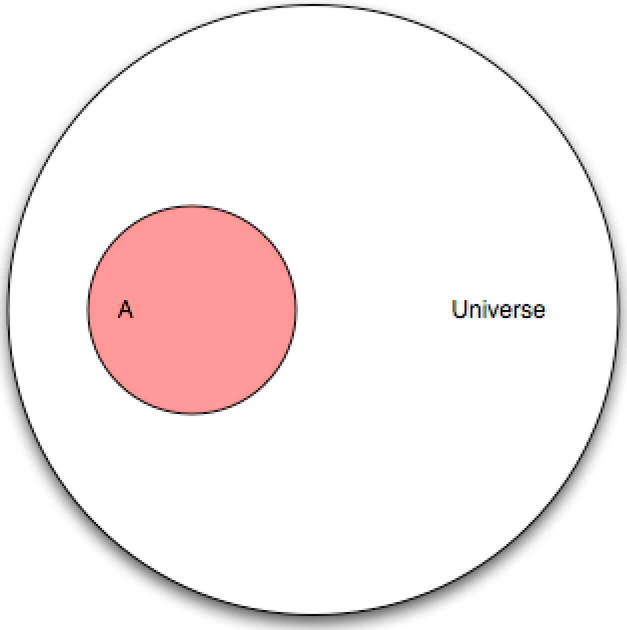
GENERAL ASSEMBLY

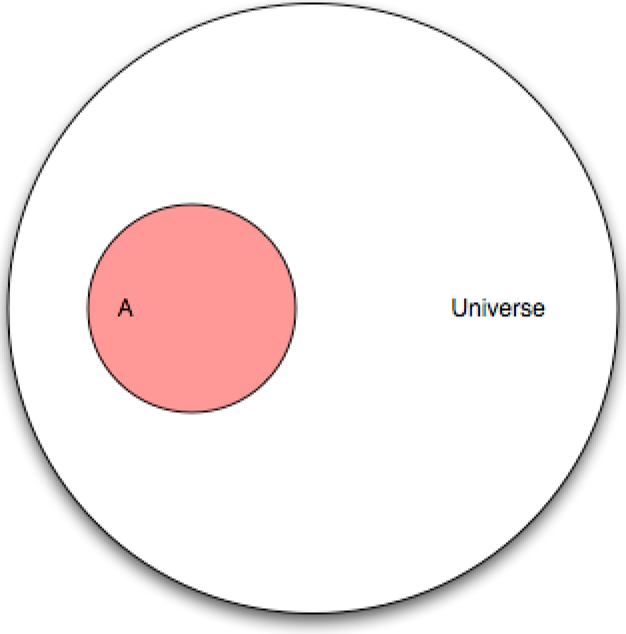
DATA SCIENCE

PROBABILITY AND BAYES' THEOREM



Let's pretend you flipped a coin and haven't looked at the result. This diagram represents the "universe" of all possible outcomes, also known as **events**. *This universe is known as the* **sample space**.

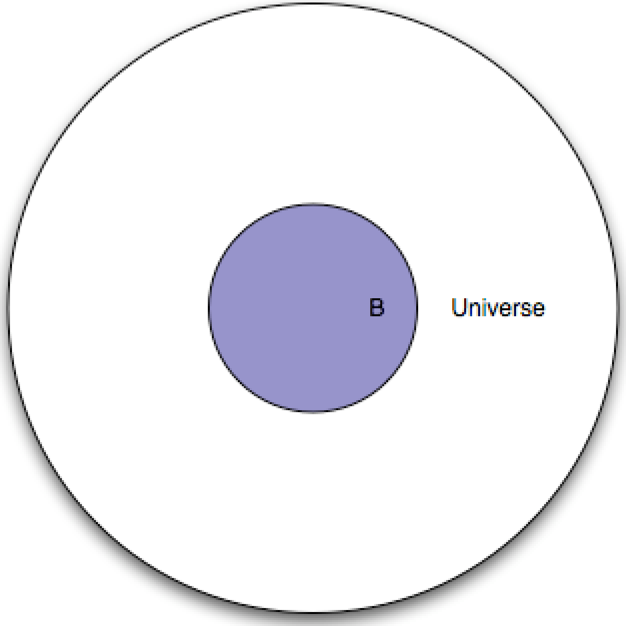
Q: What are the ***mutually exclusive events*** *that make up the* ***sample space*** *for a coin flip?*
A: Heads and tails



Let's now pretend that our universe involves a research study on humans. Event "A" is people in that study who have cancer.

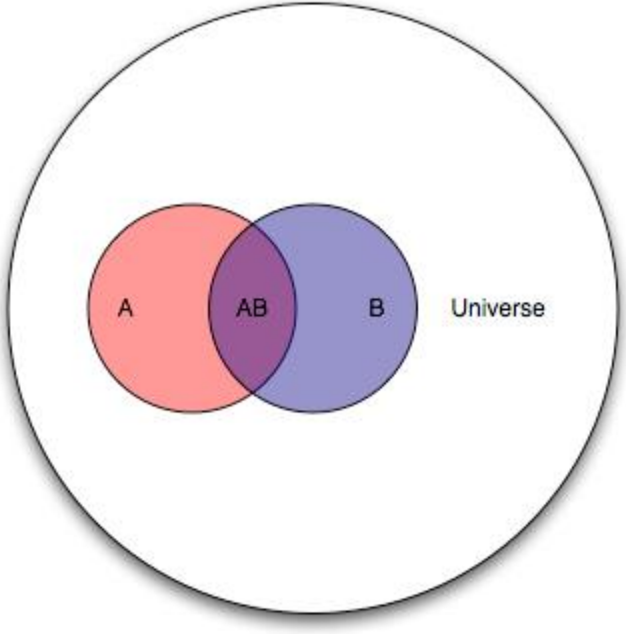
Q: If our study has 100 people and "A" has 25 people, what is the **probability** *of A?*
A: P(A) = 25/100

Q: What is the max probability of any event?
A: 1



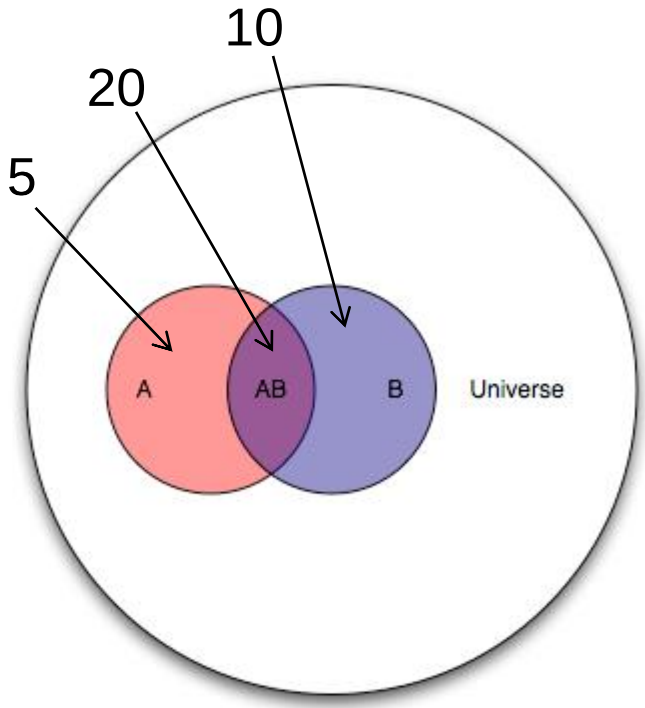
This represents the same set of people, except everyone in the study is given a test. Event "B" is everyone in the study for whom the test is positive.

Q: What portion of the diagram represents the subset of people with a negative test?
A: The white area between the smaller circle and the larger circle.



Because "A" and "B" are events from the same study, we can show them together.

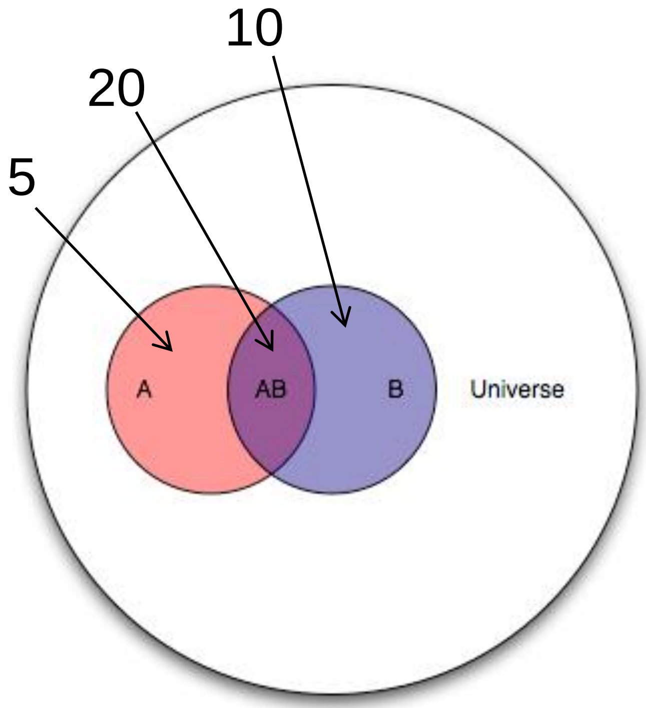
Q: How would you describe the "cancer status" and "test status" of people in each portion of the diagram (by color)?
A: Pink: cancer, negative test
Purple: cancer, positive test
Blue: no cancer, positive test
White: no cancer, negative test



The purple section is known as the **intersection** *of A and B, denoted as P(AB).*

Thinking of this test as a classifier for predicting cancer, draw the confusion matrix.

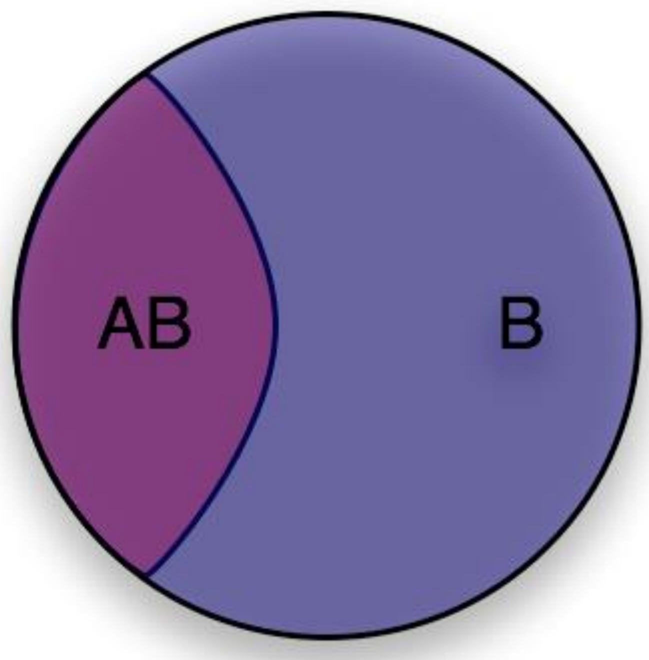
n=100	**Predicted: NO**	**Predicted: YES**
Actual: NO	65	10
Actual: YES	5	20



Q: Let's pick an arbitrary person from this study. If you were told their test result was positive, what is the probability they actually have cancer?
A: 20/30

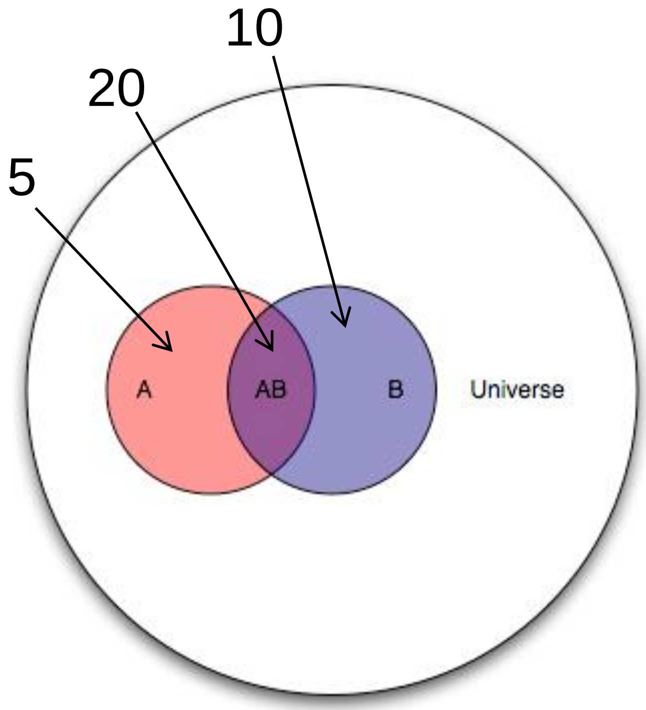
This is the **conditional probability of A given B**, *denoted as P(A|B).*

$P(A|B) = P(AB) / P(B) = (20/100) / (30/100)$



You can think of conditional probability as "changing the relevant universe." P(A|B) is a way of saying "Given that my entire universe is now B, what is the probability of A?"

This is also known as **transforming the sample space.**



Q: Let's pick another arbitrary person from this study. If you were told they have cancer, what is the probability they had a positive test result?

A: $P(B|A) = P(AB) / P(A) = 20/25$

Deriving Bayes' theorem:

We know: $P(A|B) = P(AB) / P(B)$ *and* $P(B|A) = P(AB) / P(A)$

Thus: $P(AB) = P(A|B) * P(B) = P(B|A) * P(A)$

Rearrange to get **Bayes' theorem**: $P(A|B) = P(B|A) * P(A) / P(B)$

Exercise:

1% of women at age forty who participate in routine screening have breast cancer. 80% of women with breast cancer will get positive mammograms. 9.6% of women without breast cancer will also get positive mammograms.

A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?

Part 1: Draw a confusion matrix

n=1000	Predicted: NO	Predicted: YES	
Actual: NO			
Actual: YES			

Part 1: Draw a confusion matrix

n=1000	Predicted: NO	Predicted: YES	
Actual: NO	895	95	990
Actual: YES	2	8	10
	897	103	

Given a positive test result, what is the probability of cancer?

8/103 = 7.8%

Part 2: Review of Terminology

*What is the **sensitivity** of the test?*

TP / actual yes = 80%

*What is the **specificity** of the test?*

TN / actual no = 1 - 9.6% = 90.4%

***Prevalence** = actual yes / total = 1%*

***Precision** = TP / predicted yes = 7.8%*

Part 3: Use Bayes' theorem

*P(A|B) = P(B|A) * P(A) / P(B)*

Event A is "has cancer." Event B is "positive test."
What is P(A|B)?

P(B|A) = 0.80
P(A) = 0.01
P(B) = 0.103
*P(A|B) = 0.80 * 0.01 / 0.103 = 7.8%*