"Big data" skills for online data

Suzy Moat
Data Science Lab
Warwick Business School
The big data explosion

A map of the world built only from GPS locations of Flickr photos
International Passenger Survey
Flickr: global travel tracking

http://dx.doi.org/10.1371/journal.pone.0128470

Barchiesi, Moat, Alis, Bishop & Preis (2015)
Other case studies:
http://www.wbs.ac.uk/about/person/suzy-moat/
Good projects with online data
Good projects with online data

Good question
Good projects with online data

- Good question
- Statistical methodology
Good projects with online data

- Good question
- Statistical methodology
- Access to data
Good projects with online data

Good question

Statistical methodology

Access to data
What's the problem?

Wikipedia article traffic statistics

What do Wikipedia's readers care about? Is Britney Spears more popular than Brittany? Is Asia Carrera more popular than Asia? How many people looked at the article on Santa Claus in December? How many looked at the article on Ron Paul?

What can you find?

Enter a wikipedia article title and press Go

English  

201401  

Go Top

News:

- 2008-02-29: Non-English statistics added (available from February)
- 2008-03-06: Added top list for English Wikipedia. Other projects will be added soon.
- 2008-03-19: Top 1000 articles in February for all projects.

About these stats. The raw data is available here. This is very much a beta service and may disappear or change at any time.
What's the problem?

Wikipedia article traffic statistics

Christmas has been viewed 2345449 times in the last 90 days. This article ranked 6064 in traffic on en.wikipedia.org.

“Christmas”
What's the problem?

Wikipedia article traffic statistics

Christmas has been viewed 2345449 times in the last 90 days. This article ranked 6064 in traffic on en.wikipedia.org.

“This page in json format”
What's the problem?
What's the problem?

Not tabular
What's the problem?

http://stats.grok.se/json/en/latest90/Christmas

Not tabular
What's the problem?

Not tabular

Year: 2013, Month: 12, Page title: Christmas

http://stats.grok.se/json/en/201312/Christmas
What's the problem?

Not tabular

Many pieces

http://stats.grok.se/json/en/201312/Christmas

Year
Month
Page title
What's the problem?

Not tabular

Many pieces

"APIs"
What's the problem?

Not tabular

Many pieces

"APIs"

Easy to overcome – if you can program
Myths about learning to program
Myths about learning to program

Need a computer science degree
Myths about learning to program

Need a computer science degree
Myths about learning to program

- Need a computer science degree
- Takes too long
Myths about learning to program

- Need a computer science degree
- Takes too long
Myths about learning to program

- Need a computer science degree
- Takes too long
- I can't do it – I just get errors
Don’t worry about errors!

Programming questions and answers website
Myths about learning to program

- Need a computer science degree
- Takes too long
- I can't do it – I just get errors
Myths about learning to program

- Need a computer science degree
- Takes too long
- I can't do it – I just get errors
- Can just use graphical interfaces instead
install.packages(c("RCurl", "RJSONIO", "twitteR"))

library(RCurl)

Goes to get data from the Internet for you
install.packages(c("RCurl", "RJSONIO", "twitteR"))

library(RCurl)

library(RJSONIO)

Goes to get data from the Internet for you

Turns "JSON" into a table you can understand
install.packages(c("RCurl", "RJSONIO", "twitteR"))

library(RCurl)

library(RJSONIO)

library(twitteR)

Goes to get data from the Internet for you

Turns "JSON" into a table you can understand

Goes to get data from Twitter for you
install.packages(c("RCurl", "RJSONIO", "twitteR"))

library(RCurl)

library(RJSONIO)

library(twitteR)

Goes to get data from the Internet for you

Turns "JSON" into a table you can understand

Goes to get data from Twitter for you

... and lots more!
Myths about learning to program

- Need a computer science degree
- Takes too long
- I can't do it – I just get errors
- Can just use graphical interfaces instead

Takes too long

Can just use graphical interfaces instead
Myths about learning to program

- Need a computer science degree
- Takes too long
- I can't do it – I just get errors
- Can just use graphical interfaces instead
- Only of use for new forms of data
What to learn: R and RStudio

http://www.r-project.org/
http://www.rstudio.com/
What to learn: Python

Python
Continue your learning by starting Python.
Start

https://www.codecademy.com/learn/python
R Programming

There are hundreds of websites that can help you learn R. Here's how you can use some of the best to become a productive R programmer.

Start by downloading R and RStudio.

Learn the basics

Visit DataCamp or Try R to learn how to write basic R code. Both sites provide interactive lessons that will get you writing real code in minutes. They are a great place to make mistakes and test out new skills. You are told immediately when you go wrong and given a chance to fix your code.

Or try Leada. Leada gets you programming in your own environment with videos and exercises. The first few courses are free and cover how to install R and RStudio.
Big Data: Measuring and Predicting Human Behaviour

Join us to explore how the vast amounts of data generated today can help us understand and even predict how humans behave.

Go to course – started 20 Apr

https://www.futurelearn.com/courses/big-data

Over 25,000 learners to date

Runs again in 2017
Programming skills will unlock huge volumes of interesting social data for you – great return on investment