

Clustering & Retrieval: A machine learning perspective

Emily Fox & Carlos Guestrin
Machine Learning Specialization
University of Washington

Part of a specialization

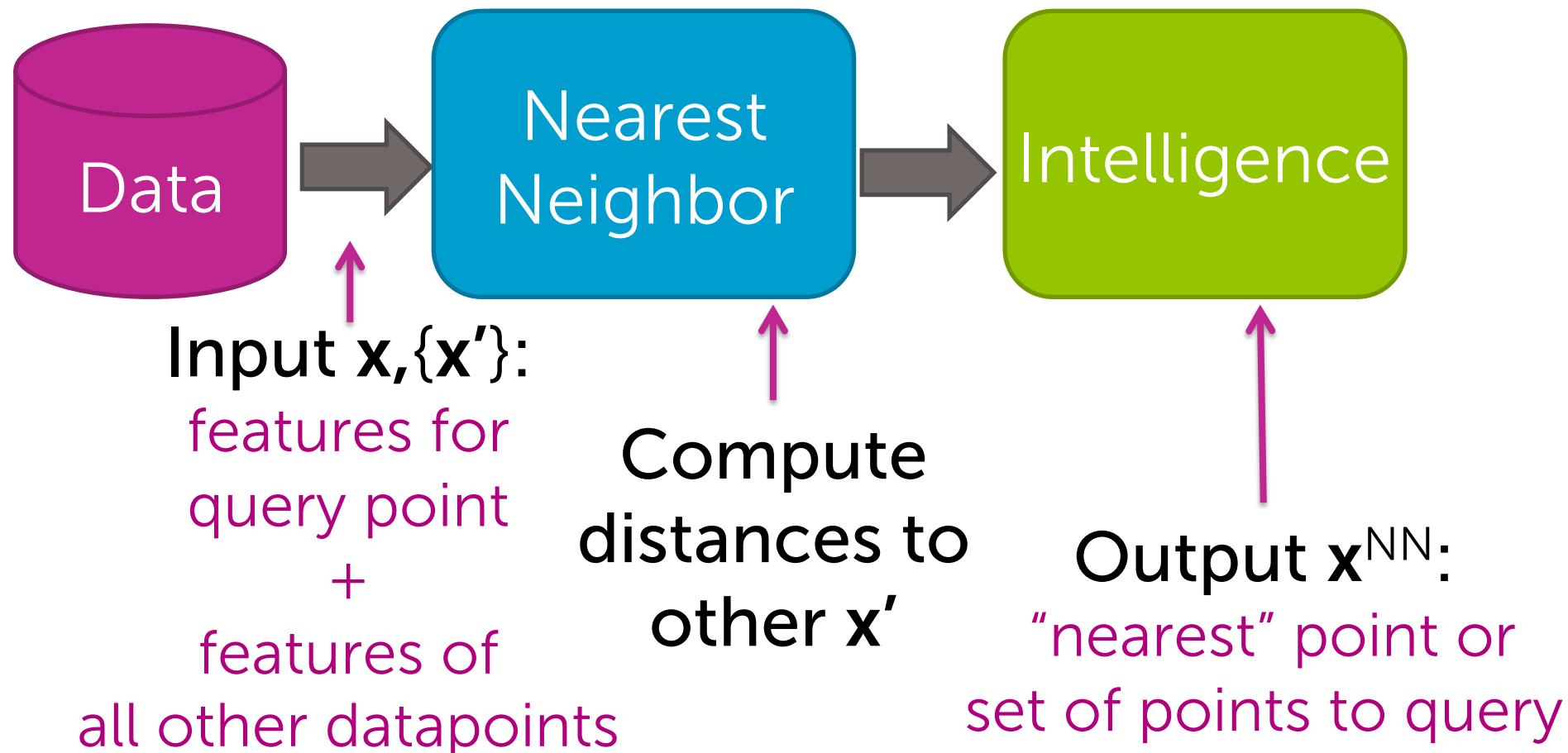
This course is a part of the Machine Learning Specialization



What is the course about?

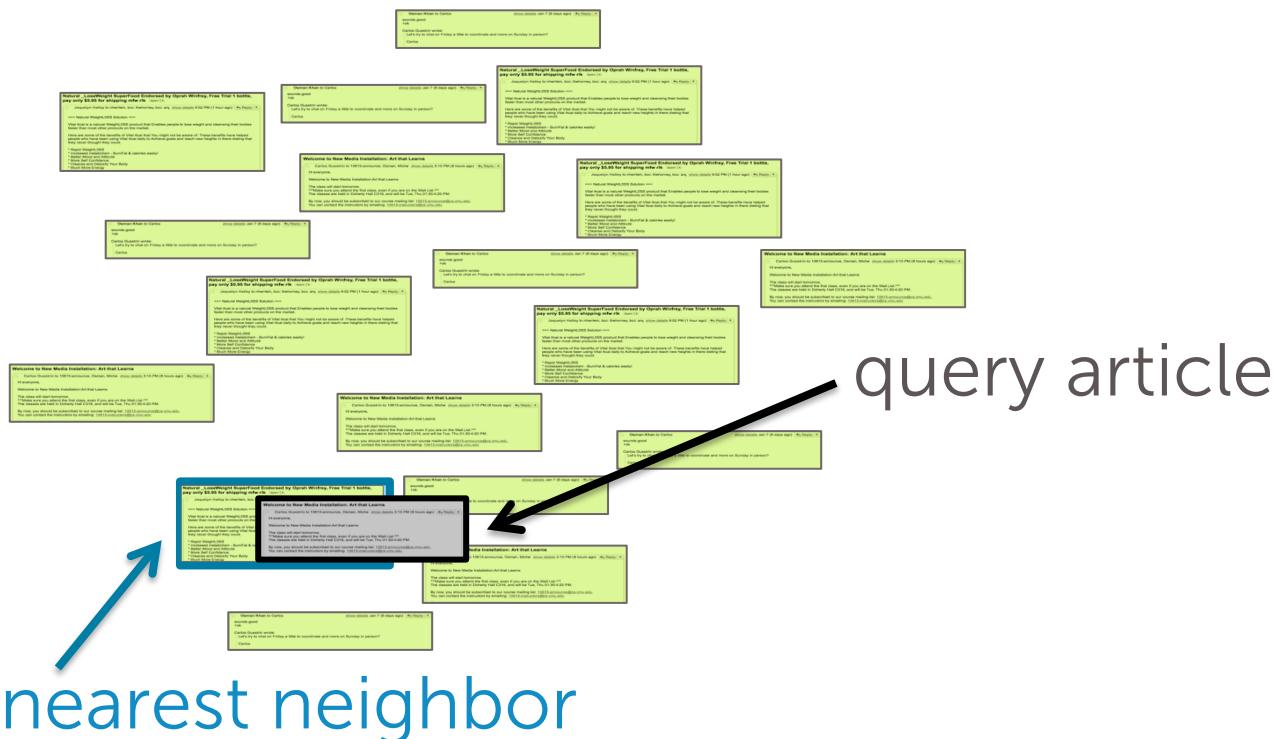
What is retrieval?

Search for related items



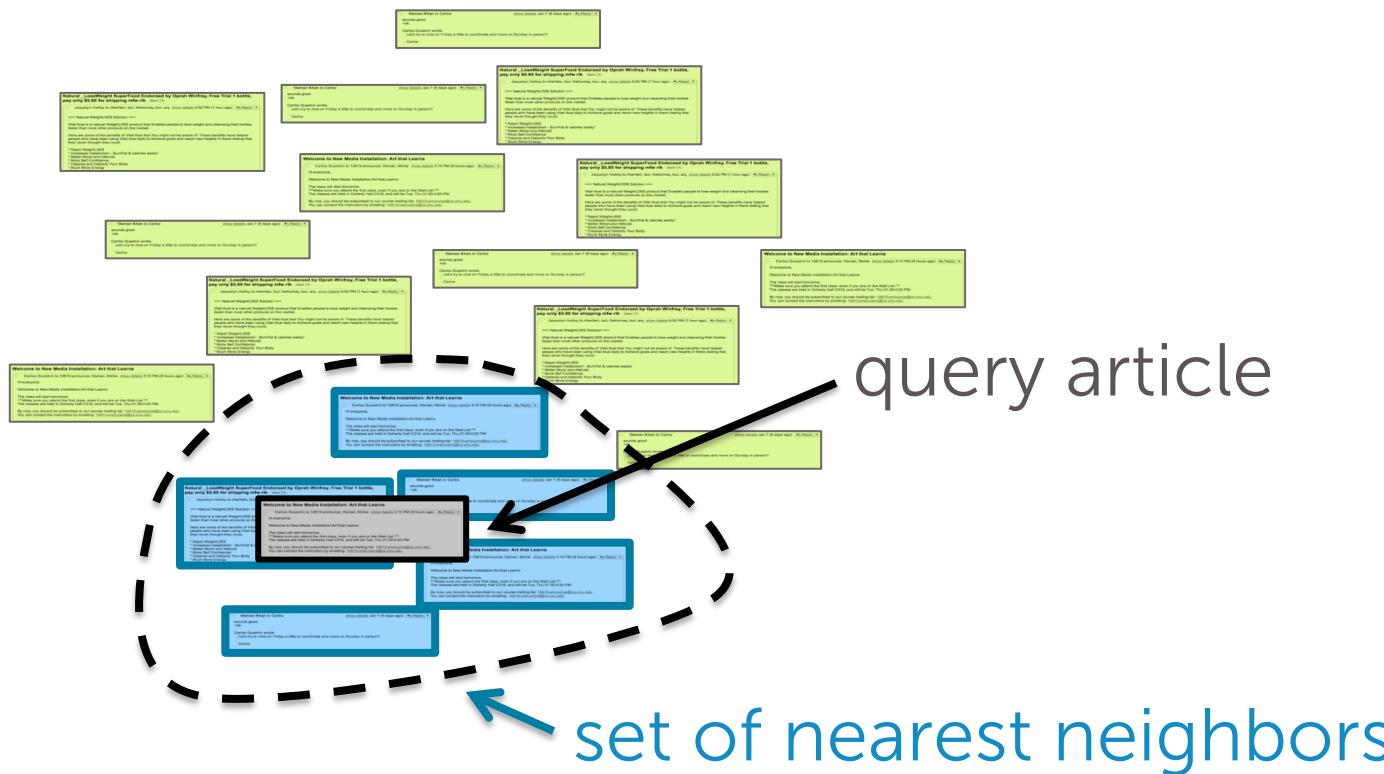
Retrieve “nearest neighbor” article

Space of all articles,
organized by similarity of text



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



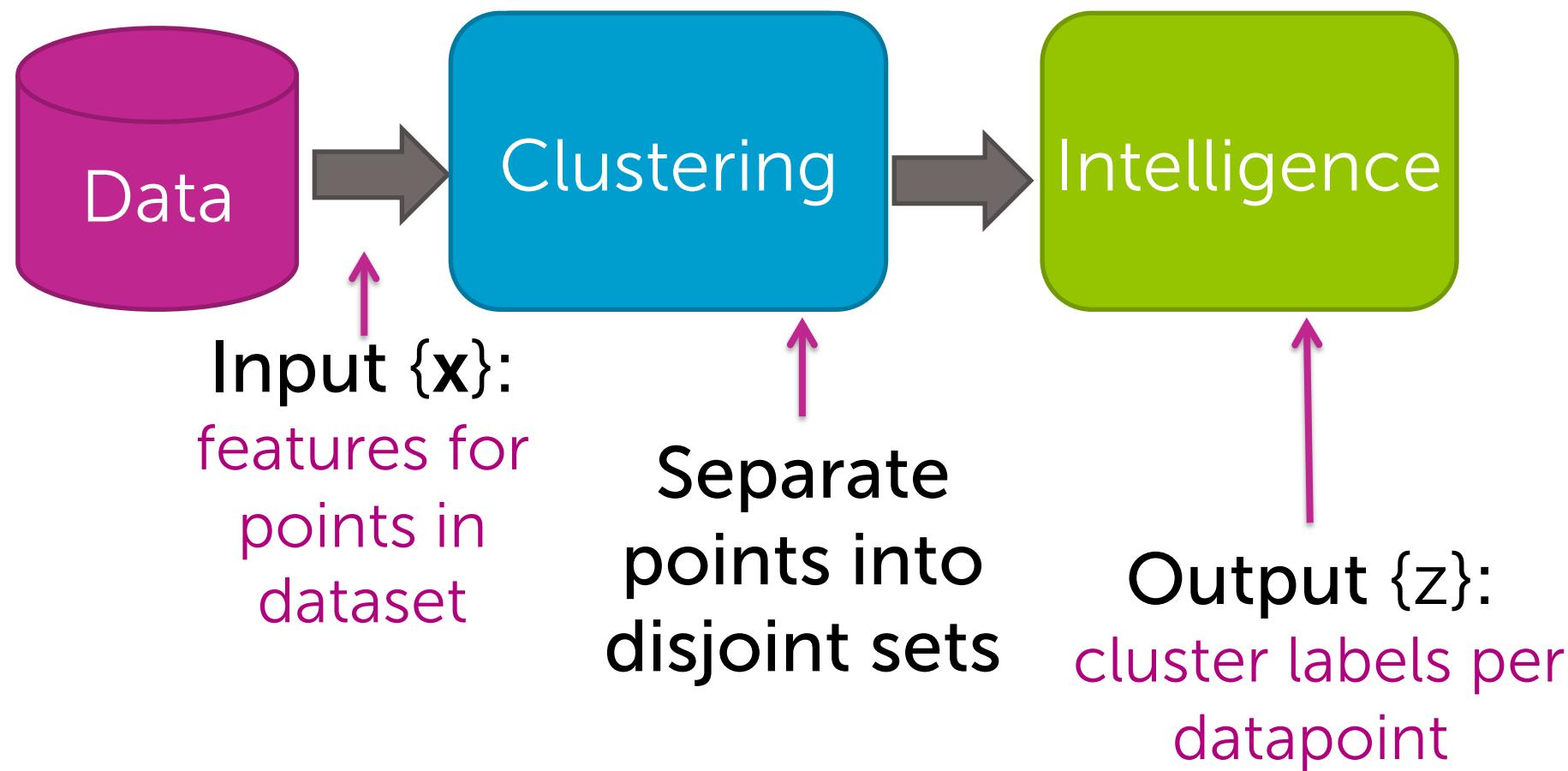
Social networks

(people you might want
to connect with)



What is clustering?

Discover groups of similar inputs



Case Study: Clustering documents by “topic”

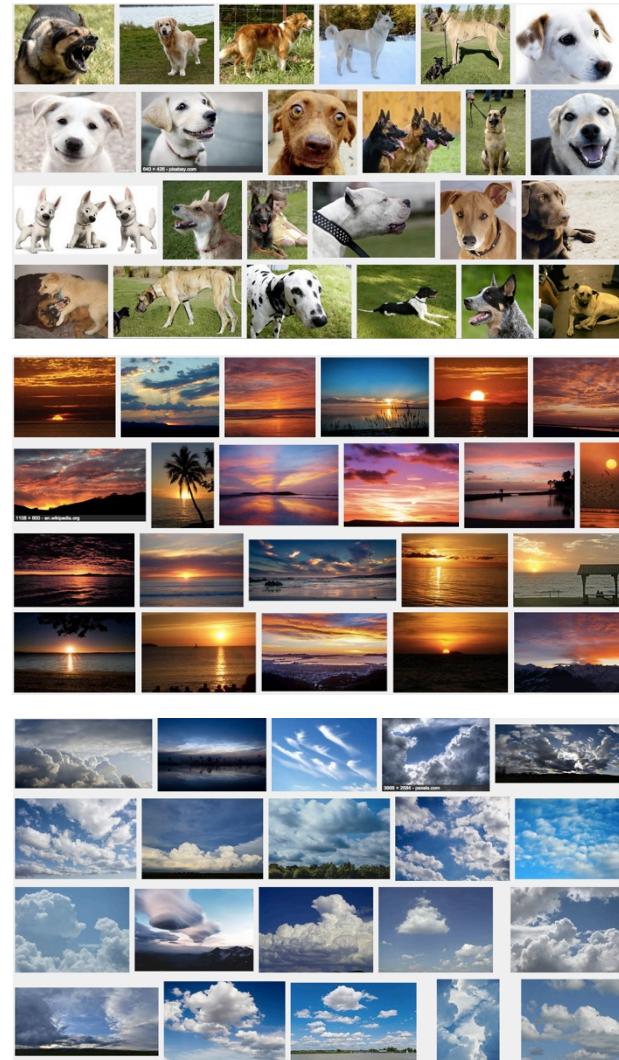
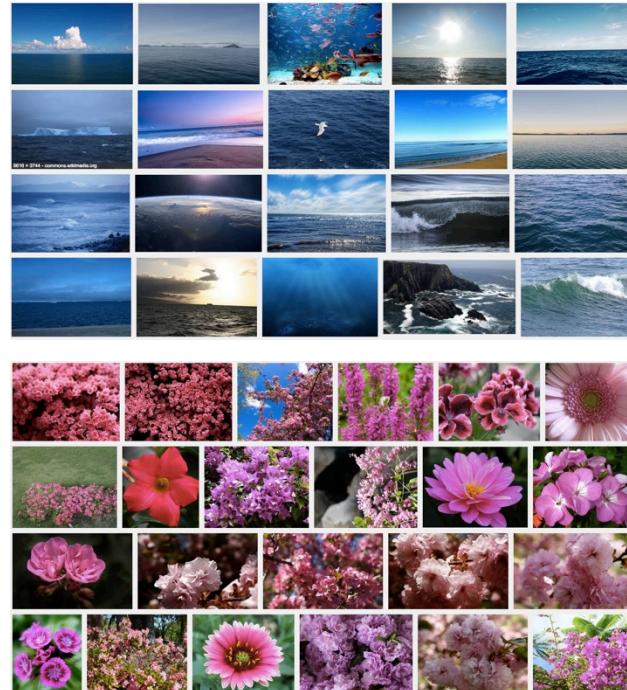


Just like retrieval, clustering has applications almost everywhere

Clustering images

For search, group as:

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



Or Coursera learners...

Discover groups of learners for better targeting of courses



Impact of retrieval & clustering

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)

Course overview

Course philosophy: Always use case studies & ...

Core
concept

Visual

Algorithm

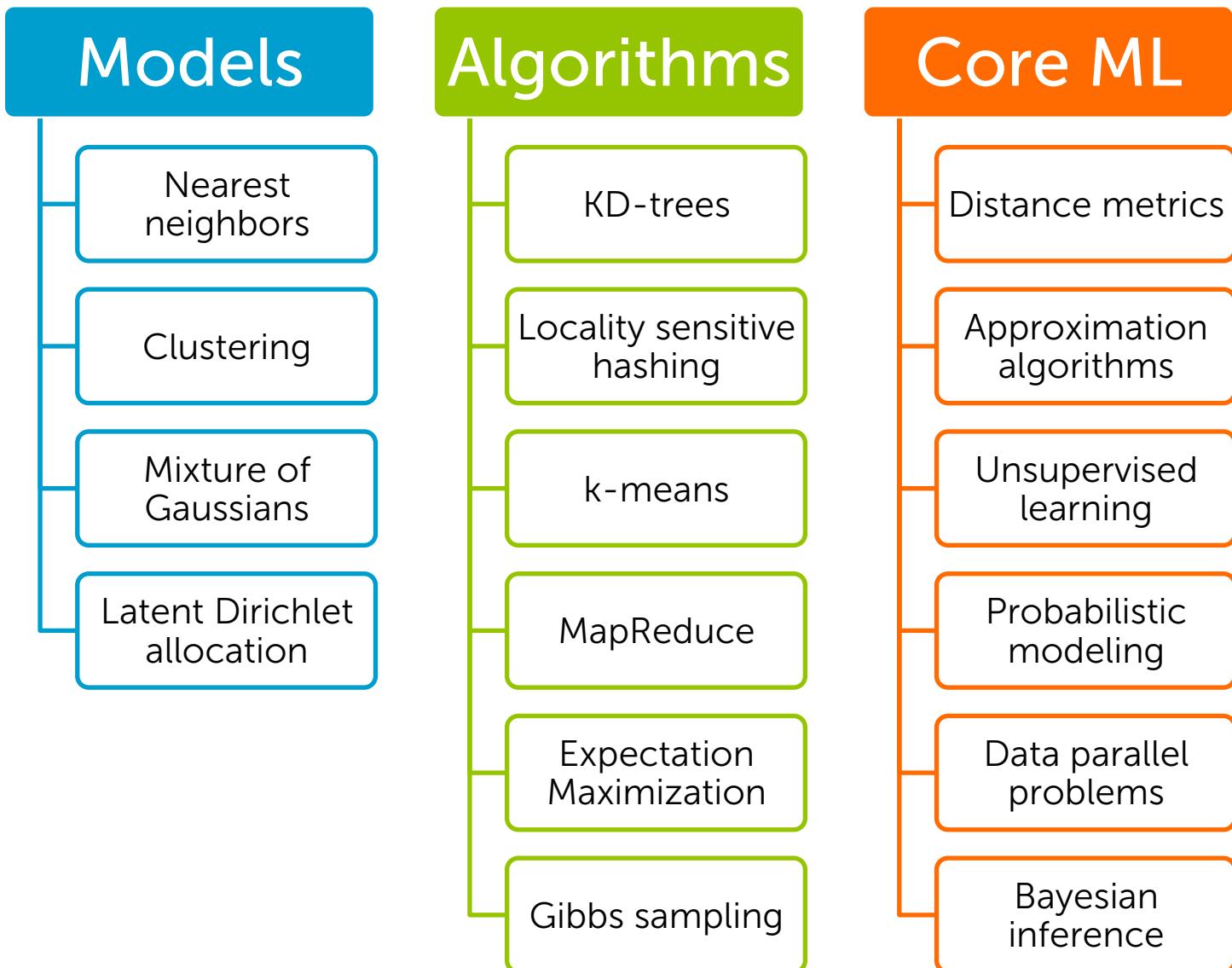
Practical

Implement

Advanced
topics

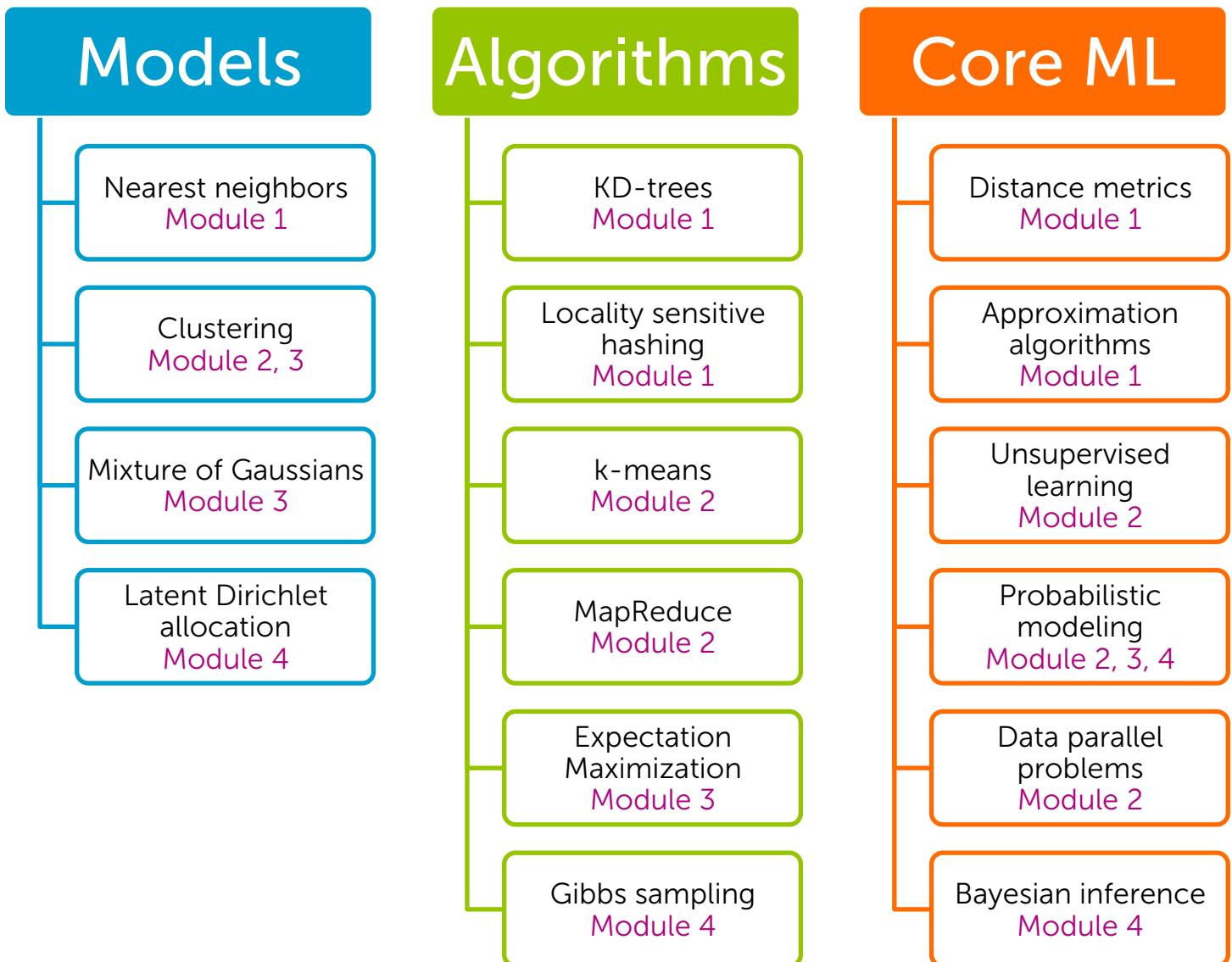
OPTIONAL

Overview of content



Course outline

Overview of content



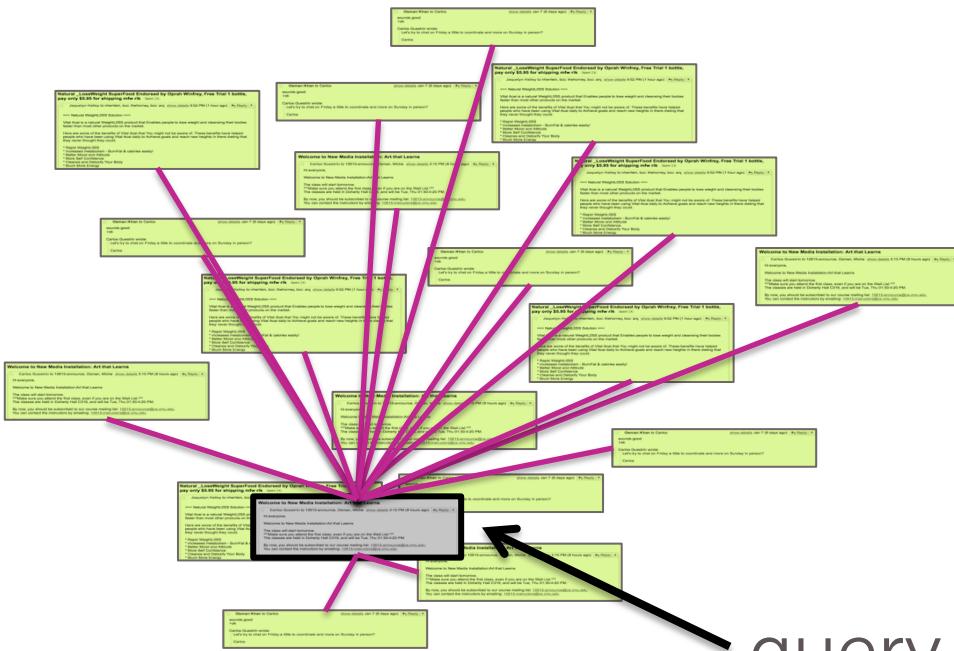
Module 1: Nearest neighbor search



Reading doc
and want to
find related
doc

Module 1: Nearest neighbor search

Compute distances to all other documents and return closest



Critical elements:

- Doc representation
- Distance measure

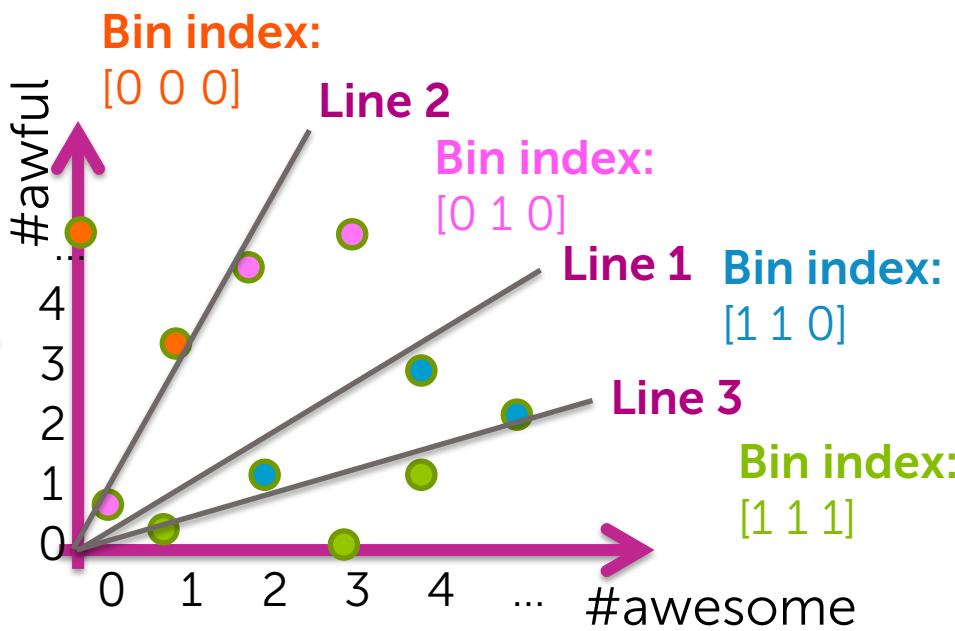
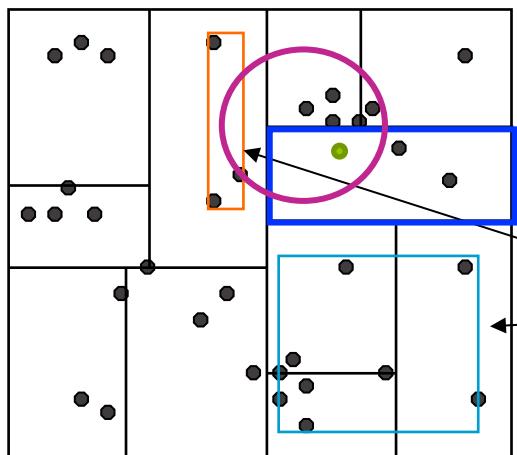
query article

Module 1: Nearest neighbor search

Efficient and
approximate
NN search

KD-trees

LSH →



Module 2: k-means and MapReduce

Discover *clusters* of related documents



Cluster 1



Cluster 2



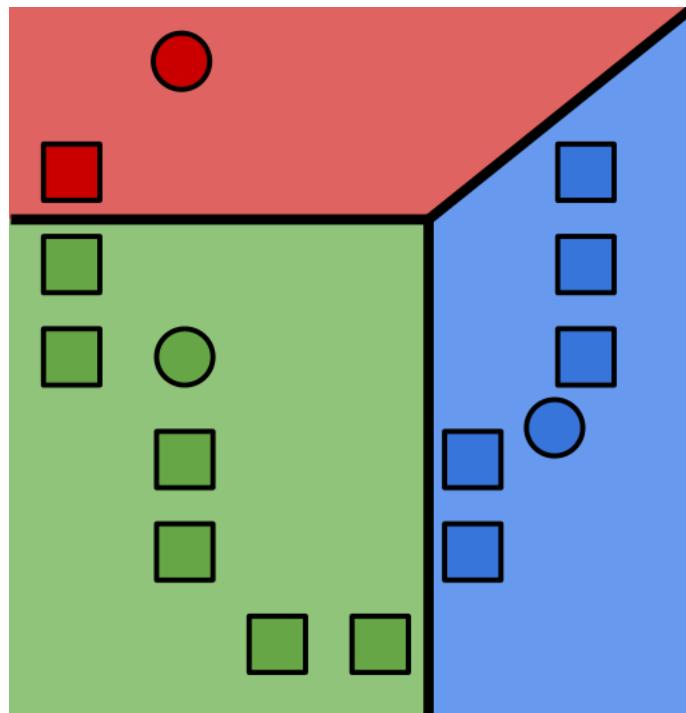
Cluster 3



Cluster 4

Module 2: k-means and MapReduce

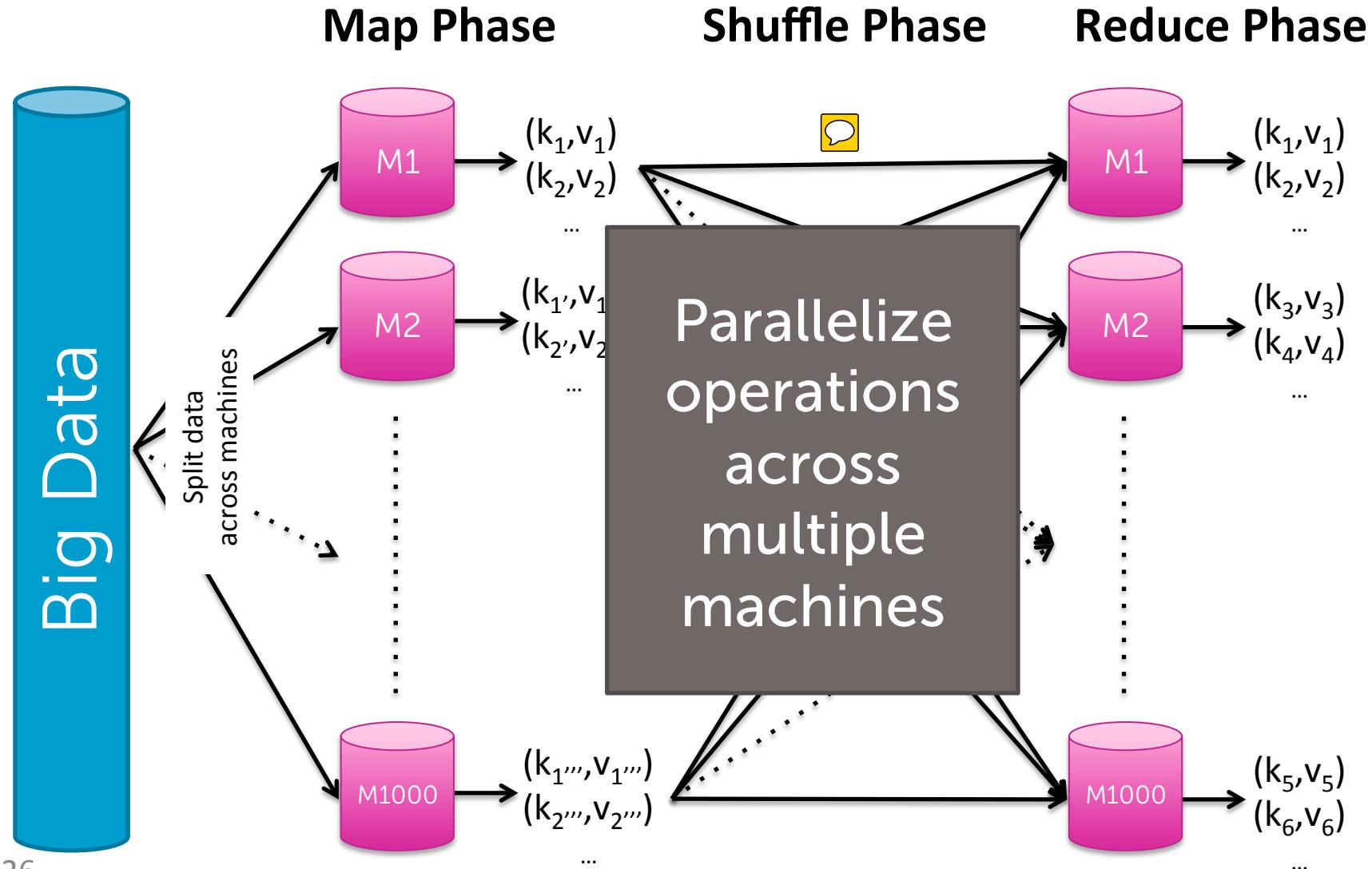
k-means aims to minimize sum of square distances to cluster centers



Makes **hard assignments** of data points to clusters

Unsupervised learning task

Module 2: k-means and MapReduce



Module 3: Mixture Models

Probabilistic clustering model



Cluster 1



captures
uncertainty
in clustering

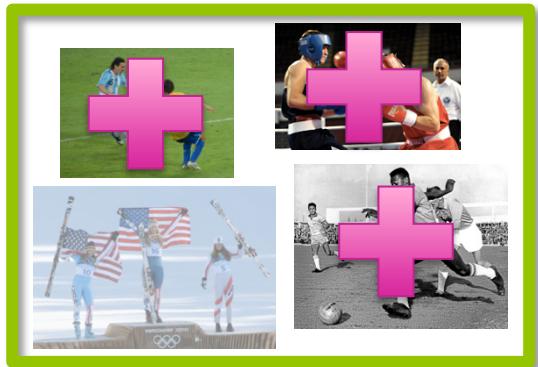


Cluster 3

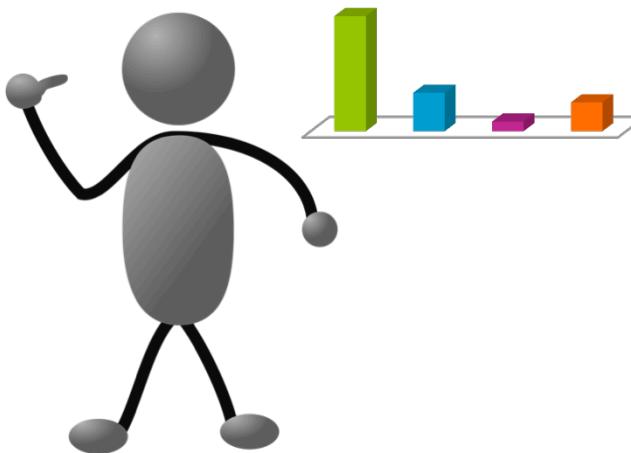


Cluster 4

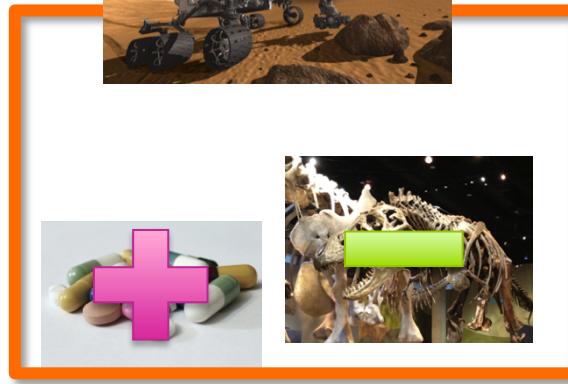
Module 3: Mixture Models



Cluster 1



Cluster 3

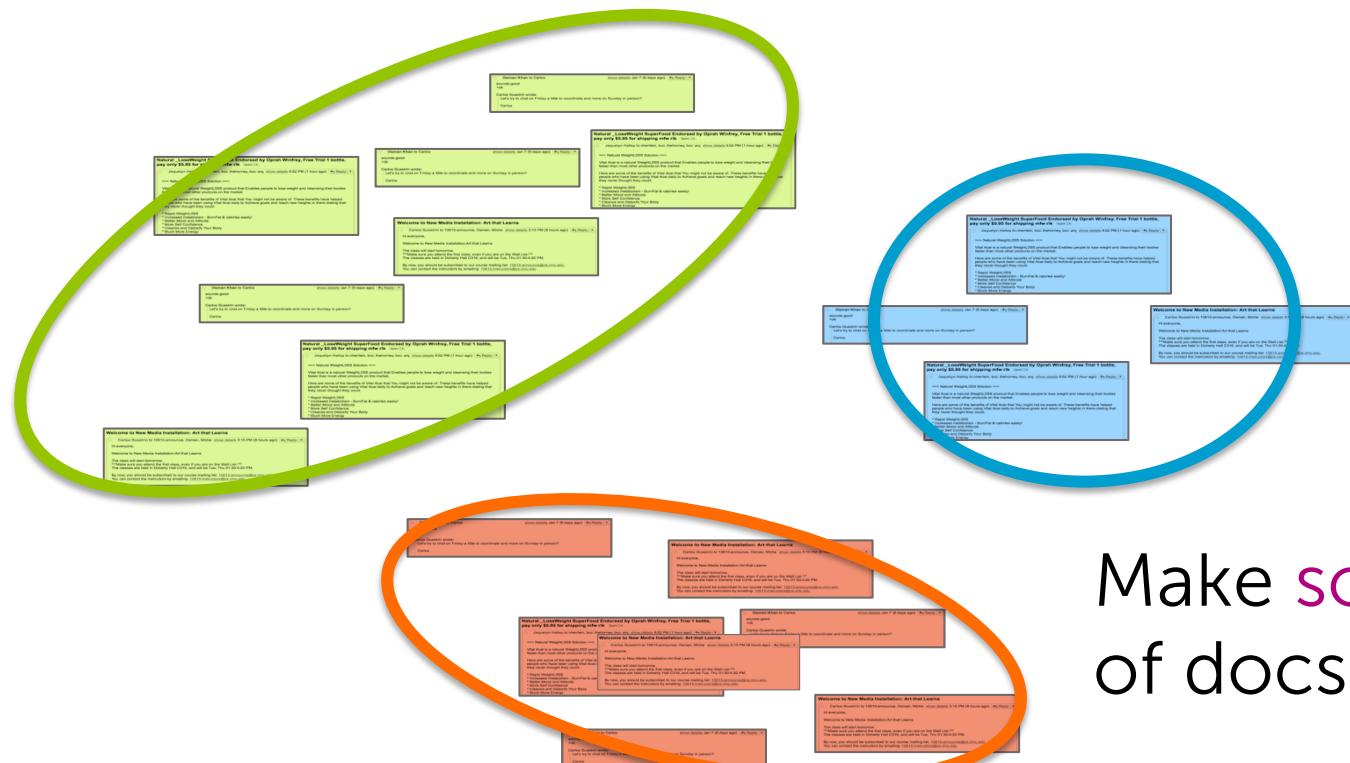


Cluster 4

Learn user
topic
preferences

Module 3: Mixture Models

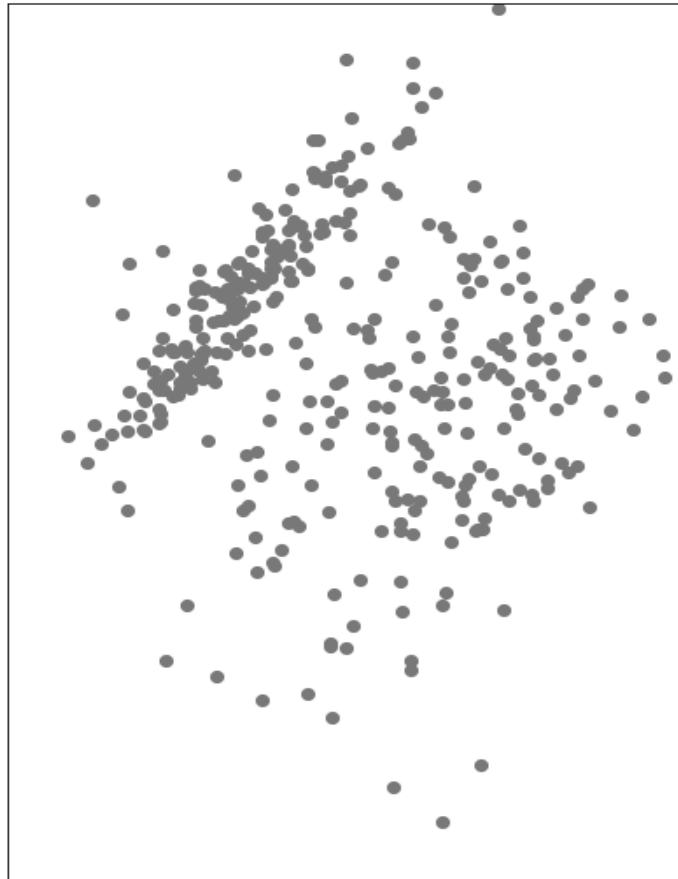
Assignments of docs to clusters based on location *and shape*, not just location



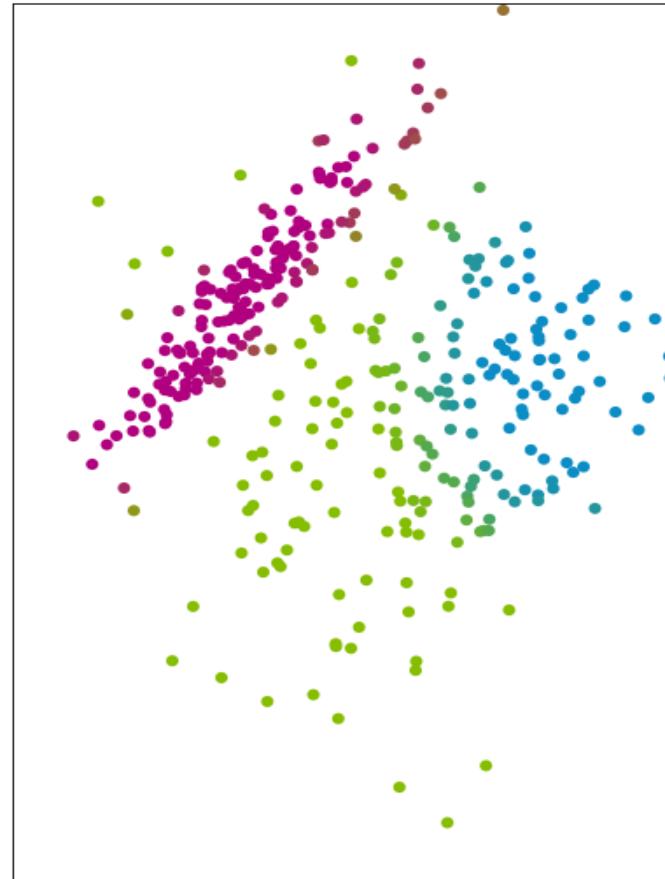
Make **soft assignments**
of docs to each cluster

Module 3: Mixture Models

Data



EM algorithm →
soft assignments



Module 4: Latent Dirichlet Allocation

^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

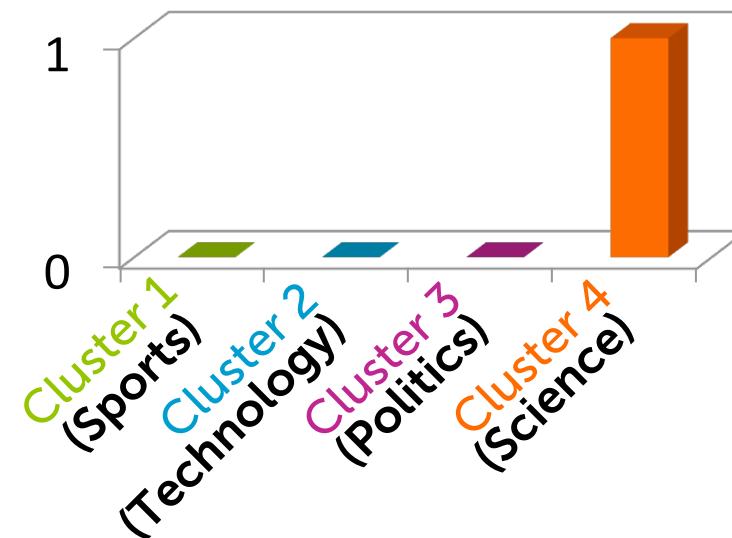
Patients with epilepsy can manifest short, sub-clinical epileptic “bursts” in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics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

Based on science related words, maybe doc in cluster 4



Module 4: Latent Dirichlet Allocation

Danilo P. Mandic, Emily B. Fox, Brian P. O'Leary^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

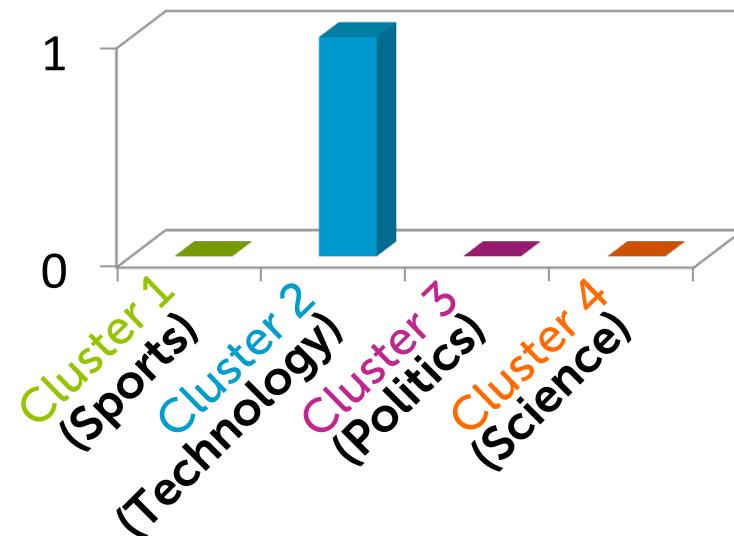
Patients with epilepsy can manifest short, sub-clinical epileptic “bursts” in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics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

Or maybe cluster 2
(technology) is a better fit



Module 4: Latent Dirichlet Allocation

^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

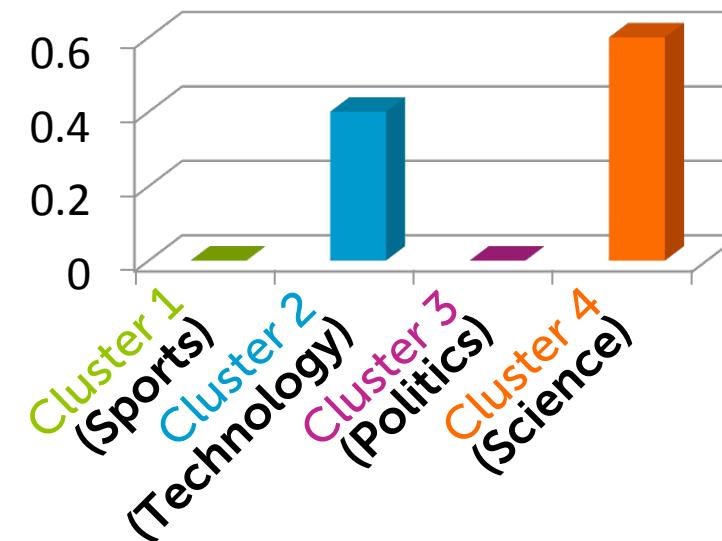
Patients with epilepsy can manifest short, sub-clinical epileptic “bursts” in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics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

Really, it's about science
and technology



Module 4: Latent Dirichlet Allocation

Each cluster/topic defined by probability of words in vocab

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
...	...

...

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:



Unsupervised learning task

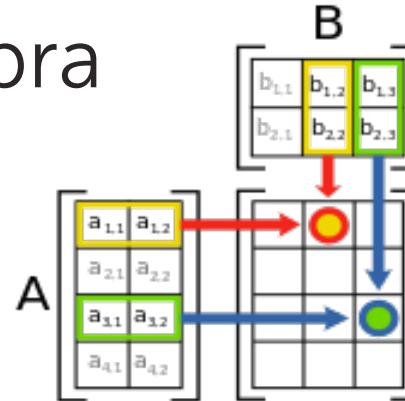
Assumed background

Courses 1, 2, & 3 in this ML Specialization

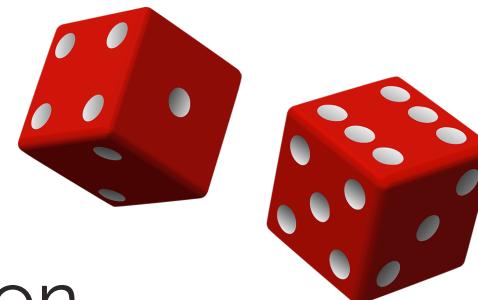
- Course 1: Foundations
 - Overview of ML case studies
 - Black-box view of ML tasks
 - Programming & data manipulation skills
- Course 2: Regression
 - Data representation (input, output, features)
 - Basic statistical concepts: mean/variance
 - Basic ML concepts:
 - ML algorithm
 - Coordinate ascent
 - Overfitting
 - Regularization
- Course 3: Classification
 - Distributions and conditional distributions
 - Maximum likelihood estimation
 - References to:
 - Linear classifier
 - Multiclass classification
 - Boosting

Math background

- Basic linear algebra
 - Vectors
 - Matrices
 - Matrix multiply



- Basic probability
 - Fundamental laws
 - Distribution and conditional distribution



Programming experience

- Basic Python used
 - Can pick up along the way if knowledge of other language

```
get_user(self, user):
    """
    Returns a QuerySet of connections for user.
    """
    set1 = self.filter(from_user=user).select_related(depth=1)
    set2 = self.filter(to_user=user).select_related(depth=1)
    return set1 | set2

def are_connected(self, user1, user2):
    if self.filter(from_user=user1, to_user=user2).count() > 0:
        return True
    if self.filter(from_user=user2, to_user=user1).count() > 0:
        return True
    return False

def remove(self, user1, user2):
    """
    Deletes proper object regardless of the order of users in argument
    """
    connection = self.filter(from_user=user1, to_user=user2)
    if not connection:
        connection = self.filter(from_user=user2, to_user=user1)
    connection.delete()
---:--- models.py Top L1 (Python AC yas)---
```



Reliance on GraphLab Create

- SFrames will be used, though not required
 - open source project of Dato
(creators of GraphLab Create)
 - can use pandas and numpy instead
- Assignments will:
 1. Use GraphLab Create to explore high-level concepts
 2. Ask you to implement most algorithms without GraphLab Create
- Net result:
 - learn how to code methods in Python



Computing needs

- Using your own computer:
 - Basic desktop or laptop
 - 64-bit required if using SFrame
 - Access to internet
 - Ability to:
 - Install and run Python (and Numpy, GraphLab Create,...)
 - Store a few GB of data
- Will also provide alternative, pre-configured machine in Cloud





Let's get started!