Data, Models, and First Steps

Digging into Data: Jordan Boyd-Graber

University of Maryland

January 27, 2014

COLLEGE OF INFORMATION STUDIES

Slides adapted from Dave Blei and Lauren Hannah

Roadmap

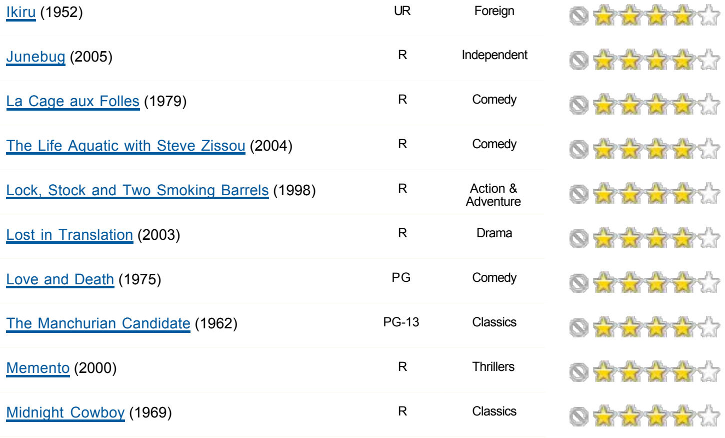
- The goals and ideas of the course
- Administrivia
- Getting started with Rattle and R

Outline

1. What can we do with data?
2. How this course is organized
3. Administrivia and Introductions
4. Introducing R and Rattle
5. Finding and using data
6. Showing off Rattle
7. Wrapup

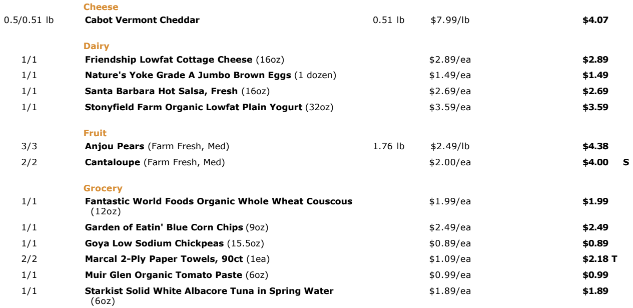
Data are everywhere.

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	⊘ ★★★★☆

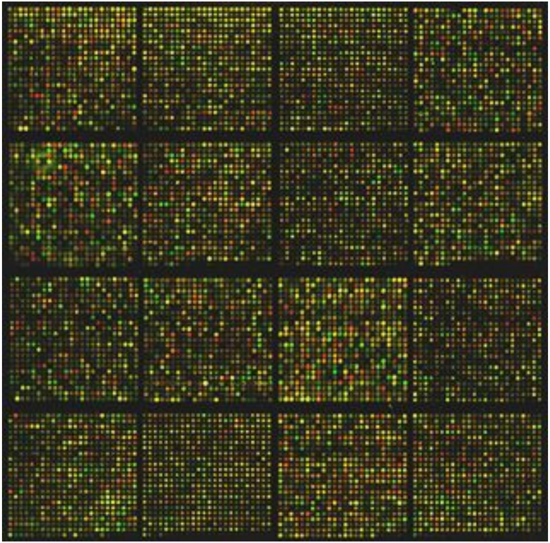
Purchase histories



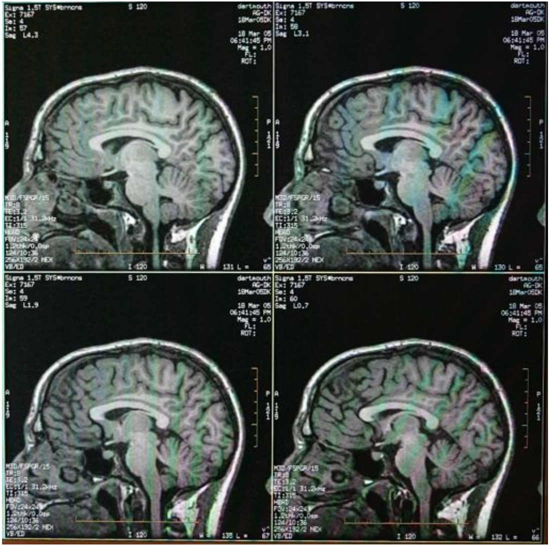
	Cheese				
0.5/0.51 lb	**Cabot Vermont Cheddar**	0.51 lb	$7.99/lb	**$4.07**	
	Dairy				
1/1	**Friendship Lowfat Cottage Cheese** (16oz)		$2.89/ea	**$2.89**	
1/1	**Nature's Yoke Grade A Jumbo Brown Eggs** (1 dozen)		$1.49/ea	**$1.49**	
1/1	**Santa Barbara Hot Salsa, Fresh** (16oz)		$2.69/ea	**$2.69**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

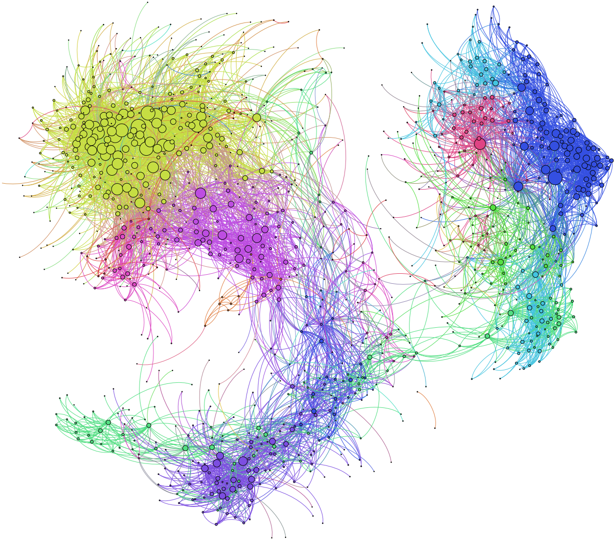
Genomics



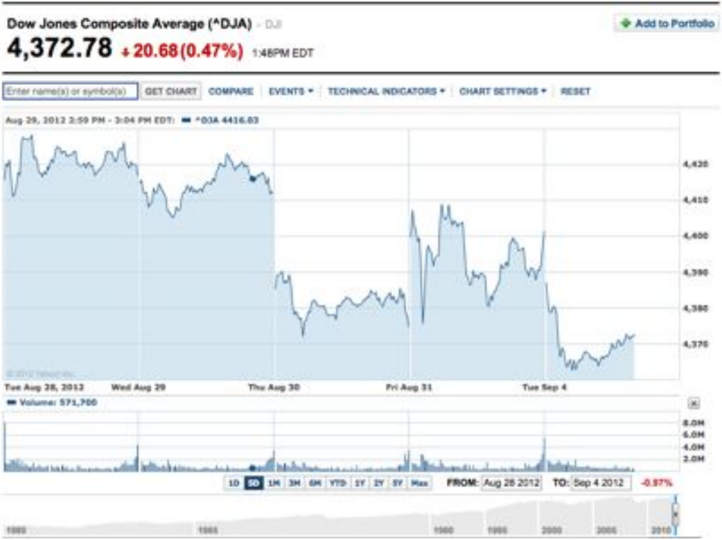
Neuroscience



Social networks



Finance



Dow Jones Composite Average (^DJA) - DJI

4,372.78 ↓20.68(0.47%) 1:48PM EDT

Data can help us solve problems.

Will NetFlix user 493234 like Transformers?

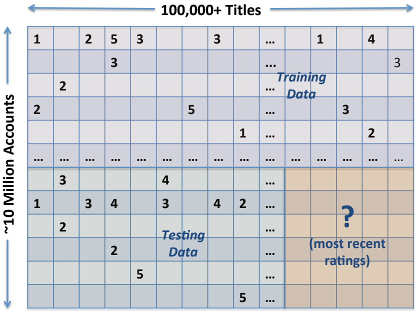
Will NetFlix user 493234 like Transformers?

★★☆☆☆

★★★★★

★★★★★

How do you know?



100,000+ Titles

~10 Million Accounts

1		2	5	3			3		...		1		4	
			3						...					3
	2								...					
2						5			...			3		
								1	...				2	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
	3				4				...					
1		3	4		3		4	2	...					
	2								...					
			2						...					
				5					...					
								5	...					

Training Data

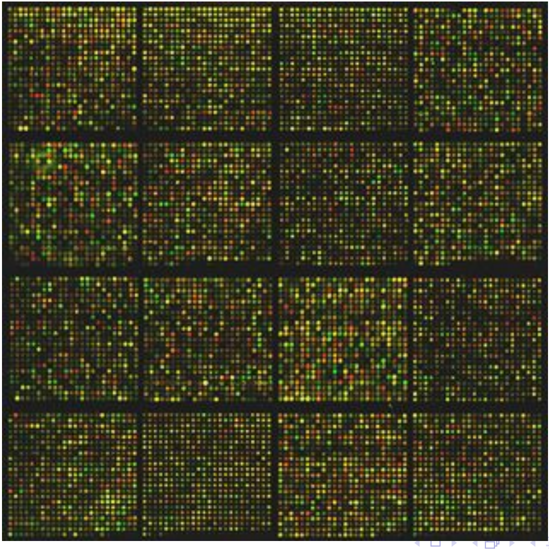
Testing Data

?

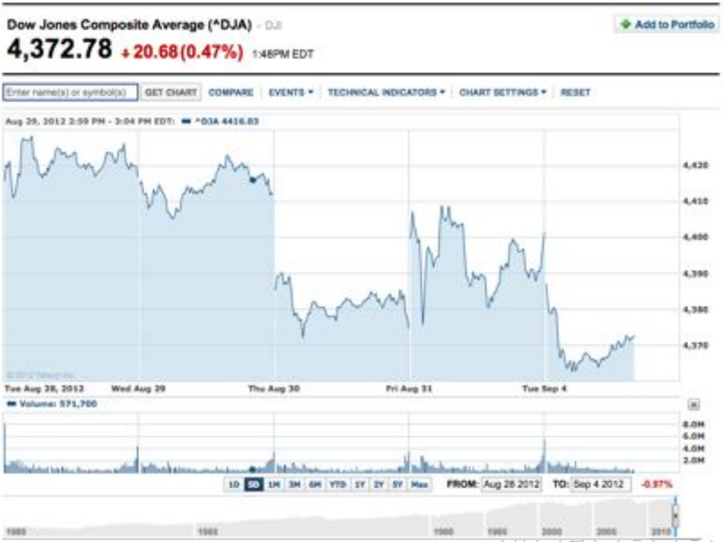
(most recent ratings)

Group many images and determine the number of groups

Which genes are associated with a disease? How can expression values be used to predict survival?



Is it likely that this stock was traded based on illegal insider information?



Dow Jones Composite Average (^DJA) - DJI

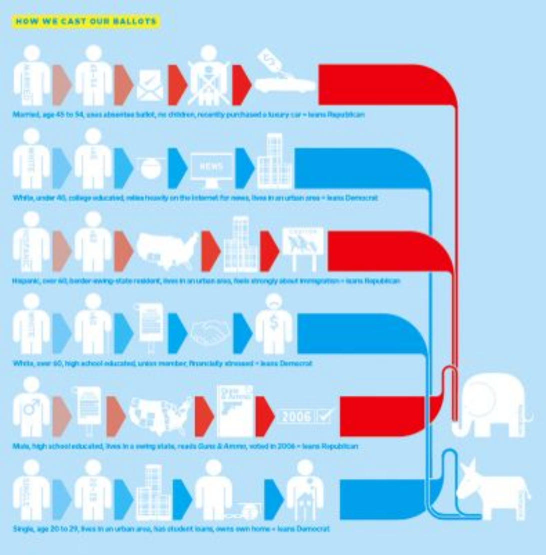
4,372.78 +20.68(0.47%) 1:48PM EDT

Aug 29, 2012 2:59 PM - 3:04 PM EDT: ^DJA 4416.83

Volume: 571,700

FROM: Aug 28 2012 TO: Sep 4 2012 -0.97%

Who will vote and for whom?


HOW WE CAST OUR BALLOTS
Married, age 45 to 54, uses absentee ballot, no children, recently purchased a luxury car = leans Republican
NEWS
White, under 40, college educated, relies heavily on the Internet for news, lives in an urban area = leans Democrat
Hispanic, over 40, border-swing-state resident, lives in an urban area, feels strongly about immigration = leans Republican
White, over 60, high school educated, union member, financially stressed = leans Democrat
Guns & Ammo
2006
Male, high school educated, lives in a swing state, reads Guns & Ammo, voted in 2006 = leans Republican
Single, age 20 to 29, lives in an urban area, has student loans, owns own home = leans Democrat

Is this spam?

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

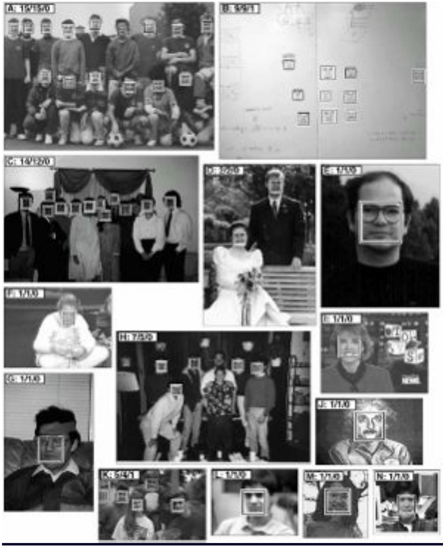
Dear Beloved,
My name is Mrs. Susan Polla, from ITALY. If you are a christian and interested in charity please reply me at : (s.polla@yahoo.fr) for insight.
Respectfully,
Mrs Susan Polla.

How about this one?

From: [snipped]
Subject: Superbowl?
Date: January 28, 2013 8:09:00 PM EST
To: jbg@umiacs.umd.edu, [snipped]

Anyone interested in coming by to watch the game? Beer and pizza, I'd imagine. Should be an exciting game!

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

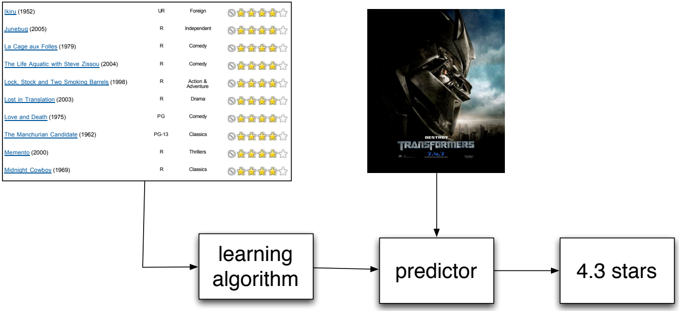
This Course (Digging into Data)

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	⊘ ★★★★☆

Outline

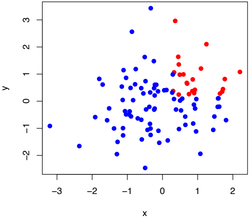
1. What can we do with data?
2. How this course is organized
3. Administrivia and Introductions
4. Introducing R and Rattle
5. Finding and using data
6. Showing off Rattle
7. Wrapup

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:

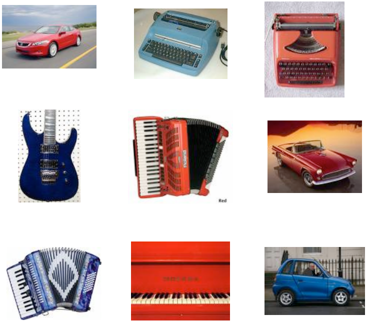
- Probabilistic foundations
- Supervised learning (more of this)
- Unsupervised learning (less of this)
- Methods that operate on discrete data (more of this)
- Methods that operate on continuous data (less of this)
- Representing data / feature engineering
- Evaluating models
- Understanding the assumptions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

Supervised vs. unsupervised methods



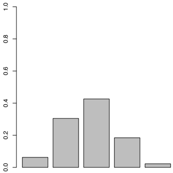
- **Supervised methods** find patterns in **fully observed** data and then try to predict something from **partially observed** data.
- For example, we might observe a collection of emails that are categorized into *spam* and *not spam*.
- After learning something about them, we want to take new email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

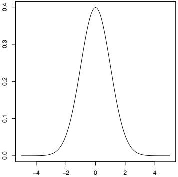
Supervised vs. unsupervised methods



- **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**

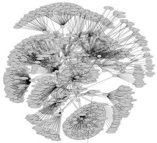
Disclaimer: most of my research falls under the “discrete” column (language), combining supervised and unsupervised methods

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...

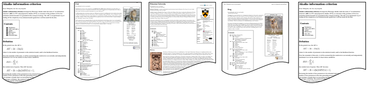
$\longrightarrow \langle 1, 0, 0, 0, 5, 0, 9, 3, 1, \ldots, 0 \rangle$



(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



- 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

Outline

1. What can we do with data?
2. How this course is organized
3. Administrivia and Introductions
4. Introducing R and Rattle
5. Finding and using data
6. Showing off Rattle
7. Wrapup

What you need for this course

- You need to use R and Rattle
- Helps to have a laptop to bring to class
- Math background
  - Not a machine learning course
  - Won't ask you to: prove anything, do integrals
  - You **do** need to be comfortable with some notation (sums, variables)
  - Will ask you to: add, divide, count, take logs
- Computer / programming skills
  - Don't need to know how to program (might help)
  - But you do need to be comfortable with assigning objects to variables
  - Need to be comfortable with the concept of functions (variables, return, etc.)
  - We'll use the command line (but you don't need to be a ninja)

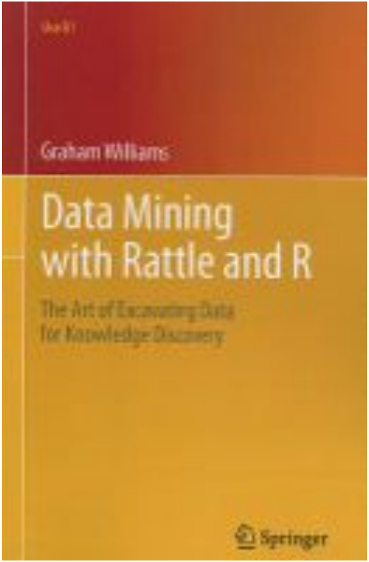
Flipped Classroom

- Last year: not enough hands-on practice
- My responsibility: record lectures before class
- In class: you help each other, and we work through examples
- Your responsibility: come to class with questions from lecture (I'll randomly call on you—part of participation)

Administrivia

- Keep track of course webpage
- Three homeworks: 5 late days
- Midterm
- Project
- Let me know about special needs

Course reading



Use R!

Graham Williams

Data Mining with Rattle and R

The Art of Excavating Data for Knowledge Discovery

Springer

- We will provide reading materials, mostly from the book.
- Slightly different focus: same concepts, use book as starting point

Communicating with Piazza

We will use Piazza to manage all communication

`https://piazza.com/umd/spring2014/inst737/home`

- Questions answered within 1 day (hopefully sooner)
- Hosts discussions among yourselves
- Use for any kind of technical question
- Use for **most** administrative questions
- Can use to send us private questions too
- Will be a factor in participation

How to ask for help

- Explain what you're trying to do
- Give a minimal example
  - Someone else should be able to replicate the problem easily
  - Shouldn't require any data / information that only you have
- Explain what you **think** should happen
- Explain what you get instead (copy / paste or screenshot if you can)
- Explain what else you've tried

Me

- Fourth year assistant professor
    - iSchool and UMIACS
    - Offices: 2118C Hornbake / 3126 AV Williams
- Second time teaching the class
- Born in Colorado (where all my family live)
- Grew up in Iowa (hometown: Keokuk, Iowa)
- Went to high school in Arkansas
- Undergrad in California
- Grad school in New Jersey
- Brief jobs in between:
    - Working on electronic dictionary in Berlin
    - Worked on Google Books in New York
- ying / jbg / jordan / boyd-graber

Outline

1. What can we do with data?
2. How this course is organized
3. Administrivia and Introductions
4. Introducing R and Rattle
5. Finding and using data
6. Showing off Rattle
7. Wrapup

Why R?

- It's Free
- Standard for statistical data science
- Used by major corporations (Facebook and Google)
- You can go very deep (if you need to)

Why Rattle?

- It's easy
- Introduces the power of R through a GUI
- Does 90% of what most users need
- Slowly eases you in to the other 10%

Installing R

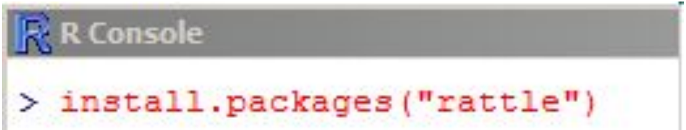
Download Installation File

http://watson.nci.nih.gov/cran_mirror/

- Particularly for OS X, download version 2-14.2
- Otherwise, you will get errors

Installing Rattle

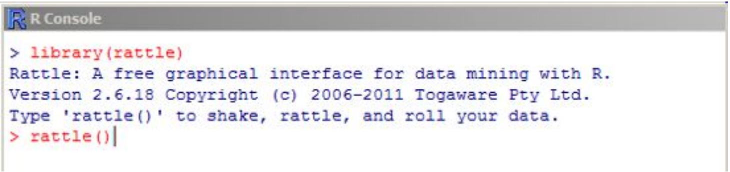
- Start R
- You'll see a command line



```
> install.packages("rattle")
```

- This tells it too look for the package "rattle" and install it
- It will ask you to choose a mirror to download the file from; choose an MD one (it's in Bethesda)

Running Rattle for the First Time



```
> library(rattle)
Rattle: A free graphical interface for data mining with R.
Version 2.6.18 Copyright (c) 2006-2011 Togaware Pty Ltd.
Type 'rattle()' to shake, rattle, and roll your data.
> rattle()
```

- It will ask you to install a bunch of things
- Just say “yes”
- If you have problems, try exiting R and trying again

`http://rattle.togaware.com/rattle-install-troubleshooting.html`

Homework 0 (not for credit)

Install R and Rattle to try it out!

Outline

1 What can we do with data?

2 How this course is organized

3 Administrivia and Introductions

4 Introducing R and Rattle

5 **Finding and using data**

6 Showing off Rattle

7 Wrapup

Where to get data?

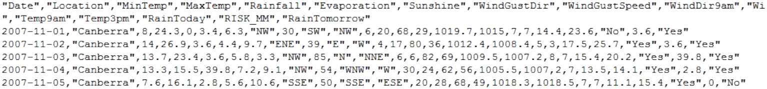
- data.gov - Obama initiative to get all government data in one place
- gapminder.org/data/ - Global development data
- infochimps.org - Pointers to interesting data
- http://bitly.com/bundles/hmason/1 - A set of links to data
- http://www.ncbi.nlm.nih.gov/ - National Center for Biotechnology Information
- http://www.ldc.upenn.edu/ - Linguistic Data Consortium
- Wild, Wild, Web
- Devices
- Research

Where to get data?

- data.gov - Obama initiative to get all government data in one place
- gapminder.org/data/ - Global development data
- infochimps.org - Pointers to interesting data
- http://bitly.com/bundles/hmason/1 - A set of links to data
- http://www.ncbi.nlm.nih.gov/ - National Center for Biotechnology Information
- http://www.ldc.upenn.edu/ - Linguistic Data Consortium
- Wild, Wild, Web
- Devices
- Research
- First Homework: Find some data and describe it

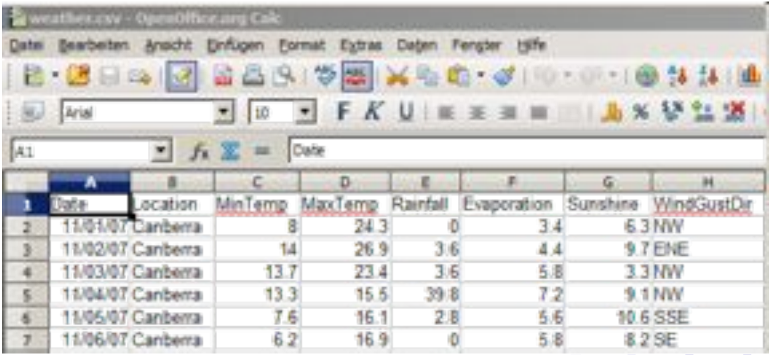
Let's get some data

- Download data from a weather station `http://goo.gl/X6EpS`
- Open it in a text application



```
"Date","Location","MinTemp","MaxTemp","Rainfall","Evaporation","Sunshine","WindGustDir","WindGustSpeed","WindDir9am","Wi
","Temp9am","Temp3pm","RainToday","RISK_MM","RainTomorrow"
2007-11-01,"Canberra",8,24.3,0,3.4,6.3,"NW",30,"SW","NW",6,20,68,29,1019.7,1015,7,7,14.4,23.6,"No",3.6,"Yes"
2007-11-02,"Canberra",14,26.9,3.6,4.4,9.7,"ENE",39,"E","W",4,17,80,36,1012.4,1008.4,5,3,17.5,25.7,"Yes",3.6,"Yes"
2007-11-03,"Canberra",13.7,23.4,3.6,5.8,3.3,"NW",85,"N","NNE",6,6,82,69,1009.5,1007.2,8,7,15.4,20.2,"Yes",39.8,"Yes"
2007-11-04,"Canberra",13.3,15.5,39.8,7.2,9.1,"NW",54,"WNW","W",30,24,62,56,1005.5,1007,2,7,13.5,14.1,"Yes",2.8,"Yes"
2007-11-05,"Canberra",7.6,16.1,2.8,5.6,10.6,"SSE",50,"SSE","ESE",20,28,68,49,1018.3,1018.5,7,7,11.1,15.4,"Yes",0,"No"
```

- Open it up in Excel or your favorite Spreadsheet



	A	B	C	D	E	F	G	H
1	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir
2	11/01/07	Canberra	8	24.3	0	3.4	6.3	NW
3	11/02/07	Canberra	14	26.9	3.6	4.4	9.7	ENE
4	11/03/07	Canberra	13.7	23.4	3.6	5.8	3.3	NW
5	11/04/07	Canberra	13.3	15.5	39.8	7.2	9.1	NW
6	11/05/07	Canberra	7.6	16.1	2.8	5.6	10.6	SSE
7	11/06/07	Canberra	6.2	16.9	0	5.8	8.2	SE

Digression: Comma Separated Value Files

- Carryover from punchcards (easier to type)
- Each data item is separated by comma (or another character)
- Just about everything can use it (lowest common denominator)
  - Libraries in programming languages (starting with Fortran)
  - Spreadsheet
  - Exports from applications / devices

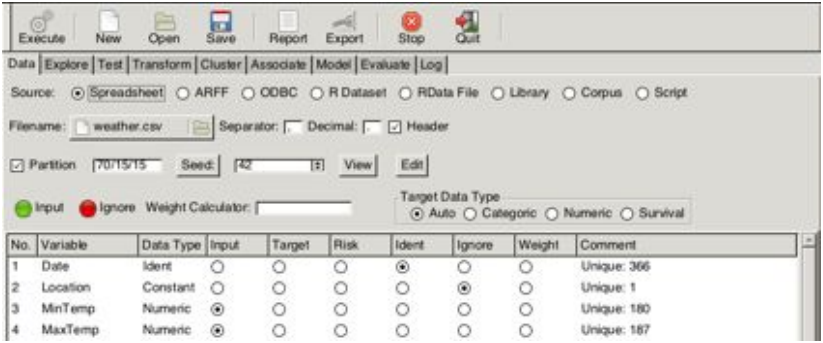
Outline

1. What can we do with data?
2. How this course is organized
3. Administrivia and Introductions
4. Introducing R and Rattle
5. Finding and using data
6. **Showing off Rattle**
7. Wrapup

Play with the weather data

- We'll only be showing off "coolness"
- Explanations later
- Goal: Get a sense of the data
- Goal: Predict when it will rain

Play with the weather data


Execute
New
Open
Save
Report
Export
Stop
Quit
Data | Explore | Test | Transform | Cluster | Associate | Model | Evaluate | Log

Source: ● Spreadsheet ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library ○ Corpus ○ Script

Filename: weather.csv Separator: , Decimal: . ☑ Header

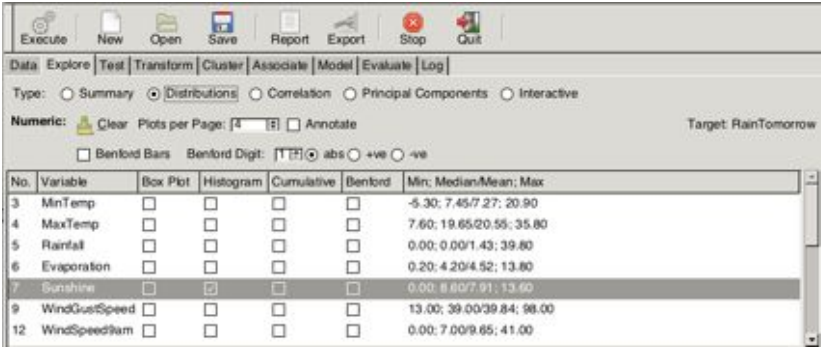
☑ Partition 70/15/15 Seed: 42 View Edit

🟢 Input 🔴 Ignore Weight Calculator:

Target Data Type ● Auto ○ Categoric ○ Numeric ○ Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	Date	Ident	○	○	○	●	○	○	Unique: 366
2	Location	Constant	○	○	○	○	●	○	Unique: 1
3	MinTemp	Numeric	●	○	○	○	○	○	Unique: 180
4	MaxTemp	Numeric	●	○	○	○	○	○	Unique: 187

Finding connections . . .



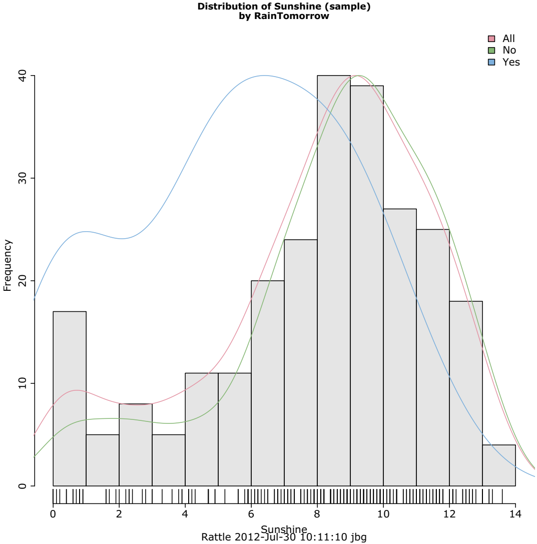
Type: ○ Summary ● Distributions ○ Correlation ○ Principal Components ○ Interactive

Numeric: Clear Plots per Page: 4 □ Annotate Target: RainTomorrow

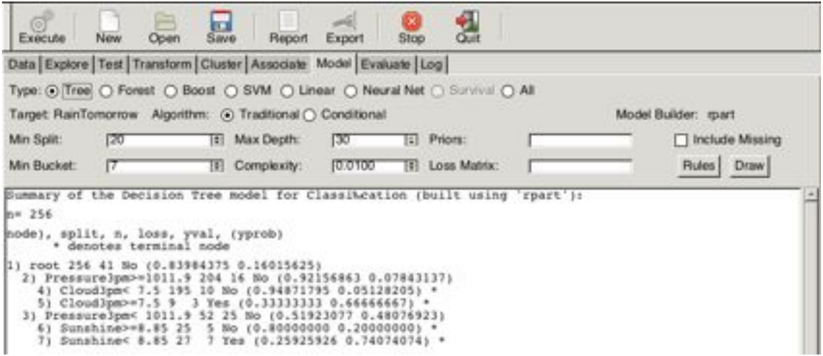
□ Benford Bars Benford Digit: 1 ● abs ○ +ve ○ -ve

No.	Variable	Box Plot	Histogram	Cumulative	Benford	Min; Median/Mean; Max
3	MinTemp	□	□	□	□	-5.30; 7.45/7.27; 20.90
4	MaxTemp	□	□	□	□	7.60; 19.65/20.55; 35.80
5	Rainfall	□	□	□	□	0.00; 0.00/1.43; 39.80
6	Evaporation	□	□	□	□	0.20; 4.20/4.52; 13.80
7	Sunshine	□	☑	□	□	0.00; 8.60/7.91; 13.60
9	WindGustSpeed	□	□	□	□	13.00; 39.00/39.84; 98.00
12	WindSpeed9am	□	□	□	□	0.00; 7.00/9.65; 41.00

Finding connections . . .



Making predictions . . .



Type: ● Tree ○ Forest ○ Boost ○ SVM ○ Linear ○ Neural Net ○ Survival ○ All

Target: RainTomorrow Algorithm: ● Traditional ○ Conditional Model Builder: rpart

Min Split: 20 Max Depth: 30 Priors: □ Include Missing

Min Bucket: 7 Complexity: 0.0100 Loss Matrix:

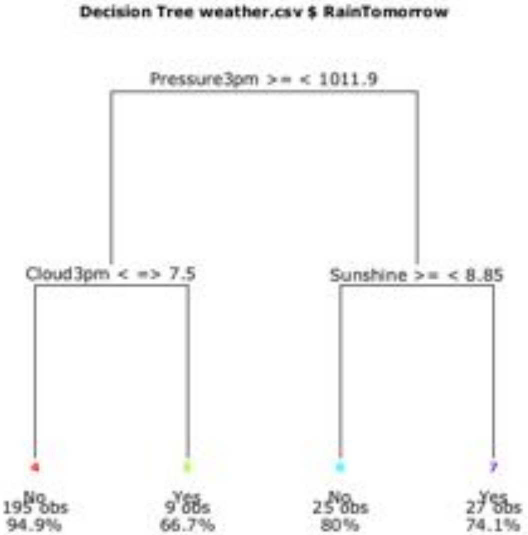
```
Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 256

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 256 41 No (0.83984375 0.16015625)
  2) Pressure3pm>=1011.9 204 16 No (0.92156863 0.07843137)
    4) Cloud3pm< 7.5 195 10 No (0.94871795 0.05128205) *
    5) Cloud3pm>=7.5 9  3 Yes (0.33333333 0.66666667) *
  3) Pressure3pm< 1011.9 52 25 No (0.51923077 0.48076923)
    6) Sunshine>=8.85 25  5 No (0.80000000 0.20000000) *
    7) Sunshine< 8.85 27  7 Yes (0.25925926 0.74074074) *
```

Making predictions . . .



Outline

1. What can we do with data?
2. How this course is organized
3. Administrivia and Introductions
4. Introducing R and Rattle
5. Finding and using data
6. Showing off Rattle
7. Wrapup

A statistician's manifesto

(From T. Hastie, via J. McAuliffe)

- Understand the ideas behind the statistical methods, so you know how to use them, when to use them, when *not* to use them.
- Complicated methods build on simple methods. Understand simple methods first.
- The results of a method are of little use without an assessment of how well or poorly it is doing.

Next time . . .

- What are probability distributions
- How to compute probabilities
- Properties of distributions