AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 590-1: Trusted AI

Lecture 3: AI: Data to Decisions Perspective, Prep, KG

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

26TH AUG 2021

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 3

- Introduction Segment
  - Recap from Lecture 2
- Main Segment
  - Importance of data
  - Types
    - By format: Structured, Semi-structured, Unstructured,
    - By media: text, audio, video, multi-media;
    - By source: enterprise, open data, social, private
  - Open Data
  - Representation: Catalog to ontology/knowledge graphs
- Concluding Segment
  - Project discussion
  - About Next Lecture – Lecture 4
  - Ask me anything

Introductory Segment

Recap of Lecture 2

- We looked at issues of Trust and Ethics
  - Trusted: competent, reliable, human values, human-AI interaction
  - Trust in technology and human are related but latter is more complex
  - AI ethics: AI leads to unique considerations
- Quality of decisions: lot of scope to improve
- Decisions during COVID-19 are complex
  - Decisions are often not trustable

AI Explanation Experience

- AI methods can be used to generate explanations for the predictions they are making. To give a sense of predictions and explanations generated, consider participating in a survey how good such methods are for images.
  - Task: object detection in images and its explanation.
  - Form: https://forms.gle/JiAhXrsRW46e8nTF7.
- Discussion on experience

Main Segment

Context: Sub-areas of AI

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

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

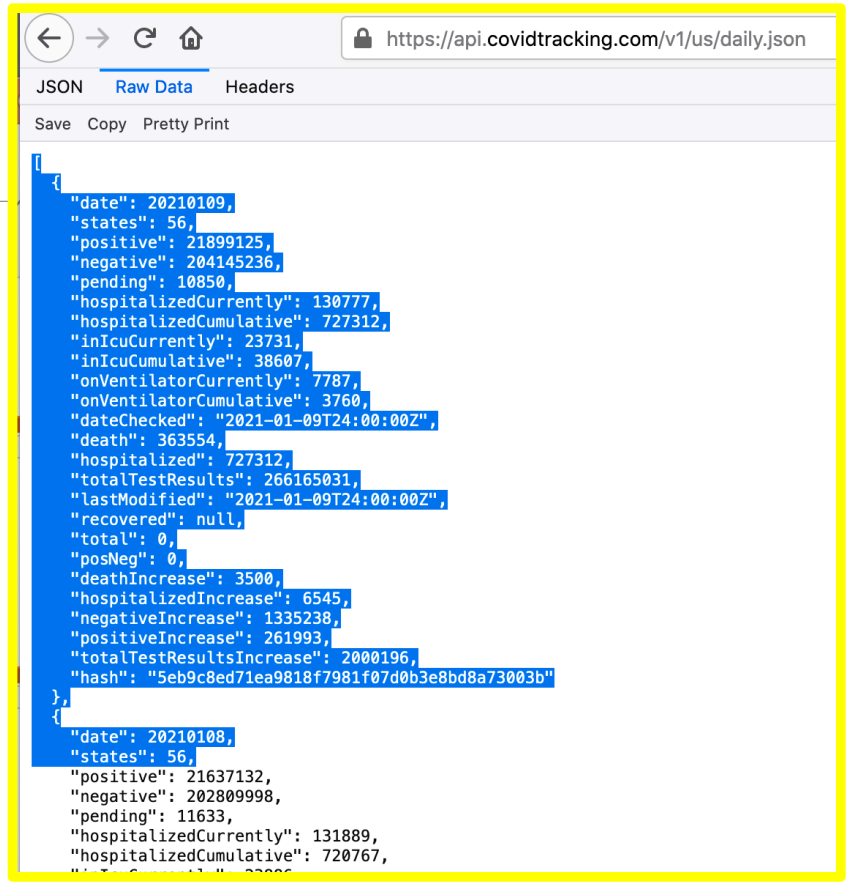
…
France,,weekly,2020-04-27,2020-05-03,2020,4,18,10498,10357,141,2010-2018 weekly average
…

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

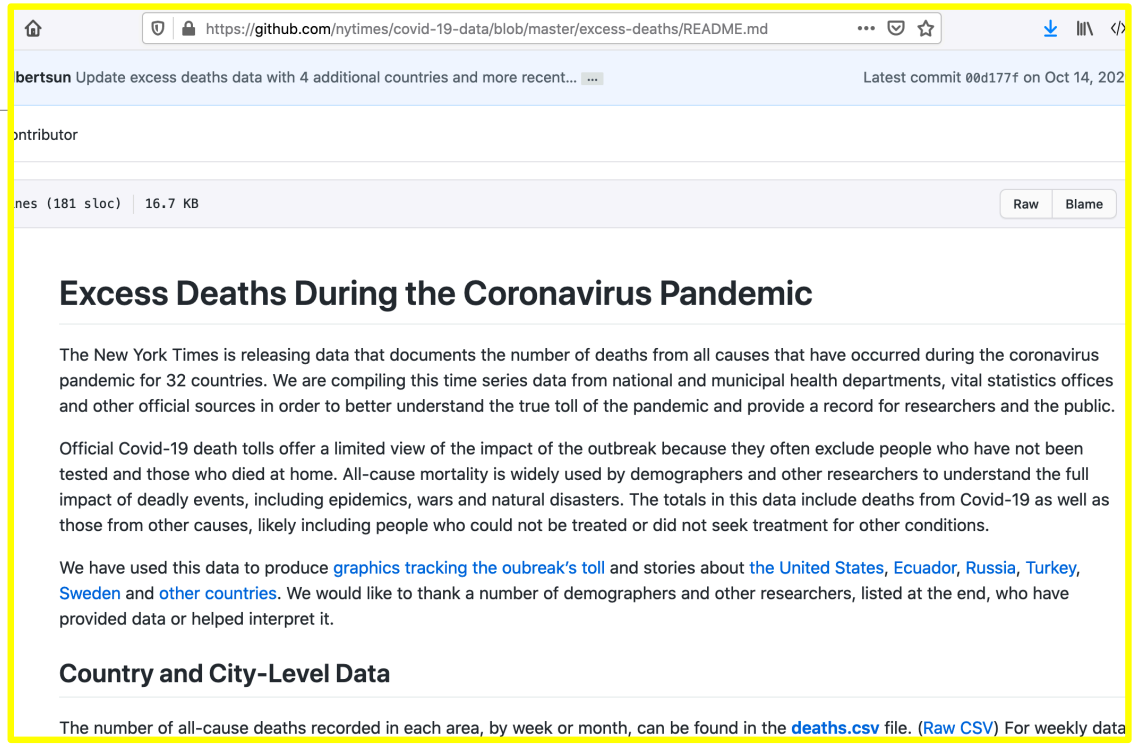
- The structure of meta-data is fixed, but the structure of data is allowed to change. Example: XML, JSON
- Benefits
  - Relatively easy to analyze, e.g., commands similar to SQL in langauges like OQL or Xquery
  - Structure of data easy to extend
- Disadvantage
  - Size of data is larger than structured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



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

Types of Data – Unstructured

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



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

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

Textual Data

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

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

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

Sound

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

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

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

Visual

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

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

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

Code Time: Processing Data Types

https://github.com/biplav-s/course-nl/tree/master/l2-languages

Data by Source

- Enterprise data is most common
  - Created and used by an individual or organization

- Social is promising data
  - People are anyway generating it (People-as-sensors)
  - However, social sites have varying data reuse permissions, license costs, access limits
  - Big data techniques already being used here

- Use sensor data if available
  - Internet of Things (IoT) and big data techniques are relevant
  - Most prevalent in health, environment and transportation

Open Data

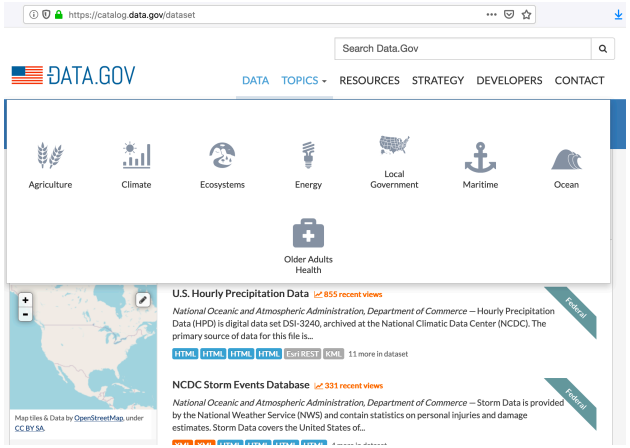
"Open data and content can be **freely used, modified, and shared** by **anyone** for **any purpose**"

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

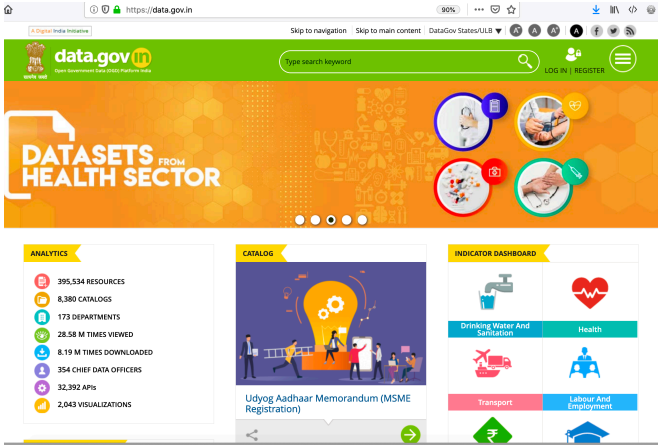
Often easiest to get but with issues (e.g., at aggregate level, with gaps, imprecise semantics)

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

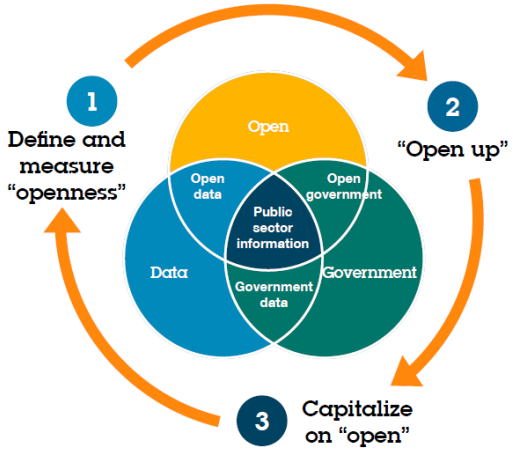


USA

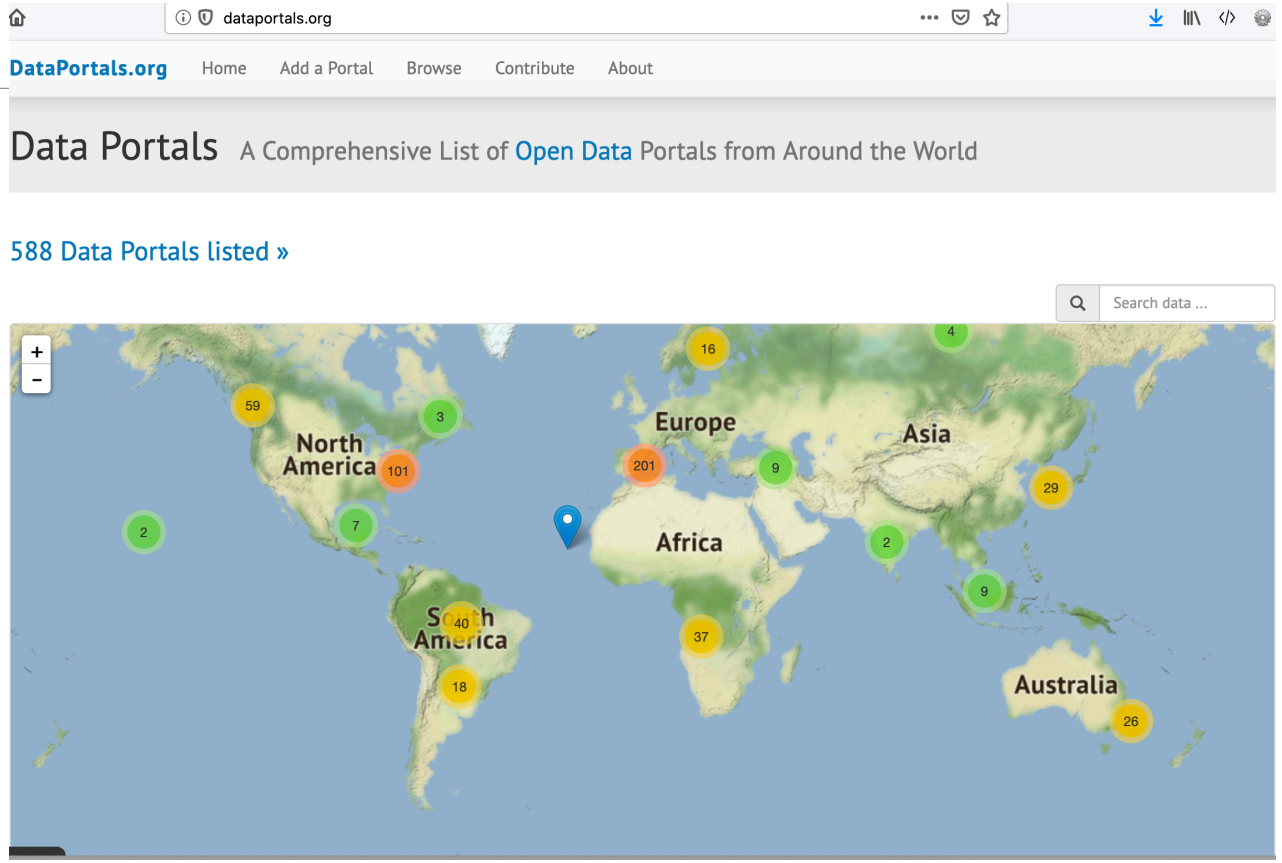


India

~600 Data Catalogs of Open Data



Source: IBM Institute for Business Value.



- 588 catalogs as on 20 June 2020
- 590 catalogs as on 24 Aug 2021

Demo and Exercise: US Open Data

- Site: https://data.gov

- Tools: https://resources.data.gov/categories/data-tools/

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

Volume

Variety

Velocity

Veracity

…

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

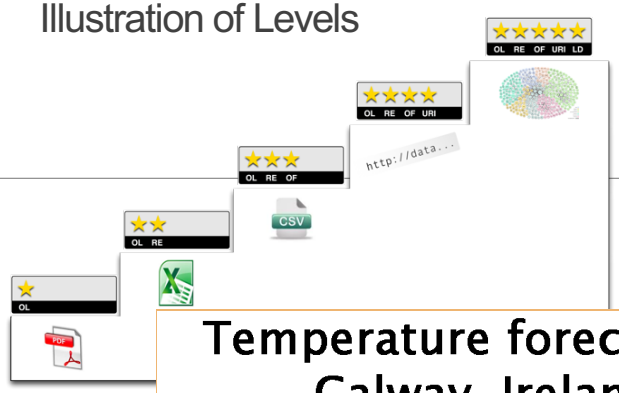
Cartoon critical of big data application, by T. Gregorius.

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

Quality of Data

Does Opening Data Make It Reusable? No

Illustration of Levels



5

Temperature forecast for Galway, Ireland

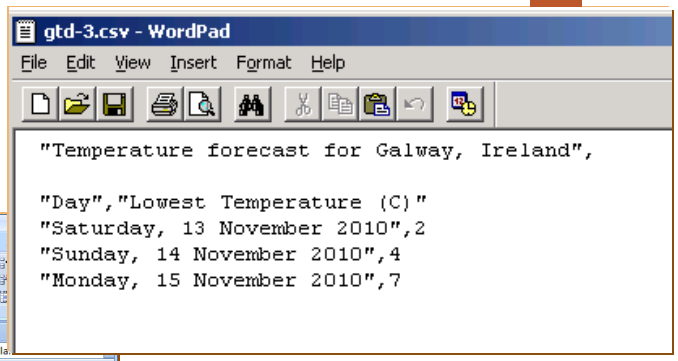
Day	Lowest Temperature (°C)
Saturday, 13 November 2010	2
Sunday, 14 November 2010	4
Monday, 15 November 2010	7

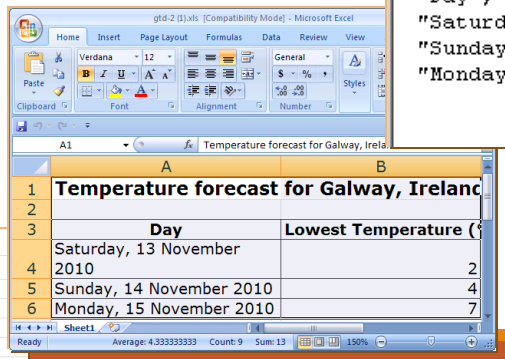
Temperature forecast for Galway, Ireland

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

4

3





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

2

1

Accessing Data

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

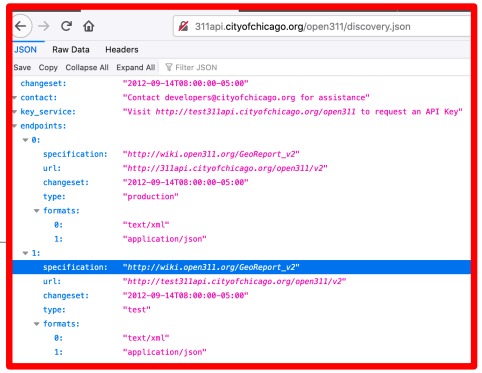
Refers to non-emergency events like graffiti, garbage, down trees, abandoned car, …

- 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"



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

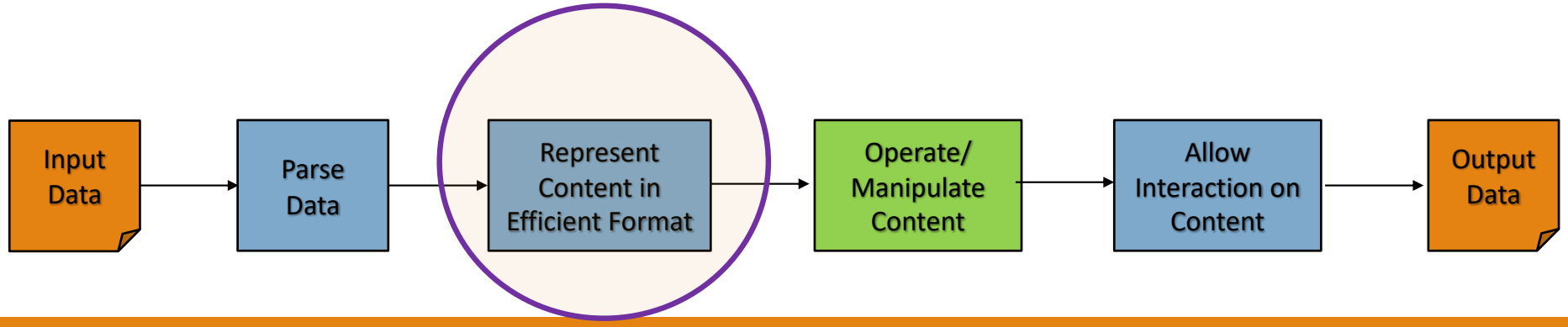
{"service_code":"4ffa4c69601827691b000018",
"attributes":
[{"variable":true,"code":"FQSKA1",
"datatype":"singlevaluelist","required":false,"order":1,
"description":"Vehicle Make/Model",
"values":
[{"key":"ASVEAV","name":"(Assembled From Parts,Homemade)"},
{"key":"HOMDCYL","name":"(Homemade Motorcycle, Moped.Etc.)"},
{"key":"HMDETL","name":"(Homemade Trailer)"}, …]
…]}

Code Time: Processing Open311 Data

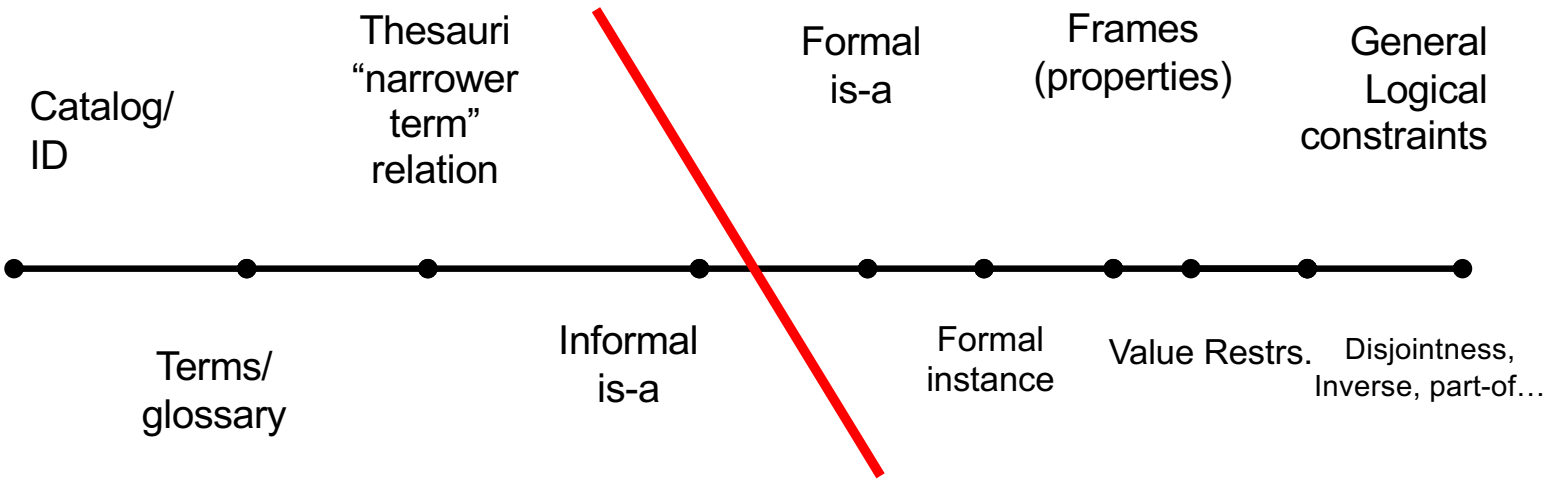
https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l2-opendata/Explore%20OpenData.ipynb

Data (Internal) Representation

A Simplified Data Analysis Process



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

Lecture 3: Concluding Comments

- We explored data in detail
  - Structure
  - Source
  - Quality and access
- Understood why data is crucial to good AI results
  - Representation

Concluding Segment

Concluding Discussion: Course Project

- **Thoughts on project idea**

- A **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
  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 works? (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

About Next Lecture – Lecture 4

Lecture 4: Supervised Learning / Structured Data

- Introduction to Machine Learning
- Methods and tools
- Data preparation