

DATA SCIENCE

DATA FROM API'S AND WEB SCRAPING

LAST TIME:

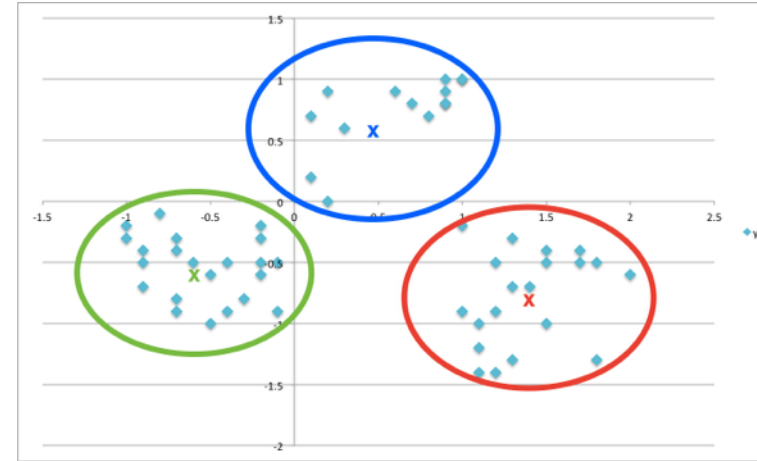
I. CLUSTER ANALYSIS

II. K-MEANS CLUSTERING

III. CLUSTER VALIDATION

EXERCISE:

IV. K-MEANS CLUSTERING IN PYTHON



QUESTIONS?

WHAT WAS THE MOST INTERESTING THING YOU LEARNT?

WHAT WAS THE HARDEST TO GRASP?

I. DATA FORMATS

II. APIS

EXERCISES:

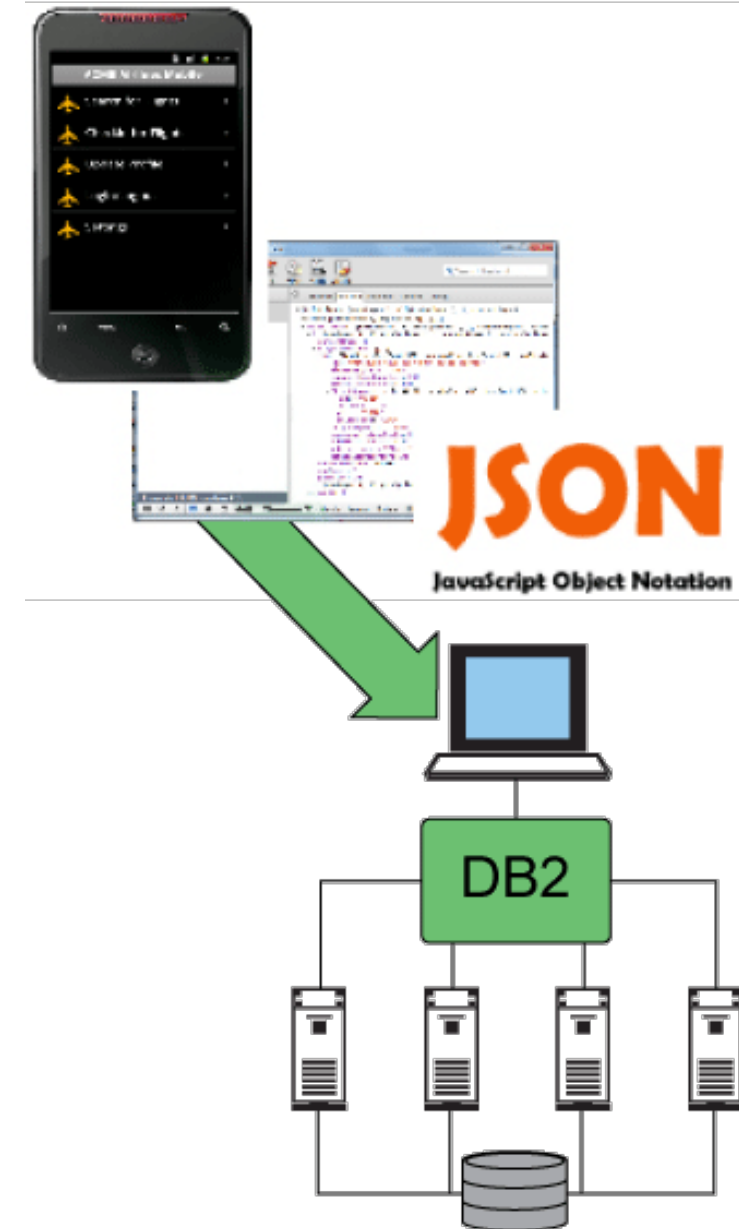
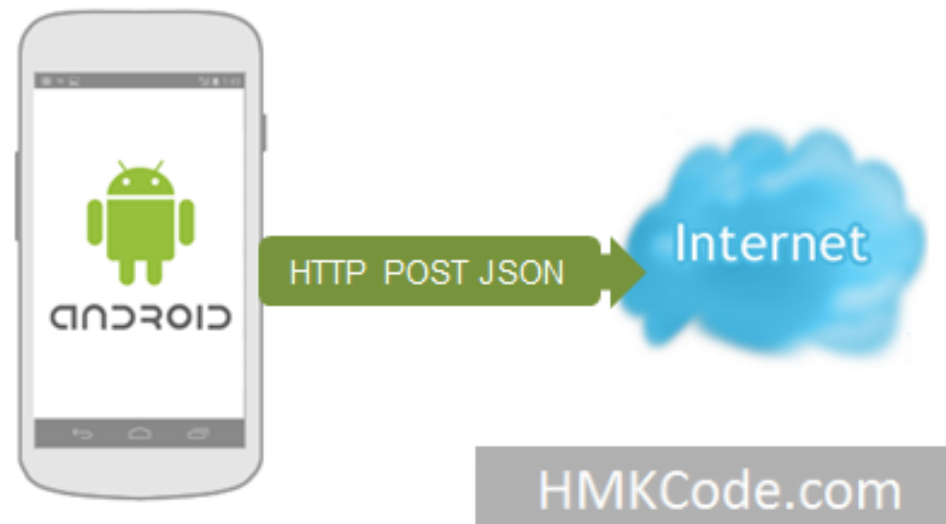
III. EXTENDED HANDS-ON LAB

DATA FORMAT, ACCESS & TRANSFORMATION

JSON, CSV, ETC...

JSON (JavaScript Object Notation) is:
a lightweight **data-interchange format**
a **string**

JSON can be passed
between **applications**
easy for **machines** to parse and generate



JSON are passed through applications
as **strings**
and converted into native objects per language.

JSON

JSON are passed through applications
as **strings**
and converted into native objects per language.

```
{ "empinfo" :  
  {  
    "employees" : [  
      {  
        "name" : "Scott Philip",  
        "salary" : f44k,  
        "age" : 27,  
      },  
      {  
        "name" : "Tim Henn",  
        "salary" : f40k,  
        "age" : 27,  
      },  
      {  
        "name" : "Long Yong",  
        "salary" : f40k,  
        "age" : 28,  
      }  
    ]  
  }  
}
```

```
import json  
  
py_object = [ { 'a':'A', 'b':(2, 4), 'c':3.0 } ]  
  
json_string = json.dumps(py_object)  
  
print 'JSON:', json_string
```

JSON: [{"a": "A", "c": 3.0, "b": [2, 4]}]

```
decoded_py_object = json.loads(json_string)
```

<https://docs.python.org/2/library/json.html>

CSV (Comma Separated Values):

```
name,game,points  
John,basketball,3  
Mary,volleyball,5  
James,ping pong,2  
...
```

CSV (Comma Separated Values):

- easy to read and write
- structured like a table
- very common
- can export to/from MS Excel

<https://docs.python.org/2/library/csv.html>

DATA FORMAT, ACCESS & TRANSFORMATION

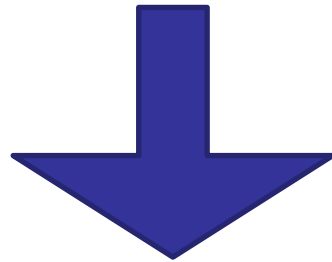
APIs

APIs (Application Programming Interface)
allow people to **interact** with the structures of
an application

- get
- put
- delete
- update
- ...

Best practices for APIs are to
use **RESTful** principles.

Best practices for APIs are to
use **RESTful** principles.



Representational State Transfer (REST)

RESTFUL EXAMPLE

RESTful API HTTP methods

Resource	GET	PUT	POST	DELETE
Collection URI, such as <code>http://example.com/resources/</code>	List the URIs and perhaps other details of the collection's members.	Replace the entire collection with another collection.	Create a new entry in the collection. The new entry's URI is assigned automatically and is usually returned by the operation. ^[9]	Delete the entire collection.
Element URI, such as <code>http://example.com/resources/item17</code>	Retrieve a representation of the addressed member of the collection, expressed in an appropriate Internet media type.	Replace the addressed member of the collection, or if it does not exist, create it.	Not generally used. Treat the addressed member as a collection in its own right and create a new entry in it. ^[9]	Delete the addressed member of the collection.

http://en.wikipedia.org/wiki/Representational_state_transfer

- The Base URL
- An interactive media type (usually JSON)
- Operations (GET, PUT, POST, DELETE)
- Driven by http requests

REST API EXAMPLE

Collection



GET <https://api.instagram.com/v1/users/10>



Operation

REST API EXAMPLE

**GET https://api.instagram.com/v1/users/
search/?q=andy**



Querystring

APIS

Okay, so what if we execute the following, but there is no item 18 in the users collection?

GET <https://api.instagram.com/v1/users/18>

APIS



<https://dev.twitter.com/rest/public>

<https://developer.linkedin.com/docs/signin-with-linkedin>

<http://www.pythonapi.com/>

PAIR EXERCISE:

<http://www.pythonapi.com/>

- 1) CHOOSE 1 API: WHAT DATA YOU CAN GET?**
- 2) INSTALL PYTHON MODULE, TRY TO EXTRACT DATA**
- 3) DISCUSS: HOW COULD YOU LEVERAGE THAT API? HOW COULD YOU USE THE DATA?**

DATA FORMAT, ACCESS & TRANSFORMATION

QUESTIONS?

INTRO TO DATA SCIENCE

EX: SCRAPING WITH PYTHON