



CSCE 581: Introduction to Trusted AI

Lectures 5 and 6: AI, ML and Supervised ML

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

28TH AND 30TH JAN 2025

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

Credits: Copyrights of all material reused acknowledged

Organization of Lectures 5, 6

- Introduction Section
 - Recap from Week 2 (Lectures 3 and 4)
- Main Section
 - L5: Complete discussion on open data and knowledge graph
 - L6: Common ML methods
 - L6: Supervised ML
 - Project Discussion
- Concluding Section
 - About next week – Lectures 7, 8
 - Ask me anything

Introduction Section

Recap from Week 2 (Lectures 3, 4)

Week 1: Introduction to Trusted AI

- Week 2
 - Expectations survey
 - Trusted decisions, data
 - Paper reading (Joy Buolamwini, Timnit Gebru. [Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification](#). In Sorelle A. Friedler, Christo Wilson, editors, Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA. Volume 81 of Proceedings of Machine Learning Research, pages 77-91, PMLR, 2018.)

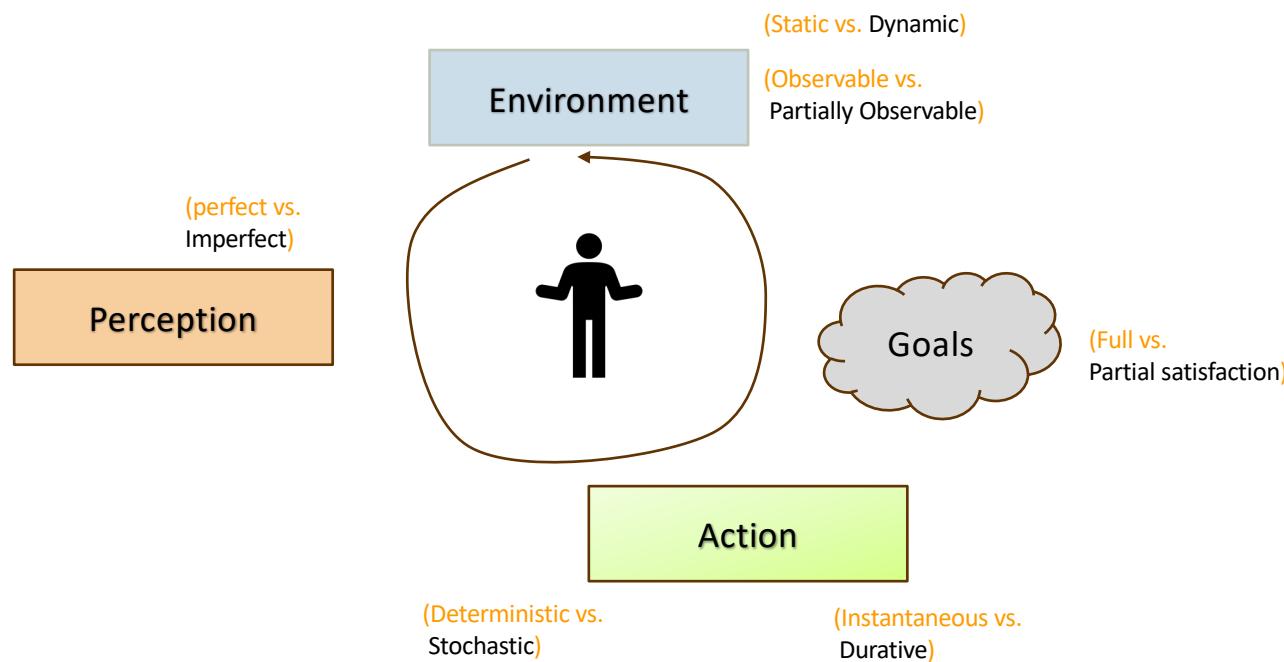
AI News

- DeepSeek R1
- Technical interest, business interest

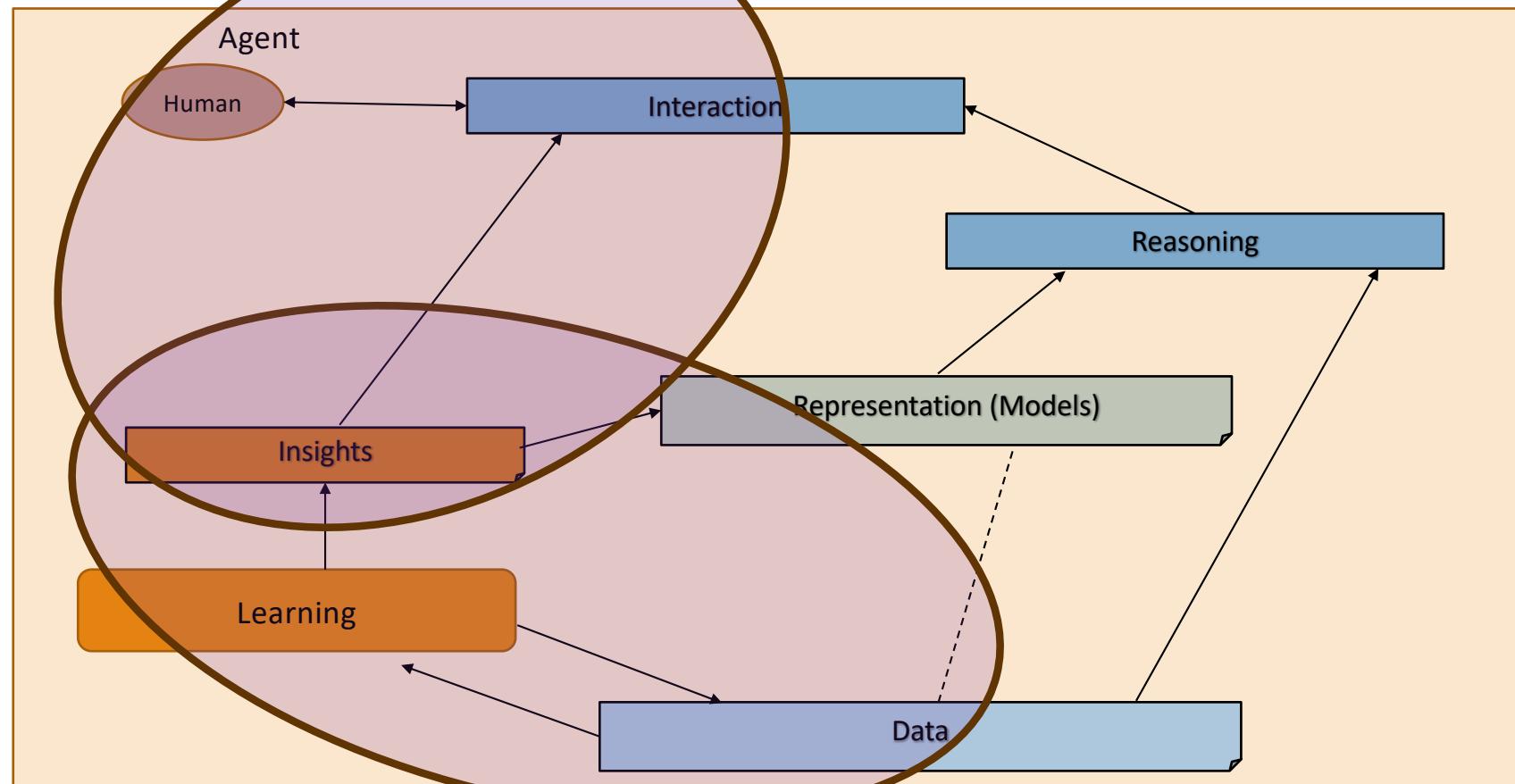
The screenshot shows a comparison chart titled "THE CHEAT SHEET" for AI models. The chart compares GPT-4, DeepSeek-R1, Claude 3, and Gemini across various metrics.

FEATURE	GPT-4	DeepSeek-R1	Claude 3	Gemini
QUALITY SCORE	86.4% MMLU	90.8% MMLU	High (Sonnet model)	Top performer
CONTEXT WINDOW	8,192 tokens	128K tokens	Not specified	2M tokens
INPUT COST (PER 1M TOKENS)	\$90.00	\$0.14	Not specified	Free for some versions
OUTPUT COST (PER 1M TOKENS)	\$60.00	\$0.28	Not specified	Free for some versions
KEY STRENGTHS	Creative writing, coding, versatility	Math and reasoning tasks	Clever and insightful responses	Search integration, image/video capabilities
OPEN SOURCE	No	Yes	No	No

Intelligent Agent Model



Relationship Between Main AI Topics (Covered in Course)



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

Main Section

Gender Shades paper

1. Joy Buolamwini, Timnit Gebru. [Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification](#). In Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA. Volume 81 of Proceedings of Machine Learning Research, pages 77-91, PMLR, 2018.
2. <http://gendershades.org/overview.html>

Dataset

- a. Megaface, which to date is the largest publicly available set of facial images, was composed utilizing Head Hunter to select one million images from the Yahoo Flicker 100M image dataset
- b. LFW, a dataset composed of celebrity faces which has served as a gold standard benchmark for face recognition, was estimated to be 77.5% male and 83.5% White. Performance not broken down by gender or race
- c. Intelligence Advanced Research Projects Activity (IARPA) released the IJB-A dataset as the most geographically diverse set of collected face.
- d. As of 2017, The National Institute of Standards and Technology is starting another challenge to spur improvement in face gender classification by expanding on the 2014-15 study.

Pilot Parliament Benchmark

- They also used IJB-A and Adience for comparison.
- One author labeled each image with one of six Fitzpatrick skin types and provided gender annotations for the IJB-A dataset.
- The Adience benchmark was already annotated for gender.
- These preliminary skin type annotations on existing datasets were used to determine if a new benchmark was needed. For PPB, 3 annotators including the authors provided gender and Fitzpatrick labels.
- A board-certified surgical dermatologist provided the definitive labels for the Fitzpatrick skin type.
- Gender labels were determined based on the name of the parliamentarian, gendered title, prefixes such as Mr or Ms, and the appearance of the photo.

Inequality in Misclassification

- a. The gender misclassification rates on the Pilot Parliaments Benchmark replicate this trend across all classifiers. The differences between female and male classification error rates range from 8.1% to 20.6%.
- b. **Even though darker females make up 21.3% of the PPB benchmark, they constitute between 61.0% to 72.4.1% of the classification error**
- c. COTS1 and COTS2 APIs solely output single labels indicating whether the face was classified as female or male. COTS3's API outputs an additional number which indicates the confidence with which the classification was made. The authors note that giving crisp class labels does not give users the ability to analyze true positive (TPR) and false positive (FPR) rates for various subgroups if different thresholds were to be chosen.
- d. Errors do not seem to happen because of image quality. They consider South African photos of similar image quality as Europeans.

Discussion: What is Right (Fair)?

- Equal errors?
- Equal accuracy?
- Errors on individual faces?
- ...

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. The top navigation bar includes links for DATA, TOPICS, RESOURCES, STRATEGY, DEVELOPERS, and CONTACT. Below the navigation is a grid of icons representing various sectors: Agriculture, Climate, Ecosystems, Energy, Local Government, Maritime, Ocean, and Older Adults Health. A featured dataset is "U.S. Hourly Precipitation Data" with 855 recent views, described as a digital data set from the National Climatic Data Center (NCDC). Another dataset, "NCDC Storm Events Database", is also shown. The footer includes a note about OpenStreetMap tiles and a CC BY SA license.

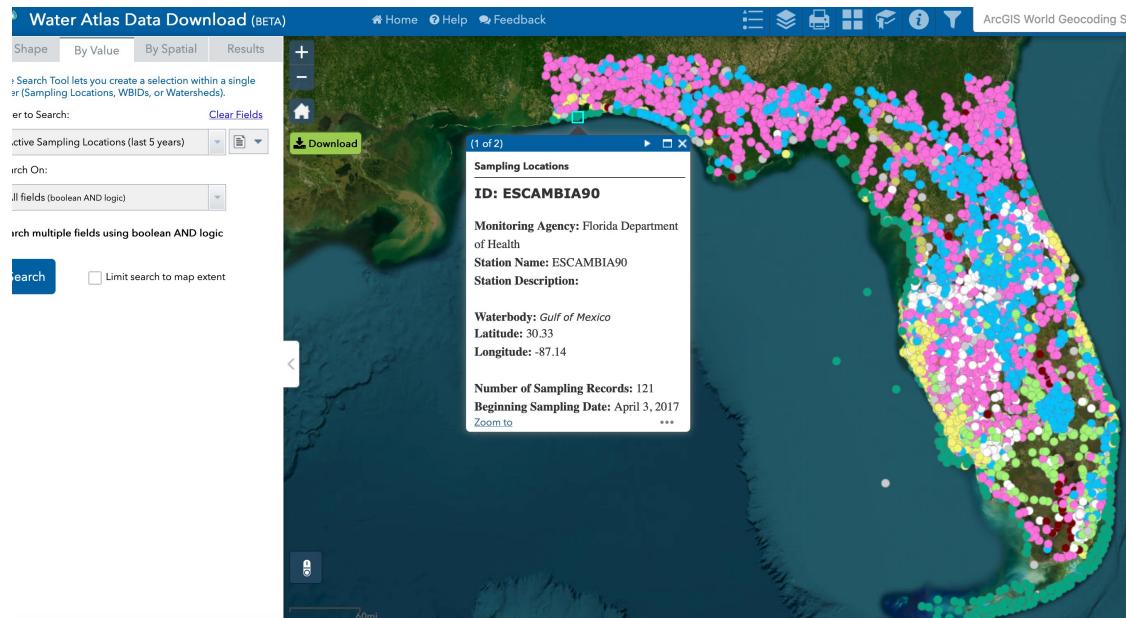
USA

A screenshot of the India data.gov.in website. The top navigation bar includes links for Skip to navigation, Skip to main content, DataGov States/ULB, and a search bar. The main banner highlights "DATASETS FROM HEALTH SECTOR". Below the banner are sections for 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 (showing a lightbulb icon and user figures), and INDICATOR DASHBOARD (showing icons for Drinking Water And Sanitation, Health, Transport, and Labour And Employment). A specific dataset, "Udyog Aadhaar Memorandum (MSME Registration)", is highlighted.

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>

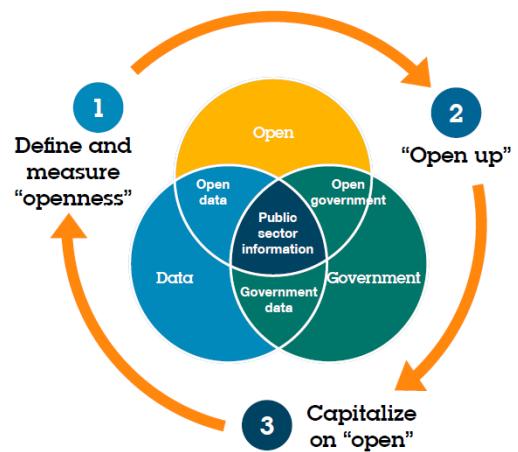


"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

Cartoon critical of big data application,
by T. Gregorius

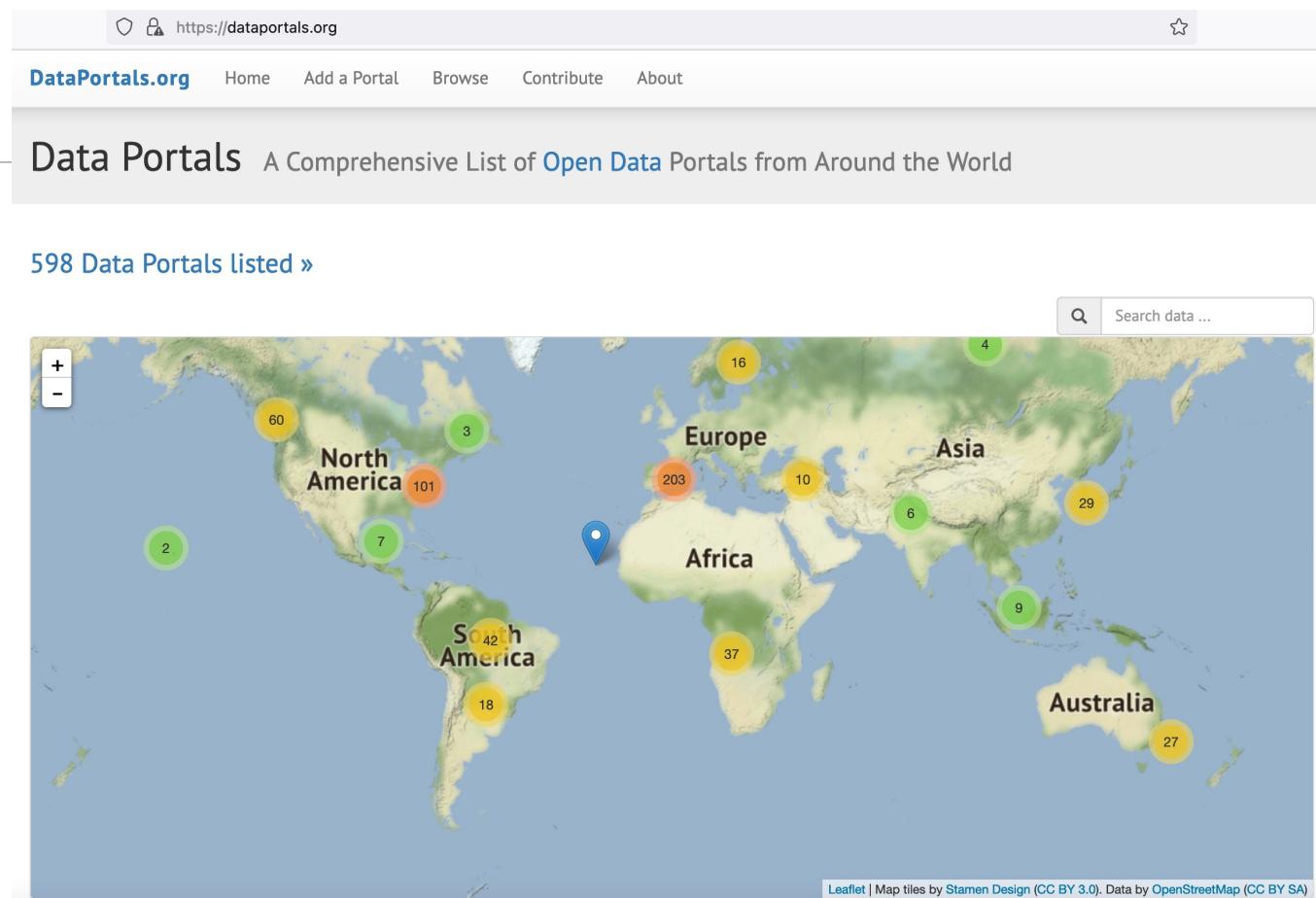
http://upload.wikimedia.org/wikipedia/commons/thumb/b/b3/Big_data_cartoon_t_gregorius.jpg/220px-Big_data_cartoon_t_gregorius.jpg

~600 Data Catalogs of Open Data



Source: IBM Institute for Business Value.

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"
1
                    "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"
1
                    "application/json"
```

The screenshot shows a JSON viewer interface with the URL 311api.cityofchicago.org/open311/discovery.json. The JSON data is displayed in a hierarchical tree view. The root object contains fields like changeset, contact, key_service, and endpoints. The endpoints array has two items, each with specification, url, changeset, type, and formats fields. The 'type' field for the first endpoint is highlighted in blue, indicating it's the current selection. The 'specification' and 'url' fields for both endpoints also have links.

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

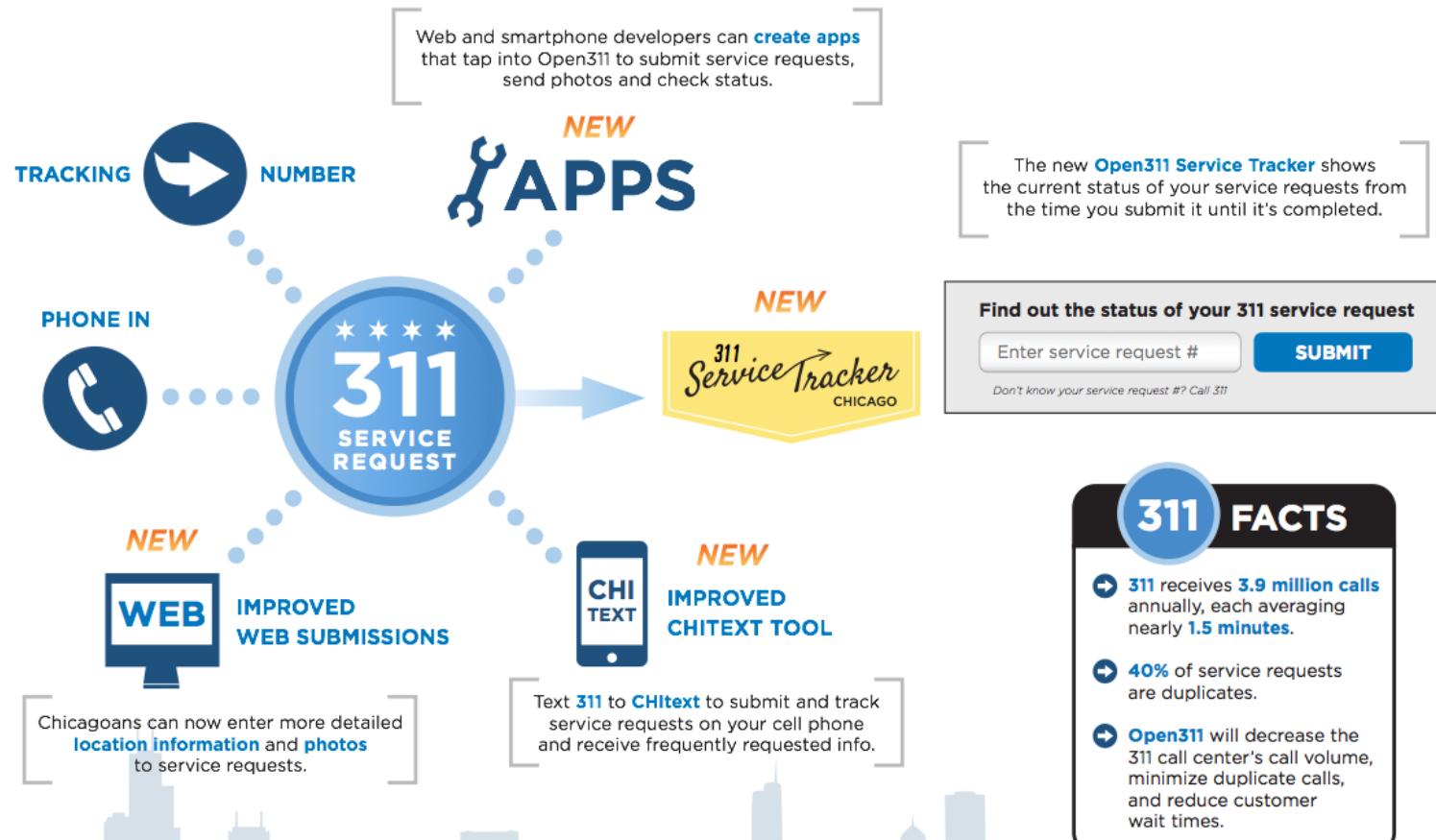
```
{"service_code":"4ffa9cad6018277d4000007b",
 "attributes":
 [{"variable":true,"code":"ISTHELI2",
 "datatype":"singlevaluelist","required":true,"order":1,
 "description":"Is the light located in your alley or the street?",
 "values":[{"key":"ALLEY","name":"Alley"},
 {"key":"STREET","name":"Street"}]},

 {"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:

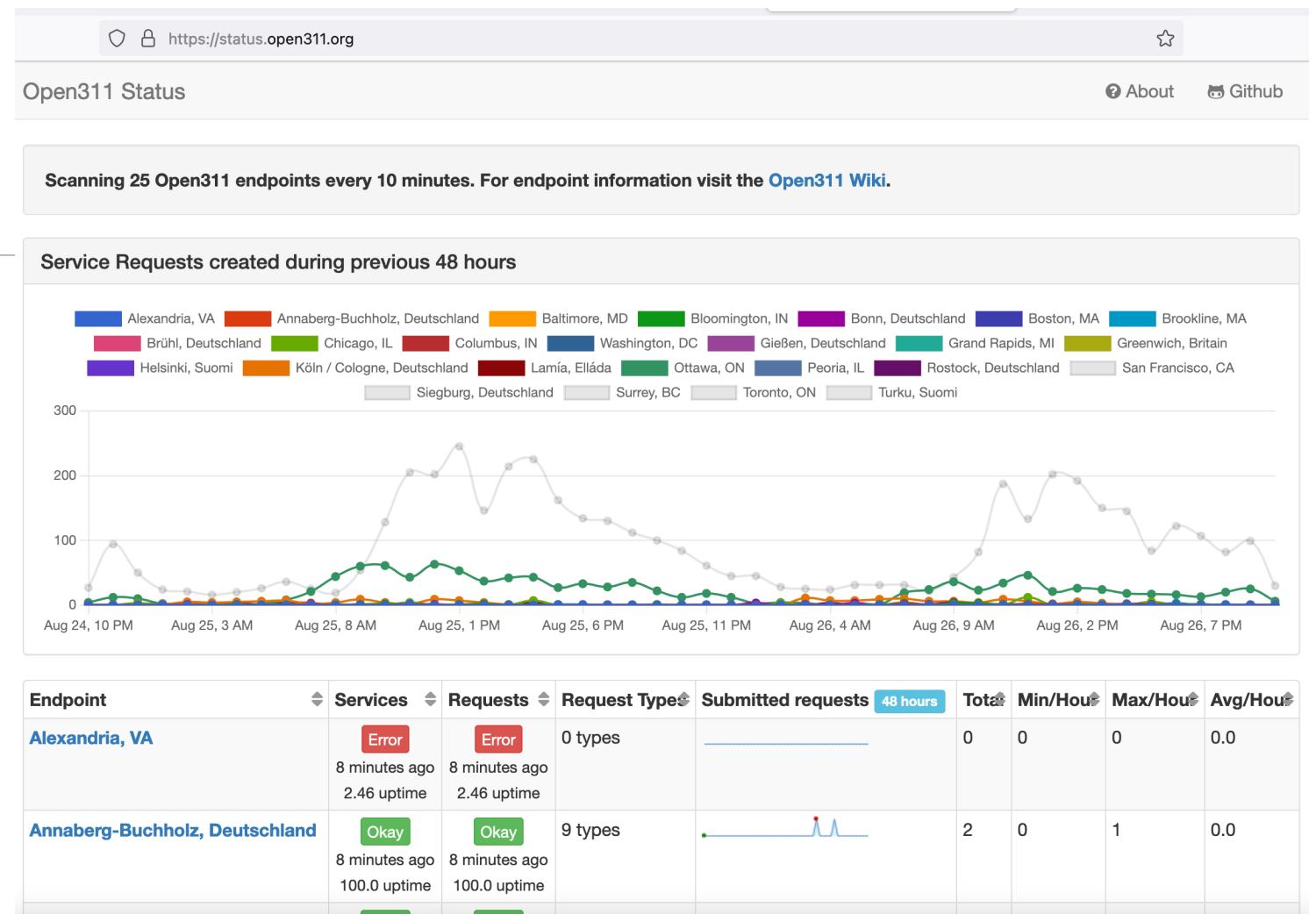
- #13-00210540**
- Address:** 1502 N Wicker Park Ave
- Created:** February 23, 2013
- Received via:** Other

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

---

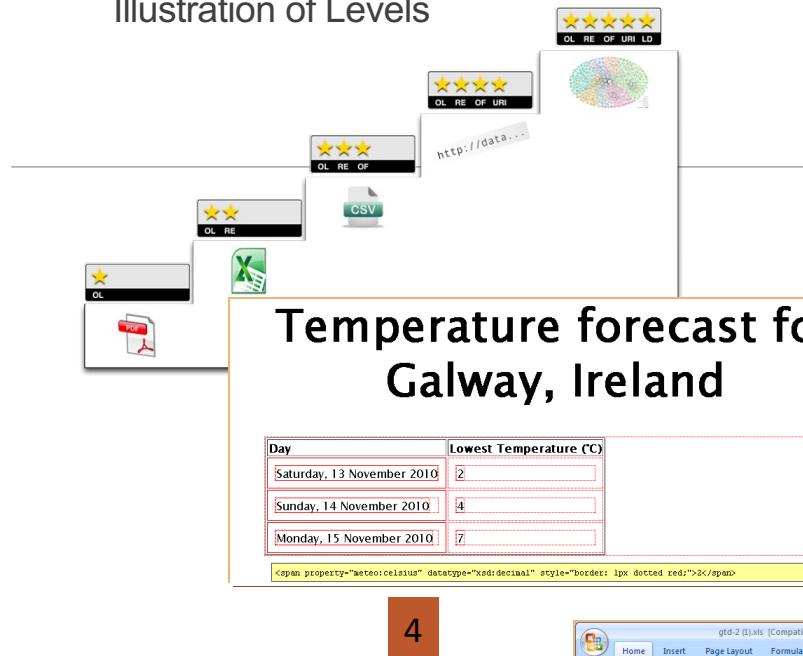
- Text of legislations - LegiScan, <https://legiscan.com/>
- Voter FAQs - <https://github.com/ai4society/election-dataset>
- Compendium of benchmarks and datasets:
  - <https://zilliz.com/learn/popular-datasets-for-natural-language-processing>,
  - UCI dataset – <https://archive.ics.uci.edu/datasets?search=&Types=Text>
  - 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

35

## 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

### gtd-3.csv - WordPad

File Edit View Insert Format Help



"Temperature forecast for Galway, Ireland",

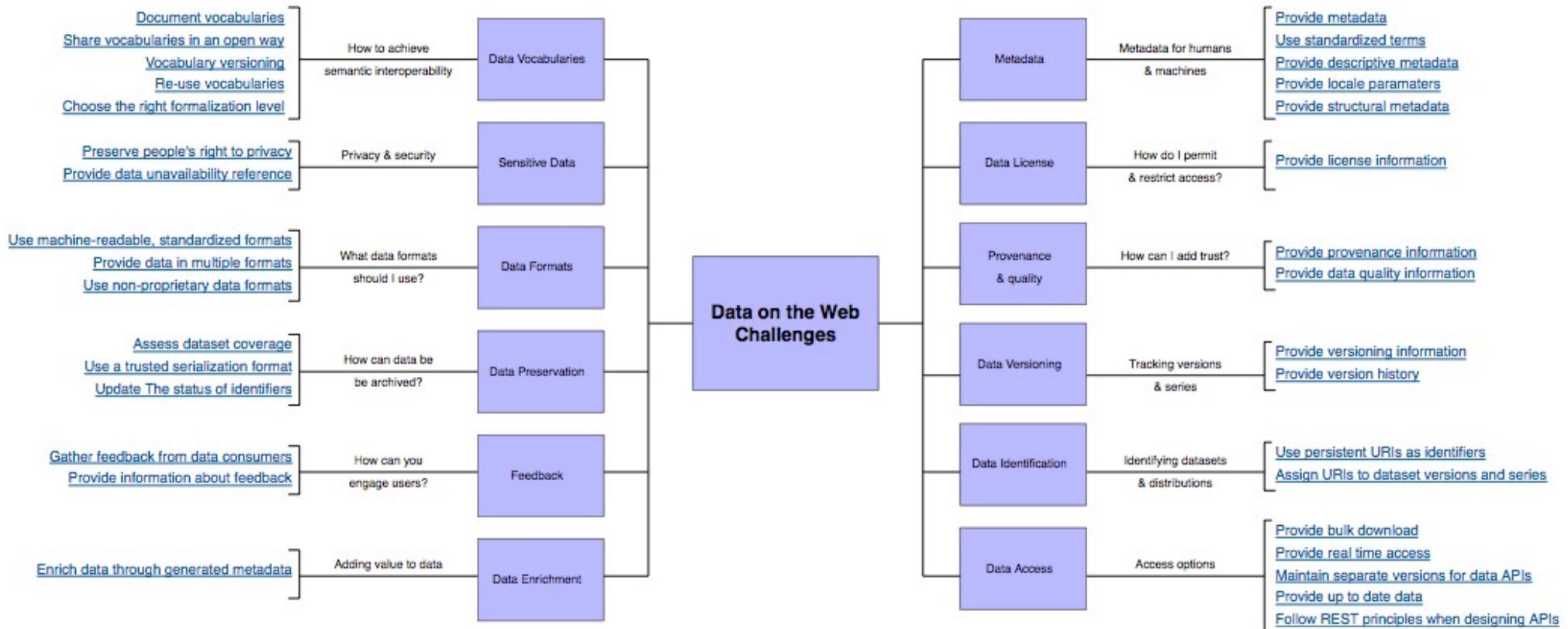
"Day", "Lowest Temperature (C)"  
 "Saturday, 13 November 2010", 2  
 "Sunday, 14 November 2010", 4  
 "Monday, 15 November 2010", 7

| A                                               | B                    |
|-------------------------------------------------|----------------------|
| <b>Temperature forecast for Galway, Ireland</b> |                      |
| 1                                               | Day                  |
| 2                                               | Lowest Temperature ( |
| 3                                               | 2010)                |
| 4                                               | 2                    |
| 5                                               | 4                    |
| 6                                               | 7                    |

2

# 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

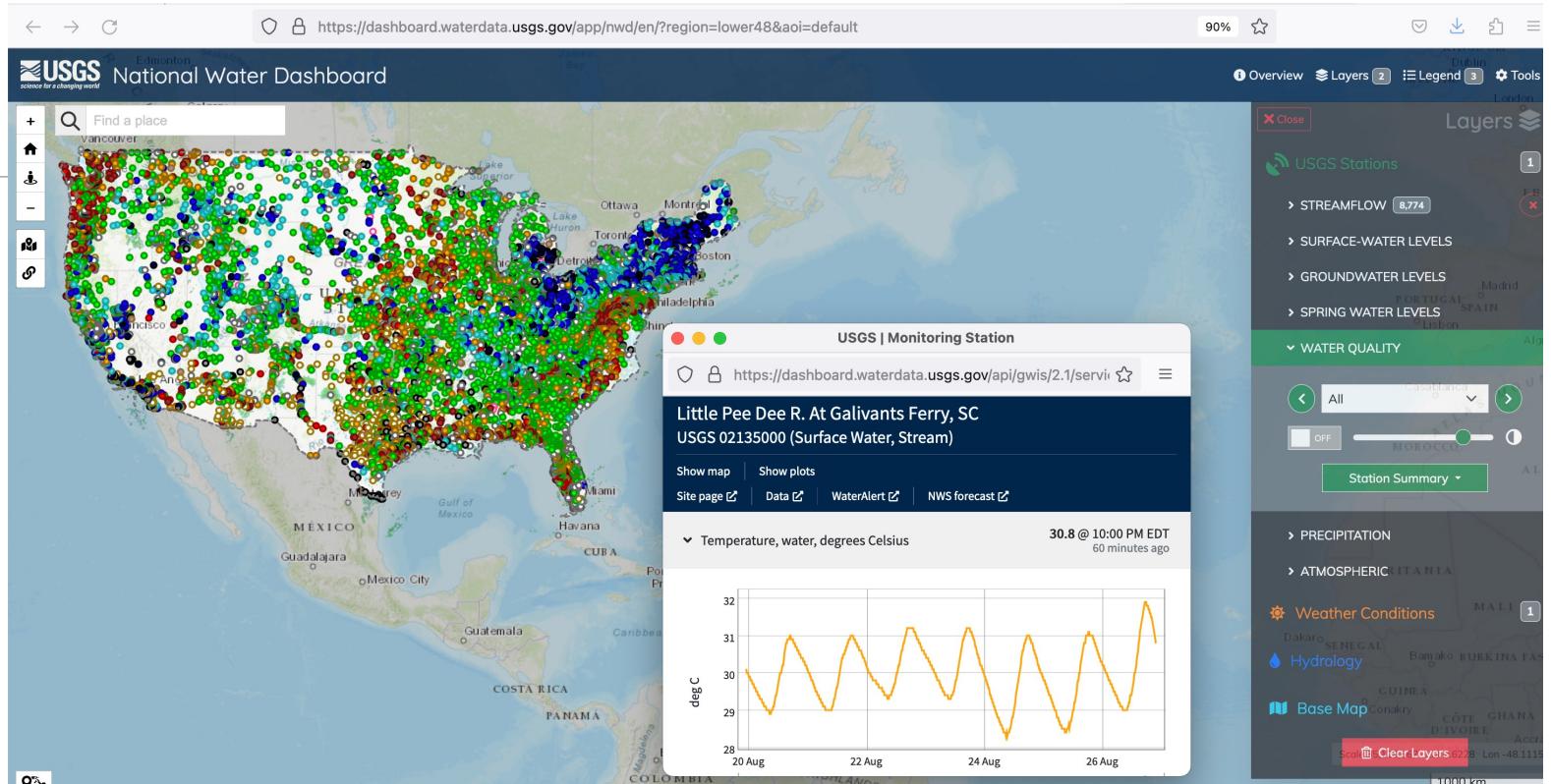
---

- What does a (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?

---

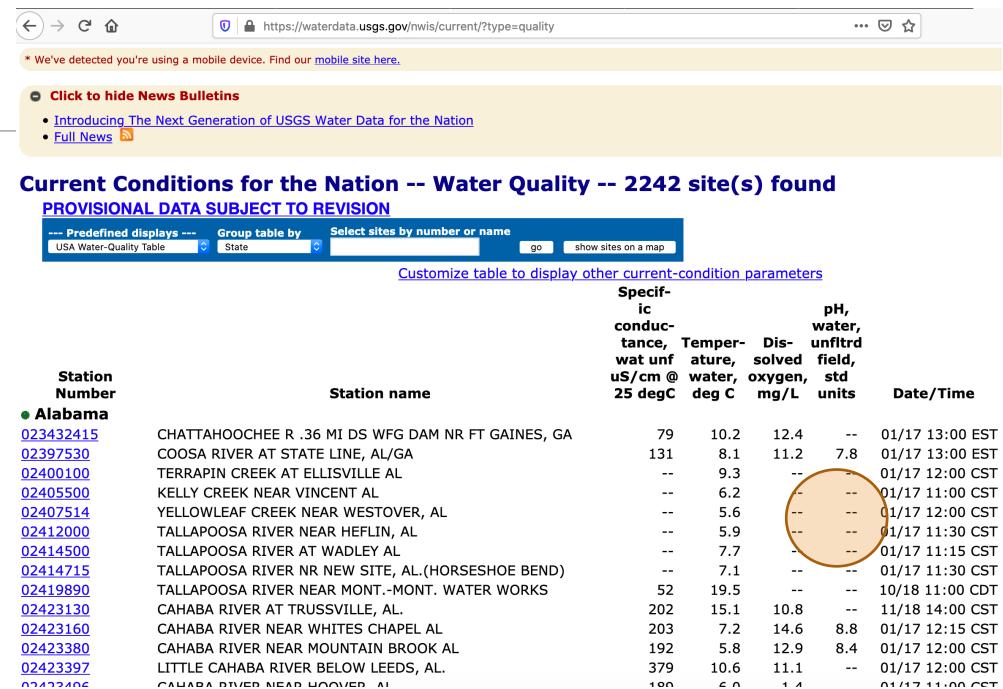
- Access and licensing (Class 3)
- Cleaning, organizing and finding related information (Class 4 – this class)
- Representing formally (in logic) to draw insights (using inferencing) – (Class 4 – this class)

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, and Date/Time. A red circle highlights the "Dissolved oxygen" column for the first few rows.

| Station Number | Station name                                       | Specific conductance at 25 degC | Temperature at 25 degC | Dissolved oxygen mg/L | 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

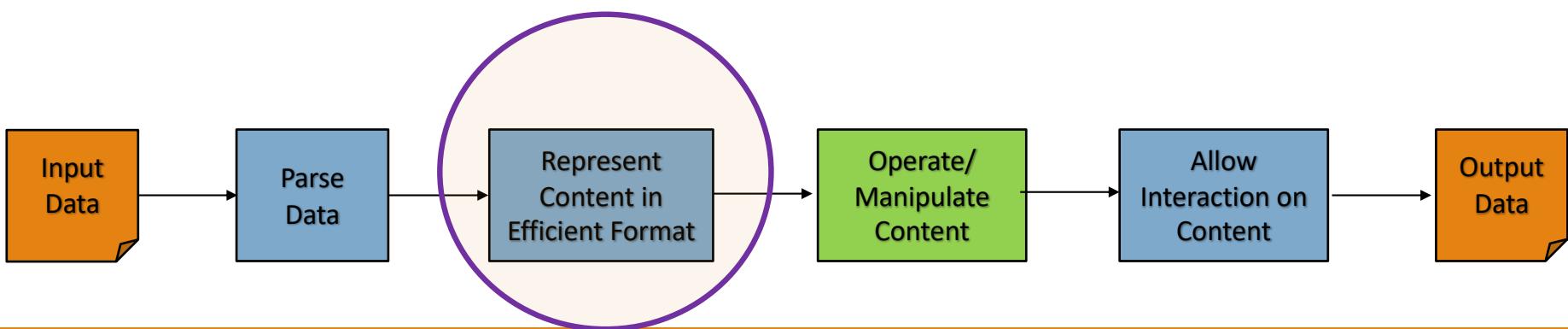
---

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

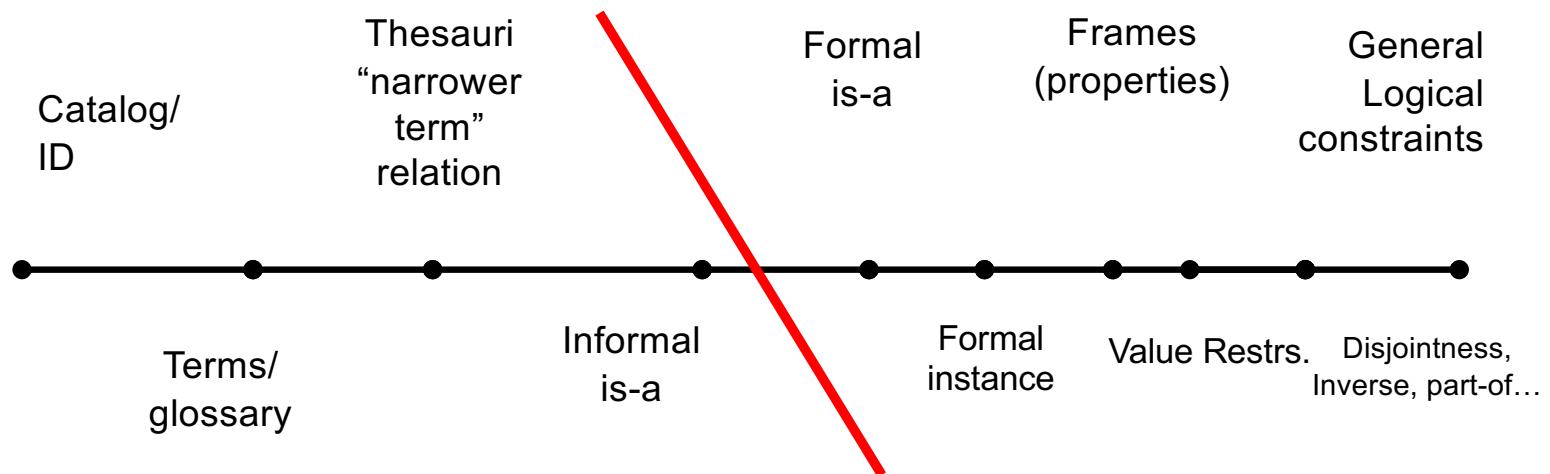
- Basic concepts: **DataPreparation-Numeric.ipynb**
- An illustration: **Clean-RealSample.ipynb**

# 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:

```

 Save A Lot Monitors - $1250

 Jon Doe's Gadgets - $1350
...
```

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:

Save A Lot Monitors - $1250

Jon Doe's Gadgets - $1350
...

```

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:

 Save A Lot Monitors - $1250

 Jon Doe's Gadgets - $1350
...

```

## 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

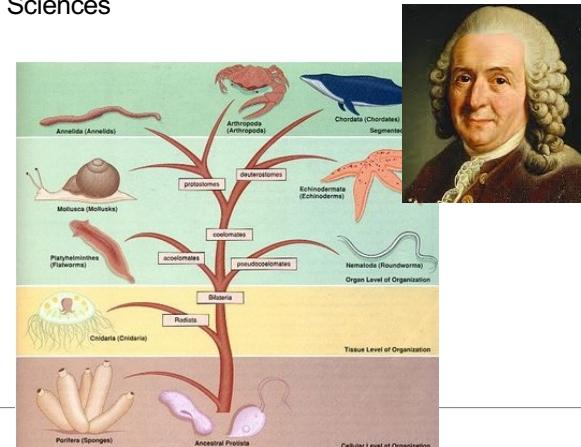
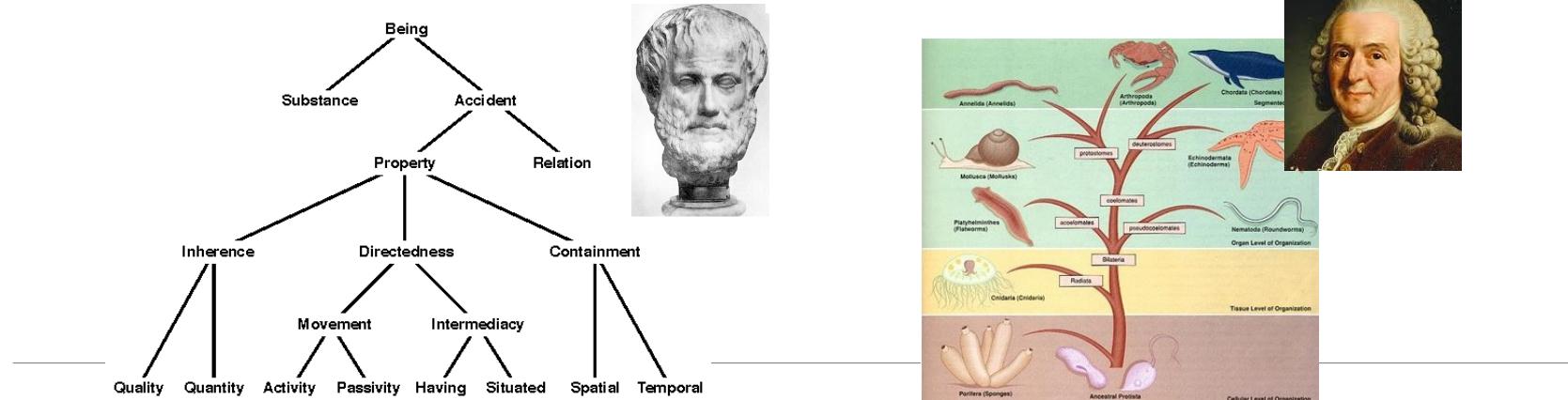
SCRIBE presentation by Rosario Usceda Sosa, Biplav Srivastava, Bob Schloss

- <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

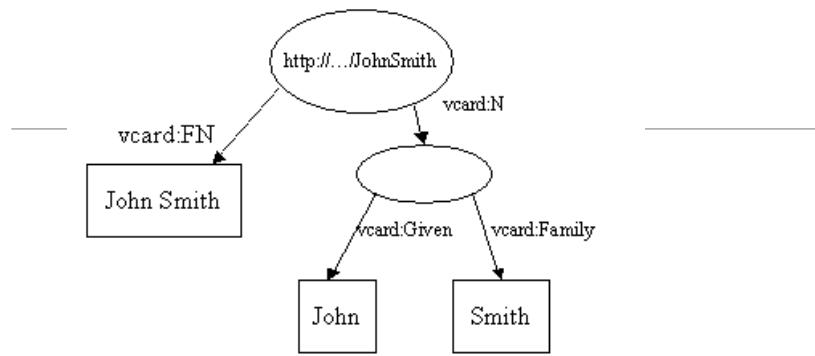


In a Smart City domain, we’re concerned with modeling the *city data* (city activity data, city departments, assets, KPIs), not the city itself (the full set of spatial and temporal relations between people and objects in the city). Ontologies help us to structure and reason about city events, entities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

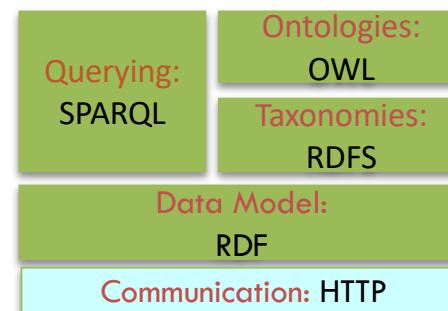
- Range, domain
- Local, global
- Existential
- Cardinality

## + Combinators

- Union, Intersection
- Complement
- Symmetric, transitive

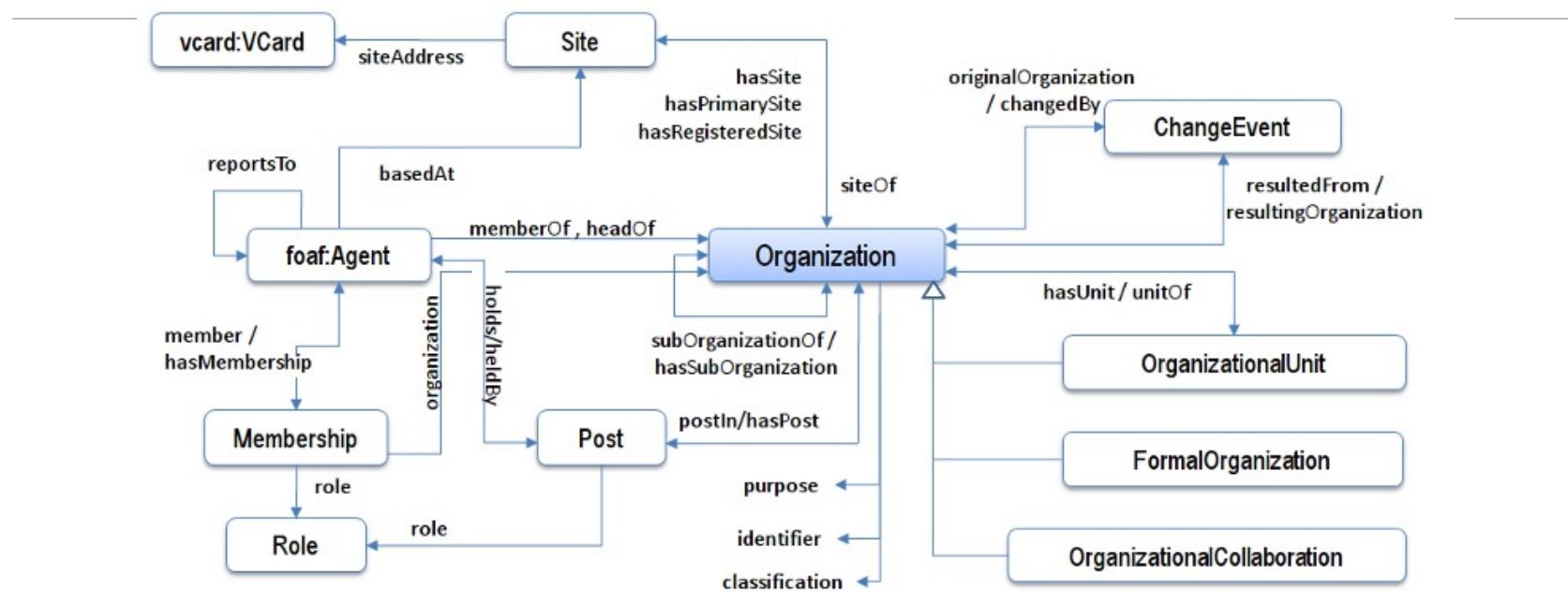
## + Mapping

- 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>

```

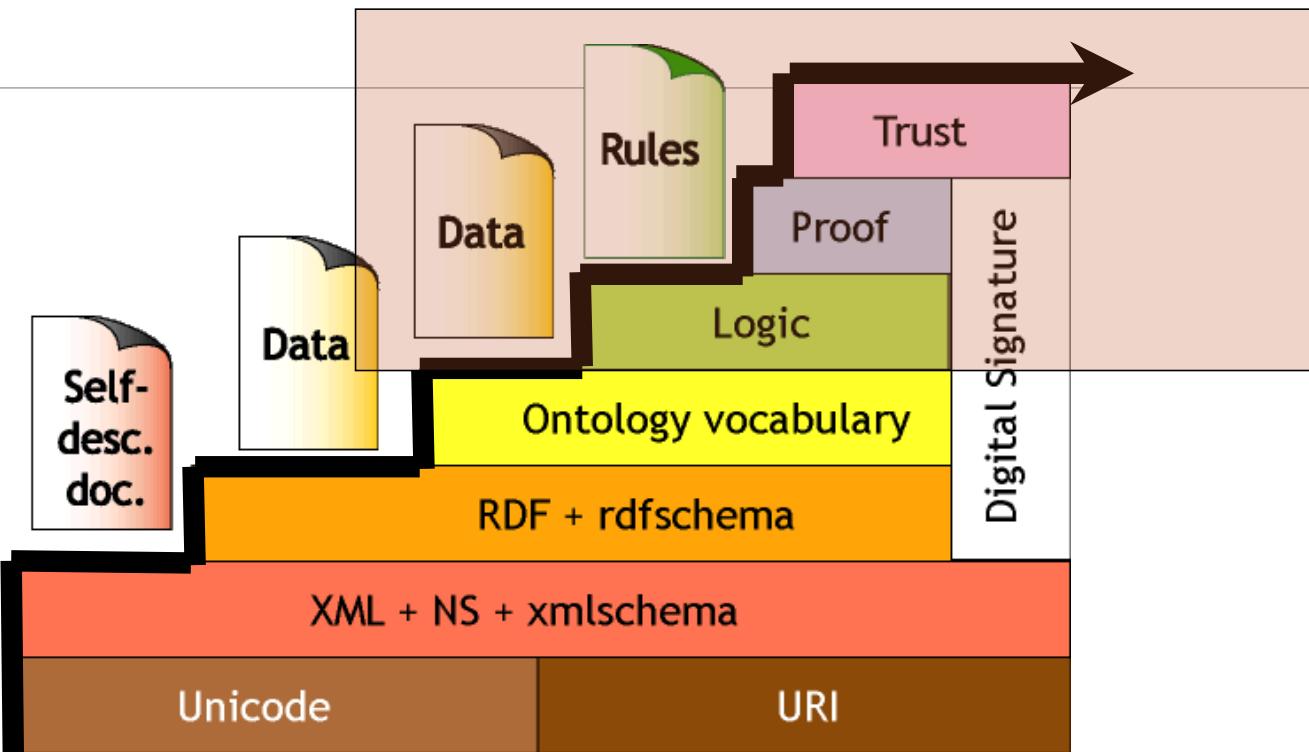
- <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:skos="http://www.w3.org/2004/02/skos/core#" xmlns:foaf="http://xmlns.com/foaf/0.1/"
 xmlns:org="http://www.w3.org/ns/org#" xmlns:gr="http://purl.org/goodrelations/v1#"
 xmlns:owl="http://www.w3.org/2002/07/owl#" xmlns:dct="http://purl.org/dc/terms/"
 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" />
 <rdf: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

```

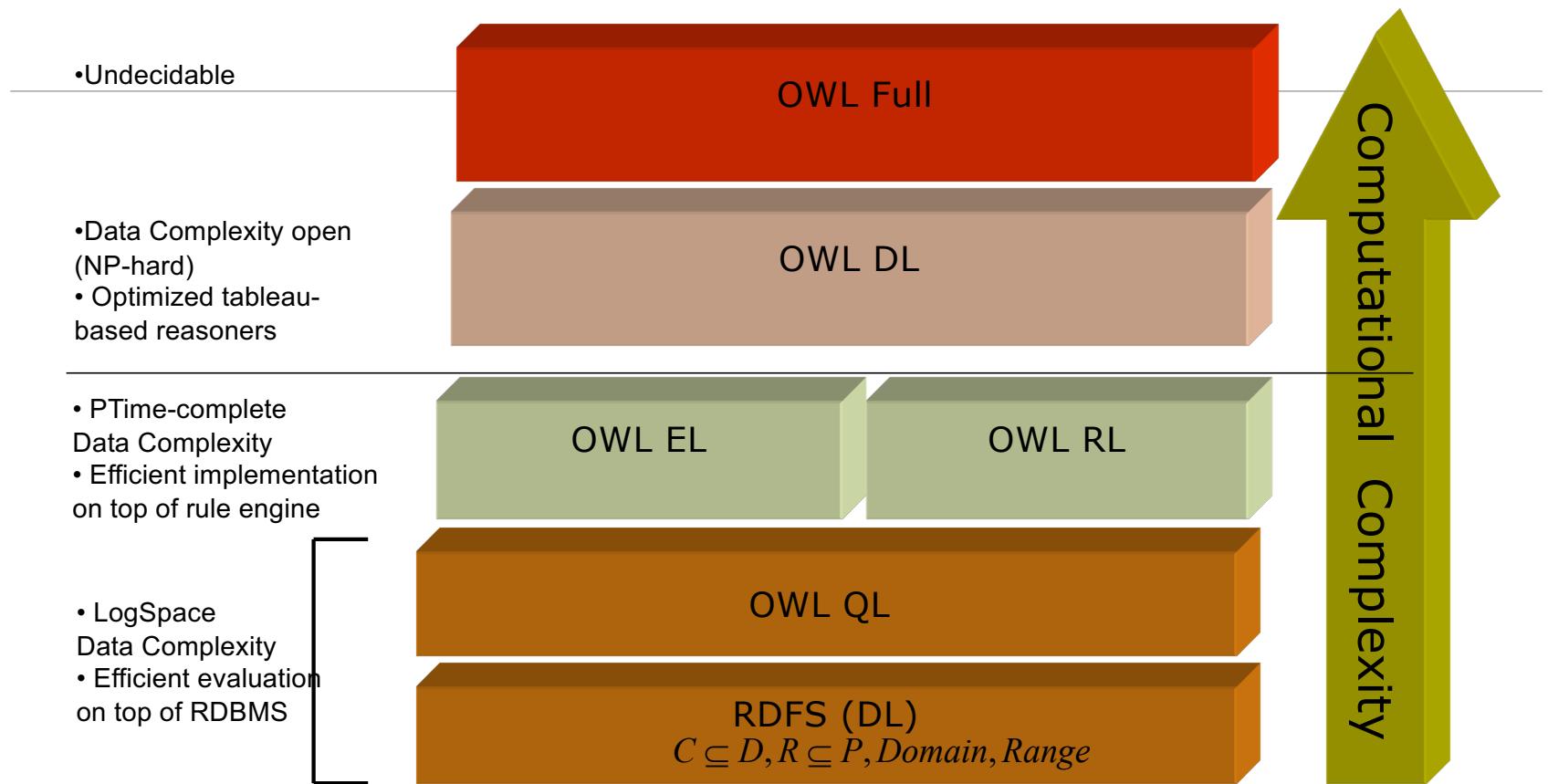
<http://www.w3.org/ns/org>

# 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>](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

# Machine Learning

---



Credit: Retrieved from internet

# Machine Learning – Insights from Data

---

- Descriptive analysis
  - Describe a past phenomenon
  - **Methods:** classification (feedback from label), clustering, dimensionality reduction, anomaly detection, neural methods, reinforcement learning (feedback from hint/ reward)
- Predictive analysis
  - Predict about a new situation
  - **Methods:** time-series, neural networks
- Prescriptive analysis
  - What an agent should do
  - **Methods:** simulation, reinforcement learning, reasoning
- New areas
  - Counterfactual analysis
  - Causal Inferencing
  - Scenario planning

# Nomenclature

---

Row, Item →

Column, Attribute, Feature

1	PID	ST_NUM	ST_NAME	OWN_OCCUPIED	NUM_BEDROOMS	NUM_BATH	SQ_FT
2	100001000	104	PUTNAM	Y	3	1	1000
3	100002000	197	LEXINGTON	N	3	1.5	--
4	100003000		LEXINGTON	N	n/a	1	850
5	100004000	201	BERKELEY	12	1	NaN	700
6		203	BERKELEY	Y	3	2	1600
7	100006000	207	BERKELEY	Y	NA	1	800
8	100007000	NA	WASHINGTON		2	HURLEY	950
9	100008000	213	TREMONT	Y	1	1	
10	100009000	215	TREMONT	Y	na	2	1800

# Types of Attributes/ Columns

---

- Numeric: has number as value in computational sense; all mathematical functions are valid.
  - Example: SQ\_FT
- Categorical: has distinct values
  - Nominal: each value is incomparable with other
    - Example: OWN\_OCCUPIED, ST\_NAME
  - Ordinal: the values can be ordered
    - Example: ST\_NUM, NUM\_BEDS
- Comment:
  - Q: what type is a binary variable?
  - A: depends on the semantics – nominal (gender), ordinal (number basements).

	PID	ST_NUM	ST_NAME	OWN_OCCUPIED	NUM_BEDROOMS	NUM_BATH	SQ_FT
1	100001000	104	PUTNAM	Y	3	1	1000
2	100002000	197	LEXINGTON	N	3	1.5	--
3	100003000		LEXINGTON	N	n/a	1	850
4	100004000	201	BERKELEY	12	1	NaN	700
5		203	BERKELEY	Y	3	2	1600
6	100006000	207	BERKELEY	Y	NA	1	800
7	100007000	NA	WASHINGTON		2	HURLEY	950
8	100008000	213	TREMONT	Y	1	1	
9	100009000	215	TREMONT	Y	na	2	1800

# Why is Type of Variable Important

---

- Handling of missing values
- Distance between
  - Values
  - Data items
- Used for measuring accuracy, error
- Guiding the learning process
  - Selection of algorithms

# Concepts

---

- **Input data:** data available
  - **Training data:** used for training a learning algorithm and get a model
  - [Optional] **Validation data:** used to tune parameters
- **Test data:** used to test a learning model

- **Classification problem**

- Separating data into classes (also called labels, categorical types)
- One of the attributes is the class label we are trying to learn
- Class label is the **supervision**

- **Clustering problem**

- We are trying to learn grouping of data
- There is no attribute indicating membership in the groups (hence, **unsupervised**)

- **Prediction problem**

- Learning value of a continuous variable

Reference: <https://machinelearningmastery.com/difference-test-validation-datasets/>

<https://www2.seas.gwu.edu/~bell/csci243/lectures/classification.pdf>

# Sample Learning Task

---

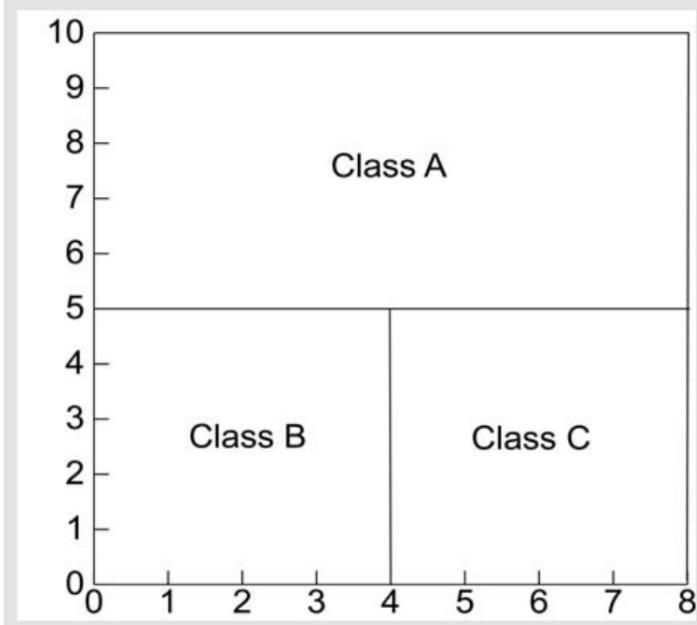
- COVID-19 data

Notebook: <https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l6-l7-l8-supervised-ml/Supervised-Regression-Classification.ipynb>

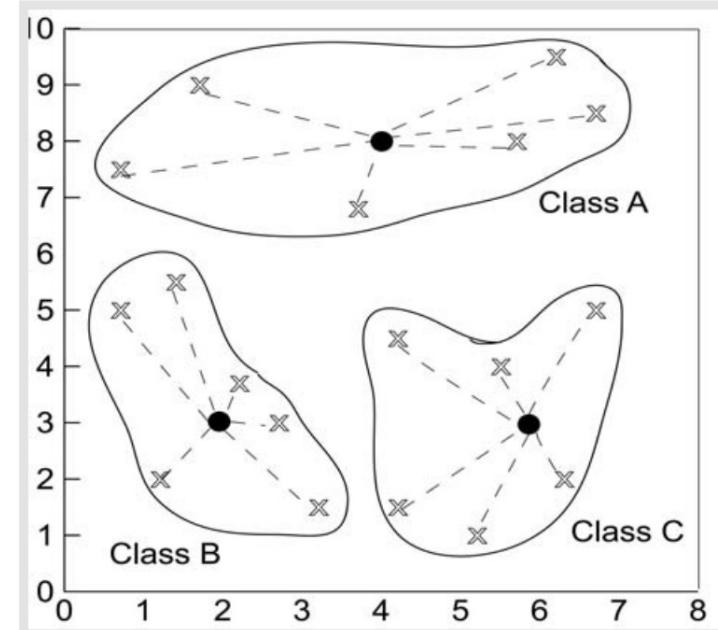
-

# Methods for Classification

Partitioning Based



Distance Based

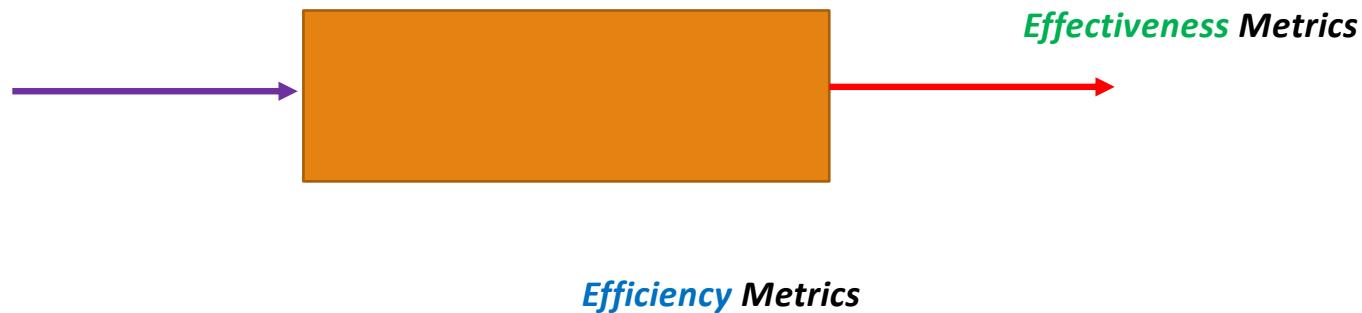


Source: <https://www2.seas.gwu.edu/~bell/csci243/lectures/classification.pdf>

# Metric Types

---

- **Effectiveness**: what the user of a system sees, primarily cares about
- **Efficiency**: what the executor in a system sees, primarily cares about



# Example: Predicting COVID cases

---

- **Effectiveness:** what the user of a system sees, primarily cares about
  - *How accurate (high) is the prediction?*
  - *How low is the error?*
- **Efficiency:** what the executor in a system sees, primarily cares about
  - *How low is the error?*
  - *How fast was prediction made?*
  - *How stable is the prediction to change in data?*

# Example: Detecting Spam in Email

---

- **Effectiveness:** what the user of a system sees, primarily cares about
  - *How many spams identified?*
  - *How many spams missed?*
- **Efficiency:** what the executor in a system sees, primarily cares about
  - *How fast were spams detected?*
  - *How much memory was used per million emails processed ?*

# Comparing Classification Methods

---

- Predictive accuracy
  - Interpretability: providing insight
  - Robustness: handling noisy data
- 
- Speed
  - Scalability: large volume of data

Source: Data Mining: Concepts and Techniques, by Jiawei Han and Micheline Kamber

# Metrics: Accuracy, Precision, Recall

---

		Predicted class	
Actual Class		Class = Yes	Class = No
		Class = Yes	True Positive
	Class = No	False Positive	True Negative

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+FN+TN)}$$

$$\text{Precision} = \frac{(TP)}{(TP+FP)}$$

$$\text{Recall} = \frac{(TP)}{(TP+FN)}$$

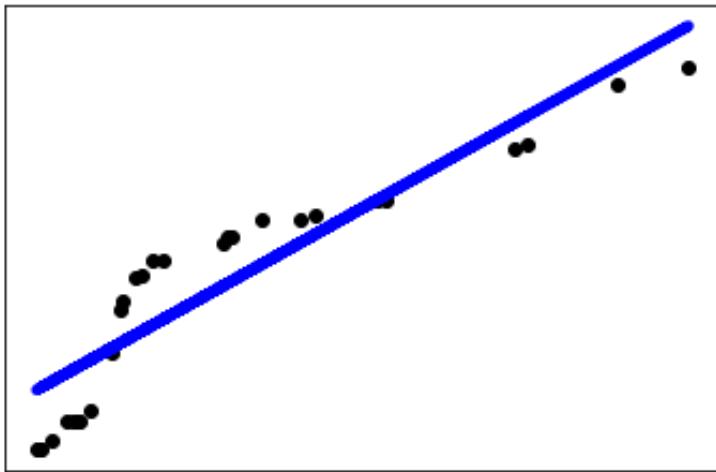
**F1 Score: Harmonic Mean**

$$1/F1 = 1/\text{Precision} + 1/\text{Recall}$$

$$F1 = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

# Linear Regression

---



Notebook: <https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l6-l7-l8-supervised-ml/Supervised-Regression.ipynb>

# Reference and Demo

- Data: UCI Datasets

- <https://archive.ics.uci.edu/datasets>
- Browse or search

The screenshot shows the homepage of the Weka 3 website. At the top, there is a navigation bar with links for Project, Software, Book, Courses, Publications, People, and Related. Below the navigation bar, the title "Weka 3: Machine Learning Software in Java" is displayed. A brief introduction states: "Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization." A note below says: "Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive nature. The name is pronounced like this, and the bird sounds like this." Another note mentions: "Weka is open source software issued under the GNU General Public License." It also links to free online courses and mentions deep learning support.

The screenshot shows the "Browse Datasets" page of the UCI Machine Learning Repository. The URL is https://archive.ics.uci.edu/datasets. On the left, there are filters for Keywords, Data Type (Multivariate, Sequential, Tabular, Text, Time-Series, Other), Subject Area, and Task. The main area is titled "Browse Datasets" and shows a list of datasets. The first dataset listed is "Iris", described as a small classic dataset from Fisher, 1936. It is a Classification task using Tabular data with 150 instances and 4 features. The second dataset is "Heart Disease", a Classification task using Multivariate data with 303 instances and 13 features. The third dataset is "Adult", a Classification task using Multivariate data with 48,84K instances and 14 features. The fourth dataset is "Wine", a Classification task using Tabular data with 178 instances and 13 features.

- Tools:

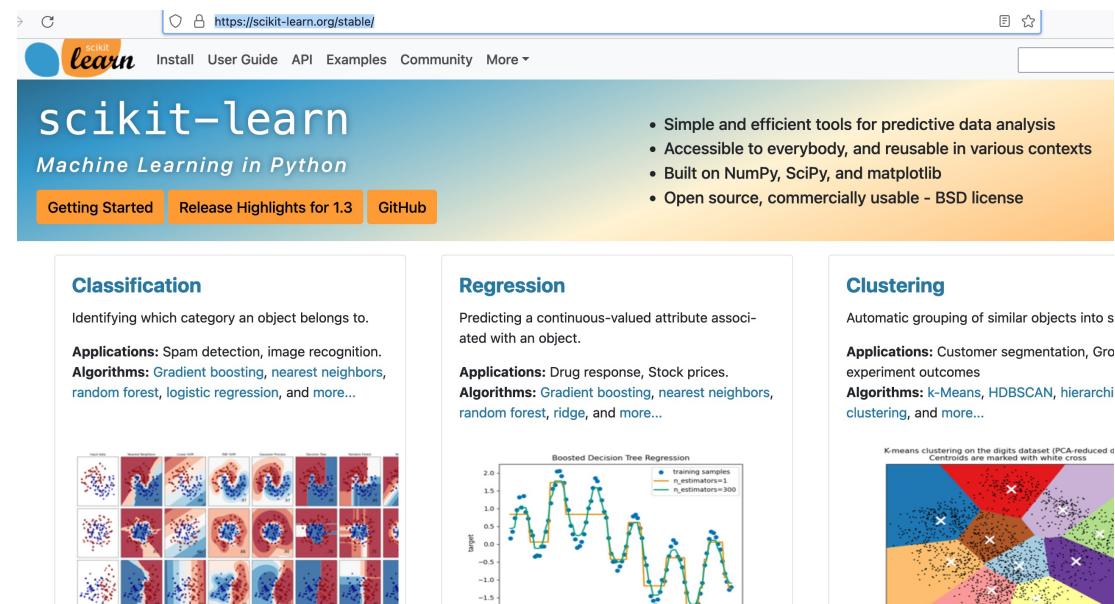
- Weka - <https://www.cs.waikato.ac.nz/ml/weka/>
- Download tool and dataset

- Libraries

- Scikit - <https://scikit-learn.org/stable/>

# Reference and Demo

- Data: UCI Datasets
  - <https://archive.ics.uci.edu/datasets>
  - Browse or search
- Tools:
  - Weka -  
<https://www.cs.waikato.ac.nz/ml/weka/>
  - Download tool and dataset
- Libraries
  - Scikit - <https://scikit-learn.org/stable/>



The screenshot shows the official website for scikit-learn (<https://scikit-learn.org/stable/>). The header includes the scikit-learn logo, navigation links for Install, User Guide, API, Examples, Community, and More, and a search bar. The main content area features the title "scikit-learn" and subtitle "Machine Learning in Python". Below this are three main sections: "Classification", "Regression", and "Clustering". Each section contains a brief description, a list of applications and algorithms, and a corresponding visual example. The "Classification" section shows a 3x3 grid of small plots illustrating various classification models. The "Regression" section shows a line plot of a regression model's performance. The "Clustering" section shows a scatter plot with colored regions representing different clusters.

# Exercise: German Credit

---

- Check in UCI
- Look at variants
  - <https://archive.ics.uci.edu/dataset/573/south+german+credit+update>

# Project Discussion

---

# Course Project

---

## •Framework

1. (Problem) Think of a problem whose solution may benefit people (e.g., health, water, air, traffic, safety)
2. (User) Consider how the primary user (e.g., patient, traveler) may be solving the problem today
3. (AI Method) Think of what the solution will do to help the primary user
  1. Solution => ML task (e.g. classification), recommendation, text summarization, ...
  2. Use a foundation model (e.g., LLM-based) solution as the baseline
4. (Data) Explore the data for a solution to work
5. (Reliability: Testing) Think of the evaluation metric we should employ to establish that the solution will work? (e.g., 20% reduction in patient deaths)
6. (Holding Human Values) Discuss if there are fairness/bias, privacy issues?
7. (Human-AI) Finally, elaborate how you will explain the primary user that your solution is trustable to be used by them

# Project Discussion: What to Focus on ?

---

- Problem: you should care about it
- Data: should be available
- Method: you need to be comfortable with it. Have at least two – one serves as baseline
- Trust issue
  - Due to Users
    - Diverse demographics
    - Diverse abilities
    - Multiple human languages
  - Or other impacts
- What one does to mitigate trust issue

# Rubric for Evaluation of Course Project

---

## Project

- Project plan along framework introduced (7 points)
- Challenging nature of project
- Actual achievement
- Report
- Sharing of code

## Presentation

- Motivation
- Coverage of related work
- Results and significance
- Handling of questions

# Project Discussion

1. Create a private Github repository called “**CSCE581-Spring2025-<studentname>-Repo**”. Share with Instructor (biplav-s)
2. Create a folder called “Project”. Inside, create a text file called “ProjectPlan.md” (or “ProjectPlan.txt”) and have details by the next class (Jan 30, 2025)

1. Title:
2. Key idea: (2-3 lines)
3. Who will care when done:
4. Data need:
5. Methods:
6. Evaluation:
7. Users:
8. Trust issue:

# Concluding Section

---

# Week 3 (L5 and L6): Concluding Comments

---

- We looked at
  - Data and characteristics
  - ML background and methods
- Prepares us for understanding trust issues

# About Next Week – Lectures 5, 6

---

# Lectures 7, 8: AI / ML Methods and Trust

---

- Supervised ML
- Trust issues

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