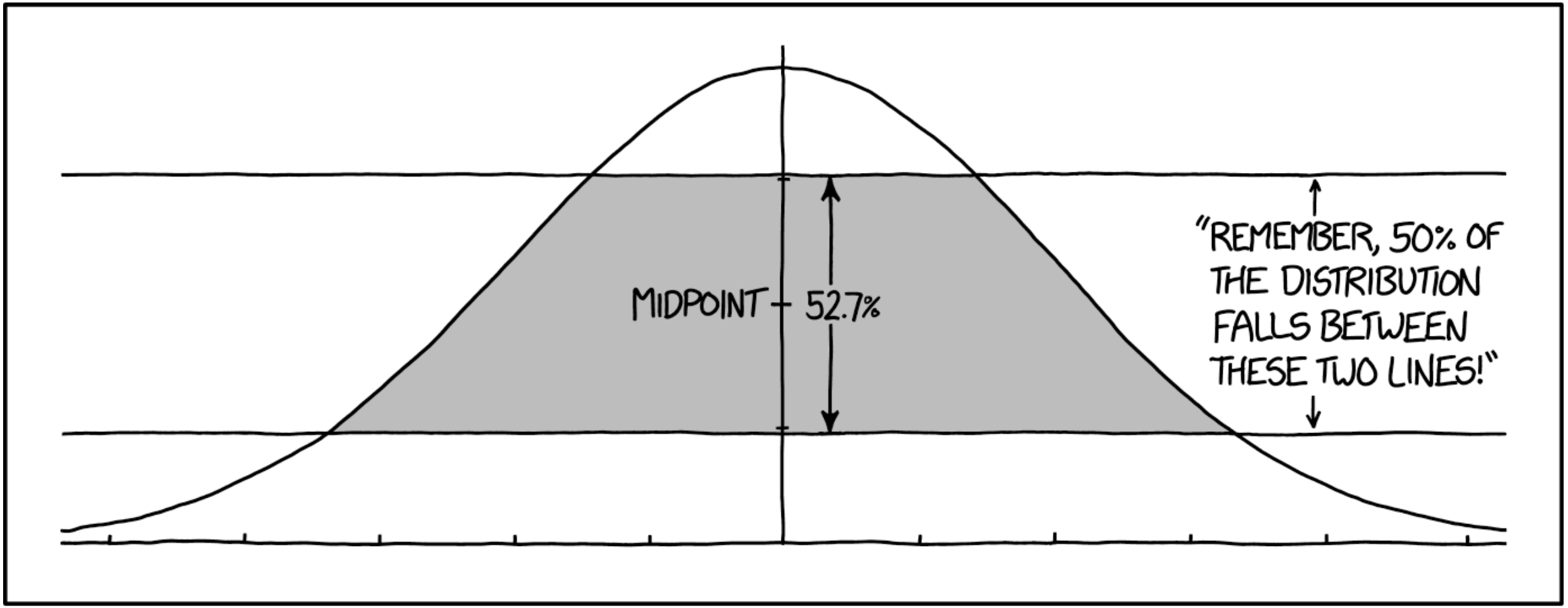


HOW TO ANNOY A STATISTICIAN

Image from https://xkcd.com/2118/

Douglas W. Higinbotham

Let’s Make A Deal

- Behind Two of the Doors Are Goats …
- Behind One of the Doors Is A Fabulous Prize!

Monty Hall Problem

Door #1

Door #2

Door #3

Monty Hall Problem

Door #1

Door #2

Door #3

Monty Hall Problem

Door #1

Door #2

Door #3

Monty Hall Problem

Door #1

Door #2

Door #3

Probability Says Switch

- Your human reaction is to stay with your original choice.
- I setup the problem a long time ago (i.e. you have 1/3 chance to pick correctly at the start)
- My opening a door is NOT random, so the remaining door has a 2/3 chance to be the winner!

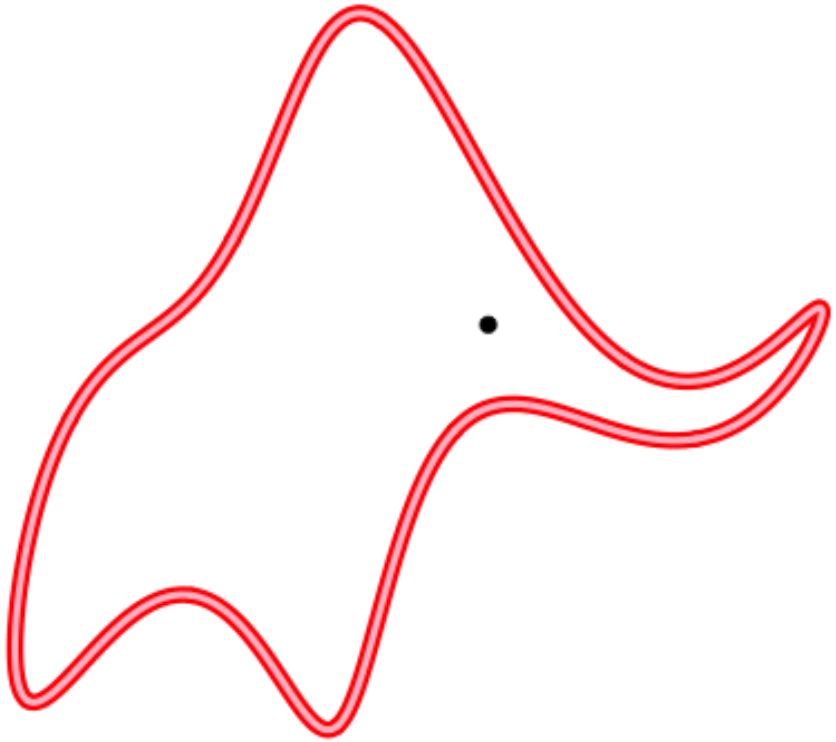
If you are still not convinced: https://en.wikipedia.org/wiki/Monty_Hall_problem

Fermi's Rejection of Dyson's Work

- https://www.webofstories.com/people/freeman.dyson/94?o=SH
- Also written as a Nature article: https://www.nature.com/articles/427297a

The Five Parameter Elephant

"Drawing an elephant with four complex parameters"
by Jurgen Mayer, Khaled Khairy, and Jonathon Howard, Am. J. Phys. 78 (2010) 648.

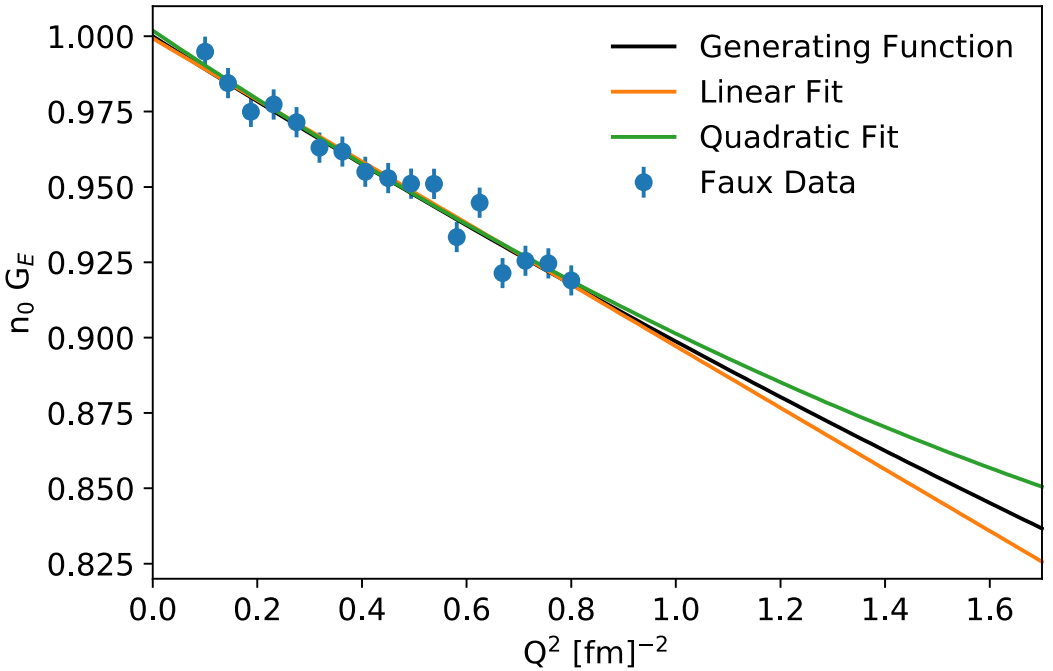


https://www.johndcook.com/blog/2011/06/21/how-to-fit-an-elephant/

Classic Example Monte Carlo Study

F. Borkowski *et al.*, Zeitschrift für Physik **275** (1976) 29.

- Use Standard Dipole For G_E
- Generate Random Data In A Simple Monte Carlo Code
  - 0.05 fm^{-2} Spacing
  - 0.005 Random Uncertainty
- With Several Different Ranges
  - 0.1 to 0.4 fm^{-2}
  - 0.1 to 0.8 fm^{-2}
  - 0.1 to 1.2 fm^{-2}
  - 0.1 to 1.6 fm^{-2}
- Fit Each Set and Range With
  - $f(q^2) = a_0 + a_1 q^2$ (linear)
  - $f(q^2) = a_0 + a_1 q^2 + a_2 q^4$ (quadratic)
- Repeat one millions times.
- Shown is one outcome of generating the pseudo data and doing the fits.



Radius of Standard Dipole is 0.81fm

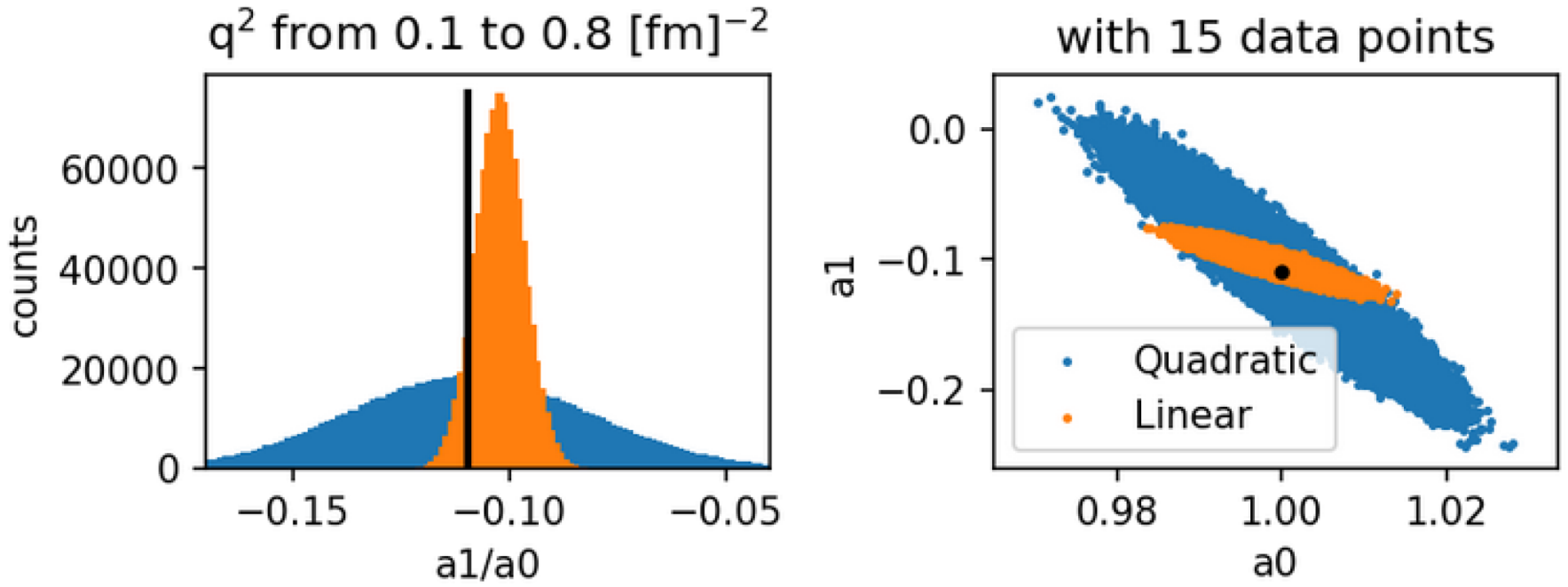
Table regenerated using Python is the same as the Z. Physik.

interval fm^{-2}	linear fit		quadratic fit	
	a_0	radius	a_0	radius
$0.1 - 0.4$	1.000	0.79	1.000	0.81
$0.1 - 0.8$	0.999	0.78	1.000	0.81
$0.1 - 1.2$	0.997	0.77	1.000	0.81
$0.1 - 1.6$	0.996	0.76	1.000	0.81

Clearly the linear fits are biased, so the authors of the paper use quadratic fits along with three floating normalization parameters to got a proton radius of 0.87(2) fm.

Visualization of the Results

Each count/dot is an one outcome from the millions of Monte Carlo fits.

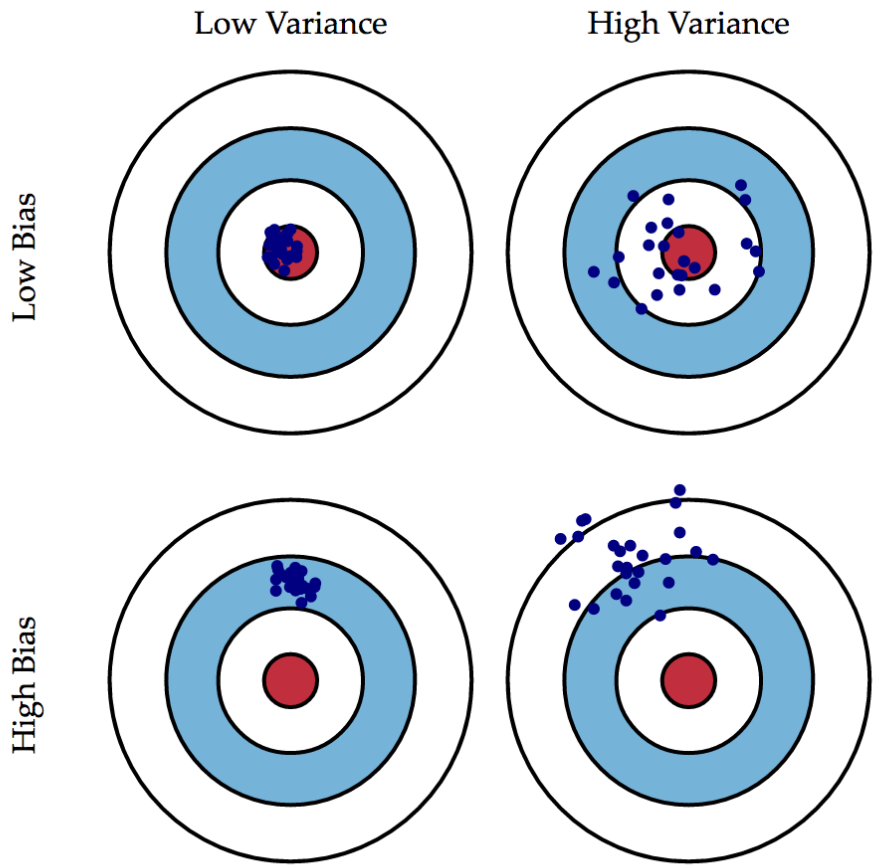


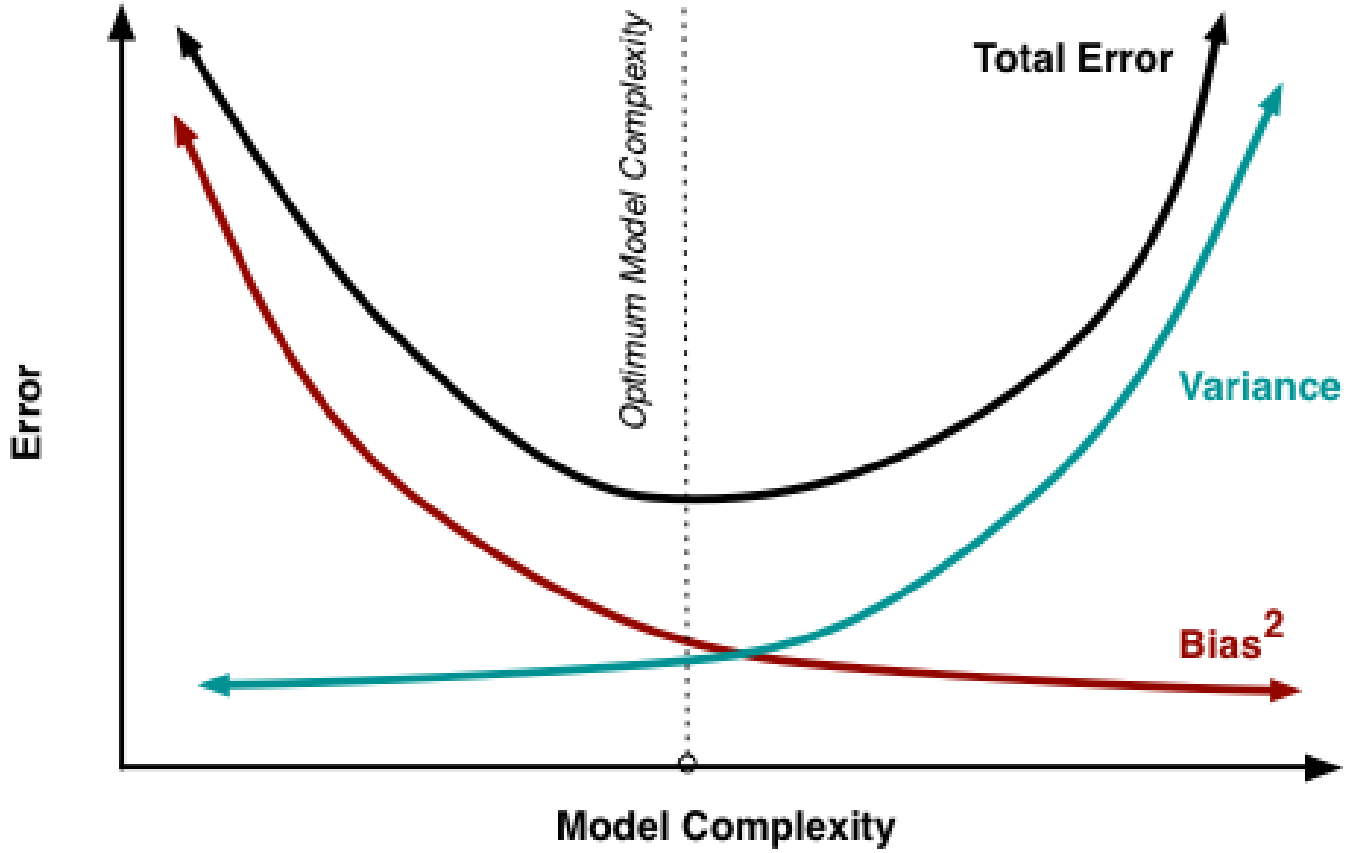
While it is true the linear model has a **high bias**, the quadratic model has an **high variance**.

We do the experiment ONCE not the average of millions of tries!

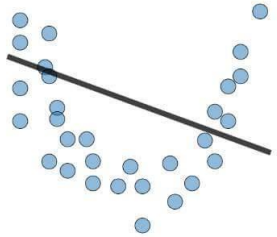
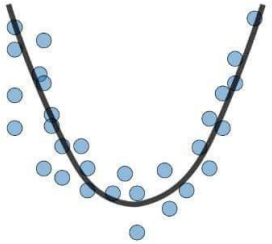
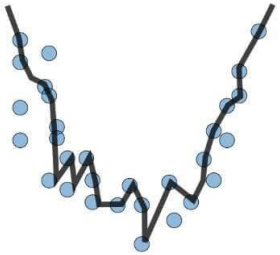
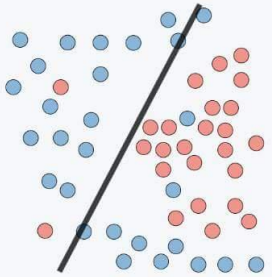
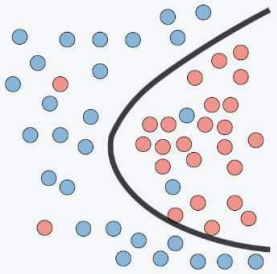
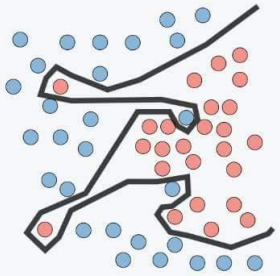
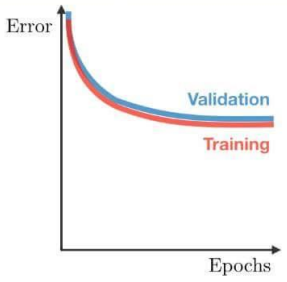
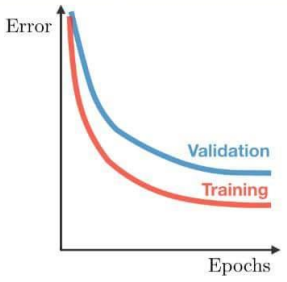
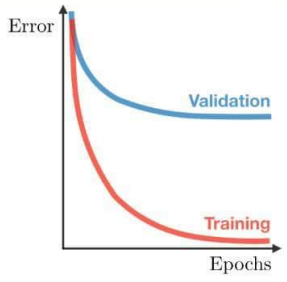
Also, here we know the bias only because we know the true function, which is not typical true for a real experiment.

Bias vs. Variance

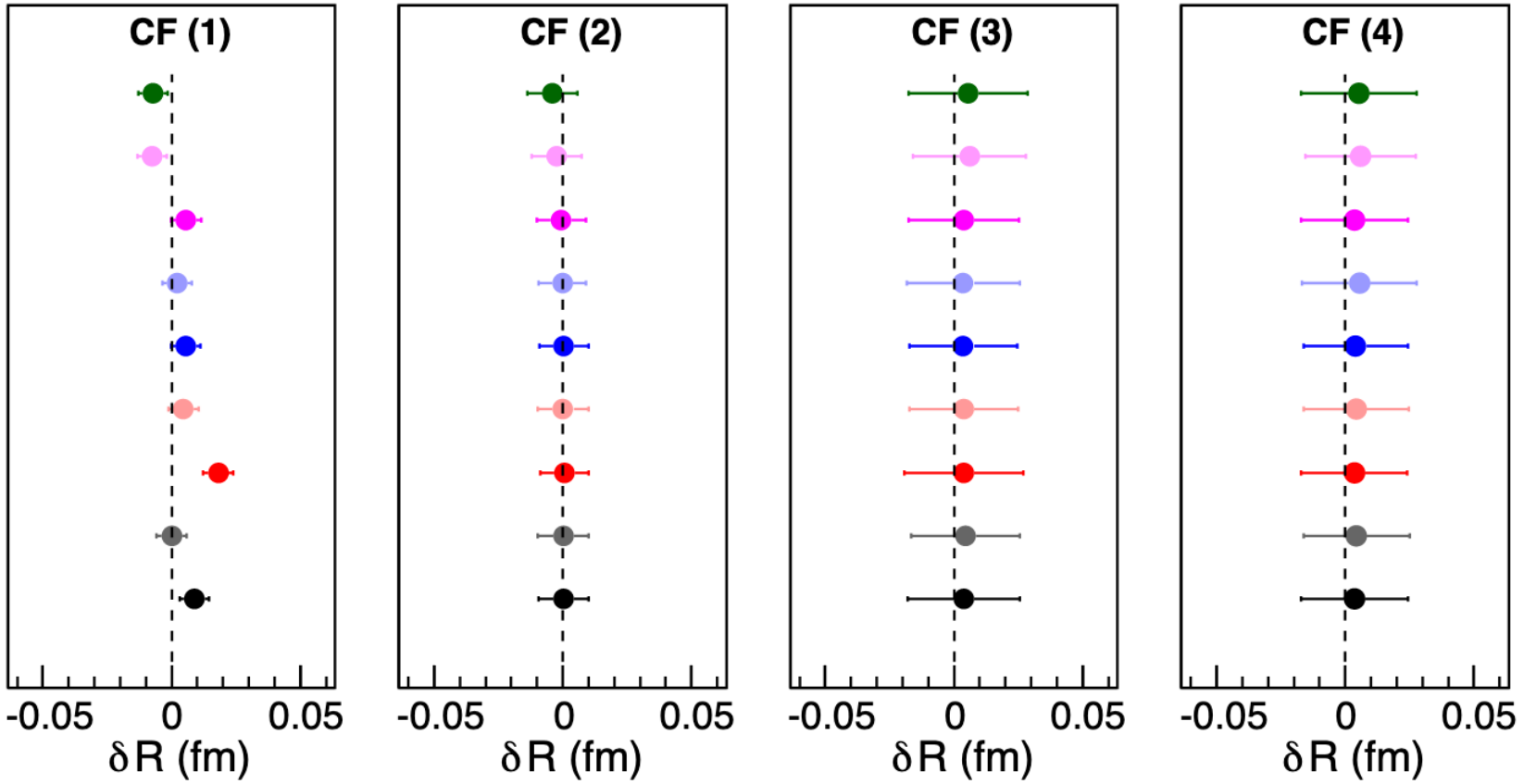




Important Consept In Machine Learning Classes

	Underfitting	Just right	Overfitting
Symptoms	• High training error • Training error close to test error • High bias	• Training error slightly lower than test error	• Very low training error • Training error much lower than test error • High variance
Regression illustration			
Classification illustration			
Deep learning illustration			
Possible remedies	• Complexify model • Add more features • Train longer		• Perform regularization • Get more data

Bias-Variance Trade-off For Expected PRad Data



Ye-2018
Bernauer-2014
Alarcón-2017
Arrington-2007
Arrington-2004
Kelly-2004
Gaussian
Monopole
Dipole

$$f_{\mathrm{CF}}(Q^2) = p_0 G_E(Q^2) = p_0 \frac{1}{1 + \frac{p_1 Q^2}{1 + \frac{p_2 Q^2}{1+\cdots}}}$$

Results from millions of simulations of the expected statistical results. We will do the experiment once.

NOTE: As Ingo Sick correctly pointed out, CF2 and Rational N=M=1 are a special case and are equivalent.

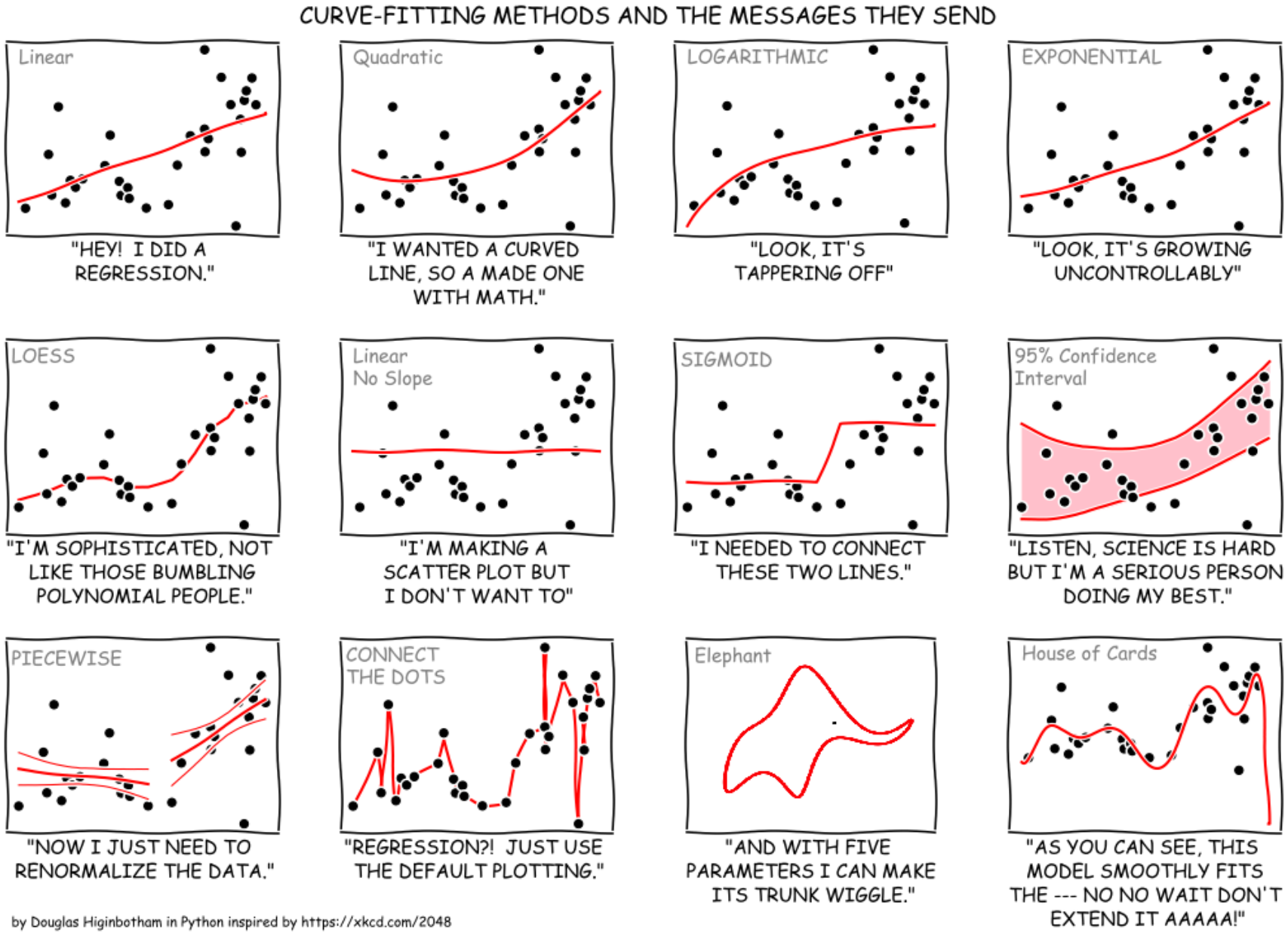
Robust Analysis For Expected Compass Data

- All that is needed is your expected binning and uncertainties
- Code from Alarcon and Weiss should be ready for public use next week
  - then you can try any radius you like
  - **Perfect for practicing blinding pulling out the radius from SINGLE sets of data**
- The PRad C++ code is on github: https://github.com/saberbud/Proton_radius_fit_class

Further Reading & Source For Python & R Software

- To Explain or Predict? (Descriptive, Predictive and Explanatory Modeling)
  - https://projecteuclid.org/euclid.ss/1294167961
  - Beautiful example of the true function NOT always being the best predictive model data.
- **The Interpretation of Errors – Fredrick James (most of you use MINUIT, so please read!)**
  - **http://seal.cern.ch/documents/minuit/mnerror.pdf**
- Data Analysis Textbooks
  - Data Reduction and Error Analysis – Philip Bevington
  - Statistical Methods in Experimental Physics – Fredrick James
  - Computation Methods in Physics – Simon Širca
  - Probability for Physicists – Simon Širca (https://www.springer.com/us/book/9783319316093)
- Anaconda Open Source Software (https://www.anaconda.com/download)
- And one of my favorites: https://en.wikipedia.org/wiki/How_to_Lie_with_Statistics
  - This book shows you the lies
  - Though it is old, you can see example after example that is discussed in this book on social media sites!

And just for fitting fun!



https://twitter.com/dwhiginbotham/status/1043837703991832577
and the source code: http://doi.org/10.5281/zenodo.1436555