

Nearest Neighbor Search: Retrieving Documents



Emily Fox & Carlos Guestrin

Machine Learning Specialization

University of Washington

Retrieving documents of interest

Document retrieval

- Currently reading article you like



Document retrieval

- Currently reading article you like
- **Goal:** Want to find similar article



Document retrieval



Challenges

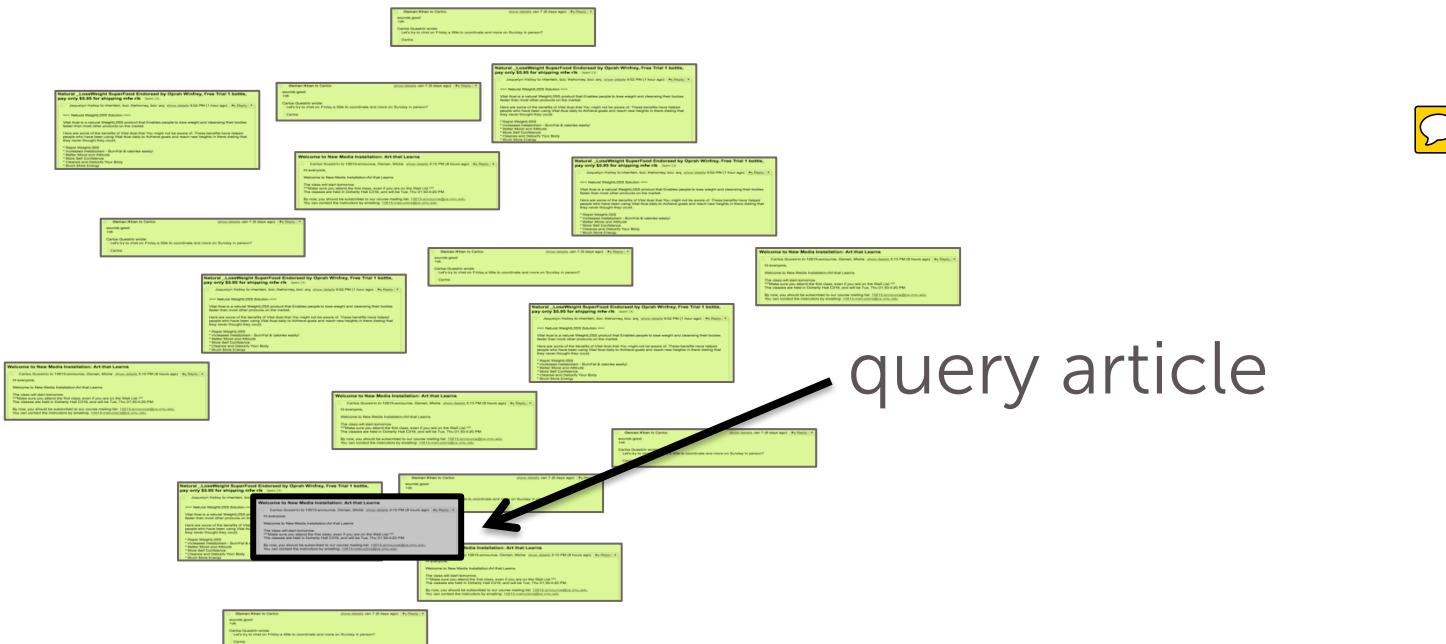
- How do we measure similarity?
- How do we search over articles?



Retrieval as k-nearest neighbor search

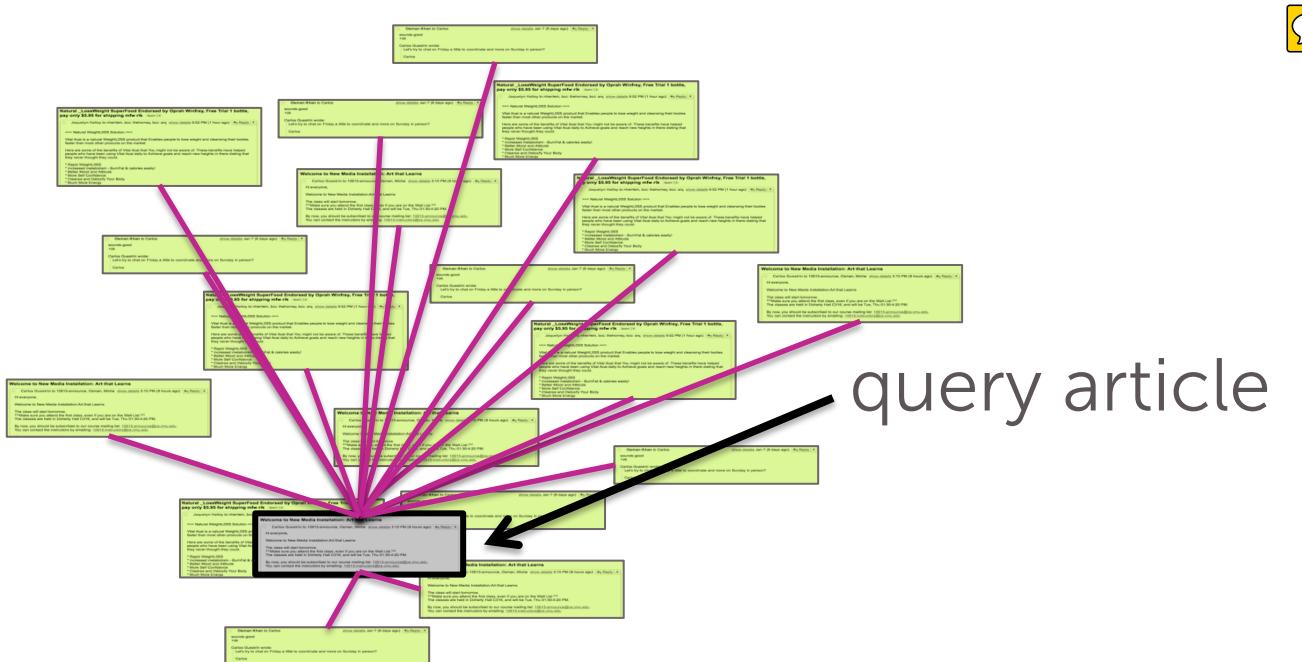
1-NN search for retrieval

Space of all articles,
organized by similarity of text



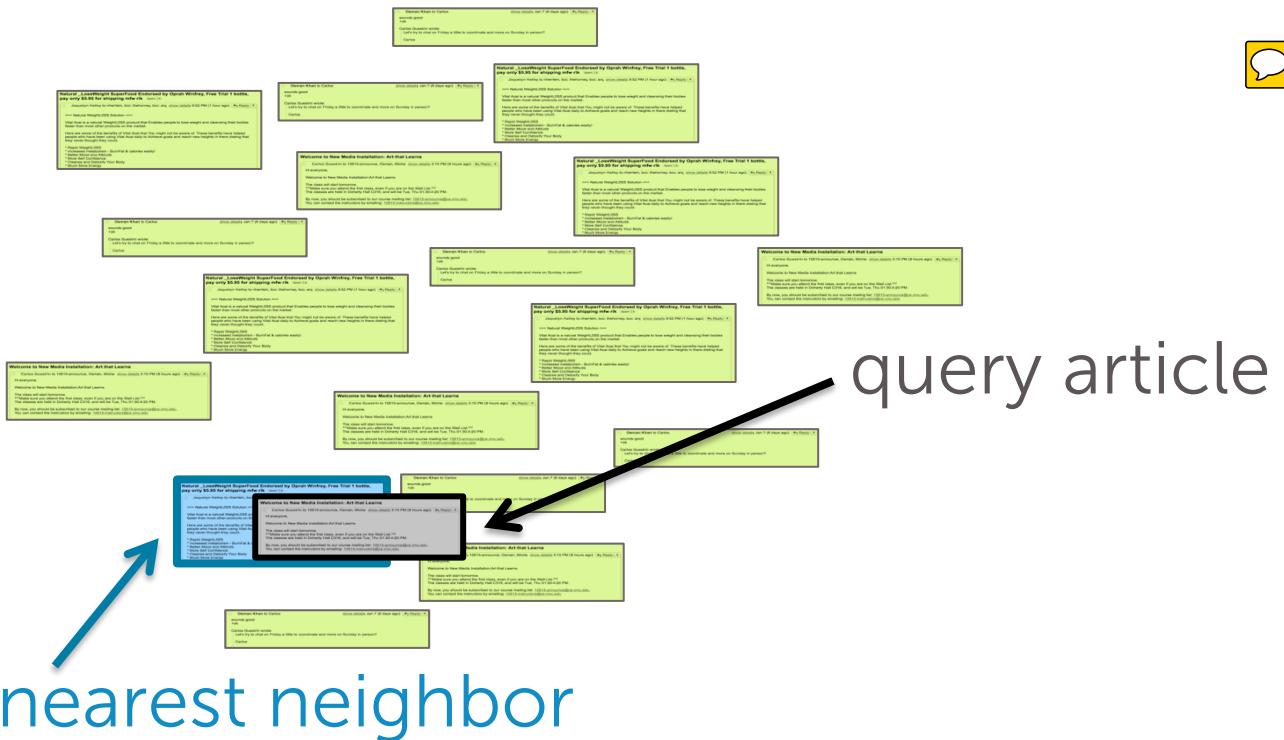
Compute distances to all docs

Space of all articles,
organized by similarity of text



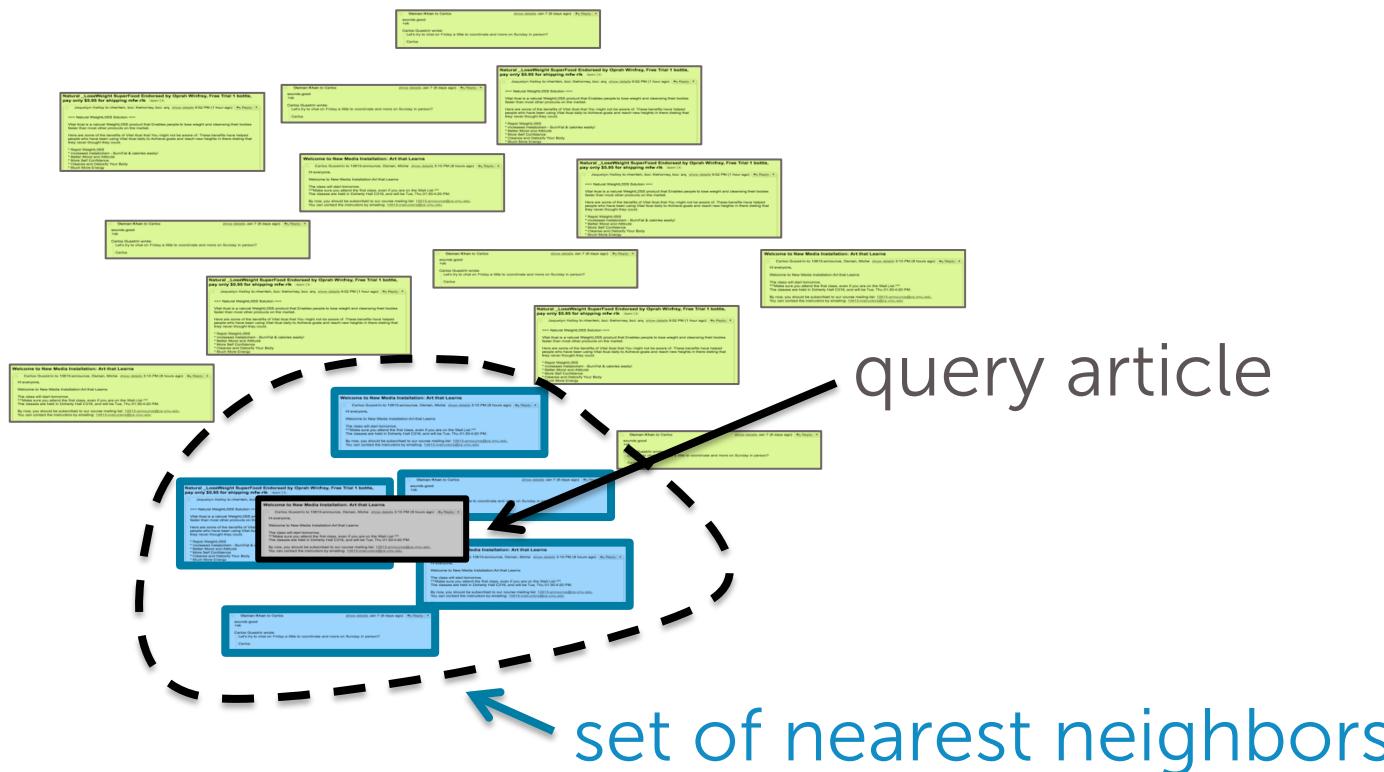
Retrieve “nearest neighbor”

Space of all articles,
organized by similarity of text



Or set of nearest neighbors

Space of all articles,
organized by similarity of text



1-NN algorithm



1 – Nearest neighbor

- **Input:** Query article  : $\underline{\mathbf{x}}_q$
Corpus of documents $(N \text{ docs})$



- **Output:** *Most* similar article  $\leftarrow \mathbf{x}^{NN}$

Formally:

$$\mathbf{x}^{NN} = \min_{\mathbf{x}_i} \text{distance}(\mathbf{x}_q, \mathbf{x}_i)$$

1-NN algorithm



Initialize **Dist2NN** = ∞ ,

For $i=1,2,\dots,N$

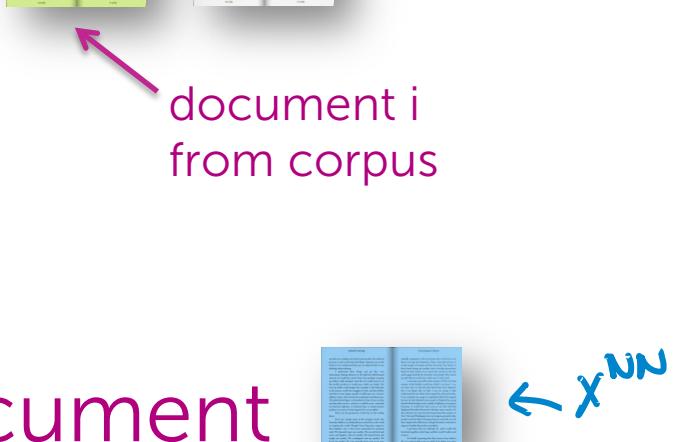
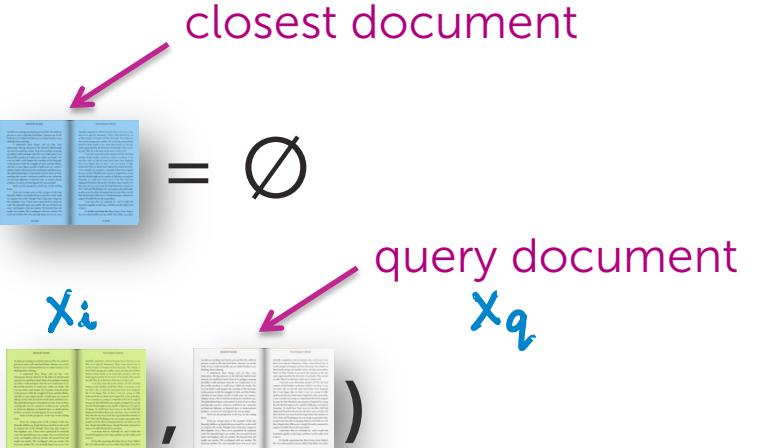
Compute: $\delta = \text{distance}(x_i, x_q)$

If $\delta < \text{Dist2NN}$

set $x_i =$

set **Dist2NN** = δ

Return most similar document



k-NN algorithm

k – Nearest neighbor

- **Input:** Query article : \mathbf{x}_q



Corpus of documents



: $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$

- **Output:** *List of k* similar articles



Formally:

$$\mathcal{X}^{NN} = \{\mathbf{x}^{NN_1}, \dots, \mathbf{x}^{NN_k}\}$$

for all \mathbf{x}_i not in \mathcal{X}^{NN} , $\text{distance}(\mathbf{x}_i, \mathbf{x}_q) \geq \max_{\mathbf{x}^{NN_j}, j=1..k} \text{distance}(\mathbf{x}^{NN_j}, \mathbf{x}_q)$



k-NN algorithm

Initialize $\text{Dist2kNN} = \text{sort}(\delta_1, \dots, \delta_k)$ ← list of sorted distances
=     

For $i=k+1, \dots, N$

Compute: $\delta = \text{distance}(\text{document}_i, \text{query doc}_q)$

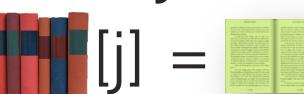
If $\delta < \text{Dist2kNN}[k]$ ← distance to k^{th} NN (furthest NN in set)

find j such that $\delta > \text{Dist2kNN}[j-1]$ but $\delta < \text{Dist2kNN}[j]$

remove furthest ~~house~~ and shift queue:



$\text{Dist2kNN}[j+1:k] = \text{Dist2kNN}[j:k-1]$

set $\text{Dist2kNN}[j] = \delta$ and   closest k docs to query doc

Return k most similar articles



Critical elements of NN search

Item (e.g., doc) **representation**

$$\mathbf{x}_q \leftarrow$$



Measure of **distance** between items:

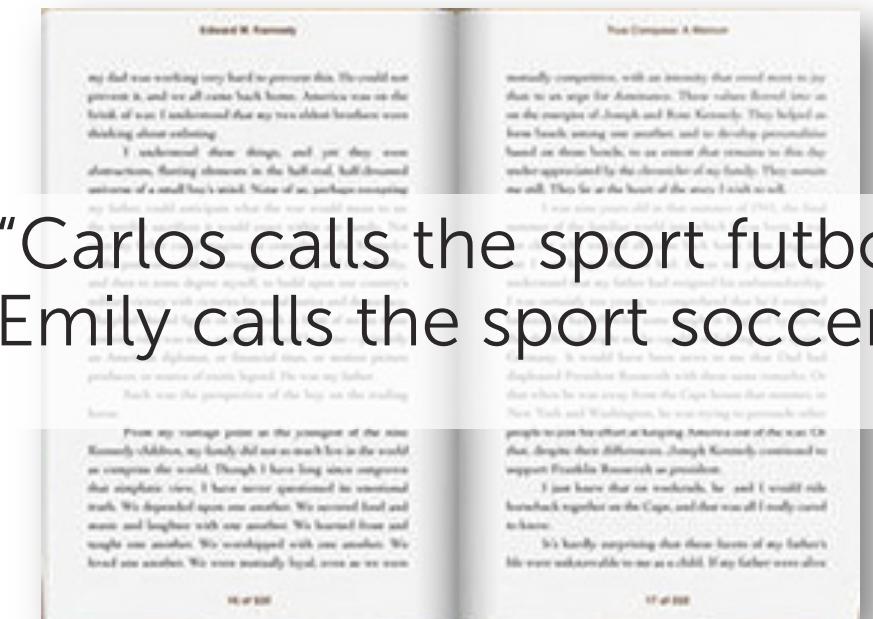
$$\delta = \text{distance}(\mathbf{x}_i, \mathbf{x}_q)$$

Document representation

Word count document representation

Bag of words model

- Ignore order of words
- Count # of instances of each word in vocabulary



"Carlos calls the sport futbol.
Emily calls the sport soccer."

Issues with word counts – Rare words



Common words in doc: “the”, “player”, “field”, “goal”

Dominate rare words like: “futbol”, “Messi”



TF-IDF document representation

Emphasizes important words

- Appears frequently in document (common locally)



- Appears rarely in corpus (rare globally)

TF-IDF document representation

Emphasizes **important words**

- Appears frequently in document (**common locally**)

Term frequency =  word counts

- Appears rarely in corpus (**rare globally**)



TF-IDF document representation

Emphasizes **important words**

- Appears frequently in document (**common locally**)

Term frequency =  word counts

- Appears rarely in corpus (**rare globally**)

Inverse doc freq. = $\log \frac{\# \text{ docs}}{1 + \# \text{ docs using word}}$



TF-IDF document representation

Emphasizes **important words**

- Appears frequently in document (**common locally**)

Term frequency = 

- Appears rarely in corpus (**rare globally**)

Inverse doc freq. = $\log \frac{\# \text{ docs}}{1 + \# \text{ docs using word}}$



Trade off: **local frequency vs. global rarity**

$tf * idf$

Distance metrics

Distance metrics: Defining notion of “closest”

In 1D, just Euclidean distance:

$$\text{distance}(x_i, x_q) = |x_i - x_q|$$

In multiple dimensions:

- can define many interesting distance functions
- most straightforwardly, might want to weight different dimensions differently



Weighting different features

Reasons:

- Some features are more relevant than others



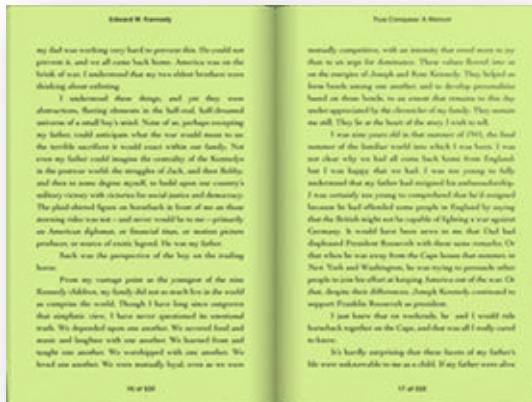
bedrooms
bathrooms
sq.ft. living
sq.ft. lot
floors
year built
year renovated
waterfront



Weighting different features

Reasons:

- Some features are more relevant than others



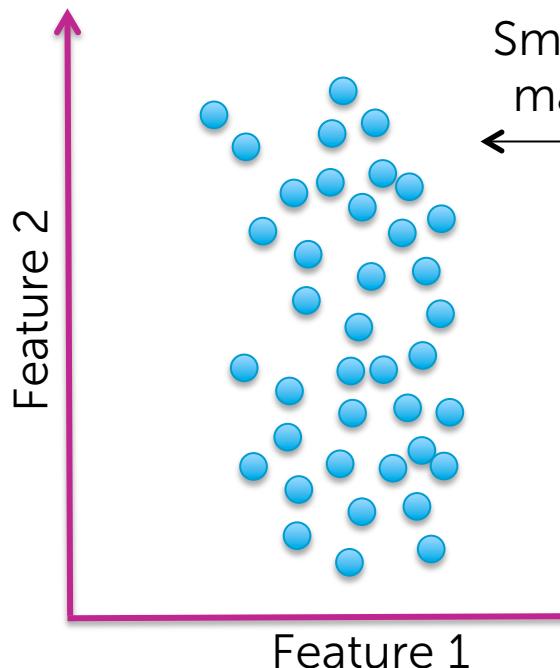
title
abstract
main body
conclusion



Weighting different features

Reasons:

- Some features are more relevant than others
- Some features vary more than others



Small changes matter more
Big changes matter less

Specify weights as a function of feature spread

For feature j :

$$\frac{1}{\max_i(\mathbf{x}_i[j]) - \min_i(\mathbf{x}_i[j])}$$



Scaled Euclidean distance

Formally, this is achieved via

$$\text{distance}(\mathbf{x}_i, \mathbf{x}_q) = \sqrt{a_1(\mathbf{x}_i[1]-\mathbf{x}_q[1])^2 + \dots + a_d(\mathbf{x}_i[d]-\mathbf{x}_q[d])^2}$$



weight on each feature
(defining relative importance)



Effect of binary weights

$$\text{distance}(\mathbf{x}_i, \mathbf{x}_q) = \sqrt{a_1(\mathbf{x}_i[1]-\mathbf{x}_q[1])^2 + \dots + a_d(\mathbf{x}_i[d]-\mathbf{x}_q[d])^2}$$

Setting weights as 0 or 1
is equivalent to
feature selection

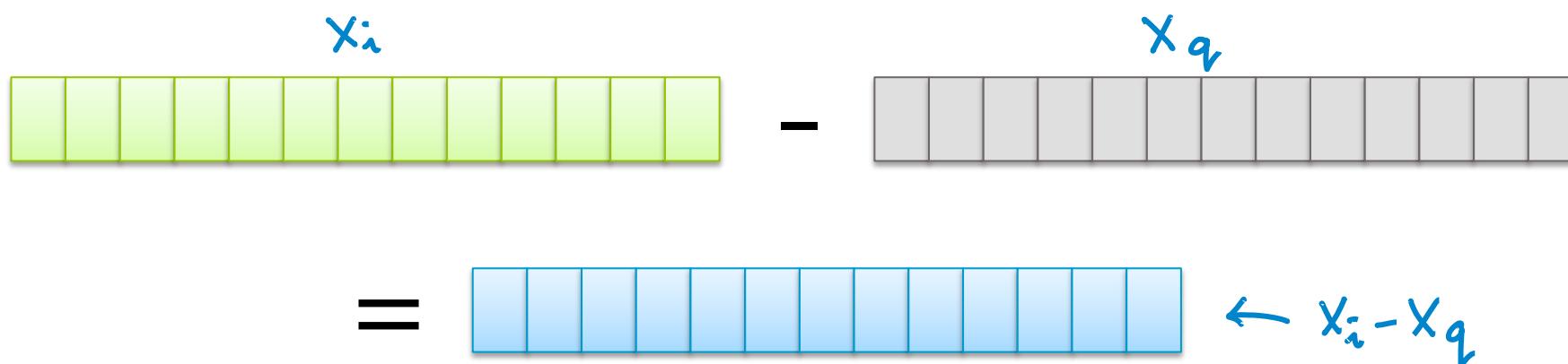
Feature engineering/
selection is
important, but hard



(non-scaled) Euclidean distance

Defined in terms of inner product

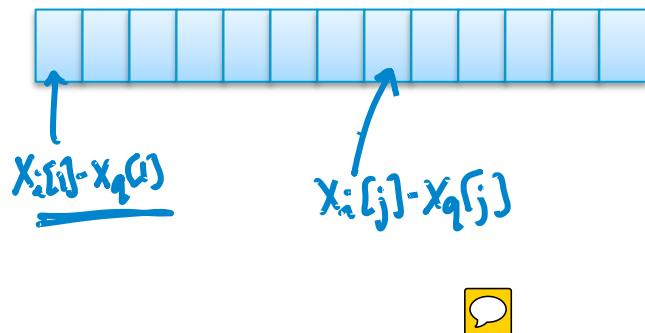
$$\text{distance}(\mathbf{x}_i, \mathbf{x}_q) = \sqrt{(\mathbf{x}_i - \mathbf{x}_q)^T (\mathbf{x}_i - \mathbf{x}_q)}$$
$$= \sqrt{(\mathbf{x}_i[1] - \mathbf{x}_q[1])^2 + \dots + (\mathbf{x}_i[d] - \mathbf{x}_q[d])^2}$$

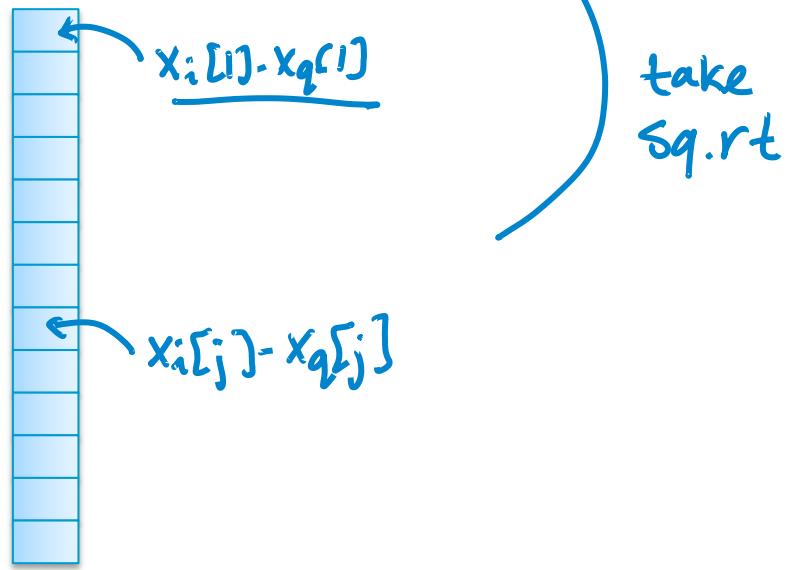


(non-scaled) Euclidean distance

Defined in terms of inner product

$$\text{distance}(\mathbf{x}_i, \mathbf{x}_q) = \sqrt{(\mathbf{x}_i - \mathbf{x}_q)^T (\mathbf{x}_i - \mathbf{x}_q)}$$
$$= \sqrt{(x_i[1] - x_q[1])^2 + \dots + (x_i[d] - x_q[d])^2}$$

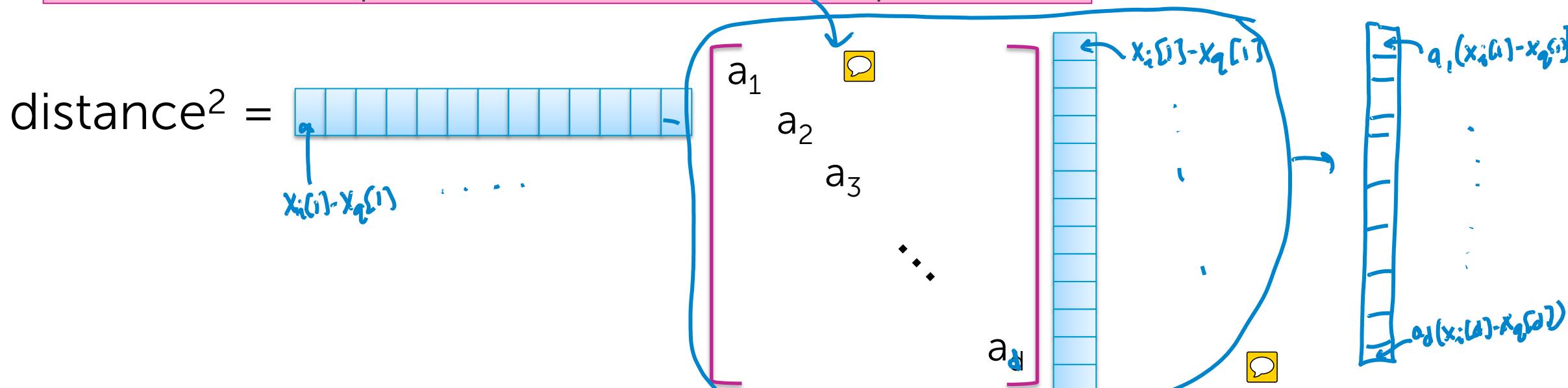
$$\text{distance}^2 =$$

$$\underline{x_i[1]-x_q[1]}$$
$$\underline{x_i[2]-x_q[2]}$$



Scaled Euclidean distance

Defined in terms of inner product

$$\text{distance}(\mathbf{x}_i, \mathbf{x}_q) = \sqrt{(\mathbf{x}_i - \mathbf{x}_q)^T \mathbf{A} (\mathbf{x}_i - \mathbf{x}_q)}$$
$$= \sqrt{a_1(x_i[1] - x_q[1])^2 + \dots + a_d(x_i[d] - x_q[d])^2}$$



Another natural inner product measure

 x_q  x_i

Similarity

$$= \mathbf{x}_i^T \mathbf{x}_q$$

$$= \sum_{j=1}^d \mathbf{x}_i[j] \mathbf{x}_q[j]$$



$$= 13$$



Another natural inner product measure

Edward W. Kennedy
my dad was working very hard to prevent that. He could not
go home, and we all came back home. America was on the
hook for us. We had to go to school, and we had to go to work.
Shaking their shoulders.

From my range point of view the things, and yet they were
distracting. Playing chess in the hall and half-dressed
and the like. I have never questioned what the war would move to
my father could imagine what the war would move to
the terrible sacrifice it would cost us one family. But
we are the ones who are the ones who are the ones who are
in the present world the strength of each, and then family,
military victory with elements for social justice and democracy.

The placed dead figure on Standard in front of us at three
memorials. We have to go to school, and we have to go to work
as American diplomats, or financial firms, or marine products
producing companies. We have to go to school, and we are now
back to the perspective of the boy on the reading
books.

From my range point of view the perspective of the nine
Family children, we did not do much for the world
in comparison. This is the reason why I have never
questioned what the war would move to
our people's views. I have never questioned in contrast
with the like. Playing chess in the hall and half-dressed
and the like. We learned from and
taught one another. We worked with one another. We
had one another. We were really good at what we are now

10 of 100
11 of 100



Similarity
= 0



Edward W. Kennedy
my dad was working very hard to prevent this. He could not
go home, and we all came back home. America was on the
hook for us. We had to go to school, and we had to go to work.
Shaking their shoulders.

From my range point of view the things, and yet they were
distracting. Playing chess in the hall and half-dressed
and the like. I have never questioned what the war would move to
my father could imagine what the war would move to
the terrible sacrifice it would cost us one family. But
we are the ones who are the ones who are the ones who are
in the present world the strength of each, and then family,
military victory with elements for social justice and democracy.

The placed dead figure on Standard in front of us at three
memorials. We have to go to school, and we have to go to work
as American diplomats, or financial firms, or marine products
producing companies. We have to go to school, and we are now
back to the perspective of the boy on the reading
books.

From my range point of view the perspective of the nine
Family children, we did not do much for the world
in comparison. This is the reason why I have never
questioned what the war would move to
our people's views. I have never questioned in contrast
with the like. Playing chess in the hall and half-dressed
and the like. We learned from and
taught one another. We worked with one another. We
had one another. We were really good at what we are now

10 of 100
11 of 100



Cosine similarity – normalize

Similarity =

$$\frac{\sum_{j=1}^d \mathbf{x}_i[j] \mathbf{x}_q[j]}{\sqrt{\sum_{j=1}^d (\mathbf{x}_i[j])^2} \sqrt{\sum_{j=1}^d (\mathbf{x}_q[j])^2}}$$

$$\mathbf{x}_i^\top \mathbf{x}_q = \cos(\theta)$$

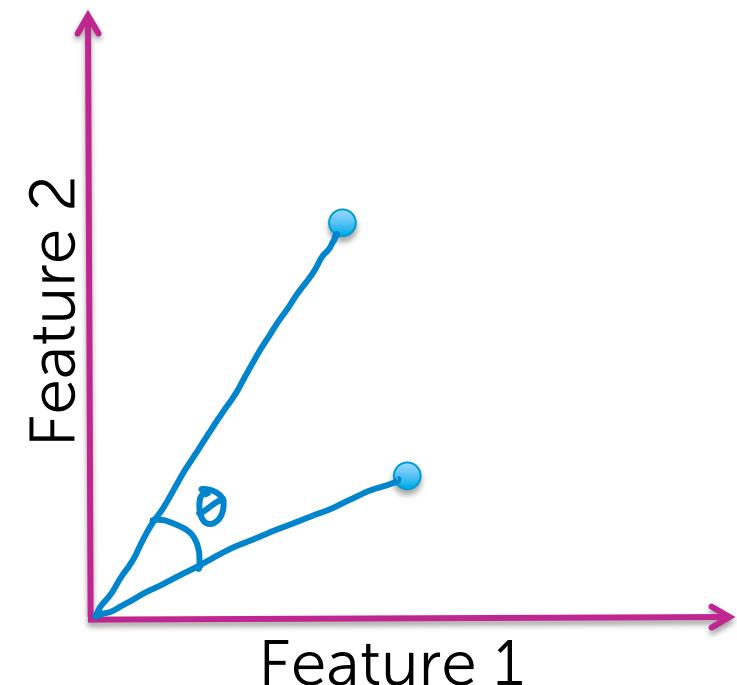
$$\|\mathbf{x}_i\| \|\mathbf{x}_q\|$$

$$= \left(\frac{\mathbf{x}_i}{\|\mathbf{x}_i\|} \right)^\top \left(\frac{\mathbf{x}_q}{\|\mathbf{x}_q\|} \right)$$

first normalize

- Not a proper distance metric
- Efficient to compute for sparse vecs

$$\mathbf{a}^\top \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos(\theta)$$



Normalize



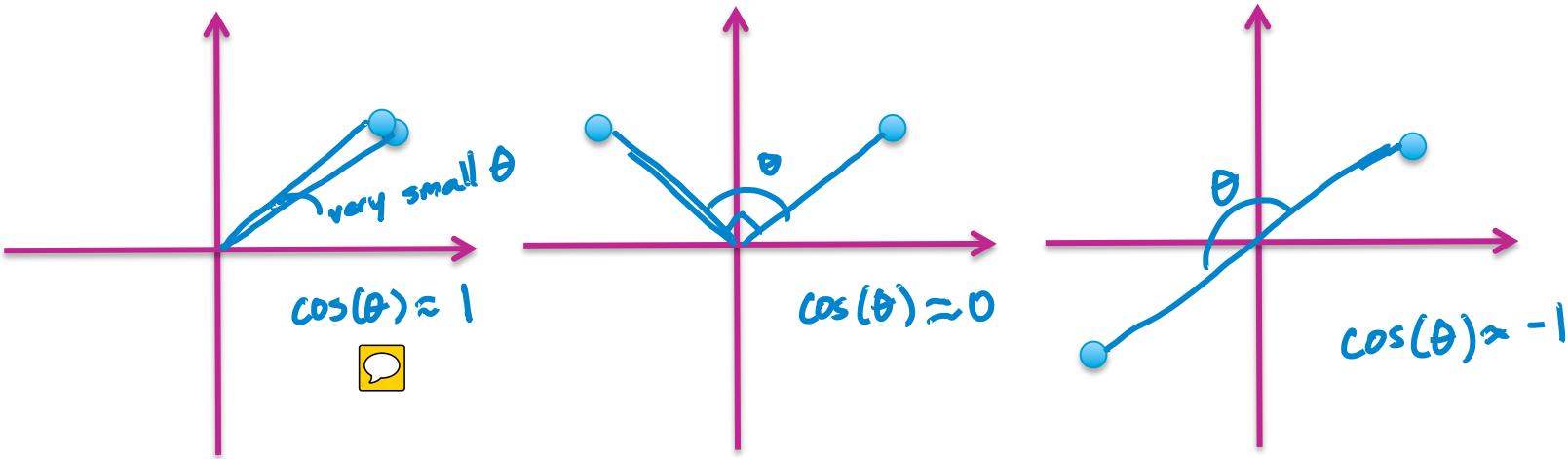
1	0	0	0	5	3	0	0	1	0	0	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---

$$\sqrt{1^2 + 5^2 + 3^2 + 1^2} \leftarrow \|x_i\| = \sum_{j=1}^d x_i[j]^2$$

$\leftarrow x_i$

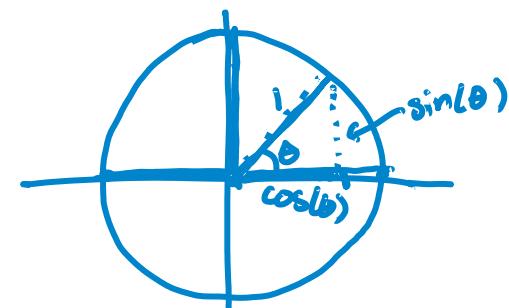
1					5	3		1				
/	0	0	0	/	/	0	0	/	0	0	0	0
6				6	6		6					

Cosine similarity



In general, $-1 < \text{similarity} < 1$

For positive features (like tf-idf)
 $0 < \text{similarity} < 1$



Define **distance = 1-similarity**

To normalize or not?

Edward W. Kennedy	New Company & Mission	Edward W. Kennedy	New Company & Mission
<p>my dad was working very hard to prevent this. He could not prevent it, and we all came back home. America was the country he wanted us to live in. He had a long time coming to terms with the fact that we were going to stay here.</p> <p>—Then they all got into the car, half-fallen, silent. I think a small boy's voice said, "Well, perhaps carrying the whole world on your shoulders is too much for you." The words never settled in until we sat outside our family. Not until we sat outside our family did we realize that we were in the greatest world the strength of, and that, and their Bibles, and that in some degree accepted, had split open our country's culture.</p> <p>The plot about figures on Newell is based on one of our most important figures in our country, a man who was an American diplomat, an American friend, or someone's personal preference of name. I guess the man was a good soldier, from what I can tell.</p> <p>From my memory, part of the purpose of the story Kennedy himself, my family did as much as we could in the campaign the world over. He had a long time coming to terms with the fact that we were going to stay here.</p> <p>We depended on one another. We worked hard and we worked together. We worked together with one another. We lived one another. We worked mostly hard, even as we were</p>	<p>mentally competitive, with intensity that could bring us to our knees if we tried to do something. But we did our best to keep our heads held up, and to develop personal relationships with those around us, so that we could be the best at the job, and the best at our job of finding friends.</p> <p>They say, "Be at the heart of the society it is to help." I think that's what my dad was. He was the leader of the family, the leader of his wife, and he was the leader of the neighborhood. He was the leader of the world, and he was the leader of the world's children. He would want us to work outside our family. Not until we sat outside our family did we realize that we were in the greatest world the strength of, and that, and their Bibles, and that in some degree accepted, had split open our country's culture.</p> <p>It would have been nice to see that Dad had developed French-English with some measure. Or, if he had developed French-English with some measure. In New York and Washington, he was trying to promote other people's interests, and he was trying to promote our own.</p> <p>He was trying to promote other people's interests, and he was trying to promote our own. And despite these differences, though Kennedy remained committed to our country.</p> <p>I just have to say, my dad and I could ride together across the Cape, and we had it all ready to go.</p> <p>In his simplicity, that these four of my father's life are inseparable to me. If it's true.</p>	<p>that my dad was trying very hard to prevent this. He could not prevent it, and we all came back home. America was the country he wanted us to live in. He had a long time coming to terms with the fact that we were going to stay here.</p> <p>—Then they all got into the car, half-fallen, silent. I think a small boy's voice said, "Well, perhaps carrying the whole world on your shoulders is too much for you." The words never settled in until we sat outside our family. Not until we sat outside our family did we realize that we were in the greatest world the strength of, and that, and their Bibles, and that in some degree accepted, had split open our country's culture.</p> <p>The plot about figures on Newell is based on one of our most important figures in our country, a man who was an American diplomat, an American friend, or someone's personal preference of name. I guess the man was a good soldier, from what I can tell.</p> <p>From my memory, part of the purpose of the story Kennedy himself, my family did as much as we could in the campaign the world over. He had a long time coming to terms with the fact that we were going to stay here.</p> <p>We depended on one another. We worked hard and we worked together. We worked together with one another. We lived one another. We worked mostly hard, even as we were</p>	<p>mentally competitive, with intensity that could bring us to our knees if we tried to do something. But we did our best to keep our heads held up, and to develop personal relationships with those around us, so that we could be the best at the job, and the best at our job of finding friends.</p> <p>They say, "Be at the heart of the society it is to help." I think that's what my dad was. He was the leader of the family, the leader of his wife, and he was the leader of the neighborhood. He was the leader of the world, and he was the leader of the world's children. He would want us to work outside our family. Not until we sat outside our family did we realize that we were in the greatest world the strength of, and that, and their Bibles, and that in some degree accepted, had split open our country's culture.</p> <p>It would have been nice to see that Dad had developed French-English with some measure. Or, if he had developed French-English with some measure. In New York and Washington, he was trying to promote other people's interests, and he was trying to promote our own.</p> <p>He was trying to promote other people's interests, and he was trying to promote our own. And despite these differences, though Kennedy remained committed to our country.</p> <p>I just have to say, my dad and I could ride together across the Cape, and we had it all ready to go.</p> <p>In his simplicity, that these four of my father's life are inseparable to me. If it's true.</p>
10 of 100	17 of 100	18 of 100	17 of 100



3	1	0	0	2	0	0	1	0	1	0	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---

Similarity = 13

6 2 0 0 4 0 0 2 0 2 0 0 0
Similarity = 52

Normalize



1	0	0	0	5	3	0	0	1	0	0	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---

$$\sqrt{1^2 + 5^2 + 3^2 + 1^2}$$

1					5	3		1				
/	0	0	0	/	/	0	0	/	0	0	0	0
6				6	6		6					

In the normalized case

my dad was working long hours to help procure this. He could not get away from his work, so we had to go to him. I think he was at the end of it because I had my two older brothers come along thinking about it.

"After all those things, and yet you were sharing, floating in the half-odd, half-dream state, you still had the desire to do what you wanted to do. I think my dad could anticipate what we would want to do. The world outside would be round about our family. But there was a desire to do what we wanted to do. And I think that in the present world the strength of love, and then family, and then the desire to do what you wanted to do, will always carry us through our personal and family existence."

The solidified ideal of freedom is based on one of the deepest human needs. It is the need to be free, to be independent, to be able to make one's own choices, to be American, democratic, or, most picture perfect, to be a cowboy.

Such was the perspective of the boy in the reading room.

From my vantage point as a pragmatist the sense of family, children, family did not seem to be in the world of the cowboy. The cowboy was a loner, a man who did not conform, who, I have never seen, conformed to anything. He was the fighter against the world. He fought on and fought on alone. We worshipped him and one another. We lived one another. We were immensely loyal, even in the most morally compromised, with an intensity that must have been in the blood of the sons of the cowboys of the West.

In the absence of a Joseph and Benjamin Kibbe they developed their own bushwhacking, one another, and to develop the same kind of loyalty, the same kind of love, the same kind of respect for the other person, for the family, for the community.

It was also true in the instance of the boy, the final member of the household model that it was family that was important. He was a good boy, he was a good son, he was a good brother, he was a good husband, he was a good father, he was a good son-in-law, he was a good grandfather. He was a good neighbor, he was a good citizen with integrity, with character, and with a strong sense of family and community.

The solidified ideal of freedom is based on one of the deepest human needs. It is the need to be free, to be independent, to be able to make one's own choices, to be American, democratic, or, most picture perfect, to be a cowboy.

1				5	3			1					
/	0	0	0	/	/	0	0	/	0	0	0	0	0
6				6	6			6					

3	1	0	0	1	/	0	0	1	/	0	1	0	0	0
/	/	0	0	/	0	0	/	4	0	/	4	0	0	0
4	4			2				4			4			

Similarity
= 13/24



Edward M. Kennedy	The Congressman's Memoir	Edward M. Kennedy	The Congressman's Memoir
my dad was working very hard to pay his rent. This could not possibly have been the best time for him to be in the world, but I am grateful that my mother and my dad were able to keep their spirits up.	And we did some new things, and kept some old ones. We were dancing, dancing the foxtrot in the ballroom, and I learned how to do the tango. I learned how to do the waltz. My father could anticipate what the world would move to as the terrible war would end over in our family. You never know what you're going to do when you're dancing, but it's the process while the strength of it, and that beauty, and that grace, and that elegance, and that poise, and that military victory with sacrifice for the peace and democracy.	I saw one day across the floor of the ballroom, the head of the national brotherhood who I had known since I was born. He was a tall man with a white mustache, and he was wearing a suit. I was happy that I was young and strong, and I wanted to go up to him and say hello.	My dad was working very hard to pay his rent. This could not possibly have been the best time for him to be in the world, but I am grateful that my mother and my dad were able to keep their spirits up as they did some new things, and as the terrible war would end over in our family. You never know what you're going to do when you're dancing, but it's the process while the strength of it, and that beauty, and that grace, and that elegance, and that poise, and that military victory with sacrifice for the peace and democracy.
The plated silver plate is beautiful. It is made of all these things, and it is a symbol of the strength of the United States, of American diplomats in financial times, or winter picnics, or picnics in the fall.	Such was the prosperity of the day on the trading floor.	From my vantage point as the younger of the two Kennedy children, my family had not yet been in the world as long as the older Kennedy children. But we were there, and that means, I have never appreciated so much the strength of the family bond between us. We had one foot and one hand, and one mother. We worked with one another. We had one foot and one mother. We worked with one another. We had one foot and one hand, and one mother.	My dad was working very hard to pay his rent. This could not possibly have been the best time for him to be in the world, but I am grateful that my mother and my dad were able to keep their spirits up as they did some new things, and as the terrible war would end over in our family. You never know what you're going to do when you're dancing, but it's the process while the strength of it, and that beauty, and that grace, and that elegance, and that poise, and that military victory with sacrifice for the peace and democracy.
I am grateful that my mother and my dad were able to keep their spirits up as they did some new things, and as the terrible war would end over in our family. You never know what you're going to do when you're dancing, but it's the process while the strength of it, and that beauty, and that grace, and that elegance, and that poise, and that military victory with sacrifice for the peace and democracy.	I am grateful that my mother and my dad were able to keep their spirits up as they did some new things, and as the terrible war would end over in our family. You never know what you're going to do when you're dancing, but it's the process while the strength of it, and that beauty, and that grace, and that elegance, and that poise, and that military victory with sacrifice for the peace and democracy.	I am grateful that my mother and my dad were able to keep their spirits up as they did some new things, and as the terrible war would end over in our family. You never know what you're going to do when you're dancing, but it's the process while the strength of it, and that beauty, and that grace, and that elegance, and that poise, and that military victory with sacrifice for the peace and democracy.	I am grateful that my mother and my dad were able to keep their spirits up as they did some new things, and as the terrible war would end over in our family. You never know what you're going to do when you're dancing, but it's the process while the strength of it, and that beauty, and that grace, and that elegance, and that poise, and that military victory with sacrifice for the peace and democracy.
Such was the prosperity of the day on the trading floor.	From my vantage point as the younger of the two Kennedy children, my family had not yet been in the world as long as the older Kennedy children. But we were there, and that means, I have never appreciated so much the strength of the family bond between us. We had one foot and one hand, and one mother. We worked with one another. We had one foot and one hand, and one mother.	Such was the prosperity of the day on the trading floor.	From my vantage point as the younger of the two Kennedy children, my family had not yet been in the world as long as the older Kennedy children. But we were there, and that means, I have never appreciated so much the strength of the family bond between us. We had one foot and one hand, and one mother. We worked with one another. We had one foot and one hand, and one mother.
The plated silver plate is beautiful. It is made of all these things, and it is a symbol of the strength of the United States, of American diplomats in financial times, or winter picnics, or picnics in the fall.	Such was the prosperity of the day on the trading floor.	From my vantage point as the younger of the two Kennedy children, my family had not yet been in the world as long as the older Kennedy children. But we were there, and that means, I have never appreciated so much the strength of the family bond between us. We had one foot and one hand, and one mother. We worked with one another. We had one foot and one hand, and one mother.	Such was the prosperity of the day on the trading floor.
Such was the prosperity of the day on the trading floor.	From my vantage point as the younger of the two Kennedy children, my family had not yet been in the world as long as the older Kennedy children. But we were there, and that means, I have never appreciated so much the strength of the family bond between us. We had one foot and one hand, and one mother. We worked with one another. We had one foot and one hand, and one mother.	Such was the prosperity of the day on the trading floor.	From my vantage point as the younger of the two Kennedy children, my family had not yet been in the world as long as the older Kennedy children. But we were there, and that means, I have never appreciated so much the strength of the family bond between us. We had one foot and one hand, and one mother. We worked with one another. We had one foot and one hand, and one mother.

1				5	3			1					
/	0	0	0	/	/	0	0	/	0	0	0	0	0
6				6	6			6					

3	1			1			1		1				
/	/	0	0	/	0	0	/	0	/	0	0	0	0
4	4			2			4		4				

Similarity
= 13/24

But not always desired...

Edward W. Kennedy
New Compromise A Memoir

my dad was working very hard to prove this. He could not prove it, so we all came back home. America was on the hook of us. I understood that my two older brothers were thinking about returning.

I understood these things, and yet they were dissimilar. Having choices in the half-cod, half-dressed manner of a child boy is a child. Not so for Kennedy. His brothers wanted to return to America. He wanted to stay in the world he knew. It would exist with one family. You can't do that with two families. You can't do that with the past. You can't do that with the struggle of it, and then being and doing some things required to hold onto one's culture's values.

The plain-faced figure on household in front of us on these pages is Edward W. Kennedy. He is the son of the man in the picture. He is the son of the man in the portrait. He is the son of the man in the photograph. He is the son of the man in the memoir. He is the son of the man in the document. We depended upon one another. We depended upon each other. We worked with one another. We lived our families. We were morally loyal, even as we were.

From my vantage point as the younger of the two Kennedy children, my family did not seem to be in the world we were in. That was not true. Those who were in the same room, we were in. They were separated from us. We depended upon one another. We depended upon each other. We worked with one another. We lived our families. We were morally loyal, even as we were.

10 of 600 10 of 600

long document

Edward W. Kennedy
New Compromise A Memoir

my dad was working very hard to prove this. He could not prove it, so we all came back home. America was on the hook of us. I understood that my two older brothers were thinking about returning.

I understood these things, and yet they were dissimilar. Having choices in the half-cod, half-dressed manner of a child boy is a child. Not so for Kennedy. His brothers wanted to return to America. He wanted to stay in the world he knew. It would exist with one family. You can't do that with two families. You can't do that with the past. You can't do that with the struggle of it, and then being and doing some things required to hold onto one's culture's values.

The plain-faced figure on household in front of us on these pages is Edward W. Kennedy. He is the son of the man in the picture. He is the son of the man in the portrait. He is the son of the man in the photograph. He is the son of the man in the memoir. He is the son of the man in the document. We depended upon one another. We depended upon each other. We worked with one another. We lived our families. We were morally loyal, even as we were.

From my vantage point as the younger of the two Kennedy children, my family did not seem to be in the world we were in. That was not true. Those who were in the same room, we were in. They were separated from us. We depended upon one another. We depended upon each other. We worked with one another. We lived our families. We were morally loyal, even as we were.

10 of 600 10 of 600

long document

short tweet

Edward W. Kennedy
New Compromise A Memoir

myself competitive, with an intensity that could never be ignored. There was no place for me in the world. Those who were in the same room, we were in. They were separated from us. We depended upon one another. We worked with one another. We lived our families. We were morally loyal, even as we were.

10 of 600 10 of 600

Normalizing can
make dissimilar
objects appear
more similar

Common
compromise:

Just cap maximum
word counts

long document

Other distance metrics

- Mahalanobis
- rank-based
- correlation-based
- Manhattan
- Jaccard
- Hamming
- ...



Combining distance metrics

Example of document features:

1. Text of document
 - Distance metric: Cosine similarity
2. # of reads of doc
 - Distance metric: Euclidean distance



Add together with user-specified weights



Scaling up k-NN search by storing data in a KD-tree

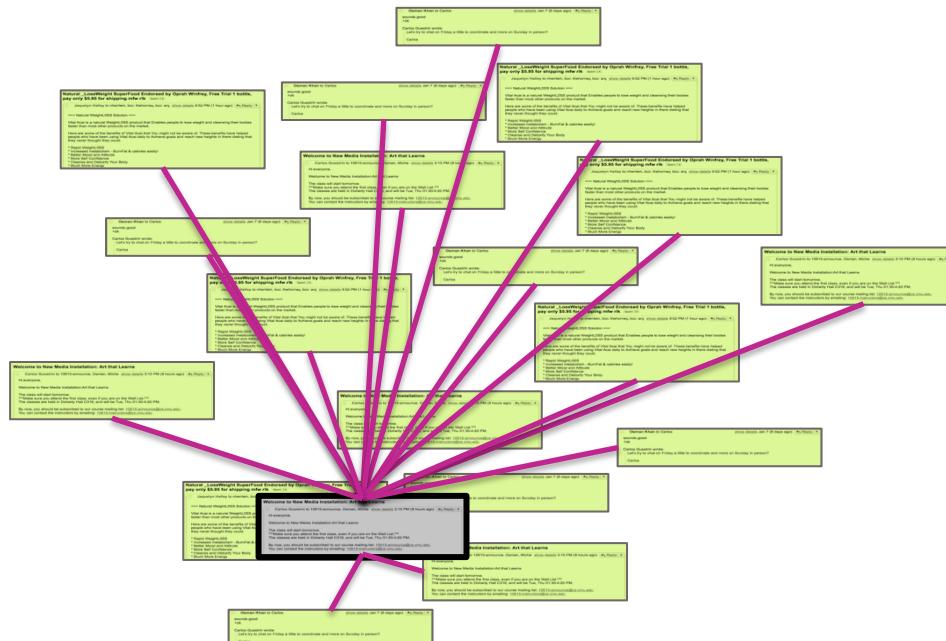
Complexity of search



Complexity of brute-force search

Given a query point, scan through each point

- $O(N)$ distance computations per 1-NN query!
- $O(N \log k)$ per k -NN query!



What if N is huge???
(and many queries)



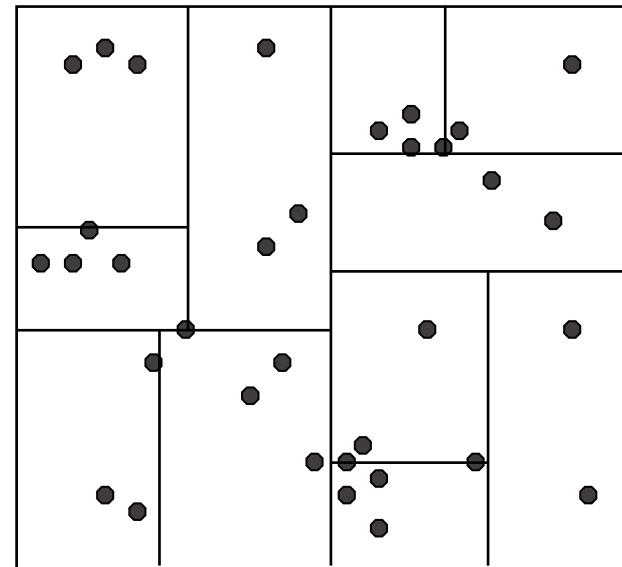
KD-tree representation

KD-trees

Structured organization of documents

- Recursively partitions points into axis aligned boxes.

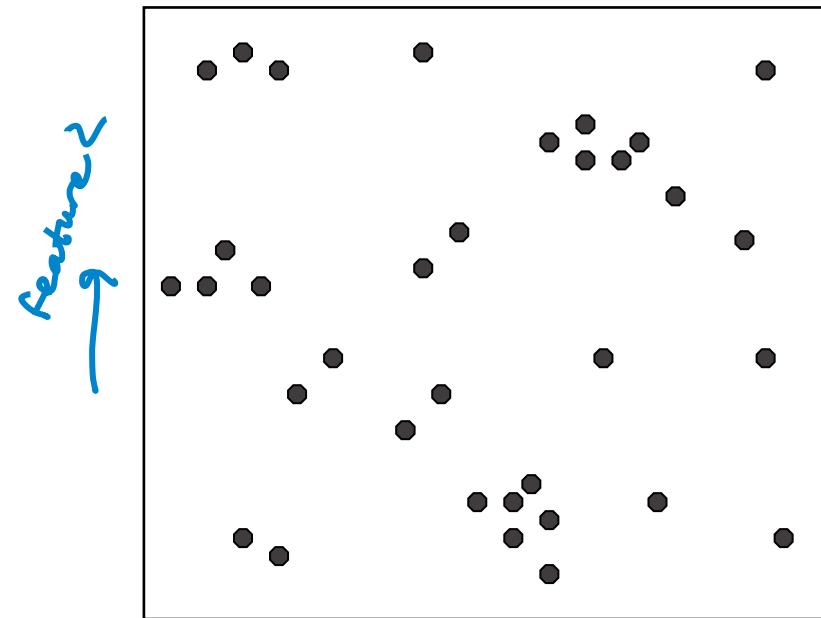
Enables more efficient pruning of search space



Works “well” in “low-medium” dimensions

- We'll get back to this...

KD-tree construction



feature 1

Start with a list of
d-dimensional points.

Pt	x[1]	x[2]
1	0.00	0.00
2	1.00	4.31
3	0.13	2.85
...

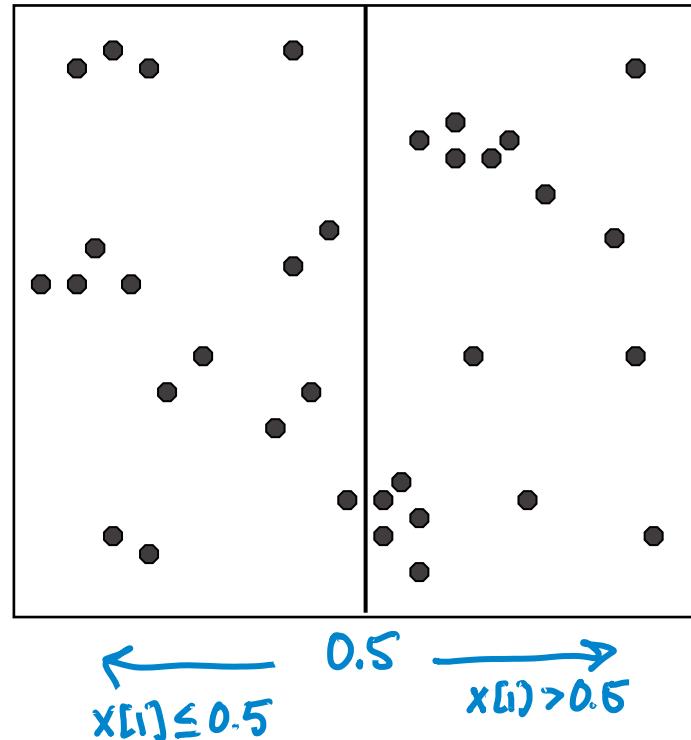


↑
obs.
indices

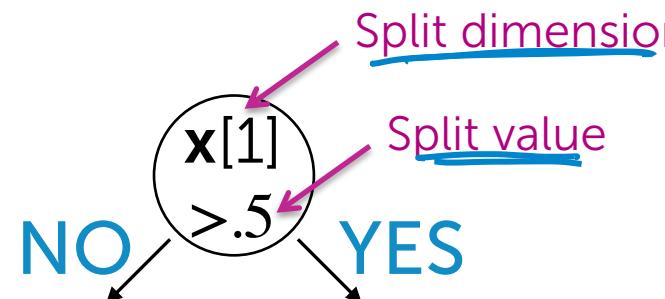
↑
feat. 1
(word 1)

↑
Feat. 2
(word 2)

KD-tree construction

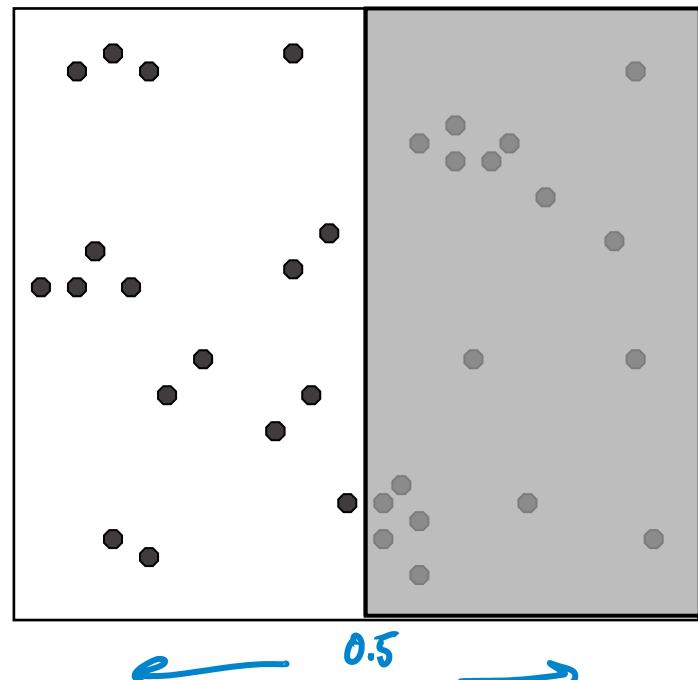


Split points into 2 groups

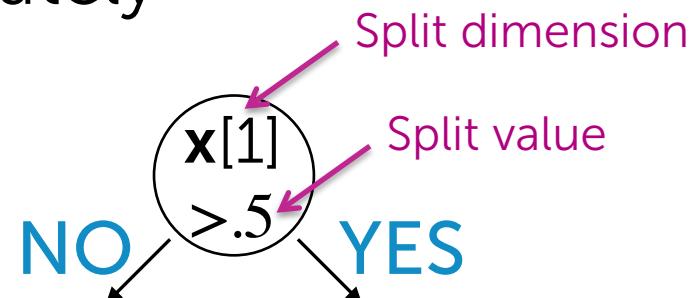


Pt	$x[1]$	$x[2]$	Pt	$x[1]$	$x[2]$
1	0.00	0.00	2	1.00	4.31
3	0.13	2.85
...			

KD-tree construction

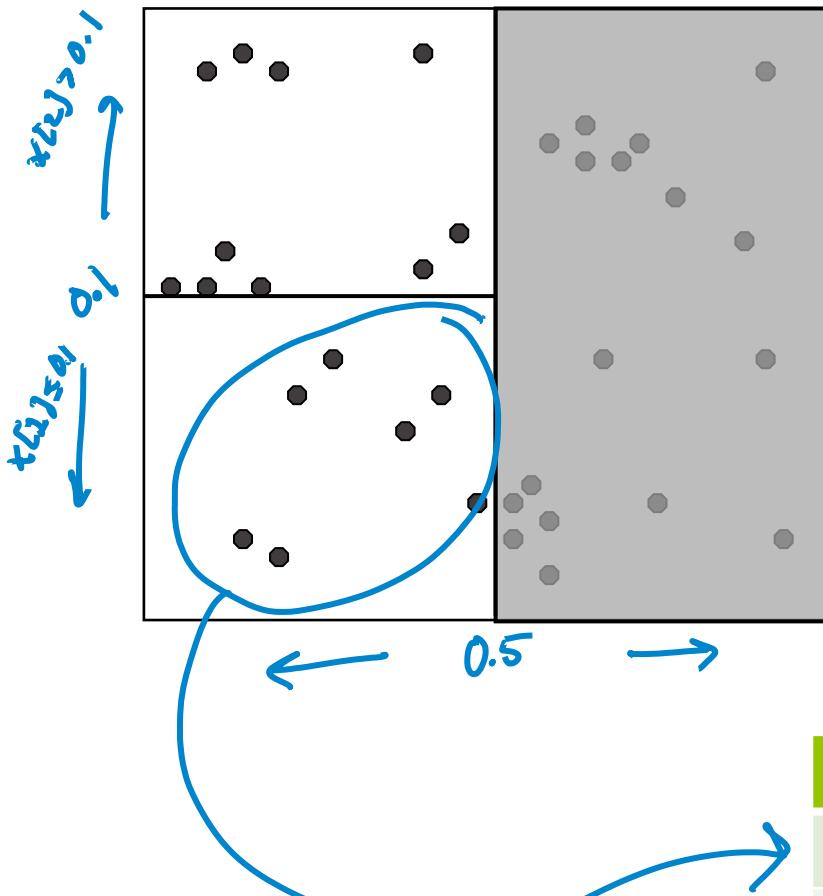


Recurse on each group separately

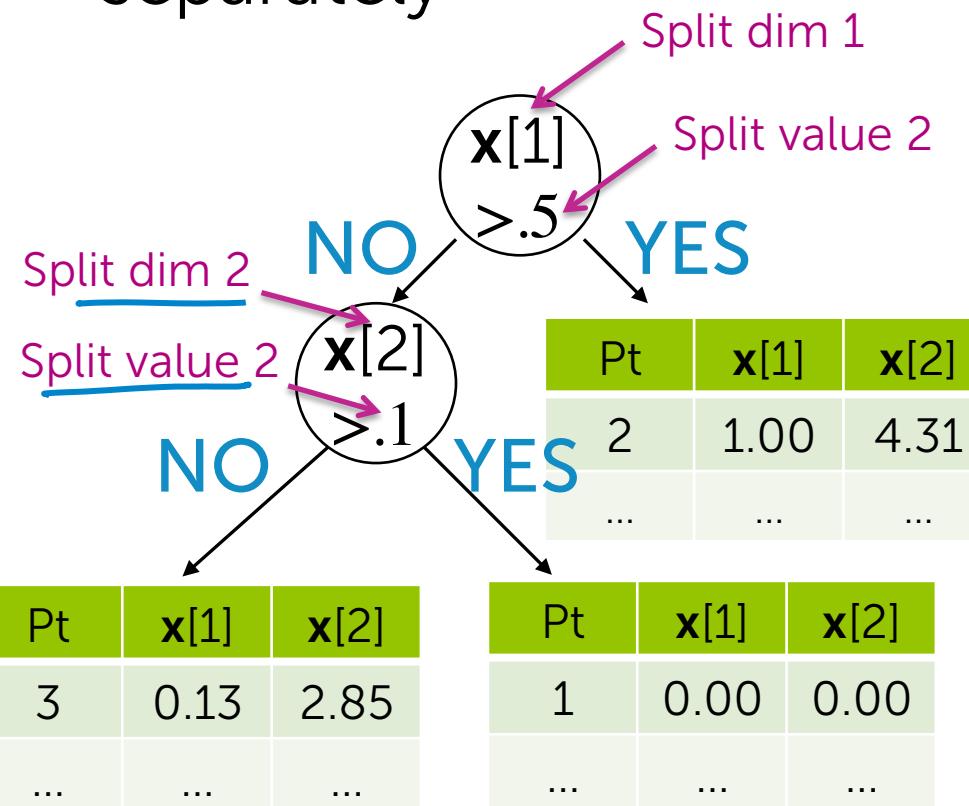


Pt	$x[1]$	$x[2]$	Pt	$x[1]$	$x[2]$
1	0.00	0.00	2	1.00	4.31
3	0.13	2.85
...			

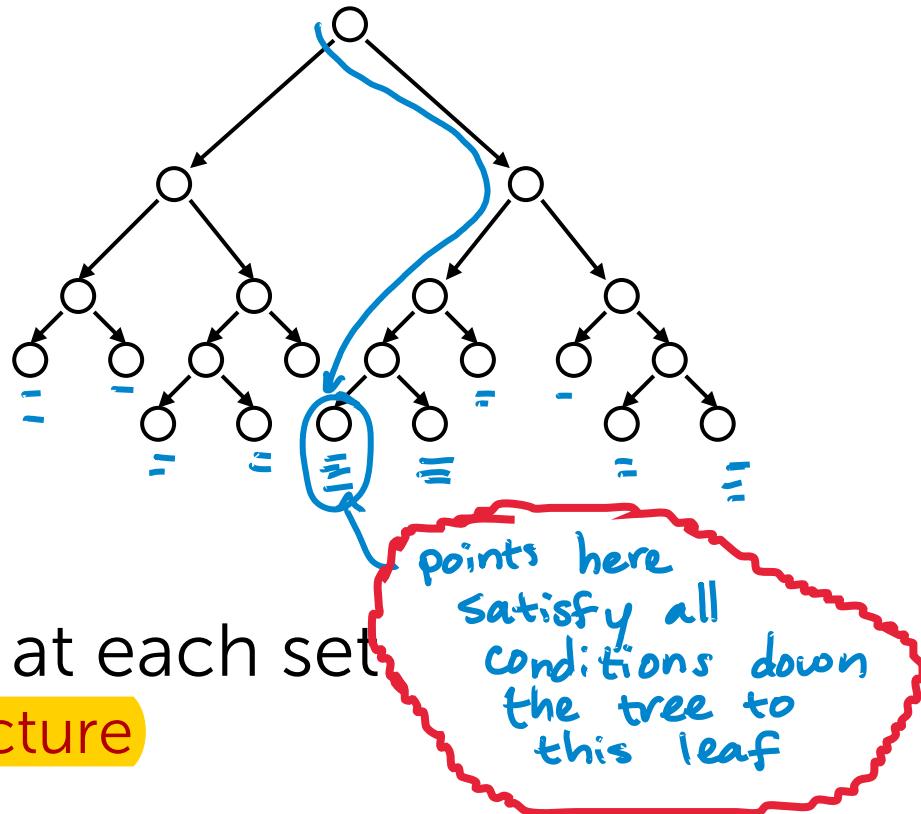
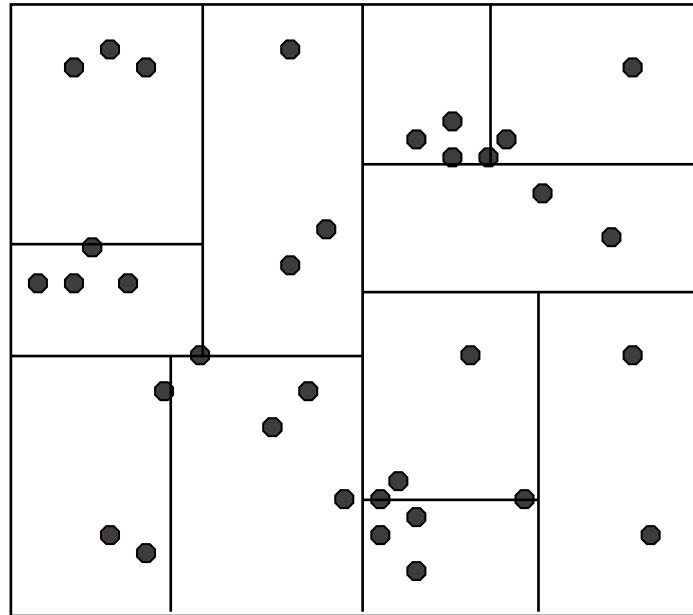
KD-tree construction



Recurse on each group separately



KD-tree construction

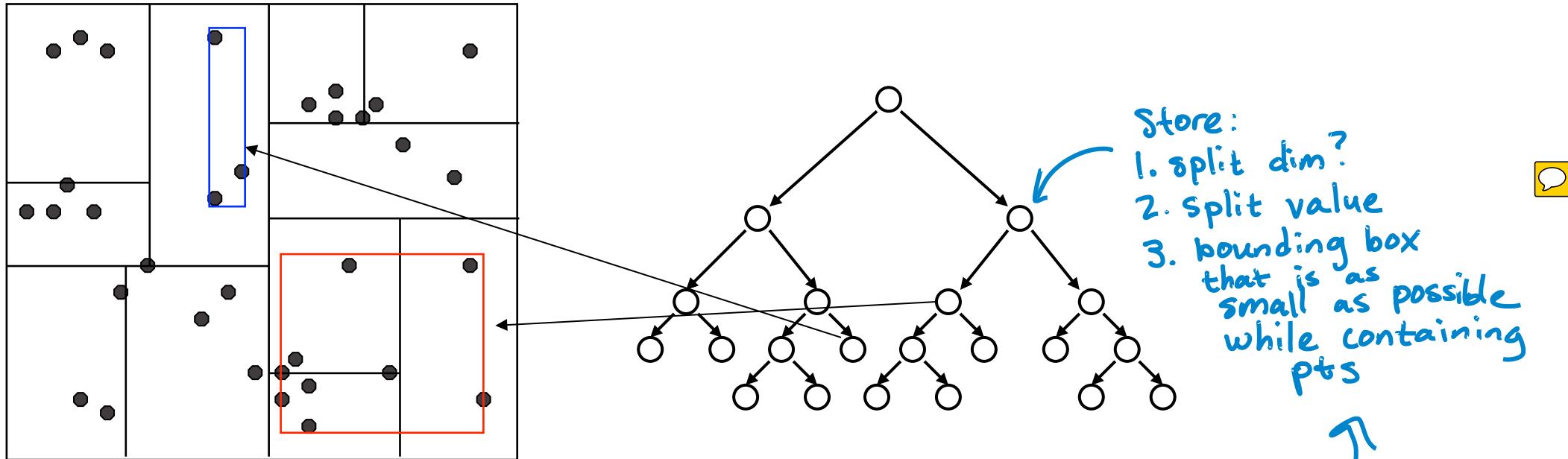


Continue splitting points at each set

- Creates a binary tree structure

Each leaf node contains a list of points

KD-tree construction



Keep one additional piece of info at each node:

#3 - The (tight) bounds of points at or below node

KD-tree construction choices

Use heuristics to make splitting decisions:

- Which dimension do we split along?

widest (or alternate) 

- Which value do we split at?

median (or center point of box,
ignoring data in box)

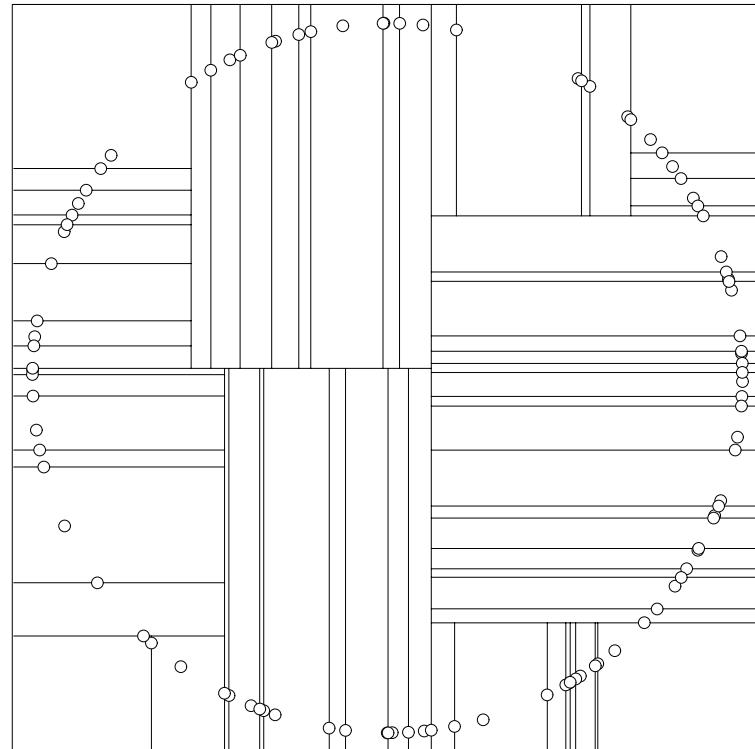
- When do we stop?

fewer than m pts left

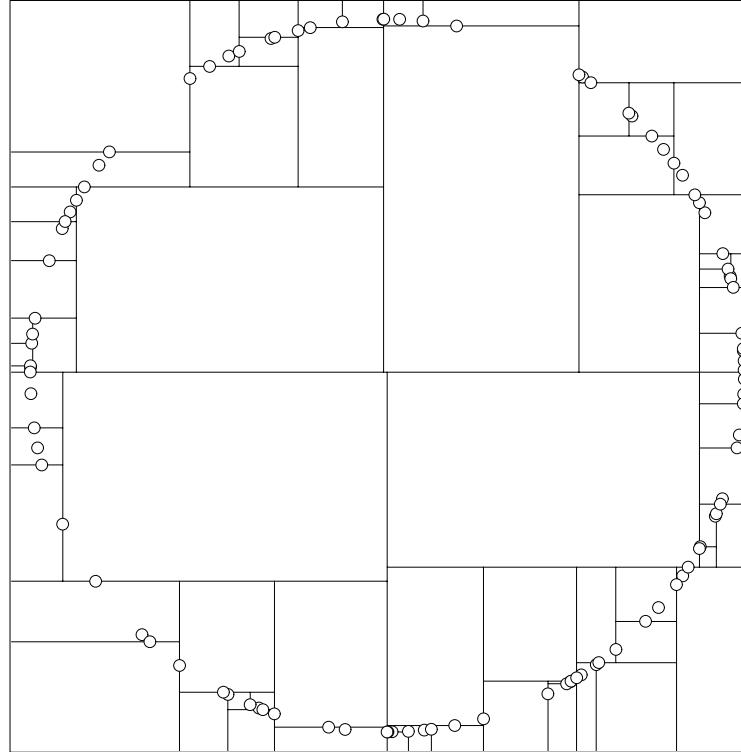
or

box hits minimum width

Many heuristics...



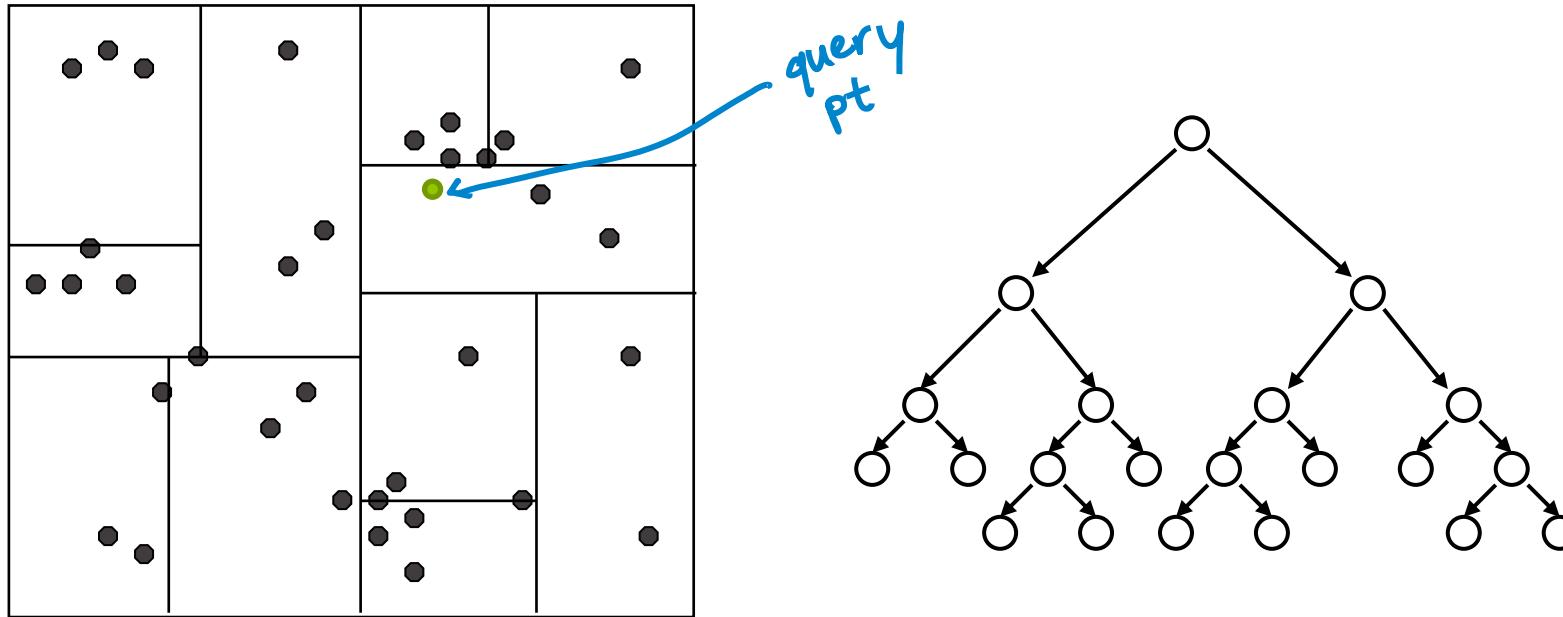
median heuristic



center-of-range
heuristic

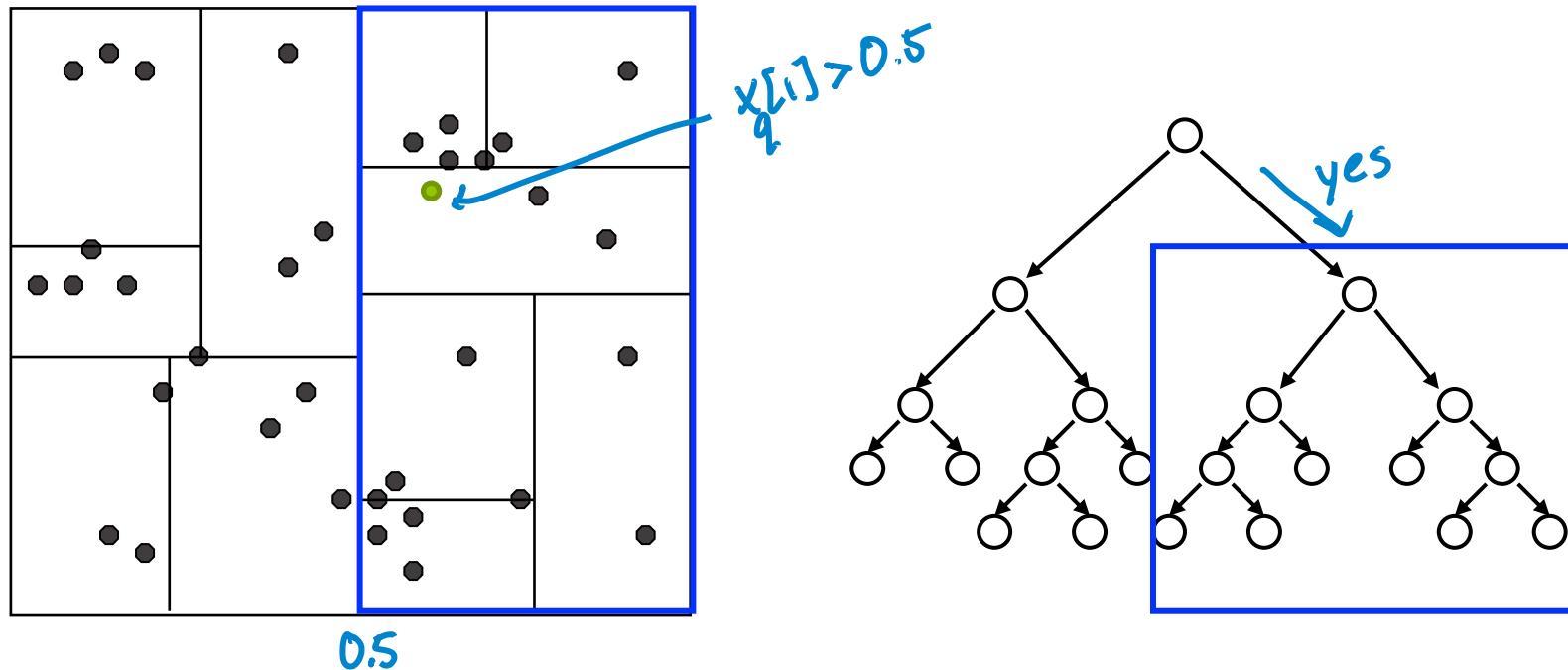
NN search with KD-trees

Nearest neighbor with KD-trees



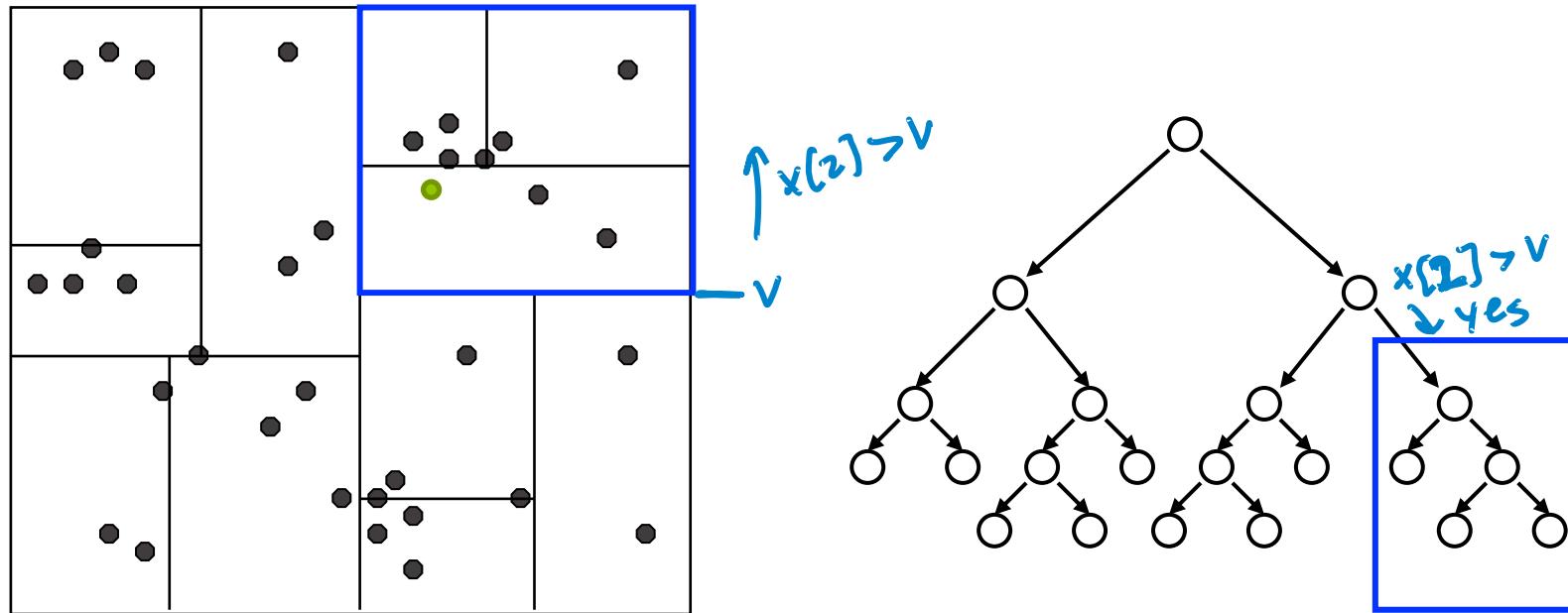
Traverse tree looking for nearest neighbor to query point

Nearest neighbor with KD-trees



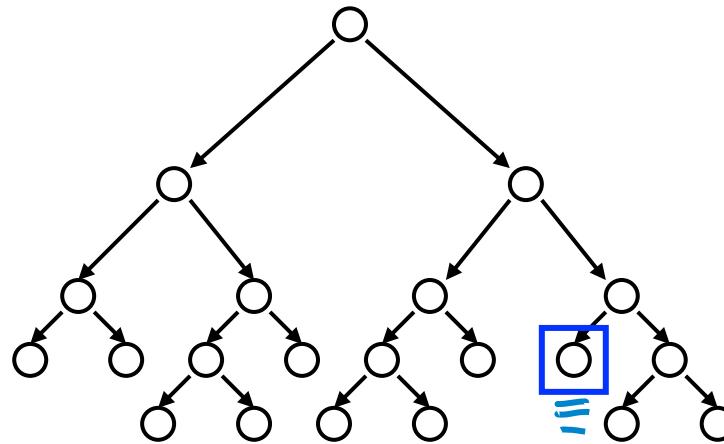
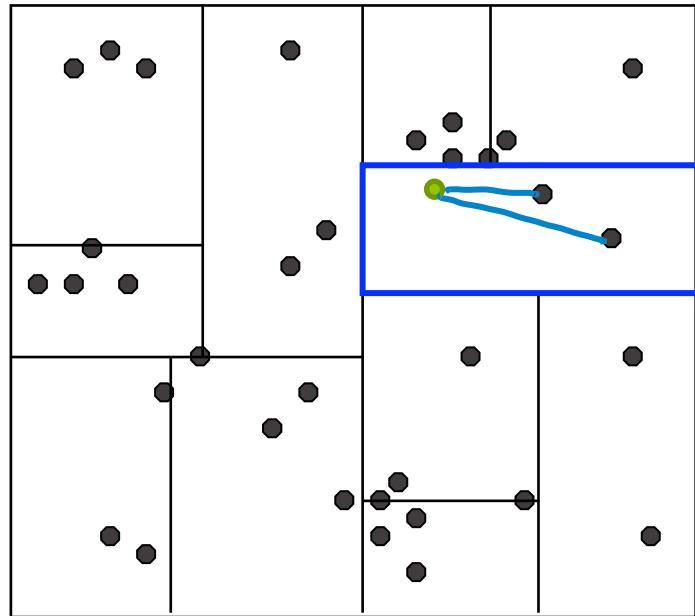
1. Start by exploring leaf node containing query point

Nearest neighbor with KD-trees



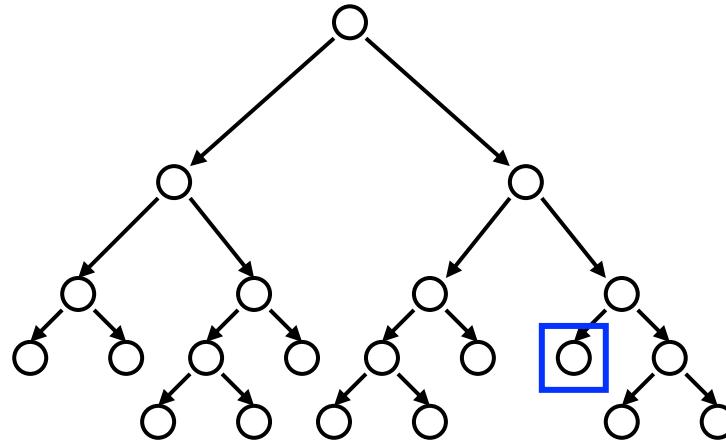
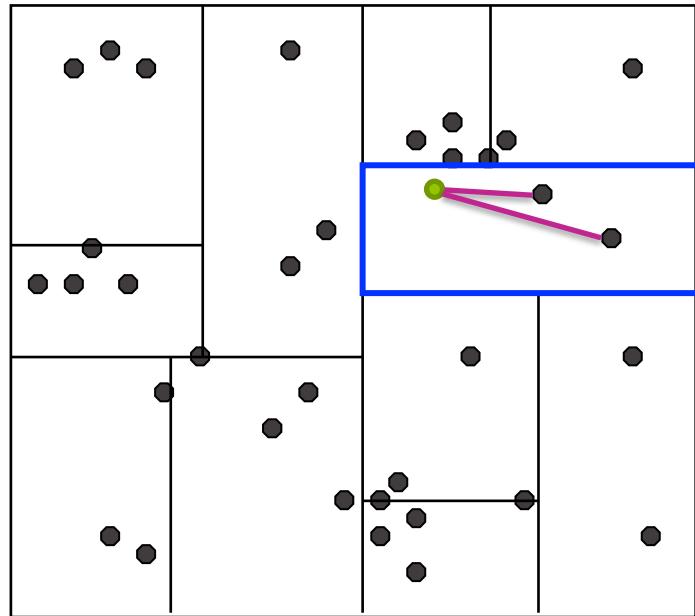
1. Start by exploring leaf node containing query point

Nearest neighbor with KD-trees



1. Start by exploring leaf node containing query point

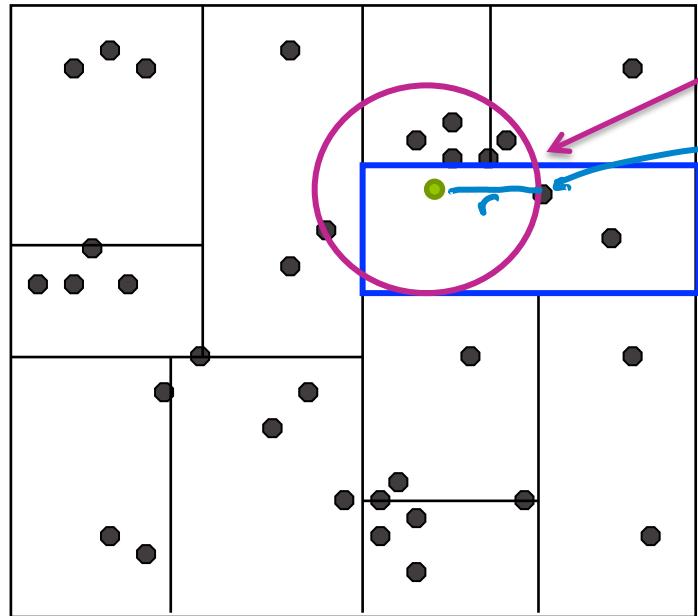
Nearest neighbor with KD-trees



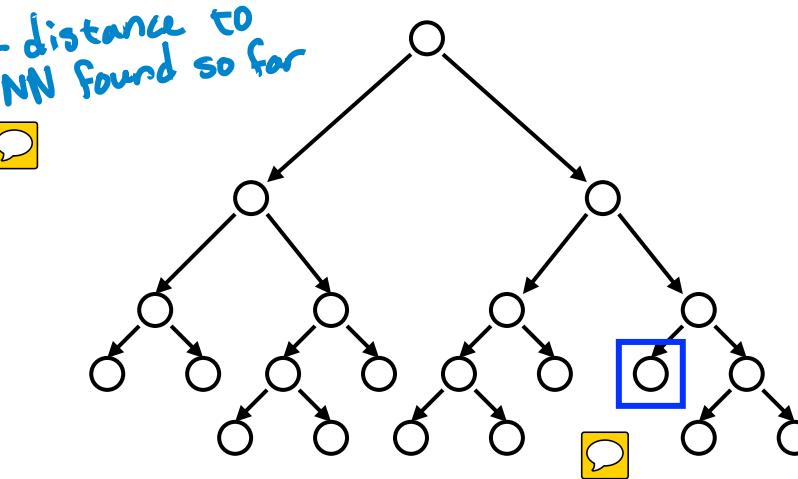
1. Start by exploring leaf node containing query point
2. Compute distance to each other point at leaf node



Nearest neighbor with KD-trees

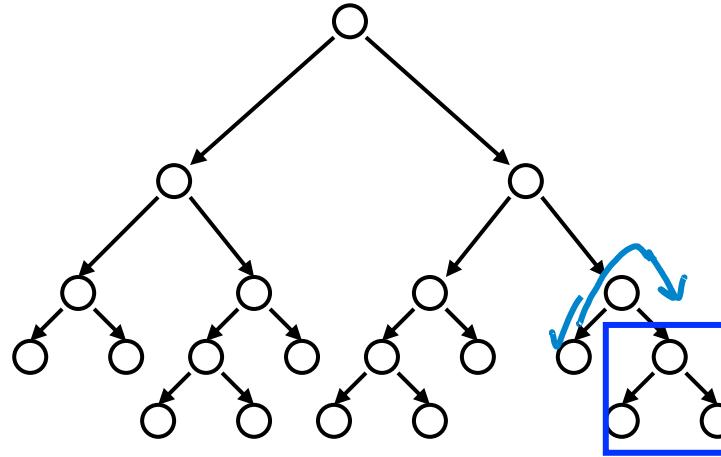
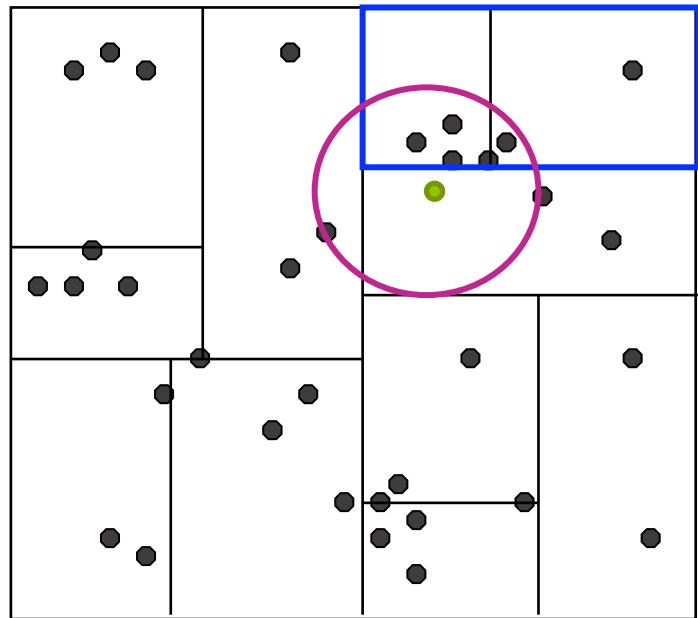


Does nearest neighbor have to live at
leaf node containing query point?



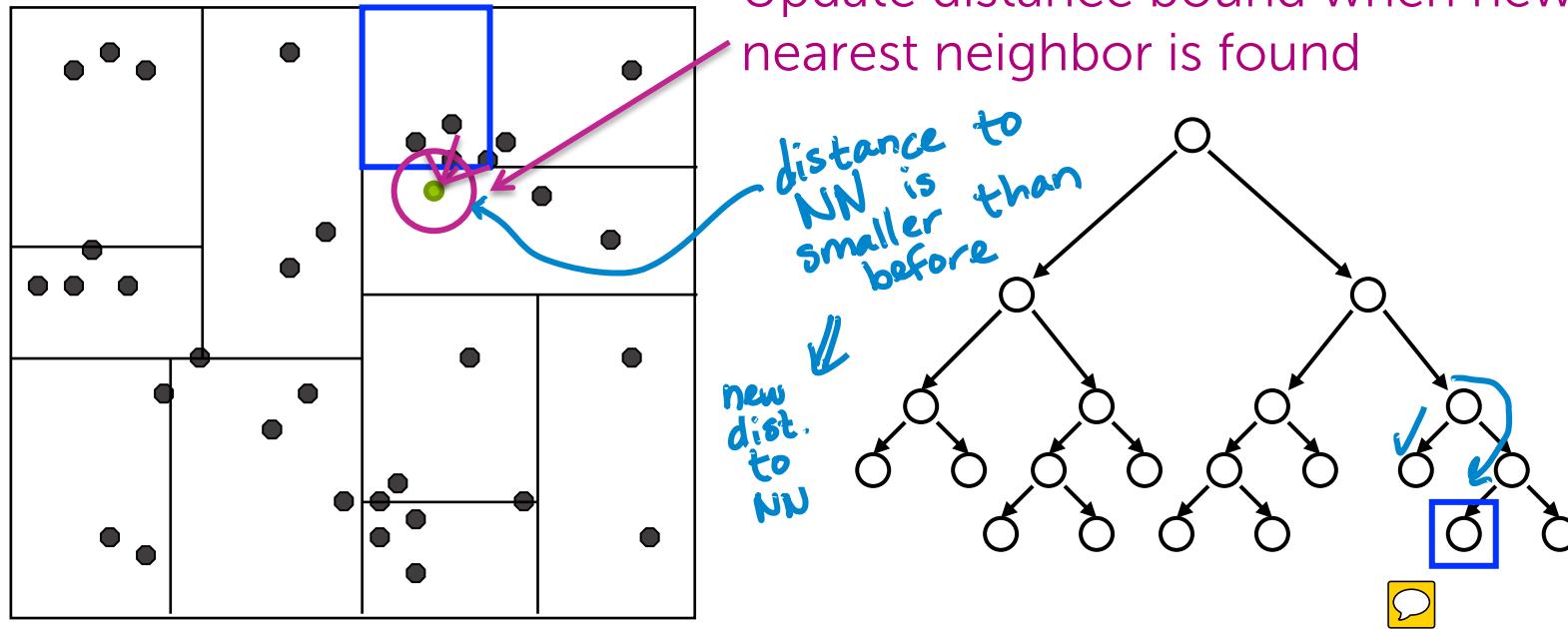
1. Start by exploring leaf node containing query point
2. Compute distance to each other point at leaf node

Nearest neighbor with KD-trees



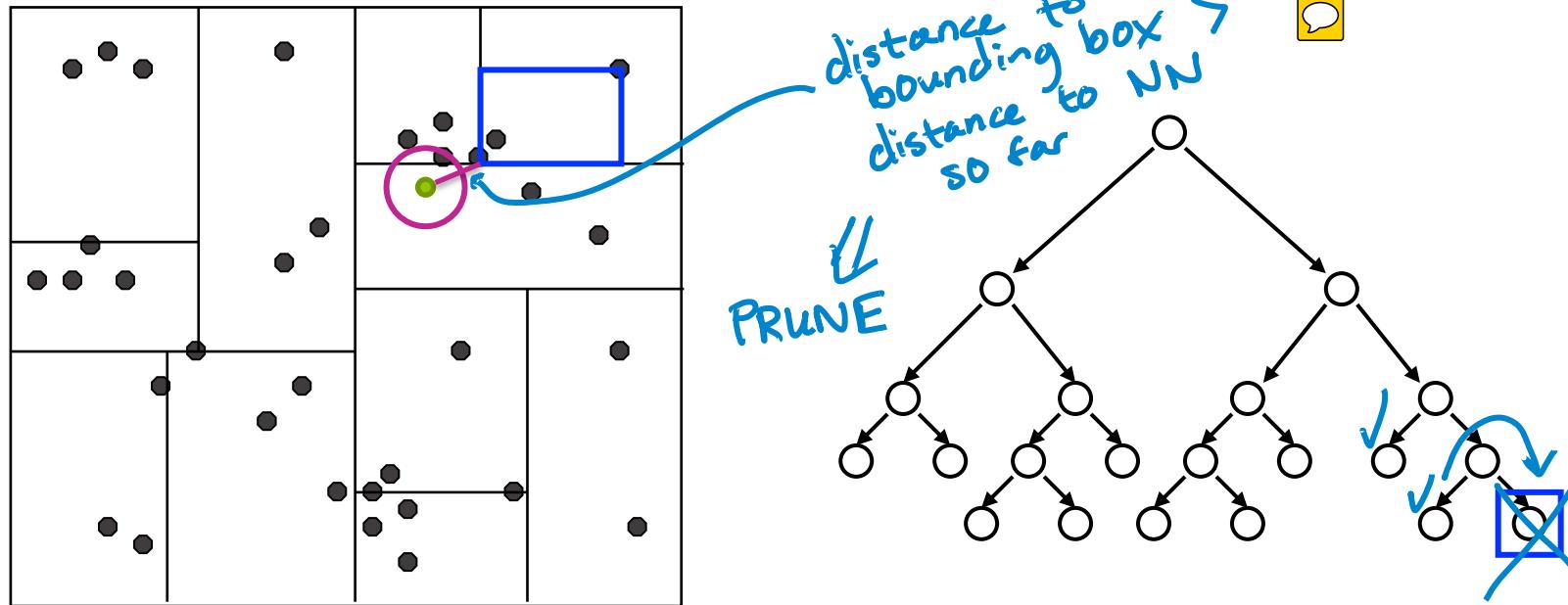
1. Start by exploring leaf node containing query point
2. Compute distance to each other point at leaf node
3. Backtrack and try other branch at each node visited

Nearest neighbor with KD-trees



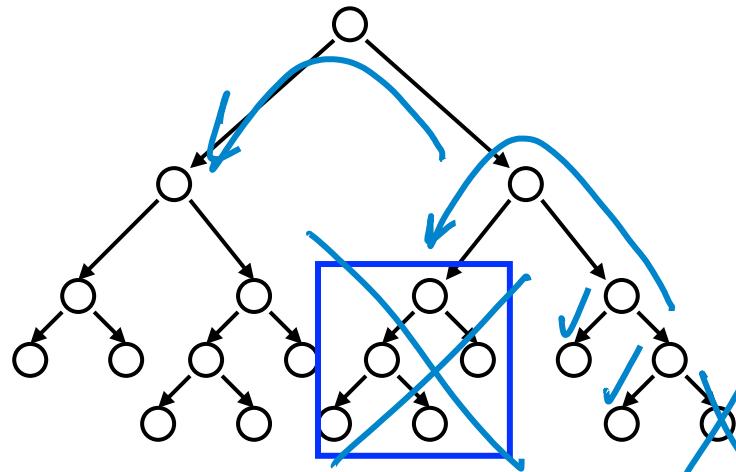
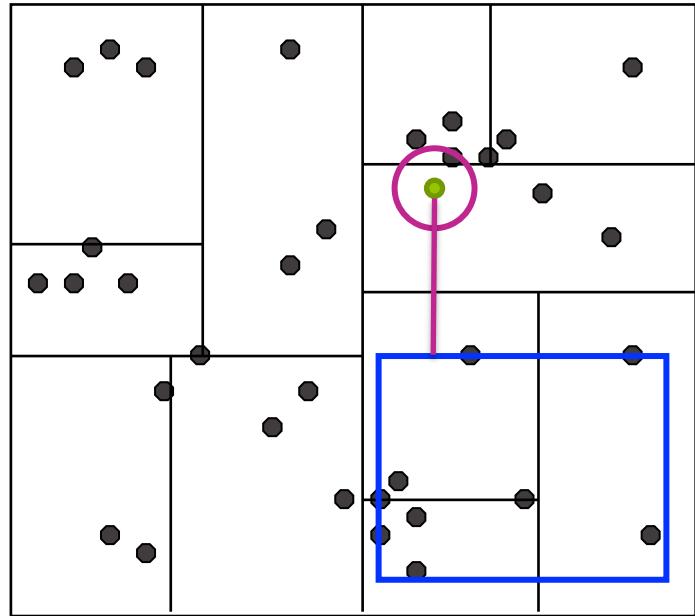
1. Start by exploring leaf node containing query point
2. Compute distance to each other point at leaf node
3. Backtrack and try other branch at each node visited

Nearest neighbor with KD-trees



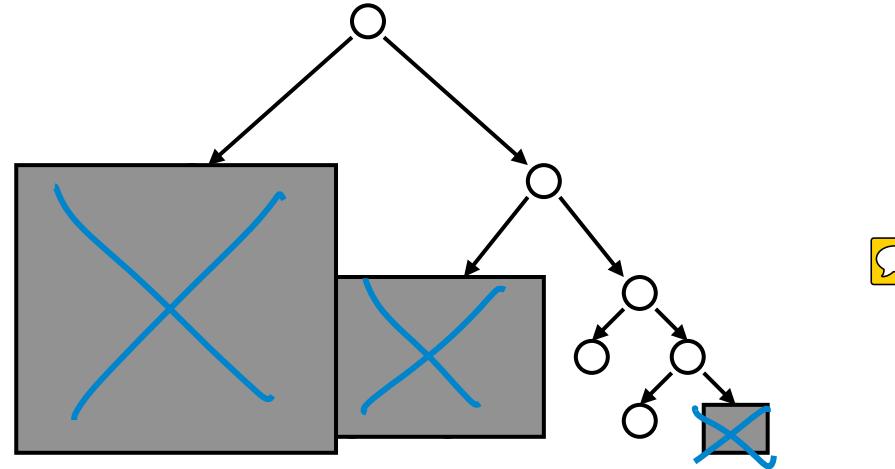
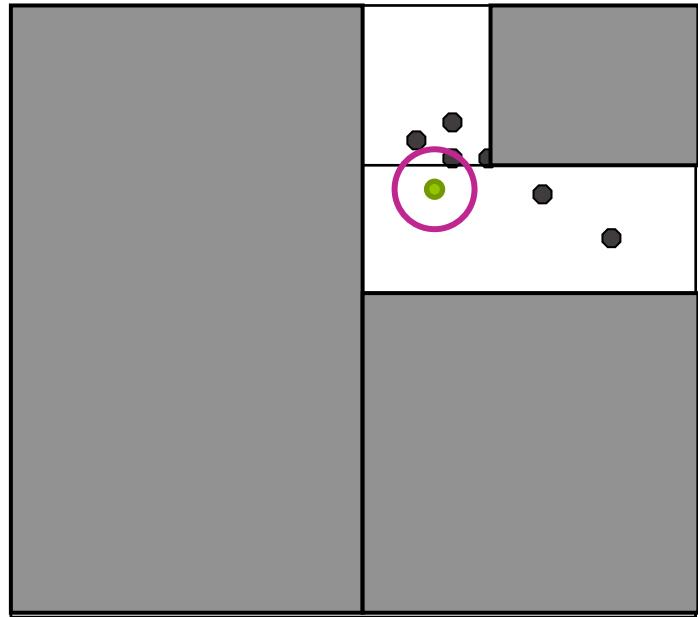
Use distance bound and bounding box of each node to
prune parts of tree that cannot include nearest neighbor

Nearest neighbor with KD-trees



Use distance bound and bounding box of each node to
prune parts of tree that cannot include nearest neighbor

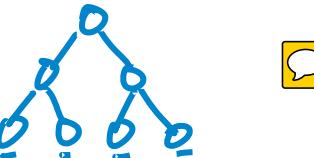
Nearest neighbor with KD-trees



Use distance bound and bounding box of each node to
prune parts of tree that cannot include nearest neighbor



Complexity

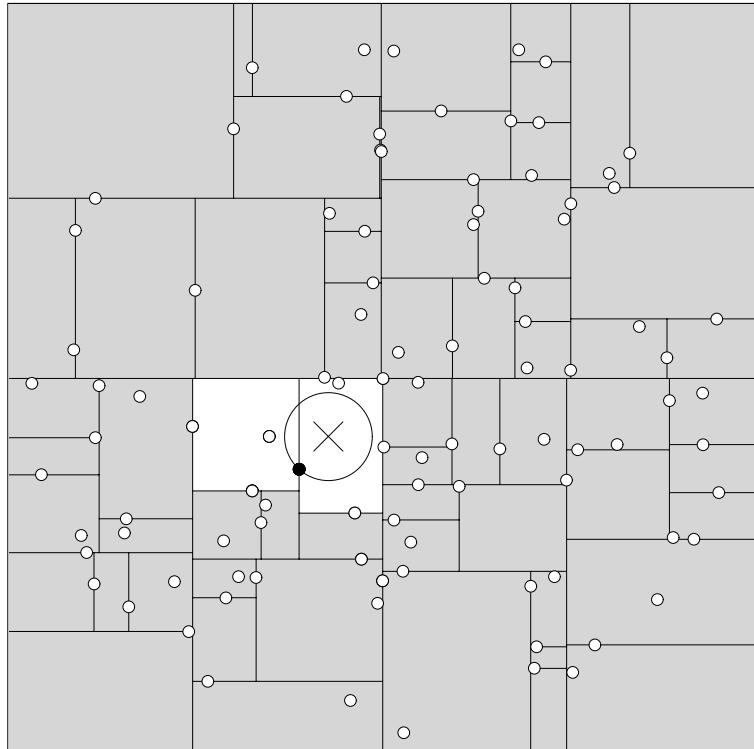


For (nearly) balanced, binary trees...

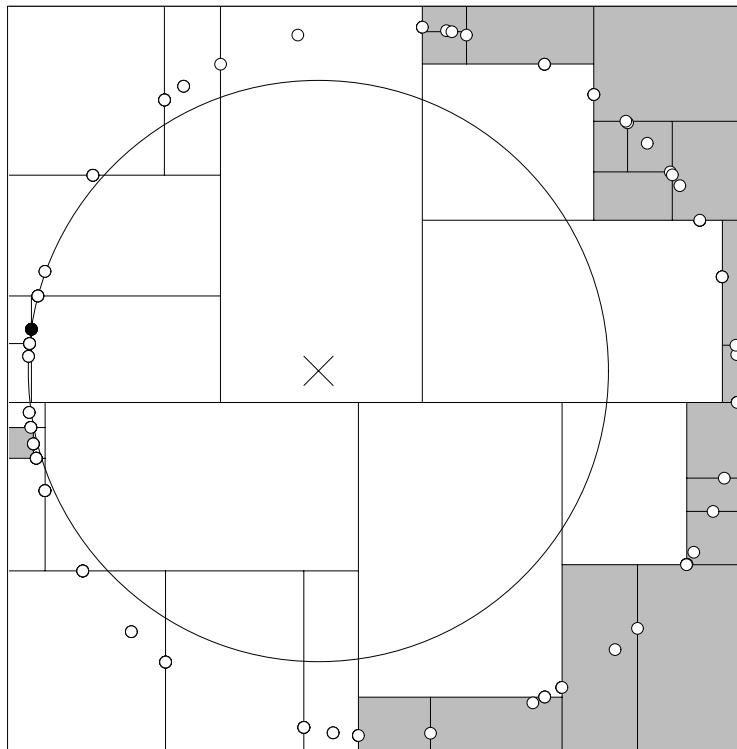
- Construction
 - Size: $2N-1$ nodes if 1 datapt at each leaf $\rightarrow O(N)$
 - Depth: $O(\log N)$
 - Median + send points left right: $O(N)$ at every level of the tree
 - Construction time: $O(N \log N)$
- 1-NN query
 - Traverse down tree to starting point: $O(\log N)$
 - Maximum backtrack and traverse: $O(N)$ in worst case
 - Complexity range: $O(\log N) \rightarrow O(N)$

Under some assumptions on distribution of points,
we get $O(\log N)$ but exponential in d

Complexity



pruned many
(closer to $O(\log N)$)



pruned few
(closer to $O(N)$)

Complexity for N queries

- Ask for nearest neighbor to each doc

N queries

- Brute force 1-NN:

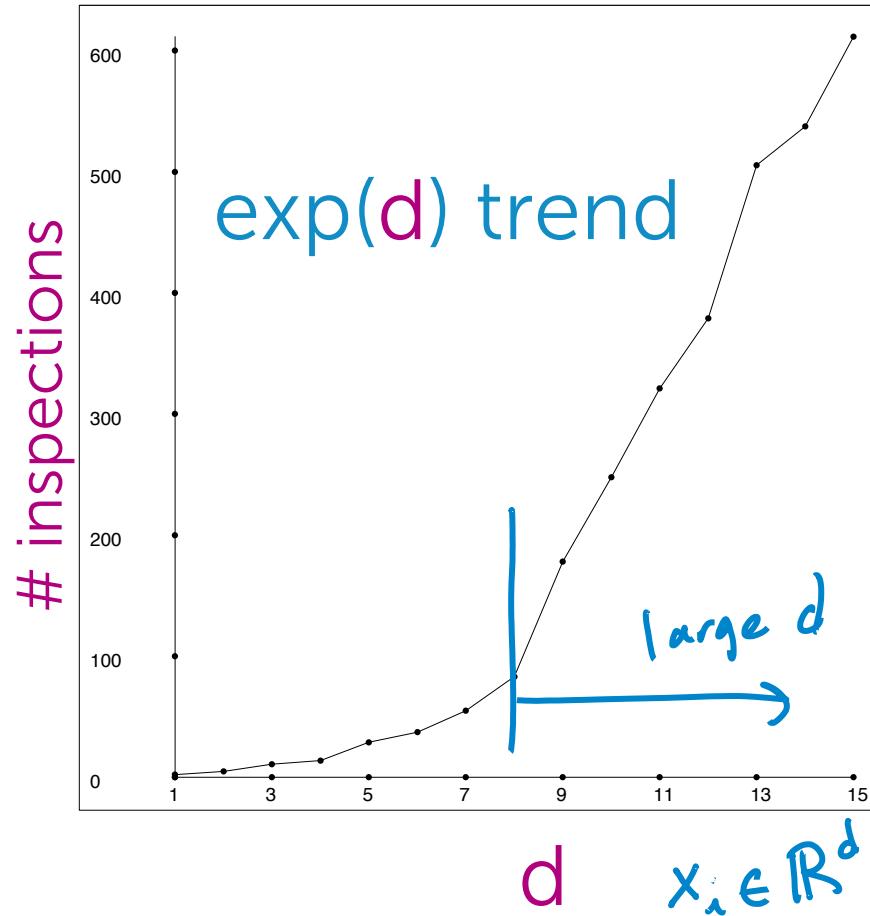
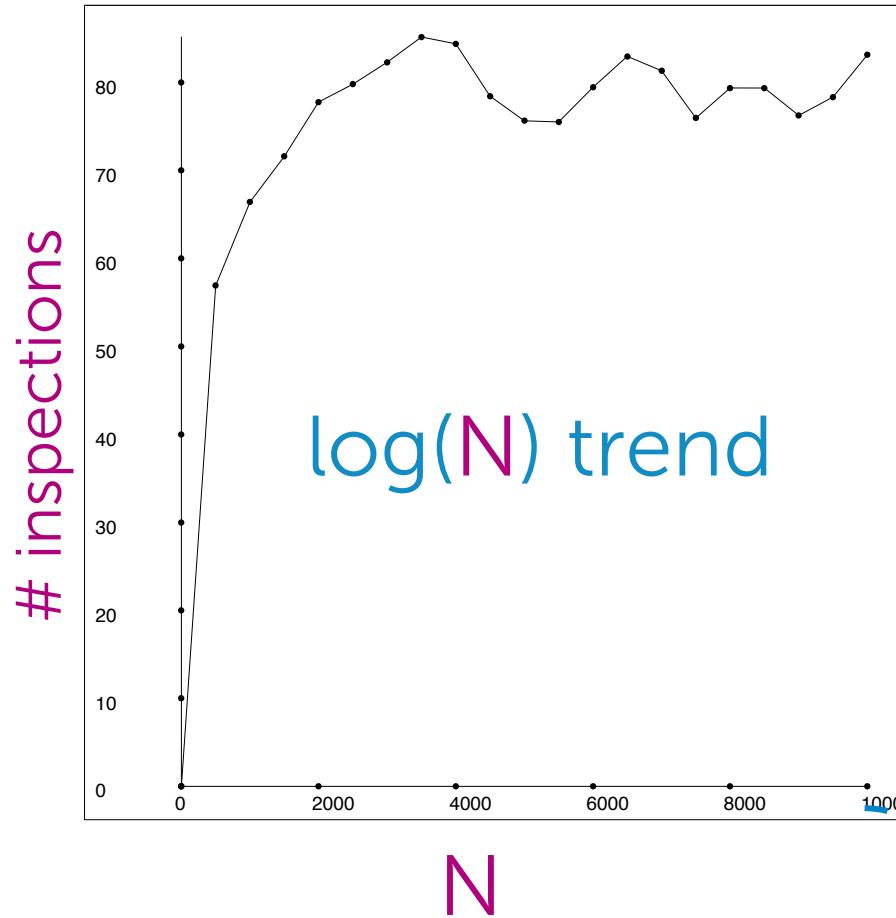
$O(N^2)$

- kd-trees:

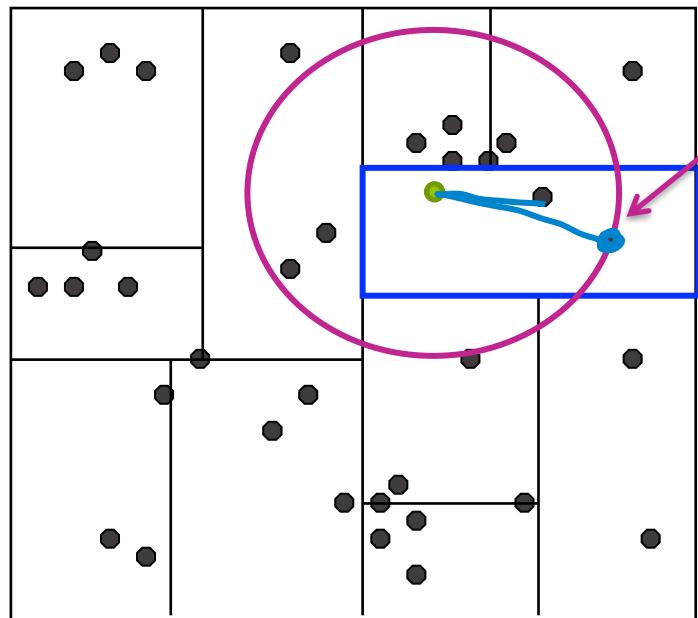
$O(N \log N) \rightarrow O(N^2)$



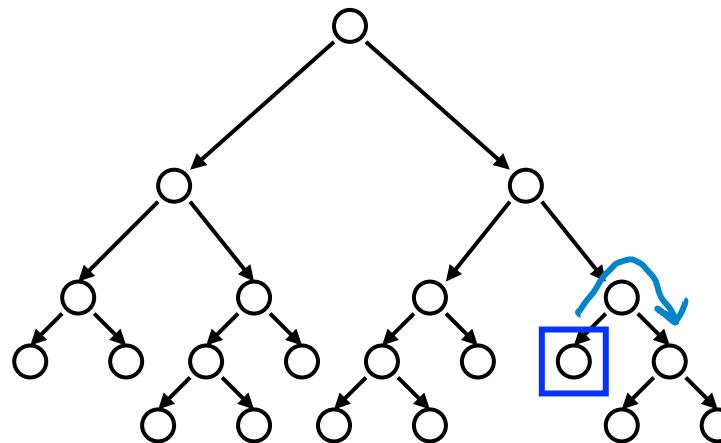
Inspections vs. N and d



k-NN with KD-trees



distance to 2nd nearest neighbor
in 2-NN example

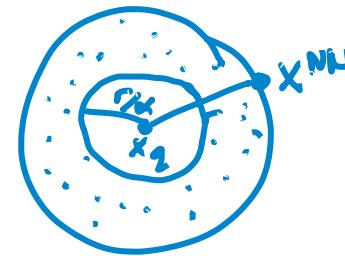
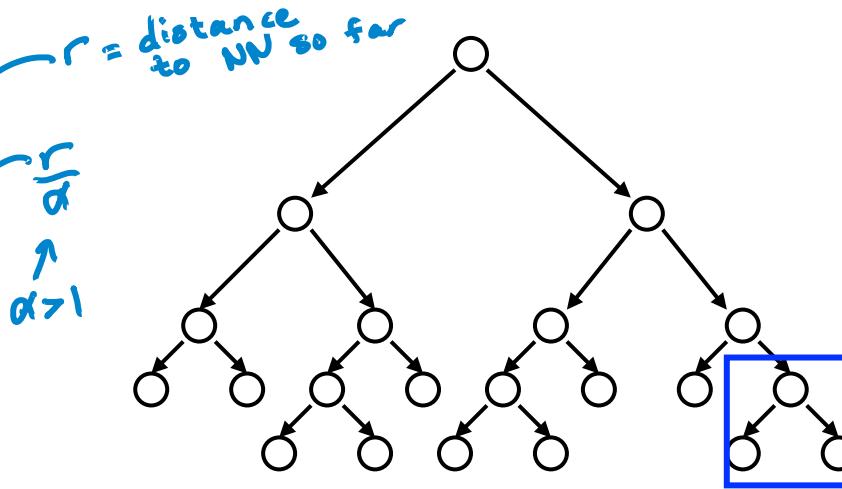
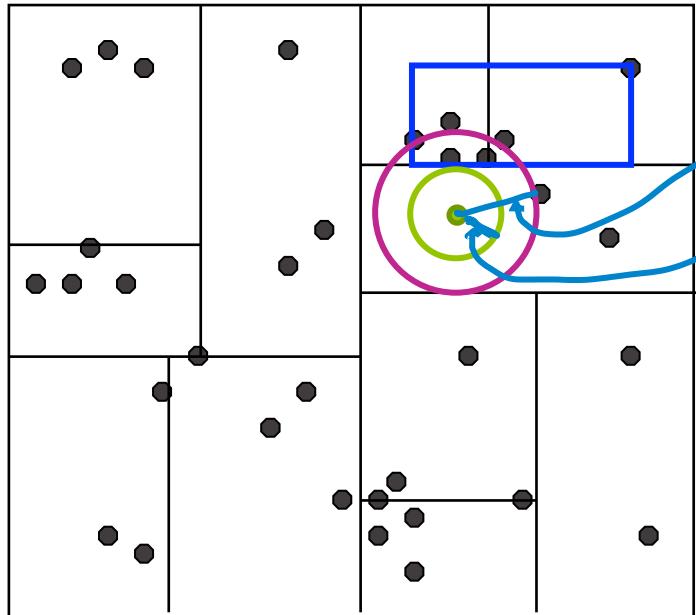


Exactly same algorithm, but maintain distance to
furthest of current k nearest neighbors

Approximate k-NN search



Approximate k-NN with KD-trees



Before: Prune when distance to bounding box $> r$



Now: Prune when distance to bounding box $> r/\alpha$

Prunes more than allowed, but can **guarantee** that if we return a neighbor at distance r , then there is **no neighbor closer** than r/α



Bound loose...In practice, often closer to optimal.

Saves lots of search time at little cost in quality of NN!

Closing remarks on KD-trees

Tons of variants of kd-trees

- On construction of trees
(heuristics for splitting, stopping, representing branches...)
- Other representational data structures for fast NN search
(e.g., ball trees,...)

Nearest Neighbor Search

- Distance metric and data representation crucial to answer returned

For both, high-dim spaces are hard!

- Number of kd-tree searches can be exponential in dimension
 - Rule of thumb... $N \gg 2^d$... **Typically useless for large d.**
- Distances sensitive to irrelevant features
 - Most dimensions are just noise → everything is far away
 - Need technique to learn which features are important to given task



Acknowledgements

Thanks to Andrew Moore
(<http://www.cs.cmu.edu/~awm/>)
for the KD-trees slide outline

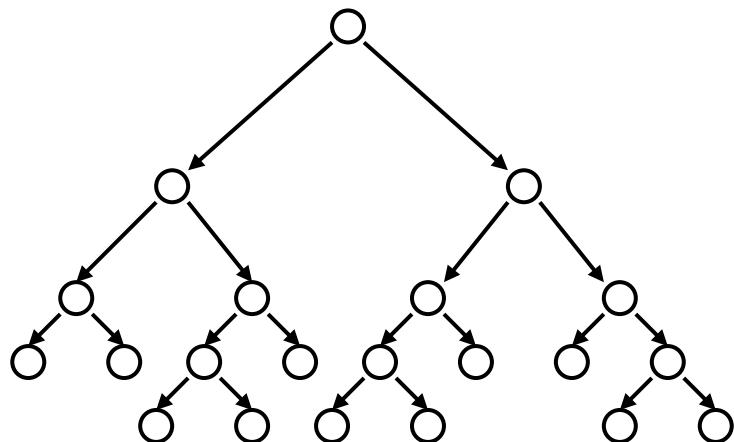


Locality sensitive hashing for approximate NN search



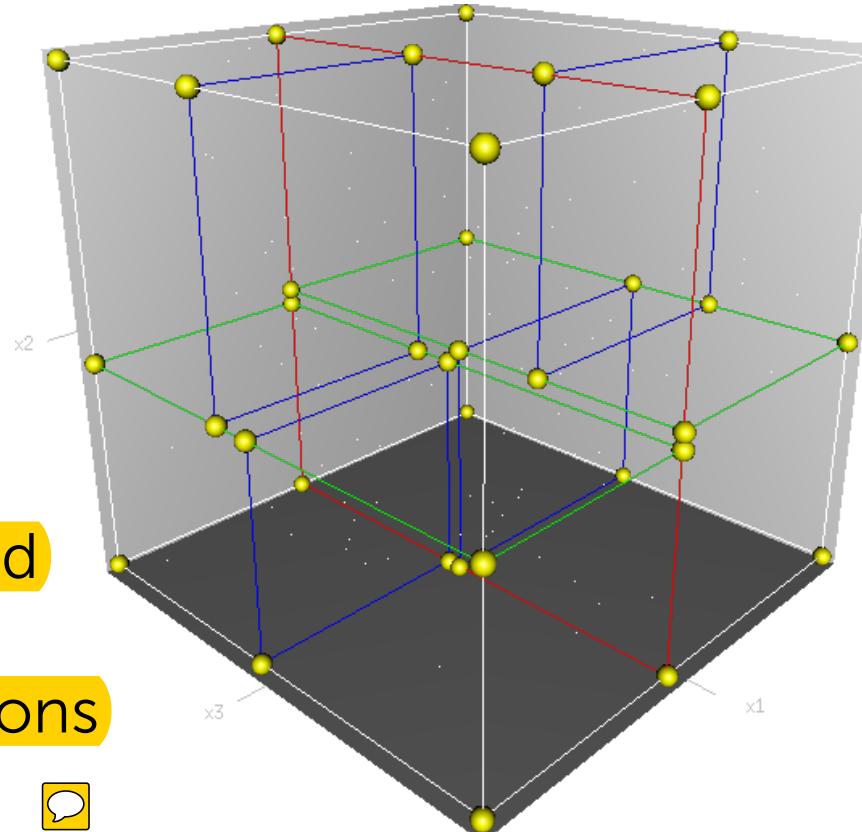
Motivating alternative approaches to approximate NN search

- KD-trees are cool, but...
 - Non-trivial to implement efficiently
 - Problems with high-dimensional data



KD-trees in high dimensions

- Unlikely to have any data points close to query point
- Once “nearby” point is found, the search radius is likely to **intersect many hypercubes** in at least one dim
- Not many nodes can be pruned
- Can show under some conditions that you **visit at least 2^d nodes**



Moving away from exact NN search

- Approximate neighbor finding...
 - Don't find exact neighbor, but **that's okay** for many applications

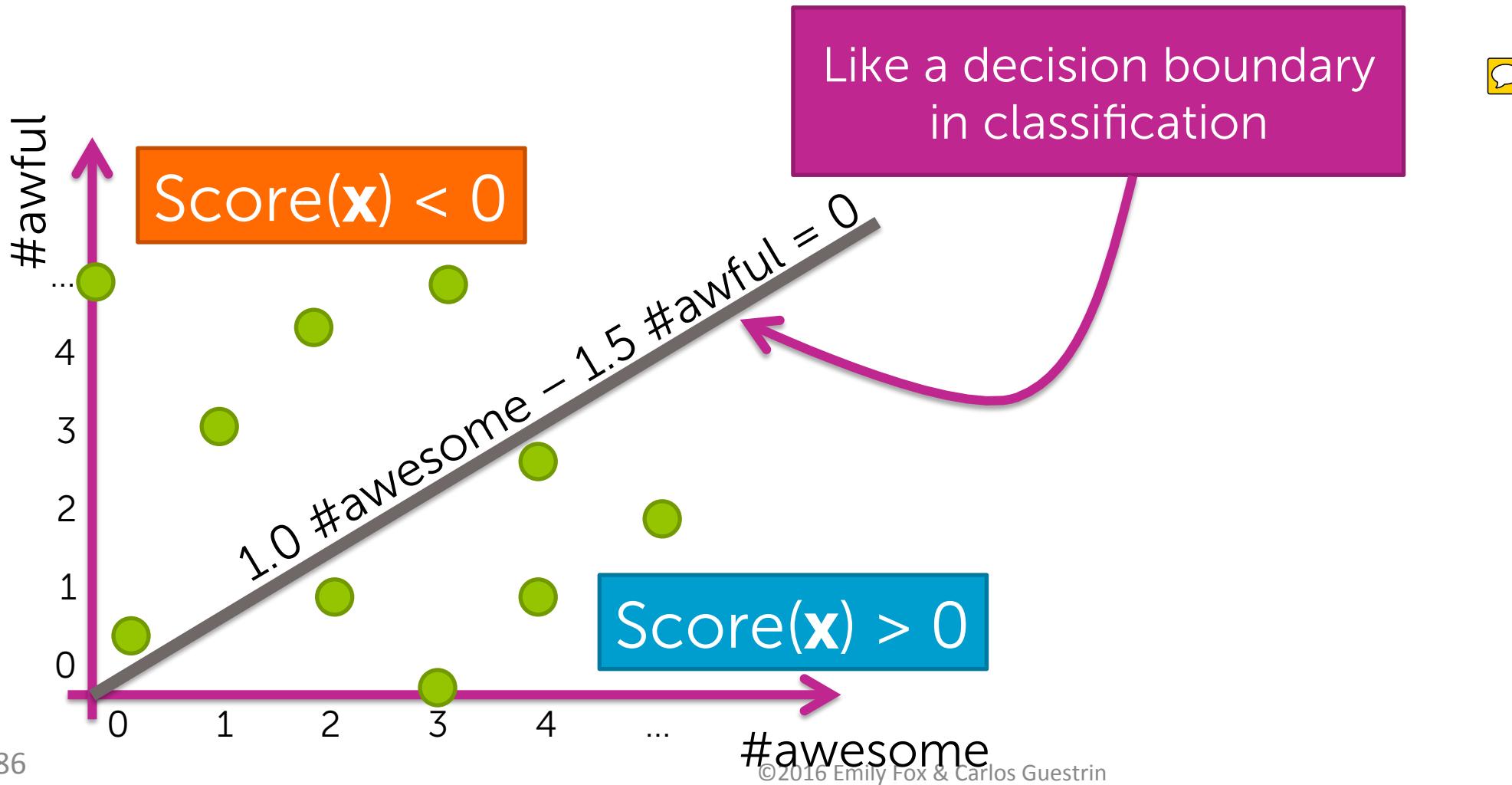


- **Focus on methods that provide good probabilistic guarantees on approximation**

LSH as an alternative to KD-trees

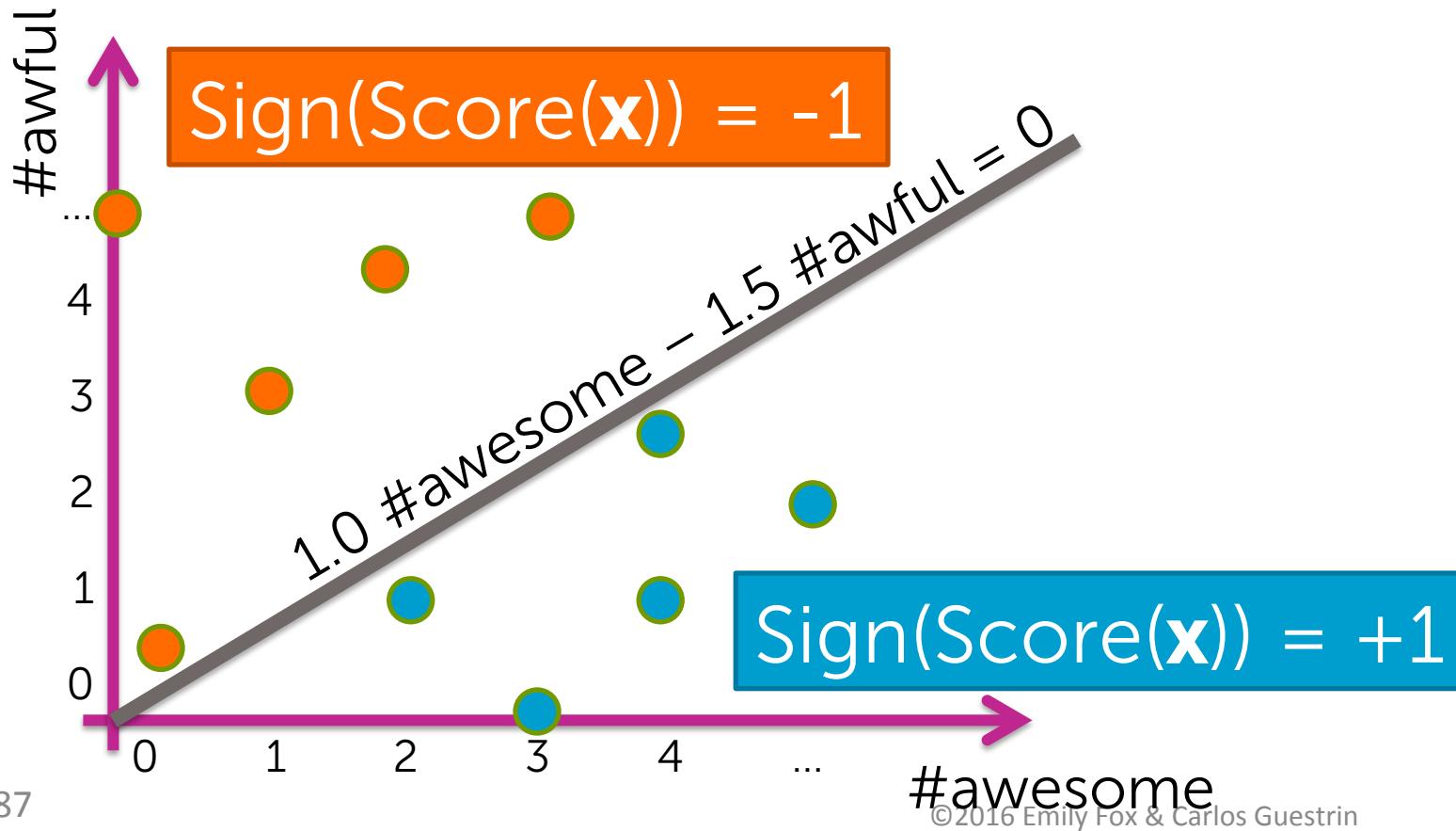
Simple “binning” of data into 2 bins

$\text{Score}(\mathbf{x}) = 1.0 \# \text{awesome} - 1.5 \# \text{awful}$

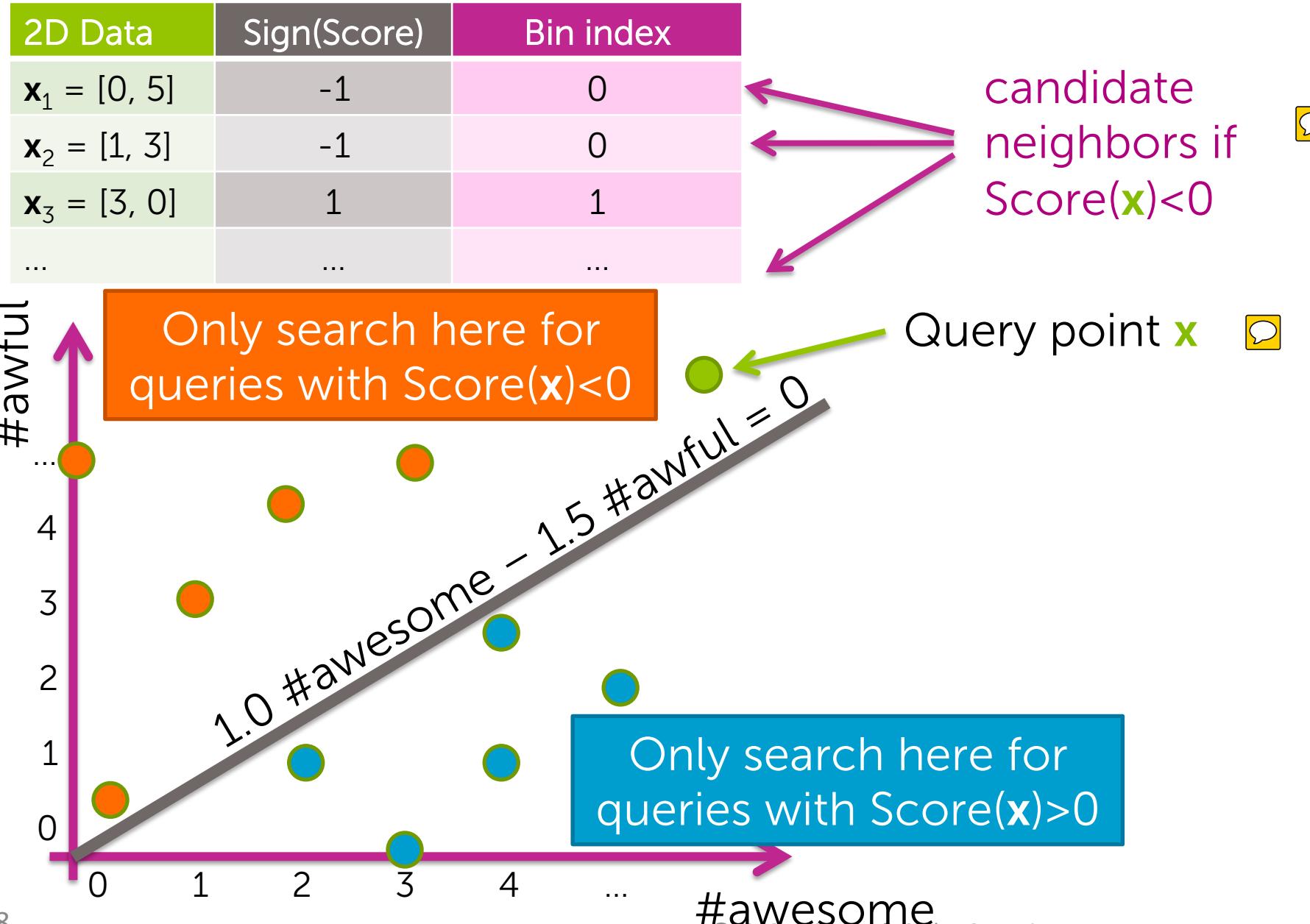


Simple “binning” of data into 2 bins

2D Data	Sign(Score)
$\mathbf{x}_1 = [0, 5]$	-1
$\mathbf{x}_2 = [1, 3]$	-1
$\mathbf{x}_3 = [3, 0]$	1
...	...



Using bins for NN search



Using score for NN search

2D Data	Sign(Score)	Bin index
$\mathbf{x}_1 = [0, 5]$	-1	0
$\mathbf{x}_2 = [1, 3]$	-1	0
$\mathbf{x}_3 = [3, 0]$	1	1
...

candidate
neighbors if
 $\text{Score}(x) < 0$

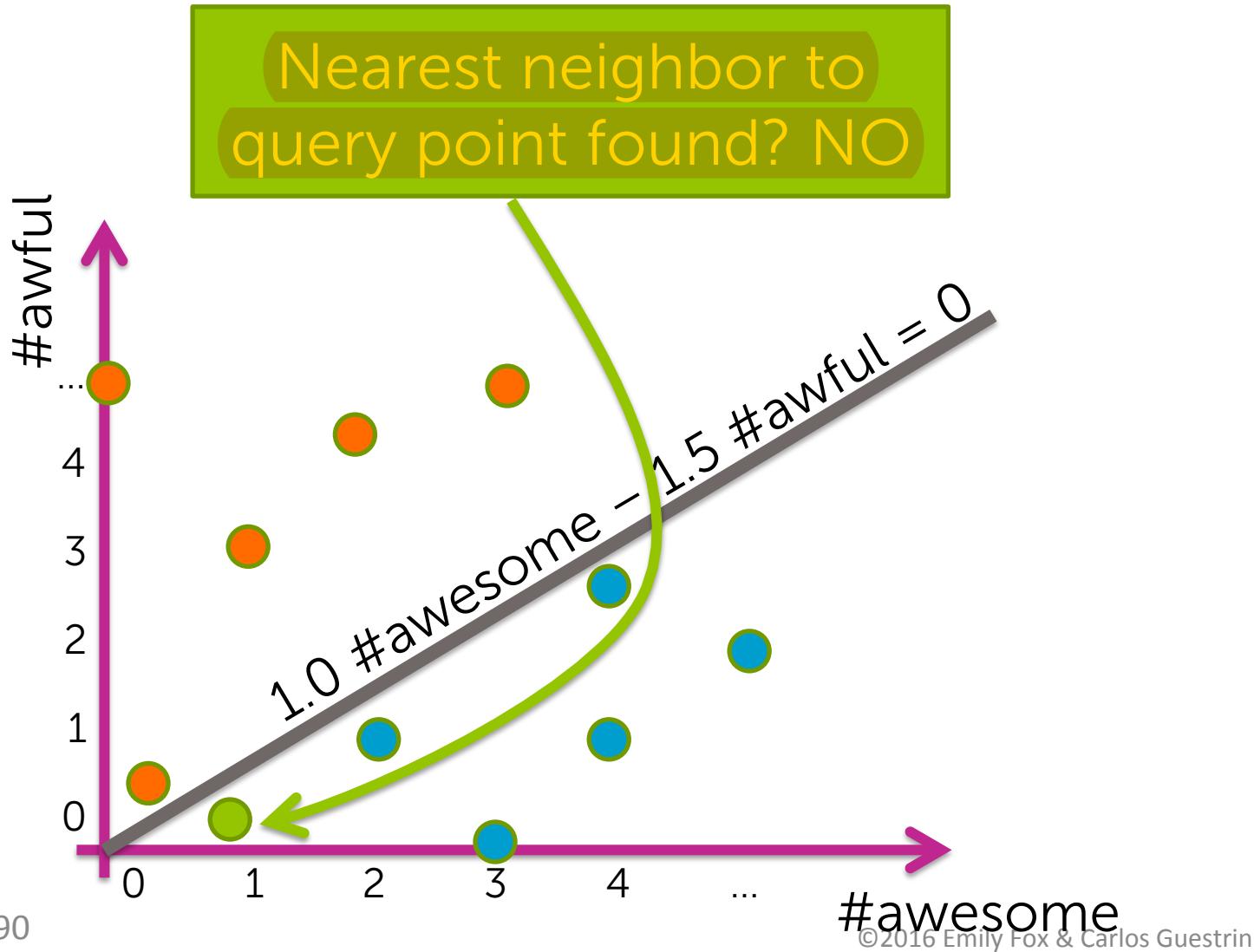
Bin	0	1
List containing indices of datapoints:	{1,2,4,7,...}	{3,5,6,8,...}

HASH
TABLE



search for NN
amongst this set

Provides approximate NN



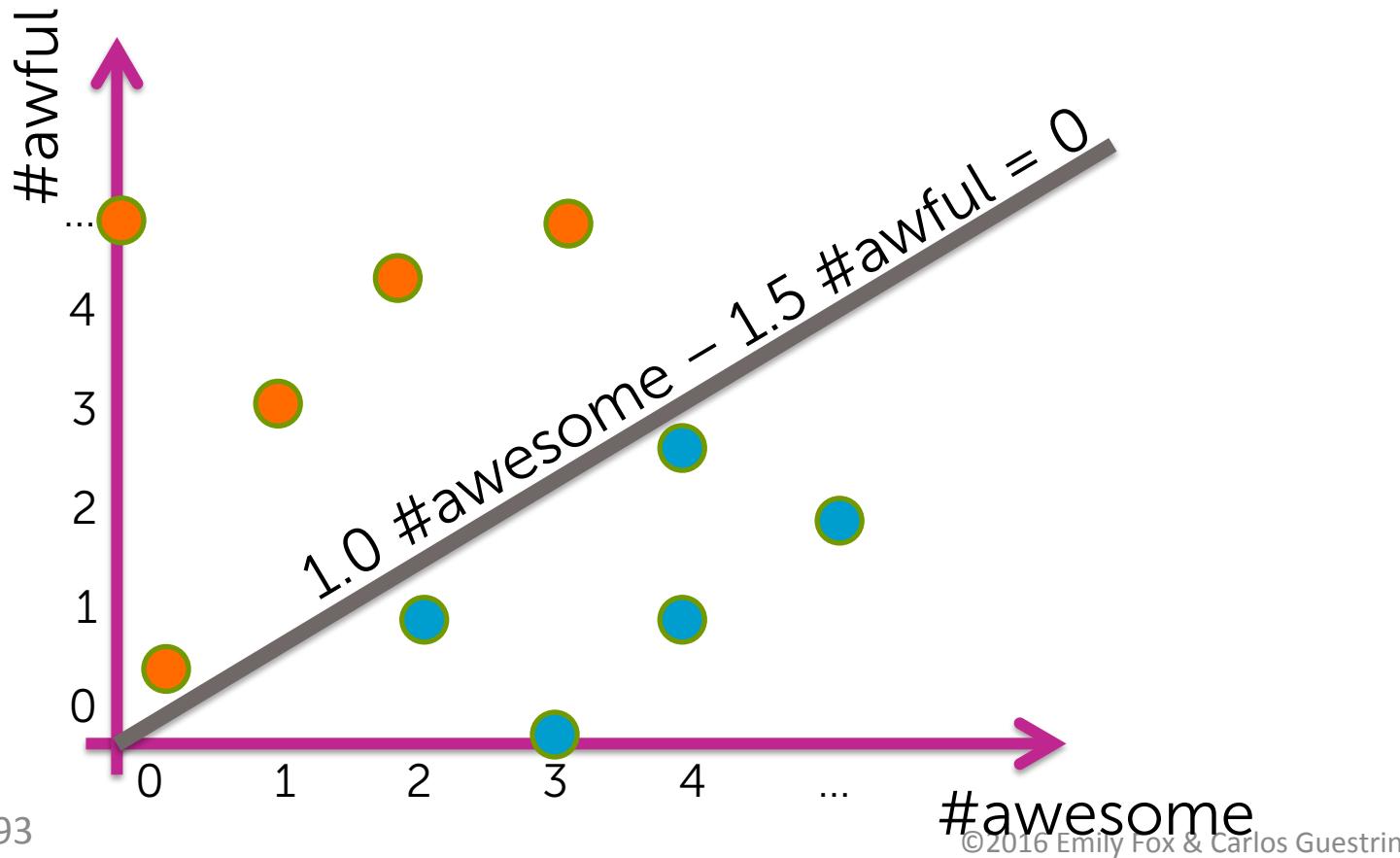
A practical implementation

Three potential issues with simple approach

1. Challenging to find good line
2. Poor quality solution:
 - Points close together get split into separate bins
3. Large computational cost:
 - Bins might contain many points, so still searching over large set for each NN query

How to define the line?

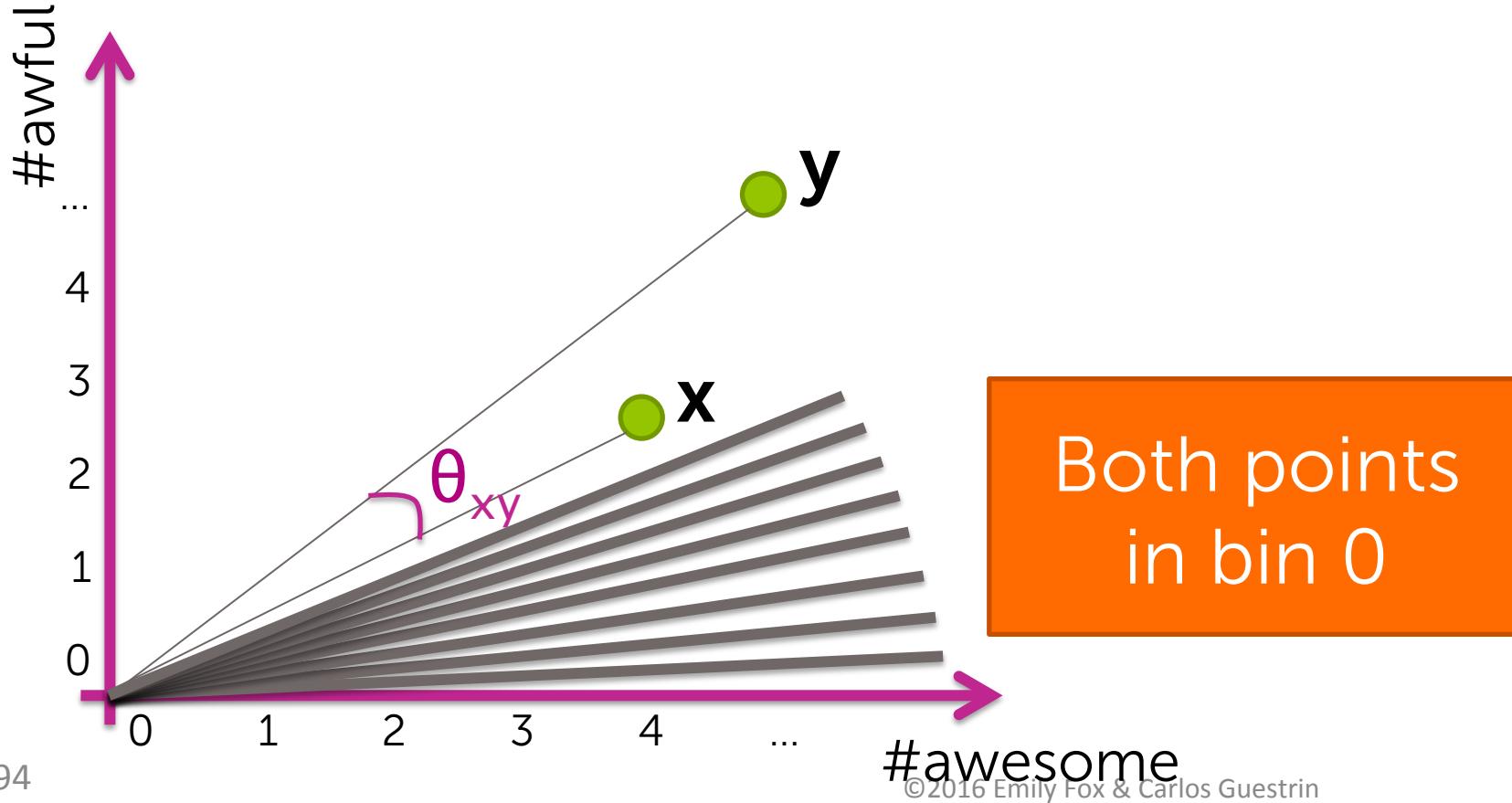
Crazy idea:
Define line randomly!



How bad can a random line be?

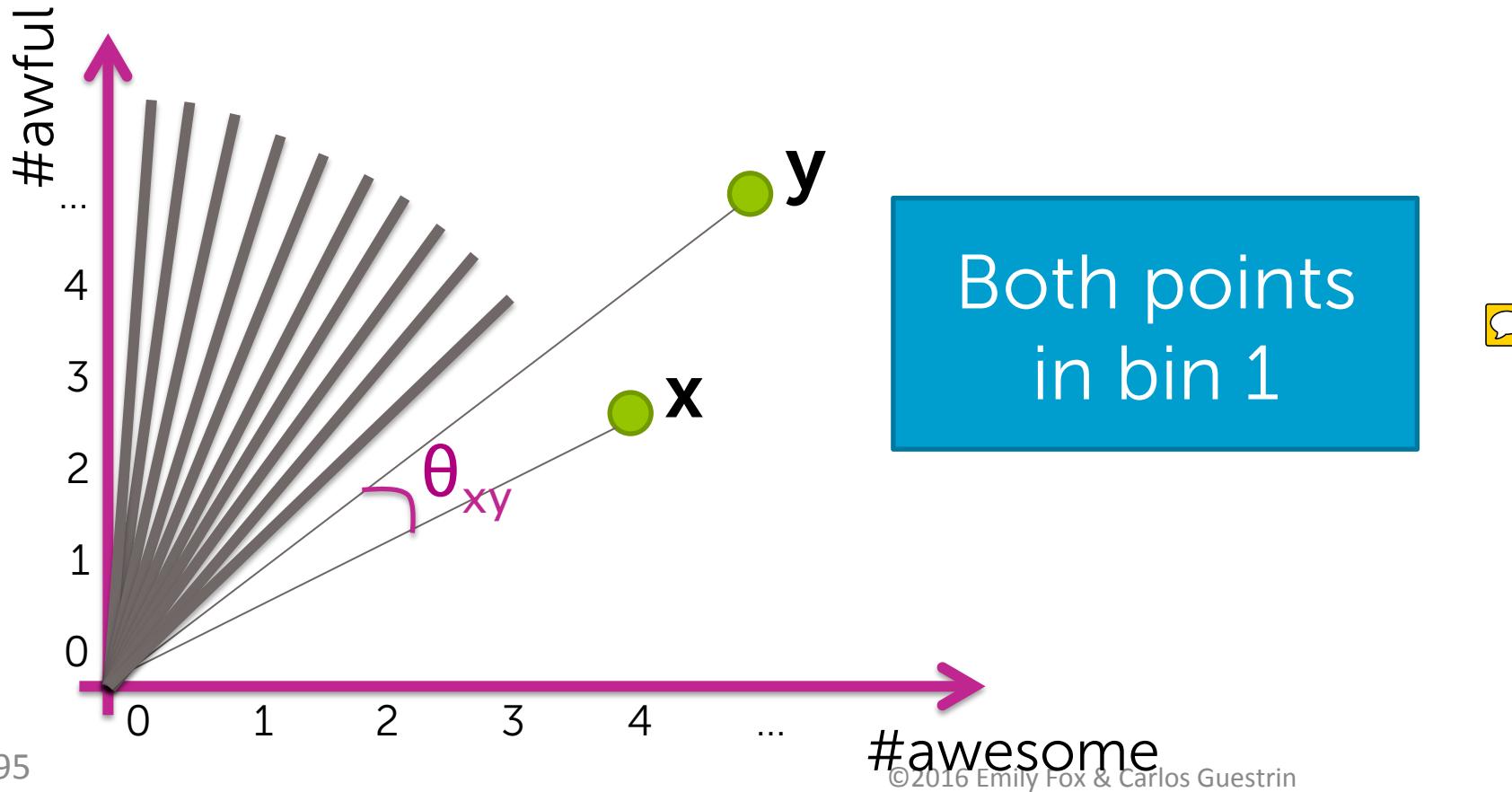


Goal: If x, y are close (according to cosine similarity),
want binned values to be the same.



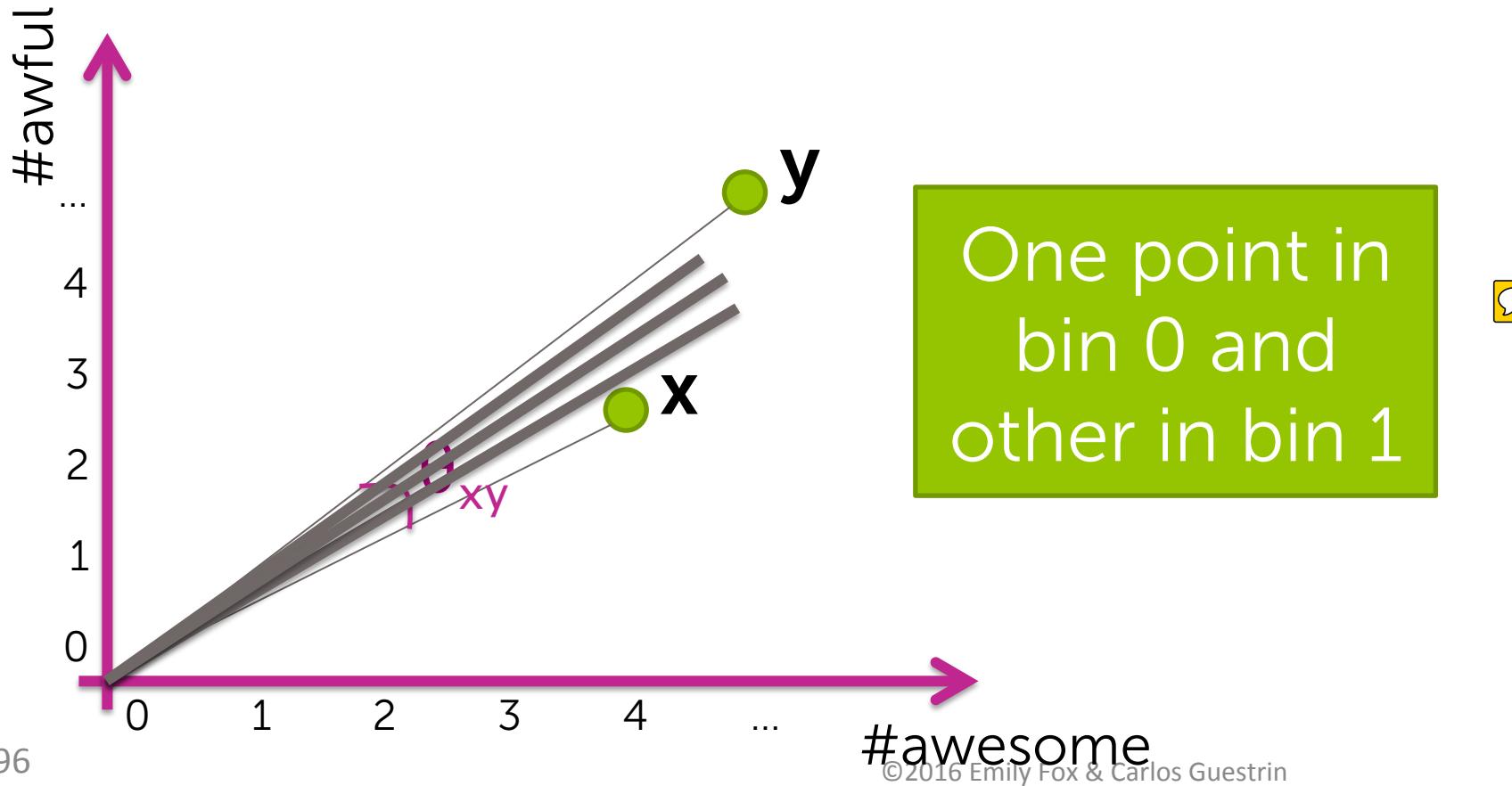
How bad can a random line be?

Goal: If x, y are close (according to cosine similarity),
want binned values to be the same.



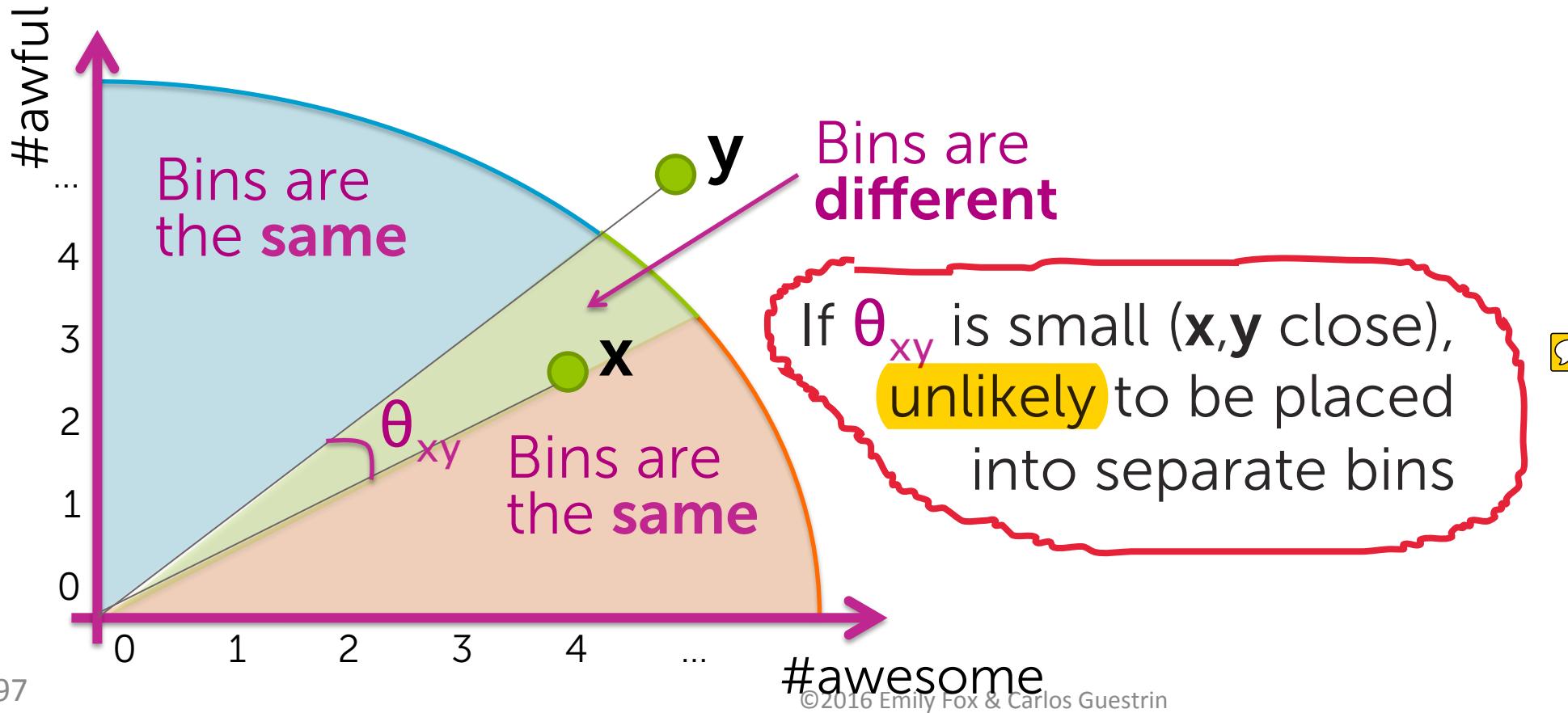
How bad can a random line be?

Goal: If x, y are close (according to cosine similarity),
want binned values to be the same.



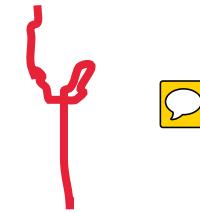
How bad can a random line be?

Goal: If x, y are close (according to cosine similarity), want binned values to be the same.



Three potential issues with simple approach

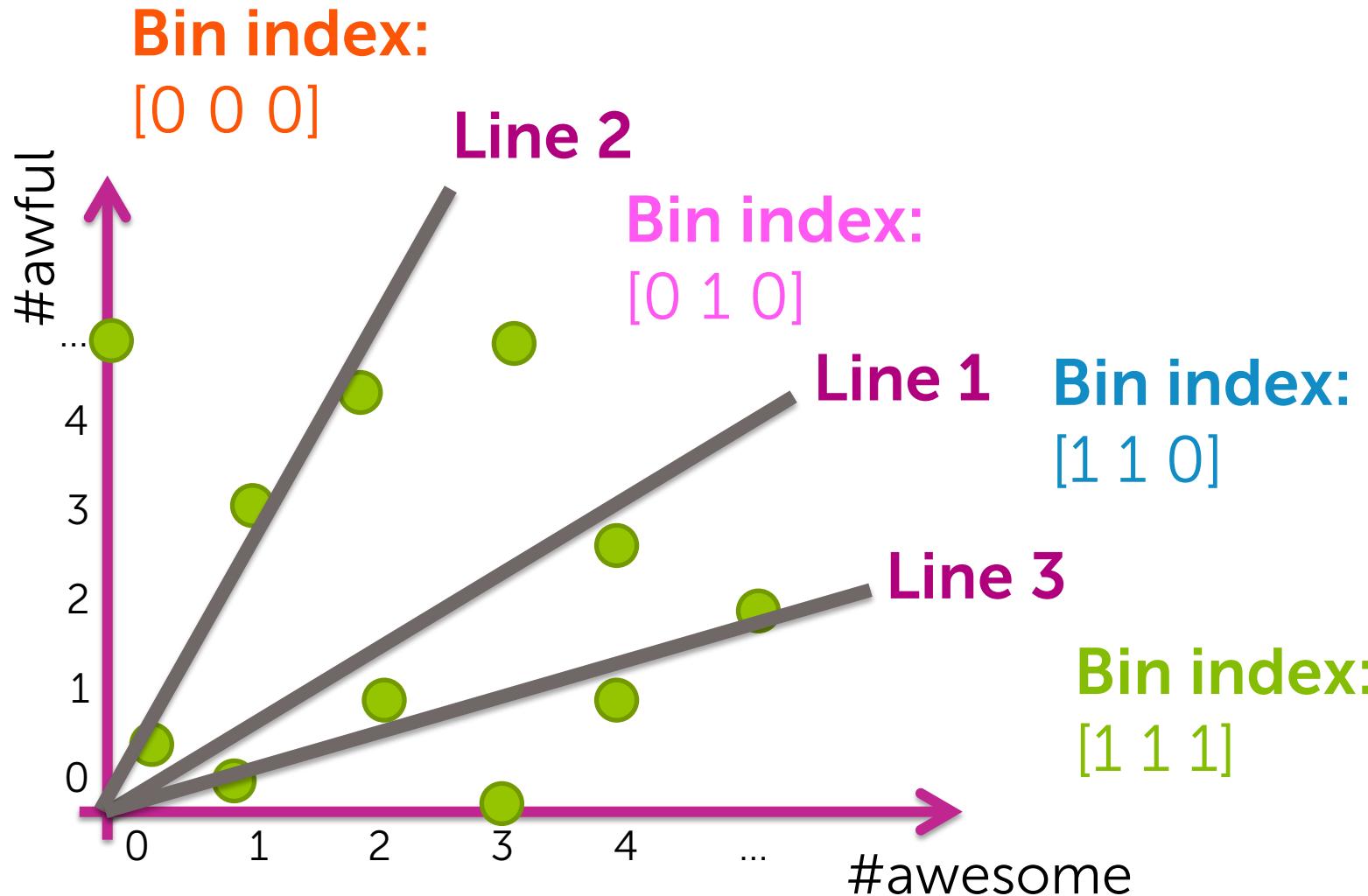
1. Challenging to find good line
2. Poor quality solution:
 - Points close together get split into separate bins
3. Large computational cost:
 - Bins might contain many points, so still searching over large set for each NN query



Bin	0	1
List containing indices of datapoints:	{1,2,4,7,...}	{3,5,6,8,...}

Improving efficiency:
Reducing # points examined per query

Reducing search cost through more bins



Using score for NN search

2D Data	Sign (Score ₁)	Bin 1 index	Sign (Score ₂)	Bin 2 index	Sign (Score ₃)	Bin 3 index
$\mathbf{x}_1 = [0, 5]$	-1	0	-1	0	-1	0
$\mathbf{x}_2 = [1, 3]$	-1	0	-1	0	-1	0
$\mathbf{x}_3 = [3, 0]$	1	1	1	1	1	1
...

Bin	[0 0 0] = 0	[0 0 1] = 1	[0 1 0] = 2	[0 1 1] = 3	[1 0 0] = 4	[1 0 1] = 5	[1 1 0] = 6	[1 1 1] = 7
Data indices:	{1,2}	--	{4,8,11}	--	--	--	{7,9,10}	{3,5,6}

search for NN
amongst this set



Improving search quality by searching neighboring bins

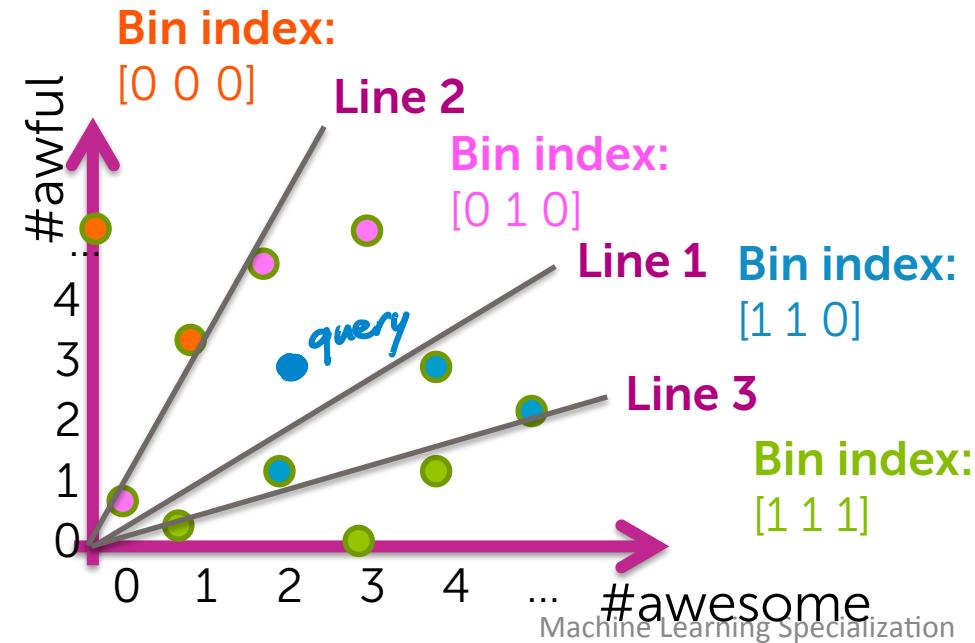
Bin	[0 0 0] = 0	[0 0 1] = 1	[0 1 0] = 2	[0 1 1] = 3	[1 0 0] = 4	[1 0 1] = 5	[1 1 0] = 6	[1 1 1] = 7
Data indices:	{1,2}	--	{4,8,11}	--	--	--	{7,9,10}	{3,5,6}



Query point here,
but is NN?

Not necessarily

Even worse than before...Each line can split pts.
Sacrificing accuracy for speed

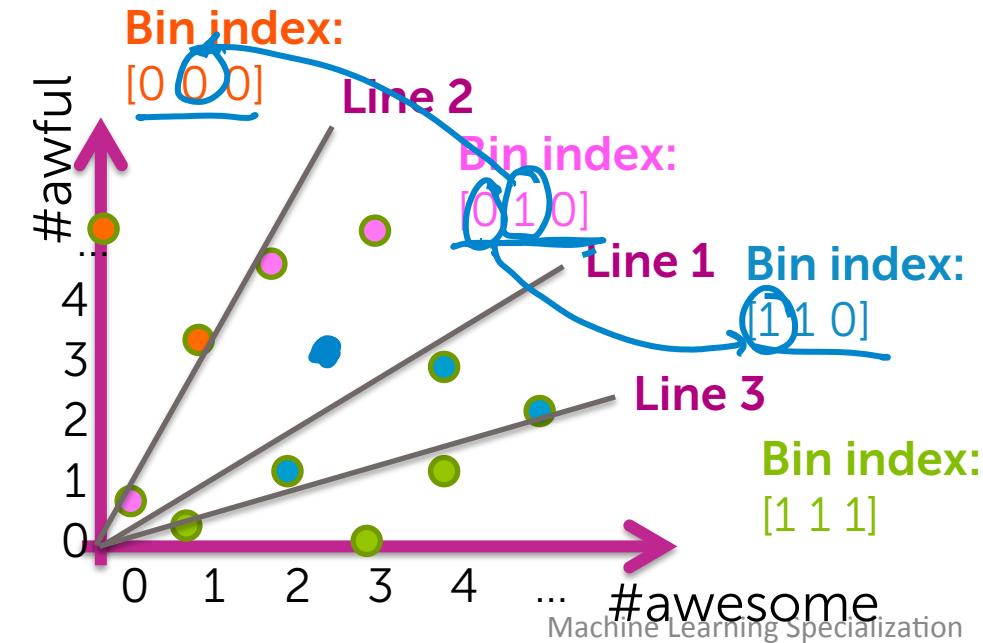


Improving search quality by searching neighboring bins



Bin	[0 0 0] = 0	[0 0 1]	[0 1 0] = 2	[0 1 1]	[1 0 0]	[1 0 1]	[1 1 0] = 6	[1 1 1]
Data indices:	{1,2}	--	{4,8,11}	--	--	--	{7,9,10}	{3,5,6}

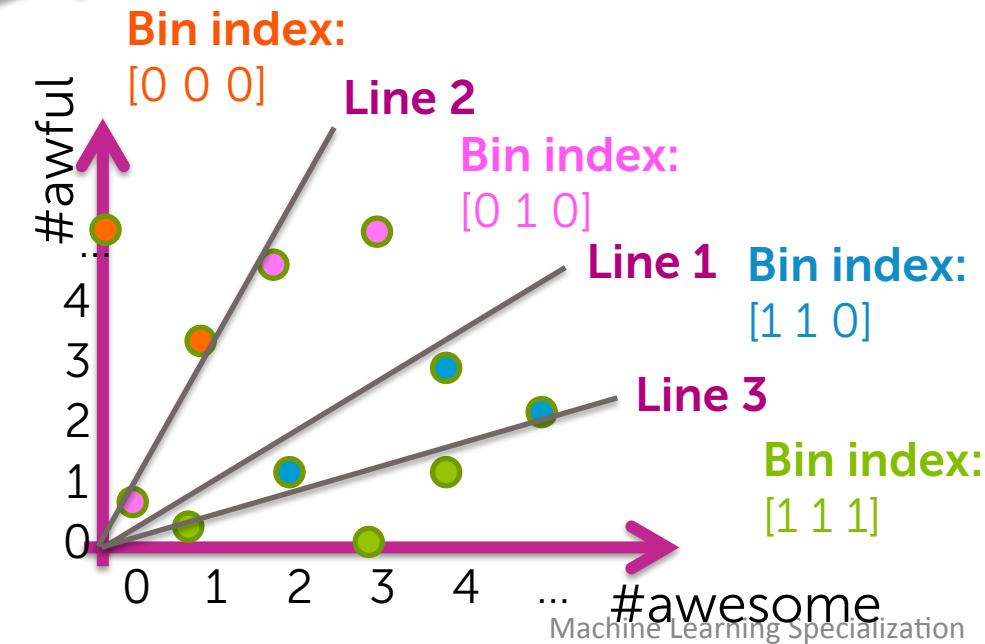
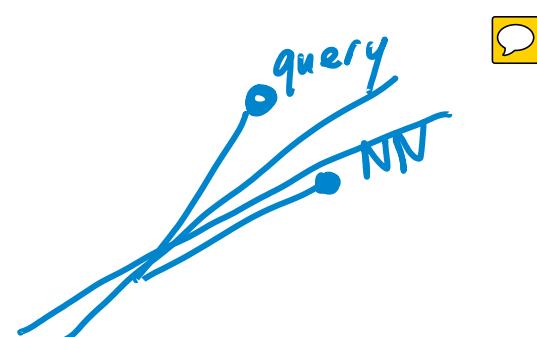
Next closest
bins
(flip 1 bit)



Improving search quality by searching neighboring bins

Bin	[0 0 0] = 0	[0 0 1] = 1	[0 1 0] = 2	[0 1 1] = 3	[1 0 0] = 4	[1 0 1] = 5	[1 1 0] = 6	[1 1 1] = 7
Data indices:	{1,2}	--	{4,8,11}	--	--	--	{7,9,10}	{3,5,6}

Further bin
(flip 2 bits)



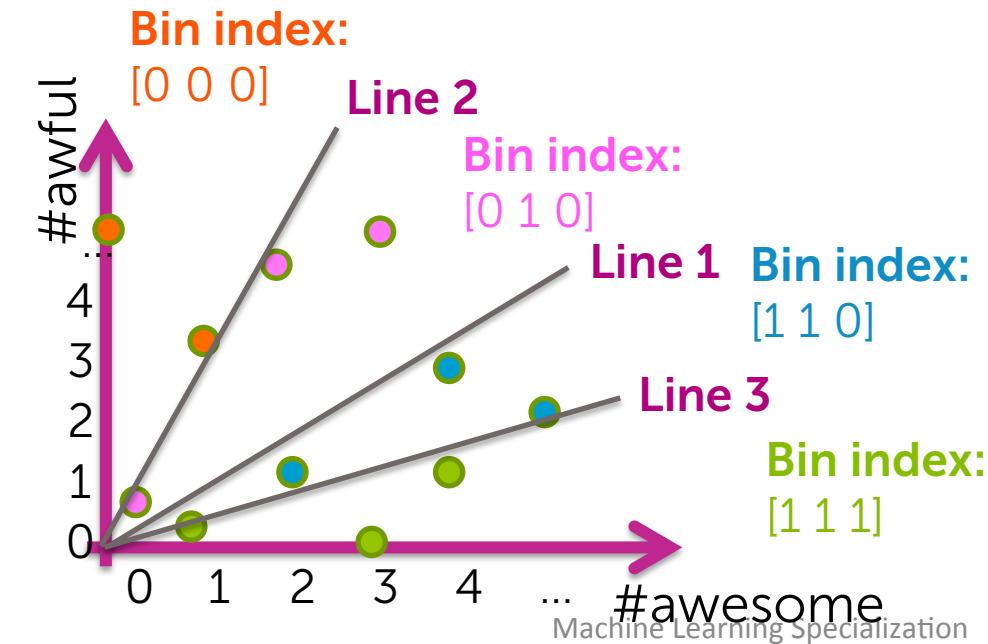
Improving search quality by searching neighboring bins

Bin	[0 0 0] = 0	[0 0 1]	[0 1 0] = 2	[0 1 1]	[1 0 0]	[1 0 1]	[1 1 0]	[1 1 1] = 7
Data indices:	{1,2}	--	{4,8,11}	--	--	--	{7,9,10}	{3,5,6}

Quality of retrieved NN can only improve with searching more bins

Algorithm:

Continue searching until computational budget is reached or quality of NN good enough



LSH recap



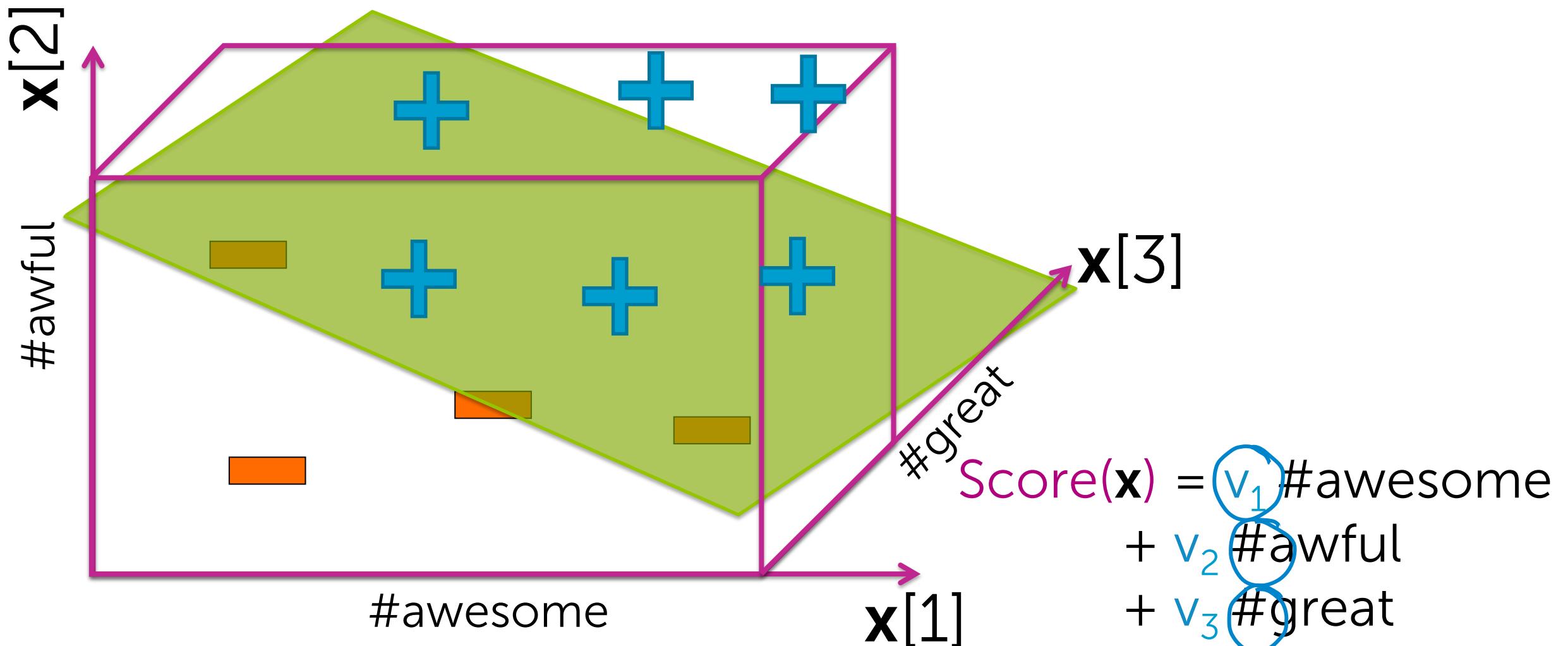
kd-tree competitor
data structure

- Draw h random lines
 - Compute “score” for each point under each line and translate to binary index
 - Use h -bit binary vector per data point as bin index
 - Create hash table
-
- For each query point \mathbf{x} , search $\text{bin}(\mathbf{x})$, then neighboring bins until time limit

Moving to higher dimensions d



Draw random planes



Cost of binning points in d-dim

$$\text{Score}_i(\mathbf{x}) = v_1^{(i)} \# \text{awesome} + v_2^{(i)} \# \text{awful} + v_3^{(i)} \# \text{great}$$

↑
ith hyperplane

Per data point,
need **d** multiplies
to determine bin
index per plane

In high-dim, (and some applications)
this is often a sparse mult.

One-time cost offset if many
queries of fixed dataset



Using multiple tables for even greater efficiency in NN search

OPTIONAL



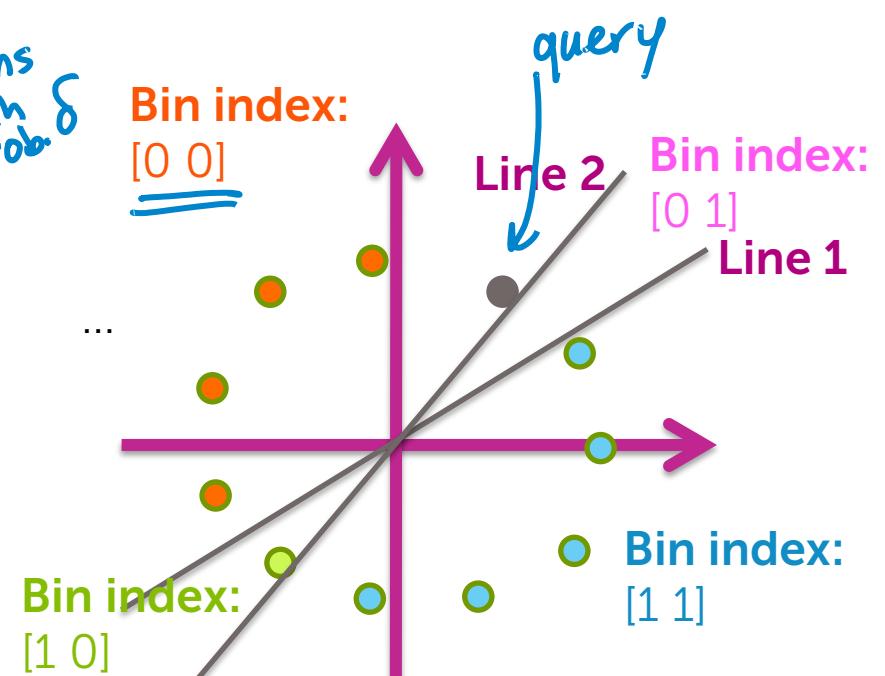
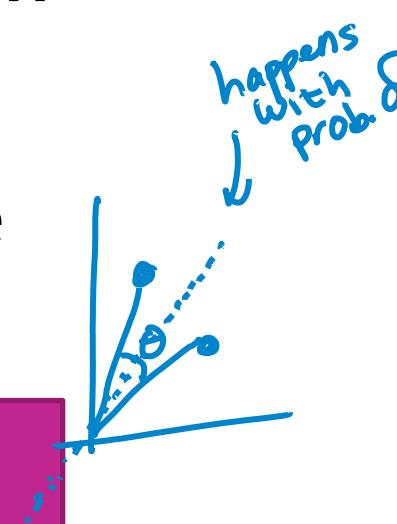
If I throw down 2 lines...

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3

For simplicity, assume we **search bins 1 bit off** from query

Let δ be the probability of a line falling between points θ apart

Search 3 bins and do not find NN with probability δ^2

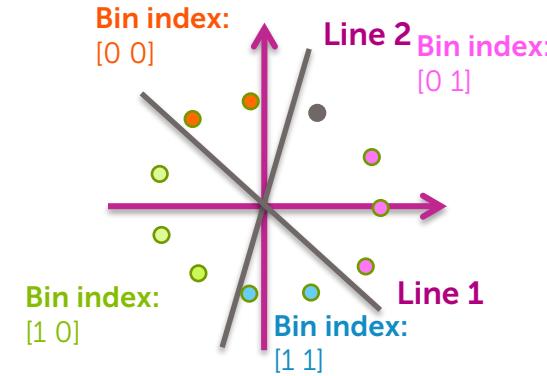
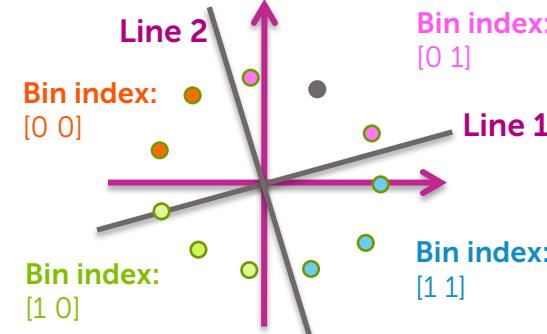
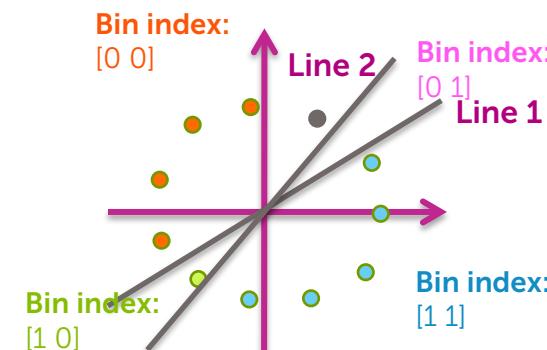


What if I repeat the 2-line binning?

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3

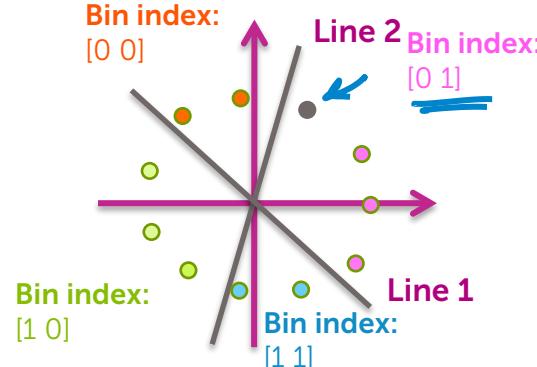
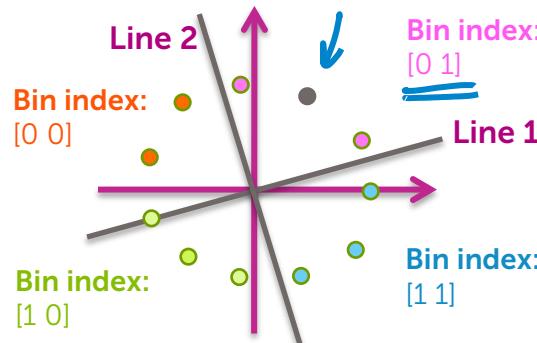
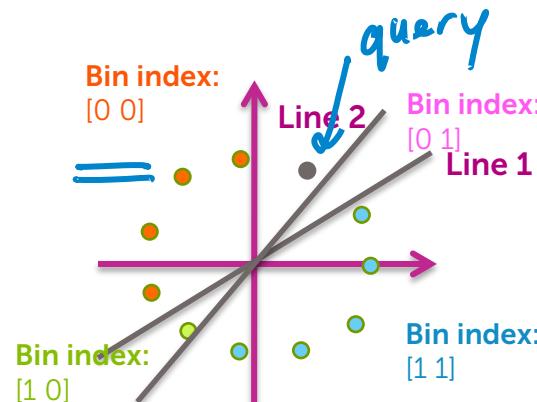


What if I repeat the 2-line binning?

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3



Now, search
only query bin
per table

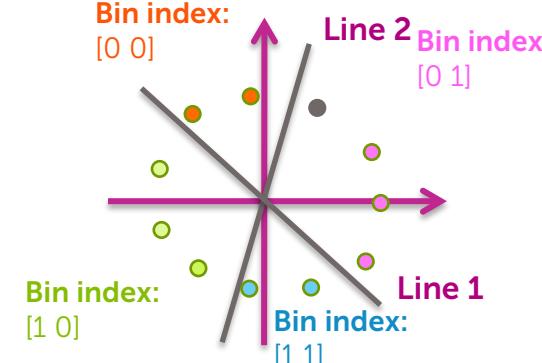
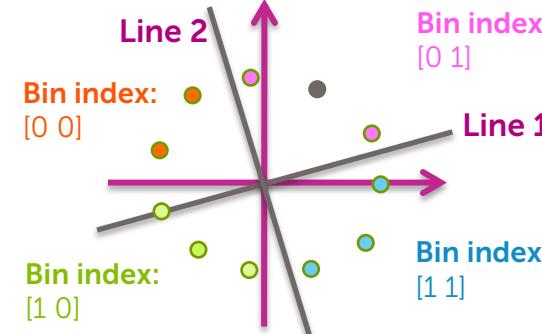
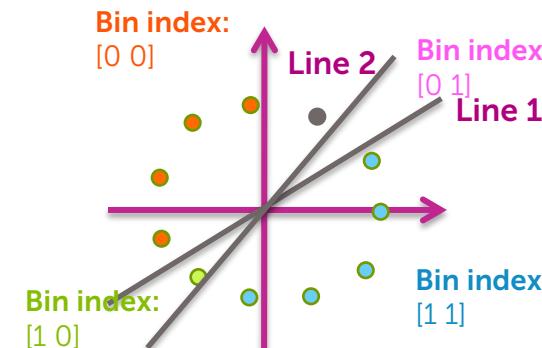
Still searching 3
bins, but what is
chance of not
finding NN?

What if I repeat the 2-line binning?

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3



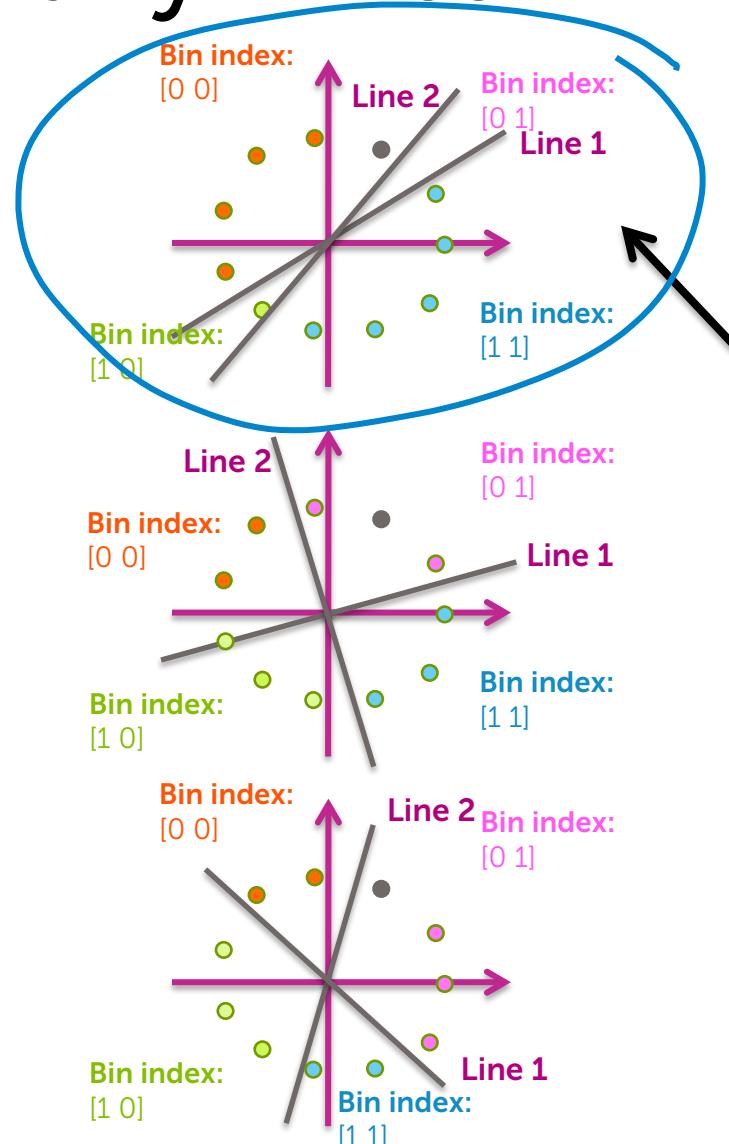
What is chance
that query pt
and NN are split
in **all tables**?

Probability of splitting neighboring points many times

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3



Probability NN is
in different bin:

$$\begin{aligned} \text{Prob} &= 1 - \text{Pr}(\text{same bin}) \\ &= 1 - (1 - \delta)^2 \\ &= 2\delta - \delta^2 \end{aligned}$$

$1 - \delta$ = prob. that 1 line
does not split
query + NN

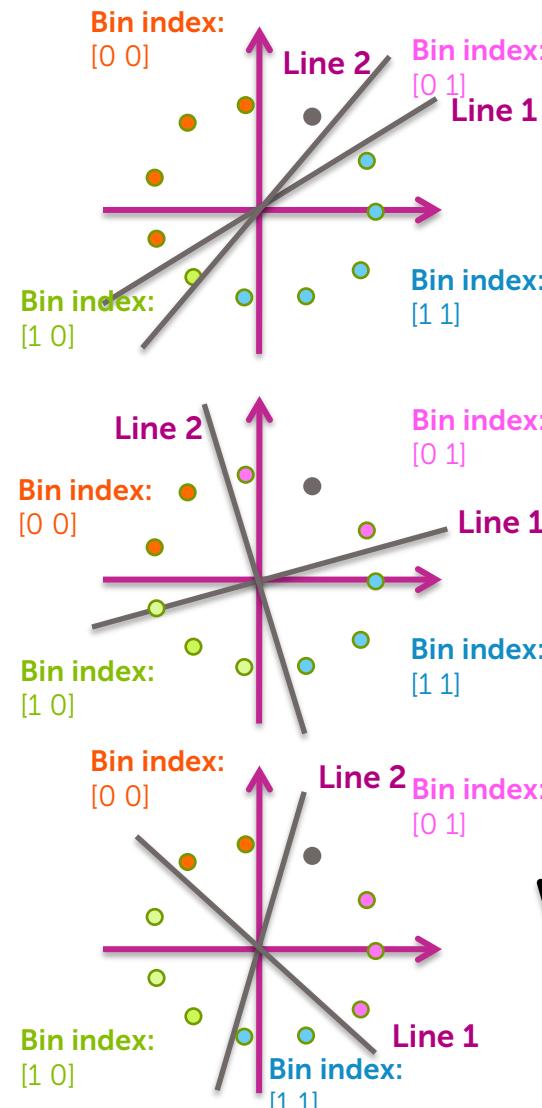
$(1 - \delta)^2$ = prob. that 2 lines
don't split pts

Probability of splitting neighboring points many times

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3



Probability NN is
in different bin in
all 3 tables:

$$\text{Prob} = (2\delta - \delta^2)^3$$

of hash tables

Comparing approaches for 2-bit tables

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3

bins
searched

3

prob. of
no NN

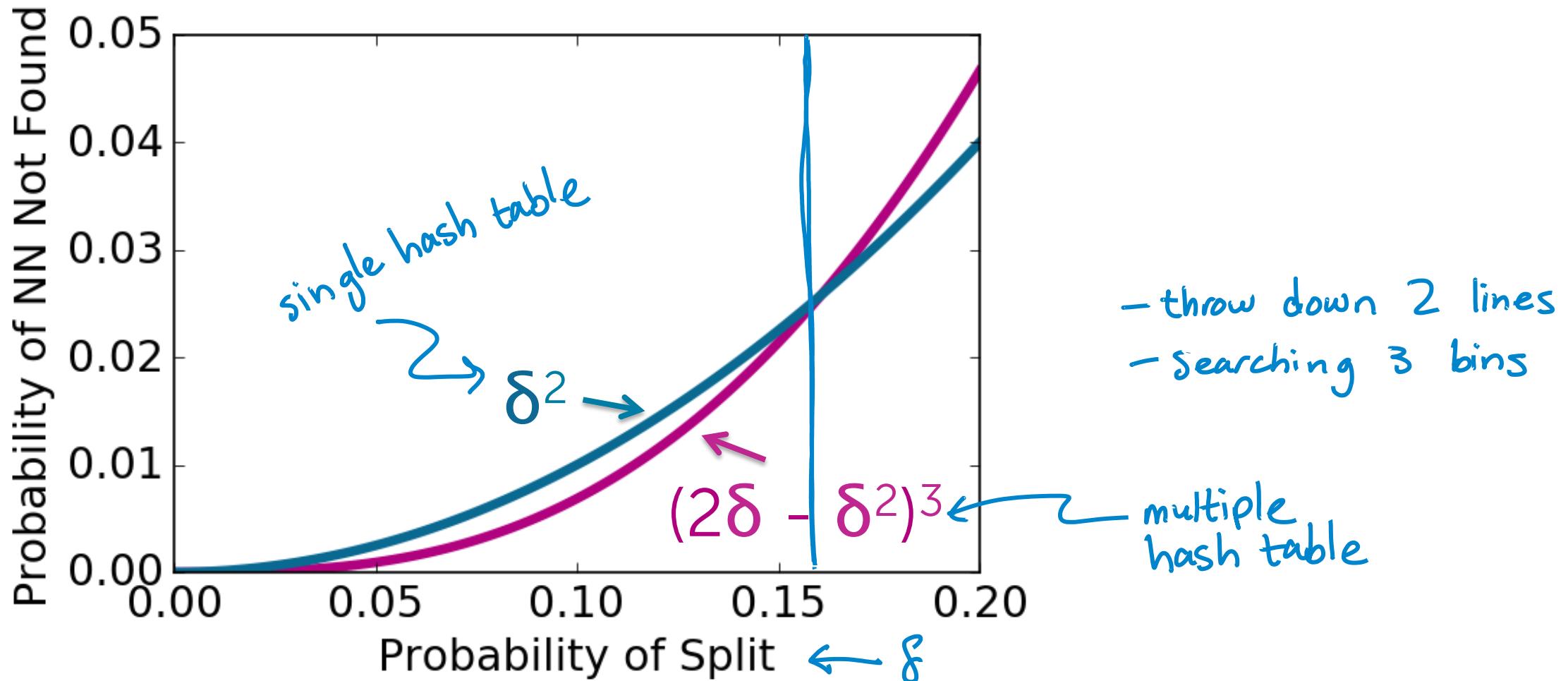
$$\underline{\delta^2}$$

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3

3

$$(2\underline{\delta} - \underline{\delta^2})^3$$

Comparing probabilities



If I throw down h lines...

Bin	[0 0 0] = 0	[0 0 1] = 1	[0 1 0] = 2	[0 1 1] = 3	[1 0 0] = 4	[1 0 1] = 5	[1 1 0] = 6	[1 1 1] = 7
indices	L0	L1	L2	L3	L4	L5	L6	L7

Still assume we search bins 1 bit off from query

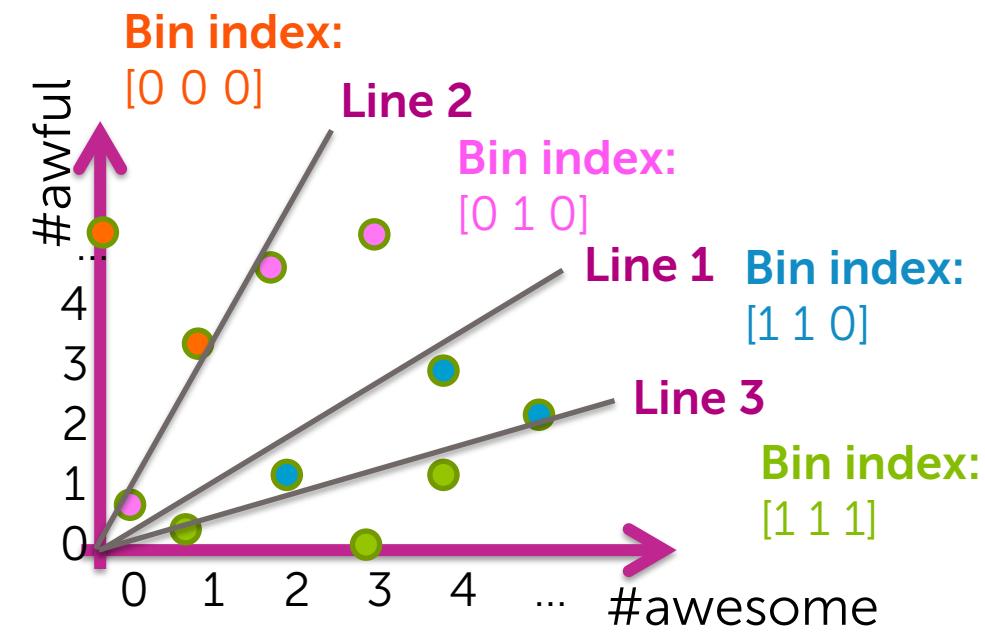
Prob. of being > 1 bit away

$$= 1 - \Pr(\text{same bin}) - \Pr(1 \text{ bin away})$$

$$= 1 - \Pr(\text{no split lines}) - \Pr(1 \text{ split line})$$

$$= 1 - (1 - \delta)^h - h\delta(1 - \delta)^{h-1}$$

↑ Prob. of no split
is $1 - \delta$
+ we have h
lines
→ Prob. none of
 h lines split is $(1 - \delta)^h$



If I throw down h lines...

Bin	[0 0 0] = 0	[0 0 1] = 1	[0 1 0] = 2	[0 1 1] = 3	[1 0 0] = 4	[1 0 1] = 5	[1 1 0] = 6	[1 1 1] = 7
indices	L0	L1	L2	L3	L4	L5	L6	L7

Still assume we **search bins 1 bit off** from query

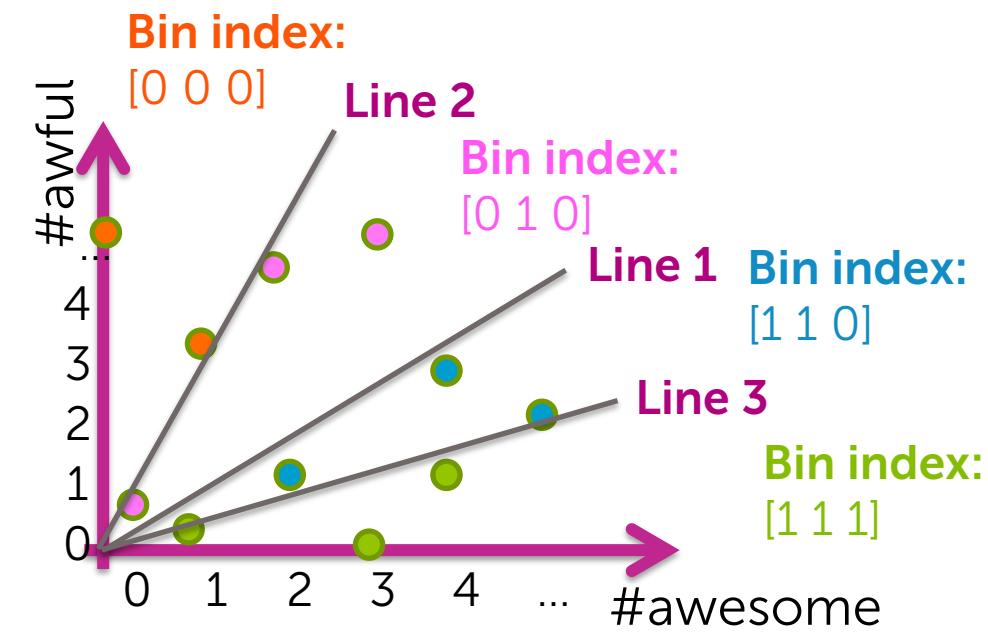
Prob. of being > 1 bit away

$$= 1 - \Pr(\text{same bin}) - \Pr(1 \text{ bin away})$$

$$= 1 - \Pr(\text{no split lines}) - \Pr(1 \text{ split line})$$

$$= 1 - (1 - \delta)^h - h\delta(1 - \delta)^{h-1}$$

Search $h+1$ bins and do not find NN
with probability $1 - (1 - \delta)^h - h\delta(1 - \delta)^{h-1}$



Probability of splitting neighboring points many times

Bin	[0 0 0] = 0	[0 0 1] = 1	[0 1 0] = 2	[0 1 1] = 3	[1 0 0] = 4	[1 0 1] = 5	[1 1 0] = 6	[1 1 1] = 7
indices	L0	L1	L2	L3	L4	L5	L6	L7

Bin	[0 0 0] = 0	[0 0 1] = 1	[0 1 0] = 2	[0 1 1] = 3	[1 0 0] = 4	[1 0 1] = 5	[1 1 0] = 6	[1 1 1] = 7
indices	L0	L1	L2	L3	L4	L5	L6	L7

Bin	[0 0 0] = 0	[0 0 1] = 1	[0 1 0] = 2	[0 1 1] = 3	[1 0 0] = 4	[1 0 1] = 5	[1 1 0] = 6	[1 1 1] = 7
indices	L0	L1	L2	L3	L4	L5	L6	L7

Bin	[0 0 0] = 0	[0 0 1] = 1	[0 1 0] = 2	[0 1 1] = 3	[1 0 0] = 4	[1 0 1] = 5	[1 1 0] = 6	[1 1 1] = 7
indices	L0	L1	L2	L3	L4	L5	L6	L7

Probability NN is
in different bin in
all $h+1$ tables

$$= (1 - \Pr(\text{same bin}))^{h+1}$$

$$= (1 - \Pr(\text{no split line}))^{h+1}$$

$$= (1 - (1 - \delta)^h)^{h+1}$$

prob. of no
split from any
of h lines thrown
down

holds for all
hash tables

Comparing approaches for h-bit tables

throw down h lines

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3

bins
searched

$h+1$

prob. of
no NN

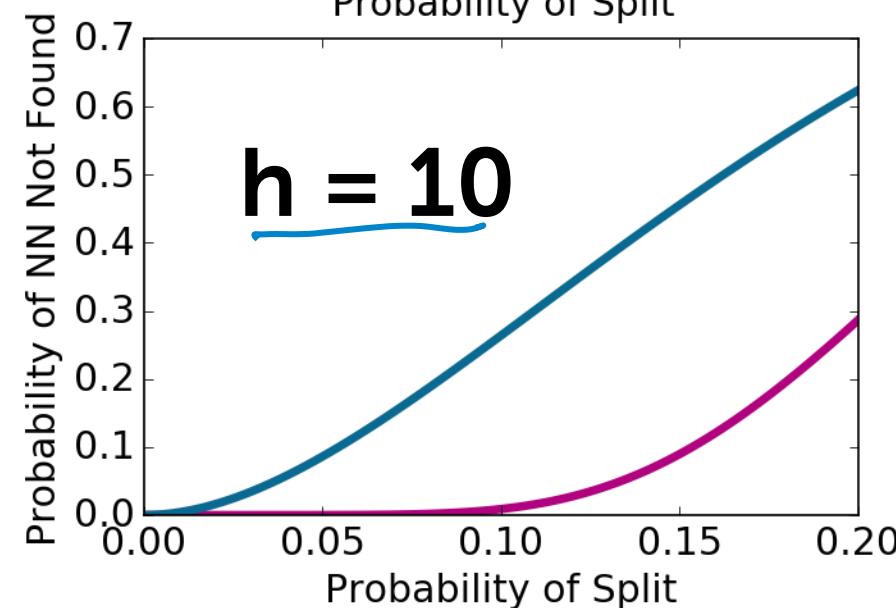
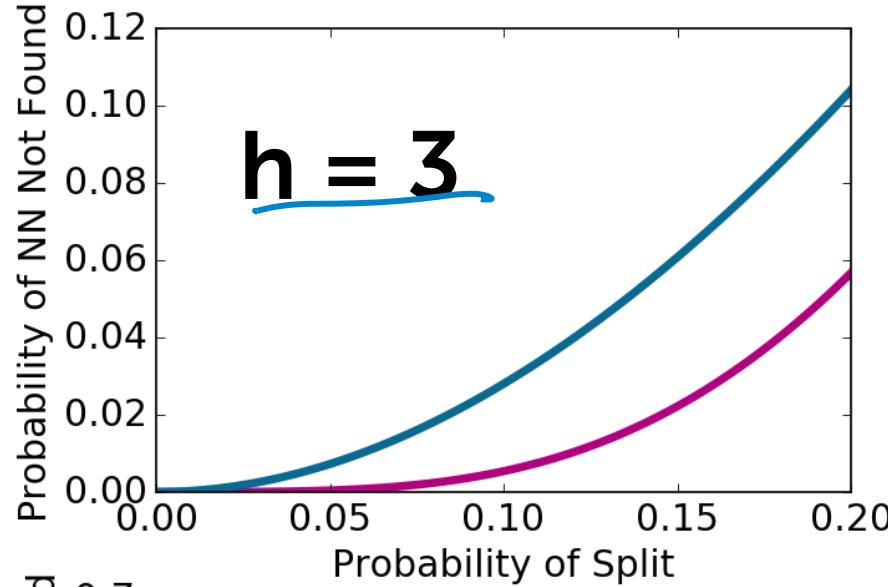
$$1 - (1 - \delta)^h - h\delta(1 - \delta)^{h-1}$$

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3

$h+1$

$$(1 - (1 - \delta)^h)^{h+1}$$

Comparing probabilities



one hash table

$$1 - (1 - \delta)^h - h\delta(1 - \delta)^{h-1}$$

$$(1 - (1 - \delta)^h)^{h+1}$$

multiple hash table

Fix #bits and increase depth

Bin	[0 0 0] = 0	[0 0 1] = 1	[0 1 0] = 2	[0 1 1] = 3	[1 0 0] = 4	[1 0 1] = 5	[1 1 0] = 6	[1 1 1] = 7
indices	L0	L1	L2	L3	L4	L5	L6	L7

Bin	[0 0 0] = 0	[0 0 1] = 1	[0 1 0] = 2	[0 1 1] = 3	[1 0 0] = 4	[1 0 1] = 5	[1 1 0] = 6	[1 1 1] = 7
indices	L0	L1	L2	L3	L4	L5	L6	L7

Bin	[0 0 0] = 0	[0 0 1] = 1	[0 1 0] = 2	[0 1 1] = 3	[1 0 0] = 4	[1 0 1] = 5	[1 1 0] = 6	[1 1 1] = 7
indices	L0	L1	L2	L3	L4	L5	L6	L7

Bin	[0 0 0] = 0	[0 0 1] = 1	[0 1 0] = 2	[0 1 1] = 3	[1 0 0] = 4	[1 0 1] = 5	[1 1 0] = 6	[1 1 1] = 7
indices	L0	L1	L2	L3	L4	L5	L6	L7

Bin	[0 0 0] = 0	[0 0 1] = 1	[0 1 0] = 2	[0 1 1] = 3	[1 0 0] = 4	[1 0 1] = 5	[1 1 0] = 6	[1 1 1] = 7
indices	L0	L1	L2	L3	L4	L5	L6	L7

Probability NN is
in different bin in
all tables falls off
exponentially fast

hash tables
↓

$$\begin{aligned} \text{Prob} &= (1 - \Pr(\text{same bin}))^{m \cdot b} \\ &= (1 - \Pr(\text{no split line}))^{m \cdot b} \\ &= (1 - (1 - \delta)^b)^{m \cdot b} \end{aligned}$$

of lines
(cbits)

Fix #bits and increase depth

Bin	[0 0 0] = 0	[0 0 1] = 1	[0 1 0] = 2	[0 1 1] = 3	[1 0 0] = 4	[1 0 1] = 5	[1 1 0] = 6	[1 1 1] = 7
indices	L0	L1	L2	L3	L4	L5	L6	L7

Typically higher probability of finding NN than searching m bins in 1 table

Summary of LSH approaches

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3

Cost of binning points is **lower**,
but likely need to search
more bins per query

Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
Bin	[0 0] = 0	[0 1] = 1	[1 0] = 2	[1 1] = 3
indices	L0	L1	L2	L3

Cost of binning points is **higher**,
but likely need to search
fewer bins per query

Summary for retrieval using nearest neighbors, KD-trees, and locality sensitive hashing

What you can do now...

- Implement nearest neighbor search for retrieval tasks
- Contrast document representations (e.g., raw word counts, tf-idf,...)
 - Emphasize important words using tf-idf
- Contrast methods for measuring similarity between two documents
 - Euclidean vs. weighted Euclidean
 - Cosine similarity vs. similarity via unnormalized inner product
- Describe complexity of brute force search
- Implement KD-trees for nearest neighbor search
- Implement LSH for approximate nearest neighbor search
- Compare pros and cons of KD-trees and LSH, and decide which is more appropriate for given dataset