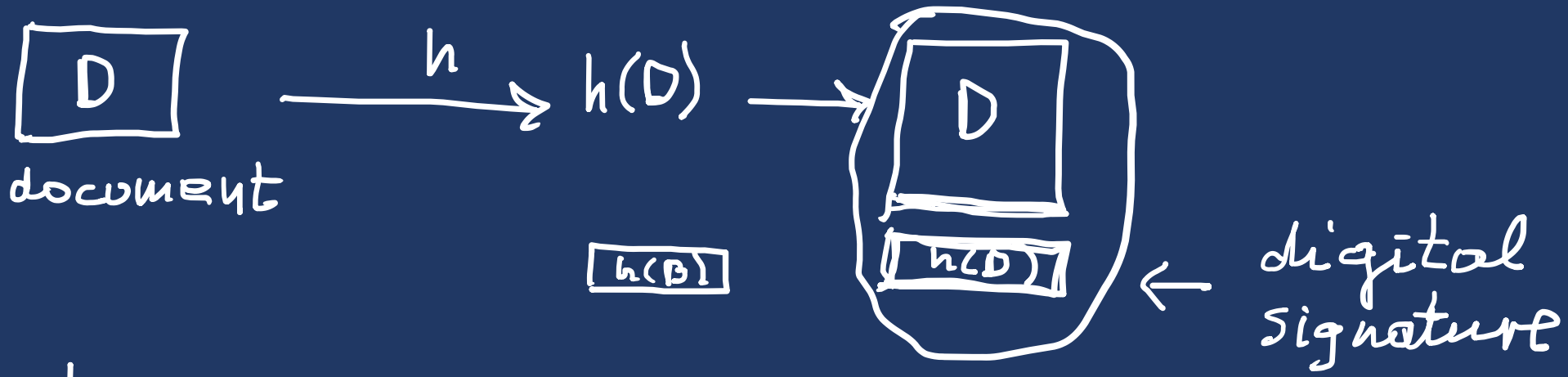
Locality Sensitive Hash Functions Families

Standard hash functions, e.g. $h \sim$ MD5

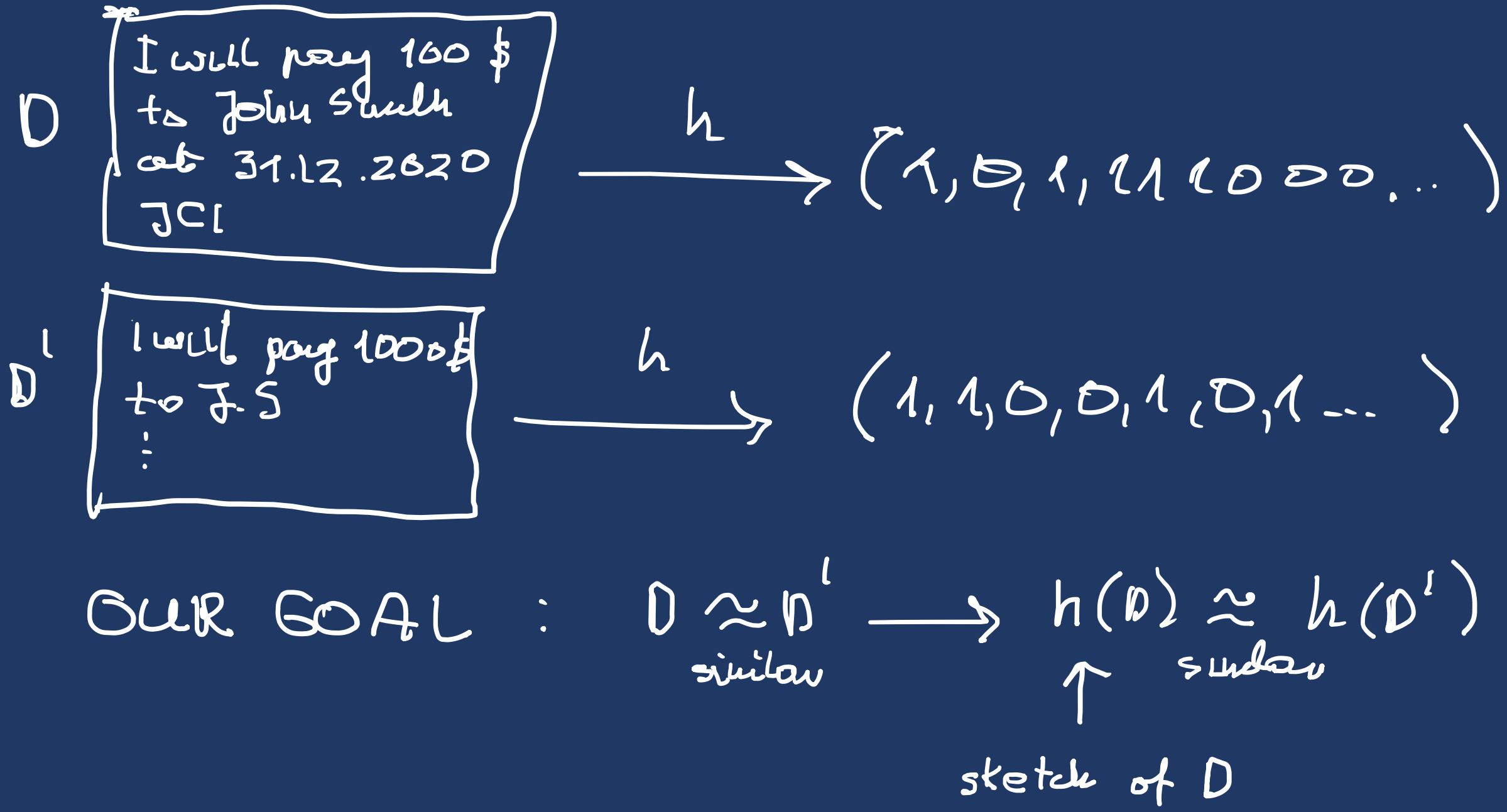


D' ← small modif. of D

$rng(h) \subseteq \{0,1\}^{256}$

$h(D)$ $h(D')$ ← they are very different

$\sum_{i=1}^{256} [\![h(D)_i \neq h_i(D')]\!] \simeq \frac{256}{2}$


D
I will pay 100 $
to John Smith
at 31.12.2020
JCI
h
(1, 0, 1, 1, 1, 0, 0, 0, ...)
D'
I will pay 1000$
to J.S
⋮
h
(1, 1, 0, 0, 1, 0, 1, ...)
OUR GOAL : D ≈ D'
similar
h(D) ≈ h(D')
similar
sketch of D

Metric space.

(X, d) ← metric space if

1) $d : X \times X \longrightarrow [0, \infty)$ $\quad (d(x,y) \geq 0)$
2) $d(x,y) = 0 \iff x = y$
3) $d(x,y) = d(y,x)$
4) $d(x,z) \leq d(x,y) + d(y,z)$

triangle inequality

EX 1 $X = \mathbb{R}^n$, $p \geq 1$.

$$\|\bar{x}\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}$$

Thm For each $p \geq 1$, d_p is a metric

$$d_p(\bar{x}, \bar{y}) = \|x - y\|_p$$

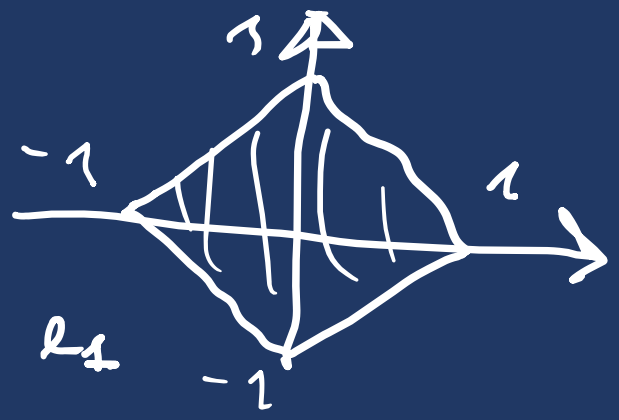
$p = 2$: $d_2(\bar{x}, \bar{y}) = \sqrt[2]{(x_1 - y_1)^2 + \dots + (x_n - y_n)^2}$ ← euclidean distance

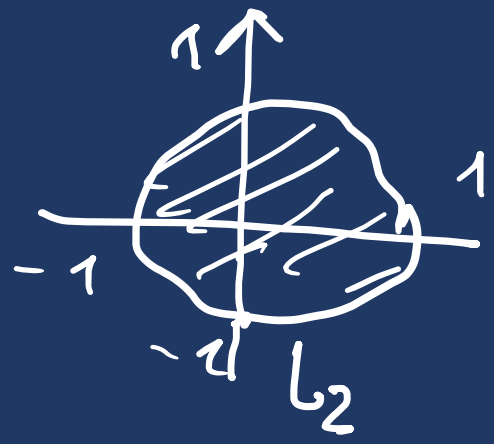
$p = 1$: $d_1(\bar{x}, \bar{y}) = \sum_{i=1}^{n} |x_i - y_i|$

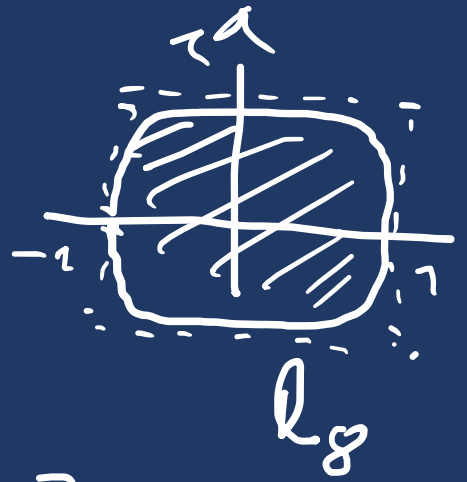
$$\lim_{p \to \infty} \|\bar{x}\|_p = \max \{ |x_i| : i = 1 \dots n \}$$

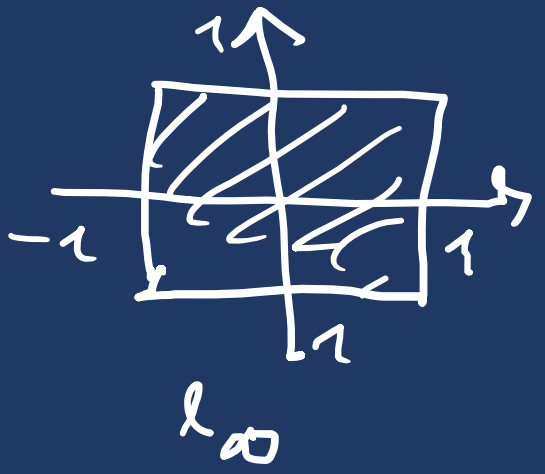
$$\overline{B}(x,r) = \{y \in X : d(x,y) \leq r\}$$

$$\overline{B}_p(0,1) = \{\bar{y} \in \mathbb{R}^n : \|y\|_p \leq 1\}$$









$(n=2)$ $\overline{B}_1(0,1) = \{(y_1,y_2) : |y_1| + |y_2| \leq 1\}$

EX 2 $X = \{0,1\}^n \; ; \; \bar{x}, \bar{y} \in X$

$$d_H(\bar{x},\bar{y}) = \sum_{i=1}^{n} |\bar{x}_i - \bar{y}_i|$$

(ℓ_1 - distance on $\{0,1\}^n$)

Hamming distance

$$= \sum_{i=1}^{n} [\![x_i \neq y_i]\!] = \sum_i \|i \in A_x \triangle A_y\| = |A_x \triangle A_y|$$

$$[\![\varphi]\!] = \begin{cases} 1 : \varphi \\ 0 : \neg\varphi \end{cases}$$

$$x_i \neq y_i \equiv (x_i = 1 \wedge y_i = 0) \vee (x_i = 1 \wedge y_i = 0)$$
$$\equiv (i \in A_x \setminus A_y) \vee (i \in A_y \setminus A_x) \equiv$$

$$\{0,1\}^n \ni \bar{x} \longrightarrow A_{\bar{x}} = \{i \in \{1..n\} : \bar{x}_i = 1\} \subseteq \{1,\dots,n\}$$

$\hookrightarrow i \in A_x \triangle A_y$

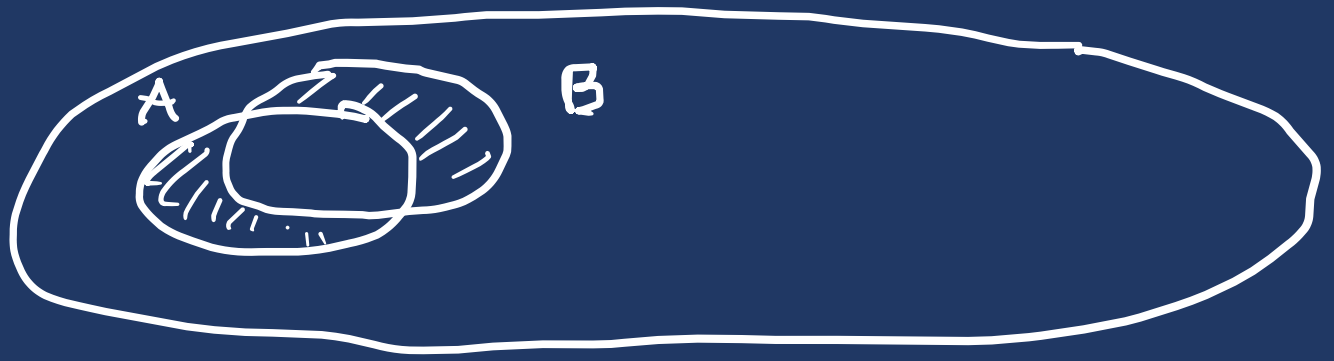
- $(\mathcal{P}(\{1,\dots,u\}), d)$ ← metric space

$$d(A,B) = |A \triangle B|$$

- Fix any set X.

$$P_{fin}(X) = \{D \subseteq X : |D| < \infty\}$$

$$d(A,B) = |A \triangle B|$$



$X = \text{ASCII}^*$

$A \underset{fin}{\subseteq} X$: documents

Main problem : our metrics are unbounded.

We want use such metrics d that

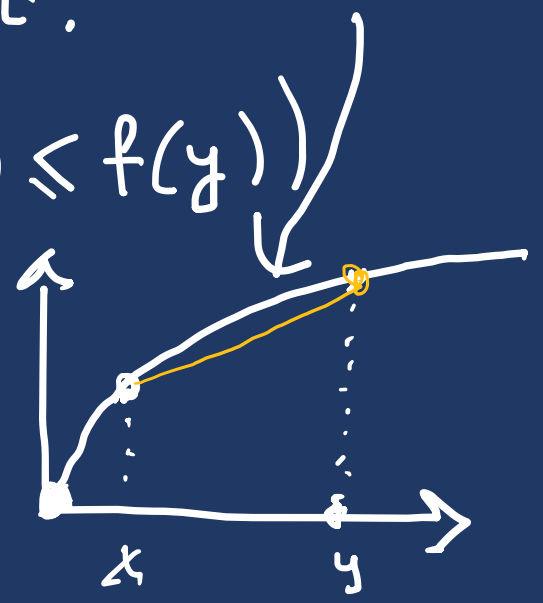
$$0 \leq d \leq 1.$$

standard way.

above line

THM. Suppose $f: [0,\infty) \to [0,\infty)$ s.t.

1) $f(0) = 0$
2) f is monotonic ($0 \leq x < y \Rightarrow f(x) \leq f(y)$)
3) f is <u>concave</u>



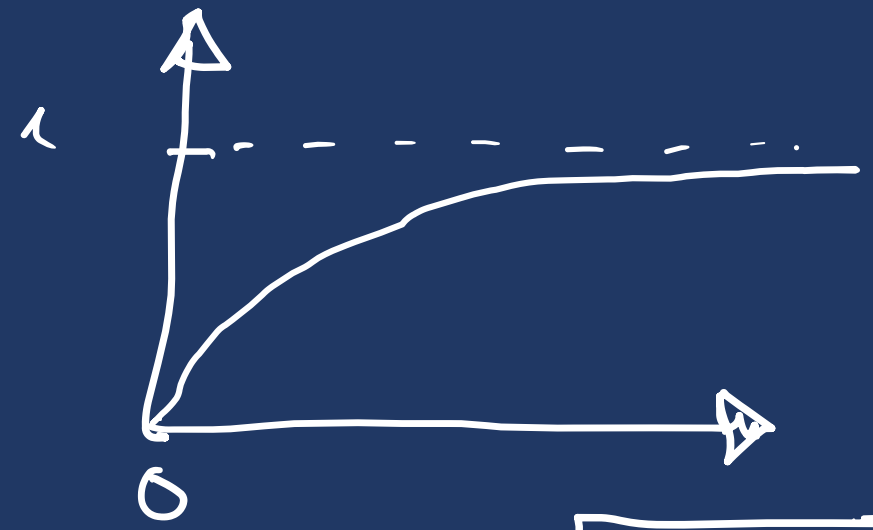
Let (X, d) be a metric space. Then

$$\rho(x,y) = f(d(x,y))$$

is a metric on X.

E1. $f(x) = \min\{x, 1\}$

E2. $f(x) = \dfrac{x}{1+x}$



$f'(x) = \dfrac{1}{(1+x)^2} > 0$

$f''(x) = \dfrac{-2}{(1+x)^3} < 0$

$\Downarrow$

f is concave

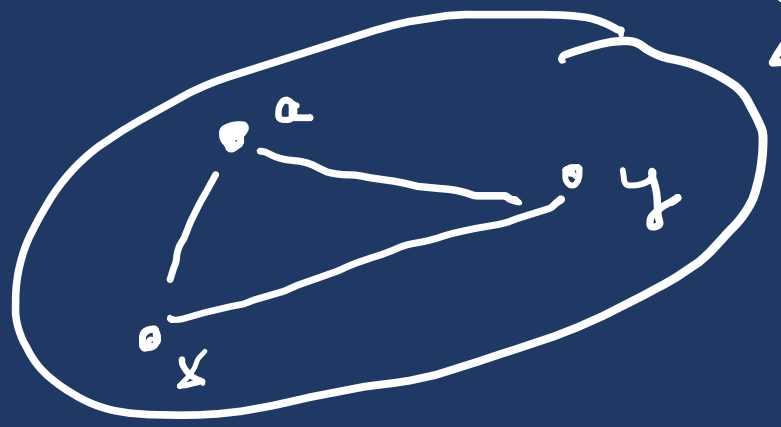
$$d(\bar{x},\bar{y}) = \frac{\sqrt{\sum (x_i - y_i)^2}}{1 + \sqrt{\sum (x_i - y_i)^2}}$$

$\mathbb{R}^n$

THM (Steinhaus) Let (X, d) be metric space, let $a \in X$. Let

$$\rho(x,y) = \frac{2 \cdot d(x,y)}{d(x,a) + d(y,a) + d(x,y)}$$

Then (X, ρ) is a metric space.



X

$$d(x,a) + d(a,y) + d(x,y) \geq d(x,y) + d(x,y)$$

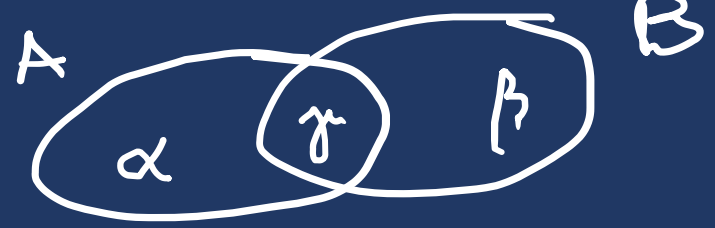
$$\rho(x,y) \leq \frac{2 \cdot d(x,y)}{2 \cdot d(x,y)} = 1$$

EX. Suppose $x \neq a$. Calculate $\rho(x,a)$!!!

Application : $d(A,B) = |A \triangle B|$, $a = \phi$

$$\rho(A,B) = \frac{2 \cdot |A \triangle B|}{|A \triangle \phi| + |B \triangle \phi| + |A \triangle B|} =$$

$$= \frac{2 \cdot |A \triangle B|}{|A| + |B| + |A \triangle B|} = \frac{2(\alpha + \beta)}{(\alpha+\gamma)+(\beta+\gamma)+(\alpha+\beta)}$$



$$= \frac{2 \cdot (\alpha + \beta)}{2(\alpha+\beta+\gamma)} = \frac{|A \triangle B|}{|A \cup B|}$$

$$d_J(A,B) = \frac{|A \triangle B|}{|A \cup B|}$$

Jaccard distance

- $d_J(A,B) \leq 1$
- $(d_J(A,B) = 1) \equiv (A \cap B = \phi)$

SIMILARITY: $S : X \times X \longrightarrow [0,1]$

- $s(x,x) = 1$
- $s(x,y) = s(y,x)$

PERFECT SIMILARITY: $s(x,y) = 1 - d(x,y)$
where d is a metric s.t $0 \leq d \leq 1$.

$$1 - d_J(A,B) = 1 - \frac{|A \triangle B|}{|A \cup B|} =$$

$$= \frac{|A \cup B| - |A \triangle B|}{|A \cup B|} = \frac{(\alpha + \beta + \gamma) - (\alpha + \beta)}{|A \cup B|}$$

$$= \frac{\gamma}{|A \cup B|} = \frac{|A \cap B|}{|A \cup B|}$$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Jaccard similarity

Estimation of $\mathcal{J}(A,B)$.

$\Omega = \{1, 2, \ldots, n\}$ $\qquad A, B \subseteq \Omega$.

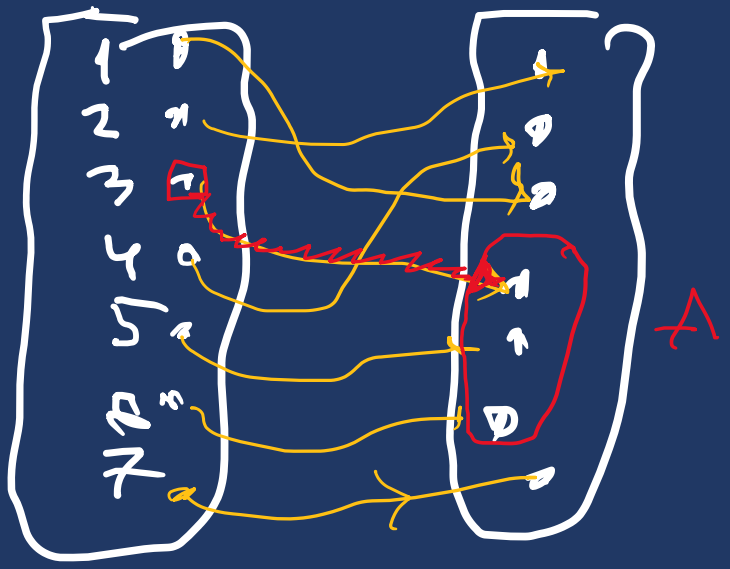
S_n = all permutations of Ω ; $|S_n| = n!$,

$\sigma \in S_n$, $A \subseteq \Omega$; $A \neq \phi$

$h_A(\sigma) = \min\{k : \sigma_k \in A\}$

$h_A(\sigma) \in \{1, 2, \ldots, n\}$

$\sigma : \{1 \ldots n\} \xrightarrow[\text{onto}]{1-1} \{1 \ldots n\}$



Look at S_n (all perm) as a prob. space:

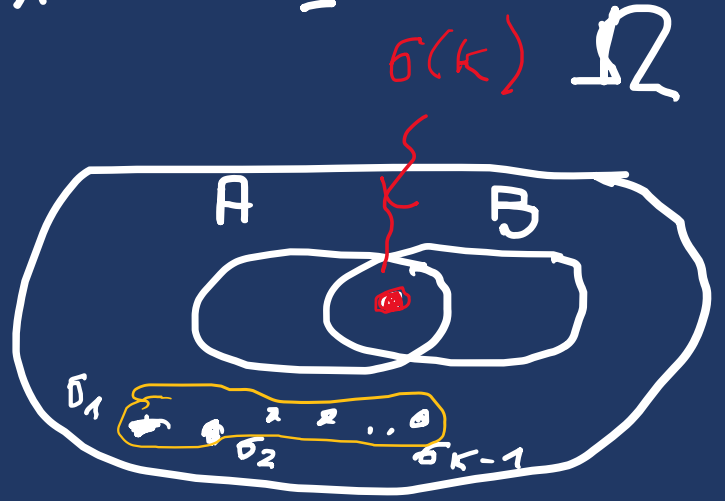
$$P(\{\sigma\}) = \frac{1}{n!}.$$

$$\Pr_\sigma\left(h_A(\sigma) = h_B(\sigma)\right) = \sum_{k=1}^{n} \Pr_\sigma\left[h_A(\sigma) = h_B(\sigma) \mid h_A(\sigma) = k\right] \cdot \Pr\left[h_A(\sigma) = k\right]$$

$$= \sum_{k=1}^{n} \Pr_\sigma\left[h_B(\sigma) = k \mid h_A(\sigma) = k\right] \cdot \Pr\left[h_A(\sigma) = k\right]$$

$$= \sum_{k=1}^{n} \frac{|A \cap B|}{|A \cup B|} \cdot \Pr\left[h_A(\sigma) = k\right]$$

$$= \frac{|A \cap B|}{|A \cup B|}$$



$\{\sigma(1), \ldots, \sigma(k-1)\} \cap (A \cup B) = \emptyset$

!!!

$$\Pr_{\sigma}\left[h_A(\sigma) = h_B(\sigma)\right] = \gamma(A,B)$$