



CSCE 580: Introduction to AI *CSCE 581: Trusted AI*

Lecture 2: Data for AI

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

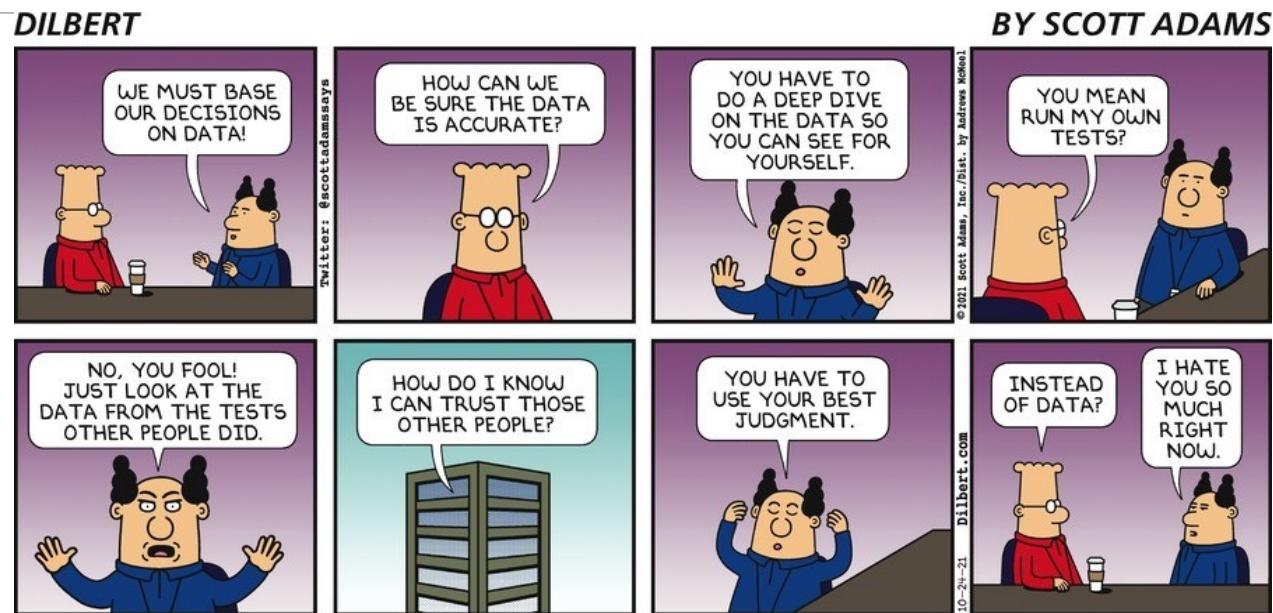
29TH AUG 2023

Carolinian Creed: “I will practice personal and academic integrity.”

Credits: Copyrights of all material reused acknowledged

Organization of Lecture 2

- Introduction Segment
 - Recap of Lecture 1
- Main Segment
 - Types
 - Structured,
 - Semi-structured,
 - Unstructured,
 - By media: text, audio, video, multi-media;
 - Open Data
 - City data: access
 - Data quality
- Concluding Segment
 - About Next Lecture – Lecture 3
 - Ask me anything



Credit: Dilbert

Introduction Section

Exercise: Session with ChatGPT

- Ask questions about Water usage
 - Experience
- Ask questions about Finance
 - Experience
- Hint:
 - Demand / supply questions: “can I drink water of Lake Murray”?, “will US have money to pay debt next year”
 - Decision questions: “which water should I choose between a bottled one and tap”?
 - Factoid questions: “is pH of 7 good for drinking water?”

Exercise: Solving Games with AI

- Popular way to learn AI is via games
 - <https://github.com/biplav-s/course-ai-tai-f23/blob/main/sample-code/Class1-games.md>

Main Section

Data – The Fuel for AI

Overview: Types of Data

- By content structure: Structured, unstructured and semi-structured
- By media: text, audio, visual, multi-media

- By source
 - Open data
 - Social data
 - Sensor data
 - Proprietary data

- Value is by fusing data across all types
 - sources, content structure and media

Types of Data - Structured

- The structure of data is fixed. Example: columns in a database
- Benefits
 - Can be stored and queried efficiently, e.g., by commercial databases
 - Easy to analyze, e.g., by SQL or programs – pandas in Python
- Disadvantage
 - Hard to handle data's structural changes. E.g., adding a new column. Complex data migration procedures

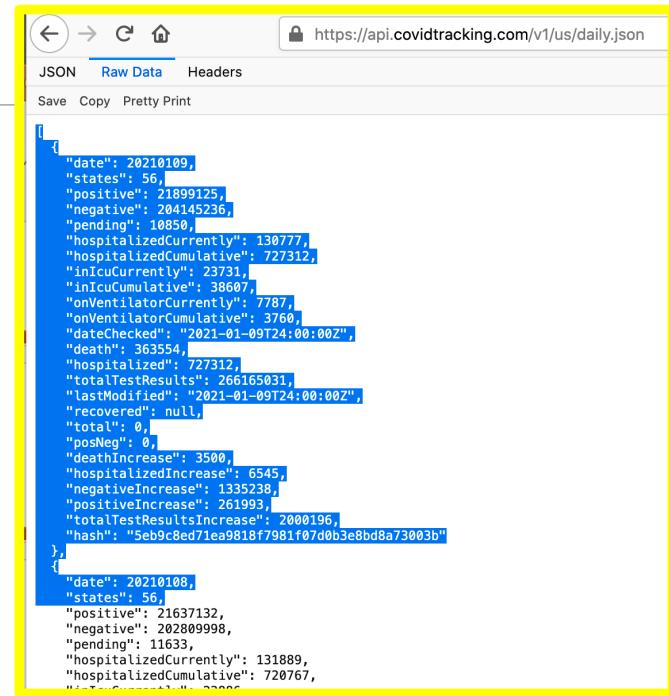
```
country,placename,frequency,start_date,end_date,year,month,week,deaths,expected_deaths,excess_deaths,baseline  
...  
France,,weekly,2020-04-27,2020-05-03,2020,4,18,10498,10357,141,2010-2018 weekly average  
...
```

Source: <https://github.com/nytimes/covid-19-data/tree/master/excess-deaths>

NYT COVID datasets: <https://github.com/nytimes/covid-19-data/>

Types of Data – Semi-Structured

- The structure of meta-data is fixed, but the structure of data is allowed to change. Example: XML, JSON
- Benefits
 - Relatively easy to analyze, e.g., commands similar to SQL in languages like OQL or Xquery
 - Structure of data easy to extend
- Disadvantage
 - Size of data is larger than structured representation as metadata is added with each record



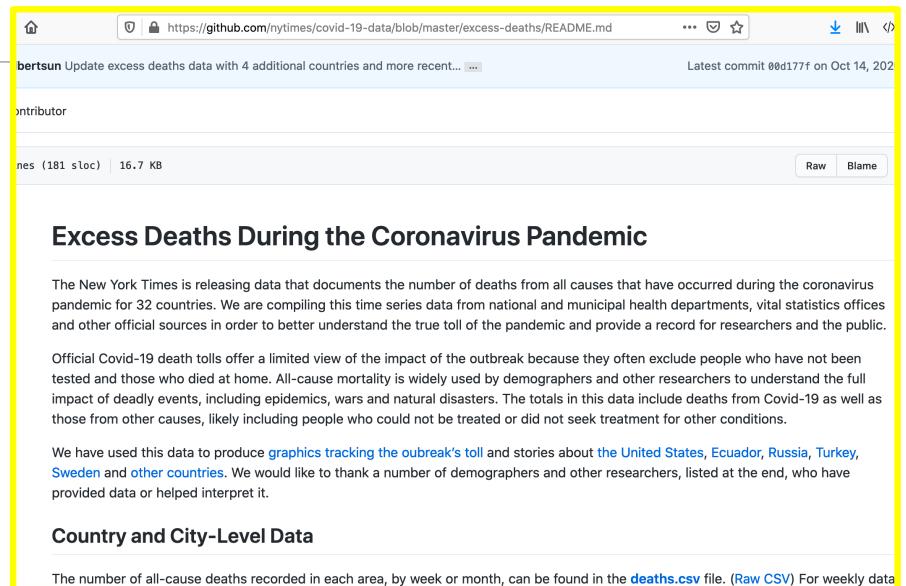
The screenshot shows a browser window displaying JSON data from the URL <https://api.covidtracking.com/v1/us/daily.json>. The browser interface includes a back button, forward button, refresh button, and a home icon. Below the address bar, there are tabs for 'JSON', 'Raw Data', and 'Headers'. Underneath the tabs are buttons for 'Save', 'Copy', and 'Pretty Print'. The main content area displays two arrays of objects. The first array represents data for January 10, 2021, with fields like 'date', 'states', 'positive', 'negative', 'pending', 'hospitalizedCurrently', 'hospitalizedCumulative', 'inICUCurrently', 'inICUCumulative', 'onVentilatorCurrently', 'onVentilatorCumulative', 'death', 'dateChecked', 'deathIncrease', 'hospitalizedIncrease', 'negativeIncrease', 'positiveIncrease', 'totalTestResults', 'lastModified', 'recovered', 'total', 'posNeg', and 'hash'. The second array represents data for January 10, 2021, with similar but slightly different field names and values.

```
[{"date": "20210109", "states": 56, "positive": 21899125, "negative": 204145236, "pending": 10850, "hospitalizedCurrently": 130777, "hospitalizedCumulative": 727312, "inICUCurrently": 23731, "inICUCumulative": 38607, "onVentilatorCurrently": 7787, "onVentilatorCumulative": 3760, "death": 36354, "dateChecked": "2021-01-09T24:00:00Z", "deathIncrease": 3500, "hospitalizedIncrease": 6545, "negativeIncrease": 1335238, "positiveIncrease": 261993, "totalTestResults": 266165031, "lastModified": "2021-01-09T24:00:00Z", "recovered": null, "total": 0, "posNeg": 0, "deathIncrease": 3500, "hospitalizedIncrease": 6545, "negativeIncrease": 1335238, "positiveIncrease": 261993, "totalTestResultsIncrease": 2000196, "hash": "5eb9c8ed71ea9818f7981fb0d0b3e8bd8a73003b"}, {"date": "20210108", "states": 56, "positive": 21637132, "negative": 202809998, "pending": 11633, "hospitalizedCurrently": 131889, "hospitalizedCumulative": 720767, "inICUCurrently": 23731, "inICUCumulative": 38607, "onVentilatorCurrently": 7787, "onVentilatorCumulative": 3760, "death": 36354, "dateChecked": "2021-01-08T24:00:00Z", "deathIncrease": 3500, "hospitalizedIncrease": 6545, "negativeIncrease": 1335238, "positiveIncrease": 261993, "totalTestResults": 266165031, "lastModified": "2021-01-08T24:00:00Z", "recovered": null, "total": 0, "posNeg": 0, "deathIncrease": 3500, "hospitalizedIncrease": 6545, "negativeIncrease": 1335238, "positiveIncrease": 261993, "totalTestResultsIncrease": 2000196, "hash": "5eb9c8ed71ea9818f7981fb0d0b3e8bd8a73003b"}]
```

Source: <https://api.covidtracking.com/v1/us/daily.json>

Types of Data – Unstructured

- The data has no structure.
Example: text
- Benefits
 - Easy to change structure
 - Content can be compactly stored
- Disadvantage
 - Hard to analyze content. Example: word analysis, sentiments, topic, ...



Source: <https://github.com/nytimes/covid-19-data/blob/master/excess-deaths/README.md>

NYT COVID datasets: <https://github.com/nytimes/covid-19-data/>

Textual Data

- Media: text
- Components: characters, words, paragraph
- Representation
 - Uncompressed / encoding – ASCII, UTF-8, UTF-16
 - Compressed - .zip
 - Lossy compression -
- Language: English, French, ...
- Programming libraries: nltk, spacy

Filename extension	.txt
Internet media type	text/plain
Type code	TEXT
Uniform Type Identifier (UTI)	public.plain-text
UTI conformation	public.text
Type of format	Document file format , Generic container format

Details: https://en.wikipedia.org/wiki/List_of_file_formats

Sound

- Media: sound
- Components: phoneme
- Representation
 - Uncompressed - .wav, .aiff
 - Compressed lossless -
 - Lossy compression - .mp3, .aac (iTunes)
- Programming libraries: [playsound](#), [simpleaudio](#), [winsound](#), [python-sounddevice](#), [pydub](#), [pyaudio](#)

Details: https://en.wikipedia.org/wiki/Audio_file_format

Filename extension	.wav .wave
Internet media type	audio/vnd.wave, ^[1] audio/wav, audio/wave, audio/x-wav ^[2]
Type code	WAVE
Uniform Type Identifier (UTI)	com.microsoft.waveform-audio
Developed by	IBM & Microsoft
Initial release	August 1991; 29 years ago ^[3]
Latest release	Multiple Channel Audio Data and WAVE Files (7 March 2007; 13 years ago (update) ^{[4][5]})
Type of format	audio file format , container format
Extended from	RIFF
Extended to	BWF , RF64

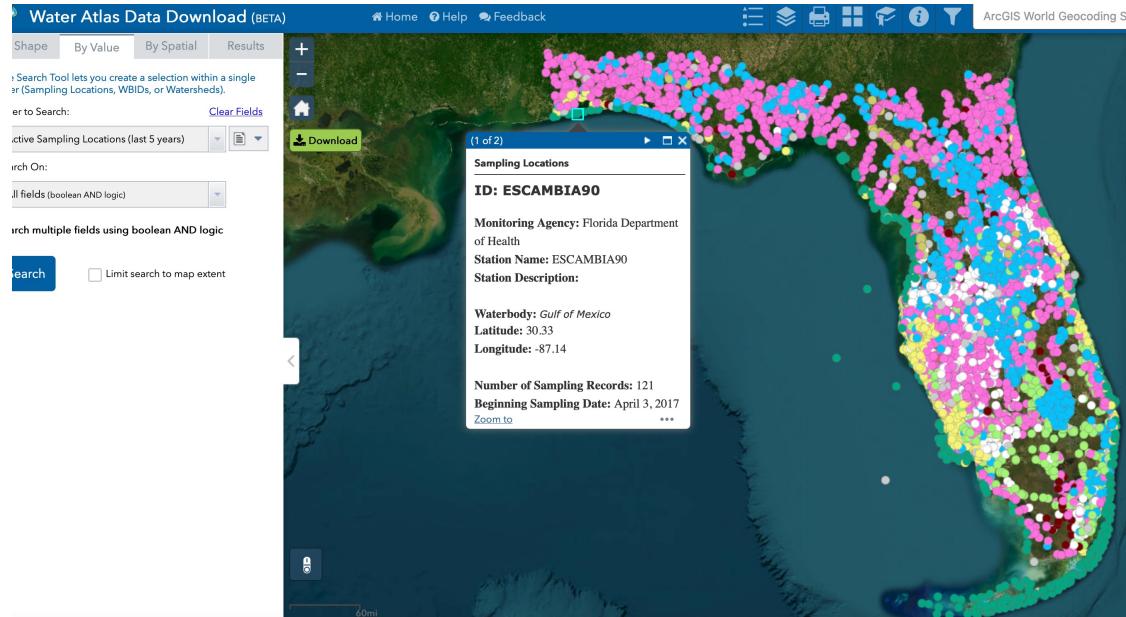
Visual

- Media: image, video
- Components: pixel, frame
- Representation
 - Uncompressed – bitmap
 - Compressed lossless - .gif
 - Lossy compression - .jpeg
 - Containers: AVI (.avi) and QuickTime (.mov)
- Programming libraries: PIL, OpenCV

<u>Filename extension</u>	.avi
<u>Internet media type</u>	video/vnd.avi ^[1]
<u>Type code</u>	'Vfw '
<u>Uniform Type Identifier (UTI)</u>	public.avi
Developed by	Microsoft
Initial release	November 1992; 27 years ago
<u>Container for</u>	Audio, Video
Extended from	Resource Interchange File Format

Open Data Should Not to Be Confused With Orthogonal Trend – Big Data

Volume
Variety
Velocity
Veracity
...



Data: <https://github.com/biplav-s/course-tai/tree/main/sample-code/common-data/water>



"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

Cartoon critical of big data application,
by T. Gregorius

http://upload.wikimedia.org/wikipedia/commons/thumb/b/b3/Big_data_cartoon_t_gregorius.jpg/220px-Big_data_cartoon_t_gregorius.jpg

Open Data

“Open data and content can be **freely used, modified, and shared by anyone for any purpose**”

<http://opendefinition.org/od/2.1/en/>

Open Data is an Old Concept in a New Setting

- Open data is the notion that data should not be hidden, but made available to everyone to **reuse**. **The idea is not new.**
- Scientific publications follow this: “standing on the shoulders of giants”
- Data quality and open publishing process is critical

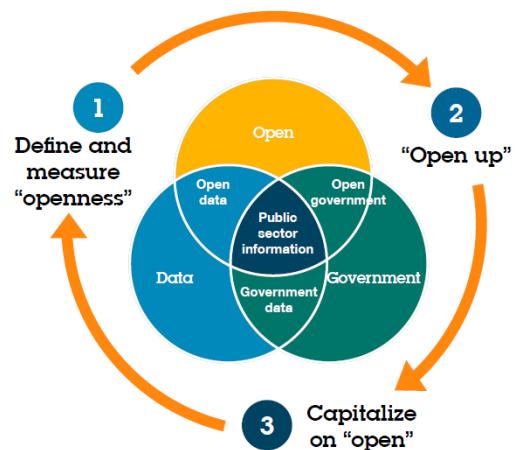
A screenshot of the US Data.gov website (<https://catalog.data.gov/dataset>). The page features a navigation bar with links for DATA, TOPICS, RESOURCES, STRATEGY, DEVELOPERS, and CONTACT. Below the navigation is a grid of icons representing various data categories: Agriculture, Climate, Ecosystems, Energy, Local Government, Maritime, Ocean, and Older Adults Health. Two specific datasets are highlighted: "U.S. Hourly Precipitation Data" (855 recent views) and "NCDC Storm Events Database" (331 recent views). Each dataset entry includes a brief description, a map, and download links in various formats (HTML, JSON, CSV, etc.).

USA

A screenshot of the India data.gov.in website (<https://data.gov.in>). The page has a green header with the site's name and a search bar. The main content area features a large banner for "DATASETS FROM HEALTH SECTOR". Below the banner, there are three main sections: "ANALYTICS" (showing 395,534 resources, 8,380 catalogs, 173 departments, 28.58 M times viewed, 8.19 M times downloaded, 354 chief data officers, 32,392 APIs, and 2,043 visualizations), "CATALOG" (with a thumbnail of a lightbulb icon), and "INDICATOR DASHBOARD" (with cards for Drinking Water And Sanitation, Health, Transport, and Labour And Employment).

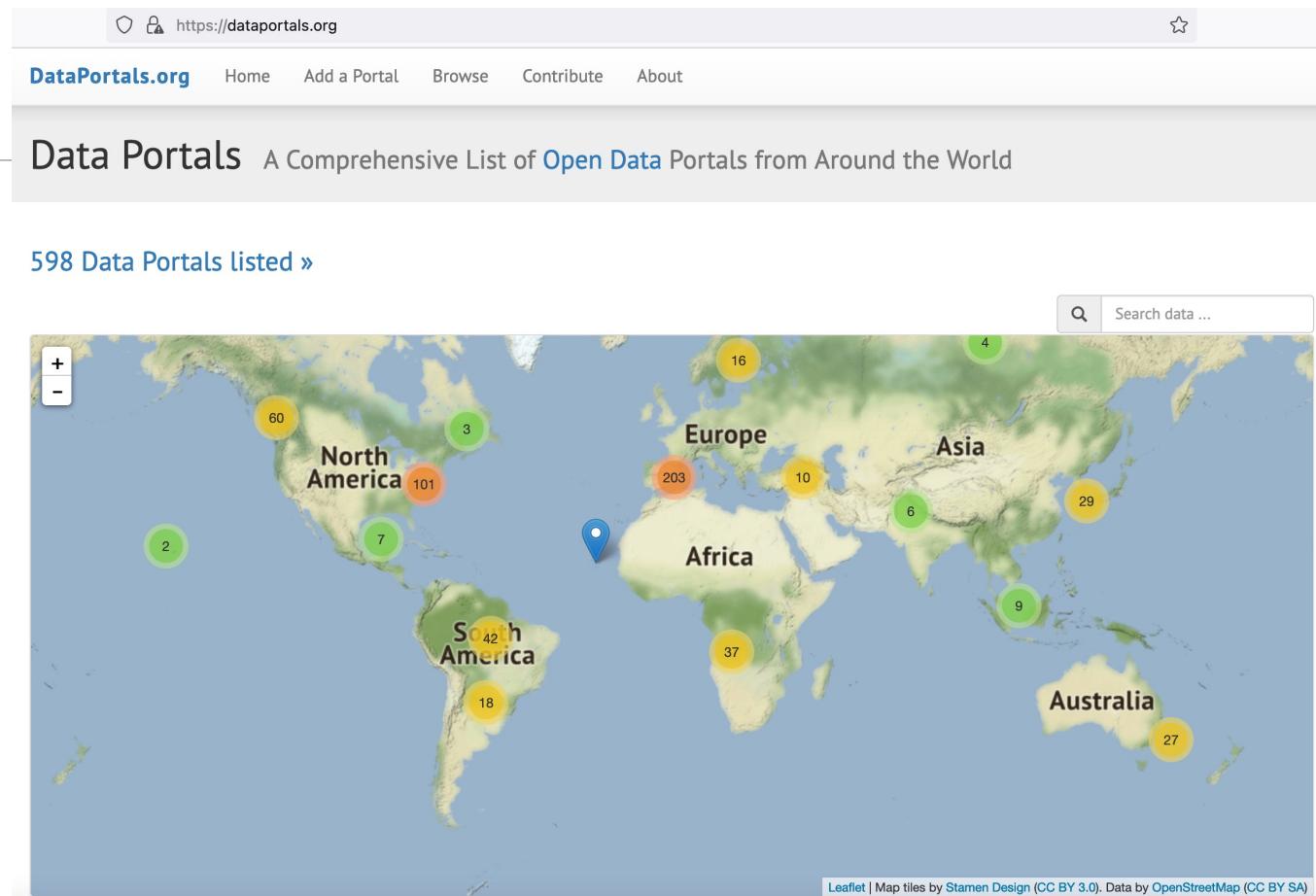
India

~600 Data Catalogs of Open Data



Source: IBM Institute for Business Value.

As on 26 Aug 2023



Demo: US Open Data

- Site: <https://data.gov>
- Tools: <https://resources.data.gov/categories/data-tools/>

Open Datasets

- data.gov OF ANY COUNTRY
 - Portal: <https://dataportals.org/>
 - US: <https://www.data.gov/> or any US state
 - India: <https://data.gov.in>
- Text of legislations - LegiScan, <https://legiscan.com/>
- Kaggle datasets: <https://www.kaggle.com/datasets>
- Google datasets search:
<https://datasetsearch.research.google.com/>

City Dashboard - London

CityDashboard aggregates simple spatial data for cities around the UK and displays the data on a dashboard and a map.

<http://citydashboard.org/london/>
<http://citydashboard.org/about.php>

[Birmingham](#)
[Brighton](#)
[Cardiff](#)
[Edinburgh](#)
[Glasgow](#)
[Leeds](#)
[London](#)
[Manchester](#)

Sat 26 Aug @ 22:23:19
Go to Map - Go to Grid - Change City

WEATHER STATION (CASA TEAM) 12
STATION WIND SPEED WIND GUSTS DIRECTION TEMPERATURE HUMIDITY RAIN TODAY PRESSURE FORECAST
CASA Office: Bloomsbury W1 Data not updated for 11442 hours

WEATHER (METAR) 871
London City Winds W-280 at 8kt, Vis 10km, Scattered clouds at 4500ft SW at 6 mph 14 C

TRAFFIC CAMERAS (TfL) 3
York Road/Leake Street Camera 00001.04226 unobtainable

TUBE LINE STATUS (TfL) 1
Bakerloo Good Service
Central Good Service
Circle Good Service
District Good Service
DLR Good Service
Elizabeth Good Service
H & C Good Service
Jubilee Good Service
Overground Part Closure
Metropolitan Good Service
Northern Good Service
Piccadilly Part Closure
Trams Good Service
Victoria Good Service

LONDON CYCLE HIRE (TfL) 61
NAN % NAN %
Stations Full Stations Empty
0 0
Bikes Available Bikes or Docks Faulty

IN SERVICE (TfL) 1
6092 London buses
322 Underground trains

AIR POLLUTION (DEFRA) 1771
µg/m³ TIME AVG OZONE NO₂ SO₂ PM_{2.5} PM₁₀
Bloomsbury
Marylebone Rd
N Kensington

BICYCLES (LBH) 3571
Goldsmiths' Row 4012 yesterday

STOCKS (YAHOO) 8
FTSE 100 Index 7121.88 91.22 (1.28%)

TRAFFIC CAMERAS (TWO AT RANDOM) (TfL) 12
75 Knightsbridge/Williams St Sun 27 Aug 03:11 Camera 00001.06730 unobtainable
London Rd/Arragon Rd Sun 27 Aug 02:43
A4 Knightsbridge by Albert Gate London Rd/Arragon Rd

BBC LONDON NEWS (BBC) 71
Bow fire: Homes 'severely damaged' in east London
blaze Fresh dates for London hot air balloon event after summer cancellations Superloop: West London express Heathrow to Harrow bus service launched

OPENSTREETMAP UPDATES (OSM) 271
Edit to future cycle route Edit to future cycle route
Edit to future cycle route Edit to future cycle route Mapped planned C35 route at Peckham Rye Update addresses in SW19 postal dist. kxplus kxplus

[Tweet](#) [About](#)

Attempt for Dashboards - Amsterdam



[2016] <http://citydashboard.waag.org/>

Exercise 1 - Explore

1. Google data search tool: <https://datasetsearch.research.google.com/>
2. US open data: <https://www.data.gov/>
3. Select a problem domain and search for data
4. Discuss your experience

Accessing Data

Example: Open 311 (<http://open311.org/>)

Refers to non-emergency events like graffiti, garbage, down trees, abandoned car, ...

- Not human life threatening
- 60+ cities support it world-wide

Discovering Open 311 of a City

<http://311api.cityofchicago.org/open311/discovery.json>

```
changeset          "2012-09-14T08:00:00-05:00"
contact            "Contact developers@cityofchicago.org for assistance"
key_service         "Visit http://test311api.cityofchicago.org/open311 to request an API Key"
endpoints          0
specification      "http://wiki.open311.org/GeoReport\_v2"
url                "http://311api.cityofchicago.org/open311/v2"
changeset          "2012-09-14T08:00:00-05:00"
type               "production"
formats            0
                   "text/xml"
                   "application/json"
1
specification      "http://wiki.open311.org/GeoReport\_v2"
url                "http://test311api.cityofchicago.org/open311/v2"
changeset          "2012-09-14T08:00:00-05:00"
type               "test"
formats            0
                   "text/xml"
                   "application/json"
1
```

The screenshot shows a JSON viewer interface with the URL 311api.cityofchicago.org/open311/discovery.json. The JSON data is displayed in a hierarchical tree view. The root object contains fields like changeset, contact, key_service, and endpoints. The endpoints array has two items (index 0 and index 1). Each endpoint item contains fields such as specification, url, changeset, type, and formats. The 'specification' field for both items points to http://wiki.open311.org/GeoReport_v2. The 'url' field for index 0 points to <http://311api.cityofchicago.org/open311/v2>, and for index 1 it points to <http://test311api.cityofchicago.org/open311/v2>. The 'changeset' field for both items is "2012-09-14T08:00:00-05:00". The 'type' field for index 0 is "production" and for index 1 is "test". The 'formats' field for both items contains two entries: "text/xml" and "application/json". A red box highlights the second endpoint entry (index 1) in the JSON tree.

Demonstration: Open 311

List of services

- <http://311api.cityofchicago.org/open311/v2/services.json>
 - Result
-

```
[{"service_code": "4ffa4c69601827691b000018", "service_name": "Abandoned Vehicle", "description": "Abandoned vehicles are taken to auto pound 3S or 3N where they are -- if not redeemed by the owners -- sold for scrap.", "metadata": true, "type": "batch", "keywords": "code:SKA", "group": "Streets & Sanitation"},
```

```
{"service_code": "4ffa9cad6018277d4000007b", "service_name": "Alley Light Out", "description": "One or more alley lights out, on a wooden pole in the alley itself, are reported under this service request type. Important information needed when reporting alley lights out includes: the exact address that the light/lights are behind, how many lights are out, and if the light(s) are completely out or if they blink on and off intermittently. Alley light repairs are done during the day when the lights are not on, so this information is essential to expedite the repair work.", "metadata": true, "type": "batch", "keywords": "code:SFA", "group": "Transportation"},
```

```
...]
```

Details of a service

- <http://311api.cityofchicago.org/open311/v2/services/4ffa4c69601827691b000018.json>
 - Result
- ```
{"service_code": "4ffa4c69601827691b000018",
"attributes": [
{"variable": true, "code": "FQSKA1",
"datatype": "singlevaluelist", "required": false, "order": 1,
"description": "Vehicle Make/Model",
"values": [
{"key": "ASVEAV", "name": "(Assembled From Parts,Homemade)" },
 {"key": "HOMDCYL", "name": "(Homemade Motorcycle, Moped.Etc.)" },
 {"key": "HMDETL", "name": "(Homemade Trailer)" }, ...
]
...
]}}
```

# Demonstration: Open 311

---

<http://311api.cityofchicago.org/open311/v2/services/4ffa9cad6018277d4000007b.json>

Result

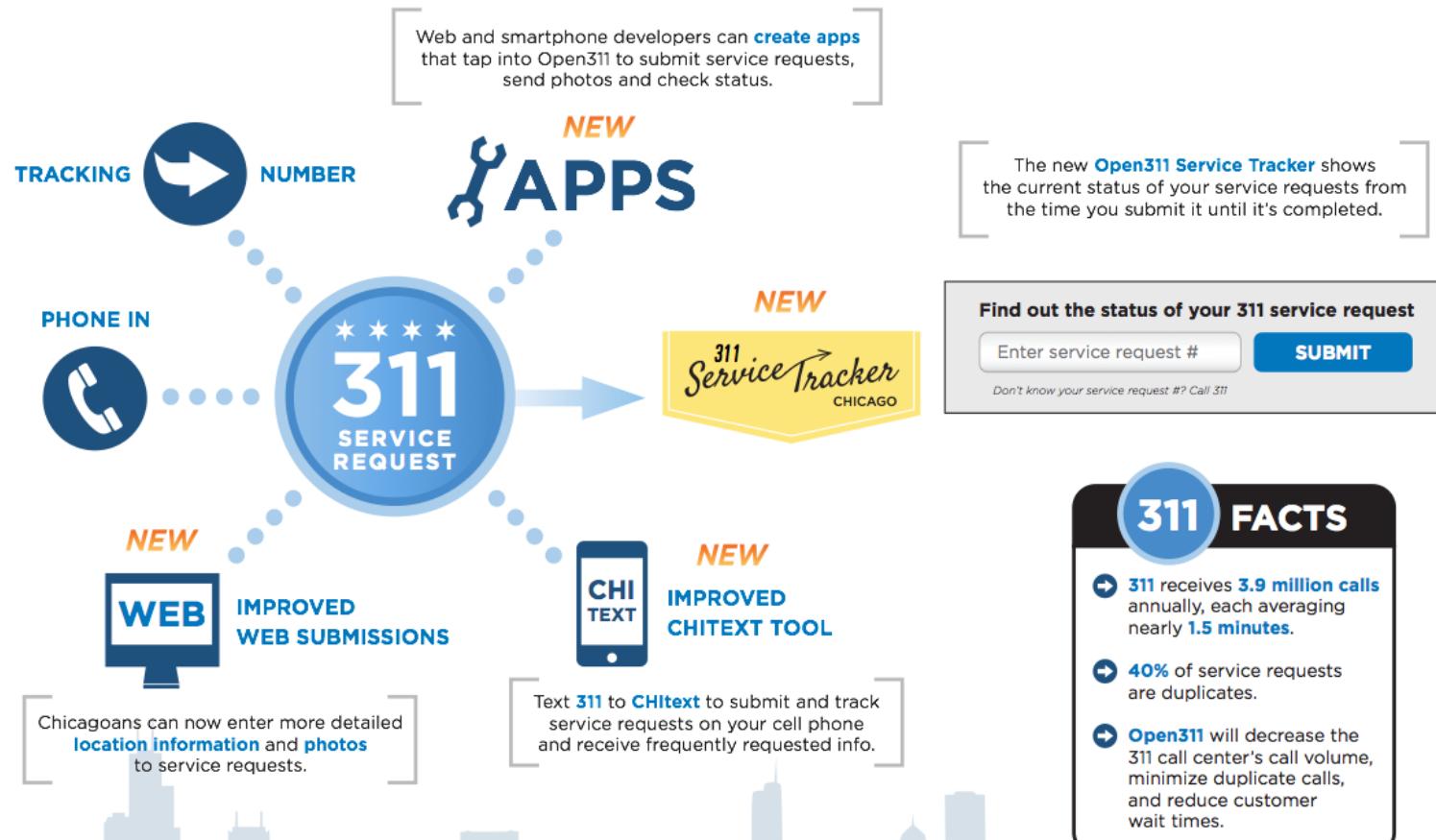
```
{"service_code":"4ffa9cad6018277d4000007b",
 "attributes":
 [{"variable":true,"code":"ISTHELI2",
 "datatype":"singlevaluelist","required":true,"order":1,
 "description":"Is the light located in your alley or the street?",
 "values":[{"key":"ALLEY","name":"Alley"},
 {"key":"STREET","name":"Street"}]},

 {"variable":true,"code":"POLEWORM",
 "datatype":"singlevaluelist","required":true,"order":2,
 "description":"Is the pole wooden or metal?",
 "values":[{"key":"METAL","name":"Metal"},
 {"key":"WOODEN","name":"Wooden"}]},

 {"variable":true,"code":"ISTHELI3",
 "datatype":"singlevaluelist","required":true,"order":3,
 "description":"Is the light directly behind this address?",
 "values":[{"key":"NO","name":"No - Light Not Directly Behind Address"},
 {"key":"YES","name":"Yes - Light Directly Behind Address"}]},

 {"variable":true,"code":"A511OPTN",
 "datatype":"string","required":false,
 "datatype_description":"Enter number as 999-999-9999","order":4,
 "description":"Input mobile # to opt-in for text updates. If already opted-in, add mobile # to contact info."}]}
```

# Chicago: Service Tracking



# Example: Application over Open Data (Chicago)

The screenshot shows a web browser displaying the Chicago 311 Service Tracker website at [servicetracker.cityofchicago.org/requests/13-00210540](http://servicetracker.cityofchicago.org/requests/13-00210540). The page title is "Rodent Baiting / Rat Complaint". Key details include:

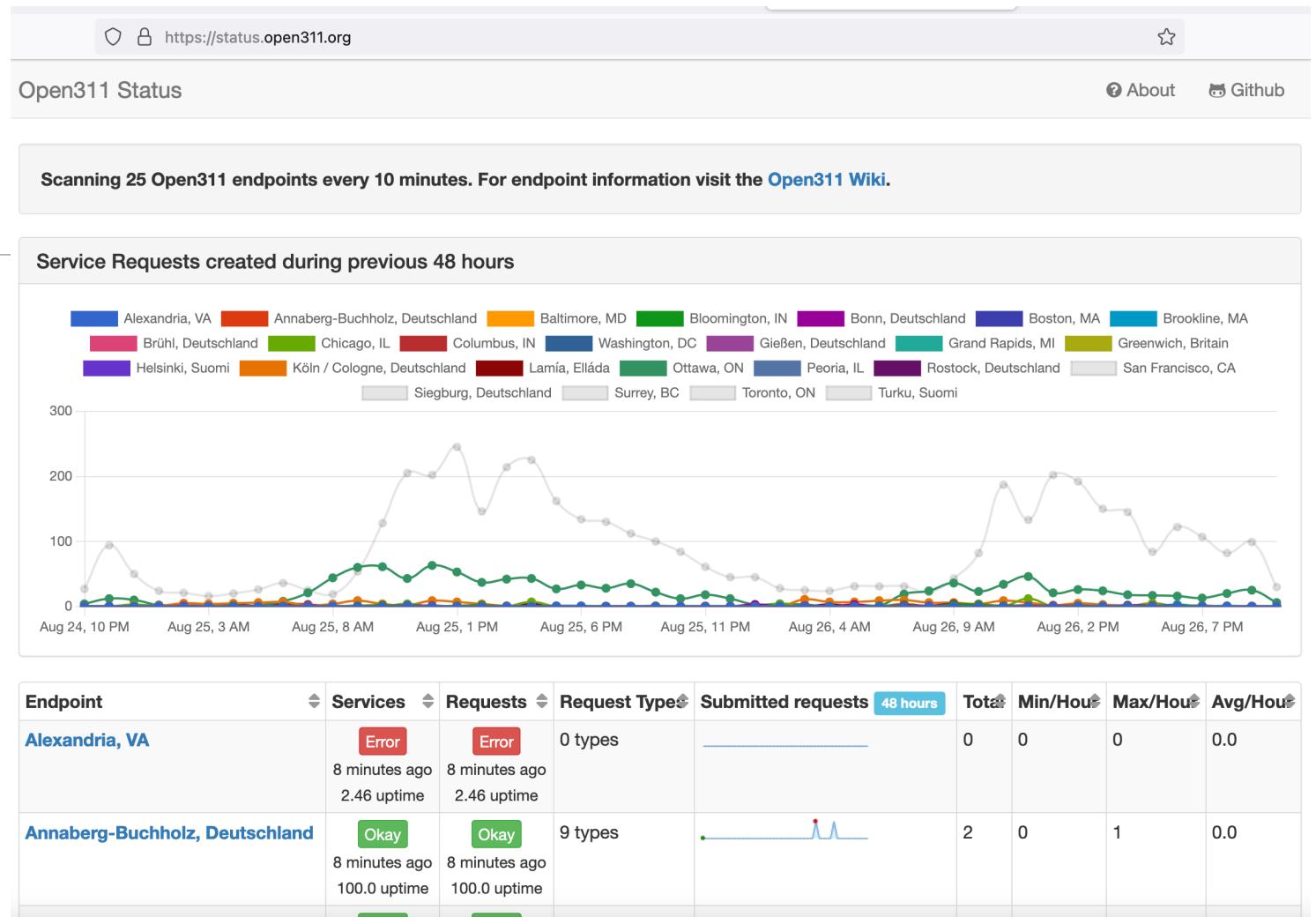
- #13-00210540**
- Address:** 1502 N Wicker Park Ave
- Created:** February 23, 2013
- Received via:** Other

A green ribbon on the right indicates the status is **Closed**.

**Activity**

| Date                    | Action                                                                                    |
|-------------------------|-------------------------------------------------------------------------------------------|
| 05-Mar-2013<br>10:04 AM | Request closed                                                                            |
| 05-Mar-2013<br>10:04 AM | Dispatch Crew Completed                                                                   |
| 23-Feb-2013<br>10:16 PM | Rodent Baiting / Rat Complaint<br>Department: Bureau of Rodent Control - S/S<br>via Other |

# Scaling with Open 311



# Exercise 2 – Programmatically Access Data

---

1. See sample code on GitHub:

- <https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/I2-opendata/Explore%20OpenData.ipynb>

2. Explore APIs of another city of your choice

# Exercise 3 – Programmatically Access Data

---

1. Water data
2. Text data

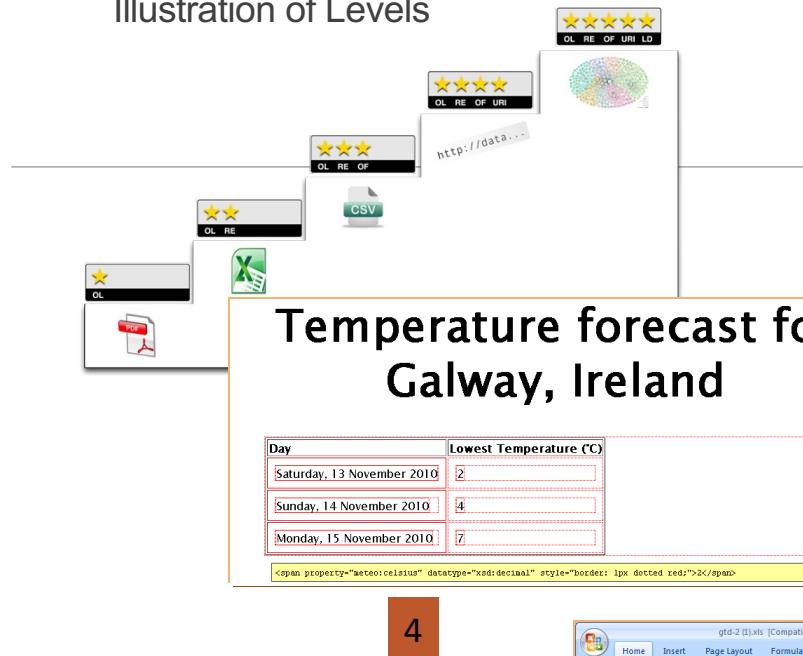
Code samples: <https://github.com/biplav-s/course-ai-tai-f23/blob/main/sample-code/Class2-data.md>

# Quality of Data

---

## Does Opening Data Make It Reusable? No

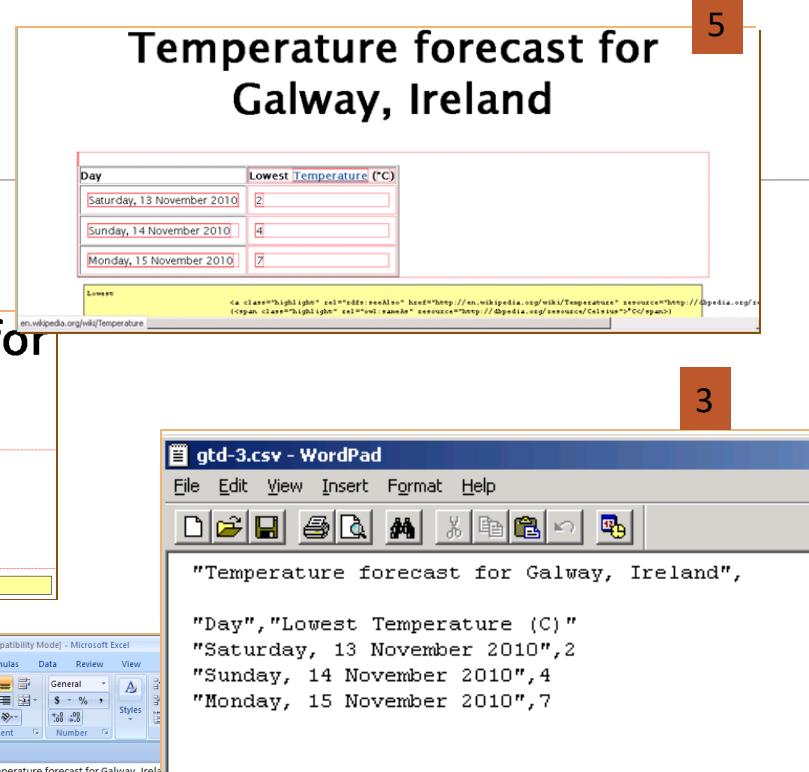
Illustration of Levels



Source: <http://5stardata.info/>

| Temperature forecast for Galway, Ireland |                         |
|------------------------------------------|-------------------------|
| Day                                      | Lowest Temperature (°C) |
| Saturday, 13 November 2010               | 2                       |
| Sunday, 14 November 2010                 | 4                       |
| Monday, 15 November 2010                 | 7                       |

1



34

# Data Quality of Public Data in India



## Right to Information

- Not even 1\*
- Information available to requester, but no one else

## Data.gov.in

- 2-3\*
- Available in CSV, etc but not uniquely referenceable

Open data movements are moving to linked data form for semantics

# Annotated – Indian Open Data

---

Vocabulary services: <http://vocab.nic.in/index.php>

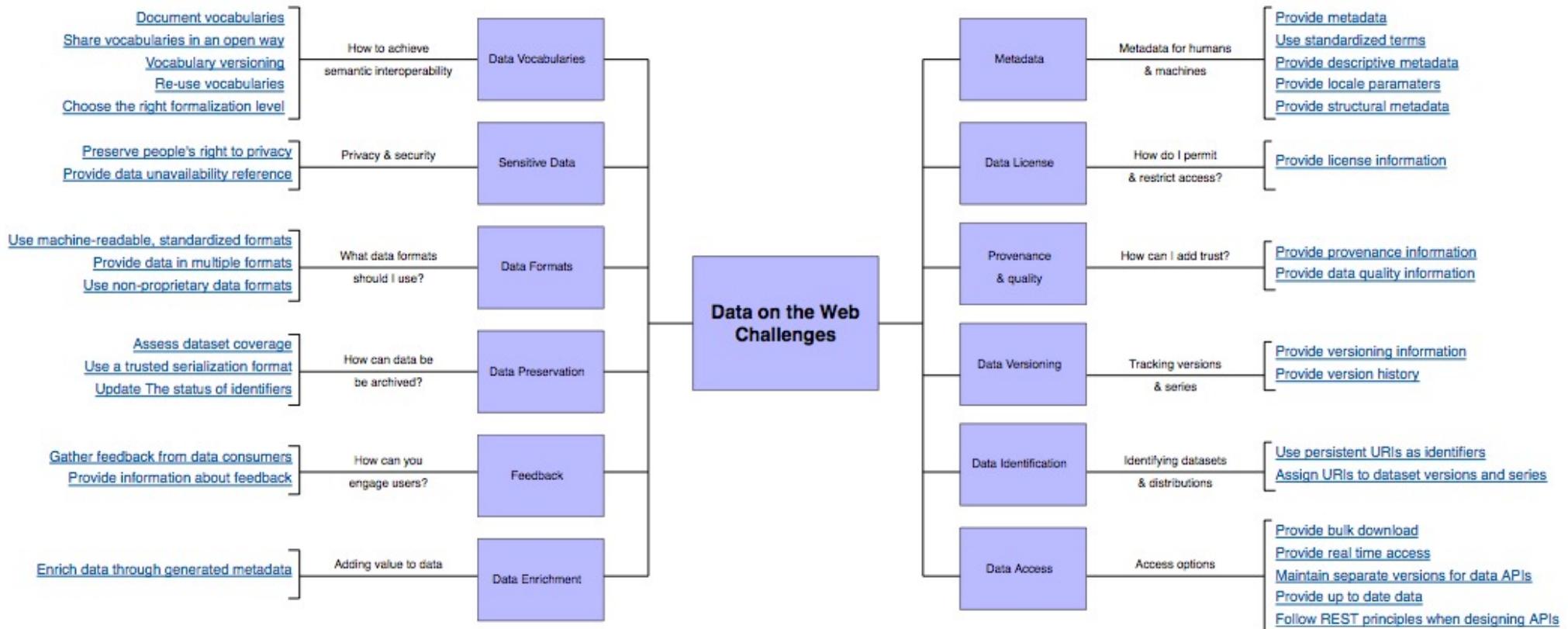
- Authoritative
- Standardized codes

## Examples

- States in the Union: <http://vocab.nic.in/rest.php/states/json>
- Districts in a state (“UP”): <http://vocab.nic.in/rest.php/district/up/json>
- State legislatures: <http://vocab.nic.in/rest.php/orgn/sg/legislature/json>
- Union government offices in a state (“TN”): <http://vocab.nic.in/rest.php/orgn/ug/state/tn/json>

# Helping Publish Good Quality Open Data is Key

Have data policy in place  
 Publish with best practices, have semantics, promote reuse  
 Figure courtesy: <http://www.w3.org/TR/2015/WD-dwbp-20150625/>



# Concluding Section

---

# Lecture 2: Concluding Comments

---

- We talked about
  - Data formats
  - Big data v/s open data
  - Open data
    - City data
    - Data access via Open311
    - Publishing data systematically

# About Next Lecture – Lecture 3

---

# Project Discussion: What Problem Fascinates You ?

---

- Data
  - Water
  - Finance
- Analytics
- Application
  - Building chatbot
- Users
  - Diverse demographics
  - Diverse abilities
  - Multiple human languages

Project execution in sprints

- Sprint 1:
  - **Solving:** Choose a decision problem, identify data, work on solution methods
  - **Human interaction:** Develop a basic chatbot (no AI), no problem focus
- Sprint 2:
  - **Solving:** Evaluate your solution on problem
  - **Human interaction:** Integrated your choice of chatbot (rule-based or learning-based) and methods
- Sprint 3:
  - **Evaluation:** Comparison of your solver chatbot with an LLM-based alternative, like ChatGPT

# Project Discussion

---

# Lecture 3: Representing and Organizing Data

---

- Data preparation
- Knowledge representation/ graph
- Ontology