CLUSTERING

GIVEN A SET OF INSTANCES (ALL FACEBOOK USERS)

DIVIDE THOSE INSTANCES INTO CLUSTERS, (DISJOINT COMMUNITIES OF FACEBOOK USERS)
SO THAT INSTANCES WITHIN A CLUSTER ARE
MORE SIMILAR TO EACH OTHER THAN TO
INSTANCES IN OTHER CLUSTERS

CLUSTERING IS VERY CLOSELY RELATED
TO CLASSIFICATION - BOTH CLUSTERING AND CLASSIFICATION
DIVIDE A SET OF INSTANCES INTO
DISJOINT GROUPS

CLASSIFICATION IS A BIT MORE FOCUSED
ON CLASSIFYING A PROBLEM INSTANCE (A NEW USER HAS SIGNED UP -
WHAT COMMUNITY WILL SHE MOST
LIKELY BELONG TO?)

CLUSTERING ON THE OTHER HAND
IS LARGELY FOCUSED ON THE PROCESS
OF DIVVYING UP THE INSTANCES WE
ALREADY HAVE

CLUSTERING IS A PROTOTYPICAL
EXAMPLE OF

UNSUPERVISED LEARNING

CLUSTERING ALGORITHMS

K-MEANS CLUSTERING

HIERARCHICAL CLUSTERING

DENSITY-BASED CLUSTERING

DISTRIBUTION-BASED CLUSTERING

IMAGINE THAT WE ARE RESEARCHERS
IN A NATIONAL PARK IN THE AFRICAN SAVANNA

OUR AIM IS TO DIVIDE ALL OF
THE SPECIES OF ANIMALS INTO
CLUSTERS

TO START WITH, SAY WE ASSUME THAT ALL
ANIMALS OF ALL TYPES BELONG TO A SINGLE
CLUSTER (THE ALL ANIMALS CLUSTER)

THEN, WE SEE THAT THERE ARE SOME
OBVIOUS DISTINCTIONS - BIRDS, MAMMALS,
REPTILES, INSECTS -

WE DIVIDE THE ALL CLUSTER INTO
4 CLUSTERS FOR EACH OF THESE TYPES

NEXT IT STRIKES US THAT THERE ARE
OBVIOUS DISTINCTIONS WITHIN MAMMALS -
WE DIVIDE INTO PREDATORS AND PREY

..AND KEEP GOING THIS WAY UNTIL
NO MORE OBVIOUS DISTINCTIONS ARE
APPARENT

THIS IS EXACTLY HOW

HIERARCHICAL CLUSTERING

WORKS - THIS SPECIFICALLY
IS CALLED TOP-DOWN (OR DIVISIVE)
HIERARCHICAL CLUSTERING

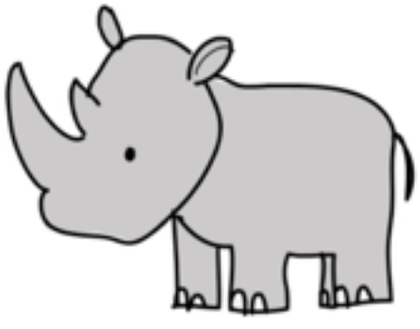
THE KEY TO THE ACTUAL IMPLEMENTATION
OF A TOP-DOWN HIERARCHICAL CLUSTERING
IS A SIMILARITY FUNCTION

A SIMILARITY FUNCTION COULD BE THE INVERSE
OF THE SUM OF ALL INTRA-CLUSTER DIFFERENCES
(THE FURTHER APART THE POINTS IN A CLUSTER
ARE, THE LESS SIMILAR THE CLUSTER)

SPLITTING A CLUSTER MAKES SENSE
ONLY IF THE NEW CLUSTERS ARE IN SOME
SENSE "TIGHTER" THAN THE PREVIOUS
CLUSTER

LET'S SAY THAT WE SOMEHOW KNEW –
MAGICALLY – THAT THERE ARE 3 PROTOTYPICAL
MAMMALS IN THE AFRICAN SAVANNA

LIONS



RHINOS

ZEBRAS

WE COULD THEN TAKE THESE
ANIMALS AS THREE "CENTROIDS"
IN OUR HYPERPLANE OF ANIMALS –

AND CONSTRUCT 3 CLUSTERS AROUND
THESE 3 CENTROIDS

THE LION CLUSTER
THE RHINO CLUSTER
THE ZEBRA CLUSTER

ANY NEW ANIMAL WOULD
SIMPLY GO TO THE CLUSTER
WHERE IT IS CLOSEST TO THE
CENTROID

THE WHOLE DIFFICULTY THEN LIES
IN IDENTIFYING THE 3 ANIMALS TO
USE AS CENTROIDS

"WHY NOT AN ELEPHANT CLUSTER
RATHER THAN A RHINO CLUSTER?"

THIS IS THE

K-MEANS CLUSTERING ALGORITHM

AT WORK

DENSITY-BASED CLUSTERING

RELIES ON THE IDEA THAT POINTS ARE
DENSELY CONCENTRATED INSIDE CLUSTERS,
BUT SPARSE AND FAR APART BETWEEN CLUSTERS

THE ALGORITHM SCANS THE DATA LOOKING
FOR DROPS IN DENSITY, AND MARKS THESE
DROPS AS CLUSTER BOUNDARIES

POINTS ARE THEN ASSIGNED TO
THE CLUSTER WHOSE BOUNDARIES
THEY LIE WITHIN

DISTRIBUTION-BASED CLUSTERING

LET'S SAY WE KNEW FOR EACH POINT,
WHAT PROBABILITY DISTRIBUTION FUNCTION
IT WAS (MOST LIKELY) DRAWN FROM

"BOYS HEIGHTS ARE NORMALLY DISTRIBUTED
WITH MEAN OF 175CMS, AND STANDARD
DEVIATION OF 3 CMS"

"GIRLS HEIGHTS ARE NORMALLY DISTRIBUTED
WITH MEAN OF 165CMS, AND STANDARD
DEVIATION OF 3 CMS"

THEN, WE COULD HAVE ONE CLUSTER FOR
EACH PROBABILITY DISTRIBUTION, AND
ASSIGN POINTS TO THE DISTRIBUTION THAT
THEY (MOST LIKELY) WERE DRAWN FROM

(MAYBE OUR POINTS ALSO INCLUDE
RHINOS, DRAWN FROM A HEIGHT
DISTRIBUTION MEAN = 150 CMS AND
STANDARD DEVIATION = 20 CMS)

THE KEY DIFFICULTY IS IN KNOWING THE
NUMBER AND PROPERTIES OF THOSE
PROBABILITY DISTRIBUTIONS