MODULE 10

Classification

PREDICTION

GUESSING THE VALUE OF AN ATTRIBUTE

- Based on incomplete information
- One way of making predictions:
  - To predict an outcome for an individual,
  - find others who are like that individual
  - and whose outcomes you know.
  - Use those outcomes as the basis of your prediction.

- Two Types of Prediction
  - Classification = Categorical; Regression = Numerical

PREDICTION EXAMPLE: SPAM OR NOT?

You made a Wells Fargo payment - wellsfargo.com You recently submitted a payment The ...

BUSINESS TRUST - -- I have a legal business proposal for you worth $23,000,000. If you kn...

Hi - Today???!!!! What a wonderful day! Congrats again! I am definitely not doing s...

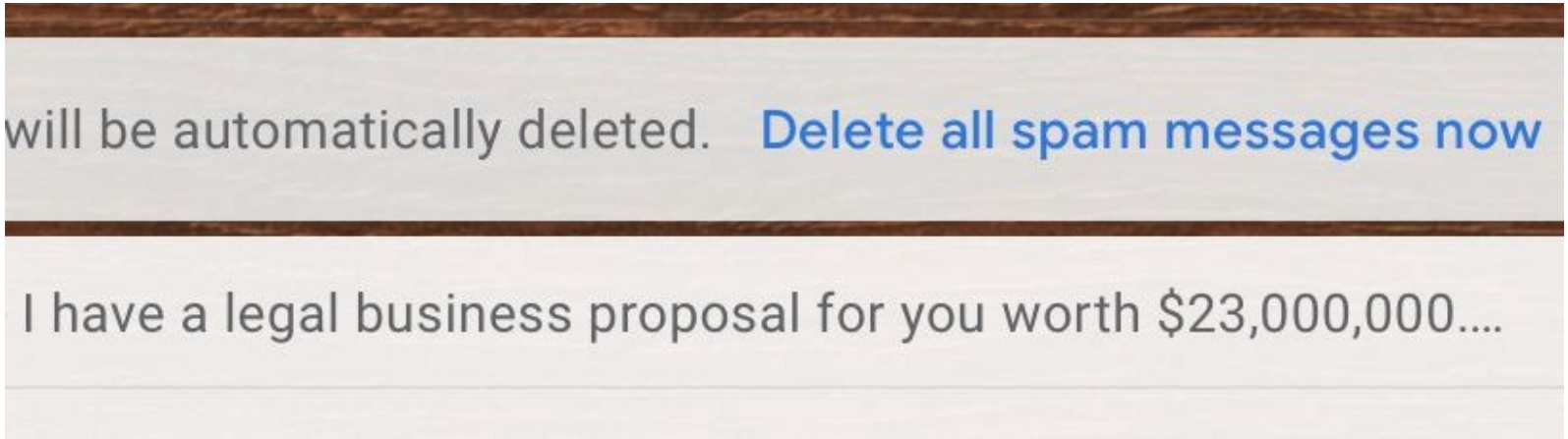
Michael Kors Handbags Up To 84% Plus Free Shipping! - Shop Handbags Online & In Store...

MACHINE LEARNING ALGORITHM

- A mathematical model
- calculated based on sample data ("training data")
- that makes predictions or decisions without being explicitly programmed to perform the task

CLASSIFICATION

CLASSIFICATION EXAMPLES



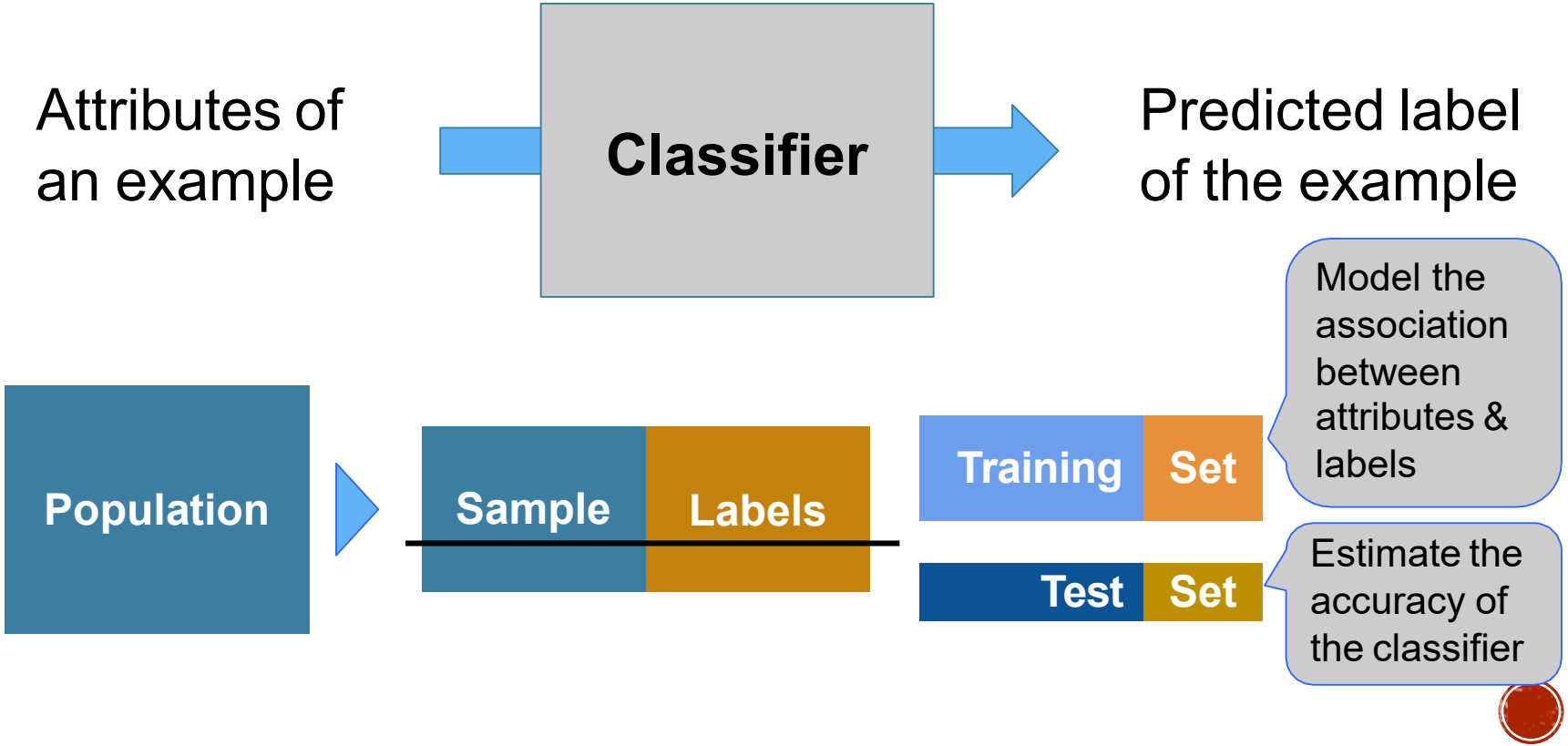
CLASSIFICATION EXAMPLES

Classification Examples

(Demo)

CLASSIFIERS

TRAINING A CLASSIFIER

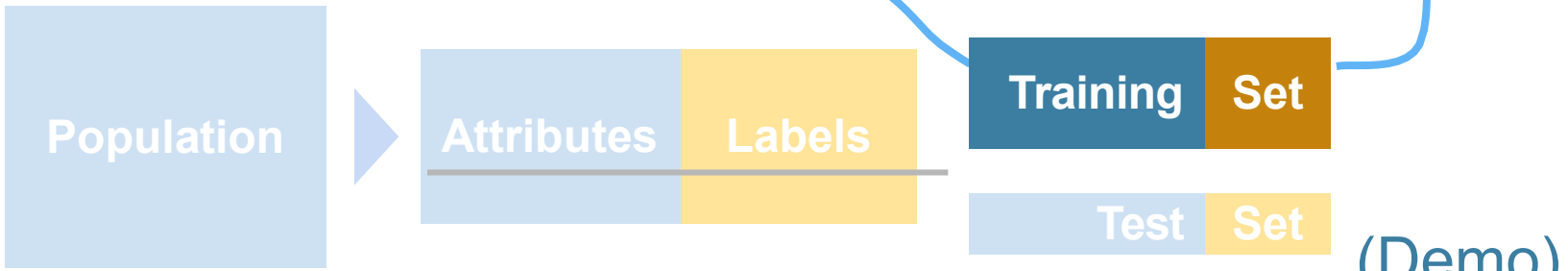


NEAREST NEIGHBOR CLASSIFIER

Attributes of an example

NN Classifier
Use the label of the most similar training example

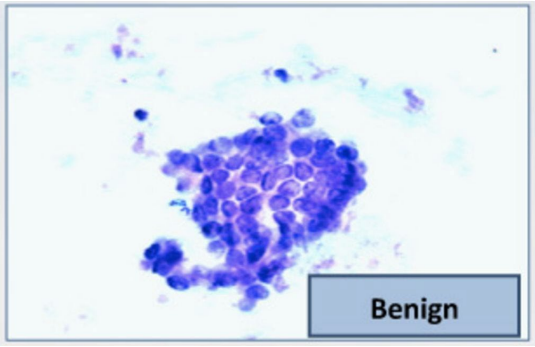
Predicted label of the example

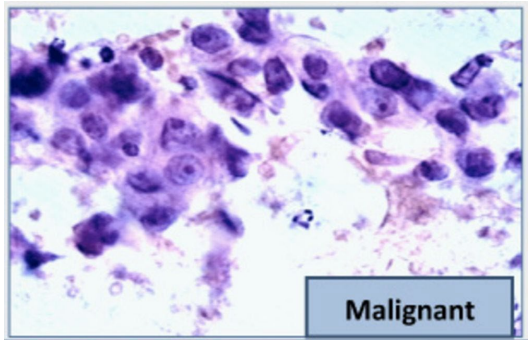


(Demo)

The Google Science Fair

- Brittany Wenger, a 17-year-old high school student in 2012
- Won by building a breast cancer classifier with 99% accuracy





(Demo)

ROWS

Rows of Tables

Each row contains all the data for one individual

- `t.row(i)` evaluates to `i`th row of table `t`
- `t.row(i).item(j)` is the value of column `j` in row `i`
- If all values are numbers, then `np.array(t.row(i))` evaluates to an array of all the numbers in the row.
- To consider each row individually, use

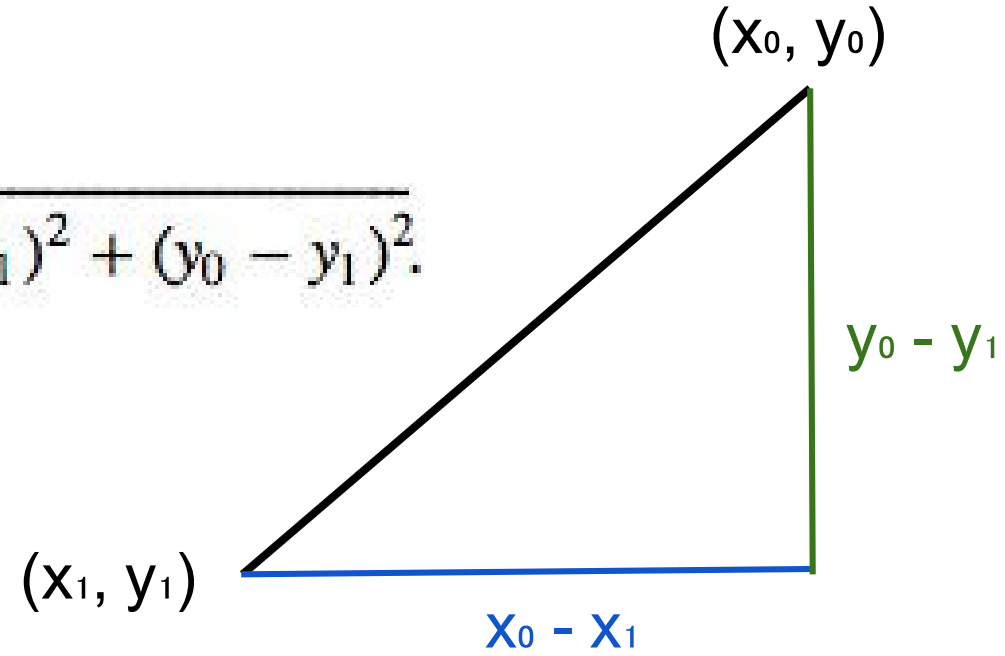
```
for row in t.rows:
    ... row.item(j) ...
```

- `t.exclude(i)` evaluates to the table `t` without its `i`th row

DISTANCE

PYTHAGORAS' FORMULA

$$D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}.$$



DISTANCE BETWEEN TWO POINTS

- Two attributes *x* and *y*:

$$D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}.$$

- Three attributes *x*, *y*, and *z*:

$$D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2 + (z_0 - z_1)^2}$$

  - and so on ...

(Demo)

NEAREST NEIGHBORS

FINDING THE *K* NEAREST NEIGHBORS

To find the *k* nearest neighbors of an example:

- Find the distance between the example and each example in the training set
- Augment the training data table with a column containing all the distances
- Sort the augmented table in increasing order of the distances
- Take the top *k* rows of the sorted table

THE CLASSIFIER

To classify a point:

- Find its *k* nearest neighbors
- Take a majority vote of the *k* nearest neighbors to see which of the two classes appears more often
- Assign the point the class that wins the majority vote

(Demo)

EVALUATION

Accuracy of a classifier

The accuracy of a classifier on a labeled data set is the proportion of examples that are labeled correctly.

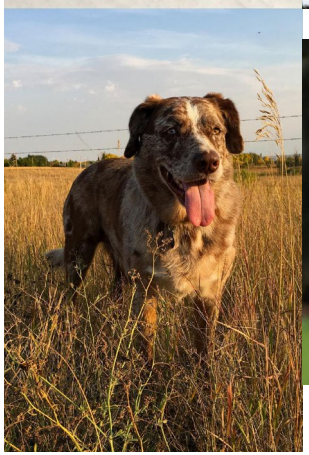
Need to compare classifier predictions to true labels.

If the labeled data set is sampled at random from a population, then we can infer accuracy on that population.

(Demo)

BEFORE CLASSIFYING

DOG OR WOLF?



START WITH A REPRESENTATIVE SAMPLE

- Both the training and test sets must accurately represent the population on which you use your classifier

- **Overfitting** happens when a classifier does very well on the training set, but can't do as well on the test set

STANDARDIZE IF NECESSARY

Chronic Kidney Disease data set

Glucose	Hemoglobin	White Blood Cell Count	Class
117	11.2	6700	1
70	9.5	12100	1
380	10.8	4500	1
157	5.6	11000	1

- If the attributes are on very different numerical scales, distance can be affected
- In such a situation, it is a good idea to convert all the variables to standard units

(Demo)

DECISIONS

DECISIONS UNDER UNCERTAINTY

- *Interpretation by Physicians of Clinical Laboratory Results* (1978)
  - "We asked 20 house officers, 20 fourth-year medical students and 20 attending physicians, selected in 67 consecutive hallway encounters at four Harvard Medical School teaching hospitals, the following question:
  - "If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs?"

DECISIONS UNDER UNCERTAINTY

- *Interpretation by Physicians of Clinical Laboratory Results* (1978)

  - "Eleven of 60 participants, or 18%, gave the correct answer. These participants included 4 of 20 fourth-year students, 3 of 20 residents in internal medicine and 4 of 20 attending physicians. The most common answer, given by 27, was that [the chance that a person found to have a positive result actually has the disease] was 95%.

CONDITIONAL PROBABILITY

SCENARIO 1

- Scenario:
  - Class consists of second years (60%) and third years (40%)
  - 50% of the second years have declared their major
  - 80% of the third years have declared their major
- **I pick one student at random.**
- Which is more likely: Second year or Third year?
  - Second year, because they are 60% of the class

SCENARIO 2

- Slightly different scenario:
  - Class consists of second years (60%) and third years (40%)
  - 50% of the second years have declared their major
  - 80% of the third years have declared their major
- **I pick one student at random...**
  That student has declared a major!
- Which is more likely: Second Year or Third Year?

(Demo)

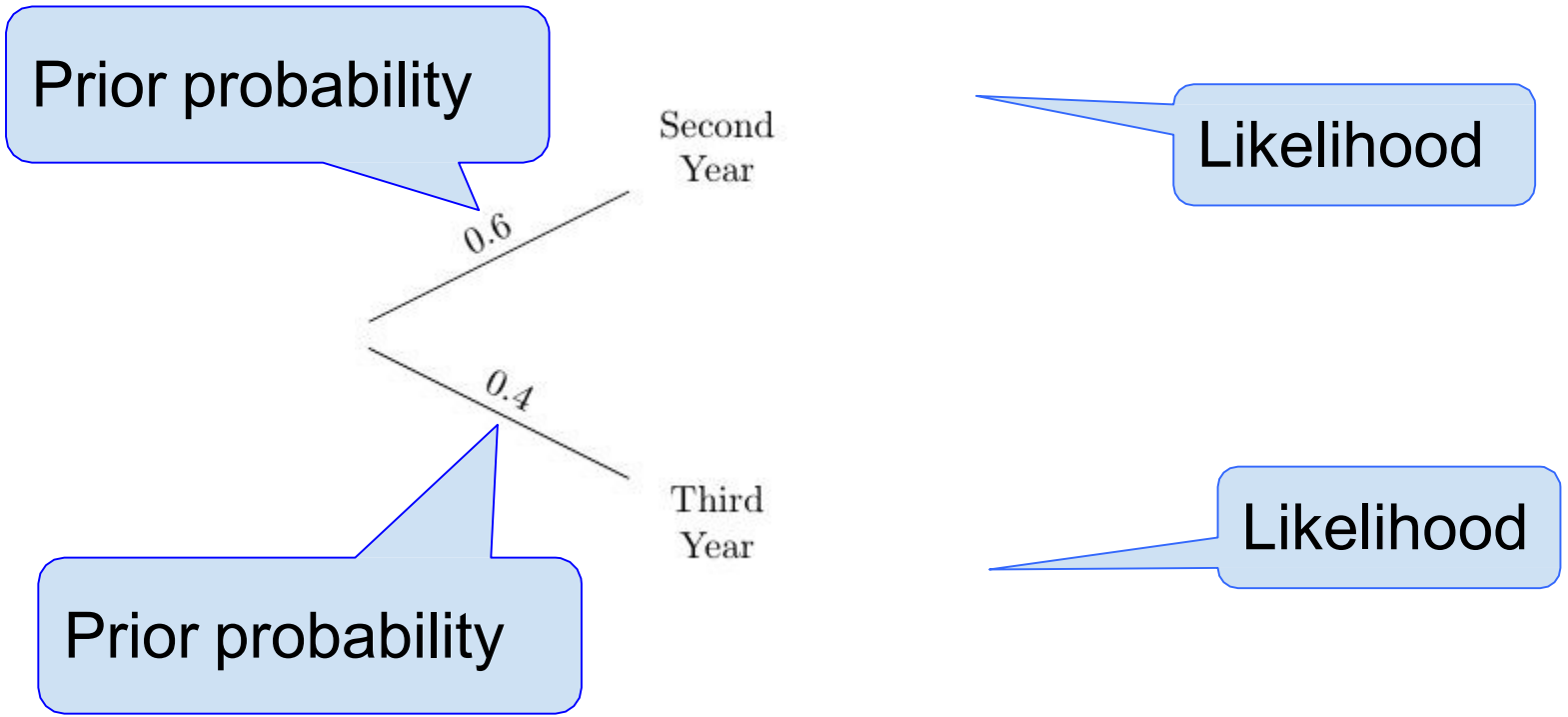
BAYES' RULE

PURPOSE OF BAYES' RULE

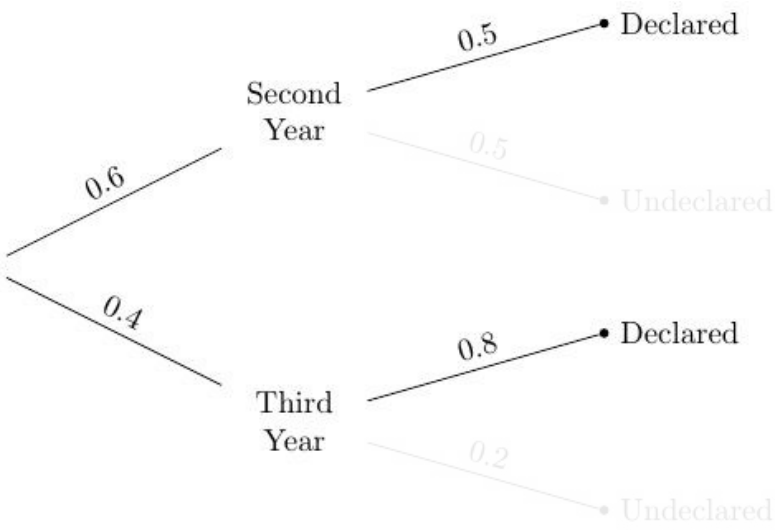
- Update your prediction based on new information

- In a multi-stage experiment, find the chance of an event at an earlier stage, given the result of a later stage

DIAGRAM AND TERMINOLOGY



DATA & CALCULATION



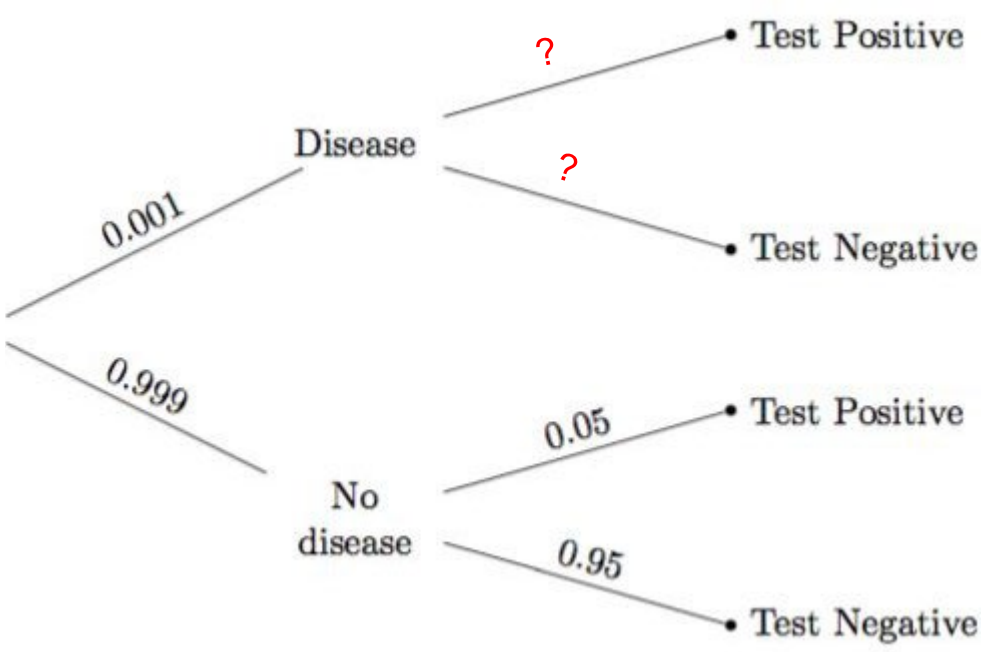
Pick a student at random.

Posterior probability:

$$P(\text{Third Year} \mid \text{Declared}) = \frac{0.4 \times 0.8}{(0.6 \times 0.5) + (0.4 \times 0.8)}$$

$$= 0.5161\ldots$$

EXAMPLE: DOCTORS & CLINICAL TESTS

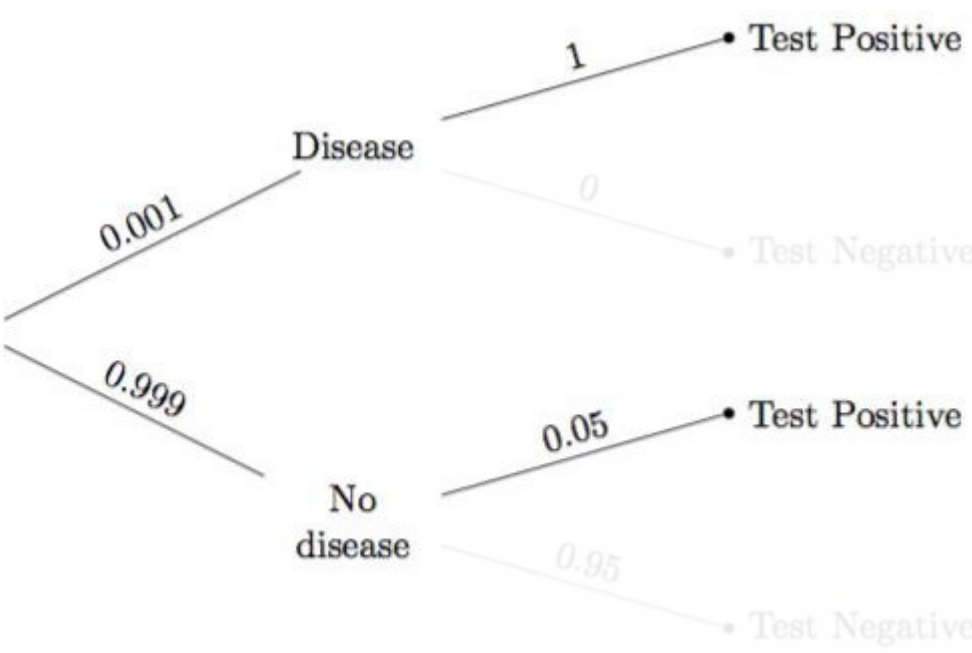


Problem did not give the *true positive* rate.

That's the chance the test says "positive" if the person has the disease.

It was assumed to be 100%.

DATA AND CALCULATION



P(Disease | Test +)

=

$$\frac{0.001 * 1}{(0.001 * 1) + (0.999 * 0.05)}$$

= 0.0196270…

(Demo)

SUBJECTIVE PROBABILITIES

SUBJECTIVE PROBABILITIES

A probability of an outcome is…

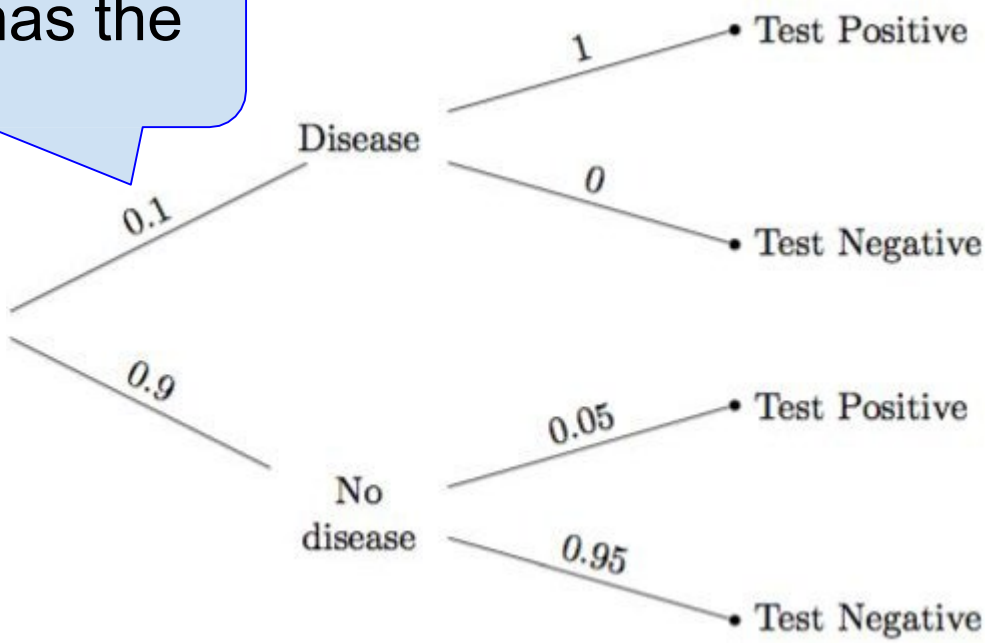
- The frequency with which it will occur in repeated trials, *or*
- The subjective degree of belief that it will (or has) occurred

Why use subjective priors?

- In order to quantify a belief that is relevant to a decision
- If the subject of your prediction was not selected randomly from the population

A SUBJECTIVE OPINION

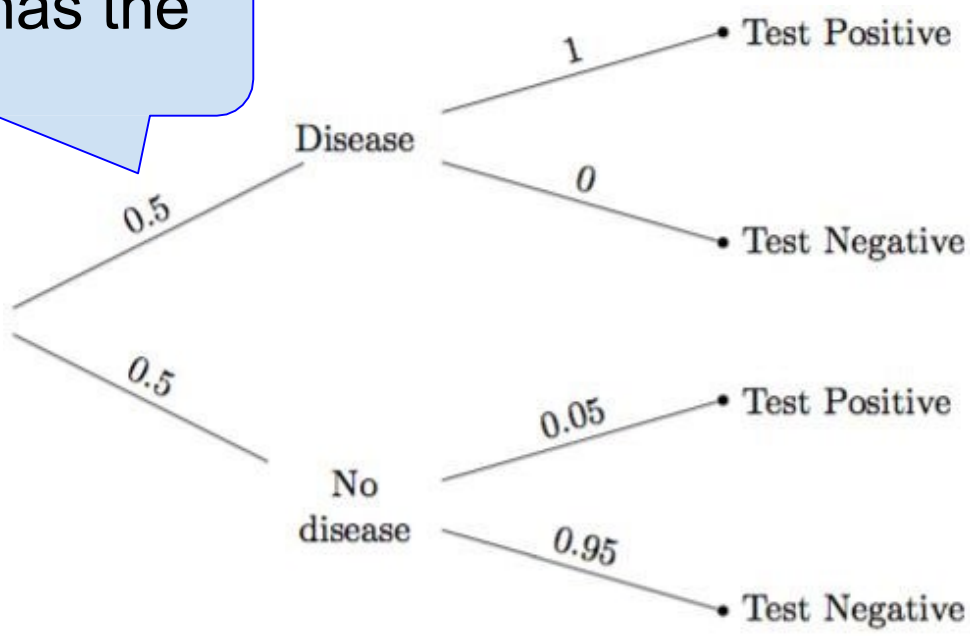
prior probability that the person has the disease



(Demo)

A DIFFERENT SUBJECTIVE OPINION

prior probability that the person has the disease



(Demo)