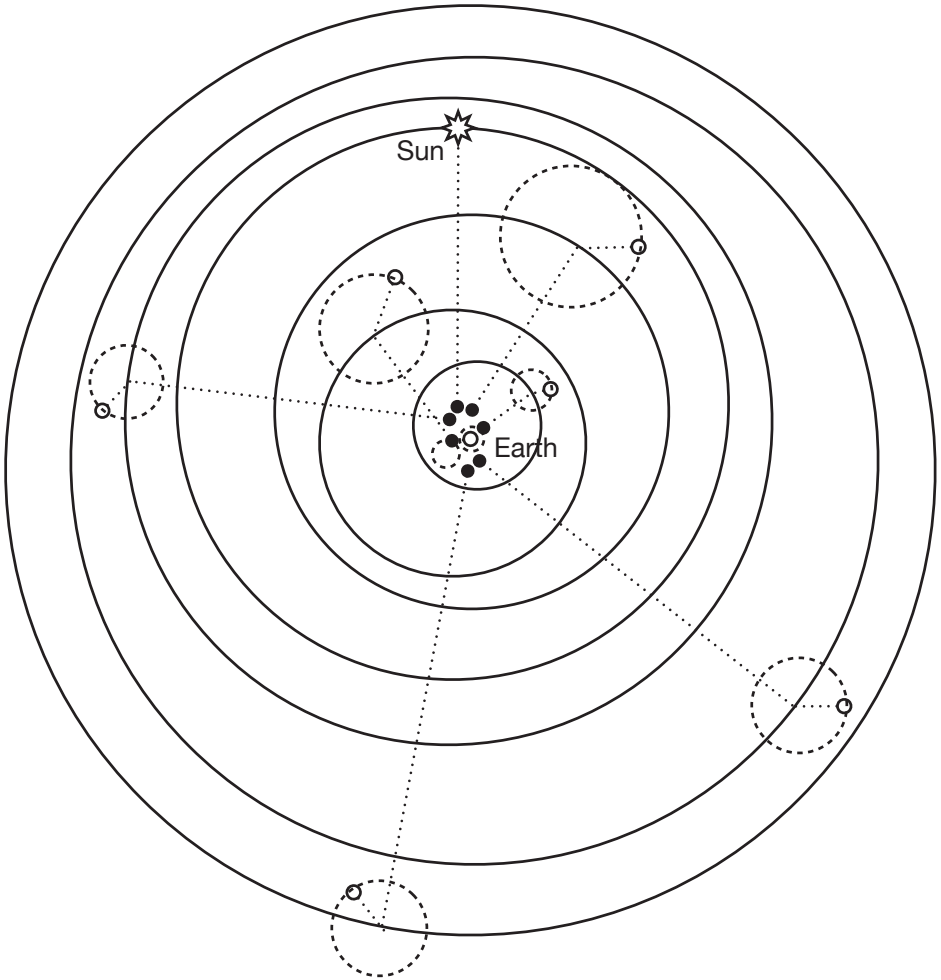
Statistical Rethinking

Week 4: Ockham, Ulysses, and the Model

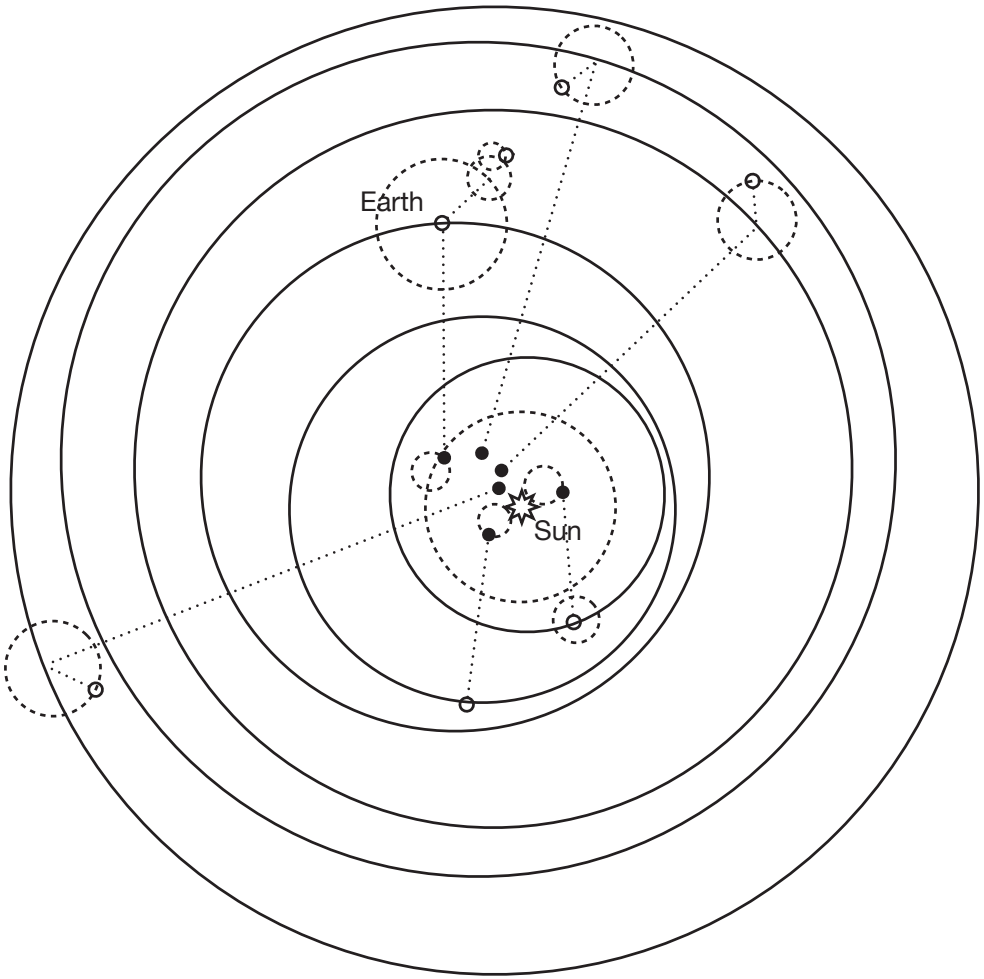
Richard McElreath

Mikołaj Kopernik (1473–1543)

Ptolemaic Model



Copernican Model

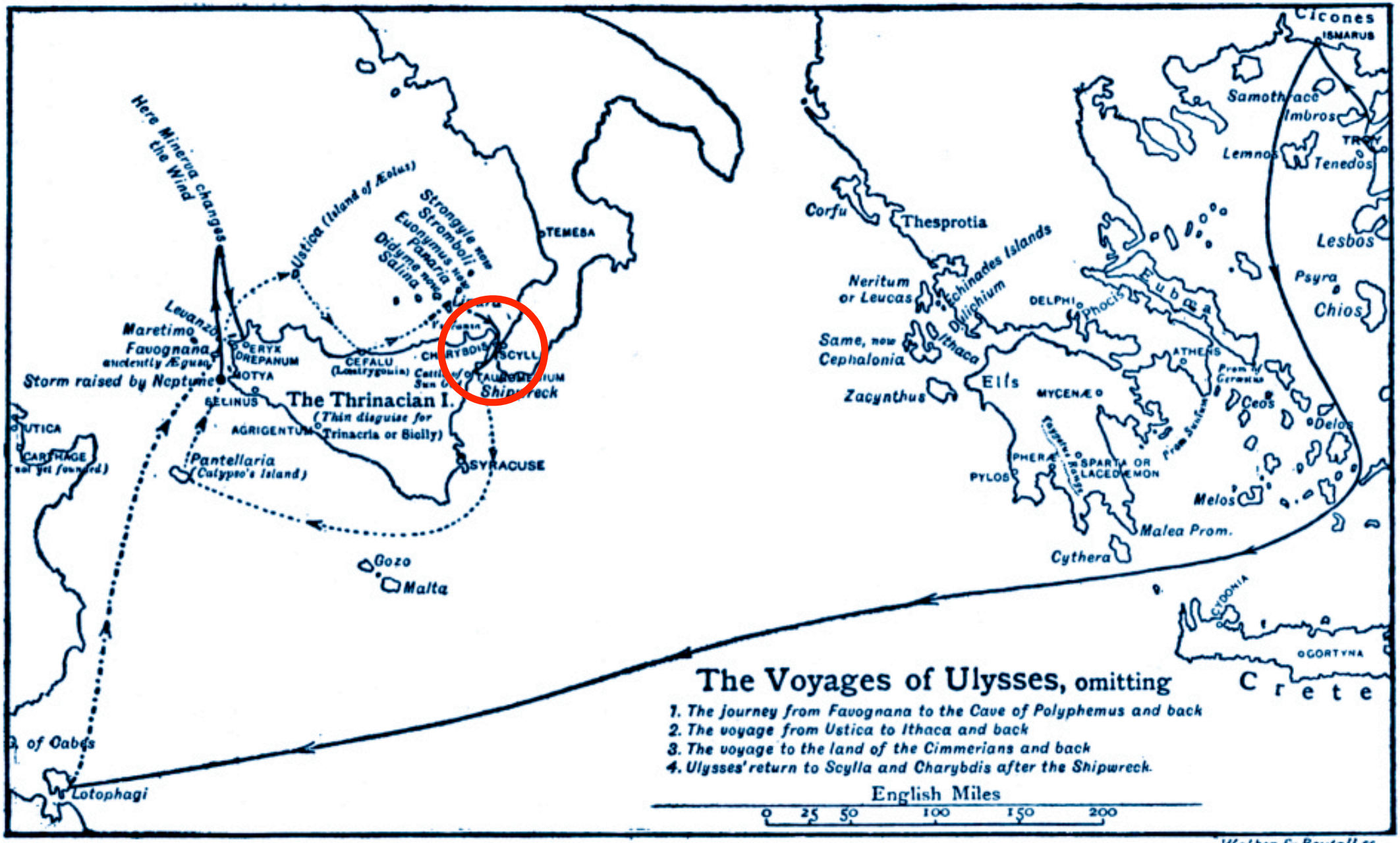


Ockham's Razor?

Numquam ponenda est pluralitas sine necessitate.

(Plurality should never be posited without necessity.)

William of Ockham
(c.1288–c.1348)


Here Minerva changes the Wind
Ustica (Island of Æolus)
Stromgyle now Stromboli
Euonymus now Panaria
Didyme now Salina
TEMESA
Levanzo
Maretimo
Favognana
anciently Ægusa
ERYX
DREPANUM
MOTYA
CEFALU
(Lemmygonia)
CHARYBDIS
SCYLLA
TAURO
Shipwreck
Storm raised by Neptune
SELINUS
The Thrinacian I.
(Thin disguise for Trinacria or Sicily)
AGRIGENTUM
UTICA
CARTHAGE
Pantellaria
(Calypso's Island)
SYRACUSE
Gozo
Malta
of Gades
Lotophagi
Cicones
ISMARUS
Samothrace
Imbros
Lemnos
TROY
Tenedos
Corfu
Thesprotia
Lesbos
Psyra
Chios
Neritum
or Leucas
Echinades Islands
Dulichium
DELPHI
Phocis
Same, now
Cephalonia
Ithaca
ATHENS
Zacynthus
Elis
MYCENÆ
Ceos
Delos
PHERÆ
SPARTA OR
LACEDÆMON
PYLOS
Melos
Malea Prom.
Cythera
CYDONIA
GORTYNA
Crete
The Voyages of Ulysses, omitting
1. The journey from Favognana to the Cave of Polyphemus and back
2. The voyage from Ustica to Ithaca and back
3. The voyage to the land of the Cimmerians and back
4. Ulysses' return to Scylla and Charybdis after the Shipwreck.
English Miles
0 25 50 100 150 200
Walker & Boutall sc.

Ulysses' Compass

- Two major hazards:
  1. Too simple
  2. Too complex

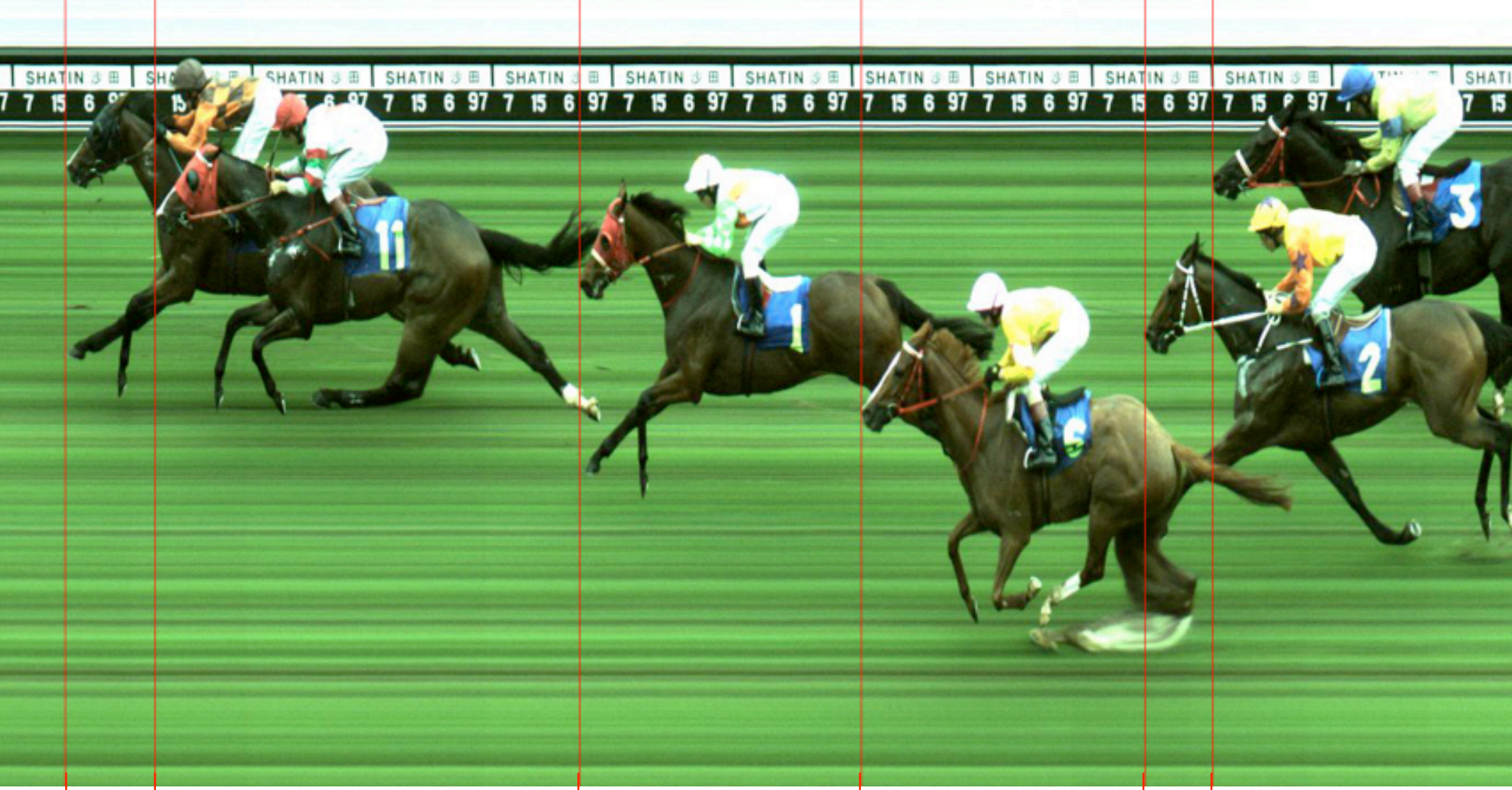
Stargazing

- *Stargazing*: Using asterisks ($p < 0.05$) to decide which variables improve prediction
- Arbitrary 5% is arbitrary

```
Coefficients:
       Estimate  Std. Error     z value     Pr(z)
a    1.5699e+02  9.3802e-16  1.6736e+17 < 2.2e-16 ***
b1   1.6540e-01  6.6628e-14  2.4825e+12 < 2.2e-16 ***
b2  -4.7063e-02  3.2586e-13 -1.4443e+11 < 2.2e-16 ***
b3   1.9168e-03  5.6805e-11  3.3743e+07 < 2.2e-16 ***
b4  -1.4002e-05  6.6694e-11 -2.0994e+05 < 2.2e-16 ***
b5  -4.7965e-07  4.7818e-08 -1.0031e+01 < 2.2e-16 ***
b6   6.6002e-09  9.5819e-10  6.8882e+00 5.651e-12 ***
tau  1.2132e-01  5.2829e-20  2.2965e+18 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```

Goals this week

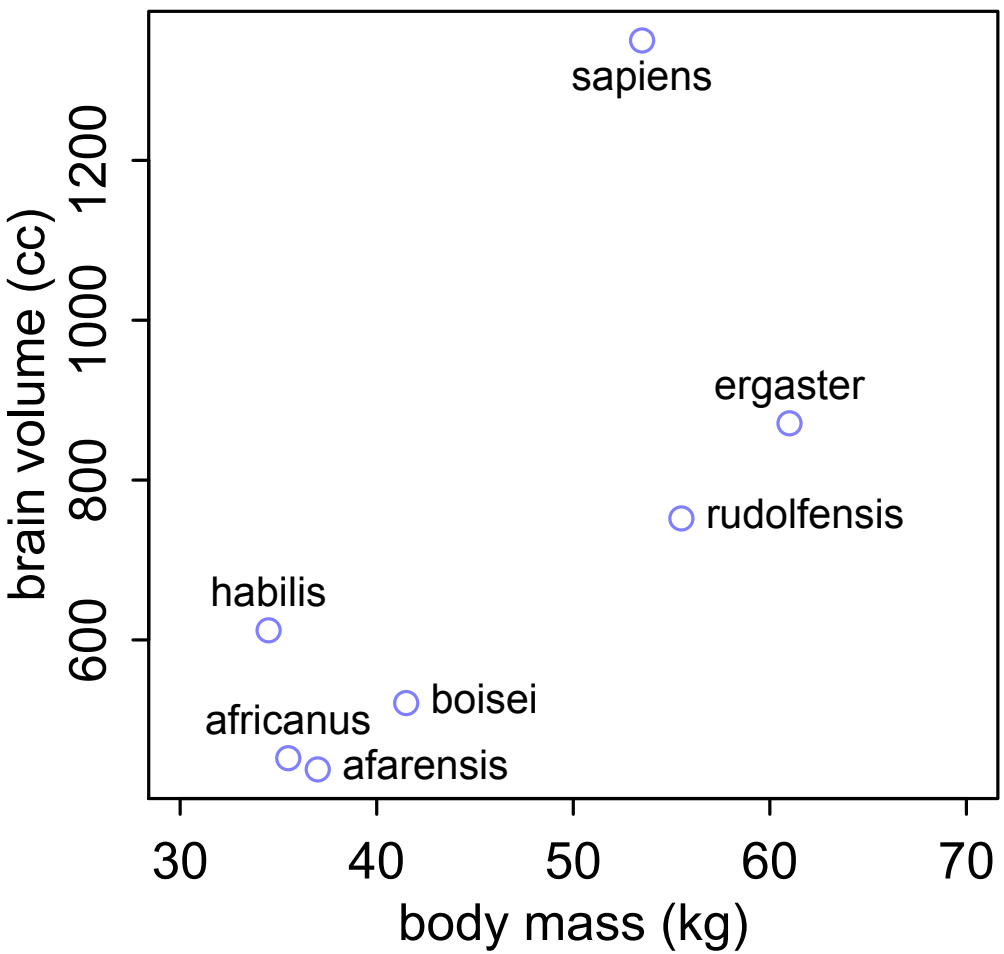
- Understand *overfitting* and *underfitting*
- Learn AIC/DIC/WAIC as ways to:
  - guard against overfitting and underfitting
  - explicitly compare models
- Introduce *regularizing* priors as complementary strategy
- Learn how to average *predictions* across models


A
B
C
D
E
F

The Problem with Parameters

- *Underfitting*: Learning too little from the data. Too simple models both fit and predict poorly.
- *Overfitting*: Learning too much from the data. Complex models always fit better, but often predict worse.
- Need to find a model that navigates between underfitting and overfitting

The Problem with Parameters



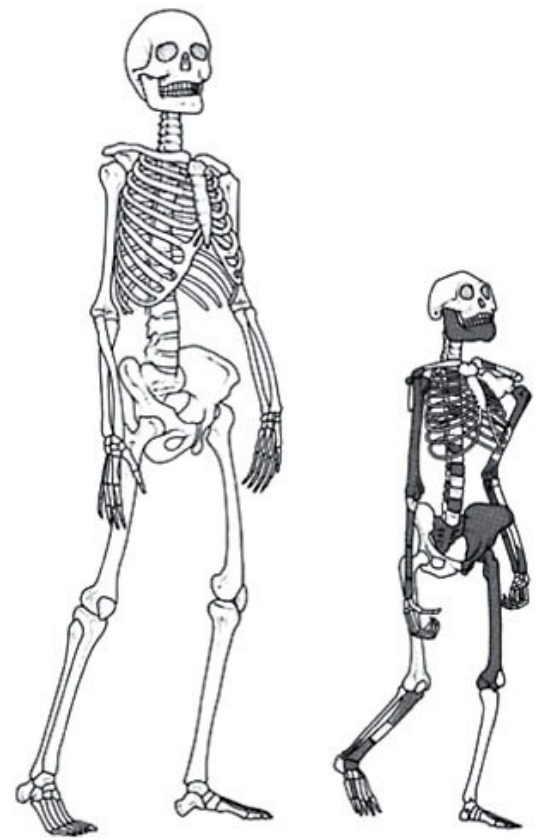
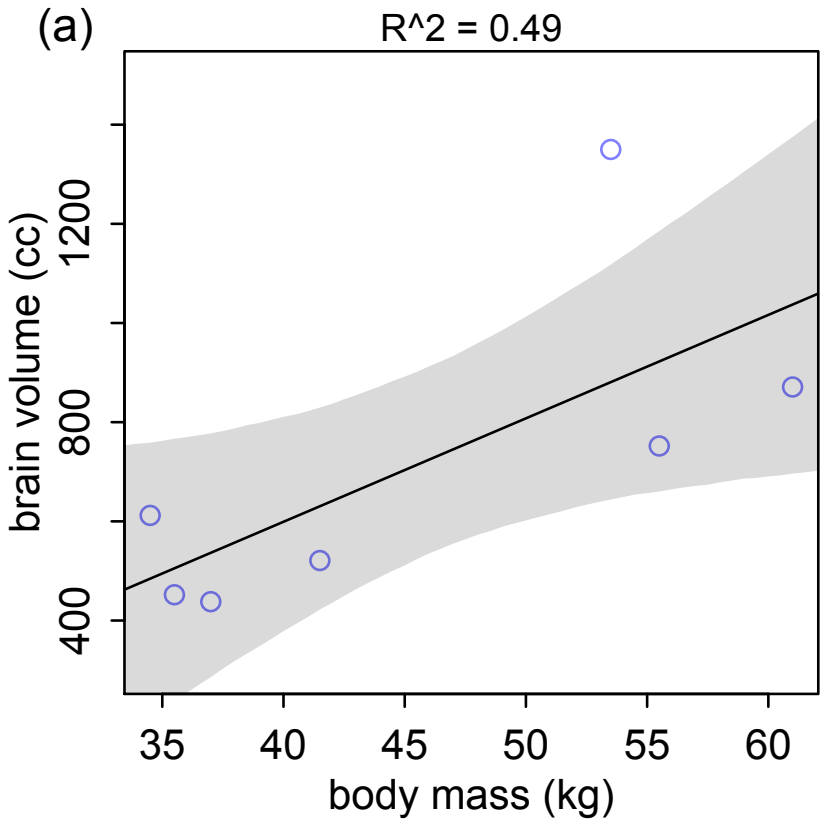


Figure 6.2

Hominin brains

- Simplest model:

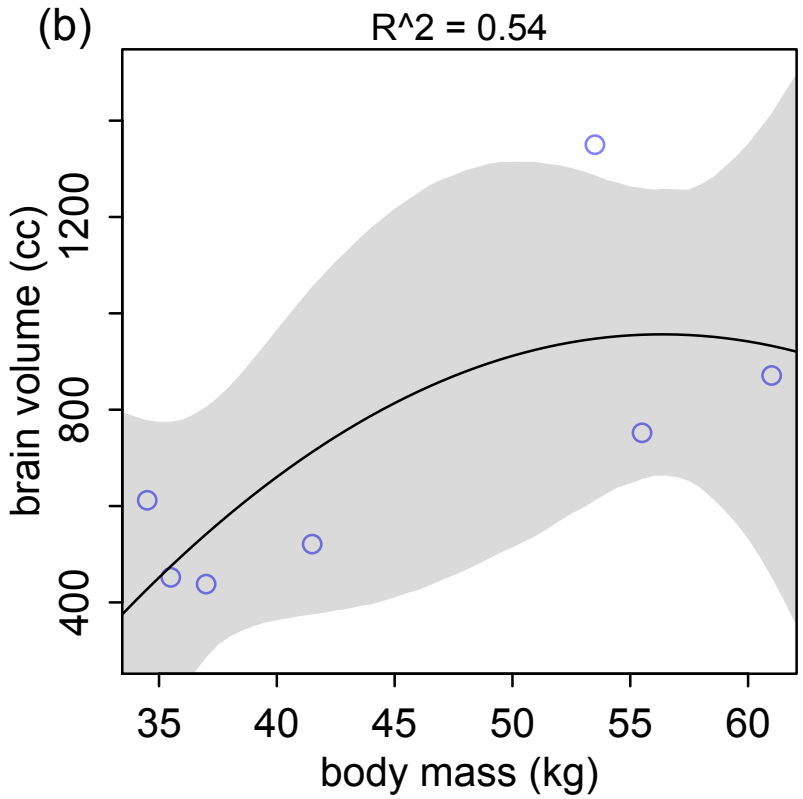
$$v_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_1 m_i$$



Hominin brains

- Why not parabola?

$$v_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_1 m_i + \beta_2 m_i^2$$



Hominin brains

- Why not higher order polynomials?

$$\mu_i = \alpha + \beta_1 m_i + \beta_2 m_i^2 + \beta_3 m_i^3 + \beta_4 m_i^4 + \beta_5 m_i^5 + \beta_6 m_i^6$$

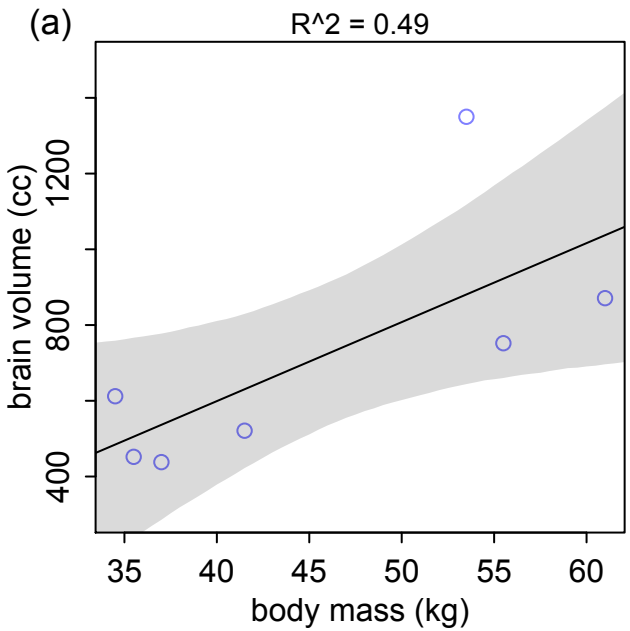

(a)
R^2 = 0.49
brain volume (cc)
400
800
1200
35
40
45
50
55
60
body mass (kg)

Figure 6.3

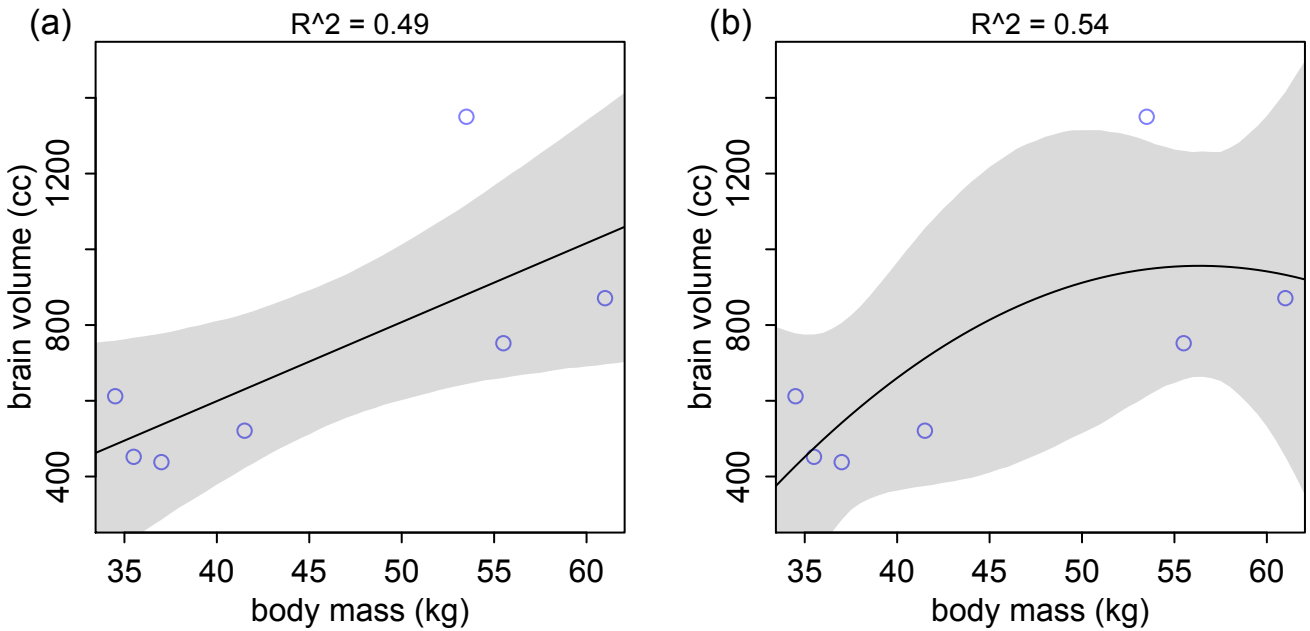

(a)
R^2 = 0.49
brain volume (cc)
400
800
1200
35
40
45
50
55
60
body mass (kg)
(b)
R^2 = 0.54
brain volume (cc)
400
800
1200
35
40
45
50
55
60
body mass (kg)

Figure 6.3

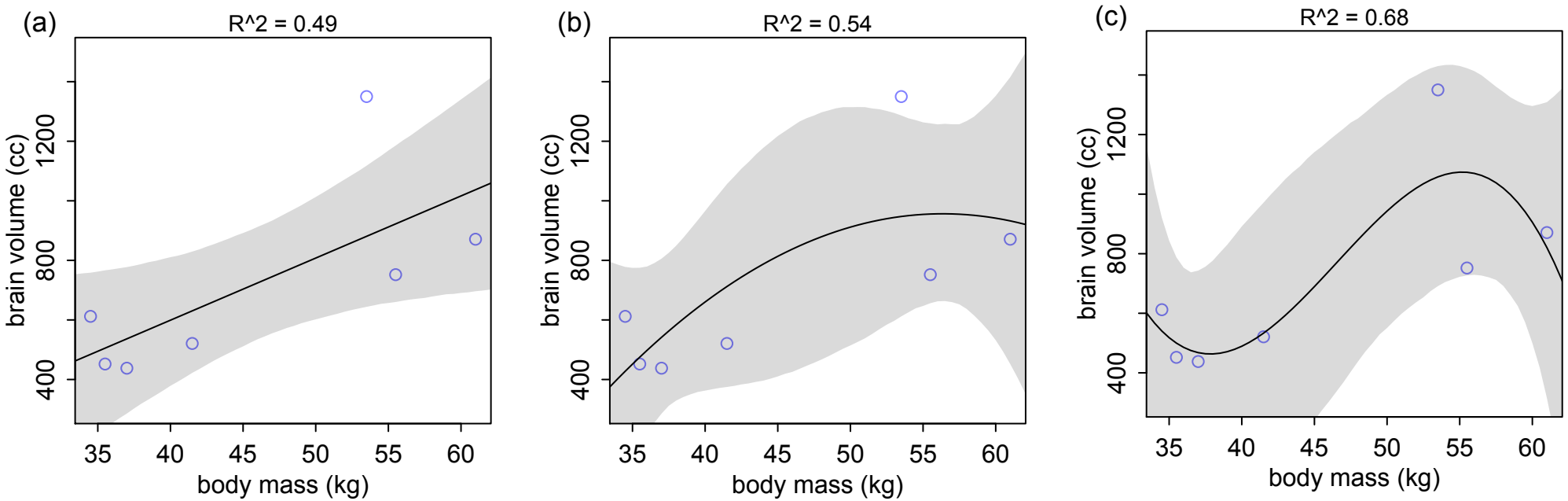

(a)
R^2 = 0.49
brain volume (cc)
400
800
1200
35
40
45
50
55
60
body mass (kg)
(b)
R^2 = 0.54
brain volume (cc)
400
800
1200
35
40
45
50
55
60
body mass (kg)
(c)
R^2 = 0.68
brain volume (cc)
400
800
1200
35
40
45
50
55
60
body mass (kg)

Figure 6.3

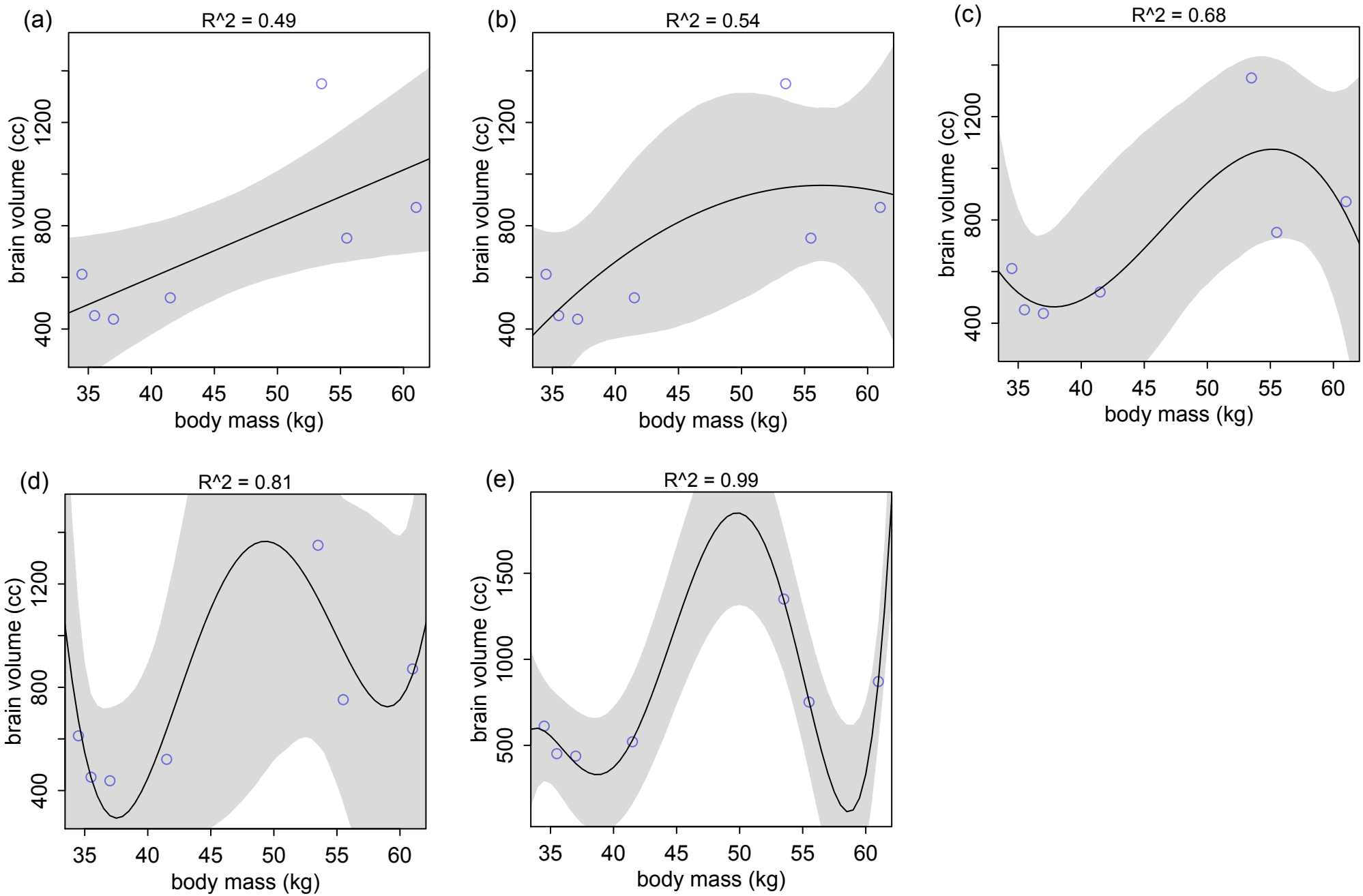

(a)
R^2 = 0.49
brain volume (cc)
400
800
1200
35
40
45
50
55
60
body mass (kg)
(b)
R^2 = 0.54
brain volume (cc)
400
800
1200
35
40
45
50
55
60
body mass (kg)
(c)
R^2 = 0.68
brain volume (cc)
400
800
1200
35
40
45
50
55
60
body mass (kg)
(d)
R^2 = 0.81
brain volume (cc)
400
800
1200
35
40
45
50
55
60
body mass (kg)
(e)
R^2 = 0.99
brain volume (cc)
500
1000
1500
35
40
45
50
55
60
body mass (kg)

Figure 6.3

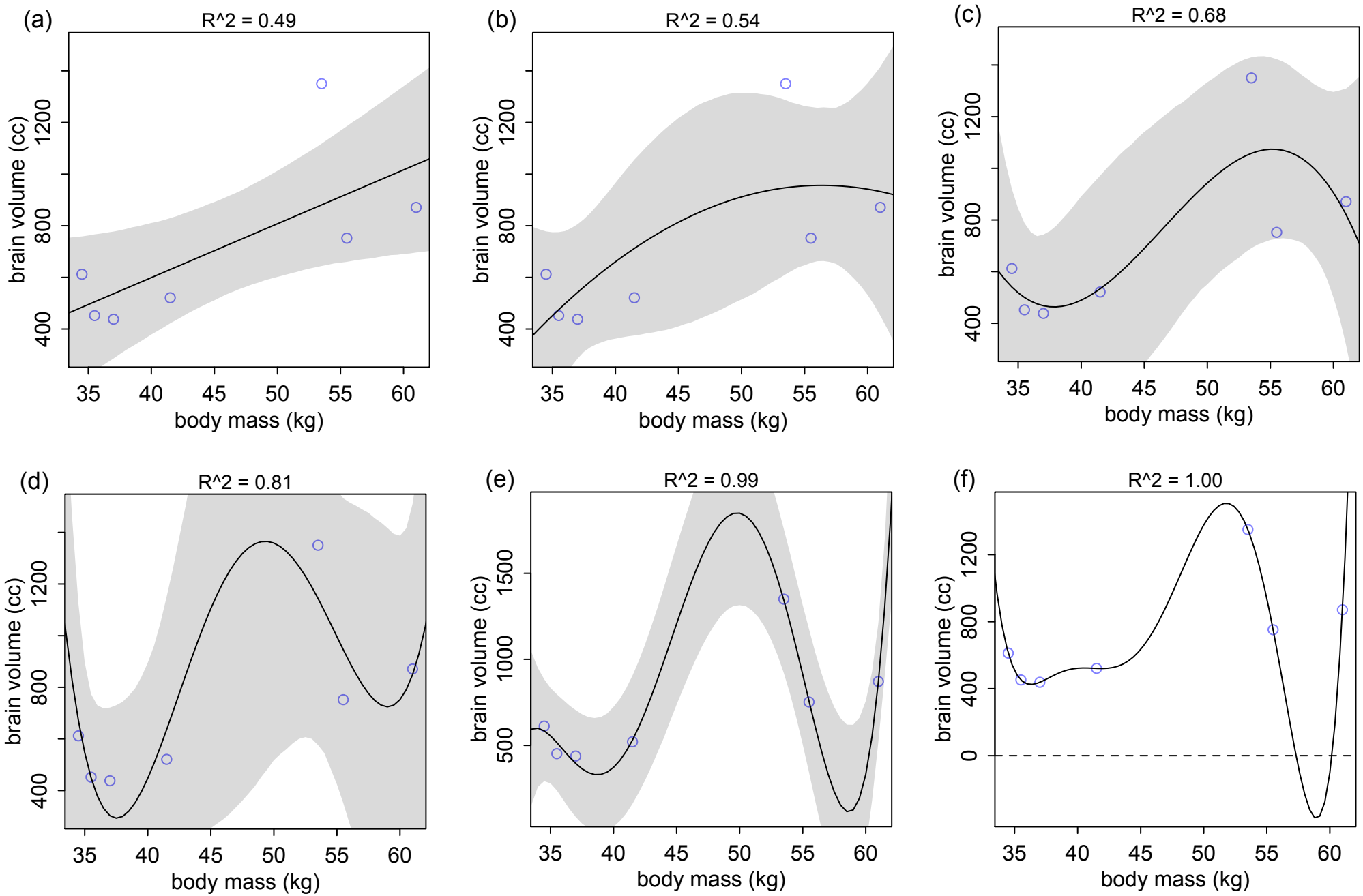

(a)
R^2 = 0.49
(b)
R^2 = 0.54
(c)
R^2 = 0.68
(d)
R^2 = 0.81
(e)
R^2 = 0.99
(f)
R^2 = 1.00
brain volume (cc)
body mass (kg)
35
40
45
50
55
60
400
800
1200
500
1000
1500
0

Figure 6.3

Underfitting
Insensitive to
exact data

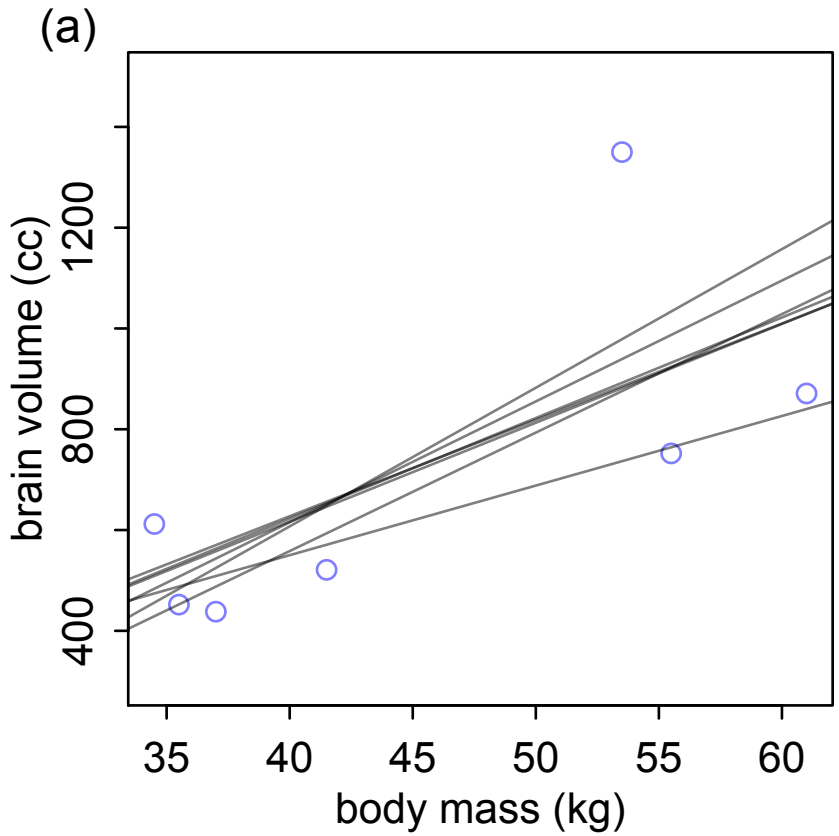


Figure 6.5

Underfitting
Insensitive to
exact data

Overfitting
Very sensitive to
exact data

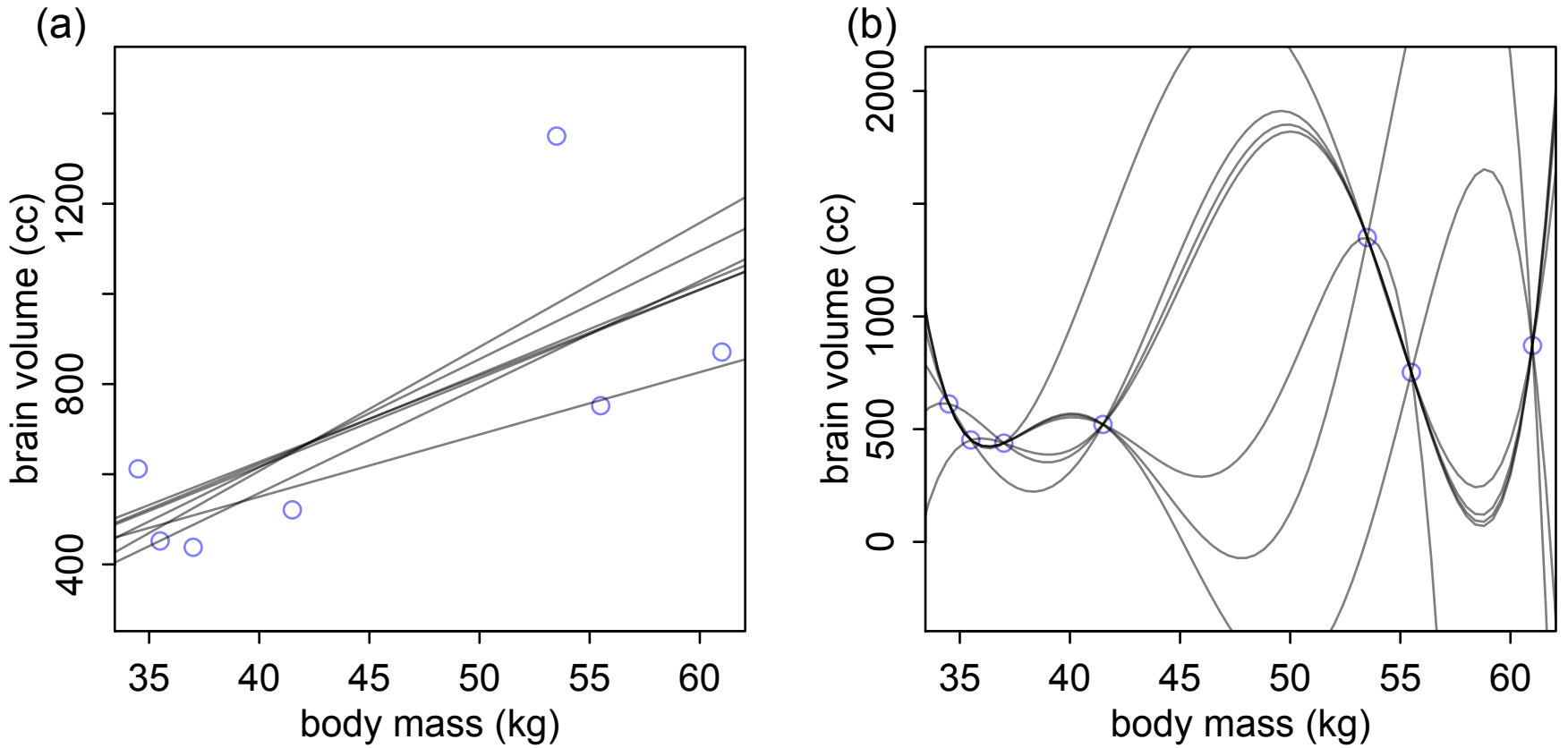


Figure 6.5

Importance of being *regular*

- Want the *regular* features of the sample
- Strategies
  - Cross-validation
  - Regularizing priors (penalized likelihood)
  - Information criteria
  - Science! (iterative group learning)
- Proper approach depends upon purpose

The road to AIC/DIC/WAIC

- What's a good target?
  - average correct?
  - average probability correct?
  - average log probability correct?
- How measure distance from the target?
- How can we estimate that distance?
- How can we adjust that estimate to account for overfitting?

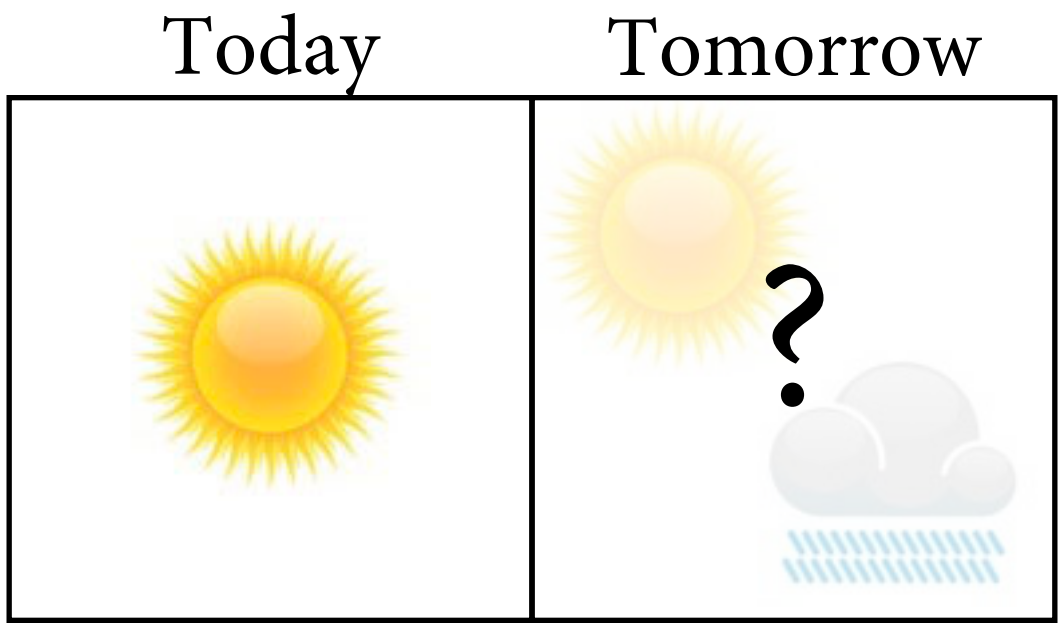
How far from truth?

- *Truth*: The real joint probability of events
  - *Truth* defines probability distribution
  - *Model* defines another
- Need a way to measure distance of a model from truth
- Distance needs to accommodate complexity of prediction task

Day	1	2	3	4	5	6	7	8	9	10
Observed										
Points										
Current	−1	−1	−1	−0.6	−0.6	−0.6	−0.6	−0.6	−0.6	−0.6
Newcomer	−5	−5	−5	0	0	0	0	0	0	0

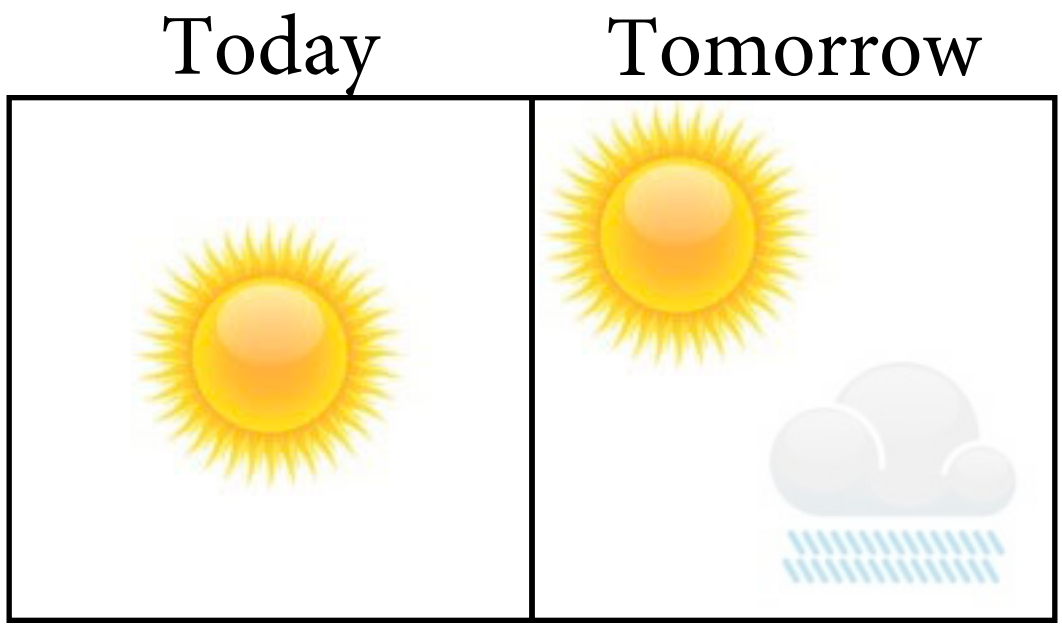
Information theory

- *Information*: Reduction in uncertainty caused by learning an outcome.



Information theory

- *Information*: Reduction in uncertainty caused by learning an outcome.



	Today	Tomorrow	
Los Angeles		?	
Glasgow		?	
New York		?	

Information theory

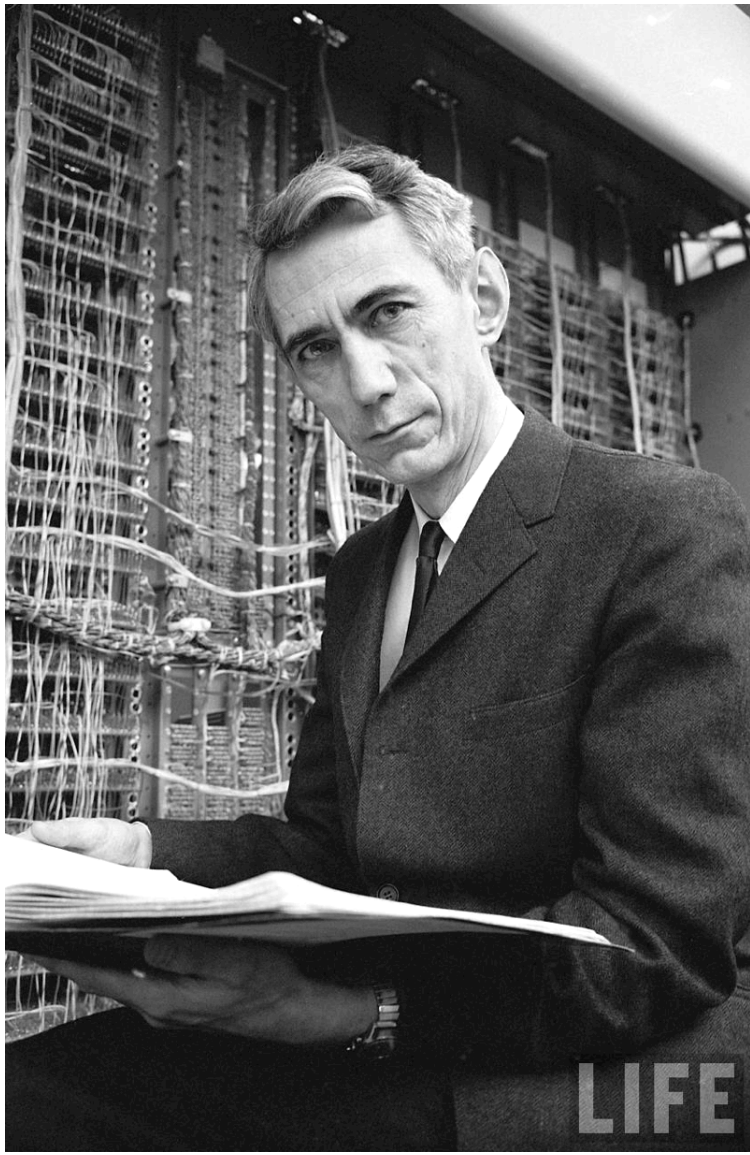
- *Information*: Reduction in uncertainty caused by learning an outcome.
- How to quantify *uncertainty*? Should be:
  1. Continuous
  2. Increasing with number of possible events
  3. Additive
- These criteria intuitive, but effectiveness is why we keep using them
  - Like Bayes: intuitive, but effectiveness is reason to use

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.



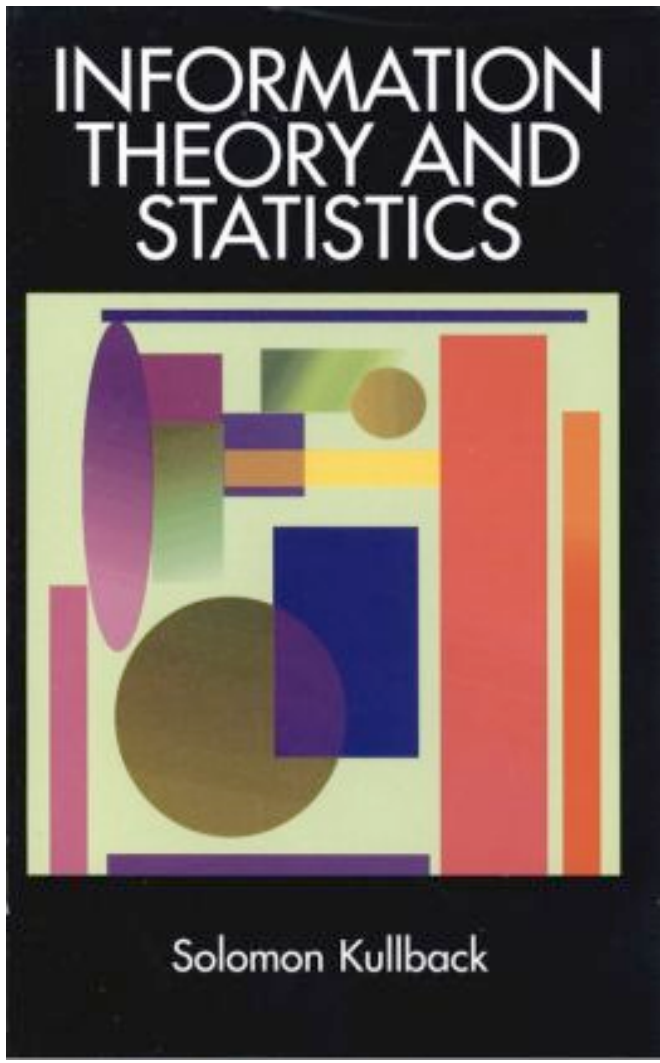
Shannon (1916–2001)

Entropy to accuracy

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

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

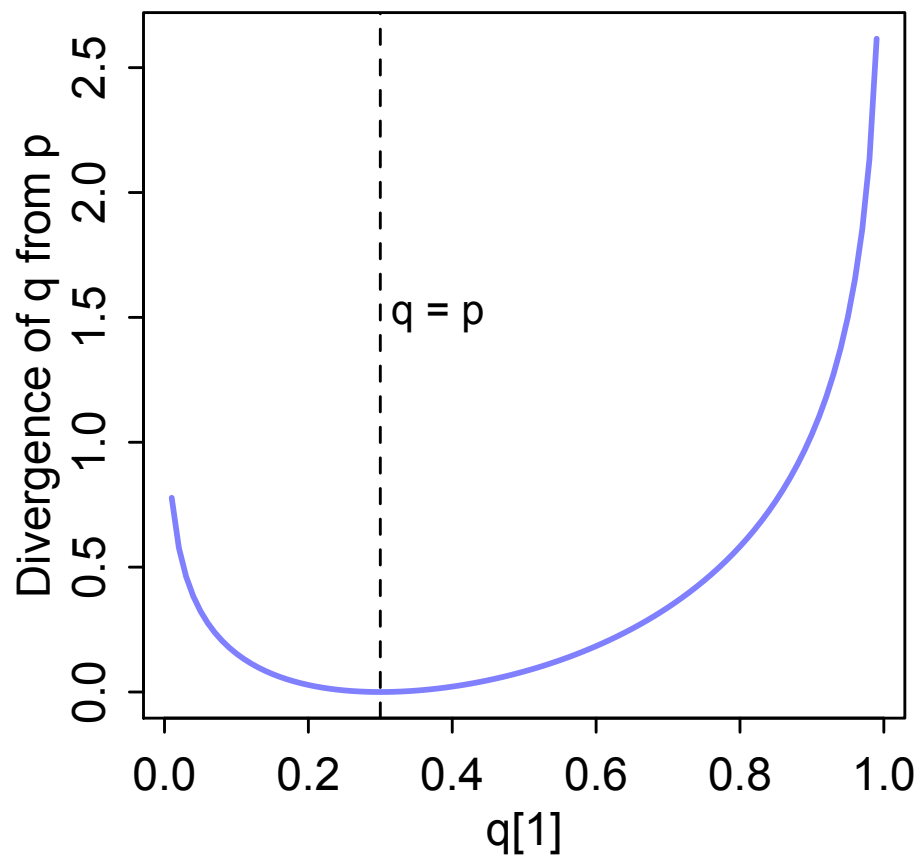
Distance from q to p is the average difference in log-probability.



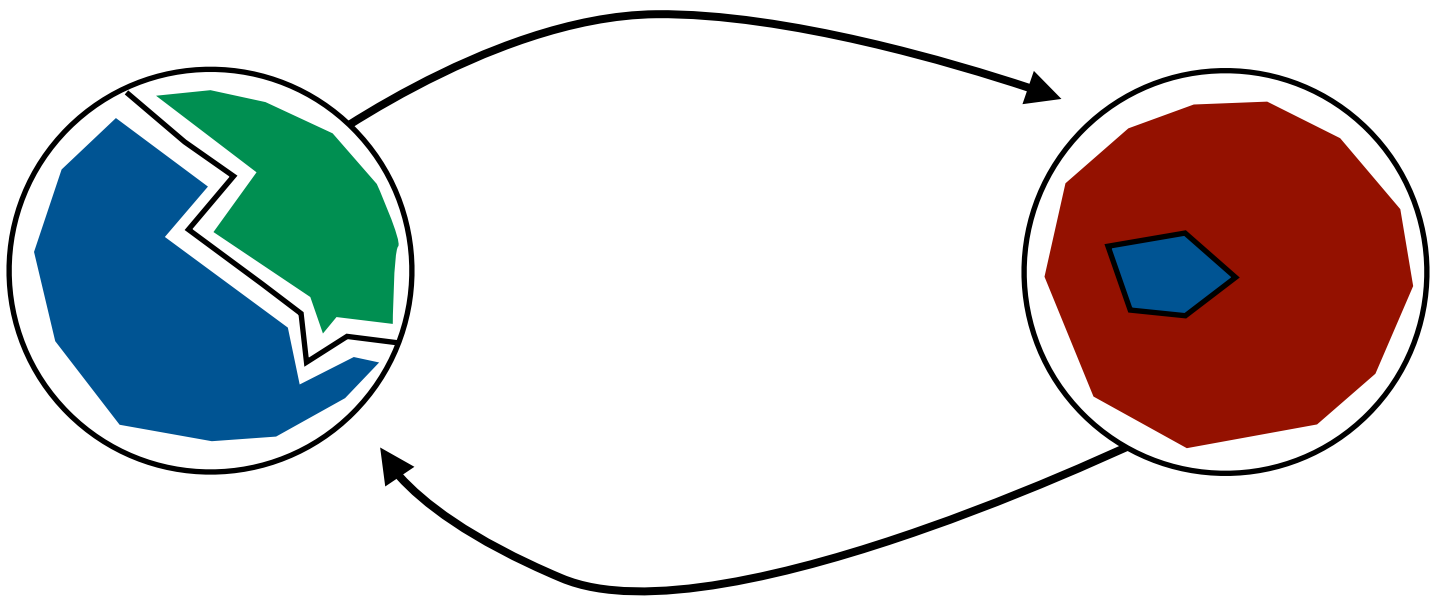
Entropy to accuracy

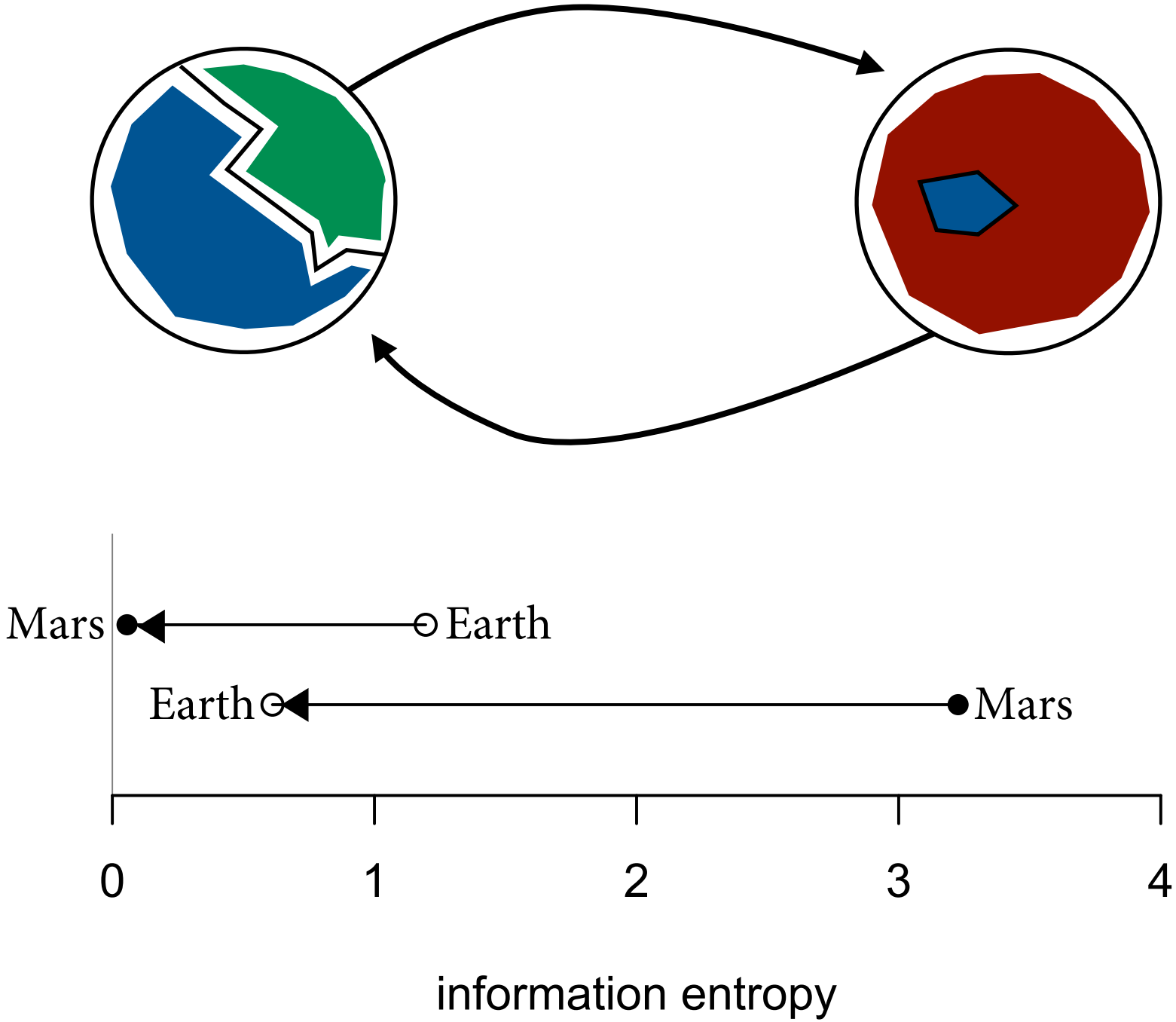
$$D_{\mathrm{KL}}(p,q) = \sum_i p_i \big(\log(p_i) - \log(q_i) \big)$$

```
p <- c(0.3,0.7)

DKL <- function(p,q)
    sum(p*(log(p)-log(q)))

q1seq <- seq(from=0.01,to=0.99,by=0.01)
DKLseq <- sapply(q1seq,
    function(q1) DKL(p,c(q1,1-q1)) )
plot( q1seq , DKLseq )
```



Direction matters




Mars
Earth
Earth
Mars
0
1
2
3
4
information entropy

Estimating divergence

- How to estimate D_{KL}?

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

"*truth*"

model

Estimating divergence

- Don't know p! Don't need it. Focus on difference between two approximating models:

$$D_{\mathrm{KL}}(p, q) - D_{\mathrm{KL}}(p, r) = -\sum_i p_i(\log q_i - \log p_i) - \left(-\sum_i p_i(\log r_i - \log p_i)\right)$$

model A

model B

Estimating divergence

- Don't know p! Don't need it. Focus on difference between two approximating models:

$$D_{\mathrm{KL}}(p,q) - D_{\mathrm{KL}}(p,r) = -\sum_i p_i(\log q_i - \log p_i) - \left(-\sum_i p_i(\log r_i - \log p_i)\right)$$

$$= -\sum_i p_i(\log q_i - \log r_i) = -(\mathrm{E}\log q_i - \mathrm{E}\log r_i)$$

model A (↑ q)

model B (↑ r)

- log-probability scores (deviance e.g.) provide estimate of $\mathrm{E}\log q_i$

Deviance (classic estimate)

$$D(q) = -2 \sum_i \log(q_i)$$

- How *bad* the model is, not how *good*
- Compute it:
  - Compute log probability of each observation
  - Sum all of these log probabilities
  - Multiply by –2
- Common to use MAP estimates for probabilities, but can use entire posterior
  - Will do so later, when compute WAIC as estimate of deviance

The road to AIC/DIC/WAIC

- ✓ What's a good prediction?
- ✓ How far is the model from the target?
- ✓ How can we estimate that distance?
- • How can we adjust that estimate to account for overfitting?

Everybody overfits

- A meta-model of forecasting:
  - Two samples: *training* and *testing*, size N
  - Fit model to *training* sample, get D_{train}
  - Use fit to *training* to compute D_{test}
  - Difference $D_{\text{test}} - D_{\text{train}}$ is overfitting

Everybody overfits

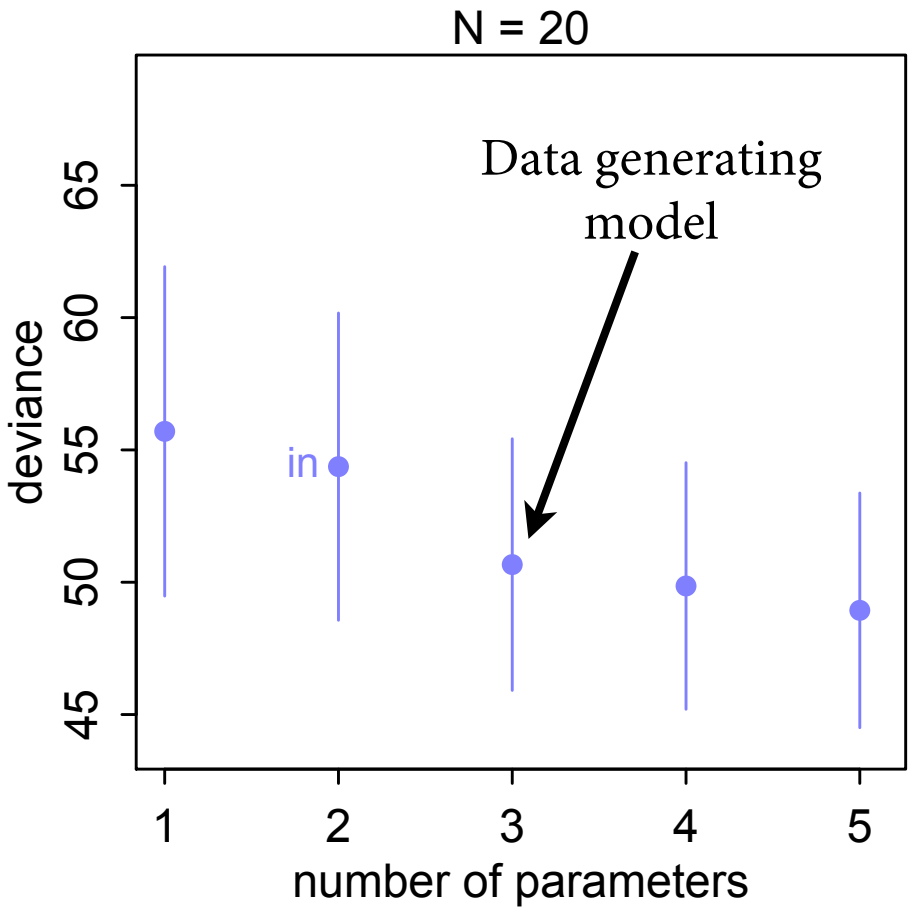
Data generating model:

$$y_i \sim \text{Normal}(\mu_i, 1)$$
$$\mu_i = (0.15)x_{1,i} - (0.4)x_{2,i}$$

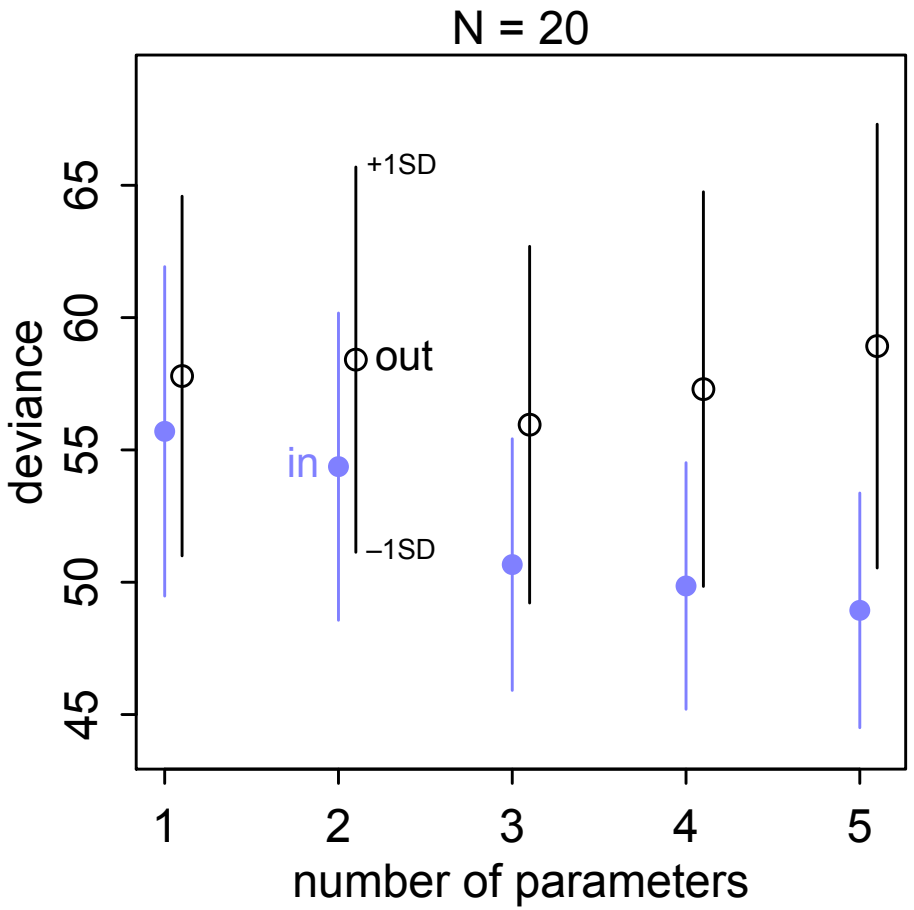
Models fit to data:
(flat priors)

$$\mu_i = \alpha$$
$$\mu_i = \alpha + \beta_1 x_{1,i}$$
$$\mu_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i}$$
$$\mu_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i}$$
$$\mu_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i}$$

Everybody overfits



Everybody overfits



Everybody overfits