INTRODUCTION TO DATA SCIENCE

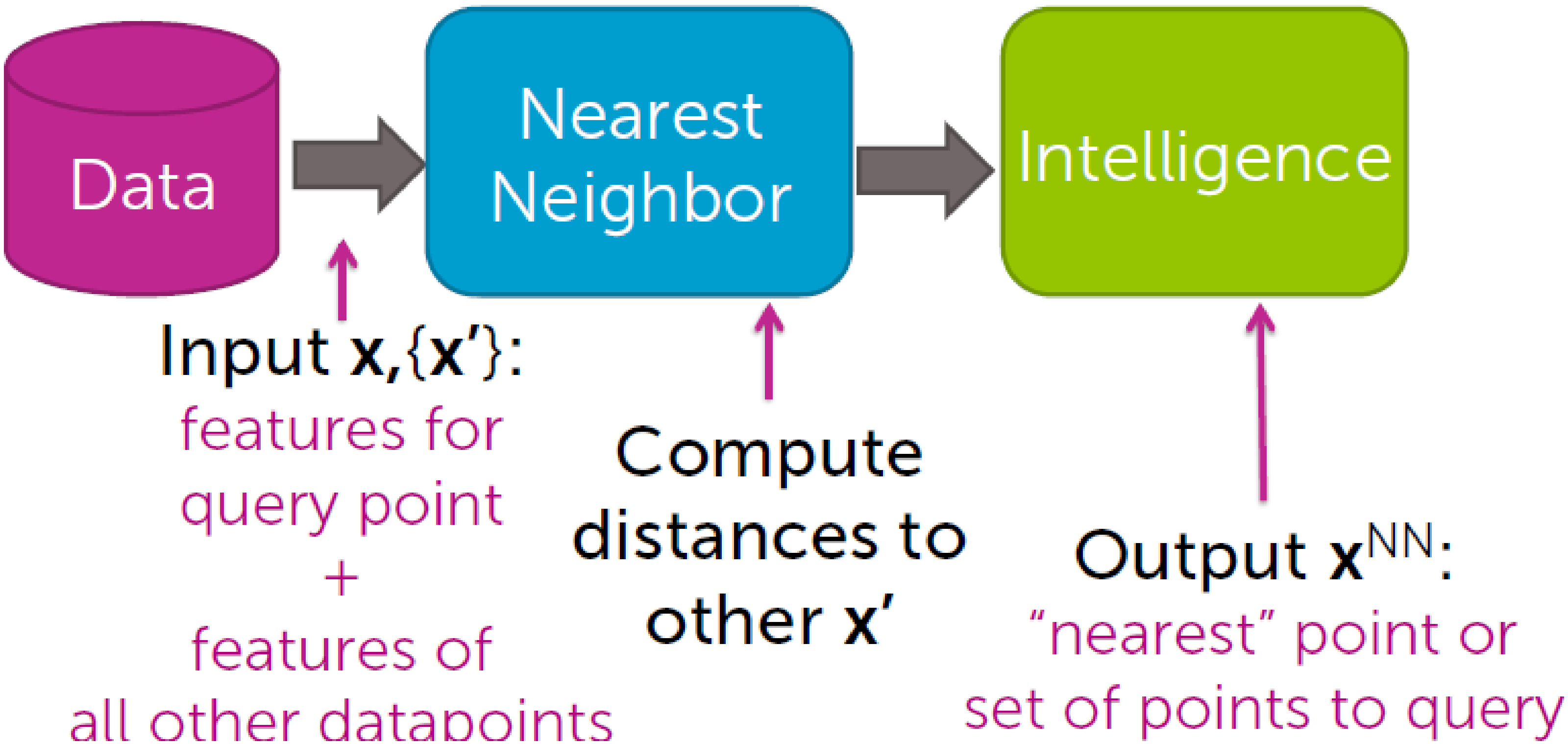
This lecture is
based on course by E. Fox and C. Guestrin, Univ of Washington

26/11, 3/12, 10/12/2019

WFAiS UJ, Informatyka Stosowana
I stopień studiów

What is retrieval?

Search for related items



What is retrieval?

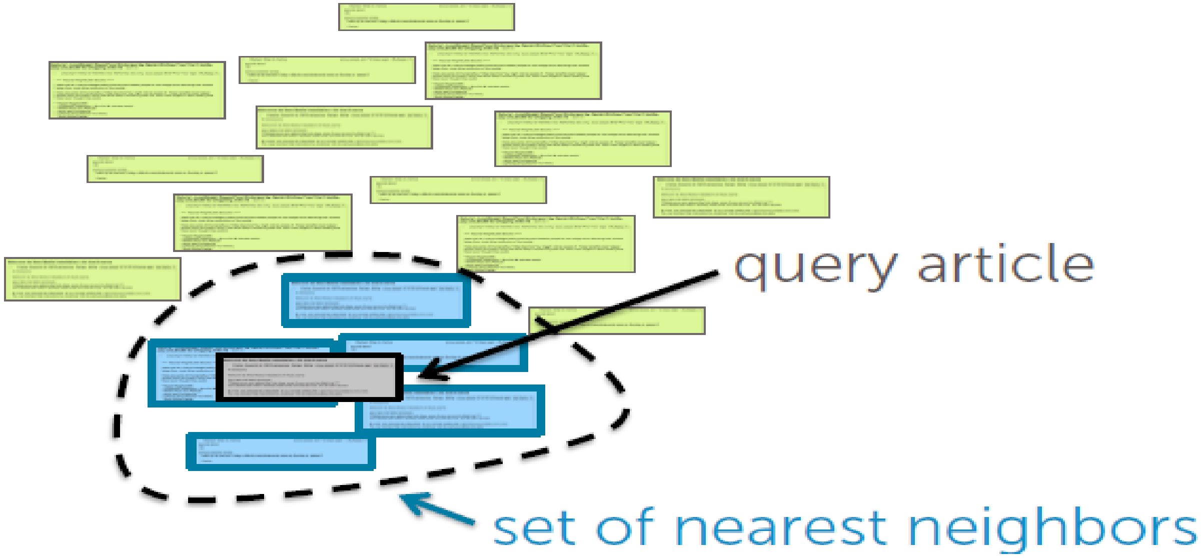
Retrieve "nearest neighbor" article

Space of all articles,
organized by similarity of text

What is retrieval?

Or set of nearest neighbors

Space of all articles,
organized by similarity of text

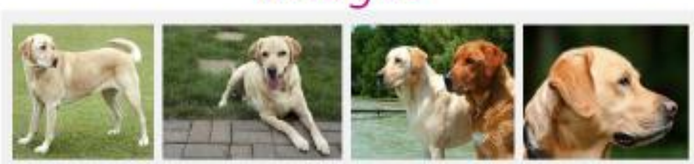


Retrieval applications

Just about everything…

Products

Images



Streaming content:

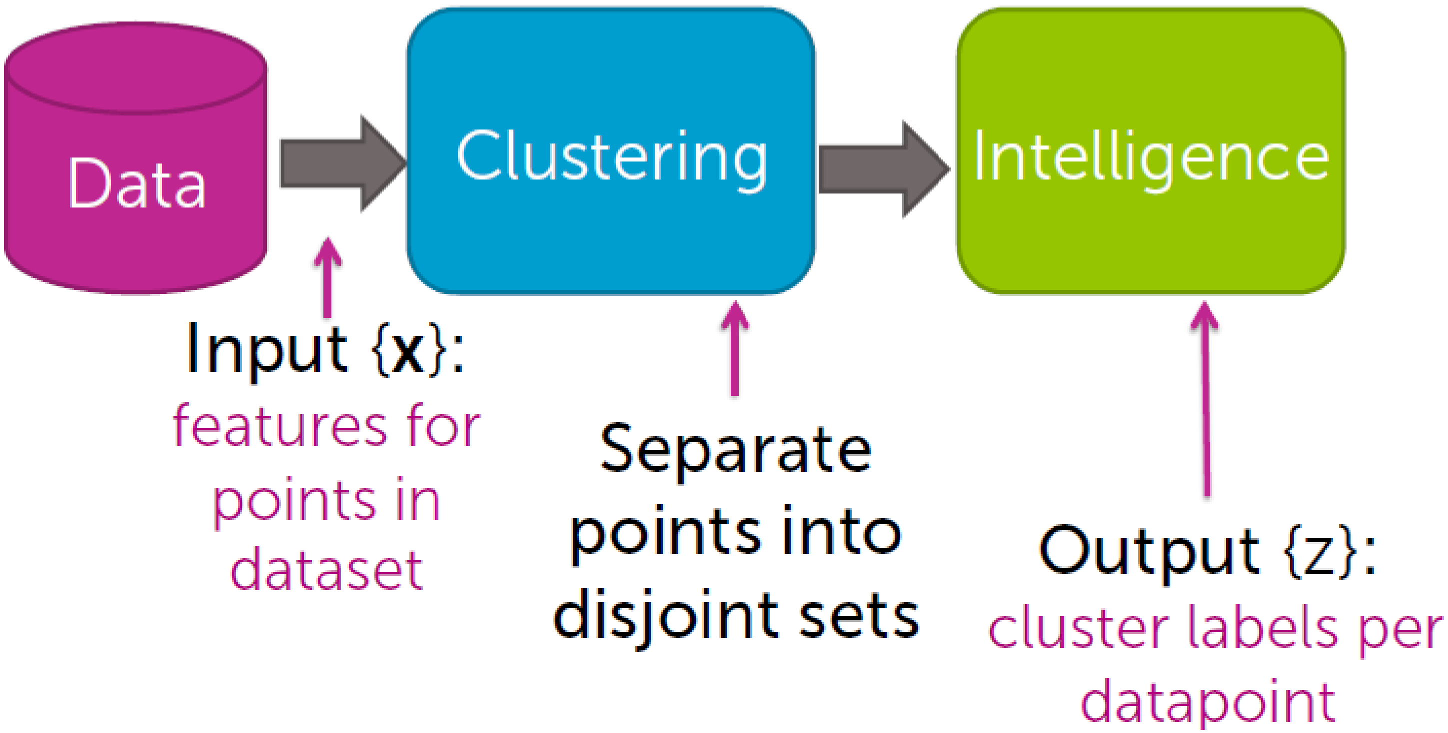
- Songs
- Movies
- TV shows
- …

News articles

Social networks
(people you might want to connect with)

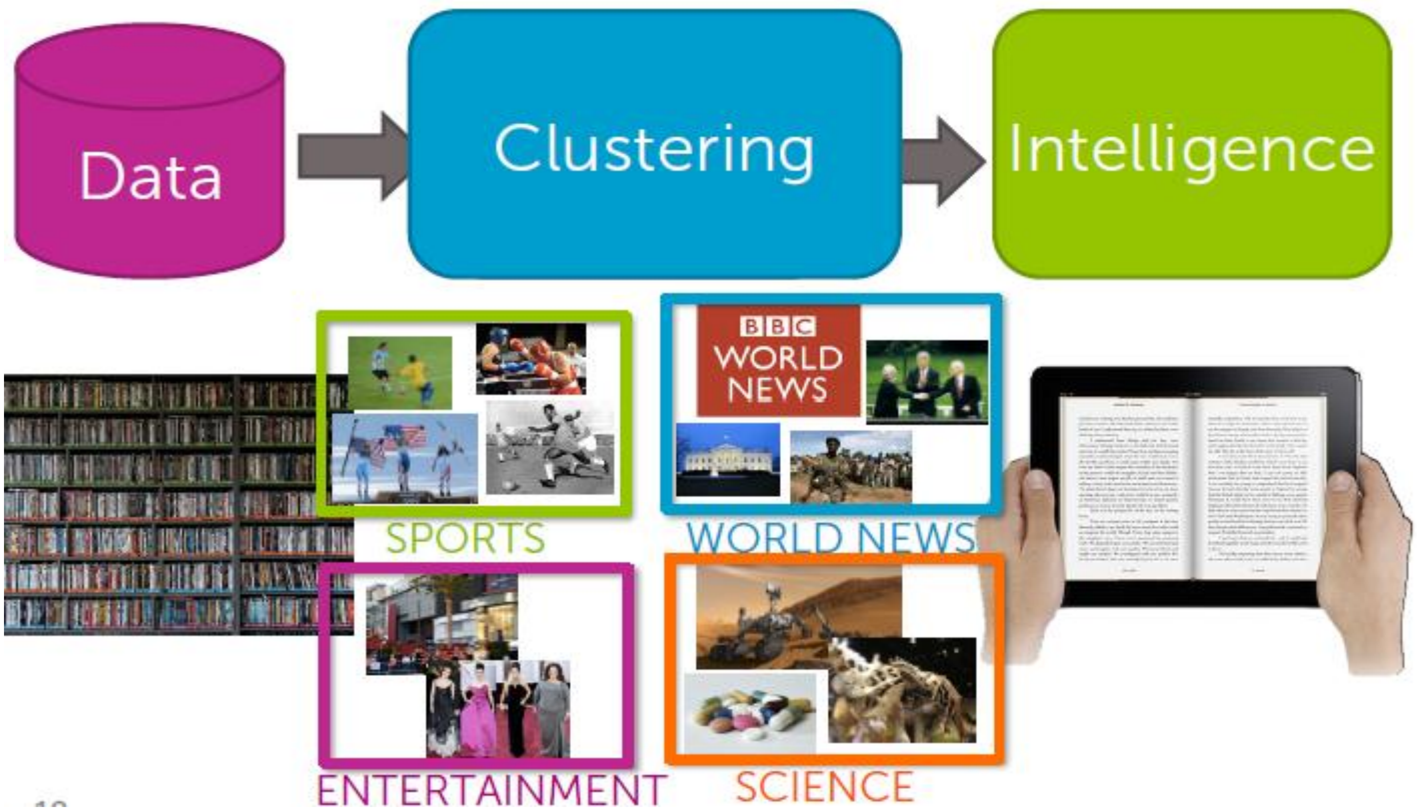
What is clustering?

Discover groups of similar inputs



Clustring applications

Clustering documents by "topic"

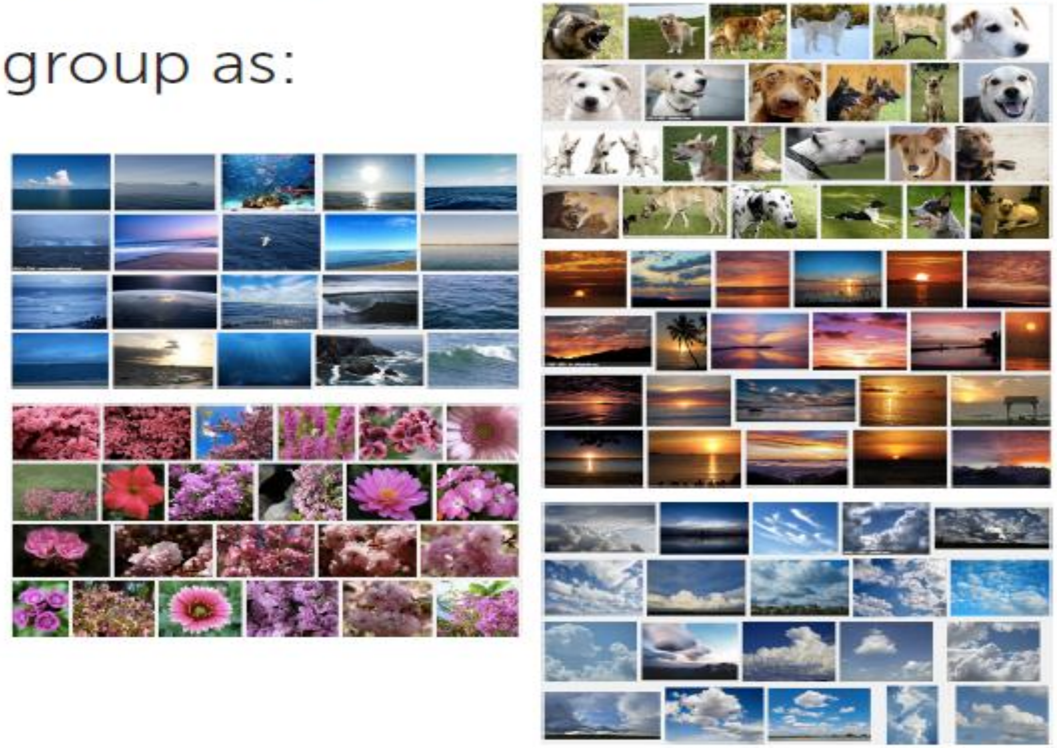


Clustering applications

Clustering images

For search, group as:

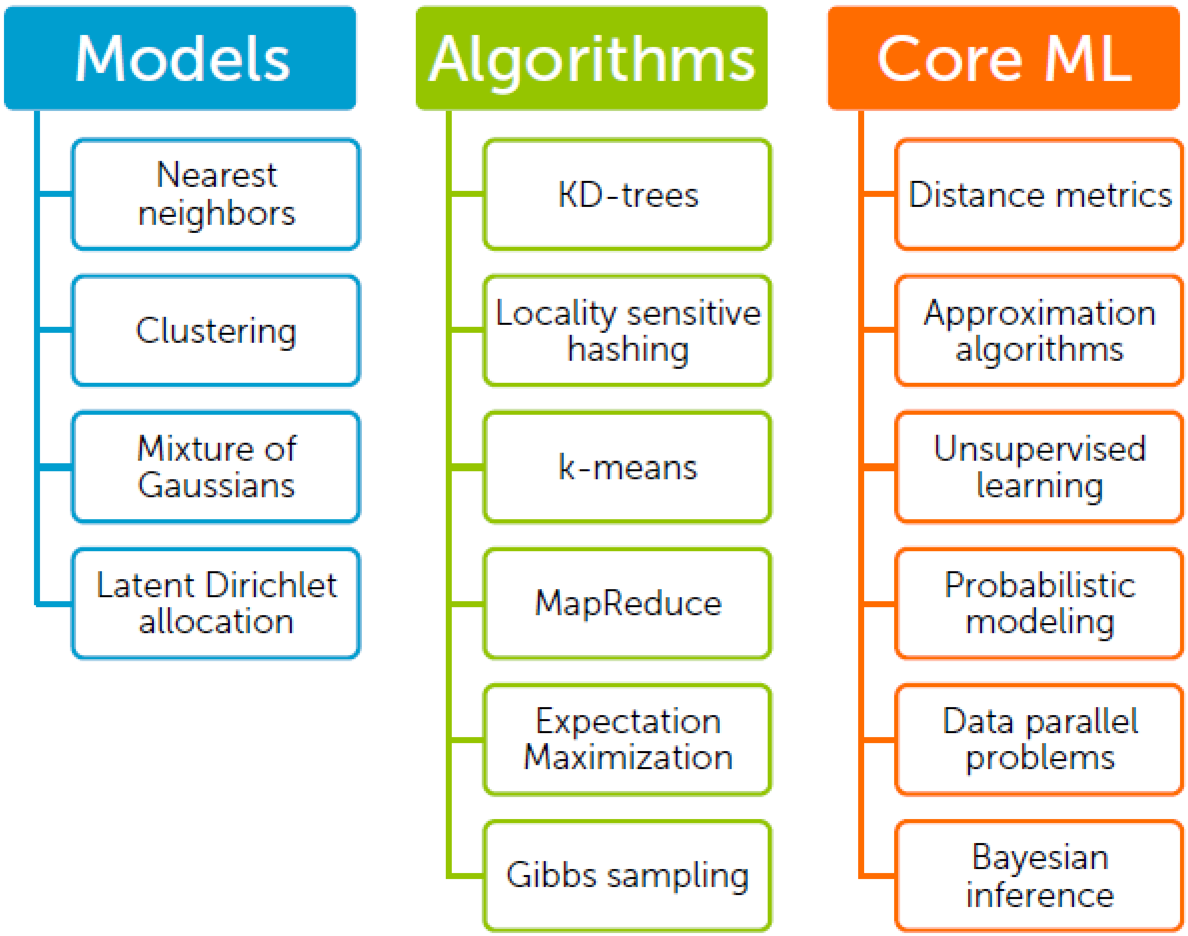
- Ocean
- Pink flower
- Dog
- Sunset
- Clouds
- ...



Impact of retrieval & clustering

- Foundational ideas
- Lots of information can be extracted using these tools (exploring user interests and interpretable structure relating groups of users based on observed behavior)

Overwiew of content

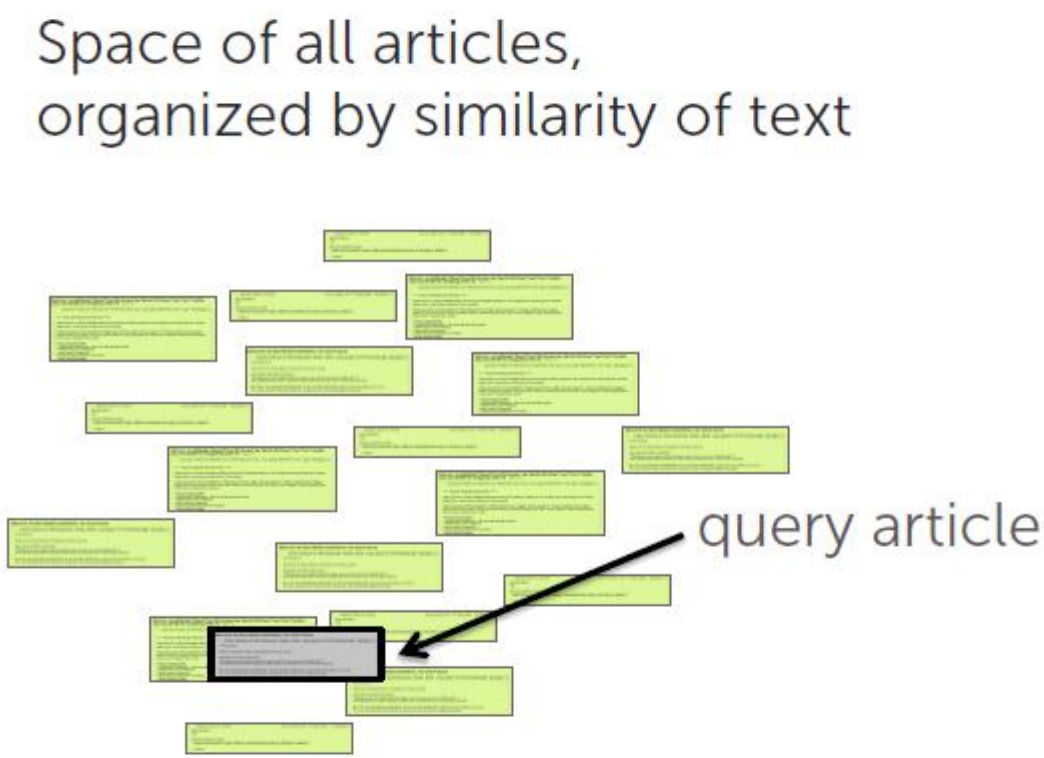


Retrieval

as

k-nearest neighbor search

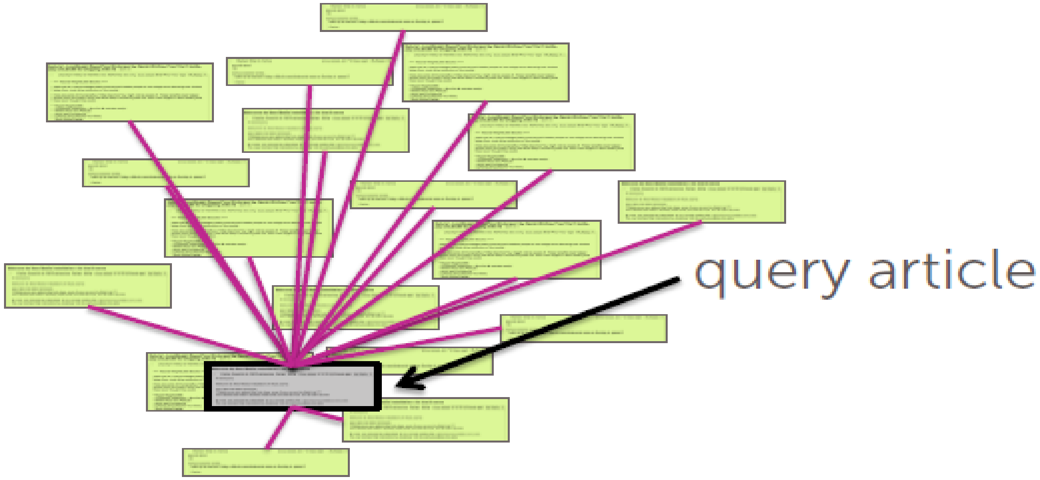
1-NN search for retrieval



1-NN search for retrieval

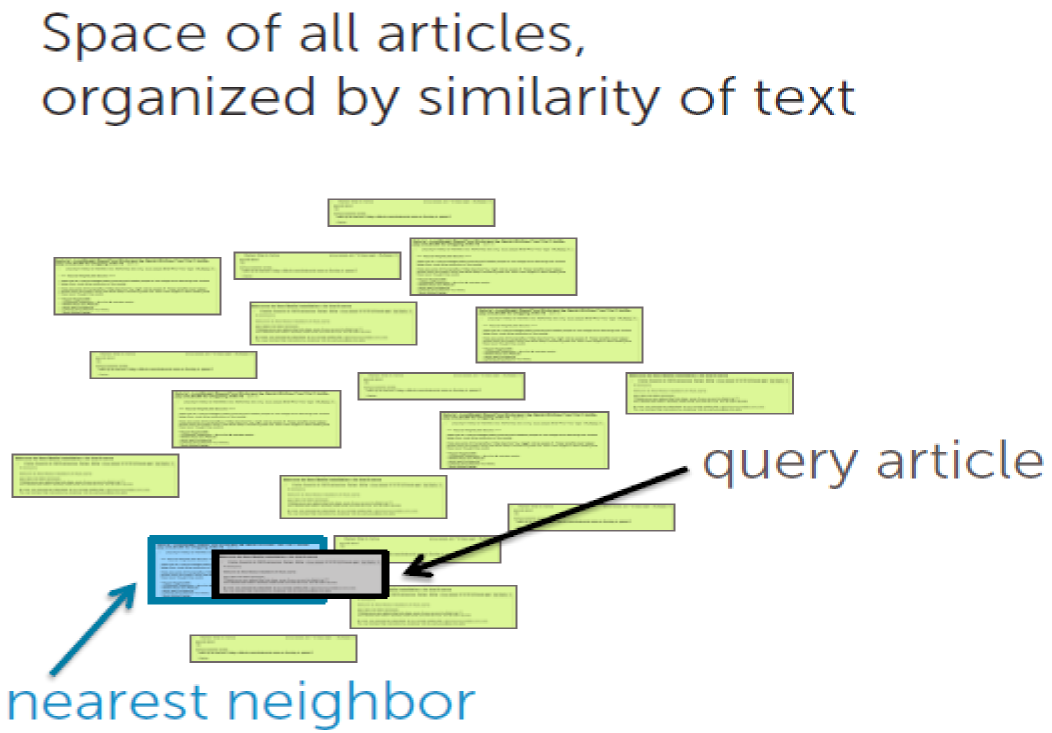
Compute distances to all docs

Space of all articles,
organized by similarity of text



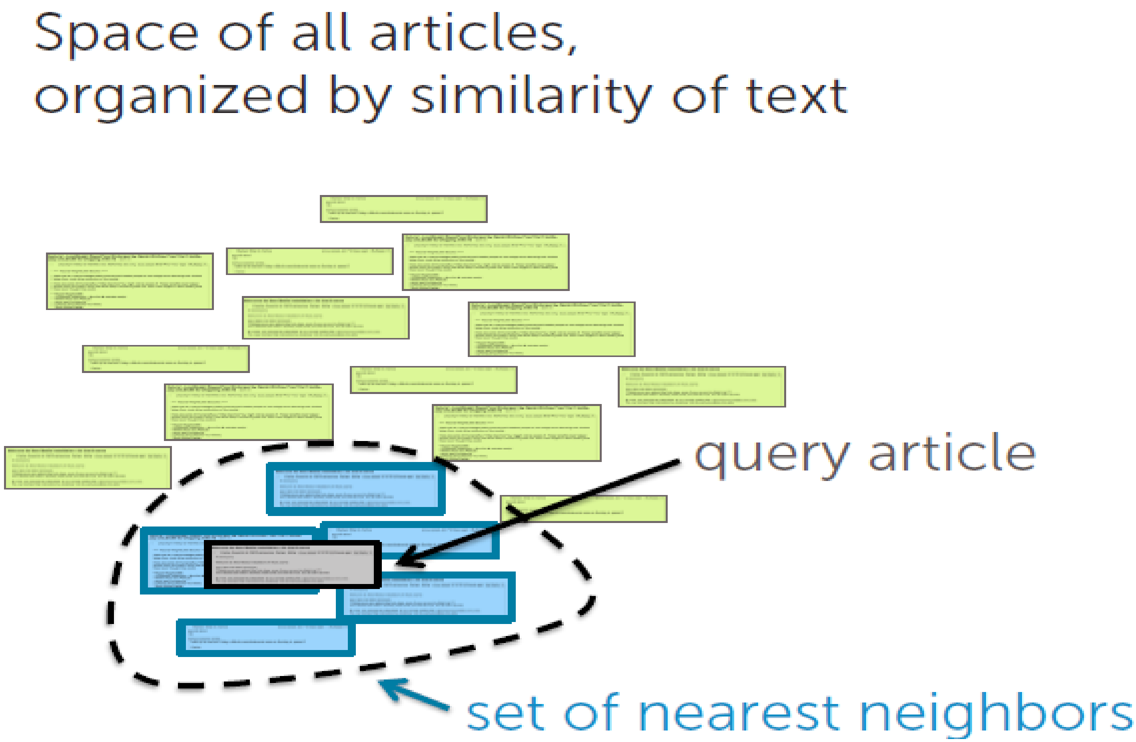
1-NN search for retrieval

Retrieve "nearest neighbor"



1-NN search for retrieval

Or set of nearest neighbors

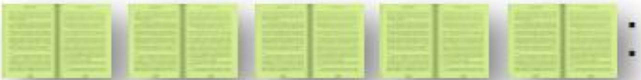


1-NN algorithm

1 – Nearest neighbor

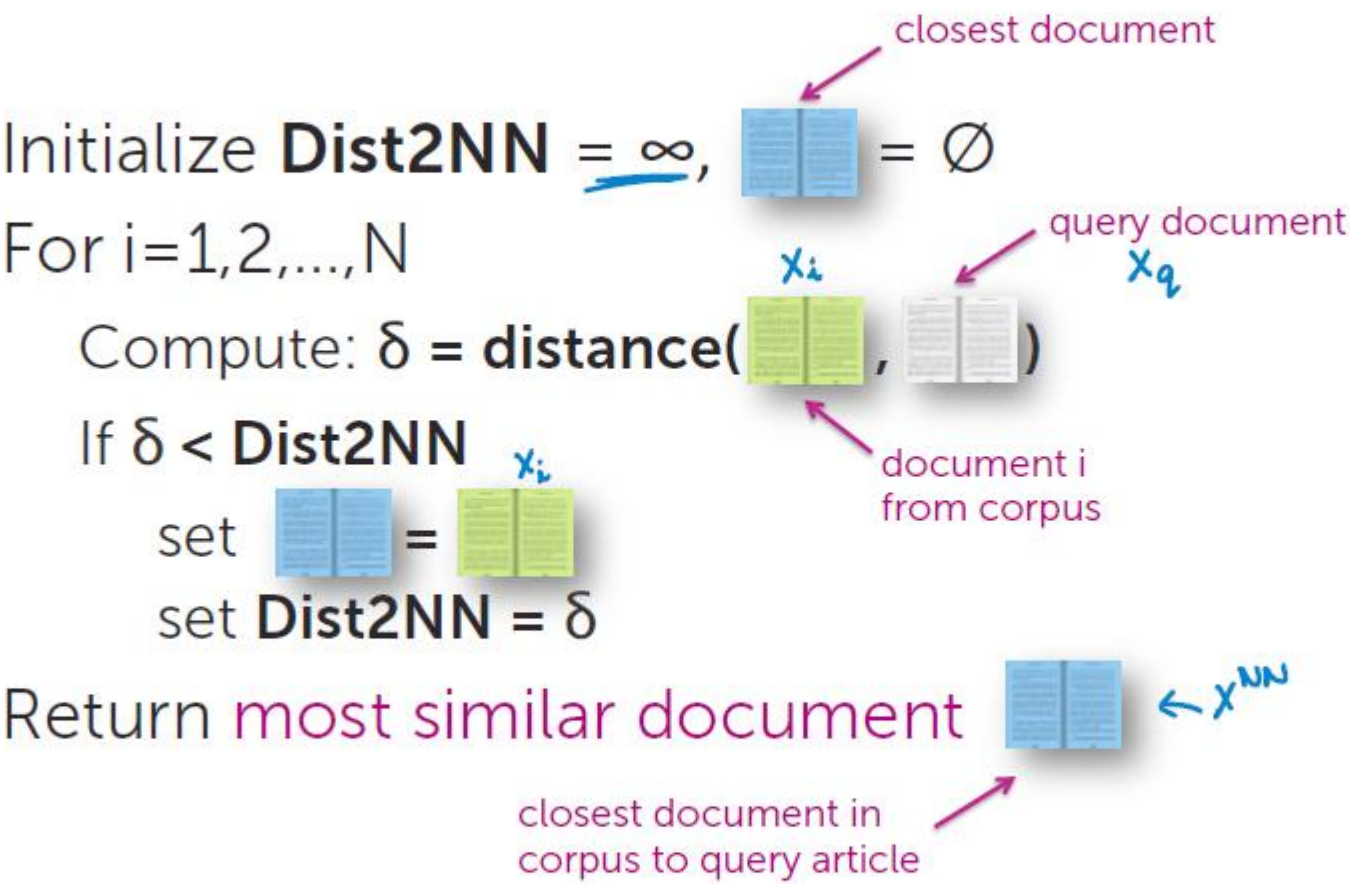
- **Input:** Query article  : $\underline{\mathbf{x}}_q$

  Corpus of documents (N docs)

   : $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$
- **Output:** *Most* similar article  $\leftarrow x^{NN}$

Formally: $x^{NN} = \min_{x_i} \text{distance}(x_q, x_i)$

1-NN algorithm



k-NN algorithm

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

  Corpus of documents

   : $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$
- **Output:** *List of k* similar articles

Formally:

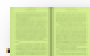
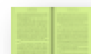
$X^{NN} = \{ X^{NN_1}, \ldots, X^{NN_k} \}$

For all x_i not in X^{NN}, $\text{distance}(x_i, x_q) \geq \max_{x^{NN_j},\, j=1,\ldots,k} \text{distance}(x^{NN_j}, x_q)$

k-NN algorithm

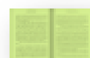
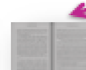
sort first **k documents**
by distance to query doc

Initialize **Dist2kNN** = sort($\delta_1,\ldots,\delta_k$) ← list of sorted distances

=rt(,..1) k ← list of sorted docs

For i=k+1,...,N

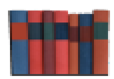
query doc

Compute: δ = **distance(** i **,** q **)**

If δ < **Dist2kNN**[k] ← distance to k^th NN (furthest NN in set)

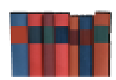
find j such that δ > **Dist2kNN**[j-1] but δ < **Dist2kNN**[j]

remove furthest house and shift queue:

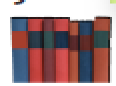
1:k] = -1]

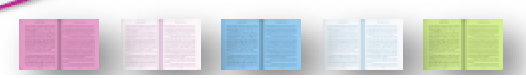
inserting new article

Dist2kNN[j+1:k] = **Dist2kNN**[j:k-1]

set **Dist2kNN**[j] = δ and [j] = i

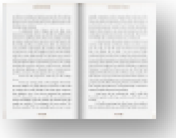
closest k docs
to query doc

Return k most similar articles 



Critical elements of NN search

Item (e.g., doc) representation

$\mathbf{x}_q \leftarrow$ 

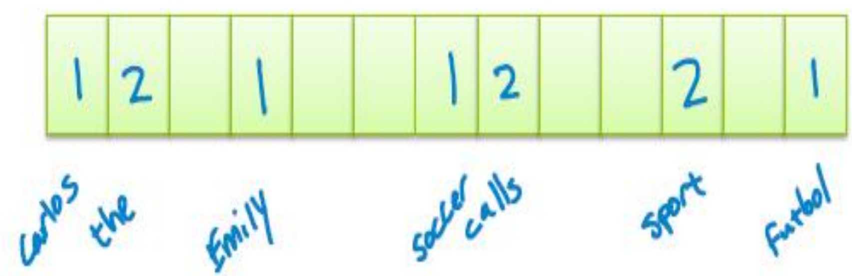
Measure of distance between items:

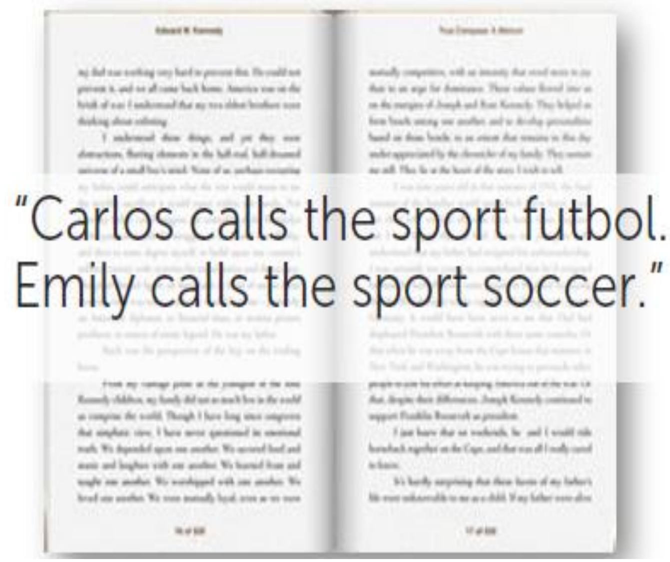
$\delta = \text{distance}(\mathbf{x}_i , \mathbf{x}_q)$

Document representation

Bag of words model

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





Document representation

Issues with word counts – Rare words

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

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

Document representation

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 = word counts

- Appears rarely in corpus (rare globally)

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

tf * idf

Trade off: local frequency vs. global rarity

Distance metrics:

Distance metrics: Defining notion of "closest"

In 1D, just Euclidean distance:

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

In multiple dimensions:

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

Distance metrics:

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

Distance metrics:

Weighting different features

Reasons:

– Some features are more relevant than others

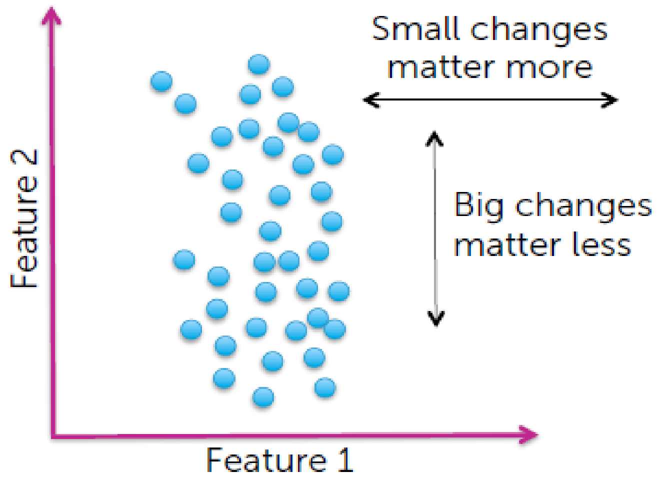
title
abstract
main body
conclusion

Distance metrics:

Weighting different features

Reasons:

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



Specify weights as a function of feature spread

For feature j:

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

Distance metrics:

Scaled Euclidean distance

Formally, this is achieved via

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

weight on each feature
(defining relative importance)

Distance metrics:

Effect of binary weights

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

$$\sqrt{a_1(\mathbf{x}_i[1]-\mathbf{x}_q[1])^2 + \ldots + a_d(\mathbf{x}_i[d]-\mathbf{x}_q[d])^2}$$

Setting weights as 0 or 1
is equivalent to
feature selection

Feature engineering/
selection is
important, but **hard**

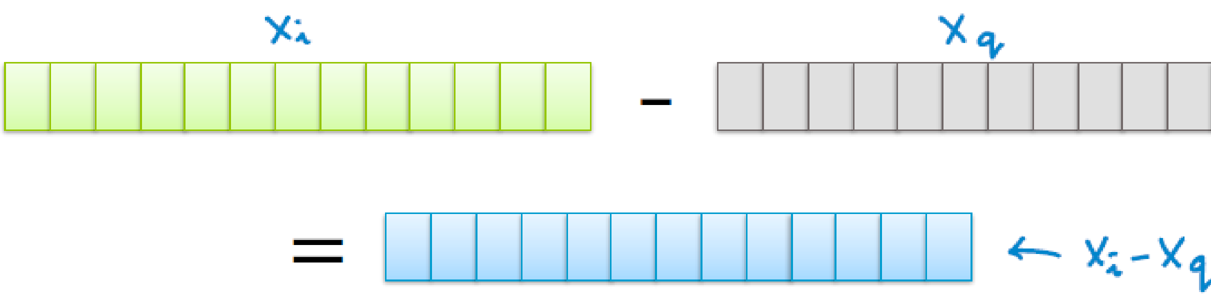
Distance metrics:

(non-scaled) Euclidean distance

Defined in terms of inner product

$$\text{distance}(\mathbf{x}_i, \mathbf{x}_q) = \sqrt{(\mathbf{x}_i - \mathbf{x}_q)^T(\mathbf{x}_i - \mathbf{x}_q)}$$

$$\sqrt{(\mathbf{x}_i[1] - \mathbf{x}_q[1])^2 + \ldots + (\mathbf{x}_i[d] - \mathbf{x}_q[d])^2}$$



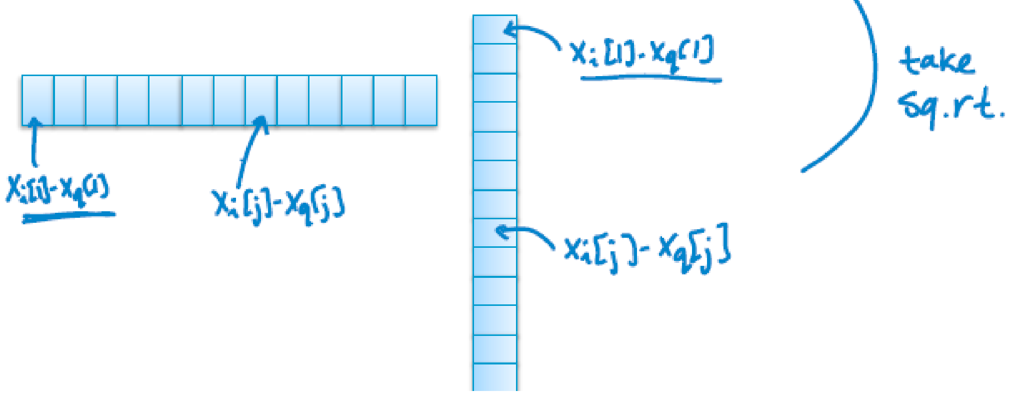
Distance metrics:

(non-scaled) Euclidean distance

Defined in terms of inner product

$$\boxed{\begin{aligned}&\text{distance}(\mathbf{x}_i, \mathbf{x}_q) = \sqrt{(\mathbf{x}_i - \mathbf{x}_q)^T(\mathbf{x}_i - \mathbf{x}_q)} \\ &\sqrt{(\mathbf{x}_i[1]-\mathbf{x}_q[1])^2 + \ldots + (\mathbf{x}_i[d]-\mathbf{x}_q[d])^2}\end{aligned}}$$

distance^2 =



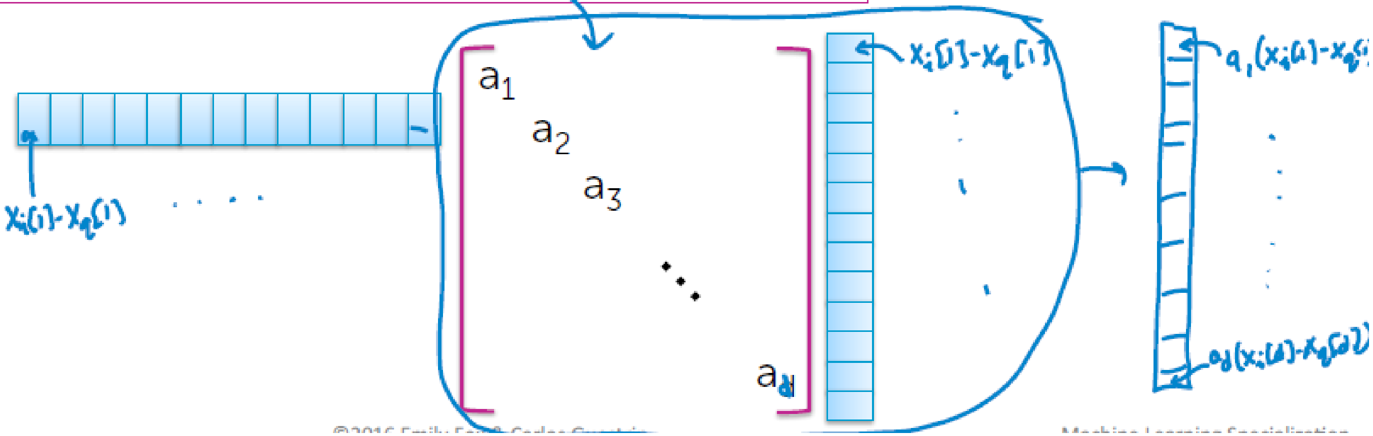
Distance metrics:

Scaled Euclidean distance

Defined in terms of inner product

$$\boxed{\begin{aligned} \text{distance}(\mathbf{x}_i, \mathbf{x}_q) &= \sqrt{(\mathbf{x}_i - \mathbf{x}_q)^T \mathbf{A} (\mathbf{x}_i - \mathbf{x}_q)} \\ &\sqrt{a_1(\mathbf{x}_i[1]-\mathbf{x}_q[1])^2 + \ldots + a_d(\mathbf{x}_i[d]-\mathbf{x}_q[d])^2} \end{aligned}}$$

$\text{distance}^2 =$



Distance metrics:

Another natural inner product measure

x_q

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

Similarity

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

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

$$= 13$$

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

x_i

Distance metrics:

Another natural inner product measure

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

Similarity = 0

0	0	1	0	0	0	9	0	0	6	0	4	0

Distance metrics

Cosine similarity – normalize

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

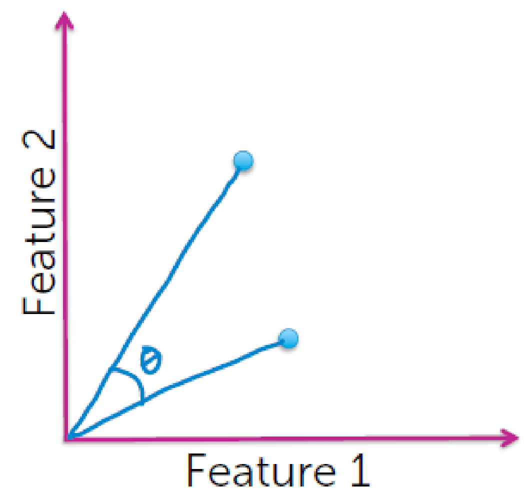
$$\frac{\mathbf{x}_i^T \mathbf{x}_q}{||\mathbf{x}_i||\,||\mathbf{x}_q||} = \cos(\theta)$$

$$= \left(\frac{x_i}{||x_i||}\right)^T \left(\frac{x_q}{||x_q||}\right)$$

first normalize

$a^T b = ||a||\,||b|| \cos(\theta)$

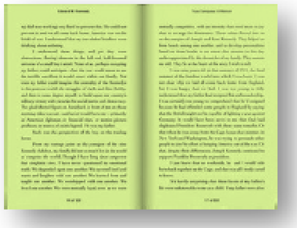
- Not a proper distance metric
- Efficient to compute for sparse vecs



26/11, 3/12, 10/12/2019

Distance metrics

Normalize



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

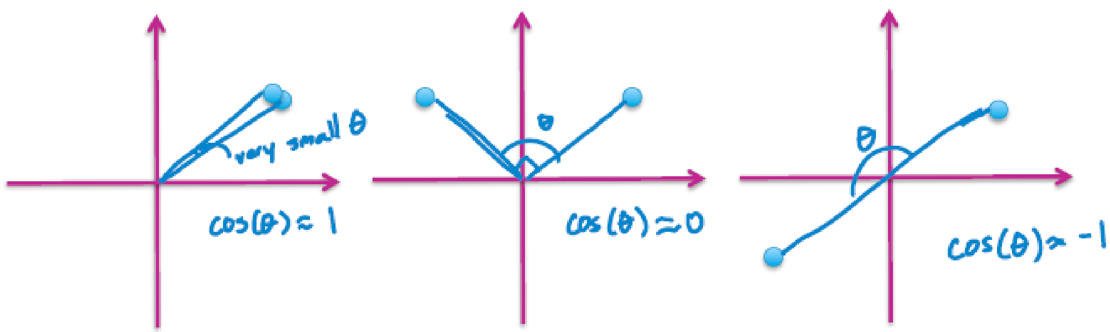
$\leftarrow x_i$

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

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

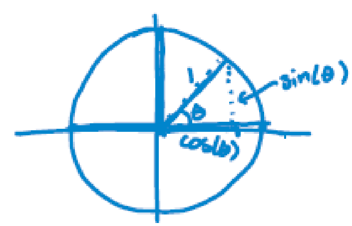
Distance metrics

Cosine similarity



In general, $-1 <$ similarity < 1

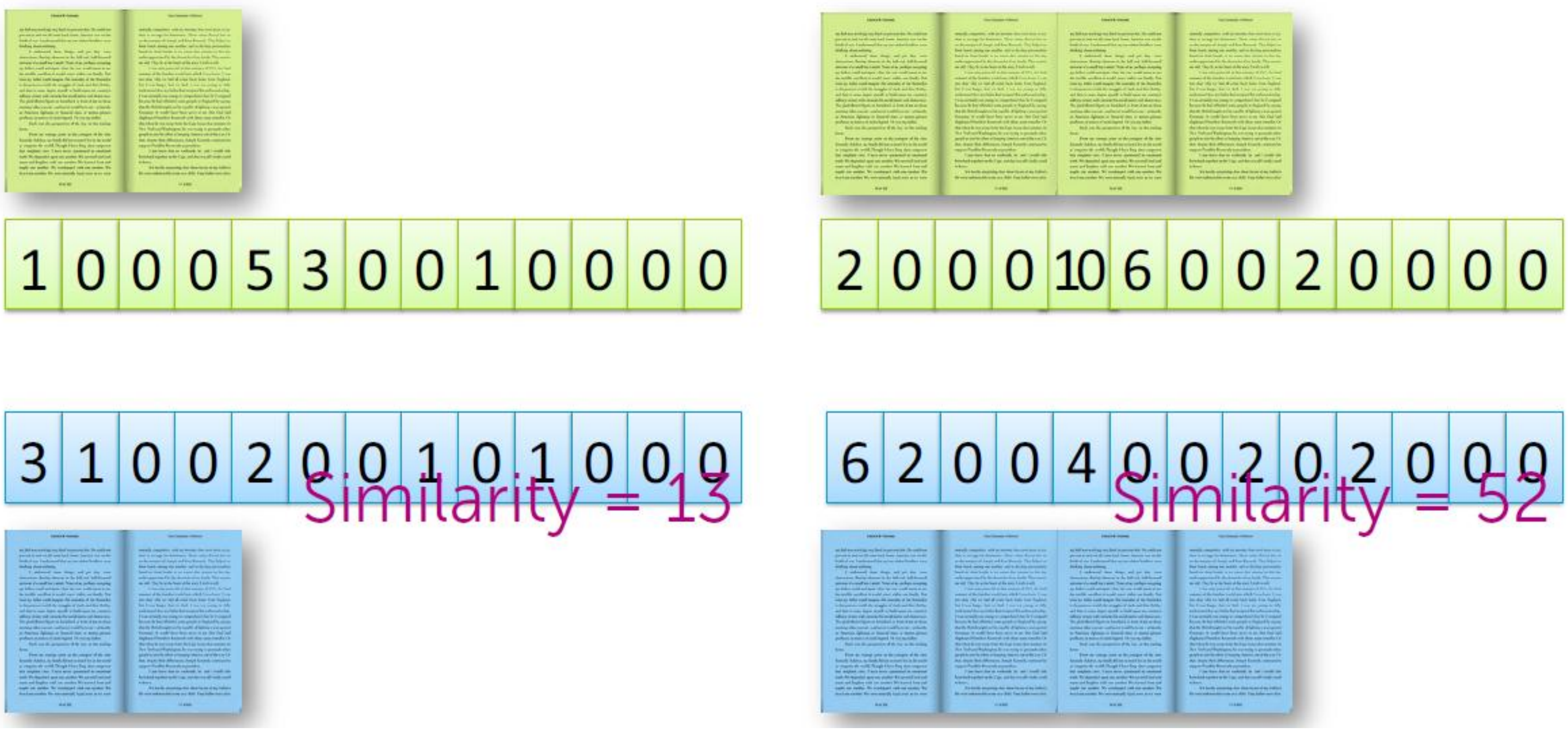
For positive features (like tf-idf)
$0 <$ similarity < 1 } our focus



Define **distance = 1-similarity**

Distance metrics

To normalize or not?



Distance metrics

In the normalized case

1/6	0	0	0	5/6	3/6	0	0	1/6	0	0	0	0
3/4	1/4	0	0	1/2	0	0	1/4	0	1/4	0	0	0

Similarity = 13/24

1/6	0	0	0	5/6	3/6	0	0	1/6	0	0	0	0
3/4	1/4	0	0	1/2	0	0	1/4	0	1/4	0	0	0

Similarity = 13/24

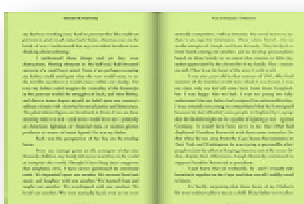
Distance metrics

But not always desired...

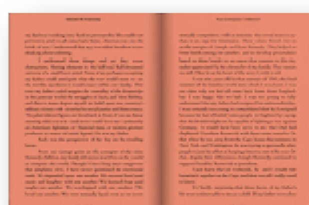
long document

short tweet

Normalizing can make dissimilar objects appear more similar



long document



long document

Common compromise:
Just cap maximum word counts

Distance metrics

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

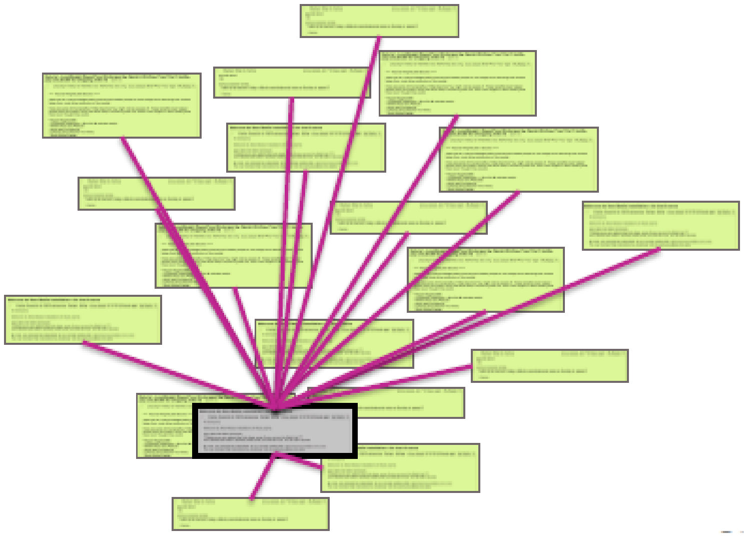
Add together with user-specified weights

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

Complexity of 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!

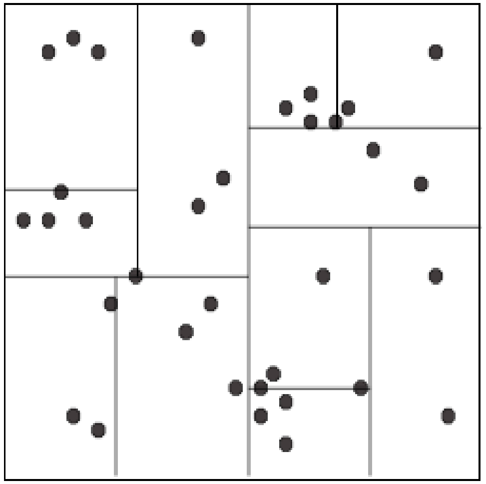


What if N is huge??? (and many queries)

KD-trees

Structured organization of documents

- Recursively partitions points into axis aligned boxes.



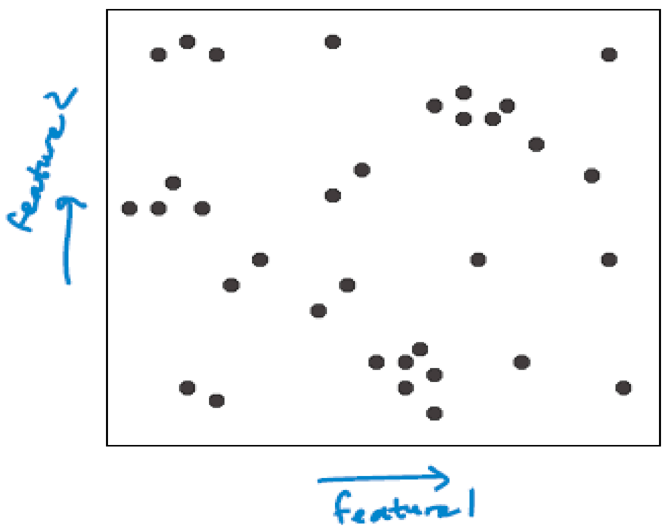
Enables more efficient pruning of search space

Works "well" in "low-medium" dimensions

- We'll get back to this...

KD-trees

KD-tree construction



Start with a list of d-dimensional points.

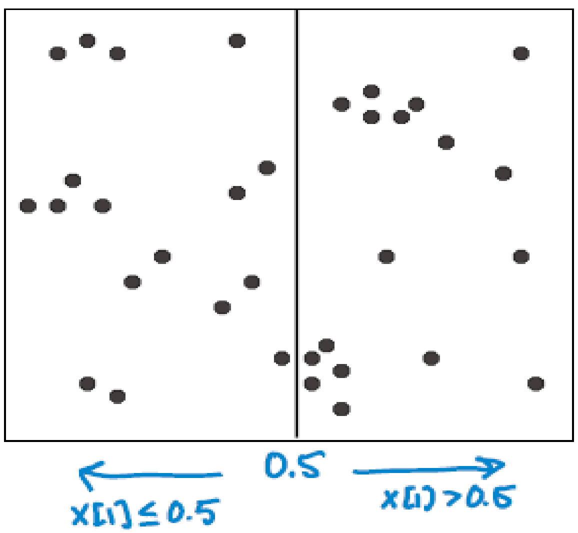
Pt	**x**[1]	**x**[2]
1	0.00	0.00
2	1.00	4.31
3	0.13	2.85
...	...	...

obs. indices

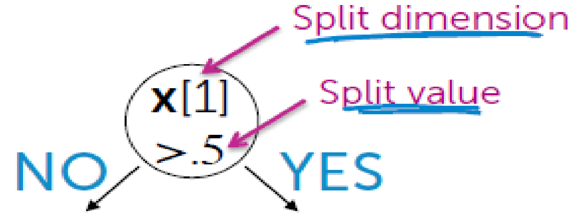
feat. 1 (word 1)

feat. 2 (word 2)

KD-tree construction



Split points into 2 groups



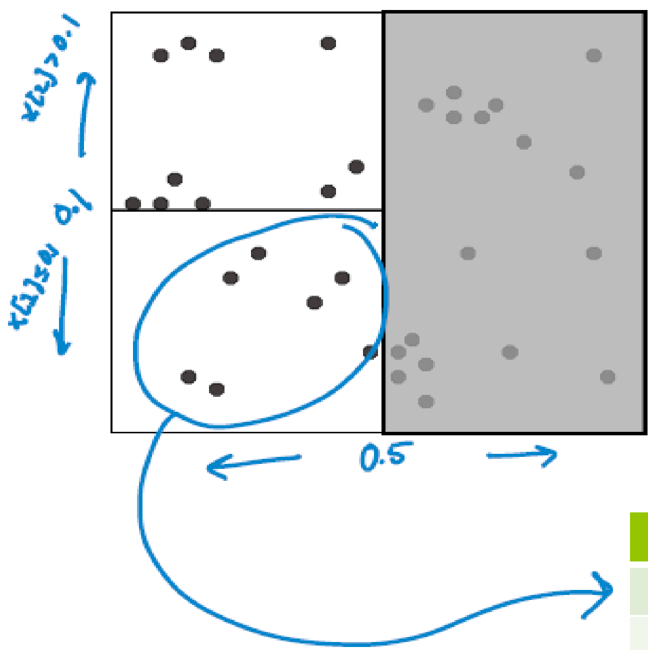
Pt	x[1]	x[2]
1	0.00	0.00
3	0.13	2.85
...	...	...

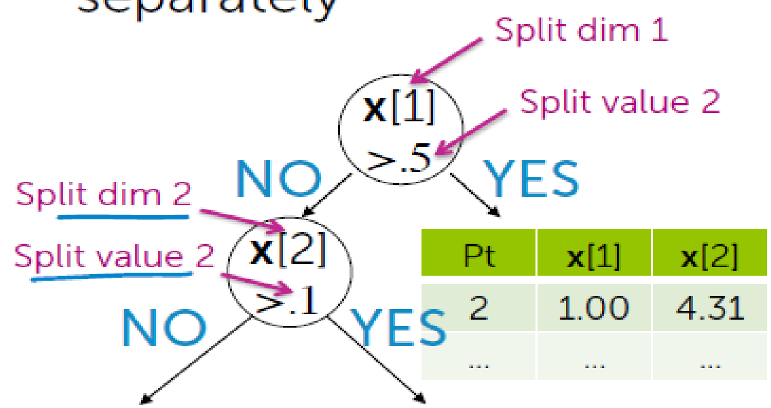
Pt	x[1]	x[2]
2	1.00	4.31
...	...	...

KD-trees

KD-tree construction

Recurse on each group separately





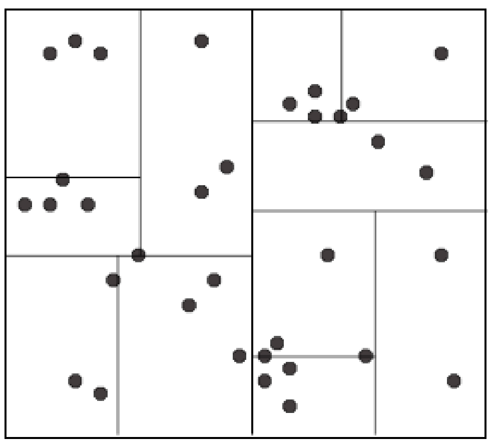
Pt	x[1]	x[2]
2	1.00	4.31
...	...	...

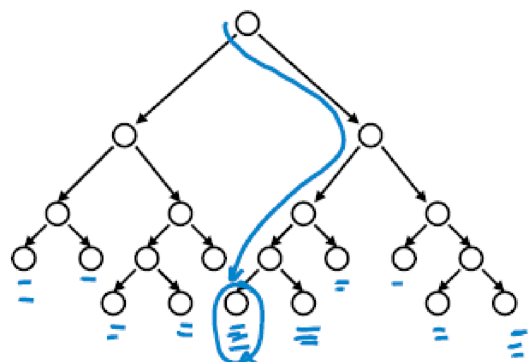
Pt	x[1]	x[2]
3	0.13	2.85
...	...	...

Pt	x[1]	x[2]
1	0.00	0.00
...	...	...

KD-trees

KD-tree construction





points here satisfy all conditions down the tree to this leaf

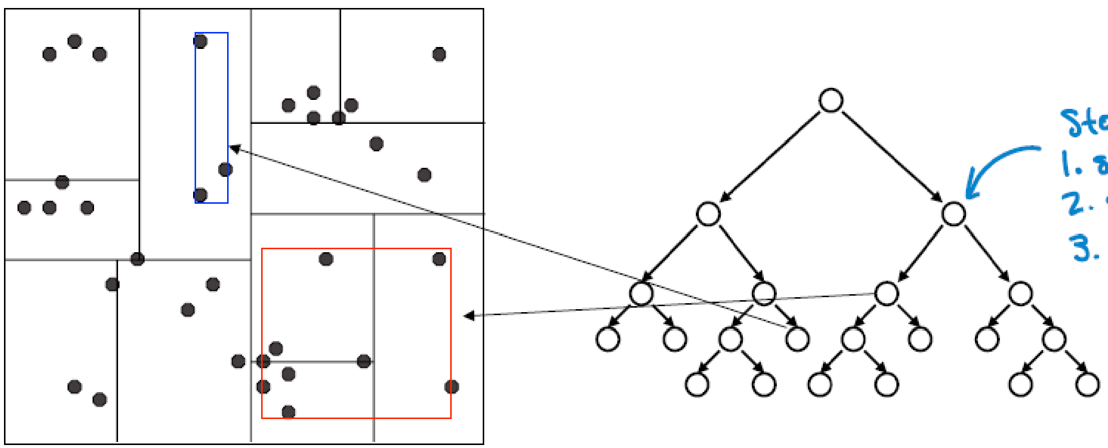
Continue splitting points at each set

- Creates a binary tree structure

Each leaf node contains a list of points

KD-trees

KD-tree construction



Store:
1. split dim?
2. split value
3. bounding box that is as small as possible while containing pts

Keep one additional piece of info at each node:

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

KD-trees

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?

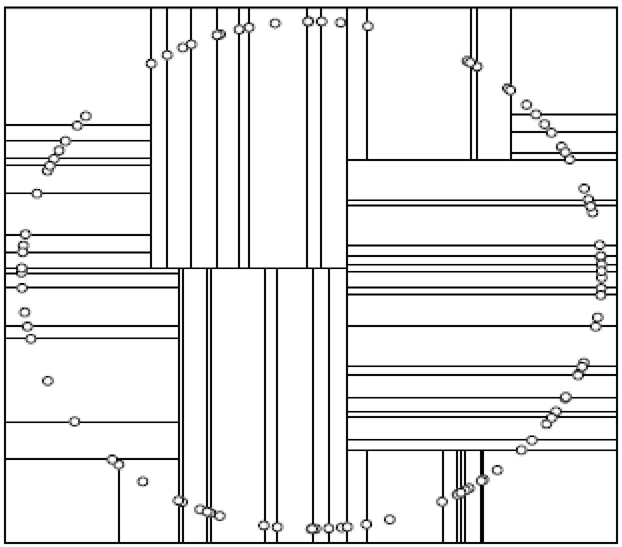
  Fewer than m pts left

  or

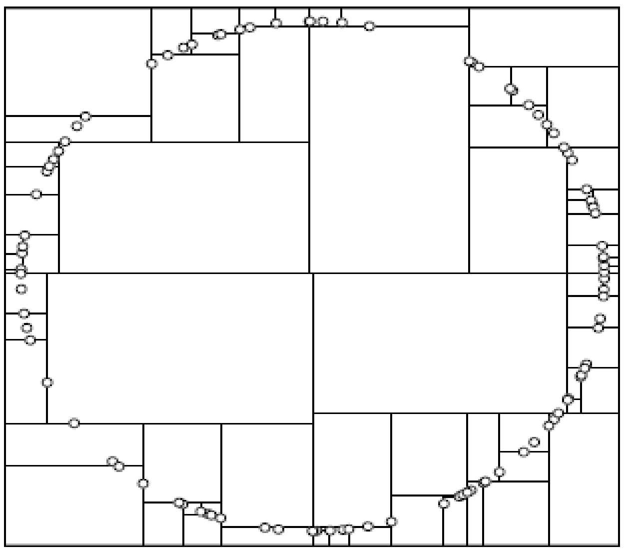
  box hits minimum width

KD-trees

Many heuristics…

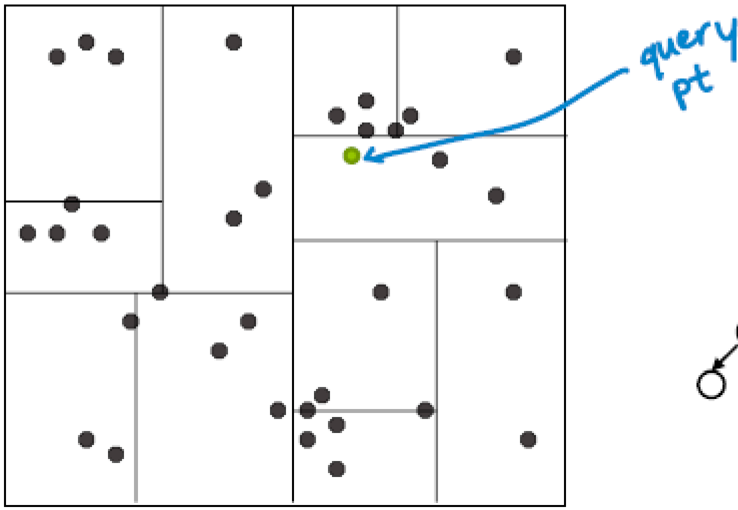


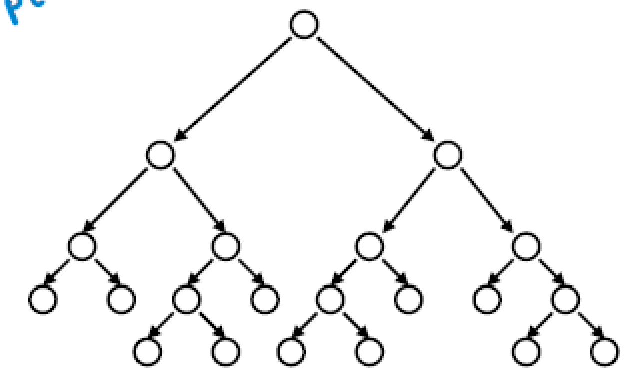
median heuristic



center-of-range heuristic

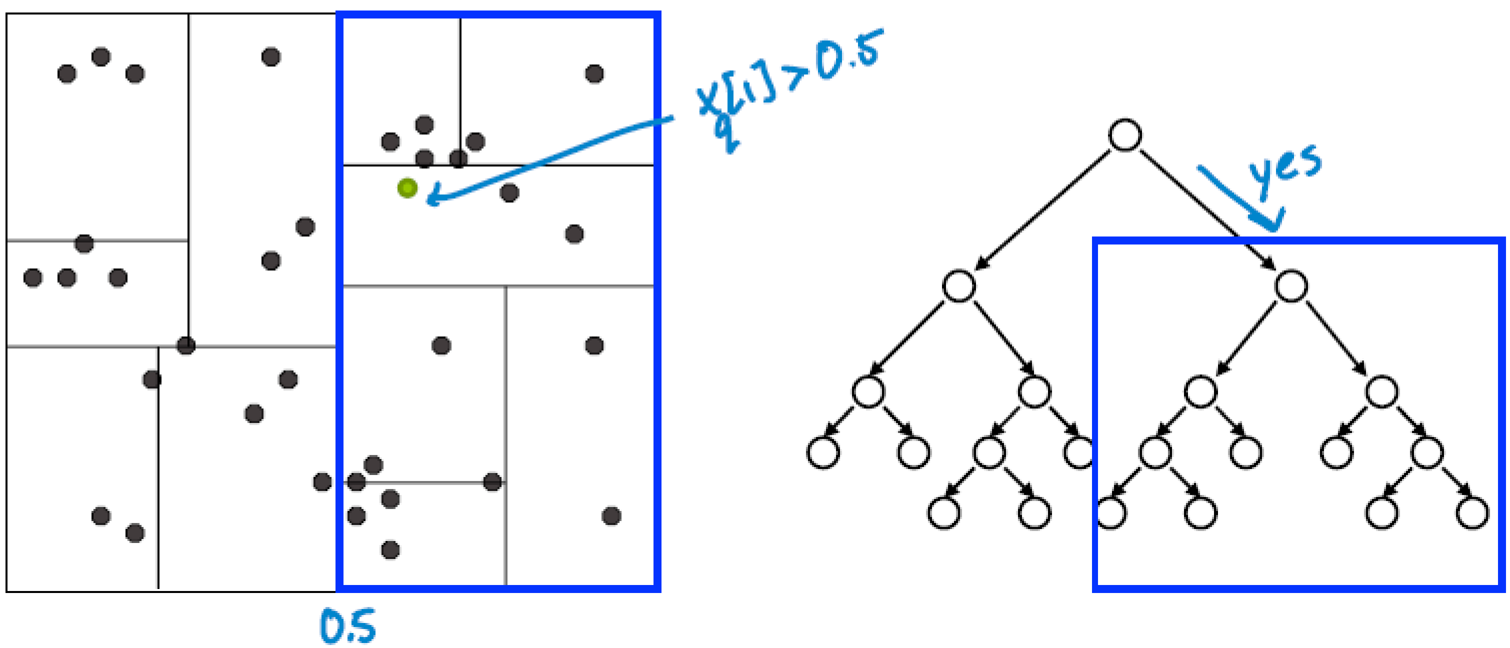
Nearest neighbor with KD-trees





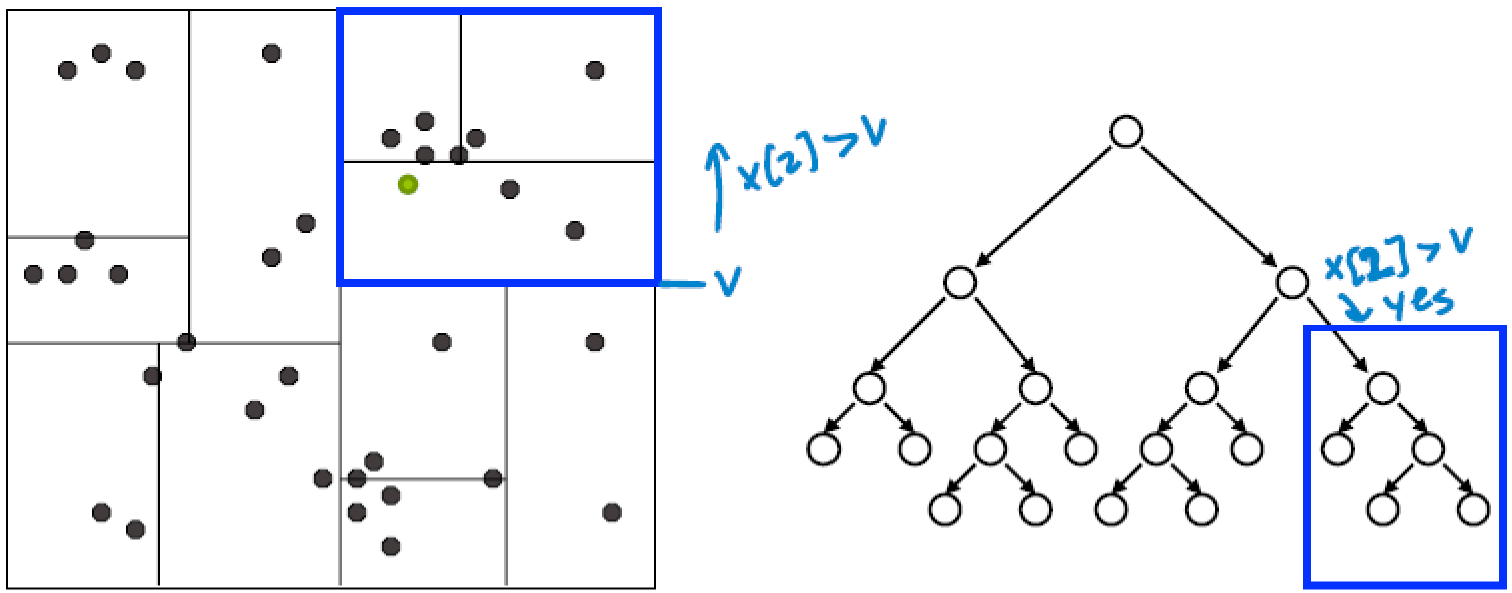
Traverse tree looking for nearest neighbor to query point

Nearest neighbor with KD-trees



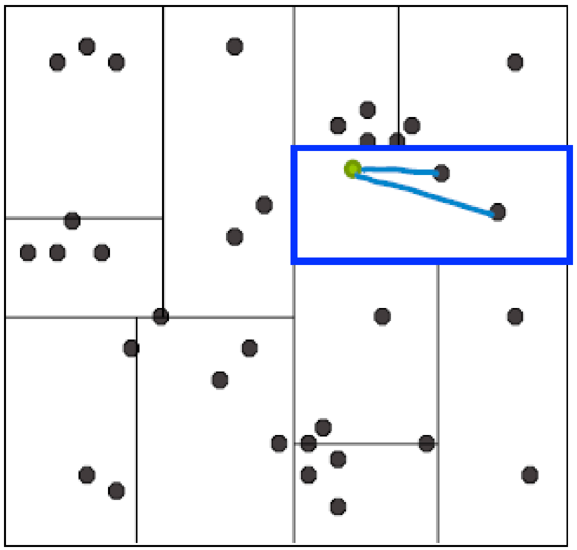
1. Start by exploring leaf node containing query point

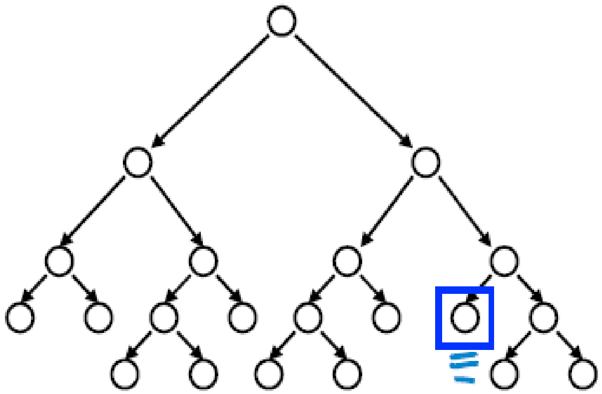
Nearest neighbor with KD-trees



1. Start by exploring leaf node containing query point

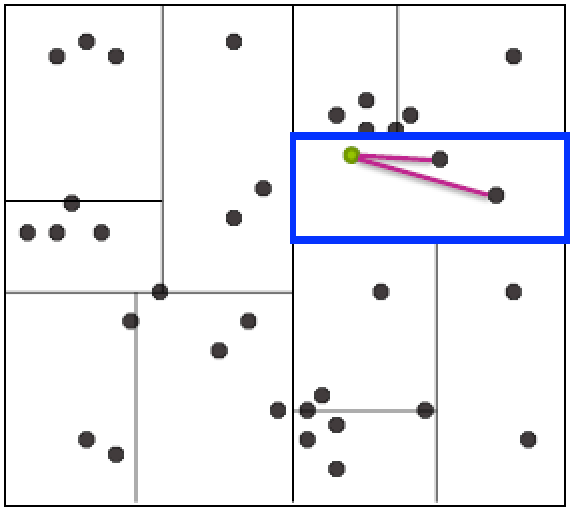
Nearest neighbor with KD-trees

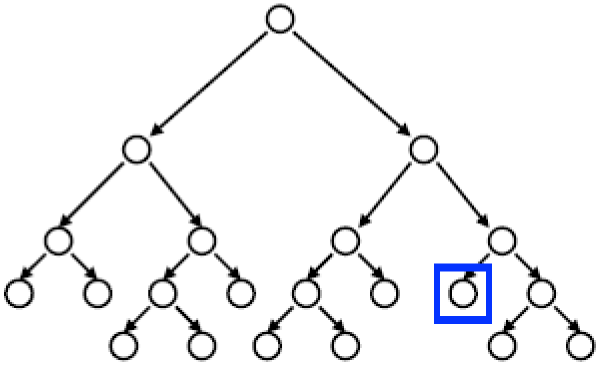




1. Start by exploring leaf node containing query point

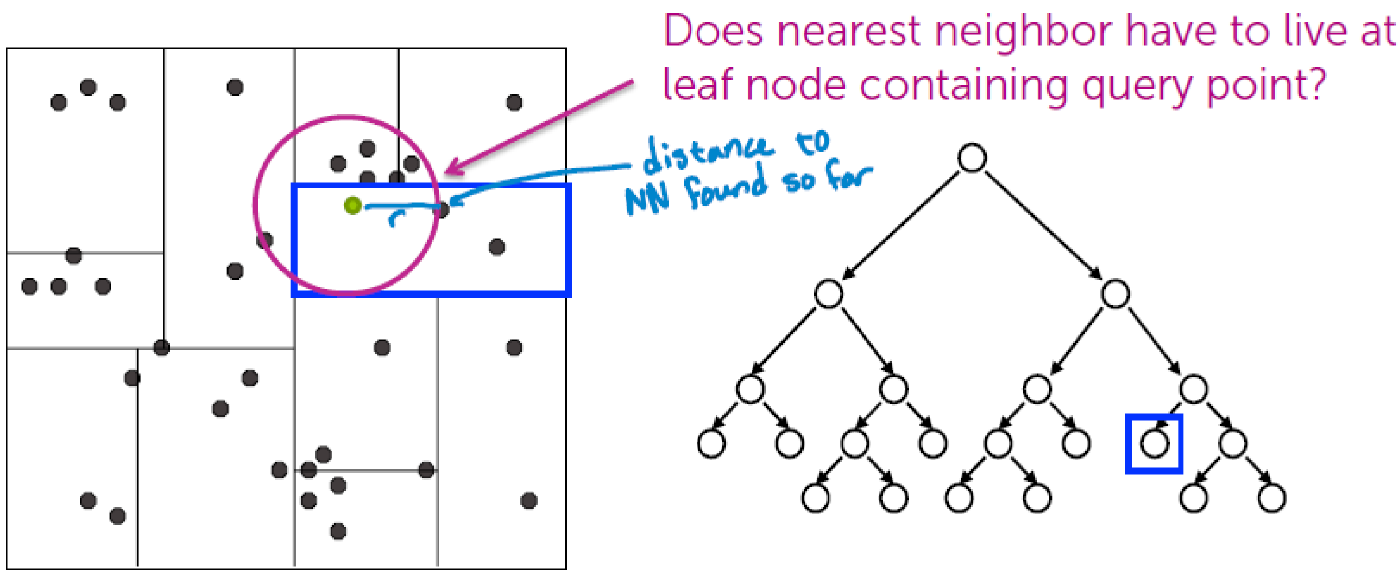
Nearest neighbor with KD-trees





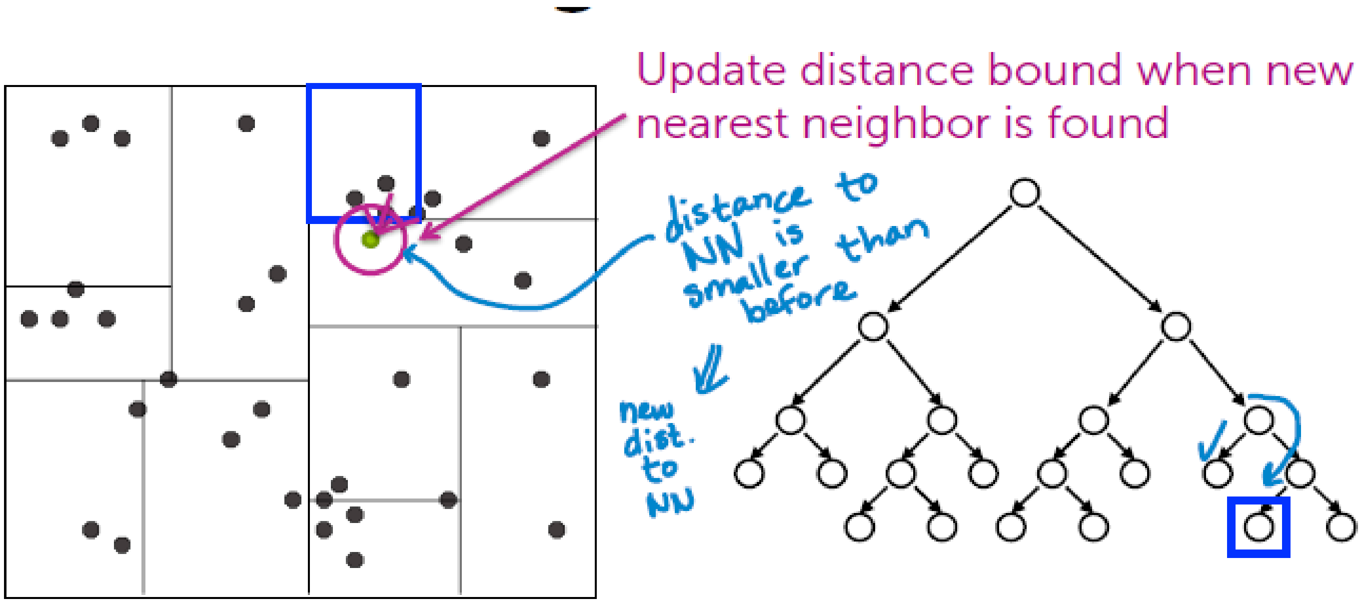
1. Start by exploring leaf node containing query point
2. Compute distance to each other point at leaf node

Nearest neighbor with KD-trees



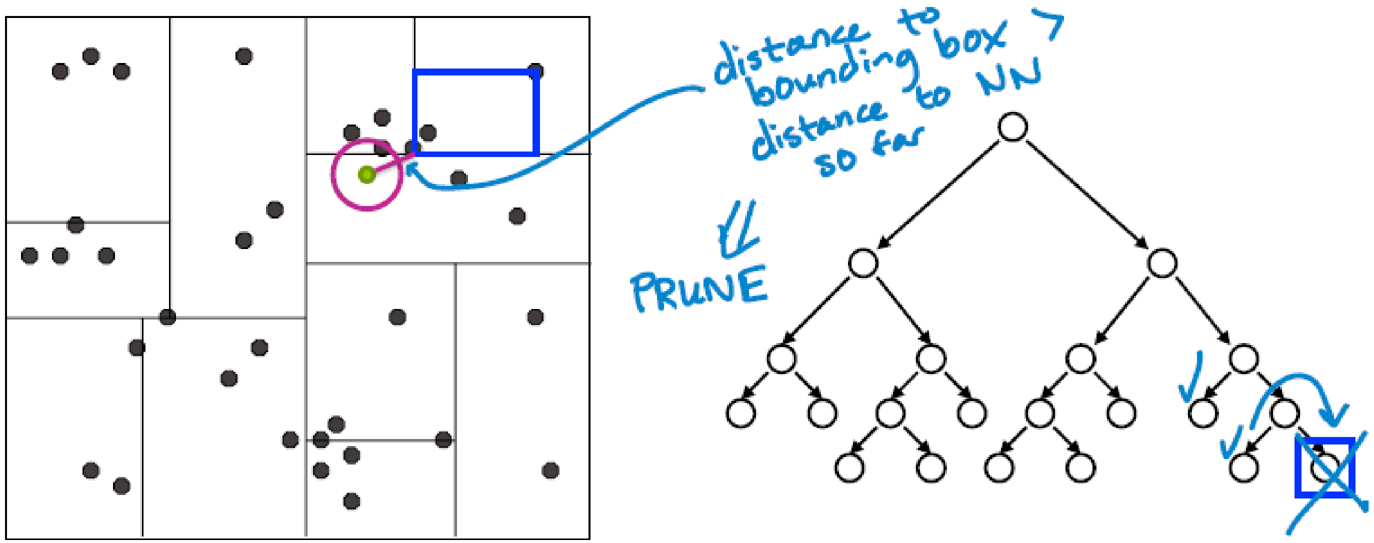
1. Start by exploring leaf node containing query point
2. Compute distance to each other point at leaf node

Nearest neighbor with KD-trees



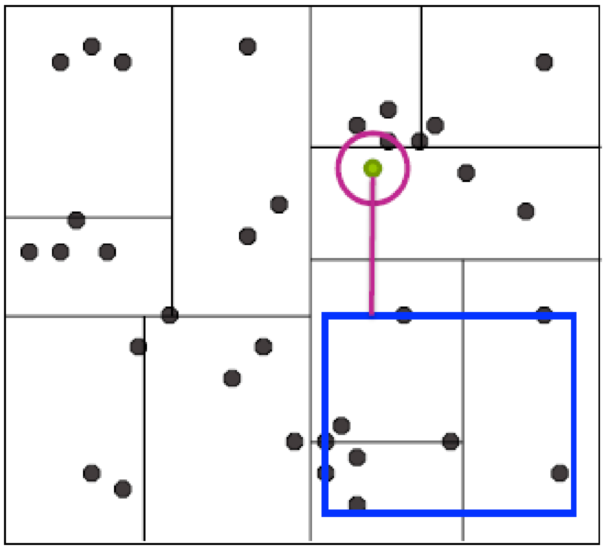
1. Start by exploring leaf node containing query point
2. Compute distance to each other point at leaf node
3. Backtrack and try other branch at each node visited

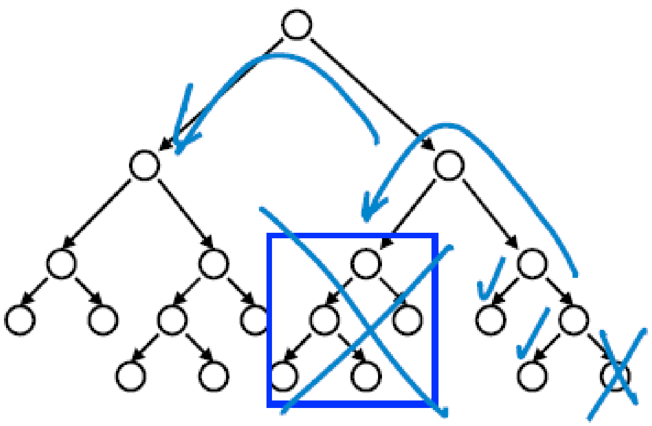
Nearest neighbor with KD-trees



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

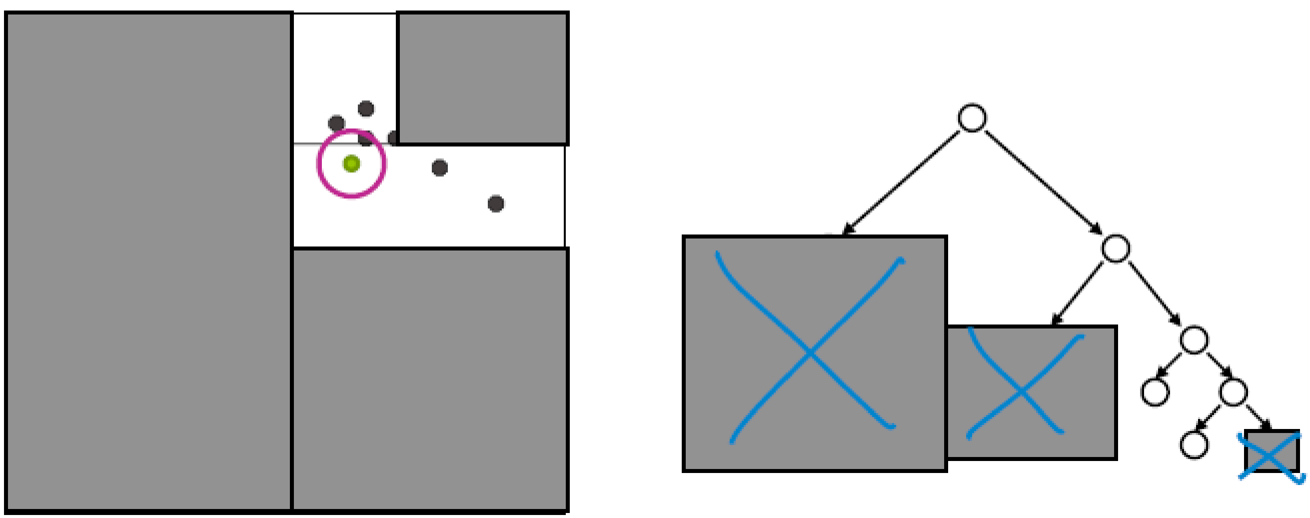
Nearest neighbor with KD-trees





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

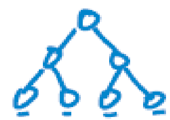
Nearest neighbor with KD-trees



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

Nearest neighbor with KD-trees

Complexity



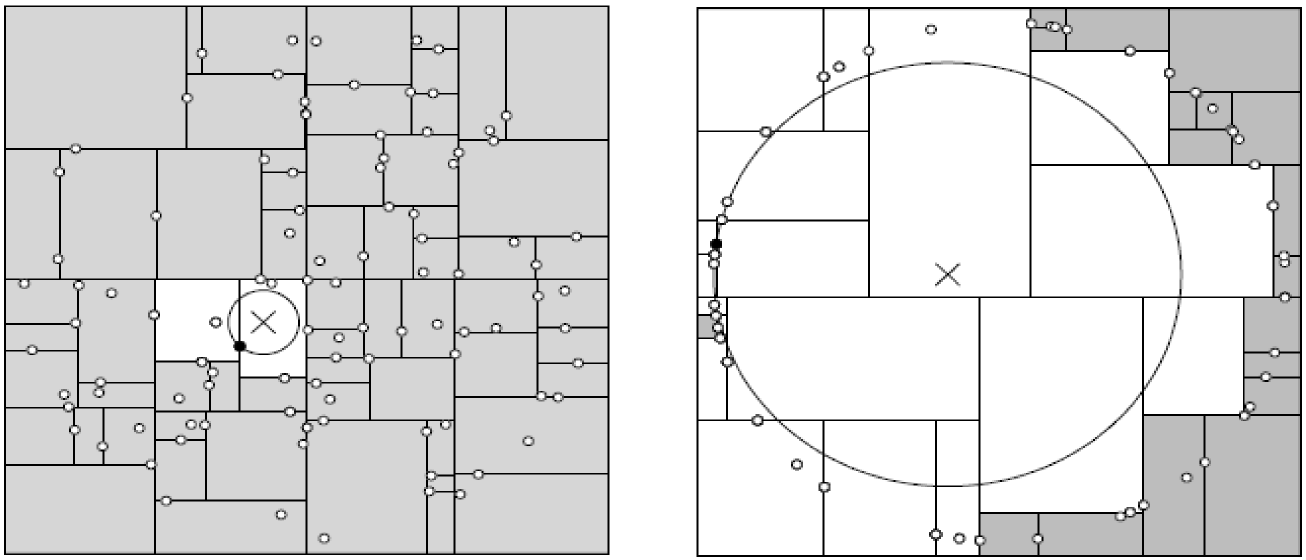
For (nearly) balanced, binary trees...

- Construction
  - Size: 2N−1 nodes if 1 datapt at each leaf → 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) → O(N)

Under some assumptions on distribution of points, we get O(logN) but exponential in d

Nearest neighbor with KD-trees

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:

  $O(N^2)$

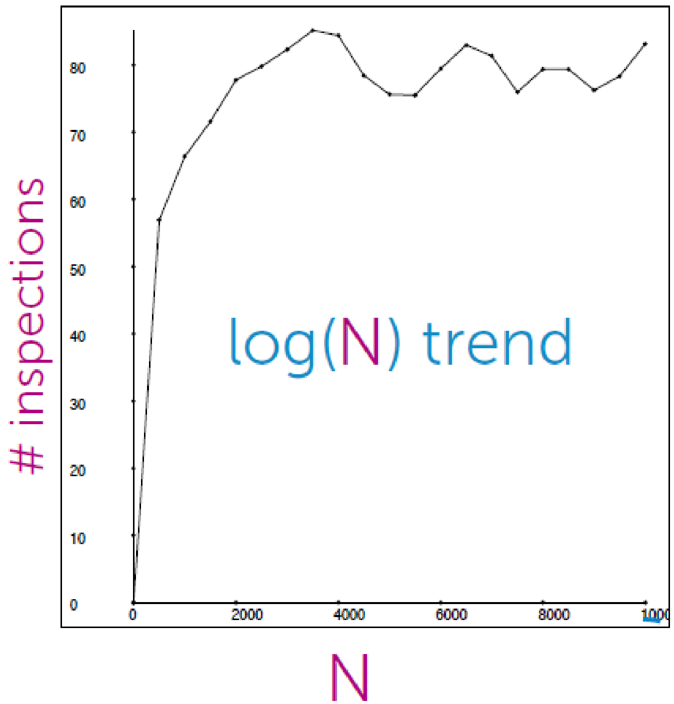
- kd-trees:

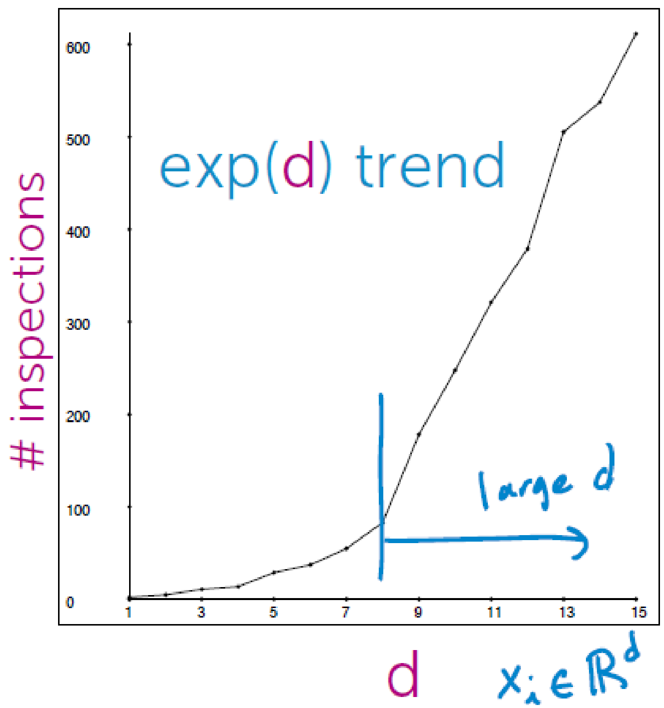
  $O(N \log N) \longrightarrow O(N^2)$

  potentially very large savings for large N!

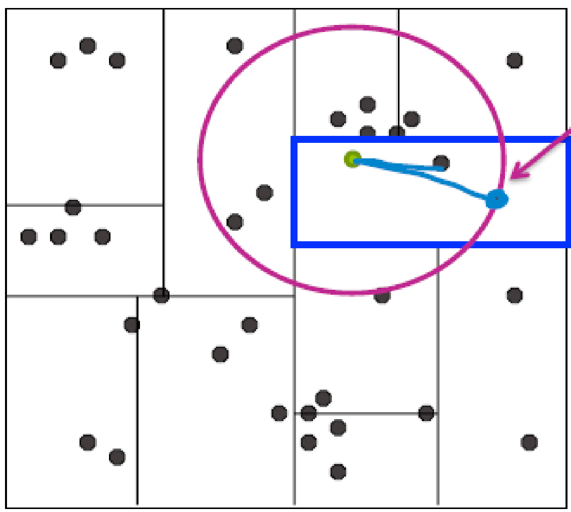
Complexity for N queries

Inspections vs. N and d

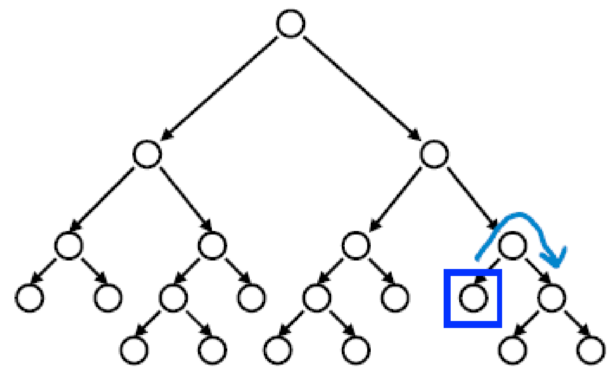




k-NN with KD-trees

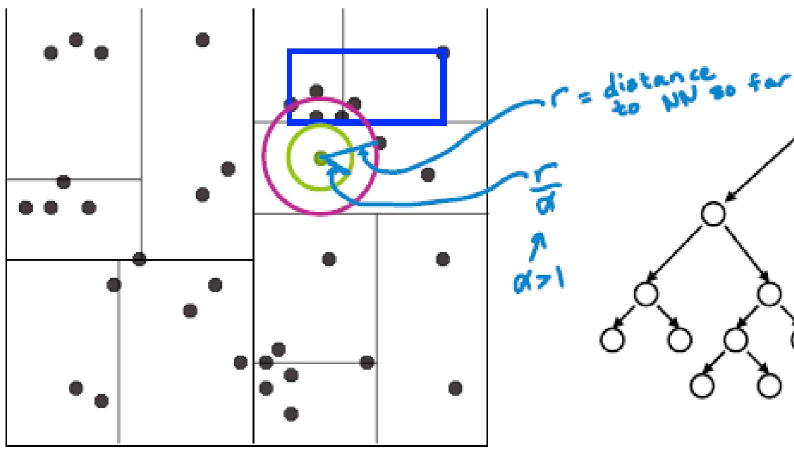


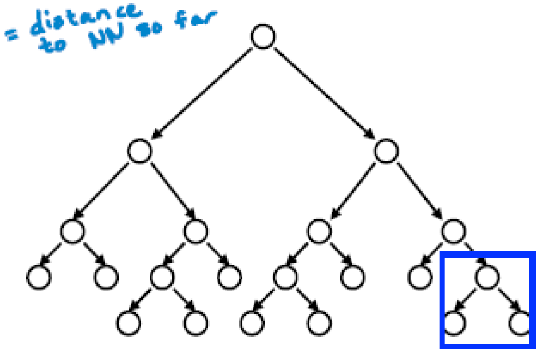
distance to 2nd nearest neighbor
in 2-NN example

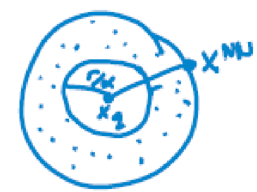


Exactly same algorithm, but maintain distance to
furthest of current *k* nearest neighbors

Approximate k-NN with KD-trees







Before: Prune when distance to bounding box > r

Now: Prune when distance to bounding box > r/α

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

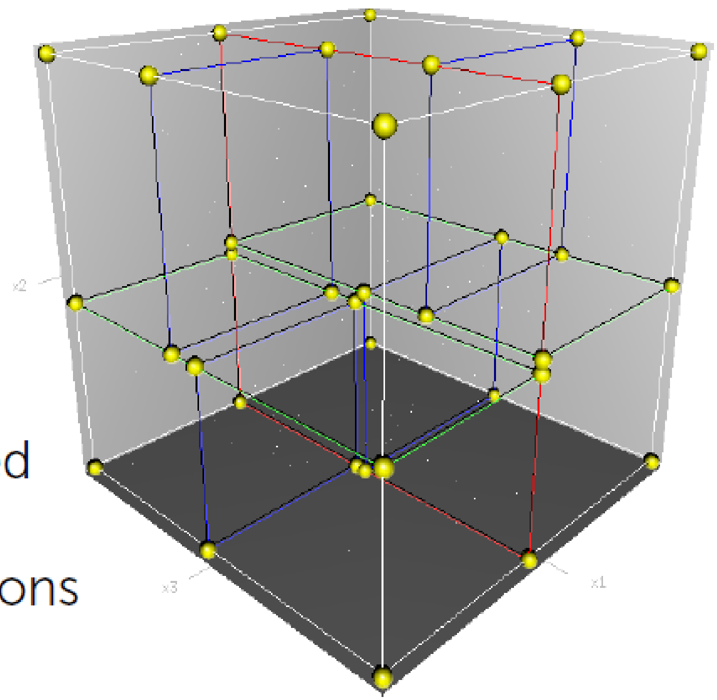
- Distance metric and data representation crucial to answer returned

For both, high-dim spaces are hard!

- Number of kd-tree searches can be exponential in dimension
  - Rule of thumb... $N >> 2^d$... 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

KD-tree in high dimmensions

- 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

Out of millions of articles, do we need the closest article or just one that's pretty similar?

Do we even fully trust our measure of similarity???

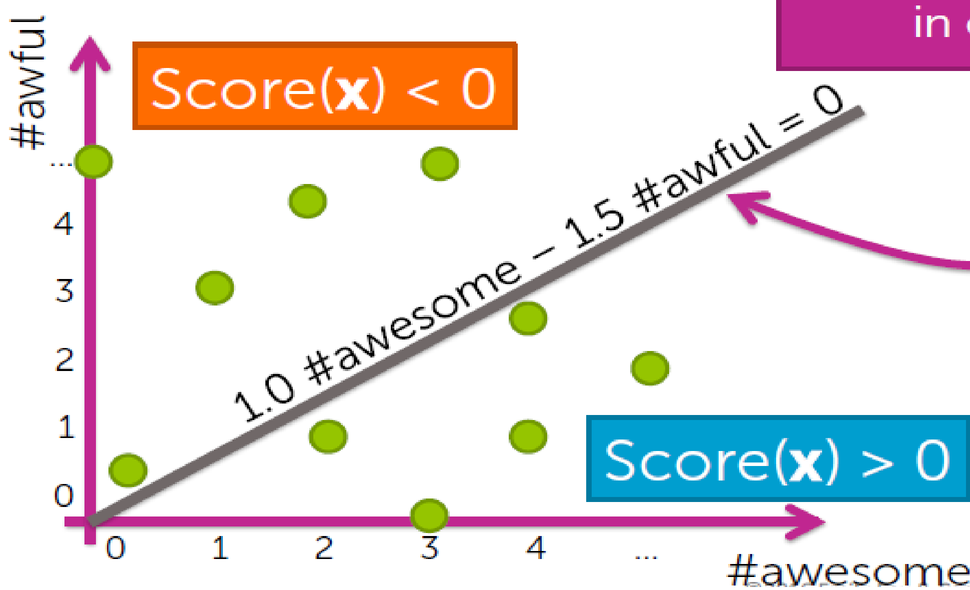
- Focus on methods that provide good probabilistic guarantees on approximation

Locality Sensitive Hashing (LHS) as alternative to KD-trees

Simple “binning” of data into 2 bins

Score(**x**) = 1.0 #awesome – 1.5 #awful

Like a decision boundary in classification



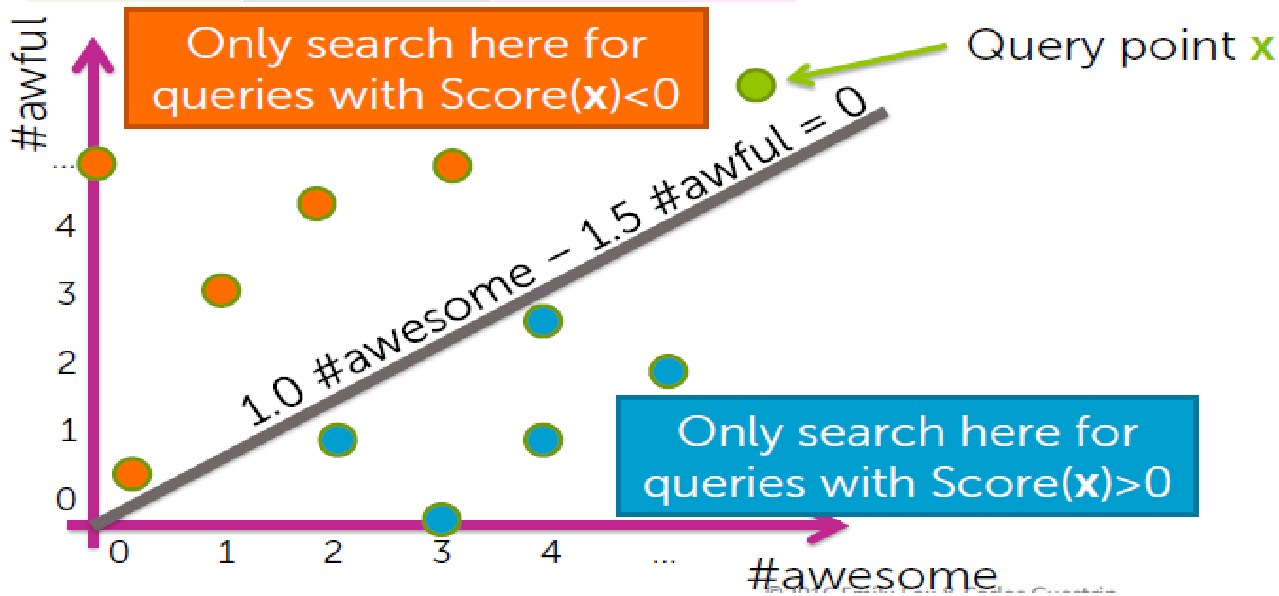
2D Data	Sign(Score)
$\mathbf{x}_1$ = [0, 5]	-1
$\mathbf{x}_2$ = [1, 3]	-1
$\mathbf{x}_3$ = [3, 0]	1
...	...

Locality sensitive hashing

Using bins 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 Score($\mathbf{x}$)<0



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 Score(x)<0

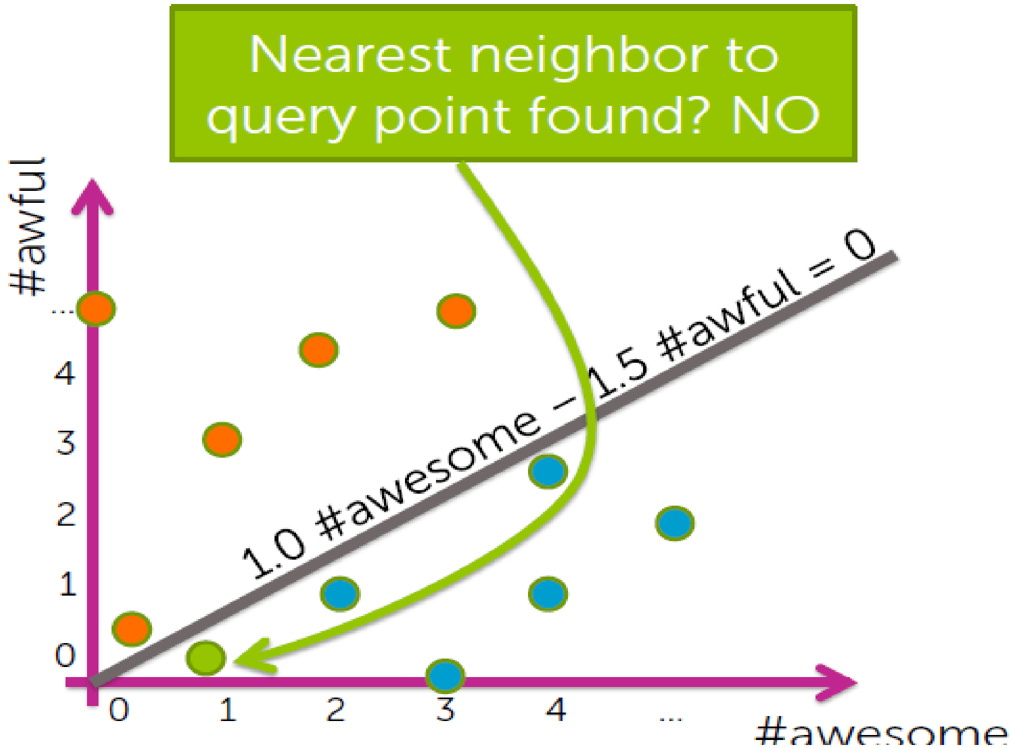
Bin	0	1
List containing indices of datapoints:	{1,2,4,7,...}	{3,5,6,8,...}

HASH TABLE

search for NN amongst this set

Locality sensitive hashing

Provides approximate NN



Locality sensitive hashing

Three potential issues with simple approach

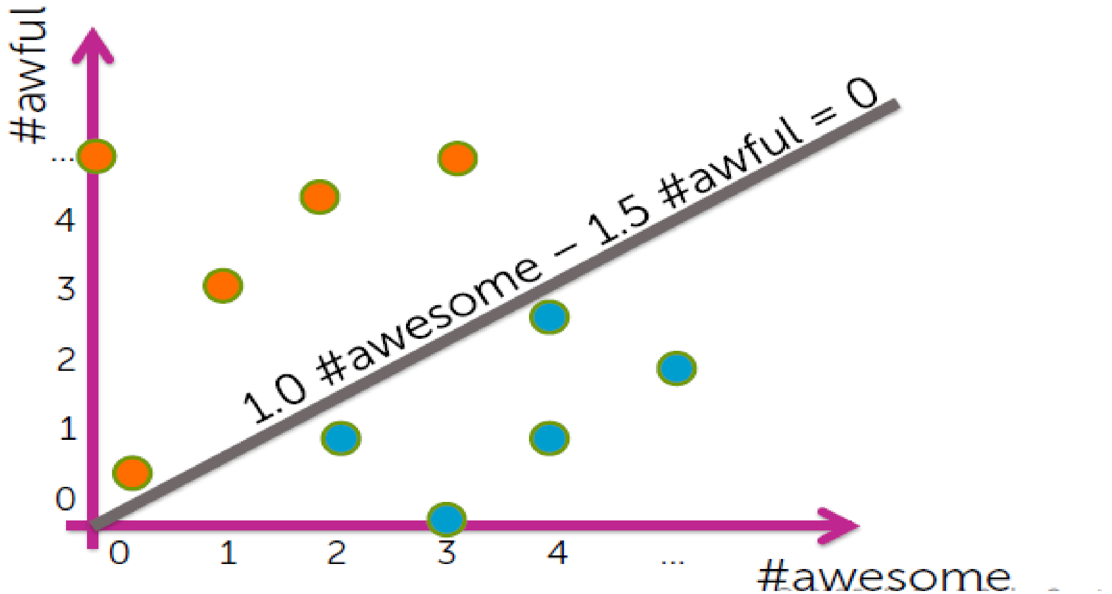
1. Challenging to find good line
2. Poor quality solution:
   - Points close together get split into separate bins
3. Large computational cost:
   - Bins might contain many points, so still searching over large set for each NN query

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

Locality sensitive hashing

How to define the line?

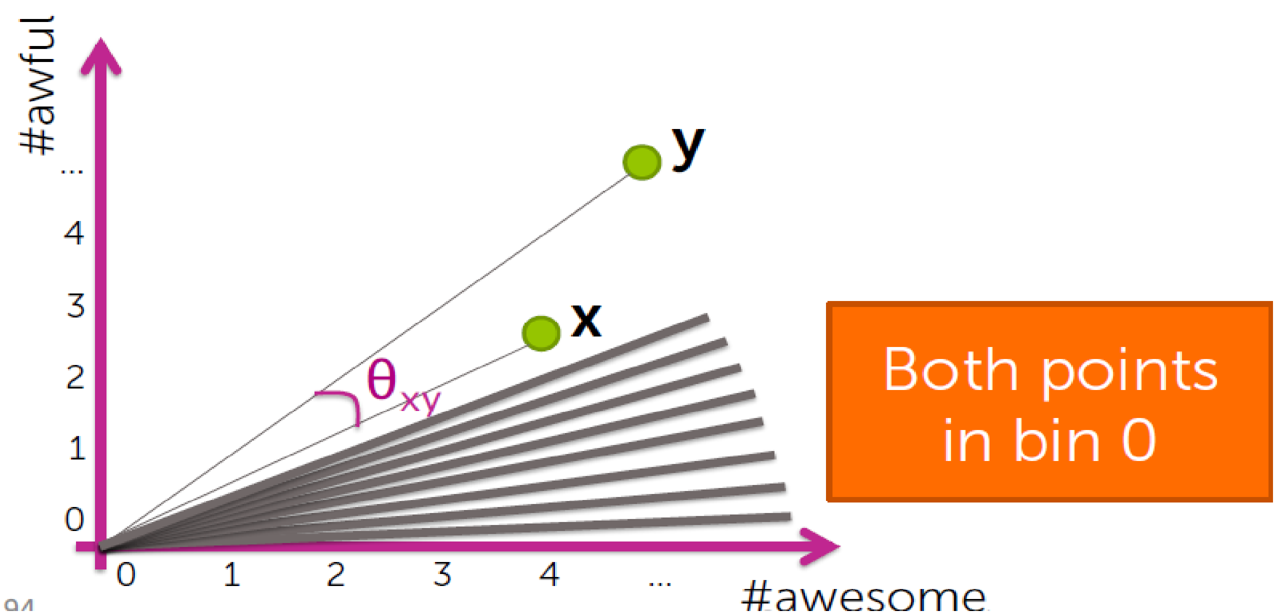
Crazy idea:
Define line randomly!



Locality sensitive hashing

How bad can a random line be?

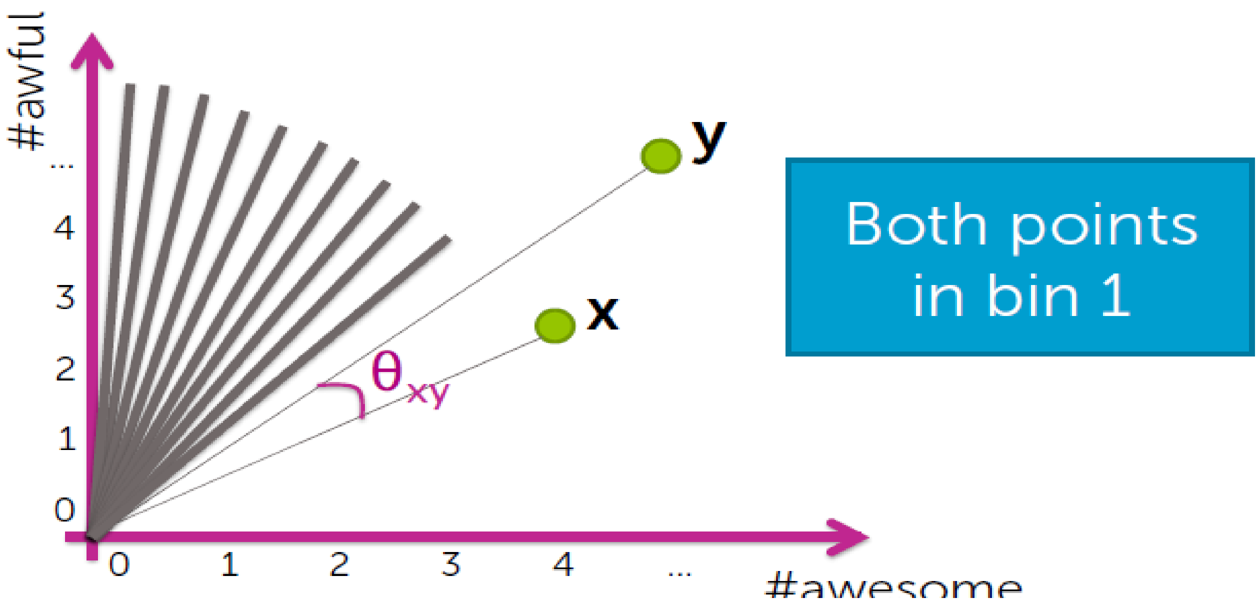
Goal: If **x**,**y** are close (according to cosine similarity), want binned values to be the same.



Locality sensitive hashing

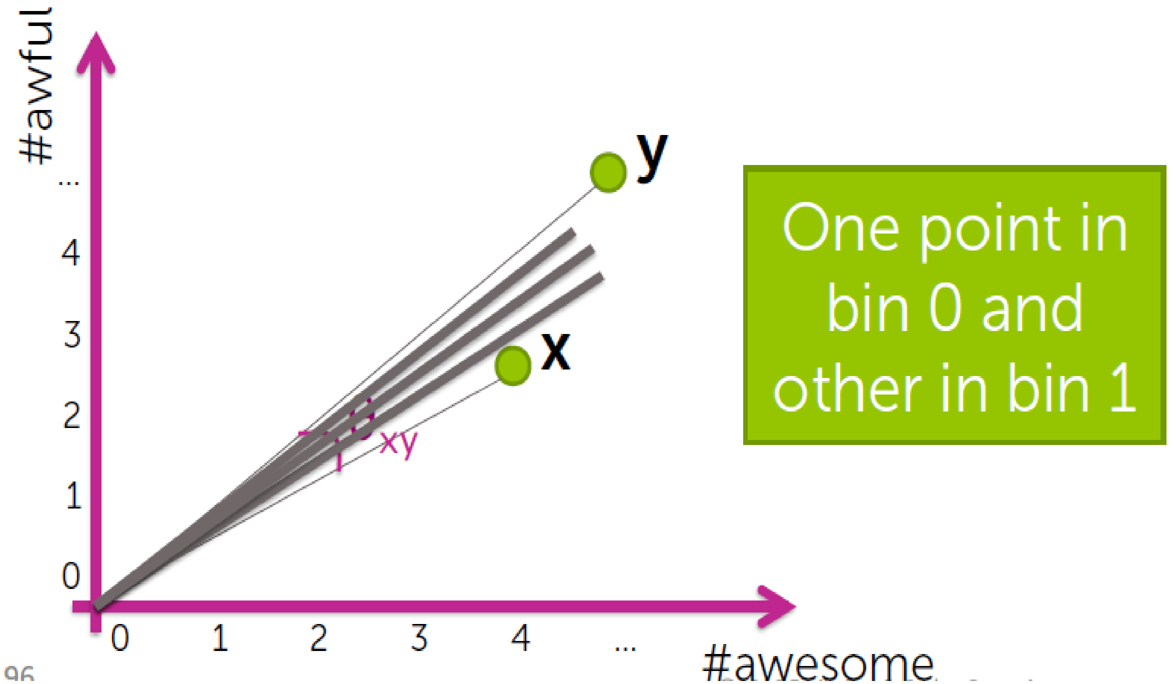
How bad can a random line be?

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



Locality sensitive hashing

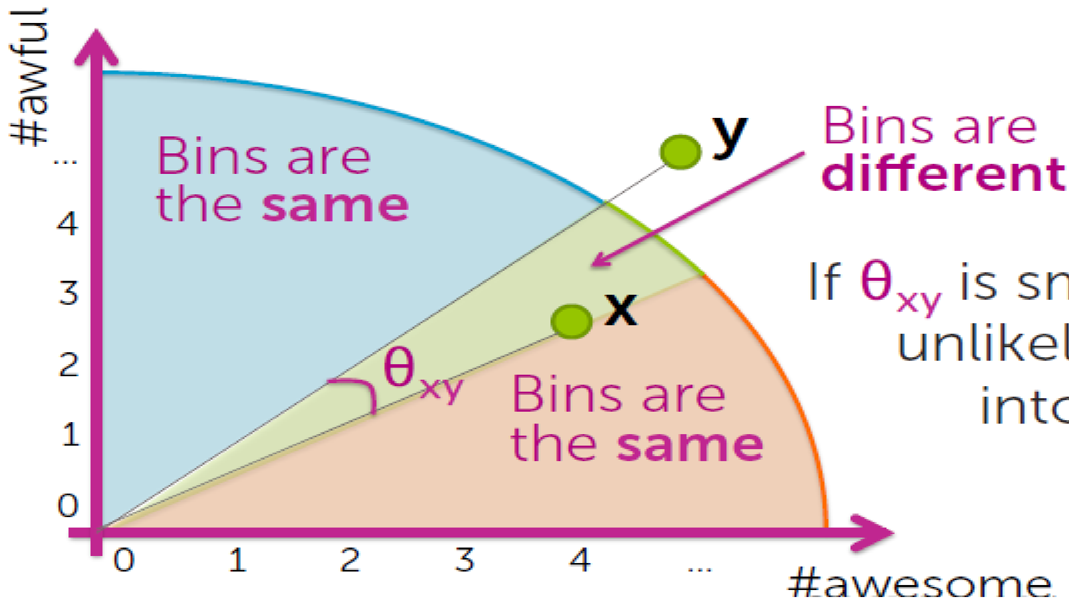
Goal: If **x**,**y** are close (according to cosine similarity), want binned values to be the same.



Locality sensitive hashing

How bad can a random line be?

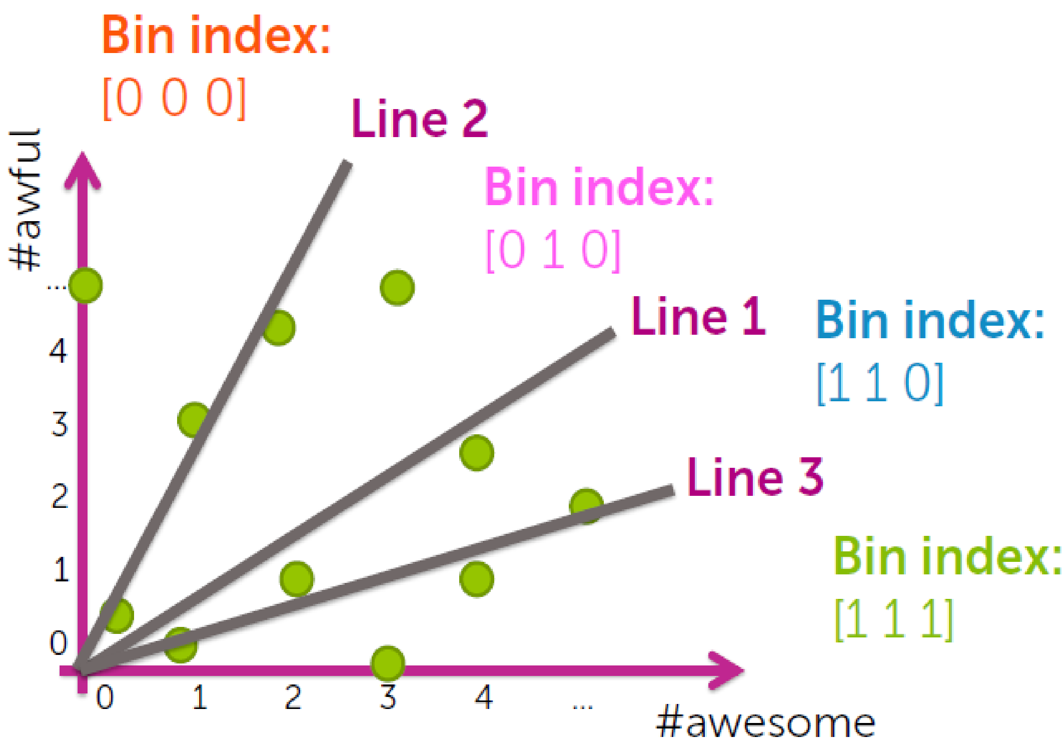
Goal: If **x**,**y** are close (according to cosine similarity), want binned values to be the same.



If θ_{xy} is small (**x**,**y** close), unlikely to be placed into separate bins

LSH: improving efficiency

Reducing search cost through more bins



LSH: improving efficiency

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

LSH: improving efficiency

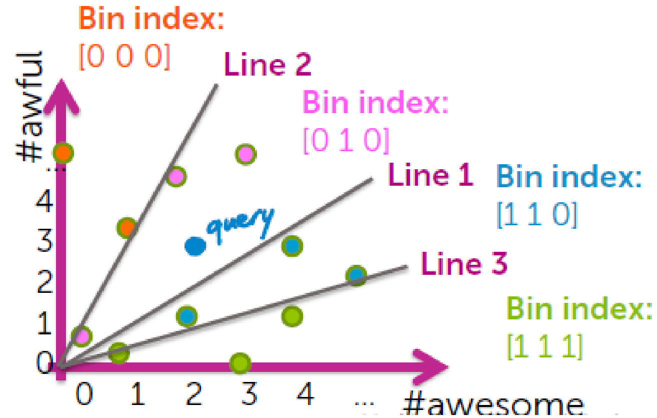
Improving search quality by searching neighboring bins

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

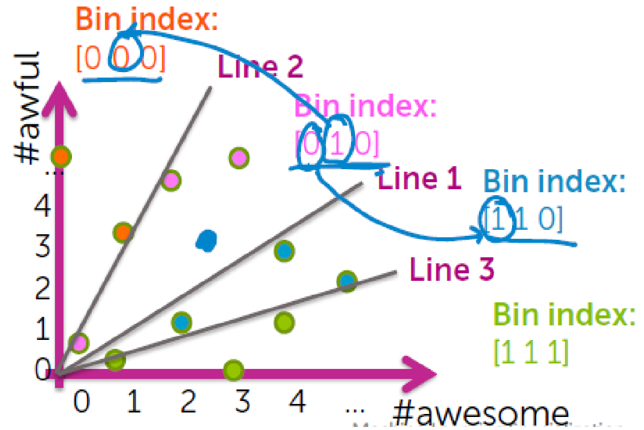


LSH: improving efficiency

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}

Next closest bins
(flip 1 bit)

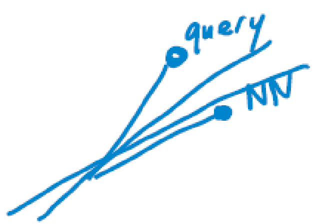


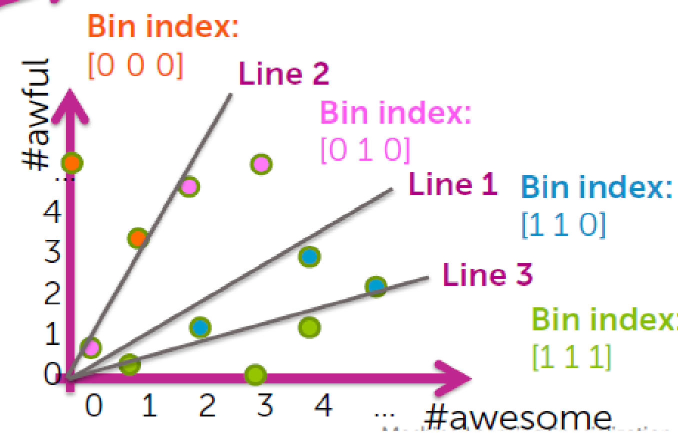
LSH: improving efficiency

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)





LSH: improving efficiency

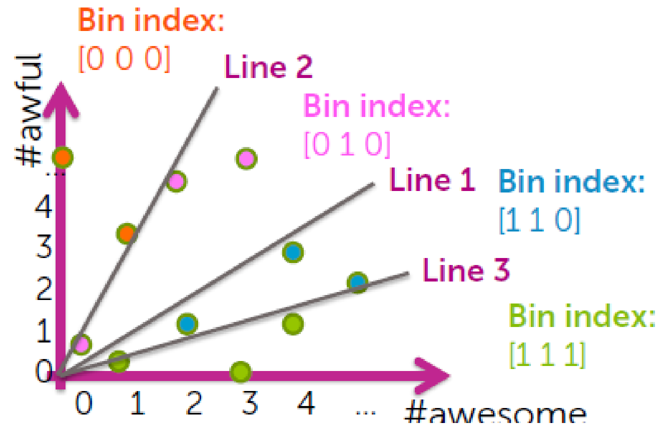
Improving search quality by searching neighboring bins

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}

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

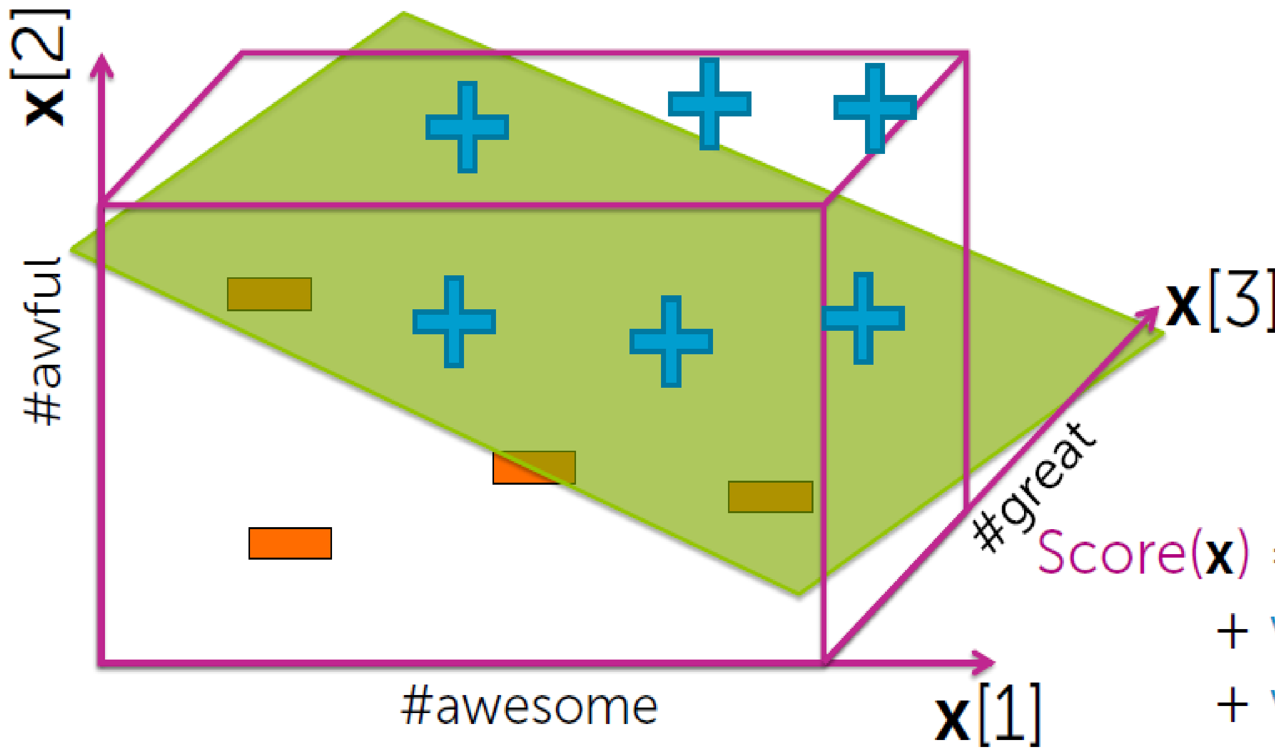
kd-tree competitor data structure

- 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 **x**, search bin(**x**), then neighboring bins until time limit

LSH: moving to higher dimmensions d

Draw random *planes*



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

LSH: moving to higher dimmensions d

Cost of binning points in d-dim

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

i th 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

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

Clustering:
An unsupervised learning task

Motivation

Goal: Structure documents by topic

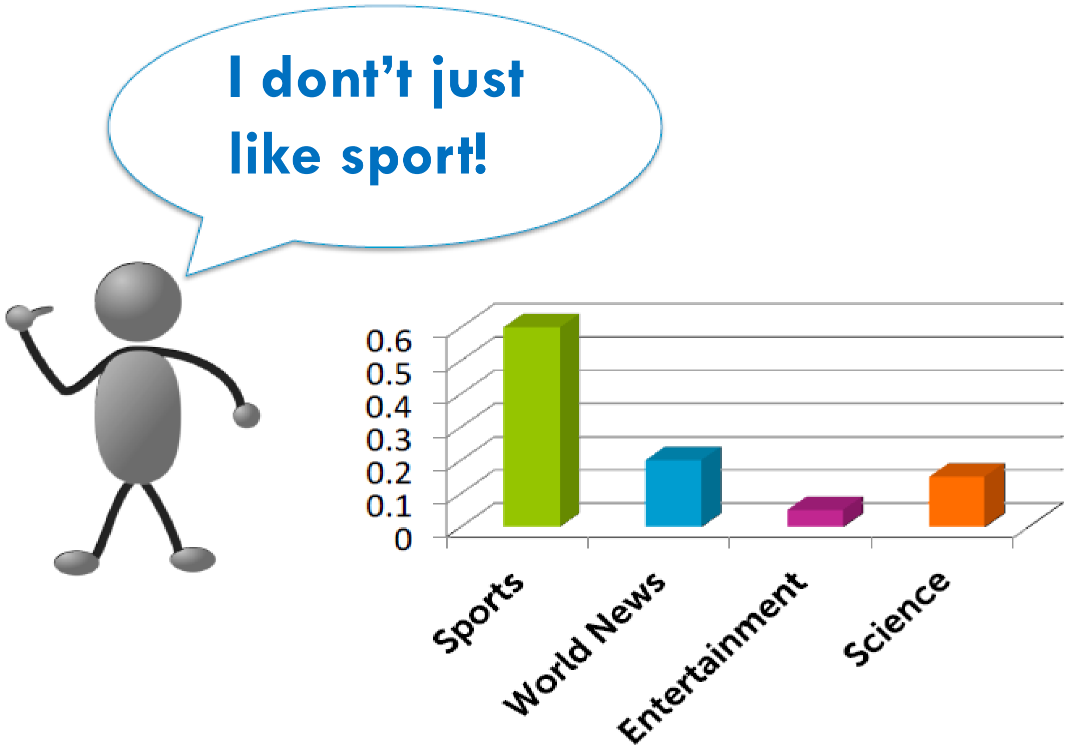
Discover groups (*clusters*) of related articles

SPORTS

WORLD NEWS

Motivation

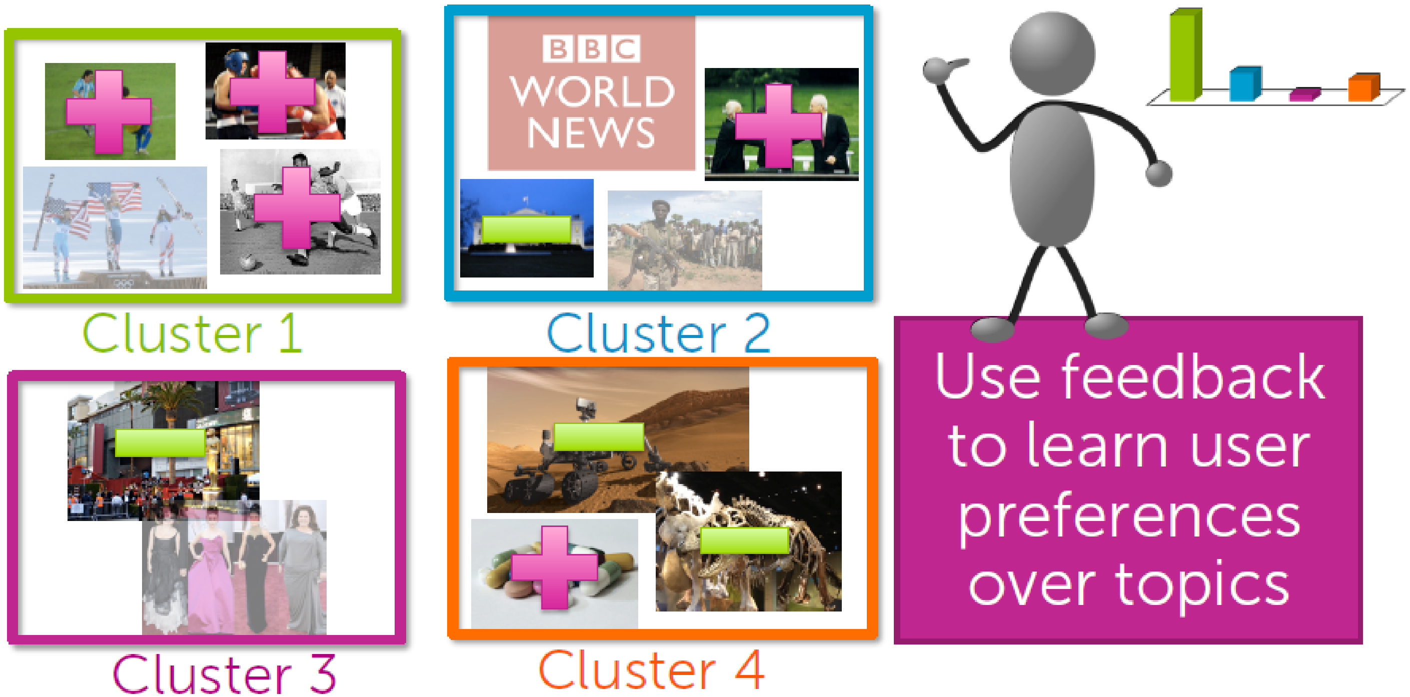
Why might clustering be useful?



Motivation

Learn user preferences

Set of clustered documents read by user

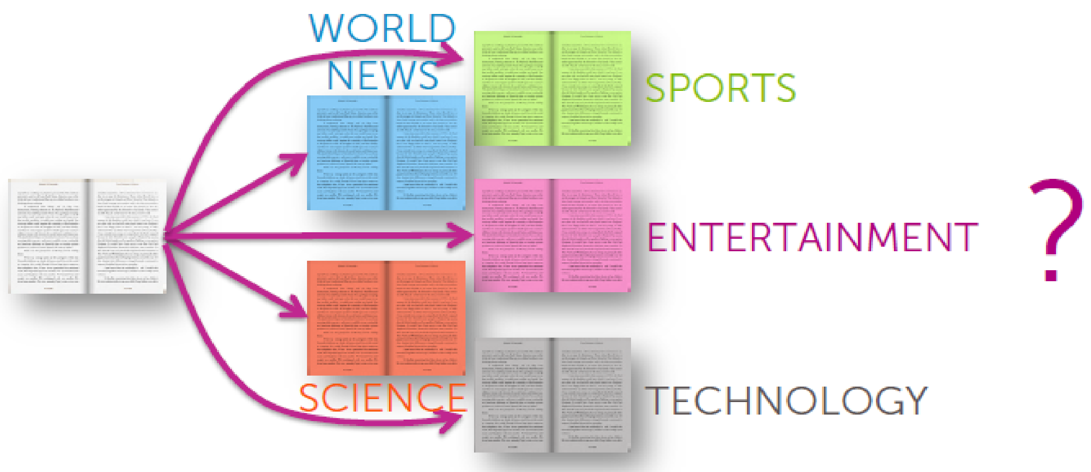


Clustering: a supervised learning

What if some of the labels are known?

Custering: a supervised learning

Multiclass classification problem



Example of
supervised learning

Clustering: an unsupervised learning

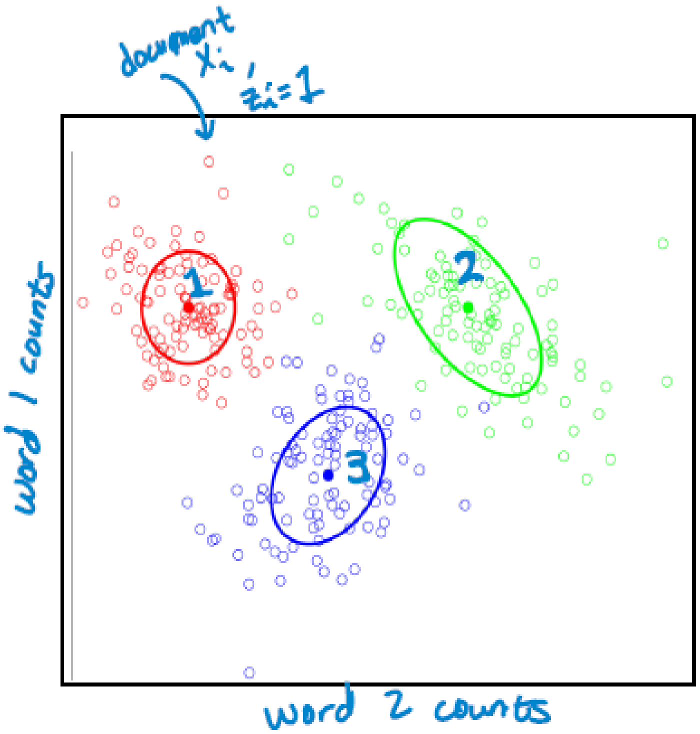
No labels provided

...uncover cluster structure from input alone

Input: docs as vectors $\mathbf{x}_i$

Output: cluster labels z_i

An unsupervised learning task

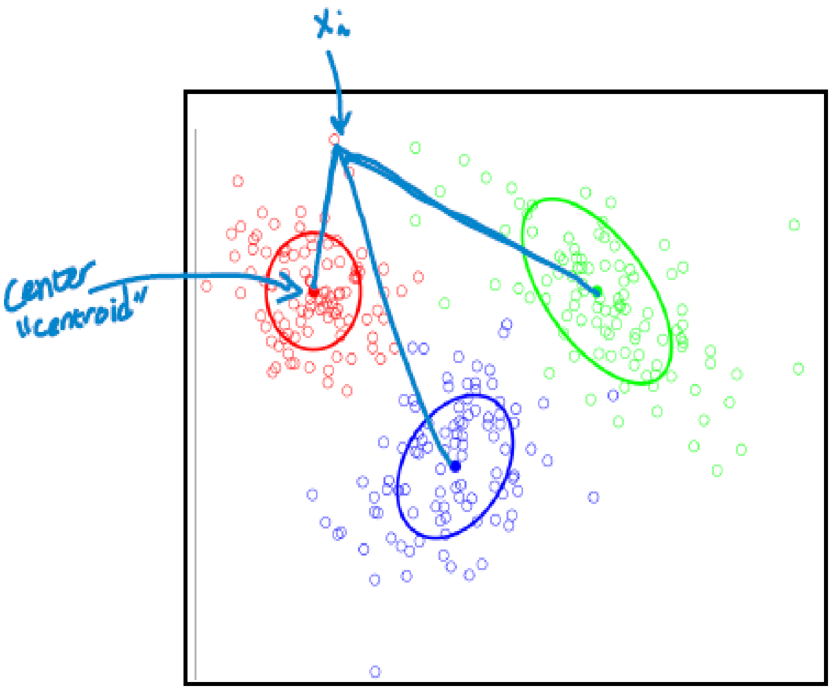


What defines a cluster ?

Cluster defined by
center & shape/spread

Assign observation $\mathbf{x}_i$ (doc) to cluster k (topic label) if

- Score under cluster k is higher than under others
- For simplicity, often define score as distance to cluster center (ignoring shape)

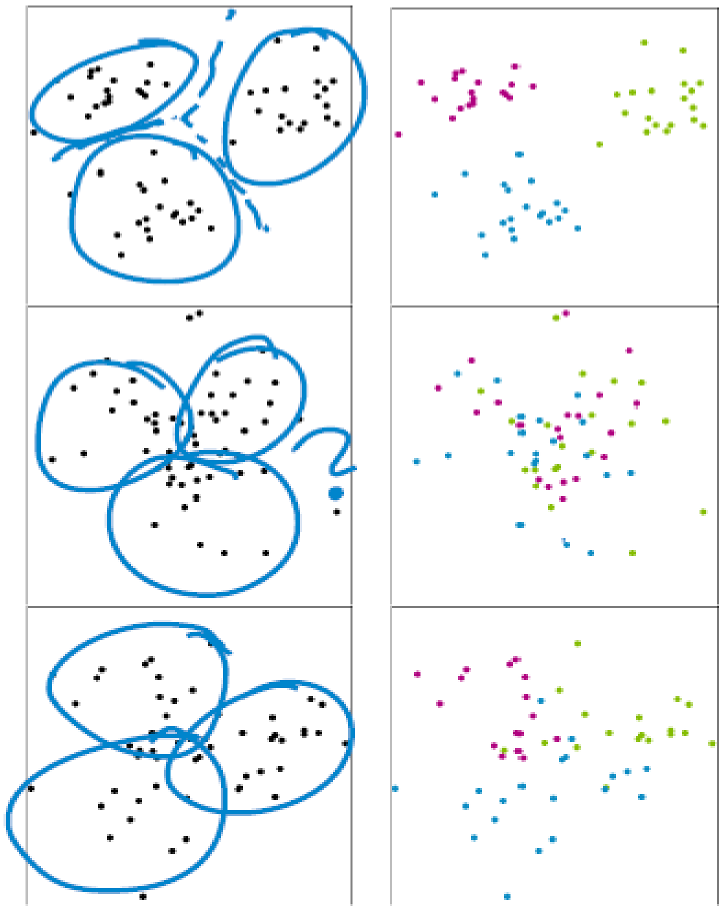


Hope for unsupervised learning

Easy

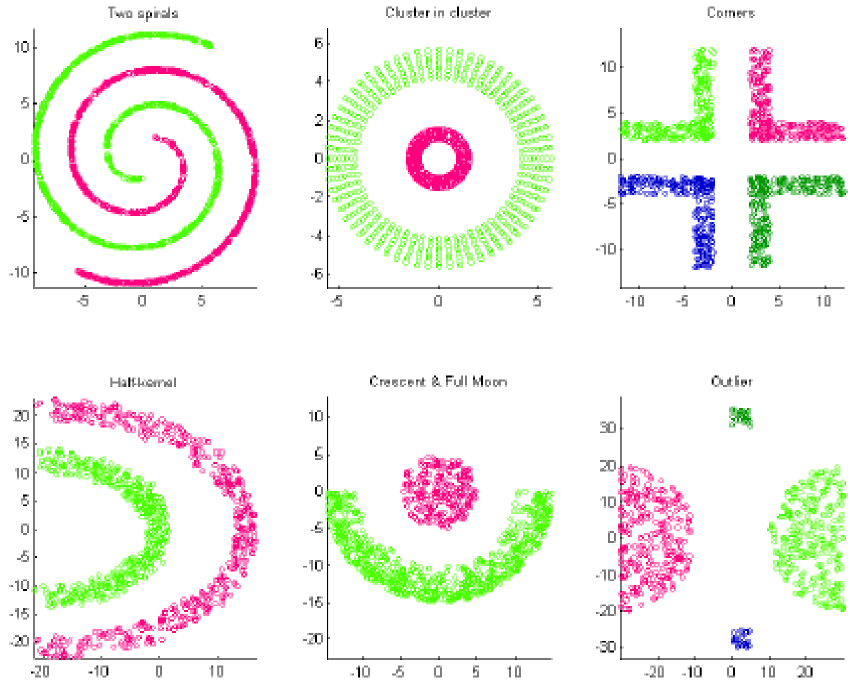
Impossible

In between



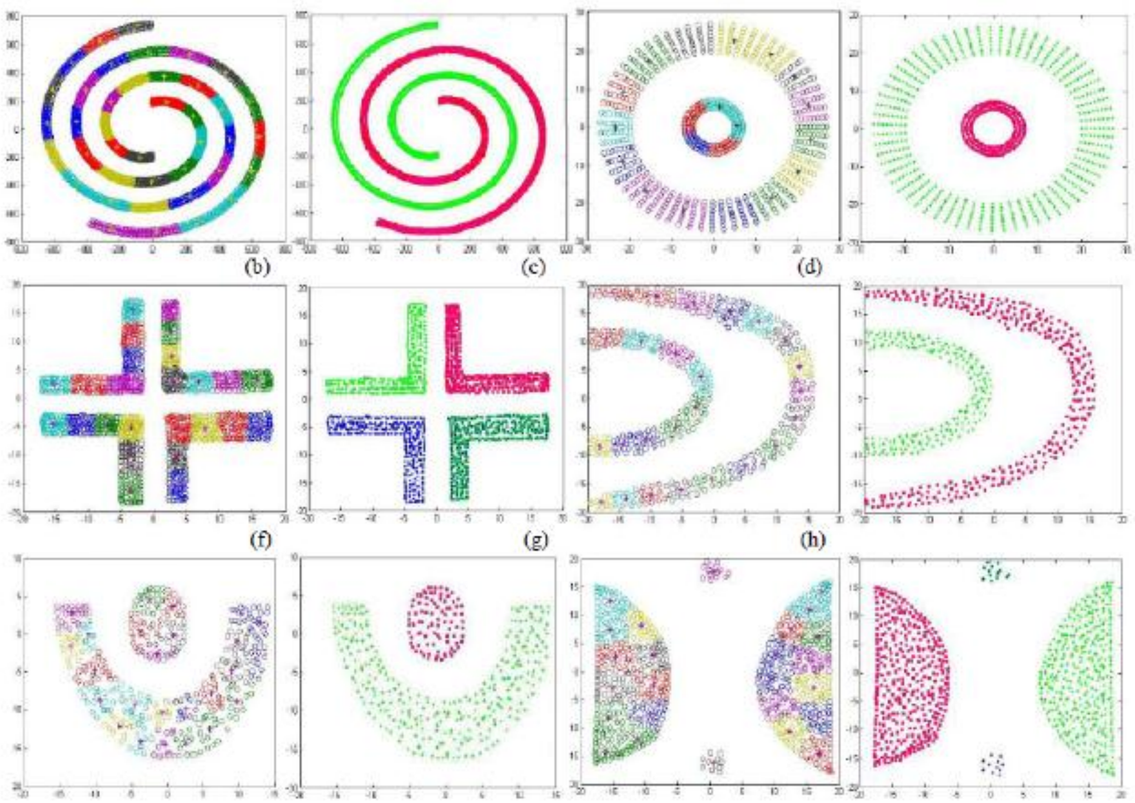
Other (challenging!) clusters to discover

Analysed by your eyes



Other (challenging!) clusters to discover

Analysed by clustering algorithms

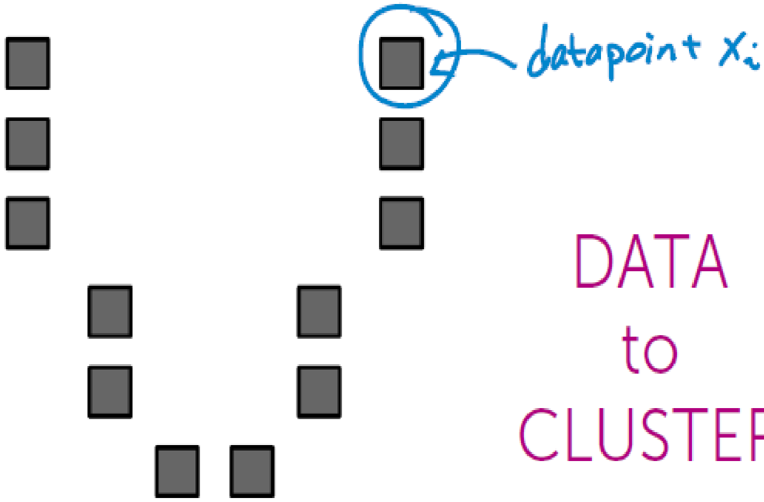


k-means clustering algorithm

k-means clustering algorithm

Assume

– Score= distance to cluster center (smaller better)

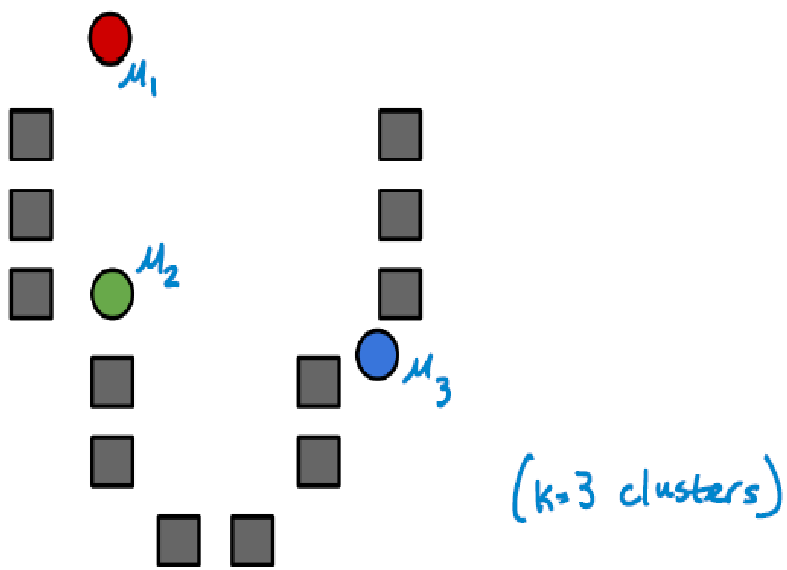


DATA to CLUSTER

k-means clustering algorithm

0. Initialize cluster centers

$$\mu_1, \mu_2, \ldots, \mu_k$$



k-means clustering algorithm

0. Initialize cluster centers
1. Assign observations to closest cluster center

$$z_i \leftarrow \arg\min_j ||\mu_j - \mathbf{x}_i||_2^2$$

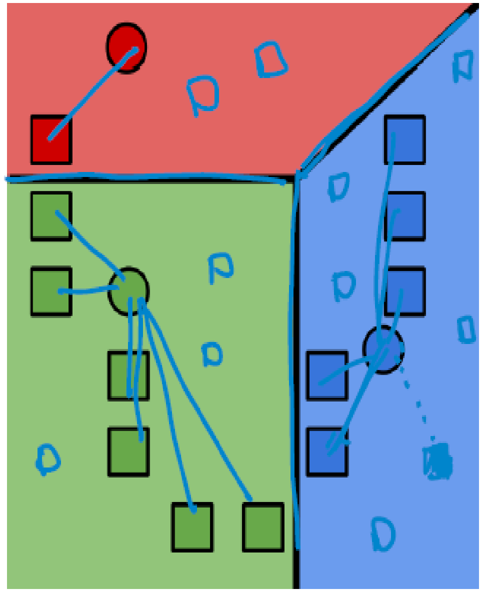
jth cluster center (varying)

ith obs. (fixed)

Inferred label for obs i, whereas supervised learning has given label y_i

return index j of the cluster whose center is closest to obs x_i

(whereas min returning minimum value of $\|\cdot\|_2^2$)



Voronoi tesselation

(for visualization only ... you don't need to compute this)

26/11, 3/12, 10/12/2019

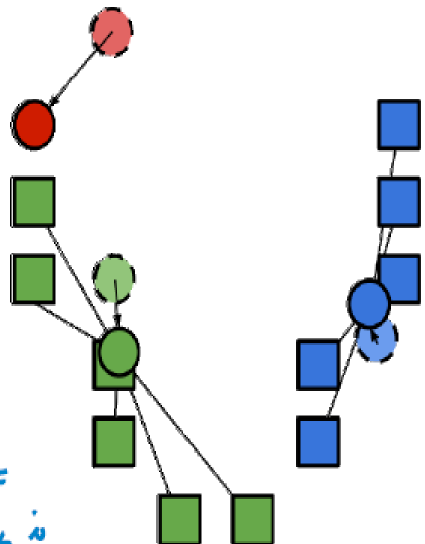
k-means clustering algorithm

0. Initialize cluster centers
1. Assign observations to closest cluster center
2. Revise cluster centers as mean of assigned observations

$$\mu_j = \frac{1}{n_j} \sum_{i:z_i=j} \mathbf{x}_i$$

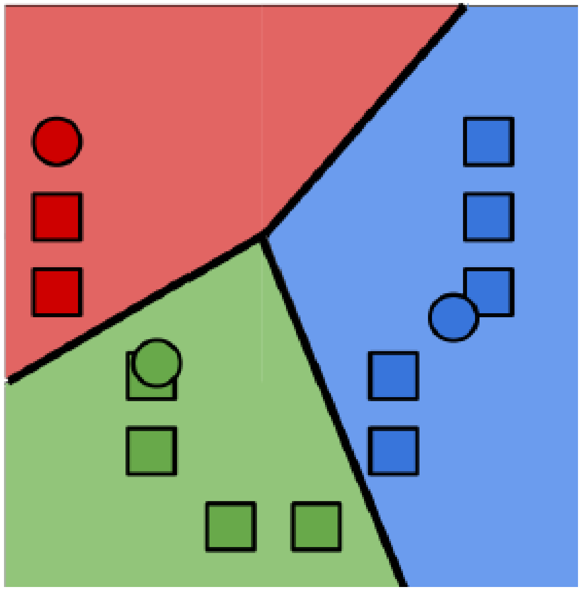
all obs. i such that $z_i = j$ (obs i is in cluster j)

# of obs. in cluster j



k-means clustering algorithm

0. Initialize cluster centers
1. Assign observations to closest cluster center
2. Revise cluster centers as mean of assigned observations
3. Repeat 1.+2. until convergence



k-means as coordinate descent algorithm

1. Assign observations to closest cluster center

$$z_i \leftarrow \arg\min_j ||\mu_j - \mathbf{x}_i||_2^2$$

2. Revise cluster centers as mean of assigned observations

$$\mu_j = \frac{1}{n_j} \sum_{i:z_i=j} \mathbf{x}_i$$

equivalent to

$$\mu_j \leftarrow \arg\min_\mu \sum_{i:z_i=j} ||\mu - \mathbf{x}_i||_2^2$$

K-means as coordinate descent algorithm

1. Assign observations to closest cluster center

$$z_i \leftarrow \arg\min_j ||\mu_j - \mathbf{x}_i||_2^2$$

2. Revise cluster centers as mean of assigned observations

$$\mu_j \leftarrow \arg\min_\mu \sum_{i:z_i=j} ||\mu - \mathbf{x}_i||_2^2$$

Alternating minimization
1. (z given μ) and 2. (μ given z)
= **coordinate descent**

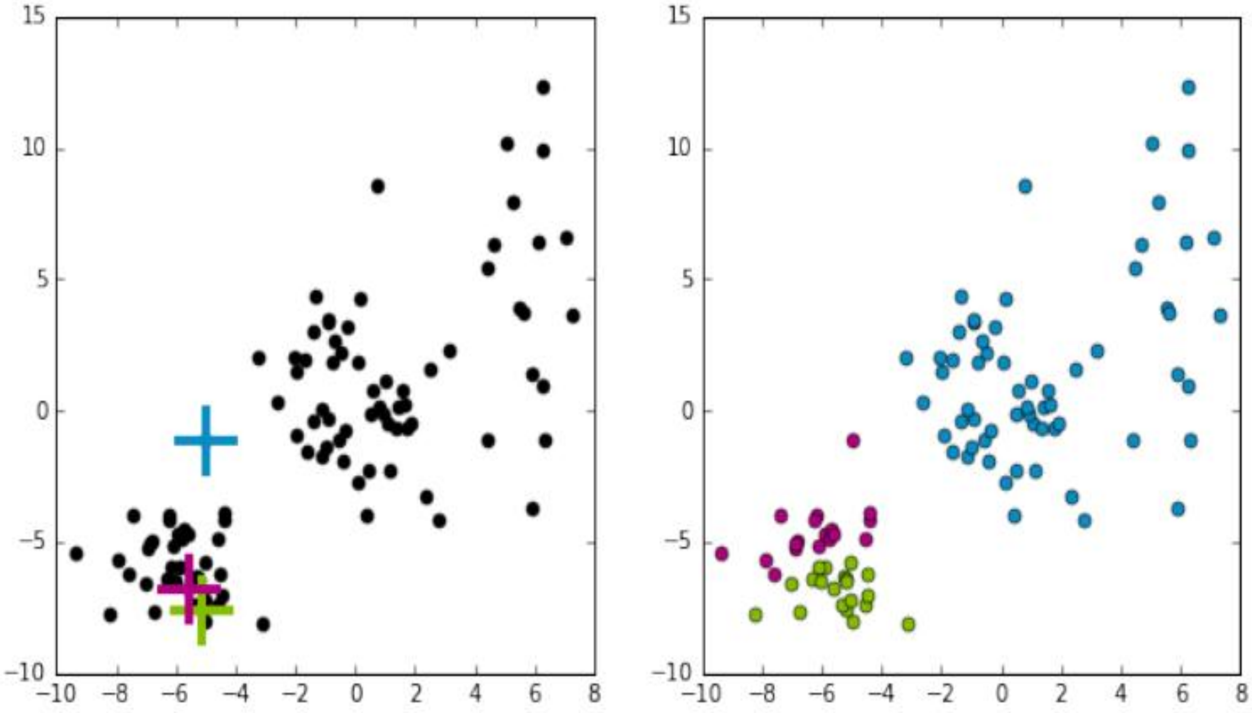
Convergence of k-means

Converges to:

- ~~Global optimum~~
- $\boxed{\text{Local optimum}}$
- ~~neither~~

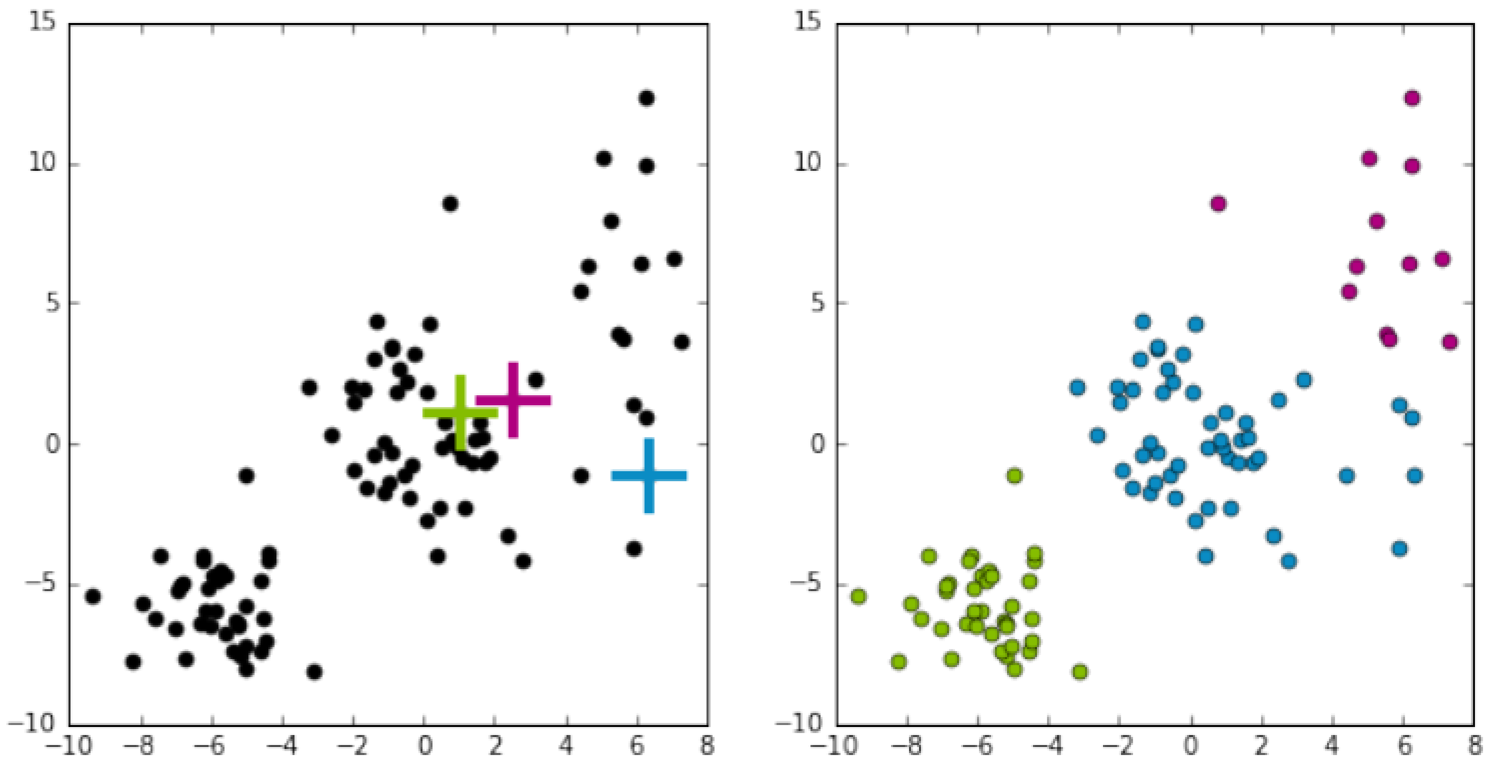
Because we can cast k-means as coordinate descent algorithm we know that we are converging to local optimum

Convergence of k-mans to local mode



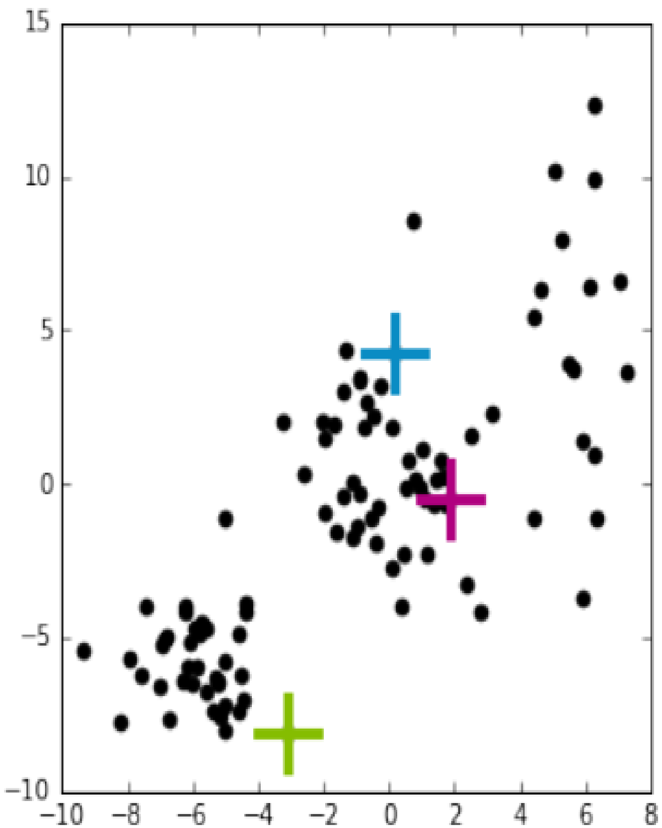
Crosses: initialised centers

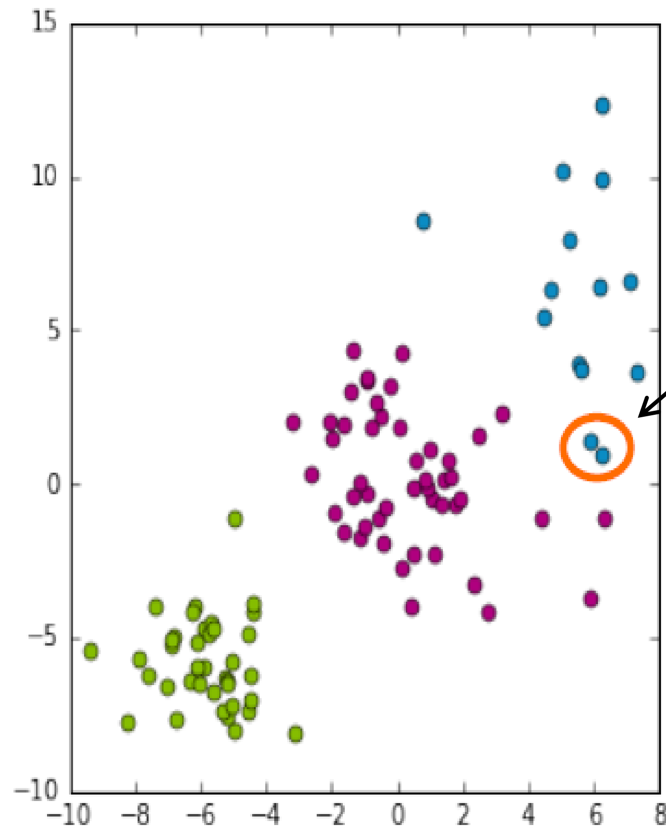
Convergence of k-mans to local mode



Crosses: initialised centers

Convergence of k-mans to local mode





Assigment to which group has changed

k-means very sensitive to initiased centers

Crosses: initialised centers

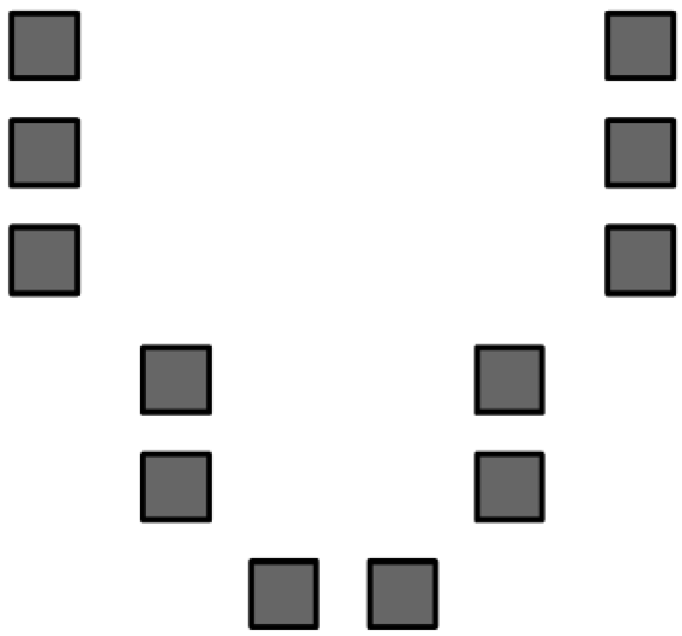
Smart initialisation: k-means++ overwiew

Initialization of k-means algorithm is critical to quality of local optima found

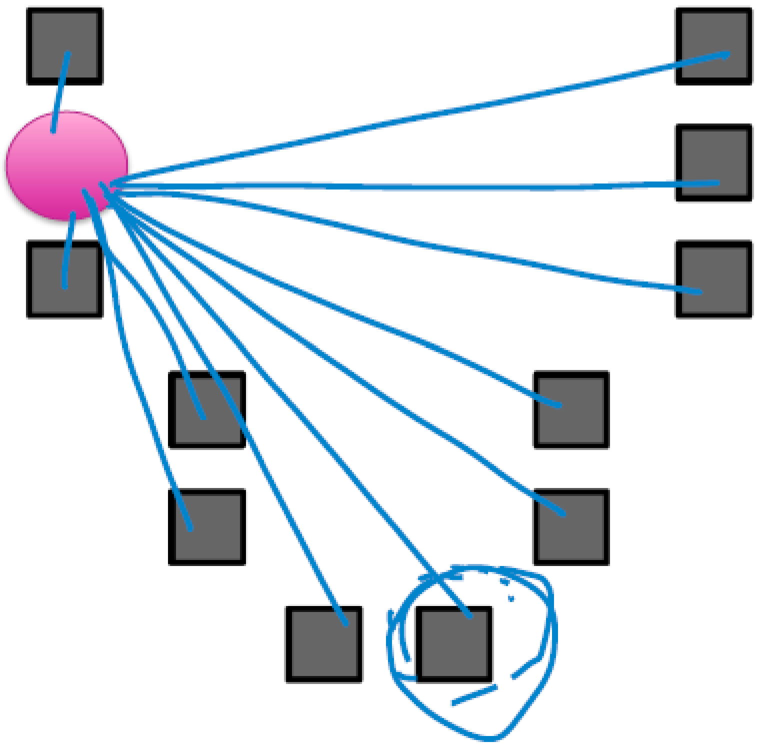
Smart initialization:

1. Choose first cluster center uniformly at random from data points
2. For each obs **x**, compute distance d(**x**) to nearest cluster center
3. Choose new cluster center from amongst data points, with probability of **x** being chosen proportional to $d(\mathbf{x})^2$
4. Repeat Steps 2 and 3 until k centers have been chosen

k-means++ visualised

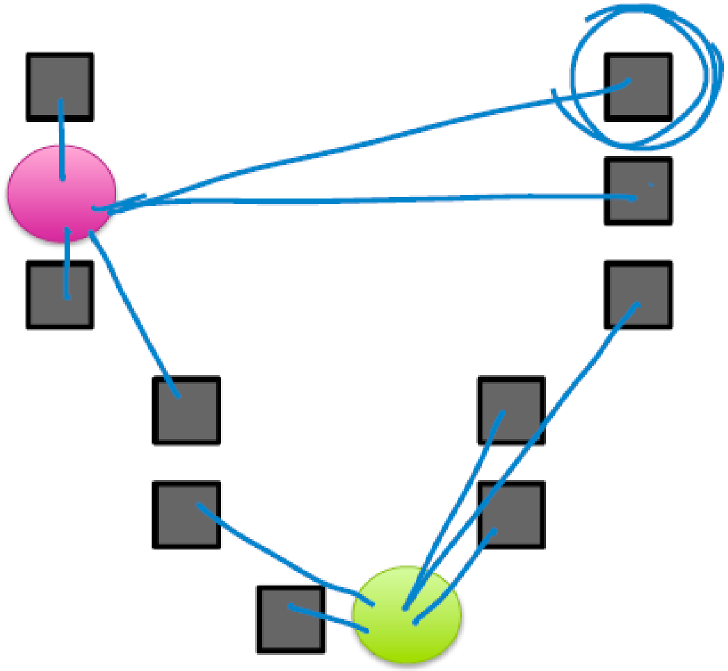


k-means++ visualised

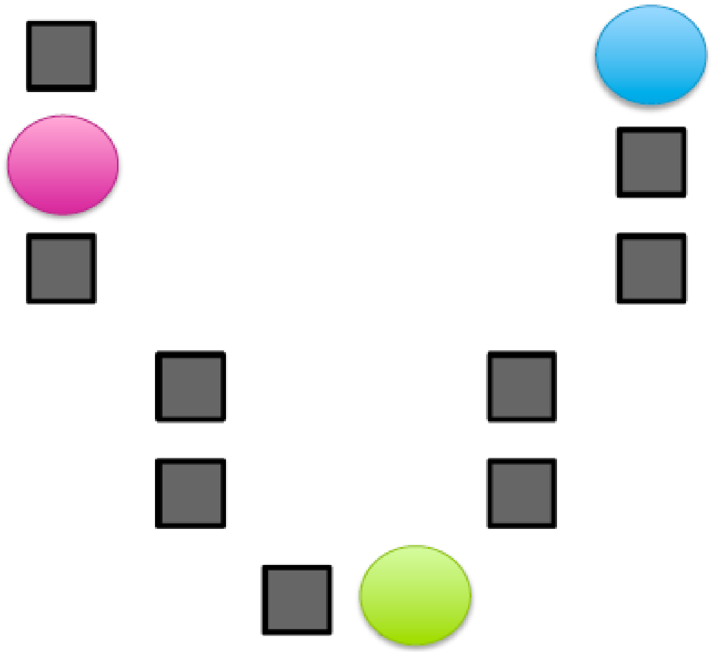


more likely to select a datapoint as a cluster center if that datapoint is far away ($dist^2$ increases this effect)

k-means++ visualised



k-means++ visualised



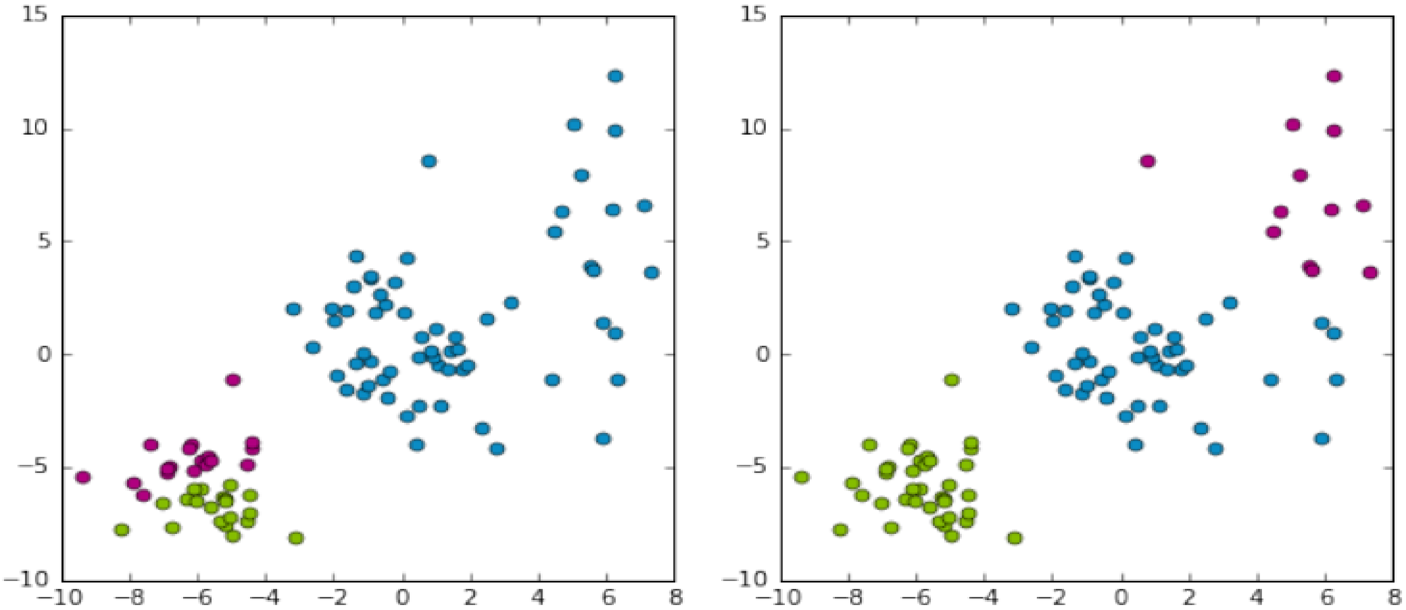
k-means++ pros/cons

Computationally costly relative to random initialization, but the subsequent k-means often converges more rapidly

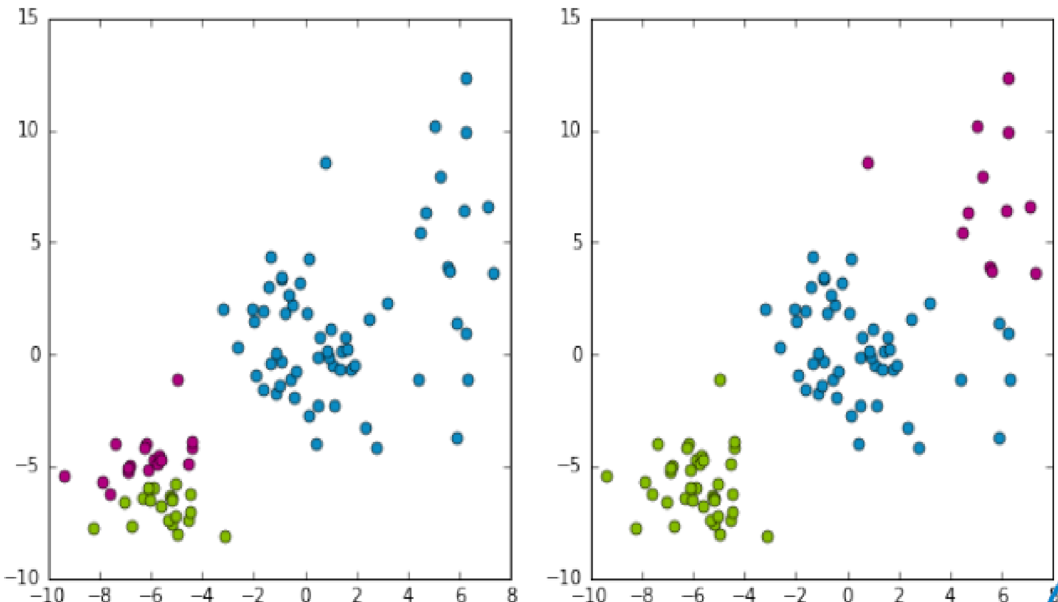
Tends to improve quality of local optimum and lower runtime

Assessing quality of the clustering

Which clustering do I prefer?



k-means objective



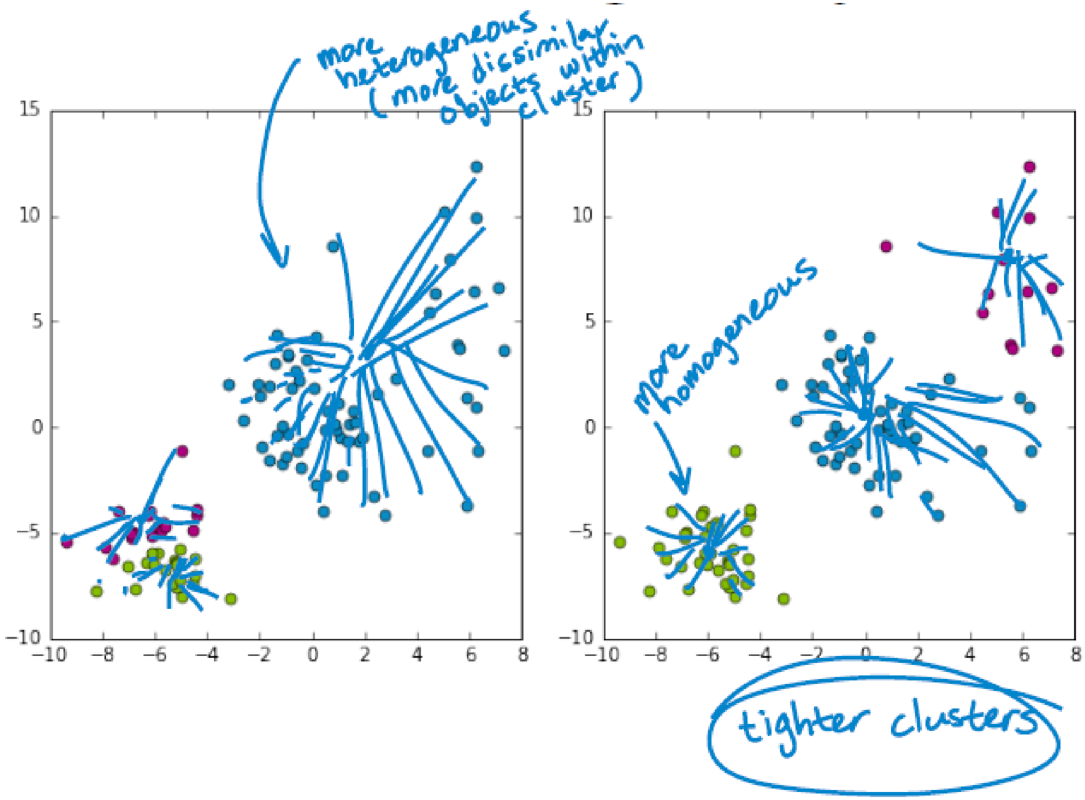
k-means is trying to minimize the sum of squared distances:

sum over all clusters

sum of squared distances in cluster j

$$\min_{\{z_i\},\{\mu_j\}} \sum_{j=1}^{k} \sum_{i:z_i=j} ||\mu_j - \mathbf{x}_i||_2^2$$

Cluster heterogeneity



Measure of quality of given clustering:

$$\sum_{j=1}^{k} \sum_{i:z_i=j} ||\mu_j - \mathbf{x}_i||_2^2$$

Lower is better!

What happens to heterogeneity as k increases?

Can refine clusters more and more to the data

→ overfitting!

# of observations

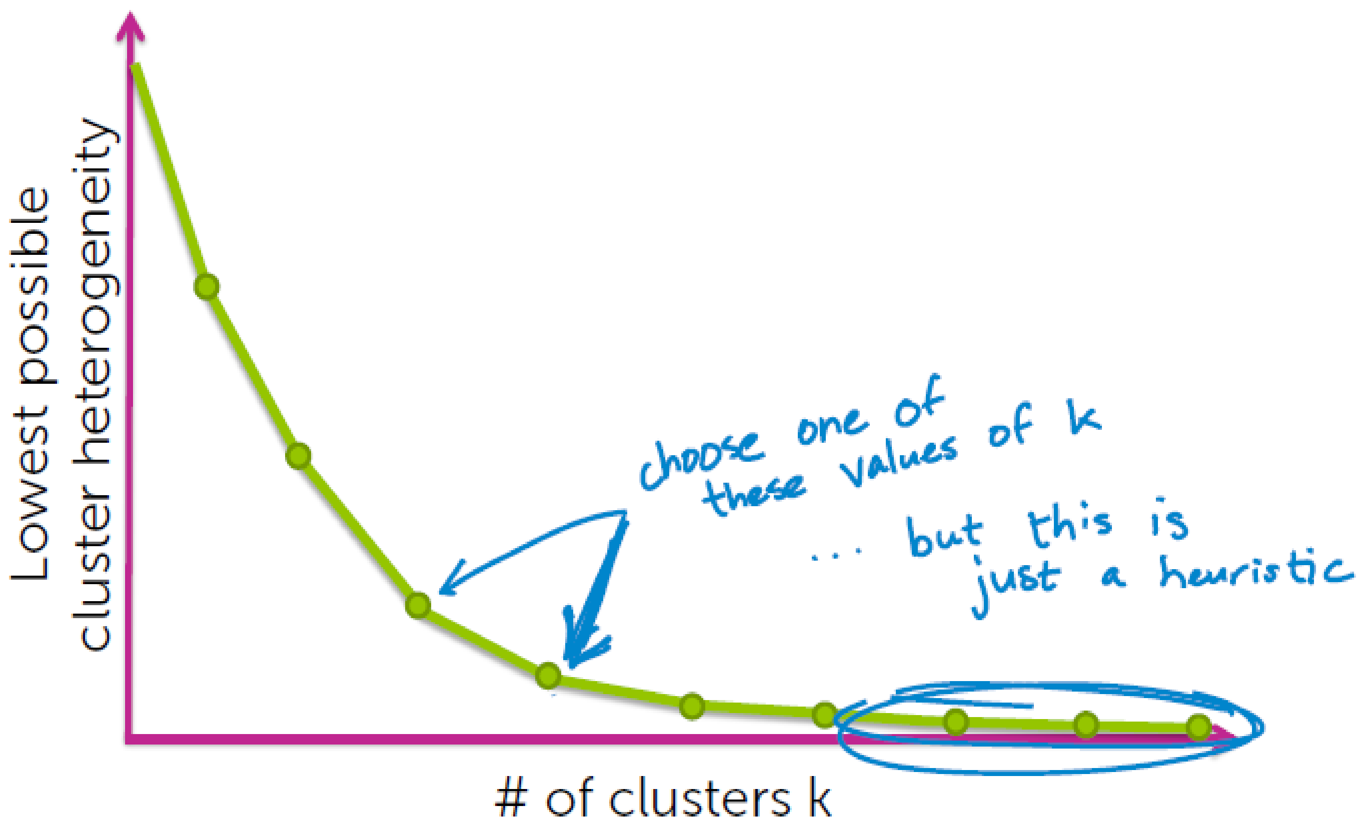
Extreme case of k=N:

– can set each cluster center equal to datapoint

– heterogeneity = 0 ! (all distances to cluster centers are 0)

Lowest possible cluster heterogeneity decreases with increasing k

How to choose k?



MapReduce

Counting words on a single processor

(The "Hello World!" of MapReduce)

Suppose you have 10B documents and 1 machine and want to count the # of occurrences of each word in the corpus

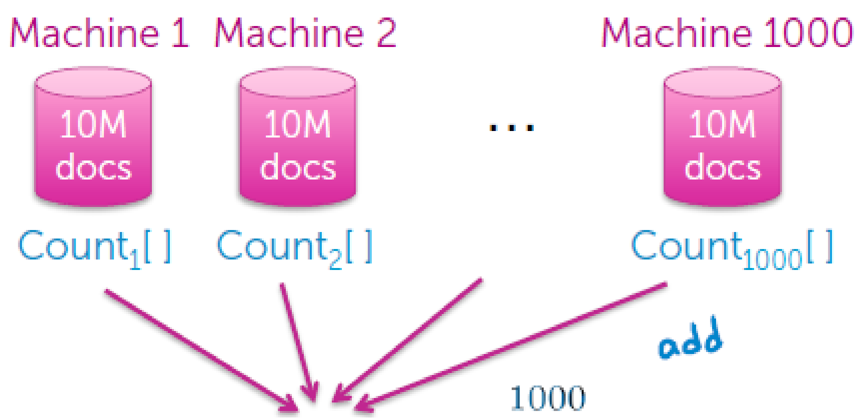
Code:

```
count[ ] ← init hash table
   for d in documents
      for word in d
         count[word] += 1
```

Naive parallel word counting

- Word counts are independent across documents $\boxed{\text{(data parallel)}}$
- Count occurrences in sets of documents separately, then merge

evenly distribute data



combine local hash tables:

$$\text{Count[word]} = \sum_{i=1}^{1000} \text{Count}_i\text{[word]}$$

How do we do this for all words in vocab?

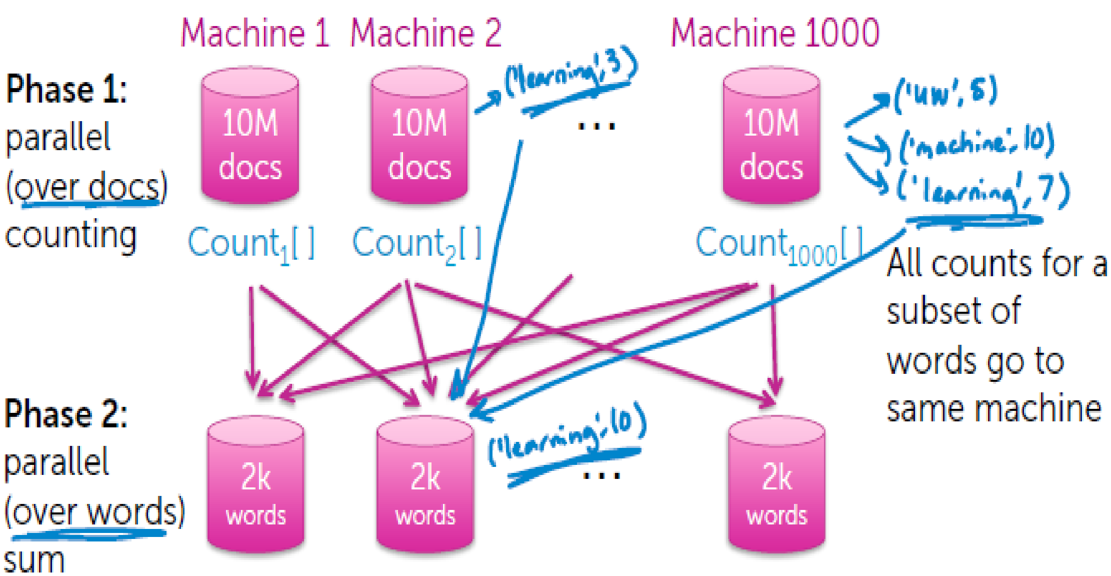
Back to sequential problem to merge counts…

have to cycle through all words in vocab… ugh.

26/11, 3/12, 10/12/2019

Counting words & merging tabels

1. Generate pairs (word,count) in parallel
2. Merge counts for each word in parallel



How to map words to machines? Use a hash function!

h(word index) → machine index

Which words go to machine i ?

$h: \underset{\text{vocab}}{V} \rightarrow [1, 2, \ldots, \#\text{machines}]$

Send counts of 'learning' to machine h['learning']

MapReduce abstraction

Map:

- Data-parallel over elements, e.g., documents
- Generate (key,value) pairs
    - "value" can be any data type

('uw',1)
('machine',1)
('uw',1)
('learning',1) ...

Reduce:

- Aggregate values for each key
- Must be commutative-associative operation

a+b = b+a   (a+b)+c = a+(b+c)

- Data-parallel over keys
- Generate (key,value) pairs

reduce('uw',[1,17,0,0,12,2])
emit('uw',32)

Word count example:

```
map(doc)
    for word in doc
    emit(word,1)
```

key   value

key   list of values

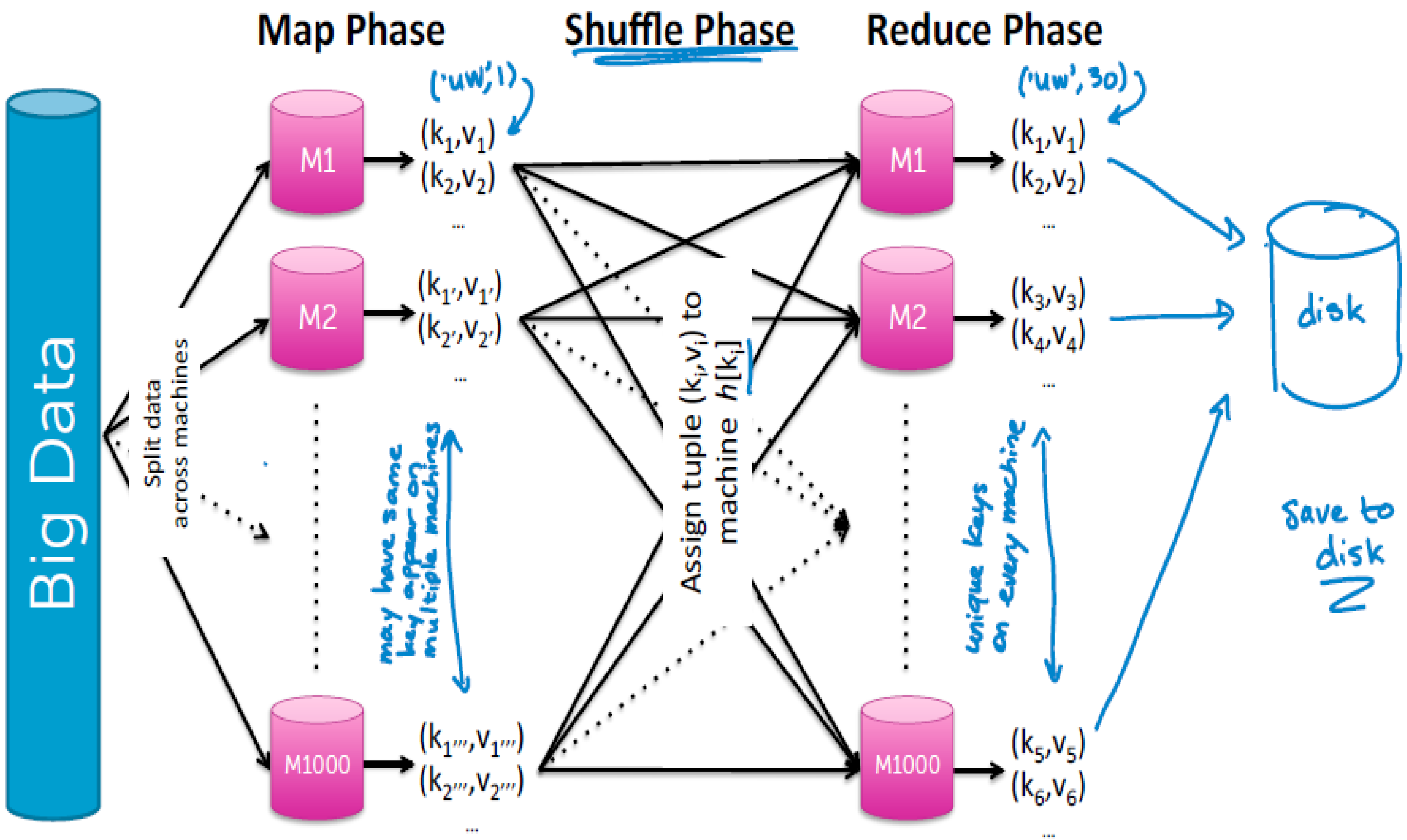
```
reduce(word, counts_list)
    c = 0
    for i in counts_list
    c += counts_list[i]
    emit(word, c)
```

key   value

MapReduce has long history in functional programming

- Popularized by Google, and subsequently by open-source Hadoop implementation from Yahoo!

MapReduce – Execution overwiew

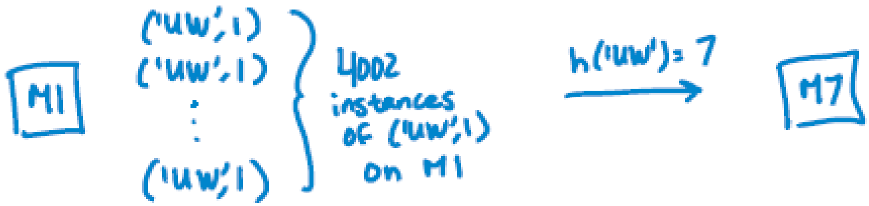


26/11, 3/12, 10/12/2019

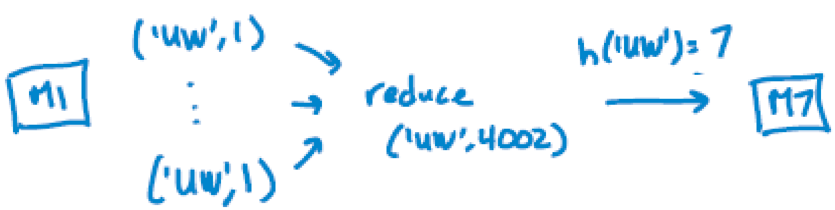
Improving performance

Combiners

- Naïve implementation of MapReduce is very wasteful in communication during shuffle:



- Combiner: Simple solution...Perform reduce locally before communicating for global reduce
  - Works because reduce is commutative-associative



Scaling up k-means via MapReduce

MapReducing 1 iteration of k-means

Classify: Assign observations to closest cluster center

$$z_i \leftarrow \arg\min_j ||\mu_j - \mathbf{x}_i||_2^2$$

Map: For each data point, given ($\{\mu_j\}$,$\mathbf{x}_i$), emit(z_i,$\mathbf{x}_i$)

Recenter: Revise cluster centers as mean of assigned observations

$$\mu_j = \frac{1}{n_j} \sum_{i:z_i=k} \mathbf{x}_i$$

Reduce: Average over all points in cluster j (z_i=k)

Scaling up k-means via MapReduce

Classification step as Map

Classify: Assign observations to closest cluster center

$$z_i \leftarrow \arg\min_j ||\mu_j - \mathbf{x}_i||_2^2$$

set of cluster centers

a datapoint

map([$\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$,…, $\boldsymbol{\mu}_k$], $\mathbf{x}_i$)

$$z_i \leftarrow \arg\min_j ||\mu_j - \mathbf{x}_i||_2^2$$

emit(z_i,$\mathbf{x}_i$)

datapoint

cluster label

$z_i=2$ (assigned to cluster 2)

datapoint x_i

e.g. emit (2, [17, 0, 1, 7, 0, 0, 5])

Scaling up k-means via MapReduce

Recenter step as Reduce

Recenter: Revise cluster centers as mean of assigned observations

$$\mu_j = \frac{1}{n_j} \sum_{i:z_i=k} \mathbf{x}_i$$

cluster label (key)

datapoints assigned to cluster j (have key j)

reduce(j, x_in_clusterj : [$\mathbf{x}_1$, $\mathbf{x}_3$,...,])
    sum = 0 ← total mass in cluster
    count = 0 ← total # of obs. in cluster
    for **x** in x_in_clusterj
        sum += **x**
        count += 1
    emit(j, sum/count)

cluster label

$\frac{\text{total mass}}{\text{total \# obs}}$

Scaling up k-means via MapReduce

Some practical considerations

k-means needs an iterative version of MapReduce

- Not standard formulation

Mapper needs to get data point and all centers

- A lot of data!
- Better implementation:
  mapper gets many data points

Parallel k-means via MapReduce

Map: classification step;

data parallel over data points

Reduce: recompute means;

data parallel over centers

What you can do now …

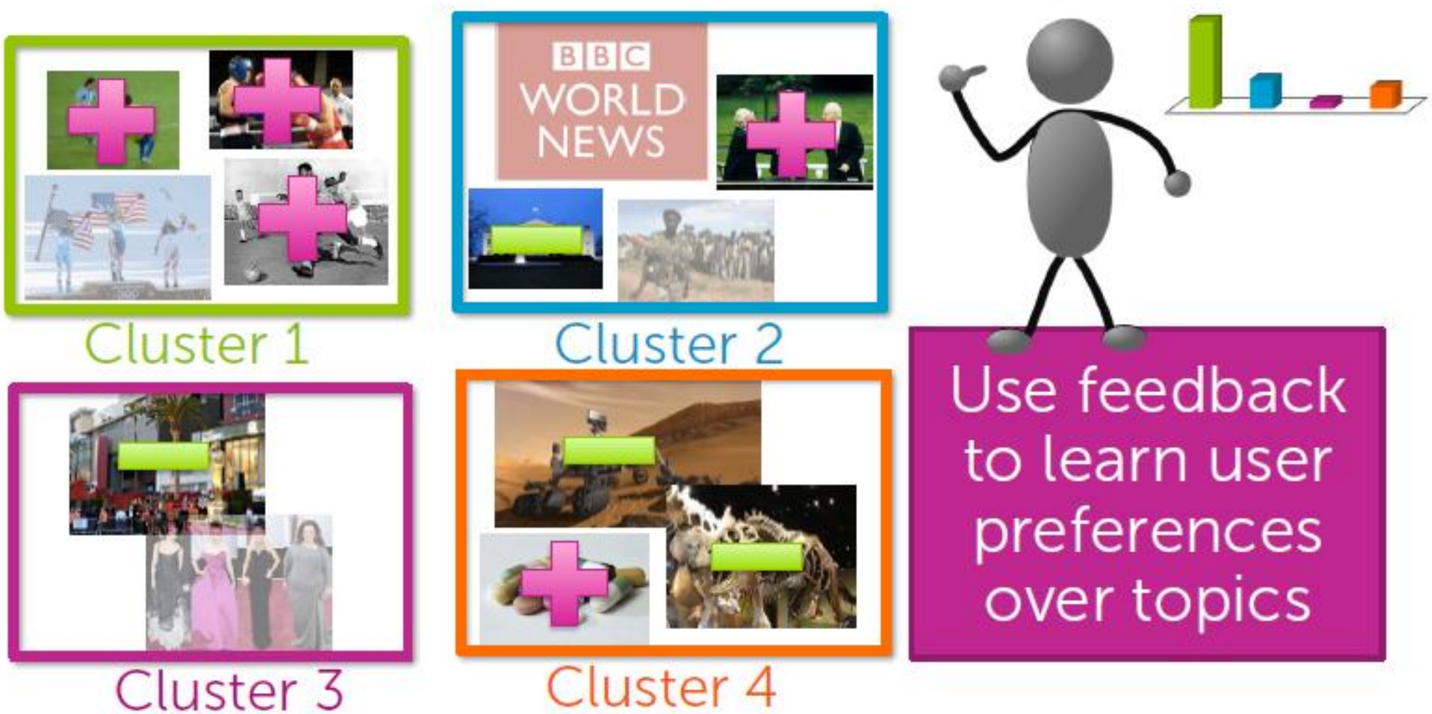
- Describe potential applications of clustering
- Describe the input (unlabeled observations) and output (labels) of a clustering algorithm
- Determine whether a task is supervised or unsupervised
- Cluster documents using k-means
- Interpret k-means as a coordinate descent algorithm
- Define data parallel problems
- Explain Map and Reduce steps of MapReduce framework
- Use existing MapReduce implementations to parallelize k-means, understanding what's being done under the hood

Probabilistic approach: mixture model

Why probabilistic approach?

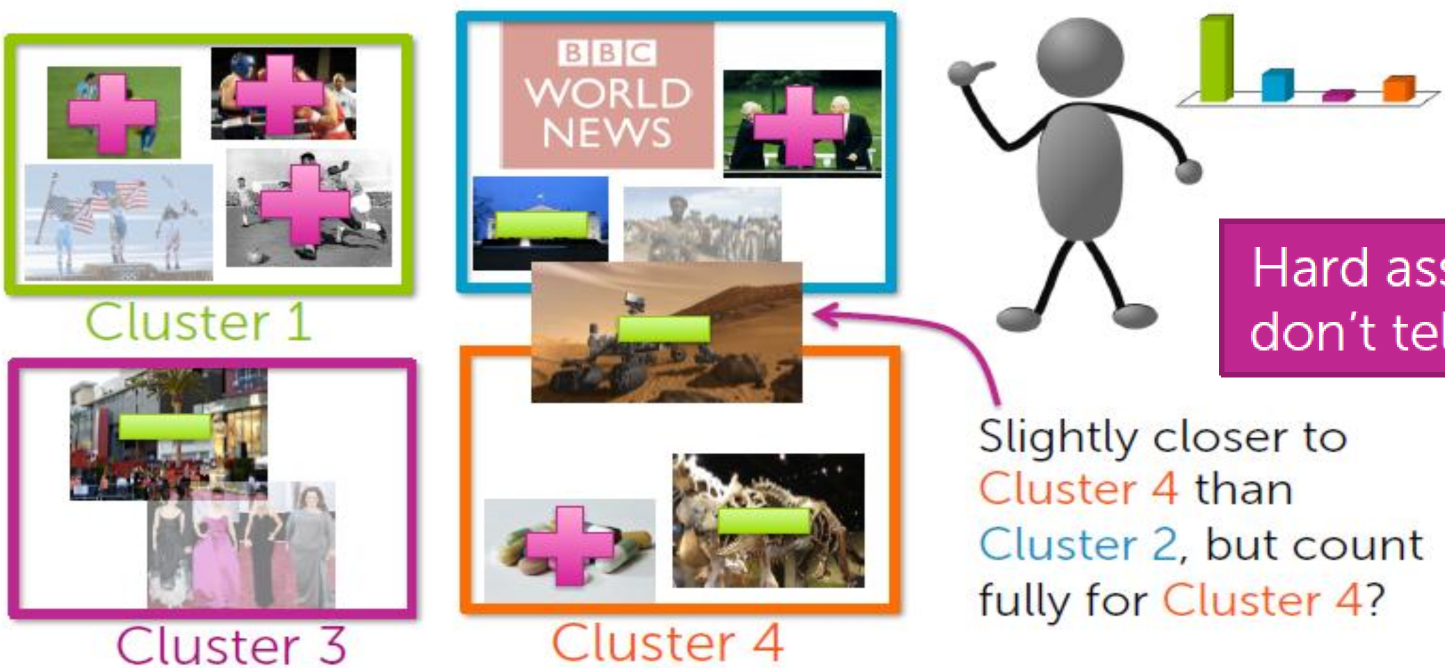
Learn user preferences

Set of clustered documents read by user



Why probabilistic approach?

Uncertainty in cluster assignments



Hard assignments don't tell full story

Slightly closer to Cluster 4 than Cluster 2, but count fully for Cluster 4?

Why probabilistic approach?

Other limitations of k-means

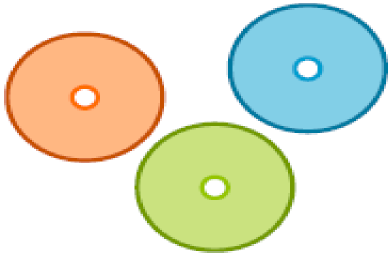
Assign observations to closest cluster center

$$z_i \leftarrow \arg\min_j ||\mu_j - \mathbf{x}_i||_2^2$$

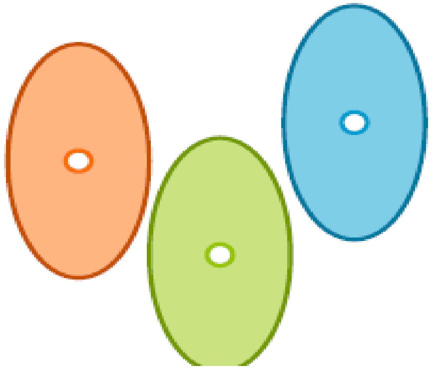
Can use weighted Euclidean, but requires *known* weights

Only center matters

Equivalent to assuming spherically symmetric clusters

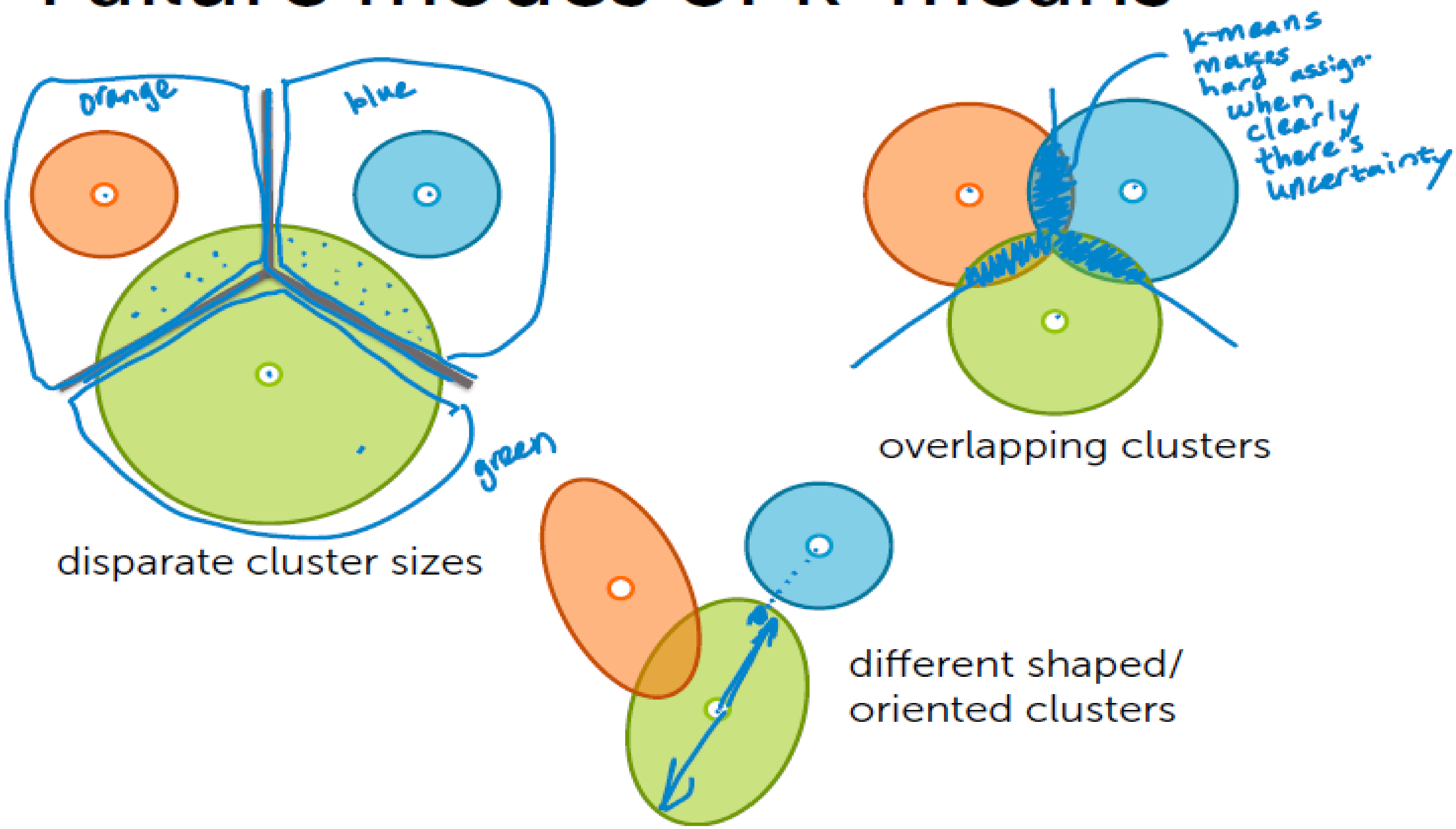


Still assumes all clusters have the same axis-aligned ellipses



Why probabilistic approach?

Failure modes of k-means



26/11, 3/12, 10/12/2019

Mixture models

- Provides **soft assignments** of observations to clusters (uncertainty in assignment)
  - e.g., 54% chance document is world news, 45% science, 1% sports, and 0% entertainment
- Accounts for cluster shapes not just centers
- Enables learning weightings of dimensions
  - e.g., how much to weight each word in the vocabulary when computing cluster assignment

Application: clustering images

Discover groups of similar images

- Ocean
- Pink flower
- Dog
- Sunset
- Clouds
- ...

Application: clustering images

Simple image representation

Consider average red, green, blue pixel intensities

[R = 0.05, G = 0.7, B = 0.9]

[R = 0.85, G = 0.05, B = 0.35]

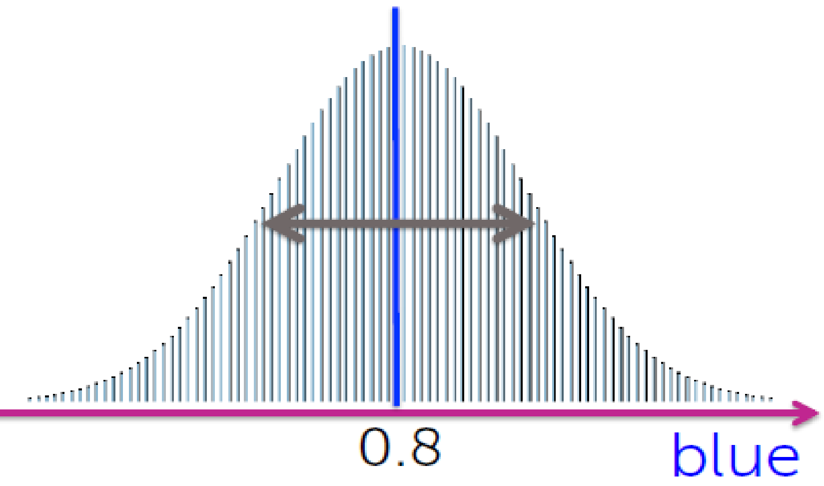
[R = 0.02, G = 0.95, B = 0.4]

Single RGB vector per image

Application: clustering images

Distribution over all cloud images

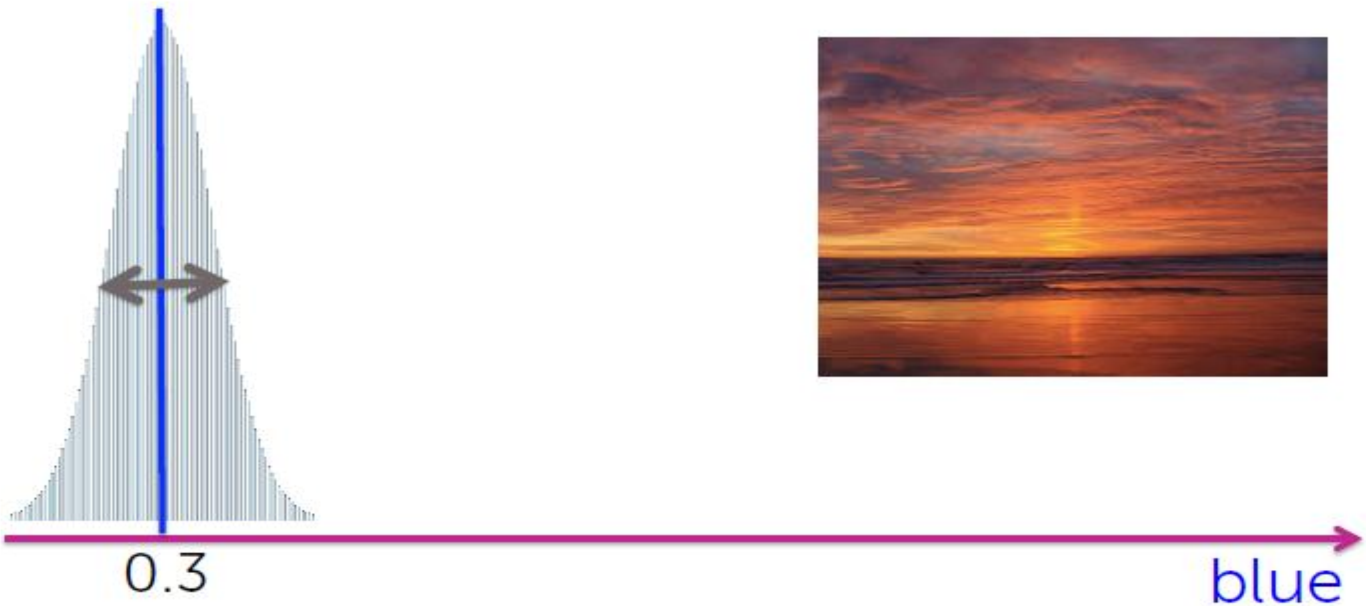
Let's look at just the blue dimension



Application: clustering images

Distribution over all sunset images

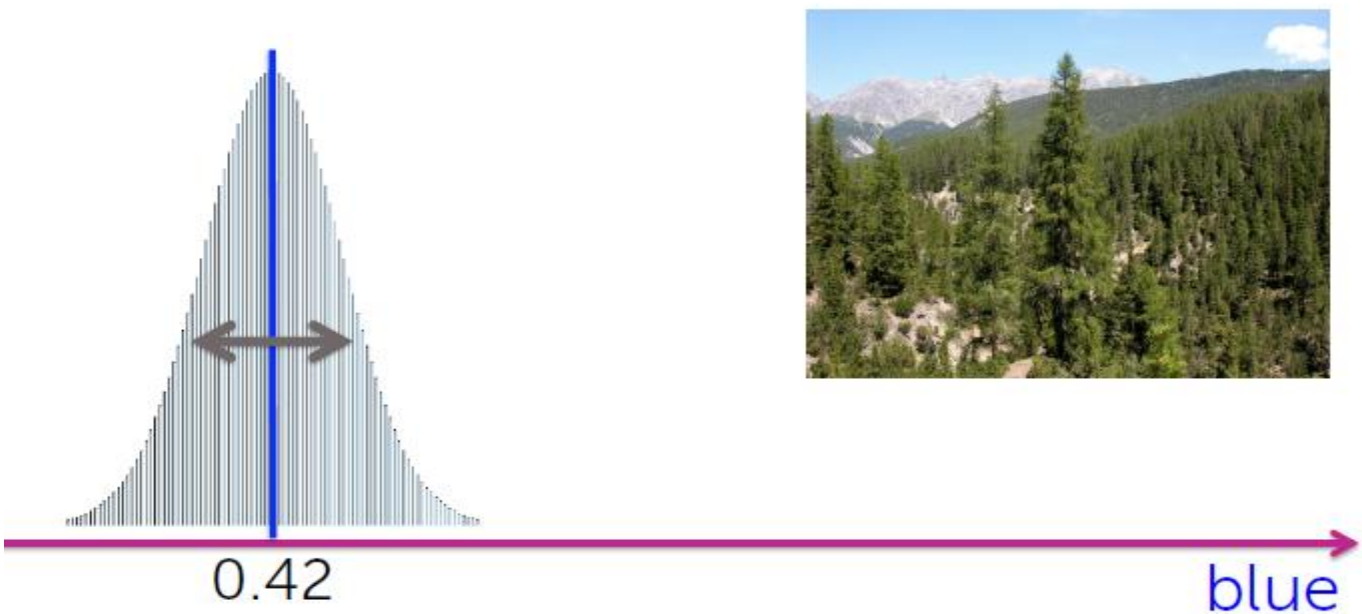
Let's look at just the blue dimension



Application: clustering images

Distribution over all forest images

Let's look at just the blue dimension



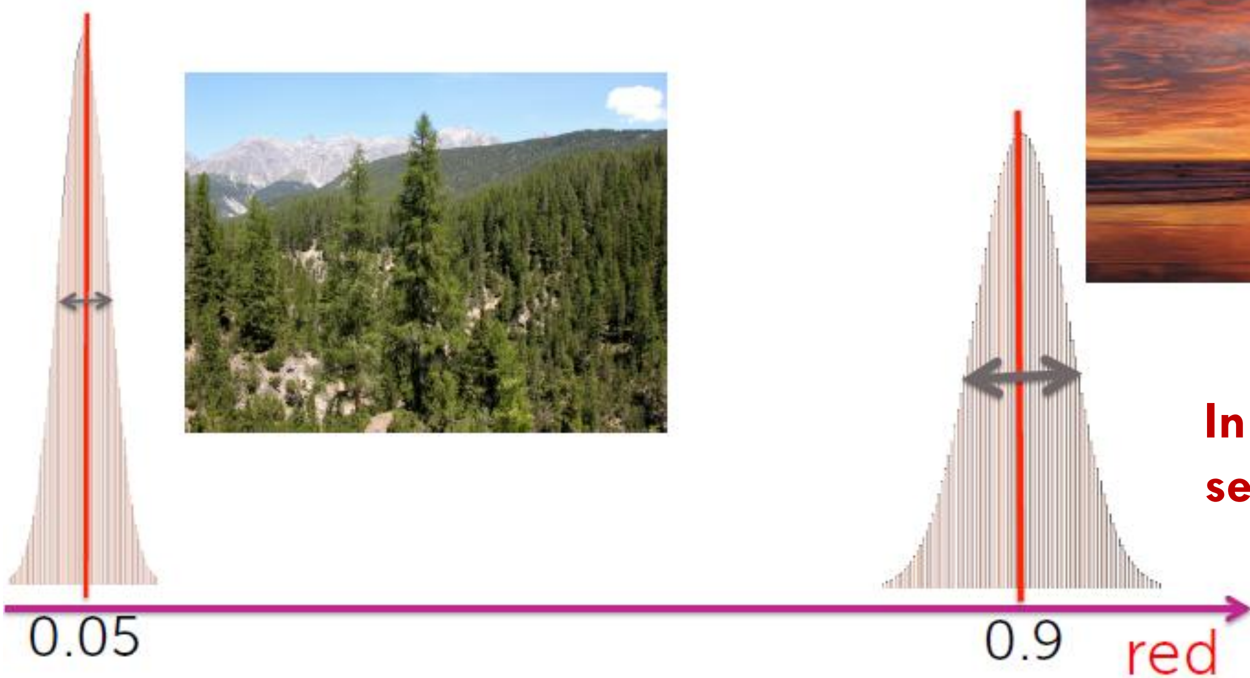
Application: clustering images

Distribution over all images

Application: clustering images

Can be distinguished along other dim

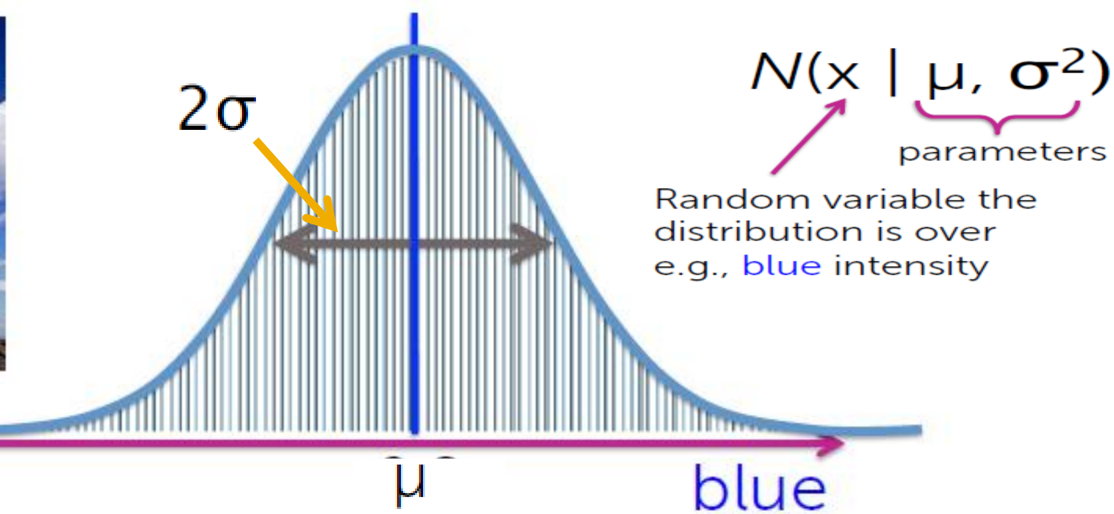
Now look at the red dimension



In this dimmension separable groups!

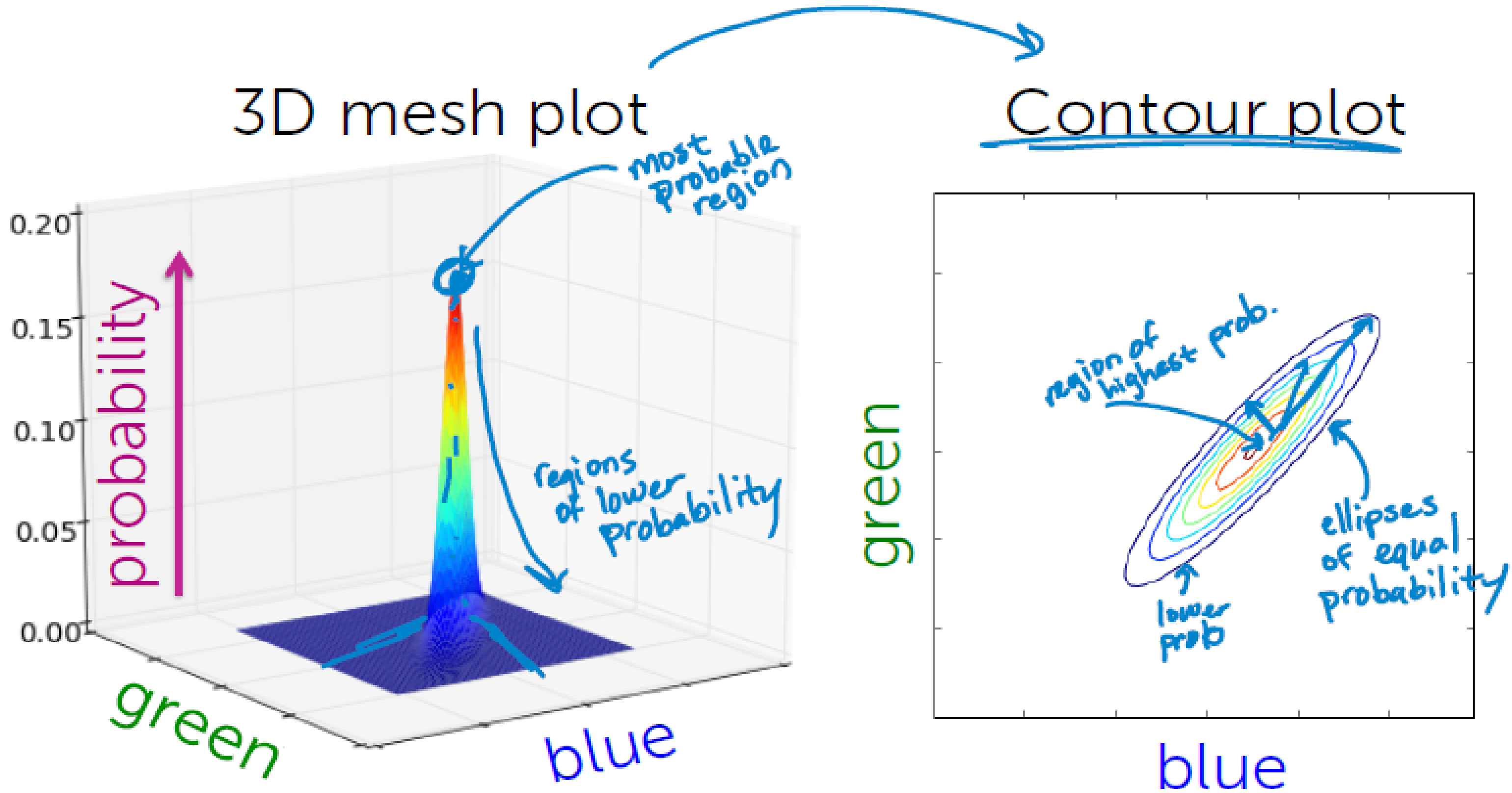
Model for a given image type

For **each dimension** of the [R, G, B] vector, and **each image type**, assume a Gaussian distribution over color intensity



Model for a given image type

2D Gaussians – Bird's eye view

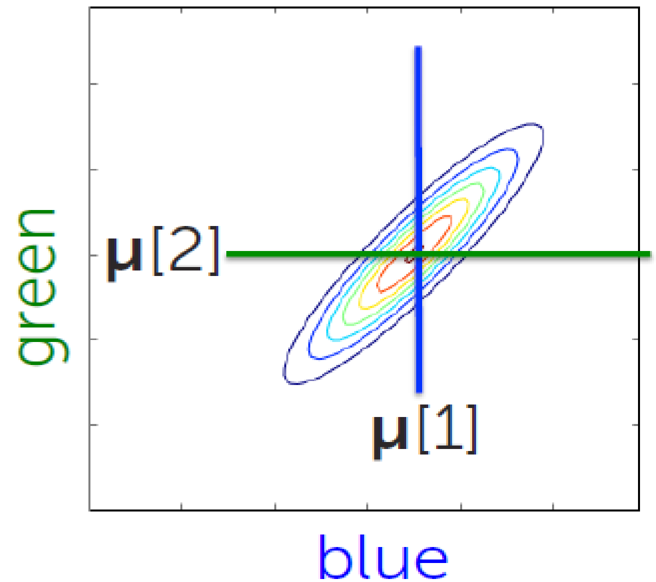


2D Gaussians – Parameters

Fully specified by **mean μ** and **covariance Σ**

$\boldsymbol{\mu} = [\mu_{blue}, \mu_{green}]$

mean centers the distribution in 2D



Application: clustering images

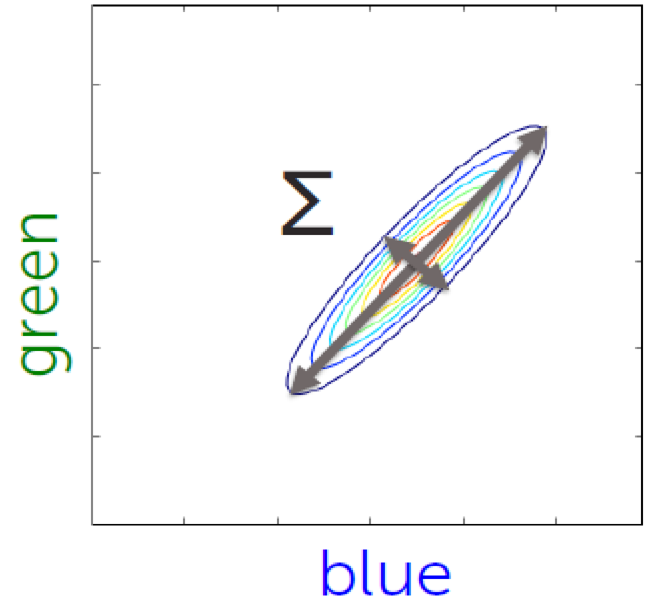
2D Gaussians – Parameters

Fully specified by **mean** μ and **covariance** Σ

$\boldsymbol{\mu} = [\mu_{blue}, \mu_{green}]$

$$\Sigma = \begin{pmatrix} \sigma_{blue}^2 & \sigma_{blue,green} \\ \sigma_{green,blue} & \sigma_{green}^2 \end{pmatrix}$$

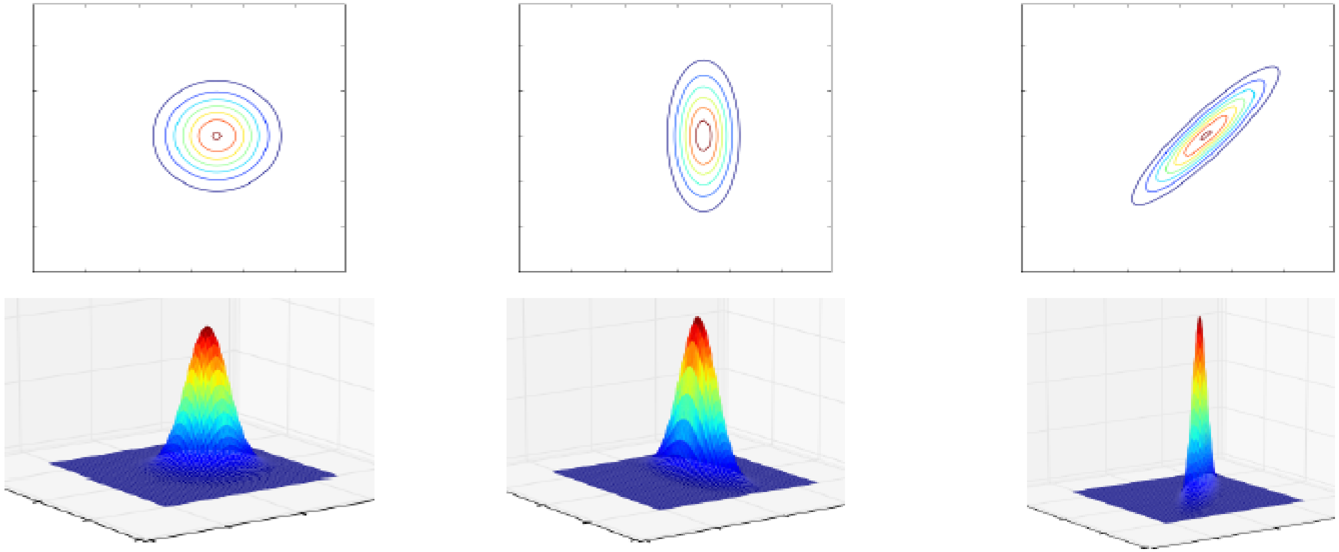
covariance determines orientation + spread



Application: clustering images

Covariance structures

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_B{}^2 & 0 \\ 0 & \sigma_G{}^2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_B{}^2 & \sigma_{B,G} \\ \sigma_{G,B} & \sigma_G{}^2 \end{pmatrix}$$

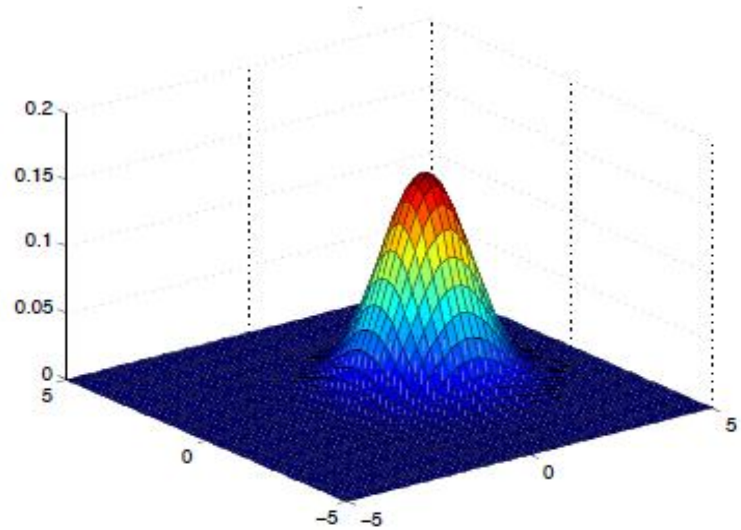


Application: clustering images

Notating a multivariate Gaussian

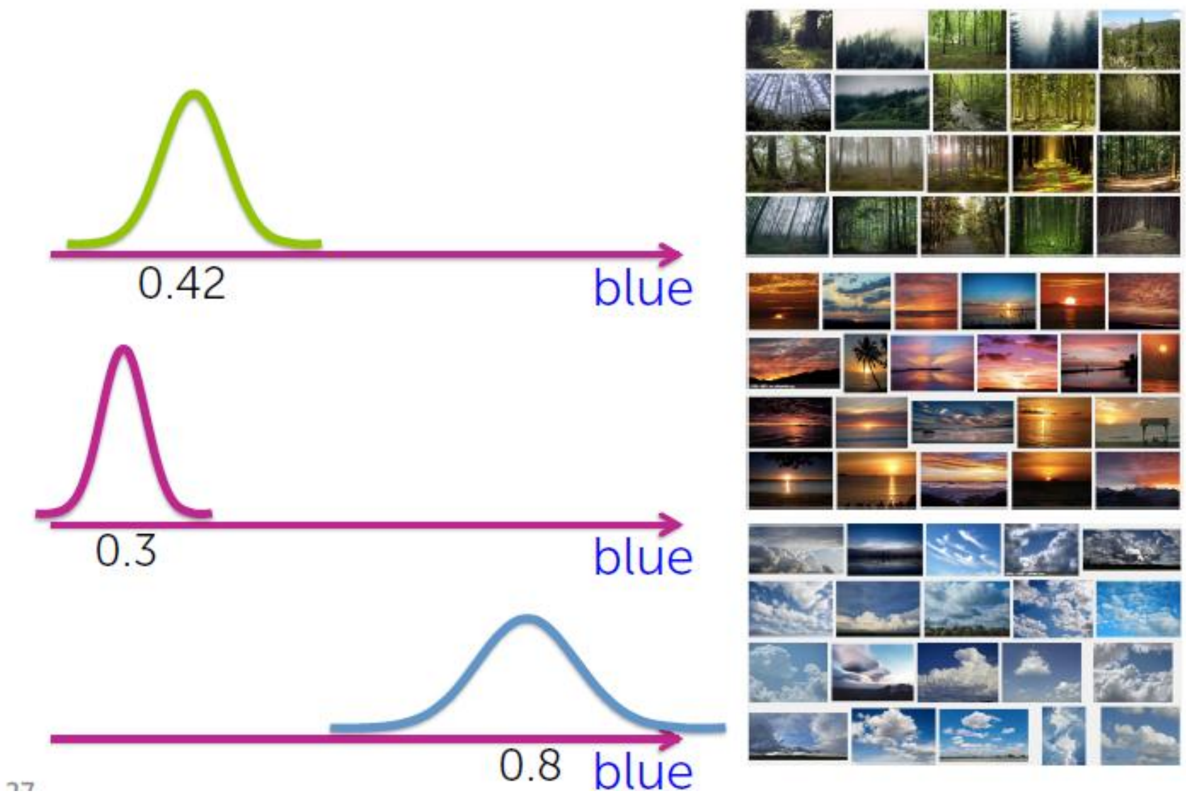
$$N(\mathbf{x} \mid \underbrace{\boldsymbol{\mu}, \Sigma}_{\text{parameters}})$$

Random vector
e.g., [R, G, B] intensities



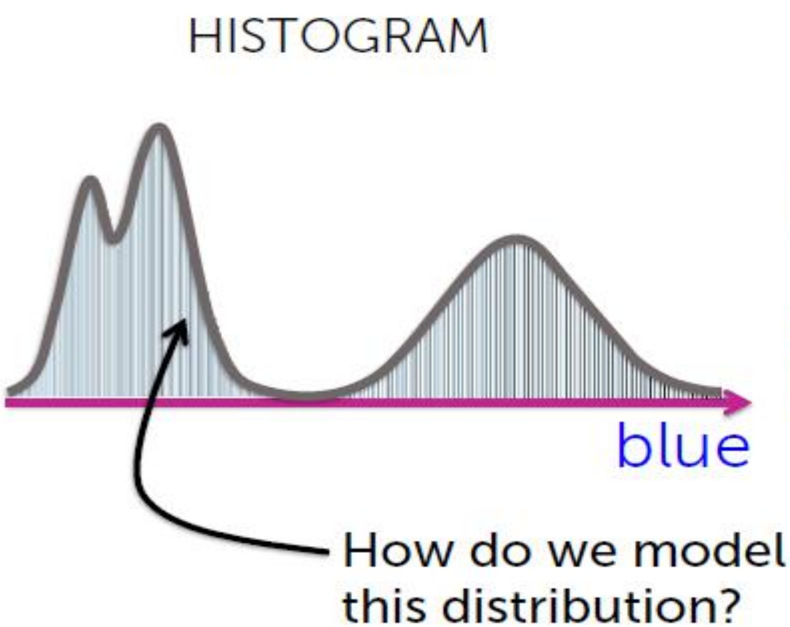
Mixture of Gaussians

Model as Gaussian per category/cluster



Mixture of Gaussians

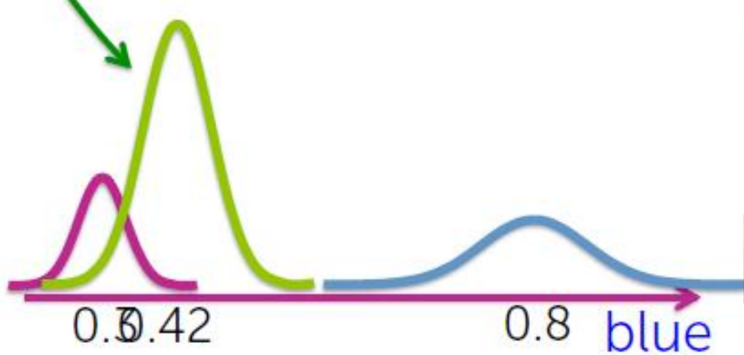
Jumble of unlabeled images



Mixture of Gaussians

What if image types not equally represented?

e.g., forest images are very likely in the collection



Mixture of Gaussians

Combination of weighted Gaussians

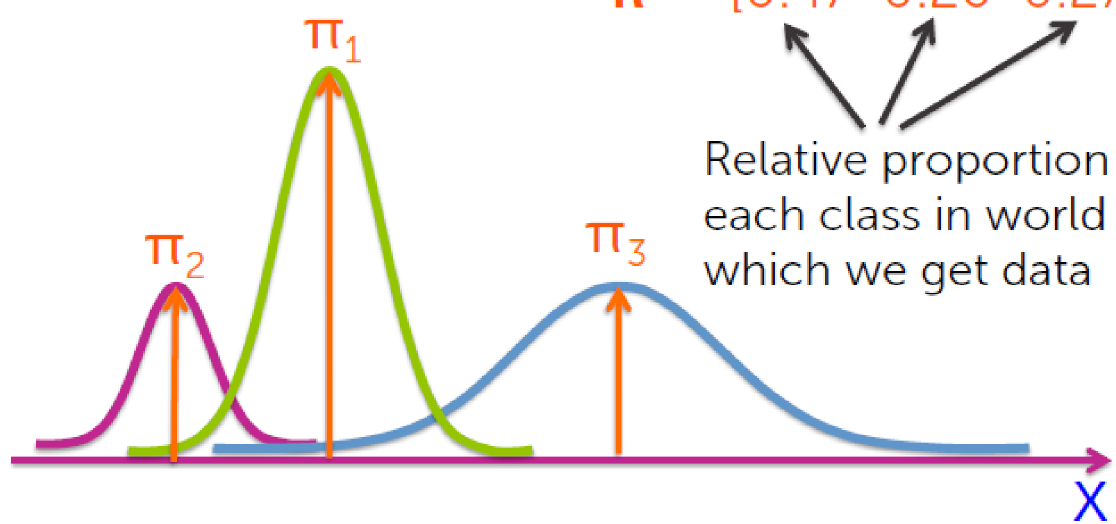
Associate a weight π_k with each Gaussian component

$$\pi = \begin{array}{ccc} \pi_1 & \pi_2 & \pi_3 \\ [0.47 & 0.26 & 0.27] \end{array}$$

Relative proportion of each class in world from which we get data

$$0 \le \pi_k \le 1$$

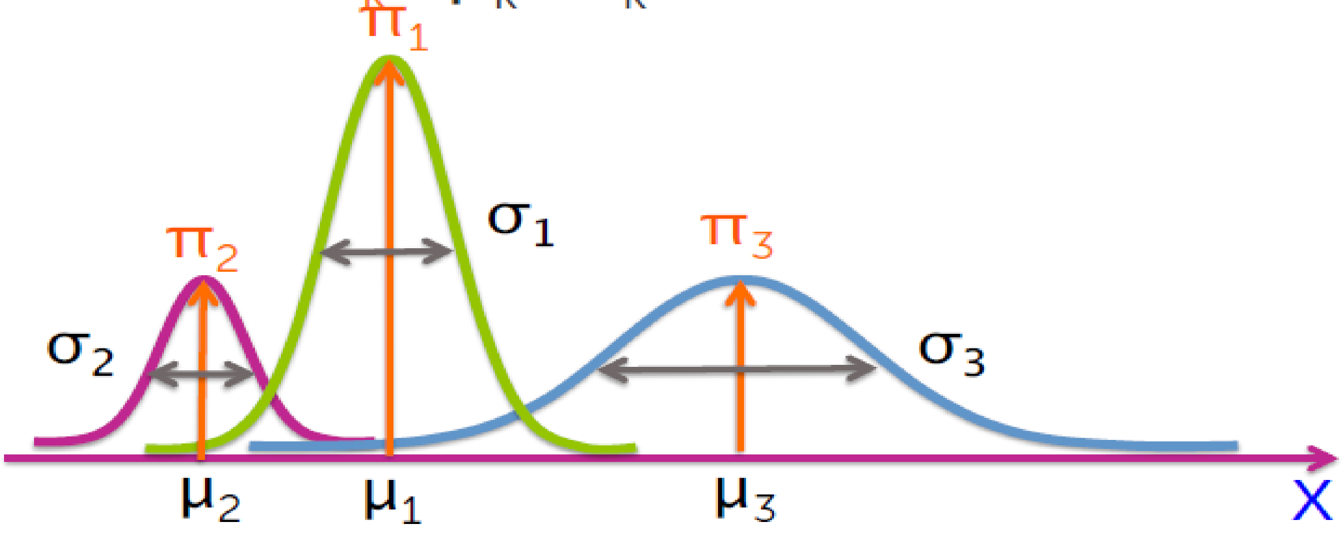
$$\sum_{k=1}^{K} \pi_k = 1$$



Mixture of Gaussians (1D)

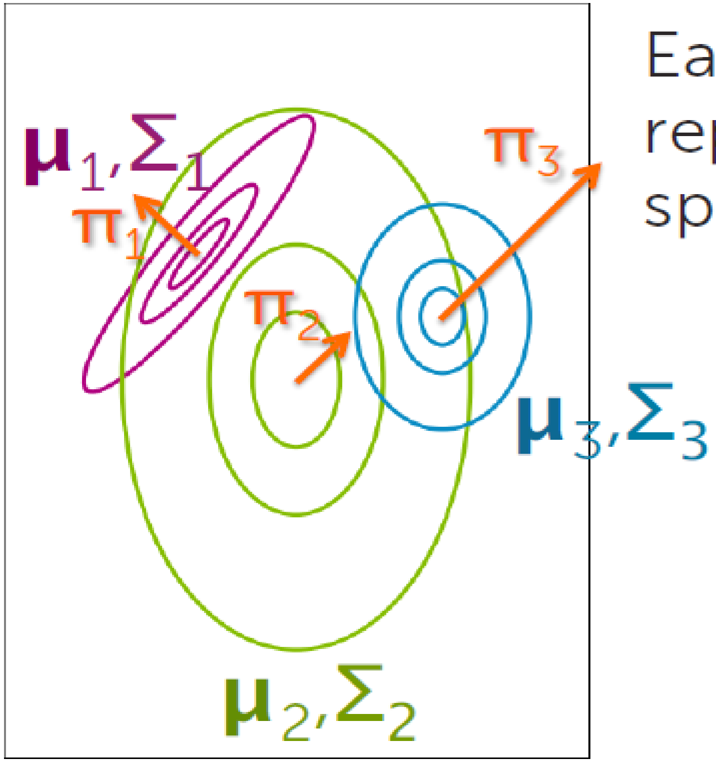
Each mixture component represents a unique cluster specified by:

$\{\pi_k, \mu_k, \sigma_k^2\}$



Mixture of Gaussians

Mixture of Gaussians (general)



Each mixture component represents a unique cluster specified by:

$$\{\pi_k , \boldsymbol{\mu}_k , \Sigma_k\}$$

Mixture of Gaussians

According to the model...

Without observing the image content, what's the probability it's from cluster k? (e.g., prob. of seeing "clouds" image)

cluster assignment for obs. x_i

$$p(z_i = k) = \pi_k$$

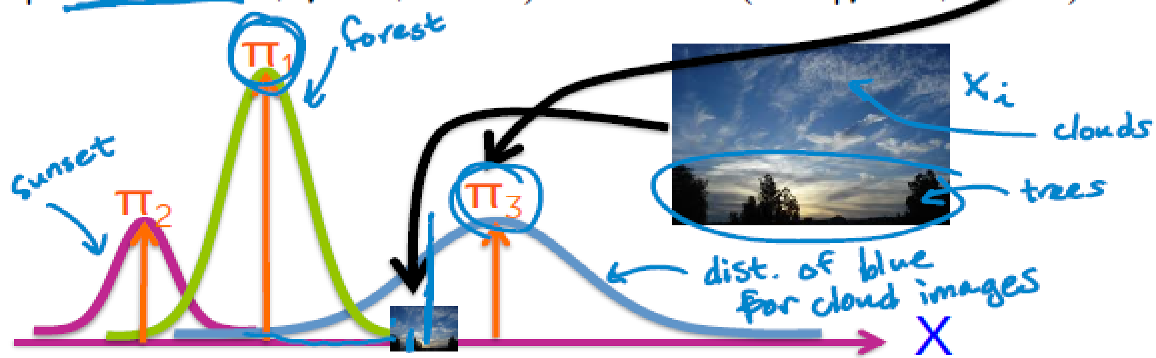
prior

Given observation $\mathbf{x}_i$ is from cluster k, what's the likelihood of seeing $\mathbf{x}_i$? (e.g., just look at distribution for "clouds")

$$p(x_i \mid z_i = k, \mu_k, \Sigma_k) = N(x_i|\mu_k, \Sigma_k)$$

likelihood

$[RGB]_i$



Application: clustering documents

Discover groups of related documents

Application: clustering documents

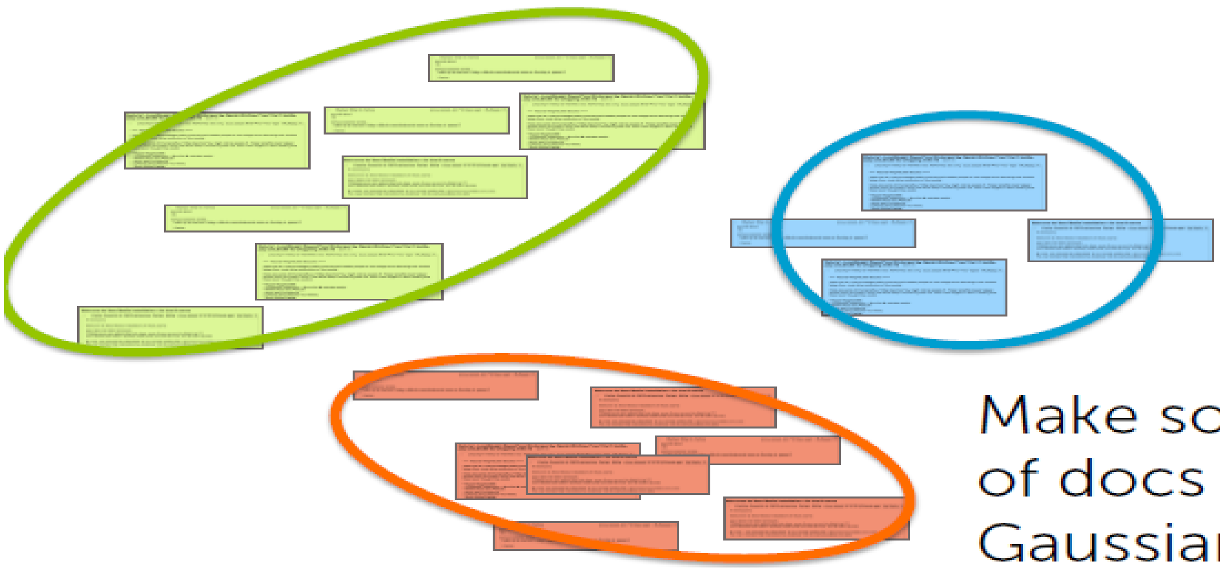
Document representation

$\mathbf{x}_i$ = tf-idf vector

Application: clustering documents

Mixture of Gaussians for clustering documents

Space of all documents
(really lives in $\mathbf{R}^V$ for vocab size V)

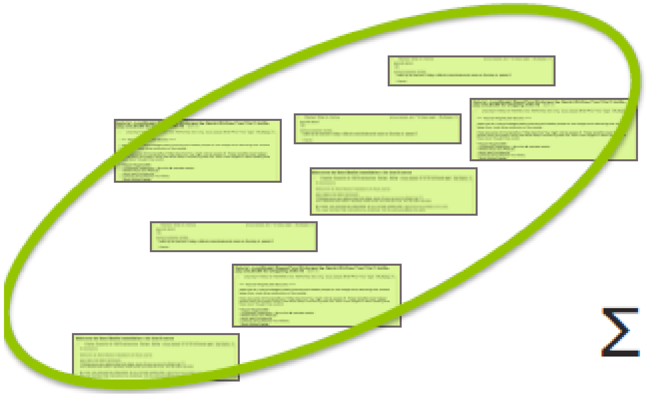


Make soft assignments of docs to each Gaussian

Application: clustering documents

Counting parameters

Each cluster has $\{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}$



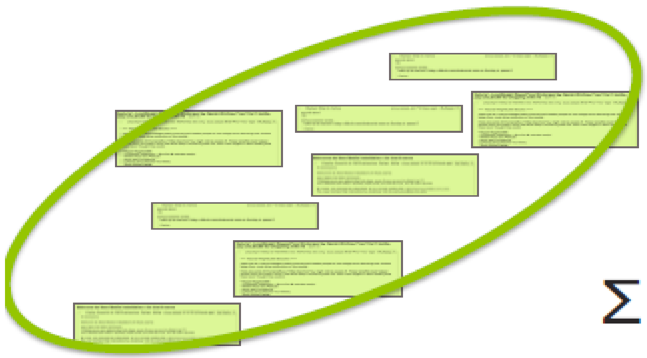
In 2D:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_2^2 \end{pmatrix}$$

Application: clustering documents

Counting parameters

Each cluster has $\{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}$



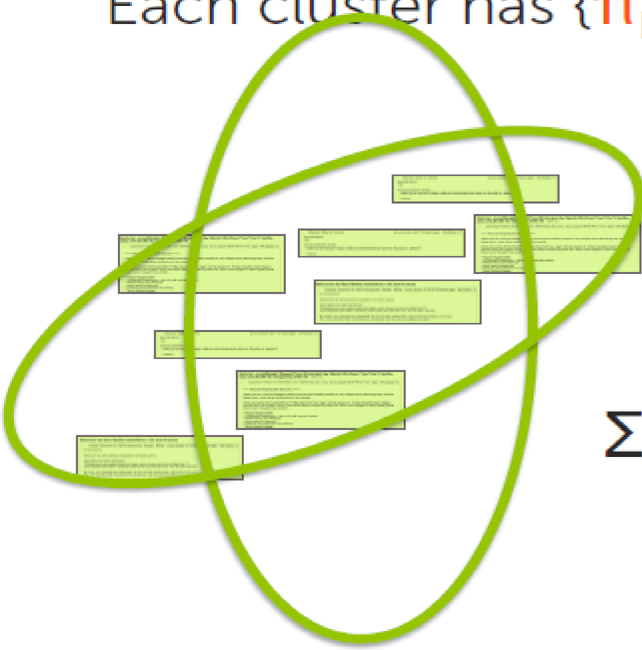
In V (vocab size) dims:

$$\Sigma = \begin{pmatrix} \cdots \\ \dfrac{V(V+1)}{2} \\ \cdots \end{pmatrix}$$

Application: clustering documents

Restricting to diagonal covariance

Each cluster has $\{\pi_k, \boldsymbol{\mu}_k, \Sigma_k$ diagonal $\}$

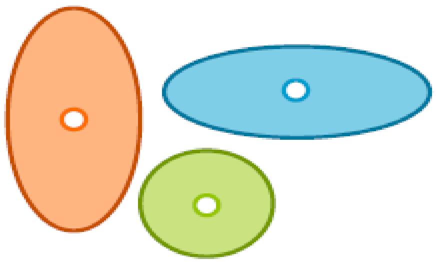


V params

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & & & \\ & \sigma_2^2 & & 0 & \\ & & \sigma_3^2 & & \\ & 0 & & \ddots & \\ & & & & \sigma_V^2 \end{pmatrix}$$

Application: clustering documents

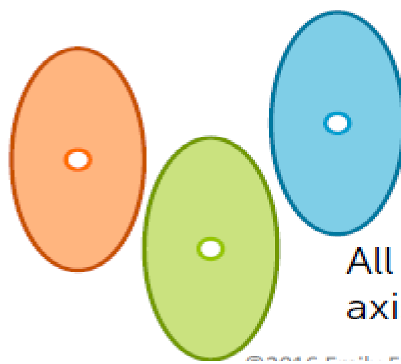
Restrictive assumption, but…



- Can **learn** weights on dimensions (e.g., weights on words in vocab)
- Can learn **cluster-specific** weights on dimensions

Still more flexible than k-means

Spherically symmetric clusters



Specify weights…

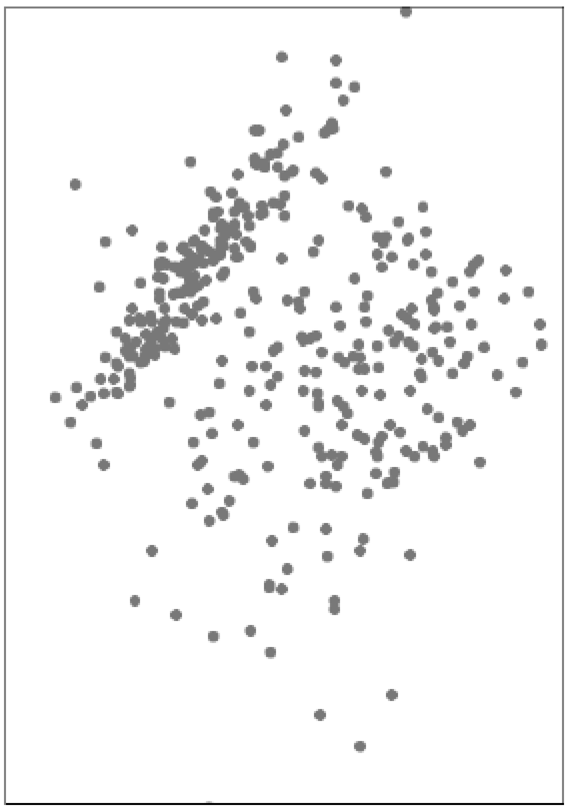
All clusters have same axis-aligned ellipses

44

Inferring soft assignments with expectation maximization (EM)

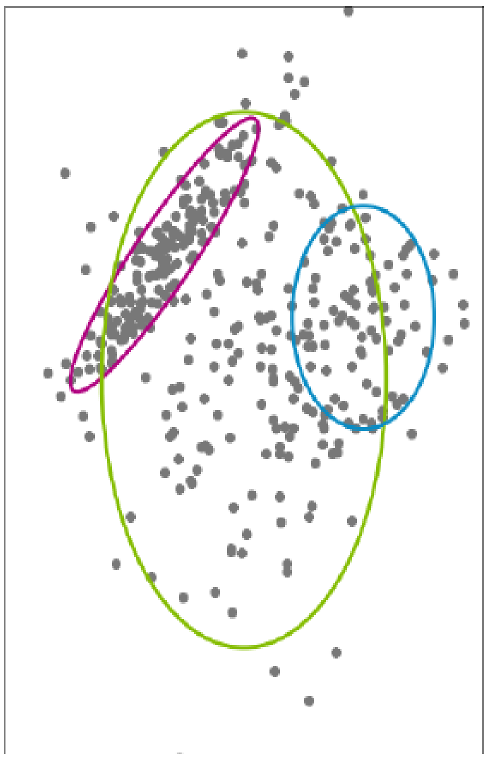
Inferring cluster labels

Data



Desired soft assignments

Compute responsibilities



$r_i = [r_{i1}\ r_{i2} \ldots\ r_{iK}]$ # clusters

fixed values defining the distribution

Responsibility cluster k takes for observation i

"given"

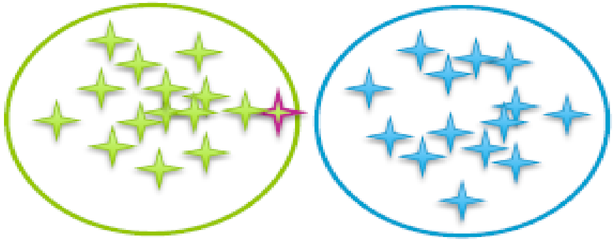
$$r_{ik} = p(z_i = k \mid \{\pi_j, \mu_j, \Sigma_j\}_{j=1}^{K}, x_i)$$

random variable

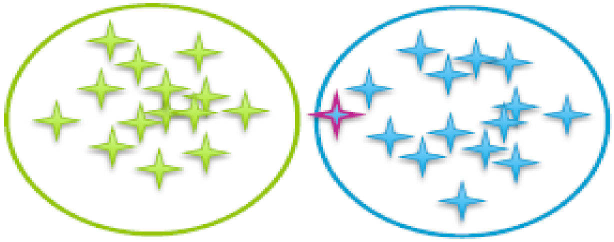
probability of assignment to cluster k

given model parameters and observed value

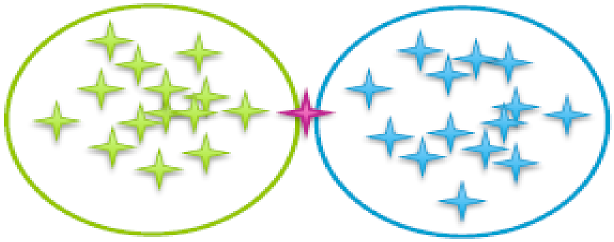
Responsibilities in pictures



Green cluster takes more responsibility



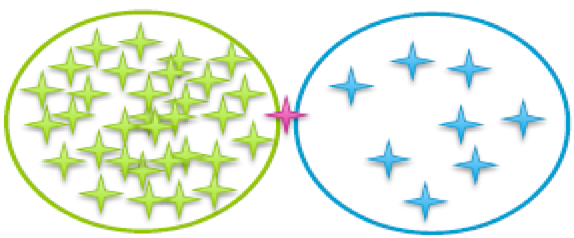
Blue cluster takes more responsibility



Uncertain… split responsibility

Responsibilities in pictures

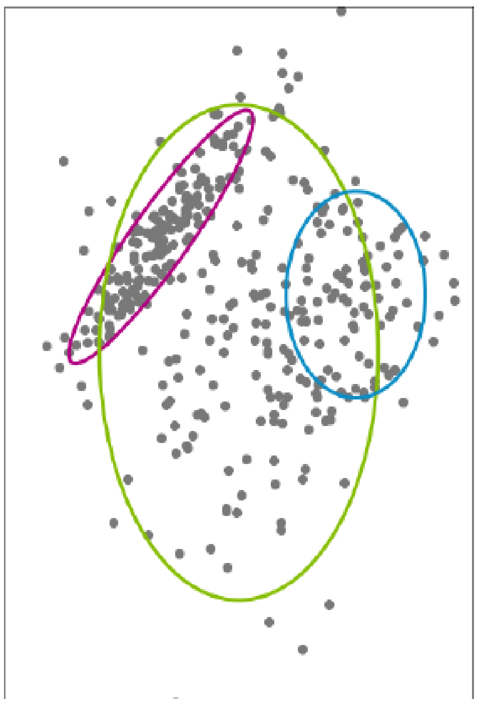
Need to weight by cluster probabilities,
not just cluster shapes



Still uncertain,
but green cluster seems
more probable…
takes more responsibility

What if we knew the cluster parameters $\{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}$?

Responsibilities in equations

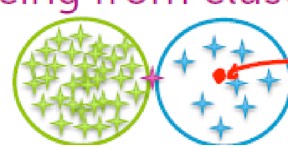


Responsibility cluster k takes for observation i

$$r_{ik} = \pi_k \, N(x_i \mid \mu_k, \Sigma_k)$$

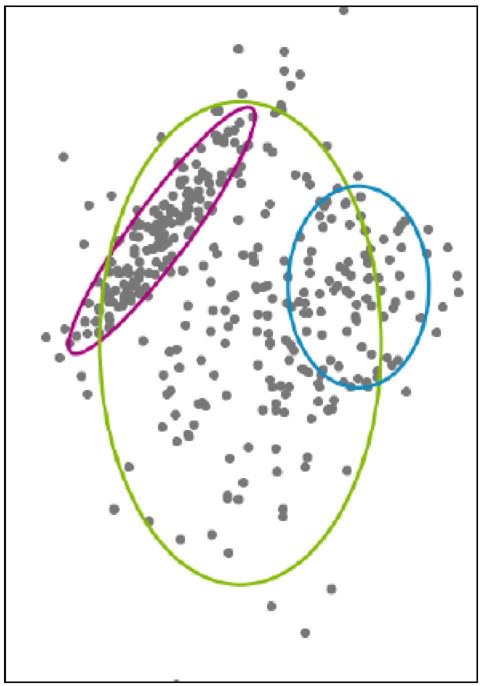
Initial probability of being from cluster k

How likely is the observed value $\mathbf{x}_i$ under this cluster assignment?



very unlikely under the green cluster, even though the prior on green is higher

Responsibilities in equations



Responsibility cluster k takes for observation i

$$r_{ik} = \frac{\pi_k \, N(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_i \mid \mu_j, \Sigma_j)}$$

Normalized over all possible cluster assignments

What if we knew the cluster parameters $\{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}$?

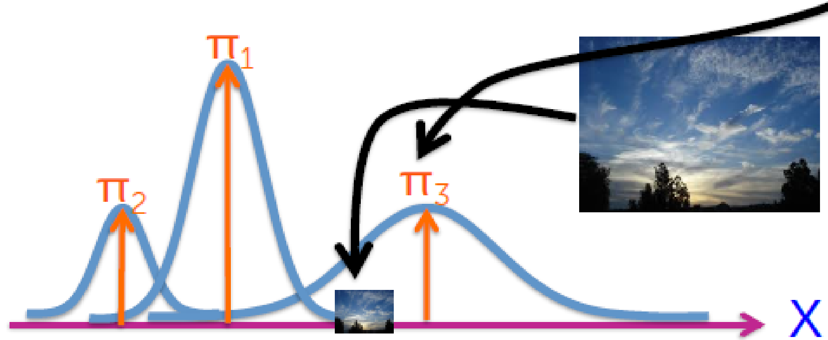
Recall: According to the model...

Without observing the image content, what's the probability it's from cluster k? (e.g., prob. of seeing "clouds" image)

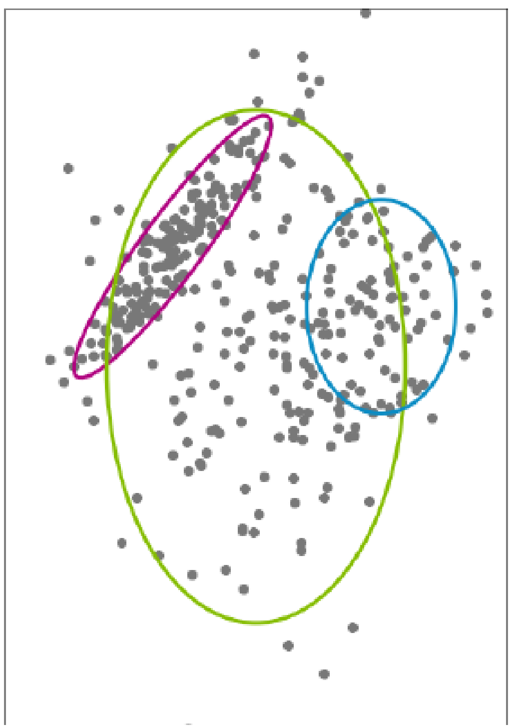
$$p(z_i = k) = \pi_k$$

Given observation $\mathbf{x}_i$ is from cluster k, what's the likelihood of seeing $\mathbf{x}_i$? (e.g., just look at distribution for "clouds")

$$p(x_i \mid z_i = k, \mu_k, \Sigma_k) = N(x_i \mid \mu_k, \Sigma_k)$$



Part 1: Summary



Desired soft assignments (responsibilities) are **easy** to compute when cluster parameters $\{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}$ are known

But, we don't know these!

Estimating cluster parameters

Imagine we know the cluster assignments

Estimation problem decouples across clusters

Is green point informative of fuchsia cluster parameters?

NO!

Data table decoupling over clusters

R	G	B	Cluster
$x_1[1]$	$x_1[2]$	$x_1[3]$	3
$x_2[1]$	$x_2[2]$	$x_2[3]$	3
$x_3[1]$	$x_3[2]$	$x_3[3]$	3
$x_4[1]$	$x_4[2]$	$x_4[3]$	1
$x_5[1]$	$x_5[2]$	$x_5[3]$	2
$x_6[1]$	$x_6[2]$	$x_6[3]$	2

Then split into separate tables and consider them independently.

Maximum likelihood estimation

R	G	B	Cluster
$x_1[1]$	$x_1[2]$	$x_1[3]$	3
$x_2[1]$	$x_2[2]$	$x_2[3]$	3
$x_3[1]$	$x_3[2]$	$x_3[3]$	3

Estimate $\{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}$ given data assigned to cluster k

maximum likelihood estimation (MLE)

Find parameters that maximize the score, or *likelihood*, of data

Imagine we knew the cluster (hard) assignments z_i

Mean/covariance MLE

Sum these vectors

R	G	B	Cluster
$x_1[1]$	$x_1[2]$	$x_1[3]$	3
$x_2[1]$	$x_2[2]$	$x_2[3]$	3
$x_3[1]$	$x_3[2]$	$x_3[3]$	3

divide by 3 (the total # of obs.)

denotes "estimate"

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{i \text{ in } k} x_i$$

← average data points in cluster k

# of obs. in cluster

$$\hat{\Sigma}_k = \frac{1}{N_k} \sum_{i \text{ in } k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

Scalar case: $\hat{\sigma}_k^2 = \frac{1}{N_k} \sum_{i \text{ in } k} (x_i - \hat{\mu}_k)^2$

Imagine we knew the cluster (hard) assignments z_i

Cluster proportion MLE

R	G	B	Cluster
$x_4[1]$	$x_4[2]$	$x_4[3]$	1

R	G	B	Cluster
$x_5[1]$	$x_5[2]$	$x_5[3]$	2
$x_6[1]$	$x_6[2]$	$x_6[3]$	2

R	G	B	Cluster
$x_1[1]$	$x_1[2]$	$x_1[3]$	3
$x_2[1]$	$x_2[2]$	$x_2[3]$	3
$x_3[1]$	$x_3[2]$	$x_3[3]$	3

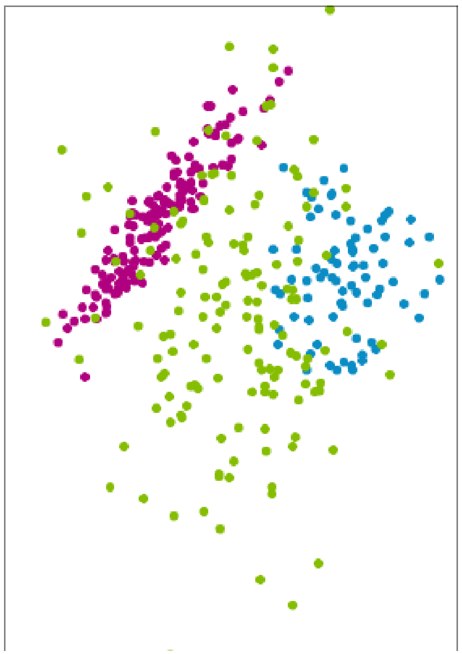
obs in cluster k

$$\hat{\pi}_k = \frac{N_k}{N}$$

total # of obs

True for general mixtures of i.i.d. data, not just Gaussian clusters

Part 2a : Summary

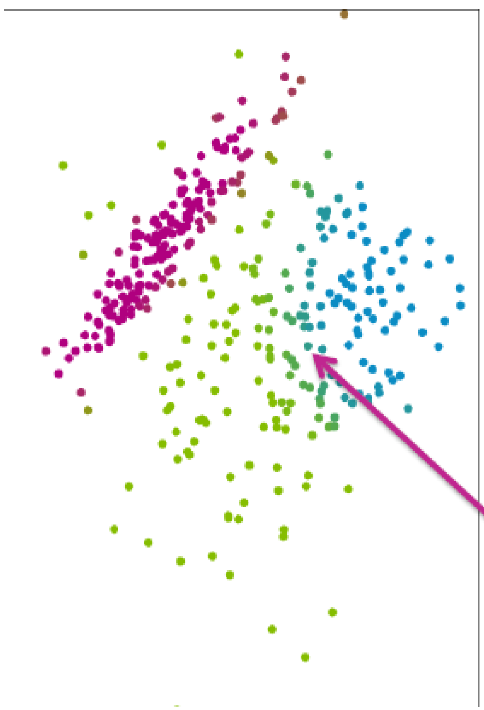


needed to compute soft assignments

Cluster parameters are simple to compute if we know the cluster assignments

But, we don't know these!

Estimating cluster parameters from soft assignments



Instead of having a full observation $\mathbf{x}_i$ in cluster k, just allocate a portion r_{ik}

$\mathbf{x}_i$ divided across all clusters, as determined by r_{ik}

What can we do with just soft assignments r_{ij}?

Maximum likelihood estimation from soft assignments

Just like in boosting with weighted observations…

R	G	B	r_{i1}	r_{i2}	r_{i3}
$x_1[1]$	$x_1[2]$	$x_1[3]$	0.30	0.18	0.52
$x_2[1]$	$x_2[2]$	$x_2[3]$	0.01	0.26	0.73
$x_3[1]$	$x_3[2]$	$x_3[3]$	0.002	0.008	0.99
$x_4[1]$	$x_4[2]$	$x_4[3]$	0.75	0.10	0.15
$x_5[1]$	$x_5[2]$	$x_5[3]$	0.05	0.93	0.02
$x_6[1]$	$x_6[2]$	$x_6[3]$	0.13	0.86	0.01

52% chance this obs is in cluster 3

Total weight in cluster: (effective # of obs)

1.242	**2.8**	**2.42**

What can we do with just soft assignments r_{ij}?

Maximum likelihood estimation from soft assignments

R	G	B	Cluster 1 weights
$x_1[1]$	$x_1[2]$	$x_1[3]$	0.30
$x_2[1]$			
$x_3[1]$			
$x_4[1]$			
$x_5[1]$			
$x_6[1]$			

R	G	B	Cluster 2 weights
$x_1[1]$	$x_1[2]$	$x_1[3]$	0.18
$x_2[1]$			
$x_3[1]$			
$x_4[1]$			
$x_5[1]$			
$x_6[1]$			

R	G	B	Cluster 3 weights
$x_1[1]$	$x_1[2]$	$x_1[3]$	0.52
$x_2[1]$	$x_2[2]$	$x_2[3]$	0.73
$x_3[1]$	$x_3[2]$	$x_3[3]$	0.99
$x_4[1]$	$x_4[2]$	$x_4[3]$	0.15
$x_5[1]$	$x_5[2]$	$x_5[3]$	0.02
$x_6[1]$	$x_6[2]$	$x_6[3]$	0.01

What can we do with just soft assignments r_{ij}?

Cluster-specific location/shape MLE

R	G	B	Cluster 1 weights
$x_1[1]$	$x_1[2]$	$x_1[3]$	0.30
$x_2[1]$	$x_2[2]$	$x_2[3]$	0.01
$x_3[1]$	$x_3[2]$	$x_3[3]$	0.002
$x_4[1]$	$x_4[2]$	$x_4[3]$	0.75
$x_5[1]$	$x_5[2]$	$x_5[3]$	0.05
$x_6[1]$	$x_6[2]$	$x_6[3]$	0.13

$$\hat{\mu}_k = \frac{1}{N_k^{\text{soft}}} \sum_{i=1}^{N} \boxed{r_{ik}} x_i$$

$$\hat{\Sigma}_k = \frac{1}{N_k^{\text{soft}}} \sum_{i=1}^{N} \boxed{r_{ik}} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

$$\boxed{N_k^{\text{soft}} = \sum_{i=1}^{N} r_{ik}}$$

Total weight in cluster k = effective # obs

Compute cluster parameter estimates with weights on each row operation

What can we do with just soft assignments r_{ij}?

MLE of cluster proportions $\hat{\pi}_k$

r_{i1}	r_{i2}	r_{i3}
0.30	0.18	0.52
0.01	0.26	0.73
0.002	0.008	0.99
0.75	0.10	0.15
0.05	0.93	0.02
0.13	0.86	0.01

Total weight in cluster:	**1.242**	**2.8**	**2.42**
Total weight in dataset:	6		

6 ← # datapoints N

$$\hat{\pi}_k = \frac{N_k^{\text{soft}}}{N}$$

$$N_k^{\text{soft}} = \sum_{i=1}^{N} r_{ik}$$

Total weight in cluster k = effective # obs

Estimate cluster proportions from relative weights

What can we do with just soft assignments r_{ij}?

Defaults to hard assignment case when r_{ij} in {0,1}

Hard assignments have:

$$r_{ik} = \begin{cases} 1 & i \text{ in } k \\ 0 & \text{otherwise} \end{cases}$$

R	G	B	r_{i1}	r_{i2}	r_{i3}
$x_1[1]$	$x_1[2]$	$x_1[3]$	0	0	1
$x_2[1]$	$x_2[2]$	$x_2[3]$	0	0	1
$x_3[1]$	$x_3[2]$	$x_3[3]$	0	0	1
$x_4[1]$	$x_4[2]$	$x_4[3]$	1	0	0
$x_5[1]$	$x_5[2]$	$x_5[3]$	0	1	0
$x_6[1]$	$x_6[2]$	$x_6[3]$	0	1	0
Total weight in cluster:			**1**	**2**	**3**

One-hot encoding of cluster assignment

What can we do with just soft assignments r_{ij}?

Equating the estimates…

$$\hat{\pi}_k = \frac{N_k^{\text{soft}}}{N} \qquad \boxed{N_k^{\text{soft}} = \sum_{i=1}^{N} r_{ik}}$$

if {0,1} just count obs i in cluster k if $r_{ik}=1$ $= N_k$ ✓

$$\hat{\mu}_k = \frac{1}{N_k^{\text{soft}}} \sum_{i=1}^{N} r_{ik} x_i$$

only add x_i if i in k ($r_{ik}=1$) $\rightarrow \frac{1}{N_k} \sum_{i \text{ in } k} x_i$ ✓

$$\hat{\Sigma}_k = \frac{1}{N_k^{\text{soft}}} \sum_{i=1}^{N} r_{ik} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$ ✓

same as above

$$= \frac{1}{N_k} \sum_{i \text{ in } k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

Part 2b: Summary

Still straightforward to compute cluster parameter estimates from soft assignments

Expectation maximization (ME)

An iterative algorithm

Motivates an iterative algorithm:

1. **E-step:** estimate cluster responsibilities given current parameter estimates

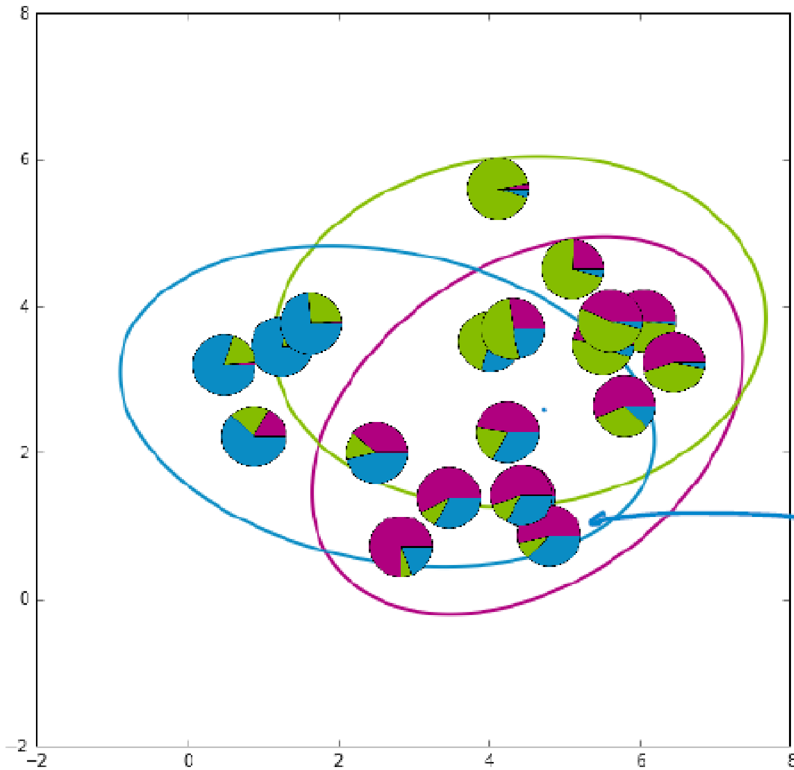
$$\hat{r}_{ik} = \frac{\hat{\pi}_k N(x_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{j=1}^{K} \hat{\pi}_j N(x_i \mid \hat{\mu}_j, \hat{\Sigma}_j)}$$

2. **M-step:** maximize likelihood over parameters given current responsibilities

$$\hat{\pi}_k, \hat{\mu}_k, \hat{\Sigma}_k \mid \{\hat{r}_{ik}, x_i\}$$

Expectation maximization (ME)

EM for mixtures of Gaussians in pictures – initialization



Initialize

iter counter

$\{\pi_k^{(0)}, \hat{\mu}_k^{(0)}, \hat{\Sigma}_k^{(0)}\}$

Then compute

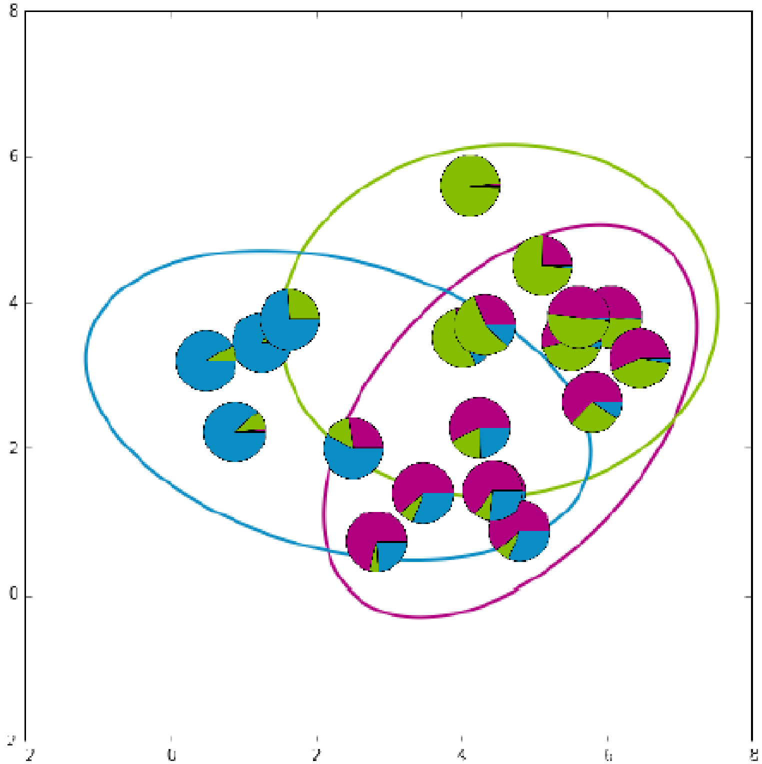
$\hat{r}_{ik}^{(1)}$

fucsia blue green

$\hat{r}_i^{(1)} = [0.52 \quad 0.4 \quad 0.08]$

Expectation maximization (ME)

EM for mixtures of Gaussians in pictures – after 1st iteration



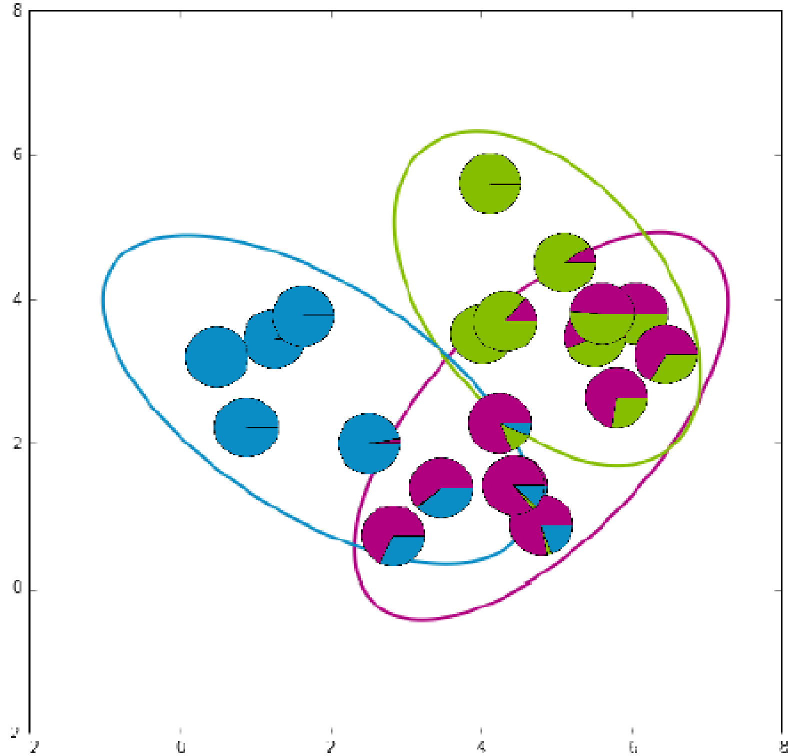
Maximize likelihood given soft assign. $r_{ik}^{(1)}$

$\rightarrow \{ \hat{\pi}_k^{(1)}, \hat{\mu}_k^{(1)}, \hat{\Sigma}_k^{(1)} \}$

Then recompute responsibilities $\hat{r}_{ik}^{(2)}$

Expectation maximization (ME)

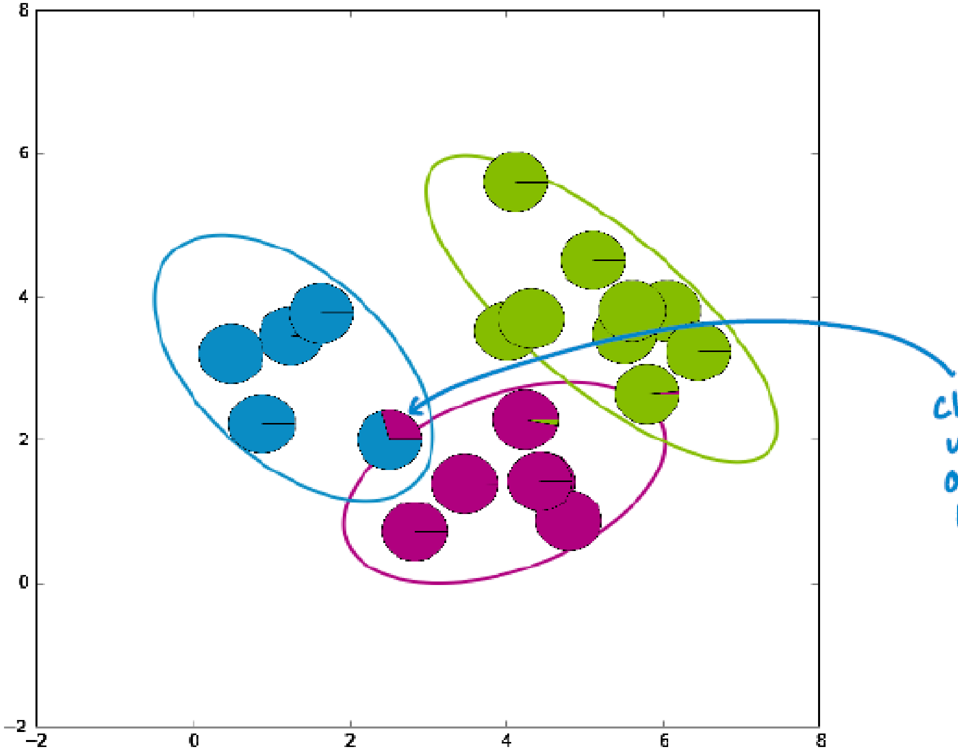
EM for mixtures of Gaussians in pictures – after 2nd iteration



rinse
+
repeat
until convergence

Expectation maximization (ME)

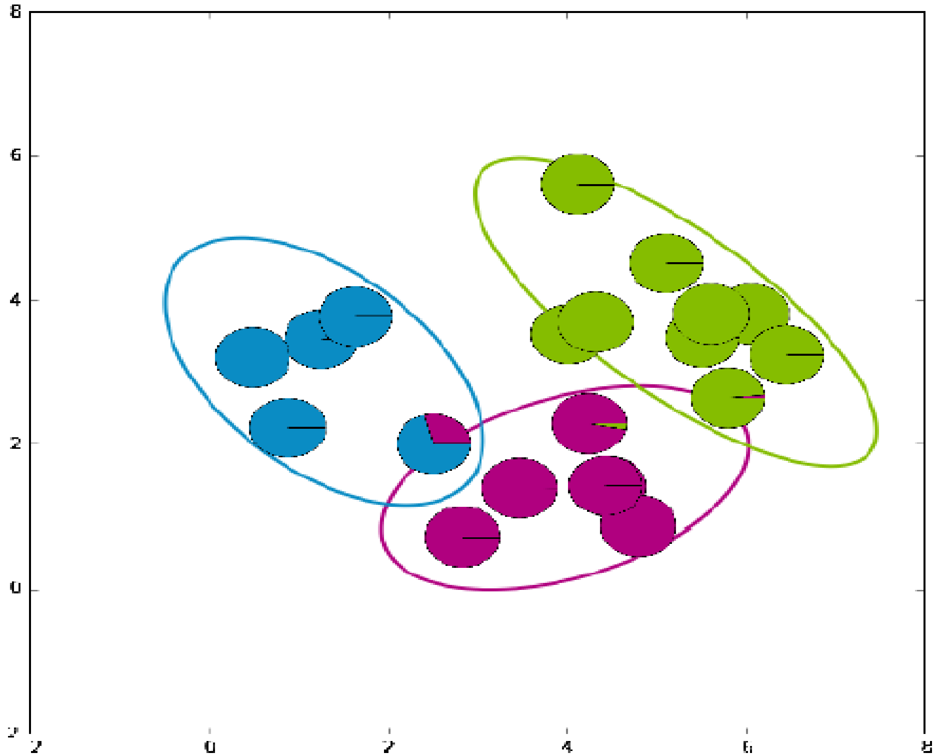
EM for mixtures of Gaussians in pictures – converged solution



Clearly see uncertainty in assignment of obs. to blue or fuchsia cluster, even in final assignments.

Expectation maximization (ME)

EM for mixtures of Gaussians in pictures - replay



Expectation maximization (ME)

Convergence of EM

- EM is a coordinate-ascent algorithm
  - Can equate E-and M-steps with alternating maximizations of an objective function
- Convergences to a local mode

- We will assess via (log) likelihood of data under current parameter and responsibility estimates

Expectation maximization (ME)

Initialization

- Many ways to initialize the EM algorithm
- Important for convergence rates and quality of local mode found
- Examples:
  - Choose K observations at random to define K "centroids". Assign other observations to nearest centriod to form initial parameter estimates.
  - Pick centers sequentially to provide good coverage of data like in k-means++
  - Initialize from k-means solution
  - Grow mixture model by splitting (and sometimes removing) clusters until K clusters are formed

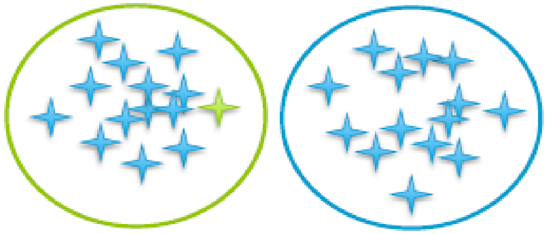
Expectation maximization (ME)

Overfitting of MLE

Maximizing likelihood can overfit to data

Imagine at K=2 example with one obs assigned to cluster 1 and others assigned to cluster 2

- What parameter values maximize likelihood?



Set center equal to point and shrink variance to 0

Likelihood goes to ∞ !

Expectation maximization (ME)

Overfitting in high dims

Doc-clustering example:
Imagine only 1 doc assigned to cluster k has word w (or all docs in cluster agree on count of word w)

Likelihood maximized by setting $\boldsymbol{\mu}_k[w] = \mathbf{x}_i[w]$ and $\sigma^2_{w,k} = 0$

Likelihood of any doc with different count on word w being in cluster k is 0!

Simple regularization of M-step for mixtures of Gaussians

Simple fix: Don’t let variances → 0!

Add small amount to diagonal of covariance estimate

Alternatively, take Bayesian approach and place prior on parameters.

Similar idea, but all parameter estimates are “smoothed” via cluster pseudo-observations.

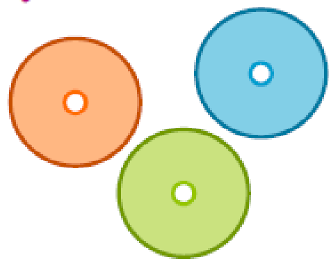
Expectation maximization (ME)

Relationship to k-means

Consider Gaussian mixture model with

$$\Sigma = \begin{pmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \sigma^2 & \\ & & & \ddots \end{pmatrix}$$

Spherically symmetric clusters



and let the variance parameter $\sigma \to 0$

Datapoint gets fully assigned to nearest center, just as in k-means

- Spherical clusters with equal variances, so relative likelihoods just function of distance to cluster center
- As variances→0, likelihood ratio becomes 0 or 1
- Responsibilities weigh in cluster proportions, but dominated by likelihood disparity

$$\hat{r}_{ik} = \frac{\hat{\pi}_k N(x_i \mid \hat{\mu}_k, \sigma^2 I)}{\sum_{j=1}^{K} \hat{\pi}_j N(x_i \mid \hat{\mu}_j, \sigma^2 I)}$$

Expectation maximization (ME)

Infinitesimally small variance EM = k-means

1. **E-step:** estimate cluster responsibilities given current parameter estimates

$$\hat{r}_{ik} = \frac{\hat{\pi}_k N(x_i \mid \hat{\mu}_k, \sigma^2 I)}{\sum_{j=1}^{K} \hat{\pi}_j N(x_i \mid \hat{\mu}_j, \sigma^2 I)} \in \{0, 1\}$$

Infinitesimally small

Decision based on distance to nearest cluster center

2. **M-step:** maximize likelihood over parameters given current responsibilities (hard assignments!)

$$\hat{\pi}_k, \hat{\mu}_k \mid \{\hat{r}_{ik}, x_i\}$$

What you can do now …

- Interpret a probabilistic model-based approach to clustering using mixture models
- Describe model parameters
- Motivate the utility of soft assignments and describe what they represent
- Discuss issues related to how the number of parameters grow with the number of dimensions
  - Interpret diagonal covariance versions of mixtures of Gaussians
- Compare and contrast mixtures of Gaussians and k-means
- Implement an EM algorithm for inferring soft assignments and cluster parameters
  - Determine an initialization strategy
  - Implement a variant that helps avoid overfitting issues

Mixed membership models for documents

Clustering model

So far, clustered articles into groups

Doc labeled with a topic assignment

Clustering goal: discover groups of related docs

Clustering model

Are documents about just one thing?

Clustering model

Soft assignments capture uncertainty

Soft assignment r_{ik} tells us this doc could be about world news **or** science

But, clustering model still specifies each doc belongs to **1 topic**

Soft assignments

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

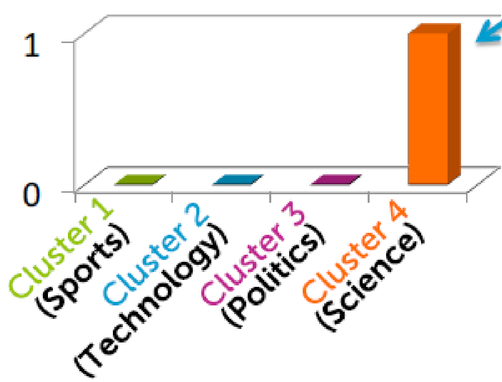
Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

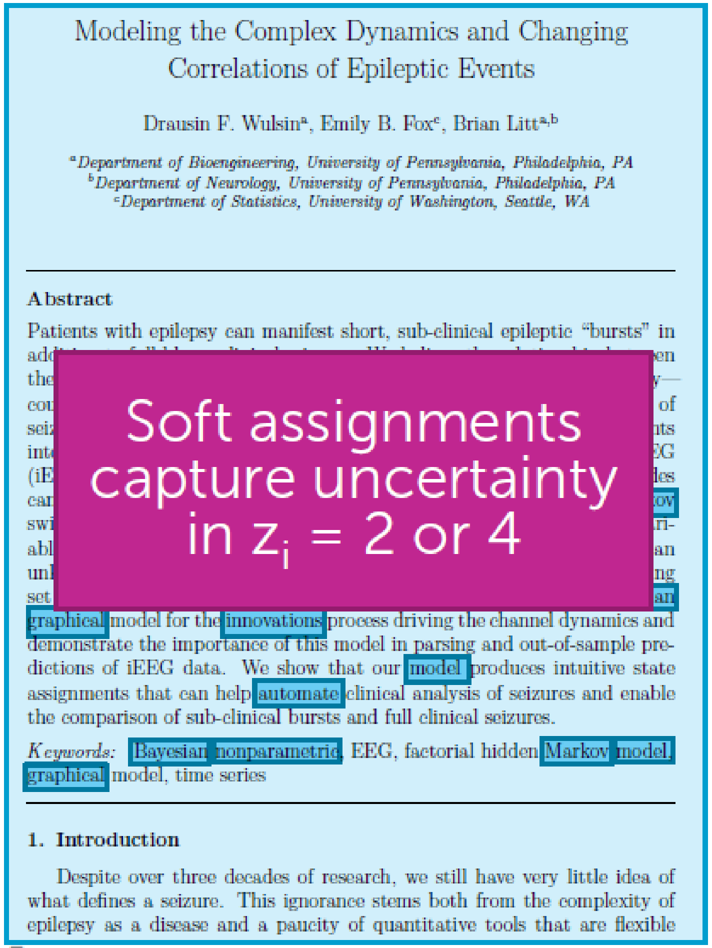
Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

Encoding of cluster membership $z_i = 4$

Based on science related words, maybe doc in cluster 4



Soft assignments



Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in

graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

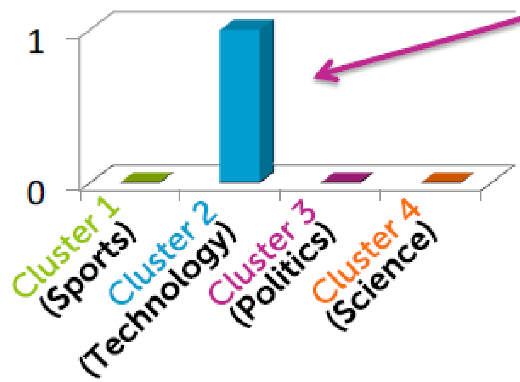
1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

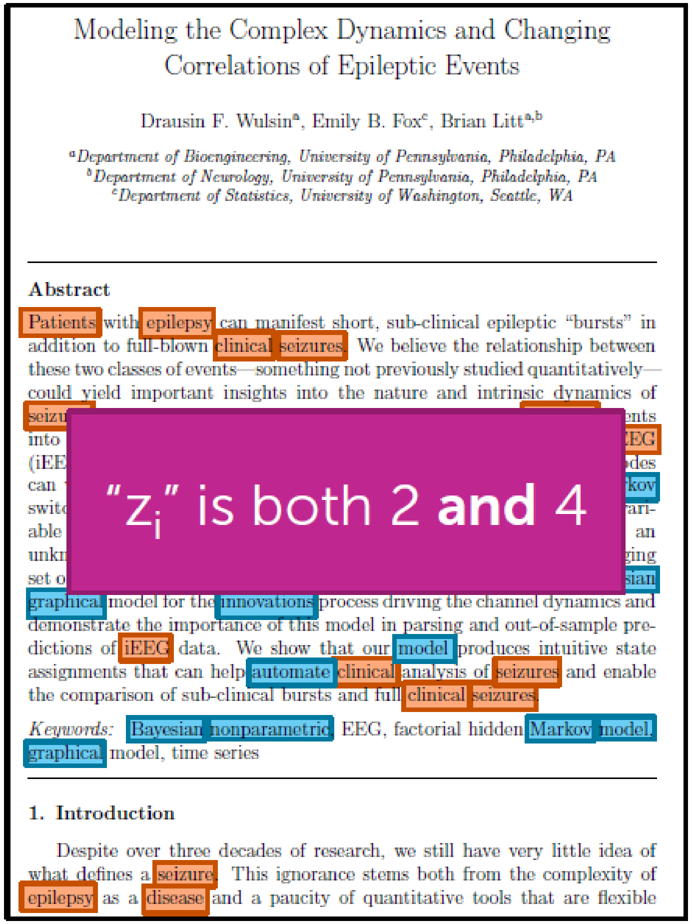
Soft assignments capture uncertainty in z_i = 2 or 4

Encoding of cluster membership z_i = 2

Or maybe cluster 2 (technology) is a better fit



Soft assignments


Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizu…

graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

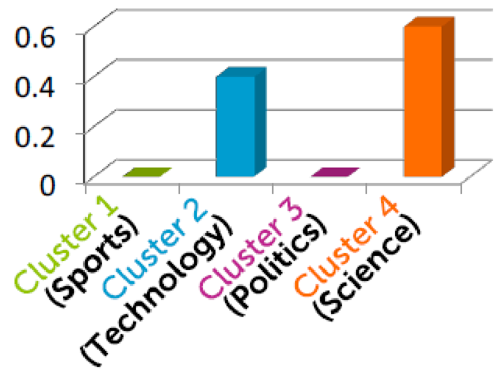
Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

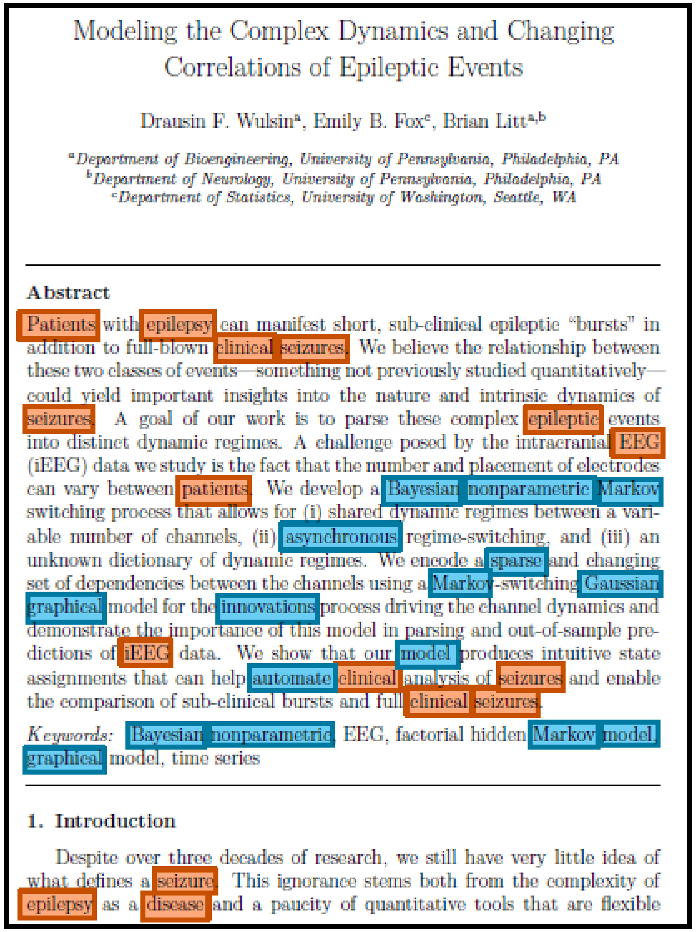
Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

"z_i" is both 2 **and** 4

Really, it's about science **and** technology



Mixed membershio models



Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

Mixed membership models

Want to discover a **set** of memberships

(In contrast, cluster models aim at discovering a single membership)

Building alternative model

An alternative document clustering model

(Back to clustering, not mixed membership modeling)

Building an alternative model

So far, we have considered…

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

$\mathbf{x}_i =$

			tf-idf vector									

Building an alternative model

Bag-of-words representation

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights int[illegible] dynamics of seizures. A goal of our work is to parse these complex epileptic events into di[illegible]allenge posed by the intracranial EEG (iEEG) [illegible] the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared [illegible] a variable number of channels, (ii) asynchrono[illegible] an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencie[illegible] using a Markov-switching Gaussian graphical model for [illegible] driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assign[illegible]ate clinical analysis of seizures and enable the co[illegible]ursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

Building an alternative model

Bag-of-words representation

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]*Department of Bioengineering, University of Pennsylvania, Philadelphia, PA*
[b]*Department of Neurology, University of Pennsylvania, Philadelphia, PA*
[c]*Department of Statistics, University of Washington, Seattle, WA*

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

---

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

$\mathbf{x}_i$ = {modeling, complex, epilepsy, modeling, Bayesian, clinical, epilepsy, EEG, data, dynamic…}

multiset

= unordered set of words with duplication of unique elements mattering

Model for „bag-of-words”

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic “bursts” in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

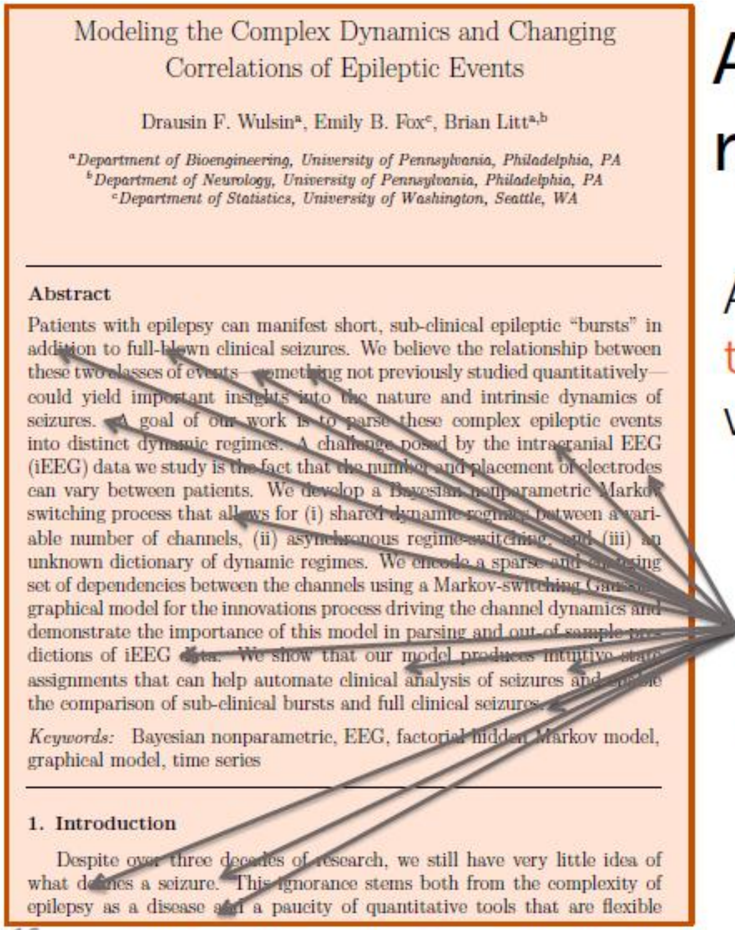
A model for bag-of-words representation

As before, the “prior” probability that doc i is from topic k is:

$$p(z_i = k) = \pi_k$$

$\boldsymbol{\pi} = [\pi_1 \ \ \pi_2 \ldots \pi_K]$
represents corpus-wide topic prevalence

Model for „bag-of-words”



Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic “bursts” in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

A model for bag-of-words representation

Assuming doc i **is** from topic k, words occur with probabilities:

SCIENCE	
patients	0.05
clinical	0.01
epilepsy	0.002
seizures	0.0015
EEG	0.001
...	...

words in vocab

Model for „bag-of-words”

Topic-specific word probabilities

Distribution on words in vocab for **each topic**

SCIENCE	
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01
…	…

TECH	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02
…	…

SPORTS	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01
…	…

…

(table now organized by decreasing probabilities
showing top words in each category)

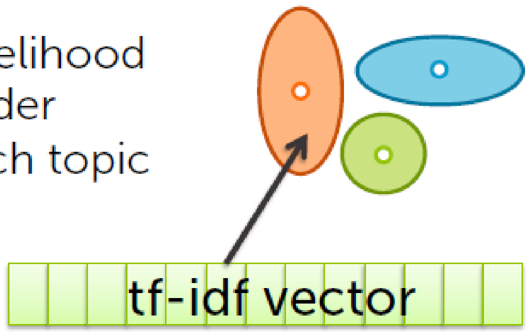
Model for „bag-of-words”

Comparing and contrasting

Previously

Prior topic probabilities $p(z_i = k) = \pi_k$

Likelihood under each topic



compute likelihood of **tf-idf** vector under each **Gaussian**

Now

$p(z_i = k) = \pi_k$

SCIENCE		TECH		SPORTS	
experiment	0.1	develop	0.18	player	0.15
test	0.08	computer	0.09	score	0.07
discover	0.05	processor	0.032	team	0.06
hypothesize	0.03	user	0.027	goal	0.03
climate	0.01	internet	0.02	injury	0.01
...	...	...	...	...	...

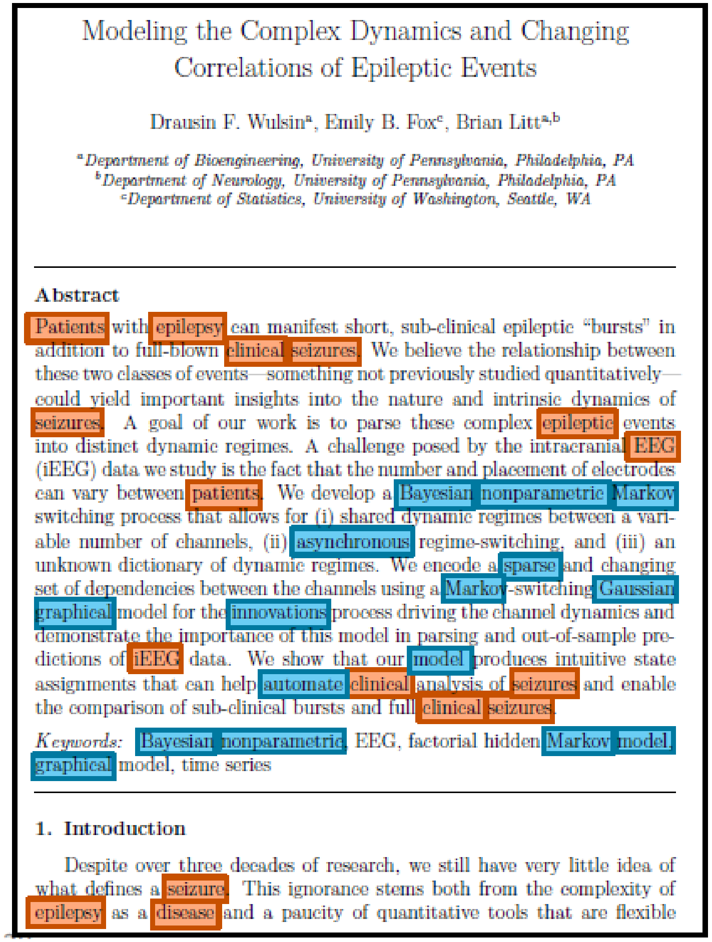
...

{modeling, complex, epilepsy, modeling, Bayesian, clinical, epilepsy, EEG, data, dynamic…}

compute likelihood of the **collection of words** in doc under each **topic distribution**

Latent Dirichlet allocation (LDA)

Latent Dirichlet allocation



Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

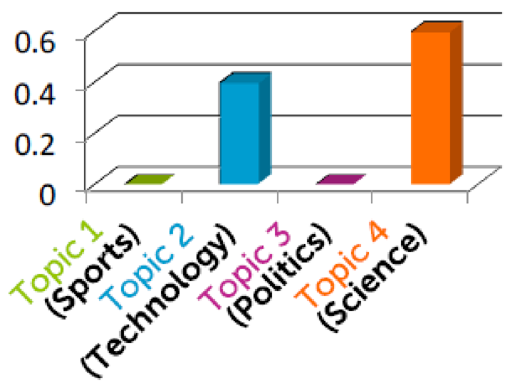
Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

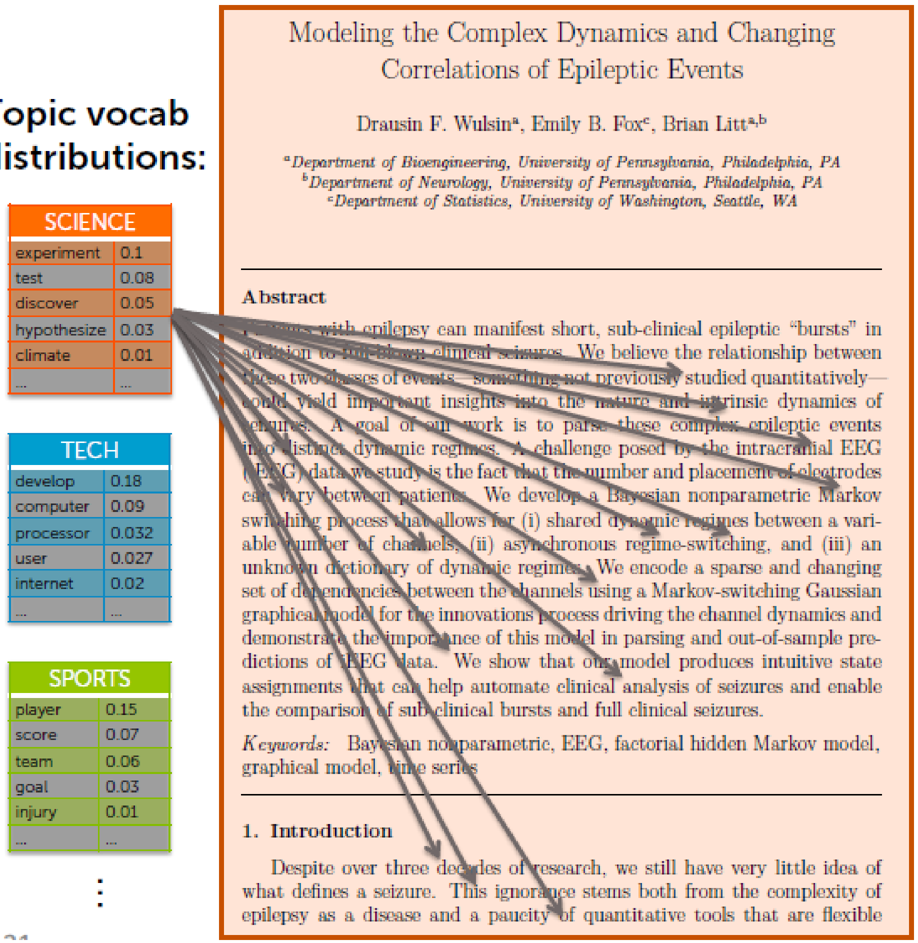
LDA is a mixed membership model

Want to discover a **set** of topics



Latent Dirichlet allocation

Topic vocab distributions:



SCIENCE	
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01
...	...

TECH	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02
...	...

SPORTS	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01
...	...

⋮

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

Clustering:

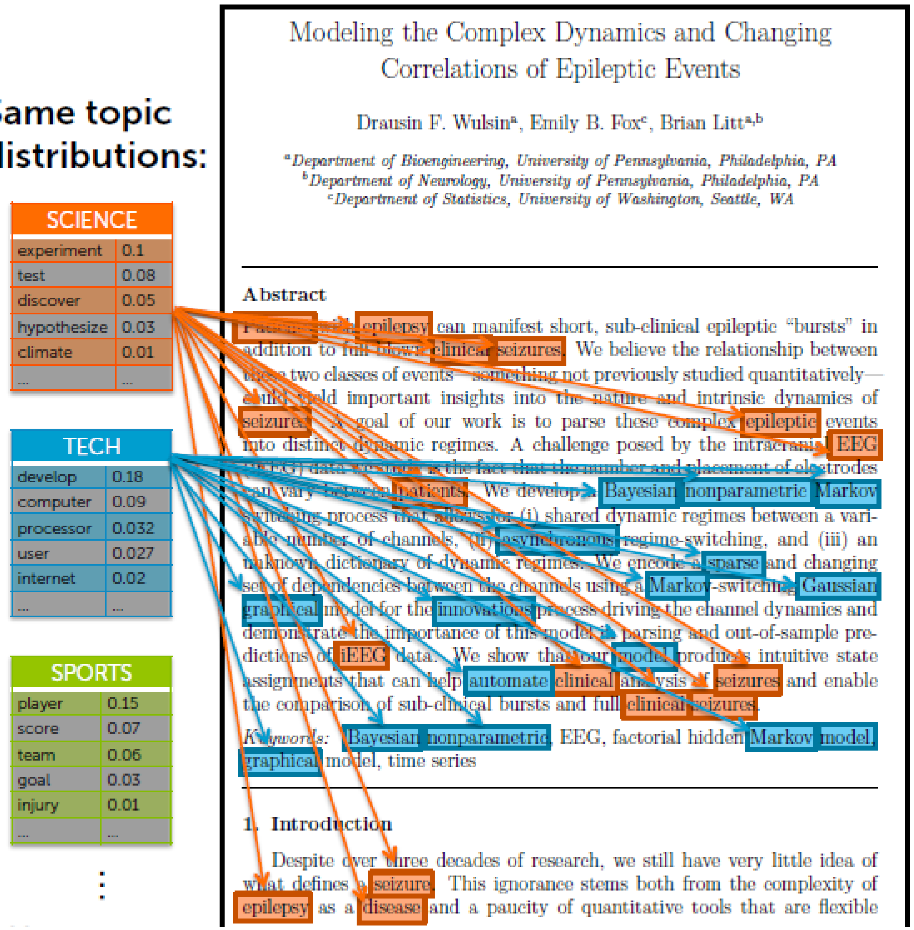
One topic indicator z_i per **document** i

All words come from (get scored under) same topic z_i

Distribution on prevalence of topics in **corpus**
$\boldsymbol{\pi} = [\pi_1\ \pi_2 \dots \pi_K]$

Latent Dirichlet allocation

Same topic distributions:



In LDA:

One topic indicator z_{iw} per **word** in doc i

Each word gets scored under its topic z_{iw}

Distribution on prevalence of topics in **document** $\boldsymbol{\pi}_i = [\pi_{i1}\ \pi_{i2} \ldots \pi_{iK}]$

Inference in LDA models

Topic vocab distributions:

SCIENCE	
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01
...	...

TECH	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02
...	...

SPORTS	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01
...	...

⋮

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]*Department of Bioengineering, University of Pennsylvania, Philadelphia, PA*
[b]*Department of Neurology, University of Pennsylvania, Philadelphia, PA*
[c]*Department of Statistics, University of Washington, Seattle, WA*

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

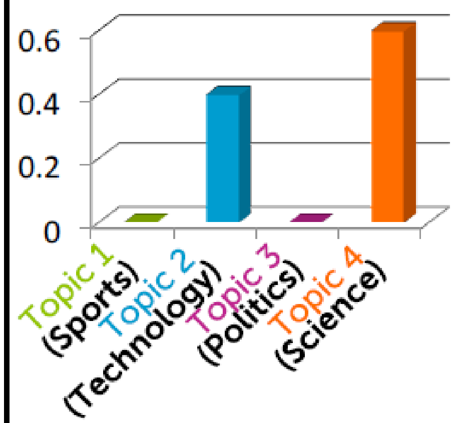
Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

Document topic proportions:

$\boldsymbol{\pi}_i = [\pi_{i1} \ \ \pi_{i2} \ \ldots \ \pi_{iK}]$



Inference in LDA models

Topic vocab distributions:

TOPIC 1	
Word 1	?
Word 2	?
Word 3	?
Word 4	?
Word 5	?
...	...

TOPIC 2	
Word 1	?
Word 2	?
Word 3	?
Word 4	?
Word 5	?
...	...

TOPIC 3	
Word 1	?
Word 2	?
Word 3	?
Word 4	?
Word 5	?
...	...

⋮

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]*Department of Bioengineering, University of Pennsylvania, Philadelphia, PA*
[b]*Department of Neurology, University of Pennsylvania, Philadelphia, PA*
[c]*Department of Statistics, University of Washington, Seattle, WA*

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

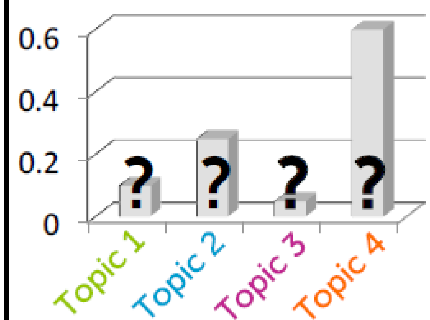
Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

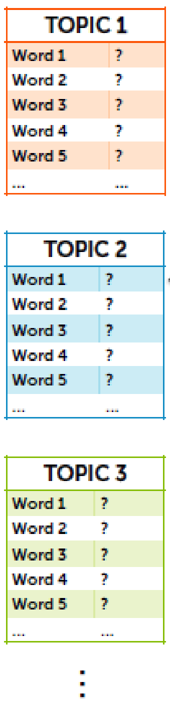
Document topic proportions:

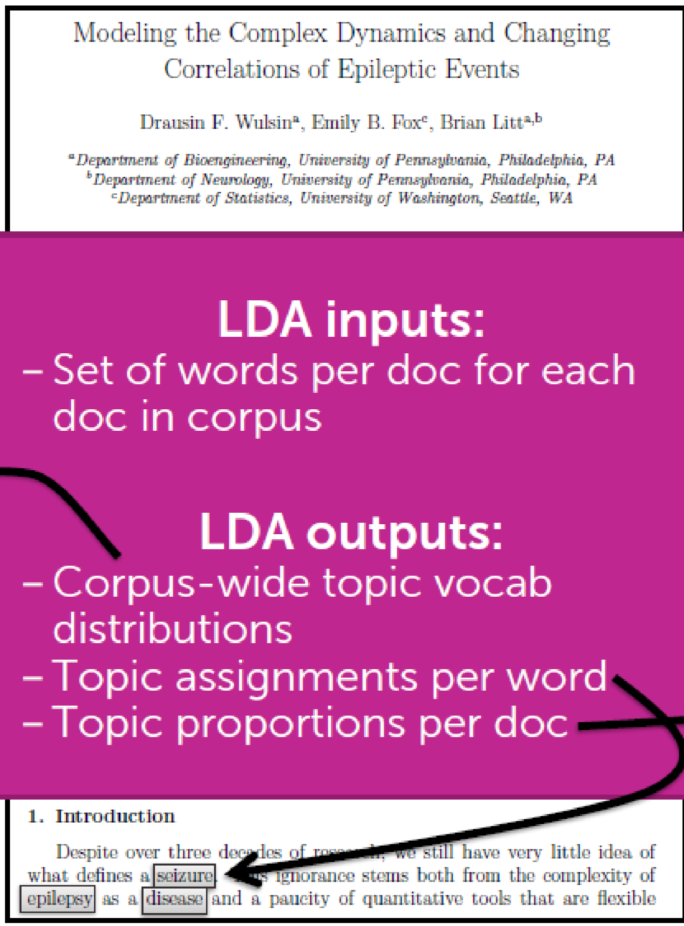
$\boldsymbol{\pi}_i = [\pi_{i1} \ \ \pi_{i2} \ \dots \ \pi_{iK}]$



Inference in LDA models

Topic vocab distributions:





Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

LDA inputs:

- Set of words per doc for each doc in corpus

LDA outputs:

- Corpus-wide topic vocab distributions
- Topic assignments per word
- Topic proportions per doc

Document topic proportions:

$\pi_i = [\pi_{i1} \; \pi_{i2} \; \ldots \; \pi_{iK}]$

Inference in LDA models

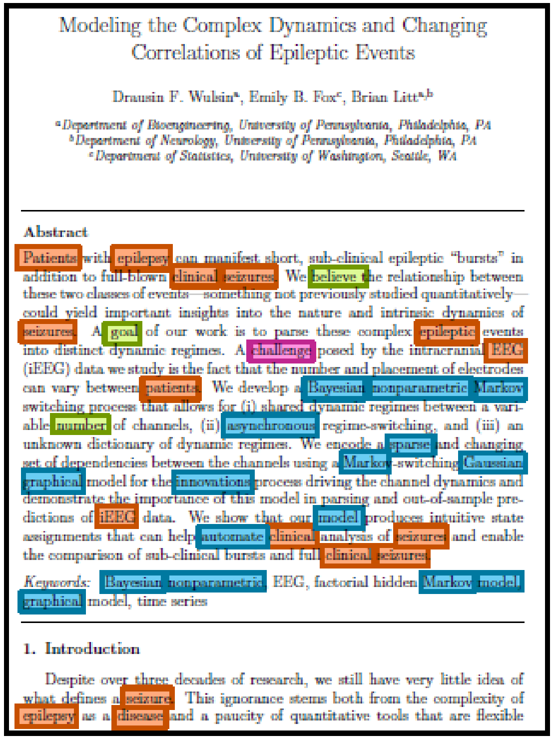
Interpreting LDA outputs

TOPIC 1	
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01
...	...

TOPIC 2	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02
...	...

TOPIC 3	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01
...	...

⋮


Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a] Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b] Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c] Department of Statistics, University of Washington, Seattle, WA

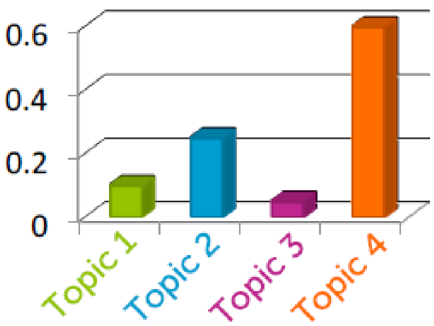
Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

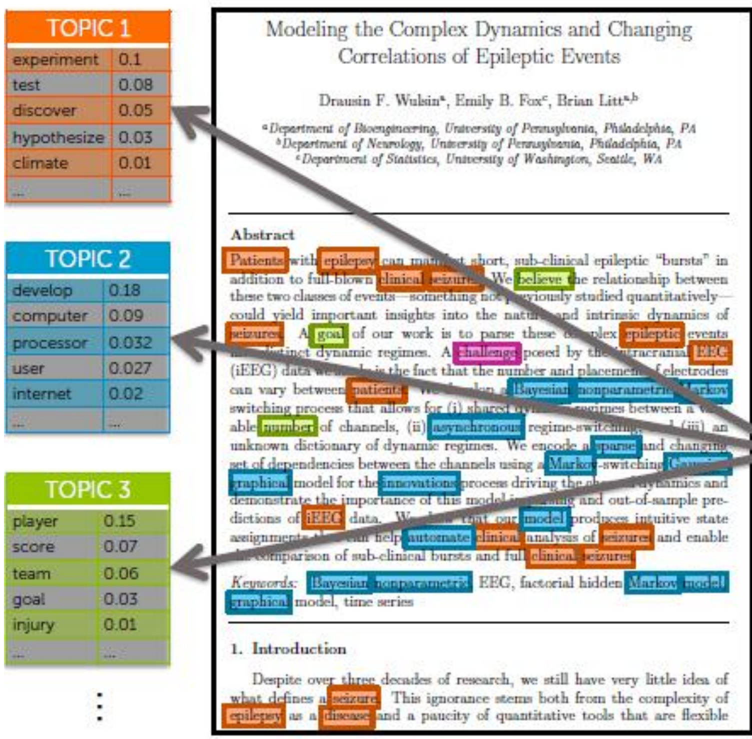
1. Introduction

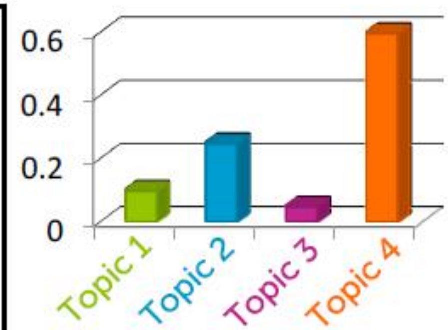
Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible



Inference in LDA models

Interpreting LDA outputs





Examine **coherence** of learned topics

- What are top words per topic?
- Do they form meaningful groups?
- Use to post-facto label topics (e.g., science, tech, sports,...)

Inference in LDA models

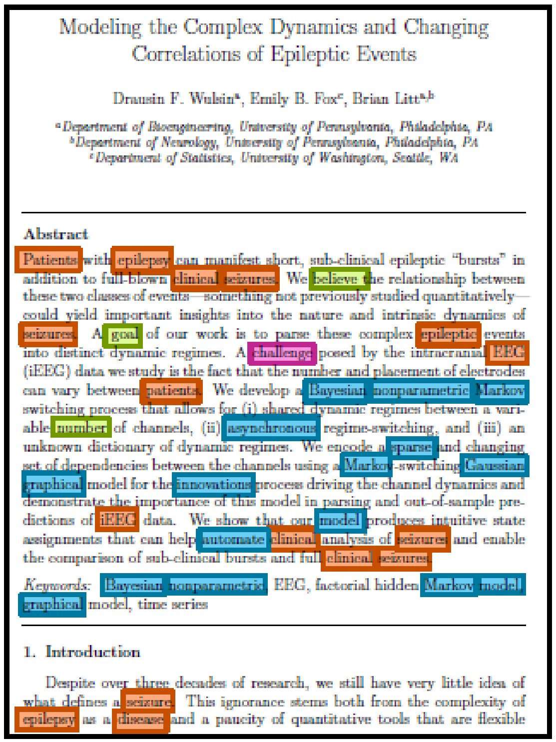
Interpreting LDA outputs

TOPIC 1	
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01
...	...

TOPIC 2	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02
...	...

TOPIC 3	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01
...	...

⋮


Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

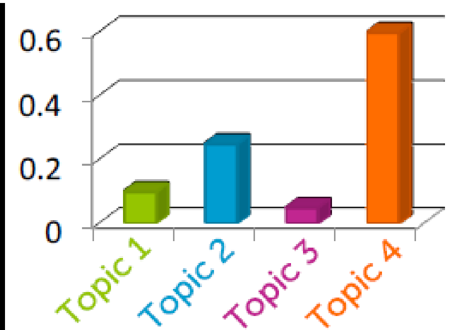
Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible



↑

Doc-specific topic proportions can be used to:

- Relate documents
- Study user topic preferences
- Assign docs to multiple categories

Inference in LDA models

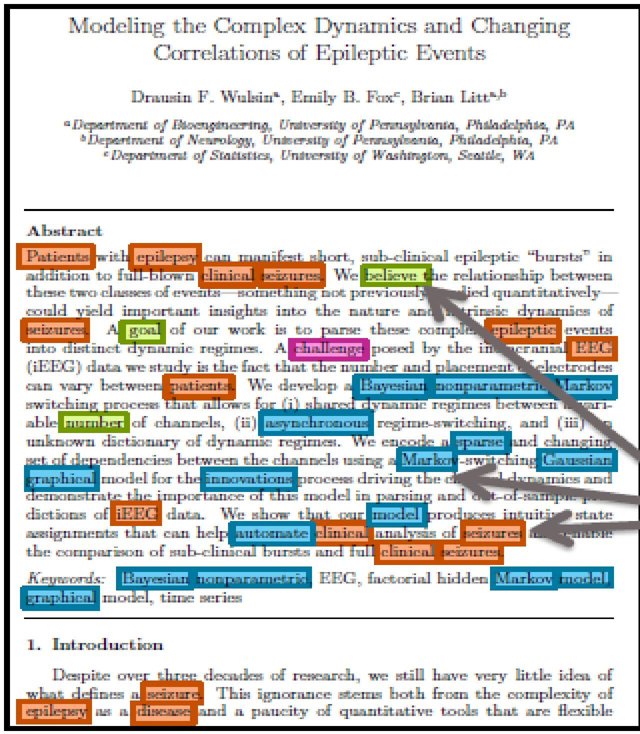
Interpreting LDA outputs

TOPIC 1	
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01
...	...

TOPIC 2	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02
...	...

TOPIC 3	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01
...	...

⋮



Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

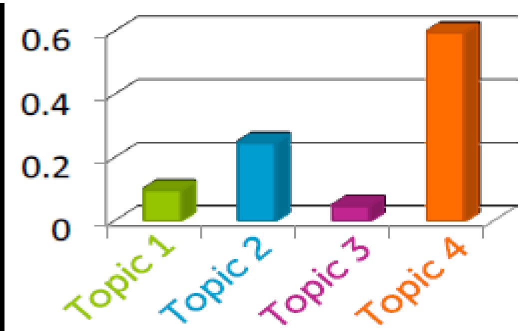
Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously ... quantitatively—could yield important insights into the nature and ... dynamics of seizures. A goal of our work is to parse these compl... epileptic events into distinct dynamic regimes. A challenge posed by the in... EEG (iEEG) data we study is the fact that the number and placement ... electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between ... variable number of channels, (ii) asynchronous regime-switching, and (iii) ... unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the c... dynamics and demonstrate the importance of this model in parsing and ...-of-sample ... dictions of iEEG data. We show that our model produces intuiti... state assignments that can help automate clinical analysis of seizures ... the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible



Typically **not** interested in word assignments

An inference algorithm for LDA: Gibbs sampling

Clustering so far

k-means

Assign observations to closest cluster center

$$z_i \leftarrow \arg\min_j ||\mu_j - \mathbf{x}_i||_2^2$$

Revise cluster centers

$$\mu_j \leftarrow \arg\min_\mu \sum_{i:z_i=j} ||\mu - \mathbf{x}_i||_2^2$$

Iterative **hard** assignment to max objective

EM for MoG

E-step: estimate cluster responsibilities

$$\hat{r}_{ik} = \frac{\hat{\pi}_k N(x_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{j=1}^{K} \hat{\pi}_j N(x_i \mid \hat{\mu}_j, \hat{\Sigma}_j)}$$

M-step: maximize likelihood over parameters

$$\hat{\pi}_k, \hat{\mu}_k, \hat{\Sigma}_k \mid \{\hat{r}_{ik}, x_i\}$$

Iterative **soft** assignment to max objective

What can we do for our bag-of-words models?

Part 1: Clustering model

SCIENCE	
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01
...	...

TECH	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02
...	...

SPORTS	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01
...	...

⋮

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

One topic indicator z_i per **document** i

All words come from (get scored under) same topic z_i

Distribution on prevalence of topics in **corpus**
$\boldsymbol{\pi} = [\pi_1 \ \pi_2 \ldots \pi_K]$

What can we do for our bag-of-words models?

Part 1: Clustering model

SCIENCE	
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01
...	...

TECH	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02
...	...

SPORTS	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01
...	...

⋮

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a] Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b] Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c] Department of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

Can derive EM algorithm:

- Gaussian likelihood of tf-idf vector

↓

multinomial likelihood of word counts (m_w successes of word w)

- **Result:** mixture of multinomial model

What can we do for our bag-of-words models?

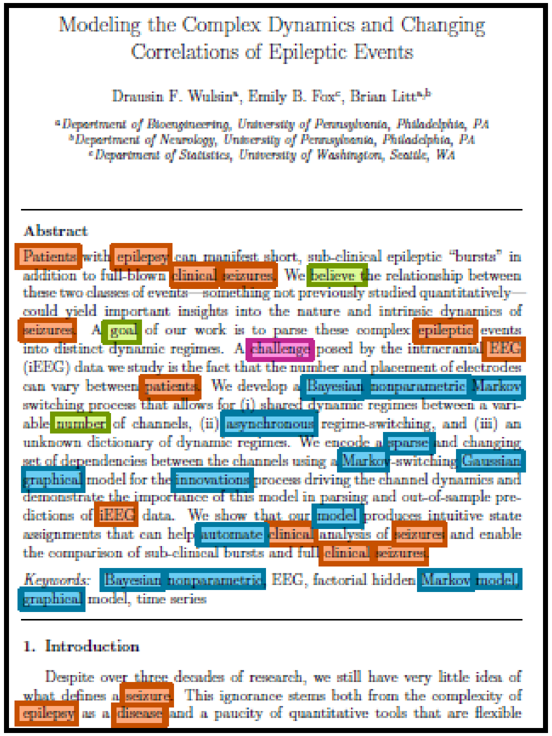
Part 2: LDA model

TOPIC 1	
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01
...	...

TOPIC 2	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02
...	...

TOPIC 3	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01
...	...

⋮



Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a] *Department of Bioengineering, University of Pennsylvania, Philadelphia, PA*
[b] *Department of Neurology, University of Pennsylvania, Philadelphia, PA*
[c] *Department of Statistics, University of Washington, Seattle, WA*

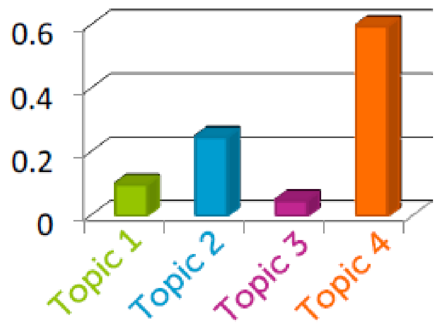
Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible



Can derive
EM algorithm,
but not common
(performs poorly)

An inference algorithms

Typical LDA implementations

Normally LDA is specified as a **Bayesian model**
(otherwise, "probabilistic latent semantic analysis/indexing")

- Account for uncertainty in parameters when making predictions
- Naturally regularizes parameter estimates in contrast to MLE

EM-like algorithms (e.g., "variational EM"), or...

Algorithm for Bayesian inference

Gibbs sampling

Iterative **random** hard assignment!

Benefits:

- Typically intuitive updates
- Very straightforward to implement

Gibbs sampling for LDA

TOPIC 1	
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01
...	...

TOPIC 2	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02
...	...

TOPIC 3	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01
...	...

⋮

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a] *Department of Bioengineering, University of Pennsylvania, Philadelphia, PA*
[b] *Department of Neurology, University of Pennsylvania, Philadelphia, PA*
[c] *Department of Statistics, University of Washington, Seattle, WA*

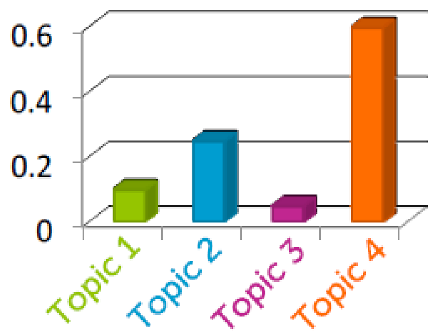
Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible



Current set of assignments

Gibbs sampling for LDA

TOPIC 1	
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01
...	...

TOPIC 2	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02
...	...

TOPIC 3	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01
...	...

⋮

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

Abstract

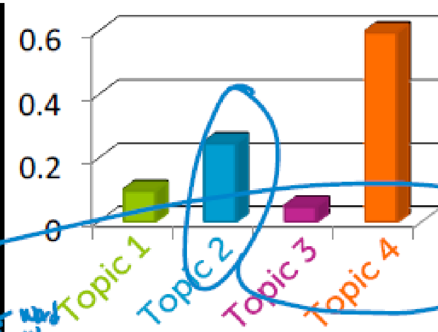
Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

word w in doc i



prior of $z_{iw}=2$

prob. $z_{iw}=2$

$$r_{iw2} = \frac{\pi_{i2} \cdot p(\text{"EEG"} \mid z_{iw}=2)}{\sum_{j=1}^{K} \pi_{ij}\, p(\text{"EEG"} \mid z_{iw}=j)}$$

prob. of assigning $z_{iw}=2$

Step 1: Randomly reassign all z_{iw} based on

- doc topic proportions
- topic vocab distributions

Draw randomly from responsibility vector $[r_{iw1}\ r_{iw2}\ \ldots\ r_{iwK}]$

Gibbs sampling for LDA

TOPIC 1	
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01
...	...

TOPIC 2	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02
...	...

TOPIC 3	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01
...	...

⋮

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a] Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b] Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c] Department of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

Step 2: Randomly reassign doc topic proportions based on assignments z_{iw} in **current doc**

Gibbs sampling for LDA

TOPIC 1	
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01
...	...

TOPIC 2	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02
...	...

TOPIC 3	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01
...	...

⋮

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

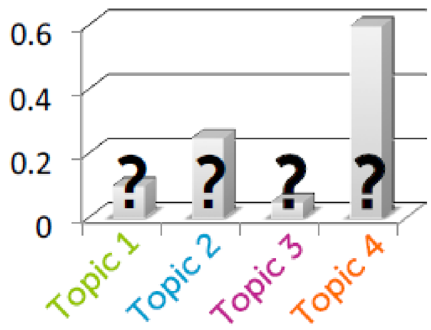
Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible



Step 3: Repeat for all docs

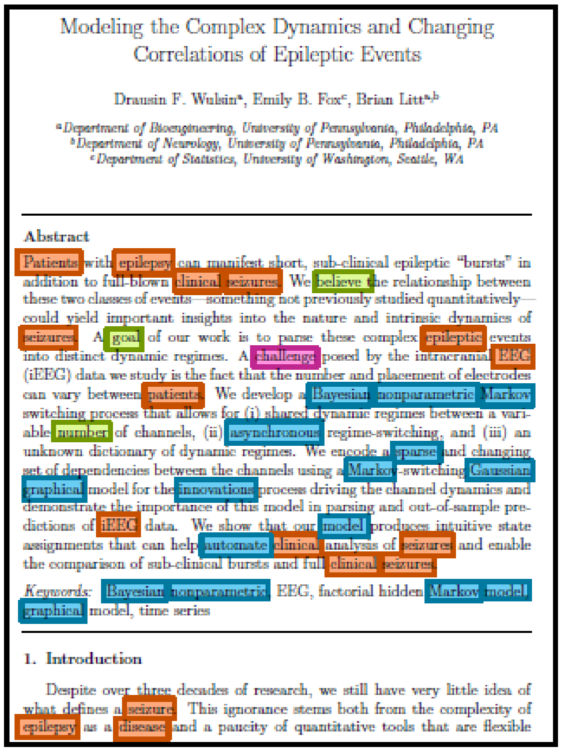
Gibbs sampling for LDA

TOPIC 1	
Word 1	?
Word 2	?
Word 3	?
Word 4	?
Word 5	?
...	...

TOPIC 2	
Word 1	?
Word 2	?
Word 3	?
Word 4	?
Word 5	?
...	...

TOPIC 3	
Word 1	?
Word 2	?
Word 3	?
Word 4	?
Word 5	?
...	...

⋮



Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a] Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b] Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c] Department of Statistics, University of Washington, Seattle, WA

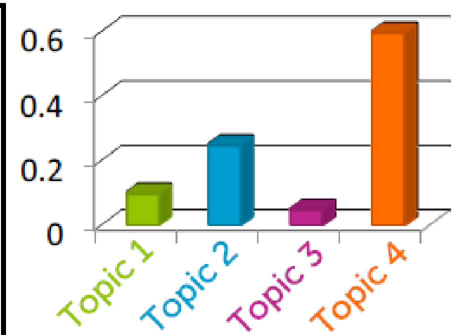
Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible



Step 4: Randomly reassign topic vocab distributions based on assignments z_{iw} in **entire corpus**

An inference algorithm: Gibbs sampling

Gibbs sampling for LDA

TOPIC 1	
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01
...	...

TOPIC 2	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02
...	...

TOPIC 3	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01
...	...

⋮

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a] *Department of Bioengineering, University of Pennsylvania, Philadelphia, PA*
[b] *Department of Neurology, University of Pennsylvania, Philadelphia, PA*
[c] *Department of Statistics, University of Washington, Seattle, WA*

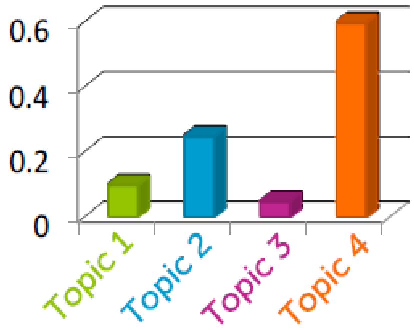
Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible



Repeat Steps 1-4 until max iter reached

An inference algorithm: Gibbs sampling

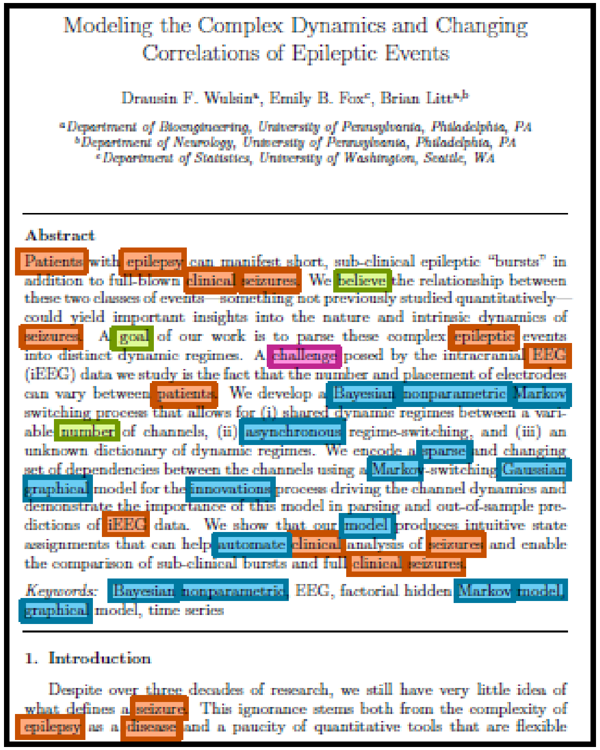
Random sample #10000

TOPIC 1	
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01
...	...

TOPIC 2	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02
...	...

TOPIC 3	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01
...	...

⋮



Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a] Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b] Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c] Department of Statistics, University of Washington, Seattle, WA

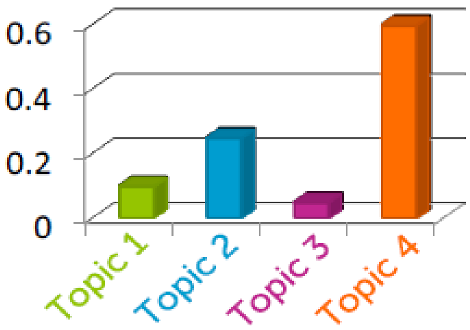
Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible



Current set of assignments

An inference algorithm: Gibbs sampling

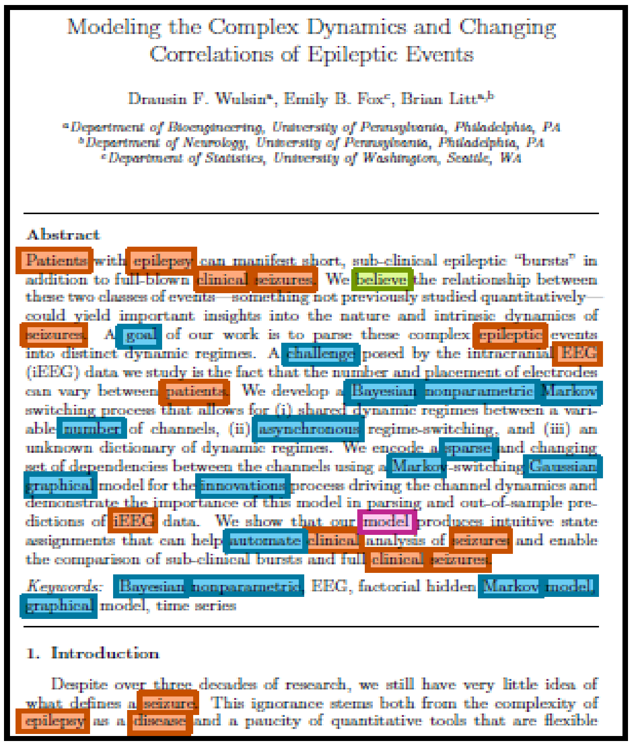
Random sample #10001

TOPIC 1	
experiment	0.12
test	0.06
hypothesize	0.042
discover	0.04
climate	0.011
...	...

TOPIC 2	
develop	0.16
computer	0.11
user	0.03
processor	0.029
internet	0.023
...	...

TOPIC 3	
player	0.15
score	0.07
team	0.06
offense	0.02
defense	0.018
...	...

⋮


Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

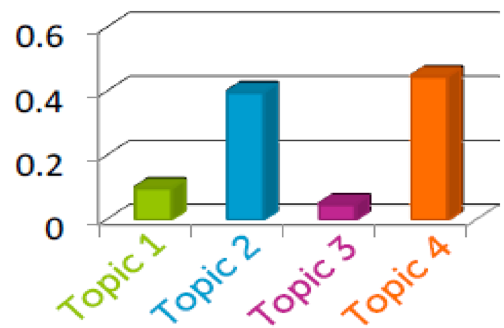
Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible



Current set of assignments

An inference algorithm: Gibbs sampling

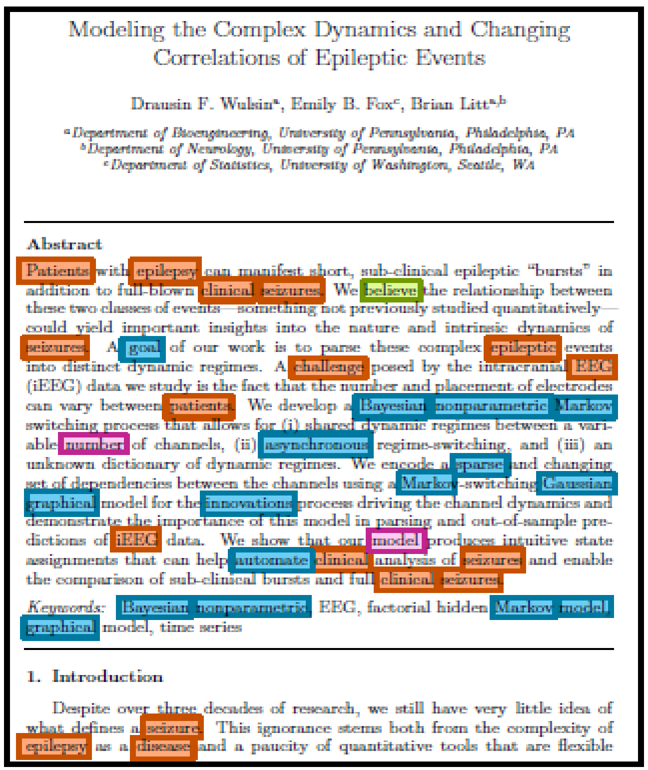
Random sample #10002

TOPIC 1	
experiment	0.10
discover	0.055
hypothesize	0.043
test	0.042
examine	0.015
...	...

TOPIC 2	
computer	0.12
develop	0.115
user	0.031
device	0.022
cloud	0.018
...	...

TOPIC 3	
player	0.17
score	0.09
game	0.062
team	0.043
win	0.03
...	...

⋮



Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

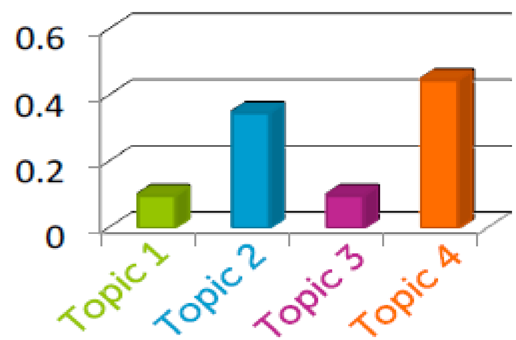
Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametrics, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

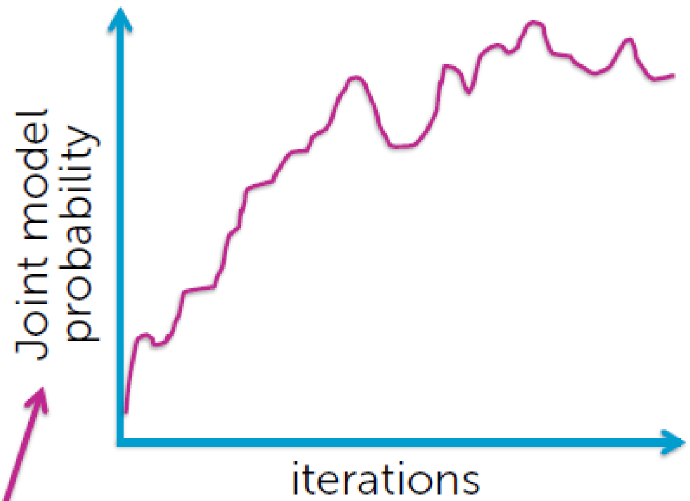


Current set of assignments

An inference algorithm: Gibbs sampling

What do we know about this process?

Not an optimization algorithm



Eventually provides "correct" Bayesian estimates...

probability of observations given variables/parameters and probability of variables/parameters themselves

An inference algorithm: Gibbs sampling

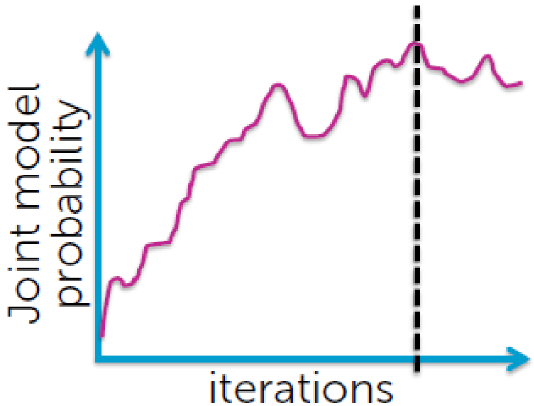
What to do with sampling output?

Predictions:

1. Make prediction for each snapshot of randomly assigned variables/parameters (full iteration)
2. Average predictions for final result

Parameter or assignment estimate:

– Look at snapshot of randomly assigned variables/parameters that maximizes "joint model probability"



Gibbs sampling algorithm

Iterative **random** hard assignment!

Assignment variables and model parameters treated similarly

Iteratively draw variable/parameter from conditional distribution having fixed:

- all other variables/parameters
  - values randomly selected in previous rounds
  - changes from iter to iter
- observations
  - always the same values

„Collapsed" Gibbs sampling for LDA

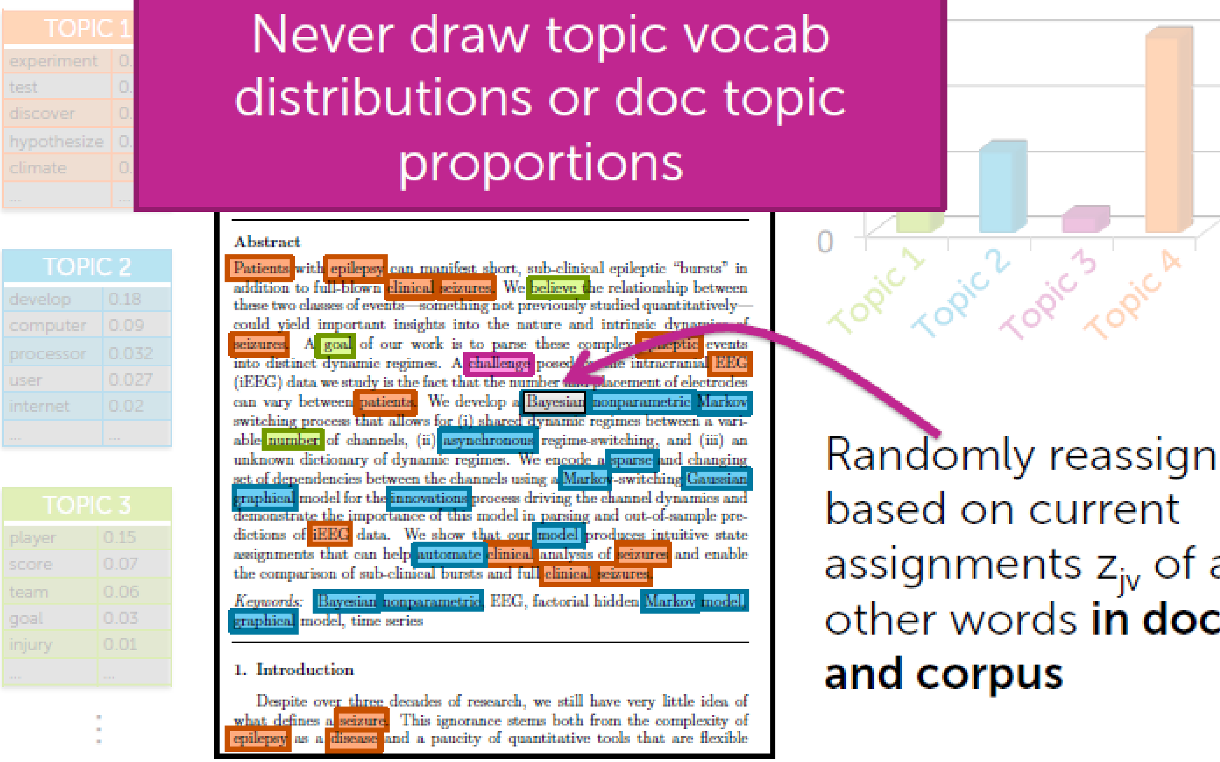
Based on special structure of LDA model, can sample **just** indicator variables z_{iw}

- No need to sample other parameters
  - corpus-wide topic vocab distributions
  - per-doc topic proportions

Often leads to much better performance because examining uncertainty in smaller space

Collapsed Gibbs sampling for LDA

Never draw topic vocab distributions or doc topic proportions



Randomly reassign z_{iw} based on current assignments z_{jv} of all other words **in document and corpus**

Collapsed Gibbs sampling for LDA

Select a document

epilepsy	dynamic	Bayesian	EEG	model

5 word document

Collapsed Gibbs sampling for LDA

Randomly assign topics

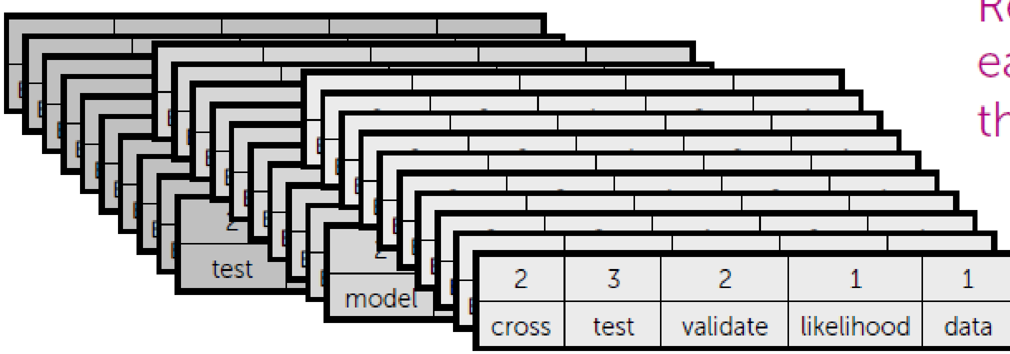
3	2	1	3	1
epilepsy	dynamic	Bayesian	EEG	model

(one possible approach)

Collapsed Gibbs sampling for LDA

Randomly assign topics

3	2	1	3	1
epilepsy	dynamic	Bayesian	EEG	model



Repeat for each doc in the corpus

Collapsed Gibbs sampling for LDA

Maintain local statistics

3	2	1	3	1
epilepsy	dynamic	Bayesian	EEG	model

	Topic 1	Topic 2	Topic 3
Doc i	2	1	2

Collapsed Gibbs sampling for LDA

Maintain global statistics

3	2	1	3	1
epilepsy	dynamic	Bayesian	EEG	model

	Topic 1	Topic 2	Topic 3
Doc i	2	1	2

	Topic 1	Topic 2	Topic 3
epilepsy	1	0	35
Bayesian	50	0	1
model	42	1	0
EEG	0	0	20
dynamic	10	8	1
…			

Total counts from **all** docs

Collapsed Gibbs sampling for LDA

Randomly reassign topics

3	~~2~~	1	3	1
epilepsy	dynamic	Bayesian	EEG	model

	Topic 1	Topic 2	Topic 3
Doc i	2	0 ~~1~~	2

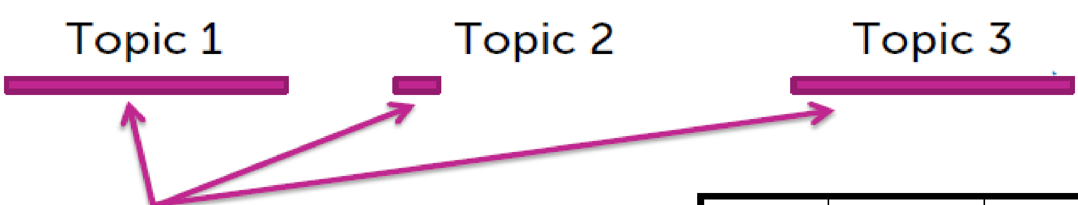
	Topic 1	Topic 2	Topic 3
epilepsy	1	0	35
Bayesian	50	0	1
model	42	1	0
EEG	0	0	20
dynamic	10	7 ~~8~~	1
...			

decrementing counts after removing current assignment $z_{iw}=2$

Collapsed Gibbs sampling for LDA

Probability of new assignment

3	?	1	3	1
epilepsy	dynamic	Bayesian	EEG	model



How much doc "likes" each topic based on other assignments in doc

	Topic 1	Topic 2	Topic 3
Doc i	2	0	2

current assignments to topic k in doc i → $n_{ik} + \alpha$ ← smoothing param from Bayes prior

words in doc i → $N_i - 1 + K\alpha$

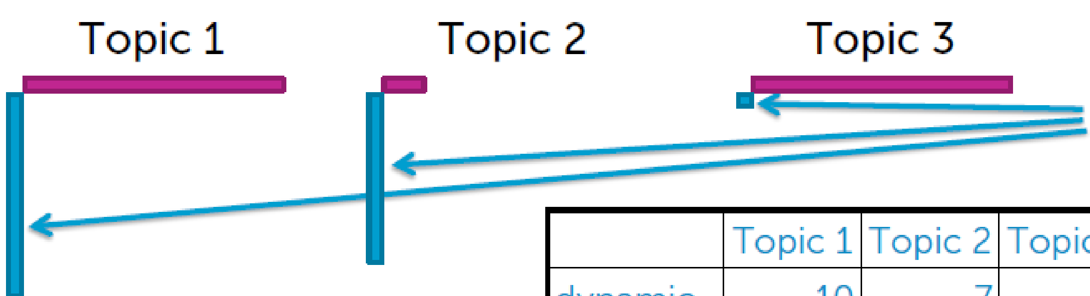
ignore current word

26/11, 3/12, 10/12/2019

Collapsed Gibbs sampling for LDA

Probability of new assignment

3	?	1	3	1
epilepsy	dynamic	Bayesian	EEG	model



How much each topic likes the word "dynamic" based on assignments in other docs in corpus

	Topic 1	Topic 2	Topic 3
dynamic	10	7	1

assignments **corpus-wide** of word "dynamic" to topic k

$$\frac{m_{\text{dynamic},k} + \gamma}{\sum_{w \in V} m_{w,k} + V\gamma}$$

smoothing param from Bayes prior

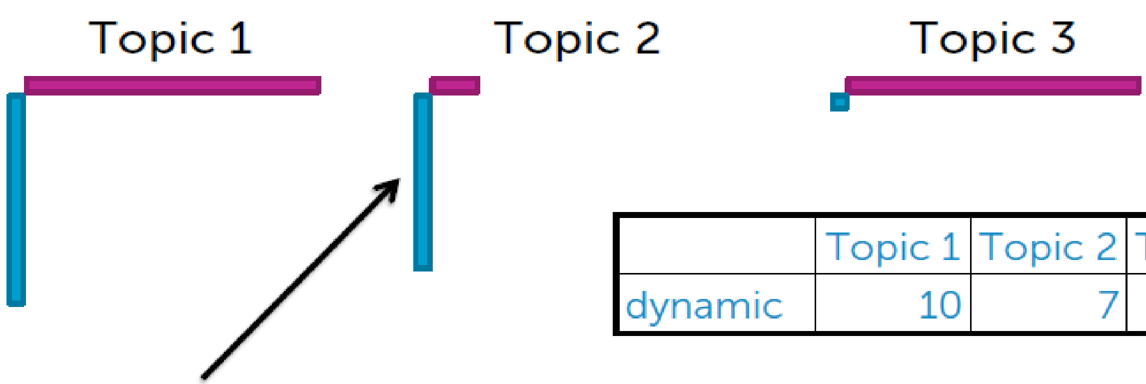
size of vocab

64

Collapsed Gibbs sampling for LDA

Probability of new assignment

3	?	1	3	1
epilepsy	dynamic	Bayesian	EEG	model



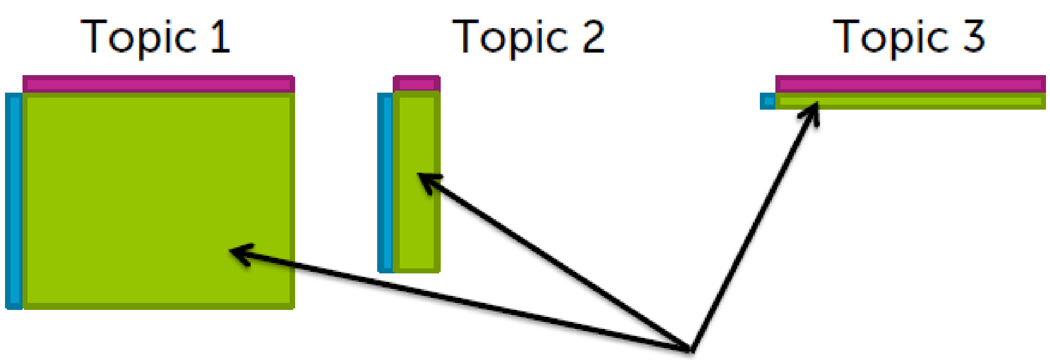
	Topic 1	Topic 2	Topic 3
dynamic	10	7	1

Topic 2 also really likes "dynamic",
but in a different context…
e.g., a topic on fluid dynamics

Collapsed Gibbs sampling for LDA

Probability of new assignment

3	?	1	3	1
epilepsy	dynamic	Bayesian	EEG	model



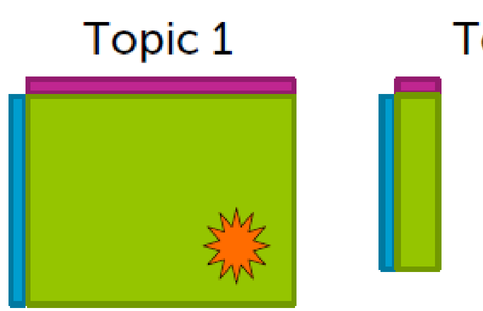
How much doc likes topic

$$\frac{n_{ik} + \alpha}{N_i - 1 + K\alpha} \; \frac{m_{\text{dynamic},k} + \gamma}{\sum_{w \in V} m_{w,k} + V\gamma}$$

How much topic likes word

Randomly draw a new topic indicator

3	?	1	3	1
epilepsy	dynamic	Bayesian	EEG	model



To draw new topic assignment (equivalently):
- roll K-sided die with these probabilities
- throw dart at these regions

Normalize this product of terms over K possible topics!

How much doc likes topic

$$\frac{n_{ik} + \alpha}{N_i - 1 + K\alpha} \quad \frac{m_{\text{dynamic},k} + \gamma}{\sum_{w \in V} m_{w,k} + V\gamma}$$

How much topic likes word

Collapsed Gibbs sampling for LDA

Update counts

3	1	1	3	1
epilepsy	dynamic	Bayesian	EEG	model

	Topic 1	Topic 2	Topic 3
epilepsy	1	0	35
Bayesian	50	0	1
model	42	1	0
EEG	0	0	20
dynamic	11 ~~10~~	7	1
…			

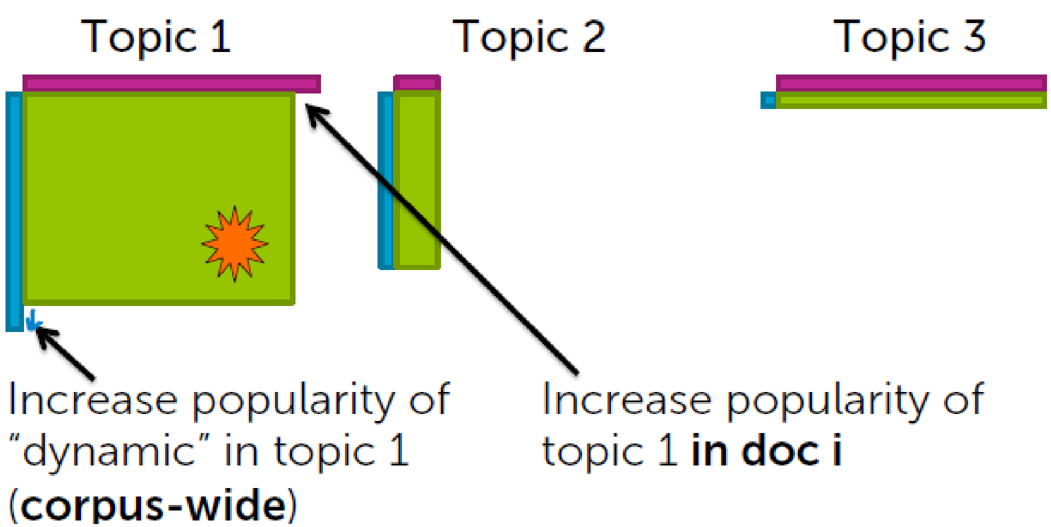
	Topic 1	Topic 2	Topic 3
Doc i	3 ~~2~~	0	2

increment counts based on new assignment of $z_{iw}=1$

Collapsed Gibbs sampling for LDA

Geometrically…

3	1	1	3	1
epilepsy	dynamic	Bayesian	EEG	model



Collapsed Gibbs sampling for LDA

Iterate through all words/docs

3	1	1	3	1
epilepsy	dynamic	Bayesian	EEG	model

	Topic 1	Topic 2	Topic 3
Doc i	2	0	2

	Topic 1	Topic 2	Topic 3
epilepsy	1	0	35
Bayesian	50	0	1
model	42	1	0
EEG	0	0	20
dynamic	10	7	1
...			

Collapsed Gibbs sampling for LDA

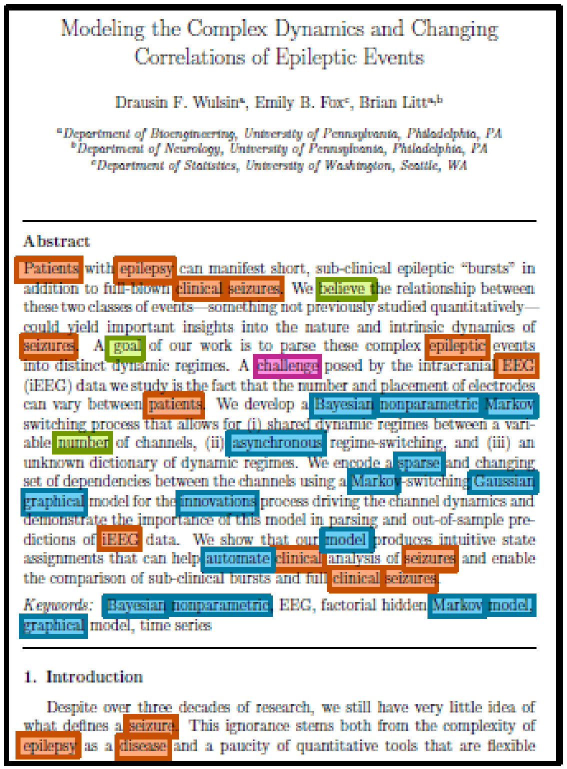
What to do with the collapsed samples?

TOPIC 1	
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01
...	...

TOPIC 2	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02
...	...

TOPIC 3	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01
...	...

⋮



Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

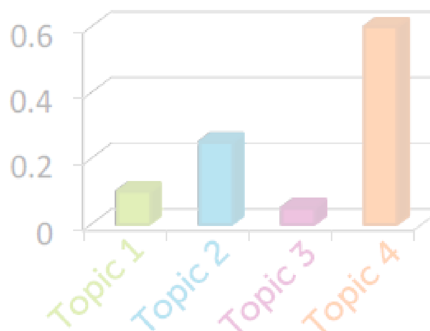
Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible



From "best" sample of $\{z_{iw}\}$, can infer:

1. Topics from conditional distribution…
   need corpus-wide info

Collapsed Gibbs sampling for LDA

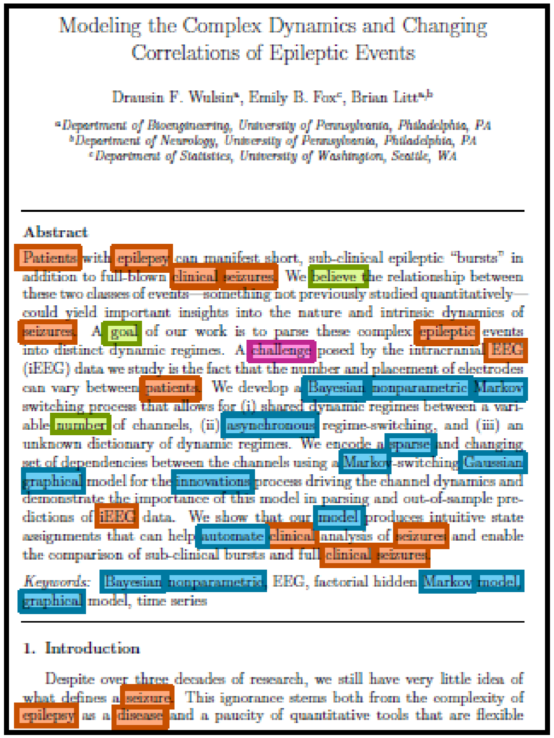
What to do with the collapsed samples?

TOPIC 1	
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01
...	...

TOPIC 2	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02
...	...

TOPIC 3	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01
...	...

⋮


Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b]Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c]Department of Statistics, University of Washington, Seattle, WA

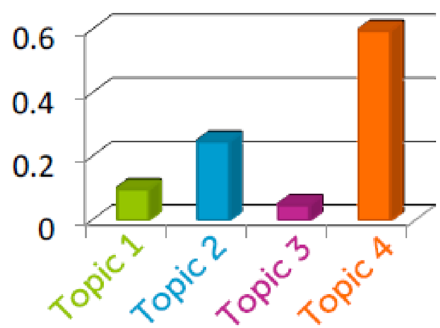
Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametrics, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

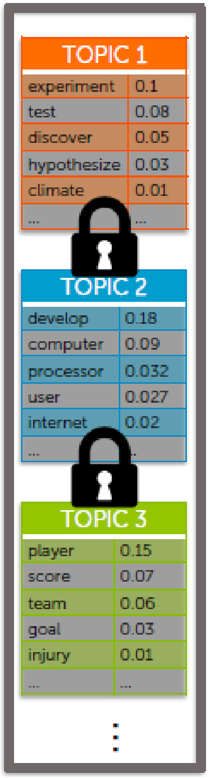


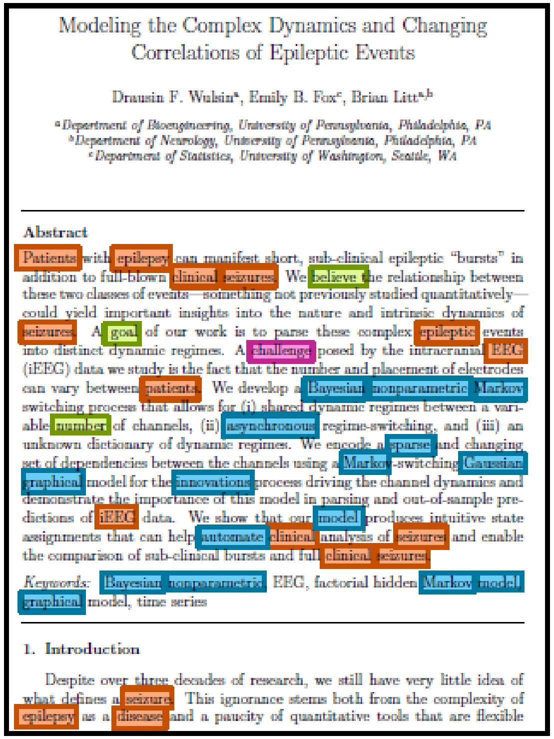
From "best" sample of $\{z_{iw}\}$, can infer:

1. Topics from conditional distribution…
   need corpus-wide info
2. Document "embedding"…
   need doc info only

Collapsed Gibbs sampling for LDA

Embedding new documents





Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin[a], Emily B. Fox[c], Brian Litt[a,b]

[a] Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[b] Department of Neurology, University of Pennsylvania, Philadelphia, PA
[c] Department of Statistics, University of Washington, Seattle, WA

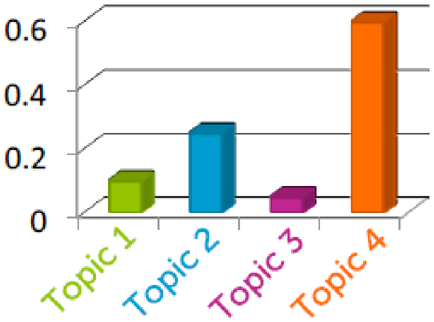
Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic “bursts” in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible



Simple approach:

1. Fix topics based on training set collapsed sampling
2. Run uncollapsed sampler on new doc(s) only

What you can do now

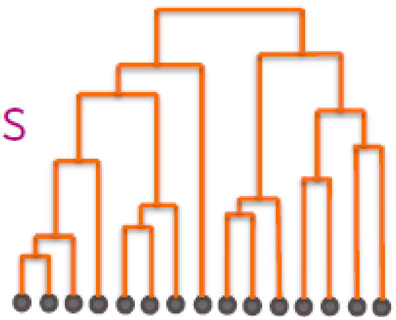
- Compare and contrast clustering and mixed membership models
- Describe a document clustering model for the bag-of-words doc representation
- Interpret the components of the LDA mixed membership model
- Analyze a learned LDA model
  - Topics in the corpus
  - Topics per document
- Describe Gibbs sampling steps at a high level
- Utilize Gibbs sampling output to form predictions or estimate model parameters
- Implement collapsed Gibbs sampling for LDA

Hierarchical clustering

Why hierarchical clustering

- Avoid choosing # clusters beforehand

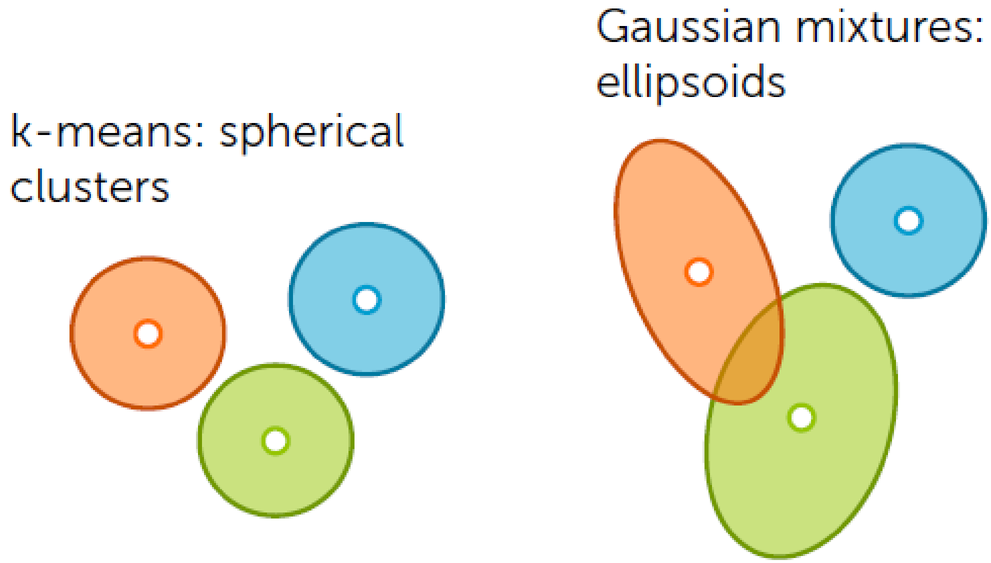
- Dendrograms help visualize different clustering granularities
  - No need to rerun algorithm



- Most algorithms allow user to choose any distance metric
  - k-means restricted us to Euclidean distance

Why hierarchical clustering

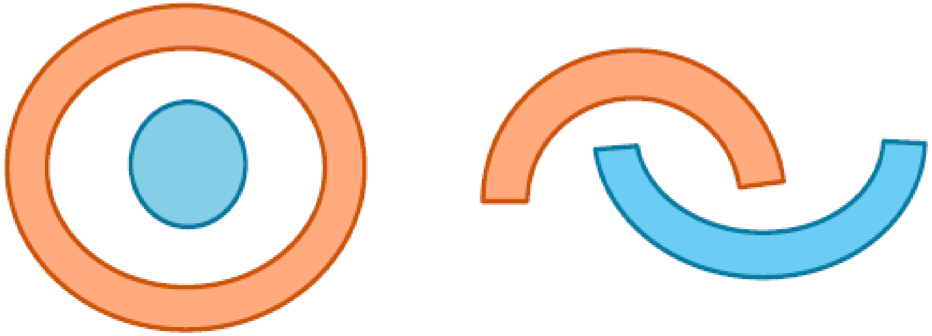
Can often find more complex shapes than k-means or Gaussian mixture models



Why hierarchical clustering

Can often find more complex shapes than k-means or Gaussian mixture models

What about these?



Two main types of algorithms

Divisive, *a.k.a top-down:* Start with all data in one big cluster and recursively split.

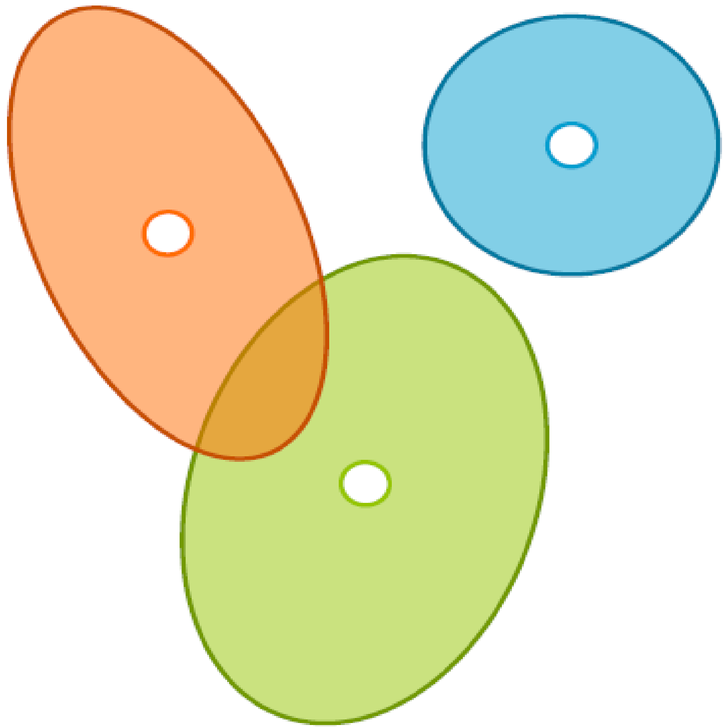
- Example: recursive k-means

Agglomerative *a.k.a. bottom-up*: Start with each data point as its own cluster. Merge clusters until all points are in one big cluster.

- Example: single linkage

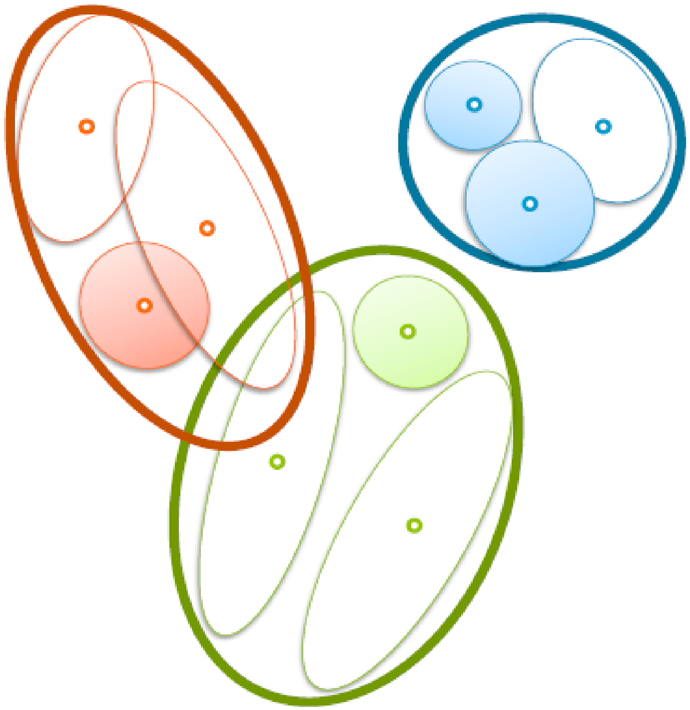
Divisive clustering

Divisive in pictures – level 1

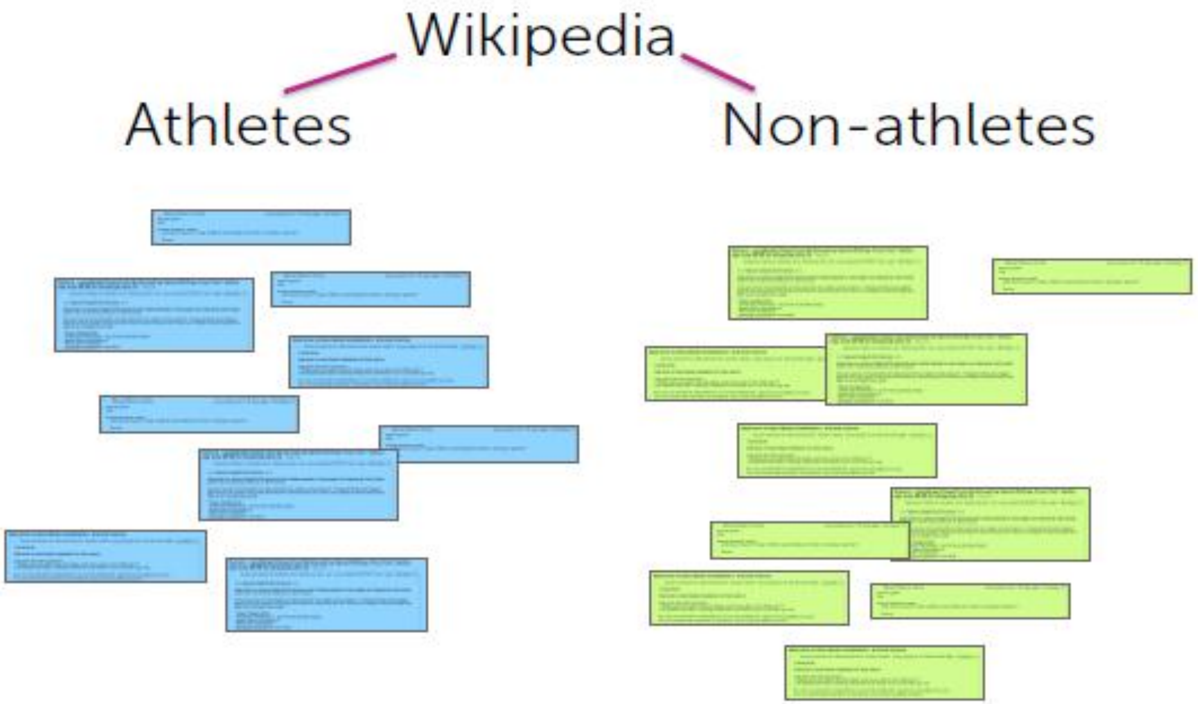


Divisive clustering

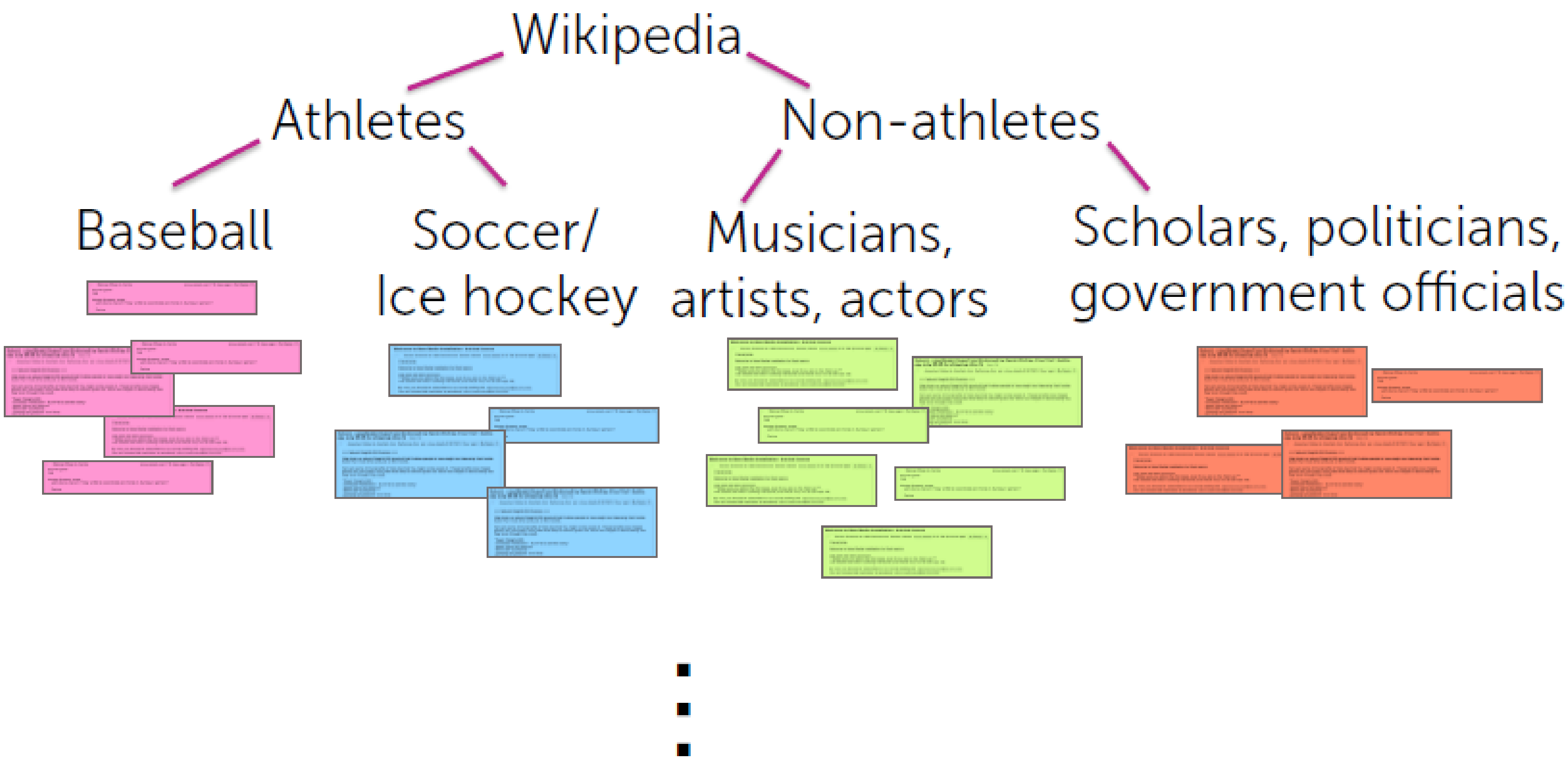
Divisive in pictures – level 2



Divisive: Recursive k-means



Divisive: Recursive k-means

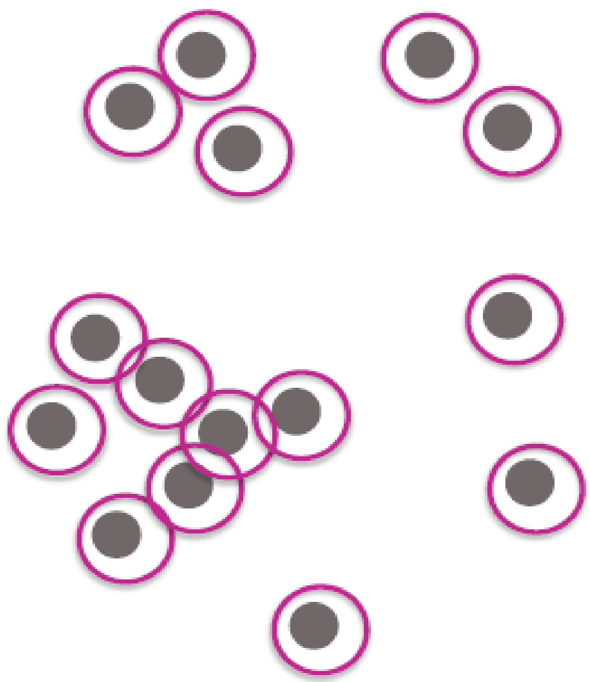


Divisive: choices to be made

- Which algorithm to recurse
- How many clusters per split
- When to split vs. stop
  - Max cluster size:
    number of points in cluster falls below threshold
  - Max cluster radius:
    distance to furthest point falls below threshold
  - Specified # clusters:
    split until pre-specified # clusters is reached

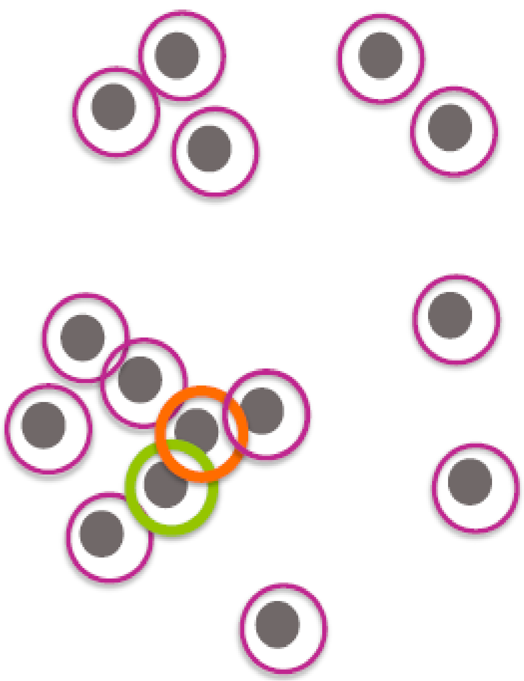
Aglomerative: Single linkage

1. Initialize each point to be its own cluster



Aglomerative: Single linkage

2. Define distance between clusters to be:



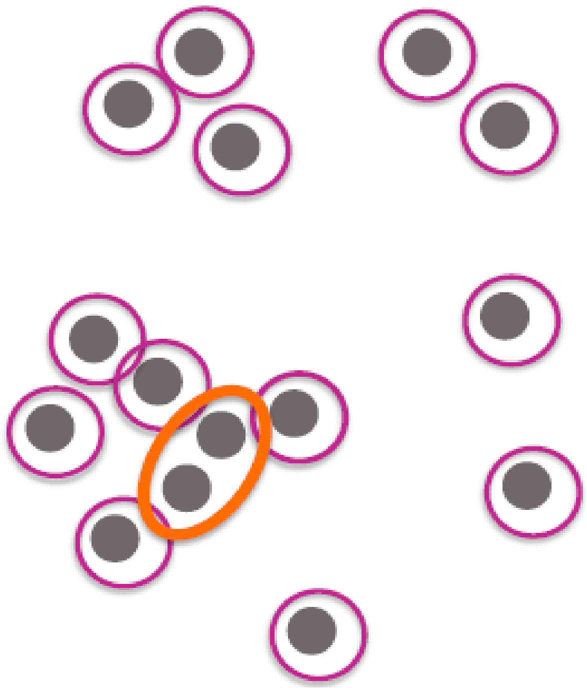
$$\text{distance}(C_1,C_2) = \boxed{\min_{\substack{\mathbf{x}_i \text{ in } C_1,\\ \mathbf{x}_j \text{ in } C_2}} d(\mathbf{x}_i, \mathbf{x}_j)}$$

specified pairwise distance function

Linkage criteria

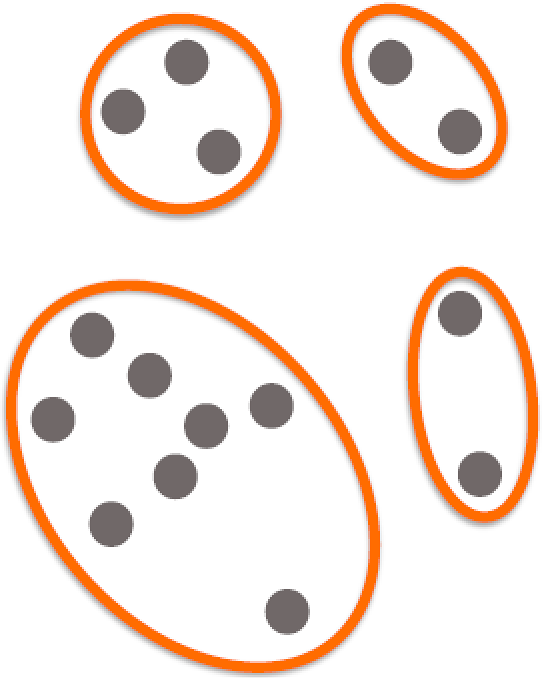
Aglomerative: Single linkage

3. Merge the two closest clusters



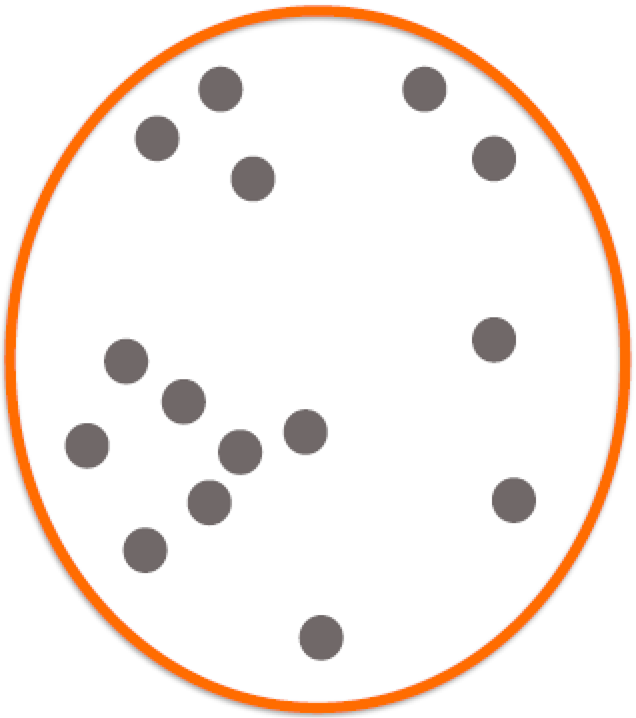
Aglomerative: Single linkage

4. Repeat step 3 until all points are in one cluster



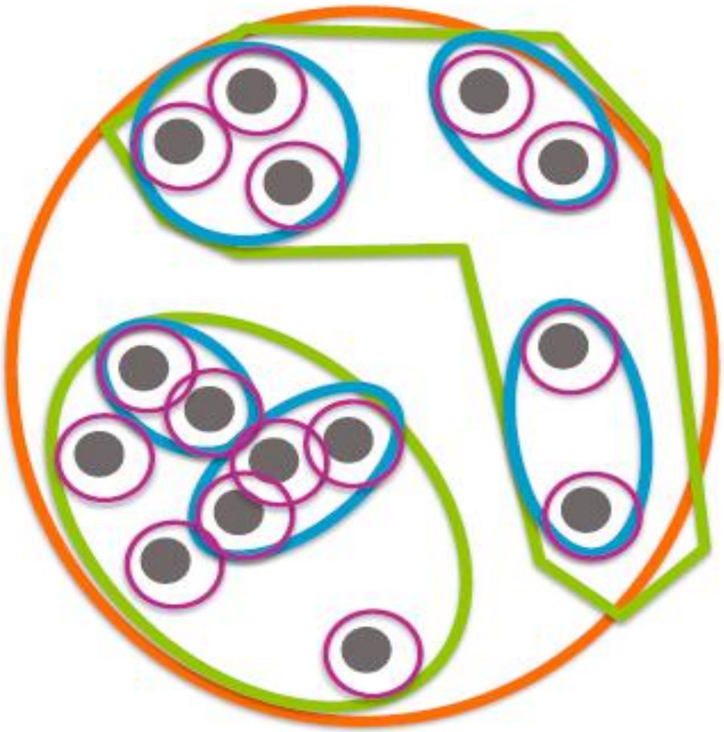
Aglomerative: Single linkage

4. Repeat step 3 until all points are in one cluster



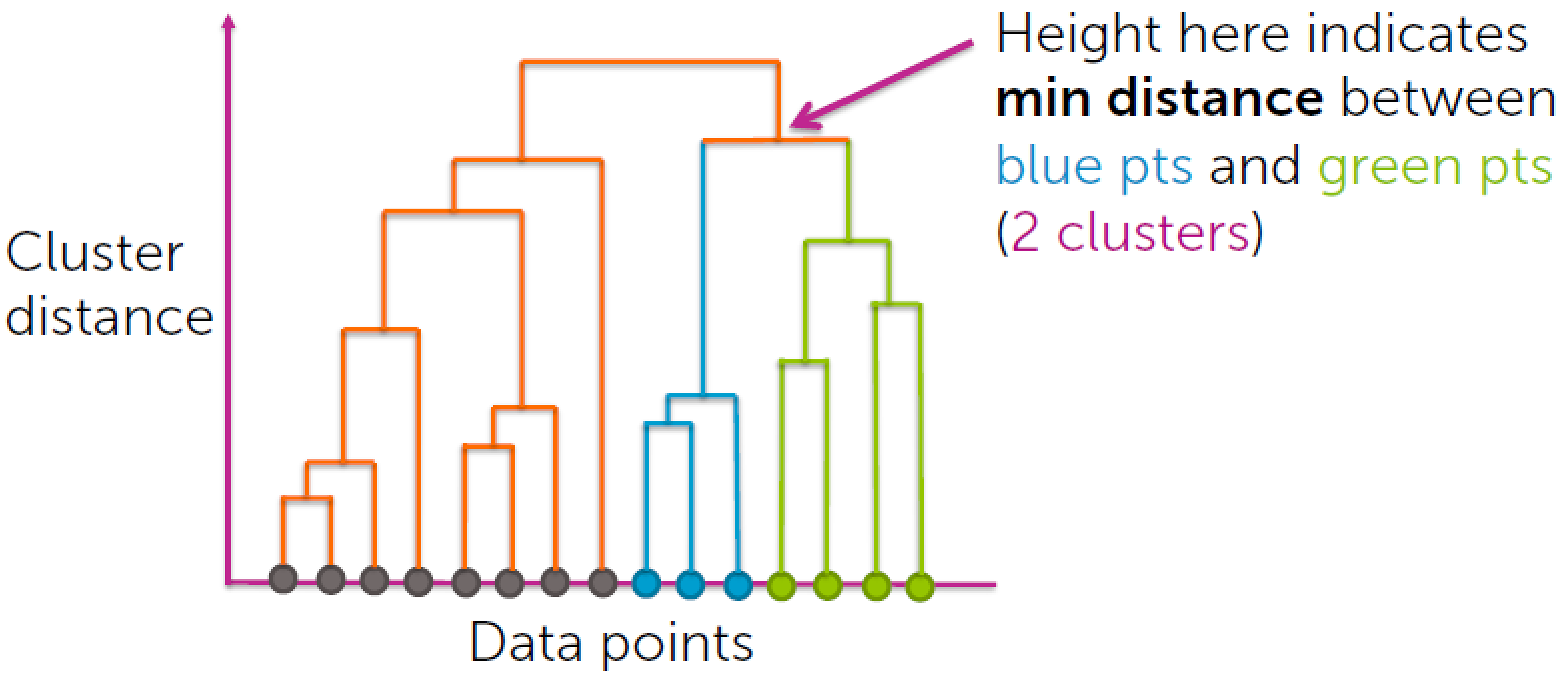
Cluster of clusters

Just like our picture for divisive clustering...



The dendrogram

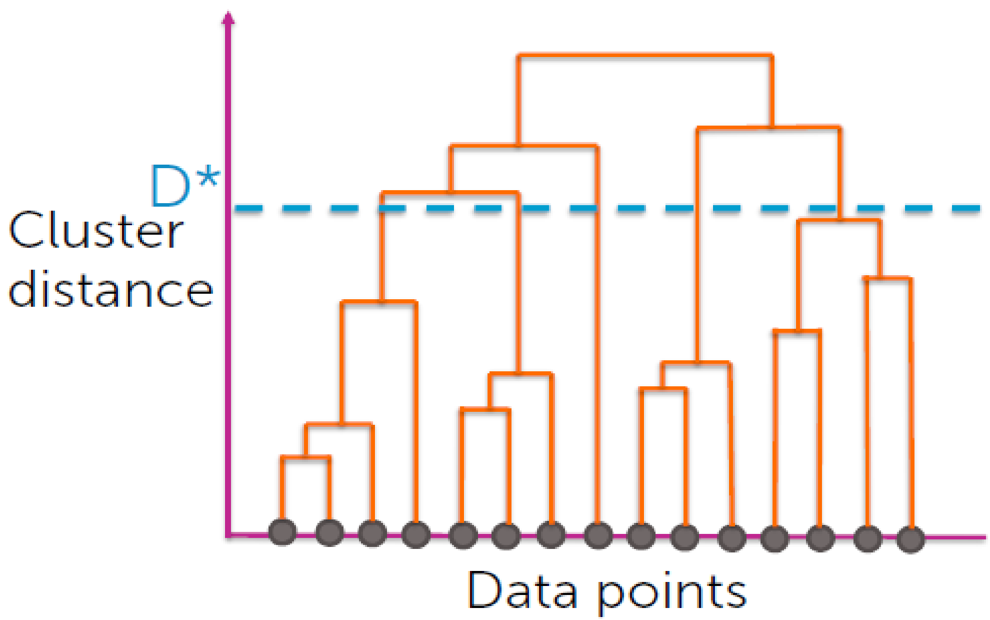
- x axis shows data points (carefully ordered)
- y-axis shows distance between pair of clusters



Extracting a partition

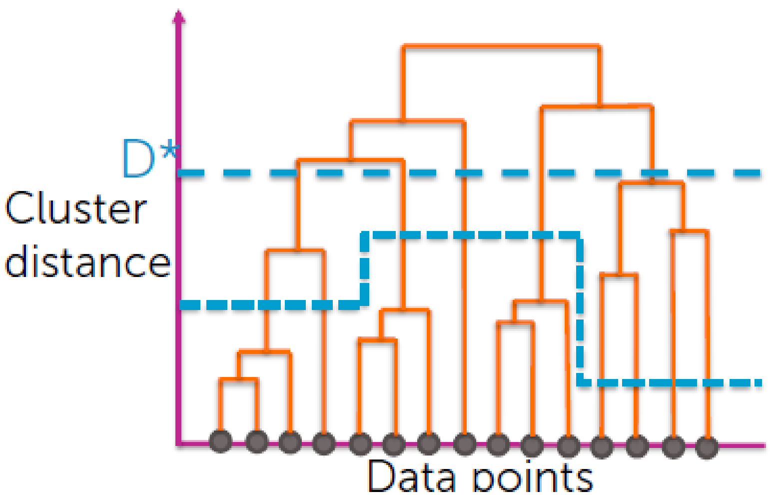
Choose a distance D* at which to cut dendogram

Every branch that crosses D* becomes a separate cluster



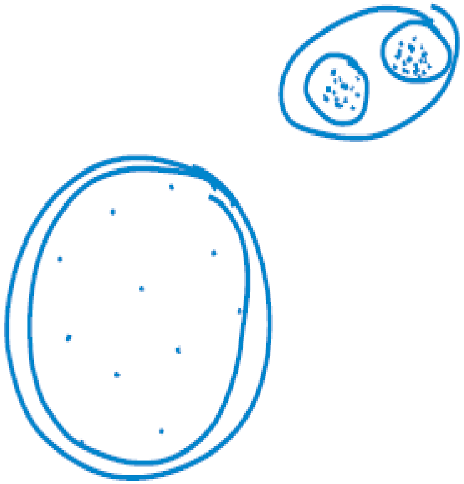
Agglomerative: choices to be made

- Distance metric: $d(\mathbf{x}_i, \mathbf{x}_j)$
- Linkage function: e.g., $\min_{\mathbf{x}_i \text{ in } C_1,\ \mathbf{x}_j \text{ in } C_2} d(\mathbf{x}_i, \mathbf{x}_j)$
- Where and how to cut dendrogram



More on cutting dendrogram

- For visualization, smaller # clusters is preferable
- For tasks like outlier detection, cut based on:
  - Distance threshold
  - Inconsistency coefficient
    - Compare height of merge to average merge heights below
    - If top merge is substantially higher, then it is joining two subsets that are relatively far apart compared to the members of each subset internally
    - Still have to choose a threshold to cut at, but now in terms of "inconsistency" rather than distance
- No cutting method is "incorrect", some are just more useful than others



Computational considerations

- Computing all pairs of distances is expensive
  - Brute force algorithm is $O(N^2 \log(N))$

    # datapoints

- Smart implementations use triangle inequality to rule out candidate pairs

- Best known algorithm is $O(N^2)$