

# Biostats 597E

Week 9 - Introduction to Web Scraping

# Web Request

Web request method:

- **GET:** Requests data from a specified resource. It is what you use when visiting a webpage (focus more on this method)
- **POST** Submits data to be processed to a specified resource. For example, registering new user.

# Web API

- Using web request to obtain data
- It provides nice and convenient interface
- It usually return well structured data in JSON or XML format
- The data is much easier to process

Example: <http://freegeoip.net>

```
library(httr)
res <- GET("http://freegeoip.net/csv/50.78.253.58")
content(res, "text")
```

# Authentication

Some web APIs do not use authentication, so anyone can make API call for free.

Some APIs require authentication, such as Twitter, to charge money or limit usage.

# Obtain Twitter Authentication

- Register Twitter account
- Go to <https://apps.twitter.com>
- Create a New App and fill the information

 Application Management



## Create an application

### Application Details

**Name \***

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

**Description \***

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

**Website \***

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens.  
(If you don't have a URL yet, just put a placeholder here but remember to change it later.)

**Callback URL**

Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth\_callback URL on the request token step, regardless of the value given here. To restrict your application from using callbacks, leave this field blank.

# Obtain Twitter Authentication

- Go to "Keys and Access" tab to obtain the API key and API secret
- Also create access token

## biostat597e

[Details](#) [Settings](#) [Keys and Access Tokens](#) [Permissions](#)

### Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	3yQN[REDACTED]qKhbilu
Consumer Secret (API Secret)	kfT6bbk[REDACTED]uD6oWzJPEbJFC6vtPIKomqxV
Access Level	Read and write ( <a href="#">modify app permissions</a> )
Owner	ustcgxdgu
Owner ID	960801247

### Application Actions

[Regenerate Consumer Key and Secret](#) [Change App Permissions](#)

# User TwitterR Package

see `Twitter.R` for example

# Job Search API Using indeed.com

indeed.com provides easy to use API to do your job search without scraping the page

First, we need obtain authentication

- Go to <http://www.indeed.com/publisher>
- Create an publisher account
- Log in account and go to 'XML Feed' to see your publisher id

See indeed.R for example