“If you can’t explain it simply, you don’t understand it well enough.”

- Albert Einstein
What is it worth?

1. Companies choose celebrities to represent their companies

2. Given a celebrity, which company would they best represent based on what people are saying about them?

Let’s use a simple correlation analysis
LOUIS VUITTON 0.353
Microsoft 0.359
TAG HEUER 0.407
MACY’S 0.324
STARBUCKS COFFEE 0.320
BWM 0.320
Coca-Cola 0.294
GE 0.220

BIG DATA – MINING SOCIAL MEDIA
DATA SOCIETY © 2015
Value of applications

- Data-driven approach to validating ideas
- Makes conclusions obvious
- Improved decision making
McDonalds picked up JT in 2003

Justin Timberlake has signed a multi-tiered sponsorship deal with McDonald’s. The alliance is tied to the restaurant chain’s new worldwide marketing campaign, which launches this month. Timberlake will
Outline

1. What is data science and its tools?
2. What is Twitter?
3. Setting up Twitter
4. Applications
   - Twitter
   - LinkedIn
5. Being thoughtful about mining social media
6. Resources
   - Accessing Twitter Feeds
   - Getting LinkedIn data
What is data science?

• Data science applies the scientific method to analyzing data

• It lies at the intersection of several disciplines

• It draws on domain specific knowledge that makes the analysis of Big Data possible
Data science control cycle

1. Ask
2. Research
3. Model
4. Validate
5. Test
6. Interpret

Sanity check

© 2015 Data Society
What is Twitter?
What is Twitter?

- Online social networking service

- Users send and read short 140-character messages called "tweets"

- Tweets happen at the speed of thought
  - Broad cross-section of society at an international level
  - Connects people to dialogue, as well as areas of interest
  - People are curious, want to express themselves, and like instant gratification
  - 100 million active users
What data can we get from Twitter?

- Text
- Links
- Author
- Blurb about author
- Location of author at time of tweet
- Number of followers of author (including names)

- People followed
- User retweeted
- Time the tweet was made
- Trending topics
Some terminology

• # (Hashtag): relates to topics mentioned by other users

• @ (at): person who’s tweeting or person being tweeted at
Required for OAuth

- API key
- API Secret
- Access Token
- Access Token Secret

*Store these as variables in R*
Access Twitter from R

- Needs to install and load 2 packages
  1. “devtools”
  2. “twitteR”

- devtools needs to be loaded first,
- twitter is downloaded from github

```r
### Setting up packages/ libraries ####

## Install and load the "devtools" package ##
install.packages("devtools")
library(devtools)

## Install and load the "twitter" package ##
install_github("twitter", username="geoffjentry")
library(twitter)
```
# Authentication process

# Setting up twitter access using OAuthenticate

api_key="enter api key here"

api_secret="enter api secret here"

access_token="enter access token here"

access_secret="enter access secret here"

setup_twitter_oauth(api_key, api_secret, access_token, access_secret)
Applications

Twitter
5 big questions

1. What are people talking about?

2. What patterns exist in tweets?
   - Geography, time, topic

3. What relationships can be extracted from the tweets?

4. What's the sentiment/opinion of the people?

5. How does information spread?
Understanding fashion

- Step 1: Fashion questions
- Step 2: Get data
- Step 3: Clean data
- Step 4: Format data (in tdm)
- Step 5: Visualize data:
  - Comparison word cloud
  - Commonality word cloud
Step 1: What are the right questions?

- Questions:
  - How are fashion companies similar/different?
  - What are fashion companies focusing on?
- Companies:
  - H&M
  - Zara
  - United Colors of Benetton
  - Calvin Klein
Step 5: Visualize data

#Comparison cloud
> comparison.cloud(tdm_fashion, colors = brewer.pal(5, "Dark2"))

#Commonality cloud
> commonality.cloud(tdm_fashion, colors = brewer.pal(5, "Dark2"))
Conclusions - Comparison

• Zara engages it’s audience better than any other company
  – Key Identifiers: pictures, newthisweek

• H & M focuses on fast trends
  – Key Identifiers: fashion, style
  – Calvin Klein promotes shopping online and has a revealing campaign?
  – Key Identifiers: shop, collection, revealmore

• UCBenetton supports campaigns having to do with women
  – unwomen, orangeurhood, supporting
UCB: UNwomen and Orangeurhood

END VIOLENCE AGAINST WOMEN NOW!
#orangeurhood
Conclusion - Commonality

- Winter trends?
  - Denim

- General promotion words
  - Last
  - Campaign
  - Night
  - Weekend
Commercial Significance

• Retail brands sell a lifestyle

• Companies can use Twitter analysis (and other social media) to make sure that they present the kind of lifestyle that they are aiming to sell

• Public perception
  – Companies can monitor what people are saying about them in social media by locality and in real time

• Retail brands can ensure that they differentiate themselves through their messaging and public perception to best appeal to their clientele
What’s going on: Charlie Hebdo

Picture of Je Suis Charlie (I am Charlie)
Picture: Wikipedia (Elya)
Uncovering meaning

- **Step 1:** What political events are likely to unravel?
- **Step 2:** Get user data
- **Step 3:** Clean data
- **Step 4:** Format data
  - tdm
  - Filter sparse terms
- **Step 5:** Visualize data
  - Hierarchical clustering
What’s happening?

Using complete hierarchical clustering with 3 clusters
What’s happening?

Using complete hierarchical clustering with 8 clusters
What’s happening?

Using average hierarchical clustering with 8 clusters
What’s happening?

• Two different clustering algorithms (complete and average)
• 8 clusters
• Consistency between clusters:
  – muslim, protests kill, windsor, violent, star, reportedly
  – speech, expression, free
  – terror, french, around, magazine, pakistan, rally
  – pak, death, shoot, protesters, cartoonists, demand, race, twitter
  – paris, new, sale
Possible interpretations

• There may be violent protests held involving Muslims

• People are supporting the expression of free speech

• There is terror around a French magazine and there’s a rally about it in Pakistan

• May have been a shooting of cartoonists that resulted in death, led to protests, and motivated by a racial undertone

• There is a news sale in Paris
Validating assumptions

10 killed, churches torched in protests over Charlie Hebdo

By Josh Levs
Updated 7:51 AM ET, Wed January 21, 2015

Shooting at Charlie Hebdo

Complete coverage: Terror in Paris
The investigation: Following
Significance

• Analyze current events in real time and understand their impact

• Understand how people are interpreting current events and their reactions

• Unrest can be anticipated (Ferguson, MO) and security and emergency services deployed

• Companies and politicians can tailor their messaging to best engage their audiences as current events unfold
Politics: Jeb Bush

Framework:
• Step 1: Jeb Bush?
• Step 2: Get data
• Step 3: Clean data
• Step 4: Format data
  – i) In tdm
  – ii) Remove sparse terms
• Step 5: Visualize data
  – i) Clustering
  – ii) Association Graph
Step 1: Who is Jeb Bush

• Seriously
  – If you just saw the name, what could you say about him?
• What associations?
• Meaningful connections?
Step 5: Plotting
Step 5: Plotting

Graph of Tweets for Jeb Bush $k=7$
Step 5: Plotting

Graph of Tweets for Jeb Bush k=8
What the data says

• 4 Stable clusters:
  – Cluster 1: raising, derail, bid, conservatives
  – Cluster 2: golf, bag
  – Cluster 3: speaks, political, loyalties, split, oconnor, block, republican, Melissa
  – Cluster 4: kurtz,says,press, starts, scrutinizing, wont, says ryan
Possible conclusions

• **Cluster 1**: conservatives are raising a bid to derail Jeb Bush

• **Cluster 2**: Jeb Bush goes golfing

• **Cluster 3**: Someone named Melissa and O’Connor speak about splitting political loyalties in the Republican party

• **Cluster 4**: Ryan and Kurtz in the press won’t start scrutinizing Jeb Bush
Conservatives want to derail Jeb

Cluster 1:
Conservatives are raising a bid to derail Jeb Bush
Mitt and Jeb in the golf bag together
The GOP establishment lifts its weary head

By ROGER SIMON | 1/13/15 5:35 AM EST

Today, Mitt Romney is faced with a simple yet profound question: Which of his failed former advisers will he rehire to fail to make him president once again?

Past advisers are pushing hard for Romney to run for president again in 2016. They stand to make millions if he does. And Mitt doesn’t care about the money. He can find millions in his couch cushions.

But the national press corps is less happy with the idea. It is the duty of the national press corps to descend from the hills after the battle is over to shoot the wounded.
Mitt and Jeb could split the party

Mitt Romney, Jeb Bush Could Split Republican Loyalties

JANUARY 13, 2015   4:24 PM ET

Listen to the Story
All Things Considered

4 min 15 sec

Melissa Block speaks with Patrick O’Connor, political reporter for the Wall Street Journal about Mitt Romney telling donors he wants to run again for president in 2016. O’Connor says Romney and former Florida Gov. Jeb Bush have the advantage of not being in office and have the ability to raise more money via superPACs before they declare their candidacy.

Cluster 3: Someone named Melissa and O’Connor speak about splitting political loyalties in the Republican party
Importance

1. Find connections more readily

2. Plots the connections and shows differences

3. Can be used to mine:
   - “Secret information”
   - Latent structure
Polarities: Elizabeth vs. Hillary

Elizabeth Warren

Hillary Clinton
Polarities

- **Step 1**: Hillary vs Warren
- **Step 2**: Get data
- **Step 3**: Clean data
- **Step 4**: Format data
  - i) In tdm
  - ii) Remove sparse terms
- **Step 5**: Visualize data
  - i) Clustering
  - ii) Association Graph
Step 8: Polarity Plot

Conway's Word Cloud, Tweets (Hillary Clinton vs. Elizabeth Warren)
Possible conclusions

• Hillary Clinton:
  – Words: watch, families, people, help, stop, economy

• Possible conclusions:
  – Family is important
  – Focused on improving the economy
Possible conclusions

• Elizabeth Warren:
  – **Words:** new, happy, year, birthday, women

• Possible conclusions:
  – Uses Twitter to wish people happy birthday and happy new year
  – Uses Twitter more as a social rather than a way to disseminate information compared to Hillary Clinton
  – Focused more on women’s rights
Importance

- Polarity ranking
  - Political campaigns
  - Upcoming elections
  - Referendums
- Rivalries
  - Messi vs. Ronaldo
  - Tech Crunch vs. Wired
  - Mac OS vs. Windows
Sentiment Analysis

Bill Maher  VS.  Bill O’Reilly
Sentiment Analysis

• **Step 1:** Who do people like more?
• **Step 2:** Get data for “bill maher” and “bill oreilly”
• **Step 3:** Clean data
• **Step 4:** Format data
  1. tdm -> matrix
  2. Setup sentiments (positive, negative)
• **Step 5:** Score data
  – Sentiment Analysis
Step 2: read in data

• We read in 2 sets of words
  – Words from Twitter
  – Positive, negative and neutral words

\[
\text{Score} = \frac{\# \text{ Times Positive Adjective Mentioned for individual } k}{\# \text{ Times Negative Adjectives Mentioned for individual } k}
\]
Step 5: scoring data

<table>
<thead>
<tr>
<th># Time Positive Words Mentioned</th>
<th># Time Negative Words Mentioned</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>382</td>
<td>293</td>
<td>89</td>
</tr>
</tbody>
</table>

For $n=1000$

<table>
<thead>
<tr>
<th># Time Positive Words Mentioned</th>
<th># Time Negative Words Mentioned</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>154</td>
<td>428</td>
<td>-274</td>
</tr>
</tbody>
</table>
Possible conclusions and applications

• Conclusions:
  – More people have an affinity for Bill Maher than Bill O'Reilly

• Applications:
  – Review of consumer products
  – Monitoring reputation
  – Financial markets
Applications
LinkedIn
Getting association data
Setup network on a 2d plot
Remove links
Remove names
Use clustering
Restore links and names
Information diffusion in a network

- Step 1: How does information flow?
- Step 2: Modeling
- Step 3: Get data
- Step 4: Generate model
- Step 5: Animate
Step 2: Google SI Model

Not a *Sports Illustrated* model
but rather a *Susceptible Infected* model
Step 2: Susceptible-Infected model

- Mathematical model
  - Susceptible-infected (SI) model
  - Based on randomness
- Simulates information propagation
  - Marketing
  - Viral videos
  - Spread of disease
Step 5: Animated model
Be thoughtful about data mining
Be thoughtful about data mining

• What we saw:
  – Mining data is only as good as the interpretation
  – Corroborate findings
  – Twitter and LinkedIn are media for acquiring data, not interpreting it

• Why it’s important:
  – Data generated approach to validating ideas
  – Makes conclusions obvious
  – Improved decision making
Final words

• This is just an overview of what’s possible with a few simple lines of code

• This is the direction in which the world is heading (i.e. data-driven approaches)

• Thinking about data in a different way

• What you can do to learn more about these techniques and how to use them
“If you can’t explain it simply, you don’t understand it well enough.”

- Albert Einstein
Resources
Search: text

- **Text:** All text containing "datascience"

```r
> tweets = searchTwitter("datascience", lang = "en")
> head(tweets, n=1)
[[1]]
[1] "arvind_Ilamaran: RT @SpazioDati: Apache Spark’s journey from academia to industry http://t.co/B8Pg2gs9ui via @bigdata datascience apachesc..."
```
Search: user

User:
Returns tweets from user “datascienceco”

> userTimeline("datascietyco", n=3)

[[1]]
[1] "datascietyco: After @facebook's research went public, should #datascientists create their own code of ethics? #datascience http://t.co/75Y92XMP4F"
Search: location, time and dates

**Location:**
Returns tweets within for a given set of coordinates for term “datascience”

```r
> searchTwitter("datascience", geocode = "long, lat, 10mi")
```

**Time/Dates:**
Returns tweets for a given time interval coordinates for term “#datascience”

```r
> searchTwitter("#datascience", since="2014-19-14", until= "2014-20-14")
```

* search goes back only so far
Search: trend

Trend:
Returns tweets that are trending at woeid location

```r
> trending_washington = closestTrendLocations(lat = 39, long = -77)

> getTrends(woeid = trending_washington$woeid)$name
[1] "David Luiz"                 "#GoldenGlobeAwards2015" "TOTY"
[4] "US Central Command"         "#HouseofCards"        "Peyton"
[7] "George Clooney"             "#RHOA"                  "Valentine's Day"
[10] "#WorldCup"

* woeid = “Where on Earth Identification”
User information functions

```r
# Get info about Data Society
ds=getUser("datasocietyco")

# User description
ds$getDescription()

# User followers
ds$getFollowersCount()

# User friends
ds$getFriends(n=5)

# User favorites
ds$getFavorites(n=5)
```

"Data Society is an online data science education company that is dedicated to teaching data science in a simple, fun and applicable way. To everyone."
Getting LinkedIn data

- [www.socilab.com](http://www.socilab.com) – provides you with an association graph
- You can request a download of your LinkedIn data from your Privacy & Settings page.
  - To request a download of your data:
  - Move your cursor over your profile photo at the top right of your homepage and select Privacy & Settings.
    - You may be prompted to sign in.
  - Click the Account tab near the bottom of the page.
  - Click Request an archive of your data under the Helpful Links section.