Multi- and Social Media

LBSC 690: Jordan Boyd-Graber

University of Maryland

November 14, 2011

COLLEGE OF INFORMATION STUDIES

Adapted from Jimmy Lin's Slides

Take-Away Messages

- Review Assignment 3
- Human senses are gullible
  - Images, video, and audio are all about trickery
- Compression: storing a lot of information in a little space
  - So that it fits on your hard drive
  - So that you can send it quickly across the network
- Data mining: How you can get money / utility out of big data

Outline

Outline

Question 1.1

```
select * from authors where year_born % 100 = 82;
select * from authors where year_born like '%%82';
select * from authors where (year_born - 1982) % 100 = 0;
```

- Doesn't matter how as long as it's a single query

Question 1.2

```
select * from authors where year_born % 100 = 82 and year_died < 1900;
select * from authors where year_born % 100 = 82 and year_died between
'0000' and '1900';
```

Question 2.1

```
select count (*) from categories;
```

Question 2.2

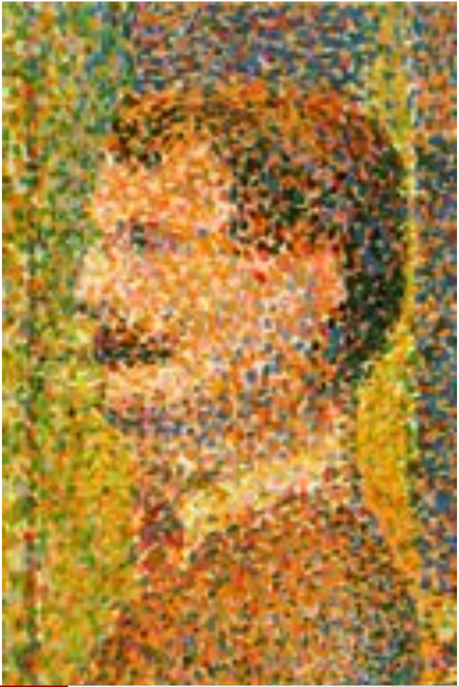
```
select count (*) as num_books, category_name from category_map INNER JOIN categories on categories.categoryID=category_map.categoryID group by category_name order by num_books desc limit 20;
```

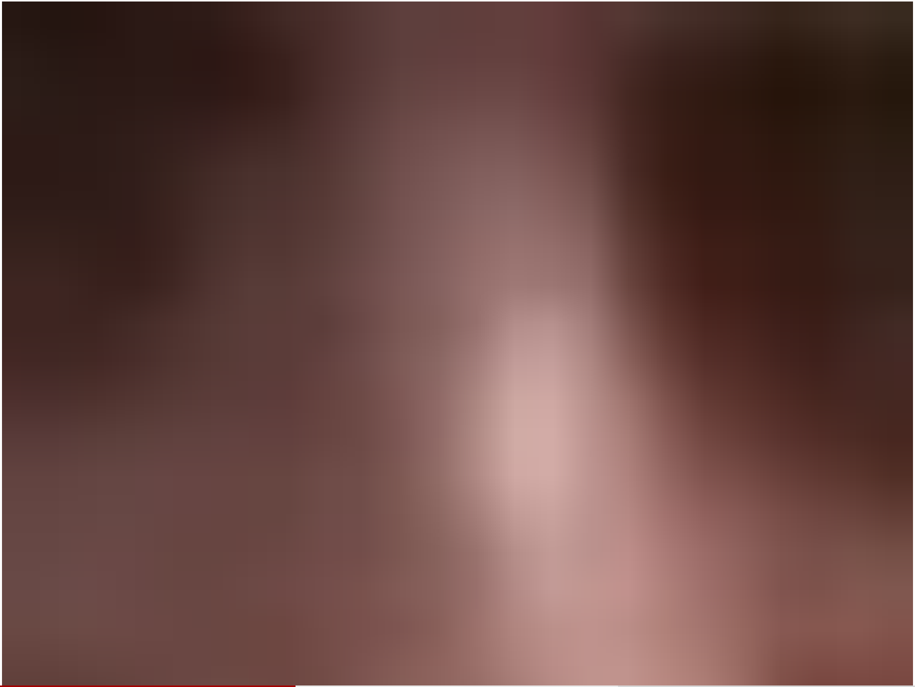
Question 2.3

```
select title from category_map INNER JOIN books INNER JOIN categories
on categories.categoryID=category_map.categoryID AND
category_map.bookID=books.bookID where category_name="History";
```

- Can also restrict based on categoryID
- People got tripped up by not specifying which categoryID is meant

Outline





What's in an image?

- **pixel** A dot of a single color in an image
- **resolution** How many dots are in in an image (e.g. 100 dpi = 100 dots per inch)

What's in an image?

- **pixel** A dot of a single color in an image
- **resolution** How many dots are in in an image (e.g. 100 dpi = 100 dots per inch)
- Once you get to the fundamental resolution, you can't go any further

Color

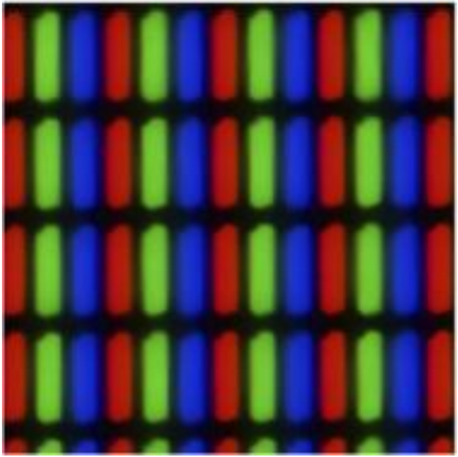
Color Wheel

Color

HTML Color

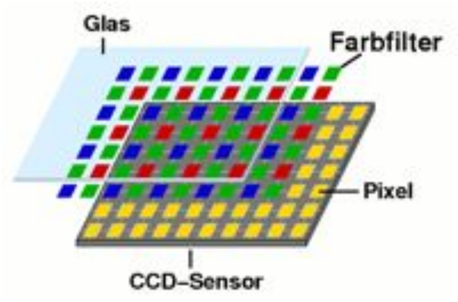
#99FF66
#9999FF

Color



LCD

Color



Camera

Is a picture worth a thousand words?

Megapixels

$2,048x1,536 = 3,145,7283$ MP
$2,560x1,920 = 4,915,2005$ MP
$3,264x2,448 = 7,990,2728$ MP
$3,648x2,736 = 9,980,92810$ MP

- Each pixel has at least one byte
- How many words would it take to match an image with 1024 x 768 resolution?

Compression

- Goal: represent the same information using fewer bits
- Two basic types of data compression:
  - Lossless: can reconstruct exactly
  - Lossy: cant reconstruct, but looks the same
- Two basic strategies:
  - Reduce redundancy
  - Throw away stuff that doesnt matter

Run-Length Encoding

- Opportunity:
  - Large regions of a single color are common

  Approach:
  - Record # of consecutive pixels for each color
- An example with text:

Run length

Sheep go baaaaaaaaaa and cows go mooooooooooo
→ Sheep go ba<10> and cows go mo<10>

Using Dictionaries

- Opportunity:
  - Data often have shared substructure, e.g., patterns
- Approach:
  - Create a dictionary of commonly seen patterns
  - Replace patterns with shorthand code
- An example with text:

Dictionary

The rain in Spain falls mainly on the plain
$\rightarrow$ The r& % sp& falls m&ly on the & (& = ain, %=in)

Palette Selection

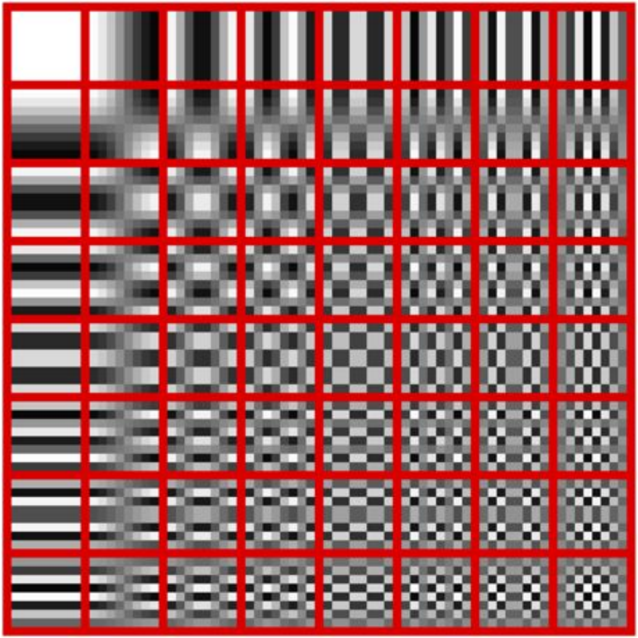
- Opportunity: No picture uses all 16 million colors
- Approach:
  - Select a palette of 256 colors
  - Indicate which palette entry to use for each pixel
  - Look up each color in the palette
- What happens if there are more than 256 colors?

GIF and PNG

- Both use limited pallette
- GIF = Graphics Interchange Format
  - Popularized by CompuServe
  - GIF uses Lempel Ziv Welch compression
  - Unisys owned patent (now expired)
- PNG = Portable Network Graphics
  - Protest against patent
  - Handles some technical details better
- GIF can be transparent
- PNG can be alpha transparent (e.g. ghosting)
- GIF can be animated

Discrete Cosine Transform

- Opportunity:
  - Images can be approximated by a series of patterns
  - Complex patterns require more information than simple patterns
  - Humans can't tell the difference between high-frequency patterns
- Approach:
  - Break an image into little blocks (8 x 8)
  - Represent each block in terms of basis images
- This is JPEG = Joint Photographics Expert Group




J.P.G.
Full quality (Q = 100): 83,261 bytes
Medium quality (Q = 25): 9,553 bytes
Average quality (Q = 50): 15,138 bytes
Low quality (Q = 10): 4,787 btes

Outline

Raster vs. Vector Graphics

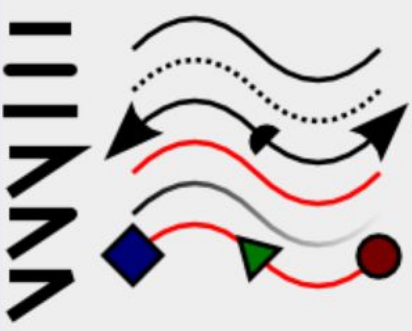
- Raster images = bitmaps
  - Actually describe the contents of the image
- Vector images = composed of mathematical curves
  - Describe how to draw the image
  - Example: SVG (Scalable Vector Graphics)

SVG Example

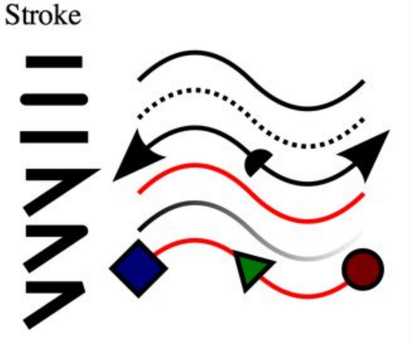
```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE svg PUBLIC "-//W3C//DTD SVG 1.1//EN"
"http://www.w3.org/Graphics/SVG/1.1/DTD/svg11.dtd">

<svg width="100\%" height="100\%" version="1.1"
xmlns="http://www.w3.org/2000/svg">

<path d="M153 334
C153 334 151 334 151 334
C151 339 153 344 156 344
C164 344 171 339 171 334
C171 322 164 314 156 314
C142 314 131 322 131 334
C131 350 142 364 156 364
C175 364 191 350 191 334
C191 311 175 294 156 294
C131 294 111 311 111 334
C111 361 131 384 156 384
C186 384 211 361 211 334
C211 300 186 274 156 274"
style="fill:white;stroke:red;stroke-width:2"/>

</svg>
```

SVG

Raster



Vector



Defining SVG Shapes

- Can **m**ove, **c**ontinue, return (**z**), **s**mooth path to, etc.
- Like describing the moves of a pen
- Can also describe how to color shapes, transparency, etc.

Pitfalls

- One of the most common pitfalls is going form vector to raster
- Don't do it if you can help it
- If you must, make sure you have a very high resolution

Outline

Basic Video Coding

- Display a sequence of images ...
  - Fast enough to trick your eyes (At least 30 frames per second)
- NTSC Video
  - 60 "interlaced" half-frames/sec, 720x486
- PAL Video
  - 25 frames/sec, 720x576
- HDTV
  - 30 "progressive" full-frames/sec, 1280x720
  - Or higher (depends on whom you ask)

To Interlace or Not?

- Interlacing: update all even lines of pixels, update all odd lines of pixels
  - Human vision is persistent
  - TVs used to work by shooting electrons at glass screen
  - Phosphor glow takes a while to wear off
- Progressive
  - Update all lines at once
  - Consistent with how LCDs work
  - Displaying on LCDs can cause problems with deinterlacing (right)

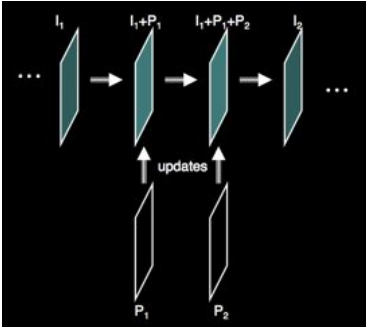
Video Example

- Typical low-quality video:
  - 640 x 480 pixel image
  - 3 bytes per pixel (red, green, blue)
  - 30 frames per second
- Storage requirements:
  - 26.4 MB/second!
  - A CD-ROM would hold 25 seconds
  - 30 minutes would require 46.3 GB
- Some form of compression required!

Video Compression

- Opportunity: One frame looks very much like the next
- Approach: Record only the pixels that change

Frame Reconstruction



- I frames provide complete image
- P frames provide series of updates to most recent I frame

Embedding a video in a webpage

Option 1: Embed

```
<embed src="intro.swf" height="200" width="200"/>
```

- The “embed” tag is unknown to HTML 4. Your page will not validate correctly.
- If the browser does not support Flash, your video will not play.
- iPad and iPhone cannot display Flash videos.
- If you convert the video to another format, it will still not play in all browsers.

Embedding a video in a webpage

Option 2: Object

```
<object data="intro.swf" height="200" width="200"/>
```

- If the browser does not support Flash, your video will not play.
- iPad and iPhone cannot display Flash videos.
- If you convert the video to another format, it will still not play in all browsers.

Embedding a video in a webpage

Option 3: Video

```
<video width="320" height="240" controls="controls">
  <source src="movie.mp4" type="video/mp4" />
  <source src="movie.ogg" type="video/ogg" />
  <source src="movie.webm" type="video/webm" />
Your browser does not support the video tag.
</video>
```

- You must convert your videos to many different formats.
- The “video” element does not work in older browsers.
- The “video” element does not validate in HTML 4 and XHTML.

Embedding a video in a webpage

Option 4: All of the above

```
<video width="320" height="240" controls="controls">
  <source src="movie.mp4" type="video/mp4" />
  <source src="movie.ogg" type="video/ogg" />
  <source src="movie.webm" type="video/webm" />
<object data="movie.mp4" width="320" height="240">
<embed src="movie.swf" width="320" height="240">
 Your browser does not support video
</object>
</video>
```

- You must convert your videos to many different formats
- The video / embed element does not validate in HTML 4 and XHTML (use DOCTYPE)
- Falls back to object, then embed (older browsers)

Embedding a video in a webpage

Option 5: Outsource

Outline

Sound

- What's sound
- How does hearing work
- How does a speaker work
- How does a microphone work

Sound

- What's sound
- How does hearing work
- How does a speaker work
- How does a microphone work
- All vibrations and waves!

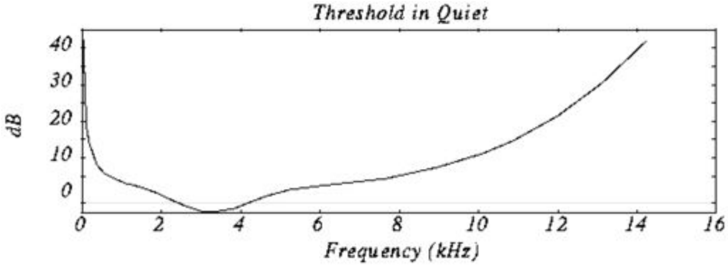
Basic Audio Coding

- Sample at twice the highest frequency (8 bits or 16 bits per sample)
- Speech (0-4 kHz) requires 8 KB/s: Standard telephone channel (8-bit samples)
- Music (0-22 kHz) requires 172 KB/s: Standard for CD-quality audio (16 bit samples)

How do MP3s work?

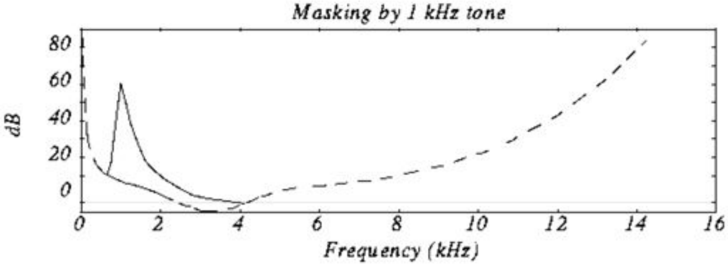
- Opportunity: The human ear cannot hear all frequencies at once, all the time
- Approach: Dont represent things that the human ear cannot hear
- Aside: **Encoding** MP3s requires licensing a patent; unlike GIF, people put up with this (But there are alternatives like OGG Vobis)

Human Hearing Response



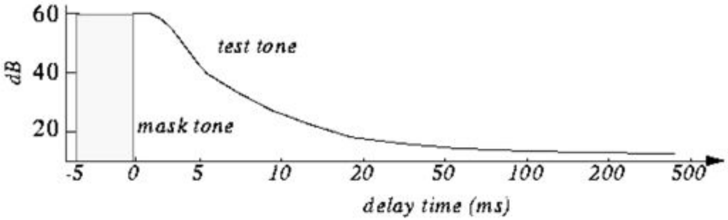
Experiment: Put a person in a quiet room. Raise level of tone at a given frequency until just barely audible. Vary the frequency and plot the results.

Frequency Masking



Experiment: Play 1kHz tone (masking tone) at fixed level (60 db). Play test tone at a different level and raise level until just distinguishable. Vary the frequency of the test tone and plot the threshold when it becomes audible.

Temporal Masking



If we hear a loud sound, then it stops, it takes a while until we can hear a soft tone at about the same frequency.

MP3s: Psychoacoustic compression

- Eliminate sounds below threshold of hearing
- Eliminate sounds that are frequency masked
- Eliminate sounds that are temporally masked
- Eliminate stereo information for low frequencies

Streaming Audio and Video

- Simultaneously:
  - Receive downloaded content in buffer
  - Play current content of buffer
  - Analogy: filling and draining a basin concurrently

Outline

Data are everywhere.

User ratings

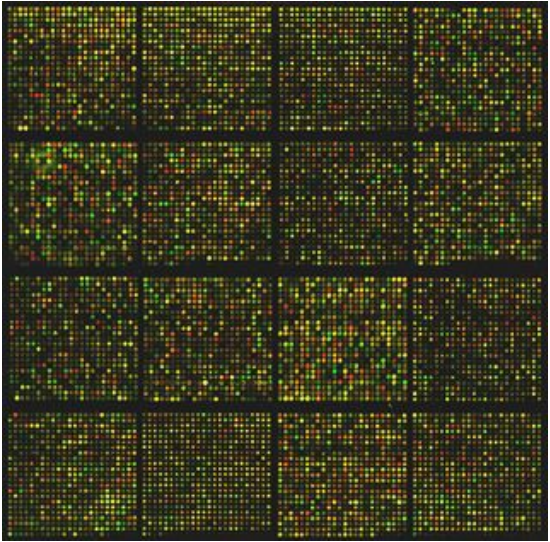
Ikiru (1952)	UR	Foreign
Junebug (2005)	R	Independent
La Cage aux Folles (1979)	R	Comedy
The Life Aquatic with Steve Zissou (2004)	R	Comedy
Lock, Stock and Two Smoking Barrels (1998)	R	Action & Adventure
Lost in Translation (2003)	R	Drama
Love and Death (1975)	PG	Comedy
The Manchurian Candidate (1962)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

Purchase histories

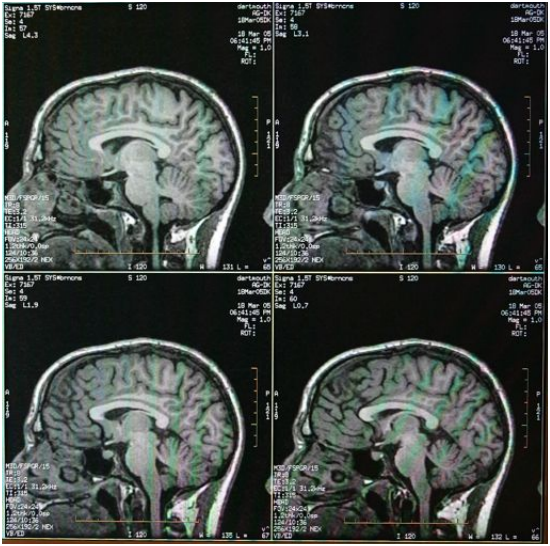
	Cheese				
0.5/0.51 lb	**Cabot Vermont Cheddar**	0.51 lb	$7.99/lb	**$4.07**	
	Dairy				
1/1	**Friendship Lowfat Cottage Cheese** (16oz)		$2.89/ea	**$2.89**	
1/1	**Nature's Yoke Grade A Jumbo Brown Eggs** (1 dozen)		$1.49/ea	**$1.49**	
1/1	**Santa Barbara Hot Salsa, Fresh** (16oz)		$2.69/ea	**$2.69**	
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

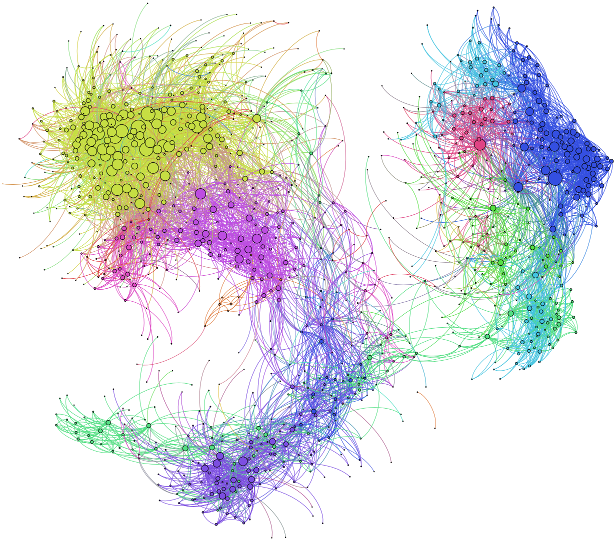
Genomics



Neuroscience



Social networks

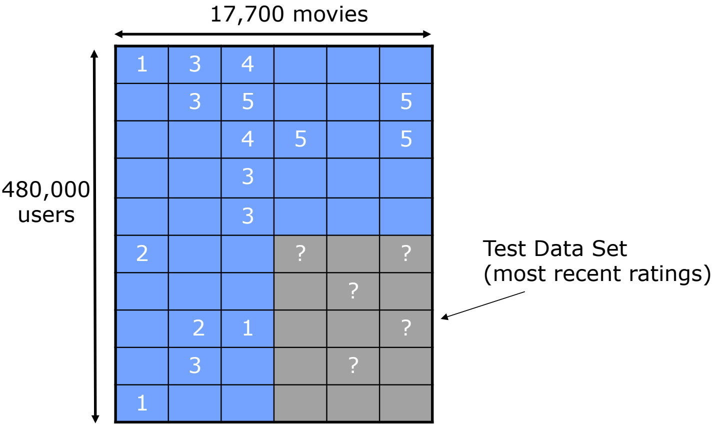


Data can help us solve problems.

Will NetFlix user 493234 like Transformers?

Will NetFlix user 493234 like Transformers?

How do you know?



Group these images into 3 groups



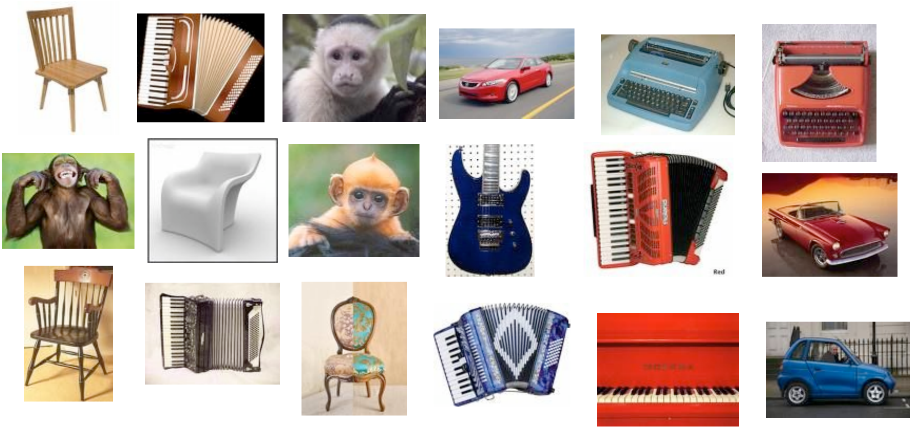






Group many images and determine the number of groups

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

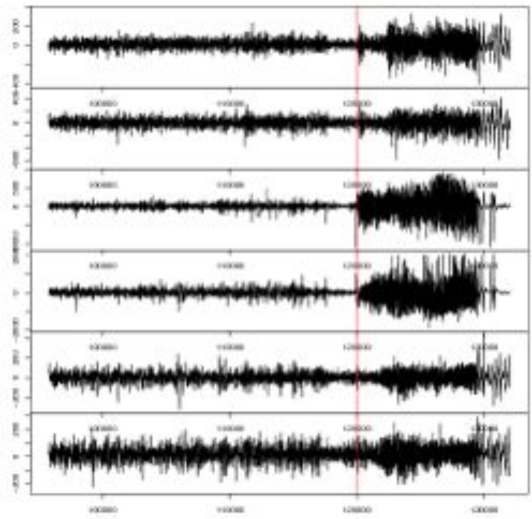
Dear Beloved,
My name is Mrs. Susan Polla, from ITALY. If you are a christian and interested in charity please reply me at : (s.polla@yahoo.fr) for insight.
Respectfully,
Mrs Susan Polla.

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.

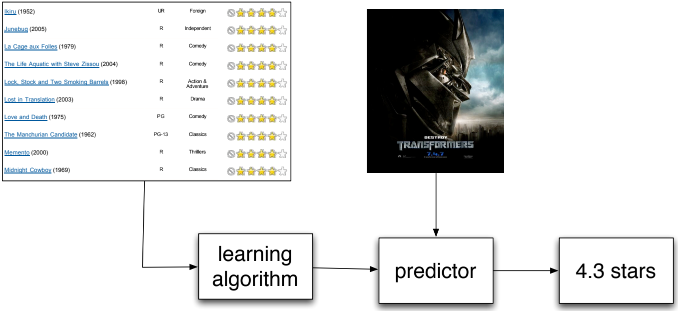
Applications

Data mining: the study algorithms that find and exploit patterns in data.

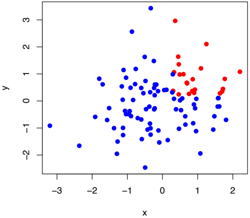
- These algorithms draw on ideas from statistics and machine learning.
- Applications include
  - natural science (e.g., genomics, neuroscience)
  - web technology (e.g., Google, NetFlix)
  - finance (e.g., stock prediction)
  - policy (e.g., predicting what intervention X will do)
  - and many others

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.**

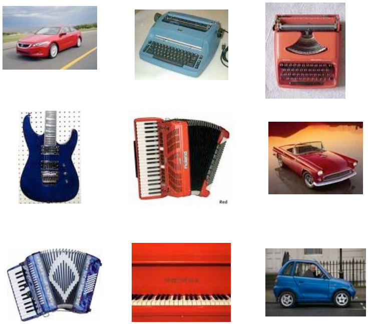


Supervised vs. unsupervised methods



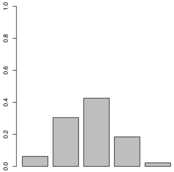
- **Supervised methods** find patterns in **fully observed** data and then try to predict something from **partially observed** data.
- For example, we might observe a collection of emails that are categorized into *spam* and *not spam*.
- After learning something about them, we want to take new email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

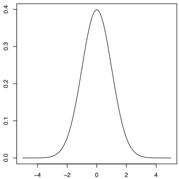
Supervised vs. unsupervised methods



- **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

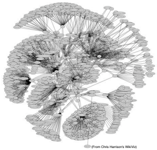
	discrete	*continuous*
supervised	**classification**	**regression**
unsupervised	**clustering**	**dimensionality reduction**

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...

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



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

This is an art

- A lot like compression
- The more intuitively you can represent data, the better you can do
  - Images: Closer to human perception (edges, features invariant to scale)
  - Sound: Using frequency features (like JPGs)
  - Text: Use syntax and chains of words "n-grams"
- For more, take "Digging into Data" in the Spring

Outline

Recap

- Storing data smart (aka resorting to trickery) can improve efficiency
- Storing data smart can let you do cool stuff with data