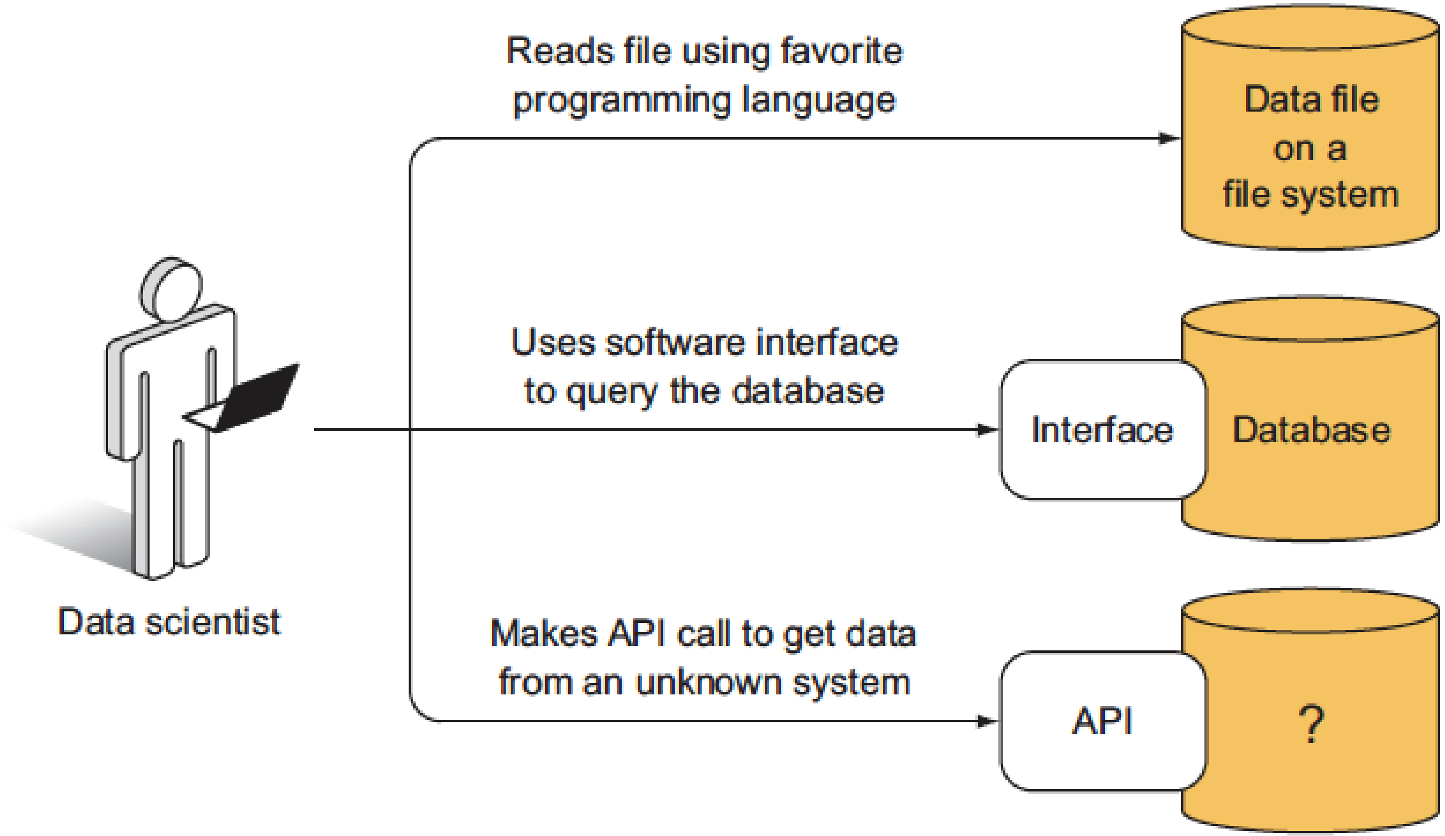
Data Search

CDS-492 | Dr. Slamani

(original slides by Dr. Ron Mahabir)

Common ways to access data



Reads file using favorite programming language
Data file on a file system
Uses software interface to query the database
Interface
Database
Data scientist
Makes API call to get data from an unknown system
API
?

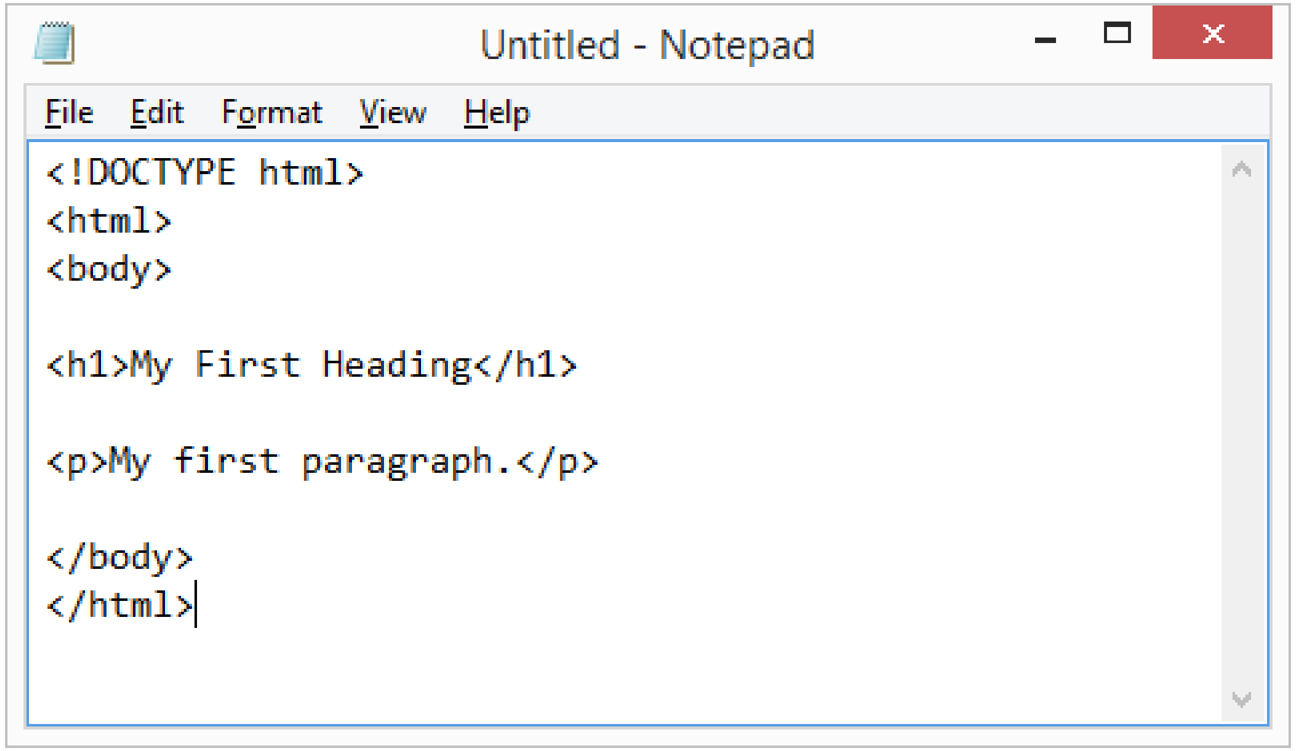
https://www.manning.com/books/think-like-a-data-scientist

Data formats - Flat files

- Just text (e.g. csv or tsv)
- Minimal
- Word documents are not flat files
  - They contain additional overhead such as style etc.
- Many popular programming languages have packages to read these
- Can be slow to read if very large
  - Database may be an option here
- Some files may be compressed

Data formats - HTML

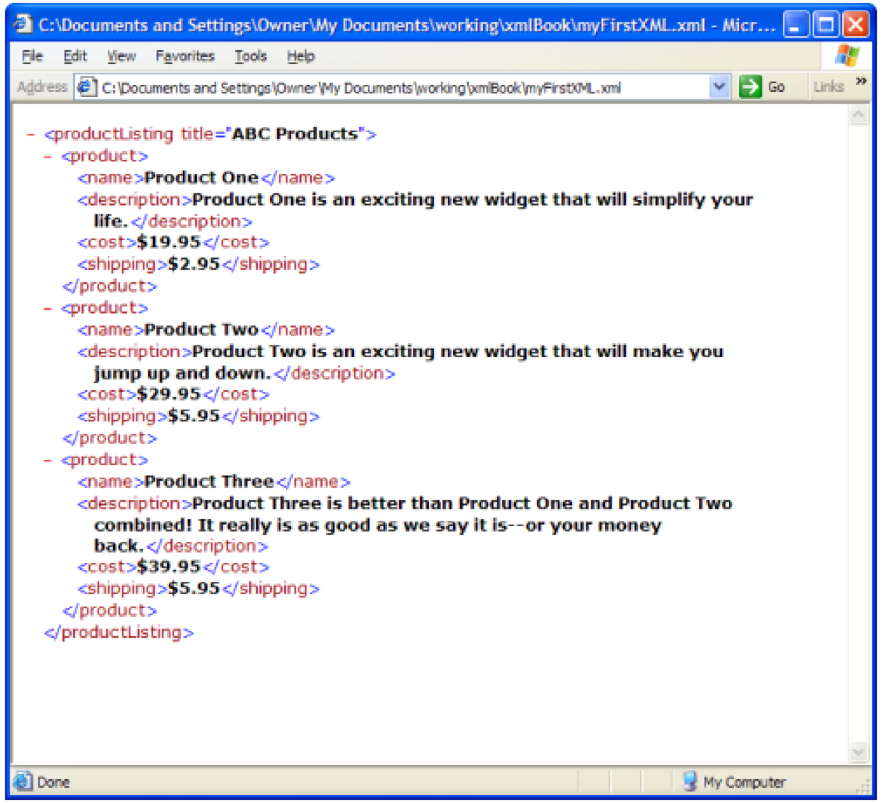
- Plain text marked up with tags or specially denoted instructions for how the text should be interpreted.



```
<!DOCTYPE html>
<html>
<body>

<h1>My First Heading</h1>

<p>My first paragraph.</p>

</body>
</html>
```

Data formats - XML

- Extension of HTML
- More suitable for storing and transmitting documents and data other than web pages

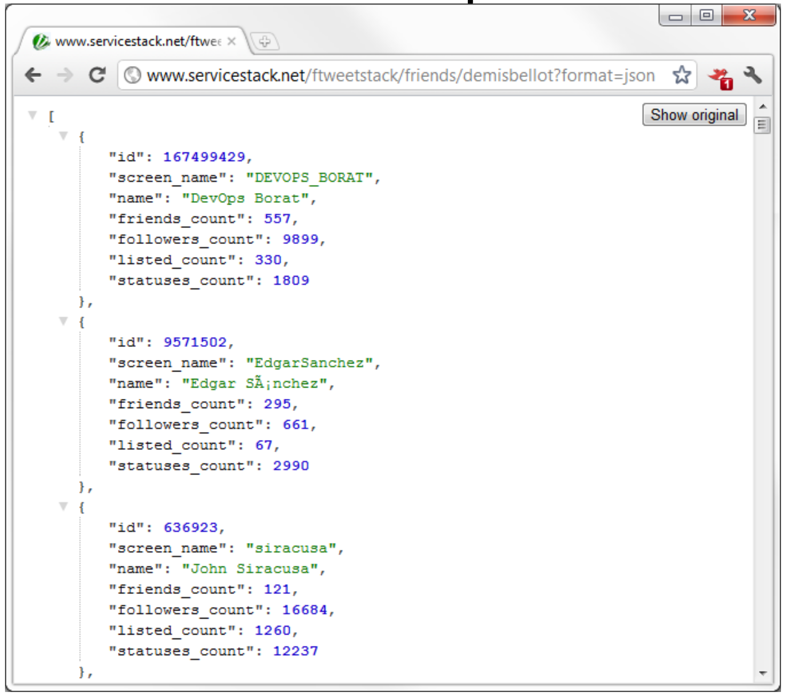


C:\Documents and Settings\Owner\My Documents\working\xmlBook\myFirstXML.xml - Micr...
File Edit View Favorites Tools Help
Address C:\Documents and Settings\Owner\My Documents\working\xmlBook\myFirstXML.xml
Go
Links
Done
My Computer

```
- <productListing title="ABC Products">
  - <product>
      <name>Product One</name>
      <description>Product One is an exciting new widget that will simplify your
        life.</description>
      <cost>$19.95</cost>
      <shipping>$2.95</shipping>
    </product>
  - <product>
      <name>Product Two</name>
      <description>Product Two is an exciting new widget that will make you
        jump up and down.</description>
      <cost>$29.95</cost>
      <shipping>$5.95</shipping>
    </product>
  - <product>
      <name>Product Three</name>
      <description>Product Three is better than Product One and Product Two
        combined! It really is as good as we say it is--or your money
        back.</description>
      <cost>$39.95</cost>
      <shipping>$5.95</shipping>
    </product>
  </productListing>
```

Data formats - JSON

- JavaScript Object Notation
- Typically describes something more like a data structure
  - such as a list, map, or dictionary in many popular programming languages
  - Leaner in number of characters compared to XML



www.servicestack.net/ftwee
www.servicestack.net/ftweetstack/friends/demisbellot?format=json
Show original

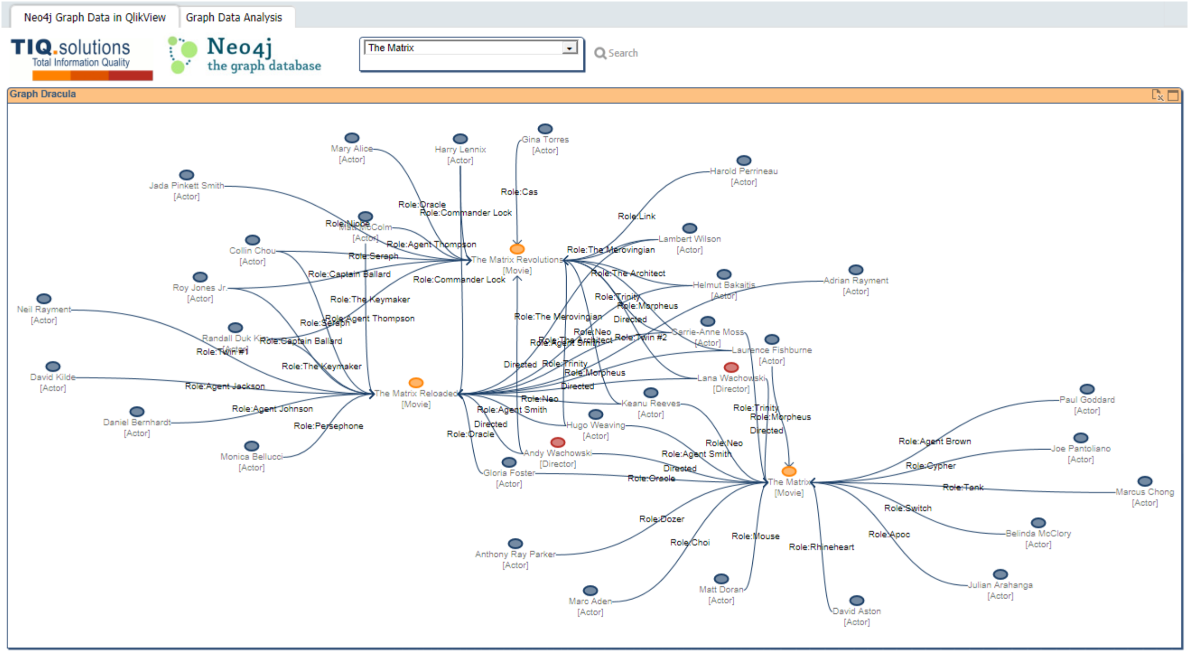
```
[
  {
    "id": 167499429,
    "screen_name": "DEVOPS_BORAT",
    "name": "DevOps Borat",
    "friends_count": 557,
    "followers_count": 9899,
    "listed_count": 330,
    "statuses_count": 1809
  },
  {
    "id": 9571502,
    "screen_name": "EdgarSanchez",
    "name": "Edgar SÃ¡nchez",
    "friends_count": 295,
    "followers_count": 661,
    "listed_count": 67,
    "statuses_count": 2990
  },
  {
    "id": 636923,
    "screen_name": "siracusa",
    "name": "John Siracusa",
    "friends_count": 121,
    "followers_count": 16684,
    "listed_count": 1260,
    "statuses_count": 12237
  },
```

Relational databases

- Storage systems used to optimize retrieval, writing and management of data
    - Indexing
- Queries
- Relationship between tables
- Some analytical capabilities

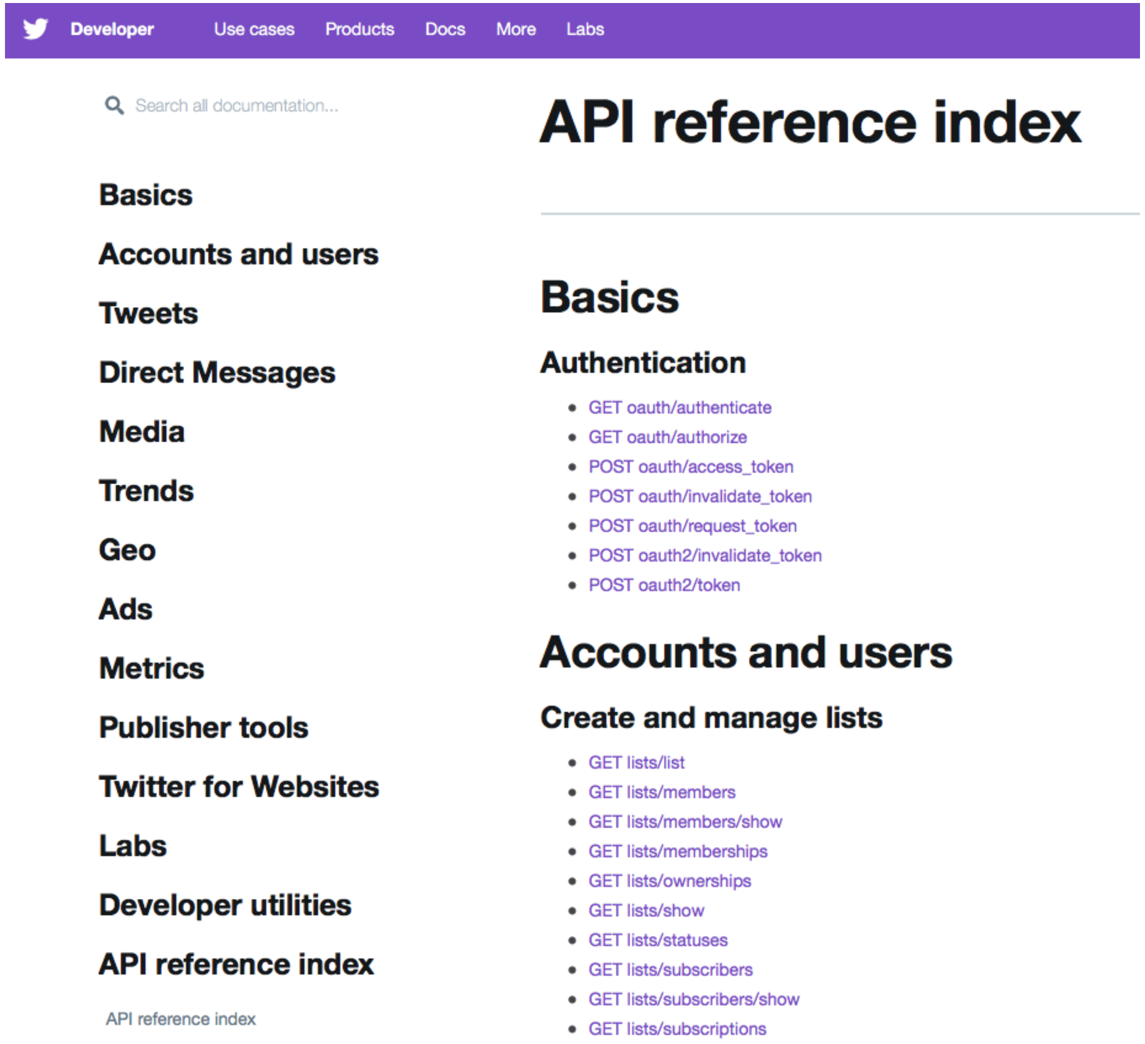
Non-relational databases

- Not only SQL (NoSQL)
- Don’t conform to the norm
- Greater flexibility
- Examples include:
  - MongoDB
  - Elasticsearch – document oriented database
    - Full-text search engine with an HTTP web interface and schema-free JSON documents.
  - Graph database
    - Graph structures for semantic queries with nodes, edges, and properties to represent and store data instead of tables, or documents.



API – Application Programming Interface

- Set of rules for communicating with a piece of software to access data
- The gatekeeper
- Most programming language have packages to assist



Basics

Accounts and users

Tweets

Direct Messages

Media

Trends

Geo

Ads

Metrics

Publisher tools

Twitter for Websites

Labs

Developer utilities

API reference index

API reference index

API reference index

Basics

Authentication

- GET oauth/authenticate
- GET oauth/authorize
- POST oauth/access_token
- POST oauth/invalidate_token
- POST oauth/request_token
- POST oauth2/invalidate_token
- POST oauth2/token

Accounts and users

Create and manage lists

- GET lists/list
- GET lists/members
- GET lists/members/show
- GET lists/memberships
- GET lists/ownerships
- GET lists/show
- GET lists/statuses
- GET lists/subscribers
- GET lists/subscribers/show
- GET lists/subscriptions

Find a work around

- MS Word
- MS Excel
- PDF
- .mbox
- netcdf – Maybe a software (network Common Data Form - file format for storing multidimensional scientific data (variables) such as temperature, humidity, pressure)
- Then again
  - For a quick peek
  - For very small files

Finding the DATA

- Google
  - General vs Specific search terms
- Repositories
  - Github
  - Kaggle
- Web scraping
  - Maybe against the terms of service for many websites
    - REMEMBER IF YOU DO WEBSCRAPE – Humans are erratic at some things! ☺
- Download an entire website
  - YIKES!!!
- Measure or collect it yourself

Copyright and licensing

- Data may have licensing, copyright, or other restrictions that can make it illegal to use the data for certain purposes.
  - University – research
  - Twitter – don’t distribute
- Without confirming that your use case is legal, you remain at risk of losing access to the data or, even worse, a lawsuit
- Other ethical considerations
  - Hand, D.J., 2018. Aspects of data ethics in a changing world: Where are we now?. *Big data*, *6*(3), pp.176-190.

Got the data, is it ENOUGH?

- On the surface things can be all bright and shiny
- Below the surface, no happy faces
- What to do
  - Dive in
  - Work on a sample
  - Combine another dataset?

Links to data

- https://www.dataquest.io/blog/free-datasets-for-projects/
- https://flowingdata.com/2009/10/01/30-resources-to-find-the-data-you-need/
- https://dsc.gmu.edu
- Uber
  - https://data.cityofnewyork.us/Transportation/uber-Data/3jeu-mn7j
  - https://help.uber.com/riders/article/download-your-data?nodeId=2c86900d-8408-4bac-b92a-956d793acd11
  - https://data.world/datasets/uber
  - https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city
  - https://www.reddit.com/r/uber/comments/74c2mn/export_uber_trips_to_csv/

Links to data

- Medicare
  - https://data.medicare.gov
  - https://www.medicare.gov/download/downloaddb.asp
  - https://www.cms.gov/newsroom/data
  - https://www.kaggle.com/cms/cms-medicare
  - https://catalog.data.gov/dataset?q=medicare
  - https://www.ama-assn.org/practice-management/medicare/medicare-claims-data-release
  - https://www.cms.gov/OpenPayments/Explore-the-Data/Dataset-Downloads.html
  - https://data.medicare.gov/download-data
  - https://data.medicaid.gov
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5433516/
  - https://fcw.com/articles/2018/03/06/cms-blue-button-api.aspx

Links to data

- https://data.humdata.org/group/bhs
- https://catalog.data.gov/dataset?tags=bahamas
- https://github.com/GBPA/gis-datafiles
- https://www.digitalglobe.com/ecosystem/open-data
- https://www.nga.mil/MediaRoom/PressReleases/Pages/dorian.aspx
- Where else?
    - FEMA
    - NOAA
    - USAID
    - USGS EarthExplorer

Auxiliary

- UCI repositories: archive.ics.uci.edu/ml/datasets.php
- reddit.com/r/datasets
- Health
    - ibi.gmu.edu/faculty-directory/niloofar-ramezani