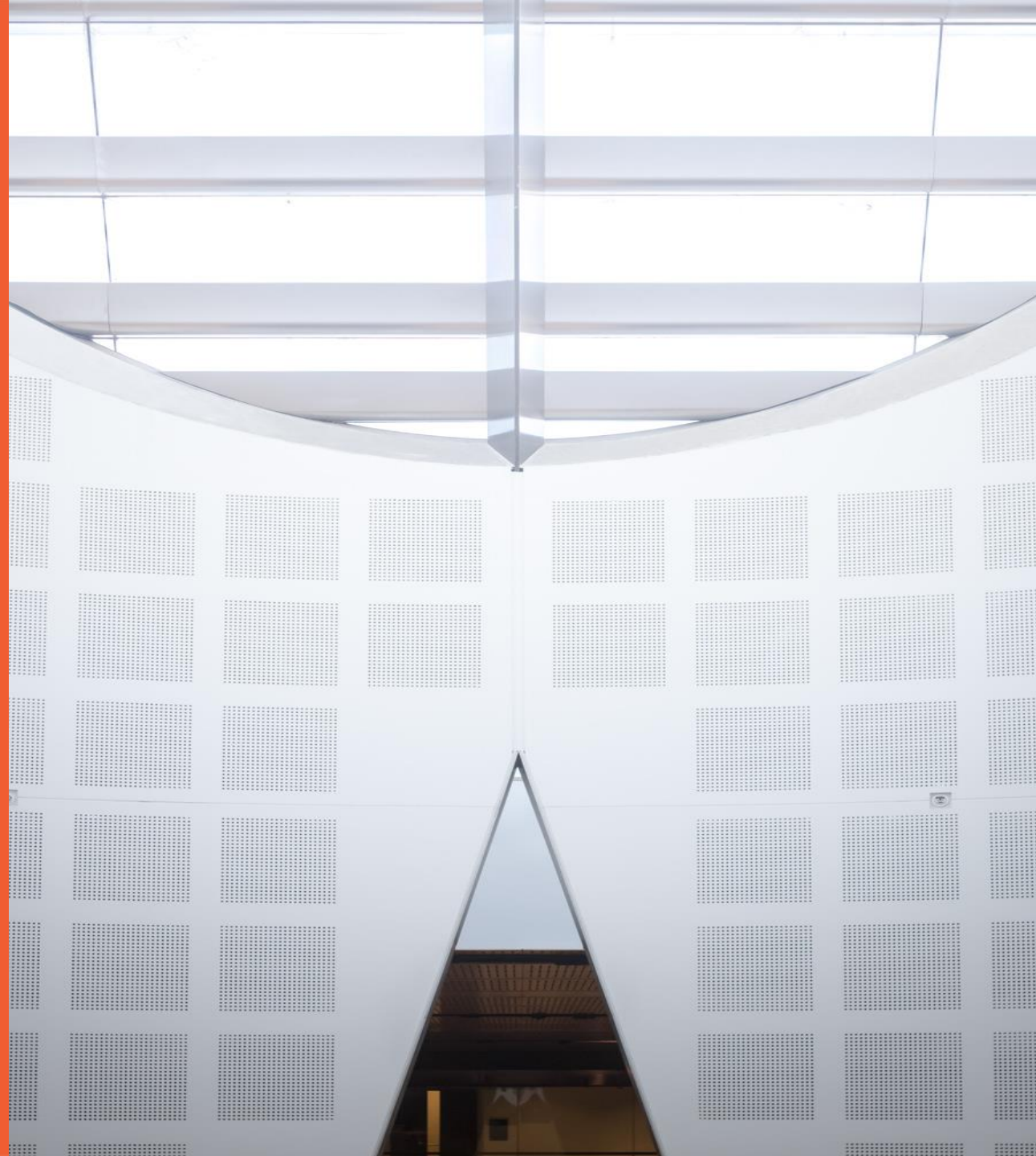
COMPx270: Randomised and Advanced Algorithms

Lecture 7: Nearest Neighbours and dimensionality reduction

Clément Canonne

School of Computer Science

THE UNIVERSITY OF SYDNEY

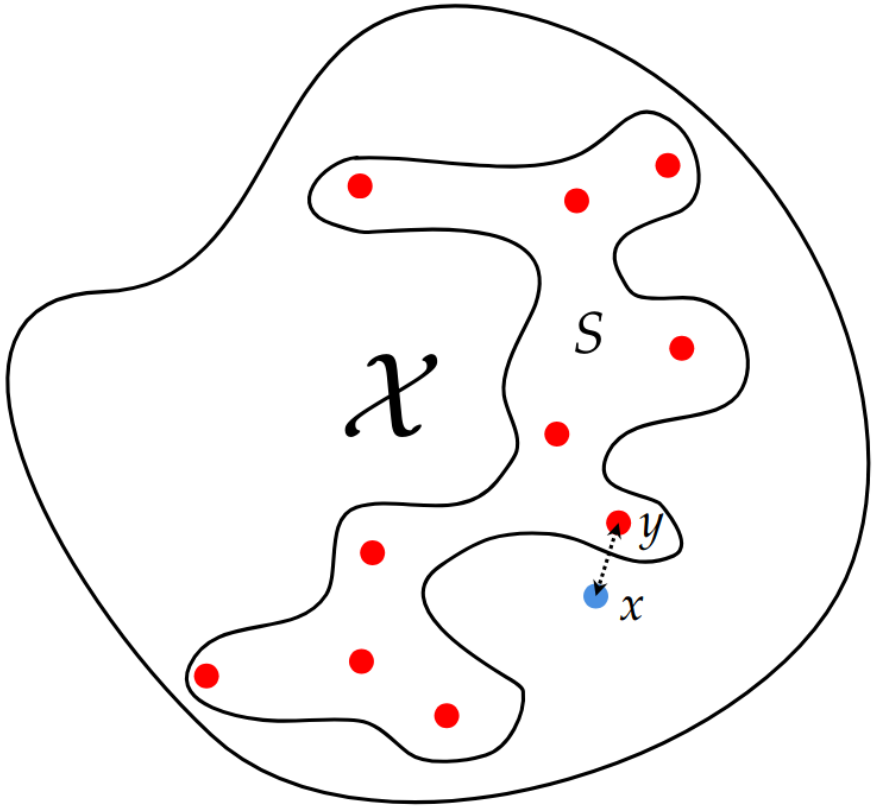


A question 

You have n pictures, each 4096x4096 pixels, of venomous spiders. Someone finds a spider in their kitchen and sends you a photo, asking which type of spider it is and if it is venomous, **because they just have been bitten.**

How long will it take you?

Nearest Neighbour Search

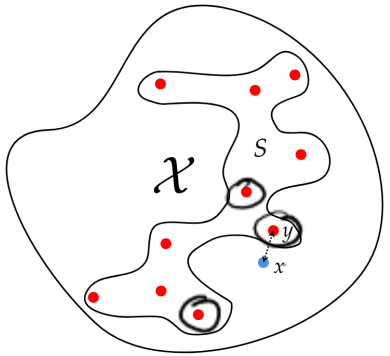


Nearest Neighbour Search

Lists? Voronoi? K-d trees? Hash tables?

Bad news...

Approximate Nearest Neighbour Search



QUERY(x): given an element $x \in \mathcal{X}$, return an element $y \in S$ sort-of-minimising $\text{dist}(x, y)$, that is, $\text{dist}(x, y) \leq C \cdot \min_{y' \in S} \text{dist}(x, y')$.

Dimensionality Reduction: the JL Lemma (Euclidean space)

JL Lemma and ANN

Beyond JL Lemma: Hashing!

Locality-Sensitive Hashing

Definition 36.1. Let $0 \leq q < p \leq 1, r > 0, C > 1$, and $(\mathcal{X}, \text{dist})$ be a metric space. Then a family of functions $\mathcal{H}$ from $\mathcal{X}$ to $\mathcal{Y}$ is a (r, C, p, q)-*Locality Sensitive Hash family* (LSH) if, for every $x, x' \in \mathcal{X}$,

- If $\text{dist}(x, x') \leq r$, then $\Pr_{h \sim \mathcal{H}}[\, h(x) = h(x') \,] \geq p$;
- If $\text{dist}(x, x') \geq Cr$, then $\Pr_{h \sim \mathcal{H}}[\, h(x) = h(x') \,] \leq q$;

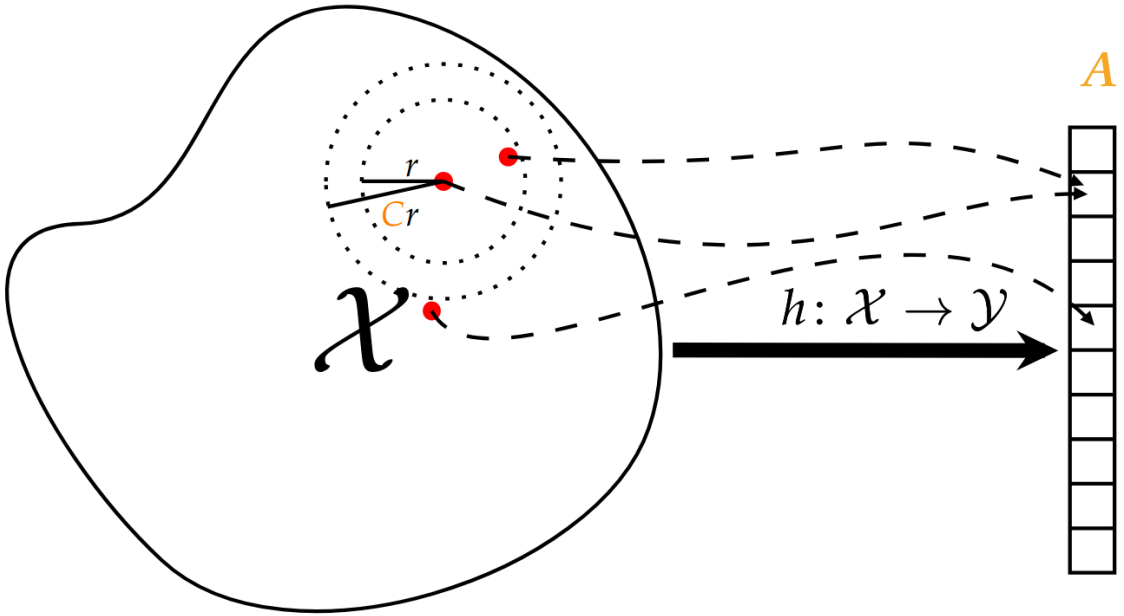
and we say $\rho := \dfrac{\log(1/p)}{\log(1/q)} > 1$ is the *sensitivity parameter* of $\mathcal{H}$.

Locality-Sensitive Hashing

Definition 36.1. Let $0 \leq q < p \leq 1$, $r > 0$, $C > 1$, and $(\mathcal{X}, \text{dist})$ be a metric space. Then a family of functions $\mathcal{H}$ from $\mathcal{X}$ to $\mathcal{Y}$ is a (r, C, p, q)-*Locality Sensitive Hash family* (LSH) if, for every $x, x' \in \mathcal{X}$,

- If $\text{dist}(x, x') \leq r$, then $\Pr_{h \sim \mathcal{H}}[\, h(x) = h(x') \,] \geq p$;
- If $\text{dist}(x, x') \geq Cr$, then $\Pr_{h \sim \mathcal{H}}[\, h(x) = h(x') \,] \leq q$;

and we say $\rho := \frac{\log(1/p)}{\log(1/q)} > 1$ is the *sensitivity parameter* of $\mathcal{H}$.



Locality-Sensitive Hashing: "Baby version"

$\textsc{Query}_r(x)$: given an element $x \in \mathcal{X}$, return an element $y \in S$, or $\perp$, such that:

- If there exists $y^* \in S$ such that $\text{dist}(x, y^*) \leq r$, then, with probability at least $9/10$, $\textsc{Query}_r(x)$ returns an element $y \in S$ such that $\text{dist}(x, y^*) \leq C \cdot r$;
- If $\text{dist}(x, y) > C \cdot r$ for *every* $y \in S$, then, with probability 1, $\textsc{Query}_r(x)$ returns $\perp$.
- Otherwise, any output in $S \cup \{\perp\}$ is allowed.

Locality-Sensitive Hashing: "Baby version" (1/4)

$\textsc{Query}_r(x)$: given an element $x \in \mathcal{X}$, return an element $y \in S$, or $\perp$, such that:

- If there exists $y^* \in S$ such that $\text{dist}(x, y^*) \leq r$, then, with probability at least $9/10$, $\textsc{Query}_r(x)$ returns an element $y \in S$ such that $\text{dist}(x, y^*) \leq C \cdot r$;
- If $\text{dist}(x, y) > C \cdot r$ for *every* $y \in S$, then, with probability 1, $\textsc{Query}_r(x)$ returns $\perp$.
- Otherwise, any output in $S \cup \{\perp\}$ is allowed.

Locality-Sensitive Hashing: "Baby version" (2/4)

$\textsc{Query}_r(x)$: given an element $x \in \mathcal{X}$, return an element $y \in S$, or $\perp$, such that:

- If there exists $y^* \in S$ such that $\text{dist}(x, y^*) \leq r$, then, with probability at least $9/10$, $\textsc{Query}_r(x)$ returns an element $y \in S$ such that $\text{dist}(x, y^*) \leq C \cdot r$;
- If $\text{dist}(x, y) > C \cdot r$ for *every* $y \in S$, then, with probability 1, $\textsc{Query}_r(x)$ returns $\perp$.
- Otherwise, any output in $S \cup \{\perp\}$ is allowed.

Locality-Sensitive Hashing: "Baby version" (3/4)

$\textsc{Query}_r(x)$: given an element $x \in \mathcal{X}$, return an element $y \in S$, or $\perp$, such that:

- If there exists $y^* \in S$ such that $\text{dist}(x, y^*) \leq r$, then, with probability at least $9/10$, $\textsc{Query}_r(x)$ returns an element $y \in S$ such that $\text{dist}(x, y^*) \leq C \cdot r$;
- If $\text{dist}(x, y) > C \cdot r$ for *every* $y \in S$, then, with probability 1, $\textsc{Query}_r(x)$ returns $\perp$.
- Otherwise, any output in $S \cup \{\perp\}$ is allowed.

Locality-Sensitive Hashing: "Baby version" (4/4)

$\textsc{Query}_r(x)$: given an element $x \in \mathcal{X}$, return an element $y \in S$, or $\perp$, such that:

- If there exists $y^* \in S$ such that $\text{dist}(x, y^*) \leq r$, then, with probability at least $9/10$, $\textsc{Query}_r(x)$ returns an element $y \in S$ such that $\text{dist}(x, y^*) \leq C \cdot r$;
- If $\text{dist}(x, y) > C \cdot r$ for *every* $y \in S$, then, with probability 1, $\textsc{Query}_r(x)$ returns $\perp$.
- Otherwise, any output in $S \cup \{\perp\}$ is allowed.

Locality-Sensitive Hashing: "Baby version" (👶)

Locality-Sensitive Hashing: "They grow up so fast" (👱)

Locality-Sensitive Hashing: But... do they exist?

Locality-Sensitive Hashing: But... do they exist?