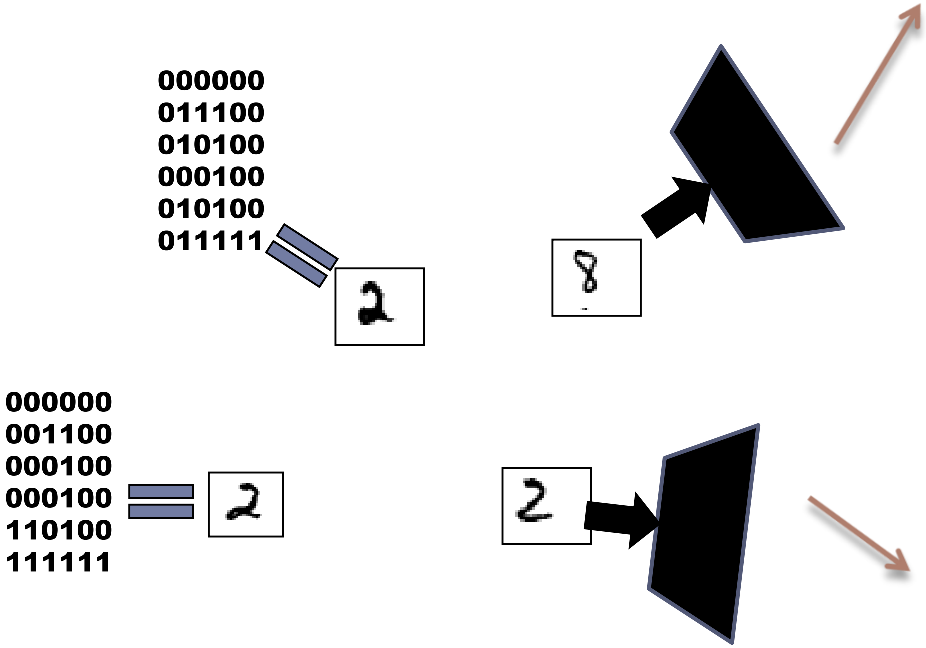
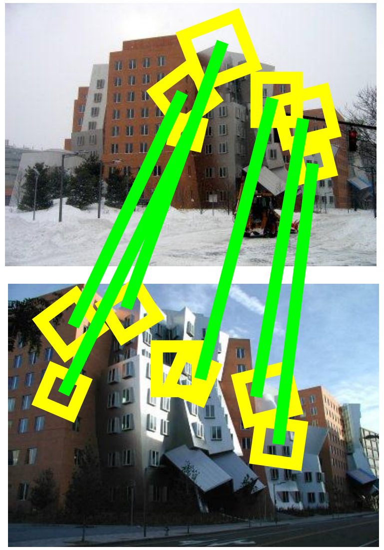
Geometry of Similarity Search

Alex Andoni
(Columbia University)

Find pairs of similar images

Measuring similarity





objects ⇒ high-dimensional vectors

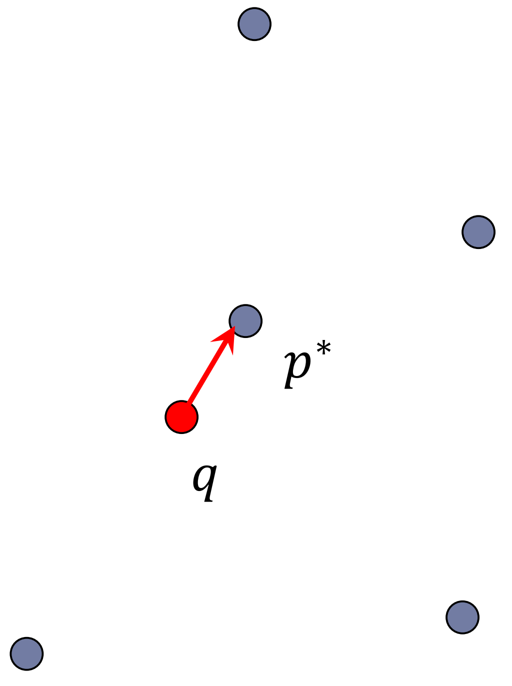
similarity ⇒ distance b/w vectors

$\{0,1\}^d$	R^d	Sets of points
Hamming dist.	Euclidean dist.	Earth-Mover Distance

Image courtesy of Kristen Grauman

Problem: Nearest Neighbor Search (NNS)

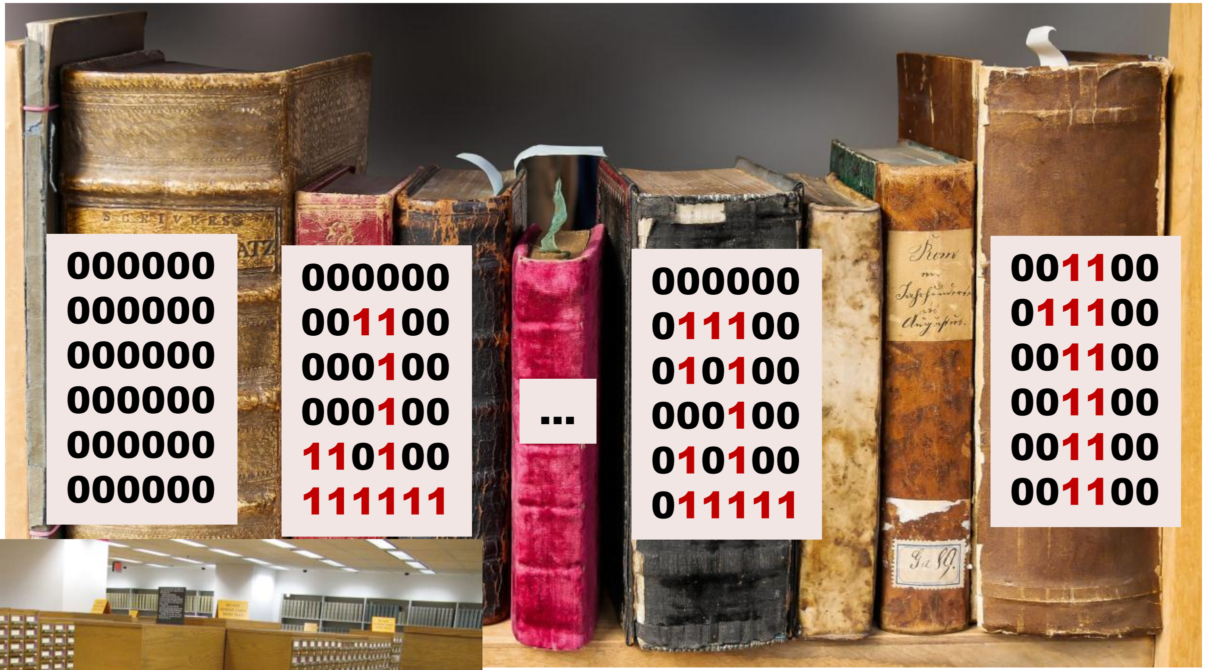
- Preprocess: a set P of points
- Query: given a query point q, report a point $p^* \in P$ with the smallest distance to q
- Primitive for: finding all similar pairs
  - But also clustering problems, and many other problems on large set of multi-feature objects
- Applications:
  - speech/image/video/music recognition, signal processing, bioinformatics, etc…



n: number of points
d: dimension

Preamble: How to check for an exact match ?

just pre-sort !



Preprocess:

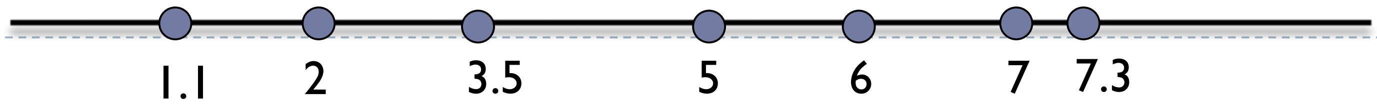
Sort the points

Query:

Perform *binary search*

Query time	Space
$O(\log n)$	$O(n)$

Also works for NNS for 1-dimensional vectors…



High-dimensional case

$\{0,1\}^d$
Hamming dist.

$n = 1{,}000{,}000{,}000$
$d = 400$

Underprepared: no preprocessing

Overprepared: store an answer for every *possible* query

Algorithm	Query time	Space
No indexing	$O(n \cdot d)$	$O(n \cdot d)$
Full indexing	$O(d)$	2^d

unaffordable if $d \gg \log n$

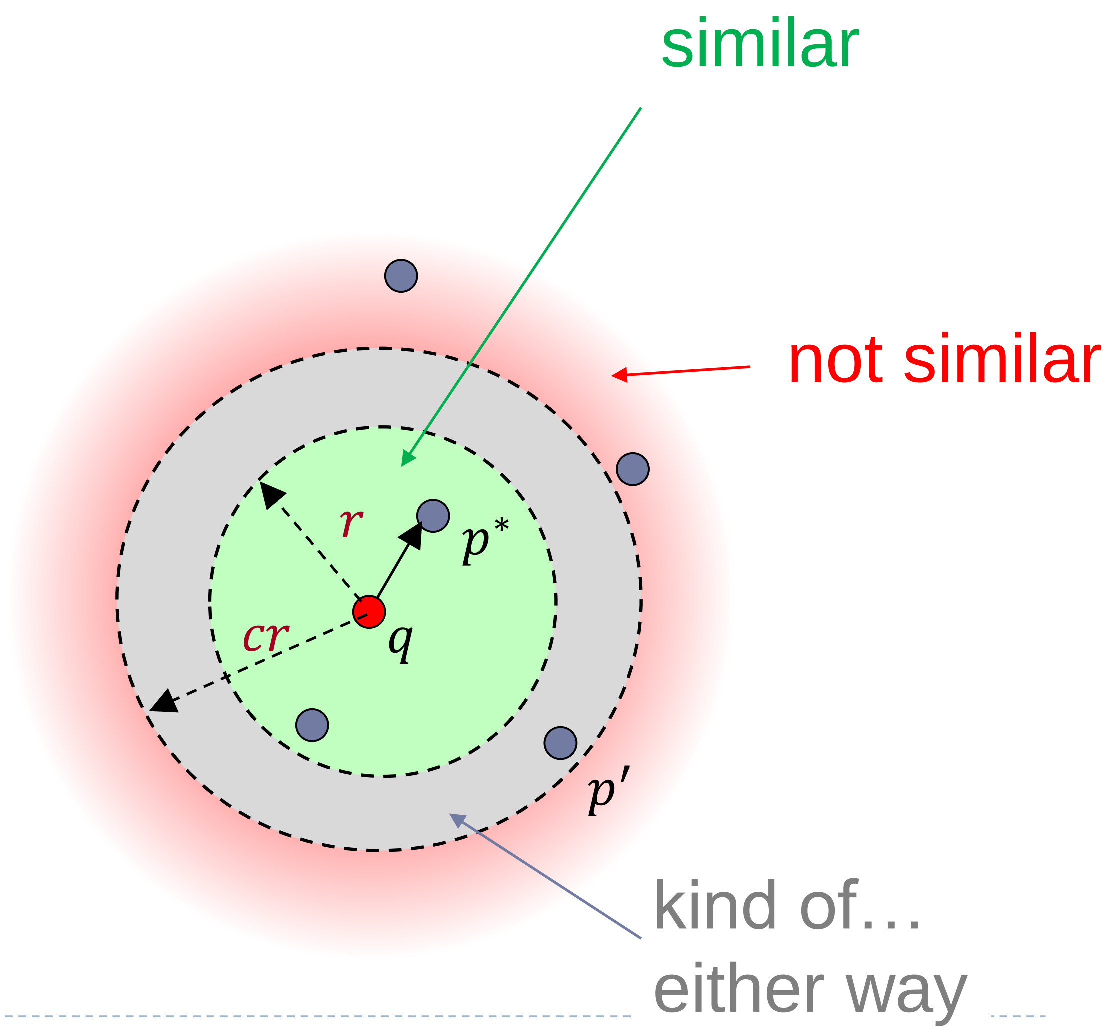
Best indexing ?	$O(d)$	$O(n \cdot d)$
A little better indexing ?	$n^{0.99}$	$O(n^2)$

Curse of dimensionality: would refute a (very) strong version of $P \neq NP$ conjecture [Williams'04]

Relaxed problem: Approximate Near Neighbor Search

c-approximate

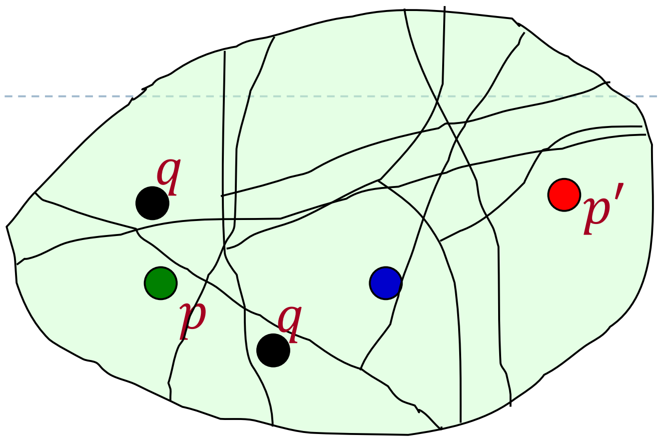
- r-near neighbor: given a query point q, report a point $p' \in P$ s.t. $||p' - q|| \leq cr$
  - as long as there is some point within distance r
- Remarks:
  - In practice: used as a filter
  - Randomized algorithms: each point reported with 90% probability
  - Can use to solve nearest neighbor too [HarPeled-Indyk-Motwani'12]



Approach: Locality Sensitive Hashing

[Indyk-Motwani'98]

Map: points → codes: s.t.
"similar" ⇔ "exact match"



randomized

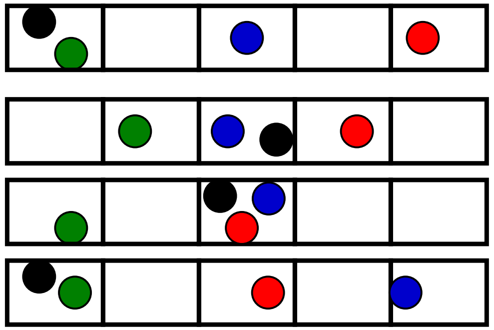
Map g on R^d s.t. for any points p, q

- for *similar pairs* (when $\|q-p\| \leq r$)

$P_1 =$ $\Pr[g(q)=g(p)]$ is not-too-low

- for *dissimilar pairs* (when $\|q-p'\| > cr$)

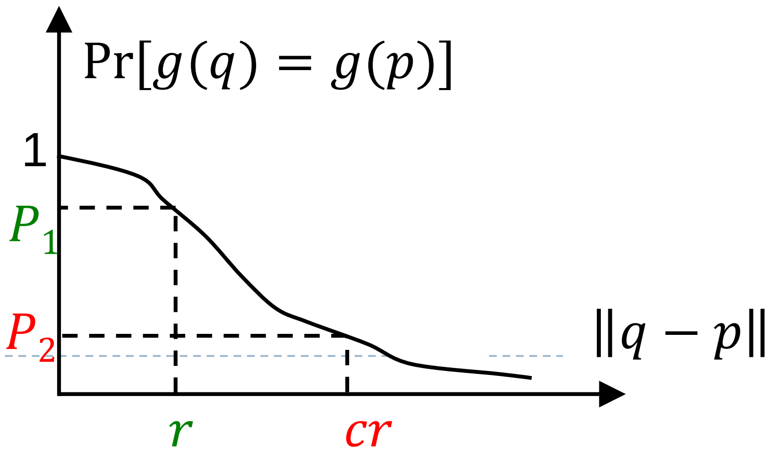
$P_2 =$ $\Pr[g(q)=g(p)]$ is low



several indexes

Use ~~an index~~ on $g(p)$ for $p \in P$

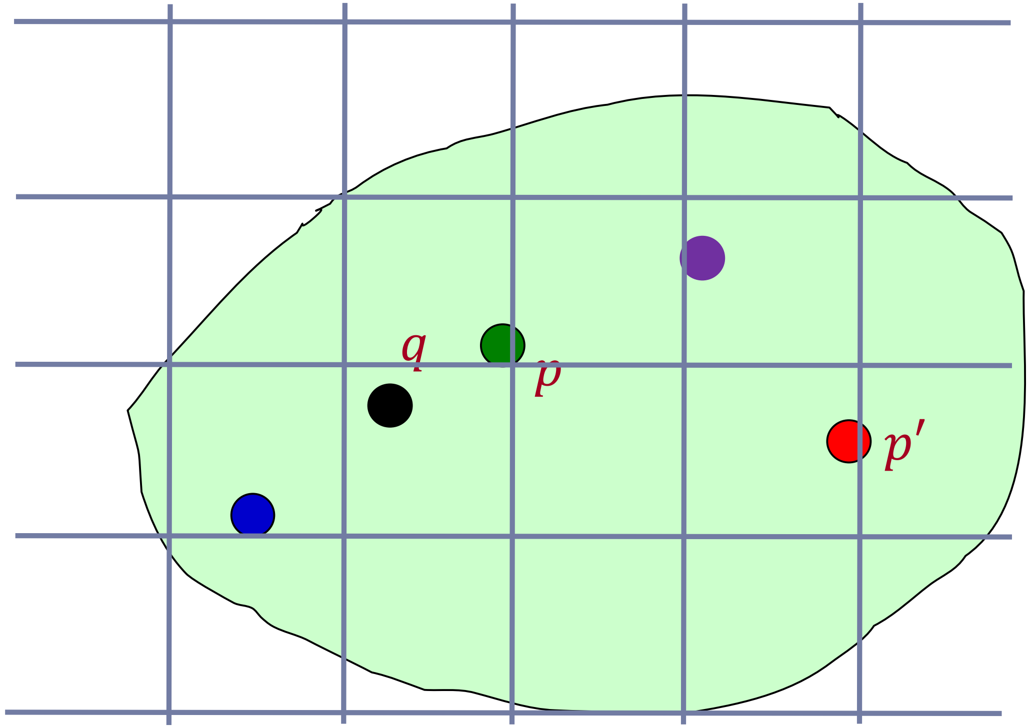
n^ρ, where $\rho = \frac{\log 1/P_1}{\log 1/P_2}$



How to construct good maps?

Map #1 : random grid

[Datar-Indyk-Immorlica-Mirrokni'04]



Map g:

- partition in a regular grid
- randomly shifted
- randomly rotated

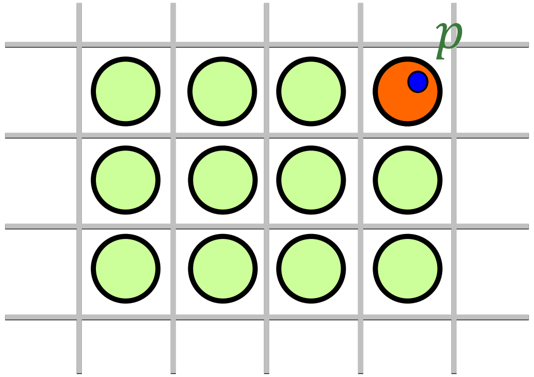
Space	Time	Exponent	$c = 2$
$n^{1+\rho}$	n^{ρ}	$\rho = 1/c$	$\rho = 1/2$

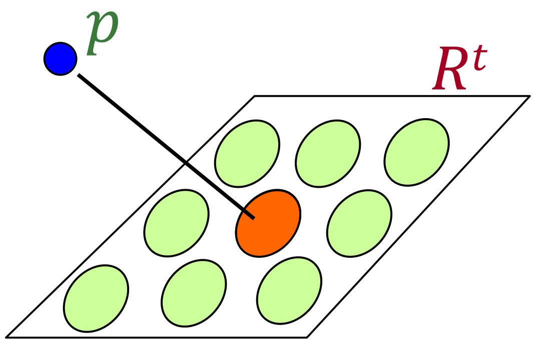
Can we do better?

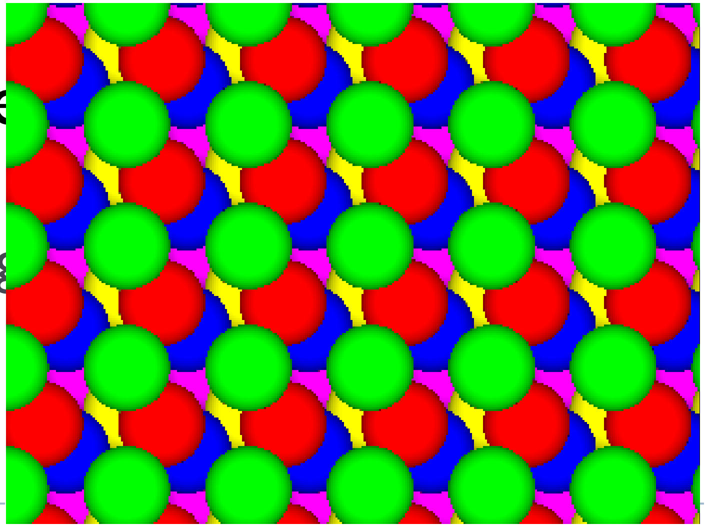
Map #2 : ball carving

[A-Indyk'06]

- Regular grid → grid of balls
  - p can hit empty space, so take more such grids until p is in a ball
- How many grids?
  - about d^d
  - start by projecting in dimension t
- Choice of re
  - ρ closer to
  - Number of g







Space	Time	Exponent	$c = 2$
$n^{1+\rho}$	n^{ρ}	$\rho \to 1/c^2$	$\rho \to 1/4$

Similar space partitions ubiquitous:

- Approximation algorithms [Goemans, Williamson 1995], [Karger, Motwani, Sudan 1995], [Charikar, Chekuri, Goel, Guha, Plotkin 1998], [Chlamtac, Makarychev, Makarychev 2006], [Louis, Makarychev 2014]
- Spectral graph partitioning [Lee, Oveis Gharan, Trevisan 2012], [Louis, Raghavendra, Tetali, Vempala 2012]
- Spherical cubes [Kindler, O'Donnell, Rao, Wigderson 2008]
- Metric embeddings [Fakcharoenphol, Rao, Talwar 2003], [Mendel, Naor 2005]
- Communication complexity [Bogdanov, Mossel 2011], [Canonne, Guruswami, Meka, Sudan 2015]

LSH Algorithms for Euclidean space

Space	Time	Exponent	$c = 2$	Reference
$n^{1+\rho}$	n^{ρ}	$\rho = 1/c$	$\rho = 1/2$	[IM'98, DIIM'04]
		$\rho \approx 1/c^2$	$\rho = 1/4$	[AI'06]

Is there even better LSH map?

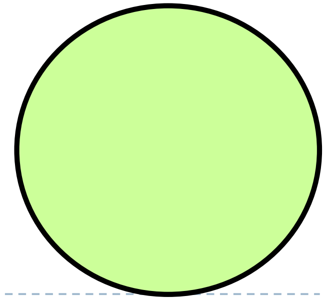
NO: any map must satisfy

$$\rho \geq 1/c^2$$

[Motwani-Naor-Panigrahy'06, O'Donell-Wu-Zhou'11]

Example of **isoperimetry**, example of which is question:

- Among bodies in R^d of volume 1, which has the lowest perimeter?
- A ball!



Some other LSH algorithms

- Hamming distance
  - g: pick a random coordinate(s) [IM'98]
- Manhattan distance:
  - g: cell in a randomly shifted grid
- Jaccard distance between sets:
  - $J(A,B) = \frac{A \cap B}{A \cup B}$
  - g: pick a random permutation π on the words

$$g(A) = \min_{a \in A} \pi(a)$$

min-wise hashing

[Broder'97, Christiani-Pagh'17]

To be or not to be

To search or not to search

be not or sketch to

...11101...

be not or search to

...01111...

...21102...

...01122...

{be,not,or,to}

{not,or,to,search}

be

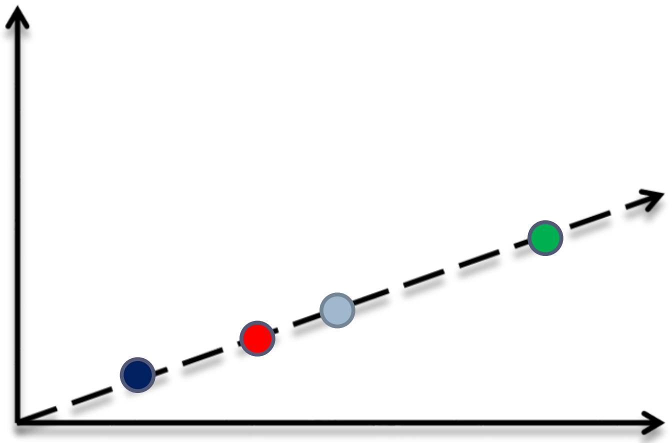
to

for π=be,to,search,or,not

LSH is tight… what's next?

Datasets with additional structure

[Clarkson'99,
Karger-Ruhl'02,
Krauthgamer-Lee'04,
Beygelzimer-Kakade-Langford'06,
Indyk-Naor'07,
Dasgupta-Sinha'13,
Abdullah-A.-Krauthgamer-Kannan'14,…]



Space-time trade-offs…

[Panigrahy'06, A.-Indyk'06, Kapralov'15,
A.-Laarhoven-Razenshteyn-Waingarten'17]

Are we really done with basic NNS algorithms?

Beyond Locality Sensitive Hashing?

Can get better maps, if allowed to *depend* on the dataset!

- Non-example:
  - define $g(q)$ to be the identity of closest point to q
  - computing $g(q)$ is as hard as the problem-to-be-solved!

" I'll tell you where to find *The Origin of Species* once you recite **all** existing books

Can get better, efficient maps, if *depend* on the dataset!

Space	Time	Exponent	$c = 2$	Reference
$n^{1+\rho}$	n^{ρ}	$\rho \approx 1/c^2$	$\rho = 1/4$	[AI'06]
		$\rho \approx \frac{1}{2c^2 - 1}$	$\rho = 1/7$	[A.-Indyk-Nguyen-Razenshteyn'14, A.-Razenshteyn'15]

best LSH algorithm

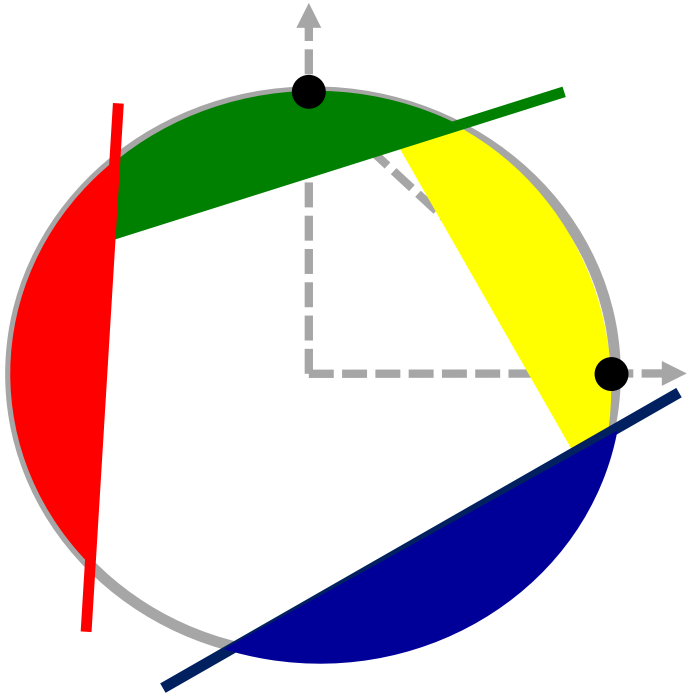
New Approach: Data-dependent LSH

[A-Razenshteyn'15]

- Two new ideas:
  - 1) a nice point configuration ← has LSH with better quality ρ
  - 2) can always reduce to such configuration ← data-dependent

1) a nice point configuration

- As if vectors chosen randomly from Gaussian distribution
- Points on a unit sphere, where
  - $cr \approx \sqrt{2}$, i.e., dissimilar pair is (near) orthogonal
  - Similar pair: $r = \sqrt{2}/c$



Map g:

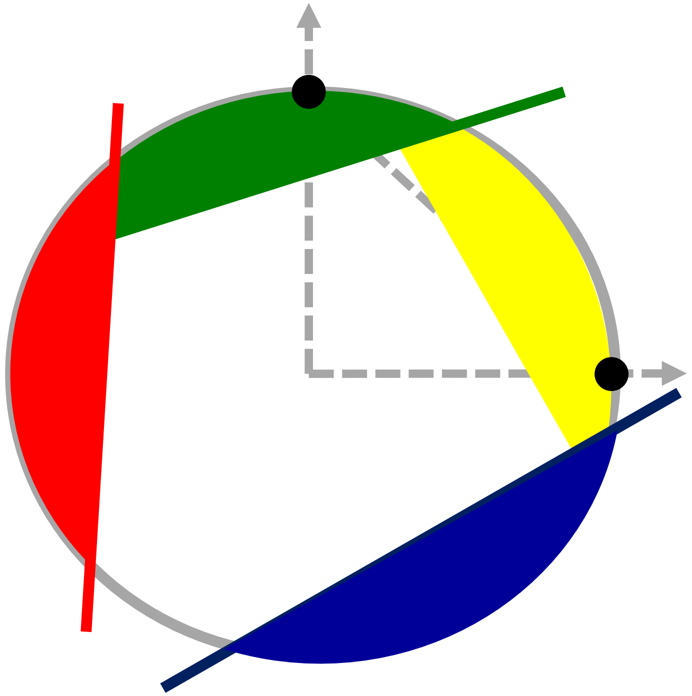
- Randomly slice out caps on sphere surface

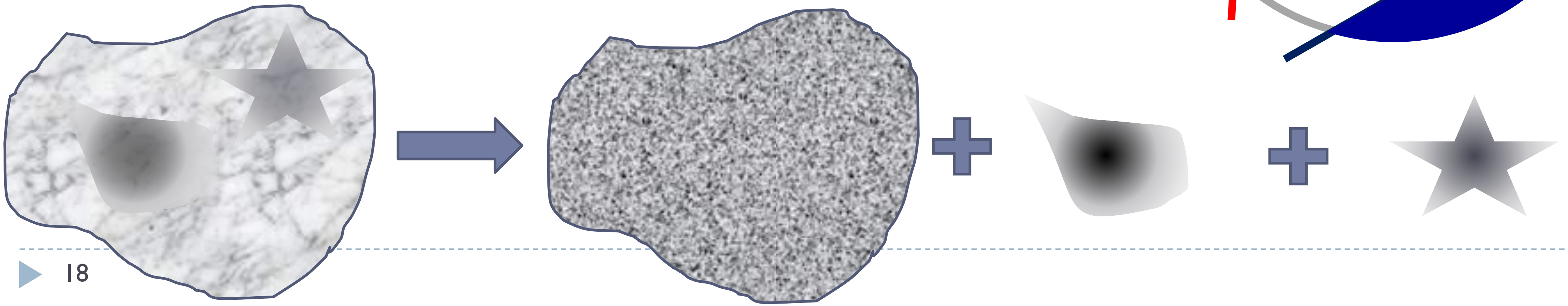
  - Like ball carving
  - Curvature helps get better quality partition

1) a nice point configuration

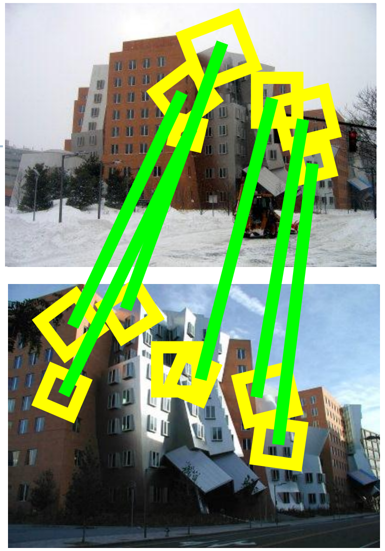
2) can always reduce to such configuration

- A worst-case to (pseudo-)random-case reduction
  - a form of "regularity lemma"
- Lemma: any pointset $P \in R^d$ can be decomposed into clusters, where one cluster is pseudo-random and the rest have smaller diameter





Beyond Euclidean space



- Data-dependent hashing:
  - Better algorithms for Hamming space
  - Also algorithms for distances where vanilla LSH does not work!
    - E.g.: distance $||x - y||_\infty = \max_{i=1..d} |x_i - y_i|$ [Indyk'98, …]

- Even more beyond?
- Approach 3: metric embeddings
  - Geometric reduction b/w different spaces
  - Rich theory in Functional Analysis

Sets of points

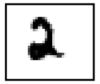
Earth-Mover Distance (Wasserstein space)

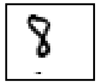
→

$\{0,1\}^d$

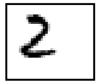
Hamming dist.

[Charikar'02, Indyk-Thaper'04, Naor-Schechtman'06, A-Indyk-Krauthgamer'08…]





Summary: Similarity Search



- objects → high-dimensional vectors
- similarity → distance b/w vectors
- Similarity Search → Nearest Neighbor Search

} Geometry

- Different applications lead to different geometries
- Connects to rich mathematical areas:
  - Space partitions and isoperimetry: what's the body with least perimeter?
  - Metric embeddings: can we map some geometries into others well?
- Only recently we (think we) understood the Euclidean metric
  - Properties of many other geometries remain unsolved!

To search or not to search