Introduction to cluster analysis

Lecture 18

What is Cluster Analysis?

Finding groups of objects such that the objects in each group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

Labeling objects with group label

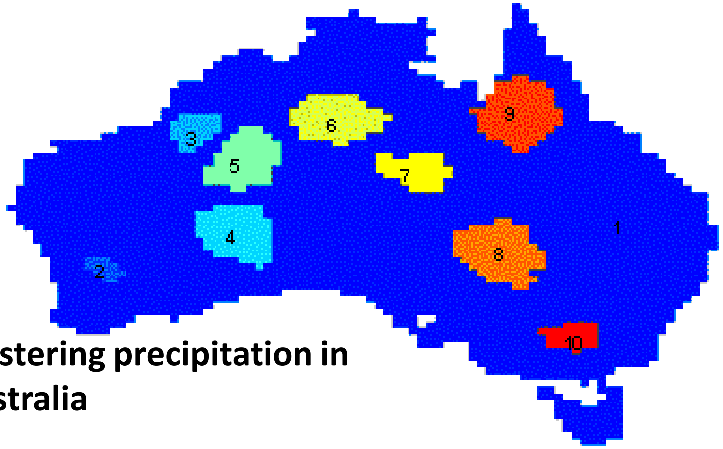
- Classes – conceptually meaningful groups of objects that share common characteristics
- Humans are skilled at dividing objects into groups (clustering) and assigning new objects to one of the groups (classification)
- Clusters are potential classes, and cluster analysis it a technique for automatically discovering classes from data

Applications of Cluster Analysis

- **Clustering for Understanding**
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
  - Segment customers into a small number of groups for additional analysis and marketing activities.

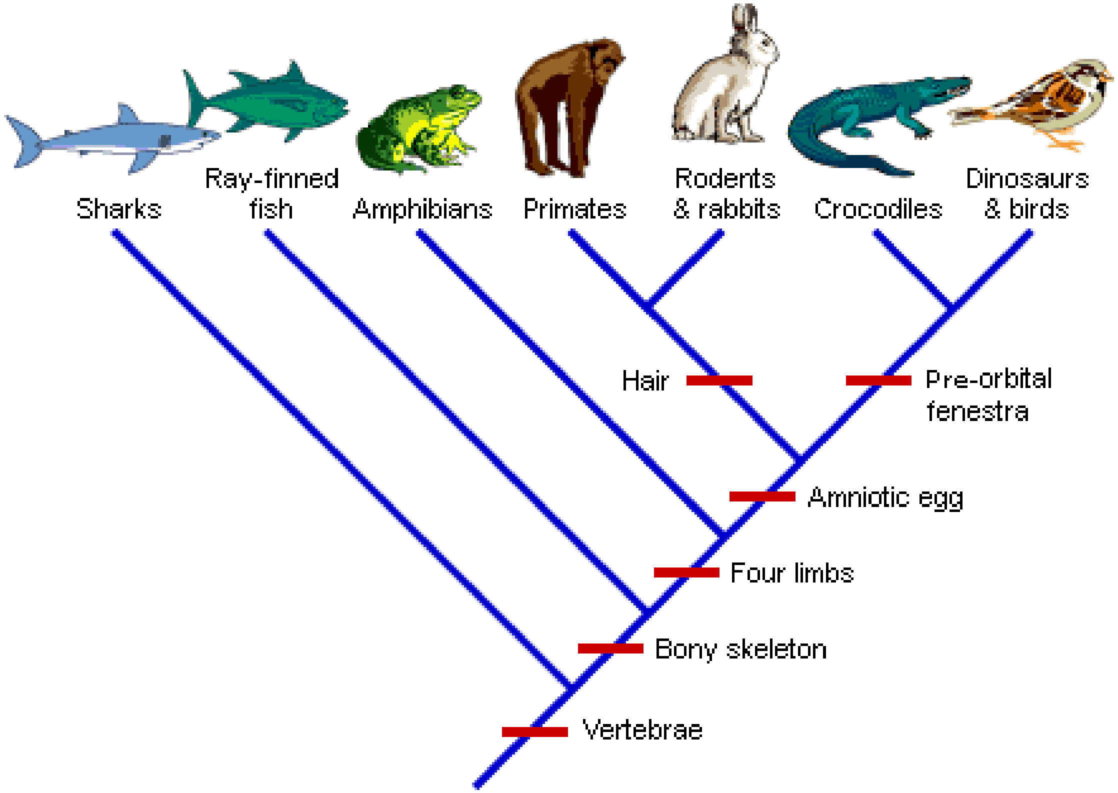
- **Clustering for Summarization**
  - Reduce the size of large data sets

	Discovered Clusters	***Industry Group***
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP



Clustering precipitation in Australia

Grouping animals into clusters: biological systematics

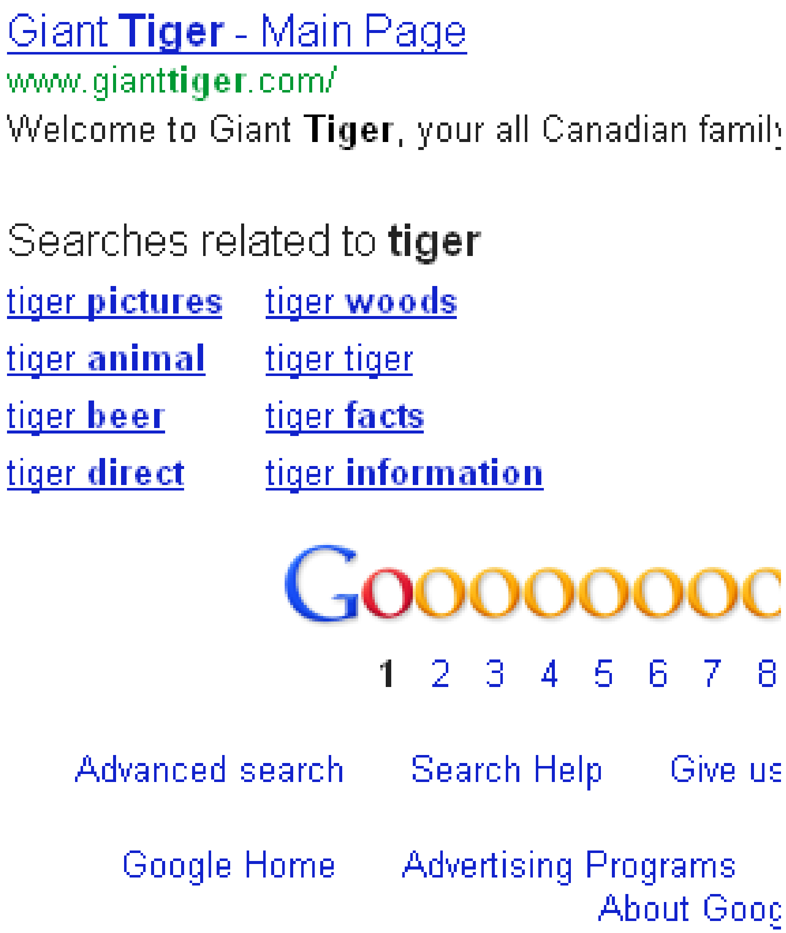


- Grouping animals into hierarchical groups to better understand evolution

Grouping documents into clusters: information retrieval

- Grouping WEB query results into small number of clusters, each capturing a particular aspect of a query

Search for *tiger*



Giant **Tiger** - Main Page
www.giant**tiger**.com/
Welcome to Giant **Tiger**, your all Canadian famil

Searches related to **tiger**

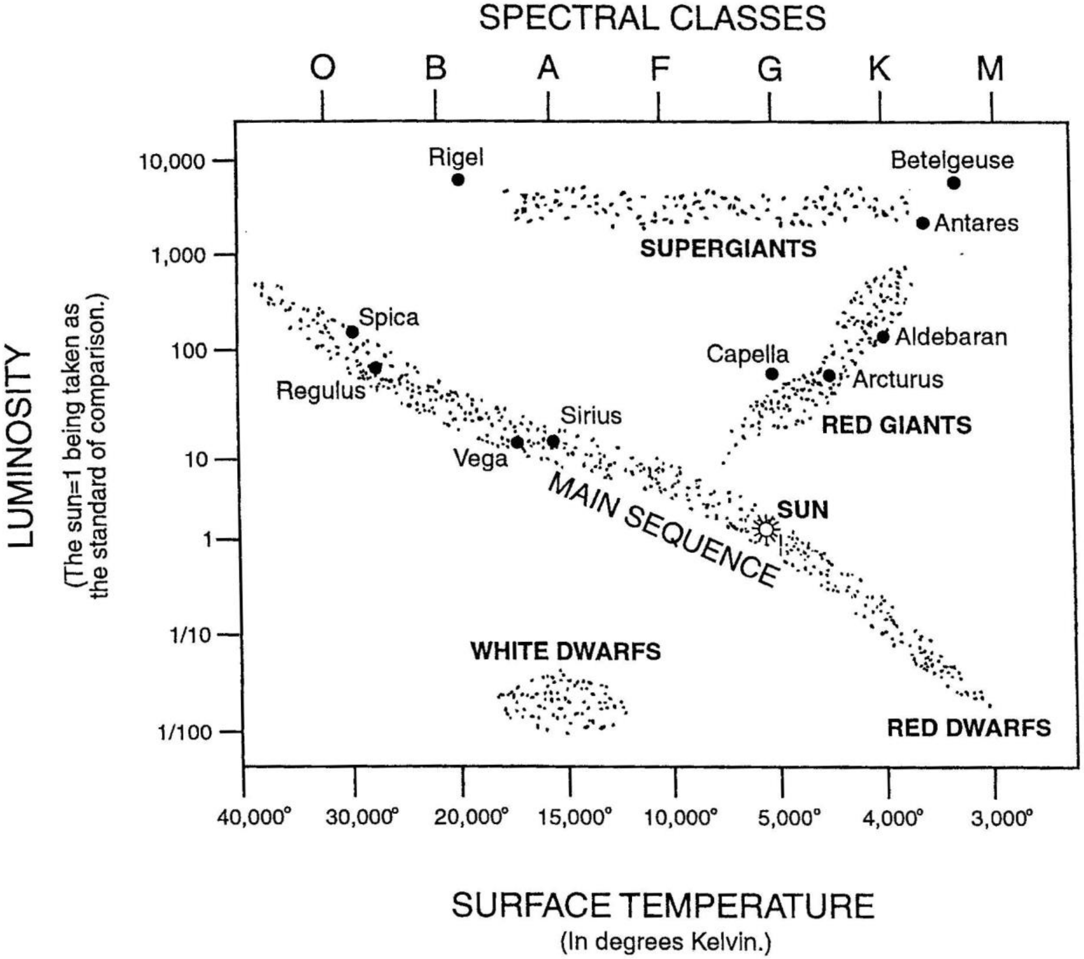
tiger **pictures**
tiger **animal**
tiger **beer**
tiger **direct**
tiger **woods**
tiger tiger
tiger **facts**
tiger **information**

1 2 3 4 5 6 7 8

Advanced search Search Help Give us

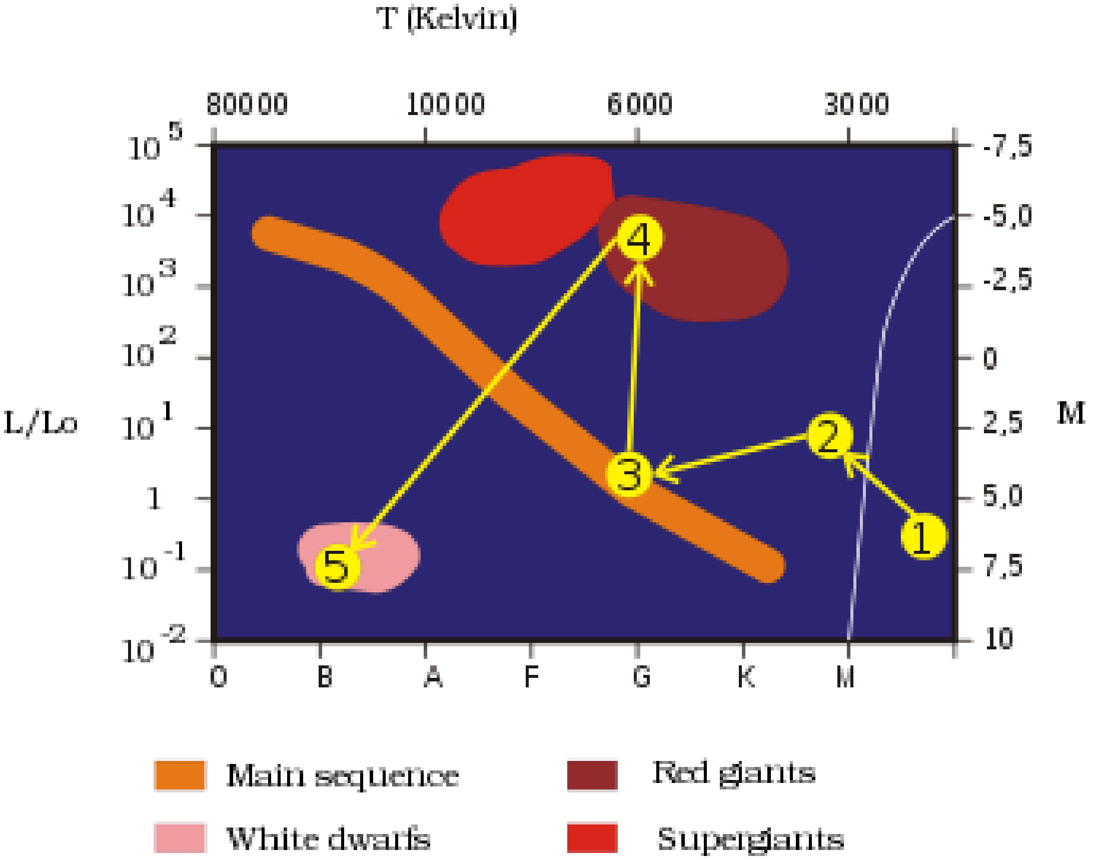
Google Home Advertising Programs
About Goog

Clustering leads to discoveries: Galaxies evolution



The Hertzsprung-Russel diagram clusters stars by temperature and luminosity

Clustering leads to discoveries: Galaxies evolution



Main sequence stars generate energy by fusing Hydrogen to Helium

When the hydrogen is used up, Helium fusion occurs, the star expands - > red giant

The outer layer of gases is stripped away, the star cools -> white dwarf

Automatic clustering

- Discovering groups (classes) of objects from unlabeled data
- *Unsupervised learning*

Formulation for computer program

- Computer asks:
  - What is similarity (dissimilarity=distance)?
  - What exactly am I looking for?

- We need to:
  - define similarity as a numeric value
  - define the notion of a cluster

  - prescribe the algorithm for finding defined clusters

Numeric *proximity* (similarity or distance) between data records

- Combination of proximity measures for each attribute
- Each attribute is a separate and independent (in this approach) dimension of the data

- First step: translate all fields into numeric variables, to make similarities (distances) numeric

Types of attributes

1. True measures (continuous)
2. Ranks (ordinal)
3. Categorical (nominal)

The distances are
Increasingly harder
to convert into a
numeric scale

How do we define the proximity measure for a single attribute of each type?

1. True measures

- True measures measure the value from a meaningful “0” point. The ratio between values is meaningful, and the distance is just an absolute difference of values.

- Examples: age, weight, length

2. Ordinal (Ranks)

- These values have an order, but the distance between different ranks is not defined

2. Ordinal (Ranks)

Example 1:

quality attribute of a product : {poor, fair, OK, good, wonderful}

Order is important, but exact difference between values is undefined

Solution: map the values of the attribute to successive integers

{poor=0, fair=1, OK=2, good=3, wonderful=4}

Dissimilarity

$d(p,q) = |p - q| / (\text{max\_d} - \text{min\_d})$

e.g. d(wonderful, fair) = |4-1| / (4-0) = .75

Not always meaningful, but the best we can do

Similarity

$s(p,q) = 1 - d(p,q)$ **e.g.** d(wonderful, fair) = .25

2. Ordinal (Ranks)

Example 2:

Top 10 swimmers - 50m Fly

1	KONOVALOV, Nikita	88	RUS	22.70
2	GOVOROV, Andriy	92	UKR	22.70
3	LEVEAUX, Amaury	85	FRA	22.74
4	CZERNIAK, Konrad	89	POL	22.77
5	KOROTYSHKIN, Evgeny	83	RUS	22.88
6	EIBLER, Steffen	87	GER	22.89
7	FESIKOV, Sergey	89	RUS	22.96
8	HEERSBRANDT, Francois	89	BEL	22.98
9	MUNOZ PEREZ, Rafael	88	ESP	23.07
10	JAMES, Antony	89	GBR	23.14

Distance between 3 and 1 (0.04 sec) is not the same as distance between 10 and 8 (0.16). It is better to use the attributes which contributed to this ranking

3. Categorical (nominal) attributes

- Each value is one of a set of unordered categories. We can only tell that X≠Y, but not how much X is greater than Y.

- Example: ice cream pistachio is not equal to butter pecan, but we cannot tell which one is greater and which one is closer to black cherry ice cream
- The general approach: if equal then similarity =1, if not equal then similarity = 0

Summary on proximity measures for a single attribute

Attribute type	Distance (dissimilarity)	Similarity
True measures	$d=\lvert x-y\rvert$	$s=-d$, $\boxed{s=1/(1+d),}$ $s=1-(d-\text{min\_d})/(\text{max\_d}-\text{min\_d})$
Ordinal	$d=\lvert x-y\rvert/(n-1)$ (values mapped to integers 0 to n-1 where n is the number of values)	$s=1-d$
Nominal	$d=0$ if $x=y$ $d=1$ if $x\neq y$	$s=1$ if $x=y$ $s=0$ if $x\neq y$

Combining measures for single attributes into proximity measure for data records

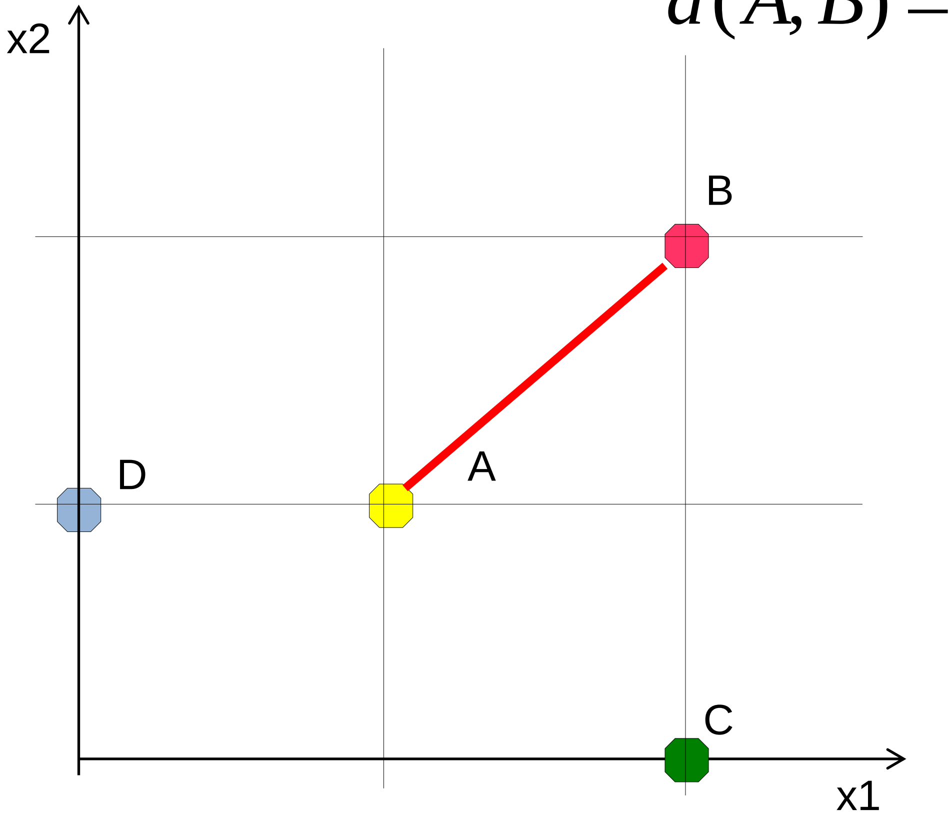
- Hundreds of similarity measures were proposed. We will look at:
  - Euclidean distance
  - Jaccard index
  - Cosine similarity

Proximity measures for data records

1. Euclidean distance (all attributes are numeric)
2. Matching coefficients (all attributes are binary– from categorical attributes transformed into binary)
3. Cosine similarity (true values as vectors)

All attributes are numeric: Euclidean distance

$$d(A,B)=\sqrt{|A_X-B_X|^2+|A_Y-B_Y|^2}$$



For N dimensions:

$$d(A,B)=\sqrt{\sum_{i=1}^{N}|A_i-B_i|^2}$$

Similarity:

$s(A,B)=1/(1+d(A,B))$

It is hard to visualize points in more than 3 dimensions, but for computer it is not the problem.

Transformed attributes

- Be careful with non true measures transformed into numeric attributes
- If gender and college degree are among the variables, then, all other variables being roughly the same, a woman with college degree will be very similar to a man without college degree (because female gives a 0 and male gives a 1, whereas college is 1 and not college is 0)

2. Matching coefficients (all attributes are binary)

	Y	Y
X	M_{11}	M_{10}
X	M_{01}	M_{00}

M_{11}: number of attributes with value 1 in both X and Y
M_{10}: number of attributes with value 1 in X and 0 in Y
M_{01}: number of attributes with value 0 in X but 1 in Y
M_{00}: number of attributes with value 0 in both X and Y

2. Matching coefficients

	Y	Y
X	M_{11}	M_{10}
X	M_{01}	M_{00}

Jaccard index is used for
asymmetric binary attributes,
where only value 1 is important

Simple Matching Coefficient

SMC = number of matches / number of attributes

$= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

Jaccard Index

J = number of M_{11} matches / number of not-both-zero attributes values

$= (M_{11}) / (M_{01} + M_{10} + M_{11})$

SMC and Jaccard example

x=(	1	0	0	0	0	0	0	0	0	0)
y=(	0	0	0	0	0	0	1	0	0	1)

SMC= $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00}) = (0+7)/10=0.7$

J= $M_{11} / (M_{01} + M_{10} + M_{11}) = (0)/3=0.0$

The choice is application-dependent.

SMC and Jaccard example

x=(	1	0	0	0	0	0	0	0	0	0)
y=(	0	0	0	0	0	0	1	0	0	1)

SMC= $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00}) = (0+7)/10=0.7$

J= $M_{11} / (M_{01} + M_{10} + M_{11}) = (0)/3=0.0$

The choice is application-dependent.
Which measure to choose for:

Comparing documents by common words?
Comparing transactions by common items?
Comparing students by knowledge of 10 topics?

Cosine similarity

- Sometimes it makes more sense to consider two records closely associated because of similarities in the way the fields *within each record are related*

- Example: sardines should cluster with cod and tuna, while kittens cluster with cougars and lions, but if we use the Euclidean distance of body-part lengths, the sardine is closer to a kitten than it is to a catfish.

Cosine similarity

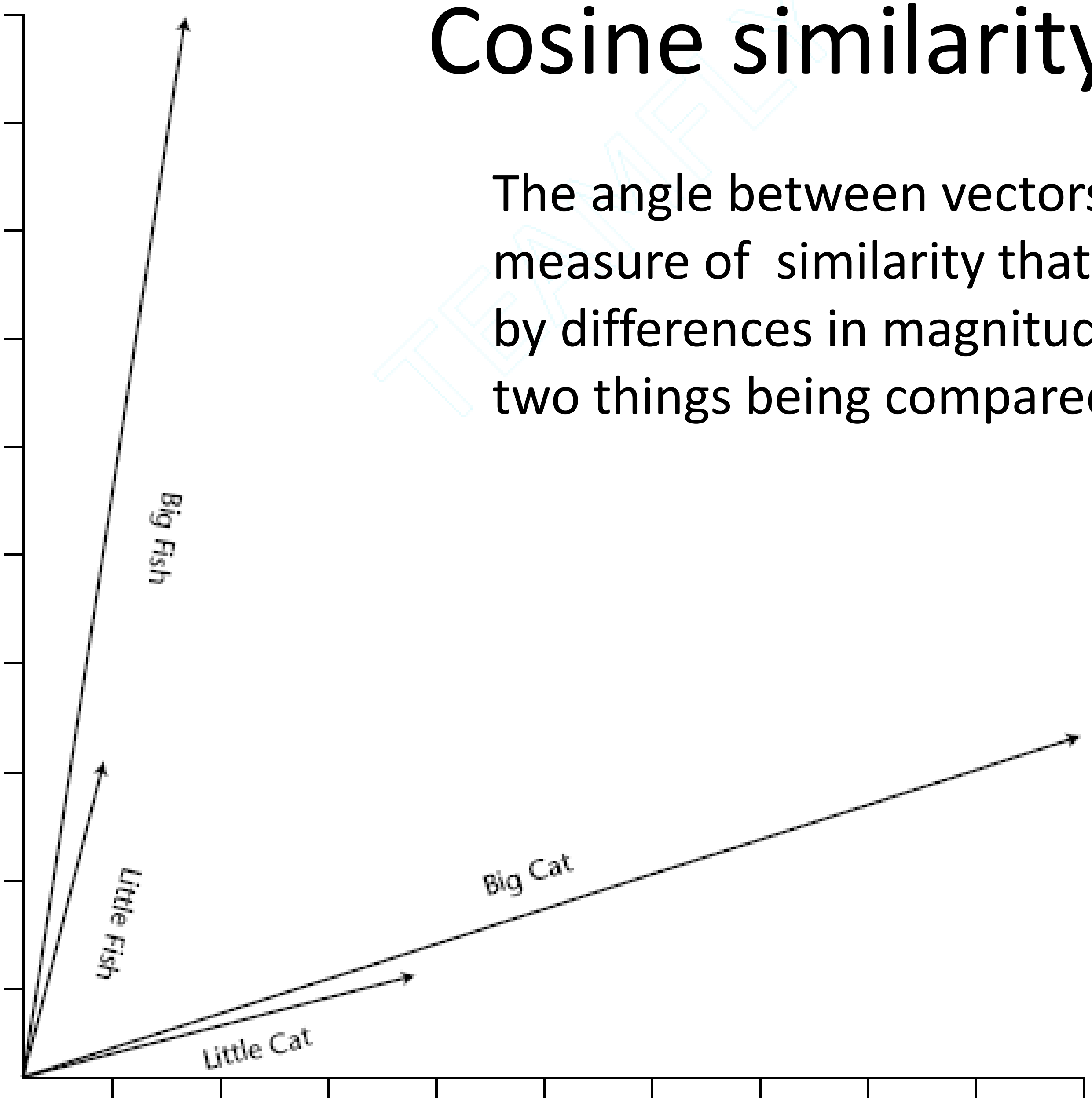
- Sometimes it makes more sense to consider two records closely associated because of similarities in the way the fields *within each record are related*

- Solution: use a different geometric interpretation. Instead of thinking of *X and Y as points in space*, think of them as *vectors and measure the angle between them.*

- In this context, a vector is the line segment connecting the origin of a coordinate system to the point described by the vector values.

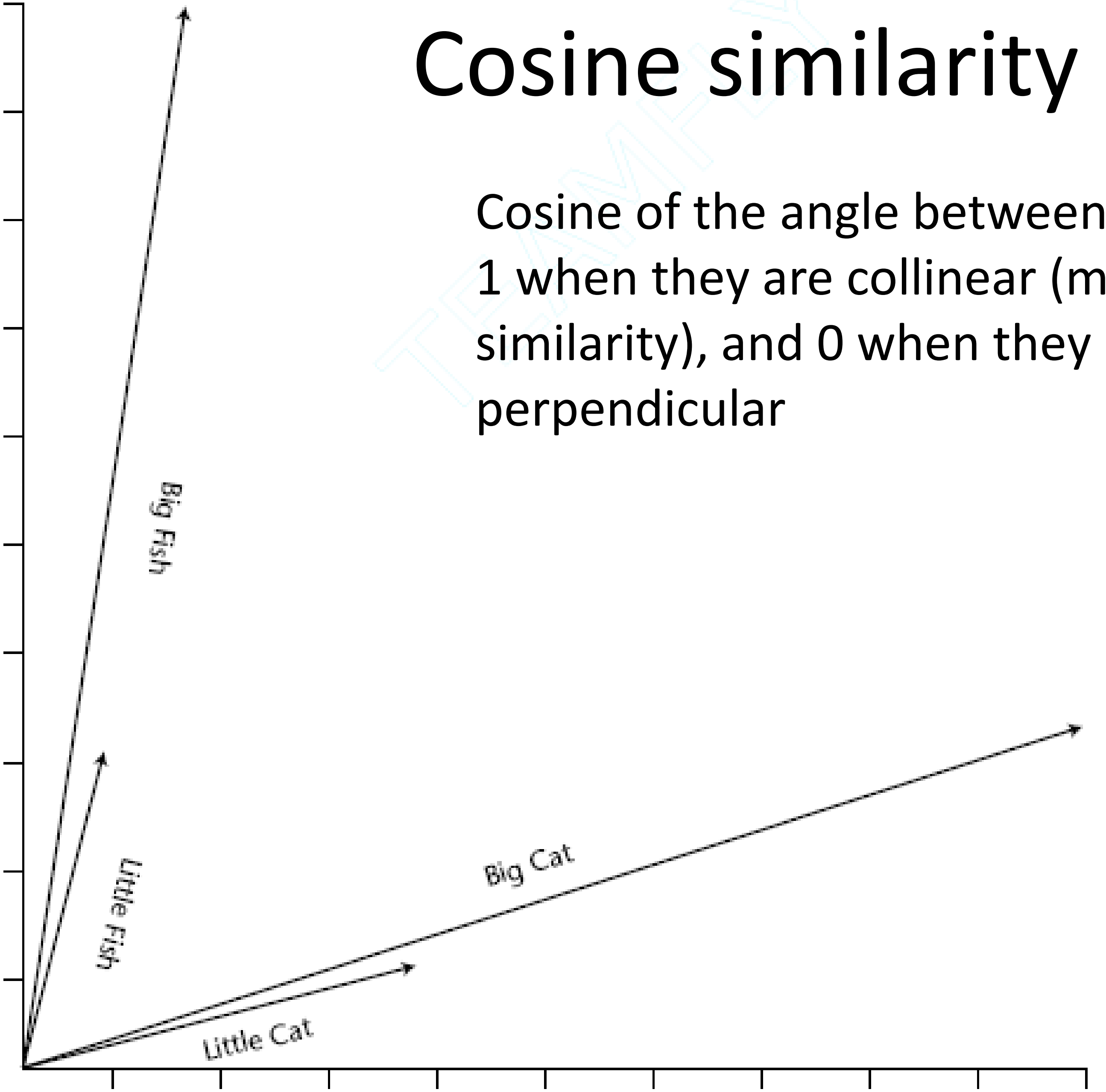
Cosine similarity

The angle between vectors provides a measure of similarity that is not influenced by differences in magnitude between the two things being compared



Cosine similarity

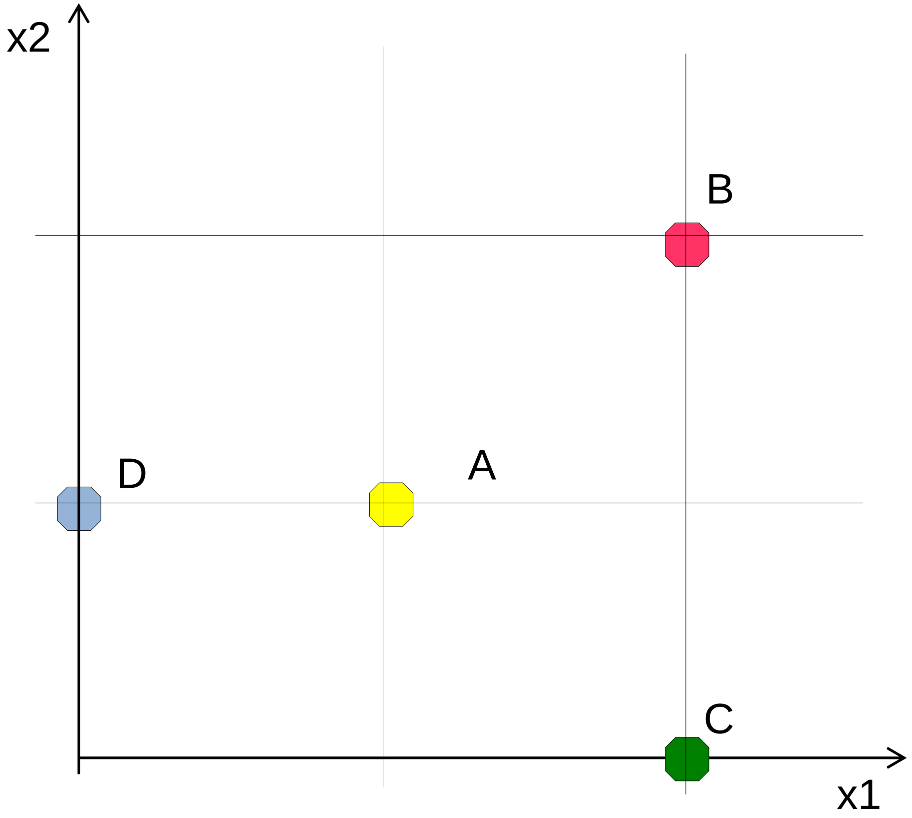
Cosine of the angle between two vectors is 1 when they are collinear (maximum similarity), and 0 when they are perpendicular



Cosine similarity

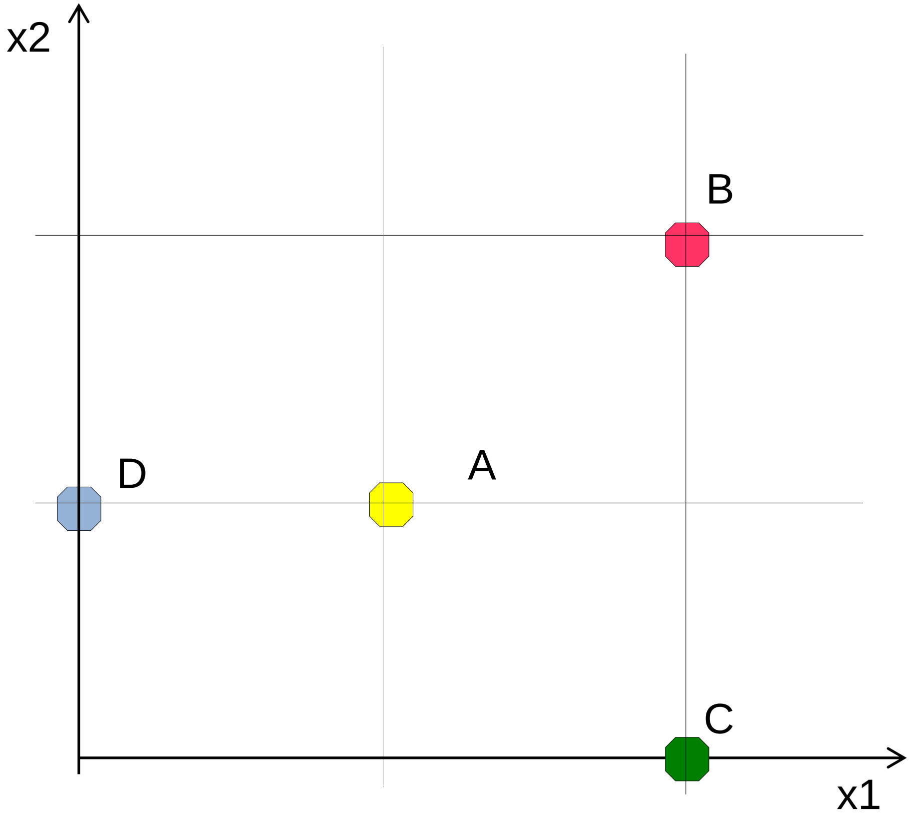
$s(A,B)=cos(\mathbf{A}, \mathbf{B}) = (\mathbf{A} \bullet \mathbf{B}) / ||\mathbf{A}||.||\mathbf{B}||$

Dot-product of vectors



Cosine similarity

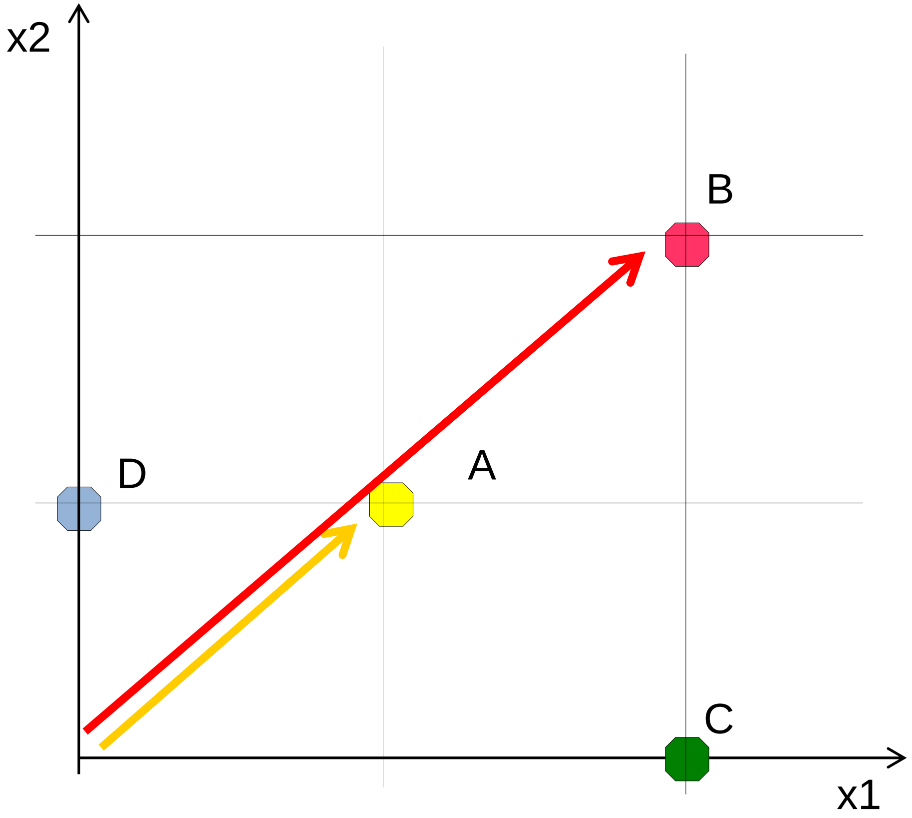
$$s(A,B)=cos(\,\mathbf{A},\,\mathbf{B}\,) = (\mathbf{A} \bullet \mathbf{B}) \,/\, ||\mathbf{A}||.||\mathbf{B}||$$



Absolute length of vector A

Cosine similarity

$s(A,B) = cos(\mathbf{A}, \mathbf{B}) = (\mathbf{A} \bullet \mathbf{B}) / ||\mathbf{A}||.||\mathbf{B}||$



$\mathbf{A}=(1,1)$

$\mathbf{B}=(2,2)$

$\mathbf{A} \bullet \mathbf{B}=1*2+1*2=4$

$||\mathbf{A}||=sqrt(1+1)$

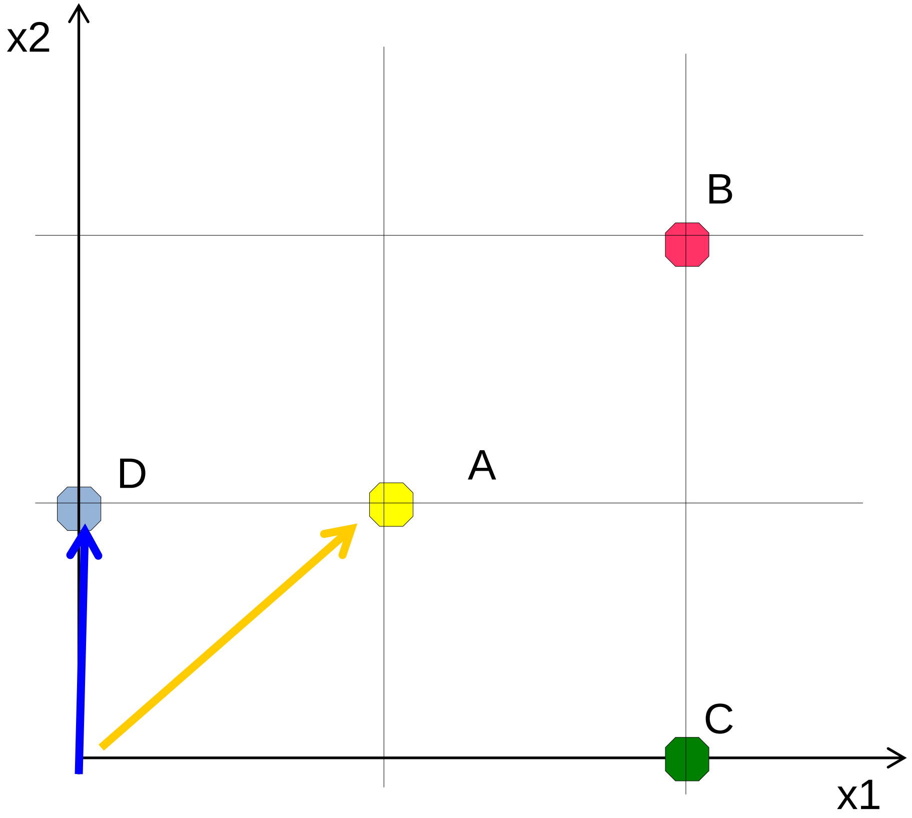
$||\mathbf{B}||=sqrt(4+4)$

$||\mathbf{A}||.||\mathbf{B}||=sqrt(16)$

$s(A,B)=cos(\mathbf{A}, \mathbf{B})=1$

Cosine similarity

$s(\mathrm{A,D})=cos(\ \mathbf{A},\ \mathbf{D}\ ) = \ (\mathbf{A} \bullet \mathbf{D}) \ /\ ||\mathbf{A}||.||\mathbf{D}||$



$\mathbf{A}$=(1,1)

$\mathbf{D}$=(0,1)

$\mathbf{A} \bullet \mathbf{D}$=0+1=1

$||\mathbf{A}\ ||$=sqrt(2)

$||\mathbf{D}\ ||$=1

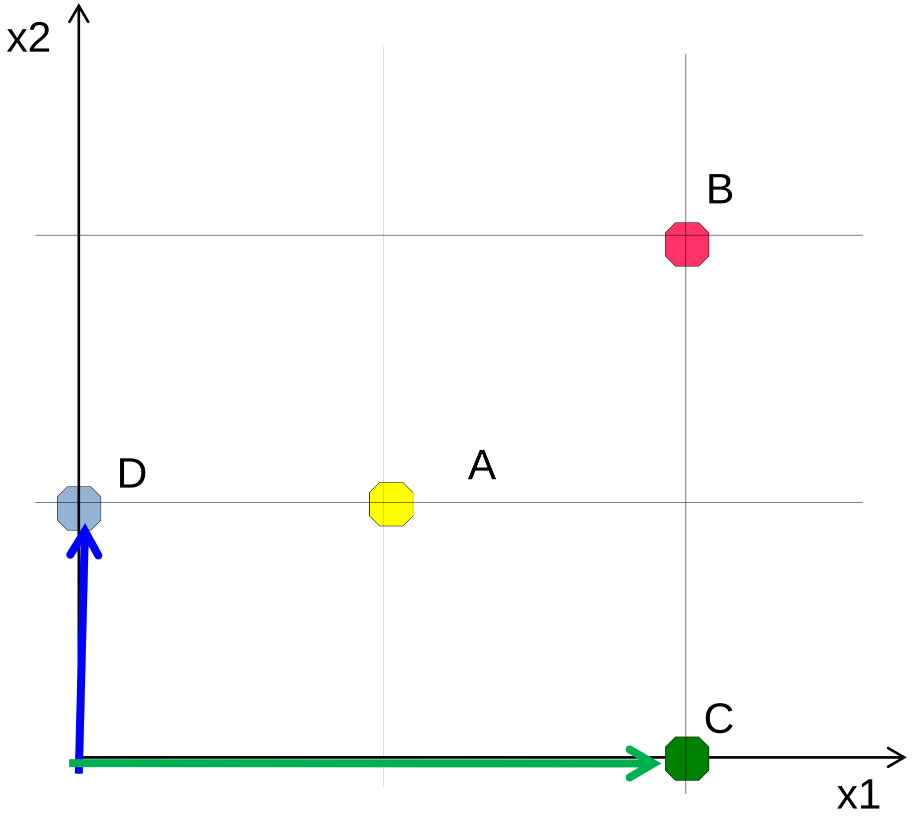
$||\mathbf{A}\ ||.||\mathbf{D}||$=sqrt(2)

$s(\mathrm{A,D})=cos(\ \mathbf{A},\ \mathbf{D}\ )$

=sqrt(1/2)≈0.7

Cosine similarity

$s(\mathrm{C,D}) = cos(\,\mathbf{C},\,\mathbf{D}\,) = (\mathbf{C} \bullet \mathbf{D}) \,/\, ||\mathbf{C}||.||\mathbf{D}||$



$\mathbf{C} = (2,0)$

$\mathbf{D} = (0,1)$

$\mathbf{C} \bullet \mathbf{D} = 0$

$s(\mathrm{C,D}) = cos(\,\mathbf{C},\,\mathbf{D}\,) = 0$

Cosine Similarity for document vectors

	w1	w2	w3	w4	w5	w6	
x=(	1	0	0	0	0	0	)
y=(	0	0	0	1	2	0	)
z=(	0	0	0	4	8	0	)

Cosine between **x** and **y** is 0 (dot-product is 0). These documents are not similar.

Cosine between **y** and **z** is 1: though the number of times each word occurs in y and z is different, these documents are about the same topic

Combining Similarities

- Sometimes attributes are of many different types, but an overall similarity/dissimilarity is needed.
- For the *k*-th attribute, compute a similarity s_k
- Then,

$$\text{similarity}(p,q) = \frac{\sum_{k=1}^{n} s_k}{n}$$

- Similar formula for dissimilarity

Scaling attributes for consistency

- X- in yards, Y in cm
- X- number of children, Y – income

Difference in 1 dollar = difference in 1 child?

Scaling: map all variables to a common range 0-1

Scaling for consistency

$$a_i = \frac{v_i - \min(all\ v)}{\max(all\ v) - \min(all\ v)}$$

For numeric attributes: convert all values into the range between 0 and 1

Scaling vectors

- Vector normalization – changes the vector values so that the length of the vector is 1, only the direction is compared
- X={Debt=200,000 equity=100,000}
- Y={Debt=2,000 equity=1,000}

Emphasizes internal relation between different attributes of each record

Encode expert knowledge with weights

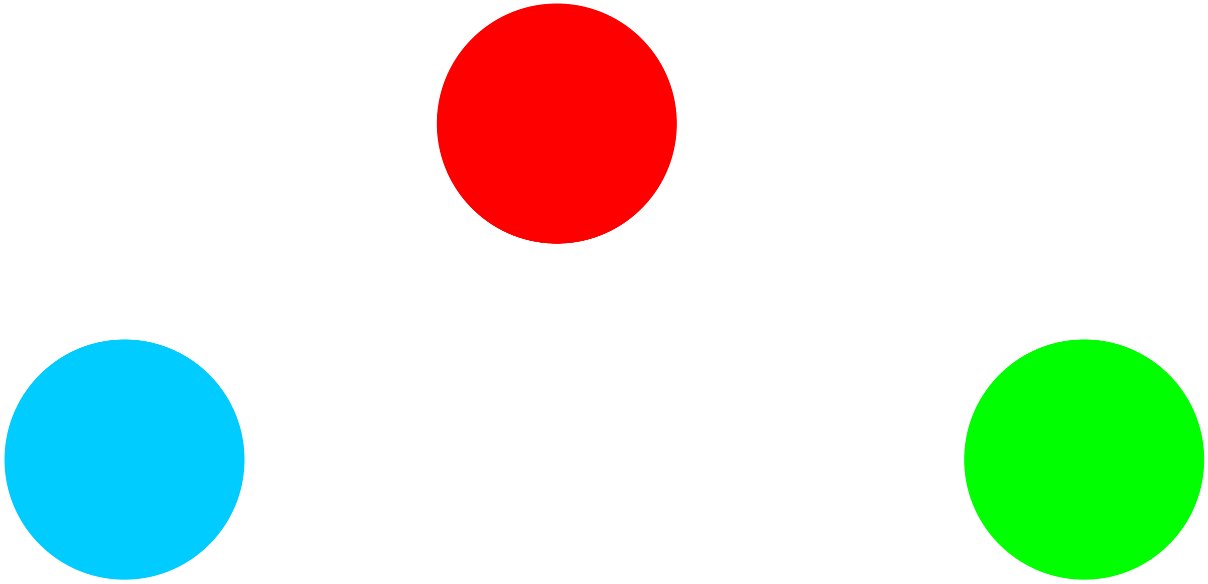
- Changes in one variable should not be more significant only because of differences in magnitudes of values

- After scaling to get rid of bias due to the units, use weights to introduce bias based on expert knowledge of context:
  - 2 families with the same income and number of children are more similar than 2 families living in the same neighborhood
  - Number of children is more important than the number of credit cards

How do we define cluster

Types of Clusters: Well-Separated

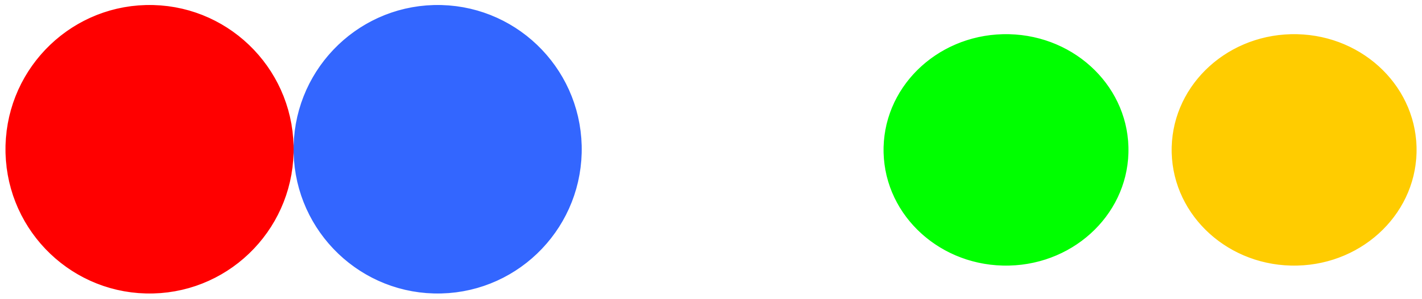
- Any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.



3 well-separated clusters

Types of Clusters: Center-Based

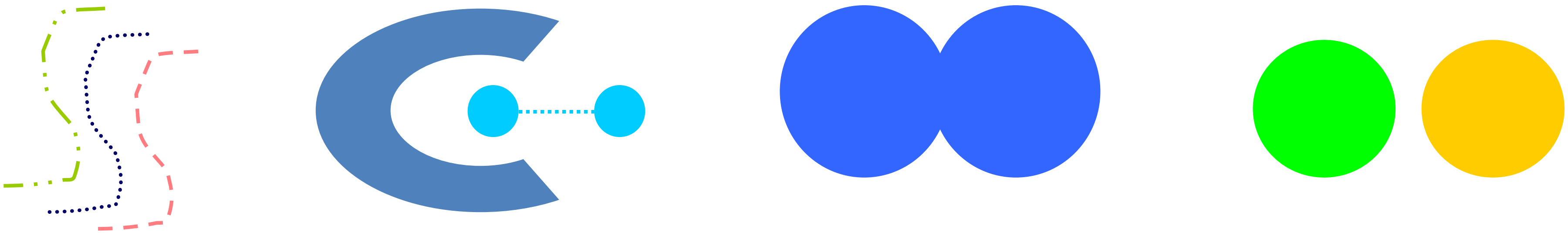
- An object in a cluster is closer (more similar) to the “center” of a cluster, than to the center of any other cluster
- The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most “representative” point of a cluster



4 center-based clusters

Types of Clusters: Contiguity-Based

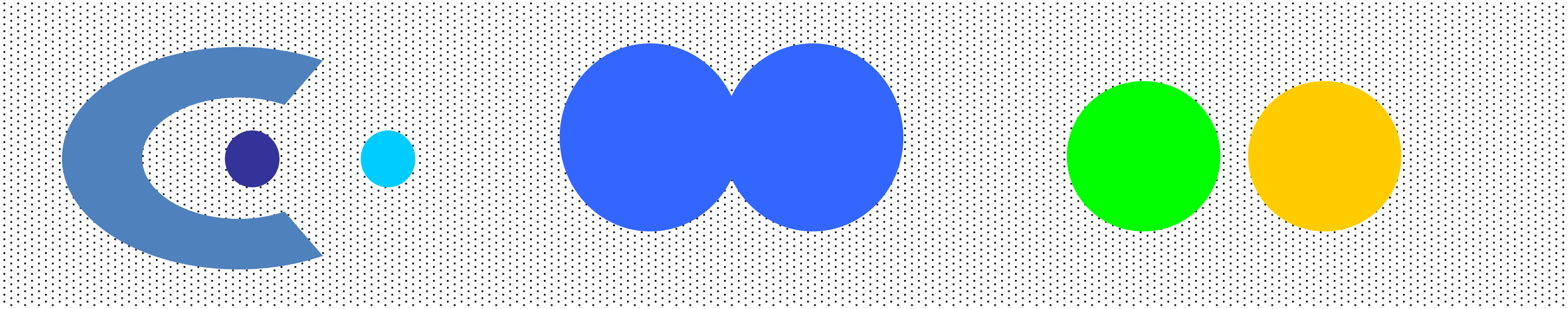
- Contiguous Cluster (Nearest neighbor or Transitive)
- A point in a cluster is closer to at least one point in the cluster than to any point not in the cluster. The group of objects that are connected to one another.



8 contiguous clusters

Types of Clusters: Density-Based

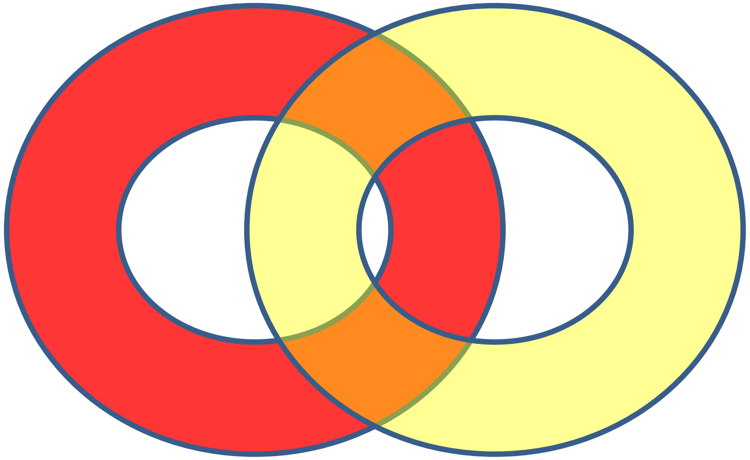
- A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
- Used when the clusters are irregular or intertwined, and when noise and outliers are present.



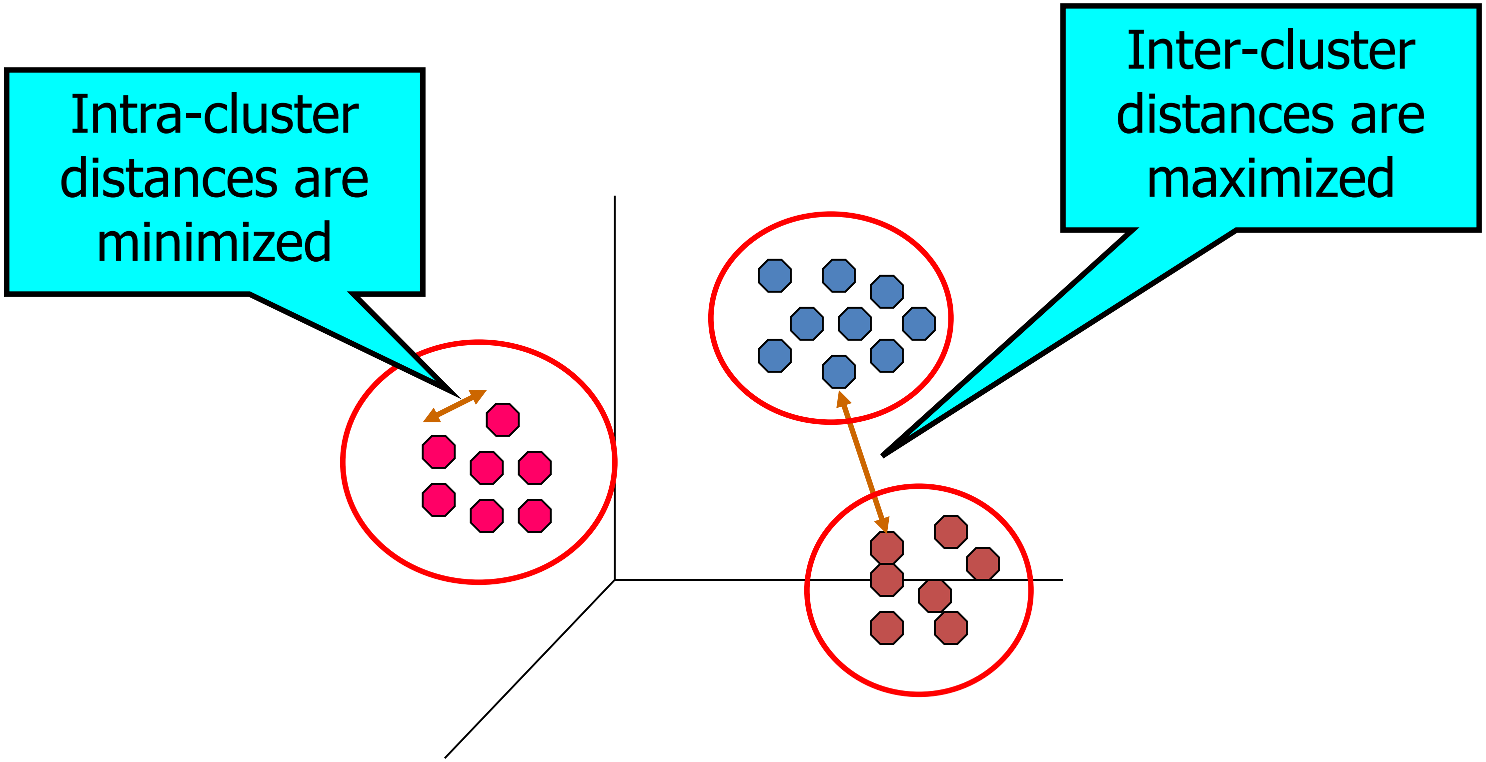
6 density-based clusters

General definition: conceptual clusters

- A set of objects that share some property. This includes all previous cluster types
- In addition it includes clusters defined by a *concept*. Such clusters are used in pattern recognition. To discover such clusters automatically, the concept should be defined first.



Objective of clustering



Intra-cluster distances are minimized

Inter-cluster distances are maximized