Approximate Nearest Neighbor:

Given metric \((X, d)\) and a subset \(S \subseteq X\),

\[ n := |S| \]

Want to preprocess \(S\) to answer approximate nearest neighbor queries:

\[ \text{NN}(q, r, c) \]

\( q \in X, r, c \in \mathbb{R}_{\geq 0} \)
If $\exists x \in S$ sit.

\[ d(q, x) \leq r, \text{ then report some } yes \text{ sit.} \]

\[ d(q, y) \leq cr. \]

If no $x \in S$ sit.

\[ d(q, x) \leq cr, \text{ report failure.} \]

o.w. report $x \in S$ sit.

\[ d(q, x) = cr \text{ or } fail. \]

Either is fine.
If $X \in \mathbb{R}^d$, can answer $\text{NNI}(q, r, (1+\epsilon))$ for any constant $\epsilon > 0$ in $O(\log n)$ time using $O(n \log n)$ space BBD-tree.
Locality Sensitive Hashing (for higher dim. metrics)

A probability distribution $H$ over hash functions is a hash family.
Given parameter $c > 1$, probabilities $p_1 > p_2 + \epsilon$ a distance $r \geq 0$, hash family $H$ is $(r, c, p_1, p_2)$-locality sensitive (LSH) if $\forall x, x' \in X$ $\forall y, y' \in S$. 
If $d(x, q) \leq r$, then
\[ Pr \left[ h(x) = h(q) \right] \geq p_1. \]

If $d(x, q) > cr$, then
\[ Pr \left[ h(y) = h(q) \right] \leq p_2. \]

Want $p_1 \gg p_2$. 

---

**Important**: The above text seems to be discussing a probabilistic scenario, possibly related to hashing or error correction in coding theory. The conditions and probabilities are used to infer a desirable inequality between two probability thresholds, $p_1$ and $p_2$.
Hamming Distance

\( x : m \text{-dim. bit vectors} \)
\( d(x, y) : \# \text{ positions where their bits disagree} \)

\( H : \text{Each } h \text{ is assigned a single coordinate index, } h(x) : \text{ bit at } h's \text{ index} \)
Choose \( h \) uniformly at random.

\[
Pr \left[ h(x) = h(y) \right] = 1 - \frac{d(x, y)}{m}
\]

is \((r, cr, p_1, p_2)\)-LSH

\[
\text{For } p_1 = 1 - \frac{r}{m}
\]
\[
p_2 = 1 - cr \frac{r}{m}
\]
\textbf{Jaccard Distance}

\( U : \) some universe of elements

\( X \subseteq U \)

\textbf{Jaccard similarity coefficient} of \( S_1, S_2 \subseteq U \)

\( J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \)
Jacard distance
\[ d(S_1, S_2) = 1 - J(S_1, S_2) \]
is a metric.

Let each \( h_i \) be based on a different permutation of \( V \).
\( h_i(S) \) is the earliest element of \( S \) according to \( h_i \)'s permutation.
\[
\begin{align*}
p_r \left[ h(s_1) = h(s_2) \right] &= \vartheta(s_1, s_2) \\
&= 1 - d(s_1, s_2) \\
p_r &= 1 - r \\
p_r &= 1 - cr
\end{align*}
\]
Angular Distance

\[ \mathbf{x}: \text{vectors in } \mathbb{R}^m \]

\[ d(x, y) := \cos^{-1} \left( \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} \right) \]

(angle between \( \mathbf{x} + \mathbf{y} \))

Each \( h \in \mathcal{H} \) based on a different unit vector \( \mathbf{w} \),

\[ \text{Pr}[h(x) = h(y)] = 1 - \frac{d(x, y)}{\pi} \]
$H$ is LSD for any $r \in [0, \pi)$ and $c > 0$ with $cr \leq \pi$.

\[ p_1 = 1 - \frac{r}{\pi} \]

\[ p_2 = 1 - \frac{cr}{\pi} \]
The LSH Data Structure:

Any $(x, d) \oplus (r, c, r, p_1, p_2) \rightarrow \text{LSH } H$

Want structure for

$\text{NN}(q, r, c)$ queries.

$k, l$: two parameters we'll fix later
For $\omega \leq l$ to $l$

For $j \leq l$ to $k$

pull $h_{\omega,j}$ from $H$

For $x \in S$

For $\omega \leq l$ to $l$

store $x$ in bucket $g_{\omega}(x) =$

$< h_{\omega}(x), h_{\omega}(x), ...$

$\omega_1, h_{\omega_1}^{(x)}, ... >$
(so l buckets for \( x \), each indexed by a \( k \)-dim. vector)

Query time: \( \text{NNL}(q, r_c) \): find \( g_1(q), g_2(q), \ldots, g_l(q) \)

For \( i \leftarrow 1 \) to \( l \)

check elements of \( S \) in \( g_w(q) \) in any order
Return any $x$ we check s.t. $d(q, x) \leq cr$.

Give up after checking all elements in $q'$s buckets - or - after 4h elements, whichever is sooner.
Query time: \( O(kd) \).
Space: \( O(ned) \).

Suppose \( \exists x^* \in S \) s.t.
\[ d(q, x^*) \leq r. \]
We want

1) For some \( \hat{u}, g_{\hat{u}}(x^*) = g_{\hat{u}}(q) \)

2) There are \( \leq 4L \) elements \( x \in S \) s.t.
\[ d(x, q) > cr \] for some \( \hat{u}, \quad g_{\hat{u}}(x) = g_{\hat{u}}(q) \).
Will guarantee both happen with constant probability.

\[ \varphi := \frac{\ln(p_1)}{\ln(p_2)} = \log_{p_2} p_1. \]

\[ \varphi \approx \gamma_0 \text{ in each case we saw.} \]
Thm: Let \( l_i = n^\phi \) and
\[
 k_i = \frac{\log n}{\log (\gamma p_i)} = -\log \rho_i n
\]

Both properties above hold with constant probability.

Proof (for 2).
Let \( x' \in S \) s.t. \( d(x', q) > cn \).
\[
 Pr [ g_{\hat{u}} (x') = g_{\hat{u}} (q) ] \leq p_z^k
\]
\[
p_z^k = p_z^{\log \rho_z n} = \sqrt{n}
\]
For any fixed $i$, expected $\#$ of such $x$ is $\leq n \cdot \gamma n = 1$. Total expected $\#$ bad $x'$ across $i \leq L$. Markov's inequality: $\text{Prob. that } > 4L \text{ bad elements is } \leq \frac{1}{4}$. 
Proof for $D$: \( \Pr \left( x^* \in S \cap d(x^*, q) \leq r \right) \)

\[
\Pr \left[ g_{\hat{u}}(x^*) \neq g_{\hat{u}}(q) \right] \\
\leq 1 - p_1^k \\
= 1 - \log p_2 n \\
= 1 - \log p_2 \delta \\
= 1 - 1/n^{\delta} \\
l = n^{\delta}, \text{ so } \Pr \left[ g_{\hat{u}}(x^*) = g_{\hat{u}}(q) \right] \\
\text{for all } \hat{u} \in n^{\delta} \text{ is } \\
\leq \left( 1 - 1/n^{\delta} \right)^n \leq 1/e
Both properties hold with prob. \( \geq 1 - \frac{1}{e} \cdot \frac{1}{n} \geq \frac{1}{3}. \)

Example: If \( c = 2 \), can do \( \tilde{O}(\sqrt{n}) \) time queries with \( O(n^{1.5}) \) space.

Can boost success to prob. \( 1 - \frac{1}{n^6} \) by checking \( O(\log n) \) independent structures.