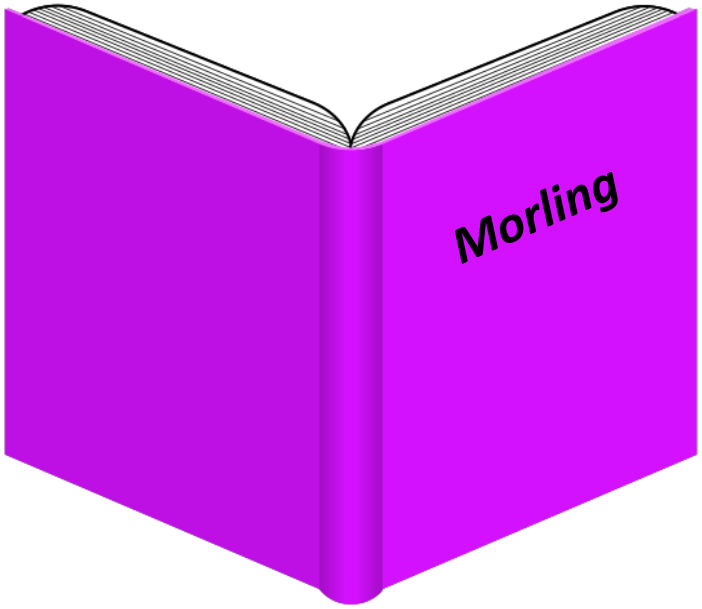


Research Methods and Statistics

Lecture 5: Probability

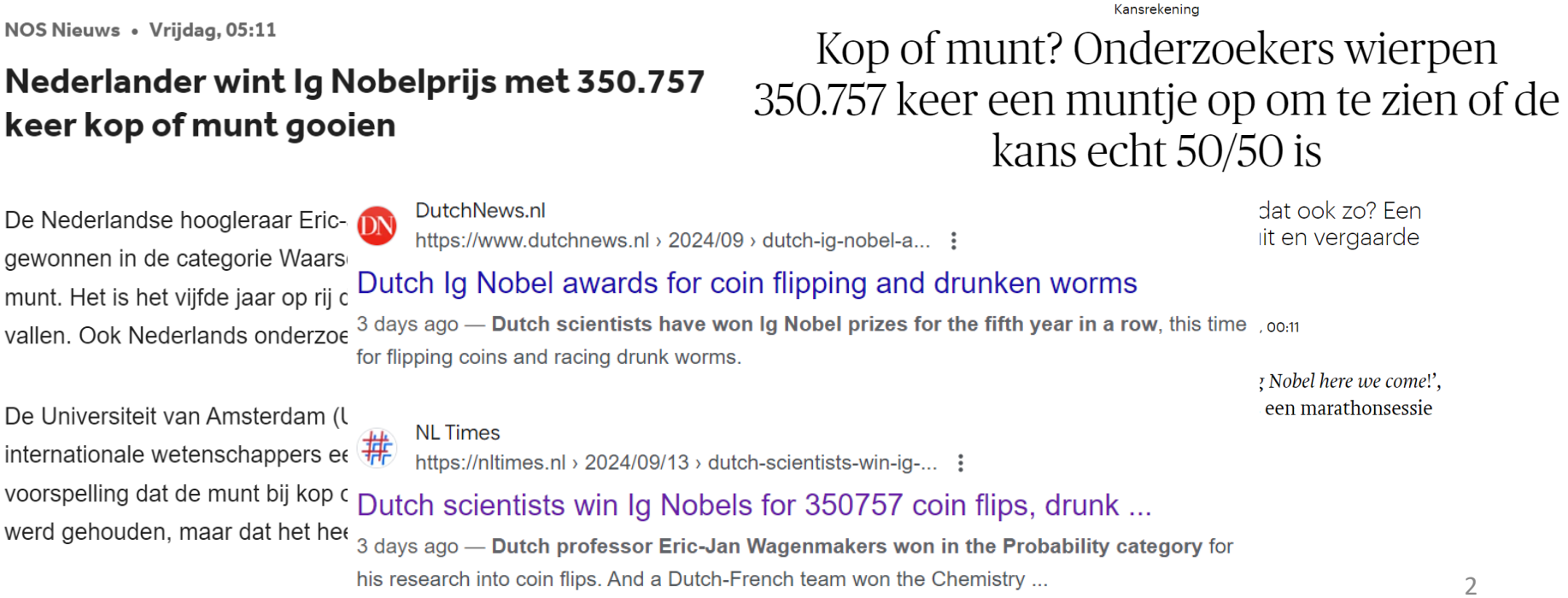
Riet van Bork

Pictures source: pixabay.org

Probability in the news

The **Ig Nobel Prize** is a satiric prize awarded to "honor achievements that first make people laugh, and then make them think.“

https://improbable.com/ig/winners/



NOS Nieuws • Vrijdag, 05:11

Nederlander wint Ig Nobelprijs met 350.757 keer kop of munt gooien

De Nederlandse hoogleraar Eric-
gewonnen in de categorie Waars
munt. Het is het vijfde jaar op rij c
vallen. Ook Nederlands onderzoe

De Universiteit van Amsterdam (U
internationale wetenschappers ee
voorspelling dat de munt bij kop c
werd gehouden, maar dat het hee

Kansrekening

Kop of munt? Onderzoekers wierpen 350.757 keer een muntje op om te zien of de kans echt 50/50 is

dat ook zo? Een
it en vergaarde

, 00:11

; Nobel here we come!',
een marathonsessie

DutchNews.nl
https://www.dutchnews.nl › 2024/09 › dutch-ig-nobel-a...

Dutch Ig Nobel awards for coin flipping and drunken worms

3 days ago — **Dutch scientists have won Ig Nobel prizes for the fifth year in a row**, this time for flipping coins and racing drunk worms.

NL Times
https://nltimes.nl › 2024/09/13 › dutch-scientists-win-ig-...

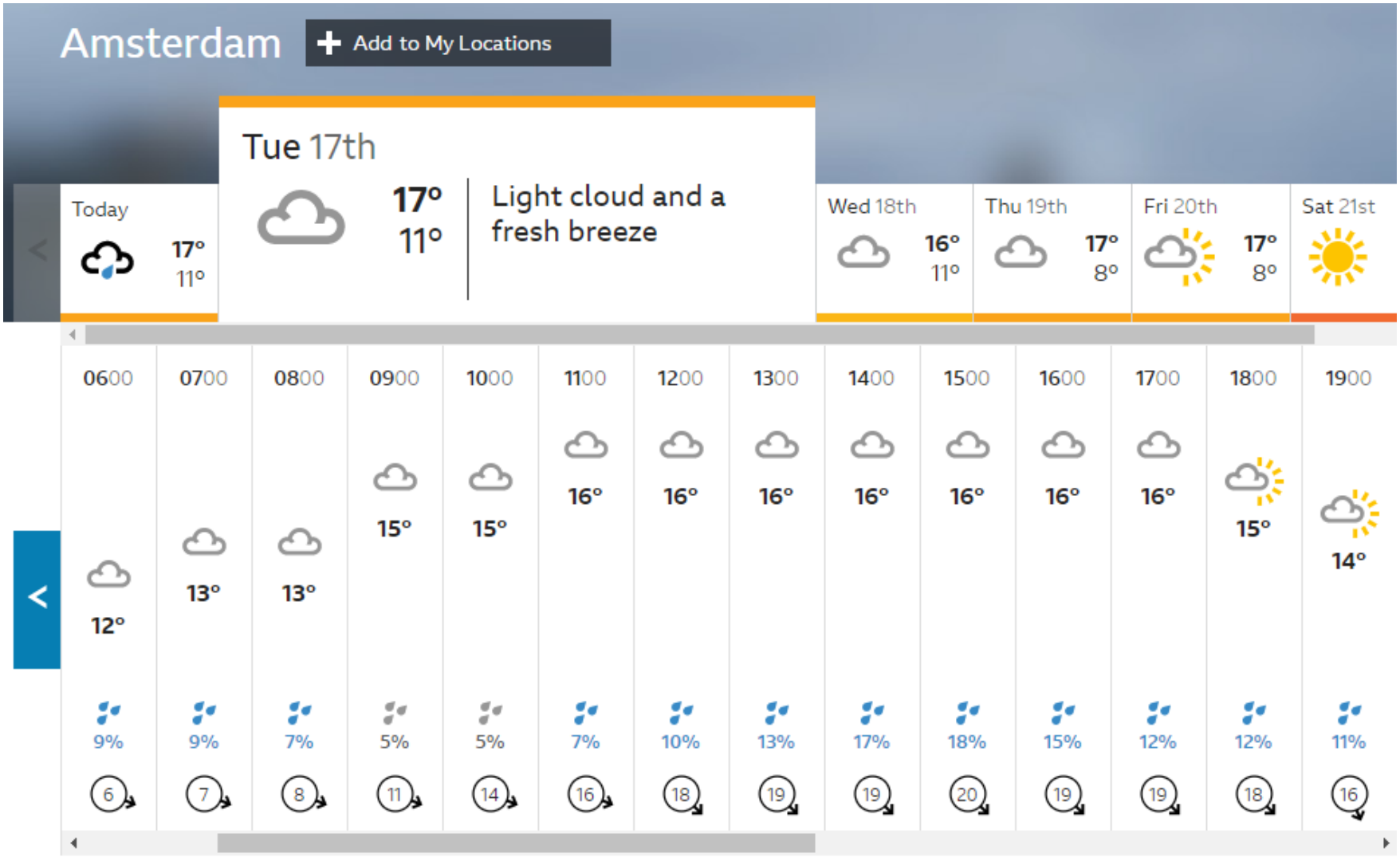
Dutch scientists win Ig Nobels for 350757 coin flips, drunk ...

3 days ago — **Dutch professor Eric-Jan Wagenmakers won in the Probability category** for his research into coin flips. And a Dutch-French team won the Chemistry ...

Probability in the news

Eric-Jan Wagenmakers (left) and František Bartoš (right) getting the prize awarded: https://www.youtube.com/watch?v=ukBwV9Lap2A&t=3529s&ab_channel=ImprobableResearch (See minute 56)

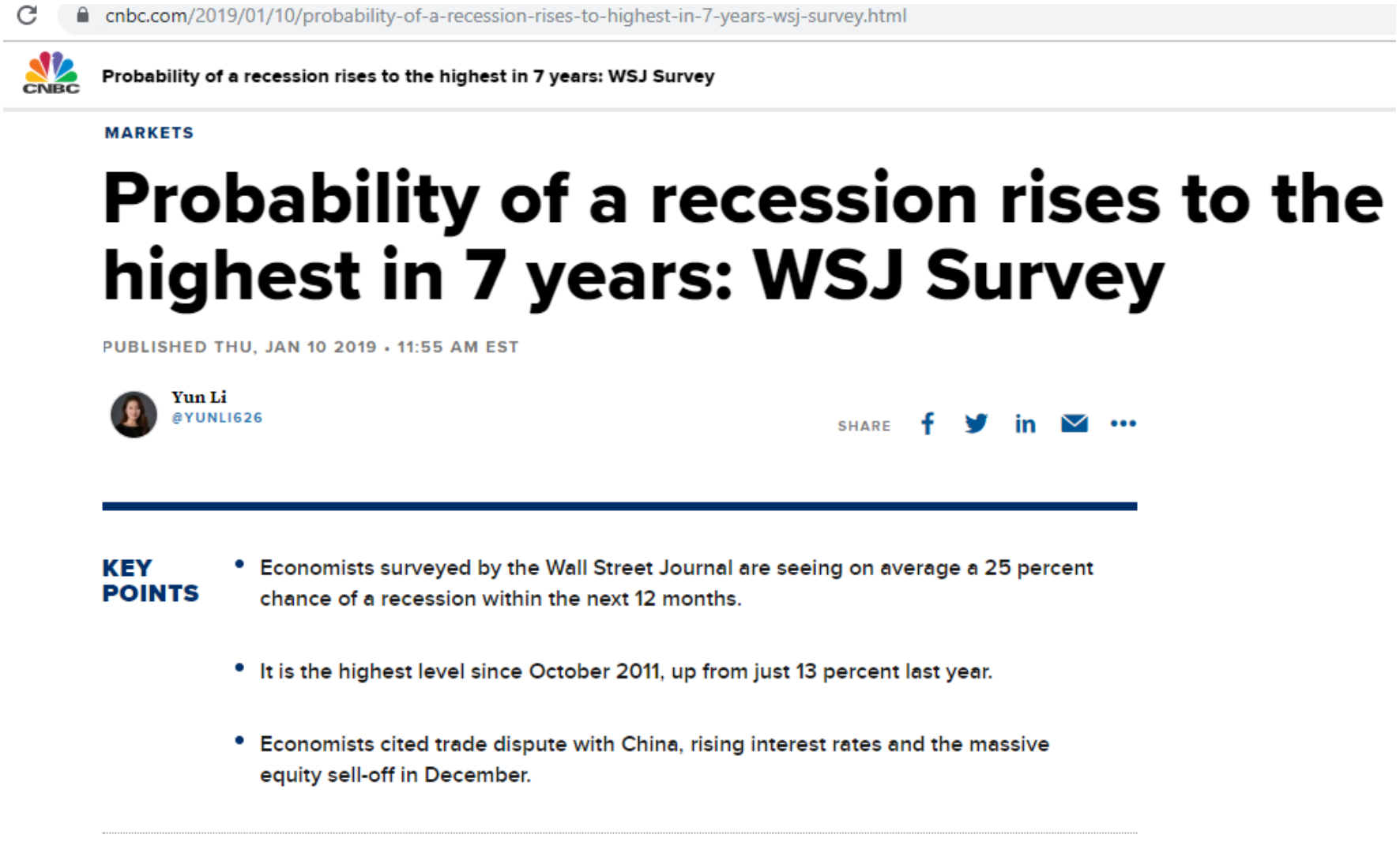
Weather forecast: “7% chance of rain”



Source: bbc.com/weather

Stormy Daniels:
“there’s a 50-50 shot what I know could bring down the president”

News:
“probability of a recession is 25 percent”



Today

What is probability?

Finding probability

What is probability?

The proportion of times that a particular outcome occurs in the long run over independent trials of a random phenomenon

What is probability?

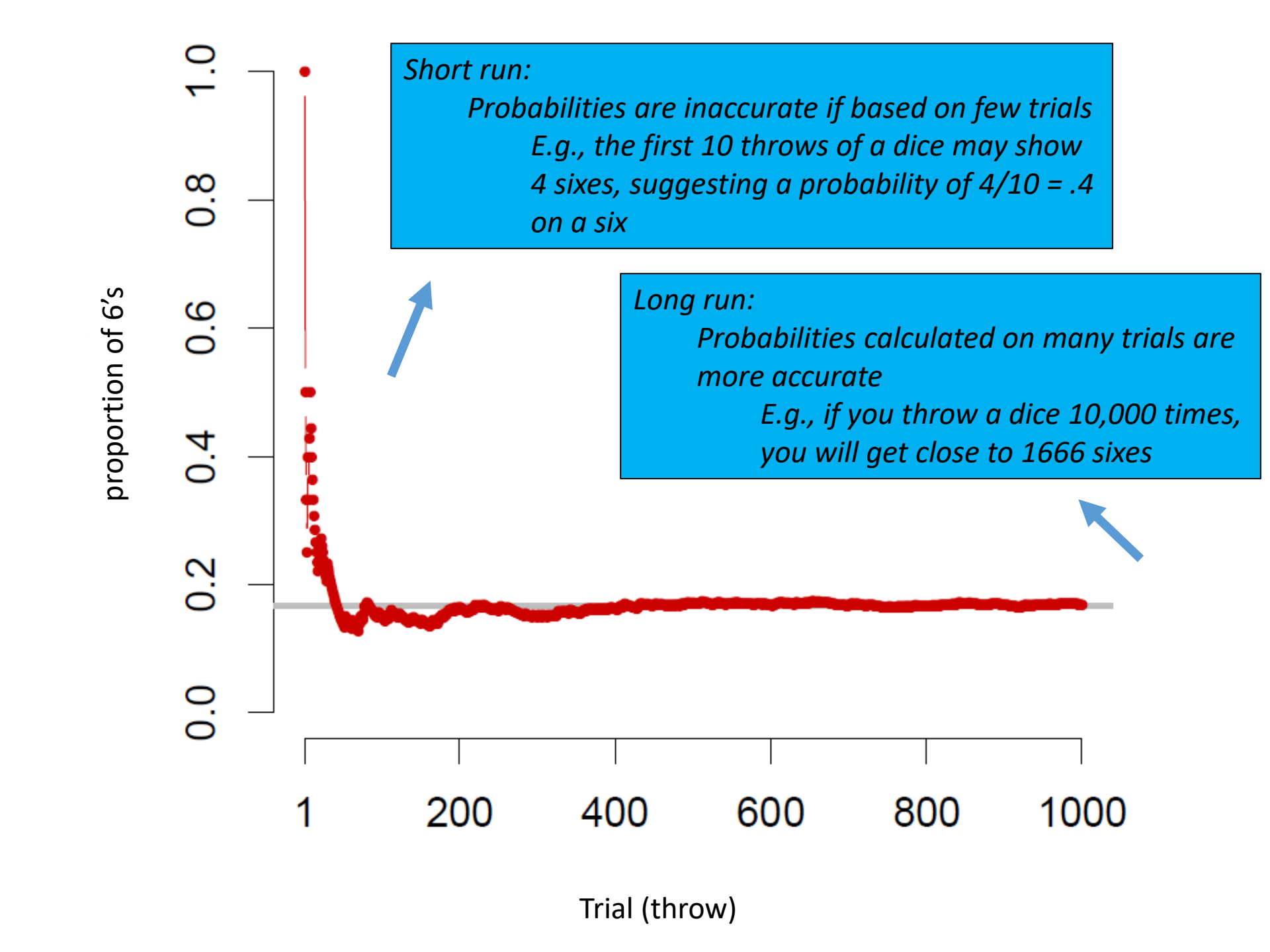
The proportion of times that a particular outcome occurs in the long run over independent trials of a random phenomenon

Where does probability apply to?

- It applies to **Random phenomena**
  - occurances for which the outcome is uncertain
    - Weather: Will it be raining tomorrow?
    - Sports: Will I win my tennis match?
    - Health care: Will a given drug be effective?
    - Education: Will I pass my exam?
    - Marketing: Will this commercial increase sales?
    - Etc.
- It does not apply to
  - Things that are not occurences
    - There is no meaningful relative frequency/proportion
    - E.g., “the earth is round”
  - Occurences that are *deterministic* (only one outcome is possible)
    - The relative frequency/proportion for one outcome is 1 and for all other outcomes is 0
    - E.g., a machine that throws a dice and you can make the machine such that it always throws heads

What is probability?

The proportion of times that a particular outcome occurs in the long run over independent trials of a random phenomenon


Short run:
Probabilities are inaccurate if based on few trials
E.g., the first 10 throws of a dice may show 4 sixes, suggesting a probability of 4/10 = .4 on a six
Long run:
Probabilities calculated on many trials are more accurate
E.g., if you throw a dice 10,000 times, you will get close to 1666 sixes
proportion of 6's
0.0
0.2
0.4
0.6
0.8
1.0
1
200
400
600
800
1000
Trial (throw)

Law of large numbers

- If the number of trials increases, the proportion of occurences of any outcome approaches a given number
  - I.e., 1/6 = 0.1667 for throwing a 6 with a die
  - I.e., 0.5 for “heads” for tossing a coin

- Holds only for *independent trials* →

What is probability?

The proportion of times that a random phenomenon occurs in the long run over ***independent trials***

- Assume a fair coin (equal probabilities heads & tails). Which sequence will likely show ‘Heads’ in the next throw?

Sequence 1: H T T T T T T T

Sequence 2: H T H T H T T H

Sequence 3: H H H H H H H T

Gamblers fallacy: the mistaken belief that, if a particular event occurs more frequently than expected in previous trials, it is less likely to happen in next trials (or vice versa)

Independent trials

Different trials of a random phenomenon are independent if the outcome of any one trial is not affected by the outcome of another trial

In practice this applies to:

- Tossing dice
- Flipping a coin
- Roulette
- Selecting a random student from the audience (with replacement)

Does not apply to:

- Will I win my tennis match?
  - Because you may become better over time, so the probability of winning increases, and you face opponents of different skill
- Will a given drug be effective for me?
  - Because if a drug works the first time, it will affect the effectiveness the next time (e.g., habituation)
- Will I pass my exam?
  - Because you learn, and because passing or failing the first time will affect your attitude towards the next exam

- In these latter cases, for it to be a probability, the proportion has to be over *hypothetical* independent repeated trials..

Probability

- Thus: probability = long run proportion

- In some practical examples, it is unrealistic to think of many independent trials
  - Will I pass the exam?
  - Will I get better from taking this drug?
  - Will it rain tomorrow?
  - Will Stormy Daniels information bring down Trump?
- Then:

Either, define a hypothetical (possibly unrealistic) chance experiment

  - If I would get multiple exams with representative questions, what proportion would I pass?
    - Or: if other people who prepared for the exam like I did would all do the exam, what proportion would pass?
  - If different people with similar symptoms as mine would be sampled and given the drug, what proportion would get better?
  - For days with similar characteristics as today, what proportion is followed by a day with rain?
  - Stormy Daniels?

Or, people may adopt the so-called "subjective definition of probability" (Bayesian)

  - Probability = degree of belief

- More about Bayesian probability in the *fourth block* of this course!

Why should we care about coin flips and dice tosses?

What do coins have to do with human beings?

We often study random phenomena related to psychology,
e.g., the probability that someone will develop depression
e.g., probability that someone gives a correct response to a math item

- Do we have independent trials?
- What is “the long run”?

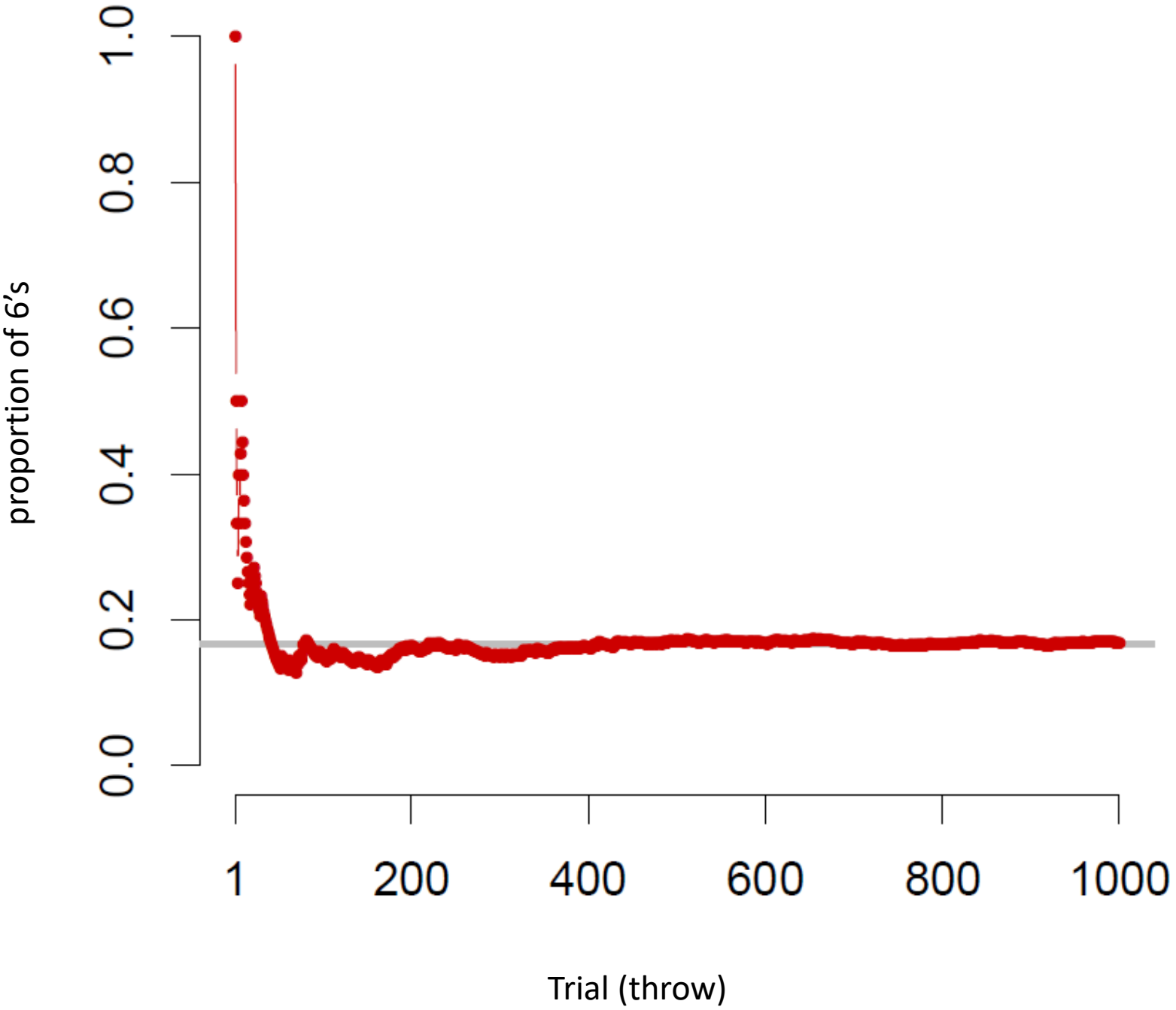
→In the social sciences, we treat each person as a new ‘trial’.

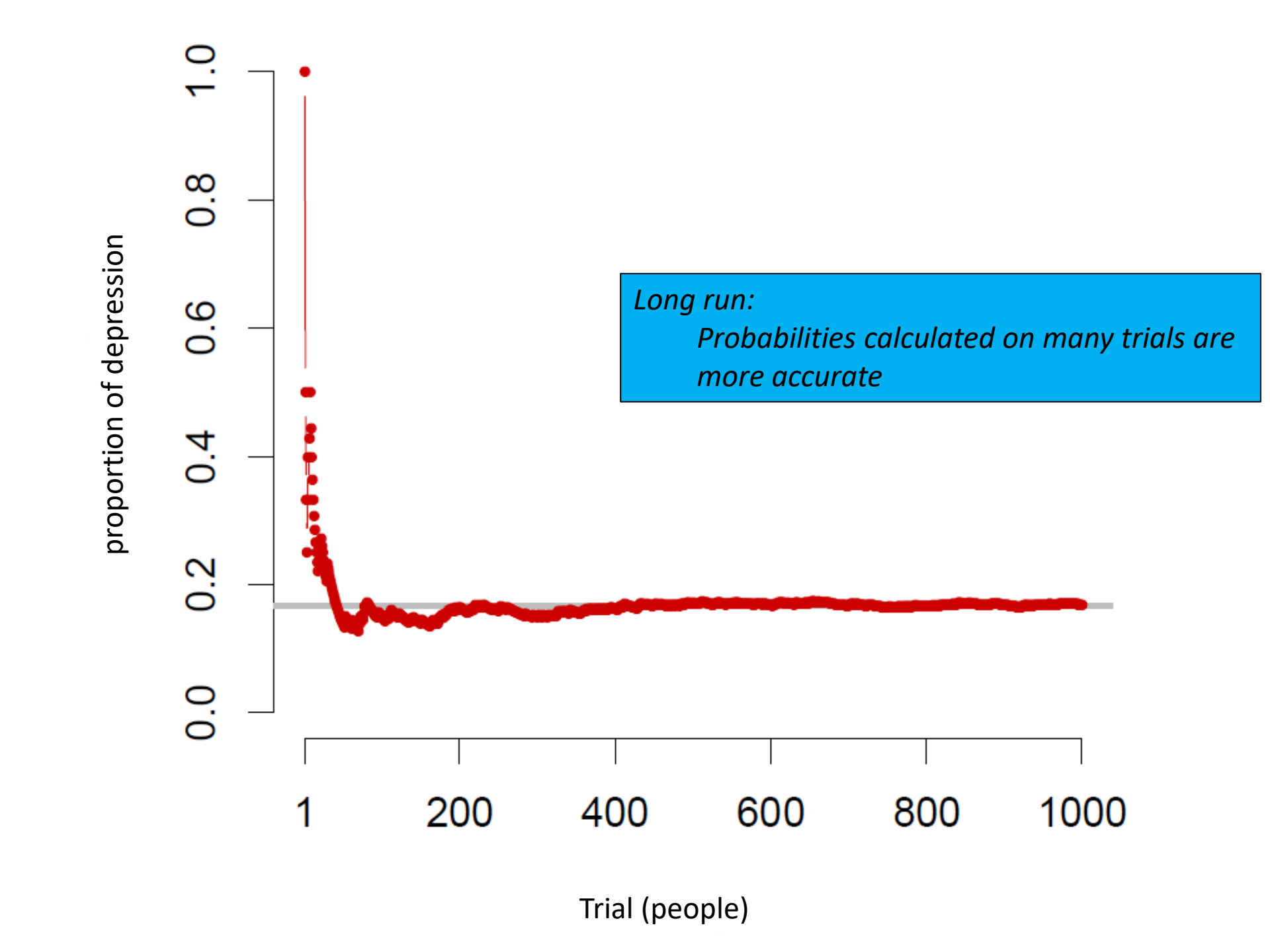
→If we select people *randomly* from a *large* population, these can be *considered* independent

→If we select a lot of people, this approximates ‘the long run’

Why considered independent?

- Suppose you are interested in “what is the probability of selecting someone with depression?”
- In a small group of people, you would have to sample with replacement for the trials to be independent.
  - For example, with 5 people of whom 3 are depressed, the first time you sample someone the probability is 3/5, but the second time not anymore! (unless you do replacement)
- In a large population, even if you do not do with replacement, we say the trials are considered independent because selecting someone hardly changes the probability of the next trial
  - For example, if you have a population of one million people, and 10% is depressed, then the first time you sample someone the probability is 0.1, and the second time it is still almost 0.1
- In a large population the influence of taking out people is so small we can ignore it.


proportion of 6's
0.0
0.2
0.4
0.6
0.8
1.0
1
200
400
600
800
1000
Trial (throw)


proportion of depression
0.0
0.2
0.4
0.6
0.8
1.0
1
200
400
600
800
1000
Trial (people)
Long run:
Probabilities calculated on many trials are more accurate

Today

What is probability?

Finding probability

Today

What is probability?

Finding probability

-Venn Diagram

-Rules for finding probability

Sample space

- For a random phenomenon, the **sample space** is the set of all possible outcomes
  - What is the sample space of:

  -tossing a die once?

  -flipping a coin twice?

  -sampling a person and registering place of birth?

  -sampling a person from this room and registering place of birth?

Sample space

- For a random phenomenon, the **sample space** is the set of all possible outcomes
  - E.g., tossing a die once: {1, 2, 3, 4, 5, 6}
  - E.g., flipping a coin once: {heads, tails}
  - E.g., place of birth: {Amsterdam, Rotterdam, Berlin, ... }
  - Flipping 2 coins: {HH, HT, TH, TT}
  - Flipping 3 coins:
    {TTT, HTT, THT, TTH, HHT, HTH, THH, HHH}

Event

- **Event**: A subset of the sample space
  - i.e., a particular outcome or a group of possible outcomes

For instance:

- A = you throw a six with a die = {6}
- B = you throw an even number with a die = {2, 4, 6}
- C = you throw exactly 2 heads with 3 coins = {HHT, HTH, THH}
- D = a randomly selected student is born in Berlin = {Berlin}

Venn diagram

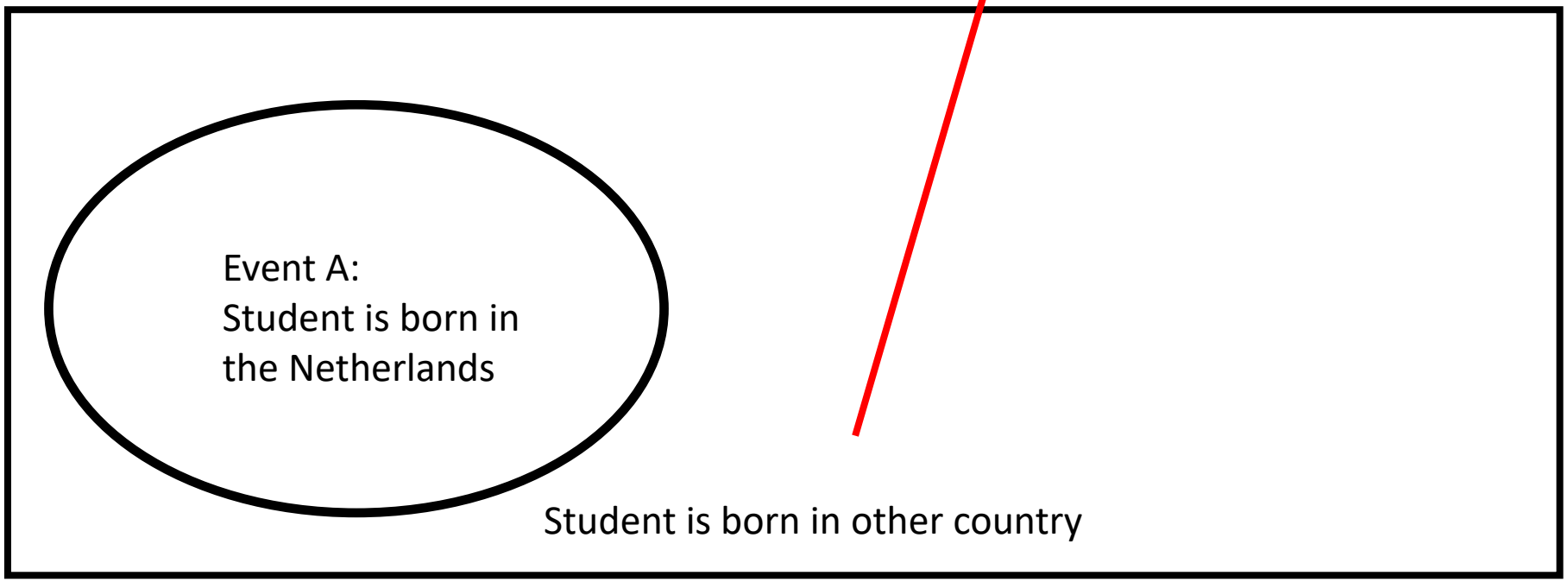
- I randomly select one of you

Sample space

Venn diagram

The complement of A denoted A^c
(i.e., “not A”)

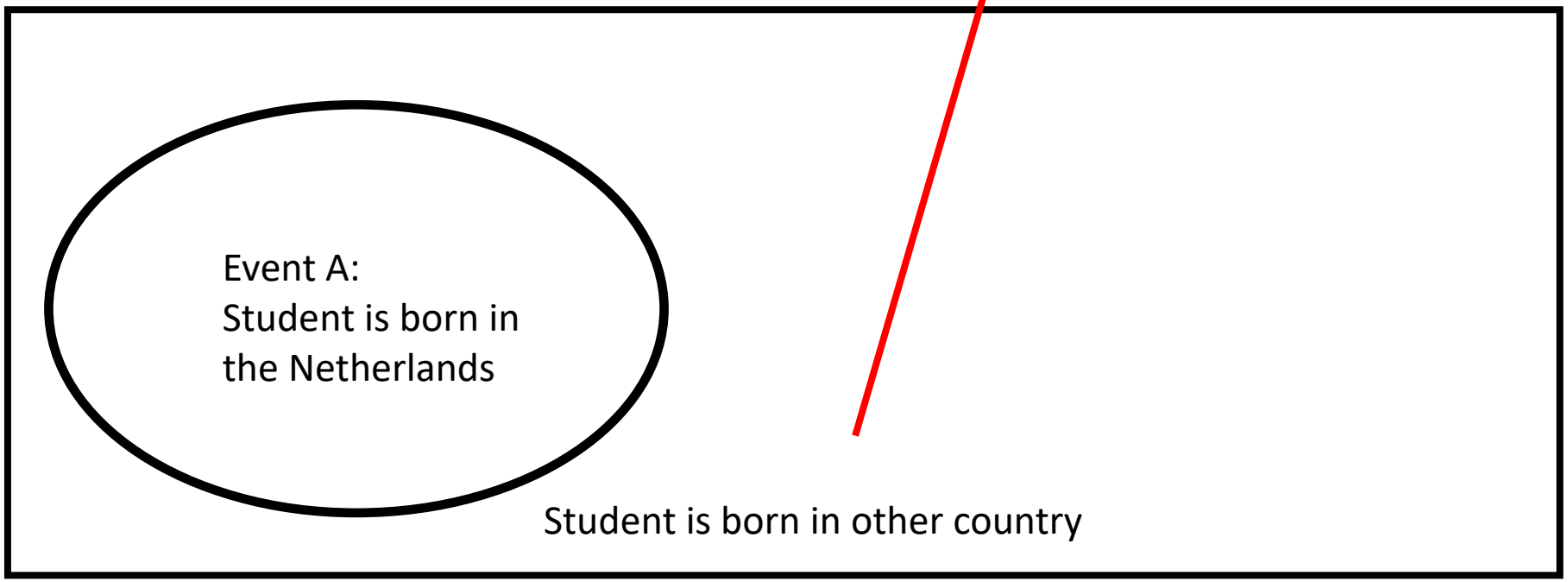
- I randomly select one of you



Venn diagram

The complement of A denoted A^c
(i.e., "not A")

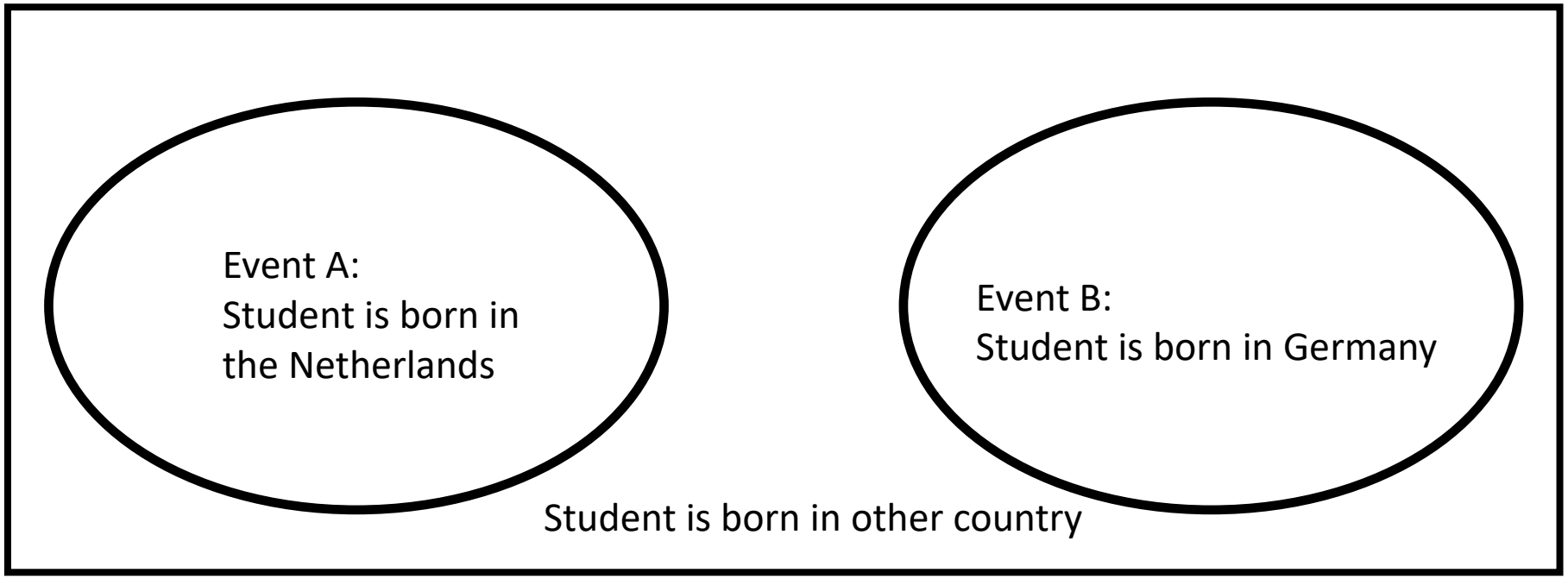
- I randomly select one of you



What is the probability of 'A or A^c' ?

Venn diagram

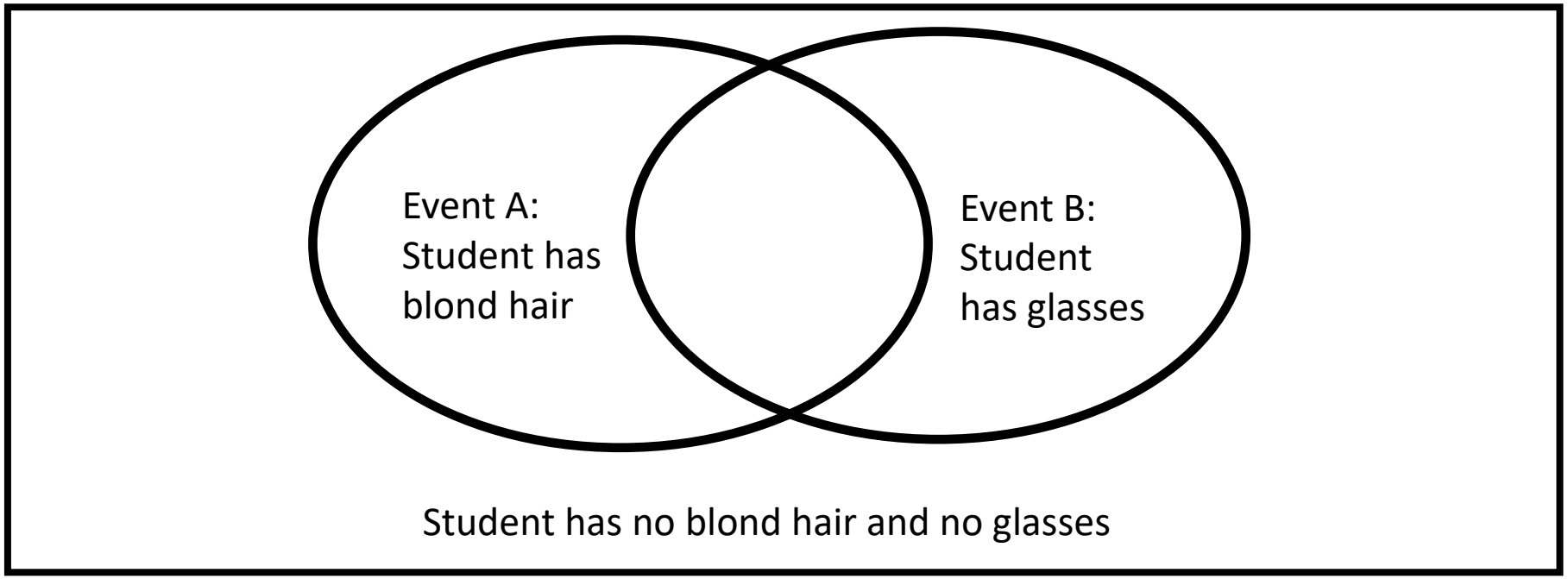
- I randomly select one of you



Disjoint: The two events do not overlap (you cannot be born in the Netherlands ánd in Germany)

Venn diagram

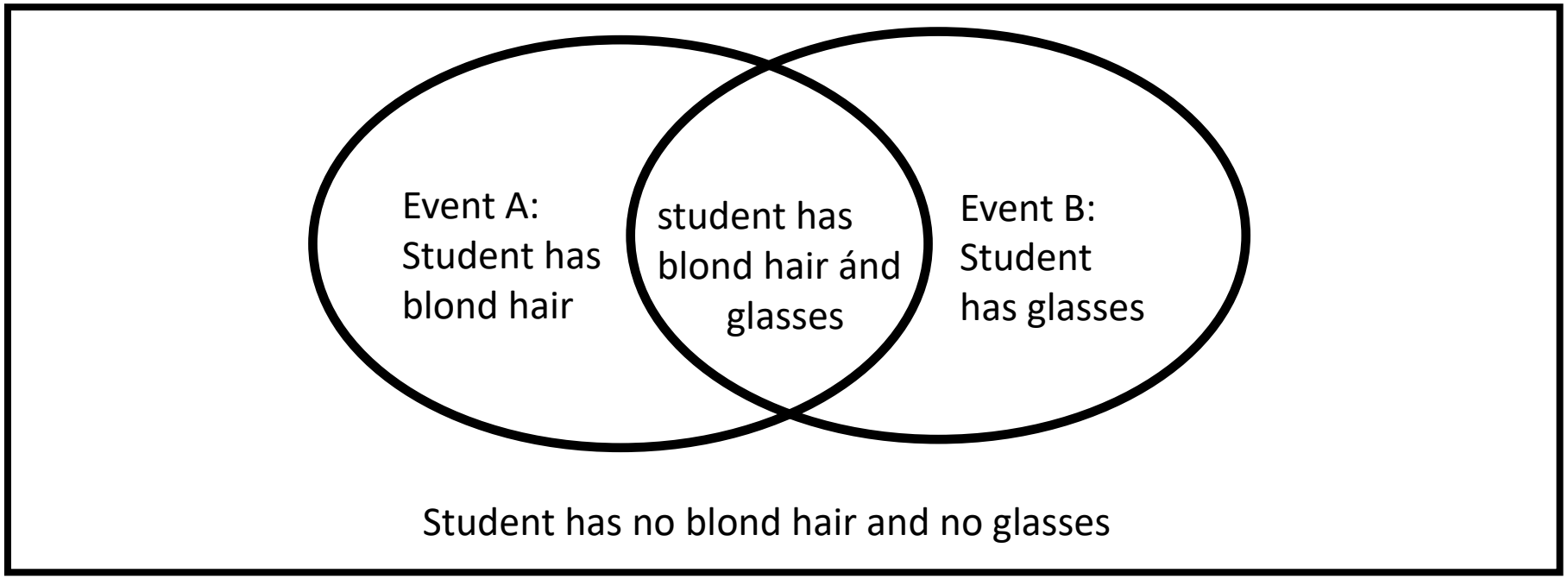
- I randomly select one of you



Venn diagram

- I randomly select one of you

Sample space

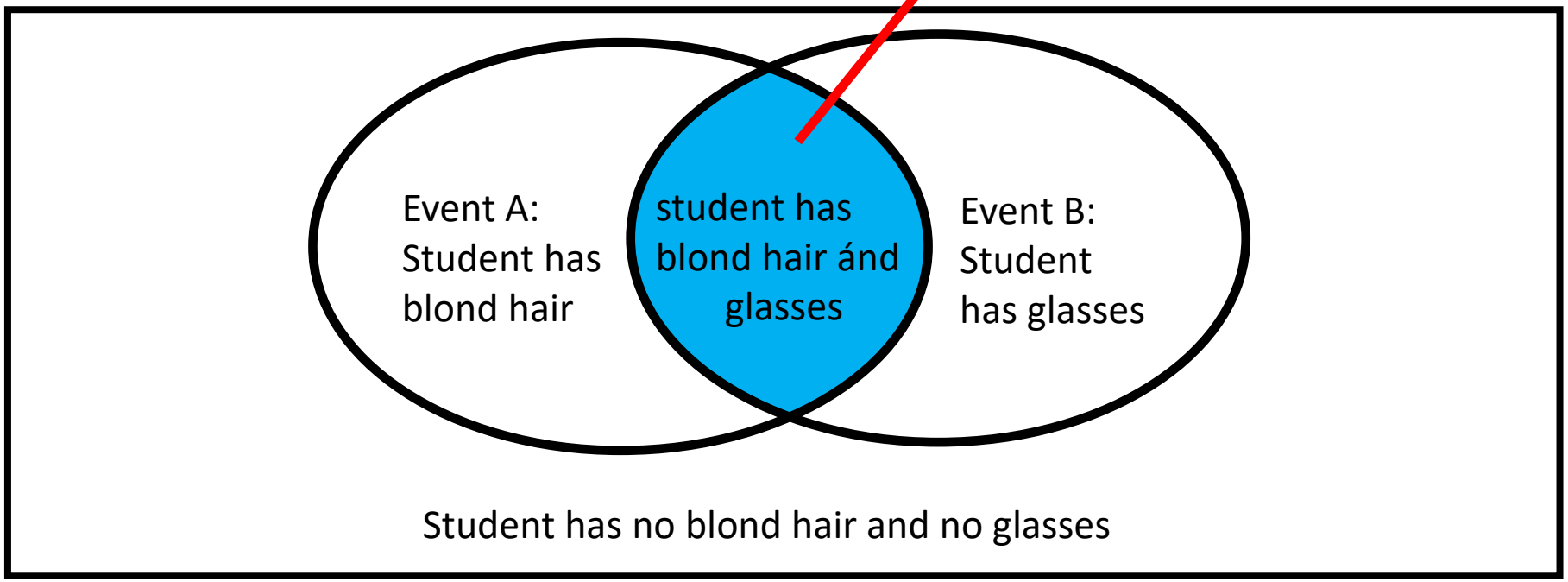


The two events overlap (you can have blond hair ánd wear glasses) → *Not* disjoint

Intersection: {"Blond student with glasses"}

Venn diagram

- I randomly select one of you

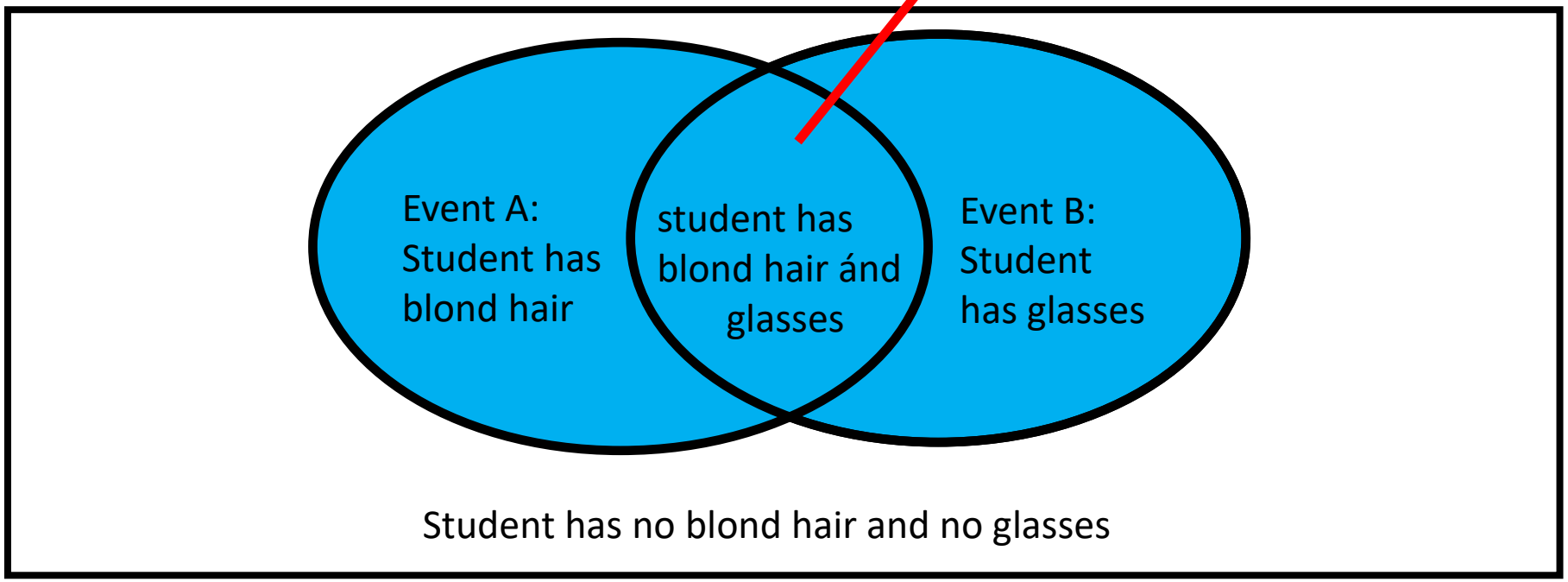


P(blond AND glasses)

Union: {“Blond students”, “Students with glasses”, “students with blond hair ánd glasses”}

Venn diagram

- I randomly select one of you



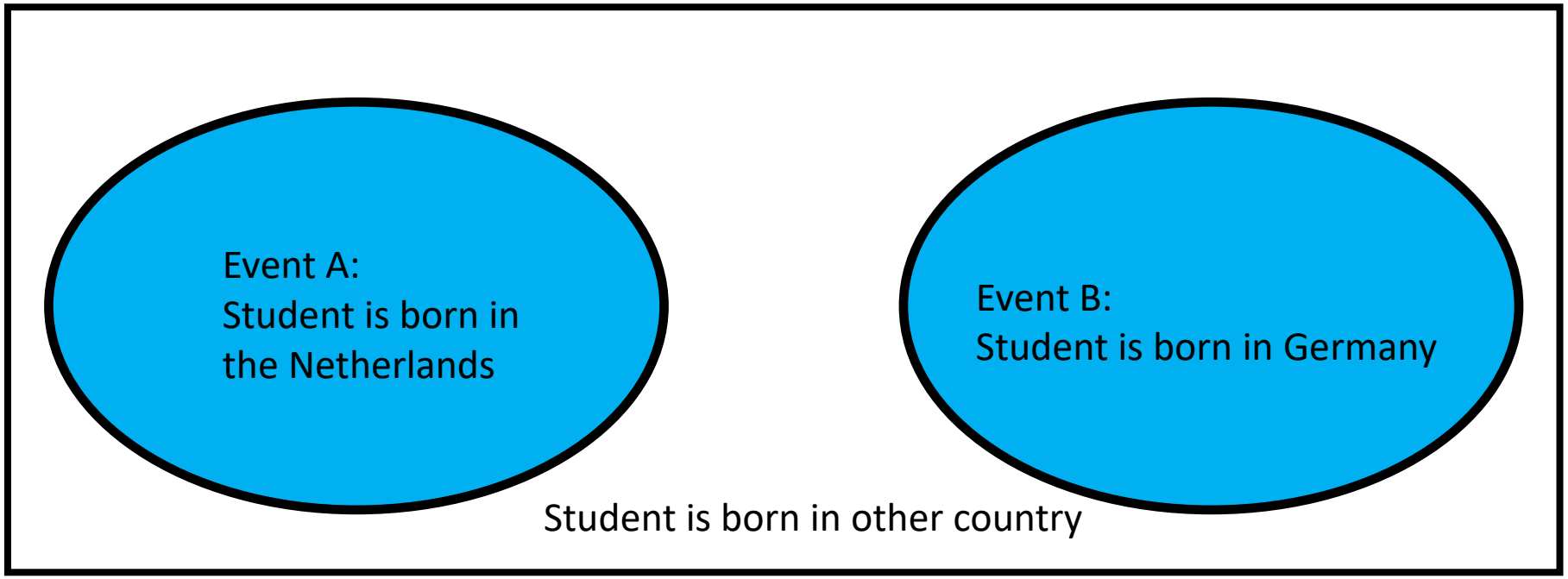
P(blond OR glasses)

Union: {“Born in the Netherlands” , “Born in Germany”}

Venn diagram

- I randomly select one of you

Sample space

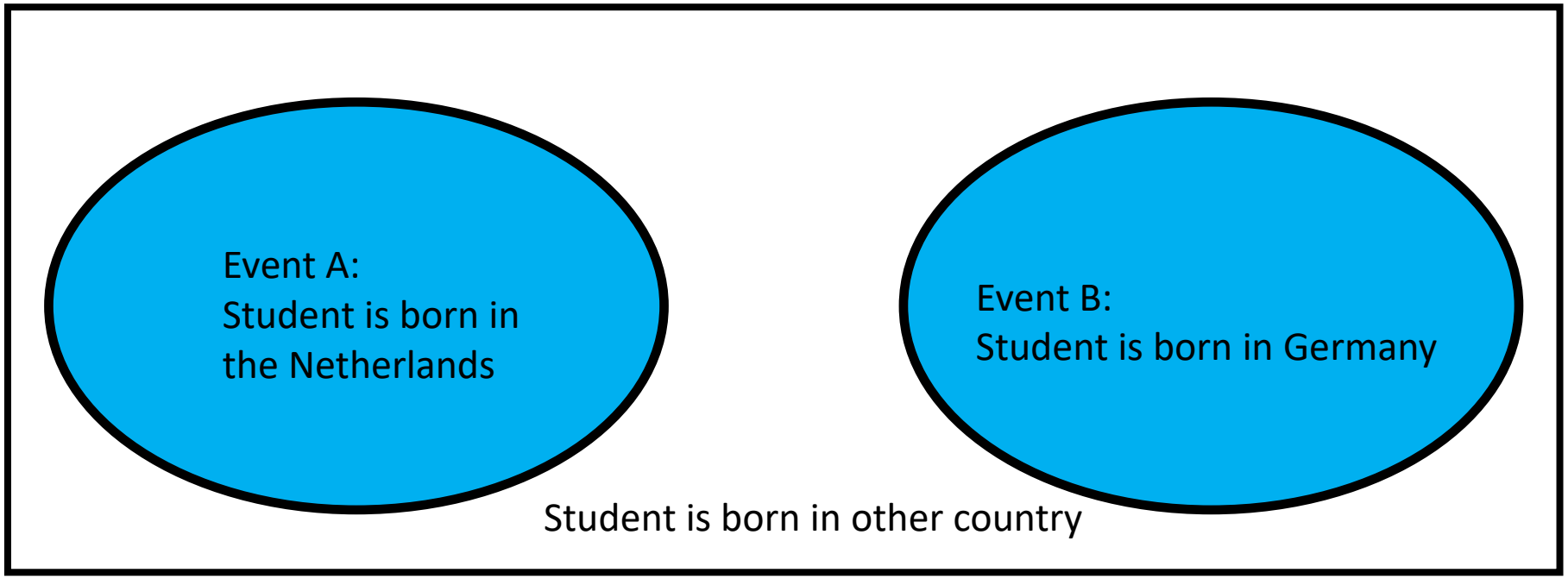


P(netherlands OR germany)

Venn diagram

What is the intersection here? **Intersection:** empty set

- I randomly select one of you



P(netherlands AND germany)

Today

What is probability?

Finding probability

-Venn Diagram

-Rules for finding probability

Rules for finding probability

- 1: Complement rule
- 2: Addition rule
  - 2A: General addition rule
  - 2B: Addition rule for disjoint events
- 3: Multiplication rule for independent events

Harvard Health Publishing
HARVARD MEDICAL SCHOOL
Trusted advice for a healthier life

HEART HEALTH | MIND & MOOD | PAIN | STAYING HEALTHY | CANCER | DI CO

Home » Harvard Health Blog » 1 in 10 Americans Depressed - Harvard Health Blog

1 in 10 Americans Depressed

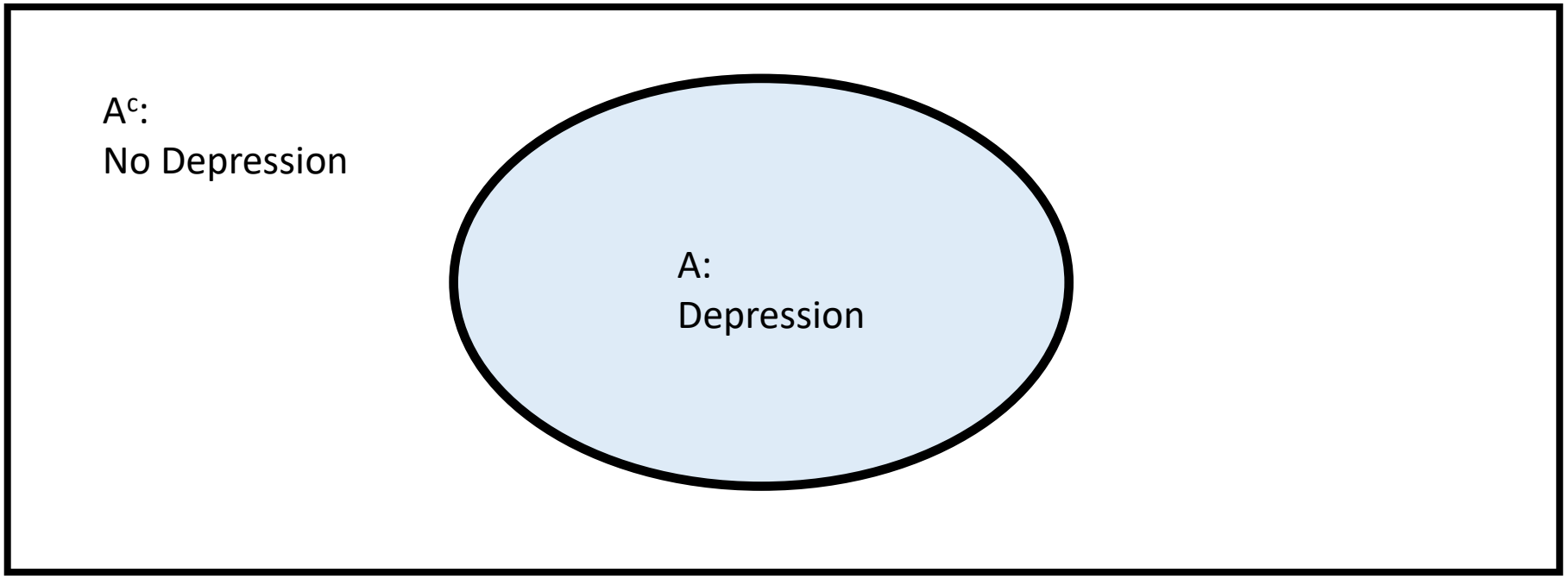
POSTED OCTOBER 02, 2010, 11:25 AM , UPDATED OCTOBER 05, 2010, 10:43 AM

Michael Craig Miller, M.D.
Senior Editor, Mental Health Publishing, Harvard Health Publishing

In time for National Depression Screening Day (October 7, 2010) and Mental Illness Awareness Week (October 3-9, 2010), the Centers for Disease Control and Prevention (CDC) published survey data on depressed mood in the United States.

This means: *if we sample a random person from the US*, the probability that this person is depressed is 0.1. We denote this P(depression)=0.1

1: Complement rule



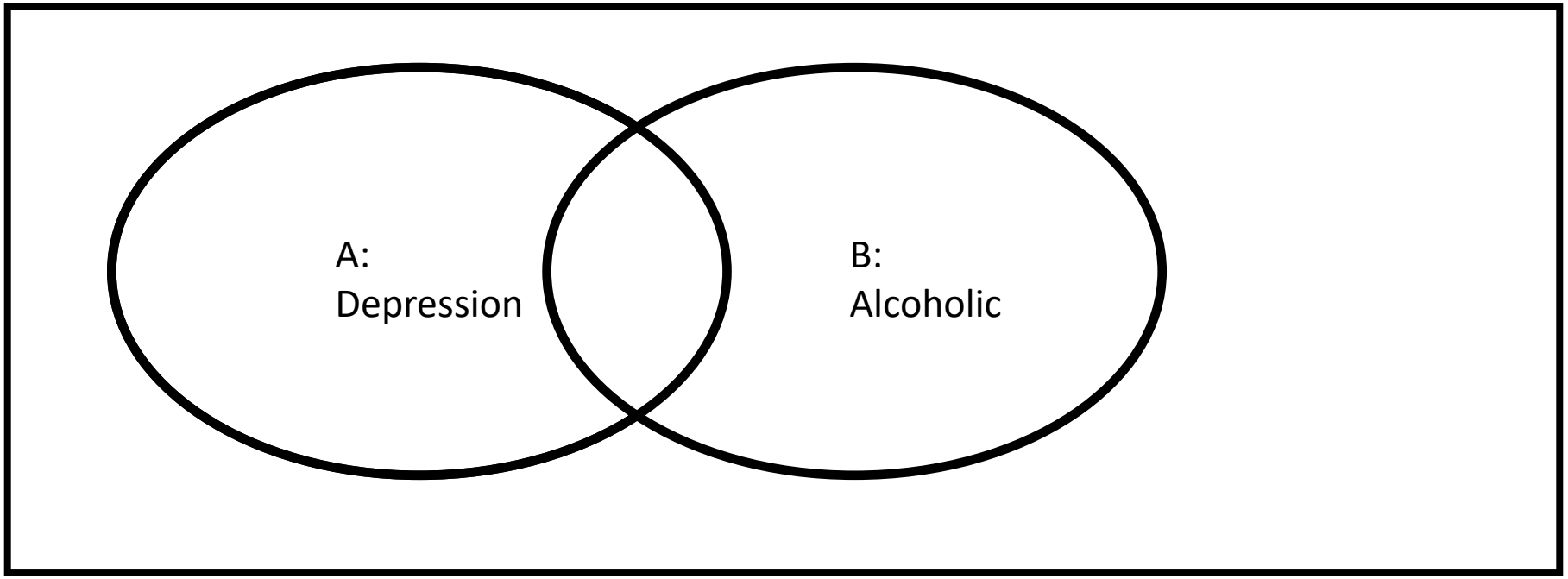
P(Depression) = P(A) = 0.10

P(No depression) = P(A^c) = 1 - P(depression) = 1 – P(A) = 1 - 0.10 = 0.90

$$P(A^c) = 1 - P(A)$$

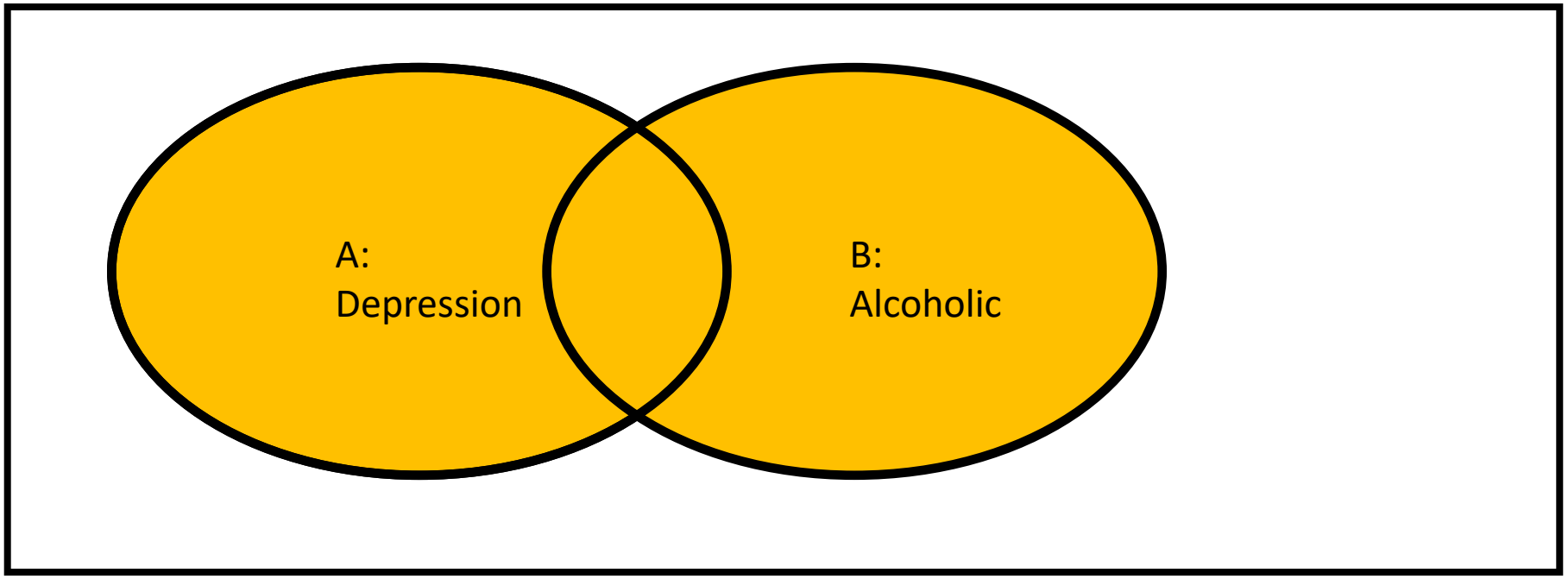
2A: General addition rule

- P(depression) = 0.10
- P(alcoholic) = 0.15
- P(depression AND alcoholic) = 0.07



2A: General addition rule

- P(depression) = 0.10
- P(alcoholic) = 0.15
- P(depression AND alcoholic) = 0.07

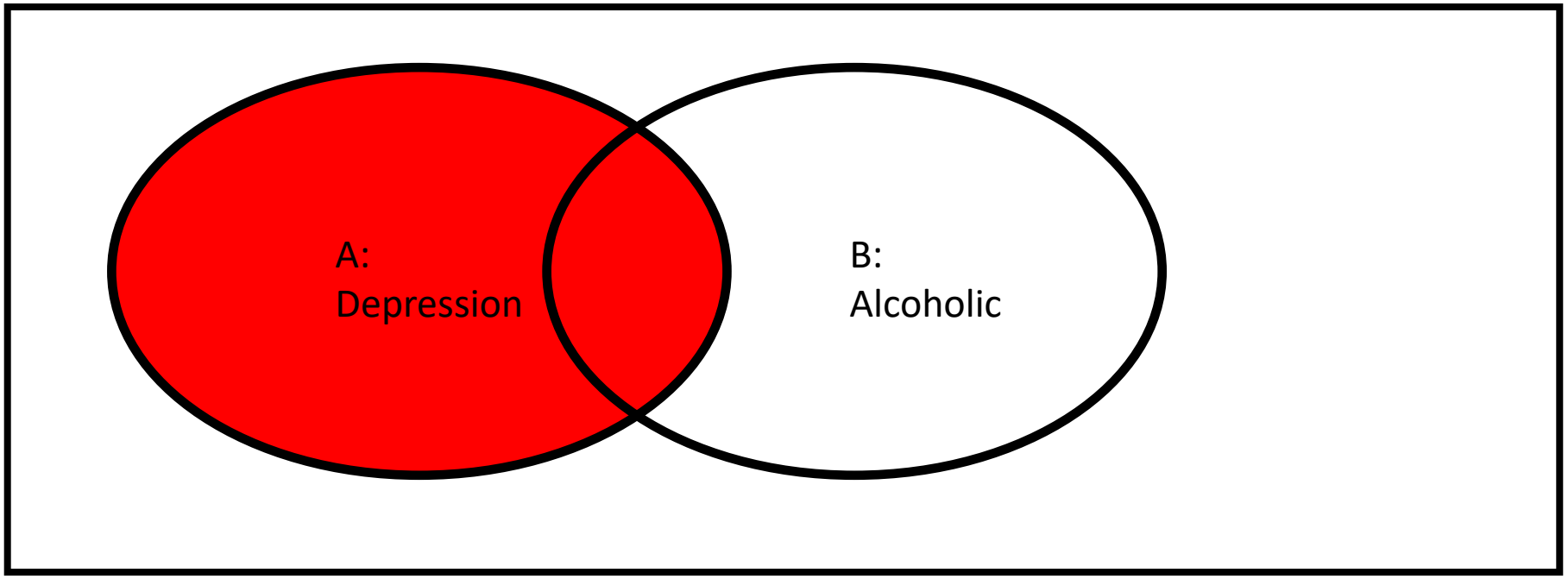


P(Depression OR alcoholic) =

i.e., “The union of the two events”

2A: General addition rule

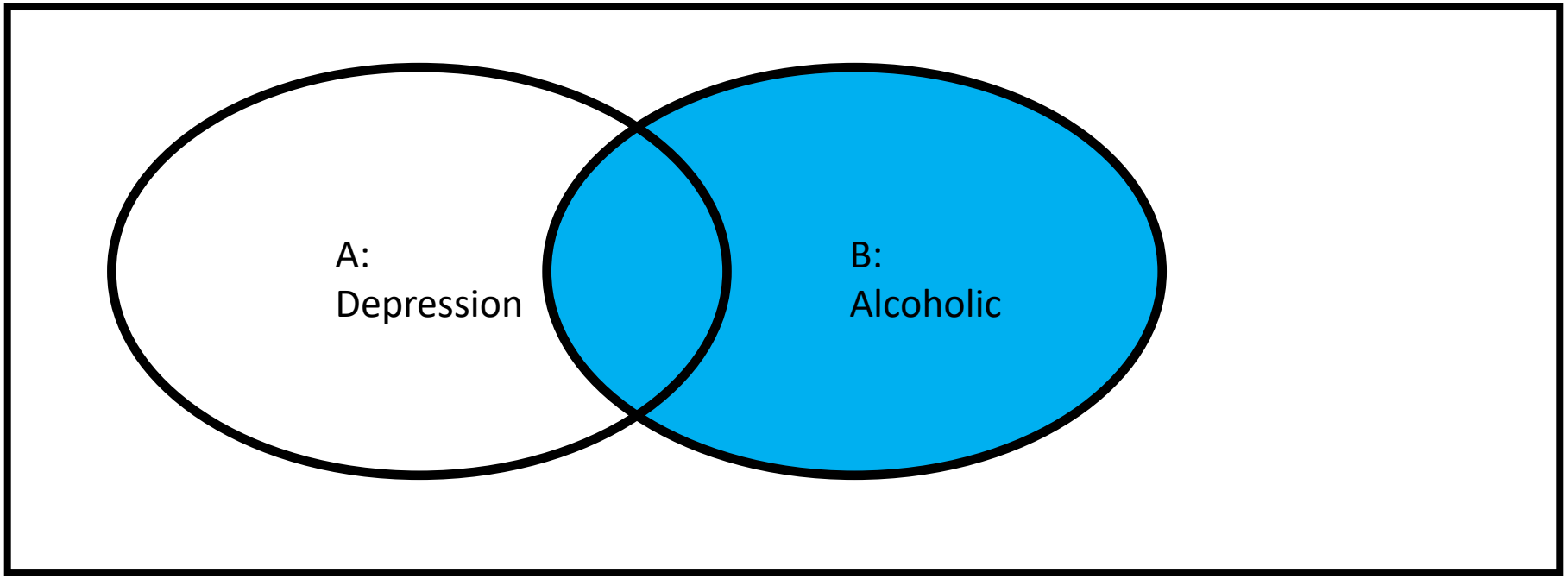
- P(depression) = 0.10
- P(alcoholic) = 0.15
- P(depression AND alcoholic) = 0.07



P(Depression OR alcoholic) = P(depression)

2A: General addition rule

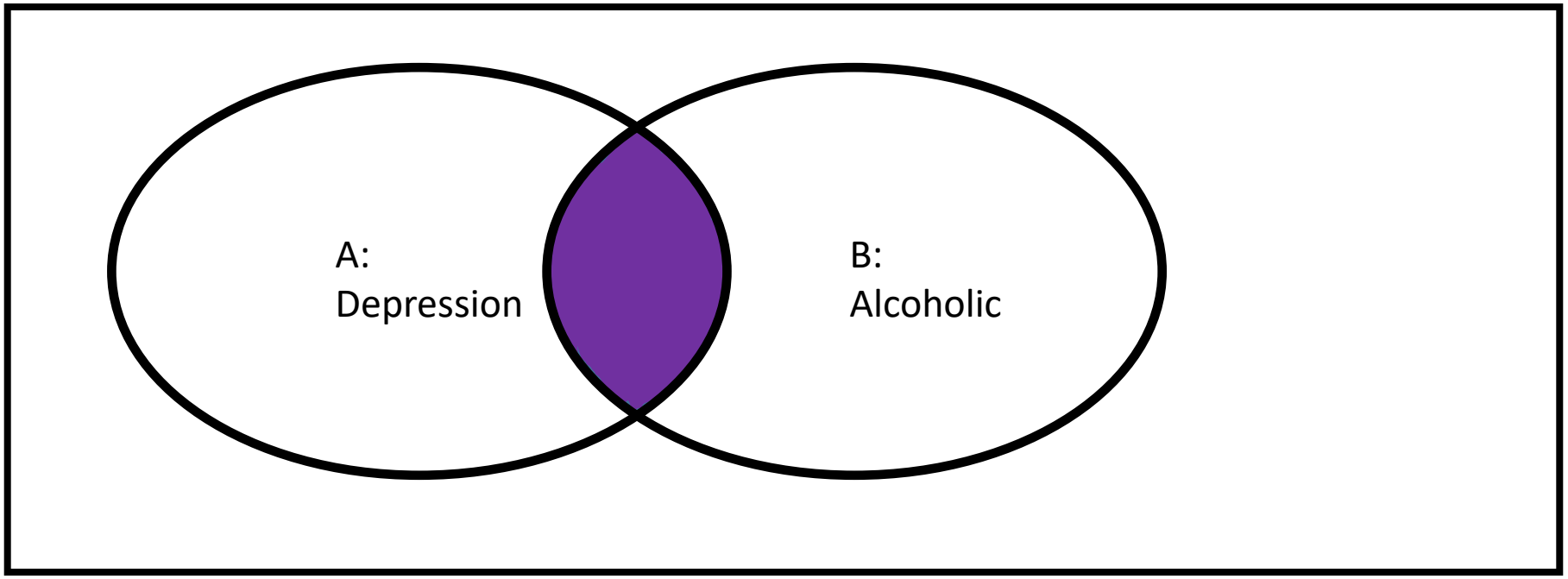
- P(depression) = 0.10
- P(alcoholic) = 0.15
- P(depression AND alcoholic) = 0.07



P(Depression OR alcoholic) = P(depression) + P(alcoholic)

2A: General addition rule

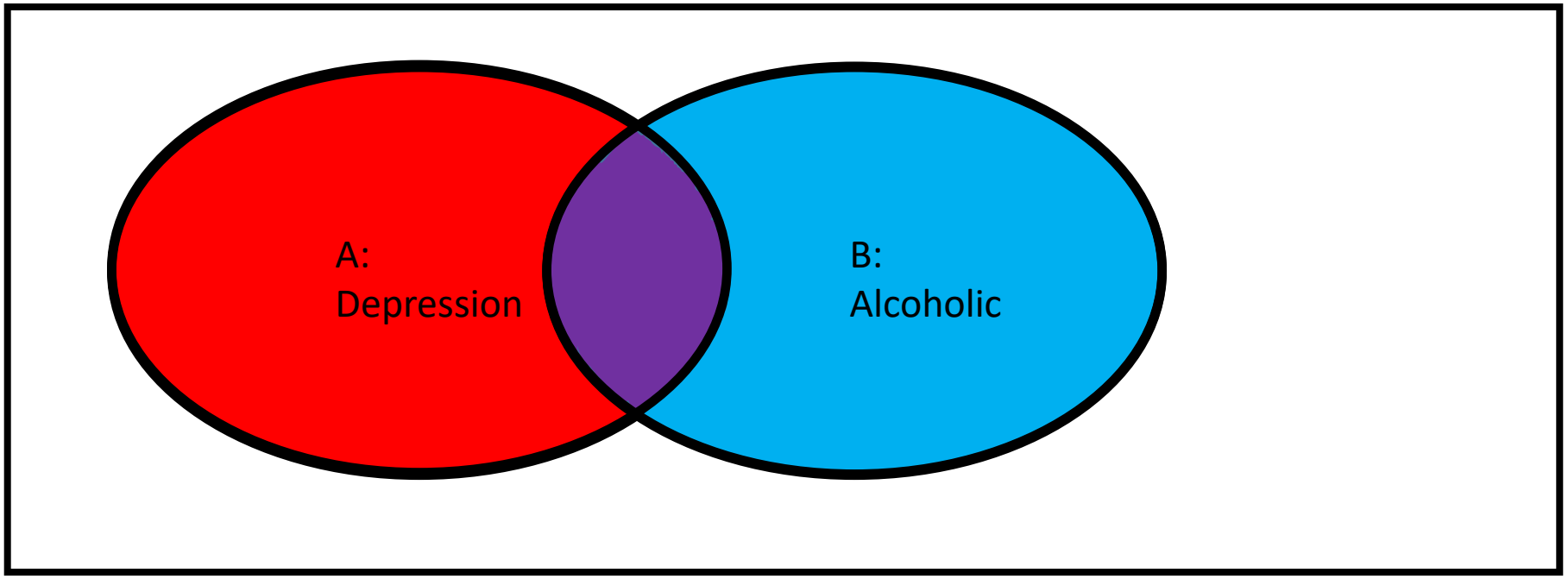
- P(depression) = 0.10
- P(alcoholic) = 0.15
- P(depression AND alcoholic) = 0.07



P(Depression OR alcoholic) = P(depression) + P(alcoholic) – P(depression AND alcoholic)

2A: General addition rule

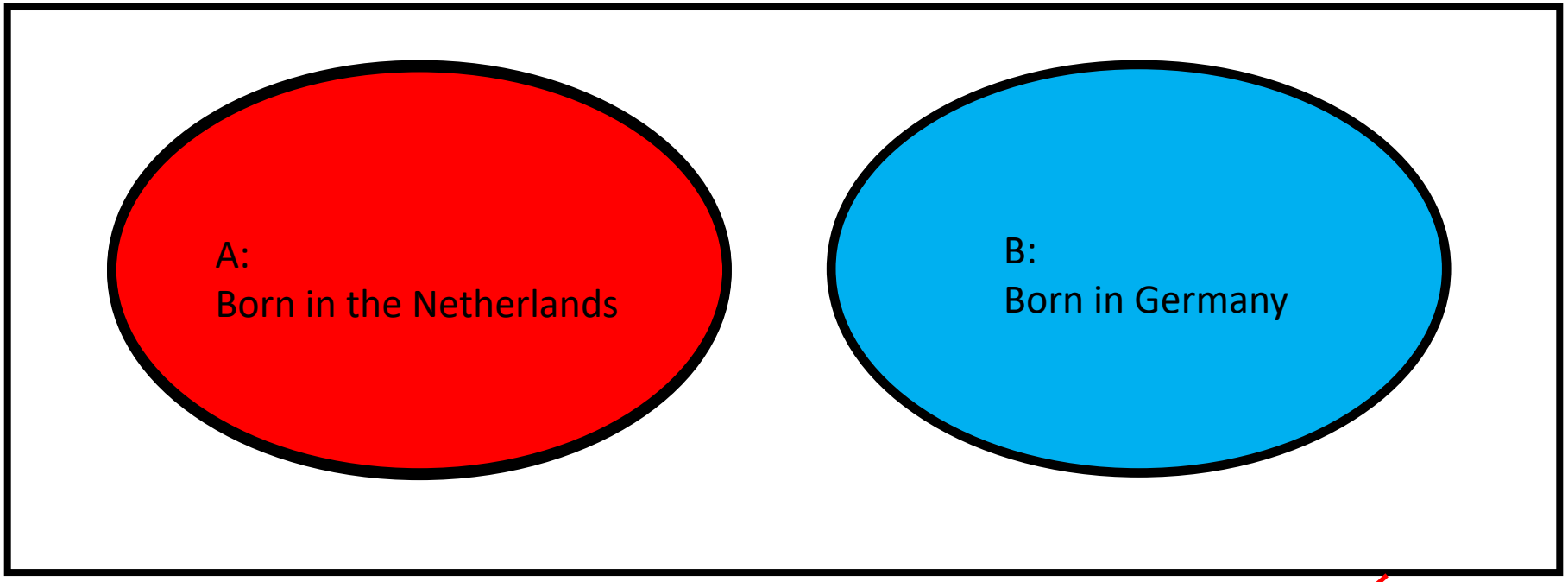
- P(depression) = 0.10
- P(alcoholic) = 0.15
- P(depression AND alcoholic) = 0.07



P(Depression OR alcoholic) = P(depression) + P(alcoholic) – P(depression AND alcoholic)
= 0.10 + 0.15 – 0.07 = 0.18

P(A or B) = P(A) + P(B) – P(A and B)

2B: Addition rule for disjoint events



P(Netherlands OR Germany) = P(Netherlands) + P(Germany) – P(Netherlands AND Germany)

P(Netherlands AND Germany) = 0

0.45 + 0.35 = 0.80

If A and B are disjoint: P(A or B) = P(A) + P(B)

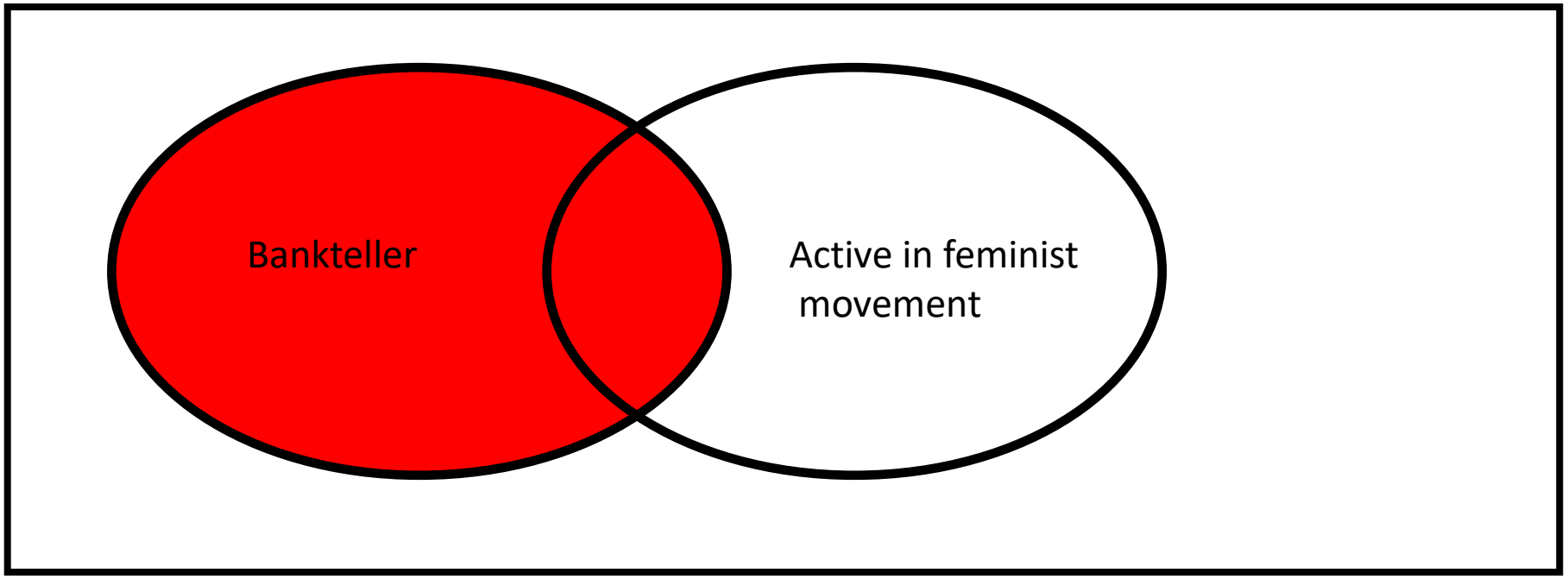
Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?

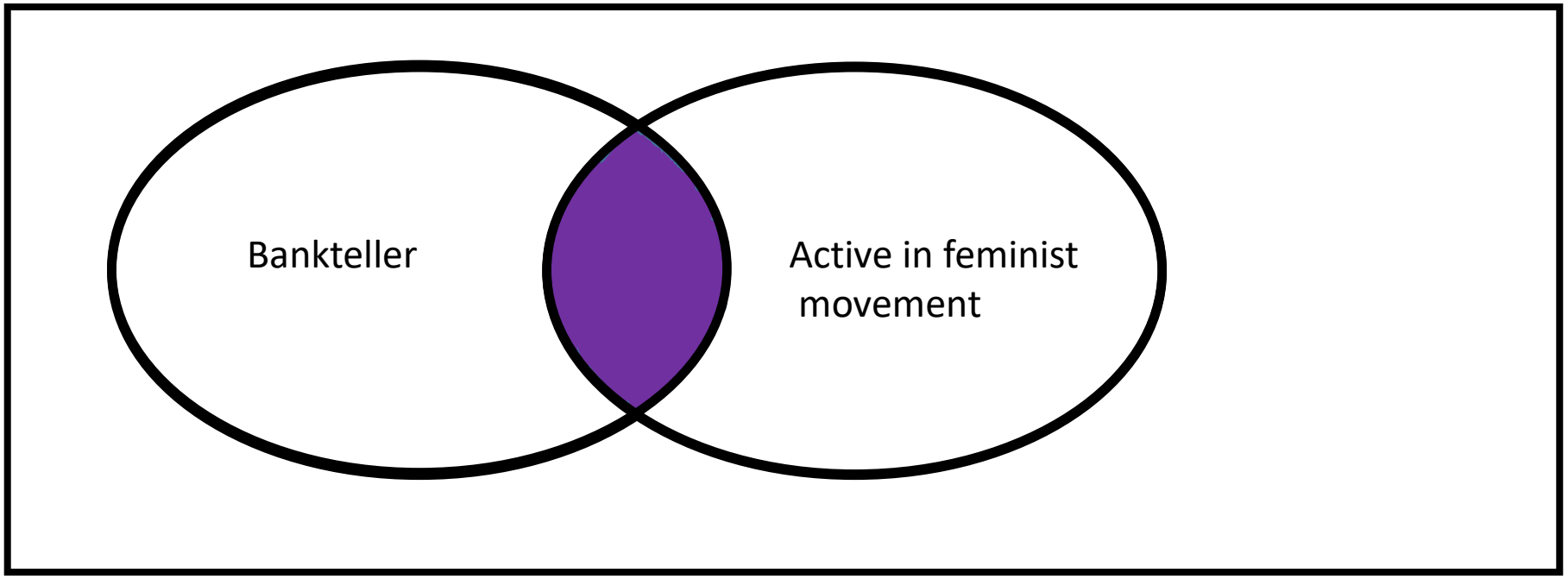
a) Linda is a bank teller.

b) Linda is a bank teller and is active in the feminist movement

Linda is a bankteller

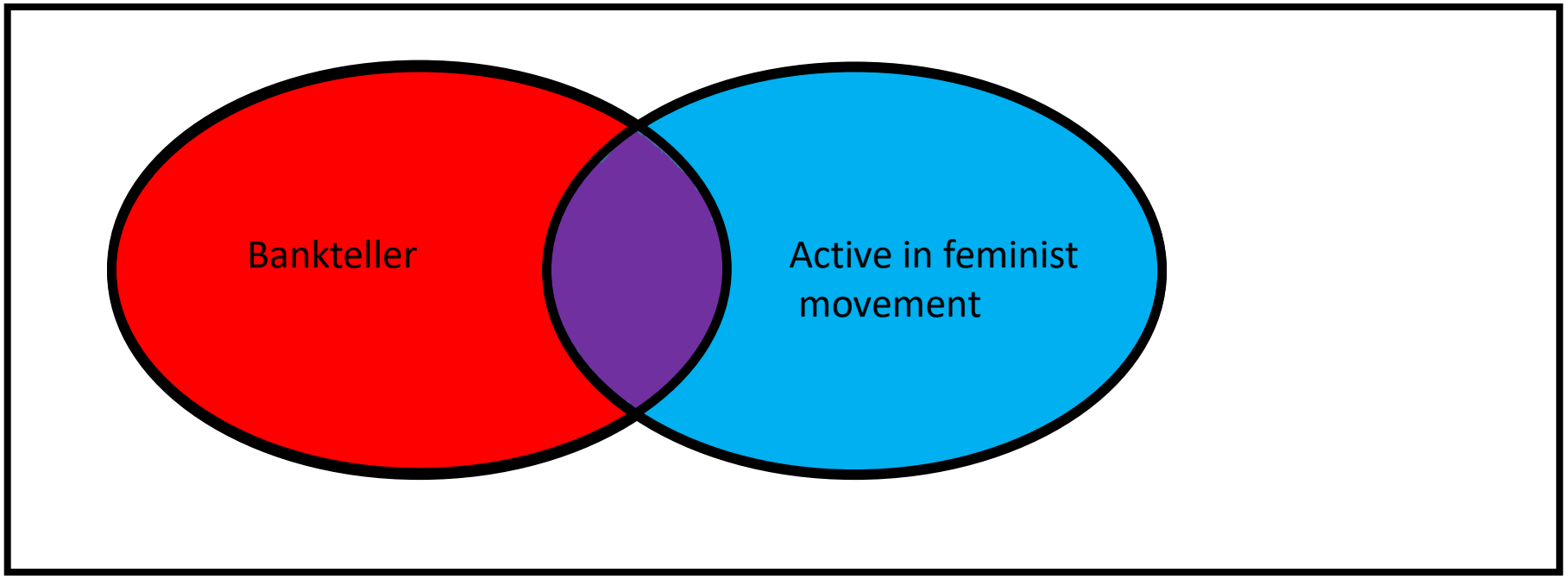


Linda is a bankteller AND feminist



P(bankteller)= 0.01
P(active in feminism) = 0.02

Probability of conjunction

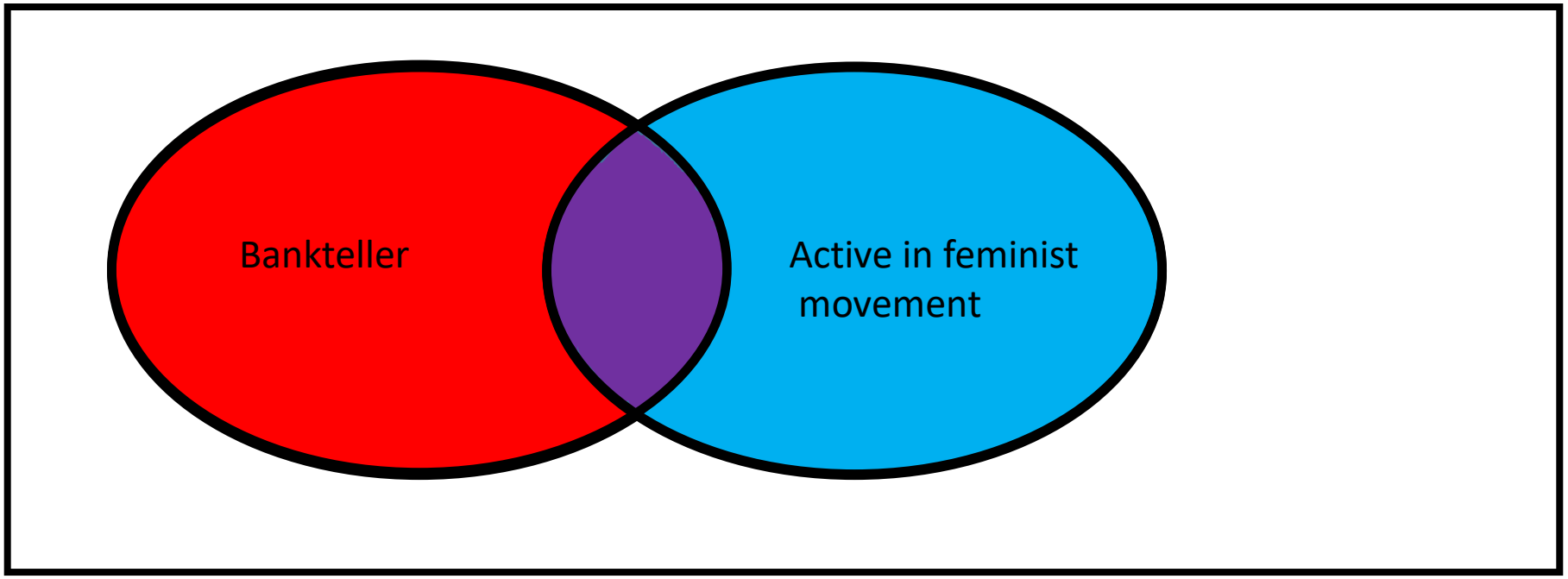


P(Bankteller AND active in feminism)

i.e., “The intersection of the two events”

3: Multiplication rule for independent events

P(bankteller)= 0.01
P(active in feminism) = 0.02



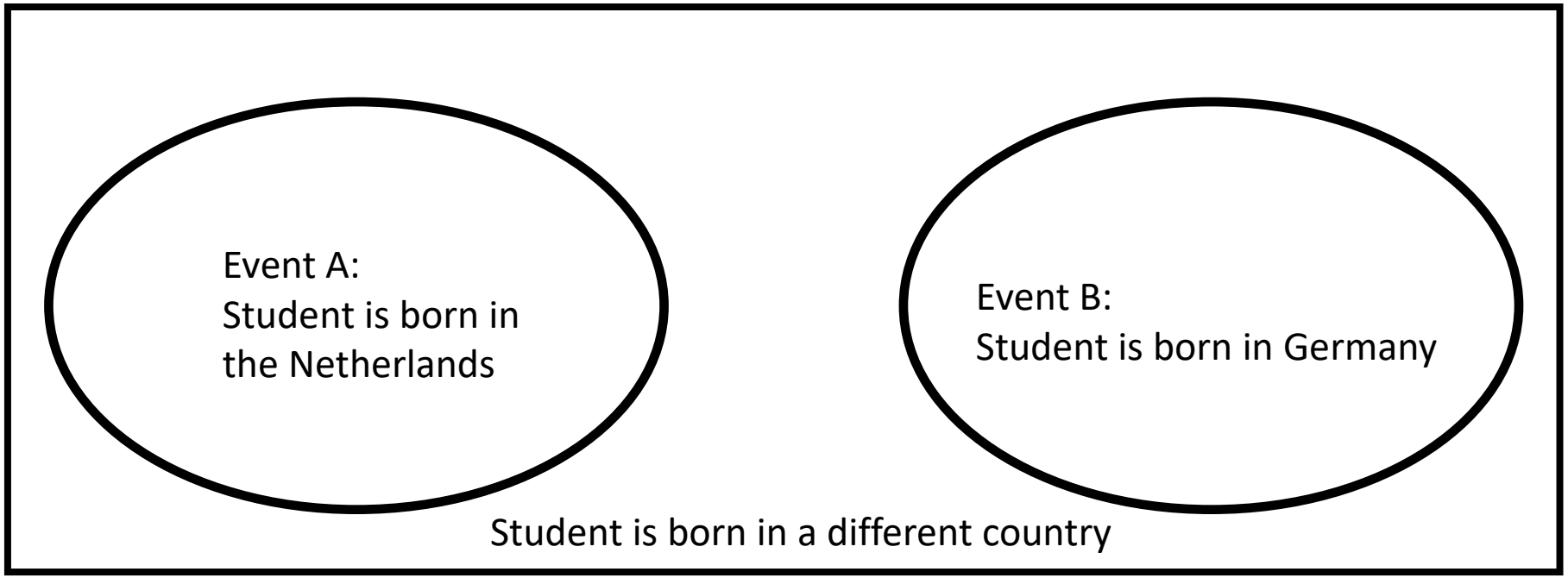
How to get the intersection?

If A and B are indepedent: P(A and B) = P(A) × P(B)

If being a bankteller is independent of being active in feminist movement:
P(Bankteller AND active in feminism) = P(bankteller) × P(active in feminism)
= 0.01 × 0.02 = 0.0002

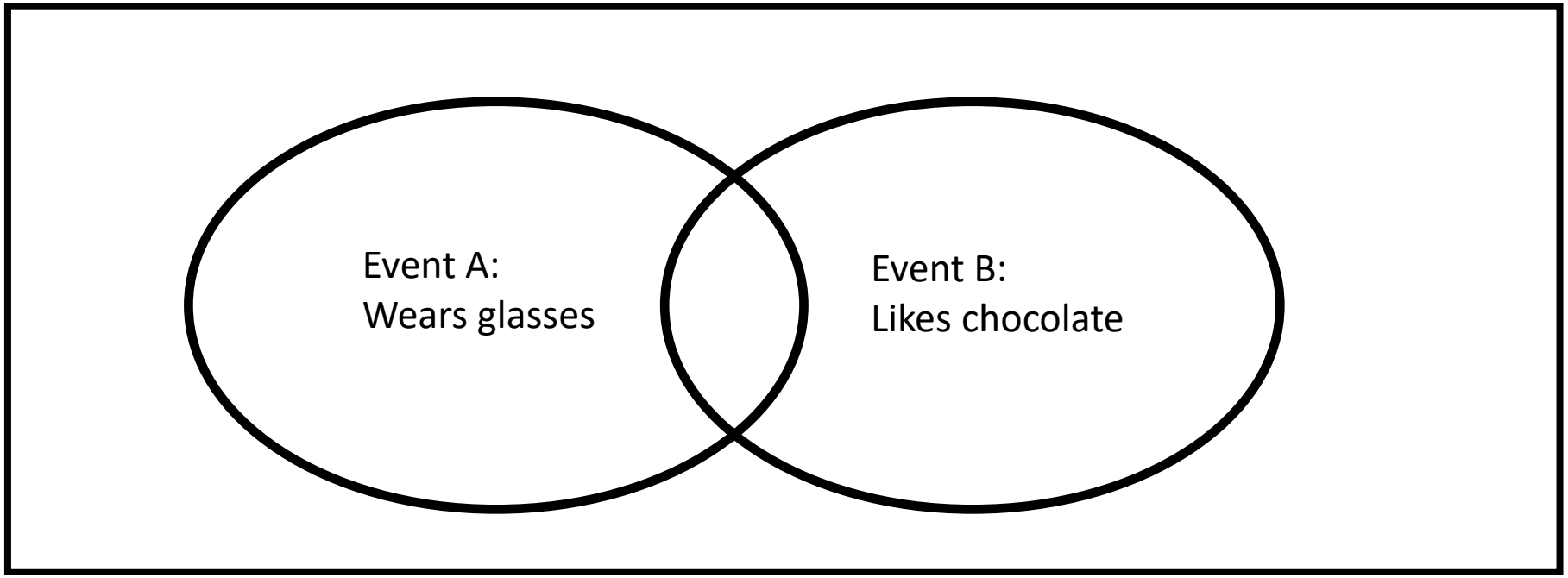
Independent events

- Two events are independent if knowing one event does not tell you anything about the other event
- Independent:
  - Wearing glasses – Liking chocolate
    - As the probabiliy of wearing glasses is the same for people who like and who dislike chocolate (and vice versa!)
  - Being an alcoholic – Having blond hair
    - As the probabiliy of having blond hair is the same for people diagnozed with alcoholism as compared to not-alcoholics (and vice versa!)
  - Coin lands heads in first toss – coin lands heads in second toss
    - The outcome of the one coin flip is independent of the other coin flip
- Dependent:
  - Being diagnosed with depression – being diagnosed with anxiety
    - Major depression and anxiety are comorbid, meaning that people being diagnosed with the one have a higher probability of being diagnosed with the other
  - Living in Amsterdam – Having a bike
    - As the probability of having a bike is larger in Amsterdam as compared to other places.



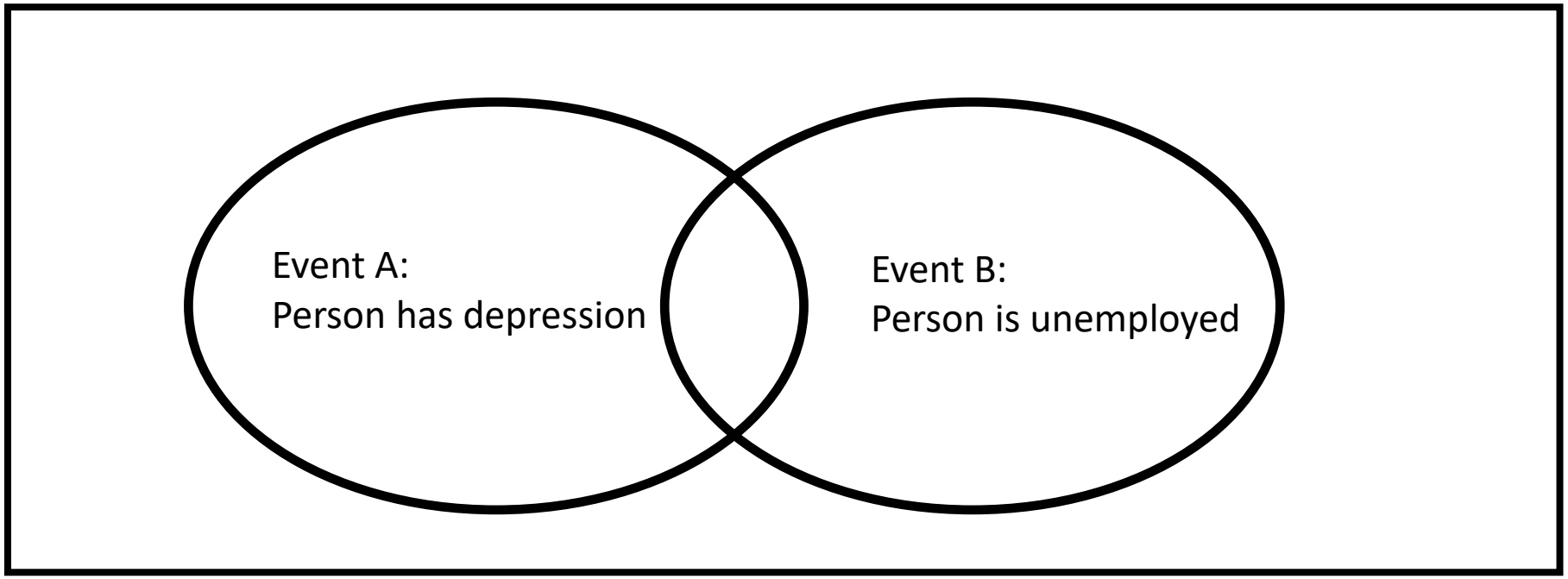
Independent?
No! → If I know that a student is born in Germany, they cannot be born in the Netherlands

For independence, ask yourself: does knowing about the one event give me any information on the probability (whether lower or higher) of the other event? Yes? → dependent



Independent?
Yes! → If I know that a student wears glasses, this does not tell me anything about whether the person will like chocolate

Independent events



Independent?

No! → Unemployed people are more often depressed than employed people

Independent events

Conclusion: From the Venn diagram, you cannot draw conclusions about independence

Except if the two events are disjoint: then you know that they are *dependent*

Rules for finding probability

1: Complement rule: $P(A^c) = 1 - P(A)$

2A: General addition rule: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

2B: Addition rule for disjoint events: $P(A \text{ or } B) = P(A) + P(B)$

3: Product rule for independent events: $P(A \text{ and } B) = P(A) \times P(B)$

Example exam question

In a nursing home, 70% of the inhabitants is female, and 25% of the inbitants has Alzheimer's disease. Assume that these variables are independent. What is the probability that a random inhabitant in this nursing home is a female ánd has Alzeimher's disease.

A. 0.125
B. 0.175
C. 0.95

Example exam question

In a nursing home, 70% of the inhabitants is female, and 25% of the inbitants has Alzheimer's disease. Assume that these variables are independent. What is the probability that a random inhabitant in this nursing home is a female ánd has Alzeimher's disease.

A. 0.125
B. 0.175
C. 0.95

Use the product rule for independent events:
0.7*0.25=0.175

Example exam question II

In a nursing home, 70% of the inhabitants is female, and 25% of the inbitants has Alzheimer's disease. Assume that these variables are independent. What is the probability that a random inhabitant in this nursing home is a female *or* has Alzeimher's disease.

A. 0.775
B. 0.125
C. 0.95

Rules for finding probability

1: Complement rule: $P(A^c) = 1 - P(A)$

2A: General addition rule: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

2B: Addition rule for disjoint events: $P(A \text{ or } B) = P(A) + P(B)$

3: Product rule for independent events: $P(A \text{ and } B) = P(A) \times P(B)$

Example exam question II

In a nursing home, 70% of the inhabitants is female, and 25% of the inbitants has Alzheimer's disease. Assume that these variables are independent. What is the probability that a random inhabitant in this nursing home is a female *or* has Alzeimher's disease.

A. 0.775

B. 0.125

C. 0.95

Use the general addition rule:
0.7+0.25-0.175=0.775

Example of question from book

- 5.16 on p.228:

Your teacher gives a true-false pop quiz with 10 questions.

a) Show that the number of possible outcomes for the sample space of possible sequences of 10 answers is 1024.

b) What is the complement of the event of getting *at least* one of the questions wrong?

c) With random quessing, show that the probability of getting *at least* one question wrong is 0.999.

Answer on next slide.

Answer to 5.16

a) The number of outcomes in the sample space is 2^10 because for each question it is possible to have True or False (2 outcomes) and there are 10 questions. 2^10=2024

b) The complement of getting at least one question wrong is getting no question wrong. After all, 1 or more questions wrong is ‘at least one question wrong’, so only ‘0 questions wrong’ is not ‘at least one questions wrong’.

c) With random guessing the probability correct for each question is 0.5. The probability of ‘0 questions wrong’ (or ‘all questions correct’) is easy to calculate because there is only one combination/sequence (CCCCCCCCCC), and so the probability is 1* 0.5^10=0.001. ‘At least one question wrong’ is the complement of ‘0 questions wrong’ and so you get its probability by subtracting it from 1: 1 - 0.001=0.999.

Example of question from book

- Excercise 5.26

You are the marketing director for a museum that raises money by selling gift items from a mail-order catalog. For each catalog sent to a potential customer, the customer's entry in the data file is Y if they ordered something and N if they did not (Y=yes, N=no). After you have mailed the fall and winter catalogs, you estimate the probabilities of the buying patterns based on those who received the catalog as follows:

Outcome (fall ,winter)	YY	YN	NY	NN
Probability	0.3	0.1	0.05	0.55

a) Display the outcomes and their probabilities in a contingency table, using the rows for the (Y,N) outcomes for the fall catalof and the columns for the (Y,N) outcomes for the winter catalog.

See next slide for questiion b, c and d.

Example of question from book

Outcome (fall ,winter)	YY	YN	NY	NN
Probability	0.3	0.1	0.05	0.55

b) Let F denote buying from the fall catalog and W denote buying from the winter catalog. Find P(F) and P(W).

c) Explain what the event “F and W” means, and find P(F and W).

d) Are F and W independent events? Explain why you would not normally expect customer choices to be independent.

Answers next slides.

Answer to 5.26

A)

		Winter	
		Y	N
Fall	Y	0.3	0.1
	N	0.05	0.55

B)
P(F) = P(YY) + P(YN) = 0.3 + 0.1 = 0.4
P(W) = P(YY) + P(NY) = 0.3 + 0.05 = 0.35

C) The event “F and W” is the event in which people both order the fall catalog and also the winter catalog. It is the intersection of the event ‘F’ and the event ‘W’. P(F and W) is in this case P(YY) which is 0.3.

For answer D see next slide

Answer to 5.26

D) F and W are independent if P(F and W) = P(F) x P(W) (see multiplication rule p.225). Here P(F) x P(W) = 0.4 * 0.35 = 0.14 which is smaller than P(F and W)=0.3. They are therefore not independent. The fact that P(F and W) is larger than the multiplication of P(F) x P(W) indicates that there is a dependency such that people who buy the one catalog are more likely to also buy the other catalog. This is something that can be expected since people who are interested in the catalog at all will likely buy both, whereas people who are not interested buy neither.

Generally customer choices can be expected to be dependent. Either because people who buy it once will buy it again. Or the other way around that if they already have one they will NOT buy it again, which also induces a dependency.

Practice with excercises!

- Some other excercises to focus on: 5.1, 5.6, 5.13, 5.15, 5.16, 5.19, 5.20, 5.23
- Also work through the examples (e.g., example 8 in 5.3)
- In doubt about how to get to the correct answer to an excercise? You can post questions on the discussion board.
- On the discussion board you can help each other! We will also check the discussion board to help out. (but not every day, so don’t start posting one day before the exam!)