



CSCE 590-1: From Data to Decisions with Open Data: A Practical Introduction to AI

Lecture 2: Data

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

14TH JAN 2021

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

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
- Concluding Segment
 - About Next Lecture – Lecture 3
 - Ask me anything

Introduction Segment

Recap of Lecture 1

- We did a quick overview of the course
- Looked at AI, Data and Decisions
- Course will focus on
 - Practical methods to derive insights from open data
 - Evaluation will be by via project, paper and quizzes
 - **Bring your ideas to your project**
- Exciting techniques to learn to impact the world around us

Questions from Lecture 1

- Math requirements
 - Linear Algebra, Differential Equations, Matrices, Discrete Mathematics
- Programming requirements
 - Data Structures, Software Engineering

Sub-areas of AI

- Representation: formal representation of knowledge.
 - Illustration: entities and their relationships, like last Russian Czar's family tree
 - Methods: Ontology, knowledge graph, word embedding, "Model"
- Reasoning: deriving conclusions from formally represented knowledge.
 - Illustration: Modus ponen – P implies Q. P is True. Hence Q must be true.
 - Methods: Deduction, Induction, Abduction, Proposition logic, First-order logic, Fuzzy logic
- Learning: drawing insights from data
 - Illustration: predict COVID cases in USA by end of the month
 - Methods: Machine Learning – Classification, Clustering, Association; Deep Neural Network

Example: Courses for a Student

- Decision: Student deciding which courses to take for their program
- Data
 - About courses
 - About faculties
 - About job opportunities
 - About research opportunities and industry trends
 - **Private**: what the student wants to do
- Analysis
 - **Descriptive**: Courses offered in different semesters; Teachers offering courses
 - **Predictive**: How full will be a particular class next semester?
 - **Prescriptive**: Should a student take a particular course?

Example: Health During a Pandemic

- Decision: Individual staying healthy during a pandemic like COVID19
- Data
 - About cases
 - About mitigation steps: e.g., mask wearing restrictions and practices, lockdowns, hospital conditions
 - **Private**: pre-existing health conditions
- Analysis
 - **Descriptive**: Regions with high and low cases
 - **Predictive**: Does wearing mask help reduce cases?
 - **Prescriptive**:
 - Whether to eat inside a restaurant?
 - How to make an urgent road trip ?
 - How to hold classes at a University?

Main Segment

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

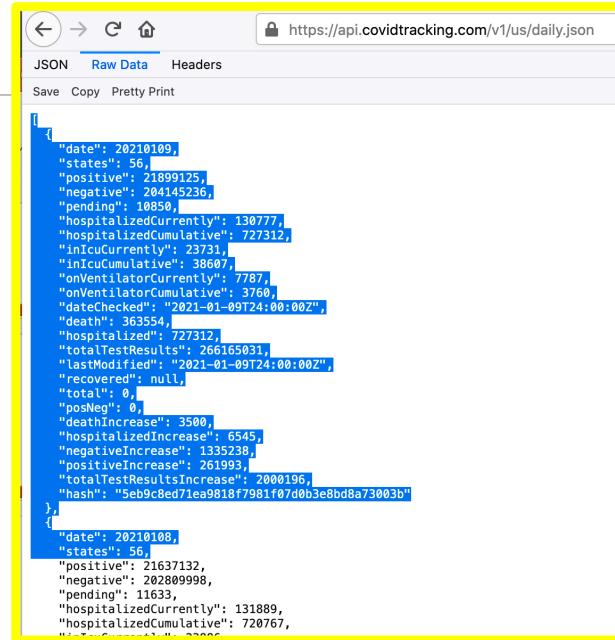
```
country,placename,frequency,start_date,end_date,year,month,week,deaths,expected_deaths,excess_deaths,baseline  
...  
France,,weekly,2020-04-27,2020-05-03,2020,4,18,10498,10357,141,2010-2018 weekly average  
...
```

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

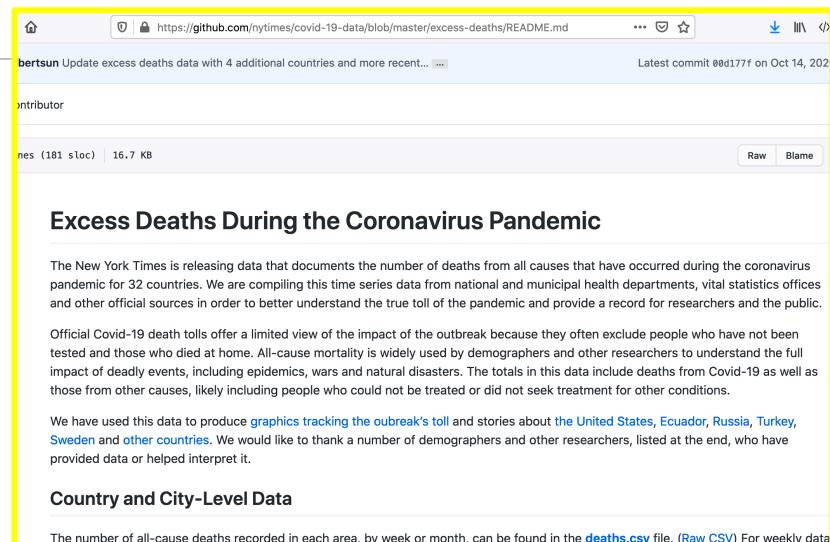


```
[{"date": "20210109", "states": 56, "positive": 21899125, "negative": 204145236, "pending": 10856, "hospitalizedCurrently": 130777, "hospitalizedCumulative": 727312, "inICUCurrently": 23731, "inICUCumulative": 38687, "onVentilatorCurrently": 7787, "onVentilatorCumulative": 3766, "dateChecked": "2021-01-09T24:00:00Z", "death": 36354, "hospitalized": 727312, "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": "5eb9c8ed71ea9818f7981f07d0b3e8bd8a73003b"}, {"date": "20210108", "states": 56, "positive": 21637132, "negative": 202809988, "pending": 11633, "hospitalizedCurrently": 131889, "hospitalizedCumulative": 720767, "inICUCurrently": 23200, "inICUCumulative": 38687, "onVentilatorCurrently": 7787, "onVentilatorCumulative": 3766, "dateChecked": "2021-01-08T24:00:00Z", "death": 36054, "hospitalized": 720767, "totalTestResults": 264165031, "lastModified": "2021-01-08T24:00:00Z", "recovered": null, "total": 0, "posNeg": 0, "deathIncrease": 3400, "hospitalizedIncrease": 6545, "negativeIncrease": 1335238, "positiveIncrease": 261993, "totalTestResultsIncrease": 2000196, "hash": "5eb9c8ed71ea9818f7981f07d0b3e8bd8a73003b"}]
```

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]
	Multiple Channel Audio Data and WAVE Files
Latest release	(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

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

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

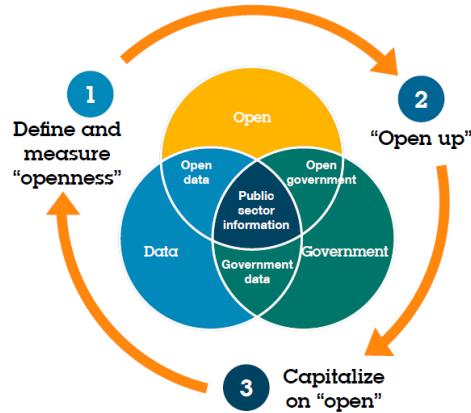
A screenshot of the Data.gov website for the USA. The top navigation bar includes links for DATA, TOPICS, RESOURCES, STRATEGY, DEVELOPERS, and CONTACT. Below the navigation are category icons for Agriculture, Climate, Ecosystems, Energy, Local Government, Maritime, and Ocean. A map of the United States is visible on the left. Two dataset cards are shown: "U.S. Hourly Precipitation Data" and "NCDC Storm Events Database". Both cards include links for HTML, JSON, XML, and CSV, along with download statistics.

USA

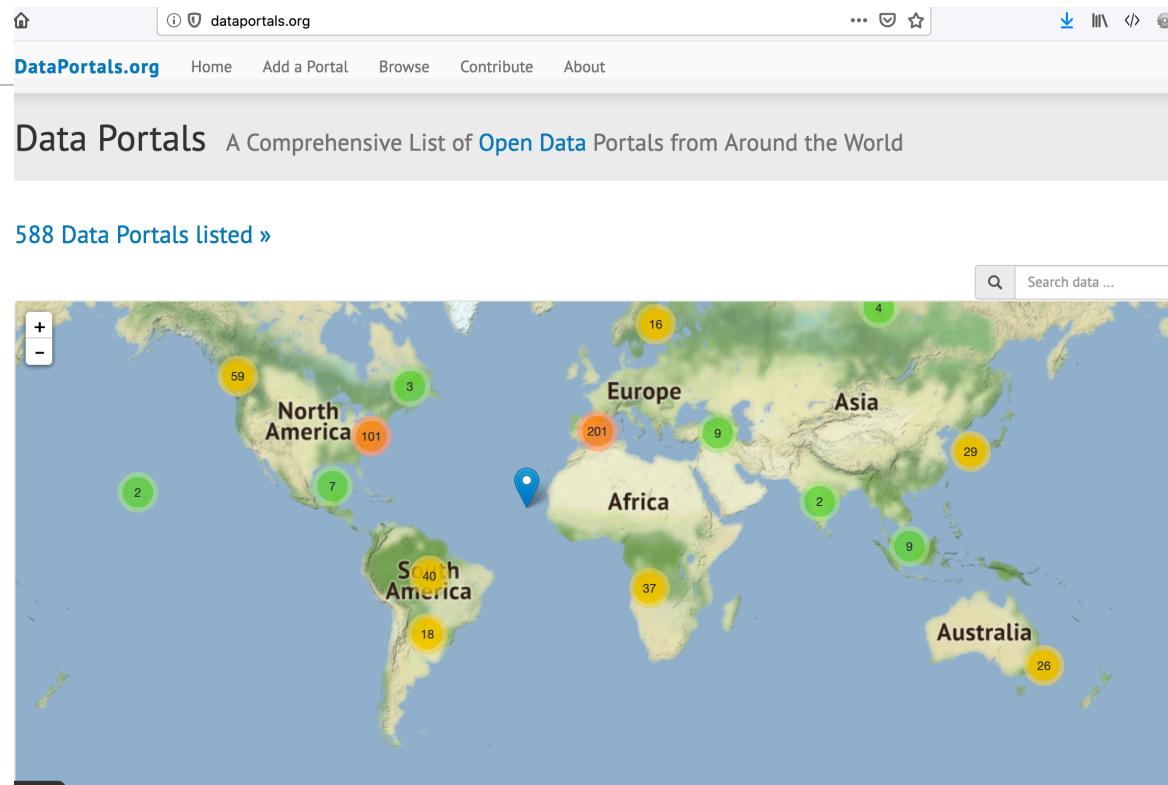
A screenshot of the data.gov.in website for India. The top navigation bar includes links for Skip to navigation, Skip to main content, DataGov Status/Help, LOG IN / REGISTER, and a search bar. A prominent banner at the top says "DATASETS FROM HEALTH SECTOR". Below the banner are three main sections: 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 lightbulb icon), and INDICATOR DASHBOARD (with icons for Drinking Water And Sanitation, Health, Transport, and Labour And Employment).

India

~600 Data Catalogs of Open Data



As on 20 June 2020



Demo: US Open Data

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

Indian Open Data

Illustration: a catalog and resources (data sets) within them

- Road Accidents; 76 datasets

The screenshot shows the data.gov.in website interface. At the top, there's a navigation bar with links for 'Skip to navigation', 'Skip to main content', 'DataGov States/ULB', and various accessibility icons. Below the header is a search bar with placeholder 'Type search keyword' and a magnifying glass icon. To the right of the search bar are 'LOG IN | REGISTER' buttons and a user profile icon.

The main content area displays a green banner titled 'Road Accidents in India 2018'. Below this, a 'Catalog Info' section provides details about the dataset, including the total number of road accidents, persons killed and injured, and other metrics. It also lists 'Released Under', 'Contributor', 'Keywords', 'Group', 'Sectors', 'Published on Data Portal', and 'Source' information.

On the right side, there are two resource cards for 'State/UT-wise Accidents Classified According to Type of Junctions during 2018' and 'State/UT-wise Accidents Classified According to Age of Impacting Vehicles during 2018'. Each card includes a large 'CSV' icon, a brief description, and a list of metadata such as granularity, file size, download count, and reference URLs.

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

Volume
Variety
Velocity
Veracity
...



"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

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>

https://citydashboard.org/london/

London

51.51 N, 0.13 W

Wed 13 Jan @ 22:00:10
Go to Map - Go to Grid - Change City

STATION	WIND SPEED	WIND GUSTS	DIRECTION	TEMPERATURE	HUMIDITY	RAIN TODAY	PRESSURE	FORECAST
CASA Office: Bloomsbury W1	2.2 mph	3.3 mph	SE ↗	1.9 °C	96%	5.146 mm	1015.0 mbar	Night Snow

WEATHER (METAR) London City
Winds E-090 at 8kt, Vis 10km, Light Rain, Broken layer at 800ft
E at 10 mph 5 C

TRAFFIC CAMERAS (TfL) High St/Collers Wd/Marlborough Rd


TUBE LINE STATUS (TfL)

Bakerloo	more
Central	more
Circle	more
District	more
DLR	more
H & C	more
Jubilee	more
Overground	Reduced Service more
Metropolitan	more
Northern	more
Piccadilly	more
TfL Rail	more
Trams	Good Service more
Victoria	more

LONDON CYCLE HIRE (TfL)
3.4 % Stations Full 3 % Stations Empty
11123 Bikes Available 290 Bikes or Docks Faulty


IN SERVICE (TfL)
570 London buses 23 Underground trains

AIR POLLUTION (DEFRAG)

μg/m³ TIME AVERD	OZONE	NO₂	SO₂	PM₂.₅	PM₁₀
Bloomsbury					
Marylebone					
Rd					
N Kensington					

BICYCLES (LBH) Goldsmiths' Row
1663 yesterday

STOCKS (YAHOO) FTSE 100 Index
6745.5 -8.6 (-0.13%)

TRAFFIC CAMERAS (TWO AT RANDOM) (TfL)


BBC LONDON NEWS (BBC)
Covid: London's virus death toll exceeds 10,000 Royal Mail delays: Londoners 'frustrated' by missing post Tower of London's 'queen' raven Merlinia missing Drill and rap music on trial

OPENSTREETMAP UPDATES (OSM)
[Add playground access](#) [Add path surfaces](#) [Add whether pedestrian crossing has an island](#) [Add road surface info](#) [Add playground #ukcp #maproulette](#) [Added shelter](#) [Updated speed limits](#) [Added source](#)


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

Attempt for Dashboards - Amsterdam



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

25

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/>)

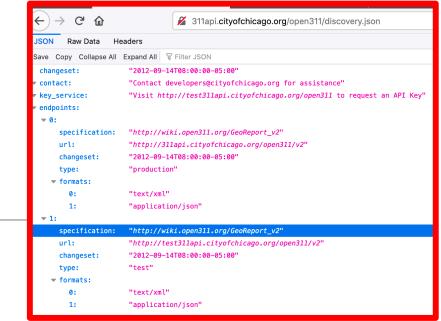
Refers to non-emergency events like graffiti, garbage, down trees, abandoned car, ...

- 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
               "text/xml"
               "application/json"
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
```



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

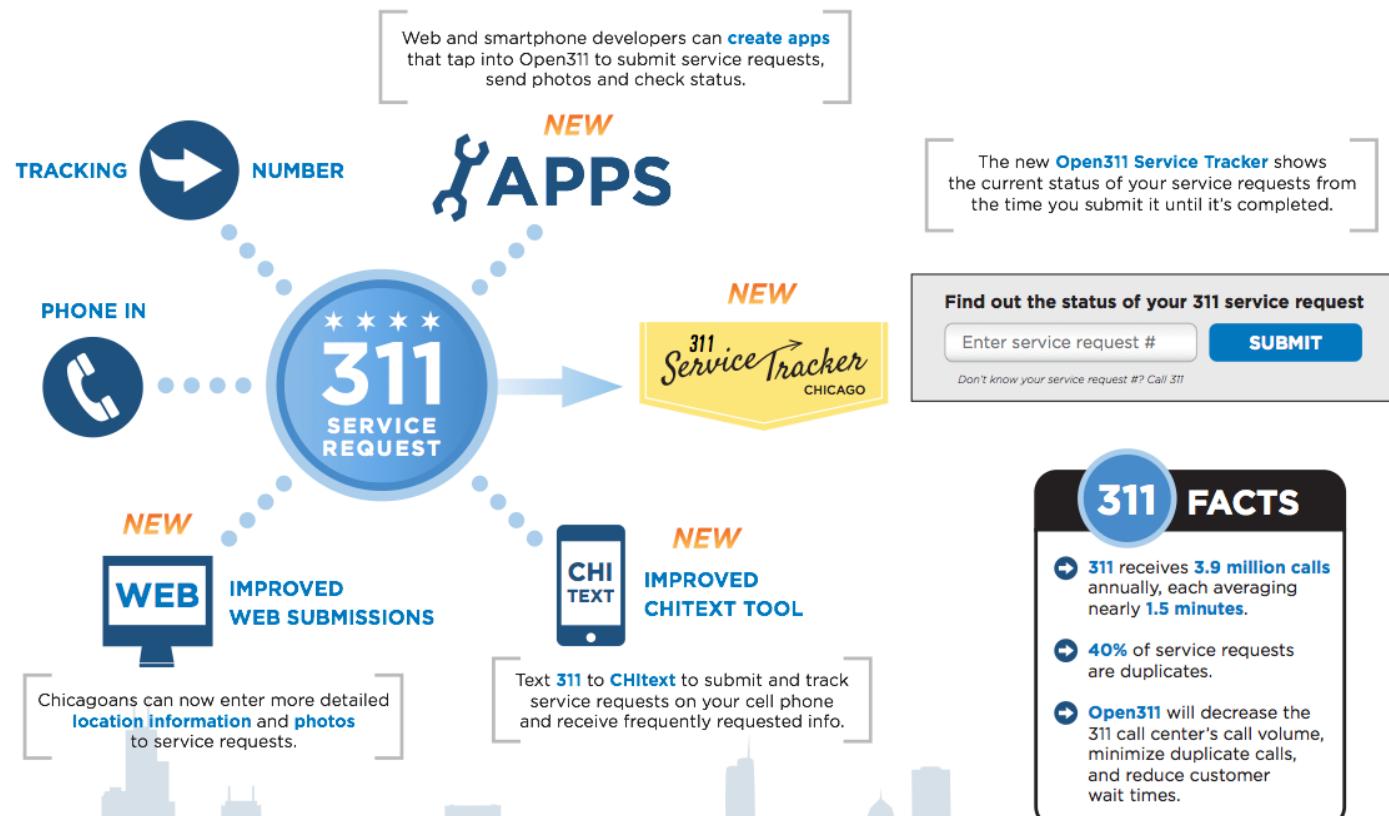
```
{"service_code":"4ffa9cad6018277d4000007b",
 "attributes":
 [{"variable":true,"code":"ISTHELI2",
 "datatype":"singlevaluelist","required":true,"order":1,
 "description":"Is the light located in your alley or the street?",
 "values":[{"key":"ALLEY","name":"Alley"},
 {"key":"STREET","name":"Street"}]},

 {"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:

- Service Request #**: #13-00210540
- Address**: 1502 N Wicker Park Ave
- Created**: February 23, 2013
- Received via**: Other
- Status**: Closed (indicated by a green badge)

The "Activity" section shows the following timeline:

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

# Scaling with Open 311

www.wiki.open311.org/GeoReport\_v2/Servers/ ▾ G Google

Open311 GeoReport v2 Servers

The following is a list of official API endpoints for jurisdictions that support the Open311 GeoReport v2 specification.

Note that the URLs require a method like /services.xml (and jurisdiction\_id if serving multiple jurisdictions). The base URL is not intended to be accessed directly. To ensure the Test and Production endpoint URLs listed here reach a valid resource, they all point to the **GET Services** method as an example

New! You can now check on the status of endpoints at: <http://open311status.herokuapp.com/>

## Jurisdiction Specific Endpoints

 Edit

 Download as YM

[Download as JSON](#)

| Name                | Country | API Discovery | API Key Request | Documentation | Production URL Example   | Test URL Example   | Gov Domain |
|---------------------|---------|---------------|-----------------|---------------|--------------------------|--------------------|------------|
| Toronto, ON         | CDN     |               |                 |               | Production /services.xml | Test /services.xml |            |
| Gießen, Deutschland | DEU     |               |                 |               | Production /services.xml |                    |            |
| Bonn, Deutschland   | DEU     |               |                 |               | Production /services.xml |                    |            |
| Helsinki, Suomi     | FIN     |               |                 |               | Production /services.xml |                    |            |
| Lamía, Elláda       | GRC     |               |                 |               | Production /services.xml |                    |            |
| Baltimore, MD       | USA     |               |                 |               | Production /services.xml | Test /services.xml |            |
| Bloomington, IN     | USA     |               |                 |               | Production /services.xml | Test /services.xml |            |
| Boston, MA          | USA     |               |                 |               | Production /services.xml | Test /services.xml |            |
| Brookline, MA       | USA     |               |                 |               | Production /services.xml |                    |            |
| Chicago, IL         | USA     |               |                 |               | Production /services.xml | Test /services.xml |            |
| Columbus, IN        | USA     |               |                 |               | Production /services.xml |                    |            |

# Exercise 2 – Programmatically Access Data

---

1. See sample code on GitHub:

- <https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/I1-intro/Explore%20OpenData.ipynb>

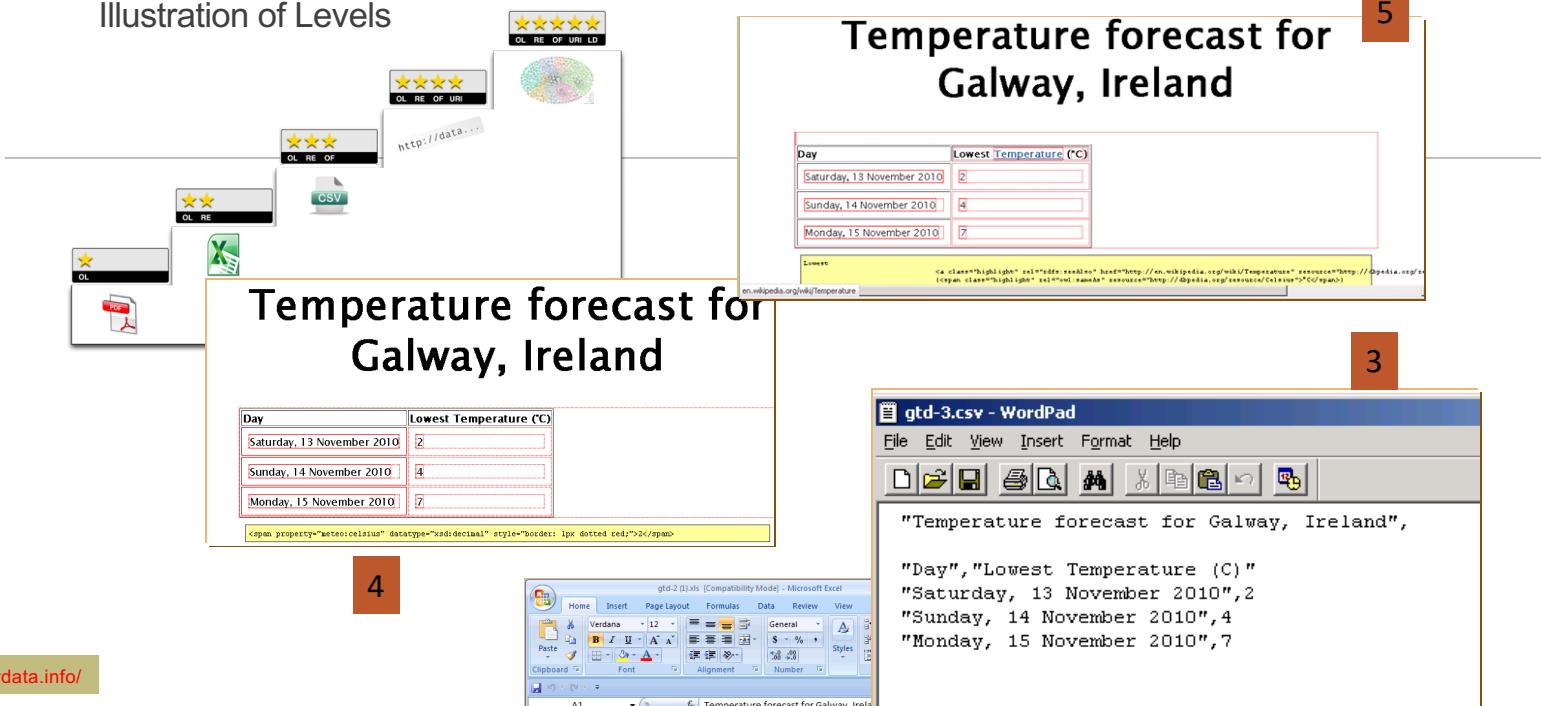
2. Explore APIs of another city of your choice

# Quality of Data

---

## Does Opening Data Make It Reusable? No

Illustration of Levels



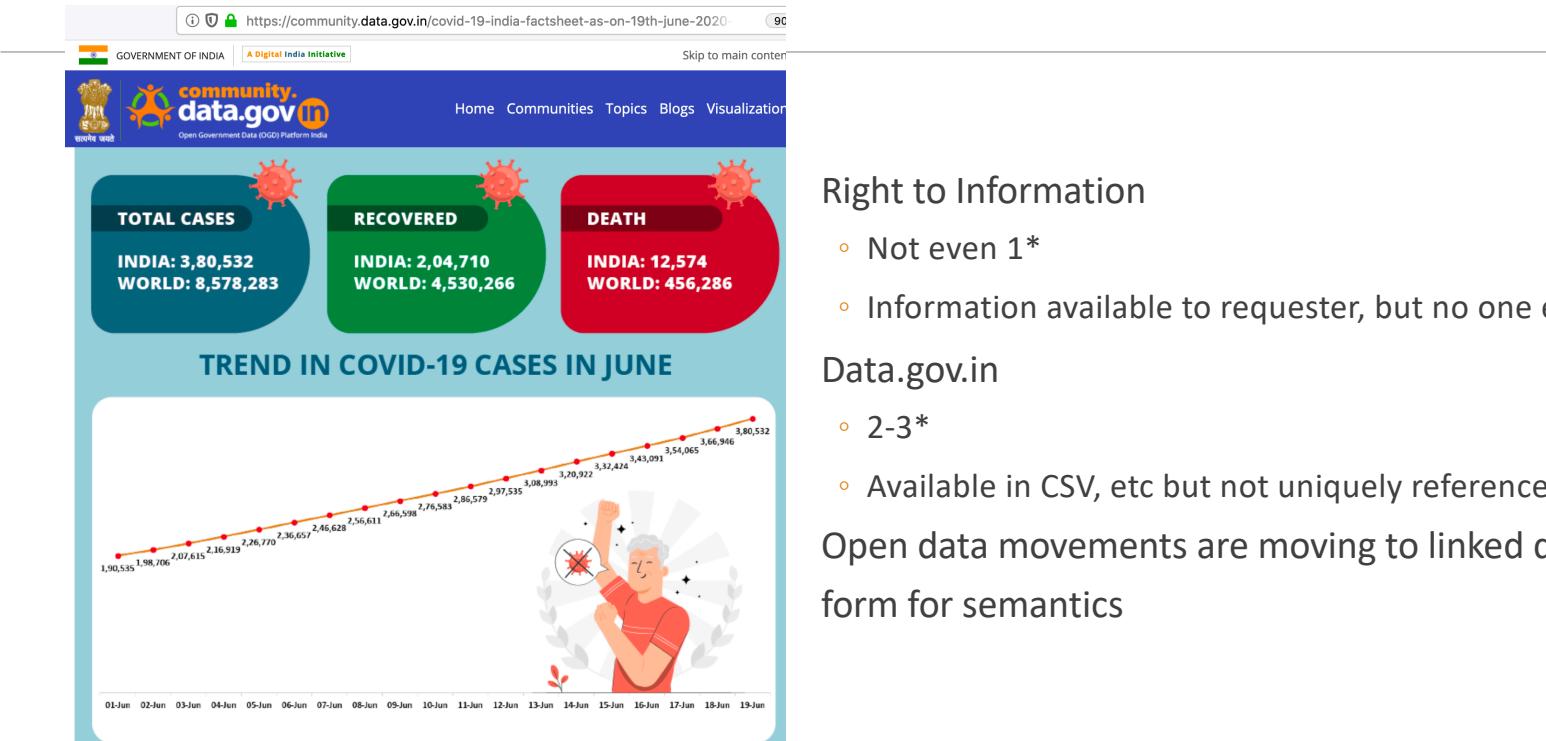
Source: <http://5stardata.info/>

| 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**

IM DATA TO DECISIONS WITH OPEN DATA: A PRACTICAL INTRODUCTION TO AI

# 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

---

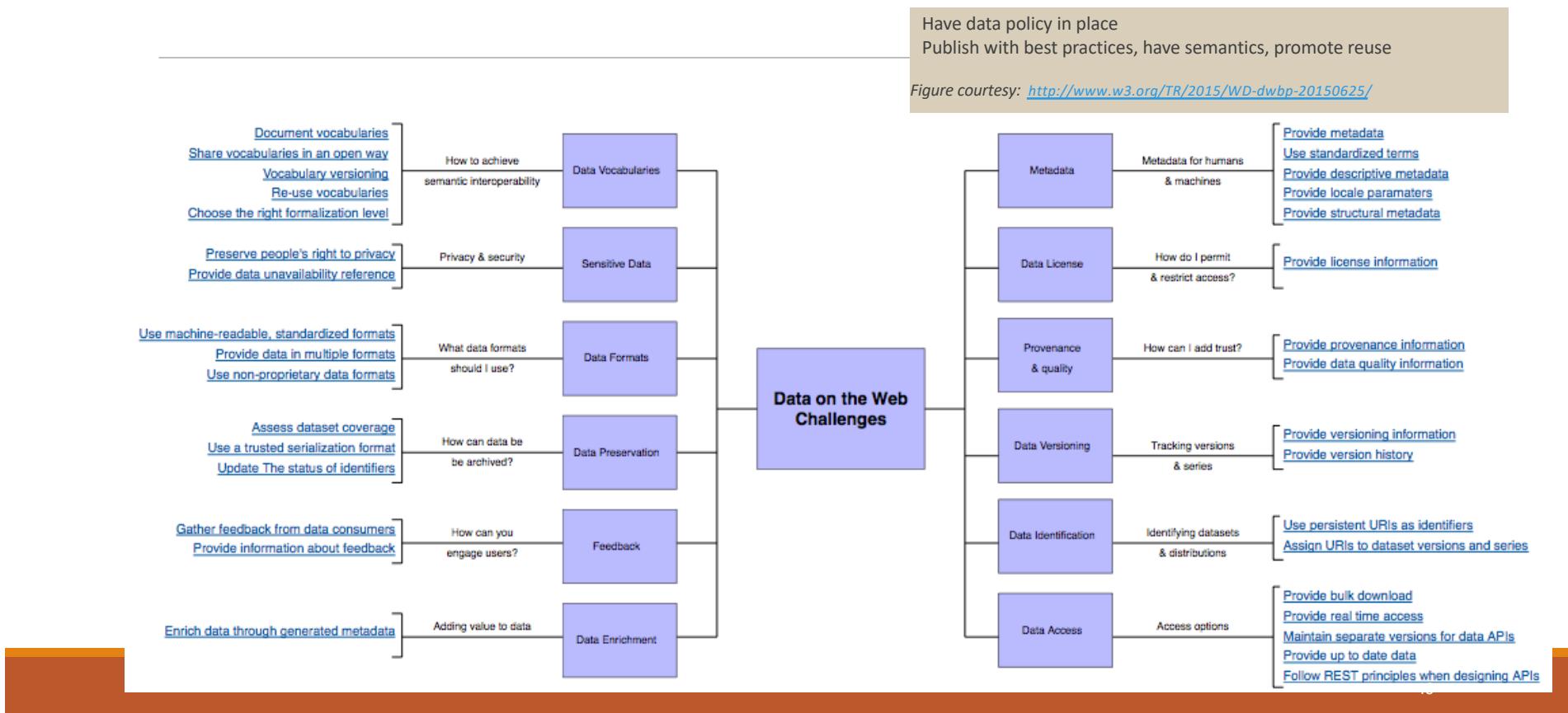
Vocabulary services: <http://vocab.nic.in/index.php>

- 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



# Lecture 2: Concluding Comments

---

- We looked at data of different types: by structure, content type and source
- Open data is useful
- Accessing open data is quite easy, Open 311 helps
- Annotations are important for discovery
- Publishing good data is important for useful AI

# Concluding Segment

---

# About Next Lecture – Lecture 3

---

# Lecture 3: A Systematic Approach

---

- Real-World Problems
- A Systematic Approach
  - Value of decision: before and after
  - Data-needed
  - Method
  - Evaluation
  - Integrating with overall process