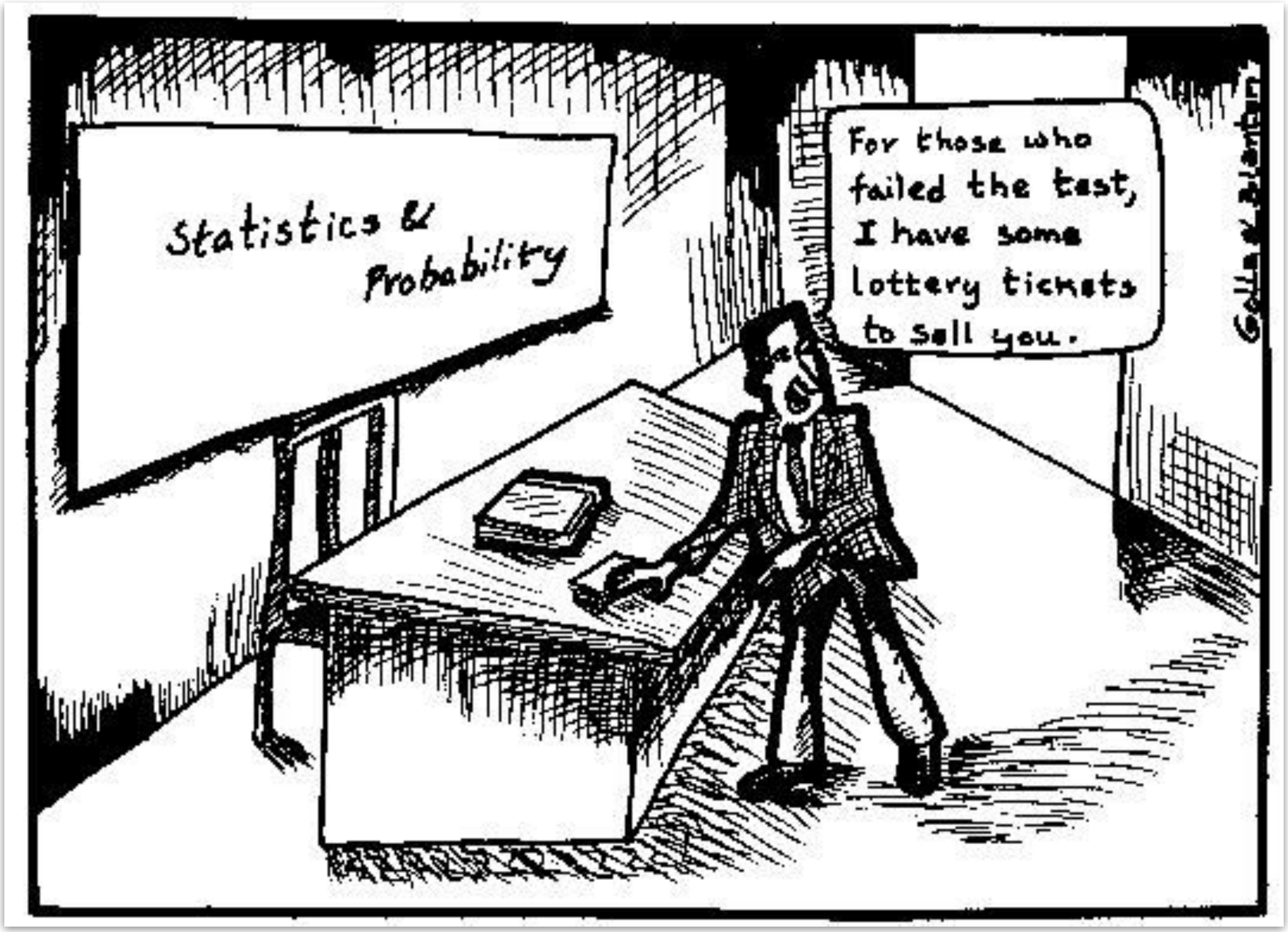
no need to keep
rows 3 and 6 free

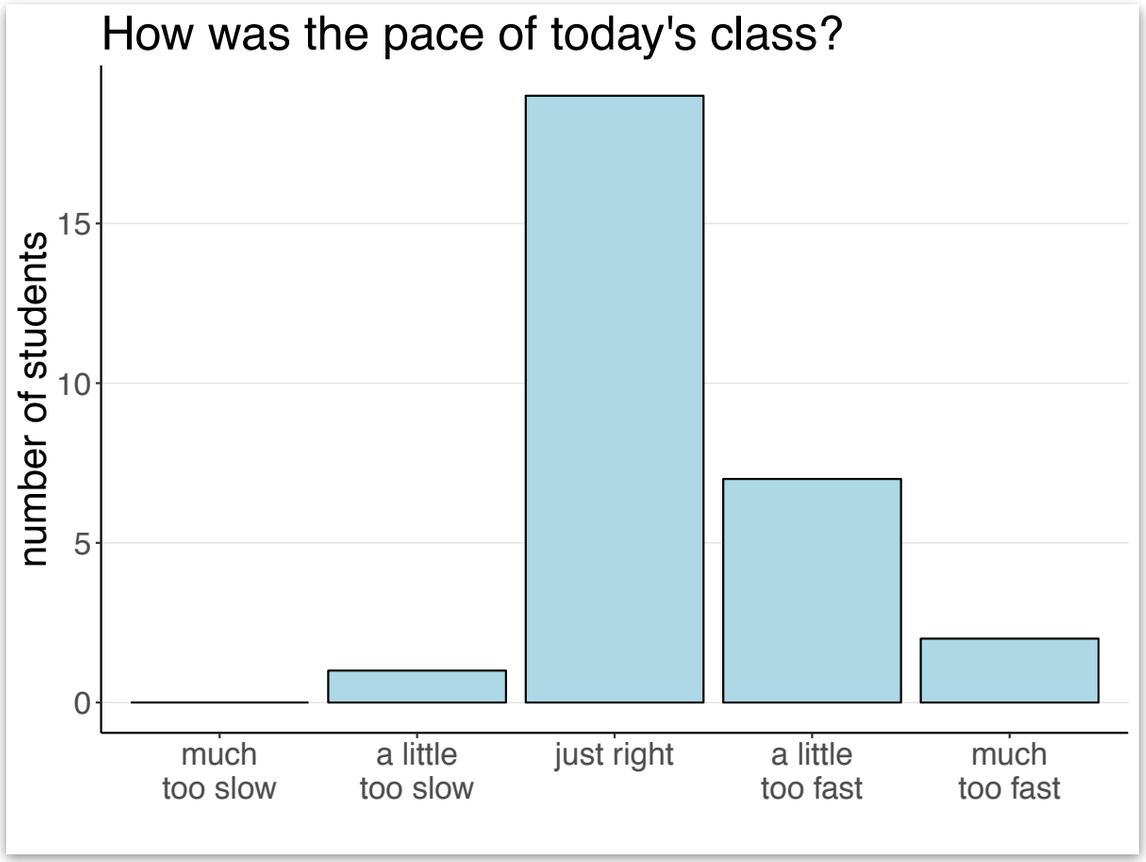
Probability & Causality

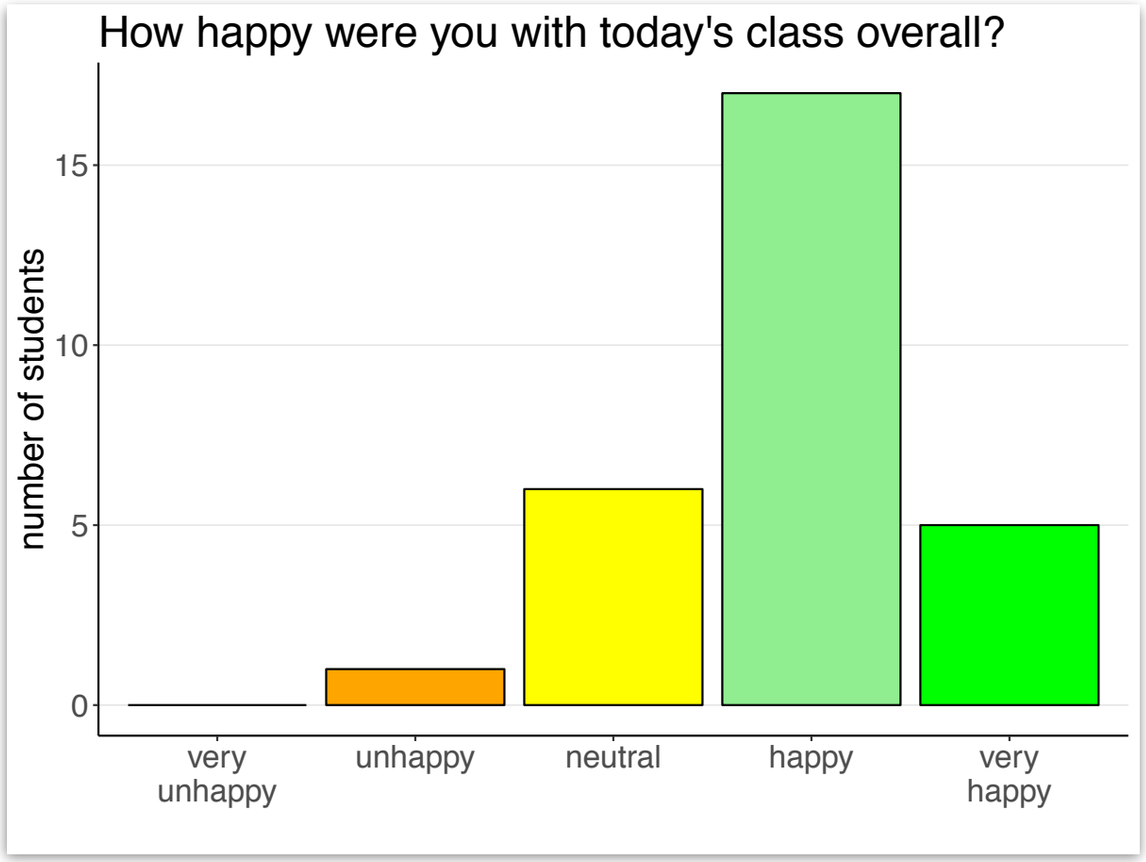


01/18/2019

Your feedback

Your feedback





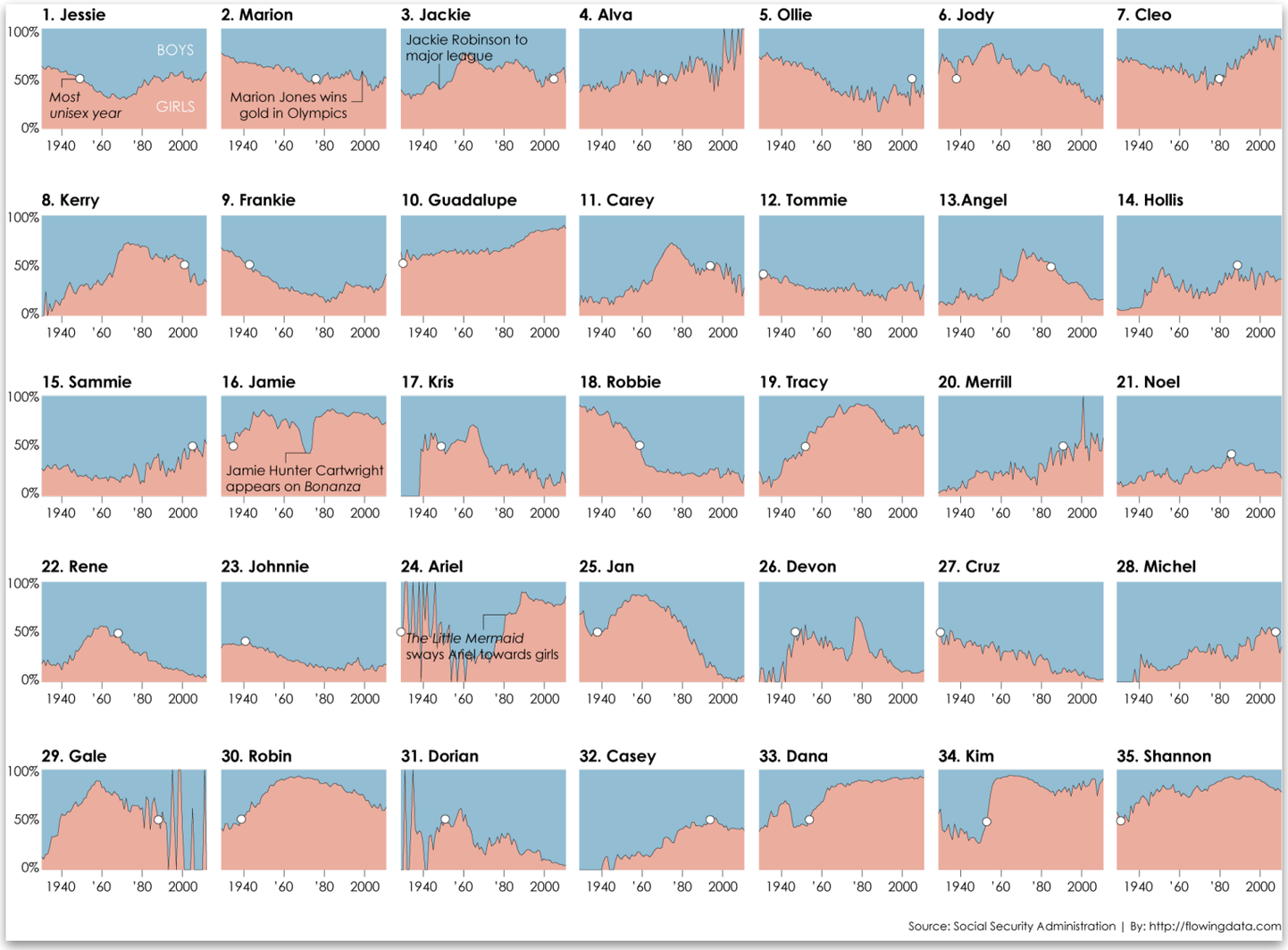
Your feedback

it was great! looking forward to more conceptual lectures haha

alright, let's do it!

Logistics

Homework 2



- *computing* the correct set of names, rather than specifying them manually (problem 2)
- don't worry if your names don't match exactly, as long as your computation makes sense, you'll get full points

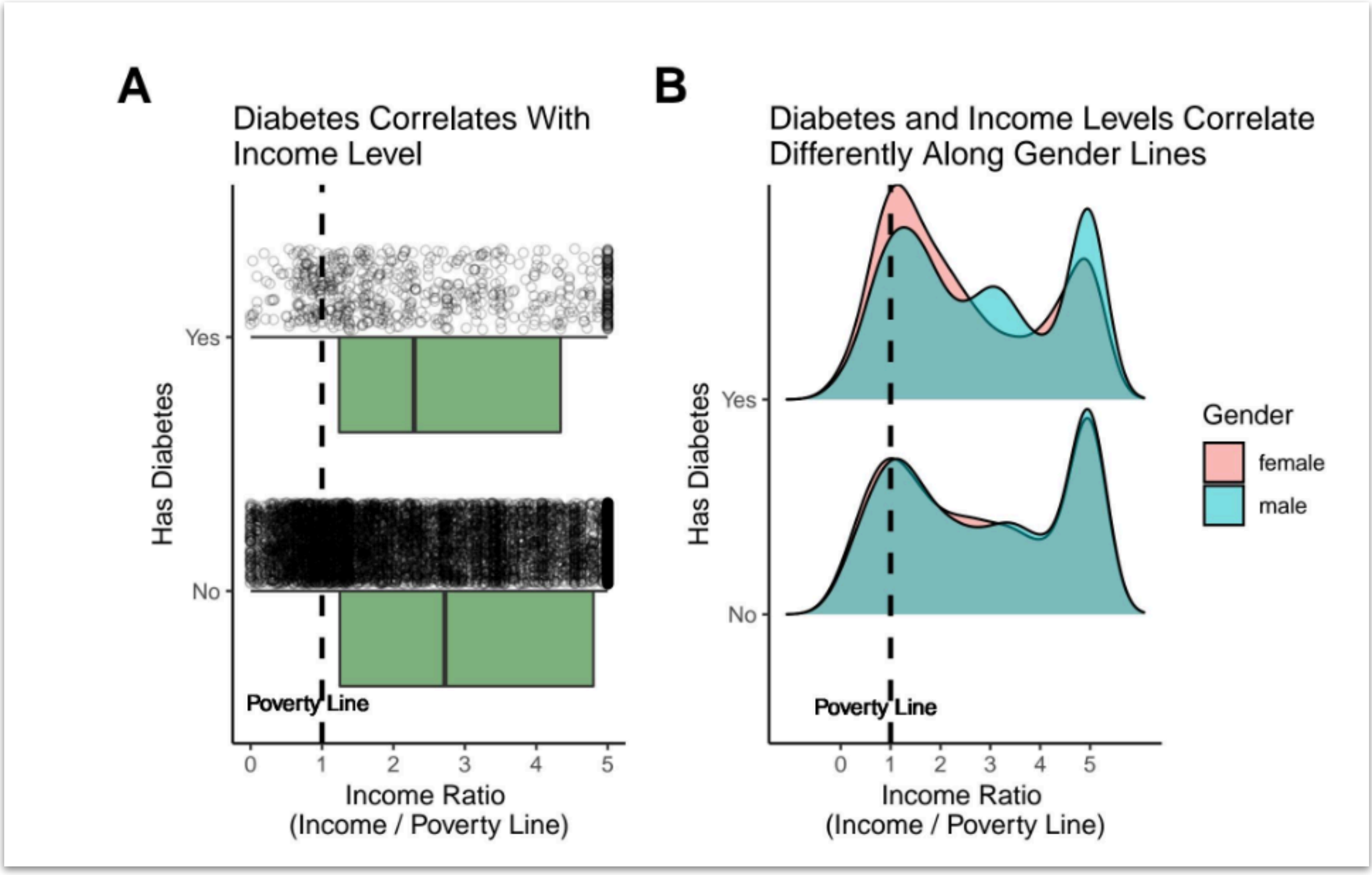
Homework 2

Instructions

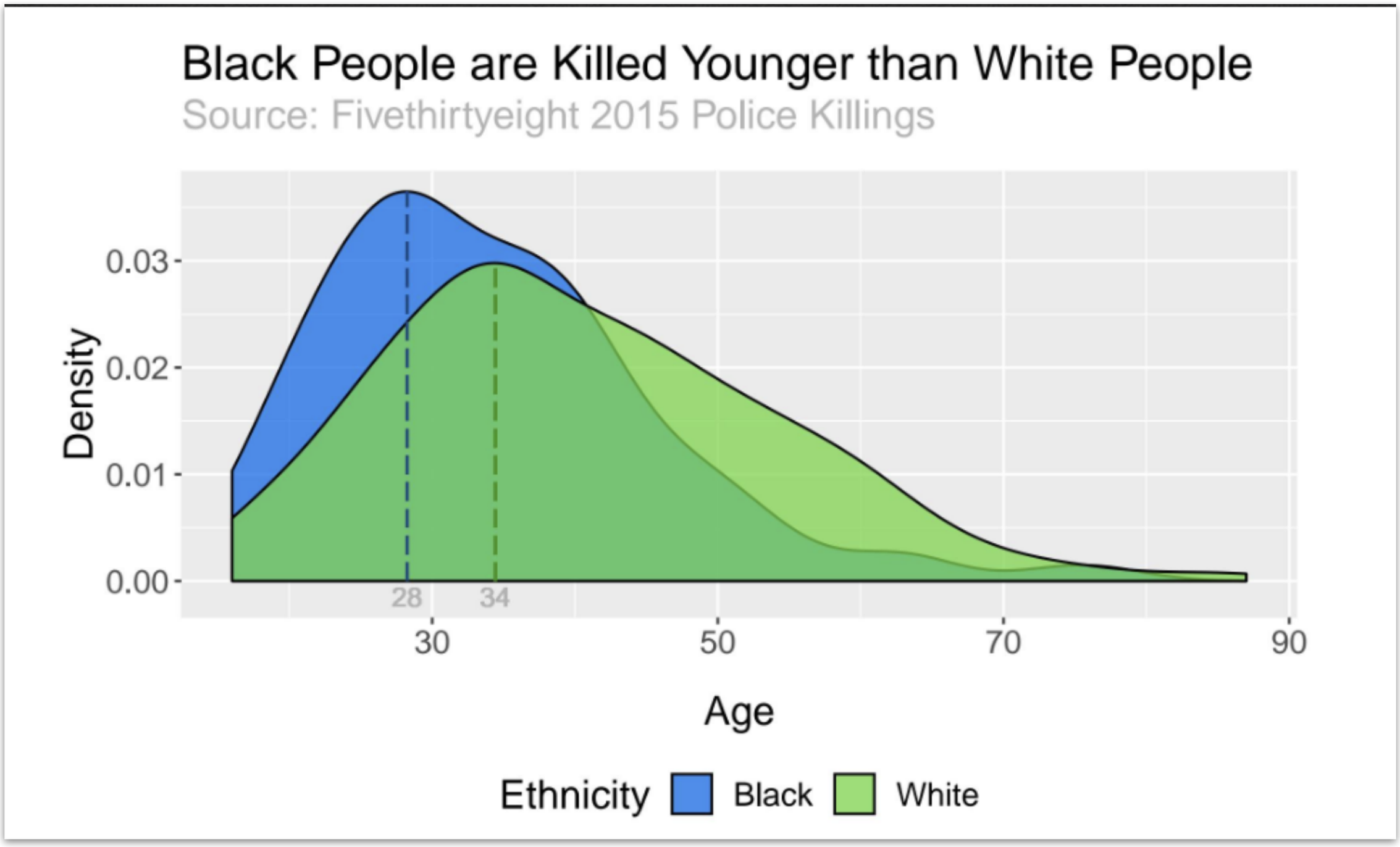
> This homework is due by **Tuesday, January 22nd, 8:00pm**.

> Upload **only the pdf file containing the code and plots** and name it: `yourname_wrangling_homework.pdf`

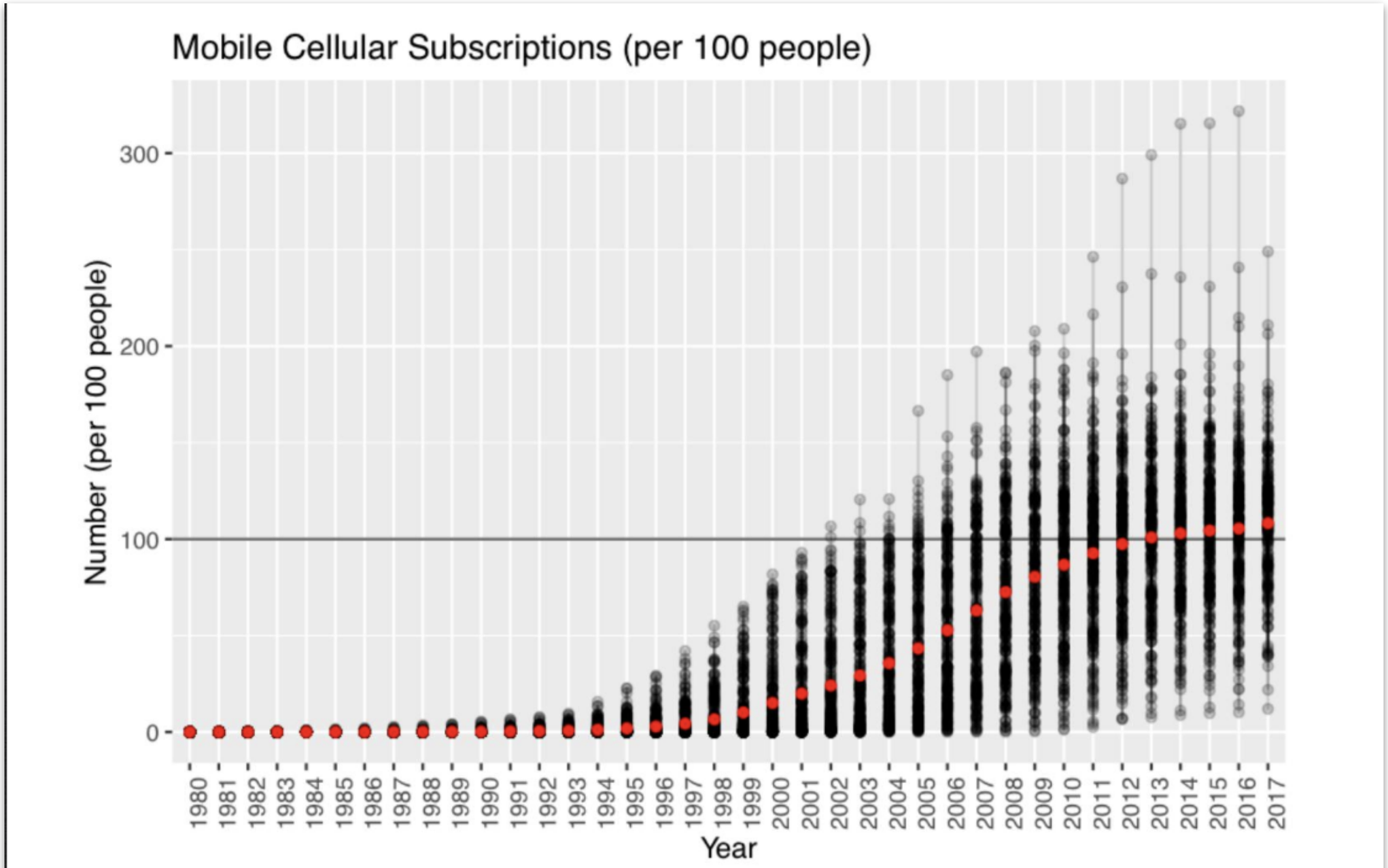
Homework 1


A
Diabetes Correlates With Income Level
Has Diabetes
Yes
No
Poverty Line
0
1
2
3
4
5
Income Ratio
(Income / Poverty Line)
B
Diabetes and Income Levels Correlate Differently Along Gender Lines
Has Diabetes
Yes
No
Poverty Line
0
1
2
3
4
5
Income Ratio
(Income / Poverty Line)
Gender
female
male

Homework 1

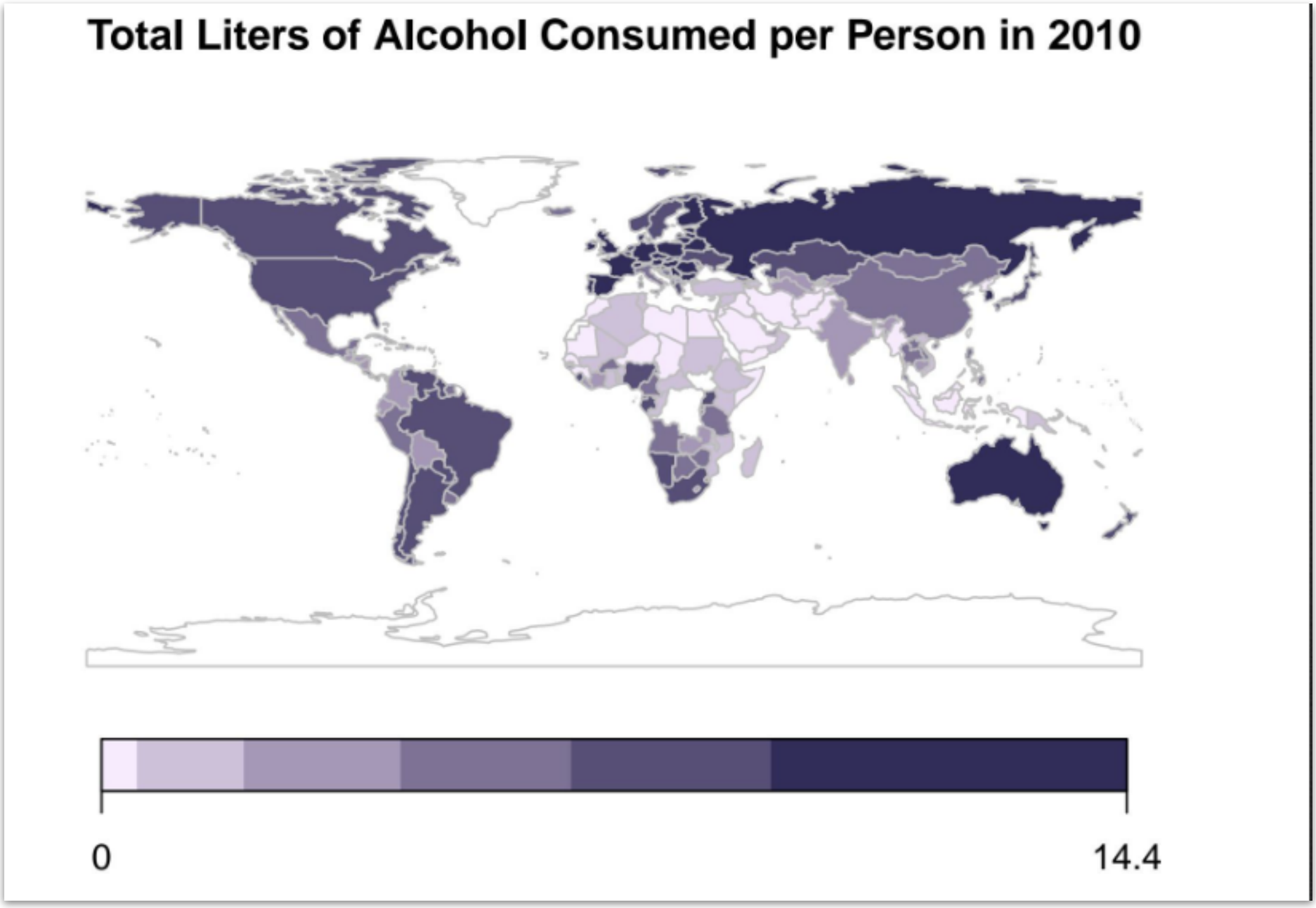

Black People are Killed Younger than White People
Source: Fivethirtyeight 2015 Police Killings
Density
0.03
0.02
0.01
0.00
28
34
30
50
70
90
Age
Ethnicity
Black
White

Homework 1

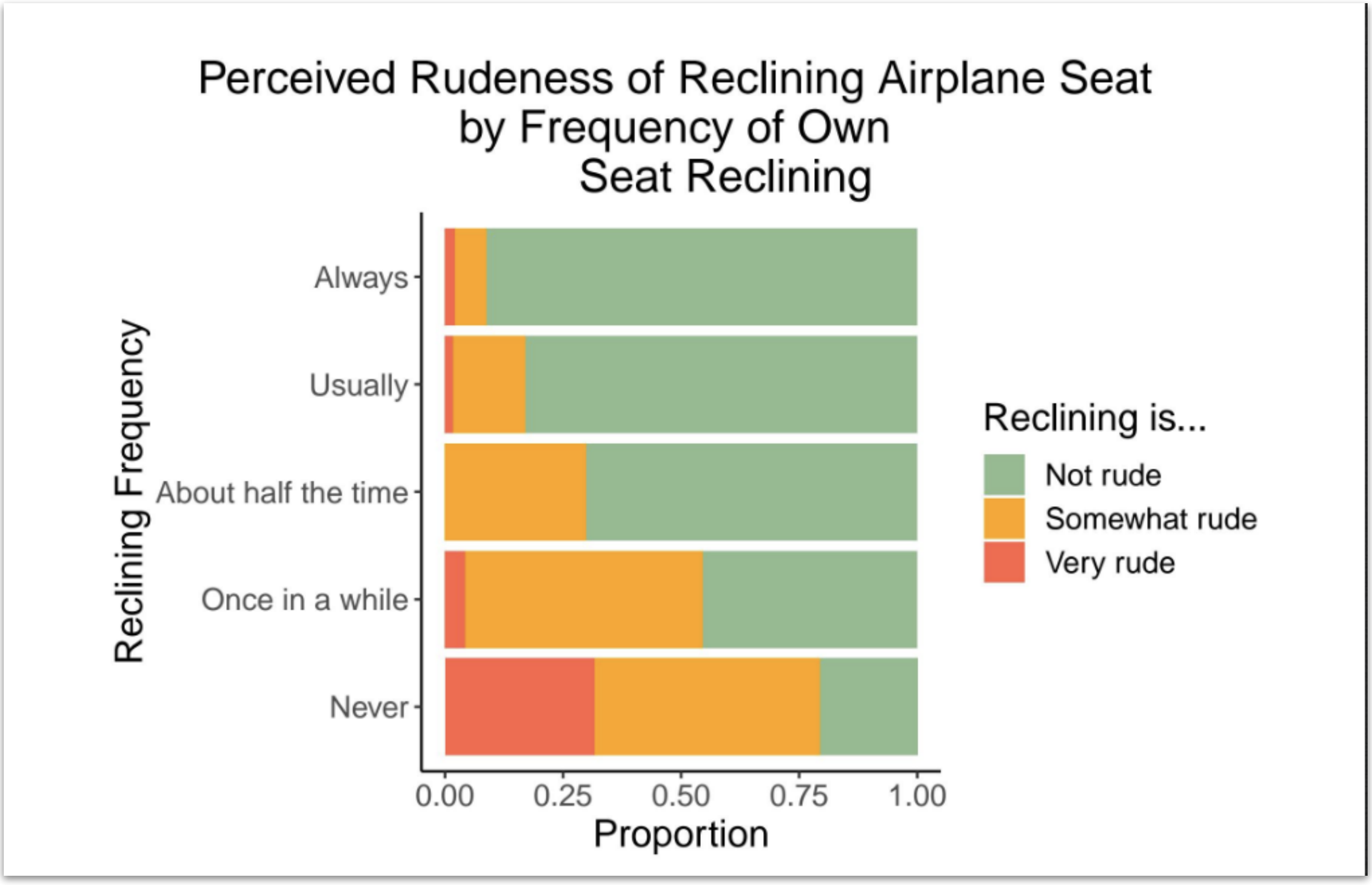


Mobile cellular subscriptions (per 100 people) from 1980 to 2017. Black point represent the average number of subscriptions for a country, red points represent the mean across all countries. The gray horizontal line represents the point at which each person on average has a single cell phone subscription.

Homework 1



Homework 1

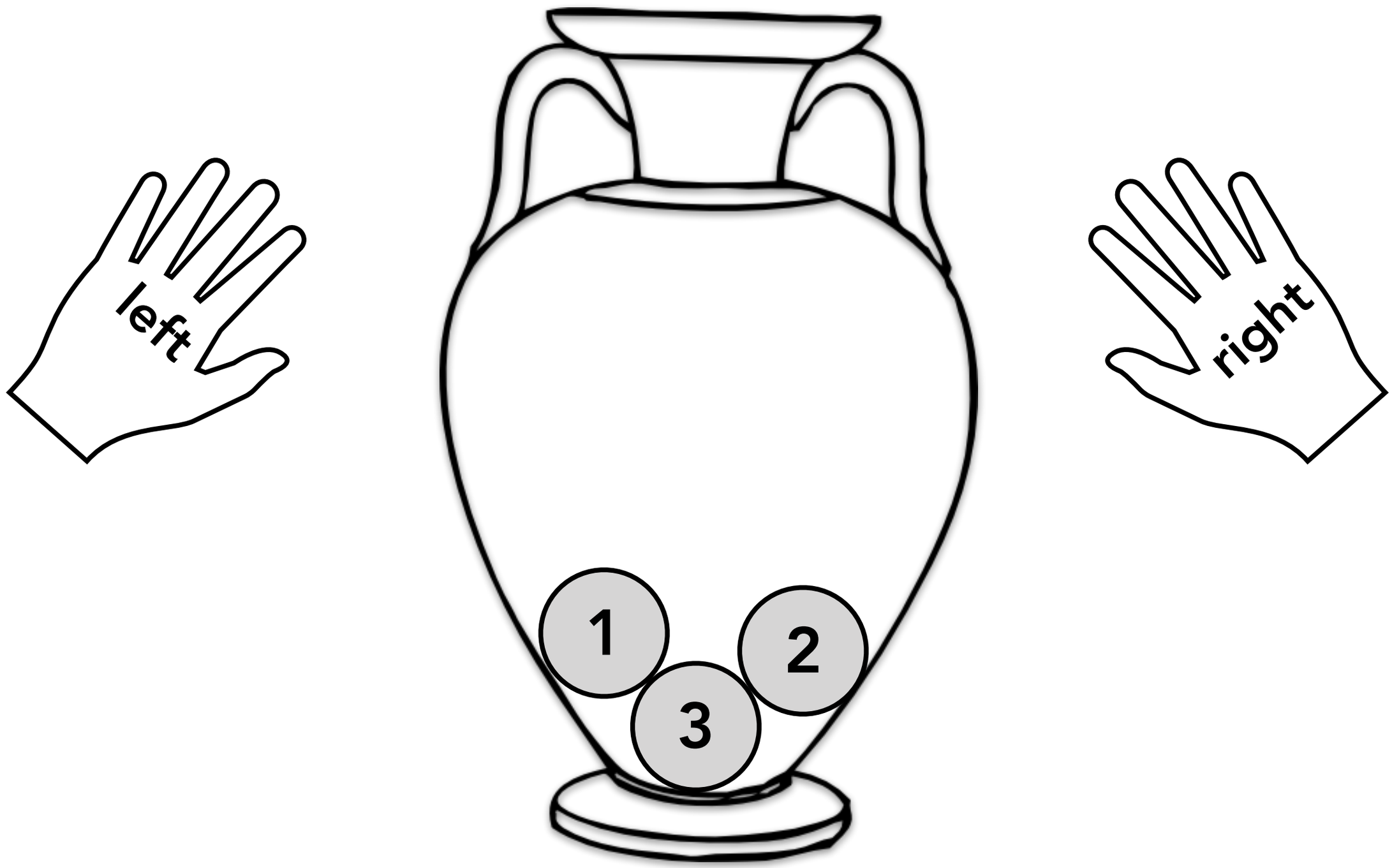

Perceived Rudeness of Reclining Airplane Seat
by Frequency of Own
Seat Reclining
Reclining Frequency
Always
Usually
About half the time
Once in a while
Never
0.00
0.25
0.50
0.75
1.00
Proportion
Reclining is...
Not rude
Somewhat rude
Very rude

Outline

- Introduction to probability / Recap
  - Counting possibilities
  - Interpretation of probability
  - Clue guide to probability
- Bayesian Networks
  - representation
  - inference
  - (un-)conditional (in-)dependence
- Causal Bayes nets

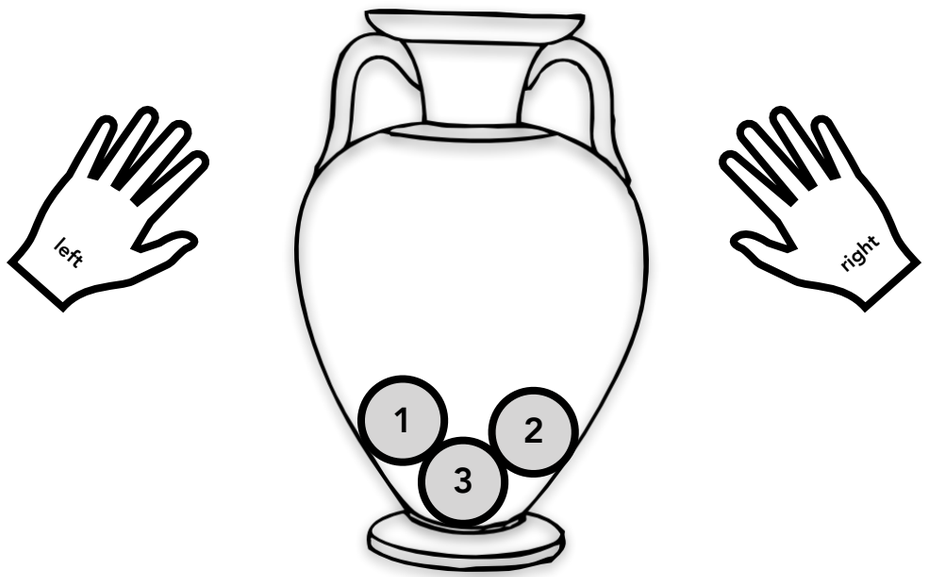
Counting possibilities

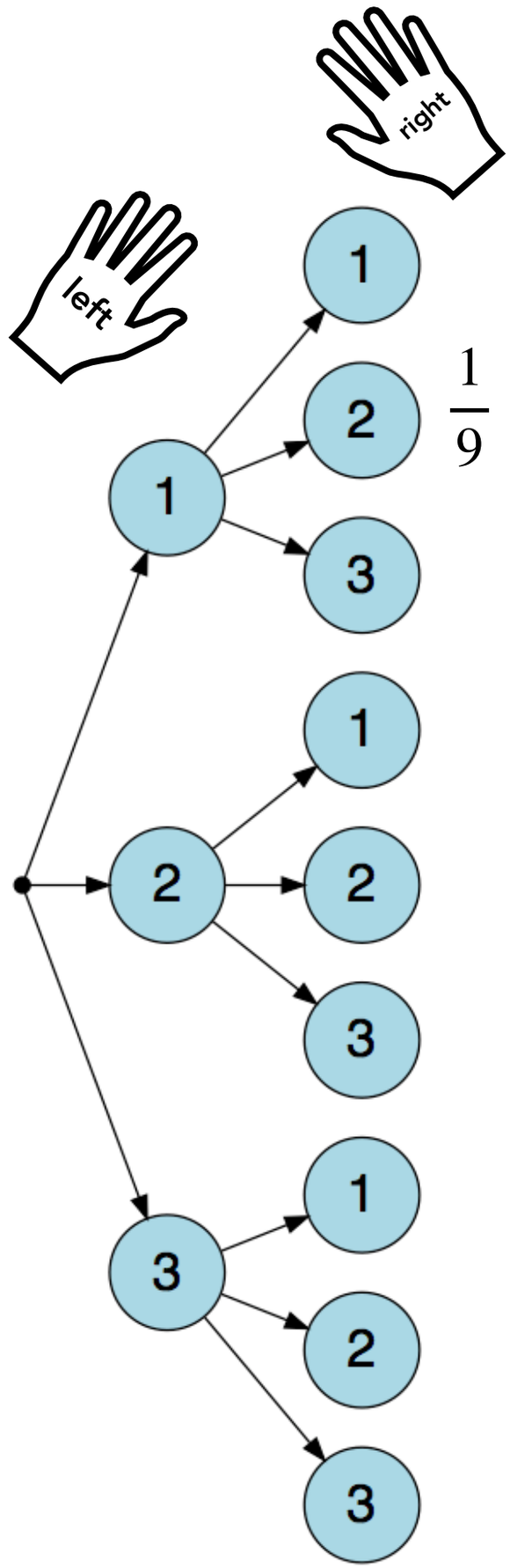
no stats class without urns!



Sampling **with** replacement

$$p(\text{left} = 1, \text{right} = 2) = ?$$



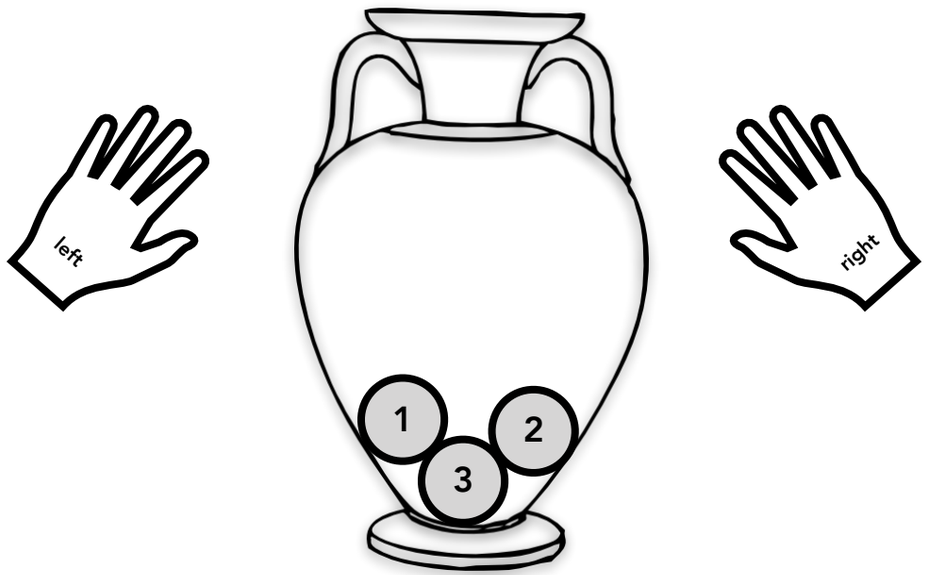


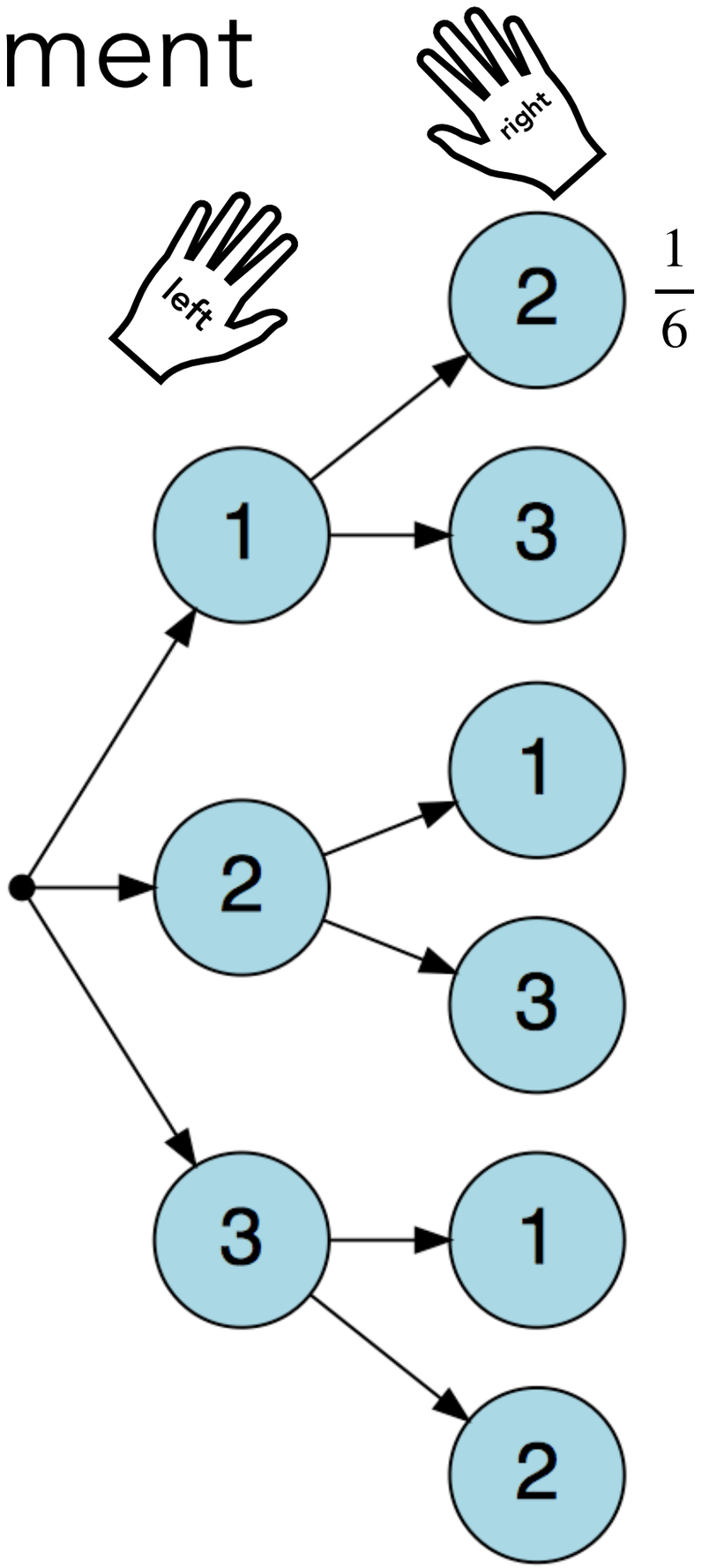
$\frac{1}{9}$

```
1 library("arrangements")
2 balls = 1:3 # number of balls in urn
3 ndraws = 2 # number of draws
4
5 # order matters, with replacement
6 permutations(balls, ndraws, replace = T)
```

Sampling **without** replacement

$p(\text{left} = 1, \text{right} = 2) = ?$





$\frac{1}{6}$

```
1 library("arrangements")
2 balls = 1:3 # number of balls in urn
3 ndraws = 2 # number of draws
4
5 # order matters, without replacement
6 permutations(balls, ndraws)
```

Naive definition of probability

$$P_{\text{naive}}(A) = \frac{\text{number of outcomes favorable to A}}{\text{number of outcomes}}$$

if all outcomes are equally likely!

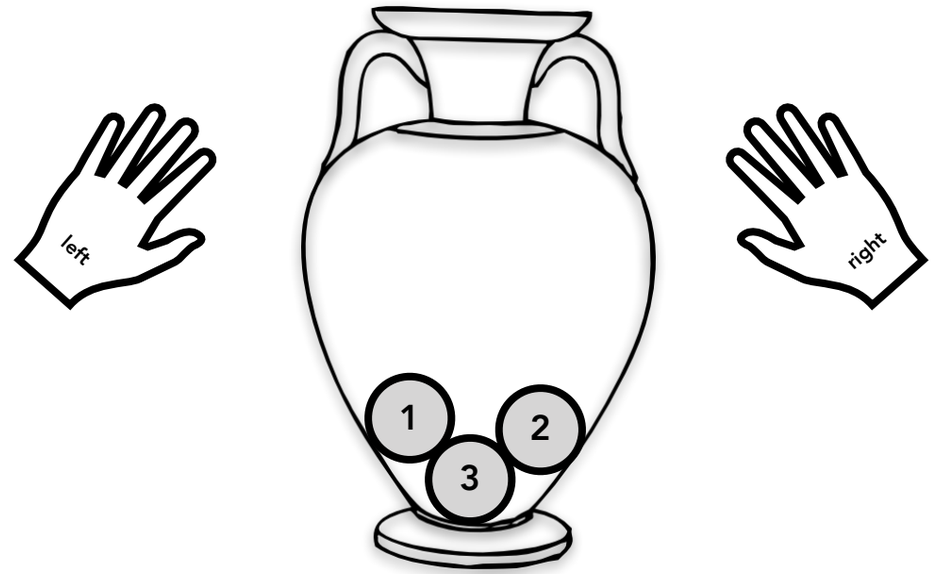
Definitions

Experiment: Activity that produces or observes an outcome.

Drawing 2 marbles from the urn with replacement, and noting the order.

Sample Space: Set of possible outcomes for an experiment.

$\Omega = \{(1, 1), (1, 2), \ldots, (2,1), \ldots, (3, 3)\}$



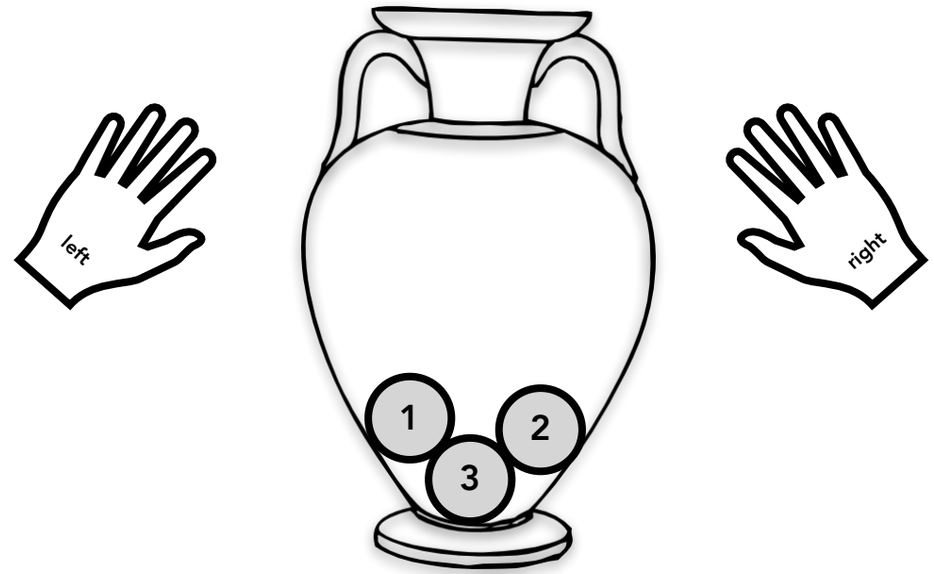
Event: Subset of the sample space. $(1, 1)$

Definitions

If $P(X_i)$ is the probability of event X_i

1. Probability cannot be negative.

$$P(X_i) \geq 0$$



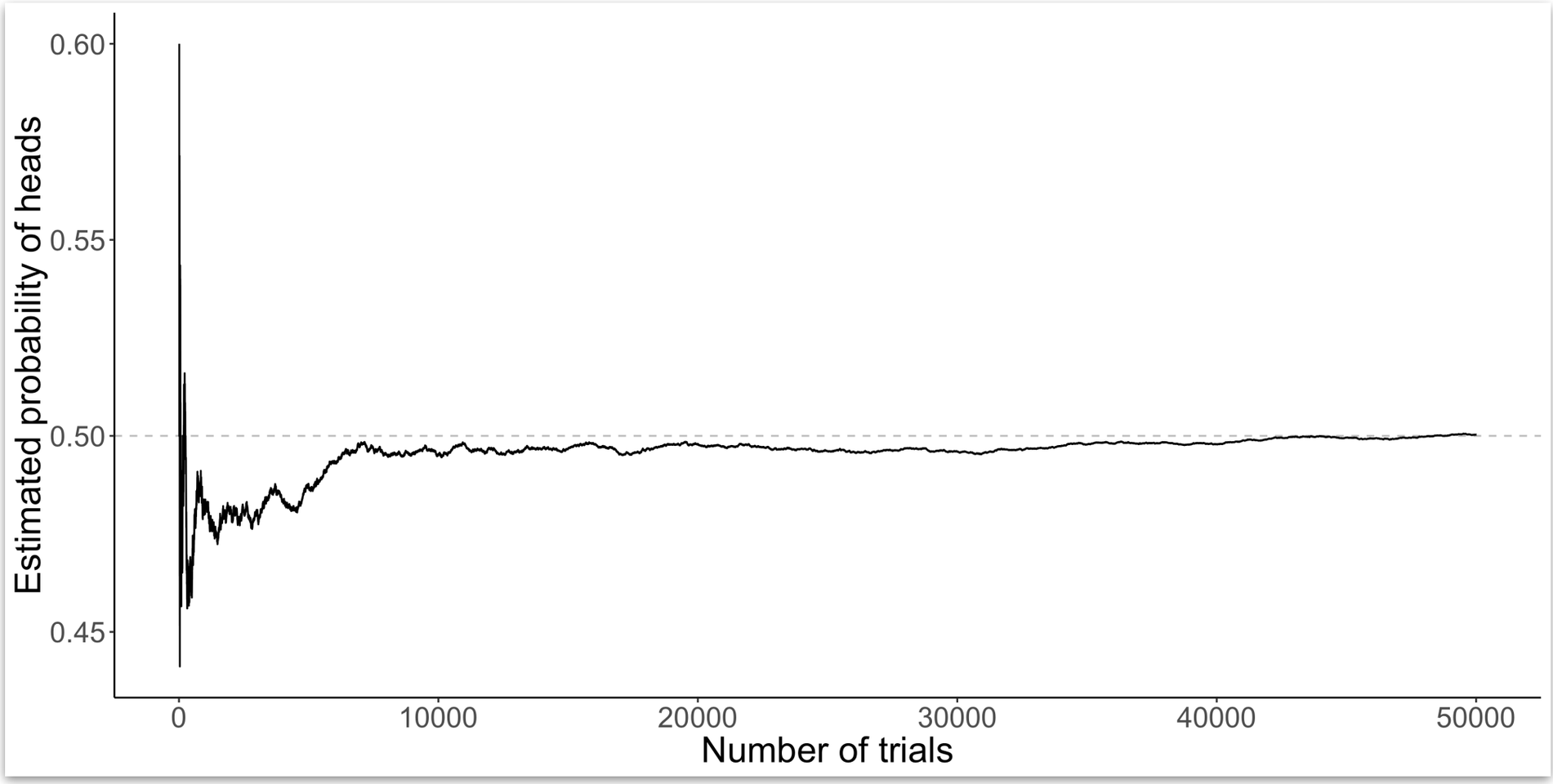
2. Total probability of all outcomes in the sample space is 1.

$$\sum_{i=1}^{N} P(X_i) = P(X_1) + P(X_2) + \ldots + P(X_N) = 1$$

Interpretations of probability

Frequentist interpretation

Probabilities = **long-range frequencies**



law of large numbers = empirical probability will approximate the true probability as the sample size increases

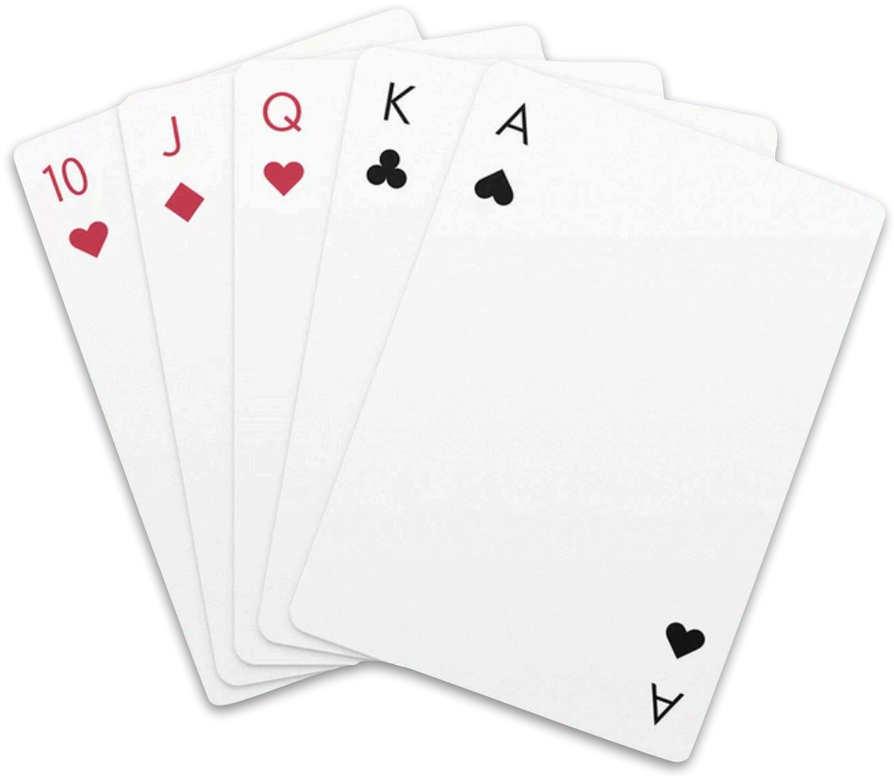
Subjective interpretation

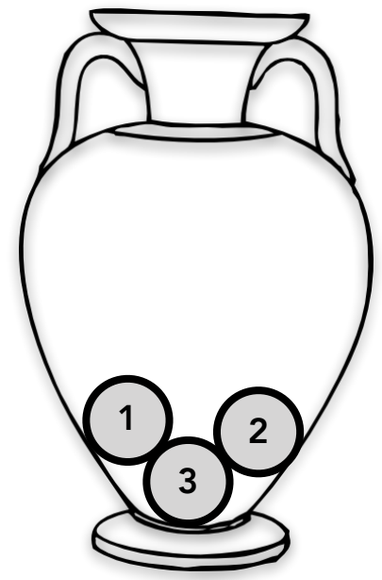
Probabilities = **subjective degrees of belief**

- applies to events which may only happen once
- **"What's the probability that humans will land on Mars someday?"**
- probabilities are not a property of the world, but of a person's beliefs about the world
- at the heart of Bayesian data analysis

Classical interpretation

Probabilities = **computed based on our knowledge of the situation**





developed by analyzing games of chance

Clue guide to probability

Clue guide to probability

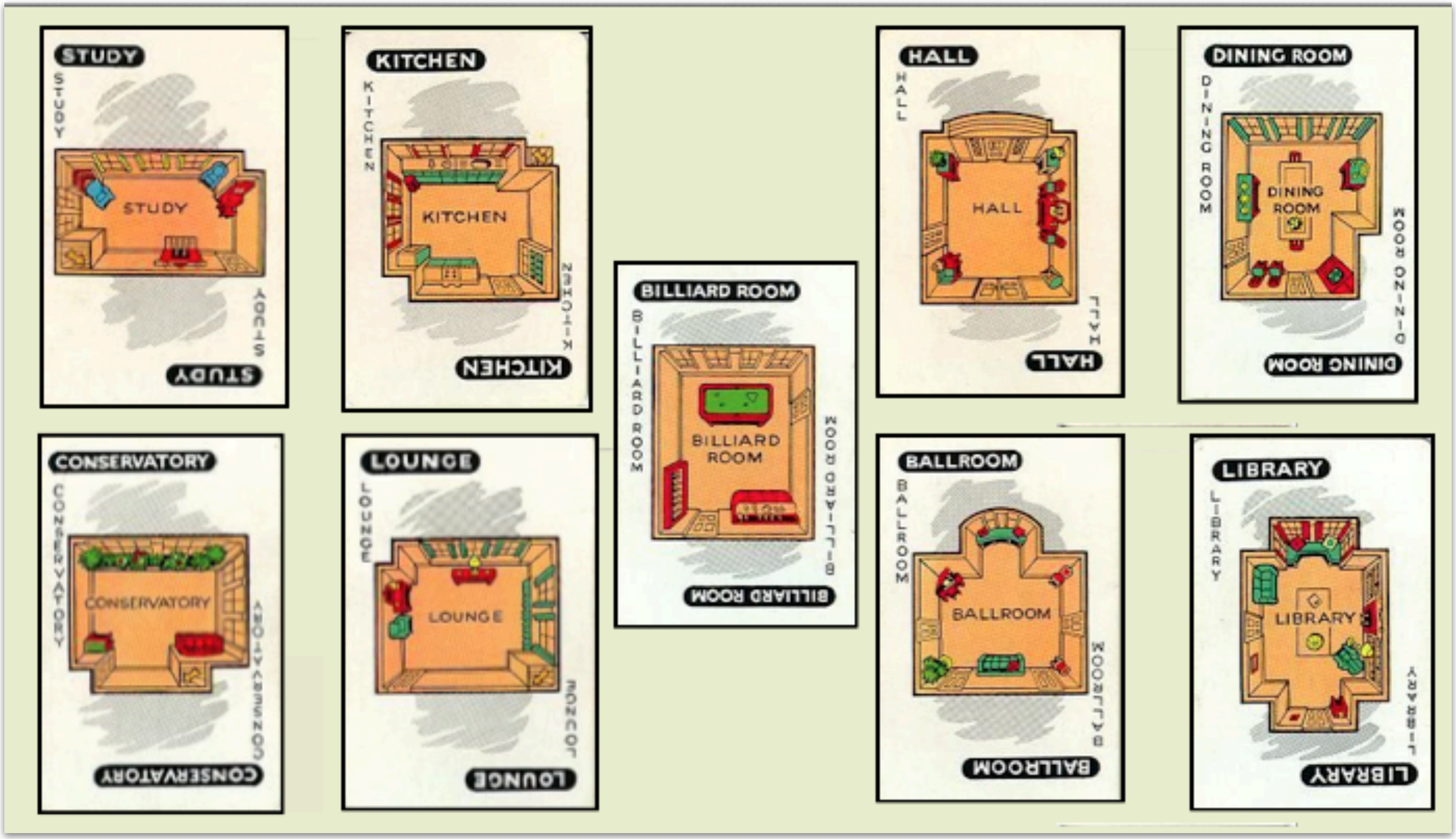
Who killed Mr Boddy?

Clue guide to probability

Who killed Mr Boddy, **with what?**

Clue guide to probability

Who killed Mr Boddy, with what, **and where?**



Clue guide to probability

Clue guide to probability

```
who = c("ms_scarlet", "col_mustard", "mrs_white",
        "mr_green", "mrs_peacock", "prof_plum")
what = c("candlestick", "knife", "lead_pipe",
         "revolver", "rope", "wrench")
where = c("study", "kitchen", "conservatory",
          "lounge", "billiard_room", "hall",
          "dining_room", "ballroom", "library")

df.clue = expand.grid(who = who,
                      what = what,
                      where = where) %>%
  as_tibble()
```

Ω

who	what	where
ms_scarlet	candlestick	study
ms_scarlet	candlestick	kitchen
ms_scarlet	candlestick	conservatory
ms_scarlet	candlestick	lounge

`nrow(df.clue)` = 324

Clue guide to probability

Who?

- 6 suspects
- mutually exclusive and exhaustive
- $p(\text{Murderer} = \text{one of the six}) = 1$

- each equally likely a priori
- $p(\text{who} = \text{Prof. Plum}) = \frac{1}{6}$

for mutually exclusive events

otherwise

- $p(A \cup B) = p(A) + p(B)$

$$p(A \cup B) = p(A) + p(B) - p(A, B)$$

- $p(\text{who} = \text{Prof. Plum} \cup \text{Mrs. White}) = \frac{2}{6}$

Clue guide to probability

- *conditional probability*:
- $p(A \mid B)$ (probability of A given B)
- **Definition**: $p(A \mid B) = \dfrac{p(A, B)}{p(B)}$

- $p(\text{Prof. Plum} \mid \text{male}) = \dfrac{1/6}{1/2} = 1/3$

Who?

who	gender
col_mustard	male
mr_green	male
prof_plum	male
ms_scarlet	female
mrs_white	female
mrs_peacock	female

```
1 df.suspects = df.clue %>%
2   distinct(who) %>%
3   mutate(gender = ifelse(
4     test = who %in% c("ms_scarlet",
                        "mrs_white",
                        "mrs_peacock"),
5     yes = "female",
6     no = "male")
7   )
```

Clue guide to probability

- *conditional probability*:
- $p(A \mid B)$ (probability of A given B)
- **Definition**: $p(A \mid B) = \dfrac{p(A, B)}{p(B)}$

- $p(\text{Prof. Plum} \mid \text{male}) = \dfrac{1/6}{1/2} = 1/3$

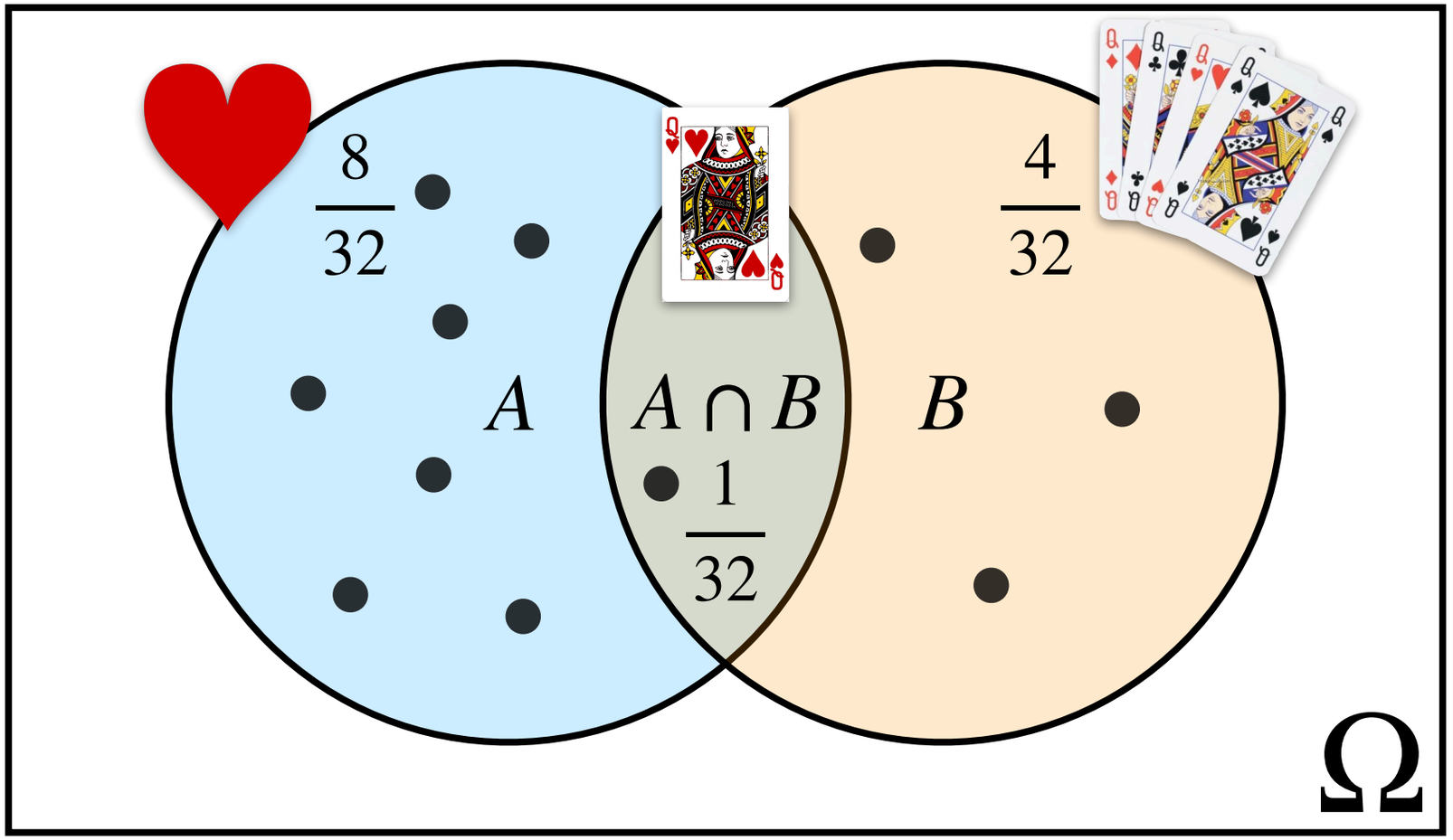
Who?

who	gender
col_mustard	male
mr_green	male
prof_plum	male
ms_scarlet	female
mrs_white	female
mrs_peacock	female

```
1 df.suspects %>%
2   summarize(p_prof_plum_given_male =
3                 sum(gender == "male" &
                        who == "prof_plum") /
4                 sum(gender == "male"))
```

use naive definition of probability

Clue guide to probability

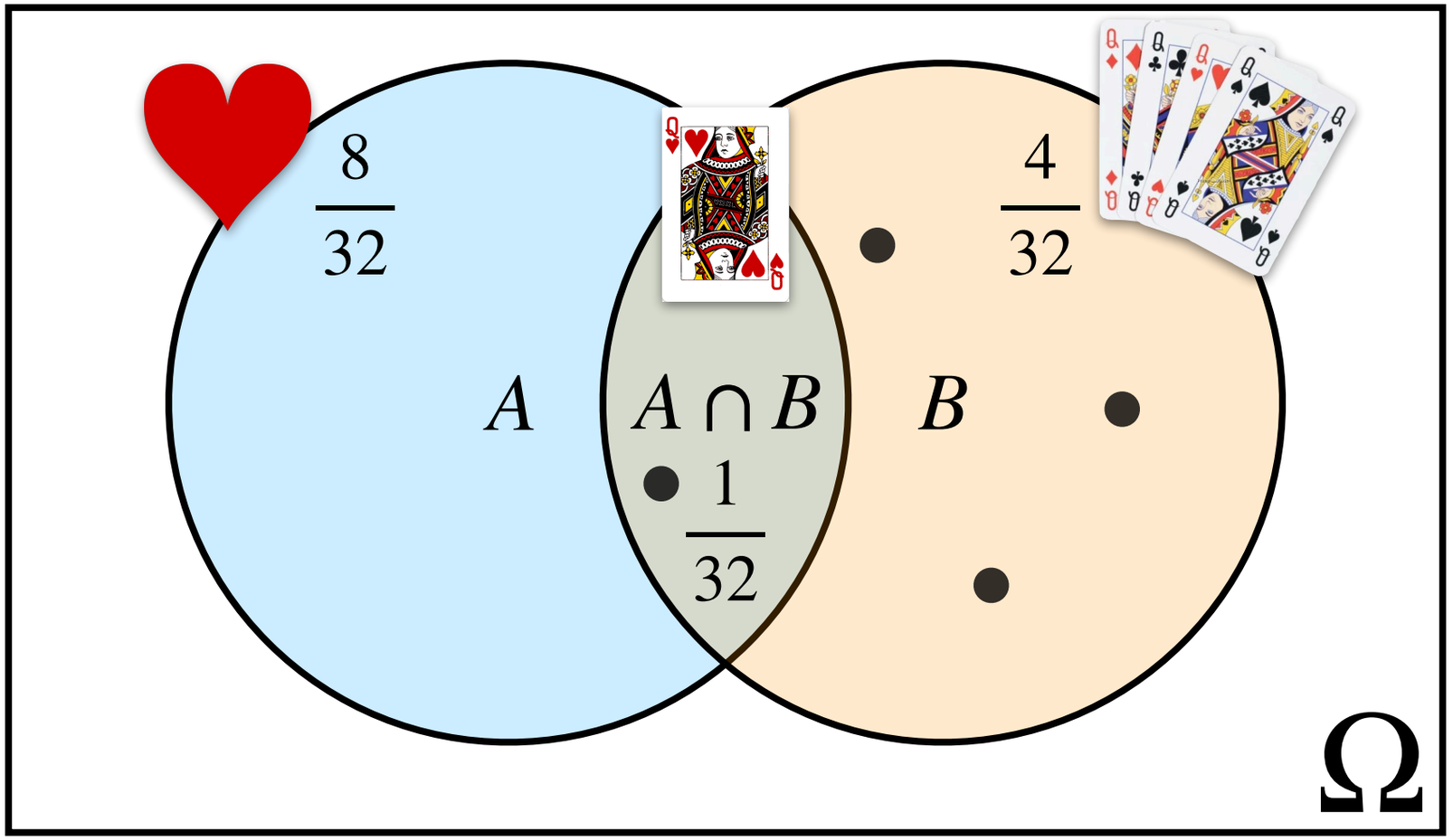


Probability of drawing a hearts, given that it's a queen?

Definition: $p(A \mid B) = \dfrac{p(A, B)}{p(B)}$

$$p(A) = \frac{8}{32} \qquad p(A, B) = \frac{1}{32} \qquad p(B) = \frac{4}{32}$$

Clue guide to probability

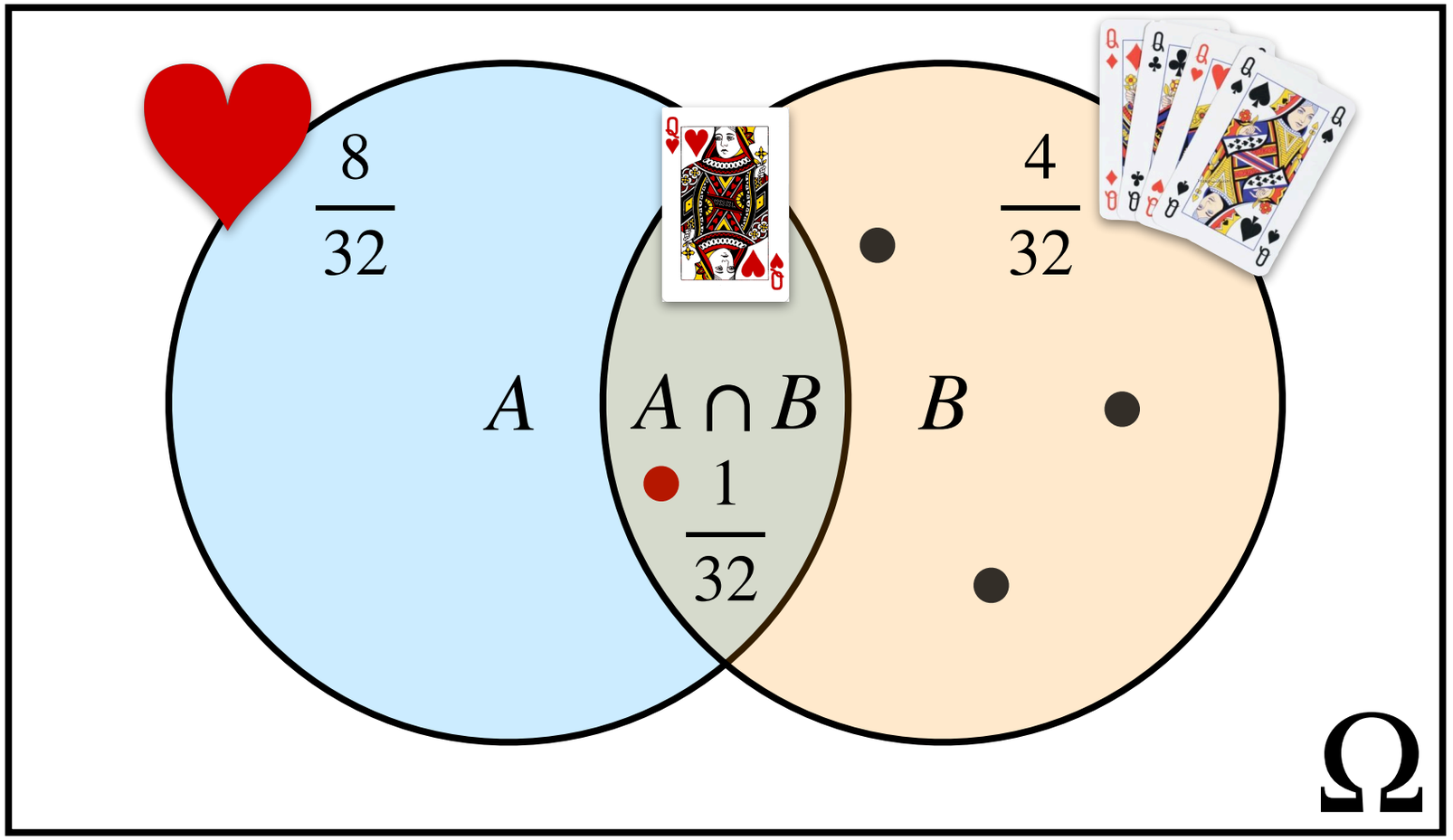


Probability of drawing a hearts, given that it's a queen?

$$p(A \mid B) = \frac{p(A, B)}{p(B)}$$

$$p(A) = \frac{8}{32} \qquad p(A, B) = \frac{1}{32} \qquad p(B) = \frac{4}{32}$$

Clue guide to probability

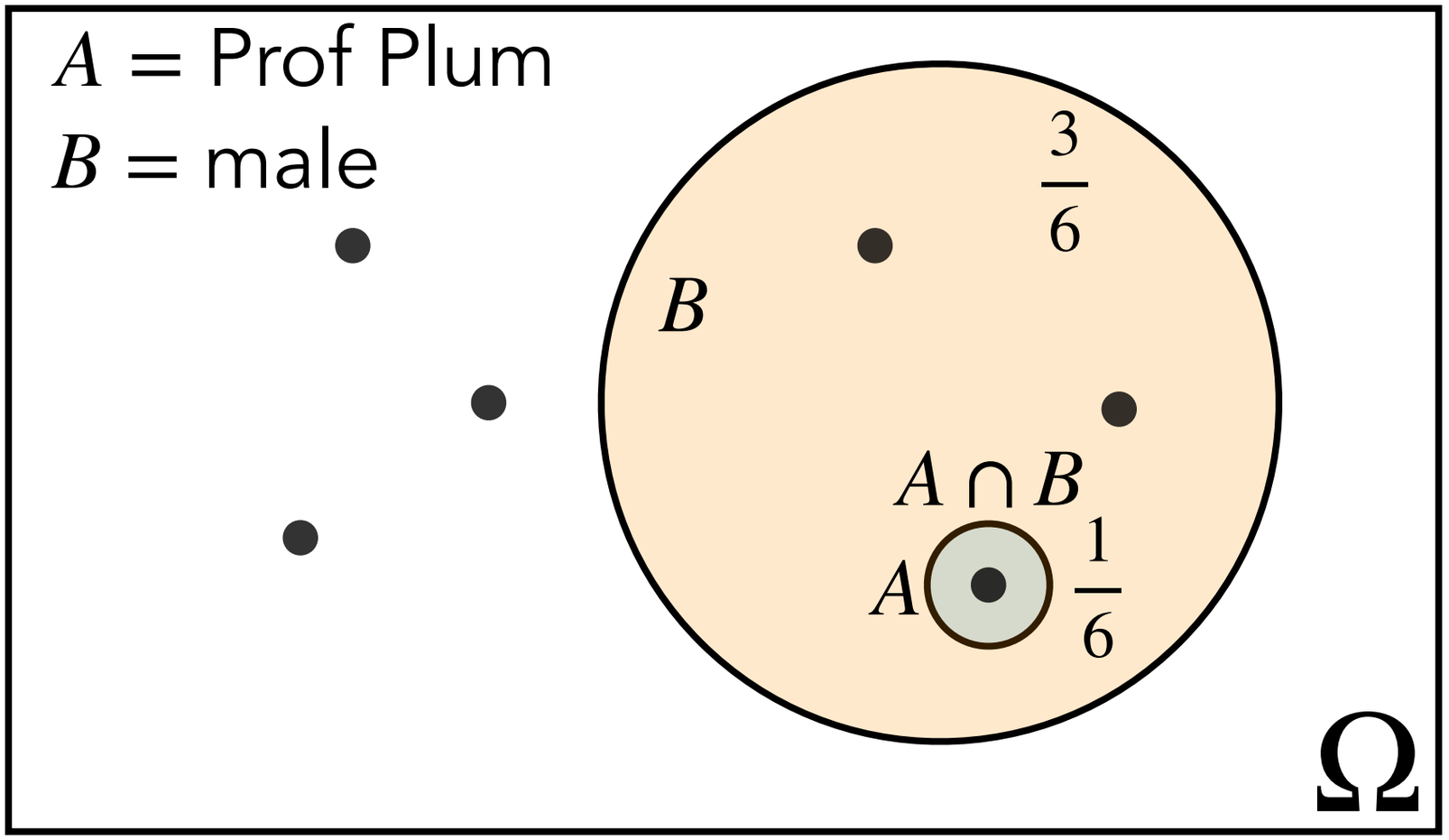


Probability of drawing a hearts, given that it's a queen?

$$p(A \mid B) = \frac{p(A, B)}{p(B)} = \frac{1/32}{4/32} = \frac{1}{4}$$

$$p(A) = \frac{8}{32} \qquad p(A, B) = \frac{1}{32} \qquad p(B) = \frac{4}{32}$$

Clue guide to probability



Probability that it was Prof Plum, given that the murderer was male?

Definition: $p(A \mid B) = \dfrac{p(A, B)}{p(B)} = \dfrac{1}{3}$

$$p(A) = \frac{1}{6} \qquad p(A, B) = \frac{1}{6} \qquad p(B) = \frac{3}{6}$$

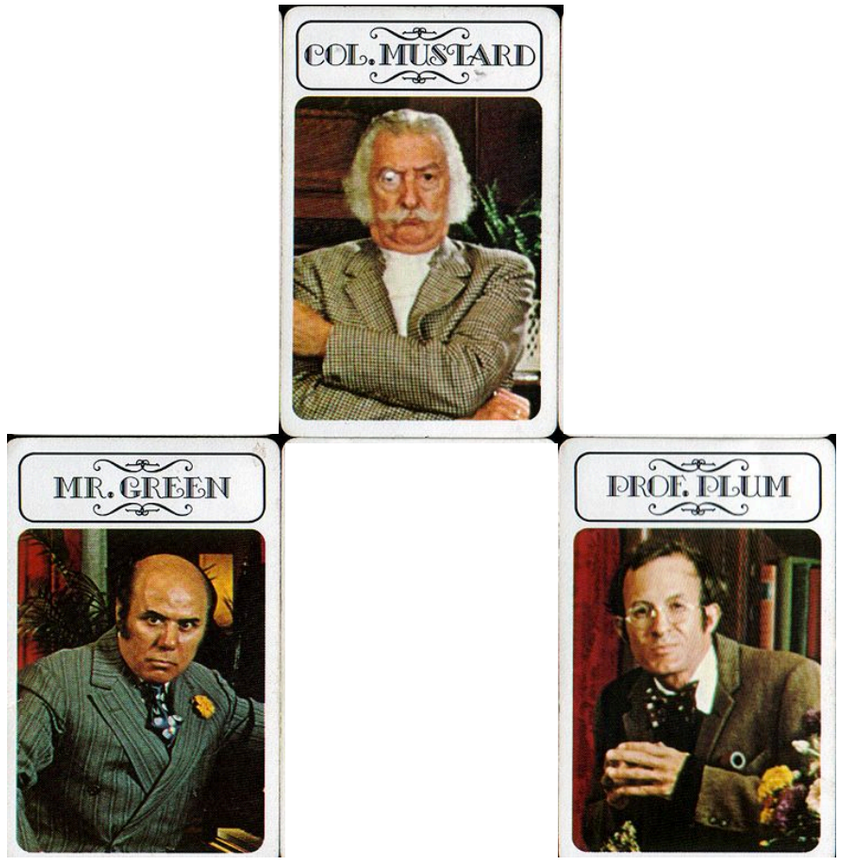
Clue guide to probability

- *conditional probability*:
- $p(A \mid B)$ (probability of A given B)
- **Definition**: $p(A \mid B) = \dfrac{p(A, B)}{p(B)}$

$$1/6$$

- $p(\text{Prof. Plum} \mid \text{male}) = \dfrac{1/6}{1/2} = 1/3$

Who?



who	gender
col_mustard	male
mr_green	male
prof_plum	male

```
1 df.suspects %>%
2   filter(gender == "male") %>%
3   summarize(p_prof_plum_given_male =
4                 sum(who == "prof_plum") /
5                 n())
```

Clue guide to probability

- *independence*:
- A and B are independent if
- **Definition**: $p(A \mid B) = p(A)$
- (probability of A does not change if you know B)

- $p(\text{Prof Plum} \mid \text{candle stick}) = p(\text{Prof Plum})$
- each card (who and what) is drawn from a separate pack of cards

Who?

What?

Clue guide to probability

- *joint probability*:
- if A and B are independent then
- **Definition**: $p(A, B) = p(A) \cdot p(B)$

- $p(\text{Prof Plum, candle stick}) =$
  $p(\text{Prof Plum}) \cdot p(\text{candle stick}) =$
  $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$

Who?

What?

Clue guide to probability

- *dependence*:
- **Definition**: $p(A \mid B) \neq p(A)$
- **Definition**: $p(A, B) = p(A) \cdot p(B \mid A)$

Who?

What?

- if women were more likely than men to use the revolver then
- p(Mrs. White | Revolver) > p(Mrs. White)

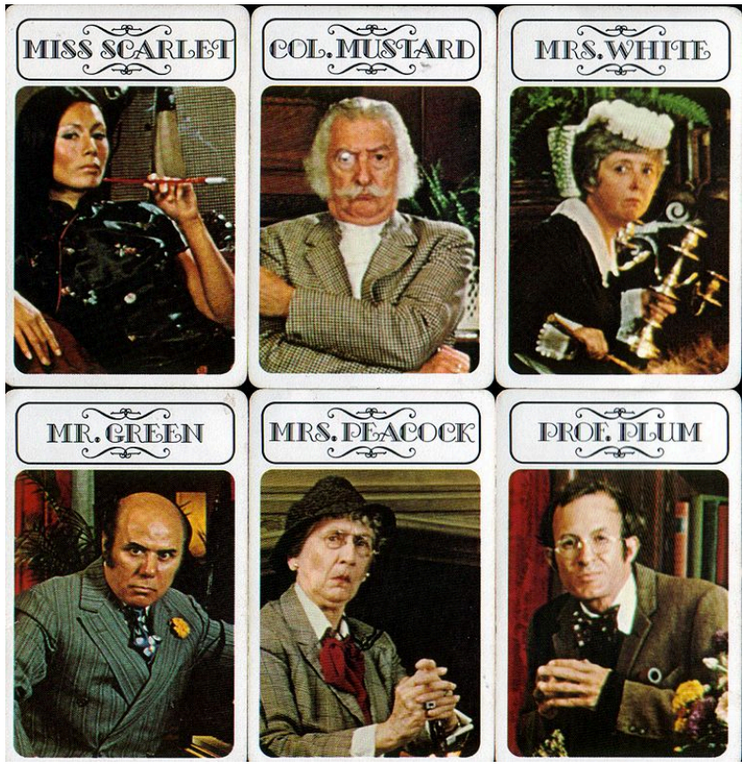
Clue guide to probability

- *law of total probability*
- **Definition**:

$$p(A) = p(A \mid B) \cdot p(B) + p(A \mid \neg B) \cdot p(\neg B)$$

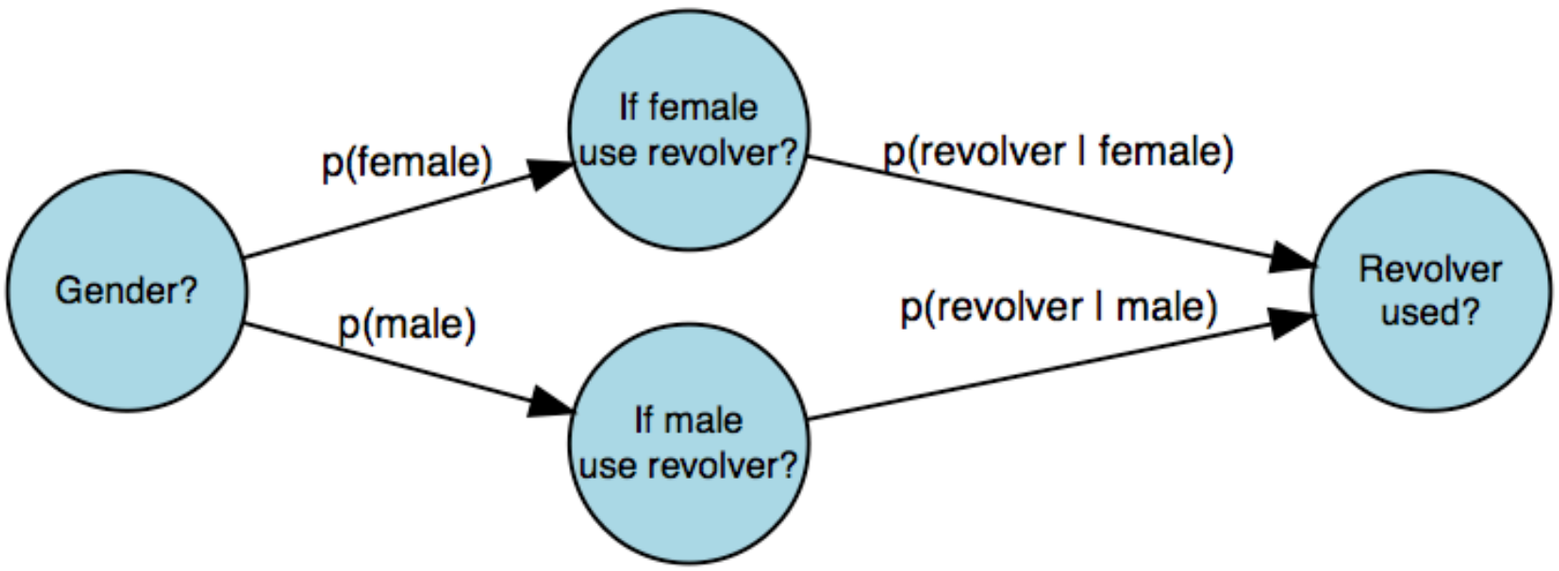
$$p(A) = \sum_{i=1}^{n} p(A \mid B_i) \cdot p(B_i)$$

Who?



What?

$p(\text{what} = \text{Revolver}) = ?$



Clue guide to probability

- *Bayes' rule (derivation)*

$$p(B \mid A) = \frac{p(A, B)}{p(A)} \qquad p(A \mid B) = \frac{p(A, B)}{p(B)}$$

$$p(A, B) = p(B \mid A) \cdot p(A) = p(A \mid B) \cdot p(B)$$

$$p(B \mid A) = \frac{p(A \mid B) \cdot p(B)}{p(A)}$$

Clue guide to probability

$$p(B \mid A) = \frac{p(A \mid B) \cdot p(B)}{p(A)}$$

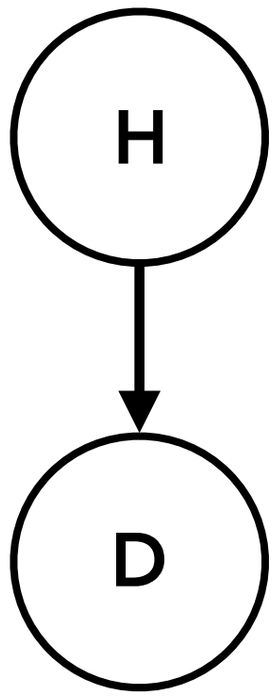
posterior **likelihood** **prior**

$$p(H \mid D) = \frac{p(D \mid H) \cdot p(H)}{p(D)}$$

subjective probability interpretation

H = Hypothesis

D = Data



formal framework for learning from data

updating our prior belief $p(H)$, to a posterior belief $p(H \mid D)$ given some data

Clue guide to probability

$$p(H \mid D) = \frac{p(D \mid H) \cdot p(H)}{p(D)}$$

posterior — likelihood — prior

H = Hypothesis

D = Data

probability of the data?!

law of total probability:

$$p(D) = \sum_{i=1}^{n} p(D \mid H_i) \cdot p(H_i)$$

Clue guide to probability

A patient named Fred is tested for a disease called *conditionitis*, a medical condition that afflicts **1% of the population**. The test result is positive, i.e., the test claims that Fred has the disease. Let **D** be the event that Fred has the disease and **T** be the event that he tests positive.

Suppose that the test is "95% accurate"; there are different measures of the accuracy of a test, but in this problem it is assumed to mean that **P(T|D) = 0.95** and **P(¬T|¬D) = 0.95**. The quantity P(T|D) is known as the *sensitivity* (= true positive rate) of the test, and P(¬T|¬D) is known as the *specificity* (= true negative rate).

Find the conditional probability that Fred has *conditionitis*, given the evidence provided by the positive test result.

Clue guide to probability

What's the probability that Fred has conditionitis?

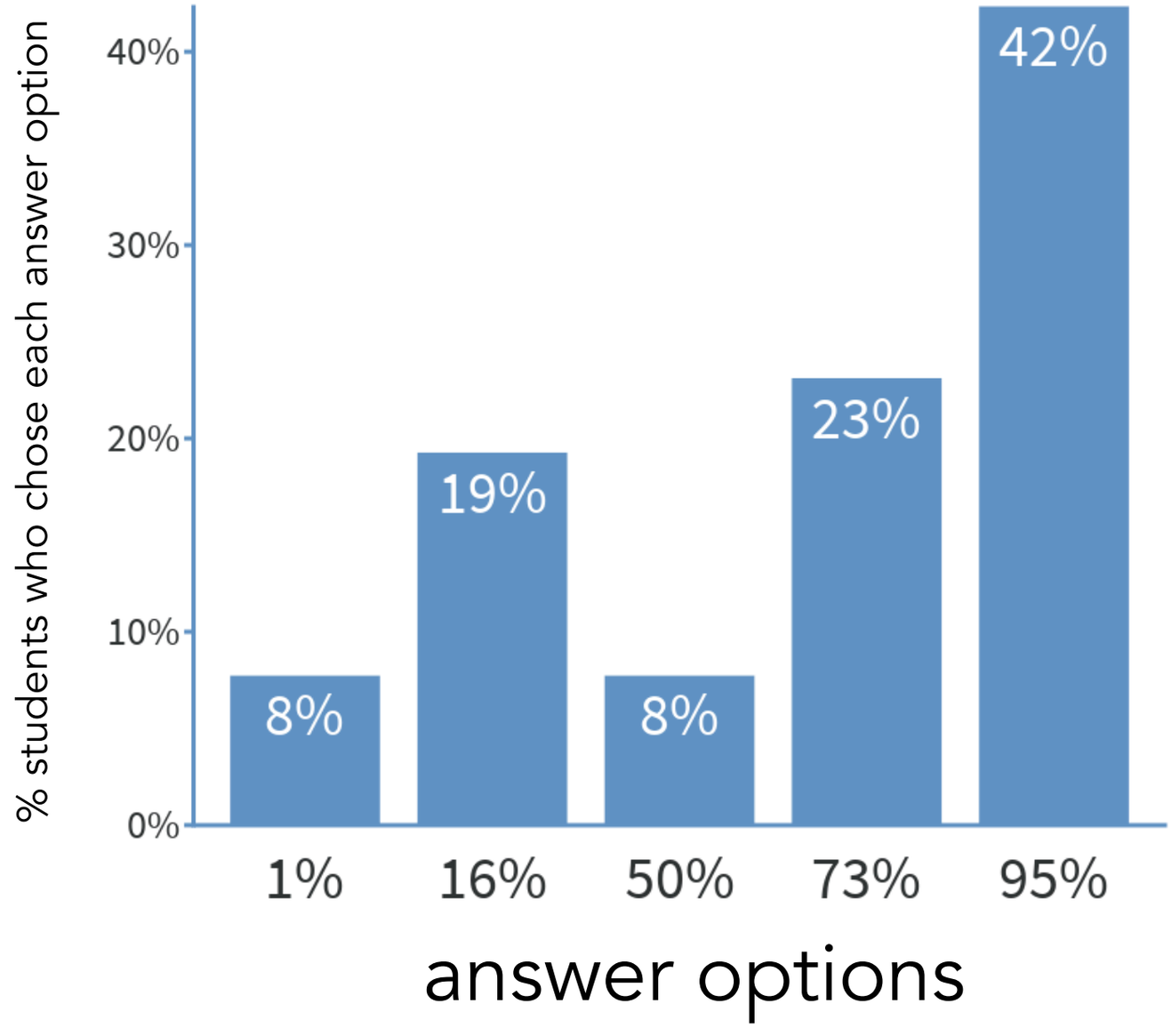
When poll is active, respond at **PollEv.com/psych252**

Answers to this poll are anonymous

A patient named Fred is tested for a disease called *conditionitis*, a medical condition that afflicts **1% of the population**. The test result is positive, i.e., the test claims that Fred has the disease. Let **D** be the event that Fred has the disease and **T** be the event that he tests positive.

Suppose that the test is "95% accurate"; there are different measures of the accuracy of a test, but in this problem it is assumed to mean that **P(T|D) = 0.95** and **P(¬T|¬D) = 0.95**. The quantity P(T|D) is known as the *sensitivity* (= true positive rate) of the test, and P(¬T|¬D) is known as the *specificity* (= true negative rate).

Find the conditional probability that Fred has *conditionitis*, given the evidence provided by the positive test result.



Clue guide to probability

what we know

$$P(D) = 0.01$$

$$P(T \mid D) = 0.95$$

$$P(T \mid \neg D) = 0.05$$

what we want to know

$$P(D \mid T) = ?$$

$$p(D \mid T) = \frac{p(T \mid D) \cdot p(D)}{p(T)}$$ **Bayes' rule**

$$p(D \mid T) = \frac{p(T \mid D) \cdot p(D)}{p(T \mid D) \cdot p(D) + p(T \mid \neg D) \cdot p(\neg D)}$$ ← **law of total probability**

$$p(D \mid T) = \frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.05 \cdot 0.99} \approx 0.16$$

Clue guide to probability

what we know

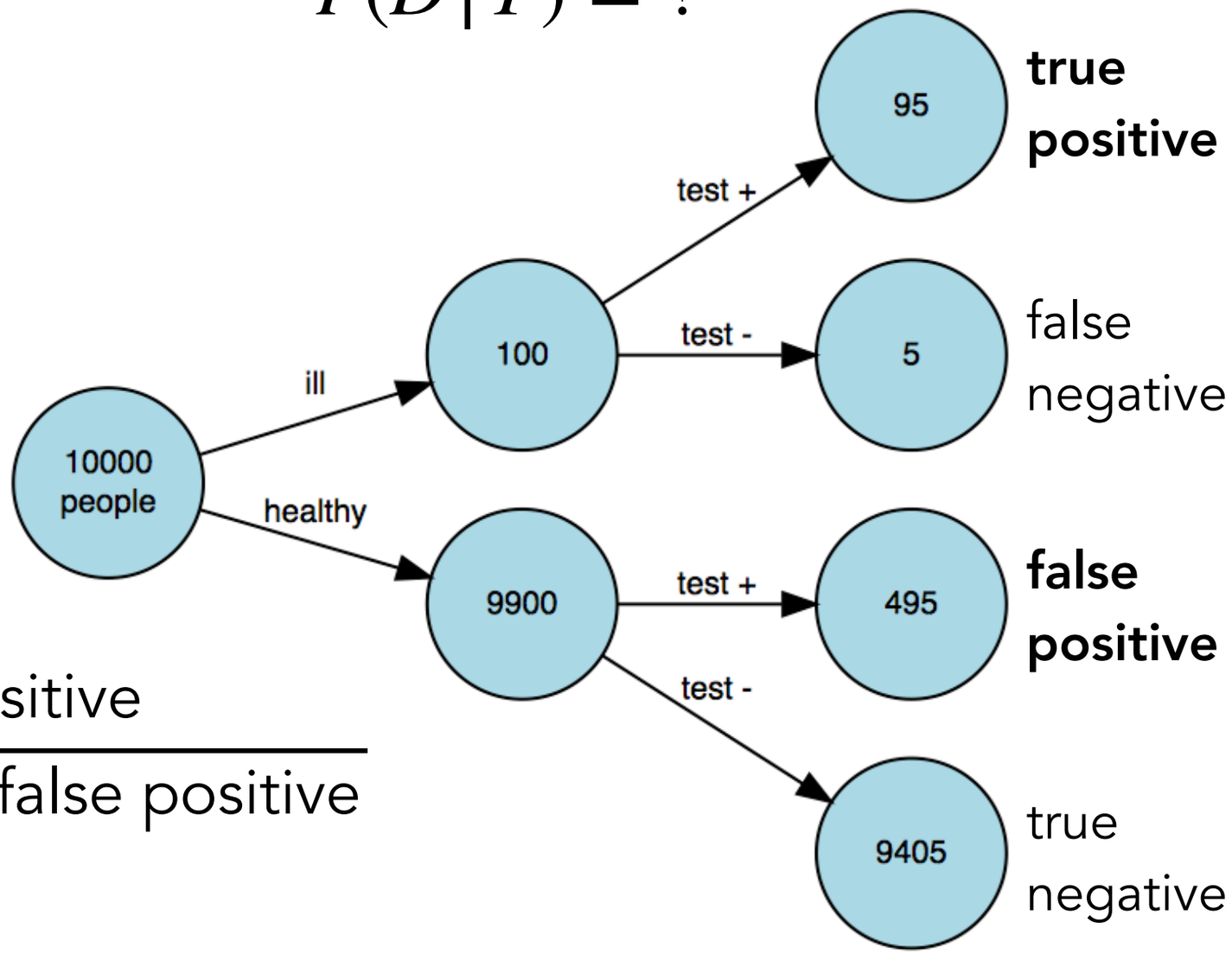
$$P(D) = 0.01$$

$$P(T \mid D) = 0.95$$

$$P(T \mid \neg D) = 0.05$$

what we want to know

$$P(D \mid T) = ?$$



$$P(D \mid T) = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

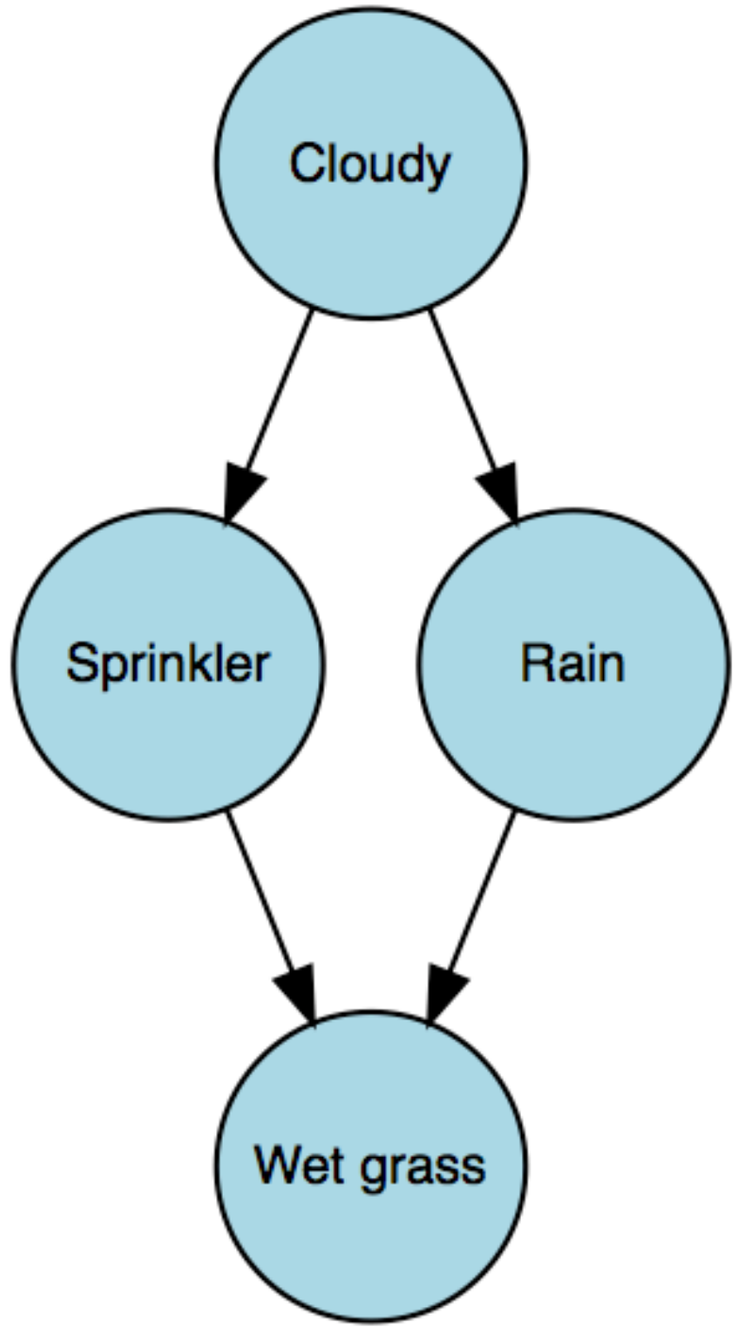
$$= \frac{95}{95 + 495}$$

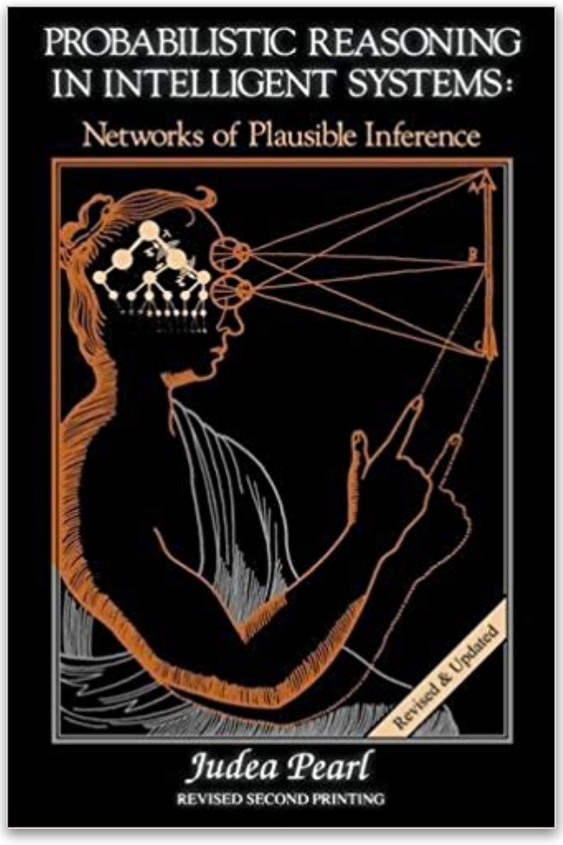
$$\approx 0.16$$

Bayesian Networks

Representation

- **nodes** represent variables of interest
- **links** represent direct dependencies between variables
- **conditional probability tables** parameterize the model



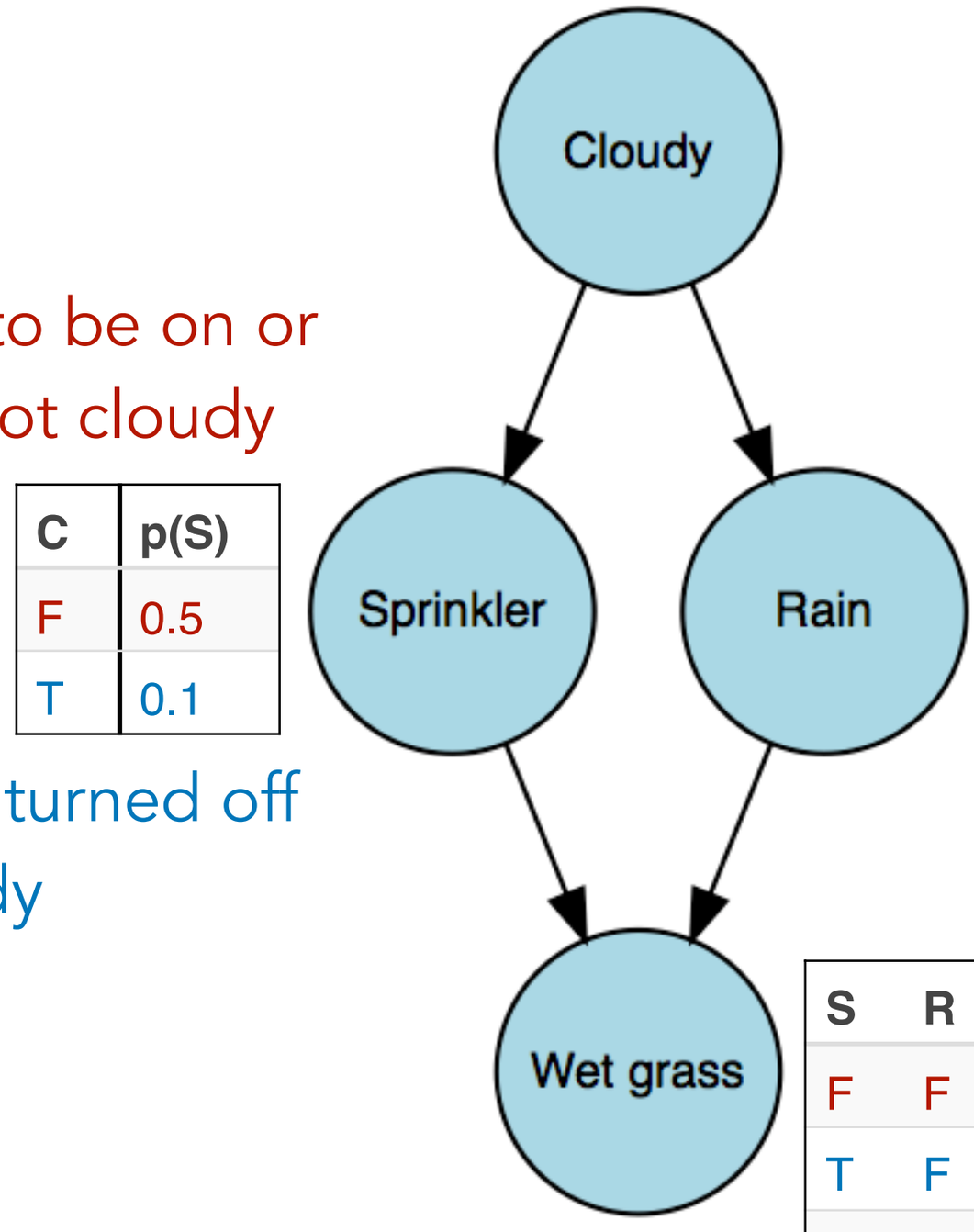


Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of plausible inference. San Francisco, CA: Morgan Kaufmann.

Representation

p(C)
0.5

equally likely to be cloudy or not



equally likely to be on or off when it's not cloudy

C	p(S)
F	0.5
T	0.1

normally gets turned off when its cloudy

it doesn't rain when it's not cloudy

C	p(R)
F	0
T	0.3

it sometimes rains when it's cloudy

S	R	p(W)
F	F	0.1
T	F	0.90
F	T	0.90
T	T	0.99

wet grass unlikely when neither

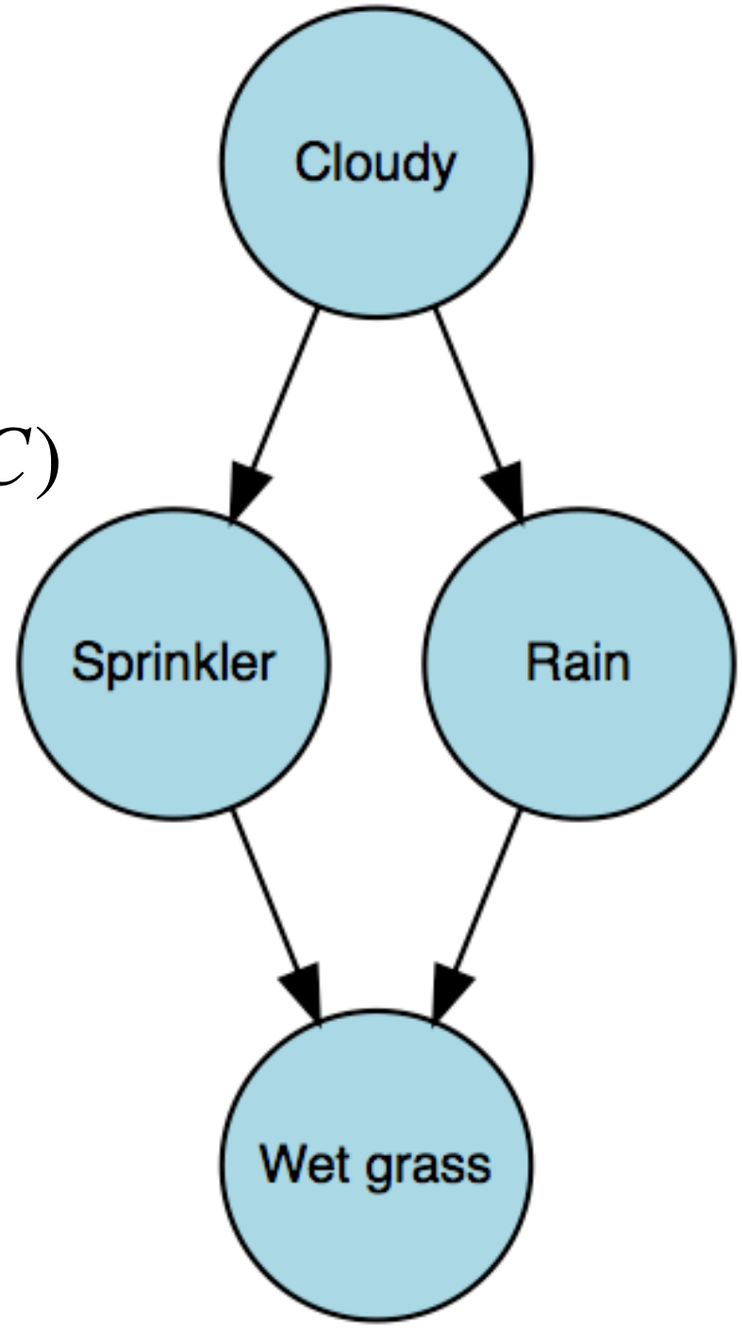
most likely wet when either

Compact representation of the joint probability distribution

chain rule

$$
\begin{aligned}
p(C,S,R,W) &= p(W \mid CSR) \cdot p(CSR) \\
&= p(W \mid CSR) \cdot p(R \mid CS) \cdot p(CS) \\
&\vdots \\
&= p(W \mid CSR) \cdot p(R \mid CS) \cdot p(S \mid C) \cdot p(C)
\end{aligned}
$$

number of parameters $8 + 4 + 2 + 1 = 15$



considering independence

$$p(C,S,R,W) = p(W \mid CR) \cdot p(S \mid C) \cdot p(R \mid C) \cdot p(C)$$

number of parameters $4 + 2 + 2 + 1 = 9$

Inference by conditioning

- supports *predictive* reasoning:

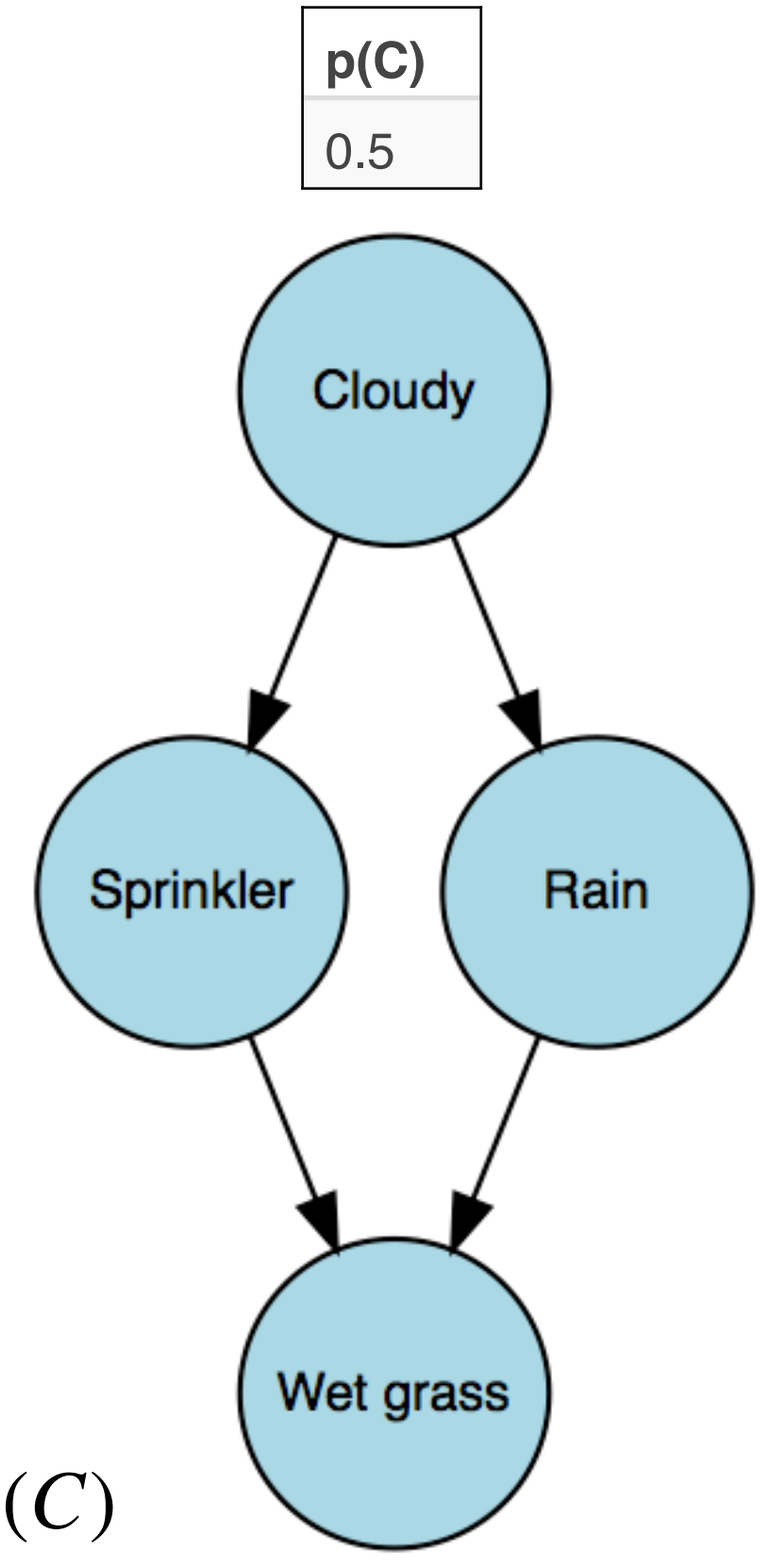
$p(W \mid C) = ?$ **probability of wet grass given that it's cloudy?**

$$p(W \mid C) = \frac{p(W, C)}{p(C)}$$

$$p(C) = 0.5$$

$$p(W, C) = \sum_{s=0}^{1} \sum_{r=0}^{1} p(C, S, R, W)$$

$$= \sum_{s=0}^{1} \sum_{r=0}^{1} p(W \mid CR) \cdot p(S \mid C) \cdot p(R \mid C) \cdot p(C)$$



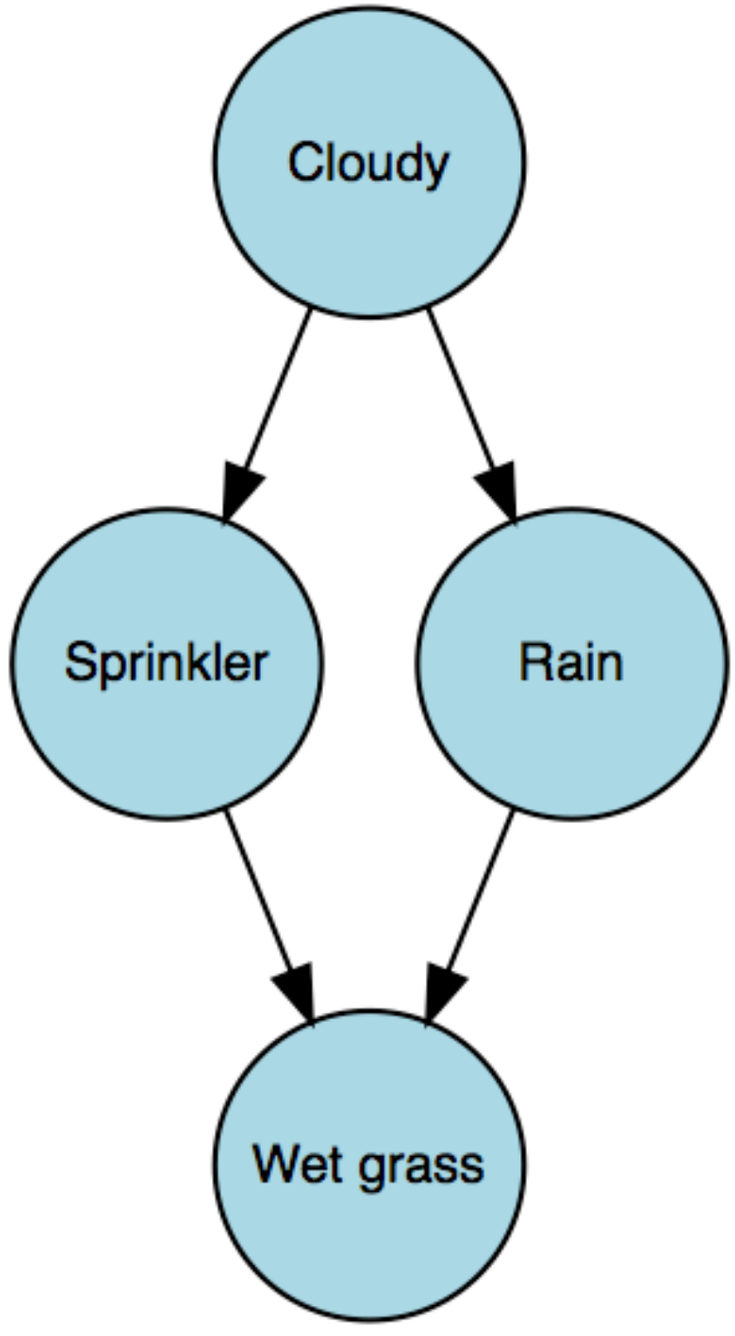
Inference by conditioning

- supports *predictive* reasoning:

$p(W \mid C) = ?$ probability of wet grass given that it's cloudy?

- and *diagnostic* reasoning:

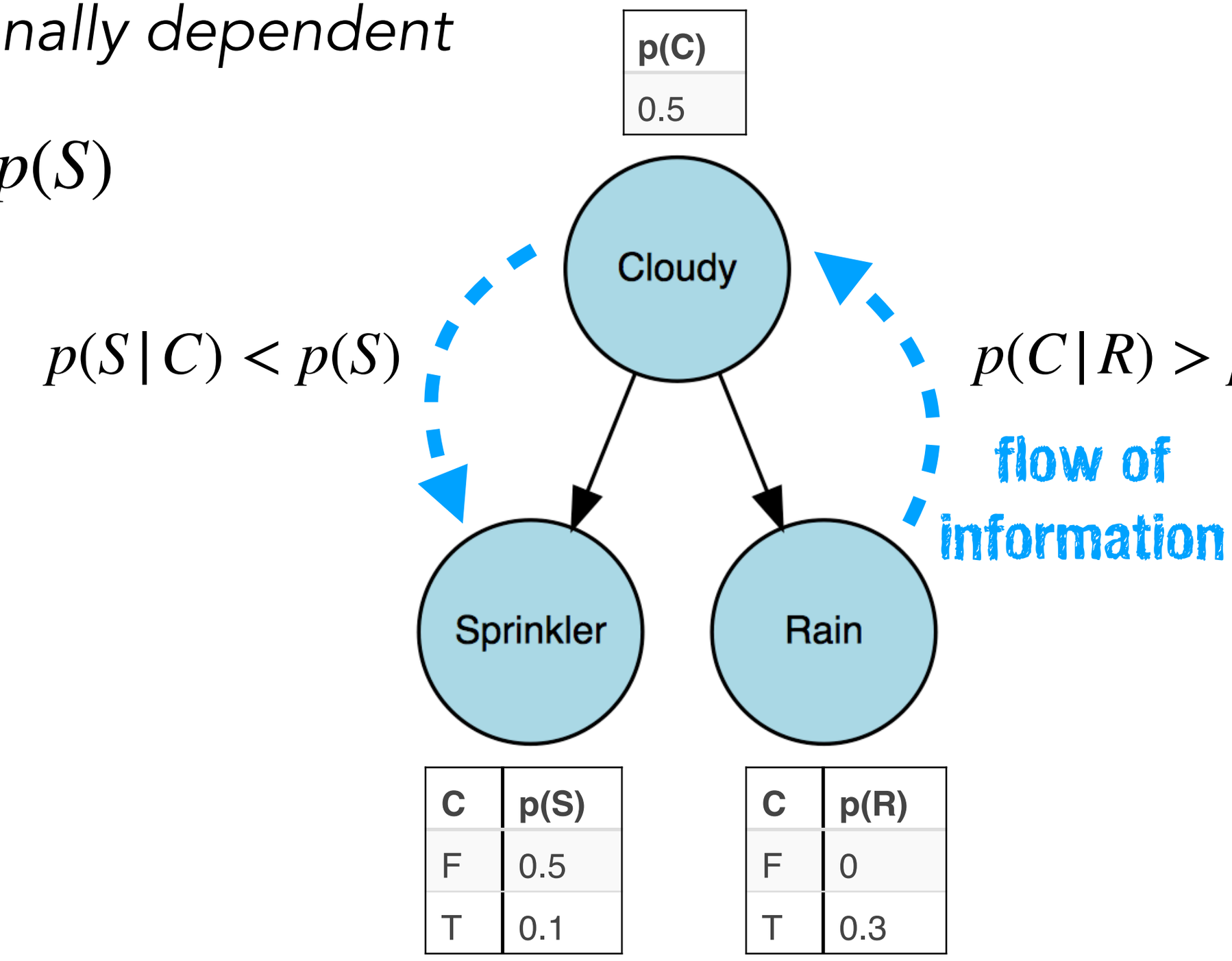
$p(R \mid W) = ?$ probability that it rained given that the grass is wet?



Patterns of inference: **Common cause**

- effects of a common cause are *unconditionally dependent*

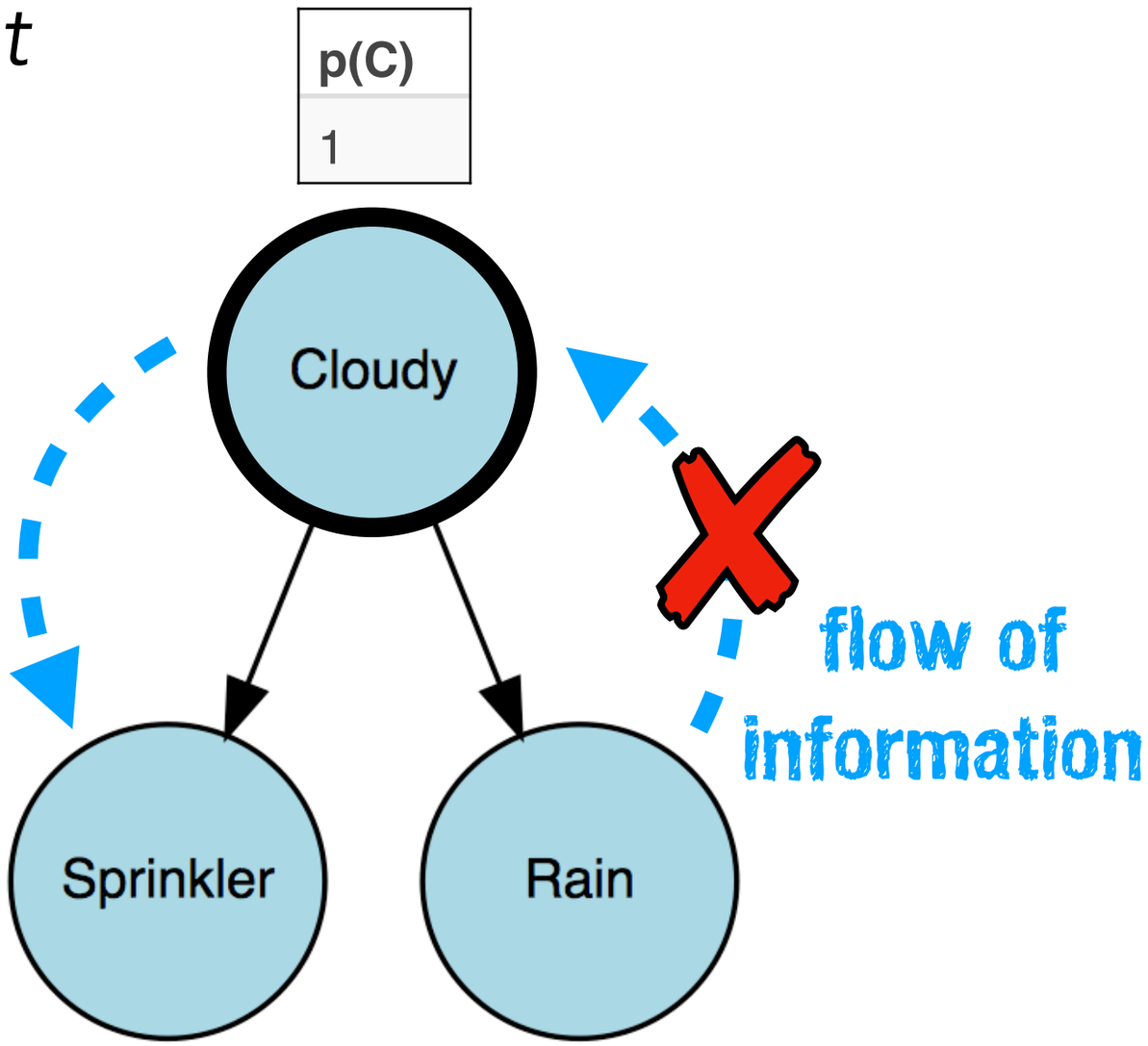
$$p(S \mid R) \neq p(S)$$



Patterns of inference: **Common cause**

- effects of a common cause are *conditionally independent given the cause*

$$p(S \mid R, C) = p(S \mid C)$$



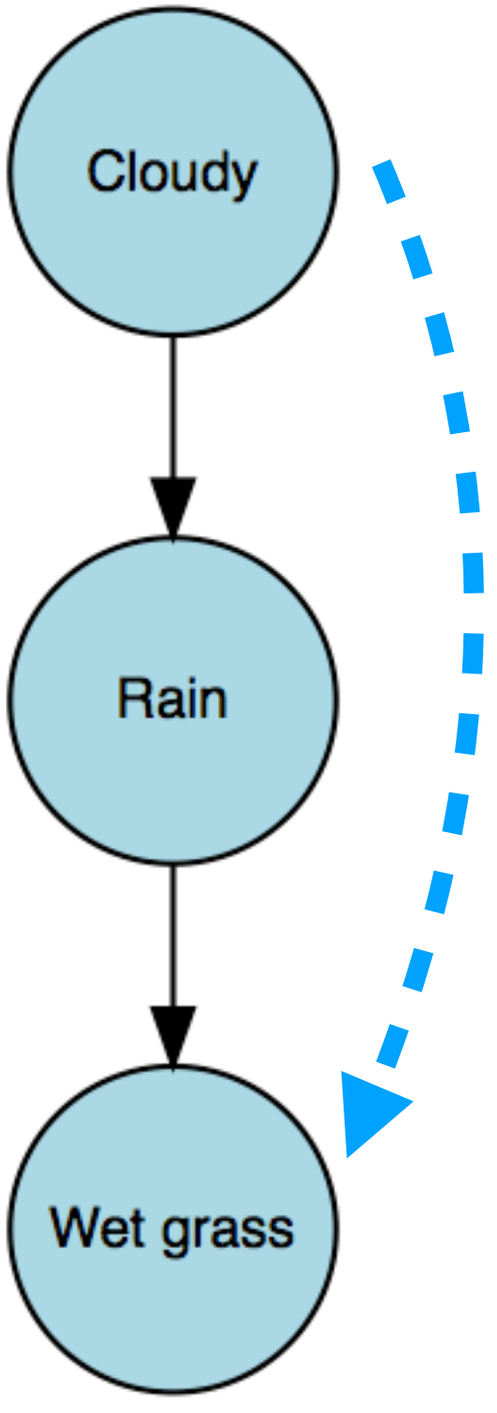
C	p(S)
F	0.5
T	0.1

C	p(R)
F	0
T	0.3

Patterns of inference: **Causal chain**

- cause and effect in a causal chain are *unconditionally dependent*

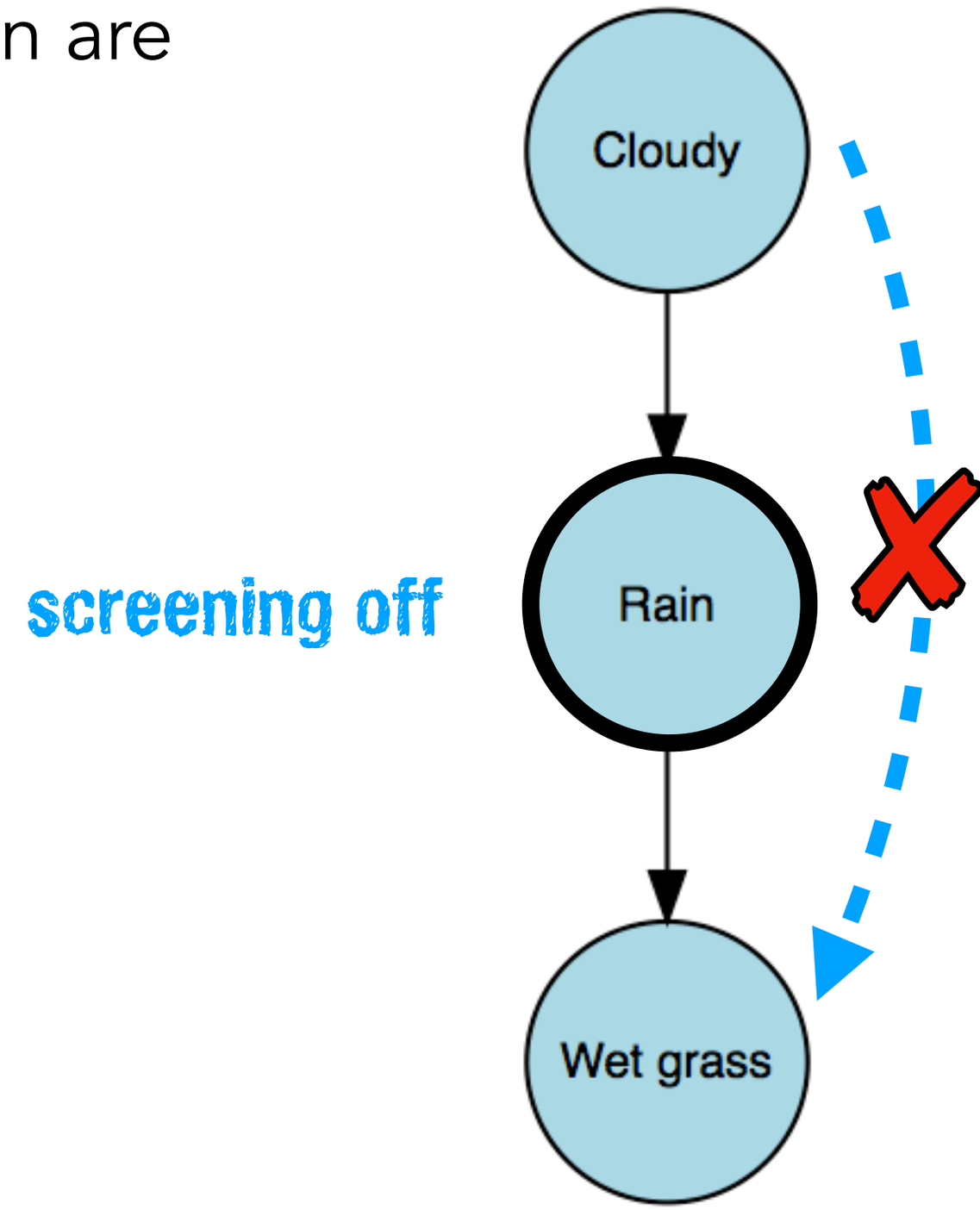
$$p(W \mid C) \neq p(W)$$



Patterns of inference: **Causal chain**

- cause and effect in a causal chain are *conditionally independent*

$$p(W \mid C, R) = p(W \mid R)$$

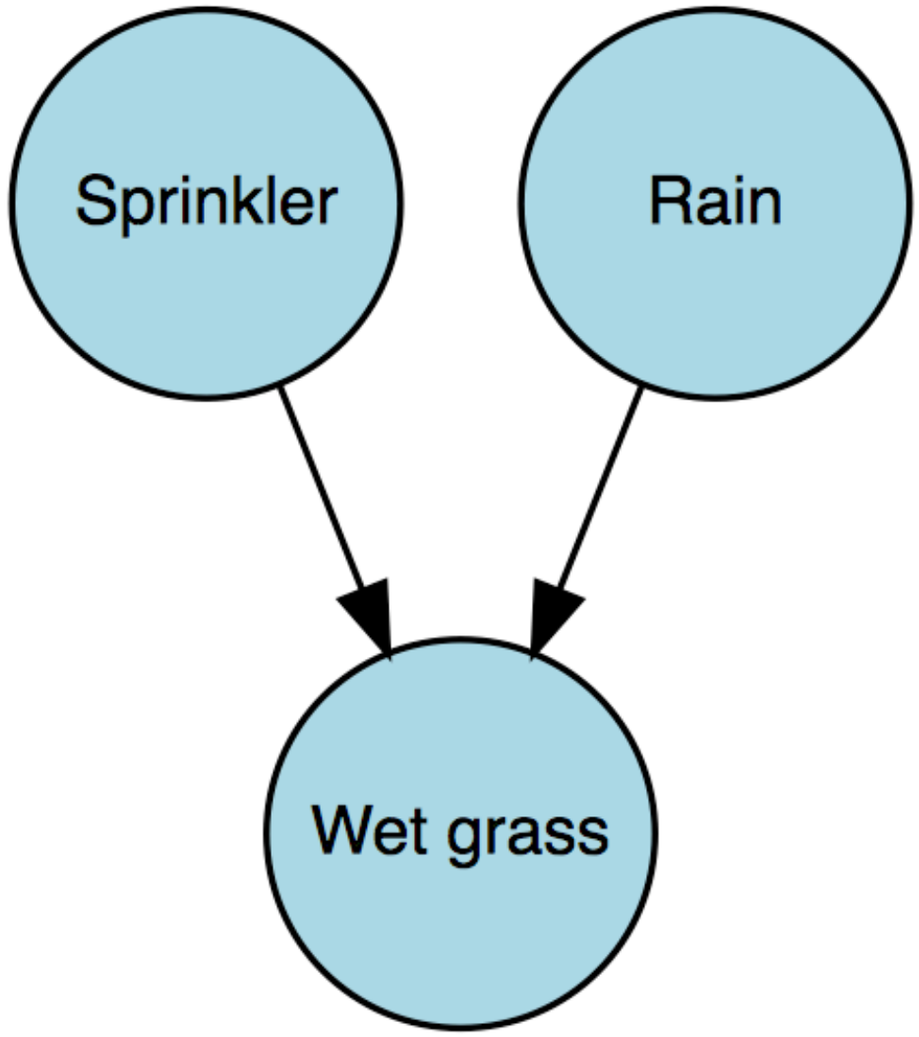


Patterns of inference: **Common effect**

- two causes of a common effect are *unconditionally independent*

$$p(S \mid R) = p(S)$$

(e.g. Sprinklers are set by a timer)

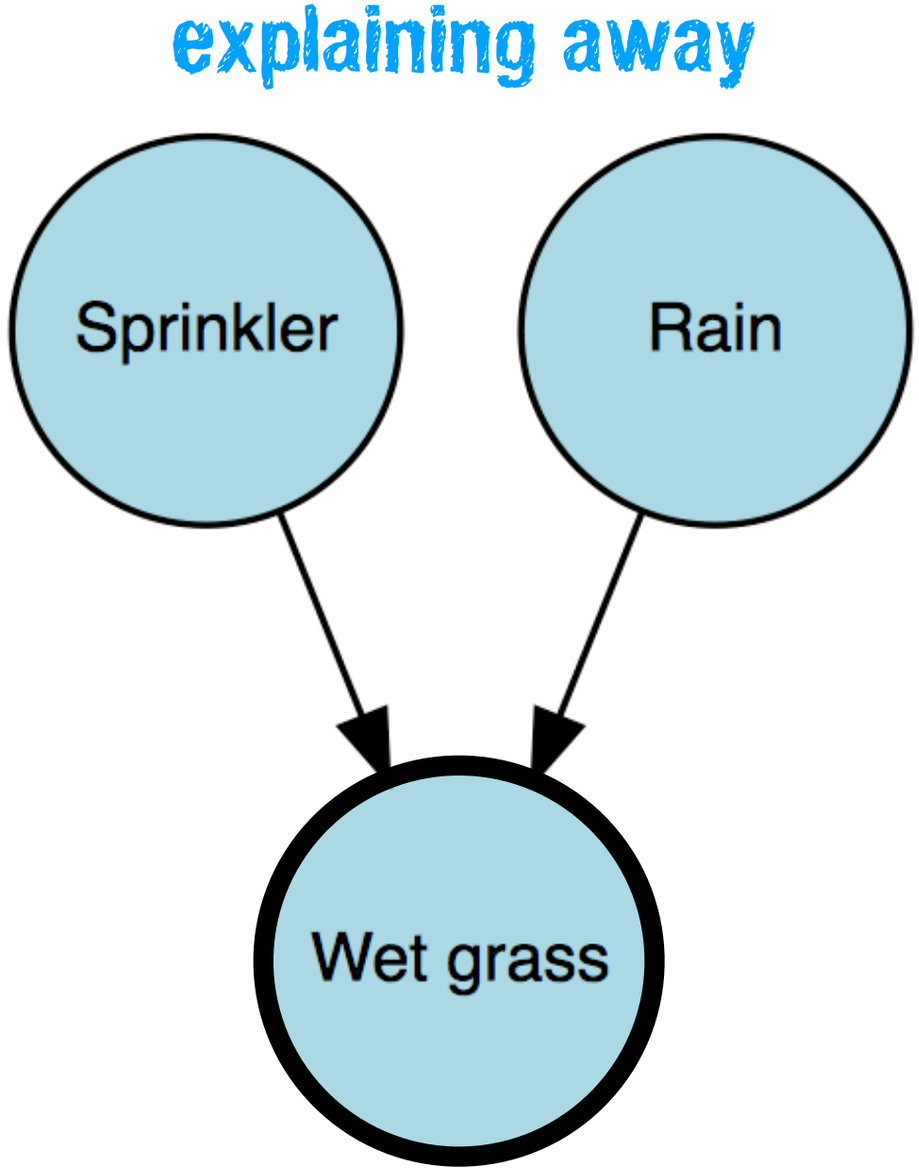


Patterns of inference: **Common effect**

- two causes of a common effect are *conditionally dependent given the effect*

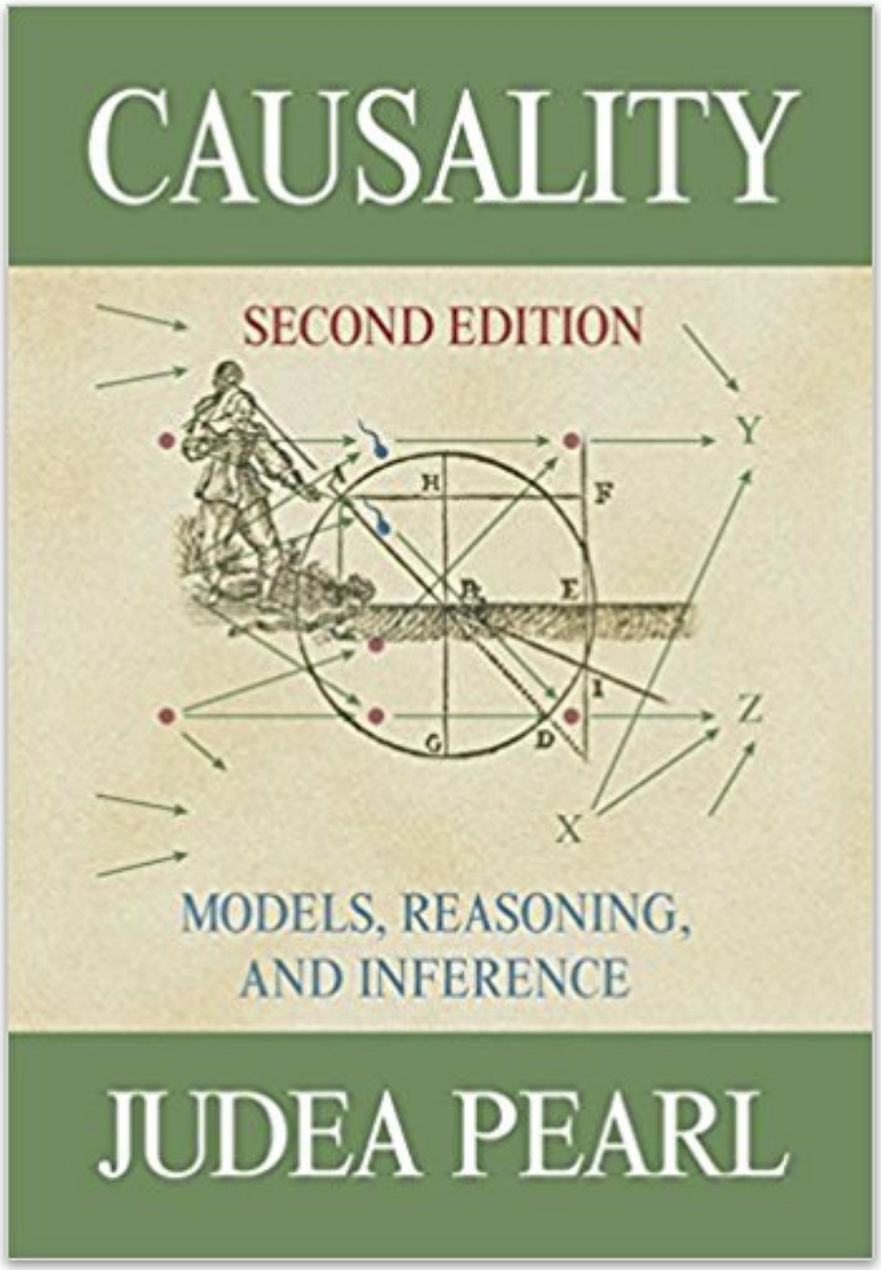
$$p(S \mid R, W) \neq p(S \mid W)$$

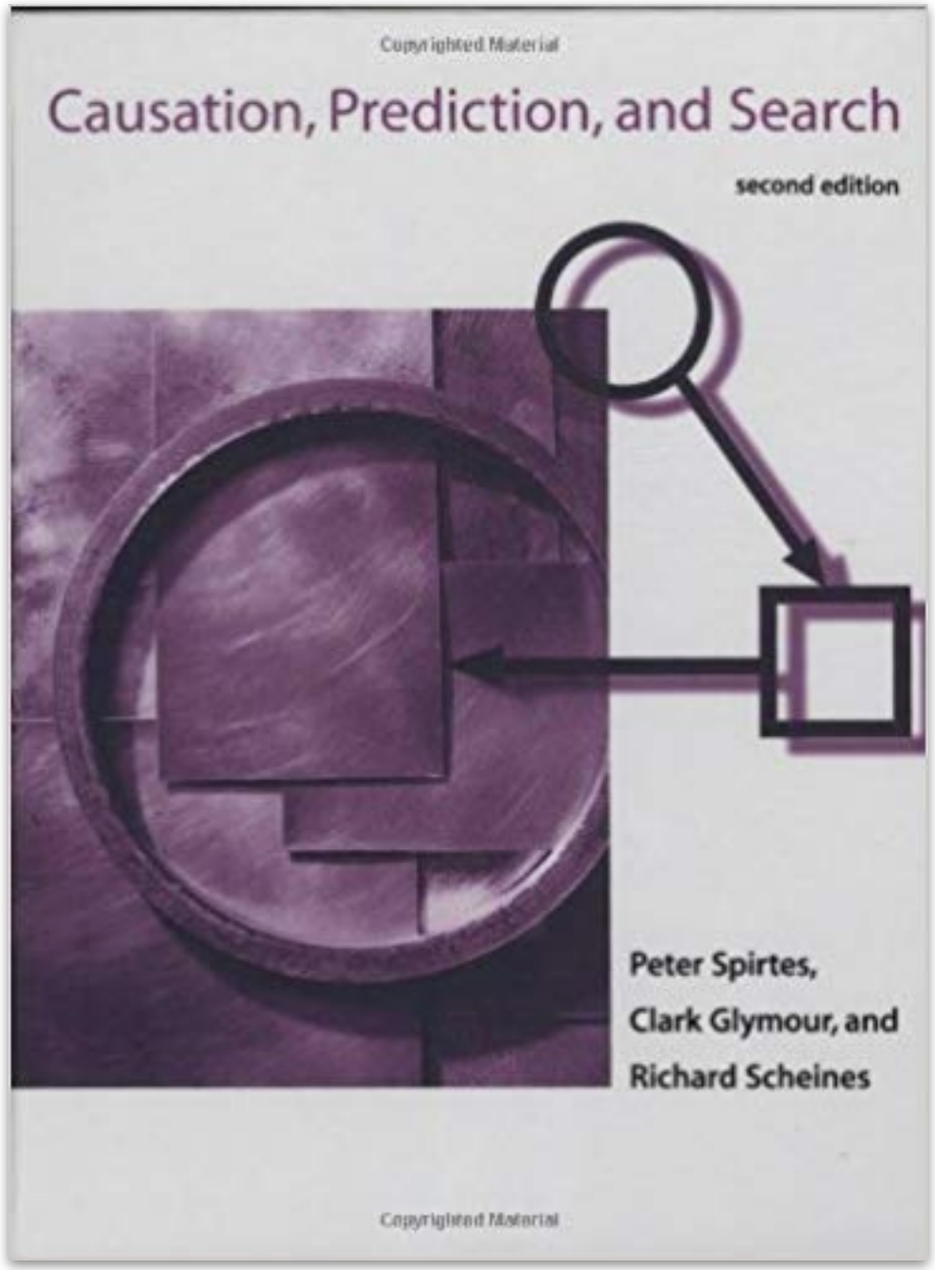
- intuitively: both causes compete to explain the effect



Note: The pattern of inference depends on the structural form which captures how `Sprinkler` and `Rain` jointly affect `Wet grass`. Explaining away holds for the commonly used noisy-or integration function.

Causal Bayesian Networks



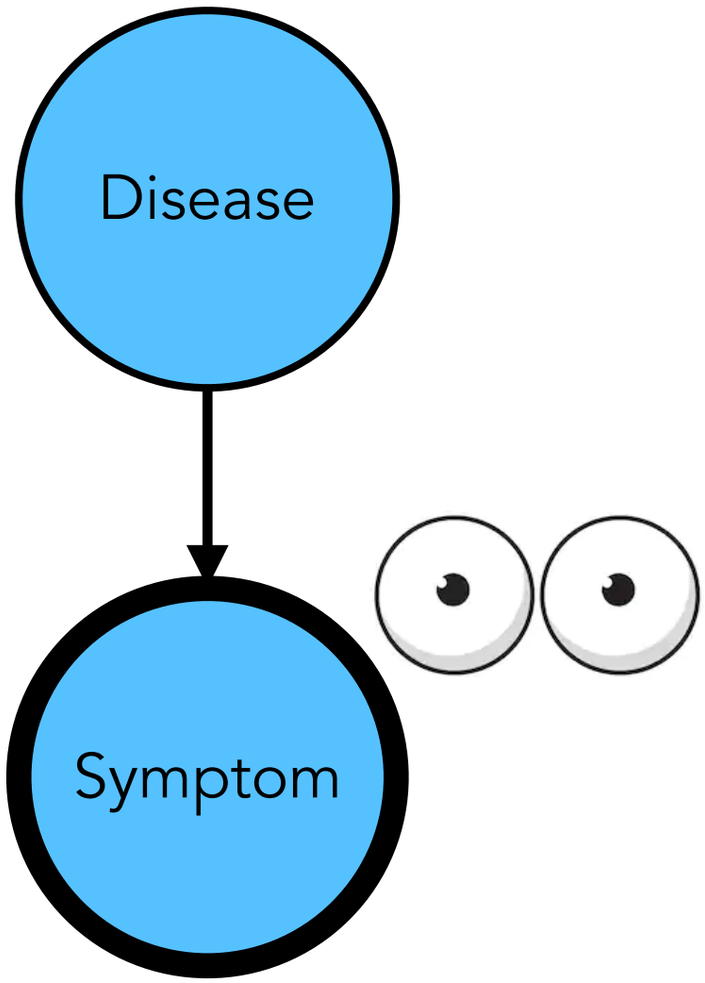


Pearl, J. (2000). Causality: Models, reasoning and inference. Cambridge, England: Cambridge University Press.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). Causation, prediction, and search. The MIT Press.

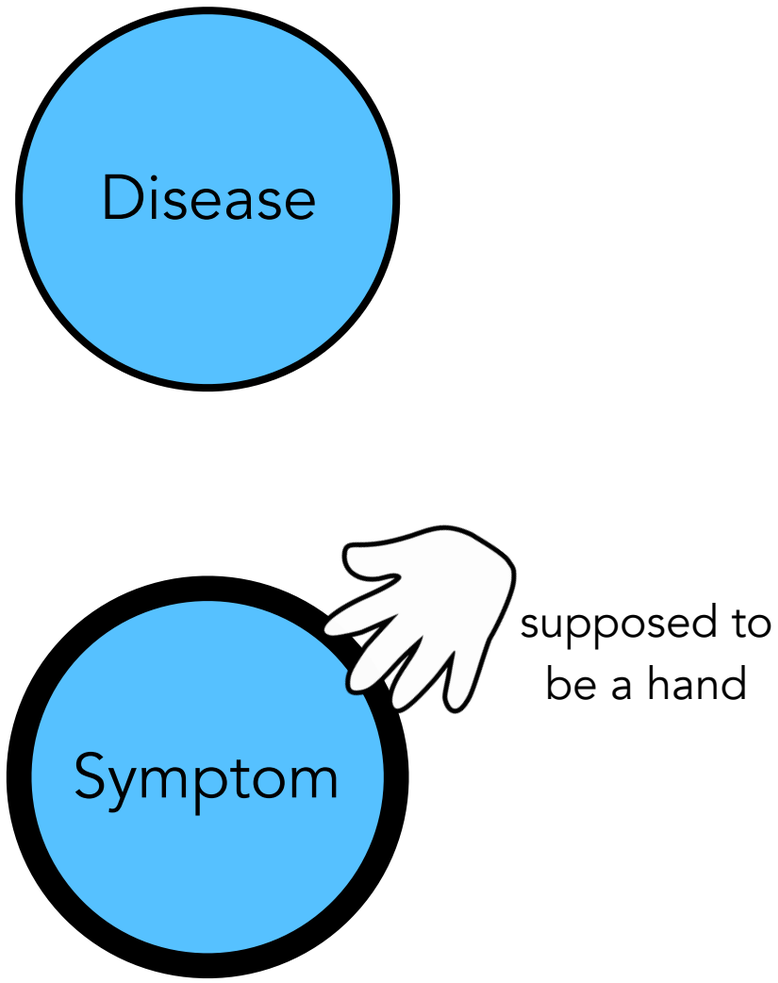
Observation vs. Intervention

seeing



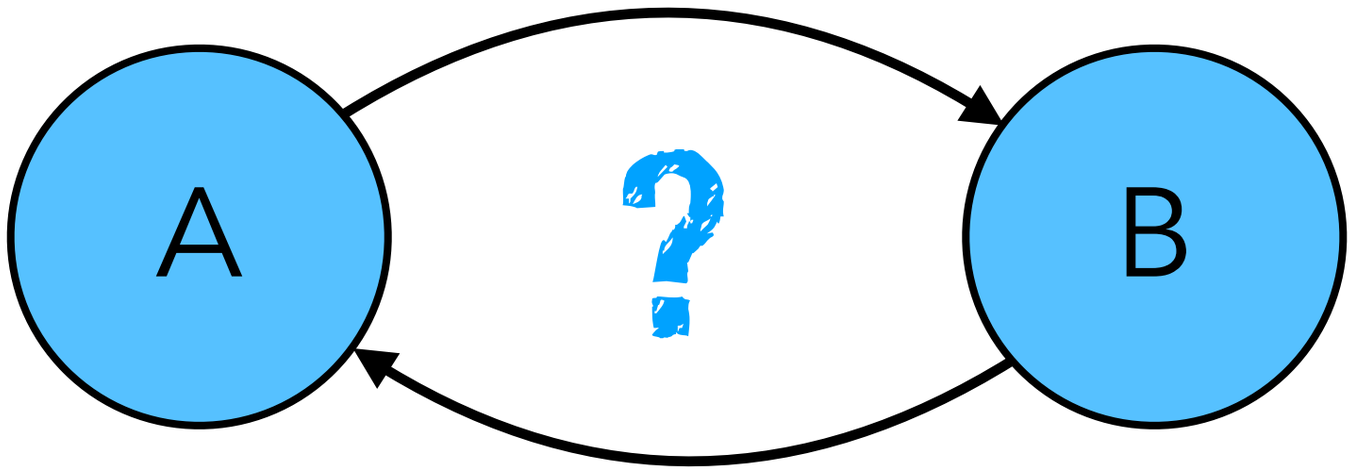
$$p(D \mid S) > p(D)$$

doing

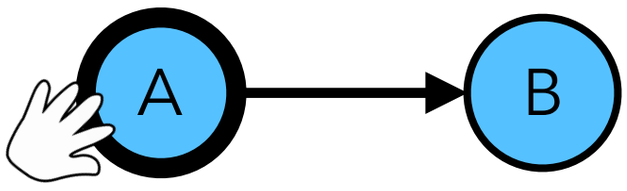


$$p(D \mid \mathrm{do}(S)) = p(D)$$

Inferring causal structure through intervention



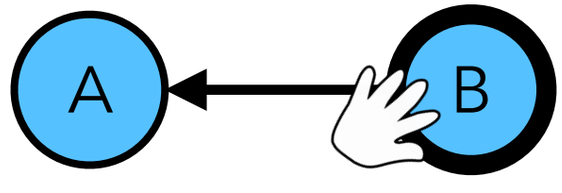
$$p(B \mid \mathrm{do}(A)) = p(B \mid A)$$



$$p(B \mid \mathrm{do}(A)) = p(B)$$

$$p(A \mid \mathrm{do}(B)) = p(A)$$

$$p(A \mid \mathrm{do}(B)) = p(A \mid B)$$



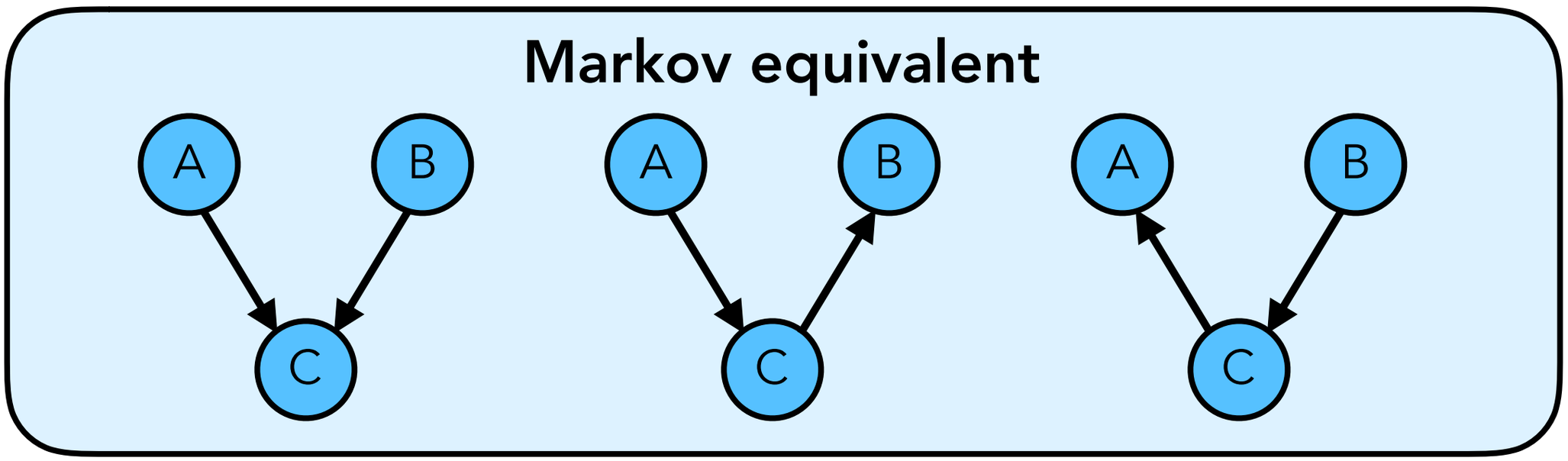
The three layer causal hierarchy

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in$Y$?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Pearl (2018) Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution. *Proceedings of the Eleventh International Conference on Web Search and Data Mining*

Important take home message

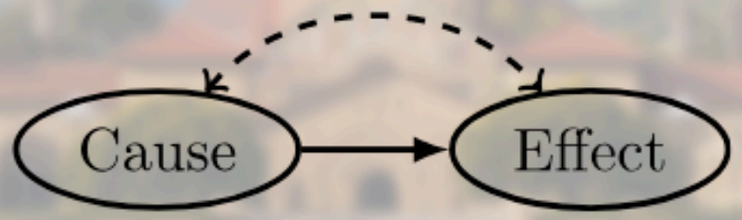
- correlation is not causation
- correlation (= probabilistic dependence) suggests that there is *some* causal relationship
- but we don't know which one it is



- **causal interventions** / experiments can reveal the underlying causal structure

Beyond Curve Fitting: Causation, Counterfactuals, and Imagination-based AI

AAAI Spring Symposium, March 25-27, 2019, Stanford, CA



Motivation

In recent years, Artificial Intelligence and Machine Learning have received enormous attention from the general public, primarily because of the successful application of deep neural networks in computer vision, natural language processing, and game playing (more notably through reinforcement learning). We see AI recognizing faces with high accuracy, Alexa answering English spoken questions efficiently, and Alpha-Zero beating Go grandmasters. These are impressive achievements, almost unimaginable a few years ago. Despite the progress, there is a growing segment of the scientific community that questions whether these successes can be extrapolated to create general AI without a major retooling. Prominent scholars voice concerns that some critical pieces of the AI-puzzle are still pretty much missing. For example, Judea Pearl, who championed probabilistic reasoning in AI and causal inference, recently said in an interview: "To build truly intelligent machines, teach them cause and effect" (link). In a recent OpEd in the New York Times, Cognitive Scientist Gary Marcus noted: “Causal relationships are where contemporary machine learning techniques start to stumble” (link).

These and other critical views regarding different aspects of the machine learning toolbox, however, are not a matter of speculation or personal taste, but a product of mathematical analyses concerning the intrinsic limitations of data-centric systems that are not guided by explicit models of reality. Such systems may excel in learning highly complex functions connecting input X to an output Y, but are unable to reason about cause and effect relations or environment changes, be they due to external actions or acts of imagination. Nor can they provide explanations for novel eventualities, or guarantee safety and fairness. This symposium will focus on integrating aspects of causal inference with those of machine learning, recognizing that the capacity to reason about cause and effect is critical in achieving human-friendly AI. Despite its centrality in scientific inferences and commonsense thinking, this capacity has been largely overlooked in ML, most likely because it requires a language of its own, beyond classical statistics and standard logics. Such languages are available today and promise to yield more explainable, robust, and generalizable intelligent systems.

Summary

- Introduction to probability / Recap
  - Counting possibilities
  - Interpretation of probability
  - Clue guide to probability
- Bayesian Networks
  - representation
  - inference
  - (un-)conditional (in-)dependence
- Causal Bayes nets

Thank you!