Understanding thy neighbors: 
Practical perspectives from modern analysis

Sanjoy Dasgupta and Samory Kpotufe
Key questions

1 **Statistical issues**: under what conditions does NN produce good predictions, and how should it be run?
   - When is 1-NN enough?
   - If using $k$-NN, what should $k$ be, roughly?
   - Is there a curse of dimension?
   - Does it adapt to latent structure: clusters, manifolds, etc?

2 **Algorithmic issues**: how to find nearest neighbors?
   - Data structures for fast NN
   - Parallelizing NN
   - Geometric tasks that build upon nearest neighbors: hierarchical clustering, minimum spanning tree, etc
Outline

1. Statistical properties of nearest neighbor
2. Algorithmic approaches to nearest neighbor search
Nearest neighbor classification

Given:

- \textit{training points} \((x_1, y_1), \ldots, (x_n, y_n) \in X \times \{0, 1\}\)
- \textit{query point} \(x\)

predict the label of \(x\) by looking at its nearest neighbor(s) among the \(x_i\).

- 1-NN returns the label of the nearest neighbor of \(x\) amongst the \(x_i\).
- \(k\)-NN returns the majority vote of the \(k\) nearest neighbors.
- \(k_n\)-NN lets \(k\) grow with \(n\).
The data space

Data points lie in a space $\mathcal{X}$ with distance function $\rho : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$.

- Most common scenario: $\mathcal{X} = \mathbb{R}^d$ and $\rho$ is Euclidean distance.
- Common more general setting: $(\mathcal{X}, \rho)$ is a metric space.
  - $\ell_p$ distances
  - Metrics obtained from user preferences/feedback
- Also of interest: more general distances.
  - KL divergence
  - Domain-specific dissimilarity measures
**Statistical learning theory setup**

**Training points come from the same source as future queries.**
- Underlying measure $\mu$ on $\mathcal{X}$ from which all points are generated.
- We call $(\mathcal{X}, \rho, \mu)$ a **metric measure space**.
- Label of $x$ is a coin flip with bias $\eta(x) = \Pr(Y = 1|X = x)$.

**Question:** why wouldn’t $\eta(x)$ always be either 0 or 1?

A classifier is a rule $h : \mathcal{X} \rightarrow \{0, 1\}$.
- Misclassification rate, or risk: $R(h) = \Pr(h(X) \neq Y)$.
- The **Bayes-optimal classifier**

$$h^*(x) = \begin{cases} 
1 & \text{if } \eta(x) > 1/2 \\
0 & \text{otherwise}
\end{cases},$$

has minimum risk, $R^* = R(h^*) = \mathbb{E}_X \min(\eta(X), 1 - \eta(X))$. 
Statistical questions

Let \( h_n \) be a classifier based on \( n \) labeled data points from the underlying distribution. \( R(h_n) \) is a random variable.

- **Consistency**: does \( R(h_n) \) converge to \( R^* \)?
  - 1-NN is not consistent. e.g. \( \mathcal{X} = \mathbb{R} \) and \( \eta \equiv 1/4 \).
Statistical questions

Let $h_n$ be a classifier based on $n$ labeled data points from the underlying distribution. $R(h_n)$ is a random variable.

- **Consistency**: does $R(h_n)$ converge to $R^*$?
  
  - 1-NN is not consistent. e.g. $\mathcal{X} = \mathbb{R}$ and $\eta \equiv 1/4$.
  - Neither is $k$-NN for fixed $k$.
  - Therefore, take $k_n$-NN classifier with $k_n \to \infty$.

What are minimal assumptions for consistency?
Statistical questions

Let $h_n$ be a classifier based on $n$ labeled data points from the underlying distribution. $R(h_n)$ is a random variable.

- **Consistency**: does $R(h_n)$ converge to $R^*$?
  - 1-NN is not consistent. e.g. $\mathcal{X} = \mathbb{R}$ and $\eta \equiv 1/4$.
  - Neither is $k$-NN for fixed $k$.
  - Therefore, take $k_n$-NN classifier with $k_n \to \infty$.

What are minimal assumptions for consistency?

- **Rates of convergence**: how fast does convergence occur?
  Rates depend upon smoothness of $\eta(x) = \Pr(Y = 1|X = x)$:

What is a suitable notion of smoothness, and rates?
Consistency results under continuity

Assume $\eta(x) = P(Y = 1|X = x)$ is continuous.
Let $h_n$ be the $k_n$-classifier, with $k_n \uparrow \infty$ and $k_n/n \downarrow 0$.

- Fix and Hodges (1951): Consistent in $\mathbb{R}^d$.

Proof outline: Let $x$ be a query point and let $x(1), \ldots, x(n)$ denote the training points ordered by increasing distance from $x$.

Training points are drawn from $\mu$, so the number of them in any ball $B$ is roughly $n \cdot \mu(B)$.

- Therefore $x(1), \ldots, x(k_n)$ lie in a ball centered at $x$ of probability mass $\approx k_n/n$. Since $k_n/n \downarrow 0$, we have $x(1), \ldots, x(k_n) \to x$.
- By continuity, $\eta(x(1)), \ldots, \eta(x(k_n)) \to \eta(x)$.
- By law of large numbers, when tossing many coins of bias roughly $\eta(x)$, the fraction of 1s will be approximately $\eta(x)$. Thus the majority vote of their labels will approach $h^*(x)$. 
Universal consistency in $\mathbb{R}^d$

Stone (1977): consistency in $\mathbb{R}^d$ assuming only measurability.
Universal consistency in $\mathbb{R}^d$

Stone (1977): consistency in $\mathbb{R}^d$ assuming only measurability.

Lusin’s thm: for any measurable $\eta$, for any $\epsilon > 0$, there is a continuous function that differs from it on at most $\epsilon$ fraction of points.

Training points in the red region can cause trouble. What fraction of query points have one of these as their nearest neighbor?
Universal consistency in $\mathbb{R}^d$

Stone (1977): consistency in $\mathbb{R}^d$ assuming only measurability.

Lusin’s thm: for any measurable $\eta$, for any $\epsilon > 0$, there is a continuous function that differs from it on at most $\epsilon$ fraction of points.

Training points in the red region can cause trouble. What fraction of query points have one of these as their nearest neighbor?

Geometric result: at most a constant number! And this yields consistency.
A key geometric fact

Pick any $n$ points in $\mathbb{R}^d$. Pick one of these points, $x$. At most how many of the remaining points can have $x$ as its nearest neighbor?
A key geometric fact

Pick any $n$ points in $\mathbb{R}^d$. Pick one of these points, $x$. At most how many of the remaining points can have $x$ as its nearest neighbor?
At most $5^d$ [Stone].
A key geometric fact

Pick any $n$ points in $\mathbb{R}^d$. Pick one of these points, $x$. At most how many of the remaining points can have $x$ as its nearest neighbor?
At most $5^d$ [Stone].
A key geometric fact

Pick any $n$ points in $\mathbb{R}^d$. Pick one of these points, $x$. At most how many of the remaining points can have $x$ as its nearest neighbor? At most $5^d$ [Stone].

But this argument fails in general metric measure spaces $(\mathcal{X}, \rho, \mu)$. 
Universal consistency in metric spaces [Chaudhuri-D’ 14]

Preiss [80’s]: An infinite-dimensional space in which consistency fails
Cerou-Guyader ’06: Conditions for universal consistency in metric spaces
Universal consistency in metric spaces [Chaudhuri-D’ 14]

Preiss [80’s]: An infinite-dimensional space in which consistency fails

Cerou-Guyader ’06: Conditions for universal consistency in metric spaces

Let \((X, d, \mu)\) be a separable metric measure space in which the Lebesgue differentiation property holds: for any bounded measurable \(f\),

\[
\lim_{r \downarrow 0} \frac{1}{\mu(B(x, r))} \int_{B(x, r)} f \, d\mu = f(x)
\]

for almost all (\(\mu\)-a.e.) \(x \in X\).
Universal consistency in metric spaces [Chaudhuri-D’ 14]

Preiss [80’s]: An infinite-dimensional space in which consistency fails
Cerou-Guyader ’06: Conditions for universal consistency in metric spaces

Let \((\mathcal{X}, d, \mu)\) be a separable metric measure space in which the Lebesgue
differentiation property holds: for any bounded measurable \(f\),

\[
\lim_{r \downarrow 0} \frac{1}{\mu(B(x, r))} \int_{B(x, r)} f \, d\mu = f(x)
\]

for almost all (\(\mu\)-a.e.) \(x \in \mathcal{X}\).

- If \(k_n \to \infty\) and \(k_n/n \to 0\), then \(R_n \to R^*\) in probability.
- If in addition \(k_n/\log n \to \infty\), then \(R_n \to R^*\) almost surely.
Universal consistency in metric spaces [Chaudhuri-D’ 14]

Preiss [80’s]: An infinite-dimensional space in which consistency fails  
Cerou-Guyader ’06: Conditions for universal consistency in metric spaces

Let $(\mathcal{X}, d, \mu)$ be a separable metric measure space in which the Lebesgue differentiation property holds: for any bounded measurable $f$,

$$
\lim_{r \downarrow 0} \frac{1}{\mu(B(x, r))} \int_{B(x, r)} f \, d\mu = f(x)
$$

for almost all $(\mu$-a.e.) $x \in \mathcal{X}$.

- If $k_n \to \infty$ and $k_n/n \to 0$, then $R_n \to R^*$ in probability.
- If in addition $k_n/\log n \to \infty$, then $R_n \to R^*$ almost surely.

Examples of such spaces: finite-dimensional normed spaces; doubling metric measure spaces.
Universal consistency in metric spaces

Query $x$; training points by increasing distance from $x$ are $x_{(1)}, \ldots, x_{(n)}$. 
Universal consistency in metric spaces

Query \( x \); training points by increasing distance from \( x \) are \( x^{(1)}, \ldots, x^{(n)} \).

1. Since \( k_n/n \to 0 \), we have \( x^{(1)}, \ldots, x^{(k_n)} \to x \).
Universal consistency in metric spaces

Query $x$; training points by increasing distance from $x$ are $x_{(1)}, \ldots, x_{(n)}$.

1. Since $k_n/n \to 0$, we have $x_{(1)}, \ldots, x_{(k_n)} \to x$.

2. Earlier argument using continuity: $\eta(x_{(1)}), \ldots, \eta(x_{(k_n)}) \to \eta(x)$.
   In this case, the $k_n$-NN are coins of roughly the same bias as $x$. 
Query \( x \); training points by increasing distance from \( x \) are \( x_{(1)}, \ldots, x_{(n)} \).

1. Since \( k_n/n \to 0 \), we have \( x_{(1)}, \ldots, x_{(k_n)} \to x \).
2. Earlier argument using continuity: \( \eta(x_{(1)}), \ldots, \eta(x_{(k_n)}) \to \eta(x) \).
   In this case, the \( k_n \)-NN are coins of roughly the same bias as \( x \).
3. It suffices that \( \text{average}(\eta(x_{(1)}), \ldots, \eta(x_{(k_n)})) \to \eta(x) \).
Universal consistency in metric spaces

Query $x$; training points by increasing distance from $x$ are $x_{(1)}, \ldots, x_{(n)}$.

1. Since $k_n/n \to 0$, we have $x_{(1)}, \ldots, x_{(k_n)} \to x$.

2. Earlier argument using continuity: $\eta(x_{(1)}), \ldots, \eta(x_{(k_n)}) \to \eta(x)$. 
   In this case, the $k_n$-NN are coins of roughly the same bias as $x$.

3. It suffices that average($\eta(x_{(1)}), \ldots, \eta(x_{(k_n)})$) $\to \eta(x)$.

4. $x_{(1)}, \ldots, x_{(k_n)}$ lie in some ball $B(x, r)$. 
   For suitable $r$, they are random draws from $\mu$ restricted to $B(x, r)$. 


Universal consistency in metric spaces

Query \( x \); training points by increasing distance from \( x \) are \( x_1, \ldots, x_n \).

1. Since \( k_n/n \to 0 \), we have \( x_1, \ldots, x_{k_n} \to x \).

2. Earlier argument using continuity: \( \eta(x_1), \ldots, \eta(x_{k_n}) \to \eta(x) \).
   In this case, the \( k_n \)-NN are coins of roughly the same bias as \( x \).

3. It suffices that \( \text{average}(\eta(x_1), \ldots, \eta(x_{k_n})) \to \eta(x) \).

4. \( x_1, \ldots, x_{k_n} \) lie in some ball \( B(x, r) \).
   For suitable \( r \), they are random draws from \( \mu \) restricted to \( B(x, r) \).

5. \( \text{average}(\eta(x_1), \ldots, \eta(x_{k_n})) \) is close to the average \( \eta \) in this ball:

\[
\frac{1}{\mu(B(x, r))} \int_{B(x, r)} \eta \, d\mu.
\]
Universal consistency in metric spaces

Query $x$; training points by increasing distance from $x$ are $x_{(1)}, \ldots, x_{(n)}$.

1. Since $k_n/n \to 0$, we have $x_{(1)}, \ldots, x_{(k_n)} \to x$.
2. Earlier argument using continuity: $\eta(x_{(1)}), \ldots, \eta(x_{(k_n)}) \to \eta(x)$.
   In this case, the $k_n$-NN are coins of roughly the same bias as $x$.
3. It suffices that average($\eta(x_{(1)}), \ldots, \eta(x_{(k_n)})$) $\to \eta(x)$.
4. $x_{(1)}, \ldots, x_{(k_n)}$ lie in some ball $B(x, r)$.
   For suitable $r$, they are random draws from $\mu$ restricted to $B(x, r)$.
5. average($\eta(x_{(1)}), \ldots, \eta(x_{(k_n)})$) is close to the average $\eta$ in this ball:
   \[
   \frac{1}{\mu(B(x, r))} \int_{B(x, r)} \eta \, d\mu.
   \]
6. As $n$ grows, this ball $B(x, r)$ shrinks. Thus it is enough that
   \[
   \lim_{r \downarrow 0} \frac{1}{\mu(B(x, r))} \int_{B(x, r)} \eta \, d\mu = \eta(x).
   \]
Rates of convergence

**Bad news:** curse of dimension

**Good news:** adaptive to
  - Intrinsic low dimension (e.g. manifold structure)
  - Smoothness of boundary
Smoothness and margin conditions

- The usual smoothness condition in $\mathbb{R}^d$: $\eta$ is $\alpha$-Holder continuous if for some constant $L$, for all $x, x'$,

$$|\eta(x) - \eta(x')| \leq L\|x - x'\|^{\alpha}. $$
Smoothness and margin conditions

- The usual smoothness condition in \( \mathbb{R}^d \): \( \eta \) is \( \alpha \)-Holder continuous if for some constant \( L \), for all \( x, x' \),

  \[
  |\eta(x) - \eta(x')| \leq L \|x - x'\|^\alpha.
  \]

- Mammen-Tsybakov \( \beta \)-margin condition: For some constant \( C \), for any \( t \), we have \( \mu(\{x : |\eta(x) - 1/2| \leq t\}) \leq Ct^\beta \).

Width-\( t \) margin around decision boundary

\[\eta(x)\]

1

1/2

x

Smoothness and margin conditions

• The usual smoothness condition in $\mathbb{R}^d$: $\eta$ is $\alpha$-Holder continuous if for some constant $L$, for all $x, x'$,

$$|\eta(x) - \eta(x')| \leq L\|x - x'\|^\alpha.$$

• Mammen-Tsybakov $\beta$-margin condition: For some constant $C$, for any $t$, we have $\mu(\{x : |\eta(x) - 1/2| \leq t\}) \leq Ct^\beta$.

Width-$t$ margin around decision boundary

• Audibert-Tsybakov: Suppose these two conditions hold, and that $\mu$ is supported on a regular set with $0 < \mu_{\text{min}} < \mu < \mu_{\text{max}}$. Then $\mathbb{E}R_n - R^*$ is $\Omega(n^{-\alpha(\beta+1)/(2\alpha+d)}).$
Smoothness and margin conditions

- The usual smoothness condition in $\mathbb{R}^d$: $\eta$ is $\alpha$-Holder continuous if for some constant $L$, for all $x, x'$,
  $$|\eta(x) - \eta(x')| \leq L\|x - x'\|^{\alpha}.$$ 

- Mammen-Tsybakov $\beta$-margin condition: For some constant $C$, for any $t$, we have $\mu(\{x : |\eta(x) - 1/2| \leq t\}) \leq Ct^\beta$.

Width-$t$ margin around decision boundary

- Audibert-Tsybakov: Suppose these two conditions hold, and that $\mu$ is supported on a regular set with $0 < \mu_{\text{min}} < \mu < \mu_{\text{max}}$. Then $\mathbb{E}R_n - R^*$ is $\Omega(n^{-\alpha(\beta+1)/(2\alpha+d)})$.
  
  Under these conditions, for suitable $(k_n)$, this rate is achieved by $k_n$-NN.
How much does $\eta$ change over an interval?

- The usual notions relate this to $|x - x'|$.
- For NN: more sensible to relate to $\mu([x, x'])$. 

**A better smoothness condition for NN** [Chaudhuri-D’14]

\[
\eta(x)
\]
A better smoothness condition for NN [Chaudhuri-D’14]

How much does $\eta$ change over an interval?

- The usual notions relate this to $|x - x'|$.
- For NN: more sensible to relate to $\mu([x, x'])$.

We will say $\eta$ is $\alpha$-smooth in metric measure space $(\mathcal{X}, \rho, \mu)$ if for some constant $L$, for all $x \in \mathcal{X}$ and $r > 0$,

$$|\eta(x) - \eta(B(x, r))| \leq L \mu(B(x, r))^\alpha,$$

where $\eta(B) = \text{average } \eta \text{ in ball } B = \frac{1}{\mu(B)} \int_B \eta \ d\mu$. 

A better smoothness condition for NN [Chaudhuri-D’14]

How much does $\eta$ change over an interval?

- The usual notions relate this to $|x - x'|$.
- For NN: more sensible to relate to $\mu([x, x'])$.

We will say $\eta$ is $\alpha$-smooth in metric measure space $(\mathcal{X}, \rho, \mu)$ if for some constant $L$, for all $x \in \mathcal{X}$ and $r > 0$,

$$|\eta(x) - \eta(B(x, r))| \leq L \mu(B(x, r))^\alpha,$$

where $\eta(B) = \text{average } \eta \text{ in ball } B = \frac{1}{\mu(B)} \int_B \eta \ d\mu$.

$\eta$ is $\alpha$-Holder continuous in $\mathbb{R}^d$, $\mu$ bounded below $\Rightarrow \eta$ is $(\alpha/d)$-smooth.
Rates of convergence under smoothness

Let $h_{n,k}$ denote the $k$-NN classifier based on $n$ training points. Let $h^*$ be the Bayes-optimal classifier.

Suppose $\eta$ is $\alpha$-smooth in $(\mathcal{X}, \rho, \mu)$. Then for any $n, k$,

1. For any $\delta > 0$, with probability at least $1 - \delta$ over the training set,
   \[
   \Pr_X(h_{n,k}(X) \neq h^*(X)) \leq \delta + \mu(\{x : |\eta(x) - \frac{1}{2}| \leq C_1\sqrt{\frac{1}{k} \ln \frac{1}{\delta}}\})
   \]
   under the choice $k \propto n^{2\alpha/(2\alpha+1)}$.

2. $\mathbb{E}_n \Pr_X(h_{n,k}(X) \neq h^*(X)) \geq C_2 \mu(\{x : |\eta(x) - \frac{1}{2}| \leq C_3\sqrt{\frac{1}{k}}\})$.  

Rates of convergence under smoothness

Let $h_{n,k}$ denote the $k$-NN classifier based on $n$ training points. Let $h^*$ be the Bayes-optimal classifier.

Suppose $\eta$ is $\alpha$-smooth in $(X, \rho, \mu)$. Then for any $n, k$,

1. For any $\delta > 0$, with probability at least $1 - \delta$ over the training set,
   \[ \Pr_X(h_{n,k}(X) \neq h^*(X)) \leq \delta + \mu(\{x : |\eta(x) - \frac{1}{2}| \leq C_1 \sqrt{\frac{1}{k} \ln \frac{1}{\delta}}\}) \]
   under the choice $k \propto n^{2\alpha/(2\alpha+1)}$.

2. $E_n \Pr_X(h_{n,k}(X) \neq h^*(X)) \geq C_2 \mu(\{x : |\eta(x) - \frac{1}{2}| \leq C_3 \sqrt{\frac{1}{k}}\}$).

These upper and lower bounds are qualitatively similar for all smooth conditional probability functions:

the probability mass of the width-$\frac{1}{\sqrt{k}}$ margin around the decision boundary.
Variants of nearest neighbor rules

1. Quantization strategies

2. Subsampling
Quantization: reduce the data

\[ \{X_i\}_{i=1}^n \]
Quantization: reduce the data

Assign $\{X_i\}$ to representatives $Q \equiv \{q\}$
Quantization: reduce the data

Pick $q's$ in $Q$ close to $x$

Kpotufe-Verma (2017): pick $Q$ to be an $\epsilon$-net.

Favorable empirical performance: small rise in error rate, significant speedup in query time.

Kontorovich-Weiss-Sabato (2017): pick $Q$ to be a suitable $\epsilon$-cover.

Then: 1-NN using $Q$ is consistent.
Quantization: reduce the data

Pick \( q \)'s in \( Q \) close to \( x \)

1. Kpotufe-Verma (2017): pick \( Q \) to be an \( \epsilon \)-net.
   Favorable empirical performance: small rise in error rate, significant speedup in query time.
Quantization: reduce the data

Pick \( q \)'s in \( Q \) close to \( x \)

1. Kpotufe-Verma (2017): pick \( Q \) to be an \( \epsilon \)-net.
   Favorable empirical performance: small rise in error rate, significant speedup in query time.

2. Kontorovich-Weiss-Sabato (2017): pick \( Q \) to be a suitable \( \epsilon \)-cover.
   Then: 1-NN using \( Q \) is consistent.
Subsampling: reduce data and parallelize

Data: \( \{(X_i, Y_i)\}_{i=1}^{n}, \ Y \in \{0, 1\} \).

Repeat for \( t = 1, 2, \ldots, N \):

- Let \( S_t \) be a random subsample of \( m \ll n \) points.

To classify \( x \): compute 1-NN wrt to each \( S_t \), take majority label.
**Subsampling: reduce data and parallelize**

Data: \( \{(X_i, Y_i)\}_{i=1}^{n}, Y \in \{0, 1\} \).

Repeat for \( t = 1, 2, \ldots, N \):

- Let \( S_t \) be a random subsample of \( m \ll n \) points

To classify \( x \): compute 1-NN wrt to each \( S_t \), take majority label.

Biau-Cerou-Guyader (2010), Samworth (2010):

- This is consistent.
- In fact, it is weighted \( k \)-NN.
  Each of \( x \)'s \( k \) nearest neighbors (in the original data set) will be its 1-NN in some fraction of \( S_t \).
- Asymptotically more accurate than \( k \)-NN.
Outline

1. Statistical properties of nearest neighbor
2. Algorithmic approaches to nearest neighbor search
The complexity of nearest neighbor search

Given a data set of $n$ points in a metric space $(X, \rho)$, build a data structure for efficiently answering subsequent nearest neighbor queries $q$.

- Data structure should take space $O(n)$
- Query time should be $o(n)$
The complexity of nearest neighbor search

Given a data set of $n$ points in a metric space $(\mathcal{X}, \rho)$, build a data structure for efficiently answering subsequent nearest neighbor queries $q$.

- Data structure should take space $O(n)$
- Query time should be $o(n)$

Unproven but common conjecture: either data structure size or query time must be exponential in the dimension of the space.

Bad case: for any $0 < \epsilon < 1$,

- Pick $2^{O(\epsilon^2 d)}$ points uniformly from the unit sphere in $\mathbb{R}^d$
- With high probability, all interpoint distances are $(1 \pm \epsilon)\sqrt{2}$
The complexity of nearest neighbor search

Given a data set of $n$ points in a metric space $(\mathcal{X}, \rho)$, build a data structure for efficiently answering subsequent nearest neighbor queries $q$.

- Data structure should take space $O(n)$
- Query time should be $o(n)$

Unproven but common conjecture: either data structure size or query time must be exponential in the dimension of the space.

Bad case: for any $0 < \epsilon < 1$,

- Pick $2^{O(\epsilon^2 d)}$ points uniformly from the unit sphere in $\mathbb{R}^d$
- With high probability, all interpoint distances are $(1 \pm \epsilon)\sqrt{2}$

How can this bad case be defeated?
NN algorithms: an impressionistic history

- 1975: The k-d tree (Bentley and Friedman).
  Widely used, but algorithmic guarantees on weak footing.
**NN algorithms: an impressionistic history**

- **1975**: The $k$-d tree (Bentley and Friedman). Widely used, but algorithmic guarantees on weak footing.
- **1980s-1990s**: More tree structures (e.g. Clarkson, Mount). Could accommodate general metric spaces.
- **1990s-**: It's okay to fail sometimes (e.g. Clarkson, Kleinberg).
- **Late 1990s-**: Locality-sensitive hashing (Indyk, Motwani, Andoni). Hashing scheme with some failure probability, widely used.
- **Recently**: Binary hashing; resurgence of trees.
**NN algorithms: an impressionistic history**

- **1975**: The $k$-d tree (Bentley and Friedman). Widely used, but algorithmic guarantees on weak footing.

- **1980s-1990s**: More tree structures (e.g. Clarkson, Mount). Could accommodate general metric spaces.

- **1990s-**: It’s okay to fail sometimes (e.g. Clarkson, Kleinberg).
**NN algorithms: an impressionistic history**

- **1975:** The $k$-d tree (Bentley and Friedman).
  Widely used, but algorithmic guarantees on weak footing.

- **1980s-1990s:** More tree structures (e.g. Clarkson, Mount).
  Could accommodate general metric spaces.

- **1990s-:** It’s okay to fail sometimes (e.g. Clarkson, Kleinberg).

- **Late 1990s-:** Locality-sensitive hashing (Indyk, Motwani, Andoni).
  Hashing scheme with some failure probability, widely used.
NN algorithms: an impressionistic history

- **1975**: The $k$-d tree (Bentley and Friedman). Widely used, but algorithmic guarantees on weak footing.

- **1980s-1990s**: More tree structures (e.g. Clarkson, Mount). Could accommodate general metric spaces.

- **1990s-**: It’s okay to fail sometimes (e.g. Clarkson, Kleinberg).

- **Late 1990s-**: Locality-sensitive hashing (Indyk, Motwani, Andoni). Hashing scheme with some failure probability, widely used.

- **Recently**: binary hashing; resurgence of trees.
The $k$-d tree [Bentley-Friedman ’75]

Defeatist search:
• Return NN in query’s leaf node; maybe not the actual NN
• Time $O(\log n) + O(\# \text{(points in each leaf)})$

Comprehensive search:
• Always returns the NN
• Can take $O(n)$ time in some cases
The k-d tree [Bentley-Friedman '75]

Defeatist search:

• Return NN in query’s leaf node; maybe not the actual NN
• Time $O(\log n) + O(\#(\text{points in each leaf}))$
The \textit{k-d tree} \cite{Bentley-Friedman '75}

\textbf{Defeatist search:}
- Return NN in query’s leaf node; maybe not the actual NN
- Time $O(\log n) + O(\#(\text{points in each leaf}))$

\textbf{Comprehensive search:}
- Always returns the NN
- Can take $O(n)$ time in some cases
Trees for general distance spaces

- Ball trees for metric spaces [Omohundro '89]
- Bregman ball trees [Cayton '08]
- Vantage-point (VP) trees [Yianilos '91; Uhlmann '91]
Trees for general distance spaces

- Ball trees for metric spaces [Omohundro ’89]

Diagram showing two overlapping circles with points inside, illustrating the concept of ball trees for metric spaces.
Trees for general distance spaces

- Ball trees for metric spaces [Omohundro ’89]
- Bregman ball trees [Cayton ’08]


Trees for general distance spaces

- Ball trees for metric spaces [Omohundro ’89]
- Bregman ball trees [Cayton ’08]
- Vantage-point (VP) trees [Yianilos ’91; Uhlmann ’91]
Controlling the complexity of NN search

Recall canonical bad case: points uniformly distributed over a $d$-dimensional unit ball.
Controlling the complexity of NN search

Recall canonical bad case: points uniformly distributed over a $d$-dimensional unit ball.

1. Methods that are adaptive to intrinsic dimension.
Recall canonical bad case: points uniformly distributed over a $d$-dimensional unit ball.

1. Methods that are adaptive to intrinsic dimension.
2. Methods that return approximate nearest neighbors.
Cover trees for metric spaces

Beygelzimer-Kakade-Langford ’06:
- Hierarchical cover of an arbitrary metric space
- Space $O(n)$, permits dynamic insertion and deletion of data points
- Query time $O(\text{poly}(c) \log n)$

A finite set $X$ in a metric space has expansion rate $c$ if for any point $x$ and any radius $r > 0$,
$$|B(x, 2r) \cap X| \leq c \cdot |B(x, r) \cap X|.$$
Cover trees for metric spaces

Beygelzimer-Kakade-Langford '06:
• Hierarchical cover of an arbitrary metric space
• Space $O(n)$, permits dynamic insertion and deletion of data points
• Query time $O(\text{poly}(c) \log n)$

A finite set $X$ in a metric space has expansion rate $c$ if for any point $x$ and any radius $r > 0$,

$$|B(x, 2r) \cap X| \leq c \cdot |B(x, r) \cap X|.$$
Variants of k-d trees with guarantees

**Random projection trees:** In each cell of the tree, pick split direction uniformly at random from the unit sphere in $\mathbb{R}^d$.

**Perturbed split:** after projection, pick $\beta \in [1/4, 3/4]$ and split at the $\beta$-fractile point.

Failure probability for defeatist search is $< 1/2$ if each leaf has $O(d)$ points, where $d$ is the doubling dimension of the data. [D-Sinha '13]
Variants of $k$-d trees with guarantees

**Random projection trees:** In each cell of the tree, pick split direction uniformly at random from the unit sphere in $\mathbb{R}^d$.

![Diagram of random projection trees]

**Perturbed split:** after projection, pick $\beta \in [1/4, 3/4]$ and split at the $\beta$-fractile point.

Failure probability for defeatist search is $< 1/2$ if each leaf has $O(d_\circ^{d_\circ})$ points, where $d_\circ$ is the **doubling dimension** of the data. [D-Sinha ’13]
Doubling dimension

[Assouad ’83; Gupta-Krauthgamer-Lee ’03]

Set $S \subset \mathbb{R}^d$ has doubling dimension $d_0$ if for any (Euclidean) ball $B$, the subset $S \cap B$ can be covered by $2^{d_0}$ balls of half the radius.
Doubling dimension

[Assouad '83; Gupta-Krauthgamer-Lee '03]

Set $S \subset \mathbb{R}^d$ has doubling dimension $d_o$ if for any (Euclidean) ball $B$, the subset $S \cap B$ can be covered by $2^{d_o}$ balls of half the radius.

Example: $S =$ line has doubling dimension 1.
Doubleing dimension

[Assouad '83; Gupta-Krauthgamer-Lee '03]

Set $S \subset \mathbb{R}^d$ has doubling dimension $d_o$ if for any (Euclidean) ball $B$, the subset $S \cap B$ can be covered by $2^{d_o}$ balls of half the radius.

1. Example: $S =$ line has doubling dimension 1.

2. A $k$-dimensional flat has doubling dimension $c_o k$ for some absolute constant $c_o$. 
Doubling dimension

[Assouad '83; Gupta-Krauthgamer-Lee '03]

Set $S \subset \mathbb{R}^d$ has doubling dimension $d_o$ if for any (Euclidean) ball $B$, the subset $S \cap B$ can be covered by $2^{d_o}$ balls of half the radius.

1. Example: $S =$ line has doubling dimension 1.

2. A $k$-dimensional flat has doubling dimension $c_0 k$ for some absolute constant $c_0$.

3. If $S$ has diameter $\Delta$ and doubling dimension $d_o$, then for any $\epsilon > 0$, it has an $\epsilon$-cover of size $\leq (2\Delta/\epsilon)^{d_o}$. 
Doubling dimension

[Assouad ’83; Gupta-Krauthgamer-Lee ’03]

Set $S \subset \mathbb{R}^d$ has doubling dimension $d_o$ if for any (Euclidean) ball $B$, the subset $S \cap B$ can be covered by $2^{d_o}$ balls of half the radius.

1. Example: $S = \text{line}$ has doubling dimension 1.

\[ \text{Example: } S = \text{line has doubling dimension } 1. \]

2. A $k$-dimensional flat has doubling dimension $c_o k$ for some absolute constant $c_o$.

3. If $S$ has diameter $\Delta$ and doubling dimension $d_o$, then for any $\epsilon > 0$, it has an $\epsilon$-cover of size $\leq (2\Delta/\epsilon)^{d_o}$.

4. If $S$ has doubling dimension $d_o$, then so does any subset of $S$. 
The doubling dimension of sparse sets

Set $S \subset \mathbb{R}^d$ has doubling dimension $d_o$ if for any (Euclidean) ball $B$, the subset $S \cap B$ can be covered by $2^{d_o}$ balls of half the radius.

1. A set of $n$ points has doubling dimension at most $\log n$.
   Proof: It can be covered by $n$ balls of any radius.
The doubling dimension of sparse sets

Set $S \subset \mathbb{R}^d$ has doubling dimension $d_o$ if for any (Euclidean) ball $B$, the subset $S \cap B$ can be covered by $2^{d_o}$ balls of half the radius.

1. A set of $n$ points has doubling dimension at most $\log n$.
   Proof: It can be covered by $n$ balls of any radius.

2. If sets $S_1, \ldots, S_m$ each have doubling dimension $\leq d_o$, then $S_1 \cup \cdots \cup S_m$ has doubling dimension $\leq d_o + \log m$.
   Proof: $S_i \cap B$ can be covered by $2^{d_o}$ balls of half the radius. Therefore, at most $m2^{d_o}$ balls are needed for the union.
The doubling dimension of sparse sets

Set $S \subset \mathbb{R}^d$ has doubling dimension $d_o$ if for any (Euclidean) ball $B$, the subset $S \cap B$ can be covered by $2^{d_o}$ balls of half the radius.

1. A set of $n$ points has doubling dimension at most $\log n$.
   Proof: It can be covered by $n$ balls of any radius.

2. If sets $S_1, \ldots, S_m$ each have doubling dimension $\leq d_o$, then $S_1 \cup \cdots \cup S_m$ has doubling dimension $\leq d_o + \log m$.
   Proof: $S_i \cap B$ can be covered by $2^{d_o}$ balls of half the radius. Therefore, at most $m2^{d_o}$ balls are needed for the union.

3. Suppose each point in $S \subset \mathbb{R}^d$ has $\leq k$ nonzero coordinates. Then $S$ has doubling dimension $\leq c_0 k + k \log d$.
   Proof: $S$ is the union of $\binom{d}{k}$ flats of dimension $k$; we’ve seen that each flat has doubling dimension $\leq c_0 k$. 


The doubling dimension of manifolds

A Riemannian submanifold $M \subset \mathbb{R}^p$ has condition number $\leq 1/\tau$ if normals to $M$ of length $\tau$ don’t intersect:

If $M \subset \mathbb{R}^p$ is a $k$-dimensional manifold of condition number $1/\tau$, then its neighborhoods of radius $\tau$ have doubling dimension $O(k)$. 

Locality-sensitive hashing [Indyk-Motwani-Andoni]

Typical hash function $h_i$:
random projection + binning

$$h_i(x) = \left\lfloor \frac{r_i \cdot x + b}{w} \right\rfloor$$

- $r_i$ is a random direction
- $b$ is a random offset
- $w$ is the bin width
Locality-sensitive hashing \[\text{[Indyk-Motwani-Andoni]}\]

Typical hash function \(h_i:\)
random projection + binning

\[
h_i(x) = \left\lfloor \frac{r_i \cdot x + b}{w} \right\rfloor
\]

- \(r_i\) is a random direction
- \(b\) is a random offset
- \(w\) is the bin width

For any data set \(x_1, \ldots, x_n\), query \(q\): probability < 1 of failing to return an approximate NN.
**Locality-sensitive hashing** [Indyk-Motwani-Andoni]

Typical hash function $h_i$: random projection + binning

$$h_i(x) = \left\lfloor \frac{r_i \cdot x + b}{w} \right\rfloor$$

- $r_i$ is a random direction
- $b$ is a random offset
- $w$ is the bin width

- For any data set $x_1, \ldots, x_n$, query $q$: probability $< 1$ of failing to return an approximate NN.
- To reduce this probability, make $t$ tables. Space: $O(nt)$. 
Approximate nearest neighbor

For data set $S \subset \mathbb{R}^d$ and query $q$, a $c$-approximate nearest neighbor is any $x \in S$ such that

$$\|x - q\| \leq c \cdot \min_{z \in S} \|z - q\|.$$
Approximate nearest neighbor

For data set $S \subset \mathbb{R}^d$ and query $q$, a $c$-approximate nearest neighbor is any $x \in S$ such that

$$\|x - q\| \leq c \cdot \min_{z \in S} \|z - q\|.$$ 

Complexity of approximate NN search in Euclidean space:

- Data structure size $n^{1+1/c^2}$
- Query time $n^{1/c^2}$
Approximate nearest neighbor

For data set $S \subset \mathbb{R}^d$ and query $q$, a $c$-approximate nearest neighbor is any $x \in S$ such that

$$\|x - q\| \leq c \cdot \min_{z \in S} \|z - q\|.$$ 

Complexity of approximate NN search in Euclidean space:

- Data structure size $n^{1+1/c^2}$
- Query time $n^{1/c^2}$

Caution: the same value of $c$ can have very different implications for different data sets.
Approximate nearest neighbor

The MNIST data set of handwritten digits:

What % of $c$-approximate nearest neighbors have the wrong label?
Approximate nearest neighbor

The MNIST data set of handwritten digits:

What % of $c$-approximate nearest neighbors have the wrong label?

<table>
<thead>
<tr>
<th>$c$</th>
<th>1.0</th>
<th>1.2</th>
<th>1.4</th>
<th>1.6</th>
<th>1.8</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error rate (%)</td>
<td>3.1</td>
<td>9.0</td>
<td>18.4</td>
<td>29.3</td>
<td>40.7</td>
<td>51.4</td>
</tr>
</tbody>
</table>

But LSH also does well on exact NN search!
Approximate nearest neighbor

The MNIST data set of handwritten digits:

What % of $c$-approximate nearest neighbors have the wrong label?

<table>
<thead>
<tr>
<th>$c$</th>
<th>1.0</th>
<th>1.2</th>
<th>1.4</th>
<th>1.6</th>
<th>1.8</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error rate (%)</td>
<td>3.1</td>
<td>9.0</td>
<td>18.4</td>
<td>29.3</td>
<td>40.7</td>
<td>51.4</td>
</tr>
</tbody>
</table>

But LSH also does well on **exact** NN search!
Hash tables versus trees

As long as these structures are randomized, can use:

- **collection of LSH tables**
- **forest of trees**

Experimental comparisons, e.g. V. Hyvonen, T. Roos et al (2016).
Relevant books
