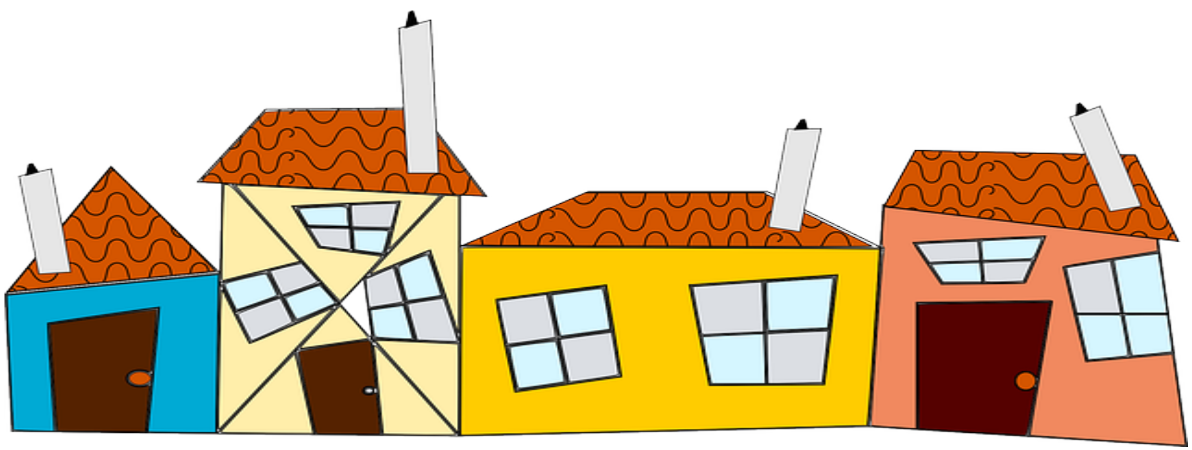
David R. Thompson
Jet Propulsion Laboratory, California Institute of Technology

Nearest Neighbors and the Curse of Dimensionality

Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology

Objectives

1. Find nearest neighbors efficiently
2. Understand the curse of dimensionality and its implications for pattern recognition
3. Know some general approaches to solve it



Nearest neighbors

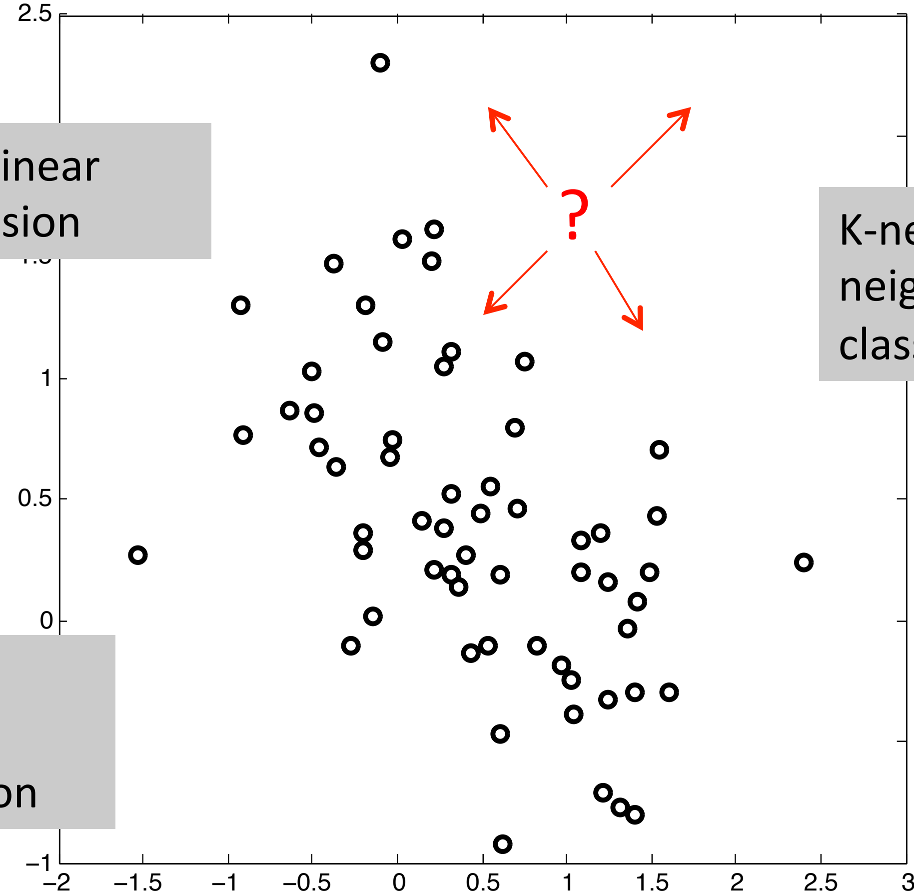
Pixabay.com, public domain CC0

Local pattern recognition

Local linear regression

K-nearest-neighbor classification

Kernel density estimation



… All rely on quickly finding points near the query!

david.r.thompson@jpl.nasa.gov

Finding nearest neighbors – 1D

How to efficiently find nearest neighbors?

$\boxed{4}$

X = [10 9 14 30 100 5 32 -4 3 72]

Finding nearest neighbors – 1D

How to efficiently find nearest neighbors?

$\boxed{4}$

Sequential search: linear time ☹

X = [10 9 14 30 100 5 32 -4 3 72]

david.r.thompson@jpl.nasa.gov

Finding nearest neighbors – 1D

How to efficiently find nearest neighbors?

$\boxed{4}$

Sequential search: linear time ☹

X = [10 9 14 30 100 5 32 -4 3 72]

SORT

X’ = [-4 3 5 9 10 14 30 32 72 100]

Finding nearest neighbors – 1D

How to efficiently find nearest neighbors?

4

Sequential search: linear time ☹

X = [10 9 14 30 100 5 32 -4 3 72]

SORT

4

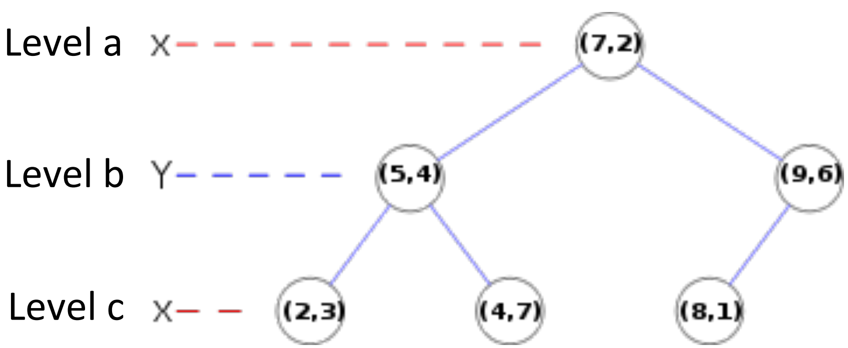
Binary search: logarithmic time ☺

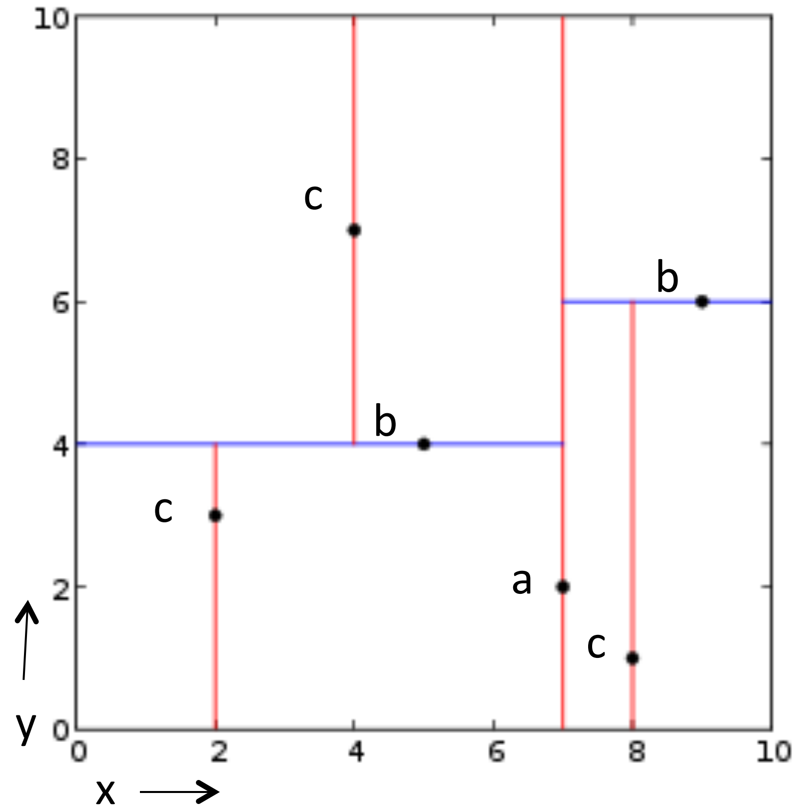
X' = [-4 3 5 9 10 14 30 32 72 100]

2 to 8 dimensions: *K-D Trees*

- Each node splits the space with a *hyperplane*
- Which one? Cycle through axis-aligned hyperplanes
- Typical search is O(log n)

Split:





Source: Wikipedia

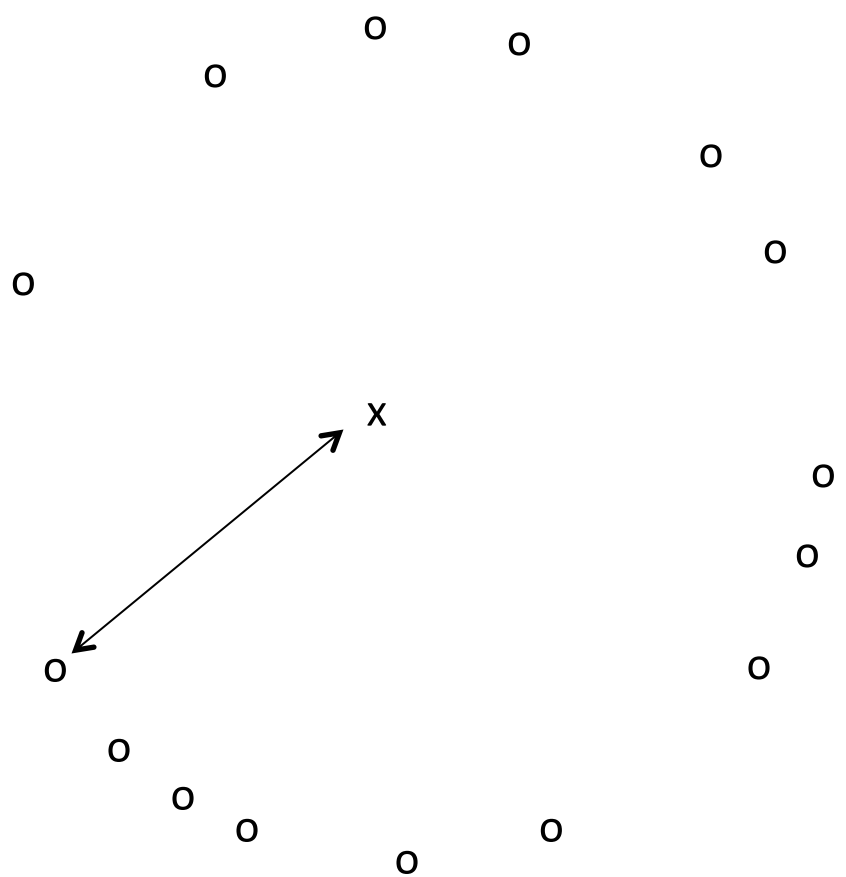
> 8 dimensions: Use *approximate nearest neighbors*

Above 8-10 dimensions, even partitioning structures are little better than brute force.

Approximate nearest neighbor methods can improve computation by orders of magnitude.

Source: Wikipedia

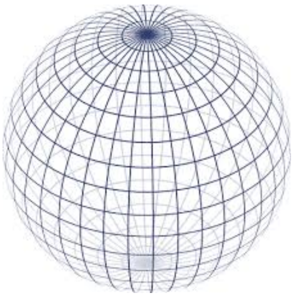
A more fundamental problem… When is nearest neighbor meaningful?



The curse of dimensionality

Volume of an N-Ball

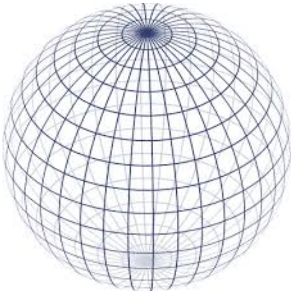
$$V_n(R) = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)} R^n,$$



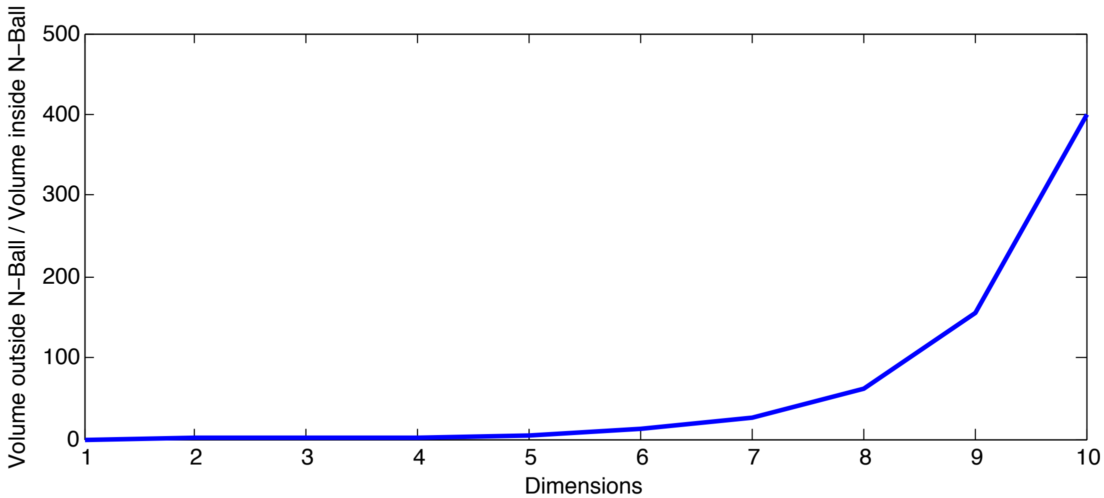
The curse of dimensionality

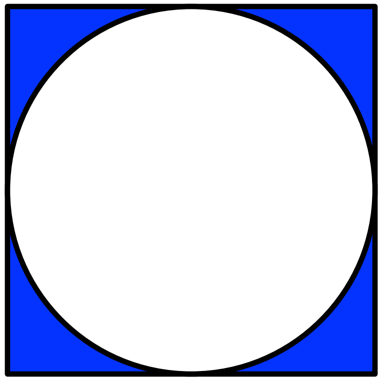
Volume of an N-Ball

$$V_n(R) = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)} R^n,$$



Inscribed inside a hypercube





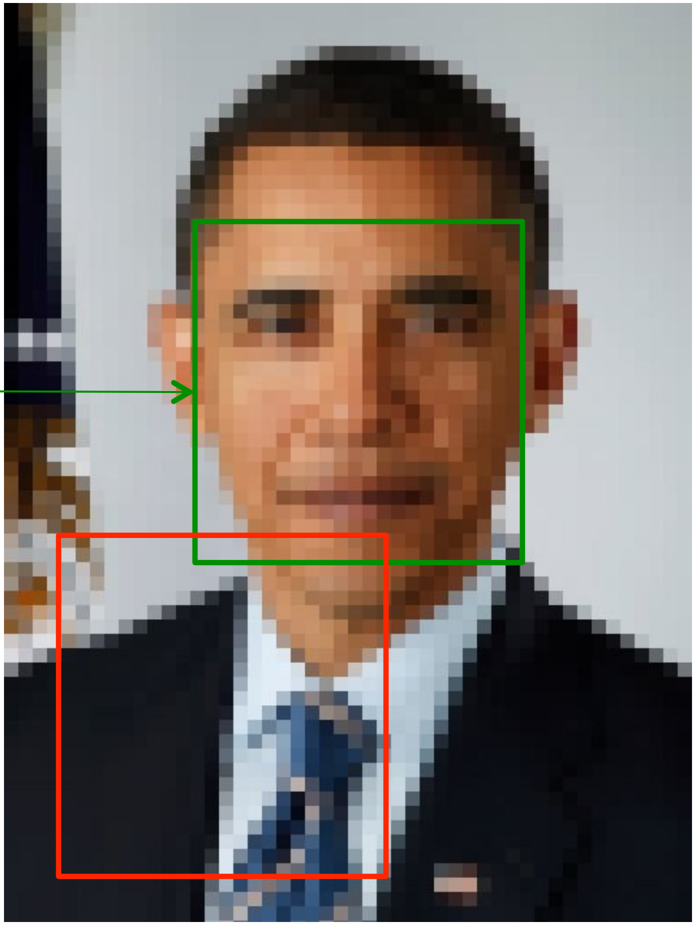
The curse of dimensionality

As dimensions increase …

- Euclidean distances become less meaningful
- Uniform distributions become exponentially harder to sample
- Many parameters become polynomially harder to estimate
- Data becomes more difficult to visualize

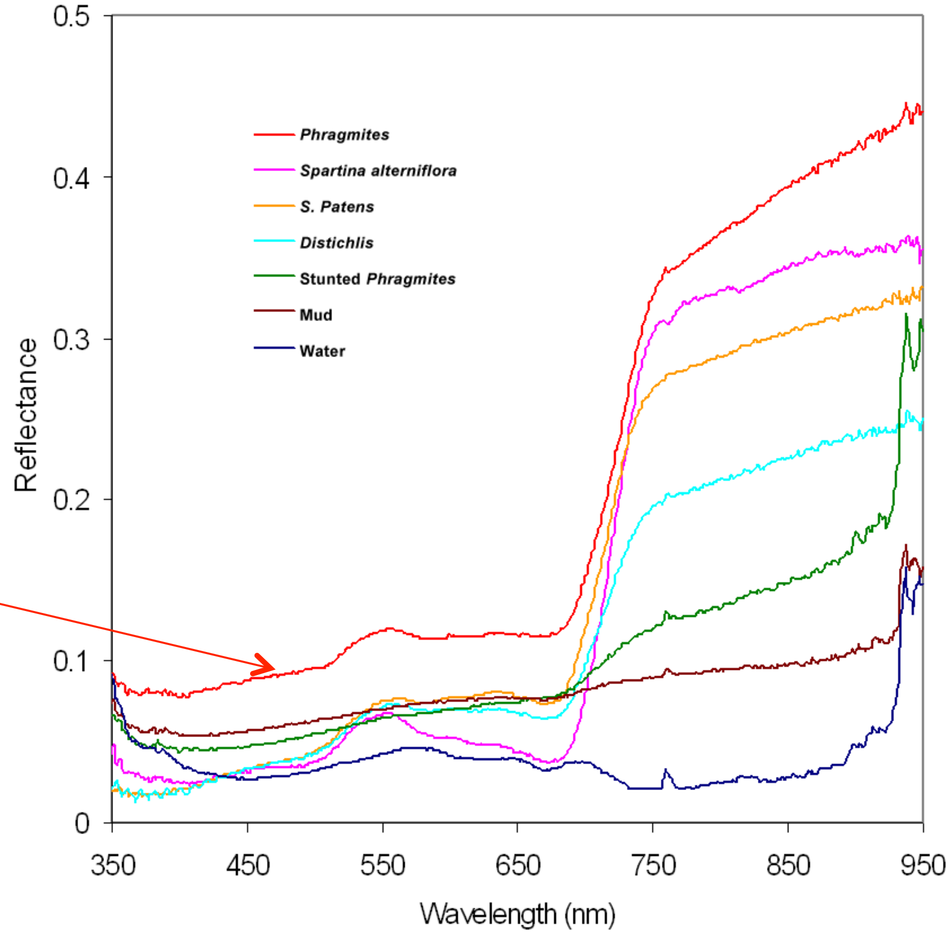
Face Recognition

20 pixels x 20 pixels =
400 dimensions



Reflectance Spectroscopy

600nm at 1nm spectral resolution = 600 dimensions



Artigas & Yang, *Urban Habitats 2004*

Solutions?

- Rely only on pattern recognition methods that are robust to high dimensions
- Represent the data differently
  - Hand-crafted features
  - Use a subset of features
  - Linear projections
  - Nonlinear projections

Summary

Finding nearest neighbors is not trivial.

- **1D?** Use a sorted list.
- **2-8D?** Use a k-d tree
- **>8D?** Use approximate nearest neghbors

High-dimensional problems are subject to the *curse of dimensionality*