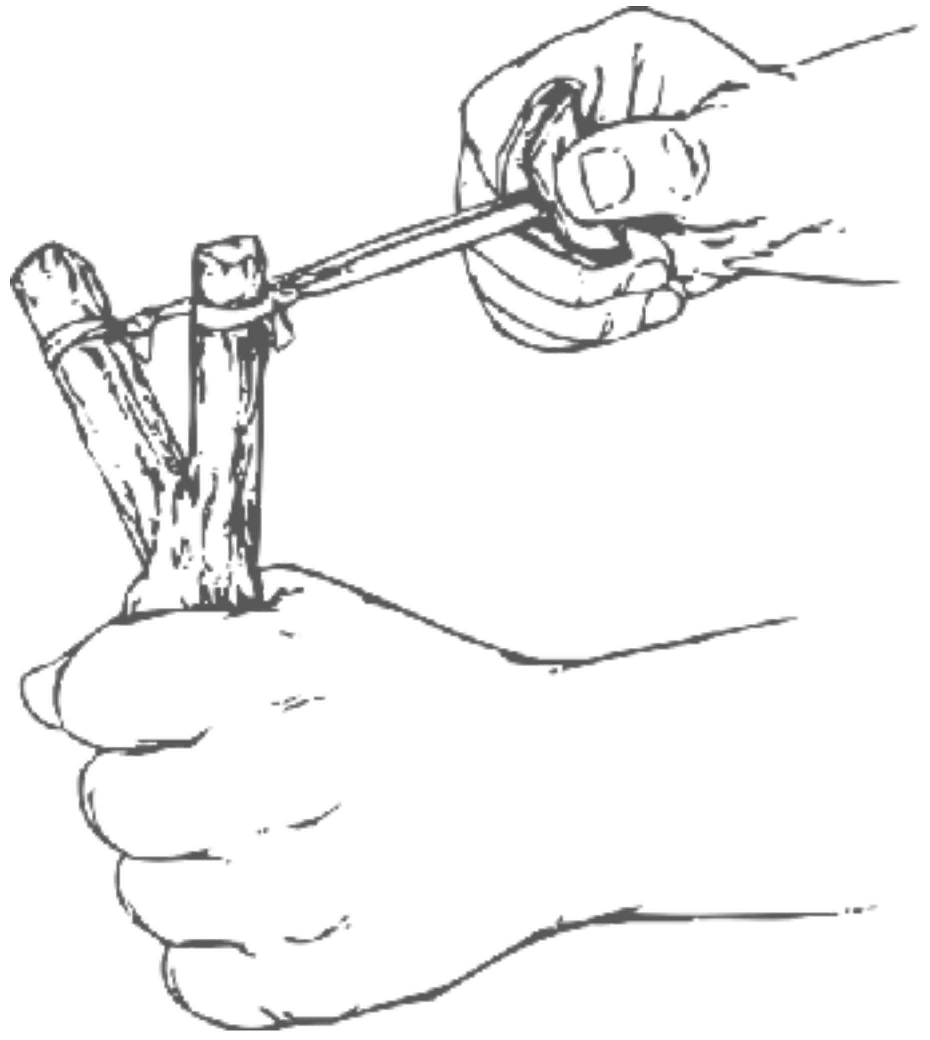
Bayesian data analysis: Theory & practice

Part 1: Bayesian basics & simple linear regression

Michael Franke

Content
First session



1. **“think Bayesian”**
    a. data-generating processes
    b. Bayesian model (prior + likelihood)
    c. updated models

2. **Big Bayesian Four**
    a. prior / posterior parameter distribution
    b. prior / posterior predictives

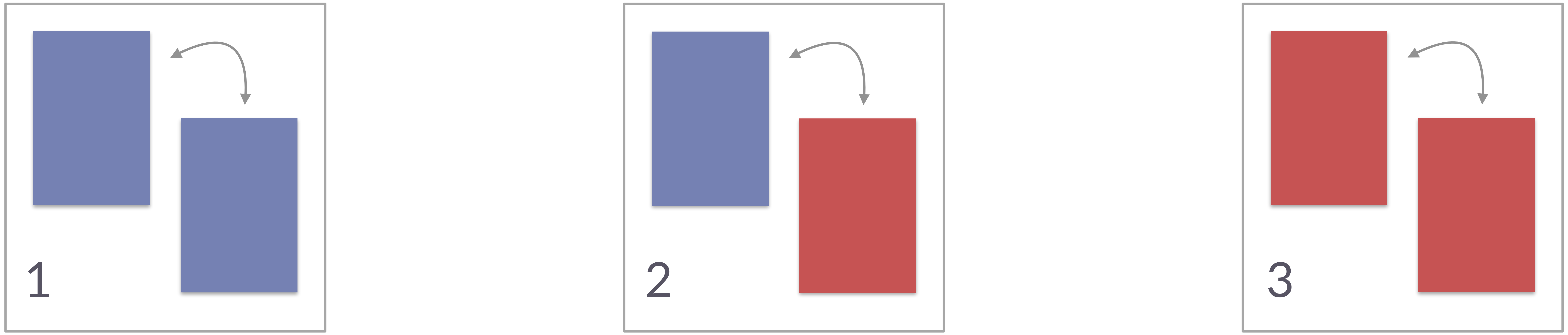
3. **(simple) linear regression**
    a. parameters & priors
    b. likelihood
    c. predictive functions

Bayesian modeling

Three-card problem

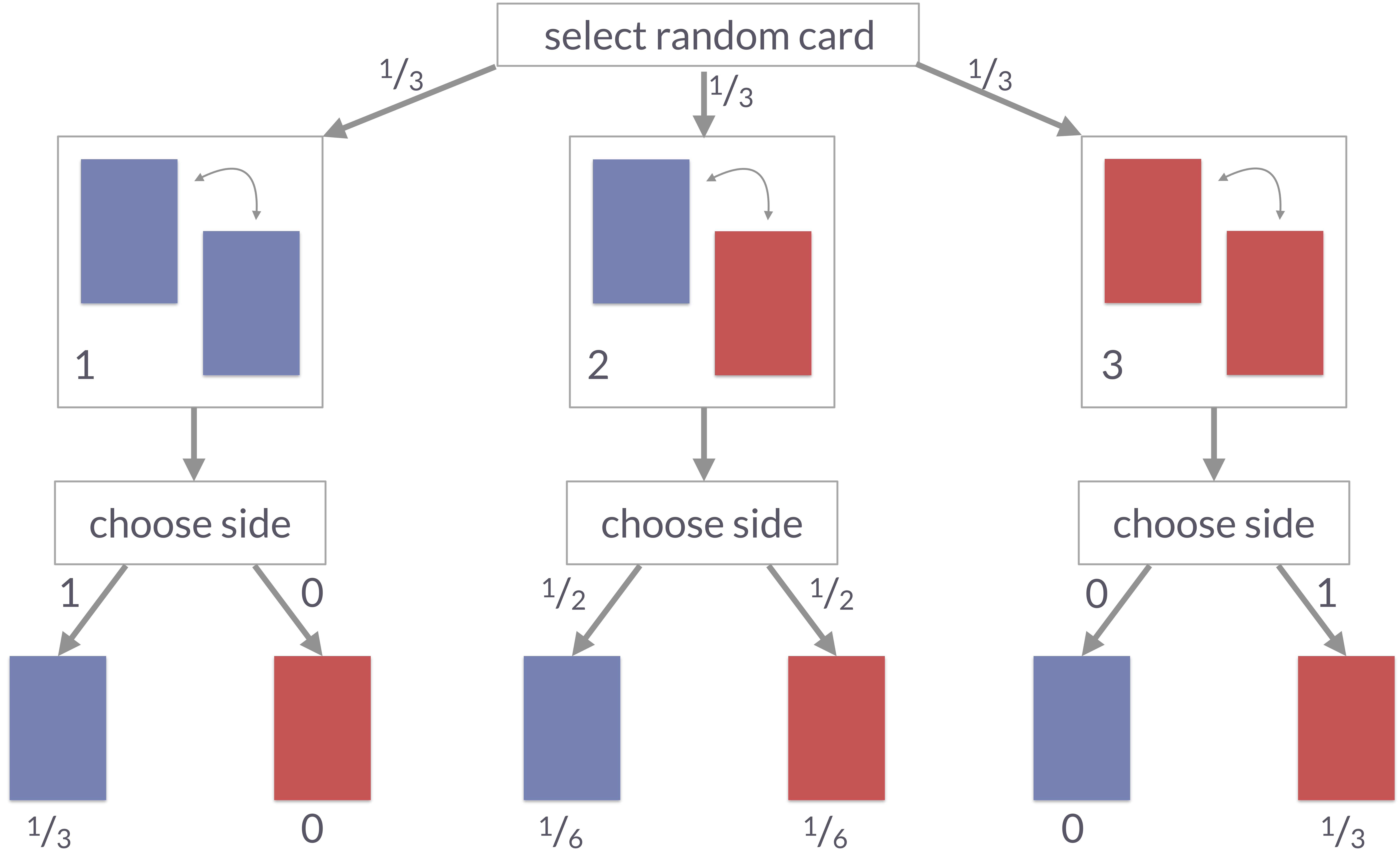
problem statement

- Sample a card (uniformly at random).
- Choose a side of that card to reveal (uniformly at random).
- What's the probability that the side you do not see is **BLUE**, given that the side you see is **BLUE**?



Three-card problem

data-generating process



Conditional probability and Bayes rule

for the three-card problem

- conditional probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- Bayes rule

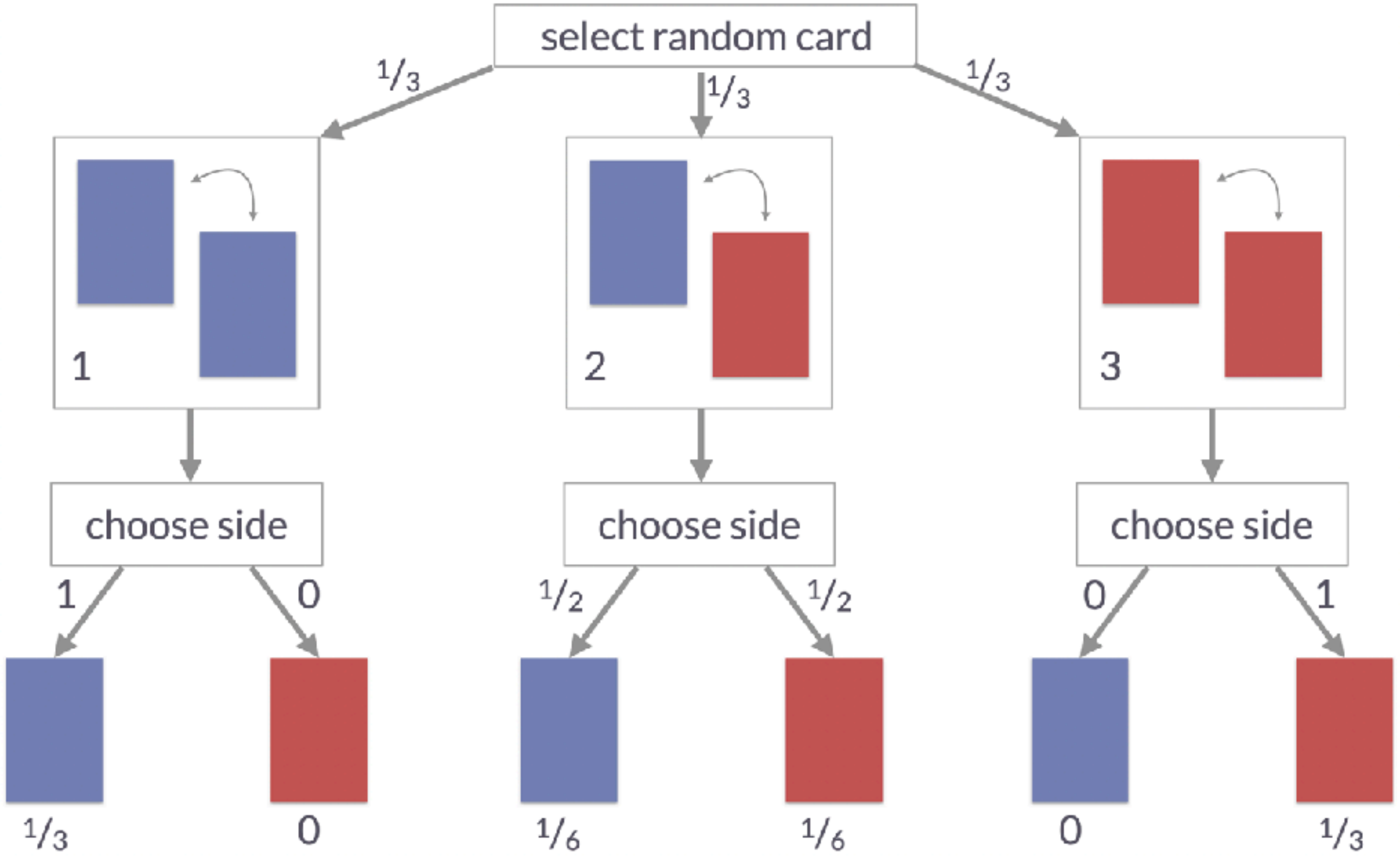
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- Applied to three-card problem:

$$P(\text{card 1} \mid \text{blue}) = \frac{P(\text{blue} \mid \text{card 1})\ P(\text{card 1})}{P(\text{blue})}$$

$$= \frac{1 \times \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

“reasoning from observed effect to latent cause via a model of the data-generating process”



demo

[3-card problem in WebPPL]

Bayes rule for parameter inference

which parameter values are likely to have generated the data?

$$P(\theta \mid D) = \frac{P(D \mid \theta)\ P(\theta)}{\int P(D \mid \theta)\ P(\theta)\ \mathrm{d}\theta}$$

read more here

Bayesian data analysis

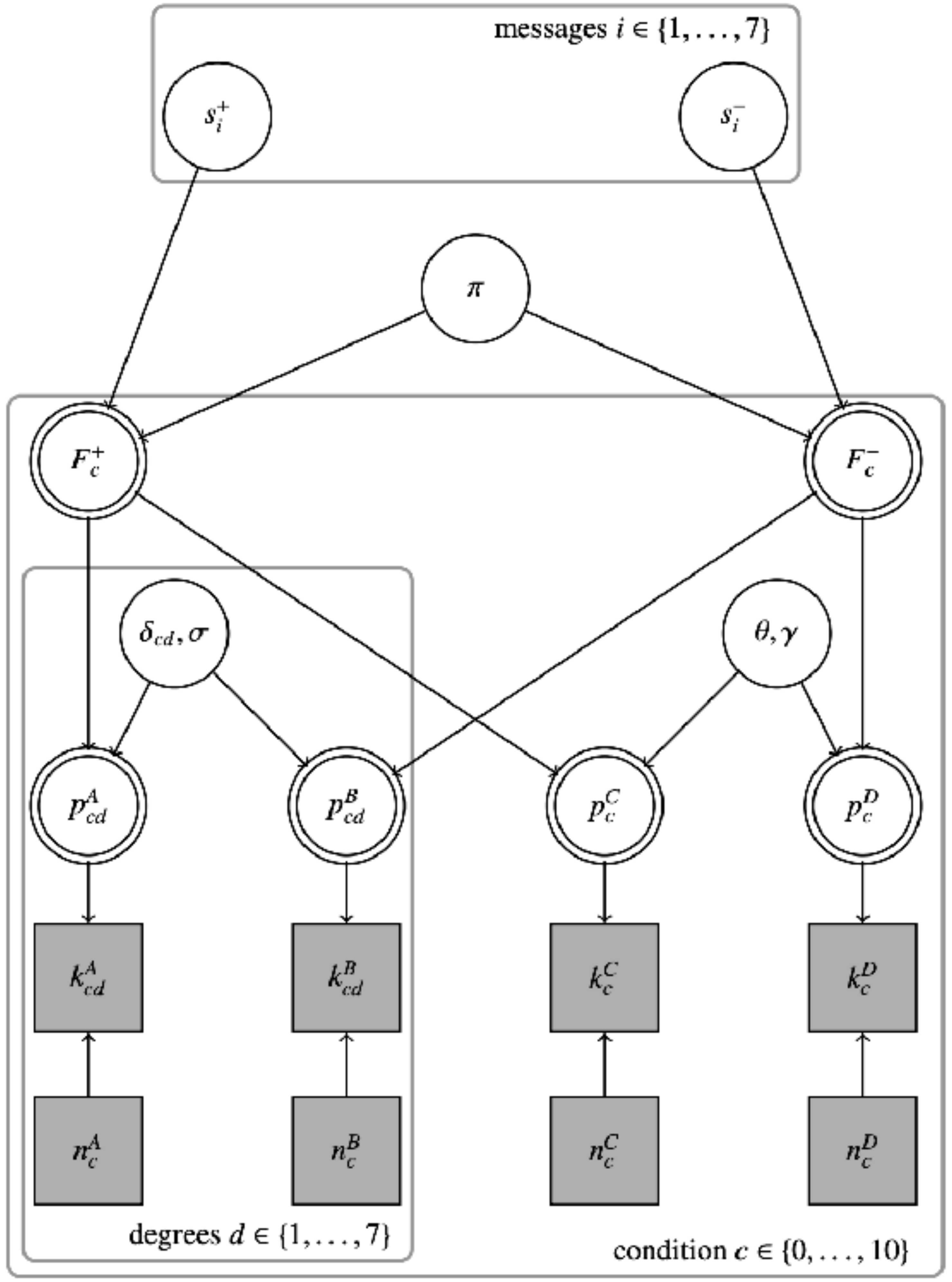
in a nutshell

- BDA is about what we *should* believe given:
  - some observable data, and
  - our model of how this data was generated (a.k.a. **the data-generating process**)
- our best friend will be **Bayes rule**
  - e.g., for **parameter inference**:

$$\underbrace{P(\theta \mid D)}_{\text{posterior}} \propto \underbrace{P(\theta)}_{\text{prior}} \times \underbrace{P(D \mid \theta)}_{\text{likelihood}}$$

  - or, for **model comparison**:

$$\underbrace{\frac{P(M_1 \mid D)}{P(M_2 \mid D)}}_{\text{posterior odds}} = \underbrace{\frac{P(D \mid M_1)}{P(D \mid M_2)}}_{\text{Bayes factor}} \underbrace{\frac{P(M_1)}{P(M_2)}}_{\text{prior odds}}$$



$$s_i^{+/-} \sim \text{Beta}(1, 1)$$

$$\tfrac{1}{\pi} \sim \text{Gamma}(0.5, 0.5)$$

$$F_c^{+/-} = \text{F}(c\,;\, \vec{s}^{\,+/-}, \pi)$$

$$\sigma \sim \text{Uniform}(0, 0.4)$$

$$\delta_{d \in \{1,\dots,6\}} \sim \text{Normal}(d/7, 14)$$
$$\delta_0 = -\infty; \;\; \delta_7 = \infty$$

$$\theta \sim \text{Normal}(0.5, 0.2)$$

$$\tfrac{1}{\gamma} \sim \text{Gamma}(1, 1)$$

$$p_{cd}^{A/B} = \int_{\delta_{cd-1}}^{\delta_{cd}} \text{Normal}(x, F_c^{+/-}, \sigma)\,\mathrm{d}x$$

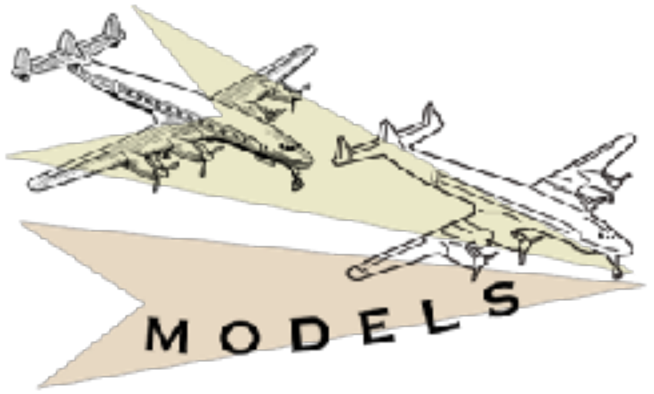
$$p_c^{C/D} = (1 + \exp(-\gamma(F_c^{+/-} - \theta)))^{-1}$$

$$k_{cd}^{A/B} \sim \text{Multinomial}(p_{cd}^{A/B}, n_c^{A/B})$$

$$k_c^{C/D} \sim \text{Binomial}(p_c^{C/D}, n_c^{C/D})$$

Statistical models

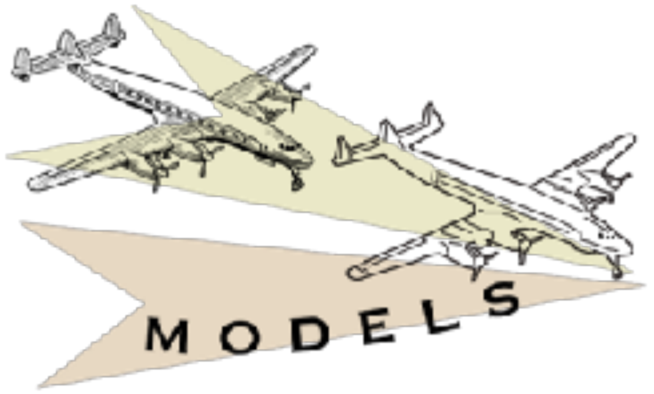
likelihoods from a data-generating process



- A **statistical model** is a condensed formal representation, following common conventional practices of formalization, of the assumptions we make about what the data is and how it might have been generated by some (usually: stochastic) process.
- "All models are wrong, but some are useful." (Box 1979)

Bayesian statistical models

parameterized likelihood + priors



- a **Bayesian statistical model** $\mathscr{M} = \langle \Theta, P_{\mathscr{D}}, P_{\Theta} \rangle$ of a stochastic process generating data D from a set of possible data $\mathscr{D}$ consists of:
  - a space of **parameter vectors** Θ
  - a (conditional) **likelihood function**: $P_{\mathscr{D}} : \Theta \rightarrow \Delta(\mathscr{D})$
  - a (prior) distribution: $P_{\Theta} \in \Delta(\Theta)$

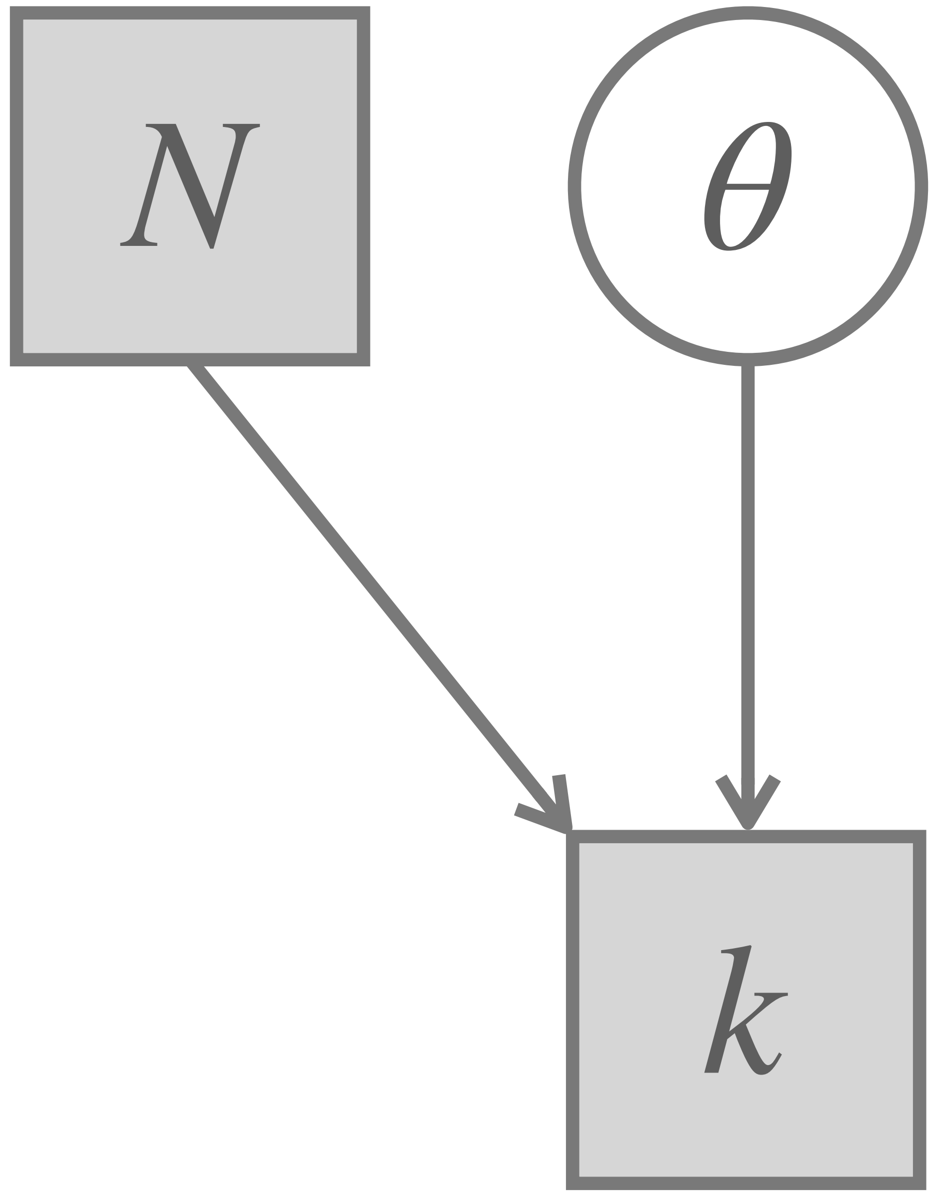
Example: The Binomial Model

the 'coin-flip' model

- data: pair of numbers $D = \{k, N\}$
  - N is the number of tosses
  - k is the number of heads (successes)
- variable:
  - θ is the number of heads (successes)
- uninformed prior:

  $\theta \sim \text{Beta}(1,1)$
- likelihood function:

  $k \sim \text{Binomial}(\theta, N)$
- conventions for model graphs:
  - circles / squares: continuous / discrete variables
  - white / gray nodes: latent / observed variables



Updating Bayesian models

training on / learning from observations / data

- let $D_{obs} \in \mathcal{D}$ be observed (training) data
- let $\mathcal{M}_1 = \langle \Theta, P_{\mathcal{D}}, P_{\Theta} \rangle$ be the initial / prior model
- the updated / posterior model is $\mathcal{M}_2 = \langle \Theta, P_{\mathcal{D}}, P_{\Theta}^{|D} \rangle$ where the new distribution over parameters $P_{\Theta}^{|D_{obs}}$ is **obtained by Bayesian parameter estimation** in the initial / prior model:

$$P_{\Theta}^{|D_{obs}}(\theta) = P_{\Theta}(\theta \mid D_{obs}) = \frac{P_{\theta}(\theta)\ P_{\mathcal{D}}(D_{obs} \mid \theta)}{C}$$

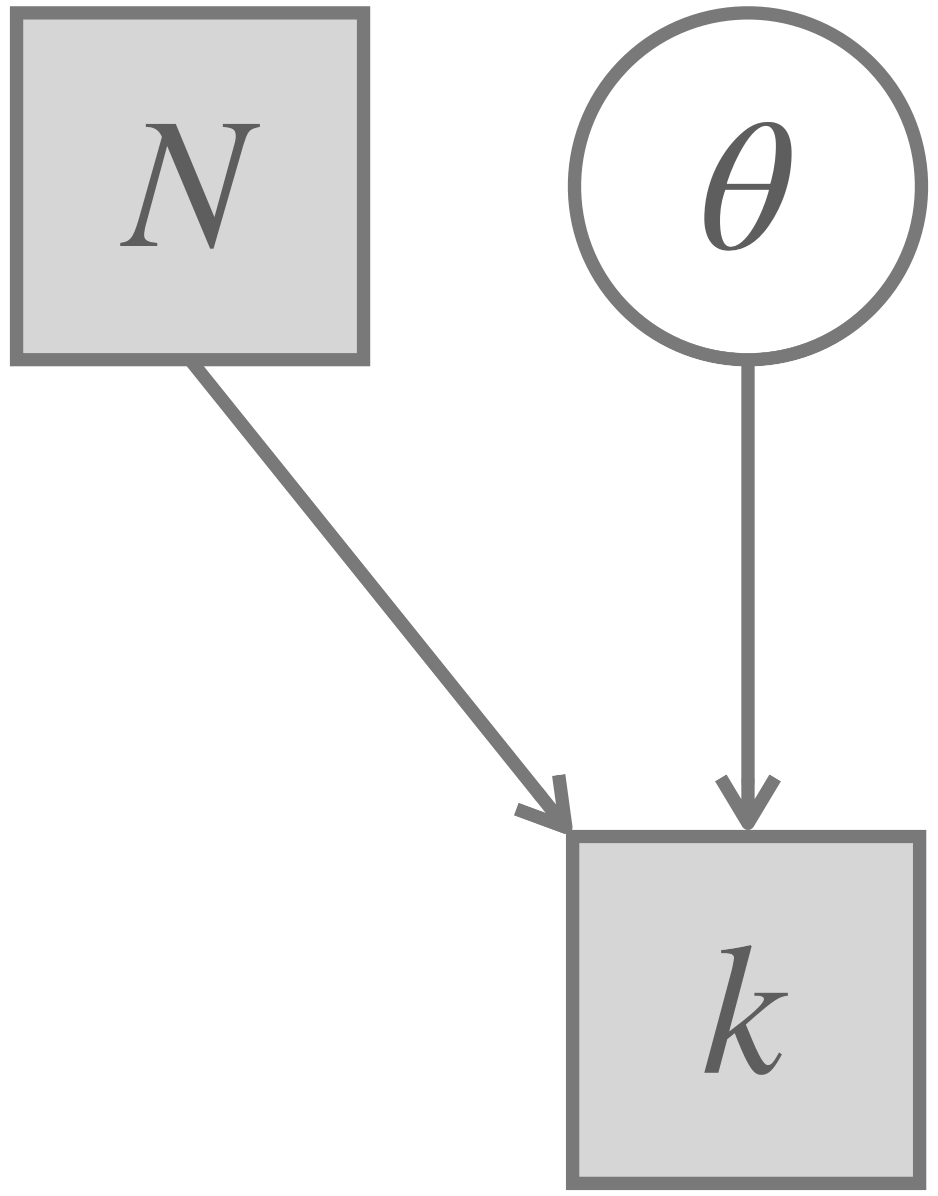
Example: The Binomial Model

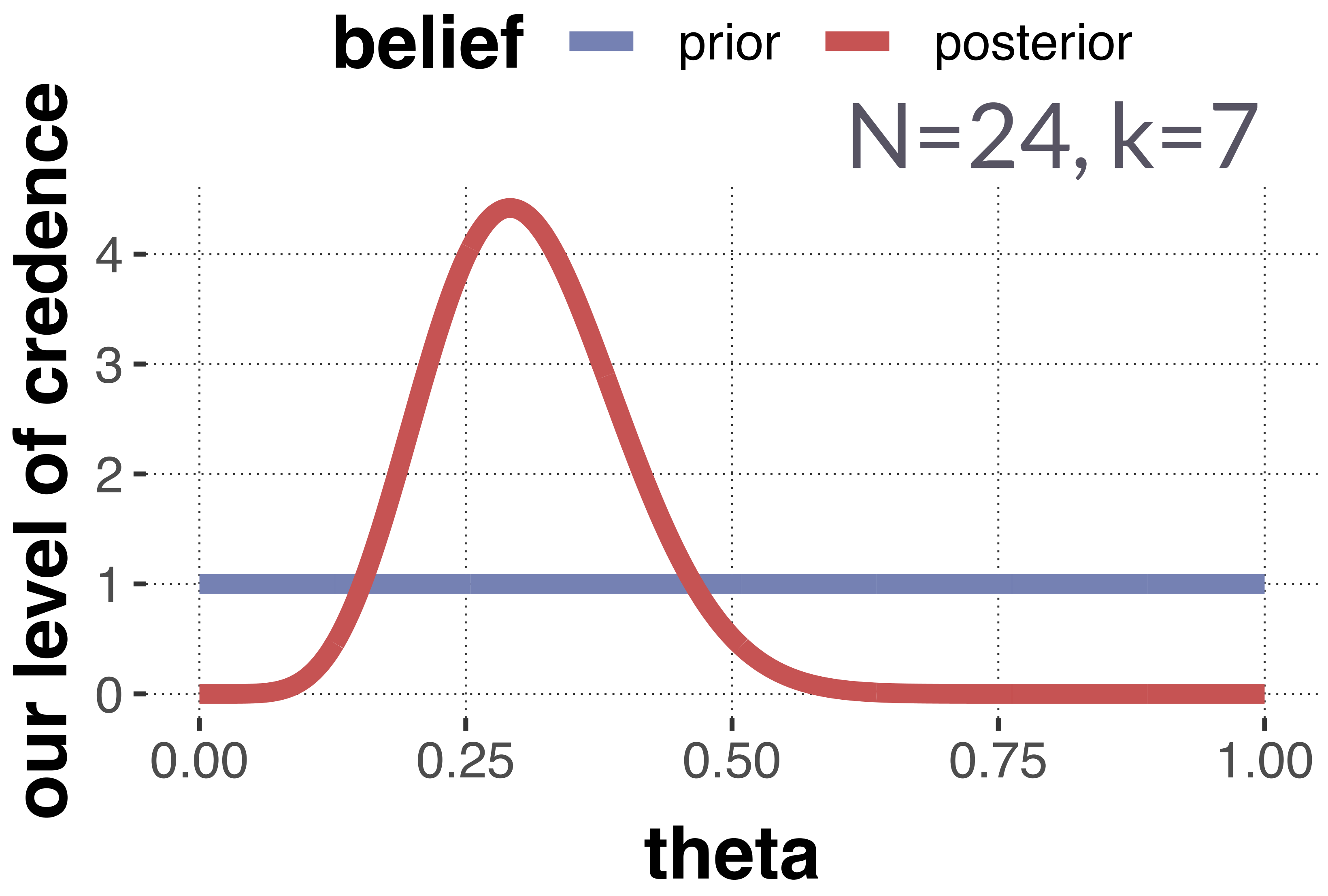
the 'coin-flip' model

- data: pair of numbers $D = \{k, N\}$
  - N is the number of tosses
  - k is the number of heads (successes)
- variable:
  - θ is the number of heads (successes)
- uninformed prior:

  $\theta \sim \text{Beta}(1,1)$
- likelihood function:

  $k \sim \text{Binomial}(\theta, N)$
- conventions for model graphs:
  - circles / squares: continuous / discrete variables
  - white / gray nodes: latent / observed variables





demo

inferring a coin bias in WebPPL

General picture

Bayesian probabilistic ML

$$\boxed{\Theta, P_{\mathscr{D}}, P_{\Theta}} \xrightarrow{D_{obs}} \boxed{\Theta, P_{\mathscr{D}}, P_{\Theta}^{|D_{obs}}}$$

non-Bayesian probabilistic ML

$$\boxed{\Theta, P_{\mathscr{D}}, \theta_1} \xrightarrow{D_{obs}} \boxed{\Theta, P_{\mathscr{D}}, \theta_2}$$

frequentist statistical model

$$\boxed{\Theta, P_{\mathscr{D}}} \xrightarrow{D_{obs}} \boxed{\Theta, P_{\mathscr{D}}, \hat{\theta}}$$

Predictions of a model

the "data-predictive distribution"

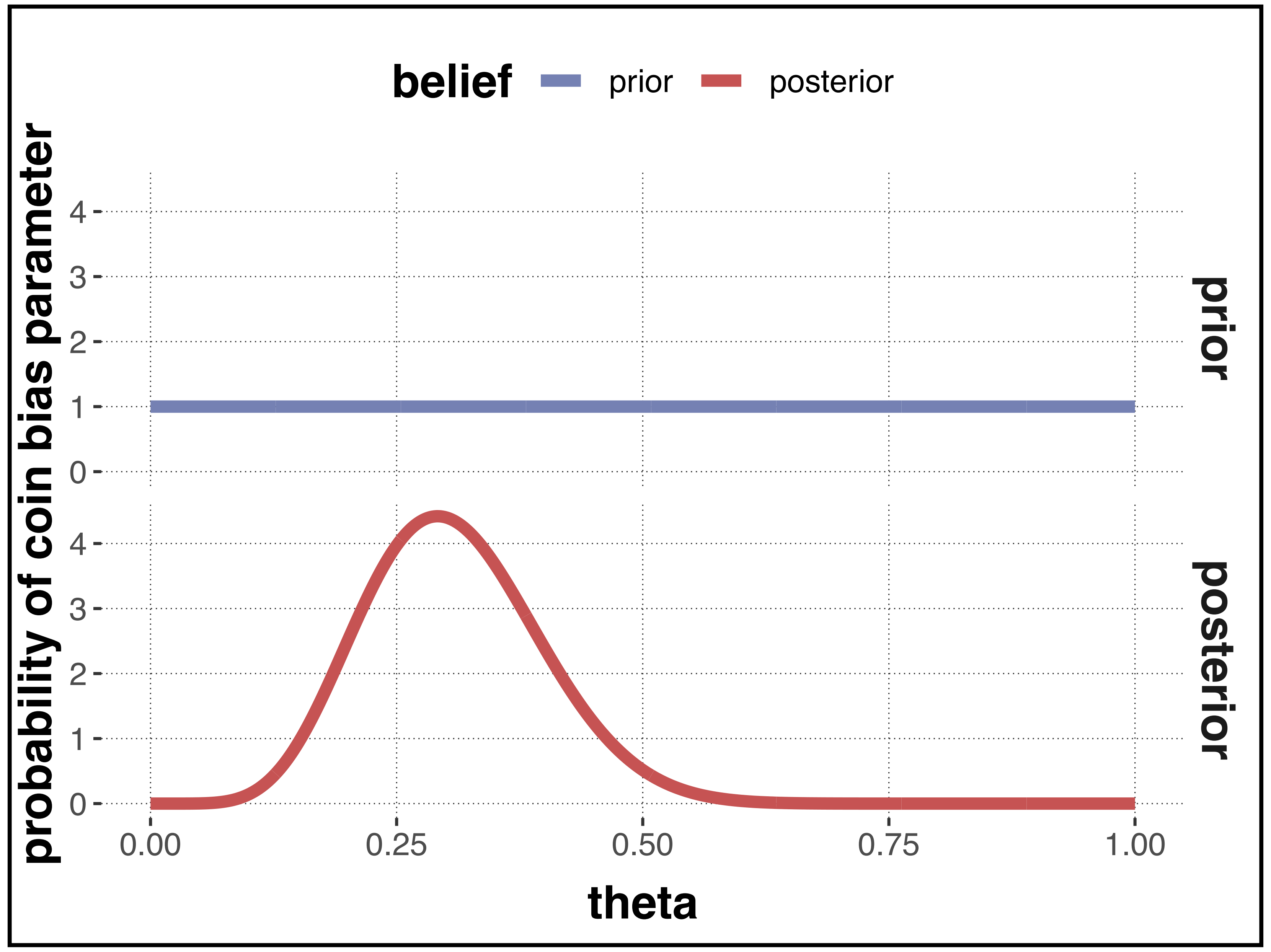
- let $\mathscr{M} = \langle \Theta, P_{\mathscr{D}}, P_{\Theta} \rangle$ be a Bayesian model
- the **predictive of $\mathscr{M}$** is the marginal likelihood:

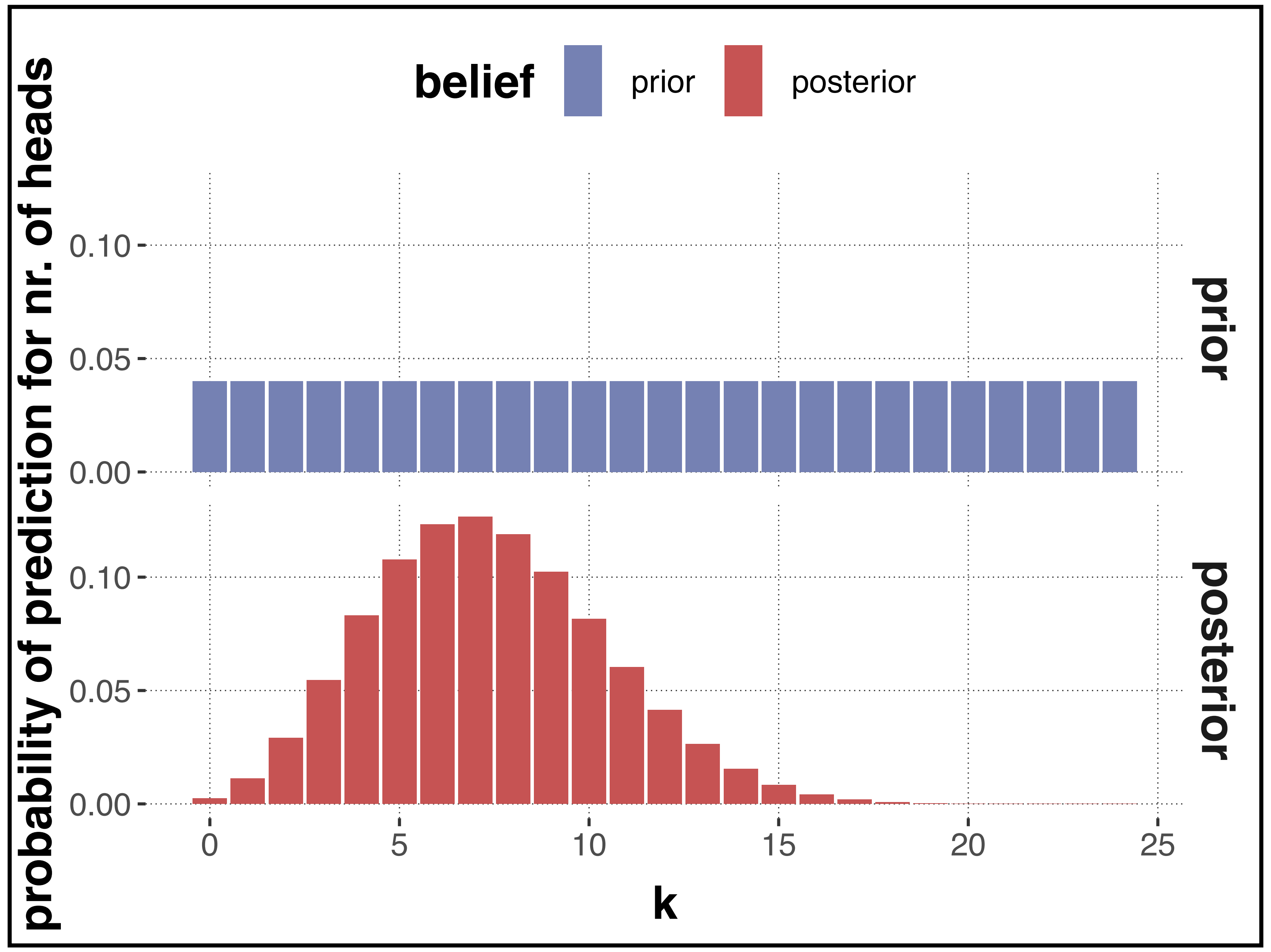
$$P_{\mathscr{D}}(D) = \int P_{\Theta}(\theta) \ P_{\mathscr{D}}(D \mid \theta) \ \mathrm{d}\theta$$

- if $\mathscr{M}_1$ is a prior model and $\mathscr{M}_2$ is the posterior model after updating with some data:
  - the predictive of $\mathscr{M}_1$ is called the **prior predictive**
  - the predictive of $\mathscr{M}_2$ is called the **posterior predictive**

The Big Bayesian 4

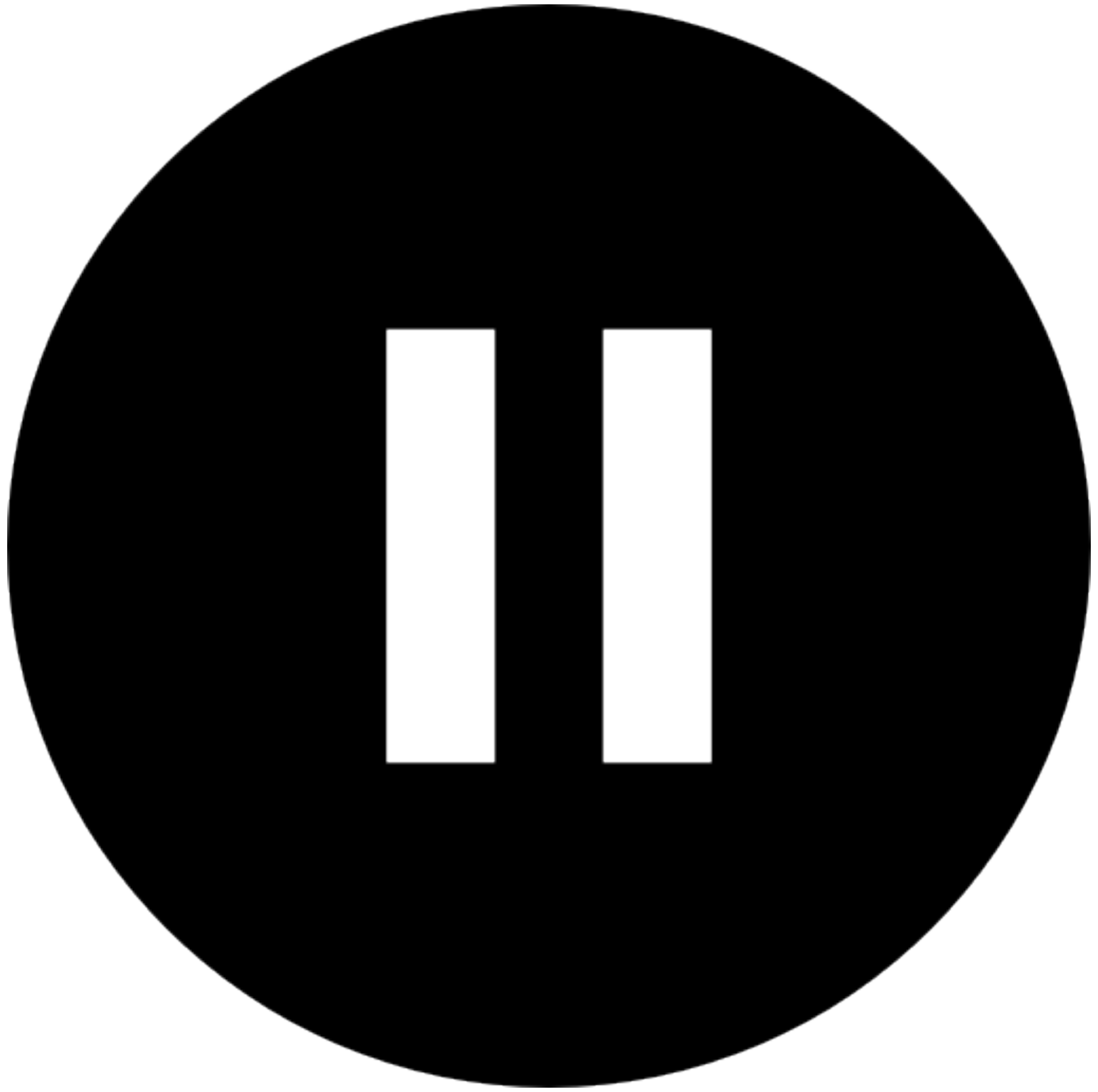
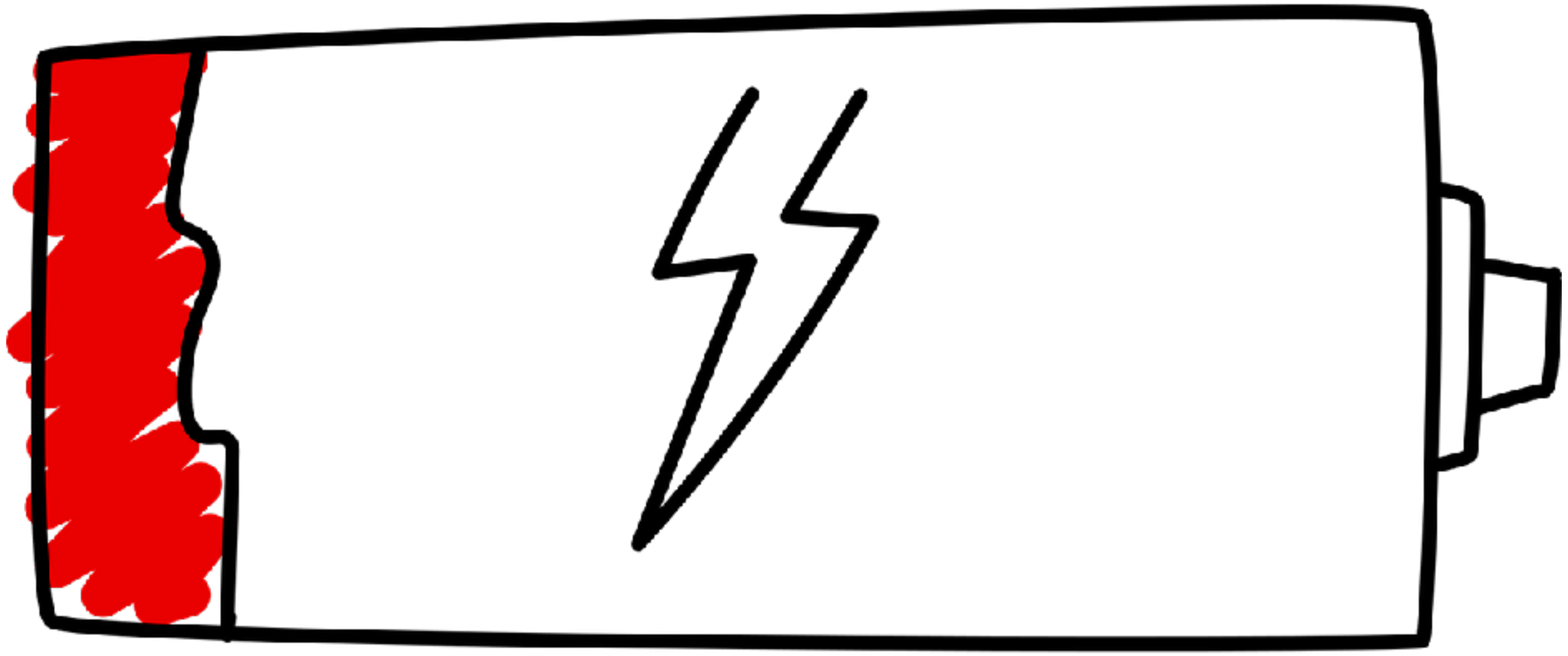
- **prior distribution**
  - uncertainty about model parameters *before* seeing the data
- **posterior distribution**
  - uncertainty about model parameters *after* seeing the data
- **prior predictive distribution**
  - distribution over likely future data points before seeing the data
- **posterior predictive distribution**
  - distribution over likely future data points before seeing the data





demo

Big Bayesian 4 (for coin flip model) in WebPPL

the multiple roles of

Priors in BDA

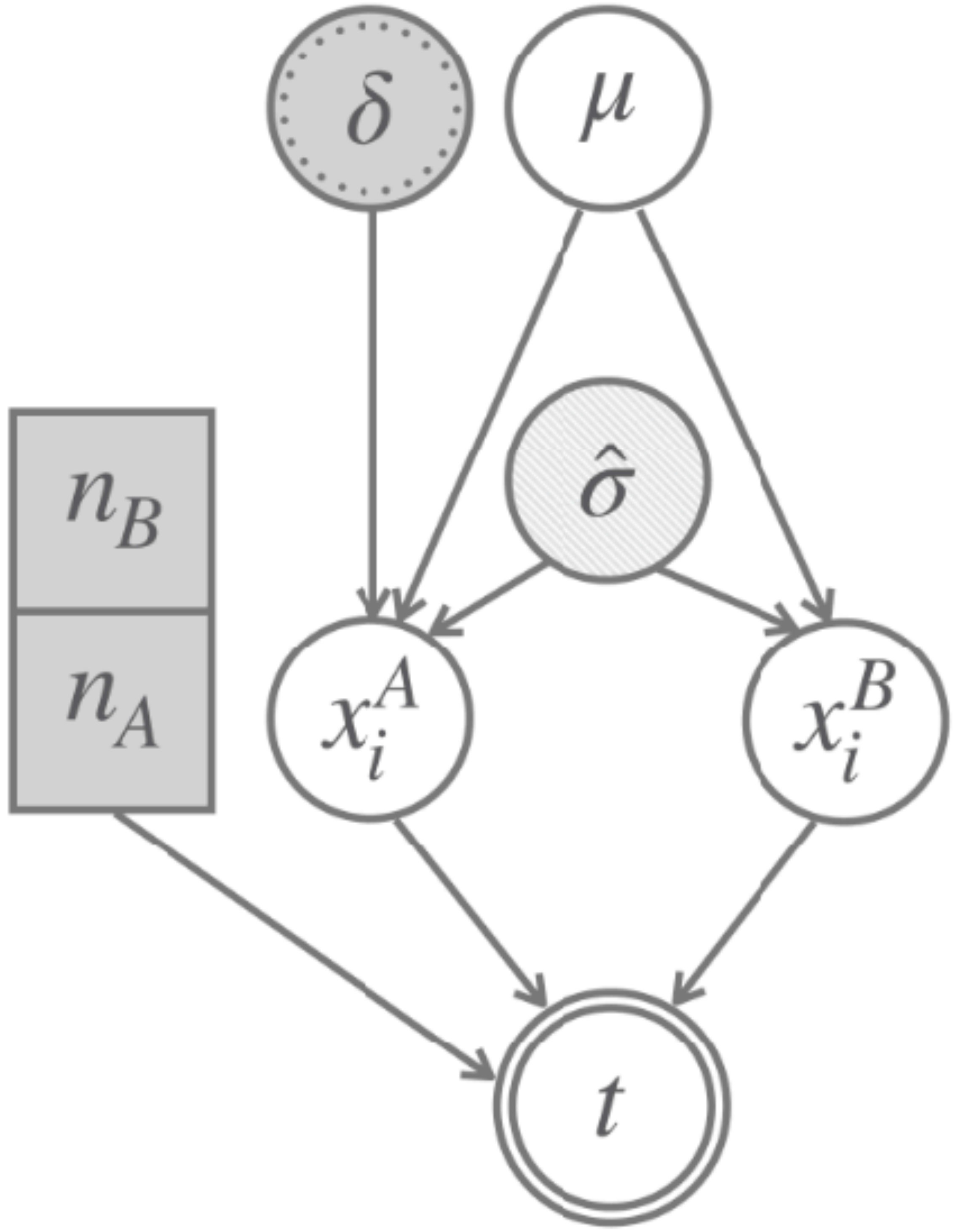
Priors in Bayesian data analysis

more on this as we go along

- subjective beliefs
  - e.g., as justified by prior research
- regularization
  - harness predictive influence of parameters
- priors in multi-level models
  - partial pooling across groups
  - reduce model complexity (↑see "regularization")
- computational considerations
  - enable efficient computation
  - avoid overfitting to sparse data
- objective priors
  - enable long-term error control (type1 & type 2 errors)
- conjugate priors
  - allow exact calculation of Bayesian posterior

Creative fun with data-generating processes

MODELS are everywhere



$$x_i^A \sim \text{Normal}(\mu + \delta, \sigma)$$

$$x_i^B \sim \text{Normal}(\mu, \sigma)$$

$$\hat{\sigma} = \sqrt{\frac{(n_A - 1)\hat{\sigma}_A^2 + (n_B - 1)\hat{\sigma}_B^2}{n_A + n_B - 2} \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

$$t = \left((\bar{x}_A - \bar{x}_B) - \delta\right) \cdot \frac{1}{\hat{\sigma}}$$

Sampling distribution:
$t \sim \text{Student-t}(\nu = n_A + n_B - 2)$

model of the data-generating process buried inside a two-sample t-test

read more here

demo

The Seven Scientists

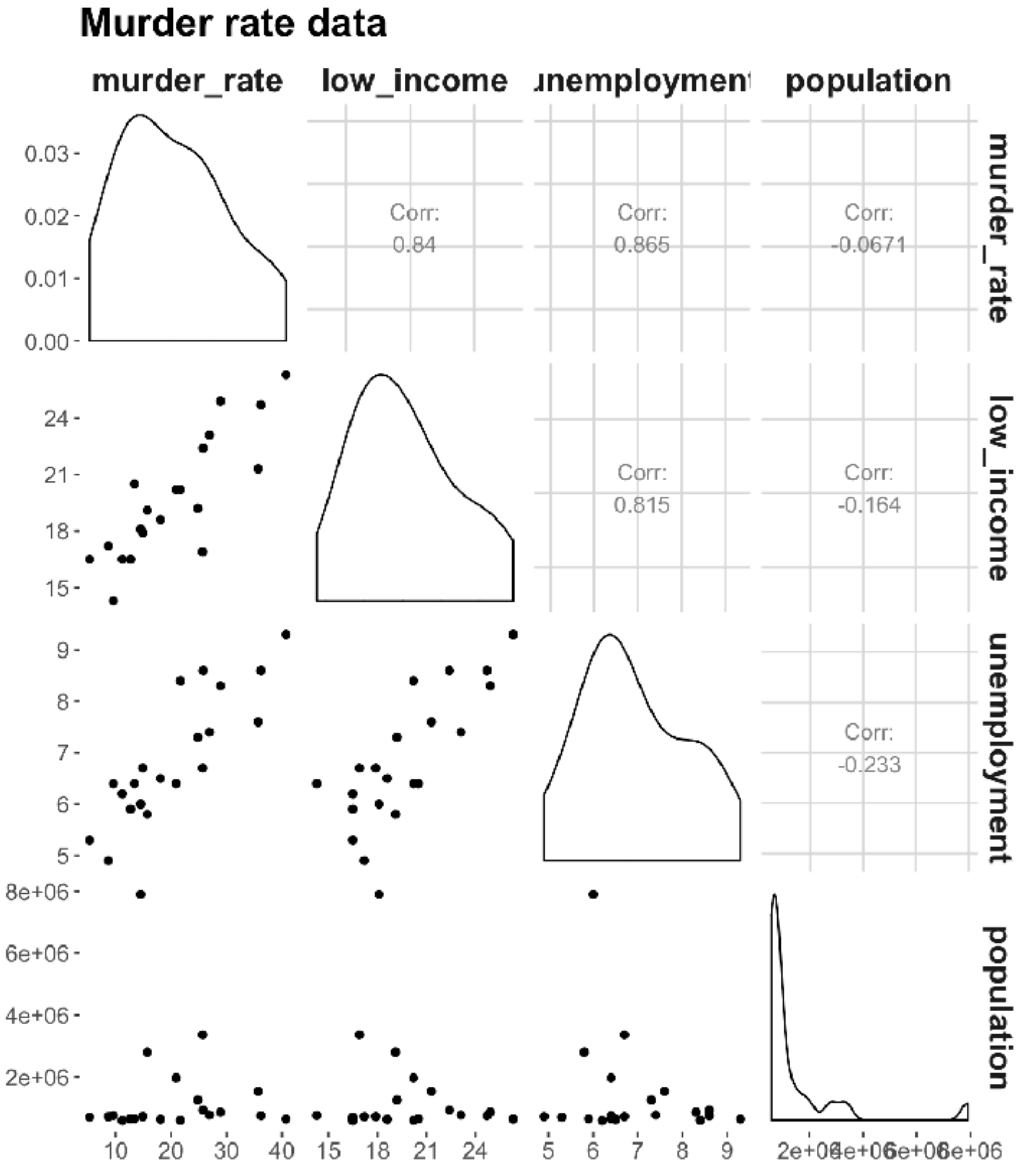
Simple linear regression

likelihood & Bayesian posterior

Murder data

annual murder rate, average income, unemployment rates and population

```
## # A tibble: 20 x 4
##    murder_rate low_income unemployment population
##          <dbl>      <dbl>        <dbl>      <dbl>
##  1        11.2       16.5          6.2     587000
##  2        13.4       20.5          6.4     643000
##  3        40.7       26.3          9.3     635000
##  4         5.3       16.5          5.3     692000
##  5        24.8       19.2          7.3    1248000
##  6        12.7       16.5          5.9     643000
##  7        20.9       20.2          6.4    1964000
##  8        35.7       21.3          7.6    1531000
##  9         8.7       17.2          4.9     713000
## 10         9.6       14.3          6.4     749000
## 11        14.5       18.1          6      7895000
## 12        26.9       23.1          7.4     762000
## 13        15.7       19.1          5.8    2793000
## 14        36.2       24.7          8.6     741000
## 15        18.1       18.6          6.5     625000
## 16        28.9       24.9          8.3     854000
## 17        14.9       17.9          6.7     716000
## 18        25.8       22.4          8.6     921000
## 19        21.7       20.2          8.4     595000
## 20        25.7       16.9          6.7    3353000
```



annual murders per million inhabitants

percentage inhabitants with low income

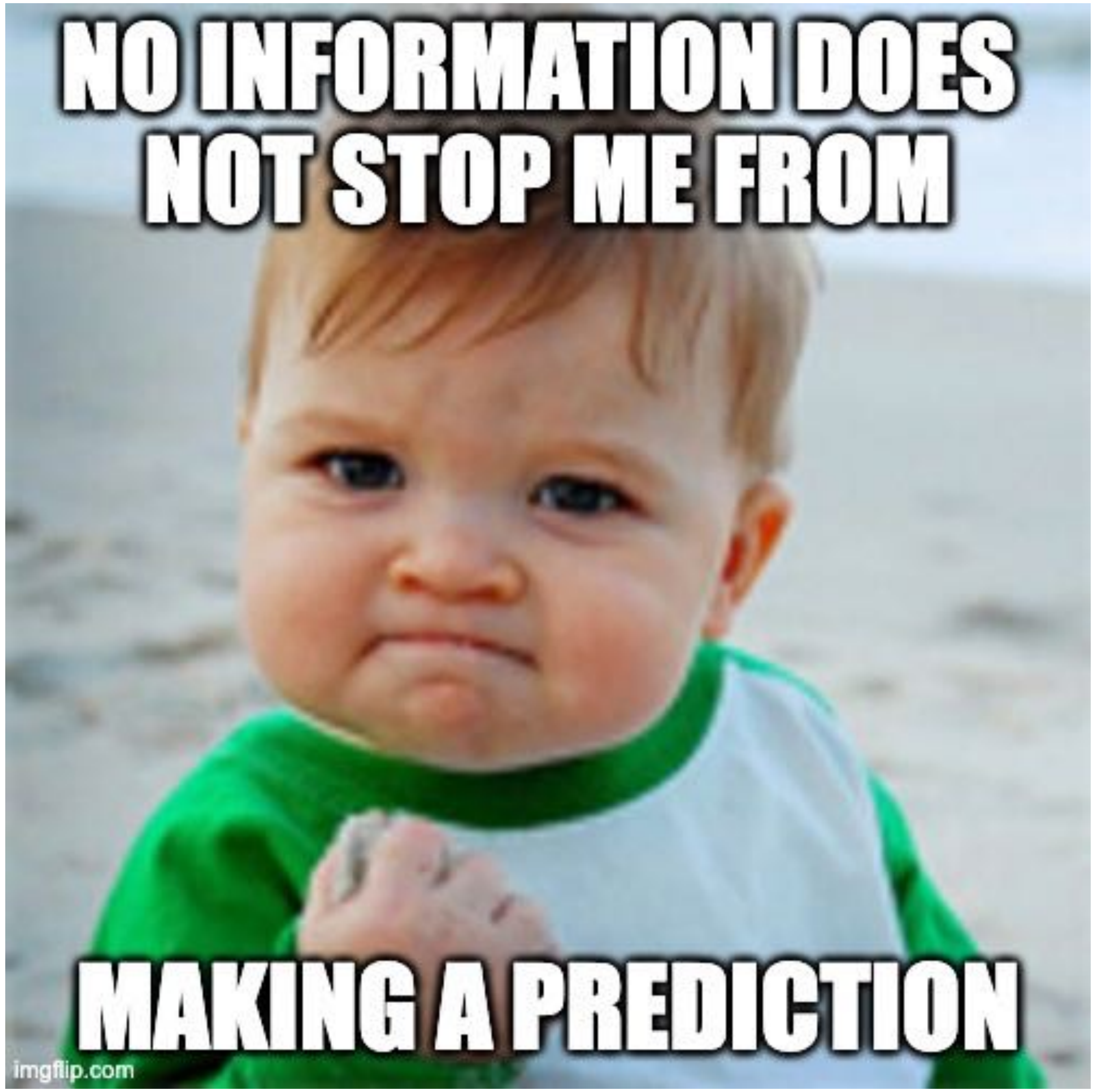
percentage inhabitants who are unemployed

total population

Predicting murder rate

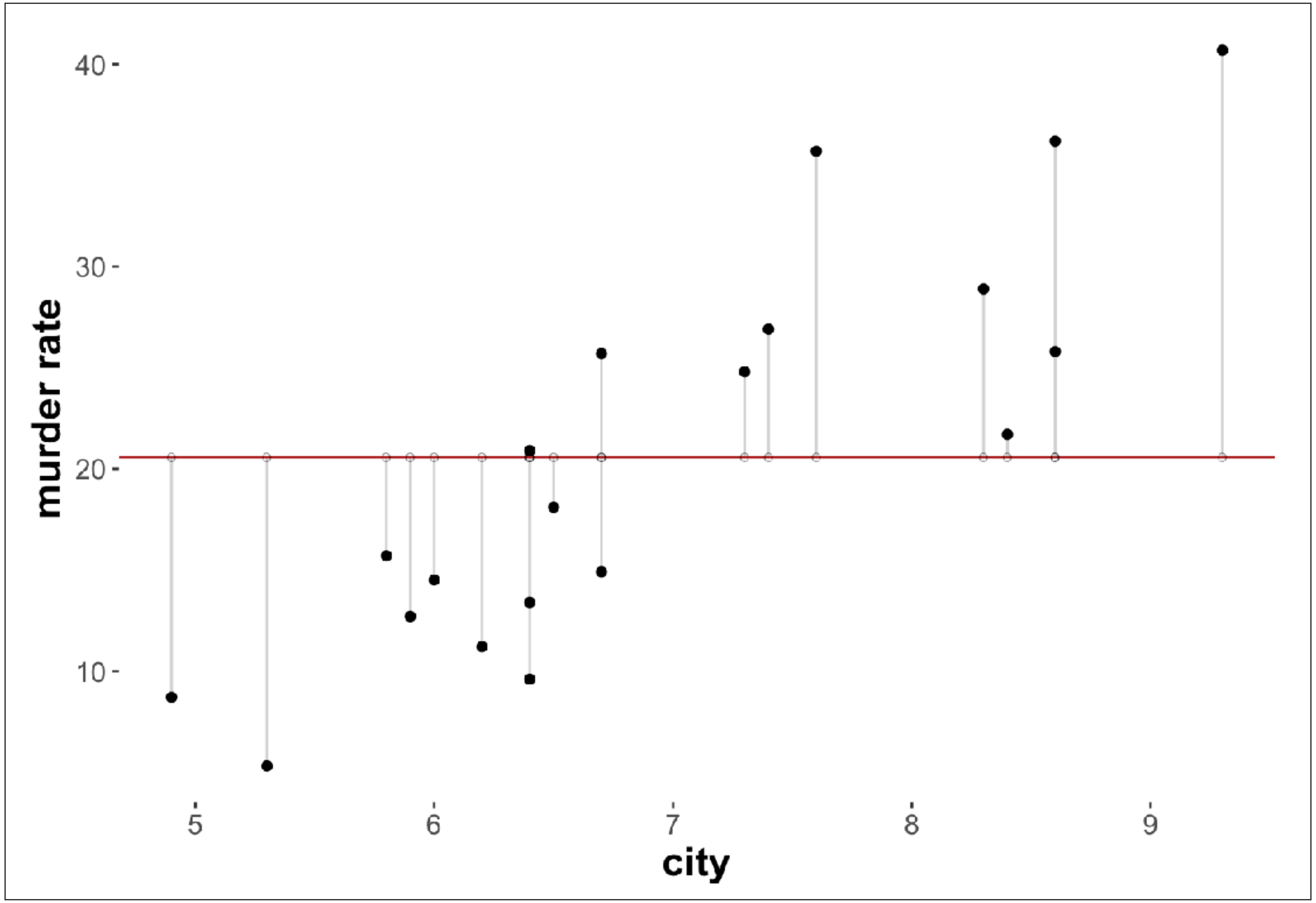
no information at all

```
## # A tibble: 20 x 4
##    murder_rate low_income unemployment population
##          <dbl>      <dbl>        <dbl>      <dbl>
##  1        11.2       16.5          6.2     587000
##  2        13.4       20.5          6.4     643000
##  3        40.7       26.3          9.3     635000
##  4         5.3       16.5          5.3     692000
##  5        24.8       19.2          7.3    1248000
##  6        12.7       16.5          5.9     643000
##  7        20.9       20.2          6.4    1964000
##  8        35.7       21.3          7.6    1531000
##  9         8.7       17.2          4.9     713000
## 10         9.6       14.3          6.4     749000
## 11        14.5       18.1          6      7895000
## 12        26.9       23.1          7.4     762000
## 13        15.7       19.1          5.8    2793000
## 14        36.2       24.7          8.6     741000
## 15        18.1       18.6          6.5     625000
## 16        28.9       24.9          8.3     854000
## 17        14.9       17.9          6.7     716000
## 18        25.8       22.4          8.6     921000
## 19        21.7       20.2          8.4     595000
## 20        25.7       16.9          6.7    3353000
```



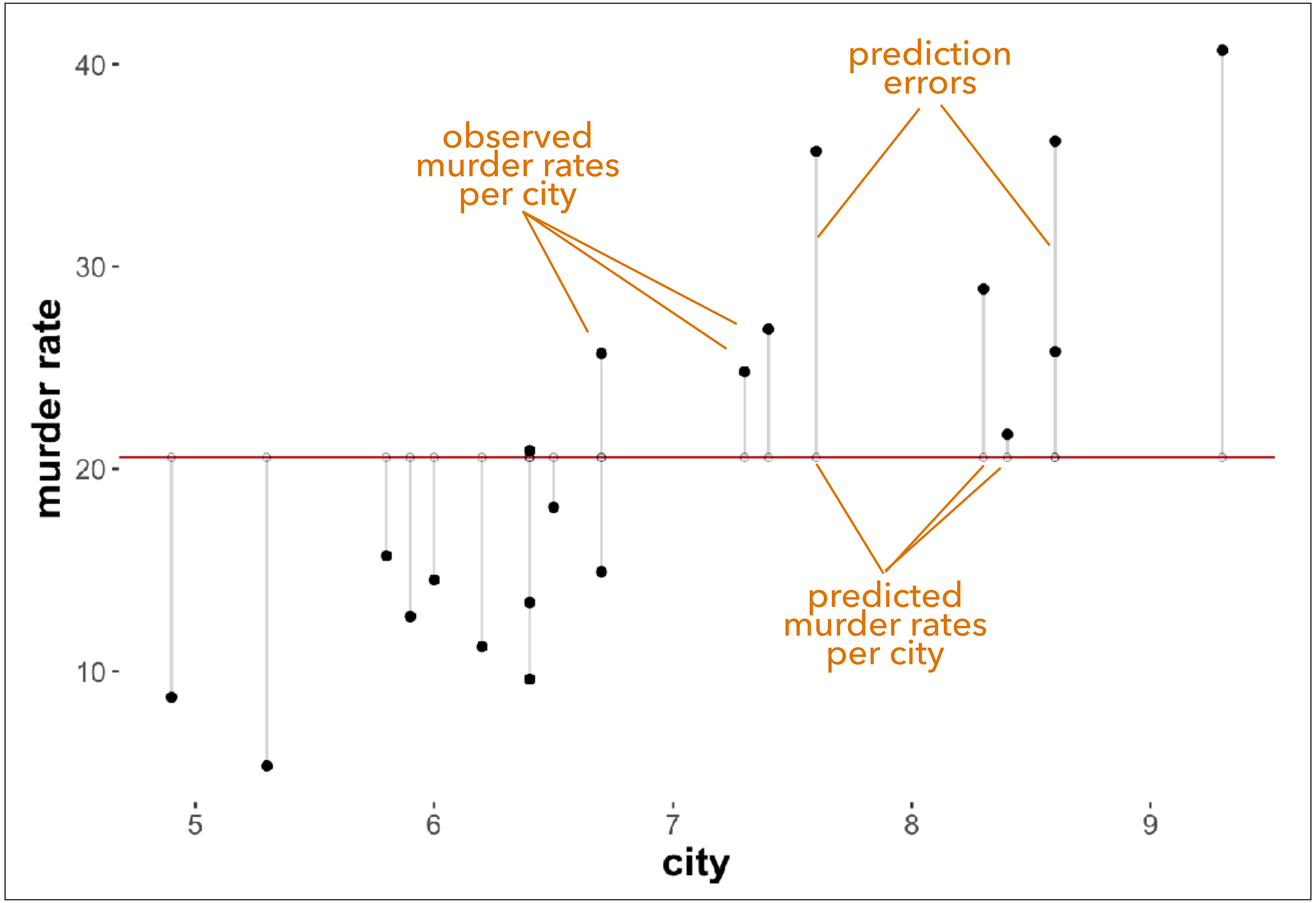
Predicting murder rate

by empirical mean



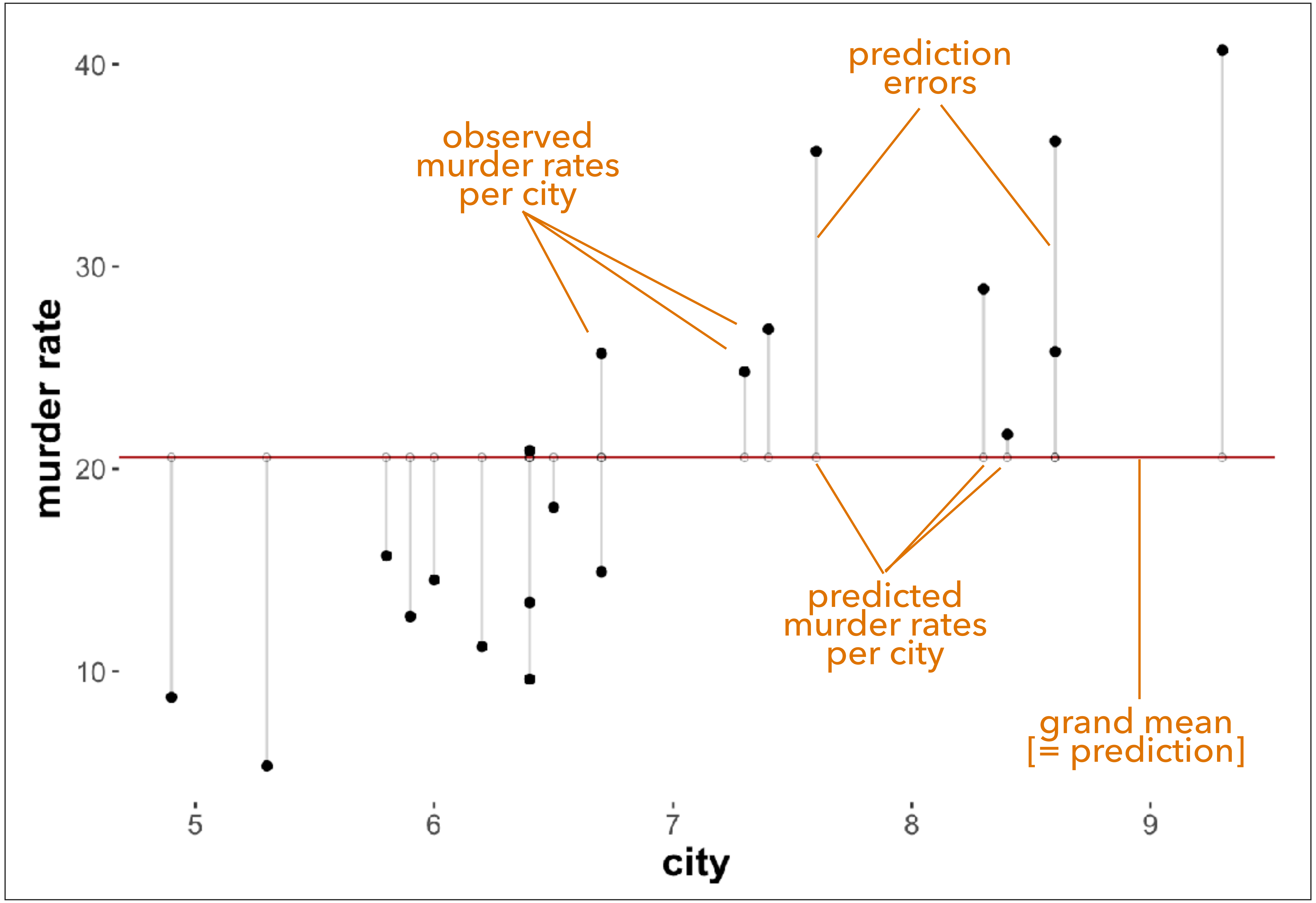
Predicting murder rate

by grand mean



Predicting murder rate

by grand mean



$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

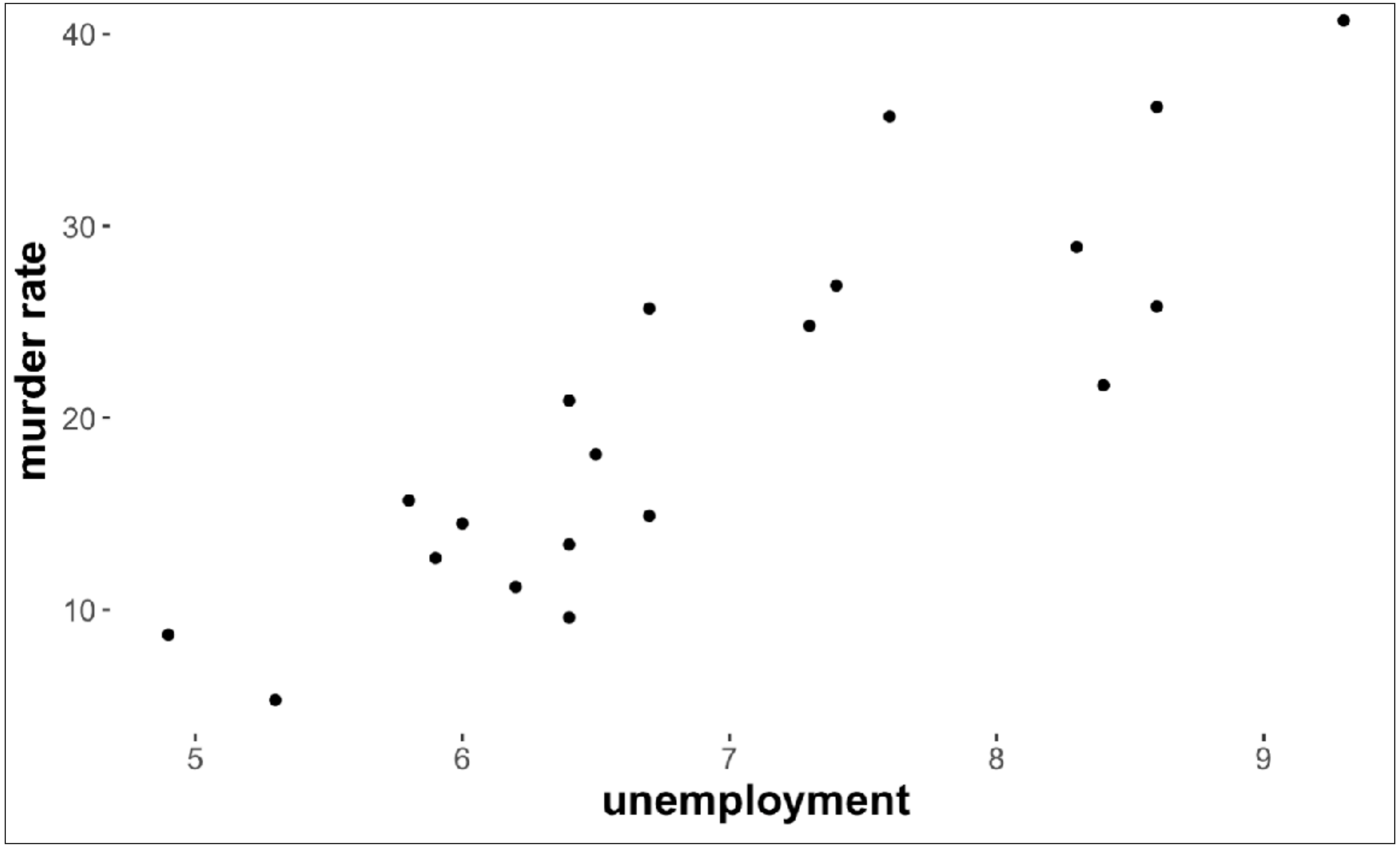
[total sum of squares]

```
y <- murder_data %>% pull(murder_rate)
n <- length(y)
tss_simple <- sum((y - mean(y))^2)
tss_simple
```

```
## [1] 1855.202
```

Predicting murder rate based on unemployment rate

some wild linear guessing



We are to predict the murder rate y_i of a randomly drawn city i. **We know that city's unemployment rate, x_i, but nothing more.**

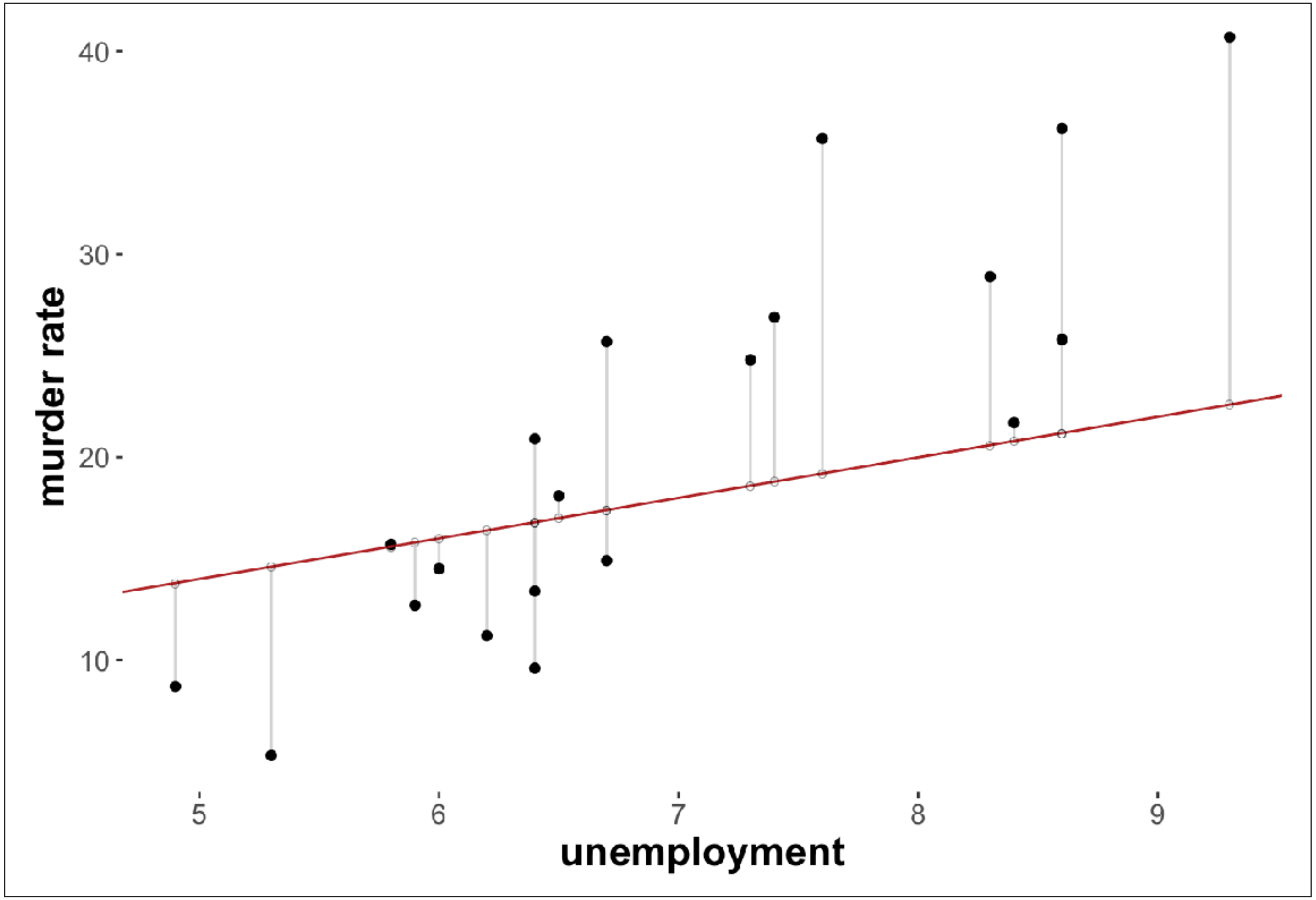
Let's just assume the following **linear relationship** to make a prediction b/c why not?!?

$$\hat{y}_i = 4 + 2x_i$$

How good is this prediction?

How good is any given prediction?

quantifying distance or likelihood



Distance-based approach

Residual Sum-of-Squares:

$$\text{RSS} = \sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2$$

- no predictions about spread around linear predictor

Likelihood-based approach:

Normal likelihood:

$$\text{LH} = \prod_{i=1}^{n} \mathcal{N}\left(y_i \mid \mu = \hat{y}_i, \sigma\right)$$

- fully predictive

Likelihood-based simple linear regression

- likelihood:

  $$y_i \sim \text{Normal}(\mu_i, \sigma)$$
  $$\mu_i = \beta_0 + x_i \cdot \beta_1$$

- differential likelihood:
  - parameter triples $\langle \beta_0, \beta_1, \sigma \rangle$ can be better or worse
  - higher vs. lower likelihood $P(D \mid \beta_0, \beta_1, \sigma)$

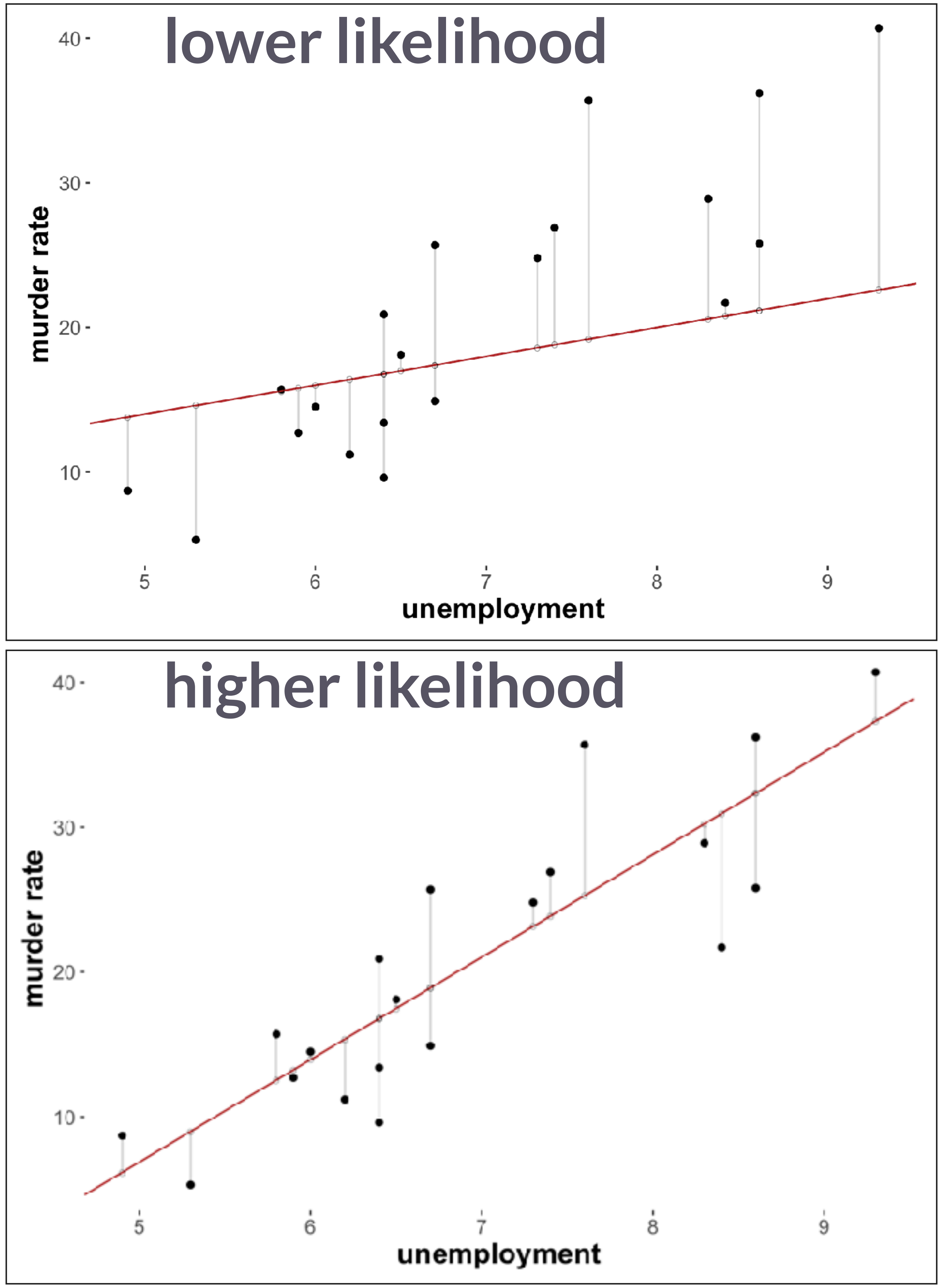
- maximum-likelihood solution:

  $$\arg\max_{\beta_0, \beta_1, \sigma} P(D \mid \beta_0, \beta_1, \sigma)$$

  - standard (frequentist) solution
  - MLE corresponds to MAP for “flat” priors

- Bayesian approach: full posterior distribution

  $$P(\beta_0, \beta_1, \sigma \mid D) \propto P(D \mid \beta_0, \beta_1, \sigma)\ P(\beta_0, \beta_1, \sigma)$$



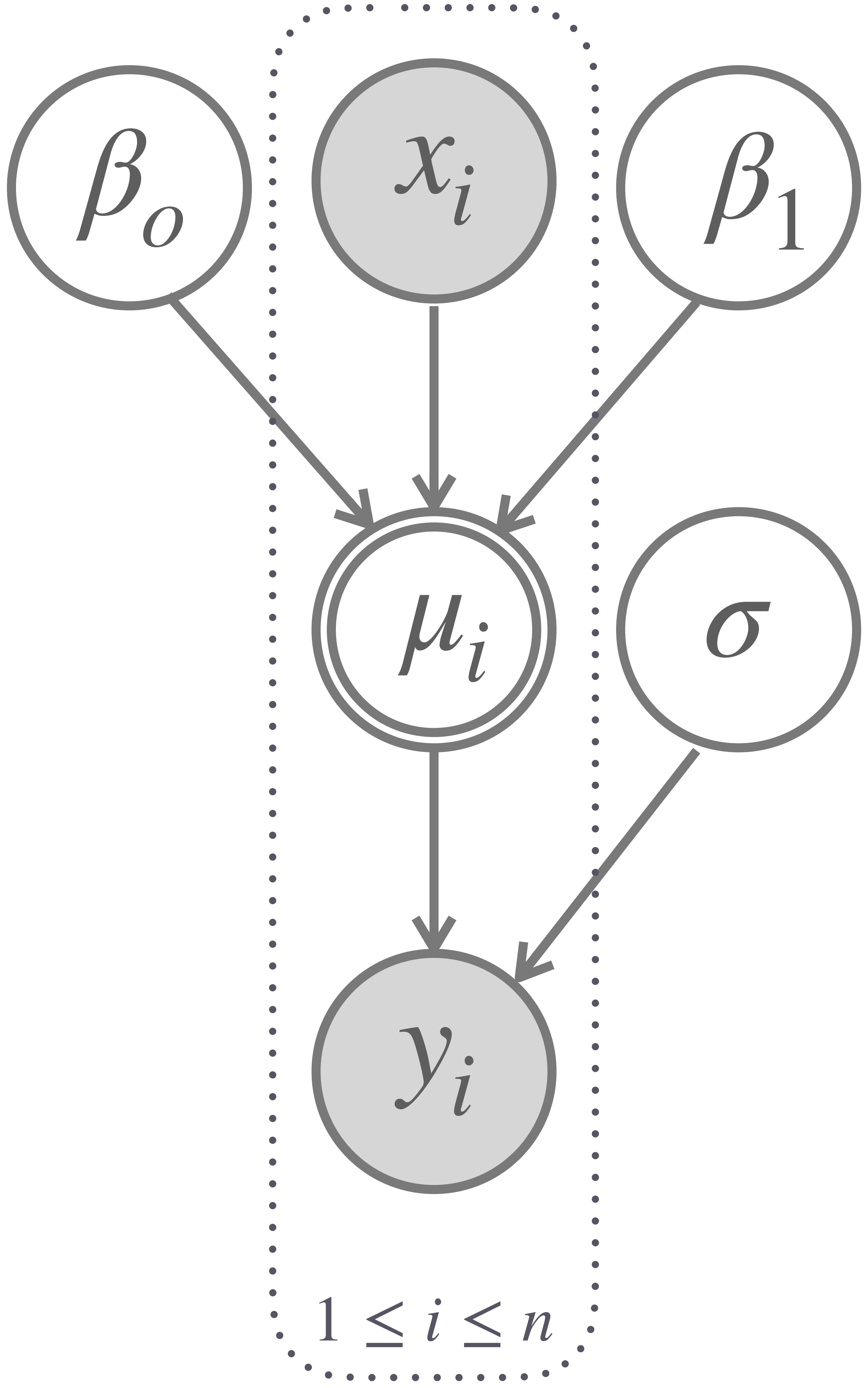
Simple linear regression model

for a single predictor variable

- data: n pairs of numbers $D = \{\langle x_1, y_1\rangle, \ldots \langle x_n, y_n\rangle\}$
  - x_i is the i-th observation of the **independent / predictor variable**
  - y_i is the i-th observation of the **dependent / response variable**
- parameters:
  - β_0 is the **intercept** parameter
  - β_1 is the **slope** parameter
  - σ is the standard deviation of a normal distribution
- derived variable: [shown in node w/ double lines]
  - μ_i is the linear predictor for observation i
- priors (uninformed):

  $\beta_0, \beta_1 \sim \text{Uniform}(-\infty, \infty) \qquad \log(\sigma^2) \sim \text{Uniform}(-\infty, \infty)$
- likelihood:

  $y_i \sim \text{Normal}(\mu_i, \sigma) \qquad \mu_i = \beta_0 + x_1 \cdot \beta_1$



demo

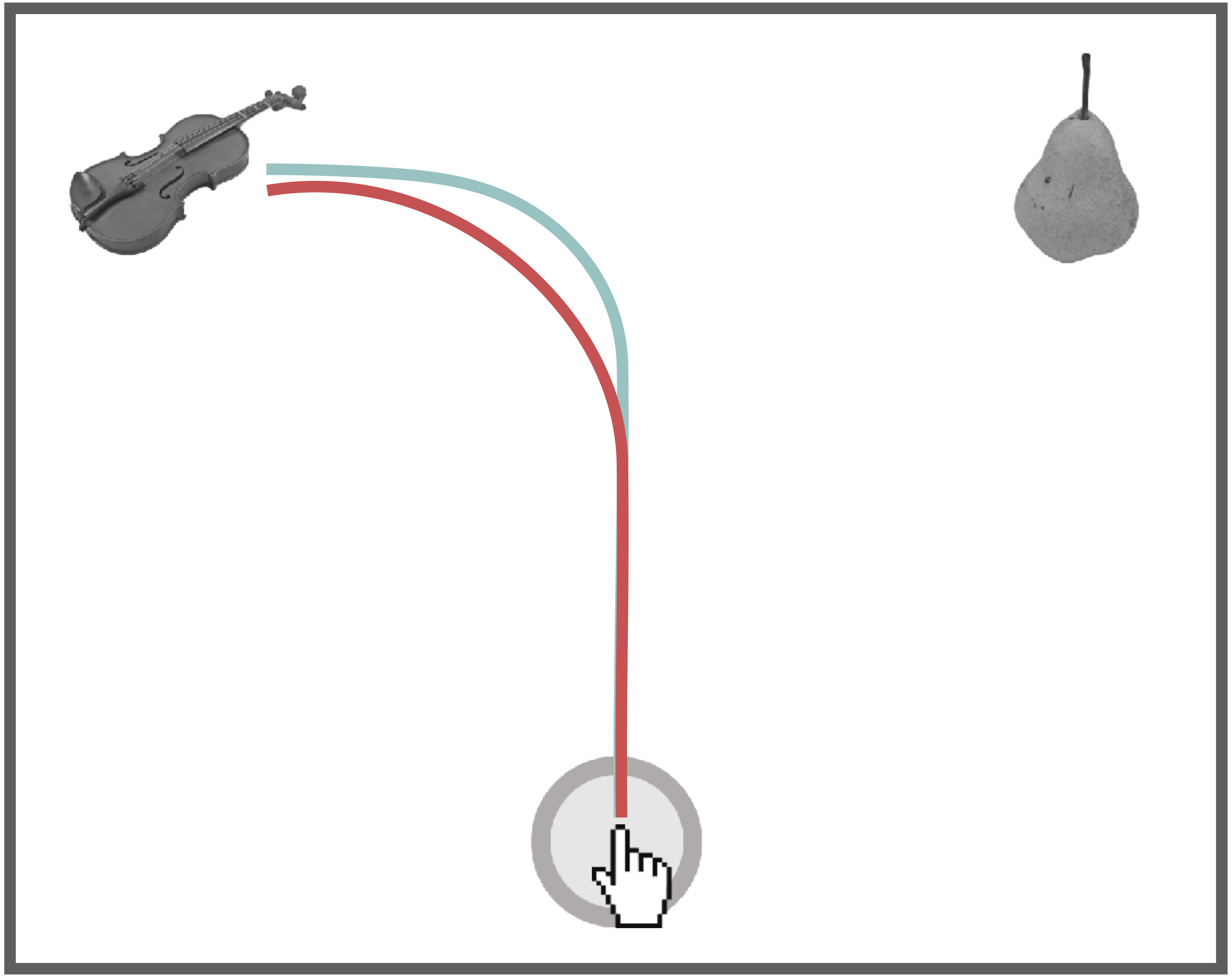
simple linear regression in WebPPL

Mouse-tracking data on typicality in category decisions

Mouse-tracking

Hand-movement during decision making

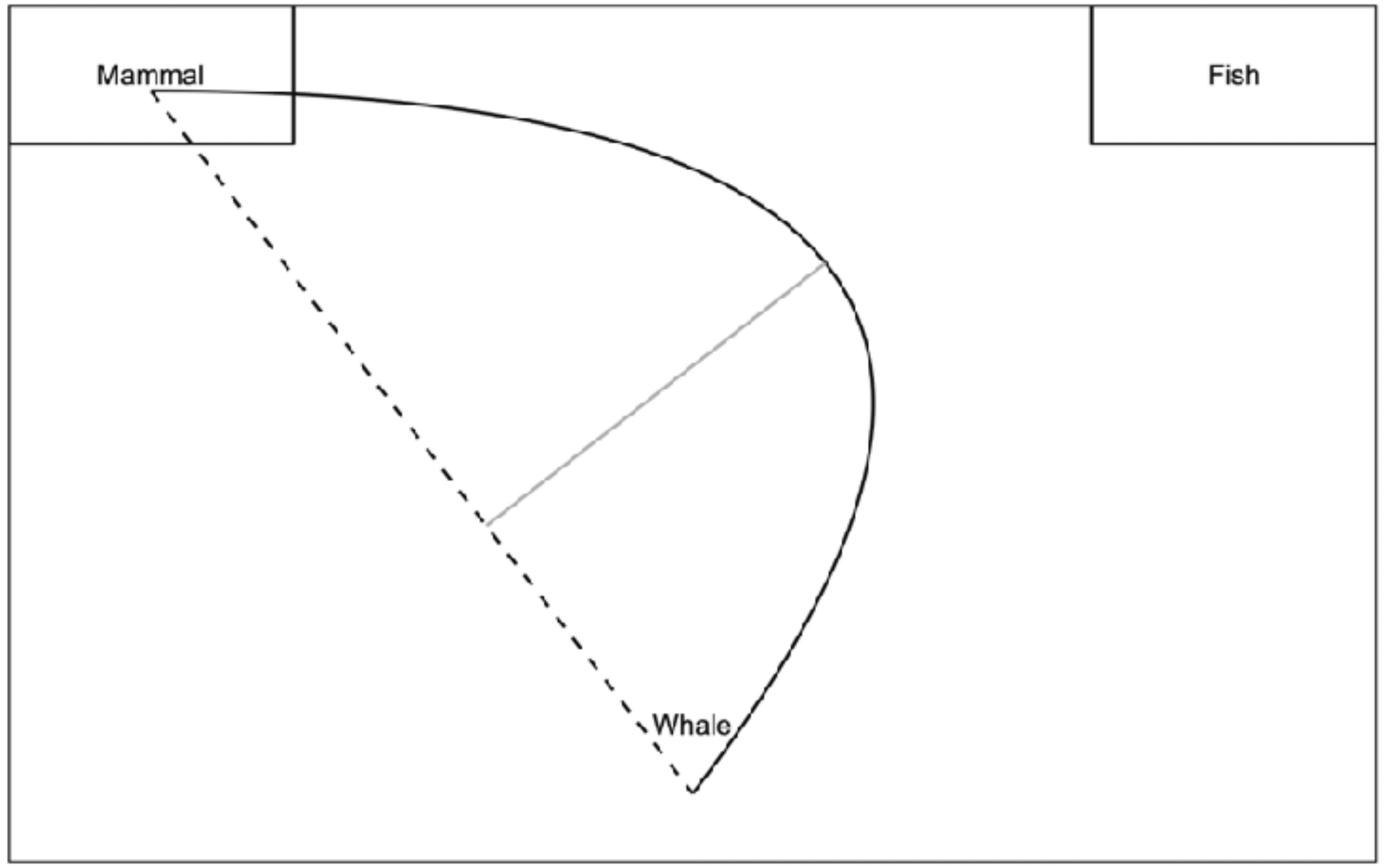
- general idea: motor-execution provides information about the ongoing decision process
  - uncertainty
  - gradual evidence accumulation
  - change-of-mind
  - time-point of decision
  - ...
- many subtle design decisions
  - click vs touch
  - move horizontally or vertically
  - ...

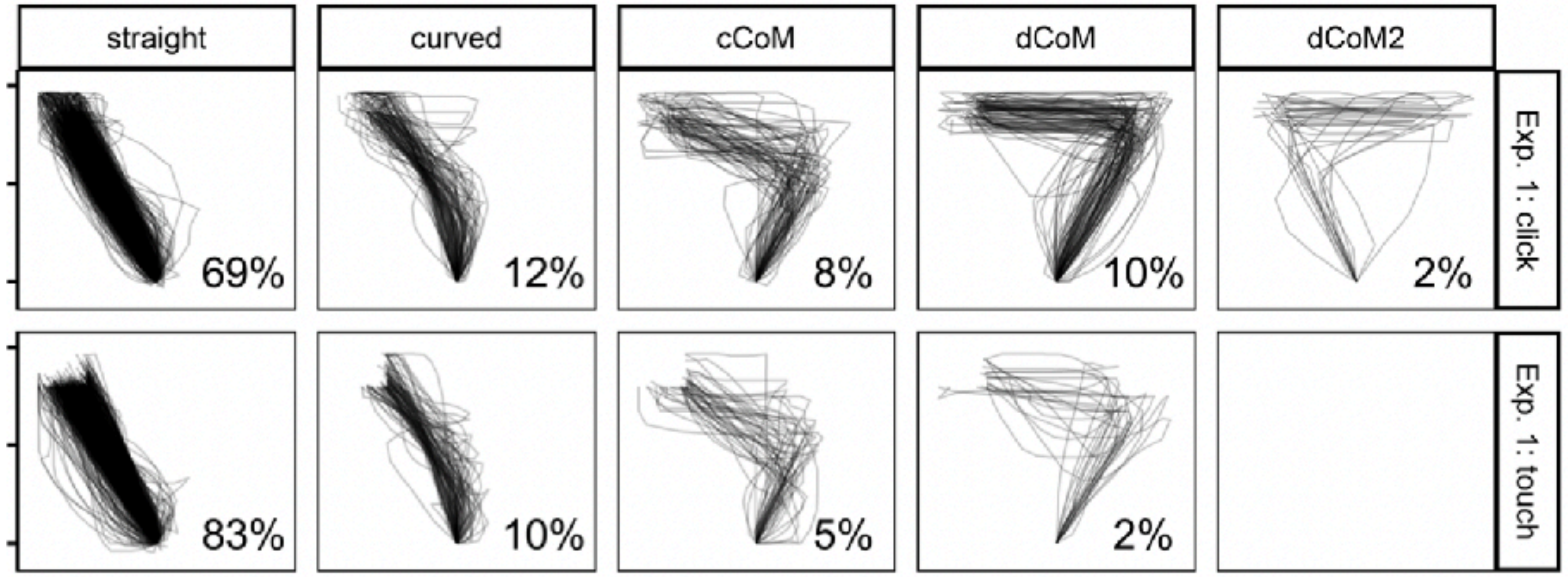


Mouse-tracking

common measures of mouse-trajectories

- raw data are lists of triples
  - (time, x-position, y-position)
- commonly used measures
  - area-under the curve (AUC)
    - area between the mouse trajectory and a straight line from start to selected option
  - maximal deviation (MAD)
    - maximum distance between trajectory and straight line from start to selected option
  - correctness
    - whether choice of option was correct or not
  - reaction time (RT)
    - how long did the movement last in total
  - type of trajectory
    - result of clustering analysis based on shape of the trajectories (usually some 3-5 categories)
  - x-flips
    - number of times the trajectory crossed the vertical middle line (at x = 0)





Running example

category recognition for typical vs atypical exemplars

- materials & procedure
  - participants read an animal name (e.g. 'dolphin')
  - they choose the true category the animal belongs to (e.g., 'fish' or 'mammal')
  - some trigger words are typical others atypical representatives of the true category
- methodological investigation:
  - two groups: **click vs touch** to select category
- **hypothesis:** typical exemplars are easier to categorize than atypical ones
  - fewer mistakes
  - smaller RTs, AUC, MAD
  - less x-flips
  - less "change-of-mind" curve types
- **research question (methods):** any differences between click & touch selection?

variables used in the data set

`trial_id` = unique id for individual trials
`MAD` = maximal deviation into competitor space
`AUC` = area under the curve
`xpos_flips` = the amount of horizontal direction changes
`RT` = reaction time in ms
`prototype_label` = different categories of prototypical movement strategies
`subject_id` = unique id for individual participants
`group` = groups differ in the response design (click vs. touch)
`condition` = category membership (Typical vs. Atypical)
`exemplar` = the concrete animal
`category_left` = the category displayed on the left
`category_right` = the category displayed on the right
`category_correct` = the category that is correct
`response` = the selected category
`correct` = whether or not the `response` matches `category_correct`

Kieslich et al.'s (2019) replication of Dale et al.'s (2003) experiment

outlook

Three pillars of BDA

1. parameter estimation / inference [which parameter values are credible given data and model?]

$$\underbrace{P(\theta \mid D)}_{\text{posterior}} \propto \underbrace{P(\theta)}_{\text{prior}} \times \underbrace{P(D \mid \theta)}_{\text{likelihood}}$$

2. predictions [which future data observations are likely given my model?]

a. prior

$$P(D_{\text{pred}}) = \int P(\theta)\ P(D_{\text{pred}} \mid \theta)\ \mathrm{d}\theta$$

b. posterior

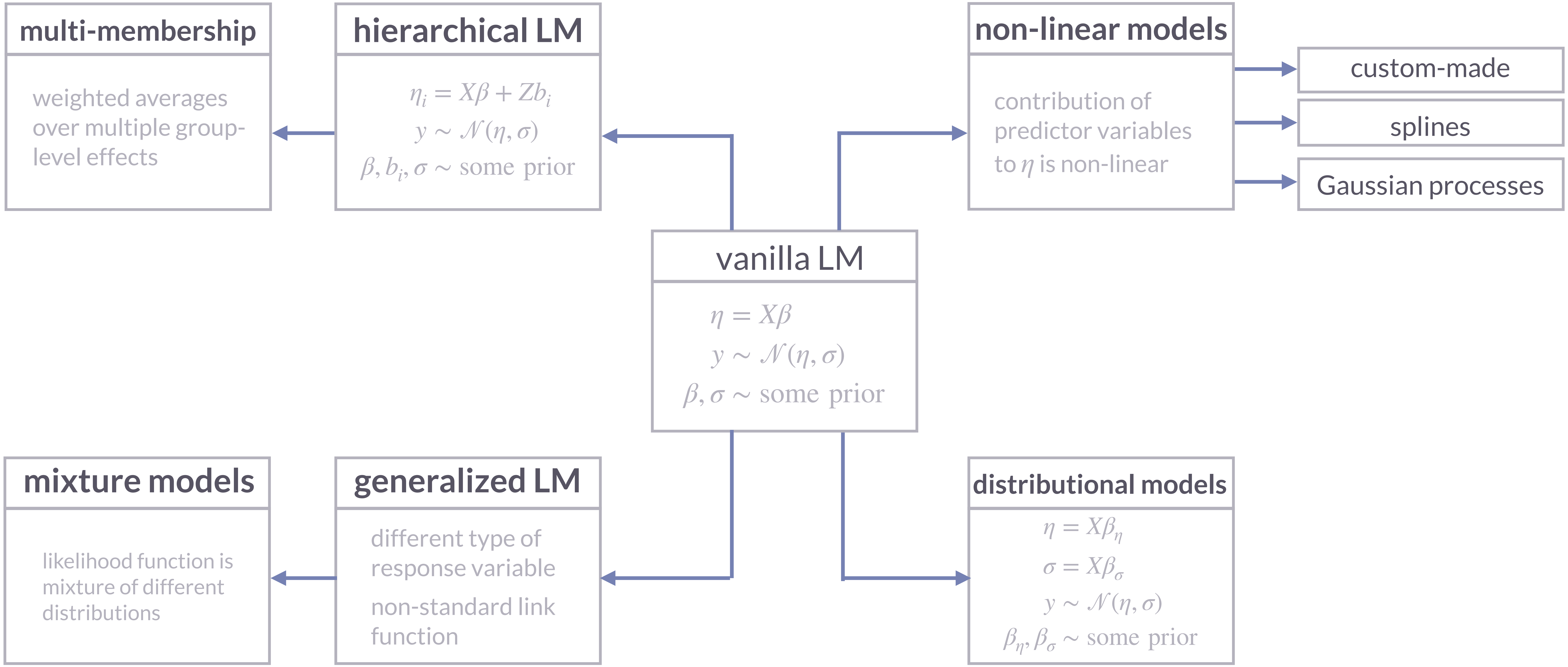
$$P(D_{\text{pred}} \mid D_{\text{obs}}) = \int P(\theta \mid D_{\text{obs}})\ P(D_{\text{pred}} \mid \theta)\ \mathrm{d}\theta$$

3. model comparison [which model of two models is more likely to have generated the data?]

$$\underbrace{\frac{P(M_1 \mid D)}{P(M_2 \mid D)}}_{\text{posterior odds}} = \underbrace{\frac{P(D \mid M_1)}{P(D \mid M_2)}}_{\text{Bayes factor}}\ \underbrace{\frac{P(M_1)}{P(M_2)}}_{\text{prior odds}}$$

Roadmap “beyond vanilla”

common extensions of linear regression modeling



recap & preparation

Recap & preparation

- check out web-book for this course
  - https://michael-franke.github.io/Bayesian-Regression/
- recap:
  - material from 1st session
    - "Thinking Bayesian"
  - basic wrangling & plotting in the `tidyverse`
    - Wrangling & Plotting
- prepare for next session:
  - Big Bayesian 4 for simple regression in BRMs
    - Regression in BRMs & prior & posterior predictives
  - BRMS cheat sheet
    - cheat sheet
  - MCMC methods
    - slides