Computational Methods for Linguists

Ling 471

Olga Zamaraeva (Instructor)
Yuanhe Tian (TA)
05/06/21

Reminders

- Assignment 3 due today
  - ...how are people doing?
- Blog responses due today
- Assignment 4 will be published soon
  - I will send out an additional announcement
  - due date moved to May 25

May 20	Working with linguistic corpora	TBA	
May 25	Visualization and Communication	TBA	Assignment 4
May 27	Visualization and Communication	TBA	Blogs 5
June 1	Presentations		
June 3	Presentations		
June 8			Assignment 5

From class syllabus

Corrections

Thank you!

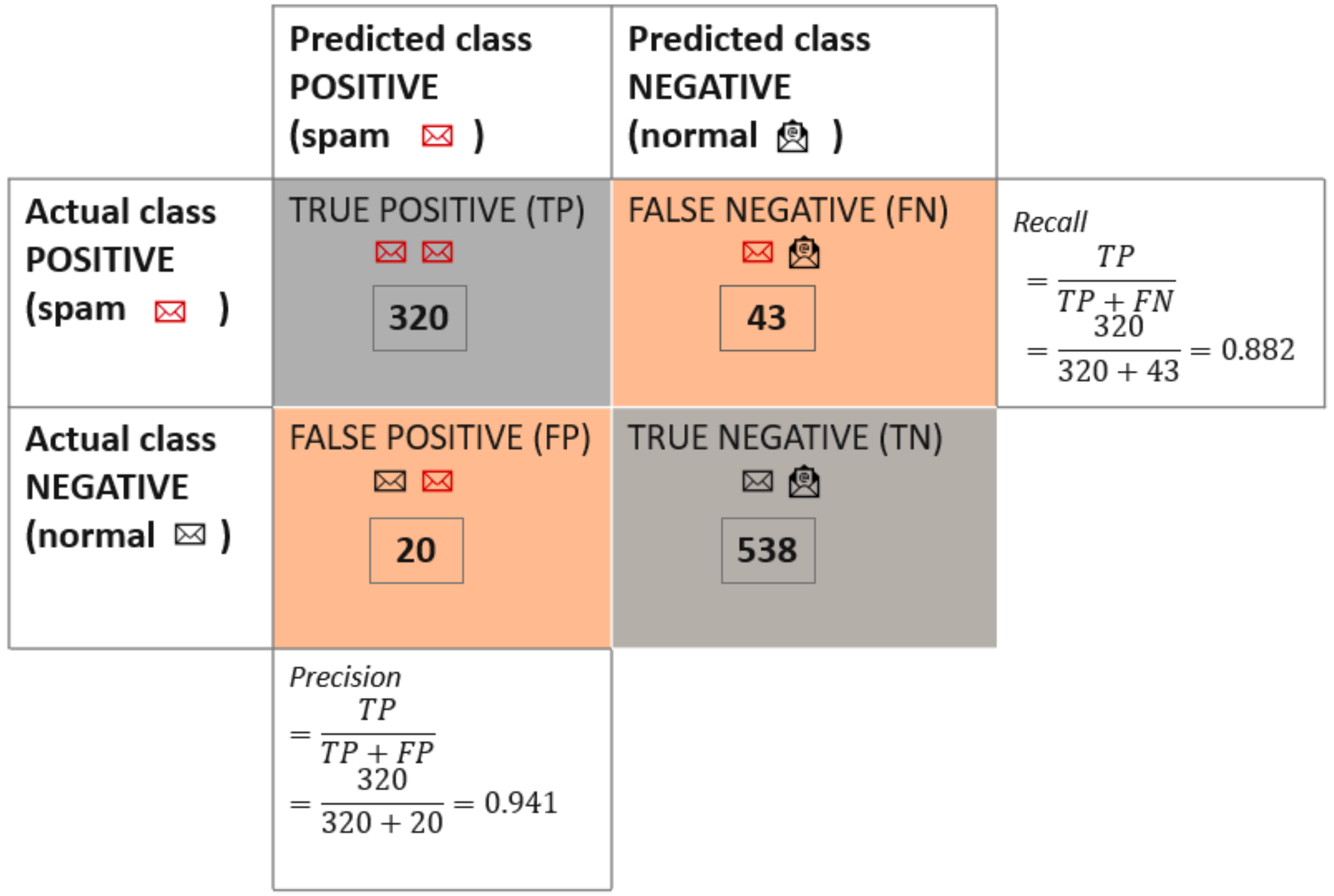
- **Age** is of course **not** a Gaussian
  - Thanks for doubting!
  - You can imagine situations where it will be but it is very different from e.g. height
  - Other examples of actually Gaussian stuff:
    - amount of hair on people’s heads
    - weight
    - age when children acquire syntax
- “**Discrete**” variable is not spelled “discreet” :)

Plan for today

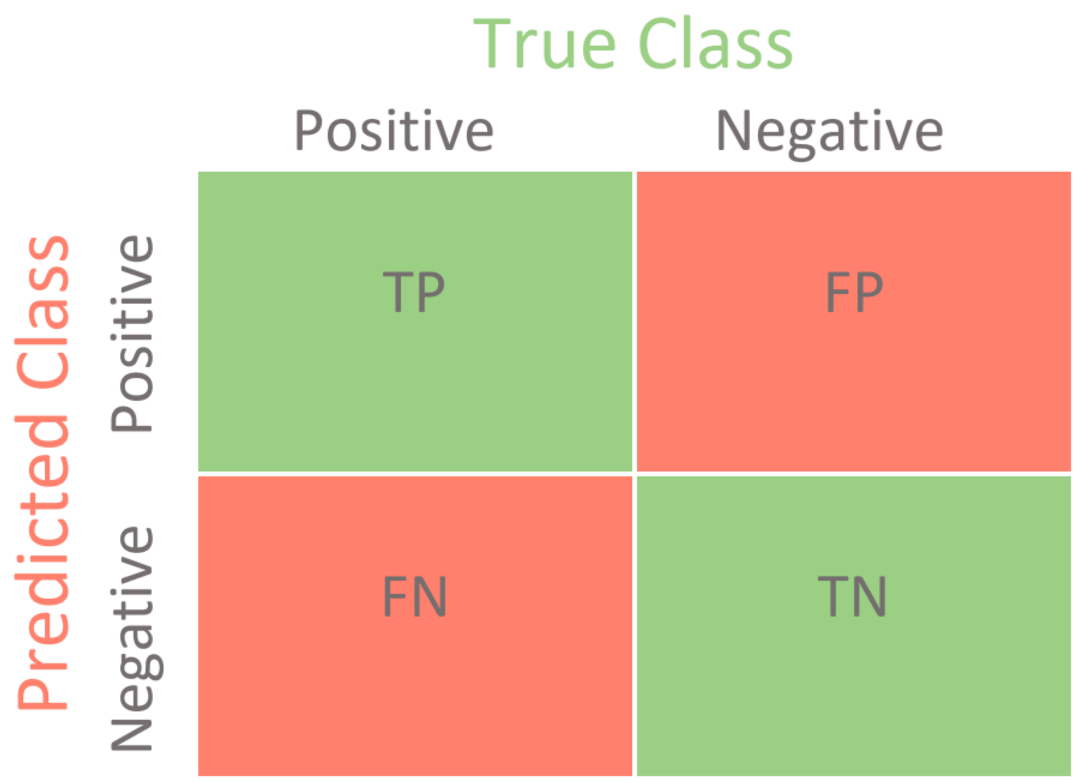
- Precision and Recall review
- Theory:
  - The Bayes Theorem
    - Activity
  - Next week: Naive Bayes classification algorithm
- Practice:
  - Packages:
    - pip
    - pandas and dataframes

Precision and Recall
review

- Context: ~~Object~~ apple 🍎 **retreival**
  - Array of objects: [0, 1, 2,3, 4,5, 6, 7]
  - Ground Truth: [🍎🍎🍎🍎🍊🍊🍊🍊]
  - Our System: [🍎🍊🍎🍎🍊🍊🍎🍊]
- Reference table for the **four types of label**
- **True Positive:** 0,2,3
- **False Positive:** 6
- **True Negative:** 4,5,7
- **False Negative:** 1
- Compute Precision and Recall as per **definitions**

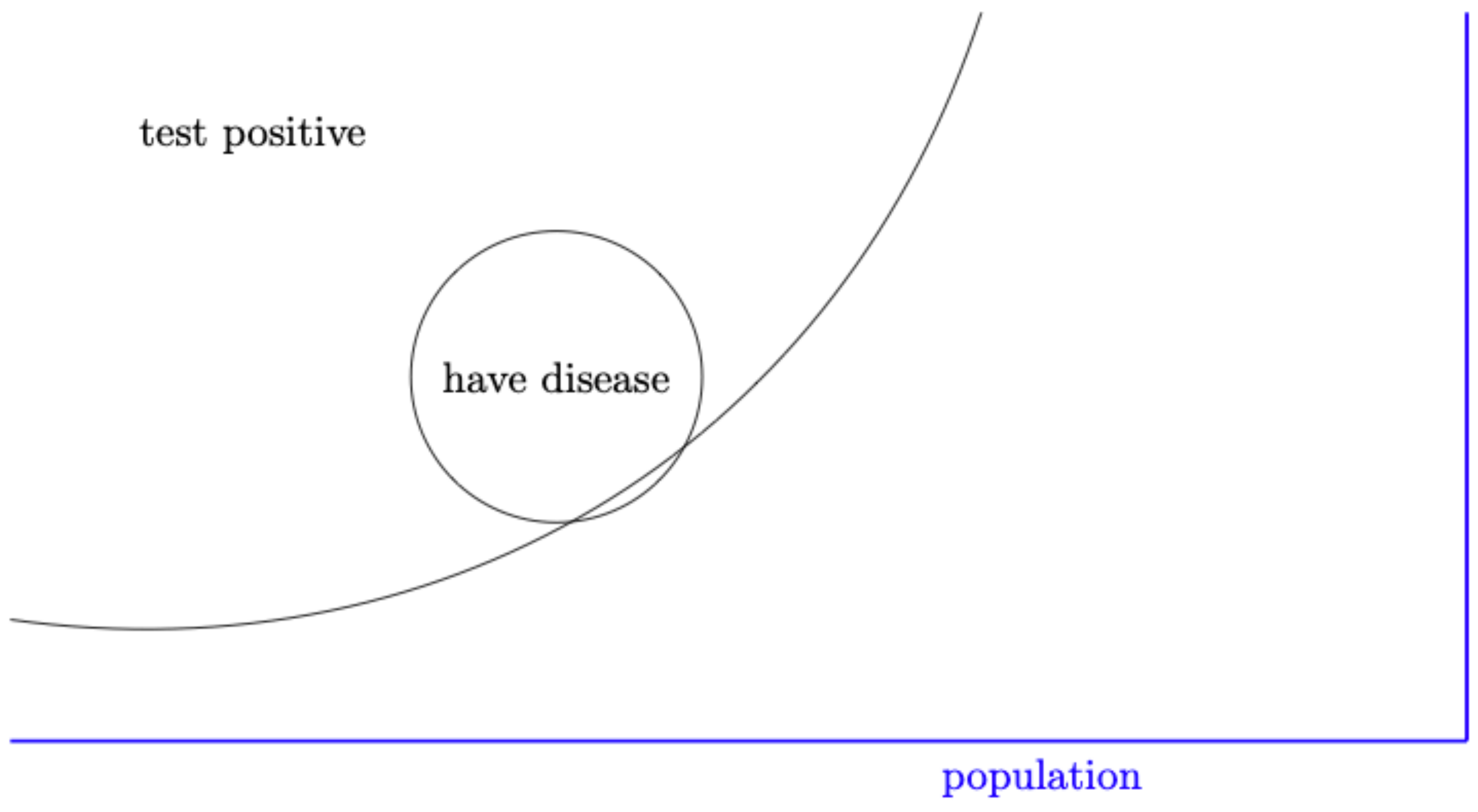


https://www.knime.com/blog/from-modeling-to-scoring-confusion-matrix-and-class-statistics



https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

A classic example
(teaser)



https://towardsdatascience.com/3-ways-to-think-about-bayes-rule-b6f5b4ef87d6

- Suppose:
  - 1% of population have cancer
  - 80% of tests detect it correctly while 20% of tests fail to detect it (“false negative”)
  - 9.6% of tests detect it when it is not there (“false positive”) while 90.4% correctly return negative
- Q: If you get a positive result, what is the probability of you having the disease?
  - Many people say “80%”
  - ...but that is not so:
    - the event of “testing” is separate from the event of “having the disease”!
    - they have different probabilities!
    - Stay tuned.

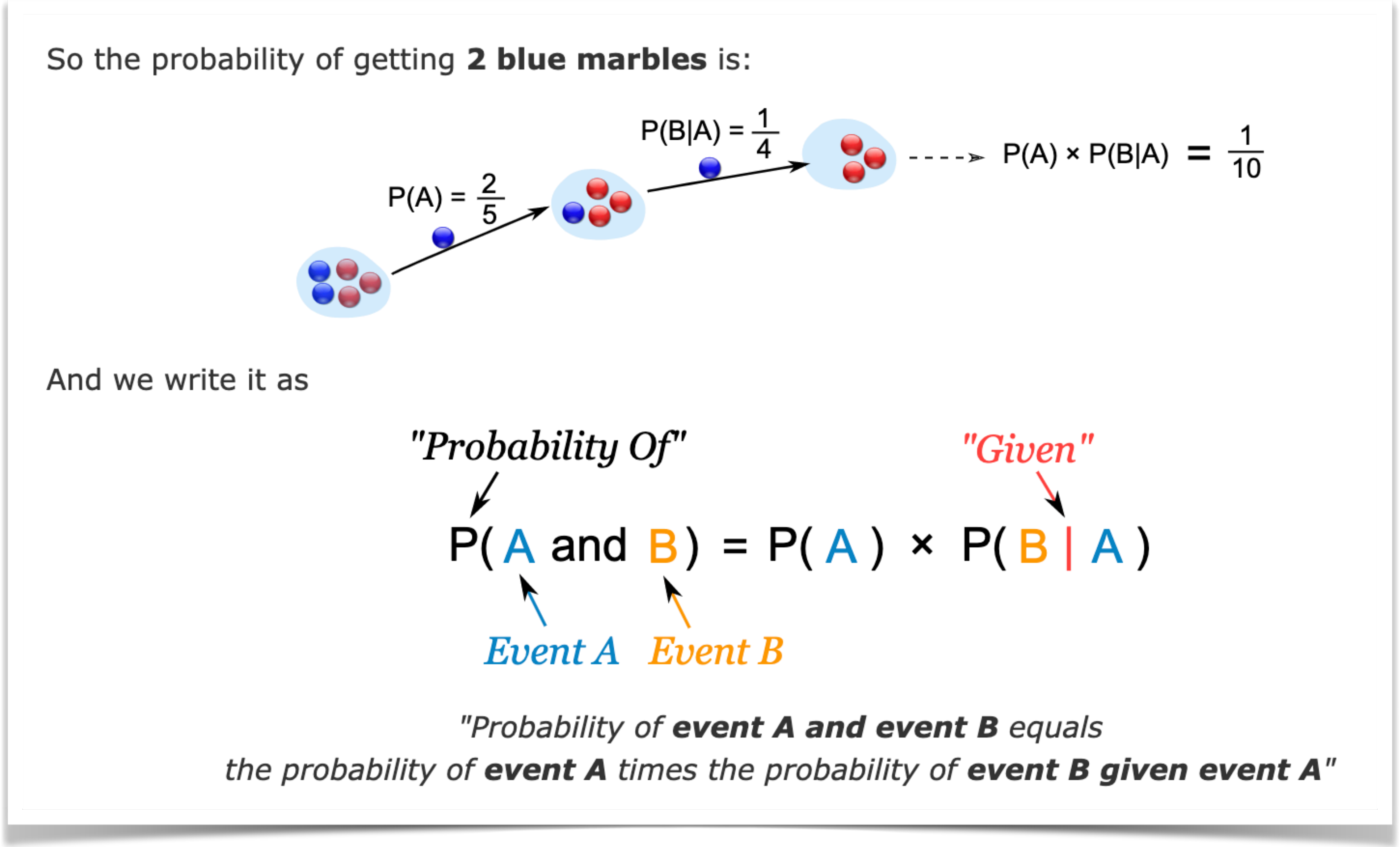
Bayes Theorem

Bayes Theorem
in probability theory

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

The Bayes Theorem

- Recall:
  - Conditional probability
    - $P(B|A) = \dfrac{P(A \cap B)}{P(A)}$
    - notation: $A \cap B$ = "A and B" both occurred
    - "Intuition":
      - How many times I saw A after I also saw B?
      - Derive the formula from the marble example
      - Sequence of A and B => product of P(A) and P(B|A)
      - ...then just rewrite the equation to express P(B|A) in terms of P(A and B) and P(A)
- By the way:
  - In a sequence of two marble draws, what's P(second marble is blue)?
    - call it **P(A)**
    - P(A) = P(second is blue)***P(first is blue) +** P(second is blue)***P(first is red)**
    - **The first marble is there!**



Conditional probability
https://www.mathsisfun.com/data/probability-events-conditional.html

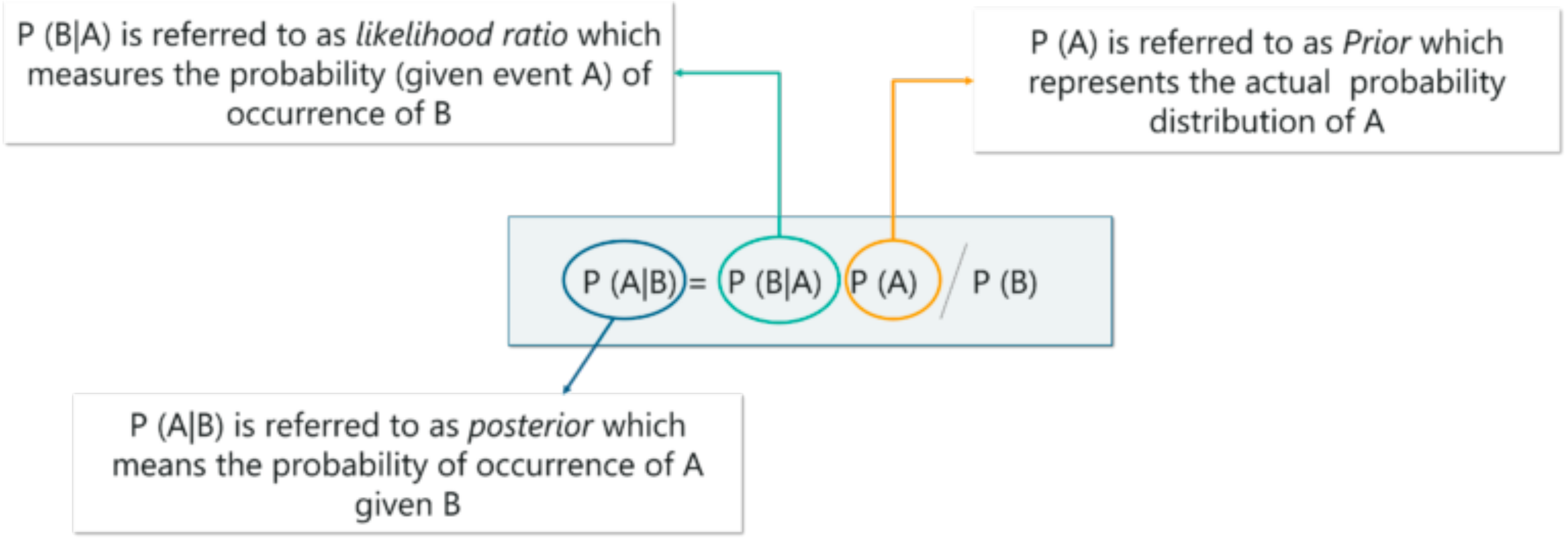
Bayes Theorem

in probability theory

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

The Bayes Theorem

- The Bayes Theorem is related to **conditional probability** and to **sequences**
- Intution:
  - “**Bayes rule** provides us with a way to update our **beliefs** based on the arrival of new, relevant pieces of **evidence**.” (Devin Soni)



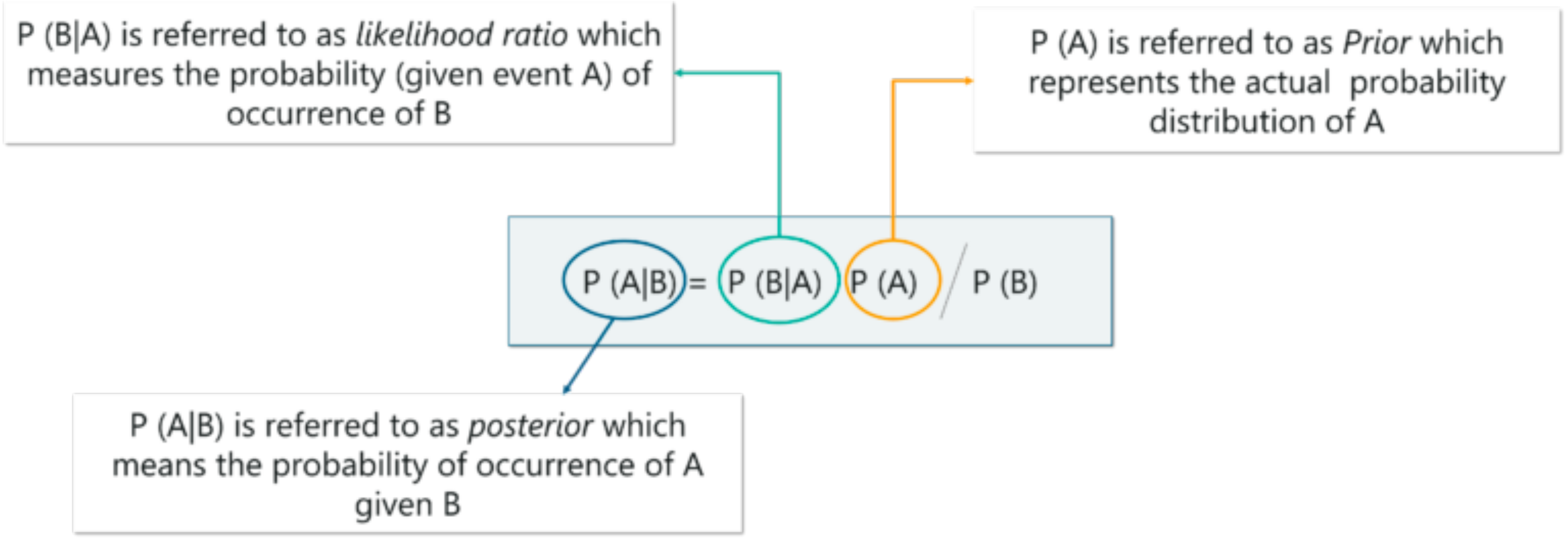
https://www.edureka.co/blog/statistics-and-probability/#Bayes%20Theorem

Bayes Theorem
derivation

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

The Bayes Theorem

- Note:
  - P(A,B) = P(A|B)*P(B)
- Note:
  - P(B,A) = P(B|A)*P(A)
- Note:
  - P(A,B) is the same as P(B,A)!
- Therefore:
  - P(A|B)*P(B) = P(B|A)*P(A)
- Therefore:
  - P(A|B) = P(B|A)*P(A) divided by P(B)!
- Why is this derivation meaningful/interesting?
  - Sometimes, we **know** P(B|A) but **not** P(A|B)!

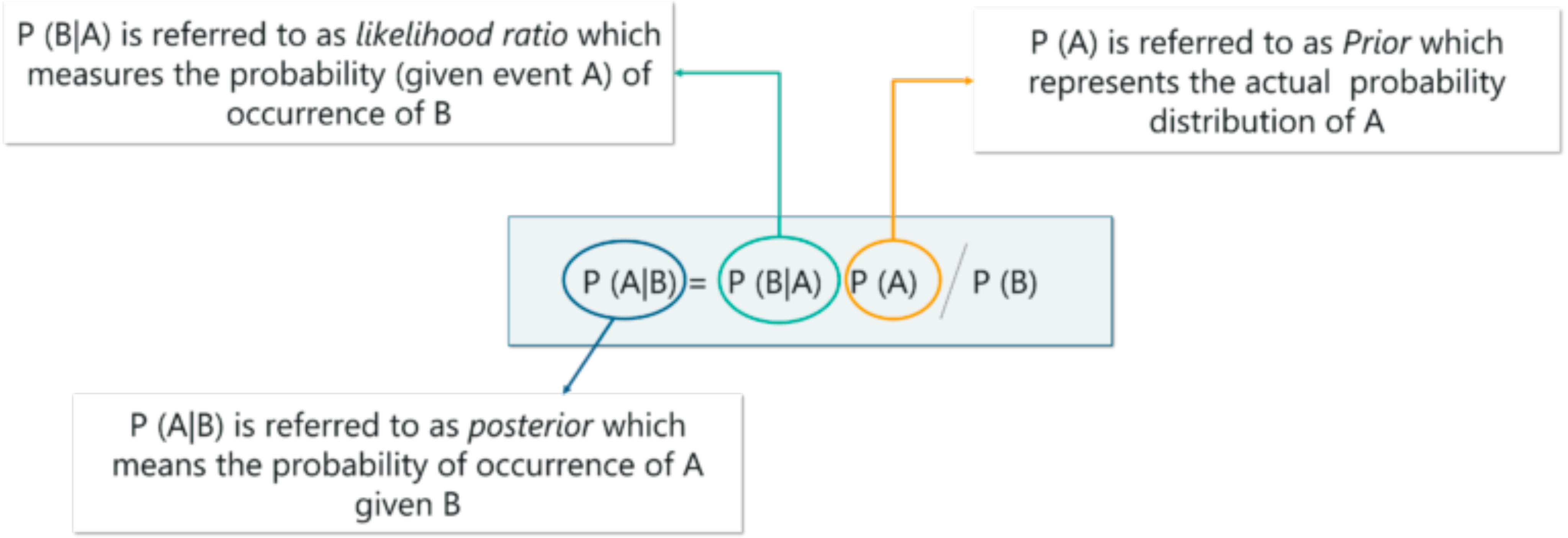


https://www.edureka.co/blog/statistics-and-probability/#Bayes%20Theorem

Bayes Theorem
an example

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

The Bayes Theorem



https://www.edureka.co/blog/statistics-and-probability/#Bayes%20Theorem

- Suppose:
- P(having cancer) = 0.05
    - (5% of people have it)
    - P(A)
- P(be a smoker) = 0.10
    - (10% of people smoke)
    - P(B)
- P(smoker|cancer) = 0.20
    - (20% of those who have cancer are smokers)
    - P(B|A)
- Find: P(cancer|smoker):
    - P(cancer|smoker) = 0.20 * 0.05 / 0.10 = 0.10

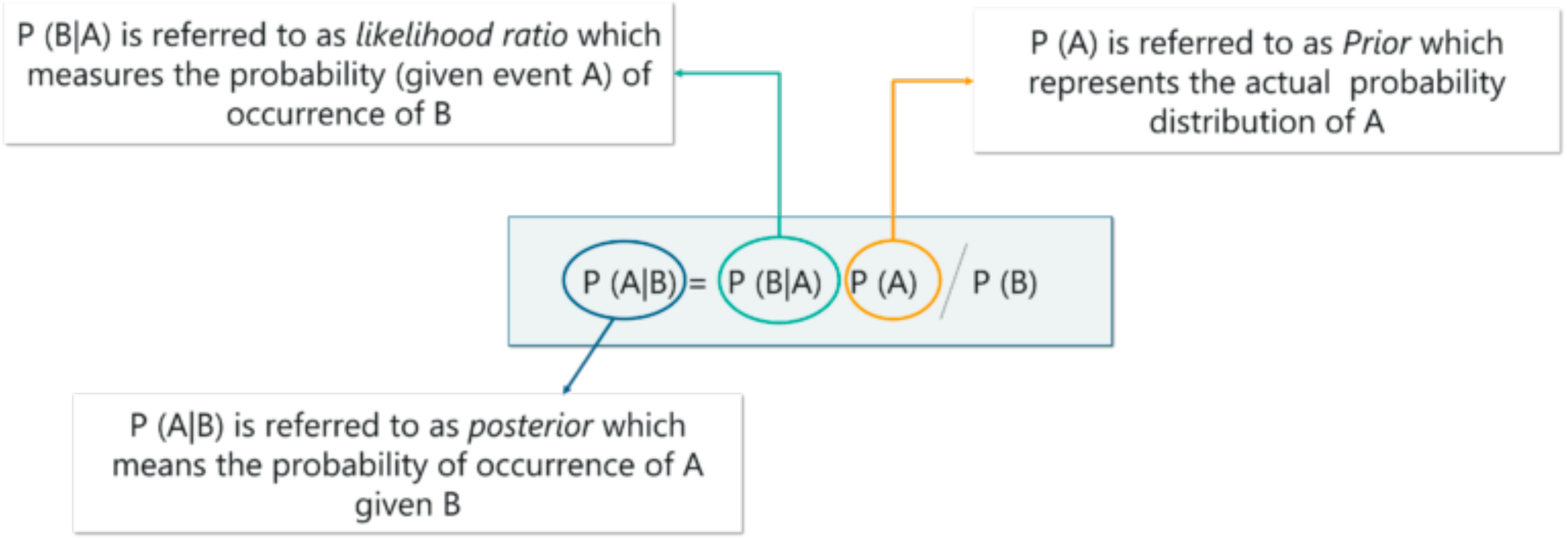
Bayes Theorem

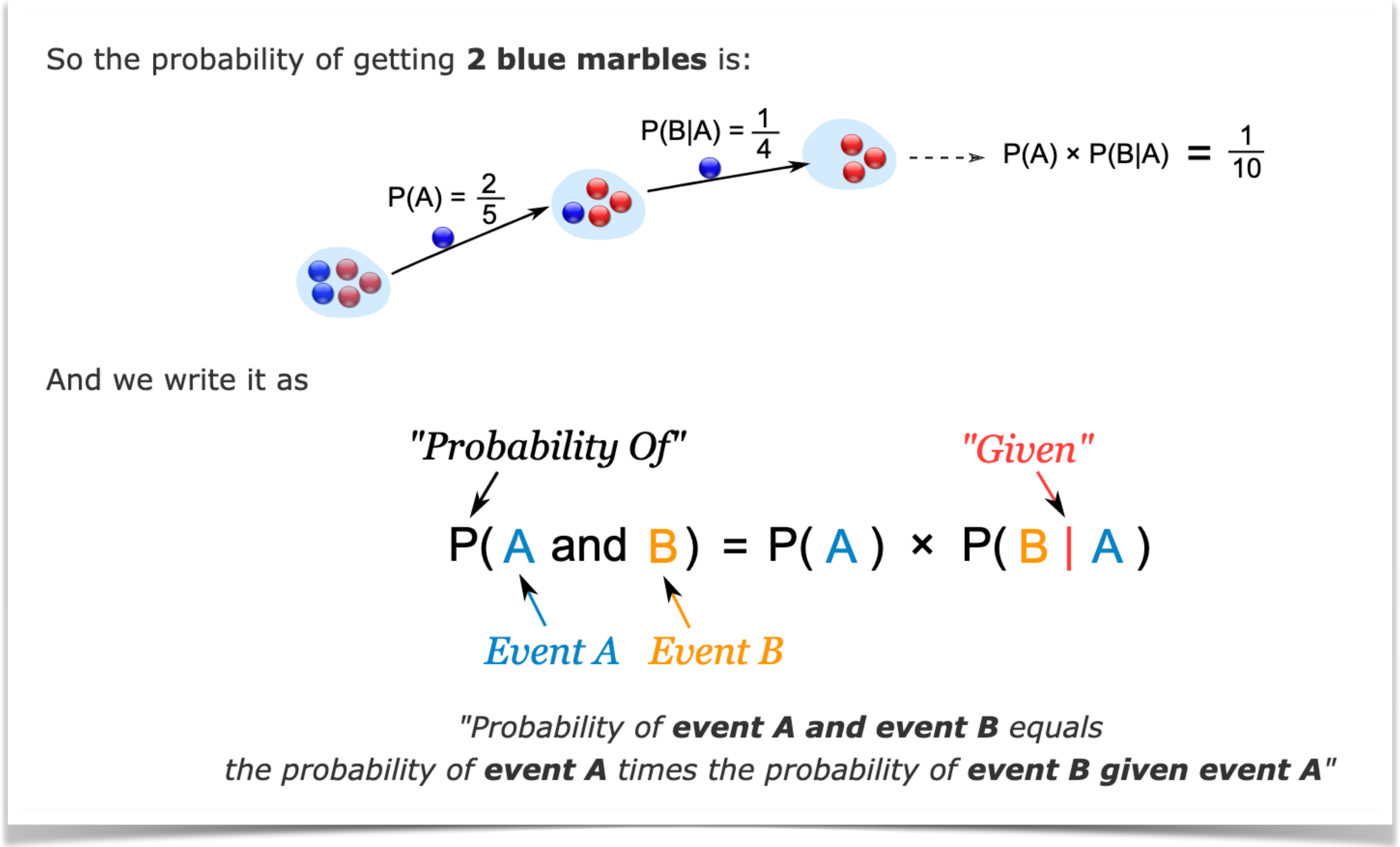
a classic example

- Suppose:
  - 1% of population have cancer
  - 80% of tests detect it correctly while 20% of tests fail to detect it ("false negative")
  - 9.6% of tests detect it when it is not there ("false positive") while 90.4% correctly return negative
- Q: If you get a positive result, what is the probability of you having the disease?
  - Work it out in a **group activity:** https://olzama.github.io/Ling471/assignments/activity-May6.html
  - Hint: "P(B) is the P(positive test). But P(positive test) is **not directly given** to you!
    - **Positive test outcome** means: [the test is positive AND person has cancer] **OR** [the test is positive and there is NO cancer!]
    - Use the marbles example: P(**two** events) is similar to P(**two** marbles)

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

The Bayes Theorem

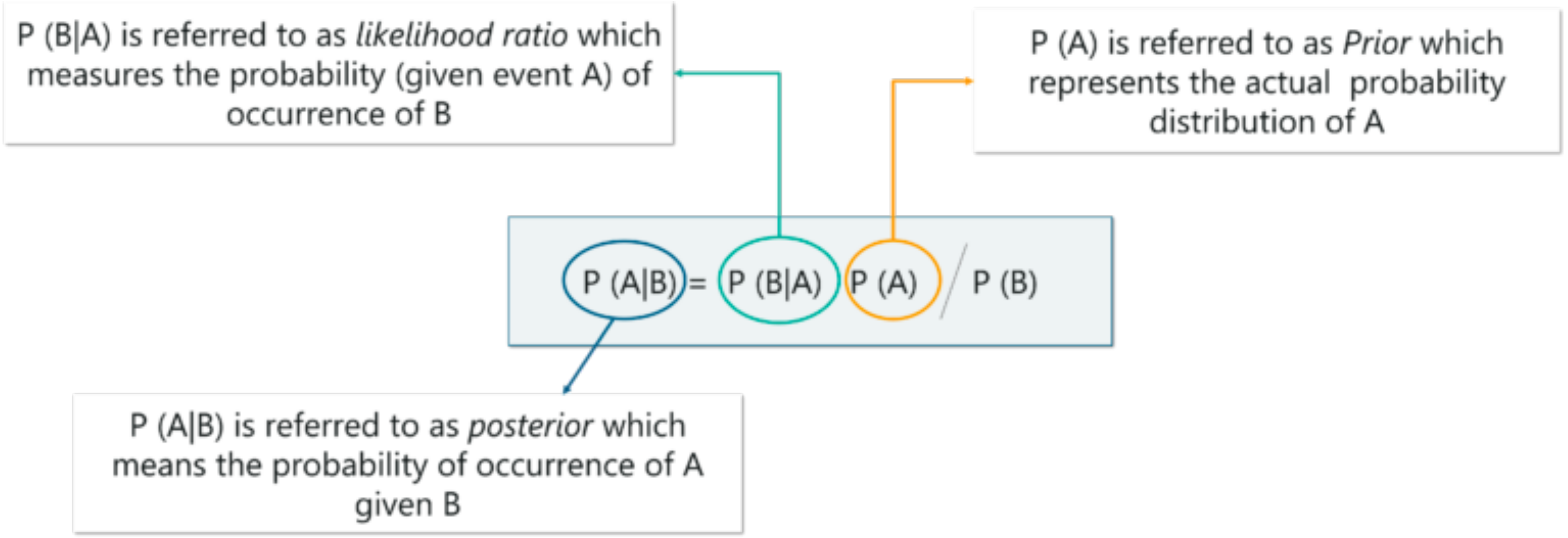




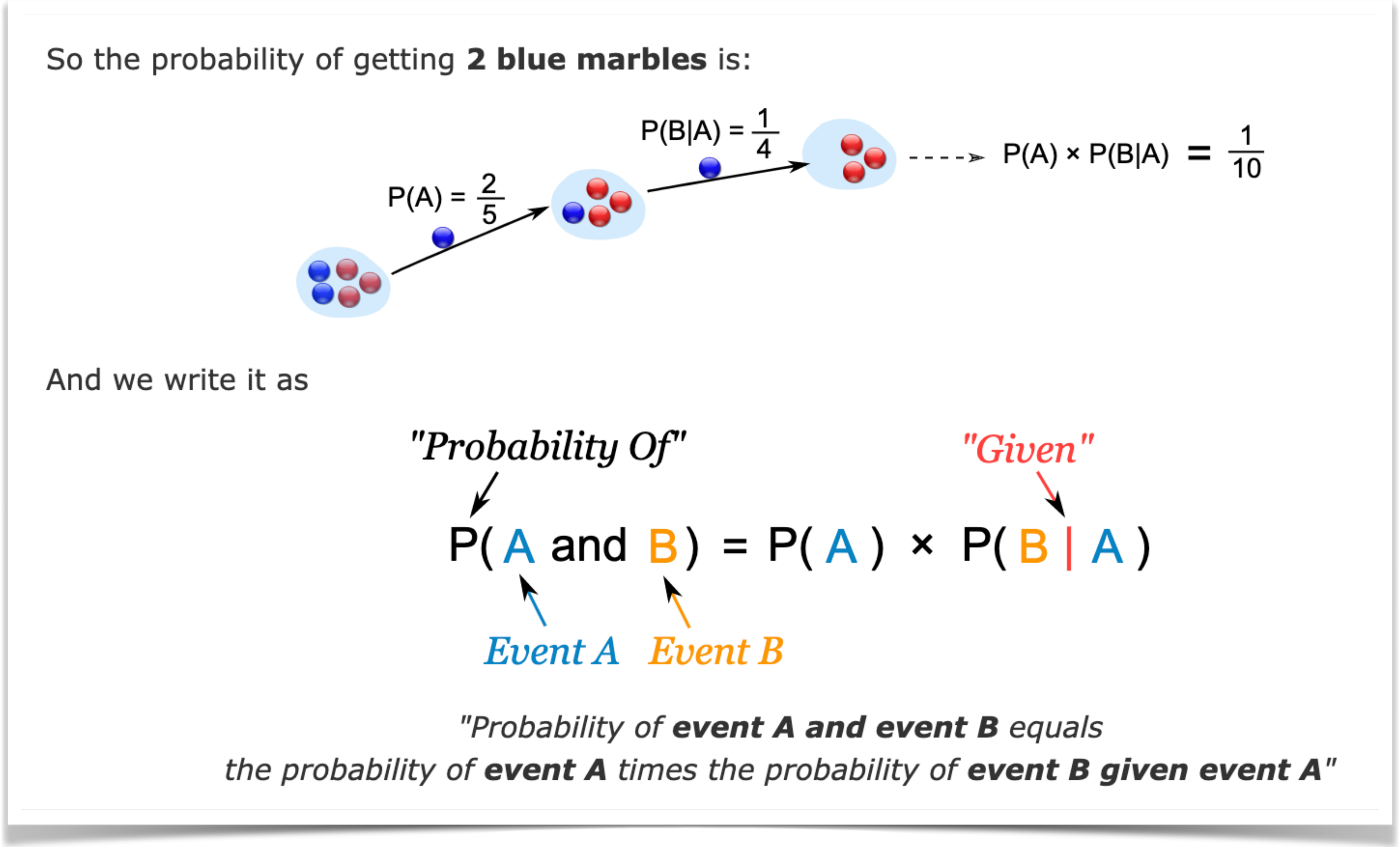
Bayes Theorem
a classic example

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

The Bayes Theorem



- Suppose:
  - 1% of population have cancer
  - 80% of tests detect it correctly while 20% of tests fail to detect it (“false negative”)
    - This 80% is out of people who **do** have the disease!
  - 9.6% of tests detect it when it is not there (“false positive”) while 90.4% correctly return negative
- **Q:** If you get a positive result, what is the probability of you having the disease?
  - **Answer: 7.8%**
  - Useful reading: https://towardsdatascience.com/3-ways-to-think-about-bayes-rule-b6f5b4ef87d6



Dataframes and pandas package

Installing packages
with pip

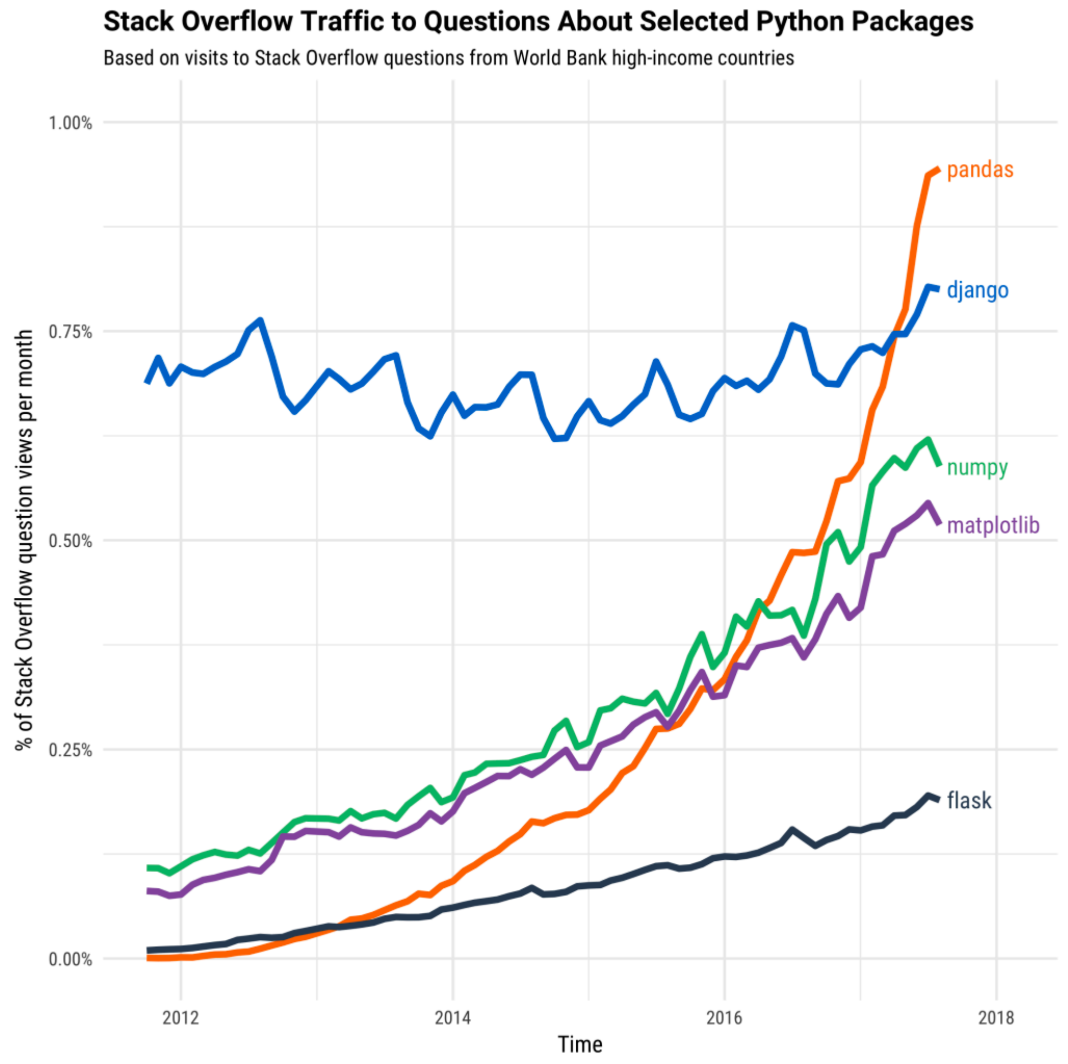
- We will need several packages for next HW
- They are best installed via pip
- pip is included in python distribution (starting from python **3.8)**
  - **Usually**, it just works
  - Some people are having issues on Windows 10
  - See instructions here:
    - https://phoenixnap.com/kb/install-pip-windows
  - ...and here:
    - https://stackoverflow.com/questions/23708898/pip-is-not-recognized-as-an-internal-or-external-command
  - In any case, start with checking whether you have pip already
    - pip --version
    - python -m pip --version
    - py -m pip —version

Pandas

a popular data science package

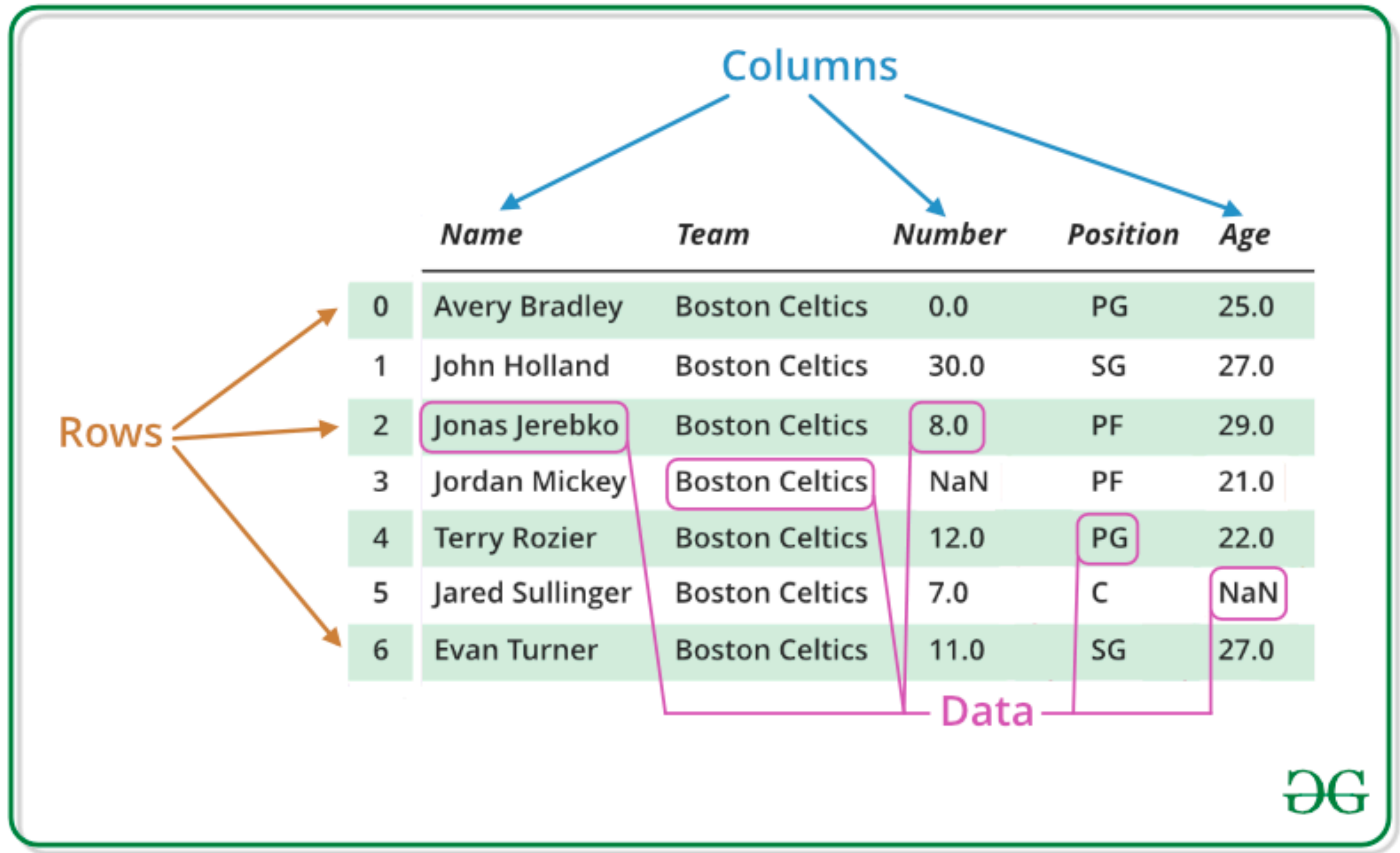
- Stores data in convenient tables
- Allows for fast data access and manipulation
- Why store data as tables?
  - In data science/statistics/machine learning:
    - you work with “observations” (=data points)
    - each data point is a row
    - What are columns?
      - data point can have different features
      - e.g. word counts!

https://www.kdnuggets.com/2020/03/python-pandas-data-discovery.html



Data as tables

- **rows**:
  - observations, datapoints
- **columns**:
  - “features”
  - can be many or few!
- Many ML algorithms involve linear algebra
  - Linear algebra includes matrix multiplication
    - Matrices are tables!



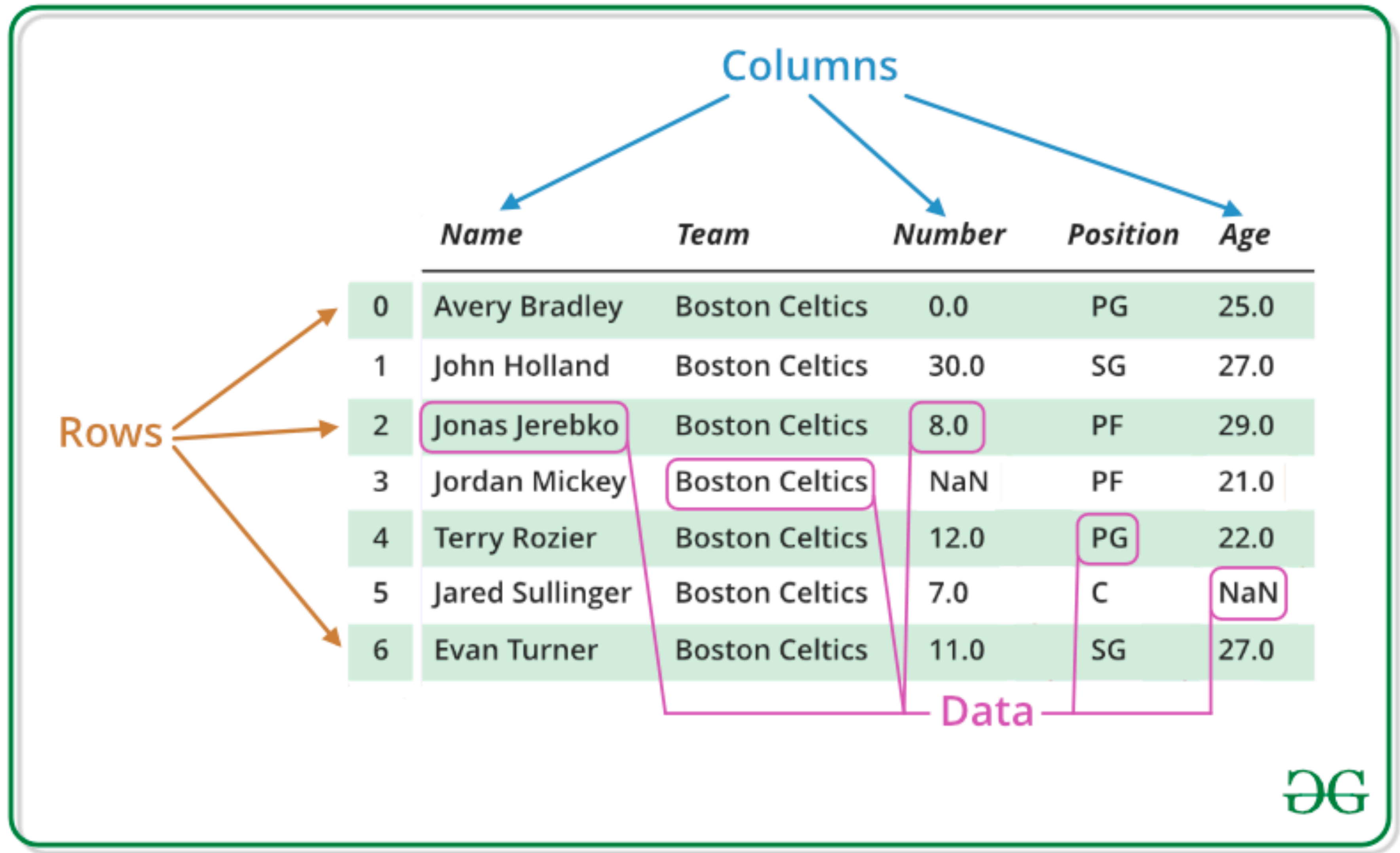
https://www.geeksforgeeks.org/python-pandas-dataframe/

```
   label                                               text
0      1  For a movie that gets no respect there sure ar...
1      1  Bizarre horror movie filled with famous faces ...
2      1  A solid if unremarkable film Matthau as Einste...
3      1  Its a strange feeling to sit alone in a theate...
4      1  You probably all already know this by now but ...
```

```
v 0: <review_vector.reviewVec object at 0x7ffc703b3be0>
  > special variables
    correct_label: 'POSITIVE'
    text: 'good good good bad bad '
```

Data as tables

- **rows**:
  - observations, datapoints
- **columns**:
  - “features”
  - can be many or few!
- The specific **dimensions** are crucial
  - For **any ML** algorithm:
    - need to know very well **how many columns** you have
      - (sometimes also rows, but that’s less important for us)
- In pandas, columns can have **names**
  - which allows convenient querying
  - the “names” row is **ignored**
    - it is not an “observation”/datapoint



https://www.geeksforgeeks.org/python-pandas-dataframe/

```
   label                                               text
0      1  For a movie that gets no respect there sure ar...
1      1  Bizarre horror movie filled with famous faces ...
2      1  A solid if unremarkable film Matthau as Einste...
3      1  Its a strange feeling to sit alone in a theate...
4      1  You probably all already know this by now but ...
```

pandas demo

Lecture survey:
in the chat