

Building Datasets with APIs

Community Data Science Workshop
Lecture 2 – Fall 2014

Frances Hocutt
@franceshocutt
frances.hocutt@gmail.com

Outline

- What is an API?
- How do we get interesting data through an API?
- What did we learn in Session 1?
- How do we write programs that use the internet?
- How can we use an API to fetch kitten pictures?
- Introduction to structured data (JSON)
- How do we use a generic API?

What is an API?

API = **A**pplication **P**rogramming **I**nterface

or:

a structured way for programs to communicate

web API = structured way to for programs to
communicate with a website

or:

just another way to get data from a website

Pedantic note:

when I say “API” today, hear “web API”

Why use an API, anyway?

Humans are good at perceiving information, patterns.
We read websites to get information/data.

Computers are good at handling (lots of) data.
We program them to get data from sites' **web APIs**.

APIs are for programs
as
webpages are for humans*

*more or less.

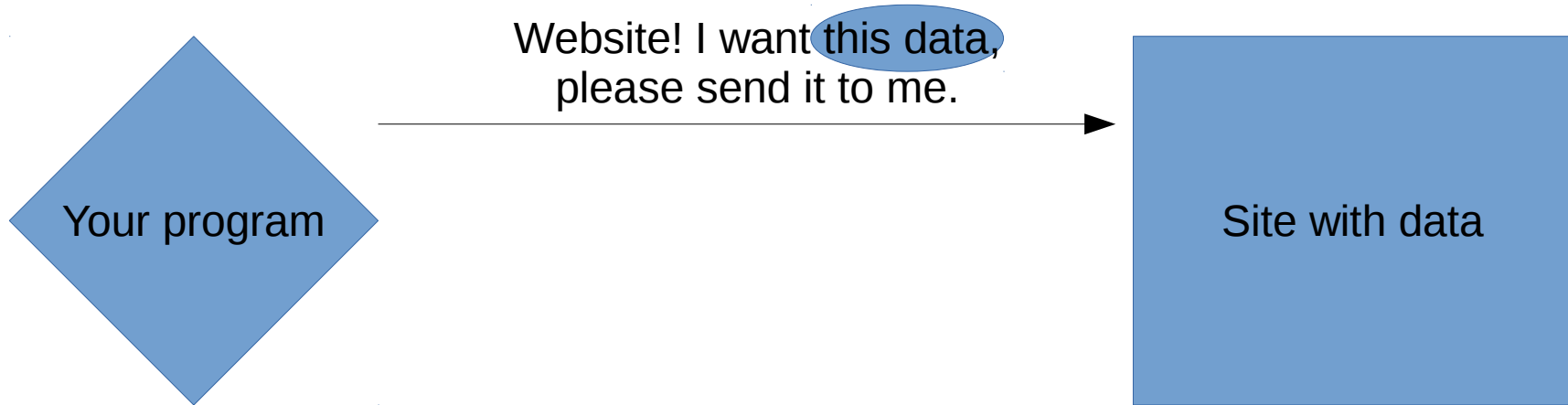
APIs make extracting data easier.

Great! How does it work?



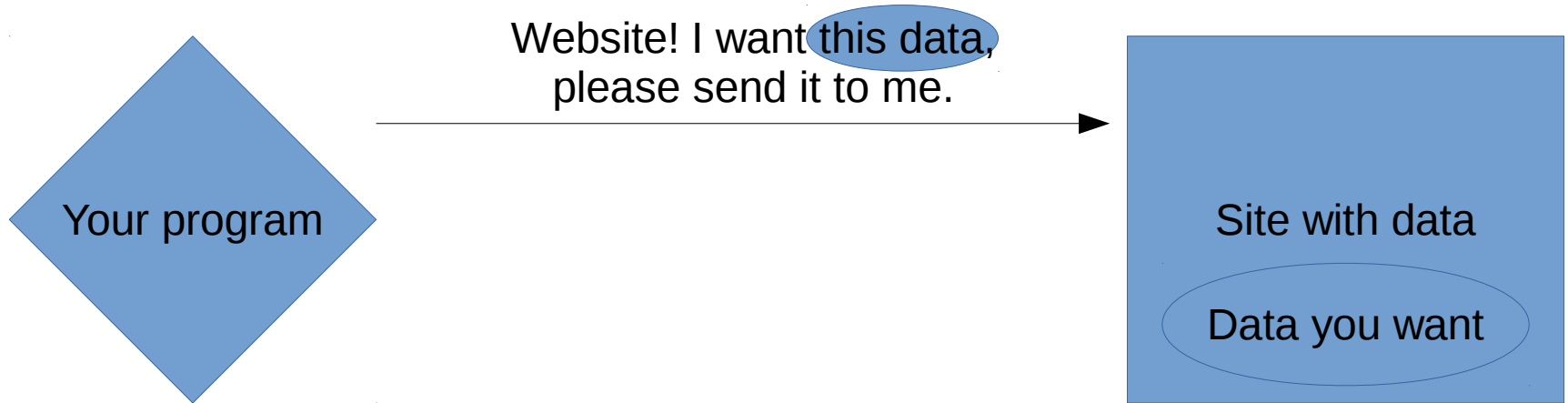
Great! How does it work?

Your program sends a carefully-phrased **request** for specific data to the site's API



Great! How does it work?

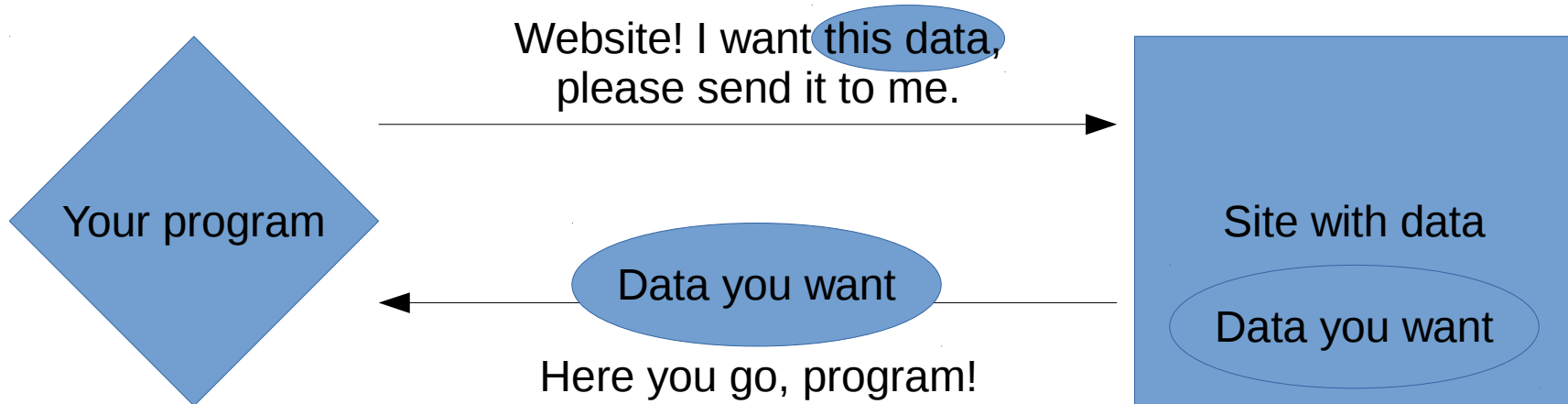
Your program sends a carefully-phrased **request** for specific data to the site's API



The site finds the data that you want and packages it in a form your program will understand.

Great! How does it work?

Your program sends a carefully-phrased **request** for specific data to the site's API



The site sends a **response** to your program that contains the data it requested.

The site finds the data that you want and packages it in a form your program will understand.



Your program now has **structured data** that it can **parse**, manipulate, analyze, etc.

To get data through an API,
you want to know:

Where do I direct my request?

How do I phrase my request?

How do I send my request?

Where do I direct my request?

The site's **endpoint**: like a website's main page.
You add your specific request to the end of it.

For example: <http://en.wikipedia.org/w/api.php>

How do I write my request?

How do I send my request?

Where do I direct my request?

How do I write my request?

Usually a specially-formatted series of inputs.
Different for different sites' APIs.
See documentation, code samples for specifics.

How do I send my request?

Where do I direct my request?

How do I write my request?

How do I send my request?

Web browser (useful for testing)

Python modules: `requests`

Example: listing links on Wikipedia

<http://en.wikipedia.org/w/api.php>

+

?action=query&prop=links&plnamespace=0&
pllimit=max&titles=Organic_chemistry

=

http://en.wikipedia.org/w/api.php?action=query&prop=links&plnamespace=0&pllimit=max&titles=Organic_chemistry

Result: structured data (JSON)

```
{... "query":{"normalized":[{"from":"Organic_chemistry",
"to":"Organic chemistry"}],"pages":{"22208":{"pageid":
22208,"ns":0,"title":"Organic chemistry","links":[{"ns":0,
"title":"5\u03b1-Dihydroprogesterone"}, {"ns":0,"title":
"Acetic acid"}, {"ns":0,"title":"Actinide chemistry"},
{"ns":0,"title":"Adhesive"}, {"ns":0,"title":"Adolf von
Baeyer"}, {"ns":0,"title":"Alcohol"}, {"ns":0,"title":"Aldol
reaction"}, {"ns":0,"title":"Alicyclic"}, {"ns":0,"title":
"Aliphatic compound"}, {"ns":0,"title":"Alkali"}, {"ns":0,
"title":"Alkali metal"}, {"ns":0,"title":"Alkaline earth
metal"}, {"ns":0,"title":"Alkaloid"}, {"ns":0,"title":
"Alkylidene"}, {"ns":0,"title":"Amine"}, {"ns":0,"title":
"Amino acid"}, {"ns":0,"title":"Analytical chemistry"},
{"ns":0,"title":"Antiaromaticity"}, {"ns":0,"title":"Applied
science"}, {"ns":0,"title":"Archibald Scott Couper"}, ... }}}
```

Getting to the data in your result

There's a module for that!

```
import json  
data = json.loads(json_data_string)
```

parses the string with your data in it and turns it
into lists + dicts

Programming skills to build datasets with APIs include:

- all our tools from last session:
 - variables, datatypes (strings, integers, lists, dicts), if statements, for loops, importing modules
- the ability to create custom URLs
- the ability to open URLs on the web
- the ability to save to files
- the ability to parse the JSON data that APIs usually give us

New programming concepts

- An easier way to put variables in the middle of strings: interpolation using `%s`, `%()s`
- Using the `requests` module to get data from websites
- Opening files and putting data in them (`write` to them)
- Parsing JSON with the `json` module

Kitten photos via API

<http://placekitten.com>

We will write a program that asks us what size kitten photo we want, uses the API to fetch the photo, and saves the photo to a file.

Using structured data (JSON)

JSON's structure is nested Python dicts + lists

Except: strings inside JSON are "string", not 'string'

Note: you can't treat a string containing JSON like it is already a list or dictionary!
You have to parse it first.

```
json_string = '{"organizers":["Frances", "Mako",  
"JMo", "Tommy"]}'
```

parses to:

```
dict_from_json = {'organizers':['Frances',  
'Mako', 'JMo', 'Tommy']}
```

API resources

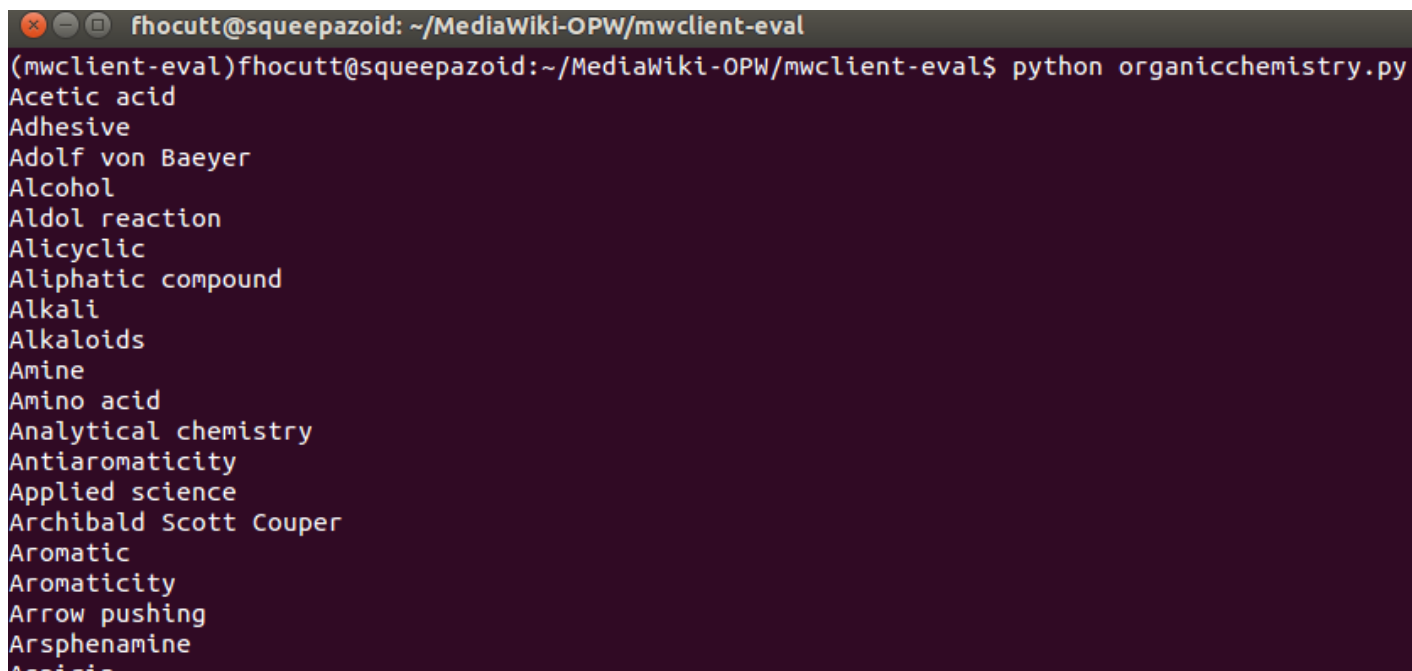
- Finding the documentation: try searching “site name + developers + api”
- Official API documentation
- Online searches for [site] + api + help
- Python modules to access that API (API client libraries; may be called the SDK, software developer's kit)

Using the mwclient module to query WP

```
import mwclient

wikipedia = mwclient.Site('en.wikipedia.org')
ochem = wikipedia.Pages['Organic chemistry']
links = ochem.links(generator=False)

for item in links:
    print item
```



```
fhocutt@squeepazoid: ~/MediaWiki-OPW/mwclient-eval
(mwclient-eval)fhocutt@squeepazoid:~/MediaWiki-OPW/mwclient-eval$ python organicchemistry.py
Acetic acid
Adhesive
Adolf von Baeyer
Alcohol
Aldol reaction
Alicyclic
Aliphatic compound
Alkali
Alkaloids
Amine
Amino acid
Analytical chemistry
Antiaromaticity
Applied science
Archibald Scott Couper
Aromatic
Aromaticity
Arrow pushing
Arsphenamine
Aspirin
```

Potential API pitfalls

- Rate limiting
 - Site limits the number of requests/hr allowed
- Authentication
 - Send identity data with your request
 - Some sites require it, some have higher rate limits for authenticated users
- Text encoding issues
 - Show up with accented and non-English characters
 - There is a difference between Unicode strings (`u' string '`) and regular strings (`' string '`)
 - Try searching for [site] + encoding problem + api

Afternoon sessions

- Building a dataset using the Wikipedia API
- Building a dataset using Twitter
- Wikipedia and Stack Exchange MySQL
- CodeAcademy practice with APIs exercises