GENERAL ASSEMBLY

DATA SCIENCE

DATA SOURCES

LAST TIME:

I. CLEANING DATA

II. MISSING DATA

EXERCISES:

III. NUMPY

IV. PANDAS

V. BOKEH & MATPLOTLIB

QUESTIONS?

WHAT WAS THE MOST INTERESTING THING YOU LEARNT?

WHAT WAS THE HARDEST TO GRASP?

I. DATA SOURCES
II. DATA FORMATS
III. APIS

EXERCISES:
IV. RETRIEVE DATA FROM VARIOUS SOURCES
V. KIMONO LABS & OTHER APIS

- EXPLORE VARIOUS DATA SOURCES
- UNDERSTAND DIFFERENT DATA FORMATS
- BE ABLE TO RETRIEVE DATA FROM APIS

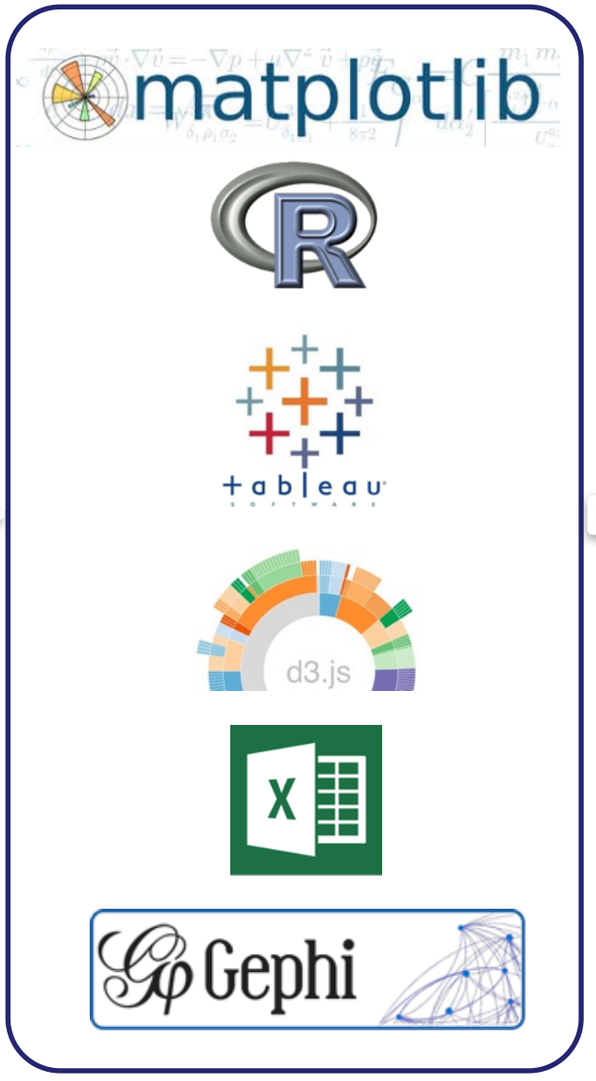
DATA SOURCES

DATA FLOW

Data Retrieval

Data ETL and Aggregation

Data Visualization



Machine Learning

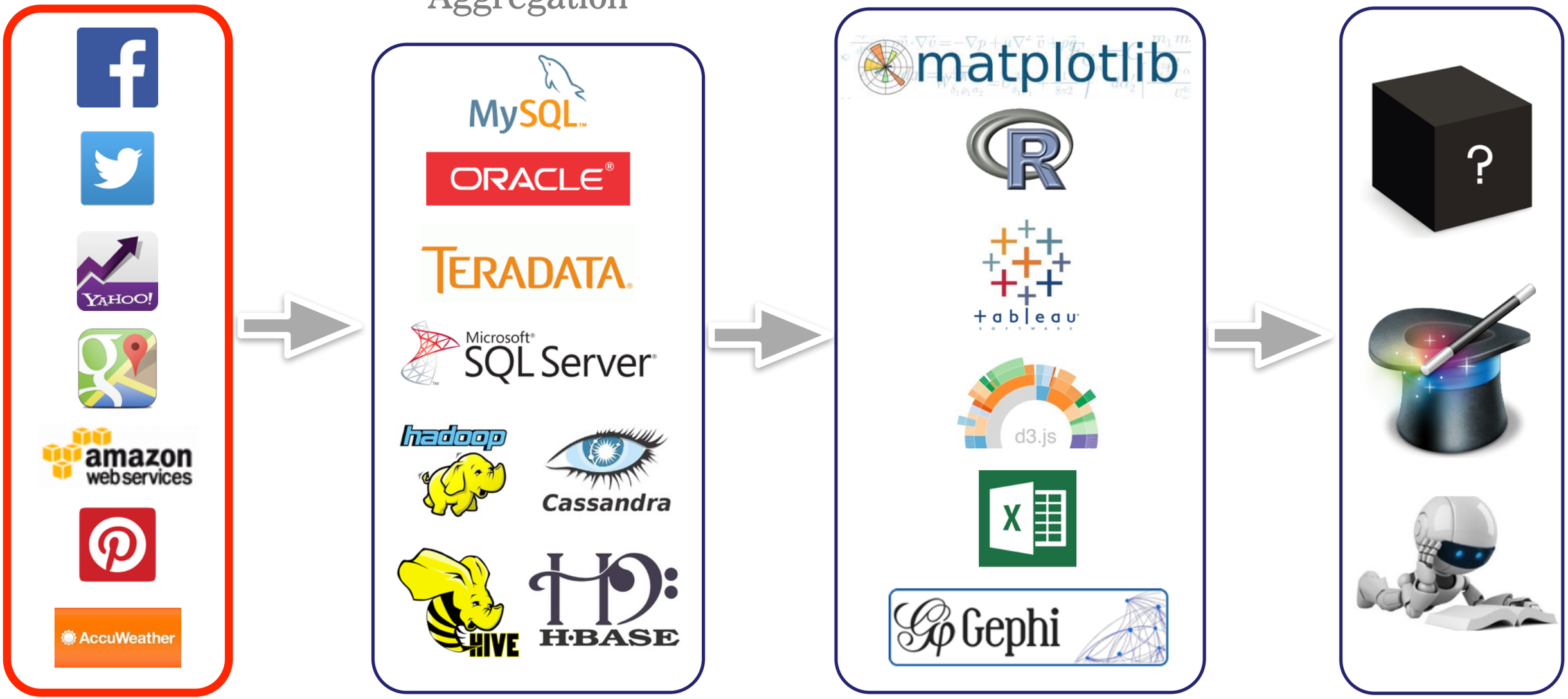
DATA FLOW

Data Retrieval

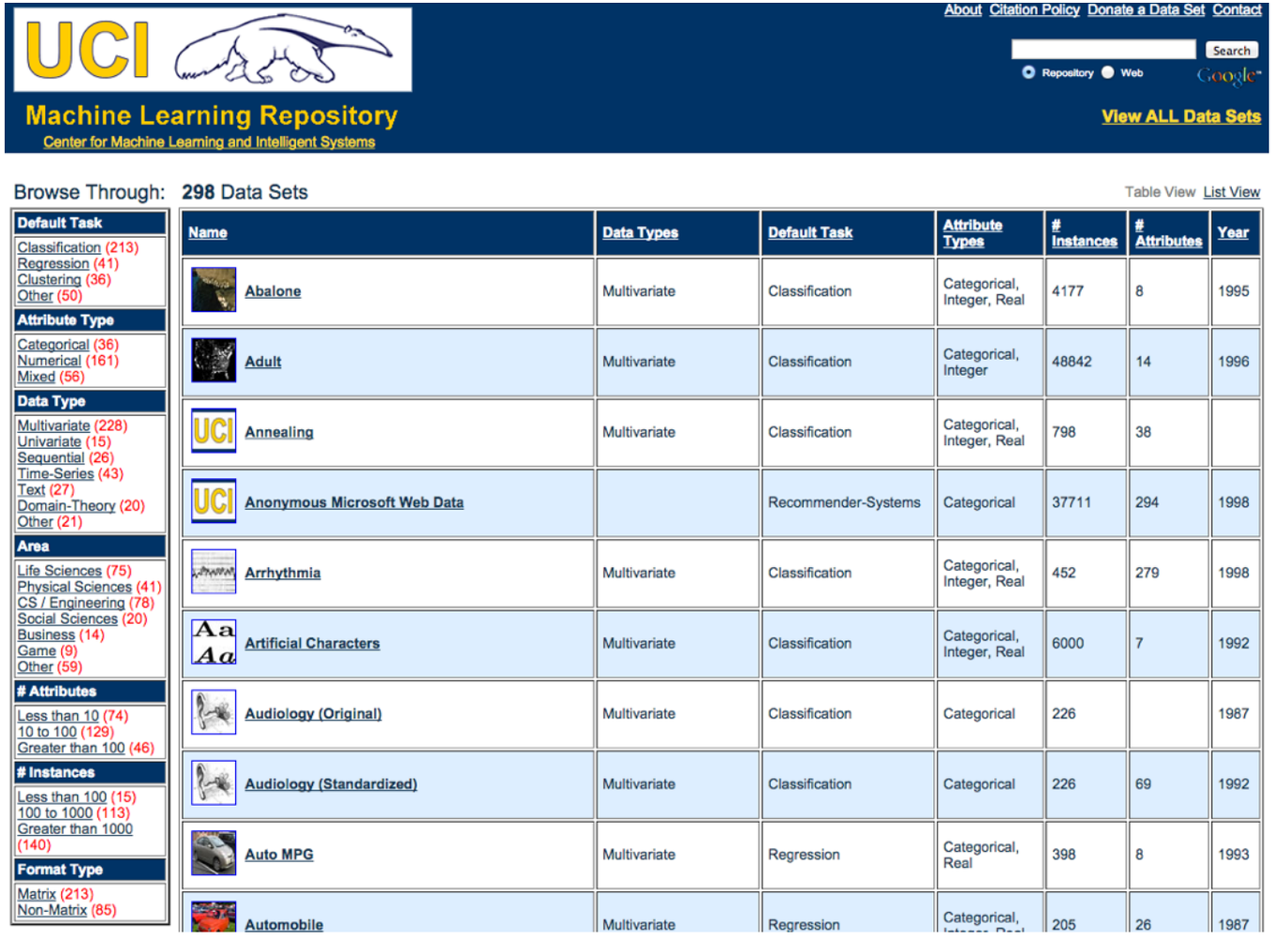
Data ETL and Aggregation

Data Visualization

Machine Learning



DATA SOURCES



UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact

Search
Repository Web
Google™

View ALL Data Sets

Browse Through: **298** Data Sets

Table View List View

Default Task
Classification (213)
Regression (41)
Clustering (36)
Other (50)

Attribute Type
Categorical (36)
Numerical (161)
Mixed (56)

Data Type
Multivariate (228)
Univariate (15)
Sequential (26)
Time-Series (43)
Text (27)
Domain-Theory (20)
Other (21)

Area
Life Sciences (75)
Physical Sciences (41)
CS / Engineering (78)
Social Sciences (20)
Business (14)
Game (9)
Other (59)

 # Attributes
Less than 10 (74)
10 to 100 (129)
Greater than 100 (46)

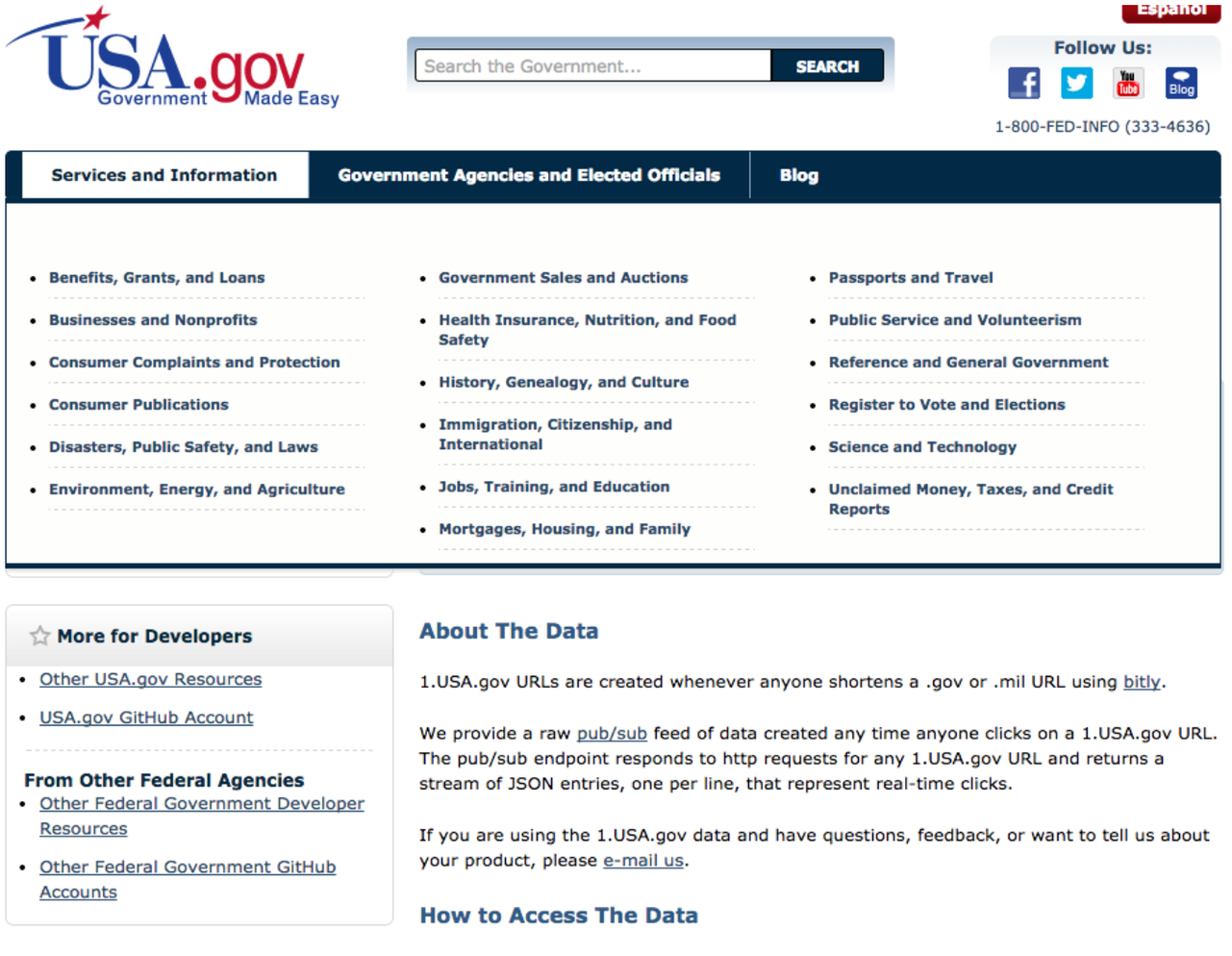
 # Instances
Less than 100 (15)
100 to 1000 (113)
Greater than 1000 (140)

Format Type
Matrix (213)
Non-Matrix (85)

Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998
Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279	1998
Artificial Characters	Multivariate	Classification	Categorical, Integer, Real	6000	7	1992
Audiology (Original)	Multivariate	Classification	Categorical	226		1987
Audiology (Standardized)	Multivariate	Classification	Categorical	226	69	1992
Auto MPG	Multivariate	Regression	Categorical, Real	398	8	1993
Automobile	Multivariate	Regression	Categorical,	205	26	1987

Source: http://archive.ics.uci.edu/ml/datasets.html

DATA SOURCES



Source: http://www.usa.gov/About/developer-resources/1usagov.shtml

DATA SOURCES

Source: http://www.kaggle.com/

1) PETE SKOMOROCH (LINKEDIN) HTTPS://DELICIOUS.COM/PSKOMOROCH/DATASET

2) HILARY MASON (ACCEL PARTNERS, BITLY) HTTPS://BITLY.COM/BUNDLES/HMASON/1

3) KEVIN CHAI (U. OF NEW SOUTH WALES, SYDNEY) HTTP://KEVINCHAI.NET/DATASETS

4) JEFF HAMMERBACHER (CLOUDERA) HTTP://WWW.QUORA.COM/JEFF-HAMMERBACHER/INTRODUCTION-TO-DATA-SCIENCE-DATA-SETS

5) JERRY SMITH (3I-MIND) HTTP://DATASCIENTISTINSIGHTS.COM/2013/10/07/DATA-REPOSITORIES-MOTHERS-MILK-FOR-DATA-SCIENTISTS/

6) GREGORY PIATETSKY-SHAPIRO (KDD) HTTP://WWW.KDNUGGETS.COM/DATASETS/INDEX.HTML

7) HTTP://WWW.QUORA.COM/DATA/WHERE-CAN-I-FIND-LARGE-DATASETS-OPEN-TO-THE-PUBLIC

8) HTTPS://GITHUB.COM/CAESAR0301/AWESOME-PUBLIC-DATASETS

9) HTTP://RS.IO/100-INTERESTING-DATA-SETS-FOR-STATISTICS

Source: http://blog.mortardata.com/post/67652898761/6-dataset-lists-curated-by-data-scientists

PAIR EXERCISE:

CHOOSE A DATA SOURCE AND LOOK AT WHAT DATA YOU CAN GET

DISCUSS HOW YOU WOULD USE THE DATA

QUESTIONS?

JSON, CSV, ETC...

JSON (JavaScript Object Notation) is:

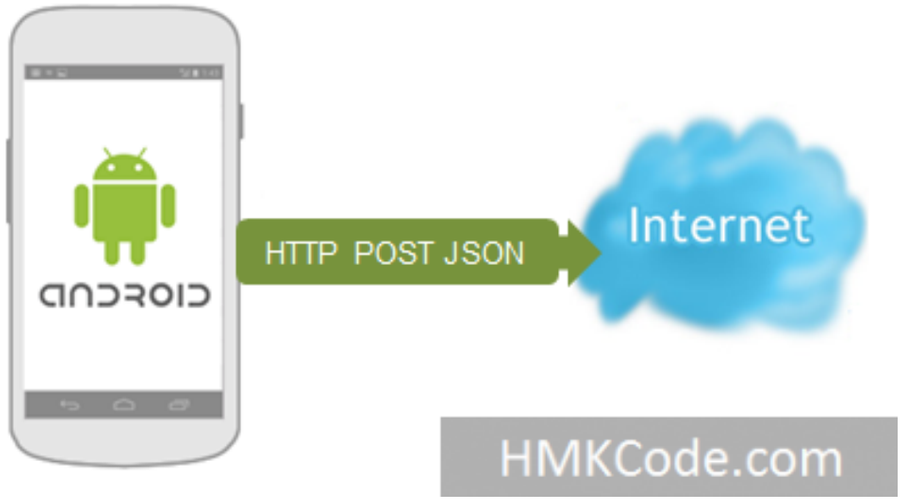
a lightweight data-interchange format

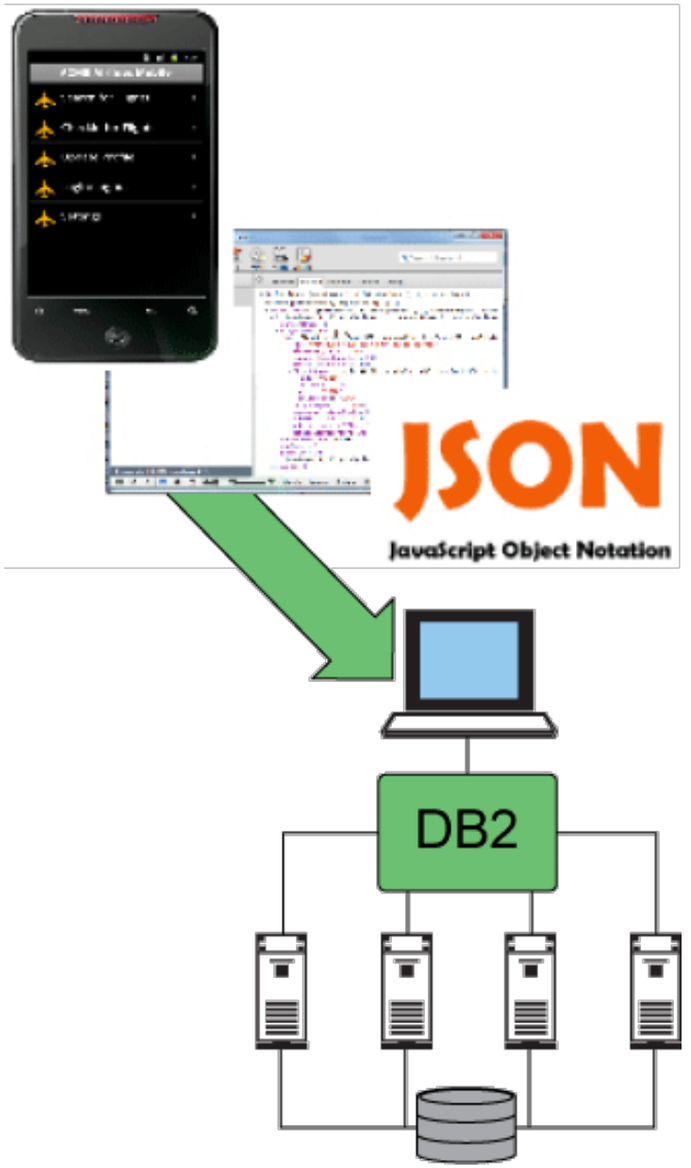
a string

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" : £44k,
            "age" : 27,

          },

          {
            "name" : "Tim Henn",
            "salary" : £40k,
            "age" : 27,
          },

          {
           “name” : “Long Yong”,
           “salary” : £40k,
           “age” : 28,

          }
                    ]
    }
}
```

PYTHON OBJECT ——> JSON STRING

```
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]}]
```

JSON STRING ——> PYTHON OBJECT

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

PYTHON JSON LIBRARY

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

PYTHON CSV LIBRARY

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

OTHER DATA FORMATS

txt

tsv

xml

dat

images

binary

etc…

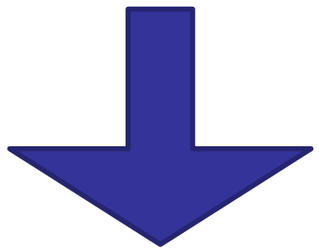
APIs

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

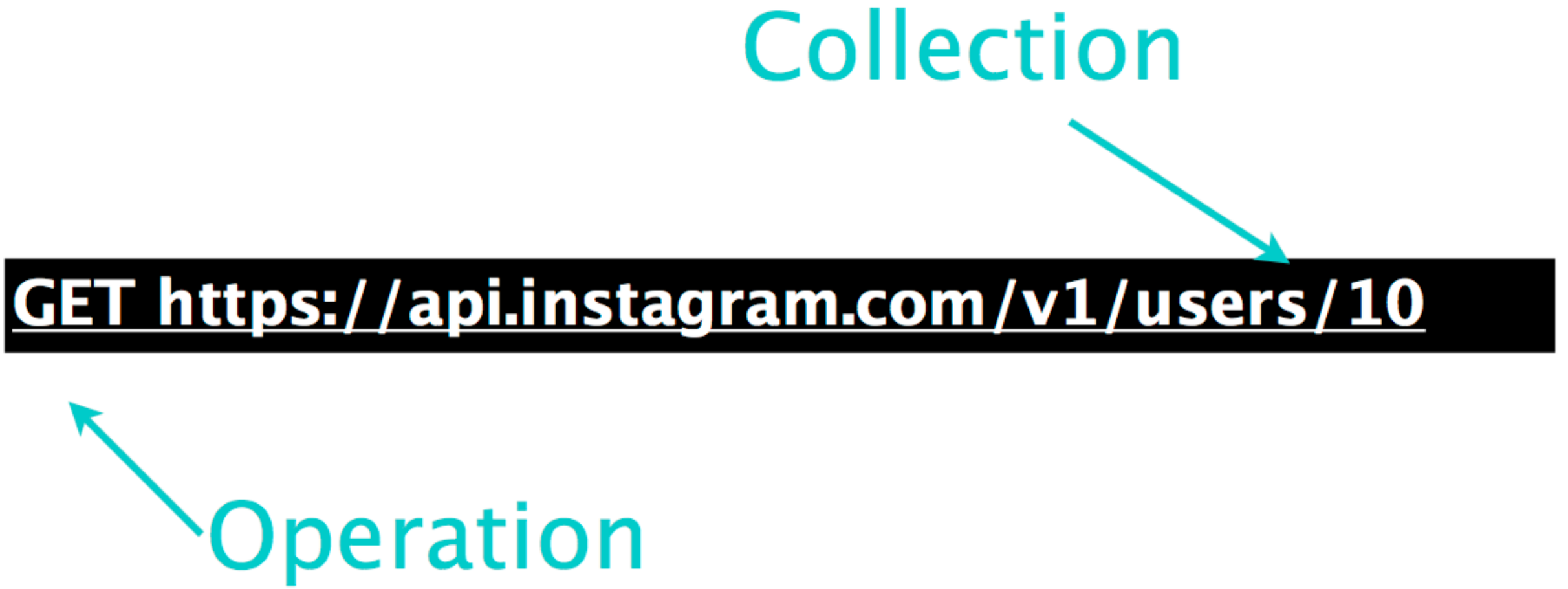
RESTful API HTTP methods

Resource	GET	PUT	POST	DELETE
Collection URI, such as `http://example.com/resources/`	**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 `http://example.com/resources/item17`	**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



REST API EXAMPLE

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

Querystring

TWITTER REST API

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

LINKEDIN REST API

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

LIST OF PYTHON APIS

http://www.pythonapi.com/

PAIR EXERCISE: http://www.pythonapi.com/

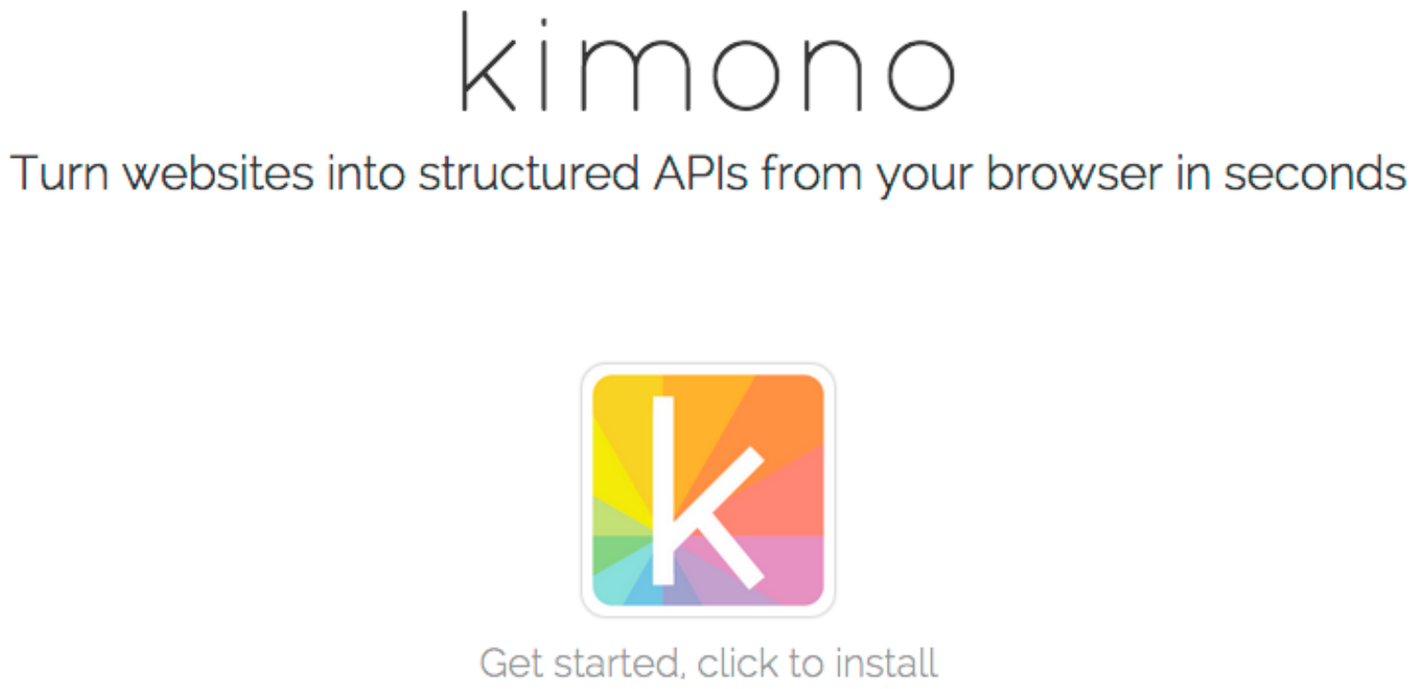
1) CHOOSE 1 API: WHAT DATA YOU CAN GET?

2) INSTALL PYTHON MODULE, TRY TO EXTRACT DATA

3) DISCUSS: HOW COULD YOU LEVERAGE THAT API? HOW COULD YOU USE THE DATA?

KIMONO LABS

www.kimonolabs.com



QUESTIONS?