UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

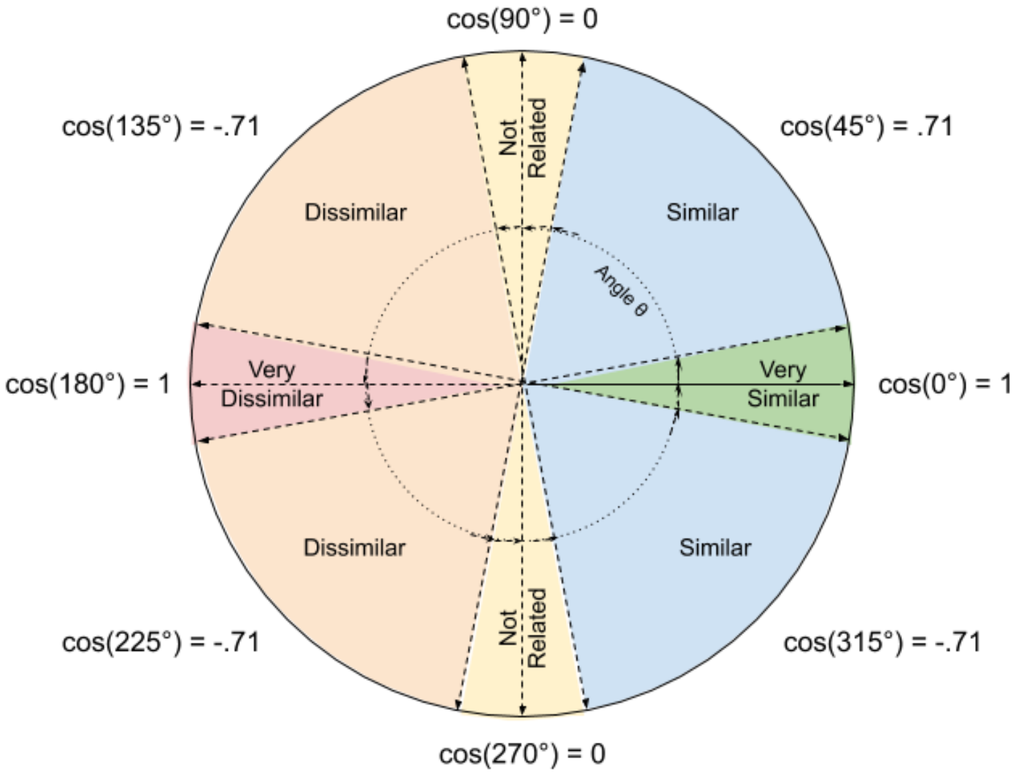
Sesión 10.2: Espacio métrico

CS3102 EDA

Índice

1. Métodos de Acceso Métrico

Similitud *Coseno*



$$d_{\cos}(x, y) = \cos\theta = \frac{x \cdot y}{\|x\|\|y\|}$$

Similitud *Coseno*

¿Similitud coseno es una distancia?

Similitud *Coseno*

$$d_{\cos}(x, y) = \cos\theta = \frac{x \cdot y}{\|x\|\|y\|}$$

Simetría:	$d(x, y) = d(y, x)$
Identidad:	$x = y \leftrightarrow d(x, y) = 0$
Desigualdad triangular:	$d(x, z) \leq d(x, y) + d(y, z)$
No negatividad:	$d(x, y) \geq 0$

Similitud *Coseno*

¿Y ahora que hacemos…?

Similitud *Coseno*

Opción 1: Usamos un equivalente a desigualdad triangular

$$d_{\cos}(x,y) \geq d_{\cos}(x,z)d_{\cos}(z,y) - \sqrt{\left(1 - d_{\cos}^2(x,z)\right) \cdot \left(1 - d_{\cos}^2(z,y)\right)}$$

Opción 2: Angular distance

$$d_\theta(x,y) = \frac{\cos^{-1} d_{\cos}(x,y)}{\pi} = \frac{\theta}{\pi}$$

Similitud *Coseno*

Opción 3: Distancia Euclidiana

Polarization identity

$$\|x\| = \sqrt{\langle x, x \rangle}$$

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2 + 2\,\mathrm{Re}\langle x, y \rangle$$

Si los vectores están normalizados

$$\|x - y\|^2 = \|x\|^2 + \|y\|^2 - 2(x \cdot y)$$

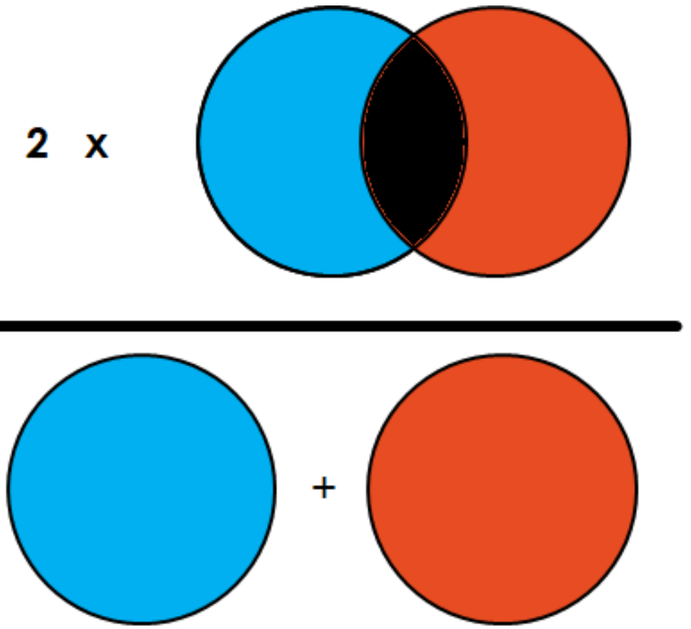
$$\|x - y\|^2 = 2 - 2(x \cdot y)$$

$$\|x - y\|^2 = 2(1 - \cos(x, y))$$

Relación monótona

Sørensen–Dice *coefficient*

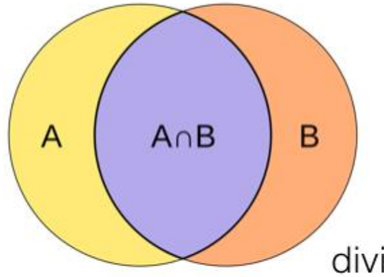
$$\text{DICE} = \frac{2|A \cap B|}{|A| + |B|} =$$

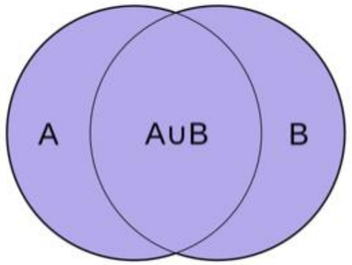


Distancia *Jaccard*

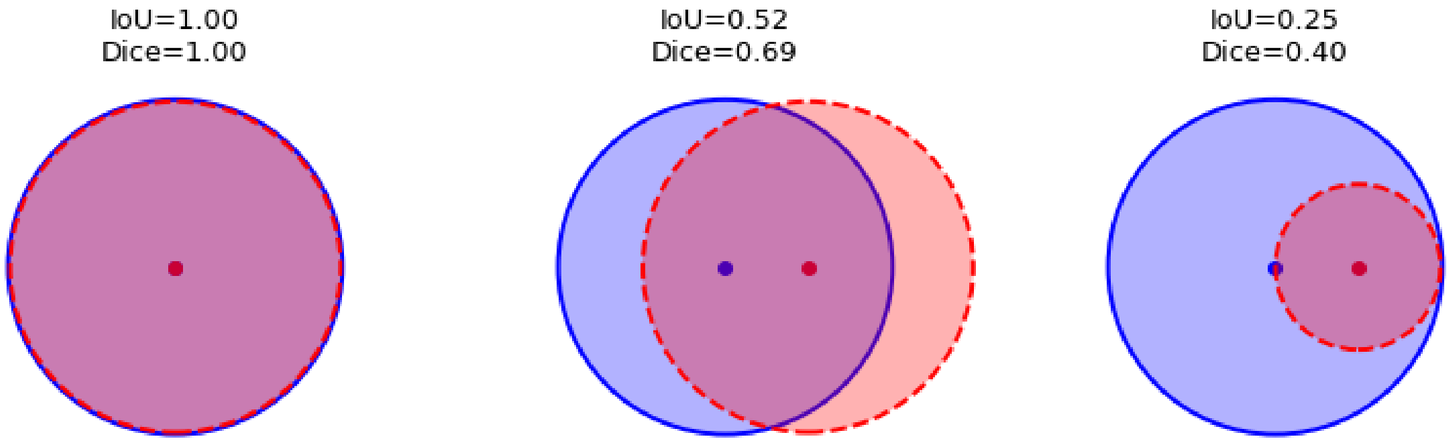
$$D(x, y) = 1 - \frac{|x \cap y|}{|y \cup x|}$$

$$J(A, B) = \frac{\text{The intersect of A \& B}}{\text{The union of A \& B}} \quad \text{division}$$

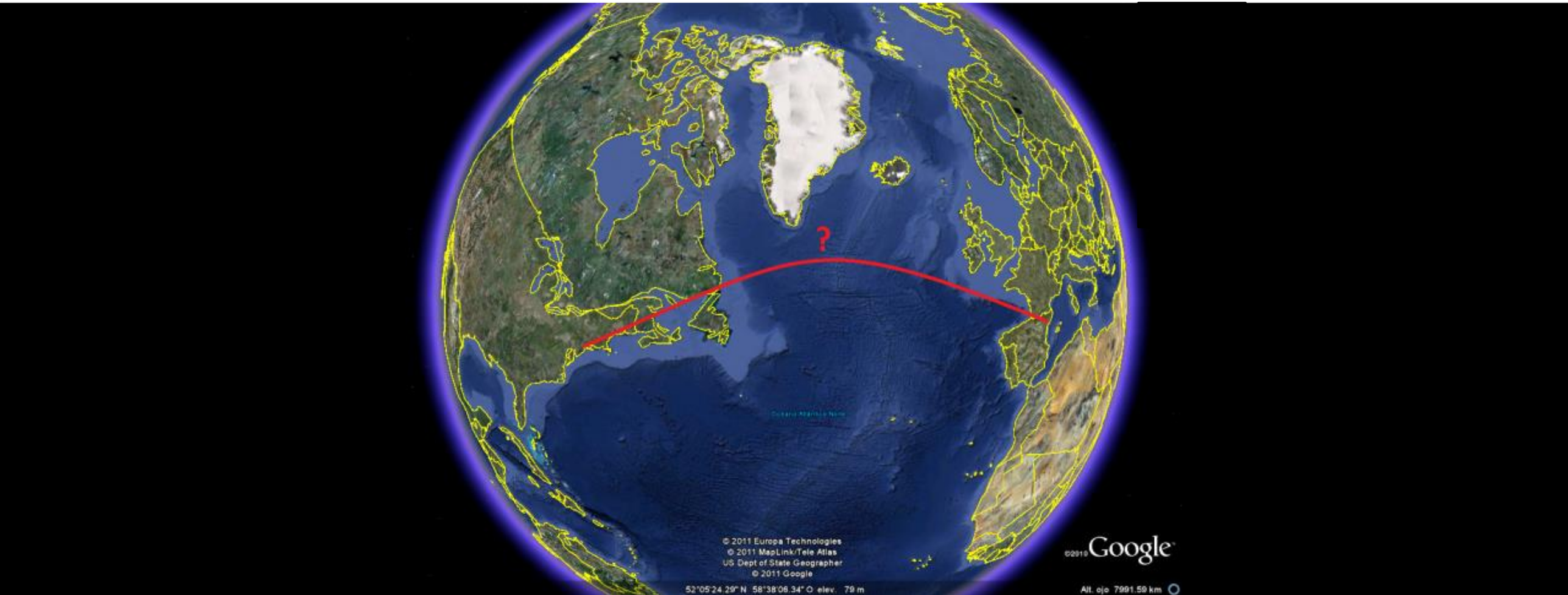




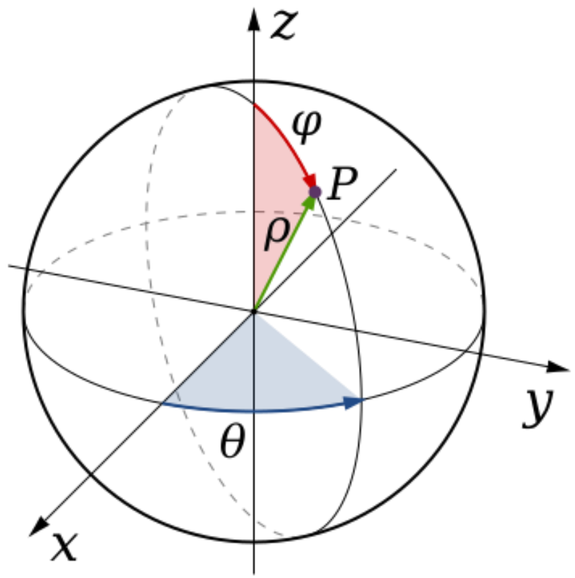
Distancia *Jaccard*



Distancia *Haversine*



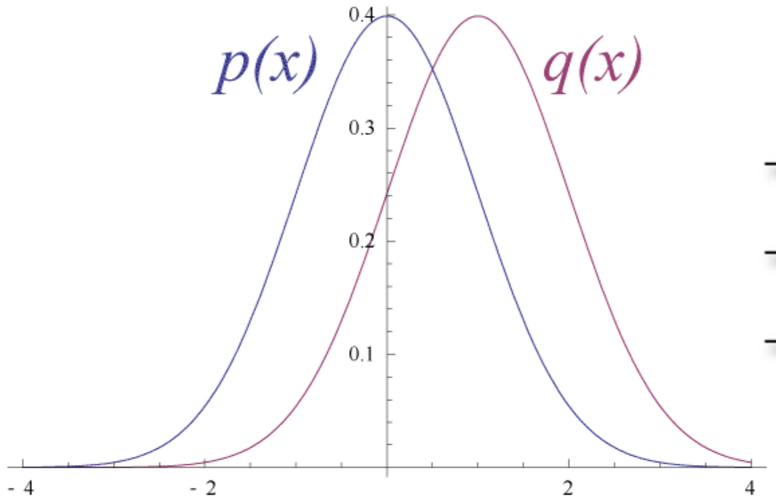
Distancia *Haversine*



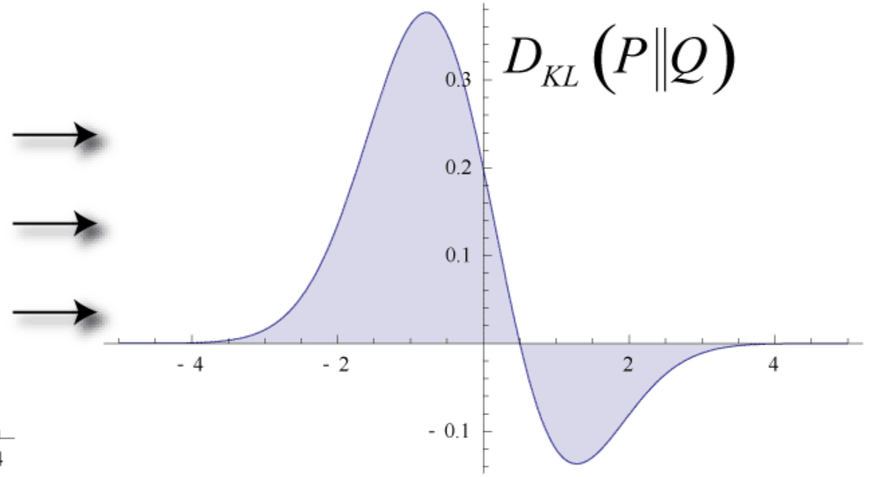
$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

Kullback–Leibler *Divergence*

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$



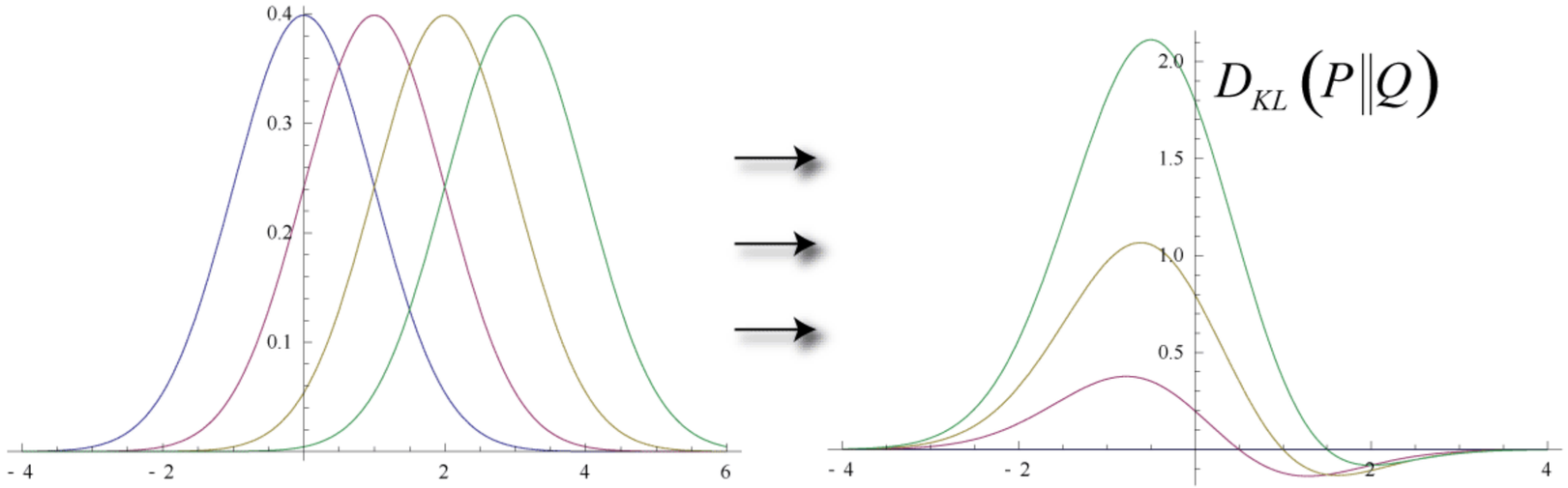
Original Gaussian PDF's



KL Area to be Integrated

Kullback–Leibler *Divergence*

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$



Original Gaussian PDF's

KL Area to be Integrated

Wasserstein *distance*

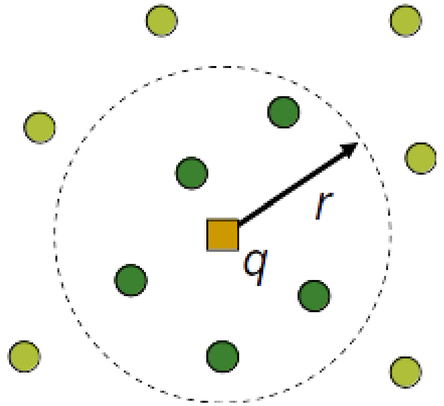
Kantorovich–Rubinstein metric

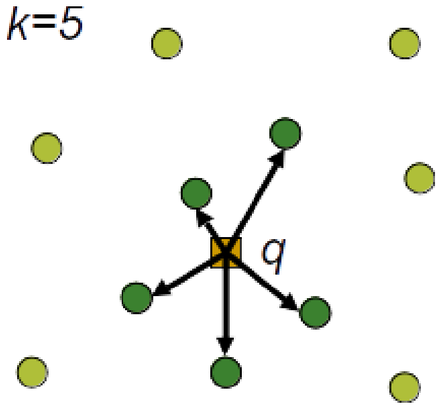
$$W_p(\mu,\nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \mathbf{E}_{(x,y)\sim\gamma} d(x,y)^p \right)^{1/p}$$

Para distribuciones unidimensionales:

$$W_p(U,Y) = \left(\int_0^1 \left| F_Y^{-1}(\omega) - F_U^{-1}(\omega) \right|^p d\omega \right)^{1/p}$$

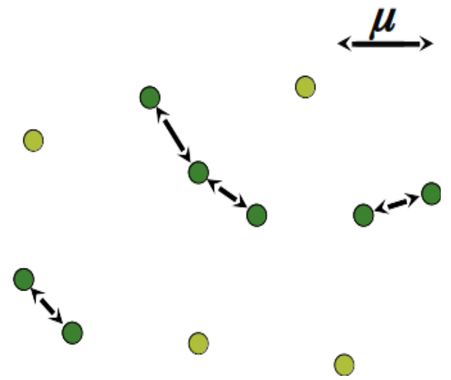
Consultas *de similaridad*



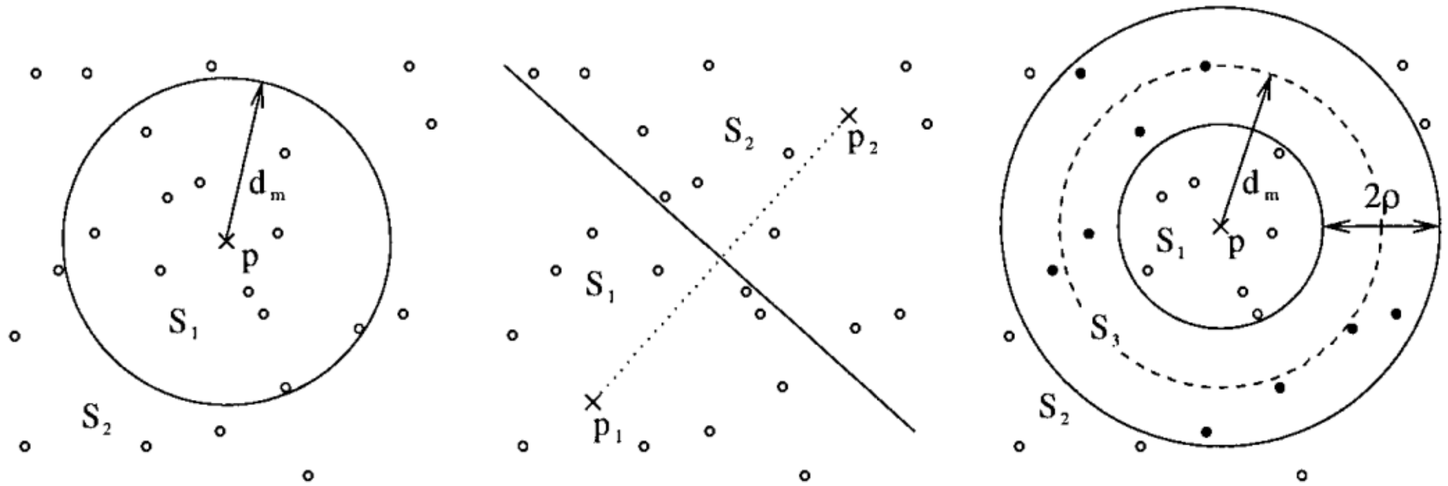


Spatial *join query*

$$J(X, Y, \mu) = \{(x, y) \in X \times Y : d(x, y) \leq \mu\}$$



Particion*amiento*

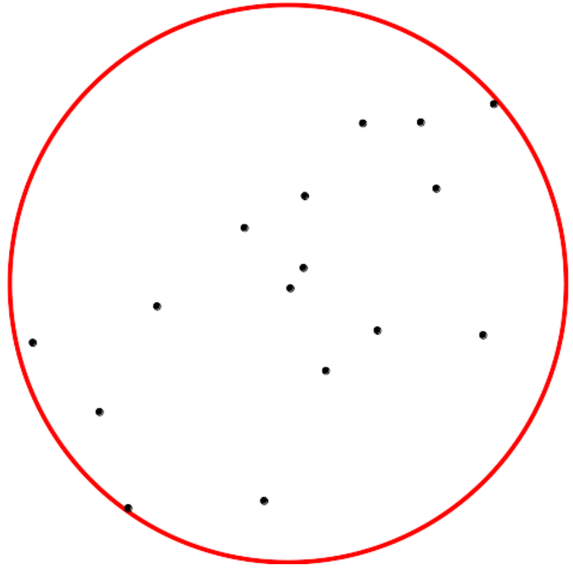


Ball

Generalized Hyperplane

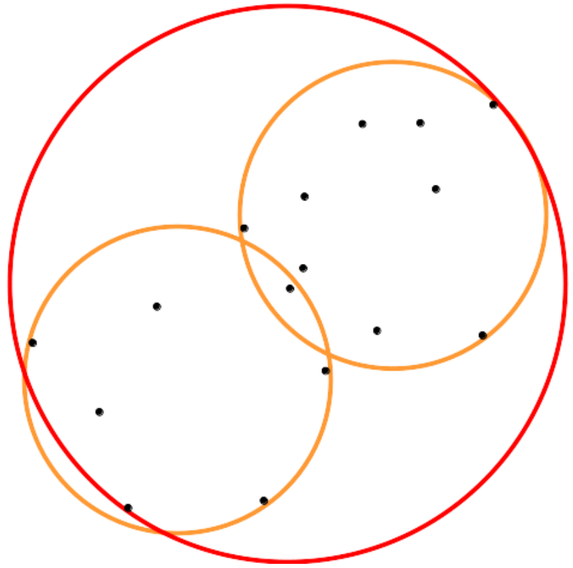
Excluded Middle

Particion*amiento*

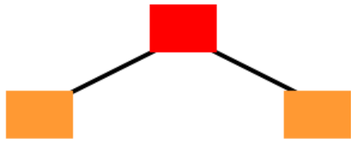


Nivel 1

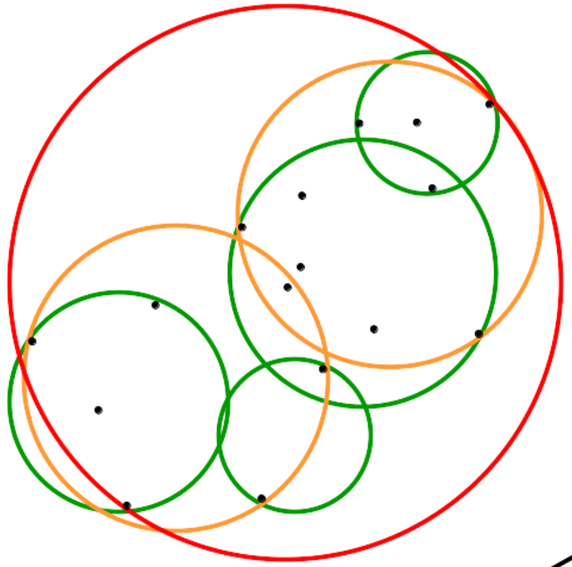
Particion*amiento*



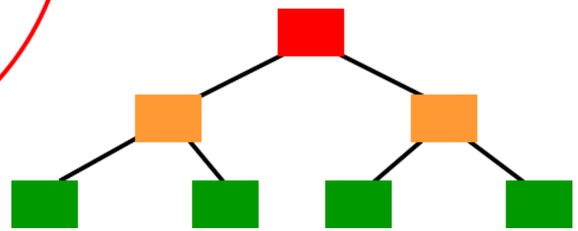
Nivel 2



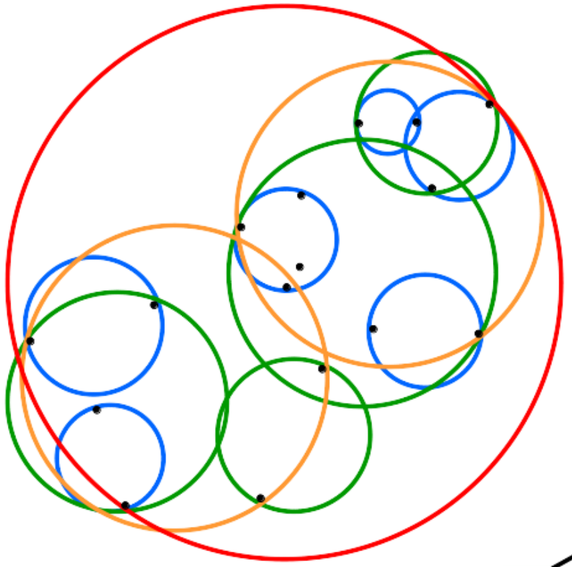
Particion*amiento*



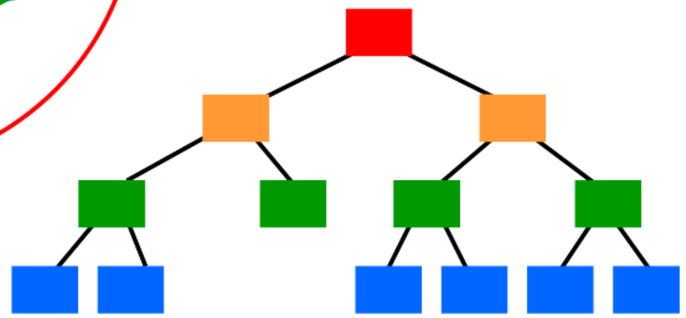
Nivel 3



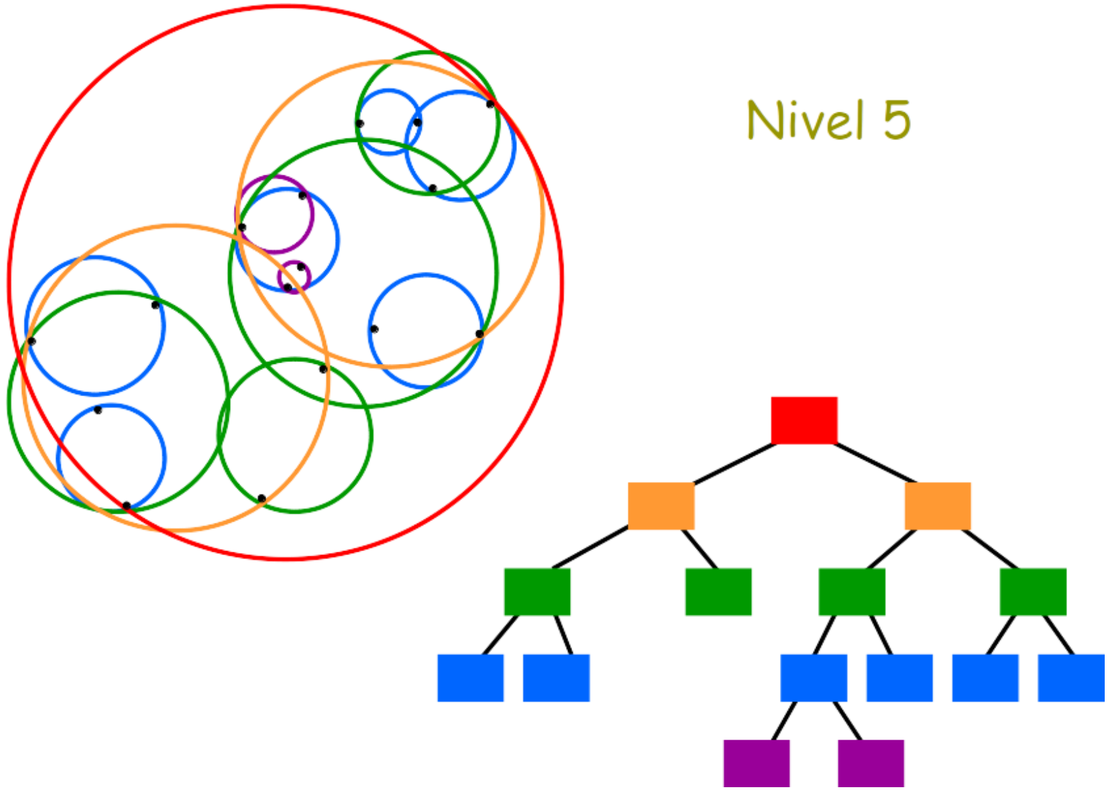
Particion*amiento*



Nivel 4



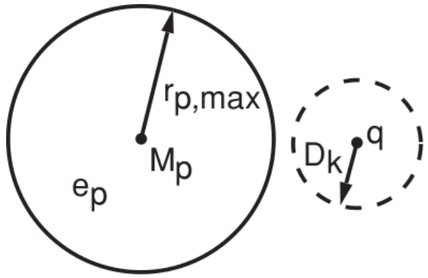
Particion*amiento*



2. Depth-First K-Nearest Neighbor

Depth-First *K-Nearest Neighbor*

Reglas 1 Nodo interno



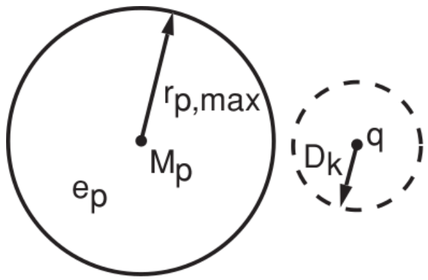
Círculo de Búsqueda

Omitir si: $D_k + r_{p,max} < d(q, M_p)$

Descartar **nodos internos**

Depth-First *K-Nearest Neighbor*

Reglas 1 Nodo interno

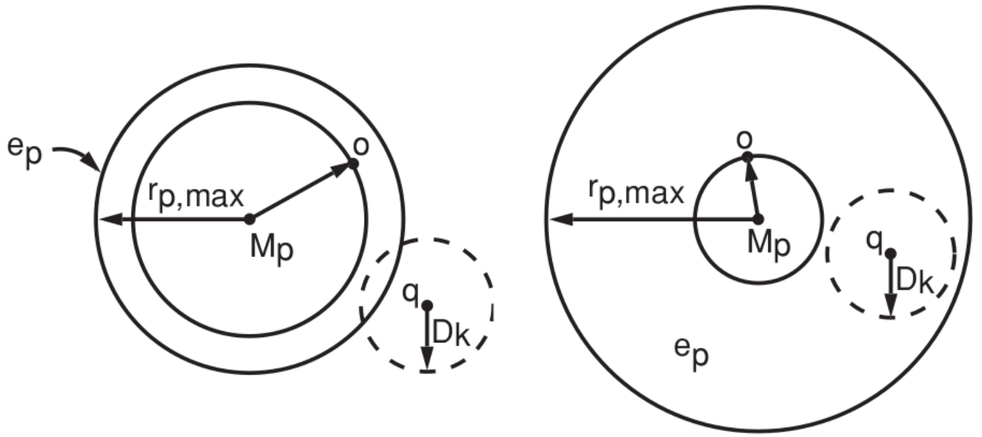


Círculo de Búsqueda

Omitir si: $D_k + r_{p,max} < d(q, M_p)$

Descartar **nodos internos**

Reglas 2 Nodo hoja



Omitir si: $D_k + d(o, M_p) < d(q, M_p)$

Descartar **puntos**

UTEC

UNIVERSIDAD DE INGENIERÍA
Y TECNOLOGÍA