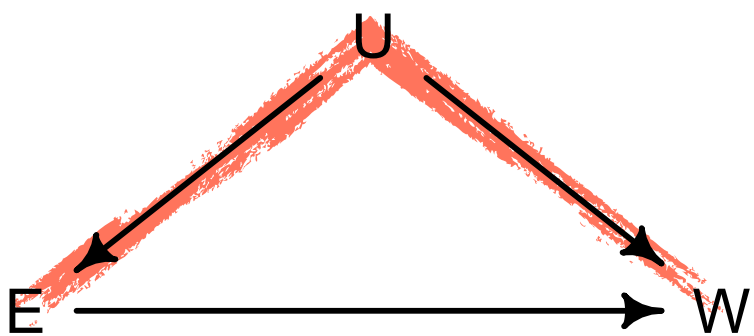
Statistical Rethinking
Winter 2019

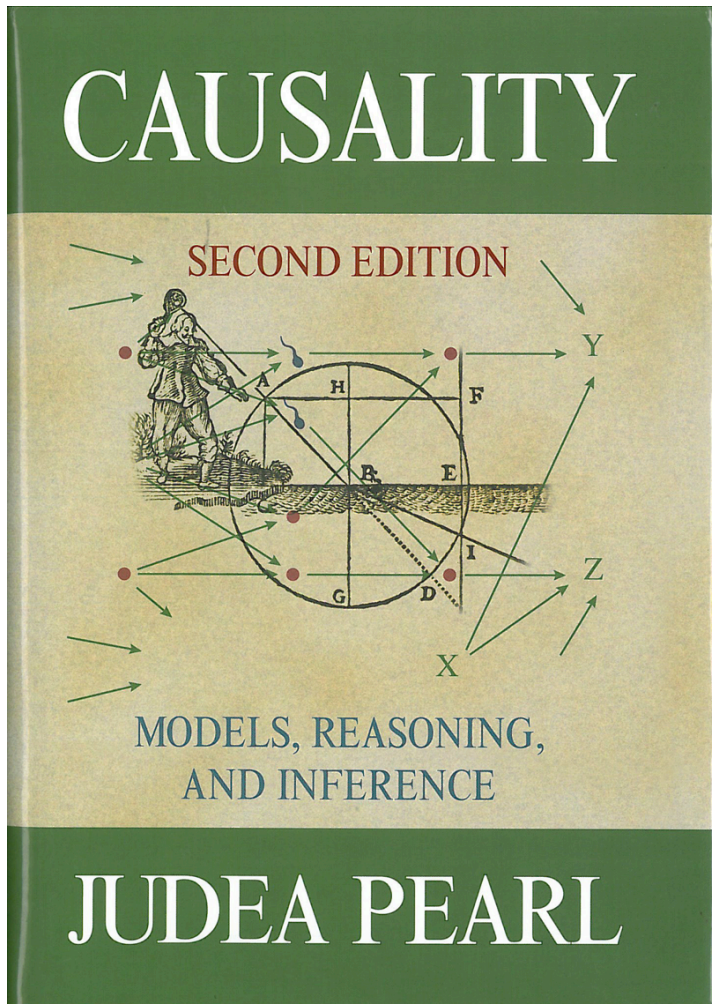
Lecture 07 / Week 4

Back-door Paths & Ulysses' Compass

Shutting the back door

- What ties these examples together:
- The **back-door criterion**: Confounding caused by existence of open back door paths from X to Y
- If you know your elements, you know how to open/close each of them





The Fork

Open unless you
condition on Z

The Pipe

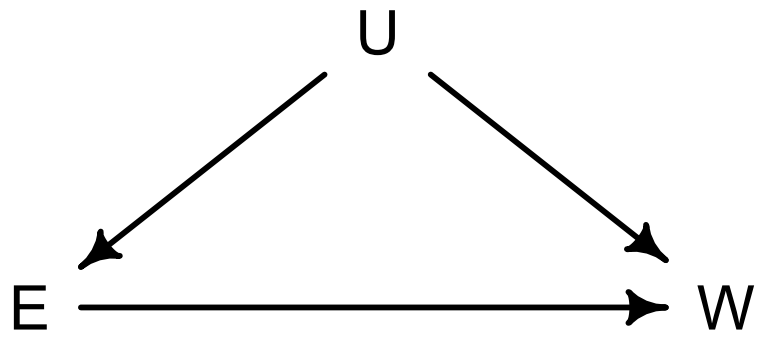
Open unless you
condition on Z

The Collider

Closed until you
condition on Z

The Descendant

Conditioning on A is
like conditioning on Z

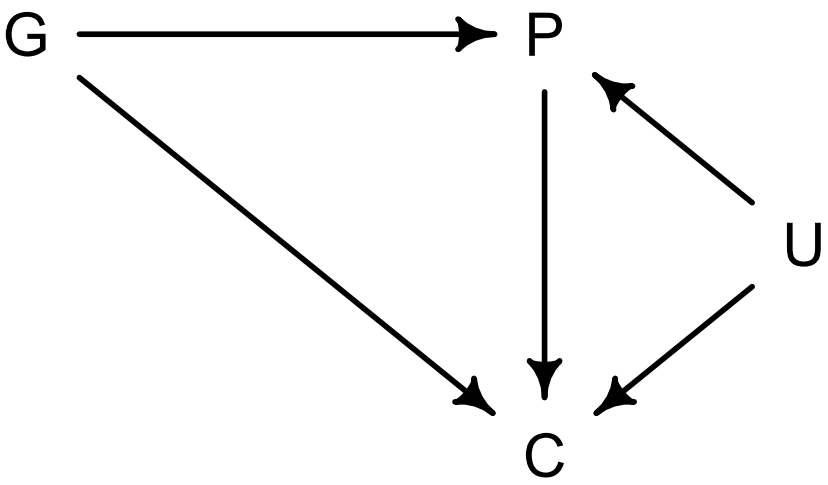


Two paths from E to W:

(1) E → W

(2) E ← U → W

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



3 paths from G to C:

(1) G → C

(2) G → P → C

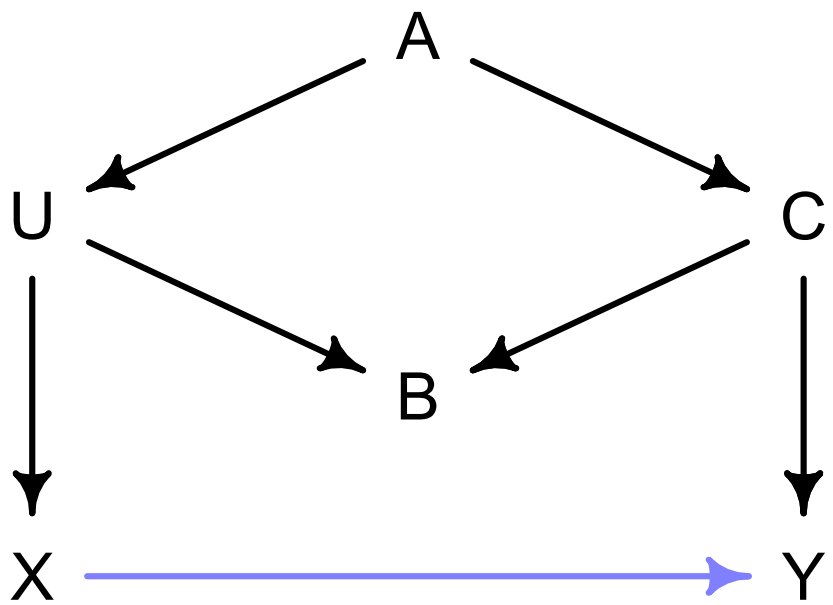
(3) G → P ← U → C

Condition on P:

Closes (2) but opens (3)

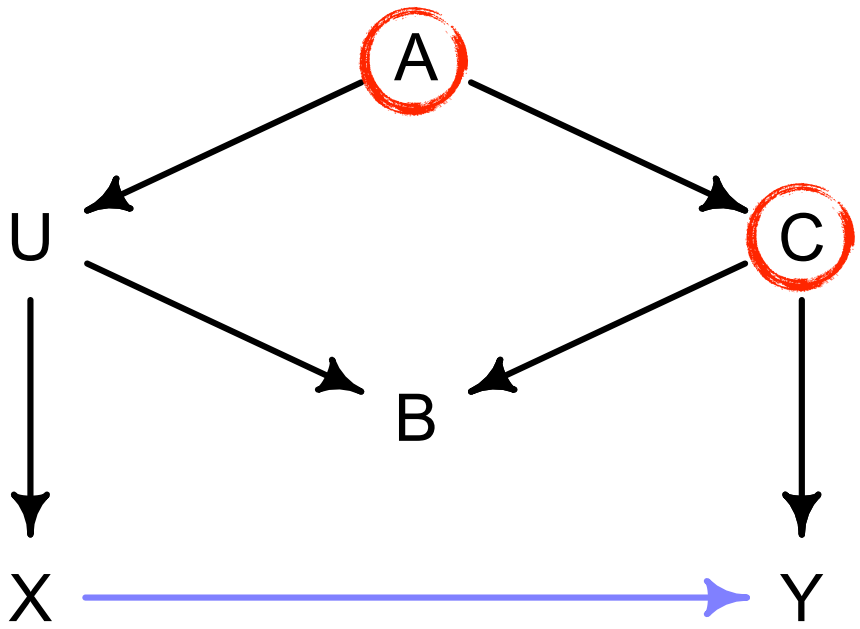
Something more interesting

- Which variables, if any, should you condition on to infer X → Y?
- Procedure: (1) Find all paths. (2) Open/close as necessary.



Something more interesting

- Which variables, if any, should you condition on to infer X → Y?
- Condition on A or C. Do not condition on B.



(1) $X \leftarrow U \leftarrow A \rightarrow C \rightarrow Y$

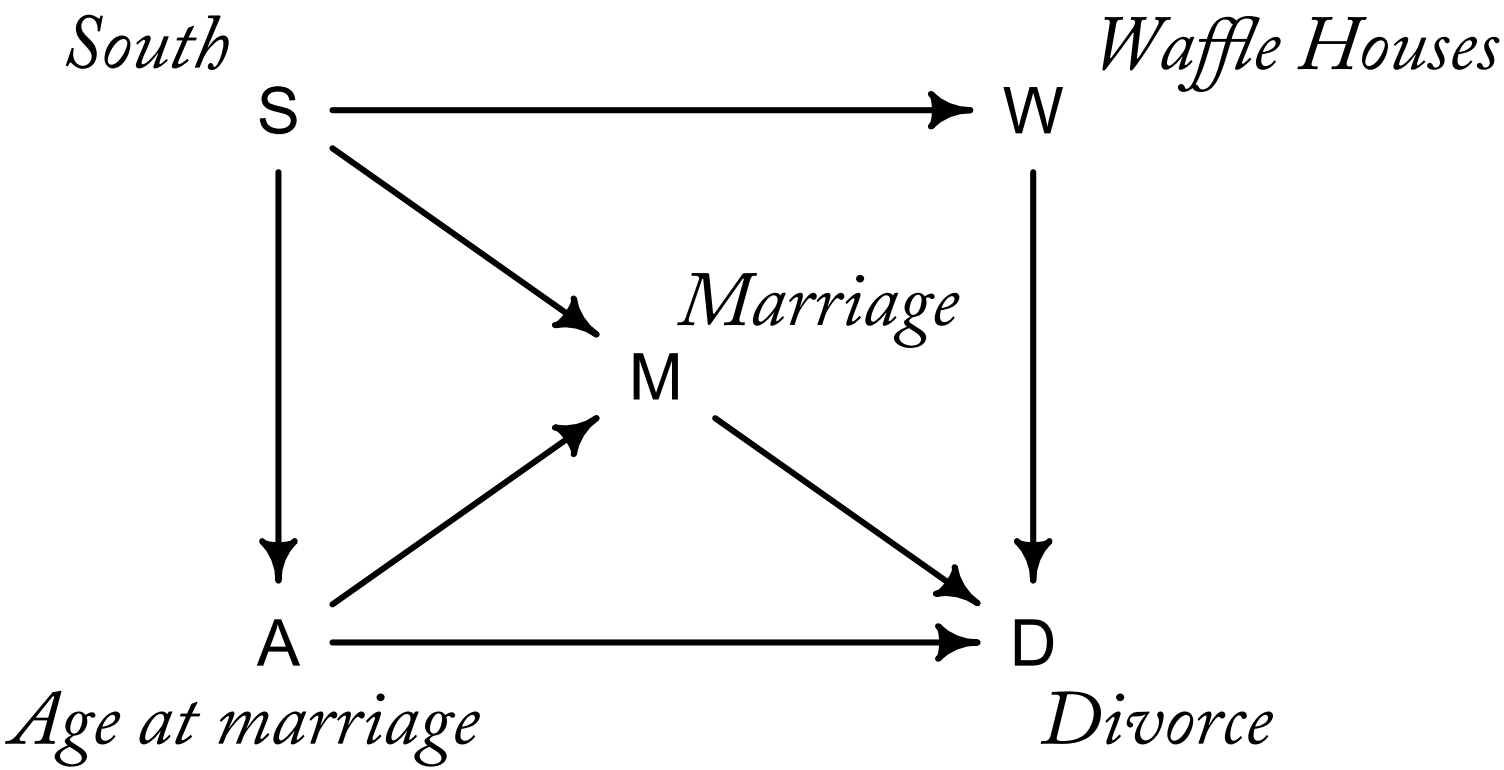
This path is open.

(2) $X \leftarrow U \rightarrow B \leftarrow C \rightarrow Y$

This path is closed.

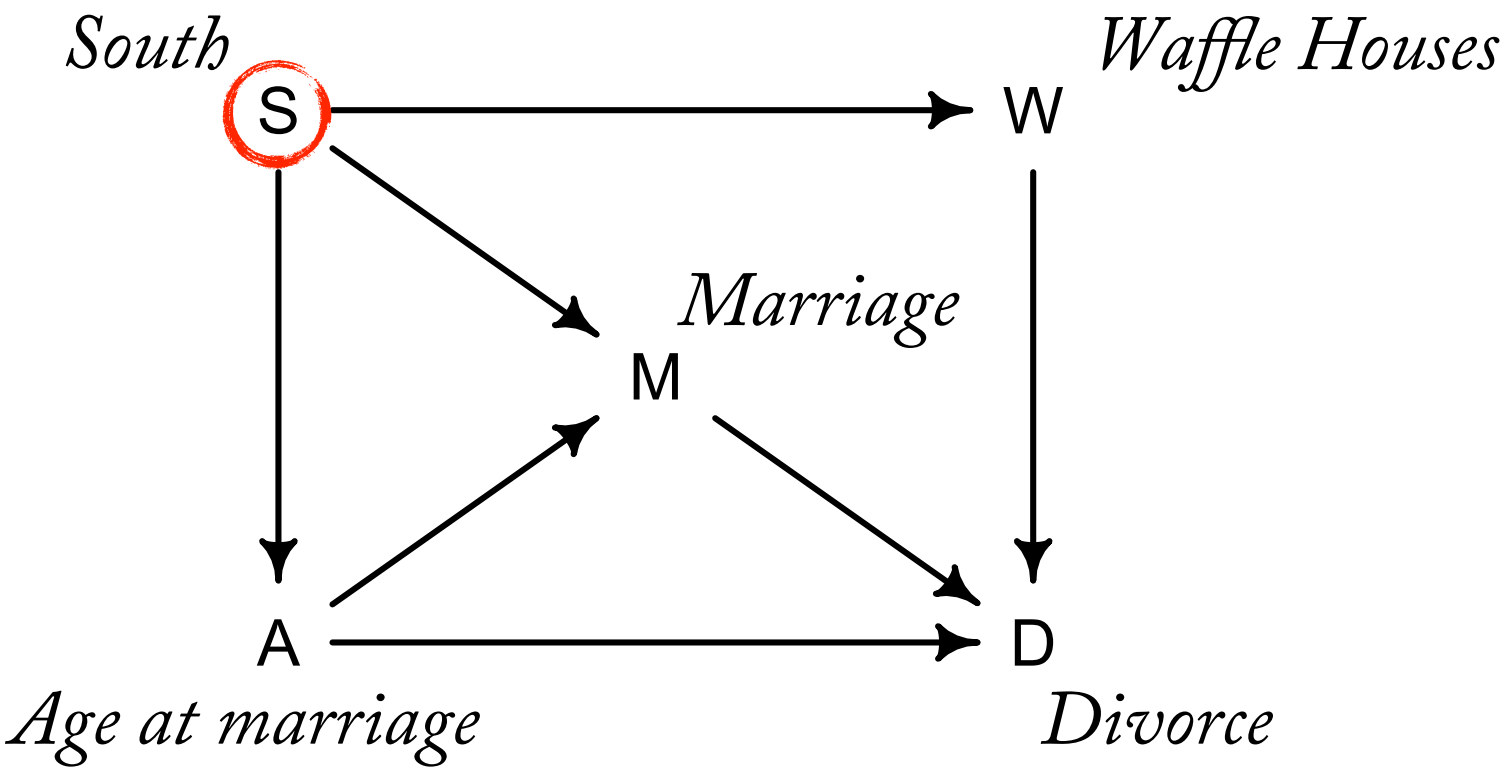
Waffles Requiem

- Remember the waffles.
- Which to control to infer W → D?



Waffles Requiem

- Remember the waffles.
- Which to control to infer W → D?



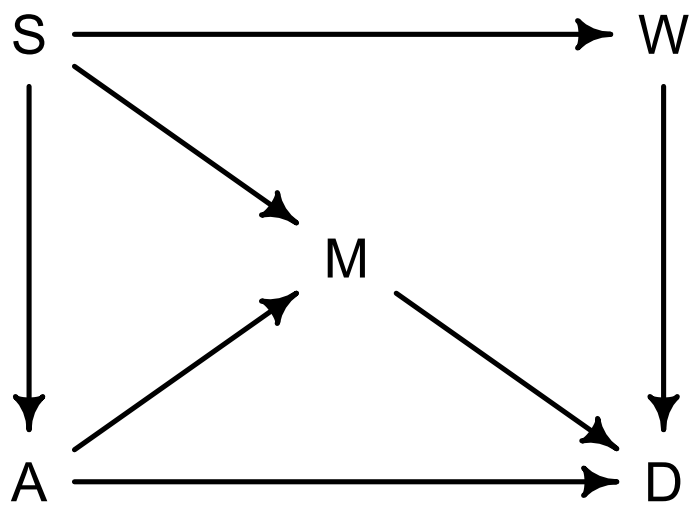
Implied conditional independence

- Given DAG, can test some implications

```
impliedConditionalIndependencies( dag_6.2 )
```

R code 6.36

```
A _||_ W | S
D _||_ S | A, M, W
M _||_ W | S
```



(1) A and W independent, conditioning on S
(2) D and S independent, conditioning on A, M, & W
(3) M and W independent, conditioning on S

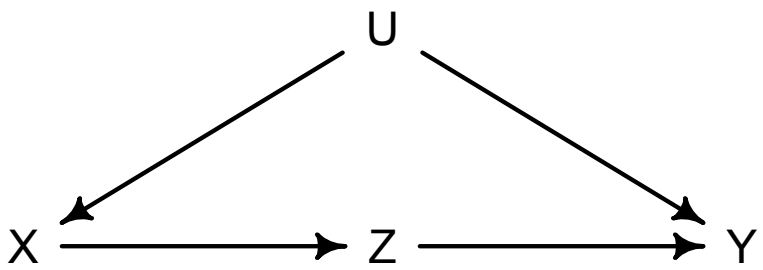
Causal inference hard but possible

- Demonstrate capable of inferring cause
- Experiments not required!
- Experiments not always practical or ethical
  - Disease, evolution, development, dynamics of popular music, global climate, war
- Experiments must choose an intervention
  - Interventions influence many variables at once
  - Experimentally manipulate obesity?

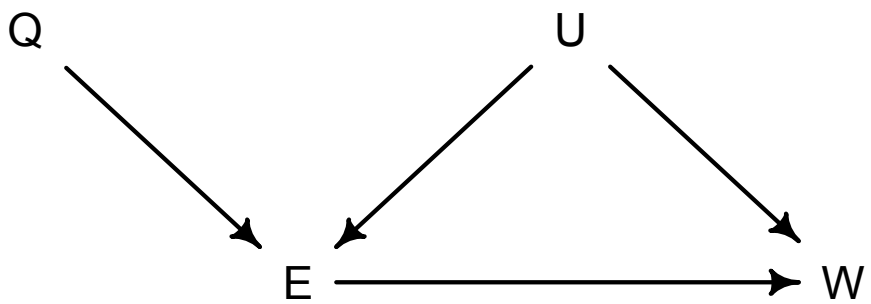
David Hume (1711–1776) rates your DAG 12/10

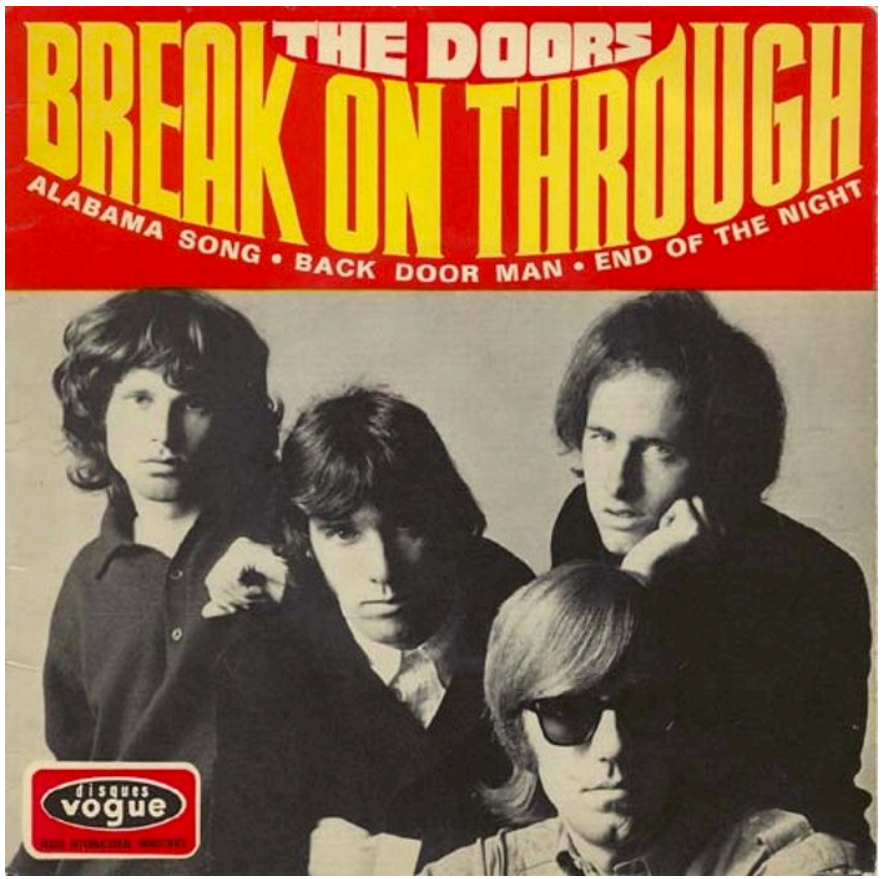
More than the Back Door

- Closing back doors is not the only option
- Front-door criterion



- Instrumental variables



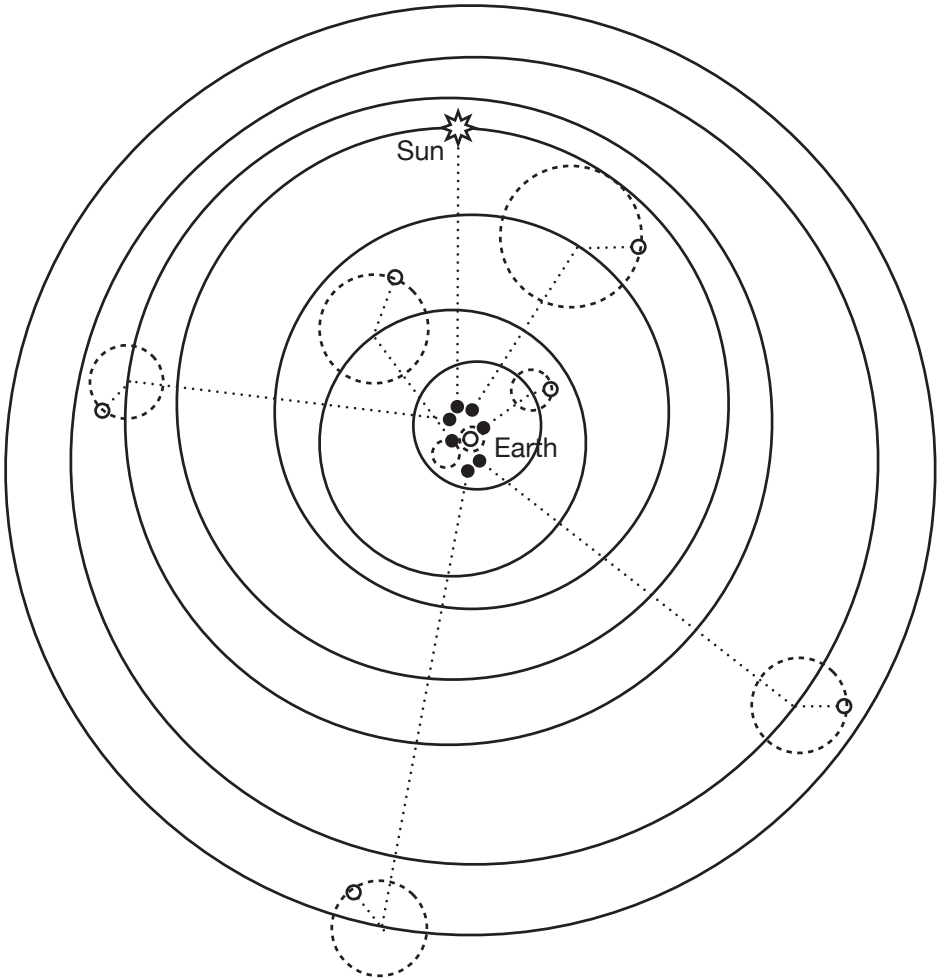


Directed Acyclic Gaffes

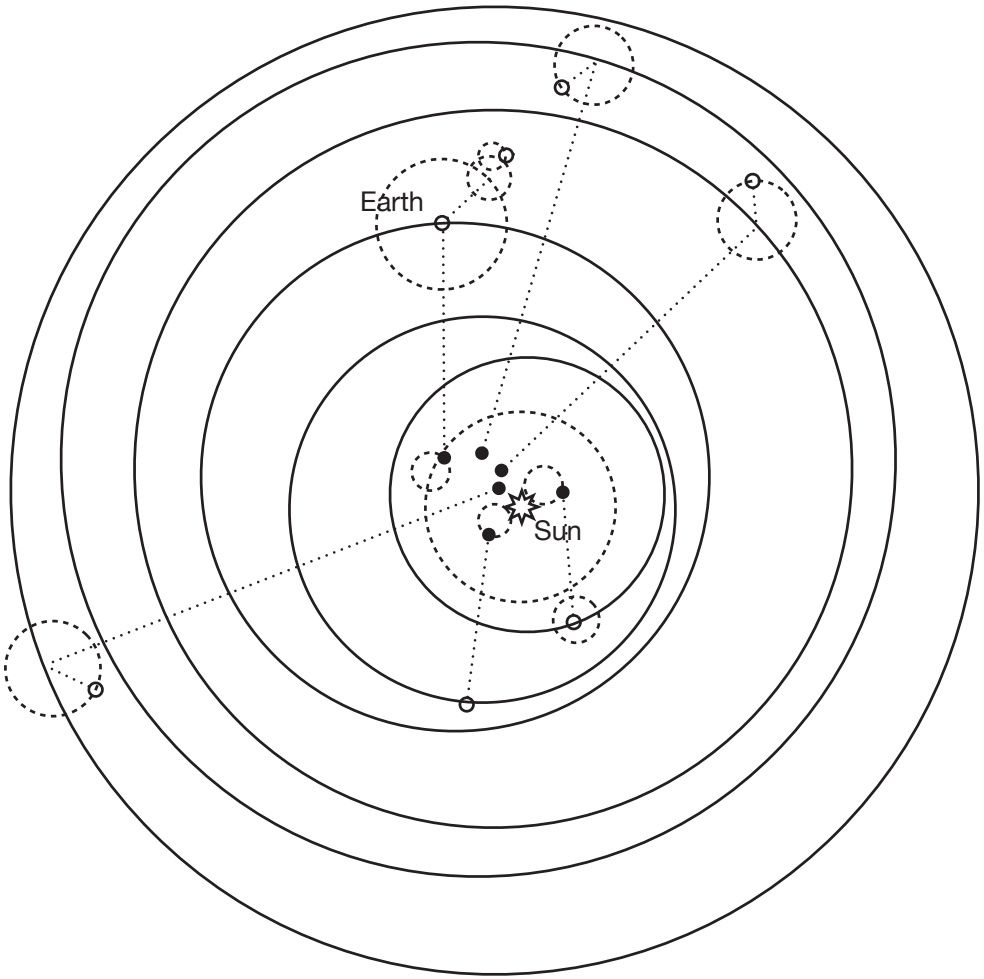
- Don't get cocky
- DAGs are small world constructs
- Residual confounding:
  - Misclassification
  - Measurement error
  - Missingness
- DAGs can accommodate these problems, but maybe tell us there are no solutions
- We will see some solutions in later week
- Eventually need *real* models of the system

Mikołaj Kopernik (1473–1543)

Ptolemaic Model



Copernican Model

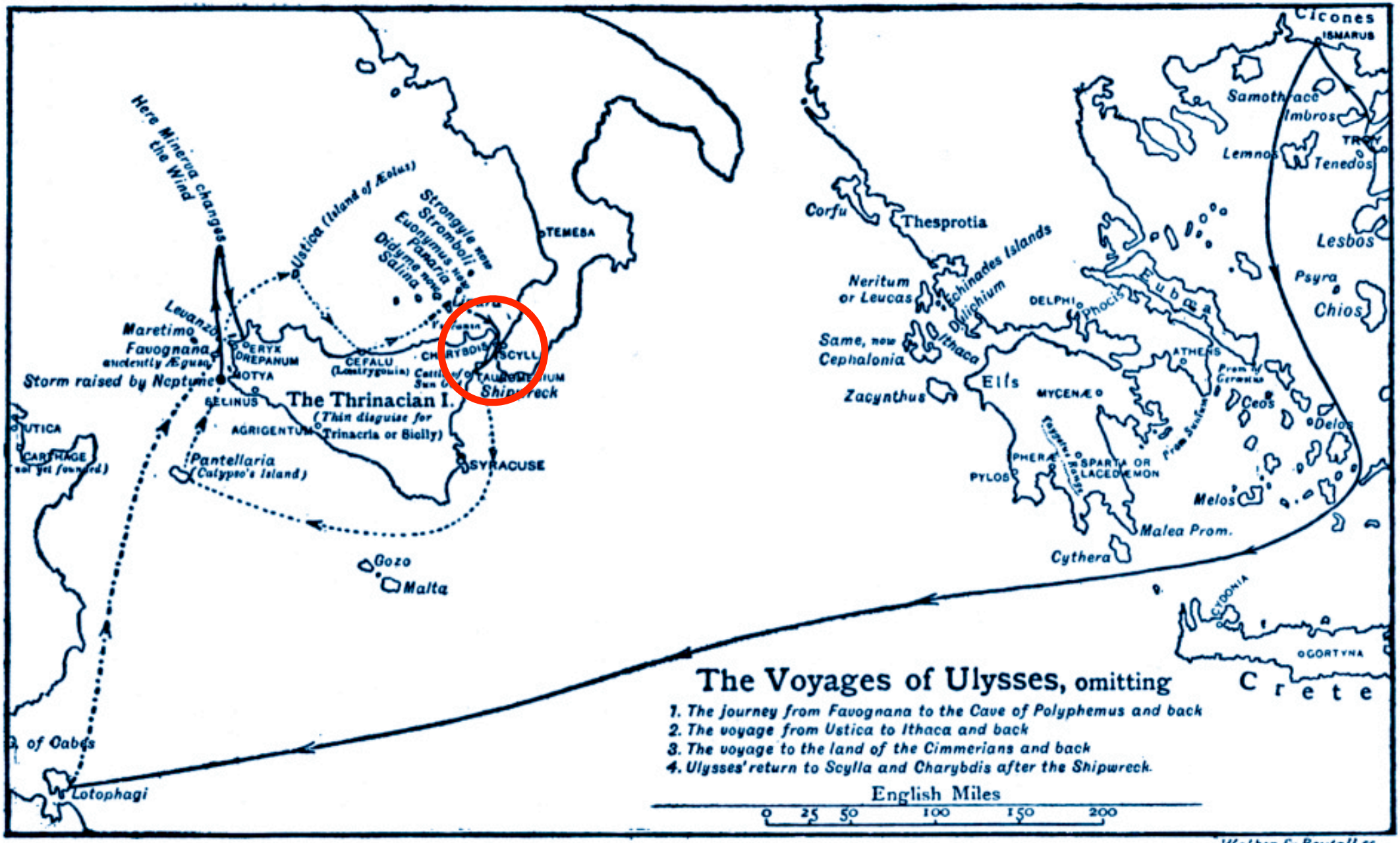


Ockham's Razor?

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

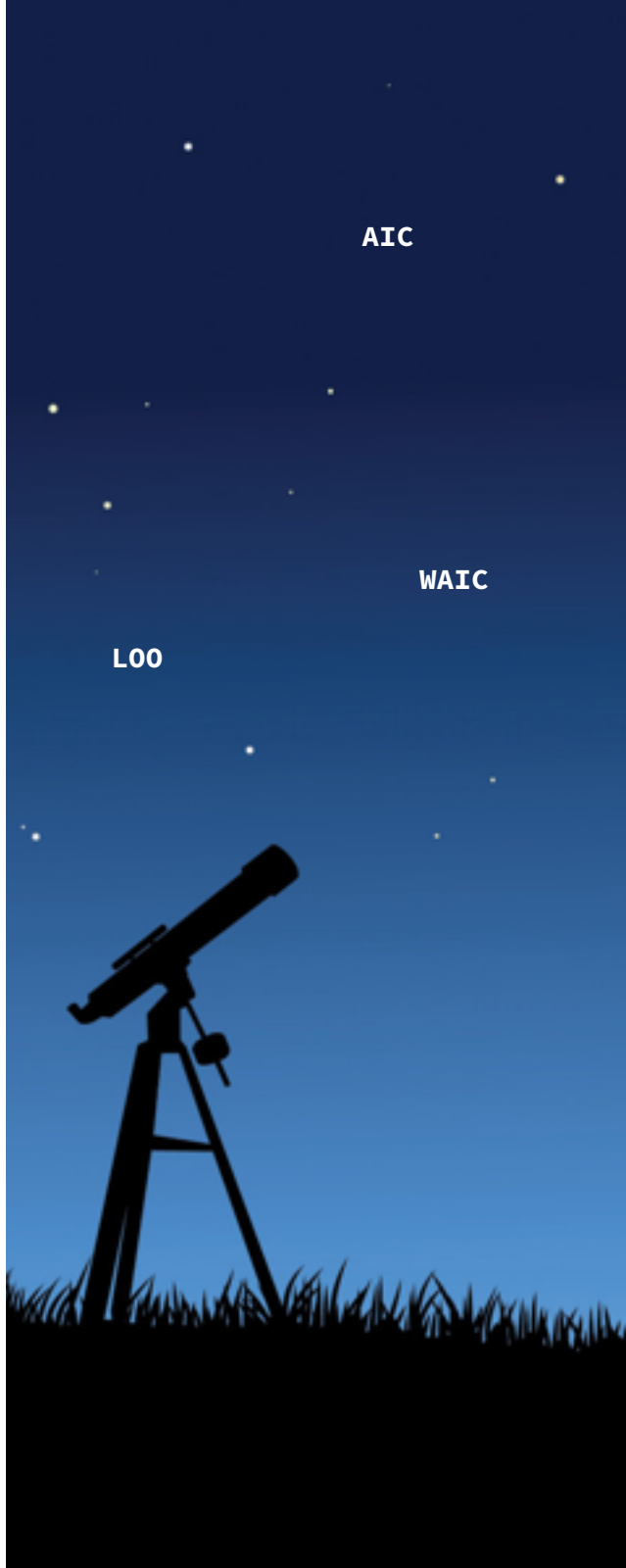
Stargazing

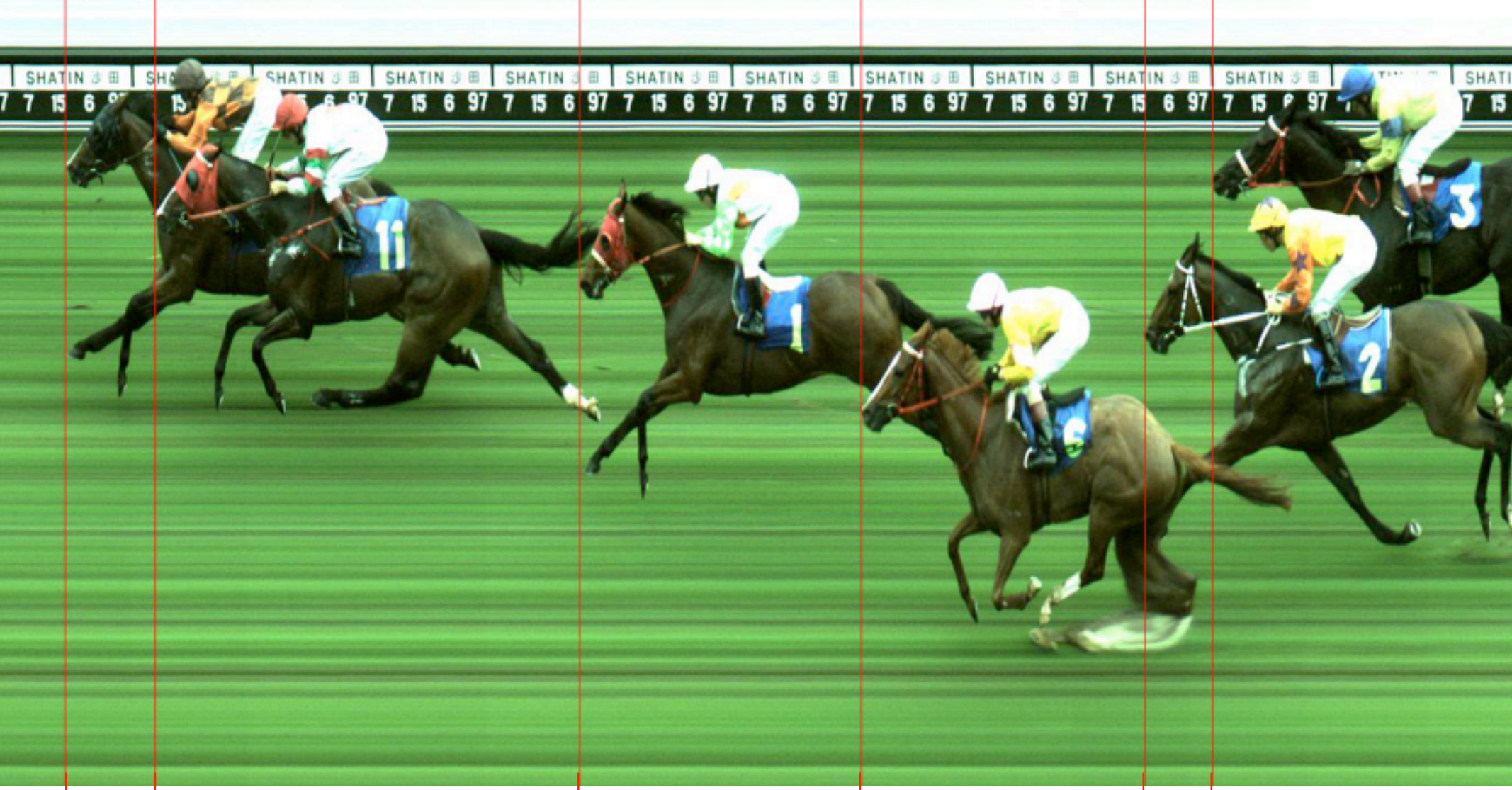
- *Stargazing*: Using asterisks ($p < 0.05$) to select variables
- Arbitrary 5% is arbitrary
- *p*-values do not regulate accuracy

```
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

- 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



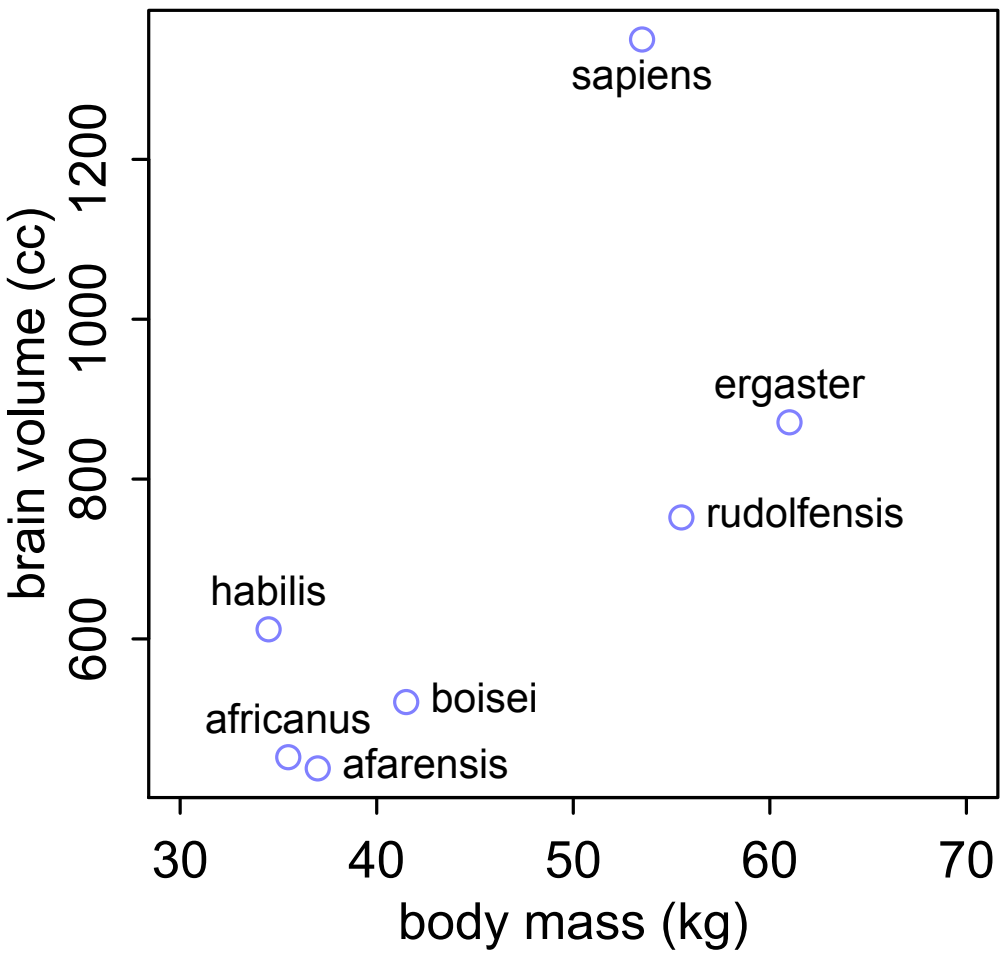

A
B
C
D
E
F

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

The Problem with Parameters



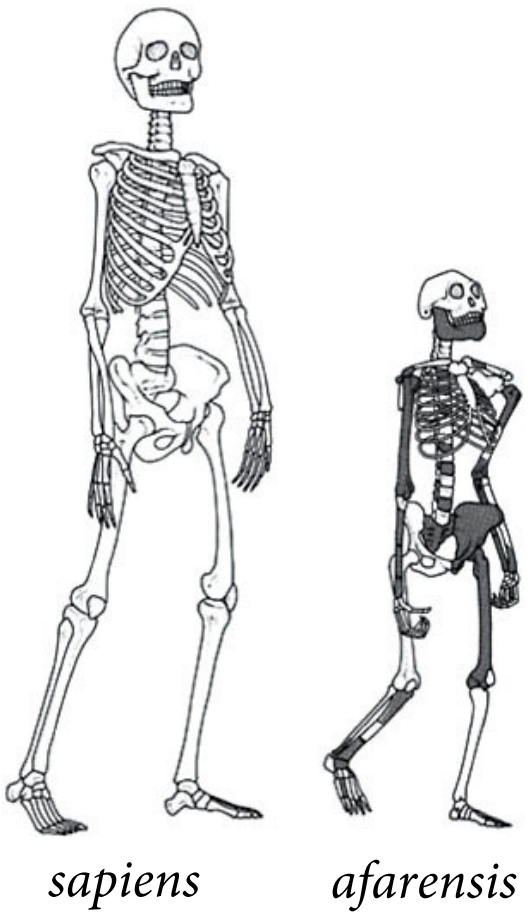


Figure 7.2

Variance "explained"

- Most common & misused measure of model fit is R-squared:

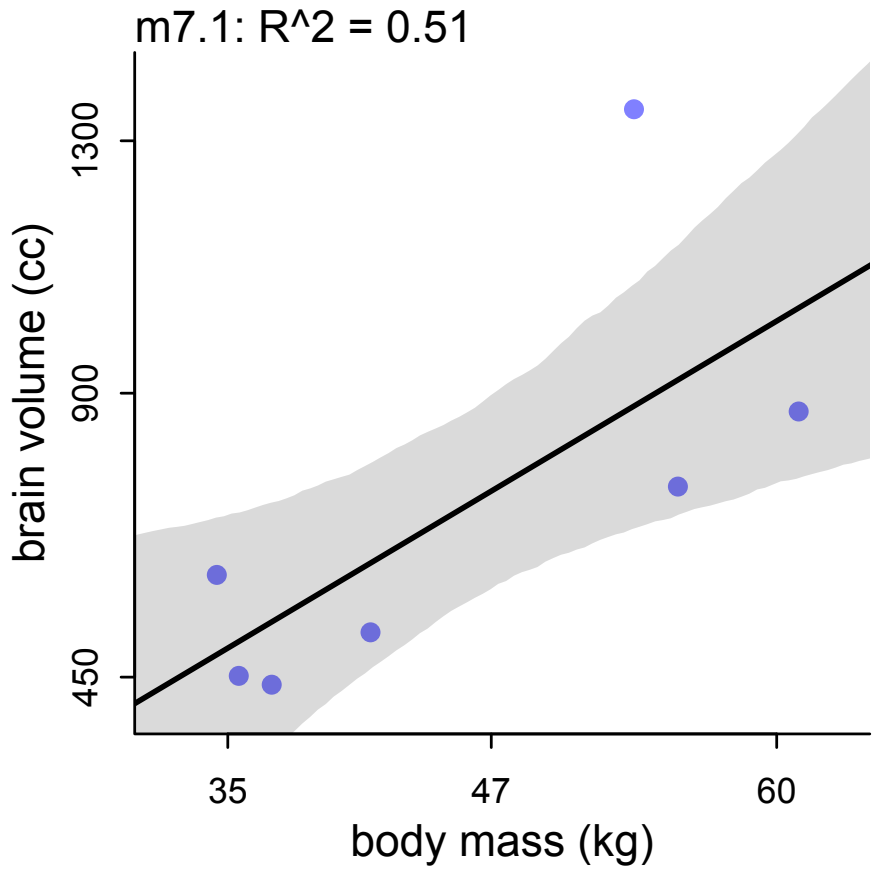
$$R^2 = \frac{\text{var}(\text{outcome}) - \text{var}(\text{residuals})}{\text{var}(\text{outcome})} = 1 - \frac{\text{var}(\text{residuals})}{\text{var}(\text{outcome})}$$

- Interpretation: Proportion of variance explained
- How does R-squared behave?

Hominin brains

- Simplest model:

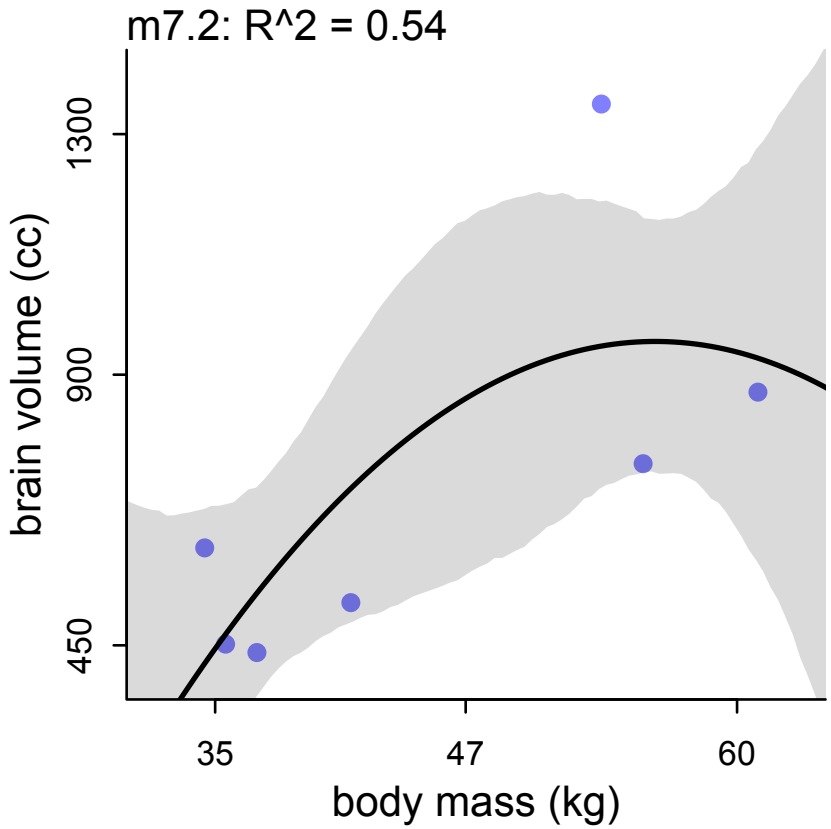
$$b_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta m_i$$
$$\alpha \sim \text{Normal}(0.5, 1)$$
$$\beta \sim \text{Normal}(0, 10)$$
$$\sigma \sim \text{Log-Normal}(0, 1)$$



Hominin brains

- Why not parabola?

$$
\begin{aligned}
b_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_1 m_i + \beta_2 m_i^2 \\
\alpha &\sim \text{Normal}(0.5, 1) \\
\beta_j &\sim \text{Normal}(0, 10) \qquad \text{for } j = 1..2 \\
\sigma &\sim \text{Log-Normal}(0, 1)
\end{aligned}
$$



Hominin brains

- Why not higher order polynomials?

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

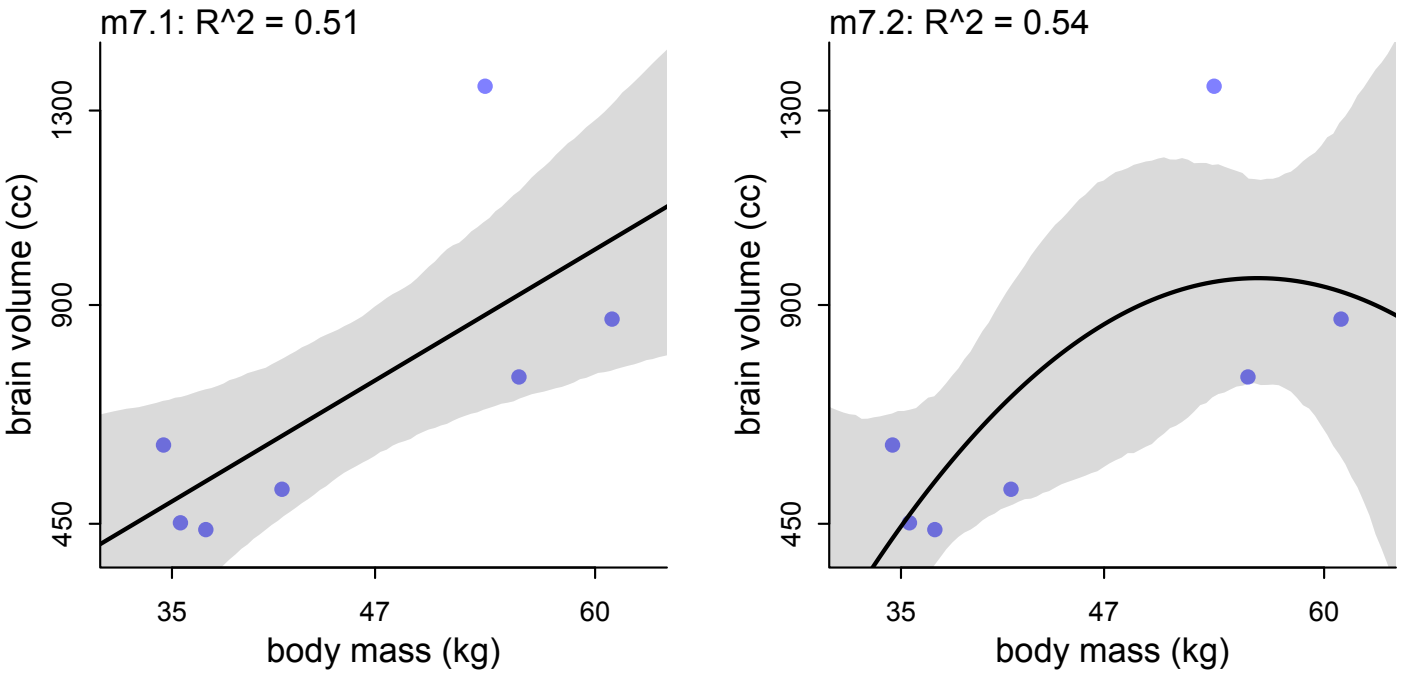

m7.1: R^2 = 0.51
brain volume (cc)
1300
900
450
35
47
60
body mass (kg)
m7.2: R^2 = 0.54
brain volume (cc)
1300
900
450
35
47
60
body mass (kg)

Figure 7.3

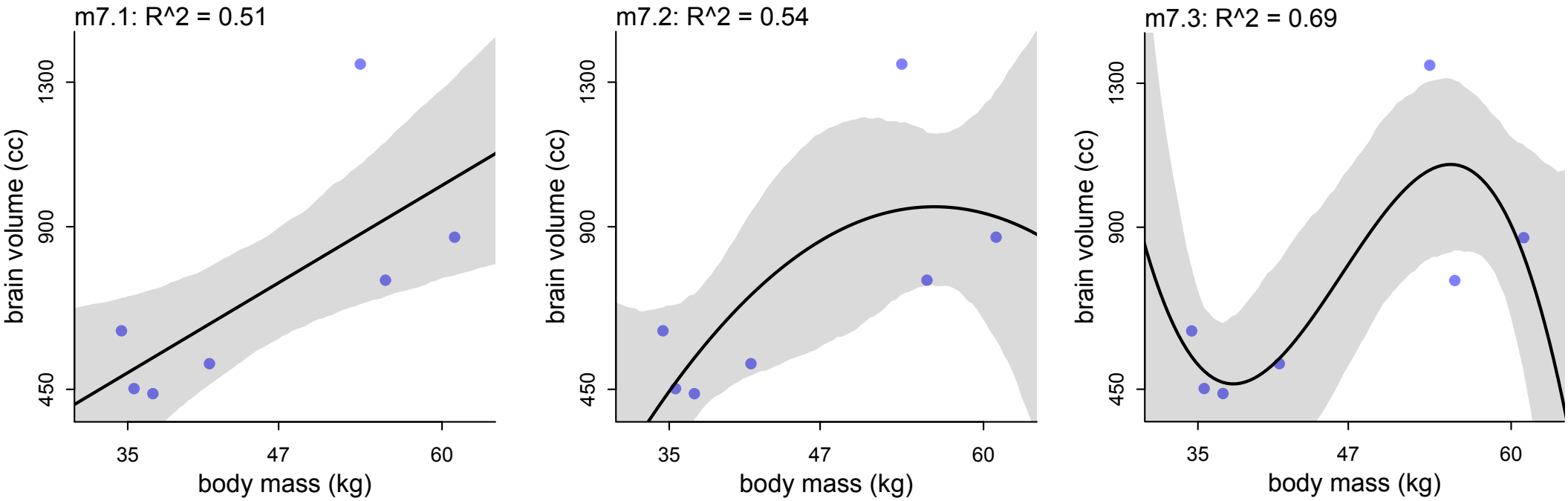
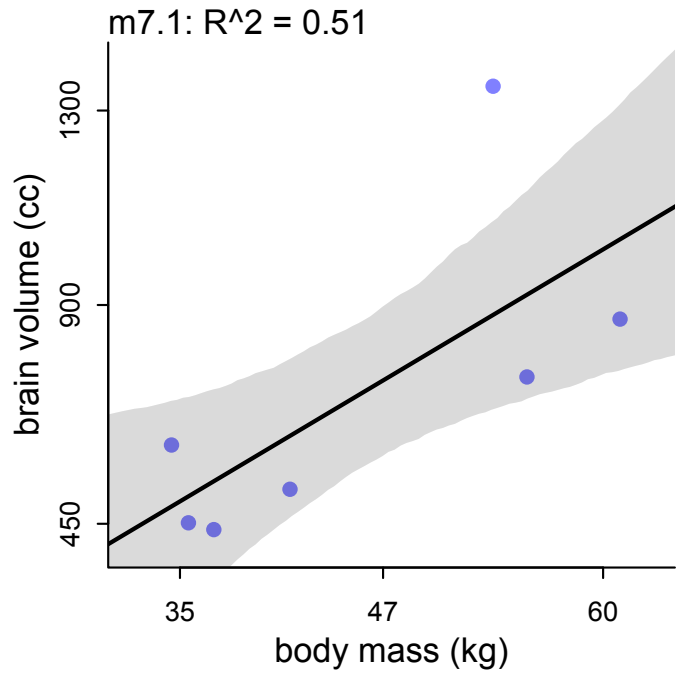
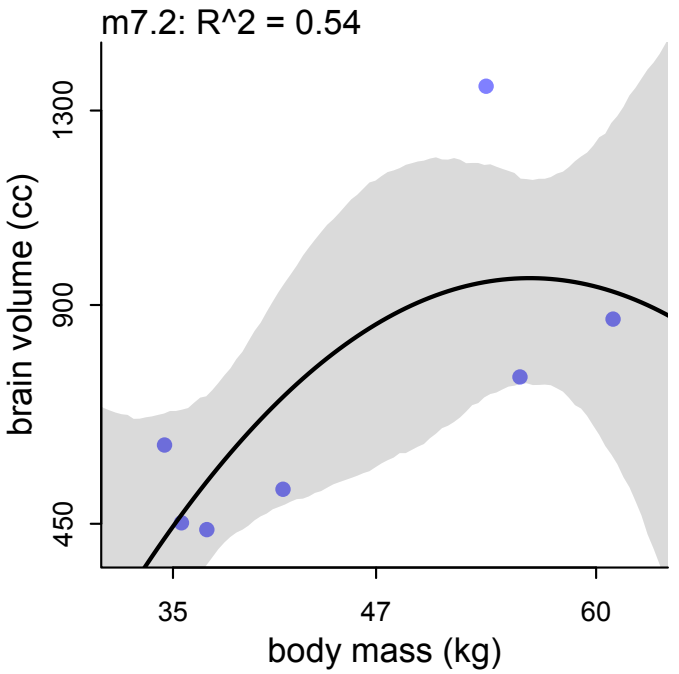
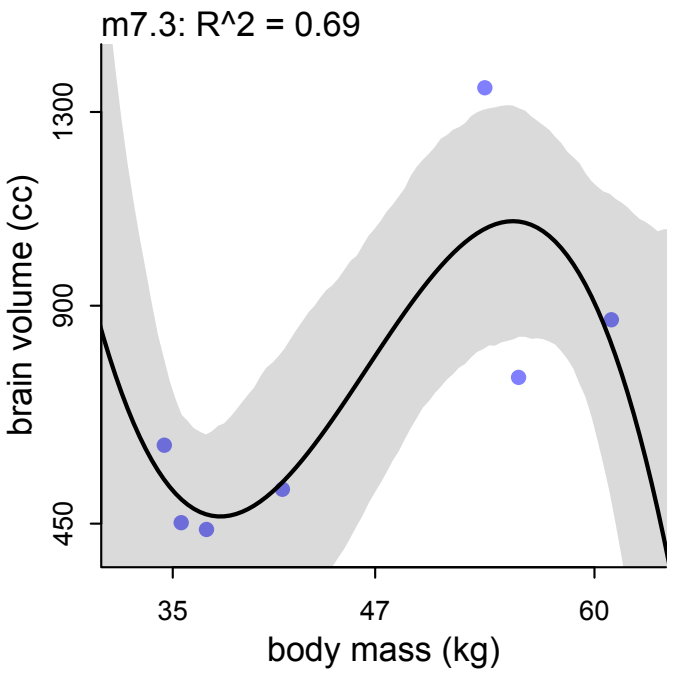

m7.1: R^2 = 0.51
m7.2: R^2 = 0.54
m7.3: R^2 = 0.69
brain volume (cc)
450
900
1300
35
47
60
body mass (kg)

Figure 7.3


m7.1: R^2 = 0.51
brain volume (cc)
450
900
1300
35
47
60
body mass (kg)


m7.2: R^2 = 0.54
brain volume (cc)
450
900
1300
35
47
60
body mass (kg)


m7.3: R^2 = 0.69
brain volume (cc)
450
900
1300
35
47
60
body mass (kg)

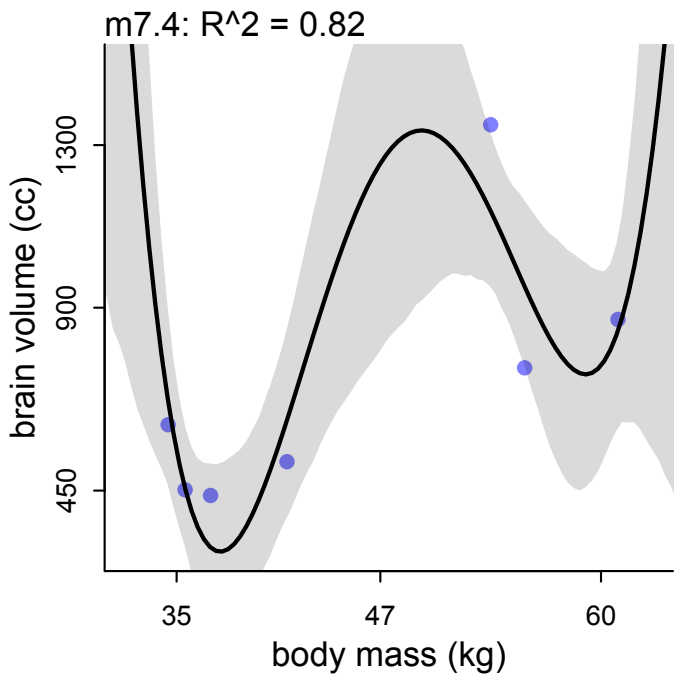

m7.4: R^2 = 0.82
brain volume (cc)
450
900
1300
35
47
60
body mass (kg)

Figure 7.3

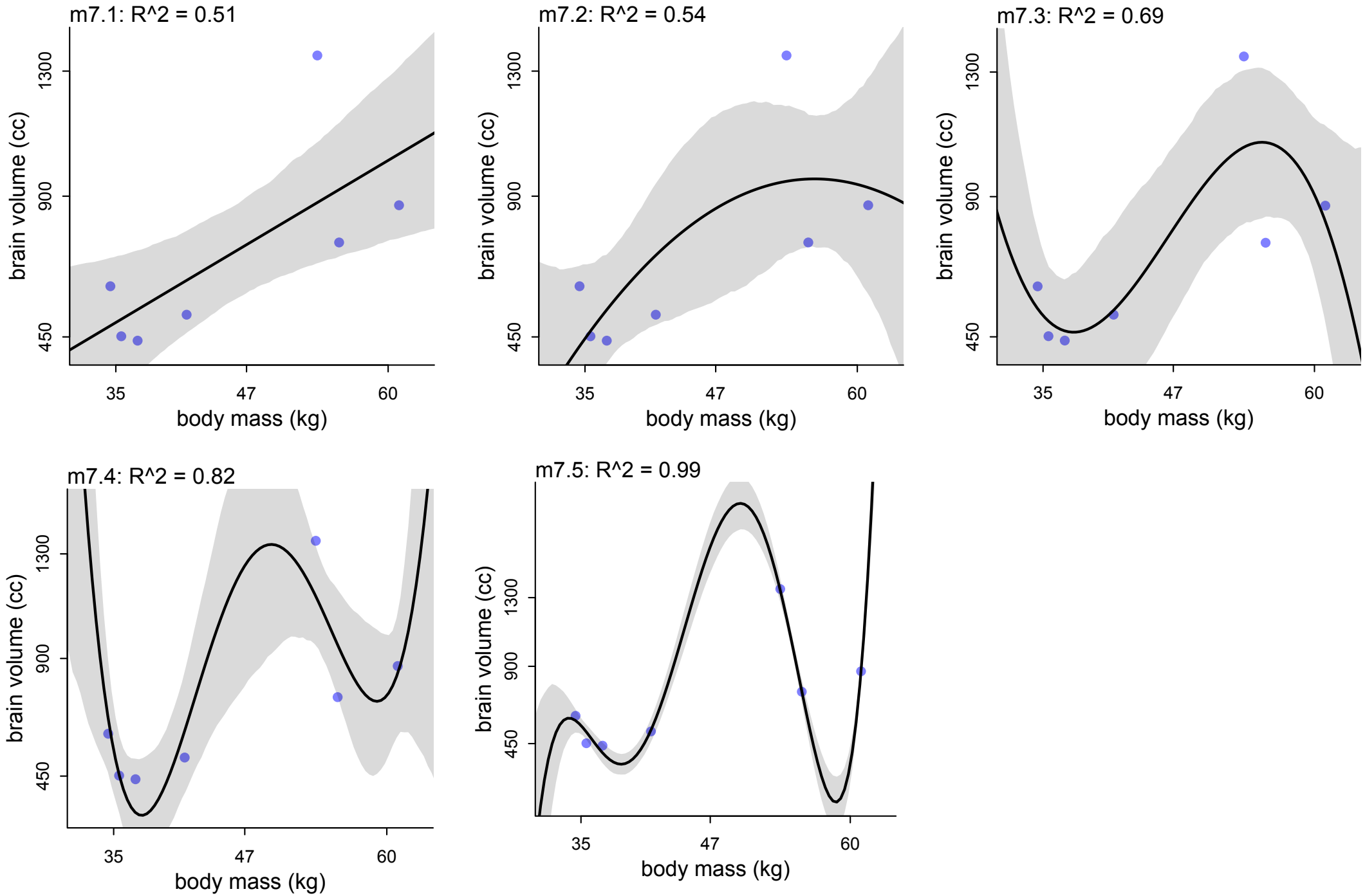

m7.1: R^2 = 0.51
m7.2: R^2 = 0.54
m7.3: R^2 = 0.69
m7.4: R^2 = 0.82
m7.5: R^2 = 0.99
brain volume (cc)
450
900
1300
35
47
60
body mass (kg)

Figure 7.3

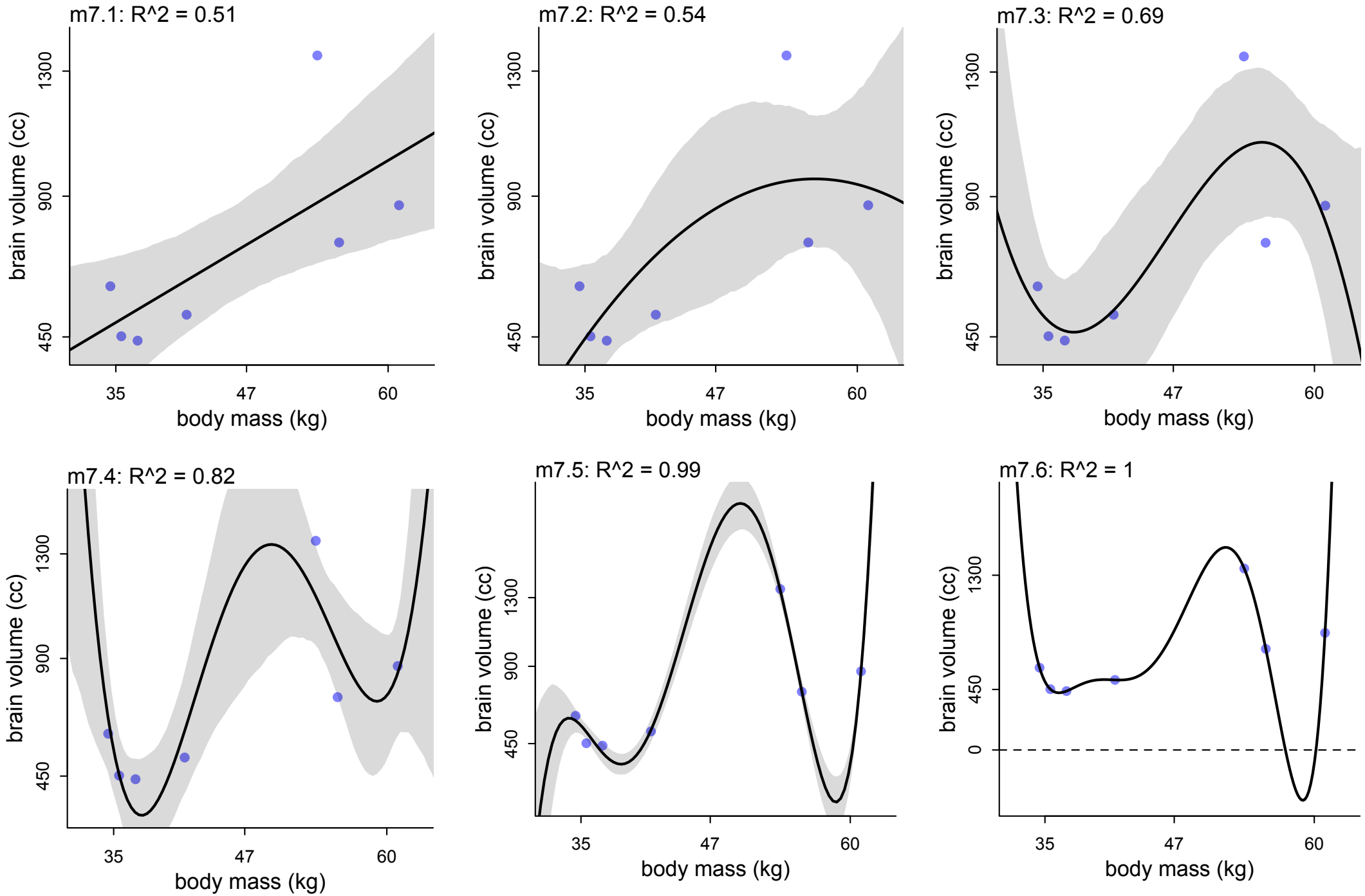

m7.1: R^2 = 0.51
m7.2: R^2 = 0.54
m7.3: R^2 = 0.69
m7.4: R^2 = 0.82
m7.5: R^2 = 0.99
m7.6: R^2 = 1
brain volume (cc)
body mass (kg)
450
900
1300
0
35
47
60

Figure 7.3

Underfitting
Insensitive to
exact data

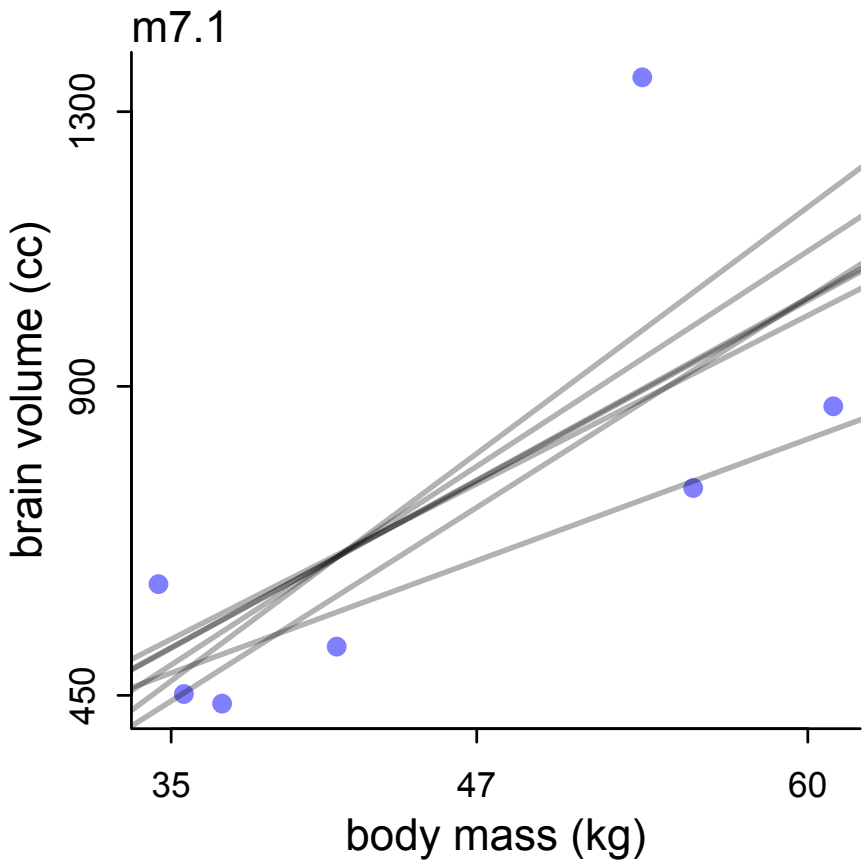
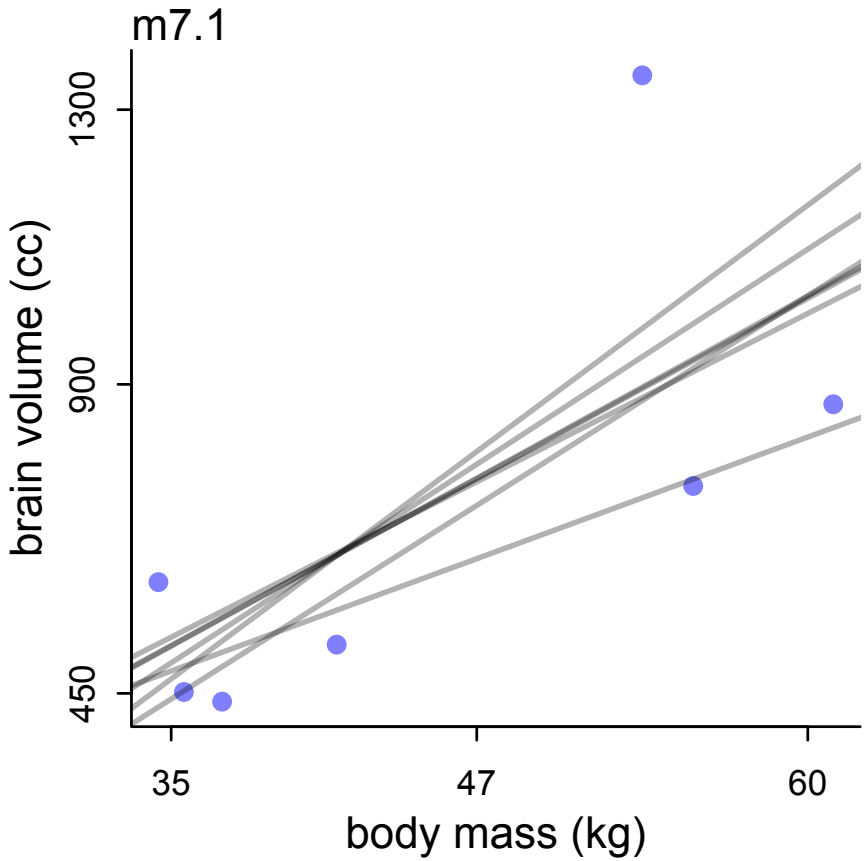


Figure 7.5

Underfitting
Insensitive to
exact data

Overfitting
Very sensitive to
exact data



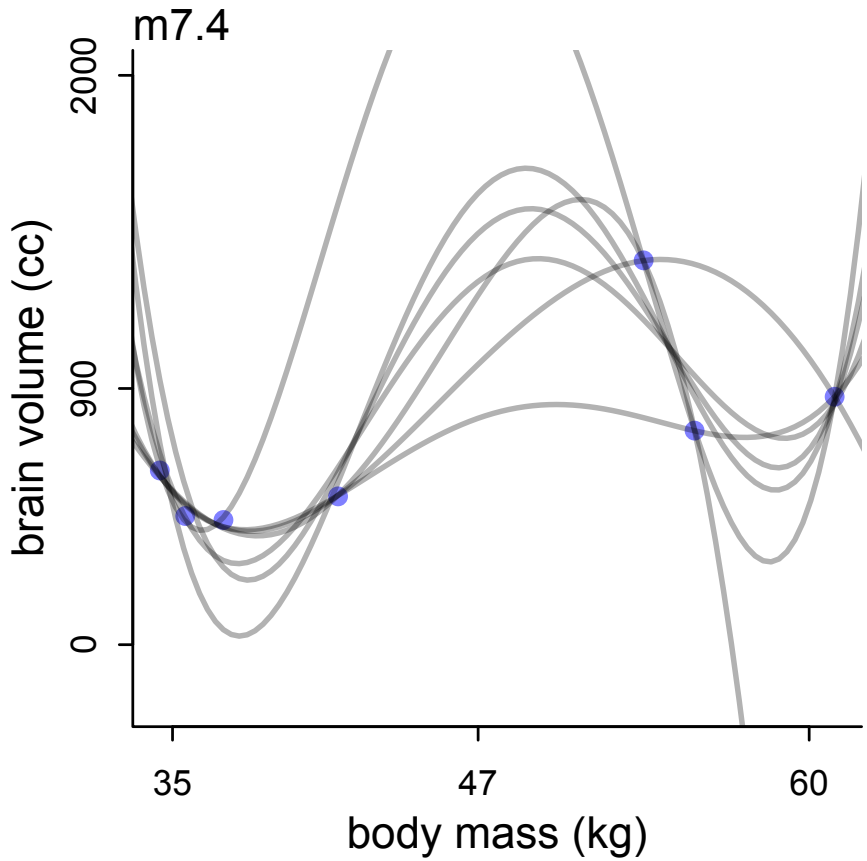


Figure 7.5

Importance of being *regular*

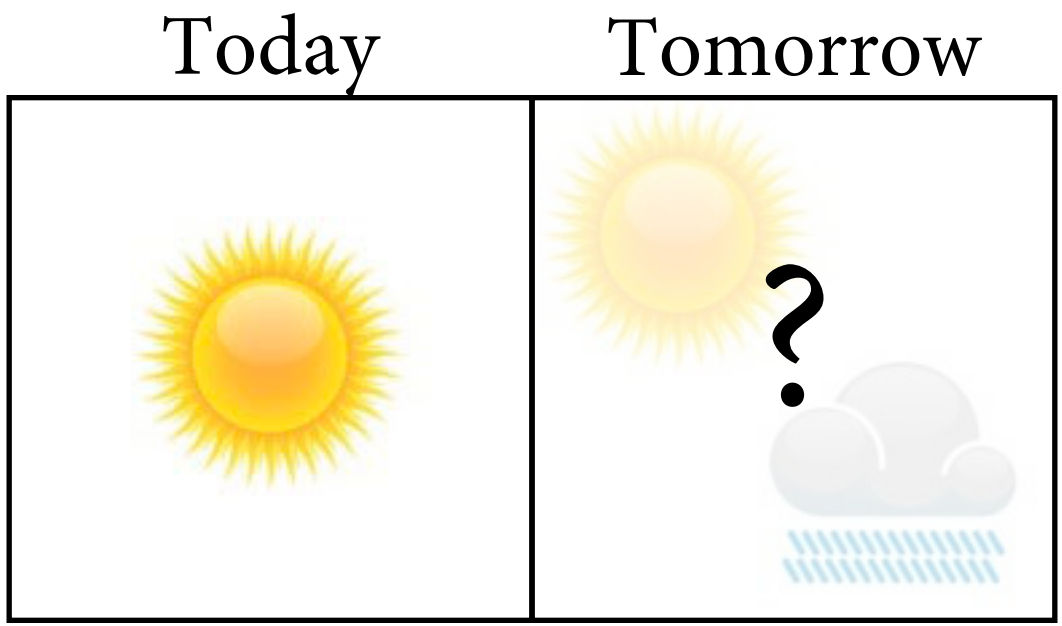
- Want the *regular* features of the sample
- Strategies
  - Regularizing priors (penalized likelihood)
  - Cross-validation
  - Information criteria
  - Science!
- Proper approach depends upon purpose
- Answers are never *only* in the data, but they do usually require data

The road to CV & WAIC

- How to measure accuracy?
- How measure distance from the target?
- How can we estimate that distance?
- How can we predict accuracy on new data?

Information theory

- Machine prediction obeys **information theory**
- *Information*: Reduction in uncertainty caused by learning an outcome.



	Today	Tomorrow	
Los Angeles		?	
Glasgow		?	
New York		?	

Information theory

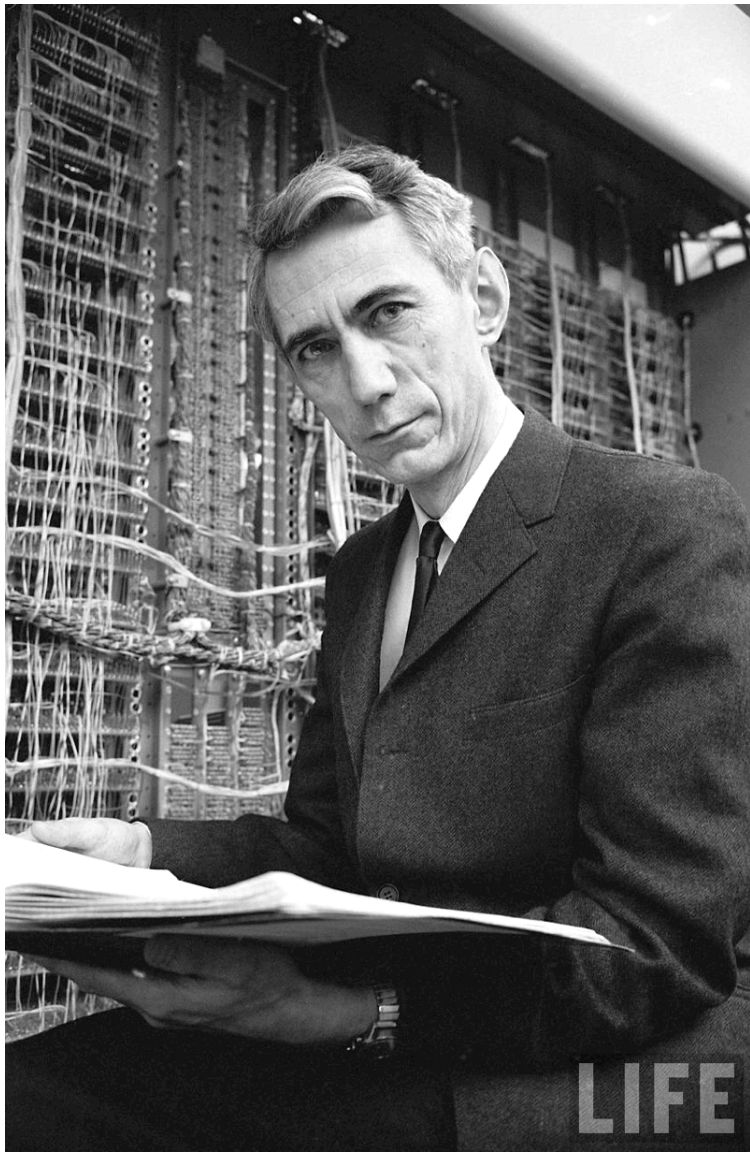
- *Information*: Reduction in uncertainty caused by learning an outcome.
- How to quantify *uncertainty*? Should be:
  1. Continuous
  2. Increasing with number of possible events
  3. Additive
- These criteria intuitive, but effectiveness is why we keep using them

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.



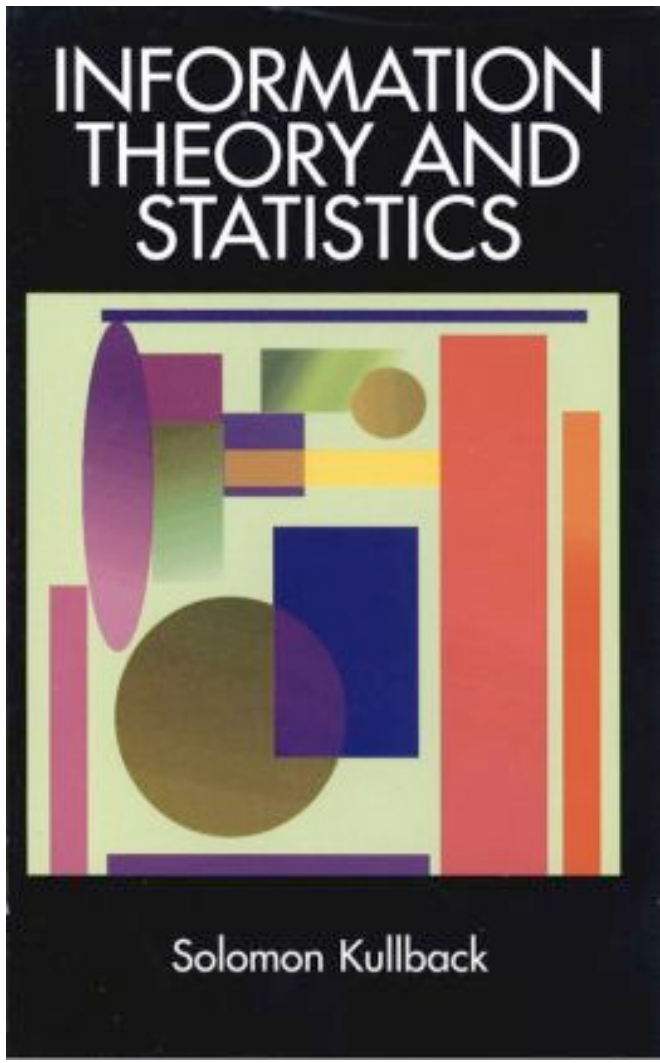
Shannon (1916–2001)

Entropy to accuracy

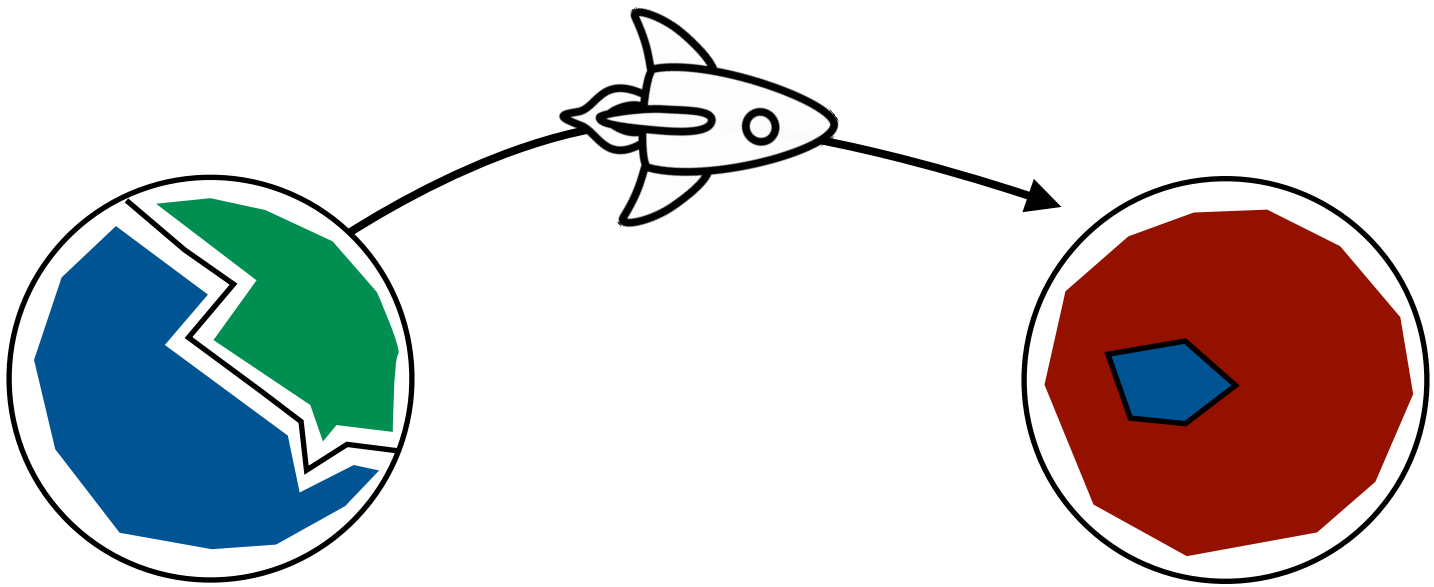
- Two probability distributions: p, q
- p is true, q is model
- How accurate is q, for describing p?
- Distance from q to p: *Divergence*

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

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



Divergence is not symmetric!



Divergence is not symmetric!