Open Data

Presentation: Jason
Discussion: Yingfeng

Paper: Renee Miller. Open Data Integration. VLDB 2018. 2130-2139.

Open Data Integration

What is open data?

- openly accessible
- easy to access
- freely available
- machine-readable

History

1940s - Robert K. Merton

- a founding father of modern sociology
- research data should be free to all for the common good [1]

History

Movements in open-source, open science, and government transparency

History

1995 - term "open data" first used in report from National Research Council

- "called for making environmental data available to the public so that scientists could study the global environment that transcends borders." [2][3]

History

2007 - group of "open-data pioneers" [4]

- including Larry Lessig (founder of Creative Commons, 2001)
- data should be complete, primary, timely, accessible, machine-processable, nondiscriminatory, nonproprietary and license-free
- (the paper mentions the first three)

Discussion (in pairs)

- Open data is really helpful for data scientists. However, what potential risks/issues will open data cause? What sort of data should be open? What sort of data needs to be discreetly disclosed or kept private?

(From Michael, Ehsan)

History

1940s - Robert K. Merton

- a founding father of modern sociology
- research data should be free to all for the common good [1]

Movements in open-source, open science, and government transparency

1995 - term "open data" first used in report from National Research Council

- "called for making environmental data available to the public so that scientists could study the global environment that transcends borders." [2]

2007 - group of "open-data pioneers"

[redacted]

The Problem

The Problem

- discoverability
- finding data that suitable

The Problem

Bad:

- raw formats
  - CSV
  - JSON
  - relational
  - XML
  - plain text

- no descriptors, no schemas
  - data is "open" (box is ticked)

The Problem

Better:

- schema available
- suitable tagging (descriptors of the data)

Still might not be in a compatible format

On to the paper…

Data Science Examples

Joinability

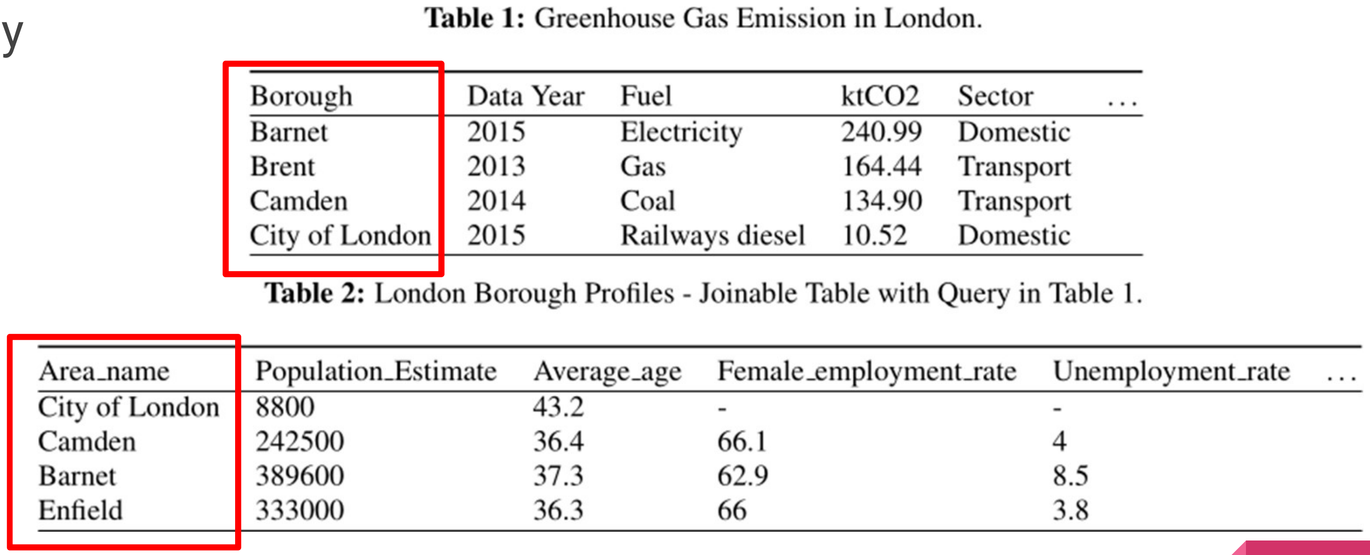


Table 1: Greenhouse Gas Emission in London.

Borough	Data Year	Fuel	ktCO2	Sector	...
Barnet	2015	Electricity	240.99	Domestic	
Brent	2013	Gas	164.44	Transport	
Camden	2014	Coal	134.90	Transport	
City of London	2015	Railways diesel	10.52	Domestic	

Table 2: London Borough Profiles - Joinable Table with Query in Table 1.

Area_name	Population_Estimate	Average_age	Female_employment_rate	Unemployment_rate	...
City of London	8800	43.2	-	-	
Camden	242500	36.4	66.1	4	
Barnet	389600	37.3	62.9	8.5	
Enfield	333000	36.3	66	3.8	

Data Science Examples

Unionability

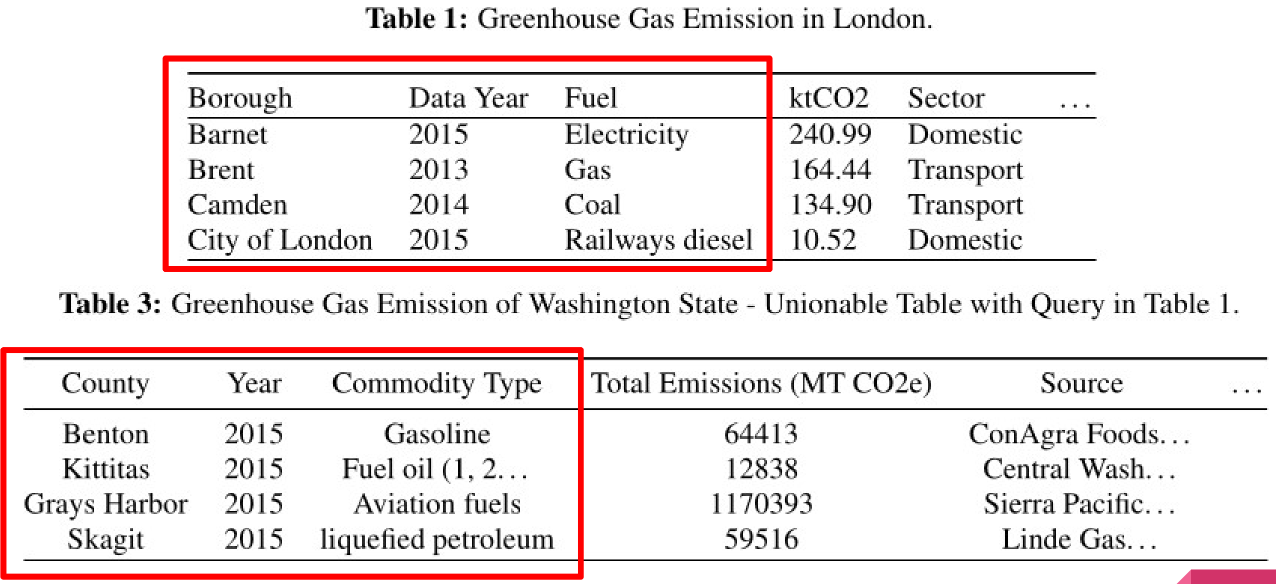


Table 1: Greenhouse Gas Emission in London.

Borough	Data Year	Fuel	ktCO2	Sector	...
Barnet	2015	Electricity	240.99	Domestic	
Brent	2013	Gas	164.44	Transport	
Camden	2014	Coal	134.90	Transport	
City of London	2015	Railways diesel	10.52	Domestic	

Table 3: Greenhouse Gas Emission of Washington State - Unionable Table with Query in Table 1.

County	Year	Commodity Type	Total Emissions (MT CO2e)	Source	...
Benton	2015	Gasoline	64413	ConAgra Foods...	
Kittitas	2015	Fuel oil (1, 2...	12838	Central Wash...	
Grays Harbor	2015	Aviation fuels	1170393	Sierra Pacific...	
Skagit	2015	liquefied petroleum	59516	Linde Gas...	

Data Science Examples

Ontologies!!

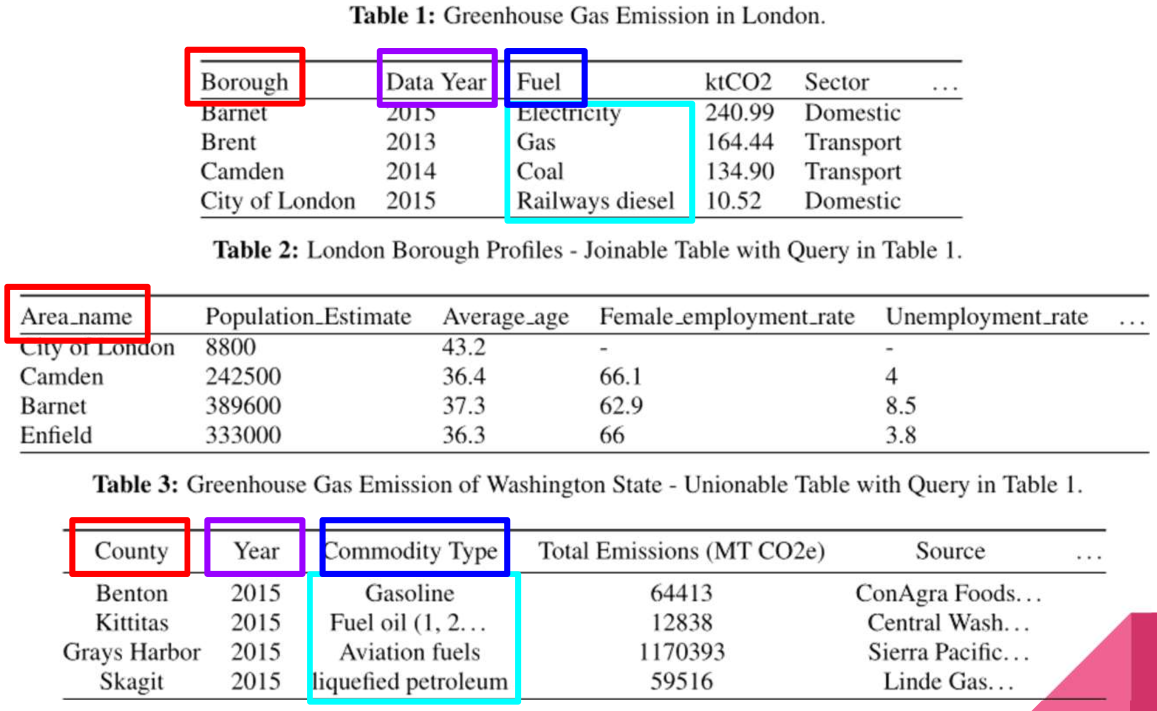


Table 1: Greenhouse Gas Emission in London.

Borough	Data Year	Fuel	ktCO2	Sector	...
Barnet	2015	Electricity	240.99	Domestic	
Brent	2013	Gas	164.44	Transport	
Camden	2014	Coal	134.90	Transport	
City of London	2015	Railways diesel	10.52	Domestic	

Table 2: London Borough Profiles - Joinable Table with Query in Table 1.

Area_name	Population_Estimate	Average_age	Female_employment_rate	Unemployment_rate	...
City of London	8800	43.2	-	-	
Camden	242500	36.4	66.1	4	
Barnet	389600	37.3	62.9	8.5	
Enfield	333000	36.3	66	3.8	

Table 3: Greenhouse Gas Emission of Washington State - Unionable Table with Query in Table 1.

County	Year	Commodity Type	Total Emissions (MT CO2e)	Source	...
Benton	2015	Gasoline	64413	ConAgra Foods...	
Kittitas	2015	Fuel oil (1, 2...	12838	Central Wash...	
Grays Harbor	2015	Aviation fuels	1170393	Sierra Pacific...	
Skagit	2015	liquefied petroleum	59516	Linde Gas...	

Data Integration for Data Science

- integration
- "Data analysis requires discovery of data that joins, unions, or aggregates with existing data in a precise way – a paradigm we call *query-driven data discovery*."
- The goal "...is to discover a query (or transformation) that translates data from one form into another."

Discussion (in pairs)

- What open data have you encountered in your life/study?
- How do you think open data integration will help your life/study? What are the challenges?

History - 1980s Data Federation

- combining small databases
- primarily within a single enterprise
- central control over the schema and mapping
- focus on
    - best global schema
    - data transformation
    - query execution across heterogeneous database systems

History - 2000s Data Exchange

- Internet -> sharing between autonomous systems
- owners retain full control of their data
- no longer necessary to have centralized or federated data
- about fitting source data with receiver's data
- known schemas
- focus on
  - best model of source data represented as target schema
  - core is schema mapping - "declarative representations of the relationship between two schemas"
  - finding joinable tables (known schemas)

History - 2020s Query-Driven Data Discovery

- shift to data science

- problem shift from integrating known data to finding the right data

Data Lakes

Data warehouses

- large amounts of structured data
- for business insights
- used by business people

Data Lakes

Data warehouses

- large amounts of structured data
- for business insights
- used by business people

Data lakes

- raw data
- may be structured, or not
- data models are created as needed
- requires specialized skills
  - data people such as data scientists
- better tools

Becoming more common

Data Lakes

Data warehouses

- large amounts of structured data
- for business insights
- used by business people

Data lakehouses

Data swamps

Data lakes

- raw data
- may be structured, or not
- data models are created as needed
- requires specialized skills
  - data people such as data scientists
- better tools

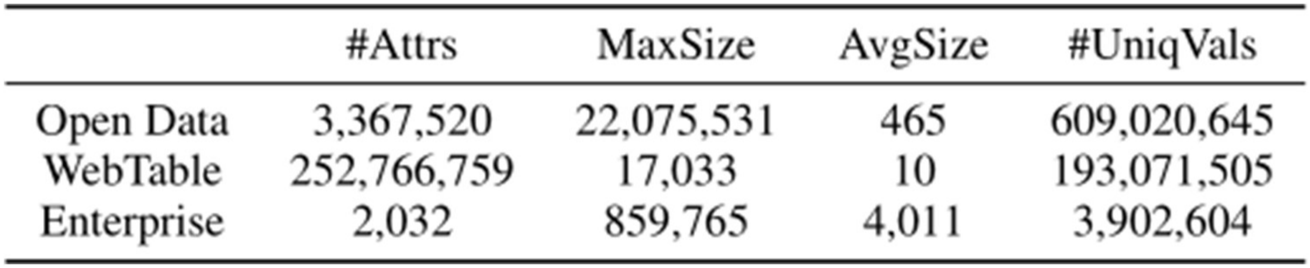
Becoming more common

Discussion (groups of 3-4)

- This paper claims that 1) machine learning may not be the desirable solution for data integration; 2) explaining integration and keeping humans in the loop are important. Do you agree with that? Why? (From Carol)

Open Data

- paper compares some open data sources



	#Attrs	MaxSize	AvgSize	#UniqVals
Open Data	3,367,520	22,075,531	465	609,020,645
WebTable	252,766,759	17,033	10	193,071,505
Enterprise	2,032	859,765	4,011	3,902,604

Open Data

- paper compares some open data sources

- experiments in open data
    - http://linkedct.org <- don't go there
    - https://github.com/oktie/linkedct

Open Data

- paper compares some open data sources
- experiments in open data
- monitoring open data availability such as from government

Open Data

- paper compares some open data sources
- experiments in open data
- monitoring open data availability such as from government entities
- observed 400% growth in open data over year to March 2017

Open Data

- paper compares some open data sources
- experiments in open data
- monitoring open data availability such as from government entities
- observed 400% growth in open data over year to March 2017
- *apparently open data growth stalled with the pandemic*

Mass Collaboration

- contribution by community members
  - WikiPedia
  - DBPedia
  - WikiData
  - WebTables
    - billions of html tables narrowed to millions containing structured data

The Modern Enterprise

- large investments in data warehouses
- integrating with data lakes
- too large for data scientists to fully understand
- pushing the limits of maintaining meta-data

- …research areas and future work

Future Work

- data discovery is a first step to data integration with data lakes
- don’t lose the lessons from data exchange and schema mapping
    - what goes around comes around…

Discussion (groups of 3-4)

- What’s the future of open data?
  - An important role in industry/academic/enterprise?
  - New research directions? (privacy, security, standard)
  - …

Group number 1, 2, 3, 4

Questions?

References

[1] “Open Data: A History,” *Data.gov*, Apr. 04, 2013. https://data.gov/blog/open-data-history/

[2] S. Badiee, J. Crowell, L. Noe, A. Pittman, C. Rudow, and E. Swanson, “Open data for official statistics: History, principles, and implementation,” *SJI*, vol. 37, no. 1, pp. 139–159, Mar. 2021, doi: 10.3233/SJI-200761.

[3] *On the Full and Open Exchange of Scientific Data*. Washington, D.C.: National Academies Press, 1995, p. 18769. doi: 10.17226/18769.

[4] “A brief history of open data,” *FCW*, Jun. 09, 2014. https://fcw.com/digital-government/2014/06/a-brief-history-of-open-data/255265/.

[5] “Open data,” *Wikipedia*. Mar. 16, 2023. Accessed: Mar. 28, 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Open_data&oldid=1144934646

[6] T. B. of C. Secretariat and T. B. S. of C. Open Government, “Open Data 101.” http://open.canada.ca/en/open-data-principles.