

R and the Data Science ToolKit (RDSTK)

Ryan Elmore
National Renewable Energy Lab

@rtelmore
rtelmore@gmail.com

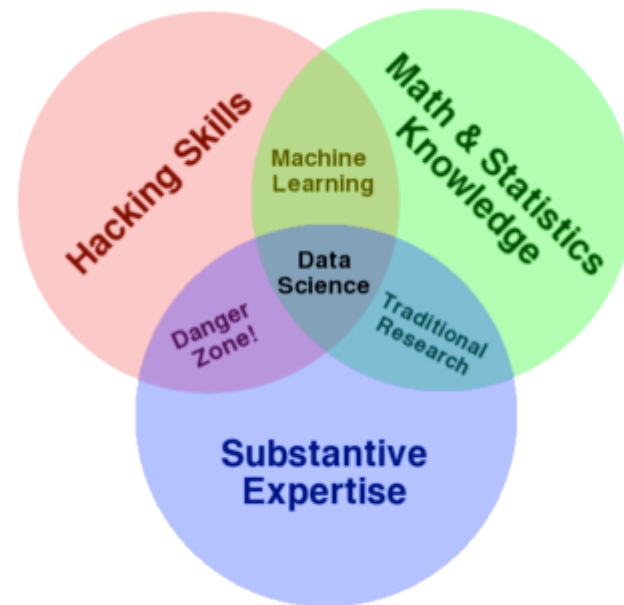
What is Data Science?

What is Data Science?

- Short Answer: Statistics!

What is Data Science?

- Short Answer: Statistics!
- Long Answer: Some hybrid of math/statistics, hacking skills, substantive expertise, possibly nunchaku and/or bow hunting skills [1]
- However nebulous, there is a lot of hype around the term.

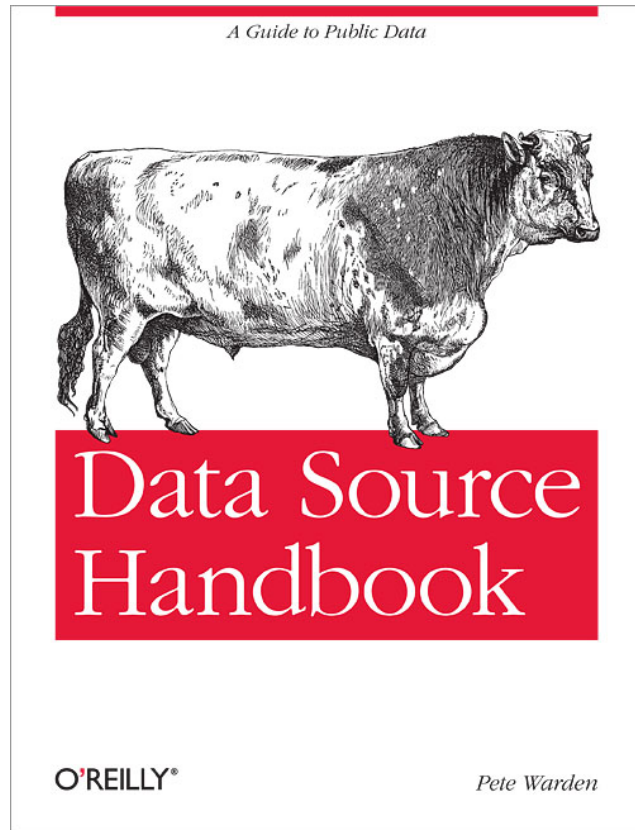


<http://www.dataists.com/2010/09/the-data-science-venn-diagram/>

[1] - "You know, like nunchaku skills, bow hunting skills, computer hacking skills... Girls only want boyfriends who have great skills." -- Napoleon Dynamite

Hype?

Hype Indeed!



Data Week: Becoming a data scientist

Data Pointed, CouchDB in the Cloud, Launching Strata

by [Edd Dumbill](#) | [@edd](#) | [Comment](#) | 2 September 2010



157



Why the term "data science" is flawed but useful

Counterpoints to four common data science criticisms.

by [Pete Warden](#) | [@petewarden](#) | [Comments: 5](#) | 9 May 2011

And Even More Hype...

The Economist May 14-20, 2011: “Corporate chefs are in demand again, office rents are soaring and the pay being offered to talented folk in fashionable fields like **data science** is reaching Hollywood levels.”



DATA**SCIENCE**TOOLKIT


- www.datasciencetoolkit.com
- Pete Warden [2] is the author
- You can download and run it as a self-contained virtual machine, python, javascript, or use the API online.
- To call the API, you can make either a GET or POST request.
- No R package...until now


[2] - You may remember him from classics such as his fight with facebook and/or the iphone tracking stuff.

RDSTK


- github.com/rtelmore/RDSTK
- All of the DSTK functionality is supported except for geodict and file2text. (hmmm, hack?)
- Dependencies: plyr, rcurl, rjson
- Typical usage involves passing in a text string and a data.frame or json string is

Github

 SOCIAL CODING

 **rtelmore** | [Dashboard](#) | [Inbox](#) 0 | [Account Settings](#) | [Log Out](#)

[Explore GitHub](#) | [Gist](#) | [Blog](#) | [Help](#) |

 **rtelmore** (Ryan Elmore) This is you! [Edit Your Profile](#)

Name

Ryan Elmore

Member Since

Aug 02, 2010

9


public repos



3

followers

Following 0 coders and watching 12 repositories [view all](#) →

Public Repositories (9)

 **RDSTK**


R   1


R wrapper for the Data Science Toolkit
Last updated about an hour ago

all commits

commits by owner

52 week participation

 **ProjectTemplate**


 Forked from [johnmyleswhite/ProjectTemplate](#)


A template utility for R projects that provides a skeletal project.
Last updated April 25, 2011



all commits

commits by owner

52 week participation

 **dstk**

 Forked from [petewarden/dstk](#)


Ruby   12


A collection of the best open data sets and open-source tools for data science
Last updated April 19, 2011

all commits


commits by owner

52 week participation


Public Activity 

 **rtelmore pushed to master at rtelmore/RDSTK** about an hour ago


87caf04 modified the help files

 **rtelmore pushed to master at rtelmore/RDSTK** May 11, 2011


e84d3de modified the ip_addys file

 **rtelmore pushed to master at rtelmore/RDSTK** May 11, 2011


e9995f0 added some examples

 **rtelmore pushed to master at rtelmore/RDSTK** May 03, 2011


c7c47c4 error in data on README

 **rtelmore uploaded a file to rtelmore/RDSTK** May 03, 2011

"RDSTK_1.0.tar.gz" is at [rtelmore/RDSTK/downloads](#)
The packaged source code.

 **rtelmore pushed to master at rtelmore/RDSTK** May 03, 2011

f7c85f6 minor stuff

 **rtelmore pushed to master at rtelmore/RDSTK** May 03, 2011

RDSTK:

The R Package for DSTK

Supported functions:

- ★ street2coordinates

- ★ coordinates2politics

- ★ ip2coordinates

- ★ text2sentences

- ★ html2text

- ★ html2story

- ★ text2people

- ★ text2times



getURL



curlPerform

street2coordinates()

```
street2coordinates <- function(address, session = getCurHandle()){  
  api <- "http://www.datasciencetoolkit.org/street2coordinates/"  
  get.addy <- getURL(paste(api, URLEncode(address), sep = ""),  
                    curl = session)  
  result <- ldply(fromJSON(get.addy), data.frame)  
  names(result)[1] <- "full.address"  
  return(result)  
}
```

address can be “5874 Green Dr., Florence, KY”

session allows the user to specify curl parameters

fromJSON and ldply R-ify everything

street2coordinates using the DSTK website

Street Address to Coordinates

API: [/street2coordinates](#)

Street Address to Location calculates the latitude/longitude coordinates for a postal address.

Currently restricted to the US and UK.

Try it for yourself. Copy and paste some addresses into the box below to see what locations it finds.

"2543 Graystone Place, Simi Valley, CA 93065" - 2543 Graystone Pl, Simi Valley, CA, United States at [34.280874,-118.766207](#)

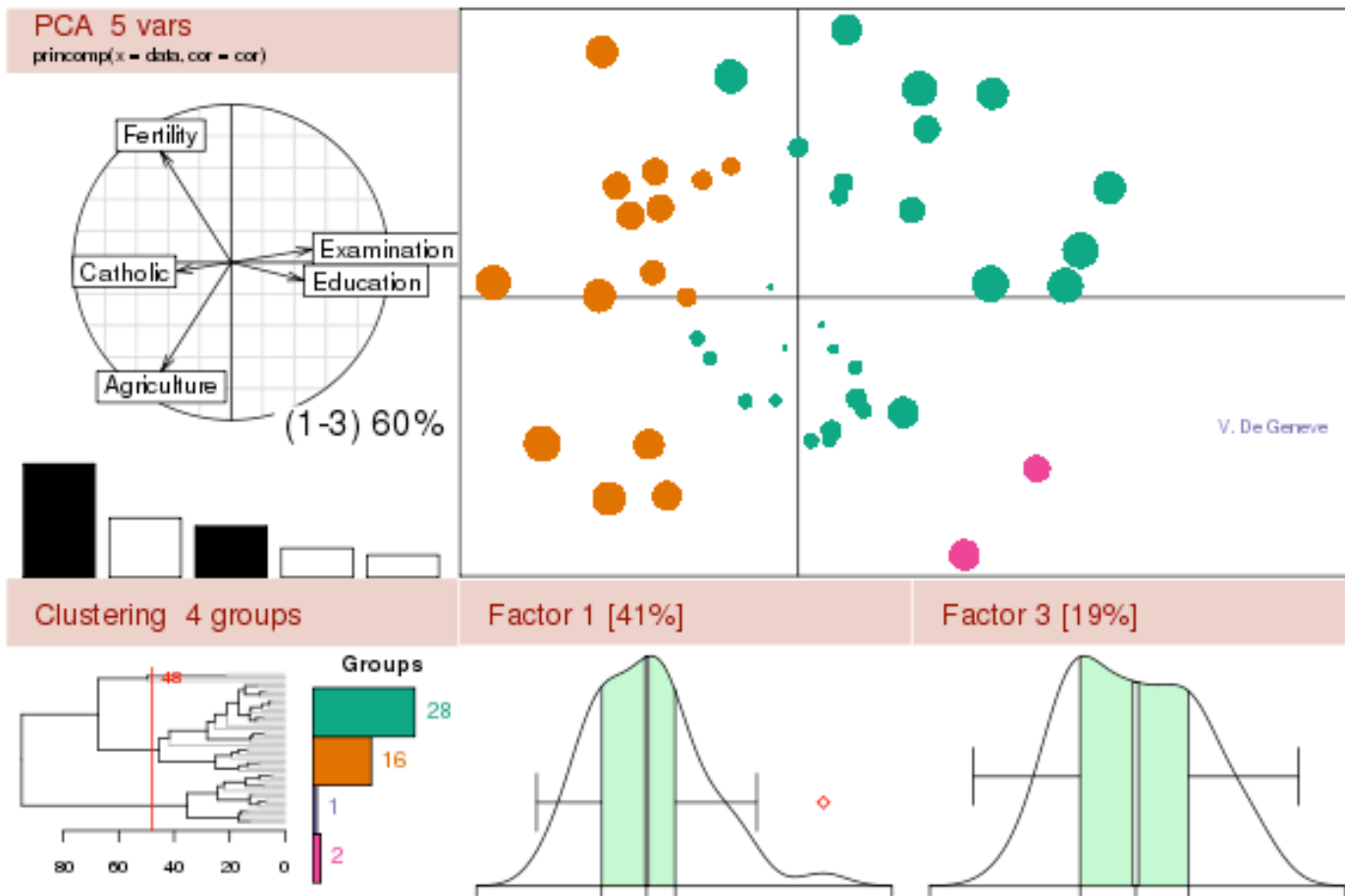
text2people()

```
text2people <- function(text, session=getCurlHandle()) {  
  api <- "http://www.datasciencetoolkit.org/text2people"  
  r = dynCurlReader()  
  curlPerform(postfields=text, url=api, post=1L,  
               writefunction=r$update, curl=session)  
  result <- ldply(fromJSON(r$value()), data.frame)  
  return(result)  
}
```

Similar to previous function in its inputs

Note that we are calling the API using a POST request (postfields=text)

Go To The R Session



Building an R Package

Resources:

- ★ Writing R Extensions
- ★ <http://cran.r-project.org/doc/manuals/R-exts.html>
- ★ Friedrich Leisch's Tutorial

Creating R Packages: A Tutorial

Friedrich Leisch

Department of Statistics, Ludwig-Maximilians-Universität München, and
R Development Core Team, *Friedrich.Leisch@R-project.org*

September 14, 2009

Package Creation

- `package.skeleton(name, list, et al.)` # RTFM
- Essentially, this will create the directory hierarchy, help files, auxiliary files, etc.
- Once everything is in working order, it's a simple R CMD BUILD `package_name` from the command line to create `tar.gz`
- Then R CMD CHECK `package_name`

Description File

Package: RDSTK

Type: Package

Title: An R wrapper for the Data Science Toolkit API

Version: 1.0

Depends: plyr, rjson, RCurl

Date: 2011-04-30

Author: Ryan Elmore

Maintainer: Ryan Elmore <rtelmore@gmail.com>

Description: This package provides an R interface to Pete Warden's Data Science Toolkit. See www.datasciencetoolkit.org for more information. The source code for this package can be found at github.com/rtelmore/RDSTK Happy hacking!

License: BSD

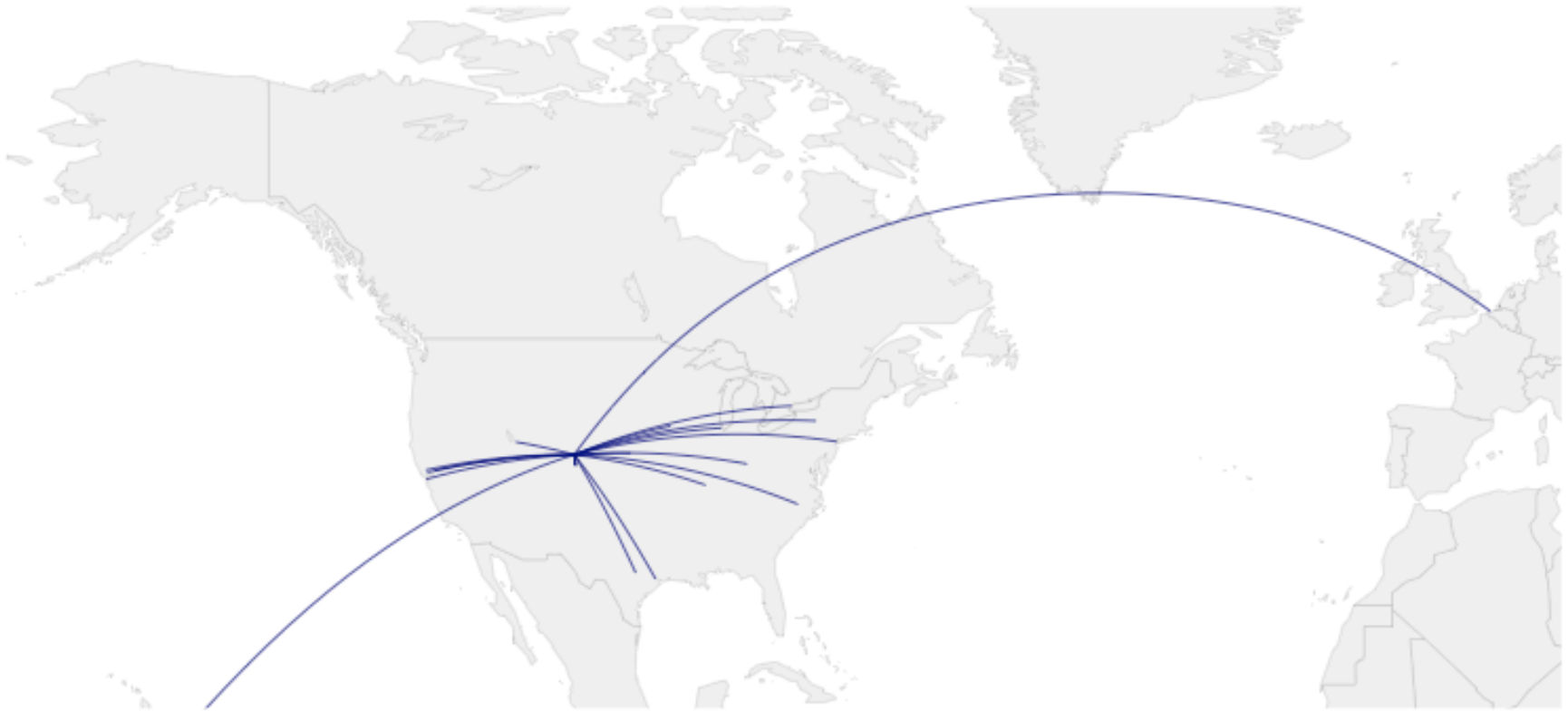
LazyLoad: yes

Data Science in Action!

- I asked everybody where they are from if not Denver.
- 16 or so respondents and I added a few myself
- Inspired by a recent post on FlowingData [3]
- Unfortunately, I asked this before I knew how to use the RDSTK!

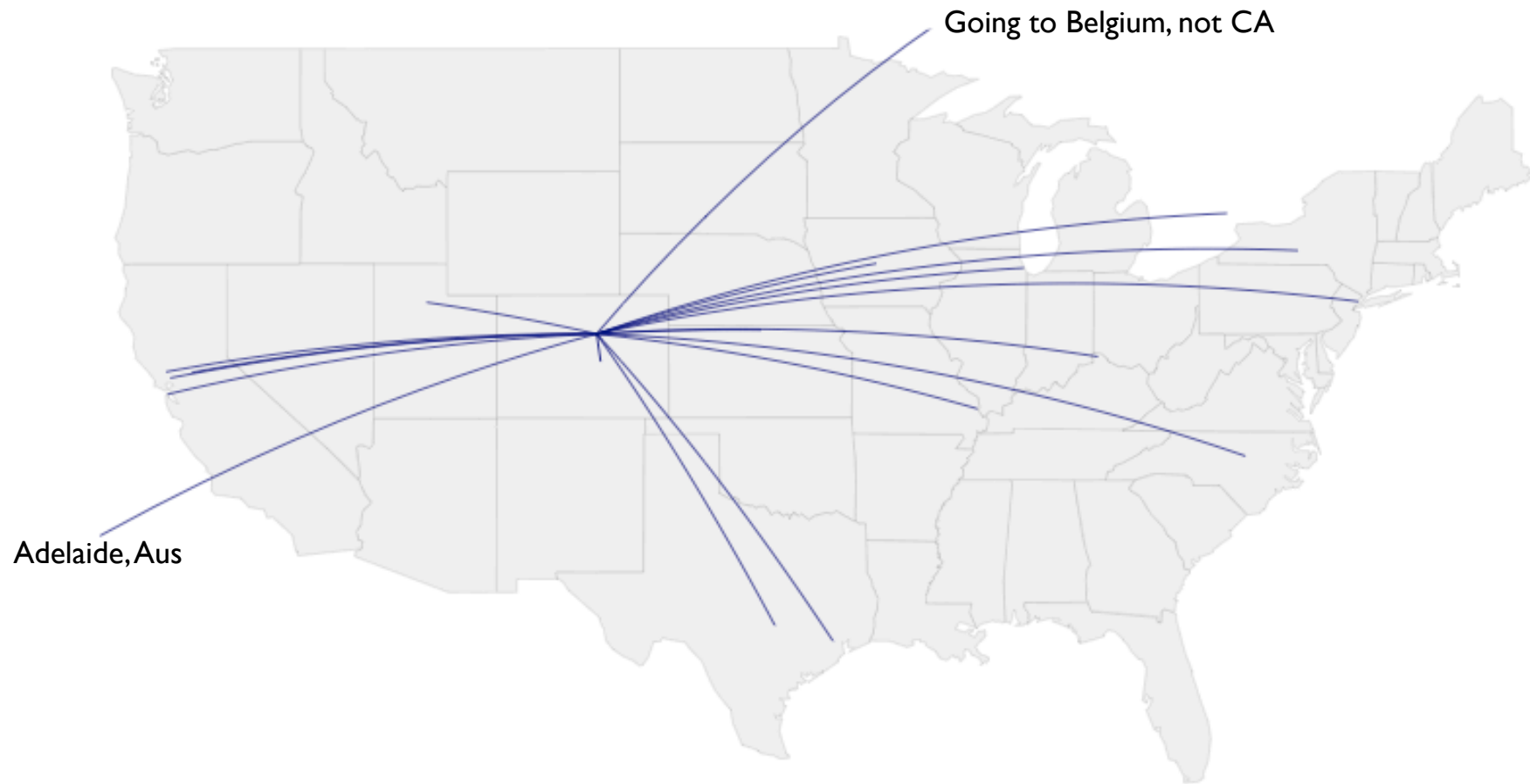
[3] - <http://flowingdata.com/2011/05/11/how-to-map-connections-with-great-circles/>

Where are we from?



Note to @RevoDavid: had to cut out Adelaide for aesthetics!

U.S.-centric Version



Code

- I'll put it on github [4] and my blog [5]
- Packages: maps & geosphere (& ggplot2)

```
map("state", col="#f2f2f2", fill=TRUE, bg="white", lwd=0.15)
for(i in 1:dim(places)[1]){
  inter <- gcIntermediate(denver[2:1], places[i, 5:4], n=50,
                          addStartEnd=TRUE)
  lines(inter, col="navy")
}
```

- I tried using ggplot2, but couldn't limit the lat and long appropriately.

[4] - github.com/rtelmore
[5] - thelogcabin.wordpress.com

Summary

- The RDSTK is available on github; I haven't even thought about packaging it up for CRAN yet.
- Feel free to add city2coordinates; the night's project!
- Unfortunately, the term 'data science' has a lot of traction.
- We all need a raise! :)

Acknowledgments

- Pete Warden for making the DSTK
- Twitter and the #rstats hashtag
- TwitteR and InfoChimps R packages
- Duncan Temple Lang for answering a question on the Rstats-help mailing list
- Andy Gayton [6] and “Noah” from the StackOverflow site.

[6] - Check out staticcloud.com for all your static website hosting needs!