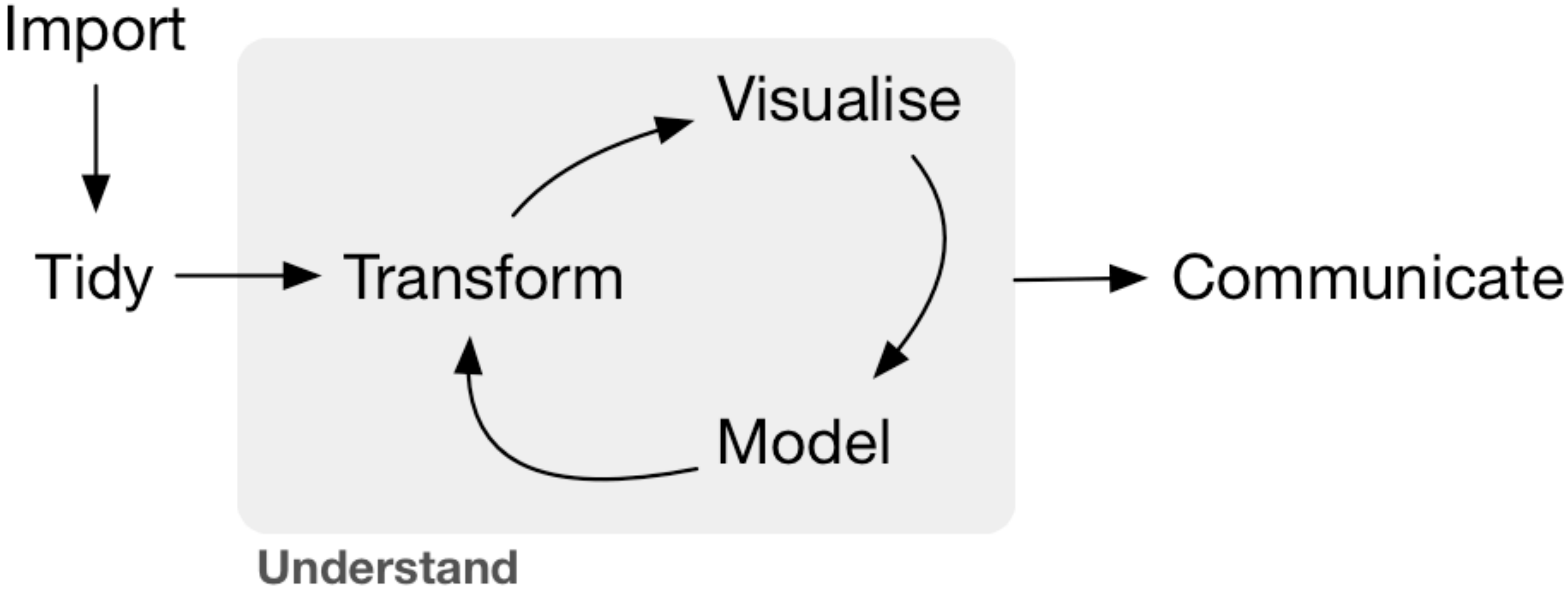
COMP10020
Introduction to Programming II
Importing Data For Data Science

Dr. Brian Mac Namee
brian.macnamee@ucd.ie
School of Computer Science
University College Dublin



https://ismayc.github.io/moderndiver-book/2-intro.html

ACCESSING DATA FROM TEXT FILES

Accessing Data From Text Files

Nothing complicated here, just load your data

You need to do all of the work to make sense of it

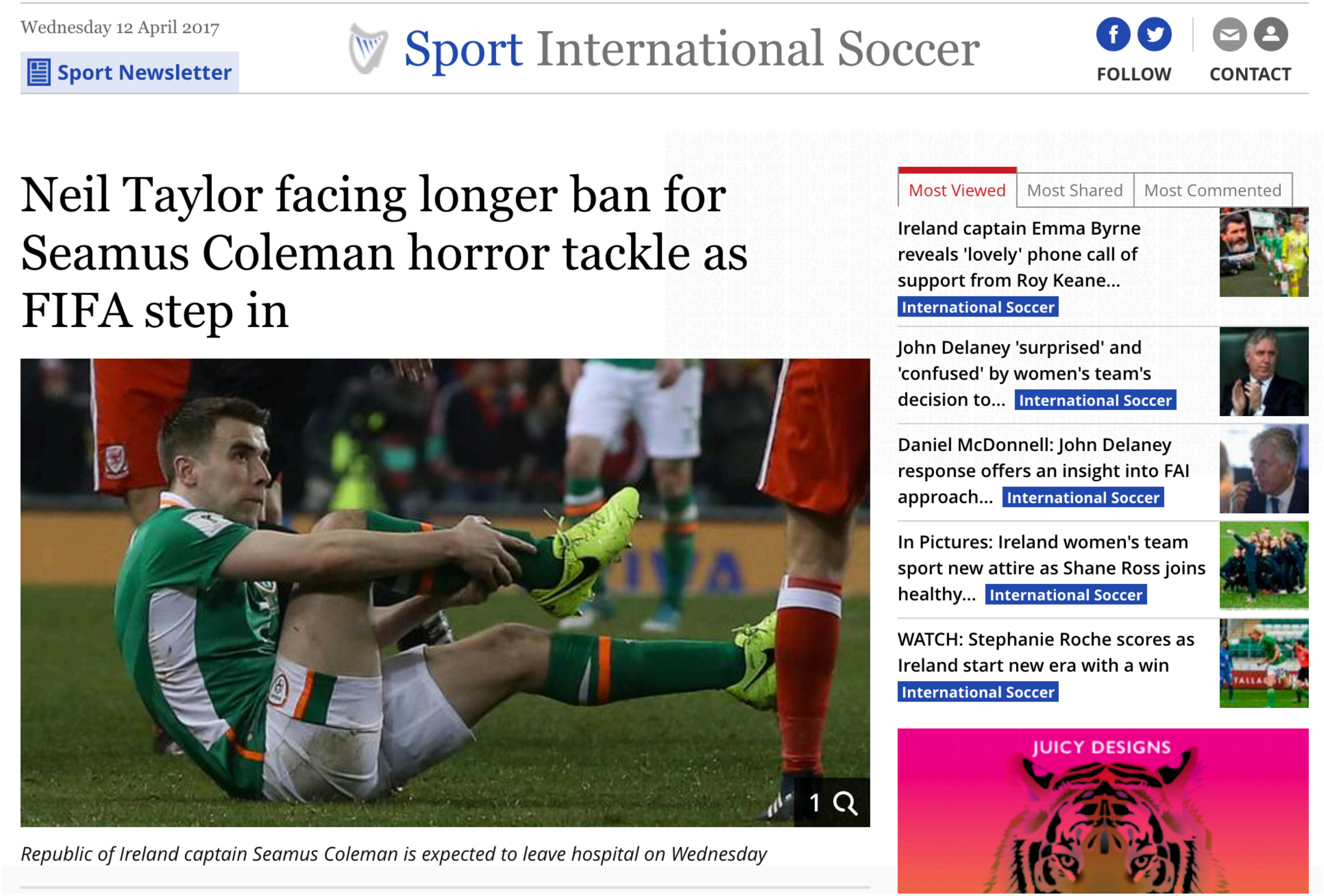
PARSING HTML FILES

Parsing HTML Files

HTML on webpages is nice and structured

```
<html>
  <head>
    <title>My HTML Page</title>
  </head>
  <body>
    <h1>My First Page</h1>
    <p>This is a page of text</p>
  </body>
</html>
```

Real Pages Are Complicated!



Wednesday 12 April 2017

Sport Newsletter

Sport International Soccer

FOLLOW

CONTACT

Neil Taylor facing longer ban for Seamus Coleman horror tackle as FIFA step in

1

Republic of Ireland captain Seamus Coleman is expected to leave hospital on Wednesday

Most Viewed | Most Shared | Most Commented

Ireland captain Emma Byrne reveals 'lovely' phone call of support from Roy Keane... International Soccer

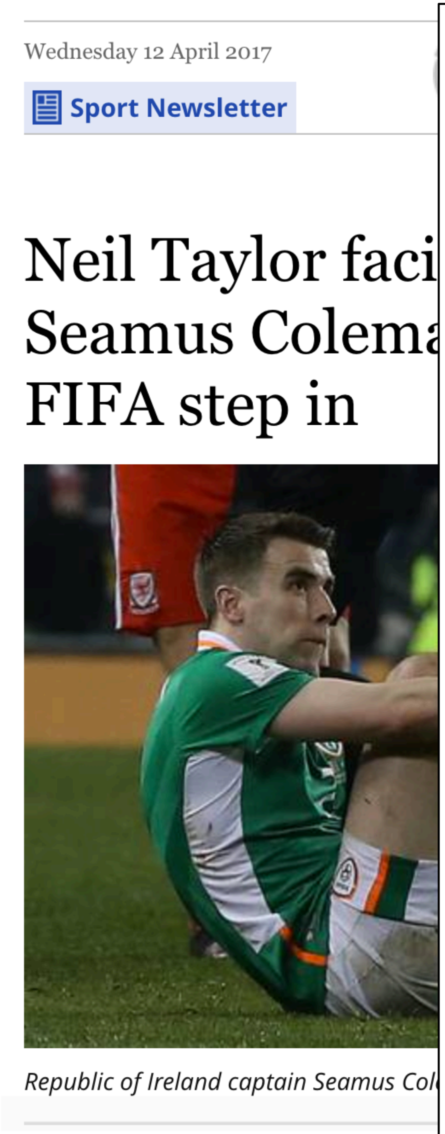
John Delaney 'surprised' and 'confused' by women's team's decision to... International Soccer

Daniel McDonnell: John Delaney response offers an insight into FAI approach... International Soccer

In Pictures: Ireland women's team sport new attire as Shane Ross joins healthy... International Soccer

WATCH: Stephanie Roche scores as Ireland start new era with a win International Soccer

Real Pages Are Complicated!



Wednesday 12 April 2017

Sport Newsletter

Neil Taylor faci
Seamus Colema
FIFA step in

Republic of Ireland captain Seamus Col

```
<!DOCTYPE html>
<html lang="en" xmlns:og="http://opengraphprotocol.org/
schema/" xmlns:fb="http://www.facebook.com/2008/fbml">
<!-- generated by inm-prod-presentation-engine-spot-174.inm.lan
2017-04-12T20:47:17.472 -->
<head>
<script type='text/javascript'>var _sf_startpt=(new
Date()).getTime()</script>
<meta http-equiv="X-UA-Compatible" content="IE=edge" />
<meta http-equiv="Content-Type" content="text/html;
charset=UTF-8"/>
<link rel="canonical" href="http://www.independent.ie/sport/
soccer/international-soccer/neil-taylor-facing-longer-ban-for-
seamus-coleman-horror-tackle-as-fifa-step-in-35578919.html" >
<link rel="alternate" media="only screen and (max-width:
640px)" href="http://m.independent.ie/sport/soccer/international-
soccer/neil-taylor-facing-longer-ban-for-seamus-coleman-horror-
tackle-as-fifa-step-in-35578919.html" >
```

OVER 4,000 LINES

BeautifulSoup

BeautifulSoup is a Python package that allows us to parse HTML

Two key functions:

- **find** find the first occurrence of a particular type of tag
- **find_all** find all occurrences of a particular kind of tag

Tags can be defined by type, style, class, ...

The "Rules" of Web Scraping

The "rules" of web scraping:

- Check a site's terms and conditions before you scrape their pages.
- Do not hammer a site with too many automated requests.
- Sites often change their layout, so scrapers often break and need to be re-written.

(*from Derek Greene*)

USING RSS FEEDS

Using RSS Feeds

RSS feeds are commonly used for news syndication

- For example: https://www.irishtimes.com/cmlink/news-1.1319192
- RSS feeds list the latest articles on a given site and are frequently updated

RSS feeds are provided in XML format

```
<?xml version="1.0"?>
<rss version="2.0">
	<channel>
		<title><![CDATA[The Irish Times - News]]></title>
		<lastBuildDate>Wed, 12 Apr 2017 20:15:05 +0000</lastBuildDate>
		<item>
			<title><![CDATA[Gardaí release suspect in investigation into Real IRA murder]]></title>
			<link>http://www.irishtimes.com/news/crime-and-law/garda%C3%AD-release-suspect-in-investigation-into-real-ira-murder-1.3046748</link>
			<description><![CDATA[Former dissident republican Aidan O’Driscoll was shot dead in Cork late last year ]]></description>
			<pubDate>Wed, 12 Apr 2017 20:15:05 +0000</pubDate>
		</item>
		<item>
			<title><![CDATA[Childcare subsidies for 9,000 families delayed over IT issues]]></title>
			<link>http://www.irishtimes.com/news/social-affairs/childcare-subsidies-for-9-000-families-delayed-over-it-issues-1.3046461</link>
			<description><![CDATA[Increase in support will not be delivered in September but is expected in new year ]]></description>
			<pubDate>Wed, 12 Apr 2017 20:01:39 +0000</pubDate>
		</item>
		...
	</channel>
</rss>
```

Using RSS Feeds

We can access RSS feeds easily in Python using the **feedparser** package

- **parse** method parses a URL to an RSS feed

Iterate across the entries in the feed and use the structure of the entries to pull out key information

ACCESSING TWITTER DATA

Accessing Twitter Data

We can use the **Tweepy** package to access the Twitter API.

- Before using Tweepy you must have Twitter **OAuth credentials** available from https://apps.twitter.com/
- Create a new application (using your own Twitter credentials) and the generate access tokens.

Accessing Twitter Data

There are 3 ways in which we can access Twitter:

- **Rest API** allows access to a Twitter account from code and programmatic posting, retweeting etc
- **Search API** allows searches of recent Twitter posts
- **Streaming API** allows a live connection to Twitter

NLTK

NLTK

The Natural Language Toolkit, NLTK (http://www.nltk.org/), is a well written, widely used, and well respected toolkit for perofmring natural langueg processing in Python.

It offers a wide range of useful functionality and data structres that make text natural langueg processing, and so text analytics, much easier.

NLTK

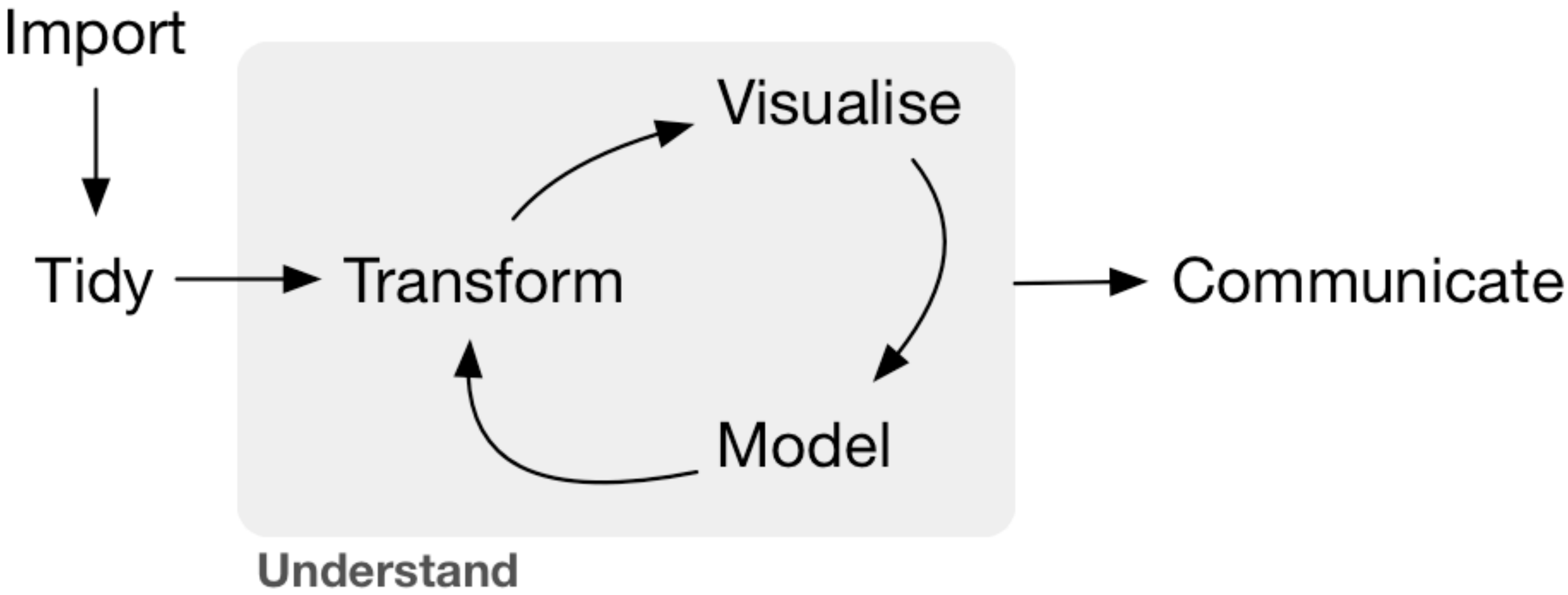
Features included in the NLTK include

- corpus management
- document classification
- colocation discovery
- part of speech tagging
- parsing
- chunking

The best reference for the NLTK is *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit* by Steven Bird, Ewan Klein, and Edward Loper

- Freely available online at http://www.nltk.org/book/

SUMMARY



https://ismayc.github.io/moderndiver-book/2-intro.html

Summary

Data is the fuel of data science

There are lots of data sources out there that we can easily connect to

Finding the right APIs and python packages are key to doing this well