The Fork

Open unless you condition on Z

The Pipe

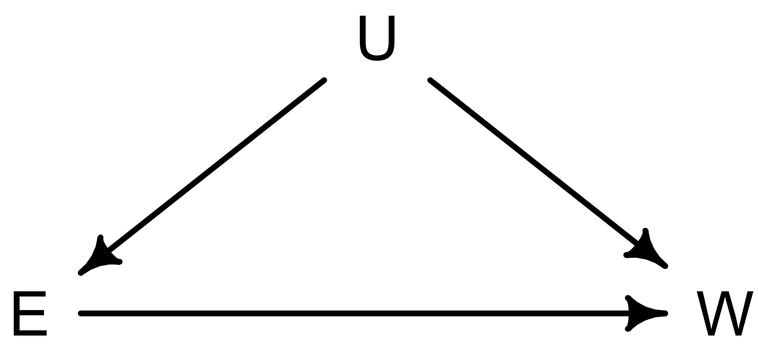
Open unless you condition on Z

The Collider

Closed until you condition on Z

The Descendant

Conditioning on A is like conditioning on Z



Two paths from E to W:

(1) E → W

(2) E ← U → W

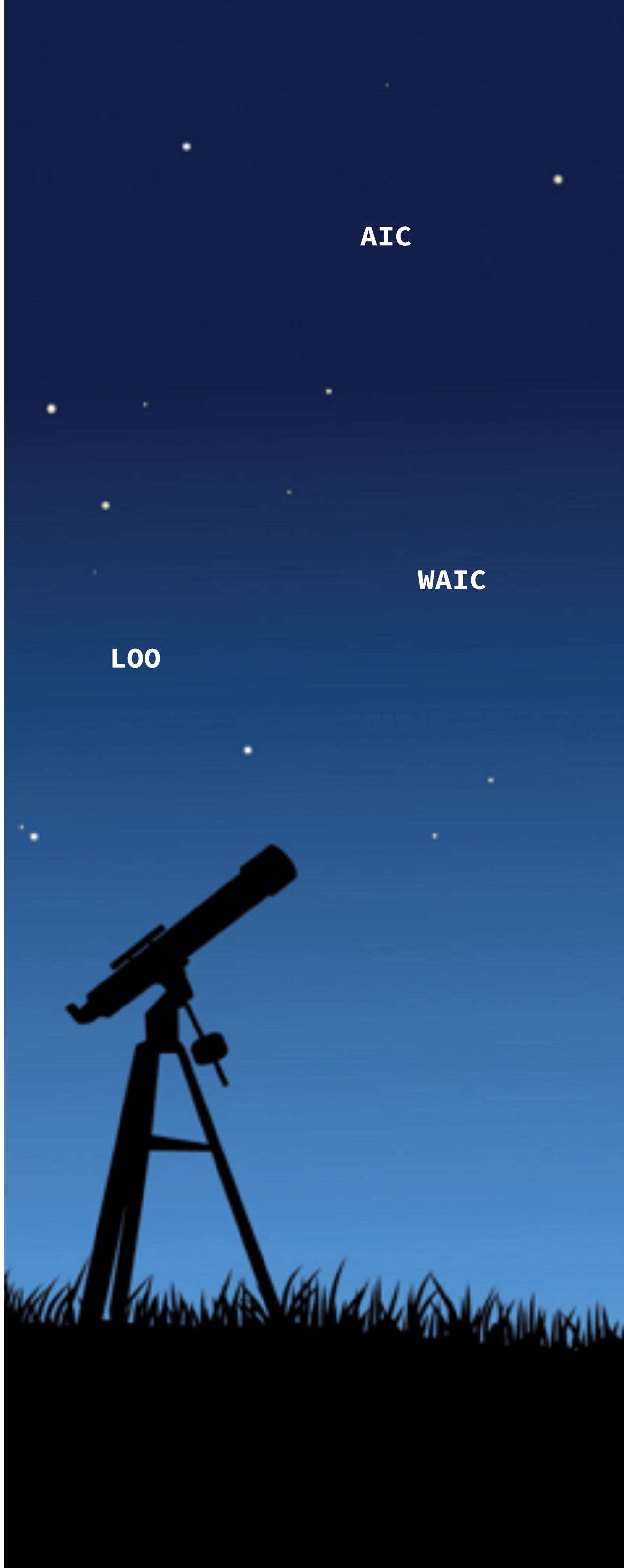
Close 2nd path by conditioning on U, closing the pipe.

Ulysses' Compass

- Two major hazards: (1) Too simple (2) Too complex

Goals

- Understand *overfitting* and *underfitting*
- Introduce *regularization*
- Cross-validation & information criteria:
  - estimate predictive accuracy
  - estimate overfitting risk
  - understand how overfitting relates to complexity
  - identify influential observations
- See that prediction and causal inference are different objectives



The Problem with Parameters

- **What should a model learn from a sample?**
- *Underfitting*: Learning too little from the data. Too simple models both fit and predict poorly.
- *Overfitting*: Learning too much from the data. Complex models tend to fit better, predict worse.
- Want to find a model that navigates between underfitting and overfitting
- Problem: Fit to sample always* improves as we add parameters

*Not true of multilevel models & other types

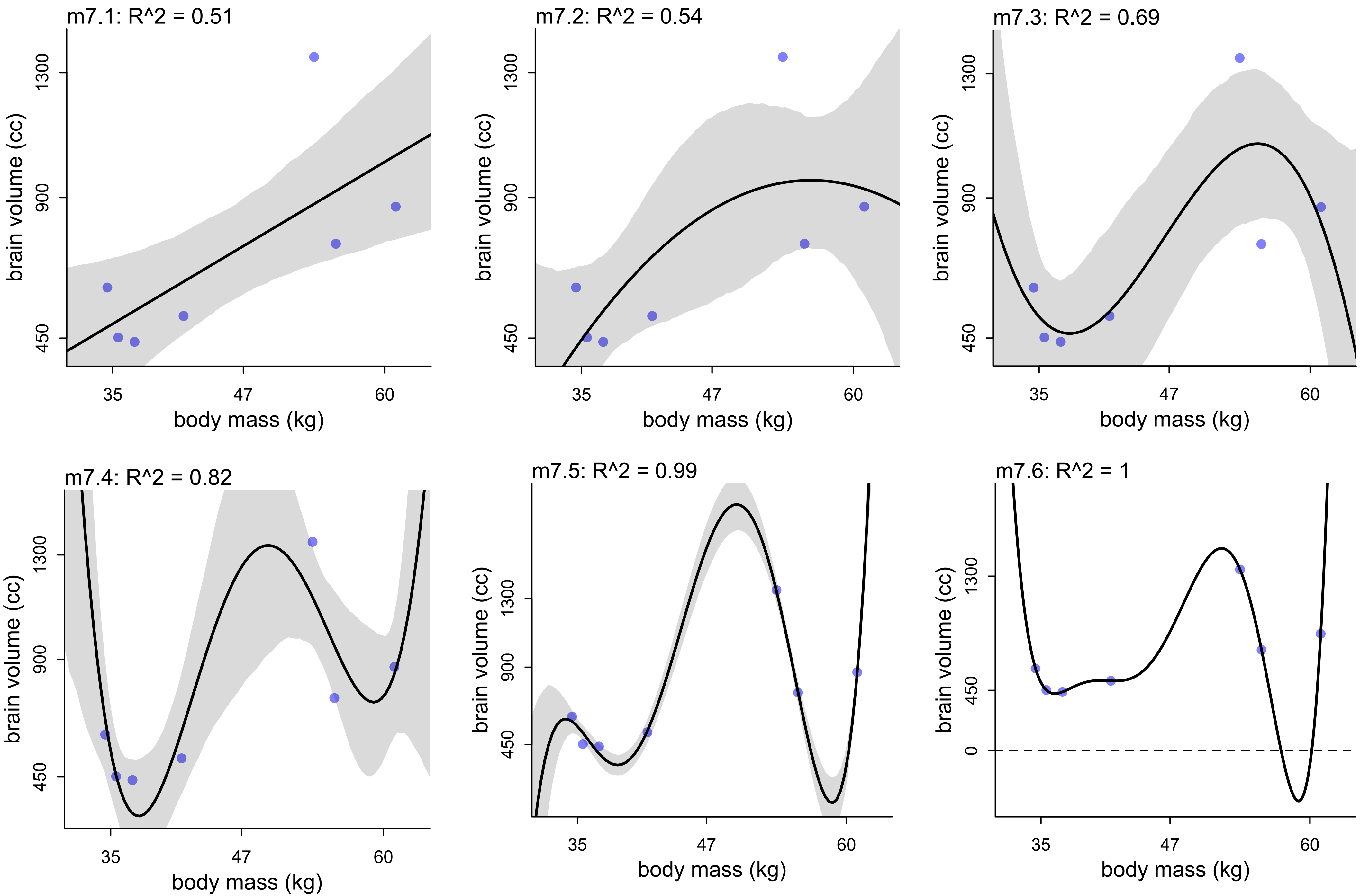

m7.1: R^2 = 0.51
m7.2: R^2 = 0.54
m7.3: R^2 = 0.69
m7.4: R^2 = 0.82
m7.5: R^2 = 0.99
m7.6: R^2 = 1
brain volume (cc)
body mass (kg)
450
900
1300
0
35
47
60

Figure 7.3

Underfitting
Insensitive to
exact data

Overfitting
Very sensitive to
exact data

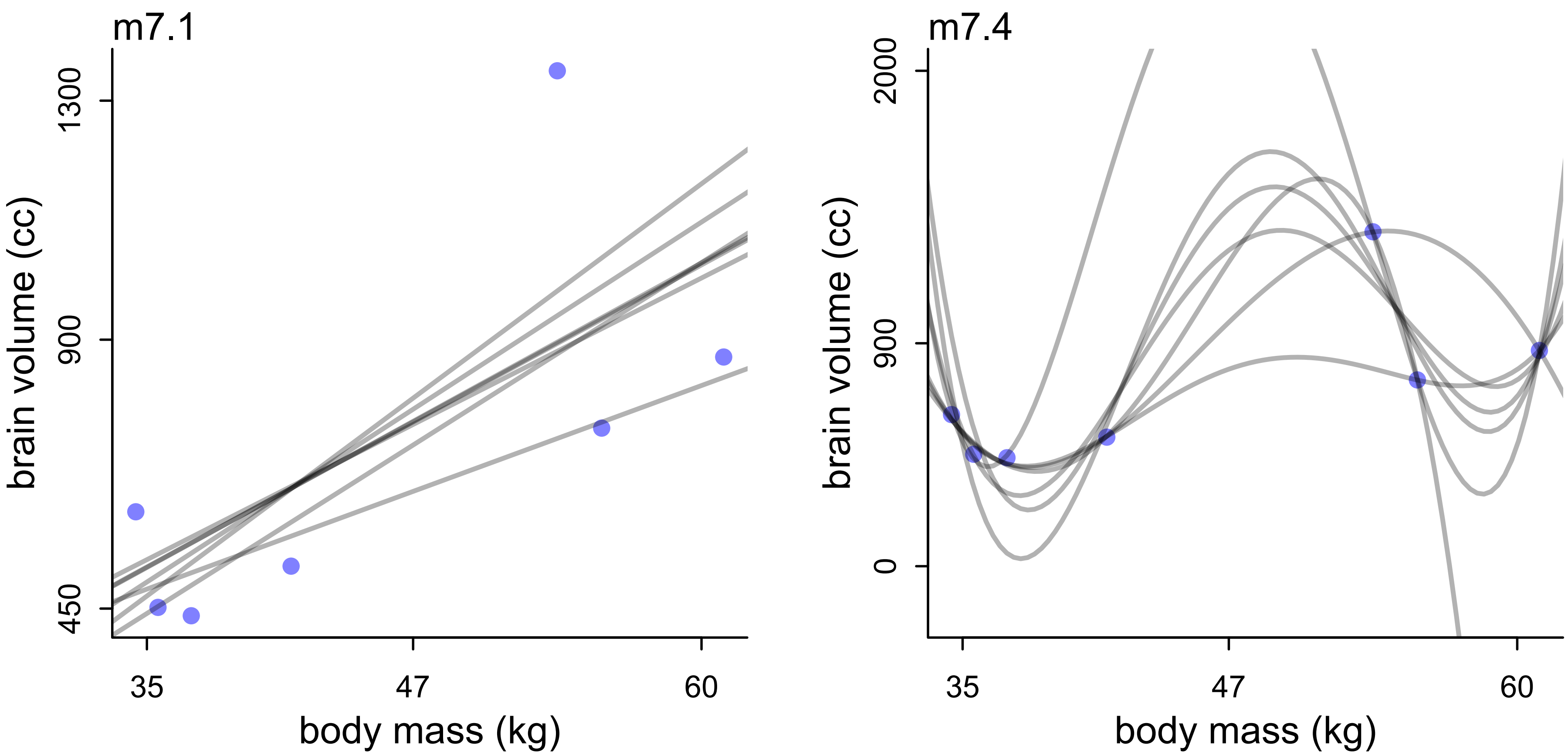


Figure 7.5

Importance of being *regular*

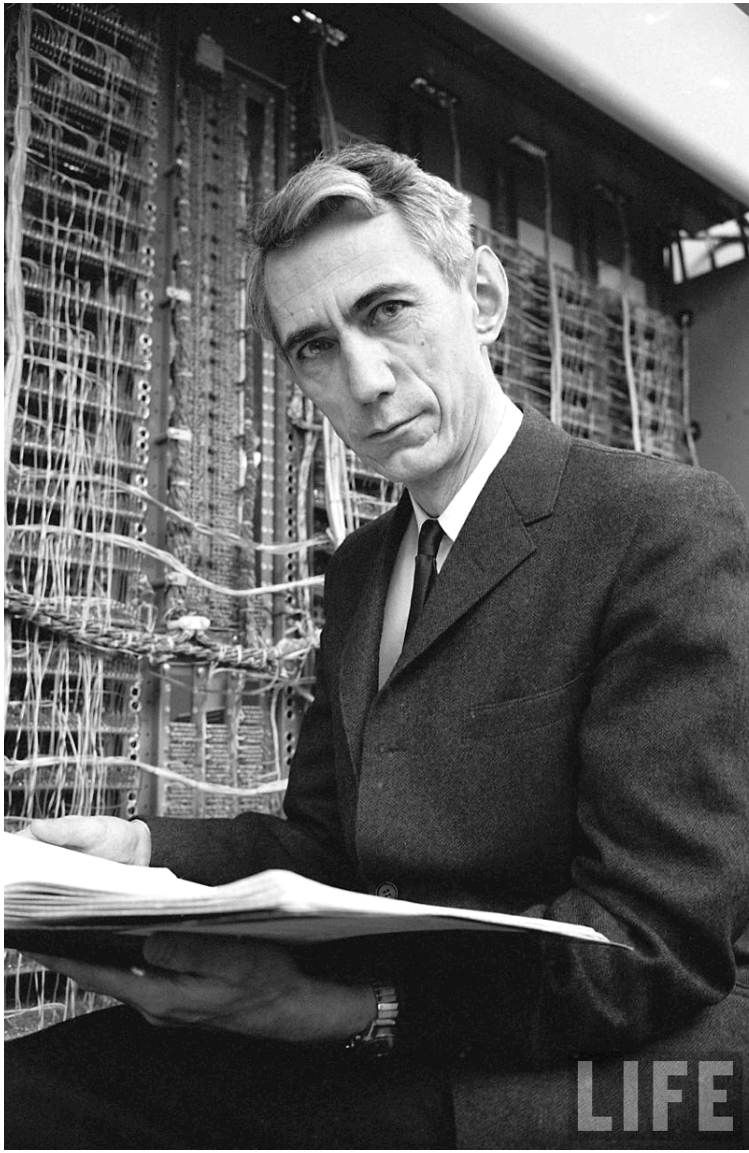
- Want the *regular* features of the sample
- Strategies
  - Regularizing priors (penalized likelihood)
  - Cross-validation
  - Information criteria
  - Science!
- Proper approach depends upon purpose
- Answers are never *only* in the data, but they do usually require data

Information entropy

- 1948, Claude Shannon derived *information entropy*:

$$H(p) = -\operatorname{E}\log(p_i) = -\sum_{i=1}^{n} p_i \log(p_i)$$

Uncertainty in a probability distribution is average (minus) log-probability of an event.



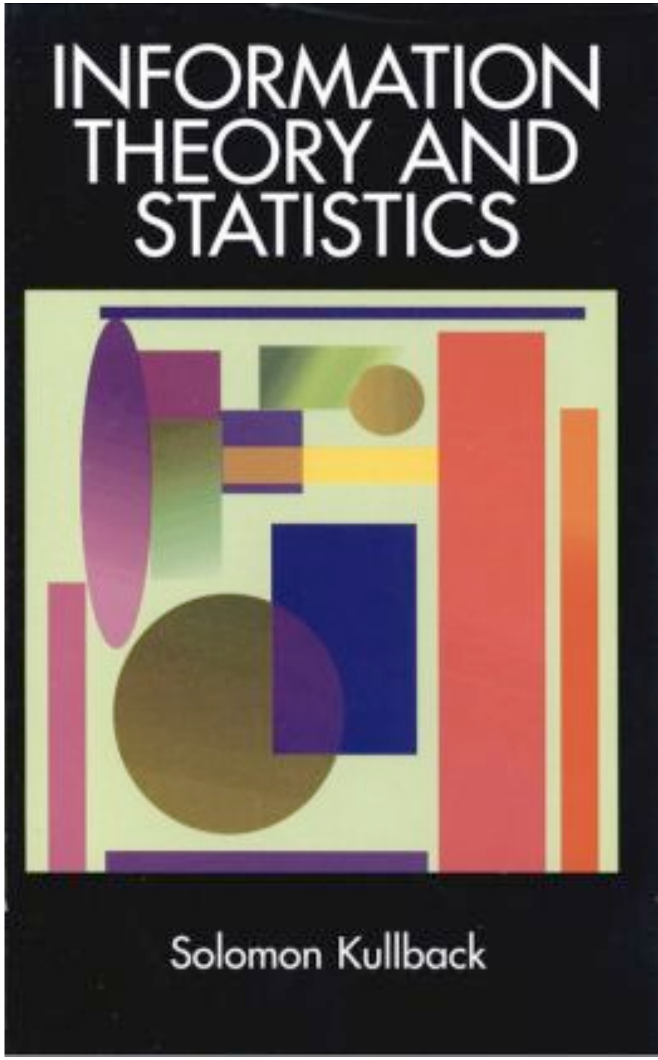
Shannon (1916–2001)

Entropy to accuracy

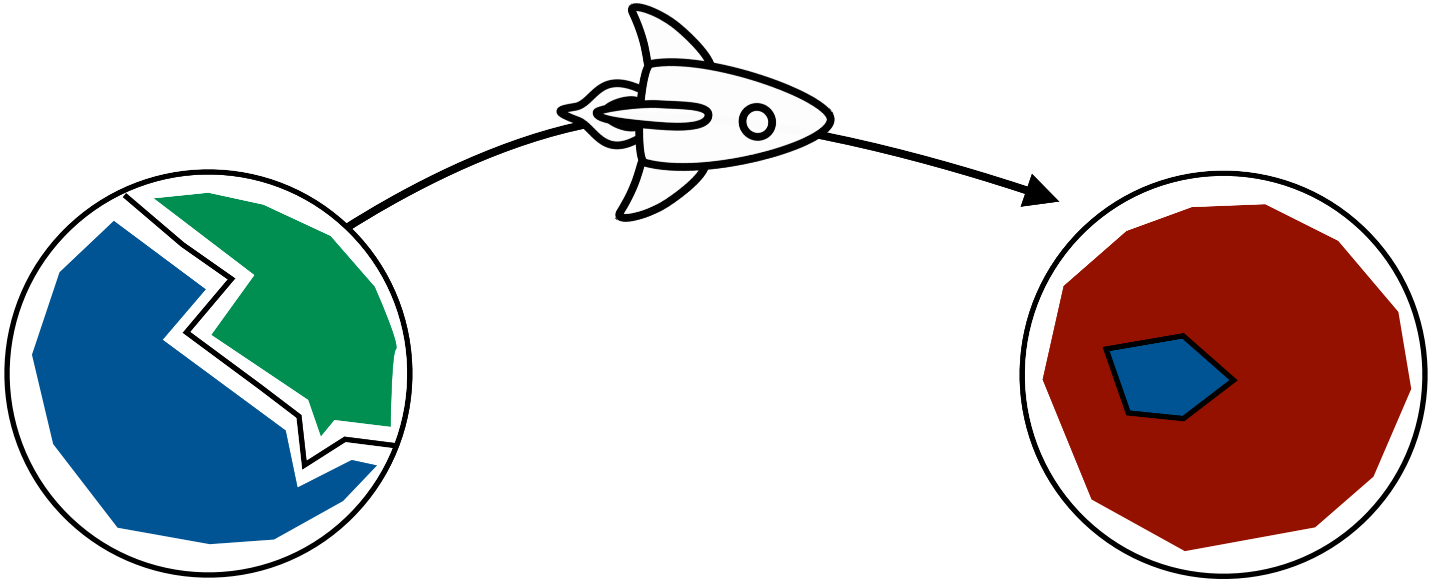
- Two probability distributions: p, q
- p is true, q is model
- How accurate is q, for describing p?
- Distance from q to p: *Divergence*

$$D_{\mathrm{KL}}(p, q) = \sum_i p_i \big(\log(p_i) - \log(q_i) \big)$$

Distance from q to p is the average difference in log-probability.



Divergence is not symmetric!



Everybody overfits

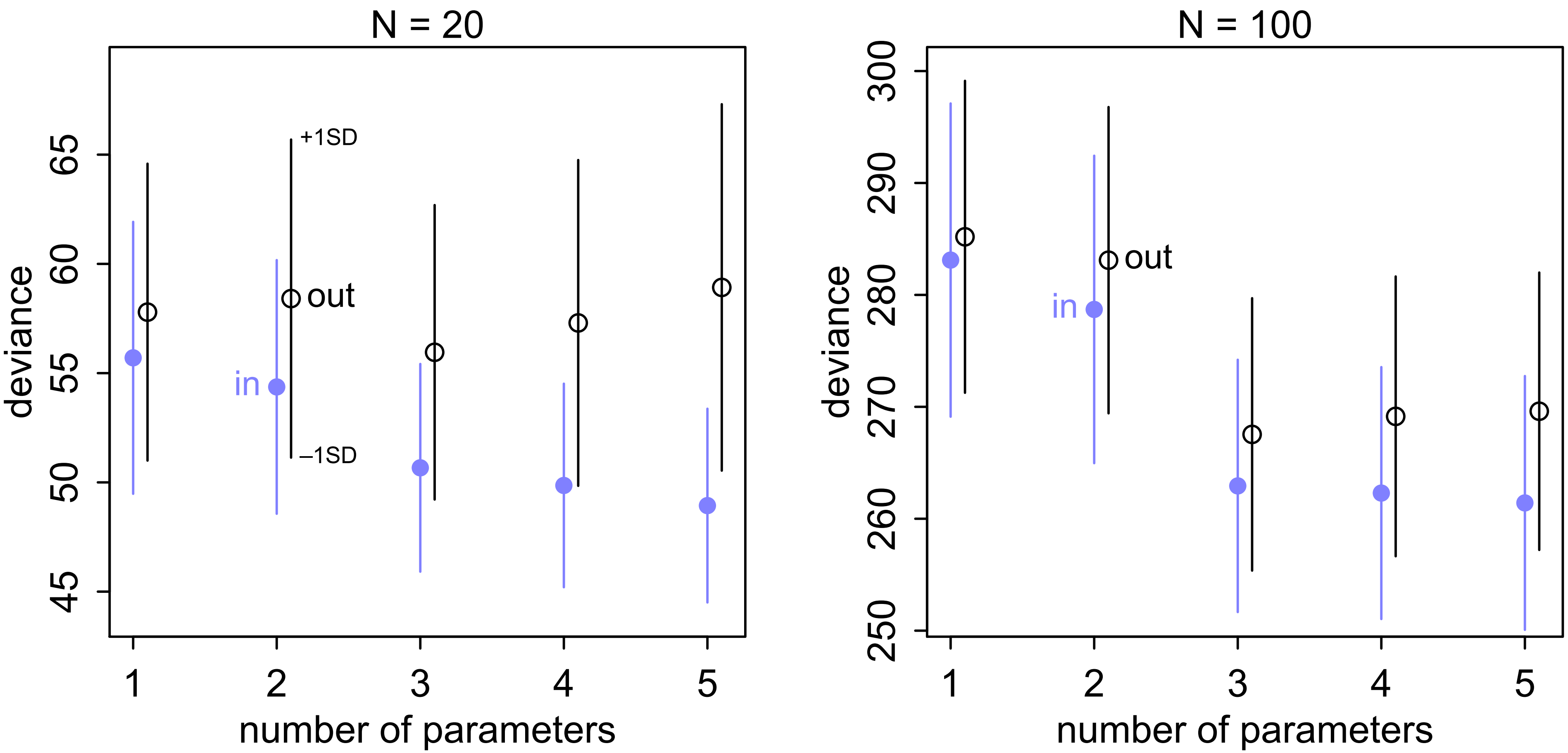


Figure 7.7

Regularization

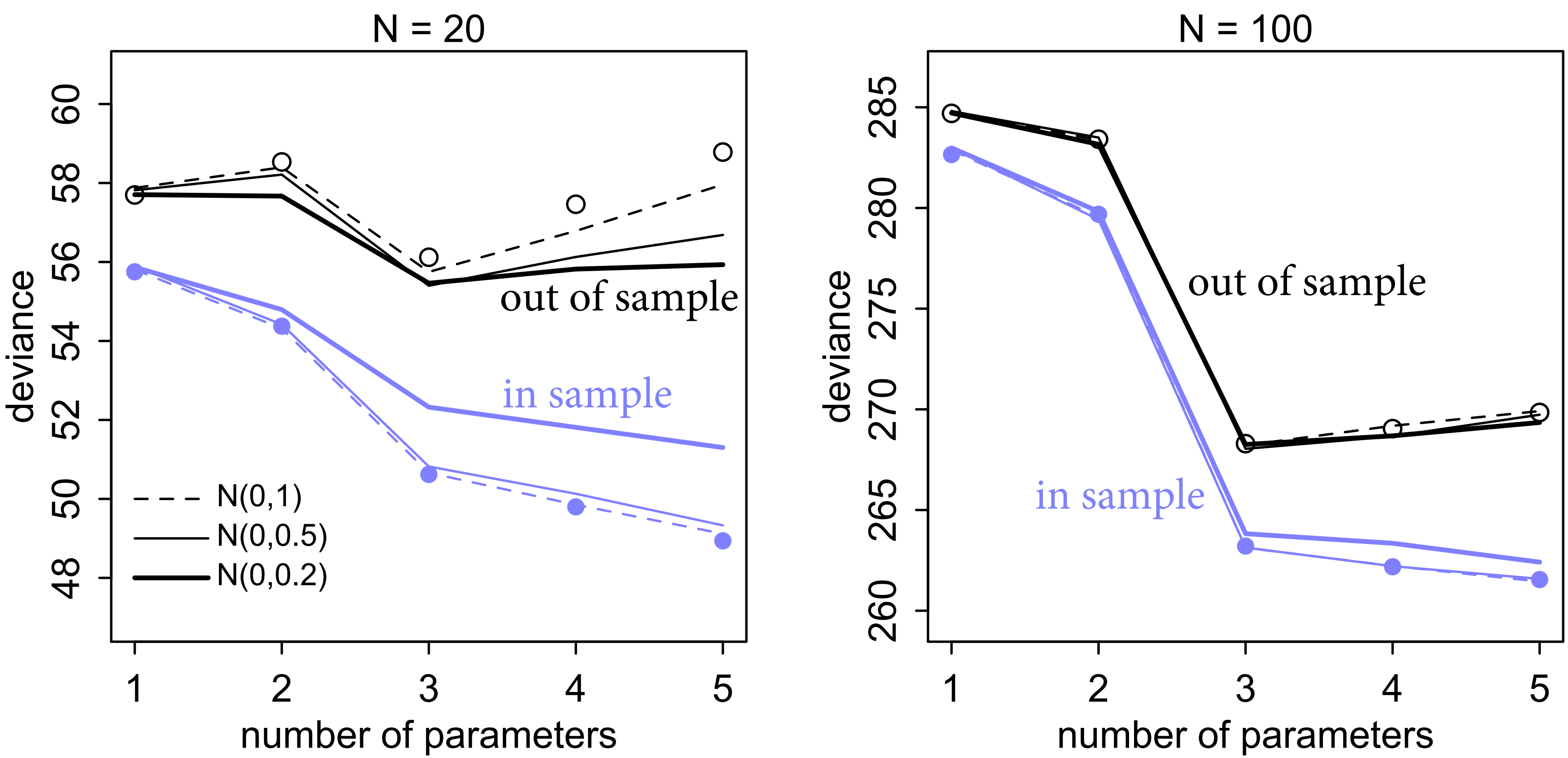


Figure 7.9

Regularization

- **Must be skeptical of the sample!**
- Use informative, conservative priors to reduce overfitting => model learns less from sample
- But if too skeptical, model learns too little
- Such priors are *regularizing*

Smooth Cross-validation

- Most common: Leave-one-out
- Very expensive!
- Useful approximation: Importance sampling (IS)
- More useful: Pareto-smoothed importance sampling (PSIS)
- PSIS-LOO accurate, lots of useful diagnostics
- `LOO` function in rethinking
- See also `loo` package

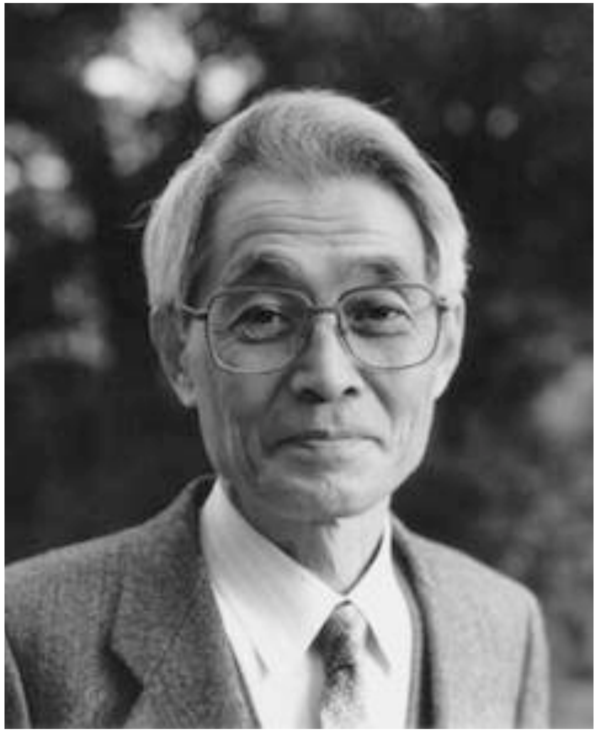
Prof Aki Vehtari (Helsinki), smooth estimator

Akaike information criterion
[ah–ka–ee–kay]

- Estimate K-L Distance in theory
- Most famous is the first, AIC
- Under some strict conditions:

$$\text{AIC} = D_{\text{train}} + 2k \approx \text{E}\, D_{\text{test}}$$

k is parameter count



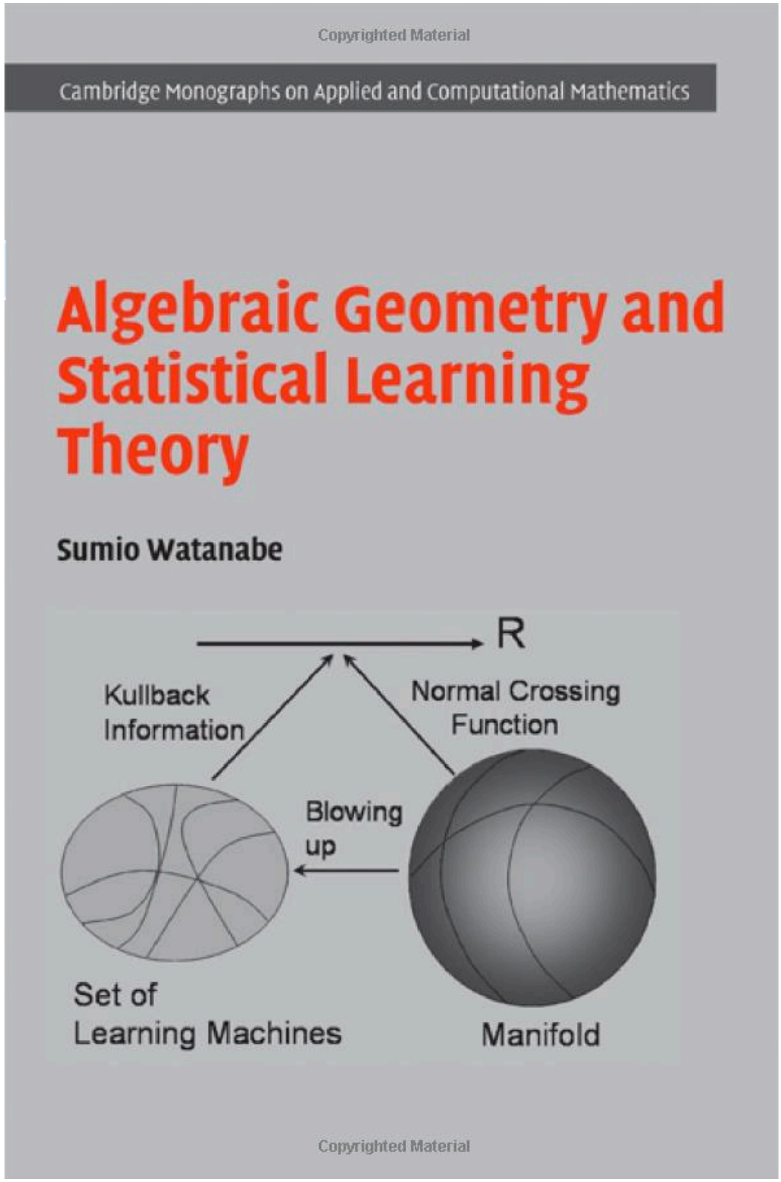
Hirotugu Akaike
赤池弘次
(1927–2009)

Widely Applicable IC

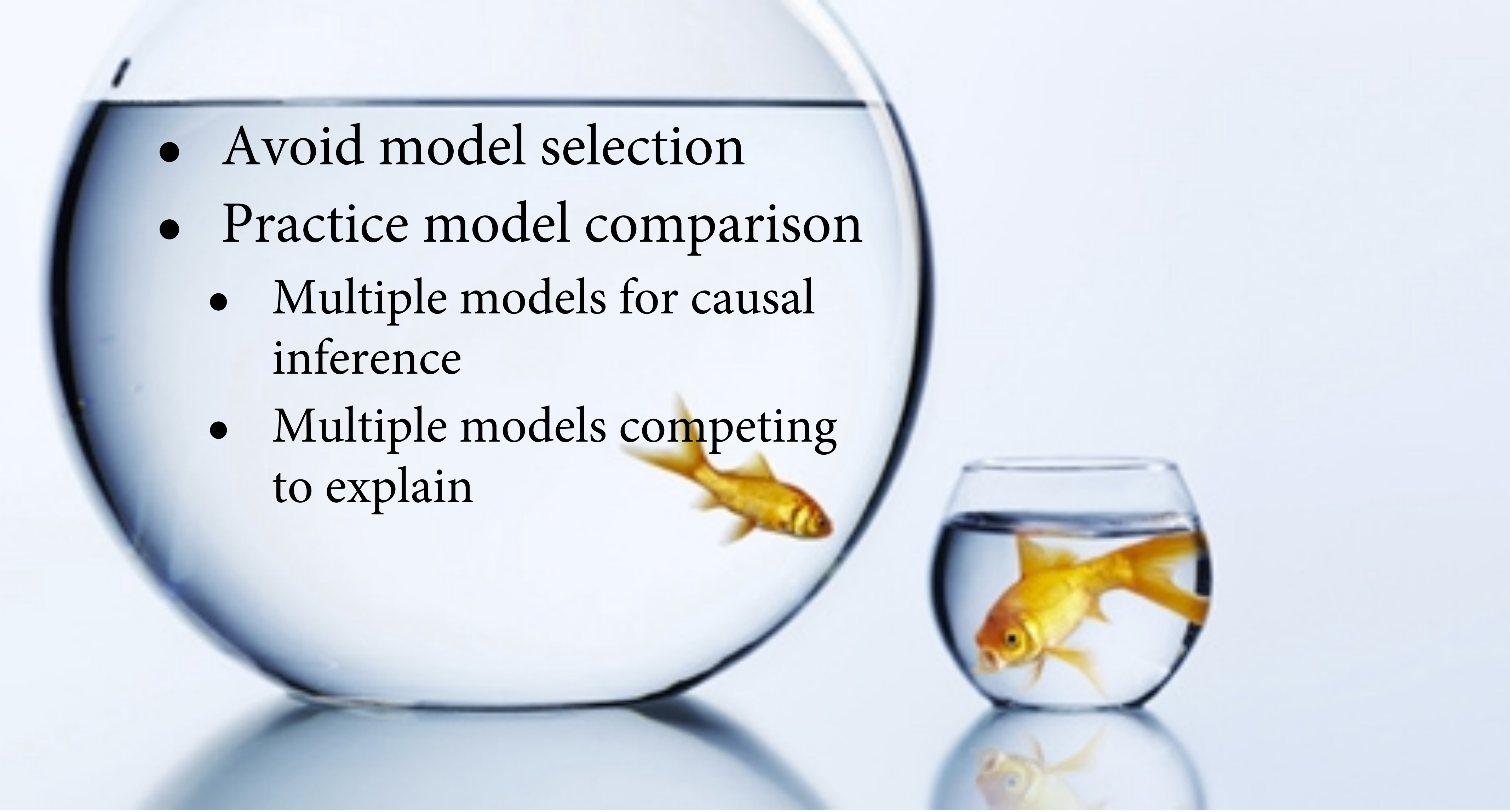
- AIC of historical interest now
- Widely Applicable Information Criterion (WAIC)
  - Sumio Watanabe 2010

$$\text{WAIC}(y, \Theta) = -2\Big(\text{lppd} - \underbrace{\sum_i \text{var}_\Theta \log p(y_i|\Theta)}_{\text{penalty term}}\Big)$$

- Does not assume Gaussian posterior
- `WAIC` function in rethinking



Using CV & WAIC



- Avoid model selection
- Practice model comparison
  - Multiple models for causal inference
  - Multiple models competing to explain

R code 7.27

```
set.seed(77)
compare( m6.6 , m6.7 , m6.8 )
```

```
                    WAIC pWAIC dWAIC weight    SE   dSE
treat + fungus m6.7 361.9   3.8   0.0      1 14.26    NA
        fungus m6.8 402.8   2.6  40.9      0 11.28 10.48
     intercept m6.6 405.9   1.6  44.0      0 11.66 12.23
```