Data, Models, and First Steps

Digging into Data: Jordan Boyd-Graber

University of Maryland

January 28, 2013

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

<table>
<thead>
<tr>
<th>Movie</th>
<th>Rating</th>
<th>Genre</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Ikiru</em> (1952)</td>
<td>UR</td>
<td>Foreign</td>
<td>5</td>
</tr>
<tr>
<td><em>La Cage aux Folles</em> (1979)</td>
<td>R</td>
<td>Comedy</td>
<td>4</td>
</tr>
<tr>
<td><em>The Life Aquatic with Steve Zissou</em> (2004)</td>
<td>R</td>
<td>Comedy</td>
<td>2</td>
</tr>
<tr>
<td><em>Love and Death</em> (1975)</td>
<td>PG</td>
<td>Comedy</td>
<td>2</td>
</tr>
<tr>
<td><em>The Manchurian Candidate</em> (1962)</td>
<td>PG-13</td>
<td>Classics</td>
<td>3</td>
</tr>
<tr>
<td><em>Midnight Cowboy</em> (1969)</td>
<td>R</td>
<td>Classics</td>
<td>2</td>
</tr>
</tbody>
</table>
**Cheese**

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Item Description</th>
<th>Weight</th>
<th>Unit Price</th>
<th>Total Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5/0.51 lb</td>
<td>Cabot Vermont Cheddar</td>
<td>0.51 lb</td>
<td>$7.99/lb</td>
<td>$4.07</td>
</tr>
</tbody>
</table>

**Dairy**

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Item Description</th>
<th>Unit Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/1</td>
<td>Friendship Lowfat Cottage Cheese (16oz)</td>
<td>$2.89/ea</td>
</tr>
<tr>
<td>1/1</td>
<td>Nature’s Yoke Grade A Jumbo Brown Eggs (1 dozen)</td>
<td>$1.49/ea</td>
</tr>
<tr>
<td>1/1</td>
<td>Santa Barbara Hot Salsa, Fresh (16oz)</td>
<td>$2.69/ea</td>
</tr>
<tr>
<td>1/1</td>
<td>Stonyfield Farm Organic Lowfat Plain Yogurt (32oz)</td>
<td>$3.59/ea</td>
</tr>
</tbody>
</table>

**Fruit**

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Item Description</th>
<th>Weight</th>
<th>Unit Price</th>
<th>Total Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>3/3</td>
<td>Anjou Pears (Farm Fresh, Med)</td>
<td>1.76 lb</td>
<td>$2.49/lb</td>
<td>$4.38</td>
</tr>
<tr>
<td>2/2</td>
<td>Cantaloupe (Farm Fresh, Med)</td>
<td></td>
<td>$2.00/ea</td>
<td>$4.00</td>
</tr>
</tbody>
</table>

**Grocery**

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Item Description</th>
<th>Unit Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/1</td>
<td>Fantastic World Foods Organic Whole Wheat Couscous (12oz)</td>
<td>$1.99/ea</td>
</tr>
<tr>
<td>1/1</td>
<td>Garden of Eatin' Blue Corn Chips (9oz)</td>
<td>$2.49/ea</td>
</tr>
<tr>
<td>1/1</td>
<td>Goya Low Sodium Chickpeas (15.5oz)</td>
<td>$0.89/ea</td>
</tr>
<tr>
<td>2/2</td>
<td>Marcal 2-Ply Paper Towels, 90ct (1ea)</td>
<td>$1.09/ea</td>
</tr>
<tr>
<td>1/1</td>
<td>Muir Glen Organic Tomato Paste (6oz)</td>
<td>$0.99/ea</td>
</tr>
<tr>
<td>1/1</td>
<td>Starkist Solid White Albacore Tuna in Spring Water (6oz)</td>
<td>$1.89/ea</td>
</tr>
</tbody>
</table>

**Vegetables & Herbs**

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Item Description</th>
<th>Weight</th>
<th>Unit Price</th>
<th>Total Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.99/ea</td>
<td>Starkist Solid White Albacore Tuna in Spring Water (6oz)</td>
<td>$1.89/ea</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

*Digging into Data: Jordan Boyd-Graber (UM) Data, Models, and First Steps January 28, 2013 6 / 57*
Document collections
Social networks

Figure 2: Link communities in US direct flight data detected by Online MMLC. Each segment is less than 500 miles resulting in regional groups. Node sizes on top represent bridgeness, and on the bottom, influence.

MMSB [4] is a Bayesian probabilistic model of network data. MMSB is a mixed membership model. It associates each unit of observation with multiple clusters rather than a single cluster via a membership probability vector. It assumes \( K \) communities and directed edges are placed independently between node pairs with probabilities determined only by the pair's community memberships.

Given the communities, the generative process defines a multinomial distribution over each node's
Data can help us solve problems.
Will Netflix user 493234 like Transformers?
Will Netflix user 493234 like Transformers?
### 100,000+ Titles

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>3</th>
<th>3</th>
<th>...</th>
<th>1</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### ~10 Million Accounts

- **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?
Who will vote and for whom?

Married, age 45 to 54, uses absentee ballot, no children, recently purchased a luxury car = leans Republican

White, under 40, college educated, relies heavily on the Internet for news, lives in an urban area = leans Democrat

Hispanic, over 60, 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

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.
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
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 solving data analysis problems. On completing the course, you will be able to learn about a new tool, apply it to 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.

---

**Digging into Data: Jordan Boyd-Graber (UMD)**

**Data, Models, and First Steps**

January 28, 2013 25 / 57
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 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

<table>
<thead>
<tr>
<th>Supervised</th>
<th>Discrete</th>
<th>Continuous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised</td>
<td>Classification</td>
<td>Regression</td>
</tr>
<tr>
<td></td>
<td>Clustering</td>
<td>Dimensionality</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reduction</td>
</tr>
</tbody>
</table>
Republican nominee George Bush said he felt nervous as he voted today in his adopted home state of Texas, where he ended...

\[ \langle 1.5, 3.2, -5.1, \ldots, 4.2 \rangle \]

\[ \langle 1, 0, 0, 0, 5, 0, 9, 3, 1, \ldots, 0 \rangle \]

\[
\begin{bmatrix}
1 & 0 & 1 & \ldots & 0 \\
0 & 1 & 1 & \ldots & 0 \\
1 & 0 & 0 & \ldots & 1 \\
& & & \ldots & \\
0 & 0 & 0 & \ldots & 0
\end{bmatrix}
\]
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)
Administrivia

- Sign up on Piazza (use a photo)
- Keep track of course webpage
- 5 late days
- Let me know about special needs
We will provide reading materials, mostly from the book. The reading will cover more than we cover in class.
Communicating with Piazza

We will use Piazza to manage all communication

http://piazza.com/class#spring2012/cos424

- 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
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

- Third year assistant professor
  - iSchool and UMIACS
  - Offices: 2118C Hornbake / 3126 AV Williams
- First 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/

- In Linux, many package managers already have it
- It is also installed on http://libcc.org
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

- 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.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.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
- Open it up in Excel or your favorite Spreadsheet
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

![Image of a data analysis interface]

- Data: Spreadsheet
- Filename: weather.csv
- Partition: 7/15/15
- Seed: 42

<table>
<thead>
<tr>
<th>No.</th>
<th>Variable</th>
<th>Data Type</th>
<th>Input</th>
<th>Target</th>
<th>Risk</th>
<th>Ident</th>
<th>Ignore</th>
<th>Weight</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Date</td>
<td>Ident</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Unique: 366</td>
</tr>
<tr>
<td>2</td>
<td>Location</td>
<td>Constant</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Unique: 1</td>
</tr>
<tr>
<td>3</td>
<td>MinTemp</td>
<td>Numeric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Unique: 180</td>
</tr>
<tr>
<td>4</td>
<td>MaxTemp</td>
<td>Numeric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Unique: 187</td>
</tr>
</tbody>
</table>
Finding connections ...
Finding connections...
Making predictions . . .

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...

Decision Tree weather.csv $ RainTomorrow

Pressure3pm $\geq$ 1011.9

Cloud3pm $<=$ 7.5

- **No**: 195 obs, 94.9%
- **Yes**: 9 obs, 66.7%

Sunshine $\geq$ 8.85

- **No**: 25 obs, 80%
- **Yes**: 27 obs, 74.1%
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.
What are probability distributions
- How to compute probabilities
- Properties of distributions