



Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**



What is Data Science

Introduction to Data Science Algorithms

Jordan Boyd-Graber and Michael Paul

AUGUST 23, 2016

We will study algorithms that find and exploit patterns in data.

- These algorithms draw on ideas from statistics and computer science.
- Applications include
 - natural science (e.g., genomics, neuroscience)
 - web technology (e.g., Google, NetFlix)
 - finance (e.g., stock prediction)
 - policy (e.g., predicting what intervention X will do)
 - and many others

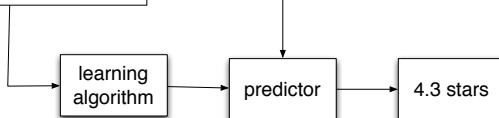
We will study algorithms that find and exploit patterns in data.

- Goal: fluency in thinking about modern data science problems.
- We will learn about a suite of tools in modern data analysis.
 - When to use them
 - The assumptions they make about data
 - Their capabilities, and their limitations
- We will learn a language and process for solving data analysis problems. On completing the course, you will be able to learn about a new tool, apply it data, and understand the meaning of the result.

Basic idea behind everything we will study

- 1 Collect or happen upon data.
- 2 Analyze it to find patterns.
- 3 Use those patterns to do something.

Rings (1952)	UR	Foreign	👍👍👍👍👍👍👍👍👍👍
Amélie (2005)	R	Independent	👍👍👍👍👍👍👍👍👍👍
Le Ciel sur l'épaule (1979)	R	Comedy	👍👍👍👍👍👍👍👍👍👍
The Life Aquatic with Steve Zissou (2004)	R	Comedy	👍👍👍👍👍👍👍👍👍👍
Lock, Stock and Two Smoking Barrels (1998)	R	Action & Adventure	👍👍👍👍👍👍👍👍👍👍
Lost in Translation (2003)	R	Drama	👍👍👍👍👍👍👍👍👍👍
Love and Death (1975)	PG	Comedy	👍👍👍👍👍👍👍👍👍👍
The Manchurian Candidate (1962)	PG-13	Classic	👍👍👍👍👍👍👍👍👍👍
Memento (2000)	R	Thriller	👍👍👍👍👍👍👍👍👍👍
Midnight Cowboy (1969)	R	Classic	👍👍👍👍👍👍👍👍👍👍

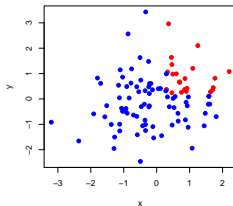


How the ideas are organized

Of course, there is no one way to organize such a broad subject. These concepts will recur through the course:

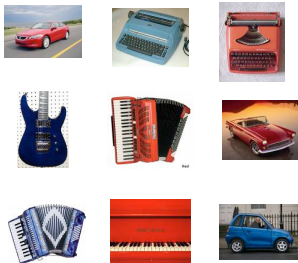
- Probabilistic foundations: distributions, approaches
- Statistical tests
- Supervised learning (more of this)
- Unsupervised learning (less of this)
- Methods that operate on discrete data (more of this)
- Methods that operate on continuous data (less of this)
- Representing data / feature engineering
- Evaluating models
- Understanding the assumptions behind the methods

Supervised vs. unsupervised methods



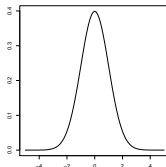
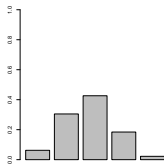
- **Supervised methods** find patterns in **fully observed** data and then try to predict something from **partially observed** data.
- For example, we might observe a collection of emails that are categorized into *spam* and *not spam*.
- After learning something about them, we want to take new email and automatically categorize it.

Supervised vs. unsupervised methods



- **Unsupervised methods** find **hidden structure** in data, structure that we can never formally observe.
- E.g., a museum has images of their collection that they want grouped by similarity into 15 groups.
- Unsupervised learning is more difficult to evaluate than supervised learning. But, these kinds of methods are widely used.

Discrete vs. continuous methods



- Discrete methods manipulate a finite set of objects
 - e.g., classification into one of 5 categories.
- Continuous methods manipulate continuous values
 - e.g., prediction of the change of a stock price.

One useful grouping

	<i>discrete</i>	<i>continuous</i>
<i>supervised</i>	classification	regression
<i>unsupervised</i>	clustering	dimensionality reduction

One useful grouping

	<i>discrete</i>	<i>continuous</i>
<i>supervised</i>	classification	regression
<i>unsupervised</i>	clustering	dimensionality reduction

Classification

logistic regression, SVM

One useful grouping

	<i>discrete</i>	<i>continuous</i>
<i>supervised</i>	classification	regression
<i>unsupervised</i>	clustering	dimensionality reduction

Clustering

k-means

One useful grouping

	<i>discrete</i>	<i>continuous</i>
<i>supervised</i>	classification	regression
<i>unsupervised</i>	clustering	dimensionality reduction

Regression

Linear Regression

One useful grouping

	<i>discrete</i>	<i>continuous</i>
<i>supervised</i>	classification	regression
<i>unsupervised</i>	clustering	dimensionality reduction

Dimensionality Reduction

...

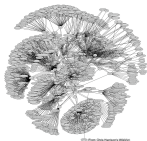
Data representation (feature engineering)



→ $\langle 1.5, 3.2, -5.1, \dots, 4.2 \rangle$

Republican nominee
George Bush said he felt
nervous as he voted
today in his adopted
home state of Texas,
where he ended...

→ $\langle 1, 0, 0, 0, 5, 0, 9, 3, 1, \dots, 0 \rangle$



→
$$\begin{bmatrix} 1 & 0 & 1 & \dots & 0 \\ 0 & 1 & 1 & \dots & 0 \\ 1 & 0 & 0 & \dots & 1 \\ \dots & & & & \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

Understanding assumptions



- The methods we'll study make **assumptions** about the data on which they are applied. E.g.,
 - Documents can be analyzed as a sequence of words;
 - or, as a “bag” of words.
 - Independent of each other;
 - or, as connected to each other
- What are the assumptions behind the methods?
- When/why are they appropriate?
- Much of this is an art