GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 2: MACHINE LEARNING

LAST TIME:

- FIRST LOOK AT DATA SCIENCE & THE DATA MINING WORKFLOW
- DATA EXPLORATION WITH UNIX
- DATA VISUALIZATION WITH R & GGPLOT2

QUESTIONS?

AGENDA

I. WHAT IS MACHINE LEARNING?

II. MACHINE LEARNING PROBLEMS

I. WHAT IS MACHINE LEARNING?

WHAT IS MACHINE LEARNING?

from Wikipedia:

“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can *learn from data*.”

source: http://en.wikipedia.org/wiki/Machine_learning

WHAT IS MACHINE LEARNING?

from Wikipedia:

“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can *learn from data*.”

“The core of machine learning deals with *representation* and *generalization*…”

source: http://en.wikipedia.org/wiki/Machine_learning

WHAT IS MACHINE LEARNING?

from Wikipedia:

“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can *learn from data*.”

“The core of machine learning deals with *representation* and *generalization*…”

- *representation* – extracting structure from data

source: http://en.wikipedia.org/wiki/Machine_learning

WHAT IS MACHINE LEARNING?

from Wikipedia:

“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can *learn from data*.”

“The core of machine learning deals with *representation* and *generalization*…”

- *representation* – extracting structure from data
- *generalization* – making predictions from data

source: http://en.wikipedia.org/wiki/Machine_learning

II. MACHINE LEARNING PROBLEMS

TYPES OF LEARNING PROBLEMS

supervised	
unsupervised	

TYPES OF LEARNING PROBLEMS

supervised	making predictions
unsupervised	discovering patterns

TYPES OF LEARNING PROBLEMS

supervised	labeled examples
unsupervised	no labeled examples

TYPES OF DATA

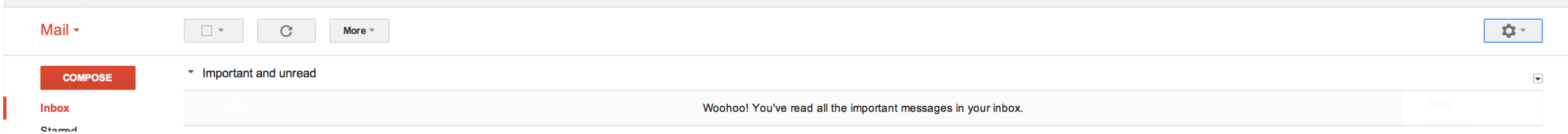
	continuous	*categorical*
	quantitative	qualitative

TYPES OF ML SOLUTIONS

	continuous	*categorical*
supervised	regression	classification
unsupervised	dimension reduction	clustering

What type of problem is this?

Priority Inbox

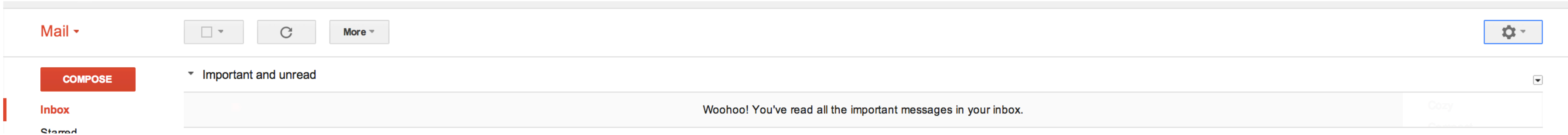


Woohoo! You've read all the important messages in your inbox.

TYPES OF ML SOLUTIONS

What type of problem is this?

Priority Inbox

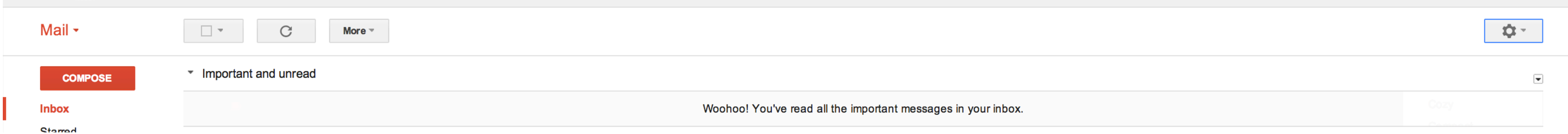


Important and unread

Woohoo! You've read all the important messages in your inbox.

Probably either.

TYPES OF ML SOLUTIONS

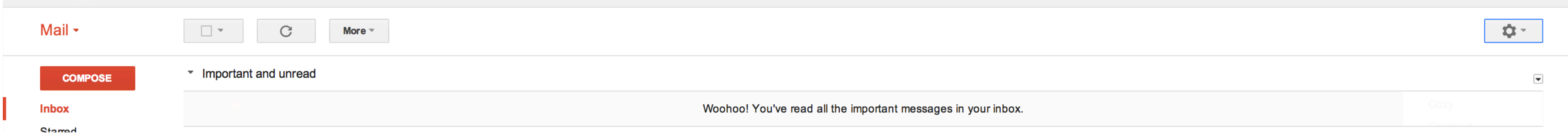


Woohoo! You've read all the important messages in your inbox.

Priority Inbox: Supervised Learning

Predict which mails users are most likely to star

TYPES OF ML SOLUTIONS



Important and unread

Woohoo! You've read all the important messages in your inbox.

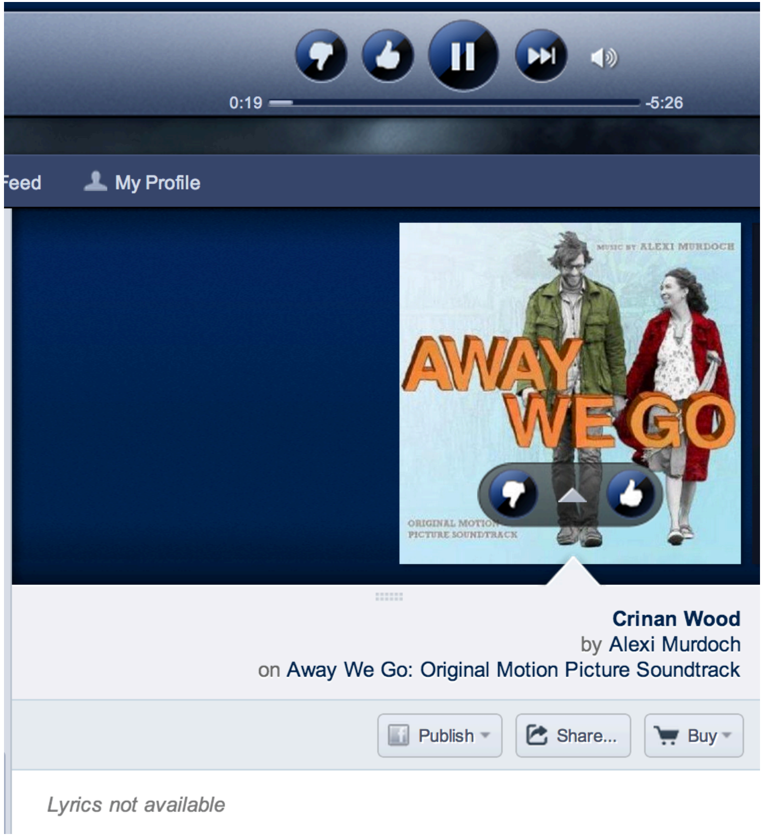
Priority Inbox: Unsupervised Learning

Group mails into groups and decide which group represents important mails

TYPES OF ML SOLUTIONS

What type of problem is this?

Music Recommendation



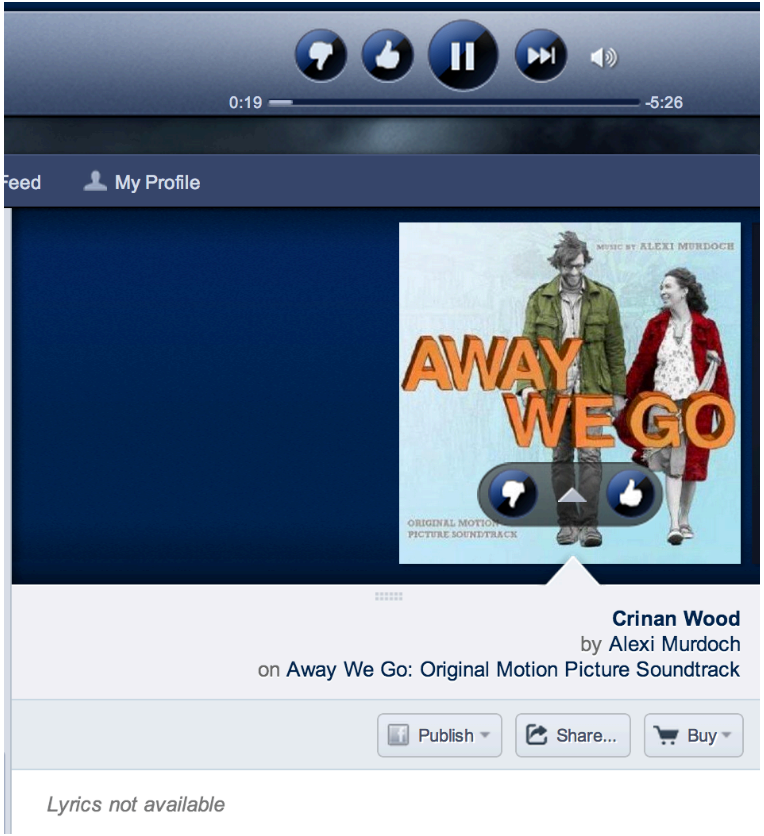
Crinan Wood
by Alexi Murdoch
on Away We Go: Original Motion Picture Soundtrack

Lyrics not available

What type of problem is this?

Music Recommendation

Probably either.



Crinan Wood
by Alexi Murdoch
on Away We Go: Original Motion Picture Soundtrack

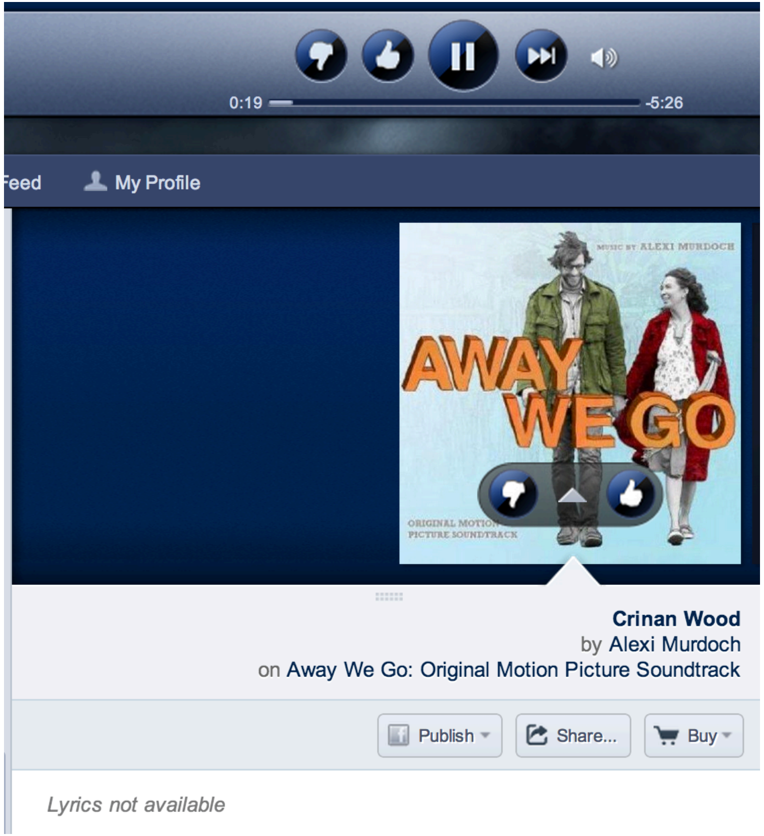
Lyrics not available

TYPES OF ML SOLUTIONS

What type of problem is this?

Music Recommendation as Supervised Learning

Predict which songs a user will ‘thumbs-up’



Crinan Wood
by Alexi Murdoch
on Away We Go: Original Motion Picture Soundtrack

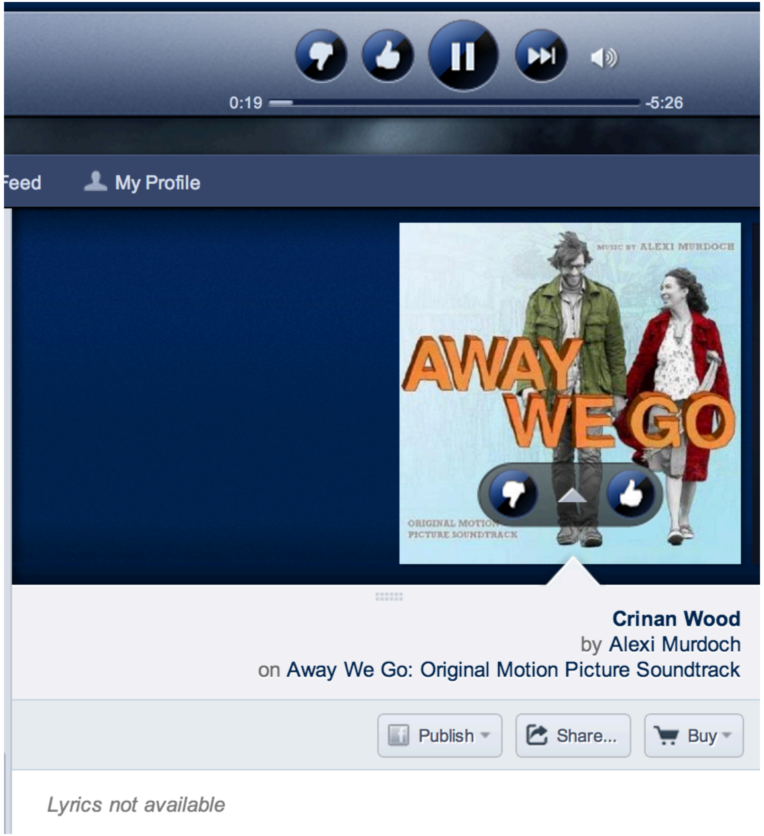
Lyrics not available

TYPES OF ML SOLUTIONS

What type of problem is this?

Music Recommendation
As Unsupervised Learning

Cluster songs based on attributes
and recommend songs in the same group



Crinan Wood
by Alexi Murdoch
on Away We Go: Original Motion Picture Soundtrack

Lyrics not available

QUESTION

HOW DO YOU DETERMINE THE RIGHT APPROACH?

APPROACHES TO ML PROBLEMS

	continuous	*categorical*
supervised	regression	classification
unsupervised	dimension reduction	clustering

ANSWER

The right approach is determined by the desired solution **and the data available.**

QUESTION

HOW DO YOU REPRESENT YOUR DATA?

TYPES OF DATA

	continuous	***categorical***
	quantitative	qualitative

	continuous	*categorical*
color	RGB-values	{red, blue}
ratings	1 – 10 rating	1-5 star rating

QUESTION

HOW DO YOU MEASURE OF QUALITY?

supervised	making predictions
unsupervised	extracting structure

ASSESSING ML PERFORMANCE

supervised	test out your predictions
unsupervised	...

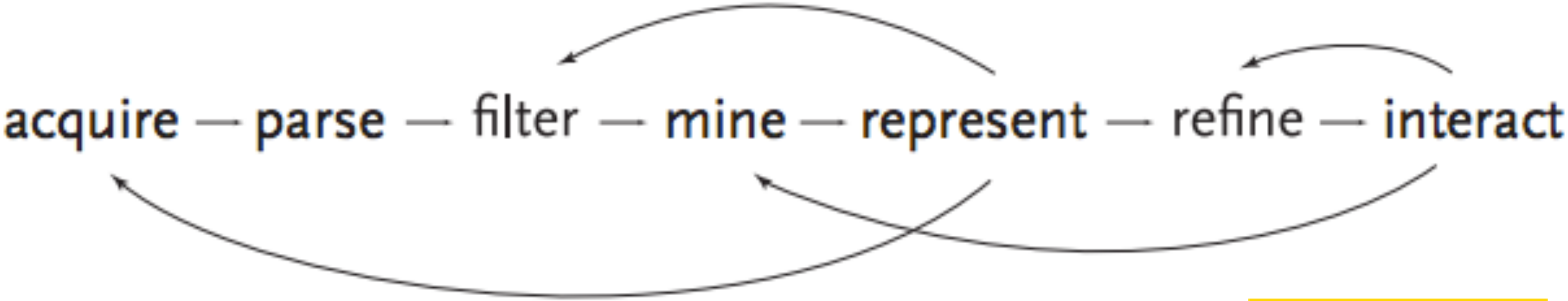
ASSESSING ML PERFORMANCE

supervised	test out your predictions

QUESTION

WHAT DO YOU DO WITH YOUR RESULTS?

THE DATA SCIENCE WORKFLOW



acquire — parse — filter — mine — represent — refine — interact

ANSWER

Interpret them and react accordingly.

source: http://benfry.com/phd/dissertation-110323c.pdf

III. SUPERVISED LEARNING

Q: How does a classification problem work?

A: Data in, predicted labels out.

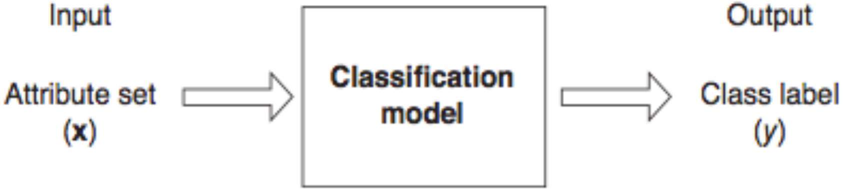
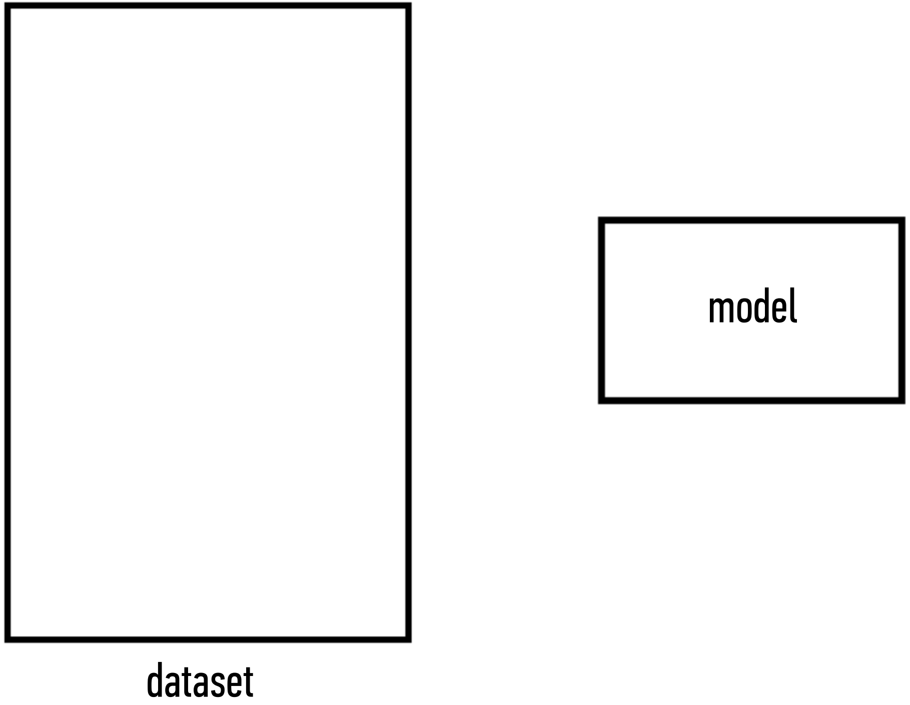


Figure 4.2. Classification as the task of mapping an input attribute set $\mathbf{x}$ into its class label y.

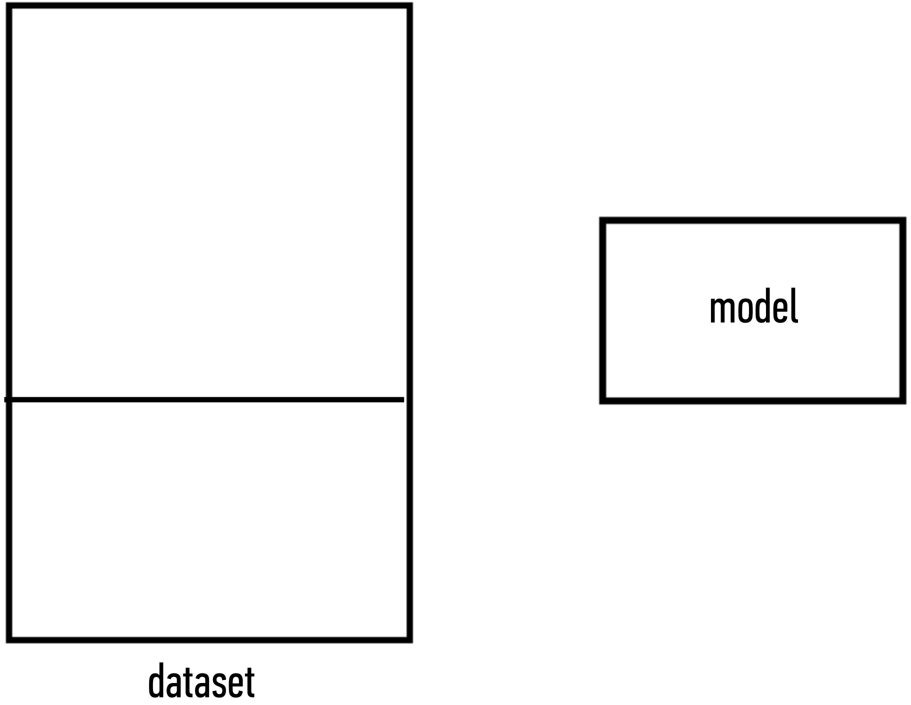
source: http://www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf

Q: What steps does a classification problem require?



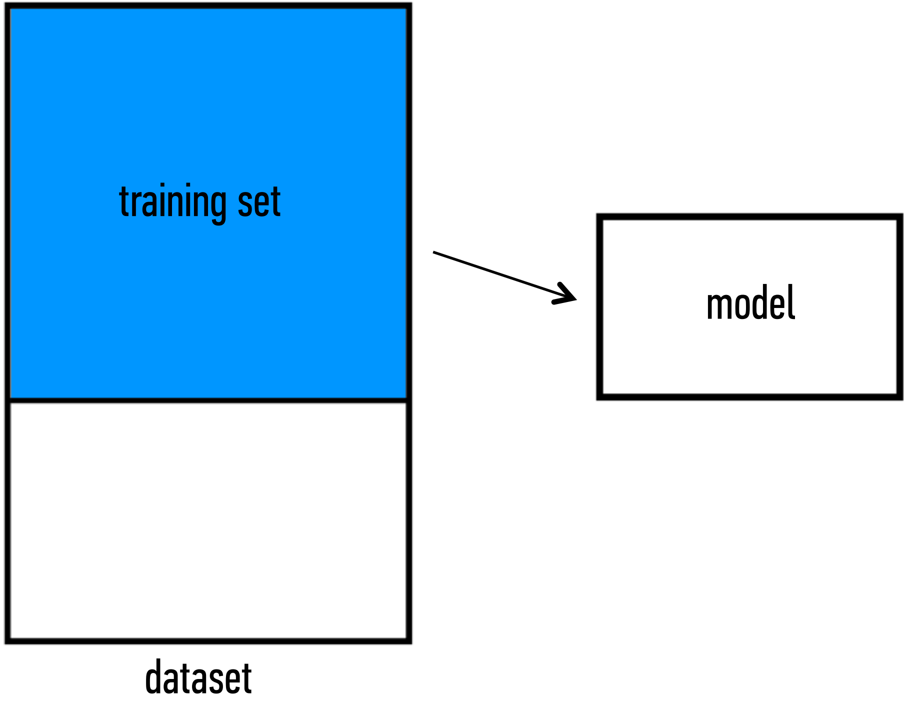
Q: What steps does a classification problem require?

1) split dataset



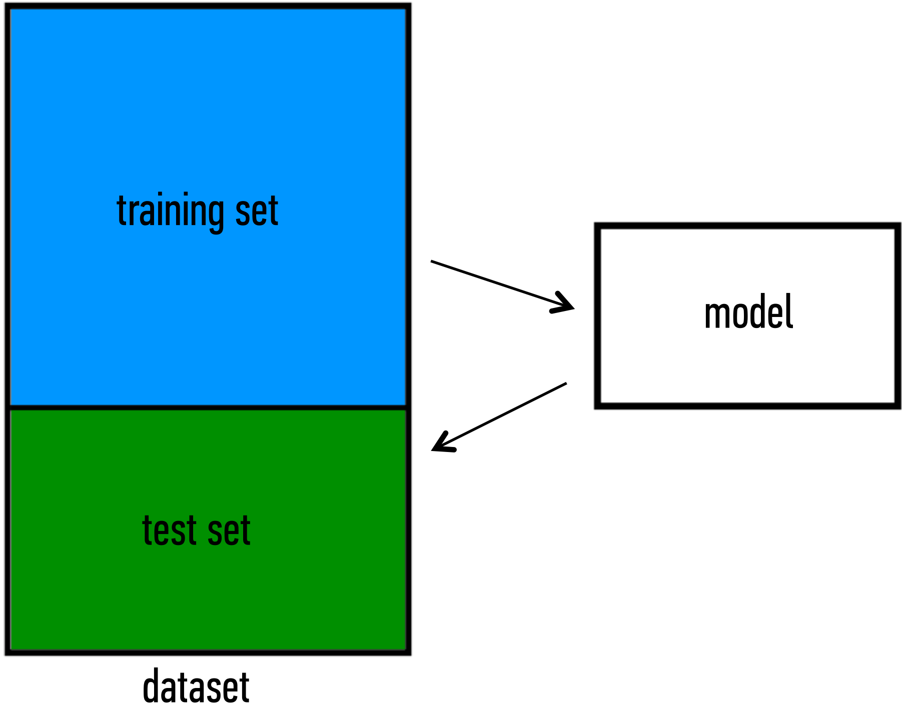
Q: What steps does a classification problem require?

1) split dataset
2) train model



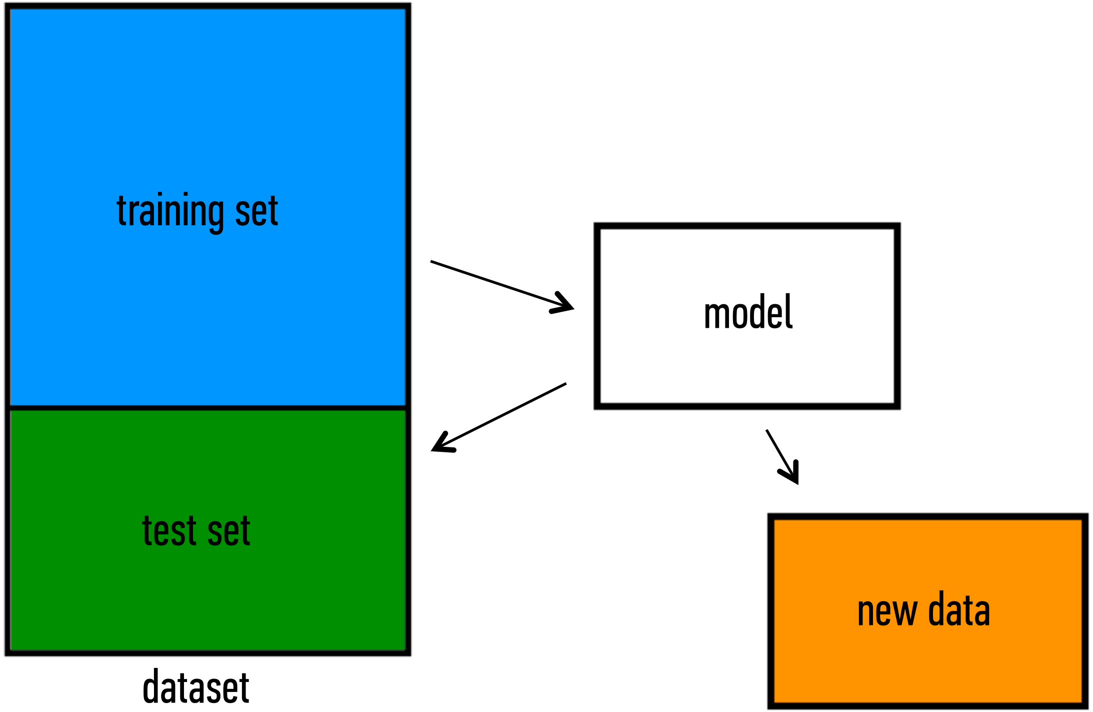
Q: What steps does a classification problem require?

1) split dataset
2) train model
3) test model



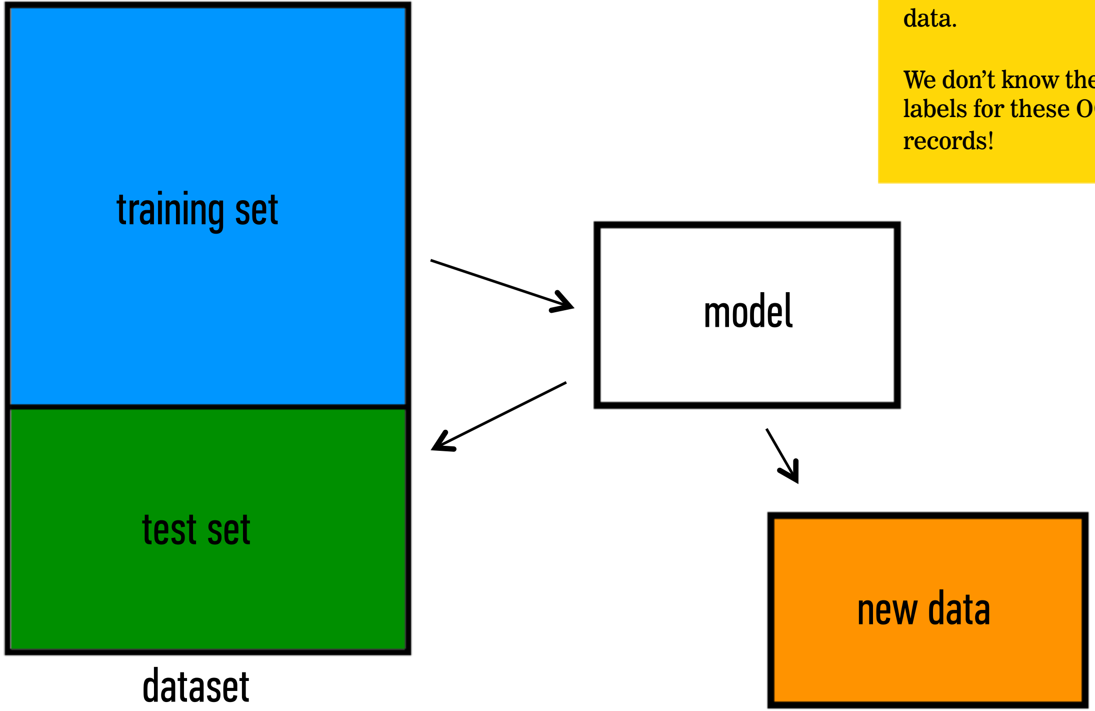
Q: What steps does a classification problem require?

1) split dataset
2) train model
3) test model
4) make predictions



Q: What steps does a classification problem require?

1) split dataset
2) train model
3) test model
4) make predictions



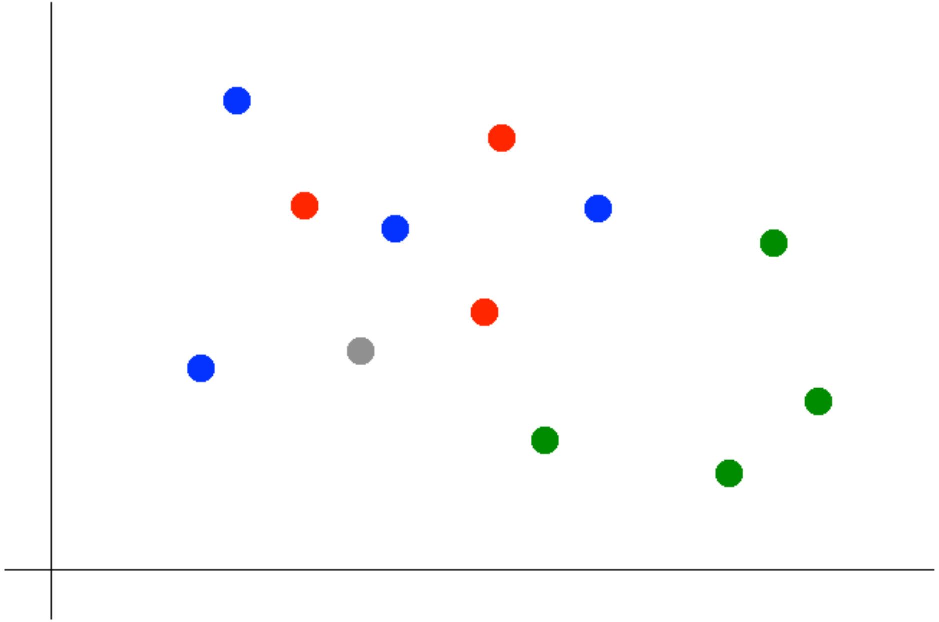
NOTE

This new data is called *out of sample* data.

We don't know the labels for these OOS records!

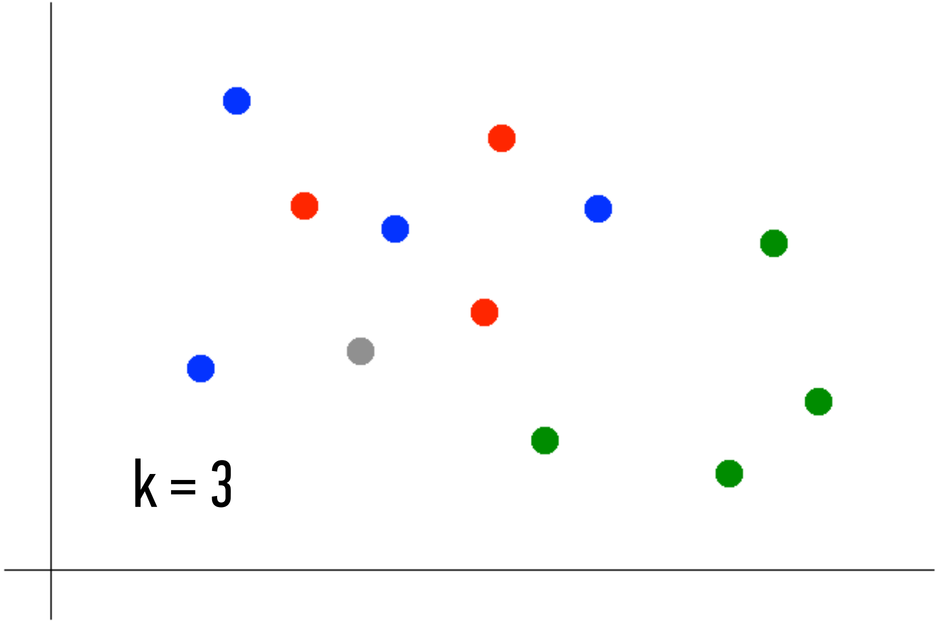
III. KNN CLASSIFICATION

Suppose we want to predict the color of the grey dot.



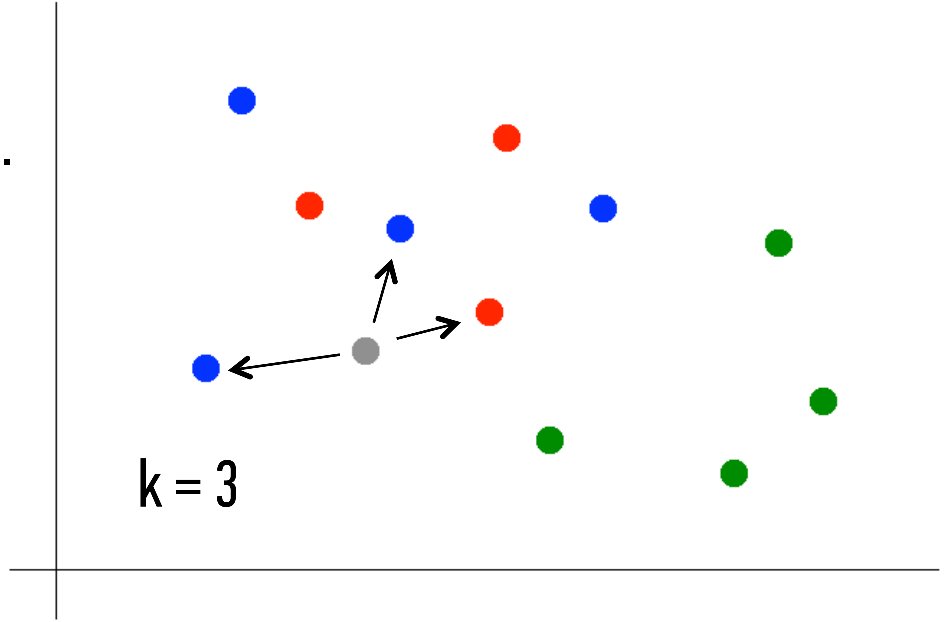
Suppose we want to predict the color of the grey dot.

1) Pick a value for k.



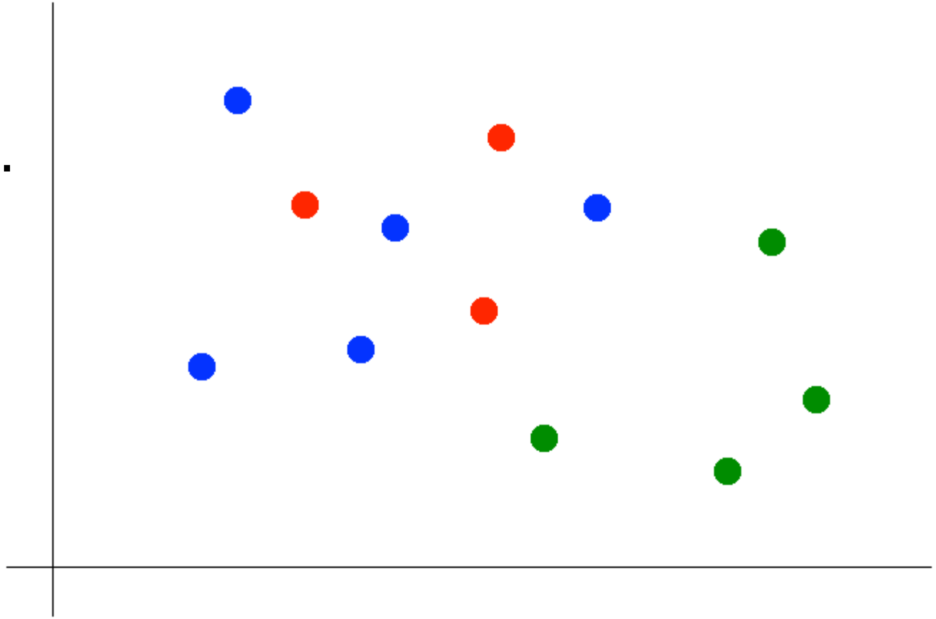
Suppose we want to predict the color of the grey dot.

1) Pick a value for k.
2) Find colors of k nearest neighbors.



Suppose we want to predict the color of the grey dot.

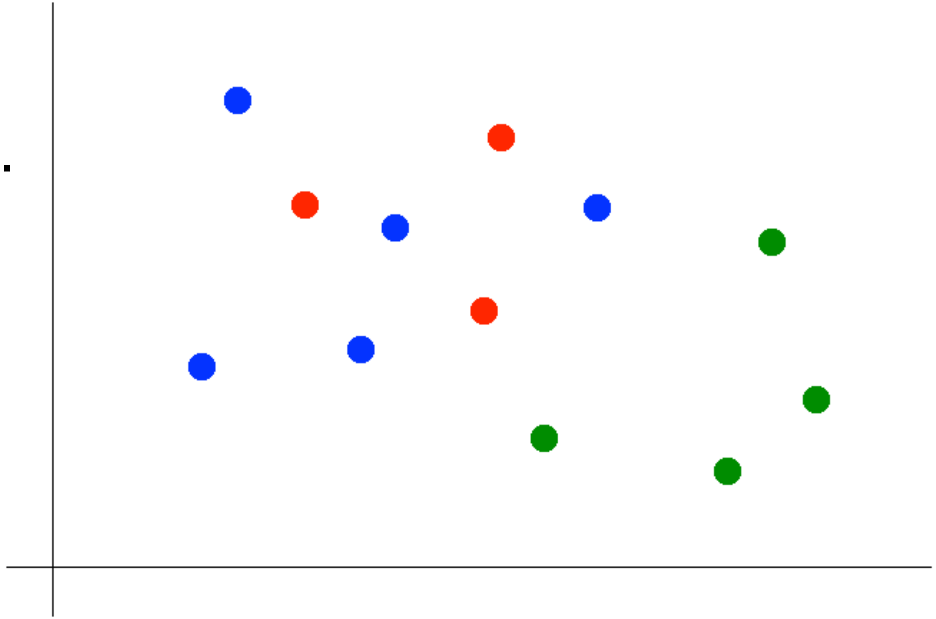
1) Pick a value for k.
2) Find colors of k nearest neighbors.
3) Assign the most common color to the grey dot.



Suppose we want to predict the color of the grey dot.

1) Pick a value for k.
2) Find colors of k nearest neighbors.
3) Assign the most common color to the grey dot.

Q: What does nearest mean?



DISCUSSION