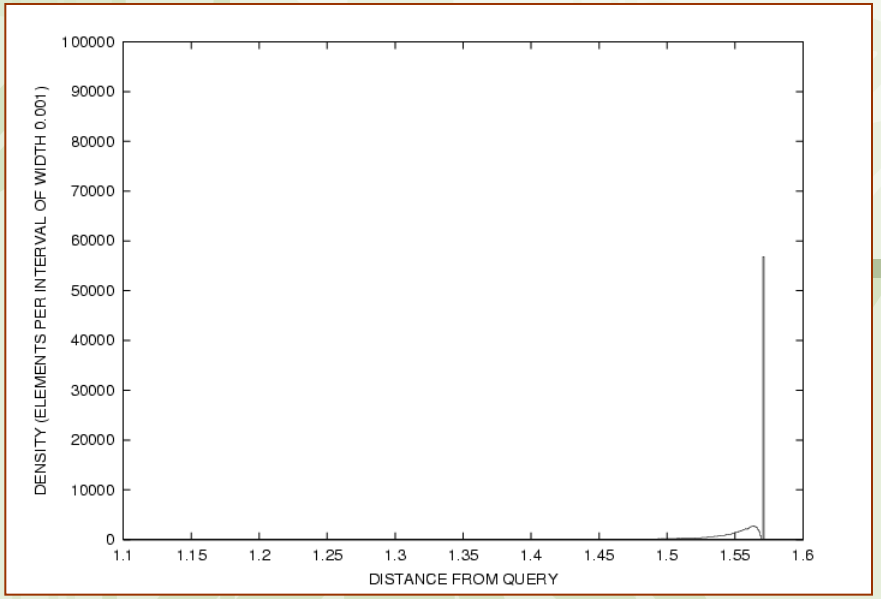
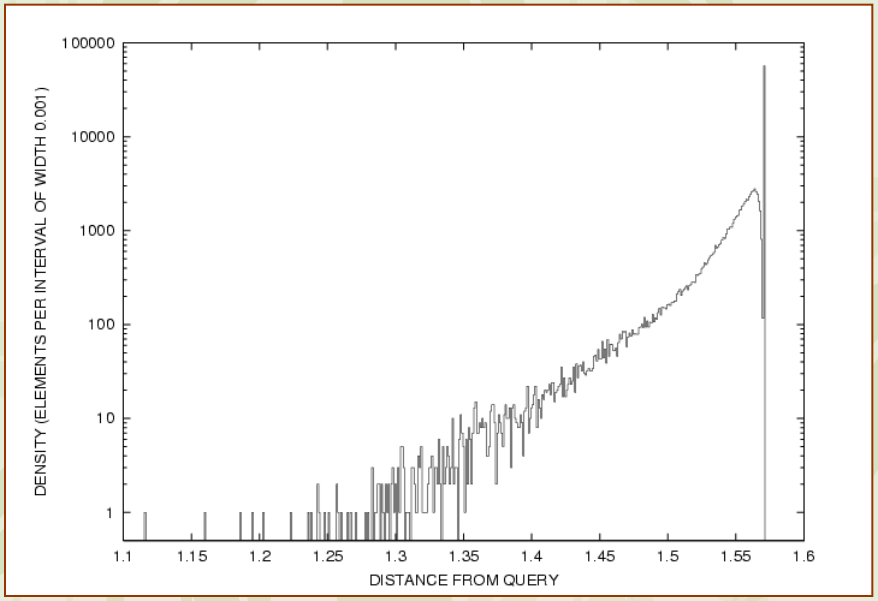
Rank Cover Trees for Nearest-Neighbor Search

Michael E. Houle
National Institute of Informatics

Michael Nett
National Institute of Informatics
University of Tokyo

The Curse of Dimensionality





Searching high-dimensional data:

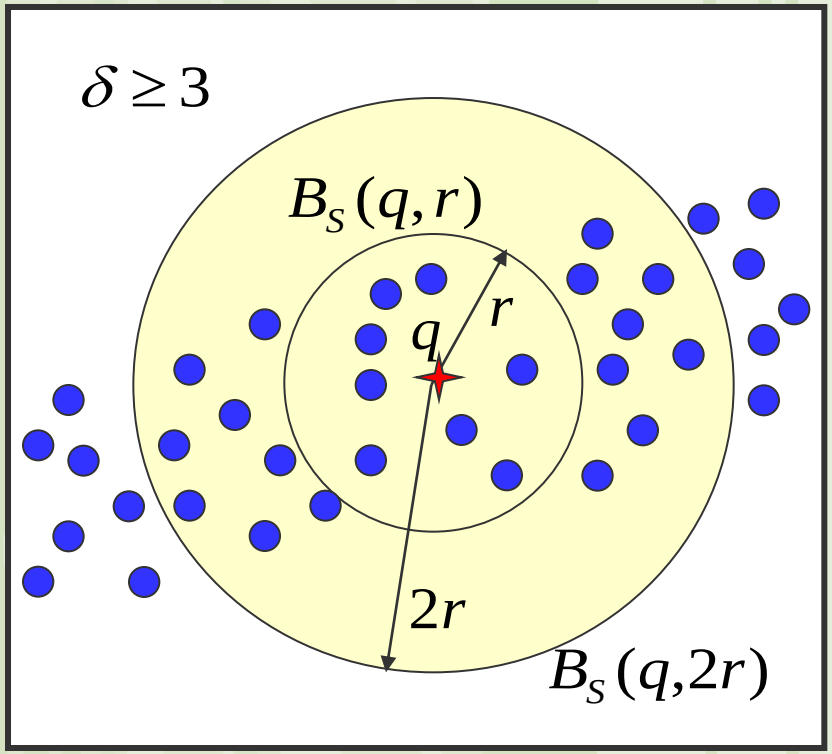
- Exact similarity queries require close to linear time.
- Data organization is a major challenge.
- 2D and 3D intuition does not apply!

Example: LA-Times 127738x6590 text data set, vector angle metric

Intrinsic Dimension

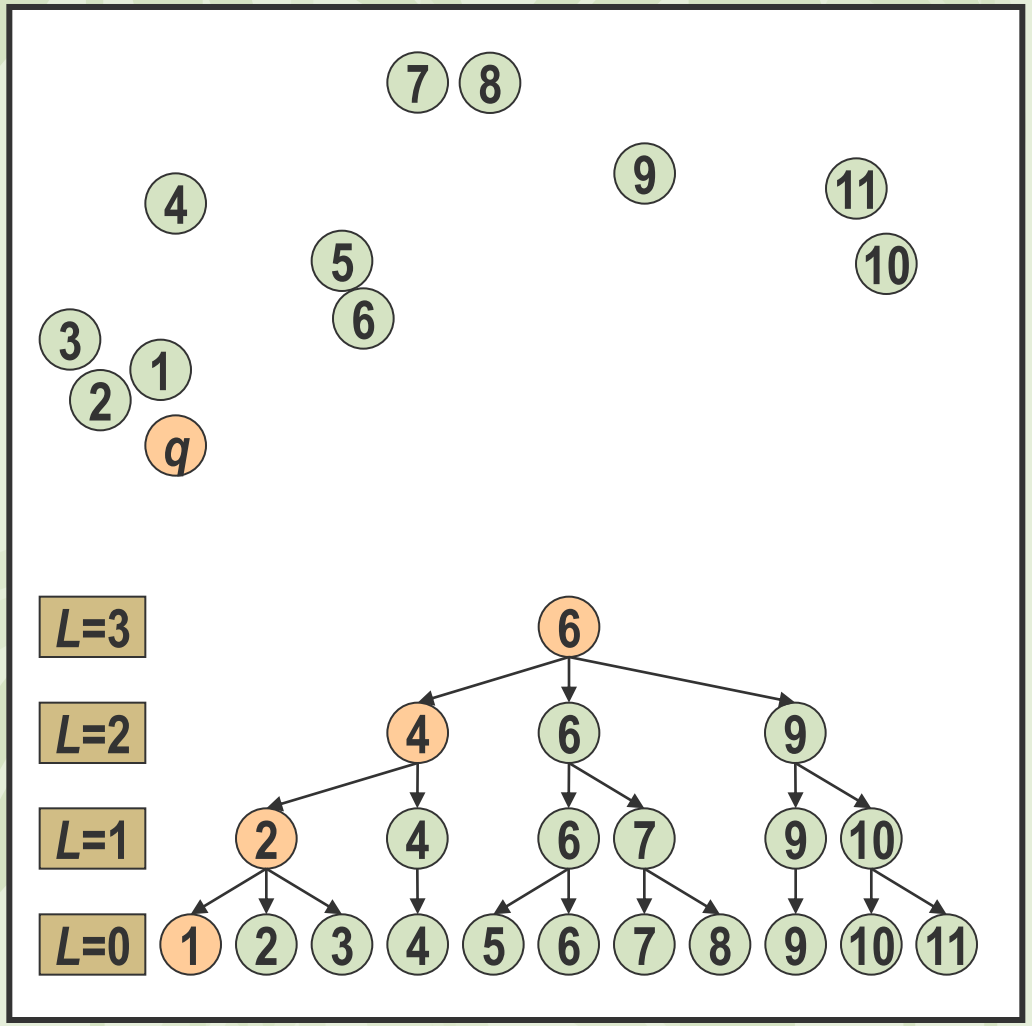
- Expansion rate of set *S*:
  - Introduced by Karger & Ruhl (2002).
  - Used to measure of the cost of an expanding search in the vicinity of a query point *q*.
  - Maximum ratio δ of the number of points in two balls centered at *q*.
- Expansion dimension: $D = \log_2 \delta$
  - Measure of intrinsic dimensionality of *S*.
  - If representational dimension is *m*…
  - Doubling the radius of a sphere → volume increases by factor 2^m.

$$|B_S(q,r)| \geq b$$
$$\Rightarrow |B_S(q,2r)| \leq \delta \cdot |B_S(q,r)|$$



Rank Cover Tree

- New tree index for similarity search based on the Cover Tree.
  - Design based on neighbor ranks to the query, instead of distances to the query.
  - Exact *k*-NN queries with extremely high probability.
  - Fast approximate queries.
- Metric space data
  - Triangle inequality satisfied.
  - No assumed knowledge of data.



Comparison of RCT with CT

- CT is exact, RCT is correct with very high probability.
- RCT achieves much smaller dependence on the intrinsic dimensionality while still being sublinear in *n*.
- CT real cost involves keeping track of nodes that lie in regions of diameters of very large length (exponential in 2) $\rightarrow$ for some distance measures, all data points could lie in these bounds until the very lowest levels of the search!
- RCT real costs are decided through the explicit choice of a coverage parameter ψ.

Operation	CT Cost	RCT Cost (*h*=3)	RCT Cost (*h*=4)	RCT Cost (*h*=8)
1-NN Query	$\delta^{12}\log_2 n$	$\delta^{4.97}n^{2/3}$	$\delta^{5.57}n^{1/2}$	$\delta^{7.01}n^{1/4}$