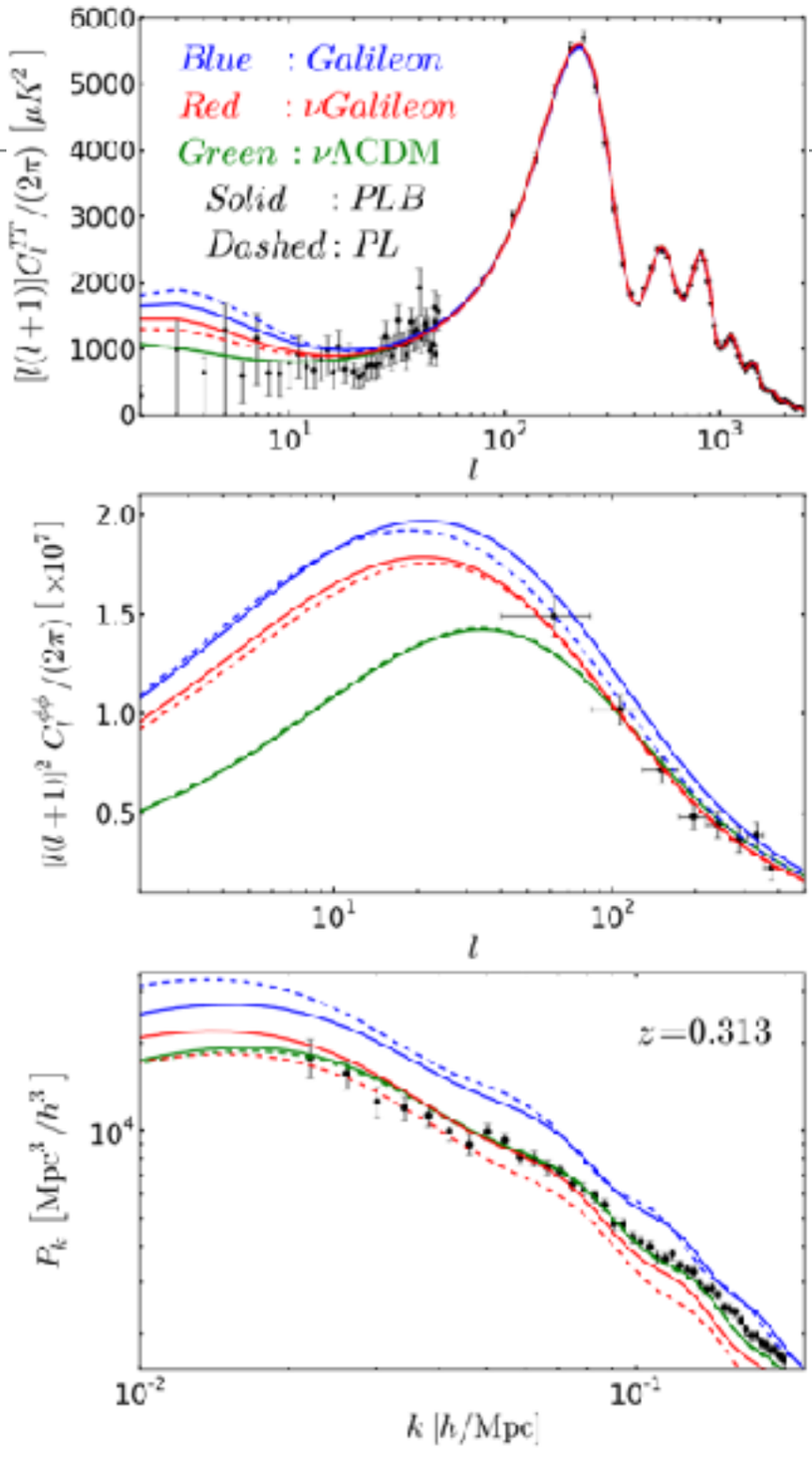
THIS IS KINDA REVOLUTIONARY

- 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 ΛCDM 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 ΛCDM 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

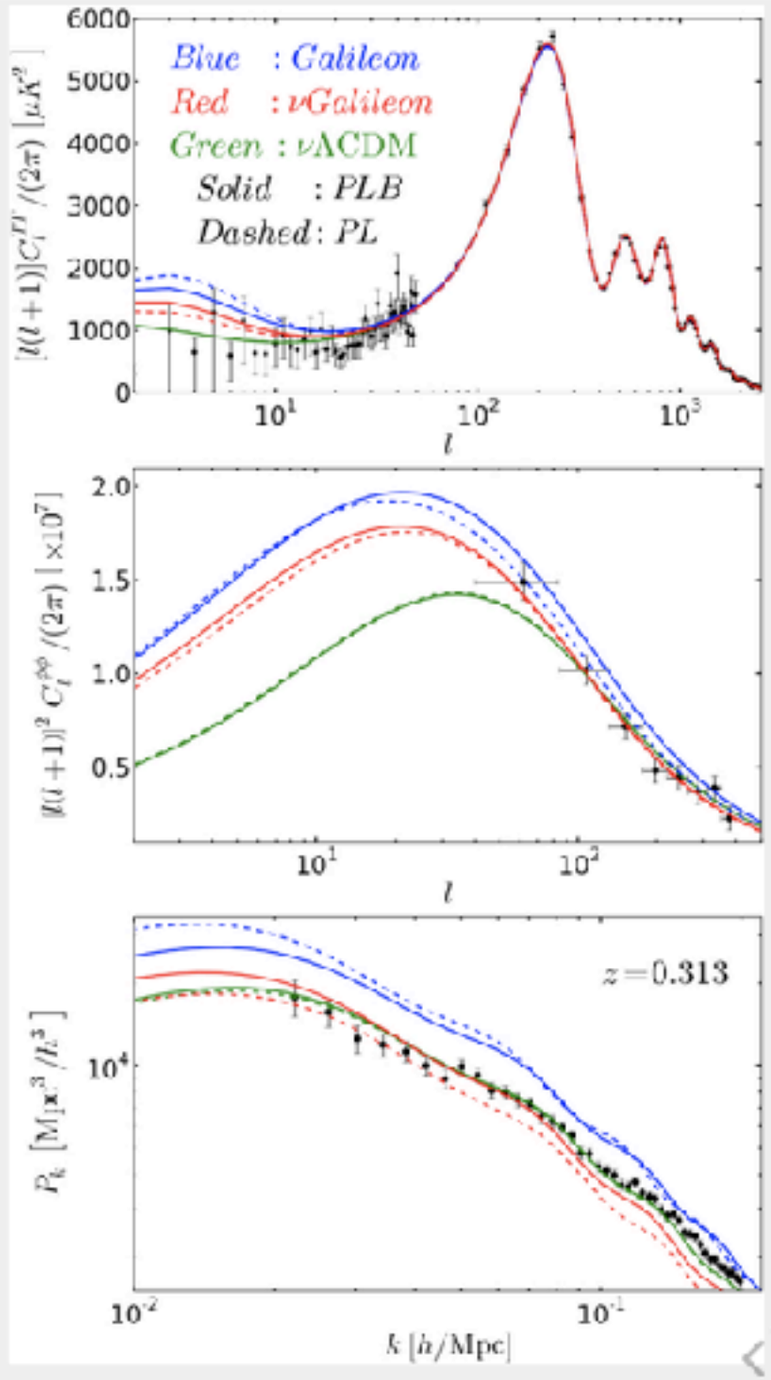


TESTING IF A MODEL IS CONSISTENT WITH DATA IN THE FREQUENTIST WORLD

- 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.

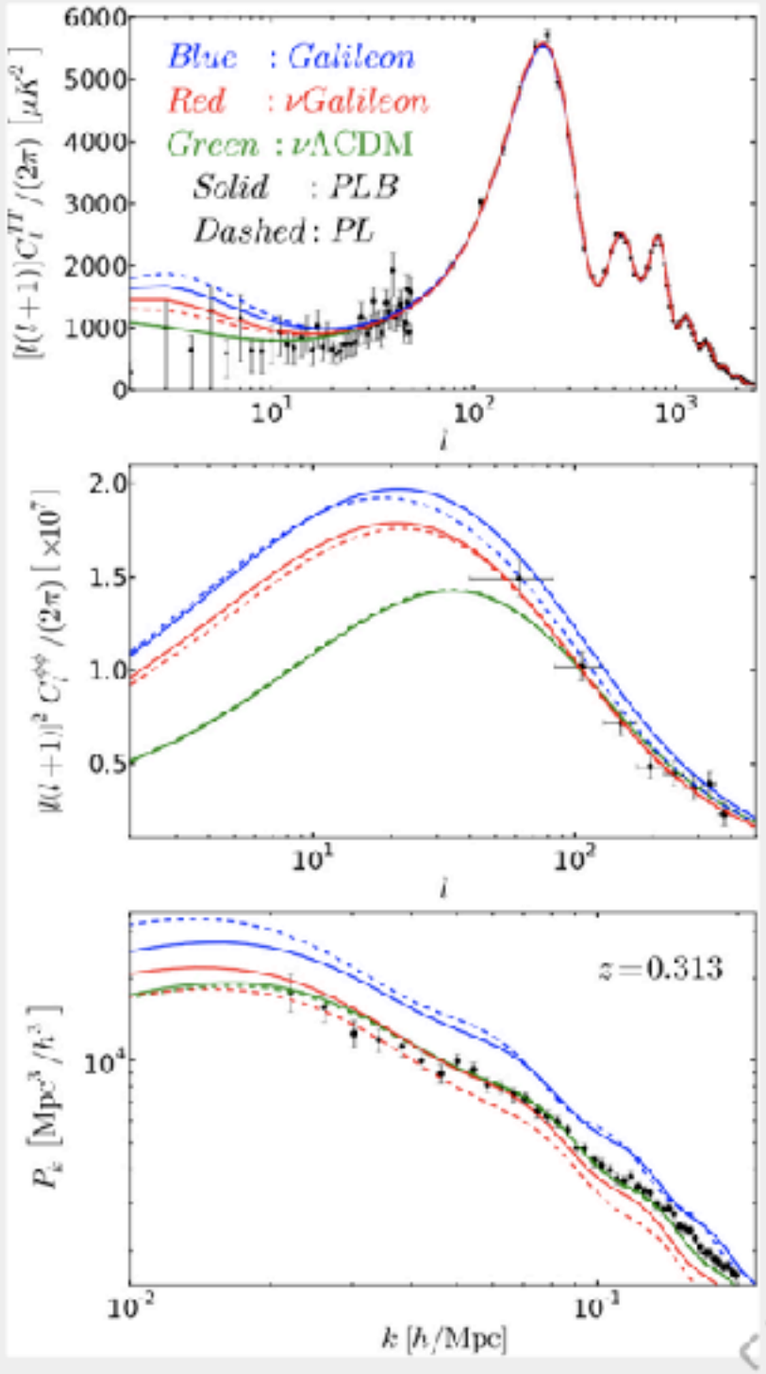
uncertainties: 1-σ

3 points out of 10 can be outside of the uncertainties

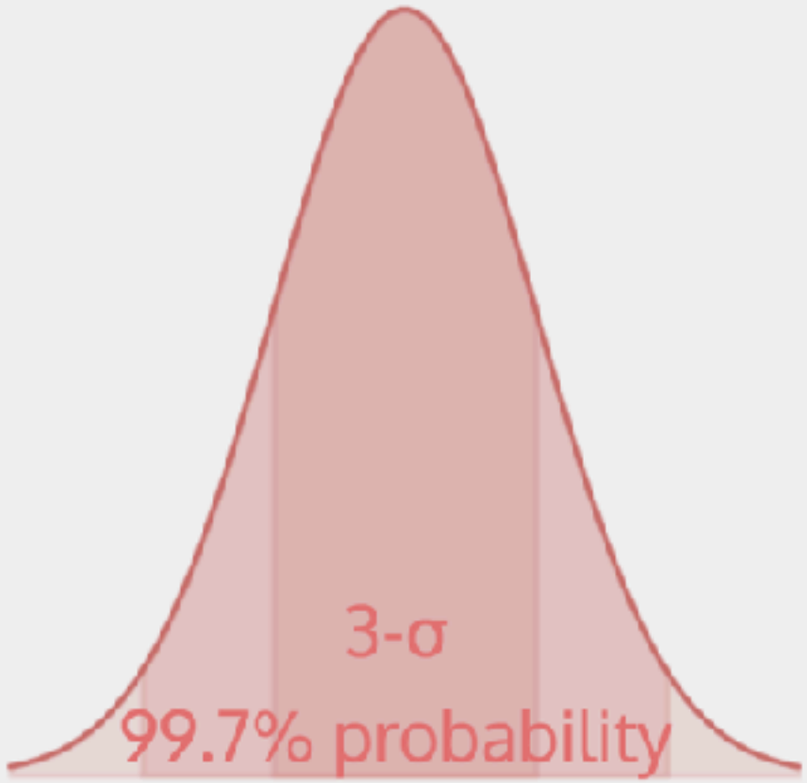


uncertainties: 1-σ

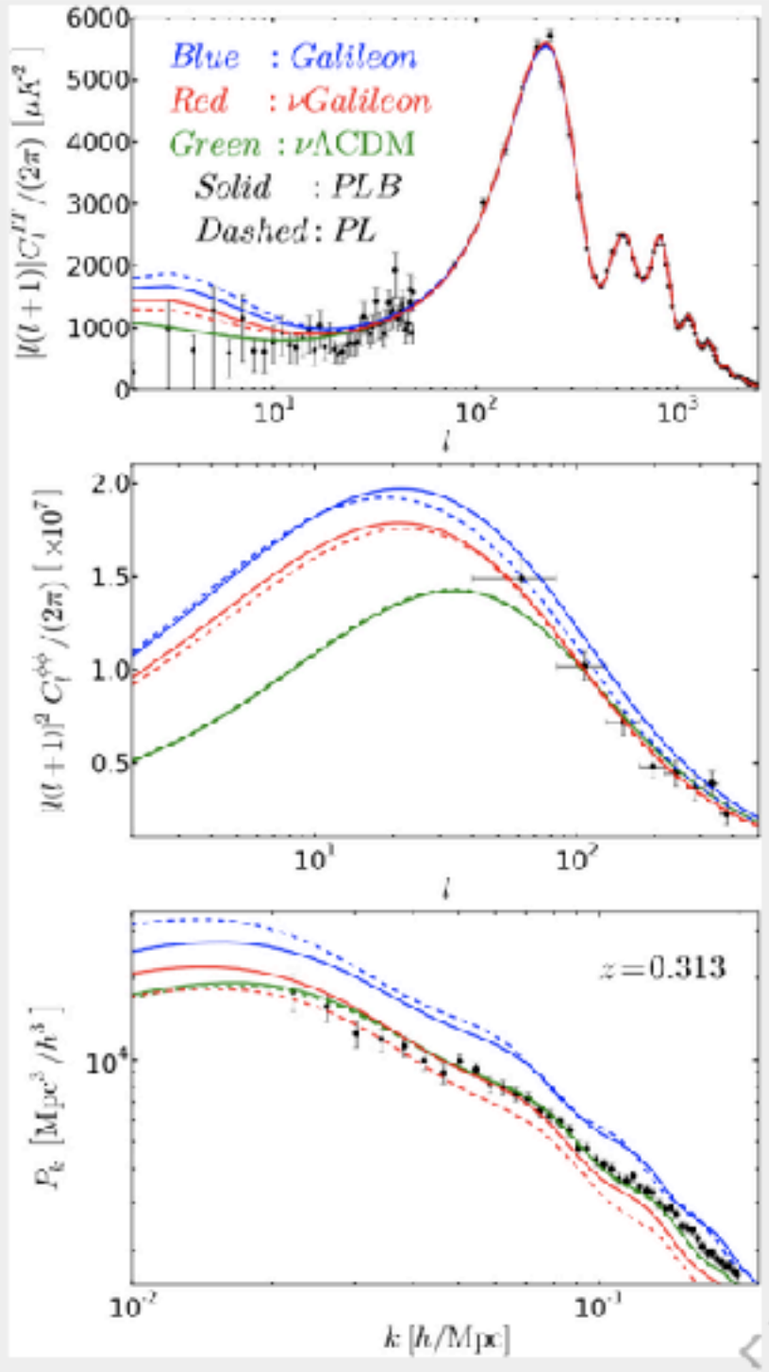
5 points out of 100 can be outside of the uncertainties



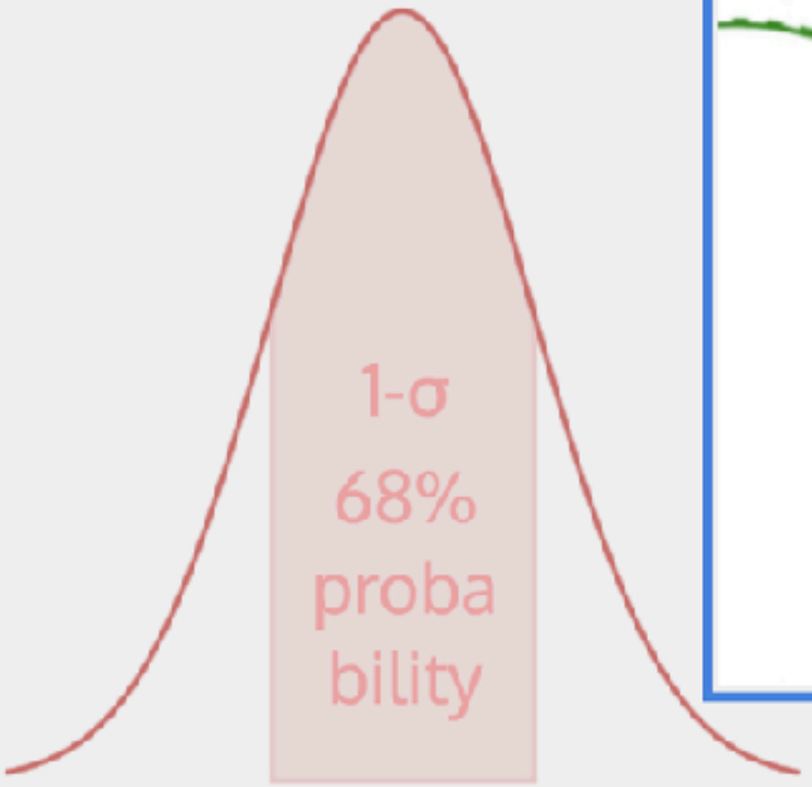
uncertainties: 1-σ



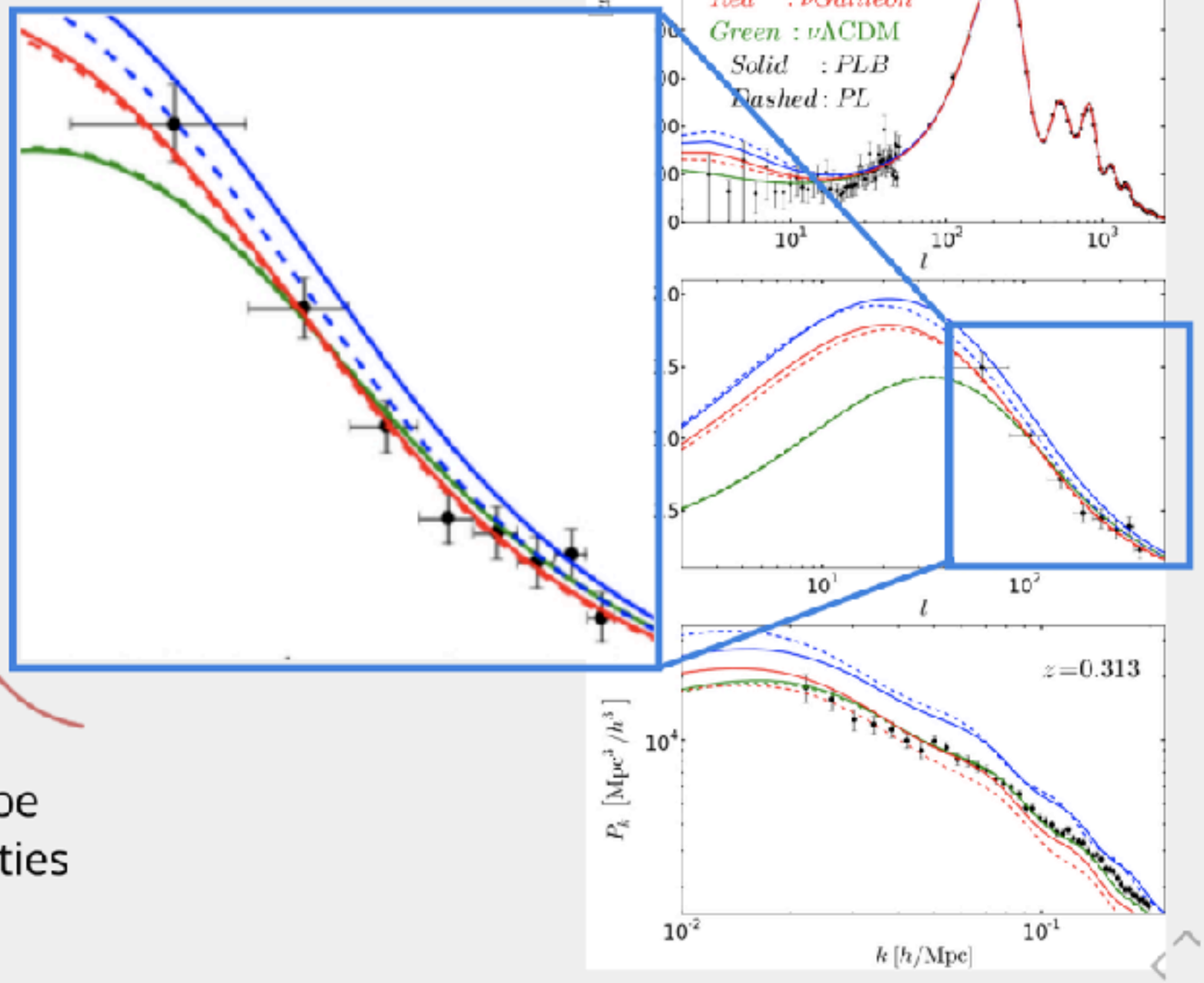
3 points out of 1000 can be outside of the uncertainties



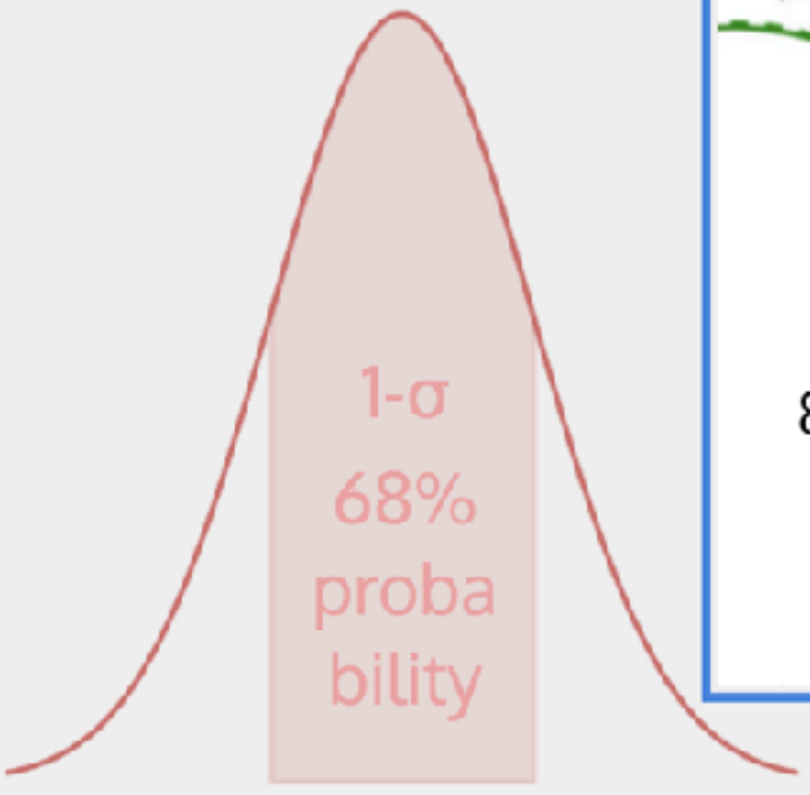
uncertainties: 1-σ



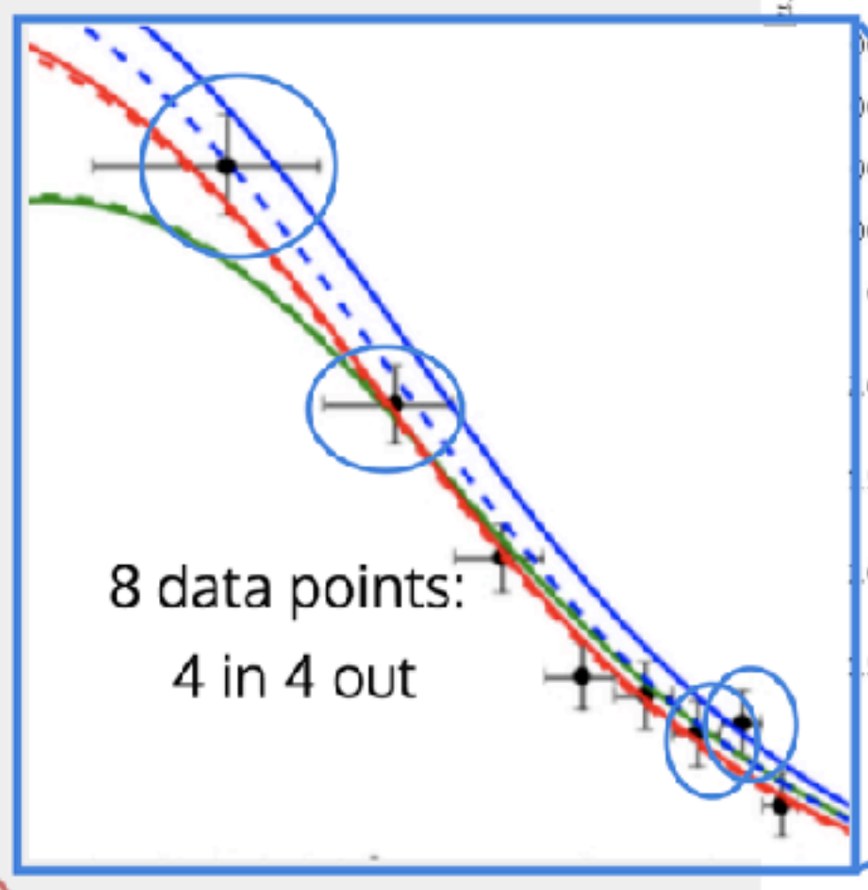
3 points out of 10 can be outside of the uncertainties

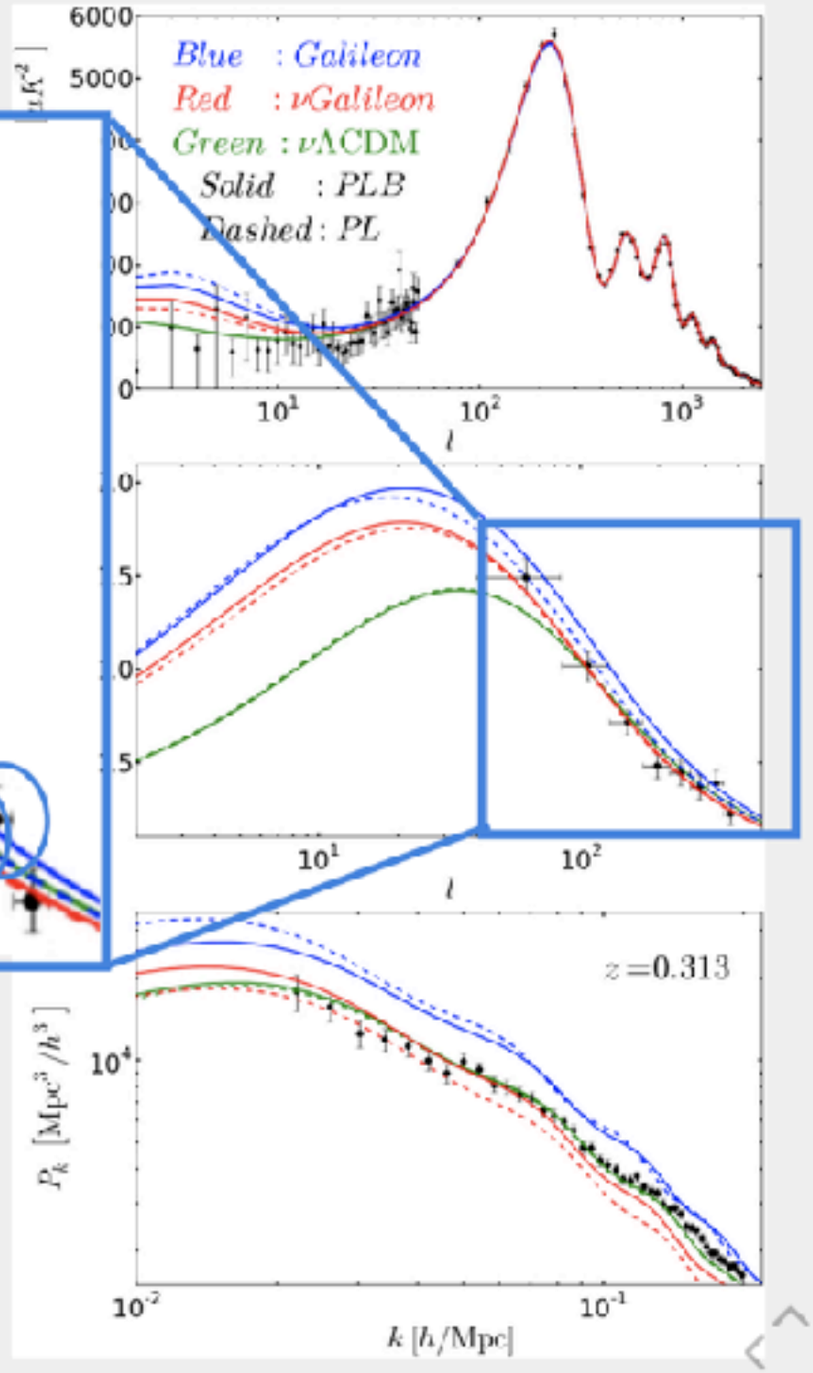


uncertainties: 1-σ

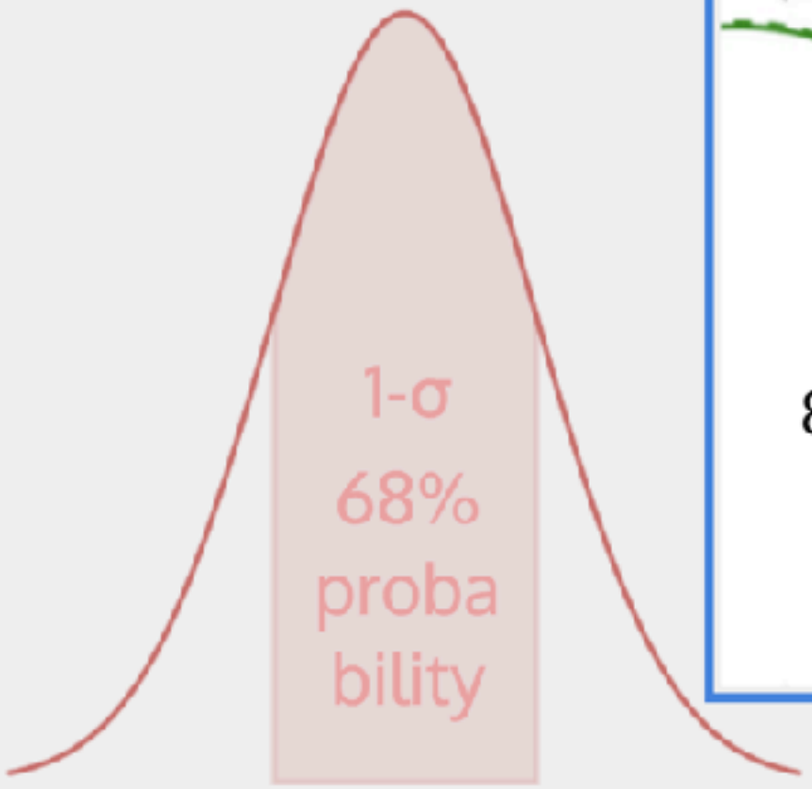


7 points out of 10 should be inside the errorbar

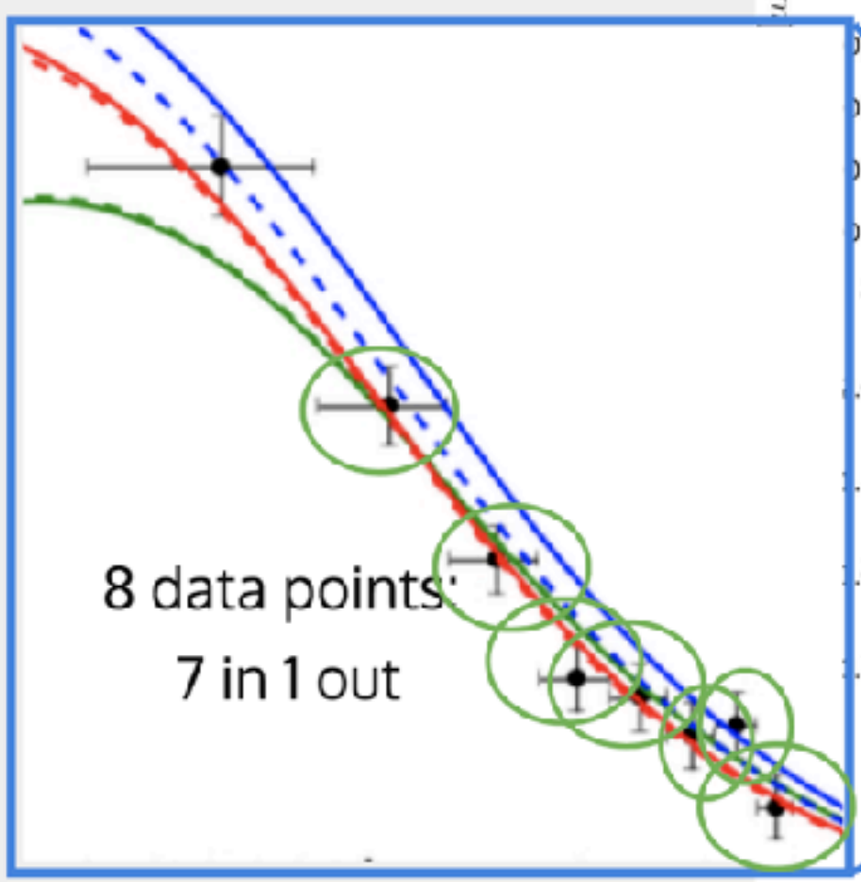


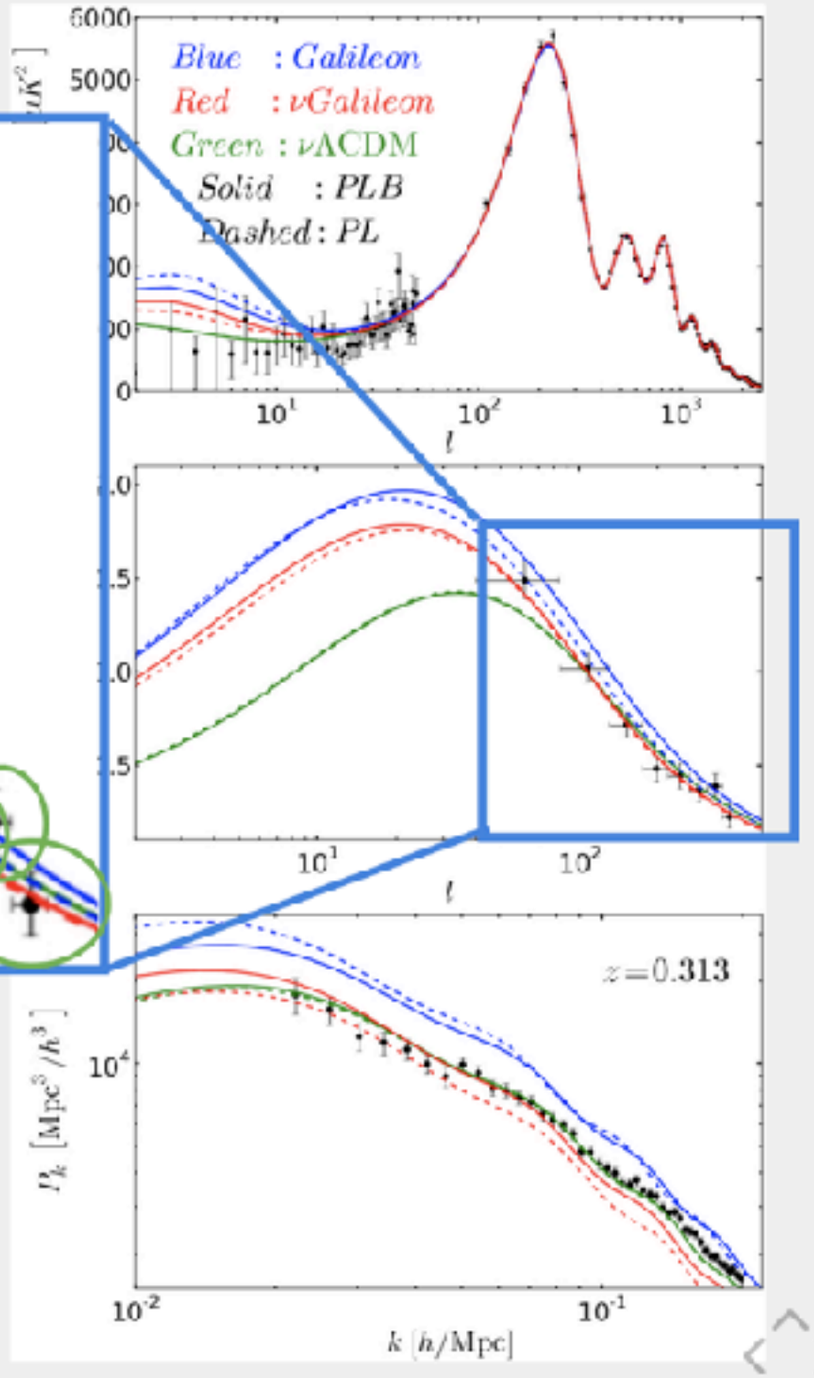


uncertainties: 1-σ

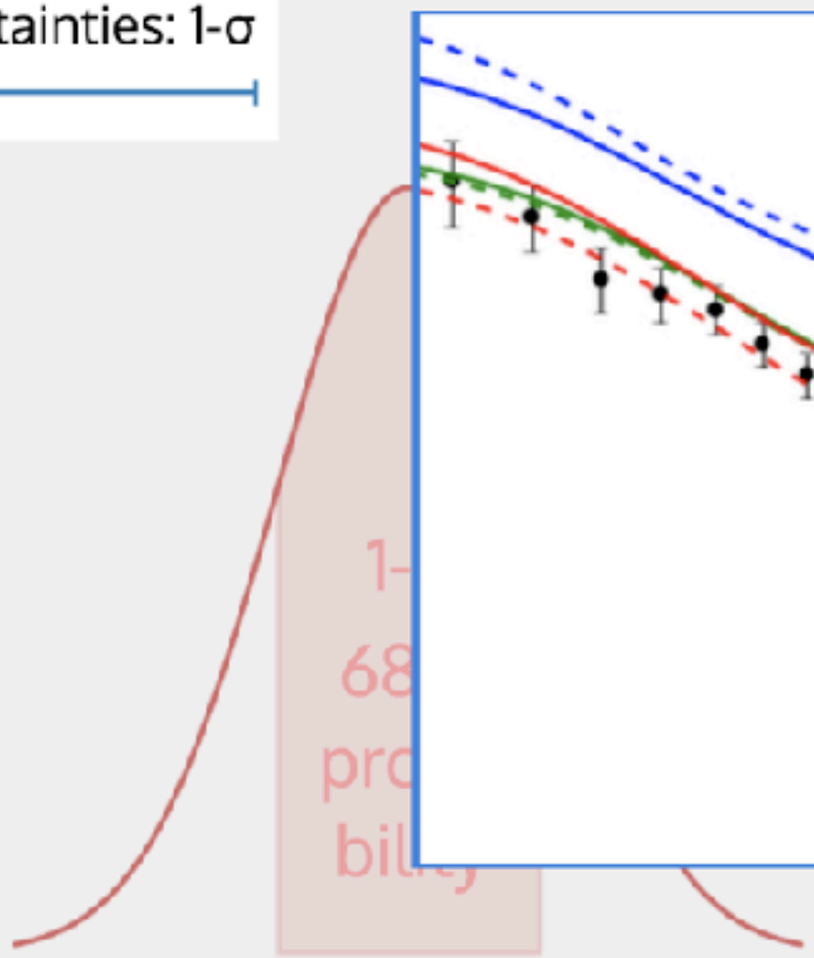


7 points out of 10 should be inside the errorbar

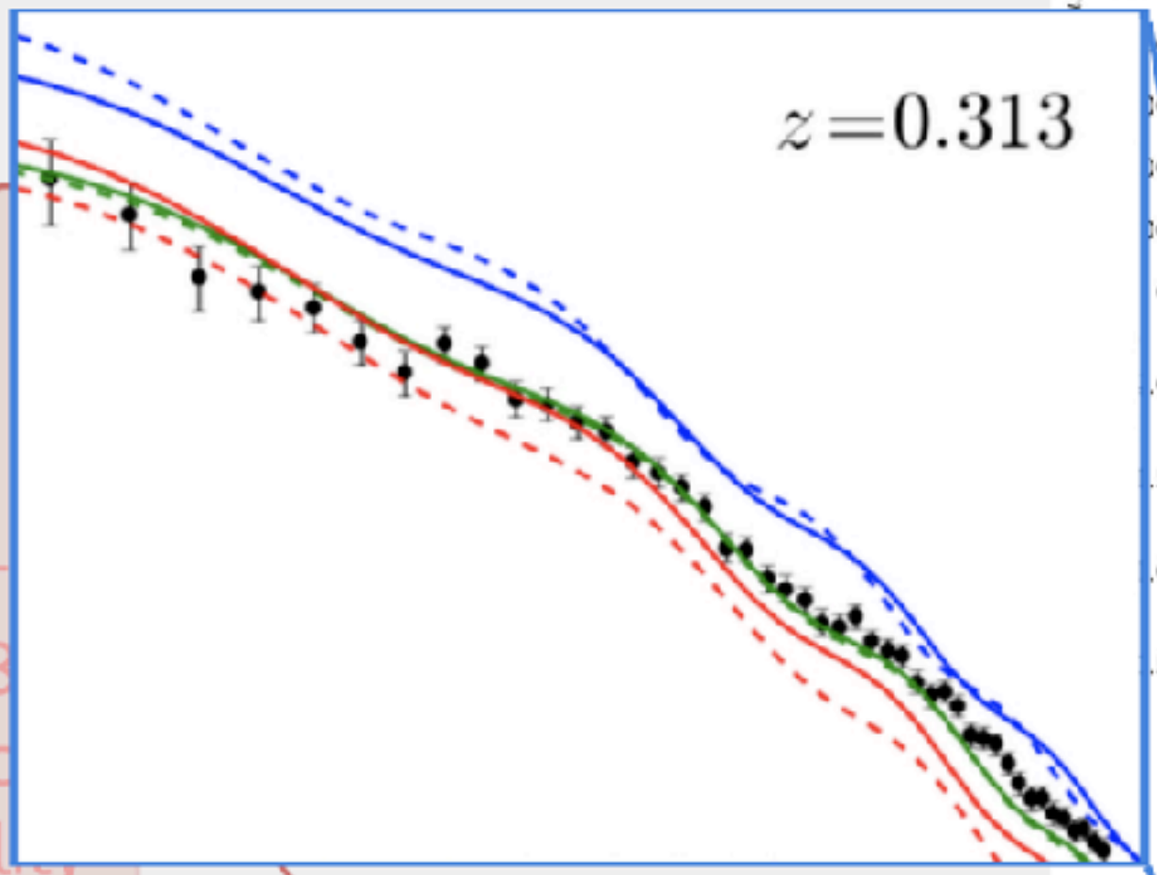


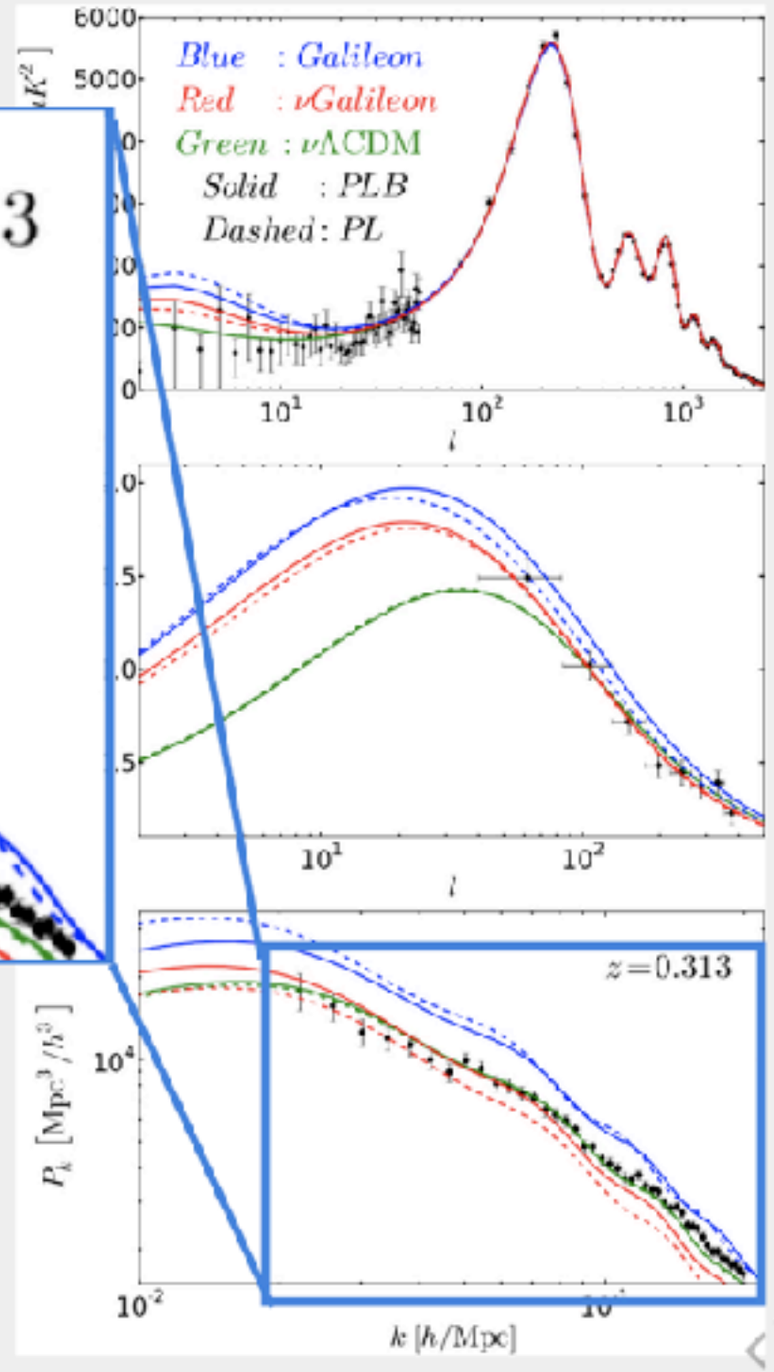


uncertainties: 1-σ



7 points out of 10 should be inside the errorbar





SO YOUR PROBLEM THEN IS HOW TO EVALUATE THE POSTERIOR

Posterior
How probable is the hypothesis given the data we observed

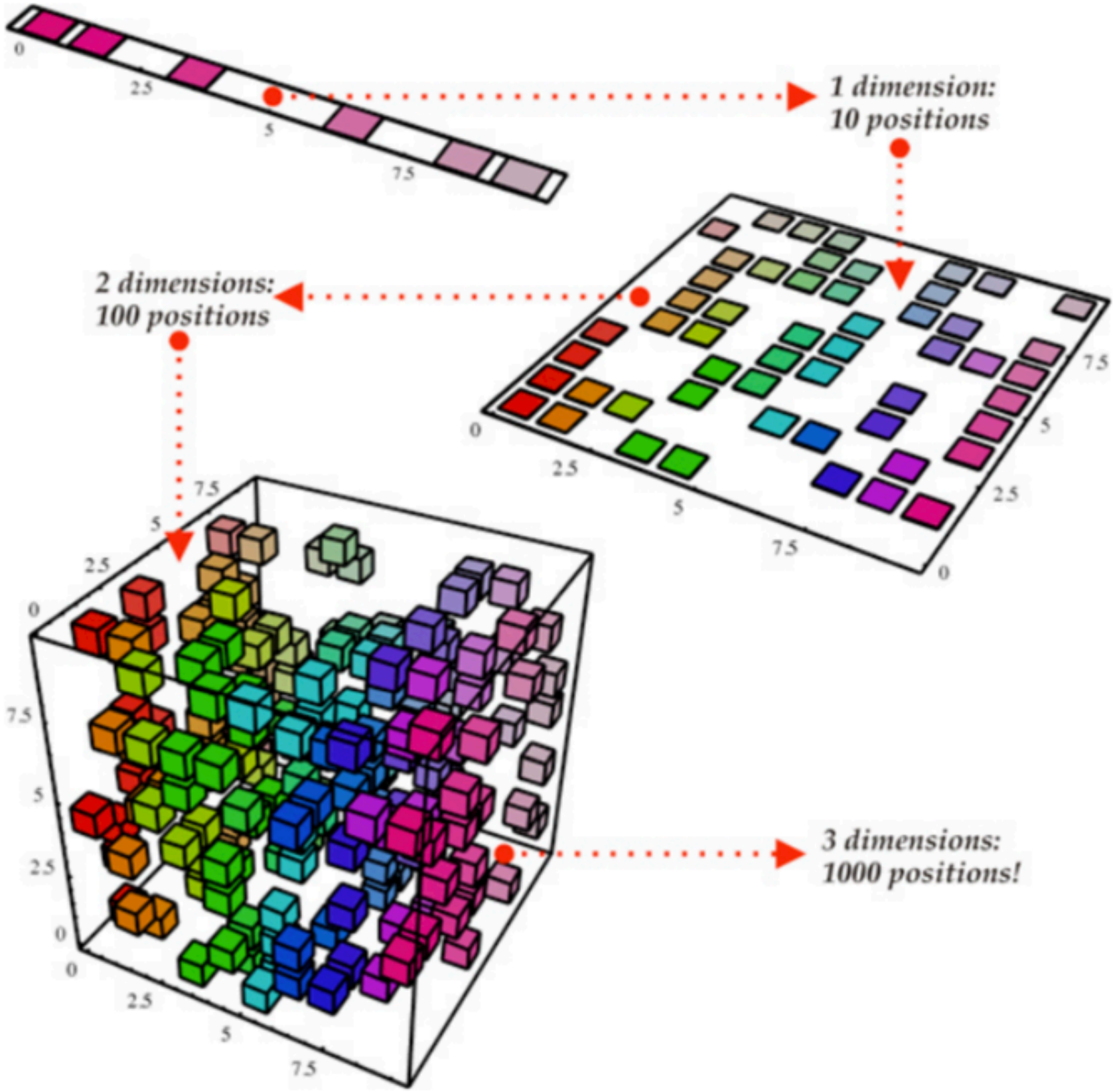
Likelihood
How probable is the data given the hypothesis is true

Prior
How probable was the hypothe
before we observed anything

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

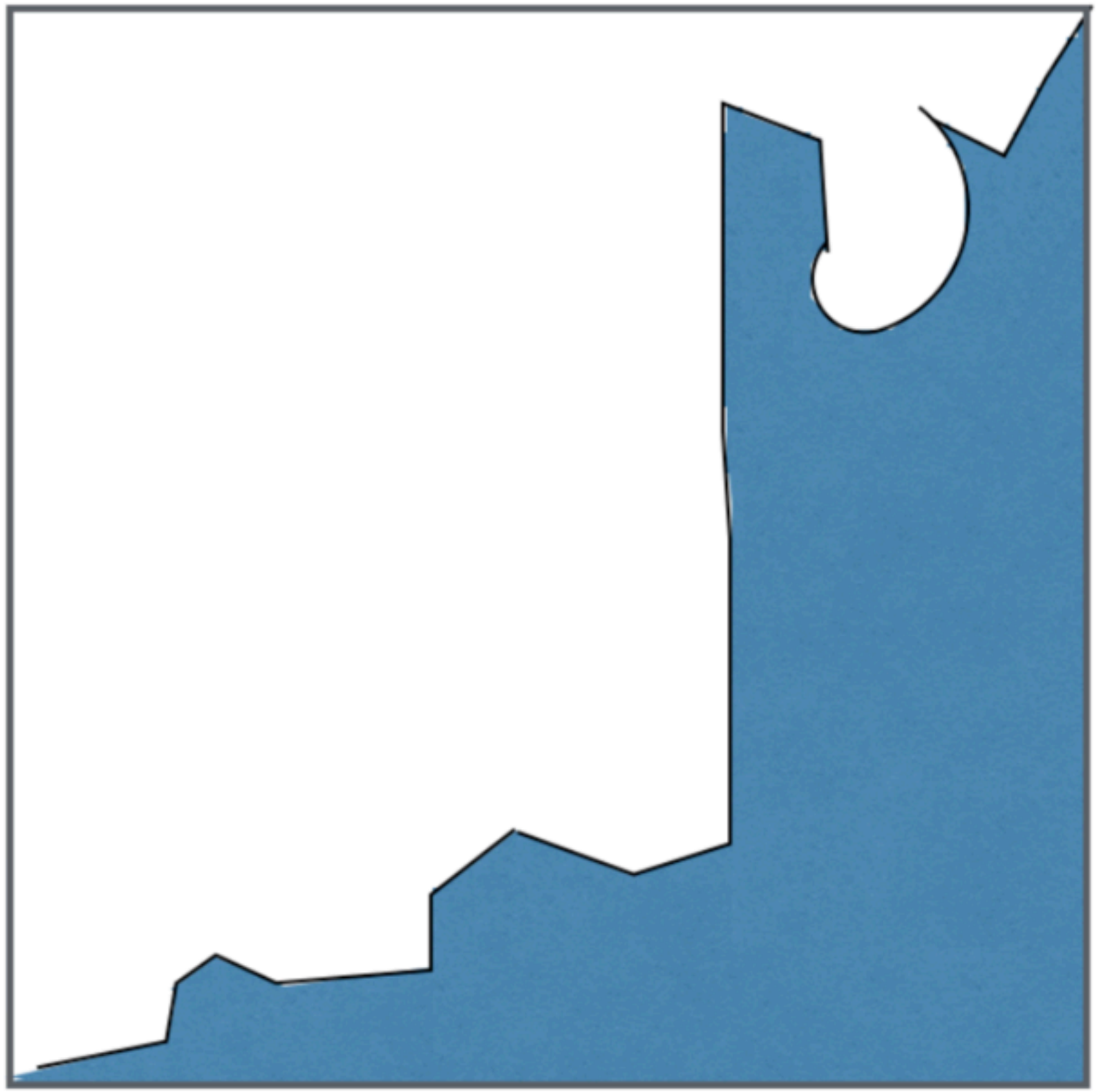
Evidence
How probable is the data over all possible hypotheses

Curse of Dimensionality

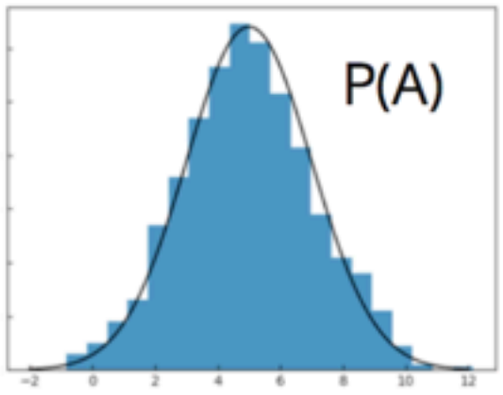


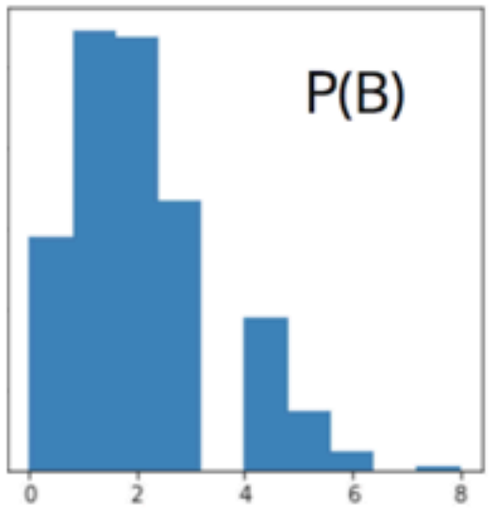
Simple Example

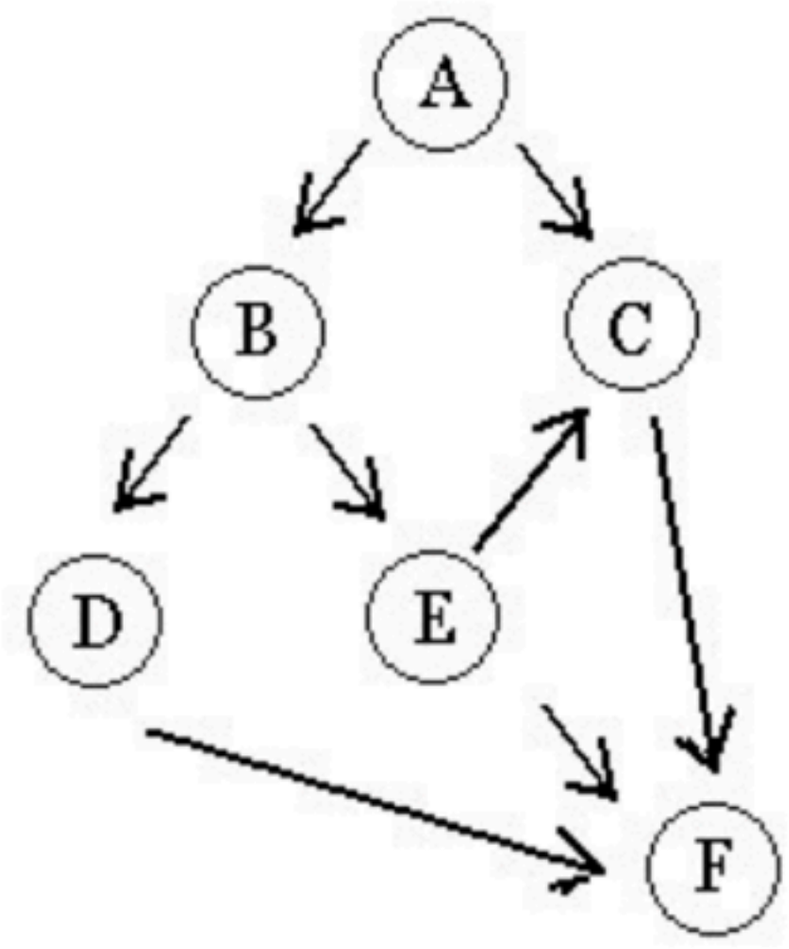
Area of a rectangle = Base × Height Area of a triangle = 1/2 Base × Height What about the area of this?



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







A ~ P(A)

B ~ P(B|A)

C ~ P(C|A,E)

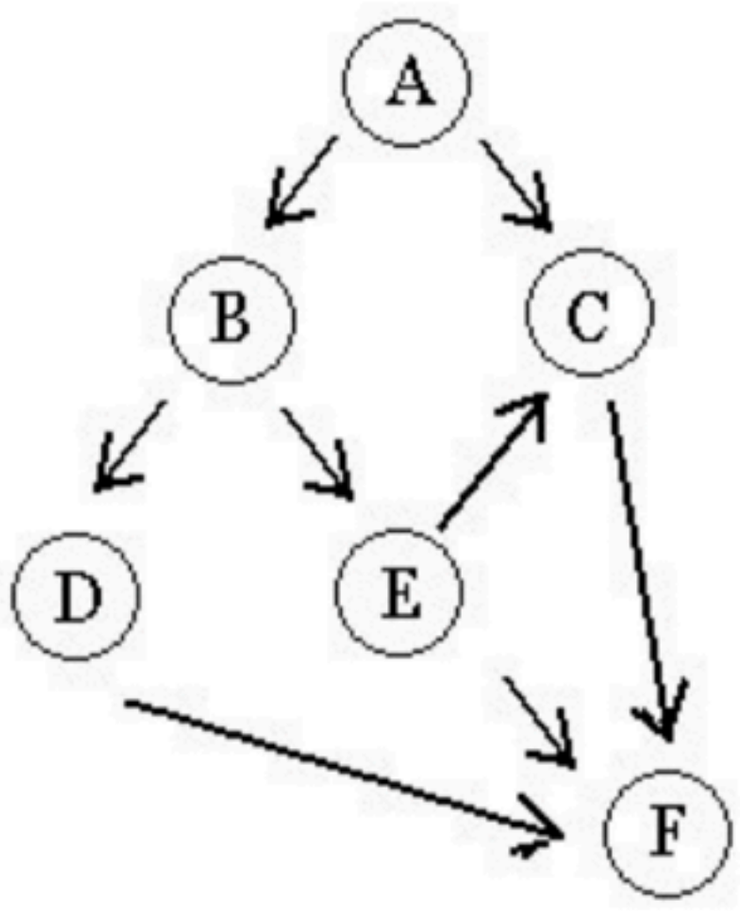
D ~ P(D|B)

E ~ P(E|B)

F ~ 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 ~ P(A)

B ~ P(B|A)

C ~ P(C|A,E)

D ~ P(D|B)

E ~ P(E|B)

F ~ P(F|C,D,E)

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

A person's depth

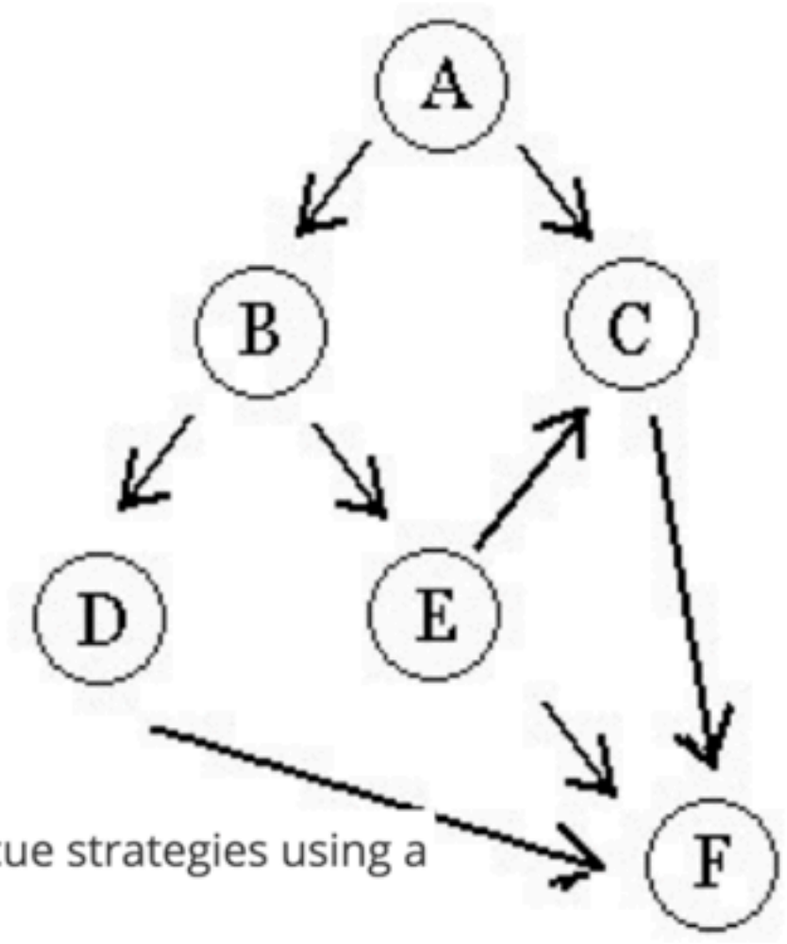
B prob to find them at time t

C they survive the avalange

D they are still alive at t

E can be resuscitated at time t

F person survives



$A \sim P(A)$

$B \sim P(B|A)$

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

$D \sim P(D|B)$

$E \sim P(E|B)$

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

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:

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

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….

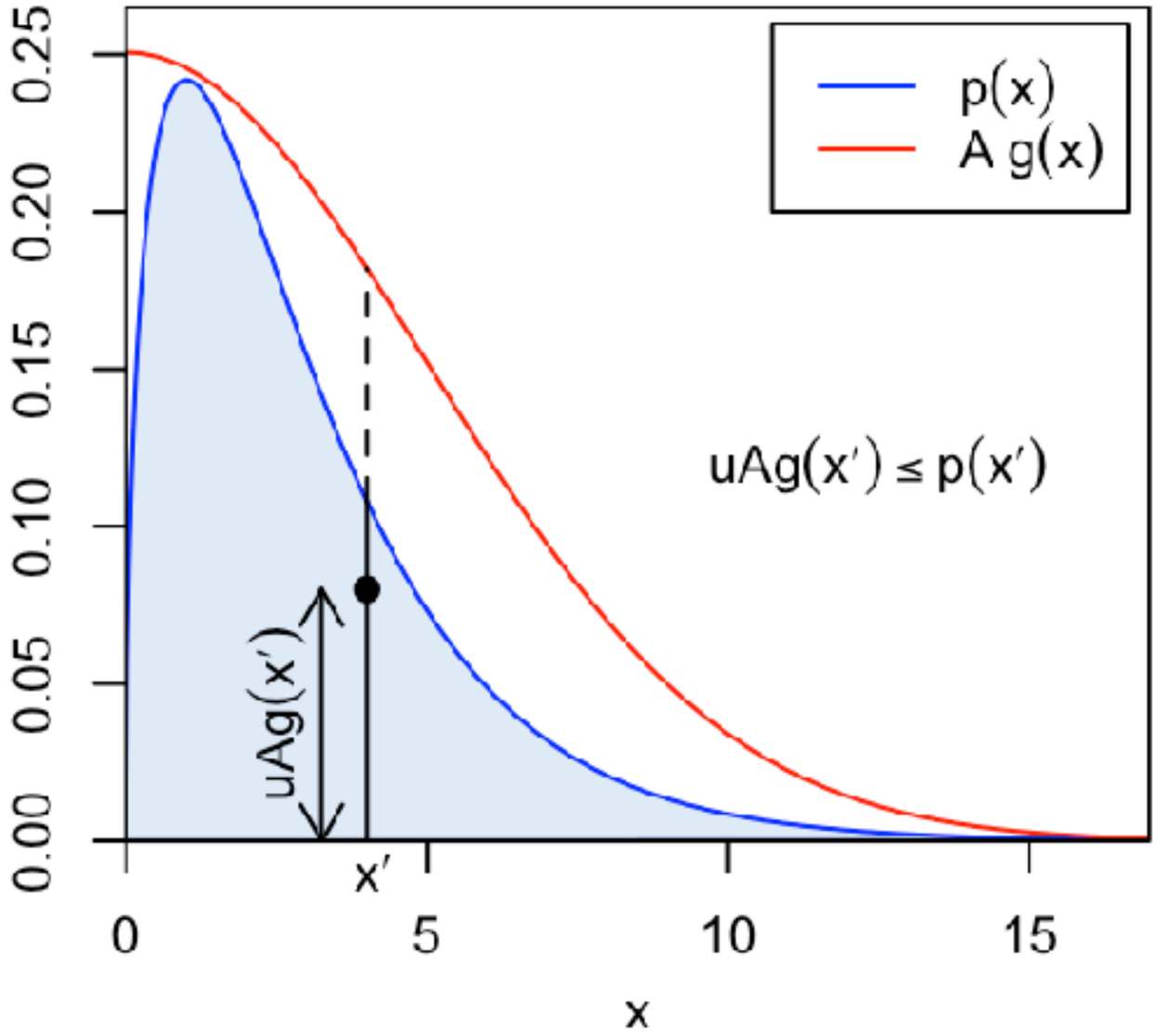
- We've come up with a model, an objective/loss/likelihood function and some priors
  - Now we actually want to evaluate the posterior P(θ|D)
  - analytically is too hard, so we resort to numerical techniques
  - Option 1: Evaluate the function on some grid of parameter values - doesn't scale
  - Option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

REJECTION SAMPLING

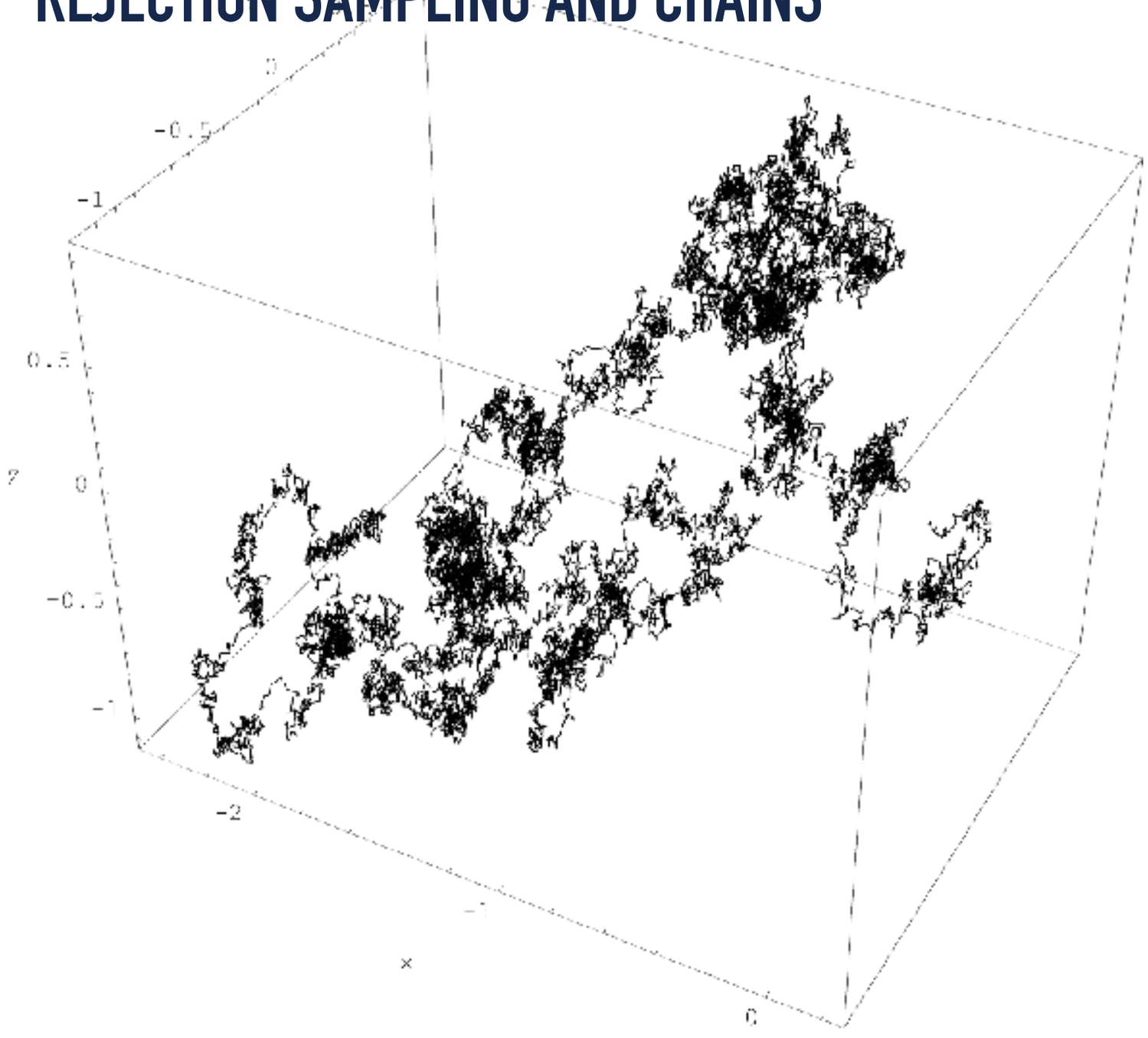
- Define an envelope function which everywhere exceeds the target PDF, $p(x)$, and can be sampled. Let this be $A.g(x)$ where A is a scaling factor and $g(x)$ is a PDF we know.

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** <= **p(x)**/**A***g(x), keep the sample x
  - otherwise, reject x



REJECTION SAMPLING AND CHAINS



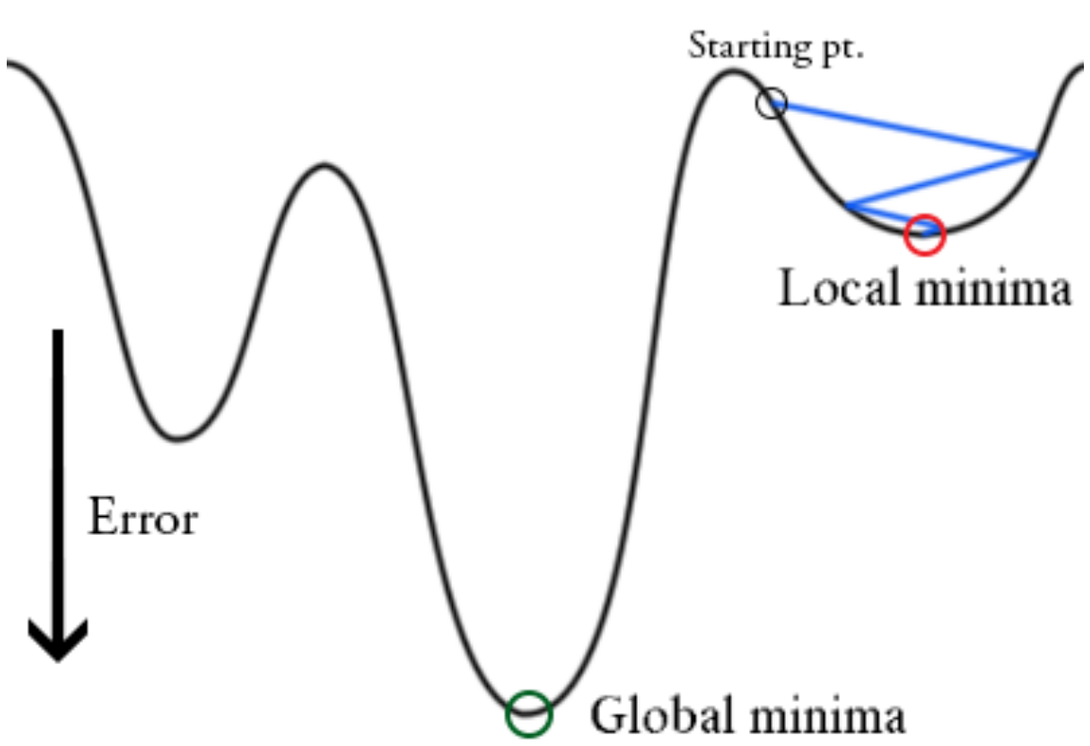
- Start sampling in your parameter space following a prior distribution at ***x***
- Compute the likelihood at this location ***p(x)***
- Move to a new location ***x′***
- Compute the likelihood at the new location ***p(x′)***
- If it's higher at the new location, keep the new sample ***x′***
- If it's 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

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(θ|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)**

MARKOV CHAIN PROPERTIES

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

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

X_{n+1} depends only on X_n

(and not on $X_1, X_2, \ldots, X_{n-1}$)

MARKOV CHAIN PROPERTIES

- **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**

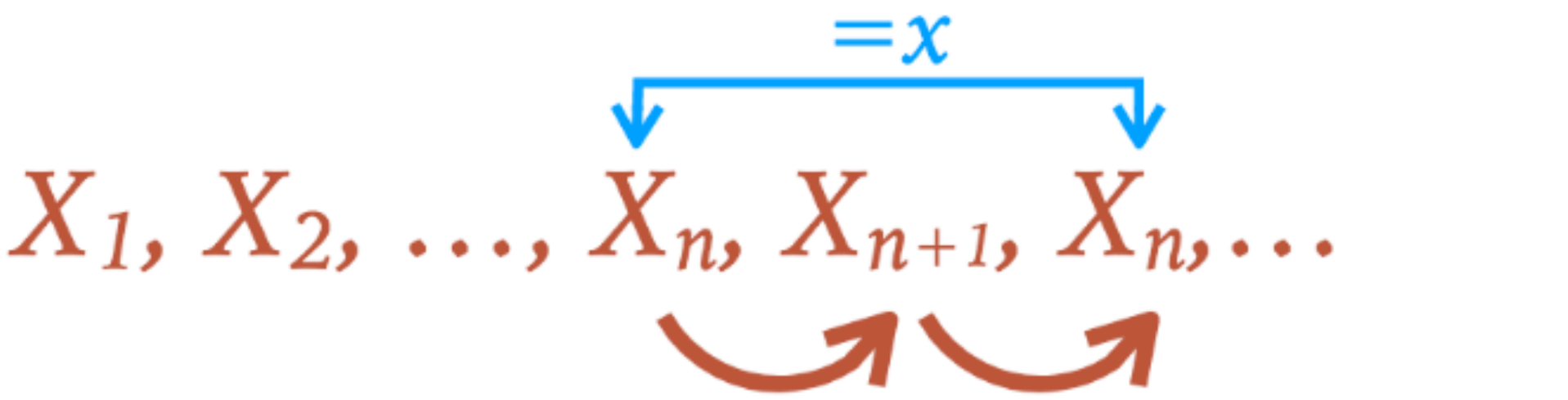
$$X_1, X_2, \ldots, X_n, X_{n+1}, \ldots, X_m, X_{m+1}, \ldots$$

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

~time invariant

MARKOV CHAIN PROPERTIES

- **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**



DETAILED BALANCE

- **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

BUILDING A CHAIN OF SAMPLES

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

$$p(x') = \int dx\, p(x)\, T(x' \mid x)$$

- where $T(x'|x)$ is the transition probability of a step from x to x'.

- **We want to have detailed balance**

$$p(x)T(x' \mid x) = p(x')T(x \mid x')$$

REARRANGING

- We'll break the transition **T(x'|x)** into two steps:
- A proposal, **g(x'| x)** and
- An acceptance ratio, **A(x'|x)**
- i.e.

$$T(x' \mid x) = A(x' \mid x)g(x' \mid x)$$

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

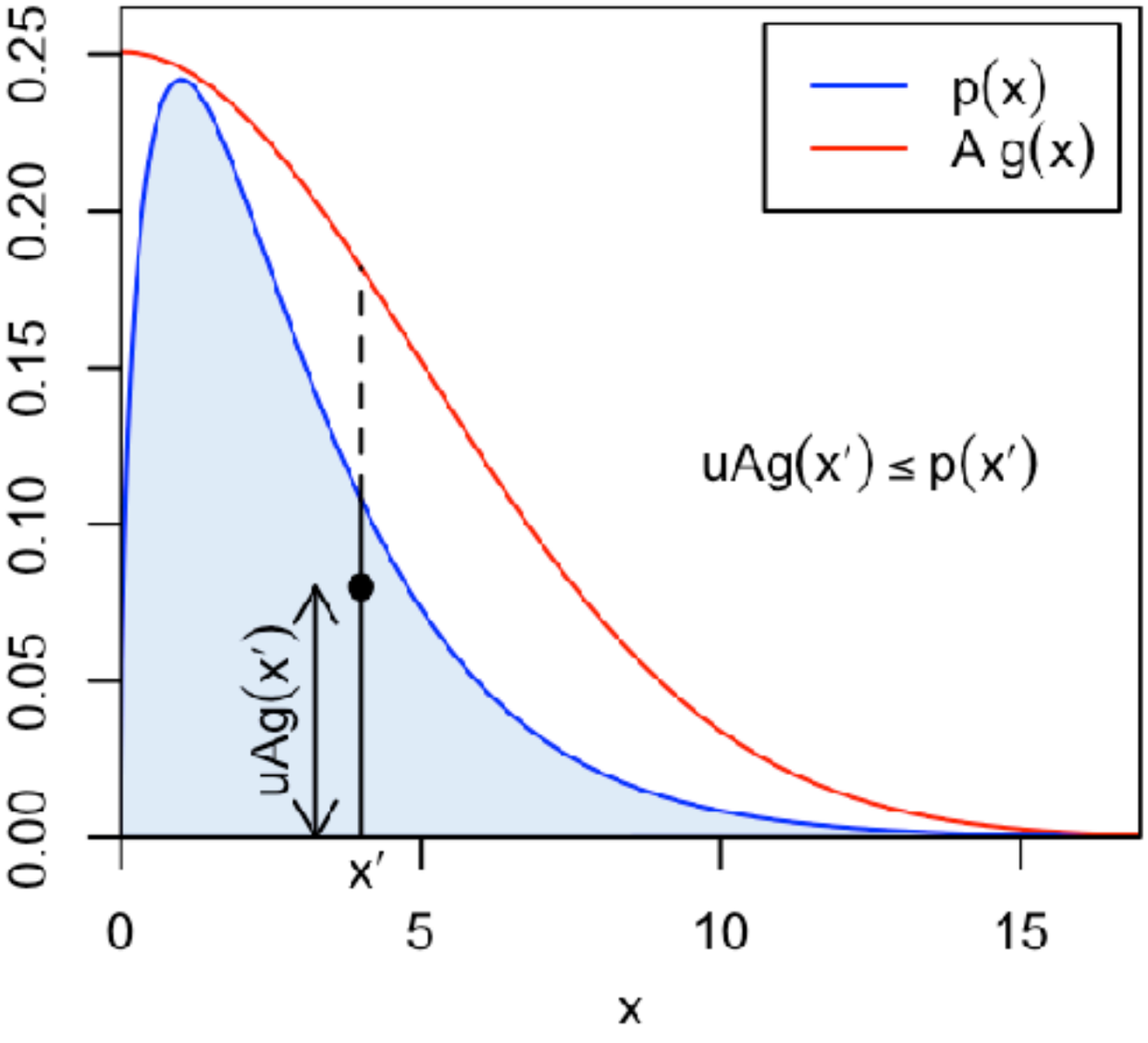
HEY THIS IS JUST REJECTION SAMPLING

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

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

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

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



HEY THIS IS JUST REJECTION SAMPLING

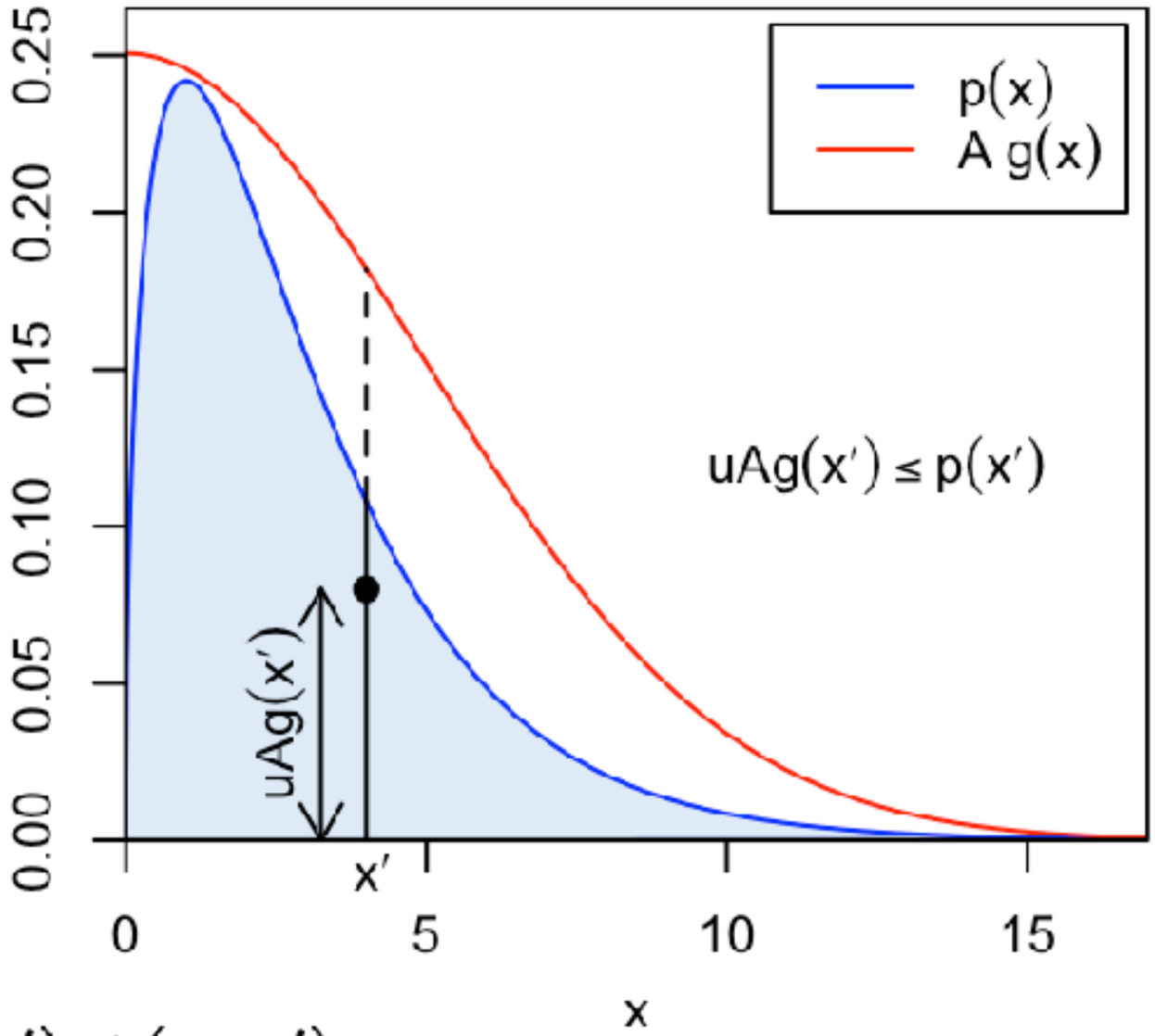
So if we identify:

$p(x)$ 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**



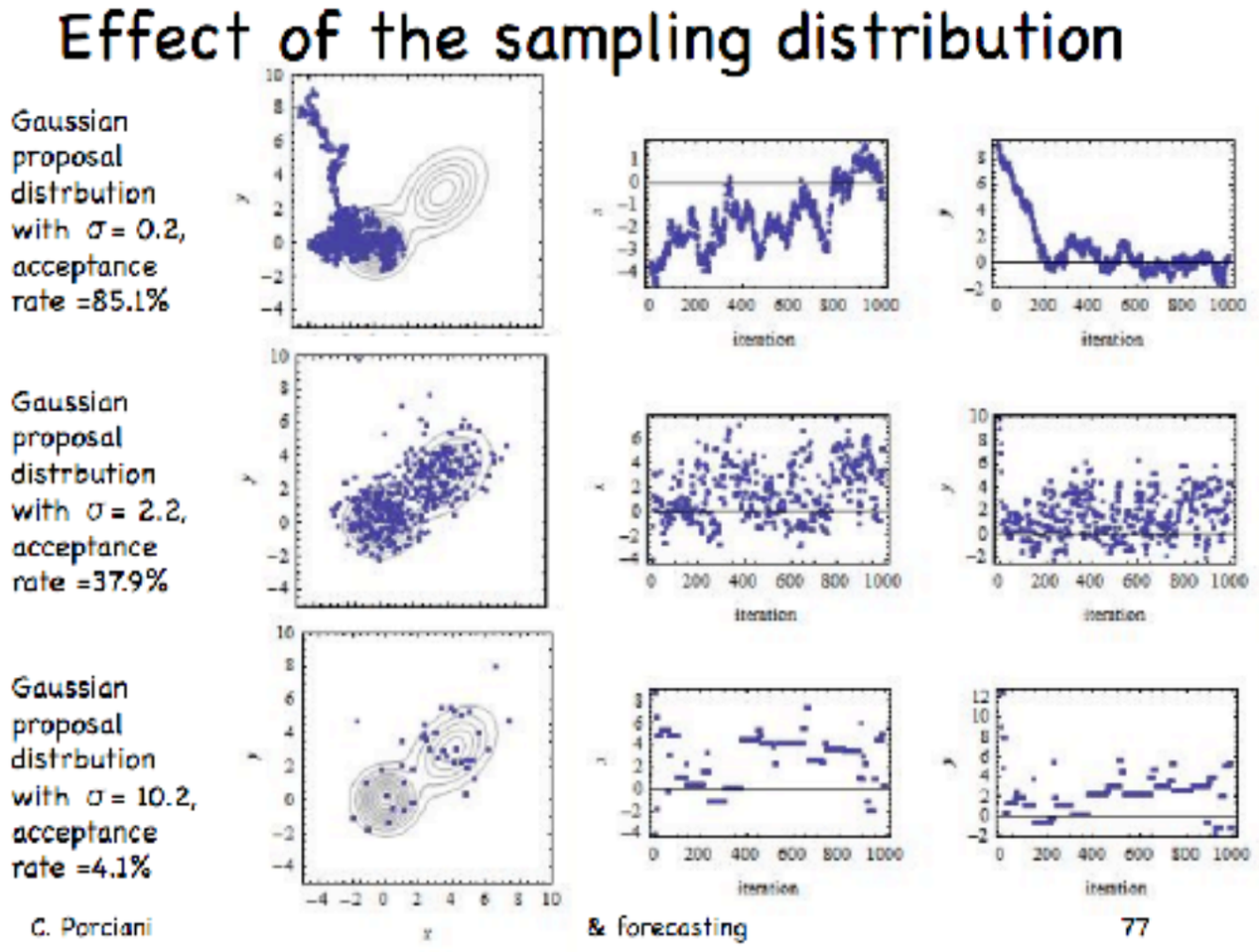
$$p(x)g(x' \mid x)A(x',x) \equiv p(x')g(x \mid x')A(x,x')$$

DETAILED BALANCE GIVES YOU REVERSIBLE MARKOV CHAINS

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

IMPORTANT:

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



MCMC PROPERTIES

- 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:

- A) How do you choose **g(x, 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 it's step size depend on a "temperature" of the system that cools - initially starts in high temperature - bad moved 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)

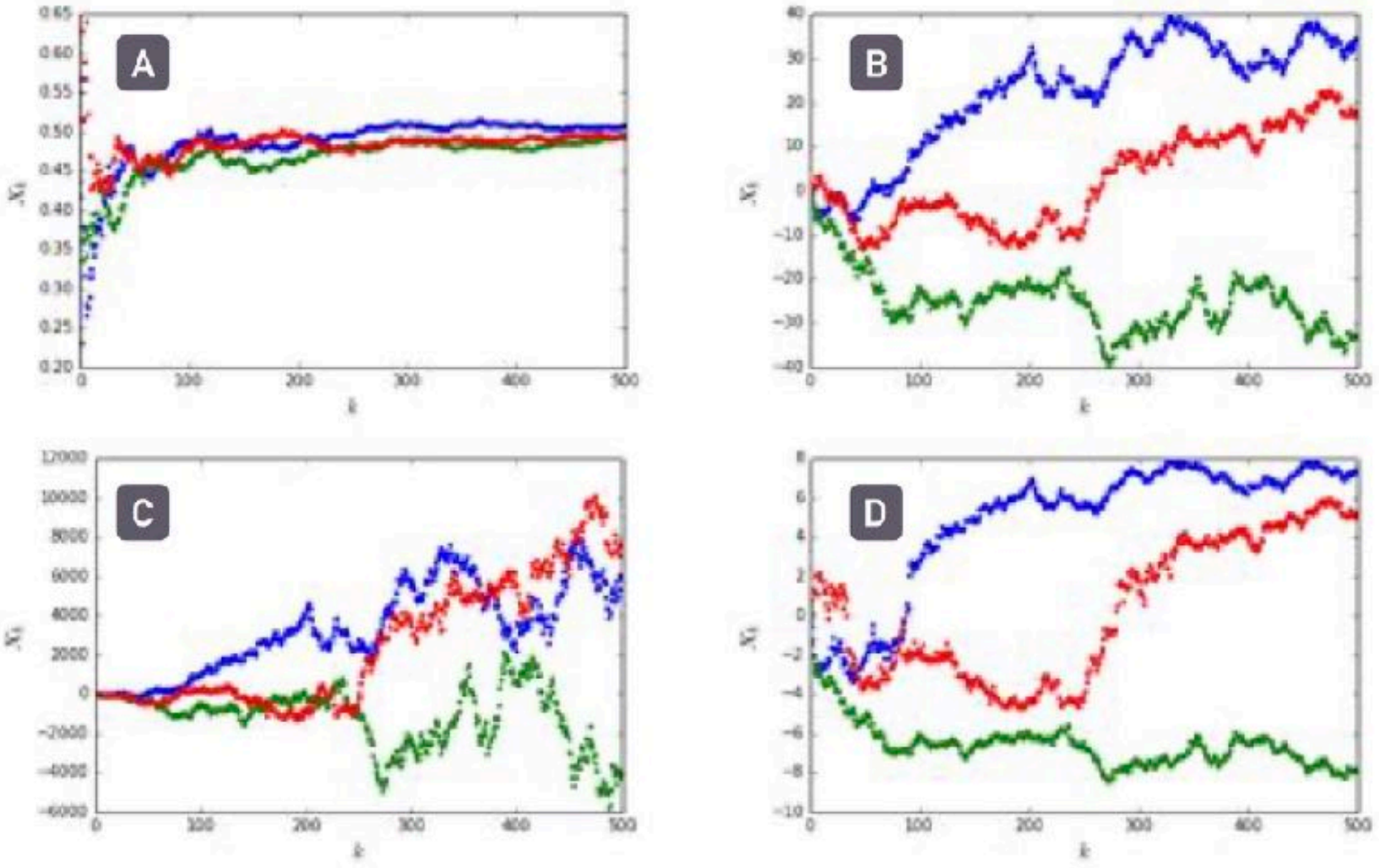
MCMC FLAVORS

- 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

MCMC DIAGNOSTICS

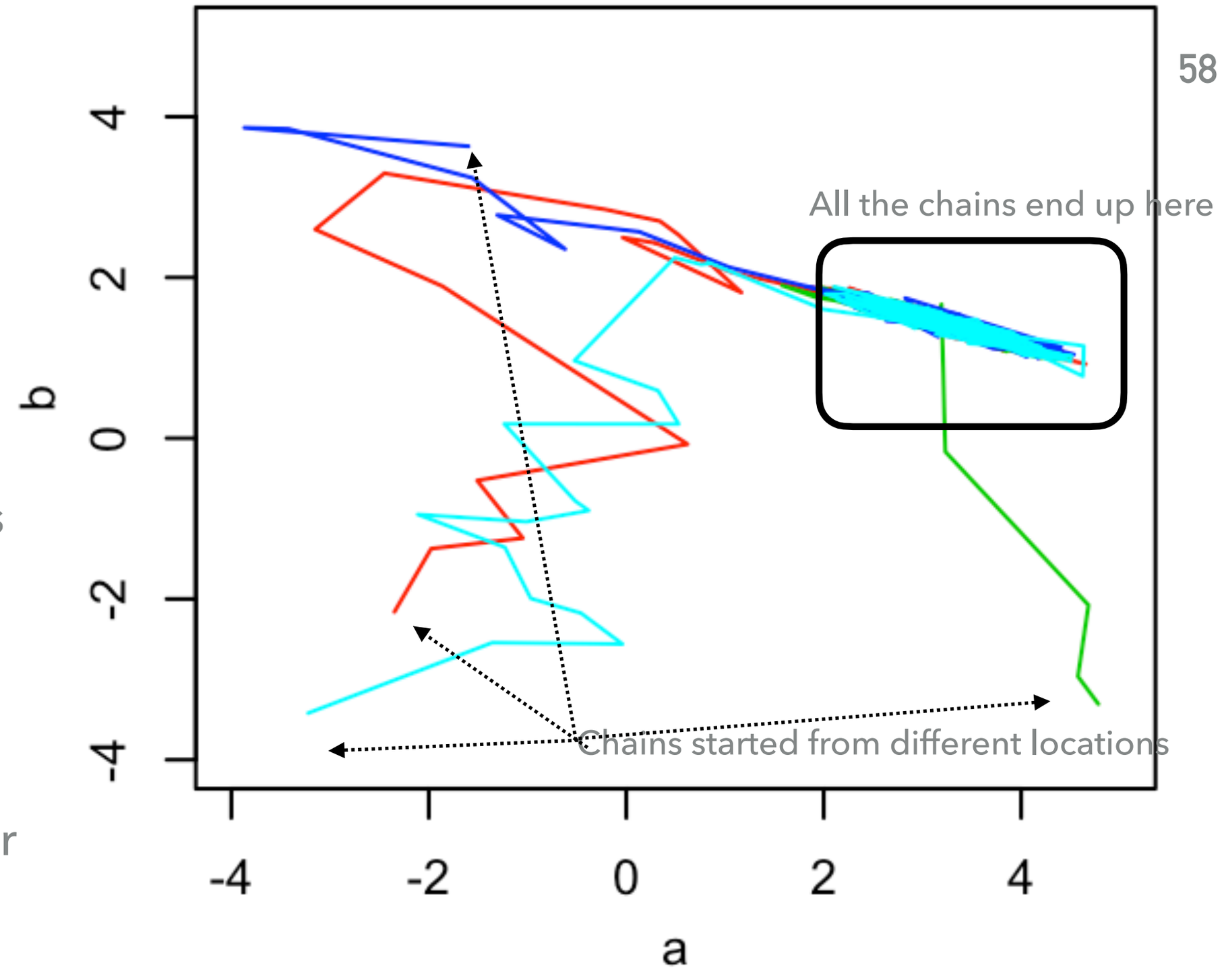
- What would make us confident of convergence?
  - Is the chain stationary?
  - Do independent chains started from overdispersed positions end up in the same stationary state?
- There's also a trick here - we wanted i.i.d samples from the posterior, but to make our Markov chain efficient, we violated the "independence" criteria a little
  - Each sample of the chain depends on the previous sample through **g(x, 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?



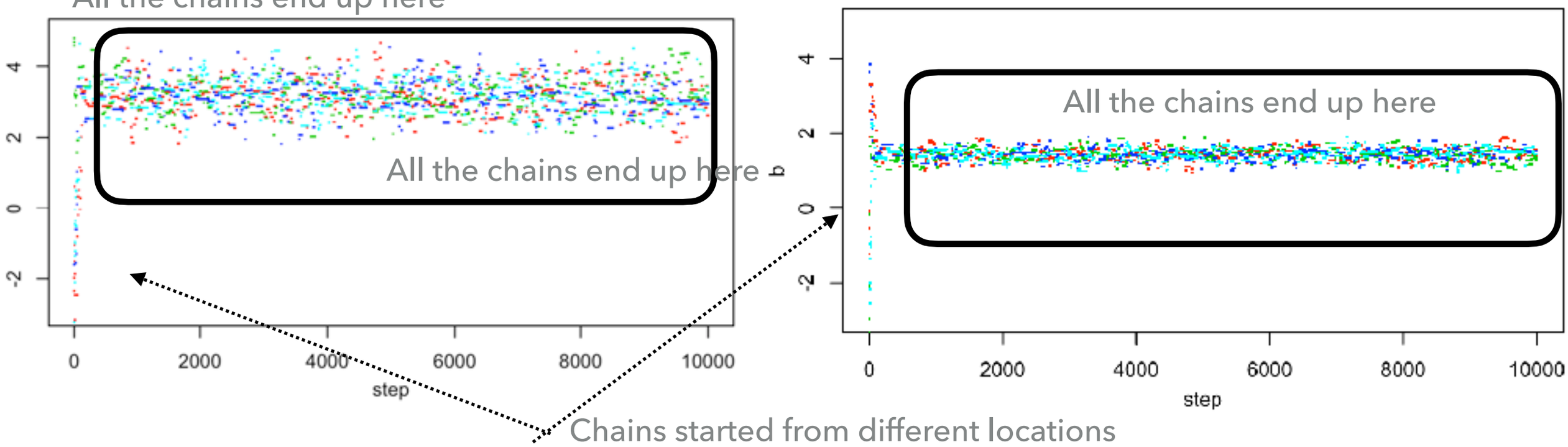
Setup: here's some chains
from a simple linear problem

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



MCMC DIAGNOSTICS

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



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

GELMAN-RUBIN

- 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 $\theta\bar{j}$ is the average θ for chain j and $\bar{\theta}$ is the global average.

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

GELMAN-RUBIN

- 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.

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

GELMAN-RUBIN

- 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

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

GELMAN-RUBIN

- 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

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

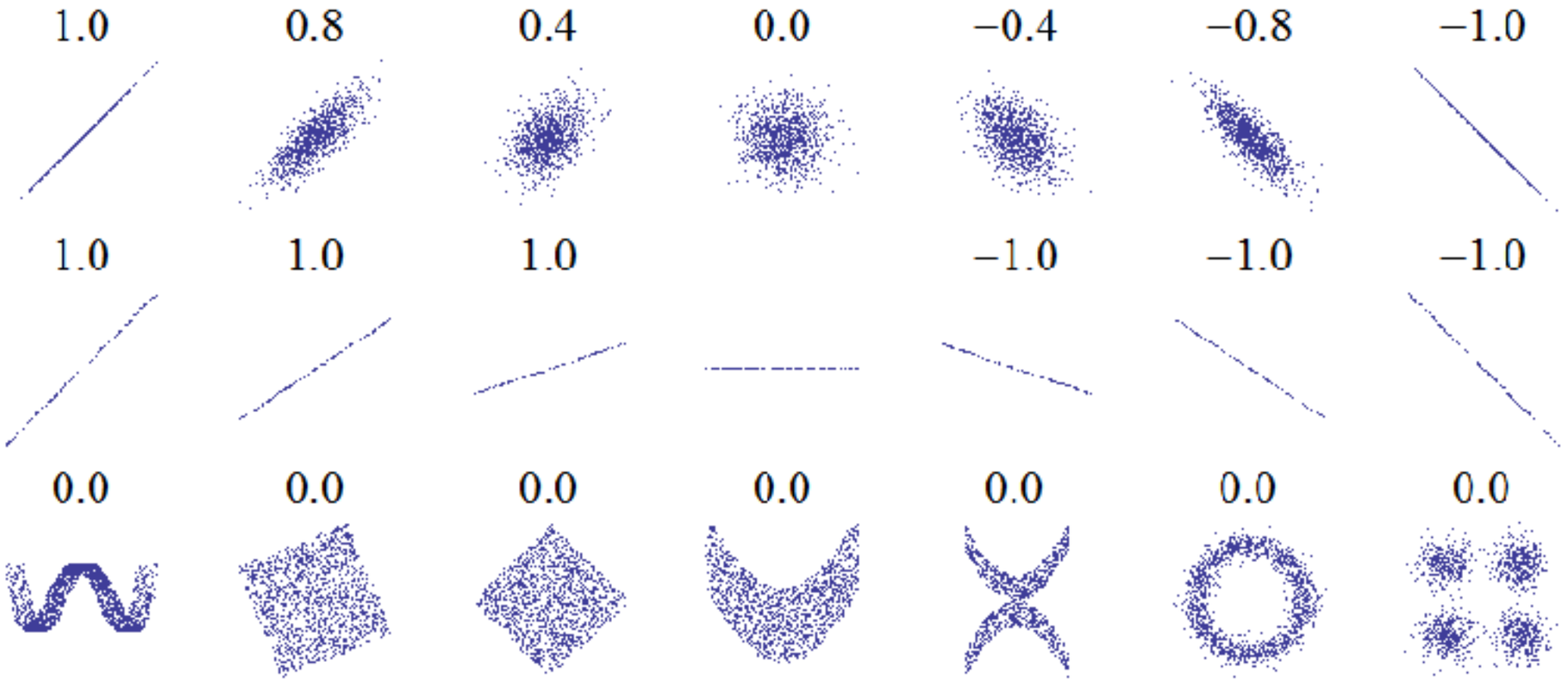
- We'd like to see $R \approx 1$ (e.g. $R < 1.1$ is often used). Note that this calculation can also be used to track convergence of combinations of parameters, or anything else derived from them. **Iff $R \approx 1$, then the chains are described as well mixed.**

RECAP

- You know you can compute the correlation between two random variables x and y

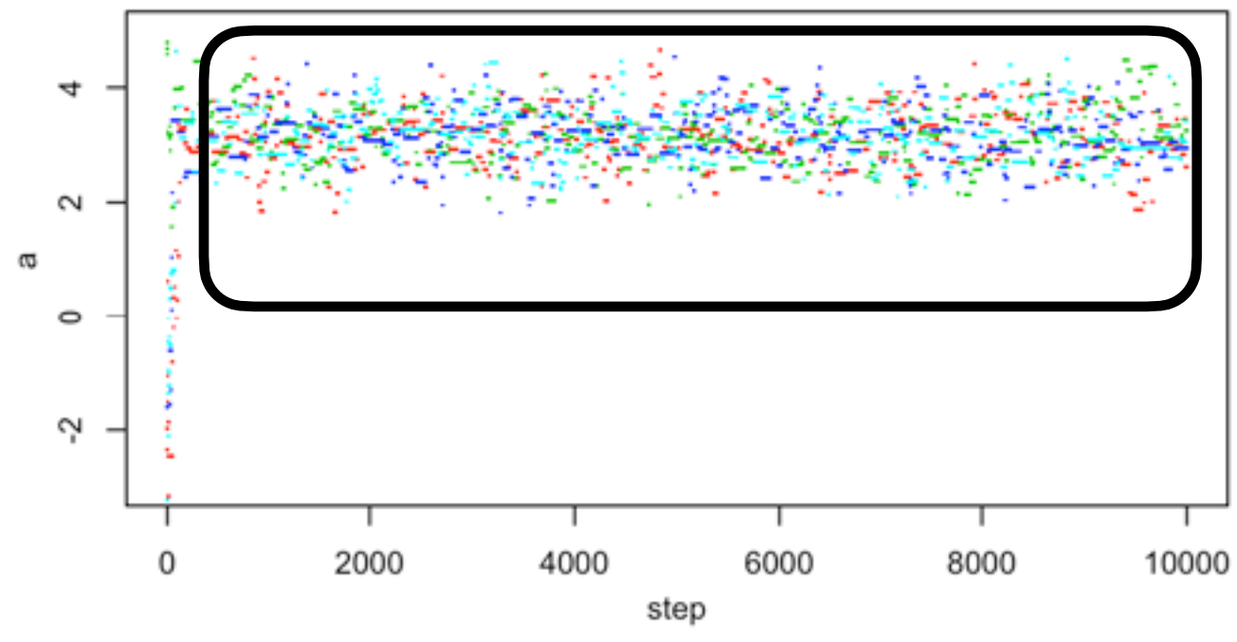
$$\mathrm{Cov}(x,y) = E([x - \langle x \rangle])E([y - \langle y \rangle])$$

$$\mathrm{Cor}(x,y) = \frac{\mathrm{Cov}(x,y)}{\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}}$$



CORRELATION TESTS

- Each Markov chain is just N_sample long vector of an M-dimensional parameter space
- 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.**



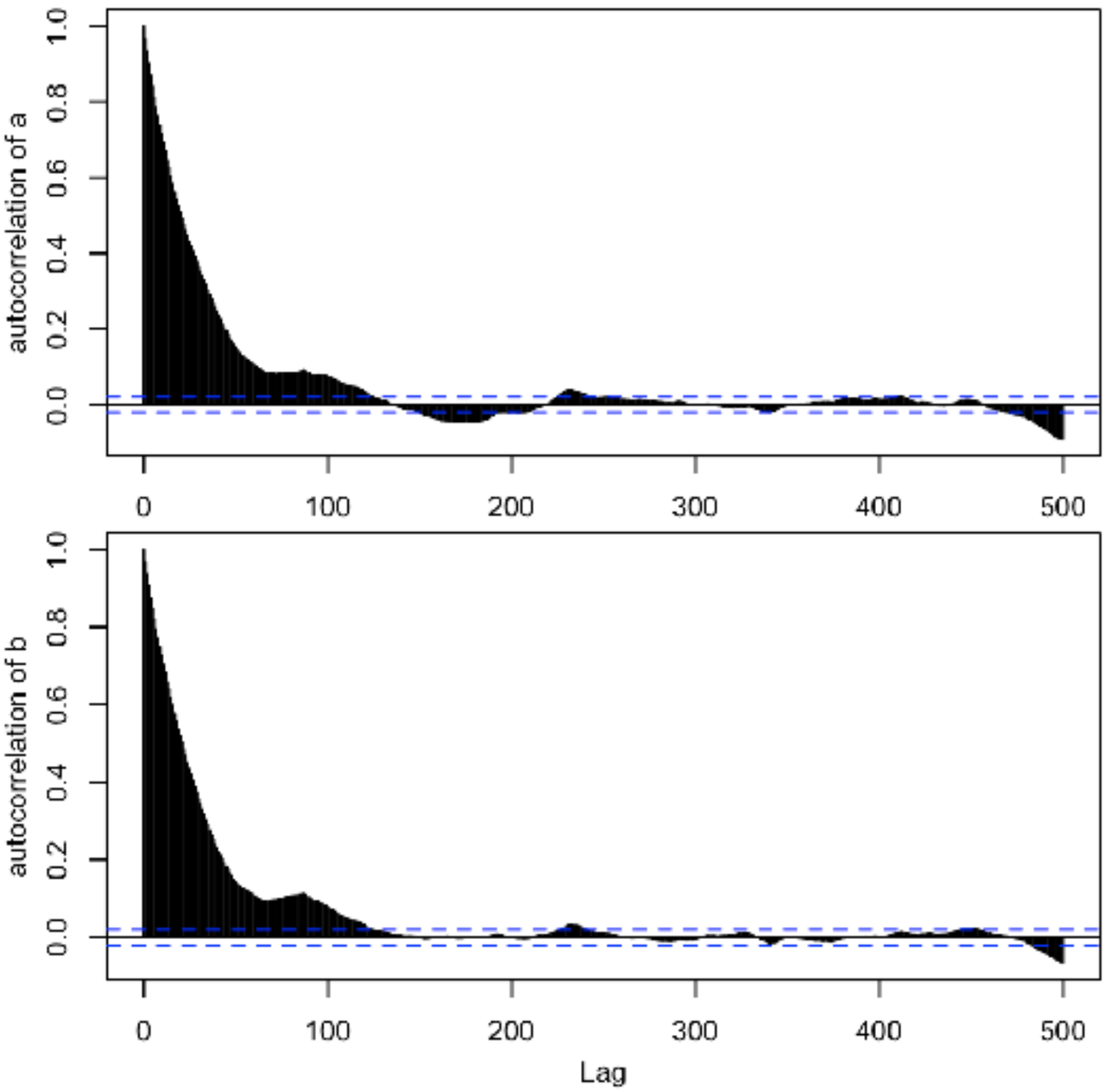
CORRELATION TESTS

- The **autocorrelation** of a sequence (after removing burn-in), as a function of lag, k, is - correlation of a series with itself k steps away

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

- 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



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 **thining** the chains by a factor of ~150, apparently! (i.e. take every 150th sample)

EFFECTIVE NUMBER OF SAMPLES

- 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 62)

$$\hat{\rho}_t = 1 - \frac{V_t}{2V}$$

- and

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

Time-Series, Fourier Transforms and the PSD

WHAT IS A TIME SERIES

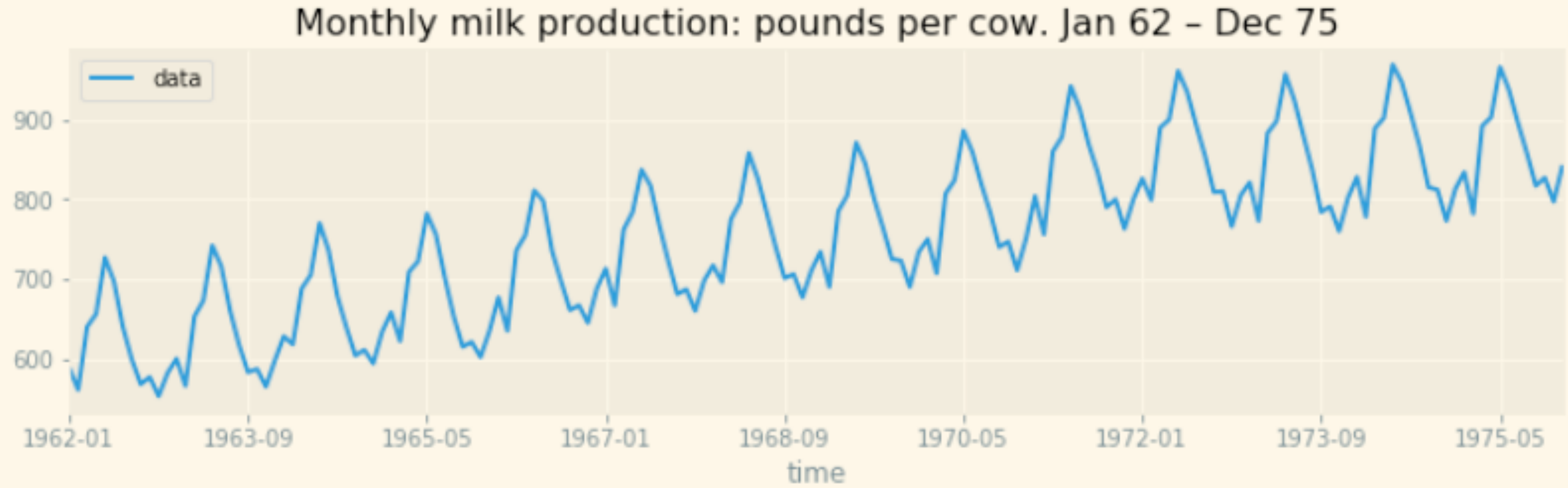
- (t1, m1), (t2, m2), (t3, m3)… (t_n, m_n) - instead of just a numerical index, the index is an **exogenous** variable - in this case time
- A time-series is any sequence of observation such that the distribution of m_k depends on m_{k-1}, m_{k-2}.
  - Time is directional - measurements only depend on the past. This is a statement of **causality**.
    - This also means that time series measurements are **not independent**
- Applies to astronomical measurements, your brain's electrical activity, the stock market, sea level in the face of a hurricane…
- These observations are typically not uniformly sampled - nor can they be from Earth even in principle (rotation, revolution).

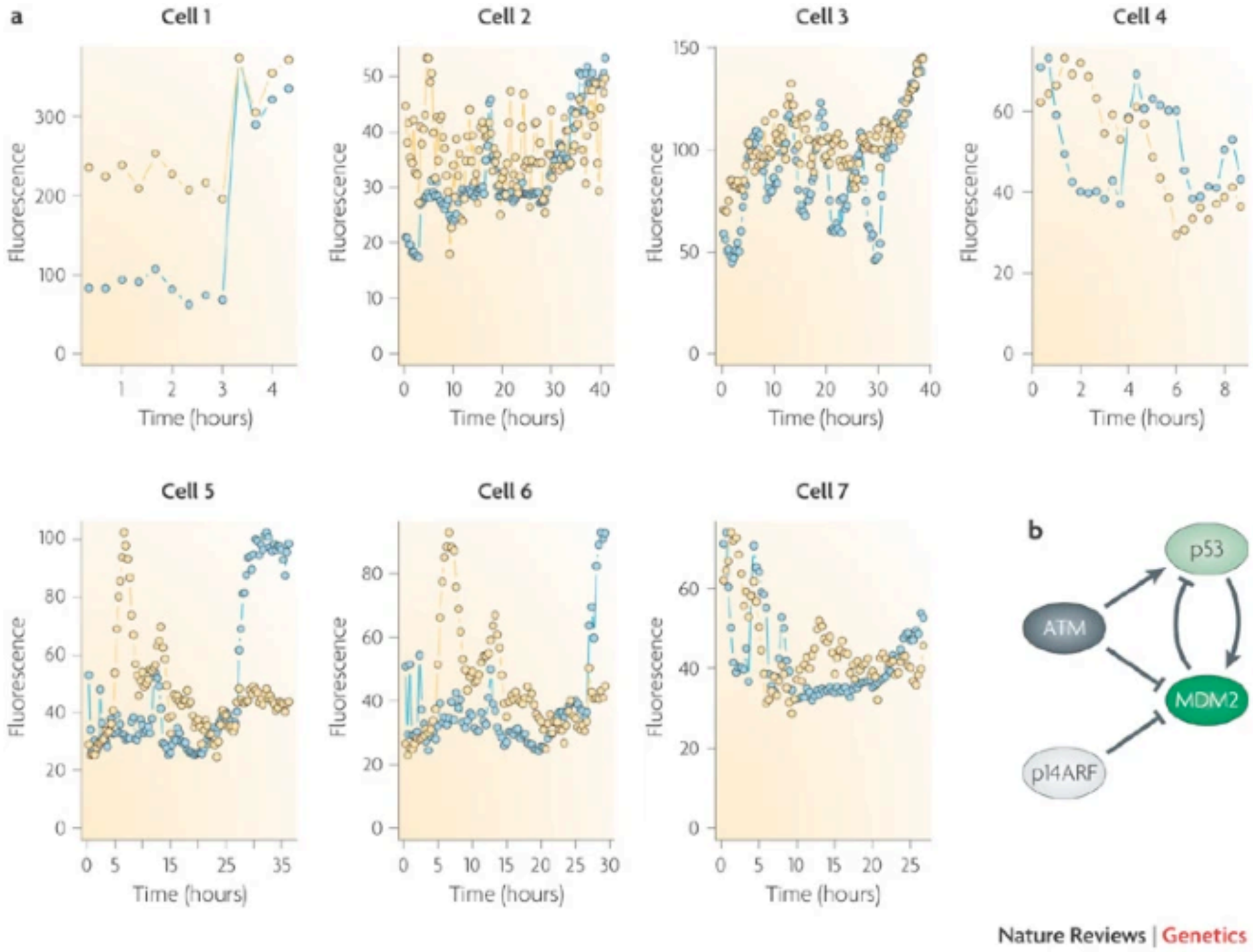
TIME SERIES HELP

- We need some mathematical formalism to work with radio data, where you record a continuous sequence of signals and have to contend with the wave nature of light
- This will ALSO help us in the optical and with images - either both as individual images or as a **time series**
- But this does mean we have to deal with some mathematical formalism…

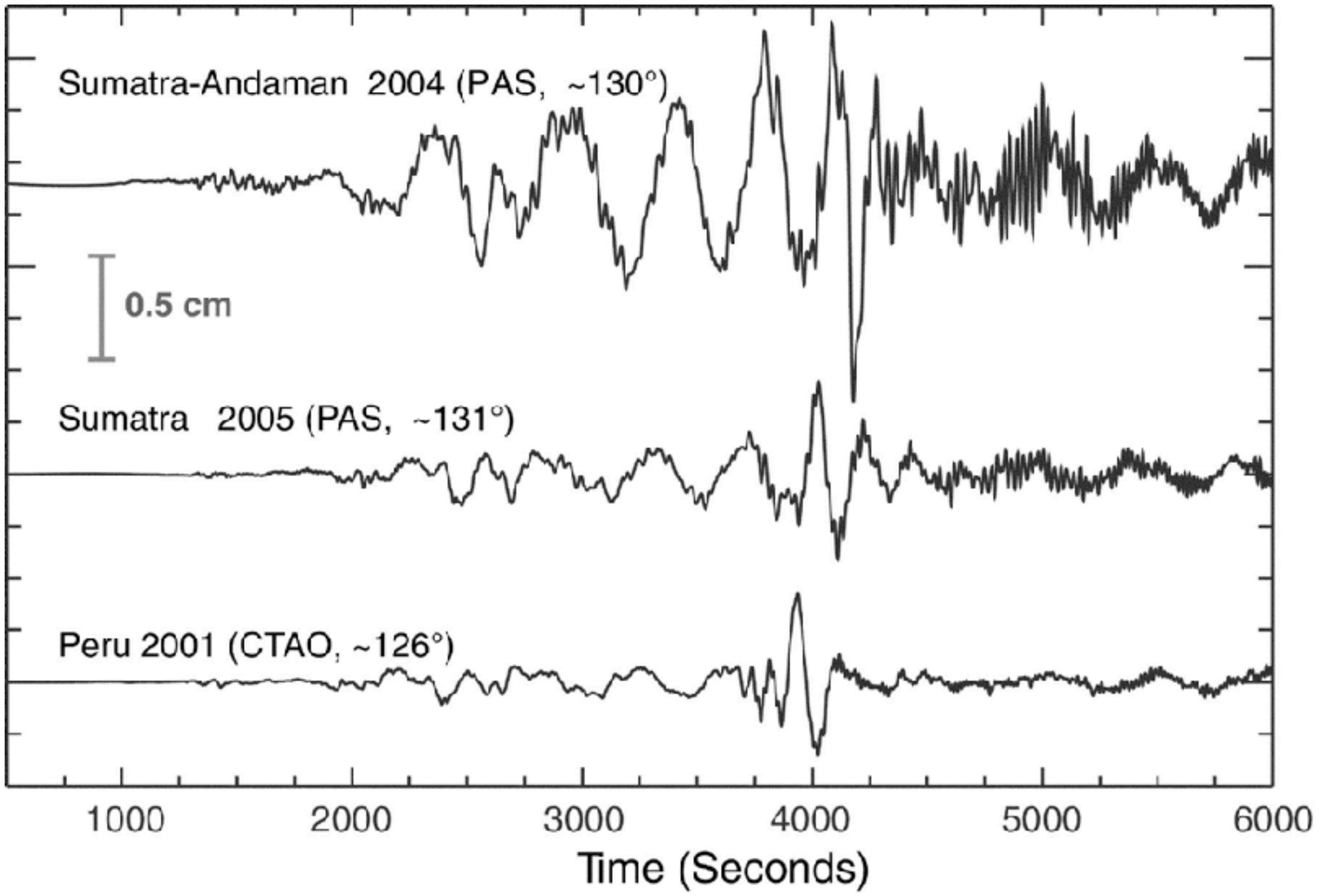
- **Fourier transform** tutorial http://www.thefouriertransform.com/
- NASA Imagine the Universe page on **timing analysis** https://imagine.gsfc.nasa.gov/science/toolbox/timing2.html
- AAVSO **time series** tutorial (Matthew Templeton) https://www.aavso.org/time-series-tutorial
- NumPy **FFT** reference https://docs.scipy.org/doc/numpy/reference/routines.fft.html
- Astropy `timeseries` module documentation https://docs.astropy.org/en/stable/timeseries/

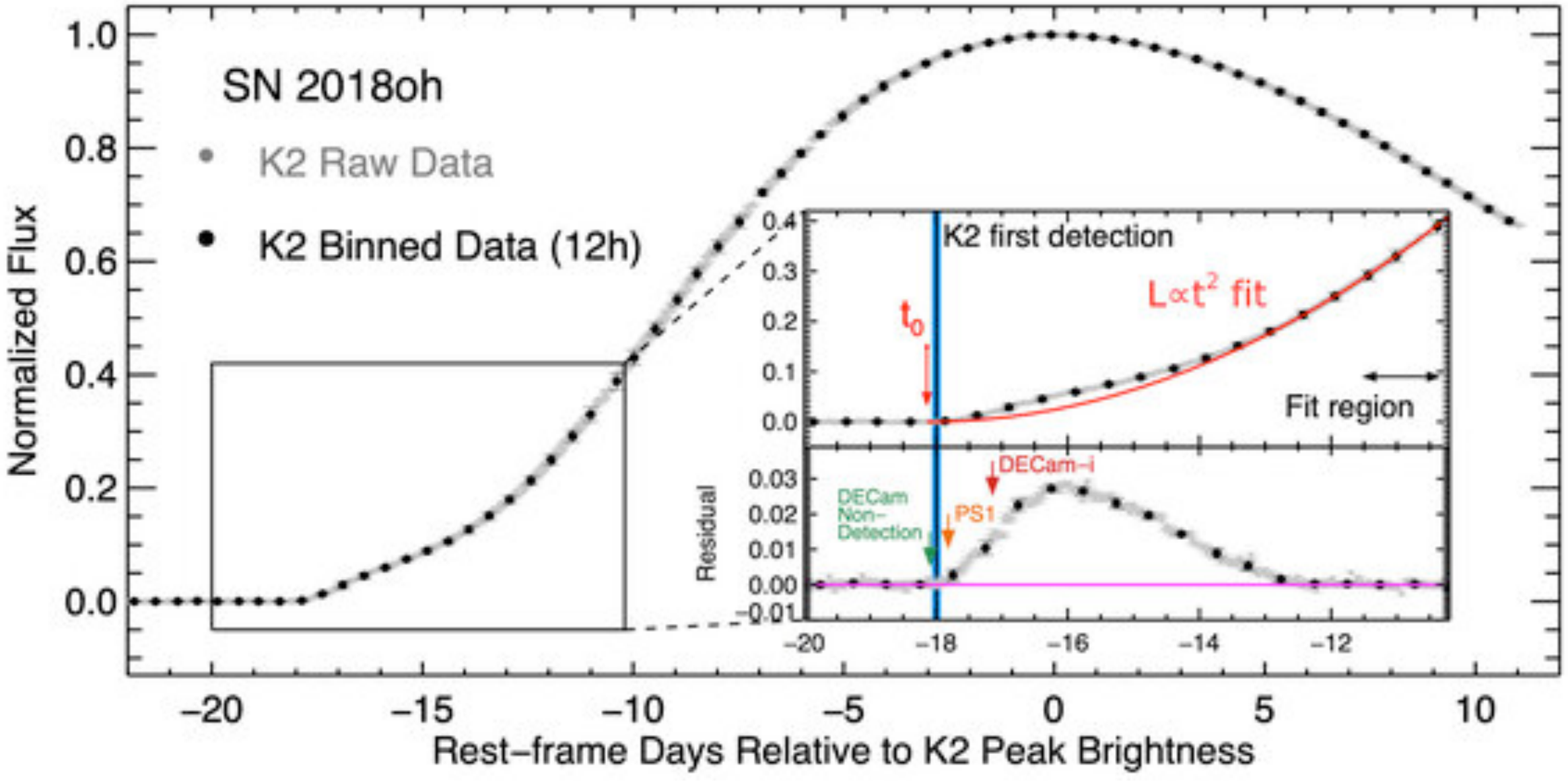
WHAT FEATURES JUMP OUT OF EACH OF THE FOLLOWING TIME SERIES?


Monthly milk production: pounds per cow. Jan 62 – Dec 75
data
900
800
700
600
1962-01
1963-09
1965-05
1967-01
1968-09
1970-05
1972-01
1973-09
1975-05
time



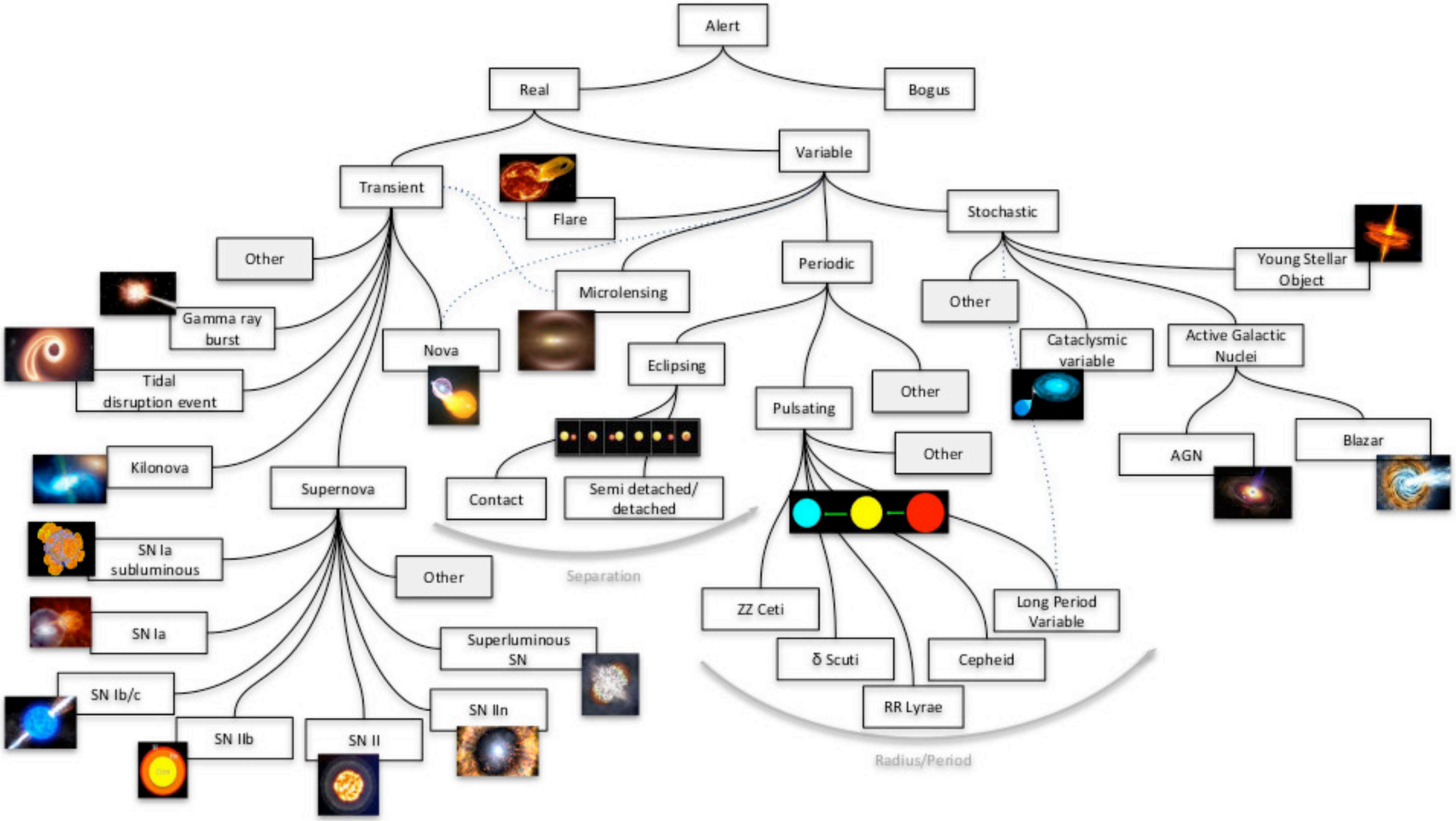
Fluctuations in p53 and MDM2 levels in single cells.


Sumatra-Andaman 2004 (PAS, ~130°)
0.5 cm
Sumatra 2005 (PAS, ~131°)
Peru 2001 (CTAO, ~126°)
1000
2000
3000
4000
5000
6000
Time (Seconds)


SN 2018oh
K2 Raw Data
K2 Binned Data (12h)
Normalized Flux
Rest−frame Days Relative to K2 Peak Brightness
K2 first detection
t_0
$L \propto t^2$ fit
Fit region
Residual
DECam Non−Detection
PS1
DECam−i

CATEGORIES OF TIME-SERIES

- We are often interested in extracting from the data:
- **Periodic signals** – source repeats with one or more characteristic periods. Goal is to tease out the different periods and their amplitudes, or to find a particular repeating signal shape to determine physical information about the source.
    - Example: depth of eclipse in an eclipsing binary helps to determine binary parameters (masses, sizes, orbital separation).
- **Burst signals** – source is a one-time (often explosive) event. Goal is to match a signal shape to determine properties of the object that caused the event.
    - Example: shape of a gravitational wave strain curve helps to determine nature of merging objects (black hole/neutron star), their masses, their spins, and the eccentricity of the orbit.
- The data always have noise in them, either due to instrumentation or because of fluctuations in the physical system involved. Statistical methods are needed to robustly extract information and determine its reliability.


Alert
Real
Bogus
Variable
Transient
Flare
Other
Gamma ray burst
Tidal disruption event
Kilonova
Supernova
SN Ia subluminous
SN Ia
SN Ib/c
SN IIb
SN II
Other
Superluminous SN
SN IIn
Nova
Microlensing
Periodic
Eclipsing
Contact
Semi detached/ detached
Separation
Pulsating
Other
ZZ Ceti
δ Scuti
RR Lyrae
Cepheid
Long Period Variable
Radius/Period
Stochastic
Other
Young Stellar Object
Cataclysmic variable
Active Galactic Nuclei
AGN
Blazar

Orthogonal polynomials

- Consider an interval (***a,b***), finite or infinite, in ***R***.
- A set of polynomials $p_n(x)$, $n=0,1,2,\ldots$, is orthogonal on (a,b) if:

$$\int_a^b p_n(x)\,p_m(x)\,dx = 0, \quad m \neq n$$

- A function expanded as a series over $p_n(x)$,

$$f(x) = \sum_n a_n\,p_n(x)$$

- The series coefficients can be obtained by exploiting orthogonality:

$$\int_a^b f(x)\,p_m(x)\,dx = \int_a^b \left(\sum_n a_n\,p_n(x) \right) p_m(x)\,dx$$
$$= a_m$$

- Related to the theory of linear function spaces on which an inner product **<*f,g*>** is defined (a normed space). Function space spanned by orthogonal basis functions for which **<*f,g*>**=*0*

CATEGORIES OF TIME-SERIES

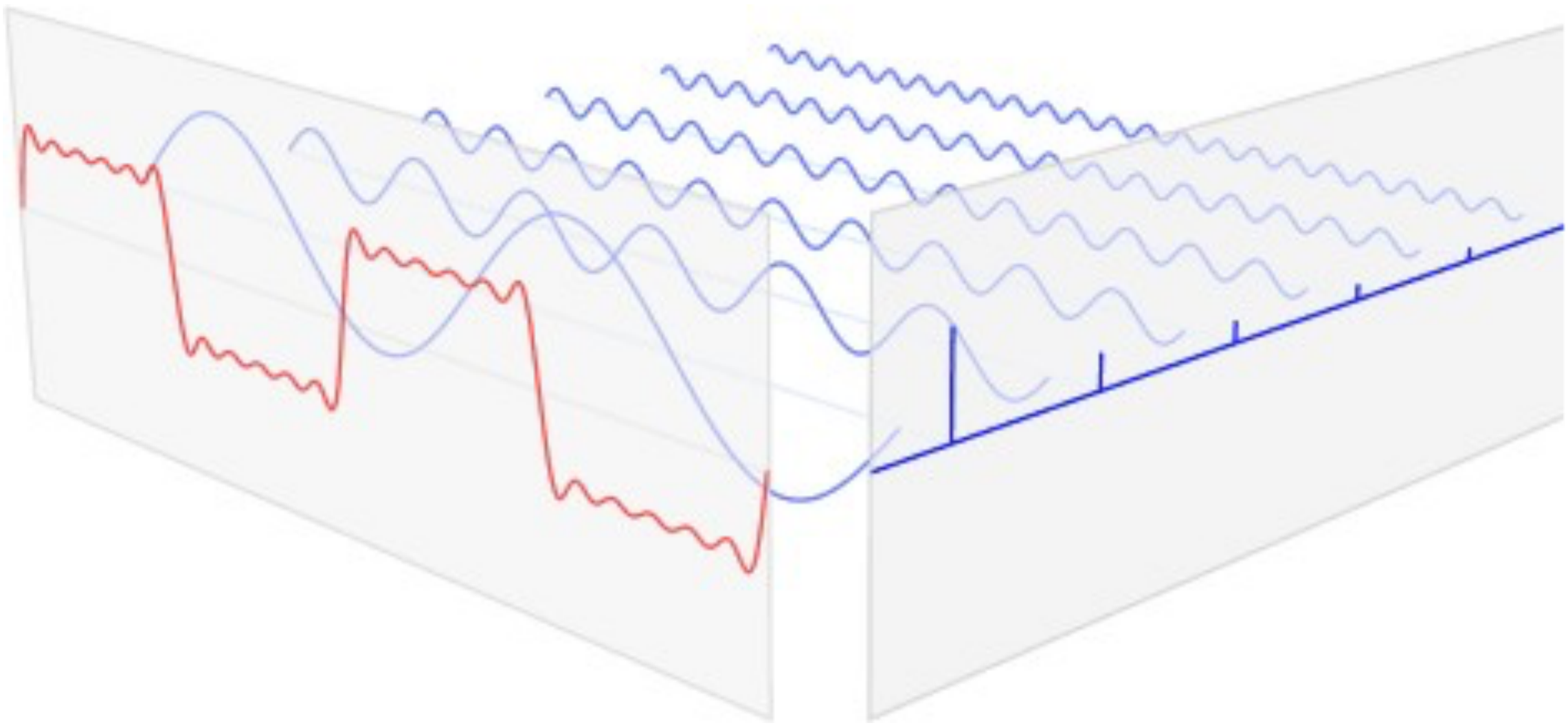
- Any "sufficiently smooth" periodic function *f(x)* can be decomposed into a sum of sine and cosine functions. This sum is called a **Fourier series**:

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty}\left[a_n \cos\left(\frac{2\pi n x}{L}\right) + b_n \sin\left(\frac{2\pi n x}{L}\right)\right]$$

- where the period of *f(x)* here is *L*. The **Fourier coefficients** are:

$$a_n = \frac{2}{L}\int_{-L/2}^{L/2} f(x)\cos\left(\frac{2\pi n x}{L}\right)\,dx \qquad b_n = \frac{2}{L}\int_{-L/2}^{L/2} f(x)\sin\left(\frac{2\pi n x}{L}\right)\,dx$$

- The quantity $k_n = \frac{2\pi n}{L}$ is the **frequency** if *x* represents time or the **wavenumber** if *x* represents space. Computing the **Fourier decomposition** of a function allows us to determine ***which frequencies dominate the signal*** represented by *f(x)*.

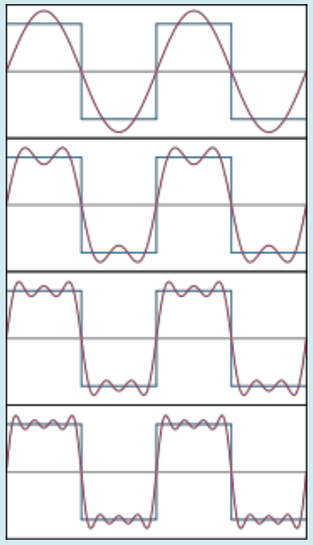


Dual Fourier domains

- The continuous Fourier transform connects two reciprocal domains of units ***x*** and $s=x^{-1}$, for example:

Domain	Reciprocal domain
Time (second)	Frequency (Hz = cycles s^{-1})
Spatial frequency (cycles m^{-1})	Spatial distance (m)

- E.g. The decomposition of a function ***f(t)*** into its spectral components ***[cos(2πvt)+ i sin(2πvt)].***



(CC)

Fourier series: arbitrary intervals

- Over an interval (-L/2, L/2):

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos\frac{2\pi nx}{L} + b_n \sin\frac{2\pi nx}{L} \right)$$

$$a_n = \frac{2}{L} \int_{-L/2}^{L/2} f(x) \cos\frac{2\pi nx}{L}\, dx$$

$$b_n = \frac{2}{L} \int_{-L/2}^{L/2} f(x) \sin\frac{2\pi nx}{L}\, dx.$$

(Sutton 2011)

- General symmetry properties:

Function type	a_n (cosine terms)	b_n (sine terms)
Even: ***f(x) = f(-x)***	☑	Zero.
Odd: ***f(x) = -f(-x)***	Zero.	☑

CATEGORIES OF TIME-SERIES

- A general (non)periodic function *f(x)* can be decomposed in a similar way:

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dk\, \hat{f}(k) e^{ikx} = \frac{1}{2\pi}\int_{-\infty}^{\infty} dk\, \hat{f}(k)(\cos kx + i \sin kx)$$

$$\hat{f}(k) = \int_{-\infty}^{\infty} dx\, f(x) e^{-ikx}$$

- The (in general complex-valued) function *f(k)* is called the **Fourier transform** of *f(x)*. The first integral above is the **inverse Fourier transform**. The two transforms are inverses of each other (hence the name!)

- Fourier transforms allow us to extract periodic information from signals that may consist of a mixture of periodic and non-periodic information.

- A particularly useful property if *f(x)* is real (bars denote complex conjugate) - real functions are by definition their own complex conjugates

$$\hat{f}(-k) = \int_{-\infty}^{\infty} dx\, f(x) e^{ikx} = \int_{-\infty}^{\infty} dx\, \overline{f(x)}\, \overline{e^{-ikx}} = \int_{-\infty}^{\infty} dx\, f(x) e^{-ikx} = \overline{\hat{f}(k)}$$

i.e. if you know the inverse function at some frequency *k*, you also know it at *-k*

We'll make use of this property in a few slides

FOURIER TRANSFORM EXAMPLE

- A square wave is defined as:

$$f(x) = \begin{cases} 1 & |x| < h/2 \\ 0 & |x| \geq h/2 \end{cases}$$

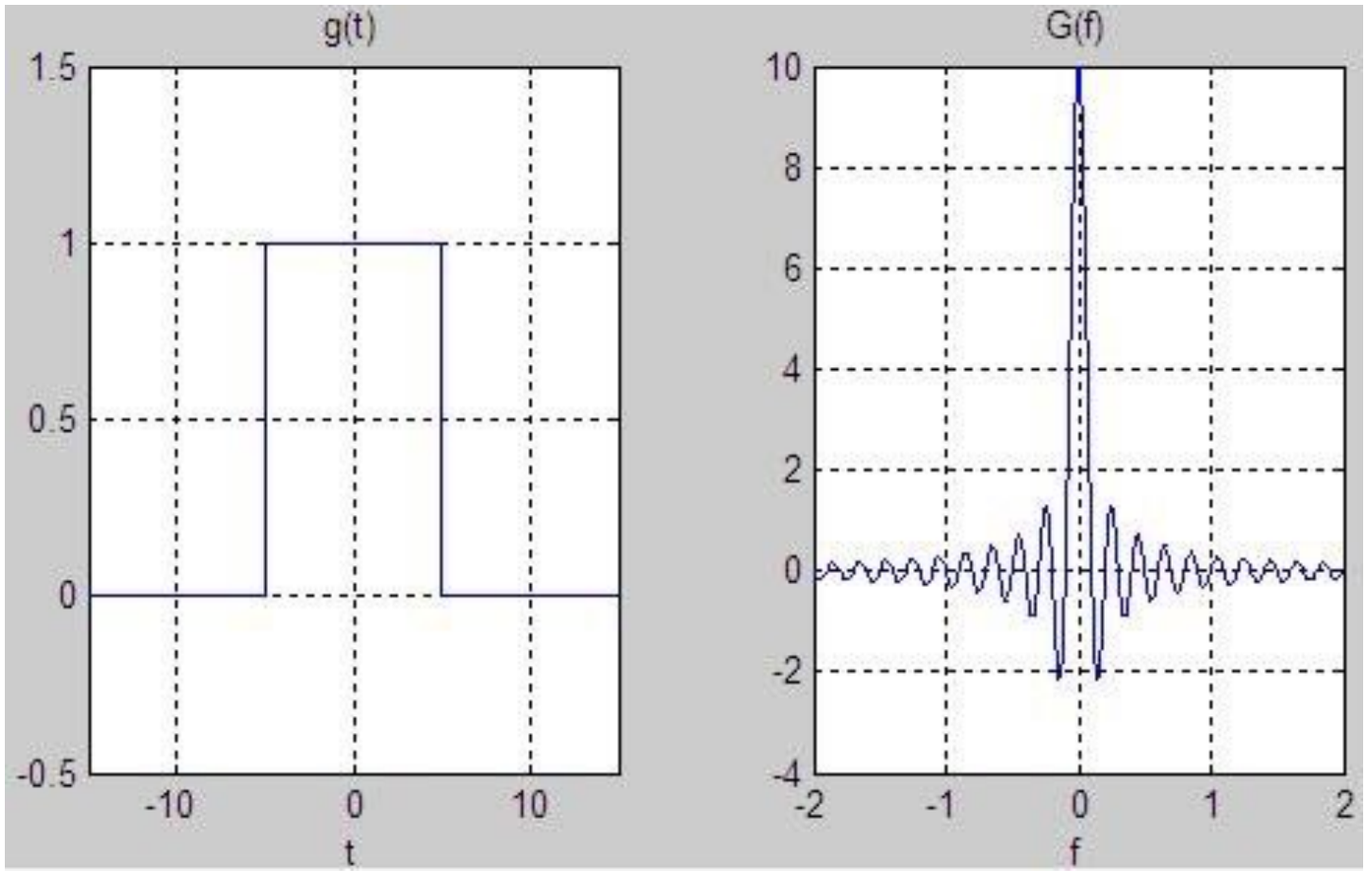
- So the Fourier transform is:

$$\hat{f}(k) = \int_{-\infty}^{\infty} dx\, f(x) e^{-ikx}$$

$$= \int_{-h/2}^{h/2} dx\, (\cos kx - i \sin kx)$$

$$= \frac{1}{k} \sin kx \Big|_{-h/2}^{h/2}$$

$$= \frac{2}{k} \sin(kh/2)$$



- In general, the “wider” a signal is in real space (x), the “narrower” it is in Fourier space (k).

THE DISCRETE FOURIER TRANSFORM

- If our function *f(x)* is known only on a ***uniformly sampled*** grid of N points with grid spacing = L/N, we can approximate its Fourier transform at a uniformly sampled grid of frequency points using the **discrete Fourier transform (DFT)**

$$f_n = f(n\Delta x) = \frac{1}{N}\sum_{m=0}^{N-1} \hat{f}_m e^{i2\pi mn/N}$$

$$\hat{f}_m = \hat{f}(2\pi m/L) = \sum_{n=0}^{N-1} f_n e^{-2\pi imn/N}$$

- Although *m* formally goes up to *N – 1*, the Nyquist frequency (*k* corresponding to *m = N/2*)

$$k_{\mathrm{Nyq}} = \frac{N}{2}\frac{2\pi}{L} = \frac{\pi}{\Delta x}$$

is the largest k-value that can be represented on the grid (you need at least two points to detect an oscillation).

The DFT treats a non-periodic *f(x)* as if it were periodic with period *L*. The implied periodic replicas produce errors at high frequencies (**aliasing**) when compared with the regular Fourier transform.

THE FAST FOURIER TRANSFORM

- Naive calculation of the DFT requires of order N^2 operations – doubling N makes the calculation take four times as long

- The fast Fourier transform (FFT), reduces this to of order N log N operations. It works by interpreting the DFT as a polynomial evaluation:

$$\begin{aligned}\hat{f}_m &= \sum_{n=0}^{N-1} f_n e^{-2\pi i m n/N} \\ &= f_0 + f_1 z + f_2 z^2 + f_3 z^3 + \ldots + f_{N-1} z^{N-1} \\ &= (f_0 + f_2 z^2 + f_4 z^4 + \ldots) + z(f_1 + f_3 z^2 + f_5 z^4 + \ldots) \\ &= \hat{f}_{\mathrm{even},m}(z^2) + z\hat{f}_{\mathrm{odd},m}(z^2)\end{aligned}$$

where

$$z \equiv [e^{-2\pi i/N}]^m = \omega^m, \quad m = 0 \ldots N-1$$

are the **Nth complex roots of unity.**

We can compute and store these once, then compute the transform recursively. The number of levels of recursion is proportional to log2 N, and we need to do this for N values of m, hence the N log N cost of the algorithm.

THE FAST FOURIER TRANSFORM

- It's important to recognize that you don't get any additional information from an FFT over a DFT - an FFT is just re-expressing these expansions to compute them faster

$$f_n = f(n\Delta x) = \frac{1}{N}\sum_{m=0}^{N-1} \hat{f}_m e^{i2\pi mn/N}$$

$$\hat{f}_m = \hat{f}(2\pi m/L) = \sum_{n=0}^{N-1} f_n e^{-2\pi imn/N}$$

$$k_{\mathrm{Nyq}} = \frac{N}{2}\frac{2\pi}{L} = \frac{\pi}{\Delta x}$$

- Still have a grid of length N frequency values with the central value being *k_Nyq*

When the input is real-valued, symmetry allows only N/2+1 values to be stored in the output, saving memory.

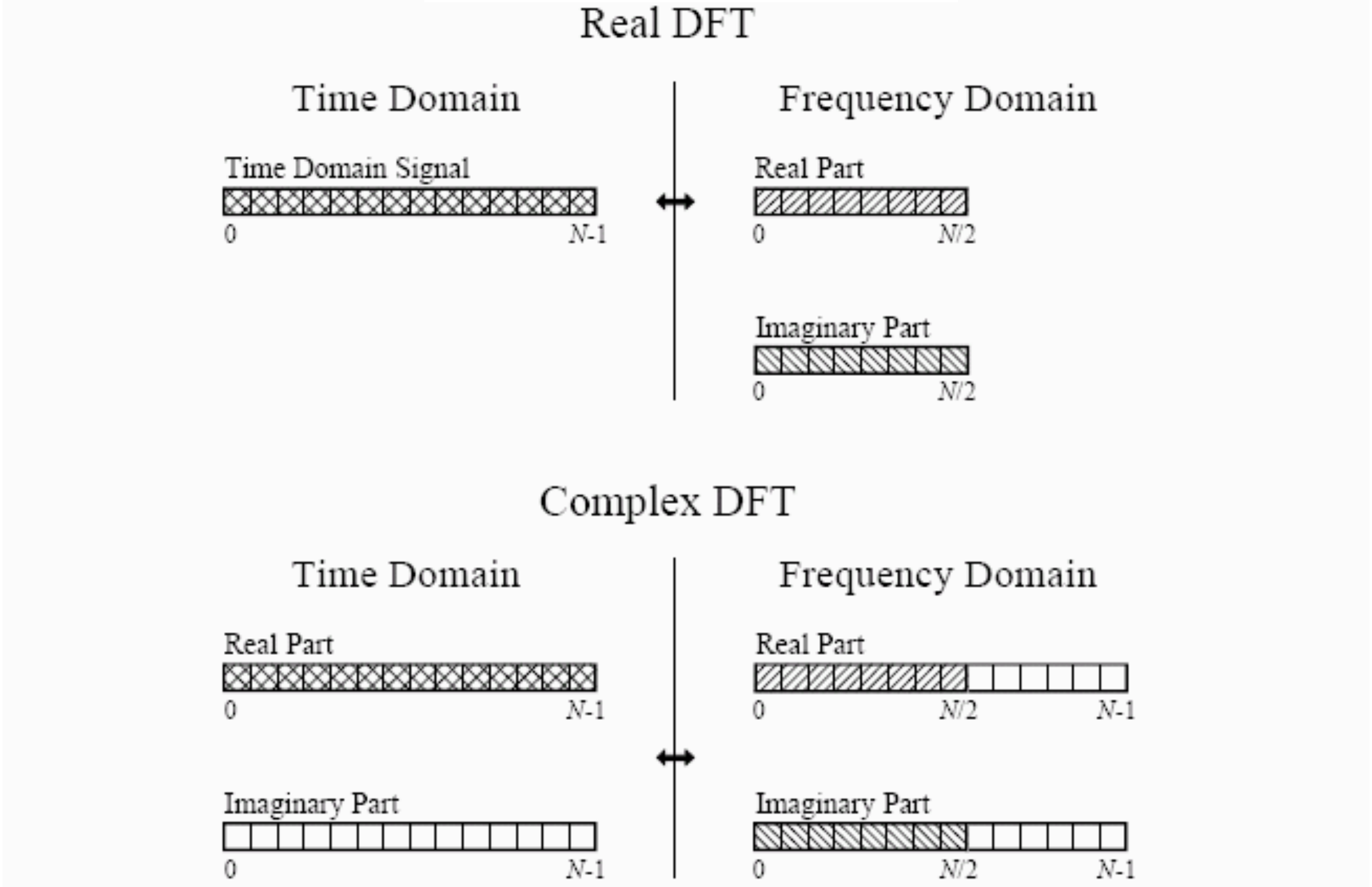


FIGURE 12-1
Comparing the real and complex DFTs. The real DFT takes an N point time domain signal and creates two $N/2 + 1$ point frequency domain signals. The complex DFT takes two N point time domain signals and creates two N point frequency domain signals. The crosshatched regions shows the values common to the two transforms.

FFTS WITH NUMPY

- The `numpy.fft` module provides FFT routines. For us the most important ones right now are:

```
rfft, irff                    1D real-complex FFT and
its inverse
```

- To make working with them easier, you can get the corresponding frequencies using

```
fftfreq(N)                frequencies for an N-
element transform
rfftfreq(N)     frequencies for an N-element
real transform


Example: (f.shape = (128,))

fhat = numpy.fft.rfft(f)
k    = numpy.fft.rfftfreq(128)
```

THE POWER SPECTRUM

- The **power spectrum** is a measure of how much each frequency/wavenumber contributes to the overall signal, regardless of phase - for our purposes "**squared amplitude of Fourier transform**"

- **Periodogram** estimator: given real-space function values

$$f_n = f(n\Delta x) \quad n = 0, \ldots, N-1$$

The discrete Fourier transform is

$$\hat{f}_m = \sum_{n=0}^{N-1} f_n e^{-2\pi i m n/N} \quad m = 0, \ldots, N-1$$

The periodogram estimate of the power spectrum is defined for nonnegative frequencies as

$$P_0 = P(0) = \frac{|\hat{f}_0|^2}{N^2}$$

Power at zero frequency

$$P_m = P(k_m) = \frac{1}{N^2}\left(|\hat{f}_m|^2 + |\hat{f}_{N-m}|^2\right) \quad m = 1, \ldots, \frac{N}{2} - 1; \; k_m \equiv \frac{2\pi m}{N\Delta x}$$

$$P_{N/2} = P(k_{\mathrm{Nyq}}) = \frac{|\hat{f}_{N/2}|^2}{N^2}$$

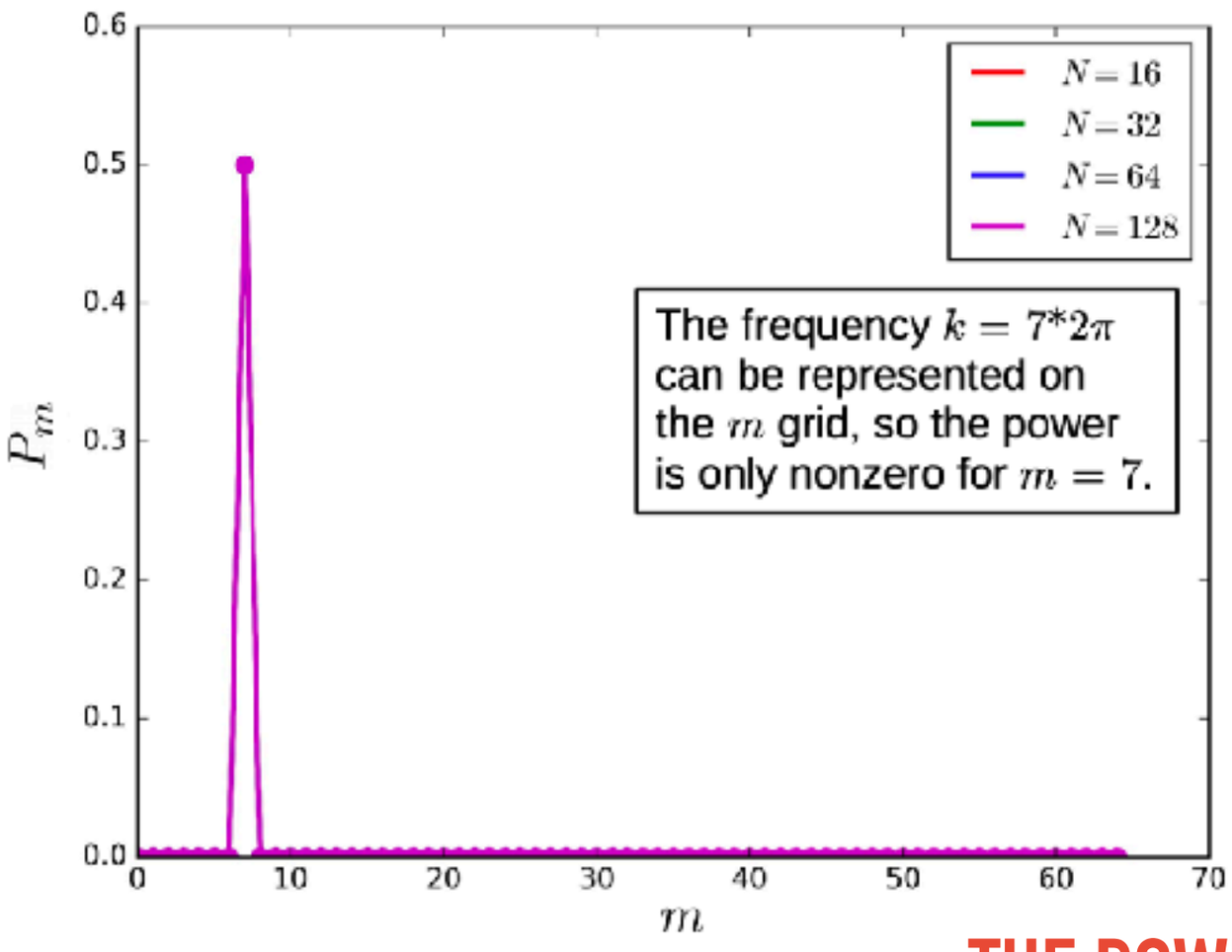
Power at Nyquist frequency

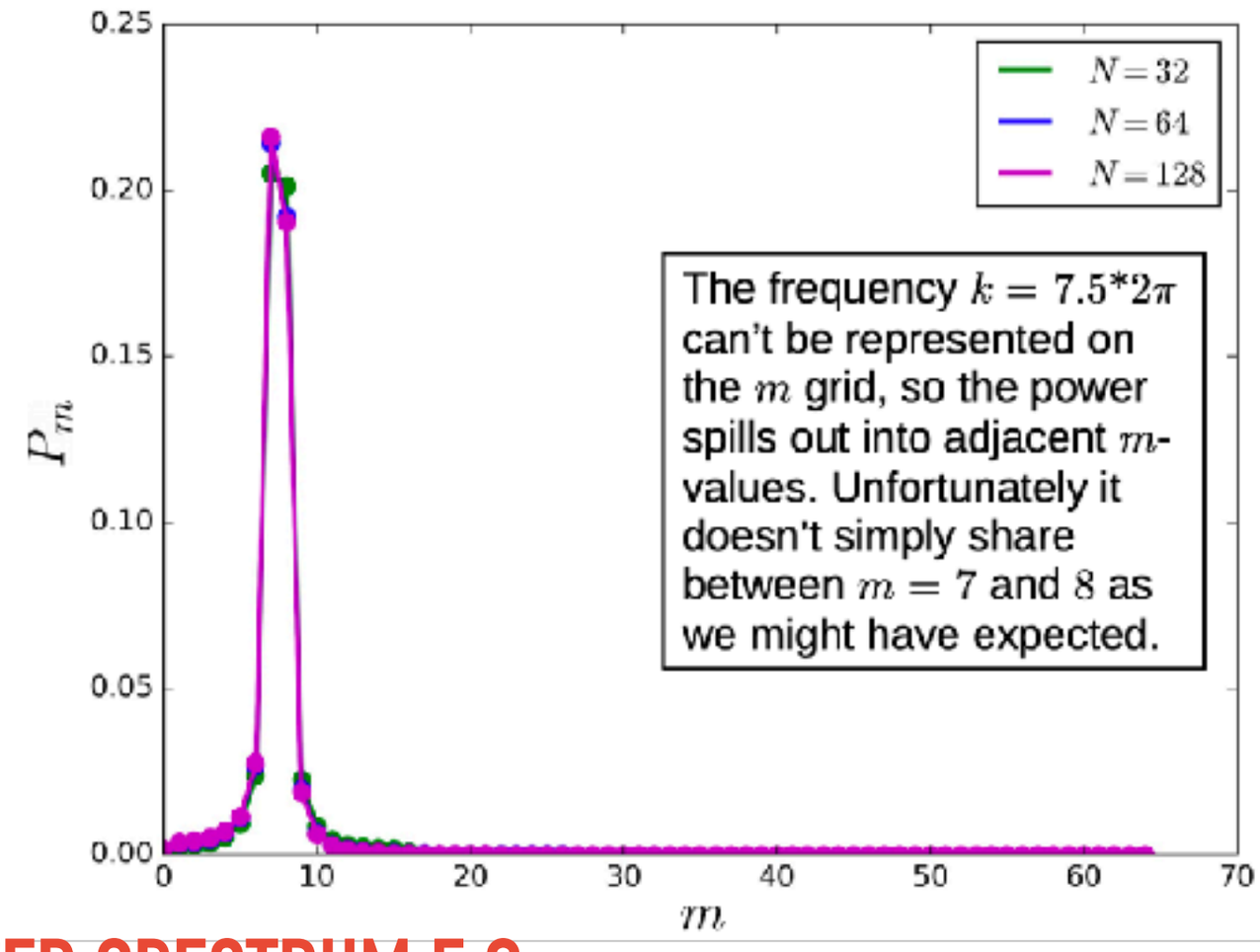
Problems with the periodogram estimator: Subject to leakage between modes

Statistical sampling variance is always 100%; as N → ∞, accuracy does not improve

$f(x) = \sin(2\pi x \cdot 7)$ $\Delta x = 1/N$ $N = 16, 32, 64, 1$

$f(x) = \sin(2\pi x \cdot 7.5)$ $\Delta x = 1/N$ $N = 32, 64, 128$

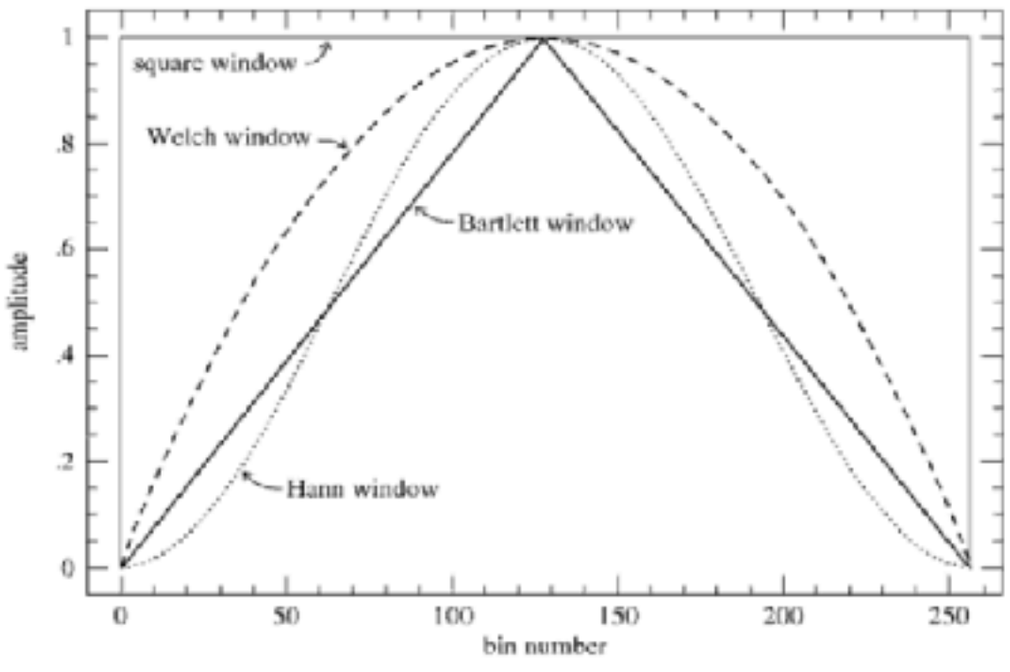


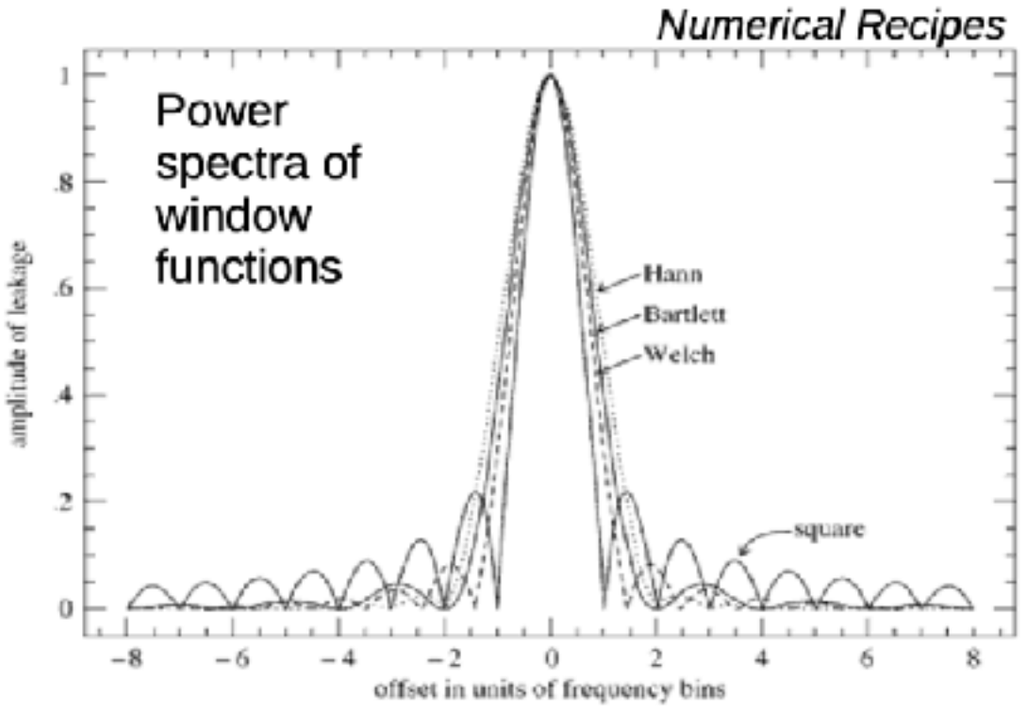


THE POWER SPECTRUM E.G.

DEALING WITH ALIASING AND LEAKAGE

- To deal with the **leakage** problem, it is necessary to **filter** the sampling window. No filtering corresponds to "square window" filtering, which is very leaky.
- Before computing the FFT, we multiply the input function *f(x)* by a window function *W(x)*, and when computing the power spectrum, we normalize by dividing by the integral of the square of the window function.





$$\hat{f}_m \to \sum_{n=0}^{N-1} W_n f_n e^{-2\pi i m n/N}$$

$$P_m \to \frac{N}{\sum_{n=0}^{N-1} W_n^2} P_m$$

Convolution

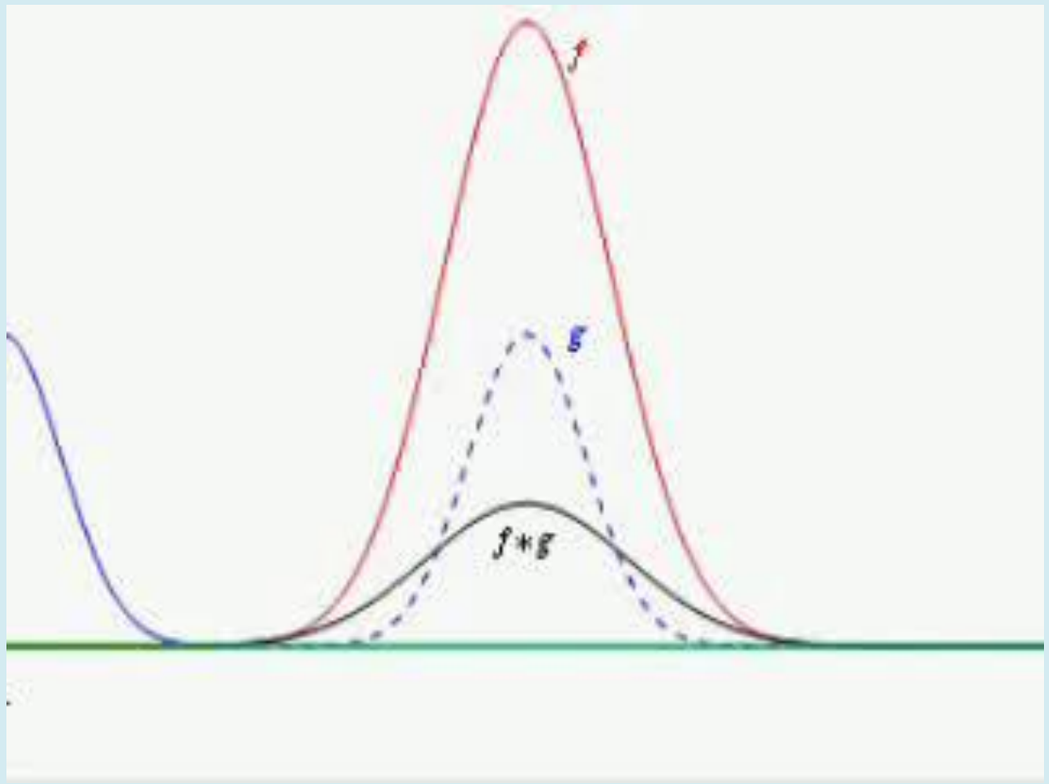
- The ***convolution*** of two functions ***f(x)*** and ***g(x)*** is denoted by ***f * g*** or ***f ** g***, and is defined as:

$$f(x) * g(x) = \int_{-\infty}^{+\infty} f(u)g(x-u)du$$

$$f * g = g * f \quad \text{(Commutative)}$$

$$f * (g * h) = (f * g) * h \quad \text{(Associative)}$$

$$f * (g + h) = f * g + f * h \quad \text{(Distributive)}$$

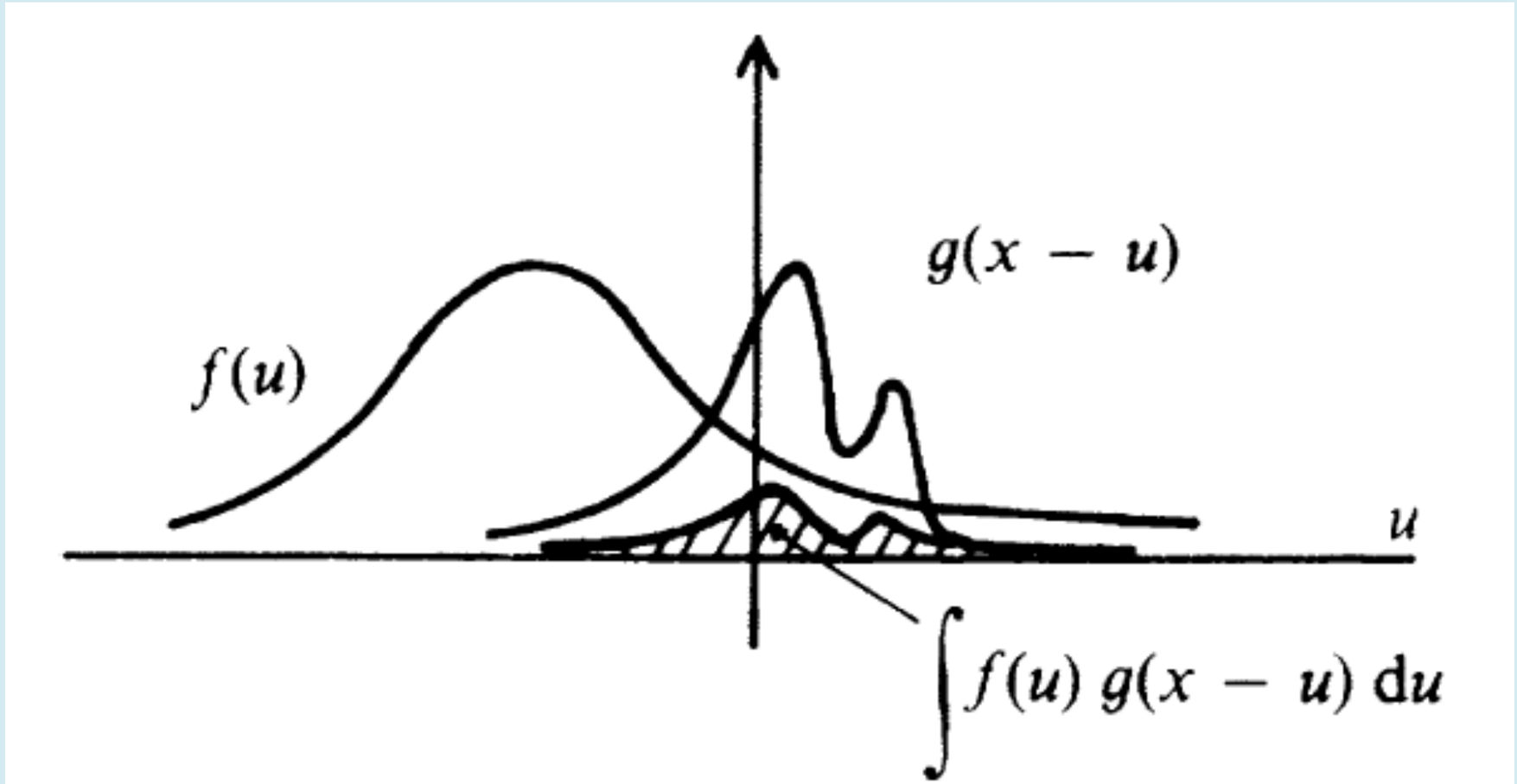


(Convolution of two Gaussian functions; (CC))

$$\sigma_{f*g} = \sqrt{\sigma_f^2 + \sigma_g^2}$$

$$\sigma = \text{Full-Width Half-Maximum (FWHM)}$$

Graphical depiction of convolution



(Léna et al. 2012)

Convolution in the Fourier domain

- The ***convolution*** of two functions ***f(x)*** and ***g(x)*** is becomes a product in the dual Fourier domain:

$$f(x) * g(x) \rightleftarrows \tilde{f}(s)\tilde{g}(s)$$

- And inversely:

$$f(x)g(x) \rightleftarrows \tilde{f}(s) * \tilde{g}(s)$$

- In higher dimensions:

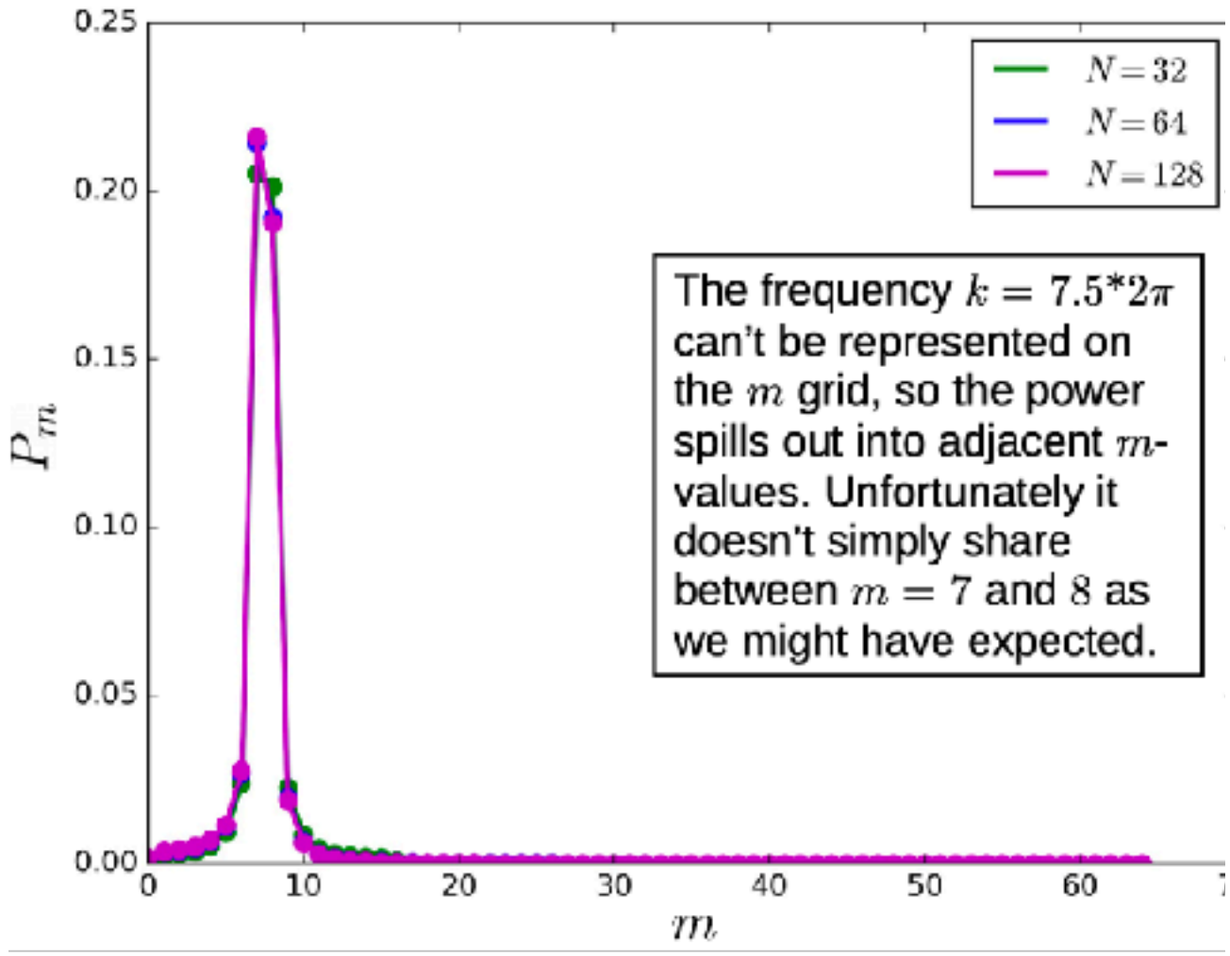
$$H(\mathbf{r}) = F(\mathbf{r}) * G(\mathbf{r}) = \iint F(\boldsymbol{\rho})G(\mathbf{r}-\boldsymbol{\rho})d\boldsymbol{\rho}$$

$$\tilde{H}(\mathbf{w}) = \tilde{F}(\mathbf{w})\tilde{G}(\mathbf{w})$$

(Léna et al. 2012)

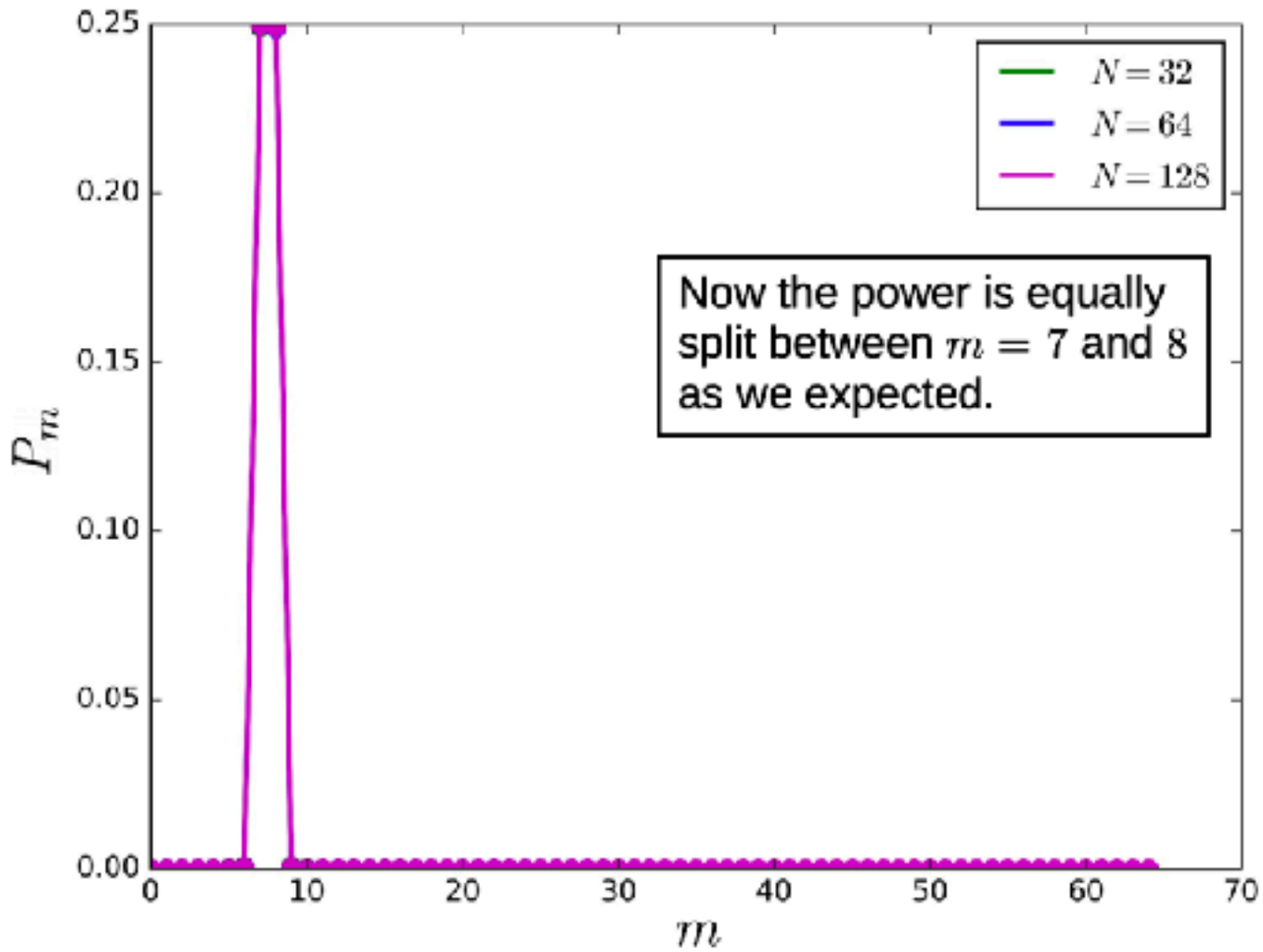
THE POWER SPECTRUM E.G.

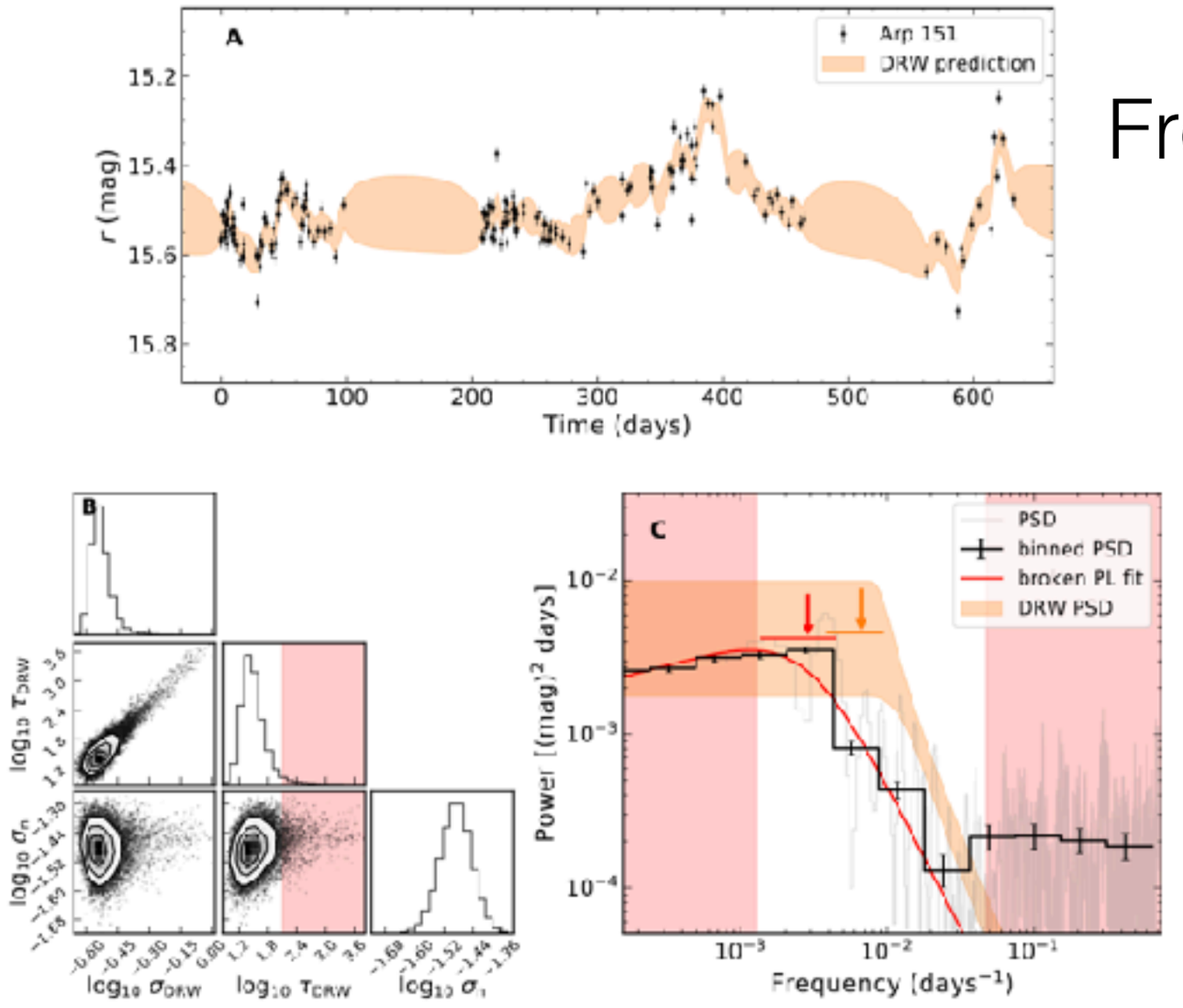
$f(x) = \sin(2\pi x \cdot 7.5)$ $\Delta x = 1/N$ $N = 32, 64,$



$f(x) = \sin(2\pi x \cdot 7.5)$ $\Delta x = 1/N$ $N = 32, 64, 128$

with Welch filter





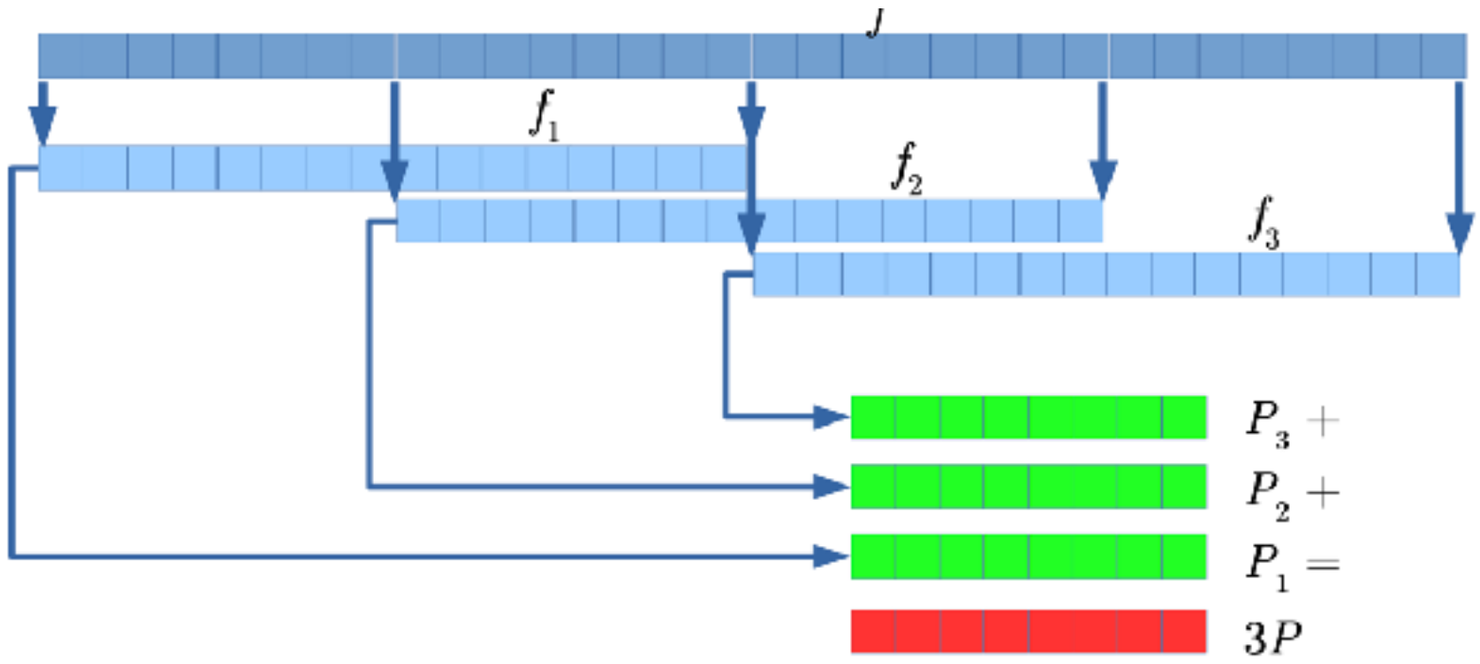
From Burke, Shen, Blaes, Gammie et al. 2022

Figure S5: **Example DRW modeling of Arp 151.** Panel A shows the r-band light curve of Arp 151 ($M_{\rm BH} = 10^{6.67\pm0.05} M_\odot$) and the best-fitting DRW model with 1σ uncertainty (orange shaded area). Panel B shows the posterior probability distributions for the fitted DRW parameters and their covariance. In the covariance panels, the contours trace the 1, 2, 3σ levels overplotted on the sample density map (black being higher density) with individual samples in the lowest-density regions shown as black points. Panel C shows the normalized PSD and binned PSD with 1σ uncertainties. The best-fitting broken power-law model is shown as a red line. The 1σ range of the DRW PSD from the posterior prediction is the orange shaded area. The corresponding break frequency $f_{\rm br}$ (from the broken power law fit) and $1/(2\pi\tau_{\rm DRW})$ (from the DRW fitting) are shown as equivalently colored arrows with the line segment below indicating the 1σ unceratinty. The red shaded regions correspond to periods greater than 20% the light curve length (in panels B and C) and less than the mean cadence (panel C), where the PSD is not well sampled.

POWER SPECTRUM ESTIMATION

- To improve the variance, can break data into shorter segments, FFT each segment, then average the power of the segments. This works best if the segments overlap by 1/2.
- For example: N = 32, three segments

For K segments, the variance is reduced by a factor of 9K/11.



- You don’t get something for nothing though - by breaking up the time-series into shorter segments, you can’t probe very long time periods i.e. high frequencies

ASTROPY AND TIME SERIES

- Astropy has a time series analysis module (`astropy.timeseries`) that provides tools to make power spectrum analysis (as well as other time series tasks) straightforward. For example:

```
from astropy.timeseries import TimeSeries, BoxLeastSquares
from astropy import units as u
import matplotlib.pyplot as plt

ts = TimeSeries.read(myfile) # contains time and flux as
columns
periodogram = BoxLeastSquares.from_timeseries(ts, 'flux')
P = periodogram.autopower(0.2 * u.day) # use a box width of
0.2 day

plt.plot(P.period, P.power, 'k-')
plt.show()
```

- A `TimeSeries` object is a specialized `Table` object. There are different types of periodogram estimators you can use. See the `astropy.timeseries` documentation for some worked examples.

Correlation

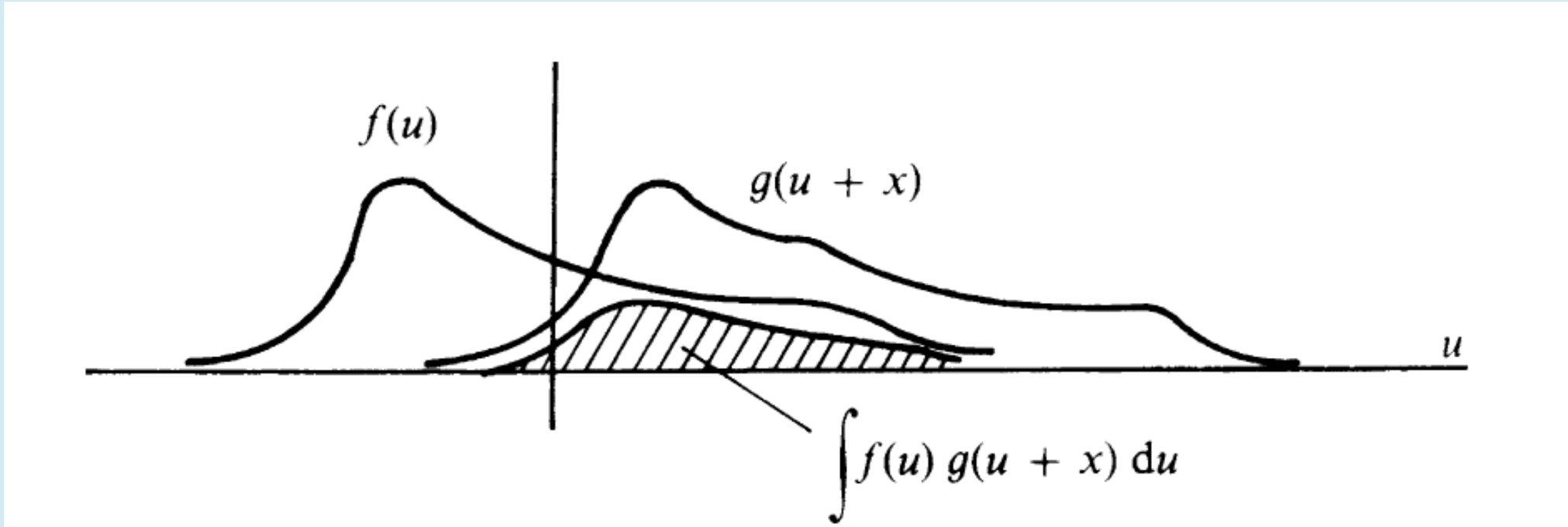
- The ***cross-correlation*** of two functions ***f(x)*** and ***g(x)*** is defined as:

$$f \otimes g = \int_{-\infty}^{\infty} f^*(u)g(u+x)\,du = \int_{-\infty}^{\infty} f^*(u-x)g(u)\,du$$

- Compare ***convolution*** (beware that two different but similar symbols are used for these two operations):

$$f(x) * g(x) = \int_{-\infty}^{+\infty} f(u)g(x-u)du$$

Graphical depiction of correlation



(Léna et al. 2012)

- The ***auto-correlation*** of function ***f(x)*** is defined as:

$$f \otimes f = \int_{-\infty}^{\infty} f^*(u) f(u+x)\, du$$

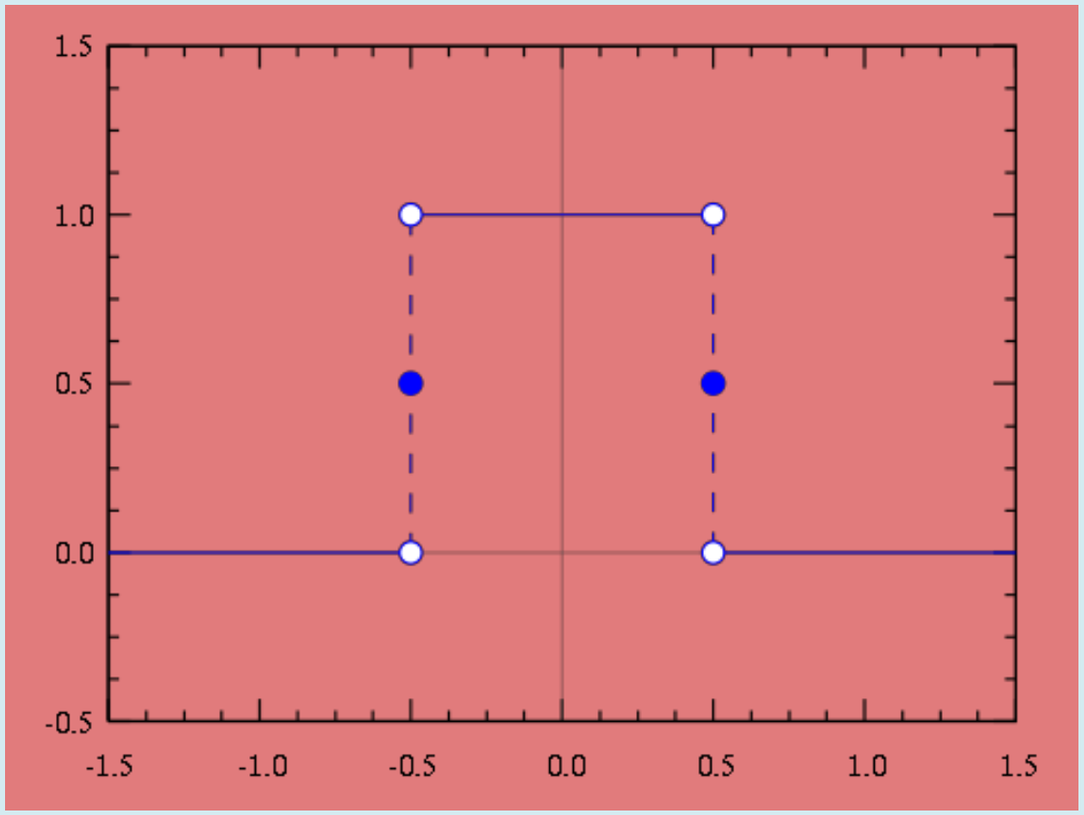
VERY LOW LIKELIHOOD OF US BEING ABLE TO GET PAST HERE ON THURSDAY

SPECIAL FUNCTIONS IN 1-D

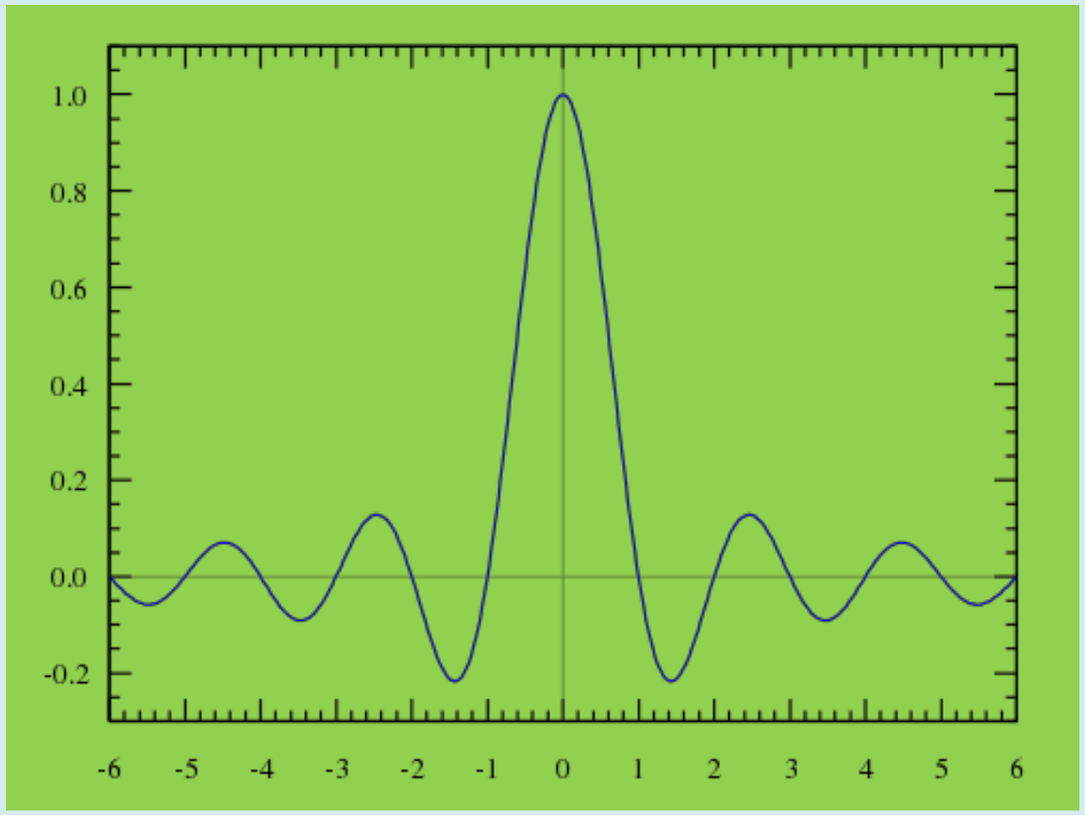
Special functions in 1-D used in Fourier transform analysis.

Boxcar and sinc functions

Boxcar function $\Pi(x)$



Sinc function $\frac{\sin(\pi s)}{\pi s}$



(CC)

Q: Is this spectrum shape intuitively sensible?

The Dirac delta function

- Mathematically, a ***Dirac delta function $\delta(x)$*** has properties:

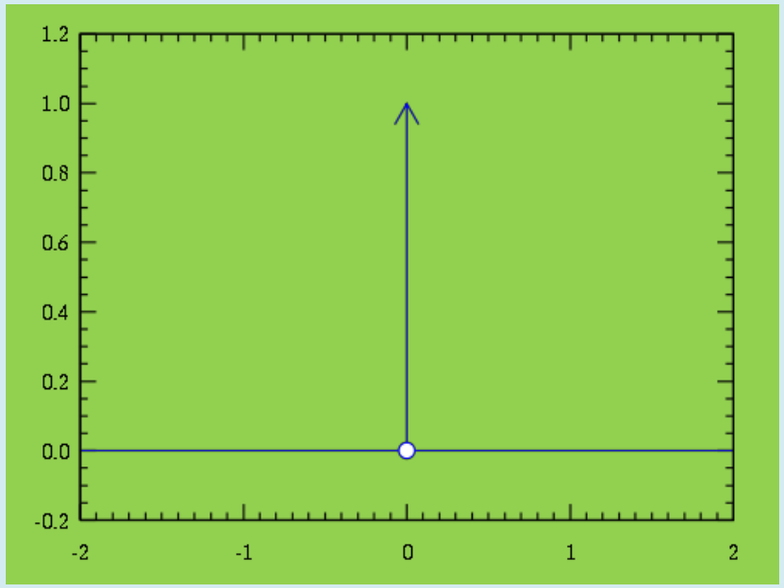
$$\delta(x - x_0) = 0 \text{ if } x \neq x_0$$

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

$$\int_{-\infty}^{\infty} f(x)\delta(x - x_0)dx = f(x_0)$$

$$\Im\big(\delta(x)\big) = 1$$

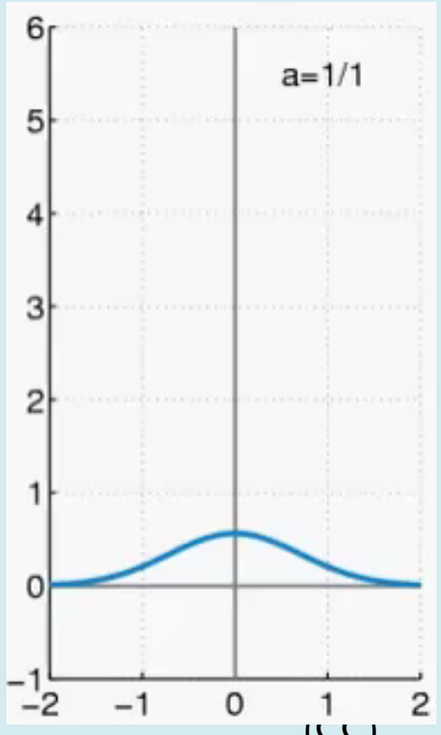
Why is this ?



Not a particularly well-behaved function

- Also known as an ***impulse function.***
- A ***generalized function*** or ***distribution:*** considered adjacent as a limiting sequence of Gaussian functions.

$$\text{Limit a} \to 0,\ \delta_a = \frac{1}{a\sqrt{\pi}} e^{-\frac{x^2}{a^2}}$$



Dirac combs

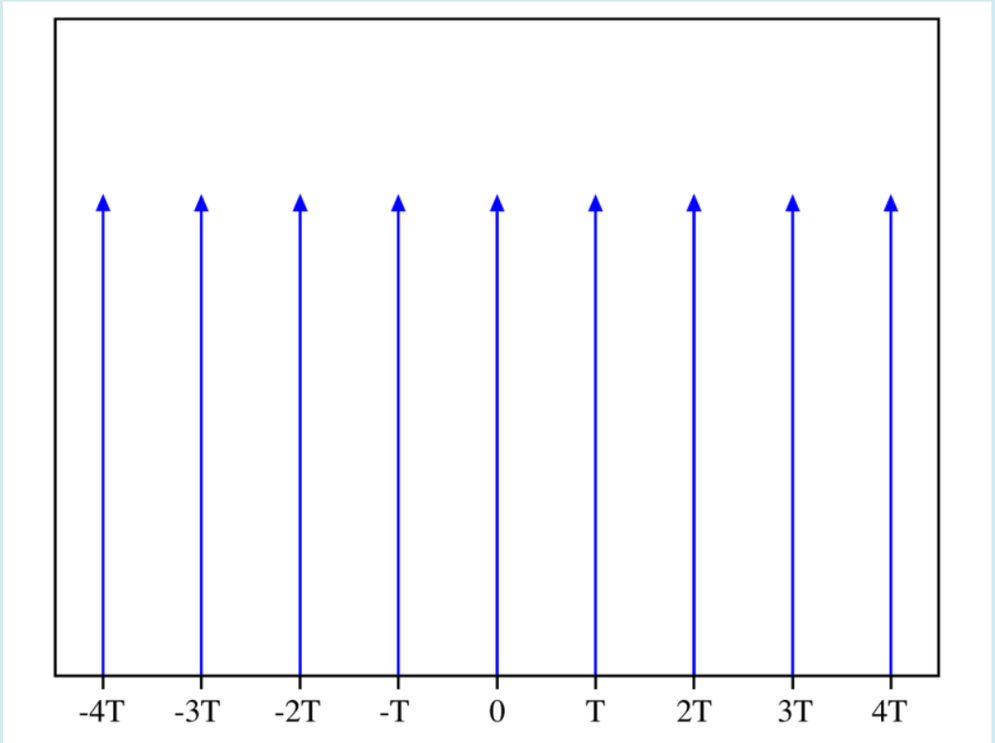
- A ***Dirac delta function δ(t)*** can be generalized to a Dirac comb function:

$$III(x) = \sum_{n=-\infty}^{\infty} \delta(x-n) \qquad \text{(unit spacing)}$$

$$III\left(\frac{x}{T}\right) = T\sum_{n=-\infty}^{\infty} \delta(x-Tn) \qquad \text{(spaced by } T\text{)}$$

- Also known as a ***shah function***.
- Very useful in modeling discrete sampling of continuous functions.
- The Fourier transform of a Dirac comb function is itself a Dirac comb function:

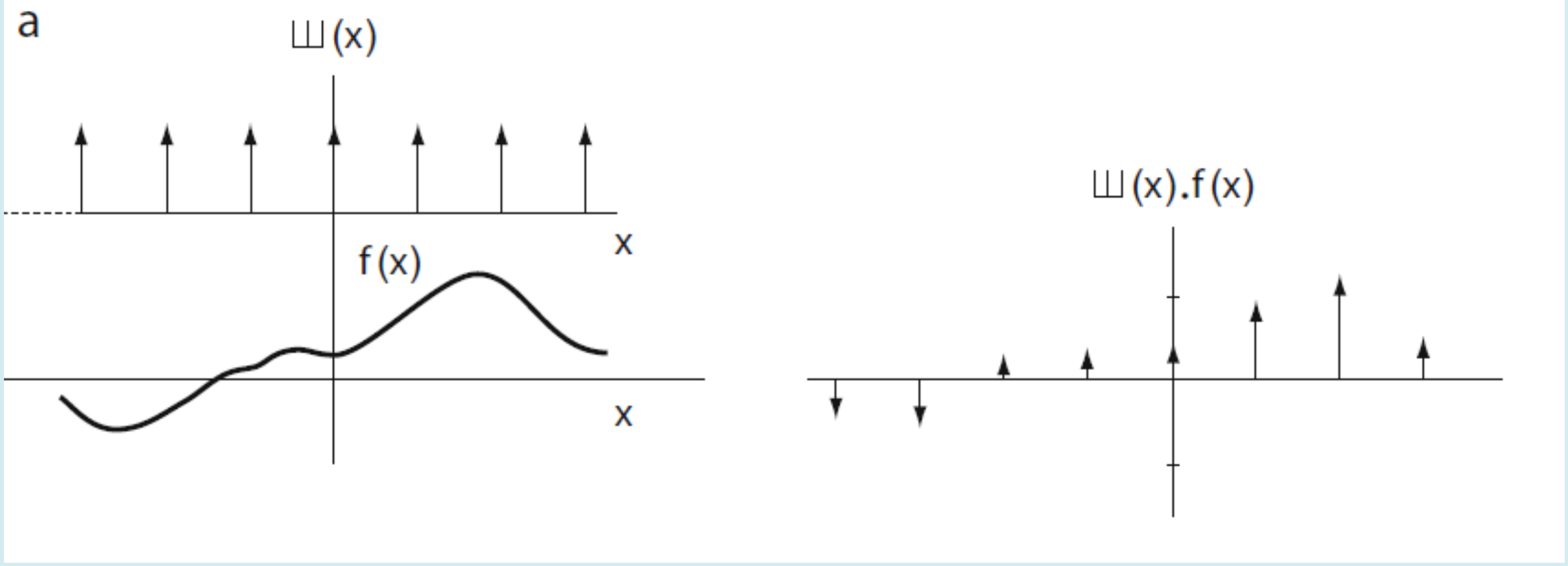
$$III(x) \rightleftarrows III(s)$$



(CC)

Q: What happens to the Fourier transform as T is decreased ?

Dirac comb as a sampling function



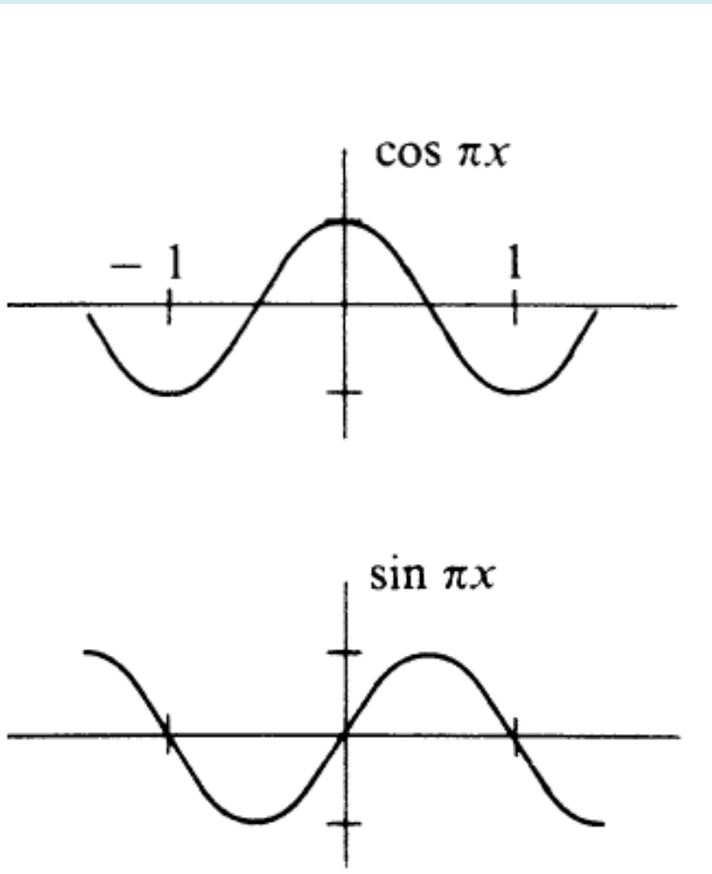
(Léna et al. 2012)

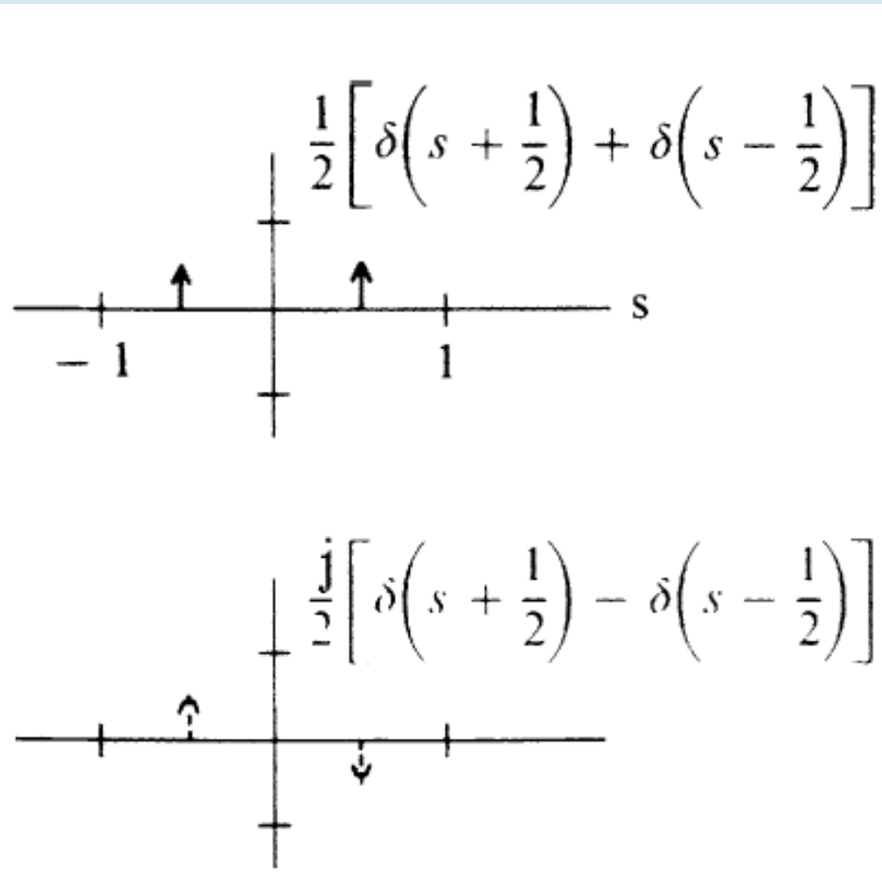
- Function $f(x)$ sampled at unit spacing.

Q: In what part of our observing system framework is this likely to be relevant?

Trigonometric functions

- Trigonometric functions are not square integrable: $\int_{-\infty}^{\infty} |f(x)|^2 \, dx$
- Their Fourier transforms are ***distributions*** (***generalized functions***).





(Léna et al. 2012)

Symbols for common Fourier transform functions

Table 4.3: Symbols for Common Functions

boxcar, top hat, or rectangle	$\Pi(\mathrm{x})$	$= \begin{cases} 1 & \lvert x\rvert < \frac{1}{2} \\ 0 & \lvert x\rvert > \frac{1}{2} \end{cases}$
triangle	$\Lambda(\mathrm{x})$	$= \begin{cases} 1 - \lvert x\rvert & \lvert x\rvert < 1 \\ 0 & \lvert x\rvert > 1 \end{cases}$
Heaviside step function	$\mathrm{H}(\mathrm{x})$	$= \begin{cases} 1 & x > 0 \\ 0 & x < 0 \end{cases}$
even impulse pair	$\mathrm{II}(\mathrm{x})$	$= \frac{1}{2}\,\delta(\mathrm{x} + \frac{1}{2}) + \frac{1}{2}\,\delta(\mathrm{x} - \frac{1}{2})$
odd impulse pair	$\mathrm{I_I}(\mathrm{x})$	$= \frac{1}{2}\,\delta(\mathrm{x} + \frac{1}{2}) - \frac{1}{2}\,\delta(\mathrm{x} - \frac{1}{2})$
comb (or shah)	$\mathrm{III}(\mathrm{x})$	$= \sum_{\mathrm{n}=-\infty}^{\infty} \delta(\mathrm{x} - \mathrm{n})$
sinc	$\mathrm{sinc}(\mathrm{x})$	$= \frac{\sin \pi \mathrm{x}}{\pi \mathrm{x}}$

(Sutton 2011)

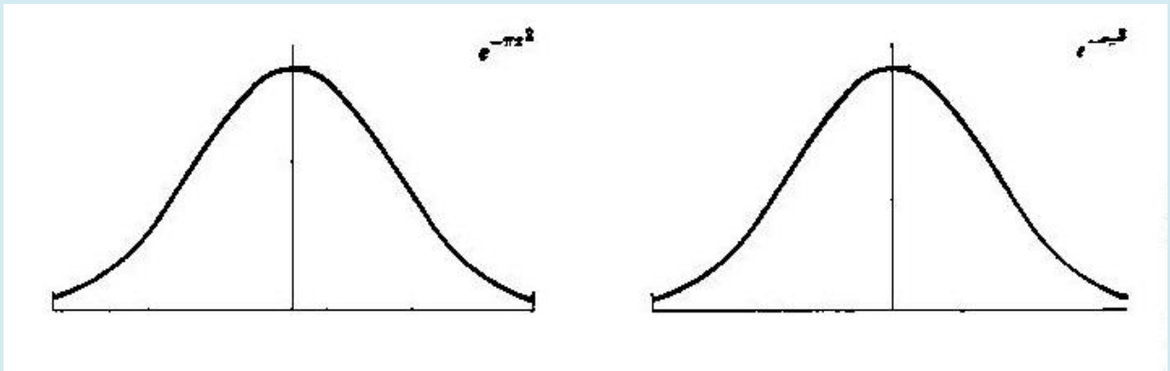
Fourier transform pairs: dictionary

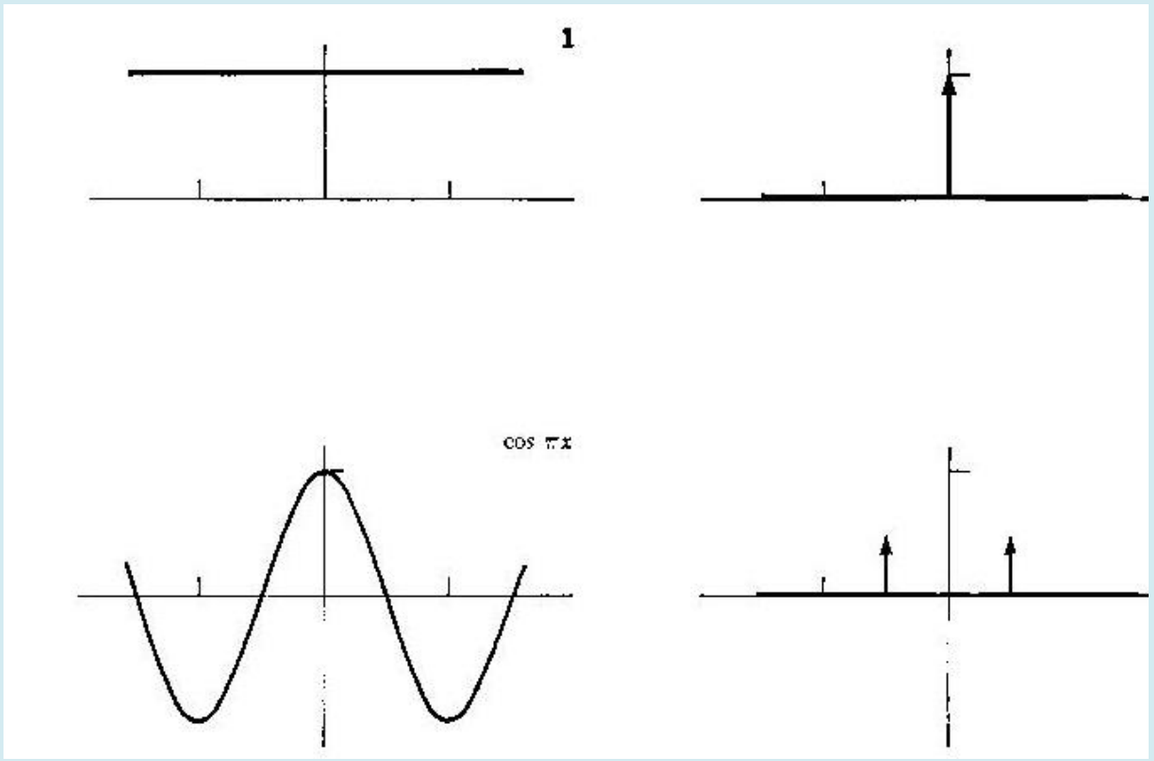
Table 4.4: Fourier Transform Pairs

$e^{-\pi x^2}$	$\rightleftharpoons$	$e^{-\pi s^2}$
1	$\rightleftharpoons$	$\delta(s)$
$\Pi(x)$	$\rightleftharpoons$	$\mathrm{sinc}(s)$
$\Lambda(x)$	$\rightleftharpoons$	$\mathrm{sinc}^2(s)$
$\cos \pi x$	$\rightleftharpoons$	$I I(s)$
$\sin \pi x$	$\rightleftharpoons$	$i\, I_I(s)$
$III(x)$	$\rightleftharpoons$	$III(s)$
$H(x)$	$\rightleftharpoons$	$\frac{1}{2}\,\delta(s) - \frac{i}{2\pi s}$

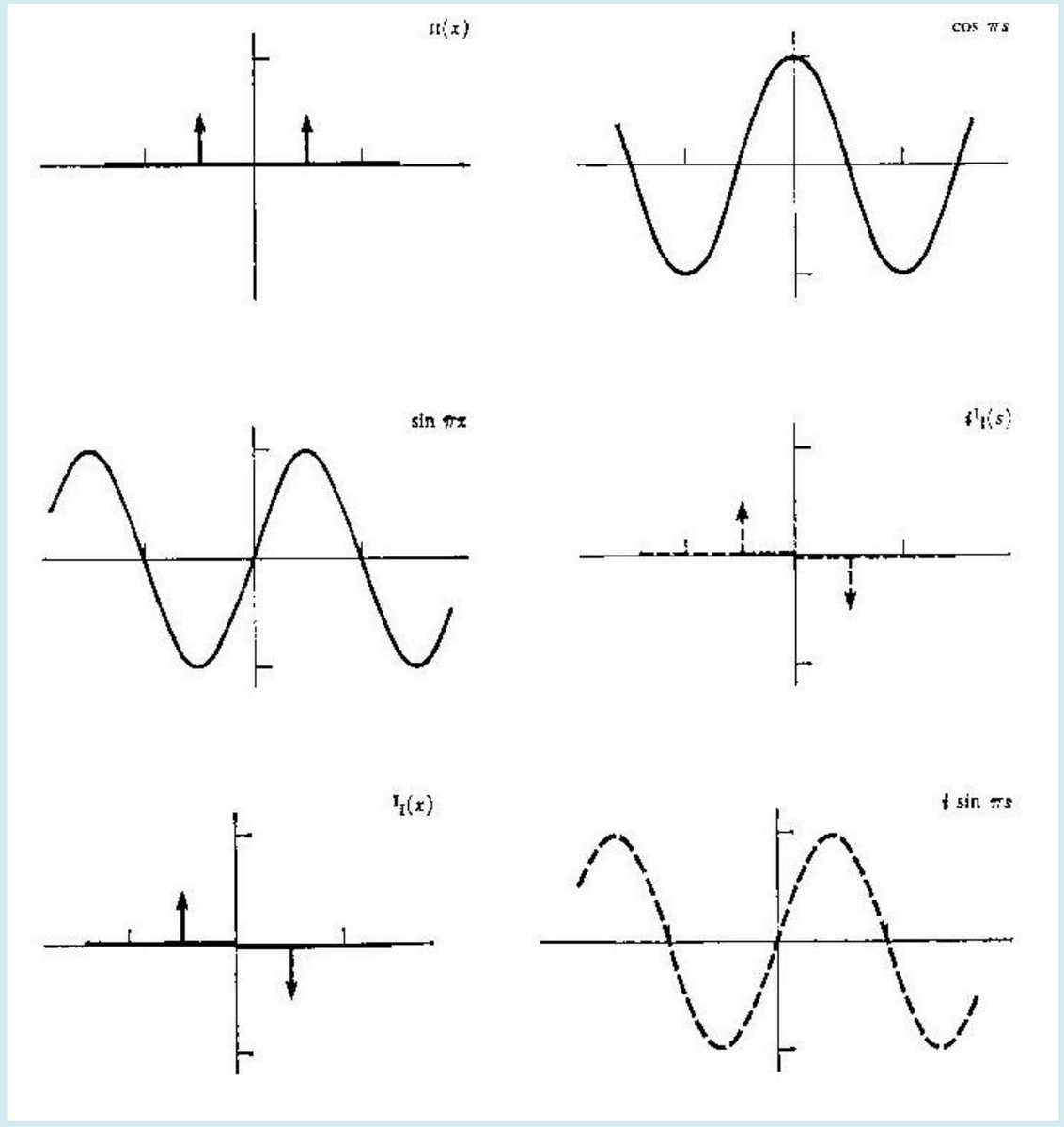
(Sutton 2011)

Pictorial dictionary of Fourier transforms

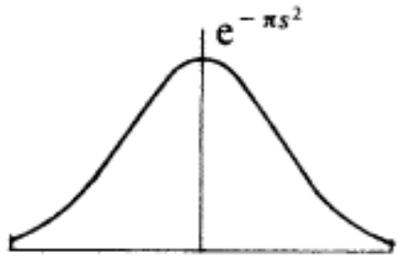


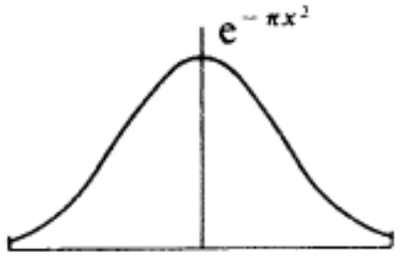


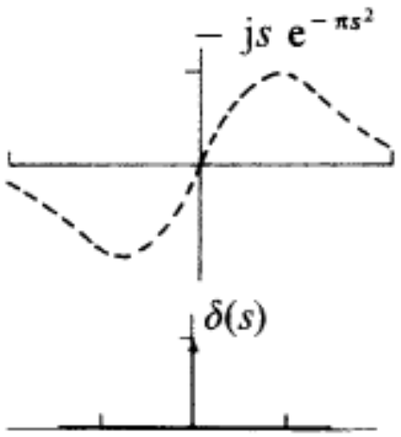
(Bracewell 1986)

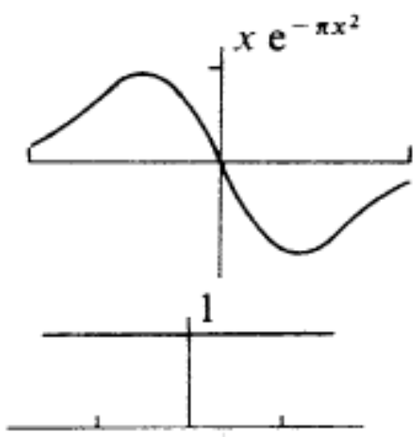


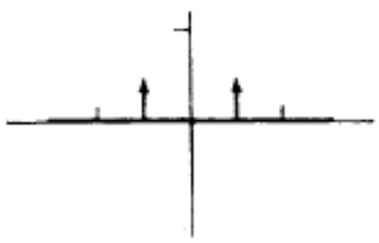
(Bracewell 1986)

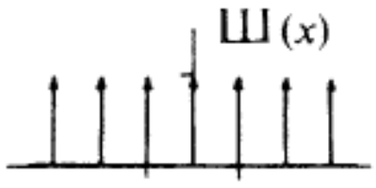


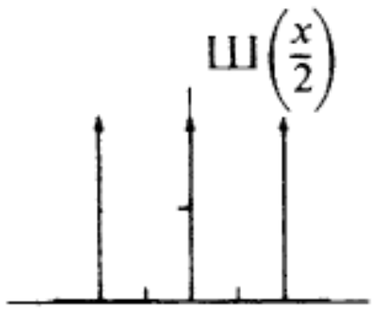


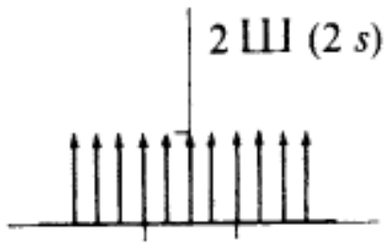


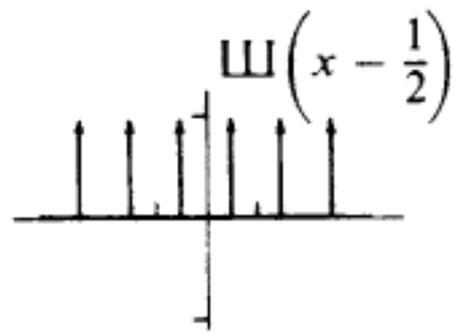


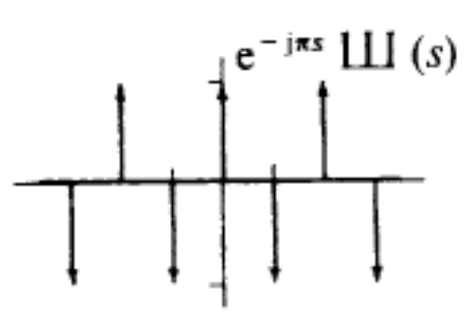


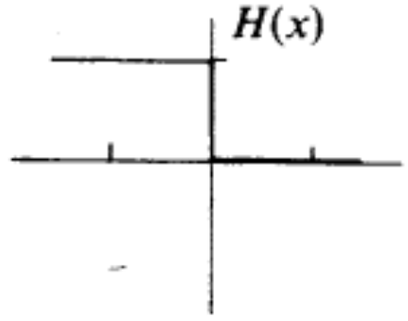


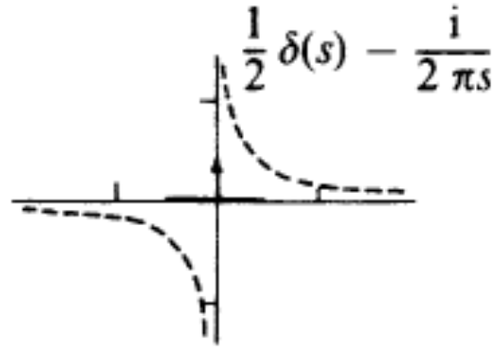


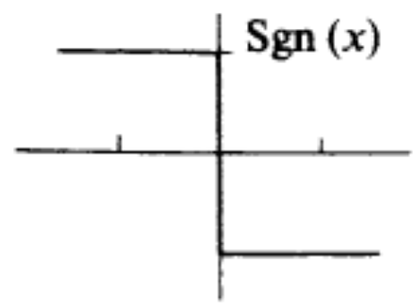


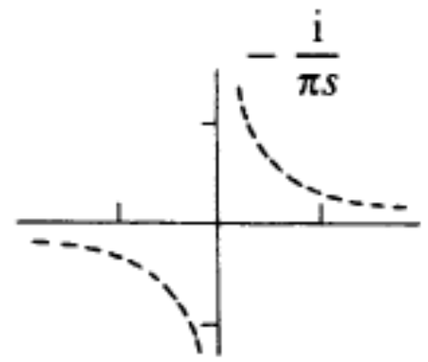


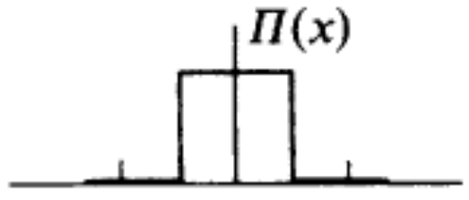


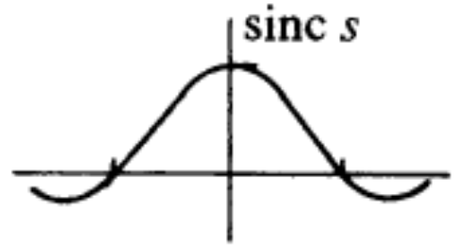


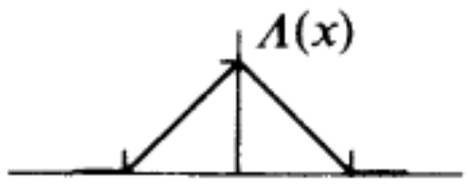


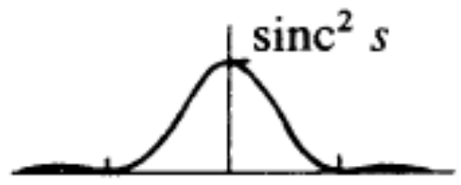












(Bracewell 1965; Léna et al. 2012)

SPECIAL FUNCTIONS IN 2-D

Special functions in 2-D used in Fourier transform analysis.

Two-dimensional Fourier transforms

From earlier vector form,

$$\tilde{f}(\mathbf{w}) = \int_{-\infty}^{\infty} f(\mathbf{r})\, e^{-2\pi i \mathbf{r}.\mathbf{w}}\, d\mathbf{r}$$

$$\mathbf{r} = (x, y) \in \mathbb{R}^2,\ \mathbf{w} = (u, v) \in \mathbb{R}^2$$

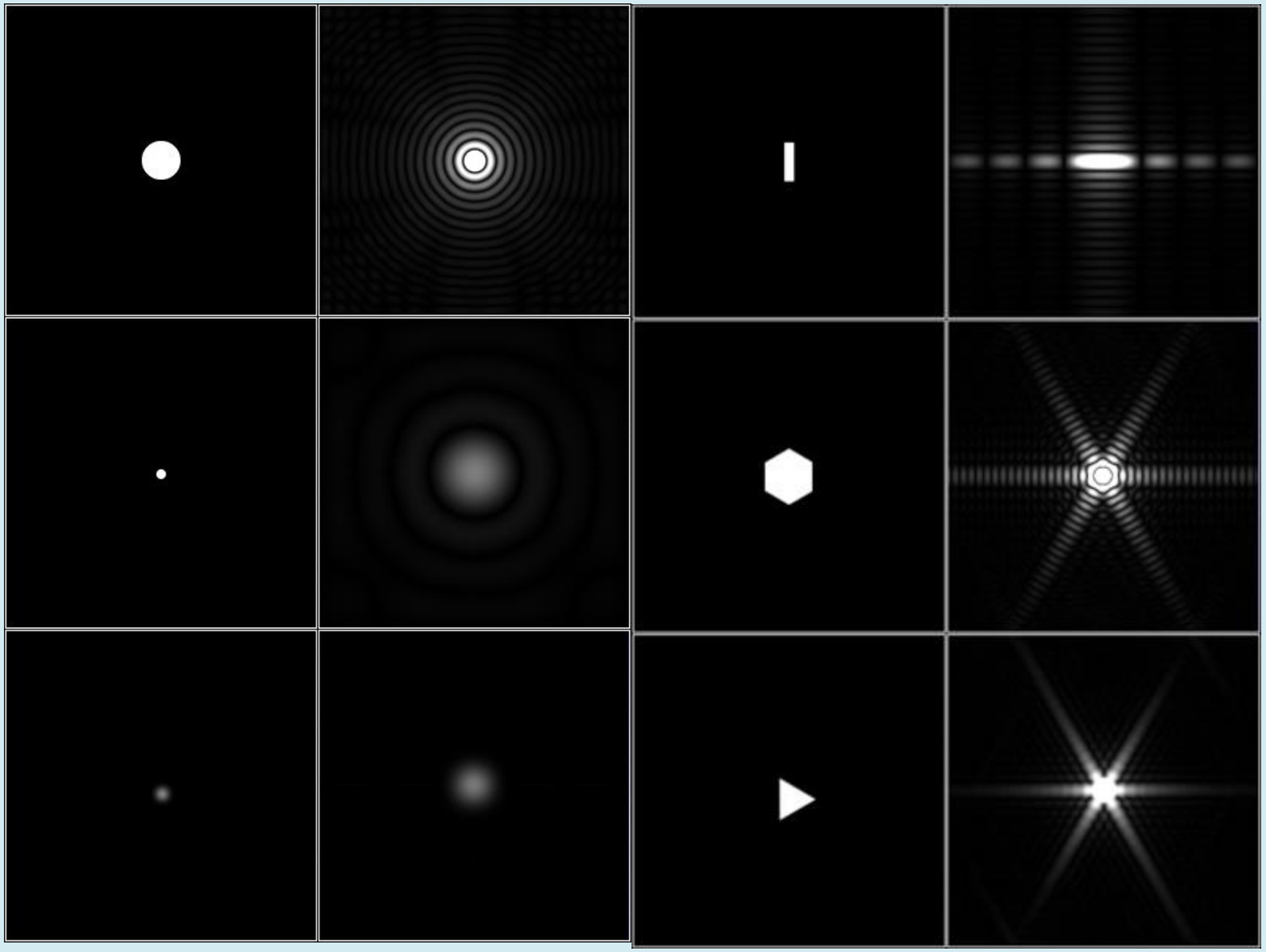
$$r^2 = x^2 + y^2$$

$$w^2 = u^2 + v^2$$

In coordinate form:

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tilde{f}(u,v)\, e^{-i2\pi(xu+yv)}\, du\, dv$$

$$\tilde{f}(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\, e^{+i2\pi(xu+yv)}\, dx\, dy$$

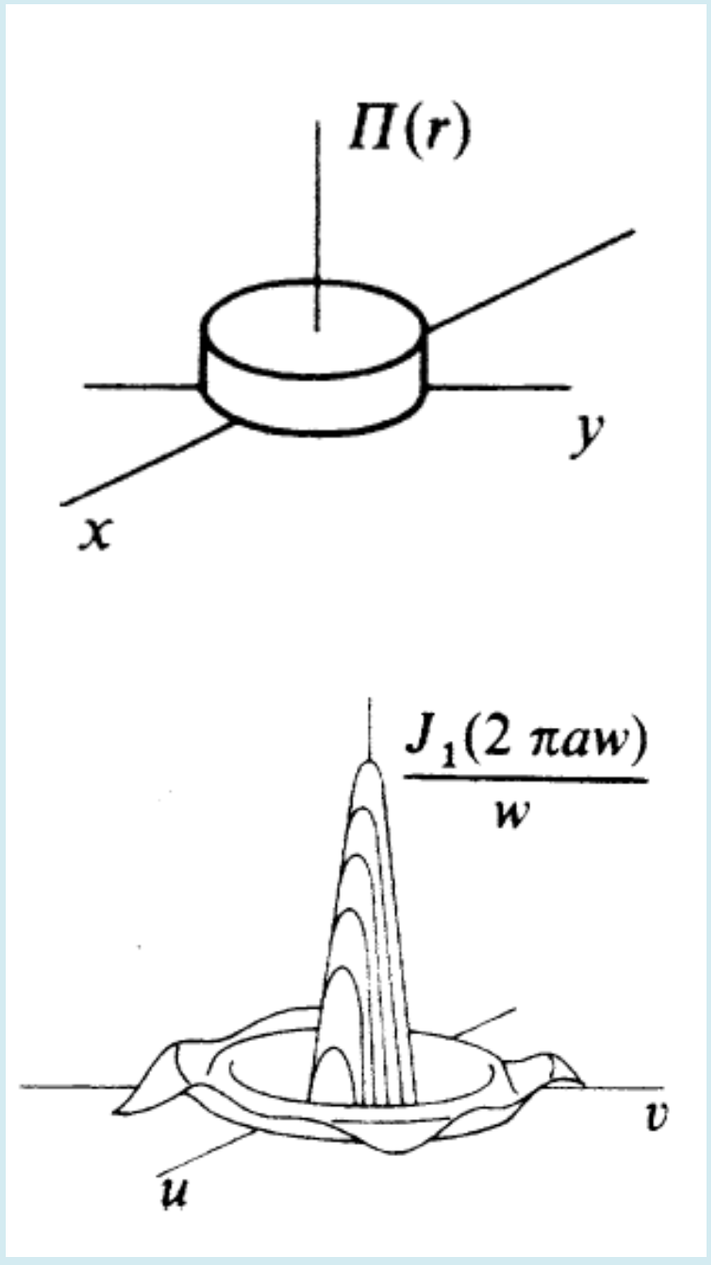


(CC)

Box function in two dimensions

- Constant inside the unit circle and zero outside:

$$\Pi\left(\frac{r}{2}\right) = \begin{cases} 1 \; r < 1 \,, \\ 0 \; r \geq 1 \,. \end{cases}$$



- Fourier transform:

$$\Pi\left(\frac{r}{2}\right) \rightleftarrows \frac{J_1(2\pi w)}{w}$$

- Similarity implies:

$$\Pi\left(\frac{r}{2a}\right) \rightleftarrows a\frac{J_1(2\pi a w)}{w}\,, \quad a > 0$$

J_1: Bessel function of order 1

(Léna et al. 2012)

Dirac distribution and sampling in 2-D

- Dirac distribution in 2-D:

$$\delta(x, y) = \delta(\boldsymbol{r}) = \iint_{\text{plane}} \mathrm{e}^{2\mathrm{i}\pi \boldsymbol{r}\cdot\boldsymbol{w}}\, \mathrm{d}\boldsymbol{w}$$

- Two-dimensional sampling function:

$$\text{Ш}(x, y) = \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} \delta(x - m, y - n)$$

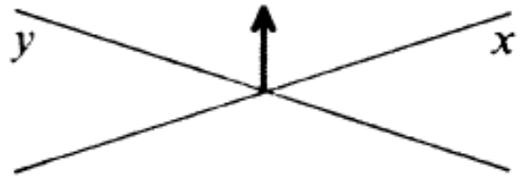
- With Fourier transform:

$$\text{Ш}(x, y) \rightleftarrows \text{Ш}(u, v)$$

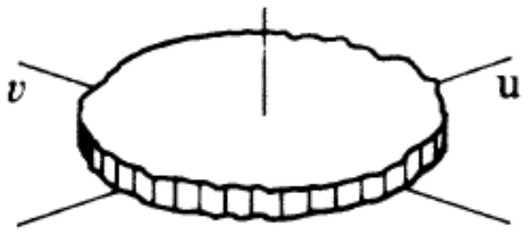
Q: What is an example from an observing system where this may be useful?

(Léna et al. 2012)

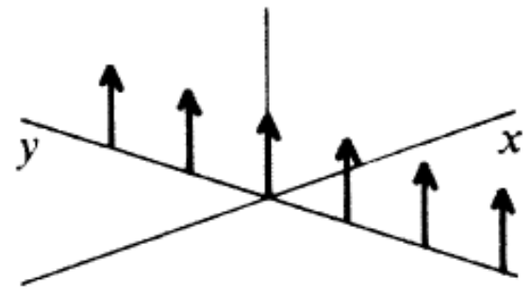
- 2-D pictorial dictionary of Fourier transforms.



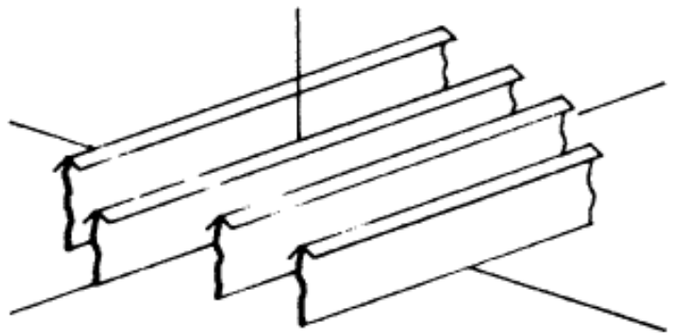
$\delta^2(x, y)$



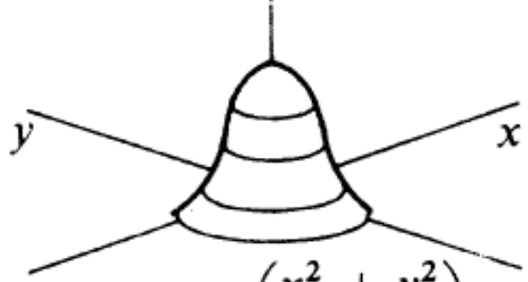
1



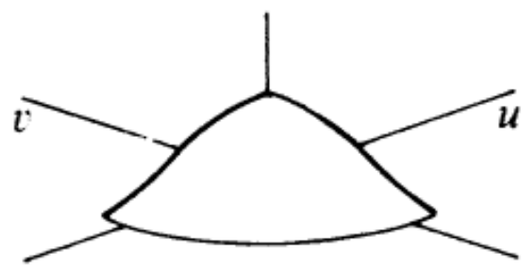
$\mathrm{III}(y)\,\delta(x)$



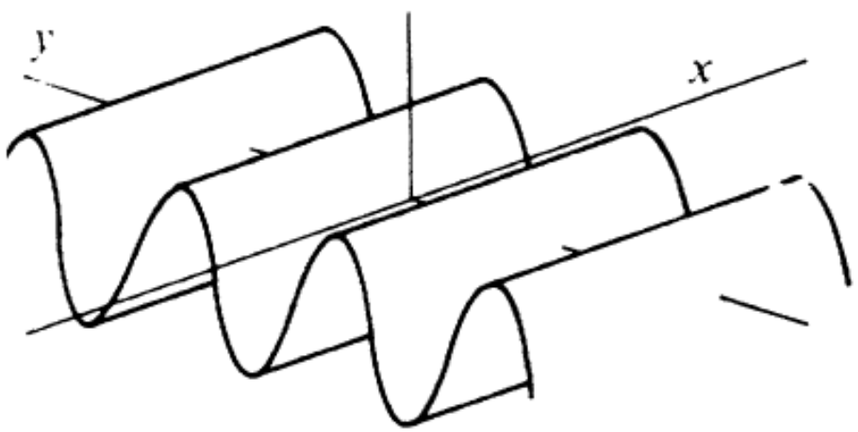
$\mathrm{III}(v)$



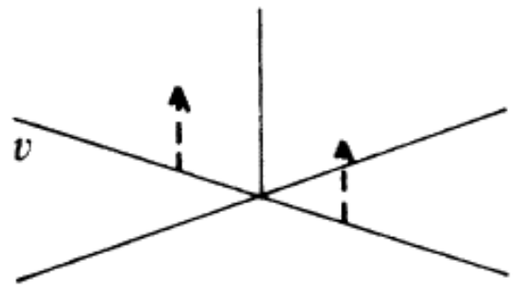
$\exp - \pi\left(\frac{x^2 + y^2}{a^2}\right)$



$a \exp - \pi a^2(u^2 + v^2)$



$\sin \pi y$



$\frac{i}{2}\,\delta^2\left(u, v-\frac{1}{2}\right) - \frac{j}{2}\,\delta^2\left(u, v+\frac{1}{2}\right)$

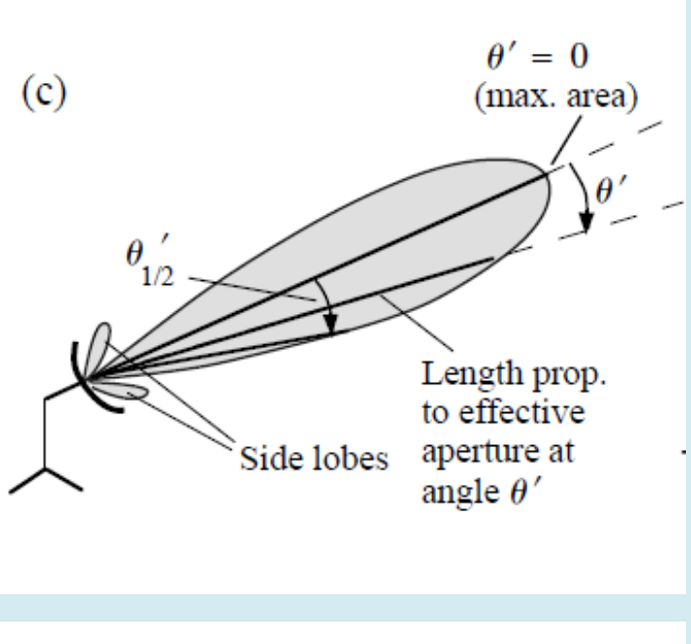
(Bracewell 1965; Léna et al. 2012)

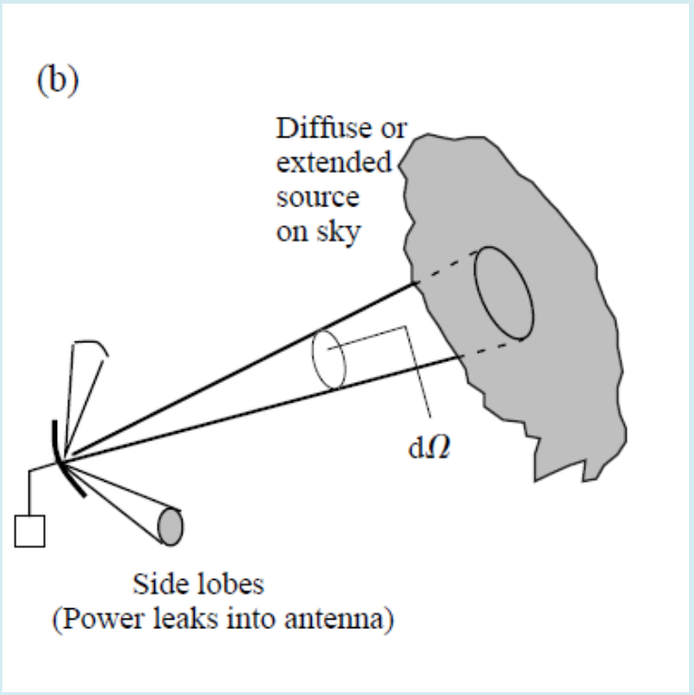
A preview: imaging as an example of convolution

- Consider pointing a telescope angular response pattern ***P(θ',φ',ν),*** toward a specific intensity distribution ***I(θ,φ,ν),*** on the sky at angular position ***(θ_0,φ_0)*** :

$$I_{obs}(\theta_0,\phi_0,\nu) = \frac{\int\int_{\theta\ \phi} I(\theta,\phi,\nu)P(\theta-\theta_0,\phi-\phi_0,\nu)\,d\Omega}{\int\int P(\theta,\phi,\nu)d\Omega}$$

- This is a ***convolution*** of functions ***P(θ,φ,ν)*** and ***I(θ,φ,ν)*** over their angular coordinates ***(θ,φ).***
- It reflects the fact that – ***even when pointing in a fixed direction*** – we sum power received from ***other*** directions on the sky toward which our telescope has a non-zero angular response.





Q: *How was this latter point discussed during the section on observing system components?*

(Bradt 2004)

Imaging as a convolution

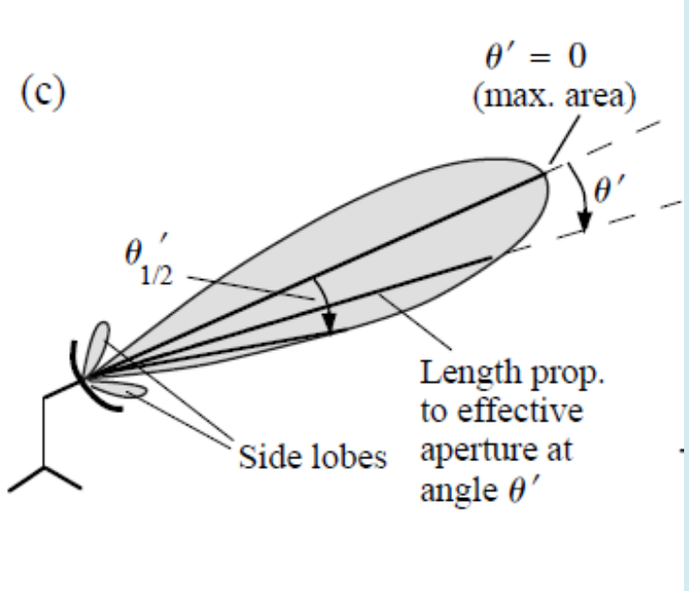
- Telescope angular response pattern ***P(θ',φ',ν)*** is pointed toward a specific intensity distribution ***I(θ,φ,ν),*** on the sky at angular position ***(θ_0,φ_0)*** :

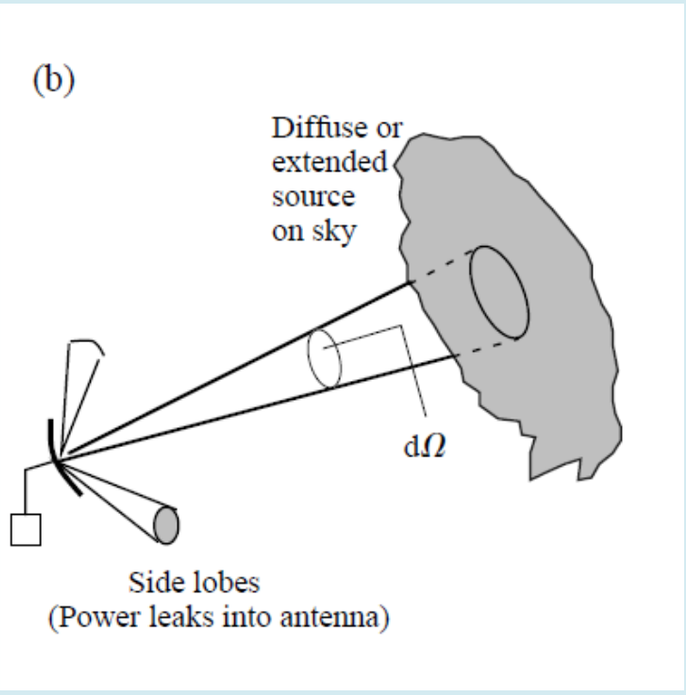
$$I_{obs}(\theta_0,\phi_0,\nu) = \frac{\int_\theta\int_\phi I(\theta,\phi,\nu)P(\theta-\theta_0,\phi-\phi_0,\nu)\,d\Omega}{\int\int P(\theta,\phi,\nu)d\Omega}$$

$I_{obs} = P * I$ over (θ,ϕ),

with normalization: $\int_\theta\int_\phi P(\theta,\phi,\nu)d\Omega = \Omega_A$,

and omitting coordinate projection effects.

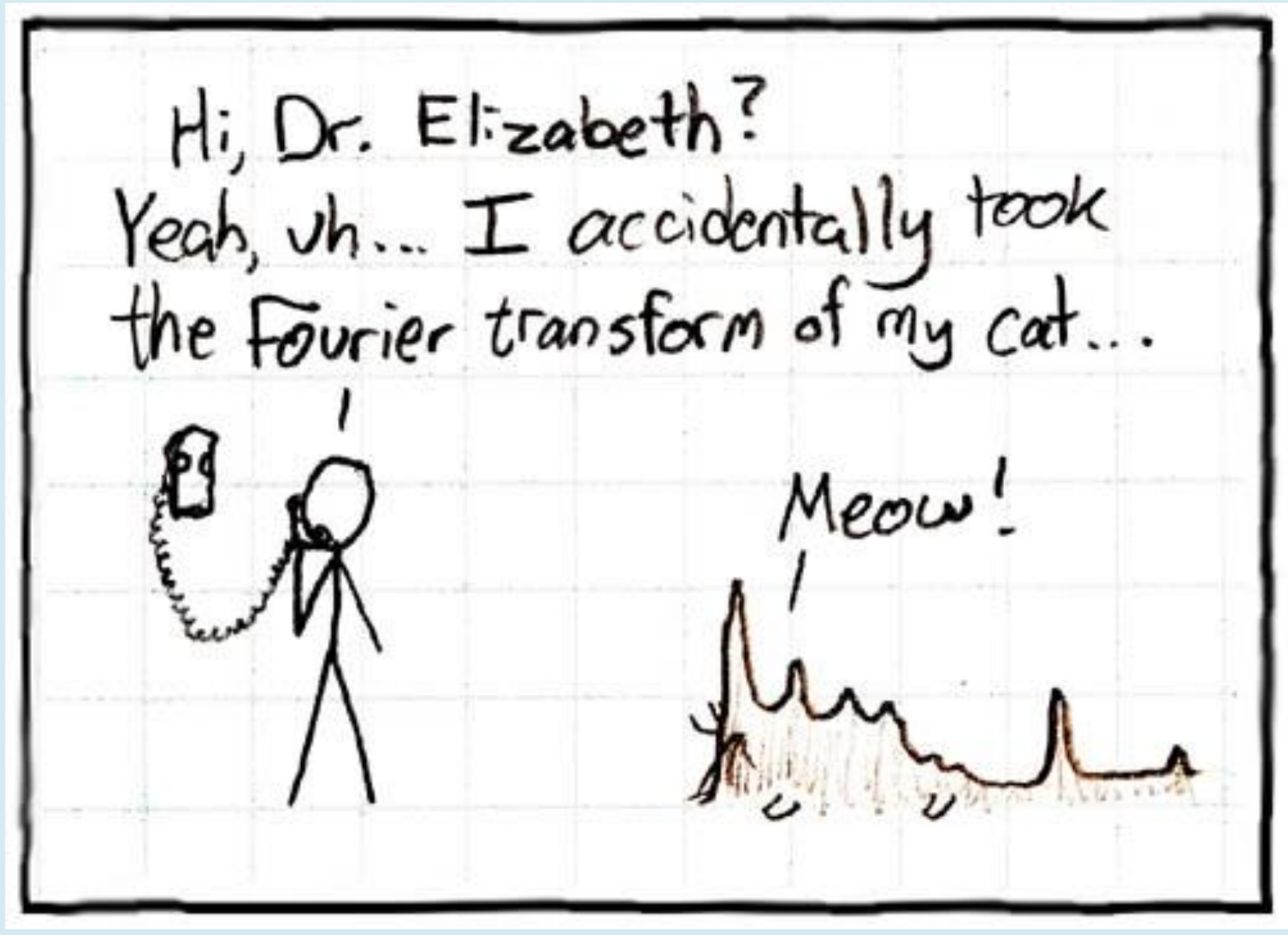




- ***Imaging is the convolution of the true specific intensity of the sky by a total intensity telescope angular response function***

(Bradt 2004)

A powerful tool – use carefully



(xkcd)

Noise

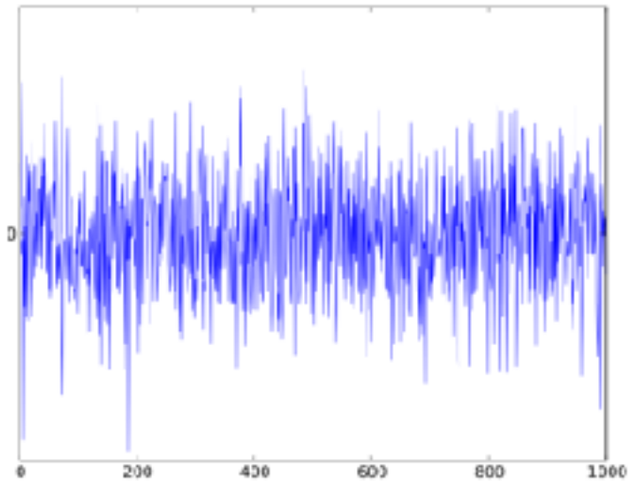
Noise

- **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

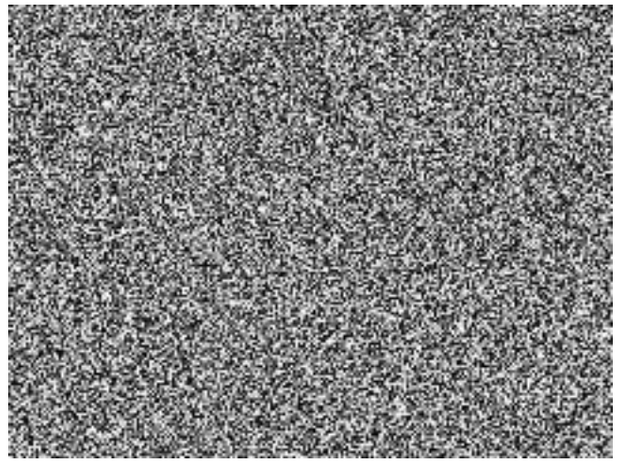
Photon Noise

Photon Noise

Photon Noise

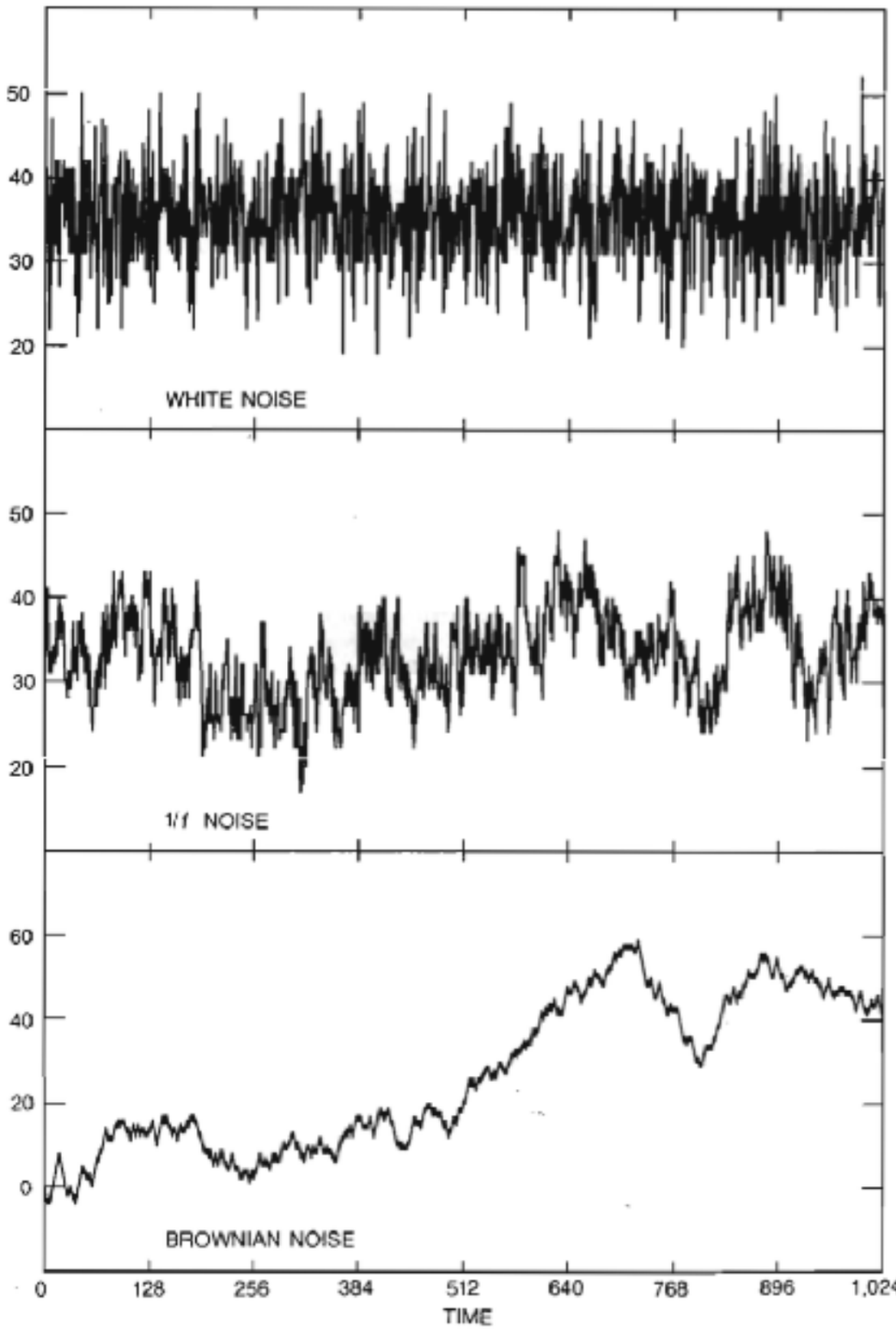


White Noise



- random signal with constant power spectral density
- each point is uncorrelated.
- Between 20 to 20,000 Hz, sounds like ssshhhhhh
- 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 $V_{noise} = \sqrt{4k_BTR}$
    - Voltage noise $V_{noise} = \sqrt{4k_BTR\Delta f}$ for RMS within a given bandwidth in $nV/\sqrt{Hz}$
    - power noise $P = V^2/R = 4k_BT\Delta f$
- Shot noise is Poisson

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



White or Poisson

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

stochastic

1/f or pink or flicker

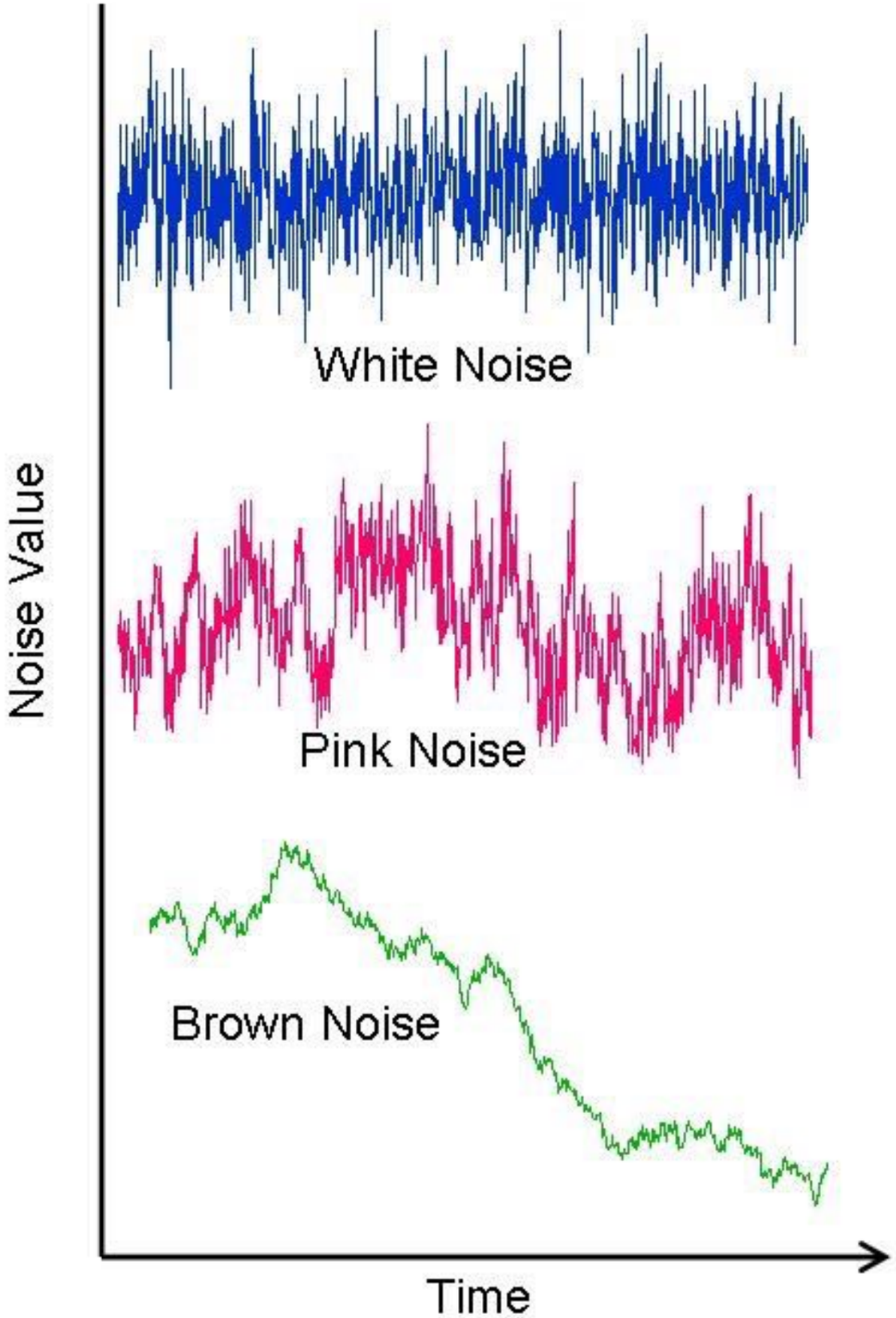
$$N_{1/\mathrm{f}}(f) \propto \frac{1}{f^1} = \frac{1}{f}$$

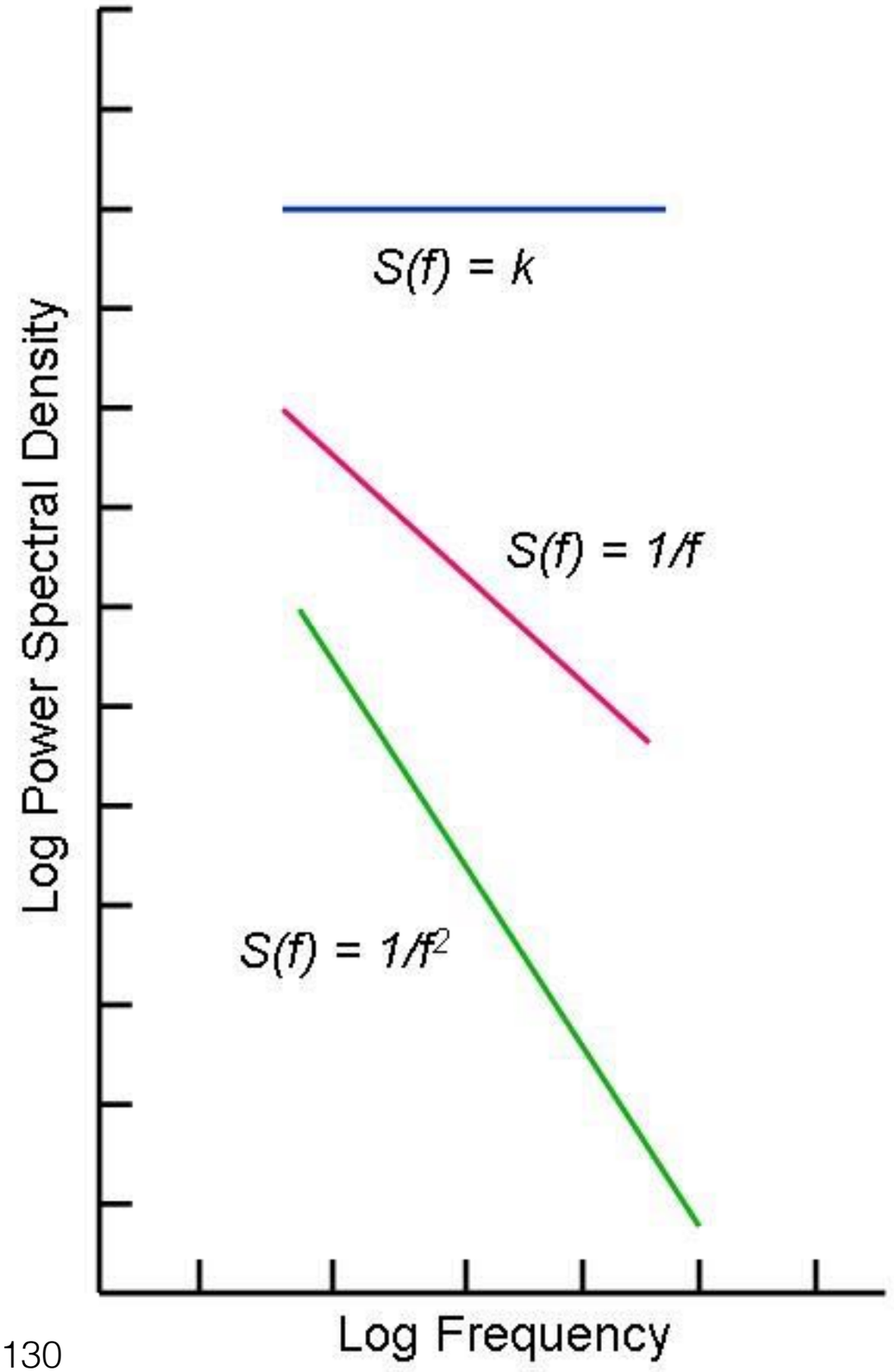
occurs in nature

Brownian

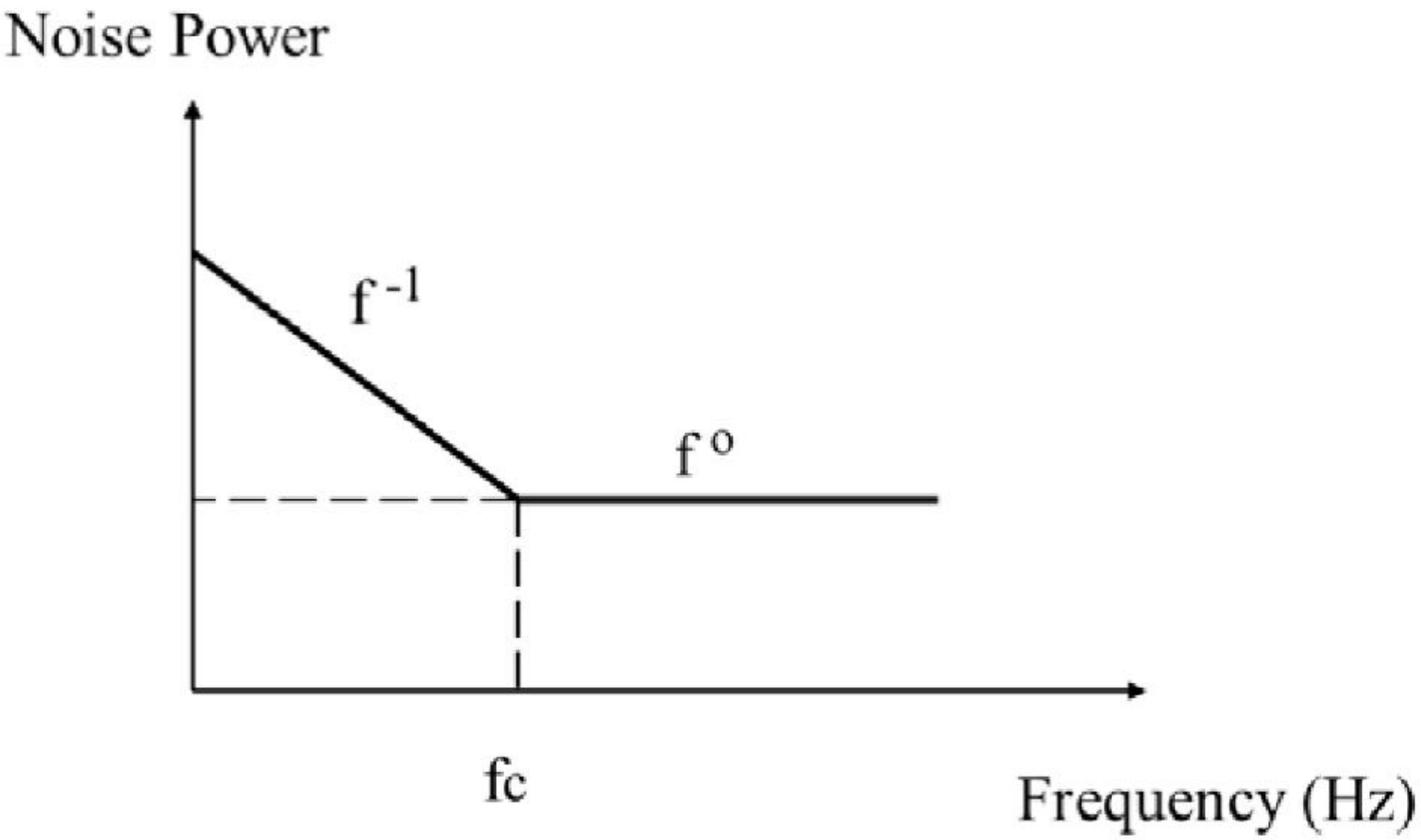
$$N_{\mathrm{Brownian}}(f) \propto \frac{1}{f^2}$$

Cumulative integral of white noise

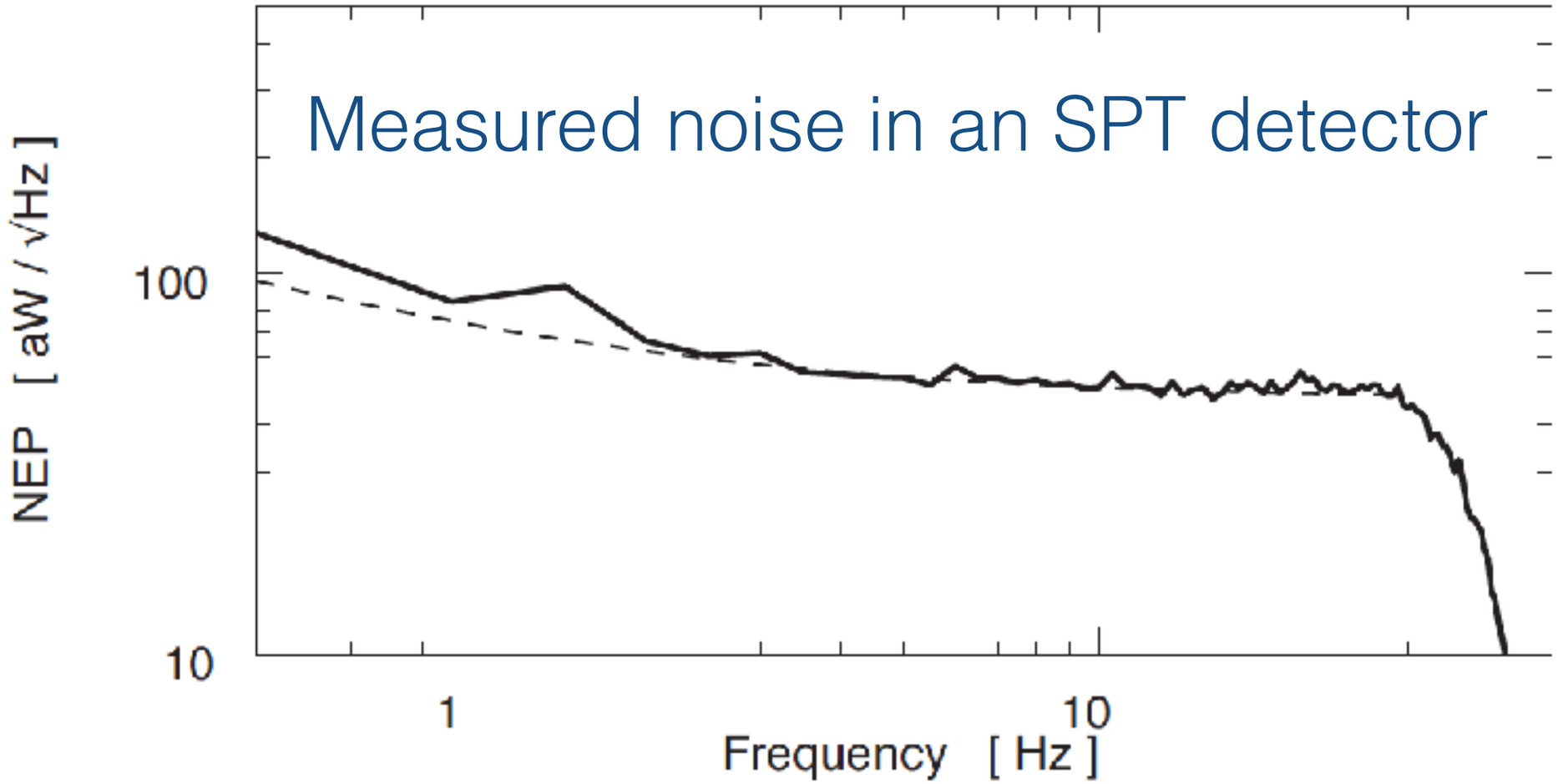

Noise Value
White Noise
Pink Noise
Brown Noise
Time

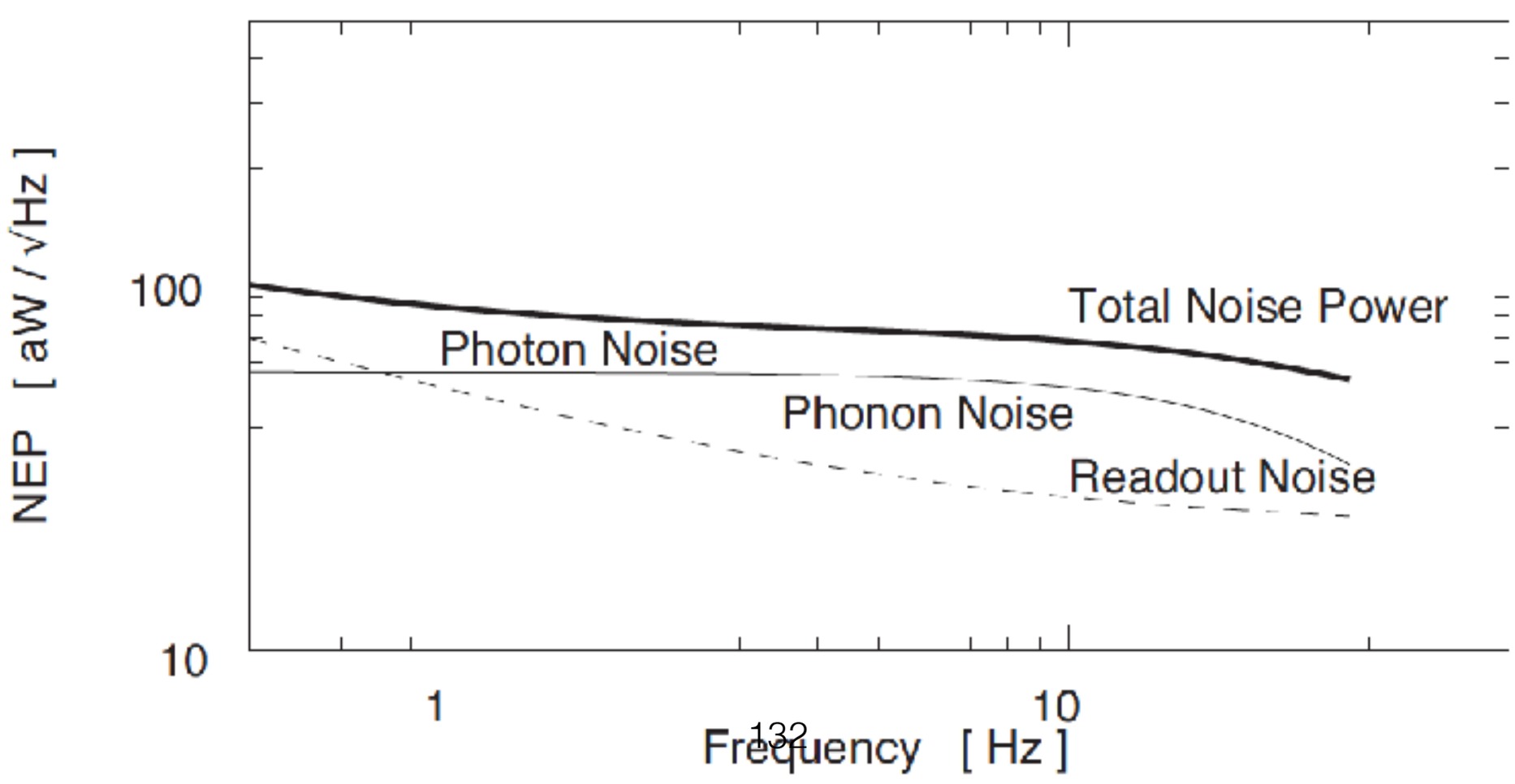

Log Power Spectral Density
S(f) = k
S(f) = 1/f
S(f) = 1/f2
Log Frequency

1/f noise



Measured noise in an SPT detector





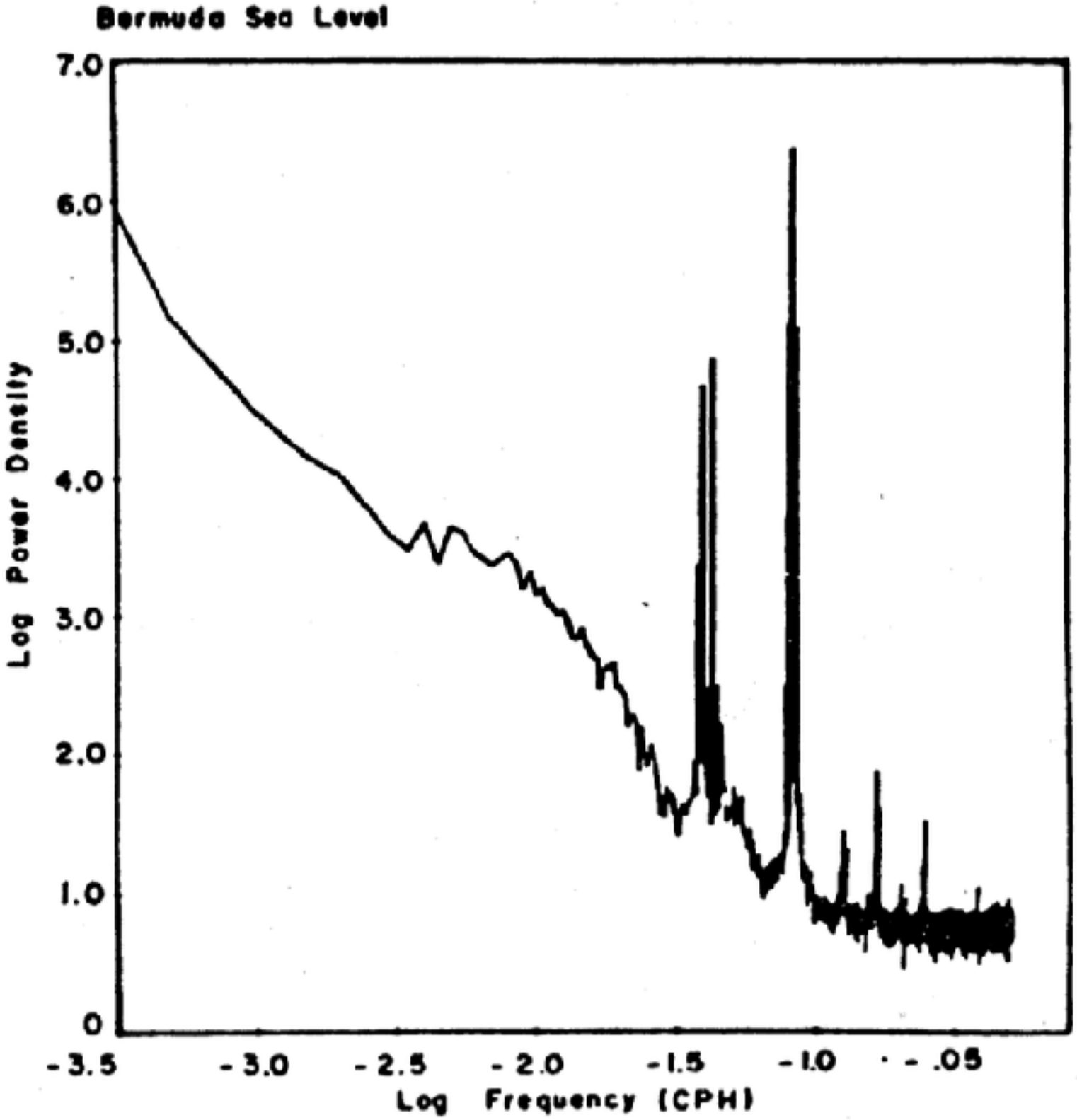
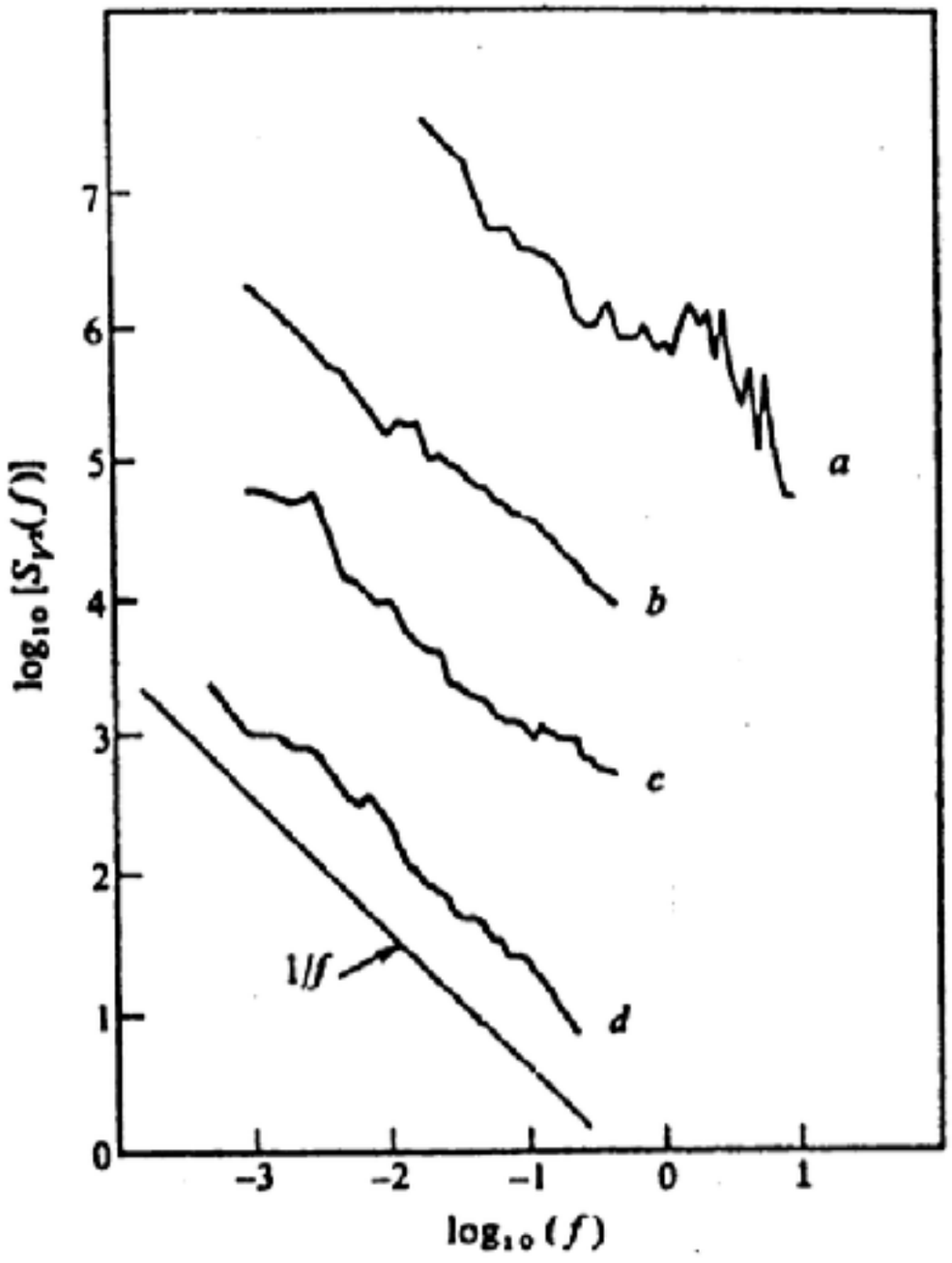


FIGURE 8 Power spectrum of sea level at Bermuda. Note the strong tidal components at 12 hours and 24 hours. The rise of power at low frequencies is here steeper than $1/f$ (and is close to $1/f^2$). (Reproduced from Wunsch.[25])



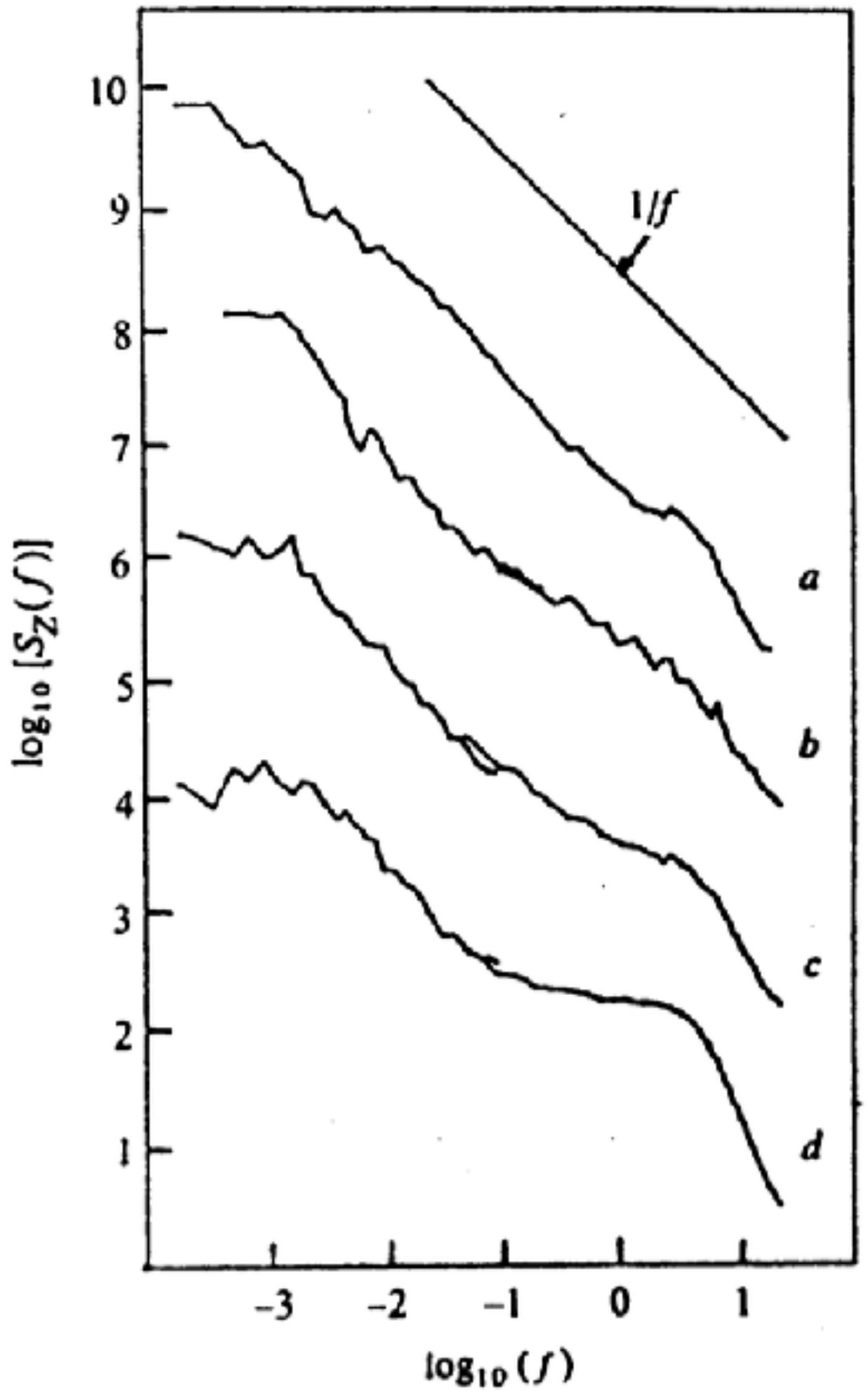
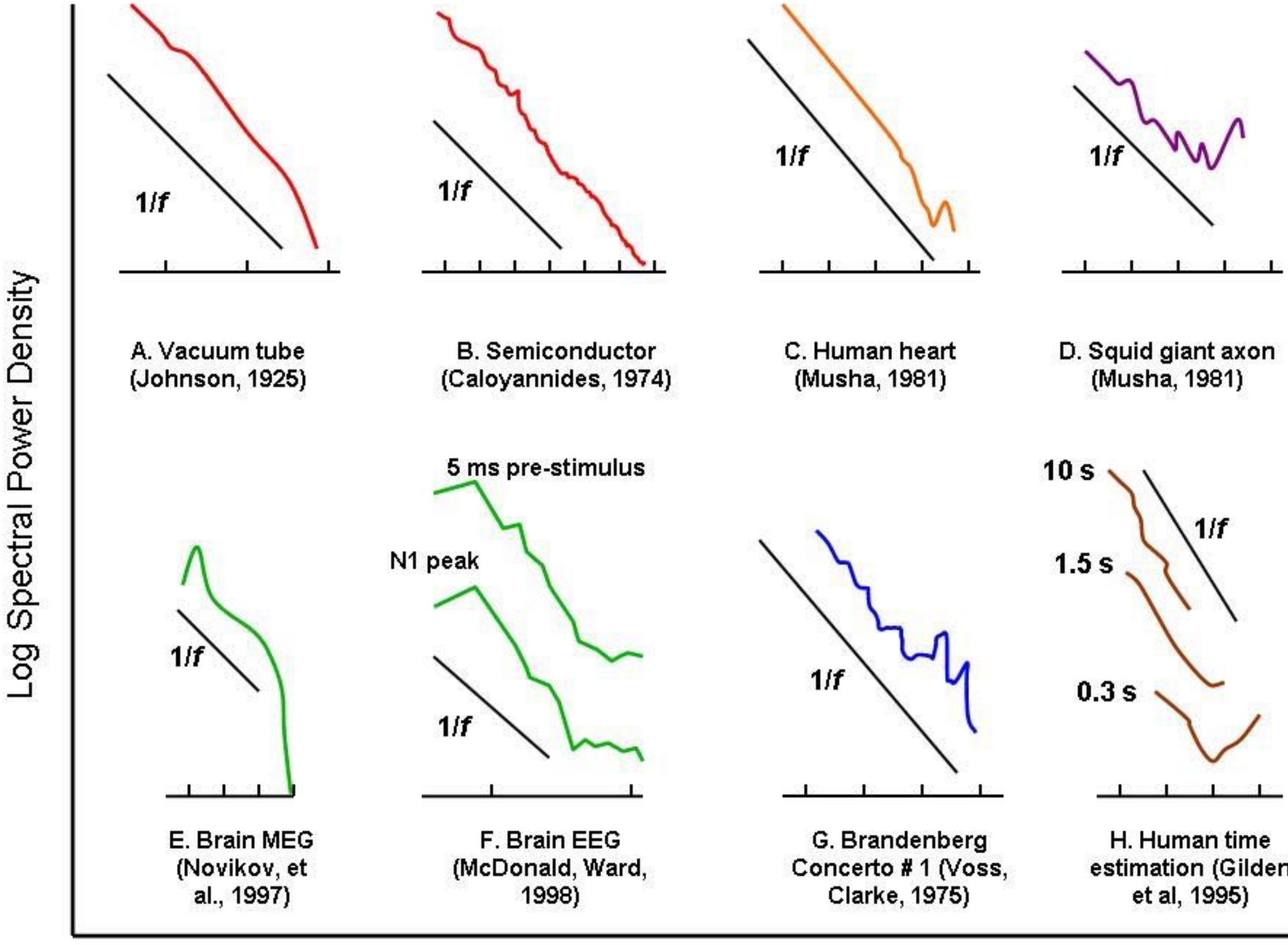
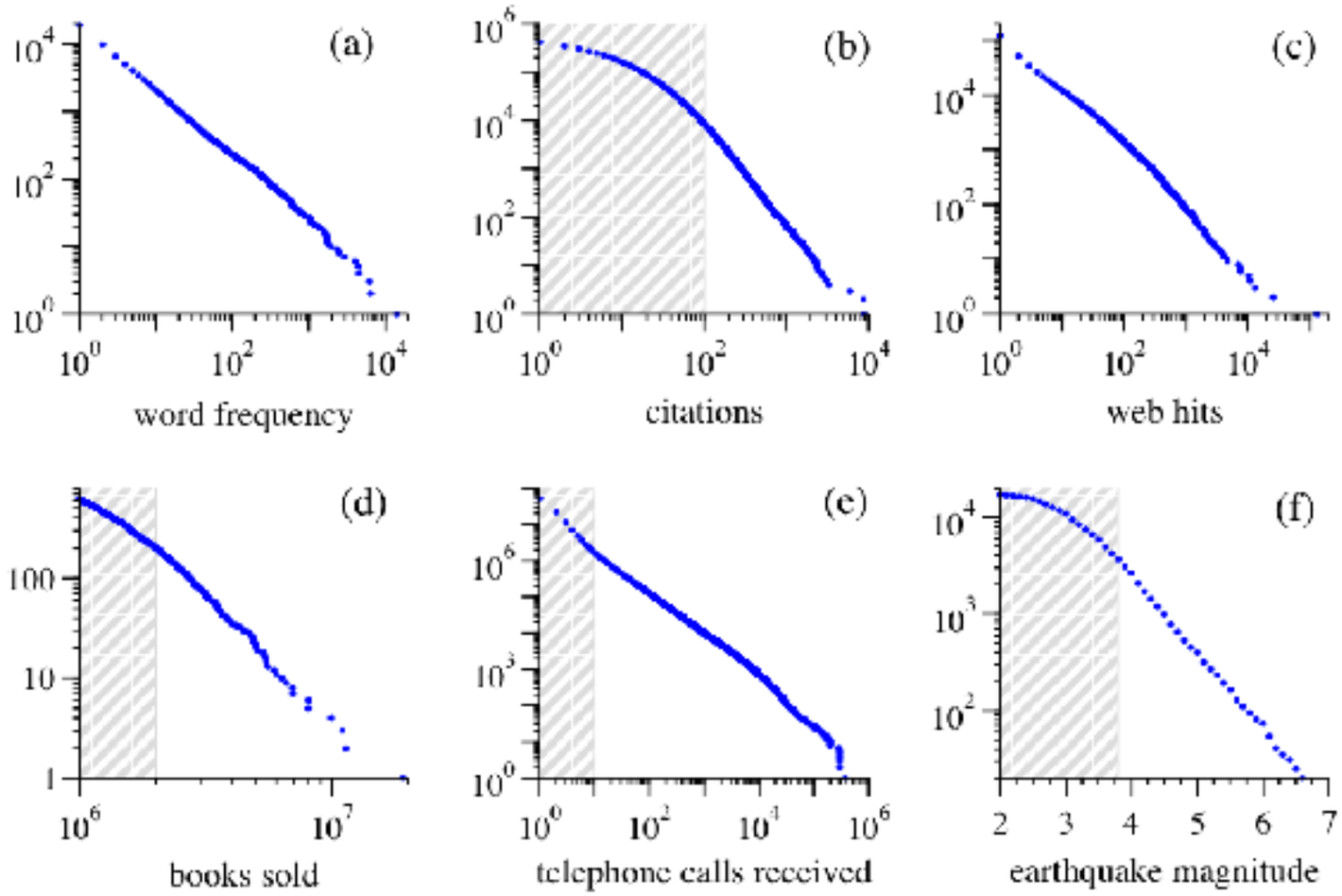


FIGURE 9 (*left*) Loudness fluctuation spectra as a function of frequency f (in Hz) for: *a*, Scott Joplin Piano Rags; *b*, classical radio station; *c*, rock station; *d*, news-and-talk station. (Reproduced from Voss and Clarke.[23]) (*right*) Power spectra of pitch fluctuations for four radio stations: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


Log Spectral Power Density
1/f
A. Vacuum tube (Johnson, 1925)
1/f
B. Semiconductor (Caloyannides, 1974)
1/f
C. Human heart (Musha, 1981)
1/f
D. Squid giant axon (Musha, 1981)
1/f
E. Brain MEG (Novikov, et al., 1997)
5 ms pre-stimulus
N1 peak
1/f
F. Brain EEG (McDonald, Ward, 1998)
1/f
G. Brandenberg Concerto # 1 (Voss, Clarke, 1975)
10 s
1/f
1.5 s
0.3 s
H. Human time estimation (Gilden et al, 1995)
Log Frequency


(a)
word frequency
(b)
citations
(c)
web hits
(d)
books sold
(e)
telephone calls received
(f)
earthquake magnitude
(g)
crater diameter in km
(h)
peak intensity
(i)
intensity
(j)
net worth in US dollars
(k)
name frequency
(l)
population of city

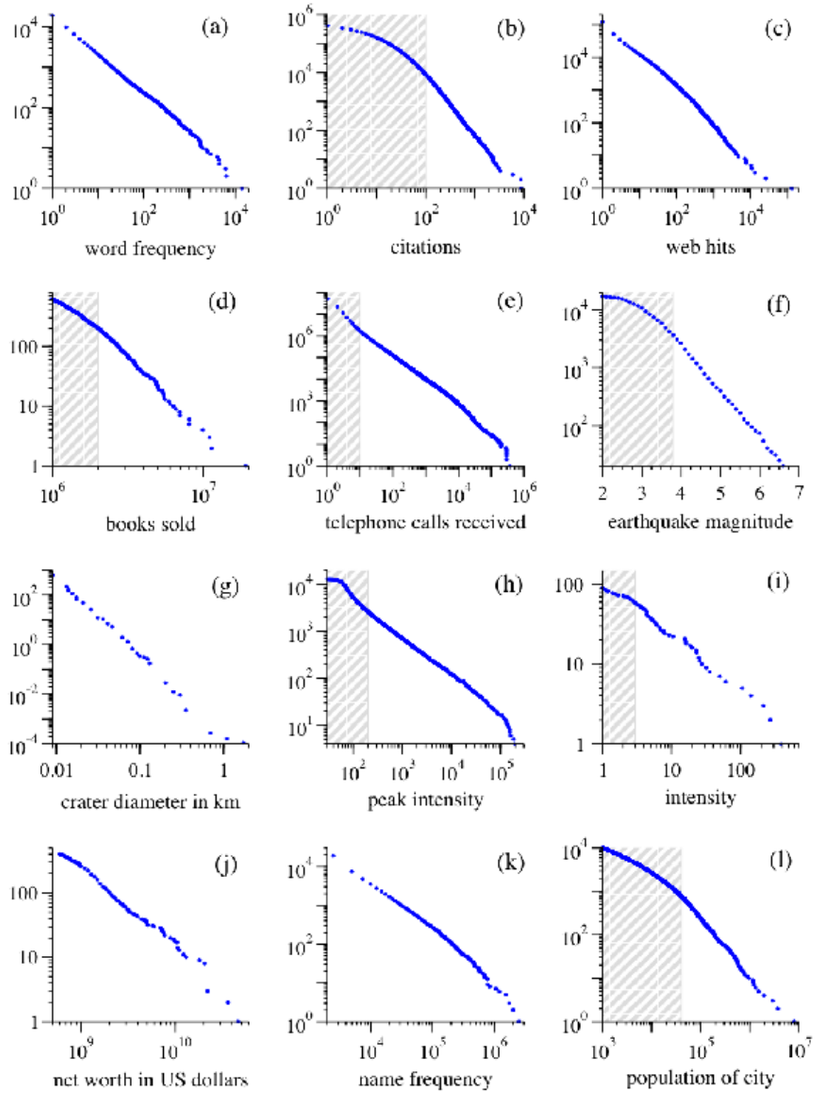
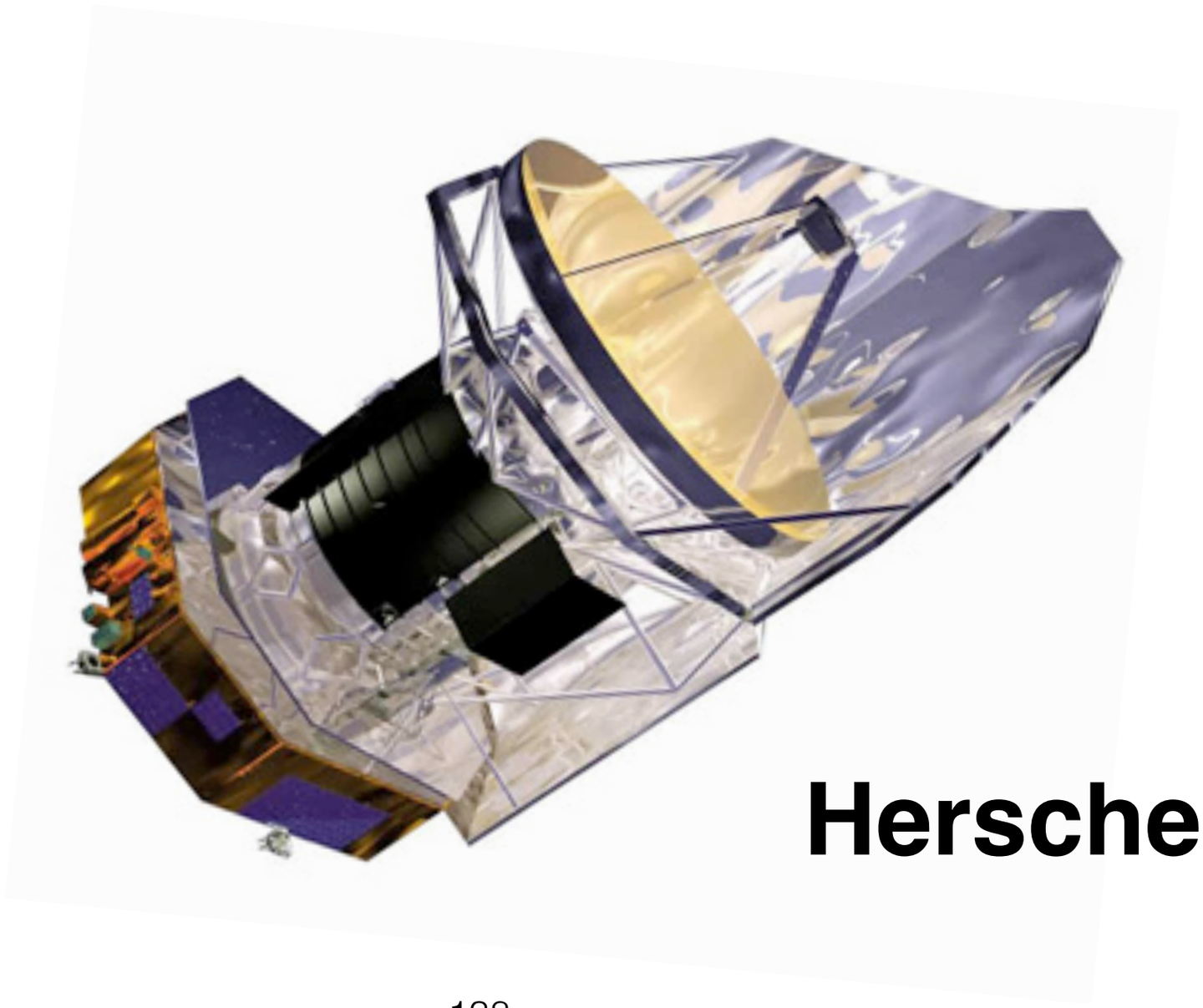
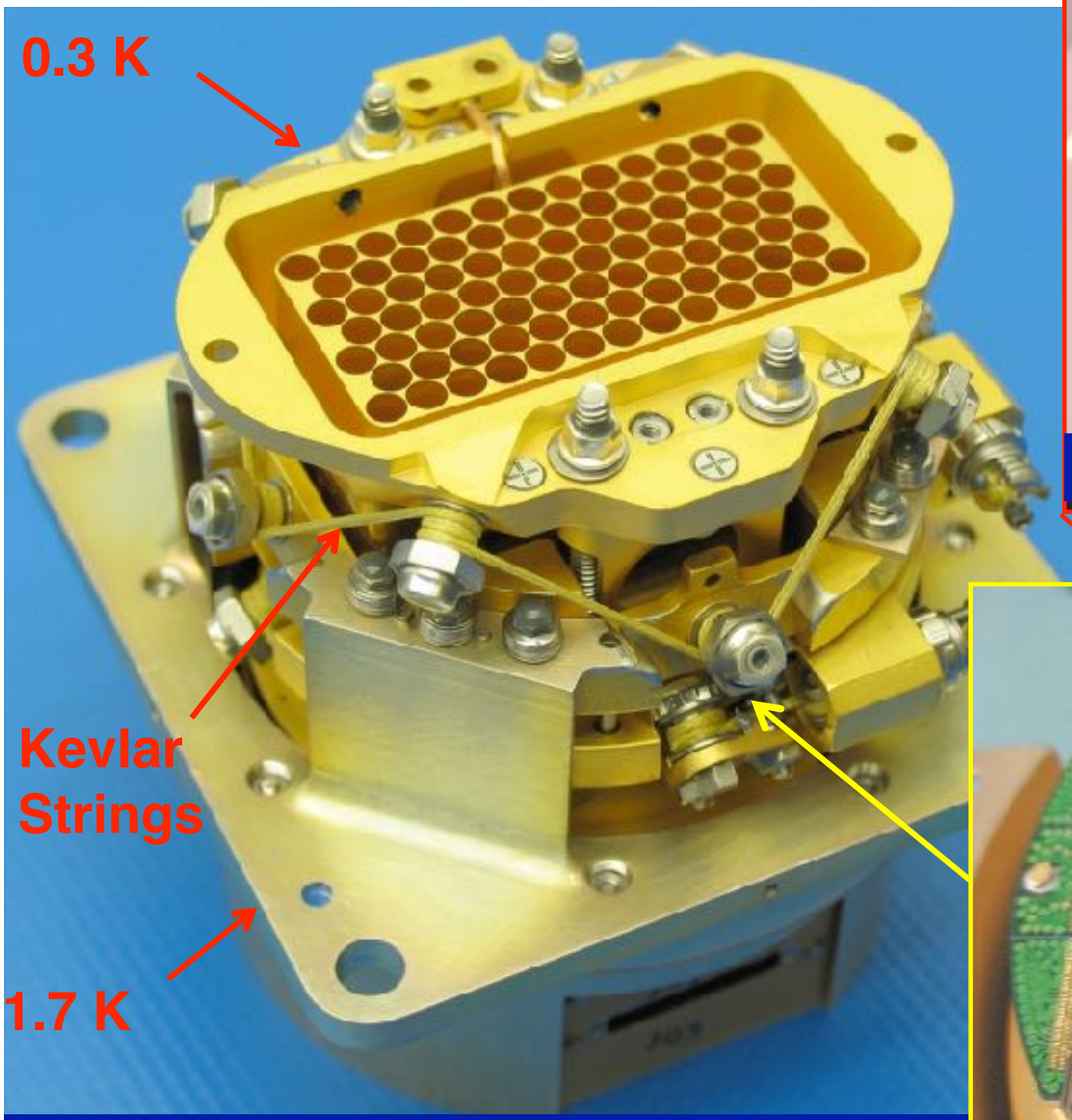


FIG. 4 Cumulative distributions or "rank/frequency plots" of twelve quantities reputed to follow power laws. The distributions were computed as described in Appendix A. Data in the shaded regions were excluded from the calculations of the exponents in Table I. Source references for the data are given 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.

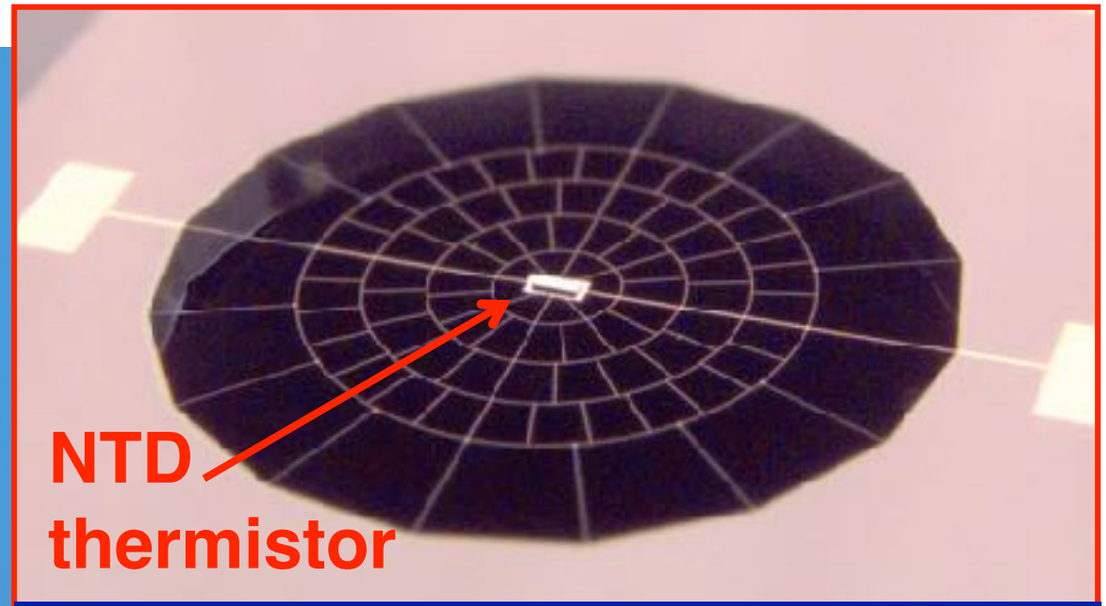
Example: Far Infrared



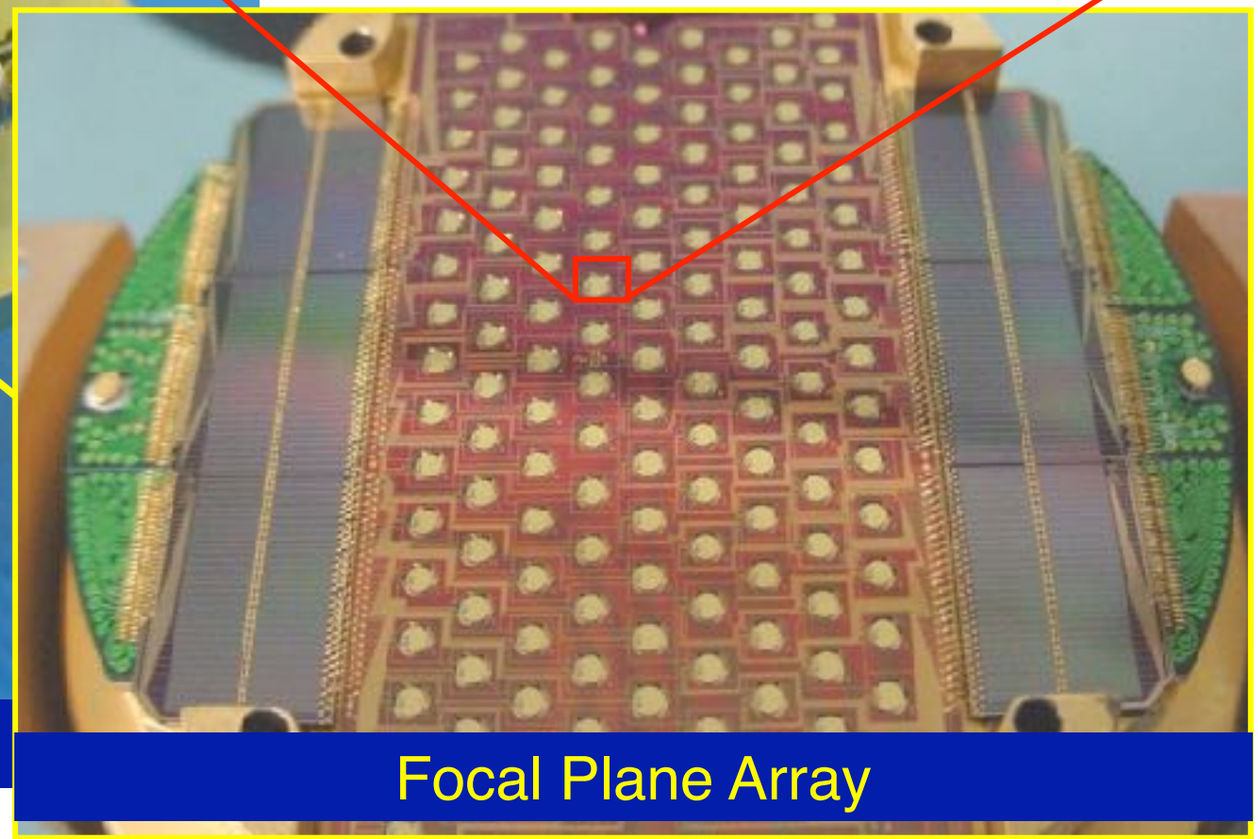
SPIRE Instrument: NASA/JPL



300 mK Detector Assembly



'Spider-Web' Bolometer



Focal Plane Array

These detectors are background-limited !

250 μm

350 μm

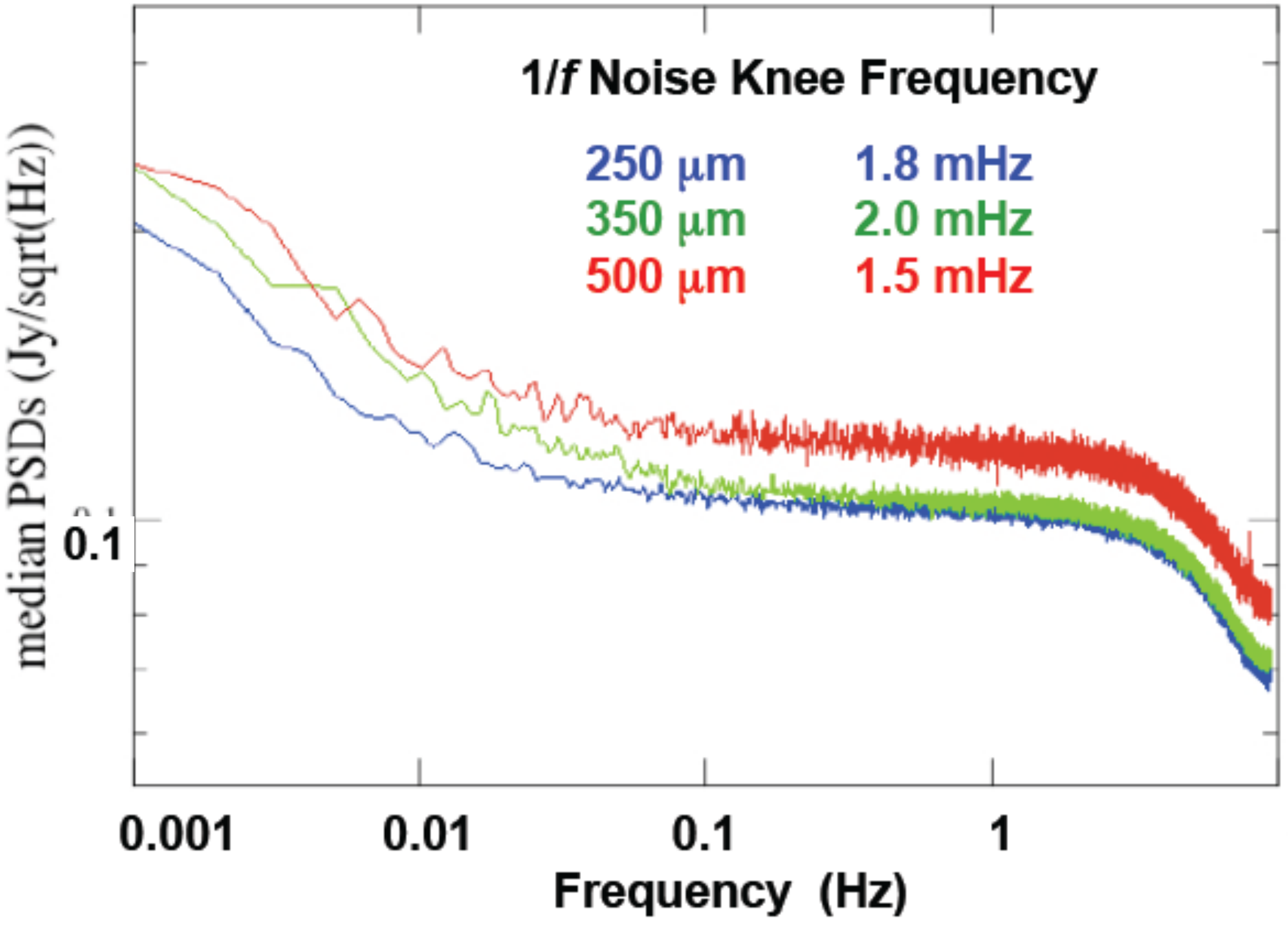
500 μm

SPIRE map of GOODS-N
c. 2009

400 arcmin^2

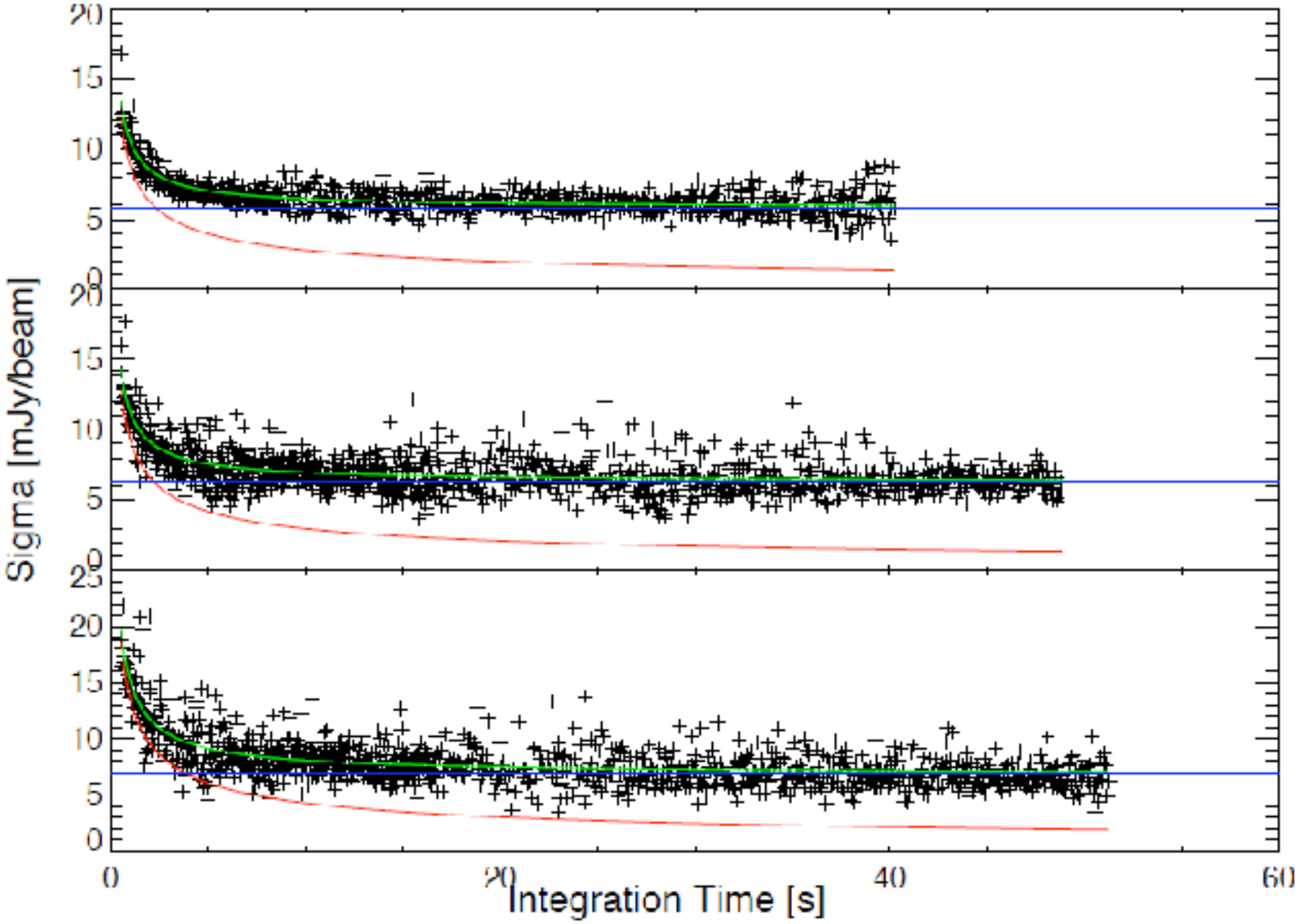
10 arcmin

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

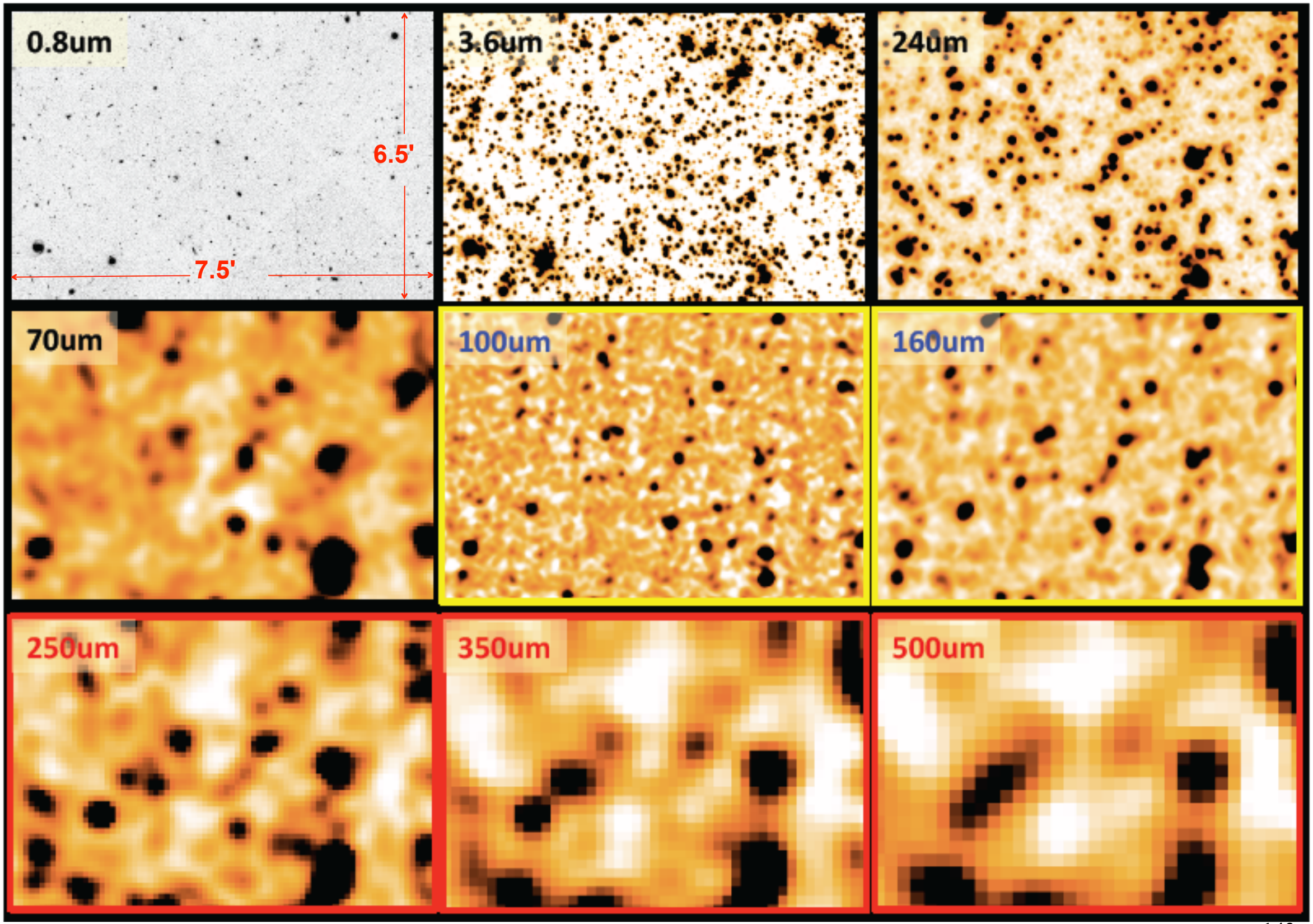


Band	Conf. (1σ*)	T/sq. deg
250 um	4.8	2.8 h
350 um	5.5	1.4 h
500 um	6.1	2.4 h

*after an iterative 5σ clip

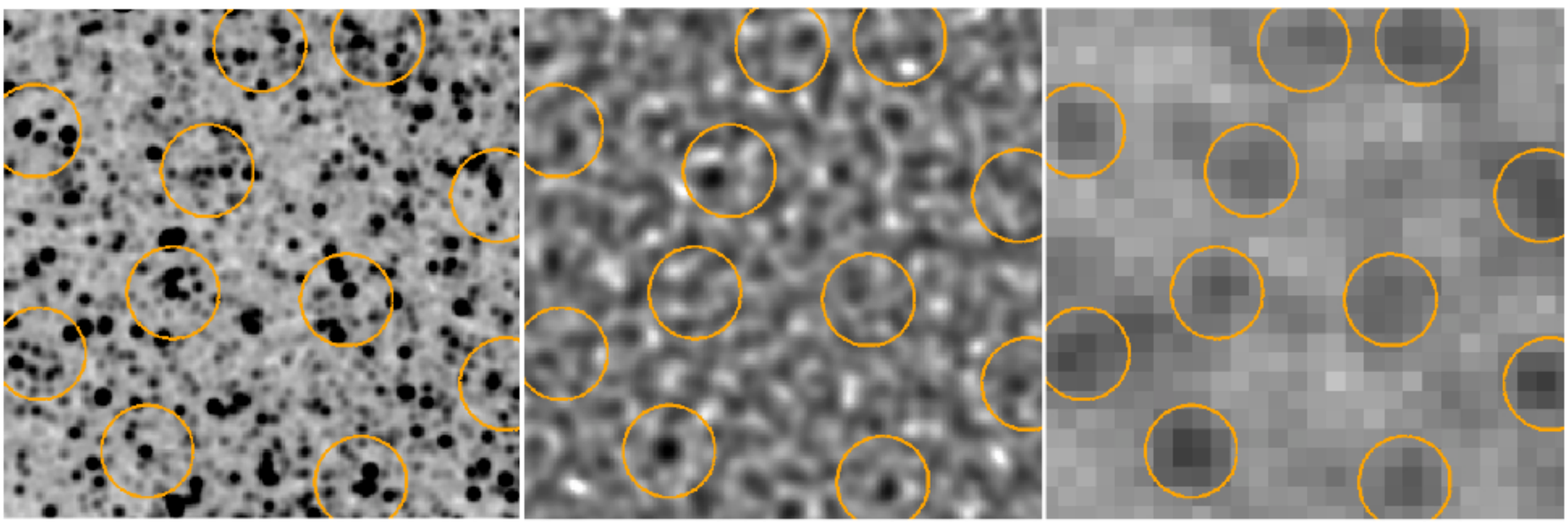
Nguyen et al. 2010

Large Beam + High Sensitivity = Confusion



D. Elbaz

Confusion noise



25 m

15 m

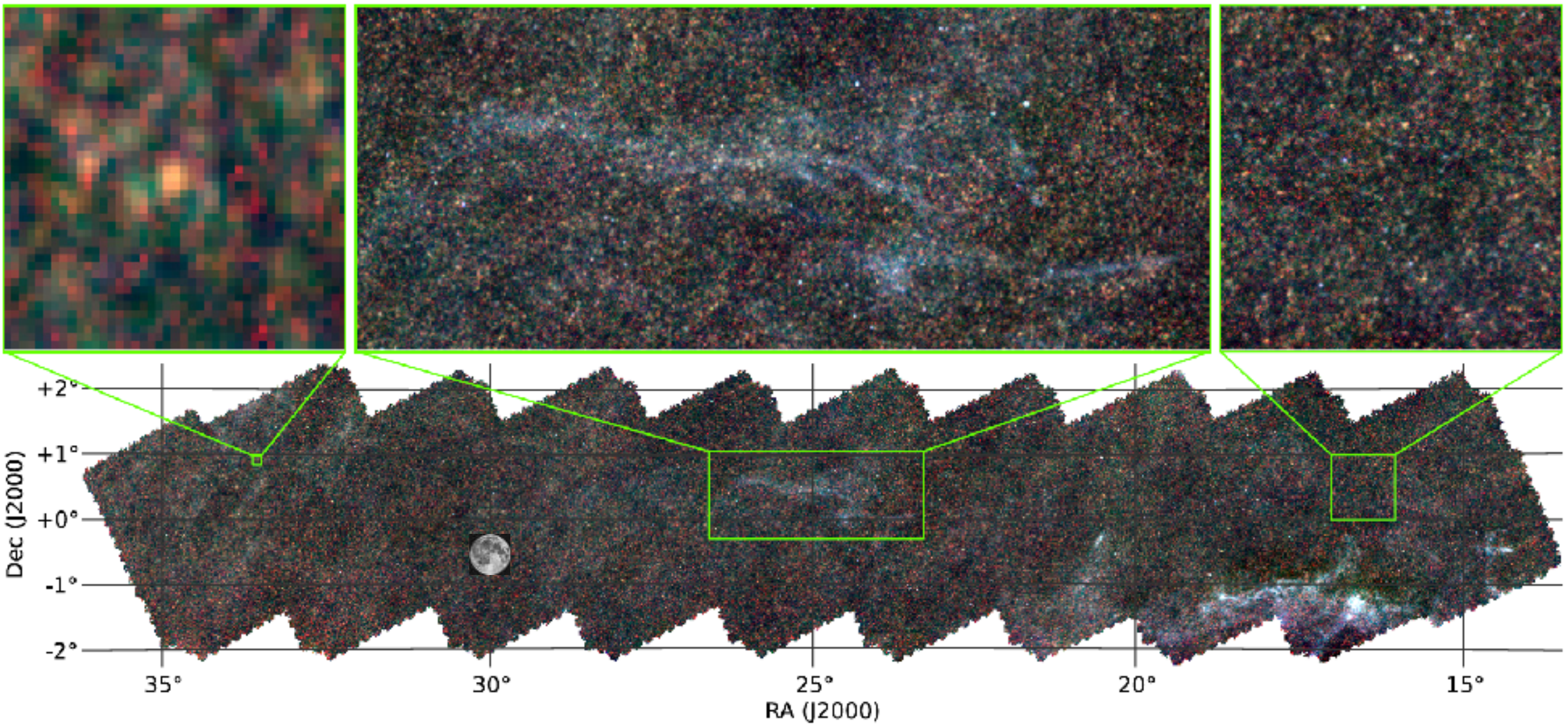
3.5 m

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 um

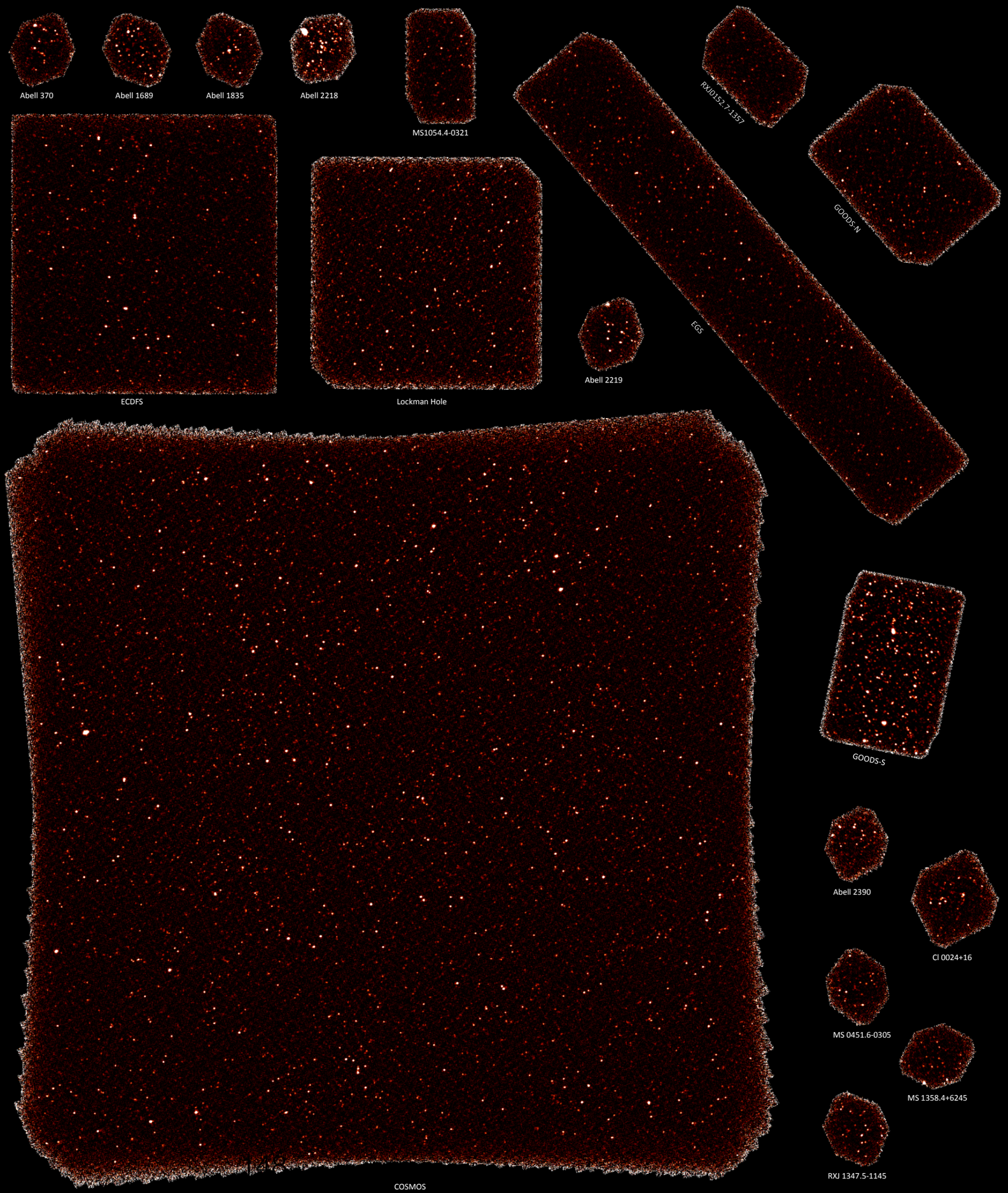
90 deg² Herschel field with cirrus



What are statistical errors? What are systematic?

Herschel/PACS

60, 100, 160 um

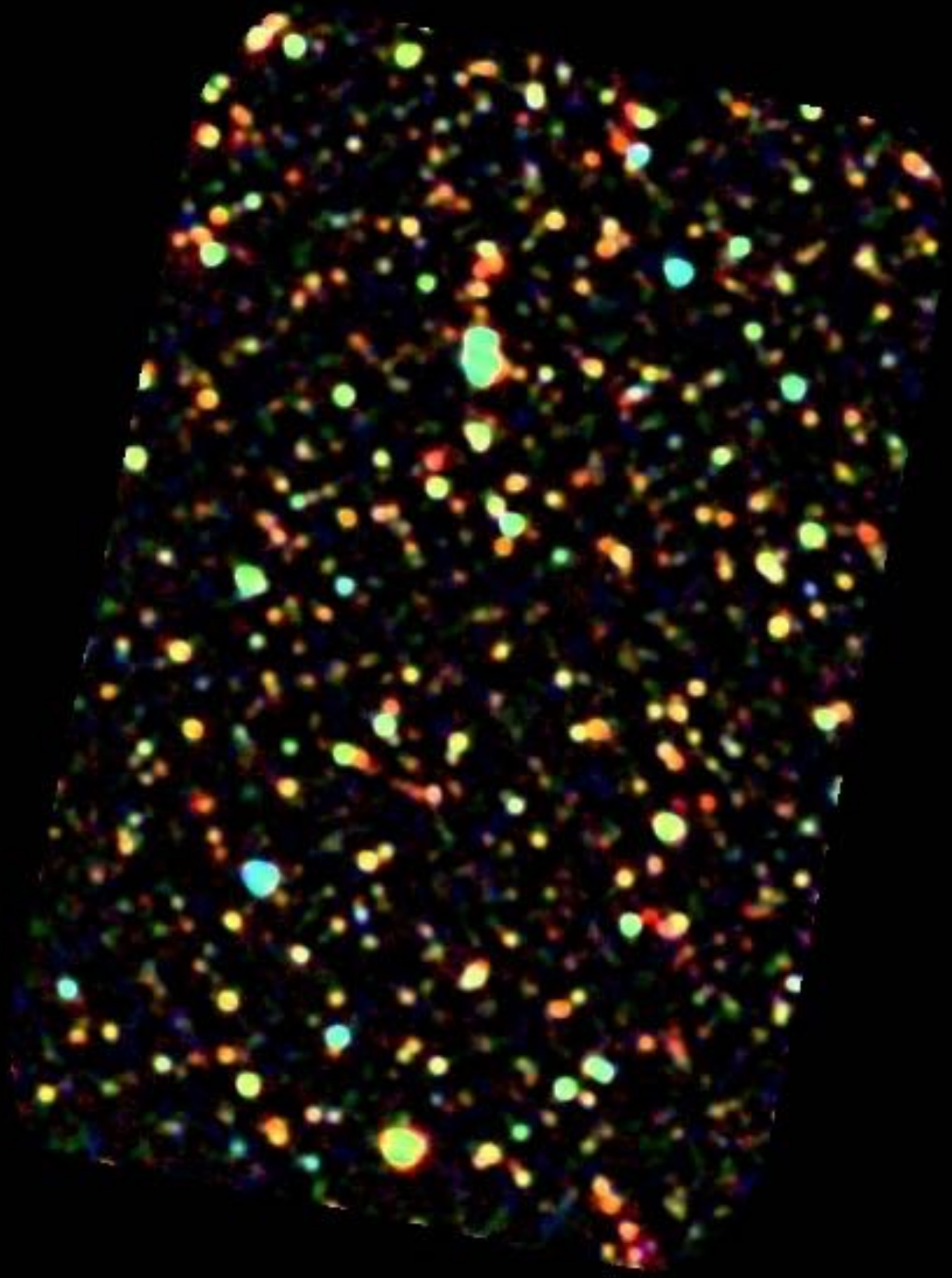


Herschel/PACS
60, 100, 160 um

Only detector noise !

NOT background limited!

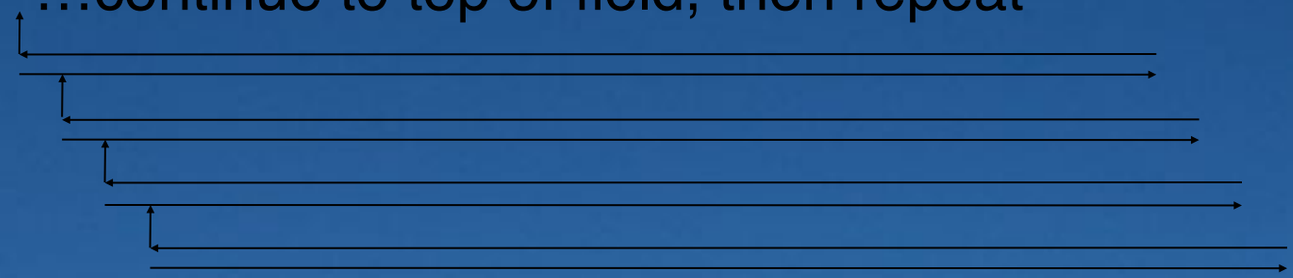
Bad !

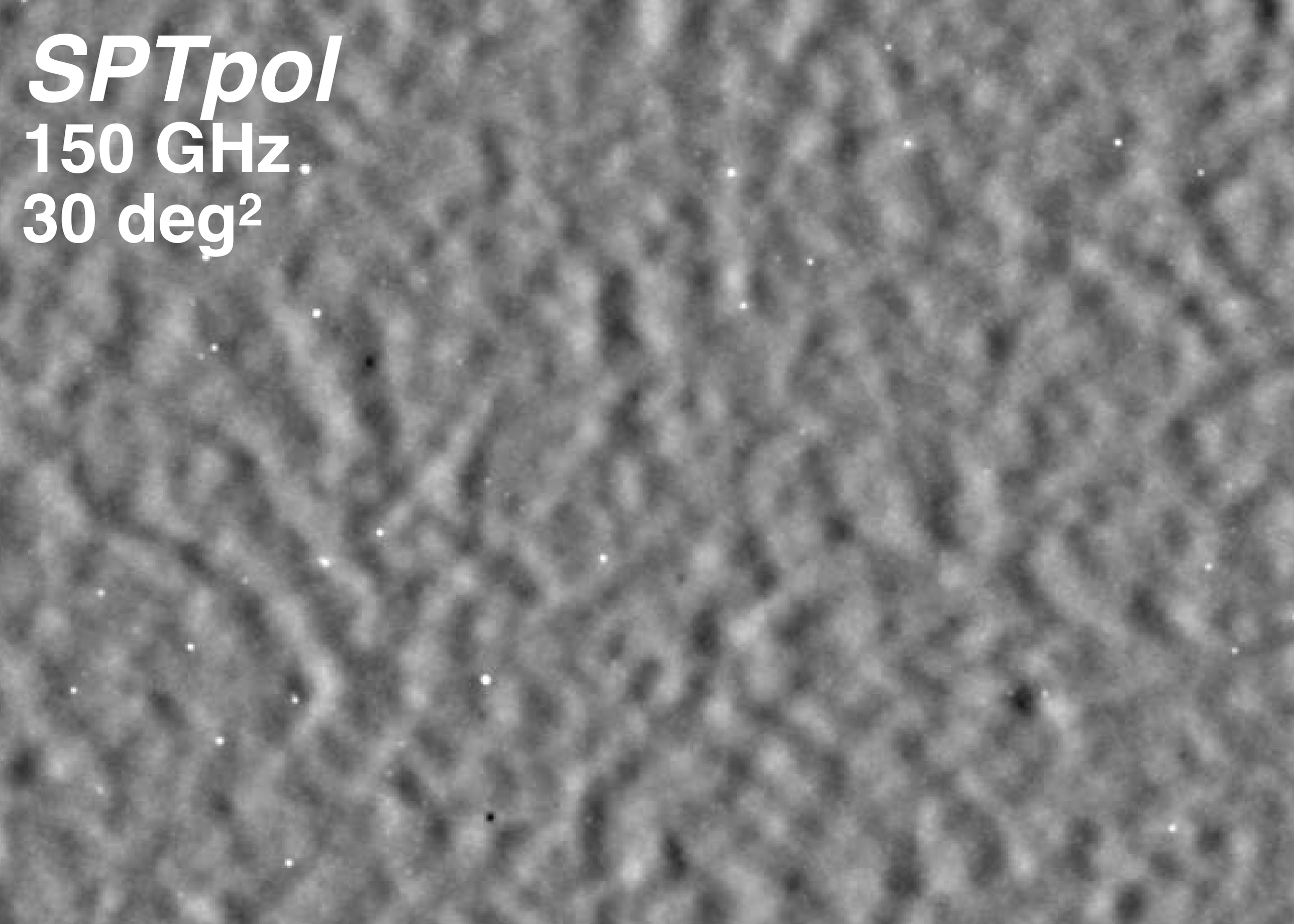


The GOODS-S field as seen by PEP

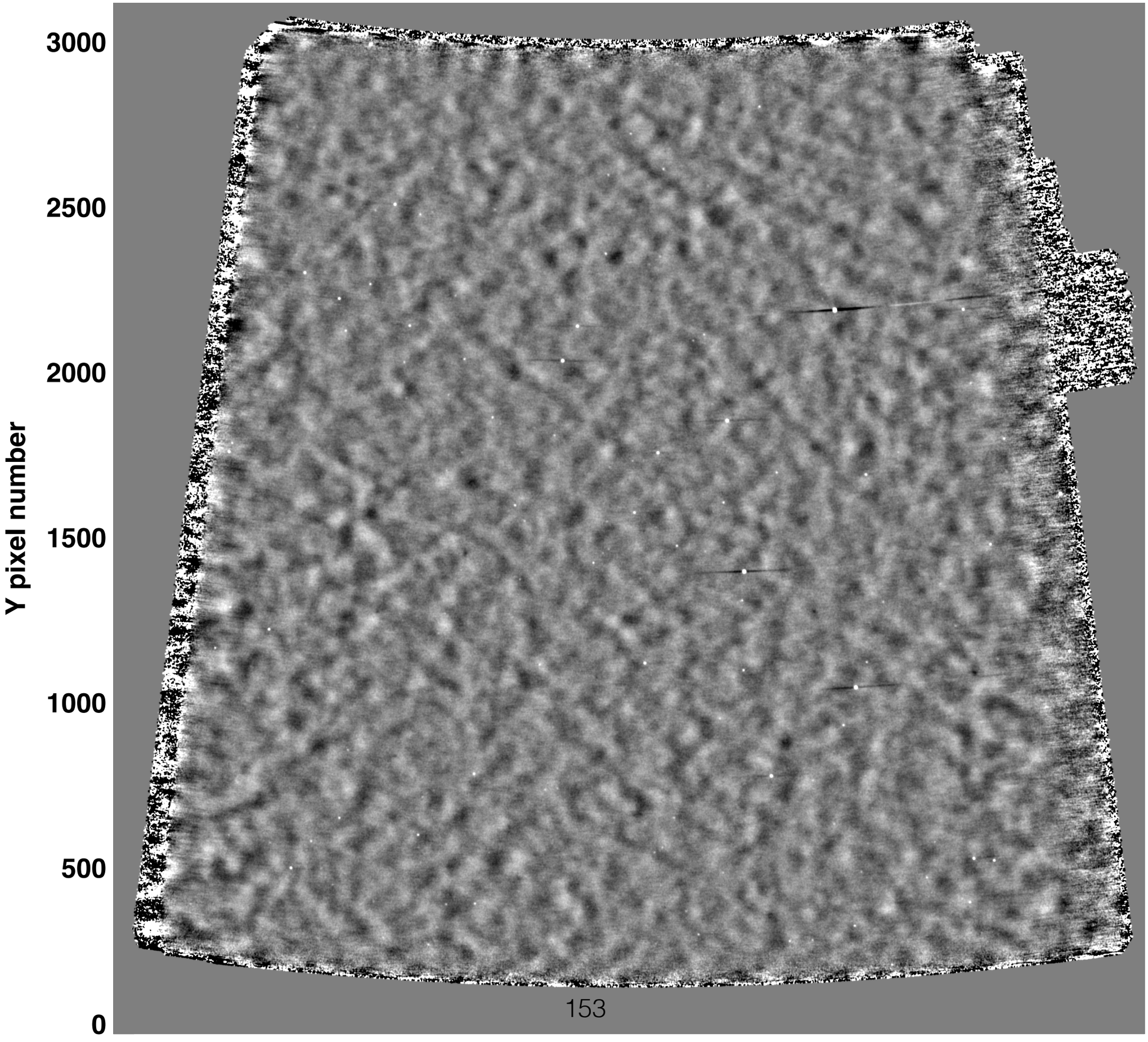
Problem: low frequency 1/f noise from the instrument and atmospheric fluctuations contaminate the astrophysical signal.

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

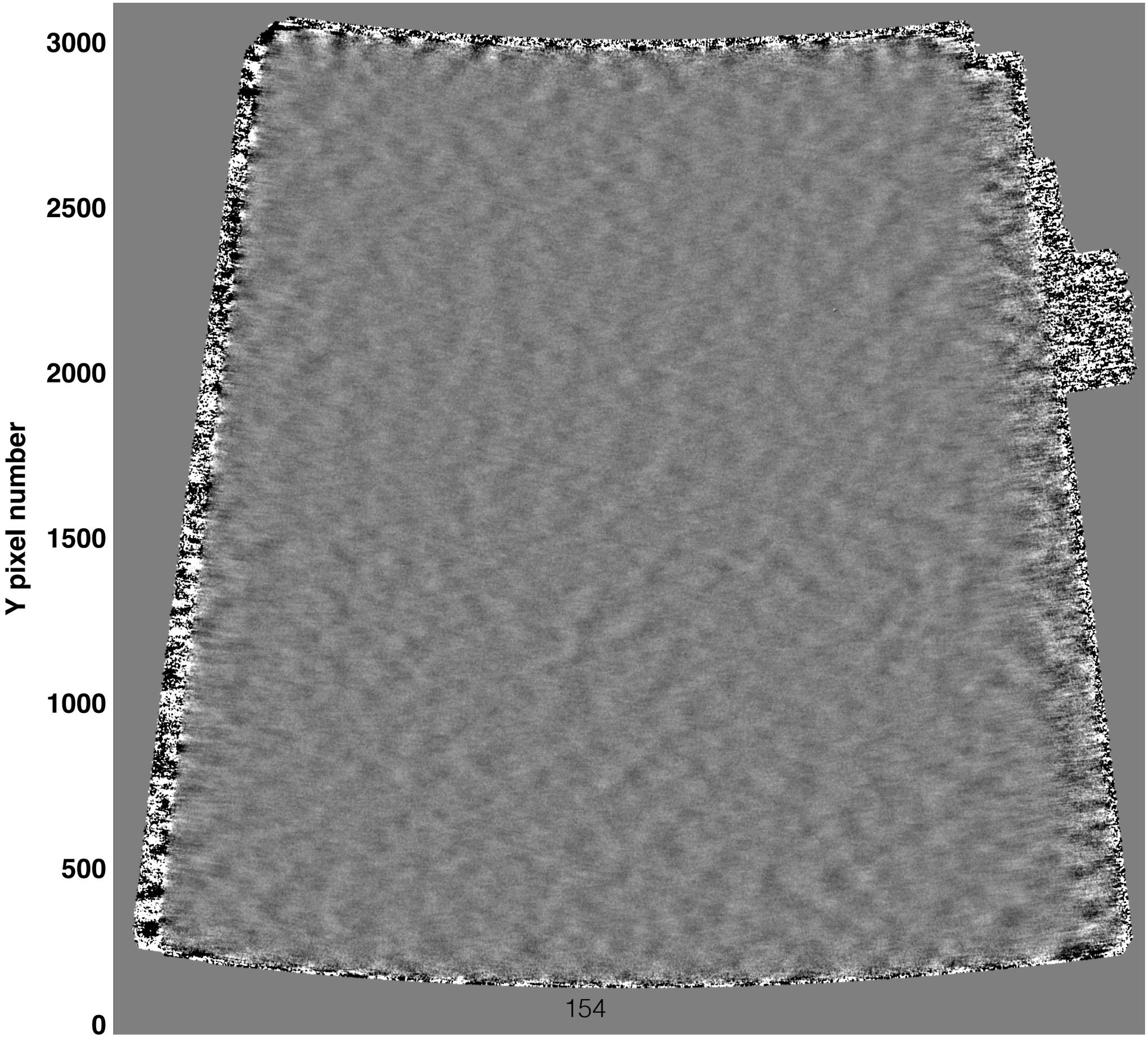



SPTpol
150 GHz
30 deg²

2.0 mm signal map

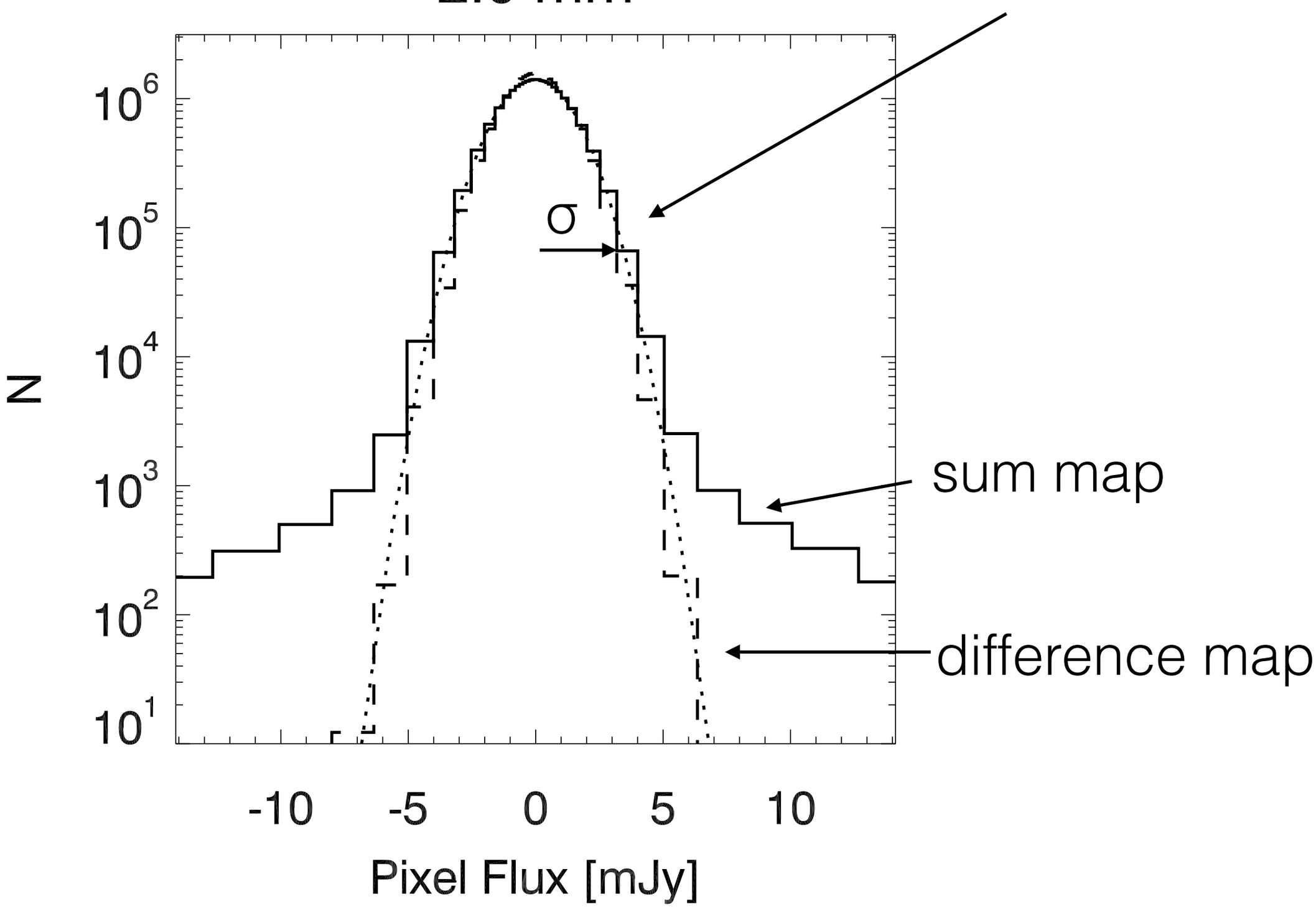


2.0 mm difference map

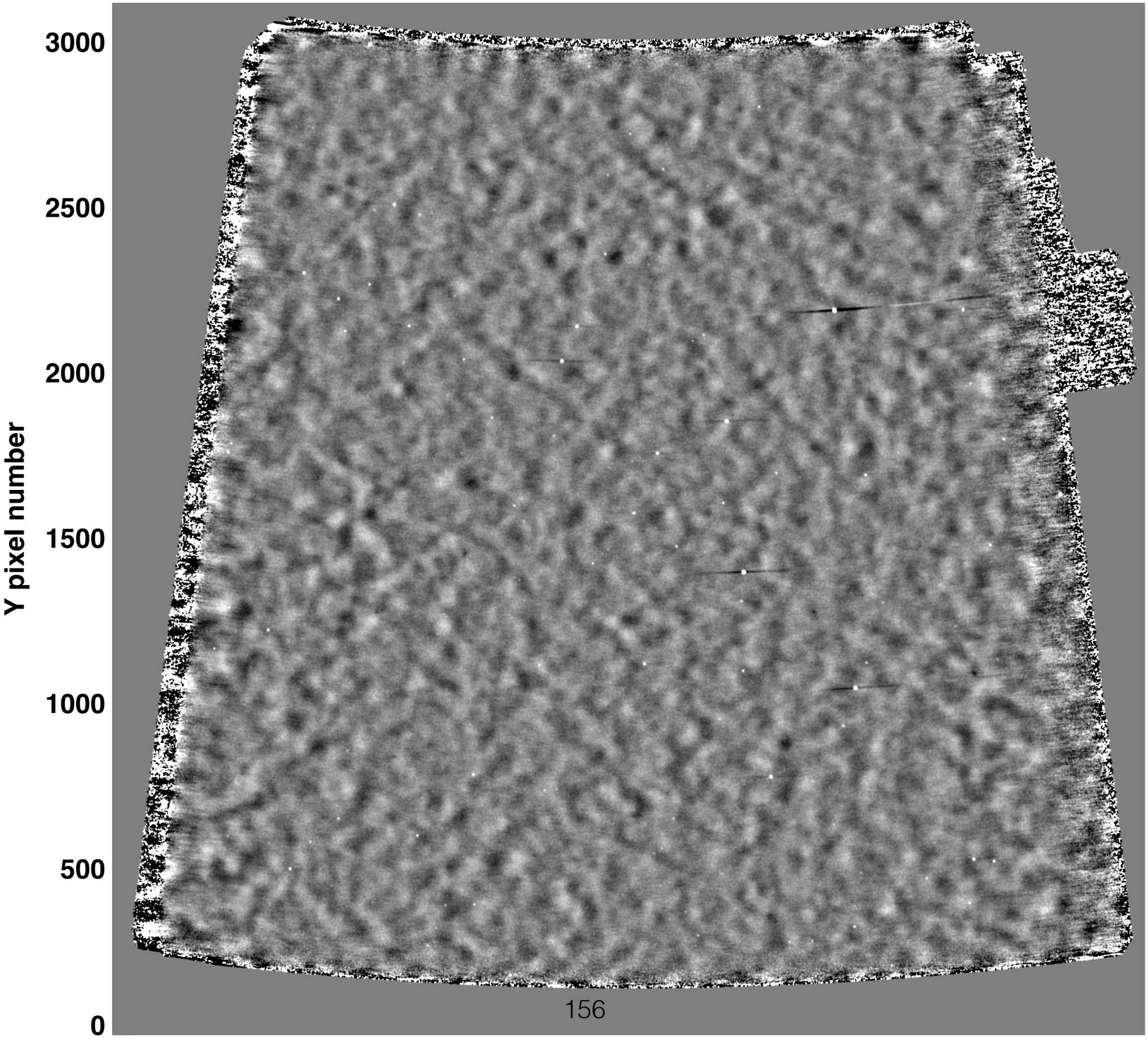


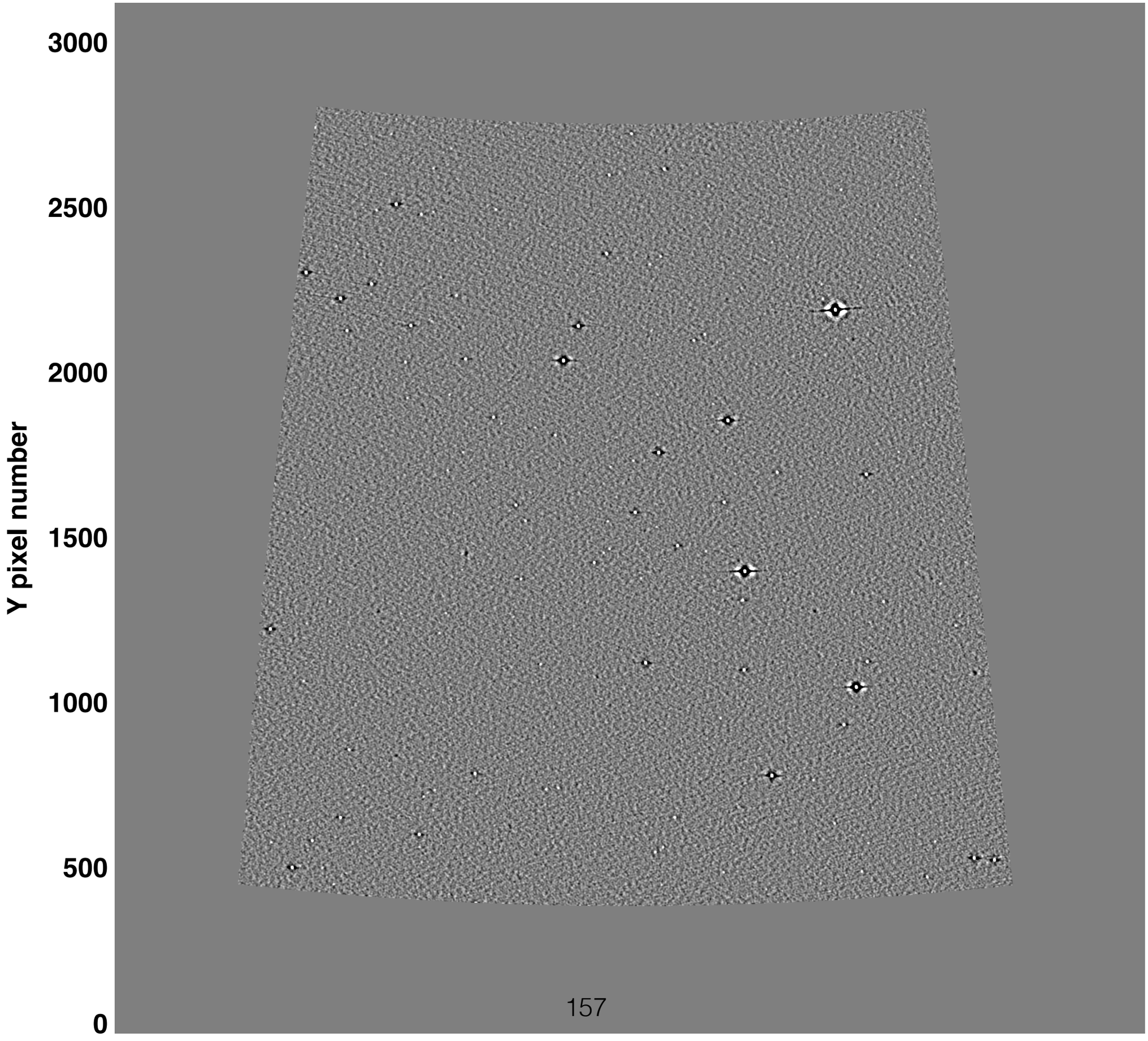
map pixel histogram

this is the noise in the map

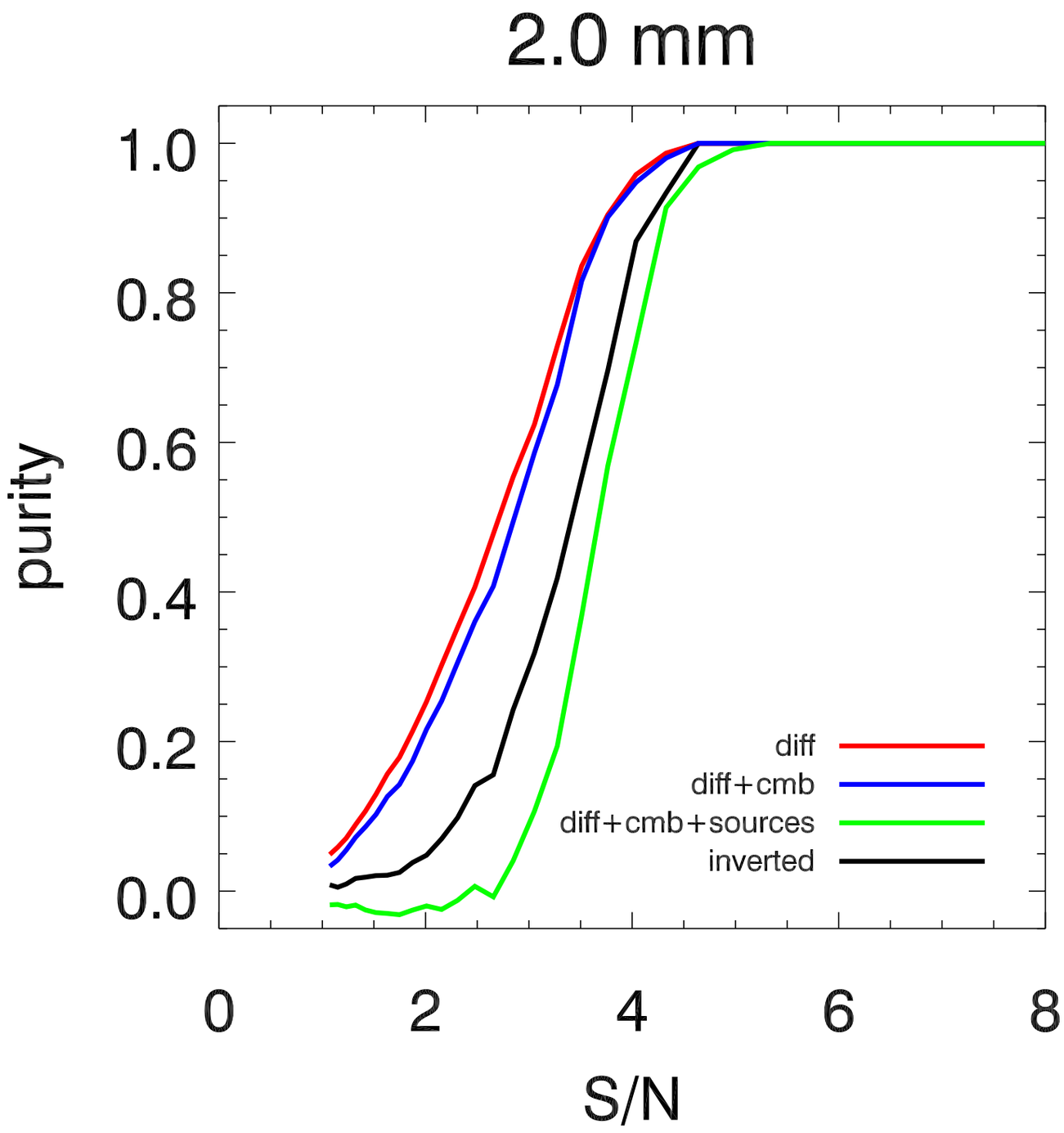


2.0 mm signal map




2.0 mm filtered map
Y pixel number
3000
2500
2000
1500
1000
500
0

Catalog Purity

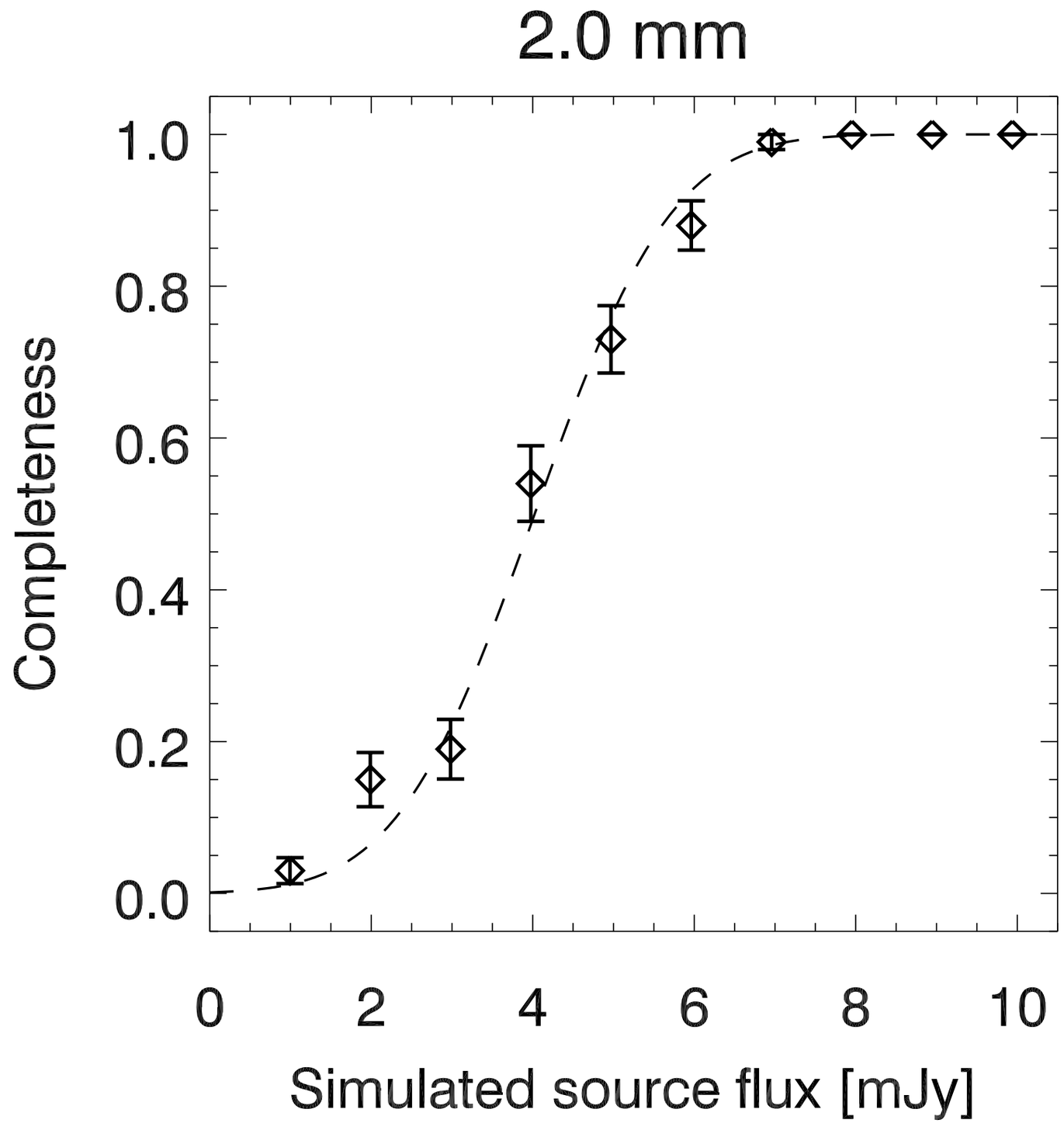


At a given S/N, what fraction of sources are real?

multiple ways of testing this.

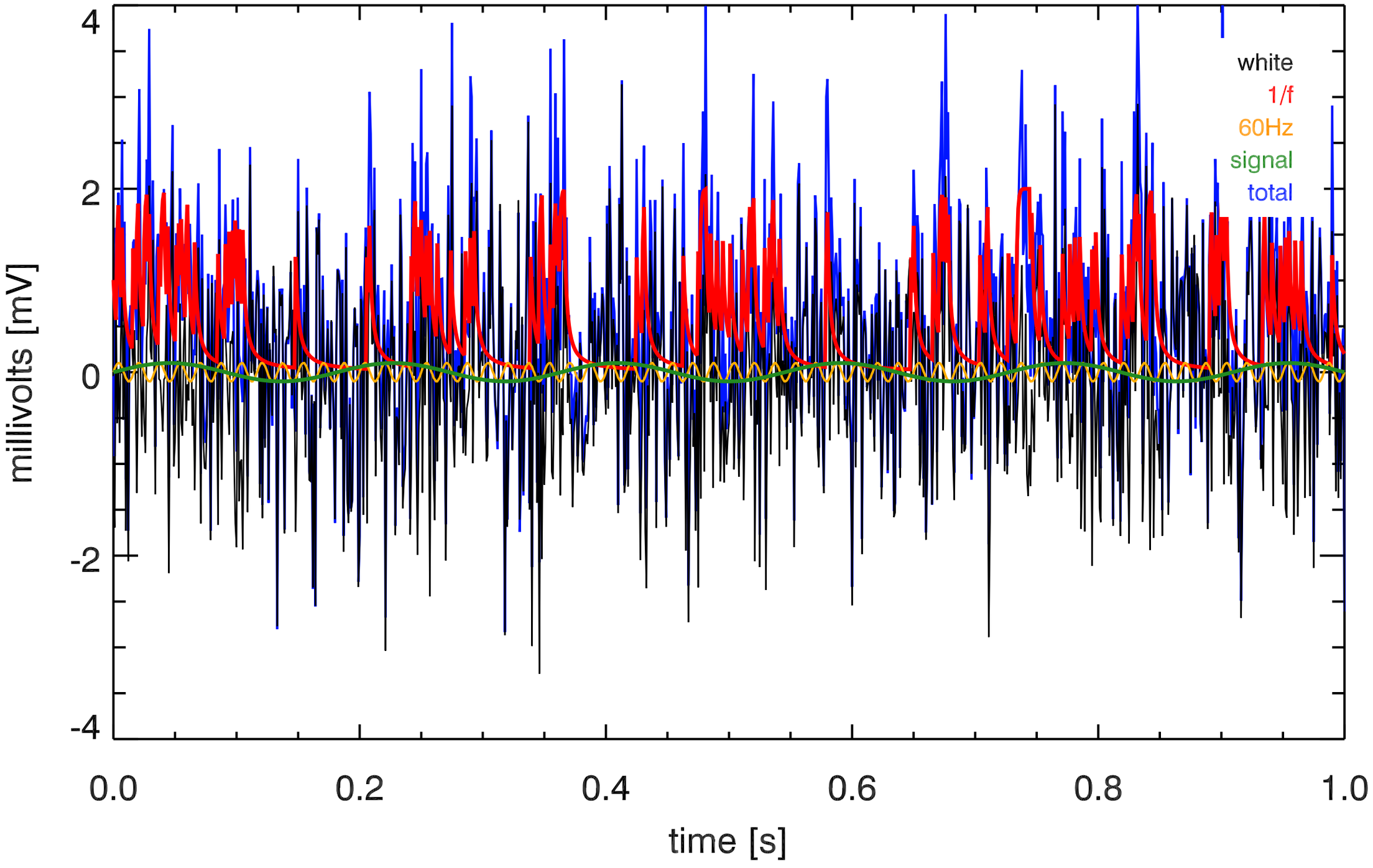
Which is best?

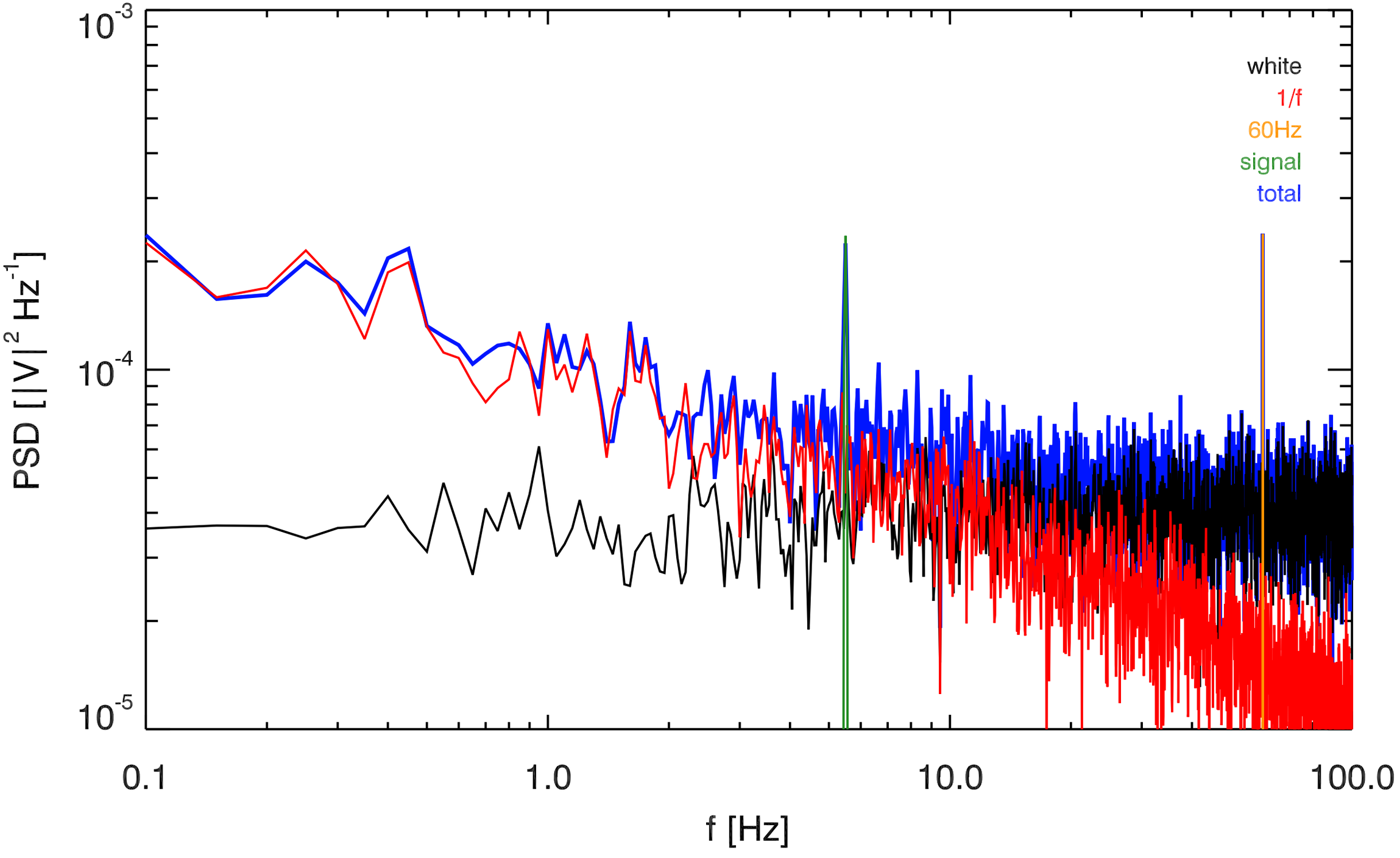
Catalog Completeness

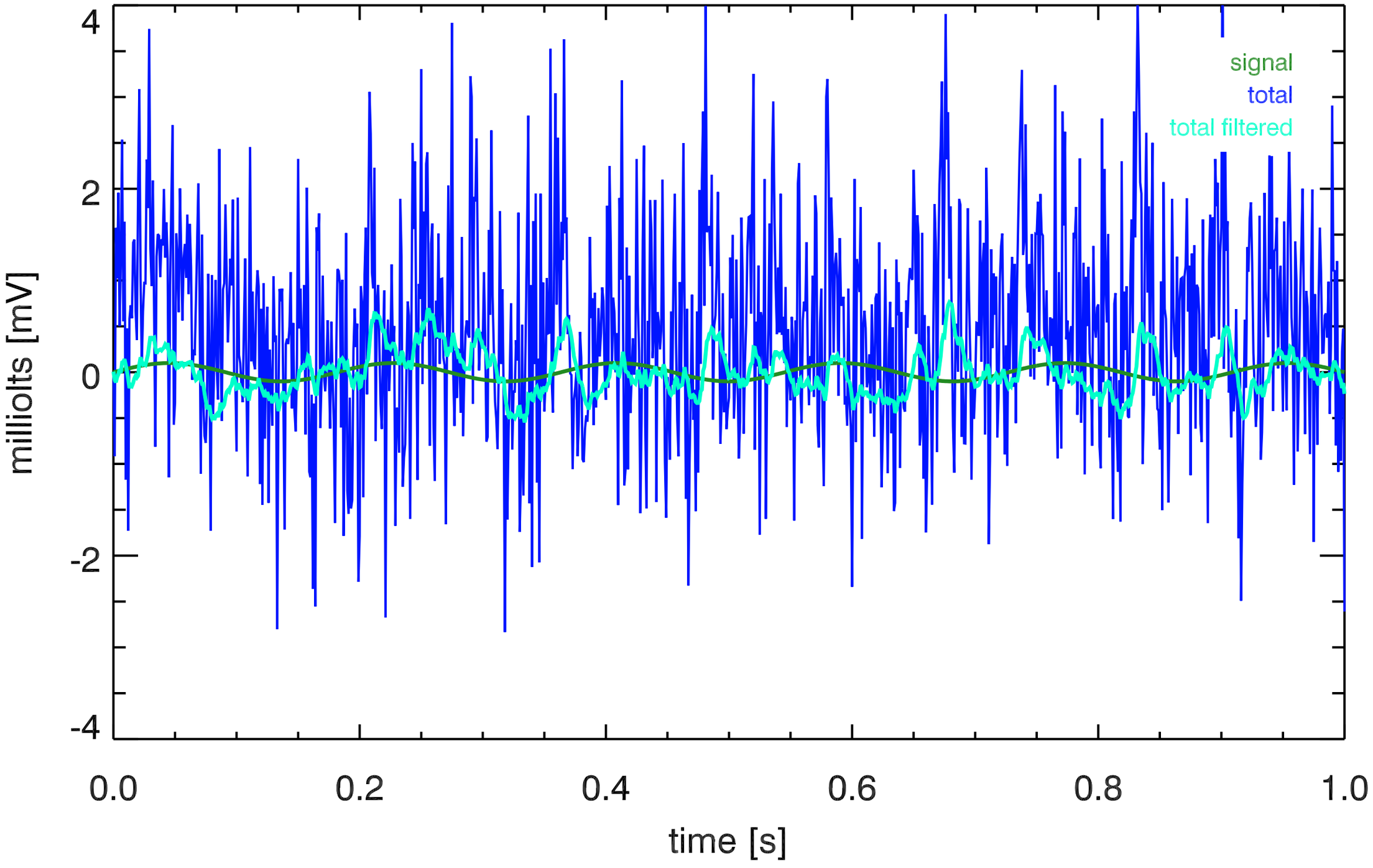


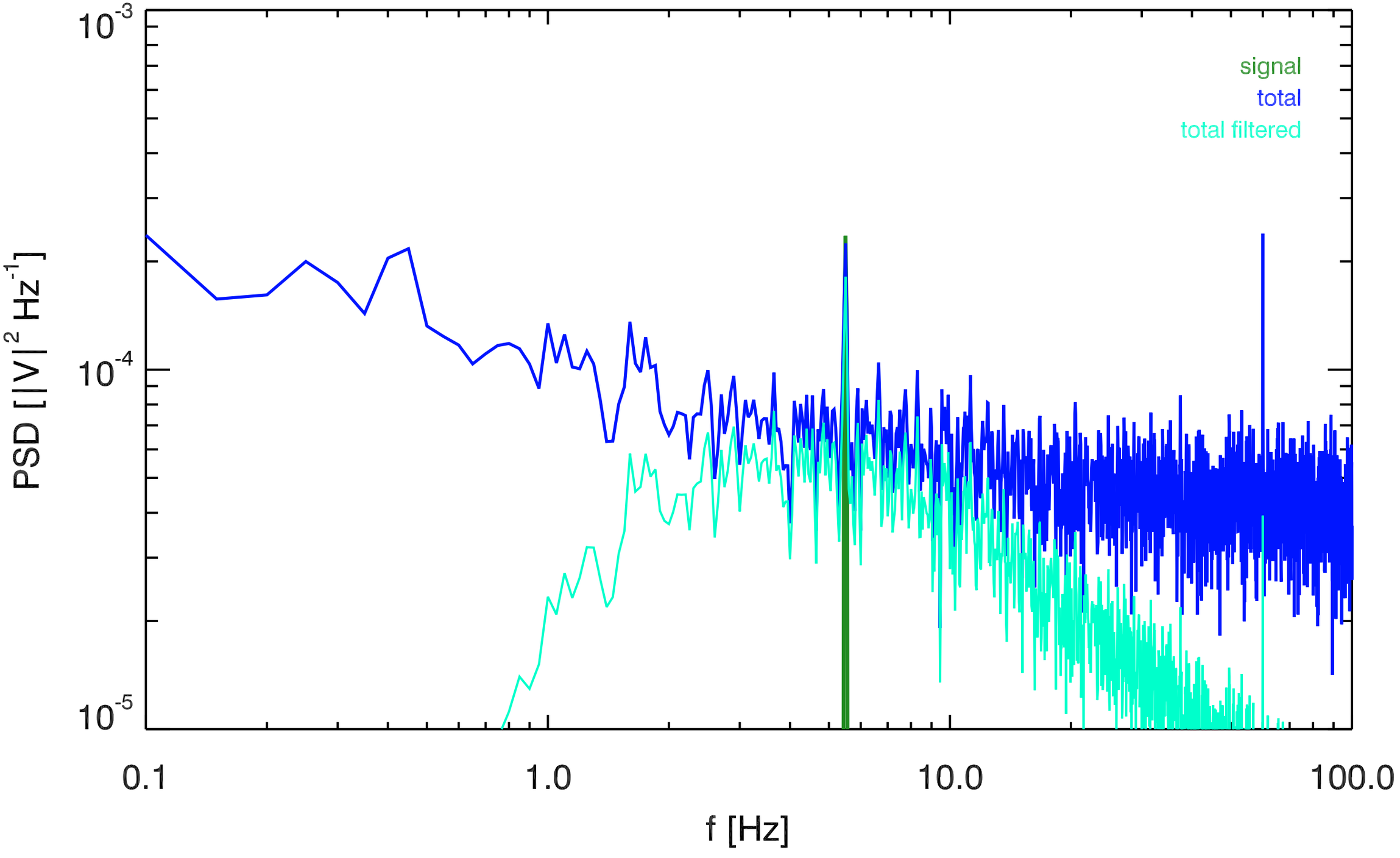
How many true sources do you expect to recover at a given flux level?

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

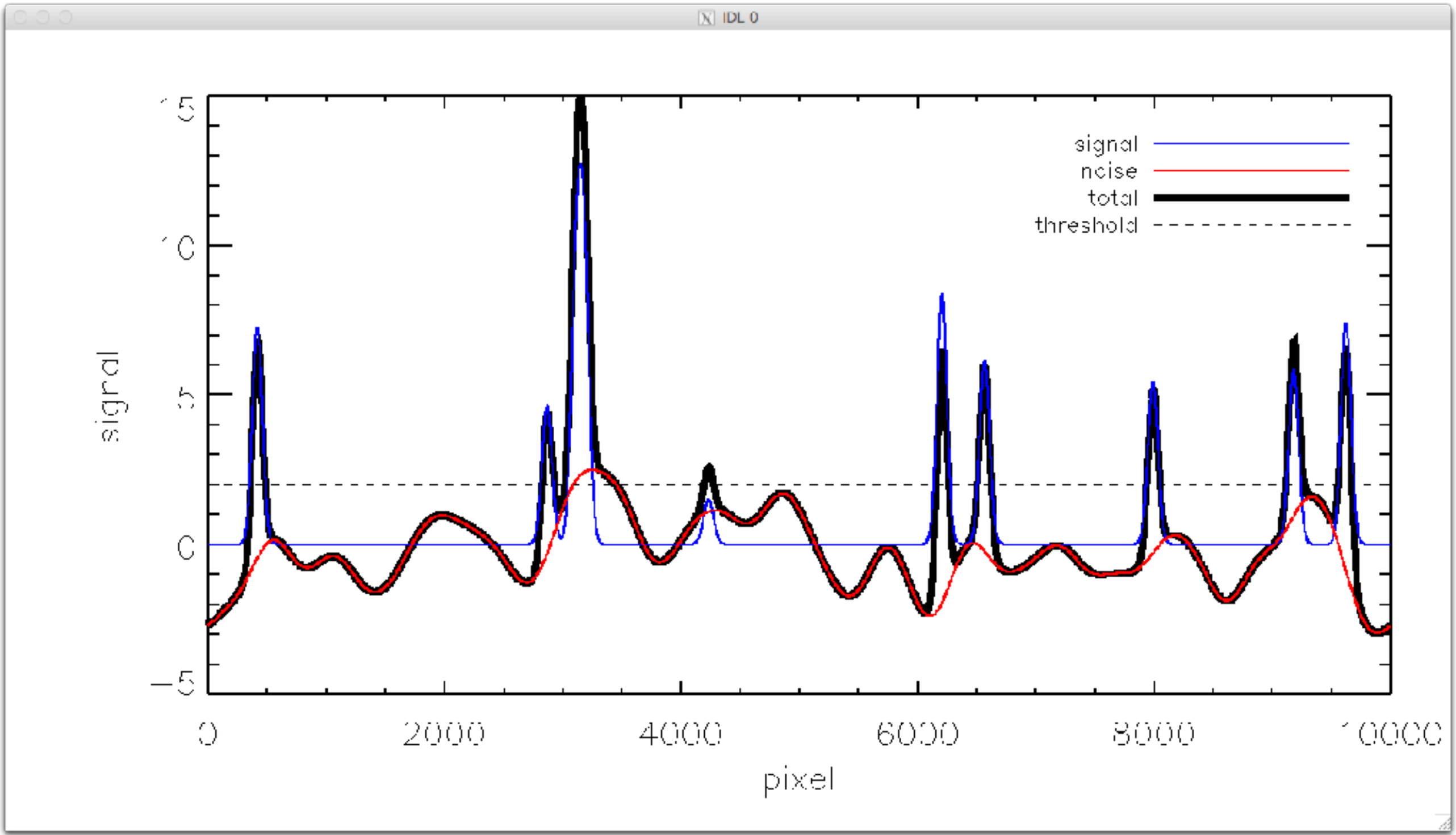

4
2
0
-2
-4
millivolts [mV]
0.0
0.2
0.4
0.6
0.8
1.0
time [s]
white
1/f
60Hz
signal
total


white
1/f
60Hz
signal
total
PSD [|V|^2 Hz^-1]
10^-3
10^-4
10^-5
0.1
1.0
10.0
100.0
f [Hz]


signal
total
total filtered
4
2
0
-2
-4
millivolts [mV]
0.0
0.2
0.4
0.6
0.8
1.0
time [s]


signal
total
total filtered
PSD [|V|^2 Hz^-1]
10^-3
10^-4
10^-5
0.1
1.0
10.0
100.0
f [Hz]

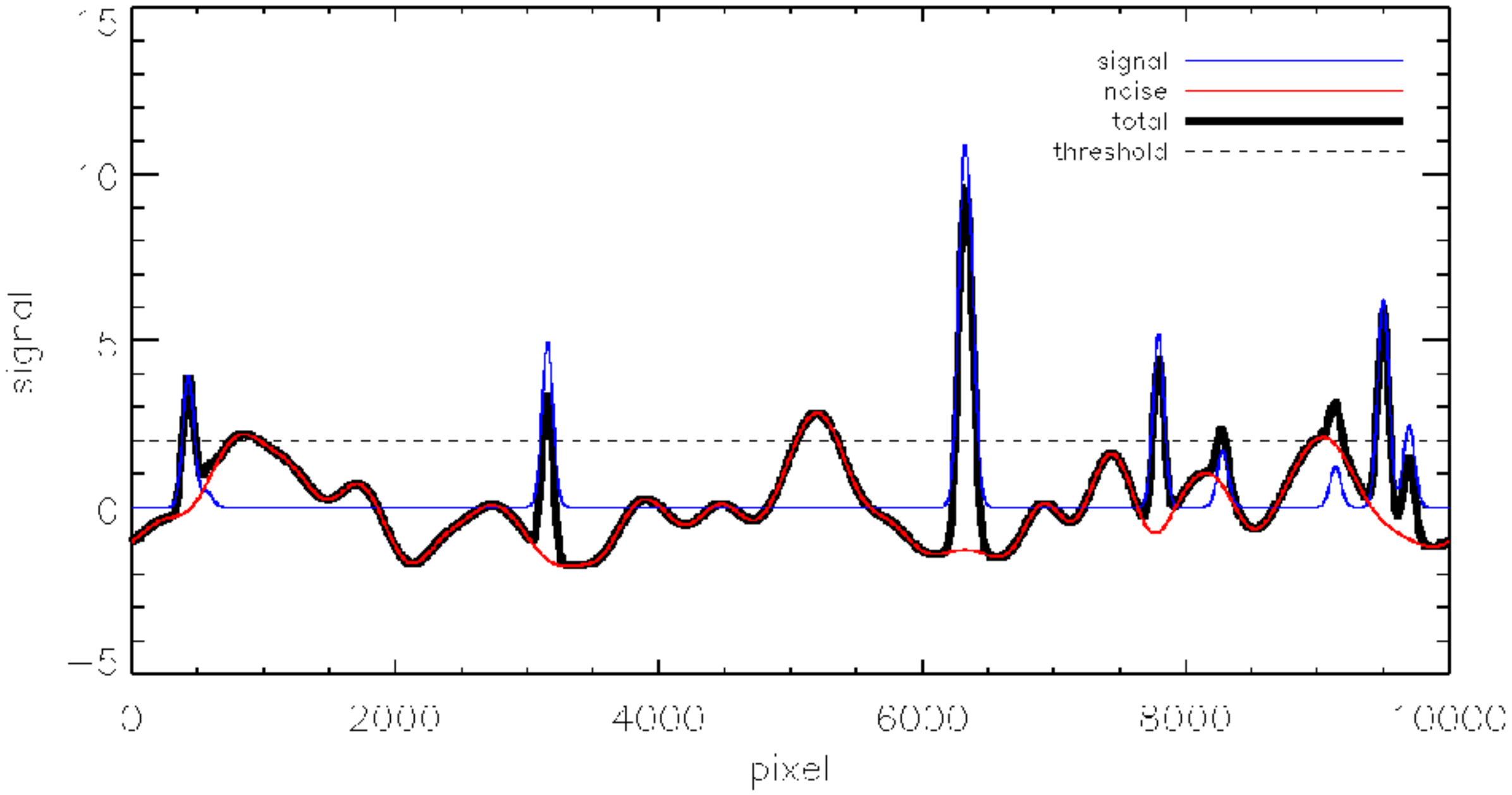
simulation exercise for Thursday



discussion question

How many sources detected?

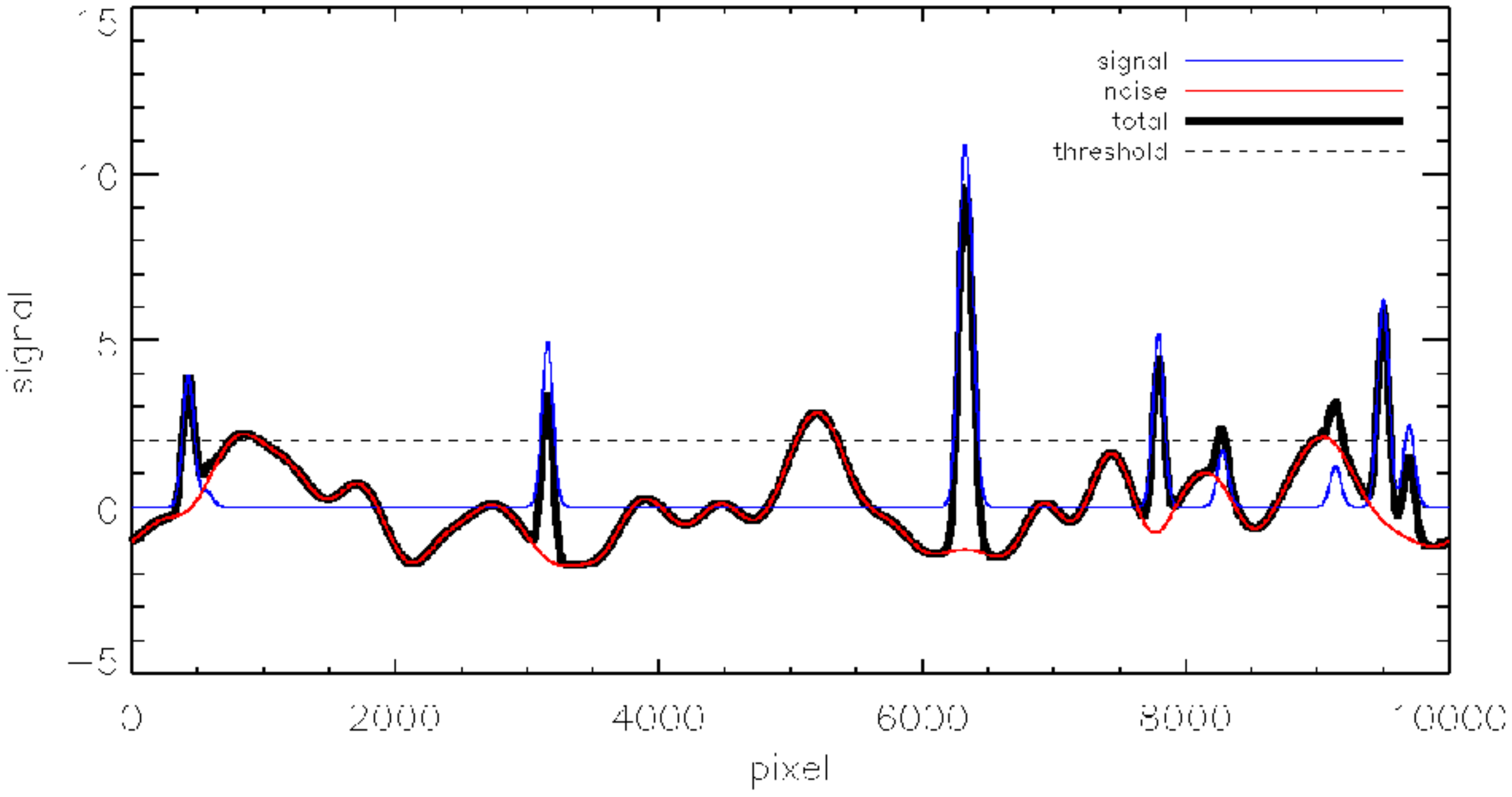
A. 6
B. 7
C. 9
D. 18



discussion question

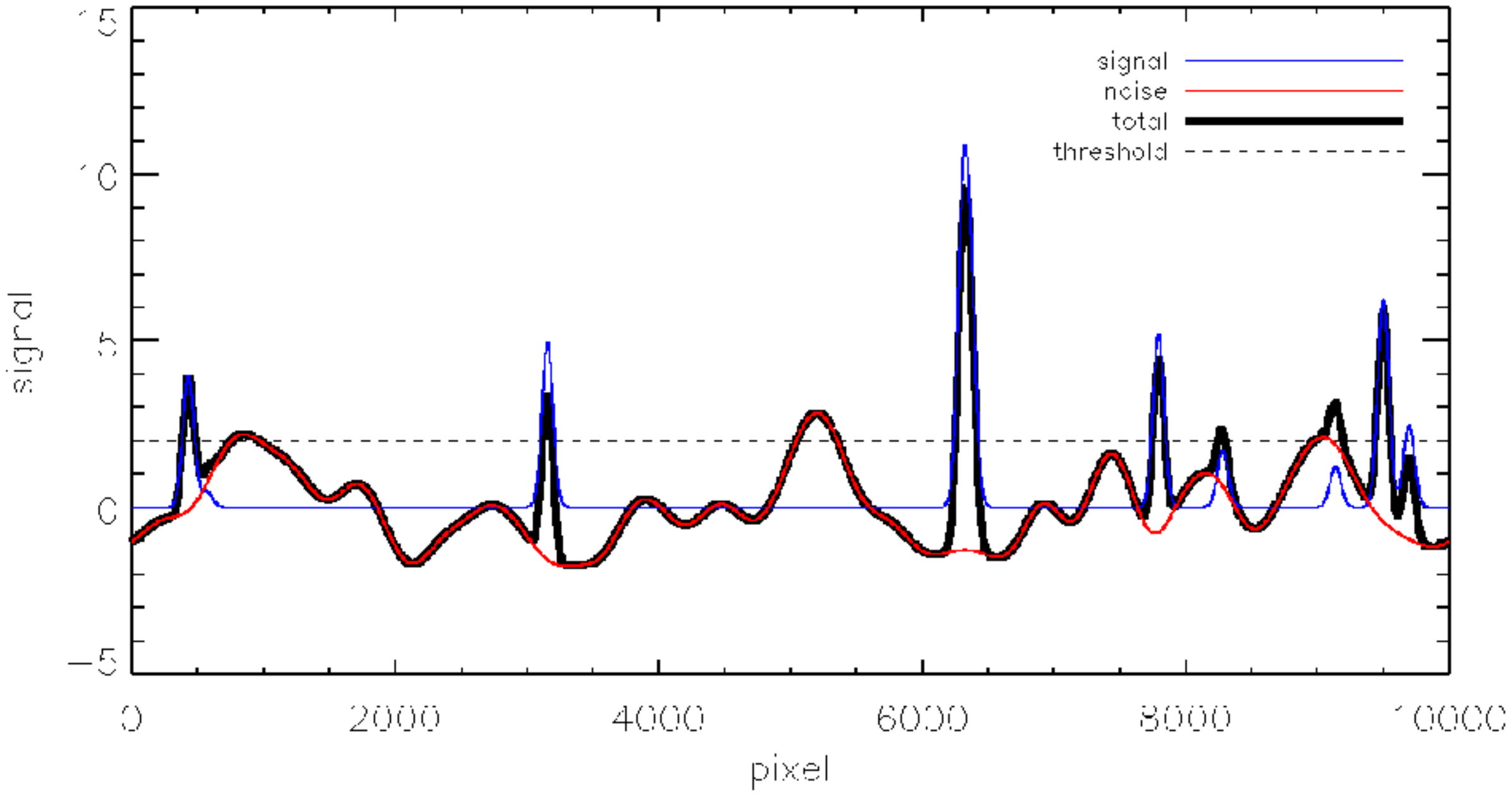
How many real sources that should have been detected are missed?

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



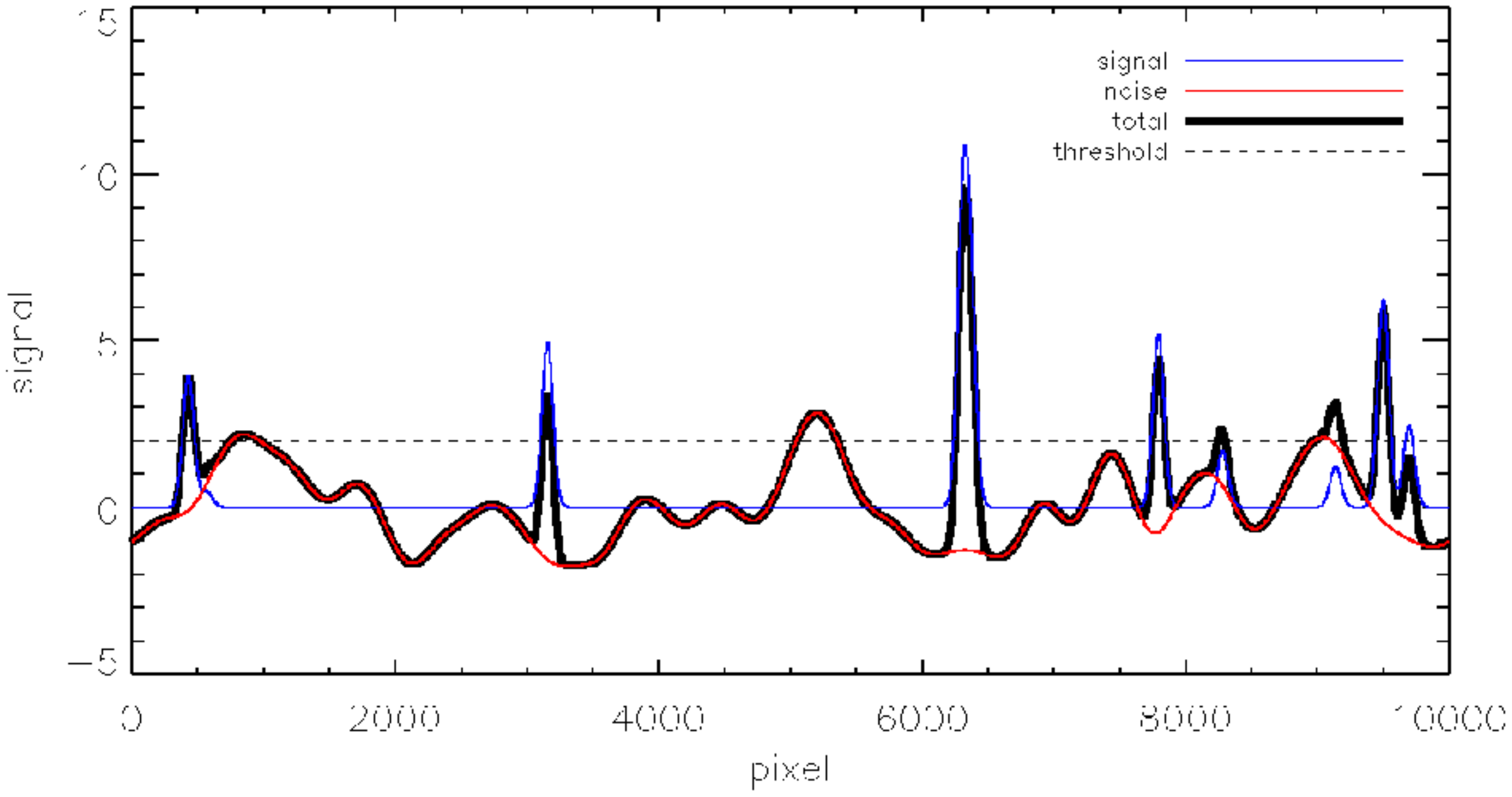
Completeness

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

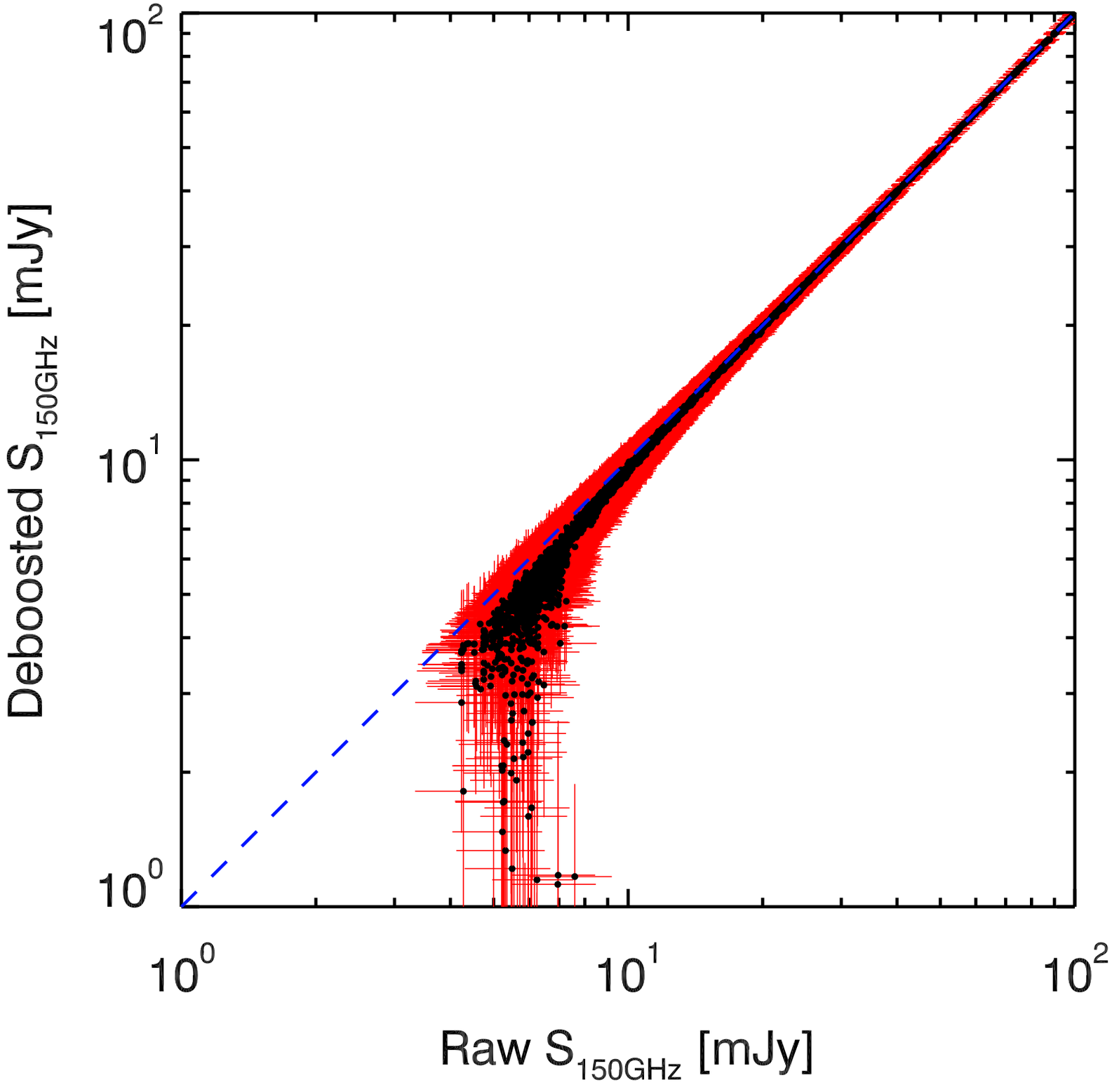


Purity

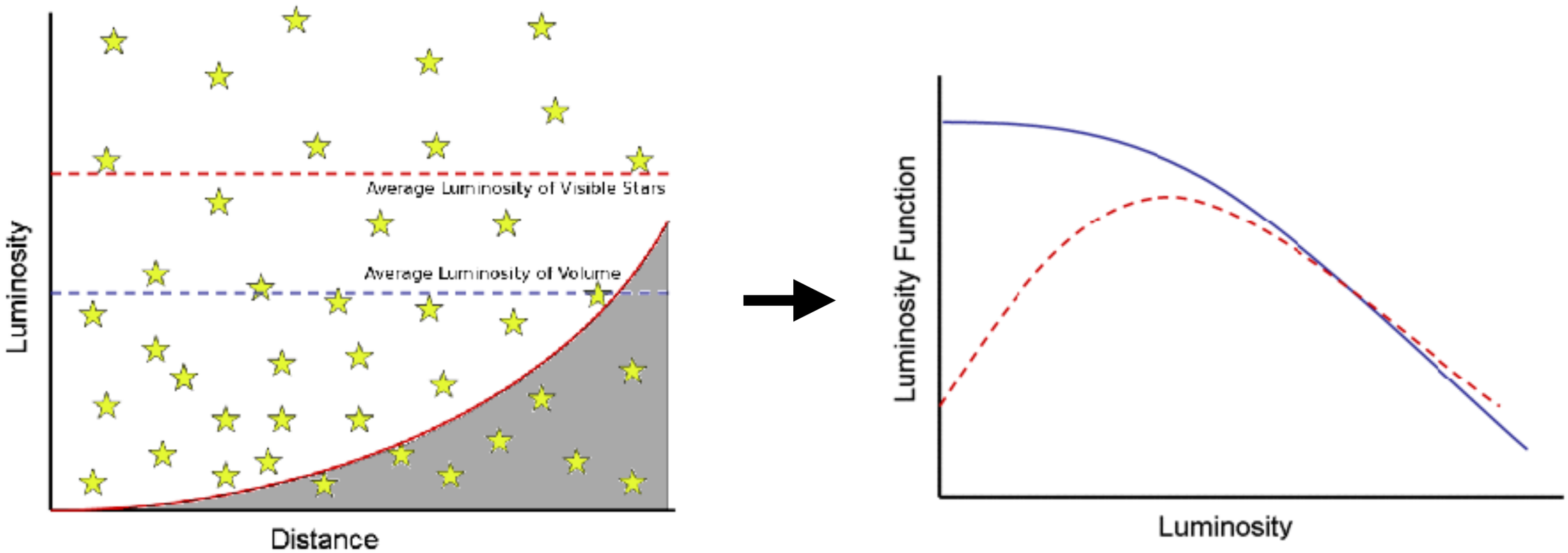
n real sources above threshold / n sources detected = 5/9 = >50%



Flux boosting bias

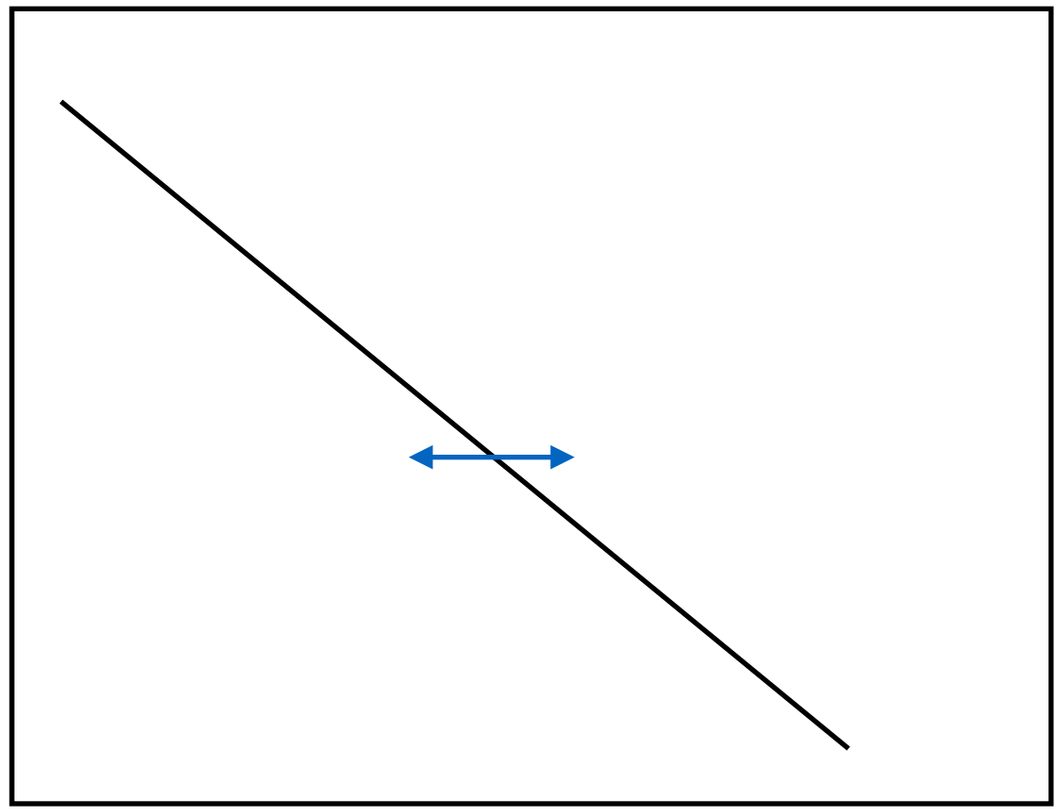


Malmquist bias



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:

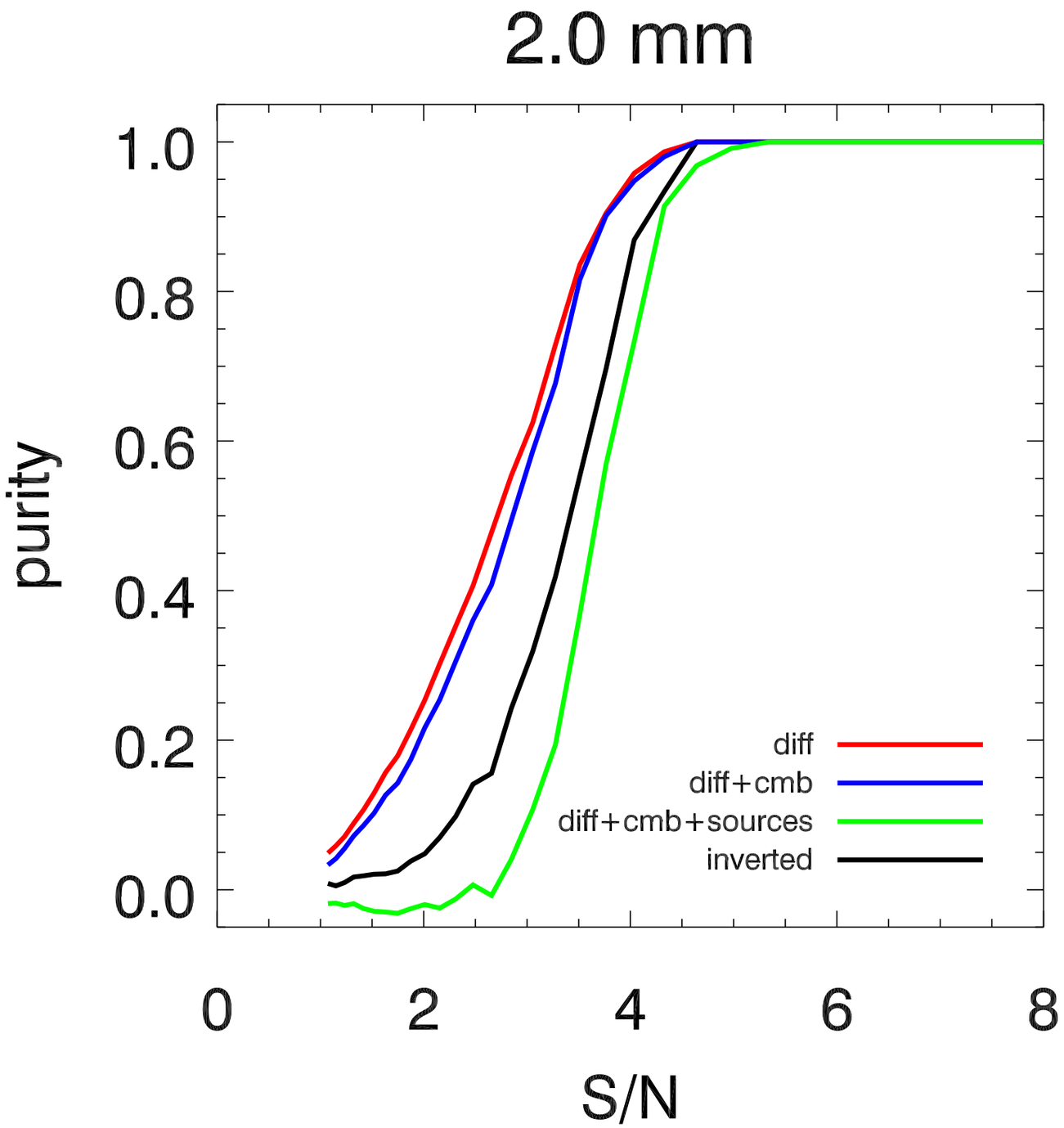
Q: What can you tell from the measurement of catalog *purity* in the plot on the right?

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>7mJy are real

D: There is a 50% chance of extracting a real source above S>4mJy.