Statistical methods for X-ray astronomy

Johannes Buchner
astrost.at/istics

28.09.2022, PHYSTAT-gamma

Workshops 2019-2022

X-ray spectral fitting methods workshop

- **When**: 24-25. September 2019
- **Where**: Max Planck Institute for extraterrestrial Physics, Garching, Germany
- **Who**: 44 participants. Speakers: Johannes Buchner, J Michael Burgess, Joer

J Michael Burgess,
Joern Wilms,
JB

VIDEO RECORDINGS, SLIDES & LINKS

BiD4BESt
School of Astro-Statistics

Successful School of Astro-Statistics

26TH JANUARY 2021 • EVENTS, STATISTICS

BiD4BESt has just enjoyed its first online school. Across two weeks five academics shared their knowledge on Bayesian Inference, X-ray Spectral Analysis, Machine Learning and its Applications to Astronomy, and Fitting the Spectral Energy Distributions of galaxies and AGN.

☆ Hol|
7 - 11 February 2022, Praha Czechia
Welcome to the X-ray Spectral Fitting (XSF) 2022 winter school page.

Peter Boorman, JB

Chandra
Data Science

Chandra Data Science:
Novel Methods in Computing and Statistics for X-ray Astronomy

Peter Boorman, JB

August 2021

Virtual (online)
Hosted by the Chandra X-ray Center

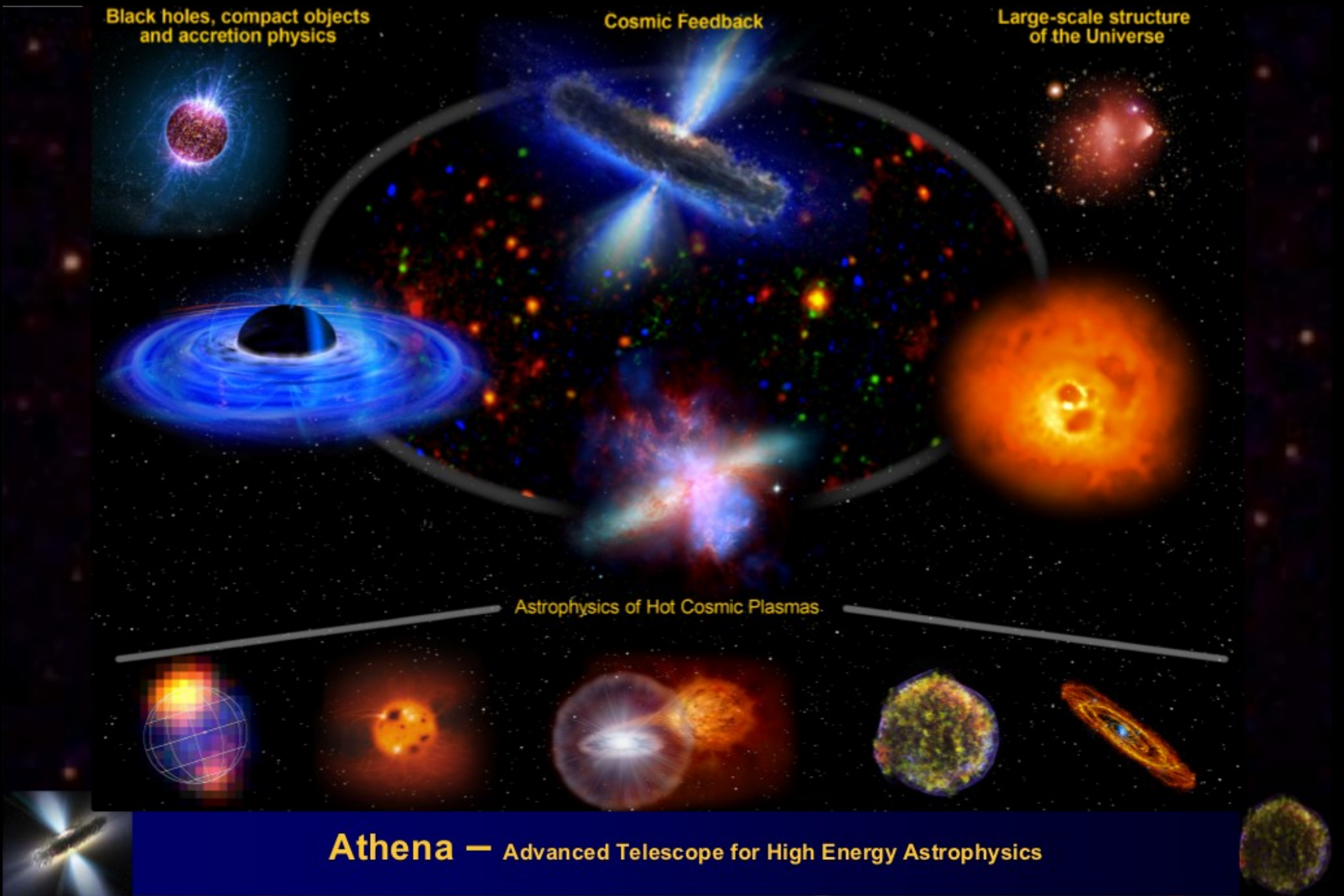
Handbook of gamma-ray and X-ray astronomy

Chapter: Statistical Aspects of X-ray Spectral Analysis

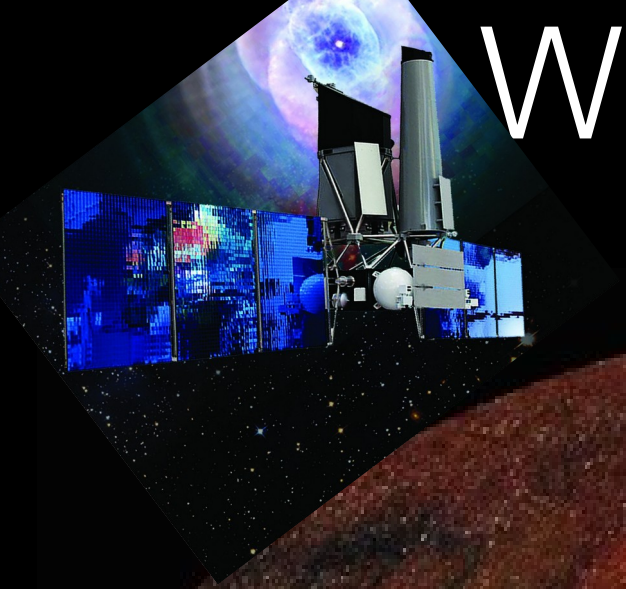
Johannes Buchner & Peter Boorman

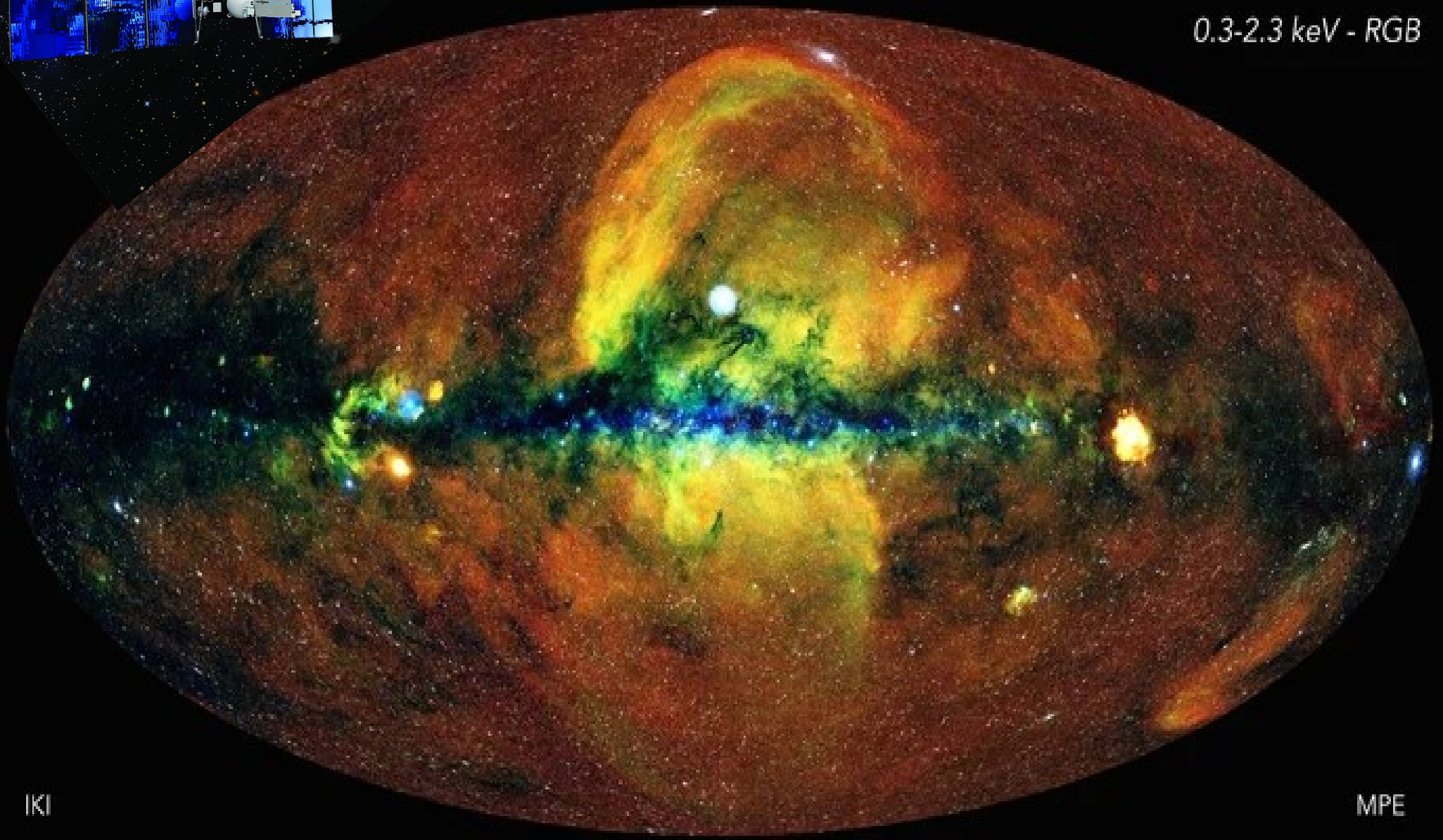
(in prep)

Why X-ray astronomy?

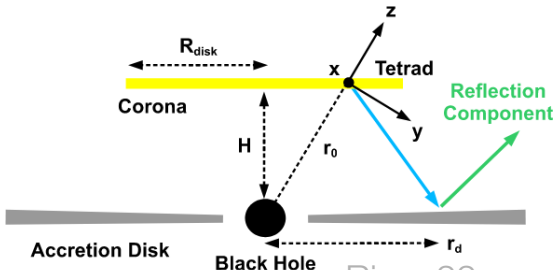


Why X-ray astronomy?

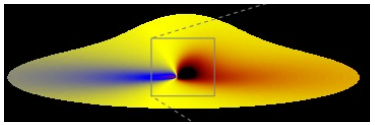




Example: Black hole spin



Riaz+22

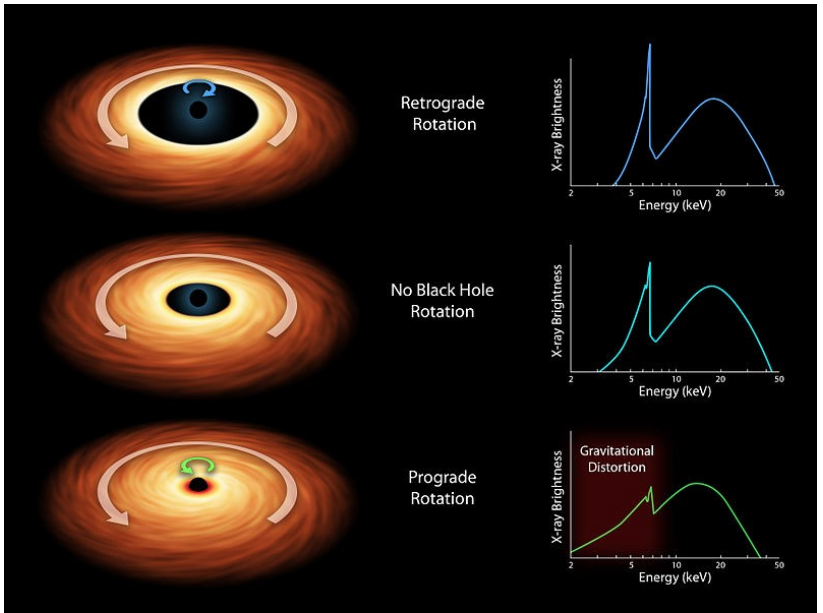


T. Dauser

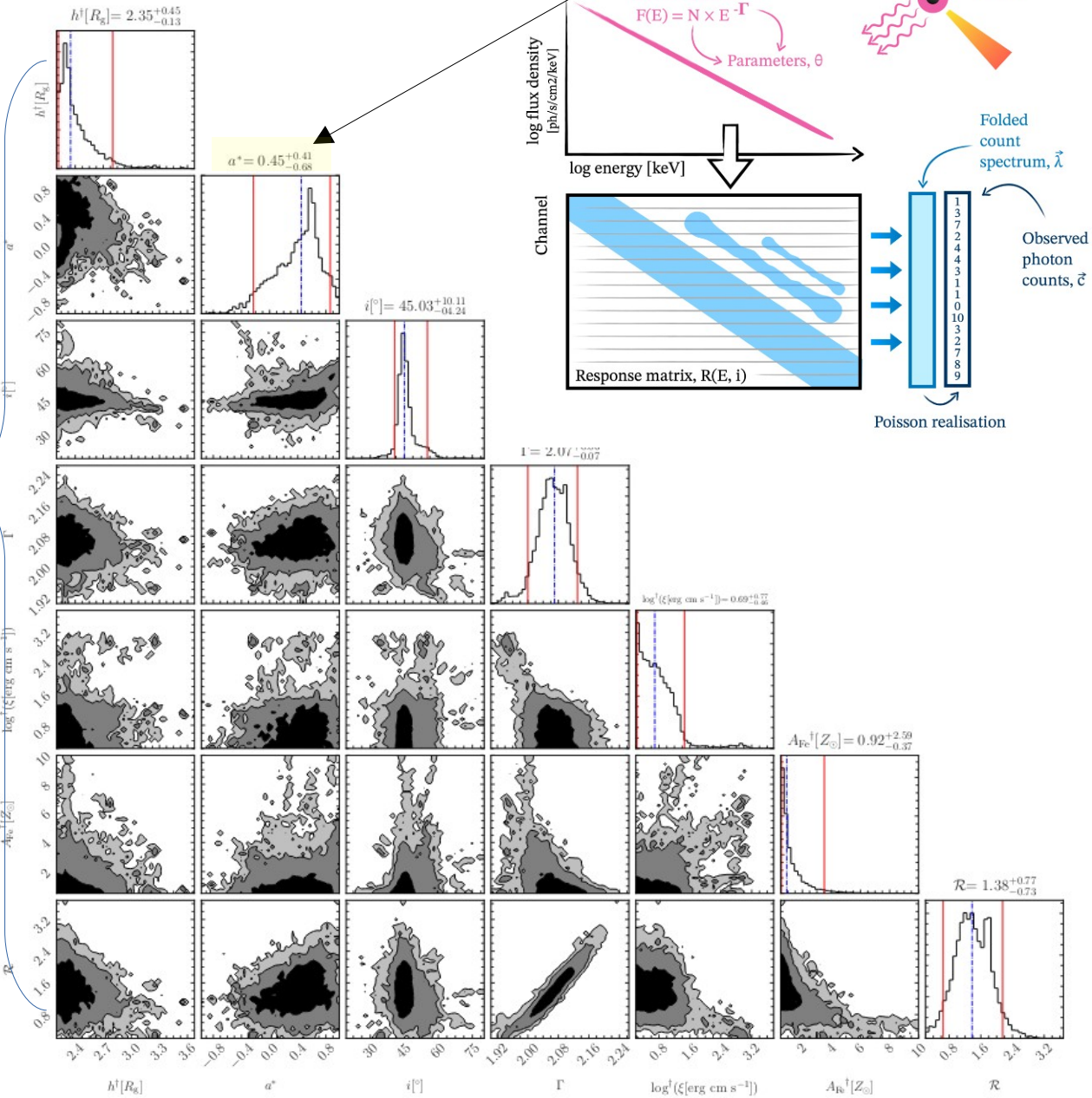
~0.2s

7 model parameters

Sisk Reynés+22



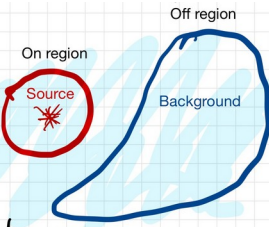
plasma height [Schwarzschild radii]



Analysis: physical models

Non-parametric

- Timing:

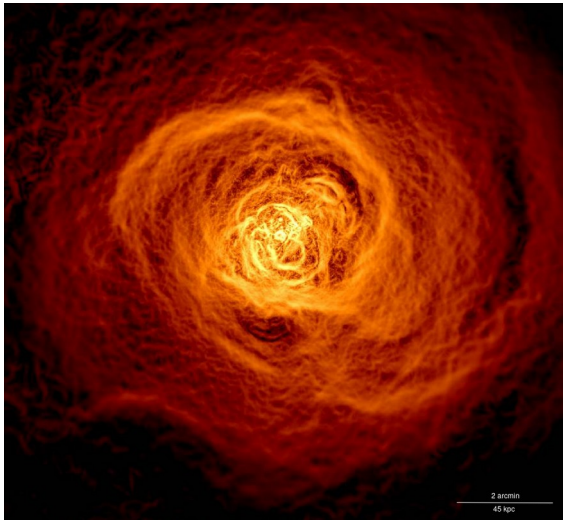


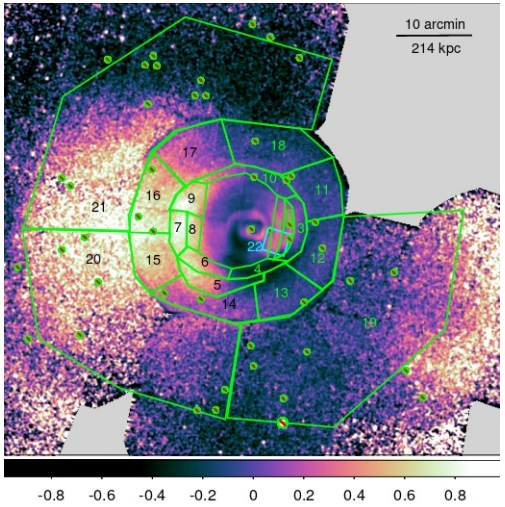
overdispersion — Buchner+21

Gaussian processes — Zoghbi+13, Wilkins+16,17

- Imaging

Sanders+16,20





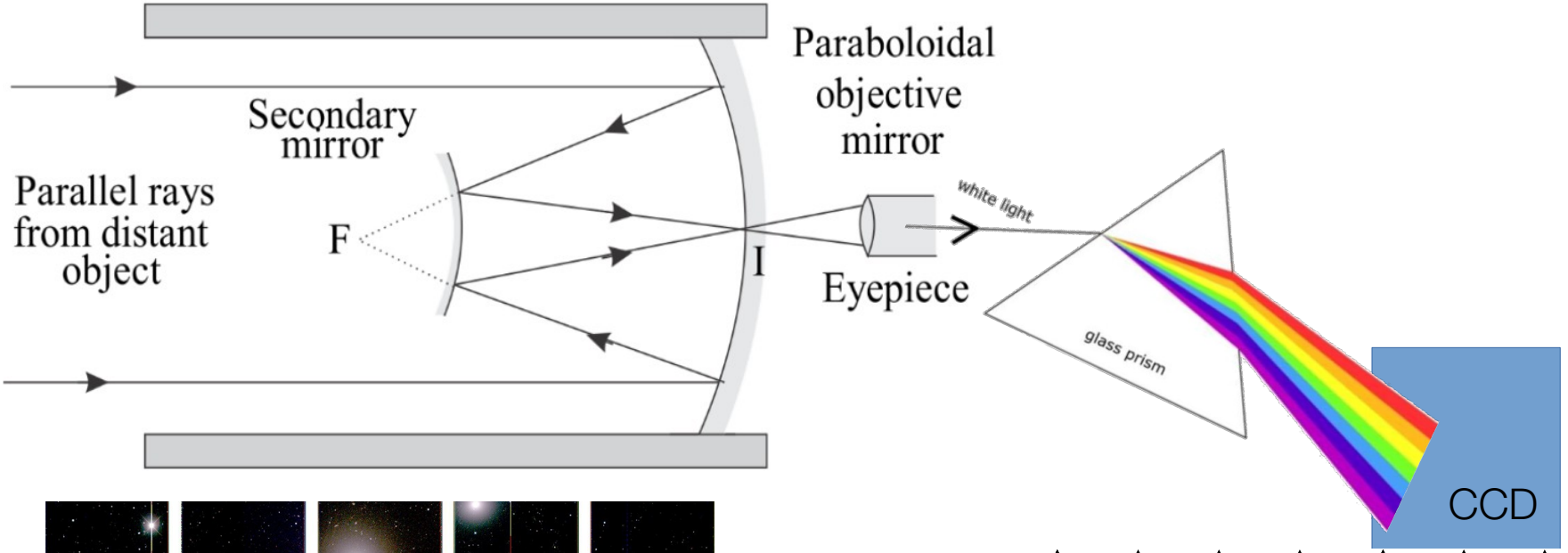
Parametric

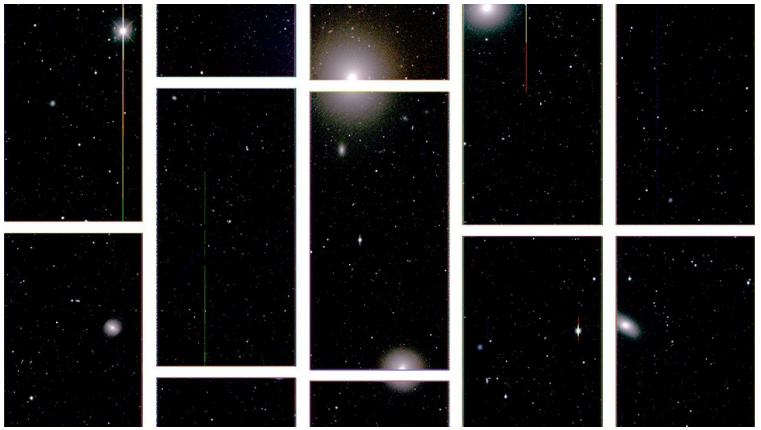
- 5-20 parameters
- Non-linear
- 1-1000ms
- non-differentiable
- Complex degeneracies
- Multi-modal

(breaks all the typical inference assumptions)

Is X-ray astronomy special?

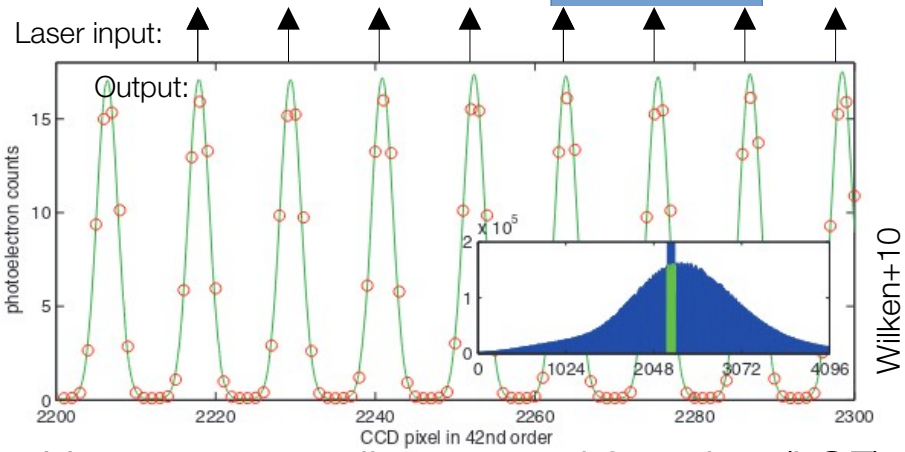
Why not use the same methods as optical astronomy?





DECAM

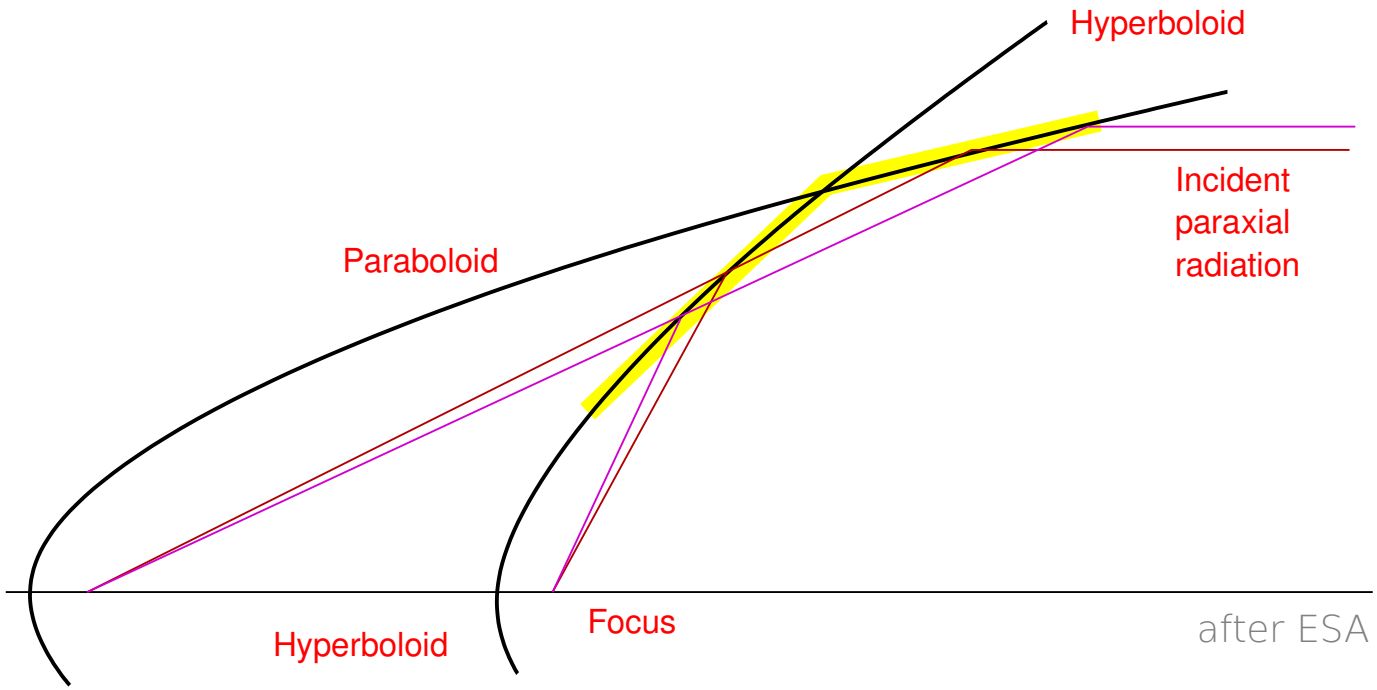
Minimal aberations,
homogeneous point-spread function (PSF)
across the focal plane



Homogeneous line-spread function (LSF)
over wavelength range

Focussing X-rays

Wolter Telescopes

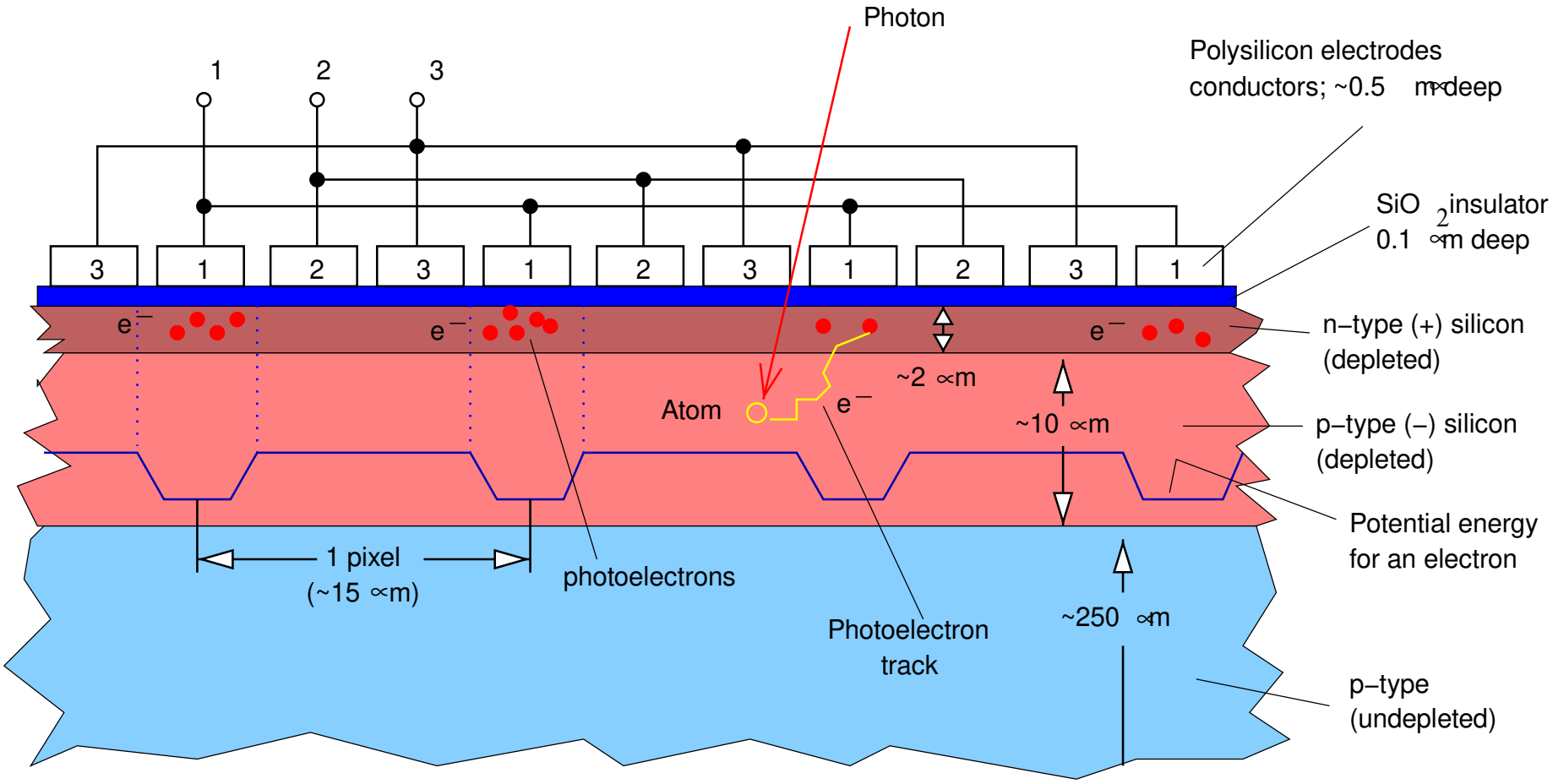


To obtain manageable focal lengths (~10 m), use two reflections on a parabolic and a hyperboloidal mirror (Wolter) type

(Wolter 1952 for X-ray microscopes, Giacconi & Rossi 1960 for UV- and X-rays).

But: small collecting area ($A \sim \pi r l/f$ where f: focal length)

Detecting X-rays

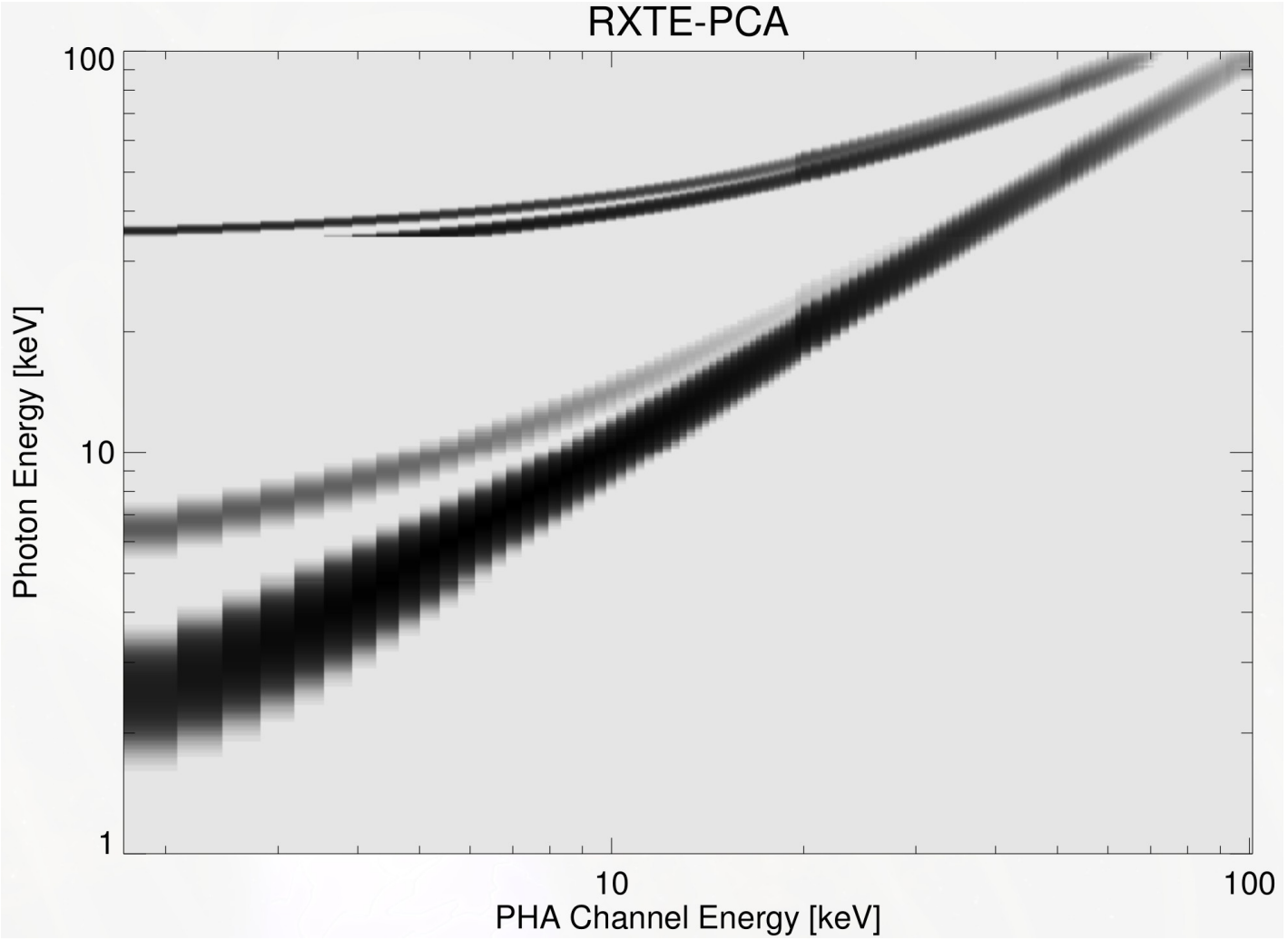


After Bradt

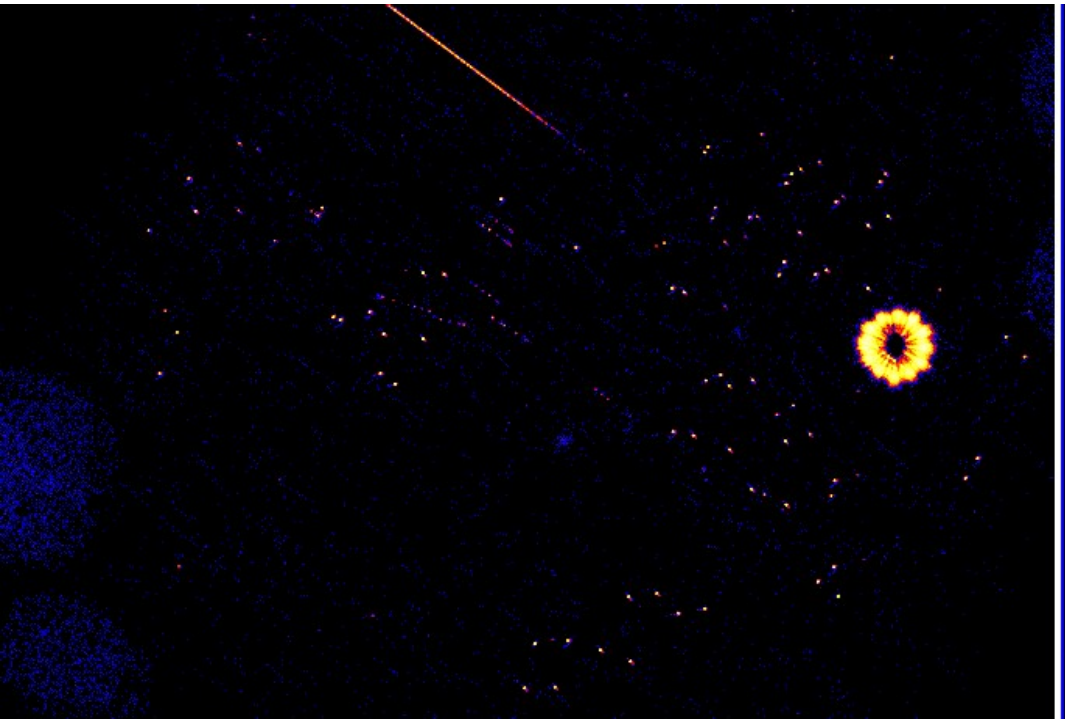
2d imaging with Charge Coupled Devices (CCDs)

Detector response matrix

(RMF)

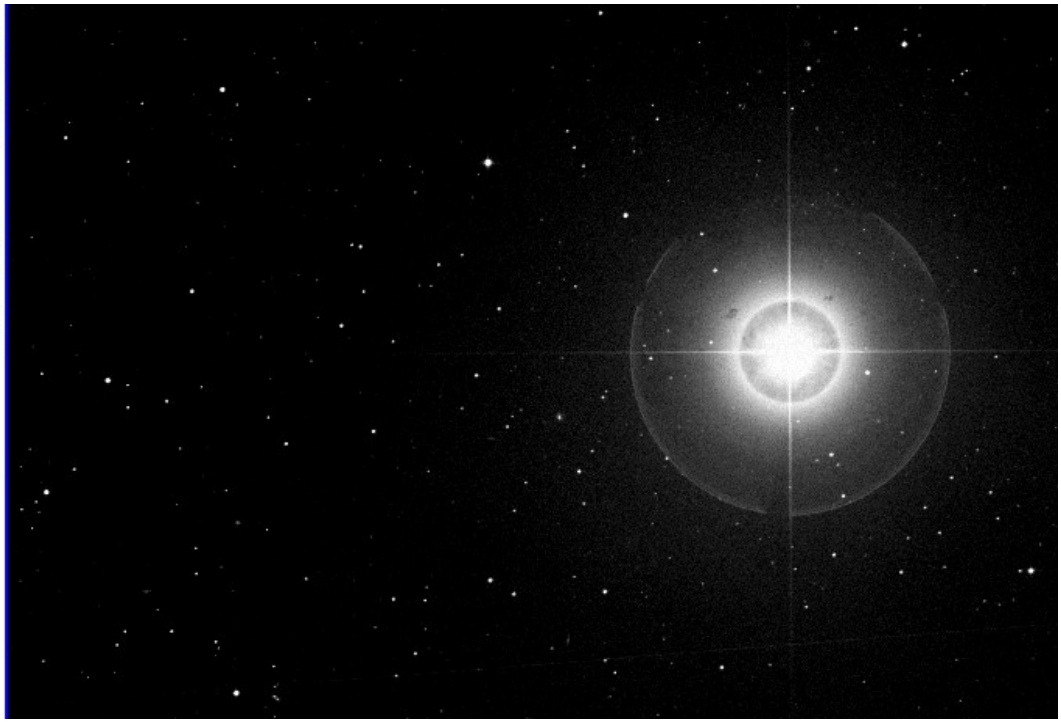


Multi-purpose data



F. Krauss

X-ray image



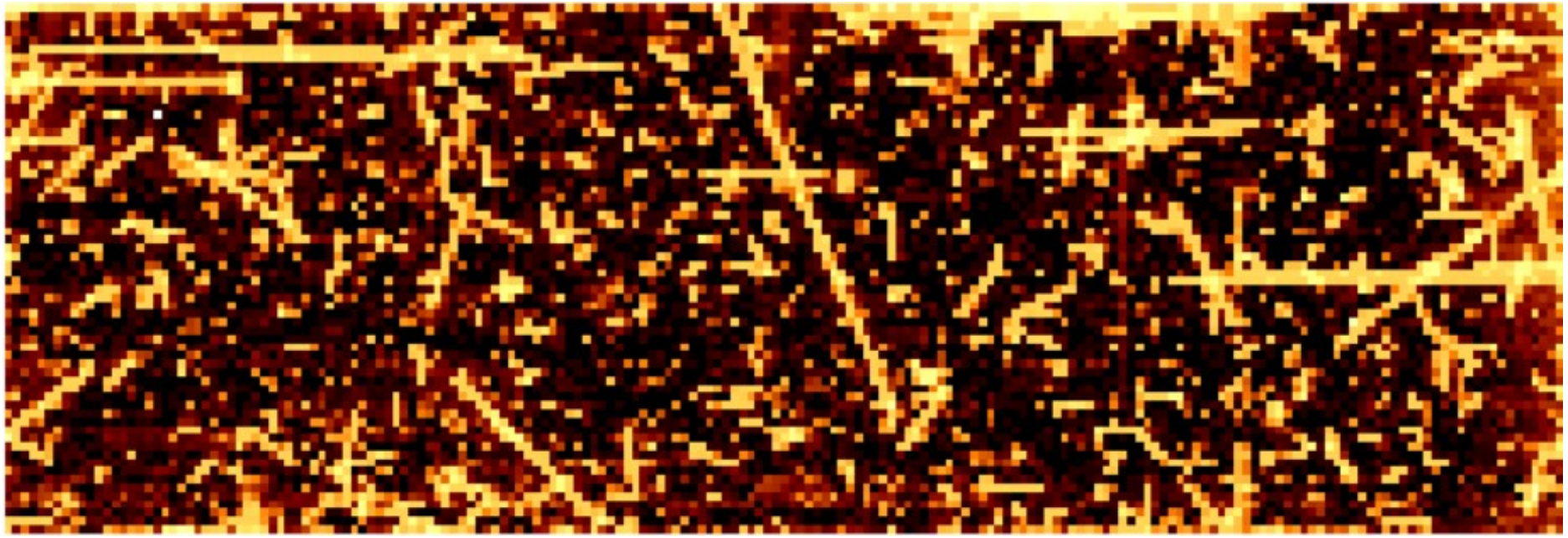
Optical image

For each detected event:
- time
- reconstructed energy
- position

Timing
Imaging
Spectroscopy
& combinations

→ Extract spectra
at image location

Background



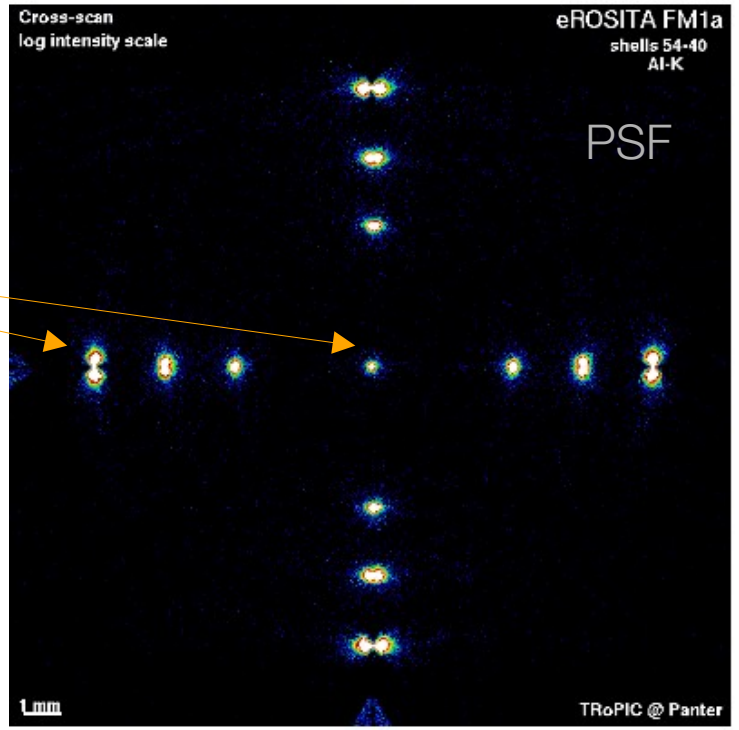
M. Wille

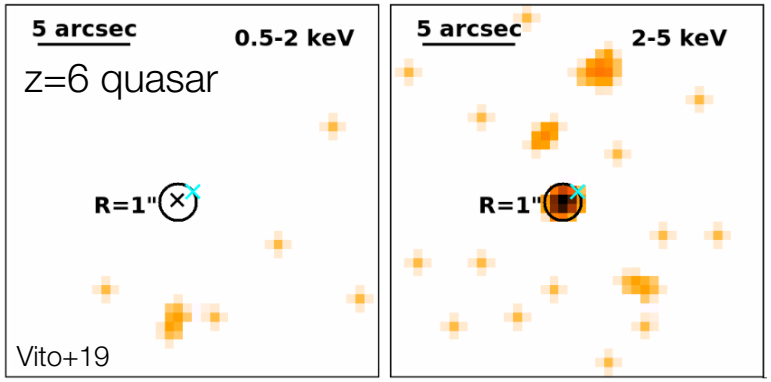
Cosmic rays & protons

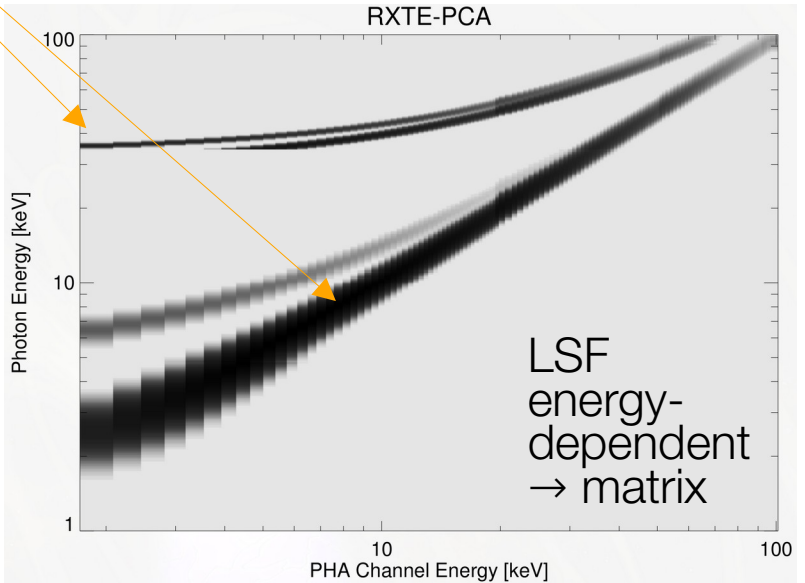
not going through the mirror

Is X-ray astronomy special?

- space-based
- Imaging response is position dependent
- Spectral response is not invertible
- Count data / shot noise

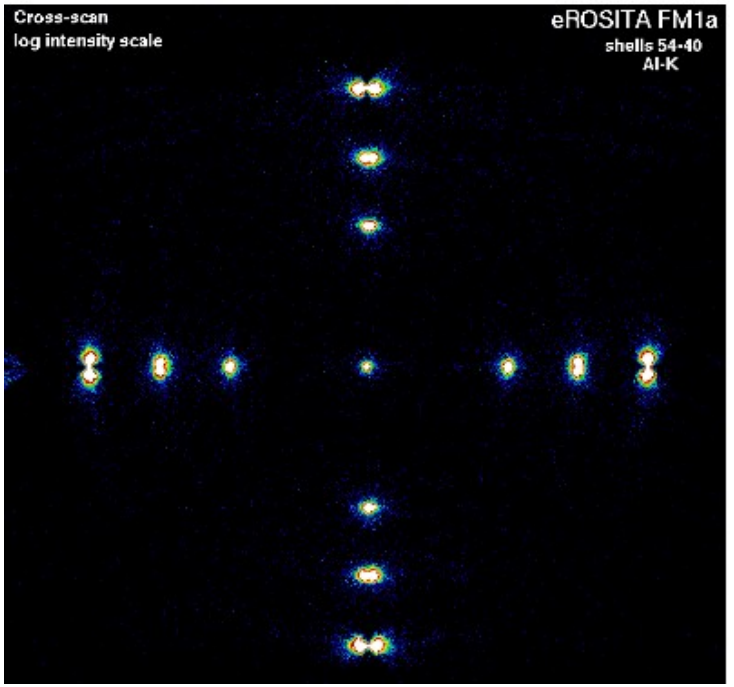


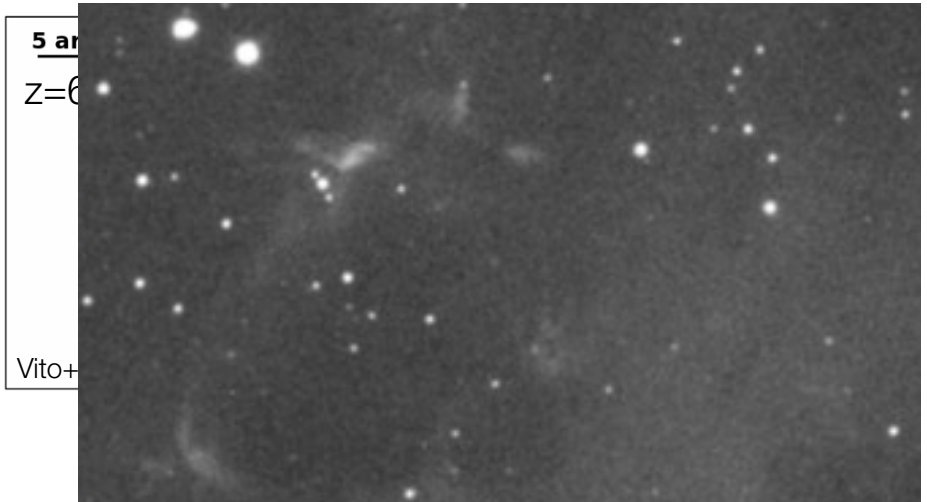


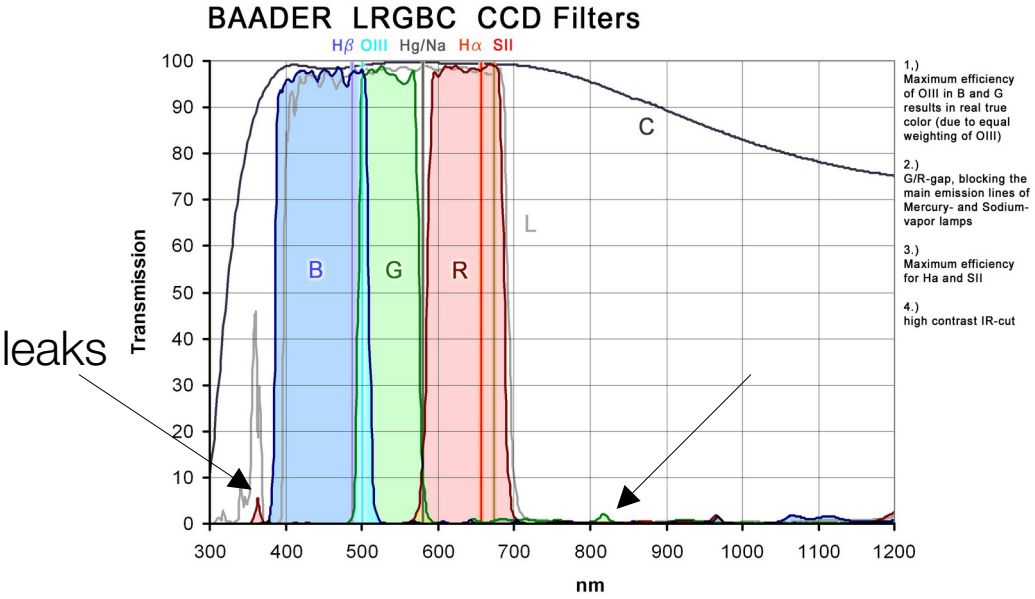


Is X-ray astronomy special?

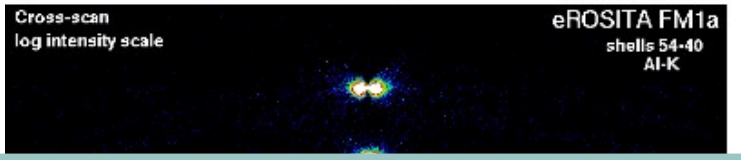
- space-based
- Imaging response is position dependent
- Spectral response is not invertible
- Count data / shot noise







Is X-ray astronomy special?

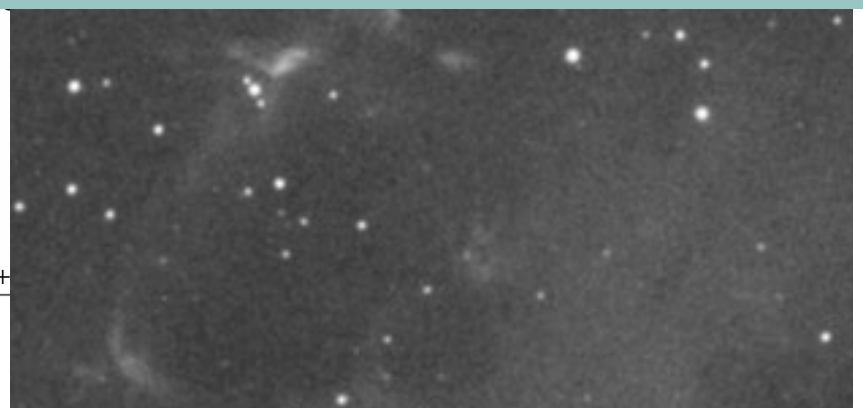


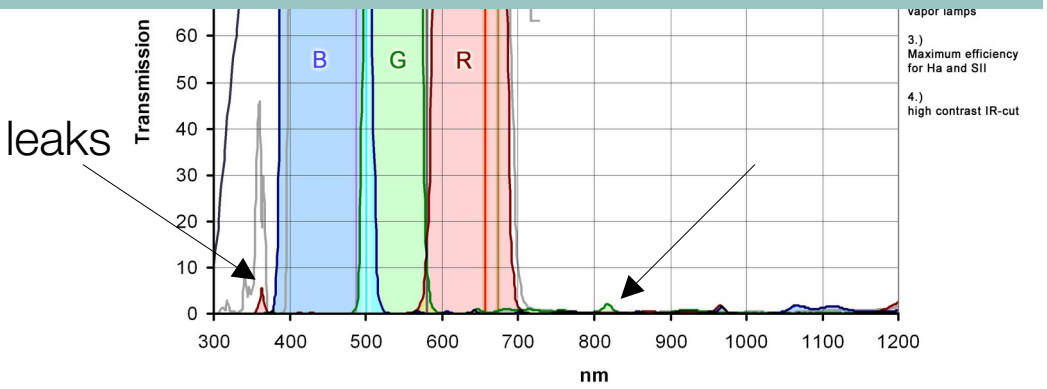
- Imaging response is

in principle, no

Optical astronomy is a special case
where you can often glance over these effects

Gaussian white noise, linear calibration

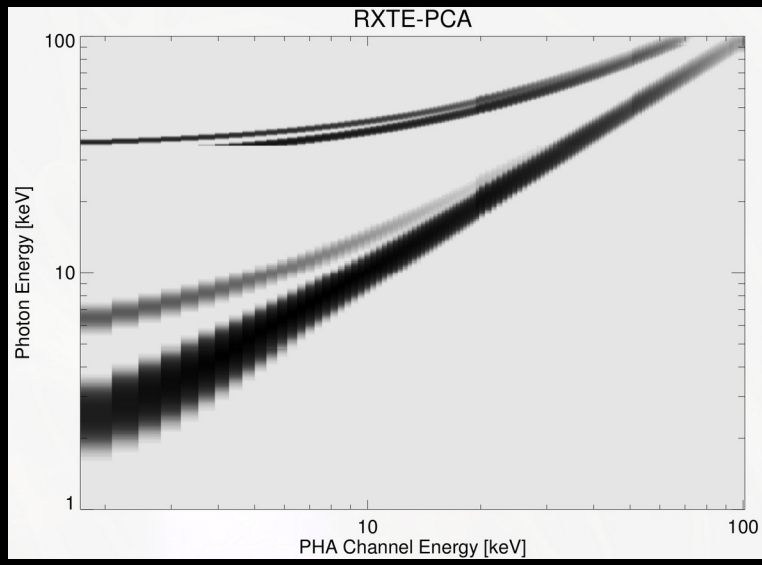




Methods

- Forward-folding
  - X-ray spectroscopy, statistical aspects
  - Likelihoods
  - Background treatment
- Workflow
- Inference methods

Story of X-ray photons



Source spectrum (physics) RMF,ARF (calibration) background Counts expected

$$(F(E) \times R(E, c) + B(E)) \times \Delta t = \lambda(c)$$

(observation)

Count spectrum observed: $k(c) \sim \mathrm{Poisson}(\lambda(c))$

Forward-fold: $P(D|\theta)$

Want: $P(\theta|D)$

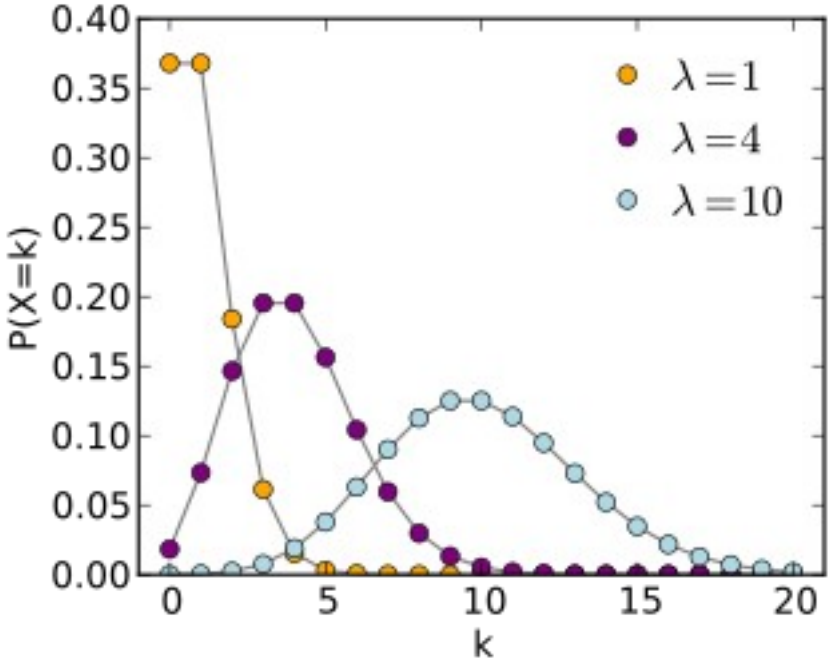
Regions of high parameter probability mass

Single spectral bin

- Poisson

  - k: integer
  - λ: real (mean&variance)
  - Asymmetric
  - Integer
  - Positive

$$P(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$



- Scaling — shape changes
- Addition — (Poisson distribution) Variability!
- Subtraction — (Skellam distribution)

Samples
Electronics (shot noise)
Photon counting (Poisson noise)

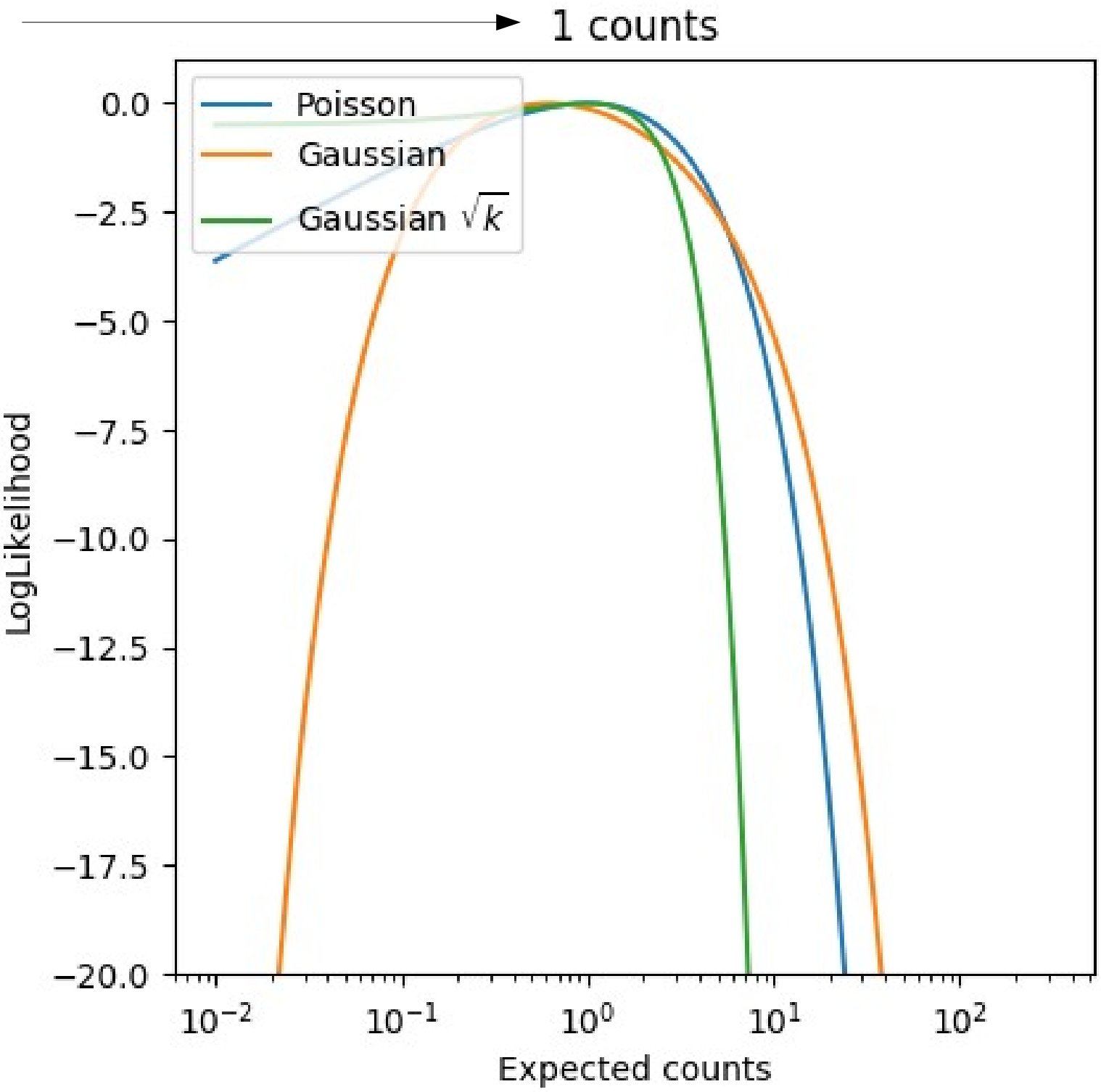
Known data

$$P(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Unknown rate

Likelihood

Probability (frequency) to produce exactly this data



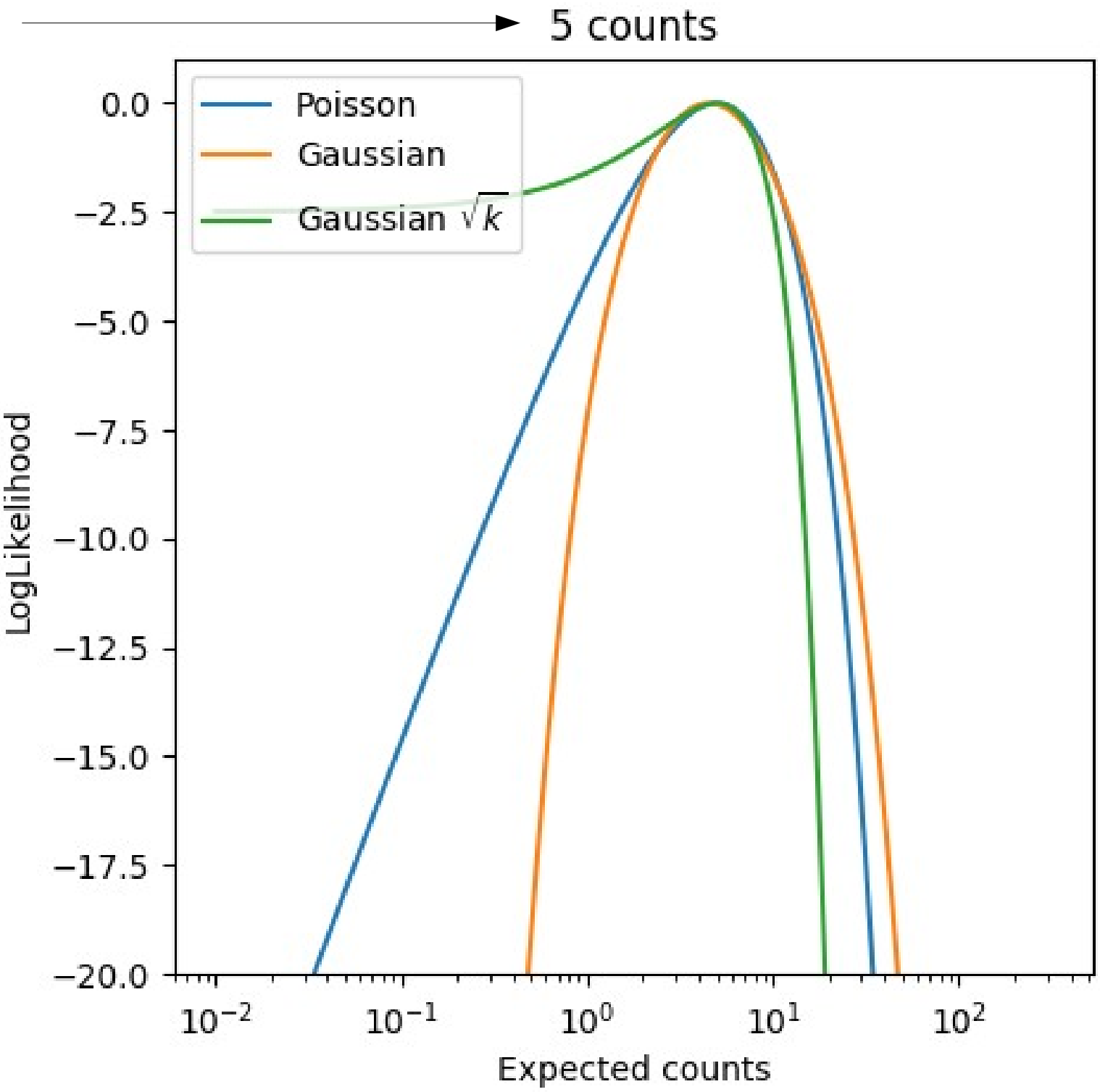
Known data

$$P(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Unknown rate

Likelihood

Probability (frequency) to produce exactly this data



Known data

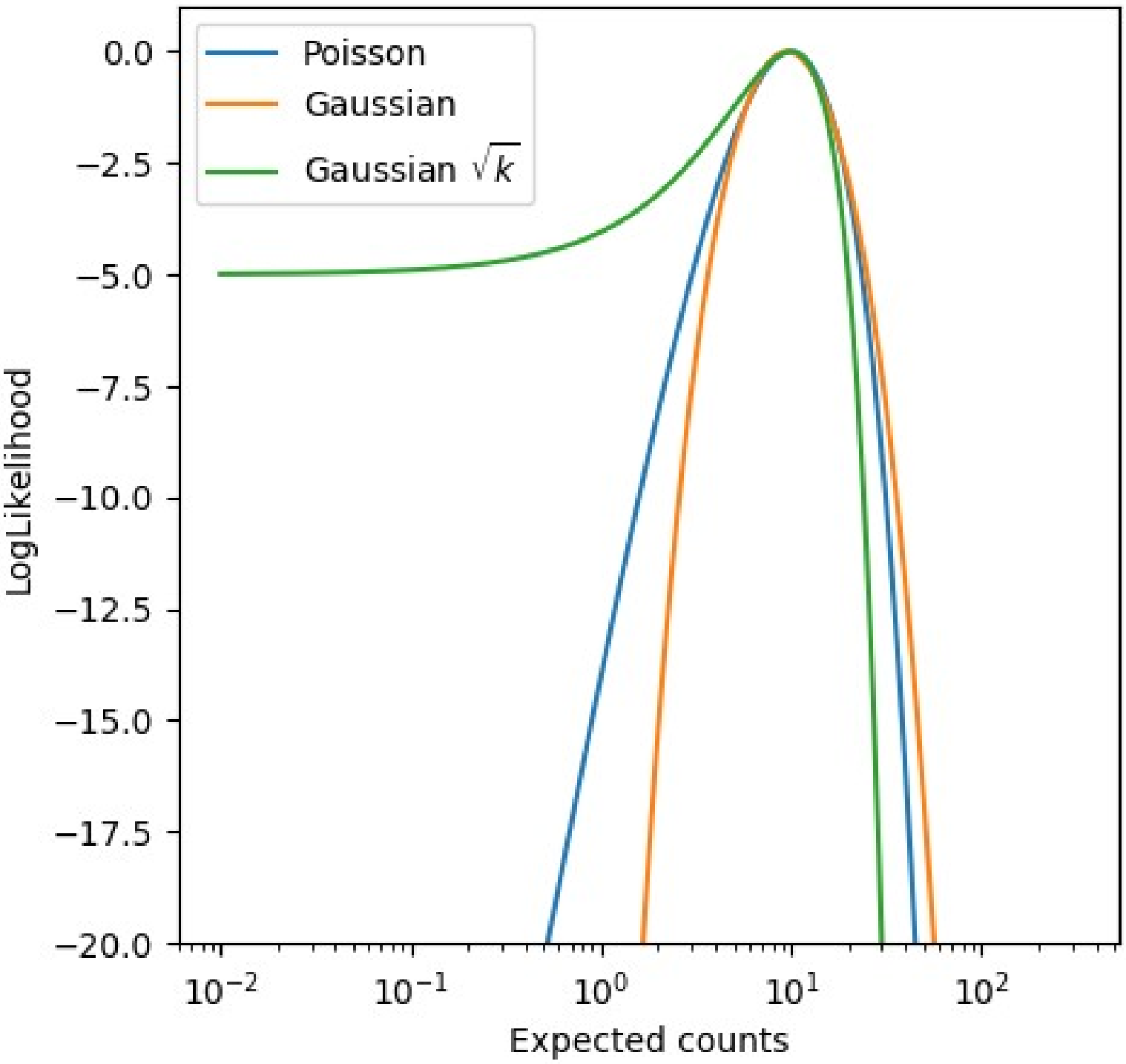
10 counts

$$P(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Unknown rate

Likelihood

Probability (frequency) to produce exactly this data



Known data

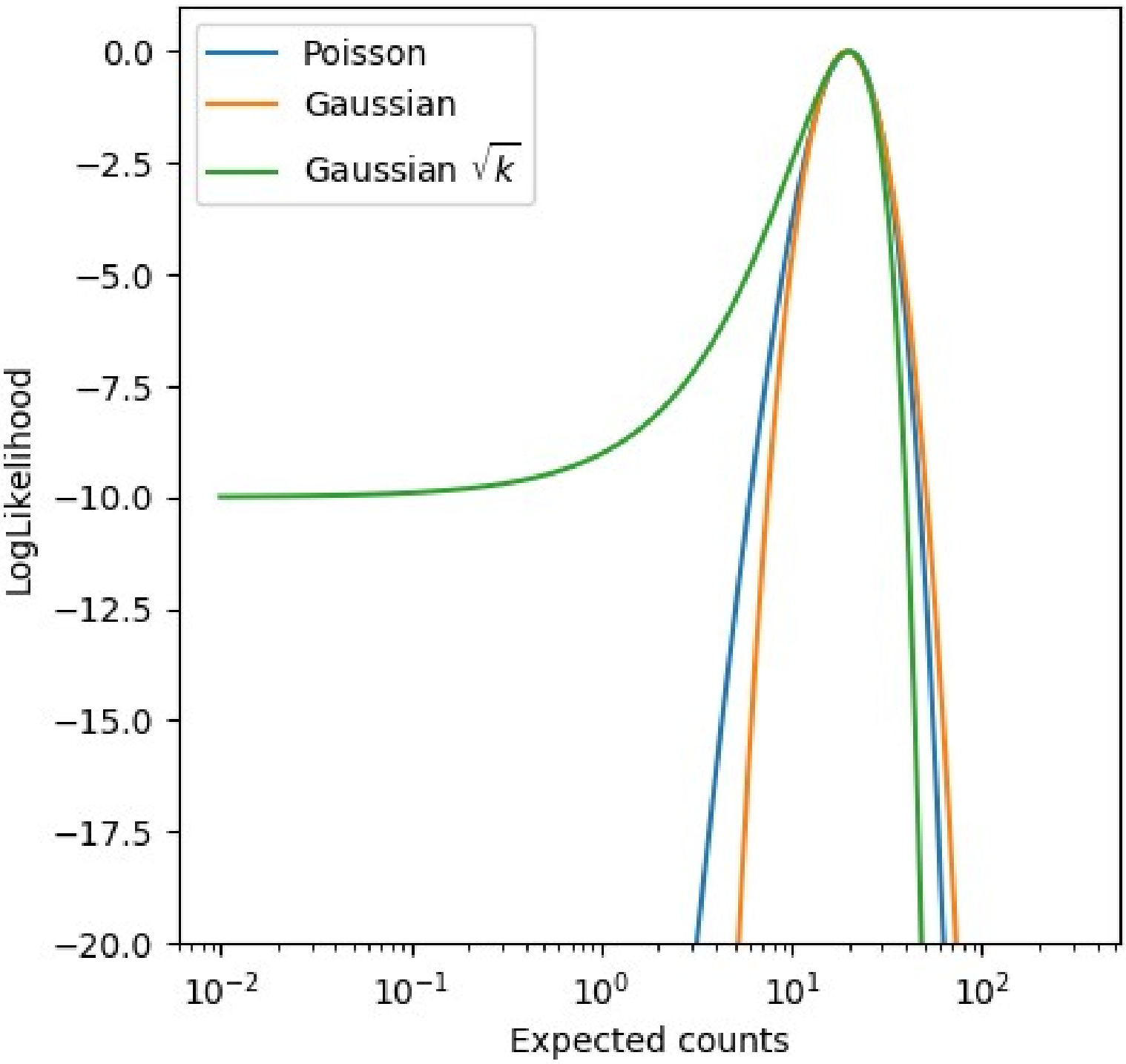
20 counts

$$P(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Unknown rate

Likelihood

Probability (frequency) to produce exactly this data



Known data

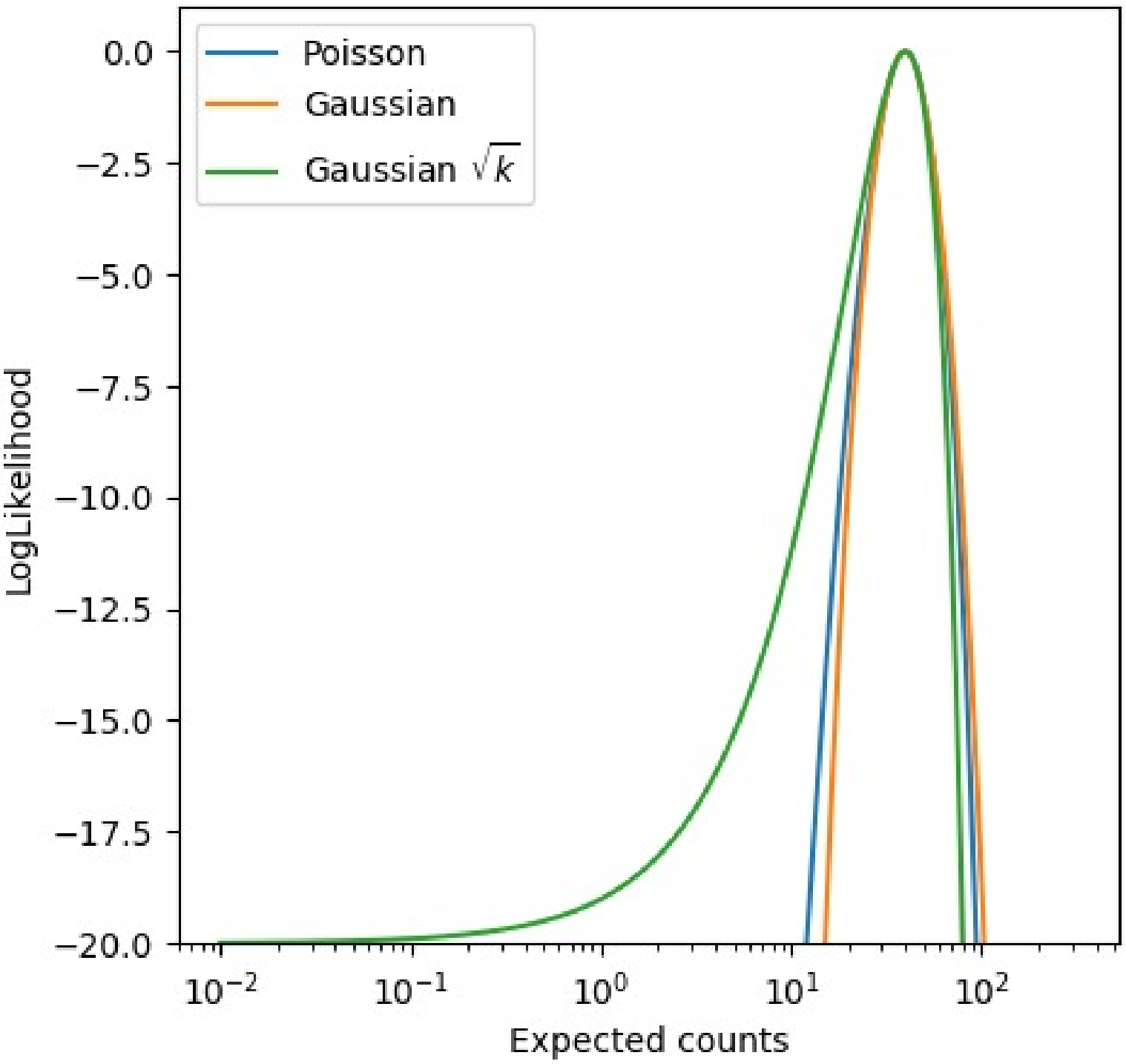
40 counts

$$P(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Unknown rate

Likelihood

Probability (frequency) to produce exactly this data



Known data

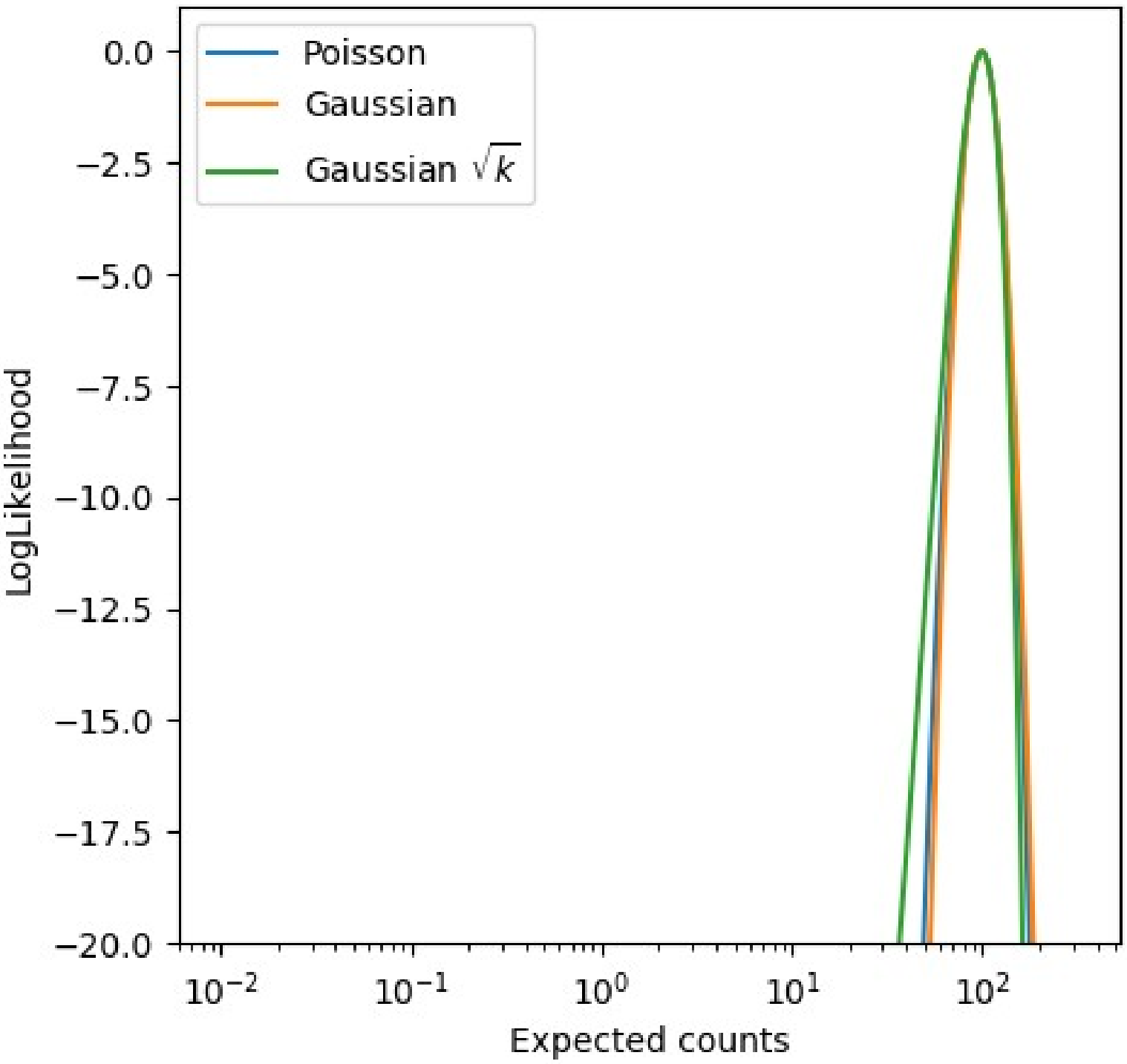
100 counts

$$P(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Unknown rate

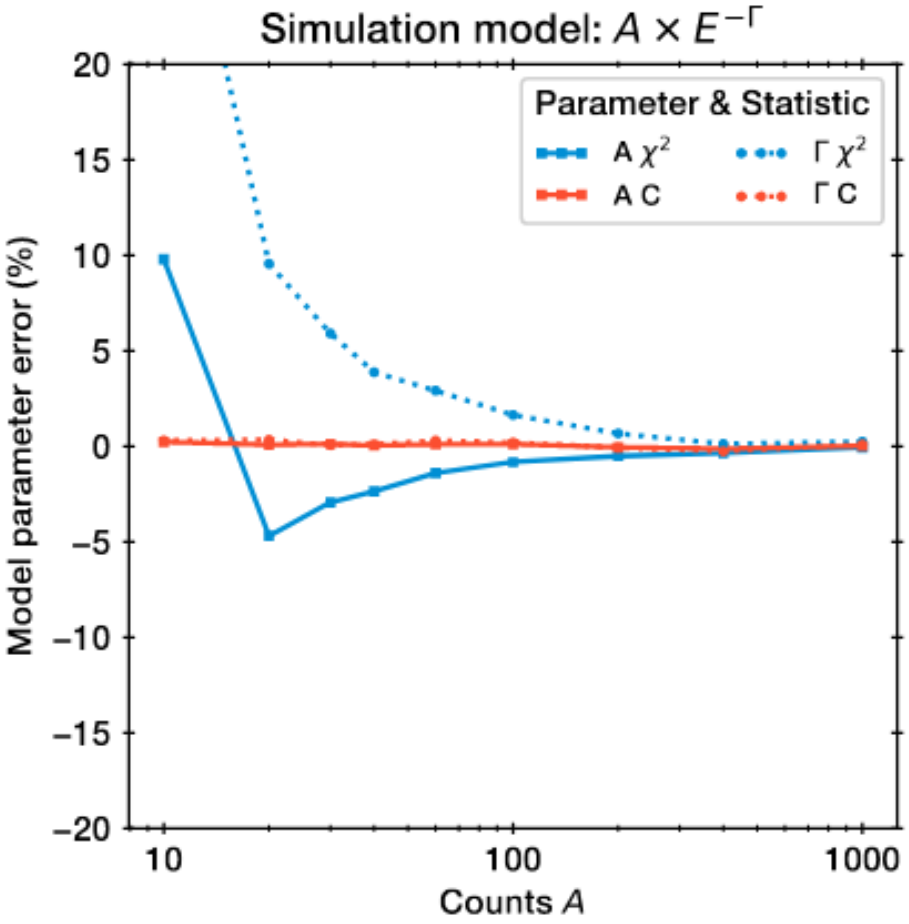
Likelihood

Probability (frequency) to produce exactly this data



Approximation quality

- Tails have different slopes
  - Gauss high-end more permissive
  - Poisson low-end more permissive
- Right way: Poisson
- Historically: Gauss faster to evaluate



Gaussian statistics are biased

Parameter biases statistically important also in the high count regime (Humphrey+2009)

Poisson statistics is unbiased (Mighell99)

→ see
Wheaton+95
Nousek&Shue1989
Mighell99
van Dyk+2001

“Statistics”

- Poisson
  - Likelihood $\mathcal{L}(k|\lambda) = e^{-\lambda}\lambda^k/k!$
    -2*log → $-2\log\mathcal{L}(k|\lambda) = 2\lambda - 2k\log\lambda + C$

- Gaussian
  - Likelihood $\mathcal{L}(k|\mu,\sigma) = \exp[-((x-\mu)/\sigma)^2/2]/\sqrt{2\pi\sigma^2}$
    -2*log → $-2\log\mathcal{L}(x|\mu,\sigma) = ((x-\mu)/\sigma)^2 + C$

“Statistics”

- Poisson
  - Likelihood $\mathcal{L}(k|\lambda) = e^{-\lambda}\lambda^k/k!$
  - -2*log → $-2\log\mathcal{L}(k|\lambda) = \underbrace{2\lambda - 2k\log\lambda + C}_{\textbf{CStat, Cash}}$

Cash (1979)

- Gaussian
  - Likelihood $\mathcal{L}(k|\mu,\sigma) = \exp[-((x-\mu)/\sigma)^2/2]/\sqrt{2\pi\sigma^2}$
  - -2*log → $-2\log\mathcal{L}(x|\mu,\sigma) = \underbrace{((x-\mu)/\sigma)^2}_{\textbf{Chi}^2} + C$

Does not mean they follow a chi² distribution!

Inference with likelihoods

-0.5 Cstat, -0.5 chi^2

$$\mathcal{L}(\overrightarrow{k}|\theta_1, \theta_2, ..., \theta_d, M, R, B, ...)$$

Higher L: model under these parameters often makes this data
Lower L: less frequently

→ Frequency of data

$$P(D|\theta)$$

Likelihood function at D, at parameter values (not a density)

Model checking

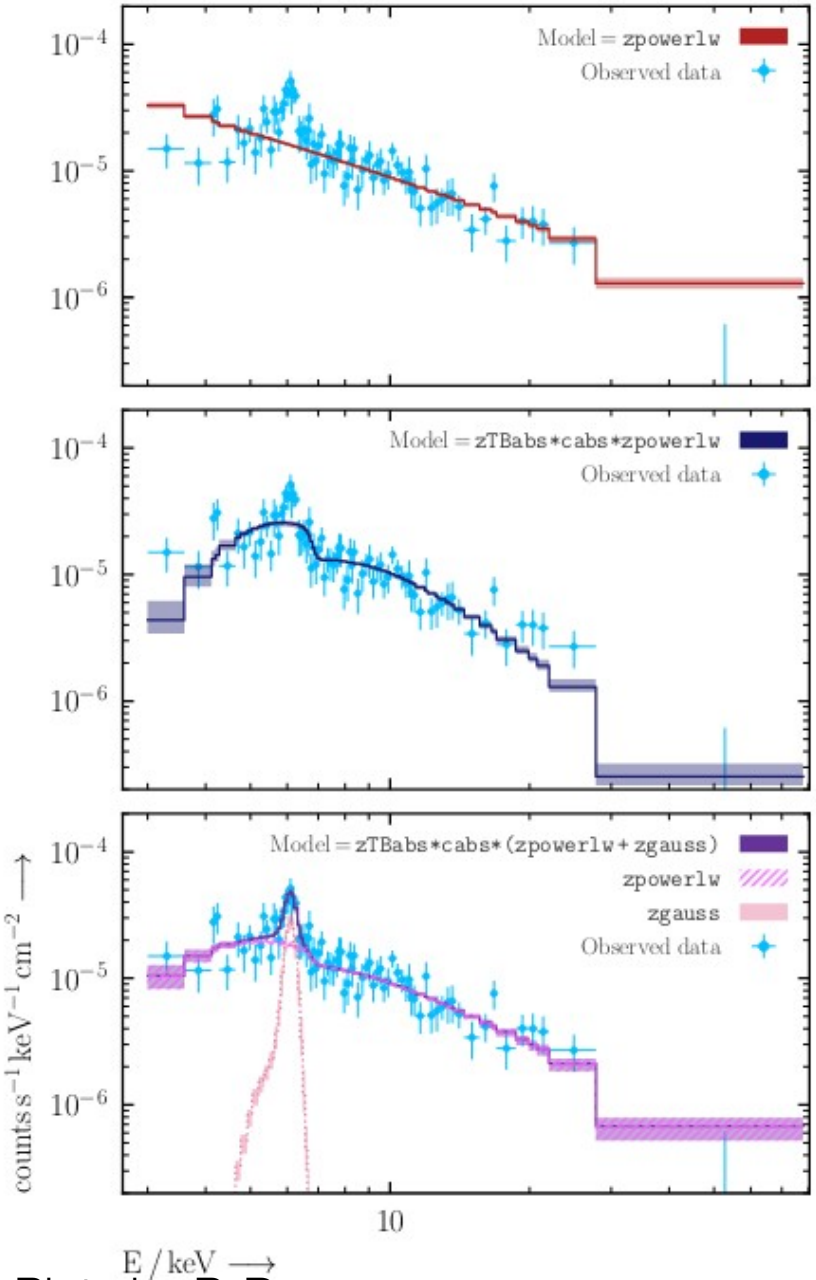
- Chi^2 test: dof in non-linear model?
- Cstat test → Kaastra2017
- Binning?

- Alternatives:

  not everything has to be a test:
  use visualisations & domain expertise

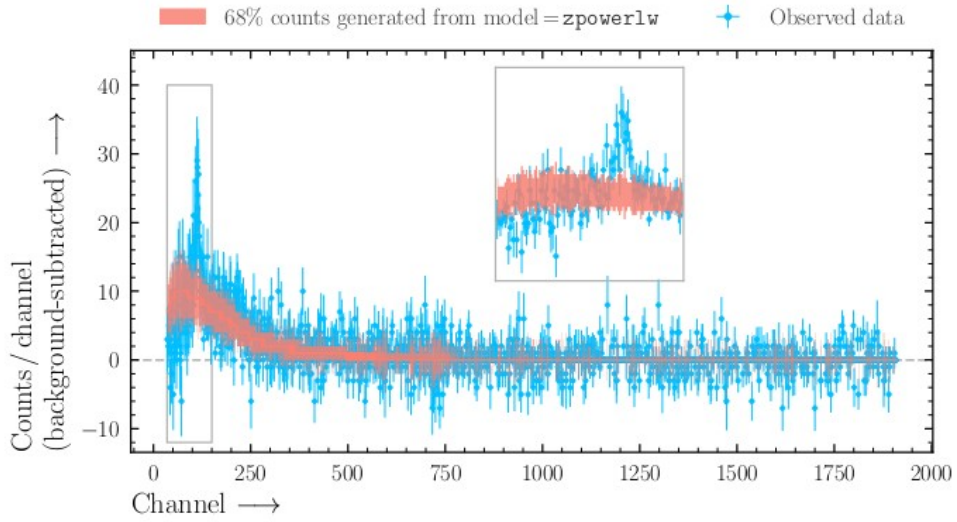
  Add flexible empirical components,
  do model comparison (easier)

Visual model checking

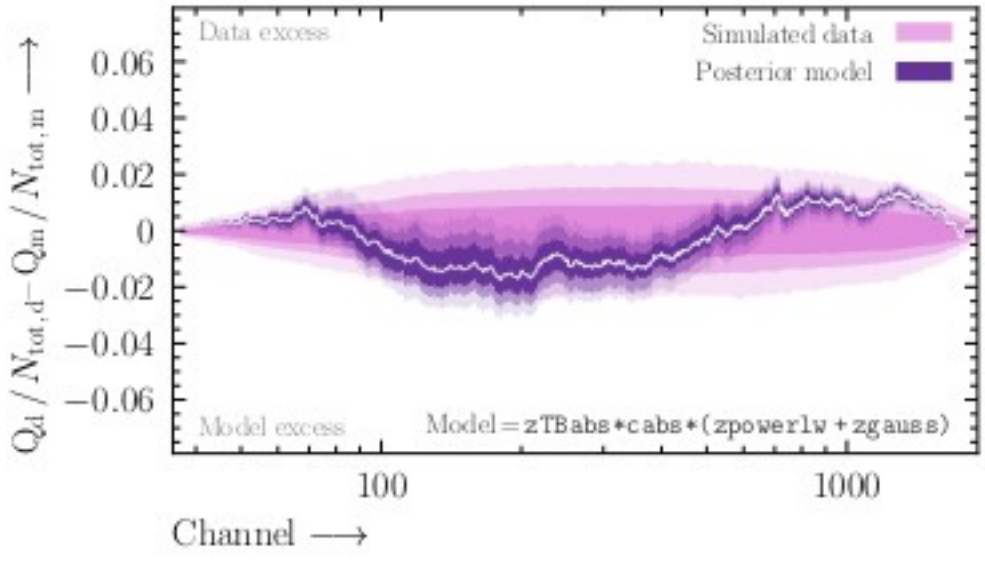


Plots by P. Boorman

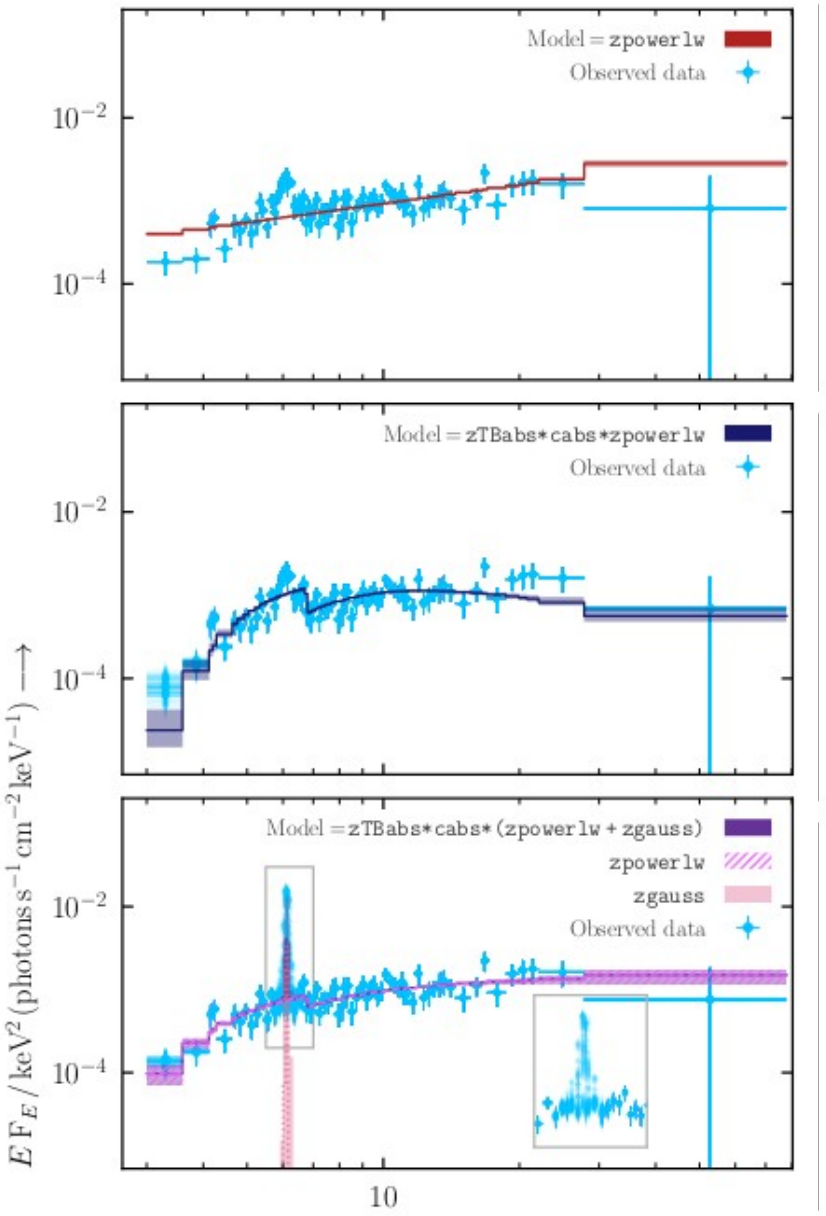
posterior predictive checks (PPC)



Q-Q plot:
cumulative model
vs cumulative data

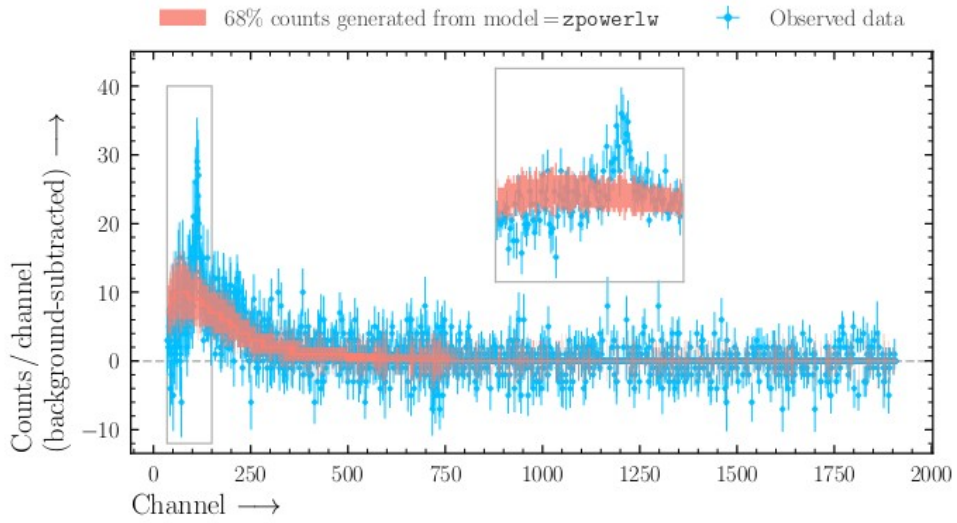


Visual model checking

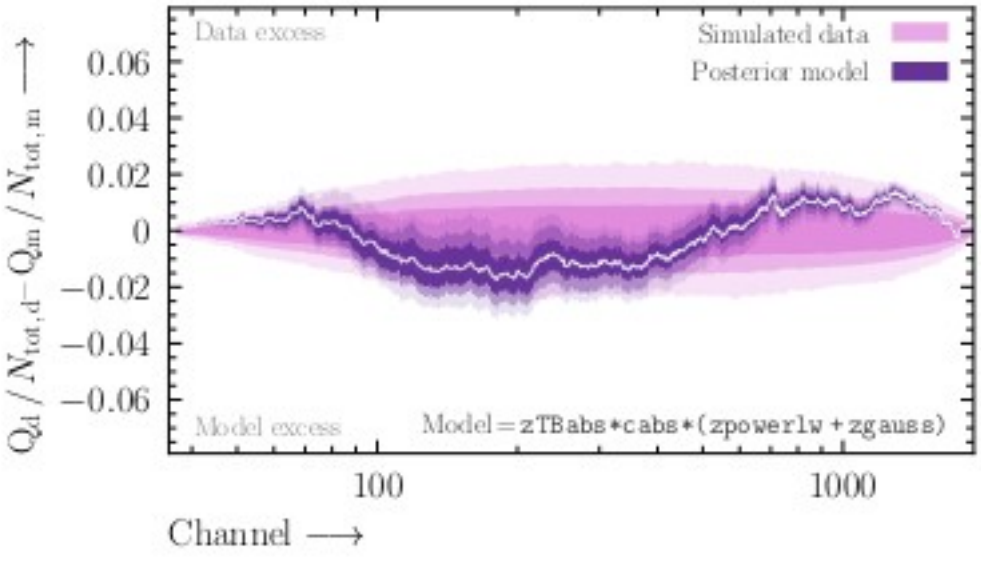


Plots by P. Boorman

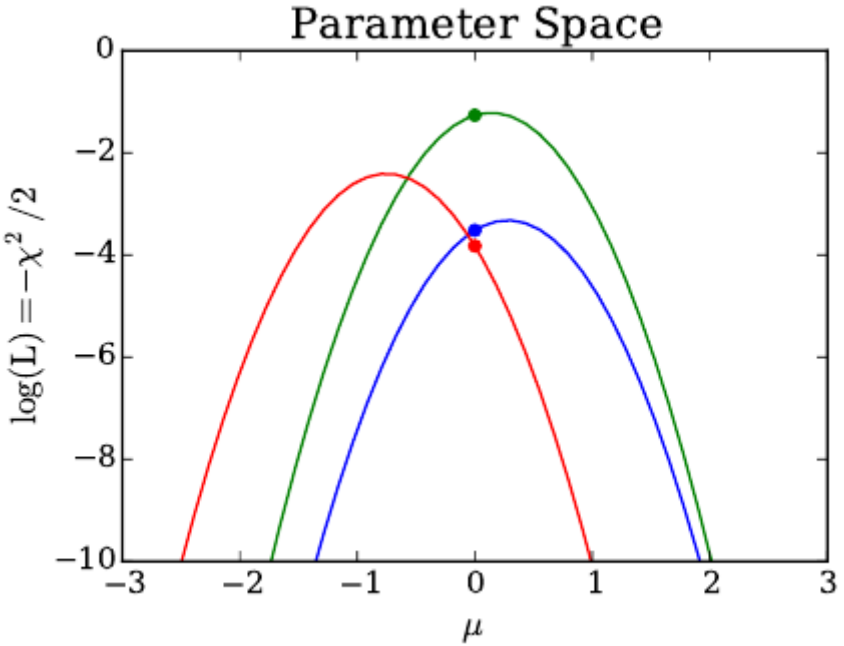
PPC: posterior predictive checks



Q-Q plot:
cumulative model
vs cumulative data



Best fit parameters



If many data are created under $\hat{\mu_D}$
logL interval -1/2 below best fit (Wilks' theorem)
contains true value 68% of realisations

$P(\hat{\mu}|\hat{\mu_D})$ Confidence interval

- If away from boundary
- If model is linear
- If ndata → high (symmetric, single gauss)
- If θ is true parameter

  → then

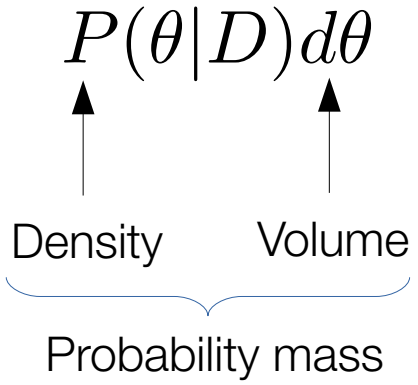
What was the question again? $P(\mu|D)$
Are conditions fulfilled?
What do unequal "errors" mean? 2d?

Inference desiderata

Probable parameter ranges of spin?

$P(\theta|D)d\theta$ Probability density

In infinitely small region: zero probability



Find regions with high probability mass

$P(D|\theta)$

- Frequentist statistics
  - How good does a procedure work
  - Properties of estimators
  - False decision rates
- Bayesian statistics
  - Consistent framework, assumptions spelled out
  - What is the probability distribution of the true parameters?
  - How probable are these physical models relative to each other?

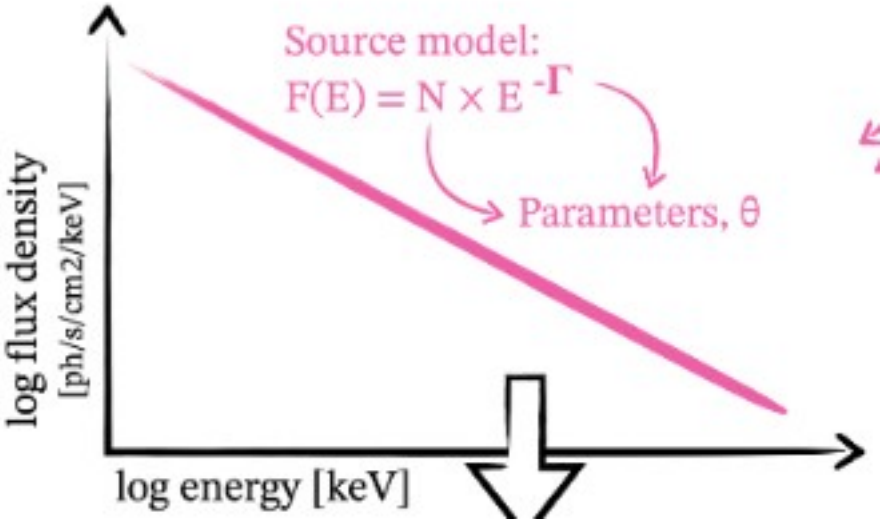
"Bayesian statistics plus"

- Do Bayesian statistics: credible intervals, posterior distributions, Bayesian model comparison
- For decisions and parameter estimation, quantify properties with Monte Carlo simulations

→ best of both worlds

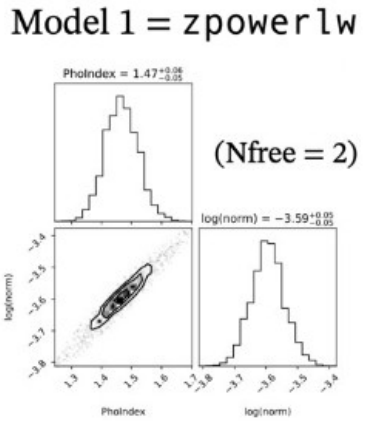
Workflow

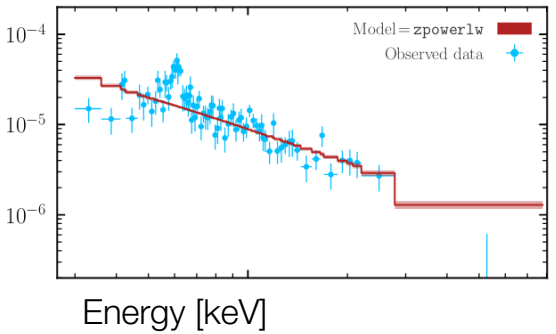
- Get data
- Assume model
- Assume parameter priors
- Produce posterior distributions
- Model checking
- Model comparison
- Vary assumptions, check robustness with simulated data
- Write paper
- Predict observations, get more data



N: log-uniform prior
Equal probability from 10^{-7} to 10^{3}

Gamma: informative prior from literature
Gaussian distribution 1.95+-0.15

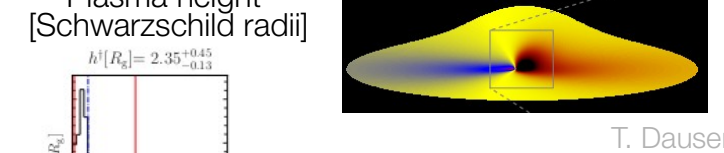




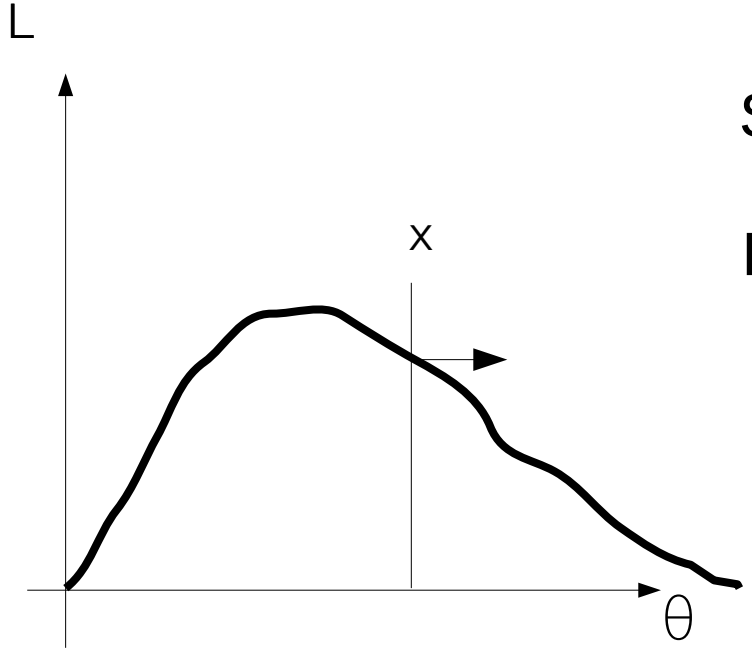
Parameter space exploration

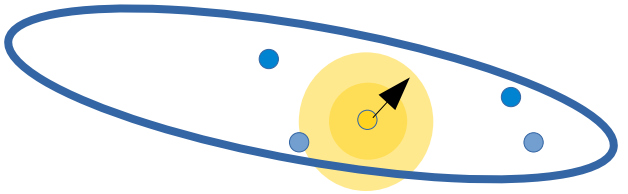
5-20 parameters
1-1000ms evaluation
non-linear → degeneracies
non-differentiable
multi-modal

- Local optimization
  - LM, simplex, … (many)
- Local sampling: MCMC
  - gradient-free
- Global optimization
  - Genetic algorithms (DE)
- Global sampling
  - Nested sampling



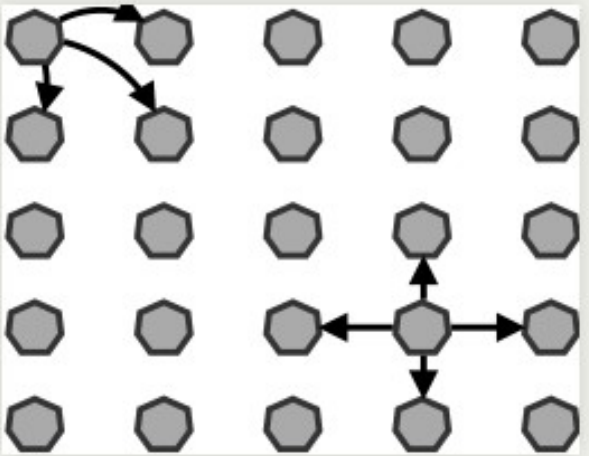
Markov Chain Monte Carlo





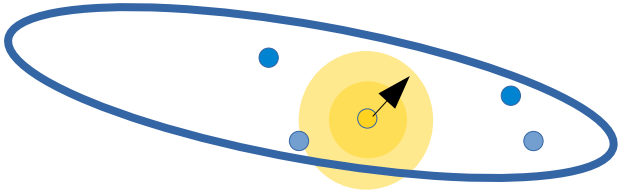
```
Starting point θ

Loop forever:
   θ' = Normal(θ, sigma_p)
   if P(θ'|D)/P(θ|D) > U():
      θ = θ'
   add θ to chain
```

- Missing ingredient: transition kernel
- tune to the problems
- Fraction of visits ~ converges to ~ probability of hypothesis
- Where does chain spend 90% of its visits



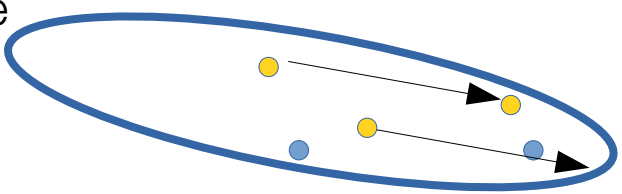
MCMC transition kernels

- Metropolis Random Walk
  - Adv: simple
  - Disadv: poor mixing



- Affine-invariant ensemble
  - Adv: auto-tuning for gaussian L
  - Disadv: poor mixing in bananas, collapses in high-d (Huijser+15)

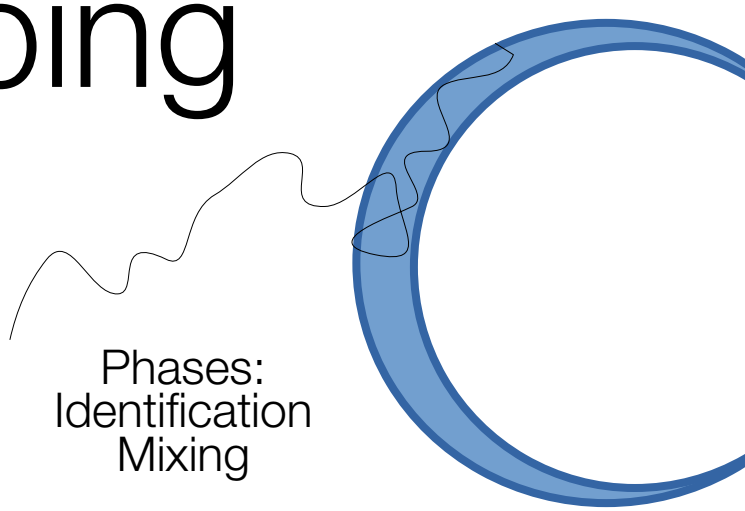
Goodman & Weare (2010)
emcee



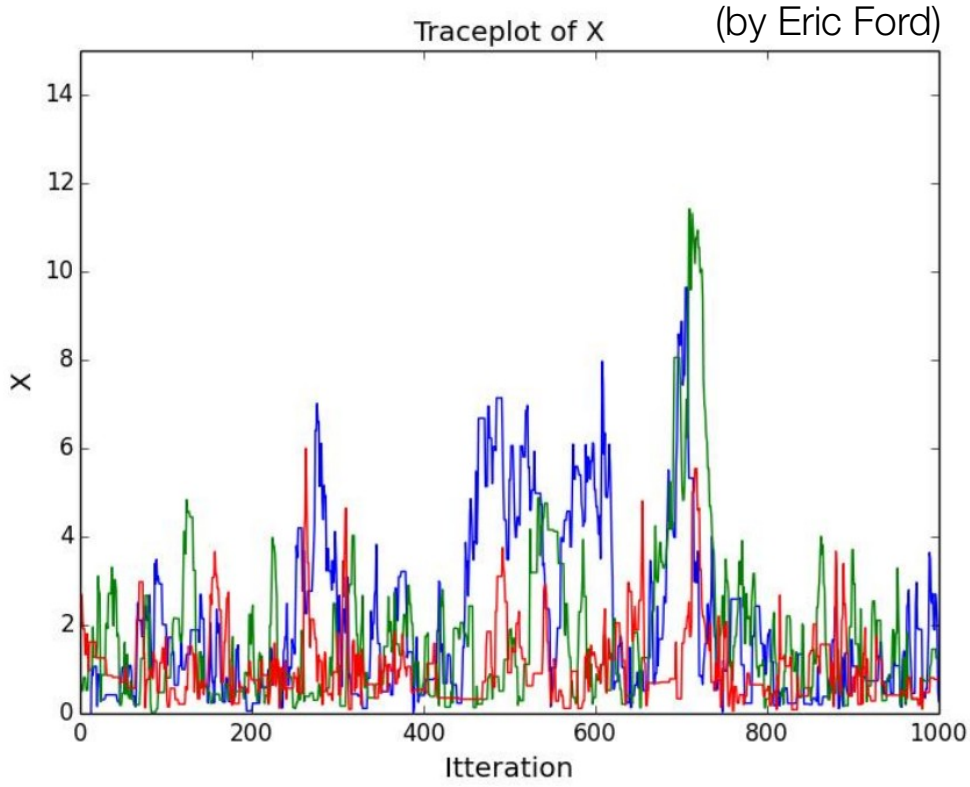
- HMC (Hamiltonian Monte Carlo)
  - Adv: tunes itself to surface
  - Disadv: need gradients of models

MCMC stopping

- MCMC theory: n→inf
- Trace plots
- Autocorrelation length
- Convergence tests
  - Detect if unreliable
  - Gelman-Rubin diagnostic
  - (many more)



(by Eric Ford)



For Goodman-Weare or Ensemble Slice Samplers, see:

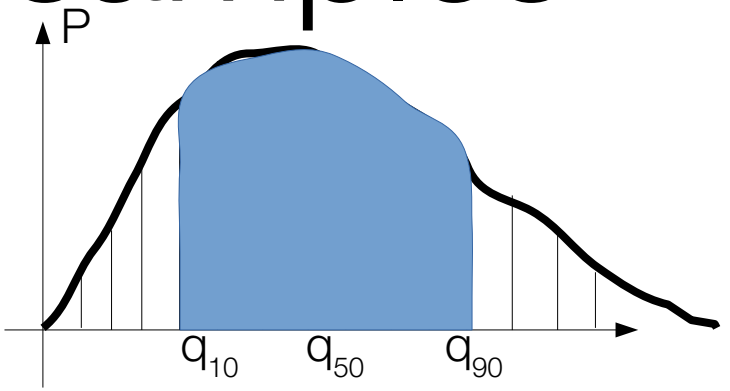
https://johannesbuchner.github.io/autoemcee/mcmc-ensemble-convergence.html

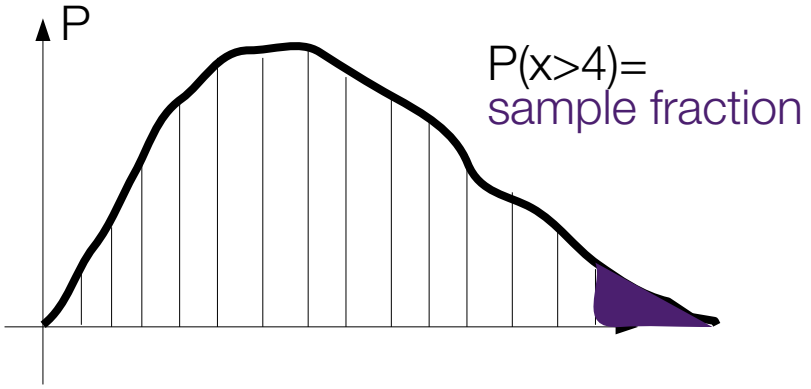
Using posterior samples

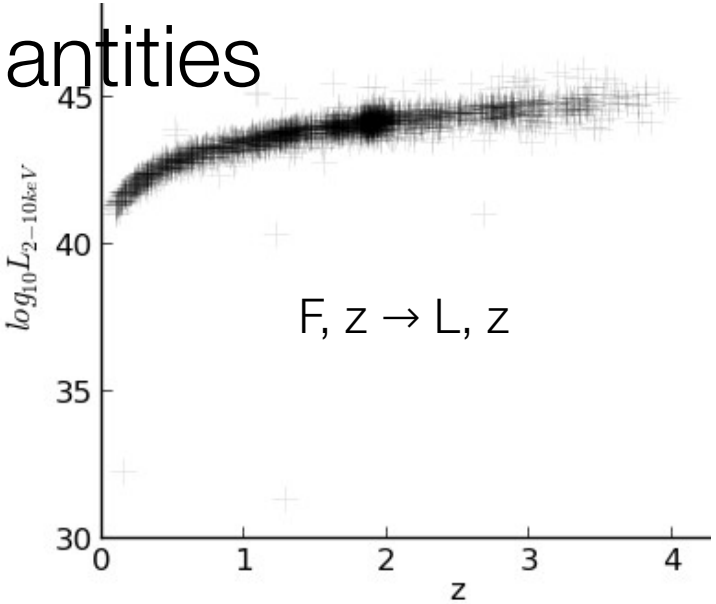
- Posterior samples

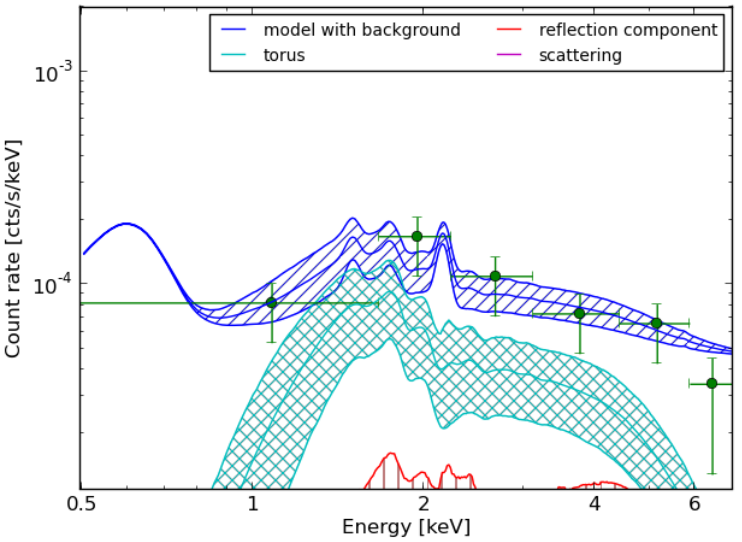
  θ_1 θ_2 θ_3 ....

- Find regions with high prob
- Compute prob. of regions
- Posterior predictions
- Derived quantities





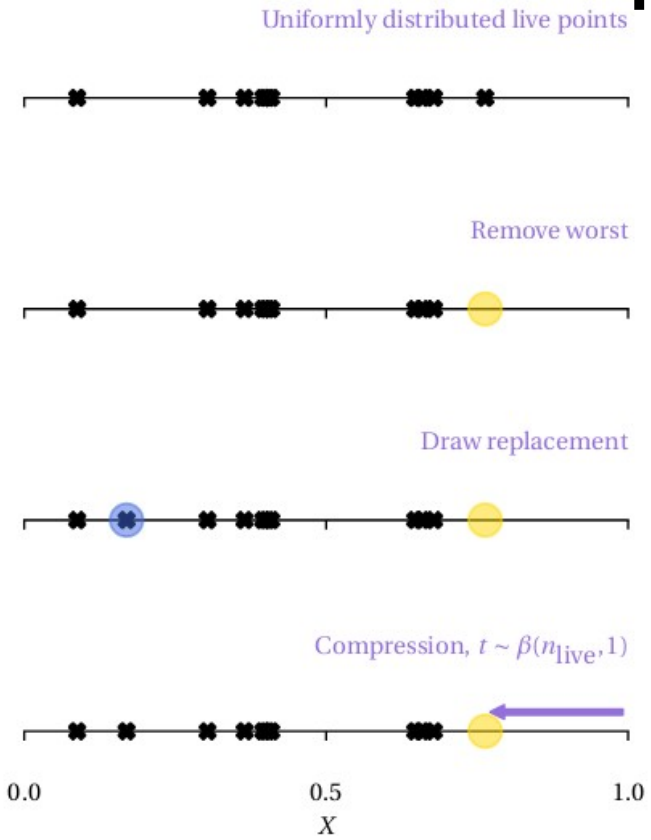




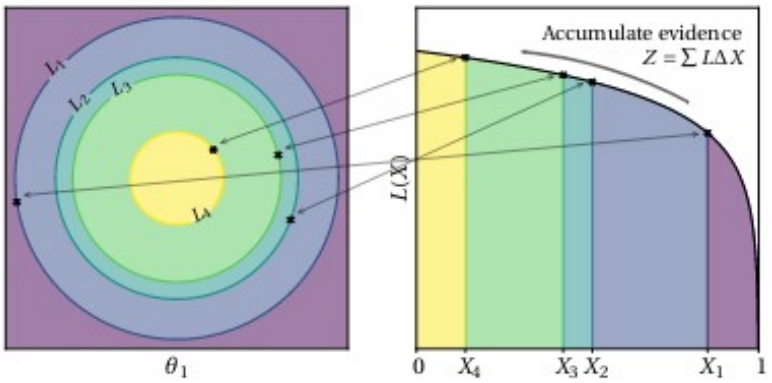
Nested sampling

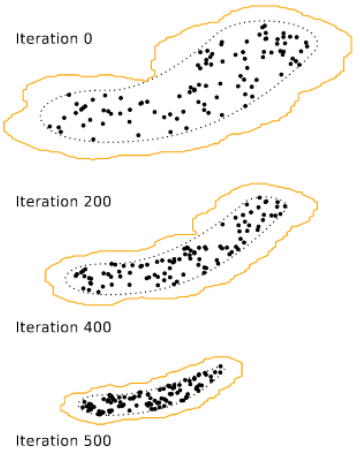
Monte Carlo Integration algorithm for Bayesian inference

- Robust, global sampling algorithm
- Simulates posterior samples
- Estimates marginal likelihood Z



Exponential compression, recursive





$V \to 0$
remainder negligible

Theory & convergence proofs:

Skilling ‘04,06,09
Evans ‘07
Chopin&Robert ‘07,10
Walter ‘14

Review of methods for statisticians & physicists:
https://arxiv.org/abs/2101.09675

Review of concept, applications & softwares:
https://arxiv.org/abs/2205.15570

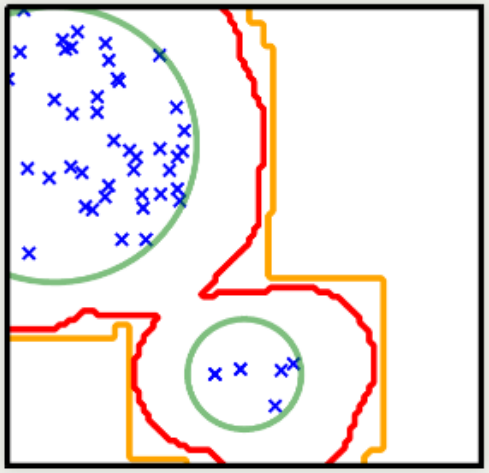
Nested Sampling for physical scientists

Greg Ashton,[1,2] Noam Bernstein,[3] Johannes Buchner,[4] Xi Chen,[5] Gábor Csányi,[6] Farhan Feroz,[7] Andrew Fowlie,[8] Matthew Griffiths,[9] Michael Habeck,[10] Will Handley,[11,12] Edward Higson,[13] Michael Hobson,[12] Anthony Lasenby,[11,12] David Parkinson,[14] Livia B. Pártay,[15] Matthew Pitkin,[16] Doris Schneider,[17] Leah South,[18] Joshua S. Speagle (沈佳士),[19,20,21] John Veitch,[22] Philipp Wacker,[17] David J Wales,[23] David Yallup[11,12]

We review Skilling's nested sampling (NS) algorithm for Bayesian inference and more broadly multi-dimensional integration.

Missing ingredients

- MCMC: optimized transition kernel
- NS: likelihood-restricted prior sampling



There are general solutions:

- Step samplers: slice sampling (Jasa&Xiang12), CHMC/Galilean (Betancourt11, Skilling12), …
- Region samplers: MultiNest (Feroz+09), MLFriends (Buchner14,19)

Review: “Nested Sampling Methods” arxiv:2101.09675

Animation:

https://johannesbuchner.github.io/mcmc-demo/app.html#RadFriends-NS,standard(via chi-feng.github.io)

Bayesian X-ray Astronomy

Buchner+14
https://johannesbuchner.github.io/BXA/

parallelisation, resuming

sophisticated, robust

inference engine

based on nested sampling

community models

data formats

fully-fledged
fitting environment

BXA
(724 citations)

MultiNest
UltraNest

sherpa
pyxspec
(threeml)
(spex)

+ background models
+ some visualisation tools

What can BXA do?

- Given any model and data supported by the fitting environment
- and priors
- can produce
  - posterior probability plots

    useful for:
    - Parameter constraints
    - Sample distributions
  - Evidence Z

    useful for model comparison

(no “good starting point” needed)
no minimal counts needed
no rebinning needed

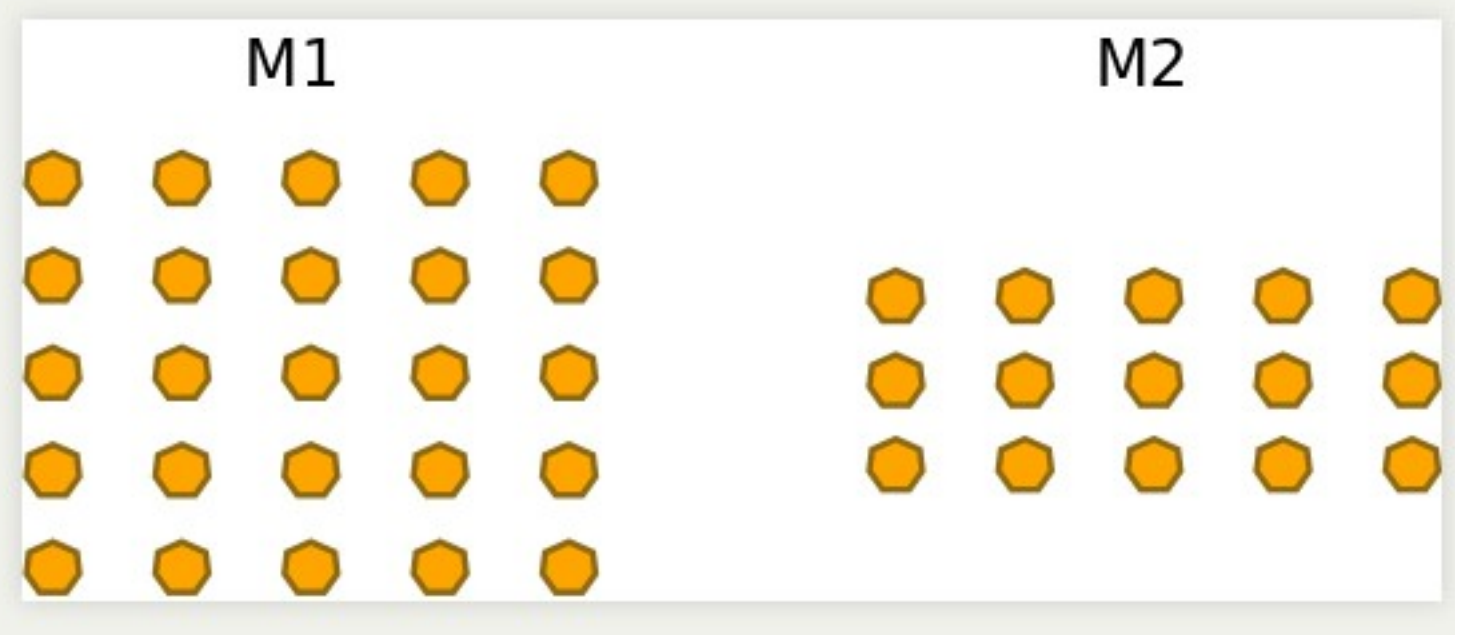
What to do with Z

- Z1, Z2

$$\underbrace{\frac{p(M1|D)}{p(M2|D)}}_{\text{Posterior odds ratio}} = \underbrace{\frac{Z1}{Z2}}_{\text{Bayes factor}} \underbrace{\frac{p(M1)}{p(M2)}}_{\text{Prior odds ratio}}$$

Punishing prediction diversity

(not number of parameters)

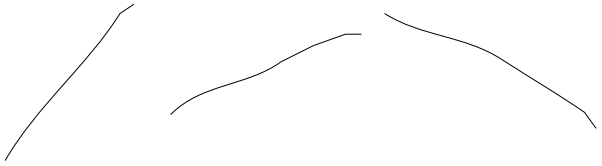


Flexible model

Inflexible model

Data

L high, V tiny



L medium, V medium

What to do with Z

- Z1, Z2

$$\frac{p(M1|D)}{p(M2|D)} = \frac{Z1 \cdot p(M1)}{Z2 \cdot p(M2)}$$

$$\frac{p(M_1|D)}{\sum p(M_i|D)} = \frac{Z_1 \cdot p(M_i)}{\sum_i Z_i \cdot p(M_i)}$$

- model priors: leave to reader or motivated by theory
- Discard highly improbable model or marginalise
- Does $\frac{p(M1|D)}{p(M2|D)} = 3/1$ mean M2 is correct in a quarter of the cases?

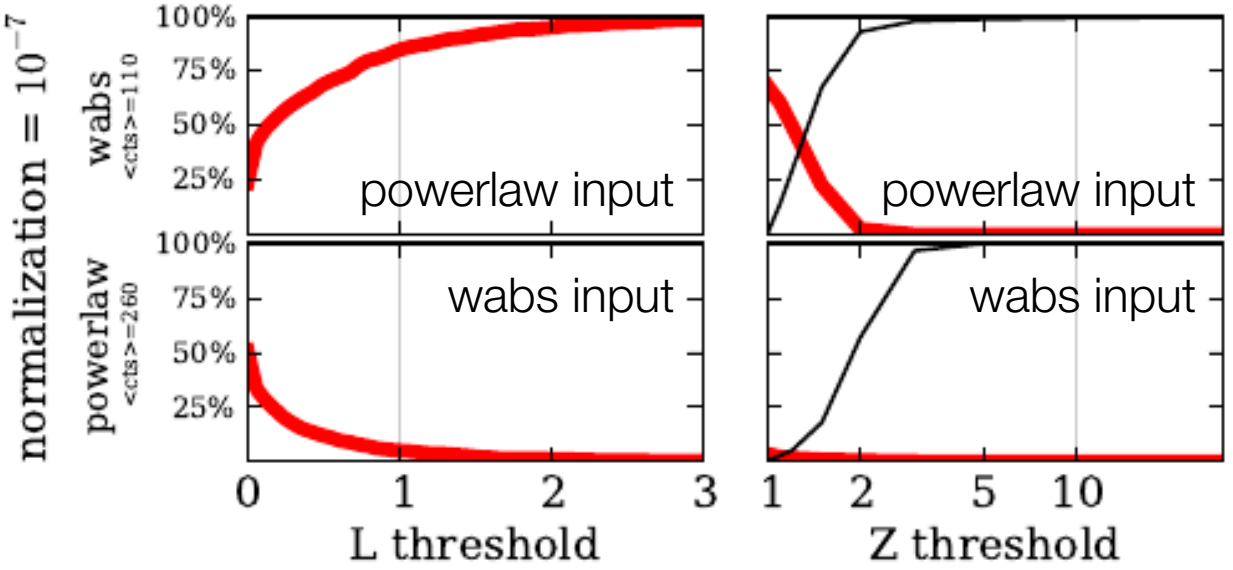
Frequentist properties of Bayesian methods

- Make decisions
  - Is parameter greater than X?
  - Should I continue working with model A or model B?
- Monte Carlo simulation (Parametric bootstrap)
  - allow arbitrary complexity to test

assume model A, generate data

→ how often would model B erroneously be selected?
→ what threshold is safe ($\alpha<0.01$)?

Calibrating model decisions



Buchner+14
Baronchelli+19

False negatives
Non-decisions

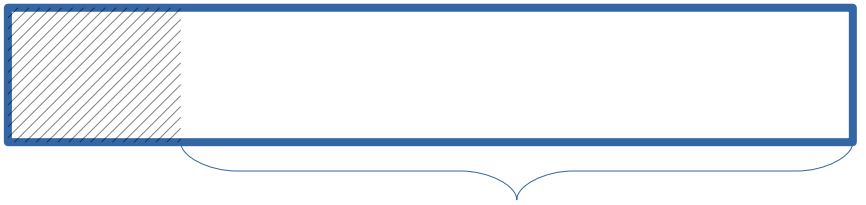
Advantages:
- Get rid of parameter prior dependencies
- Have frequentist properties of Bayesian method
- Completely Bayesian treatment + decisions

Disadvantages:
- Can be computationally expensive

Model comparison

Buchner+14

- Empirical models
  - Information content
  - Prediction quality
- Component presence
  - Regions of practical equivalence
- Physical effects
  - Bayesian model comparison
  - Priors often well-justified



https://arxiv.org/abs/1506.02273
Betancourt (2015)

Information criteria

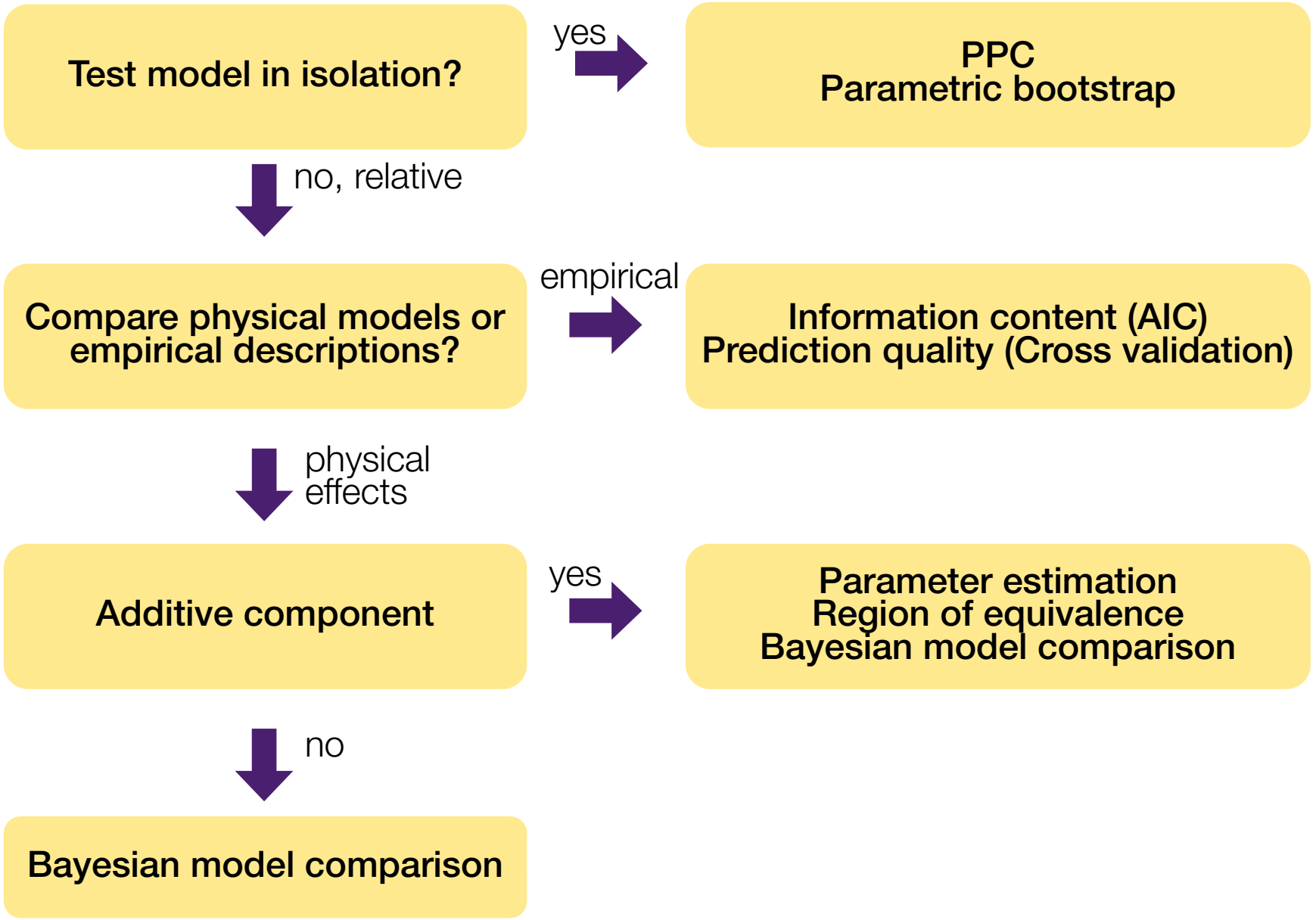
- Akaike information criterion Akaike (1973)
- Is more complex worth storing?

$$AIC = 2 * d - 2 * L_{max}$$
$$AIC = 2 * d + CStat$$

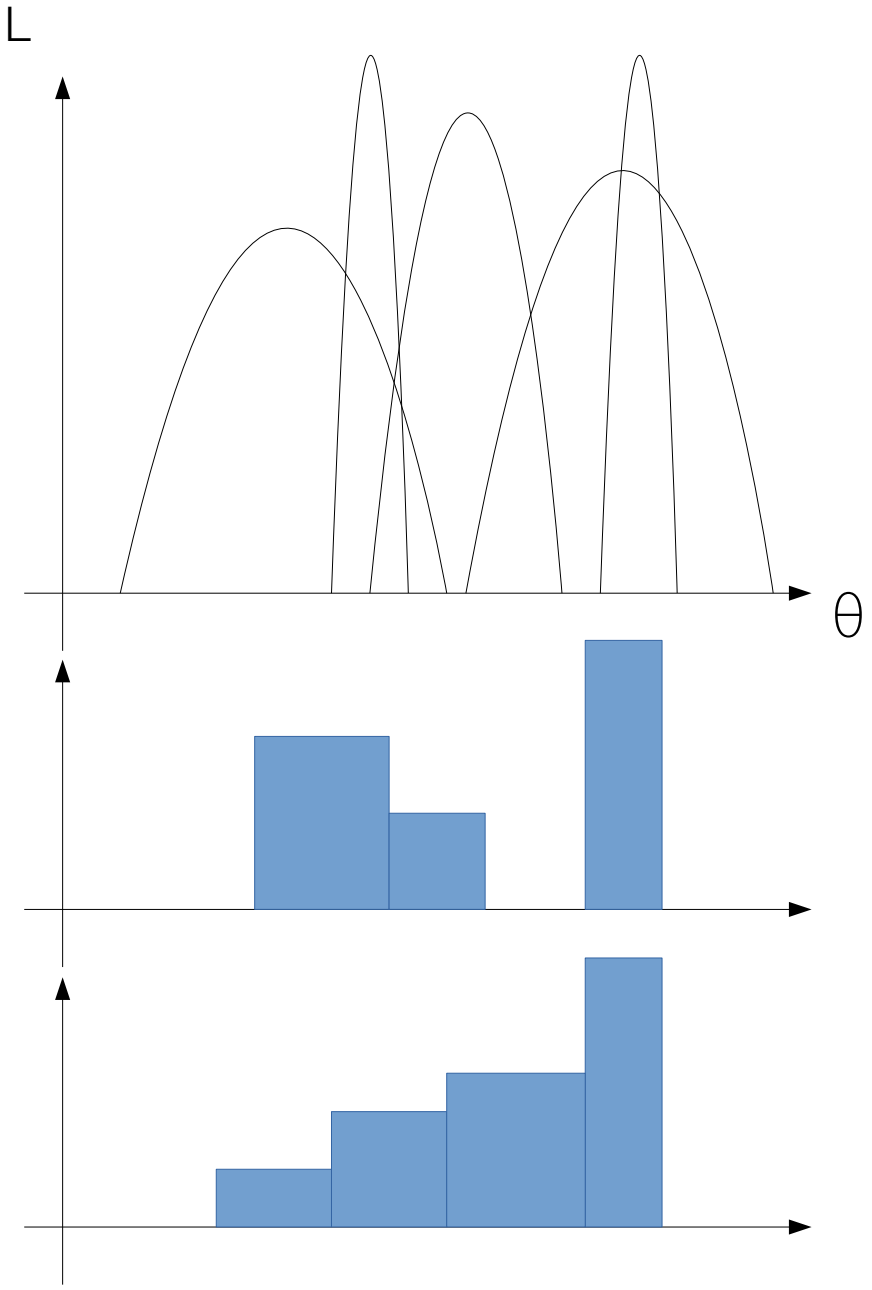
Advantages:
- rooted in information theory
- independent of prior

Disadvantages:
- No uncertainties, thresholds unclear
- …

Model comparison



Best fit distributions



Convolution of

True parameter distribution +
Measurement error & analysis method

Confidence intervals

Histogram of best fits

Meaning?
Upper limits?

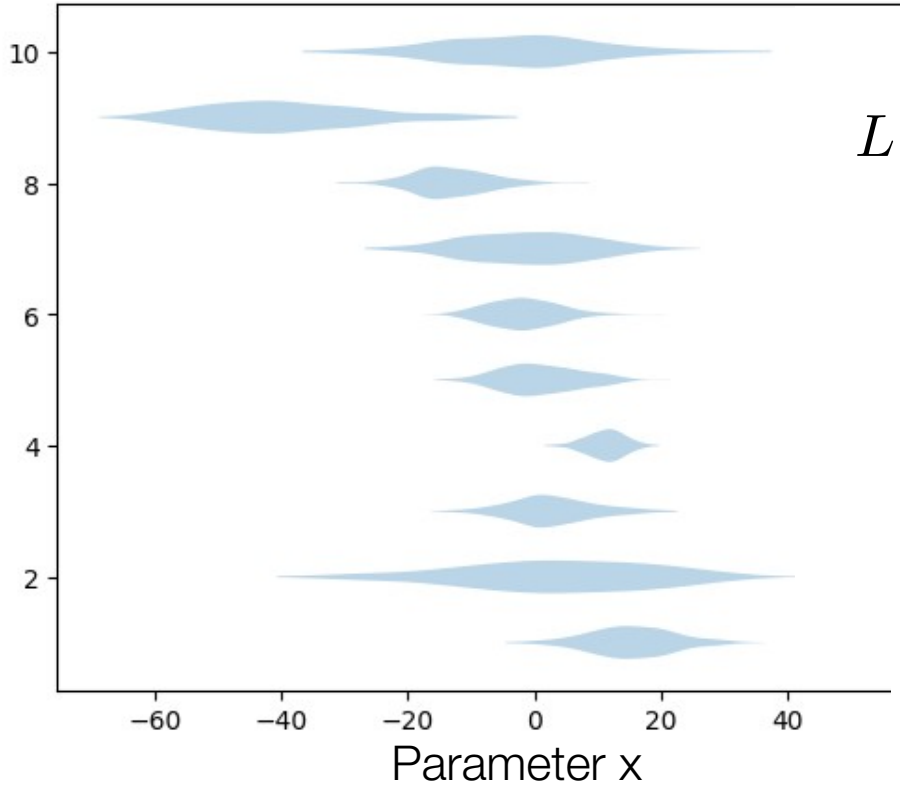
Cumulative distribution

Clean solution:
Model the distribution (HBM)

e.g. Buchner+17a
Baronchelli+19
→ PosteriorStacker

Sample distributions

Hierarchical Bayesian models

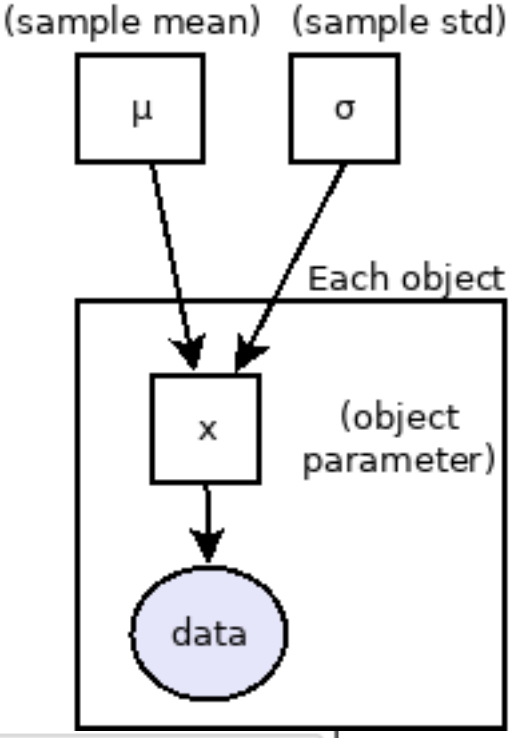


Parametric sample distribution

$$L = \prod_j \frac{1}{N} \sum_i P(\theta_{ij})$$

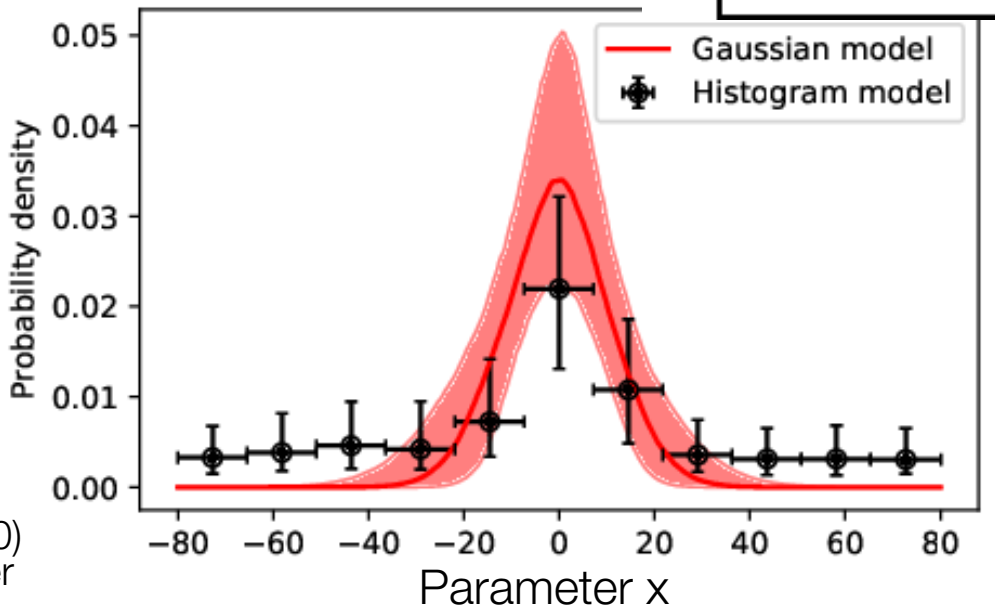
One of the posterior samples

All of the objects



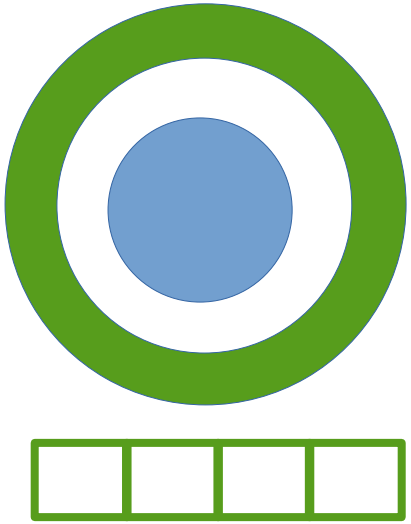
Objects with large uncertainties do not wash out the signal!

Self-consistent, Bayesian analysis



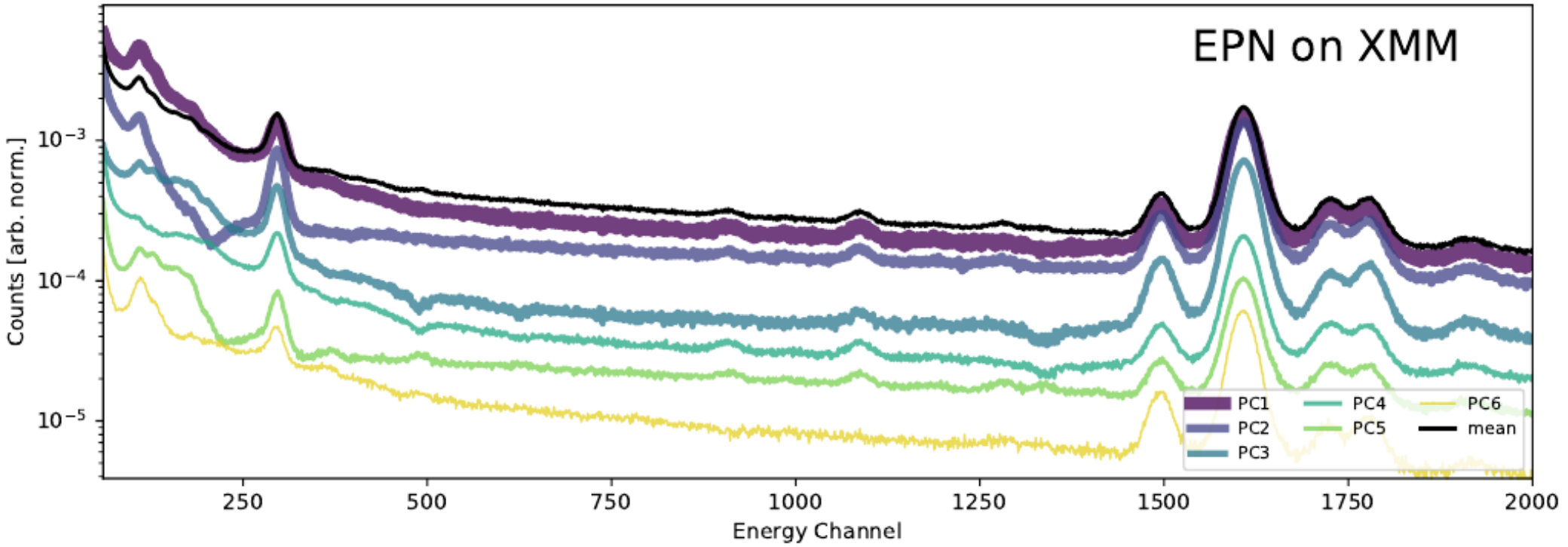
PosteriorStacker (Baronchelli, Nandra & Buchner, 2020)
https://github.com/JohannesBuchner/PosteriorStacker

Empirical background models



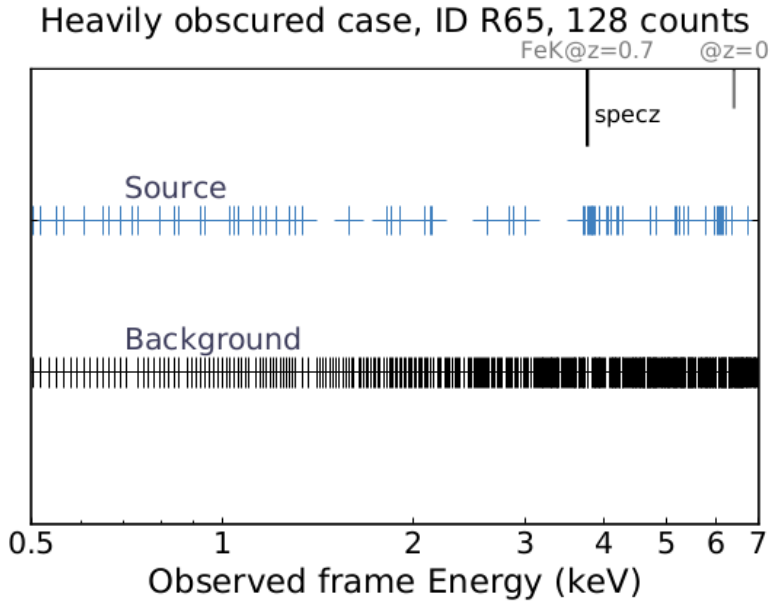
Automated shape finding
Simmonds, Buchner et al. (2017)

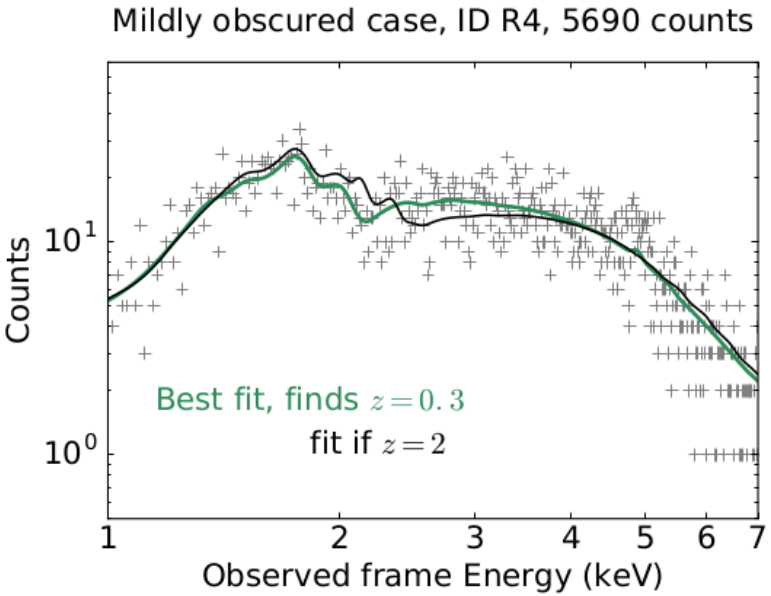
XMM/PN,MOS, Chandra/ACIS, NuSTAR,
Suzaku, RXTE, Swift/XRT

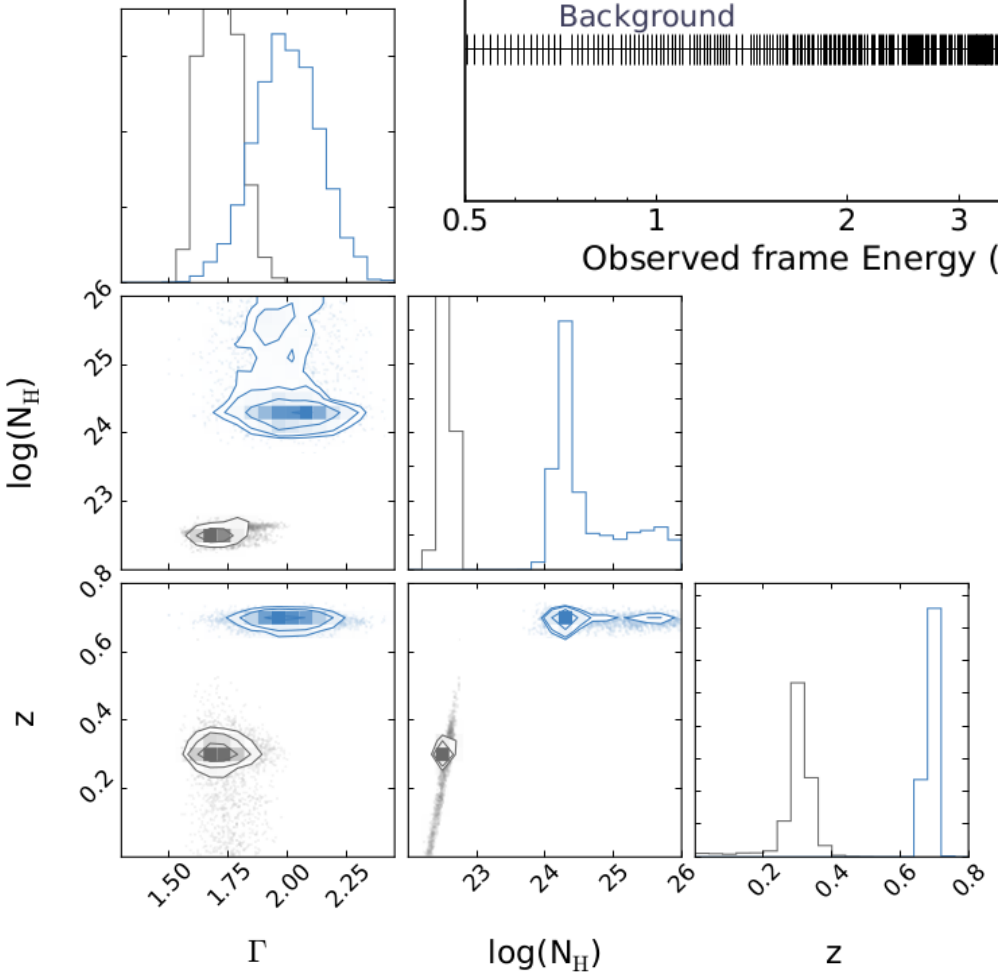


XZ: X-ray redshifts

AGN in the Chandra deep field south







Simmonds, JB et al. 2018

Classical: Peca et al. 2021
requires tuning thresholds for each instrument

Methods for X-ray astronomy

- Forward-folding instrument response
  - Poisson statistics
  - Nuisance background modeling
  - Bayesian framework for inferring parameters and models
  - Monte Carlo simulations to verify robustness
- Practical algorithms

  Nested sampling & Goodman-Weare MCMC
- Short overview of techniques for
  - Model checking
  - Model comparison

Stay tuned for book chapter (or request a copy from me)