Data Collection from Syndicated Web Sources - APIs

Session 4 @ CBA Batch 12
April 2019

sudhir_voleti@isb.edu

1

Session Outline

- Session 2 Continued: Webscraping Basics
  - [1] Atis.org revisited [2] Scraping Cricbuzz [3] Scraping Amazon Reviews
- Py Scraping Google search results
- Scraping Dynamic pages with Webdrivers
  - [1] Scraping python.org
- Intro to DC from APIs
  - DarkSky weather API, openMovieDB API
- DC from Finance APIs – Quandl

2

Session 2 Webscraping Basics

Continued

3

Atis Glossary Scraping revisited

- **Background:**
- Atis.org website's changed in the past few months. Old codes weren't working.
- So I and my RA, Ms Anisha both rewrote the codes in parallel.
- LMS contained her version which was different from mine.
- Let me briefly walk you through my (simpler) version.
- Aim is to demo the use of user-defined functions in webscraping
- Pls open: **Webscraping basics with rvest.Rmd**

4

Atis Glossary Scraping Recap

- First 2 steps were common to what we did previously.
- Step 3 onwards we see a change.
- We wrote 1 function for 1 letter and looped over all letter links.
- More generally:
- We write a functions (for 1 link) but that link contains a set of other (secondary) links.
- So we write another functions for those secondary links and so on.
- We'll use this approach going forward in the next 2 examples.

5

Scraping IPL 2019 match schedules from Cricbuzz

- Open: **Another rvest exercise - cricbuzz.Rmd**
- The webpage looks like this - what all attributes can we easily scrape?
- Use SelectorGadget and ID the tags for those attributes. For example:
- See the photos tab on the page. How to scrape images?
- Exercise done. Learnings? Applications? Etc.

6

Scraping Structured Webpages - Amazon reviews

- Why might anyone want to scrape product reviews from Amazon?
- Recall what structured webpages are.
- Consider how we discerned structure in Amazon review URLs.
- Then we wrote a function to scrape one page.
- We constructed URLs for required page range, and looped our func over it.
- Exercise done. Learnings? Applications? Etc.

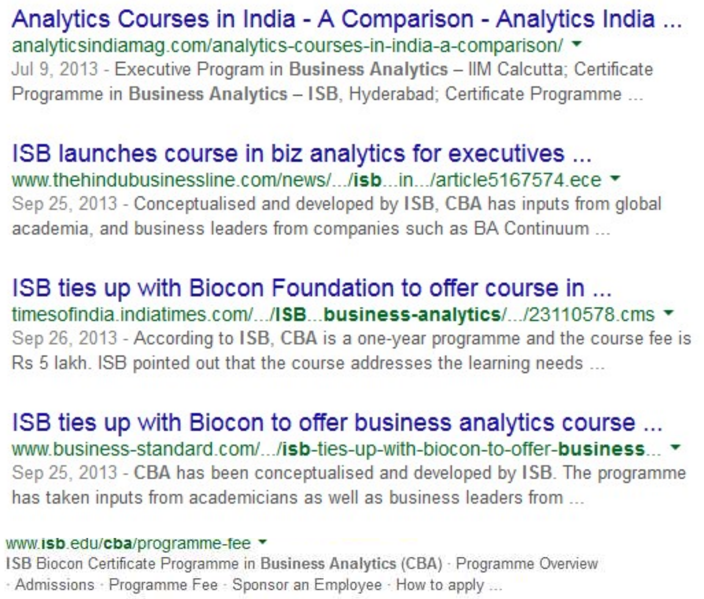
7

Py-scraping Google Search Results

- Suppose you're running a kebab joint… You want to know where you stand vis-à-vis local (city based) competition.
- Maybe you want to know what shows up when someone Google-searches for, say, 'Kebab Hyderabad' etc …
- What pages would show up, in which order, what content would they contain, what sentiment would they express, etc.
- *Why* focus on Google search only? *
- And what all can you do once that data are in?*
- Python makes it super-simple to scrape Google search results → save it as a structured object → that you can analyze in R (for example).

8

Google Search Content for ‘ISB CBA’



Analytics Courses in India - A Comparison - Analytics India ...
analyticsindiamag.com/analytics-courses-in-india-a-comparison/ ▾
Jul 9, 2013 - Executive Program in **Business Analytics** – IIM Calcutta; Certificate Programme in **Business Analytics** – **ISB**, Hyderabad; Certificate Programme ...

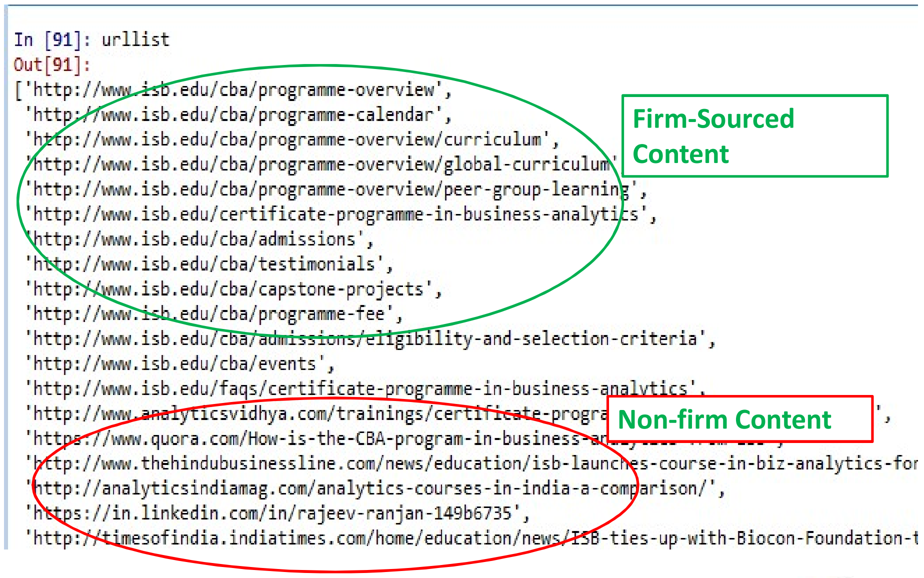
ISB launches course in biz analytics for executives ...
www.thehindubusinessline.com/news/.../**isb**...in.../article5167574.ece ▾
Sep 25, 2013 - Conceptualised and developed by **ISB**, **CBA** has inputs from global academia, and business leaders from companies such as BA Continuum ...

ISB ties up with Biocon Foundation to offer course in ...
timesofindia.indiatimes.com/.../**ISB**...**business-analytics**/.../23110578.cms ▾
Sep 26, 2013 - According to **ISB**, **CBA** is a one-year programme and the course fee is Rs 5 lakh. ISB pointed out that the course addresses the learning needs ...

ISB ties up with Biocon to offer business analytics course ...
www.business-standard.com/.../**isb**-ties-up-with-biocon-to-offer-**business**... ▾
Sep 25, 2013 - **CBA** has been conceptualised and developed by **ISB**. The programme has taken inputs from academicians as well as business leaders from ...

www.**isb**.edu/**cba**/programme-fee ▾
ISB Biocon Certificate Programme in **Business Analytics** (**CBA**) · Programme Overview · Admissions · Programme Fee · Sponsor an Employee · How to apply ...

Extracting Google Search Results in Python

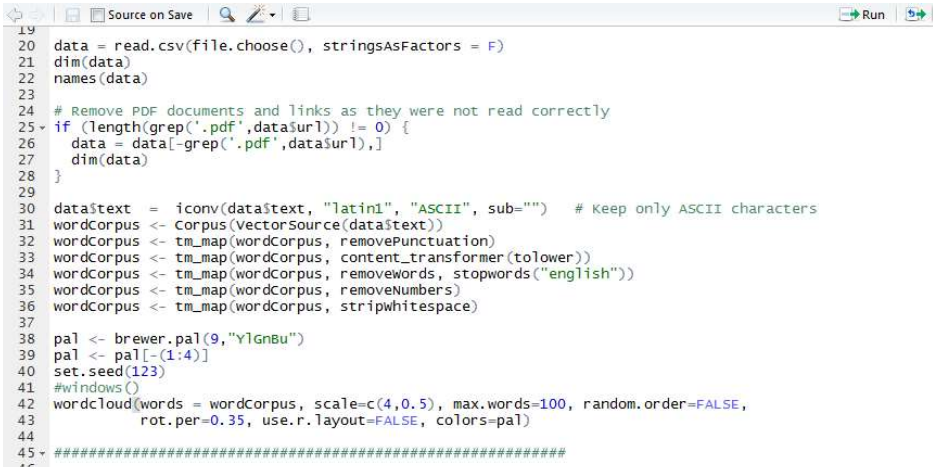


```
In [91]: urllist
Out[91]:
['http://www.isb.edu/cba/programme-overview',
 'http://www.isb.edu/cba/programme-calendar',
 'http://www.isb.edu/cba/programme-overview/curriculum',
 'http://www.isb.edu/cba/programme-overview/global-curriculum'
 'http://www.isb.edu/cba/programme-overview/peer-group-learning',
 'http://www.isb.edu/certificate-programme-in-business-analytics',
 'http://www.isb.edu/cba/admissions',
 'http://www.isb.edu/cba/testimonials',
 'http://www.isb.edu/cba/capstone-projects',
 'http://www.isb.edu/cba/programme-fee',
 'http://www.isb.edu/cba/admissions/eligibility-and-selection-criteria',
 'http://www.isb.edu/cba/events',
 'http://www.isb.edu/faqs/certificate-programme-in-business-analytics',
 'http://www.analyticsvidhya.com/trainings/certificate-progra                           ',
 'https://www.quora.com/How-is-the-CBA-program-in-business-an
 'http://www.thehindubusinessline.com/news/education/isb-launches-course-in-biz-analytics-for
 'http://analyticsindiamag.com/analytics-courses-in-india-a-comparison/',
 'https://in.linkedin.com/in/rajeev-ranjan-149b6735',
 'http://timesofindia.indiatimes.com/home/education/news/ISB-ties-up-with-Biocon-Foundation-t
```

Firm-Sourced Content

Non-firm Content

Basic Text Analysis on the Scraped Text in R



```
data = read.csv(file.choose(), stringsAsFactors = F)
dim(data)
names(data)

# Remove PDF documents and links as they were not read correctly
if (length(grep('.pdf',data$url)) != 0) {
  data = data[-grep('.pdf',data$url),]
  dim(data)
}

data$text  =  iconv(data$text, "latin1", "ASCII", sub="")   # Keep only ASCII characters
wordCorpus <- Corpus(VectorSource(data$text))
wordCorpus <- tm_map(wordCorpus, removePunctuation)
wordCorpus <- tm_map(wordCorpus, content_transformer(tolower))
wordCorpus <- tm_map(wordCorpus, removeWords, stopwords("english"))
wordCorpus <- tm_map(wordCorpus, removeNumbers)
wordCorpus <- tm_map(wordCorpus, stripWhitespace)

pal <- brewer.pal(9,"YlGnBu")
pal <- pal[-(1:4)]
set.seed(123)
#windows()
wordcloud(words = wordCorpus, scale=c(4,0.5), max.words=100, random.order=FALSE,
          rot.per=0.35, use.r.layout=FALSE, colors=pal)

############################################################
```

11

Webdrivers for Web-scraping

12

Using Selenium in Py

- Open the HTML file 'Intro to webdrivers - Selenium in Py'
- Let's walk though it step by step.
- What I'll show next is fairly basic. However, ...
- If you're aware of alternatives, better ways to do the same thing etc., pls speak up and share with the class.

13

Using Webdrivers: Recap

- What are webdrivers and where are they most used?
- What modules did we invoke for using webdrivers?
- What main functions were called? What did they do?
- What further possibilities come to mind with webdriver use?
- Ready for some basic homework involving py and selenium?

14

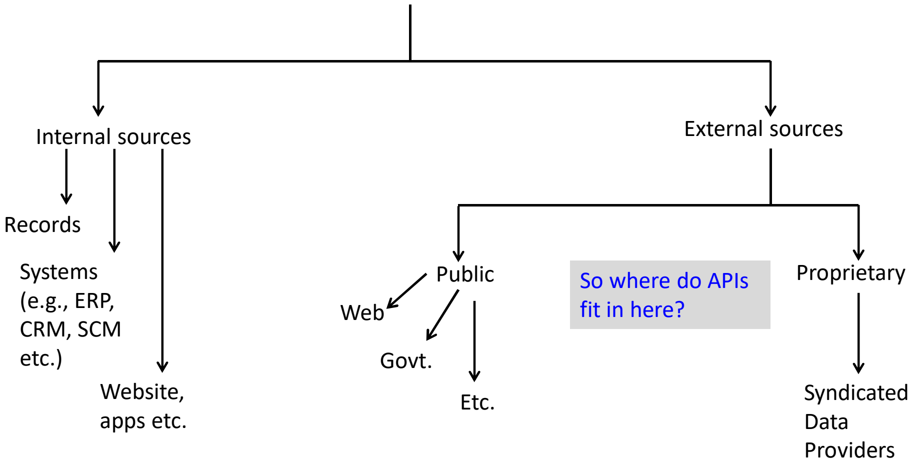
Some API

Preliminaries

15

Some Common Secondary Data Sources

- Let’s quickly organize common secondary data sources:



Internal sources

Records

Systems (e.g., ERP, CRM, SCM etc.)

Website, apps etc.

External sources

Public

Web

Govt.

Etc.

So where do APIs fit in here?

Proprietary

Syndicated Data Providers

16

APIs: Some Preliminaries

- What is an API?

  Application Programming Interfaces are *interfaces* between 2 *services* …

- Examples?

- “There’s an API for that”.

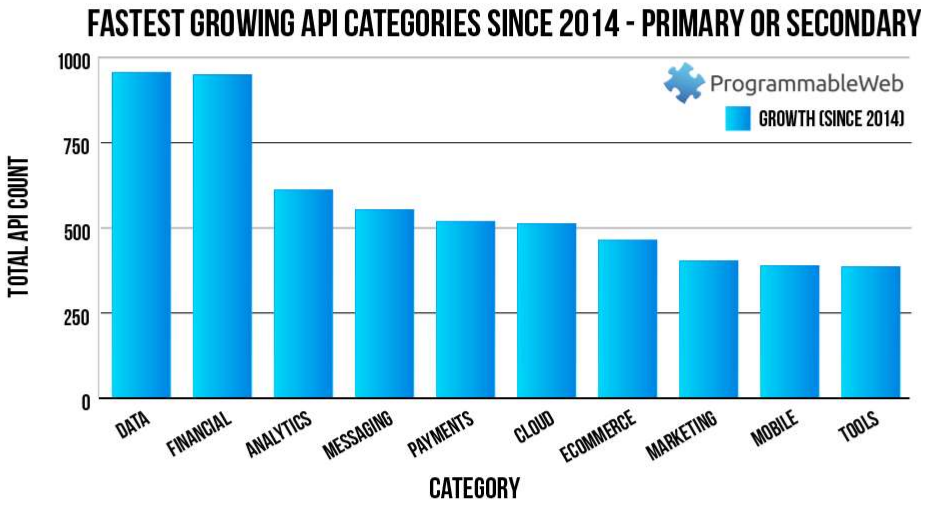
- *Why* do firms like FB or Google put out APIs?

  Potential source of revenue since data is currency→ monetize the data asset; Invite developers to deploy cool stuff through their platform

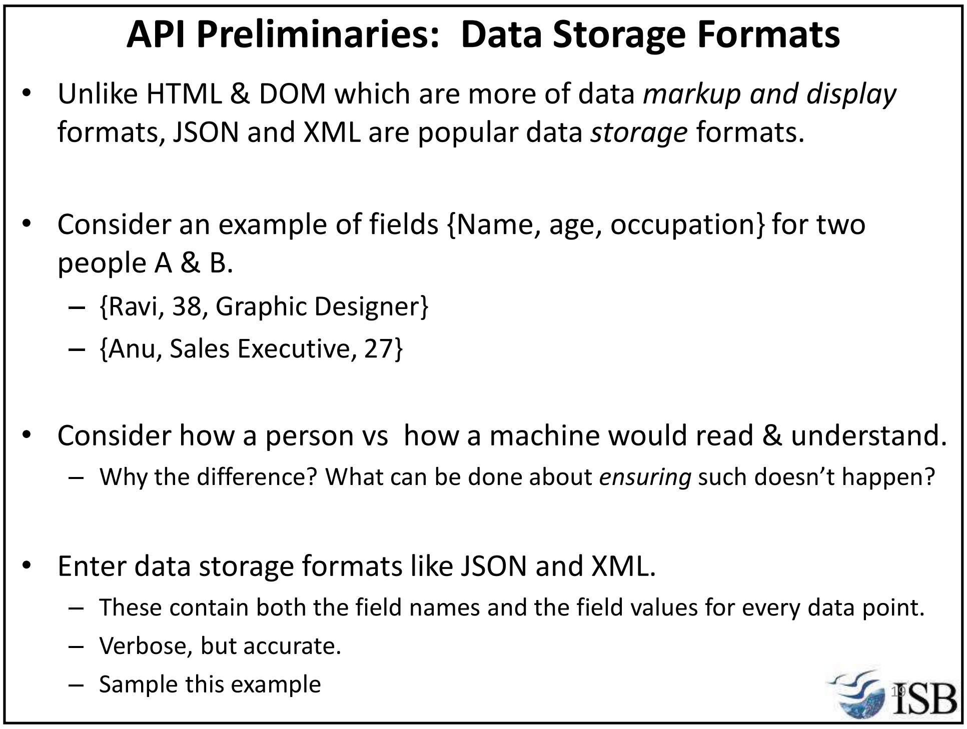
- In which domains might APIs be likely to be found? How many might be there?

17

APIs: The Growth Story across Domains



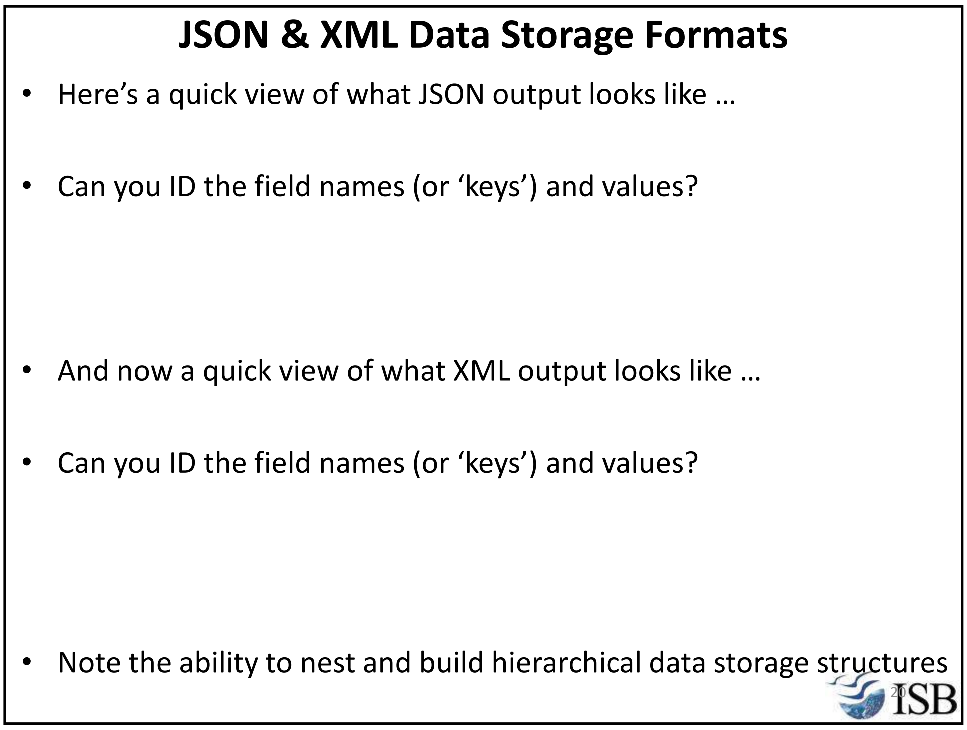
18



API Preliminaries: Data Storage Formats

- Unlike HTML & DOM which are more of data *markup and display* formats, JSON and XML are popular data *storage* formats.
- Consider an example of fields {Name, age, occupation} for two people A & B.
  - {Ravi, 38, Graphic Designer}
  - {Anu, Sales Executive, 27}
- Consider how a person vs how a machine would read & understand.
  - Why the difference? What can be done about *ensuring* such doesn't happen?
- Enter data storage formats like JSON and XML.
  - These contain both the field names and the field values for every data point.
  - Verbose, but accurate.
  - Sample this example

19



JSON & XML Data Storage Formats

- Here's a quick view of what JSON output looks like …
- Can you ID the field names (or 'keys') and values?
- And now a quick view of what XML output looks like …
- Can you ID the field names (or 'keys') and values?
- Note the ability to nest and build hierarchical data storage structures

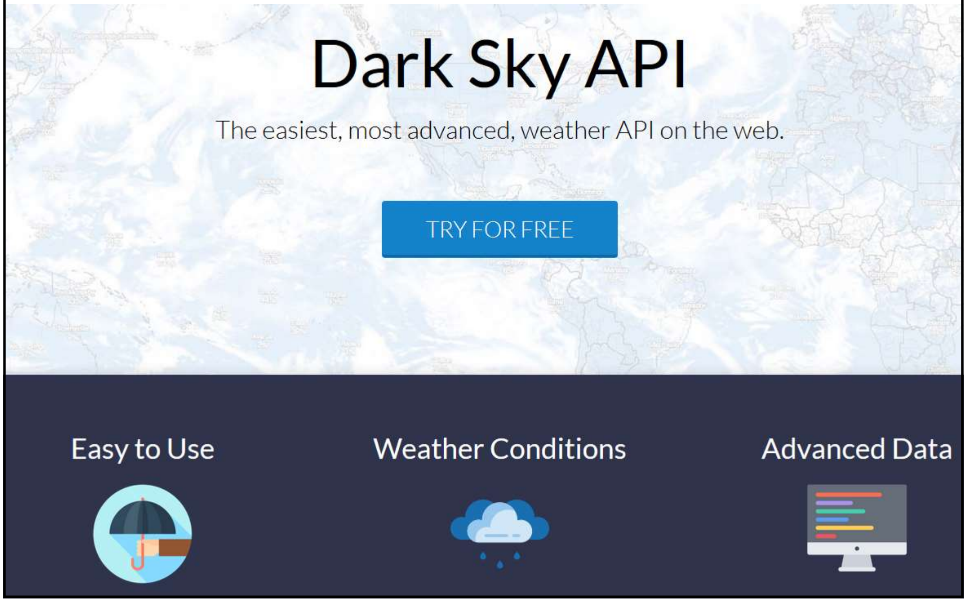
20

Intro to DC from APIs

21

DC from APIs: A Simple Weather API…

- https://darksky.net/dev



Dark Sky API

The easiest, most advanced, weather API on the web.

22

DC from APIs: A Simple Weather API…

- Who in business might want to use a weather API? Why?
- Examine what the documentation says:
- What the API gives in terms of output data fields
- Which among those fields are required vs optional
- How to construct a query for those fields
- Some sample output
- Let’s run the queries we’ve built and examine output

23

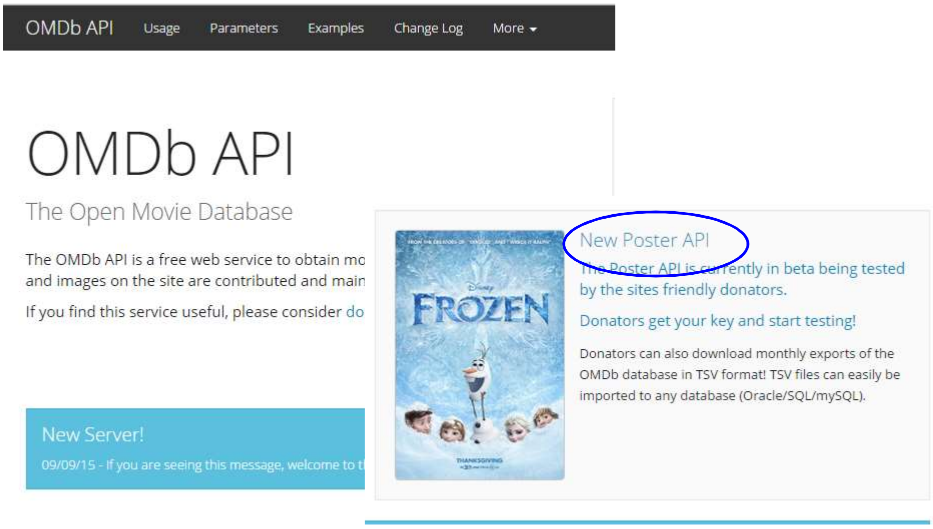
Recapping our Weather API journey

- What is an API? How to connect to one?
- How to know what all data fields are available from an API?
- How to know what the API query construction should be like?
- How to know the *pricing* of data and/or services?
- What *after* getting the data from an API?

24

Data from APIs: A Simple Movies DB API…

- http://www.omdbapi.com/



OMDb API

The Open Movie Database

The OMDb API is a free web service to obtain mo
and images on the site are contributed and main

If you find this service useful, please consider do

New Server!

09/09/15 - If you are seeing this message, welcome to t

New Poster API

The Poster API is currently in beta being tested by the sites friendly donators.

Donators get your key and start testing!

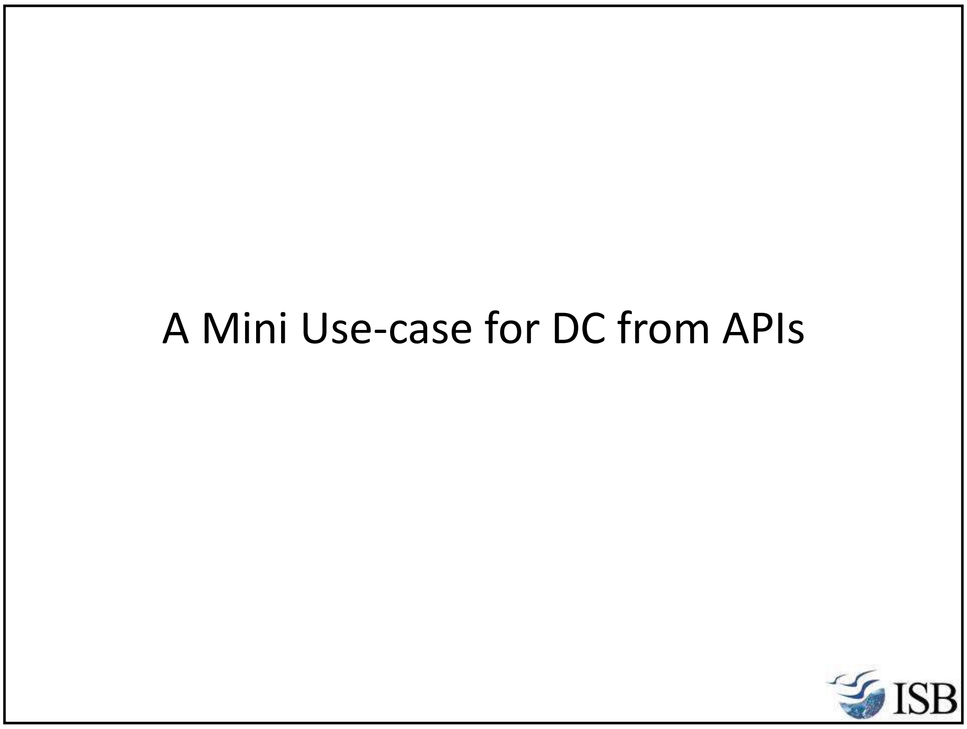
Donators can also download monthly exports of the OMDb database in TSV format! TSV files can easily be imported to any database (Oracle/SQL/mySQL).

25

Data from APIs: A Simple Movies DB API…

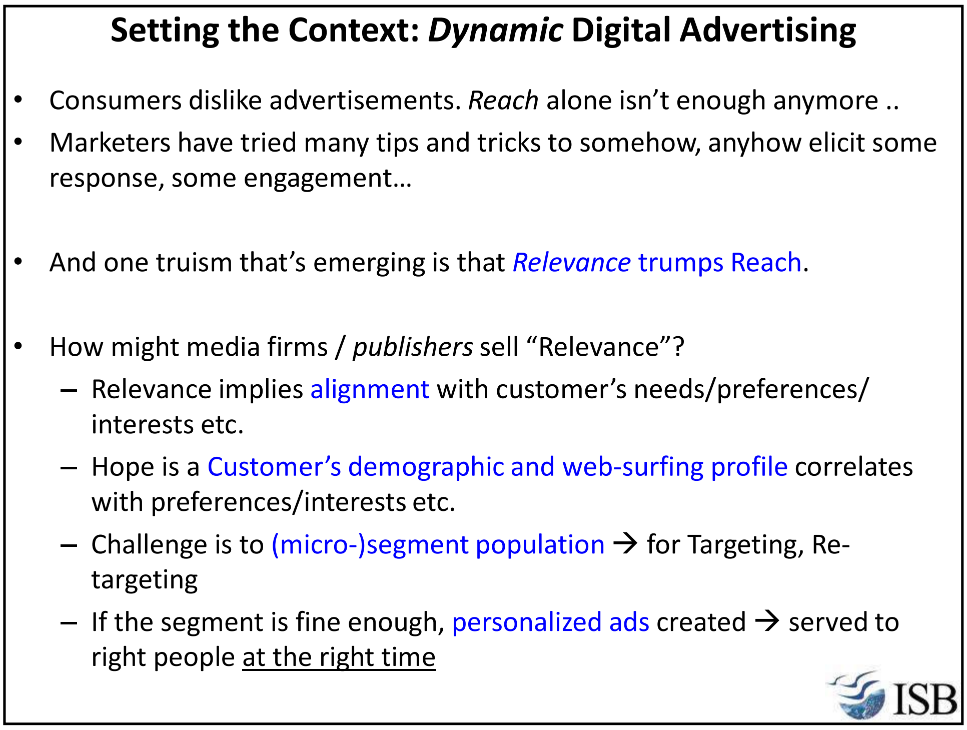
- Goto http://www.omdbapi.com/ and explore the page a bit …
- Qs to ponder upon:
- How is the API key delivered? What are the usage restrictions?
- How does this site make money?
- What does their legal section say?
- Open the file ‘’ and browse through the results of an API query on omdb

26



A Mini Use-case for DC from APIs

27



Setting the Context: *Dynamic* Digital Advertising

- Consumers dislike advertisements. *Reach* alone isn't enough anymore ..
- Marketers have tried many tips and tricks to somehow, anyhow elicit some response, some engagement…
- And one truism that's emerging is that *Relevance* trumps Reach.
- How might media firms / *publishers* sell "Relevance"?
  - Relevance implies alignment with customer's needs/preferences/ interests etc.
  - Hope is a Customer's demographic and web-surfing profile correlates with preferences/interests etc.
  - Challenge is to (micro-)segment population → for Targeting, Re-targeting
  - If the segment is fine enough, personalized ads created → served to right people at the right time

28

A Simple *Dynamic* Advertising Campaign

- At its simplest, dynamic ad campaigns require **data signals** (from APIs), a set of *event rules* or *"trigger" conditions* and ad copy/messaging to go along.
- Consider this illustrative example…

Brands	**Some data signals received from APIs…**					
	Weather	Profile	Sports	Entertainment	Geography	Social
Audi	✓	✓	✓	✓		✓
Lipitor		✓	✓	✓	✓	
L'Oreal	✓	✓		✓	✓	✓
Wells fargo		✓			✓	✓
Starbucks	✓					✓

Discretion and judgment are important. Some sports (say, football) gel better with some brands (Bud) than others.

- *What* events might trigger *which* ad copies to get served to *whom*?

29

A Simple Dynamic Advertising Campaign

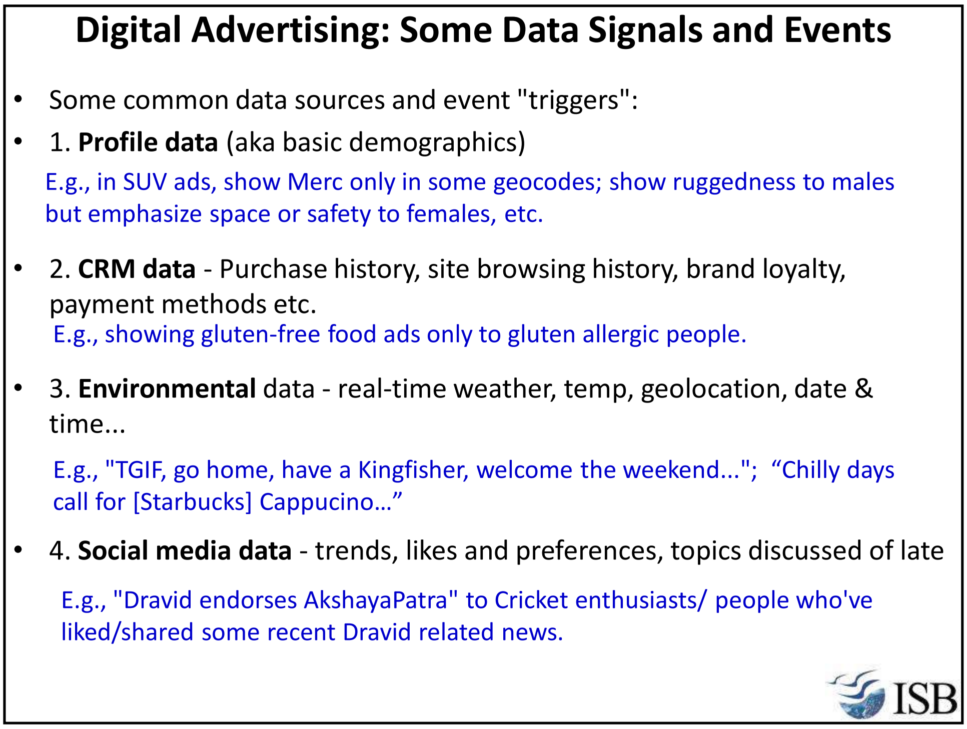
- Again, illustrative only…

Data Signal(s)	**Trigger Condition**	**Message**	**Images**
Weather, City	Weather = Rainy, City = Seattle	Hey, Seattle, it's been pouring. Time to hit the Gym.	Animated Gym scene, treadmill, Nike jogging shoe
Weather, City	Weather = Cold, City = Seattle	Hey, Seattle, it's freezing. Warm up with a workout.	Animated image of huddled person vs person on a treadmill, Nike jogging shoe.
Weather, City	Weather = Hot, City = Seattle	Hey, Seattle, it's a glorious sunny day.	Golf range, Golfer in Nike Golf shoes.

- Marketers write "event rules" and creatives build messaging copy (including the default copy) in tandem.

- Modularization of ad-copy into elements that can be mixed and matched to dynamically create ads is a reality today.

30



Digital Advertising: Some Data Signals and Events

- Some common data sources and event "triggers":
- 1. **Profile data** (aka basic demographics)

  E.g., in SUV ads, show Merc only in some geocodes; show ruggedness to males but emphasize space or safety to females, etc.
- 2. **CRM data** - Purchase history, site browsing history, brand loyalty, payment methods etc.

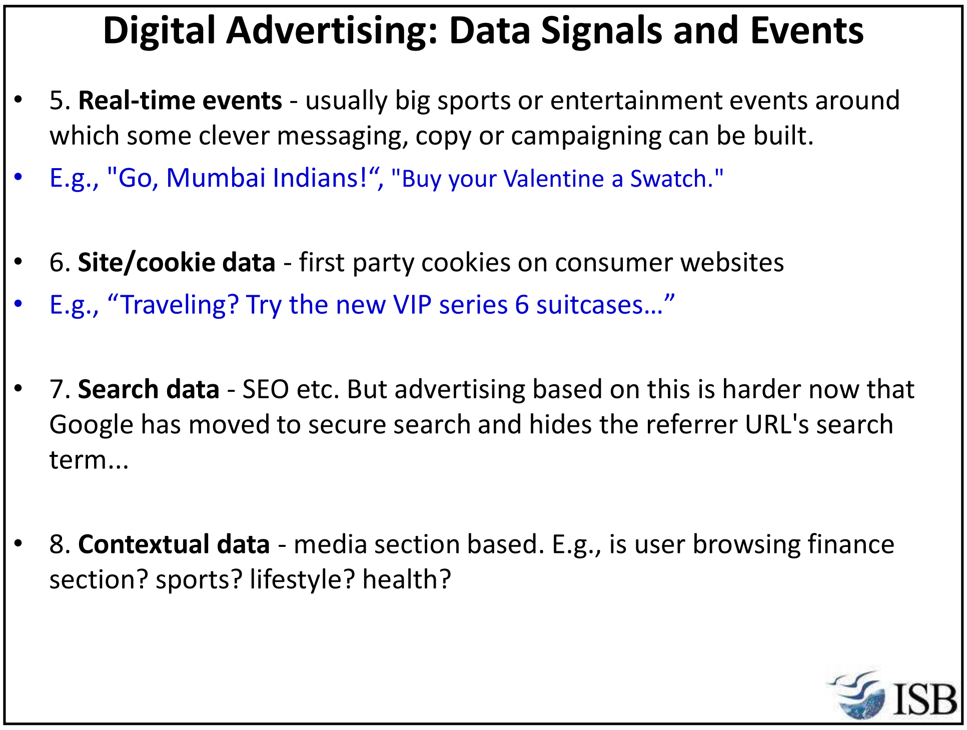
  E.g., showing gluten-free food ads only to gluten allergic people.
- 3. **Environmental** data - real-time weather, temp, geolocation, date & time...

  E.g., "TGIF, go home, have a Kingfisher, welcome the weekend..."; “Chilly days call for [Starbucks] Cappucino…”
- 4. **Social media data** - trends, likes and preferences, topics discussed of late

  E.g., "Dravid endorses AkshayaPatra" to Cricket enthusiasts/ people who've liked/shared some recent Dravid related news.

ISB

31



Digital Advertising: Data Signals and Events

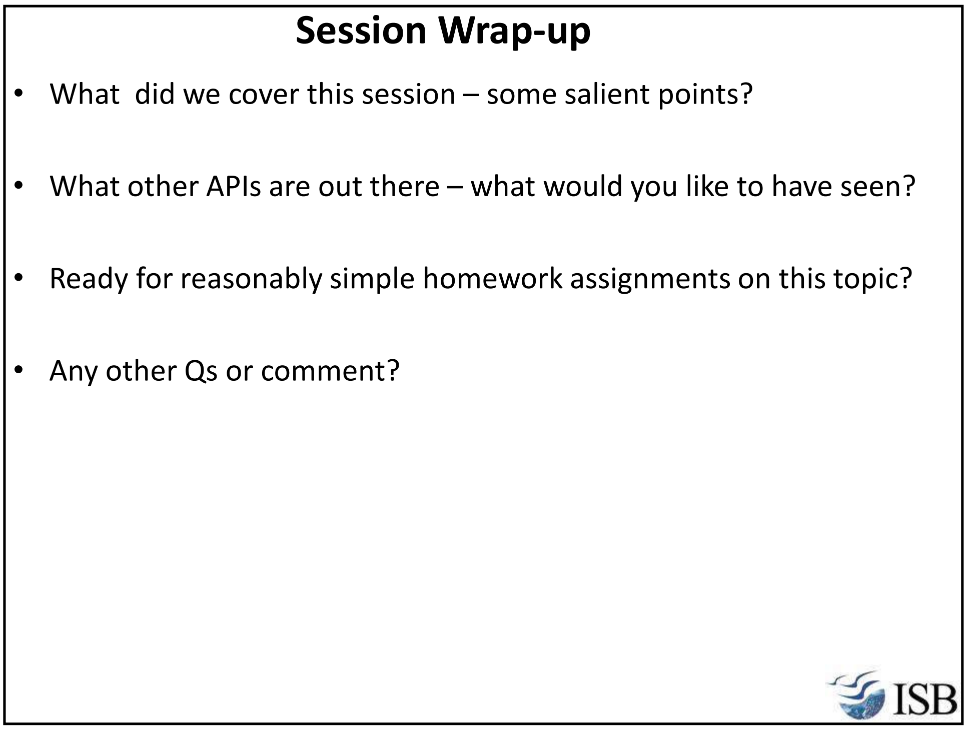
- 5. **Real-time events** - usually big sports or entertainment events around which some clever messaging, copy or campaigning can be built.
- E.g., "Go, Mumbai Indians!“, "Buy your Valentine a Swatch."
- 6. **Site/cookie data** - first party cookies on consumer websites
- E.g., “Traveling? Try the new VIP series 6 suitcases…”
- 7. **Search data** - SEO etc. But advertising based on this is harder now that Google has moved to secure search and hides the referrer URL's search term...
- 8. **Contextual data** - media section based. E.g., is user browsing finance section? sports? lifestyle? health?

ISB

32

Session Wrap-up

33



Session Wrap-up

- What did we cover this session – some salient points?
- What other APIs are out there – what would you like to have seen?
- Ready for reasonably simple homework assignments on this topic?
- Any other Qs or comment?

34