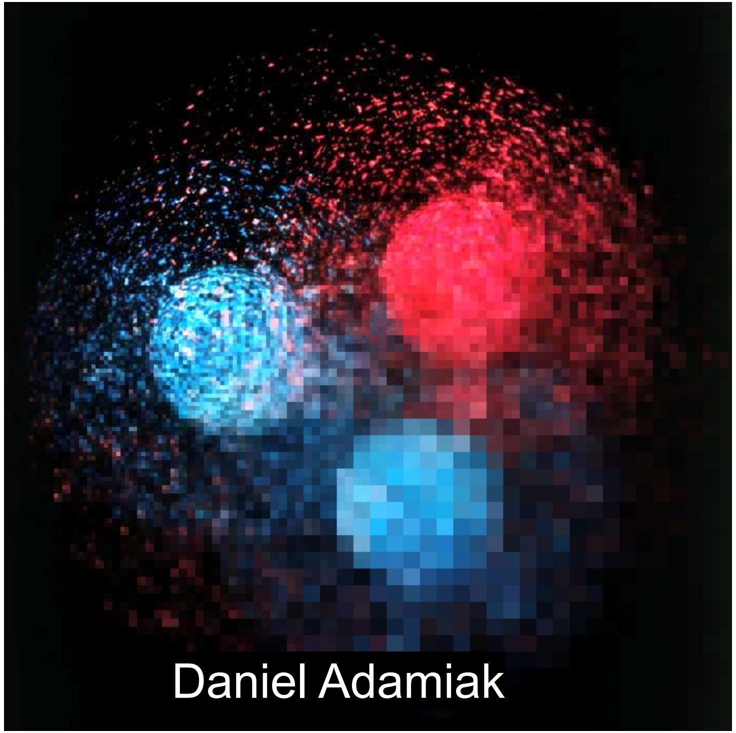
The Resolution of Proton Imaging

How much does data really tell us about the proton?



Daniel Adamiak

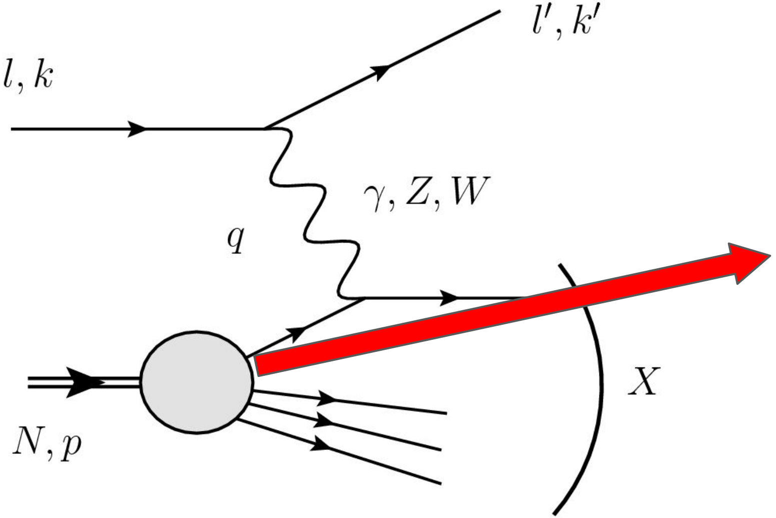
PDFLattice 2024

With Nobuo Sato, Chris Cocuzza, Kevin Braga

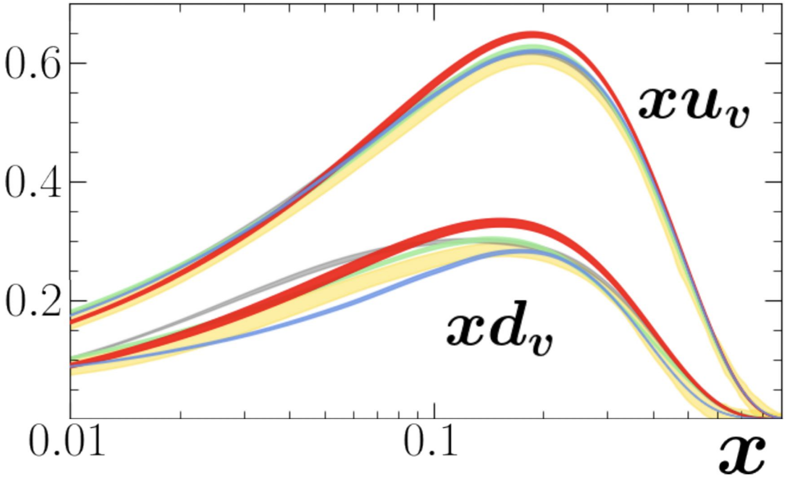
The internal composition and dynamics of the proton is described by Quantum Correlation Functions (QCFs):

For example, the parton distribution functions (PDFs):

Deep Inelastic Scattering (DIS):

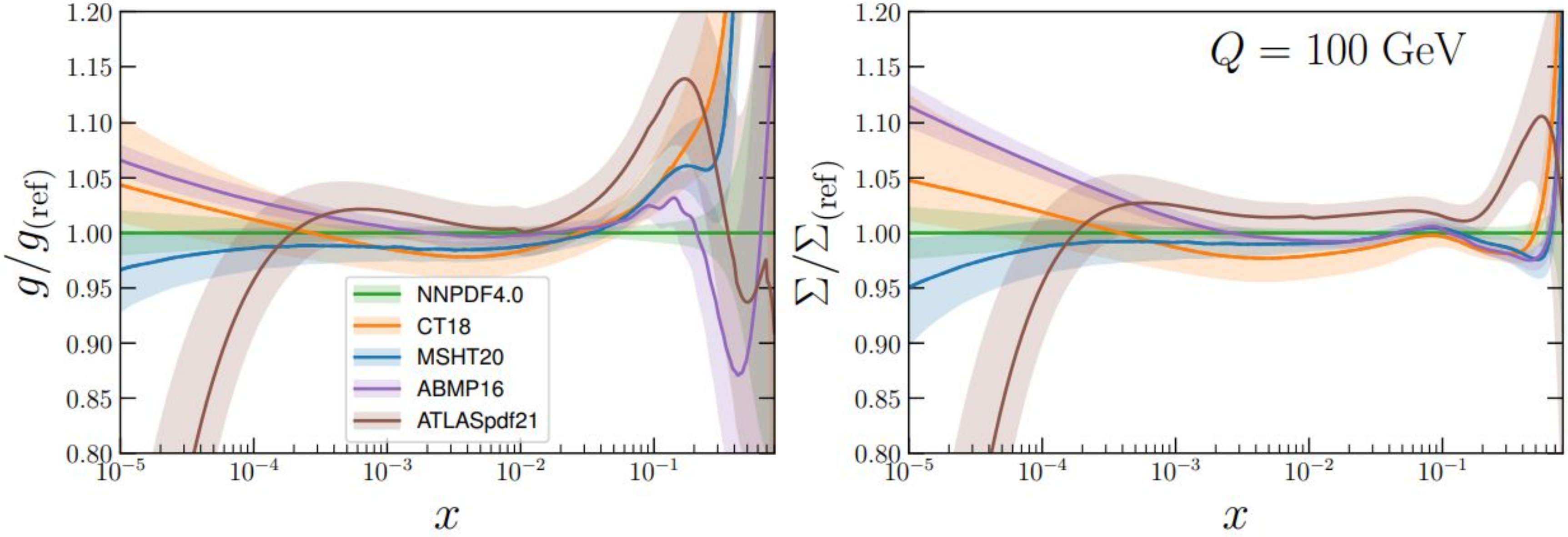


$$x = \frac{-q^2}{2p \cdot q}$$



Measures quark number density as a function of longitudinal momentum fraction, *x*

Comparison of PDF extractions



Precision PDFs (Snowmass 21 WP) [2203.13923v2]

Where do these uncertainty bands come from?

- The data has uncertainty
- The data has a distribution
- Models are used to fit the data

All three sources are combined to give the uncertainty band. But where were the data on the previous plots?

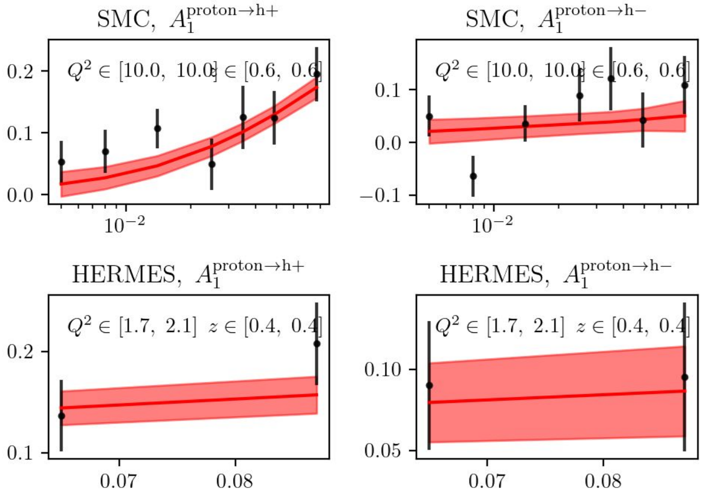
Let’s look at a toy example to explore how each point affects the posterior distribution

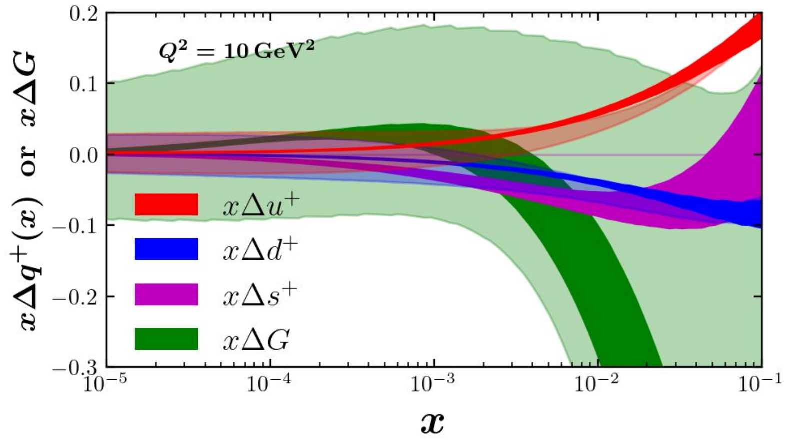
How does this all apply to PDFs?

Observables computed through convolutions

$$F_2(x,Q^2) = x\sum_q e_q^2 \int_x^1 \frac{dz}{z} C_q\left(\frac{x}{z},Q^2\right) f_q(z,Q^2)$$

$$\frac{\partial}{\partial \ln Q^2} f_i(x,Q^2) = \frac{\alpha_s}{2\pi}\int_x^1 \frac{dz}{z} p_{ij}\left(\frac{x}{z},Q^2\right) f_j(z,Q^2)$$





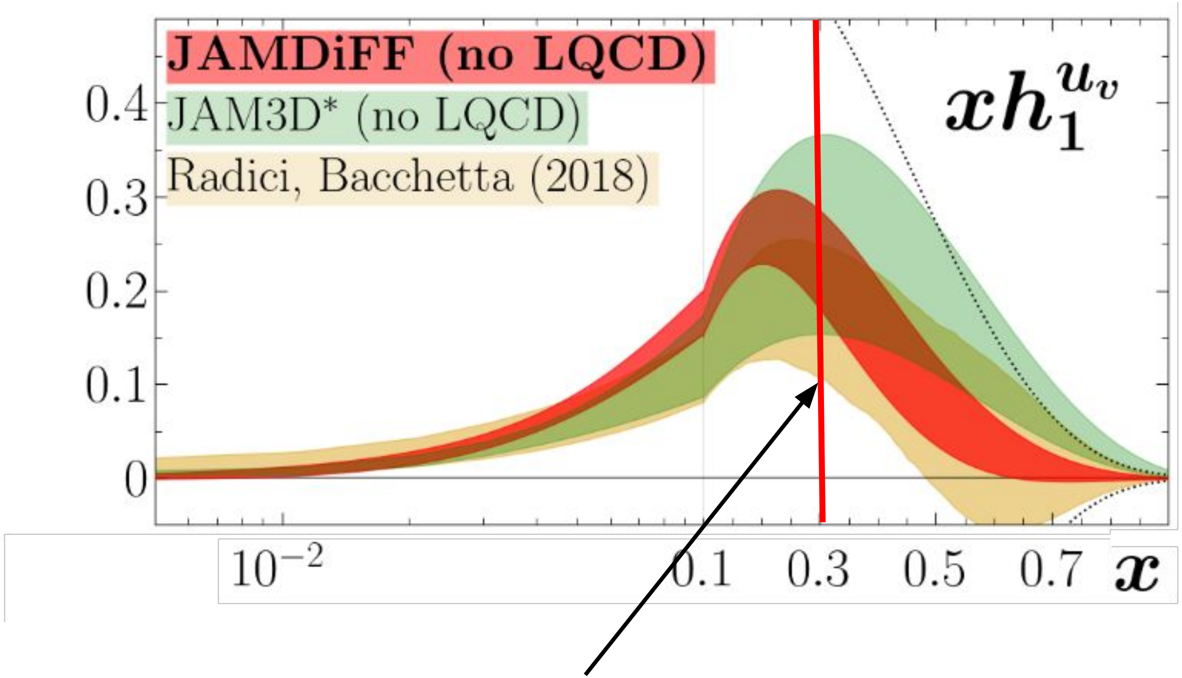
PDFs can only be inferred!

Motivating Example: Transversity PDFs

We model PDFs with

$$f(x, Q_0^2) = x^\alpha (1-x)^\beta$$

Leading to high confidence at large-*x*, despite lack of data



Cocuzza et al. 2308.14857

Data cutoff (Di-hadron production)

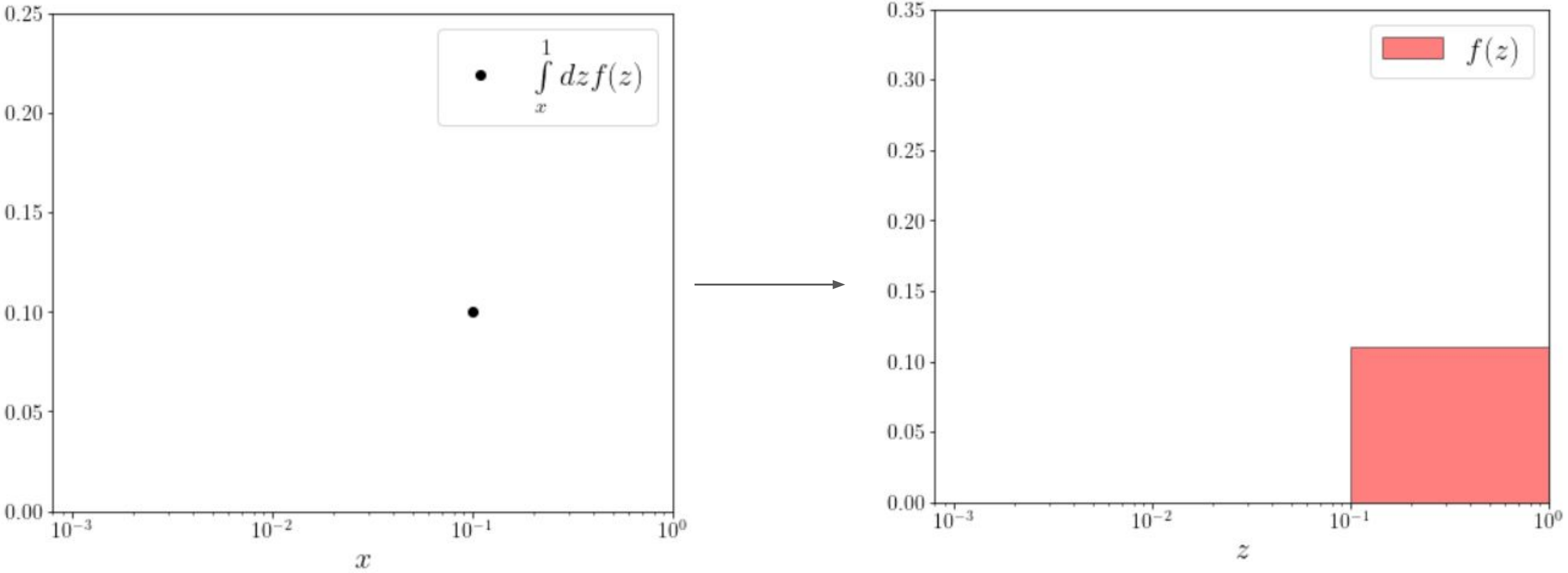
What is the constraining power of the data on the PDFs?

How might we explore the impact data has on input scale PDF without having to worry about model bias?

Let’s look at another toy example

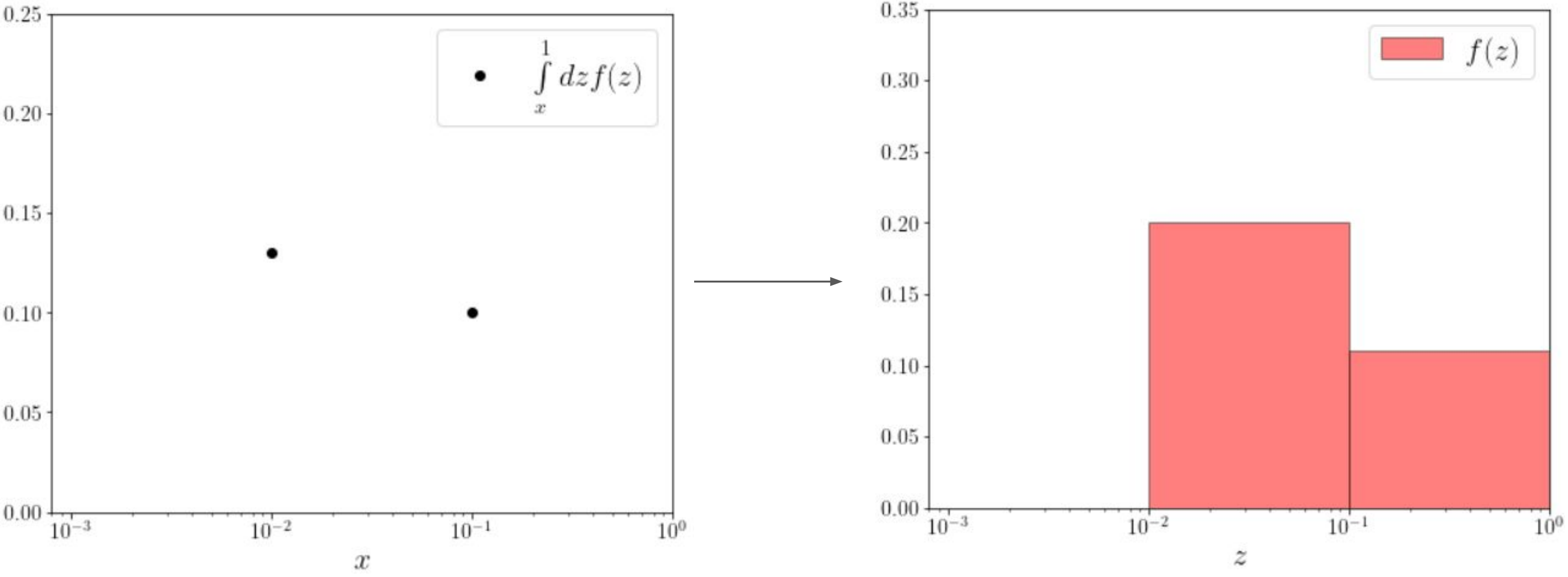
In convoluted observables, what can we learn?

Consider the most trivial convolution: the partial moment. What can we infer from it?



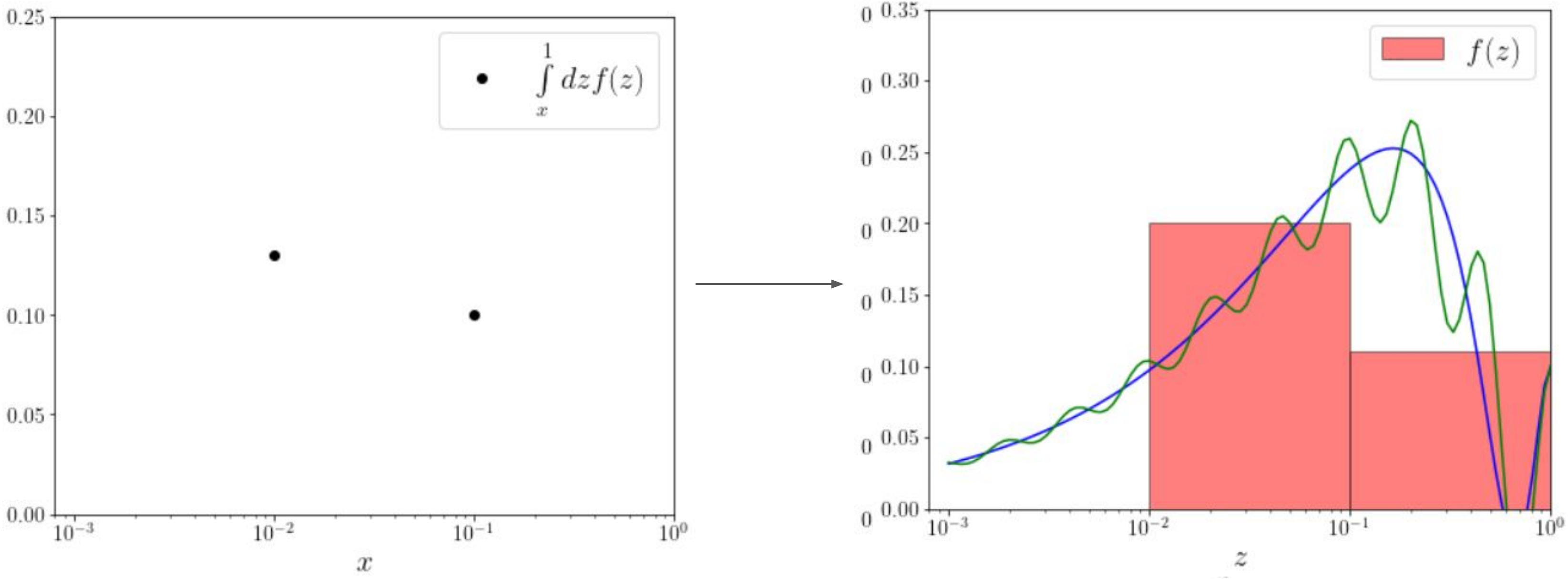
In convoluted observables, what can we learn?

Consider the most trivial convolution: the partial moment. What can we infer from it?



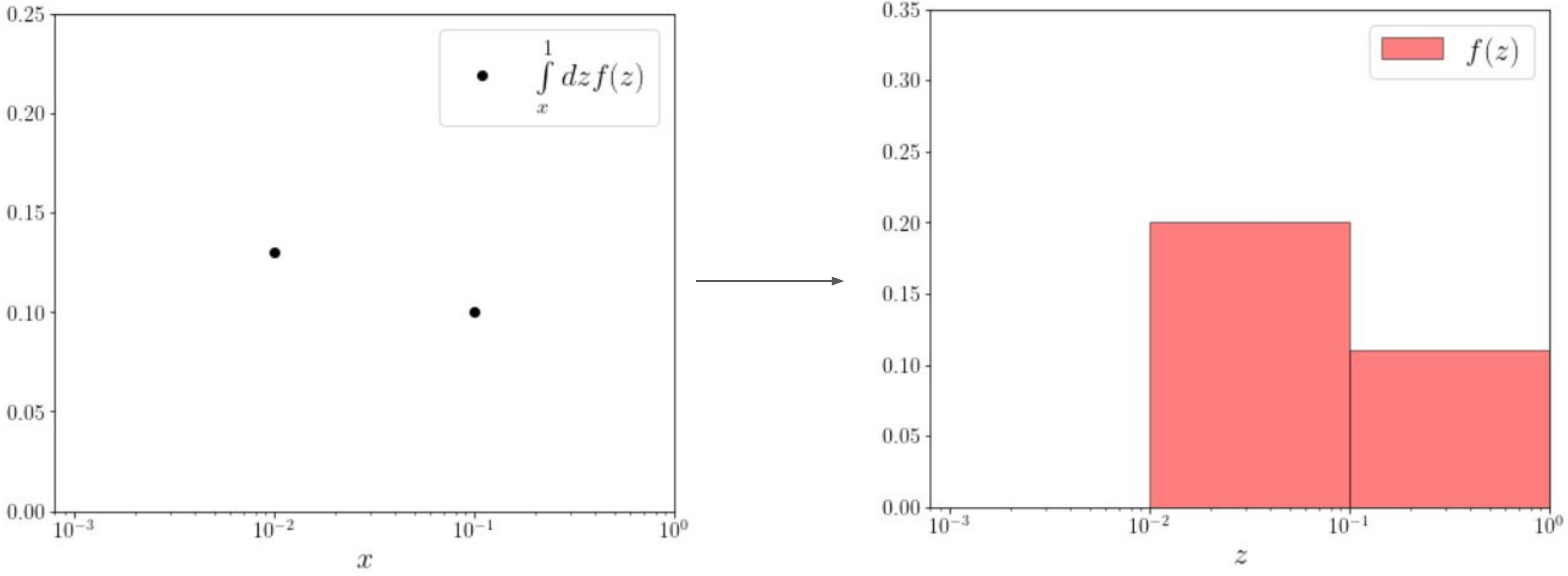
The area under the curve is constrained, not the exact form

Isomorphic to the average value of the function between two *x* points being constrained



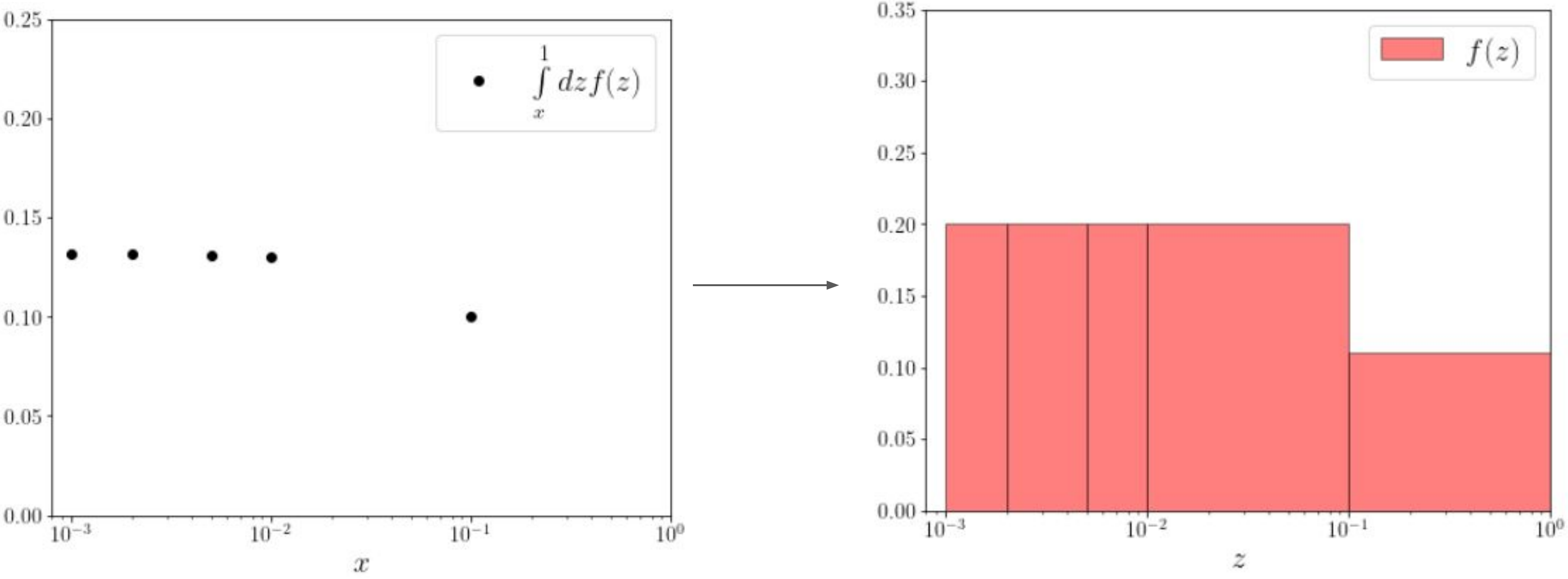
Data tell us about the resolution of PDFs: $\langle f(x_i<x<x_{i+1})\rangle$

These histograms are the real constraints of data



Data tell us about the resolution of PDFs: $\langle f(x_i<x<x_{i+1})\rangle$

These histograms are the real constraints of data



How can we infer the PDF resolution from DIS data?

- The convolutions involved in computing observables are much more complex

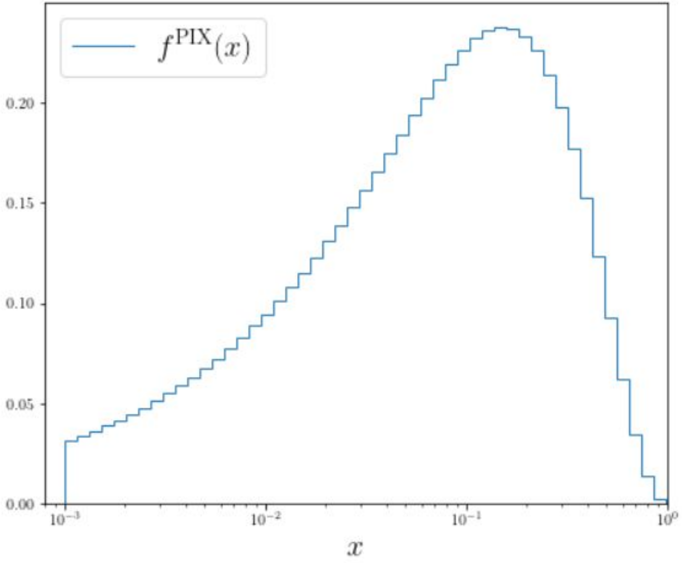
$$F_2(x,Q^2) = x\sum_q e_q^2 \int_x^1 \frac{dz}{z} C_q\left(\frac{x}{z},Q^2\right) f_q(z,Q^2)$$

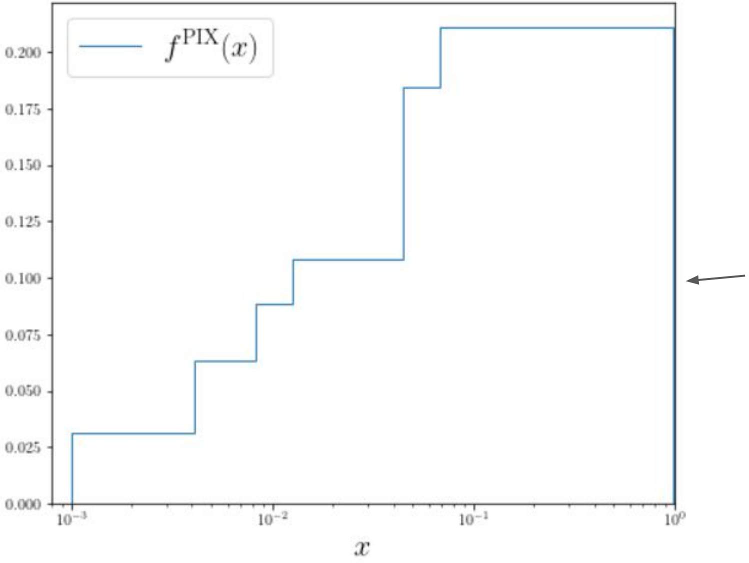
$$\frac{\partial}{\partial \ln Q^2} f_i(x,Q^2) = \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} p_{ij}\left(\frac{x}{z},Q^2\right) f_j(z,Q^2)$$

- I don’t know how to infer the resolution analytically
- But I have an algorithm that starts with a high resolution fit and gradually lowers the resolution until the quality of the fit begins to degrade

How to determine the lowest resolution

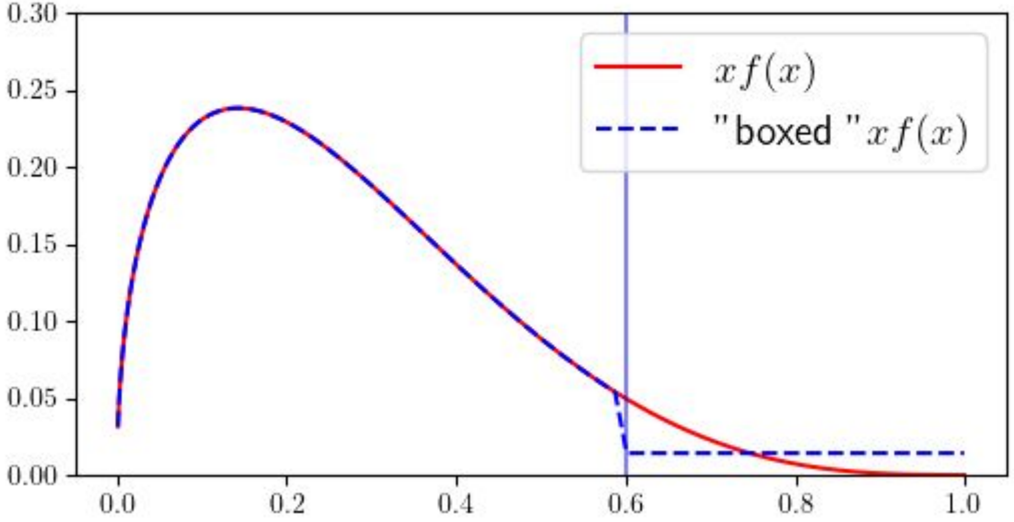
Compare high and low resolution fit, if both give good χ^2 (within some tolerance) then we have a better idea of how low the resolution really is





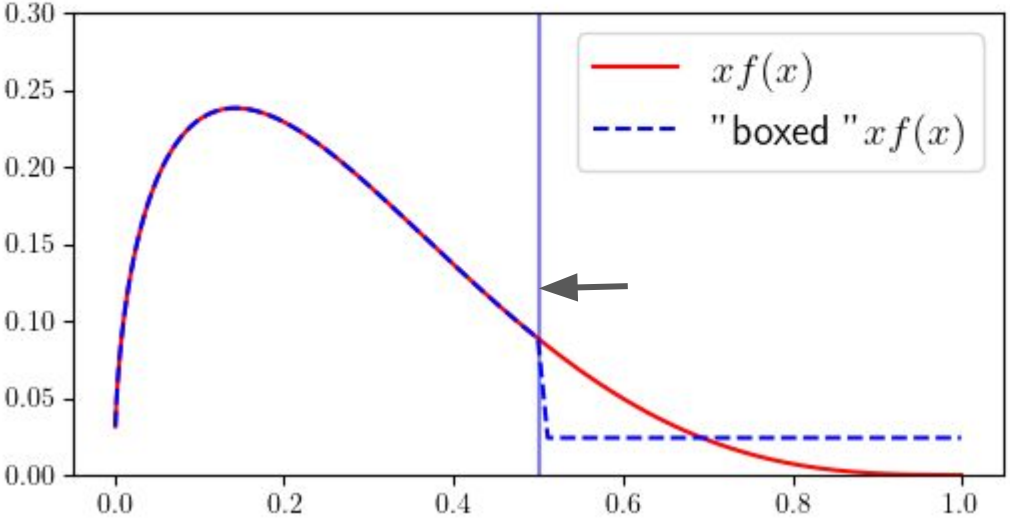
More informative about our true state of knowledge

Focus only on large-*x* for now: One Boxing



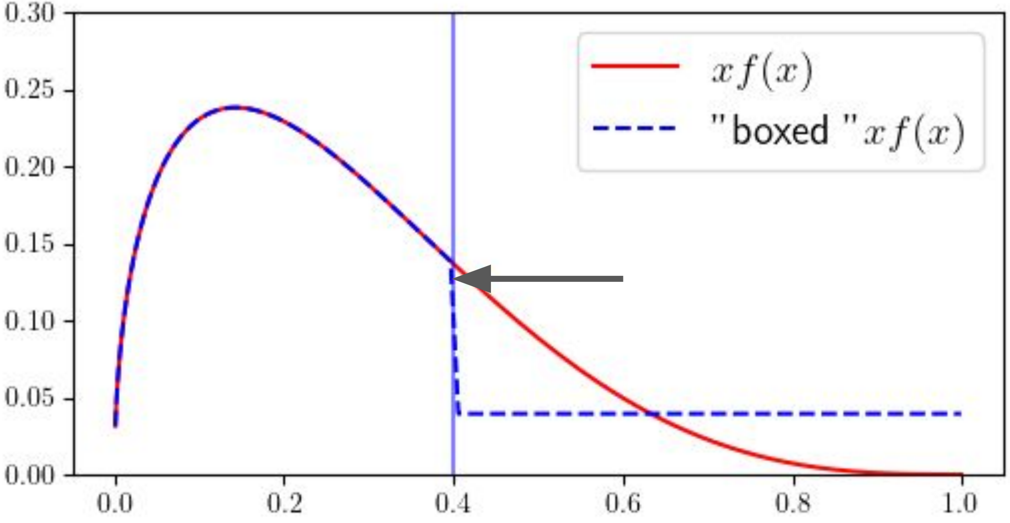
- Draw a box from x=x0 to x=1
- Average PDF within that box
- Compute chi2
- Reject box if change in chi2 is too large

Focus only on large-*x* for now: One Boxing



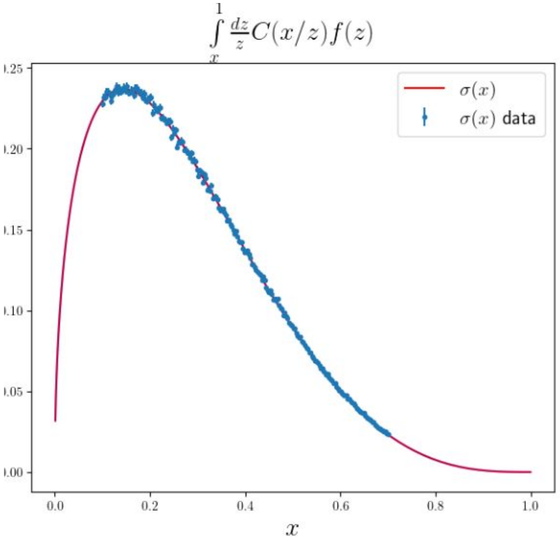
- Increase the size of the box until the chi2 tolerance is reached

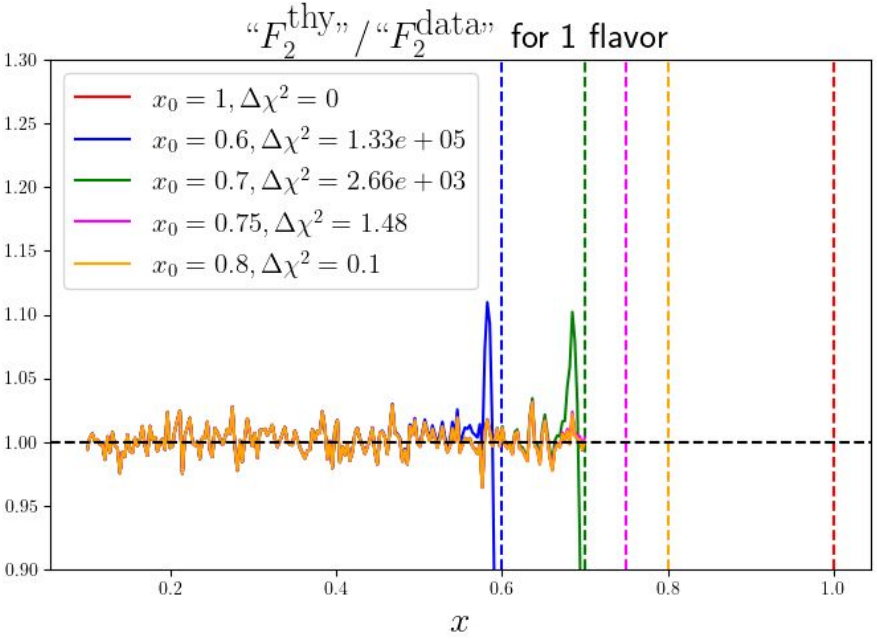
Focus only on large-*x* for now: One Boxing



- Increase the size of the box until the chi2 tolerance is reached
- The larger we can make the box, the less the data is constraining us

Toy Example: “F2” = $\int\limits_x^1 \frac{dz}{z} C(x/z) f(z)$

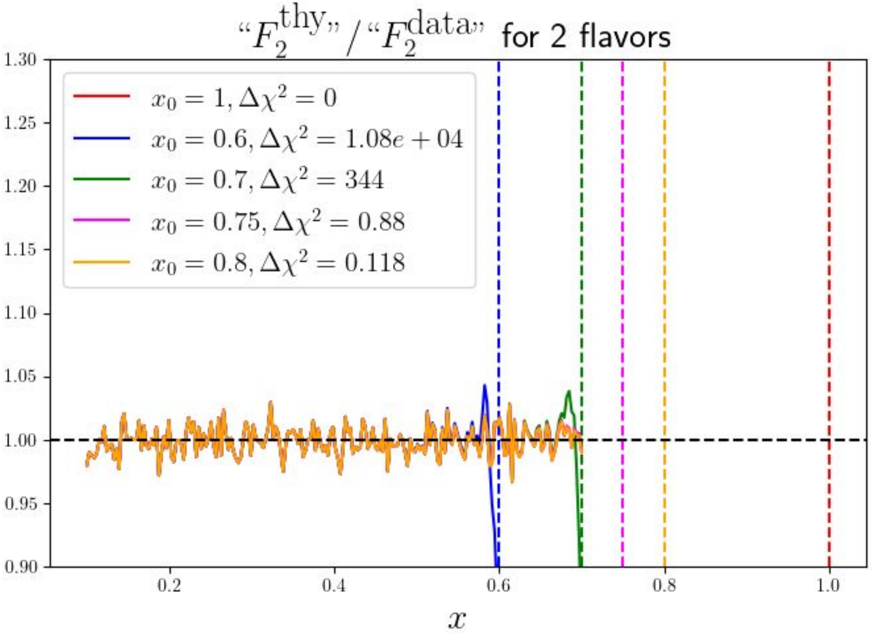




Toy Example: “F2” = $\sum_i \int_x^1 \frac{dz}{z} C(x/z) f_i(z)$

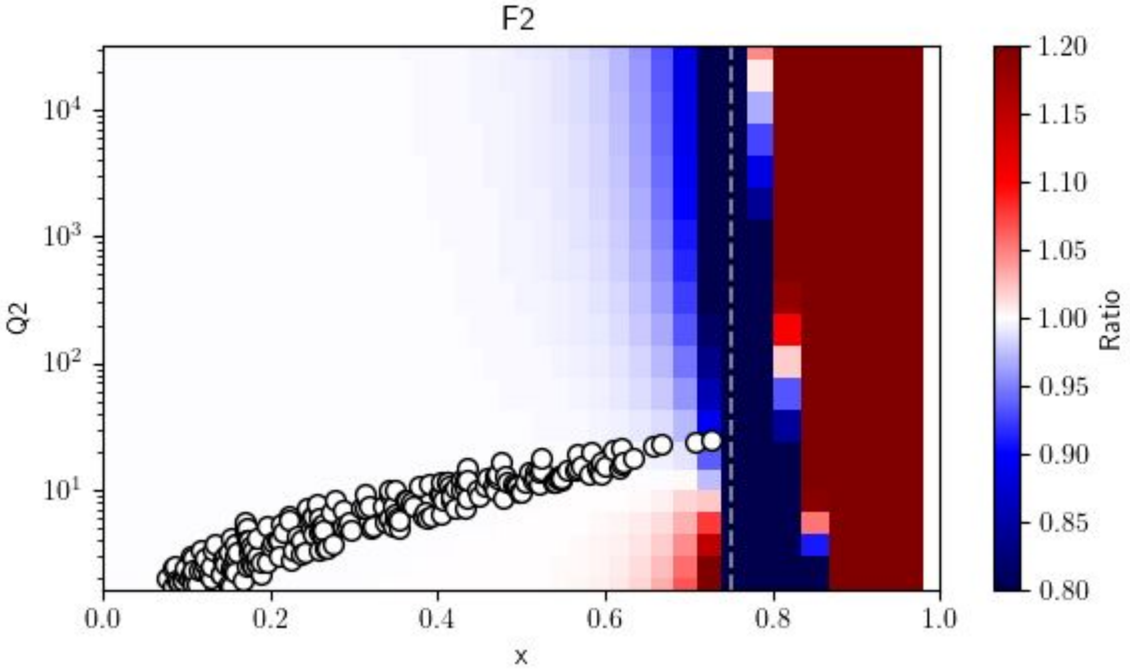
Add a second ‘suppressed’ flavor

The data is a less sensitive to this second flavor

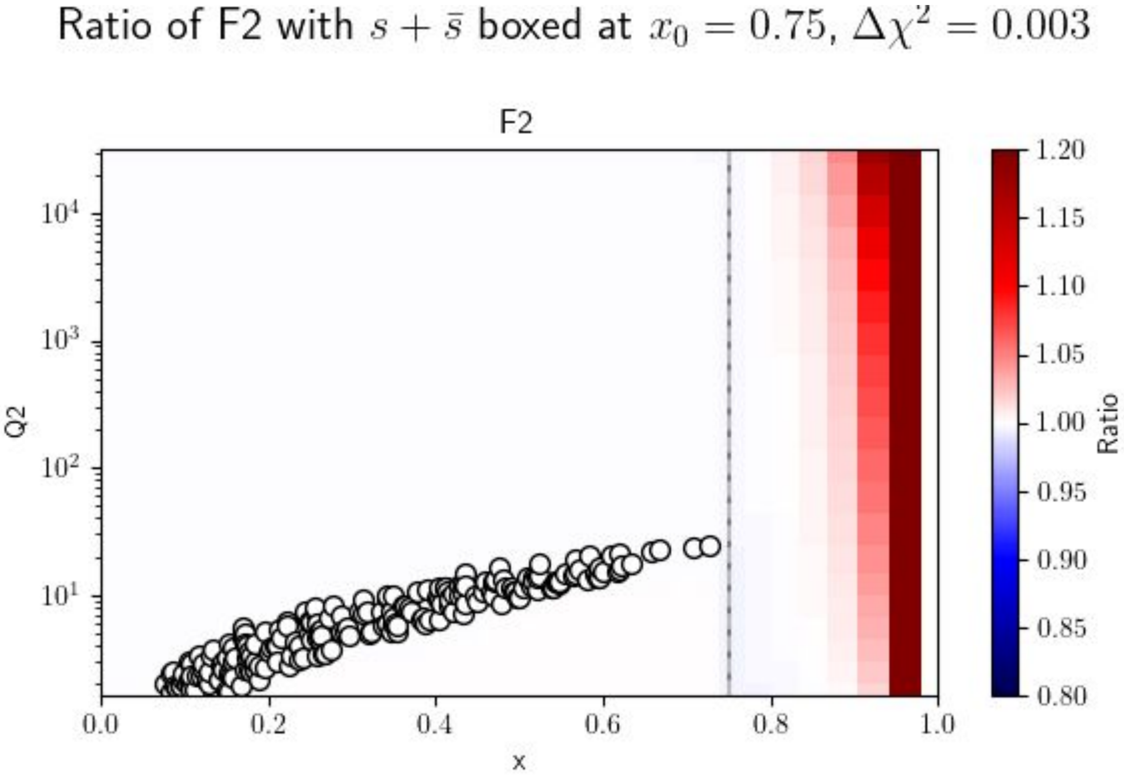


Real Example: SLAC DIS proton data: Boxing u-PDF

Ratio of F2 with $u(uv\&\bar{u})$ boxed at $x_0 = 0.75$, $\Delta\chi^2 = 3.771$



Real Example: SLAC DIS proton data: Boxing s+sbar-PDF



Conclusions

- Model flexibility is somewhat isomorphic to replica uncertainty, so we should
- be careful about comparing different models
- Data can only constrain the resolution , or the average value, or integrals of the PDFs
- The resolution of PDFs is related to the distribution of data (in *x* (mostly))
- The uncertainty of data can then be turned into uncertainty of the average value of the PDFs within a bin - Decorolating (mostly) data distribution from data uncertainty
- Boxing can be an important diagnostic tool for understanding when we are extrapolating QCFs

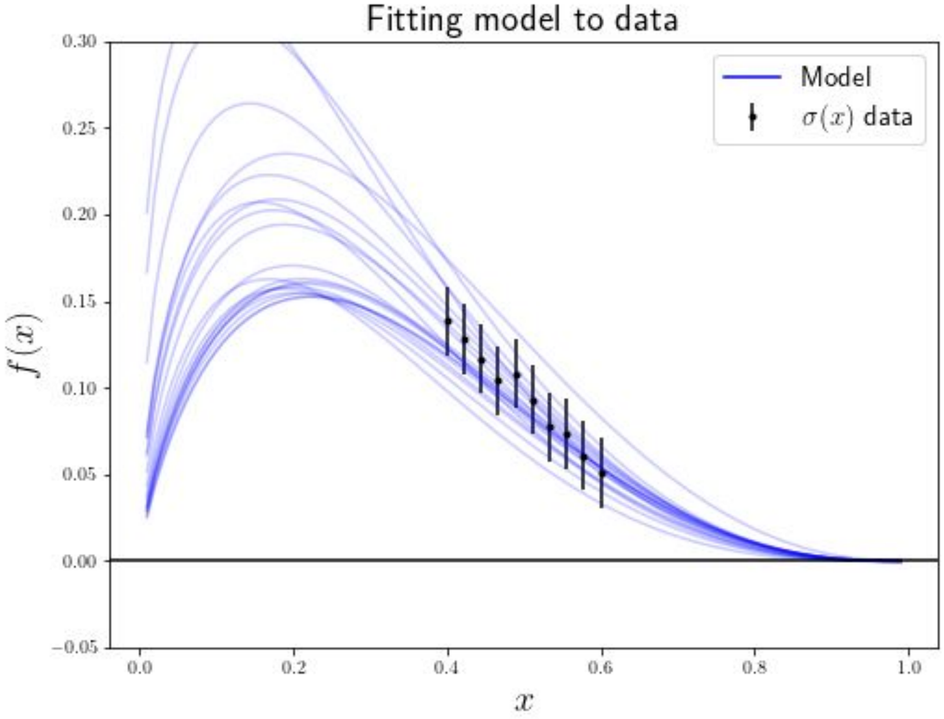
Thank you!

Backup Slides

What can data tell us about what we’re measuring?

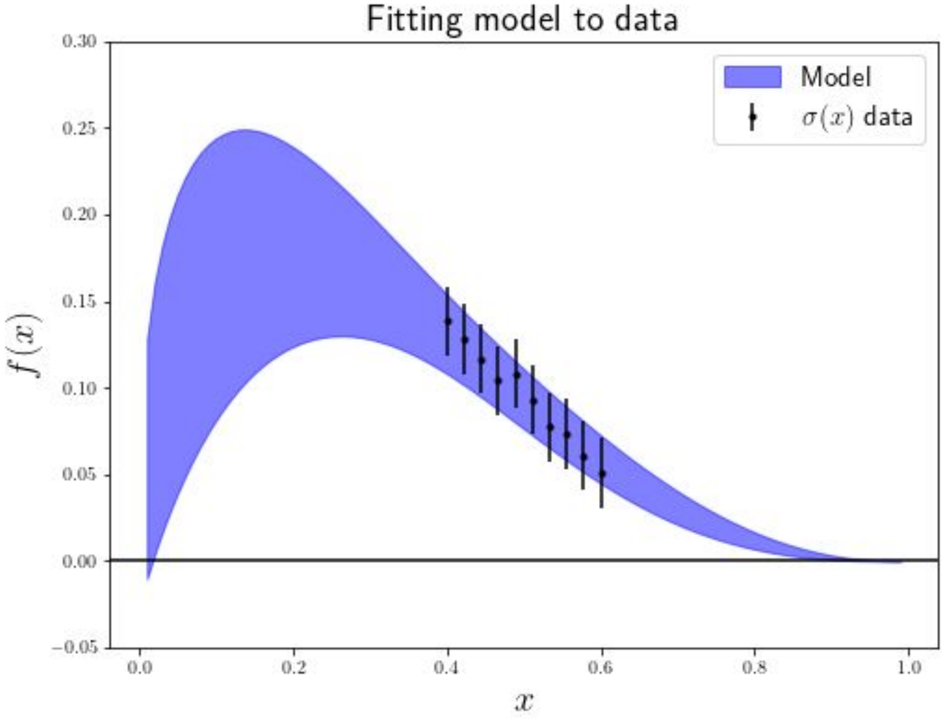
Many fits are done using a model with free parameters.

Different parameters that give a good χ^2 result in ‘replicas’



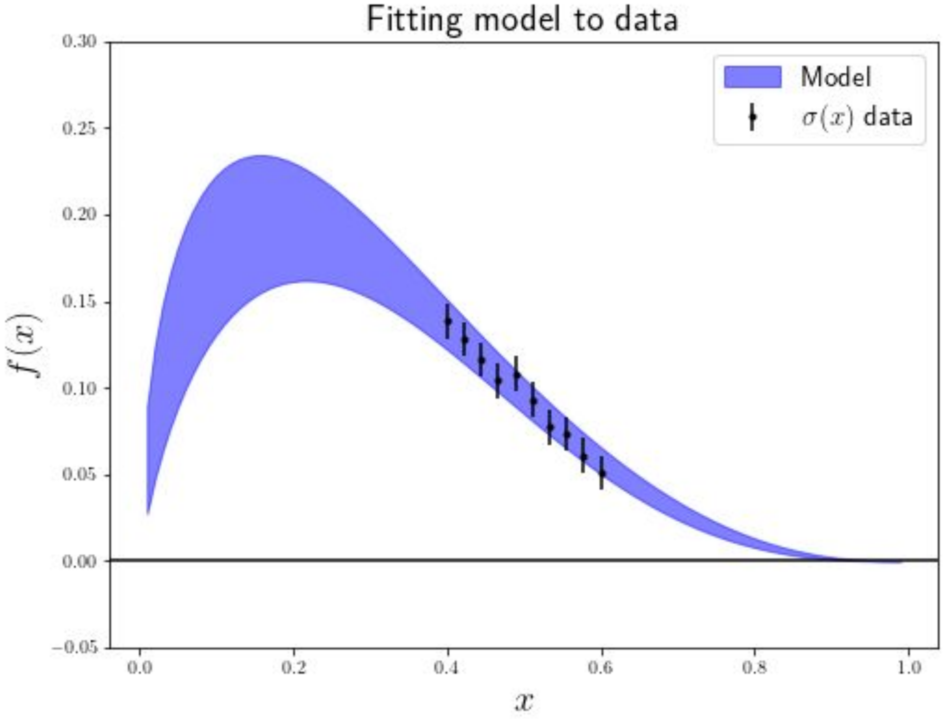
What can data tell us about what we’re measuring?

The uncertainty band is the distribution of replicas



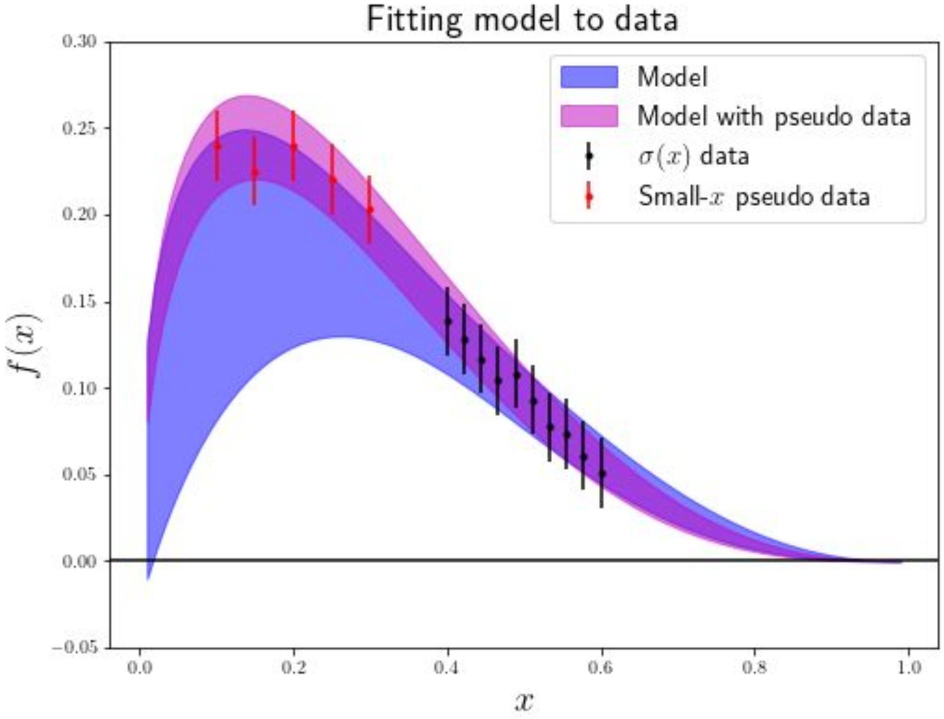
What can data tell us about what we’re measuring?

The more precise the data, the more constrained the fit



What can data tell us about what we’re measuring?

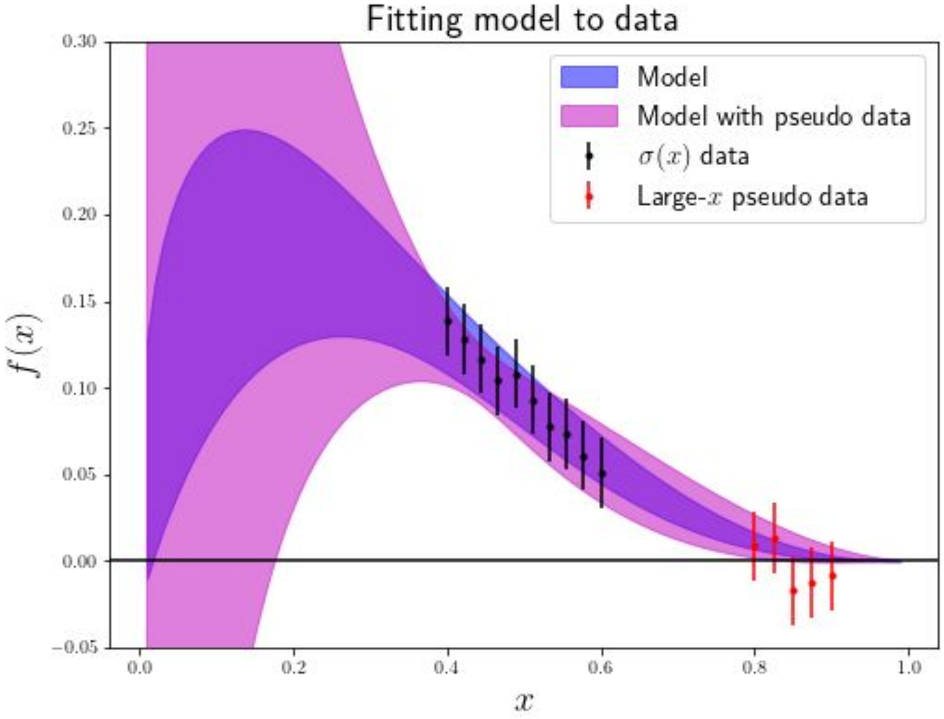
Additional data can also constrain the fit



What can data tell us about what we’re measuring?

But sometimes additional data does not constrain the fit

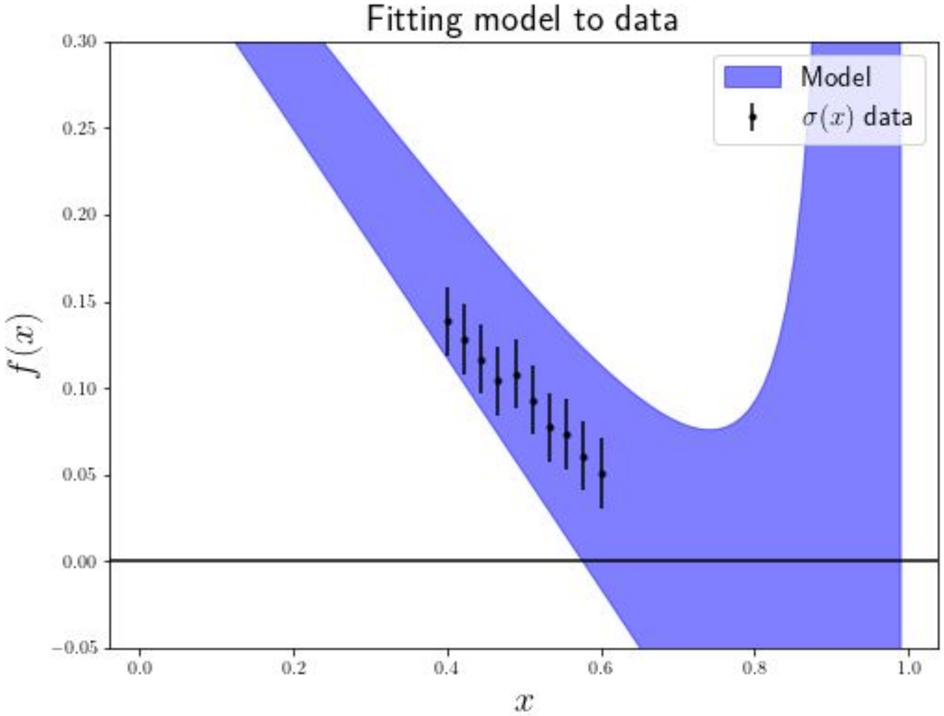
This happens when the model is already very confident



What can data tell us about what we're measuring?

Changing the model also changes the posterior uncertainty

In general, the more flexible the model, the larger the uncertainties (a problem for neural nets?)



Genetic Algorithm (GA)

- Means with which to maximize fidelity $\mathcal{F}$

$$\mathcal{F} = \frac{1}{|\beta|\chi^2}$$

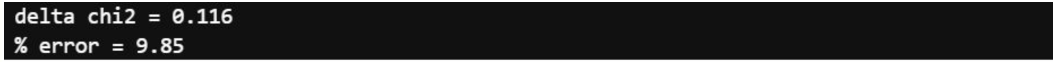
$$|\beta| \equiv \text{number of boxes}$$

- Method
  a. Generate random guesses for where to box the PDF(s)
  b. Evaluate fidelity for all guesses
  c. Select best guesses to reproduce
  d. From those best “parents”, mutate them via 5 distinct types of mutations
  e. Repeat until desired chi2 or fidelity is achieved

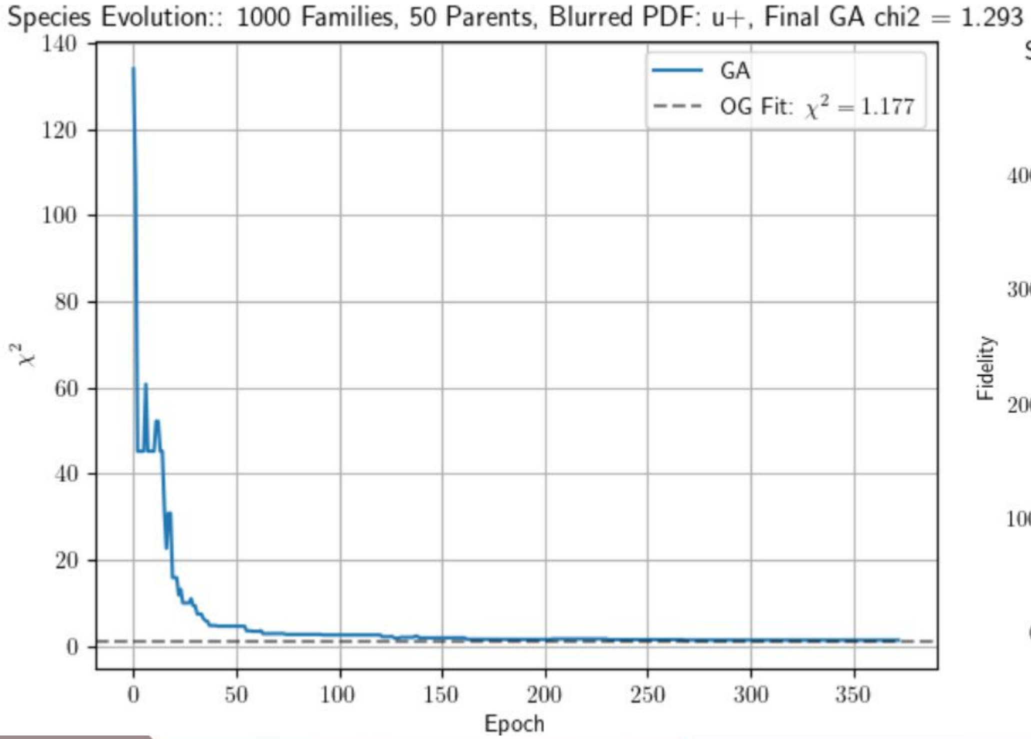
Mutation Types

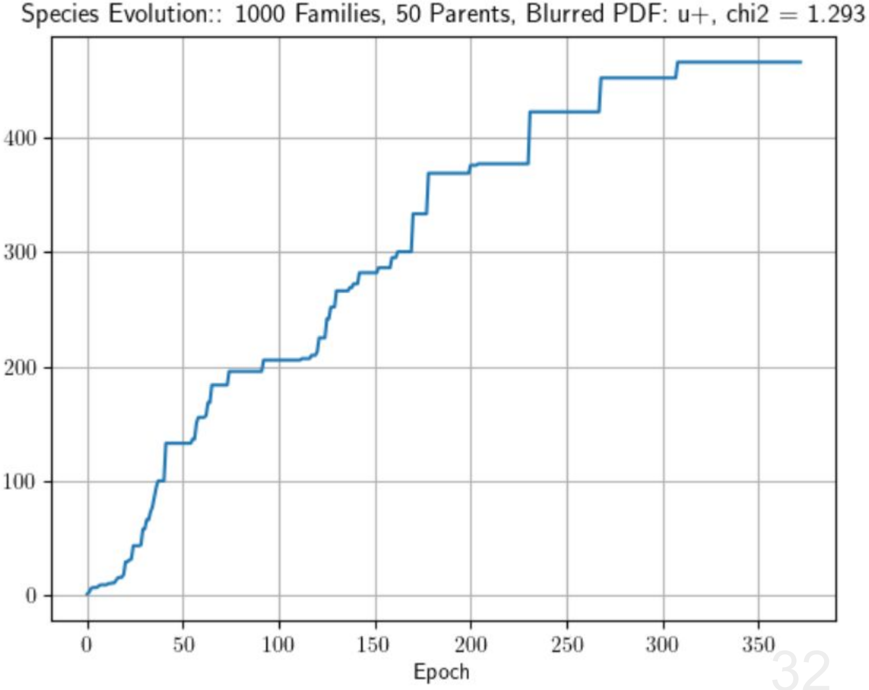
1. Add
   a. Adds a box somewhere random in the PDF
2. Remove
   a. Removes one of the existing boxes in the PDF
3. Modify
   a. Enlarges or shrinks the size of a randomly selected box
4. Swap
   a. Swaps the positions of two boxes
5. Shuffle
   a. Shuffles the sequence of all boxes

Results: Boxing u+, Training

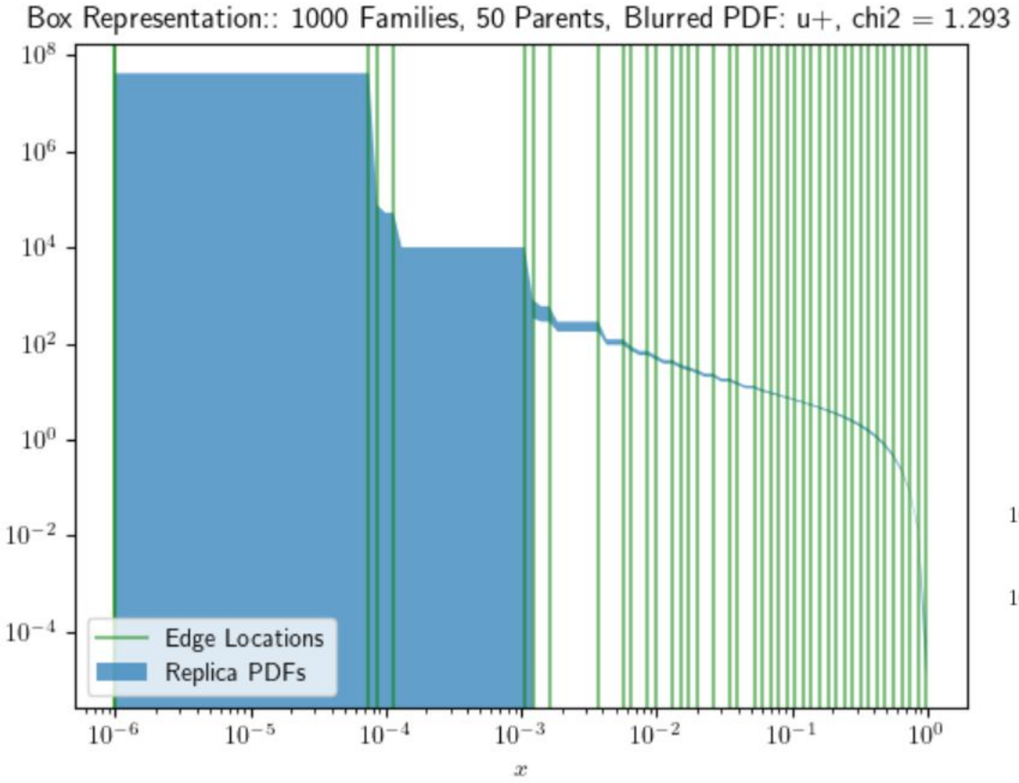


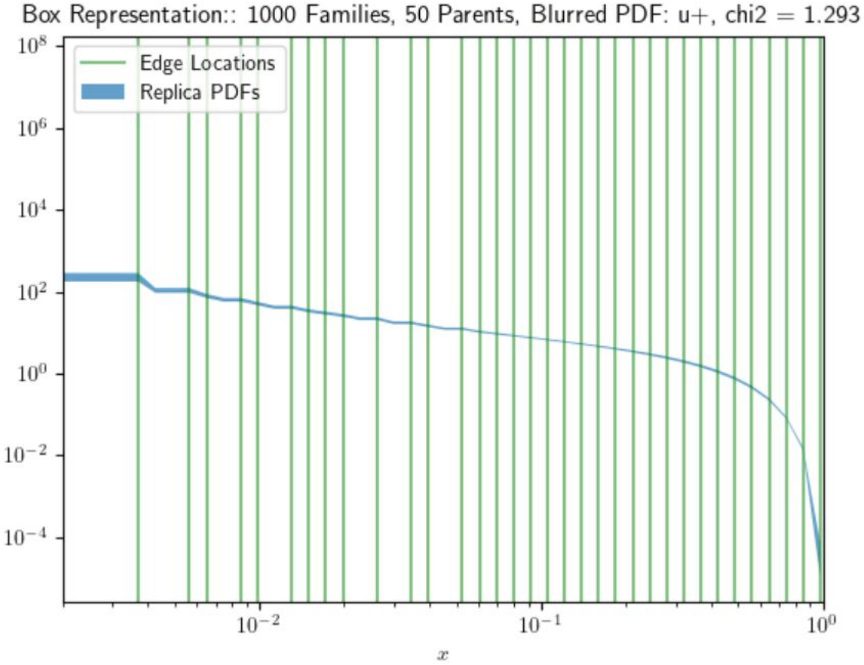
```
delta chi2 = 0.116
% error = 9.85
```

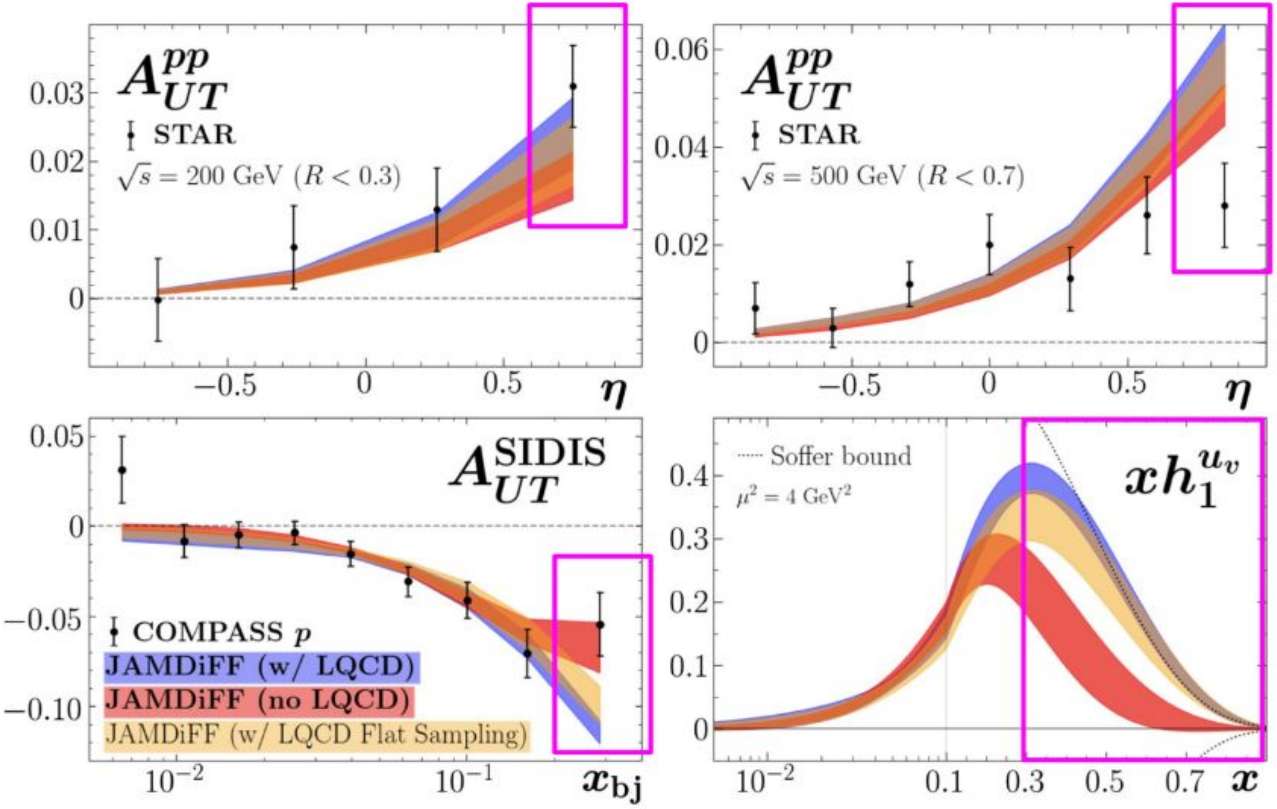




Results: Boxing u+, PDFs





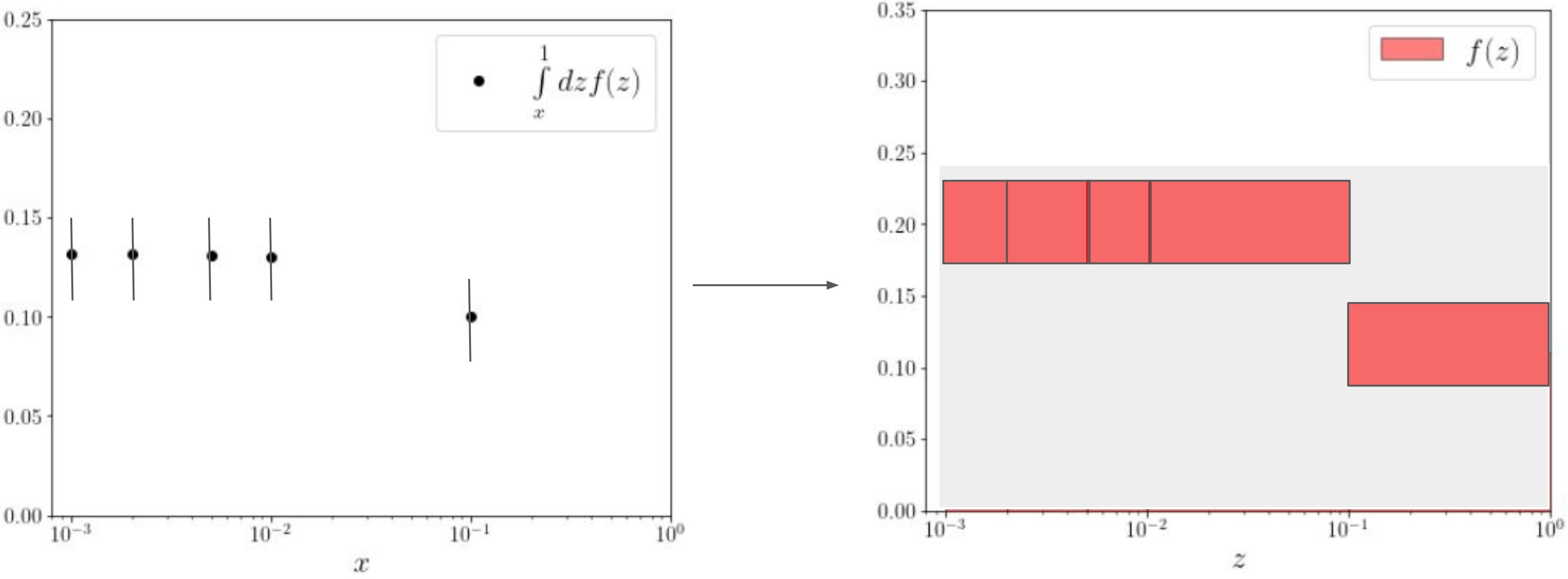


Ask Chris or Nobuo about the meaning of this

Pitonyak, Cocuzza, Metz, Prokudin, NS, `24 (PRL)
Cocuzza, Metz, Pitonyak, Prokudin, NS, Seidl `24 (PRL)
Cocuzza, Metz, Pitonyak, Prokudin, NS, Seidl `24 (PRD)

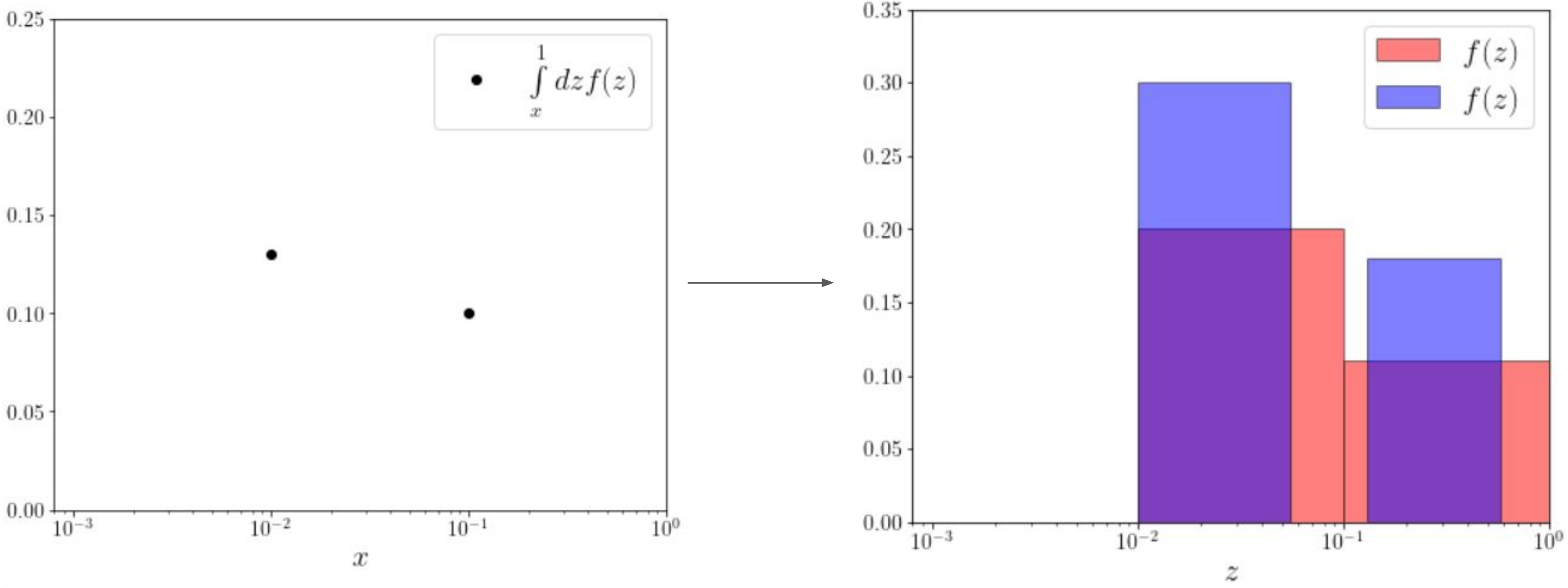
Data Uncertainty can also be incorporated

These histograms are the real constraints of data



What do we expect for the distribution of valid replicas?

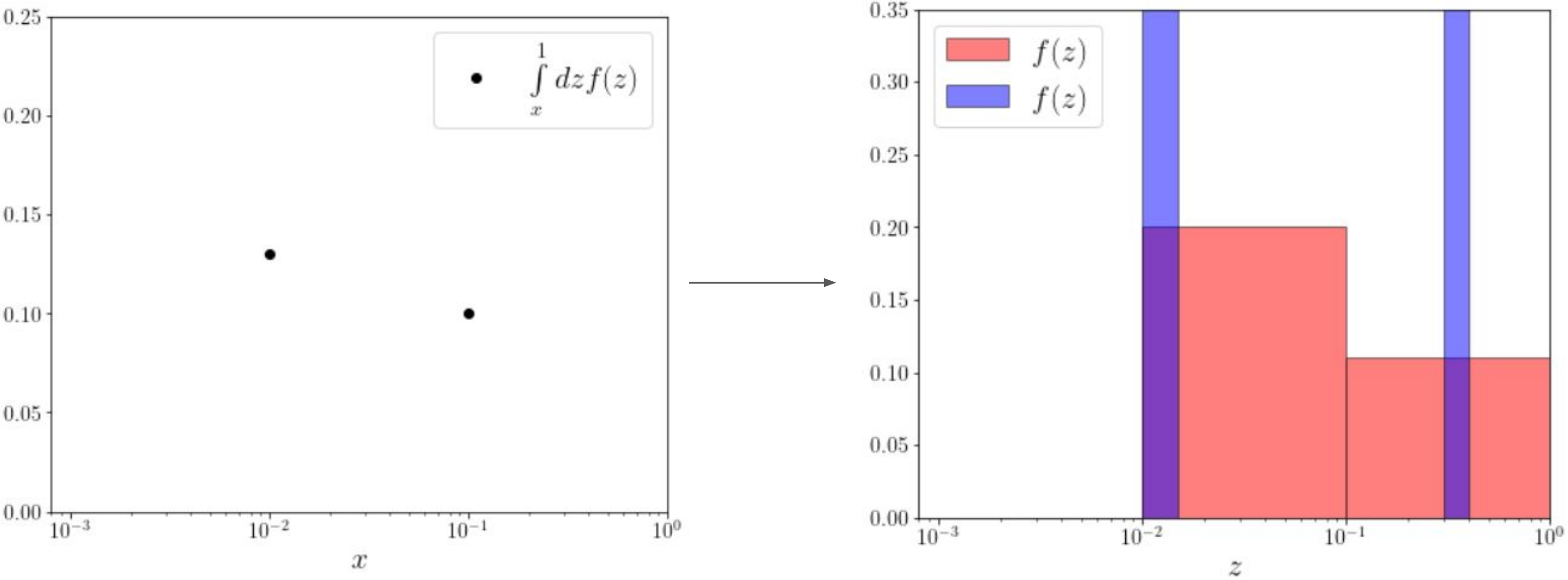
Consider loading all the information into individual pixels



Set some pixels to 0, let the other pixels pick up the rest of the area

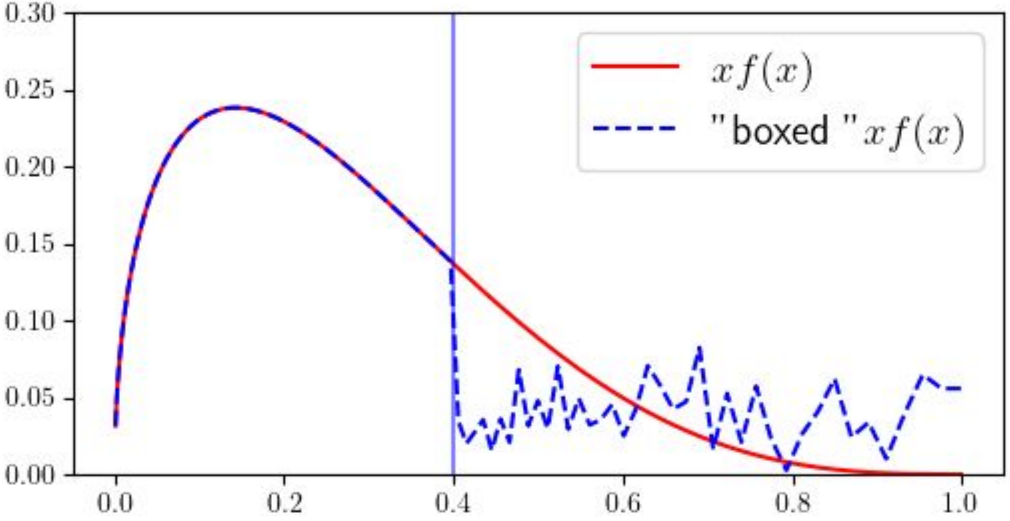
What do we expect for the distribution of valid replicas?

The finer your grid, the wider the distribution of replicas!



This implies an infinite uncertainty for infinitely flexible models, even when data has zero uncertainty!

Focus only on large-*x* for now: One Boxing



- To truly test the sensitive to the shape, we also add random noise