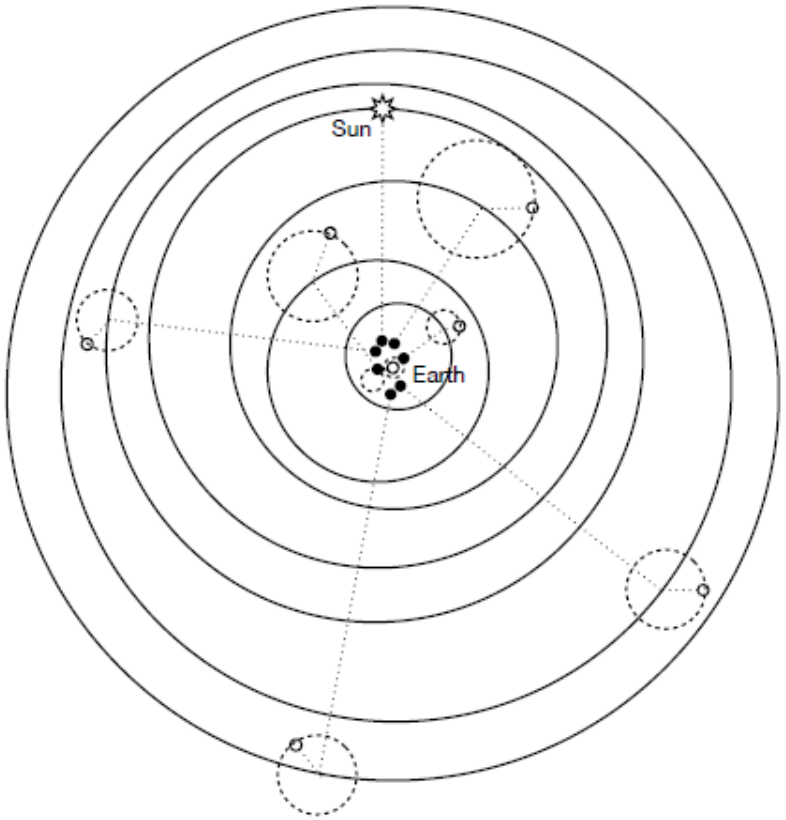
Evaluación de modelos

Dr. Héctor Nájera
Dr. Curtis Huffman

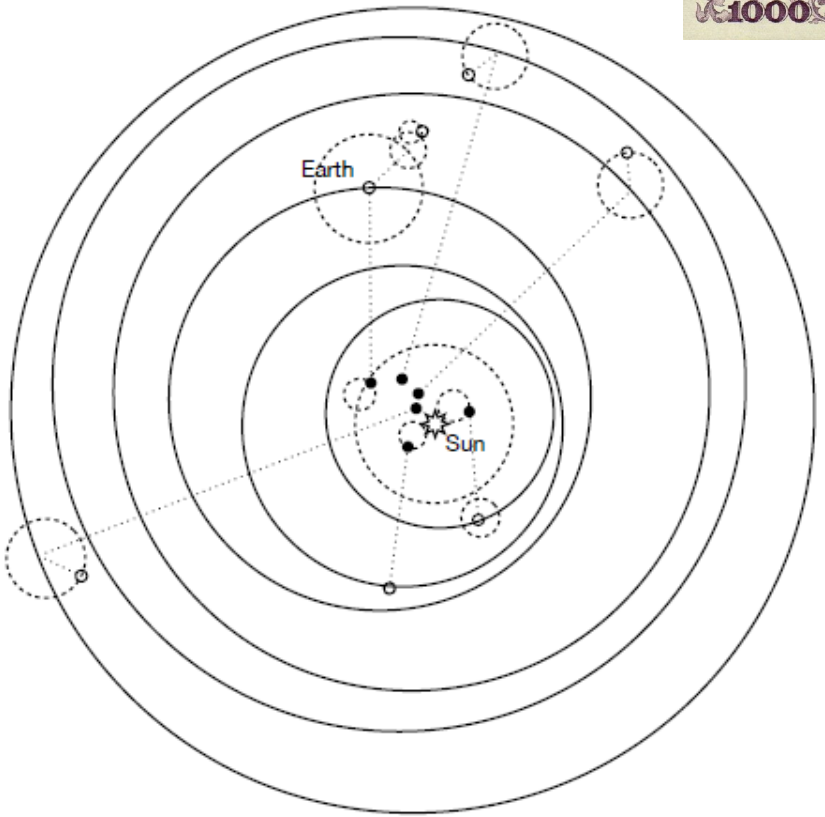
Accuracy v Simplicity

Accuracy v Simplicity

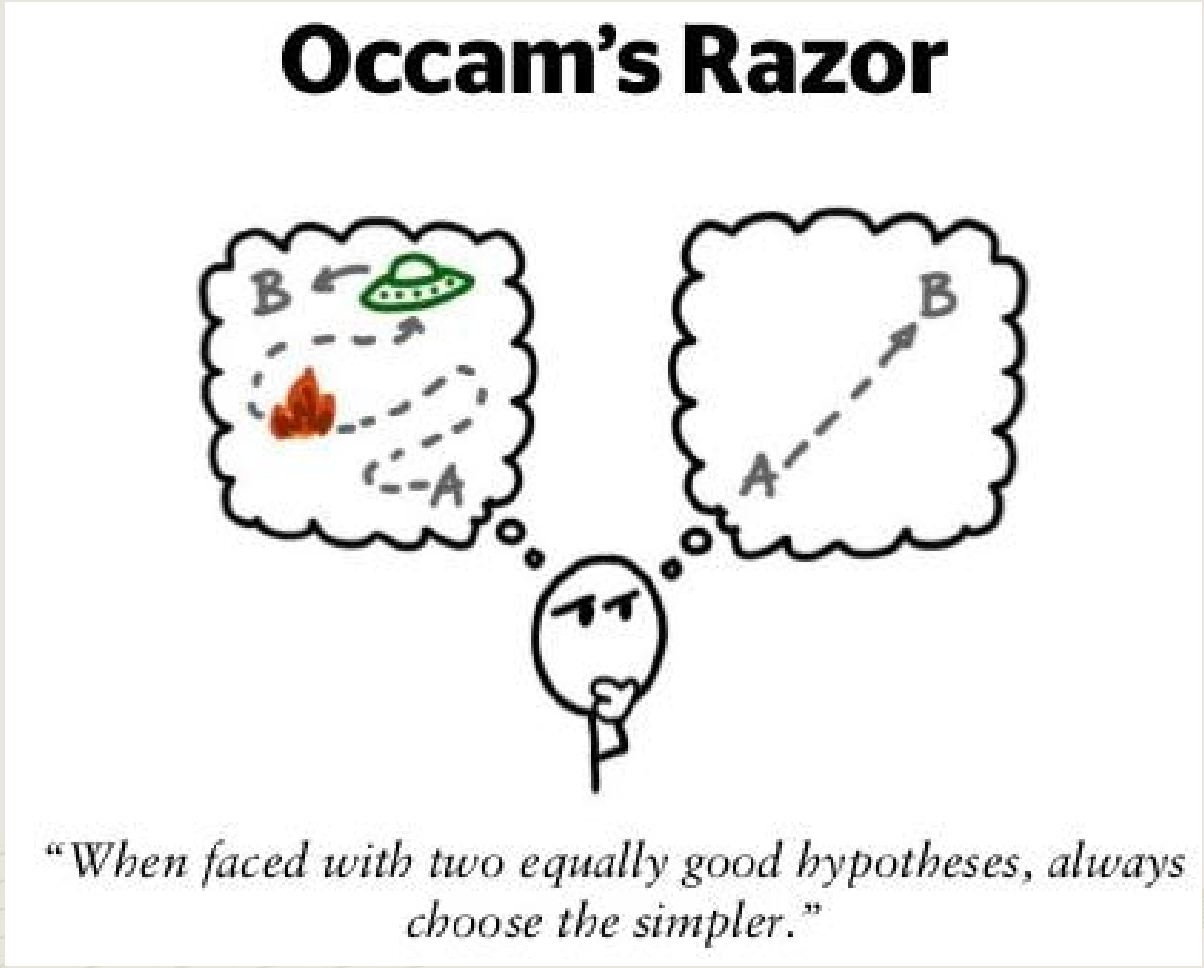
Ptolemaic Model



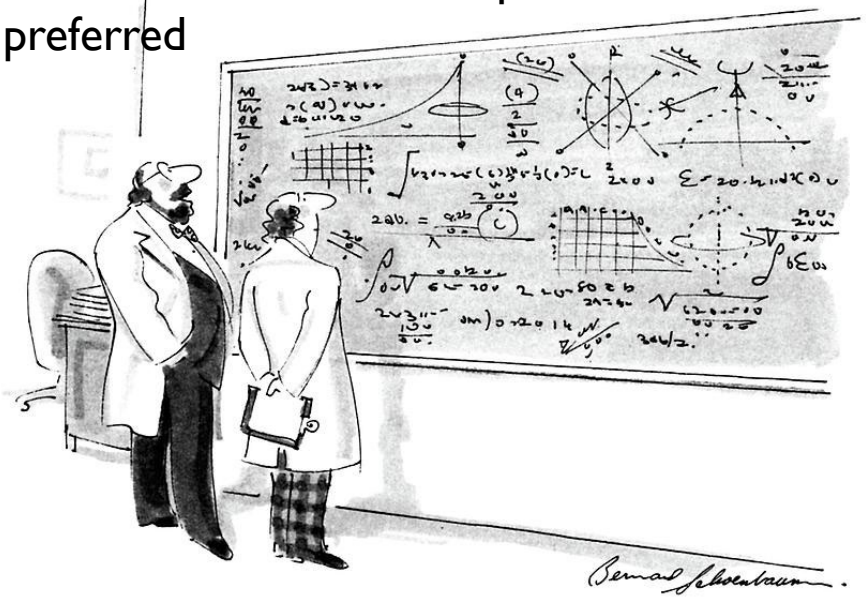
Copernican Model



Accuracy v Simplicity



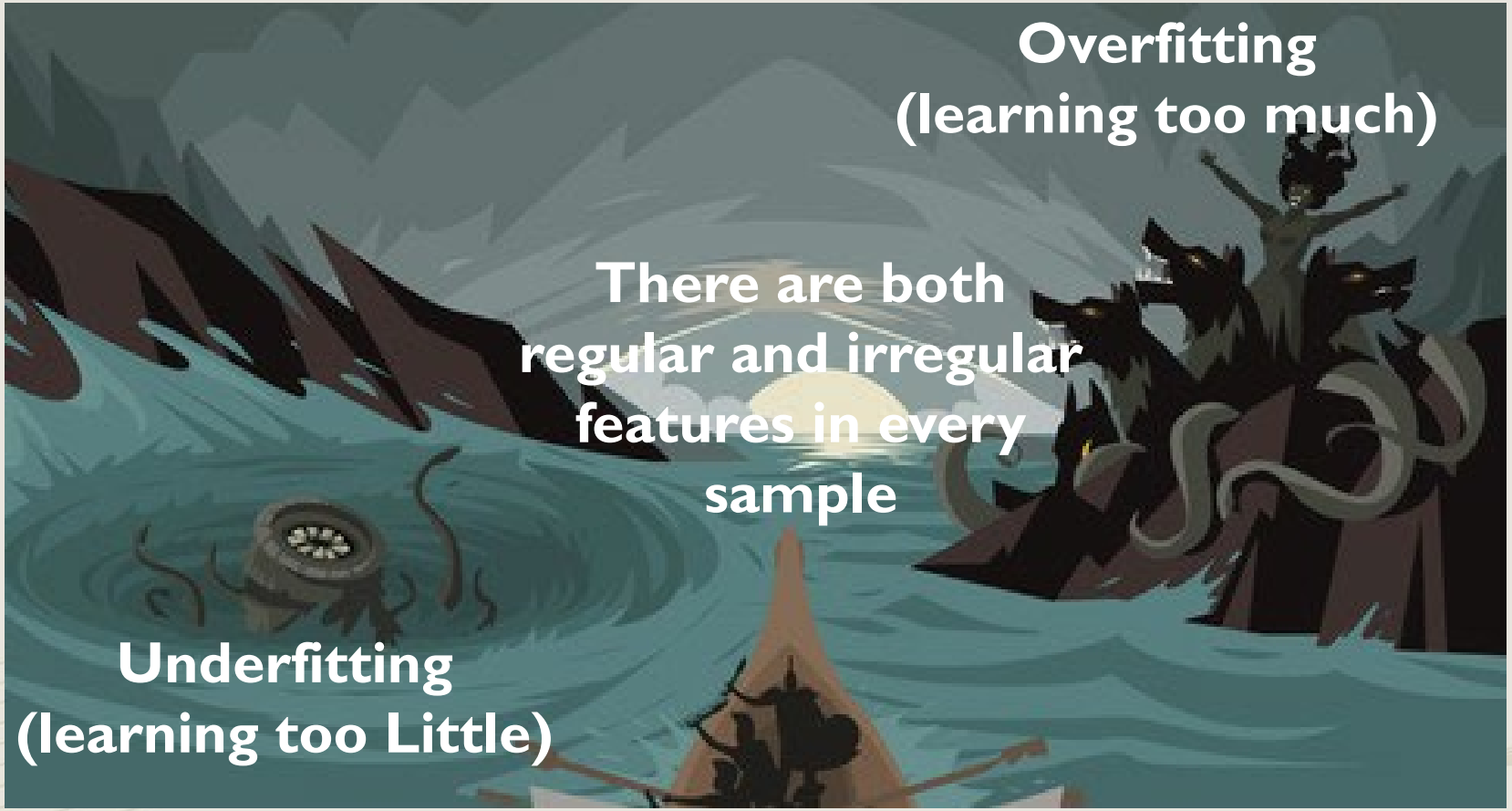
Models with fewer assumptions are to be preferred

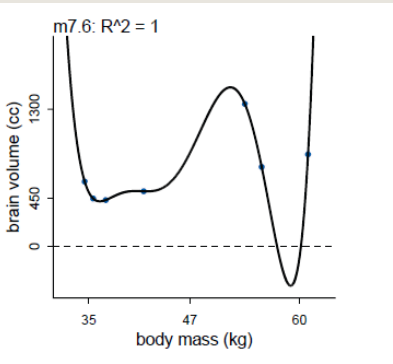


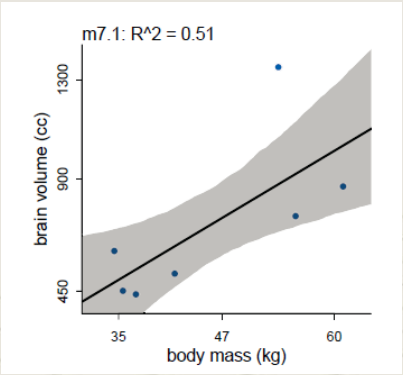
"Oh, if only it were so simple."

Accuracy v Simplicity

Making the model more complex (adding parameters) nearly always improves the fit of a model to the data (a measure of how well the model can retrodict the data used to fit the model), at the cost of predict new data worse.



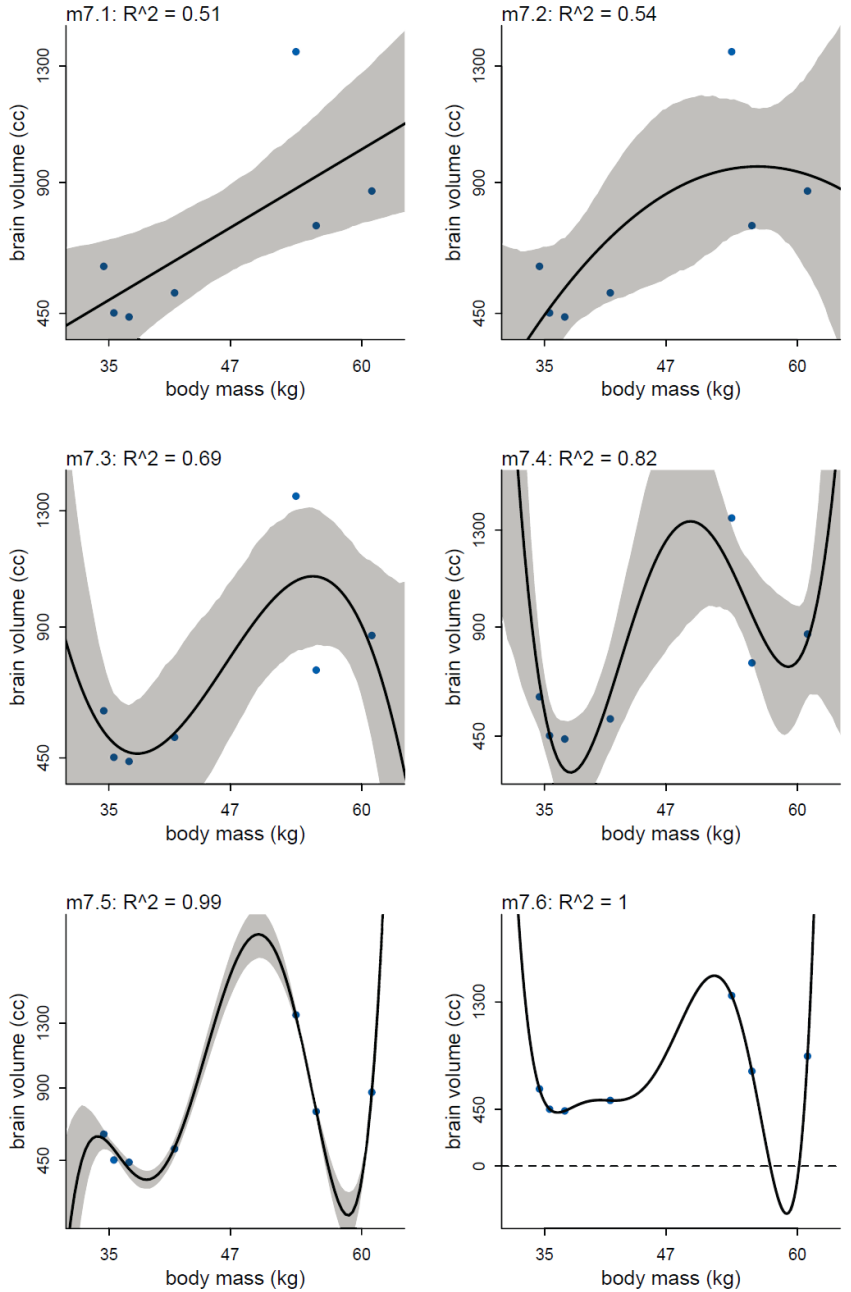




The proportion of variance "explained" by the model

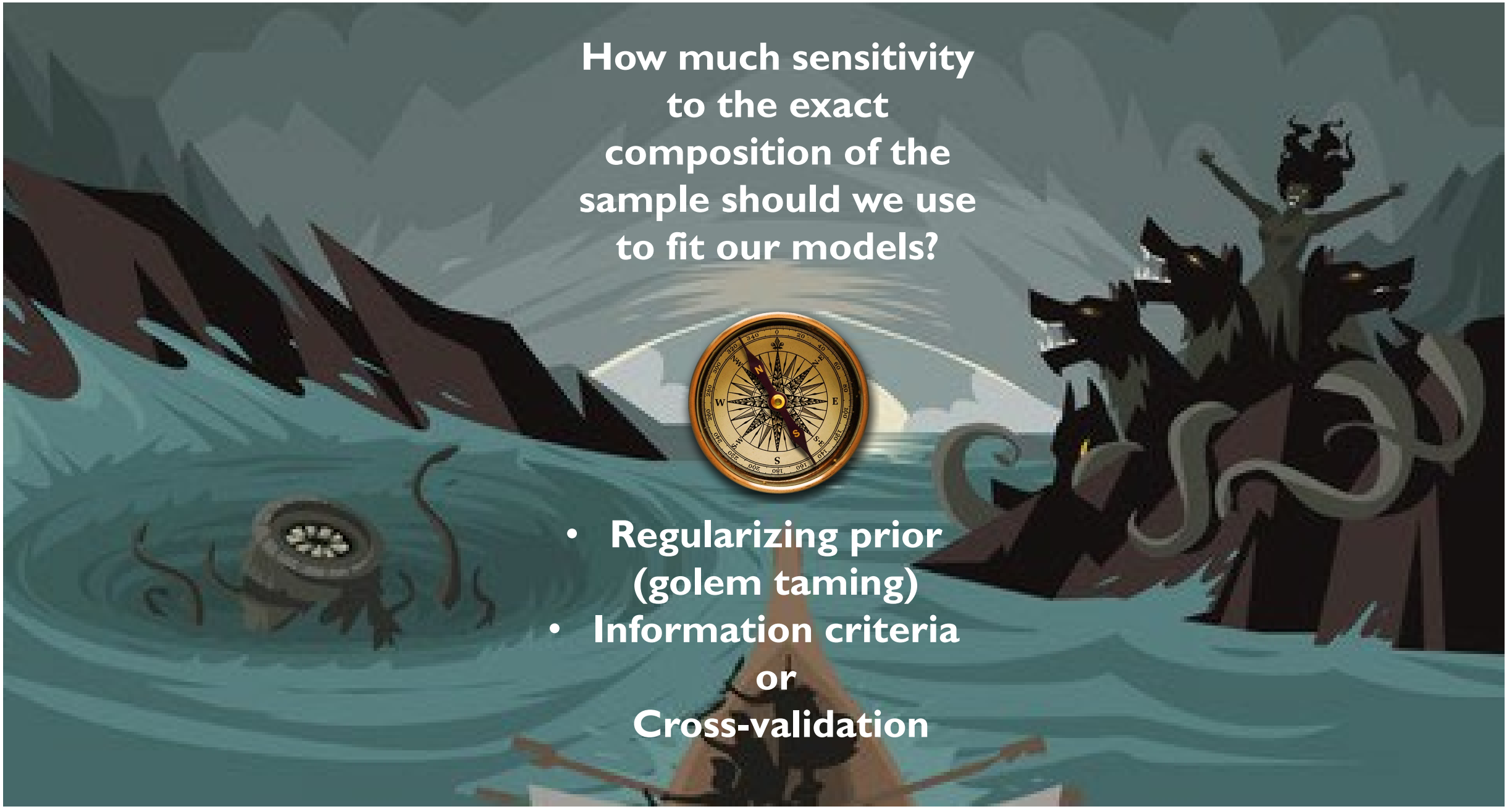
Too few parameters hurts, too

Stargazing doesn't help (***), the conventional 5% threshold is purely conventional, we shouldn't expect to optimize anything.

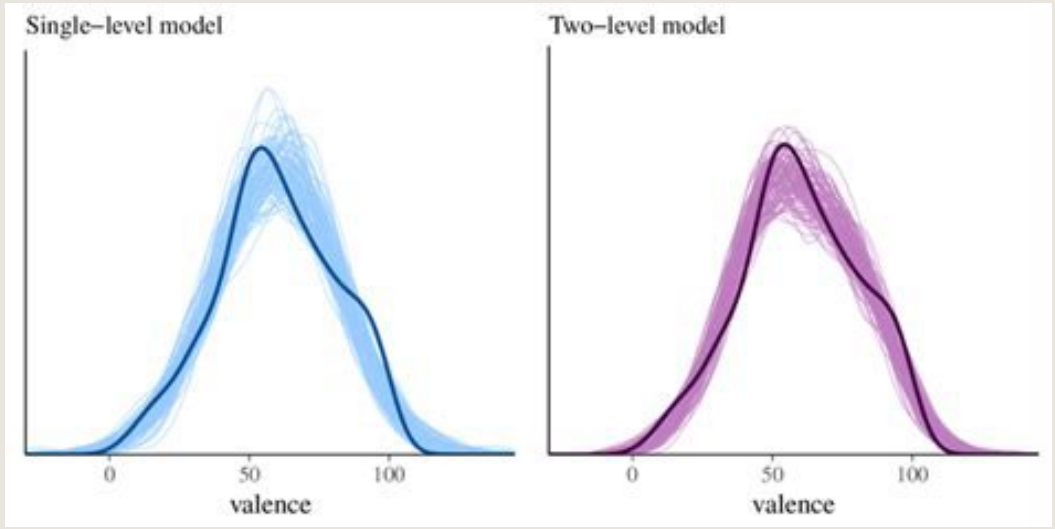


Encoding the sample is never the goal, we are not in the business of retrodiction. We don't care for irregularities.

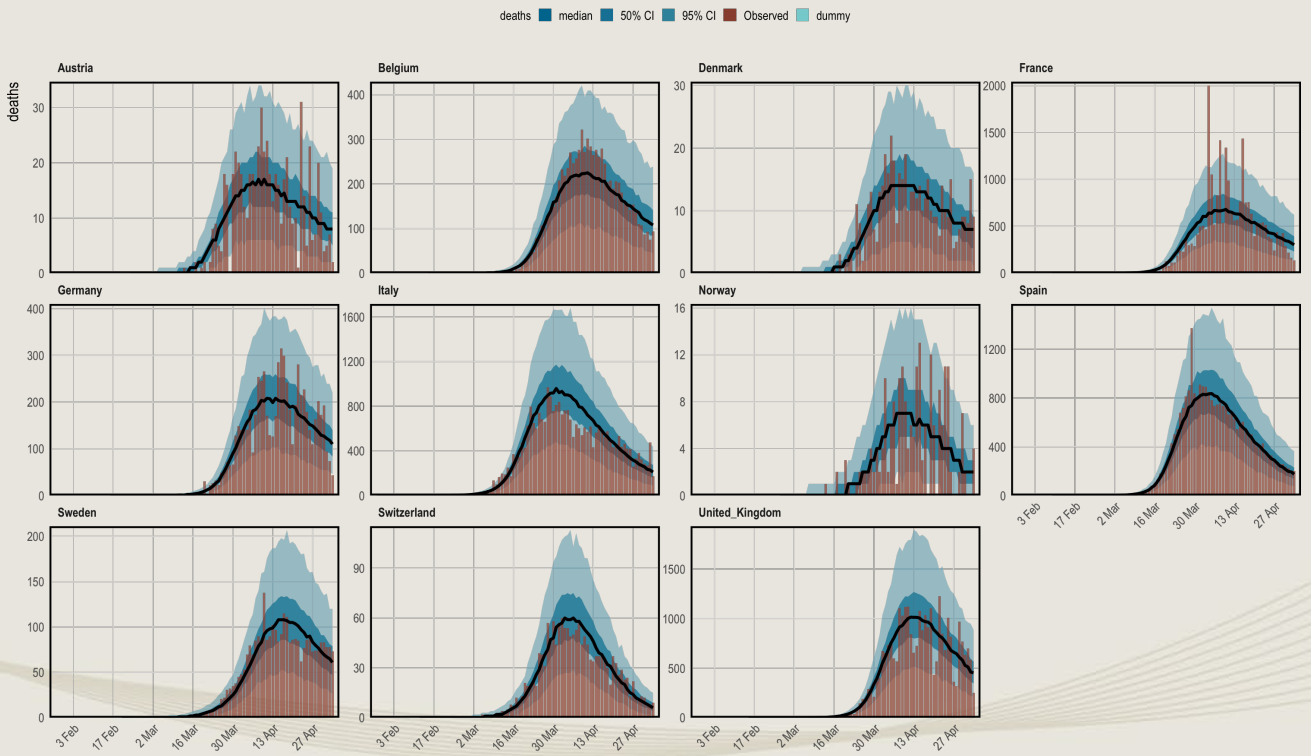
The fit is perfect, but the model is ridiculus


How much sensitivity
to the exact
composition of the
sample should we use
to fit our models?
• Regularizing prior
(golem taming)
• Information criteria
or
Cross-validation

Within-sample v Out-of-sample

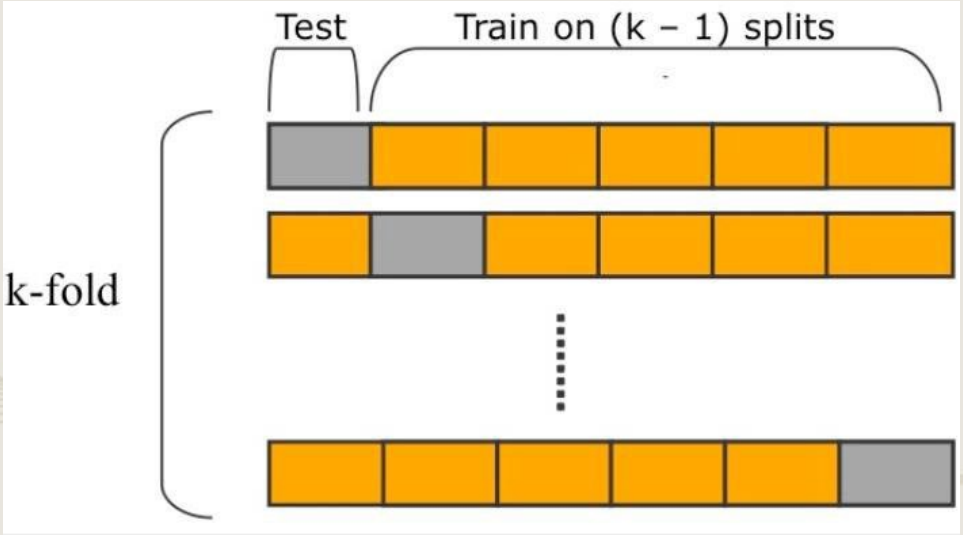


Predicción: Evaluación “formal” de un buen modelo (nunca within-sample)



Out-of-sample?

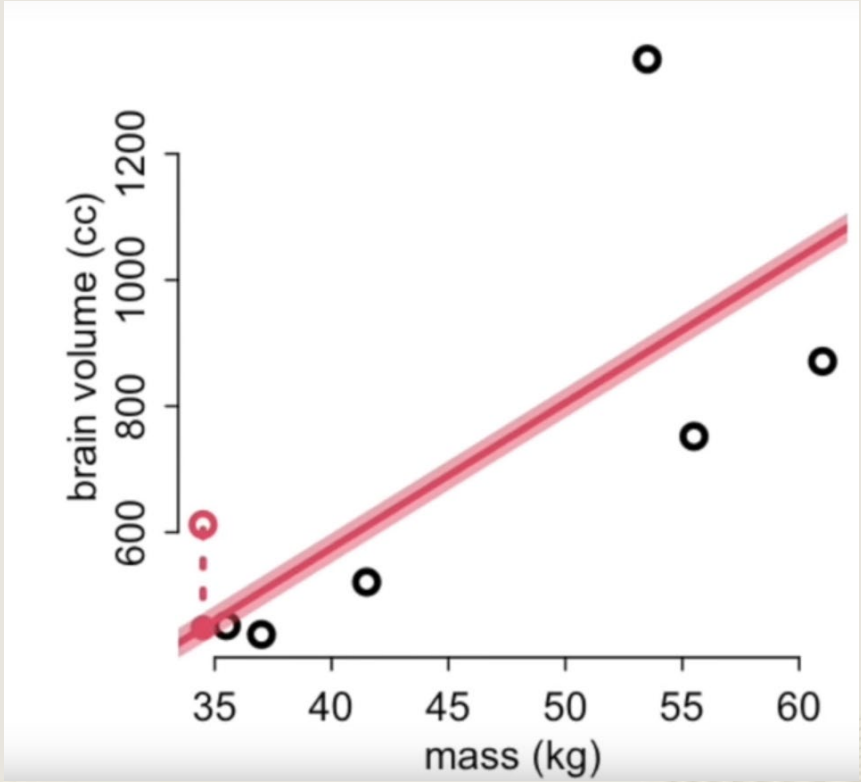
- How to evaluate your models out-of-sample?
  - Divide the sample in a number of chunks (folds). Predict each fold, after training on all the others.
    - How many chunks?
      - The maximum number of chunks (leave-one-out cross-validation)



Leave-one-out (loo) cross validation

¿Qué tan bueno es el modelo para predecir el siguiente punto –nuevas observaciones-?

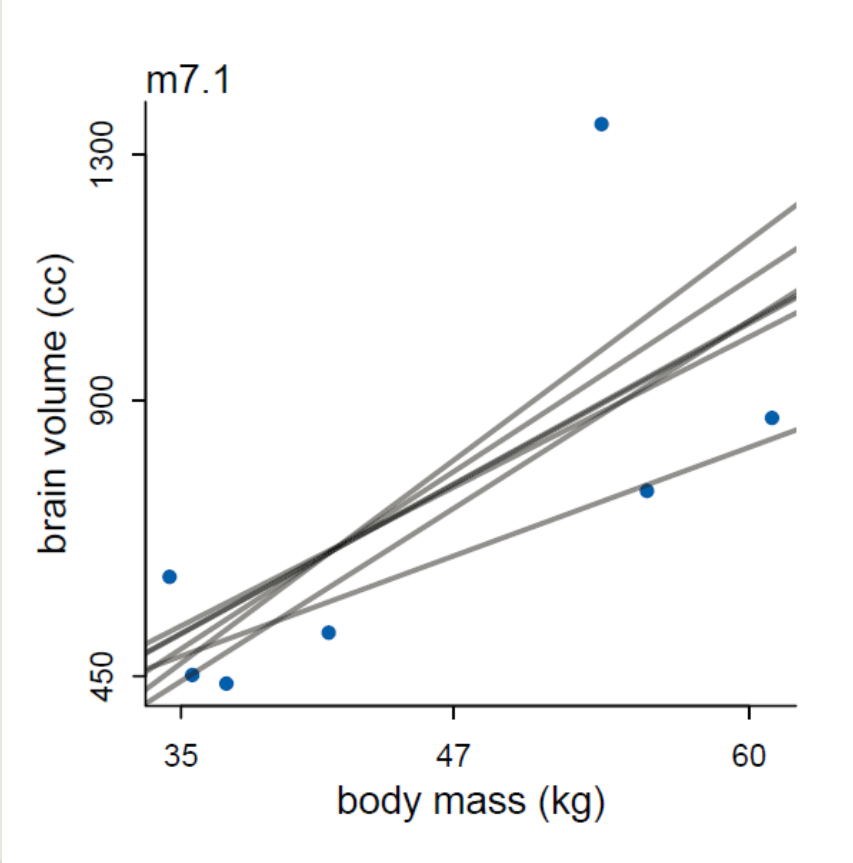
1. Tiramos un punto
2. Estimamos con el resto de la muestra
3. Predecimos el punto que tiramos -con 2-
4. Repetimos 1 con otro punto
5. Tomamos los errores de cada paso. MSE
   - Fuera
   - Dentro



Leave-one-out (loo) cross validation

¿Qué tan bueno es el modelo para predecir el siguiente punto –nuevas observaciones-?

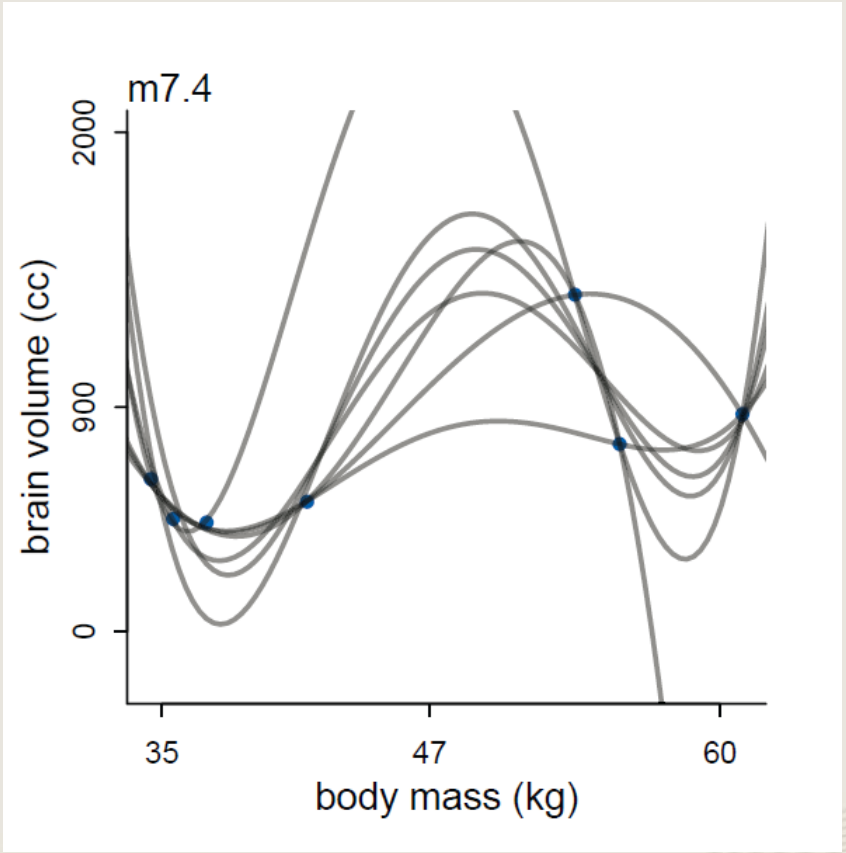
1. Tiramos un punto
2. Estimamos con el resto de la muestra
3. Predecimos el punto que tiramos -con 2-
4. Repetimos 1 con otro punto
5. Tomamos los errores de cada paso. MSE
   - Fuera
   - Dentro



Leave-one-out (loo) cross validation

¿Qué tan bueno es el modelo para predecir el siguiente punto –nuevas observaciones-?

1. Tiramos un punto
2. Estimamos con el resto de la muestra
3. Predecimos el punto que tiramos -con 2-
4. Repetimos 1 con otro punto
5. Tomamos los errores de cada paso. MSE
   - Fuera
   - Dentro



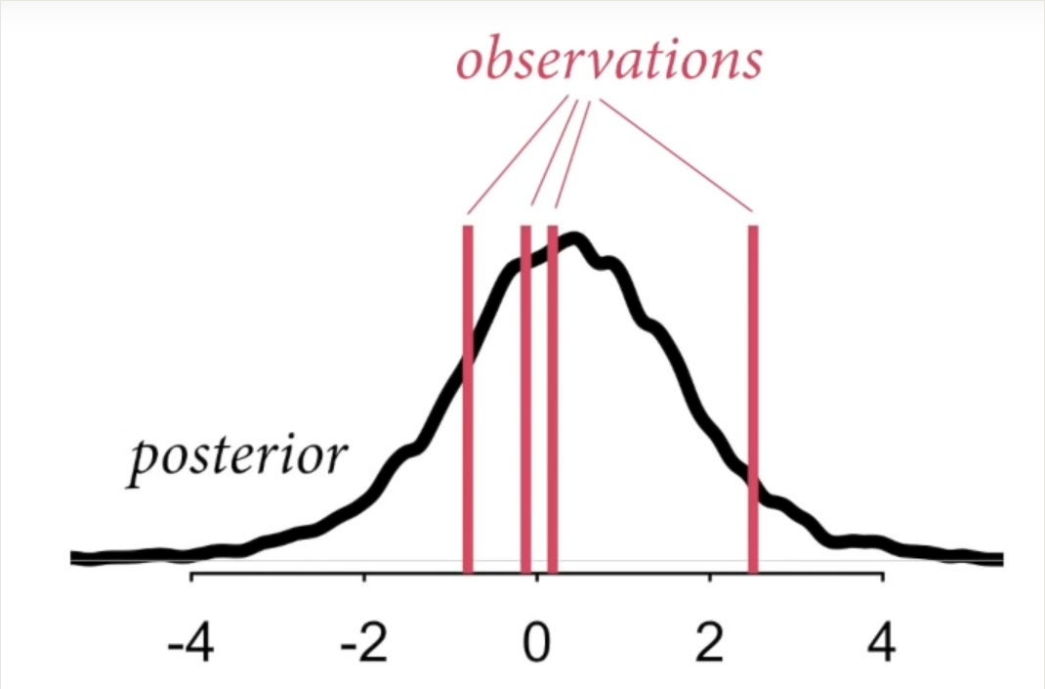
Bayesian loo and cross-validation

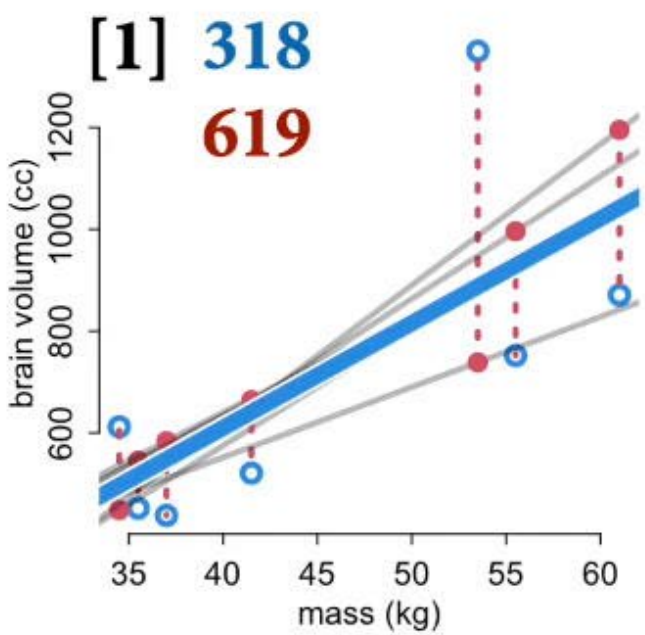
En Bayes no usamos puntos y una sola línea.

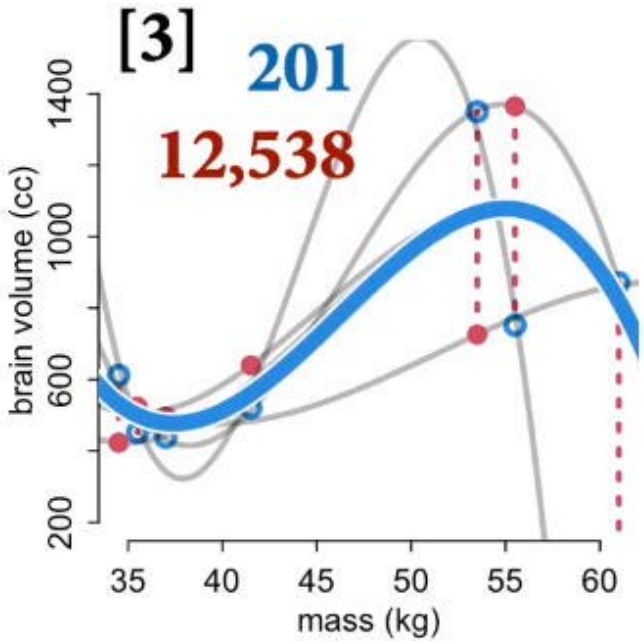
- Usamos toda la posterior, no punto a punto
- La probabilidad posterior de cada observación: la densidad predictiva de cada punto
- Obtenemos posteriores para cada conjunto menos un dato

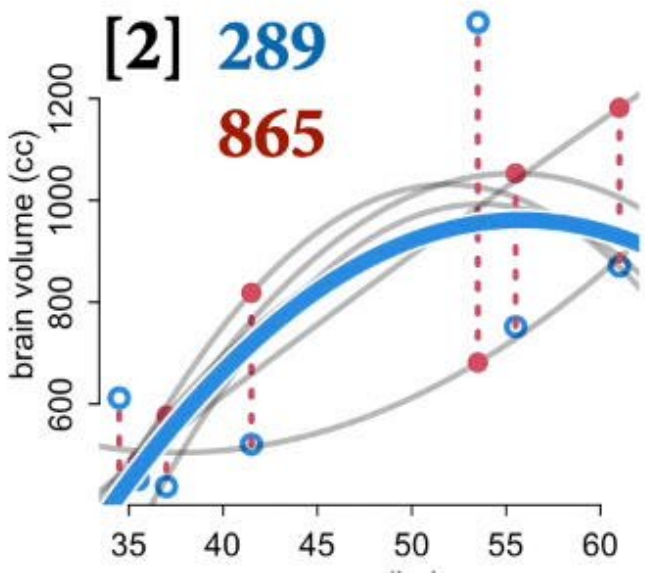
El cómputo de esto es, para todo efecto práctico, imposible!!

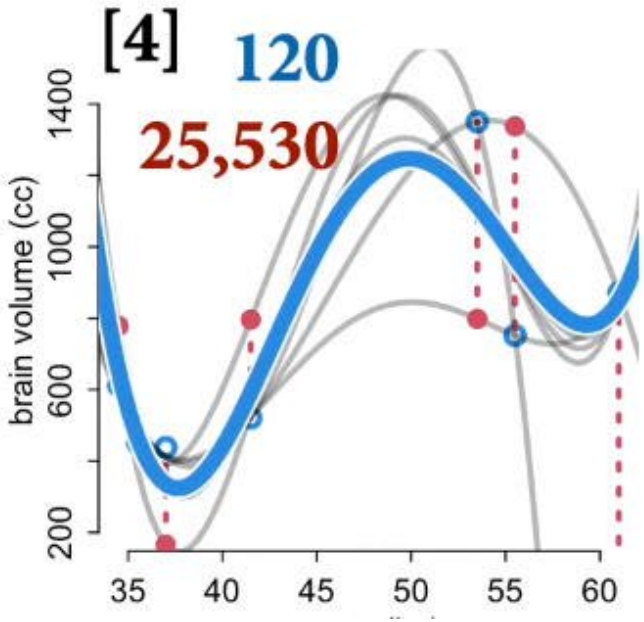
Pero si lo fuera ¿Cómo luciría?

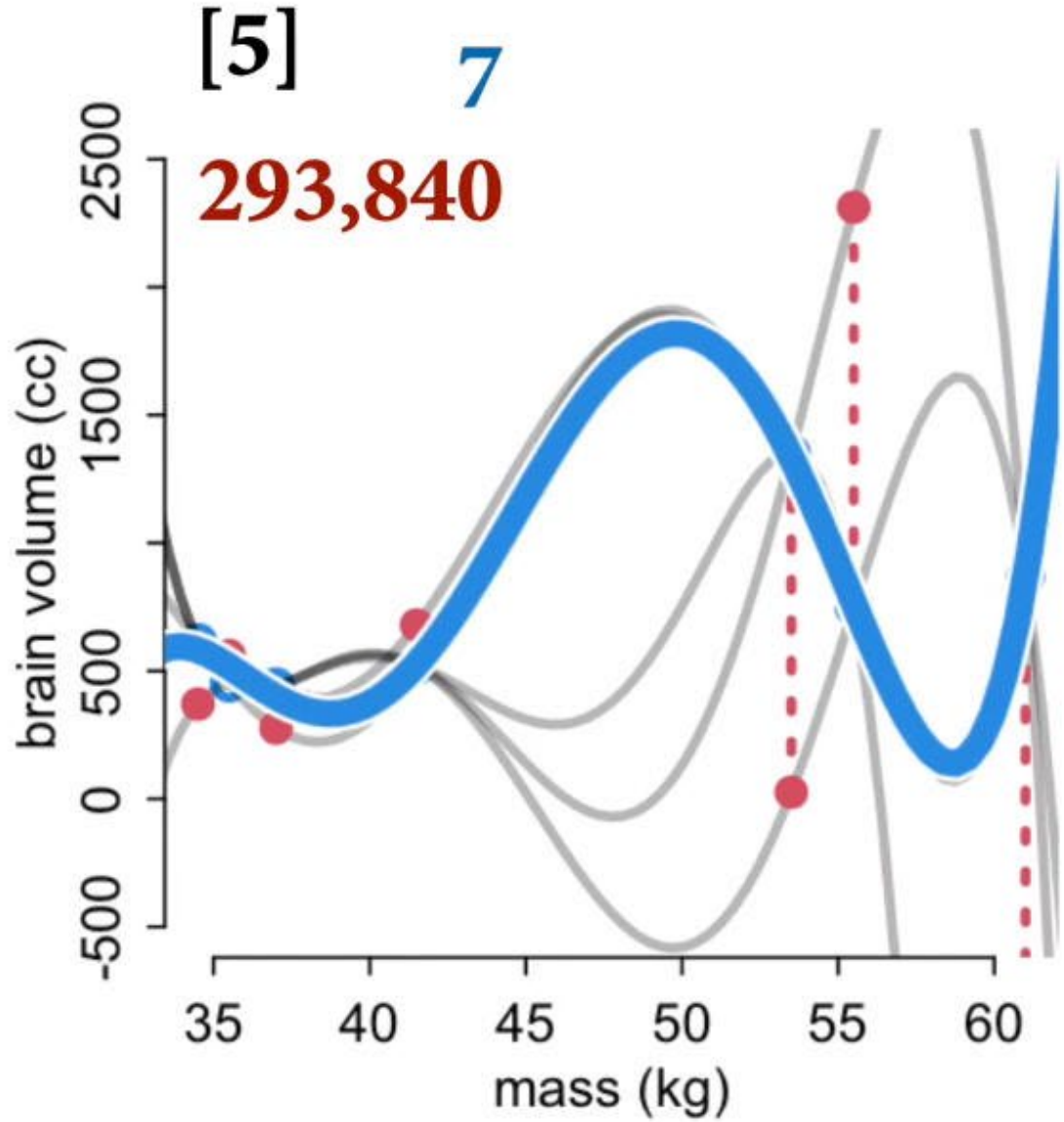



[1] 318
619
brain volume (cc)
600
800
1000
1200
35
40
45
50
55
60
mass (kg)


[3] 201
12,538
brain volume (cc)
200
600
1000
1400
35
40
45
50
55
60
mass (kg)


[2] 289
865
brain volume (cc)
600
800
1000
1200
35
40
45
50
55
60


[4] 120
25,530
brain volume (cc)
200
600
1000
1400
35
40
45
50
55
60

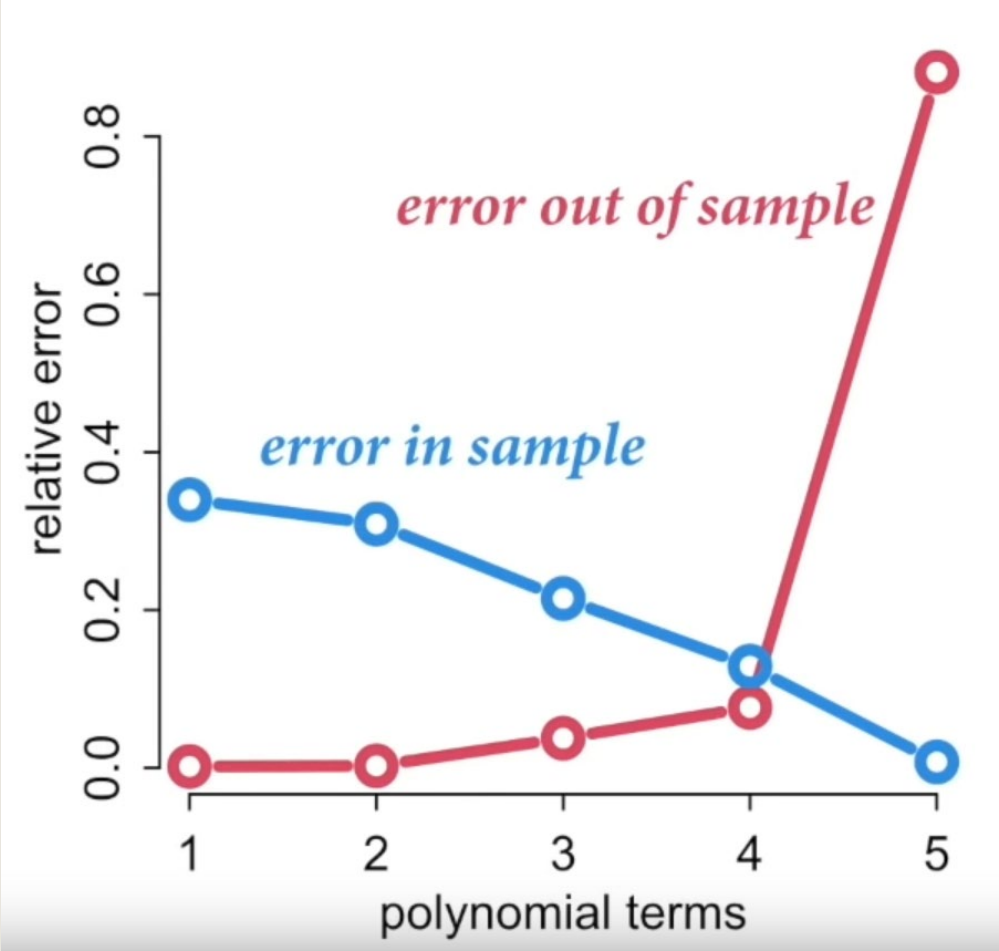

[5] 7
293,840
brain volume (cc)
-500
0
500
1500
2500
35
40
45
50
55
60
mass (kg)

Cross-validation

Comparo distintos modelos candidatos

- Error in sample > error out sample
- A medida que el modelo es más complejo, su poder predictivo fuera de la muestrea cae

Flexibilidad ~ ¡¡¡sobreajuste!!!

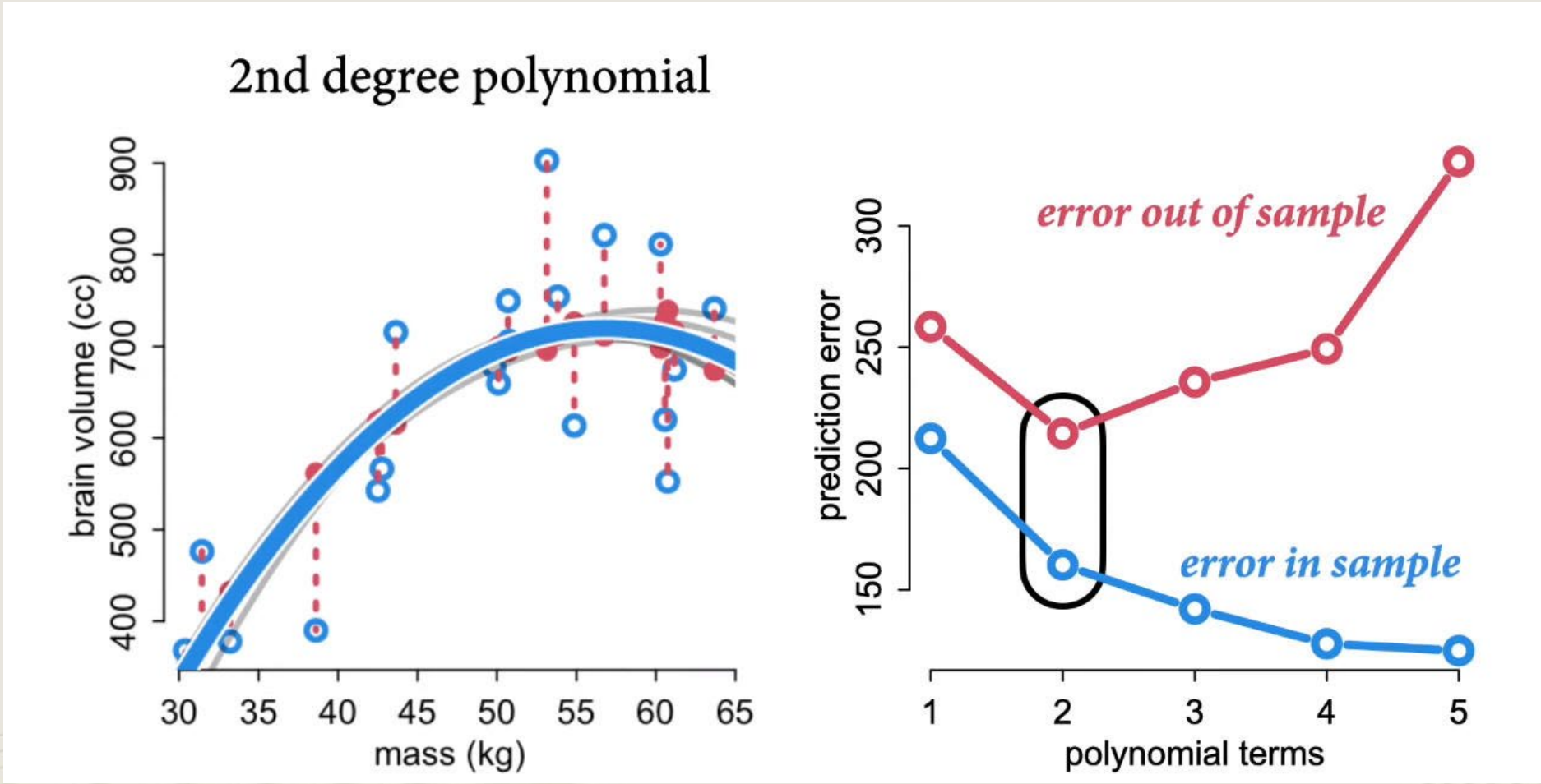


Cross-validation: fitting is easy, but prediction is hard

Cross-validation is a measure of overfitting

Regularization: function que encuentra aspectos regulares del proceso

Priors, priors, priors… reducir la flexibilidad del modelo



¡¡El cómputo de esto es, para todo efecto práctico, imposible!!

- Cross-validation es demasiado costosa e ineficiente

- ¿Qué tal si podemos calcular la brecha a partir de una sola estimación?
  - **Information criteria (WAIC): Widely applicable information criteria**. Teoría de la información
  - **Importance sampling (PSIS): Pareto smooth importance sampling.**
    - Point with low probability has a strong influence on the posterior distribution

Using tools like PSIS and WAIC is much easier than understanding them. Which makes them quite dangerous.

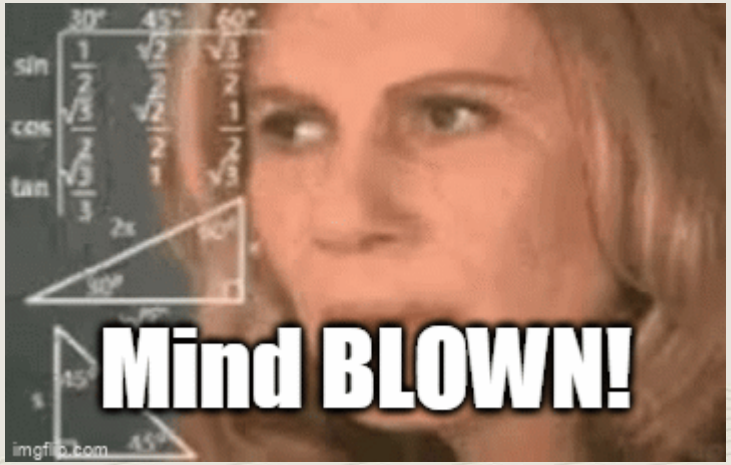
Measuring accuracy (Information Theory)

- What do you want the model to do well at?
- What do we talk about when we talk about accuracy?
- Which definition of accuracy is maximized by knowing the true model generating the data?

  The joint likelihood $P(D|\theta)$
- Information theory: a measurement scale for distance from perfect accuracy (out of sample deviance)

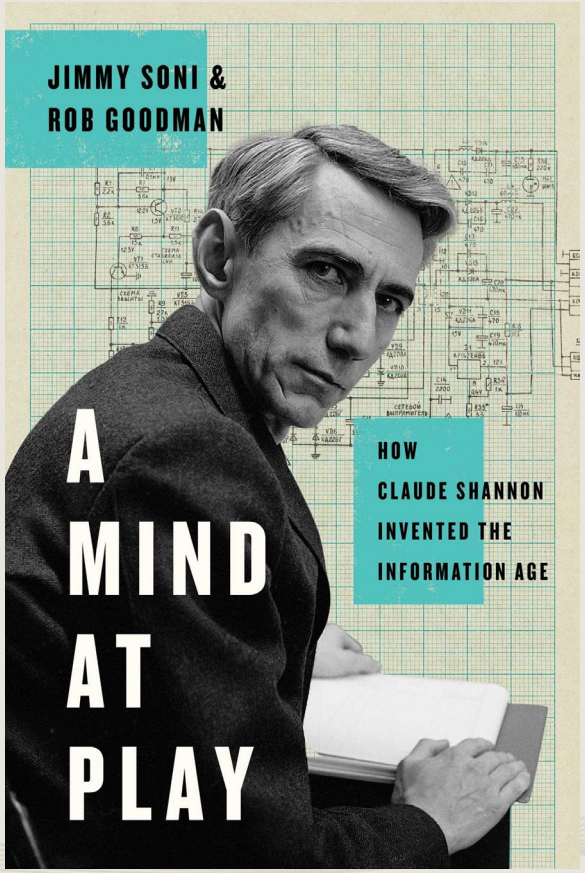
Measuring accuracy (Information Theory)

- **The true data-generating model will not have the highest hit rate**



Information criteria

- Implementing them is much easier tan understanding them (Information Theory)
  - The bit player (https://youtu.be/JP1Ljp8X6hg)



Referencias

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle", in Petrov, B. N.; Csáki, F. (eds.), 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: Akadémiai Kiadó, pp. 267–281. Republished in Kotz, S.; Johnson, N. L., eds. (1992), Breakthroughs in Statistics, vol. I, Springer-Verlag, pp. 610–624

Akaike, H. (1985), "Prediction and entropy", in Atkinson, A. C.; Fienberg, S. E. (eds.), A Celebration of Statistics, Springer, pp. 1–24.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian Statistics* 4, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 147{167. Oxford University Press.

Magnusson, M., Andersen, M., Jonasson, J., and Vehtari, A. (2019). Bayesian leave-one-out cross-validation for large data. *Proceedings of the 36th International Conference on Machine Learning*, 97:4244–4253.

Vehtari, A., Gelman, A., & Gabry J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. In Statistics and Computing, doi:10.1007/s11222-016-9696-4. arXiv preprint arXiv:1507.04544.

https://www.youtube.com/watch?v=odGAAJDlgp8