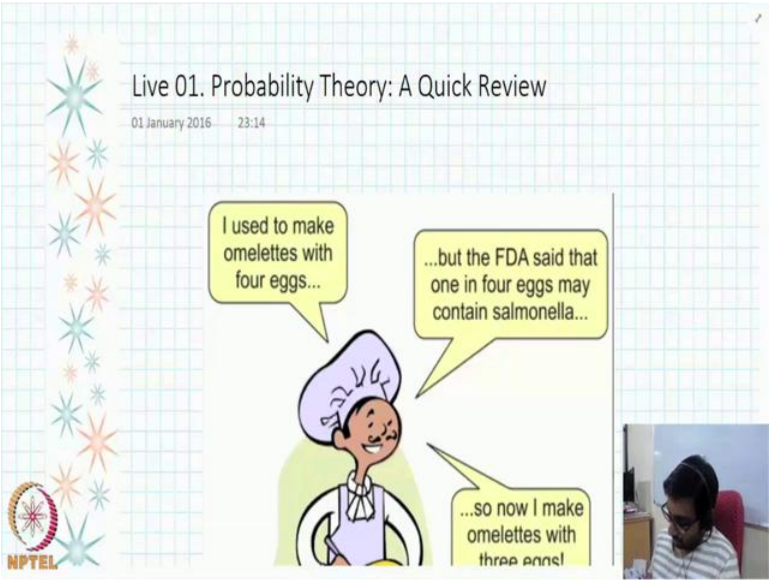
Algorithms for Big Data
Prof. John Ebenezer Augustine
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 01
Basic definitions

Hello everybody.

(Refer Slide Time: 00:19)

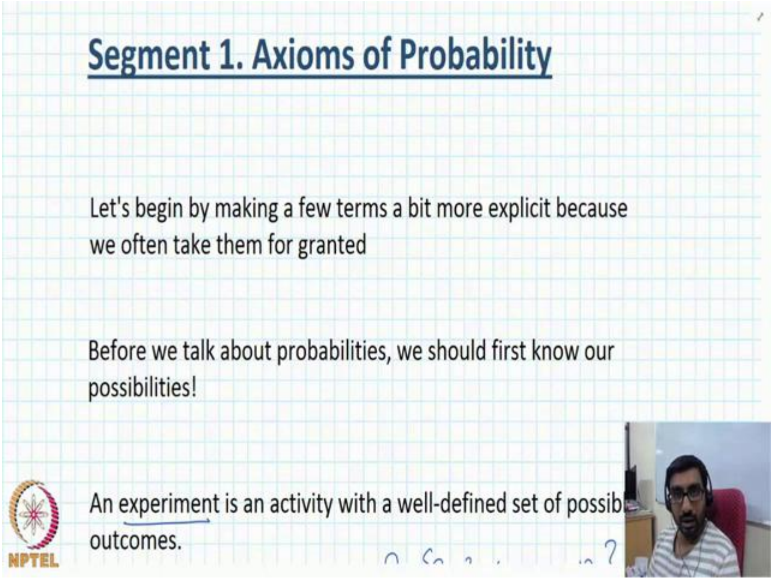


Actually you have seen Probability Theory in some context of the (Refer Time: 00:24). In today's lecture we are just going to quickly review this theory, and let us begin by looking at this cartoon which illustrates and in form the aspect of probability theory. A lot of times there are some misconceptions in this in peoples understanding of probability theory, as a result they have the tendency to apply them, in an appropriate place; this is not just true.

A particularly it is an also true or false, just as the whole and also in fact, a larger mathematics that is whole. But the cartoon, but nicely illustrate how misunderstanding can lead to funny an in appropriate a basic of (Refer Time: 01:13) theory.

So take the moment, pause enjoy the cartoon and continue.

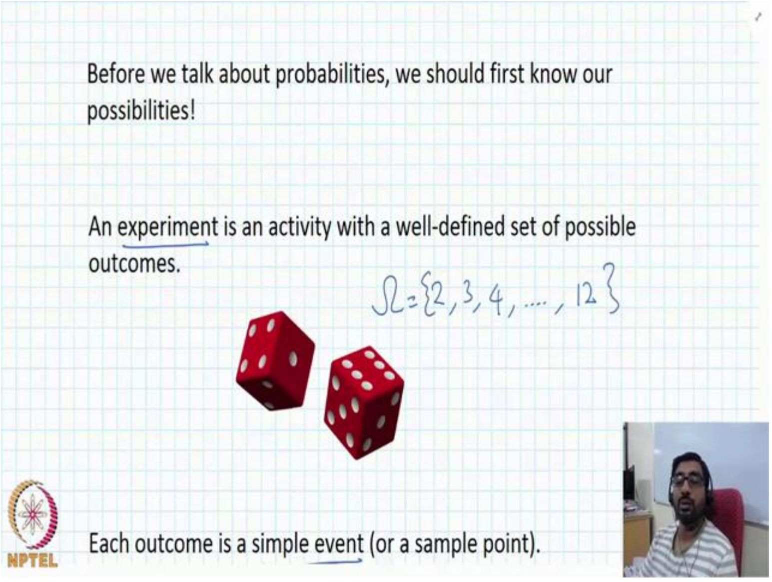
(Refer Slide Time: 01:23)



In that, so, now that to understand this theory, let us start from the very basics, let us start from the axioms of probability theory, and from that we have to make clear about some basic definitions, and let us start at the very hot probability theory in which is, an experiment. What is an experiment? Is a simple (Refer Time: 01:43) an activity that leads to a well defined a set of that occurs. So, when example of experiment will be a rolling a dice or a proximate point or running randomized algorithms and so on and so forth.

So, this outcome that we talk about the various possibilities that we can encounter as the result of the experimenters; now clear understanding of this possibilities, is very important and we, in fact a very quick example, we call it the sample space is this is simply the set up all possible outcomes, when you done an experiment. So, here is an experiment rolling two dice.

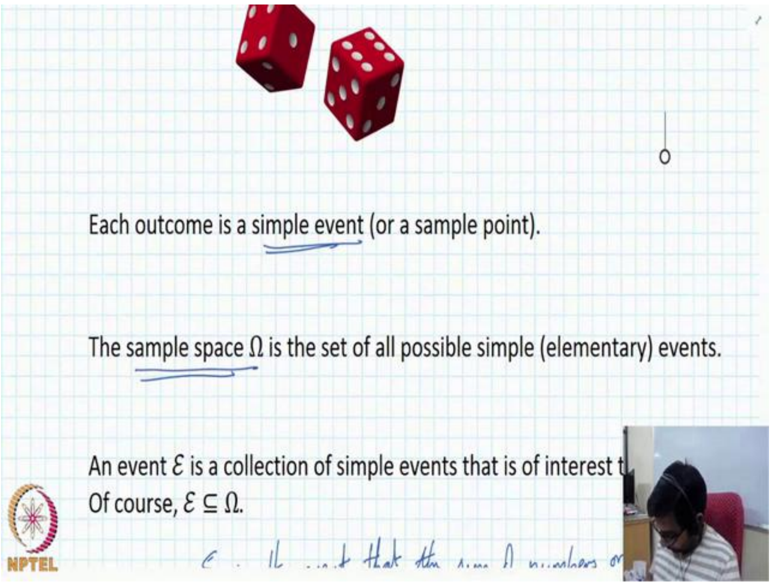
(Refer Slide Time: 02:54)



And the set of all possible outcomes is, let us suppose we are interested in the sum of the numbers that occur in both this dice has should be broken. So, outcomes possible outcomes are 2, 3, 4 up to 12 because one is not possible because, 1 in a both of dice will at least add the number of 1 on the faces. So, the some of them is replacement with 2.

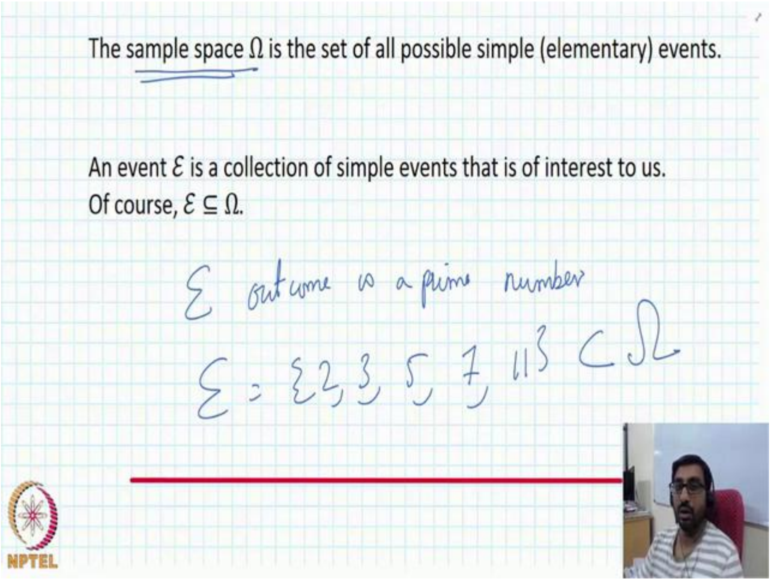
So, each of these outcomes is a basically 8, what is called as the simple event. So, in this particular experiment, we have 11 simple, possible simple events in some time this simple events are also called a sample points. As we mentioned the sample space is simply be a set of all possible outcomes.

(Refer Slide Time: 03:31)



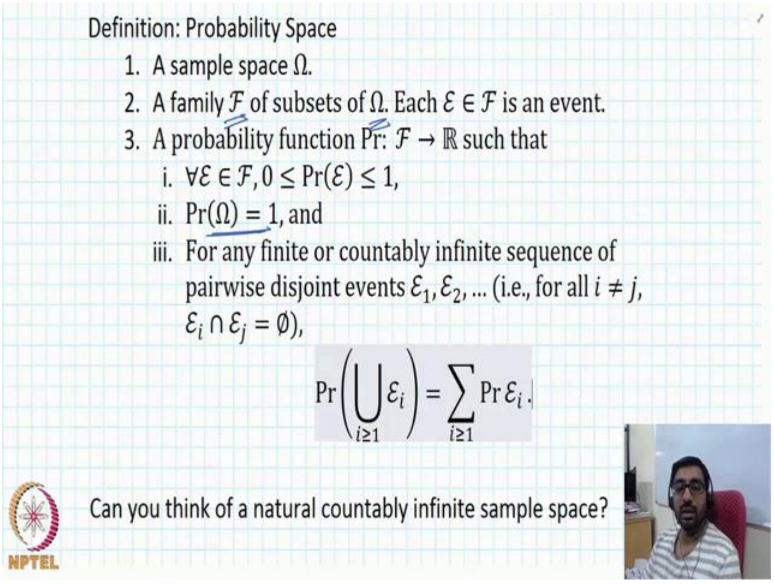
Now, sometimes we may not be interested in, the simple events the elements of the sample space. Sometimes we are interested in other events that are basically subsets of the sample space. So, let us give (Refer Time: 03:52) of some examples.

(Refer Slide Time: 03:53)



Suppose considered the event e is the event that outcome, is a prime number. So, this e is now 2, 3, 5, 7 or 11 and of course, this is a subset of the sample space. These events are useful because now we have the freedom to define these events and we can define it appropriate to the context in which we are interested in.

(Refer Slide Time: 04:31)



Now, that we have introduced the theories possible these sample space, now we had to term introduced the notion of probabilities the likely to which, each of events in sample space occur. So, for this we will have to define a probability space, and the probability space consist of a three compounds the first 20 of which we already seen the first compound, it is a simply the sample space we should have a very clear understand it, of what the experiment is and therefore, what the a set of all outcome says?

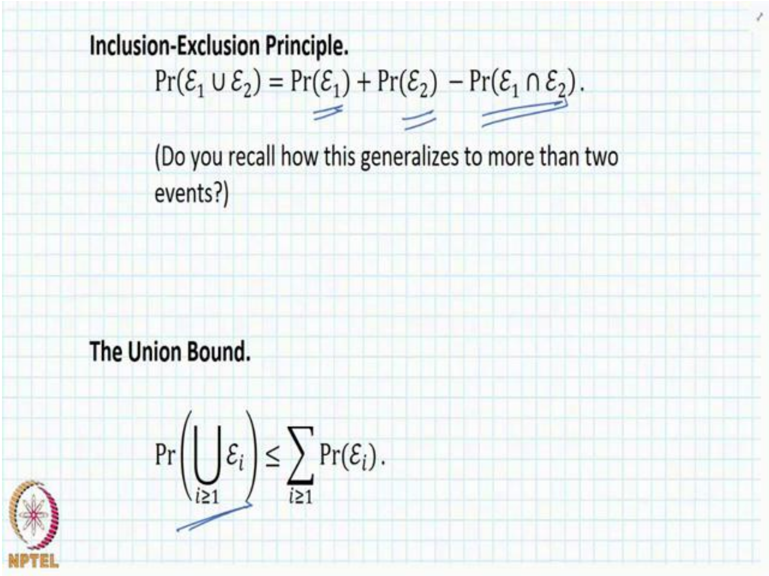
Then knowing the sample space, we are we must also know the set of all (Refer Time: 05:15), that we are interested in and these (Refer Time: 05:19), then it is captured as a family of subsets family, f of subsets of the sample space omega. And each element of this family is event there is of interest to us. Now we need to define a probability function, it is basically a function that maps that each event to a real value number, and this probability function must obey some a basic (Refer Time: 05:56), first of all a probabilities are measure of likely hood, and they range from 0 to 1, this is 0.

Then it is impossible for that. (Refer Time: 06:05), occur and it is a 1 the event must occur. So, for example, that the probability. So, let us say the event be an (Refer Time: 06:17), is. In fact, the entire sample space omega, then one of the outcomes must necessarily occur that is for the probability of the sample space itself must always be 1, and any other event must have a probability that is either 0 or something in between, 0 and 1, or 1. It can never go outside these this range 0 to 1.

And finally, we want these probabilities to also obey this rule. Now if you look at disjoint events. The probabilities of the union of this disjoint event, so here disjoint events are e 1, e 2 and so on. The probability of the union of the events must, equal the sum of the individual probabilities. Now, if we has such a probability function defined on top of a sample space and a set of (Refer Time: 07:32), and then what we have what we call a probability space, and the entire probability theory based on understanding probability space these are a very fundamental notions and everybody passed to clear have clear understanding of this probability space and I just quick exercise, we talked about these events being finite of accountably infinite.

We have also seen a probabilities are or a sample space where it had a finite number of events that simply (Refer Time: 08:13), rolling to that is can you think of a natural experiment, where the outcome is accountably infinite sequence of events.

(Refer Slide Time: 08:25)



So, before we end this segment, let us look at a couple of fundamental principles are, very useful for probability theory. First one is the Inclusion-Exclusion Principle, which actually comes from inclusion principle, exclusion principle and set theory. Let us say we have 2 events e1 and e2. Probability of their union is equal to the probability of the first event, e1 by itself what is the probability of e2 by itself, but then we done a little bit of over counting. So, let us do this small correction, this as more fraction and. So, we have to subtract how the probability of their intersections. So, this is class inclusion exclusion principles. And this I am sure you this context of probability theory, are the set theory and. Secondly, appear in probability theory as well of course.

And then you may recall that this principle can be generalized more than 2 sets, or in this case more than 2 events. So, put exercise e2, write down the formula when you have not just 2 events, but some kind of events. As simplification of the Inclusion-Exclusion Principle called the union bound.

Secondly, the very, very use full to motivate the usefulness considered a set of events, e1, e2 etcetera that are bad events, you do not want these times to up it will showed up the probability which these occur is very, very small also we are; obviously, interested in upper bounding the probability of the union of these events that is would be have on the left hand side. We could use the inclusion exclusion principle, but the formula gets pretty messy, especially when the number of events is smoother 2.

However, this these nice simplification the right hand side need not be as complicated or we want to enough account and for lot purposes this is could bound is sufficient or we want and all we need on the right hand side is, the summation of the individual a probabilities, this if you sum the individual probabilities there might be some over counting, but if we are careful with bounding the individual probabilities for a lot of applications such sufficient in order to give us a good upper bound thus in other words it is a smalls upper bounds. So, throughout this position you will be applying the union bounded the several context. Context this distances to be a very useful principle in a lot of algorithmic analysis.