Uncertainties in global analyses: Parton distributions need representative sampling

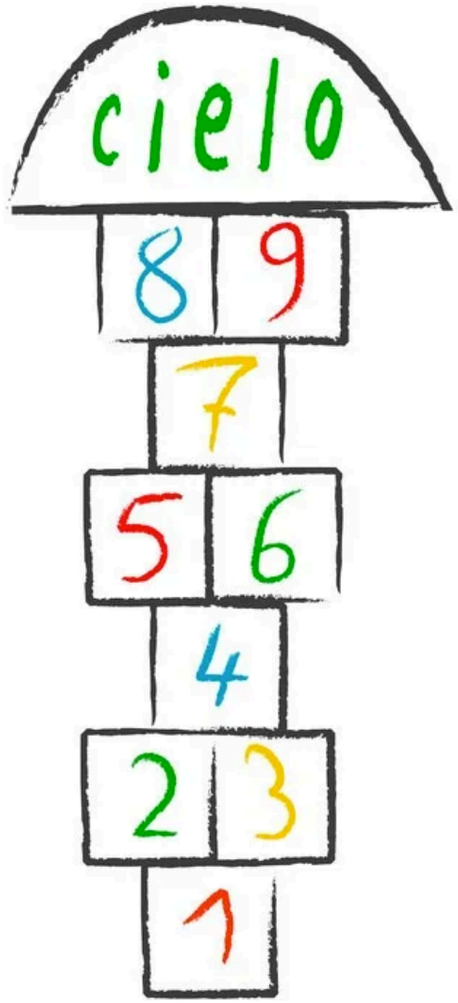
Aurore Courtoy

Instituto de Física
National Autonomous University of Mexico (UNAM)

based on 2205.10444

in collaboration with J. Huston, P. Nadolsky, K. Xie, M. Yan & C.-P. Yuan

2nd detector incubator meeting — December 2022



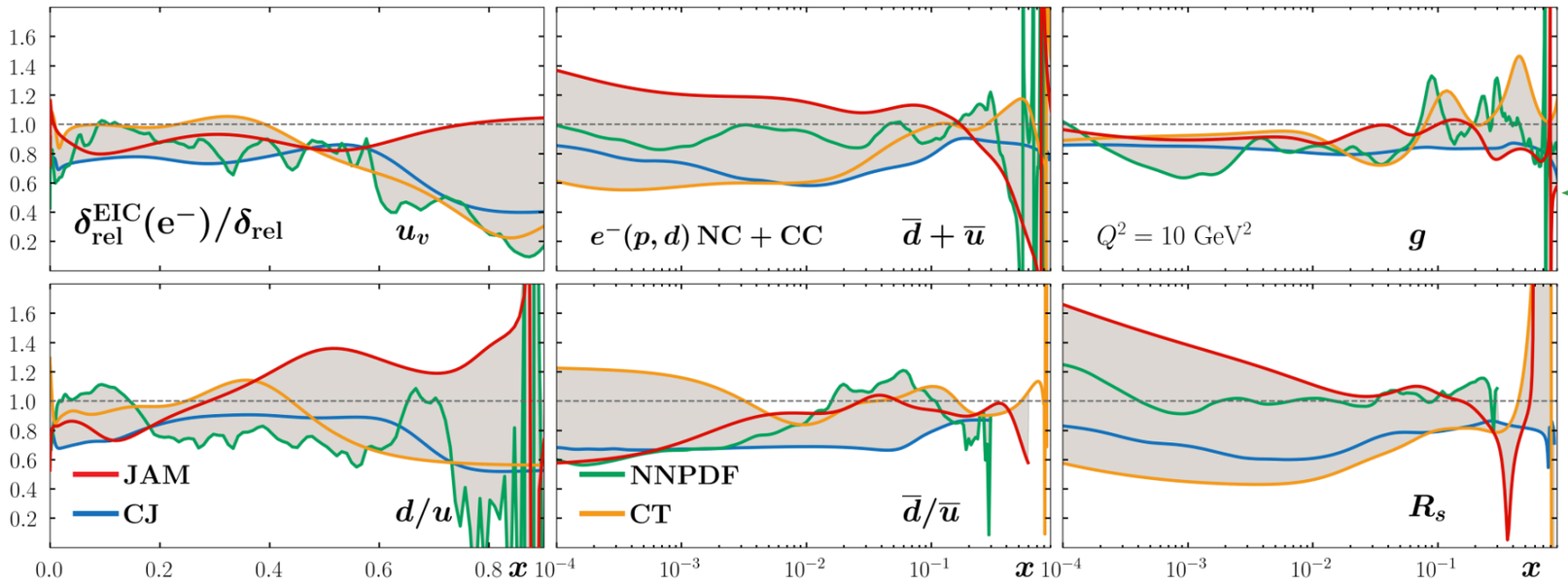
Instituto de Física UNAM

CONACYT
Consejo Nacional de Ciencia y Tecnología

dgapa
Dirección General de Asuntos del Personal Académico

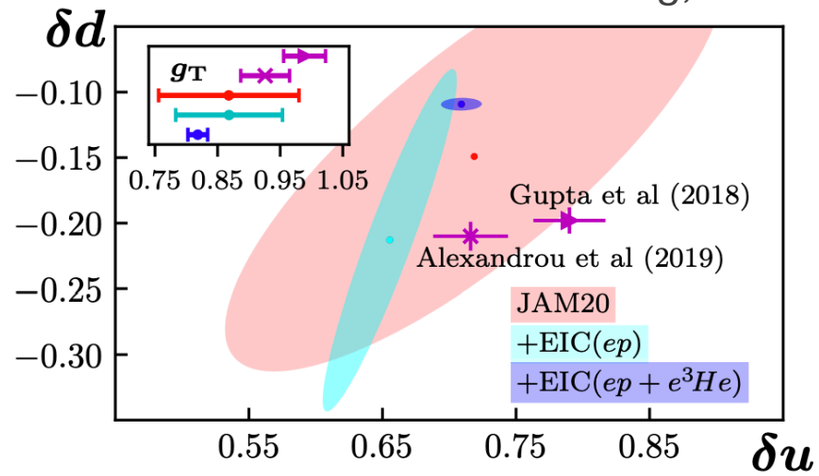
Do diversity and complementarity matter for the EIC?

Examples from
EIC for high-energy physics (Snowmass 21 WP)
[2203.13199]



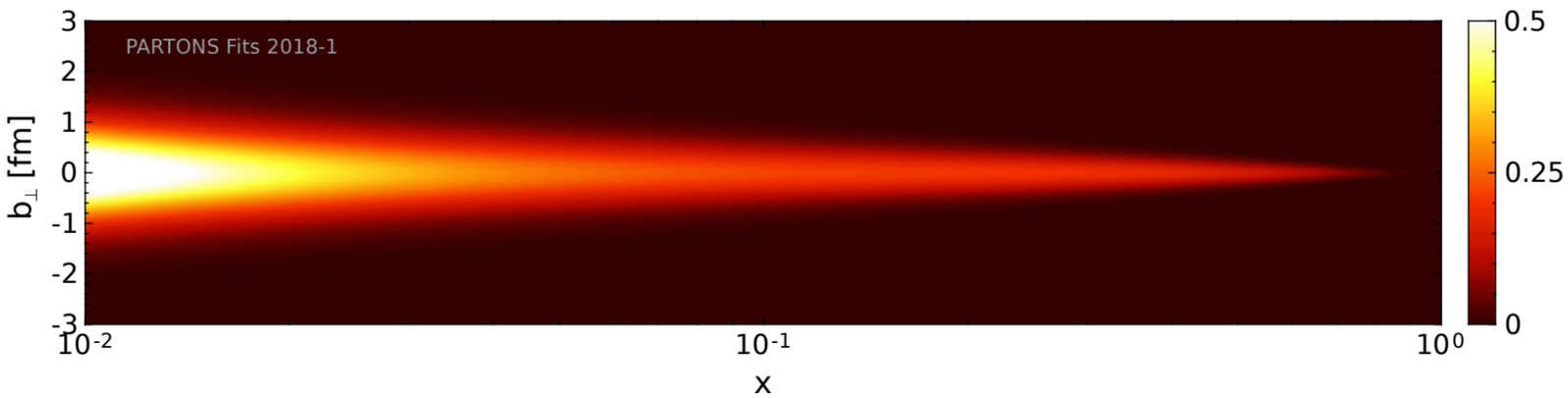
Unpolarized PDFs

see R. Seidl talk this morning, too



Polarized PDFs and TMDs

see B. Kriesten talk yesterday, too



GPDs and form factors

Consistently growing move towards (pre) Monte Carlo and Machine Learning techniques

Uncertainty prescription (tolerance) in different groups

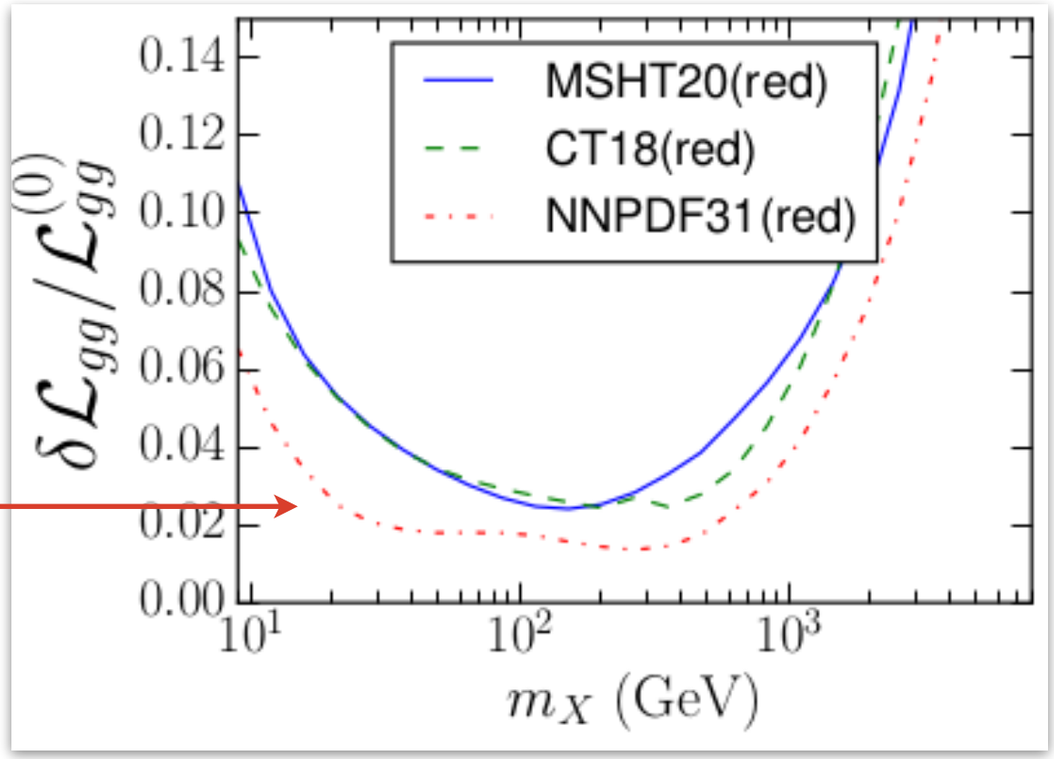
PDF4LHC21 benchmarking exercise:
comparison of uncertainties for same sets of data and QCD settings.

The uncertainties for CT18 (Hessian), MSHT20 (Hessian) and NNPDF3.1 (MC) reduced sets are still different.

Key role played by methodology.

NNPDF31(red) smaller by a factor of ~2

Monte-Carlo global analyses seem to lead to smaller uncertainties *wrt* Hessian fits.
Trend sustained by NNPDF4.0 set.



PDF4LHC21 [J.Phys.G 49]

The tolerance puzzle and the big-data paradox

Outside of HEP and NP, there is significant interest in statistical problems that are similar to the PDF tolerance problem. These studies introduce a fundamental distinction between the **fitting uncertainty** and **sampling uncertainty**, often overlooked in the PDF fits.

Article

Unrepresentative big surveys significantly overestimated US vaccine uptake

Nature **v. 600** (2021) 695

https://doi.org/10.1038/s41586-021-04198-4
Received: 19 June 2021

Valerie C. Bradley[1,6], Shiro Kuriwaki[2,6], Michael Isakov[3], Dino Sejdinovic[1], Xiao-Li Meng[4] & Seth Flaxman[5,✉]

SCIENCE ADVANCES | RESEARCH ARTICLE

MATHEMATICS

Models with higher effective dimensions tend to produce more uncertain estimates

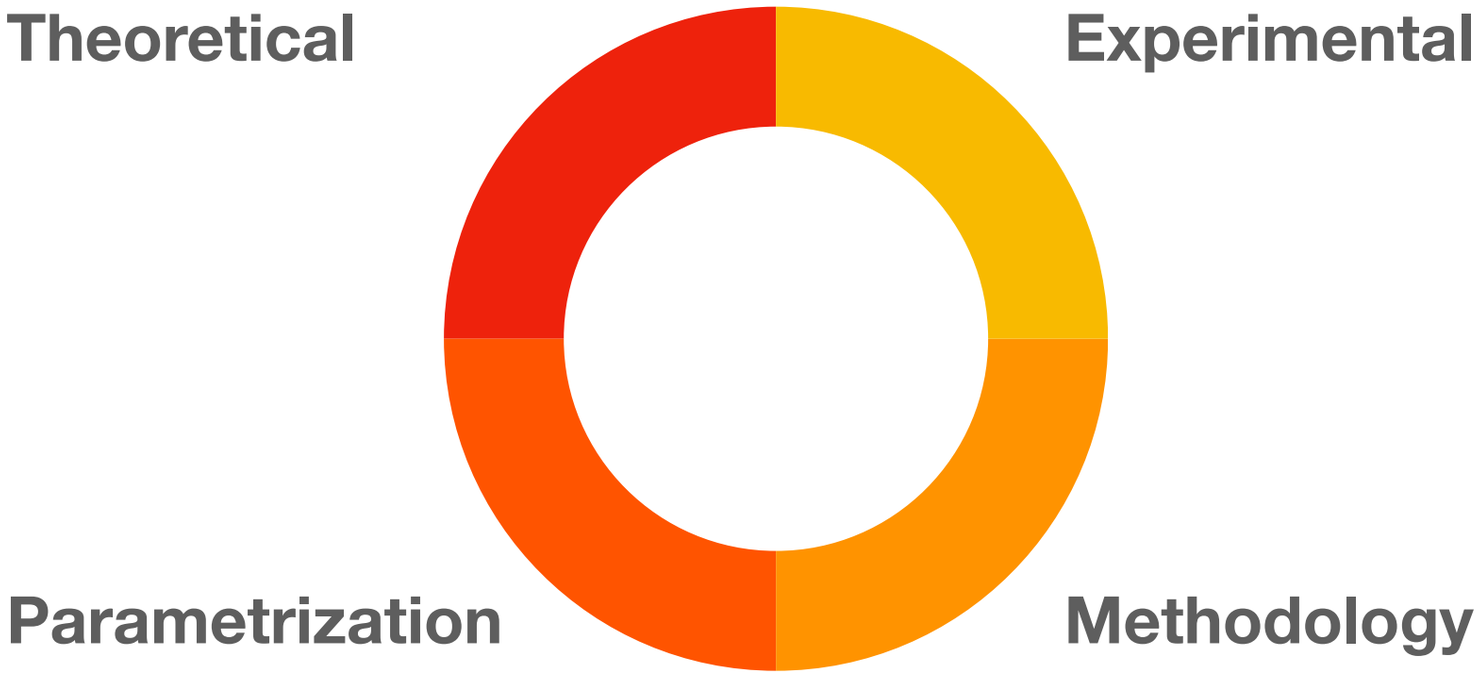
Arnald Puy[1,2,3]*, Pierfrancesco Beneventano[4], Simon A. Levin[2], Samuele Lo Piano[5], Tommaso Portaluri[6], Andrea Saltelli[3,7]

The Big Data Paradox in Clinical Practice

Pavlos Msaouel

To cite this article: Pavlos Msaouel (2022) The Big Data Paradox in Clinical Practice, Cancer Investigation, 40:7, 567-576, DOI: 10.1080/07357907.2022.2084621

A new avenue to understand PDF tolerance



In all four categories of uncertainties, we can further distinguish *PDF fitting accuracy* from *PDF sampling accuracy*.

Goodness-of-fit applies to an individual best fit.

[Kovarik et al, Rev.Mod.Phys. 92 (2020)]

Sampling accuracy applies either to the tolerance or the number of error sets in a PDF ensemble.

This talk.

Sampling accuracy vs. fitting accuracy

Differences of tolerance prescriptions are in part due to sampling conventions.
There is evidence that PDF fitting groups doing the same analysis arrive at different conclusions because of their tolerance criteria.

PDF sampling takes place over experimental data sets, parametrization forms, hyperparameters, settings of fits, model approximations.

Biases in sampling are particularly risky in large-scale analyses.

PDFs need representative sampling

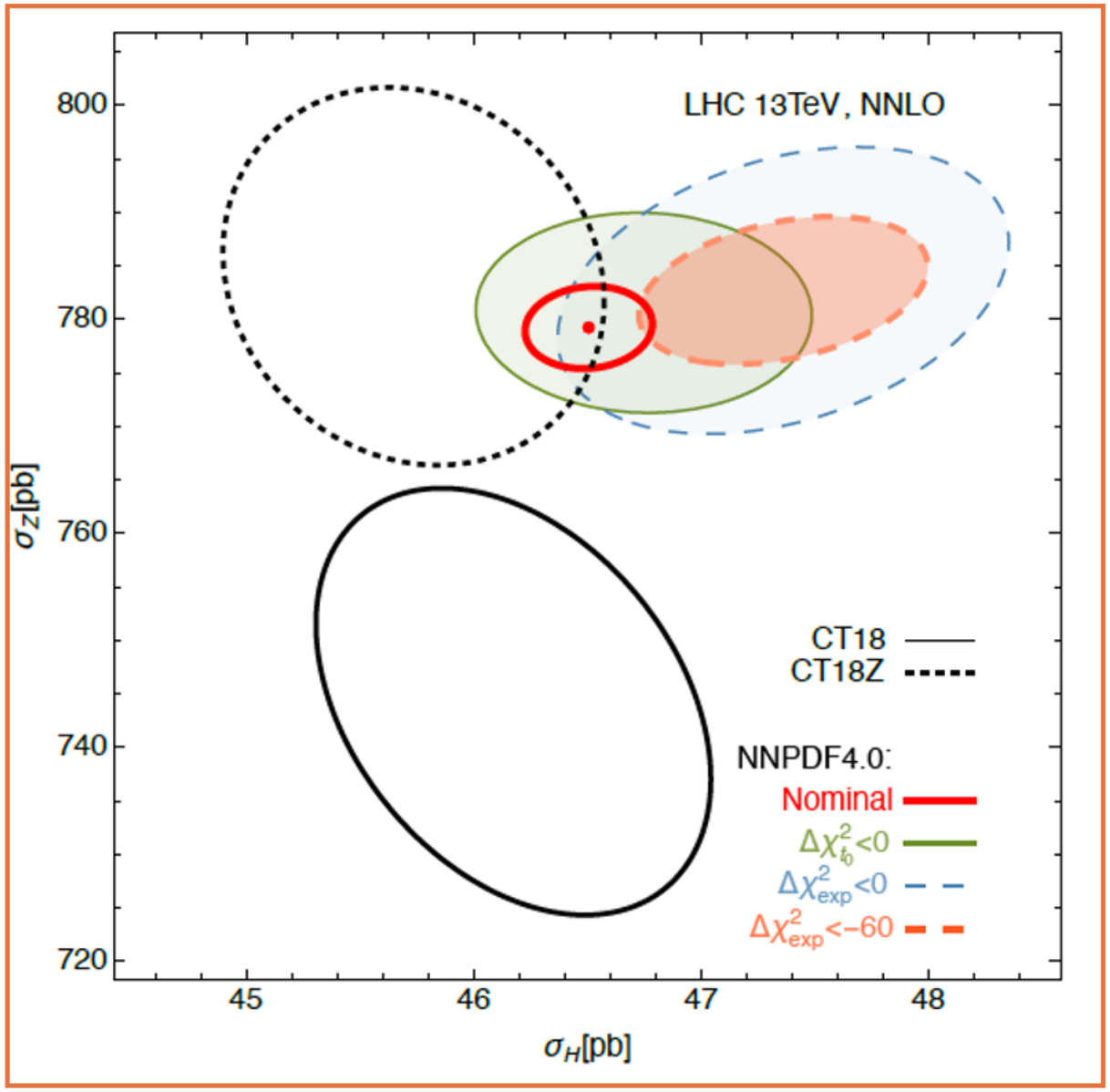
Sampling uncertainty can be estimated for specific predictions in targeted low-dimensional searches, called the **hopscotch scan,** that we designed in [2205.10444].

Black ellipses:
CT18(Z) correspond to a large tolerance (T~6) that accounts for sampling biases.

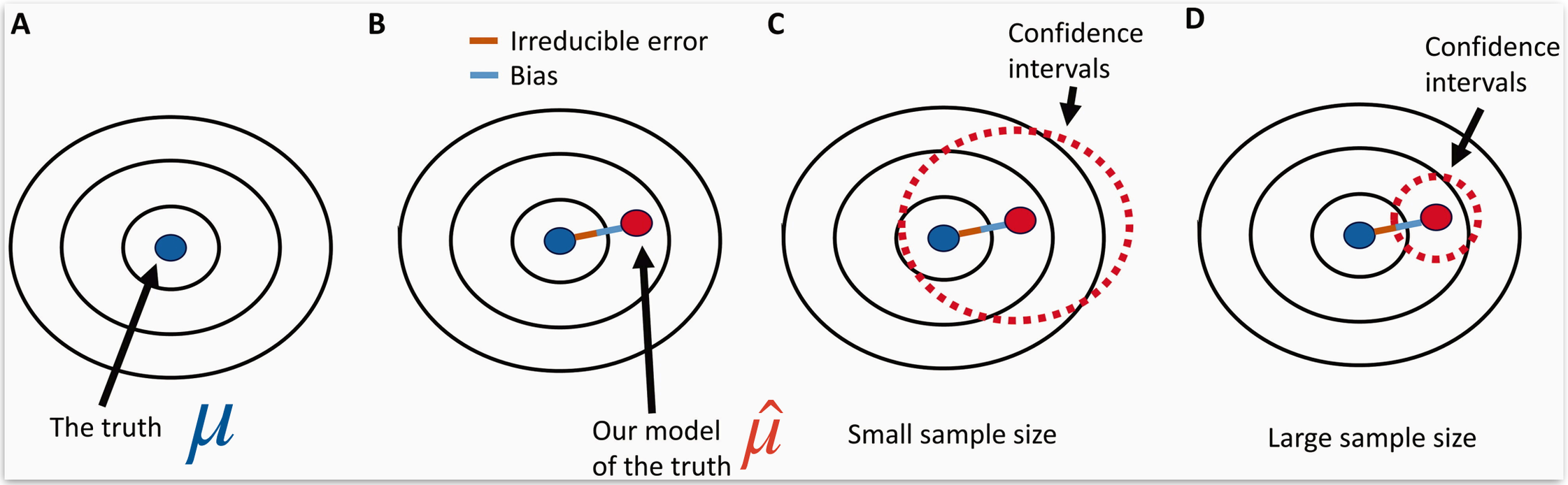
Red ellipse:
nominal NNPDF4.0

Filled color ellipses:

- areas of possible solutions corresponding to lower ($\Delta\chi^2 < 0$) w.r.t. the nominal NNPDF4.0 solution
- found through the hopscotch scan — a dimensionality reduction method.



From small to big data sets — sampling uncertainties



With an increasing size of sample $n \to \infty$, under a set of hypotheses, it is usually expected that the *deviation* on an observable decreases like $\left(\sqrt{n}\right)^{-1}$.

That's the law of large numbers.

What uncertainties keep us from including *the truth,* μ?

The law of large numbers disregards the *quality of the sampling*, Irreducible error, Bias.

Illustration from:
Pavlos Msaouel (2022)
Cancer Investigation, 40:7, 567-576

Trio identity

Xiao-Li Meng
The Annals of Applied Statistics
Vol. 12 (2018), p. 685

$$\mu - \hat{\mu} = (\text{data+sampling defect}) \times (\text{measure discrepancy}) \times (\text{inherent problem difficulty})$$

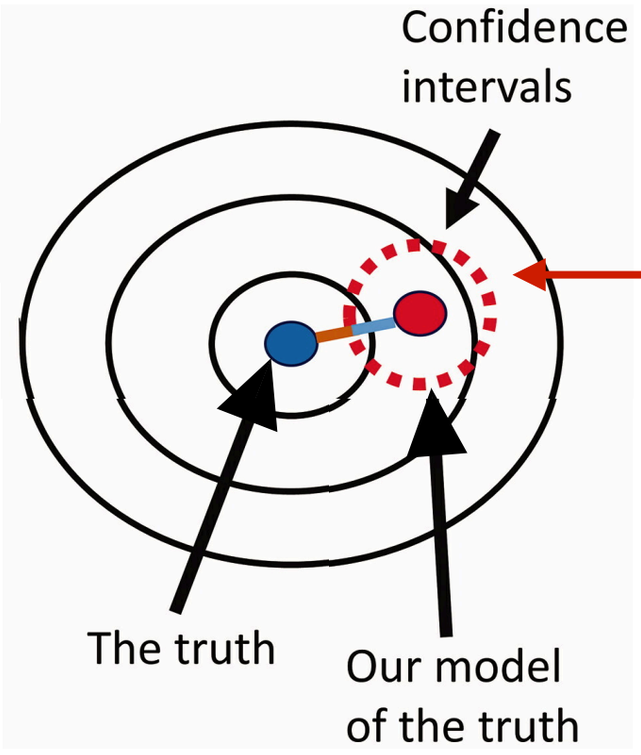
depends on the sampling algorithm

can tend to $\left(\sqrt{n}\right)^{-1}$ for random sampling

Irreducible error
Bias

≡ statistical model, quality of data,…

Large sample size



Conclusions

Assess the uncertainties in Hessian vs. Monte Carlo approaches for a more diverse and faithful understanding.

Study of the **sampling uncertainties** — a complementing source to the fitting uncertainty.

Highlights on the sampling uncertainties:

1. Tolerance criteria related to sampling choices. A PDF fit with few parameters and $\Delta\chi^2 = 1$ tolerance probably underestimates the parametric uncertainty. Large tolerance accounts for confounding sources.

2. Concept of effective large dimensions. Difficult to sample the full parameter space with many parameters without biases. A **hopscotch scan** intelligently reduces dimensionality of the relevant PDF parameter space for an observable under consideration.

3. Validating the final PDFs may be easier than understanding the respective fitting algorithm. Hopscotch algorithm is a **test outside the fit** to verify the PDF uncertainty for a specific QCD cross section or observable.