College of Media, Communication and Information
UNIVERSITY OF COLORADO BOULDER

What is Data Science

INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

JANUARY 23, 2017

This Course (Data Science)

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

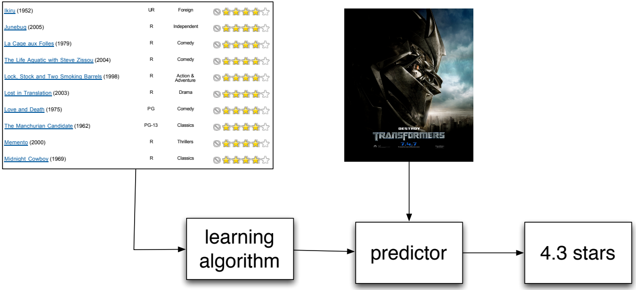
This Course (Data Science)

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 of 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.**



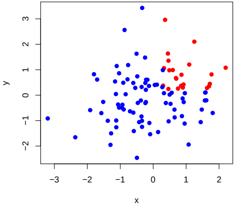
Ikiru (1952)	UR	Foreign
Junebug (2005)	R	Independent
La Cage aux Folles (1979)	R	Comedy
The Life Aquatic with Steve Zissou (2004)	R	Comedy
Lock, Stock and Two Smoking Barrels (1998)	R	Action & Adventure
Lost in Translation (2003)	R	Drama
Love and Death (1975)	PG	Comedy
The Manchurian Candidate (1962)	PG-13	Classics
Memento (2000)	R	Thrillers
Midnight Cowboy (1969)	R	Classics

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:

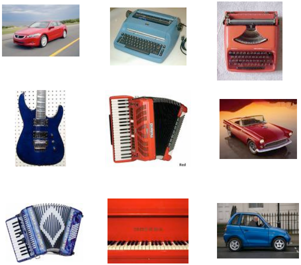
- Probabilistic foundations: distributions, approaches
- Statistical tests
- Supervised learning (more of this)
- Unsupervised learning (less of this)
- Methods that operate on discrete data (more of this)
- Methods that operate on continuous data (less of this)
- Representing data / feature engineering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

Supervised vs. unsupervised methods



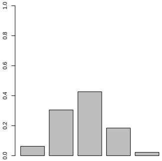
- **Supervised methods** find patterns in **fully observed** data and then try to predict something from **partially observed** data.
- For example, we might observe a collection of emails that are categorized into *spam* and *not spam*.
- After learning something about them, we want to take new email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

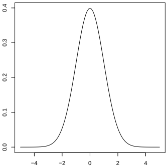
Supervised vs. unsupervised methods



- **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

	discrete	*continuous*
supervised	**classification**	**regression**
unsupervised	**clustering**	**dimensionality reduction**

One useful grouping

	discrete	*continuous*
supervised	**classification**	**regression**
unsupervised	**clustering**	**dimensionality reduction**

Classification

logistic regression, SVM

One useful grouping

	discrete	*continuous*
supervised	**classification**	**regression**
unsupervised	**clustering**	**dimensionality reduction**

Clustering

k-means

One useful grouping

	discrete	*continuous*
supervised	**classification**	**regression**
unsupervised	**clustering**	**dimensionality reduction**

Regression

Linear Regression

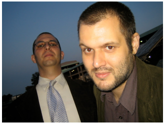
One useful grouping

	discrete	*continuous*
supervised	**classification**	**regression**
unsupervised	**clustering**	**dimensionality reduction**

Dimensionality Reduction

...

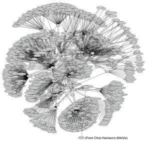
Data representation (feature engineering)



$\longrightarrow \langle 1.5, 3.2, -5.1, \ldots, 4.2 \rangle$

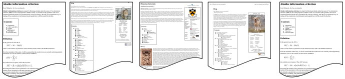
Republican nominee George Bush said he felt nervous as he voted today in his adopted home state of Texas, where he ended...

$\longrightarrow \langle 1, 0, 0, 0, 5, 0, 9, 3, 1, \ldots, 0 \rangle$



$$\longrightarrow \begin{bmatrix} 1 & 0 & 1 & \ldots & 0 \\ 0 & 1 & 1 & \ldots & 0 \\ 1 & 0 & 0 & \ldots & 1 \\ & \ldots & & & \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix}$$

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