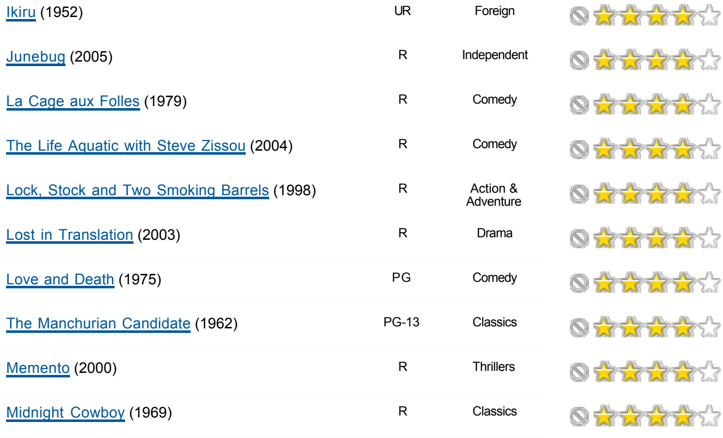
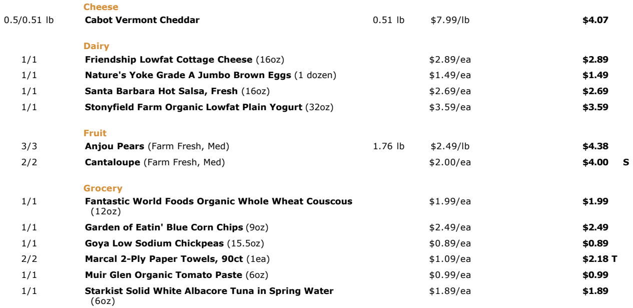
Data are everywhere.

User ratings



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	⊘ ★★★★☆

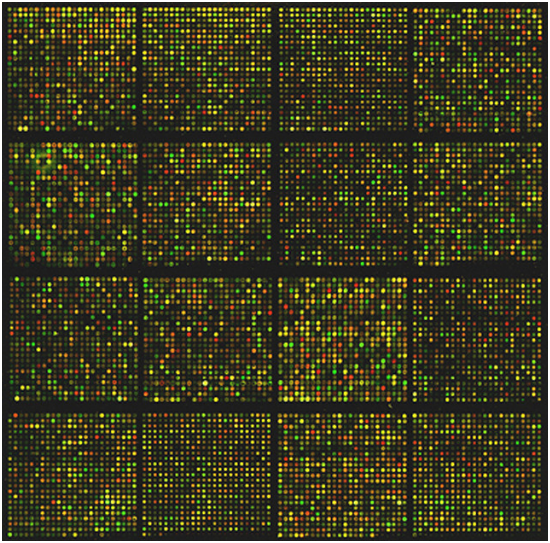
Purchase histories



	Cheese				
0.5/0.51 lb	**Cabot Vermont Cheddar**	0.51 lb	$7.99/lb	**$4.07**	
	Dairy				
1/1	**Friendship Lowfat Cottage Cheese** (16oz)		$2.89/ea	**$2.89**	
1/1	**Nature's Yoke Grade A Jumbo Brown Eggs** (1 dozen)		$1.49/ea	**$1.49**	
1/1	**Santa Barbara Hot Salsa, Fresh** (16oz)		$2.69/ea	**$2.69**	
1/1	**Stonyfield Farm Organic Lowfat Plain Yogurt** (32oz)		$3.59/ea	**$3.59**	
	Fruit				
3/3	**Anjou Pears** (Farm Fresh, Med)	1.76 lb	$2.49/lb	**$4.38**	
2/2	**Cantaloupe** (Farm Fresh, Med)		$2.00/ea	**$4.00**	**S**
	Grocery				
1/1	**Fantastic World Foods Organic Whole Wheat Couscous** (12oz)		$1.99/ea	**$1.99**	
1/1	**Garden of Eatin' Blue Corn Chips** (9oz)		$2.49/ea	**$2.49**	
1/1	**Goya Low Sodium Chickpeas** (15.5oz)		$0.89/ea	**$0.89**	
2/2	**Marcal 2-Ply Paper Towels, 90ct** (1ea)		$1.09/ea	**$2.18**	**T**
1/1	**Muir Glen Organic Tomato Paste** (6oz)		$0.99/ea	**$0.99**	
1/1	**Starkist Solid White Albacore Tuna in Spring Water** (6oz)		$1.89/ea	**$1.89**	

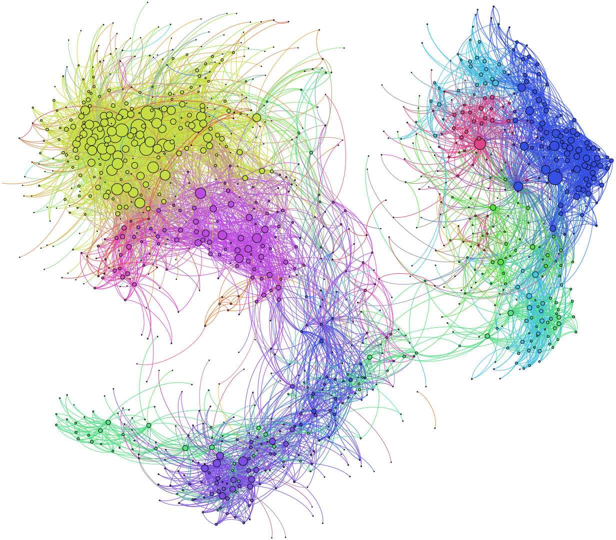
Document collections

Genomics



Neuroscience

Social networks



Data can help us solve problems.

Will NetFlix user 493234 like Transformers?

Will NetFlix user 493234 like Transformers?

DESTROY
TRANSFORMERS
7.4.7

★★☆☆☆

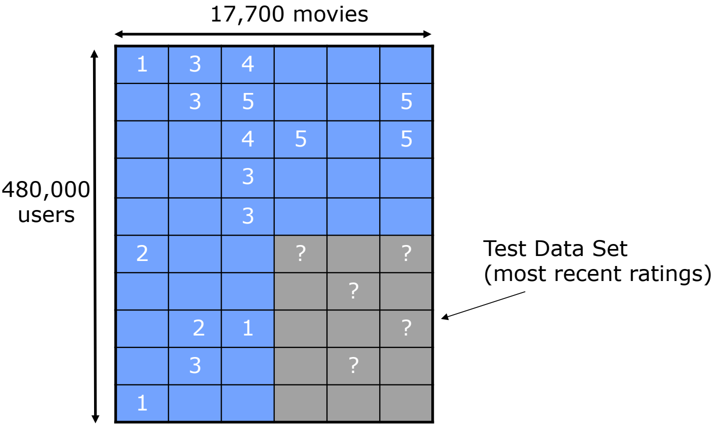
THE STAR WARS SAGA CONTINUES
THE EMPIRE STRIKES BACK
MARK HAMILL · HARRISON FORD · CARRIE FISHER
BILLY DEE WILLIAMS · ANTHONY DANIELS

★★★★★

ALIEN
In space no one can hear you scream.

★★★★★

How do you know?



17,700 movies

480,000 users

Test Data Set (most recent ratings)

1	3	4			
	3	5			5
		4	5		5
		3			
		3			
2			?		?
				?	
	2	1			?
	3			?	
1					

Group these images into 3 groups

Group many images and determine the number of groups

Rank these images...

- ...according to relevance to instrument.
- ...according to relevance to machine

Is this spam?

Subject: CHARITY.
Date: February 4, 2008 10:22:25 AM EST
To: undisclosed-recipients:;
Reply-To: s.polla@yahoo.fr

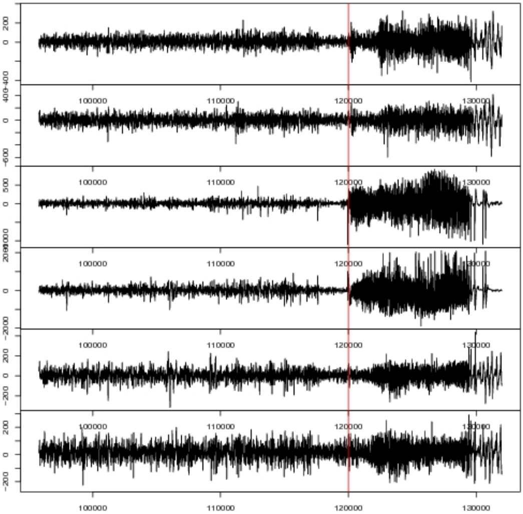
Dear Beloved,
My name is Mrs. Susan Polla, from ITALY. If you are a christian and interested in charity please reply me at : (s.polla@yahoo.fr) for insight.
Respectfully,
Mrs Susan Polla.

How about this one?

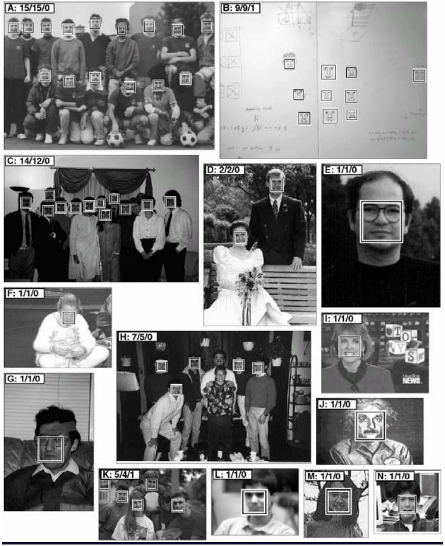
From: [snipped]
Subject: Superbowl?
Date: January 30, 2008 8:09:00 PM EST
To: jbg@cs.princeton.edu, [snipped]

Anyone interested in coming by to watch the game? Beer and pizza, I'd imagine. If anyone wants, we could get together earlier, play a board game or cards or roll up characters or something. Takers?

When did the seizure begin?



Where are the faces?



Data contain patterns
that can help us solve problems.

This Course (Digging into Data)

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

- These algorithms draw on ideas from statistics and machine learning.
- 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

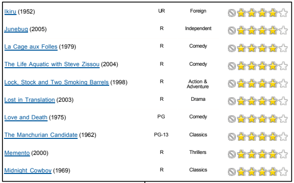
This Course (Digging into Data)

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

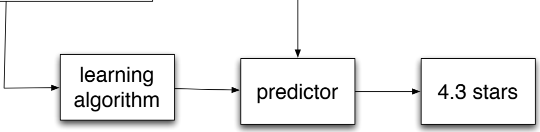
- Goal: fluency in thinking about modern data analysis problems.
- We will learn about a suite of tools in modern data analysis.
  - When to use them
  - The assumptions they make about data
  - Their capabilities, and their limitations
- We will learn a language and process for of solving data analysis 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	⊘★★★★☆

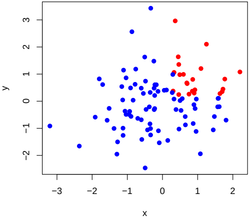


How the ideas are organized

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

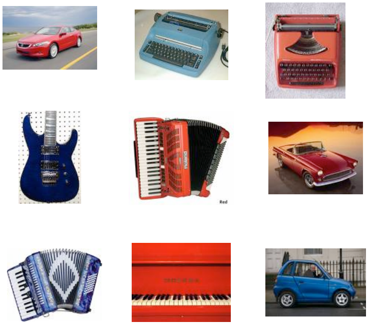
- Supervised learning
- Unsupervised learning
- Methods that operate on discrete data
- Methods that operate on continuous data
- Representing data
- Understanding the assumptions behind the methods

Supervised vs. unsupervised methods



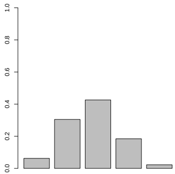
- **Supervised methods** find patterns in **fully observed** data and then try to predict something from **partially observed** data.
- For example, we might observe a collection of emails that are categorized into *spam* and *not spam*.
- After learning something about them, we want to take new email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

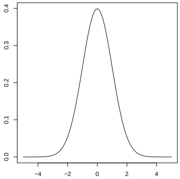
Supervised vs. unsupervised methods



- **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



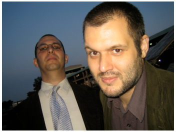


- Discrete methods manipulate a finite set of objects
  - e.g., classification into one of 5 categories.
- Continuous methods manipulate continuous values
  - e.g.,prediction of the change of a stock price.

One useful grouping

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

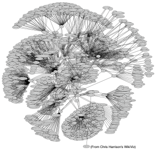
Data representation



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

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

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



(From Chris Harrison's WikiViz)

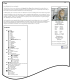
$$\longrightarrow \begin{bmatrix} 1 & 0 & 1 & \ldots & 0 \\ 0 & 1 & 1 & \ldots & 0 \\ 1 & 0 & 0 & \ldots & 1 \\ & \ldots & & & \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix}$$

Understanding assumptions

Akaike information criterion

Contents

Definition


Cat

Contents

Princeton University

Contents

History

Dog

Contents

Akaike information criterion

Contents

Definition

- 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?