LET'S CONSIDER THE PROBLEM OF

FRAUD DETECTION

YOU WORK AT A LARGE BANK OR PAYMENT SERVICE (SAY AMERICAN EXPRESS OR PAYPAL)

YOU WANT TO IDENTIFY FRAUDULENT CREDIT CARD TRANSACTIONS

THIS CAN BE FRAMED AS A CLASSIFICATION PROBLEM

CLASSIFY TRANSACTIONS AS

FRAUD OR NOT

AS WITH ANY CLASSIFICATION PROBLEM - WE NEED TO PICK

FEATURES

EACH TRANSACTION WOULD BE REPRESENTED AS A LIST OF THESE FEATURES (FEATURE VECTOR)

AMOUNT SPENT

IP ADDRESS

NUMBER OF FAILED ATTEMPTS

TIME SINCE LAST TRANSACTION

LOCATION OF TRANSACTION

RANDOM VARIABLES

AMOUNT SPENT [0, INFINITY]

IP ADDRESS SET OF ALL IP ADDRESSES IN THE WORLD

NUMBER OF FAILED ATTEMPTS [0,1,2,3.....]

TIME SINCE LAST TRANSACTION [0, INFINITY]

LOCATION OF TRANSACTION (MUMBAI, BANGALORE, CHENNAI ....)

EACH OF THESE IS A VARIABLE WHOSE VALUE CANNOT BE DETERMINED BEFOREHAND

HOWEVER THE RANGE OR SET OF VALUES IT CAN TAKE IS PREDETERMINED

EACH VARIABLE'S VALUE WOULD BE DIFFERENT FOR EACH TRANSACTION, BUT THE EXACT VALUE IT WILL TAKE IS SUBJECT TO CHANCE

TEMPERATURE

DISTANCE BETWEEN A PERSON'S EARS

A PERSON'S BLOOD TYPE

RANDOM VARIABLES ARE EVERYWHERE

NUMBER OF LEAVES ON A TREE

NUMBER OF TIMES A USER VISITS FACEBOOK IN A DAY

LENGTH OF A TWEET

(ALSO KNOWN AS
A STOCHASTIC VARIABLE)

A RANDOM VARIABLE

IS A VARIABLE WHOSE VALUE
IS SUBJECT TO VARIATIONS DUE TO CHANCE
I.E. RANDOMNESS

TYPES OF RANDOM VARIABLES

DISCRETE [0, 1, 2, 3, 4...]

CAN TAKE ONLY INTEGER VALUES

CONTINUOUS [0,1] ; [0, INFINITY)

CAN TAKE ANY VALUE FROM A RANGE OF VALUES

CATEGORICAL (RED, BLUE, GREEN)
(CATEGORIES/GROUPS)

CAN TAKE ONE OF A LIMITED, FIXED SET OF VALUES

RANDOM VARIABLES

AMOUNT SPENT [0, INFINITY]

CONTINUOUS RANDOM VARIABLE

IP ADDRESS SET OF ALL IP ADDRESSES IN THE WORLD

CATEGORICAL RANDOM VARIABLE

NUMBER OF FAILED ATTEMPTS [0,1,2,3.....]

DISCRETE RANDOM VARIABLE

TIME SINCE LAST TRANSACTION [0, INFINITY]

CONTINUOUS RANDOM VARIABLE

LOCATION OF TRANSACTION {MUMBAI, BANGALORE, CHENNAI ....}

EACH OF THESE IS A VARIABLE
WHOSE VALUE CANNOT BE
DETERMINED BEFOREHAND

HOWEVER THE RANGE
OR SET OF VALUES IT CAN TAKE
IS PREDETERMINED

EACH VARIABLE'S VALUE
WOULD BE DIFFERENT FOR
EACH TRANSACTION, BUT THE
EXACT VALUE IT WILL TAKE
IS SUBJECT TO CHANCE

PROBABILITY DISTRIBUTION

HERE IS A TABLE THAT TELLS US THE NATIONALITIES OF FACEBOOK USERS

Country	% of Facebook users
USA	10%
India	7%
Brazil	6%
Indonesia	4%
Mexico	3%
United Kingdom	2%
Japan	1.6%
France	1.6%
Others	65%

P(the person picked is from USA)

= P(X = USA)

= 0.1

A PROBABILITY DISTRIBUTION IS A TABLE OR AN EQUATION THAT LINKS EACH OUTCOME OF A STATISTICAL EXPERIMENT WITH ITS PROBABILITY OF OCCURRENCE

A STATISTICAL EXPERIMENT

PICK A USER AT RANDOM FROM THE ENTIRE GROUP OF FACEBOOK USERS

WHAT COUNTRY ARE THEY LIKELY TO BE FROM?

A RANDOM VARIABLE X

TOSSING A DIE IS A PROTOTYPICAL EXAMPLE OF A STATISTICAL EXPERIMENT

THE OUTCOME OF THE TOSS IS X

IT CAN TAKE ANY VALUE FROM THE SET {1,2,3,4,5,6}

A DISCRETE RANDOM VARIABLE

THIS IS IT'S PROBABILITY DISTRIBUTION

n	P (X = n)
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

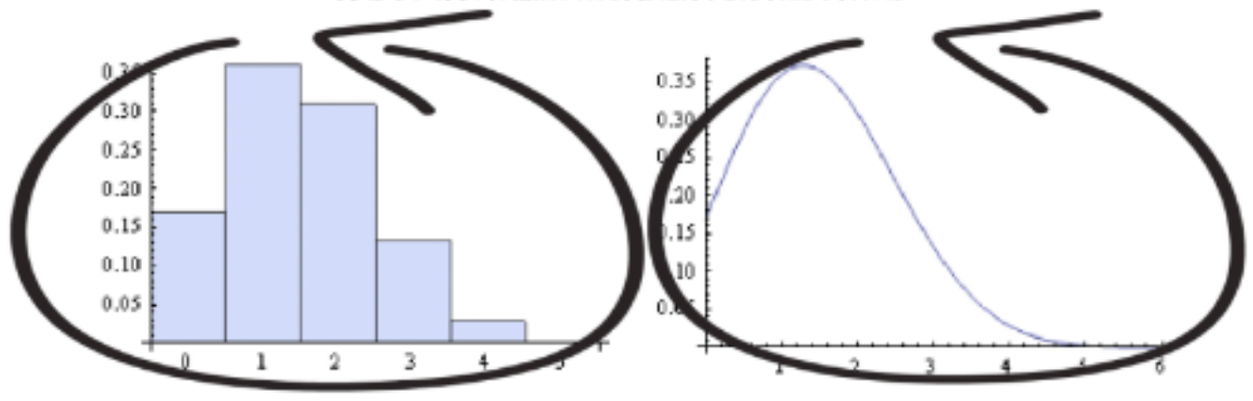
ALL THE VALUES HAVE EQUAL PROBABILITY

A UNIFORM PROBABILITY DISTRIBUTION

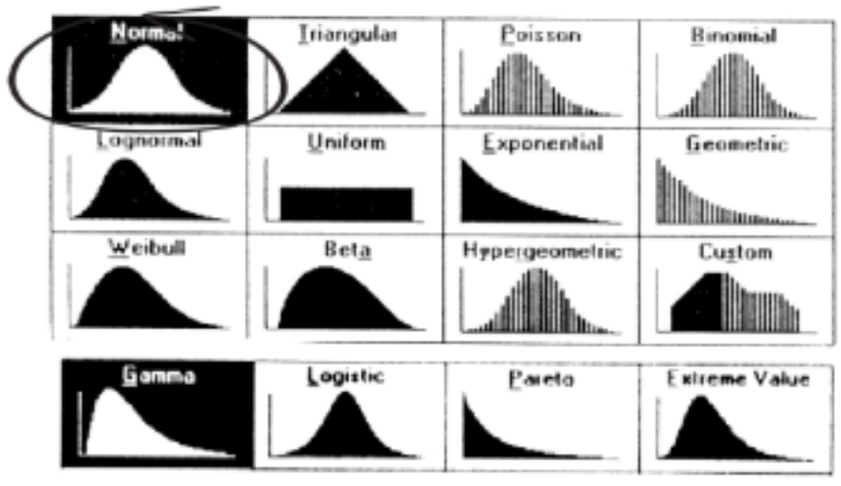
IS A DISTRIBUTION THAT HAS CONSTANT PROBABILITY

RANDOM VARIABLES CAN BE **DISCRETE** OR **CONTINUOUS**

AND SO CAN THEIR PROBABILITY DISTRIBUTIONS



STATISTICIANS AND MATHEMATICIANS HAVE STUDIED A LOT OF DIFFERENT RANDOM VARIABLES IN NATURE AND REALIZED THAT THERE ARE SOME RECURRING THEMES



THEY HAVE DEFINED SOME STANDARD DISTRIBUTIONS AND MOST RANDOM VARIABLES THAT YOU WOULD EVER ENCOUNTER WOULD FALL INTO ONE OF THESE DISTRIBUTIONS

LET'S TALK ABOUT ONE OF THESE THAT IS VERY COMMONLY SEEN IN MANY INSTANCES OF MACHINE LEARNING AND STATISTICAL PROBLEMS

THE NORMAL DISTRIBUTION

THIS IS A DISTRIBUTION PATTERN THAT HAS BEEN SEEN TO OCCUR IN MANY NATURAL PHENOMENA

HEIGHT OF A PERSON,
BLOOD PRESSURE,
LENGTHS OF OBJECTS PRODUCED BY MACHINES,
PERFORMANCE OF STUDENTS IN A CLASS

SAY YOU ARE DOING A HIGH SCHOOL SCIENCE EXPERIMENT

MEASURING THE DIAMETER OF A BALL BEARING USING SOME CALLIPERS

YOU WOULD TAKE A NUMBER OF MEASUREMENTS

AND THEN USE THE AVERAGE AS THE DIAMETER OF THE BALL BEARING

A MEASUREMENT IS A RANDOM VARIABLE

MOST OF THE MEASUREMENTS WOULD BE CONCENTRATED NEAR A CENTRAL POINT

THESE MEASUREMENTS ARE DRAWN FROM A PROBABILITY DISTRIBUTION THAT LOOKS LIKE THIS

THE MEAN AND STANDARD DEVIATION ARE SUFFICIENT TO COMPLETELY DESCRIBE A NORMAL DISTRIBUTION

GIVEN THESE 2 NUMBERS YOU CAN LOOK UP THE PROBABILITY USING STANDARD TABLES

THE MEAN OR AVERAGE VALUE

THE SPREAD OF THE DISTRIBUTION IS DESCRIBED BY A PARAMETER CALLED

STANDARD DEVIATION