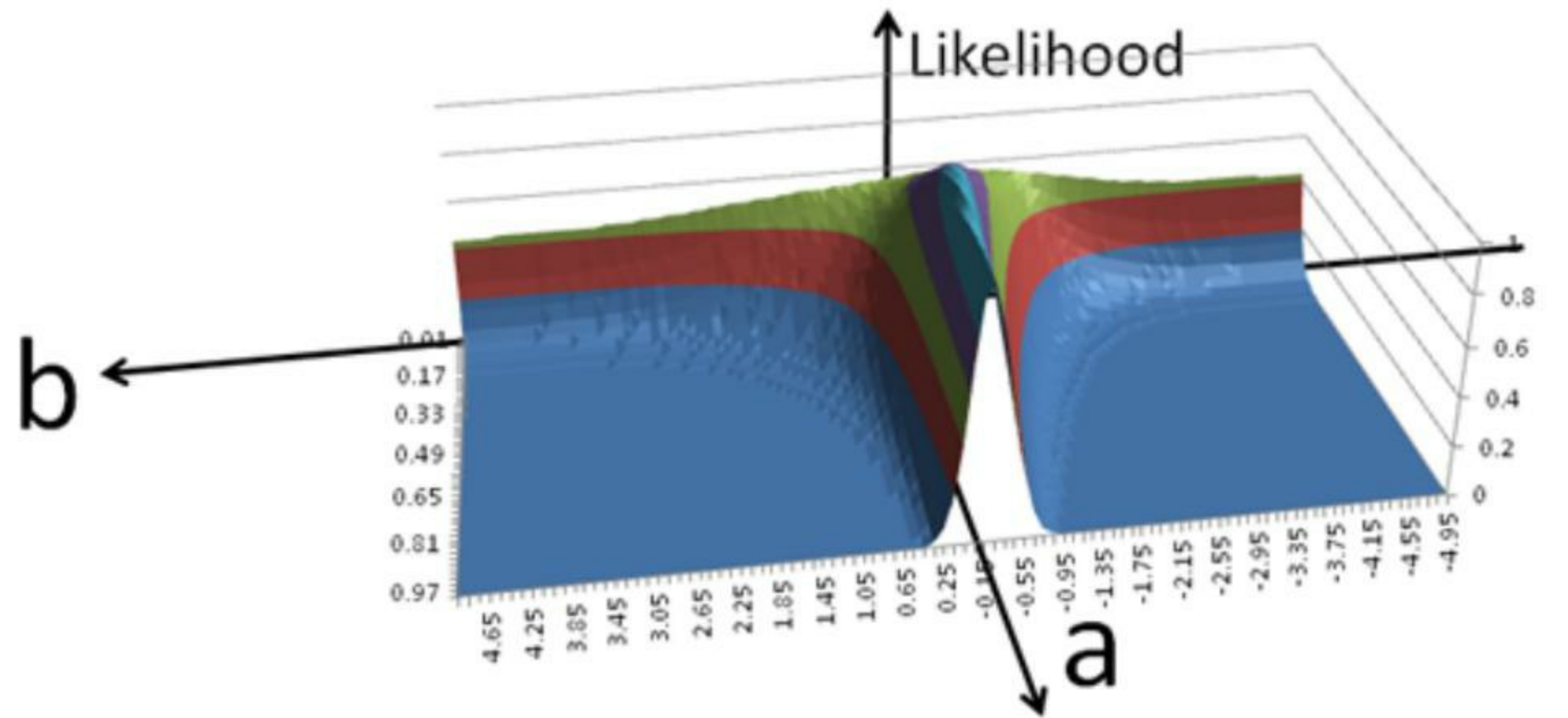


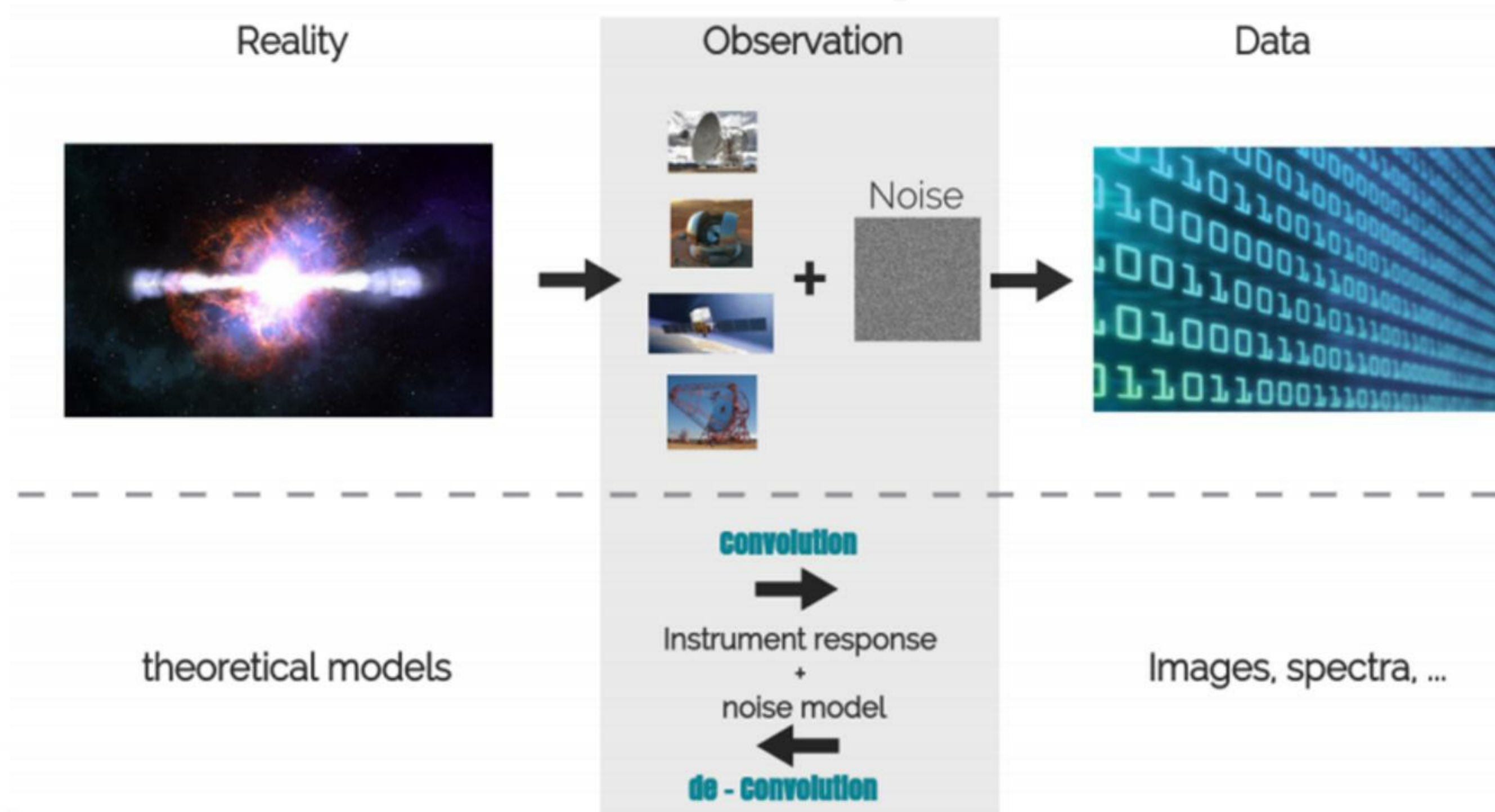
# (MAXIMUM) LIKELIHOOD ANALYSIS

Introduction



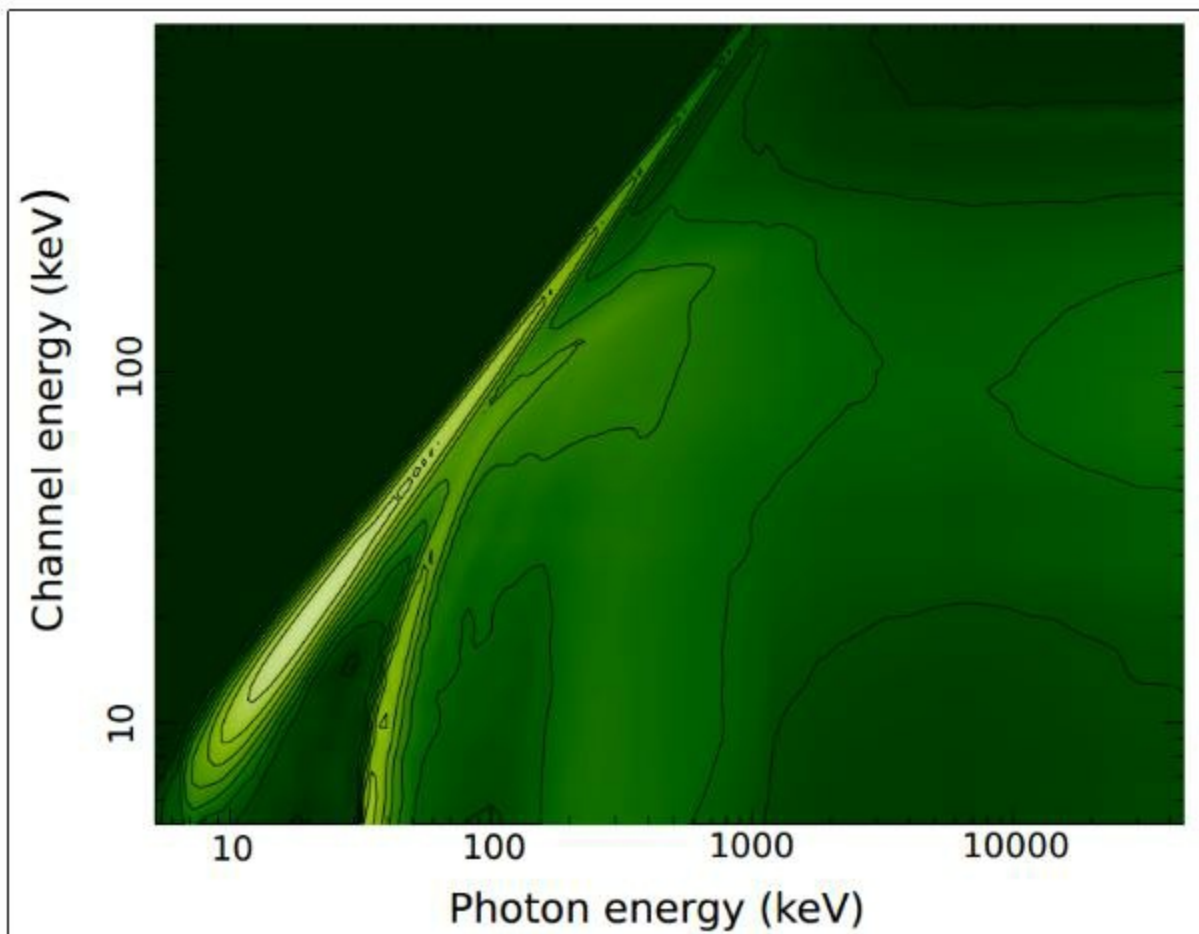
GIACOMO VIANELLO  
(STANFORD UNIVERSITY)

# The observation process

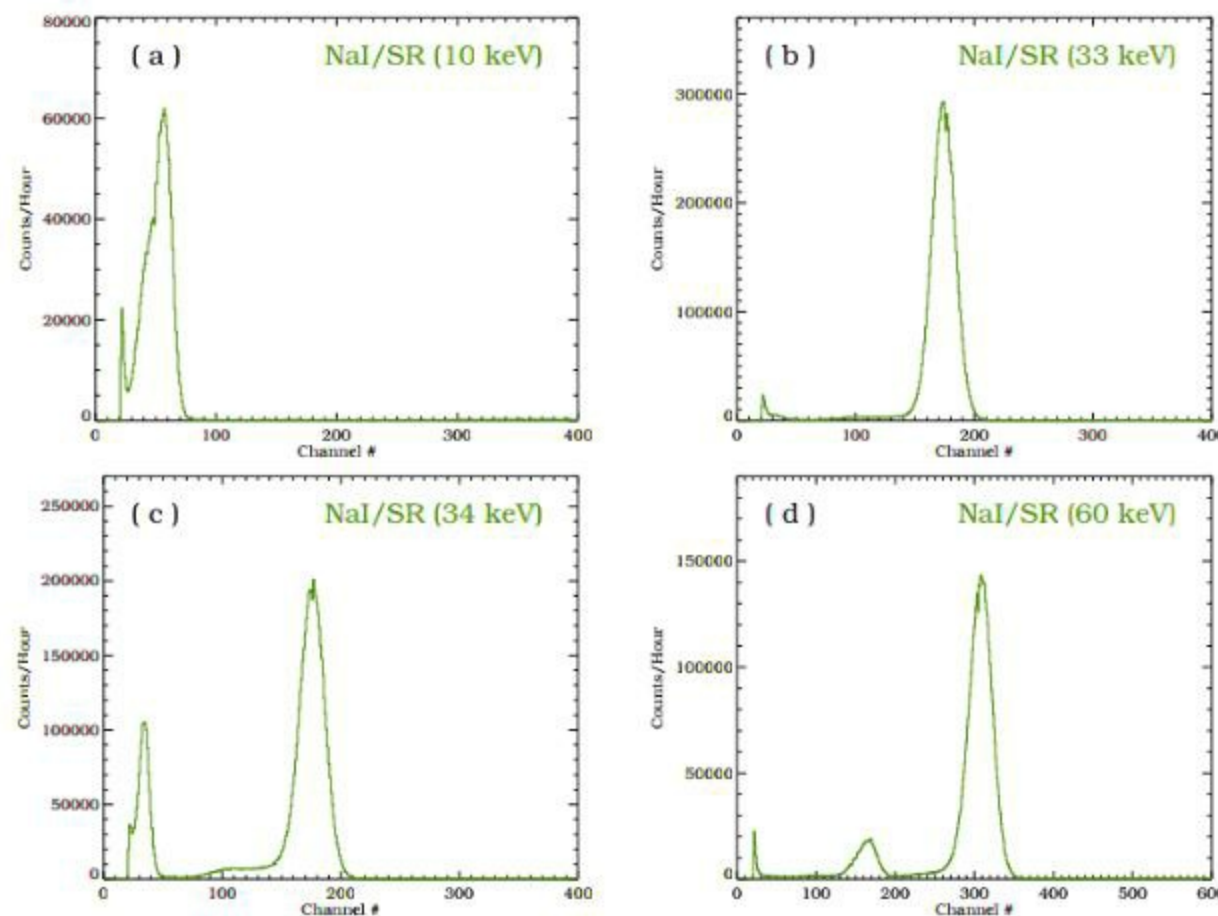


(Vianello et al. 2015)





Energy response of a Fermi/GBM NaI detector



# The instrument

- react to the incoming signal  $S$  producing an analog output  $D$  related to the input in some known way:

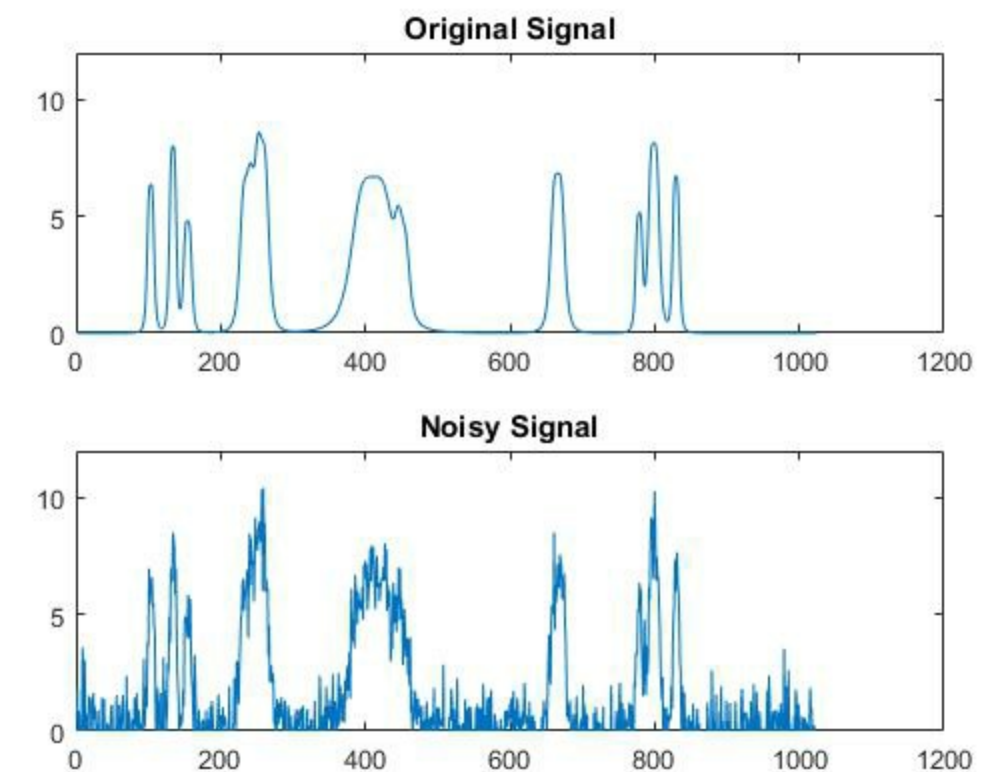
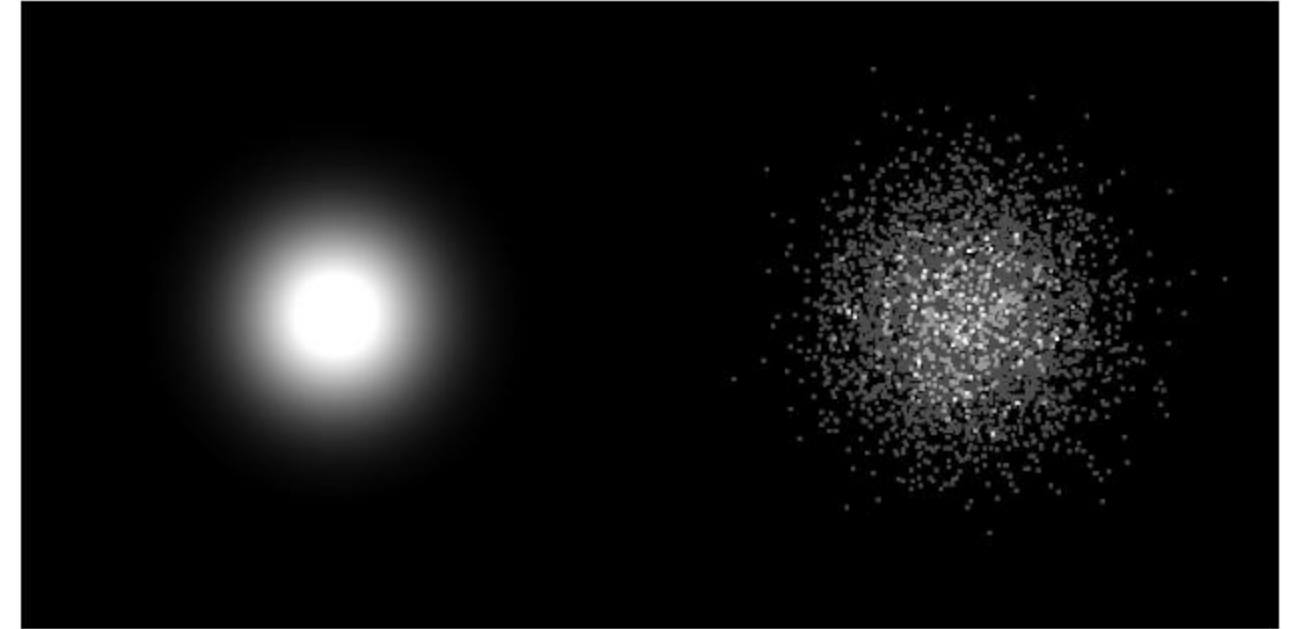
$$D(\vec{p}', E', \dots) = S(\vec{p}, E, \dots) \circ R(\vec{p}, E, \vec{p}', E', \dots)$$

- The dispersion in energy is called energy dispersion, the dispersion in space is called Point Spread Function (PSF)
- For an ideal instrument  $E=E'$  (no energy dispersion),  $p = p'$  (infinite spatial resolution) so that  $D(p, E) = S(p, E) \times R(p, E)$
- Without knowing the response, the output of an instrument is of little interest
- You know  $R$  up to a certain level (systematic uncertainties)



# Noise (randomness)

- Random processes producing noise can be in the source and/or in the detector
- destroys information: in general makes the equation in the previous slide not invertible (solution becomes non-unique)
- different types
  - Gaussian noise
  - Poisson noise
  - ....
- Different types of noises can contribute in different part of the observation process
- Noise model(s): the type of noise assumed to be at play in the analysis at hand



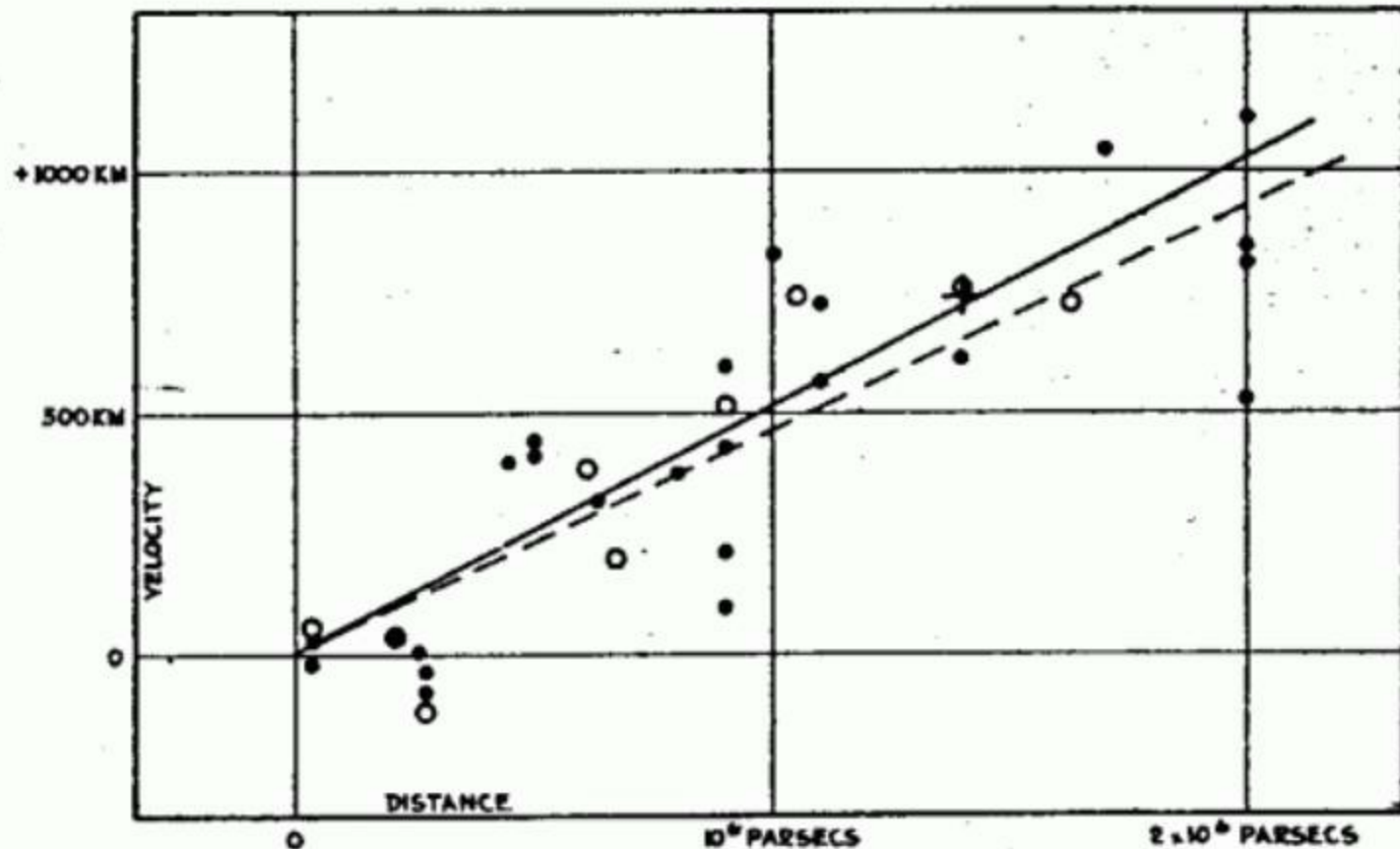


FIGURE 1

Hubble's law (from his paper, 1929)

# Model

- "a mathematical representation of a theory"
  - "The speed of a galaxy is directly proportional to their distance from the Milky Way"
- "theory" in a broad sense:
  - complete theory: General Relativity
  - phenomenological theory: a linear relationship
- often we start with a phenomenological theory, only later we are able to include discoveries in a broader physical theory
- sometimes we have a prediction from a theory which we want to verify
- A model has parameters which are adjusted to give the best fit on the dataset (more on this later)

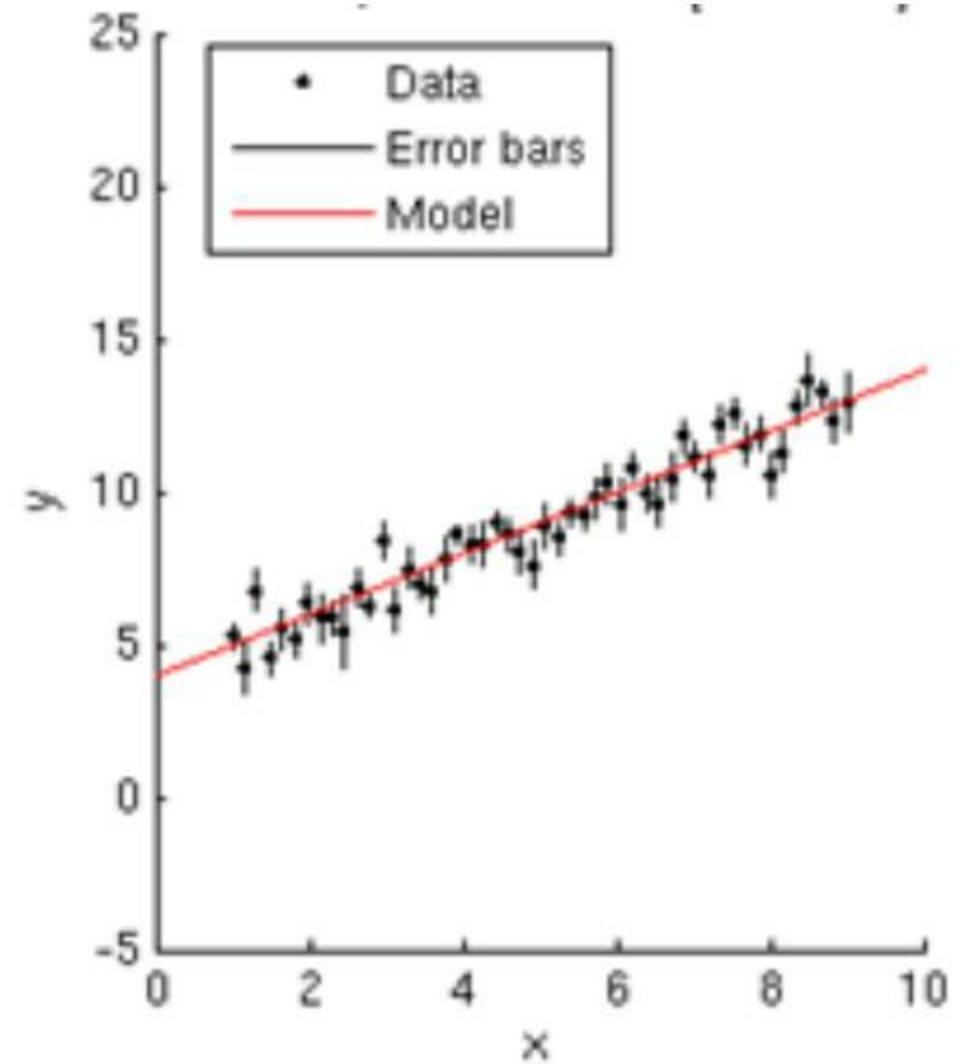


you already used Maximum Likelihood!

# LINEAR REGRESSION WITH ERRORS

Find the line which passes through the data and minimizes the objective function:

$$\chi^2 = \sum_{i=0}^N \frac{(y_i - ax_i - b)^2}{\sigma_i^2}$$



# Justification (1/2)

- N points  $(x_i, y_i)$
- Noise model is Gaussian, each point is characterized by its error bar  $\sigma_i$
- The model is assumed to be a line  $y = ax + b$
- the prediction at each  $x_i$  is  $M(a,b) = ax_i + b$
- Maximum likelihood idea: "what are the values for the parameters so that the model is most likely to have generated our data?"
- The probability that the  $y_i$  point has been generated by the model M with some values for its parameters  $(a,b)$  is:

$$P(y_i|a, b) = \frac{1}{\sqrt{2\pi} \sigma_i} e^{-\frac{(y_i - ax_i - b)^2}{2\sigma_i^2}}$$

- For independent events the probability  $P(A \text{ and } B) = P(A)P(B|A) = P(A)P(B)$ . So if the points are all independent, the probability of observing the data given the model (likelihood) is:

$$L(y|a, b) = \prod_{i=0}^N P(y_i|a, b)$$

- The Maximum Likelihood Estimate (MLE) for  $a,b$  is the pair  $(a,b)_{MLE}$  for which the likelihood function is at its maximum:

$$(a, b)_{MLE} = \operatorname{argmax}_{a,b} L(y|a, b)$$



# Justification (2/2)

- Let's take the logarithm. The point  $(a,b)$  for which  $L(y|a,b)$  is maximum is the same for which  $\log(L(y|a,b))$  is maximum because  $\log(\cdot)$  is a monotonically-increasing function:

$$\log L(y|a, b) = \sum_{i=0}^N \log P(y_i|a, b)$$

- Let's plug in the Gaussian distribution and develop some algebra:

$$\begin{aligned} \log L(y|a, b) &= \sum_{i=0}^N \log \left( \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(y_i - ax_i - b)^2}{2\sigma_i^2}} \right) \\ &= \sum_{i=0}^N -\frac{(y_i - ax_i - b)^2}{2\sigma_i^2} - \log(\sqrt{2\pi}\sigma_i) \\ &= -\frac{1}{2}\chi^2(a, b) - \text{const} \end{aligned}$$

- of course maximizing this function is the same as minimizing  $\chi^2$ :

$$(a, b)_{MLE} = \underset{a,b}{\operatorname{argmax}} L(y|a, b) = \underset{a,b}{\operatorname{argmin}} \chi^2(a, b)$$

- (note that the  $1/2$  factor in front of chi square is NOT unsequential for confidence intervals. If you build a c.i. using  $\chi^2$  you need to need to double the delta you are looking for with respect to the likelihood case)



# Likelihood and MLE

- Let's generalize the definition of likelihood:
  - we have a model  $M$  with a set of parameters  $\vec{\Omega}$  :
  - we have a dataset  $D$ , and  $d_{[i]}$  is a single datum characterize by its probability distribution  $P_i$  which depends  $\vec{\Omega}$  or a subset of it:

$$L(D|\vec{\Omega}) = \prod_{i=0}^N P_i(d_i|\vec{\Omega})$$
$$\log L(D|\vec{\Omega}) = \sum_{i=0}^N \log \left( P_i(d_i|\vec{\Omega}) \right)$$

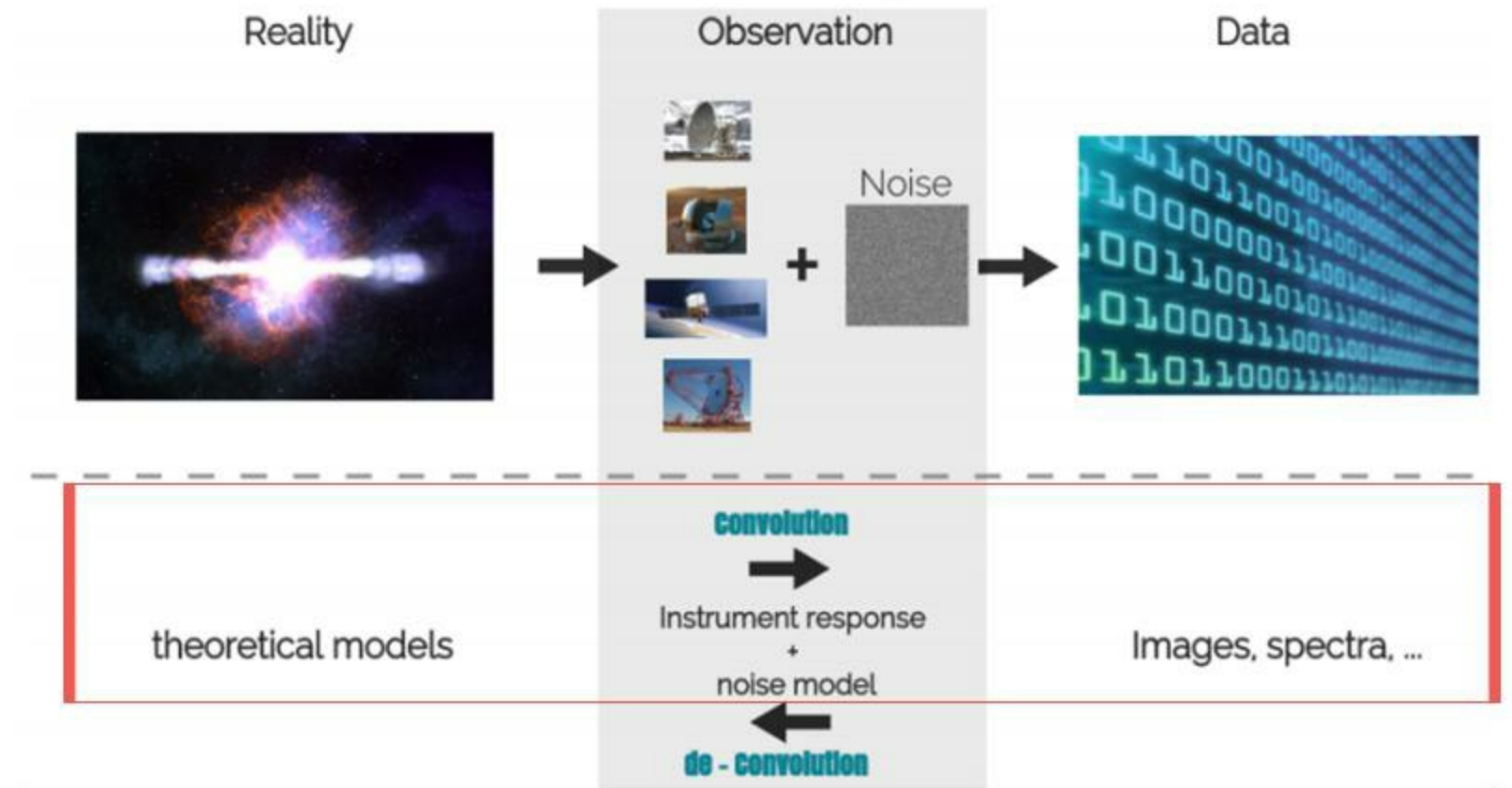
- The Maximum Likelihood Estimate for the parameters is:

$$\vec{\Omega}_{MLE} = \underset{\vec{\Omega}}{\operatorname{argmax}} L(D|\vec{\Omega})$$

- Some notes:
  - $0 < P_i < 1$ , so  $L > 0$ ,  $\log L < 0$
  - each point can have its own probability distribution

# MAXIMUM LIKELIHOOD IN ASTRONOMY

- To link our model to the data we need to pass through the instrument!
- Idea: bring our theoretical model  $M$  to the data space (forward-folding)
  - for a given set of parameters, compute how our instrument would see  $M$ , i.e., convolve  $M$  with the response of the instrument, then compute the likelihood
  - maximize the likelihood to find MLE





# Photon counting and Poisson noise

- Radiation from an astronomical source is a random process:
  - for instance, populations of particles (atoms, electrons...) emitting photons
- Let's take a main-sequence star (blackbody radiation with stable temperature)
  - The emission of a single photon is governed by quantum mechanics and happens at a random time
  - The emission of one photon does not influence the emission of another (independent processes)
  - The rate of emitted photons  $r$  is constant

• homogeneous Poisson process:  $P(n|\lambda) = \frac{\lambda^n}{n!} e^{-\lambda}$  with  $\lambda = r\Delta t$

- If the rate is not constant (but randomness and independence are still satisfied), then we have a non-homogeneous Poisson process with  $r = r(t)$
- Poisson process can be generalized to other quantities (energy) and also to more than 1 dimension




# Poisson likelihood:

- Fermi/LAT and the detectors of the GBM are photon-counting experiments
- Let  $m$  be our theoretical model and  $M = m \circ R$  (model prediction, more on this tomorrow)
- Let's divide our data in a certain number of bins  $i=1\dots N$  in  $n$  dimensions (space, time, energy...), and be  $n_i$  the number of photons recorded in the  $i$ -th bin
- divide the model prediction the same way so that we have the prediction of the model for each bin ( $M_i$ )
- For a given bin the Poisson distribution would be:

$$P_i(n_i|\vec{\Omega}) = \frac{\left(M_i(\vec{\Omega})\right)^{n_i}}{n_i!} e^{-M_i(\vec{\Omega})}$$

- The log-likelihood function becomes:

$$\begin{aligned}\log L(D|\vec{\Omega}) &= \sum_{i=0}^N n_i \log (M_i(\Omega)) - M_i(\vec{\Omega}) - \log (n_i!) \\ &= -N_{pred}(\vec{\Omega}) + \sum_{i=0}^N n_i \log (M_i(\Omega)) - \log (n_i!)\end{aligned}$$




**EXERCISE**

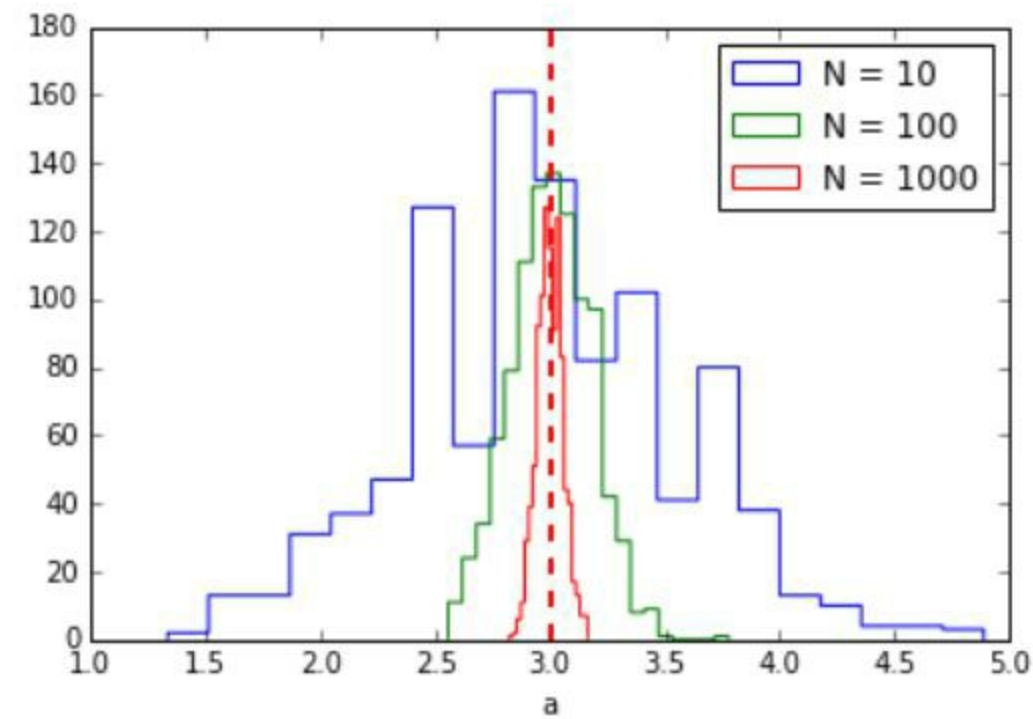
# INSTALL AND RUN

- open virtual machine
- if you don't have the code already, open a terminal then:
  - `> cd`
  - `> git clone https://github.com/giacomov/fermi_school_like.git`
  - `> cd fermi_school_like`
  - `> ipython notebook Summer_school_2016.ipynb`
- if you have the code already get the updates:
  - `> cd`
  - `> cd fermi_school_like`
  - `> git fetch --all`
  - `> git reset --hard origin/master`
  - `> ipython notebook Summer_school_2016.ipynb`



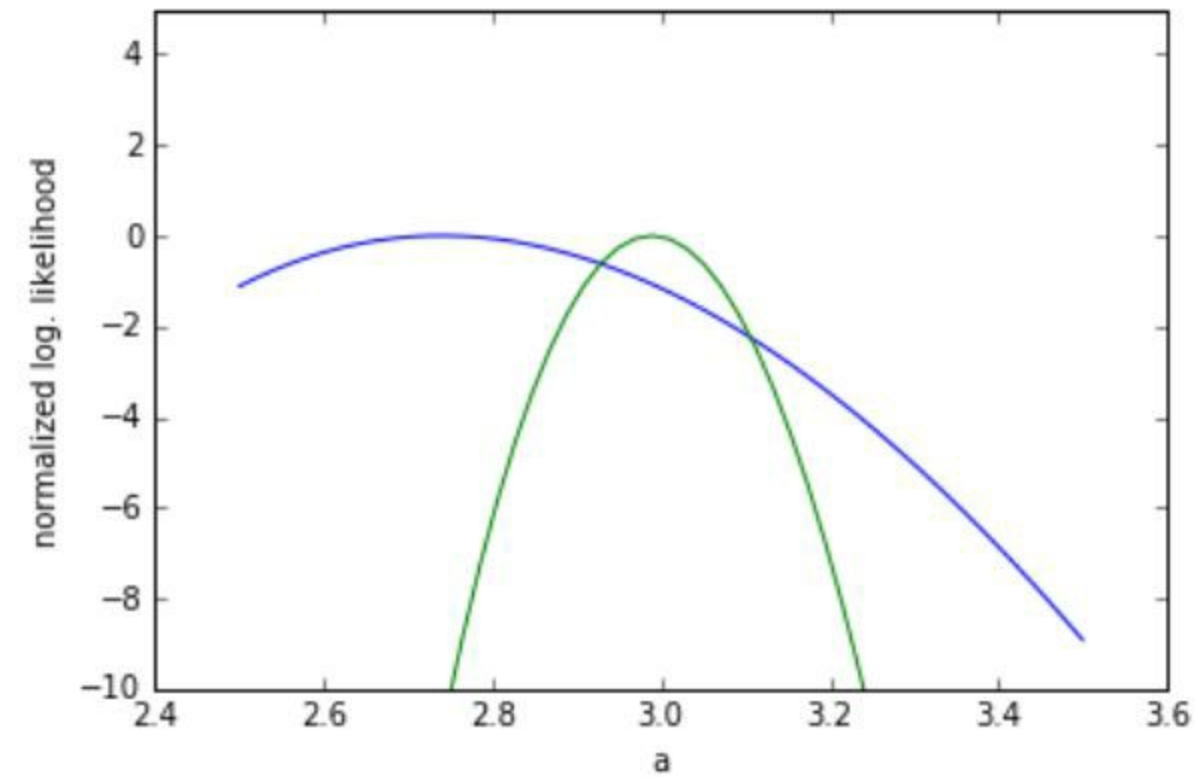
# Error computation on single measurement

The more data you have, the smaller the variance around the true value (width of the distribution is prop. to error):



This happens because the likelihood is more peaked around the true value.

# Error computation on single measurement



There is a theorem (Wilks' theorem) that under certain hypothesis guarantees that the true value is contained in the interval between  $\max(L) - 0.5$  and  $\max(L) + 0.5$  for 68% of the time. This is how an error bar in likelihood analysis is computed.

If you have more than one parameter, you have to repeat the fit for each point in the grid (profile likelihood)



# Model bias and variance

Modeling error

Intrinsic variance

$$Err(x) = \underbrace{\left(E[\hat{f}(x)] - f(x)\right)^2}_{\text{Bias}} + \underbrace{E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right]}_{\text{Variance}} + \sigma_e^2$$

Bias: how much the average prediction is far from the truth

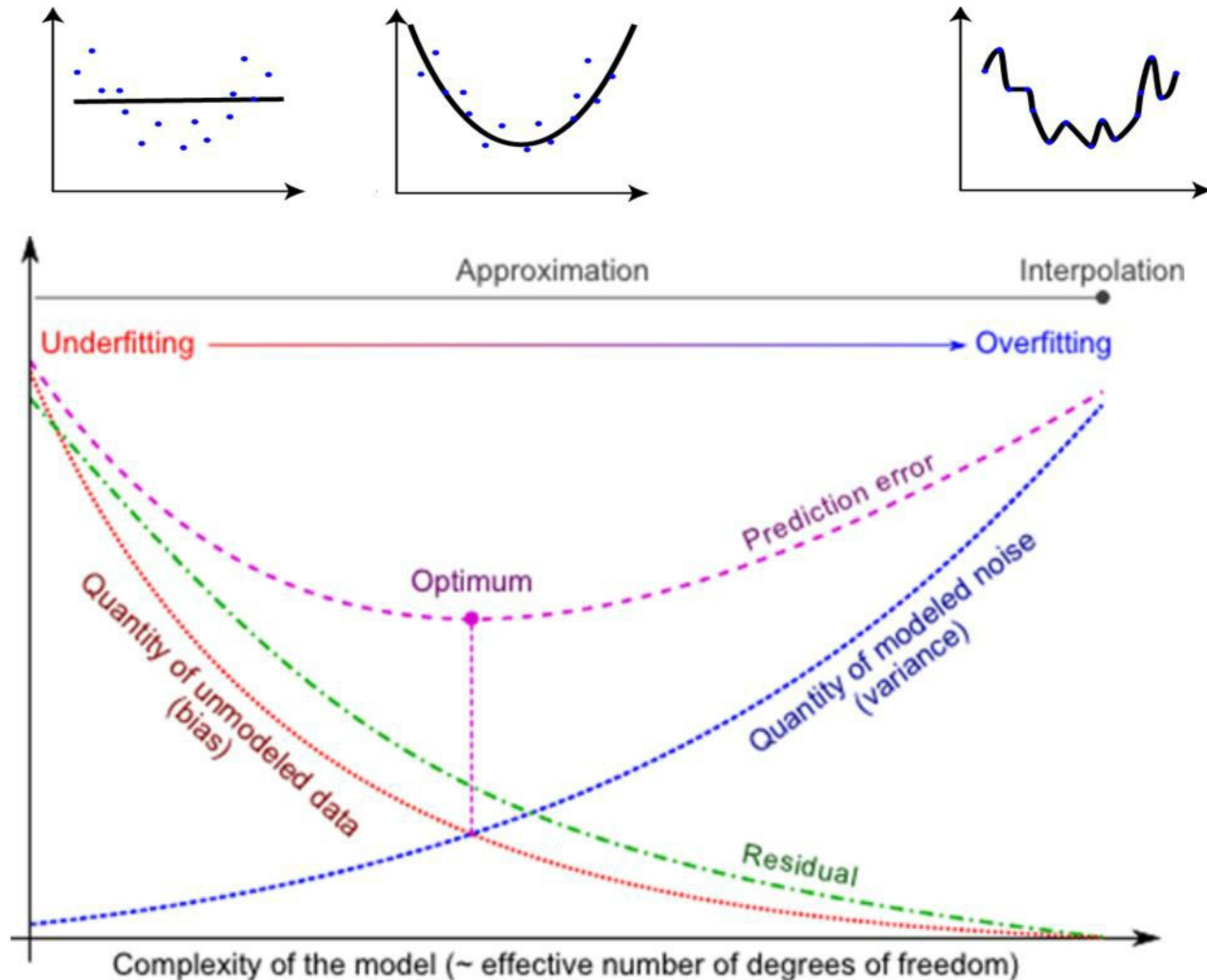
Variance: how much each prediction is different from the average prediction

$f$  Generative function ("truth")

$\hat{f}$  Model optimized for one experiment

$E$  Average over many (-> inf) experiments

# The bias vs variance tradeoff



- Both a model with high bias and a model with high variance will fail to generalize beyond our current dataset
- In practice, the truth is (of course) not accessible, so how to reach the tradeoff?
- If you have a lot of data generated by the same process:
  - divide in training and cross-validation data: optimize your model on the training set, check its performances on the cross-validation set
- In Astronomy this seldom happens
  - still, we need a model which explains all significant features of our data (no underfitting), but which does not model noise (no overfitting)



# Make sure you can explain these in terms of bias vs variance

- "better fit" not always "better model", i.e., "better fit"  $\neq$  "smaller modeling error"
- adding one or more additive features to an existing model (for example, adding a spectral line to a power law) will always give you a better fit
- confirmation of an effect from a different experiment increases your confidence on the result
- In general, between a reasonable model with many parameters and another with fewer parameters, you cannot say which one is better (=gives a smaller modeling error) unless they are nested (hint: number of parameters between non-nested model does not necessarily measure the difference in complexity. Need effective degrees of freedom)



# Significance of a source

- so how to decide if a feature is significant?
- Likelihood Ratio Test:
  - compare two models: a simple one  $H_0$  (null hypothesis) and a more complex one  $H_1$  (alternative hyp.)
  - $H_0$  and  $H_1$  need to be nested: there is a set of parameters which  $H_1 \rightarrow H_0$
- Under certain circumstances (Protasov et al. 2002) Wilks' theorem holds:
$$2(\log L_0(D|\vec{\Omega}_{0,MLE}) - \log L_1(D|\vec{\Omega}_{1,MLE})) \sim \chi^2_{(\dim(\vec{\Omega}_0) - \dim(\vec{\Omega}_1))}$$
- The difference between the logLs (which is the ratio between the Ls) is called simply TS for Test Statistic
- The meaning is: the improvement in the  $f(TS)$  obtained with  $H_0 \rightarrow H_1$  is a random coincidence with probability  $\frac{1}{\chi^2_{(\dim(\vec{\Omega}_0) - \dim(\vec{\Omega}_1))}}$
- if the probability is below a threshold, we can reject the null hypothesis

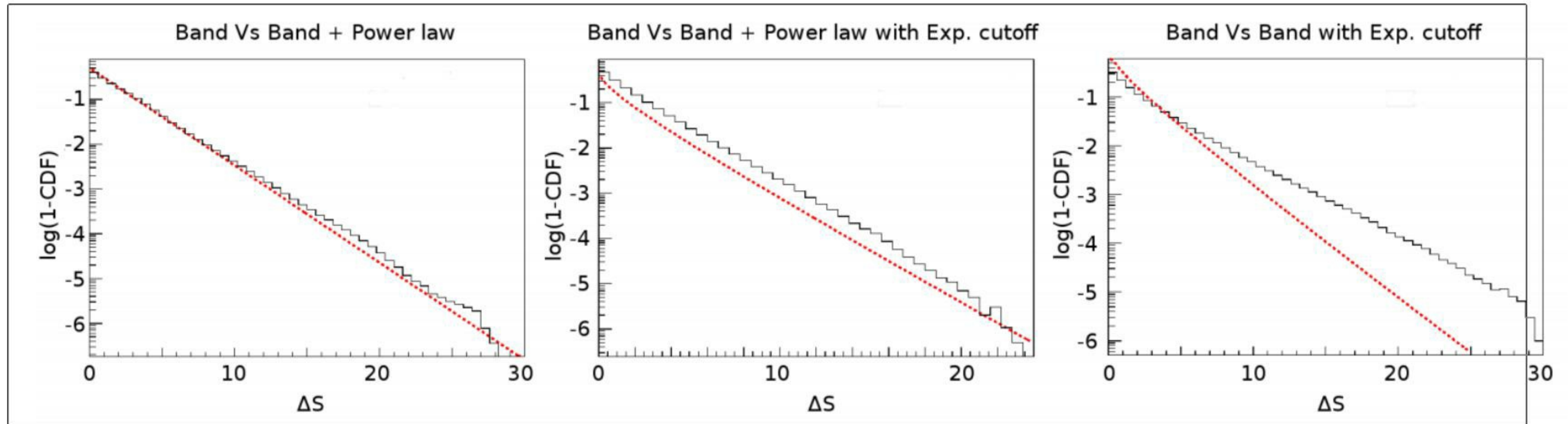


# When does Wilks' theorem hold?



- Conditions:
  - the point where  $H_1 \rightarrow H_0$  must be in a region of the parameter space not at the boundaries of the allowed values
  - and the information matrix must be finite and positive definite. In lay terms, the point where  $H_1 \rightarrow H_0$  must not cause degeneracy in the parameters
  - Example: adding a point source does not satisfy this requirement
    - $H_1 \rightarrow H_0$  for a normalization = 0 for the source
    - but normalization = 0 is at the boundary (no negative source), and the spectrum and the position of the source do not matter anymore if norm = 0 (complete degeneracy)
- Pay a visit to Monte Carlo!
  - simulate many times  $H_0$ , fit it with both  $H_0$  and  $H_1$ , record TS
  - make a histogram of TS and verify if it follows  $\chi^2$
  - if not, use the MC to "calibrate" TS

# Example



From the first LAT GRB Catalog  
(Ackerman et al. 2013)