Unlock Unstructured Data

Hui Lin

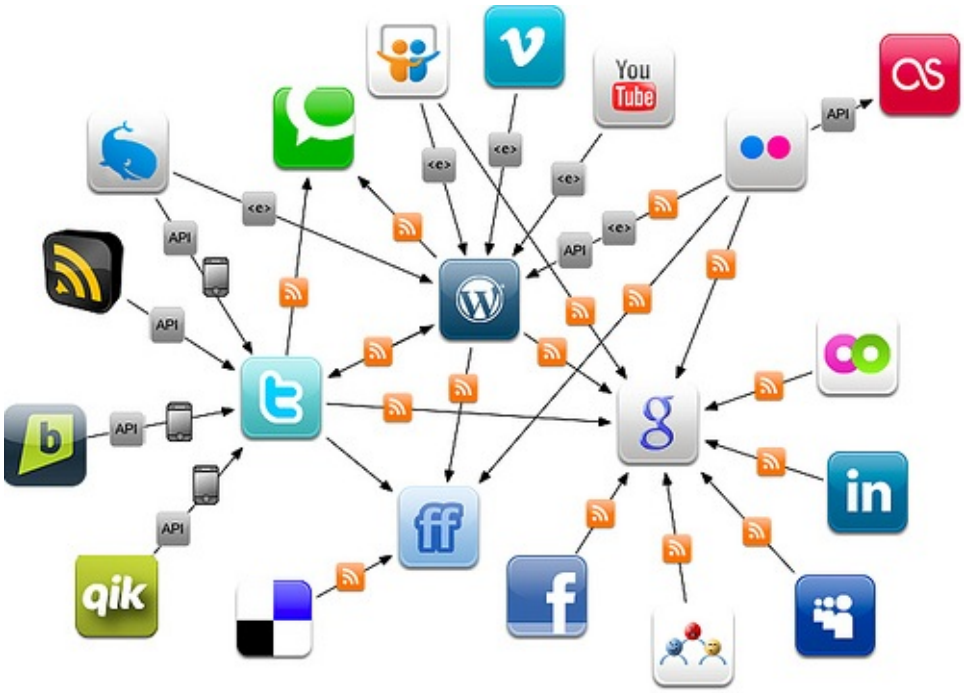
2016-12-16

What I do?

Outline

- What?
- Why?
- Where?
- Automatic data pipeline
- Data analysis
- Visualization and Report
- Cross functional collaboration

What is Unstructured Data?



Why?—It is everywhere

- Open government data
- Search engine data
- Services that track social behavior
- Social media data

Why?—Practical arguments

- Social media is impactful! [Amusing Ourselves to Death: Public Discourse in the Age of Show Business (1985)]
  - *"the medium is the message"*
  - *"the medium is the metaphor"*
  - *"the form excludes the content"*
- Financial resources are sparse
- … and so is our time
- Reproducibility

Where to get the data?

- API: Twitter/Google/Wikipedia…
- Webpage: Forum, Reviews
- Survey
- Interviews

API: Trump v.s Clinton Wikipedia View

Static Web: Wikipedia

Show 10 entries Search:

	name	n
1	Ronald Fisher	36
2	Karl Pearson	28
3	Jerzy Neyman	17
4	Udny Yule	13
5	Abraham Wald	12
6	Milton Friedman	11
7	Alexander Alexandrovich Chuprov	10
8	Arthur Lyon Bowley	10
9	Dennis Lindley	10
10	Harold Hotelling	10

Showing 1 to 10 of 621 entries Previous 1 2 3 4 5 … 63 Next

Static Web: buzzfeed.com

Show 10 entries Search:

	Titles
1	Download The BuzzFeed App Now!
2	23 Cheap Ways To Seem Like You're Actually Fancy
3	17 Stunning Women Who Make 'The Big Chop' Look So Damn Good
4	The Internet Has Decided On Its Favorite Cursed Child Couple
5	How Normal Are Your Burger Choices?
6	Can You Get Through This Post Without Spending $50
7	What's The Deal With Those Men's Gymnastics "Onesies"?
8	Raise Your Kitchen Game With The BuzzFeed Food Newsletter!
9	17 Things All Secretly Judgmental People Will Understand
10	Oh My God, Beach Volleyball Is Magnificent

Showing 1 to 10 of 98 entries Previous 1 2 3 4 5 … 10 Next

Static Web: buzzfeed.com

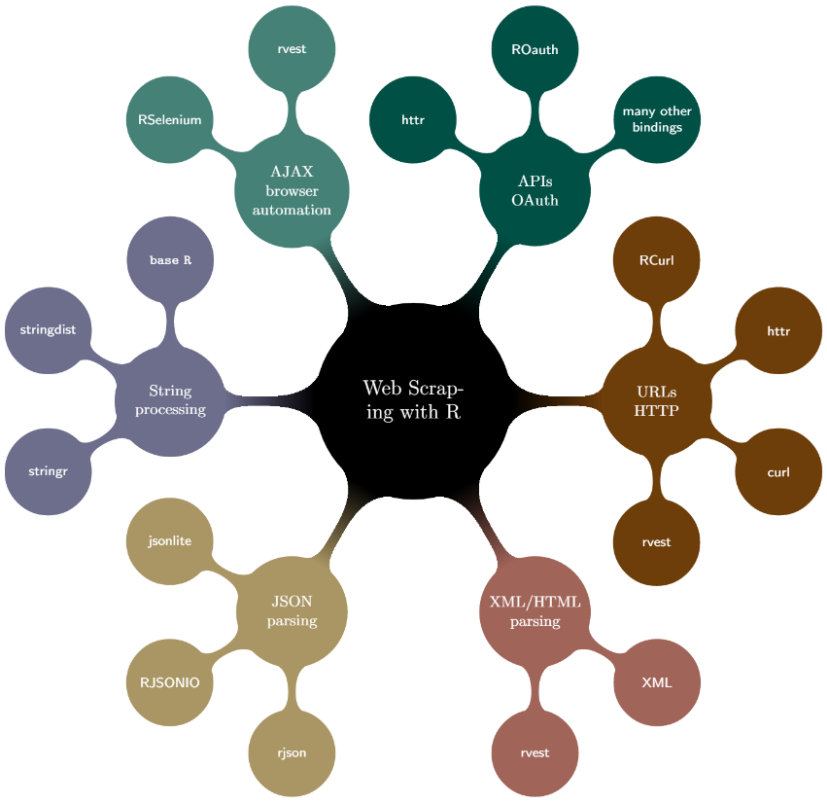
Show 10 entries Search:

	authors	Freq
1	David Mack	5
2	Alicia Melville-Smith	3
3	Andy Neuenschwander	3
4	Adam Davis	2
5	Adolfo Flores	2
6	Brian Galindo	2
7	Casey Rackham	2
8	Erin Chack	2
9	Erin La Rosa	2
10	Gena-mour Barrett	2

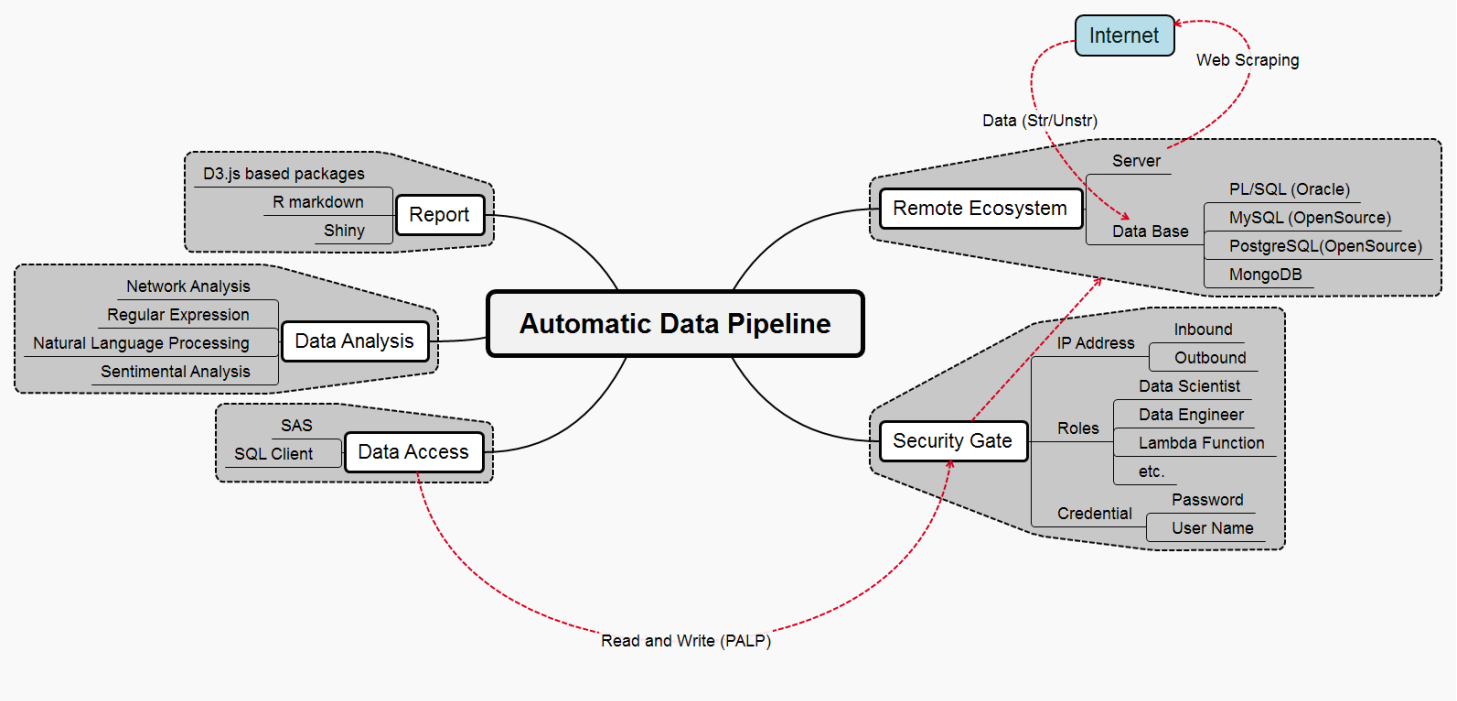
Summary of packages

- https://github.com/ropensci/user2016-tutorial

R tools



Automatic Data Pipeline



Data Analytics: Regular Expression

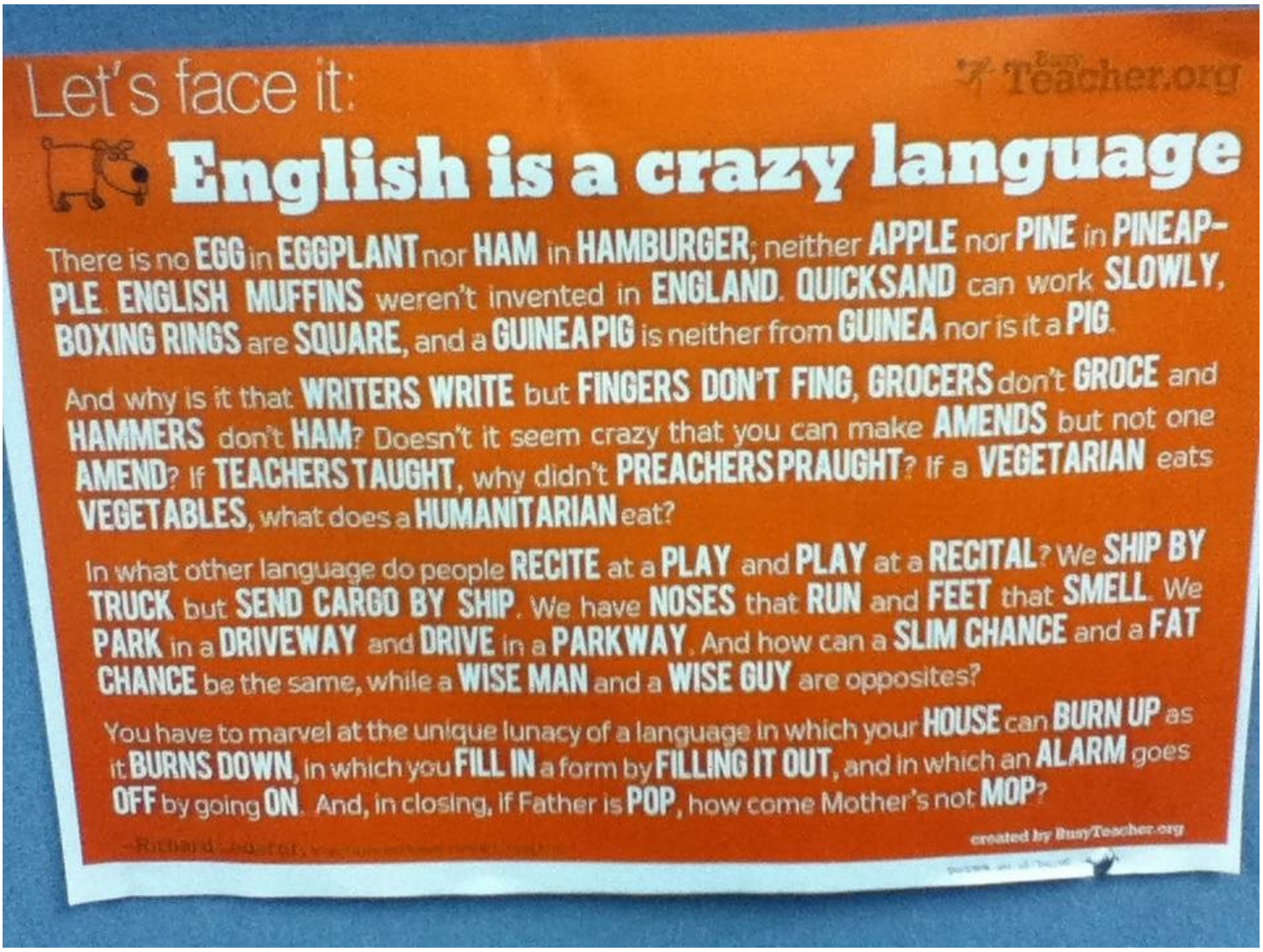
Data Analytics: Regular Expression

- http://stat545.com/block022_regular-expression.html

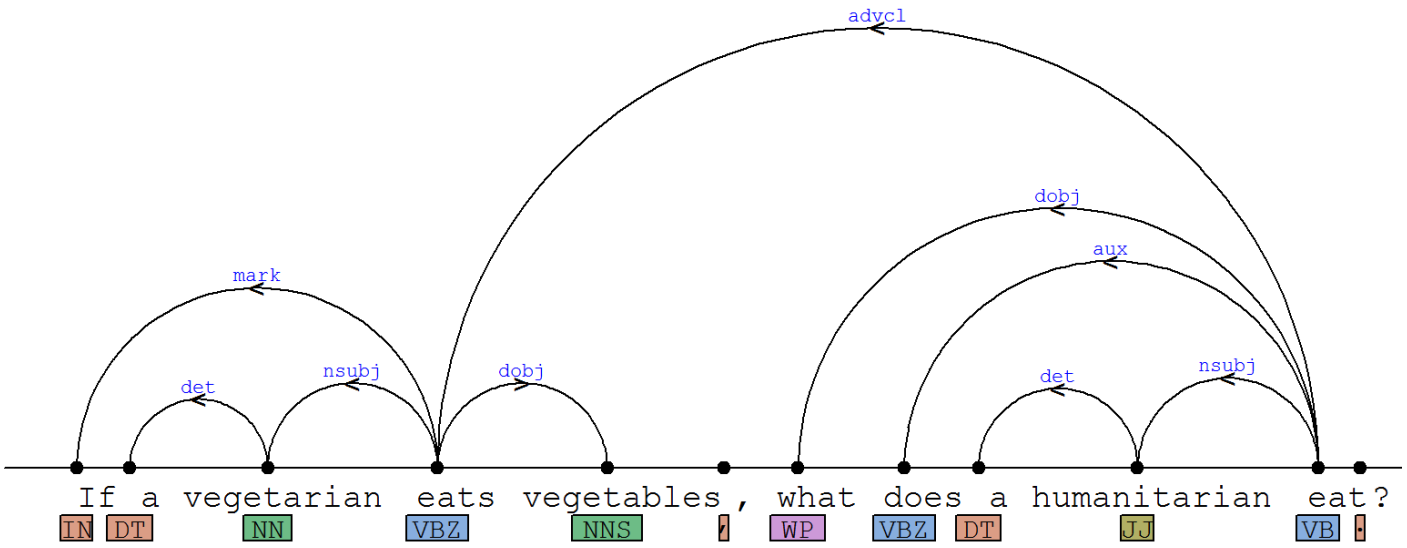
```
x <- c("here", "is", "P9929AMXT", "a", "P9703AM", "baby",
    "P0506AM", "example", "P1197AM", "P1271AM")
idx <- grep("(^P)[[:digit:]]+", x)
x[idx]
```

```
## [1] "P9929AMXT" "P9703AM"   "P0506AM"   "P1197AM"   "P1271AM"
```

Data Analytics: Natural Language Processing



NLP: How does computer understand language?



Issue driven

- If you don't know where to go ……

Issue driven

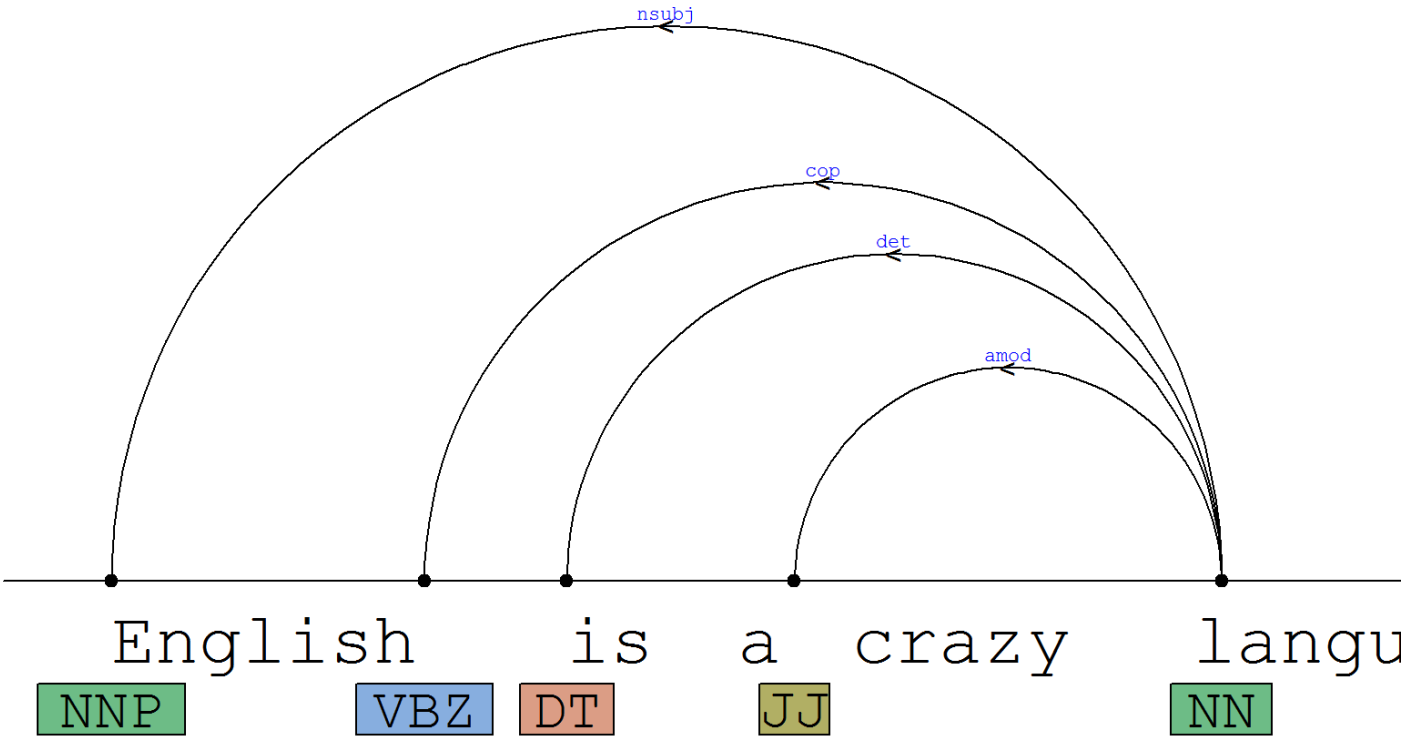
- If you don't know where to go …

- If you know where to go …

NLP: What are you interested in?

NLP: What are you interested in?

```
## [1] "English is a crazy language"
## [1] "English muffins"
```



Marketing Campaign: `#yieldhero`

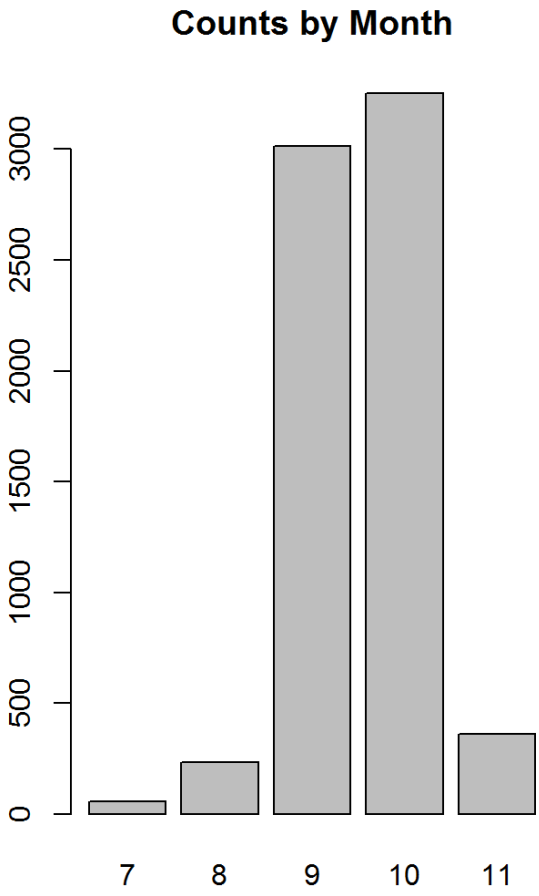
- When is the best time to tweet
- Who to target
- Product mentioned
- Sentiment score

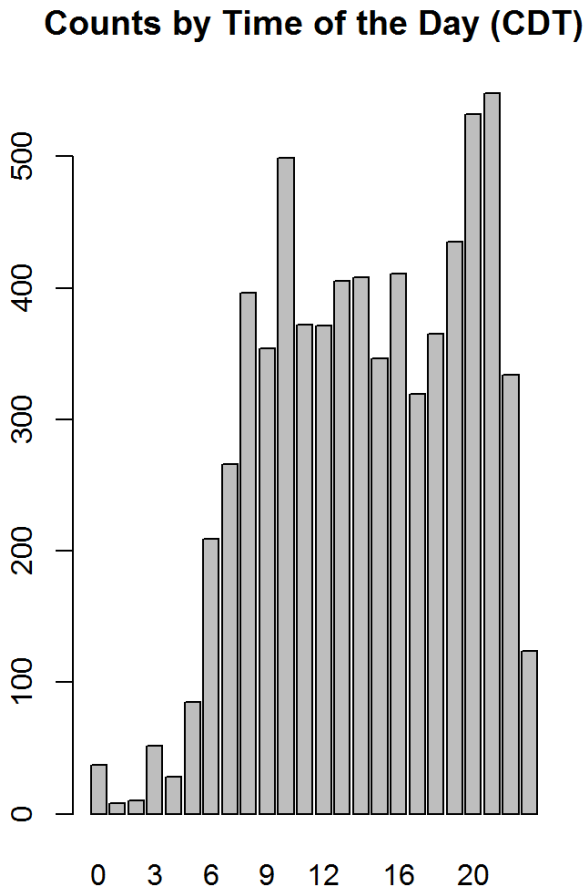
#yieldhero Summary Statistics

- From 2016-07-28 to 2016-11-18
- There are 6914 tweets, 1930 original tweets
- Products mentioned > 20 times
  - *P1197AM*
  - *P0157AMX*
  - *P22T73R*
  - *P28T08R*

When is the best time to tweet?

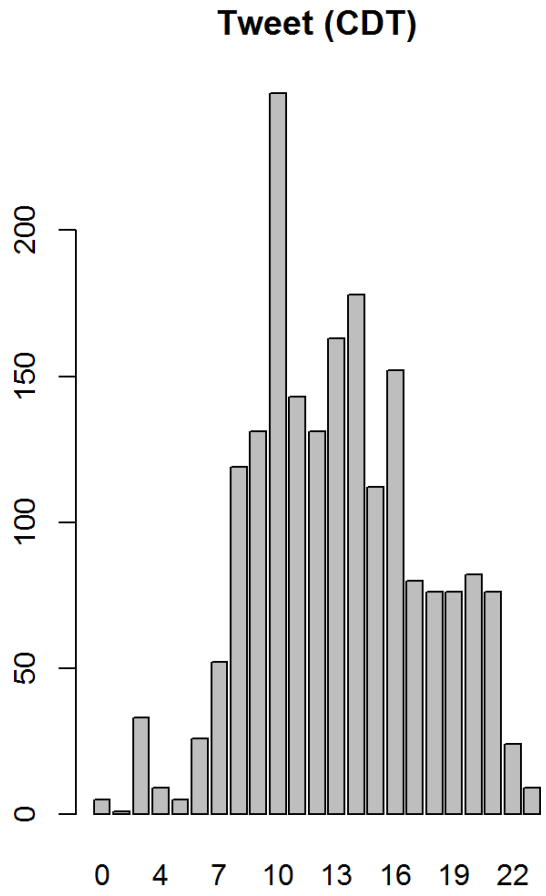
- Total Tweet Counts

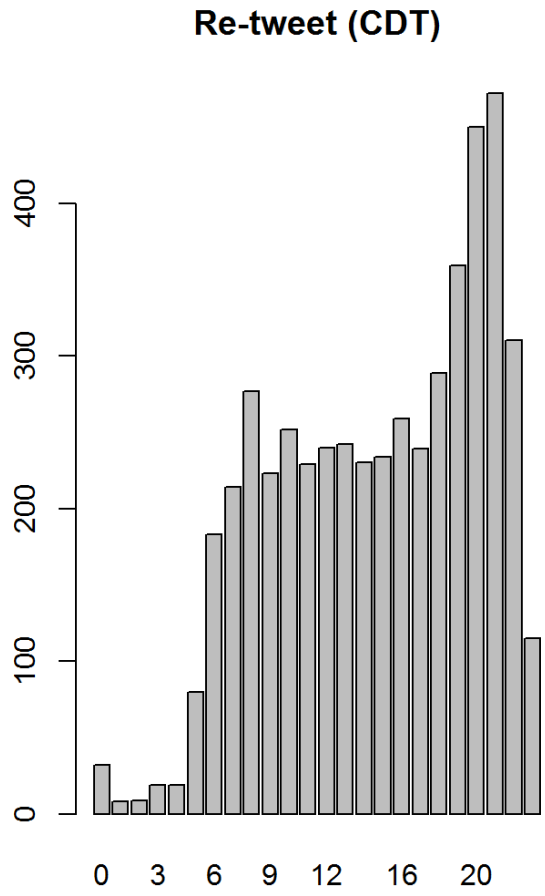




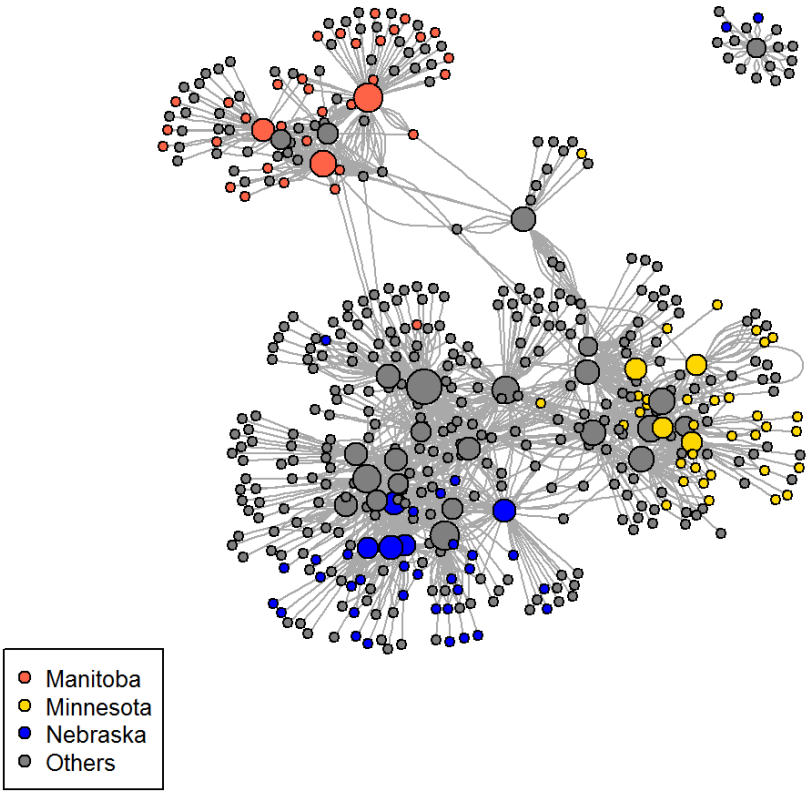
When is the best time to tweet?

- Tweet and Re-tweet counts by time of the day

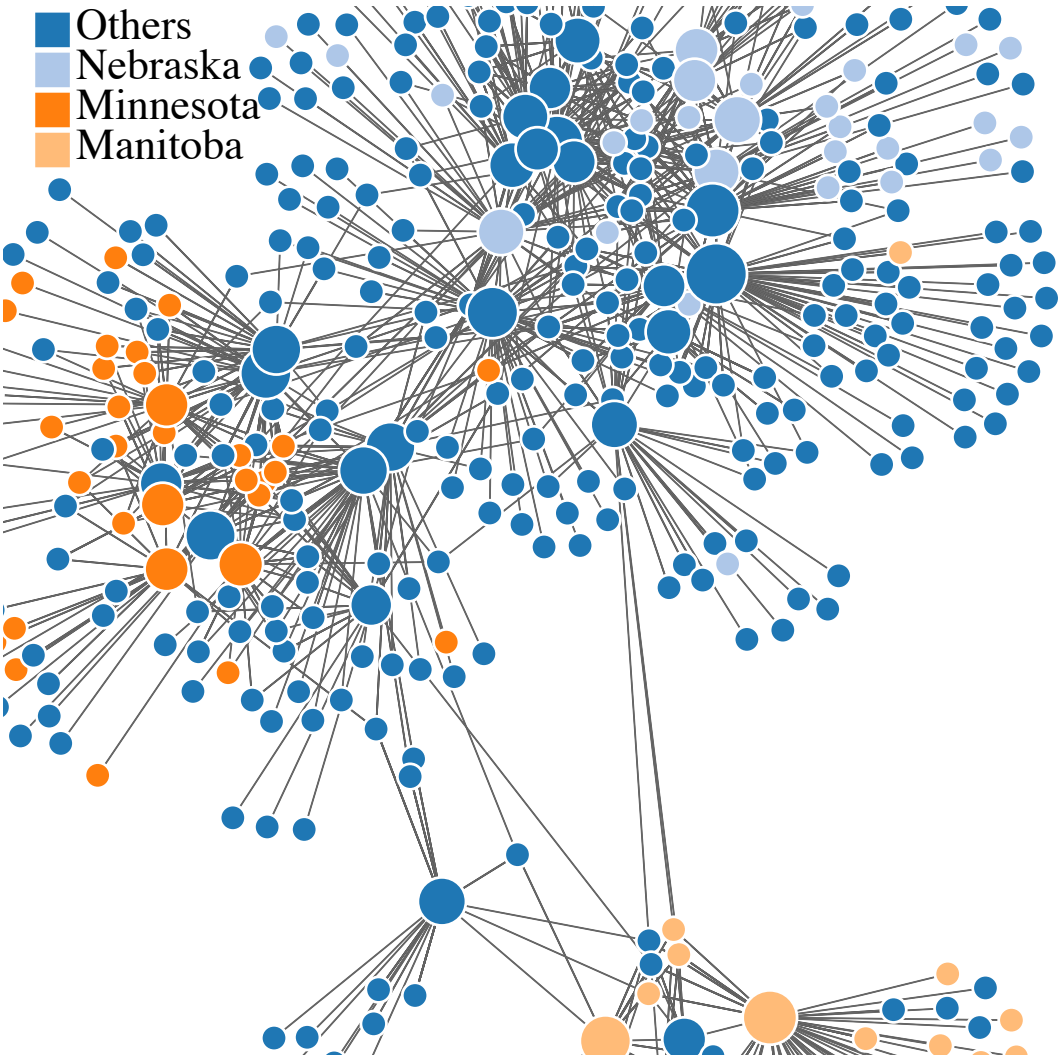




Who to target?



Network



Who to target?

Show 5 entries Search:

	name	size	state
	All	All	All
1	PallasSeeds	1	Unknown
2	jrjr58	1	Indiana
3	frenchrw	1	Texas
4	Pioneer_IL	1	Illinois
5	DarrenVanness	1	Nebraska

Showing 1 to 5 of 482 entries

Previous 1 2 3 4 5 ... 97 Next

Products Mentioned

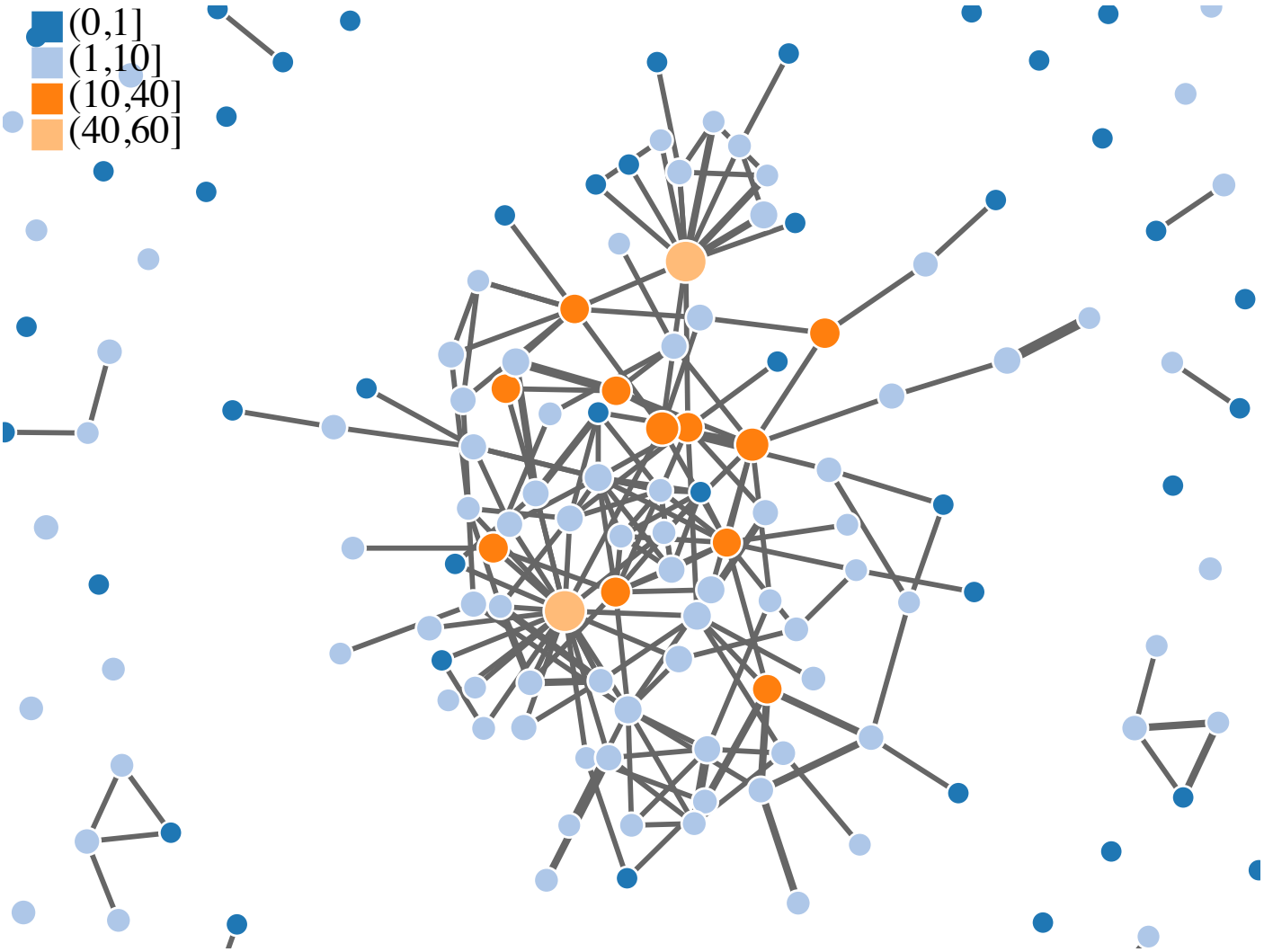


Table of Products

Show 7 entries Search:

	PROD_NM	size
	All	All
1	P1197AM	48
2	P0157AMX	47
3	P22T73R	21
4	P28T08R	21
5	P9188AM	14
6	P0506AM	13
7	P0589AMXT	13

Showing 1 to 7 of 209 entries Previous 1 2 3 4 5 … 30 Next

Shiny App Example

```
library(shiny)
runApp('Rcode/Shiny_NLP')
```

Is web scraping legal?

- No unambiguous yes or no in any country according to current jurisdiction
- So far, court cases (especially in the US) often dealt with commercial interest and often huge masses of data
  - *eBay vs. Bidder's Edge*
  - *AP vs. Meltwater*
  - *Facebook vs. Pete Warden*
  - *United States vs. Aaron Swartz*

Recommendation for your work

- Encrypt sensitive personal identifiable information
- YOU take all the responsibility for your web scraping work
- If you publish data, do not commit copyright fraud
- If in doubt, ask the author/creator/provider of data for permission
- Consult current jurisdiction

Trick: robots.txt

- What is robots.txt?

  "Robots Exclusion Protocol", informal protocol to prohibit web robots from crawling content

- Located in the root directory of a website, e.g. http://baidu.com/robots.txt
- Documents which bot is allowed to crawl which resources (and which not)
- Not a technical barrier, but a sign that asks for compliance
- Syntax in robots.txt
- Scraping etiquette

Data and Code

- All data are available in the package I am developing:

```
library(devtools)
install_github("happyrabbit/DataScienceR")
library(DataScienceR)
```

- Slides and R code can be found here:
- http://scientistcafe.com/