

Recap & Look ahead

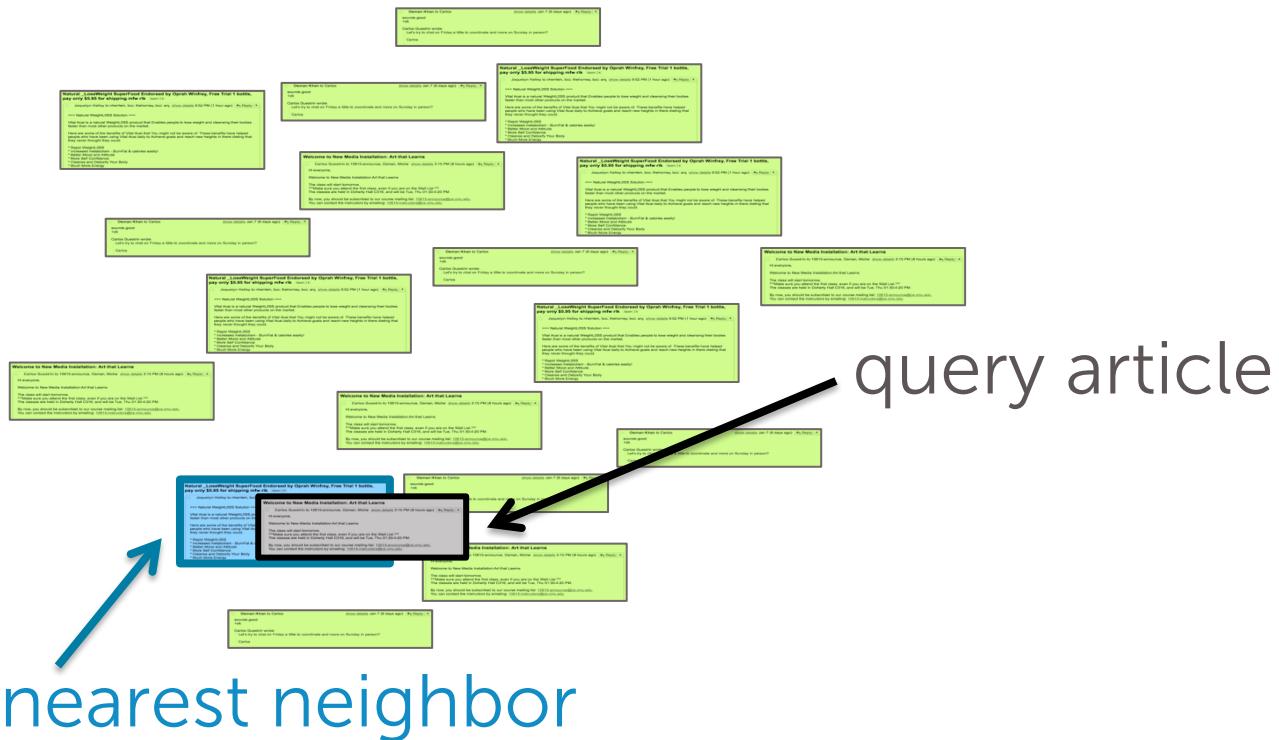
Emily Fox & Carlos Guestrin
Machine Learning Specialization
University of Washington

What we've learned

Module 1: Nearest neighbor search

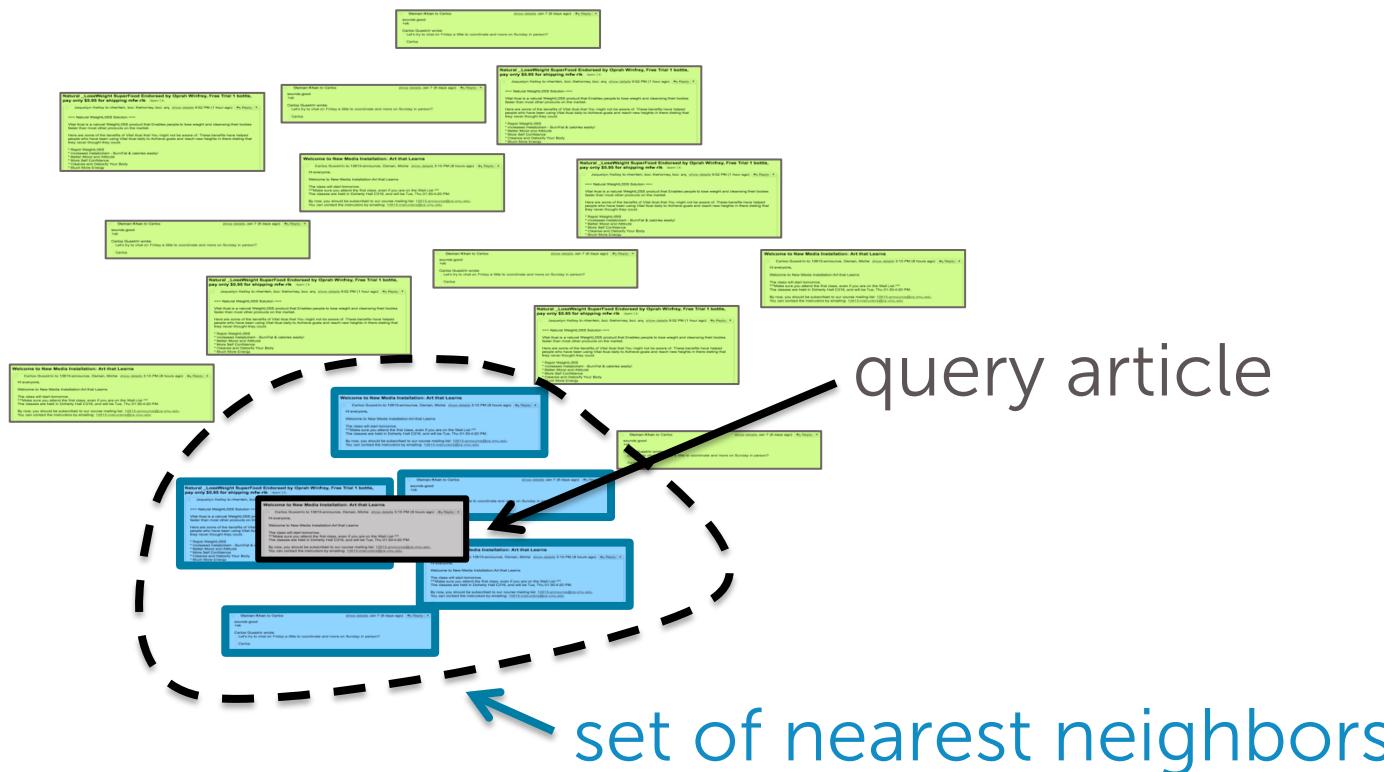
1-NN search

Space of all articles,
organized by similarity of text



k-NN search

Space of all articles,
organized by similarity of text



TF-IDF document representation

Emphasizes **important words**

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

Term frequency = 

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

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



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

$tf * idf$

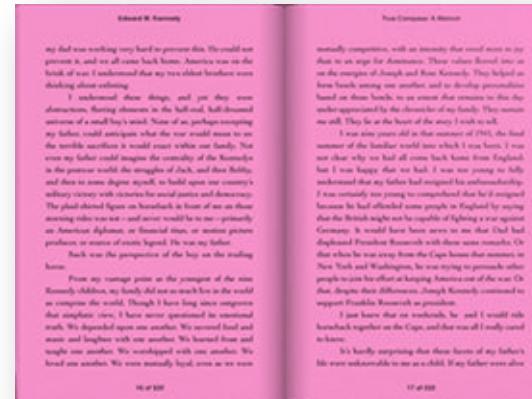
Scaled Euclidean distance

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

weight on each feature



**title
abstract
main body
conclusion**

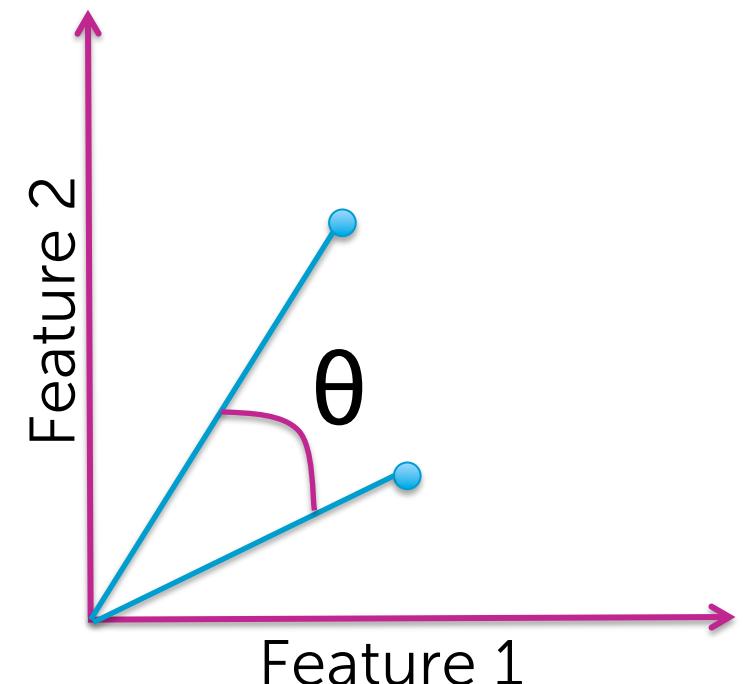


Cosine similarity – normalize

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

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

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



To normalize or not?

short tweet

Normalizing can make dissimilar objects appear more similar

long document

my dad was working very hard to get this done. He could not prove it, and we all knew he was right. America was the last country to give us our freedom. We were the last ones to stand up and say "no more".

My dad was a good man, and yet his own shortcomings showed through in the self, and the self-centered nature of a small boy's mind. Now as of today, we are still here, and we are still here because of the love and support of the people around us. The world would not be the same without our family. Our family is the most important thing in the world. We are the most important thing in the world, no argument of such, and this Melvin and I have been here to help build our country's culture, and to help our country grow and move forward.

The plain old truth is here because of all of us as there are millions of us here, and now we will be here to protect our freedom, and to keep our country safe from any kind of producer, or even of course, terrorist. I was my dad's favorite son.

Such was the progression of time.

Now, my comparison to the rest of the family members of ours, is that we are here to help the world complete the war. We have long since completed the war, and we are here to help the world complete the war. We helped open our doors. We helped feed and care, and laughter with him. We learned from him, and he learned from us. We are here to help the world complete the war. We are here to help the world complete the war.

It is really surprising how much love we are given.

long document

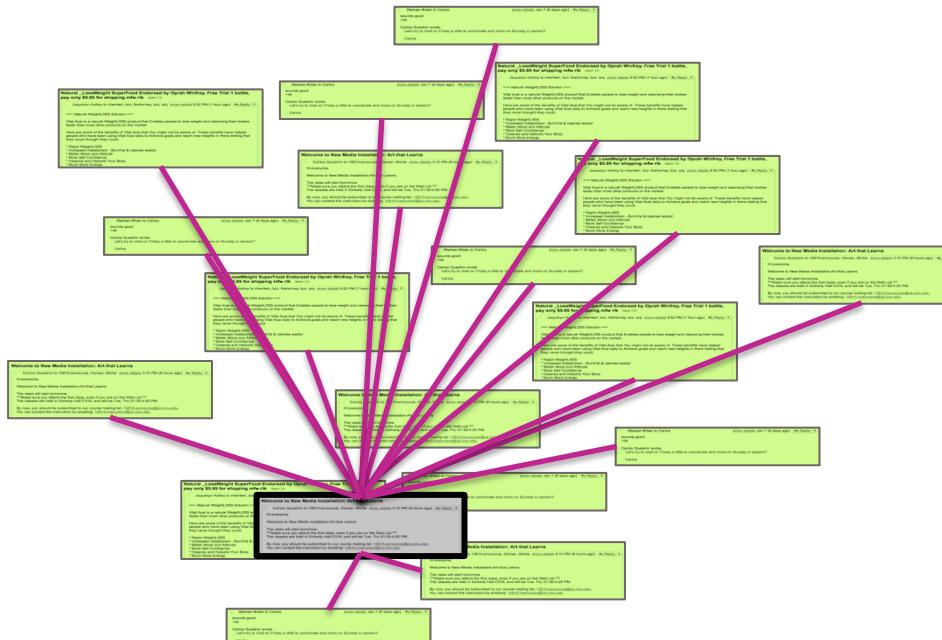
long document

Common compromise: Just cap maximum word counts

Complexity of brute-force search

Given a query point, scan through each point

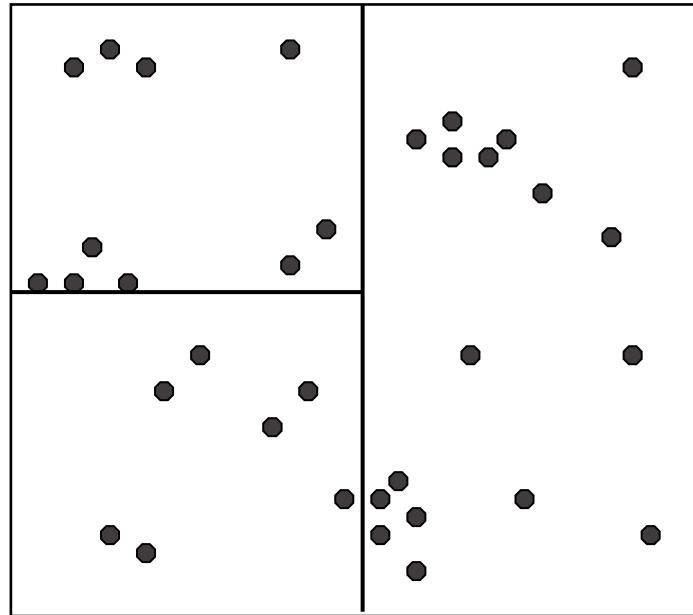
- $O(N)$ distance computations per 1-NN query!
 - $O(N \log k)$ per k -NN query!



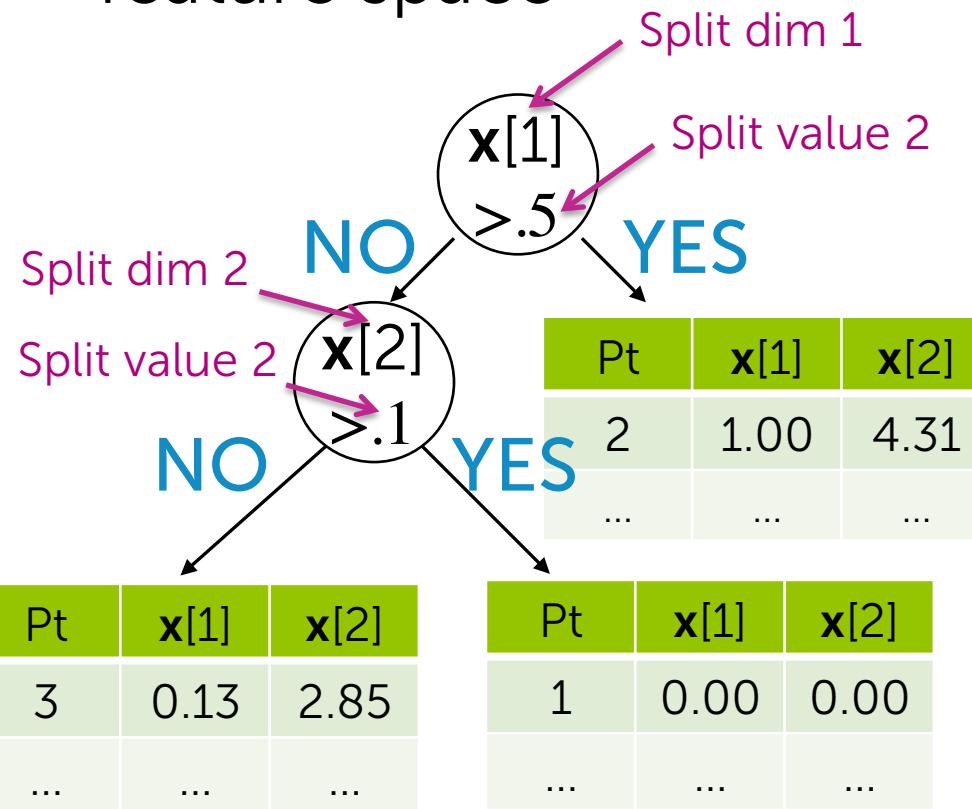
What if N is huge???

(and many queries)

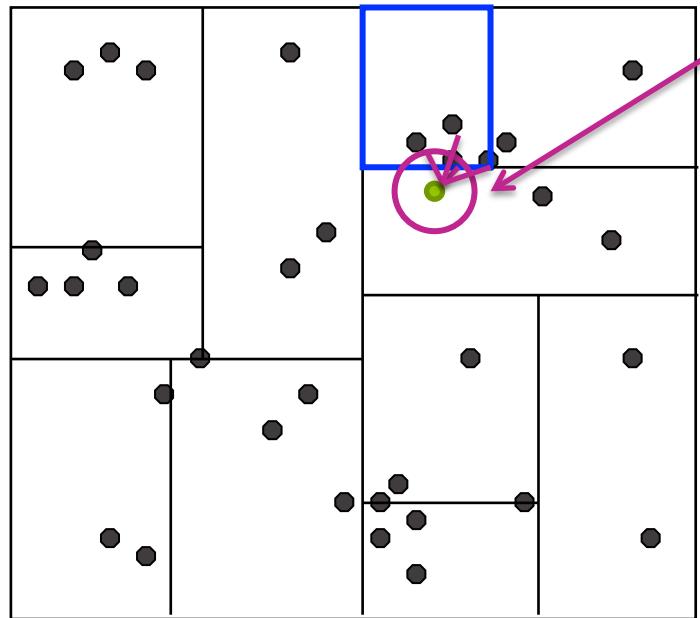
KD-trees



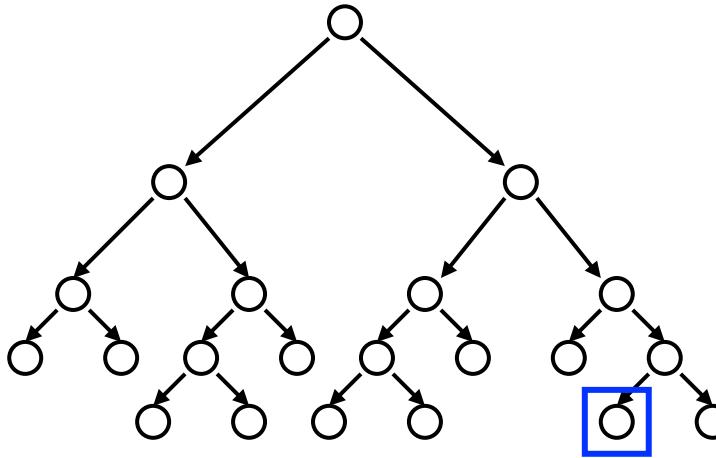
Recursively partition the feature space



Nearest neighbor with KD-trees

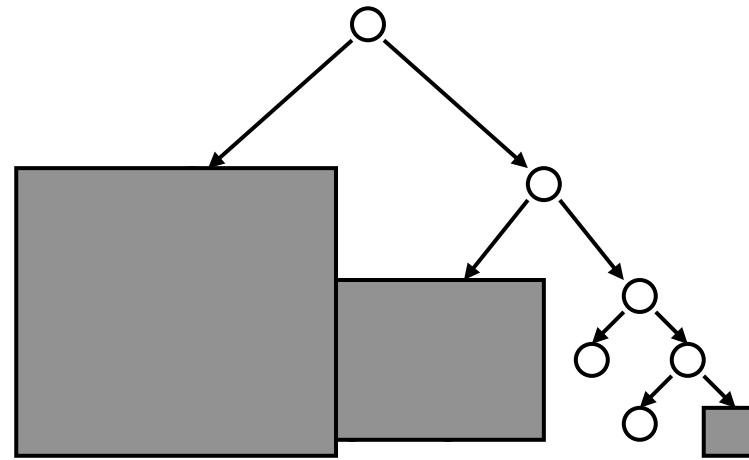
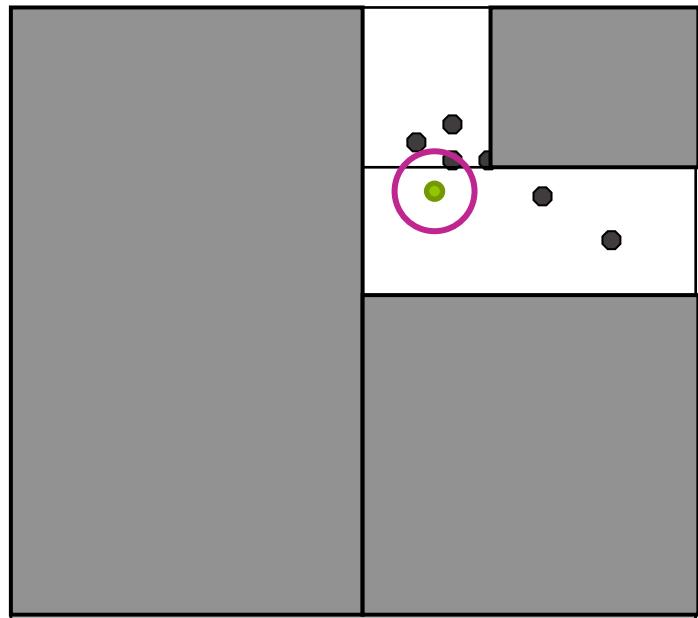


Update distance bound when new nearest neighbor is found



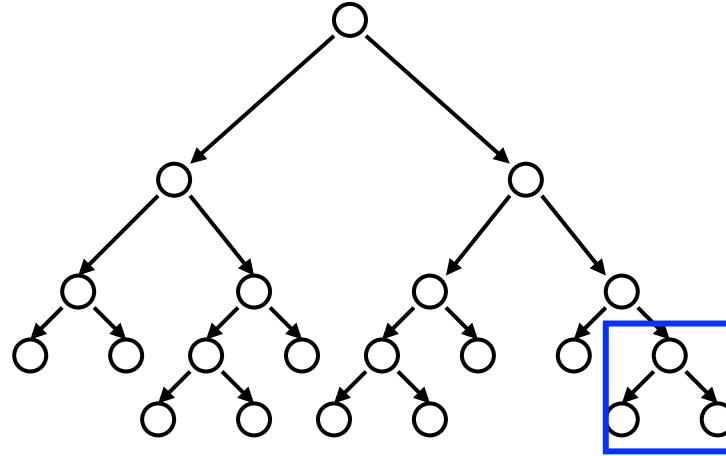
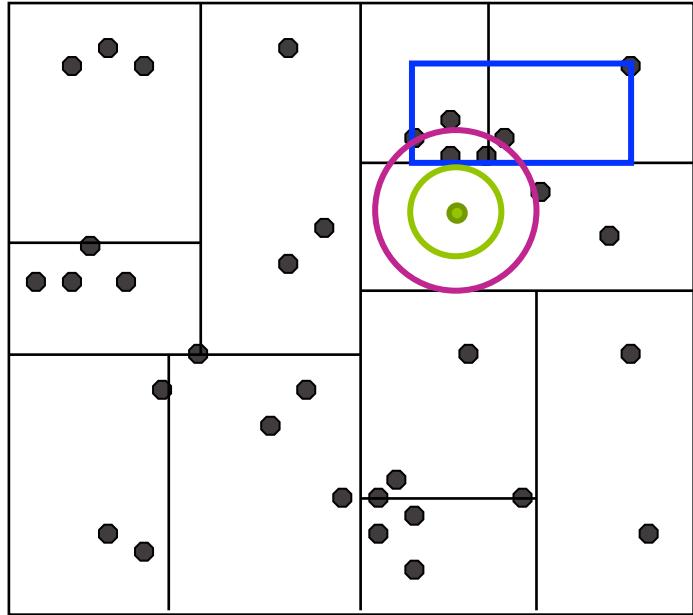
1. Start by exploring leaf node containing query point
2. Compute distance to each other point at leaf node
3. Backtrack and try other branch at each node visited

Nearest neighbor with KD-trees



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

Approximate k-NN with KD-trees



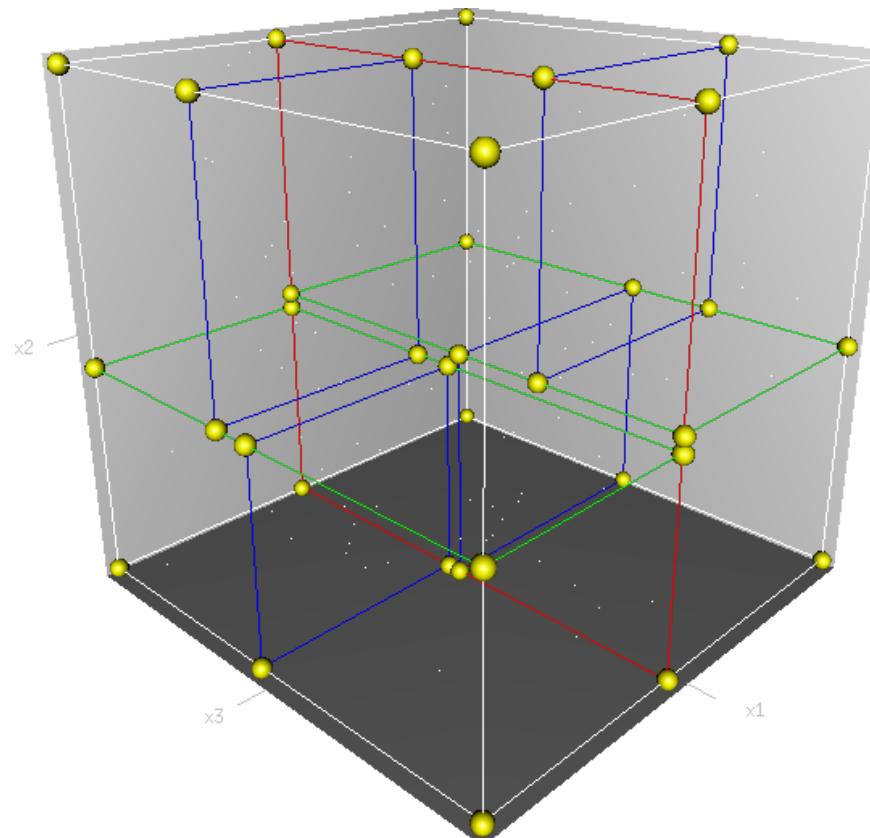
Before: Prune when distance to bounding box $> r$

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

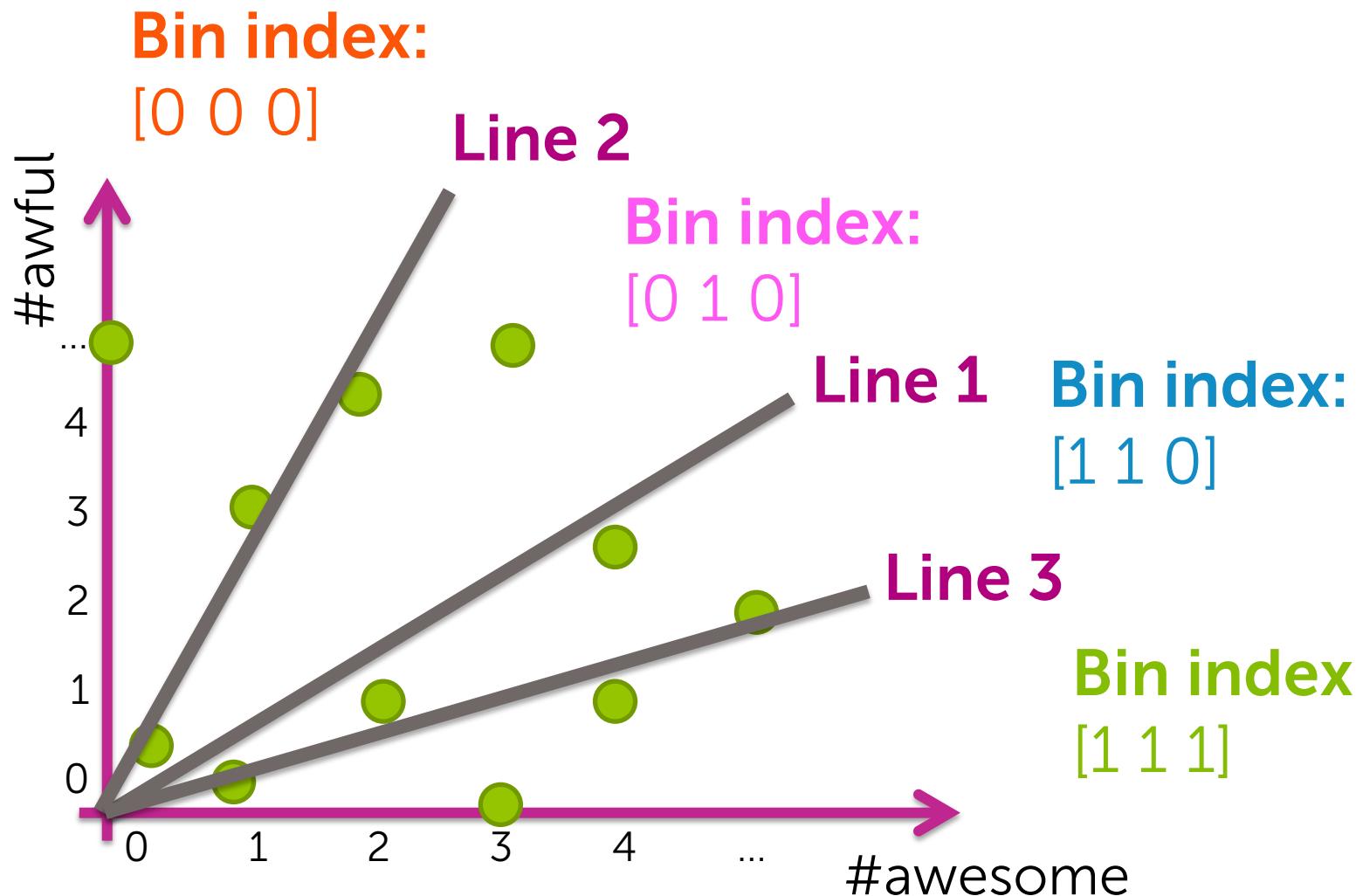
Saves lots of search time at little cost in quality of NN!

Limitations of KD-trees

- Difficult to implement
- Don't tend to perform well in high dimensions
 - Under some conditions, visit at least 2^d nodes



Locality sensitive hashing

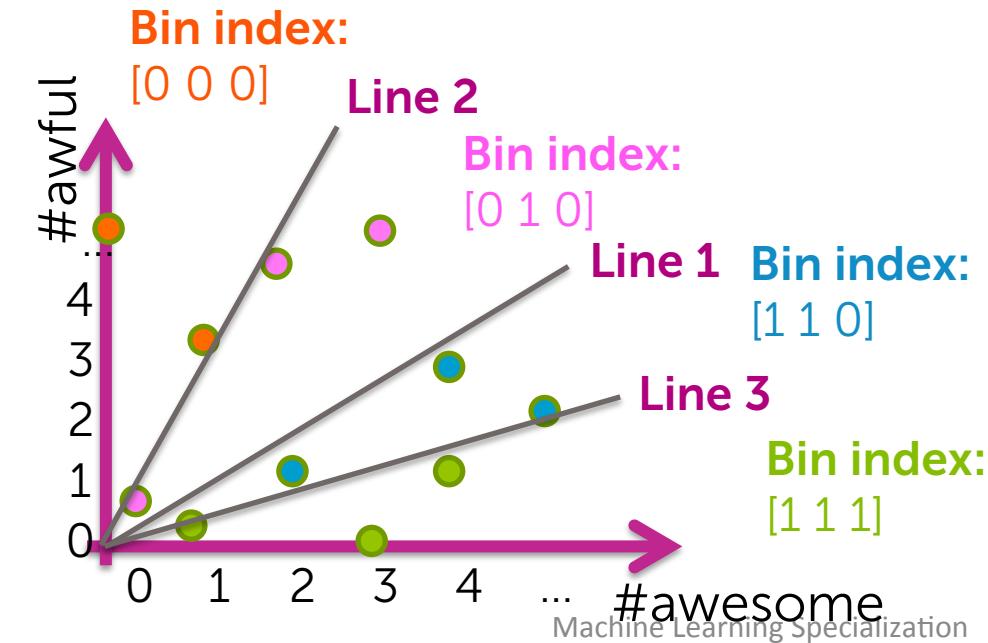


LSH for approximate NN search

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?

Next closest
bins (flip 1 bit)



Module 2: k-means and MapReduce

Module 2: k-means and MapReduce

Discover *clusters* of related documents



Cluster 1



Cluster 2



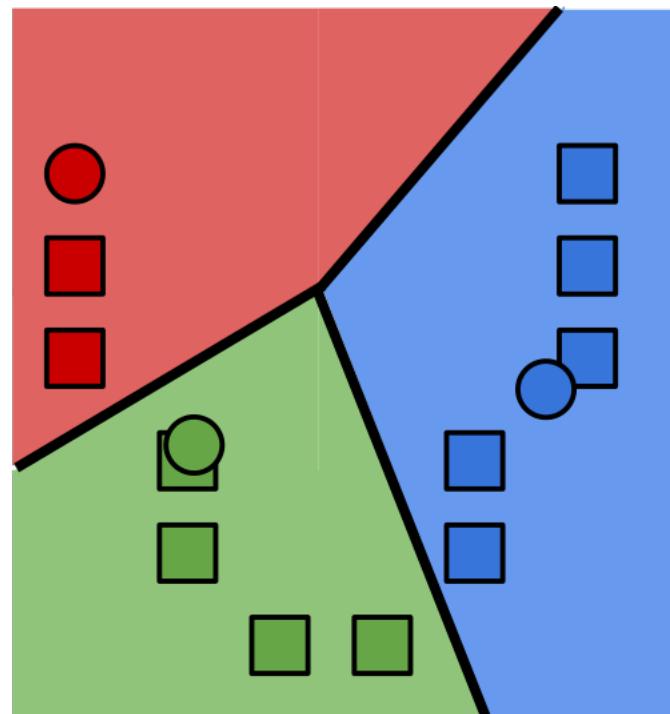
Cluster 3



Cluster 4

k-means 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



A coordinate descent algorithm

1. Assign observations to closest cluster center

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

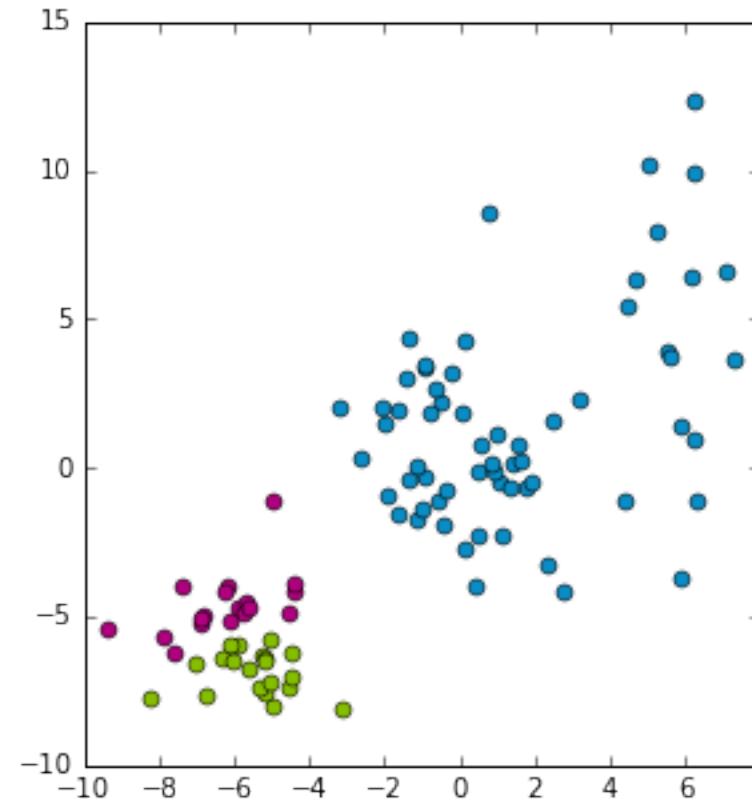
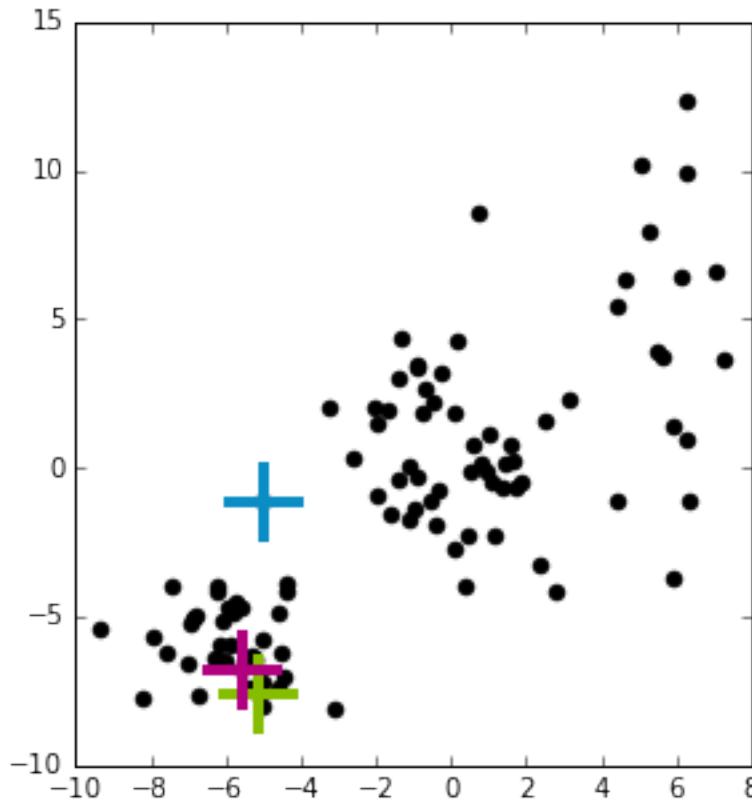
2. Revise cluster centers as mean of assigned observations

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

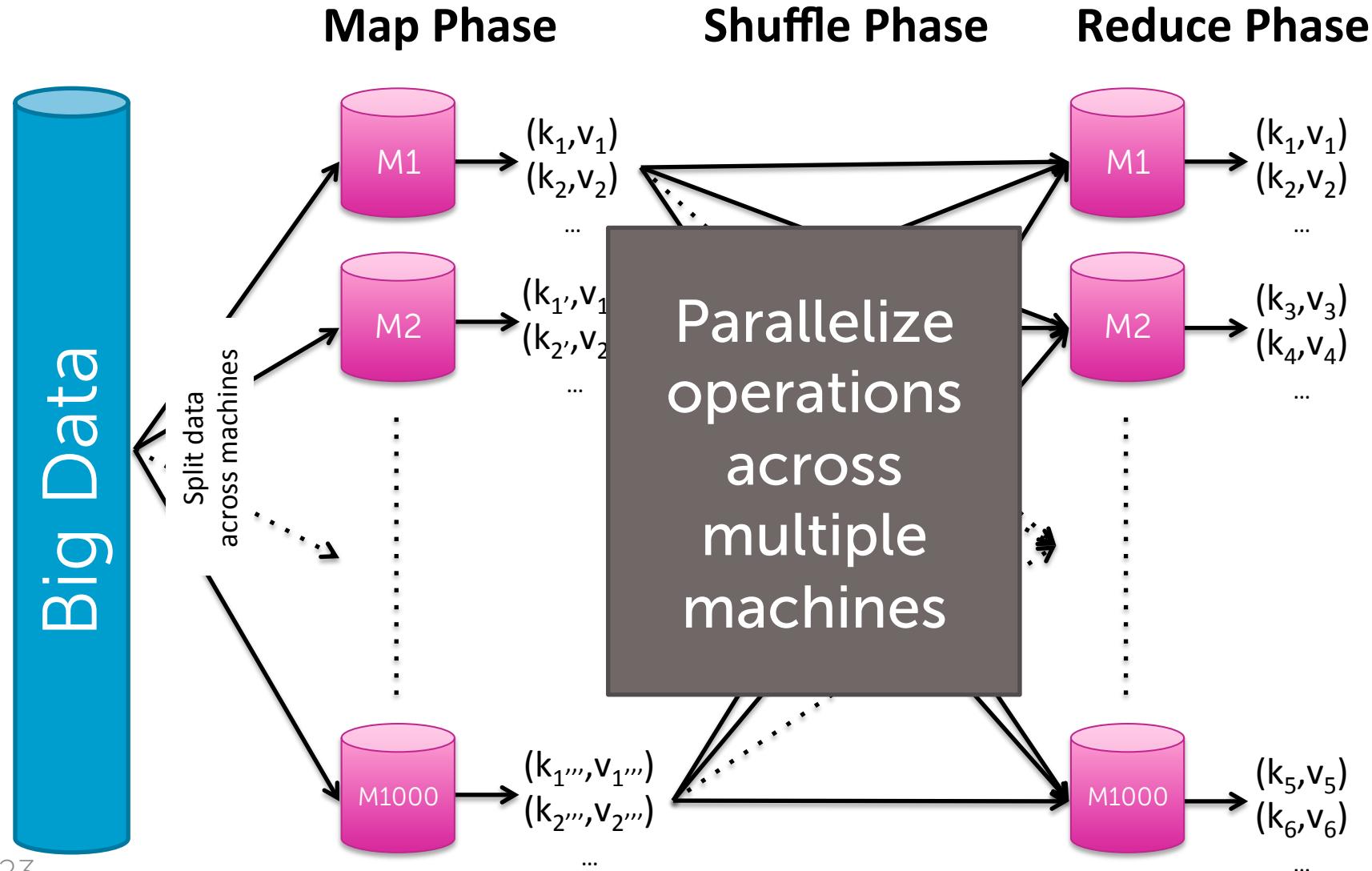
Alternating minimization

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

Convergence of k-means to local mode



MapReduce framework



MapReduce abstraction

Map:

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

Word count example:

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

Reduce:

- Aggregate values for each key
- Must be commutative-associative operation
- Data-parallel over keys
- Generate (key,value) pairs

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

MapReduce has long history in functional programming

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

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 ($z_i=k$)

Module 3: Mixture models

Mixture models

Probabilistic clustering model



Cluster 1



captures
uncertainty
in clustering

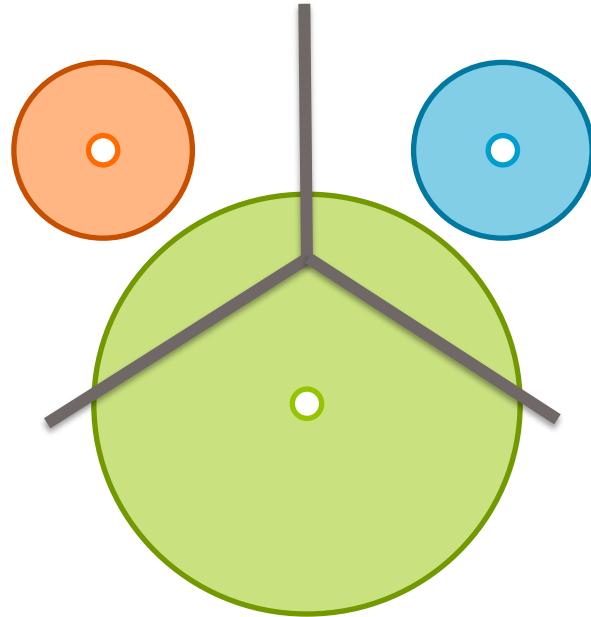


Cluster 3

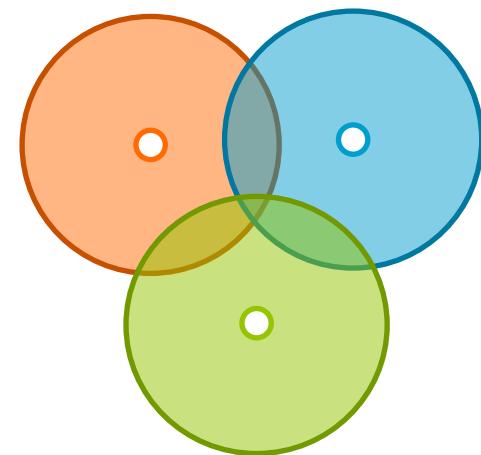


Cluster 4

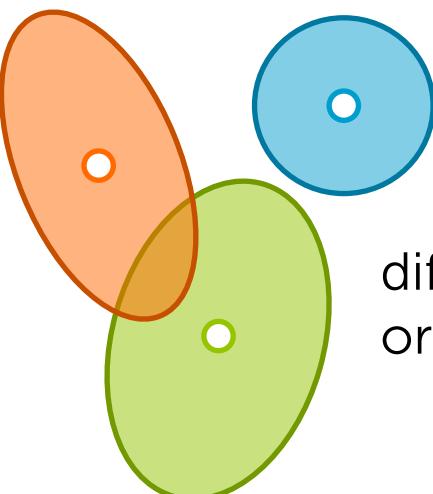
Failure modes of k-means



disparate cluster sizes

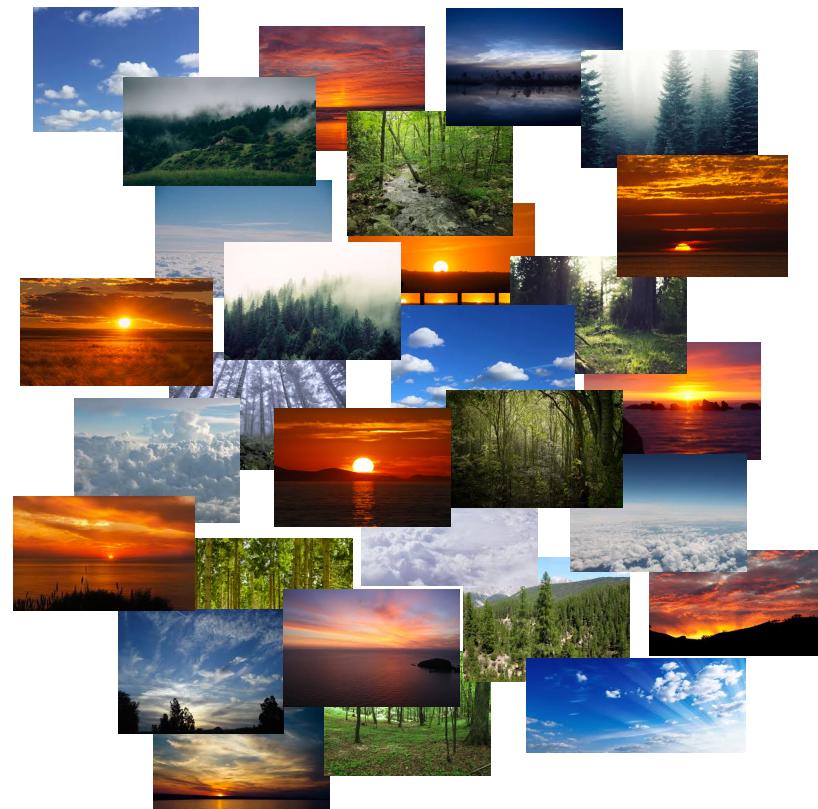
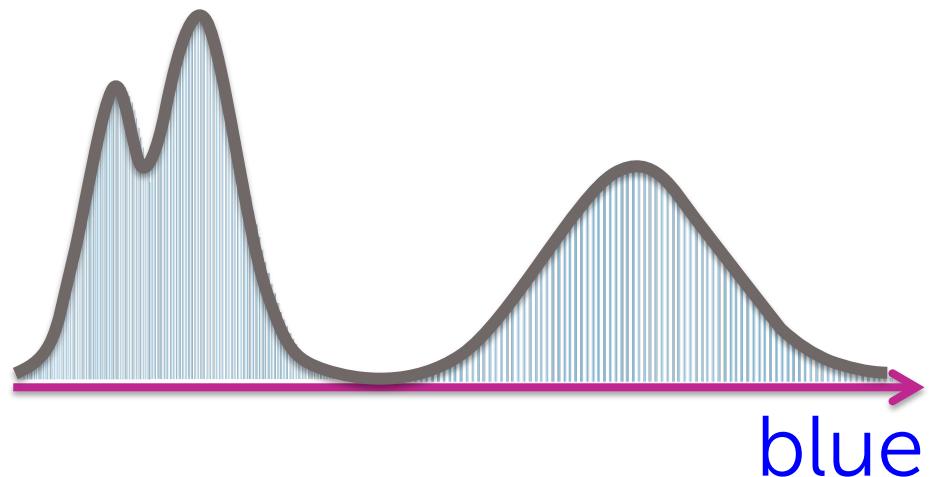


overlapping clusters

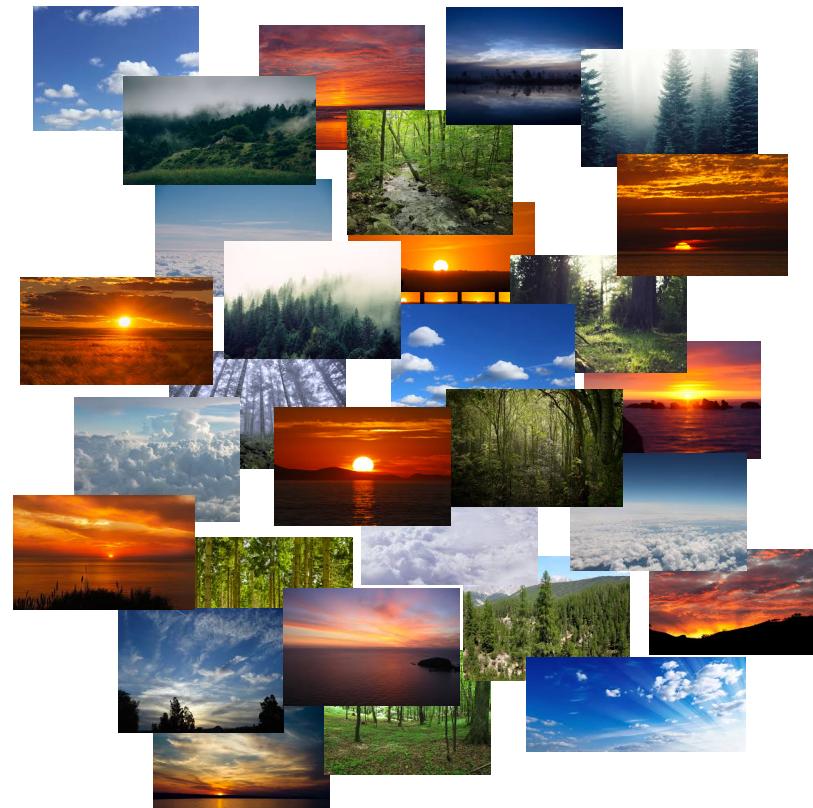
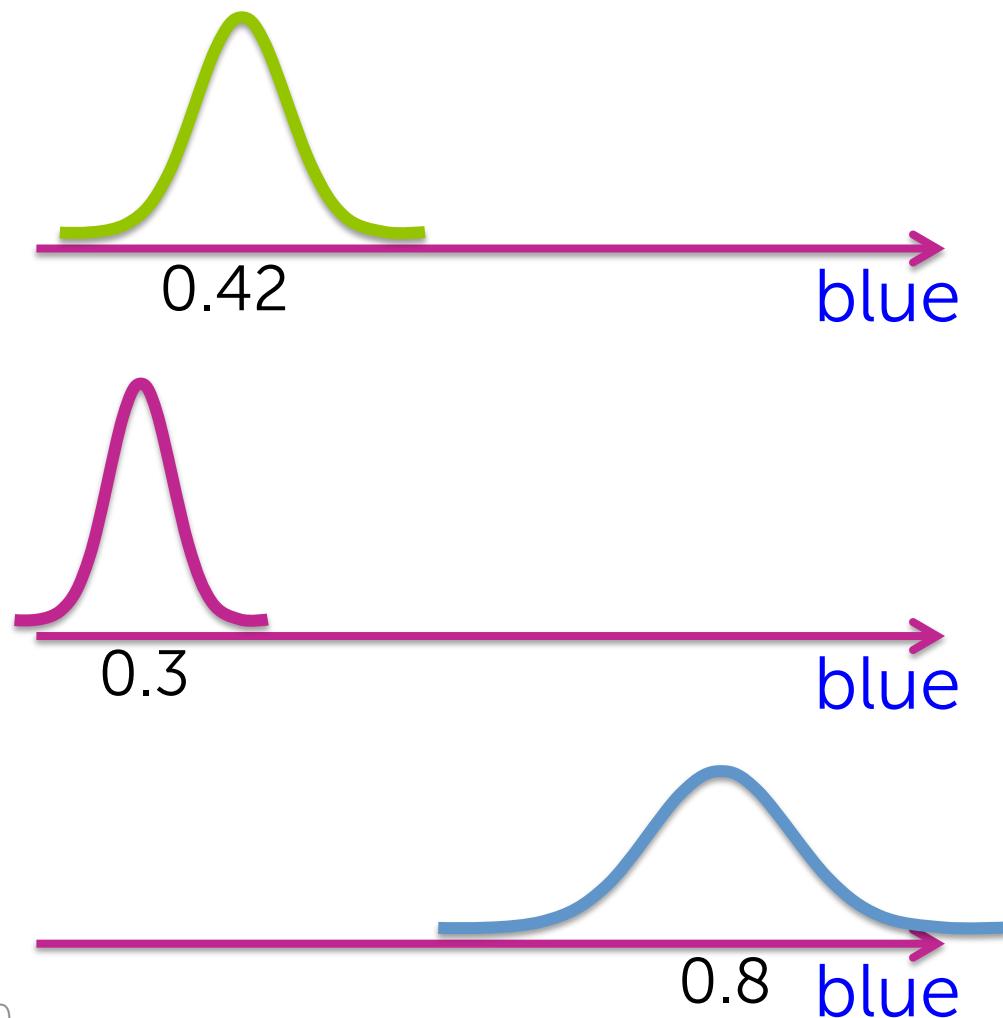


different shaped/
oriented clusters

Jumble of unlabeled images

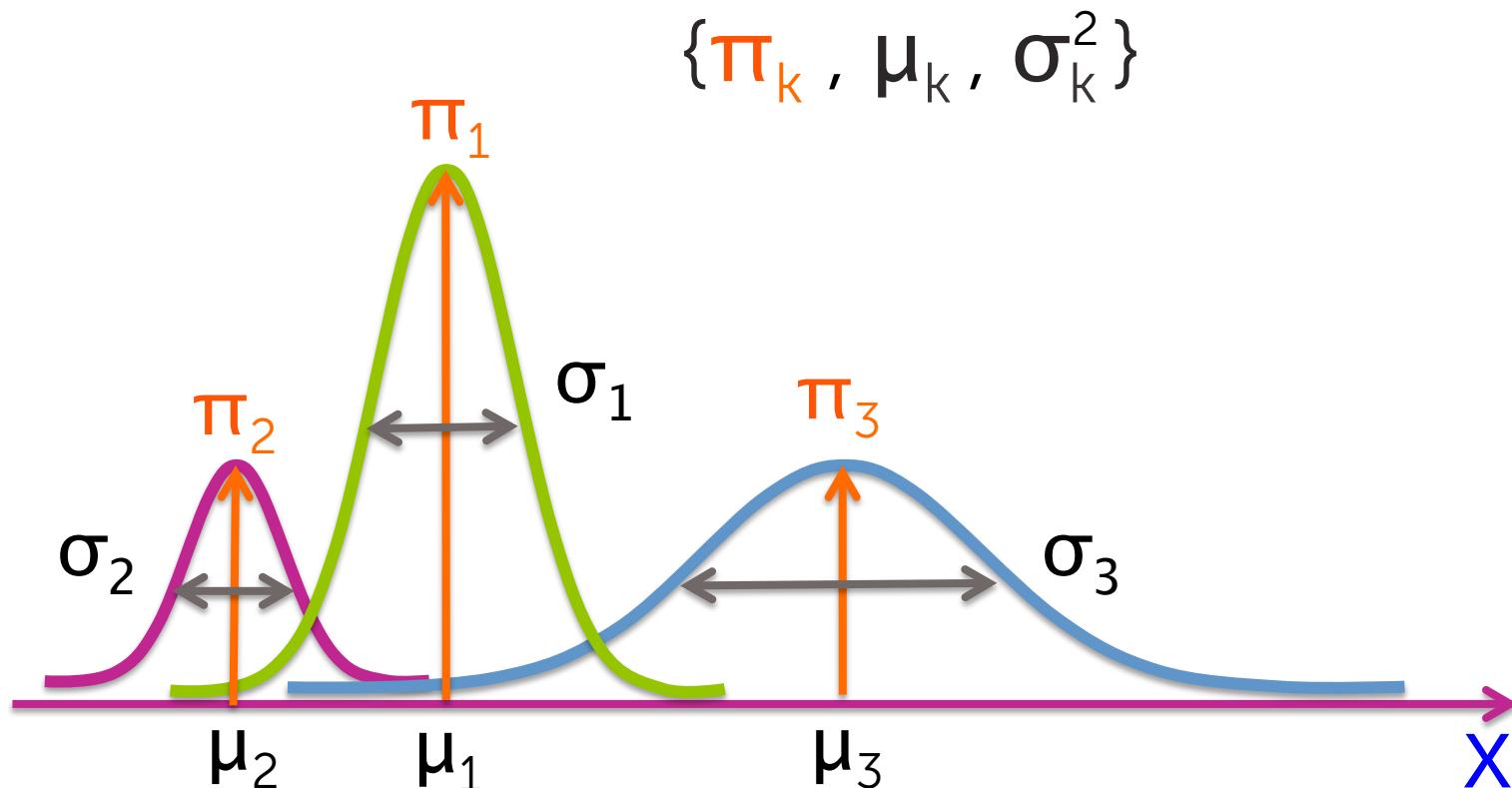


Model of jumble of unlabeled images



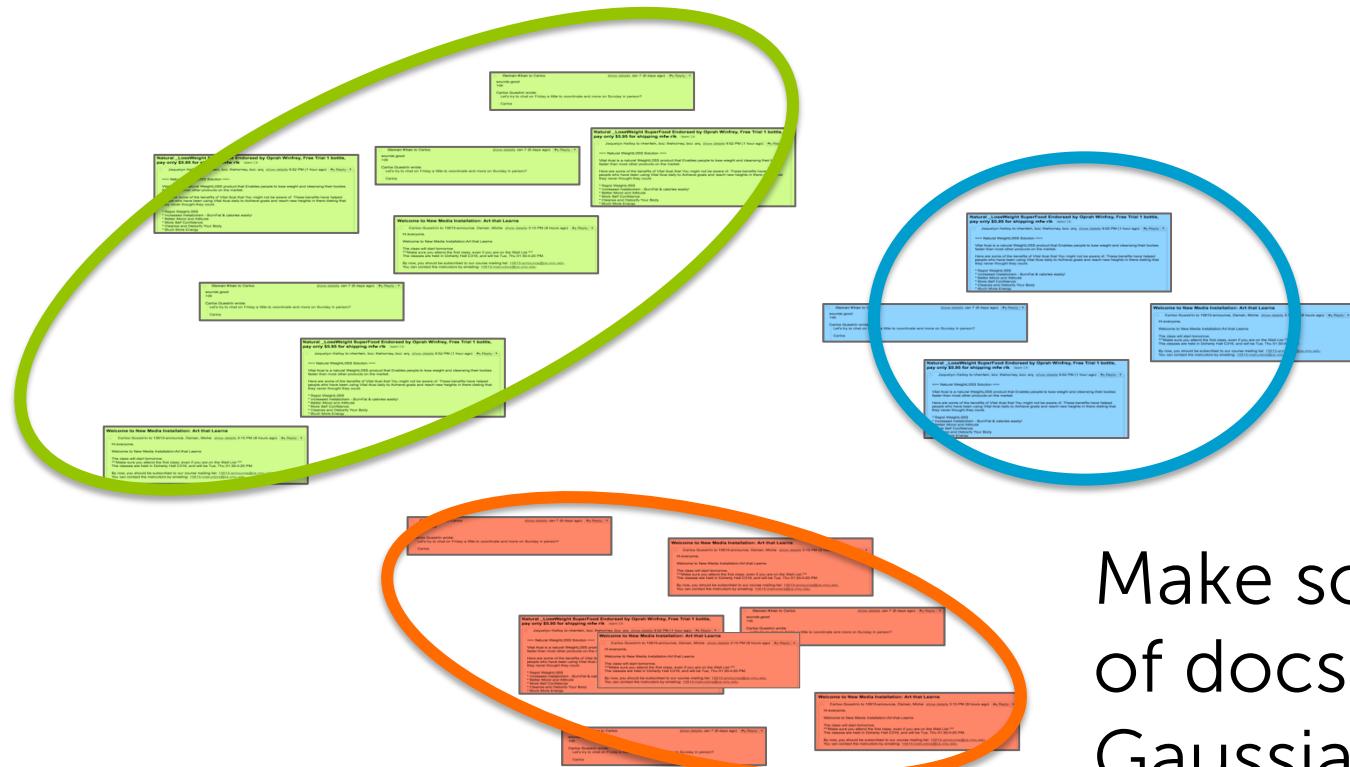
Mixture of Gaussians (1D)

Each mixture component represents a unique cluster specified by:



Mixture of Gaussians for clustering documents

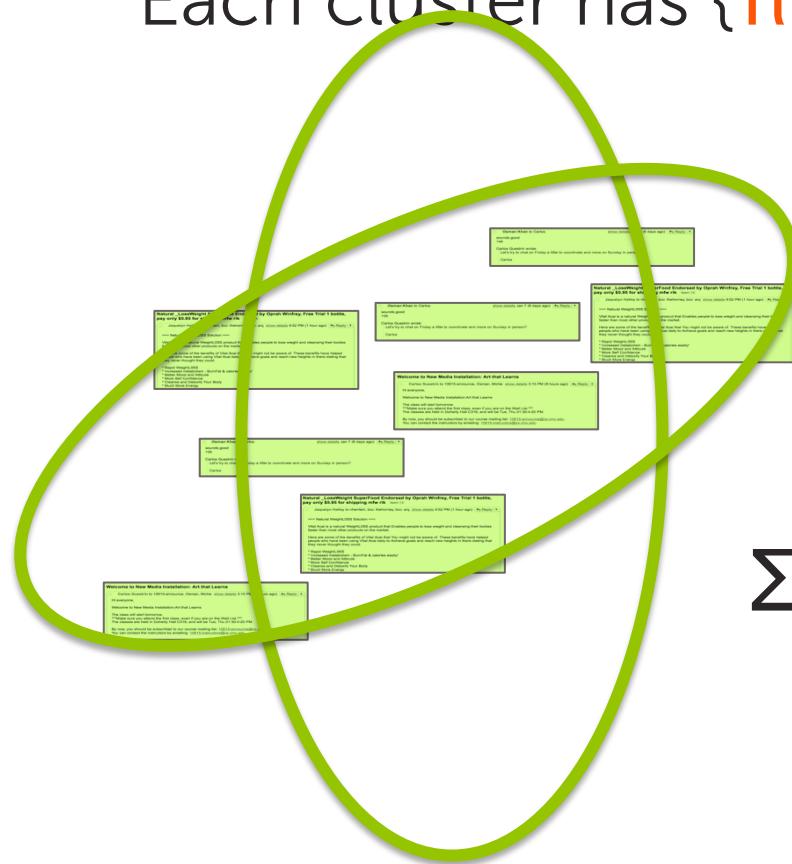
Space of all documents
(really lives in \mathbb{R}^V for vocab size V)



Make soft assignments
of docs to each
Gaussian

Restricting to diagonal covariance

Each cluster has $\{\pi_k, \mu_k, \Sigma_k \text{ diagonal}\}$

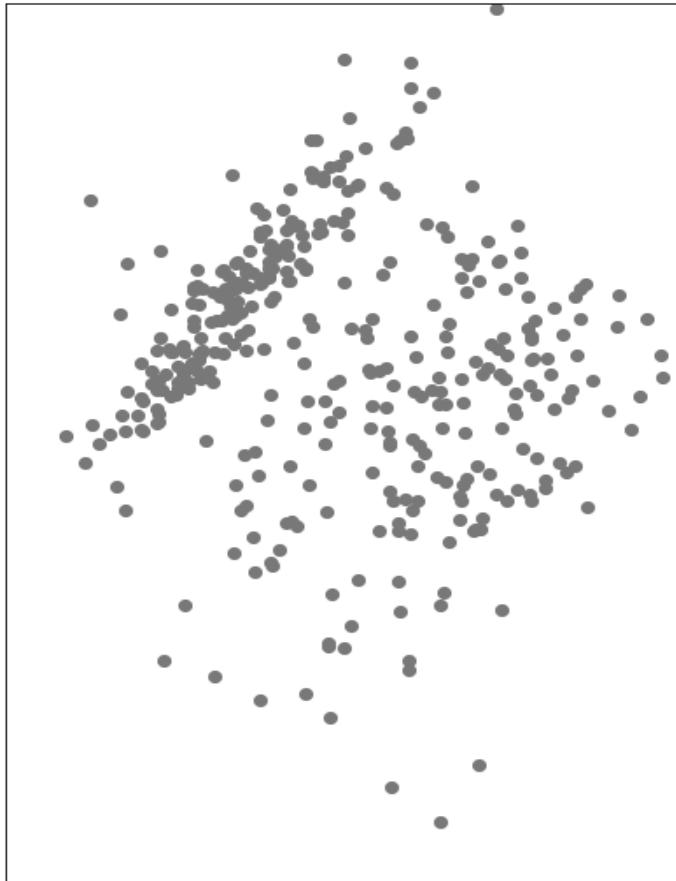


V params

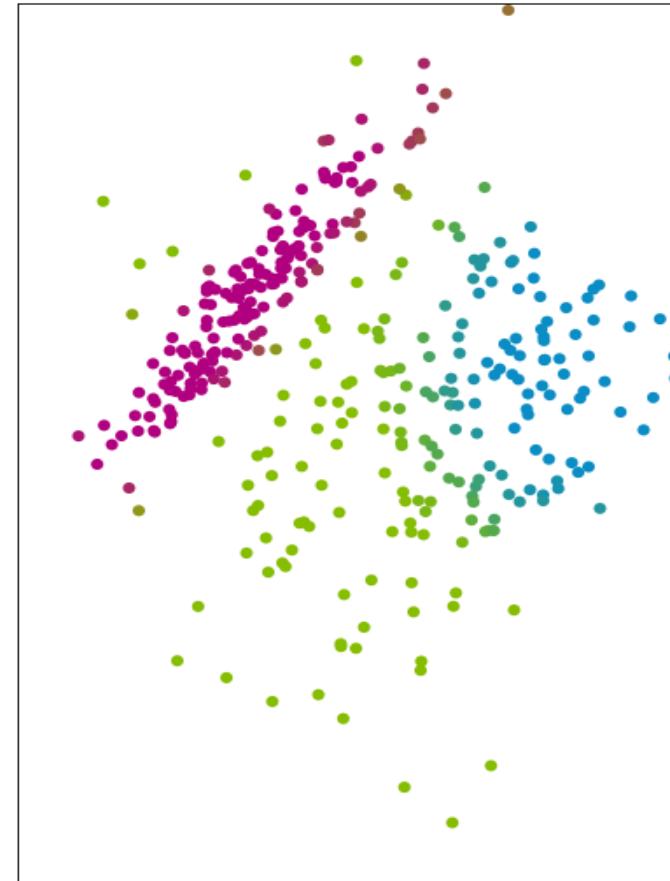
$$\Sigma = \begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \sigma_3^2 & \\ & & & 0 \\ 0 & & & \\ & & \ddots & \\ & & & \sigma_V^2 \end{pmatrix}$$

Inferring cluster labels

Data



EM algorithm →
soft assignments



Expectation maximization (EM): An iterative algorithm

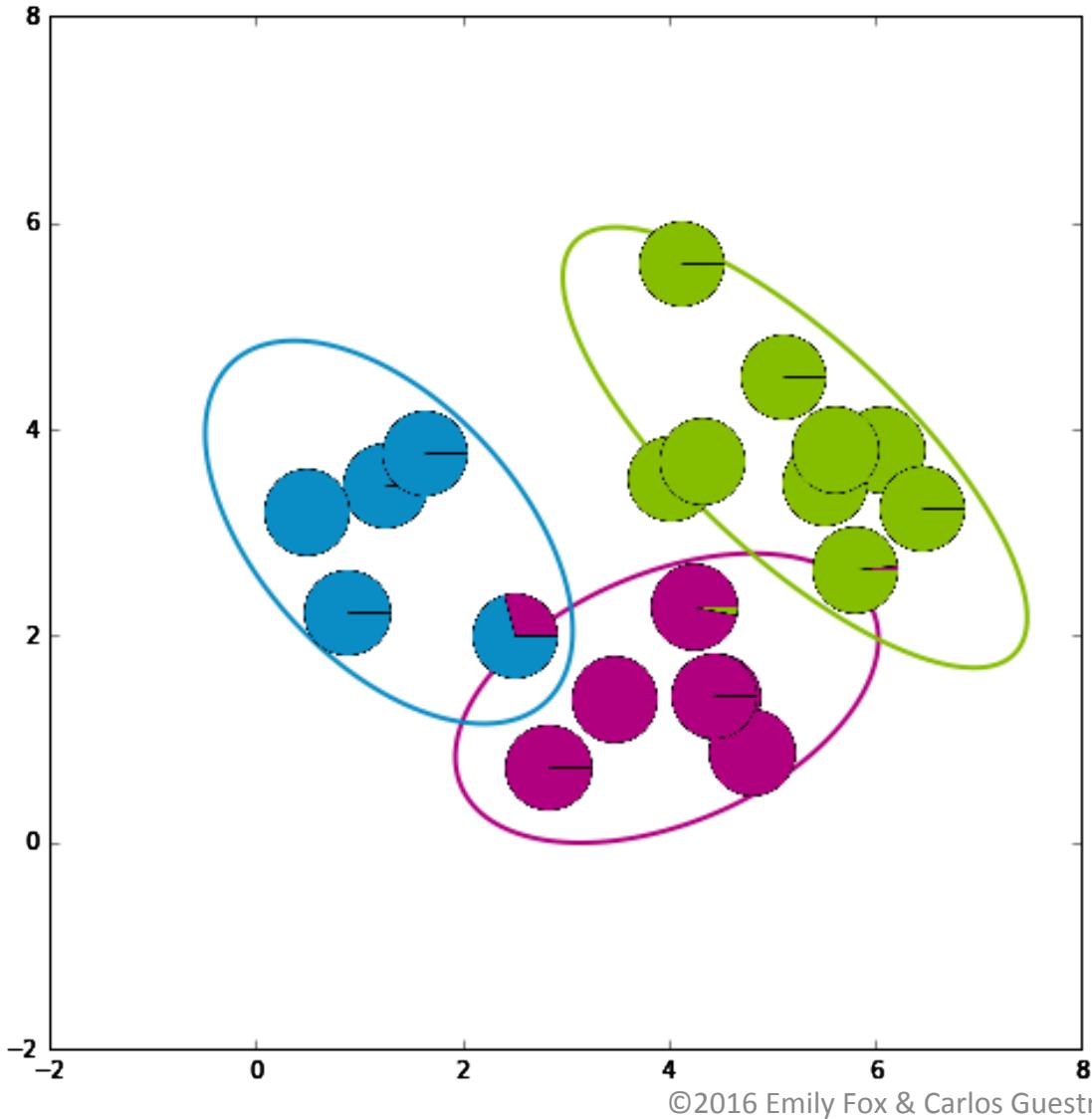
1. **E-step:** estimate cluster responsibilities given current parameter estimates

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

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

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

EM for mixtures of Gaussians in pictures - replay

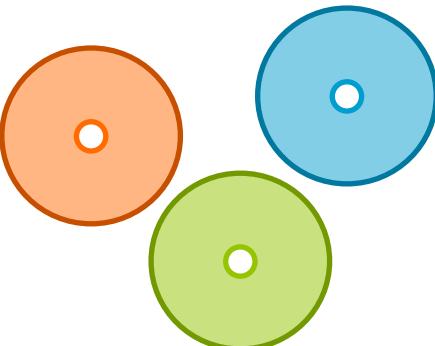


Relationship to k-means

Consider Gaussian mixture model with

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

Spherically
symmetric clusters



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

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

Module 4: 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}

^aDepartment of Bioengineering, University of Pennsylvania, Philadelphia, PA

^bDepartment of Neurology, University of Pennsylvania, Philadelphia, PA

^cDepartment 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

$$\boldsymbol{\pi} = [\pi_1 \ \pi_2 \ \dots \ \pi_K]$$

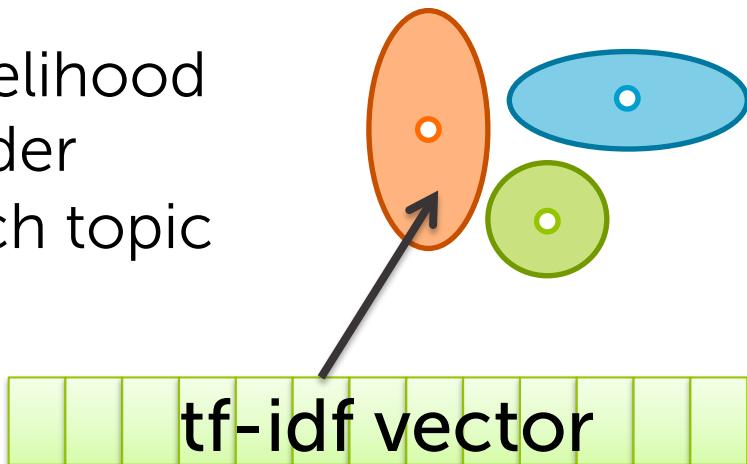
Comparing and contrasting

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**

Same topic 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}

^aDepartment of Bioengineering, University of Pennsylvania, Philadelphia, PA

^bDepartment of Neurology, University of Pennsylvania, Philadelphia, PA

^cDepartment 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

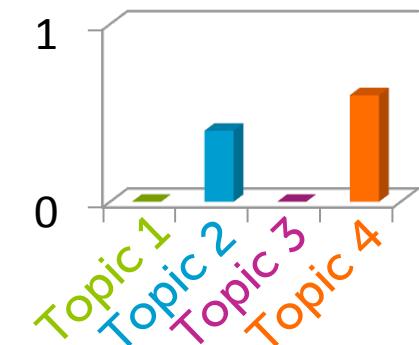
In LDA:

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

Each word scored
under topic z_{iw}

Distribution on
topics in document

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



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}

^aDepartment of Bioengineering, University of Pennsylvania, Philadelphia, PA

^bDepartment of Neurology, University of Pennsylvania, Philadelphia, PA

^cDepartment 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:

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



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}

^aDepartment of Bioengineering, University of Pennsylvania, Philadelphia, PA

^bDepartment of Neurology, University of Pennsylvania, Philadelphia, PA

^cDepartment of Statistics, University of Washington, Seattle, WA

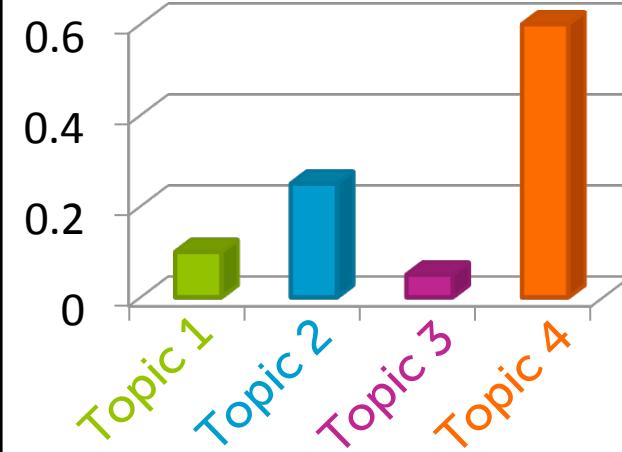
Abstract

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

- doc topic proportions
- topic vocab distributions

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

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}

^aDepartment of Bioengineering, University of Pennsylvania, Philadelphia, PA

^bDepartment of Neurology, University of Pennsylvania, Philadelphia, PA

^cDepartment 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

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}

^aDepartment of Bioengineering, University of Pennsylvania, Philadelphia, PA
^bDepartment of Neurology, University of Pennsylvania, Philadelphia, PA
^cDepartment of Statistics, University of Washington, Seattle, WA

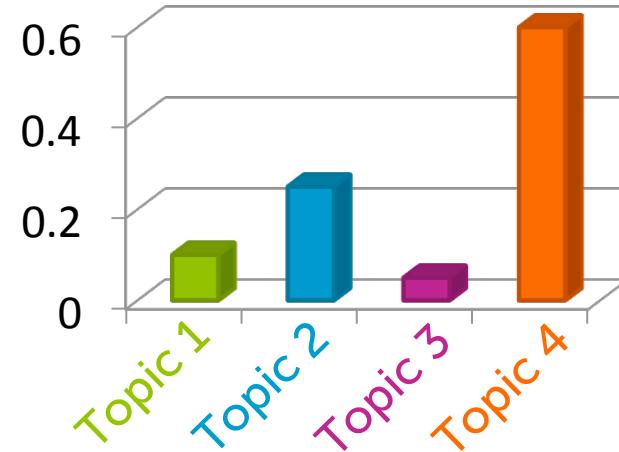
Abstract

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

Collapsed 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}

^aDepartment of Bioengineering, University of Pennsylvania, Philadelphia, PA
^bDepartment of Neurology, University of Pennsylvania, Philadelphia, PA
^cDepartment of Statistics, University of Washington, Seattle, WA

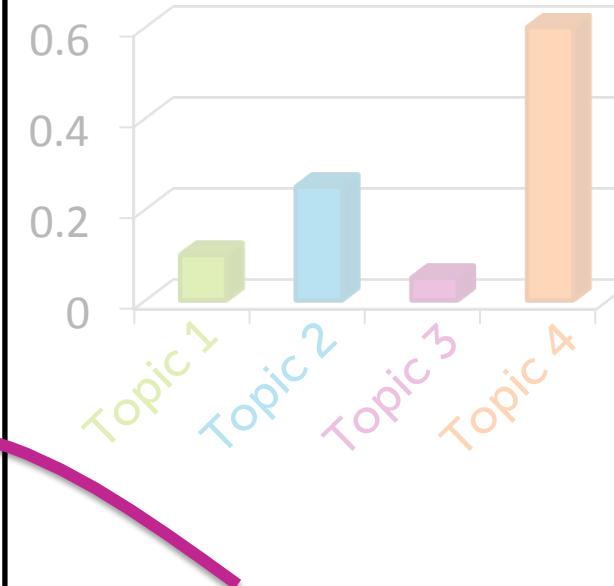
Abstract

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



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

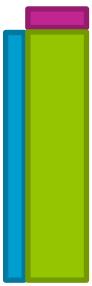
Collapsed conditional distribution

3	?	1	3	1
epilepsy	dynamic	Bayesian	EEG	model

Topic 1



Topic 2



Topic 3



Probability of assignment of word
in doc i to topic k proportional to:

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

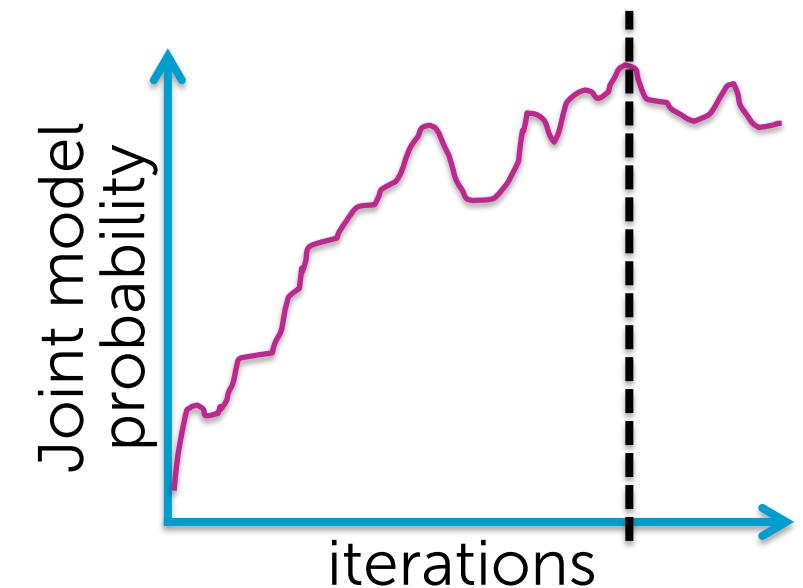
What to do with sampling output?

Predictions:

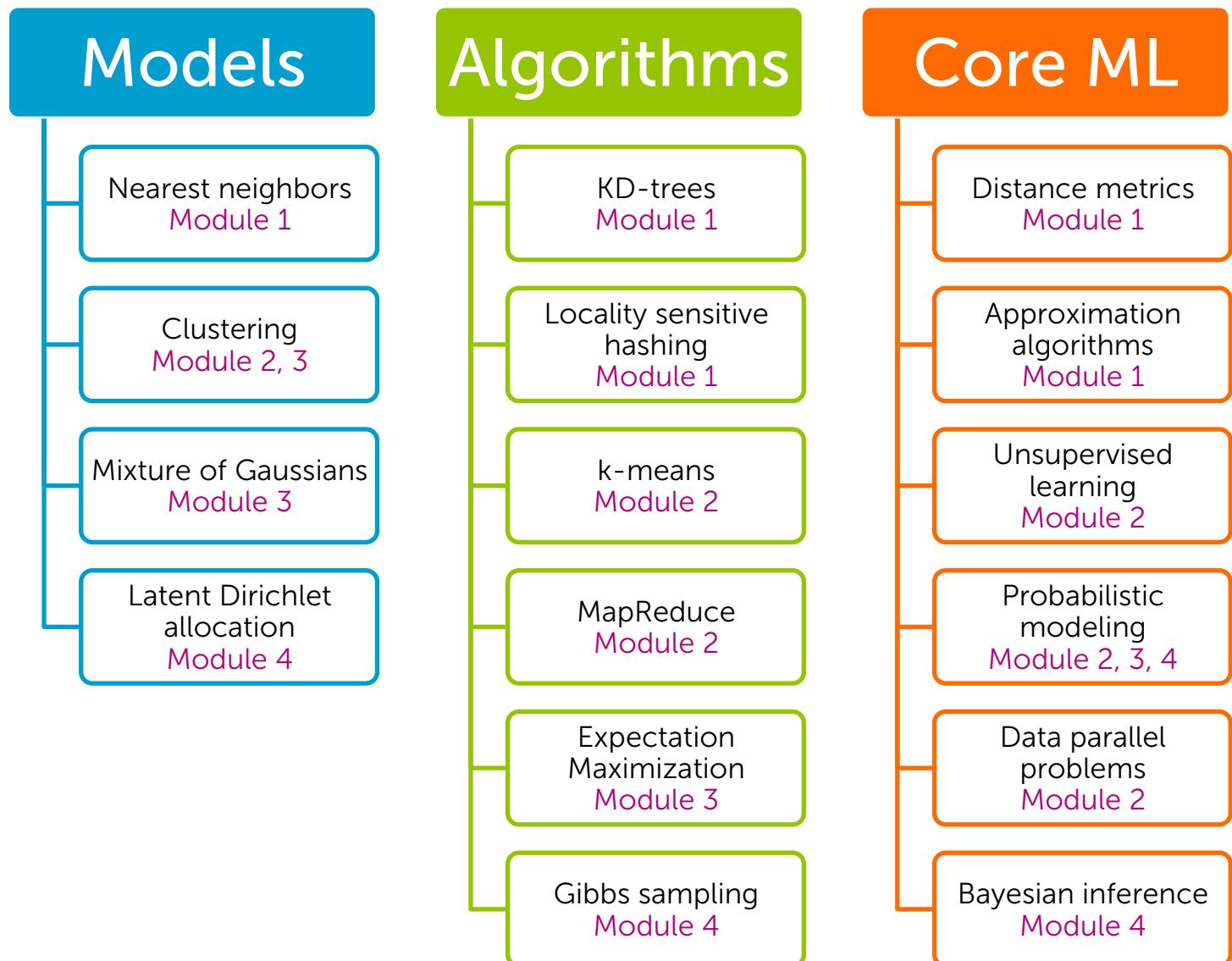
1. Make prediction for each snapshot of randomly assigned variables/parameters (full iteration)
2. Average predictions for final result

Parameter or assignment estimate:

- Look at snapshot of randomly assigned variables/parameters that maximizes “joint model probability”



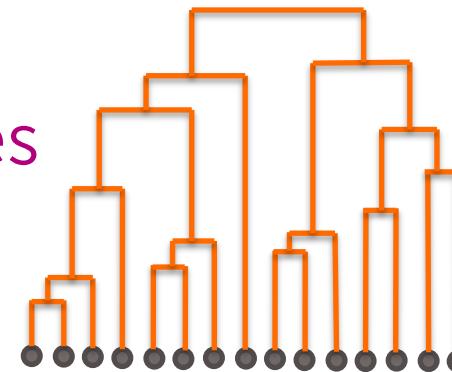
Summary of what we learned



Bonus content: Hierarchical clustering

Why hierarchical clustering?

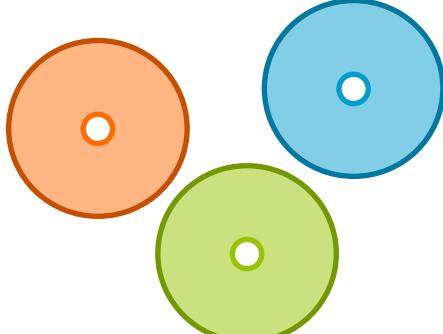
- Avoid choosing # clusters beforehand
- Dendograms help visualize different clustering granularities
 - No need to rerun algorithm
- Most algorithms allow user to choose any distance metric
 - k-means restricted us to Euclidean distance



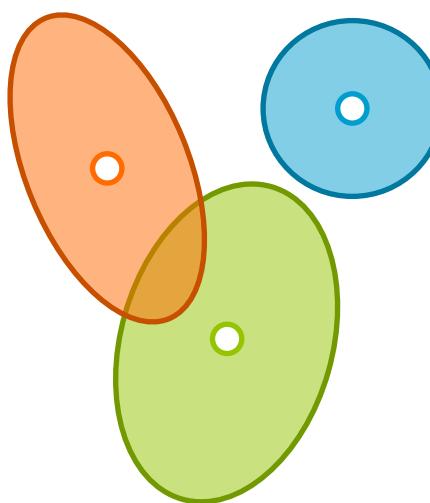
Why hierarchical clustering?

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

k-means: spherical clusters



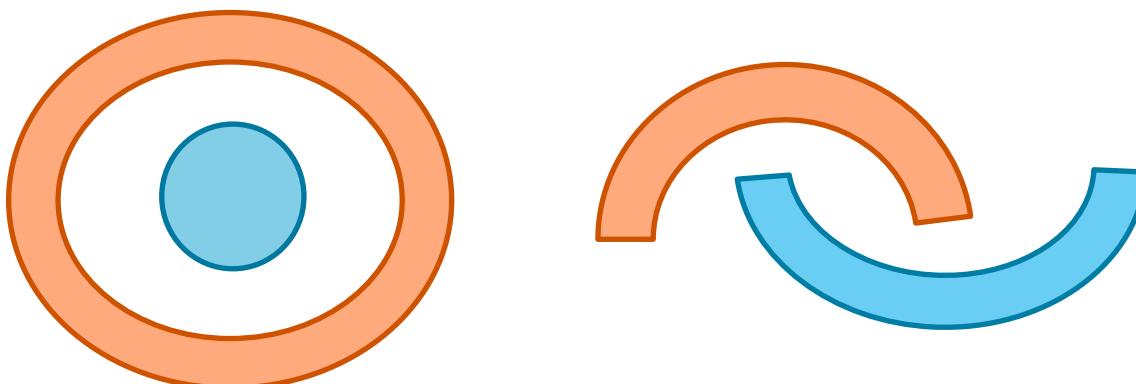
Gaussian mixtures:
ellipsoids



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.

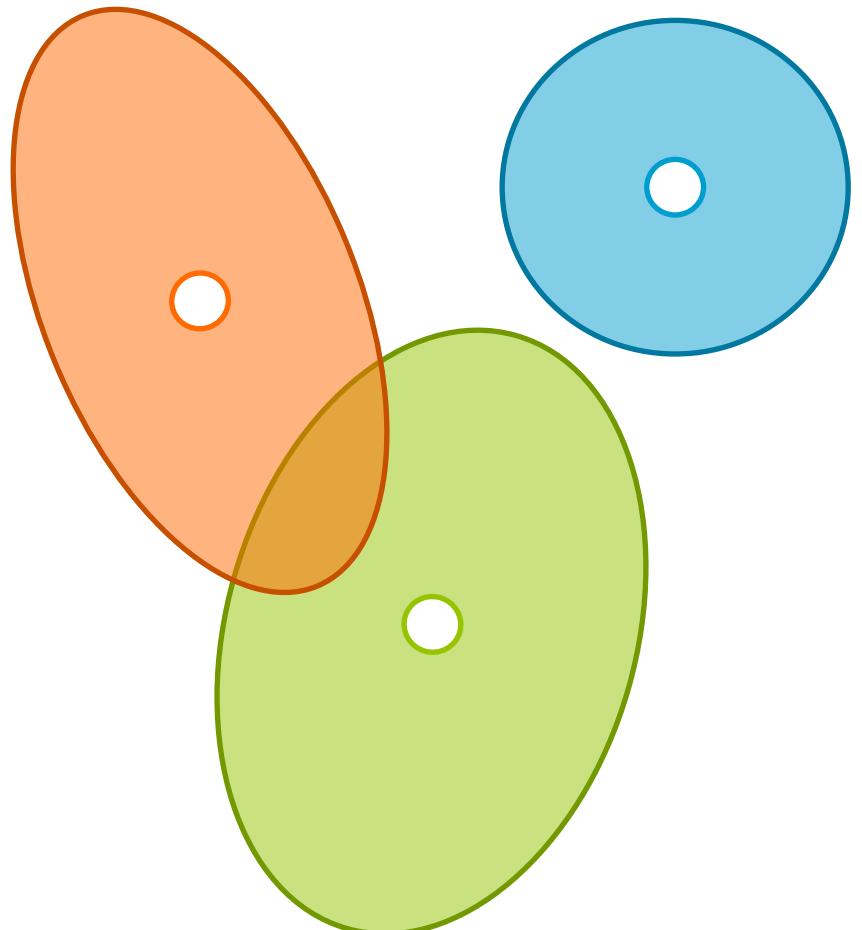
- Example: **recursive k-means**

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.

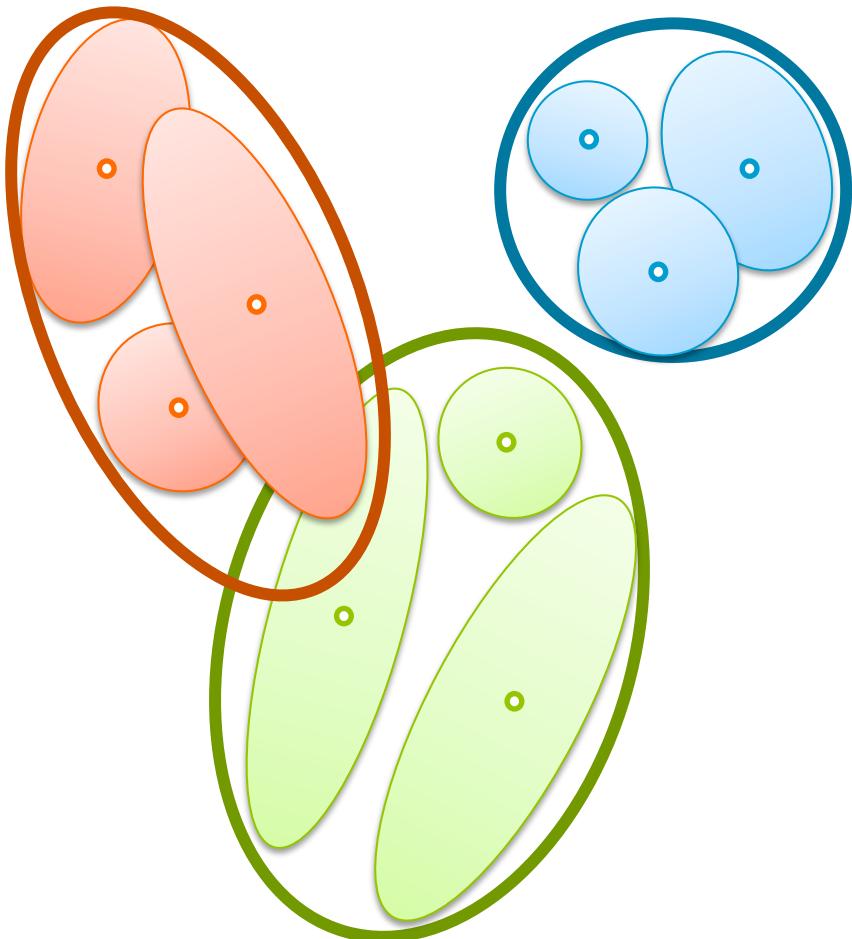
- Example: **single linkage**

Divisive clustering

Divisive in pictures – level 1

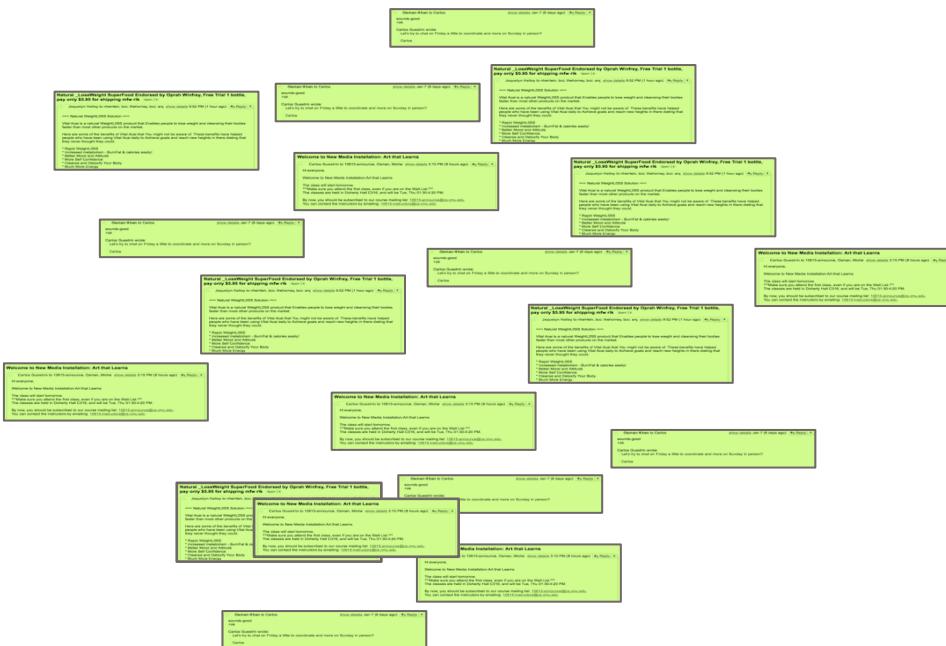


Divisive in pictures – level 2



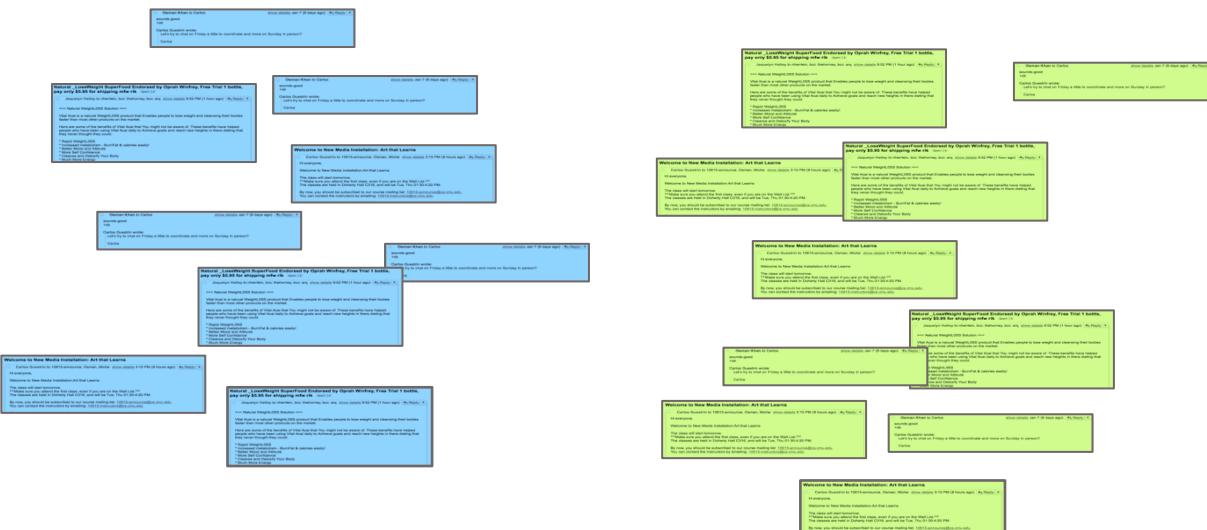
Divisive: Recursive k-means

Wikipedia

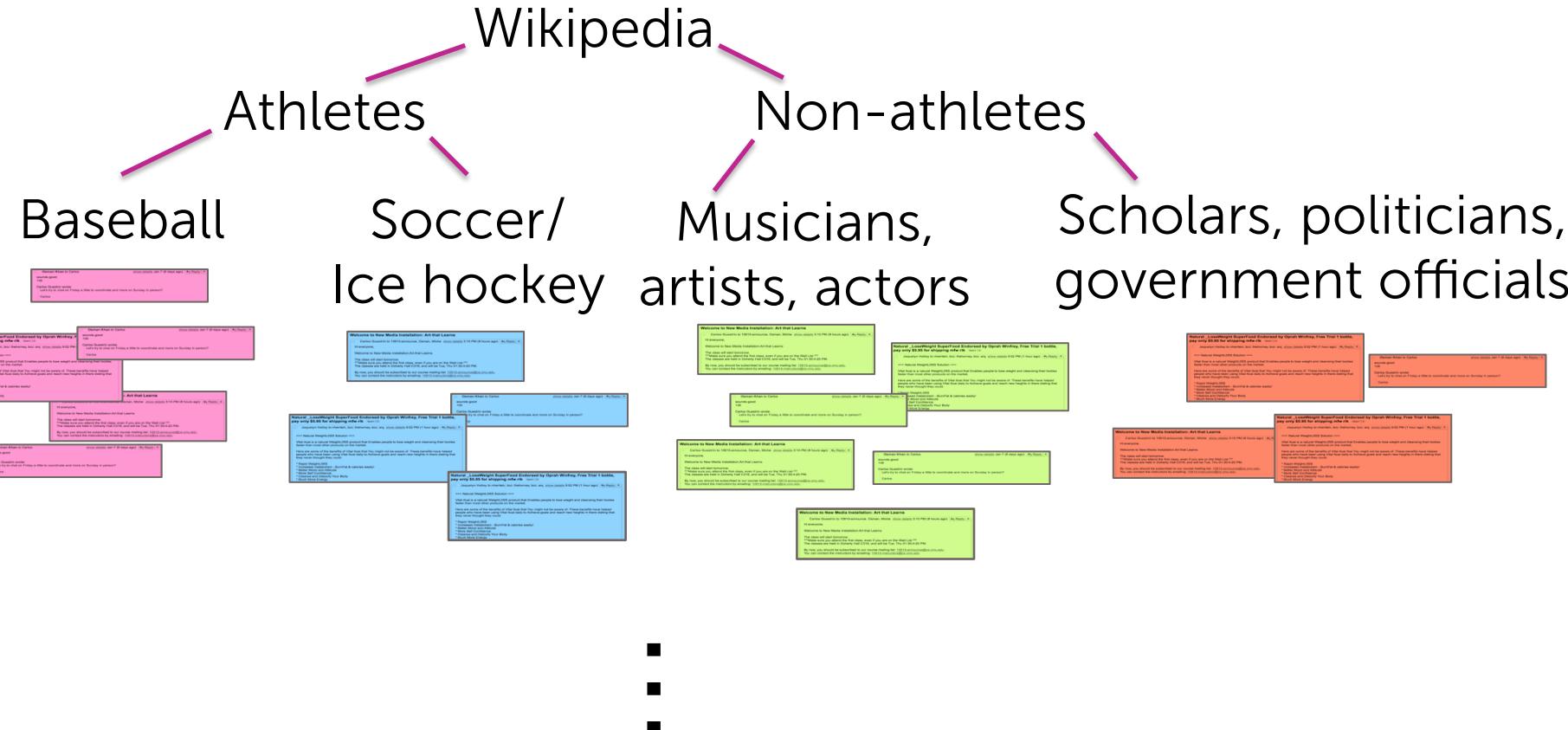


Divisive: Recursive k-means

Athletes Non-athletes



Divisive: Recursive k-means



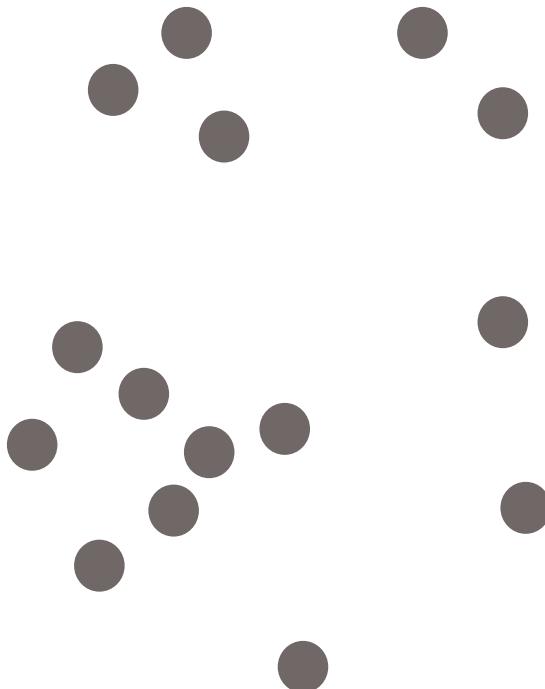
Divisive choices to be made

- 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

Agglomerative clustering

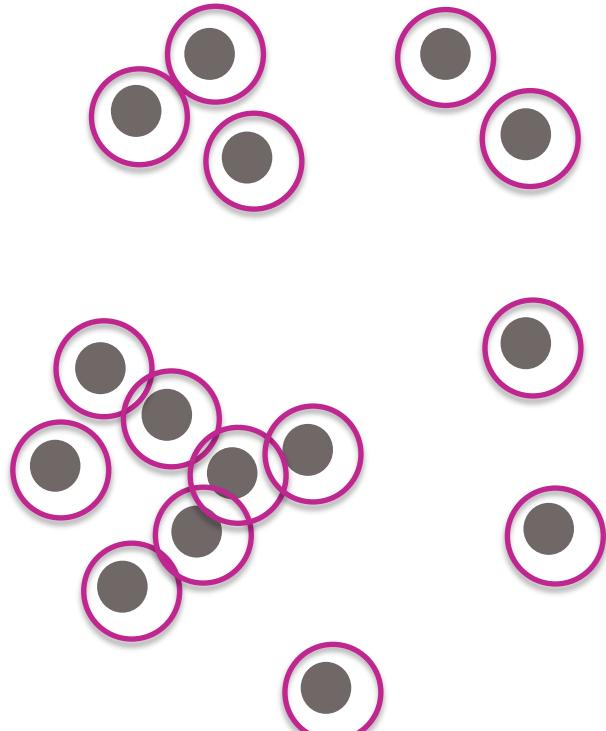
Agglomerative: Single linkage

1. Initialize each point to be its own cluster



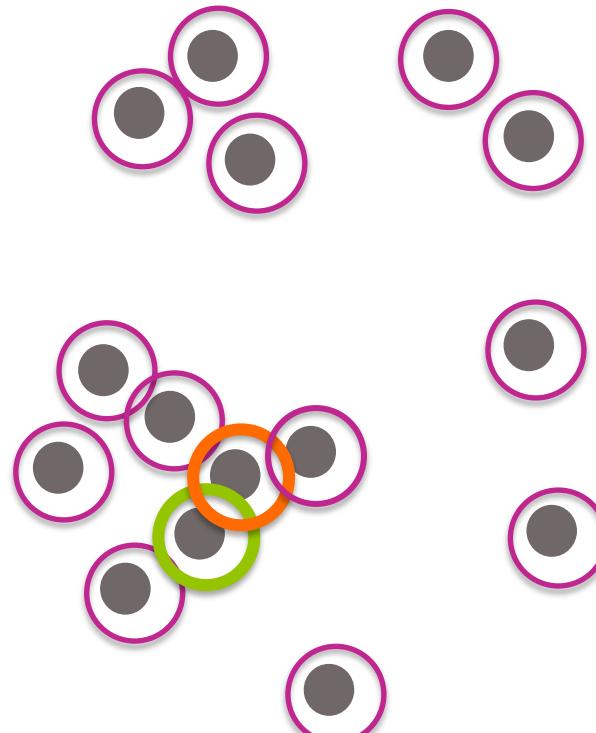
Agglomerative: Single linkage

1. Initialize each point to be its own cluster



Agglomerative: Single linkage

2. Define distance between clusters to be:



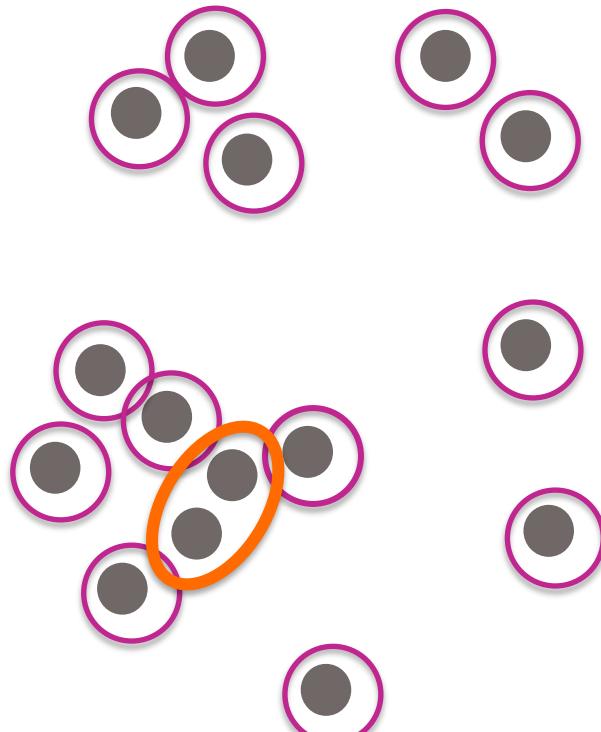
$\text{distance}(C_1, C_2) = \min_{\substack{x_i \text{ in } C_1, \\ x_j \text{ in } C_2}} d(x_i, x_j)$

specified pairwise distance function

Linkage criteria

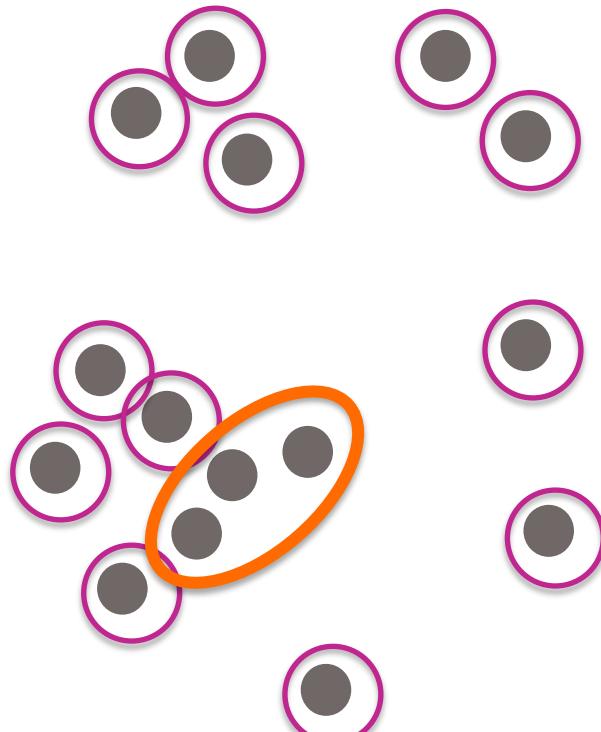
Agglomerative: Single linkage

3. Merge the two closest clusters



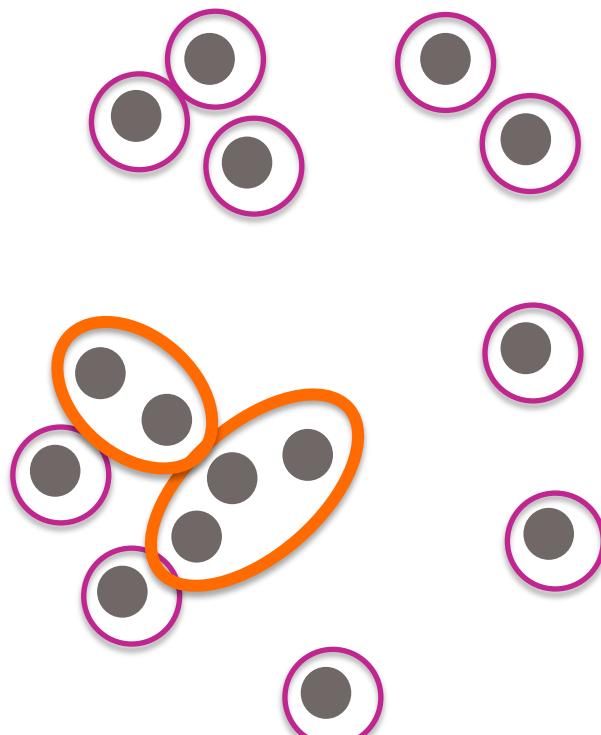
Agglomerative: Single linkage

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



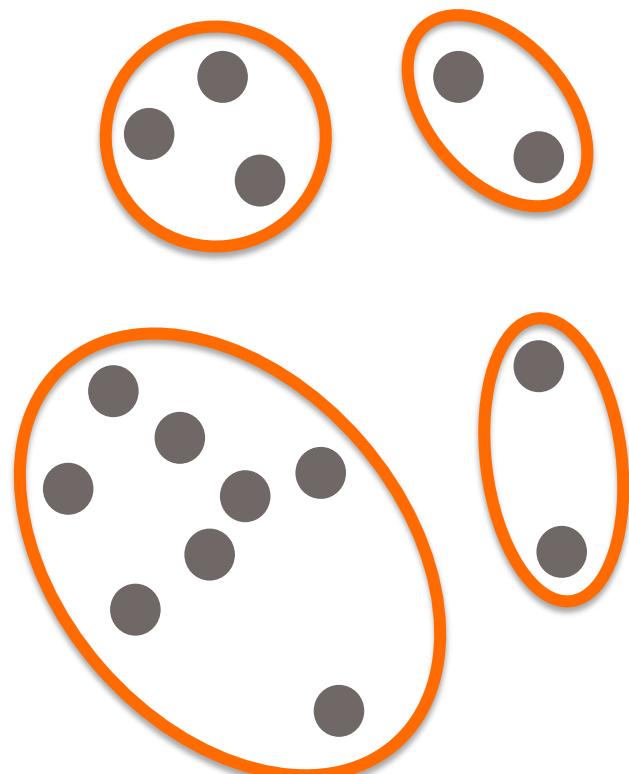
Agglomerative: Single linkage

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



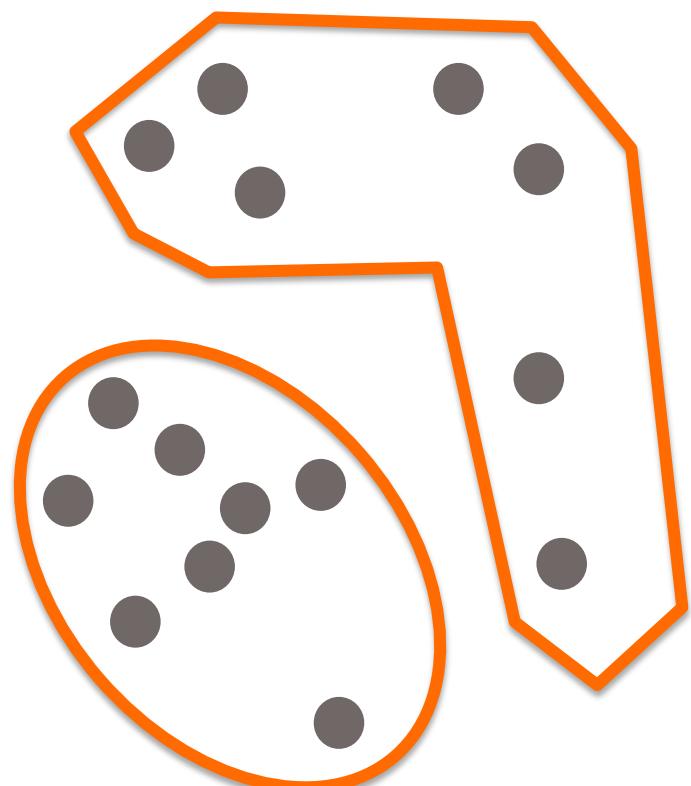
Agglomerative: Single linkage

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



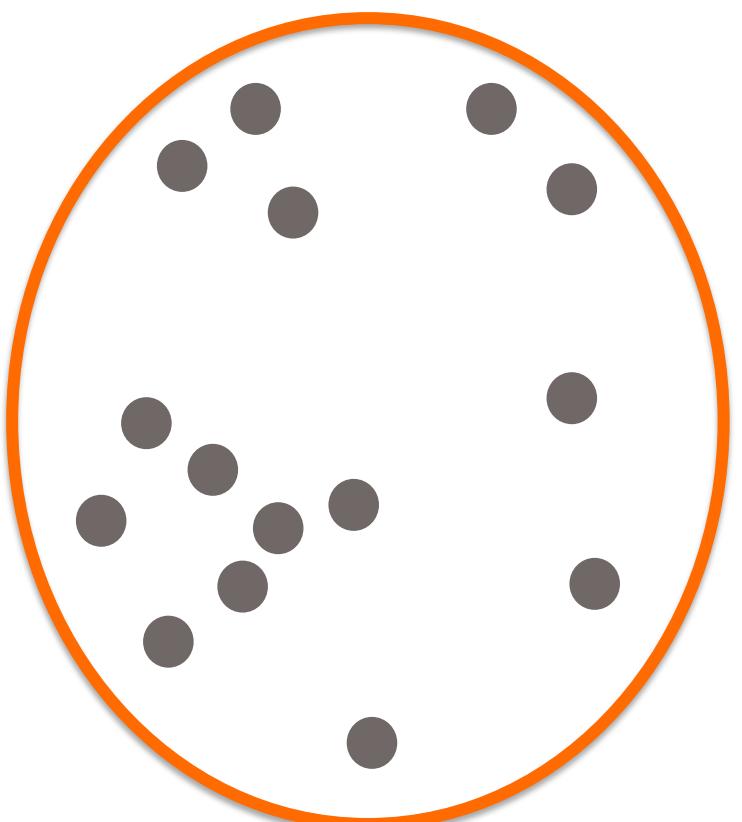
Agglomerative: Single linkage

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



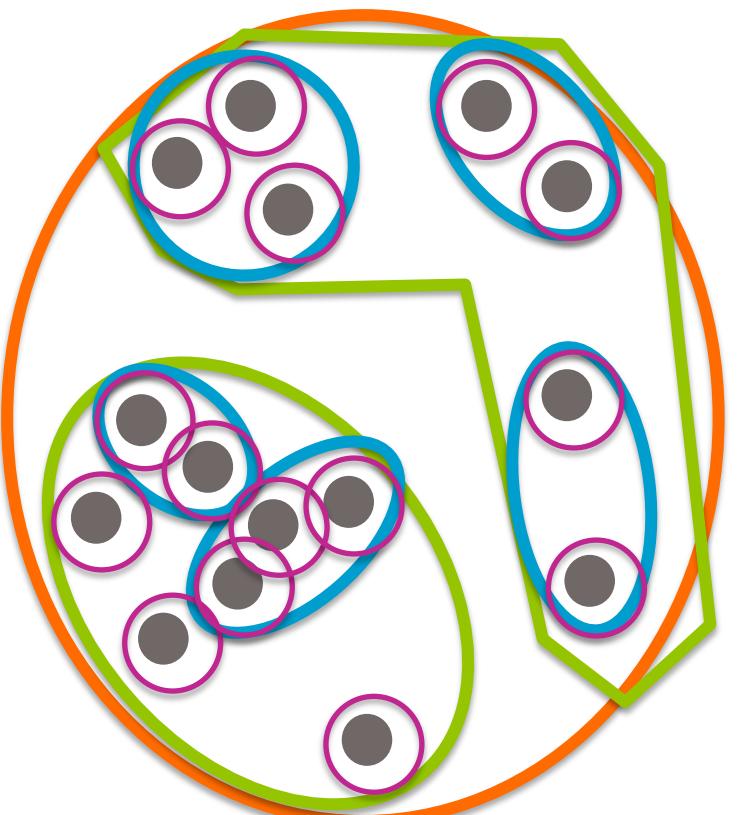
Agglomerative: Single linkage

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



Clusters of clusters

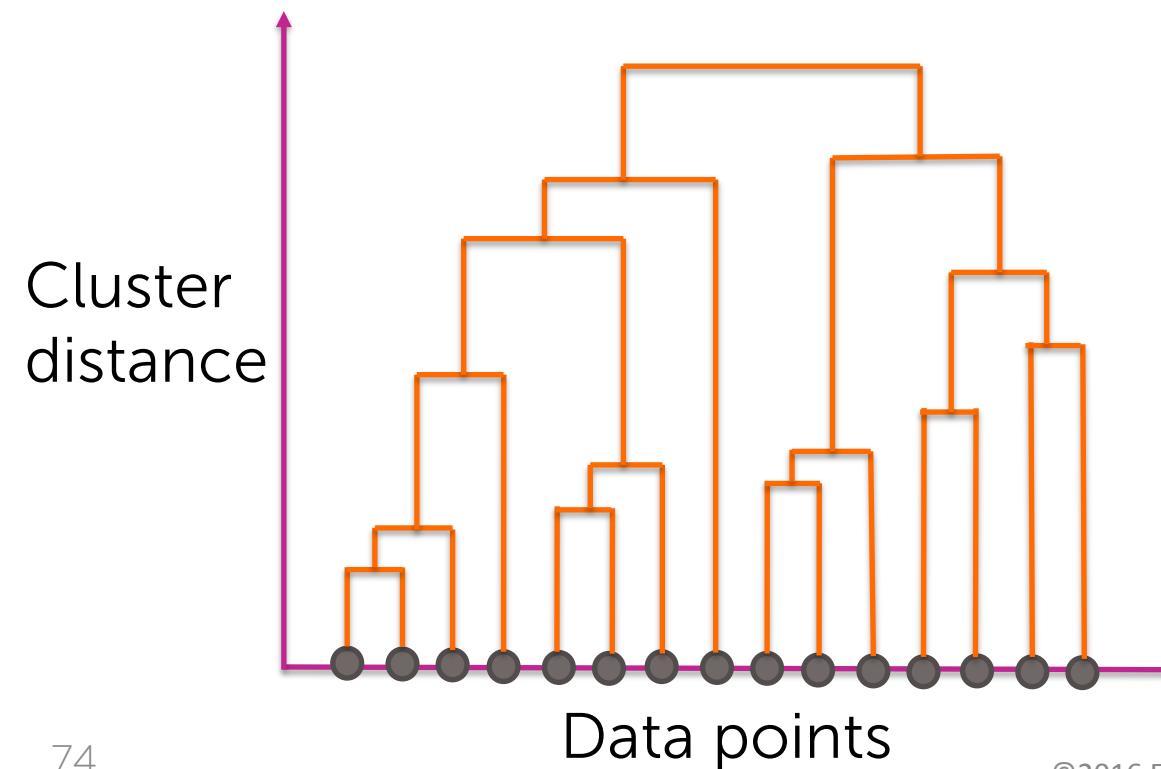
Just like our picture for divisive clustering...



The dendrogram for agglomerative clustering

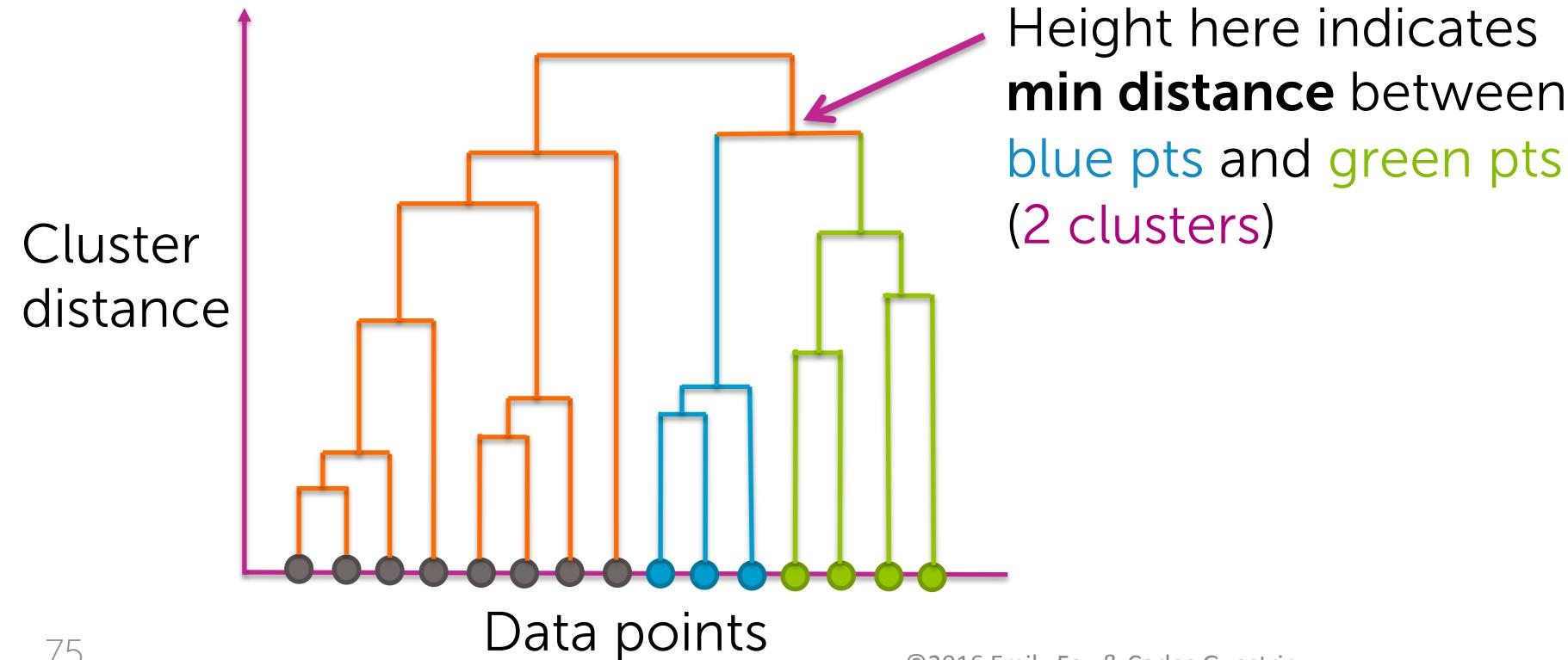
The dendrogram

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



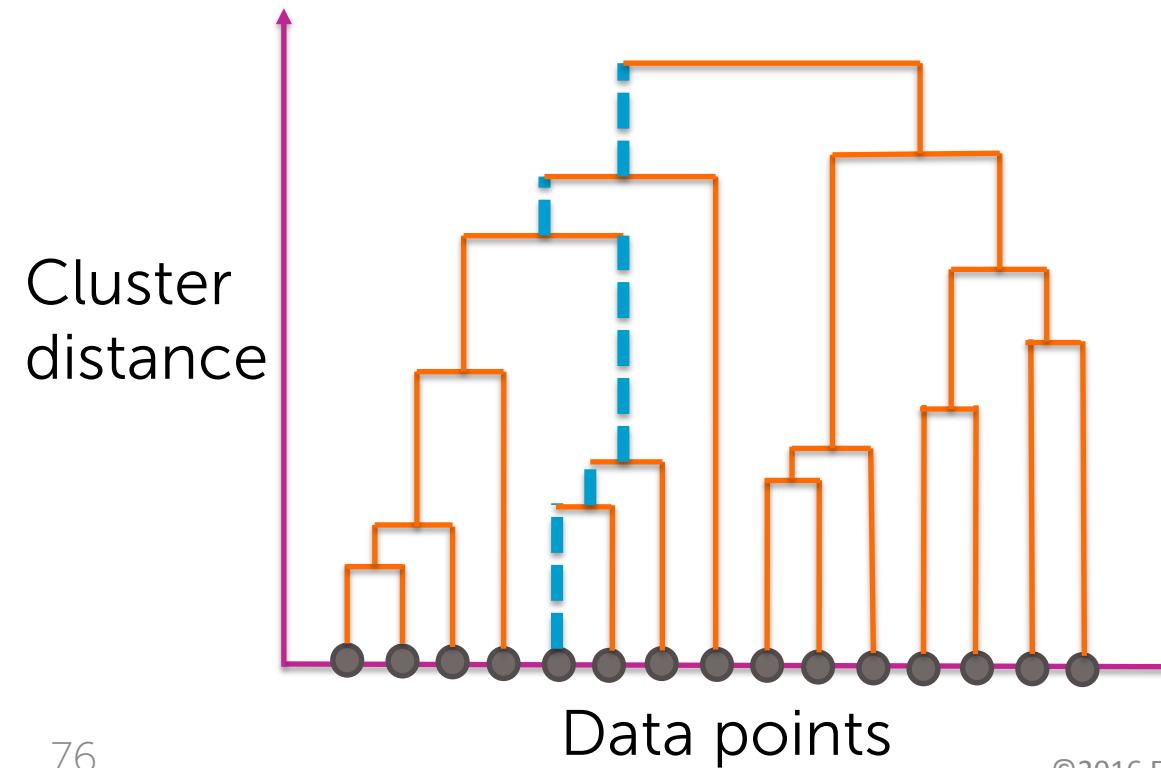
The dendrogram

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



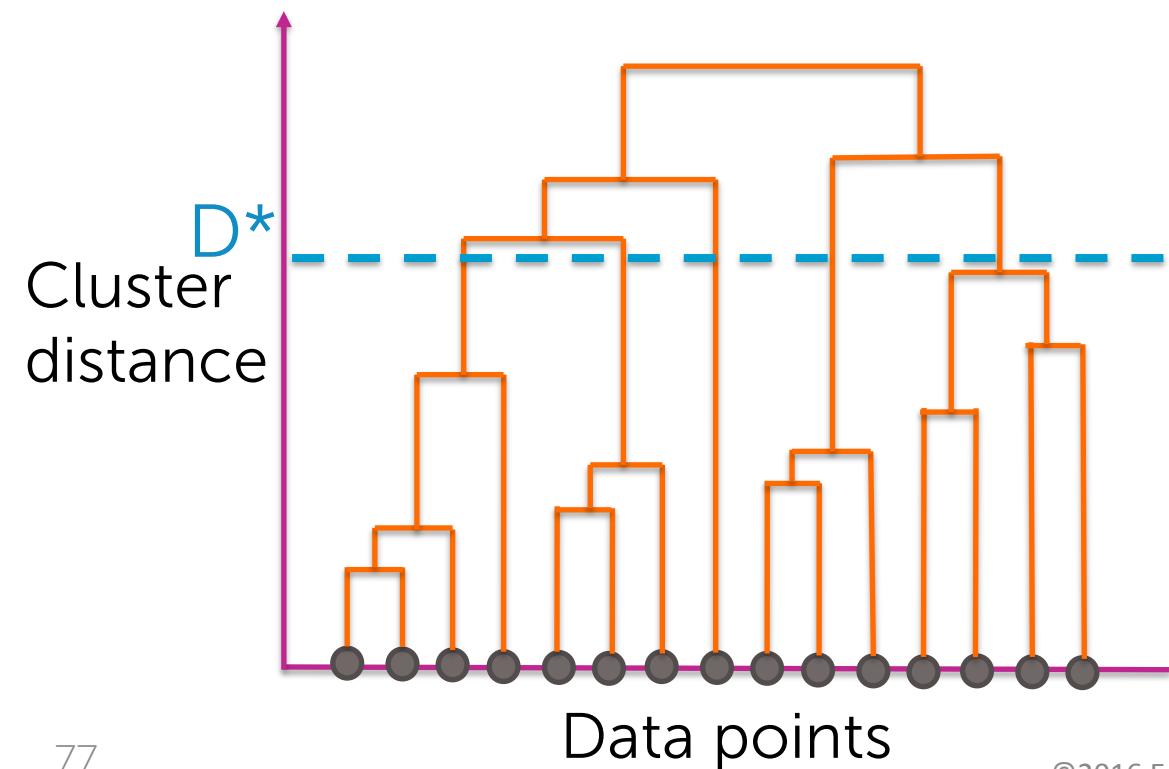
The dendrogram

Path shows all clusters to which a point belongs
and the order in which clusters merge



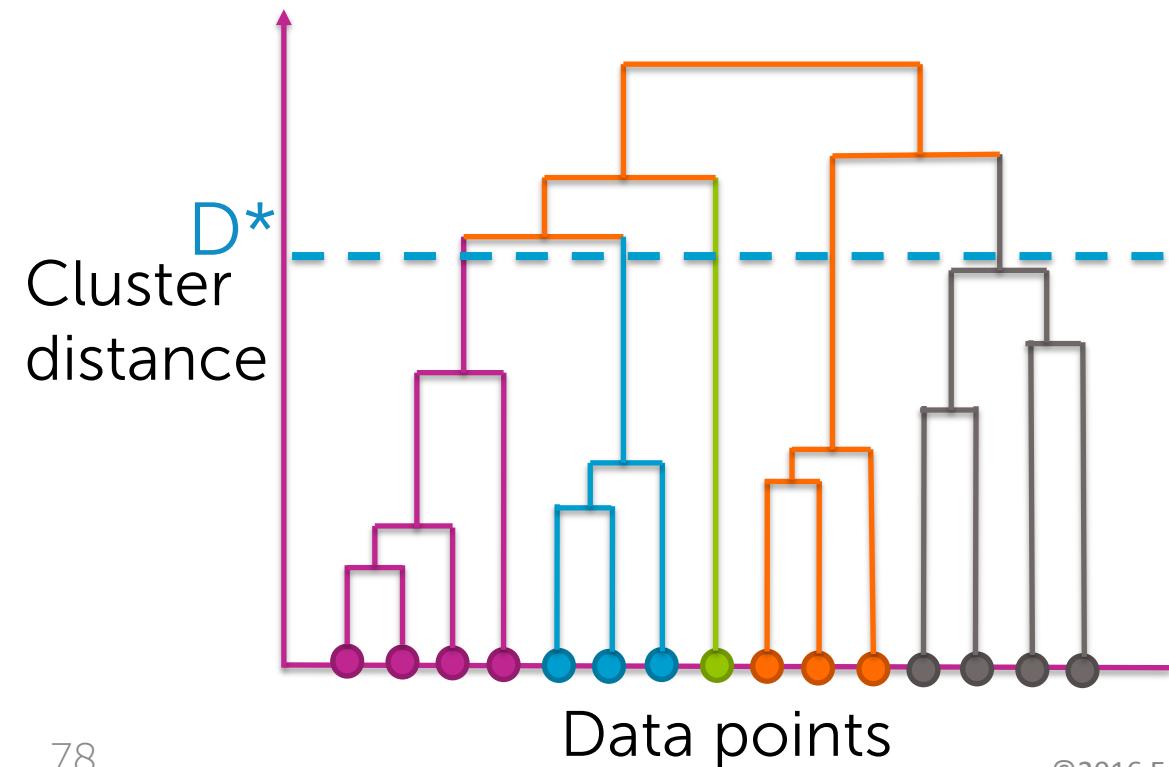
Extracting a partition

Choose a distance D^* at which to cut dendrogram



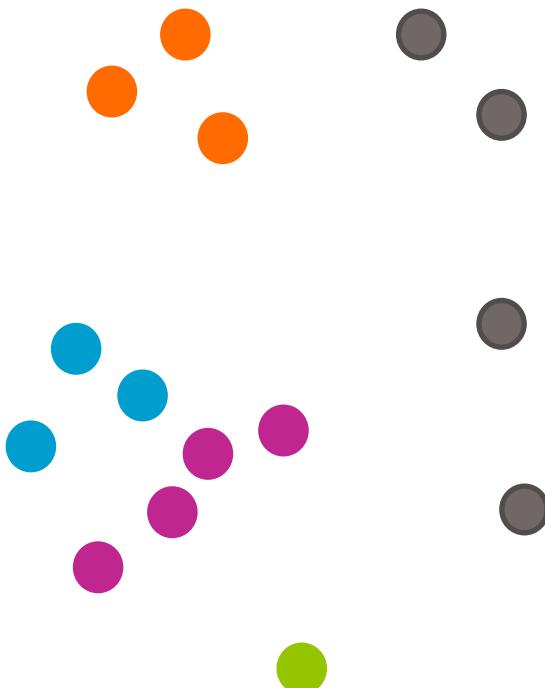
Extracting a partition

Every branch that crosses D^*
becomes a separate cluster



Extracting a partition

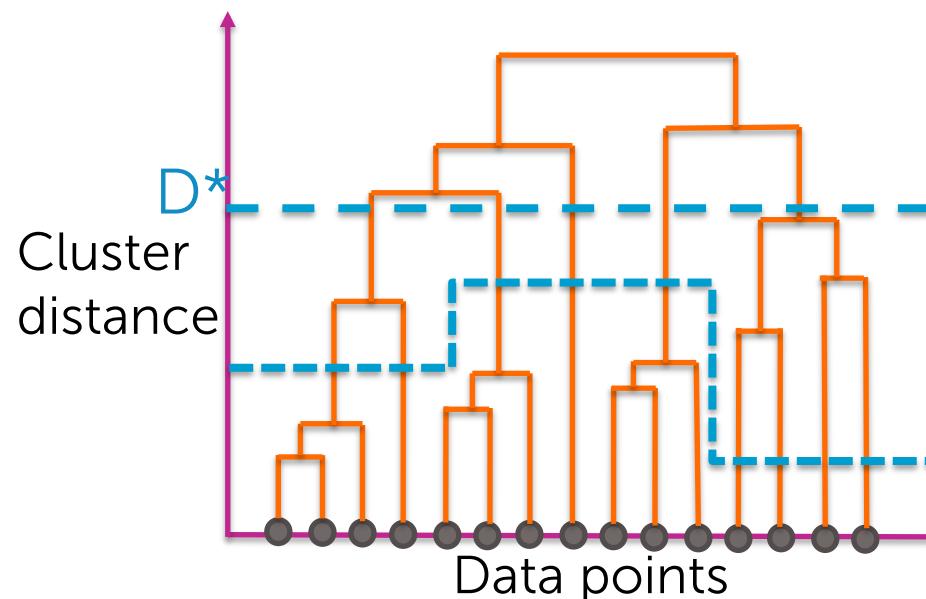
Every branch that crosses D^*
becomes a separate cluster



Agglomerative clustering details

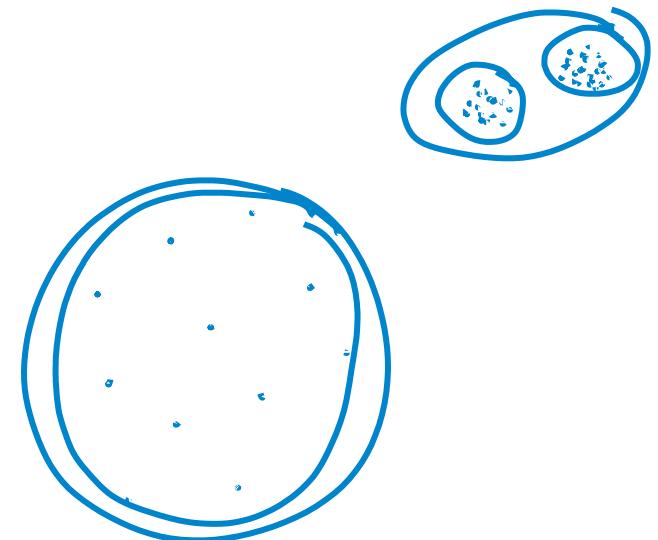
Agglomerative choices to be made

- Distance metric: $d(\mathbf{x}_i, \mathbf{x}_j)$
- Linkage function: e.g., $\min_{\substack{\mathbf{x}_i \text{ in } C_1, \\ \mathbf{x}_j \text{ in } C_2}} d(\mathbf{x}_i, \mathbf{x}_j)$
- 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



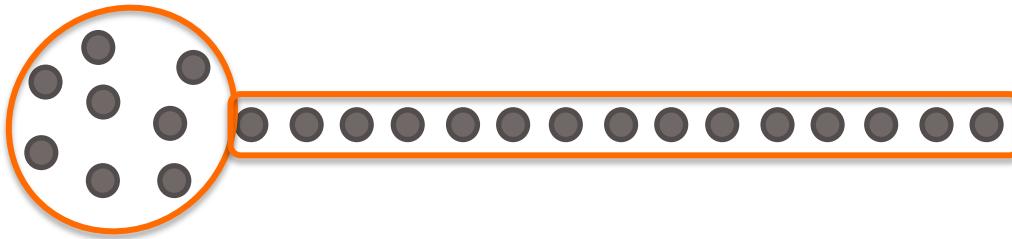
Computational considerations

- 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 $O(N^2)$

Statistical issues

Chaining: Distant points clustered together if there is a chain of pairwise close points between

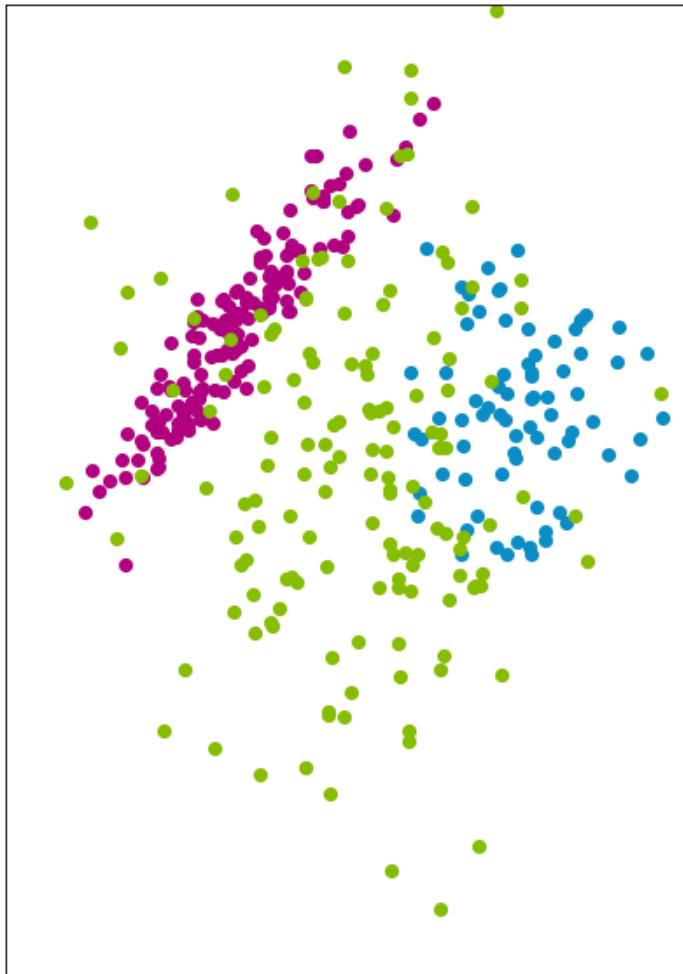


Other **linkage functions** can be more robust, but **restrict the shapes** of clusters that can be found

- **Complete linkage:**
max pairwise distance between clusters
- **Ward criterion:**
min within-cluster variance at each merge

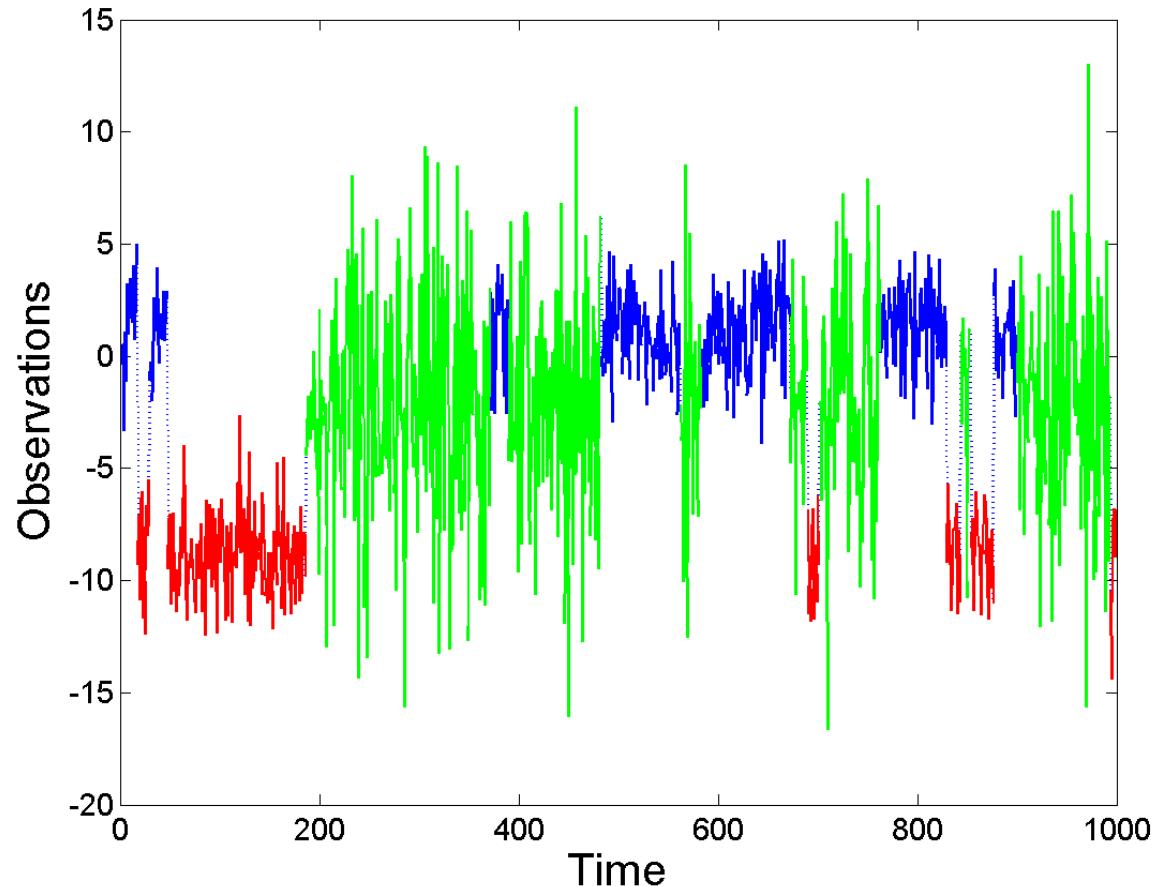
Hidden Markov models (HMMs): Another notion of “clustering”

So far, looked at clustering unordered data



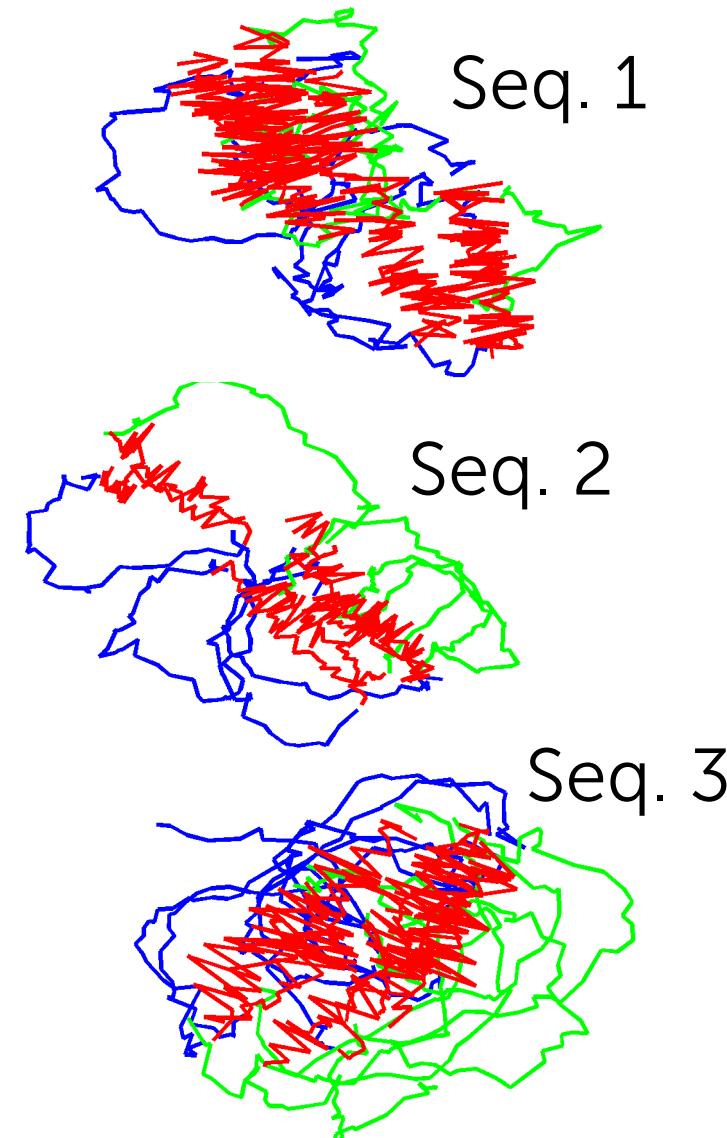
Data index (i.e., when observation was recorded) does not influence clustering

What if we have time series data?



Would be hard to distinguish **red**, **blue**, and **green** clusters if we ignored order of data

Example: Honey bee dances

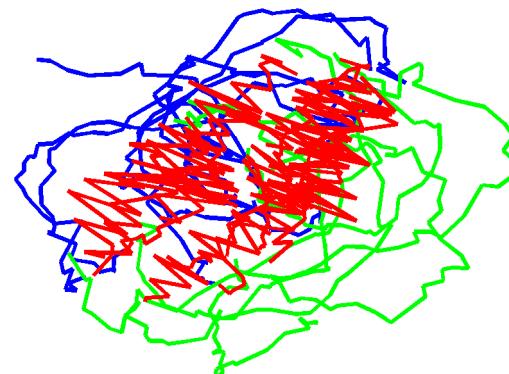


Repeated patterns of dance transitions

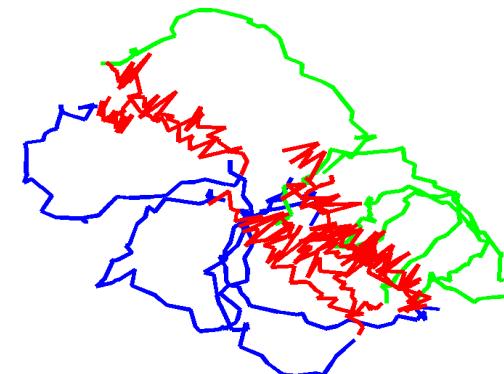
Sequence 1



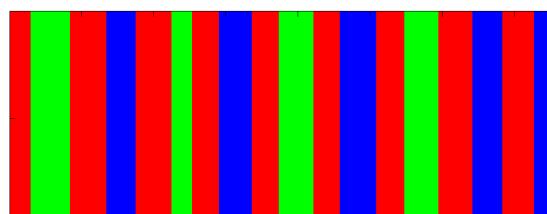
Sequence 2



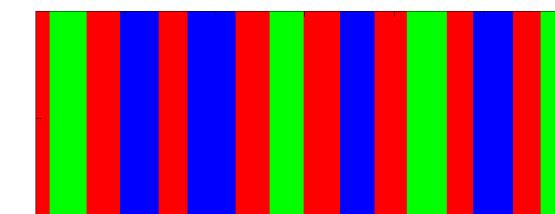
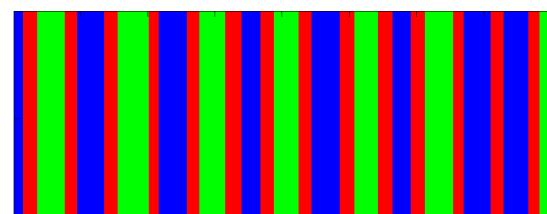
Sequence 3



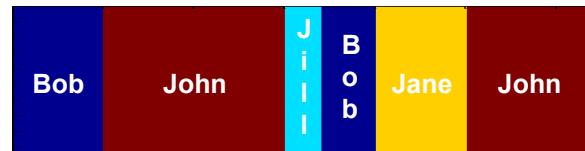
Cluster labels over time



waggle
dance turn right turn left



Similar ideas appear in many applications

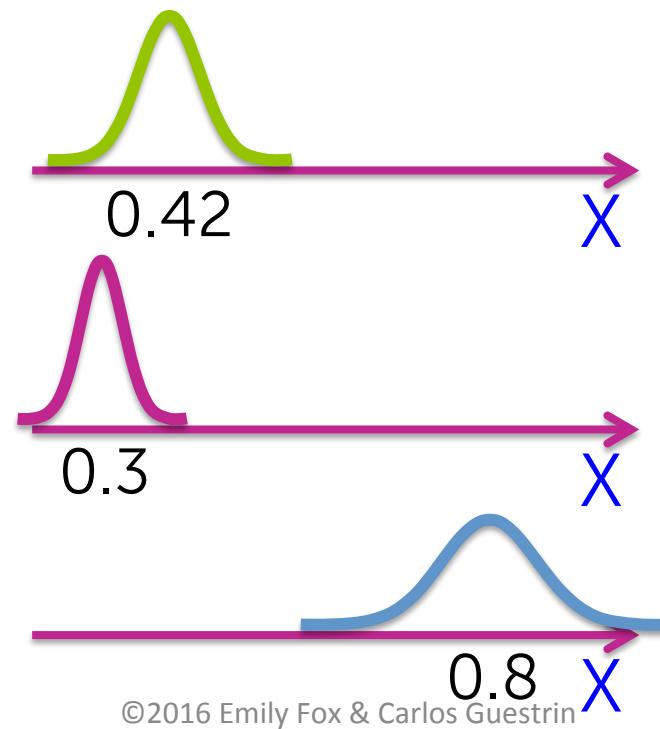


Hidden Markov model (HMM)

As in mixture model...

Every observation x_t is associated with cluster assignment variable z_t

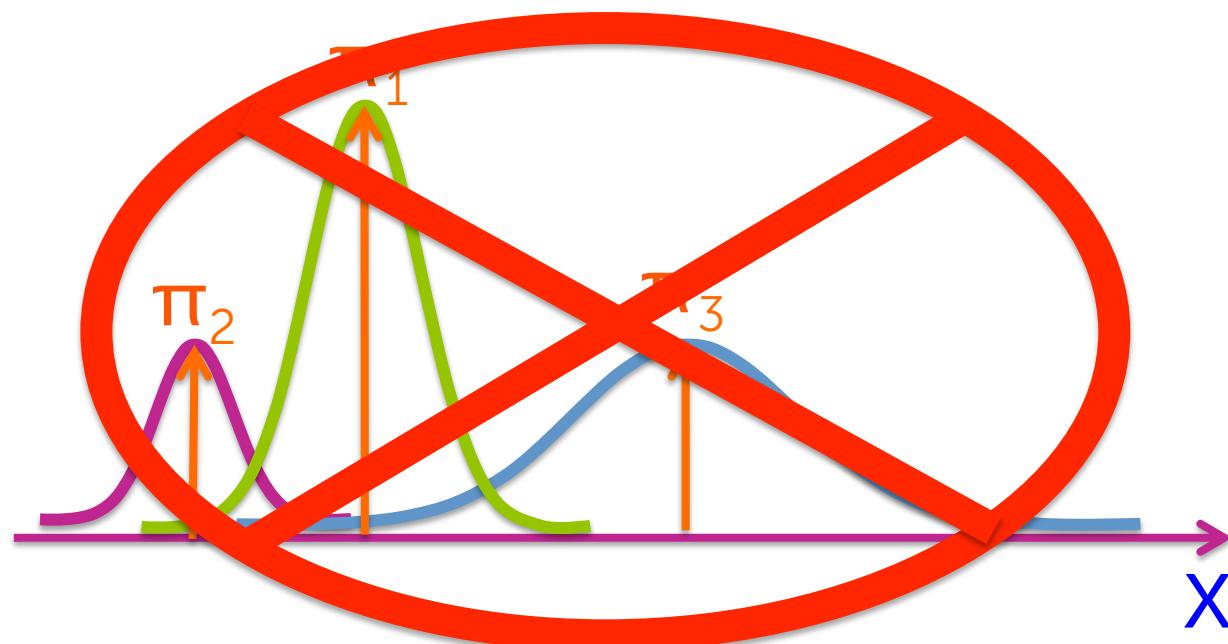
Each cluster has a distribution over observed values



Hidden Markov model (HMM)

Difference from mixture model:

Probability of ($z_t = k$) depends on previous cluster assignment z_{t-1}



Inference in HMMs

- Learn MLE of HMM parameters using EM algorithm = **Baum Welch**
- Infer MLE of state sequence given fixed model parameters using dynamic programming = **Viterbi algorithm**
- Infer soft assignments of state sequence using dynamic programming = **forward-backward algorithm**

What we didn't cover

Other clustering + retrieval topics

Retrieval:

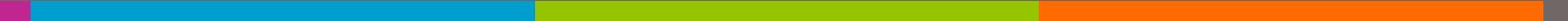
- Other distance metrics
- Distance metric learning

Clustering:

- Nonparametric clustering
- Spectral clustering

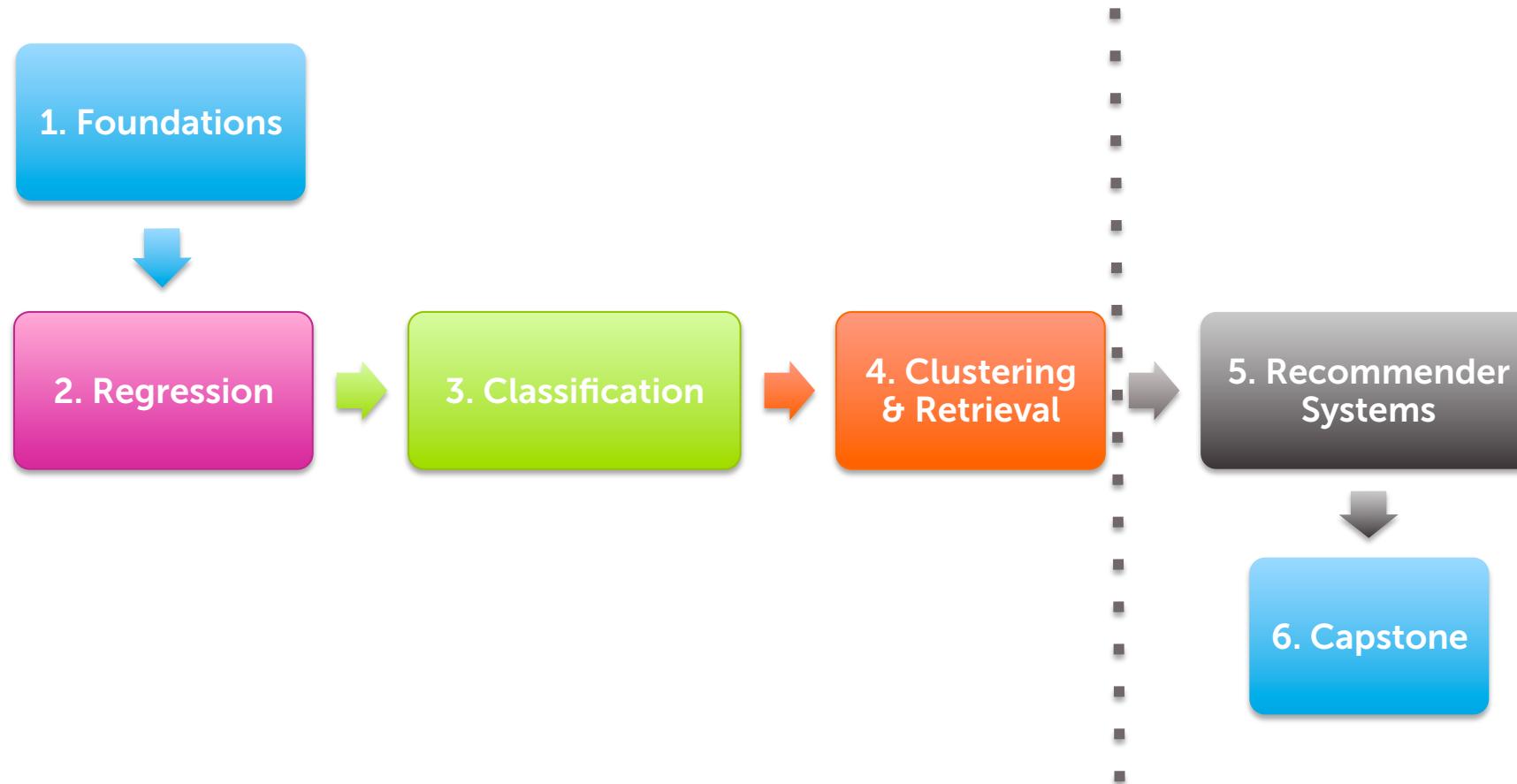
Related ideas:

- Density estimation
- Anomaly detection



What's ahead in this specialization

This course is a part of the Machine Learning Specialization



5. Recommender Systems & Dimensionality Reduction

Case study: Recommending Products

Models

- Collaborative filtering
- Matrix factorization
- PCA

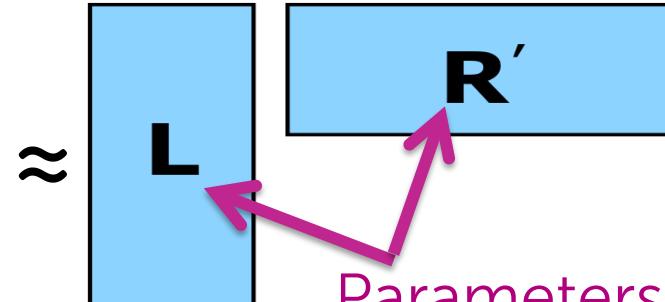
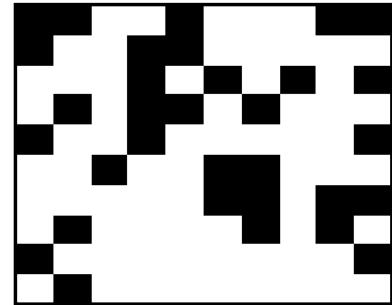
Algorithms

- Coordinate descent
- Eigen decomposition
- SVD

Concepts

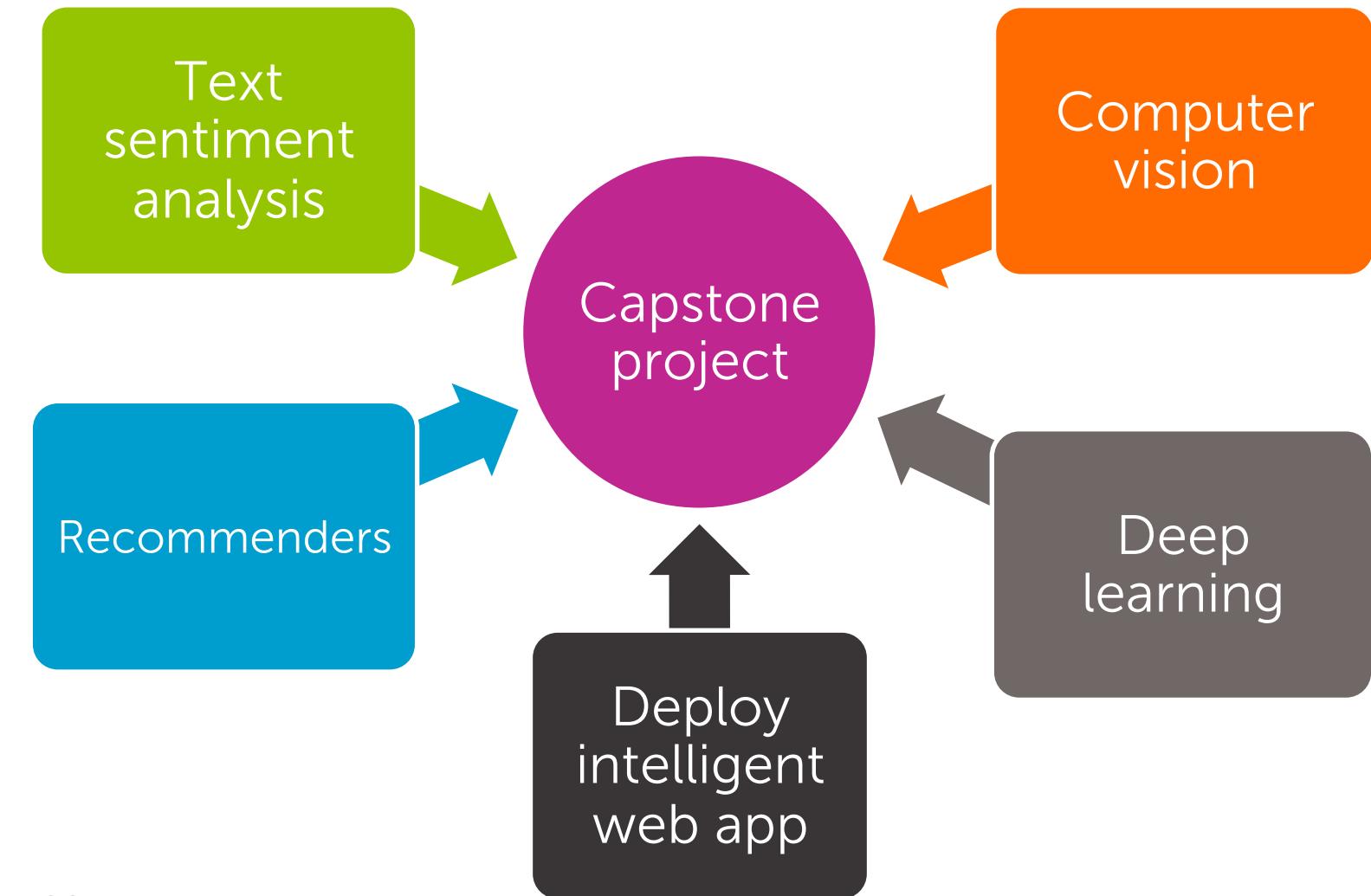
- Matrix completion, eigenvalues, cold-start problem, diversity, scaling up

Rating =



Parameters
of model

6. Capstone: *Build and deploy an intelligent application with deep learning*





Thank you...