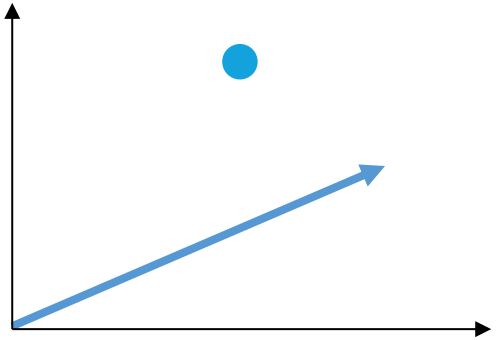
How many features we need?

…just 1?

…maybe 2?

…or 3?



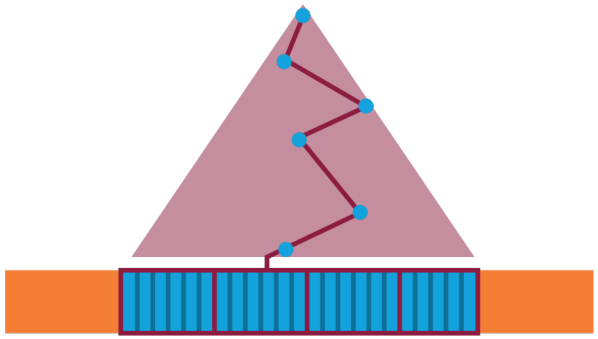
…or, many many more (100s, 1000s) ???

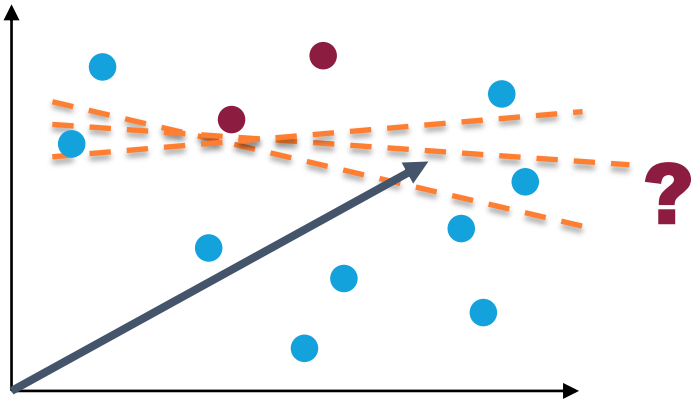
Dimensionality curse!!!

Dimensionality curse: The more dimensions we have, the less efficient and effective search and analysis becomes

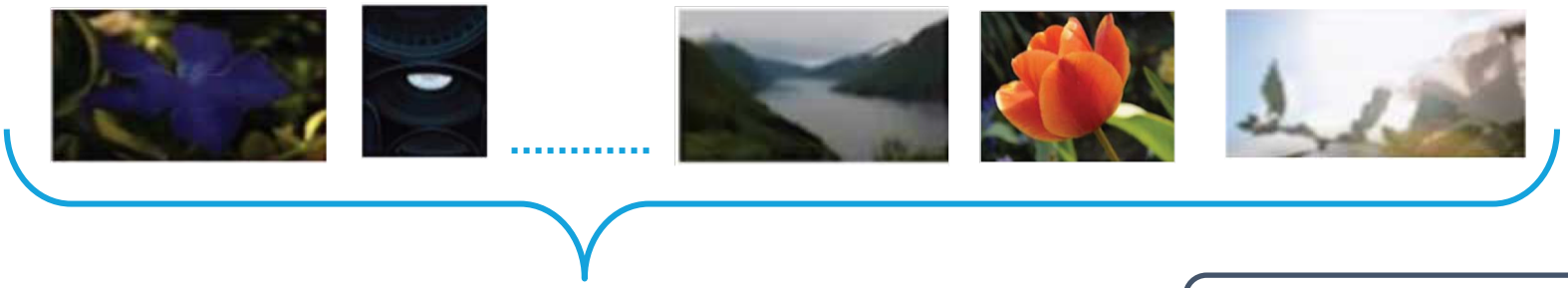
Efficiency: Search data structures are not very efficient at high dimensions

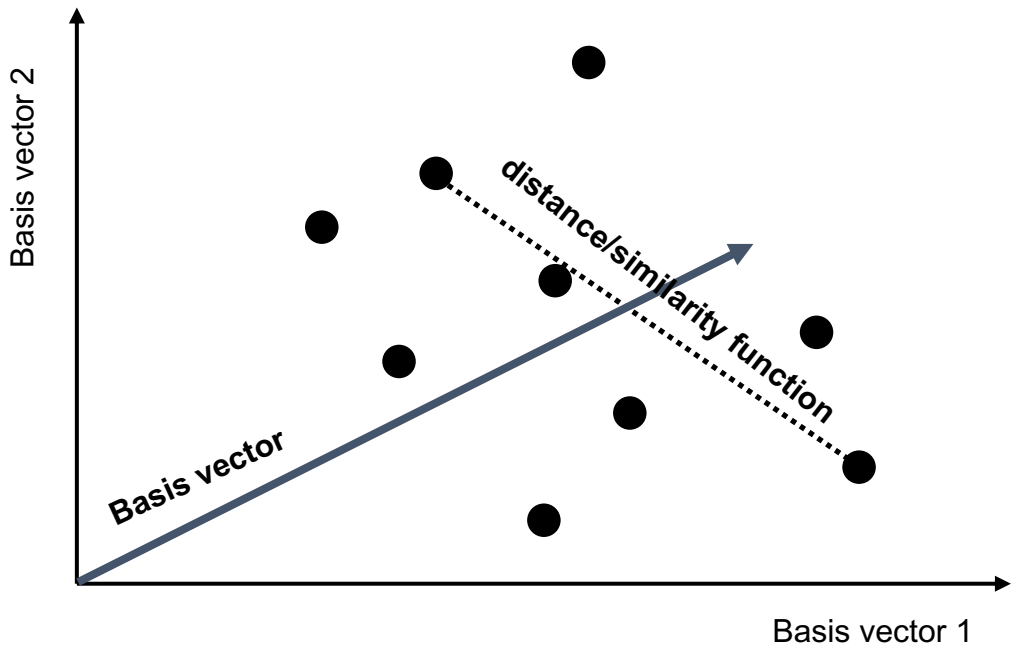
Effectiveness: The more dimensions we have, the more data we need to discover patterns (prevent overfitting)





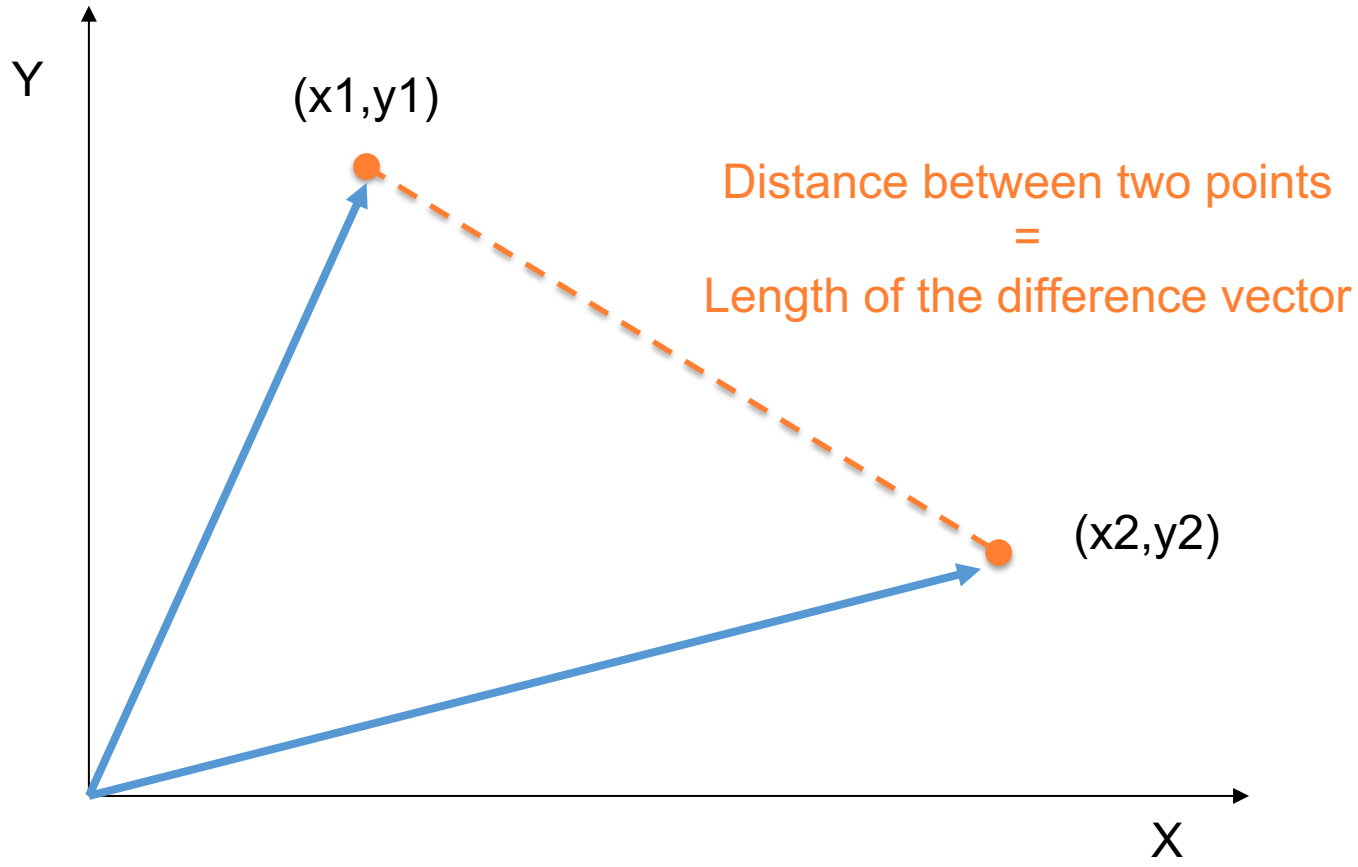
Vector Spaces





- What are good features to use as basis vectors?
- How many features do we need as basis vectors?
- What is a good distance/similarity function?

Distance

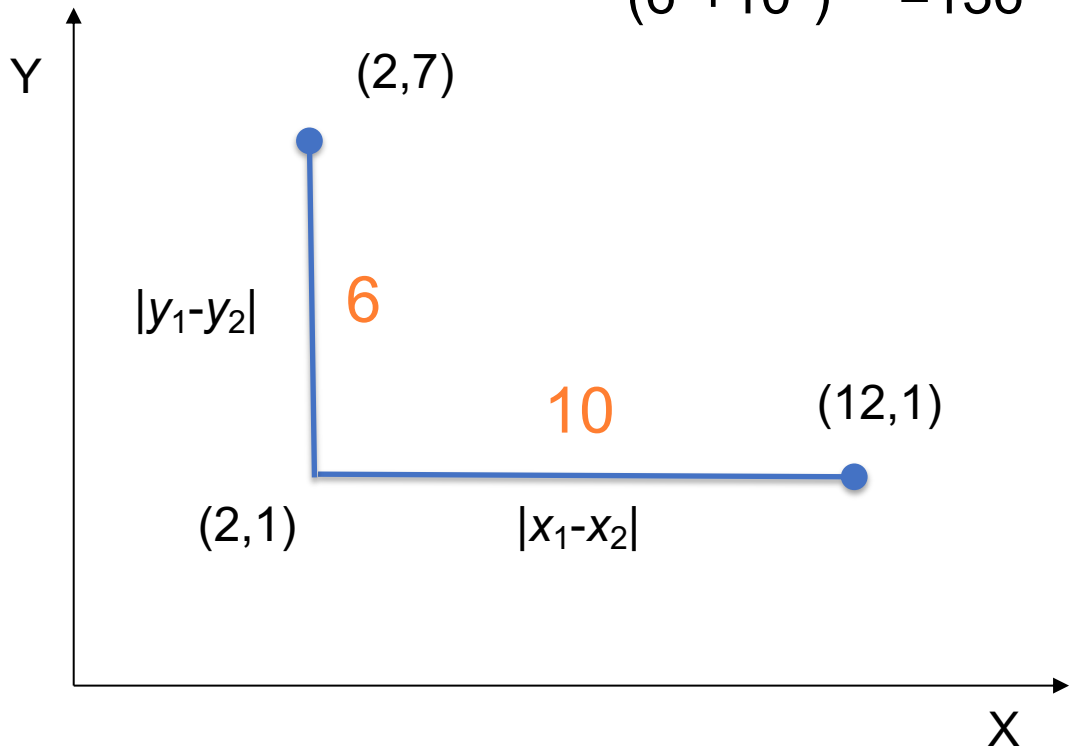

Y
(x1,y1)
Distance between two points
=
Length of the difference vector
(x2,y2)
X

P-Norms

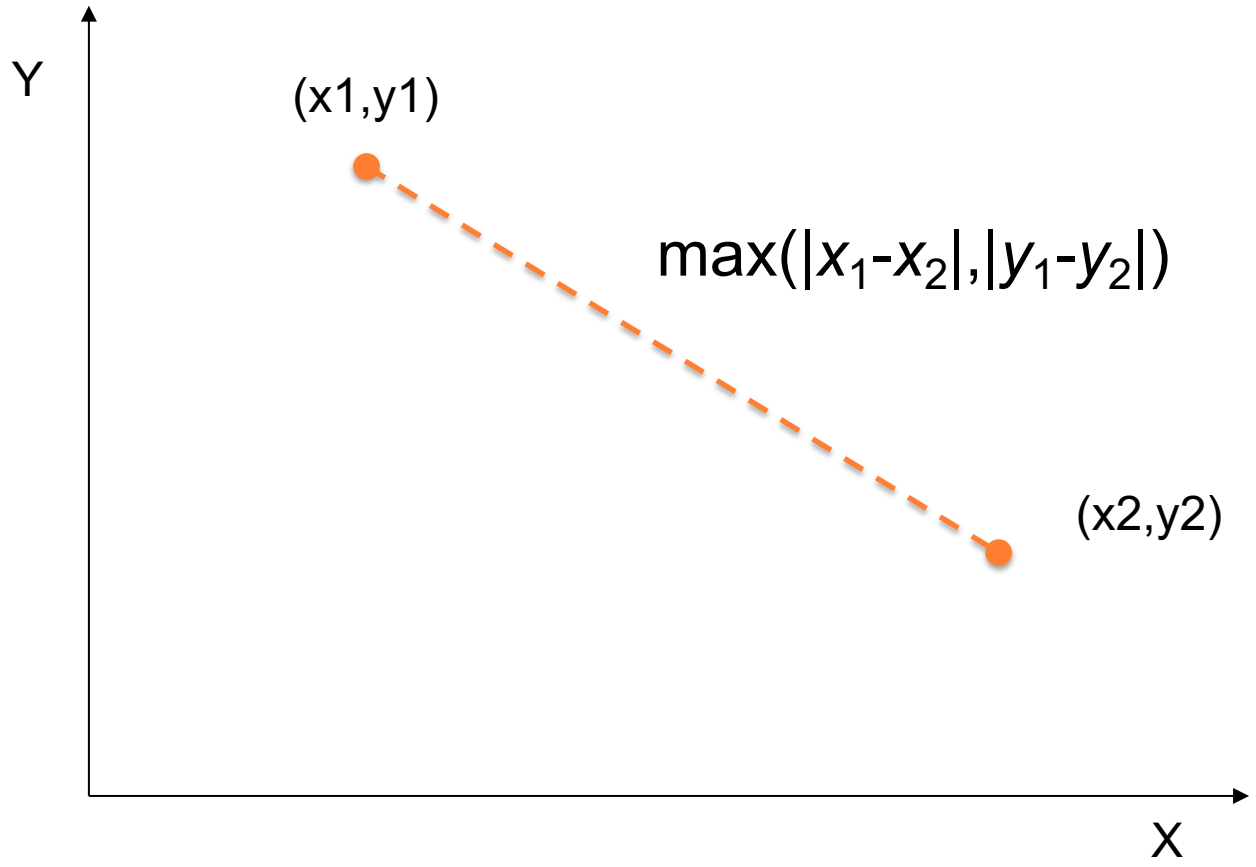
- 1-norm
- 2-norm

$(6+10)=16$

$(6^2+10^2)^{1/2} = 136^{1/2} = 11.66$



∞-norm (L∞ distance)



P-Norms

- 1-norm
- 2-norm
- …
- ∞-norm

$(6+10)=16$

$(6^2+10^2)^{1/2} = 136^{1/2} = 11.66$