



## *CSCE 581: Introduction to Trusted AI*

### Lectures 3 and 4: Introduction to AI, Trust and Real-World Applications

---

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

21<sup>ST</sup> AND 23<sup>RD</sup> JAN 2025

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

**Credits:** Copyrights of all material reused acknowledged

# Organization of Lectures 3, 4

---

- Introduction Section
  - Recap from Week 1 (Lectures 1 and 2)
  - Results of expectation survey
  - Announcements, course scope
- Main Section
  - L3: Trusted Decisions, Data
  - L4: Data Prep, Knowledge Graph
- Concluding Section
  - About next week – Lectures 5, 6
  - Ask me anything

**Credit:** Dilbert

# Introduction Section

---

# Recap from Week 1

---

## Week 1: Introduction to Trusted AI

- Lecture 1: AI and Trust
  - AI as intelligent agent
  - Trust problem
  - Flexibility to adjust between AI methods and handling trust based on students' expectations
    - CSCE 581: AI/ ML topics and with a focus on fairness, explanation, Data privacy, reliability
    - CSCE 580: Classical AI topics and a focus on implementation
- Lecture 2: Case studies
  - Data analysis for traffic improvement (South Carolina), Trust –  
<https://ai4society.github.io/projects/traffic-page/index.html>
  - Recommendations and Trust [Fairness and Teaming Recommendation] –  
[https://ai4society.github.io/projects/group\\_rec/index.html](https://ai4society.github.io/projects/group_rec/index.html)

# Expectations Survey

- 60% response rate
- AI background (1-5 scale): 2 (50%), 3 (33%)

## One thing want to learn

- I'd like to learn how researchers are working towards mitigating hallucinations in LLMs
- How can the recipients (readers) of AI-generated information think that the information they see is trustworthy
- How reliable is AI and how far will it go?
- LLMs
- I aim to explore Trustworthy AI from an interdisciplinary conceptual perspective, focusing on its applications across domains such as healthcare. My interests include human-in-the-loop use cases emphasizing explainability for both the developers(us) and users, alongside examining our work's social contexts and implications.
- I want to learn the ways in which we can test the accuracy of our models in the scope of computer vision

## Trust topics to cover

- News platforms and posts in social media platforms
- job replacements? and privacy concerns ?

## AI problems to cover

- Unsure (All sound fun to learn)
- I think using AI to simulate the development of a small society is an interesting topic. I've read that some game developers are trying to apply AI to non-player characters in order to achieve more realistic experiences and self-evolving stories. I also heard that a team (from Stanford if I remembered correctly) built a computer-simulated community, using GPT-3 to simulate the behavior and conversations of individual residents to simulate a self-evolving mini-society.
- ethical concerns?
- My interest primary is Machine Learning
- I am primarily interested in social computing and healthcare applications.
- computer vision! I work for Boeing doing Augmented Reality research and would love to learn more details about the development of Computer Vision.

# Expectations Survey and Action

- 60% response rate
- AI background (1-5 scale): 2 (50%), 3 (33%) => AI topics will be explained before trust issues covered

## One thing want to learn

- I'd like to learn how researchers are working towards mitigating hallucinations in LLMs
- How can the recipients (readers) of AI-generated information think that the information they see is trustworthy
- How reliable is AI and how far will it go?
- LLMs
- I aim to explore Trustworthy AI from an interdisciplinary conceptual perspective, focusing on its applications across domains such as healthcare. My interests include human-in-the-loop use cases emphasizing explainability for both the developers(us) and users, alongside examining our work's social contexts and implications.
- I want to learn the ways in which we can test the accuracy of our models in the scope of computer vision

Legends: a) LLMs covered in the context of trust, b) Computer vision only touched in passing, c) All in green will be covered

## Trust topics to cover

- News platforms and posts in social media platforms
- job replacements? and privacy concerns ?

## AI problems to cover

- Unsure (All sound fun to learn)
- I think using AI to simulate the development of a small society is an interesting topic. I've read that some game developers are trying to apply AI to non-player characters in order to achieve more realistic experiences and self-evolving stories. I also heard that a team (from Stanford if I remembered correctly) built a computer-simulated community, using GPT-3 to simulate the behavior and conversations of individual residents to simulate a self-evolving mini-society.
- ethical concerns?
- My interest primary is Machine Learning
- I am primarily interested in social computing and healthcare applications.
- computer vision! I work for Boeing doing Augmented Reality research and would love to learn more details about the development of Computer Vision.

Legends: All in green will be covered, games and computer vision will be touched upon

# High Level Semester Plan (Adapted, Approximate)

## CSCE 581 –

- Week 1: Introduction
- Week 2: Background: AI - Common Methods
- Week 3: The Trust Problem
- Week 4: Machine Learning (Structured data) - Classification
- Week 5: Machine Learning (Structured data) - Classification – Trust Issues
- Week 6: Machine Learning (Structured data) – Classification – Mitigation Methods
- Week 7: Machine Learning (Structured data) – Classification – Explanation Methods
- Week 8: Machine Learning (Text data, **vision**) – Classification,

### **Large Language Models**

- Week 9: Machine Learning (Text data) - Classification – Trust Issues, LLMs
- Week 10: Machine Learning (Text data) – Classification – Mitigation Methods
- Week 11: Machine Learning (Text data) – Classification – Explanation Methods
- Week 12: Emerging Standards and Laws, **Real world applications**
- Week 13: Project presentations
- Week 14: Project presentations, Conclusion

AI/ ML topics and with a focus on fairness, explanation, Data privacy, reliability

# AI Trust News

---

- Dr Kush Varshney, IBM Research, will give invited lecture on March 18, 2025 (Tuesday) !
  - We are using his book !
- OpenAI and lack of trust
  - OpenAI's o3 and comparison with Theranos scandal (Jan 2025); showing exceptional results on EpochAI's FrontierMath benchmark while having access to much of the test data, and funding the same.
    - Claim of results, <https://www.nature.com/articles/d41586-025-00110-6>
    - Controversy: <https://analyticsindiamag.com/ai-news-updates/openai-just-pulled-a-theranos-with-o3/>,  
<https://content.techgig.com/technology/is-openai-misleading-the-ai-world-the-o3-benchmark-controversy/articleshow/117392266.cms>
  - Earlier, OpenAI and other LLM companies gave up on providing information for elections!,
    - Source: <https://openai.com/index/how-openai-is-approaching-2024-worldwide-elections/>
    - Rozado, D. 2024. The Political Preferences of LLMs. arXiv:2402.01789.

# 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

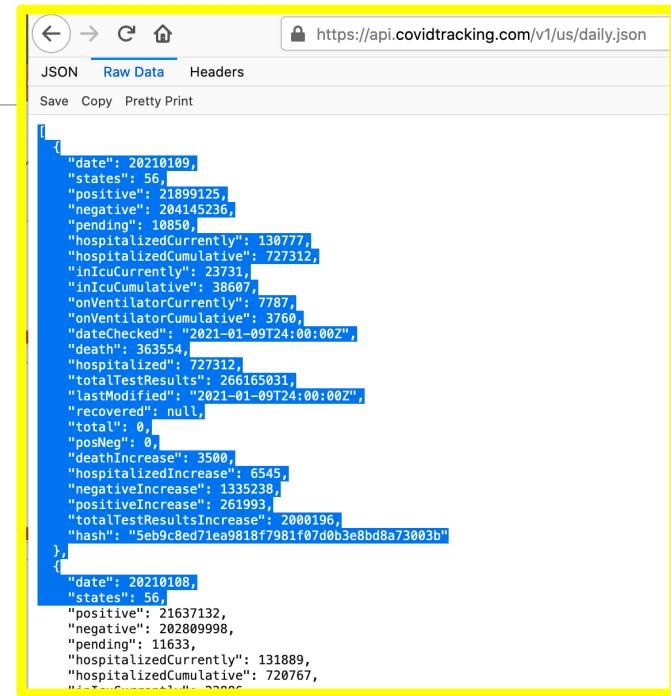
```
country,placename,frequency,start_date,end_date,year,month,week,deaths,expected_deaths,excess_deaths,baseline  
...  
France,,weekly,2020-04-27,2020-05-03,2020,4,18,10498,10357,141,2010-2018 weekly average  
...
```

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



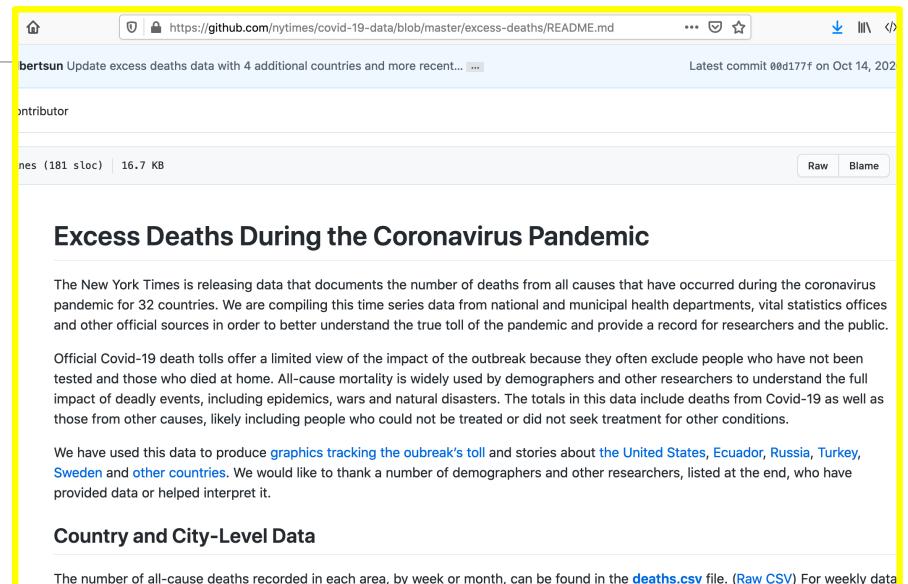
A screenshot of a web browser displaying JSON data from the URL <https://api.covidtracking.com/v1/us/daily.json>. The browser interface includes a back button, forward button, refresh button, and a home icon. Below the address bar, there are tabs for 'JSON', 'Raw Data' (which is currently selected), and 'Headers'. Underneath the tabs are buttons for 'Save', 'Copy', and 'Pretty Print'. The main content area shows two JSON objects. The first object has a date of 2021-01-09, 56 states, and various counts for positive, negative, and pending cases. The second object has a date of 2021-01-10, 56 states, and similar counts. A yellow box highlights the JSON data.

```
[{"date": "2021-01-09", "states": 56, "positive": 21899125, "negative": 204145236, "pending": 10850, "hospitalizedCurrently": 130777, "hospitalizedCumulative": 727312, "inICUCurrently": 23731, "inICUCumulative": 38607, "onVentilatorCurrently": 7787, "onVentilatorCumulative": 3760, "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": "5eb9c8ed71ea9818f7981fb0d0b3e8bd8a73003b"}, {"date": "2021-01-10", "states": 56, "positive": 21637132, "negative": 202809998, "pending": 11633, "hospitalizedCurrently": 131889, "hospitalizedCumulative": 720767, "inICUCurrently": 23731, "inICUCumulative": 38607, "onVentilatorCurrently": 7787, "onVentilatorCumulative": 3760, "dateChecked": "2021-01-10T24:00:00Z", "death": 36354, "hospitalized": 720767, "totalTestResults": 268475031, "lastModified": "2021-01-10T24:00:00Z", "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

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

Details: [https://en.wikipedia.org/wiki/List\\_of\\_file\\_formats](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](https://en.wikipedia.org/wiki/Audio_file_format)

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

# Visual

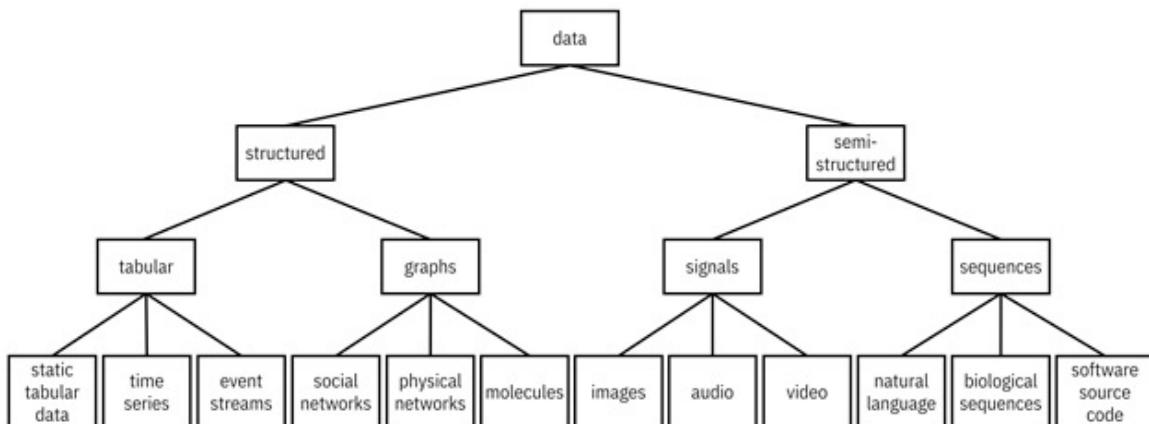
- Media: image, video
- Components: pixel, frame
- Representation
  - Uncompressed – bitmap
  - Compressed lossless - .gif
  - Lossy compression - .jpeg
  - Containers: AVI (.avi) and QuickTime (.mov)
- Programming libraries: PIL, OpenCV

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

# Types of Data

---

- By media: Text, Sound (speech), Visual (image, video), Multi (modal, media)
- By structure: unstructured, semi-structured, structured
- By features: time-series, labeled/ unlabeled, spatio-temporal,



**Image credit:**

<http://www.trustworthymachinelearning.com/trustworthymachinelearning-04.htm>

# 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, REST API, KML).

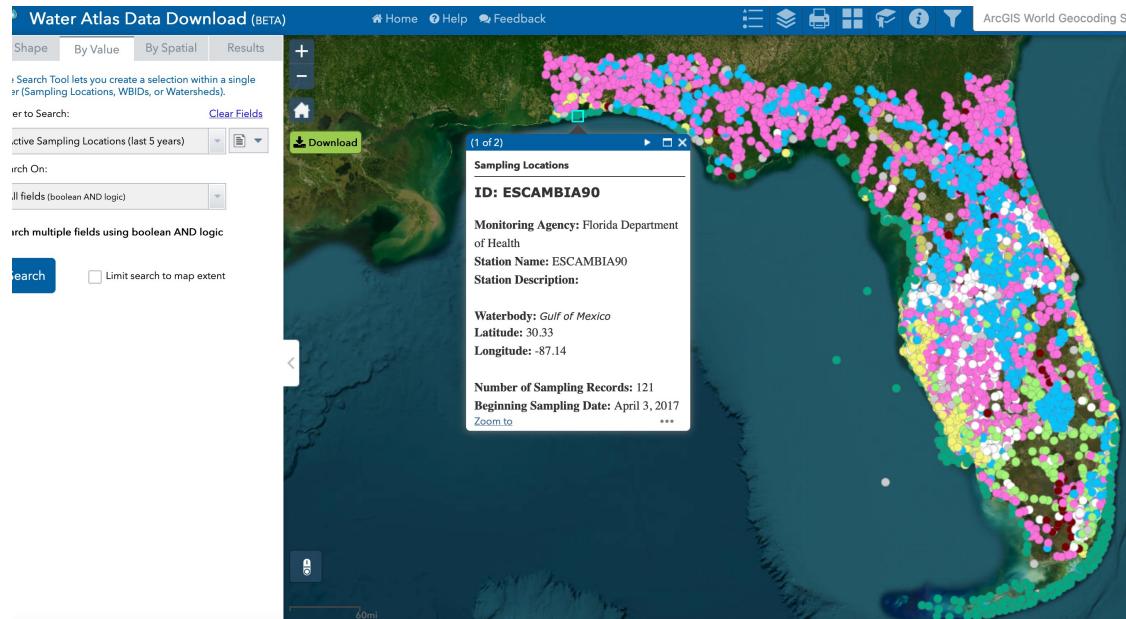
USA

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 are three sections: "ANALYTICS" (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, 2,043 visualizations), "CATALOG" (a lightbulb icon with people around it), and "INDICATOR DASHBOARD" (with cards for Drinking Water And Sanitation, Health, Transport, and Labour And Employment). A footer at the bottom right shows a map of India with state names.

India

# 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>

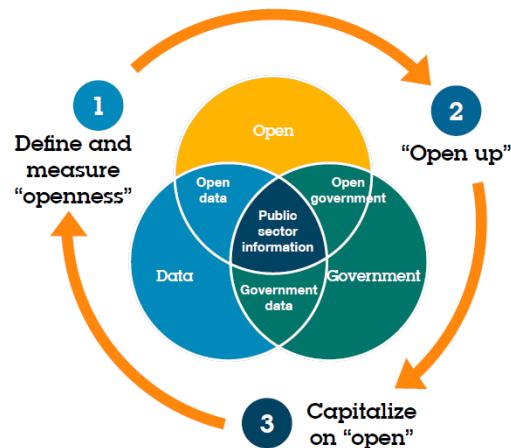


"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

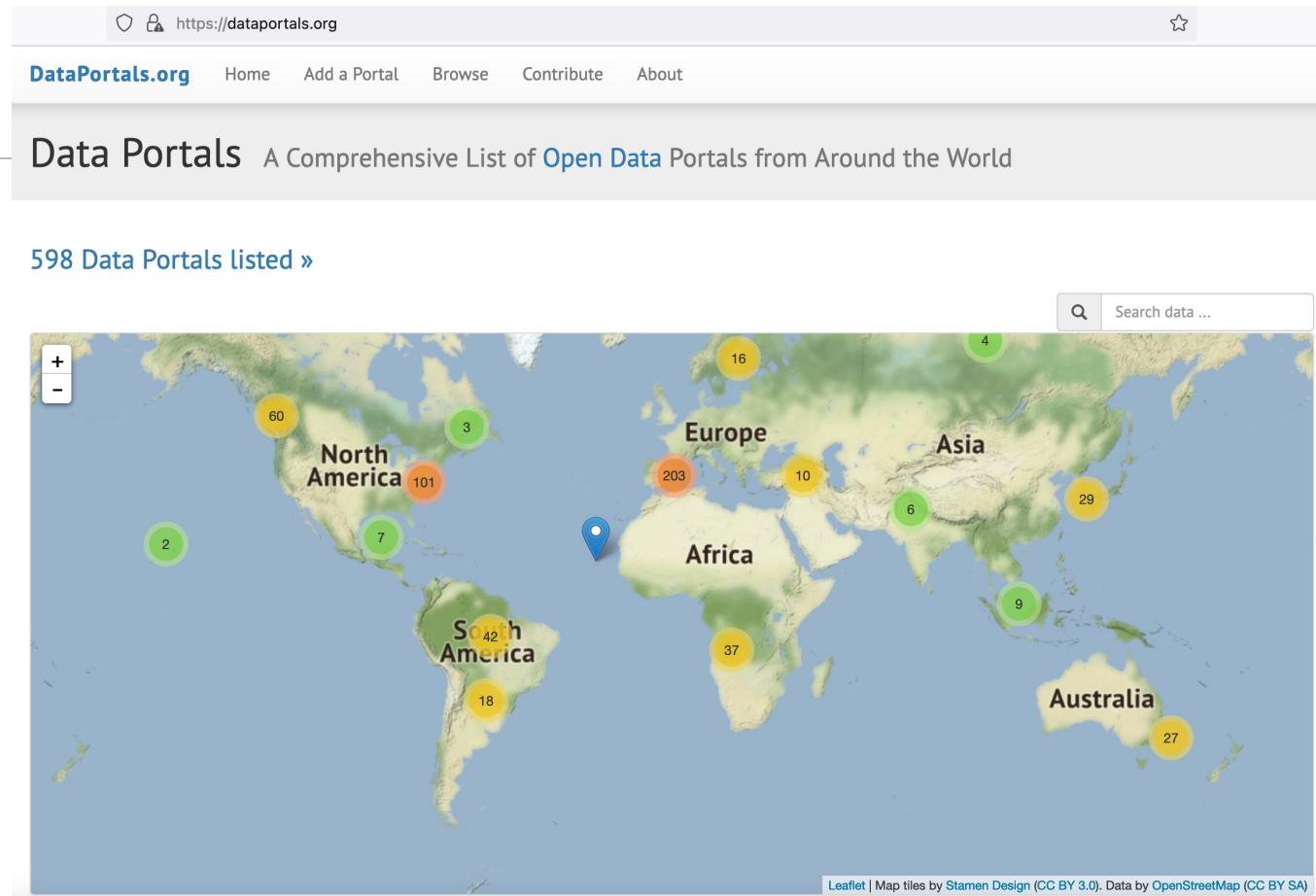
Cartoon critical of big data application,  
by T. Gregorius

[http://upload.wikimedia.org/wikipedia/commons/thumb/b/b3/Big\\_data\\_cartoon\\_t\\_gregorius.jpg/220px-Big\\_data\\_cartoon\\_t\\_gregorius.jpg](http://upload.wikimedia.org/wikipedia/commons/thumb/b/b3/Big_data_cartoon_t_gregorius.jpg/220px-Big_data_cartoon_t_gregorius.jpg)

# ~600 Data Catalogs of Open Data



As on 26 Aug 2024



# 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 PM₂.₅ 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
                   "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](http://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 (0 and 1). Each item has fields like specification, url, changeset, type, and formats. The 'specification' field for both items points to [http://wiki.open311.org/GeoReport\\_v2](http://wiki.open311.org/GeoReport_v2). The 'url' field for item 0 points to <http://311api.cityofchicago.org/open311/v2>, and for item 1 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 item 0 is "production" and for item 1 is "test". The 'formats' field for both items contains "text/xml" and "application/json".

# 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

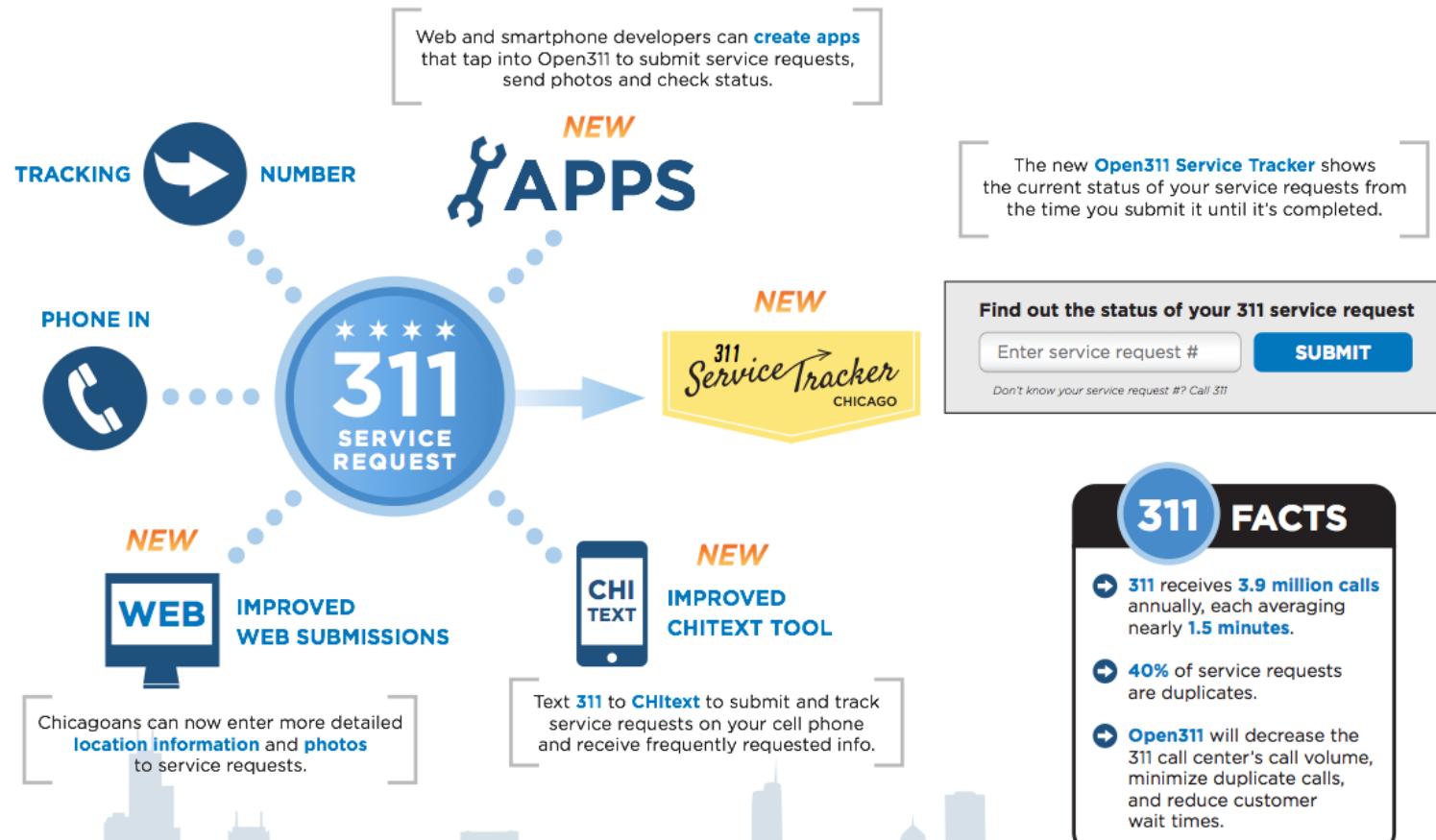
```
{"service_code":"4ffa9cad6018277d4000007b",
 "attributes":
 [{"variable":true,"code":"ISTHELI2",
   "datatype":"singlevaluelist","required":true,"order":1,
   "description":"Is the light located in your alley or the street?",
   "values":[{"key":"ALLEY","name":"Alley"},
             {"key":"STREET","name":"Street"}]},

 {"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:

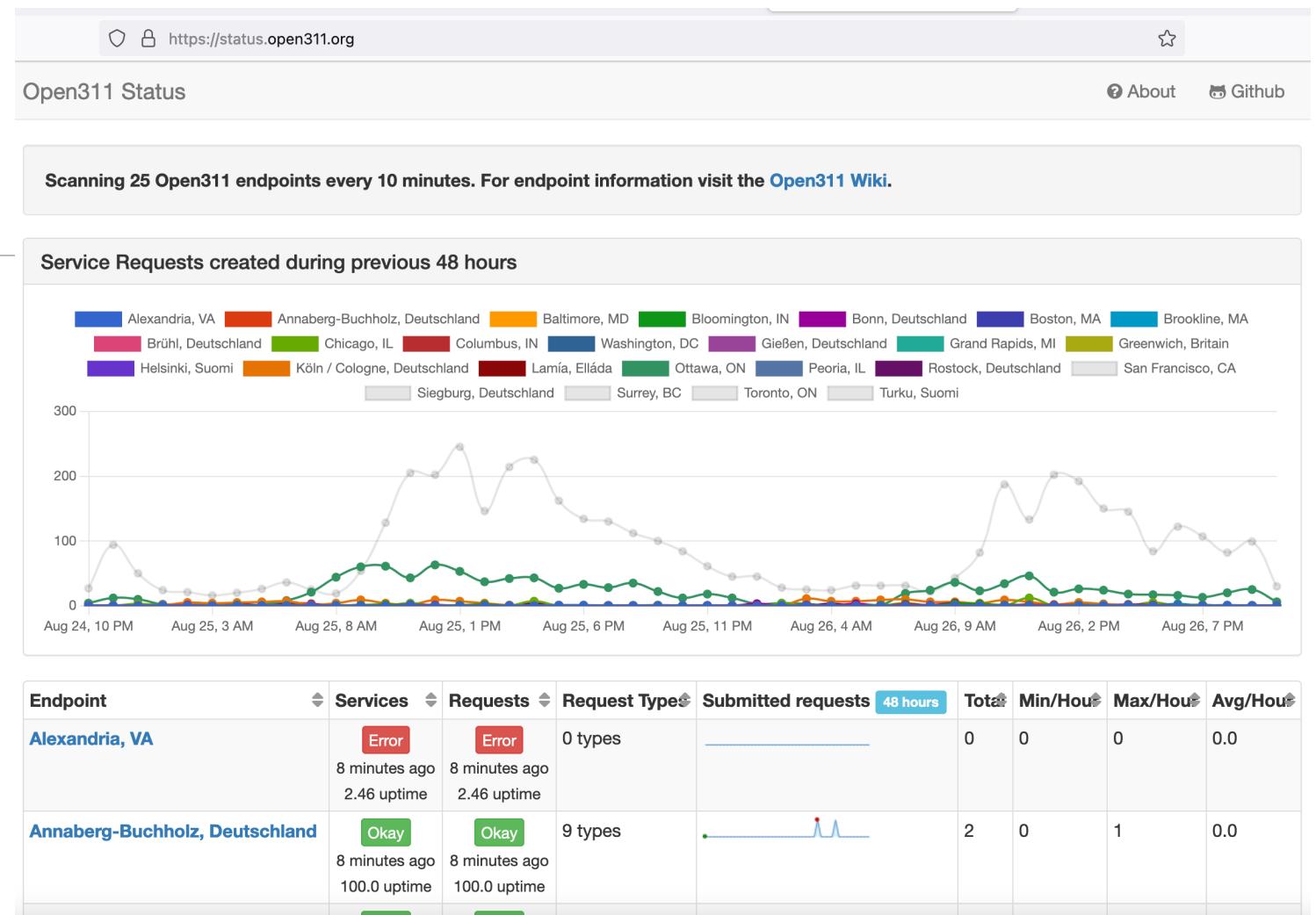
- #13-00210540**
- Address:** 1502 N Wicker Park Ave
- Created:** February 23, 2013
- Received via:** Other

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>

# Text Data

---

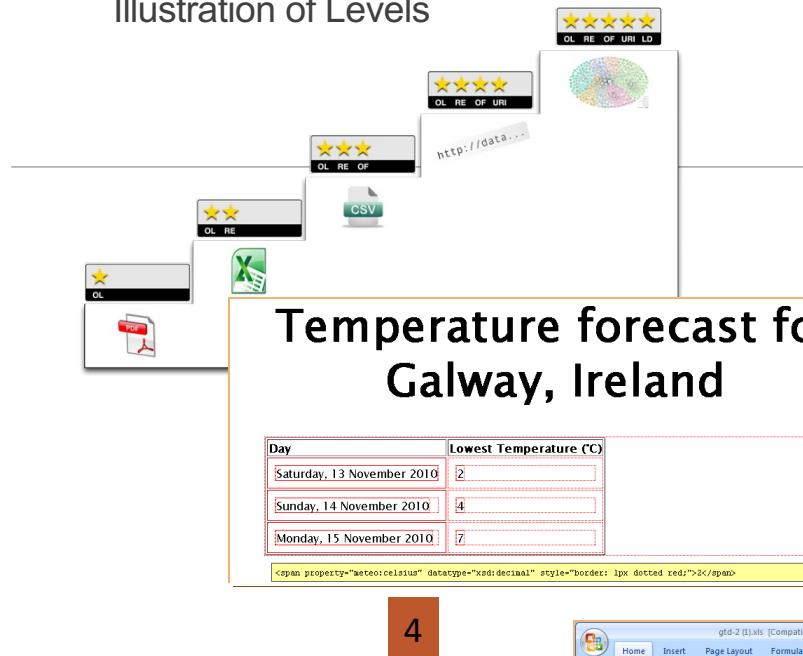
- Text of legislations - LegiScan, <https://legiscan.com/>
- Voter FAQs - <https://github.com/ai4society/election-dataset>
- Compendium of benchmarks and datasets:
  - <https://zilliz.com/learn/popular-datasets-for-natural-language-processing>,
  - UCI dataset – <https://archive.ics.uci.edu/datasets?search=&Types=Text>
  - Kaggle - <https://www.kaggle.com/datasets?search=text>
- NLP task specific -
  - <https://paperswithcode.com/task/named-entity-recognition-ner/>
  - ...

# 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

2

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

1

39

## Temperature forecast for Galway, Ireland

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

en.wikipedia.org/wiki/Temperature

<span class="highlight" style="border: 1px dotted red;">>2</span>

### gtd-3.csv - WordPad

File Edit View Insert Format Help



"Temperature forecast for Galway, Ireland",

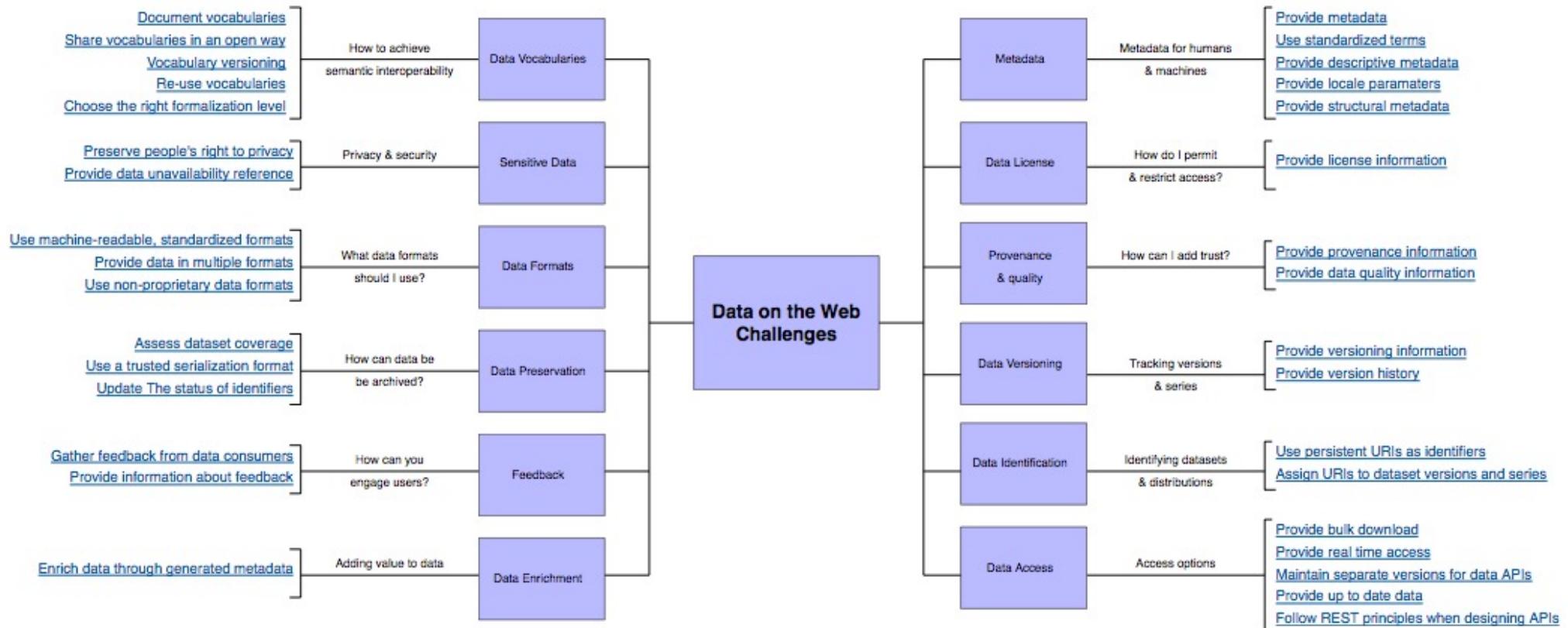
"Day", "Lowest Temperature (C)"  
 "Saturday, 13 November 2010", 2  
 "Sunday, 14 November 2010", 4  
 "Monday, 15 November 2010", 7

2

5

# 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/>



# 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>

# Quality of Data in SC

---

- Data
  - <https://sc.gov/data-and-transparency>
  - <https://rfa.sc.gov/data-research/population-demographics/census-state-data-center/housing-units-in-structure-2015-2019>
  - Comment: Lots of pds and reports: combines/ confounds data with presentation
- Quality of data
  - 1-3 star
  - Not easily amenable for analysis

# Guideline: Human Impact of Data and AI

---

- We study technology (AI) but it works with data
- Data, when from people or about people, can have issues like bias
  - **Example:** data reveals a view which is influenced by data collection practices
  - **Difference:** **World as it is**, world according to data and **world as it should be**
- The course and instructor believes in
  - Not promoting bias of any kind
  - Respecting everyone regardless of background

# Discussion Exercise: Your Resumes

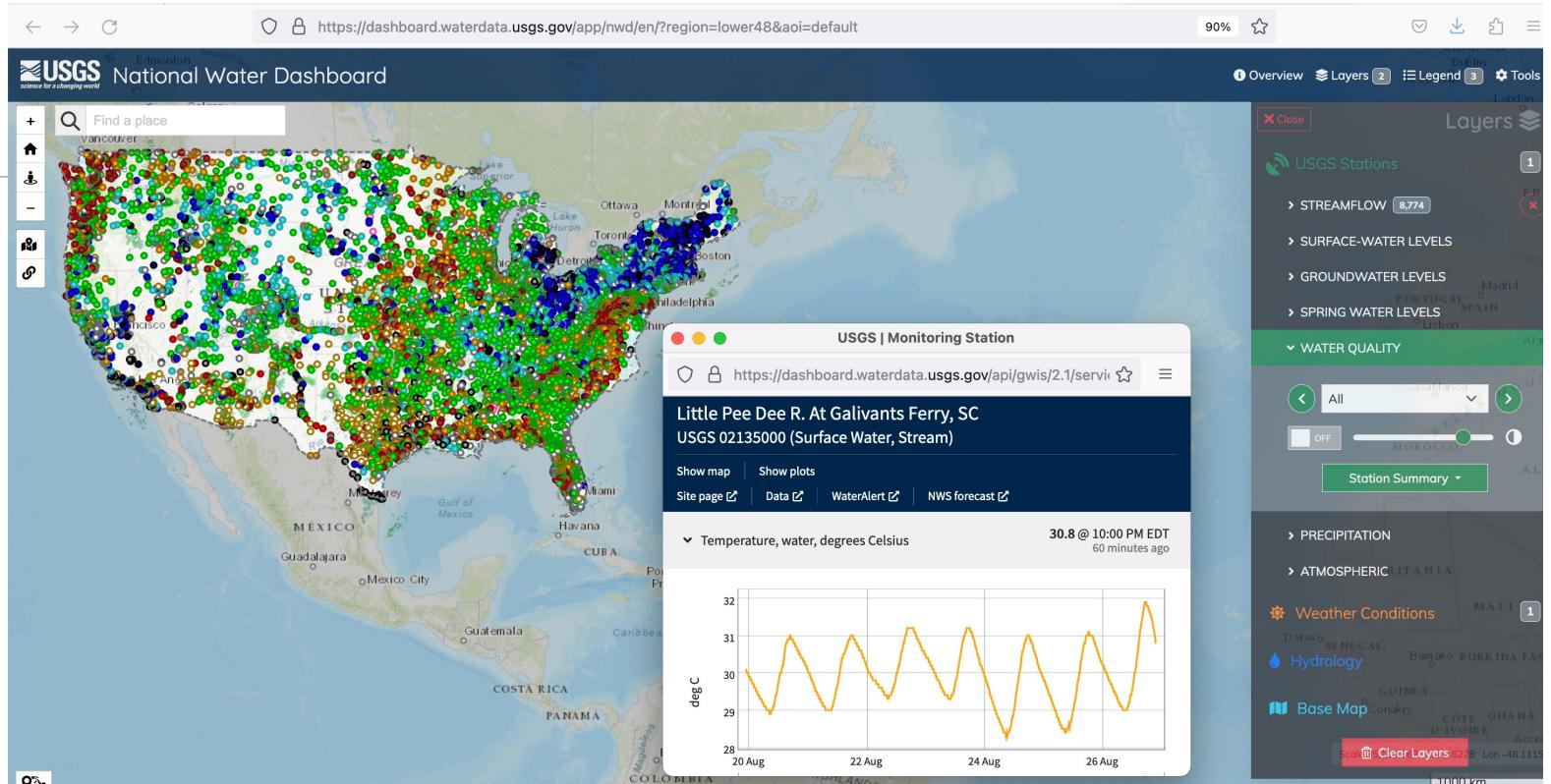
---

- What does a search (Google search) tell about you?
- What does a LLM/ ChatGPT tell about you?
- Task:
  - Put your resume at: <TBD>
- Course task: We will analyze them as part of AI/ data science activity in a later class

---

# Working With Data – Preparing and Organizing Information

# Water Data



<https://dashboard.waterdata.usgs.gov/app/nwd/en/?region=lower48&aoi=default>

Claims data from 13,000 locations online on 26 Aug 2023

# How Do We Start Working With This?

---

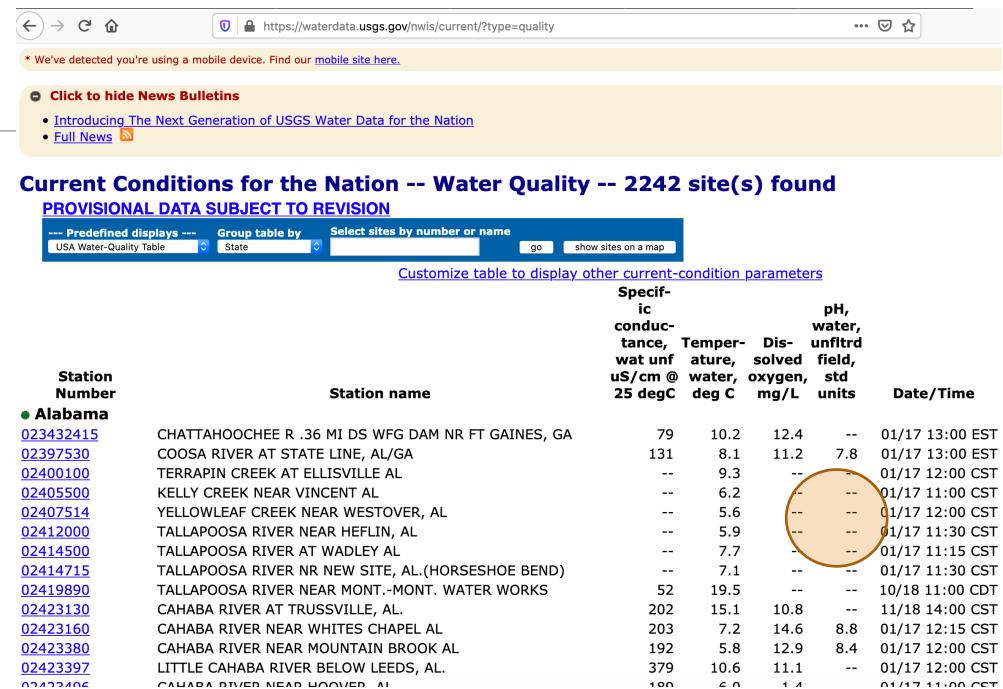
- Access and licensing (Class 3)
- Cleaning, organizing and finding related information (Class 4 – this class)
- Representing formally (in logic) to draw insights (using inferencing) – (Class 4 – this class)

Is this important ? YES !

- Understanding impact of hurricanes
- Planning during regular times – homes, schools, roads; hospital services; electricity, ...
- Economic development

# Common Problem: Missing Value

- Occurrence
  - Missing completely at random
  - Missing at random (a group not wanting to participate)
  - Missing not at random (a group not able to participate)
- What does it mean?
  - The value was not provided
  - The value does not exist or has no practical interpretation
  - The value is being hidden (redaction)
  - Others: The value is not reliable, ...
- How to detect it?
  - By checking for specific values: NA, Not applicable, out-of-range value, 0, -1, "".



The screenshot shows a web browser displaying the USGS Water Data for the Nation website. The URL is https://waterdata.usgs.gov/nwis/current/?type=quality. The page title is "Current Conditions for the Nation -- Water Quality -- 2242 site(s) found". The main content is a table of water quality data for Alabama. The columns include Station Number, Station name, Specific conductance at 25 degC, Temperature at 25 degC, Dissolved oxygen mg/L, pH, and Date/Time. A red circle highlights the "Date/Time" column for the last row of data.

| Station Number | Station name                                       | Specific conductance, 25 degC | Temperature, 25 deg C | Dissolved oxygen, mg/L | pH, water, unfiltered field, std units | Date/Time       |
|----------------|----------------------------------------------------|-------------------------------|-----------------------|------------------------|----------------------------------------|-----------------|
| 023432415      | CHATTahoochee R .36 MI DS WFG DAM NR FT GAINES, GA | 79                            | 10.2                  | 12.4                   | --                                     | 01/17 13:00 EST |
| 02397530       | COOSA RIVER AT STATE LINE, AL/GA                   | 131                           | 8.1                   | 11.2                   | 7.8                                    | 01/17 13:00 EST |
| 02400100       | TERRAPIN CREEK AT ELLISVILLE AL                    | --                            | 9.3                   | --                     | --                                     | 01/17 12:00 CST |
| 02405500       | KELLY CREEK NEAR VINCENT AL                        | --                            | 6.2                   | --                     | --                                     | 01/17 11:00 CST |
| 02407514       | YELLOWLEAF CREEK NEAR WESTOVER, AL                 | --                            | 5.6                   | --                     | --                                     | 01/17 12:00 CST |
| 02412000       | TALLAPOOSA RIVER NEAR HEFLIN, AL                   | --                            | 5.9                   | --                     | --                                     | 01/17 11:30 CST |
| 02414500       | TALLAPOOSA RIVER AT WADLEY AL                      | --                            | 7.7                   | --                     | --                                     | 01/17 11:15 CST |
| 02414715       | TALLAPOOSA RIVER NR NEW SITE, AL.(HORSESHOE BEND)  | --                            | 7.1                   | --                     | --                                     | 01/17 11:30 CST |
| 02419890       | TALLAPOOSA RIVER NEAR MONT.-MONT. WATER WORKS      | 52                            | 19.5                  | --                     | --                                     | 10/18 11:00 CDT |
| 02423130       | CAHABA RIVER AT TRUSSVILLE, AL                     | 202                           | 15.1                  | 10.8                   | --                                     | 11/18 14:00 CST |
| 02423160       | CAHABA RIVER NEAR WHITES CHAPEL AL                 | 203                           | 7.2                   | 14.6                   | 8.8                                    | 01/17 12:15 CST |
| 02423380       | CAHABA RIVER NEAR MOUNTAIN BROOK AL                | 192                           | 5.8                   | 12.9                   | 8.4                                    | 01/17 12:00 CST |
| 02423397       | LITTLE CAHABA RIVER BELOW LEEDS, AL.               | 379                           | 10.6                  | 11.1                   | --                                     | 01/17 12:00 CST |
| 02423406       | CAHABA RIVER NEAR HOOVER, AL                       | 190                           | 5.0                   | 1.4                    | --                                     | 01/17 11:00 CST |

# Missing Value – Handling

---

- Ignoring missing value (Omission)
  - Reduces available data
- Impute new value (Imputation)
  - Mean or median
  - Default value
- Analysis techniques which are robust against missing value
  - Expectation maximization

# Code Examples

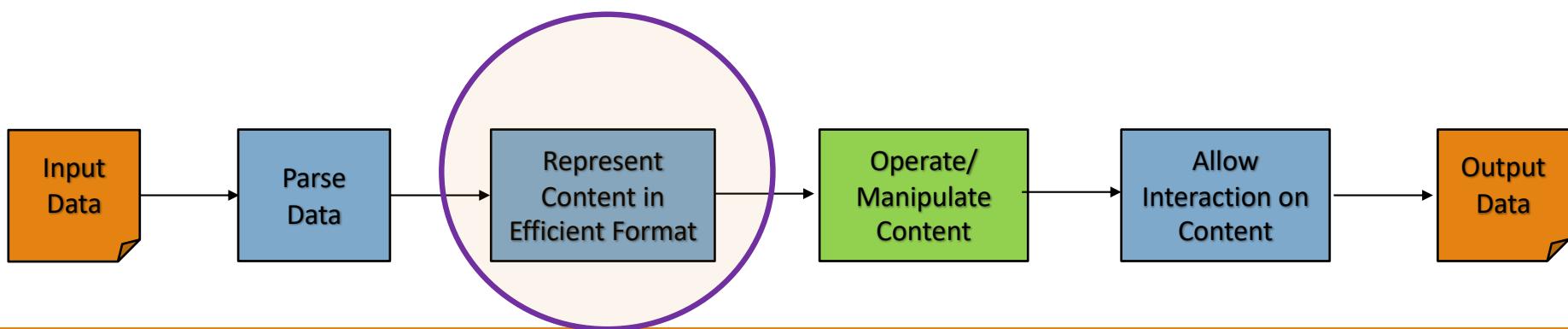
---

<https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l5-dataprep/>

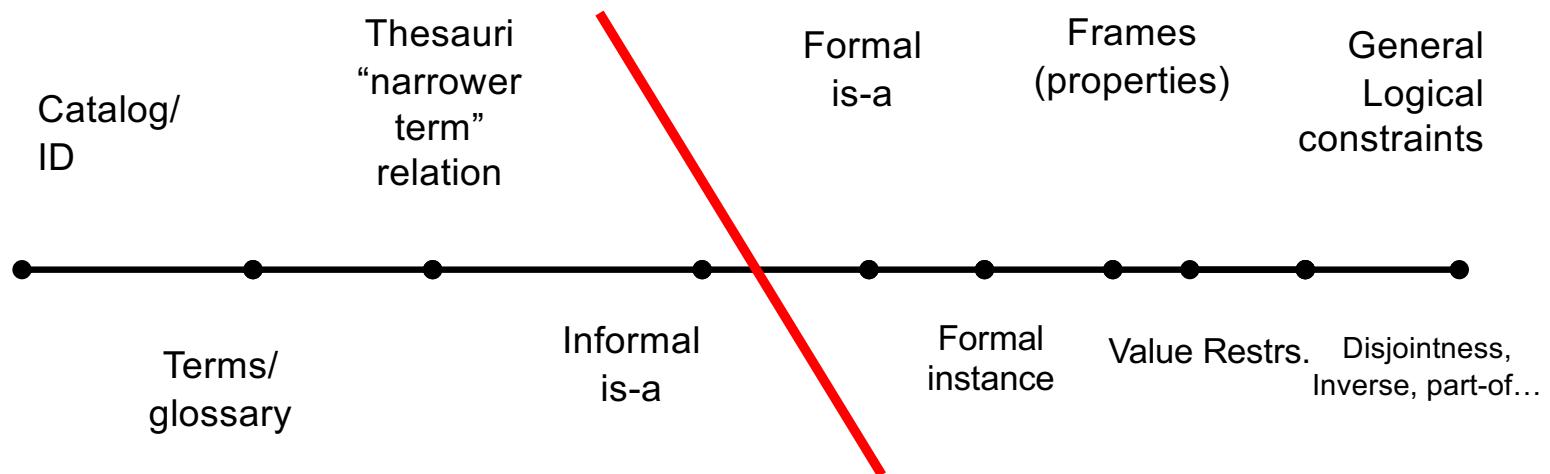
- Basic concepts: **DataPreparation-Numeric.ipynb**
- An illustration: **Clean-RealSample.ipynb**

# Annotation: Knowledge Graphs and Ontology

---



# The Spectrum of Annotation Methods



Ontologies Come of Age McGuinness, 2001, and From AAAI Panel 99 – McGuinness, Welty, Uschold, Gruninger, Lehmann  
Plus basis of Ontologies Come of Age – McGuinness, 2003

# Thesaurus – Authoritative Entities and Relationships

---

Countries: [https://en.wikipedia.org/wiki/List\\_of\\_ISO\\_3166\\_country\\_codes](https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes)

| ISO 3166 <sup>[1]</sup>                                |                                                           |                                      |                             | ISO 3166-1 <sup>[2]</sup>   |                             |                                       | ISO 3166-2 <sup>[3]</sup>     |  |
|--------------------------------------------------------|-----------------------------------------------------------|--------------------------------------|-----------------------------|-----------------------------|-----------------------------|---------------------------------------|-------------------------------|--|
| Country name <sup>[5]</sup>                            | Official state name <sup>[6]</sup>                        | Sovereignty <sup>[6]</sup><br>[7][8] | Alpha-2 code <sup>[5]</sup> | Alpha-3 code <sup>[5]</sup> | Numeric code <sup>[5]</sup> | Subdivision code links <sup>[3]</sup> | Internet ccTLD <sup>[9]</sup> |  |
| Afghanistan                                            | The Islamic Republic of Afghanistan                       | UN member state                      | AF                          | AFG                         | 004                         | ISO 3166-2:AF                         | .af                           |  |
| <b>Akrotiri and Dhekelia – See United Kingdom, The</b> |                                                           |                                      |                             |                             |                             |                                       |                               |  |
| Åland Islands                                          | Åland                                                     | Finland                              | AX                          | ALA                         | 248                         | ISO 3166-2:AX                         | .ax                           |  |
| Albania                                                | The Republic of Albania                                   | UN member state                      | AL                          | ALB                         | 008                         | ISO 3166-2:AL                         | .al                           |  |
| Algeria                                                | The People's Democratic Republic of Algeria               | UN member state                      | DZ                          | DZA                         | 012                         | ISO 3166-2:DZ                         | .dz                           |  |
| American Samoa                                         | The Territory of American Samoa                           | United States                        | AS                          | ASM                         | 016                         | ISO 3166-2:AS                         | .as                           |  |
| Andorra                                                | The Principality of Andorra                               | UN member state                      | AD                          | AND                         | 020                         | ISO 3166-2:AD                         | .ad                           |  |
| Angola                                                 | The Republic of Angola                                    | UN member state                      | AO                          | AGO                         | 024                         | ISO 3166-2:AO                         | .ao                           |  |
| Anguilla                                               | Anguilla                                                  | United Kingdom                       | AI                          | AIA                         | 660                         | ISO 3166-2:AI                         | .ai                           |  |
| Antarctica <sup>[a]</sup>                              | All land and ice shelves south of the 60th parallel south | Antarctic Treaty                     | AQ                          | ATA                         | 010                         | ISO 3166-2:AQ                         | .aq                           |  |
| Antigua and Barbuda                                    | Antigua and Barbuda                                       | UN member state                      | AG                          | ATG                         | 028                         | ISO 3166-2:AG                         | .ag                           |  |
| Argentina                                              | The Argentine Republic                                    | UN member state                      | AR                          | ARG                         | 032                         | ISO 3166-2:AR                         | .ar                           |  |

# (Unique) US Counties Information

In COVID sample code: <https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/I3-health/CovidExploration.ipynb>,

reference made to **FIPS** code

## References:

- [https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143\\_013697](https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143_013697)
- [https://github.com/kjhealy/fips-codes/blob/master/county\\_fips\\_master.csv](https://github.com/kjhealy/fips-codes/blob/master/county_fips_master.csv)

**Question:** how many Richland counties are there in US ?

**Answer:** 14

## County FIPS Codes

| FIPS  | Name     | Stat |
|-------|----------|------|
| 01001 | Autauga  | AL   |
| 01003 | Baldwin  | AL   |
| 01005 | Barbour  | AL   |
| 01007 | Bibb     | AL   |
| 01009 | Blount   | AL   |
| 01011 | Bullock  | AL   |
| 01013 | Butler   | AL   |
| 01015 | Calhoun  | AL   |
| 01017 | Chambers | AL   |
| 01019 | Cherokee | AL   |
| 01021 | Chilton  | AL   |
| 01023 | Choctaw  | AL   |
| 01025 | Clarke   | AL   |
| 01027 | Clay     | AL   |
| 01029 | Cleburne | AL   |
| 01031 | Coffee   | AL   |
| 01033 | Colbert  | AL   |
| 01035 | Conecuh  | AL   |

# Is-a Relationship

---

# List of Countries, States, ... (County), City

---

- United Nations: <https://unece.org/trade/cefact/unlocode-code-list-country-and-territory>
- US Source: <https://github.com/grammakov/USA-cities-and-states>

# Schema.org

---

- Website: <https://schema.org/docs/about.html>
- GitHub: <https://github.com/schemaorg/schemaorg>
- An organization of metadata information for entities found on the web. Mostly backed by web search companies.
- Explore
  - Thing: <https://schema.org/Thing>
  - Product:

# Schema.org

## Example 2

No Markup   Microdata   RDFa   JSON-LD   Structure

*Example notes or example HTML without markup.*

```

Dell UltraSharp 30" LCD Monitor

87 out of 100 based on 24 user ratings

$1250 to $1495 from 8 sellers

Sellers:
<a href="save-a-lot-monitors.com/dell-30.html">
  Save A Lot Monitors - $1250</a>
<a href="jondoe-gadgets.com/dell-30.html">
  Jon Doe's Gadgets - $1350</a>
...

```

No structure

# Schema.org

Example 2

No Markup Microdata RDFa JSON-LD Structure

Example notes or example HTML without markup.

```

Dell UltraSharp 30" LCD Monitor
87 out of 100 based on 24 user ratings
$1250 to $1495 from 8 sellers
Sellers:
<a href="save-a-lot-monitors.com/dell-30.html">
Save A Lot Monitors - $1250</a>
<a href="jondoe-gadgets.com/dell-30.html">
Jon Doe's Gadgets - $1350</a>
...

```

No structure

Structure in JSON-LD format

## Example 2

No Markup Microdata RDFa JSON-LD Structure

Example encoded as JSON-LD in a HTML script tag.

```
<script type="application/ld+json">
{
  "@context": "https://schema.org",
  "@type": "Product",
  "aggregateRating": {
    "@type": "AggregateRating",
    "bestRating": "100",
    "ratingCount": "24",
    "ratingValue": "87"
  },
  "image": "dell-30in-lcd.jpg",
  "name": "Dell UltraSharp 30\" LCD Monitor",
  "offers": {
    "@type": "AggregateOffer",
    "highPrice": "$1495",
    "lowPrice": "$1250",
    "offerCount": "8",
    "offers": [
      {
        "@type": "Offer",
        "url": "save-a-lot-monitors.com/dell-30.html"
      },
      {
        "@type": "Offer",
        "url": "jondoe-gadgets.com/dell-30.html"
      }
    ]
  }
}</script>
```

# Schema.org

Example 2

No Markup Microdata RDFa JSON-LD Structure

*Example notes or example HTML without markup.*

```

Dell UltraSharp 30" LCD Monitor

87 out of 100 based on 24 user ratings
$1250 to $1495 from 8 sellers

Sellers:
<a href="save-a-lot-monitors.com/dell-30.html">
  Save A Lot Monitors - $1250</a>
<a href="jondoe-gadgets.com/dell-30.html">
  Jon Doe's Gadgets - $1350</a>
...

```

## No structure

Example 2

No Markup Microdata RDFa JSON-LD Structure

*Example encoded as JSON-LD in a HTML script tag.*

```
<script type="application/ld+json">
{
  "@context": "https://schema.org",
  "@type": "Product",
  "aggregateRating": {
    "@type": "AggregateRating",
    "bestRating": "100",
    "ratingCount": "24",
    "ratingValue": "87"
  },
  "image": "dell-30in-lcd.jpg",
  "name": "Dell UltraSharp 30\" LCD Monitor",
  "offers": [
    {
      "@type": "AggregateOffer",
      "highPrice": "$1495",
      "lowPrice": "$1250",
      "offerCount": "8",
      "offers": [
        {
          "@type": "Offer",
          "url": "save-a-lot-monitors.com/dell-30.html"
        },
        {
          "@type": "Offer",
          "url": "jondoe-gadgets.com/dell-30.html"
        }
      ]
    }
  ]
}</script>
```

## Structure in JSON-LD format

### Example 2

No Markup Microdata RDFa JSON-LD Structure

*Structured representation of the JSON-LD example.*

|                        |                                                         |
|------------------------|---------------------------------------------------------|
| <b>@type</b>           | Product                                                 |
| <b>name</b>            | Dell UltraSharp 30" LCD Monitor                         |
| <b>offers</b>          |                                                         |
| <b>@type</b>           | AggregateOffer                                          |
| <b>offerCount</b>      | 8                                                       |
| <b>lowPrice</b>        | \$1250                                                  |
| <b>highPrice</b>       | \$1495                                                  |
| <b>offers</b>          |                                                         |
| <b>@type</b>           | Offer                                                   |
| <b>url</b>             | http://example.org/jondoe-gadgets.com/dell-30.html      |
| <b>offers</b>          |                                                         |
| <b>@type</b>           | Offer                                                   |
| <b>url</b>             | http://example.org/save-a-lot-monitors.com/dell-30.html |
| <b>image</b>           | http://example.org/dell-30in-lcd.jpg                    |
| <b>aggregateRating</b> |                                                         |
| <b>@type</b>           | AggregateRating                                         |
| <b>ratingValue</b>     | 87                                                      |
| <b>ratingCount</b>     | 24                                                      |
| <b>bestRating</b>      | 100                                                     |

## Induced Structure

# Schema.org - continued

---

- **Exploration Exercise**

- Services: <https://schema.org/Service>
- Event: <https://schema.org/Event>

- Benefit:

- Easy to incorporate annotations
- Uses popular development tools and technologies (JSON, Microformat)

- Disadvantage

- Cannot perform deep inferencing
- Popular in certain communities

# Formalizing Knowledge in an Ontology

---

## Sources:

Achille Fokoue, Anastasios Kementsietsidis Tutorial

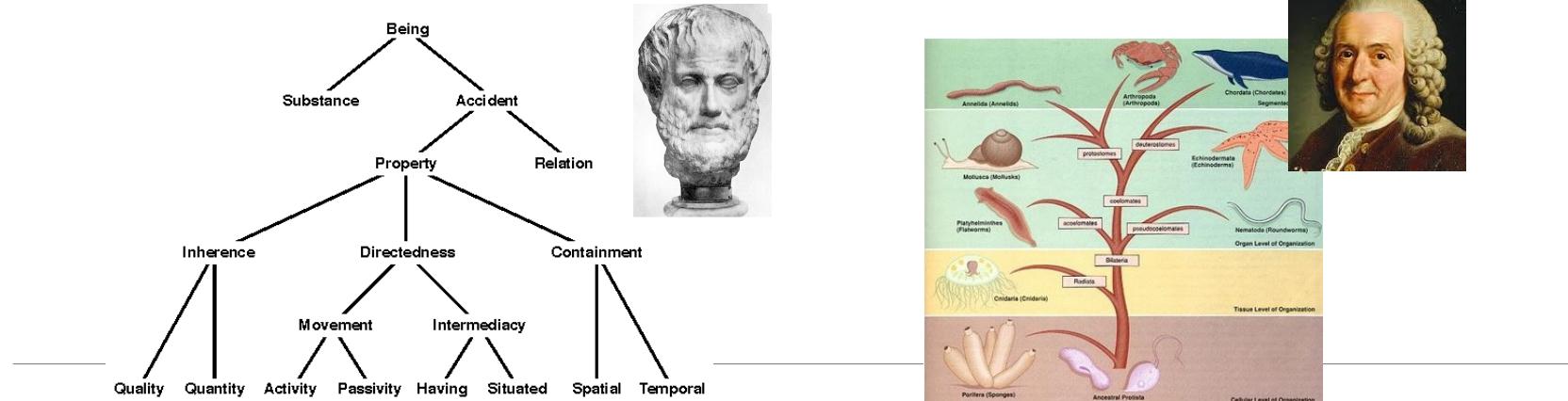
SCRIBE presentation by Rosario Usceda Sosa, Biplav Srivastava, Bob Schloss

- <https://github.com/rschloss/ismp>,
- [https://researcher.watson.ibm.com/researcher/view\\_group.php?id=2505](https://researcher.watson.ibm.com/researcher/view_group.php?id=2505)

## What is an ontology, anyway?

In Computer Science, “An ontology is a formal explicit description of concepts in a domain of discourse (**classes** (sometimes called concepts)), **properties** of each concept describing various features and **attributes** of the concept (slots (sometimes called roles or properties)), and **restrictions** on slots (facets (sometimes called role restrictions)). An ontology together with a set of individual instances of classes constitutes a knowledge base. In reality, there is a fine line where the ontology ends and the knowledge base begins.” [Noy, 2000]

Not to be confused with ontologies (and/or taxonomies) in Philosophy or Life Sciences

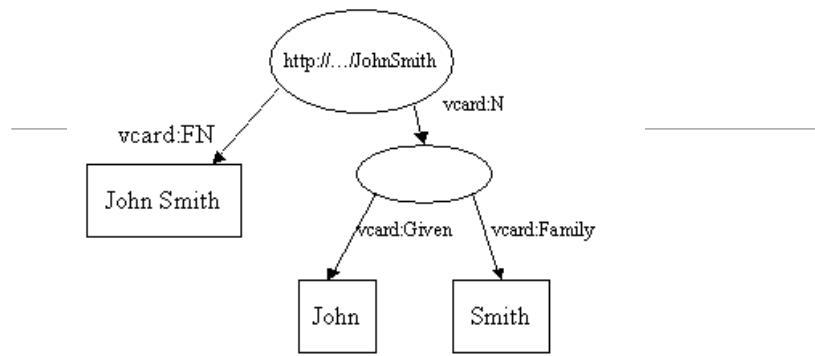


In a Smart City domain, we’re concerned with modeling the *city data* (city activity data, city departments, assets, KPIs), not the city itself (the full set of spatial and temporal relations between people and objects in the city). Ontologies help us to structure and reason about city events, entities and services.

**Ontology = Class + Relations + Constraints**

**Knowledge Base = Ontology + instances + (Standard) Inference and rules**

# RDF / Turtle Example



---- Turtle ----

```
<http://somewhere/JohnSmith>
  <http://www.w3.org/2001/vcard-rdf/3.0#FN>
    "John Smith" ;
  <http://www.w3.org/2001/vcard-rdf/3.0#N>
    [ <http://www.w3.org/2001/vcard-
      rdf/3.0#Family>
        "Smith" ;
      <http://www.w3.org/2001/vcard-
      rdf/3.0#Given>
        "John"
    ] .
```

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
syntax-ns#"
  xmlns:vcard="http://www.w3.org/2001/vcard-
rdf/3.0#" >
  <rdf:Description rdf:nodeID="A0">
    <vcard:Given>John</vcard:Given>
    <vcard:Family>Smith</vcard:Family>
  </rdf:Description>
  <rdf:Description
    rdf:about="http://somewhere/JohnSmith">
    <vcard:FN>John Smith</vcard:FN>
    <vcard:N rdf:nodeID="A0"/>
  </rdf:Description>
</rdf:RDF>
```

# OWL extends RDF...

---

## RDF-schema

- Class, subclass
- Property, subproperty

## + Restrictions

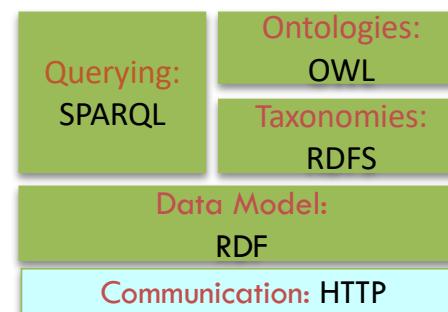
- Range, domain
- Local, global
- Existential
- Cardinality

## + Combinators

- Union, Intersection
- Complement
- Symmetric, transitive

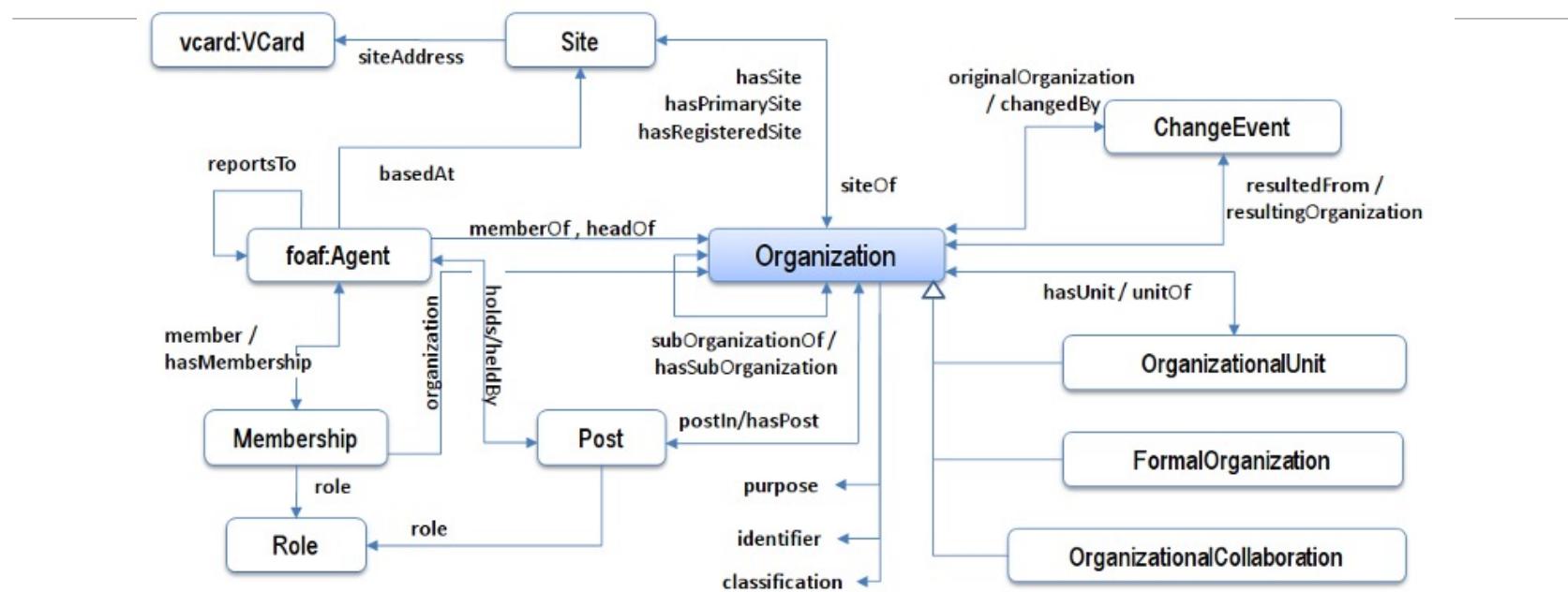
## + Mapping

- Equivalence
- Inverse



**Source:** Achille Fokoue, Anastasios Kementsietsidis Tutorial

# Larger Example: Organization Ontology



Ontology description: <http://www.w3.org/TR/vocab-org/>

Ontology: <http://www.w3.org/ns/org.ttl>

# Larger Ontology

```

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

...
@prefix : <http://www.w3.org/ns/org#> .

# -- Meta data ----

<http://www.w3.org/ns/org#>
  a owl:Ontology;
  owl:versionInfo "0.7";
  rdfs:label "Core organization ontology"@en;
  rdfs:comment "Vocabulary for describing organizational structures, specializable to a broad variety of types of organization."@en;
  dct:created "2010-05-28"^^xsd:date;
  dct:modified "2010-06-09"^^xsd:date;
  dct:modified "2010-10-08"^^xsd:date;
  ...
  rdfs:seeAlso <http://www.w3.org/TR/vocab-org/> ;
  .

# -- Organizational structure ----

org:Organization a owl:Class, rdfs:Class;
  rdfs:subClassOf foaf:Agent;
  owl:equivalentClass foaf:Organization;
  rdfs:label "Organization"@en;
  rdfs:label "Organisation"@fr;
  owl:hasKey (org:identifier);
  rdfs:comment """Represents a collection of people organized together into a community or other social, commercial or political structure. ... Alternative names: _Collective_ _Body_ _Org_ _Group """@en;
  rdfs:comment """Représente un groupe de personnes organisées en communauté où tout autre forme de structure sociale, commerciale ou politique. ... code provenant d'une liste de code."""@fr;
  rdfs:isDefinedBy <http://www.w3.org/ns/org> ;
  .

```

<http://www.w3.org/ns/org.ttl>

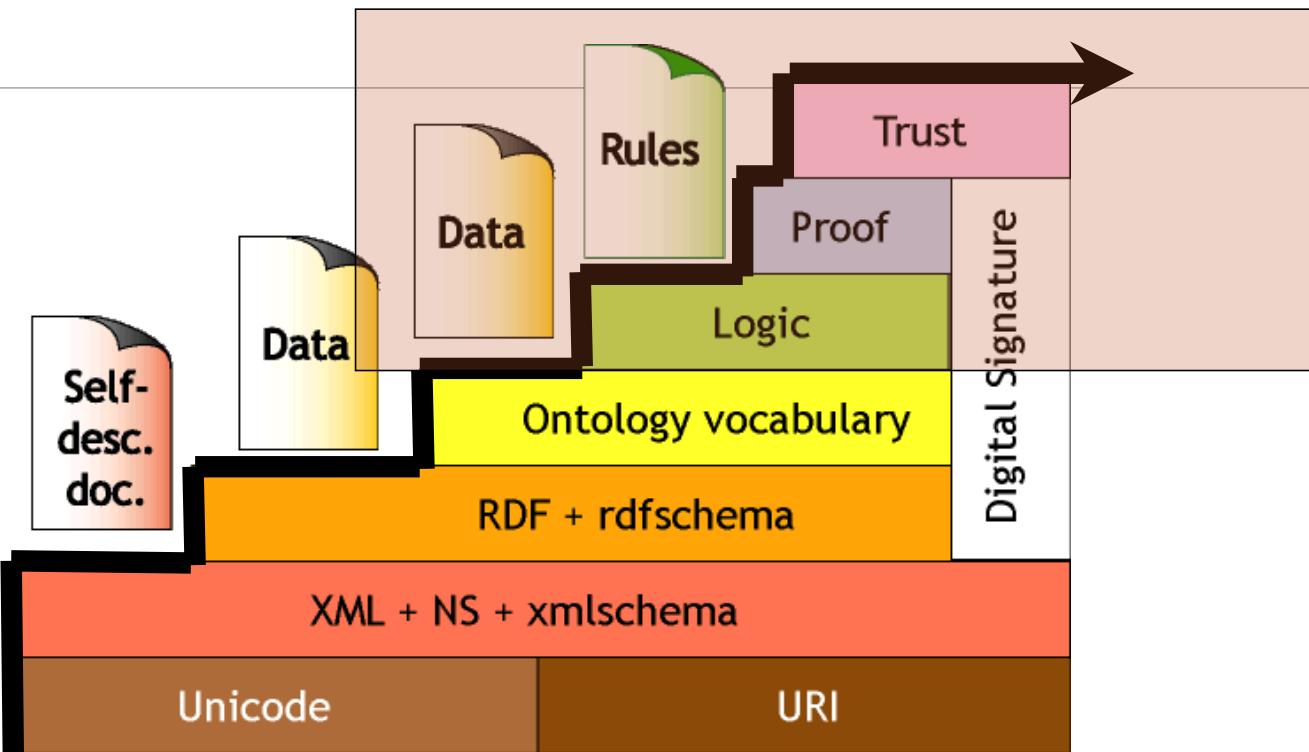
```

- <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#" xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:org="http://www.w3.org/ns/org#" xmlns:gr="http://purl.org/goodrelations/v1#"
  xmlns:owl="http://www.w3.org/2002/07/owl#" xmlns:dct="http://purl.org/dc/terms/"
  xmlns:prov="http://www.w3.org/ns/prov#" xmlns:owlTime="http://www.w3.org/2006/time#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#" xmlns:vcard="http://www.w3.org/2006/vcard/ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
# -- Meta data ----

+ <owl:Ontology rdf:about="http://www.w3.org/ns/org#">
+ <rdfs:Class rdf:about="http://www.w3.org/ns/org#Organization">
- <rdfs:Class rdf:about="http://www.w3.org/ns/org#Role">
  <rdfs:label xml:lang="fr">Rôle</rdfs:label>
- <owl:disjointWith>
  <owl:Class rdf:about="http://www.w3.org/ns/org#ChangeEvent" />
  <owl:disjointWith>
    <rdfs:subClassOf rdf:resource="http://www.w3.org/2004/02/skos/core#Concept" />
- <owl:disjointWith>
  <owl:Class rdf:about="http://www.w3.org/ns/org#Site" />
  <owl:disjointWith>
    <rdfs:comment xml:lang="fr">Indique le rôle qu'une Personne ou un autre Agent peut avoir dans une Organisation. Les instances de cette classe décrivent le rôle dans l'absolu; pour indiquer une personne ayant ce rôle spécifique dans une Organisation, utilisez une instance de `org:Membership`. Il est courant que les rôles soient organisés dans une sorte de taxonomie, ce qui peut être représenté avec SKOS. Les propriétés de libellés standards de SKOS devraient être utilisées pour libeller le Rôle. D'autres propriétés additionnelles pour ce rôle, comme une fourchette de Salaire peuvent être ajoutées par une extension de ce vocabulaire.</rdfs:comment>
- <owl:disjointWith>
  <owl:Class rdf:about="http://www.w3.org/ns/org#Membership" />
  <owl:disjointWith>
    <rdfs:label xml:lang="en">Role</rdfs:label>
    <rdfs:isDefinedBy rdf:resource="http://www.w3.org/ns/org" />
    <rdfs:type rdf:resource="http://www.w3.org/2002/07/owl#Class" />
    <rdfs:comment xml:lang="en">Denotes a role that a Person or other Agent can take in an organization. Instances of this class describe the abstract role; to denote a specific instance of a person playing that role in a specific organization use an instance of `org:Membership`. It is common for roles to be
  
```

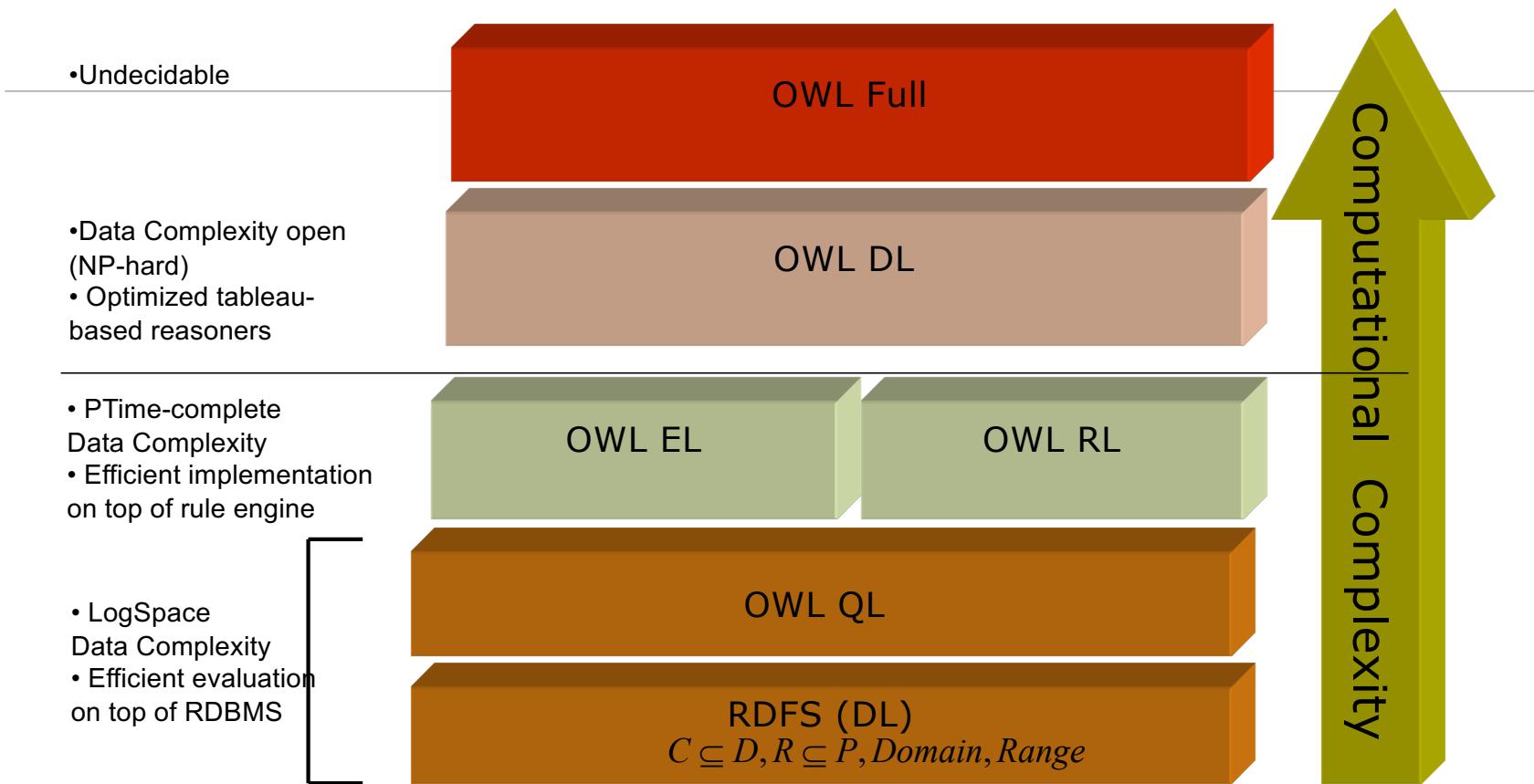
<http://www.w3.org/ns/org>

# Moving to the future of the web



Semantic Web LayerCake (Berners-Lee, 99; Swartz-Hendler, 2001)

# Challenge of Reasoning on Ontologies



## What makes a good ontology for data integration?

A *good* ontology is a *useful* ontology, an ontology that *both* humans and systems can process.

### Human Usability

**Communicable.** Naming, natural language support, etc.

**Concise.** A simple way to describe the key entities of the model and yet able to infer many facts

**Consistent.** Naming conventions and modeling patterns

**Authoritative** to domain experts

**Documented**, not just descriptions, but also provenance

**Managed and maintained** by people throughout the model lifecycle.

**Reusable** in similar domains, for similar instances.

- ❑ *Formal representation of knowledge in a particular domain*
- ❑ *Formally defines key concepts and relations in the domain*
- ❑ *Specifies relationships between those key concepts and relations*
- ❑ *Supports automated reasoning about entities in the domain*

### System Usability

**Scalable** so large amounts of data can be parsed, stored and retrieved.

**Efficient** query and inferencing

**Programmable** solutions, both in open and closed data paradigms.

**Open** infrastructure and tools

# Using Ontology

---

- Visually via tools like Protégé - <https://protege.stanford.edu/>
- Programmatically with APIs like
  - Jena (Java) - <https://jena.apache.org/documentation/ontology/>
  - OwlReady2 (Python) - <https://bitbucket.org/jibalamy/owlready2/src/master/>
  - Rdflib (Python) - <https://github.com/RDFLib/OWL-RL>
- A compendium of resources - <https://github.com/totogo/awesome-knowledge-graph>

# Code Illustration

---

On Github:

<https://github.com/biplav-s/course-nl/blob/master/l11-ontology/Exploring%20ontologies.ipynb>

# Knowledge Graph

---

- No clear definition
  - ["Towards a Definition of Knowledge Graphs," by Lisa Eherlinger and Wolfram Wöß, CEURWorkshop Proceedings.](#) 2016, <http://ceur-ws.org/Vol-1695/paper4.pdf>
  - For practical purposes, concepts and their relationships; not constraints
  - Driven by applications in search and information integration
  - See discussion at: <http://accidental-taxonomist.blogspot.com/2019/05/knowledge-graphs-and-ontologies.html>
- But ontology as knowledge graph widely used in industries
  - Industry-Scale Knowledge Graphs: Lessons and Challenges, CACM 2019,  
<https://cacm.acm.org/magazines/2019/8/238342-industry-scale-knowledge-graphs/fulltext>

# KG Usage

|                  | <b>Data model</b>                                                                                                                                   | <b>Size of the graph</b>                                                                                | <b>Development stage</b>                   |
|------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|--------------------------------------------|
| <b>Microsoft</b> | The types of entities, relations, and attributes in the graph are defined in an ontology.                                                           | ~2 billion primary entities, ~55 billion facts                                                          | Actively used in products                  |
| <b>Google</b>    | Strongly typed entities, relations with domain and range inference                                                                                  | 1 billion entities, 70 billion assertions                                                               | Actively used in products                  |
| <b>Facebook</b>  | All of the attributes and relations are structured and strongly typed, and optionally indexed to enable efficient retrieval, search, and traversal. | ~50 million primary entities, ~500 million assertions                                                   | Actively used in products                  |
| <b>eBay</b>      | Entities and relation, well-structured and strongly typed                                                                                           | Expect around 100 million products, >1 billion triples                                                  | Early stages of development and deployment |
| <b>IBM</b>       | Entities and relations with evidence information associated with them.                                                                              | Various sizes. Proven on scales documents >100 million, relationships >5 billion, entities >100 million | Actively used in products and by clients   |

Figure courtesy: Industry-Scale Knowledge Graphs: Lessons and Challenges, CACM 2019

# Concluding Section

---

# Week 2: Concluding Comments

---

- We looked at
  - Trusted decisions
  - Data and characteristics
  - Common ways to prepare data
  - How to organize content for inferencing / reasoning
- Prepares us for understanding trust issues

# About Next Week – Lectures 5, 6

---

# Lecture 5, 6: AI / ML Methods

---

- Common AI methods
- ML landscape
- Supervised ML

|    |             |                                                                                                                  |
|----|-------------|------------------------------------------------------------------------------------------------------------------|
| 1  | Jan 14 (Tu) | Introduction, Trusted AI                                                                                         |
| 2  | Jan 16 (Th) | Case Studies: Data Analysis for AI, Analysis for Trust [Traffic], Recommendations and Trust [Fairness and ULTRA] |
| 3  | Jan 21 (Tu) | Review: Trusted Decisions, Expectations, Course Scope; Data                                                      |
| 4  | Jan 23 (Th) | AI: Data Prep, Knowledge Graph                                                                                   |
| 5  | Jan 28 (Tu) | Common AI methods: ML Landscape                                                                                  |
| 6  | Jan 30 (Th) | AI - Structured: Analysis – Supervised ML                                                                        |
| 7  | Feb 4 (Tu)  | AI - Structured: Analysis – Supervised ML                                                                        |
| 8  | Feb 6 (Th)  | AI - Structured: Analysis – Supervised ML – Trust Issues                                                         |
| 9  | Feb 11 (Tu) | AI - Structured: Analysis – Supervised ML – Trust Issues                                                         |
| 10 | Feb 18 (Th) | AI - Structured: Analysis – Supervised ML – Mitigation Methods                                                   |
| 11 | Feb 18 (Tu) | AI - Supervised ML: Explanation Tools                                                                            |